Supporting and Transforming High-Stakes Investigations with Expert-Led Crowdsourcing

Sukrit Venkatagiri

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science and Applications

Kurt Luther, Chair
Christopher L. North
Ha Rim Rho
Kate Starbird
Tanushree Mitra

September 30, 2022
Falls Church, Virginia

Keywords: Human–Computer Interaction, Crowdsourcing, Experts

Copyright 2022, Sukrit Venkatagiri
Supporting and Transforming High-Stakes Investigations with Expert-Led Crowdsourcing

Sukrit Venkatagiri

(ABSTRACT)

Expert investigators leverage their advanced skills and deep experience to solve complex investigations, but they face limits on their time and attention. In contrast, crowds of novices can be highly scalable and parallelizable, but lack expertise and may engage in vigilante behavior. In this dissertation, I introduce and evaluate the framework of expert-led crowdsourcing through three studies across two domains, journalism and law enforcement. First, through an ethnographic study of two law enforcement murder investigations, I uncover tensions in a real-world crowdsourced investigation and introduce the expert-led crowdsourcing framework. Second, I instantiate expert-led crowdsourcing in two collaboration systems: GroundTruth and CuriOSINTy. GroundTruth is focused on one specific investigative task, image geolocation. CuriOSINTy expands the flexibility and scope of expert-led crowdsourcing to handle more complex and multiple investigative tasks: identifying and debunking misinformation. Third, I introduce a framework for understanding how expert-led crowdsourced investigations work and how to better support them. Finally, I conclude with a discussion of how expert-led crowdsourcing enables experts and crowds to do more than either could alone, as well as how it can be generalized to other domains.
Supporting and Transforming High-Stakes Investigations with Expert-Led Crowdsourcing

Sukrit Venkatagiri

(GENERAL AUDIENCE ABSTRACT)

Expert investigators leverage their advanced skills and deep experience to solve complex investigations, but they face limits on their time and attention. In contrast, there is growing interest among non-professional members of the public to participate in investigations, but they lack the expertise or may engage in harmful behavior. In this dissertation, I introduce a new concept called, expert-led crowdsourcing, that allows professionals and non-professionals to work together on a high-stakes investigations in two domains: journalism and law enforcement. First, I explored how expert-led crowdsourcing played out in CrowdSolve, a real-world investigation of two decades-old murder cases. At CrowdSolve, over 250 amateur sleuths supported eight law enforcement experts to uncover new leads two for the two cases. Second, I build two software applications, GroundTruth and CuriOSINTy, to better support expert-led crowdsourced investigations. GroundTruth helps investigators work with a crowd to find the exact geographic location where a photo was taken. CuriOSINTy extends GroundTruth’s capabilities to help investigators with more complex and multiple investigative tasks involved in identifying and debunking misinformation on social media. Third, I compared and contrasted the three prior studies to develop a more detailed understanding of expert-led crowdsourced investigations and how to better support them. Finally, I conclude with a discussion of how expert-led crowdsourcing enables experts and crowds to do more than either could alone, as well as how it can be used in other professions.
Acknowledgments

Writing this dissertation about crowdsourcing was itself an exercise in crowdsourcing. It was also a successful one thanks to a wonderful set of mentors, colleagues, friends, and family.

First, I am eternally grateful to my advisor, Kurt Luther, for his endless feedback, encouragement, and support during my time at Virginia Tech. If I had not connected with him during his Investigative Technologies in Society class in my first semester, I would not be the researcher that I am today. He exemplifies many of the qualities that I admire most in an advisor and scholar: virtue, curiosity, patience, and kindness; as well as the highest standards for ethical conduct and research.

I am also deeply indebted to my dissertation committee, Kate Starbird, Chris North, Eugenia Rho, and Tanushree Mitra, were quick to provide feedback and ask challenging and impactful questions that helped improve my dissertation and my own thought process. I have also been fortunate to have been mentored by Siddharth Suri, Matthew Lease, Jacob Thebault-Spicker, and Laura Malkiewich, who have helped me to become a better researcher and colleague. Jake — the early years of my Ph.D. were shaped by your guidance and feedback, as well as the intellectually challenging discussions we have had on experimental design.

I truly appreciate the dozens of participants and hundreds of crowd workers who have taken part in my research. Thank you for working with me, and for making this research possible. I hope my work has, in some small way, improved yours as well.
A dissertation about sociotechnical systems would be incomplete without acknowledging the immense role that these systems have played in this dissertation. Thank you, to all of the applications, tools, and websites, without which conducting my research projects and writing this dissertation would have been an insurmountable task. Relatedly, I want to acknowledge the hundreds if not thousands of hours that the developers of these systems have invested — thank you!

I also want to acknowledge two sources of financial support: a CAREER award from the U.S. National Science Foundation and a Commonwealth Cyber Initiative grant from Virginia. They provided me with the freedom and peace of mind, allowing me to focus on research, mentorship, and service.

I want to acknowledge the immense amount of support I have received from the administrative staff at Virginia Tech. Roxanne Paul, Jessica Mullins, Sharon Kinder-Potter, Melanie Darden, Teresa Hall, and Robert Marcum — thank you for helping me overcome the logistical hurdles involved in finishing a Ph.D., including a 250-mile move from Blacksburg to the D.C.-area and as a global pandemic.

Special thanks to my labmates at the Crowd Intelligence Lab, including Vikram Mohanty, Yasmine Belghith, Tianyi Li, Anirban Mukhopadhyay, Nai-Ching Wang, Joey Yu, Aditya Bharadwaj, Rifat Sabbir Mansur, and Lee Lisle. All of you have shaped my research, provided a sympathetic ear, and been a friendly face I could turn to on campus or at conferences, no matter the time of day. I would not have gotten through my Ph.D. without all of you.

The friends that I had coming into Virginia Tech as well as those that I made along the way at my internships and conferences: all of you have given me inspiration and compan-
ionship no matter where I was (or will be). Thank you: Aakash Gautam, Varsha Koushik, Ashish Baghudana, Suprita Patil, Zachary Levonian, Dave Holtz, Martin Riedl, Timothy Stelter, Azeez Humayun, Melanie Trammell, Akshita Jha, Nurendra Choudhary, Nikhil Muralidhar, Sathappan Muthiah, Debanjan Datta, Ryan Evans, Jordan Huffaker, Jane Im, Sai Gouravajhala, Penny Trieu, Lindsay Blackwell, Harmanpreet Kaur, Alex Williams, Joon Sung Park, Amy Zhang, Himanshu Zade, Lindah Kotut, Rachel Moran, Maddie Jalbert, Morgan Wack, Andrew Beers, Shruti Phadke, Tiffany Anne, Xiao Ma, Patrick Carrington, Haley MacLeod, Varun Murthy, Vaishak Vellore, Zaenab Fatima, Shweta Sridhar, Reetika Reddy, and Priyanka Agrawal.

Finally, thank you to my family for their endless love and support, and for shaping me into the person I am today. My parents, Seema and Sirish, sacrificed so much to provide a better life for me and my sister. I am forever grateful for all that you do for us. Thank you to my sister, Saiesha, as well as my extended family, Poornima, Sridhar, Sunny, Deepika, Mayra, Shankar, Sangitha, Malaika, and Siddharth for your encouragement. Knowing that you were all there for my parents, and for each other, kept me sane despite being 9,000 miles away with a 10-hour time difference. I owe my success to all of you.

Tim, I truly appreciate your infinite patience as I wrote my dissertation — I finished it! I could not have thrived in graduate school during a global pandemic without your constant encouragement, immeasurable support, and endless supply of cookies (as well as your occasional insistence that I was spending too much time on writing and not getting enough exercise). Thank you for you being the best friend, fellow foodie, travel companion, debate opponent, copy editor, sounding board, co-author, and partner that I could have wished for.
Contents

List of Figures xv

List of Tables xvii

1 Introduction 1

1.1 Research Questions and Outline 3

1.1.1 Study 1: Identify Challenges and Opportunities in a Co-located Expert-led Crowdsourced Investigation 3

1.1.2 Study 2: Design and Evaluate a Software Tool to Support One Type of Investigative Task 4

1.1.3 Study 3: Design and Evaluate a Software Tool to Support a Broader Range of Investigative Tasks 5

1.1.4 Study 4: Synthesize and Explain the Expert-Led Crowdsourcing Framework 5

1.2 Outline of This Proposal 7

1.2.1 Statement of Contributions 9

2 Review of Literature 10

2.1 Investigations as Sensemaking 10

2.1.1 The Sensemaking Loop 10
3.3 RQ1. Organizing CrowdSolve ............................................. 36
  3.3.1 Organizers’ Goals for CrowdSolve ................................. 37
  3.3.2 Choosing a Case to Investigate .................................. 37
  3.3.3 Putting It All Together ........................................... 39
  3.3.4 The Event ......................................................... 43
3.4 RQ2. Other Stakeholders’ Motives and Definitions of Success ...... 45
  3.4.1 Law Enforcement Experts and TCPD ............................. 46
  3.4.2 Families ......................................................... 47
  3.4.3 Attendees ......................................................... 48
3.5 RQ3. Interactions and Challenges Between the Stakeholders ...... 50
  3.5.1 Interactions Between Attendees and Experts .................... 50
  3.5.2 Interactions Among the Attendees ............................... 55
  3.5.3 Interactions Between Attendees and the Families ............... 59
3.6 RQ4. Managing Tensions Between Stakeholders ..................... 62
  3.6.1 Experts vs. Novice Crowds ..................................... 64
  3.6.2 Security and Privacy vs. Openness ............................... 65
  3.6.3 Entertainment vs. Reality ....................................... 69

4 GroundTruth: Augmenting Expert Image Geolocation through Shared Representations and Crowdsourcing 72
4.1 Motivation .............................................................. 72
4.2 RQ5. Designing GroundTruth ........................................ 75
  4.2.1 Expert Image Geolocation Workflows Today ............... 75
  4.2.2 Design Rationale for Shared Representations ............... 76
  4.2.3 System Description and Scenario ............................. 78
  4.2.4 Expert Interface .................................................. 79
  4.2.5 Crowd Worker Interface ................................. 81
  4.2.6 Implementation Details ......................................... 83

4.3 Methods ................................................................. 84
  4.3.1 Image Dataset Selection ........................................ 85
  4.3.2 Expert Recruitment .............................................. 85
  4.3.3 Crowd Worker Recruitment ..................................... 86
  4.3.4 Procedure ........................................................ 87
  4.3.5 Data Collection and Analysis .................................. 89
  4.3.6 Limitations ........................................................ 90

  4.4.1 Image Difficulty ................................................... 91
  4.4.2 Overall Performance ............................................. 92
  4.4.3 Effect of Aerial Diagram on Crowd Performance ............ 93
  4.4.4 Effect of Heatmap on Expert Performance .................... 94

4.5 RQ7. What Were Experts’ Experiences of Using GroundTruth? .... 96
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5.1</td>
<td>Aerial Diagram</td>
<td>97</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Gridded Map Overlay</td>
<td>100</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Heatmap of Crowd Feedback</td>
<td>103</td>
</tr>
<tr>
<td>4.6</td>
<td>Discussion</td>
<td>107</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Designing Shared Representations for Image Geolocation</td>
<td>107</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Generalizing Shared Representations</td>
<td>111</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Broader Impacts</td>
<td>112</td>
</tr>
<tr>
<td>5</td>
<td>CuriOSINTy: Designing a Platform for Combating Misinformation through Collaborative Capture the Flag Competitions</td>
<td>114</td>
</tr>
<tr>
<td>5.1</td>
<td>Motivation</td>
<td>114</td>
</tr>
<tr>
<td>5.2</td>
<td>Background: Combating Mis- and Disinformation Through Open Source Intelligence</td>
<td>119</td>
</tr>
<tr>
<td>5.3</td>
<td>RQ8. Developing CuriOSINTy Using Research Through Design</td>
<td>121</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Phase 1: Framing the Problem</td>
<td>122</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Phase 2: Prototyping Workflows and Interfaces</td>
<td>124</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Phase 3: Clarify Design Goals</td>
<td>127</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Phase 4: Deploy and Iterate on Design</td>
<td>130</td>
</tr>
<tr>
<td>5.4</td>
<td>The CuriOSINTy CoCTF Platform</td>
<td>135</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Team Member Interface</td>
<td>136</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Staff Interface</td>
<td>139</td>
</tr>
</tbody>
</table>
5.5 RQ9. Students’ Experiences ................................................. 141
  5.5.1 Participants’ Backgrounds, Motives, and Definitions of Success ... 141
  5.5.2 Students’ Initial Reactions to OSINT CoCTFs .......................... 143
  5.5.3 Collaboration Styles During the CoCTFs ............................. 145
  5.5.4 CuriOSINTy Was Enjoyable to Use and Helped to Better Structure Investigations ................................................. 148

5.6 RQ10. Students’ Performance ........................................... 151
  5.6.1 Point System Was Effective But Revealed New Tensions ......... 151
  5.6.2 Judging Improved Quality But May Be Difficult to Scale ....... 168
  5.6.3 API Enabled Appropriation ........................................... 172

5.7 Discussion ................................................................. 176
  5.7.1 Designing with Appropriation in Mind ............................. 176
  5.7.2 Learning From Appropriation: Evaluating CuriOSINTy Against Our Design Goals ................................................. 177

5.8 Summary ................................................................. 181

6 A Framework for Expert-Led Crowdsourcing in High-Stakes Investigations 182
  6.1 Motivation ................................................................. 182
  6.2 Case Studies ............................................................. 184
    6.2.1 Defining Expert-Led Crowdsourcing ............................... 185
  6.3 Designing the Expert-Led Crowdsourcing Framework ................. 187
7.2.1 Give the Crowd (More) Agency ........................................ 262
7.3 Incorporate AI-Infused Systems ........................................ 265
7.4 Consider Long-Term Software Maintenance and Support ............ 267

8 Conclusion ................................................................. 269
8.1 Contributions ......................................................... 269
  8.1.1 Introduced the Concept of Expert-Led Crowdsourcing ............ 270
  8.1.2 Identified How to Manage the Sociotechnical Challenges in ELC Investigations ......................................................... 270
  8.1.3 Developed ELC Systems to Support a Range of Complex Investigative Tasks ......................................................... 271
  8.1.4 Examined How Expert-Led Crowdsourcing Works to Develop an ELC Framework ......................................................... 272

Bibliography ............................................................... 274
List of Figures

1.1 Overview of Work Completed ........................................ 7
2.1 The Sensemaking Loop: Investigations as Sensemaking .......... 11
3.1 CrowdSolve Schedule of Events ....................................... 45
3.2 CrowdSolve Attendees and Group Work Sessions ................... 50
4.1 Expert Interface in GroundTruth ....................................... 79
4.2 Crowd Worker Interface in GroundTruth ............................. 82
4.3 Participants in the GroundTruth Evaluation ......................... 87
4.4 GroundTruth Expert and Crowd Session and Performance Details ... 92
5.1 Research through Design (RtD) Process ............................ 122
5.2 Participant Table ......................................................... 132
5.3 The structure of the CuriOSINTy system. ............................ 135
5.4 Number of Evidence Pieces and Flags ............................... 153
5.5 Team Rankings for Each Event ....................................... 156
5.6 Percent Collaboration and Tasks ..................................... 158
6.1 Expert-Led Crowdsourcing Framework .............................. 203
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2 Stakeholder Groups</td>
<td>211</td>
</tr>
<tr>
<td>6.3 Crowd Worker Motivations</td>
<td>217</td>
</tr>
<tr>
<td>6.4 Crowd Worker Availability and Turnover</td>
<td>220</td>
</tr>
<tr>
<td>6.5 Spectrum of Coordinated Action</td>
<td>222</td>
</tr>
<tr>
<td>6.6 Security, Privacy, Control, and Agency</td>
<td>226</td>
</tr>
<tr>
<td>6.7 Workflows</td>
<td>230</td>
</tr>
<tr>
<td>6.8 Duration, Synchronicity, Co-locatedness, and Communication</td>
<td>235</td>
</tr>
<tr>
<td>6.9 Sociotechnical Infrastructure</td>
<td>241</td>
</tr>
<tr>
<td>6.10 Articulation Work of Expert-Led Crowdsourcing</td>
<td>249</td>
</tr>
<tr>
<td>7.1 Expert-Led Crowdsourcing vs. Top-Down and Bottom-Up Crowdsourcing</td>
<td>259</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Participants in the CrowdSolve Study</td>
<td>35</td>
</tr>
<tr>
<td>6.1</td>
<td>The expert-led crowdsourcing framework extends and modifies Malone’s CI genome.</td>
<td>190</td>
</tr>
<tr>
<td>6.2</td>
<td>The expert-led crowdsourcing framework incorporates “how” elements of Starbird’s crowd work continuum framework.</td>
<td>194</td>
</tr>
<tr>
<td>6.3</td>
<td>Incorporating Liu’s framework into ELC</td>
<td>198</td>
</tr>
<tr>
<td>6.4</td>
<td>Incorporating Lee and Paine’s framework into ELC</td>
<td>200</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Investigations are broadly defined as the systematic process of collecting and schematizing disparate pieces of information to arrive at a theory or conclusion. They are among the most difficult yet impactful undertakings. Successful investigations can bring about significant real-world impact: catching criminals and finding missing persons, uncovering fraudulent or discriminatory practices in governments and corporations, and combating false narratives and correcting history. Investigations can be so fascinating that the “true crime” genre is among the most popular form of entertainment today [16]. Yet, investigators face formidable analytical and sociotechnical challenges, and many investigations go unsolved or lead to incorrect conclusions. For example, over 6,000 homicides go unsolved every year [12] and journalists are often misled by false evidence [295].

Professional investigators in fields such as journalists, law enforcement detectives, and intelligence analysts bring substantial training and experience to bear in solving high-stakes investigations. Professionals must trace and verify the provenance of every piece of evidence, corroborate facts using a diverse set of tools, and trace the narratives of dozens of individuals [146, 286]. As a result, their investigations may fail because they prioritize the wrong leads or miss important clues. These labor-intensive tasks also do not scale easily, meaning that successful investigations are limited by individual experts’ time, resources, or breadth of expertise [22, 81].

To overcome these challenges, investigators often turn to the public for help. This includes
early “wanted” posters [9], to the television show, *America’s Most Wanted*, responsible for capturing over 1,000 fugitives based on viewers’ tips. When Jim Gray, a computer scientist, went missing in 2007, the U.S. Coast Guard worked with researchers at Amazon, Microsoft, and NASA to lead a crowdsourcing effort on Amazon Mechanical Turk to search through thousands of satellite images for Gray’s ship [8]. Similar efforts have successfully helped find missing persons [11] and support crisis response efforts [90].

Simultaneously, the emergence of new information and communication technologies, including online forums and powerful, free investigative tools (e.g., satellite imagery, face recognition), have given rise to crowdsourced investigations organized by members of the public [315]. These investigations seldom involve professional investigators [152, 243, 310]. Novices among the public seeking to become more involved in professional investigations have clashed with professionals’ desires for greater privacy, further complicating their interactions [184]. Often a lack of skills, competing norms and values, and coordination challenges have resulted in failure and real-world harm [243, 316]. This includes misidentifying and innocent individuals as suspects in the 2013 Boston Marathon bombing [243, 294] and the 2020 George Floyd protests [194]. These are not isolated incidents, however. Numerous incidents of cyberstalking and doxxing in crowdsourced investigations continue to occur [66, 69, 84, 194, 210].

These successes and failures point to an urgent need to more deeply understand the challenges and opportunities that crowdsourced investigations present. Despite law enforcement agencies’ and the media’s discouragement [184, 243, 315], crowdsourced investigations will continue to emerge and evolve [67, 316]. Thus, there is opportunity to develop sociotechnical systems that enable more ethical, effective, and efficient crowdsourced investigations by involving professional investigators.

In this dissertation, I investigate how the design of technology can support and transform
crowdsourced investigations to support professional investigators with high-stakes investigations in journalism and law enforcement. I design and evaluate a framework that leverages the complementary strengths of professional investigators and novice crowds, which I call the expert-led crowdsourcing framework. This hybrid approach benefits from the domain knowledge and expertise of professionals and the speed and large scale of crowds.

This work makes the following four contributions: 1) a detailed understanding of sociotechnical and ethical challenges in crowdsourced investigations involving professionals, 2) two systems to enable efficient and effective coordination between professionals and crowds, 3) an evaluation of the effectiveness of these two systems, and 4) a framework that compares and contrasts expert-led crowdsourcing across three different settings.

1.1 Research Questions and Outline

Below, I describe the four studies that comprise this dissertation and the research questions that each of them addresses:

1.1.1 Study 1: Identify Challenges and Opportunities in a Co-located Expert-led Crowdsourced Investigation

To understand challenges and opportunities in an expert-led crowdsourced investigation, I first conducted an ethnographic study of CrowdSolve, a real-world crowdsourcing event [322]. At CrowdSolve, over 250 true crime enthusiasts received training from and were guided by eight professional investigators to generate new leads for two decades-old murder investigations. To design the initial expert-led crowdsourcing (ELC) framework, I framed the following research questions, explored in detail in Chapter 3:
• RQ1: What is involved in organizing a co-located ELC investigation?

• RQ2: How is success defined in a co-located ELC investigation?

• RQ3: What are the benefits of and challenges in a co-located ELC investigation?

• RQ4: How are the challenges managed in a co-located ELC investigation?

1.1.2 Study 2: Design and Evaluate a Software Tool to Support
One Type of Investigative Task

After understanding the challenges faced in an ELC investigation and how they are managed, I instantiated the expert-led crowdsourcing framework in a sociotechnical system. I designed and developed a software tool called GroundTruth to enable expert investigators to leverage the abilities of distributed, online crowds [321]. GroundTruth focused on a specific investigative task — image geolocation — and the research questions are oriented around how to design GroundTruth, how experts and crowds perform together, and what experts’ experiences are of using GroundTruth. I explored the following research questions, described in detail in Chapter 4:

• RQ5: How can we design a system to enable experts and crowds to work together on a specific investigative task?

• RQ6: How well do experts and novice crowds perform?

• RQ7: What are experts’ experiences?
1.1.3 Study 3: Design and Evaluate a Software Tool to Support a Broader Range of Investigative Tasks

GroundTruth successfully enabled experts to work with crowds to geolocate images. Next, I sought to expand the scope of the expert-led crowdsourcing framework to support greater flexibility in and variety of investigative tasks. I also sought to use a different type of motivation (enjoyment-based instead of immediate-payoff), and focused on a different coordination style (competition instead of collaboration). I built CuriOSINTy to support capture the flag competitions that allows teams to rapidly identify and debunk misinformation on social media [324]. CuriOSINTy also introduces collaborative features to mitigate the challenges of competition, such as duplication of effort and siloed information. I conducted a deployment and evaluation of CuriOSINTy in a semester-long class with 46 students. I explored the following research questions, described in detail in Chapter 5:

- RQ8: How can we design a system to enable experts and a trained crowds to work together on a broader set of investigative tasks?
- RQ9: What are crowd worker’s experiences?
- RQ10: How well does a trained crowd perform?

1.1.4 Study 4: Synthesize and Explain the Expert-Led Crowdsourcing Framework

In this dissertation, I introduced and expanded the scope of expert-led crowdsourcing. First, in CrowdSolve, I studied the sociotechnical and ethical challenges posed by ELC investigations in a high-stakes setting: a law enforcement investigation of two murder cases. Second,
to address the expert–crowd coordination challenges that I found in CrowdSolve, I introduced the concept of shared representations within ELC. I then instantiated shared representations in GroundTruth, an ELC system to enable efficient and effective coordination between experts and crowds for one complex task: image geolocation. Third, I studied how to support greater flexibility in and variety of ELC tasks through CuriOSINTy. I also explored a different way to motivate the crowd — collaborative competition — versus altruism in CrowdSolve and monetary payment in GroundTruth. Collectively, these studies describe beneficial yet complex collaborations between experts and crowds.

Having studied ELC investigations in these three different settings, what is missing is a framework to show how expert-led crowdsourced investigations work in high-stakes investigations. I thus explored the following questions:

- **RQ11:** How do expert-led crowdsourced investigations work and what types of investigations can it be applied to?
- **RQ12:** Who are the stakeholders that need to be brought in, what motivates them, and how can they contribute?
- **RQ13:** What is the nature of the collaboration between stakeholders, and what factors determine the success (or failure) of an investigation?
- **RQ14:** What sociotechnical infrastructure is needed to facilitate this collaboration?

To answer these questions, I first highlight the applicability of existing crowdsourcing frameworks [106, 209, 221, 292] in the context of expert-crowd collaboration in high-stakes settings. Second, I summarize the findings and implications from my three prior studies. Third, I present a conceptual framework for expert-led crowdsourced investigations that compares and contrasts these three studies, and contend that expert-led crowdsourcing allows experts
and crowds to do more than either could by themselves. I describe this in more detail in Chapter 6:

## 1.2 Outline of This Proposal

This dissertation is structured as follows:

<table>
<thead>
<tr>
<th>Milestones</th>
<th>Research Questions</th>
<th>Study Details</th>
<th>Status</th>
</tr>
</thead>
</table>
| Study 1 (Ch. 3): Initial empirical work to generate framework. | • RQ1: What is involved in organizing a co-located ELC investigation?  
• RQ2: How is success defined in a co-located ELC investigation?  
• RQ3: What are the benefits and challenges of a co-located ELC?  
• RQ4: How are the challenges managed in a co-located ELC investigation? | CrowdSolve. Ethnographic study of a 250+ person crowdsourcing event, focusing on the challenges of organizing the event and the coordination challenges that professionals and crowds faced. Major challenges include structuring the event to ensure information security and effectiveness, competing norms and values among participants, and difficulties scaling up leadership and expertise. These challenges were managed through four regulatory mechanisms. | Published at ACM CSCW 2021 [1]. |
| Study 2 (Ch. 4): System design and evaluation to implement framework. | • RQ5: How can we design a system to enable experts and crowds to work together on a specific investigative task?  
• RQ6: How well do experts and novice crowds perform?  
• RQ7: What are experts’ experiences? | GroundTruth. Iterative design of GroundTruth, a web-based crowdsourcing tool to support expert-crowd collaboration for a specific investigative task. Describes how design implications from related work influenced specific features in GroundTruth. Summarizes the process of development, participant recruitment, and mixed-methods lab-based evaluation. | Published at ACM CSCW 2019 [4]. |
| Study 3 (Ch. 5): System design and deployment to expand scope of framework. | • RQ8: How can we design a system to enable experts and a trained crowds to work together on a broader set of investigative tasks?  
• RQ9: How well does a trained crowd perform?  
• RQ10: What are crowd worker’s experiences? | CuriOSINTy. Iterative design and evaluation of CuriOSINTy, a web-based software tool to support a broader range of investigative tasks through capture-the-flag competitions. Describes how design implications from prior and related work influenced specific features in CuriOSINTy. Summarizes the process of development, participant recruitment, and mixed-methods evaluation in a classroom deployment. | Completed. To be submitted to ACM CHI 2023 on 9/16/22. |
| Study 4 (Ch. 6): Synthesize takeaways from prior three studies to explain the framework. | • RQ11: How do expert-led crowdsourced investigations work and what types of investigations can it be applied to?  
• RQ12: Who are the stakeholders that need to be brought in, what motivates them, and how can they contribute?  
• RQ13: What is the nature of the collaboration between stakeholders, and what factors determine the success (or failure) of an investigation?  
• RQ14: What sociotechnical infrastructure is needed to facilitate this collaboration? | ELC Framework. Synthesizes prior studies (Studies 1-3) to present the ELC framework for investigations. Framework compares and contrasts these studies, with a focus on the investigative domain, stakeholder groups involved, sociotechnical infrastructure required, the nature of the coordinated action between stakeholders, and the articulation work required to enable the coordinated action. | Completed. To be submitted to ACM TOCHI or ACM CSCW 2023. |

Figure 1.1: Overview of work completed.

### Chapter 1: Introduction

Motivation for the thesis work, research questions organized in four milestones, this outline, and the timeline for completing the proposed work.

### Chapter 2: Review of the Literature
First describes how investigations can be viewed as sensemaking. Then reviews prior research on crowdsourced sensemaking and associated challenges. Next discusses the two types of crowdsourced investigations: bottom-up and top-down, and the benefits and challenges of each. Next, provides an overview of approaches for motivation crowds. Finally reviews prior research on designing systems to coordinate complex work.

**Chapter 3: CrowdSolve: Managing Tensions in an Expert-Led Crowdsourced Investigation**

Details an ethnographic study of a real-world crowdsourcing event [322]. Describes findings, such as efforts required to organize the event, stakeholder groups’ motives and definitions of success, challenges between stakeholder groups, and how these challenges were managed.

**Chapter 4: GroundTruth: Augmenting Expert Image Geolocation with Crowdsourcing and Shared Representations**

Details the design and evaluation of GroundTruth, a web-based software tool to enable experts to guide crowds in geolocating images in real-time [321]. Describes evaluation, experts’ and crowds’ performance, and experts’ experiences.

**Chapter 5: CuriOSINTy: Designing a Platform for Combating Misinformation through Collaborative Capture the Flag Competitions**

Details the design and evaluation of CuriOSINTy, a web-based collaborative capture the flag (CoCTF) platform to enable experts and novice crowds to quickly identify and debunk a large volume of potential misinformation on social media [324]. Describes evaluation, crowd workers’ performance, and crowd workers’ experiences.

**Chapter 6: A Framework for Expert-Led Crowdsourced Investigations** Synthesizes prior studies (Chapters 3-5) to present the ELC framework for investigations. Framework
1.2. **Outline of This Proposal**

compares and contrasts these studies, with a focus on the investigative domain, stakeholder groups involved, sociotechnical infrastructure required, the nature of the coordinated action between stakeholders, and the articulation work required to enable the coordinated action.

**Chapter 7: Discussion** I discuss how to apply and evaluate the expert-led crowdsourcing framework in other high-stakes settings, the ethical implications to consider, and avenues for future work.

**Chapter 8: Conclusion** I conclude with a summary of contributions that this work makes.

### 1.2.1 Statement of Contributions

This dissertation includes first-authored, peer-reviewed articles that have been previously published in Association for Computing Machinery (ACM) venues. The ACM’s policy on reuse of previously published articles is as follows:\(^1\):

> “Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included.”

The above paragraphs include references to first-authored articles that are previously published or under review [321, 322, 324].

\(^1\)https://authors.acm.org/author-resources/author-rights
Chapter 2

Review of Literature

2.1 Investigations as Sensemaking

Investigations involve a series of sensemaking tasks, in which the goal is to gather and analyze large amounts of diverse, unstructured information to arrive at a theory or conclusion [117, 257, 269]. Within CSCW and HCI, there exists a large body of work studying and supporting sensemaking in investigations [e.g., 117, 206, 255, 270, 296]. In this work, I focus on investigations in two domains: journalism and law enforcement, which are classic examples of sensemaking [81, 92, 94, 235, 257].

2.1.1 The Sensemaking Loop

One influential model by Pirolli and Card [257] characterizes sensemaking as a 16-step, iterative process with two key subloops, one focused on gathering, searching, and extracting information (foraging); the other on building a mental model that best fits the information (synthesis); with structure and effort increasing in later steps. Using the example of image geolocation, a common investigative task, I relate the steps in an investigation back to the sensemaking steps as defined by Pirolli and Card [257] (see Fig. 2.1). Half the steps (15, 12, 9, 5, 3) are top-down, moving from theory to data, while the other half (2, 5, 8, 11, 14) are bottom-up, creating a “dual search” cycle between adjacent steps.
2.1. Investigations as Sensemaking

Figure 2.1: The Sensemaking Loop: Investigations as Sensemaking. The orange dotted arrow indicates one possible set of steps that an investigator would undertake to debunk or verify visual media through image geolocation. I relate these steps to Pirolli and Card’s sensemaking loop for intelligence analysts [257] (inset). Replicated from [321] with author’s permission.

Experts in many investigative fields, such as journalism, human rights, and military intelligence, perform image geolocation as a key step in the broader task of verifying photos and videos shared on social media. The goal is to identify the precise location where the image was captured, to help support or refute claims about its provenance and meaning.

Applying this model to an image geolocation example, suppose an expert investigates a photo
that they encountered on social media purporting to show evidence of a violent protest in an unidentified German city. This claim may provide they with an initial top-down hypothesis. However, as they inspect the visual content, clues in the road signage and building architecture instead suggest Austria, providing a bottom-up challenge to the initial hypothesis. On one building, the expert notices a distinctive but unfamiliar logo (step 5: read and extract). They run a reverse image search (3: search for information) which yields a large number of results (1: external data sources). Sorting through these (2: search and filter), they find a match for a business with multiple offices in three cities in Austria (6: search for relations). They then draw an aerial diagram of the photo of interest (8: schematize) and manually search satellite imagery from each city (9: search for evidence), one at a time. Four hours later, they locate the matching building in one of the cities. This discovery, along with other contextual evidence (11: build case), allows them to debunk the original claim of Germany (14: tell story), based on which they writes a news report (16: presentation).

Successfully geolocating an image often requires experts to use a manual, brute-force process. Looking through large areas can take hours or even days, with no guarantee of success. Fatigue, coupled with the time-crunched nature of this work and limited resources, is a major challenge because constant vigilance is required to find a needle in a haystack.

Beyond image geolocation, investigators face constraints on their time and available resources for other sensemaking tasks. Investigators also face sociotechnical challenges [225], due to the changing nature of online platforms. Tools and data that investigators rely upon may become inoperable or inaccessible. New tools and data sources may also, emerge that investigators must quickly learn how to leverage.

There is some evidence that investigators seek assistance through crowdsourcing [81, 180, 343]). However, little technological support exists to mediate interactions between pro-
fessional investigators and crowds, and their success rates are unknown. To address this research gap, I explore how experts respond to crowdsourced support in investigations, with a focus on the entire sensemaking loop.

### 2.1.2 Scaling Up Investigations with Crowdsourced Sensemaking

Collaboration has the potential to speed up and scale up sensemaking by dividing up foraging and synthesis tasks, as well as providing multiple perspectives on schematizing and theorizing about connections, among other benefits [117]. However, collaboration also creates coordination challenges. For example, collaborators may have different skills and backgrounds, be geographically separated, need to externalize their thoughts for others, and have access to different parts of the dataset [116, 123, 137, 167].

Most of these efforts focus on small groups, i.e., dyads [137, 138, 139, 167] or triads [273, 341]. However, less work has explored scaling up collaborative sensemaking to a larger number of people, such as crowd workers, where coordination challenges are amplified [307]. **My work contributes to this growing body of research to support crowdsourced sensemaking.**

Some projects have focused on crowdsourcing specific sensemaking steps as microtasks, such as searching and filtering [251], reading and extracting [64], or schematizing [70, 178, 216]. Others crowdsourced the *entire sensemaking loop* to perform complex tasks like solving mysteries [205, 307] or writing articles [143, 175]. **Most of these efforts focus on how either novice crowds alone or crowd–AI hybrids can complete sensemaking tasks, my work differs by exploring how crowds can augment an expert’s sensemaking process.** Further, while the majority of crowd sensemaking research focuses on text data [e.g., 32, 64, 178, 216, 307? ], my work includes a focus on visual data.
2.2 Control in Crowdsourced Investigations

Trottier [315] posits that there are two primary investigative models: top-down crowdsourced investigations and bottom-up crowdsourced investigations. Both involve collecting, processing, and analyzing information. In law enforcement, the goal is an arrest or criminal proceedings [141], while in journalism the goal is a news article [319].

Top-down and bottom-up crowdsourced investigations differ along three primary dimensions: 1) which group is in control of the investigation, 2) what actions each group can perform, and 3) how information flows among the various groups. Below, we review prior studies of both models and draw parallels to citizen science, another type of crowdsourced investigation involving novices and experts.

2.2.1 Top-down Crowdsourced Investigations

Top-down investigations are the most common form of investigation. They are led and sanctioned by professionals, while the public plays a relatively passive role in generating leads as information sensors [122] and sources [81], rather than information processors. The flow of information is also highly restricted and unidirectional — from individual members of the public to professionals. Professionals actively collects information through open calls to the public, such as websites [319], tip hotlines [224], and community policing events [162, 204], or passively by monitoring activities online and offline [212].

Within HCI, Erete and Lewis [204] surfaced the importance of studying engagement between law enforcement professionals and the public in online and offline settings because the two formats may afford different types of interactions in both character and content. In another domain, Lalone et al. [184] studied how crisis-response experts and novices interacted on
Reddit. They found tensions between the experts’ desires for privacy and the public’s desire for information access, the rigid and hierarchical nature of expert investigations compared to the loose and easily reconfigurable public investigations, and what types of information each group considers relevant.

This top-down investigative model is similar to traditional citizen science projects where expert scientists direct novice volunteers to collect, categorize, and transcribe data, along with other basic tasks, while the scientists take on the advanced roles of data validation, synthesis, and analysis [192].

2.2.2 Bottom-up Crowdsourced Investigations

A second, more recent model enabled by the growth of networked digital technologies is the bottom-up crowdsourced investigation [315]. Here, crowds of non-professionals (i.e., the general public) take on a more active role in conducting their own investigations that do not involve — and may not be sanctioned by — professionals. Crowds coordinate their efforts to share, collect, and analyze information with others [293, 343] and even administer social and procedural justice [66, 151].

Although prior work has shown that crowds have the potential to scale up and speed up investigations [e.g., 81, 150, 321], undirected novice crowds may engage in potentially harmful behavior.

This includes conspiracizing about [231, 276] and misidentifying targeted individuals [243]. More dangerously, it can involve digilantism [316] — a combination of online doxxing (publicly releasing sensitive details such as the target’s home address) and embodied vigilantism (e.g., following a target or visiting their home). For example, crowds on 4chan tried to uncover a whistleblower associated with the impeachment trial of Donald Trump, even
misidentifying and harassing individuals online [310]. Because of these and other similar incidents [243], Huey et al. [158] found that law enforcement officials are increasingly wary of soliciting help from the public, further complicating interactions between the two groups.

A related trend has occurred within citizen science, where novices can now more deeply participate in scientific investigations through bottom-up data validation practices [336] and crowdsourced data analyses [311]. However, according to Law et al. [192], leaders of such projects must “negotiate perceptions amongst peers and superiors, constantly demonstrating that they [are] doing real science and not public service.” Similarly, crowdsourced investigations focused on crime may also need to establish their legitimacy in the eyes of law enforcement.

In this work, I propose a novel framework that combines bottom-up and top-down crowdsourced investigations, which I call expert-led crowdsourcing. Expert-led crowdsourcing combines professional investigators’ domain knowledge and ethos with the speed and scale of crowds — to achieve greater results than either could by themselves.

### 2.3 Approaches for Motivating Crowds to Participate

In this section, I provide an overview of different approaches to motivating crowds to participate in crowdsourced investigations, as well as how my three studies leverage a combination of these approaches.

Self-determination Theory [87] describes two types of motivations that individuals can possess: intrinsic motivations and extrinsic motivations. Intrinsic motivations act within an individual and are personally rewarding (e.g., enjoyment). Extrinsic motivations act on the
individual from outside, and any result in an external reward (e.g., monetary compensation).

Within the field of crowdsourcing, Kaufmann et al. [170] conducted a survey of the literature to categorize the intrinsic and extrinsic motives that crowd workers had for participating in crowdsourcing systems. This includes idea [200] and design competitions [52], content markets [51], Amazon Mechanical Turk [96], and other crowdsourcing domains [248]. They found two types of intrinsic motivations: enjoyment-based and community-based motivations; and three types of extrinsic motivations: immediate payoff, delayed payoff, social motivations.

Both intrinsic and extrinsic motivations can be leveraged to motivate crowds to participate in crowdsourcing initiatives, including investigations. However, different individuals are motivated by the two types to different extents. Further, intrinsic motivations are more difficult to design for than extrinsic motivations, since the latter are often external factors.

**In this work, the three settings that I explore consider a variety of these motivations, ranging from true crime fandoms (community-based) and capture-the-flag competitions (enjoyment-based), to payment on Amazon Mechanical Turk (immediate payoff) and real-world investigations (social motivations).**

### 2.3.1 Intrinsic Motivations

There are two types of intrinsic motivations that exist within crowdsourcing platforms: enjoyment-based and community-based motivations [170]. Enjoyment-based motivations include skill variety, task identity, task autonomy, direct feedback, and pastime or hobby. Community-based motivations include identifying with a particular community and social contact [169]. All but one of my studies (GroundTruth) take intrinsic motivations into consideration.

CrowdSolve and CuriOSINTy leverage enjoyment-based motivation in the form
of alternate reality games (ARGs). ARGs combine physical and digital artifacts to set up cryptic mysteries that are intended to be solved by crowds. To succeed, crowds must quickly share information and solutions, and leverage their varied expertise [124]. ARGs are also designed to be highly immersive [227], and some ARG designers even attempt to create learning environments [246].

CrowdSolve also leverages community-based motivation in the form of true crime fandom. Fandom refers to communities that are socially organized around their shared appreciation of an element or genre of popular culture. Fandoms can be powerful and transformative sites of social support [107, 153], learning [111], and creativity [114, 214], as well as sites of toxic behavior and targeted harassment [110, 165]. True crime fandom revolves around media that contain detailed, often sensationalized, fact-based accounts of criminal acts, such as murders and sexual assaults [120]. True crime, unlike many online fandoms focused on consuming and creating fiction, is unusual in that it is focused on real-life events that have social implications and pose ethical dilemmas. For instance, true crime fans have engaged in positive forms of collective action towards investigations and cases, ranging from justice reform [287] to uncovering new leads and overturning wrong verdicts [159]. However, the actions of true crime producers and consumers have also raised concerns about vigilante behavior [76].

### 2.3.2 Extrinsic Motivations

There are three types of extrinsic motivations that exist within crowdsourcing platforms: immediate pay-offs, delayed pay-offs, and social motivations [170]. Immediate pay-offs is typically monetary compensation, while delayed pay-offs includes signaling and skill advancement [96]. Social motivations include external values, external obligations and norms,
and indirect feedback [183].

GroundTruth leverages immediate pay-off, i.e., crowd workers recruited on Mechanical Turk participate in exchange for monetary compensation. In contrast, CrowdSolve and CuriOSINTy rely on social motivations, such as a desire for altruism and justice (both), and social learning (CuriOSINTy). CuriOSINTy also leverages external values and indirect feedback through a point system and judging.

2.4 Systems for Coordinating Complex Work

Researchers have created a variety of crowdsourcing systems to overcome the coordination challenges encountered when scaling up crowdsourced investigations. Traditional crowdsourcing systems rely on the requester–crowd paradigm, where a requester puts out an open call for work to be completed by crowd workers. Requesters may compensate crowd workers for their efforts, or crowd workers may be motivated by a combination of other intrinsic and extrinsic factors, as described in Section 2.3.

2.4.1 Requester-Crowd Interaction Models

Early crowdsourcing focused on a relatively simple model where a requester posted a task and returned later, after the task was completed, to review results. The introduction of the waiting room or retainer model allowed requesters to receive results in near real-time [45, 186, 187]. This fast turnaround time enabled new, richer modalities of interaction between requesters and crowds, such as clarification of ambiguous task instructions [53, 223], workflows [182], or conversations with multiple exchanges [188]. Like this last class of sys-
tems, I leverage real-time crowdsourcing (via LegionTools [135]) to quickly return crowd results to the requester, a key requirement for investigative tasks that are often time-sensitive. Most similar to my work, a subset of real-time crowdsourcing systems support what I term crowd-augmented expert work. This model assumes that (1) requesters are experts in some domain (e.g., animation, innovation, design), (2) requesters are simultaneously performing the same (or a superset) of tasks as the crowd, and (3) the expert’s own work is shaped and redirected based on the crowd results streaming in real-time. Crowdboard [31] explored whether online crowds could augment expert ideation in a hybrid physical–virtual studio. Apparition [189, 198] and SketchExpress [197] enabled crowds to prototype user interfaces and animations drawn or described by expert designers. Inspired by these prior works, I extend the crowd-augmented expert work paradigm to visual search tasks. My focus on visual search, a type of analytic task, complements prior work in this area largely focused on creative and expressive tasks, which pose distinct challenges [112].

2.4.2 Visualizations to Support Collaboration

A variety of technological solutions have been explored for improving coordination in collaborative sensemaking. One theme has been to use tools like visualizations and tabletop displays to help collaborators externalize their ideas in ways that are easily shared and aggregated with others [116, 137, 138, 139, 341], though these tools are geared towards lengthy sessions with 2–3 collaborators, not crowds and microtasks. Likewise, as crowdsourcing workflows have become more complex, new tools have helped requesters to monitor and interpret crowd work, often with dashboards and visualizations [53, 176, 182, 215, 271, 337]. Inspired by these, I also leverage visualizations to aggregate and display crowd results to the requester (expert investigator).
However, unlike these works, in which requesters typically wait extended periods of time to review work performed entirely by the crowd, I implement workflows where crowds provide feedback in real-time, while an expert investigator performs the same (and superset of) tasks. Most similar to our work, Crowdboard [31] provided two versions of an interactive workspace: “studio” for in-person experts, and “web” for online crowds; while Apparition [189] provided a “shared canvas” which showed or hid editing tools depending on roles assigned to experts and crowds. While these interfaces proved effective in their respective task domains, it is unclear how to adapt the underlying principles for analytic tasks in investigations.

Consequently, I adapt a concept from mixed-initiative systems, shared representations [149], that suggests concrete design principles for collaborative analytic work, such as data analysis and sensemaking.

### 2.4.3 Trade-off Between Predefined Tasks and Appropriability

Crowdsourcing traditionally relies upon predefined workflows [265] to decompose complex tasks into smaller microtasks that can be quickly completed by a globally distributed pool of novice workers. Completed microtasks are then combined into a final result [143, 208]. Tasks can be decomposed algorithmically [175] or manually by experts [70] or crowd workers [182], as well as a combination of two or all three [143, 176].

Algorithmic task decomposition approaches have enabled crowds to address a number of complex problems. Soylent’s Find-Fix-Verify workflow [44] enabled crowds to summarize and proofread documents. Similarly, CrowdForge’s map-reduce workflow [175] enabled crowds to collaboratively write high-quality articles and make complex decisions.

However, prior work has also found that workflows may constrain appropriations that “lie beyond a pre-specified workflow’s bounds” [71, 265]. Pre-specified workflows may thus
limit the possible ways in which requesters can intervene when instructions are incorrectly interpreted or when new types challenges emerge within a given task. Investigative tasks, such as image geolocation, often require investigators to leverage multiple approaches. For example, when one approach does not prove fruitful, an investigator may backtrack and use a new technique or tool. To this end, prior work has explored ways to build adaptive [80, 133, 208] and real-time [45, 46] workflows where requesters can provide timely feedback and dynamically modify tasks as needed. Prior work has shown that providing such guidance or shepherding can improve the quality of work [102].

Following Dix’s heuristics for software appropriation [98], I explore how to provide greater flexibility for both requesters and crowd workers through adaptable and appropriable workflows in GroundTruth and CuriOSINTy. CuriOSINTy builds in GroundTruth with investigative workflows that provide structure but do not enforce ‘hard’ constraints on the exact process or order of completion. CuriOSINTy also includes an API that enables “pluggability and configurability of tools” [98], where new tools can be integrated and configured easily. And lastly, the API enables teams to crowdsource their own work to other teams so that they may “allow the user to do more for themselves” [98].

2.4.4 Supporting Collaborative and Crowdsourced Sensemaking

Investigations have been shown to involve a complex sensemaking process [24, 33, 82, 321], where investigators (e.g., journalists, historians, etc.) must collect, analyze, and make sense of disparate sources of information to arrive at a singular theory or conclusion presented in a final report. This sensemaking process also closely maps onto the OSINT cycle [338], which involves similar steps (discovery, archival, verification, and reporting).
Prior work shows that investigators face challenges in adequately using OSINT tools and techniques. Investigators may lack time, personnel, access to vital information (e.g., reach data, follower networks, content metadata), or the programming and data science expertise to access and analyze large volumes of data [225]. Further, many OSINT tools and techniques can quickly become obsolete because of the “rapid evolution of online environments and social media platforms” [225]. These challenges, coupled with the complex and dynamic nature of misinformation investigations [297], result in many investigations being under- or unexplored.

One solution to scaling up investigators’ work practice is through collaboration. Collaboration can speed up sensemaking by dividing up discovery and verification tasks and provide diverse perspectives when creating schemas and generating reports [117]. Yet, collaboration can also result in coordination challenges. For instance, collaborators may be physically separated, have different skills and backgrounds, and have access to different data sources and tools [116, 123, 137, 167]. Prior research in collaborative sensemaking has largely focused on studying small, co-located teams (2–10 people) [171, 236, 327], or large, distributed crowds [33, 81, 89, 117, 205]. Scaling up collaborative sensemaking to involve a larger number of people can further amplify coordination challenges [307].

Most closely related to our work, Goyal et al. studied how small, co-located teams solve mysteries [138, 139], and Alharthi et al. studied how a team of search and rescue responders engaged in collaborative sensemaking during crises [24]. To better support collaborative sensemaking Alharth et al. suggest flexible rules and roles, sharing information obtained individually with others, and fostering social interaction within the group. Dailey and Stabird looked at large, distributed crowds that engaged in collective sensemaking and rumorizing during crises on social media [81], and Arif et al. studied how similar crowds corrected rumors [33]. In addition, Venkatagiri et al. studied how a large, physically co-located crowd of 250
people engaged in synchronous sensemaking to help law enforcement investigators uncover new leads on two decades-old investigations [322].

Prior work has focused on small group sensemaking for co-located teams, or large group sensemaking for geographically and temporally distributed crowds. Instead, we focus on large group sensemaking for a synchronous, virtually co-located crowd. Specifically, we study a semester-long field deployment with 40 students who conduct investigations into misinformation online in a virtually co-located classroom.

Some prior projects have focused on crowdsourcing specific sensemaking steps via microtasks, such as collecting [251], extracting [64, 321], or schematizing [70, 178, 216] data. Other projects have crowdsourced all parts of the sensemaking loop to support more complex work, such as unraveling mysteries [205, 206], or drafting articles [26, 39, 44, 143]. Like this latter group of systems, I seek to support the entire sensemaking loop, with a focus on high-stakes investigations.

CuriOSINTy differs from both types of crowdsourcing systems in three ways. First, while prior work has focused on enabling and enhancing collaborative sensemaking environments, our work begins with a largely competitive sensemaking environment in which we introduce collaborative elements. Second, unlike traditional crowdsourcing systems where the range of possible tasks is well-defined and limited, CuriOSINTy provides more agency to the crowd. Crowd workers have the freedom to choose what to collect and archive, which tools and techniques to use to conduct their verifications, and how the final report is structured. Third, traditional crowdsourcing systems are often fixed in their functionality, relying on the developers of the system to introduce new features. Expert-led crowdsourcing, though, is designed for complexity by leveraging appropriability [98, 132].

In CrowdSolve, GroundTruth, and CuriOSINTy, roles and workflows are flexible
and dynamic, and additional automated tools and crowdsourcing techniques can be easily integrated and configured.

2.4.5 Introducing Collaboration Into Competitive Environments

Competition can benefit endeavors through providing an increased sense of urgency, immersion, and motivation [252, 345]. There are several types of competitive environments, ranging from hackathons [259], capture the flag competitions [58, 73], and innovation contests [183, 308], to games [199, 237] and even self-competitions [232]. While competitions such as innovation contests and hackathons have been leveraged to address real-world problems, capture the flag competitions (CTFs) have largely been theoretical in nature.

In capture the flag competitions (CTFs), teams compete against each other by “capturing flags,” either from other competitors (attack/defense-style) or from the organizers (jeopardy-style) [58]. The team that captures the most flags or the highest cumulative value of flags — where different flags are worth different points — wins the CTF. CTFs are perhaps most well-known in the field of cybersecurity [57, 58, 226, 331], often requiring programming expertise to solve cryptographic puzzles, make database and network queries, or uncover exploits in operating systems.

Prior work has shown that CTFs are best-suited for settings focused on collecting new information or uncovering new problems [168]. Researchers have also shown that CTFs introduce a sense of urgency and strongly motivate novices to participate [58, 226], while also helping novices learn through hands-on experience [74].

However, most CTFs are designed to be theoretical in nature [168], with little emphasis on solving real-world problems. This may partly be due to the origins of CTFs in the cybersecurity field [74], where unauthorized access of real-world computer systems is illegal.
in the U.S. [104]. Still, CTFs attract tens thousands of participants yearly, engaging in millions of hours of collective effort. For example, DEF CON’s CTF attracted 3,229 teams from 2018 to 2021, each logging 276 hours of active game time [17]. Assuming three people per team, this amounts to over 2.5 million hours of effort spent on a theoretical competition that could have directly addressed a real-world problem.

Most closely related to our work, TraceLab’s Search Party CTF is a rare instance of a real-world oriented CTF, with the goal of collecting information to help law enforcement find missing persons [78]. In 2022, the Search Party CTF attracted 250 teams that collectively made nearly 4,000 submissions in four hours [6]. CuriOSINTy builds on Trace Lab’s jeopardy-style model of assigning flags of greater strategic importance more points. In Trace Lab’s model, flags are independent of each other (i.e., there are no successive challenges to be completed). Whereas in CuriOSINTy, flags build on each other — starting with a discovery flag and ending with a reporting flag. Relatedly, researchers [41] have studied OSINT organizations that exhibited elements of competition and collaboration — a term they call Social OSINT — and found two other limitations of CTFs, and competitions more generally. They are: 1) duplication of effort and 2) siloed information. These limitations are acceptable in theoretical environments, but may be less desirable in real-world investigations. TraceLab’s CTF, despite its real-world orientation, still exhibits these two limitations.

Prior work has found that introducing collaboration within a competition can help overcome these two limitations. Two types of collaboration can exist in a competition: intra-team and inter-team. While intra-team collaboration naturally benefits a single team, research has shown that inter-team collaboration can be beneficial even in competitive environments [308]. Analyzing 25 data science innovation competitions (ICs) on Kaggle, Tausczik and Wang also found that sharing code between teams improved each individual teams’ performance [308].
In a design IC, Hutter et al. found that teams who engaged in “communitition,” community-level collaboration among competing teams, made higher quality submissions and were more likely to win [160].

Different from prior CTFs, CuriOSINTy includes beneficial elements of communitition found in ICs, such as sharing work products and building on others’ solutions. CuriOSINTy also differs from most prior CTFs in that it is focused on addressing a real-world problem: combating misinformation online.
Chapter 3

CrowdSolve: Managing Tensions in an Expert-Led Crowdsourced Investigation

3.1 Motivation

Law enforcement officials have sought the help of the public in solving crimes for centuries [144, 279]. In the United States, government agents such as federal marshals and county sheriffs distributed posters and handbills of wanted criminals since at least the mid-nineteenth century [9]. As technology advanced, these “wanted” posters were augmented with photographs, published in newspapers, and eventually aired on television in shows like America’s Most Wanted, which has claimed responsibility for capturing over 1,000 fugitives from viewers’ tips. Today, law enforcement agencies use websites and social media to request the public’s help in ongoing investigations [e.g., 8, 10, 11].

While the format of these requests has evolved to new types of media over the years, the nature of the request remains relatively unchanged. That is, law enforcement officials seek pieces of information (“tips”) from the public to further an investigation. Roles are predefined, with law enforcement controlling all major elements of the investigation, from selecting
the targets to deciding which lines of inquiry should be followed up, while the public’s contribution is limited to providing raw data. Trottier [315] refers to these law enforcement-led investigations as “top-down.”

Simultaneously, with the emergence of new information and communication technologies, including online forums, ubiquitous digital photography, and powerful, free investigative tools (e.g., satellite imagery maps, reverse image search) have given rise to a new model of emergent, “bottom-up” investigations [315]. These efforts are organized online by members of the public to conduct investigations without the involvement of law enforcement professionals [e.g., 152, 243, 310]. Novices among the public seeking to become more involved in law enforcement experts’ investigations have clashed with experts’ desires for greater privacy, further complicating their interactions [184]. While some crowdsourced investigations have resulted in successes, such as locating missing persons [11], catching criminals [90], and supporting crisis response efforts [184], they are perhaps better known for high-profile failures involving vigilantism [316], or its online form, “digilantism” [243]. These include misidentifying individuals involved in the 2013 Boston Marathon bombing [243, 294], the 2017 Unite the Right rally [84], and the 2020 George Floyd protests [194], among others.

Despite this criticism, crowdsourced investigations continue to flourish and evolve [241, 343]. Therefore, it is important to understand the contexts in which they can operate ethically and successfully. In this study [322], I report on an ethnographic study of CrowdSolve,\(^1\) an expert-led crowdsourcing event that blended top-down guidance by law enforcement experts with bottom-up participation by a crowd of more than 250 amateur sleuths. Over the course of four days in October 2019, experts and the novice crowd collaborated in a co-located setting in Seattle, Washington, USA to investigate two decades-old unsolved murder cases.

Using Lee et al.’s lens of human infrastructure [196], I examine the motives, experiences,
and interactions that enabled success for all of its stakeholders, including law enforcement, the victims’ families, and attendees. I also highlight how the event’s organizers managed three key design tensions [306] within these infrastructures: experts vs. crowd, openness vs. security and privacy, and entertainment vs. reality. Finally, I discuss sociotechnical design recommendations for improving crowdsourced investigations in particular, and crowdsourcing more generally.

3.2 Methods

Initial contact. One month prior to the start of CrowdSolve, the organizers of the event contacted us based on a referral from a colleague, due to our prior experience with crowdsourced investigations. The organizers asked us to recommend tools and techniques that could they could use to leverage the 250+ attendees’ efforts in analyzing hundreds of pages of case files and generating useful leads for law enforcement. Over several phone calls and email exchanges, we provided recommendations from related work in crowdsourced sense-making. Based on the nature of the two cases, we also suggested that organizers use GIS software (e.g., Google Maps) and collaboration tools (e.g., Google Docs and Sheets, wikis). However, given limited time and concerns about digitizing case files (detailed below), the organizers implemented only one of our suggestions: giving attendees specific and actionable sensemaking tasks.

Due to the unprecedented nature of the event, we believed that studying it would provide valuable insights for CSCW and crowdsourcing research. Through our conversations with the organizers, we learned more about their plans for the event and obtained their permission to study it formally. We determined that an ethnographic study [163] would be best suited to understanding the emergent norms and interactions surrounding the event. Our
3.2. Methods

study was approved by our university’s Institutional Review Board (IRB) and took place between October 2019 and January 2020. The first two authors attended the entire four-day event. We also familiarized ourselves with true crime culture through news articles, academic papers, and podcasts associated with the event and true crime more generally.

Goals. We used Lee et al.’s lens of human infrastructure to examine CrowdSolve as an example of collaboration between law enforcement and the public. Human infrastructure is defined as “the people, organizations, networks and arrangements that constitute [a] site as a collective entity” [196]. While Lee et al. focused on the human infrastructure of a cyberinfrastructure, we found that human infrastructure is also a useful lens with which to examine contexts where digital technologies play a smaller role [e.g., 105], including CrowdSolve’s intentionally limited cyberinfrastructure. Of particular relevance to CSCW, we expected this lens would help reveal how, despite the minimal technological infrastructure, CrowdSolve enabled the crowdsourcing of sensitive information by leveraging the human infrastructure at the event. We also sought to examine the motives, experiences, and interactions of multiple stakeholders during the event — organizers, experts, victims’ families, and attendees — especially their individual and collective contributions towards the greater infrastructure of CrowdSolve.

Reflexivity. All members of our research team are interested in identifying ethical and effective ways to enable novices to support experts in domains such as law enforcement and journalism. Our prior work has focused on issues of efficacy, power, and fairness in social-technical and crowdsourcing systems.
3.2.1 Data Collection

Two members of the research team physically attended CrowdSolve from October 17–20, 2019 in Seattle, WA. We used multiple methods to collect data both in-person and online: a survey, diary entries, participant and non-participant observation, semi-structured interviews, and focus groups. We also took detailed field notes throughout.

**Terminology** We use the term *stakeholders* to collectively refer to the event organizers, law enforcement experts, victims’ families, and attendees. *Attendees* refers to individuals who we observed through participant observation. We did not interview all attendees. *Participants* refers to the subset of stakeholders who participated in our study and completed the consent form, a survey, diary entries, and/or took part in the semi-structured interviews.

**At the event** With the permission and support of the organizers, we sat next to the event’s registration desk and distributed informational packets about our study. The packets contained a flyer advertising our study, consent form, survey, and instructions for submitting diary entries. They also included our contact information, as well as a link to an identical online version which participants could share with others who did not have a physical copy. The survey asked for participants’ contact information; demographic details (age, gender, and occupation); and open-ended questions such as their motivations for participating in the event, what they would consider a positive experience, and also asked for the names of similar events that they had taken part in before, if any. In total we distributed 230 packets to all stakeholders and received 45 signed physical copies and three online copies. Four other participants provided consent verbally.

At the start of the event, when all of the attendees were seated in the main conference room, the organizers introduced the first two authors as researchers and informed them of
our study. We also took care to introduce ourselves as researchers throughout the event. We conducted both participant and non-participant observation in person, totalling approximately 60 hours. We attended sessions, participated in and observed group discussions, spoke with individual stakeholders, and facilitated informal focus group discussions during meal times. The organizers also invited us to the private Facebook group set up for the event, where we conducted non-participant observation from October 17, 2019 to January 30, 2020.

We received only three diary entries (two by phone and one by email) and interviewed two of these participants. Low participation may have been due to the time-compressed nature of the event, with attendees’ schedules filled from early morning to 9:00 pm or later most nights.

With participants’ permission, we audio-recorded all interviews except for one where the surroundings were too noisy. We also took detailed notes. All but one of the interviews were conducted in person during CrowdSolve; the other was conducted online four weeks after the event. We did not conduct any interviews on the first day to give participants (and ourselves) time to settle in. To avoid disrupting participants, we conducted interviews either before that day’s program, during the break, or at the end of the day. Based on our observations of experts’ presentations, individual work practices, and the collaborative work sessions, the first two authors periodically conferred on ways to adapt the interview guide. For instance, we observed that one team, to make better use of their time, assigned specific portions of the case files and a subset of expert prompts to each teammate to read and answer. This observation led us to ask the questions: “How did your team make sense of the case files to answer the experts’ prompts? What would you do differently in the future?” We did not directly use the case files to seed our interviews. This is because we did not have access to the case files until they were handed out each day, and because they were hundreds of pages
long, making it difficult to quickly develop specific questions.

We stopped conducting interviews when we reached theoretical saturation \[277\]. In total, we conducted 27 semi-structured interviews with 30 participants, totalling 10 hours (average = 22 min., minimum = 13 min., maximum = 60 min.).

### 3.2.2 Analysis

The first author transcribed all recorded interviews. In consultation with the rest of the research team, the first author conducted an inductive thematic analysis \[55\] of all of the collected data. As previously mentioned, we used Lee et al.’s work \[196\] as an analytic lens to highlight the human infrastructure underlying this event.

We identified four primary stakeholders: organizers, attendees, experts, and the victims’ families. Some of the initial themes for organizers were conceptualizing the event, finding suitable cases, preventing information leaks; for attendees: desire for altruism, desire for entertainment, information overload, challenges in collaboration, wanting closure; for experts: providing training, keeping attendees focused; and for the families: regaining control, emergent norms.

To refine our analysis, we conducted a member-check \[130\] by presenting our initial findings to the organizers in January, 2020. This also provided the organizers with insights on the challenges and opportunities that we found, which they could address or leverage in the second version of CrowdSolve, scheduled to take place in Chicago, IL at the end of February, 2020. We also received feedback on the final version of this paper from one of the organizers in January 2021.
3.2.3 Participants

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>M/W*</th>
<th>Age</th>
<th>Role</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>Kevin Balfe</td>
<td>M</td>
<td></td>
<td>lead organizer</td>
<td>partner at Red Seat Ventures</td>
</tr>
<tr>
<td>O2</td>
<td></td>
<td>W</td>
<td></td>
<td>organizer</td>
<td>true crime producer, journalist</td>
</tr>
<tr>
<td>O3</td>
<td>James Baysinger</td>
<td>M</td>
<td></td>
<td>organizer</td>
<td>insurance agent, true crime podcast host</td>
</tr>
<tr>
<td>OE4</td>
<td>Arthur Roderick</td>
<td>M</td>
<td>63</td>
<td>expert and organizer</td>
<td>retired US Marshal, consultant</td>
</tr>
<tr>
<td>OE5</td>
<td>Karen Smith</td>
<td>W</td>
<td></td>
<td>expert and organizer</td>
<td>retired detective, forensic consultant</td>
</tr>
<tr>
<td>E6</td>
<td>Grover M Godwin</td>
<td>M</td>
<td>61</td>
<td>expert</td>
<td>forensic investigator</td>
</tr>
<tr>
<td>E7</td>
<td>Mark McClish</td>
<td>M</td>
<td>60</td>
<td>expert</td>
<td>retired US Marshal, statement analysis expert</td>
</tr>
<tr>
<td>PD†</td>
<td>Mickey Hamilton</td>
<td>M</td>
<td></td>
<td>expert</td>
<td>TCPD lead detective for both cases</td>
</tr>
<tr>
<td>F1</td>
<td>William</td>
<td>M</td>
<td></td>
<td>Moyer’s husband</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>Sam</td>
<td>W</td>
<td></td>
<td>Moyer’s daughter</td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td></td>
<td>W</td>
<td></td>
<td>Bodine’s family</td>
<td></td>
</tr>
<tr>
<td>F4</td>
<td></td>
<td>W</td>
<td></td>
<td>Bodine’s family</td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>Tanner</td>
<td>M</td>
<td></td>
<td>Bodine’s son</td>
<td></td>
</tr>
<tr>
<td>F6</td>
<td>Karlee</td>
<td>W</td>
<td></td>
<td>Bodine’s daughter</td>
<td></td>
</tr>
<tr>
<td>F7</td>
<td>Taylor</td>
<td>W</td>
<td></td>
<td>Bodine’s daughter</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td></td>
<td>W</td>
<td>27</td>
<td>attendee</td>
<td>private investigator</td>
</tr>
<tr>
<td>P2</td>
<td></td>
<td>M</td>
<td>27</td>
<td>attendee</td>
<td>musician and teacher</td>
</tr>
<tr>
<td>P3</td>
<td></td>
<td>W</td>
<td>30</td>
<td>attendee</td>
<td>speech-language pathologist</td>
</tr>
<tr>
<td>P4</td>
<td></td>
<td>M</td>
<td>35</td>
<td>attendee</td>
<td>private investigator, para educator</td>
</tr>
<tr>
<td>P5</td>
<td></td>
<td>W</td>
<td>46</td>
<td>attendee</td>
<td>tax accountant</td>
</tr>
<tr>
<td>P6</td>
<td></td>
<td>W</td>
<td>67</td>
<td>attendee</td>
<td>retired director of special education</td>
</tr>
<tr>
<td>P7</td>
<td></td>
<td>W</td>
<td>48</td>
<td>attendee</td>
<td>mechanical engineer</td>
</tr>
<tr>
<td>P8</td>
<td></td>
<td>W</td>
<td>39</td>
<td>attendee</td>
<td>clinical social worker and consultant</td>
</tr>
<tr>
<td>P9</td>
<td></td>
<td>W</td>
<td>28</td>
<td>attendee</td>
<td>forensic mental health specialist</td>
</tr>
<tr>
<td>P10</td>
<td></td>
<td>W</td>
<td>63</td>
<td>attendee</td>
<td>clinical psychologist</td>
</tr>
<tr>
<td>P11</td>
<td></td>
<td>W</td>
<td>39</td>
<td>attendee</td>
<td></td>
</tr>
<tr>
<td>P12</td>
<td></td>
<td>W</td>
<td>47</td>
<td>attendee</td>
<td>high school English teacher</td>
</tr>
<tr>
<td>P13</td>
<td></td>
<td>W</td>
<td>50</td>
<td>attendee</td>
<td>criminal justice and social science educator</td>
</tr>
<tr>
<td>P14</td>
<td></td>
<td>M</td>
<td></td>
<td>attendee</td>
<td></td>
</tr>
<tr>
<td>P15</td>
<td></td>
<td>W</td>
<td></td>
<td>attendee</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Participant codes and demographics. Some names have not been anonymized because they have been widely mentioned on the event’s website, social media pages, and related news articles. † PD’s views are his own and do not represent the views of TCPD. * We used an open-ended question to ask participants what gender they identified as; we received two response types: “man” (M) and “woman” (W).

Organizers and experts. We interviewed five of the organizers (O1, O2, O3, OE4, OE5) who were responsible for making high-level decisions that shaped CrowdSolve. O1 is one of two partners at Red Seat Ventures (RSV), a venture and consulting business. O2 works for RSV as a true crime producer, while O3, OE4, and OE5 have acted in consulting capacities and as hosts for previous RSV events. O3 runs a popular true crime podcast called Hide
CHAPTER 3. CROWDSOLVE: MANAGING TENSIONS IN AN EXPERT-LED CROWDSOURCED INVESTIGATION

and Seek, OE4 is a retired US Marshal who has hosted several episodes of the TV show, America’s Most Wanted, and OE5 is a forensic scientist who has worked on over 20,000 cases. We also interacted with, but did not interview, two other organizers who handled logistics for the event.

Apart from OE4 and OE5, we interviewed two of the four other experts at the event (E6 and E7). E6 is a criminal profiling expert who uses multivariate analysis to determine potential suspect profiles, and E7 is a former US Marshal with expertise in detecting deception based on analyzing interview statements. While OE4 and OE5 helped to organize the event and acted as expert facilitators for parts of the event, E6 and E7 only acted as expert facilitators. OE5, E6, and E7 led one or more expert sessions. PD is a member of the Thurston County (Washington) Police Department (TCPD) and the lead detective for both the cases. He did not lead any sessions but answered attendees’ questions during other experts’ sessions.

Victims’ families and attendees We interviewed the Moyer family (F1, F2) once, and the Bodine family (F3–F7) twice. We interviewed 15 attendees (P1–P15) at least once. We interviewed several attendees twice (P1 and P2 together, P8, P10, P14). Some of the interviews included more than one attendee (P1 and P2; P8, P11, and P15).

3.3 RQ1. Organizing CrowdSolve

Using Lee et al.’s [196] lens of human infrastructure we highlight: (1) the various decisions, labors, and coordination activities required to organize the event; (2) the stakeholders’ diverse backgrounds, motives, and definitions of success; and (3) how each of them played a crucial, interrelated role in the functioning of the event.

In this section, we foreground how the organizers conceptualized the event, their goals, the
3.3. RQ1. Organizing CrowdSolve

steps they took to find suitable cases, and finally how they set up the event to meet their goals.

3.3.1 Organizers’ Goals for CrowdSolve

RSV has organized CrimeCon, an annual national true crime convention, since 2016. According to O1 (one of the two partners at RSV), many CrimeCon attendees over the years expressed that they wanted the event to delve deeper into a single case file. In response, O1’s team organized CrowdSolve: a co-located event where a crowd of novice true crime enthusiasts would learn from, and be guided by, law enforcement experts as they collectively investigate a case. RSV is a for-profit company and one of its main goals is to maximize profit. For instance, the registration fees for the event ranged from $329 to $799 per attendee, with the more expensive tiers providing additional opportunities to interact with the victims’ families and law enforcement. Correspondingly, we found that the event was marketed as providing deeper engagement for true crime enthusiasts.

However, similar to the attendees (discussed in Section 3.4.3), O1 highlighted an additional, altruistic goal for the event: “My minimum benchmark for success is that our final report to [TCPD] includes a number of new ideas, leads … that are not B.S. That they feel like are worth their action and attention.”

3.3.2 Choosing a Case to Investigate

To choose a case for CrowdSolve, the organizers first needed to develop selection criteria, then obtain law enforcement’s permission, and finally get the victims’ families to agree to participate in the event.
Developing selection criteria

O1 said that finding and selecting potential cases was “probably the most challenging” part of organizing CrowdSolve. It required aligning the organizers’ goals for the event with external factors, such as the attendees’ limited expertise, the amount of case information available, and legal barriers, among others. To help manage these considerations and determine what cases would work best, O1 decided to form an advisory committee. He drew on the relationships he built over the past four years organizing CrimeCon to populate the committee with investigators and detectives:

“They provided a solvability factor in the case itself...Is there something that you [the crowd] could potentially impact? Or is this going to be a cold case forever? Or is this a case that is just completely reliant on a piece of science, like DNA?”

The committee believed that the “solvability factor” of a case was an especially important selection criterion. OE4, who was part of the advisory committee, described how he drew on his experience consulting for America’s Most Wanted — a TV show that leveraged crowdsourced tips from the public to find fugitives — to generate a three-page list of criteria that comprised the solvability factor. This included whether there was a suspect, if the district attorney’s office was on board, and whether law enforcement was willing to open up their case files.

Finding potential cases and approaching law enforcement

After defining the criteria for cases, the organizers then needed to find potential cases. O1 said that one potential source of cases was the CrimeCon community itself. Fans, attendees, and victims’ families would often pitch cases to the organizers to be presented at CrimeCon
to the extent that their “email [in]boxes are overwhelmed with pitches.” The organizers reviewed these requests to shortlist cases for CrowdSolve. O1 said that the fans, attendees, or victims’ families would then introduce them to the local police department handling the case. However, the organizers said that they faced resistance from police departments because they lacked a track record of conducting such events. Even when the committee would get further into a conversation with law enforcement, they often encountered other challenges, such as “the family hates the police, or... solvability factor is very low” (O1).

Approaching victims’ families

At times, the sheriff’s office would introduce the organizers to the victim’s family. O1 found that “in every case, the families have been the most open to any sort of experience like this.” O3 initially pitched one of the two cases at CrowdSolve, the Nancy Moyer case (described below), to the organizers. Prior to contacting the organizers, he was investigating the case for his true crime podcast, *Hide and Seek* [37]. The case was first presented at CrimeCon 2019 by O3 and PD. When the organizers were searching for a case for CrowdSolve, they quickly realized that both TCPD and the Moyer family were interested in participating.

O2, who has worked as a true crime producer for the past 10 years, was the organizer responsible for approaching the victims and working with law enforcement to organize the case files. She would obtain the victims’ family’s contact information from the police department handling the case or would message them on social media.

### 3.3.3 Putting It All Together

We found that the organizers spent considerable time and effort in planning the event. Apart from general logistics involved in setting up an event for over 250 people, the two
cases’ unique circumstances influenced the structure of the event, including the experts that were recruited and how the case files were curated.

The two cases

Below, we briefly summarize the two cold cases selected for CrowdSolve.

**Nancy Moyer**, a 36 year-old mother of two, disappeared on March 6, 2009. Her body has not been found, and her disappearance has remained a cold case since then. In July 2019, O3 investigated the Moyer case for his podcast, and during his investigation, a suspect, Eric Roberts, allegedly confessed in a 911 call that he was responsible for Nancy’s disappearance [126]. O3 later interviewed Roberts for his podcast, during which time he recanted his confession. However, this new development turned the cold case into an active investigation, causing TCPD to restrict the information that they could share with CrowdSolve organizers and attendees.

In turn, O1 said he worried that restricting information on the Moyer case would make it less immersive for attendees, since they would have less access than originally indicated. Further, Roberts’ confession might bias the attendees who were generating leads and ideas. Therefore, the organizers decided to find a second case for the event. O2 was already working with TCPD to present the Moyer case at CrowdSolve, so she worked with PD to choose another TCPD case for the event, the Karen Bodine case. O2 reached out to the family members on Facebook.

**Karen Bodine**, a 37 year-old mother of three, was found dead on January 22, 2007. According to the CrowdSolve website, Karen had a prior arrest record related to involvement with “some unsavory people.” The lead detective, PD, believed that Karen’s murder was associated with her involvement with drugs and those people.
3.3. RQ1. Organizing CrowdSolve

As of this writing, PD is the lead detective at TCPD investigating the Nancy Moyer and Karen Bodine cases. Although several detectives prior to him had been assigned to the two cases, they have since been reassigned or moved to other police departments. Both cases remain unsolved.

Recruiting experts

Since the organizers were “trying to do something that mostly is in the hands of professionals” (O1) and the attendees had a variety of backgrounds and experiences, organizers first needed to make sure that all the attendees had a common set of skills and knowledge for conducting an investigation and approaching the case files. The organizers planned to have multiple breakout sessions throughout the event where each expert would educate the attendees on a specific aspect of the case. The organizers approached experts who had previously presented at CrimeCon or whom they knew from mutual connections. Two organizers, OE4 and OE5, acted in dual capacities as experts and organizers. O1 described how the experts were specifically chosen depending on the key elements of each case:

*Here [Karen Bodine’s case] was the autopsy, the manner of death, the statement analysis, deception analysis, the crime scene itself, and the photos...[For both] we’ve brought in two former US Marshals, we have a woman who’s done 20,000 cases in her career, Dr. Godwin, a world renowned criminologist, and Dr. Smock, the only police surgeon in the world.*

Controlling access to information

To generate relevant new leads, the organizers decided to give attendees access to the case files. However, TCPD had privacy and security concerns around information being leaked
and the potential for jury tampering. Consequently, the organizers made three interrelated decisions to control access to information at the event.

First, the organizers and TCPD required all attendees — including the first two authors at the event — to sign a non-disclosure agreement (NDA) stating that they would not share or discuss any sensitive information that they were provided at the event or present in the case files. Photography was also strongly discouraged. The organizers said that they instituted the NDA to mitigate, though not eliminate, the potential for information to be leaked and thus taint a jury pool.

Second, the organizers decided to have the event be entirely in-person, and to not use technology to share any sensitive information. The only online component was a private Facebook group to make announcements and connect attendees with each other. The organizers also did not share any sensitive information on the Facebook group.

Finally, as a result of the decision to not use any technology and to further limit the possibility of information being leaked, the organizers physically printed the case files to hand out to attendees. At the end of each day, the organizers required attendees to return the case files, and counted them to make sure none was missing.

O2 said that both cases consisted of “hundreds, if not thousands of pages of interviews, and reports, and photos,” which contrasts with the many blurry photos she had encountered when working with other police departments. However, she described the files as organizationally “just a mess” and needed to manually curate the files before handing them to attendees. She found this task difficult because she needed to first organize the files and learn the structure, and second, understand what TCPD needed help with and wanted attendees to focus on. In total, O2 said that the task took her approximately 350–400 hours to complete, occasionally working with OE5 and another organizer.
With regards to vigilante behavior, OE4 saw a possibility that some attendees would take the case into their own hands and interact with potential suspects. He went on, “I think I’ve mentioned a couple times about screwing up the prosecution. But, you know, we try to warn them as much as we can. That’s always going to be an issue.” OE5, on the other hand, was more optimistic and believed that the majority of attendees held altruistic motives with no ill intent.

### 3.3.4 The Event

The event spanned approximately $3\frac{1}{2}$ days from Thursday evening to Sunday evening. Because the organizers and TCPD were limited in how much information they could share on the active Moyer case, they decided only to spend the first $1\frac{1}{2}$ days on it. The remaining 2 days were allotted to the Bodine cold case, which could be explored in more depth, since it was not an active case. The event proceeded as follows (see Fig. 3.1 in Appendix for the complete schedule):

1. OE4 presented an overview of the case, including a timeline and description of relevant evidence found.

2. The organizers divided the 250 attendees into three large groups. For the Moyer case, the organizers divided attendees into three groups for only one breakout session at the end of the second day. For the Bodine case, the organizers divided attendees into groups much earlier — within the first three hours on Saturday.

3. The organizers gave each attendee a packet containing a portion of the curated, printed case file that varied based on the group they were assigned to.

4. The three groups rotated between different conference rooms for 45-minute expert
sessions. In each session, experts discussed specific elements of the case, such as perpetrator profiles (E6), suspect statements (OE4), and in the Bodine case, the autopsy report and a walk-through of the crime scene.

5. After the expert sessions, there was a two-hour breakout session where organizers instructed attendees in each group to form smaller teams of 8–10 people. Here, attendees worked to answer questions that the experts had identified as being potentially useful to TCPD, as well as generating other theories and ideas. There was one breakout session for the Moyer case and two for the Bodine case.

6. After the breakout sessions, all the stakeholders convened in one conference room for a plenary Q&A session.

7. On the last day, the organizers recruited a professional facilitation firm to elicit feedback from all attendees, both about the two cases and the event itself. This resulted in a Case Action Report (CAR) which the organizers shared with TCPD after the event (detailed in Section 3.5.1).

Our focus was primarily on step 5, the collaborative work sessions. These sessions were heavily scaffolded by the experts, who gave attendees specific observations to focus on or questions that they should answer in smaller teams. For example, in one session, the teams were asked to analyze a set of crime scene photos and answer a set of questions, such as, “Is there anything that may give you clues about her [the victim’s] state of mind or her plans for the weekend?” The experts rotated among the different teams, answering their questions and responding to the ideas that they generated. PD also attended some of these sessions.
### 3.4 RQ2. Other Stakeholders’ Motives and Definitions of Success

In the previous section, we presented the organizers’ motives for developing CrowdSolve, the steps taken to organize the event, and the outline of the event itself. In this section, we present the motives of the law enforcement experts, TCPD, the victims’ families, and attendees for participating in the event and their definitions of a successful outcome.

---

#### Figure 3.1: This is the schedule of events at CrowdSolve. Each column represents the schedule for each of the three large groups that the 250 attendees were divided into. Photo credits: Red Seat Ventures, LLC (used with permission).
3.4.1 Law Enforcement Experts and TCPD

We found that the law enforcement experts — OE5, E6, and E7 — expressed a desire to help the lead detective (PD), TCPD, and the victims’ families. In addition, the novelty of the setting, financial compensation, and the opportunity to showcase their expertise in an event that had amassed media attention positively influenced their decision to participate. For example, E7 noted that the event was the first of its kind and that he wanted to share his expertise with attendees.

As for TCPD, PD said that the police department was willing to participate in CrowdSolve for two reasons. First, since the two cases had not been solved for over a decade, the event might act as a catalyst for helping them make progress. For example, PD felt that there was not “anything to gain for the family, the department, or anyone else by just leaving them in a room somewhere and having a detective look over them every four or five years.” OE5 described how attendees, despite being novices, could provide a new perspective on cold cases:

*I’ve always known that the more eyes you can put on a case the more chances you have of solving it, especially when they go cold, especially when leads run out. Just because we’re experts, so to speak, in the field, and we have experience, doesn’t mean that we don’t get tunnel vision, and that doesn’t mean that we see everything.*

Second, PD believed that the experts “alone are worth the money” for the training they could provide to him and the attendees. He saw the event as an opportunity for TCPD to learn new ideas and get input from other law enforcement professionals with expertise that he did not possess.
3.4.2 Families

Both families said that they were excited to participate in CrowdSolve since it was an avenue that had not been explored before, but were willing to try since over a decade had passed and the cases had not been solved. F6 was in a state of disbelief that someone was trying to help solve her mother’s case: “I was like ‘What! Someone who’s paying attention to us?’ Like, ‘Are you kidding? Wait, is this real?’” She immediately agreed to participate because she felt that the more exposure the case got, the better. F1 also said that CrowdSolve provided his family with a chance to “put [the] case right back in the spotlight where there’s more people looking at it” and help move the cases forward. Their motives for attending the event were twofold: 1) to regain control over their loved ones’ narratives, and 2) to help solve the cases.

“Setting the record straight”

CrowdSolve provided the families with the unique opportunity to take control of theirs’ and their loved ones’ narratives, both at the event with attendees, and outside of it through the media. For instance, F2 said she often reads comments posted to social media about her mother out of curiosity, but that sometimes it upsets her and she tries “to set the record straight and reply to their comments.” However, she finds it difficult to do so because people would often ignore her, not realizing she was Nancy’s daughter.

While attendees might speculate and engage in rumor-mongering at CrowdSolve, F2 said that being at the event with a platform from which to speak allowed families a chance to clarify beliefs and facts, because they had “some control to be able to let people know, ‘Okay, here’s what we believe really happened here.’”
Families’ definitions of success

During the event, both families said that the event was already a success since it could help move the cases forward in some manner. For example, it could give TCPD access to more expertise, or generate new ideas that might lead to a prosecution. They also found the risk of failure to be minimal. If nothing came of the event, they would be “no worse off than [they] were before, so why not try?” (F1). According to F1, the “ultimate success” would be to solve the crime. However, the two families were realistic in their expectations and did not believe that the event would immediately lead to the cases being solved.

3.4.3 Attendees

Attendees had diverse backgrounds and occupations that influenced their expectations of the event and what they believed would be a successful outcome. While some attendees were from the Seattle area, most had flown in from other parts of the U.S., Canada, and Europe. OE5 and O1 pointed out that the majority of attendees identified as women. Only six of the 35 survey respondents, and five of the 15 interviewees, had attended CrimeCon or a similar event. Two of the five had attended at least three of the previous four CrimeCons. Almost all attendees had a deep interest for, and fascination with, true crime media.

Attendees expressed four motives for participating and ways that they would measure the success of the event. They wanted to (1) engage in altruism, (2) have a unique experience, (3) learn more about the role of law enforcement, and (4) find like-minded individuals. We focus on the first two here, as they came up in nearly all of our interviews.

First, almost all attendees expressed a desire to engage in altruism by helping the victims and their families “so that the family gets things that all families should get” (P6) — i.e., closure. They felt that this event provided them with a way to do so. P14 said, “When I heard about
this, it was like for the first time the public — you, me — could be an investigator. Maybe I could do some good.”

Second, many chose to attend because the victims’ families would be present, which made the event feel more personal. The attendees reported feeling more invested in the event. For example, P10 said that having the families present made it personal and real, and that it served as a constant reminder to be respectful to the victims. One attendee, P14, contrasted this experience to participating in online forums, and said that she did not see herself “getting personally invested in a random crime and feeling that same passion.”

Some attendees with relevant professional backgrounds, such as psychologists or social workers, saw true crime events like CrowdSolve as a way to advocate not just for the victims of the two cases, but for all victims more broadly. P3 explained:

"The whole true crime thing, it’s not about any one event or one moment, it’s really about this upwelling of understanding that’s coming that we all need help with this. ...I think [success is] more complex than just... “Does it solve the crime or not?”

Even though CrowdSolve brought together attendees with varied backgrounds, almost all envisioned themselves contributing towards a form of social justice, resulting in a shared definition of what a successful event meant to them.
3.5 RQ3. Interactions and Challenges Between the Stakeholders

While the organizers were responsible for envisioning, setting up, and running CrowdSolve, our findings also show that the interactions between the other stakeholders in the group were crucial for it to function properly. We focus on the most salient relationships: (1) between attendees and experts, (2) between the attendees themselves, and (3) between the attendees and the families.

3.5.1 Interactions Between Attendees and Experts

We find that the event acted as a sort of crowdfunding model, where the attendees’ registration fees directly paid for the experts to be at the event. During the event, experts taught attendees new techniques that could be used to generate useful leads for TCPD. However, the relationship between them evolved in complex and occasionally strained ways.
Experts helped attendees learn and generate leads

OE4 said that CrowdSolve bringing together experts was “very beneficial” to all stakeholders involved. Experts not only helped PD and TCPD, but also the attendees themselves. Because the experts were chosen based on important aspects of each case, O1 said that they were able to provide attendees with a baseline level of training when looking through the case files. Many attendees said that the experts’ guidance was valuable and ensured that they stayed focused on the task. P5 said that “having an expert [meant] that you can just bounce ideas off of [them], they really get it because it’s their field, and they’re bouncing ideas back.” P7, an engineer by profession, found it “very useful” when one of the experts corrected a misconception she had learned from television about strangulation by “pointing out the science of why it’s not true,” helping her to better understand the case.

Attendees also wished that they had access to additional experts beyond those at the event. For example, in Karen Bodine’s case — in which Karen and many of the suspects were reportedly drug users — two attendees wished there was more instruction about the behavioral effects of drug use and addiction. This would have helped them to “only ask the questions about her drug use that were really necessary to understand how to solve this crime” (P8).

Some experts’ presentations were distressing

Many attendees, as well as the Bodine family, found that some experts handled gory details and imagery better than others. P8 noted that the predominantly female audience (“90 to 95% women” (OE5)) may have included many survivors of violence. Some experts were more respectful and mindful of this audience, while other presentations were more insensitive and distressing. P12 described the variance in presentations:

[OE5] said ‘I’m cropping these pictures and being very respectful, the family is
here.’ And so we did see some crime scene photos, but the way that she cropped them was very sensitive…And then we went into the second one where it was about strangulation…and he had pictures that showed her private parts that didn’t need to be shown. So I could feel the tone in the room was very different. One of the family members got up and left…I thought that was really, really insensitive, and made a lot of us uncomfortable.

This example demonstrates the tension between attendees’ desire to be immersed in every detail of a case and the distress of being exposed to such macabre imagery. Further, not all of the experts may have had experience presenting to an audience of novices.

Experts were overwhelmed

Although the three experts said that interacting directly with the attendees was helpful, they were also overwhelmed by the 75 people split into 10 groups that were vying for their attention. P5 described the difficulty of getting experts’ attention:

[E]ven when the experts were talking to us, people kept coming in and interrupting us…you get one table who just talks to them for the whole time. And everyone else is like “I didn’t get a chance to ask them this or that.”

We observed that experts would walk between groups and answer multiple questions, but often about the same topic. Further, experts did not spend the same amount of time with each group, causing some groups to feel left out. Thus, experts’ time and attention were not only spread thin, but may not have been allocated optimally.
Attendees provided useful leads

Experts found the format of the event was similar to, but better than, a tip hotline for multiple reasons. First, OE4 explained that the event was focused on people analyzing evidence and generating new insights to help police. In contrast, PD said that tip hotlines rely on an eyewitness account and that often people don’t know what information law enforcement may need. Second, the attendees were not anonymous and experts interacted with them in real-time. Therefore, experts said that they were more willing to trust the attendees’ insights that were strongly grounded in the evidence. PD added that with anonymized tip hotlines, communication is one-way, and attempting to follow up on a tip can be frustrating.

According to the experts, the baseline training that attendees received allowed them to provide new perspectives and generate potentially helpful leads. PD said that some of the ideas suggested to him by the attendees and experts were useful, since they were things he had not thought of yet. This was one of the main reasons TCPD had agreed to participate in CrowdSolve in the first place (see Section 3.4.1). For example, PD planned to follow up on an attendee’s suggestion of a forensic paint database:

> Somebody pointed out that the Royal Canadian Mounted Police have a paint database, and they can do some spectrometry on one of the paint flakes that we’ve collected in this case and possibly tell me the make and model of the car that it came from.

Despite years of training, PD said that he learned valuable techniques from the experts, e.g., he pointed out how E7’s statement analysis technique would be useful in interrogating suspects.
Synthesizing attendees’ leads

Apart from the facilitation session on the last day, there was no structured way for attendees to submit their ideas to experts or the organizers during the event. During the single facilitation session, attendees rotated among groups of different sizes and shared their ideas with each other. Some of this information was written on large sheets of paper stuck to the walls, or using Post-it notes grouped into three categories: ideas, actions, and questions. The facilitators also used a mobile website to survey and quantify attendees’s thoughts and project them on a large screen as dynamic word clouds for everyone to see. After the event, the organizers asked attendees to answer questions in a survey about specific suspects and other thoughts that they may have. The organizers then synthesized notes from the facilitation session and the two surveys into a Case Action Report (CAR). The CAR was shared with TCPD after the event.

Although the facilitation session allowed attendees to discuss their thoughts and ideas in a structured manner with others, it occurred at the end of the event. Some attendees likely would not have remembered every valuable idea that they had, and multiple attendees had already left the event at that point of time.

Because we signed the same NDA as attendees, we are unable to describe the CAR in detail. At a high level, it mostly contained attendees’ thoughts about potential perpetrators and next steps for law enforcement to take, ranked based on how many attendees felt that way. The organizers said in a Facebook post after the event that the CAR was well-received by PD and was “generating quite a bit of enthusiasm among the detectives and the cold case unit.”
Attendees wanted closure

Six weeks after the event, the CAR was shared with attendees on the Facebook group, and included responses from PD to their questions. However, there was no way for attendees to follow up with the experts. Many wished that they had more information about what would happen after the event, such as whether they would be able to communicate directly with PD or the other experts about the two cases in the future. P1 proposed an online forum for this purpose:

> You know, I keep thinking what happens after this weekend...do we just kind of go back to our regular lives? ...If there’s somewhere where [PD is] like, “Okay, I’m focusing on this particular case, Karen’s case, I’m going to login to the server. Okay, people have questions. Yes, I checked the DNA for that.” And now we can continue having that dialogue in a way where he can also maintain it.

Apart from wanting to receive periodic updates, attendees wanted a way to continue working on the two cases even after the event. Indeed, on the Facebook group, one attendee said that they would create a Google News alert for the cases. On the other hand, experts were financially compensated for their participation only during the event, and may have been reluctant to participate after the event ended.

3.5.2 Interactions Among the Attendees

The attendees started to form a community, but faced challenges when it came to the collaborative work sessions. These challenges ranged from teams lacking the requisite expertise, context, or technology when reading through the case files, to collaboration challenges, such as sharing information between teams and structuring group work. Further, a few attendees
engaged in behavior bordering on vigilantism during the event.

Teams lacked organization and expertise

Within the three larger groups, attendees were asked to form small teams of about 10 people. However, we observed that without explicit guidance, they were often formed in an arbitrary manner. For example, P15 said that her team’s group discussion felt unstructured and rushed at times because her team did not choose a group leader. As a result, some teams struggled to stay focused on the task at hand. P12 explained how she refocused her team’s attention by pointing out when the topic of discussion drifted: “That was the way to silence those voices, because we’re not helping them by having that conversation, we’re helping them by answering questions that they [experts] have asked us.”

Some attendees also wished for more people with relevant expertise in criminology and psychology in their teams. They suggested offering a scholarship or special rate to encourage students in those fields to attend CrowdSolve.

Attendees experienced information overload

Attendees were enthusiastic about having access to real police case files. We observed that some took detailed notes and often discussed their findings and summaries with each other in person. Despite their enthusiasm, many attendees reported experiencing information overload when reading through over 100 pages of case files and analyzing multiple witness statements. In P9’s words: “Oh my God. It was driving me crazy. I had to start writing names down because I couldn’t keep people straight.”

Even though the case files had been organized prior to being handed out, some attendees felt that they lacked context about what they and other groups were given. For example, P14
said he felt limited without having access to the entire set of case files: “I can’t do anything if I don’t have the whole puzzle.” Although it might have been prohibitively expensive to provide each attendee with a printed copy of the entire set of case files, P14 instead wished for a shared Google Drive where attendees could read and search the case files digitally, but public access was restricted, attendees’ actions were tracked, and printing and saving local copies was disabled.

Although attendees were each given a small portion of the case files, we noticed that they struggled to synthesize and make sense of what they were given. Complicating matters further, attendees did not always know how long they would be able to keep the case files, and wished that they were informed earlier about how much time they were allotted. P1 and P5 wanted to have access to case files even prior to the event starting. P8 explained that having case files and setting expectations prior to the event would be beneficial because many of the other attendees she spoke with did not have (or make) time to read through the case files.

**NDA made attendees wary of using technology**

A few teams reported using software like Google Docs, Google Earth, and spreadsheets to work on the cases. One team used Google Docs to synthesize multiple conflicting witness statements, while another team used Google Earth to determine the distance from a potential suspect’s home and the location where the victim’s body was found. Other teams also used spreadsheet software, like Google Sheets and Excel, to create a timeline and synthesize their findings. In contrast, we observed that teams that did not use any software at all faced difficulties in collaborating effectively.

Despite these efforts, many attendees said that they were unsure about what types of soft-
ware, if any, they were allowed to use. This confusion may have arisen because the NDA prohibited attendees from retaining case files verbatim, but allowed attendees to retain information that they had synthesized from the case files. However, some attendees demonstrated awareness of, and concern about, the security of different cloud-based tools. One attendee explained how their team struggled with this issue:

*We were using Google Docs and realized maybe that’s not secure. So we were trying to figure out how to compile information properly other than just sticky notes. And the best we came up with was like an Excel spreadsheet that we could print.*

**Sharing information between teams was challenging**

Attendees wished for a way to easily share information between teams during the group sessions. Because attendees in each team did not know what information others had, what questions they were answering, and what leads they had generated, there was duplication of effort. Further, aside from individual attendees’ responses being collected during the facilitation session at the end of the event, there was no way for attendees to view and build on information that others had submitted. P8 described an ideal scenario where each team knew what other teams had generated by “*having each group write up their findings, their report, their theory, in a way where we can all share it...in a Facebook group or something else.*”

**Some attendees’ actions were concerning**

Despite OE4’s warnings, a few attendees planned or actually engaged in behavior tending towards vigilantism. Some attendees admitted visiting the place where Karen Bodine’s body
was found, and one shared photos of the site on the Facebook group. Others visited nearby libraries and geology laboratories. These activities concerned other attendees who found out about them, including P8:

*I definitely think it could promote vigilantes to take matters into their own hands.*
*I mean, even in the Facebook group, people are spending all this time going to these neighborhoods. And I think there’s some danger in going and hanging out in front of yards of people who are persons of interest, because people are going to get paranoid.*

While P8 was primarily concerned with the physical safety of attendees who visited crime scenes and potential suspects’ homes, O1 mentioned the issue of potentially tainting a jury pool and thus hurting the prosecution’s case at trial.

### 3.5.3 Interactions Between Attendees and the Families

Although the attendees and families valued each other’s participation in CrowdSolve, social norms and expectations between the attendees and the families were initially unclear and evolved as the event progressed. On the whole, both families found the event worthwhile.

**Families made attendees feel invested and provided useful information**

Nearly all attendees said that the presence of the victims’ families made them feel heavily invested in the event. They said that the families’ presence was an integral part of CrowdSolve and reminded them that the event was “*not just an intellectual exercise*” (P10). P14 explained how the families’ physical presence reminded him exactly why he was at the event: to help solve the two cases.
Reflecting on the event, many attendees were happy to have indirectly helped the two families, both by paying for the experts to be at the event and generating new leads that might enable TCPD to solve the cases. As mentioned in Section 3.4.3, providing this assistance to the families was one of attendees’ main motives for attending the event. Some attendees, like P6, even consoled the family: “I walked up to each of them separately and hugged them. That’s just [the mother in] me.”

F6, a family member, found that the attendees were very compassionate and sympathetic when speaking to her, and welcomed these expressions of support. However, other attendees expressed skepticism about how altruistic CrowdSolve could be. P12 questioned the commercial and publicity-oriented aspects of the event, and whether it was designed to actually help the families, or if it was solely a money-making enterprise. P12 saw these as competing tensions, saying, “if you really want to help, you would avoid [the] commercial piece.”

Beyond the sympathy and motivation, the families’ presence also provided attendees with unique and potentially valuable information. P14 found that the families also helped to answer questions that no one else could answer because “they know things that maybe aren’t in the reports.”

Uncertain norms around sensitive topics

All the attendees were experiencing the event for the first time, and social norms developed as CrowdSolve progressed. Attendees were initially unsure how to interact with the families. For example, P12 recalled wondering, “What kind of social dynamic do we have, both in the room and then outside?”

Between attendees and victims’ families, we found that attendees hesitated to bring up certain potentially upsetting topics in the families’ presence, such as a victim’s drug use.
Consequently, a few attendees wished that the families were not always in the same room, so that they could ask experts questions without having to worry about being respectful towards the families. The attendees’ reticence may have been well-intended but counterproductive, since it did not necessarily align with what the families wanted. For example, F1 and F2 said that they were happy to answer any questions that would move the case forward:

\[
\text{Sam [daughter] and I have told them, there’s not much we haven’t heard at this point. And so we’ve actually had to ask people “No, please be direct with us. Don’t talk around it. If you have a question, we’re here to solve something, ask us a question directly,” and then they will. And then when they don’t get this horrified response from us, then they actually feel better.}
\]

Families’ overall reflections

Despite the families’ enthusiasm to participate, we found that CrowdSolve still took an emotional toll on both families. Not only did they have to recollect possibly painful memories of Karen Bodine and Nancy Moyer, but they also heard about and viewed distressing descriptions and imagery, and interacted with many of the 250 enthusiastic attendees. The Bodine family advised that future families who might participate in similar events that it requires a great deal of mental preparation and it would be helpful to set low expectations. One family member recommended that organizers “[give as much] notice as you can give and [set] expectations for the family on what’s going to be discussed and shown.” Overall, however, the Moyer family felt that the event helped to move Nancy’s case forward, one of their main goals. One family member said: “after 10 years, things really start to feel like they’ve stagnated...but this puts it in the forefront, and it hasn’t been in the forefront for years.” The Bodine family was also moved by the attention that Karen’s case received and how supportive the attendees were.
3.6 RQ4. Managing Tensions Between Stakeholders

As reviewed in the related work, public participation in law enforcement investigations has primarily followed two models: top-down and bottom-up. The traditional top-down model limits public participation in investigations primarily to providing tips to law enforcement. The growth of information and communication technologies has enabled novice crowds to self-organize their own bottom-up investigations [184]. However, the methods and outcomes of many bottom-up crowdsourced criminal investigations have been criticized by victims’ families, law enforcement, and the media [e.g. 76, 243, 310].

The CrowdSolve model. In the previous section, we used Lee et al.’s [196] lens of human infrastructure to magnify the “social conditions and activities” that led to the emergence of an unprecedented infrastructure: CrowdSolve. We have argued that CrowdSolve represents a third model of crowdsourced investigations, expert-led crowdsourced investigations. It consisted of two dimensions: a unique collaboration environment and contributions from several distinct stakeholder groups.

First, CrowdSolve’s highly controlled environment granted participation to only those who paid registration fees and signed confidentiality agreements, and collaboration occurred in a synchronous, co-located, and largely offline manner, over a defined time period. This setting served to build and maintain trust among all stakeholder groups involved while also providing privacy and security for sensitive evidence and discussions. Such boundedness of time and space contrasts with many forms of crowd-work and true crime fandom, as we discuss below. Second, each of the four stakeholder groups meaningfully contributed to the event. The organizers chose investigations and prepared microtasks. Novice crowds with diverse backgrounds completed the microtasks, scaling up the investigation. Experts provided crucial training and guidance to the novices. Finally, involving impacted stakeholders
motivated the crowd and provided useful information.

Together, these two dimensions enabled CrowdSolve, a crowdsourced investigation of sensitive information which — in contrast to many bottom-up investigations — all of the stakeholder groups deemed a success by their own definitions. We found that PD reported receiving useful new perspectives and leads from attendees and the other experts to follow up on, without sensitive information being leaked. Additionally, our findings show that the event satisfied attendees’ motives for helping the victims’ families and law enforcement while providing them with an immersive true crime experience. Finally, the two families felt that their voices were heard and the event moved their cases forward.

**Tensions in the design space.** However, within any human infrastructure, there is the potential for friction, because various stakeholders have different motives that can lead to conflicting actions [196]. Tatar [306] argues that tensions exist at the junction between what is and what ought to be, and in the actions taken to negotiate that difference. Foregrounding such tensions in a system can be valuable because it allows for configurations to surface that may “make or break a system” [306].

Next, we discuss three interrelated tensions that the CrowdSolve model exhibited and detail aspects that worked well and those that can be improved upon in the future. The first tension is between the control of the experts organizing the event and the crowd of true crime fans participating in it. The second tension is between the conflicting goals of opening up the two cases for a crowdsourced investigation, and preserving the privacy and security of the victims and their families. The third tension is between the entertainment aspects of the event and the reality of the murders and their real-world consequences.
3.6.1 Experts vs. Novice Crowds

We contend that CrowdSolve successfully leveraged the complementary strengths of experts and novice crowds in three ways. First, the organizers invested considerable time and effort in the planning phase to make it suitable for crowdsourcing. They worked with law enforcement to find cases that would benefit from crowdsourced attention and had the support of victims’ families. They also curated and pre-processed the case files to provide the crowd with access to relevant, high-quality information. Second, the organizers secured the participation of law enforcement experts who served dual roles. They ran training sessions to teach relevant investigative skills to the attendees. They also led discussion sessions where they provided high-level guidance and leadership to keep the crowd focused on making progress. Third, the event activities were organized in ways to take advantage of the large scale of the 250–strong crowd. Sessions were parallelized, with each group led by an expert, and focused on specific topics. As a result, we found that the experts not only helped attendees learn, but also that attendees applied this knowledge to generate new and useful leads for TCPD.

**Design Recommendation 1:** Crowdsourcing can speed up and scale-up investigations. However, not all types of work can be easily decomposed into microtasks [142, 192]. For instance, rapidly developing crises may make it infeasible to design complex crowdsourcing platforms [30] or expertise may be limited but must be quickly scaled up. CrowdSolve’s expert-led crowdsourced investigation model can be applied in such situations that require flexibility and rapid scaling-up of effort. Experts can train and guide novice crowds in a synchronous and easily appropriable [132] setting to enable novices to complete more complex macrotasks without building additional technological infrastructure.

However, there were three aspects of the event that can be improved upon in future expert-led crowdsourced investigations. First, despite organizers’ efforts to design clear microtasks,
attendees reported that their team discussions were often unstructured. Second, the short session durations and large number of teams meant that teams had to vie for experts’ attention. Further, attendees from different teams often asked experts the same questions. This resulted in an inefficient use of both attendees’ and experts’ time. Third, although the experts found attendee-generated leads useful, they did not seem to note down these leads. The only formalized attempt to capture them occurred at the end of the event when a professional facilitation team was brought in.

**Design Recommendation 2:** Crowdsourced investigations might benefit from redistributing leadership [214] from the organizers and experts by assigning a team leader to each team. This team leader could help to divide work and facilitate discussion among team members, providing more frequent opportunities to reflect on and synthesize findings. They could also coordinate with other teams to decide which questions to ask the experts, using theirs and the experts’ time more effectively. To further reduce the burden on formal leaders, team members could provide shared leadership, as seen in Wikipedia [349], by informally performing some leadership behaviors, like providing feedback.

### 3.6.2 Security and Privacy vs. Openness

Echoing Lalone et al.’s [184] findings in crisis informatics, we found that the CrowdSolve organizers had to manage the tension between openness and security and privacy. On the one hand, opening up the cases and sharing as much information with attendees as possible would simultaneously maximize the chances of discovering new leads and facilitate an immersive, true crime experience. On the other hand, the organizers also had to weigh the privacy considerations of the victims and their families, as well as law enforcement’s desire to maintain a viable legal case and avoid tainted juries. These considerations affected decisions
CHAPTER 3. CrowdSolve: Managing Tensions in an Expert-Led Crowdsourced Investigation

about what information to share, as well as how to prevent the information from leaking beyond the group.

We found that CrowdSolve’s organizers employed a variety of regulatory mechanisms that allowed law enforcement to provide the crowd with unprecedented access to case files while minimizing the risk of inappropriate behavior. Lessig’s New Chicago School theory [201] provides a framework for identifying these mechanisms and analyzing their effectiveness. His theory posits that human behavior can be regulated by laws, norms, markets, and architecture.

The organizers leveraged existing laws by having attendees sign an NDA to discourage them from sharing information beyond CrowdSolve — a tactic also used by some corporations to protect sensitive data processed on internal crowd platforms [140]. Such “leaks” could hurt the case and instigate public debates over the guilt of potentially innocent suspects. As of this writing, to the best of our knowledge, no information has been leaked outside the event, thus achieving one of PD’s and the organizers’ objectives for the event.

Next, the organizers fostered social norms for appropriate behavior within the event. For instance, attendees were instructed not to engage in vigilante behavior. Many attendees reported accepting some responsibility for creating a positive impression of the event in the eyes of law enforcement and the general public as being beneficial in helping to solve criminal investigations. However, some norms were not embraced by all attendees or were not clearly set. For example, a small number of attendees visited crime scenes and some attendees were uncertain about how to act towards victims’ families. Although visiting decades-old crime scenes is not explicitly vigilante behavior, it was against the organizers’ instructions and might further embolden other attendees.

**Design Recommendation 3:** Relying on social norms may not have been entirely suc-
cessful because the event was $3\frac{1}{2}$ days long, and norms take time to be established in a community [113]. Crowdsourced investigations should strive to establish norms through recurrent events focusing on other cases and set up private online forums for attendees to continue discussion.

Finally, the organizers used markets to make the event more attractive to potential attendees than competitors, such as bottom-up investigations on Reddit or Facebook. In exchange for paying a registration fee, attendees gained unique opportunities, such as access to case files and training from law enforcement experts. The registration fee also caused those who were willing to pay to feel more invested emotionally and financially. However, those same commitments of time, travel, and money may have also excluded many potential attendees.

**Design Recommendation 4:** To broaden participation by reducing travel requirements, events such as CrowdSolve could be held simultaneously at multiple locations. Less sensitive parts of an investigation could also be conducted virtually before or after the event. For example, in our findings, we highlighted how some attendees visited local libraries and geology laboratories to conduct further research. This decoupling also paves the way for social learning via legitimate peripheral participation (LPP), as in online communities for citizen science [239], software development [344], and, most closely related to our work here, fandom [115]. LPP allows newcomers to start with smaller time and monetary commitments and gradually increase their participation. During an event, experts could assign some low-risk tasks to attendees who attend virtually, such as finding and corroborating information through zero-touch open-source research [129].

Lessig’s fourth mode of regulating behavior is through architecture. As Lessig notes, most forms of architecture, such as laws of physics and major social and cultural forces, are beyond the influence of individuals. The notable exception is computer software [202]. However, the CrowdSolve organizers prohibited attendees from using digital technologies and instead used
physical architectures such as using printed copies of case files and limiting physical access to the room. The decision to limit the use of digital technology had downstream implications for handling the case files. Despite attendees’ enthusiasm, many reported experiencing information overload, and were unable to use software tools to support their sensemaking tasks. Further, attendees could not access outside data sources to complement the case files.

**Design Recommendation 5:** We are cautious about increasing the use of technology [131], especially online tools, given the sensitive and real-world nature of these investigations. There are two major security concerns. First, as the experts expressed, there is risk of information being leaked accidentally or intentionally by the attendees. Second, as we heard from the attendees, concerns around unauthorized access by outsiders remain. Yet, the potential of software and online networks to support collaborative sensemaking is well-established in the CSCW literature. One solution, building on the assumption that attendees are co-located as in CrowdSolve, is to provide access to digital tools using an offline intranet. This approach could allow co-located teams to leverage collaborative software for sensemaking [e.g., 63, 137, 143, 205] while creating friction against accidental or intentional data leaks enabled by unrestricted internet access. When co-location is impossible, a more complex solution may be to divide case files into smaller information slices such that the crowd is given just enough global context with which to conduct their analysis and work collaboratively, while minimizing the potential for damaging information leaks. Specific slicing techniques could include directing workers to evidence documents based on time frames [72], named entities like suspects or locations [205], semi-supervised clustering techniques [143], or privacy-preserving task assignment algorithms [61].

Critics of prominent failed crowdsourced manhunts often point to the public deliberation of potential suspects as a fundamental flaw that inevitably harms innocent people [84, 243]. Large-scale public participation is seen as inseparable from unrestrained, often damaging,
speculation, dooming the entire enterprise [218]. CrowdSolve challenges this assumption. It employs an alternative model where a potent combination of regulation mechanisms, coupled with expert training and oversight, allows large-scale novice crowds to investigate difficult topics in a controlled environment. Future work can explore how CrowdSolve’s affordances for supporting crowdsourced sensemaking while protecting sensitive data can be adapted beyond law enforcement to high-stakes investigations in other domains, such as journalism, human rights, and counter-terrorism.

### 3.6.3 Entertainment vs. Reality

We found that attendees’ enthusiasm for true crime blurred the boundary between entertainment and reality. This enthusiasm led to invested participation, but it occasionally bordered on fetishization. We also found that attendees also desired closure, but were limited due to the nature of criminal investigations.

The organizers’ unconventional yet beneficial decision to have the victims’ families at the event supported attendees’ dual desires for altruism and immersion. The families’ presence helped attendees empathize, heightened the stakes of the event, and strongly motivated attendees to work hard and generate leads. We also found that attendees were able to glean additional, valuable information from the families that was not present in the case files.

**Design Recommendation 6:** Involving representatives of the people most impacted by an investigation — in this case, the victims’ families — can improve crowd behavior in at least three ways: increased motivation, greater access to information, and greater empathy. Other domains of crowd work can also leverage this approach. For example, citizen scientists analyzing data related to pollution effects on a particular community would benefit from interacting with local residents of that community, not just experts (i.e., scientists) guiding
While many attendees witnessed the families’ presence in a positive way, some sought out entertainment and objectified those who were involved in the two cases. For instance, when experts displayed images or described gruesome details in the sessions, we noticed attendees glancing at the families to gauge their reactions.

Though problematic, this behavior at CrowdSolve was less severe compared to other fandom communities where perpetrators are often obsessed over and even sexualized, such as through fan fiction [253]. While other immersive experiences, such as ARGs, thrive on extreme immersion [227], CrowdSolve benefited from the organizers and experts constantly warning attendees that their actions had real consequences.

Events like CrowdSolve may offer few opportunities for attendees to obtain closure. Many attendees reported wanting to follow up with experts to ask more questions and with TCPD to learn about the progress of the case. Like Coombs [211], we found that requests seem to arise from attendees’ competing goals. They want to see the case solved, but — like a serialized true crime podcast — they also enjoy engaging with the twists and turns of an ongoing investigation.

Continuous engagement can be challenging for two reasons. First, experts were compensated for participating in the event, and even though they might be invested in its outcome, it would be expensive to retain their services after the event. Second, the legal and procedural constraints of criminal investigations mean that law enforcement cannot share detailed updates with attendees until the case is solved or they intentionally release information through the media. After the event, the organizers did provide some high-level findings to the attendees that TCPD allowed them to divulge, but only because the attendees were still under an NDA.
3.6. RQ4. MANAGING TENSIONS BETWEEN STAKEHOLDERS

**Design Recommendation 7:** While there is no easy solution for providing closure at the end of crowdsourced investigations — especially those dealing with cold cases — organizers should ensure that there are other avenues for long-term interaction and communication with participants. For example, as mentioned above, continued engagement can happen through LPP [115] before or after the event, or by participating in events about other cases. If there is an NDA still in force, organizers can share high-level updates with attendees on a private forum, such as the informal Facebook group created by CrowdSolve attendees.

By framing volunteer crowd work as an act of fandom, leaders and requesters can come to see crowds as more than just interchangeable human processing units (HPUs) [85] or a “human API” [161]. Instead, crowd workers often seek to form communities around shared passions and goals, such as learning new information, solving problems, and, in the case of CrowdSolve, seeking justice. As in games with a purpose [328], organized efforts, like CrowdSolve, allow participants to indulge in their passions while also contributing meaningfully to society.
Chapter 4

GroundTruth: Augmenting Expert Image Geolocation through Shared Representations and Crowdsourcing

4.1 Motivation

High-stakes settings like investigative journalism and human-rights advocacy increasingly leverage photo and video evidence from social media in their investigations [146, 286]. For instance, in 2017, Europol launched the Stop Child Abuse: Trace An Object campaign [11] that relies on volunteer crowds to help identify the origin of objects in the backgrounds of imagery (photos and videos) involving child abuse. Similarly, Bellingcat, an online open-source investigative community, uses social media imagery to investigate the credibility of claims made by and about governmental or terrorist activity [152].

Uncertainty in image provenance raises questions about whether the imagery has been altered or is being reused. This uncertainty affects the credibility of the imagery and subsequently harms its efficacy as a form of evidence, whether in the court of law or of public opinion. If these images are to be used as evidence, verification is critical [146, 286].

A key step in this verification process is image geolocation, a complex sensemaking process
that involves identifying the exact location where photo or video imagery was taken [152, 180]. If an expert investigator succeeds in geolocating an image, then it can reliably be used to make claims about an event that happened at a particular place and time. Initially, an expert inspects the image for clues, such as familiar landmarks, license plates, street signs, etc., to narrow their search. When these clues are not conclusive, they use inference and experience to conduct a manual, brute-force search through large swathes of satellite imagery, looking for the location depicted in the image [152]. This task may take hours to days, and may not prove fruitful. It also does not scale easily [22, 81], meaning that successful geolocation is limited by experts’ time and attention. Computer vision attempts at automating this process [147, 326, 333] are insufficiently accurate, placing photos within within 200km of the correct location less than 30% of the time. Further, they have constraints that may not generalize well for many real-world contexts [54].

An alternative approach that has seen success is leveraging the powerful and adaptive capabilities of geographically distributed online crowds [8, 10, 11, 13, 14, 304]. However, crowds often lack expertise in knowing where to look or how to assess relevance, which can lead to false positive rates as high as 64.1% [180]. Undirected crowds can also lead to vigilantism [343] and misidentification [243]. The question then arises, can crowds effectively augment expert work practice to geolocate images?

In this study [321], I propose an approach that combines experts’ deep domain knowledge and experience with the speed and scale of crowds. To enable this approach, I extend Heer’s idea of *shared representations* between humans and intelligent agents [149], and use it to facilitate crowd-supported expert image geolocation. Heer describes shared representations as a common language through which both humans and intelligent agents can work in tandem to achieve a shared objective, balancing the complementary strengths and weaknesses of each. This approach is relevant to crowd-supported expert work practice because it augments but
does not replace experts, while still promoting efficiency and correctness, and it requires “neither perfect accuracy nor exhaustive modeling of the user’s tasks to be useful” [149].

I explore this approach through GroundTruth, a system I developed to help experts geolocate images with a crowd. GroundTruth consists of three shared representations as system components: (1) an expert-created *aerial diagram* to help share context with the crowd, focus their attention, and overcome their spatial reasoning limitations; (2) a *gridded map overlay* specified by experts that generates microtasks for crowd workers, indicating where they should search, while providing the expert an overview of crowd progress; and (3) a *heatmap* displaying expert and crowd decisions which quickly and at-scale indicates to the expert where their own time and attention is best spent.

I conducted a mixed-methods evaluation of GroundTruth involving a think-aloud protocol, log analysis, and semi-structured interviews with 11 experts working with 567 crowd workers. I find that GroundTruth effectively merges the benefits of both expertise and crowdsourcing, demonstrating the feasibility of crowd-supported expert image geolocation using shared representations. Experts worked with crowds in real-time to narrow the search area substantially, and frequently succeeded in geolocating the image. Experts were also excited by the idea of incorporating GroundTruth into their toolset since it provides features that are not currently available in other tools. Finally, I reflect on challenges and successes in designing shared representations highlighted through our evaluation.

In summary, this work makes four contributions:

1. This work makes a technical contribution by introducing *shared representations* in crowd-supported visual search, allowing visual traits and context to be easily communicated between experts and novice crowds performing a complex sensemaking task: image geolocation.
2. GroundTruth makes a system contribution as an operationalization of shared representations that enables expert investigators to geolocate images with the help of crowds. This is done using three shared representations: an aerial diagram, a gridded map overlay, and a heatmap displaying experts’ decisions and real-time crowd feedback.

3. A mixed methods evaluation with expert investigators, who drew their own aerial diagrams and worked with crowds in real time to geolocate images. Experts expressed an overall preference for GroundTruth over current expert work practice and tools. Our evaluation finds that shared representations support new collaborative work dynamics.

4. I further develop implications for enriching expert–crowd collaboration in investigative work, and applying shared representations to complex tasks beyond visual search, which are currently only the purview of experts.

4.2 RQ5. Designing GroundTruth

4.2.1 Expert Image Geolocation Workflows Today

To provide context for our system and evaluation, we first summarize expert image verification and geolocation workflows. Although there are empirical studies of verification in general [54, 284], image geolocation has seen less scholarly attention [179, 229]. Therefore, we draw on practitioner accounts [36, 146, 152, 286] to fill gaps in the limited scholarly record. We also relate these existing practices to sensemaking theory [257] and identify challenges that motivated our design rationale and system development.

Experts in many investigative fields, such as journalism, human rights, and military intelligence, perform image geolocation as a key step in the broader task of verifying photos and videos shared on social media. The goal is to identify the precise location where the image
was captured, to help support or refute claims about its provenance and meaning.

Experts perform image geolocation using a largely manual process characterized by iteratively narrowing down possibilities to find a needle in a haystack. They start with a photo that they want to geolocate, obtained through social media or by received from clients. Next, they examine the surrounding context and metadata of the image, researching the user who posted it and the claims made about it. Most social media platforms scrub metadata, including geotags, for uploaded content as a privacy measure, which is why experts focus their attention on the actual visual content of the images [286]. Experts look for road signs, business names, phone numbers, unique landscape or architectural features, or other clues that could point to certain locations or rule out others. If this step does not sufficiently narrow down the location, experts may resort to a brute-force approach of manually searching satellite imagery in candidate locations for potential matches. Experts often draw an aerial diagram of the ground-level photo of interest to ease visual comparison [152]. This “translation” process requires expertise in spatial reasoning and mental rotation which experts develop over time [229].

4.2.2 Design Rationale for Shared Representations

Heer’s [149] shared representations, adapted from Horvitz’s guidelines for mixed-initiative systems [154], enable people and agents of varying abilities to work together to achieve a common objective. They enable a user to direct tasks, performing a superset of the agents work. In parallel, the agent can support the user’s foraging tasks. This allows for the merging of the agent’s scale and speed with the user’s deep domain knowledge and skills.

Building on these ideas, we explored whether shared representations could inform the design of a system to support image geolocation in which expert work was augmented by human
4.2. RQ5. Designing GroundTruth

crowd workers rather than AI agents. To do so, we employed an iterative design process with expert investigators over the span of eighteen months, along with pilot studies. From these efforts, we developed three types of shared representations, each motivated by one of Heer’s principles [149].

1. Heer’s first principle encourages augmentations that “provide significant value, promoting efficiency, correctness, and consideration of alternate possibilities that a user might not have otherwise considered.” We adapted this idea for expert–crowd interaction as a *shared lens*. In GroundTruth, this takes the form of an expert-drawn aerial diagram. The aerial diagram is a shared lens into what the expert believes is relevant in a satellite image search, and indicates to the crowd what to look for. It bootstraps existing expert practice of aerial diagramming and overcomes the crowd’s limited spatial reasoning skills, allowing them to support experts at-scale and promote consideration of alternatives.

2. Heer’s second principle emphasizes automated suggestions that “augment, but do not replace, user interaction” and “blend into the interactive experience in a nondisruptive manner and can be directly invoked or dismissed” by the user. We adapted this idea as a *shared environment* between experts and crowds that, for GroundTruth, takes the form of a gridded map overlay. The grid structures the search area, telling the crowd where to look, and divides it into smaller cells (microtasks) for them to easily provide feedback. The grid lets experts direct investigations, and work with the crowd in a nondisruptive, dismissable manner. That is, the expert performs tasks that are the same as—and a superset of—the crowd’s, working alongside them in real time.

3. Heer’s third principle encourages augmentations that “require neither perfect accuracy nor exhaustive modeling of the user’s task to be useful.” We adapted this idea for
expert–crowd interaction as *shared analysis*. In GroundTruth, this takes the form of a heatmap. The heatmap enables shared analysis between experts and crowds to locate the ground-level photo (and aerial diagram) within cells of satellite imagery. The heatmap aggregates crowd feedback to prioritize expert attention and allows experts to easily exclude cells as they conduct their search in parallel. Although crowds prioritize some false positives, they also rule out many irrelevant cells, providing value to experts despite imperfect accuracy.

### 4.2.3 System Description and Scenario

GroundTruth consists of two different interfaces for the three shared representations (aerial diagram, grid, heatmap). The expert interface allows an expert to define and manage a geolocation task, where the expert uploads the ground-level photo and aerial diagram, and specifies the search space (for both them and crowd workers) by drawing the grid. The crowd worker interface allows crowd workers to perform geolocation microtasks specified by the expert, using the aerial diagram.

We now describe the expert and crowd worker interfaces in detail using the following fictional scenario based on a real-life event [313]. Noor is an investigative journalist who works for an online intelligence group that investigates war crimes. Late last night, the Turkish Air Force bombed the city of Aleppo, and a few seconds of footage of this bombing was released by Turkish state-run press. Located two thousand kilometers away, Noor does not have access to drone footage; and with the recentness of the event, updated satellite imagery is unavailable. To confirm the video’s authenticity, Noor must verify that these air strikes did indeed occur in Aleppo. Under a time crunch, Noor decides to use GroundTruth to help her geolocate the imagery depicted in the video.
4.2.4 Expert Interface

![Expert Interface](image)

Figure 4.1: Expert interface during George’s session. The correct location is located in the dark green cell on the bottom right. Priority (agreement) in the color legend corresponds to the number of crowd workers (out of 3) who said that the satellite imagery in the cell matched the aerial diagram. For details, please see Section 4.2.6.

**Step 0: Drawing the aerial diagram.** Upon logging in, Noor is asked to upload two images: a photo to be geolocated, and an aerial diagram that represents a bird’s-eye view of the photo, which will help both her and crowd workers easily find matching satellite imagery. She first captures a screenshot from the video, and uploads it. Then, using Adobe Photoshop, she starts to draw an aerial diagram of the photo that she wants to geolocate (e.g., the photo on the top left in Fig. 4.1). She decides to include roads, the outlines of unique buildings, and trees (which are rare in Aleppo), and then uploads it. GroundTruth uses a digital photo format, thus allowing an expert to draw the aerial diagram by hand or using any number of digital tools. Then, Noor specifies the width of the diagram she has drawn in meters or feet.

Next, she is shown a two-panel interface (Fig. 4.1), with tooltips and interface elements
that change depending on which of three steps she is currently in. The three steps are: (1) navigating to the correct location on the map, (2) drawing the search area, and (3) filtering through crowd feedback on the heatmap. On the left, placed within a tabbed view, are the ground-level photo and aerial diagram, along with rotate, pan, and zoom controls. On the right is a Google Maps map, with controls to zoom in/out, and to toggle the satellite/map view.

Step 1: Narrowing the search area. Noor needs to verify that the air strikes took place in Aleppo. She has narrowed down the location to an area of several square kilometers, in the northwest corner of the city, based on news reports and contextual clues associated with the ground-level photo. Noor thus navigates to this location using the embedded search bar on the map.

Step 2: Drawing the search area. Next, Noor delineates the search area that she thinks contains the photo by clicking on the Draw Search Area button, and then clicking and dragging the mouse cursor to draw opposing corners of a rectangular gridded map overlay. In practice, the area can be as large, or as small as one would like, based on constraints such as time, cost, the number of crowd workers available, etc. Next, the system automatically divides the search area into a grid of equally sized regions. Each region consists of a $4 \times 4$ grid of 16 cells, each of which are the same width as the aerial diagram provided by Noor, to allow for 1:1 comparison. Three crowd workers are asked to compare each cell to the aerial diagram provided by Noor.

Step 3: Analyzing crowd feedback. Now, Noor can begin to search through the map with the grid overlaid on top of her target search area. In parallel, she recruits people on Amazon Mechanical Turk (MTurk) to help. She takes their feedback into account as it
begins to stream in and is displayed as a heatmap. On the left panel, there is a color legend that indicates the number of crowd workers that said the cell matched the aerial diagram provided. The colors correspond to the number of Yes/Maybe judgements (between 0 to 3).

Next to this is text that alternates between Pan Mode and Inspection Mode. Pan Mode is the default mode where Noor can pan across and zoom into or out of the map. In Pan Mode, she can give up on the search using the Give Up button, or hide crowd feedback with the Hide Color button. If she continues to zoom in, until only one cell takes up 80% of the map, she enters Inspection Mode. Here, two new buttons are added: clicking the Exclude Cell button excludes a cell from the search, and clicking the Found It button indicates that she has found the cell that matches her image. Even when she hides crowd feedback, her decisions are still visible, in line with Heer’s second principle.

The crowd, noticing that Noor’s diagram contains several trees, quickly eliminate cells of satellite imagery that do not contain them. When they encounter cells that contain trees, they attempt to carefully match it to the buildings that she has represented in the aerial diagram. Within minutes, the crowd has narrowed down the search area by 50%, highlighting areas that she should pay attention to and prioritize. She inspects three cells, and finds that one of them matches the aerial diagram. She clicks on the Found It button to end her search, having successfully geolocated and verified an airstrike that took place in Aleppo, Syria.

### 4.2.5 Crowd Worker Interface

The crowd worker interface consists of three columns. The left one shows only the aerial diagram drawn by Noor, and is randomly rotated to avoid imposing an orientation. The crowd is given only the aerial diagram because crowd performance is better, in comparison to being shown only the ground level photo or both [180]. A crowd worker can rotate, zoom,
Figure 4.2: Crowd worker interface. Note that this is a reproduction of what a real crowd worker would have seen during George’s session, and includes the same aerial diagram that he created.

and reset the image to its original position using the buttons underneath the image. On the right is a small mini-map of the crowd worker’s search region consisting of 16 smaller, equally-sized cells (4×4 grid). In the middle, there is a single cell, shown in a Google Maps satellite view, and workers are prevented from panning outside this cell. The worker can click on the green **Yes/Maybe** button if the satellite cell potentially matches the aerial diagram, or a red **No** button if it does not. Their judgment is reflected on the mini-map where the corresponding cell is either colored green or red, respectively. Once sure of their choice, they can click on the **Next** button. Then, the satellite view advances to the next cell following a Creeping Line search pattern used by search and rescue professionals [340]. Once the crowd worker provides feedback for all 16 cells, they can submit the task to MTurk. GroundTruth also allows experts to recruit crowd workers from social media or other sources. In either case, a single URL is provided and the system automatically directs them to a region to perform the geolocation microtask.

The crowd worker interface for GroundTruth (Fig. 4.2) is based on our prior work [180], where we designed a technique and an interface to facilitate crowd-supported image geolocation and
there was no expert interface present. The crowd worker interface here differs from our prior work \cite{180} in several ways. First, to streamline visual comparison, we increased the size of the diagram to match the size of the satellite image cell. Because of this layout change, we moved the mini-map to the top right. We also added buttons to allow the user to zoom in and out of the diagram and reset its position. All other parts of the interface are visually similar.

### 4.2.6 Implementation Details

We built GroundTruth using the Django web framework, a PostgreSQL database, and the Google Maps API for displaying satellite imagery, drawing the grid, and displaying the heatmap of expert and crowd decisions. The gridded map region is drawn using Google Maps’ Shapes API, by clicking and dragging the mouse cursor across the map to delineate a rectangular bounding box representing the area under investigation. Next, GroundTruth divides the bounding box into equally sized \textit{regions}, composed of equally sized $4 \times 4$ \textit{cells} whose height and width are defined by the expert when they specify the cell/diagram width. However, if it cannot be divided equally, it is resized so that it can be. Each region and corresponding cells are stored in the database. Due to the map projection used by Google Maps, cells do not appear square unless they are near the equator. As workers load the crowd worker interface, they are redirected to a specific region, i.e., a subset of the grid. Their feedback for each cell is stored individually in the database as a \textit{judgement}. On the expert interface, the system queries the database to see if three judgements have been made for a cell (by three workers). If so, the number of \textbf{Yes/\textit{Maybe}} judgements is calculated based on the number of workers that said the cell matched the diagram by clicking the \textbf{Yes/\textit{Maybe}} button. Each cell is colored based on how many workers clicked \textbf{Yes/\textit{Maybe}}, using the \textit{one-yes}
rule, described in Section 4.3.5. A cell is only colored once three crowd workers have provided feedback for it. The colors are red, orange, light green, and dark green that correspond to no (zero of three crowd workers said yes), low (one of three), medium (two of three), and high priority (all three), respectively. While experts were shown the color legend that used the word “agreement,” for clarity, we henceforth refer to it as crowd worker “priority.”

4.3 Methods

Heer [149] states that the evaluation of systems that use shared representations must consider both “task time and quality measures, as well as more qualitative concerns regarding participants’ perceived autonomy and creativity of action.” Along similar lines, we seek to form a holistic understanding of how GroundTruth supports experts in performing image geolocation with crowds, and posed the following research questions:

- **RQ6 Performance:** How well did experts and crowds perform image geolocation together using GroundTruth?

- **RQ7 Experience:** What were experts’ experiences of using GroundTruth to work with crowds in geolocating an image?

We conducted an exploratory evaluation in which geolocation experts worked with crowds from MTurk to geolocate images. Our mixed-methods approach included a think-aloud protocol and analysis of system log data, followed by in-depth, semi-structured interviews. Below, we describe the setup for our study, expert recruitment, the study procedure, and qualitative and quantitative data collection and analysis methods.
4.3. METHODS

4.3.1 Image Dataset Selection

To provide a controlled “playground” for experts with different skill sets and who are used to geolocating images in different contexts, we selected a diverse set of images to geolocate, in terms of possible challenges and strategies. We randomly chose coordinates—using guidelines provided by Mehta et al. [229]—that matched each of 24 triplets (6 biomes × 2 population density types × 2 language types), and excluded those that were of prominent locations or had identifying location markers. We then obtained their respective ground-level photos using Google Street View in GeoGuessr to remove road names. The resultant set of 24 images were displayed randomly in an image gallery that experts could choose from.

Now, image geolocation can involve manually searching an area as small as a town, to as large as a country [229]. Thus, due to time constraints with these experts and to ensure that they each had a similarly sized, manageable search area, we narrowed the search area to one that was 240 to 300 times larger than the area that was depicted in the ground-level photo. This varied based on the diagram width that the expert selected. The narrowed search area was in the form of a green rectangle, that encompassed the image location.

4.3.2 Expert Recruitment

We recruited 11 participants (experts) with expertise in image verification and geolocation, defined as performing verification/geolocation on a weekly basis for at least one year. To reach the widest audience possible in this relatively small population, we recruited participants using purposive sampling [238]. This involved advertising our study on Twitter with a link to a screening survey. We also reached out to experts that had previously participated in our design process, as well as others that we found online via email. We paid experts $75 for taking part in a 90-minute study, which is commensurate with their specialized skills.
After 11 experts, we reached theoretical saturation and concluded recruitment.

After recruitment, we asked experts to fill in an online consent form and pre-survey, and scheduled a date for the study. Experts self-reported belonging to a wide range of domains, including journalism, law enforcement, human rights/war zone investigation, and financial intelligence. Two experts identified as female and nine as male; and ranged in age from 18 to 45, with a median age of 32. Nine participants identified as White/Caucasian, one as Asian, and one as Asian American. We include seven participants’ full names and affiliations, while four participants expressed wanting to remain anonymous. We believe these experts are accurately able to assess the pros and cons of being identified because of their professions: as journalists who are often in the public eye, and as intelligence analysts who are aware of various security and privacy related issues. McGregor et al. [228] also found that journalists were able to accurately assess similar security and privacy issues.

4.3.3 Crowd Worker Recruitment

We used LegionTools [135] to facilitate real-time, expert–crowd interaction. The retainer mode allowed us to recruit unique crowd workers for each session from Amazon Mechanical Turk (MTurk), with no qualifications except for being USA-based. This was to ensure that a large a pool of crowd workers would be available to work in a short period of time. Prior to accepting the task, crowd workers were asked to provide consent. Once recruited, they finished a tutorial and were then pooled up at a waiting page. This pooling was necessary to facilitate a real-time experience for experts. Based on pilot studies, it takes 10–15 minutes to pool 30 crowd workers.

We paid crowd workers $7.50/hr for the time they were active in our task, i.e., completing a short 2-minute tutorial and working on a 10-minute task. While they waited for the expert
4.3. Methods

to specify the search area, we paid crowd workers $2.00/hr (for up to 15 minutes). While waiting, we informed crowd workers that they could complete other tasks, and would receive a browser pop-up notification when our task was ready. The total pay per task was at most $2.00.

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Age</th>
<th>Country</th>
<th>Profession</th>
<th>Organization</th>
<th>Yrs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lorenzo Romani</td>
<td>Male</td>
<td>30–35</td>
<td>Italy</td>
<td>OSINT Analyst</td>
<td>Anon. Intelligence Firm</td>
<td>4</td>
</tr>
<tr>
<td>Dakota Flourney</td>
<td>Female</td>
<td>26–30</td>
<td>USA</td>
<td>Journalist</td>
<td>Storyful</td>
<td>3</td>
</tr>
<tr>
<td>John*</td>
<td>Male</td>
<td>36–40</td>
<td>Europe</td>
<td>Senior Analyst</td>
<td>Anon. Military</td>
<td>15+</td>
</tr>
<tr>
<td>Jack*</td>
<td>Male</td>
<td>30–35</td>
<td>Canada</td>
<td>Investigator</td>
<td>Anon. Intelligence Org.</td>
<td>1</td>
</tr>
<tr>
<td>Aqwam Hanifan</td>
<td>Male</td>
<td>26–30</td>
<td>Indonesia</td>
<td>Reporter</td>
<td>Tirto.id</td>
<td>3</td>
</tr>
<tr>
<td>Alec*</td>
<td>Male</td>
<td>18–25</td>
<td>USA</td>
<td>Consultant</td>
<td>Anon. Intelligence Firm</td>
<td>1</td>
</tr>
<tr>
<td>Benjamin Decker</td>
<td>Male</td>
<td>30–35</td>
<td>USA</td>
<td>Researcher</td>
<td>Harvard University</td>
<td>5</td>
</tr>
<tr>
<td>George*</td>
<td>Male</td>
<td>41–45</td>
<td>USA</td>
<td>Investigator</td>
<td>Anon. Police Dept.</td>
<td>16</td>
</tr>
<tr>
<td>Neil Seward</td>
<td>Male</td>
<td>18–25</td>
<td>Canada</td>
<td>Manager</td>
<td>Anon. Bank</td>
<td>9</td>
</tr>
<tr>
<td>P. Kim Bui</td>
<td>Female</td>
<td>36–40</td>
<td>USA</td>
<td>Director</td>
<td>Anon. Newspaper</td>
<td>15</td>
</tr>
<tr>
<td>James O’Brien†</td>
<td>Male</td>
<td>46–50</td>
<td>USA</td>
<td>IT Consultant</td>
<td>Self-Employed</td>
<td>18</td>
</tr>
</tbody>
</table>

Figure 4.3: Expert participants and demographics. * = Anonymized, † = James_FP session.

4.3.4 Procedure

Before each session, experts filled in a consent form and pre-survey asking about demographic information (Table 4.3) and work practice. Then, we presented each expert with a walkthrough of the system, asked them to perform an image geolocation task, and conducted a semi-structured interview on their experience. All eleven sessions were run over video chat software (Zoom).

We first demonstrated how GroundTruth worked with a walkthrough, and provided a high-level overview of the system’s underpinnings. After clarifying any questions, we gave each expert a link to the aforementioned image gallery. Here, we instructed them to choose one image to geolocate, that was similar to what they would typically encounter in their work.
After each session, we removed the image that they picked from the gallery, so that no two experts would geolocate the same image. This was to elicit diversity in terms of techniques and challenges.

Next, we asked them to rate how difficult the image would be to geolocate using a seven-point Likert scale (Table 4.4). Then, we provided a narrowed location (240 to 300 times larger than the area depicted in the photo) to reduce the search area to a manageable size, and due to time constraints.

We used a think-aloud protocol [109] for the rest of the session, where we asked experts to externalize their thoughts. To familiarize experts with the protocol, we asked them to practice using the protocol on an unrelated task (using a search engine to find a desktop wallpaper).

We then asked experts to spend at most 10 minutes drawing the aerial diagram that the crowd would use. They went through the four-step process on the expert interface (as in Section 4.2.4), using the tooltips provided. Once they had specified the search area, the system divided it into smaller cells (based on the width chosen), and were given to the crowd to provide feedback.

In parallel, crowd workers were hired using LegionTools, who completed a tutorial and were sent to a waiting page. We hired crowd workers when the expert began to draw the aerial diagram. Once the expert had specified the search area, we routed crowd workers from the waiting page to the task page to work on the task.

As the heatmap populated crowd feedback, experts inspected each cell of satellite imagery, and provided their judgments on whether or not it contained the correct location. We gave each expert a time limit of 30 minutes, at the end of which we asked them to pick a cell that they believed contained the image, if they hadn’t already. They were informed that the
narrowed search area they were provided contained the correct location. We also provided one expert (James) with an area that did not contain the correct location. We will refer to James and his session as James_FP.

After they picked a cell that they thought contained the correct location, we conducted in-depth, semi-structured interviews with each expert, asking them about their overall experience using GroundTruth, their use and perceptions of the heatmap, how using the system compares to current practice, how they see it fitting into their work, among other questions.

### 4.3.5 Data Collection and Analysis

The audio and video of each session was recorded using Zoom, and fully transcribed. Sessions ranged from 68 to 112 minutes (median = 80). The first author of this work conducted each session and took detailed notes [290] both during and after each session on how the expert utilized GroundTruth. The notes recorded every time the expert commented on, or reacted to crowd feedback; what cells they looked at; any issues they faced with the user interface; and other points of interest. Notes were subsequently incorporated into each transcript.

Next, we conducted a deductive thematic analysis [55] of the transcripts, based on themes relevant to our research questions and three system components. These themes describe experts’ behavior while using GroundTruth and their reflective experiences afterwards (RQ2): drawing the aerial diagram, using the grid and the heatmap with crowd feedback, among others.

To analyze expert performance (RQ1), we determined how far the location that the expert selected was from the location in the ground-level photo, in terms of distance and number of cells. We also calculated how long experts took to do so, from the moment they specified the search area. We also asked experts to rate how difficult the image would be to geolocate.
To analyze each crowd’s performance (RQ1), we determined how long they took to provide on the cell that contained the ground-level photo, and on the entire expert-specified search area. We also calculated what percentage of the search area they were able to rule out while retaining the correct cell, and what percentage of cells they indicated as low, medium, and high priority. Here, the correct cell is the cell that contains the location in the ground-level photo. To analyze their feedback, we relied on the one-yes rule proposed by Kohler et al. [180] for calculating aggregated crowd worker performance in image geolocation tasks. The one-yes rule states that if at least one crowd worker said that a cell and the aerial diagram match, then that particular cell would be classified as a possible match. Conversely, a cell will turn red if all three workers said that there is no match to the aerial diagram. This helps ensure that a needle-in-a-haystack problem like geolocation—where there is only one correct answer—will not be overly aggressive in ruling out possible cells. We made this conservative decision because incorrectly discarding the correct location is worse than keeping an incorrect one. A long tail distribution among the three priorities (with fewer high priority cells) is indicative of better performance since the crowd is able to better direct an expert’s attention.

### 4.3.6 Limitations

We are unable to draw conclusions across participants because we recruited experts from different fields that involve image geolocation, and no two experts geolocated the same image. However, this allowed us to highlight the diversity of challenges faced and strategies employed by experts in geolocating images with crowds.

Further, geolocation involves considering context, searching social media for corroborating sources, and inspecting visual clues, only resorting to brute-force satellite image search when earlier steps are insufficient. In our study, images contained no context within them, which
is not typical of how experts encounter images in their work. However, this allowed us to simulate the brute-force step of image geolocation, when other approaches have been exhausted.

Finally, due to experts’ time constraints, we limited each session to about 90 minutes and experts did not go through the full image geolocation process, focusing only on the brute-force step. However, this created a time-compressed situation where experts might normally seek others’ help. Experts’ time constraints, coupled with the typical duration of an image geolocation task (hours to days [146, 152]) also meant that it was infeasible to obtain a baseline. Future work should also study expert performance without crowd support, to obtain baseline performance metrics.

4.4 RQ6. How Well Did Experts and Novice Crowds Perform?

In this section, we summarize how experts rated the images in terms of difficulty to geolocate, experts’ and crowds’ overall performance, an analysis of how expert-drawn aerial diagrams affected crowd performance, and how crowd feedback affected expert performance. The correct cell is defined as the cell that contains the location depicted in the ground-level photo. Table 4.4 includes detailed information about each expert’s session.

4.4.1 Image Difficulty

Experts evaluated image difficulty on a seven-point Likert scale ranging from extremely easy (-3) to extremely difficult (+3), with a median rating of moderately difficult (+2), and only two as easy (<0).
Figure 4.4: Expert and crowd session details and performance. * = Image difficulty was assessed using a seven-point Likert Scale ranging from extremely easy (-3) to extremely difficult (+3). † = James_FP session.

Experts said that the images we presented to them in the image gallery did not have any clues within the images that would have helped them find it easily, and that it would have been difficult to geolocate without the initial clue that we gave the experts (typically, they would rely on clues associated with or present within the image). Once we provided experts with a hint that narrowed the search location to an area that was 240–300 times the width they estimated for the ground-level photo, some experts thought the image was easier to find than they initially thought, while others thought it was harder to find.

### 4.4.2 Overall Performance

Overall, experts ranged from identifying the single correct cell to selecting a cell no more than seven away. While six experts identified the correct cell, the other four were an average
of three cells away. Note that the cell widths chosen by experts here varied from 100 to 500 meters (the system allows for cell widths between 100 to 1000m.). They took between 2 and 29 minutes (mean = 13.7), with the session capped at 30 minutes.

Crowd workers took between 8m 20s and 22m 10s to provide feedback on the entire search area (mean = 13m 5s). They narrowed the search area by 27.3 to 66.5% (mean = 43%) for an area that ranged from 144 to 288 cells (1.44 to 36km$^2$, mean = 6.81km$^2$). In 7 out of 10 cases, crowd workers were able to identify the correct cell in a search area of 144 to 288 cells, taking between 5m 16s to 22m 10s (mean = 11m 37s). Crowd workers ruled out the correct cell in three instances, excluding the James_FP session where the search area intentionally did not contain the correct location.

For the three sessions where the crowd did not identify the correct cell, two experts (Aqwam, Ben) pinpointed the correct location within 10 minutes. On the other hand, there were three cases (Lorenzo, John, Alec) where experts were, on average, 350m away from the correct location. Here, the crowd marked the correct location in green within 11 minutes. In other words, there was only one session (out of ten)—Kim’s—where both the crowd and expert did not pinpoint the correct location.

### 4.4.3 Effect of Aerial Diagram on Crowd Performance

The characteristics of the aerial diagram, such as how easy it was for crowds to interpret, whether it incorporated real or imagined features, and how common those features were in the satellite imagery, affected how well the crowd performed.

In sessions where experts depicted unique architectural and structural features in their aerial diagrams, and/or where the search area consisted of satellite imagery that was easy to rule out, the crowd performed well. For example, in Lorenzo’s session, the crowd was able to
reduce the search area by 66.5% because the diagram contained roads, while large parts of the search area consisted of shrubland without roads. In addition, 73.4% of the remaining search area was marked as low priority (yellow, one crowd worker said the cell matched the diagram), 20.3% as medium (light green, two), and 6.3% as high (dark green, three). In Dakota’s session, although the crowd eliminated 40.8% of the search area from the search, we still observe a long tail in terms of priority: 2.7% of cells were marked as high priority, one of which was the correct cell.

When the expert’s aerial diagram was unclear, or when multiple cells of satellite imagery appeared to match features depicted in the aerial diagram, crowd performance was lower than average. For example, in Aqwam’s session, the crowd reduced 27% of the search area (the average was 43%), perhaps because his diagram depicted features that were not actually in the ground-level photo or the satellite imagery. In this case, no crowd worker thought the correct cell matched the aerial diagram. In George’s session, the crowd marked the correct cell as high priority. However, the crowd eliminated just 30% of the search area. This may be because in George’s aerial diagram he annotated two rectangles as “high-rise buildings” in the center of Dubai, a dense urban area with nearly 200 skyscrapers. Similarly, in John’s session, crowd workers had marked the correct cell as high priority. However, there were more high-priority cells than average (21% vs. 12%) because the building he depicted in his aerial diagram had a structural feature that was common in the dense urban area he was searching through.

4.4.4 Effect of Heatmap on Expert Performance

Experts used and interpreted the crowd feedback displayed as a heatmap in a variety of ways, which affected their performance. Four experts (Lorenzo, Aqwam, Alec, George) found the
correct location using crowd feedback. Three others (John, Alec, Kim) chose cells that looked visually similar to the correct location but were not. Here, the crowd also indicated that they looked similar (medium/high priority). Three (Jack, Ben, Dakota) found the correct location before the crowd could finish giving feedback for that cell. Finally, James_FP is in a separate category.

Lorenzo, Aqwam, Alec, and George found the correct location by inspecting areas that the crowd had marked as high priority. John also inspected the correct cell that the crowd had marked as high priority, and thought it was a possible match but then excluded it from the search. He said this was because some features in the satellite imagery did not match due to the angle from which it was taken. Towards the end of the time limit given, John picked a location that was 700m (seven cells) from the correct one; crowd workers had marked the cell he chose as high priority.

Alec and Kim identified locations that looked similar to the correct one, but were not, and the two crowds had marked them as medium priority. In both sessions, Alec and Kim directed their search based on crowd feedback, by first inspecting high and medium priority cells. The location Alec picked was within 200m (two cells) of the correct one. Alec said that the satellite imagery was not detailed enough to be 100% certain of his decision. Kim managed to narrow the location to within 400m (four cells) of the correct one. She initially disagreed with the crowd’s medium priority level, but then changed her mind. Although Kim took 25 minutes to search through 7.68km², she expressed desire for another hour to verify her choice and continue searching. In Alec’s session the crowd marked the correct cell as medium priority, while in Kim’s session, they ruled it out.

On the other hand, Jack, Ben, and Dakota knew exactly what features to look for and were able to quickly find the correct location within 0m, before crowd feedback came in for the correct cell. They took two, three, and six minutes, respectively. Both Jack and Dakota
relied on crowd feedback initially rule out locations, which let them quickly find the correct location. In Dakota’s session, a large portion of the search area consisted of forests, while the ground-level photo was that of a cylindrical structure, which is possibly why she was able to find it quickly. Jack also knew exactly what features to look for (a bridge-like structure along a canal), and in only a portion of the search area contained canals, making it easier to filter through. Similarly, Ben said that there were many distinctive features in the ground-level photo that made it easy for him to geolocate.

Finally James_FP thought that he had found the correct location after six minutes of searching, even though it was not contained within the search area that he drew. James_FP mentioned that without recent, high-quality satellite imagery, it would be difficult to directly confirm whether or not it was the correct location.

4.5 RQ7. What Were Experts’ Experiences of Using GroundTruth?

Overall, experts said that they were excited by the potential for GroundTruth to scale up expertise in fields such as journalism, human rights advocacy, and criminal investigations. They mentioned specific benefits such as training new investigators, and enabling quicker and better support from crowds, and that it may even save lives.
4.5. RQ7. What Were Experts’ Experiences of Using GroundTruth?

4.5.1 Aerial Diagram

Determining the Width

Experts varied in their ability to estimate the width of the ground-level photo. Some, like Ben and Lorenzo, were able to quickly and accurately make a reasonable guess. For example, Ben’s strategy involved extrapolating from smaller objects of known size: “so if it’s one lane, [it’s] essentially one car, which means somewhere between four and seven feet across approximately.” Other experts, like Kim, found the task more challenging. Kim said, “the width is hard for me. I’m not a good judge of distance whereas many other people are. I’m pretty terrible at knowing what 300 feet is.”

In designing GroundTruth, we included the width estimation step for the practical purpose of leveraging expert knowledge to determine an appropriate cell width for crowd workers. However, some experts found that the step also supported their own spatial reasoning by prompting them to reflect on the relative proportions of key objects. For example, Dakota said, “that’s very helpful...[to] estimate the proportions of the area. You know, in reference, I didn’t know it was next to a river and that helped me get a sense of how big of an area I need to be looking in.”

Drawing the Diagram

While drawing the diagram, experts highlighted salient features in the photo that they thought would be visible on satellite imagery. Some experts chose to label their diagrams, while others only drew the outlines of buildings and roads. Many noted that drawing aerial diagrams is a skill that is developed over time. For example, Kim said, “I know to do buildings and big roads...but they don’t always think about trees, driveways, the color of the...
roof, the smaller details. I always tell people not to look in the foreground but to look in the background of images, and that’s a bit of a training process.” In her diagram, Kim provided crowd workers key features to look for, such as “white shutters, roof tiles, and a large, green lamp.”

Eight out of eleven experts elected to draw the diagram by hand and upload a digital photo, while three used digital drawing tools: Microsoft PowerPoint (George), Microsoft Paint (Lorenzo), and Adobe Photoshop (Aqwam). Experts who expressed the preference for drawing it by hand explained that it was quicker, as well as easier to represent certain features and make annotations.

Reflecting on their experiences using the diagram, and viewing crowd feedback on the heatmap, all experts felt that they would have drawn the diagram differently and taken more time to do so. Since the GroundTruth diagram is dual-purpose in nature, experts grappled with drawing the diagram just for themselves versus also drawing for crowd workers. Experts said that determining what features to include within their diagrams, and what to exclude, was crucial. George explained:

\[
\text{I probably would have taken about an hour to do it versus 10 minutes because I think crap in equals crap out...If you don’t have a quality diagram that really conveys what you’re seeing, you’re going to get those false positives...and I take full ownership for that just based on the poor accuracy of my diagram.}
\]

Kim and Ben wanted to communicate more detail to the crowd on what to look for in the form of a list or detailed notes. Kim said, “I would have said, ‘I don’t think this is in an extremely rural area, or in the middle of a city. It seems like it’s in the outskirts of a city, keep an eye out for X, Y, Z.’”

John and Dakota both highlighted their “extremely limited artistic ability” which was suffi-
ccient for their own purposes but, they worried, not for crowds. John proposed that getting feedback from crowd workers on his diagram would help him improve. Alternatively, experts expressed a desire to have a built-in tool to draw the diagram with pre-specified symbols to reduce uncertainty for crowds and speed up the process. For example, Alec said, “It’s a good process, but to make it more streamlined you might want have a key of labels you can put on it...Maybe set symbols for certain things, because I wasn’t sure what to put for a tree.” Other possibilities suggested by experts included outsourcing this task to a graphic design expert or even another set of crowd workers.

Using the Diagram

Many experts made references to the diagrams they had drawn. John, Dakota, Alec, and Neil looked at the diagram on the interface intermittently to compare it to feedback that crowd workers had provided, whereas Ben, Aqwam, and James_FP frequently switched between using the diagram and the ground-level photo. Kim preferred to use the physical diagram she had drawn because “I can rotate it [on the interface]. But this is literally what I’m doing on paper...I’m a super tactile person.” Lorenzo said that he uses aerial diagrams in his work practice, but did not use the digital diagram during his session, relying solely on the ground-level photo.

Diagram as Privacy Protector

Experts also identified privacy benefits related to using an aerial diagram. Because the diagram can abstract away or hide certain sensitive details, experts believed it allowed for more sensitive investigations to be crowdsourced than would typically be possible. John explained:
CHAPTER 4. GROUNDTRUTH: AUGMENTING EXPERT IMAGE GEOLOCATION THROUGH SHARED REPRESENTATIONS AND CROWDSOURCING

There wasn’t much that was actually shared because it was only the diagram… There was no context that was actually given to the crowd worker. Whereas if you give someone a photograph, let’s say for a war zone or something like that, then that can contain the kinds of information that you necessarily don’t want to give out to other people.

George compared the diagram to police sketches of suspects drawn with the help of eyewitnesses:

> It’s not a photograph of the actual suspect. It’s a rendering of the interpretation from both victim and artist working cooperatively. This is the mapping equivalent of that where you’ve got an analyst who’s rendering based on a slice of imagery but then it gets provided to the crowd… I don’t think we’d run into any privacy issues with that.

4.5.2 Gridded Map Overlay

Grid as an Organizational Tool

Some experts would search in a linear fashion through the cells, ruling them out one by one, only deviating when they saw potential matches along the way. Others would search through the grid based on where crowds had (or had not) provided feedback. George and Kim said that the grid allowed them to keep track of where they had searched before, because without it, there would have been a higher tendency to wander across the map haphazardly.

Lorenzo, John, Jack, Aqwam, and Kim extensively utilized the Exclude button to mark cells they had examined and ruled out. Beyond excluding cells, Lorenzo, John, and Kim expressed the desire to have a Possibly button for them to mark cells that they had examined but not
4.5. RQ7. What Were Experts’ Experiences of Using GroundTruth?

fully ruled out.

Grid as a Coordination Tool

Experts suggested that the grid can be used not only to structure their search pattern, but also to aid in coordination between an expert and crowds. Kim and John proposed that another potential use of the Exclude button could be to rule out areas so that crowd workers would not have to review them, saving them effort and time. Dakota and John highlighted how GroundTruth’s grid could allow them to geolocate images with their colleagues in a collaborative manner. Dakota explained:

[I would say,] “I need your help geolocating, let’s go.” Then all of us focus on it, and really just to narrow it down, and they say, “Okay, now we’ve got, you know, 70% or whatever of this grid looked at a cursory glance.” I can take it from [t]here. I think that would be fantastic.

However, Neil pointed out that because of these experts’ limited time, and difficulties in coordination, it may be difficult to find experts.

John and Kim said that in the past, they had divided up a search area into a grid to collaborate with others, either by sharing coordinates or printing out an entire map and indicating who should search where. John described how he and his colleagues were trying to find Austin Tice, a US journalist and former Marine who was abducted in Syria, based off a video that was released of him where his captors were driving through a valley. He mentioned how GroundTruth’s grid would have been useful in delegating work in that investigation:

It’s several tens of square kilometers, maybe a mountain range, between Syria and Lebanon and that’s what we actually did on Google Earth. We started to
draw this grid of search areas and started to actually go through bit by bit... We’ve talked about it several times, “Oh we wish that we had something like this, that would actually help us structure the delegation of work in a smarter way.”

John added that experts do not necessarily perform the same tasks with one another, but that “it’s one person finding something, and then they call on others to actually verify that they agree with that finding.” Therefore, he wished that each cell was labeled with an identifier so that it would be easier to refer to when coordinating a search with colleagues: “Hey go look at B1 or B30 or whatever.”

Grid Non-use

Three experts (Jack, Ben, Neil) did not use the grid to structure their search. In one session, Jack, thinking that the correct location was located near a water body, immediately began searching along the length of a canal. Similarly, Ben and Neil directed their search by first looking at areas that they thought had matching architectural and geographical features.

Furthermore, Jack wished that the grid overlay could be turned off completely, while Neil, George wished that the opacity of crowd worker feedback could be varied. Jack stated a preference for searching through the map without any labels or satellite imagery, which he found distracting at times, and view only the outlines of buildings and roads. He often looks for matching shapes in the map first, and then looks to verify the location further once he has a potential match.
4.5. RQ7. What Were Experts’ Experiences of Using GroundTruth?

4.5.3 Heatmap of Crowd Feedback

Feedback Usage Patterns

Most experts directed their search based on crowd feedback, and worked in real-time alongside crowd workers. Some experts (Dakota, Kim, John, Alec) described prioritizing cells with high crowd agreement. Others (Lorenzo, James_FP, Aqwam) focused on areas that the crowd had not yet provided feedback on.

Most experts also avoided looking at areas with low crowd agreement, other than to check how crowd workers were performing. For example, Alec said:

They have all the red squares at the bottom, I didn’t waste my time looking at that. With the middle area, they said it was more likely to be in there so I looked more in there. It gave me more of a baseline to look at, we weren’t starting from scratch.

All experts used the **Hide/Show Color** buttons when closely inspecting cells in order to hide crowd feedback. Unexpectedly, some experts, such as Ben, John, and Kim, also used these buttons to deliberately—but temporarily—hide crowd feedback to avoid being ”biased” by it. These experts elaborated that if they found a likely match first, and others independently agreed with them, it could provide valuable confirmation. Ben described a prior experience geolocating an Instagram video of an Islamic State fighter with his colleagues:

We generally double blind each other and we’ll then measure up against work...So nobody even discussed what they were even hypothesizing or looking at until we were all finished...It could lead to some bias or something being overlooked.

On reflection, these experts noted that the nature of their work shifted from actually doing
geolocation towards verifying crowd feedback. This change was described as more of a spectrum rather than a dichotomy. For example, John reflected that “I was waiting for the inputs to come from others and then start verifying their findings...that kind of changed the way that I would normally run the process.”

Jack, Ben, Dakota, and Neil found the correct location before crowd workers returned feedback for that location. Jack and Ben did not rely on crowd feedback, while Neil sometimes made use of it. Dakota relied on crowd feedback initially, but once she had inspected all cells that crowd workers thought matched, she went on to inspect cells that had no feedback yet.

Feedback Accuracy and Utility

Overall, experts said that crowd workers performed well with the diagrams they were provided. Neil and Kim were initially skeptical of crowd feedback, but then grew to trust it. Others (Lorenzo, Alec, Dakota) trusted crowd feedback from the beginning.

Dakota, Ben, George, and John said that the most helpful part of crowd feedback was that it helped to rule out locations, so that these experts did not have to search there. For example, Dakota said, “The thing that [the crowd] did the best was specifically blocking out areas that didn’t include these features, or it didn’t include them all together, which is super helpful.”

Feedback Speed

Most experts were impressed with how quickly crowd feedback streamed in. However, experts had mixed opinions about whether they wanted crowd feedback to appear in real time or to be asynchronous. Dakota and George identified certain time-sensitive applications in their work that would benefit from real-time crowd feedback. Reflecting on his use of GroundTruth,
George said:

The most obvious benefit still is from saving time, being able to distill that larger map into regions where I could direct my attention much more quickly than I would have been able to do on my own...I think it’s very exciting what you guys are doing...Let me take this opportunity to tell you that what you’re doing could save lives in that type of [time-compressed] scenario and never forget that.

In contrast, Jack enjoys the challenge of geolocating images himself, and felt that he would only resort to using GroundTruth when he is stumped and wants to take a break. He explained, “I could see myself like sort of turning this on, drawing the square, going to get a cup of coffee or something, coming back and then seeing where can I start looking, where are the dark greens, where are the light greens.” Similarly, Ben juggles multiple projects at different stages, and said that it would be useful if he could submit images to be geolocated, and then review crowd feedback at a later point of time.

**Trusting the Crowd**

Experts voiced several considerations that affected their willingness to trust crowd feedback provided by GroundTruth. Building on the above discussion of cost and incentives, Kim questioned whether paid crowds would complete the task in good faith, saying, “Some people might be just like, ‘Let me do this as fast as I can so I get money,’ and then some people might be legitimately enjoying the experience. So, that’s generally the plus and minus of using [Amazon Mechanical] Turk workers.” Relatedly, Jack voiced concerns about Mechanical Turk workers providing crowd feedback because he said their anonymity limits accountability.

Taking this idea further, Ben worried about the potential for adversarial crowd workers to hijack or mislead a GroundTruth investigation:
oftentimes the geolocation happens to be around something political. Particularly if it’s a war crimes investigation, or some sort of criminal act that’s being investigated. Inherently, there’s going to be opinions that attempt to [alter] the perception of the outcome, whether that means brushing it under the rug, whether it means hyperbolizing it, etc.

John wished that the system could be set up such that he could work only with people that he trusts and have been previously vetted:

That we could actually split the works work amongst ourselves so not necessarily outsource to anybody else. But we would need a controlled platform where we can bring in trusted partners and people we know, [and] people who know what they’re doing, basically, and then actually split the tasks between them....

Cost and Incentives of the Crowd

Two experts who worked for the private sector (Alec, Neil) said they felt that the actual cost of running their sessions ($2 per crowd worker, approximately $60–120 per search area) was very reasonable given the speed with which feedback streamed in. Kim, a journalist, said that while paying workers would be useful for breaking news desks, it would depend upon each organization’s budget.

Furthermore, although several experts felt that workers should be paid a fair wage for their time, John and Jack, who work with a volunteer-based group of open-source investigators, said they doubted that paid crowds would be necessary for their work. They had the ability to mobilize a large number of volunteer crowds on social media, and had already done so for investigations in the past. Building on this idea, Ben was enthusiastic about GroundTruth’s ability to democratize open-source investigations by allowing non-experts to make valuable
contributions, as well as empowering experts:

…it demonstrates the opportunities to do verification at scale, that is truly unique. And I think that really empowering the much larger investigations, is a really significant value add for the research and journalism communities. Especially when there are a limited number of people who have these skills.

4.6 Discussion

4.6.1 Designing Shared Representations for Image Geolocation

We designed GroundTruth to help crowds augment experts’ complex sensemaking task of image geolocation, drawing inspiration from Heer’s notion of shared representations in mixed-initiative systems [149]. In this section, we reflect on the successes and challenges in adapting the principles of shared representations from AI to crowds within the context of image geolocation.

Shared Lens

GroundTruth provided a shared lens between experts and crowds to support Heer’s first principle of providing the user with significant value and promoting efficiency, correctness, and consideration of alternatives. The shared lens took the form of an aerial diagram drawn by the expert and shared with the crowd.

Diagramming for a crowd. Our prior work [180] showed that crowds using just a ground-level photo to search satellite imagery lead to unacceptably high false negatives, whereas a
perfectly drawn aerial diagram dramatically improved crowd performance. However, little was known about how well diagrams drawn by real geolocation experts would fare. The positive results of our evaluation suggest that, following Heer’s first principle, real diagrams do enable valuable crowd performance, helping the expert work more efficiently and correctly by prioritizing high-agreement cells. Aerial diagrams provided a shared lens that helped close the expertise gap between experts and crowds in two ways. First, they bootstrapped the expert’s traditional process to leverage spatial reasoning and mental rotation skills that novice crowds lack. Second, experts drew on experience to focus the crowd’s attention on key elements (i.e., permanent, unique), while omitting the rest.

More broadly, our evaluation illuminated how drawing an aerial diagram for one’s self differs from one intended to be a communication tool for an unknown crowd of novices. Experts also described additional benefits of the diagram we had not considered, such as its privacy-preserving attributes that can protect both the investigation and the crowd. However, the dual goals of a diagram in the context of GroundTruth also raised new challenges and concerns. Some wanted to include additional details and context that would not benefit themselves, but might help clarify meanings for the crowd. Prior work has found that such shared context is important when dividing a task into microtasks [274], even when performed by the same person (e.g., selfsourcing [309]).

Providing these details would require extra time and labor, so experts also suggested better tools for rapidly drawing diagrams, such as a library of symbols for common objects (e.g., buildings, intersections, bridges), as seen in creative contexts [189]. Besides speeding up the drawing task, regular use of these tools could further streamline communication between experts and crowds.
Benefits of drawing experience. While all experts had deep experience with image geolocation, not all of them were familiar with aerial diagramming, resulting in lower-quality diagrams. This gap speaks to the multi-step process of traditional image geolocation, where experts only resort to brute-force satellite imagery analysis when previous steps are insufficient [54, 180]. While our diversity of experts and images prevents direct comparisons, it is suggestive that the three experts who voiced the least confidence and experience with diagramming (John, Kim, and Alec, see Section 4.5.1) also returned the largest distances from the target location (700m, 400m, and 200m, respectively, see Table 4.4). These results point to the benefits of expert training or experience in drawing diagrams for themselves as well as the crowd.

Shared Environment

We designed GroundTruth to provide a shared environment in the form of a gridded map overlay. Heer’s second principle specifies that agents should augment but not replace user interaction, blend in nondisruptively, and be easily invoked or dismissed. Along these lines, crowd feedback is visualized on the same grid that the expert uses to search the imagery, but can be toggled on and off, while still retaining the experts’ own decisions to exclude cells.

Support selfsourcing. We designed GroundTruth’s gridded map overlay primarily to support expert–crowd interaction, explicitly prioritizing experts’ decisions over the crowd’s. Most experts emphasized how it enabled them to visualize and act on crowd feedback while retaining agency [149]. However, we were surprised by how many experts found the grid to be helpful per se in enabling them to systematically search a region and mark off cells using the Exclude button. For these experts, the grid supported a selfsourcing [309] practice not readily available in existing tools. One expert further suggested the ability to judge a cell
as Possibly a match, in addition to the current Exclude option, to flag it for closer inspection after an initial pass.

Symbols and structures. Experts suggested mechanisms within the grid that could more effectively support collaborative search. One suggested unique identifiers for each cell to enable easy reference when communicating with colleagues. If implemented, such a feature could leverage existing geographic identifier schema such as What3words [15] for integration with broader volunteered geographic information (VGI) efforts. Another expert suggested the ability to label or add notes to individual cells, either for themselves or direct the crowd in a more nuanced manner.

Shared Analysis

Heer’s third principle embraced augmentations that required neither perfect accuracy nor exhaustive modeling of the user’s task to be useful. Likewise, GroundTruth supported a shared analysis where crowds would contribute to one module of the broader image geolocation pipeline: satellite image search. Further, as our prior work suggested that an ideal setup yields a 50% false positive rate [180], we anticipated crowd analysis would be useful, but imperfect.

Crowd analysis with real experts. Our evaluation found that crowds augmented experts’ work in ways that were useful without being completely accurate or exhaustive. Even with real experts, authentic diagrams, larger search areas, and more diverse images, crowds reduced the search areas by an average of 43%, comparable to the false positives for an ideal setup seen in prior work. Indeed, there was only one session (of 10) where both the crowd and expert missed the correct location (see Section 4.4.2 for more details). Taken together,
our results not only show that the crowd’s shared analysis augmented expert performance in realistic settings, but also suggest that experts might do better to pay closer attention to the crowd’s feedback.

**Optimizing crowd allocation.** While experts valued crowd feedback, they also suggested ways that GroundTruth could better allocate worker effort. For example, the system hires multiple crowd workers to review every cell, regardless of the expert’s judgements. However, if an expert excludes a cell not yet reviewed by the crowd, the system could remove those tasks from the worker queue as redundant, allowing workers to focus on other cells. While prior work have also implemented multi-step aggregation and review pipelines [44, 173, 215], in this scenario, workers released from redundant cells would be dynamically reallocated to high priority cells to provide another layer of review prior to expert verification. An expert also requested the ability to define the initial investigation area as a polygon rather than a square, to capture nuances of geography and avoid low-probability areas like mountains or bodies of water.

### 4.6.2 Generalizing Shared Representations

While this paper focuses on image geolocation, we suggest that shared representations can be adapted to other complex tasks and domains where crowds support expert investigative work. Most similarly, shared representations could enable other types of crowdsourced satellite image analysis, such as natural disaster damage assessment [299]. They could also help with other types of “needle-in-a-haystack”-type visual search tasks, such as identifying objects to combat human trafficking [11], identifying people in historical photos [233], or finding missing pets after a crisis [334].

More broadly, shared representations may benefit sensemaking activities for which searching
is only one of many foraging and synthesis tasks. As discussed previously, image verification is a sensemaking activity encompassing not only image geolocation, but also consideration of other types of visual clues in the photo of interest, as well as the broader context about who posted the photo, when, and why. This and other sensemaking domains also require analysis of non-visual material, expanding the notion of a shared lens to other media. For example, crowds could augment expert analysis of textual data to perform bottom-up qualitative content analysis [32], to investigate evidence documents to solve a crime [137], or to compare features when shopping for an unfamiliar product [178].

Finally, it may be possible to extend the utility of shared representations beyond the analytic tasks addressed by this paper and by Heer [149], which pose unique constraints [112]. Systems like Crowdboard [31] and Apparition [189] illustrate how crowd-augmented expert work can support generative tasks like ideation and prototyping, while other models of expert–crowd interaction support creativity in domains like writing [172]. Designing such shared representations, and understanding how they must differ from those presented here, requires a detailed understanding of the activities that a user engages in during the sensemaking process [244]. We can gain this knowledge, as in GroundTruth’s example, by working closely with real experts to understand their current practice, needs, and attitudes.

4.6.3 Broader Impacts

GroundTruth has the potential to save expert investigators time and scale up their expertise, as well as possibly save lives. It contributes to democratizing the field of visual investigation, empowering newcomers and novices to help debunk misinformation, and responds to a growing need for tools to support information credibility assessment [320]. However, technology can also have negative impacts [148]. Indeed, GroundTruth could be used by oppressive
governments to geolocate images, or hijacked by troll farms to skew investigations. Further, by supporting visual investigations, we may contribute to the growing atmosphere of surveillance and indifference towards privacy. Overwhelmingly, however, we believe GroundTruth can be used to do more good than harm.
Chapter 5

CuriOSINTy: Designing a Platform for Combating Misinformation through Collaborative Capture the Flag Competitions

5.1 Motivation

Mis- and disinformation harms societies by lowering trust in civic institutions, in experts, and in one another. Functioning democracies rely on expert investigators in domains including journalism, law enforcement, and human rights activism to combat misinformation. This includes identifying foreign influence operations, holding governments and corporations accountable, and even correcting history [].

While the growth of information and communication technologies provides investigators with greater access to information [184], the rapid pace and large volume of content produced can further overwhelm investigators. Investigators also face challenges in researching misinformation due to limited personnel, tools, and access to proprietary data [225, 315]. The wide scope and highly contextual nature of misinformation also requires investigators to develop
multiple skill sets in software development, data science, and legal research, among others [225]. Investigators must also act quickly to address misinformation during emergent crises and important events (e.g., elections). If investigators fail to respond quickly, miss important evidence, or misinterpret information, it can result in the loss of lives and the destabilization of democratic institutions — such as the January 6th attack on the US Capitol [5].

Solutions for scaling up investigators’ work practice include leveraging automated or mixed-initiative tools [79, 145, 283] and crowdsourcing [1, 18, 95, 166, 256]. Automated and mixed-initiative tools have shown some promise in supporting investigative work. Examples include CrowdTangle, a platform for discovering and monitoring viral content on social media [79]; Hoaxy, a Twitter network visualization tool [283]; and Claimbuster, a fact checking API that uses natural language processing techniques [145]. However, investigators are wary of becoming overreliant on tools, especially in high-stakes settings [166]. This is because the highly contextual and often disguised nature of misinformation may result in tools being insufficiently accurate [321] or unable to detect coordinated inauthentic behavior [ ]. Tools that investigators use may also become obsolete because of two factors: 1) frequent changes in social media platform features and APIs [225], and 2) an overreliance on individual developers for long-term software development and maintenance [38, 193].

An alternative approach for addressing misinformation is to leverage the powerful and adaptive capabilities of online crowds to devise useful solutions and synthesize diverse information sources. There are three general ways to leverage crowds across the collaboration spectrum [60, 268]: 1) traditional, 2) collaborative, and 3) competitive crowdsourcing. In traditional crowdsourcing, a requester creates atomic tasks that crowd workers complete independently, with little to no interaction with each other [e.g., 18, 166, 256]. Crowds in collaborative crowdsourcing engage in long-term and frequent interactions based on shared goals, with shared decisions and resources [256]. Finally, in competitive crowdsourcing, crowds may
share a common goal but engage in short-term and infrequent interaction, where decisions are made independently by crowd workers with minimal information shared between each other [2, 4, 157, 331].

The HCI and CSCW community has extensively studied traditional [95, 175, 216], collaborative [41, 65, 249] and competitive [41, 217, 308, 331] crowdsourcing environments. Our work here is concerned with the latter two types of crowdsourcing: collaborative and competitive. While collaborations and competitions each have their own benefits and challenges, prior work in settings such as software development, machine learning, and design innovation has found benefits to intentionally combining elements of competition and collaboration [41, 160, 190, 217, 308]. For example, while competitions may be inefficient due to duplicate work and siloed information between competing teams [41], allowing teams to collaborate by sharing solutions and information with each other led to better overall performance [183]. Further, while collaborations may suffer from groupthink and inaccuracy blindness [41, 217], encouraging competition through gamification and adopting an adversarial mindset (“red-teaming”) mitigated these challenges [245].

While combining competitive and collaborative features in crowdsourcing has shown promise in other contexts, there are three challenges to adapting current approaches to the context of combating misinformation online. First, misinformation exists within a dynamic and highly contextual setting, requiring an efficient and rapid response [296]. Second, and as a result of the highly contextual setting, combating misinformation requires deep investigative expertise that novice crowds may not possess [20, 35]. Third, investigative work requires synthesizing diverse information sources that crowds may not have access to [225]. I thus framed the following two research questions:

- **RQ1:** How can we merge the complementary benefits of competition and collaboration
5.1. Motivation

to provide an efficient and rapid response to misinformation?

- **RQ2**: How can we support novice crowds in conducting investigations about misinformation that are traditionally only the purview of expert investigators?

To answer these research questions, I built CuriOSINTy [324], a platform that enables crowdsourcing through collaborative capture the flag competitions (CoCTFs). I developed the CuriOSINTy platform through a Research through Design (RtD) process [350] that involved problem framing and extensive prototyping followed by a four month-long process of iterative development and deployment with 46 participants.

To motivate crowds and provide a greater sense of urgency, CuriOSINTy borrows the concept of capture the flag competitions (CTFs) from the field of cybersecurity [58, 168, 331]. To ameliorate the disadvantages of competition, such as duplication of effort and information silos [41, 308], CuriOSINTy incentivizes information sharing and collaboration between competing teams. I augment the capabilities of novice crowds by providing expert training and guidance, as well as leveraging publicly accessible open source intelligence (OSINT).

Compared to traditional crowdsourcing approaches, CuriOSINTy fully leverage the creative and adaptive capabilities of crowds by giving them greater agency in determining how to combine techniques and tools. However, to structure and improve the quality of work, the platform includes scaffolding and rubrics from the field OSINT.

Through our mixed-methods evaluation, I found that CuriOSINTy merged the benefits of competition and collaboration, enabling useful sociotechnical interactions. For example, CuriOSINTy enabled participants to quickly discover, archive, verify, and report on hundreds of pieces of potential misinformation on social media. Participants also said that they enjoyed using CuriOSINTy and that it helped to better structure their workflows as they worked within their team and with other teams. Our RtD process and mixed-methods evaluation also
highlighted tensions between competition versus collaboration, and in-depth versus broad investigations. Finally, I reflect on our experiences conducting an RtD study, on designing with appropriation in mind, and on enabling a CoCTF. I also present implications for HCI and CSCW more broadly.

In summary, our work makes four contributions:

1. This study makes a conceptual contribution by introducing collaborative capture the flag competitions to support a rapid and efficient response to misinformation.

2. Using a Research through Design process, I developed the CuriOSINTy platform. This makes a system contribution by operationalizing collaborative capture the flag competitions (CoCTFs).

3. A mixed-methods evaluation through a semester-long deployment with 46 participants. Participants expressed an overall preference for using CuriOSINTy for structuring their investigations. Our evaluation finds that it also merges the complementary benefits of competition and collaboration, and that by designing for appropriation, I may design platforms with greater capability and longevity.

4. I present implications for building appropriable platforms, conducting crowdsourced investigations, and more broadly, designing collaborative competitions in other high-stakes settings.
5.2 Background: Combating Mis- and Disinformation Through Open Source Intelligence

Prior work has identified three approaches to combating mis- and disinformation online: agent-, message-, and interpreter-oriented [? ]. In addition, each of these approaches can be focused on either individual or a collection of: agents, messages, and receivers [? ]. Agent-oriented approaches are concerned with the specific actors that generate and spread misinformation, while receiver-oriented approaches are focused on the targets of misinformation and how receivers are affected.

Most closely related to our work here are message-oriented approaches that are concerned with: a) identifying content that is potential misinformation and b) verifying or refuting claims associated with or made by the content. While the growth of publicly available information and communication technologies has made misinformation more prevalent, it has also democratized access to large amounts of sensitive data and powerful tools to analyze it [129]. In turn, this has led to a well-established investigative field of open source intelligence (OSINT) Note that “open source” in OSINT has a different meaning than in open-source software (OSS). In OSINT it refers to using publicly available data and tools, but the code for these tools does not necessarily need to be publicly available for others to use, study, change, or distribute [289]. that is focused solely on collecting and analyzing publicly available data to generate intelligence for a specific purpose [339]. OSINT has been widely used by media agencies [? ], civil society organizations [? ? ], governments [339], and online sleuths [28] to combat misinformation. OSINT is also popular in other application domains, including business [15,75 - chi22], cybersecurity [28-chi22], and counter-terrorism [69-chi22].

Prior work [41, 339] shows that OSINT cycle consists of four steps: 1) discovering content; 2) verifying its provenance and determining the the veracity of claims made; 3) archiving the
content to prevent information from being lost; and 3) reporting on the results of the investigation. Prior work has also focused on developing systems to support investigators with each of the four steps in the OSINT cycle. For example, CrowdTangle [79] and Algorithm Tips [95] support automated content discovery; Hoaxy [283] and DejaVu [? ] support verification; Hunchly [3] and the Web Archive Workbench [156] support archiving content; and Maltego supports reporting [? ]. Given the growing number of OSINT systems, Abdullah et al. created a OSINT Explorer, a system to help identify which OSINT tool to use for a given purpose [19].

While the majority of prior systems supported individual steps in the OSINT cycle, CuriOSINTy is designed to support all four steps. CAPER is another tool that supports all four steps in the OSINT cycle. It is designed to enable collaborative information, acquisition, processing, exploitation, and reporting. CAPER enables the collection and analysis of images and video; multilingual text and audio; natural language processing (e.g., named entity recognition); and facial and audio biometric recognition.

CAPER [25] was designed for law enforcement agents with the goal of preventing organized crime, whereas CuriOSINTy is focused on combating misinformation on social media. In addition, CAPER is designed to be used by an individual — the authors use the word “collaborative” to refer to content created by multiple users. In contrast to CAPER and most OSINT tools that are catered towards individual users, CuriOSINTy supports virtually any number of users collaborating within and across teams on one or more investigations. Lastly, while investigations have been well studied within CSCW and HCI, OSINT — despite its popularity in practice — has garnered less attention [25, 41? ]. Through this work, we show that OSINT can be a valuable framework for the CSCW and HCI community.
5.3. RQ8. Developing CuriOSINTy Using Research Through Design

Misinformation on social media is a wicked problem [267] because it is a symptom of another problem (e.g., political polarization or psychological biases), it can be interpreted and solved in many different ways (e.g., social, psychological, or technological), and solving it is identical to completely understanding it and there are no clear criteria for sufficient understanding [7, 203, 234].

To address wicked problems, Zimmerman and Forlizzi [350] propose an approach to conducting research called Research through Design (RtD). RtD is the process of iteratively designing and critiquing an artifact that acts as a proposed solution to a wicked problem. Solutions to wicked problems are not right or wrong, but good or bad depending on the initial framing [43]. Engaging in RtD enables researchers to investigate what a potential future might look like so as to reframe the wicked problem [351].

Artifacts produced through RtD can also help inform new theory and future work, provided the RtD process is well-documented. The RtD process includes four phases [83, 181, 351]: 1) framing the problem; 2) creating prototypes to better understand the problem; 3) clarifying the ideal preferred state through design goals; and 4) deploying and iterating on the solution. Here, we engaged in a RtD process to develop CuriOSINTy, a platform to enable collaborative CTF competitions that leverage OSINT to rapidly and efficiently combat misinformation. Next, we describe the four phases of our RtD process, from framing the problem (Phase 1) to deployment and iteration (Phase 4). Then, we describe the resultant platform.
Figure 5.1: Our Research through Design process involved four phases: 1) framing the problem, 2) prototyping workflows and interfaces, 3) clarify design goals, and 4) deploy and iterate on design. New features introduced at each phase include: (a) Points for competition, leaderboard, judging, task system, openly accessible information, and archive and reporting flags. (b) Points for collaboration, scaffolding, self-assessment rubric, and API for tools and tasks. (c) Narrative threads, evidence pieces, nested flags, and filters.

5.3.1 Phase 1: Framing the Problem

In response to Carroll and Kellog’s criticism that the thing often proceeds theory in HCI [59], Zimmerman and Forlizzi suggest that things in RtD should be informed by current theory and practice while spawning new theory and practice through the design and evaluation process [350]. Below we describe our engagement with practice and findings from prior work.

Our Engagement with Practice. Two of the authors also have extensive experience in designing and evaluating crowdsourcing systems to support sensemaking [anonymized for review], as well as prior experience studying competitive and collaborative online communities [anonymized for review]. In addition, one of the authors has participated in large-scale collaborative investigations, and all authors have taken part in several investigation-oriented competitions, including hackathons and OSINT CTFs.

Findings from Prior Work. The research team also engaged with the literature on: combating misinformation through OSINT, crowdsourced sensemaking, and collaboration in
competitive environments (see Section ??). We found five influential themes.

First, expert investigators who seek to combat misinformation require additional resources, such as personnel [225]. Second, while competitive and collaborative crowdsourcing have been used to support investigators in combating misinformation [41], little work has explored how to combine both approaches in this context. Third, CTFs show promise in attracting and motivating large, novice crowds through non-monetary incentives [58, 68]. However, it is unclear how to adapt CTFs that are traditionally theoretical for a real-world application, as well as how to introduce elements of collaboration into a CTF. This resulted in our first research question:

**RQ1.** How can we merge the complementary benefits of competition and collaboration to provide an efficient and rapid response to misinformation?

Fourth, due to the dynamic and highly contextual nature of misinformation, combating it requires significant experience with the context, techniques, and tools that novice crowds may not possess [20]. Fifth, such investigative work may also requires synthesizing diverse information sources that novice crowds may not have access to [339]. It is unclear how to support novice crowds in conducting investigations into misinformation. This includes how to provide novices with relevant context, techniques, and tools — and more importantly — access to information. This resulted in our second research question:

**RQ2.** How can we support novice crowds in conducting investigations about misinformation that are traditionally only the purview of expert investigators?
5.3.2 Phase 2: Prototyping Workflows and Interfaces

To better inform the design of the CuriOSINTy platform, we prototyped interaction workflows and interfaces with ten members of our lab who had prior investigative experience to act as a small crowd. Our prototyping process was conducted over several weeks, where we piloted different workflows and created low- and high-fidelity prototypes.

**Pilot Collaborative and Competitive Workflows.** To choose an optimal workflow for CuriOSINTy, we piloted ways to blend collaborative and competitive approaches using Google Forms, Docs, and Sheets. In both cases, crowd workers completed three types of tasks discovering, archiving, and debunking potential misinformation.

We provided the crowd with topics to investigate such as COVID-19, election security, and financial misinformation. Then, we asked them to search for these topics on social media platforms such as Twitter and Facebook to identify potential misinformation that was recently posted and appeared to reach a wide audience (i.e., large number of shares/retweets). Crowd workers submitted content that they found through a Google Form.

Note that potential misinformation refers to content that has not been debunked (verified or refuted), but appears to be falsifiable. Archiving involves saving the content through a screenshot or by using an online archival service, such as the Internet Archive. Debunking content involves verifying or refuting the claims made in the content.

**Collaborative workflow.** To overcome the limitations traditional crowdsourcing workflows[265], we explored how to provide greater flexibility and agency to the crowd, and to allow crowd workers to share information with each other. Thus, we piloted a collaborative workflow where crowd workers worked together, with no restrictions on which tasks an individual could perform or the order of operations. Crowd workers submitted content through the
submission form, including the URL of the content, a short summary contextualizing the content, and a description of how it was found or verified.

Here, we found it difficult to simultaneously coordinate work among all ten crowd workers without reverting to using microtasks, especially when everyone was focused on investigating the same topic.

**Competitive workflow.** To overcome the limitations we found in the collaborative workflow, we explored how to incentivize teams to work quickly and efficiently through a competitive CTF workflow. This workflow consisted of four features: teams, points, a leaderboard, and judging.

First, we explored how to more easily coordinate efforts between crowd workers by dividing the crowd into teams of two or three. To help streamline each team’s efforts, we also asked them to choose a team leader and a unique topic to investigate.

Second, to make it easier to compare performance across teams and to make the evaluation more objective, we assigned point values to discovery flags (finding content relevant to the topic) and verification flags (debunking claims made within the content). Discovery and verification flags were worth five and 20 points, respectively. Third, we heightened the sense of urgency and competitiveness in the CTF by creating a leaderboard in Google Sheets. The leaderboard displayed the cumulative points earned by each team in a graph and updated every 15 minutes.

Fourth, to increase the quality of the crowd’s work through frequent feedback [102, 223], we introduced the concept of judging where experienced crowd workers provided feedback on flags created by others. This included whether a flag was relevant to the topic, if the contents were falsifiable, and if they were able to accurately debunk claims made within the content. To help judges evaluate a crowd workers’ submission, we asked crowd workers to
include a link to a Google Doc that described their investigative process in greater detail than was included in the submission form. To provide feedback, we used Google Doc’s comment feature, and the points were tabulated using Google Sheets.

From this workflow, we identified four themes. First, we learned that certain types of content were more difficult to discover, and that some verifications required more time and effort than others — indicating that they should be worth more points. Second, we found that judging flags was time consuming and decided to make judging quicker by providing greater structure to flags.

Third, some teams performed better than others due to differences in their composition (e.g., technical vs. topical expertise) and tools used (e.g., using reverse image search vs. manual searches). Fourth, we learned that most teams were often unaware of what other teams were working on.

Ultimately, we settled on a competitive workflow for two reasons: first, we found that it greatly reduced coordination challenges found in having a larger, collaborative crowd, and second, we found that the CTF motivated the crowd to work quickly and efficiently.

**Design Low- and High-Fidelity Interface Mockups** After settling on a CTF as a starting point for our workflow, we needed to overcome the limits of competitions by introducing beneficial elements of collaboration. Thus the research team engaged in a cyclic process of brainstorming and designing possible new features for a collaborative CTF (CoCTF). To help illustrate what these features would look like in practice, we created low-fidelity and high-fidelity mockups of the interface.

*Low-Fidelity mockups.* After our first brainstorming session, we created low-fidelity mockups (hand-drawn and on Balsamiq), and solicited feedback from our lab. We arrived at two ways
to support collaboration: a crowdsourcing task system and openly accessible information. First, we wanted teams to be able to support each other when a particular task proved too cumbersome for any one team. This resulted in a task system — similar to Amazon Mechanical Turk — that teams could use to crowdsource parts of their work to others.

Second, to increase awareness between teams [108, 308] and further support collaboration, we made all information on CuriOSINTy openly accessible to every user, regardless of their team membership.

**High-Fidelity mockups.** To illustrate in more detail what the CuriOSINTy platform would look like, we used Google Slides and lightweight HTML and CSS to create high-fidelity mockups. The mockups resembled Reddit’s compact card view with tabs for flags and tasks. The initial flag types were discovery and verification. After reflecting on this design and engaging closely with literature on OSINT [289, 339], we decided to structure the crowd’s sensemaking process, without introducing a more rigid workflow that prior work found constrains crowdworker creativity and adaptability [265]. We thus created four different flag types that directly mapped on to the four steps in the OSINT cycle [41, 339]: discovery, archival, verification, and reporting flags.

### 5.3.3 Phase 3: Clarify Design Goals

By combining prior work and our prior experiences, as well as engaging in a reflective and iterative prototyping process, we arrived at the following three design goals: 1) support an efficient and rapid-response to misinformation; 2) give the crowd agency; and 3) support a range of tasks and tools. However, it is challenging to design a software that meets these goals in a complex setting involving a large number of people coordinating their actions [132].
One promising approach in CSCW for navigating this complexity is designing for appropriation \[98\]. That is, instead of trying to understand, model, or anticipate all of the features of a complex system, it makes sense to design solutions that can be used in diverse and dynamically reconfigurable ways — thus creating more robust solutions for complex problems \[132\].

Following five of Dix’s heuristics for supporting software appropriation \[98\], we describe how we instantiated our design goals. Dix suggests in his first heuristic that designers expose the intentions behind the system, that is, making design assumptions and decisions explicit, and “if they are wrong then they [can] be re-examined” \[98\]. Along these lines, we explicitly exposed the intentions behind the CuriOSINTy platform through our three design goals:

**Goal 1.** Our primary goal for CuriOSINTy was to enable an efficient and rapid-response to misinformation on social media and support expert investigators. We instantiate this goal by using two of Dix’s heuristics.

Dix’s second heuristic that we incorporated is to encourage sharing between teams their various setups and appropriations of technology. Similarly, prior work studying competitions found that when communitition was encouraged (collaboration among competing teams), the individual \[308\] and collective performance of teams was higher \[49, 183\]. In CuriOSINTy, we explicitly enable teams to learn from and collaborate with each other by making all information openly available between teams. We also explicitly incentivize collaboration by rewarding teams that collaborate with additional points.

The third heuristic we incorporate is to provide visibility to users, making clear how the platform works so that the users can devise their own uses. CuriOSINTy not only makes all information accessible to all users, but it also displays the current status of various actions. For example, whether a piece of evidence is completed, or if a flag has been approved or rejected by a judge. Further, the platform shows users maximum number of points that can
be obtained through each flag type as well as how many points that judges awarded their flag.

**Goal 2.** Our second goal was to give the crowd agency. This is in contrast to typical crowdsourcing systems where crowds are given specific microtasks, with limited agency in how to complete them [53, 223]. This follows Dix’s fourth heuristic to support not control users’ actions, but rather to provide the necessary functionality so that users can achieve the desired task without detailed instructions.

In CuriOSINTy, experts provide some, but not complete, direction to the crowd. The expert specifies which misinformation narratives the crowd should track, and possibly which platforms to search. The crowd chooses how to search these platforms, which posts to investigate, and how to verify or refute a particular piece of misinformation. To enable novice crowds to conduct more complex investigations, we designed CuriOSINTy in tandem with a class to provide several weeks of training on each step in the OSINT cycle.

To further provide support, CuriOSINTy incorporates scaffolding and rubrics which have been shown by prior work to enable novice crowds to match expert-level performance [102]. CuriOSINTy provides scaffolding by dividing each narrative into multiple evidence pieces focused on a particular claim. In turn, each evidence piece is divided into four different flags (discovery, archival, verification, and reporting). However, the platform does not enforce ‘hard’ constraints on the exact process or order of completion, but encourages high quality submissions through the point system and through a self-assessment rubric. When a user submits a flag, they are also shown a rubric asking them to assess the quality of their flag along several parameters. This not only explicitly indicates to the user what a high-quality flag looks like, but also helps judges to evaluate flags quicker.
Goal 3. While providing the crowd with more agency can be beneficial, our prototyping process revealed that it was important to enable CuriOSINTy to support the crowd in using multiple tools and completing a range of complex tasks. However, prior work has shown that no single workflow can fully support complex work [265]. Dix’s fifth heuristic is relevant here.

Dix suggests that platforms be designed with *plugability and configuration* in mind to allow parts of the same (or different) platform to be combined in different ways. Along these lines, CuriOSINTy includes an *application programming interface* (API) and task system similar to Amazon Mechanical Turk. Together, they enable users to design their own automated or crowdsourcing tools that can be tied into the platform. In this way, roles that each user takes on are flexible and dynamic: users may be a member in a team competing against other teams; they may contribute to a competing teams’ work to gain points for their own team; they may design new tools for future users to use; and they may also act as traditional crowd workers completing microtasks generated by other users.

### 5.3.4 Phase 4: Deploy and Iterate on Design

RtD can create new situations and practices for researchers to investigate, producing gaps in behavioral theory and technical opportunity, while revealing design patterns around problem framings and specific interactions [350]. Thus, the primary aim of our study was to better understand emergent sociotechnical interactions enabled by CuriOSINTy as well as opportunities for improvement.

**Research and Class Setting** Due to the sensitive and highly contextual nature of investigations into misinformation, we required crowd workers that we could trust and hold accountable. In our design goals, we also noted the importance of providing the crowd with more agency and allowing them to take on complex tasks through extensive training. Given
these requirements, we deployed CuriOSINTy in a semester-long class and evaluated it using mixed-methods. All authors helped to design the class and teach students OSINT investigative skills through hands-on training. This enabled a tight coupling between the skills students learned in class and how they applied them while using CuriOSINTy.

To evaluate the CuriOSINTy platform, we required students with both topical expertise in misinformation and technical expertise in testing software and developing tools. Thus we advertised the class to senior undergraduate students and new graduate students in both the Computer Science and Political Science departments.

The class met online due to the COVID-19 pandemic. Twice a week for approximately 90 minutes, we taught students techniques and tools in each of the four steps of the OSINT cycle. Modules for each step lasted approximately three to four weeks and included both theory and practice sessions as well as additional readings.

**Participant Recruitment and Demographics** This study was approved by our university’s IRB. The first author recruited students during a guest lecture they gave in class and stated that participation was voluntary. To minimize the feeling of coercion, we decided not to provide extra credit to students for participating in the study. We provided consenting participants with $20 gift cards for completing a demographic survey, submitting weekly reflections, and taking part in at most two semi-structured interviews. For the final CTF, we provided the three top-scoring teams with prizes of $55, $45, and $35.

Twenty out of 46 students (from six out of 11 teams) consented to participate in our study. They ranged in age from 19 to 23, with a median age of 21. Fifteen (of 20) participants were majors in computer science or similar fields, and four were majors in political science or similar. Six identified as women, and 14 as men. None of the political science students had participated in CTFs prior to the class, while six of the computer science students had.
CHAPTER 5. CURIOSINTY: DESIGNING A PLATFORM FOR COMBATING MISINFORMATION THROUGH COLLABORATIVE CAPTURE THE FLAG COMPETITIONS

Figure 5.2: Table of all teams that used CuriOSINTy over the semester, their team size, the participants that we interviewed, their self-identified gender, and their degree major.

Classroom Deployment Procedure. We first demonstrated CuriOSINTy to the class in week six and described the motivation for creating it. We also asked students to engage in an iterative, participatory design process with us. Our deployment procedure involved the following three steps that we repeated every two weeks (for a total of six deployments between week six and 16).

1. **Demonstrate and use CuriOSINTy.** We presented a slide deck that described important features, as well as changes to existing features or new ones that were added since the last deployment. Then, students used CuriOSINTy to participate in a CoCTF. The research team acted as experts providing guidance to the crowd on what topics to investigate, as well as evaluating their submissions. Experts with background in human rights and vaccine misinformation led two of the CoCTFs.

2. **Reflect with students and research team.** Immediately after each CTF, we collected both verbal and written feedback from students. The research team engaged in a reflective practice of debriefing together after each CTF. The debriefs were not only for determining which student feedback to address, but also for sharing our own observations and feature suggestions.

3. **Implement new and refine existing features.** In the two weeks between deployments, the
research team implemented the features and changes that we decided upon during the debrief. We also updated the slide deck to indicate major new interface and interaction changes that we made to CuriOSINTy. This slide deck also served as a changelog for the research team to reflect on while writing this work.

Data Collection and Analysis Methods  We collected both qualitative and quantitative during and after each deployment. This included participant observation with detailed notes, notes from our weekly feedback and reflections, semi-structured interviews, and system log data.

Participant observation. The first author conducted participant observation which involved taking detailed notes of how students used and interacted with each other through CuriOSINTy and their verbal feedback on the system after each CTF.

Feedback and reflection sessions. After each deployment and during class, we asked students about techniques that worked well and did not; whether and how they contributed to other teams’ flags (engaged in collaboration); as well as what they liked about that week’s version of the CuriOSINTy system and what could be improved.

Semi-structured interviews. The first author interviewed 16 of the 20 students who provided consent (the other four were unavailable) and took detailed notes [290], which were incorporated into the transcript. The interview guide contained questions about students’ backgrounds; how they worked with their team members and other teams; their perceptions of the platform and how it had changed over time; their feedback on various system components (organization of threads, evidence, and flags, the incentive structure, collaboration, judging, the API); and reflections on what they learned.

In total there were eight interviews with members from six different teams, with one to
four team members in attendance for each. All participants were interviewed immediately before (KG6-9, SS-4, SS-5, OT-1) or after (SL-2, SL-3, KP14-17, DD-13) the final CTF. We interviewed DD-10 and DD-11 from Disinformation Defeaters before and after the final CTF because they wanted to provide additional feedback. The first author recorded all interviews using Zoom and fully transcribed the recordings. Interviews ranged from 58 to 85 minutes (average = 65).

**System log data.** To gain insight into how teams performed using CuriOSINTy, we collected system log data. This included which flags each team member worked on, their self-evaluation when submitting a flag for judging, the judge’s final evaluation of the flag and the number of points awarded, if a team member contributed a flag to another team’s evidence piece (and the details of this flag), which tools teams submitted using the API, how team members used tools to create tasks, and how other team members completed these tasks. We received IRB approval to analyze anonymized system log data from students who did not consent to participate in the interview portion of our study.

**Data analysis.** Finally, we conducted a deductive thematic analysis [55] of the transcripts, based on themes relevant to our research questions and the various system components. These themes largely aligned with the structure of the interview guide. After downloading and fully anonymizing the system log data, we analyzed log data for all but the first event using the pandas and numpy Python libraries. We omitted the very first event from our data analysis because students would not have had sufficient time to familiarize themselves with CuriOSINTy.
5.4 The CuriOSINTy CoCTF Platform

CuriOSINTy is designed to support experts and teams of crowd workers in efficiently and rapidly responding to misinformation on social media. Teams discover potential misinformation that makes falsifiable claims, verify or refute those claims, archive the potential misinformation and document verification process, and finally write up a short report on their findings.

**Implementation Details.** We built CuriOSINTy using the Python/Django web framework, a PostgreSQL database, and hosted it on Heroku. In the following sections, we describe how experts and a crowd would use CuriOSINTy to conduct an investigation into misinformation online.

CuriOSINTy consists of two main interfaces: the team member interface and the staff inter-
face. The staff interface allows staff to create new CTF events, create threads (topics) for teams to focus on, and judge the flags submitted by each team. The team member interface allows users to form a team for a specific event and invite other users to join their team, or be invited to join existing teams. From there, team members can participate in the event by creating evidence and flags for a specific thread to obtain points and compete in the CTF.

We now describe the team member and staff interfaces in detail using the following fictional scenario based on a real-life event [23]. Jane is an investigative journalist who works for a large news agency. Early last evening, a fire broke out in a warehouse in the Port of Beirut in Lebanon. Minutes later, two large explosions rocked the city, with the second explosion being felt in northern Israel and in Cyprus, 150 miles away. Soon after, a large red-orange cloud was visible over central Beirut. Located several thousand miles away, Jane does not have access to drone footage; and with the recentness of the event, updated satellite imagery is unavailable. However, residents of the city quickly took to social media to share what they had seen and heard. They quickly uploaded thousands of photos and videos of the event, and several rumors spread across social media as to the cause of the explosions. To investigate the event and debunk false and potentially harmful claims circulating on social media, Jane must act fast. Jane decides to request help from a local group of college students with whom she has worked in the past. Together, they decide to use CuriOSINTy to help structure their work and collectively investigate the explosions in Beirut.

5.4.1 Team Member Interface

Joining a Team. Most of the students Jane recruited have previously used CuriOSINTy, so they log in to the platform with their existing accounts and click on the Events tab to access the current CoCTF event. The students form teams and choose team leaders. The
team leaders create their teams on CuriOSINTy and share the invitation codes with their respective team members. The Leaderboard displays all of the newly formed teams and their total points so far (currently zero).

Creating Evidence and Flags. The students find that Jane has already generated several threads for them to focus on. Team leaders communicate with their members to decide on a high-level strategy for which threads to address. Some teams decide to assign specific team members to entire pieces of evidence to reduce context-switching and collaboration costs. That is, one team member will discover, verify, archive, and report on each evidence piece. Other teams instead choose to play off their team members’ individual strengths in discovery and verification. These teams assign two team members to discover new evidence and create discovery flags, while three other team members review the evidence through verification, archive, and reporting flags.

To create evidence pieces, a user clicks on the New Evidence button. CuriOSINTy prompts the user to choose a thread that the evidence relates to, and then specify a name and URL that links to the material that they discovered.

Simultaneously, on the same page, the user creates a discovery flag (i.e. evidence cannot exist without an initial discovery flag). First the user specifies the type of discovery (text, long-form text, image, audio-only, or video). Next, to promote transparency in the investigative process, the user describes how they found this evidence. Finally, the user self-evaluates the quality of their discovery flag using four criteria: originality, discovery, influence, and recency. Once the user clicks on the Add Evidence button, the new evidence piece is created. All users, irrespective of team membership, can now view this evidence and contribute to it.

Next, following the steps of the investigative process [103], users can add verification, archive, and reporting flags. New flags can be added in any order after the initial discov-
ery flag. A typical evidence piece will consist of one of each type, but to accommodate more complicated evidence pieces with multiple claims or links, users can submit multiple flags of each type. Verification, archive, and reporting flags also have multiple sub-types and different evaluation schemes. These evaluation schemes incentivize factors that we determined are important for investigations that focus on emergent crises with large volumes of misinformation online.

**Scoring Points.** Judges can approve flags in any order, except for reporting flags. Judges are instructed to approve reporting flags only after the other three types of flags have been submitted and approved. Once a flag has been approved, the team member who created the flag is awarded points, as shown in Figure ???. A team’s points are calculated as the sum of its team members’ points. If a judge rejects a flag, a user can create a new flag with additional details and context. The user can additionally communicate offline with the judge to clarify disagreements.

CuriOSINTy also seeks to incentivize collaboration by awarding points for inter-team collaboration. If a user submitted a flag to an evidence piece created by a user from a different team, points are calculated as follows. If the flag is approved, then the user who created the flag is awarded the number of points that the judge assessed the flag for. The user who created the flag is awarded points for supporting users outside of their team. Further, user who created the evidence is also awarded points to show that their initial contribution was valuable.

**Using Tools.** CuriOSINTy promotes appropriability and extensibility by providing an API that supports generic task, task-response, and reward formats, similar to Human Intelligence Tasks (HITs) on Amazon Mechanical Turk. The API allows users to develop custom tools that tie into CuriOSINTy’s event structure.
Tools exist outside of events, so any user or team can create tools for others to use. These tools must be registered and approved by a system administrator before they can be used by others. Any user can use pre-existing tools to create tasks during a specific event. Tasks can then be completed by the user’s own team members or members of other teams, as shown in Figure ???. To incentivize communication, CuriOSINTy awards users who complete a task a fixed number of points (currently 25), if that task was created by a different team. Users who create tasks are not awarded points, but may use the task results to create new flags and obtain points.

In this scenario, a previous group of students had created a tool to support object identification within an image. Upon uploading an image to the tool, it runs the image through an computer vision-based object detection algorithm. After this initial pass, a user can specify how many tasks to create. The tool automatically creates these tasks which are then visible to other users on CuriOSINTy. Here, Team Gamma’s leader used this tool to analyze an image of a nearby market. Using their private Discord channel, they asked other teams to help them analyze the image. Other teams’ members visited the task link accessible through CuriOSINTy’s task page to quickly help Team Gamma identify additional relevant objects. Team Gamma’s leader then used these responses to create a new evidence of a potential explosive device identified through everyone’s collaborative efforts.

5.4.2 Staff Interface

A user can be either a staff member or a team member for a given event, but not both. To access the staff interface, users can upgrade their account to a staff account on a per-event basis.

Structuring the Event. There are two types of staff members, judges and regular staff.
Both can create events, threads, and manage other activities within CuriOSINTy. Judges have additional permission to evaluate flags that teams submit.

Jane recruits one of her colleagues and another student to serve as judges for the event. She first creates the event, followed by several distinct threads for teams to focus their efforts on. This includes: potential causes of the explosion, injuries and lives lost, recent news about the port, historical information about the port, among others.

**Judging Flags.** Teams quickly went to work created multiple pieces of evidence and flags to support Jane’s efforts. As teams begin to quickly create flags, the judges start evaluating their work, as shown in Figure ???. This judging acts as a first-pass filter so that Jane can focus on the most relevant and accurate evidence that teams identify.

Judges are able to view the flag that a user submits, the evidence that the flag is part of, as well as the user’s self-evaluation. Taking all of this into account, a judge can then determine whether the flag is appropriate and provable. They can also provide their own evaluation to change the number of points that a team is awarded.

**Using Evidence.** CuriOSINTy provides Jane with a birds-eye view of the entire CTF, as teams and judges are working. She can see each flag and evidence that teams submit (e.g., Figs. ?? and ??), filter them based on the thread they are part of, see which ones have been judged, and which ones may contain potential misinformation. She can also view what types of tasks teams are creating from existing tools.

This birds-eye view helps Jane direct the event and steer teams to focus on more important topics or dive deeper into a specific topic. For example, Jane notices that Team Rho has discovered video footage of the explosion taken from a short distance from a fertilizer storage facility. She directs Team Rho to geolocate the footage to identify the exact location where it was taken. She also asks other teams to look for other footage in the nearby area by using
Twitter’s geotag search feature.

By the end of the day, teams have amassed over 700 unique pieces of evidence. Jane has already looked through half of them, and has created threads on Twitter debunking the social media posts that had gained the most traction. Jane also forwarded the photo of the explosive device to her editor to send to law enforcement. Now, she is now working to use all of this evidence to write a long-form article about the explosion in Beirut.

5.5 RQ9. Students’ Experiences

5.5.1 Participants’ Backgrounds, Motives, and Definitions of Success

Background and Motives

We found that many of the Computer Science students had prior exposure to cybersecurity courses and CTF events, [insert survey results]. DD-10, who had attended several cybersecurity CTFs, felt that using a CTF format “is a really cool way to introduce people to the field” of OSINT. Along these lines, KP-15 described how OSINT was one of several categories in a cybersecurity CTF they attended and that “those were the most fun ones, at least for me, to do. So that introduced me to OSINT as a whole. And when I saw that there was a new class about OSINT, I just thought ‘This is great. Like, this is perfect. I can explore this even further.’”

In contrast, we learned that most of the Political Science students had prior experience conducting OSINT investigations. For example, OT-1 said, “I’ve been pretty active in OSINT. I feel like it’s one of the most accessible forms of intelligence that I think anyone can get
We also found that students were drawn to the real-world nature of the class. For instance, DD-12 said he became interested in the class because he was “looking for more opportunities to connect my computer science degree to real world applications ...and this just seemed like a good opportunity to do so in a very topical area.”

Several students also expressed an interest in learning more about OSINT investigations to combat misinformation.

In addition, the Political Science students said that they were curious about the interdisciplinary nature of the class and wanted to develop more expertise in programming and with software tools.

**Defining Success**

All students said that a successful CTF, in theory, is one where they won the CTF by scoring the most points. In practice, however, students’ definitions of success were more varied and included finding actionable misinformation, achieving a “flow state” with their teammates, having an enjoyable experience, and learning new skills to grow as an investigator.

**Success is winning the CoCTF** All students said that one form of success was scoring the most points and winning the CTF. One student, SL-2, said she was “very competitive in pretty much anything” and described how she would feel successful if she could “find the way to most easily and effectively win within [the] bounds” of a competition’s rules.

Many students said their definition of success evolved over the course of the semester to the point where they said that success was more than just winning the CTF. For example, KG-6 said, “Before this class started, I would probably say success is winning [in terms of] point
value. But after we did a couple CTF events, point value wasn’t the best way to do it. Just because there were some groups who would just literally blow us out of the water, and get like eight times the amount of points as we did. I thought we were doing our best. So I just think, for us, success is finding misinformation, learning from it, and helping us gain knowledge.”

**Success is finding misinformation** Not all students were equally motivated by competing to win the CTF. For example, DD-10, DD-12, and OT-1 acknowledged that a successful CTF meant scoring a large number of points and being on top of the leaderboard. However, DD-10 enjoyed the CTFs the most when he managed to uncover misinformation and for DD-12 “piecing together a story or some sort of a coordinated campaign would be much more successful than just collecting a bunch of unrelated flags.”

**Success is having an enjoyable experience** Multiple students mentioned that a successful CTF involved achieving a “flow state” (cite mihaly csikszentmihalyi) where they were able to successfully find misinformation and submit flags that were, in turn, approved by judges. For instance, SL-3 said, “a successful CTF event would be if I’m banging out flags and stuff immediately and I was able to find accounts spreading misinformation right away […] you just get into the rhythm of putting out flags and getting them accepted.”

### 5.5.2 Students’ Initial Reactions to OSINT CoCTFs

We found that the computer science students were largely familiar with the concept of CTF competitions, whereas political science students were not. Many students also said that they enjoyed the competitive and gamified nature of CuriOSINTy, however, some students felt that it detracted from the quality of the investigation — a point we return to in below.
Finally, both groups of students mentioned how CuriOSINTy could serve as an introduction to the cybersecurity profession, as well as how CuriOSINTy helped them learn how to conduct investigations, which we describe in more detail below.

Many of the computer science students who had prior exposure to CTFs said CuriOSINTy’s format was “very familiar for people who major in CS” (OT-1). DD-10 added that our CTF was “a really cool thing that you put together” by combining elements of CTFs, OSINT, and combating misinformation.

DD-12 said that the overall format of CuriOSINTy — with teams competing against each other to capture flags and score the most points — was similar to other cybersecurity CTFs he had participated in before. However, DD-12 and KG-8 pointed out that CuriOSINTy had a real-world orientation (practical investigation vs. theoretical) with no limit on how many points a team could score given that flags were not predetermined; a different area of focus (misinformation on social media vs. cybersecurity vulnerabilities); and shorter duration (60 to 90 minutes vs. several hours or days). Students also pointed out another key difference: many cybersecurity CTFs introduce sequential challenges that build upon each other while in CuriOSINTy, flags were largely independent of each other:

“In the first round you might have to do an SQL injection. And then once you do that, it gives you a clue that you need to go on to the next flag. So [success] is more just a measure of how far you could get in the competition. [But] with our CTFs in class, it’s ‘get as many flags as you can,’ and the points will add up.” (DD-12)

In contrast with the computer science students, none of the political science students were familiar with CTFs. DD-11 said that participating in gamified and fast-paced investigations was “very much a culture shock for us” because “quite literally, everything is different.” Still, DD-11 said the format of the investigation was advantageous for political science students...
because “it isn’t just writing a paper on this topic that we’ve researched for a few weeks.” Instead, she said the CTF taught efficiency and teamwork with a focus on addressing a real-world problem. In addition, OT-1 said that the instructions we provided were clear and that participating in the CTF was “pretty easy once you get the hang of it.”

5.5.3 Collaboration Styles During the CoCTFs

We asked students how they worked within their teams, focusing on leadership and workflow types as well as how they communicated with each other. We found two different types of workflows that teams employed during the CTFs: (1) assembly line and (2) free for all.

Assembly line workflow

In contrast to the free for all workflow, two teams (DD, KGB) developed an informal leadership structure with an assembly line workflow. Team KP — which had largely used a free for all workflow — also set up an assembly line workflow for the final CTF. In the assembly line, each team member would focus on certain phases of the investigation, with two or more team members working on the same piece of evidence.

There were two ways teams approached the assembly line workflow. The first was based on what teams were used to working on, i.e., habit. The second approach was based on playing to individual team members’ strengths.

Teams DD (in all CTFs) and KP (in the final CTF) set up their assembly line based on habit. For the final CTF, KP-15 and his team decided to come up with a strategy to win the CTF because there was a monetary prize associated with placing in the top three teams. He went on to describe how each of his team members focused on one type of flag: “I focused on verification because that’s what I’m accustomed to. And Granada was archive, and Ganon
was archive and discovery.”

DD-12 said that he did not enjoy how each member of his team worked independently on each flag type instead of deeply collaborating with each other for each discovery, verification, archive, and reporting flag. Unlike DD-12, KP-14 enjoyed this assembly style workflow:

“We just were very fluid, moving very quickly, [...] and we kind of spread all that work out. And, you know, I would find evidence, Pranav would archive it, things like that. That was a really great example, at least from what I’d experienced of a really great CTF. I felt really good when the flags that I submitted would actually get approved.” (KP-14)

Team KGB created their assembly line workflow based on individual team members’ strengths. KG-6 said that the goal of their setup was to maximize the number of points they scored. KG-8 said that his team noticed after the first few CTFs that (Elena) was the best at discovering content, but that his team had not yet “figured out how to do archival.” To help his team, he decided to spend more time learning how to create high-value archive flags during one of the latter CTFs. At the end of that CTF, KG-8 said that his team “ended up actually doing better than we usually did just because we all sort of had dedicated tasks that we could repeat over and over again.”

Leadership We found that informal team leaders emerged over time for DD and KGB. For instance, DD-12 pointed out: “So DD-10 and I kind of emerged as the leaders of the team. It wasn’t done intentionally that way, it just worked itself out.” In addition, DD-10, 11, and 12’s team member, DD-13, wished for a mechanism within CuriOSINTy to assign team members to a specific flag type to make this assembly style workflow easier to manage.

Communication Teams that made use of assembly line workflows necessarily needed to communicate frequently with each other. We also noticed that these teams tended to focus
on a small set of topics before moving onto other topics. This may have been to ensure that each team member had sufficient context to contribute to the evidence, since no single team member was responsible for a whole evidence piece. Along these lines, DD-12, whose team focused on financial misinformation, would suggest to his teammates where to look and would also dynamically shift their focus: “I’d say [DD-10] go look at Reddit, r/WallStreet, r/stonks, [DD-11], you look at Twitter and look these accounts probably have good stuff, or these search queries might yield good results, etc. And then if we determine one person wasn’t really getting a lot, maybe we’d shift and do two people in one from one source.”

**Free for all workflow**

Different from the assembly line workflow, four teams (SS, SL, OT, KP) employed workflows that were largely “free for all” in nature, where decisions were quickly made in an ad hoc manner. Here, team members largely worked independently on a given evidence piece, submitting flags for each of the four phases. For example, SL-3 said, “it was mostly a free for all… We’d be talking about our flags and what we found. But we wouldn’t really collaborate on or delegate specific tasks.”

SS-5 also believed that the free for all workflow was better than an assembly style workflow. He said that his team initially followed an assembly style workflow, but quickly decided to switch to working largely independently. This was because it was difficult to communicate intention and context in an assembly line workflow:

“The first person’s discovery flag doesn’t communicate well to the other person trying to do the verification. ‘Hey, why do you actually think this is misinformation?’ or ‘Why do you think this needs to be verified in the first place?’ So you can’t really just split it up purely into those stages.’”
Leadership  Apart from SL, three of the four teams did not have a team leader. OT-1 said members of his team would take on tasks at the start of each CTF without a leader assigning them: “There’s not necessarily one person that gives the shots. It’s really just, ‘Hey, I’m good at doing this. And I could do this.’ … So we just state what we’re going to do.” For Team SL, SL-3 said that SL-2 became the de facto team leader because of her strong performance over the first few CTFs.

Communication  Unlike the assembly style workflow, we found that teams who worked in a free for all style did not communicate much with each other, perhaps because they worked independently. For instance, SL-2 said, “Occasionally, we would talk over Zoom to make sure that we weren’t both working on the same piece of disinformation. But that was about it.” Similarly, SS-4 said he and his teammates did not communicate with each other frequently, instead trying to maintain individual awareness of each other’s actions. One example was when SS-4 noticed that his team had a large number of approved discovery flags without associated verification flags, and told his team, “maybe we need to actually start delving into the information a little bit more in trying to verify it.”

5.5.4 CuriOSINTy Was Enjoyable to Use and Helped to Better Structure Investigations

Students enjoyed using CuriOSINTy

Students, including, (SL-2; SS-4; KP-15; KG-8), said they enjoyed using CuriOSINTy. For example, KP-15 mentioned that one of his favorite things about using CuriOSINTy is “once you get into the flow state, it feels very intuitive. It’s very easy to get momentum going.” He also valued the real-world impact CuriOSINTy could have (addressing the spread of
misinformation) compared to more theoretical CTFs, adding: “I think it’s easier to see an impact and feel more fulfilled than just debugging system code … you can feel there’s more of a sense of good and purpose.”

SS-4 explained how CTFs more generally manage to attract a large number of participants: “I find them enjoyable, they’re just kind of a fun thing to do. And if it’s an event or issue that someone is particularly passionate about, I can see why people would search these events out and take part in them in their free time.”

A few students, however, did not enjoy using CuriOSINTy as much. For instance, DD-11 wished that the CTFs were more focused on submitting high quality content instead of a large quantity of content — a point we return to below:

“I personally haven’t gotten much out of the CTFs thus far just because you’re trying to throw as many things on the wall, and maybe something will stick. … With how little time we were given, we weren’t able to find good pieces of [mis]information.”

**CuriOSINTy helped students learn OSINT**

We found that CuriOSINTy helped most students learn more about OSINT and taught them to more critically examine information online. Further, KG-8 claimed that CTFs are a good way to introduce students to OSINT because, “if you gamify something and give it to college kids, then they’ll learn how to do it while having a good time.”

We designed CuriOSINTy in the context of a semester-long class, with the goal of helping students apply the skills that they learned in class during the CTFs. With respect to this, SS-4 agreed that the “platform helped us grasp how these things can be processed in a real world setting.” In addition, both SS-4 and KP-14 said that CuriOSINTy tied in well with the course curriculum and not only taught them how prevalent misinformation was but also
helped them develop a mental model for conducting investigations. KP-15 explained how:

"It presents these phases of an investigation in a structured way. So I feel like just by participating in the CTF, what I'm taking away from it is that methodology of: you discover, then you verify, archive, and then report on. ... Since it divides it up that way, I think it helps you separate things out. [It’s] an easier way than considering all of it as one process."

Apart from teaching students how to better structure investigations, SS-4 said that CuriOSINTy taught him how to write clear justifications for his process, i.e., what he found, how he found them, and how he verified or refuted the claim. While SL-3 believed that CuriOSINTy did not necessarily improve his ability to identify or debunk misinformation, he mentioned how “instead of just reading about [investigations], we were able to perform it ourselves.”

**Improving CuriOSINTy**

Students suggested areas where CuriOSINTy’s evidence and flag structure can be improved. We discuss other improvements related to judging, the point system, and the API, in their respective sections.

In terms of structural changes, students wished that flags were organized in a way that they were easier to search, sort, filter, make sense of, and reuse. For instance, KG-8 said he felt confused by having four different flags for each evidence piece because this required him to read each flag to know what work was complete and what was remaining. While CuriOSINTy supports multiple verification flags for the same evidence, enabling students to verify different part of a claim, SL-2 also wished that verification flags could be reused across different evidence pieces. This is because certain claims made by the same person or the same claim made by different people could all be verified or refuted by one reusable verification flag.
KG-6 wanted flags to be organized in a better manner because "there’s just so much information, it’s [a lot of] scrolling" towards the end of the CTF when a team has submitted multiple flags. However, her teammates KG-8 and KG-9, admitted to not using the evidence view which might have made it easier to search through and make sense of. This suggests that the relationship between flags and evidence should be made clearer within CuriOSINTy.

In terms of procedural changes, students suggested ways to make it easier to submit and process flags within CuriOSINTy through additional tools and features. For example, DD-13 twice suggested a web browser extension to automatically capture information required to submit a discovery flag (e.g., creation date, number of likes or views, etc.) so that he could create a queue of discovery or verification flags that he could process in batches. SS-4, SL-2, DD-10, and KG-8 also wished for a feature to determine whether a piece of content had already been discovered and archived before because "it ended up being a hunt and peck situation where I’d be like, ‘vaccines are poison — Control + F.’ Is there an archive? Okay, there’s not an archive, I’m good to make a flag" (KG-8). This includes discoveries for the exact same URL, discoveries for the same content on different platforms, or discoveries for similar content.

5.6 RQ10. Students’ Performance

5.6.1 Point System Was Effective But Revealed New Tensions

We found that the point system (incentive structure) initially promoted a competitive environment. This is in line with prior work [41, 308]. About two thirds of students said that they enjoyed the competitive environment, while one third said that they did not. As the semester progressed, we modified the point values in line with our goal to make CuriOSINTy
more collaborative, while also taking into account students’ feedback on the relative balance of points assigned to different flag categories. However, given the fast-paced, largely competitive, and gamified structure of CuriOSINTy, we found two key tensions over the course of the semester. The first is a tension between competition and collaboration, and the second is a tension between quantity and quality. This section delves into students’ perceptions of the incentive structure and then describes the two tensions in detail.

Receptiveness to the Point System

In Fig. 5.4, we see that for Event 1, teams submitted 227 flags across 148 evidence pieces (average of 1.54 flags per evidence). By Event 5, teams submitted 597 flags across 228 evidence pieces (2.62 flags per evidence) — a 70.1% increase in the number of flags per evidence. Despite a 163% increase in the number of flags, the approval rate stays nearly constant with a slight increase by Event 5: from 78.5% at Event 1 to 80.4% for Event 5. This means that teams not only became more efficient at submitting more flags, without a decrease in the approval rating.

For Events 2, 3, and 4, we see that approval ratings were approximately 90%. This may have been because one team in Event 5 submitted a large volume of low-quality flags in an attempt to game the system — but they were not entirely successful in doing so. Further, we required one or two people from each team to serve as a judge. Because they were new to judging, students may have been more lenient that the experienced judges in Event 1 and 5. Still, we see a consistent increase in the number of flags submitted without a decrease in approval rating.

We found that students’ receptiveness to the points system was affected by a combination of motivational factors — both intrinsic (sense of achievement, competence, and learning)
and extrinsic (monetary incentives, grade incentives).

Some students — such as KG-9 and KP-15 — were intrinsically motivated to participate and develop their skills. For these students, the point-based leaderboard provided direct feedback to indicate how well they were performing and whether their strategies were effective.

For example, KG-9 said: “I’m a pretty competitive person. So I like being able to look at the scoreboard and be like, “Hey, we did pretty good today.’ Or sometimes we have bad days, too. And then you learn from the bad days, like, ‘Oh, maybe I should have found more misinformation.’”

<table>
<thead>
<tr>
<th>Type</th>
<th>Event 1</th>
<th>Event 2</th>
<th>Event 3</th>
<th>Event 4</th>
<th>Event 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total No. of Flags</strong></td>
<td>227</td>
<td>158</td>
<td>257</td>
<td>238</td>
<td>597</td>
</tr>
<tr>
<td><strong>No. of Approved Flags</strong></td>
<td>179</td>
<td>144</td>
<td>229</td>
<td>209</td>
<td>480</td>
</tr>
<tr>
<td><strong>No. of Rejected Flags</strong></td>
<td>48</td>
<td>14</td>
<td>28</td>
<td>29</td>
<td>117</td>
</tr>
<tr>
<td><strong>Flag Approval Rate</strong></td>
<td>78.9%</td>
<td>91.1%</td>
<td>89.1%</td>
<td>87.8%</td>
<td>80.4%</td>
</tr>
<tr>
<td><strong>No. of Verification Flags</strong></td>
<td>29</td>
<td>22</td>
<td>53</td>
<td>40</td>
<td>93</td>
</tr>
<tr>
<td><strong>No. of Verification Flags Identifying Misinformation</strong></td>
<td>N/A</td>
<td>7</td>
<td>37</td>
<td>27</td>
<td>83</td>
</tr>
<tr>
<td><strong>Pct. of Verification Flags Identifying Misinformation</strong></td>
<td>N/A</td>
<td>31.82%</td>
<td>69.81%</td>
<td>67.50%</td>
<td>89.25%</td>
</tr>
<tr>
<td><strong>Total No. of Evidence</strong></td>
<td>148</td>
<td>97</td>
<td>112</td>
<td>114</td>
<td>228</td>
</tr>
<tr>
<td><strong>Total No. of Flags Per Evidence</strong></td>
<td>1.54</td>
<td>1.63</td>
<td>2.3</td>
<td>2.09</td>
<td>2.62</td>
</tr>
</tbody>
</table>

Figure 5.4: The number of evidence pieces and flags created per event.
Although we found that most students were intrinsically motivated to develop their skills, some may not have enjoyed CuriOSINTy’s points system for four reasons. First, some students were more used to a traditional assignment grading structure (as opposed to “game style” OT-1). Second, one graduate student, DD-13, said that it might be an “an age thing” where they were not motivated by competing for its own sake.

Third, some students felt that the points did not accurately measure whether someone was good at conducting an investigation. For example, KP-15 said, “I don’t think points necessarily correlate to how good of an investigator you are. I think it also has a lot to do with your strategy and which things you focus on. I really enjoyed things that felt engaging to me, like maybe an original verification, archival, or a really good discovery.” Relatedly, DD-12 seemed to experience a sense of apathy as the semester progressed. They said that although they became more capable in conducting an investigation, they became “more lazy in terms of what I’m going to do to get the points” perhaps because it was nearing the end of the semester or because they were busy with other obligations.

This leads to the fourth reason that some students may not have been enthusiastic about the point system: they believed it could be easily gamed. For example, KP-15 was able to re-use the same fact check article for verification flags related to different discovery flags about whether the NASA moon landing was real. She said, “I had pretty much one fact check article that discussed the film technology ... and I reused that for a number of my verifications. You can say I exploited it a little bit because it didn’t matter I had one source that verified a number of claims.” SL-3 said once they figured out how to maximize the number of points they could be awarded, they applied that technique to every CTF. As we discussed previously, one of our main design goals was to rapidly collect and debunk potential misinformation. However, to DD-12, this felt like a “perverse incentive structure where it’s in your best interest to find something false, even if it’s only questionably so.”
This was in contrast to what he believed was an enjoyable experience — conducting in-depth investigations. However, DD-12 acknowledged that it’s “very hard to discover a coordinated campaign in an hour, which is why I think you have to do it the way that you do it, given the time constraints.”

Finally, some students suggested that they were motivated by extrinsic rewards such as extra credit (DD-13) or large monetary prizes (KP-15; DD-13). Although KP-15 said he became more motivated to compete once there were “gift cards at stake,” DD-13 said, “extra credit would have been more incentive than the money.”

**Changing Incentives Can Affect Desired Outcomes**

We were able to encourage students to focus more (or less) on certain aspects of the investigation by changing the point values for different flags and flag categories. For example, KP-15 focused more on verifications once they noticed that verification flags were worth more points than discovery flags:

“During the first few CTFs, I would mainly focus on discovery. And I don’t think I’d get a whole lot of points from discovery... I would do maybe five discoveries and then it would be a few hundred points. When I did one verification, it got me as many points as it took for five discoveries. Once I noted that, I was like, ‘What am I doing? I should focus on verification,’ because for the same amount of time I can get way more points.” (KP-15)

Possibly as a result of realizing that verification flags were worth more points, KP-15 and his team placed second in Event 5, where 61.8% of their points came from verification flags. In previous events, the percent of points that came from verification ranged from 0% in Event 1 to 54.9% in Event 4.

For Event 5, Team KP and OT (who placed fourth) were the only two teams who received
60% or more of their points from verification flags. On the other hand, the average percent of points from verification flags for all other teams was 33% for Event 5 (minimum = 0% and maximum = 47.3%).

Interestingly, the team that placed first in Event 5, Team SL, submitted only one verification flag (which was rejected). Instead, it appears that Team SL opted to obtain the majority of their points from discovery and archival flags (40.8% and 39.5%, respectively). Team MH, who placed third, chose a more evenly distributed strategy, obtaining 31.5% of points from verification flags, 20% from discovery flags, 16% from archival flags, and 8.8% from reporting flags.

<table>
<thead>
<tr>
<th>Event / Team</th>
<th>Event 1</th>
<th>Event 2</th>
<th>Event 3</th>
<th>Event 4</th>
<th>Event 5</th>
<th>Avg. Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank</td>
<td>Points</td>
<td>Rank</td>
<td>Points</td>
<td>Rank</td>
<td>Points</td>
</tr>
<tr>
<td>OT</td>
<td>5</td>
<td>1724</td>
<td>4</td>
<td>2753</td>
<td>4</td>
<td>7637</td>
</tr>
<tr>
<td>SL</td>
<td>2</td>
<td>3708</td>
<td>1</td>
<td>4826</td>
<td>1</td>
<td>9003</td>
</tr>
<tr>
<td>SS</td>
<td>4</td>
<td>1782</td>
<td>9</td>
<td>1097</td>
<td>8</td>
<td>4531</td>
</tr>
<tr>
<td>KG</td>
<td>6</td>
<td>1551</td>
<td>7</td>
<td>1681</td>
<td>10</td>
<td>3991</td>
</tr>
<tr>
<td>DD</td>
<td>1</td>
<td>3942</td>
<td>5</td>
<td>2635</td>
<td>11</td>
<td>915</td>
</tr>
<tr>
<td>KP</td>
<td>9</td>
<td>825</td>
<td>6</td>
<td>2597</td>
<td>6</td>
<td>5760</td>
</tr>
<tr>
<td>MH</td>
<td>8</td>
<td>1051</td>
<td>8</td>
<td>1421</td>
<td>5</td>
<td>5979</td>
</tr>
<tr>
<td>TL</td>
<td>11</td>
<td>366</td>
<td>3</td>
<td>3412</td>
<td>3</td>
<td>8166</td>
</tr>
<tr>
<td>BF</td>
<td>10</td>
<td>740</td>
<td>2</td>
<td>4215</td>
<td>7</td>
<td>5468</td>
</tr>
<tr>
<td>KF</td>
<td>3</td>
<td>2889</td>
<td>11</td>
<td>282</td>
<td>2</td>
<td>8445</td>
</tr>
<tr>
<td>JE</td>
<td>7</td>
<td>1492</td>
<td>10</td>
<td>470</td>
<td>9</td>
<td>4250</td>
</tr>
</tbody>
</table>

Figure 5.5: The rank and number of points that each team scored for each of the five events.

Initially DD-10 noted that “no one does reporting flags either because that takes way more work to really put together a report.” For Events 1 and 2, only one team submitted one reporting flag: Team BF. As a result, we increased the point values for submitting Reporting
flags for the third CTF. We found that students responded to this change by submitting a larger number of reporting flags, accounting for 3.9% of their total points, on average, for Event 3, 2% for Event 4, and 5% for Event 5.

Team KP and SL’s tactic seemed to be prioritizing certain actions that would maximize the number of points that they scored. Along these lines, KP-15 said he largely focused on completing his own team’s flags, however, once we added the collaboration incentive, “any flag that would pop up once I refreshed, I would just go for it, because we’re gonna get those extra points by doing another team’s flags.” In this way, Team KP consistently increased the number of points they obtained through collaboration, from 3.5% and 10.3% in Events 3 and 4, respectively, to 12.3% in Event 5 (see Fig 5.6).

In Fig. 5.6, for Event 5 we see a slight correlation with respect to how well teams ranked and whether or not they leveraged the collaboration and task features. For example, Team SL placed first receiving 13.6% of their points through collaboration and tasks, and Team KP placed second receiving 12.4% of points in a similar manner. However, Team MH did not obtain any points through collaboration and 1% of points through tasks, but still placed third.

Balance of points: fair but arbitrary? Multiple students — including OT-1, DD-10, DD-11, and KG-9 — said that the balance of points had improved over the course of the semester as we incorporated students’ feedback into the rubric. We valued work that was of greater strategic importance (more recent, more reach, etc.), of higher quality, or required more effort to do. For instance, DD-10 said, “Before the more laborious tasks weren’t rewarded nearly as much as they should have been. I think now they are [rewarded] more.” OT-1 added, “I think originally archiving was 100 points... To me that was way too much. And I think you guys lowered it to 50 or something now. So that makes sense, because
Figure 5.6: The rank and percent of points that each team scored for each of the five events. Percent collaboration refers to the percent of total points that a team scored through collaboration, while percent tasks refers to the percent of total points that a team scored by completing tasks.

Still, we found opposing perspectives around how many points should be assigned to certain types of flags. KG-8 said that verification and reporting flags should be worth more points because they were more crucial to the investigation. She also expected verifications to include two or three distinct sources that would help debunk the claim in a more authoritative manner, “but then to offset the extra workload, have it be more points” (KG-8). In contrast,
while OT-1 said that the balance of points for discovery, archival and reporting flags were fair, he believed that verification flags were worth too many points. This may be because OT-1, who has worked in the intelligence community, said it’s “really hard to say with 100% probability [that something has been debunked] … and so if you give someone 100 points for verifying it, I felt it was a little too much. And 600 points is the max you can get for verifying all of it” OT-1.

A second smaller group of students, who were opposed to the point system, said that the balance of points felt “arbitrary and subjective” (DD-13). For example, DD-13 pointed out how the points awarded to certain flag categories increased by a factor of three to ten depending upon the recency, quality, or effort required. KP-14 said that the fact that we were changing the point system made it hard to compare performance across different teams because “how good one team does compared to another is kind of relative” to what is valued in the point system.

Balancing Competition and Collaboration

As the semester progressed, we modified the point values in line with our goal to make CuriOSINTy more collaborative, while also taking into account students’ feedback on the relative balance of points assigned to different flag categories.

Still, we found a tension between competition and collaboration. Some students felt that collaboration was not incentivized enough for it to be worth the effort, or that it was unclear how they could collaborate with other teams. We conclude with ways to better incentivize collaboration.
Competitions promote efficiency and intra-team collaboration. In line with prior work, we found that the competitive environment encouraged students to work more efficiently — both own and with others.

While working individually, KP-14 said that CuriOSINTy “definitely enforced my thought process of, ‘How do I compete in a CTF? Where do I get my points from?’” SS-4 and SS-5 felt similarly, saying “It helped us to understand the process overall, but it encouraged us to be more efficient in how we look at the process.”

We also found that the competitive environment encouraged individuals to collaborate with others in their team, e.g., OT-1 and KG-8. KG-8 explained that one of the main benefits about a competition is that it encourages people in a team to collaborate with each other. He described a CTF where his team performed exceptionally well because of this:

“Elena was pumping out misinformation. We had other people doing Midland stuff. And I was just spam verifying, so that we could [make an] assembly line. I really liked that. It encourages you to like work as a little unit.”

Competitions are fun, but not for everyone. Many students said that they enjoyed the competitive environment, while some others either did not enjoy it or did not engage with it.

KG-9, who described himself as being highly competitive, said “for people with competitive personalities, which obviously isn’t everyone, there’s just this extra gratification from winning, which I think a lot of people really enjoy.” SL-2 also enjoyed competing in the CTFs and felt that competitions are “nice for motivating people to want to do OSINT investigations because competitions are fun.” Team KG’s average rank was 7.8, while Team SL’s average rank was 2.5, always placing in the top three for all events (see Fig. 5.5). However,
SL-2 also pointed out the tension that a competitive environment may sacrifice quantity for quality — which we describe in more detail below.

Although KP-15, felt that being able to apply his critical thinking skills was “its own reward,” he and his team decided to compete more eagerly when they noticed that they were performing well and might be able to win the CTF. KP-14 added that knowing the other participants in the CTF also “builds the competitive nature of it, versus if it was just strangers.”

In contrast to the students who enjoyed the competitive aspect of CuriOSINTy, some students said that they did not enjoy or engage with it. Instead, we found that they focused on skill development and conducting more in-depth investigations. This may be because they were not motivated by competitions, but by other factors that we previously mentioned in Section ??.

For instance, DD-13 pointed out how he was one of three graduate students in the class and that he was not a competitive person, perhaps because he was older than the undergraduate students. He added, “I wasn’t too concerned about other people beating me, but at the same time I was trying to get the experience out of it.”

SS-4 said that he noticed the point system in the initial CTFs, but as the semester progressed, he “forgot that we even had a point system.” Instead, he said that he was more focused on investigating what he thought was important:

“If I think it’s important, I’ll just do it. It doesn’t matter if it only gets five points. At some point you have to think, ‘Are you more worried about the points or are you more worried about the substance of your investigation?’”

We see this reflected in Team DD and team SS’s ranking across events. Although Team DD won Event 1, they consistently placed last for Events 3 through 5; and Team SS placed in the last three teams for Events 2 through 5 (see Fig. 5.5).
**Intra-team collaboration can be useful in competitions but is difficult to structure.** Apart from encouraging competition and intra-team collaboration, we observed that CuriOSINTy also promoted intra-team collaboration. Some students found the ability to collaborate with other teams useful, but others were unsure how to do so effectively.

We found that students appreciated the ability to gain points through collaboration for two reasons. First, this incentivized people to work together, and second, working on other teams’ flags gave them access to a wider variety of topics to investigate and methods to use.

For example, when KG-8 and KG-9 learned that Team KFC was performing well because they were using a Twitter scraping tool, they both decided to look into it and have their team use it as well, because otherwise “we’re totally going to get destroyed.” KG-9 liked this balance between competition and collaboration in CuriOSINTy where it is “half collaboration and half competition.”

OT-1 also said that there were multiple instances where he was contributing to another teams’ evidence piece to gain points — by Event 5, OT obtained 7.2% of their total points from collaboration. OT-1 also learned tactics that were beneficial for his own work: “I was looking at another team’s discovery post, because I was going to archive it. And that account had over 100,000 followers, and I was like, ‘Well, I haven’t heard of this account before.’ ... So it was helpful to find other accounts through the CTF.”

KP-14 said: “As people got better at the CTF, they became more competitive and more collaborative. But as far as adding the feature of actually being able to verify other people’s stuff, that definitely had a significant boost on collaboration.” KP-14’s team saw the collaboration feature as an “opportunity to be a shark,” and ended up receiving many of their points in the final CTF this way. KP-14’s team did so by searching and filtering on the evidence tab for specific teams’ evidence pieces. Then, they would inspect the status, such
as if there was an archival, verification, or reporting flag present. If there was a missing flag, someone from KP-14’s team would attempt to create one themselves. For Event 4, Team KP collaborated the most out of all teams — they obtained 373% more points through collaboration than the second-most collaborative team. For Event 5, Team KP placed second and was the second-most collaborative team at 12.41% versus 13.63% for Team SL who placed first. See Figure 5.6 for more details.

Although OT, KGB, and KP took advantage of these collaboration features during the CTFs, OT-1 said that collaboration was not common in typical OSINT investigations he had participated in. Further, SL-2 was “not really sure how to collaborate through the CuriOSINTy platform;” and DD-11 decided not to collaborate because “there wasn’t really a lot of time to understand what the other teams were working on and what their objectives are, both from the flag and evidence perspective, and then from the tool perspective.”

Finally, KP-15 pointed out a tension between competition and collaboration, noting that one may come at the cost of another: “There’s a fine line between competition and collaboration with some things, because if I verify some other group’s piece of evidence, we’re technically collaborating, but perhaps their strategy is to have them focus on their own pieces of evidence. So maybe I’m disrupting the[ir] strategy as well.”

**Further promote collaboration with incentives and norms** We find that CuriOSINTy can further promote collaboration in two ways: first, by providing a greater incentive in terms of points, and second, by changing social norms in the CTF.

SS-4 thought the idea of collaborating across competing teams was a “really cool idea and it definitely works [but] if all the teams are on board with doing that, it’ll go a lot smoother.” Further, SL-2 and DD-11 both said that collaborating with teams was not incentivized enough in the most recent version of CuriOSINTy because it was not “worth as much points
as the time that went into doing them properly, [versus] making flags yourself” (SL-2). Even though SL-2 said collaboration was not sufficiently incentivized, they still obtained 13.63% of their points through collaboration and tasks.

Apart from increasing the point incentive for inter-team collaboration, we find that changing social norms around collaboration may also help increase inter-team collaboration.

For some students, collaborating in this way required more effort and information from the individual trying to help another team. Although SS-4 felt that CuriOSINTy had improved over the semester and that “it’s really easy to see what your team is doing, what other teams are doing”, he was apprehensive about collaborating. This is because he felt that he might be taking away from another team’s points if he contributed a flag to their evidence piece. His teammate, SS-5, also described the context switching cost that one needs to overcome when working on another team’s evidence, leading him to solely work on his own teams’ evidence:

“I haven’t helped another team’s flag yet. Even with the points I’m just not inclined to, because at the end of the day, I have to read through theirs, understand what it is. That’s almost like stopping in my tracks what I’m doing already, and then trying to understand what they’re doing.”

OT-1 suggested a change in mindset for all participants at the CTF might help improve collaboration. He suggested that participants should view the CTF as teams collectively working towards a common goal, “instead of separate teams working on separate things, trying to win.”
Balancing Breadth and Depth

We mentioned previously how one of the goals for CuriOSINTy was to rapidly identify and debunk misinformation. This required casting a wide net, both in terms of covering a wide variety of topics but also collecting a large quantity of content.

From Event 2, we began tracking if a verification flag identified an instance of misinformation — that is, whether the student created a discovery flag that refuted the original claim that made in the content that was found. For example, one team member from Team JE found a video posted online that the poster claimed was an instance of voter fraud. However, this team member was able to debunk this video by finding an alternative source that had investigated the same video. In this case, that verification flag identified an instance of misinformation.

From Event 2 to Event 5, the percentage of approved verification flags with instances of misinformation rose from 31.8% to 89.25% (see Fig. 5.4), though it is to be noted that Event 5 focused on topics such as 9/11 and chemtrail conspiracy theories, which are more likely to be misinformation.

However, we found a tension between our design goals that emphasized breadth versus students’ desire to conduct in-depth investigations.

Defining a good investigation: breadth vs. depth. We previously described how the competitive environment promoted efficiency and breadth. However, OT-1, who had prior experience with OSINT investigations, believed that a good player needed to balance discovering a large quantity of content while making sure that it is also of high quality through careful research:

“I think being able to discover information quickly and timely within the CTF timeframe is
a good skill to have. This is where the problem with quantity over quality comes in, because a lot of times it’s easy to discover poor quality tweets made by bots, you know, it’s obvious, but then the real exploitable information is a little bit harder to find. So if you’re able to do that, and then archive it in the background, and then at the same time, you’re thinking about, ‘Okay, so how do I verify it — verify the account and verify the information?’ That’s a well-rounded player.” (OT-1)

Five other students also said that they preferred an environment that incentivized conducting in-depth investigations. For example, KG-8 said her team ended up finding “a lot of small pieces of misinformation [because] a lot of the bigger fish had sort of been fried already” — which, to her, did not feel rewarding from an investigation standpoint. DD-10 also pointed out how his team had not spent much time looking into any single piece of evidence, but rather “trying to just cast a huge net.” He went on to describe what he perceived as the tension between breadth and depth within a competition:

“Stuff like that, that takes a lot of time, and quantitatively it’s not very much actual result at all, is actually the most rewarding. You really have to be clever about this one image instead of finding all of them. That’s something I’m worried about the whole platform in general, because the competitive incentive is useful for making people want to work harder, but I think it strongly discourages people from really looking into things.” (DD-10)

**Rewarding and assessing depth.** Earlier, we found that students were receptive to changes in the point structure. In turn, rewarding in-depth work with more points may satisfy students’ desires for depth. For example, KP-15 said that they valued original verifications where they “extrapolated on some knowledge from a few different sources ... as opposed to just using a fact check article,” but also added that there was a big “point boost” for original verifications, which he described as a “win-win” scenario.
Still, multiple students pointed out that it was hard to determine how much effort someone put into an in-depth investigation. This included SL-2, who said:

“If you do something that you put a lot of effort into, that takes half an hour, at surface glance is going to be indistinguishable from something that you spent five hours going and researching and figuring things out. ... You’re not going to know from a judge’s point of view how much time one person spent compared to the other, and they’re graded exactly the same.” (SL-2)

Two students (SL-3, SS-5) proposed an alternative to rewarding effort: rewarding completion of flags. We designed the four flag categories to encompass all stages of an investigation, thus ensuring sufficient depth for a given piece of content. Even though we designed the point system to incentivize students to complete all four flag categories, we found that of evidence pieces were not completed. To fix this issue, SS-5 proposed additional points for completing all flag categories, or in other words, “tipping the scales towards actually having to go through the process instead of just having [inaudible] because I still feel like I rack up discovery flags on the scoreboard.”

**Longer duration can promote depth.** We designed the CTFs within CuriOSINTy to be conducted within the span of a 3-hour class session. This was both due to practical considerations — students may not be concurrently available outside of class — but also because identifying and debunking misinformation quickly was an important design consideration.

However, five students felt that the short duration made it difficult to conduct in-depth investigations. For instance, DD-10 mentioned how “it’s very hard to discover a coordinated campaign in an hour.” While SL-2 enjoyed the fast-paced competitive environment, she believed “doing it fast ... is counter to properly doing an investigation.”
Apart from better rewarding depth by increasing point values, we find that increasing the duration of the events can further enable depth. This can be done in two ways. First, DD-13 suggested allowing students to research the topics for a given CTF prior to taking part in it. Second, (SS-4, SS-5, DD-10, DD-13) suggested having the CTFs be longer in duration – ranging from one to several days, either synchronously or asynchronously. DD-10 described what this might look like:

“It’s almost like a hackathon thing where it’s way more than an hour. You get together with some friends at the conference center with everybody, you’re like, ‘Okay, you’ve got this weekend to look into this thing.’ And, you know, there’s a competition that rewards the people who do really good work [where] its just judges going over the work. But you’re not getting points in an automated way. You come up with a report of what you found. And you submit that, and then people can really say like, ‘Did you find something important?’” (DD-10)

5.6.2 Judging Improved Quality But May Be Difficult to Scale

Self-assessment rubric and judging improved quality of flags

We found three aspects of the judging process that helped improve the quality of flags which students submitted — both prior to submission and for later submissions.

First, students said that the self-assessment rubric helped improve the quality of their flags prior to submission. For example, KG-7 noted that the self-assessment rubric helped them better understand the requirements for a high-quality flag to “make sure I can get the most points and I can justify the points” (KG-7).

Second, students said that the judges’ feedback encouraged them to submit high quality and
more detailed flags that were more likely to be approved — so that they would score more points. For SS-5, many of his discovery flags were rejected by judges early on because he did not sufficiently describe why what he had found was potential misinformation. However after he resubmitted the same flags with more details, they were approved. Along these lines, KP-15 said that judging “incentivizes you to really give your all into each flag and provide as much detail as possible to ensure the highest possible chance of it going through and being approved.”

Third, we found that allowing students to be judges provided a valuable perspective that also increased the quality of their flags. OT-1 said that being a judge taught him what to look for when evaluating if flags contained misinformation — a perspective that he then leveraged as a participant submitting flags. Two students (SL-3, KG-8) also said that being a judge enabled them to learn about new techniques that other students were using. For instance, KG-8 said as a judge, she noticed how one team won a CTF because they created a script that prioritized accounts with a large number of Twitter followers (our rubric assigns more points when the potential audience for a piece of misinformation is large):

“I really liked judging actually. It was cool to see other people’s processes. I learned a lot about the team that I judged, I believe it was KFC. They had a Python script that would collect potential misinformation with a certain number of followers and then they submitted a ton of discovery flags that were super valuable. So they actually ended up winning because this account has 10,000 followers.”

Judging misinformation may be subjective

Despite our use of a rubric to make judging more fair and objective, three students (DD-13, KP-14, OT-1) pointed out that judging whether something is misinformation or not may be
a subjective task. DD-13 describes the challenges in doing so:

“It goes back to a point of subjectivity where one judge might say ‘I don’t think this misinformation.’ ... Because some people might find something to be fake news. And some people might just say it’s up to interpretation.” (DD-13)

KP-14 also worried that the evidentiary standard required for verifying or refuting a claim also differed between judges. He said that he would sometimes reject flags because he did not think that a participant had submitted enough evidence, but was not sure if another judge would have rejected the flag for the same reason, perhaps because they were “a little more timid to reject it.”

To overcome these challenges, KP-14 and KP-15 suggested introducing more anonymity into the judging process while also rotating judges between teams. This way, students would not adapt their submission style to a specific judge, but would instead be encouraged to submit high-quality flags because “you have no set standard in mind” (KP-15), while ensuring that judges’ potential biases are evenly distributed across teams. Alternatively, DD-13 suggested a tiered judging system where one judge would go over another judges’ evaluation. However, DD-13 caveated his recommendation, noting that this was “not super feasible” since it would require judges to do more work, and may lead to confirmation bias where one judge may trust another judges’ evaluation without doing their own independent evaluation.

**Judging may be difficult to scale**

The research team, including three to six research assistants, acted as judges for all of the CTFs. During the first two CTFs, we were able to maintain our rate of evaluation with students’ rate of submission. However, students soon became quicker and more adept at submitting flags, leading to two occasions where the judging team could not finish evaluating
all submitted flags by the end of the class. Six students (SL-2, SS-5, KP-15, DD-10, DD-12, DD-13) noted how the “big bottleneck during the CTF became judging” (DD-12) because there were nearly four of five times as many students as judges during the early CTFs. DD-10 believed that this may cause the judges to become overwhelmed and, in turn, becoming more lenient in their evaluations:

“But in this platform, I think you’re encouraged to just, you know, ‘whatever, close enough,’ because the judges don’t have time to really, really look into what you figured out. I haven’t spoken to any of them, but I’m very doubtful that they actually did their own analysis to corroborate mine. I’m sure they were just like ‘Yeah, you know, seems good.’ And so because of that, I feel like you’re incentivized to just toss together something ... It is very tempting, and it is very easy to cheat.” (DD-10).

To decrease judging bottlenecks and to provide students with practice being a judge, we asked students from each team to sign up as a judge for Events 3 and 4. With the students’ help, we found that we were better able to keep up with the rate of submission. In Event 3 there were 25 judges who evaluated 292 flags, and in Event 4 there were 24 judges who evaluated 238 flags. In both events, judges took on average of about 10.6 minutes between when a flag was submitted and when it was judged.

For Event 5, we wanted students to fully participate in the CTF, and recruited additional research assistants. Here, 11 judges took 20.6 minutes on average to judge 597 flags. In other words, even though the amount of work quadrupled (half the number of judges and twice the number of flags), judges only took twice as much time to evaluate flags.

In addition, KP-15 said he had a “bad habit” of constantly checking whether his flags were approved or rejected but that during the last two CTFs, the rate of evaluation was “so fast that I didn’t even need to do that anymore. I could assume that it would be approved or
Students also perceived that the self-assessment rubric enabled them as judges to evaluate flags more quickly because the rubric “streamlines the process – ‘Okay well, I need to follow this link, I need to check all these things’ – it makes it easier to approve it” (KG-9). SS-5 also noted how the rubric encouraged students to more accurately rate themselves — which meant, as a judge, he did not have to change the rating by much except in a small number of cases. Still, DD-13 did not think he was a good judge because he felt that he spent too much time carefully evaluating flags.

5.6.3 API Enabled Appropriation

We designed CuriOSINTY’s API to allow participants to easily crowdsource tasks to other participants. Most computer science students said that using the API was easy, however some pointed out that APIs can be both freeing and constraining. Here we also describe the types of tools that students created using the API.

Experiences using CuriOSINTY’s API

For most students with a programming background, using the API was “very straightforward” (KP-14) — perhaps because we provided code examples and documentation on how to use it.

CuriOSINTY was largely designed to be used by students irrespective of their programming abilities. However, effectively leveraging its API to build add-on tools was difficult for students without a programming background:

“I think across most political science students, we weren’t really able to contribute that much
to the development. ... [It] was kind of hard to pick up learning Django and Bootstrap, and at the same time trying to create a tool and learning what’s feasible or not.” (DD-13)

Three students also noted that APIs like CuriOSINTy’s are simultaneously freeing and constraining. For example, DD-12 noted how the API enabled students to easily create their own crowdsourcing tools but may limit the variety of tools that can be created:

“I thought it was definitely cool how it handled the task management for you. ... So [the tools] are not storing anything, they can just say ‘Okay, here’s my data.’ I think that’s really useful. ... But I definitely think more flexibility would be good because it would support more use cases, right? The narrower and more restrictive the API is, the more it drives you towards one specific type of software project, because you have to build something that fits within that framework. I say that understanding that it creates a lot more work for you guys, because you know, the sets of requirements then cascades from your perspective.”

Although this was not part of the project requirements, KP-14’s team created a tool that duplicated many of the API’s in-built storage features because “if one day CuriOSINTy went down, and it no longer existed, we had our own backend. It was a NodeJS backend with a MongoDB database, basically using Axios on the NodeJS server to make the request to the CuriOSINTy API. And it was designed in a way where we were storing all the information ourselves.”

**Types of tools enabled by the API**

Students created a variety of tools using CuriOSINTY’s API supporting three of the four steps in the OSINT cycle (other than reporting). This ranged from two tools to support discovery, two to support archiving, four to support verification, and one tool that replicated many of CuriOSINTy’s features in an encrypted environment. Of these eight tools, all
but two (HAL9001 and Investweetgator) leveraged crowdsourcing. No team built a tool to support reporting.

**Tools for discovery**  Students built TwitCat and CrowdSieve to support crowdsourced content discovery on Twitter and Reddit, respectively. TwitCat scraped tweets based on a user’s query and created tasks for crowd workers to annotate tweets based on whether it may be potential misinformation or it. TwitCat then aggregated, ranked, and displayed the crowd’s feedback to the user “so that they might more quickly identify pieces of potential disinformation to explore further, alongside the [crowd workers’] rationale for their decision.” KP-14, who used several other teams’ tools during the final CTF, particularly admired TwitCat: “I used a few... But there is one tool that integrated Twitter, like it embedded tweets in the page. And I thought that was really cool. It inspired me like that’s definitely something that we should have added to our platform to actually embed tweets.”

While TwitCat automatically scraped tweets, CrowdSieve relied upon crowd workers to search through subreddits manually. OT-1 recognized the importance of using CuriOSINTy’s crowdsourcing feature that several of its tools supported: “Let’s say you have three teams working on the Ukraine topic ... You’re not going to be able to find all the information [by yourself] so literally having the CTF is forcing people to work with the crowd and utilizing all their manpower.”

DD-13, whose team built CrowdSieve, said he would like to incorporate automation into the tool: “Especially for things that might feel a little bit, maybe menial, or time consuming. … Some form of human machine teaming. If you’re familiar with that, like investigators and intelligent software working in tandem.” SS-4 also suggested tools that would make it easier to discover content through keyword searches and by automatically scraping websites.
Tools for archiving  KeyHive and HAL9001 were archival tools built by two different teams. KeyHive asks users to input an image URL and automatically archives it, while creating tasks on CuriOSINTy where crowd workers are asked to annotate each image. Key-Hive then enables users to compare and search for images based on the crowd’s annotations. HAL9001 did not leverage crowdsourcing, but enabled users to query CuriOSINTy for tasks and flags through a Discord chat bot. It also included a feature where users can enter a URL and automatically archive it using an online archival service. HAL9001 also checks if the URL has already been submitted before — an additional feature that CuriOSINTY does not natively include. SL-2 also described an additional feature of her team’s tool: “it gets flags that have been discovered but haven’t been verified if other people want to then go and verify it.”

KP-14 and DD-12 experienced difficulty using existing archival tools to archive video content, perhaps because “there’s a certain cost in actually storing the video” (KP-14) and wanted to create tools to better support this. During the CTFs, KP-14 resorted to submitting screenshots of videos instead.

Tools for verification  Of the four tools that teams built to support verification, two were focused on supporting verification on Twitter: using crowds to annotate popular hashtags and to annotate tweets from a given Twitter user. One tool, Photo Verification App, allowed a user to post an image that could be verified by other crowd workers. Credibility Checker included similar features but with a broader scope, supporting verifications of any type of content. SS-5, whose team built Credibility Checker, decided to focus on verification because “that’s the most time consuming step because you have to do a little bit more research” thus allowing the “investigator to focus on finding new flags or archiving previous discoveries while their volunteers determine the credibility of the flag sent to them.”
5.7 Discussion

In this work, we engaged in a four month-long Research through Design process to develop the CuriOSINTy platform. We find that a RtD process helped us to improve and validate the design of the platform, moving closer towards our ideal preferred state described in our three design goals.

5.7.1 Designing with Appropriation in Mind

Recall that we instantiated our three design goals using five of Dix’s heuristics for software appropriation. Namely, (1) expose intentions, (2) provide visibility, (3) encourage sharing, (4) support not control, and (5) encourage pluggability and configuration. However, Dix actually provides two other heuristics which we discuss below: (6) learn from appropriation and (7) allow for interpretation.

First, Dix suggests that designers should learn from appropriation. By observing how a system has been used and appropriated, we can then redesign the system to better support users. Thus, in section (C5ii) of the discussion, we revisit our design goals and discuss areas where the platform worked well and areas for improvement. We also highlight tensions that we found in our design.

Second, Dix recommends that designers allow for interpretation of the system. In other words, not every part of the system should have a fixed meaning, but rather it should include elements where users can add their own meanings. In turn, this allows users to appropriate the system for other applications. This intentional “absence of meaning” is useful for making the system more generalizable for other applications. We discuss how CuriOSINTy can be easily adapted for other investigative settings. This includes addressing rumors during
5.7. Discussion

Elections, tracking emergent and mass-participation events (e.g., riots or protests), among others. (More details in the dissertation draft.)

5.7.2 Learning From Appropriation: Evaluating CuriOSINTy Against Our Design Goals

Our mixed-methods evaluation allowed us to assess CuriOSINTy against our design goals. We now revisit the three goals and discuss the findings from our evaluation.

Goal 1: Enable an Efficient and Rapid Response to Misinformation on Social Media

We find that Curiosity enabled a rapid and efficient response to misinformation on social media by merging the complementary benefits of competition and collaboration.

_CuriOSINTy reduced inefficiencies compared to current CTF competitions in two ways_. First, we showed that a crowd of 46 students could be motivated to quickly identify and debunk hundreds of pieces of potential misinformation in sessions as short as sixty minutes by creating a competitive environment with a points-based incentive structure. In turn, CuriOSINTy demonstrated that CTFs that have traditionally been used as theoretical games (with “wasted” effort) can also be leveraged to support real-world collective action. Second, we mitigated the limitations of competitions, including duplication of effort and siloed information by allowing competing teams to view and build upon each others’ evidence and flags.

Curiosity reduced inefficiencies compared to traditional crowdsourcing approaches in three ways. In many traditional crowdsourcing systems that involve monetary compensation,
designers implement “attention checks” to make sure crowd workers are making an honest effort to complete the work, and also aggregate crowd feedback on the same microtask to mitigate the effects of low-quality work or biases. In CuriOSINTy, however, we were able to ensure high-quality work through a trusted crowd, a point-based incentive system, self-assessment rubrics, and judges. First, because we knew the students and developed a working relationship with them over the semester, we were able to trust them to submit higher-quality work compared to anonymous crowd workers. Second, we could easily delineate and communicate low- and high work to students through the point system and rubric. Third, the self-assessment rubrics and judging mechanism served as a way to improve students’ work in the short- and long-term.

However, we also find areas where greater structure within teams could lead to greater efficiency gains. For instance, some teams organically devised assembly line workflows and team leaders emerged over time. We observed that these teams frequently placed high in the leaderboard. Another set of teams employed “free for all” workflows with minimal collaboration among team members and no explicit team leader. We found that the “free for all” teams did not perform as well.

Our findings suggest that both types of teams may benefit from more explicit structure and roles, such as delineating the responsibilities for each team leader and assigning roles to each team member. For example, the leader could mitigate unwanted redundancy by assigning team members to work on a specific topic or social media platform. To prevent judges from being overwhelmed by work, the leader could also conduct a preliminary evaluation of their flags before forwarding it to the judge. To further increase efficiency, individuals could be assigned or encouraged to focus on tasks that they preferred or excelled at, such as content discovery versus verification.
Goal 2: Give the Crowd (More) Agency

In traditional crowdsourcing systems, complex tasks are divided into microtasks that crowd workers complete independently, with little to no interaction with each other or agency in how to complete these tasks. However, CuriOSINTy builds on a growing body of literature that shows that crowds can perform more complex tasks — provided that they are sufficiently motivated, as well as given adequate scaffolding, training, and agency.

In the previous section, we found that CuriOSINTy structured students’ work so that they could more easily perform complex investigative tasks ranging from discovery and archival to verification and reporting. CuriOSINTy was flexible enough that students were able to investigate a range of topics, from COVID-19 and election misinformation to human rights violations and stock market rumors.

*Providing more agency can lead to a virtuous cycle.* We also found that CuriOSINTy helped students learn to more critically examine information online and develop a mental model for conducting investigations. This proved to be a virtuous cycle: between the first and fifth event, students submitted 65% more evidence pieces and 163% more flags, with only a 1.3% reduction in flag approvals from judges. In addition, students said that they enjoyed using CuriOSINTy — possibly motivating them to continue participating in the events.

*Allow participants to self-select based on what motivates them.* Still, we found that not all students were motivated by competition, perhaps because we did not allow students to choose between a predominantly competitive or collaborative setting, as would be typical outside of the classroom. In future work, researchers should make participants aware of the predominantly competitive nature of the event beforehand. This way, those who are motivated more by competition can self-select into the event, while still benefiting from collaborative features such as information sharing and intra-team contributions.
Goal 3: Support a Range of Complex Tasks and Tools

We found that teams developed a wide variety of tools using CuriOSINTy’s API. This included tools to support discovery, archival and verification. Most computer science students also said that the API was easy to use and were able to develop prototype tools within a few weeks.

Further, as some students pointed out, APIs can be both freeing and constraining. For example, while the computer science students could use the API, none of the political science students had a programming background, and as a result, could not use the API. In turn, this constrained how they could appropriate CuriOSINTy. We alleviated this constraint through teams that were composed of both political and computer science students to provide domain and technical expertise, respectively. Alternatively, designers could explore programmable user interfaces or natural language commands to make it easier to appropriate a system.

Another challenge common to all software projects is long-term maintainability and support. To address this, we designed CuriOSINTy with appropriation in mind to maximize the ways in which it can be used in the future. We also plan to open source the code for the CuriOSINTy platform so that others may continue to improve it.

Yet, it would be remiss to not mention the ethical concerns with making powerful tools openly available. Perhaps bad actors might use CuriOSINTy to track and harass political opponents, or authoritarian governments might use it to identify and imprison dissidents. We believe that bad actors will (and already do) engage in this type of behavior irrespective of the availability of such tools. However, by demonstrating how CuriOSINTy can be used for prosocial behavior, we believe it can help empower marginalized communities.
5.8 Summary

Using a Research through Design process, we developed CuriOSINTy, a platform for collaborative capture the flag competitions (CoCTFs) that enables a trained crowd to combat misinformation on social media. Our mixed-methods evaluation showed that CuriOSINTy leverages the complementary strengths of competition and collaboration, allowing a crowd to quickly and efficiently identify and debunk misinformation. We also highlighted tensions between competition versus collaboration and the challenges of software appropriation.
Chapter 6

A Framework for Expert-Led Crowdsourcing in High-Stakes Investigations

6.1 Motivation

Traditionally, there have been two types of investigations involving professionals and crowds [315]: top-down and bottom-up. Top-down investigations involve a one-way flow of information, from the crowd to the investigator. Here professionals control all major elements of the investigation, from selecting the targets to deciding which lines of inquiry should be followed up, while the crowd’s contribution is limited to providing data. In bottom-up investigations, an emergent crowd of novices take an active role in conducting their own investigations and rarely involve — or are sanctioned by — professionals [184]. Here, crowds coordinate their efforts to share, collect, and analyze information with others [293, 343] and even administer social and procedural justice [66, 151].

I presented three studies in this dissertation that, together, demonstrate the possibility of a novel, third approach to conducting investigations in high-stakes domains with professionals and a novice crowd. I called this approach expert-led crowdsourcing (ELC). ELC blends
top-down training and guidance by professionals with bottom-up participation by a crowd of novices. In the three prior studies, we found that ELC extends the literature’s current conceptualization of crowdsourcing, beyond traditional microtask crowdsourcing (top-down) and emergent collective behavior (bottom-up) [292, 322].

In early crowdsourcing approaches, complex work is often decomposed into microtasks for crowd workers to complete independently, with little to no interaction or awareness of each other. The resultant work is then aggregated and provided back to the requester. Here, ELC complicates the typical relationship between “requester” and “crowd” [177, 292] by building on the trajectory set by more recent work enabling greater collaboration between the requester and crowd [53, 188, 223], allowing expert–crowd collaboration to happen in real-time [45, 187], and providing more agency and training to novice crowd workers [149, 335].

In this dissertation, I introduced and expanded the scope of expert-led crowdsourcing. First, in CrowdSolve, I studied the sociotechnical and ethical challenges posed by ELC investigations in a high-stakes setting: a law enforcement investigation of two murder cases. Second, to address the expert–crowd coordination challenges that I found in CrowdSolve, I introduced the concept of shared representations within ELC. I then instantiated shared representations in GroundTruth, an ELC system to enable efficient and effective coordination between experts and crowds for one complex task: image geolocation. Third, I studied how to support greater flexibility in and variety of ELC tasks through CuriOSINTy. I also explored a different way to motivate the crowd — collaborative competition — versus altruism in CrowdSolve and monetary payment in GroundTruth. Collectively, these studies describe beneficial yet complex collaborations between experts and crowds.

Having studied ELC investigations in these three different settings, what is missing is a framework to show how expert-led crowdsourced investigations work in high-stakes investigations. Thus in this work, I explore the following questions. How do expert-led crowd-
Chapter 6. A Framework for Expert-Led Crowdsourcing in High-Stakes Investigations

sourced investigations work? What types of professional investigations can a novice crowd help with and how? Who are the stakeholders that need to be brought in, what motivates them, and how can they contribute? What is the nature of the collaboration between experts and crowds, and what factors determine the success (or failure) of an investigation? Finally, what sociotechnical infrastructure is needed to facilitate this collaboration?

To answer these questions, I first highlight the applicability of existing crowdsourcing frameworks [106, 209, 221, 292] in the context of expert-crowd collaboration in high-stakes settings. Second, I summarize the findings and implications from my three prior studies. Third, I present a conceptual framework for expert-led crowdsourced investigations that compares and contrasts these three studies, and contend that expert-led crowdsourcing allows experts and crowds to do more than either could alone. In the next chapter, I discuss how to apply and evaluate the expert-led crowdsourcing framework in other high-stakes settings, the ethical implications to consider, and avenues for future work.

6.2 Case Studies

In the three prior chapters, I presented three case studies involving expert-led crowdsourced investigations across two domains: journalism and law enforcement. Each case study relied on long-term engagement with the specific context — spanning several months of study in each case. Although much of the empirical findings have been covered in prior work, the case studies provided valuable context for the expert-led crowdsourcing framework I present here. In addition, my understanding of these have studies evolved from continued reflection and navigating between the three different contexts, as well as deeper engagement with related work. Although all three studies focus on expert-led crowdsourced investigations in high-stakes settings, they explore vastly different contexts, stakeholders, and sociotechnical
infrastructure. Together, these similarities and differences not only allow us to highlight the range of settings in which expert-led crowdsourcing can exist, but also the various sociotechnical configurations that must be brought into alignment for an expert-led crowdsourced investigation to be successful.

The first study, CrowdSolve, involved a real-world law enforcement investigation of two murder cases with six experts and 250 crowd workers. In the second study, GroundTruth, I explored how to design a crowdsourcing system to support investigators with one investigative task: image geolocation. This was followed by a lab-based evaluation with eleven experts and 567 crowd workers on Mechanical Turk. Finally the third study, CuriOSINTy, involved a three month-long iterative design and deployment process of a collaborative capture the flag (CoCTF) platform that enabled up to four experts and 46 trained crowd workers to identify and debunk misinformation online.

### 6.2.1 Defining Expert-Led Crowdsourcing

A range of terms are used to refer to crowds — a large number of people working towards a common goal, e.g., crowdsourcing, crowd work, collective intelligence, human computation, peer production, and citizen science [220, 261, 292, 328]. Modifying one of Howe’s two definitions for crowdsourcing [155] and incorporating elements from Kittur et al.’s work [177, 314], I define *crowdsourcing* as: the act of taking a job traditionally performed by an agent and designating parts or all of it to a large group of people through a sociotechnical system composed of individuals, organizations, technologies, relationships, workflows, and actions.

In the previous three chapters, I studied, implemented, and expanded upon *expert-led crowdsourcing*. The four components of expert-led crowdsourcing are: 1) one or more professional
requesters called experts, 2) a large crowd of non-professionals (from 10 to over 250 individuals), 3) complex sensemaking work that currently cannot be automated, and 4) a sociotechnical system (e.g., people, workflows, software, and hardware) that enable experts and a crowd to complete this work in a synchronous and real-time manner.

Here, the “requester” or expert — the person requesting work to be completed — possesses professional knowledge and expertise in an investigative domain (e.g., journalism, law enforcement, human rights advocacy). Similarly, the “crowd” — the people completing much of the requested work — are non-professionals and members of the general public [184]. Experts and crowds work together in a largely synchronous and real-time manner, but may be geographically distributed online or they may be physically co-located. Both experts and crowds are motivated to participate for a variety of reasons, ranging from monetary compensation and gamification to altruism and learning.

The “work” I focus on here is largely prosocial in nature. Although expert-led crowdsourcing could be applied to more antisocial crowdsourcing applications such as crowd turfing [86] or searching for and harassing individuals [66, 194].

This work to be completed is largely specified by the expert, however the crowd possesses agency in determining how or when to complete this work. The expert also provides training and guidance to the crowd so that they can complete work that was traditionally only the purview of experts. Finally, in all cases, there exists a sociotechnical system that facilitates this work. This includes individuals who engage in articulation work [300] to set up the system or customize existing systems, bring together other experts and members of the crowd, and assist with designing workflows and tasks to be completed. The technical portion of the system includes software and hardware that mediate interactions between experts and crowds. This includes an interface for experts to specify the work, assist with generating and assigning tasks, providing an interface for crowds to complete this work, aggregate and
6.3 Designing the Expert-Led Crowdsourcing Framework

Given the popularity and applicability of crowdsourcing to a large number of domains, researchers have developed numerous frameworks for characterizing crowdsourcing initiatives, largely focused on who participates, why they participate, how they participate, and what they do [221]. This includes describing the types of stakeholders present [209]; their motivations [220, 292]; the synchronicity, co-locatedness, and agency involved [195, 292]; and the complexity of the work assigned to crowds [174] as well as workflows for generating tasks, assigning tasks, and aggregating the resultant work [175].

Here, I discuss prior frameworks for crowdsourcing complex work — work that requires human creativity and innovation and is difficult to automate through software or hardware and consider how it can be incorporated into a new framework to support expert-led crowdsourced investigations.

First I discuss Malone et al.’s genome for collective intelligence that provides an overarching structure for the expert-led crowdsourcing framework. Next, I describe how the ELC framework incorporates elements from other frameworks. This includes frameworks for emergent collective behavior, microtask crowdsourcing frameworks, and macrotask crowdsourcing frameworks.
6.3.1 Malone et al.’s Genome for Collective Intelligence

Malone et al. [221] identify building blocks of crowdsourcing systems that leverage collective intelligence (CI). While there may be a small number of crowdsourcing systems that do not involve “groups of individuals doing things collectively that seem intelligent” [221], Malone et al.’s definition of collective intelligence is similar to Quinn and Bederson’s Venn Diagram [261] showing collective intelligence as an umbrella category for crowdsourcing. The examples provided in Malone et al.’s work also overlap significantly with those provided by Howe [155].

Malone et al. identify four building blocks — or genes — that exist in a variety of permutations and combinations within crowdsourcing systems: who, why, what, and how. More specifically, Starbird breaks down Malone’s framework into the following questions [292]: 1) Who does the work? Are they in a closely coupled organization or a loosely defined crowd? 2) Why do they do the work? Is it for ”money, love, or glory?” [291] 3) What are they doing? Are they generating new ideas or making decisions? 4) How is this work being accomplished? Is it independently through collections, or together through contents or collaborations?

This genome framework is useful for classifying different crowdsourcing systems because many of these systems have varied configurations that manifest as a combination of genes. Still, this framework is not without its limitations. I discuss three of the most relevant limitations.

Limitations of Malone et al.’s Genome for CI

First, the most significant limitation is the framing around “genes.” Although genes allow efficient macro-level description of a crowdsourcing system, it hinders comparisons between different crowdsourcing systems. For example, GroundTruth and CuriOSINTy may appear to have the same how (a sociotechnical system for dividing up complex work) and what
(investigating misinformation online) genes, and different who (novice crowds vs. trained crowds) and why (paid vs. volunteer crowds) genes. Examining the differences more closely, however, it is apparent that CuriOSINTy is far less structured and supports a greater number of tasks than GroundTruth. Indeed, though CuriOSINTy and GroundTruth have two matching genes, CuriOSINTy may be more similar to CrowdSolve than GroundTruth in many ways (e.g., diverse motives, different crowd worker backgrounds, and less structured work).

A second limitation of Malone et al.’s genome framework [222] is that the “why” gene conflates internal and external motivations, and obscures the various subcategories of motivation that an individual may possess. Malone et al. consider three reasons why a crowd worker may engage in work: “money, love, and glory.” However, Malone et al. do not completely unpack these motivations. For instance, ”love” conflates different motivations into one subcategory, thus hiding differences in incentive structures. According to Malone et al., ”love” includes entertainment, becoming friends with others, and contributing to a shared cause. By obscuring these various subcategories, it becomes difficult to compare between and across different crowdsourcing initiatives.

For example, in my work I found that attendees at CrowdSolve took part because of a combination of ”love” motives: education, entertainment, contributing to a cause, and socializing with others. While experts were motivated by love, money and glory; and the victims’ families were motivated almost entirely (and quite literally) by love.

In CuriOSINTy, students took part in the events because it was part of the class structure, which is neither money, glory, or love. Although some mentioned love and glory as motivation for more (or less) enthusiastic participation. Finally, in GroundTruth, while MTurk crowd workers accepted the HIT because of money, our feedback survey showed that they kept accepting the HIT because of love — our HIT was unique among those that were being posted to MTurk because it was enjoyable and broke the monotony of completing lengthy
Table 6.1: The expert-led crowdsourcing framework extends and modifies Malone’s CI genome.

<table>
<thead>
<tr>
<th>Component</th>
<th>Expert-led Crowdsourcing Framework</th>
<th>Malone et al.’s Genome of CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who</td>
<td>Experts and non-professional crowds (trusted or anonymous)</td>
<td>Requester and anonymous crowd</td>
</tr>
<tr>
<td>Why</td>
<td>Intrinsic motivation (enjoyment-based, community-based) and extrinsic motivation (immediate, delayed, social)</td>
<td>Money, love, glory</td>
</tr>
<tr>
<td>What</td>
<td>Sensemaking tasks</td>
<td>Create or decide</td>
</tr>
<tr>
<td>How</td>
<td>Collection, collaboration, competition, or hybrid</td>
<td>Collection (independent or competition) and collaboration</td>
</tr>
</tbody>
</table>

surveys or drawing innumerable bounding boxes.

These examples illustrate that Malone et al.’s why category may be too simplistic and easy to overlook. I contend that fully expanding on the various permutations and combinations for motivation is beyond the scope of this work. Yet, I also draw attention to this gap and encourage careful consideration when designing a crowdsourcing initiative.

A third limitation is that Malone et al. consider the “how” gene to include collections where crowds work independently or in contests, and collaborations where crowds work together. In GroundTruth, experts and crowds worked in real-time, and experts’ reacted based on incoming crowd feedback. While this work was not entirely collaborative, it is also not quite like fully independent microtask crowdsourcing. Further, recent work has shown that collaboration can exist within competitions and vice versa — and that there are benefits to merging competition and collaboration [40, 160, 307]. Along these lines, I introduced beneficial elements of collaboration into a competitive setting in CuriOSINTy. Thus the “how” gene involves not only collection and collaborations, but also competitions, and combinations of collaboration and competition.
6.3. Designing the Expert-Led Crowdsourcing Framework

Building on Malone et al.’s CI Genome

Still, Malone et al.’s genome [221] provides a concise way to highlight different components of a crowdsourcing system. Throughout the rest of this work, I leverage Malone’s genome framework — in combination with other, more recent work — to compare and contrast the frameworks I discuss next, as well as the expert-led crowdsourcing framework I introduce here.

To overcome Malone et al.’s limitations around motivations, I leverage prior work on motivation in crowdsourcing systems [169, 348] to include intrinsic motivations, such as enjoyment-based (e.g., fun and curiosity) and community-based (e.g., building social capital or community) motivations; and extrinsic motivations, such as immediate payoffs (e.g., money), delayed payoffs (e.g., learning and skill development), and social motivations (e.g., value alignment and recognition).

To provide greater nuance and overcome Malone et al.’s limitations around the nature of collaborations, I leverage the concept of a collaboration spectrum [60], though I refer to it as the type of coordinated action. I also make the competition subcategory its own category, and include a hybrid category that represents combinations of competition and collaboration.

6.3.2 Frameworks for Emergent Collective Behavior and Coordinated Action

Defining Emergence

Emergence is defined as the “process where actions and interactions of agents results in the (oft unexpected) global behavior of a system” [278]. Lee and Paine [195] argue that all sociotechnical systems — by virtue of also being social systems — are emergent to some
degree [278]. According to Starbird [292], emergence is a “property that manifests, not something that a system has or is.” Still, it can be reinforced or hindered by its design. Collaborations are emergent because they can form at different times, various actors work to reshape collaborations and create technical arrangements depending on their needs at that time. Lee and Paine describe this as a “a complex and rapidly shifting kaleidoscope of interactions, technologies, and collaborations from teams to multiple organizations” [195]. While Salminen [275], Starbird [292], and Lee and Paine [195] find that emergence is a key feature of crowdsourcing initiatives, it is not addressed in Howe’s original version of crowdsourcing [155] or Malone et al.’s framework [220].

Research in CSCW has studied emergence in sociotechnical systems [195, 242, 292]. For example, Pollock et al. [258] find a contradiction in software design in that, despite a great deal of standardization in software solutions, there is a great deal of diversity in organizational settings to which they can be applied. These organizations engage in a combination of appropriation [97, 132] and “gentrification” [258] work that involves extending and modifying the software to from organization to organization to fit their specific setting. This appropriation and gentrification also enables organizations to align themselves with the software – resulting in their co-evolution alongside each other [195].

**Starbird’s Crowd Work Continuum Framework**

Starbird [292] provides a framework for showing how crowd work (what I refer to as crowdsourcing) during mass disruption events exists along a continuum with four key positions, indicating both the types of work and, for certain settings, points in time representing an evolving project. These four positions along the crowd work continuum are: 1) emergent collective behavior, 2) open source collaborations, 3) virtual teamsourcing, and 4) micro-tasking.
During mass disruption events, such as crises and elections, new social, organizational, and technological configurations arise due to disruptions in normal routines, highlighting opportunities for improvisation [185]. This includes the convergence by a large number of people, such as professional and non-professional volunteers [56, 184] resulting in emergent collective behavior (ECB). At the opposite end of the spectrum exists microtasking or microtask crowdsourcing where requesters intentionally design very rigid structures, and pre-specify and assign work to be completed by the crowd. In microtask crowdsourcing, the rigid structure often severely limits the possibility of emergence. In between exists open source collaborations and virtual teamsourcing. In open source collaborations, crowds have more self-defined goals, tasks, and structures than ECBs. In virtual teamsourcing, the crowd’s goals are defined by another organization or individual (e.g., a client), are increasingly rigid in structure, and with fewer opportunities for emergence.

Starbird highlights how many crowdsourcing systems and communities (including examples within her own work) drift from emergent collective behavior towards rigid microtask crowdsourcing. Starbird attributes this to *structuration*. Giddens [128] defines structuration as the natural progression from nascent, unstructured, emergent groups to highly rigid and stratified organizations — due to frequently repeated work.

Most relevant to this work, Starbird expands upon Malone et al.’s genome framework in two ways by providing greater detail for: 1) the why genome (motivations) and 2) the how genome (emergence, progression of participation, visibility, connectivity, and ownership).

**The Why.** First, for motivation, Starbird focuses on *capital*. The money gene is largely unchanged but is referred to as economic capital (the money gene). The love gene is broken down into four categories: two forms of social capital (bridging and bonding), benevolence, and entertainment. The glory gene is referred to as symbolic capital, such as developing a reputation as a volunteering. Starbird also finds an internal motivation that Malone’s
Table 6.2: The expert-led crowdsourcing framework incorporates “how” elements of Starbird’s crowd work continuum framework.

genome does not cover: self-improvement. Finally, Starbird shows that motivations may vary across individuals and across time, even within one crowdsourcing initiative.

The How. Second, Starbird identifies various facets of the “how” gene. For instance, emergence is identified as a salient component of “how” crowdsourcing initiatives come to be and change over time. Progression of participation is another important facet that is identified, and describes how some crowd workers may progress to become leaders within an initiative. Preece and Shneiderman [260] and Lave and Wenger [191] also identify progression of participation through the reader to leader framework and the framework of legitimate peripheral participation, respectively.

Visibility, connectivity, and ownership are three other factors identified in the crowd work continuum framework [292]. Visibility is a combination of three factors: 1) Whether the “activity is public or visible to others?” 2) If visible, to whom is it visible — the broader public, other crowd workers, or a subgroup? and 3) is it ”done anonymously, pseudonymously or in a way in which the worker’s true identity is known?” [297]. Connectivity refers to whether the initiative allows crowd workers to interact with each other. Finally, ownership refers to who can determine what work is done and how that work is done.

Incorporating Starbird’s Crowd Work Continuum Framework [292] I incorporate portions of Starbird’s framework into the expert-led crowdsourcing framework. Namely: I include a more nuanced conceptualization of how experts and crowds collaborate with each other — such as whether emergence is enabled or hindered, the progression of par-
6.3. Designing the Expert-Led Crowdsourcing Framework

ticipation (i.e., **TRAINING** and **TASK DESIGN**), visibility of the work (i.e., **SECURITY AND PRIVACY**); the connectivity between experts and crowds and within the crowd (i.e., **SYNCHRONICITY** and **COMMUNICATION**); and the **CONTROL, ACCOUNTABILITY, AND AGENCY** afforded to experts and crowds.

Liu’s Crisis Crowdsourcing Framework [209]

Liu [209], extended Starbird’s framework [292] by focusing on the **articulation work** involved in crisis crowdsourcing initiatives. Liu leverages Schmit [281] and Fjuk et al.’s [119] seven dimensions of cooperative work that are a superset of Malone et al.’s four CI genes: who, why, what, how, when, where, and the **articulation work** involved. Thus, Liu’s framework differs from Malone’s and Starbird’s along five axes: 1) considering the articulation work involved to make crowdsourcing initiatives happen; 2) a more nuanced understanding of the stakeholder groups involved (“who”), their expertise, and their relationships; 3) identifying more complex workflows between different stakeholders (“what”); 4) explicitly incorporating spatiotemporal factors (“where” and “when”); and 5) providing greater nuance to social, technological, organizational, and policy interfaces that enable the initiative (“how”).

**Articulation Work.** First, Liu argues that crisis crowdsourcing is largely enabled by the articulation work that takes places within the emergent and convergent coordination of crowds, information, and resources during crises. Liu then constructs her crisis crowdsourcing framework by analyzing the development of novel interfaces that support articulation work needed to make crisis crowdsourcing initiatives possible.

Strauss introduced the term **articulation work** [300] and, according to Liu [209] it was introduced to CSCW as an “analytical framework for understanding the communication and coordination mechanisms within cooperative work that involves mutually dependent actions..."
and collaborating actors” [119, 281, 300]. Articulation work involves “a set of activities required to manage the distributed nature of cooperative work” [281] and is a “supra-type of work in any division of labor done by various actors” [300]. Others have defined it as invisible coordination and negotiation work necessary to make the (actual) work possible [50, 127, 281]. Though traditionally invisible, articulation work is rendered visible through the digital traces left behind by the use of information and communication technologies in crowdsourcing initiatives. This trace data not only includes the digital work produced, but also the metadata and system logs associated with it [125].

**The Who.** Second, while Malone’s and Starbird’s frameworks are largely focused on one type of crowd (e.g., members of the general public, Twitter users, or crisis volunteers), Liu’s framework explicitly considers the possibility of multiple stakeholder groups simultaneously engaging with each other. Namely: 1) affected populations, 2) diasporas, 3) social networks, and 4) digital volunteer communities [209]. Liu also notes that these stakeholder groups may each possess different, but still valuable, sets of expertise. In turn, each group may play a different role within the crowdsourcing initiative and have varying relationships with each other. This is similar to Meier’s [230] concept of “unbounded and bounded crowdsourcing” where crowds can be bound by varying levels of expertise.

**The What.** Third, Liu argues that typical characterizations of crowdsourcing assumed a one-way flow of information from the crowd to the requester. However, in practice, she finds a complex set of interactions and workflows, such as parallel, iterative, directed uni-directional (“crowd seeding”), bi-directional (“crowd-feeding”), and passive (“crowd harvesting”) workflows. In parallel workflows, crowds perform the same task independently to generate an output that is aggregated at a later point. In iterative workflows, each crowd worker iteratively improves upon the work of a previous worker. In directed uni-directional workflows, requesters strategically target certain crowds as opposed to an open call to an undefined
crowd. In bi-directional workflows, there is an active two-way feedback loop where the requester shares information with the crowd in a meaningful way. Finally, in passive workflows, the crowd’s data is harvested or mined without their direct knowledge or consent.

The When and Where. Fourth, Liu explicitly incorporates spatiotemporal factors into her framework — the when and the where. Liu relates the temporal factor, when, to the disaster management lifecycle. This lifecycle starts with detecting crisis events, activating response forces and responding, and iterating across multiple crisis events. Liu’s spatial factor, where, considers the location of crises, the location of crowds (co-located or distributed), and the location of tasks (remotely online or physically at the crisis location).

The How. Fifth, Liu provides greater nuance to how these crisis crowdsourcing initiatives are made possible and are conducted through social, technological, organizational, and policy (STOP) interfaces. Social interfaces include sociocultural “norms, values, beliefs, practices, and relationships with other key stakeholders” [209]. Technological interfaces include technical systems or tools to make the crowdsourcing initiative efficient, reliable, and robust. Organizational interfaces include organization “structures, conceptual schemas, standard operating protocols, and allocation of resources” [209]. Finally, Policy interfaces include legal or organizational policies and regulations on doing the work and engaging with stakeholders.

Incorporating Liu’s Crisis Crowdsourcing Framework I incorporate and adapt five elements from Liu’s crisis crowdsourcing framework:

1. I renders visible the ARTICULATION WORK involved in enabling and conducting ELC investigations.

2. I leverage different STAKEHOLDER GROUPS, who have varying levels and types of expertise,
Chapter 6. A Framework for Expert-Led Crowdsourcing in High-Stakes Investigations

<table>
<thead>
<tr>
<th>Component</th>
<th>Expert-led Crowdsourcing Framework</th>
<th>Liu’s Crisis Crowdsourcing Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articulation Work</td>
<td>Designing and running expert-led crowdsourced investigations</td>
<td>Enabling crowdsourcing during emergent crises and mass convergence events</td>
</tr>
<tr>
<td>Who</td>
<td>experts, crowds, affected stakeholders</td>
<td>Affected populations, diasporas, social networks, digital volunteer communities</td>
</tr>
<tr>
<td>What</td>
<td>Workflows and task design, skills and training</td>
<td>Varied progression of participation</td>
</tr>
<tr>
<td>When</td>
<td>Duration</td>
<td>Crisis management lifecycle</td>
</tr>
<tr>
<td>Where</td>
<td>Co-locatedness</td>
<td>Location of crises, crowds, and tasks</td>
</tr>
<tr>
<td>How</td>
<td>human, financial infrastructure, technological infrastructure, physical infrastructure, policy</td>
<td>social interfaces, technological interfaces, organizational interfaces, policy interfaces</td>
</tr>
</tbody>
</table>

Table 6.3: The expert-led crowdsourcing framework incorporates and adapts elements of Liu’s crisis crowdsourcing framework [209]: articulation work, who, what, when and where, and how.

3. Because different stakeholders may possess useful and disjoint sets of expertise and knowledge, I include workflows and task design along with skills and training into the framework.

4. When and Where — I consider the duration and co-locatedness of the collaboration.

5. I consider the various sociotechnical infrastructures that make the collaboration possible: human infrastructure, financial infrastructure, technological infrastructure, physical infrastructure, and policy.

Lee and Paine’s Model of Coordinated Action (MoCA) [195]

More recently, Lee and Paine [195] introduced a conceptual framework of and for computer-supported collaborative work called the Model of Coordinated Action (MoCA). The MoCA framework describes complex collaborations and environments that are diverse in their memberships and practices. While the MoCA framework extends to all types of coordinated actions, I find that it is equally relevant for describing crowdsourcing initiatives.

The MoCA framework [195] consists of seven dimensions that loosely map onto four of Liu’s
seven dimensions [209]: the articulation work involved in setting up the collaboration; the “who” (communities of practice, scale, and turnover); the “how” (synchronicity, nascence); the “where” (physical distribution); and the “when” (planned permanence). Like Starbird’s crowd work continuum framework, each of MoCA’s dimensions exist along a continuum. However, the MoCA framework does not consider “why” people may engage in a collaboration or “what” they may be doing, though Lee and Paine do consider that different stakeholders may be performing the same or similar actions for different motivations and to achieve different end-goals.

The Who. First, the MoCA framework considers the number of communities of practice (CoP) [191, 332] represented in a collaboration. This helps to capture the “mish mash of backgrounds, perspectives, artifacts, and tools” [195] used within a coordinated action when more than one CoP is involved. One example is a virtual product team within an organization that comprises of five individuals: a product manager, a designer, a software engineer, a data scientist, and a sales person. While the size of the team may be N = 5, each of these individuals comes from a different CoP, each with its own expectations, norms, tools, practices. Thus the number of CoP here is also N = 5.

MoCA also considers the scale of the coordinated action, that is, the number of participants (from 2 to N), and the turnover for those within the collaboration. Turnover refers to the “stability of the participant makeup within a collaboration” [195].

The How. Second, MoCA borrows from Johansen’s original time-space matrix [164] to consider the synchronicity of a collaboration. It is a “continuum of coordinated action ranging from being conducted synchronously (at the same time), to asynchronously (at different times)”, or a mix of both [195]. It also introduces the concept of nascence that differentiates un-established versus established coordinated actions.
Table 6.4: The expert-led crowdsourcing framework incorporates and adapts elements of Lee and Paine’s MOCA framework [195].

The Where and When. Third, MoCA considers the physical distribution of coordinated actions. This is similar to Liu’s conceptualization [209] of where ‘the action happens’ and where ‘the people are.’ Finally, MoCA also considers the planned permanence of the collaboration, i.e., whether and for how long the collaboration is planned or intended to exist for.

Incorporating Lee and Paine’s Model for Coordinated Action I incorporate the following elements from Lee and Paine’s MoCA framework:

1. The different **Stakeholder Groups**, may come from different communities of practice, and the **Scale** of the collaboration may range from ten to hundreds of individuals.

2. While the ELC framework already incorporates synchronicity, I also take into account the nascence of **Workflows and Task Design**.

3. While the ELC framework already incorporates co-locatedness, I incorporate the planned permanence of a collaboration as part of the **Duration**.
6.4 The Expert-Led Crowdsourcing Framework for High-Stakes Investigations

Having reviewed prior crowdsourcing frameworks, I now introduce a framework for expert-led crowdsourcing in high-stakes investigations. While Malone et al. [221], Liu [209], Fjuk et al. [119], and Schmidt [280] delineate the dimensions of cooperative work based on the Five Ws and One H (who, what, when, where, why, and how), it is both a limiting and confusing set of dimensions. For instance, while Malone et al. and Starbird [292] use “why” to describe the motivation for people (the “who”) to take part in a collaboration, Liu uses “why” as the motivation for conducting tasks in the first place (the “what”) [209]. Similarly, one may find the need to motivate the “how” with motivations. Further, one could leverage “how,” “when,” and “where” to describe the nature of a collaboration, but use “who,” “how,” and “what” to describe the underlying sociotechnical infrastructure. Instead, for the sake of simplicity and specificity, the expert-led crowdsourcing framework intentionally omits the language of Five Ws and One H.

Below, I introduce the five categories that comprise the expert-led crowdsourcing framework and discuss their connections to HCI and CSCW. The characteristics of each these categories and their constituent parts have been described in prior HCI and CSCW literature. Like Lee and Paine’s MoCA [195], the ELC framework provides a way to “tie together many of these loose threads” [195] and explain how they weave together into a coherent framework in investigative domains.

The expert-led crowdsourcing framework consists of five overarching categories, each with its own set of constituent dimensions, as shown in Figure 6.1:

- The **investigative domain** being addressed by the ELCI and the **sensemaking** work
practice that occurs within that domain — it includes three dimensions: **feasibility**, **tractability**, and **representability**.

- **The stakeholder groups** includes three different stakeholder groups who may take part in an ELCI: experts, crowds, and other affected stakeholders. These groups may belong to different communities of practice, as well have varied motivations, skills, availability and turnover rates. The scale or the number of stakeholders may also vary.

- **The nature of coordinated action** includes nine dimensions: 1) the type of coordinated action; 2) workflows and task design provided; 3) control and agency for each stakeholder group; 4) security and privacy in the collaboration; 5) communication between individuals; 6) duration of the collaboration; 7) synchronicity of actions; 8) co-locatedness of work and individuals; and 9) training for the crowd.

- **Articulation work** is the component that binds together all of the other components and is the invisible coordination and negotiation work necessary to make the (actual) work possible [50, 127, 281].

**Coordinated Action vs. Collaboration.** Throughout the rest of the framework, I use Lee and Paine’s term “coordinated action” instead of the canonical CSCW term, “collaboration.” This is because the word “coordinated” encompasses a spectrum of working styles [40, 60]: competitions, co-existence, communication, cooperation, coordination, collaboration, and integration — along with hybrids such as competitive collaborations and collaborative competitions [160, 308]. Further, “action” emphasizes goal-directedness that is implicitly suggested by the term “work,” while including “other forms of action that may not traditionally be considered work” [195], such as games with a purpose [328] or serious leisure [298].
6.4. THE EXPERT-LED CROWDSOURCING FRAMEWORK FOR HIGH-STAKES INVESTIGATIONS

Figure 6.1: The five components of the expert-led crowdsourcing framework: 1) investigative domain and sensemaking, 2) shareholder groups, 3) nature of coordinated action, 4) sociotechnical infrastructure, and 5) articulation work. (Note: This diagram was designed using resources from Flaticon.com)

Lee and Paine [195] suggest that “coordinated action is the core of modern CSCW research and is constituted by the interdependence of two or more actors who, in their individual activities, are working towards a particular goal through one or more overlapping fields of action.”

6.4.1 RQ11. Investigative Domain and Sensemaking

The **investigative domain** being addressed by the ELCI and the **sensemaking** work practice that occurs within that domain — it includes three dimensions: **feasibility**, **tractability**, and **representability**.

While many of the prior frameworks that ELC draws on are general-purpose frameworks —
such as Lee and Paine [195], Malone et al. [221], Kittur et al. [175], Kim and Robert Jr. [174], and Valentine et al. [318] — the expert-led crowdsourcing framework is tailored to a specific type of domain and tasks.

Like Starbird’s [292] and Liu’s work [209] on crisis crowdsourcing, I explicitly circumscribe the domain within which this work is done, as well as the type of work. In doing so, the ELC framework carries greater explanatory power to answer the question: “how do expert-led crowdsourced investigations work?” As a result, ELC may not be applicable to all domains and tasks. Rather, as described here, I designed the ELC framework for investigative domains and for tasks that involve sensemaking.

Investigations can be understood as a series of sensemaking tasks, in which the goal is to gather and analyze large amounts of diverse, unstructured information to arrive at a theory or conclusion [117, 257, 269]. Within CSCW and HCI, there exists a large body of work studying and supporting sensemaking in investigations [e.g., 117, 206, 255, 270, 296]. In this work, I explain the ELC framework by exploring investigations in two domains, journalism and law enforcement, which are classic examples of sensemaking [81, 92, 94, 235, 257].

Coordinated action has the potential to speed up and scale up sensemaking by dividing up foraging and synthesis tasks, as well as providing multiple perspectives on schematizing and theorizing about connections, among other benefits [117]. However, coordinated action also creates coordination challenges. For example, actors may have different skills and backgrounds, be geographically separated, need to externalize their thoughts for others, and have access to different parts of the dataset [116, 123, 137, 167].

In the three case studies, I found that ELC can be beneficial for coordinated action involving sensemaking tasks. Findings from the case studies revealed that ELC can mitigate many of the challenges of coordinated action in investigations, enabling experts and crowds to do
more than either could by themselves. However, identifying sensemaking tasks that a crowd can support an expert with requires determining its automatability, tractability, feasibility, and representability.

**Automatability**

ELC may not be applicable to tasks that can be — and perhaps, should be — automated or do not require frequent expert involvement [93, 285]. This includes tasks such as speech-to-text transcription, optical character recognition, and social network community detection, among others. Instead, ELC is applicable to sensemaking tasks that leverage human intelligence to make sense of disparate concepts or ideas and arrive at a singular theory or conclusion.

For example, in GroundTruth, the sensemaking task involved matching a hand-drawn aerial diagram to a satellite image — a task where prior work has shown automated approaches to be insufficiently accurate. While in CrowdSolve the task involved reading through hundreds of pages of case files to generate leads that were both novel and useful to law enforcement. Finally, CuriOSINTy supported multiple sensemaking tasks around identifying content on social media that could potentially be misinformation. Despite recent work in natural language processing (NLP) that claims to be able to separate fact from fiction, researchers have shown NLP approaches to be limited both in scope and their ability to adapt to the dynamic nature of misinformation online [302].

If a sensemaking task cannot be automated to a high degree of accuracy, reliability, and robustness in real-world applications, then an ELC approach may be more suitable. While automatability appears to be a binary choice — yes or no — I have found that it is more of a continuum ranging from low to high. Answering the question — “(How) is this task
‘AI-hard?’ — can help us determine whether or not to consider using ELC.

**Low Automatability.** If automatability is low, then the accuracy, reliability, or robustness may not be sufficient — such as with automated approaches to image geolocation that we piloted prior to developing the GroundTruth system. We found that computer vision attempts at automating image geolocation were insufficiently accurate, placing photos within 200km of the correct location less than 30% of the time [147, 326].

**High Automatability.** If automatability is high, then the sensemaking task may be accurately, reliably, and robustly be performed by one or more automated approaches. For example, NLP techniques have become increasingly accurate at detecting entities — i.e., locations, individuals, and organizations — in real-world datasets [34].

Still, even if automation is not an option, it does not necessarily mean that ELC is the correct choice. Instead, one must still determine whether it is feasible to leverage ELC. Further, it is possible to synergistically combine artificial intelligence and machine learning approaches into certain workflows and tasks. I discuss this more in Section 6.4.4.

**Feasibility**

I find that determining the feasibility for applying ELC involves several sociotechnical factors: 1) the security, privacy, and legal requirements of the work; 2) the expertise, data, and tools required to do the work; 3) spatiotemporal constraints; and 4) monetary constraints. I cover these in more detail in later sections.

**Low Feasibility.** Low feasibility of applying ELC may be in national security or corporate settings that require security clearances and access to sensitive or proprietary data. Alternatively, the expertise required to do this type of work may be highly limited or legally restricted (e.g., nuclear physics or quantum cryptography). Other limiting factors include
restricted access to expertise, data, and tools, or high monetary cost associated with giving a crowd access. For instance, attempting to crowdsource a medical surgery to non-medical professionals is perhaps ill-advised. Lastly, spatiotemporal constraints may also make it infeasible to leverage ELC. For instance, if a task must be done immediately and there is no existing infrastructure, or if it is physically inaccessible.

Even though feasibility may be low, it does not mean that ELC cannot be leveraged. For instance, though CrowdSolve was a first-of-its-kind event to provide a crowd of novices access to real law enforcement case files for two murder investigations, the organizers successfully mitigated these concerns with increased time and effort (e.g., organizing a co-located event, using physical case files, restricting use of technology). Similarly, Trace Lab’s OSINT Search Party CTF [78] closely works with law enforcement to choose missing persons cases, but only provides the crowd with publicly available information.

**High Feasibility.** In contrast, ELC may be highly feasible when four conditions are met. First, there are minimal or few security, privacy, and legal considerations or if they can be overcome easily. Second, the expertise, data, and tools may be readily accessible. For example, in both GroundTruth and CuriOSINTy, we worked with journalists and researchers who were relatively easy to access — although we did provide gift cards to compensate experts for their time. In addition, we used publicly available data (e.g., OSINT and social media data) and tools (e.g., Google Maps, Hoaxy, and the Web Archive).

Third, spatiotemporal constraints may be minimal, overcome easily, or even be beneficial. If the crowd is not required to be involved for a long period of time, then ELC may be more feasible, whereas if the crowd is able to be present for several days or weeks, that may make ELC even more feasible. For example, in CuriOSINTy, we trained a crowd over three months to be able to do more complex work than would have been possible if they were trained for less than a minute (like with GroundTruth).
Fourth, feasibility may determined by the availability of monetary support in some instances. For example, CrowdSolve would not have been possible without financial contributions from the crowd of amateur sleuths attending the event. We discuss this in more detail under financial infrastructure.

Though ELC may be highly feasible, it may not be fully realized unless the task is tractable and representable.

**Tractability**

Tractability involves considering whether the work at hand can be easily led, taught, and controlled. This is a crucial factor in determining the applicability of expert-led crowdsourcing.

Expert-led crowdsourcing relies upon experts to lead the crowdsourcing initiative, to teach crowds valuable skills that can further augment the quality of their work, and to control the crowd’s actions to mitigate concerns with many crowdsourced investigations. This includes mitigating information leaks as well as vigilantism, misidentification, and doxxing.

**Easy Tractability.** Easy tractability refers to instances where it is relatively easy to direct or control the crowd’s actions, as well as teach them directly applicable skills.

In GroundTruth, we conducted multiple experiments to determine that crowds could perform image geolocation with less than a minute of training. In CrowdSolve, we found that the event organizers used a variety of regulatory mechanisms — laws, norms, markets, and architecture — to mitigate the possibility of unwanted behavior such as leaked information or vigilantism (cf. Lessig’s Code and Other Laws of Cyberspace [202]). Finally, in CuriOSINTy, though we required considerable time to teach the crowd investigative skills, it was easy to teach these skills in a classroom setting. The classroom setting and focus on publicly available
6.4. The Expert-Led Crowdsourcing Framework for High-Stakes Investigations

information also made it easy to trust the crowd, requiring us to exert less “control” over the crowd.

**Difficult Tractability.** Difficult tractability refers to instances where the crowd’s actions may be difficult to direct, their actions cannot easily be controlled or are likely to become uncontrollable, and it may be difficult to teach them directly applicable skills.

For example, it may be feasible to have an expert direct and teach a crowd of citizen scientists to identify poisonous red cap mushrooms, but teaching them to *safely sequester or remove* these mushrooms may be a dangerous undertaking.

**Representability**

Representability involves considering whether the experts’ sensemaking tasks can be easily translated or represented in a form that is understood by crowd workers with varying levels of (relevant) expertise; and whether the crowds’ work can be easily represented to the expert.

**Easy to Represent.** ELC may be easier to leverage if the sensemaking task is easily representable. For example, in GroundTruth, we spent considerable effort designing *shared* representations that allowed experts to communicate complex or tacit knowledge to the crowd as well as allow the crowd’s feedback to be aggregated and displayed to the expert. GroundTruth leveraged three types of shared representations: a shared aerial image, a shared map grid, and a shared heatmap. The aerial diagram allowed the expert to tell the crowd what to search and the map grid indicated where to look. The crowd’s aggregated feedback was then displayed in the same map grid as a heatmap, allowing the expert to stay within their work environment while still incorporating the crowd’s feedback.

In CrowdSolve, however, representability was more difficult, though not improbable. The event organizers spent considerable effort and time to convert hundreds of pages of existing
case files into a format that the crowd could easily make sense of. Representability was relatively more easy in CuriOSINTy because we started with mostly *clean slate* investigations, and we leveraged training, scaffolding, and rubrics to help communicate expert work practice to the crowd.

**Difficult to Represent.** ELC may not be appropriate if it is difficult to represent the sensemaking task in a way that crowd workers can understand, or to represent crowd feedback in a way that experts can understand.

For instance, flash organizations involve coordinated action among several experts — such as a film crew composed of a director, camera crew, and video editors [318]. Each of these experts may possess one to several years of experience, and as a result, developed mental models for communicating and coordinating work with other members of a film crew. These mental models may be difficult to represent or doing so may have limited return on investment (ROI).

### 6.4.2 RQ12. Stakeholder Groups

The **stakeholder groups** includes four different stakeholder groups who may take part in an ELCI: *organizers, experts, crowds,* and other *affected stakeholders*. These groups may belong to different **communities of practice**, as well have varied **motivations, skills, availability** and **turnover** rates. The **scale** of the coordinated action may also vary.

**Organizers, Experts, Crowds, and Affected Stakeholders**

In the three case studies, I found that ELC can mitigate many of the challenges of coordinated action in investigations, enabling experts and crowds to do more than either could alone — by merging their complementary strengths. For example, GroundTruth allowed experts to
work with novice crowds geolocate images within minutes, compare to hours or days of the expert was doing this work alone. Similarly, in CrowdSolve, six experts to guide and train over 250 amateur sleuths to make sense of and generate valuable leads for a lone detective in charge of two decades-old murder investigations. CurioSINTy also enabled experts to work with 46 trained crowd workers to identify and successfully debunk hundreds of pieces of misinformation. How, then, do we determine who is an expert and who is a crowd worker? Further, who brings together experts and crowds?

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>CrowdSolve</th>
<th>GroundTruth</th>
<th>CurioSINTy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organizers</strong></td>
<td>Event organizers</td>
<td>Research team</td>
<td>Research team</td>
</tr>
<tr>
<td><strong>Expert</strong></td>
<td>Police detective, retired US Marshal, forensics analysts, speech pathologist</td>
<td>Professional journalists and human rights activists with experience in image geolocation</td>
<td>Research team, as well as researchers studying health and human rights misinformation</td>
</tr>
<tr>
<td><strong>Crowd</strong></td>
<td>Amateur sleuths with interest in true crime</td>
<td>Amazon Mechanical Turk workers</td>
<td>Trained students</td>
</tr>
<tr>
<td><strong>Affected Stakeholders</strong></td>
<td>Victims’ family members</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Figure 6.2: How the four types of stakeholder groups existed within the three studies.

**Organizers.** I refer to organizers as individuals who set up the ELC investigation, including determining the feasibility of applying ELC; developing or assembling its sociotechnical infrastructure; and specifying the nature of the coordinated action. Organizers, like other stakeholder groups, benefit from the success of an ELC investigation, though perhaps in more indirect manner.

Unlike typical emergent coordinated action, such as in crisis situations, ELC exhibits less emergence in the three studies presented in this work. Here, the three studies required more intentional, directed effort from a small number of organizers (versus more diffuse and small amounts of efforts from a large number of individuals in emergent coordinated action).

For example, in CrowdSolve, the event organizers brought together the lead detective, other law enforcement experts, the victims’ families, and the crowd. They also organized every
aspect of the event, ranging from setting the schedule, to determining what role other stakeholders perform, to synthesizing the crowd’s work. In GroundTruth and CuriOSINTy, the organizers — the research team — designed and developed the sociotechnical infrastructure (the crowdsourcing systems) and brought together experts and crowds to use the system. However, the organizers did not synthesize the crowds’ work products. Instead, the crowdsourcing systems and experts synthesized the crowd feedback.

There may be other scenarios where ELC exhibits emergence, however, such as the Sedition Hunters community on Twitter where professional investigators coordinate the actions of members of the public to identify individuals who took part in the attack on the U.S. Capitol on January 6th, 2021 [263]. In these cases, the organizers of the collective action may be also the experts. Even in CrowdSolve, two of the organizers were also experts with domain knowledge in law enforcement.

Experts. In this framework, I refer to experts as professional investigators with one or more (typically several) years of experience conducting investigations that predominantly involve sensemaking tasks. Prior work has shown that professionals develop more efficient and effective mental models for doing sensemaking work, compared to novices with little to no experience involving deliberate practice [20].

Within the context of an ELC investigation, these experts take on additional meta-tasks (or articulation work) similar to that of a “requester” in traditional crowdsourcing systems [175]. This includes training the crowd, generating micro or macrotasks for the crowd to work on, guiding and reviewing the crowd’s work through feedback, and synthesizing the crowd’s work into a final result. To support experts with this articulation work, there may be workflows, as well as sociotechnical infrastructure — i.e. human, technological, physical, and financial infrastructure — and policy.
In CrowdSolve, experts included the lone police detective in charge of the two murder investigations, six former or current law enforcement professionals (e.g., a US Marshal, and investigators with experience in strangulation, blood splatter analysis, and speech pathology). In GroundTruth, we studied how to support journalists, human rights advocates, military intelligence analysts, among others, with image geolocation. Each expert had over one year of experience conducting image geolocation, and on average, experts and over eight years of experience. Finally, in CuriOSINTy, the research team acted as experts when conducting OSINT investigations into misinformation, but we also recruited two experts who had at least one year of experience investigation human rights and health misinformation.

Identifying and recruiting experts, however, is not an easy task and requires considerable levels of articulation work, as we will describe in Section 6.4.5. The organizers of an ELC investigation face several challenges in recruiting experts. For instance, in CrowdSolve and GroundTruth, experts were compensated for their time. Without compensation, experts who are strapped for time may be less willing to test out a system where the benefits or unclear or unknown. If the system cannot be tested, then it becomes difficult to demonstrate that the system works — essentially forming a vicious cycle. Long-term engagement with experts, mutual collaboration, and, when relevant, financial compensation, may help to break this cycle to build and evaluate an ELC system.

Crowds. I refer to crowds as non-professional members of the public [184]. These crowd workers may possess varied types and levels of expertise in different domains (such as where they are employed). Though beneficial, ELC does not assume that crowd workers have expertise in an investigative domain. Instead, ELC overcomes the limitations of working with a non-professional crowd through expert guidance and training, as well as through carefully designed workflows and tasks with scaffolding and rubrics. I describe this in more detail in Section 6.4.3. ELC investigations are only possible, however, when crowd workers
are motivated to participate, which I explain below.

In an ELC investigation, crowds help scale up and speed up expert work practice by completing the tasks that experts provide to them. Depending on how the ELC investigation is set up, crowds may also perform tasks that experts did not anticipate, such as recruiting additional crowd workers or developing software tools [140, 346].

In CrowdSolve, experts gave the crowd specific sensemaking tasks that they believed would generate the most useful leads, such as analyzing transcripts, crime scene photos, and other evidence. Still, some crowd workers searched for and provided other useful information that the experts did not explicitly ask for, such as weather reports and satellite imagery. In GroundTruth, the crowds provided feedback on search areas ranging from 1.44 to 36 km², narrowing the search area by 27.3 to 66.5% (mean = 43%). In seven out of 10 cases, crowd workers were able to identify the correct location taking between 5m 16s to 22m 10s (mean = 11m 37s). CuriOSINTy provided extensive training and structure that allowed 46 crowd workers to discover and debunk 228 unique pieces of misinformation in less than 90 minutes, as well as collect 369 additional pieces of metadata.

Recruiting crowds can be as easy as hiring anonymous crowd workers through Amazon Mechanical Turk’s API (GroundTruth), or as complex as conducting a training course (CuriOSINTy) or advertising an event to true crime enthusiasts from around the world (CrowdSolve). Each of these approaches comes with their benefits and challenges, including varying motivations, levels of trust, availability, and turnover. I discuss these factors in the following subsections.

Affected Stakeholders. Affected stakeholders are those individuals who may directly be affected by the crowd’s efforts and may be able to provide specific input that could make the ELC investigation more successful. Including affected stakeholders may not be a requirement
for ELC investigations, but their participation may be highly beneficial, such as attracting a greater number of crowd workers or experts, motivating them, and providing them with previously unknown but useful information.

For example, in CrowdSolve, we found that the victims’ families presence not only led to a greater number of crowd workers taking part in the event because of their deep fascination with true crime media, but also motivated them to stay focused and work harder to support the victims’ families and help bring the perpetrators to justice. Crowd workers also said that the victims’ families provided more details about the victims that were not present in the law enforcement case files. For example, one crowd worker said that a family member told them about the victim’s daily habits — which helped them rule out potential reasons for the victim to go missing.

Like CrowdSolve, GroundTruth and CuriOSINTy could have benefited from including affected stakeholders. For instance, in GroundTruth, once the expert had narrowed the search location to a particular country or region, we could have relied upon the regional knowledge of crowd workers or other informants living in that area [134, 288]. In this way, we might have been able to recruit fewer crowd workers or geolocated the image faster. We discuss the importance of tangentially relevant skills and communities of practice next.

**Skills and Communities of Practice**

When recruiting stakeholders, it may be beneficial to recruit them based on desired skills, tools, or communities of practice that they may belong to. Wenger [332], building on the notion of legitimate peripheral participation, coined the term “communities of practice” (CoP). CoP defines a group of people with a common concern or passion, whereby newcomers increasingly interact with more experienced members of a community. Over time, these
newcomers incorporate values, norms, skills, and tools of those within the community [195]. While CoPs take time to form, ELC investigations may benefit from recruiting individuals from one or more existing CoPs to take advantage of their unique skills and tool sets.

For example, in CrowdSolve, many of the crowd workers wished that a portion of the crowd consisted of individuals with a background in criminology so that they could learn from them and also more effectively generate leads as a group. Though CrowdSolve did not require that crowd workers have any specific skill set, the second version of the event created a highly discounted rate (up to 80%) to attract those with a background in criminology. In GroundTruth, experts said that when they faced setbacks in their investigation, they would often ask experts with relevant geographic expertise to consult on the investigation. Finally in CuriOSINTy, we brought together crowd workers with both technical expertise in computer science (to develop and use tools) and topical expertise in political science (to help develop these tools and increase overall work output).

More broadly, ELC investigations may benefit from bringing stakeholders with backgrounds in leadership, data science, writing, or those with access to tacit knowledge or difficult-to-obtain data [213, 214, 225].

Though CoP appear to be a small dimension within the broader ELC framework, it is a complex multidimensional framework of its own. Like Lee and Paine’s MOCA Framework [195], I note that CoP dimension here cannot capture the complexity and nuance of Wenger’s CoP, but provides a starting place when considering how ELC investigations can be improved.

Motivations

Self-determination Theory [87] describes two types of motivations that individuals can possess: intrinsic motivations and extrinsic motivations. Intrinsic motivations act within an
individual and are personally rewarding (e.g., enjoyment). Extrinsic motivations act on the individual from outside, and any result in an external reward (e.g., monetary compensation).

Within the field of crowdsourcing, Kaufmann et al. [170] conducted a survey of the literature to categorize the intrinsic and extrinsic motives that crowd workers had for participating in crowdsourcing systems. They found two types of intrinsic motivations: enjoyment-based and community-based motivations; and three types of extrinsic motivations: immediate payoff, delayed payoff, social motivations.

Both intrinsic and extrinsic motivations can be leveraged to motivate crowds to participate in crowdsourcing initiatives, including ELC investigations. However, different individuals are motivated by the two types to different extents. Further, intrinsic motivations are more difficult to design for than extrinsic motivations, since the latter are often external factors.

If an ELC investigation is designed where crowd workers are not adequately motivated, it may lead to lower performance. For example, in CuriOSINTy, we introduced a largely competitive environment, and found that those who were motivated by competitions were more engaged and motivated to participate, while those that were motivated more by collaboration were less engaged and motivated. When designing an ELC investigation, crowd workers should be allowed to opt-in to the investigation and be informed about the nature of the coordinated action prior to participating.

<table>
<thead>
<tr>
<th>Motivations</th>
<th>CrowdSolve</th>
<th>GroundTruth</th>
<th>CuriOSINTy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enjoyment-Based</td>
<td>Pastime, direct feedback</td>
<td>Skill variety</td>
<td>Direct feedback</td>
</tr>
<tr>
<td>Community-Based</td>
<td>Particular community (true crime fandom)</td>
<td>None</td>
<td>Particular community (learning community)</td>
</tr>
<tr>
<td>Extrinsic</td>
<td>Immediate Payoffs</td>
<td>Compensation</td>
<td>Grade</td>
</tr>
<tr>
<td>Delayed Payoffs</td>
<td>Signaling</td>
<td>Unknown</td>
<td>Skill advancement</td>
</tr>
<tr>
<td>Social Motivations</td>
<td>External values, social learning</td>
<td>Unknown</td>
<td>External values, social learning</td>
</tr>
</tbody>
</table>

Figure 6.3: Crowd workers’ intrinsic and extrinsic motivations for the three studies.
Intrinsic Motivations. There are two types of intrinsic motivations that exist within crowdsourcing platforms: enjoyment-based and community-based motivations [170]. Enjoyment-based motivations include skill variety, task identity, task autonomy, direct feedback, and pastime or hobby [170]. Community-based motivations include identifying with a particular community and social contact.

CrowdSolve and CuriOSINTy leverage enjoyment-based motivation in the form of alternate reality games (ARGs). ARGs combine physical and digital artifacts to set up cryptic mysteries that are intended to be solved by crowds. To succeed, crowds must quickly share information and solutions, and leverage their varied expertise [124]. ARGs are also designed to be highly immersive [227], and some ARG designers even attempt to create learning environments [246].

CrowdSolve also leverages community-based motivation in the form of true crime fandom. Fandoms can be powerful and transformative sites of social support [107, 153], learning [111], and creativity [114, 214], as well as sites of toxic behavior and targeted harassment [110, 165]. In prior work, true crime fans have engaged in positive forms of collective action towards investigations and cases, ranging from justice reform [287] to uncovering new leads and overturning wrong verdicts [159].

Although GroundTruth largely motivated MTurk crowd workers through extrinsic factors (i.e. money), our feedback survey showed that some kept accepting the HIT because of intrinsic factors — our HIT was unique among those that were being posted to MTurk because it was enjoyable and broke the monotony of completing lengthy surveys or drawing innumerable bounding boxes.

Extrinsic Motivations. There are three types of extrinsic motivations that exist within crowdsourcing platforms: immediate payoffs, delayed payoffs, and social motivations are [170].
Immediate payoffs is typically monetary compensation, while delayed payoffs includes signaling and skill advancement [96]. Social motivations include external values, external obligations and norms, and indirect feedback [183].

GroundTruth leveraged immediate pay-off, i.e., crowd workers recruited on Mechanical Turk participate in exchange for monetary compensation. In contrast, CrowdSolve and CuriOSINTy relied on social motivations, such as a desire for altruism and justice (both), and social learning (CuriOSINTy).

Scale

In an ELC investigation, crowds help scale up and speed up expert work practice. Like Lee and Paine’s MoCA framework [195], we define scale as the number of participants involved in the ELC investigation, ranging from ten to N (where N is undefined).

Typically, we would expect the number of experts to remain relatively small compared to the number of crowd workers. The ratios were 6 experts to 250 crowd workers in CrowdSolve, 1:51 on average (11 to 567 overall) in GroundTruth, and 3:46 on average in CuriOSINTy.

While a large scale is taken as a given for crowdsourcing applications, crowdsourcing systems must explicitly be designed to be flexible and extensible. For instance, significantly scaling down a workflow or system designed for a crowd of ten thousand may result in suboptimal performance or unanticipated results. Similarly, scaling up a system meant for one hundred people may require more complicated social and technological arrangements. To put it bluntly, scale matters.
CHAPTER 6. A FRAMEWORK FOR EXPERT-LED CROWDSOURCING IN HIGH-STAKES INVESTIGATIONS

Availability and Turnover

Another important consideration when involving stakeholders is their availability and turnover. *Availability* refers to the amount of time an individual is present for a specific investigation, while *turnover* refers to the rate at which participants enter and leave.

<table>
<thead>
<tr>
<th>Motivations</th>
<th>CrowdSolve</th>
<th>GroundTruth</th>
<th>CuriOSINTy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic</td>
<td>Enjoyment-Based</td>
<td>Pastime, direct feedback</td>
<td>Skill variety</td>
</tr>
<tr>
<td></td>
<td>Community-Based</td>
<td>Particular community (true crime fandom)</td>
<td>None</td>
</tr>
<tr>
<td>Extrinsic</td>
<td>Immediate Payoffs</td>
<td>None</td>
<td>Compensation</td>
</tr>
<tr>
<td></td>
<td>Delayed Payoffs</td>
<td>Signaling</td>
<td>Unknown</td>
</tr>
<tr>
<td></td>
<td>Social Motivations</td>
<td>External values, social learning</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Figure 6.4: Crowd workers’ availability and turnover for the three studies.

Implicit within these definitions are collaborations ranging from restricted-entry collaborations to fully open and public collaborations. In restricted-entry collaborations — like CrowdSolve and CuriOSINTy — crowd workers may be highly available during an investigation and membership is relatively stable for the duration of the investigation, i.e., turnover is low. In fully open and public collaborations — like GroundTruth or many mass convergence events — not only may the scale be very large, but most crowd workers may only be available for short periods of time, but turnover rapidly.

While, Starbird [292], Lee and Paine [195], and Liu’s [209] frameworks largely consider the crowd as emergent stakeholders whose membership and participation may be transient or evolves over time. Although the ELC framework allows for crowds to be transient and interchangeable “human processing units” [85], we urge readers to consider that the benefits of ELC may only be fully realized when a crowd is fixed. Prior work has shown benefits to having a trusted and largely constant set of crowd workers, including greater efficiency and work quality, as well as an improved sense of belonging and purpose among crowd workers.
6.4. THE EXPERT-LED CROWDSOURCING FRAMEWORK FOR HIGH-STAKES INVESTIGATIONS

For example, GroundTruth relied on paid crowd workers from Amazon Mechanical Turk. Because we could not assure experts that every single crowd worker would provide high-quality work 100% of the time, we opted to aggregate work from three different crowd workers for each microtask. In contrast, for CrowdSolve and CuriOSINTy, we were not only able to trust the crowd and hold them accountable, but we could also train them for a longer duration (months in CuriOSINTy and hours in CrowdSolve, versus less than one minute in GroundTruth). This allowed us to assign more complex tasks to the crowd than would typically be possible.

However, what is not clear is what the turnover of experts and affected stakeholders should be in ELC investigations. ELC investigations explicitly consider, and perhaps necessitate, that experts and affected stakeholders are available for the entire duration of the investigation. This is because they have a (more) vested interest in the outcome of the investigation, but also because ELC expects that experts retain control and ownership over the investigation — allowing experts to be held accountable for any mistakes.

6.4.3 RQ13. Nature of Coordinated Action

The nature of coordinated action includes nine dimensions: 1) the type of coordinated action; 2) workflows and task design provided; 3) control and agency for each stakeholder group; 4) security and privacy in the collaboration; 5) communication between individuals; 6) duration of the collaboration; 7) synchronicity of actions; 8) co-locatedness of work and individuals; and 9) training for the crowd.
Type of Coordinated Action

While Lee et al. [195], Starbird [292], and Liu [209] focus on more collaborative or cooperative crowdsourcing, and Malone et al. [221] focused on co-existent or traditional crowdsourcing, the ELC framework exists along a spectrum of coordinated action. The coordinated action spectrum (Fig. 6.5) ranges from crowdsourcing competitions and co-existing or traditional crowdsourcing on one end to cooperative crowdsourcing and collaborative crowdsourcing on the other end, along with hybrid versions [40, 60, 207].

![Figure 6.5: The four types of coordinated action that may exist for expert-led crowdsourced investigations.](image-url)
Compete. Crowdsourcing competitions involve competition between individuals or teams, who collectively comprise a single crowd — often with a shared goal, and no interaction between members of the crowd. Here, crowd workers take on distinct roles but have a great deal of agency to make decisions or take action. Examples include innovation contests [e.g. 183, 308], design contests [e.g. 160], and capture the flag competitions [e.g. 226].

Co-Exist. In co-existent or traditional crowdsourcing, there is minimal or no interaction between members of the crowd. Once work is completed by crowd workers, it is sent directly to the requester, or crowd workers may use shared artifacts that allow work to be sent unidirectionally from one crowd worker to the next. Here, crowd workers have minimal agency and their roles are highly constrained by structured workflows. Examples include surveys, games with a purpose [e.g. 328], as well as unidirectional pipelines and workflows from crowd to requester [e.g. 44, 175, 205, 216].

Cooperate, Coordinate, or Communicate. In cooperative crowdsourcing, there is informal or intentionally planned interaction between members of the crowd and between the crowd and requester, and includes sharing of information and adjusting or re-aligning work. Here, crowd worker agency is constrained, but crowd workers are still given some freedom within their roles due to the use of less structured or flexible workflows. Examples include multi-directional pipelines and workflows within the crowd and between crowd and requester [e.g. 205, 206] as well as emergent collective behavior [e.g. 195, 292]. GroundTruth is an example of cooperative crowdsourcing where experts dynamically reacted to real-time crowd feedback.

Collaborate. In collaborative crowdsourcing, there are longer-term and interdependent interactions between members of the crowd and between the crowd and requester. Similar to crowdsourcing competitions, individuals here take on distinct roles but have a great deal of agency to make decisions or take action. Examples include virtual teams [e.g. 195, 292],
flash teams [e.g. 264], and flash organizations [e.g. 265, 318]. CrowdSolve is also an example of collaborative crowdsourcing where crowd workers worked together to investigate the two murder investigations, exchanging information with each other, as well as receiving feedback from experts on their findings.

**Hybrid.** There are also hybrid versions, such as competitions with collaboration [160, 308] or collaborations with competition [40]. As we reviewed in the related work, prior work has shown benefits to merging the beneficial elements of competition and collaboration. Similarly, CuriOSINTy leveraged a hybrid setup of a CTF competition with elements of collaboration to both make a CTF applicable for a real-world application and mitigate duplication of effort and knowledge silos.

**Choosing the Type of Coordinated Action.** In some instances, it may be possible to intentionally choose the type of coordinated action when designing an ELC investigation. For example, in all three studies explored here, we specified how crowd workers interacted with each other. However, for emergent collective action, it may be more difficult to specify the type of coordinated action because it is dictated by the surrounding sociotechnical infrastructure, including the tools available to coordinate work and social norms present within a community of practice.

When designing an ELC investigation, choosing the type of coordinated action also depends on three factors: 1) the communities of practice that the investigation draws stakeholder groups from; 2) factors that motivate stakeholders; 3) the training, task design, and workflows that exist or will be designed to support coordinated action. We briefly discuss these factors here, but cover them in more detail in the following sections.

First, the community of practice that stakeholders are from affects the norms and practices that they are familiar with. For example, GroundTruth relied upon workers from Amazon
Mechanical Turk who are more familiar with co-existent crowdsourcing, while TraceLabs involves a competitive setup because the crowd comes from the adversarial and competitive field of cybersecurity.

Second, factors that motivate stakeholders can also determine the type of coordinated action that is chosen. For example, crowd workers in GroundTruth were motivated by immediate payoff (monetary payment) and task independence (being able to work whenever they are available); while in CrowdSolve, crowd workers were motivated by enjoyment (sense of immersion and curiosity) and social values (altruism). If an ELC investigation is designed where crowd workers are not adequately motivated, it may lead to lower performance, such as in CuriOSINTy, where we introduced a largely competitive environment to a crowd.

Third, training, task design, and workflows may also affect the type of coordinated action. If minimal training is provided, or if highly structured tasks and workflows are used, co-existent crowdsourcing may be more appropriate. In contrast, if there is extensive training provided or workflows are unstructured, competitive, collaborative, or hybrid approaches may be more appropriate.

**Security, Privacy, Control, and Agency**

Security, privacy, control, and agency are four inter-related dimensions that can enable, encourage, or constrain coordinated action.

**Security.** Security is defined as measures taken to prevent unauthorized access or behavior. This includes measures to ensure secure access to information; protect stakeholders from physical or virtual harm; minimize opportunity for stakeholders to engage in vigilante behavior; and reduce the likelihood of interference from unwanted (outside) actors.

In CrowdSolve, security measures were explicit (e.g., security guard, ID checks, and non-
Figure 6.6: Definitions of security, privacy, control, and agency, as well as how they existed within the three studies.

disclosure agreements) given the sensitive nature of the two murder investigations. Security was more implicit in GroundTruth and CuriOSINTy, through image obfuscation and leveraging a trusted crowd in a classroom setting, respectively.

**Privacy.** Privacy is different from security in that it refers to the ability to control, access, and regulate information and is an assessment of who can access what information. Includes whether anyone (stakeholders or members of the public) can see the work being done; who can see this work; and whether stakeholders of members of the public know those taking part (eponymous, pseudonymous, or anonymous).

Privacy, though explicitly considered in all three studies, was more extensive in CrowdSolve through physical case files, restricted access, and intentionally restricted technology use. Privacy was more limited in GroundTruth and CrowdSolve. In GroundTruth, there were secure user accounts for experts, but crowd worker participation was unrestricted (anonymous crowd workers on MTurk could sign up). CuriOSINTy used secure user accounts for experts.
and crowd workers, but we restricted who could participate to a trusted, known crowd.

**Control.** Control refers to the power to determine a course of action for others, but also to be held responsible for any potential outcomes. This includes what is done, who works with whom and how, and who takes responsibility for the outcomes.

In CrowdSolve, experts and organizers had significant control: they specified questions to be answered, controlled access to case files, trained the crowd, synthesized outputs, and took some responsibility for the outcomes of the investigation. However, because crowd workers’ identities were known and they signed the non-disclosure agreement, as well as due to state and federal laws, the crowd could be held accountable for their actions in case any members leaked information or engaged in vigilante behavior.

In GroundTruth, experts had near-total control: they could tell the crowd what to look for and where to look and they controlled access to information through obfuscation of the aerial diagram. Because experts had near-total, and GroundTruth relied on anonymous crowd workers who were unlikely to be held responsible for their actions, experts had to take full responsibility for the outcomes of the investigation.

Finally, in CrowdSolve, experts had limited control: they specified general topic areas and trained the crowd. However, the crowd decided what to look for and where to look. Leveraging a known crowd, as well as a classroom setting and university policies, meant that the crowd could be held responsible for their actions.

Control and responsibility appear to be synonymous in these three scenarios — there were mechanisms to hold those with more control accountable for their actions. However, ELCs could exist where there are anonymous experts or crowd workers who are given (or take up) a significant level of control, but by being anonymous it would be difficult to hold them accountable for the outcomes of an investigation. Examples include QAnon conspiracy
theories (e.g., PizzaGate) [118, 254]; directed, crowdsourced harassment on 4Chan, 8Chan, and its successor (e.g., Gamergate) [21].

**Agency.** Agency is the freedom to determine how one does something, such as how to do the work (e.g., what tools or information to use), and where or when to do it. Agency is different from control. Control refers to who has the power to decide others’ actions or place restrictions upon others’ actions, while agency refers to how much freedom an individual upon their own actions.

In CrowdSolve, experts had more agency, while crowd workers had some agency: crowd workers can choose whether or not to do the work, or how they work within a team, but could not decide what to do. In GroundTruth, experts had complete agency, while the crowd had limited agency: crowd workers can decide whether or not to do the work, but cannot decide how to do it. In CuriOSINTy, experts and crowds had similar levels of agency: crowd could decide what work to do and how to do it, but experts could also do their own work as well as incorporate the crowd’s.

Agency is dependent upon control, but also security and privacy. For instance, if there are no security and privacy measures in place, then the crowd can choose to take control of an investigation and thus have more agency. Agency for experts and crowds are not inherently at odds with each other. That is, both experts and crowds can have high levels of agency, but for different tasks, such as in CuriOSINTy.

In addition, when crowds are given greater levels of agency, they may support experts in unanticipated ways — such as building software or recruiting additional crowd workers. Along these lines, in CrowdSolve, some crowd workers searched for and provided additional useful information (e.g., weather conditions and soil type pertaining to one of the investigations). Too much agency, however, can result in unwanted behavior. For instance, some
crowd workers visited the decades-old crime scenes, while others visited a nearby library to search for weather reports for the week that the crimes were committed.

**Regulating Human Behavior through Lessig’s New Chicago School Theory.** Lawrence Lessig, in his book, “Code and Other Laws of Cyberspace,” presents the New Chicago School theory or “pathetic dot theory” [201, 202] that posits four ways to regulate human behavior: laws, norms, markets, and architecture. In Section 6.4.4, I describe how these regulation mechanisms are instantiated in sociotechnical infrastructure.

Security, privacy, control, and agency can also be regulated through laws, norms, markets and architecture. For example, as laws, security and privacy mechanisms existed as NDAs and state or federal laws in CrowdSolve and course policies in CuriOSINTy. They can also be instantiated through norms, such as verbal warnings and a sense of common purpose that the organizers leveraged in CrowdSolve, or the “OSINT mindset” that we leveraged in CuriOSINTy. CrowdSolve and GroundTruth leveraged markets but in opposite ways. In CrowdSolve, crowd workers could only attend the event if they paid a high registration fee (several hundred dollars), while in GroundTruth, crowd workers were paid and payment could be withheld if the quality of their work was subpar. Finally, architecture can be instantiated through technological or physical restrictions (e.g., secure use accounts, obfuscated imagery, or physical case files), as well as providing separate crowd workers access to partial sets of data. Lessig’s theory [201] can similarly be applied to agency and control.

**Training, Task Design, and Workflows**

Perhaps one of the most important aspects of ELC, and one that differentiates it from most traditional crowdsourcing applications, is the training provided to the crowd, along with the design of tasks based on complementary strengths and the minimal use of workflows.
Training. *Training* in ELC refers to any investigation-specific instruction or preparation that the ELC system, organizers, or expert provides to the crowd, and that can be directly applied to the investigation at hand. By training novice crowds, it minimizes the need for more extensive expert supervision, quality control mechanisms, complex workflows, and microtasking. In contrast to ELC, traditional crowdsourcing approaches incorporate hierarchical review mechanisms [142], iteration [208], duplication and aggregation [262], complex workflows [265], and extensive microtask design [175]. It is to be noted that a trained novice may be more proficient at a particular task than an untrained novice, but still less than an expert [20].

The amount of training provided can minimize the need for overly simple microtask design and complex workflows, while also increasing the quality of work, and allowing crowds to perform more complex tasks. For example, while GroundTruth provided limited training in the form of a one to two minute-long tutorial, it also meant that we needed to extensively evaluate the single task (matching an aerial diagram to a grid of satellite imagery), aggregate crowd worker feedback, and also structure it in a way that experts could easily specify the work to the crowd and visualize their work in real-time.

In contrast, CrowdSolve and CuriOSINTy provided extensive training to the crowd, ranging
in duration from several hours to several weeks. In CrowdSolve, experts with specialized skills in forensics analysis, speech pathology, among others, provided hour-long training that walked crowd workers through every aspect of the two murder investigations. Without this training, novice crowd workers would have been unlikely to generate useful leads in a domain where they had limited relevant knowledge. In other words: training in CrowdSolve allowed the crowd to support experts with more complex work. Similarly, in CuriOSINTy, experts provided four week-long training sessions dedicated to each of the four steps in the OSINT cycle: discovery, archival, verification, and reporting. The crowd was not only able to identify and debunk large quantities of misinformation on social media, but they were also able to do this with minimal expert intervention. In both studies, we found that this training meant that the crowd needed less explicit task design and minimally structured workflows.

However, training itself can be time-consuming. Organizers and experts must prepare training modules, which they then teach to the crowd. One way to minimize the time required is to reuse the training across investigations, where possible, as well as to allow crowd workers to view it asynchronously and hold a separate synchronous session to answer questions. Another way to minimize the amount of training required is through task design and workflows, which I discuss next.

Task Design. Task design involves determining what tasks crowd workers should perform. Often — but not always — it includes determining how to divide up complex work into smaller, more manageable pieces.

One key aspect of ELC is assigning tasks based on the complementary strengths that crowd workers and experts possess. Typically, novice and trained crowds may excel at quickly scaling up repetitive tasks [175], devising creative solutions to problems [307], as well as collecting and synthesizing disparate sources of information [205, 206, 216]. What experts tend to excel at, compared to crowds, is knowing how to perform certain tasks that require
domain knowledge expertise, knowing where to find specialized information, and knowing which tactics are effective versus ineffective [20].

Organizers and experts can specify tasks in varying levels of detail, ranging from highly structured tasks (GroundTruth), to directed questions (CrowdSolve), and general evaluation parameters or heuristics (CuriOSINTy). This follows prior work [53] that shows that crowd workers can be given some freedom in choosing which tasks to work on and how.

GroundTruth leverages experts domain knowledge to create aerial diagrams, narrow the search area, and incorporate crowd feedback into their search. GroundTruth then introduces crowds who are tasked with matching the aerial diagram to specific set of satellite imagery, and asked to match the aerial diagram with satellite imagery. Together, we found that experts and crowds succeeded nine out of ten times in geolocating the image. In three instances when crowd workers ruled out the correct location, experts were still successful in geolocating the image within ten minutes for two instances. There were also three instances where experts did not correct identify the correct location, but the crowd was able to do so in eleven minutes for all three instances. In other words, there was only one instance (out of ten) where both the crowd and expert did not pinpoint the correct location.

In CrowdSolve, the organizers chose which cases to investigate and curated the set of case files to narrow the crowd workers’ focus, while the experts created structured questions for the crowd to answer along with training and feedback. The crowd workers looked at a specific set of case files and answered a narrowed set of questions. We found that this prevented them from engaging in conspiratorial behavior or veering too far off-topic in their discussions. This also enabled them to generate useful leads for law enforcement. For example, one crowd worker noticed that a paint chip that was collected as part of evidence had not been identified yet. The crowd worker informed the experts that the Royal Canadian Mounted Police’s database of automobile paint could help them identify the make and model of car
Finally, in CuriOSINTy, the experts specified narrative threads that indicated to the crowd what topics to focus on. This narrowed the crowd’s focus to only topics that the expert planned to investigate and could act upon, as compared to searching for misinformation on an endless number of topics on social media. Experts did not — and perhaps could not — specify to the crowd exactly what misinformation they were looking for. Rather, the evaluation rubric indicated to the crowd what was important when they were searching for or debunking misinformation. For example, the rubric for discovery flags focused on potential misinformation that was more recent and had a wider audience (in terms of the number of shares or retweets). Thus, in CuriOSINTy, experts did not overly specify to the crowd what to look for or where to look, rather, they specified the characteristics of the content that they wanted the crowd to find.

More extensive task design may be needed (or desirable) if the training provided is minimal, and vice versa. Task design, like training, can exhibit a wide spectrum from minimal to extensive, but in all instances leverages the complementary strengths of experts and crowds, allowing them to do more than either could by themselves.

Workflows. Workflows are “pre-specified sets of decomposed tasks that are sequenced and integrated by computation to achieve a final goal” [265]. Crowdsourcing workflows can be powerful because they embed expert domain knowledge into software, allowing large distributed crowds to meaningfully contribute to causes [47].

However, workflows also pose challenges such as poor initial work that can derail work in other stages [206], uncoordinated contributions with inconsistent or unhelpful changes [44], and crowd workers may not have sufficient global context to contribute effectively [70, 182].

Retelny et al. [265] conducted a study comparing workflow-based teams and role-based
teams. The workflow-based teams used existing workflows that separated task dependencies (i.e., traditional crowdsourcing), while the role-based teams were given weakly-specified workflows where task and their dependencies were unmentioned, and the crowd workers were required to manage the work themselves. In their study, Retelny et al. [265] found that while workflows serve as useful artifacts for coordination, and led to teams following plans exactly as specified, but they also overly constrained adaptation and creativity. In contrast, role-based teams dynamically adapted their goals, deliverable, and work structure, but lacked structure and constraints. Retelny et al. conclude that workflows are “a double-edged sword that limits crowdsourcing from engaging in the adaptations necessary to achieve complex goals” [265].

This follows Suchman’s work which recognizes that complex goals require adaptation, versus pre-specified plans that are algorithmically embedded and managed, and struggle to account for unanticipated situations and unforeseen or emergent behavior. Based on this prior work, as well as our findings from the three studies, we do not advocate for highly structured workflows and overly specified plans in ELC investigations.

Instead, ELC follows a middle ground between workflow-based and role-based approaches, where experts and crowds take on specific roles and types of work. Crowds can also also be subdivided down into smaller teams of three to ten crowd workers, allowing for greater adaptability and coordination compared to a larger crowd of tens or hundreds of crowd workers. In this way, ELC sacrifices some automation for greater flexibility. This does not mean that automation needs to be entirely done away with, instead, I propose that automation be a means to an end, and not the end itself.

For example, in GroundTruth, experts often used automated reverse image search tools such as Google Lens, TinEye, and Yandex as a first step or filter to geolocate the image. If these reverse image search tools succeed — which they rarely do — then there would be need to
work with a crowd to geolocate the image. Still, even when they do not succeed, they may provided additional information to the expert to help narrow the search — such as what country the image may be located in. In addition, automated image tagging tools could be used to identify objects or transcribe and translate text to support the experts. While the use of tools was restricted in CrowdSolve, we found that crowd workers still used satellite imagery tools, historical weather databases, and collaboration tools such as Google Docs and Sheets. Finally, in CuriOSINTy, we leveraged informal roles such as a team leader, and provided scaffolding by dividing the work into four phases (discover, archive, verify, report). Like Retelny et al. [265], we found that teams may have benefited from more structure and occasionally veered off course, but still succeeded in taking up complex work. In all three cases, automated tools enabled the work rather than controlled and constrained it.

Duration, Synchronicity, Co-locatedness, and Communication

Duration, synchronicity, co-locatedness, and communication are four inter-related dimensions that make ELC investigations easier or more difficult to conduct.

<table>
<thead>
<tr>
<th>Type</th>
<th>CrowdSolve</th>
<th>GroundTruth</th>
<th>CuriOSINTy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Duration</strong></td>
<td>4 days</td>
<td>30 minutes</td>
<td>90 minute sessions every week for 3 months</td>
</tr>
<tr>
<td><strong>Synchronicity</strong></td>
<td>Synchronous (asynchronous possible)</td>
<td>Synchronous (asynchronous possible)</td>
<td>Synchronous (asynchronous possible)</td>
</tr>
<tr>
<td><strong>Co-locatedness</strong></td>
<td>Co-located</td>
<td>Distributed</td>
<td>Distributed</td>
</tr>
<tr>
<td><strong>Communication</strong></td>
<td>Frequent</td>
<td>None</td>
<td>Constant</td>
</tr>
</tbody>
</table>

Figure 6.8: The duration, synchronicity, co-locatedness and communication dimensions for the three studies.

**Duration** *Duration* encompasses both the length of time that an individual ELC investigation takes place, as well as its planned permanence. The length of one ELC investigation
may last 30 minutes in the case of GroundTruth, 90 minutes to weeks and multiple months in the case of CuriOSINTy, and days to decades in the case of CrowdSolve.

The duration of an ELC investigation may make it easier or more difficult to conduct. Longer duration events may require more resources, such as frequent crowd worker participation; continued access to software, hardware, and possibly a physical location; and continuous involvement from other stakeholder groups. On the other hand, shorter duration investigations may require more effort to initially set up, but less continuous effort.

Planned permanence is the amount of time that the broader set of coordinated actions are intended to persist for [195]. Planned permanence is the planned or intended permanence because it is difficult to predict how long an ELC investigation or system will eventually last, and it is not straightforward to determine at what point it has become “permanent.”

Irrespective of whether an ELC investigation is temporary or permanent, stakeholder groups must “create shared practices, artifacts, and terms” [195]. With longer planned permanence, participants may strive to create more stable artifacts and practices [195]. There may also be hybrids of temporary and more permanent coordinated actions.

It must be noted that temporary investigations are “not easier than more permanent ones” [195], but rather they are different in easier or more difficult ways. For example, spinning up a temporary expert-led crowdsourced investigation, as in the case of CrowdSolve, required considerable time and effort to set up workflows and infrastructures that were only used once (for that event). In contrast, although CuriOSINTy also required effort to design and develop, the class reused the CuriOSINTy platform several times.

While CrowdSolve as an event had low planned permanence, it required considerable effort to set up. However, it resulted in the creation of more permanent artifacts (the case analysis reports) that became part of the decades-long investigations. In contrast, GroundTruth
and CuriOSINTy had longer planned permanence as software platforms, but the artifacts that crowd workers created and their practices were more temporary.

**Synchronicity.** Johansen’s taxonomy of CSCW tools [164], in particular his time-space matrix provides a way to conceptualize coordinated action across two dimensions: synchronicity and co-locatedness. Like Lee and Paine’s MoCA framework [195], synchronicity here is a continuum ranging from being conducted synchronously or asynchronously, as well as a mixture of both [99].

*Synchronicity* in ELC investigations refers to the work done between experts and crowd workers, and between crowd workers themselves. In all three studies, experts trained and guided crowds synchronously. Experts also pre-processed much of the work asynchronously before providing it to the crowd. This included curating the case files, narrowing the search area and drawing the aerial diagram, and specifying narrative threads. Experts may also synthesize the crowd’s work asynchronously, like in CrowdSolve (the case analysis report), or synchronously like in GroundTruth and CuriOSINTy (looking at the heatmap, judging submissions).

Synchronous coordinated action may involve greater cost or complexity, but can also result in the work being completed faster or simplified coordination. Unlike many of the other dimensions presented here, ELC is predicated upon synchronous coordinated action between experts and crowds, albeit some asynchronous coordinated action may be possible. This is not only because experts may require crowd workers to perform work within a short span of time, but also because experts provide training and guidance to the crowd as they conduct this work. If experts and crowds worked asynchronously, it would take a longer time for the work to be completed — which may not be ideal in high-stakes investigations with time constraints.
Synchronicity is also related to duration, availability, and turnover. If crowd workers are required to be present for a longer duration and to work synchronously, such as in CrowdSolve or CuriOSINTy, it may be more difficult or more expensive to recruit crowd workers who may be required to take time off from their other duties or places of employment. On the other hand, if the duration is short, synchronous work may be more feasible, i.e., GroundTruth. If turnover is high, synchronous work also becomes more difficult.

Co-locatedness The second dimension from Johansen’s time-space matrix [164], is co-locatedness. It is also a continuum that is concerned with “whether coordinated actions are taking place in the same [physical] location” or at completely different physical locations [195]. Further, some types of work may be more or less affected by distance than others [247].

Similar to synchronous coordinated action, co-located coordinated action may involve greater cost or complexity — especially for when the scale of participation is large. In CrowdSolve, crowd workers spent hundreds of dollars to travel to and attend the event in Seattle, WA, from around the U.S. and as far as Europe. Whereas in GroundTruth and CrowdSolve, crowd workers could be located anywhere but still contribute to the coordinated action.

However, co-location can also have benefits, as Olson and Olson point out [247]. For instance, in CrowdSolve, experts and crowd workers felt that they could dedicate more time to the event because of the time-confined and spatially co-located nature of the event. Whereas in GroundTruth and CrowdSolve, we had to rely on other mechanisms to incentivize crowd workers to continue to participate, such as monetary compensation, grades, and a gamified environment. Still, information and communication technologies, such as Zoom and Google Drive, and crowdsourcing systems, like GroundTruth and CuriOSINTy, have helped to bridge this “distance” to make distributed coordinated action more feasible.
An additional benefit of a co-located ELC investigations is that it may be easier to control the event’s security and privacy. With CrowdSolve, the organizers were easily able to control access to the case files — they were handed out to crowd workers when they entered the conference venue, and were taken back before they left. However, with GroundTruth and CuriOSINTy, it was less feasible to spatiotemporally restrict access to case files largely because of the distributed nature of the work and the fact that any crowd worker could have easily retained a copy of information provided to them without the organizer’s or expert’s knowledge.

Co-location is related to duration, availability, and turnover in many of the same ways as synchronicity. If co-location is required for a longer duration, it may be less feasible and vice versa, depending on crowd worker availability and turnover. Co-location may also be correlated to synchronicity. The three studies discussed here are largely synchronous; CrowdSolve was co-located, while GroundTruth and CuriOSINTy were distributed.

Communication Communication refers to the amount and quality of interaction between crowd workers, as well as crowd workers and experts. Communication can be direct or indirect. Direct communication involves direct contact and exchange of information between individuals, whereas indirect communication involves exchange of information through an intermediary artifact, such as a physical document or a software tool.

Communication is directly related to agency, duration, and co-locatedness of the coordinated action. If crowd workers have limited agency in terms of how they can perform tasks, then less communication is needed because there are fewer opportunities for errors to arise. If the coordinated action is longer in duration or is co-located, there may be more opportunities for disagreement or misunderstanding — meaning that direct communication may be beneficial.

The synchronicity of a coordinated action does not directly determine whether direct or in-
direct communication may be beneficial. For example, GroundTruth relied upon real-time, synchronous interaction between experts and crowd workers, but no interaction between crowd workers. To minimize direct communication, GroundTruth computationally aggregated crowd worker feedback. However, this may not be feasible in all scenarios. In contrast, CuriOSINTy relied upon indirect communication between experts and crowd workers through: 1) expert-specified narrative threads and 2) near real-time expert evaluations of crowd workers’ submission. However, crowd workers communicated with each other both directly through a Zoom meeting room and indirectly through CuriOSINTy’s task feature.

Direct communication can be beneficial, provided that the workflows and tasks are designed in a way to ensure that crowd workers are not do not veer off-topic or spend too much time deliberating (like in CrowdSolve). To mitigate these concerns, CuriOSINTy leveraged smaller teams compared to CrowdSolve to make communication easier, instilled a greater sense of urgency through short duration events, as well as scaffolding and rubrics to structure the crowd’s work. Although GroundTruth did not permit direct communication between crowd workers, prior work has found that even distributed and anonymous crowd workers may communicate with each other [346].

6.4.4 RQ14. Sociotechnical Infrastructure

According to Star and Ruhleder [291] and Lee et al. [196], infrastructure is an “underlying framework that enables a group, organization, or society to function in certain ways.” This can refer to pipes to transport water and roads to transport people. However, both Star and Ruhleder [291], and more recently, Lee et al. [196], suggest that the term “infrastructure” extends beyond physical constructs.

Star and Ruhleder identified eight aspects of an infrastructure and noted that it is a fun-
damentally relational concept, with ambiguous and multiple meanings, properties that are emergent, and where the sociotechnical aspects are interwoven — i.e., all infrastructure is sociotechnical infrastructure. In the three prior studies, I found that an important aspect of ELC investigations is its underlying sociotechnical infrastructure. Here, I define sociotechnical infrastructure to include a combination of human, technological, physical, and financial infrastructure, as well as community-level, organizational, and government policy. Table 6.9 shows how the five types of infrastructure were instantiated in the three studies.

By foregrounding the sociotechnical infrastructure that constitutes a site of coordinated action, I highlight how it enables coordinated action to emerge or to intentionally be brought into existence. Secondarily, foregrounding the sociotechnical infrastructure allows us to examine more closely how it supports or hinders coordinated action.

<table>
<thead>
<tr>
<th>Infrastrucuture</th>
<th>Definition</th>
<th>CrowdSolve</th>
<th>Study</th>
<th>CurIOSINTy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Infrastructure</td>
<td>The people, organizations, networks, and arrangements that constitute a site as a collective entity.</td>
<td>Police detective, event organizers, group facilitator, experts, victims’ families, crowed.</td>
<td>Researchers, experts, crowds.</td>
<td>Researchers, experts, crowds.</td>
</tr>
<tr>
<td>Technological Infrastructure</td>
<td>Software tools and systems used to enable the investigation, as well as any physical computing hardware or technology that the software runs on or utilizes.</td>
<td>Limited: Facebook group, and some crowd workers used software tools without explicit permission from the organizers or experts.</td>
<td>Extensive: GroundTruth system, Legion Tools software, Amazon Mechanical Turk API, web hosting.</td>
<td>Extensive: CurIOSINTy system and tools enabled by its API, web hosting, other software to conduct investigations.</td>
</tr>
<tr>
<td>Physical Infrastructure</td>
<td>The physical location where stakeholders coordinate their actions if they are co-located.</td>
<td>Extensive: Co-located event over four days spread across three large conference rooms.</td>
<td>None.</td>
<td>None.</td>
</tr>
<tr>
<td>Financial Infrastructure</td>
<td>Any form of monetary payment to set up or conduct the investigation.</td>
<td>Event organizers invested money; crowd workers paid to attend the event.</td>
<td>Research funds for software development, and to recruit experts and crowd workers.</td>
<td>Research funds for software development. Crowd workers paid to register for the course.</td>
</tr>
<tr>
<td>Policy</td>
<td>Governmental or organizational policies and regulations that affect what work can be done, how it can be done, and any consequences for doing that work.</td>
<td>Police department policies in active vs. cold cases. NDA with legal ramifications. Explicitly set norms.</td>
<td>None internally. Amazon MTurk policies enforced external constraints.</td>
<td>University policy. Explicitly set norms.</td>
</tr>
</tbody>
</table>

Figure 6.9: Definitions of various types of sociotechnical infrastructure as well as how they existed within the three studies.
Human Infrastructure

Defining Human Infrastructure. Lee et al. [196] coined the term human infrastructure to serve as an analytical lens that magnifies the “social conditions and activities that constitute the emergence of [a] infrastructure” and define human infrastructure as the people, organizations, networks and arrangements that must be brought into alignment for large-scale distributed work to be accomplished.

The term human infrastructure suggests that teams are not the fundamental unit of analysis because individuals are members of several overlapping networks, groups, and organizations. Individuals work within and navigate these structures to accomplish work. This human infrastructure benefits from other existing infrastructures (technological, physical, financial, and policy), but also contribute to building them [196].

Benefits of Human Infrastructure as a Lens. Most importantly, though traditional organization maps and structures can partially describe large-scale coordinated action, it fails to describe the whole. These maps and structures are limited because they cannot tell us how individuals navigate permission and management issues in-situ, nor can they tell us the different working styles across different disciplines and groups or the role that longer-term, informal communication and collaboration may play within a broader site of collective action. In other words, human infrastructure “achieves collective action not by making my relationship to the whole visible but by making it invisible, indeed irrelevant ... [it] does not create a distributed team; it dissolves the very need for one” [196].

By uncovering the human infrastructure through infrastructural inversion [291], we can foreground: (1) the various decisions, labors, and coordination activities required to organize a site of collective action; (2) stakeholder groups’ diverse backgrounds, motives, and definitions of success; and (3) how each of them plays a crucial, interrelated role in the functioning of
For example, CrowdSolve did not materialize overnight. It required active, concerted effort by the different stakeholder groups for it to come into existence as a site of collective action. First, over the span of more than a decade, several detectives investigated the two murder cases. Second, the event organizers devised, set up, and managed the event; recruited additional law enforcement experts, crowd workers (attendees), and the victims’ families stakeholders. Third, the experts taught and guided the crowd, and advised the lead detective on the case; while the victims’ families motivated those present at the event. Fourth, the crowd generated useful leads for law enforcement, and finally, the professional facilitator and event organizers synthesized the crowd’s work into a detailed report. The combined, inter-related efforts of all of these stakeholder groups — that is, the human infrastructure of CrowdSolve — not only made it possible, but also made it successful.

In GroundTruth, we — as researchers — built and refined the system to enable crowds to support expert investigators with image geolocation. To evaluate GroundTruth, we then recruited experts and hired MTurk crowd workers. The experts used GroundTruth in three ways: 1) generate aerial diagram to indicate to the crowd what to look for, 2) direct the crowd’s attention by specifying where to look through the gridded map, and 3) synthesize the crowd’s feedback to arrive at a final result. Lastly, the crowd used the crowd worker interface on GroundTruth to speed up and scale up the expert’s image geolocation work practice. The human infrastructure in CuriOSINTy was largely similar to GroundTruth, although the crowd played an out-sized role compared to experts.

In all three studies, the articulation work (and the “regular” work) that each of the stakeholders undertook made the ELC investigation successful. For instance, in CrowdSolve the organizers, experts, and victims’ families supported crowd workers and kept their behavior in check, preventing unwanted vigilantism. The crowd, in turn, generated useful leads for
law enforcement and made the victims’ families feel supported. In GroundTruth, the crowd was able to narrow the search area quickly and together, experts and crowds succeeded nine out of ten times. Lastly, in CuriOSINTy, we taught the crowd how to conduct OSINT investigations. In turn, the 46 crowd workers successfully identified and debunked hundreds of pieces of potential misinformation in less than 90 minutes.

**Tensions in Human Infrastructure.** However, within any human infrastructure, there is the potential for friction, because various stakeholders have different motives that can lead to conflicting actions [196]. Tatar [306] argues that these tensions exist at the junction between what is and what should be, and in the actions taken to bridge that difference. Foregrounding these tensions in a sociotechnical system can be useful because it allows for relationships and structures to surface that may “make or break a system” [306, 321].

For example, in CrowdSolve, we found tensions between: 1) the control of the experts organizing the event and the crowd of true crime fans participating in it; 2) the conflicting goals of opening up the two cases for a crowdsourced investigation, and preserving the privacy and security of the victims and their families; and 3) the entertainment aspects of the event and the reality of the murders and their real-world consequences.

In GroundTruth, we found tensions between: 1) trusting an anonymous crowd versus providing them more information to increase their performance; 2) the time required to use GroundTruth and work with the crowd versus the time taken to geolocate the images without the crowd (for certain images); and 3) the value and the costs associated with recruiting paid crowd workers.

In CuriOSINTy, we found tensions between: 1) structuring the event for individuals motivated more by competition versus those motivated more by collaboration and 2) some crowd workers’ desire to investigate multiple topics for a short span of time versus others’ desire
investigate one topic for a longer time period.

Ultimately, human infrastructure is a valuable lens with which to examine sociotechnical infrastructures — ELC investigations in particular — because it allows us to understand not only what makes ELC work but also the various ways in which ELC can break down.

Technological Infrastructure

Technological infrastructure refers to any software tools and systems used to enable the ELC investigation, as well as any physical computing hardware or technology that the software runs on or utilizes.

Technological infrastructure not only includes crowdsourcing systems that enable experts and crowds to engage in coordinated action, but also any other software or hardware they may use to conduct these investigations.

For example, in CrowdSolve, the organizers explicitly chose not to use or build software to coordinate work. Instead, crowd workers and experts dealt with physically printed case files that were handed to them each day. However they used a Facebook group to ask experts questions outside of the co-working sessions, they also used Google Docs and Sheets (and later Microsoft Word and Excel) to synthesize their work, as well as other investigative tools such as satellite imagery and mapping software.

In contrast, GroundTruth and CuriOSINTy were crowdsourcing software systems that we developed to help coordinate and structure the crowd’s work. Both tools relied on Heroku for cloud hosting (and thus their physical servers). GroundTruth also relied upon Legion Tools [135] to generate HITs, as well as recruit and queue up workers through the Amazon Mechanical Turk API.

Depending upon the type of work that needs to be completed in the ELC investigation, it
may be possible to use no technological infrastructure. However, to take full advantage of the capabilities of ELC, and as the scale of the coordinated action increases, mechanisms to support large-scale distributed work may be needed [77]. Thus, it may be necessary to use or appropriate commercially-available off-the-shelf software systems (COTS) [97, 121], to develop custom software systems, or a combination of the two.

**Physical Infrastructure**

Here, I use the term *physical infrastructure* to refer to the physical location where stakeholders coordinate their actions if they are co-located.

If stakeholders are co-located and it is a large-scale coordinated action, it make be difficult to coordinate the efforts of individuals involved [303]. In this case, it may be useful to lean more heavily on the technological infrastructure to facilitate team formation [317] as well as coordinate team work [264, 318].

For example, in CrowdSolve, experts recruited a professional group facilitator that used an online survey tool to ask crowd workers questions. This tool then generated analyses in real time (e.g., a word cloud or bar charts) to prompt reactions and solicit further information from the crowd. After the event, the facilitator synthesized the information collected from the online survey and hand-written responses into a final case report that was provided to the experts. In contrast to CrowdSolve, crowds in GroundTruth and CuriOSINTy were distributed entirely online.

**Financial Infrastructure**

*Financial infrastructure*, though oft-overlooked, refers any form of monetary payment to set up or conduct the ELC investigation. In some instances, it may be possible through
entirely free and open source software tools (e.g., Mattermost), or through social media platforms (e.g., Twitter). Yet, even when there is no money changing hands, individuals may spend a considerable amount of time. In other instances, developing or acquiring the software; modifying and evaluating the software; as well as recruiting experts, crowds, and other affected stakeholders may involve significant monetary cost.

In CrowdSolve, we found that the event organizers invested considerable time, effort, and money to set up the event and bring together experts, crowds, and the victims’ families. Further, over 250 crowd workers paid several hundred dollars each for event registration fees, travel, and accommodation during the four day-long event. In CuriOSINTy crowd workers paid to register for the course (and thus take part in the investigations). Whereas in GroundTruth, the research team paid crowd workers to participate. In both CuriOSINTy and CrowdSolve, we also expended research funds for software development and evaluation.

Policy

*Policy* refers to governmental, organizational, or community-level policies and regulations that affect what work can be done, how it can be done, and any consequences for doing that work.

Policies can support or limit an ELC investigation. For example, in CrowdSolve, when one of the two cold cases turned into an active investigation because a suspect confessed to the crime, the police department’s policies prevented the detective from sharing information about suspects with the crowd. However, in the other still-cold case, the detective could share information about prior suspects as well as transcripts from interviews with them. The event organizers set up organizational policy in two ways: through “hard” and “soft” policy. Hard policy involved non-disclosure agreements that crowd workers were required to sign prior to
participating in the investigation that limited what information they could retain or make
publicly available. Soft policy involved explicitly setting norms that signaled to the crowd
appropriate and inappropriate behavior (e.g., contacting suspects or conspiracizing on the
Facebook group).

When crowd workers have more agency and control, or security and privacy concerns are
heightened, explicit policies and norms can help mitigate the occurrence of unwanted behav-
ior.

However, in GroundTruth, we did not need to set any explicit, internal policies because we
had a great deal of control over what information was provided to the crowd workers and
what tasks they conducted. More broadly, Amazon Mechanical Turk’s policies affected what
the GroundTruth system itself could do. For example, MTurk policies prevent requesters
from asking crowd workers for any personally identifiable information. This makes it difficult
to contact crowd workers to take part in another investigation.

### 6.4.5 RQ11 Revisited: The Articulation Work of Expert-Led Crowdsourced Investigations

It would be remiss to not discuss the component that binds together all of the other com-
ponents in ELC investigations: the invisible coordination and negotiation of work necessary
to make the coordinated action possible [50, 127, 281].

Given the importance of articulation work in *getting work done*, especially in any large-scale
coordinated action, I foreground the role it plays in ELC. Articulation work for coordinated
action is dependent upon a “common field of work,” as proposed by Schmidt and Simone
[282]. This common field of work is composed of multiple, interdependent actors who,
through their actions, change its state. Fundamental to coordinated action is interaction
though a common field of work. In this way, articulation work enables and constrains the nature of interdependent actors and their complex actions, and Schmidt and Bannon argue that it is inherent to any organization [281].

<table>
<thead>
<tr>
<th>Actor</th>
<th>CrowdSolve</th>
<th>GroundTruth</th>
<th>CuriOSINTy</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>-</td>
<td>Recruited and trained crowd workers, generated micro tasks, aggregated and visualized crowd work.</td>
<td>Scaffolded work into four steps, tracked contributions, and motivated crowd through points and leaderboard.</td>
</tr>
<tr>
<td>Organizers</td>
<td>Brought together other stakeholder groups, chose cases to investigate, curated case files, organized the event, and synthesize final report.</td>
<td>Designed and developed the system, recruited experts, paid crowd workers.</td>
<td>Designed and developed the system, recruited experts and crowd workers, and trained crowd workers.</td>
</tr>
<tr>
<td>Expert</td>
<td>Trained the crowd, framed questions for crowd to answer, and provided real-time feedback on the crowd’s work.</td>
<td>Specified location, created aerial diagram, and reviewed crowd’s aggregated work.</td>
<td>Specified narrative threads and evaluate crowd work.</td>
</tr>
<tr>
<td>Crowd</td>
<td>Paid registration fees to recruit experts and fund the event, answered experts’ questions, and generated new leads.</td>
<td>Matched aerial diagram to satellite imagery.</td>
<td>Identified, archived, verified, and reported on social media misinformation.</td>
</tr>
<tr>
<td>Affected Stakeholders</td>
<td>Victims’ family members motivated crowd workers and answered their questions about the victims.</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 6.10: The articulation work performed by the systems and stakeholder groups in each of the three studies.

**Actors Engaged in Articulation Work**

Articulation work is performed by a variety of *actors*, both human and non-human. In ELC, this includes the crowdsourcing system itself, if present, as well as the four stakeholder groups: organizers, experts, crowd workers, and affected stakeholders.

While there may be a common set of articulation work that must be performed in an ELC investigation, the same set of actors do not necessarily need to perform this work. For example, while the experts trained the crowd in CrowdSolve and CuriOSINTy, the GroundTruth system provided a training tutorial to the crowd. While the organizers provided the financial infrastructure for the investigations in GroundTruth and CuriOSINTy, the crowd did so in CrowdSolve. Table 6.10 describes these differences in more detail.
Crucial to supporting articulation work are artifacts that represent the common field of work, facilitating coordination and communication. Artifacts play a greater role when the scale of coordinated action is larger, more distributed, asynchronous, and longer in duration, as well as when the work itself is more complex, and when sociotechnical infrastructure is more limited. Artifacts may be passive — in that they are changed by others’ actions, or they may be active — in that they change how others act.

While GroundTruth and CuriOSINTy explicitly leveraged crowdsourcing systems as artifacts to coordinate work, CrowdSolve did not. In GroundTruth, the system recruited and trained crowd workers, generated micro tasks, as well as aggregated and visualized crowd work. In CuriOSINTy, the system scaffolded work into four steps, tracked contributions, and motivated the crowd through points and leaderboard.

In CrowdSolve, we found that the lack of underlying technological infrastructure increased the burden on the human infrastructure, requiring all stakeholder groups to negotiate and coordinate work amongst themselves. However, this also meant that workflows were more malleable, instead of being constrained by a less flexible system [265]. The amount of articulation work required to set up and run an ELC investigation can be significantly reduced by the adoption of crowdsourcing tools.

**Recursiveness of Articulation Work**

One actor’s work may be another actor’s articulation work and vice versa [195]. In this way, the relationship between articulation work and work is recursive and complementary.

For example, in CrowdSolve, the organizers’ work involved bringing together different stakeholder groups and setting up the event, but for other stakeholders this would be considered articulation work. In a similar way, the experts trained crowd workers and framed questions
6.5 Discussion and Conclusion

In this section, I highlight the core contributions of the ELC framework; discuss the importance of experts versus the high-stakes setting in which ELC takes place; and explain how to use the ELC framework to design new investigations or augment existing ones.

6.5.1 Contributions of the Expert-Led Crowdsourcing Framework

The ELC framework described here builds on, and incorporates, several other frameworks for coordinated action: Malone’s collective intelligence genome framework [221], Lee and Paine’s MOCA framework [195], Starbird’s crowd continuum framework [292], and Liu’s crisis crowdsourcing framework [209].

While Malone’s framework is focused on crowdsourcing microtasks; Lee and Paine’s framework is focused more on the “how” of a range of coordinated action types — including crowdsourcing initiatives. Both, however, are domain-independent. In contrast, Starbird’s and Liu’s frameworks are more narrowly focused on the high-stakes domains of crisis crowdsourcing, where crowds emerge and develop expertise over time.

The ELC framework is focused on the high-stakes domain of crowdsourced investigations, and includes three elements that merge the prior four frameworks: 1) “how” coordinated action occurs; 2) a conceptualization of micro- and macro-tasks; 3) and how ethically conducting
high-stakes crowdsourced investigations requires expert leadership and careful coordination
between different stakeholder groups, each of whom contributes in their own way.

By comparing and contrasting three projects that involved expert-led crowdsourcing, I pro-
vide an overarching framework that describes how ELC investigations work.

The ELC framework introduces three novel and beneficial dimensions: 1) training, 2) control
and agency, and 3) security and privacy. These novel dimensions help to explain how expert-
led crowdsourced investigations can be conducted both effectively and ethically.

First, experts provide explicit training and guidance to the crowd unlike prior frameworks.
This intentional decision enables the crowd to better support experts by taking on more com-
plex investigative tasks compared to more traditional microtask or emergent crowdsourcing
initiatives. Second, ELC provides explicit control over the investigation to experts versus
traditional crowdsourced investigations (both top-down and bottom-up), while still allowing
the crowd greater agency. Third, a focus on security and privacy is inherent to high-stakes
investigative domains where the cost of errors or vigilante behavior is high: lives or even
democracies may be at stake. By focusing on these three dimensions, experts can make
crowdsourced investigations not only more successful but also more ethical.

6.5.2 Role of High-Stakes Context Versus Involvement of Experts

It is important to disentangle the roles that the high-stakes context and the involvement of
experts play within the ELC framework. That is, could ELC be applicable to high-stakes
settings without experts, or would ELC work if the “expert” were not a professional?

For instance, there may be other high-stakes contexts, such as crisis mapping [209, 292],
that involve crowds of varying levels of expertise [48]. Could ELC be applicable in crisis
mapping? Perhaps. However, two of its three dimensions mentioned above would not be
applicable, namely: control and agency, and security and privacy. In crisis mapping, control is distributed among various crowd workers. Security and privacy are also rarely a concern because crisis mapping relies upon open source intelligence (i.e., publicly available sources [209]).

Alternatively, would ELC work if the “expert” were not a professional? Again, the same two dimensions would not be applicable. Control not only involves directing the actions of others, but also being held responsible for those actions. The U.S. military follows a similar maxim: “A [leader] can delegate authority but not responsibility. Authority refers to who is in charge, while responsibility refers to who is accountable” [101]. In ELC investigations, professionals, such as journalists or law enforcement agents, work in the public eye and can be held accountable for their actions through professional sanctions and legal ramifications. In contrast, many emergent (bottom-up) crowdsourced investigations suffer from a lack of accountability when they engage in vigilante behavior because there are no (immediately apparent) leaders who can be held accountable [184, 194, 243, 316].

It is clear from the name that experts lead (control) the crowdsourced investigation, and are also held responsible for any outcomes (good or bad) that may arise from the investigation. Thus, it is a combination of both the high-stakes setting and the involvement of professionals who can be held accountable that make expert-led crowdsourcing uniquely beneficial and challenging.

While experts may be the ones who control the investigations, the researchers and developers of the software systems (that experts and crowds use) also play a significant role in enabling both the successes and failures of ELC investigations. What level of control and agency should researchers assume? If researchers created ELC software tools without expert involvement, experts may not find the tools useful, and they may result in more harm than good. For example, if tools like CrowdTangle and GroundTruth were easily accessible to a
novice crowd, it might exacerbate vigilantism such as that associated with Reddit’s investigation of Sunil Tripathi [218] or the “human flesh search engines” that exist in China [66]. In addition, these tools may be abused by oppressive or authoritarian governments. Though these risks may not be possible to eliminate entirely, by expecting experts and researchers to work together, the risks can be mitigated.

The findings from the three studies explored in this dissertation — along with my own direct involvement in developing software for two of them — also indicate that ELC investigations are improved when experts and researchers work together. Experts provide domain knowledge and investigative expertise in knowing where help is needed, while researchers provide scientific knowledge and technical expertise in knowing how to effectively help. Thus experts should engage in participatory design or co-design with experts and other stakeholders to develop tools to support ELC investigations. Experts retain control over and accountability for the investigations, but researchers retain control over and accountability for the software tools.

6.5.3 Using the Expert-Led Crowdsourcing Framework

Having described how ELC works through the framework, I now explain how to apply the framework to design new investigations or augment existing ones. This involves two parts: setting up the investigation and training experts to work with crowds.

Designing ELC Investigations

The first two questions to be answered in designing an ELC investigation are: “Who designs it?” and ”Who participates in it?” The answer to these could be two different groups of people or the same.
In the three case studies explored in this dissertation, the designers of the ELC and those who participate in it were two distinct groups. In CrowdSolve, the event organizers set up the event and brought together the other stakeholder groups. Similarly, in GroundTruth and CrowdSolve, I — as part of a research team — set up the investigation and recruited experts and crowd workers. Still, even though participants did not directly design the investigation, their involvement and participation (e.g., through participatory design [240]) can be beneficial in determining its success.

Alternatively, the same group of people could design and participate in the the ELC. For example, the Sedition Hunters community on Twitter [263, 323] is run by a small group of investigators who not set up the investigation and create software tools, but also direct the crowd’s efforts, respond to the crowd’s feedback, and further investigate their and submissions.

In both cases, designing an ELC involves identifying each dimension of the framework — such as determining:

- What types of investigations to support based on automatability, tractability, feasibility, representability.
- Who the stakeholder groups, and their scale, motivations, availability, and skills.
- The available sociotechnical infrastructure, e.g., technological, physical, policy, and financial.

Once these three categories have been determined, the organizers also have to determine the nature of the coordinated action, including determining the potential security and privacy concerns associated with this type of work. For example, if the work involves “hot-button” or partisan issues, those involved may be targeted because of their marginalized identities or
the type of institutions in which they work [100]. In these situations, the coordinated action may need to be more co-located and synchronous, the security measures made more strict, and the use of technology reduced (such as in CrowdSolve). In addition, the crowd may need more training in how to communicate and conduct investigations, and experts may need to exert more control over the investigation while also taking on greater accountability.

Further, the duration of the coordinated action should be determined, as well as whether it is possible to sustain it for a longer period, based on need, the availability of funds, and stakeholder interest.

Lastly, the workflows and task design would need to be created. If the crowd were fully committed for a period of time to the investigation, it may be easier to assign them to portions of the investigation. Whereas, if the crowd were only partially committed (in terms of availability), it may be better to rely on micro-contributions [328] and solicit volunteers [336].

It is important to note that not all of these factors should — or even can — be created all at once. Instead, it is important to expect, and enable emergence [278] and appropriation [98, 132]. For instance, workflows may be emergent in more loosely formed coordinated actions. Similarly, communication norms and security and privacy practices may evolve over time through a process of iteration and testing or intentional and rapid prototyping [91].

Training Experts to Work With Crowds

Similar to how leaders of an organization learn how to manage their employees, if ELC investigations are to be successful, it is important for experts to learn how to manage crowds. In this way, ELC investigations may function as flash teams or organizations [264, 318]. This leadership may be top-down or distributed [214]; and managed through social norms [201]
or through software [273]. Future work should explore how to support experts in leading crowds or distributing leadership in an ELC setting.

Apart from leading ELC investigations, experts would also benefit from learning how to generate workflows and tasks for crowd workers based on their skills and availability. Systems like GroundTruth and CuriOSINTy already provide short tutorials to onboard experts. However, researchers should also study how to design more detailed courses that teach experts to perform these tasks. In addition, experts might benefit from frameworks that could allow experts to easily specify the nature of the work, the crowds’ features, and suggest how to generate and assign smaller tasks. In this way, experts who lead ELC investigations are made not born: they can develop these skill sets through deliberate practice [29].

To promote responsible use of powerful ELC tools that support coordinated action, tool designers could require experts to pass certain certifications or tests in ethics, leadership, and workflow or task generation before experts can use these tools.

### 6.5.4 Conclusion

In this chapter, I highlighted the applicability of existing crowdsourcing frameworks [106, 209, 221, 292] in the context of expert-crowd collaboration in high-stakes settings. Second, I summarized the findings and implications from my three prior studies. Third, I presented a conceptual framework for expert-led crowdsourced investigations that compares and contrasts these three studies, and contend that expert-led crowdsourcing allows experts and crowds to do more than either could alone.
Chapter 7

Discussion and Future Work

7.1 Enriching Expert–Crowd Collaboration in Investigations

I propose three areas of future work inspired by our experiences with CrowdSolve, GroundTruth, and CuriOSINTy that suggest how learning and adaptation between experts and crowds can enable more productive collaboration on other types of investigative and analytic tasks.

7.1.1 Enrich Expert–Crowd Interaction

Above, I reviewed prior work on requester–crowd interaction models, suggesting that innovations in real-time crowdsourcing have enabled richer interactions between requesters and crowds. These interactions can be imagined along a spectrum of increased requester participation. One class of projects employs a hand-off model where requesters specify some initial criteria and then a crowd rapidly completes the tasks largely on its own [45, 186]. A second class has a requester interact sporadically with the crowd, providing facilitation and guidance as required, either with the meta-workflow [53, 223] or the task itself [44, 62, 188]. A third class, which I called expert-led crowdsourcing, has experts (i.e., requesters with valuable task-specific skills) working alongside the crowd and performing a superset of the crowd’s tasks, with crowd results streaming in and influencing the expert’s own work [31, 197, 198].
7.1. Enriching Expert–Crowd Collaboration in Investigations

Figure 7.1: Expert-led crowdsourced investigations differ from traditional bottom-up and top-down investigations. There is richer interaction between experts and crowds: experts train and guide crowds, who support the expert by collecting and analyzing data. In contrast, the top-down model exhibits limited interaction between experts and crowds, instead relying on a one-way flow of information. The bottom-up model often exhibits no direct interaction between experts and crowds.

While GroundTruth embodied the crowd-augmented expert work model within the domain of visual search, the flow of information between experts and crowds was nevertheless highly constrained. Experts defined a search area and diagram for crowds, while crowds returned search results to experts for review. A two-way flow of information might allow experts and crowds better adapt to each other’s progress, increasing efficiency and accuracy, albeit with higher collaboration costs. Prior work such as Crowdboard’s virtual sticky notes [31], Apparition’s synchronized canvas [198], and structured communication for writing tasks [274] suggest domain-specific mechanisms for richer expert-led crowdsourcing that could be adapted for analytic tasks like visual search.

CuriOSINTy introduced a more interactive, two-way flow of information between experts and crowds. The expert specified topics to investigate and provided feedback to the crowds.
in real-time, while the crowd followed the experts’ guidance and reacted to their feedback — modifying the targets of their investigations and the content that they submitted. Richer two-way communication between experts and crowds could better harness their complementary strengths. Workers might contribute insights and voice concerns about an investigation, while experts bring leadership, professionalism, and investigative expertise to guide the crowd.

Taking this idea further, a growing body of research explores how crowd collectives could address power imbalances between requesters and workers on crowdsourcing platforms [140, 272, 335]. On the other hand, crowdsourced investigations lacking expert oversight are often problematic [194, 343].

### 7.1.2 Help Crowds Learn Valuable Skills

Heer’s idea of incremental learning aligns with recent crowdsourcing scholarship on supporting more complex and creative tasks, and providing more meaningful work experiences for crowd workers beyond the scope of the current task. One thread of research has explored providing just-in-time learning for novice workers to gain task-specific skills [215, 250, 347], with recent work suggesting that crowds can gain more generalized skills and knowledge within the context of microtasks [140, 329]. Crowd workers can also learn from previous workers through a shared memory space [136, 188] — a type of shared representation between crowd workers across instances — that ensures that new crowd workers can quickly adapt to a particular expert’s diagrams and working style. Another thread of research seeks to offer novice crowds career ladders and paths to more financially and intrinsically rewarding work [305]. As in the case of CrowdSolve, enriching expert–crowd interaction requires expanding the typical dichotomy of experts and novice crowds, creating intermediate roles
7.2. Simultaneously Providing Greater Agency and Structure

that workers can claim as they learn skills and gain knowledge that helps them participate both in the task at hand, and the broader labor market.

7.1.3 Engage Diverse Crowds

A focus on crowd learning also directs attention to the composition and diversity of crowds in expert–crowd collaboration. Most real-time crowdsourcing research hires paid crowd workers, who are assumed to be novices and largely transient beyond one task session [273]. My work here was no exception, though our evaluation revealed that many experts preferred other types of crowds. Some experts worked as freelancers or in cash-strapped organizations where budgeting regular crowdsourcing payments seemed infeasible. Others had access to volunteer novice crowd labor, either through institutional programs or social media followers, that mitigated the need for payment. Still others proposed using their colleagues as crowd workers, drawing on their expertise and mutual trust to speed up searches and improve accuracy. Beyond this diversity of incentives motivating crowd work, a diversity of demographics, experiences, and geographic locations may streamline investigations by allowing experts to solicit specialized knowledge from the crowd.

7.2 Simultaneously Providing Greater Agency and Structure

In CrowdSolve and CuriOSINTy, there were areas where greater structure within teams could lead to greater efficiency gains. For instance, in CuriOSINTy, some teams organically devised assembly line workflows and team leaders emerged over time. We observed that these teams frequently placed high in the leaderboard. Another set of teams employed
“free for all” workflows with minimal collaboration among team members and no explicit team leader. In both CrowdSolve and CuriOSINTy, we found that the “free for all” teams did not perform as well. Our findings suggest that both types of teams may benefit from more explicit structure and roles [265], such as delineating the responsibilities for each team leader and assigning roles to each team member. For example, the leader could mitigate unwanted redundancy by assigning team members works on a specific topic or social media platform. To prevent judges from being overwhelmed by work, the leader could also conduct a preliminary evaluation of their flags before forwarding it to the judge. To further increase efficiency, individuals could be assigned or encouraged to focus on tasks that they preferred or excelled at, such as content discovery versus verification.

7.2.1 Give the Crowd (More) Agency

In traditional crowdsourcing systems, complex tasks are divided into microtasks that crowd workers complete independently, with little to no interaction with each other or agency in how to complete these tasks. CuriOSINTy builds on a growing body of literature — including GroundTruth — that shows that crowds can perform more complex tasks — provided that they are sufficiently motivated, as well as given adequate scaffolding, training, and agency.

In the previous chapters, we found that CrowdSolve, GroundTruth, and CuriOSINTy structured crowd workers’ work so that they could more easily perform complex investigative tasks. Further, CuriOSINTy was flexible enough that crowd workers were able to investigate a range of topics.

Future work should explore ways to design frameworks and tools that enable experts to coordinate the efforts of diverse crowds with minimal prior experience working with crowds. These tools and frameworks could enable any expert to become a “crowdsourcer” [81]. For
instance, how might an expert, who has five investigative targets, direct the best subsets of crowd workers to investigate them? How might the crowd’s feedback be aggregated and shown to the expert in real-time for a range of tasks, from image geolocation, to text and network analysis, and beyond?

Providing more agency can lead to a virtuous cycle

In GroundTruth, crowd workers were given much less agency compared to CrowdSolve and CuriOSINTy. Our survey feedback revealed that crowd workers were enthusiastic to take up the task and wanted to contribute more — both by doing the same task for different regions, but also by providing more feedback to the expert on the aerial diagram. In the other two studies, we found that providing more agency to the crowd led to a virtuous cycle.

In CrowdSolve, attendees generated new leads and provided experts with additional resources that they were not previously aware of, such as additional databases involving the weather, soil conditions, and automobile paint types. If the organizers had structured attendees tasks in a more rigid way, such as in traditional crowdsourcing approaches, they may not have provided these suggestions. This aligns with previous work that shows that rigid workflows constrain the crowd’s creativity and adaptability [265].

We also found that CuriOSINTy helped students learn to more critically examine information online and develop a mental model for conducting investigations. This proved to be a virtuous cycle: between the first and fifth event, students submitted 65% more evidence pieces and 163% more flags, with only a 1.3% reduction in flag approvals from judges. In addition, students said that they enjoyed using CuriOSINTy — possibly motivating them to continue participating in the events.
Enable crowd workers’ diverse motivations to be met

In Coase’s Penguin [42], Benkler introduces the concept of “commons-based peer production” where groups collaborate on large-scale projects by making contributions of varying sizes, but where each individual may have diverse motives (rather than purely based on economic conditions or directions from a manager).

I find similar diversity in motivations for stakeholders within the three studies in this dissertation, but in the context of high-stakes investigations vs. free and open-source software development in Benkler’s case [42]. While it may be difficult to design crowdsourcing systems that meet crowd workers’ diverse goals, these systems may be more successful if they are able to do so, similar to other peer production systems that exist today.

For example, CrowdSolve and GroundTruth might have been less effective in alternative contexts. If CrowdSolve focused on having attendees complete microtasks, minimized interaction with each other and with the experts and victims’ families, attendees might not have been as motivated to participate. Similarly, if GroundTruth leveraged a monetary incentive scheme with citizen scientists or online volunteer communities, those volunteers might be less motivated to use GroundTruth.

In CuriOSINTy we were able to observe the effects of designing to motivate one group versus another. We found that not all students were motivated by competition, perhaps because we did not allow students to choose between a predominantly competitive or collaborative setting, as would be typical outside of the classroom.

In future work, researchers should explore how to dynamically modify the affordances of a system and event to meet crowd workers’ preferences, such as enabling self-competition [232] and intra-team collaboration. For example, the crowd could complete a survey that assesses their individual personality traits to prioritize certain aspects over others [266].
If most crowd workers largely preferred working collaboratively across teams, competitive aspects could be downplayed in the interface design and vice versa. For instance, in CuriOSINTy, the ratio of points for collaboration versus competition could be changed. If most crowd workers exhibited a preference for self-improvement, self-competition with individualized badges (for accomplishing certain milestones) could be enabled, while also prompting experts to provide more individualized feedback to experts.

Lastly, if there was no clear consensus in preferences among the crowd, each crowd worker could be shown slightly different interfaces with certain elements emphasized over others, depending on their preferences. Those who preferred competition could be shown the leaderboard and their team’s score on every page; while those who prioritized collaboration could be shown an alternative leaderboard that emphasized how much they had collaborated.

7.3 Incorporate AI-Infused Systems

The three projects discussed in this dissertation focused on aspects of experts’ investigations where automated approaches were insufficient on their own. However, there are several areas where artificial intelligence and machine learning (henceforth, ‘AI’) can be infused into these systems and into expert-led crowdsourcing, more broadly.

For example, GroundTruth supported professional investigators with image geolocation once the search area had been narrowed by an expert or computer vision-based (CV) tools [147, 326]. Future work could consider ways to incorporate CV-based tools into the crowd’s or expert’s workflow. For instance, the expert could use a CV tool as a first-pass to generate an aerial diagram, while further refining it on their own. Another CV tool could help the crowd rule out areas that were unlikely to be a match (e.g., comparing a building to a parking lot or farm land).
In CrowdSolve, much of the work was intentionally “low-tech” to maintain security and privacy. Still, the organizers could have used walled-off computers that allowed crowd workers to more easily search through case files using optical character recognition. Matching algorithms could also be used to survey the crowd for their skill sets and preferred working styles to form more efficient teams [275].

Finally, in CuriOSINTy, crowds employed (and built) several AI-infused tools. For example, one team built a tool that leveraged existing APIs to detect whether an Twitter account was a bot or not; while other teams used community detection algorithms to identify different communities on social media. Future work should explore ways to help experts dynamically generate relevant keywords for the crowd to search; while enabling the crowd to more quickly identify potentially false or misleading claims using natural language processing techniques [301].

More broadly, ELC systems could be infused with AI to manage interactions between experts and crowds. This includes enabling experts to creating micro- and macro-tasks for the crowd; supporting crowd workers’ with their work; and aggregating the crowds’ feedback in a more easily-digestible manner for experts.

However, infusing AI into ELC systems requires a careful understanding of both AI systems, their behavior, and their limitations; as well as whether to design human-AI systems [342] and how to do so through guidelines for human–AI interaction [27, 149].

For example, there are certain tasks that humans excel at, such as devising creative solutions or combining disparate sources of data to arrive at a theory or conclusion [219]. In this case, AI should be used to augment human creativity. For example, in CuriOSINTy, an AI tool could be used to suggest tools to experts or the crowd, but allow them to choose which tools to use to investigate claims online.
AI excels at detecting patterns and finding similarities [326], and thus AI can be used to surface these patterns and similarities to humans who then decide what course of action to take. For example, an initial step in the GroundTruth pipeline could be to run the satellite imagery through a computer vision tool [326] to provide initial matches, but then rely on the crowd to eliminate false positives, or to search through a narrower area once certain regions or countries had been ruled out by the CV tool.

7.4 Consider Long-Term Software Maintenance and Support

One challenge common to all sociotechnical endeavors — including software projects — is long-term maintainability and support [88, 312]. The issue of (non-)maintenance in sociotechnical systems is so severe that Vinsel suggests that regulatory enforcement can help enforce maintenance (and repair) of sociotechnical systems, but may be in tension with the rapid pace of technological innovation [325].

For example, the Google Maps API version that GroundTruth uses will soon be deprecated, and would require significant work to re-implement many of its features that rely upon the Google Maps API. In CuriOSINTy, we addressed this problem by designed it with appropriation and maintainability in mind to maximize the ways in which it can be used in the future [88]. We also plan to open source the code for the CuriOSINTy platform so that others may continue to improve it [42].

Even CrowdSolve, that did not directly rely on any software tools, required some form of long-term community support and maintenance. We observed on the CrowdSolve Facebook group that attendees would frequently ask for new updates or conduct further research, while
the victims’ families also requested assistance to help further the two investigations. Experts, however, have limited time and resources, and may not be able to support this community in the long-term. Future expert-led crowdsourcing initiatives should consider not only the short-term support required, but also any long-term support necessary to make these initiatives more successful and worthwhile for all stakeholder groups. This includes maintaining the community of practice that may have developed intentionally or unintentionally through the ELC investigation [75, 330].

Finally, it would be remiss to not mention the ethical concerns with making powerful tools openly available. Perhaps bad actors might use GroundTruth or CuriOSINTy to track and harass political opponents, or authoritarian governments might use it to identify and imprison dissidents. We believe that bad actors will (and already do) engage in this type of behavior irrespective of the availability of such tools. However, by demonstrating how GroundTruth and CuriOSINTy can be used for prosocial behavior, we believe it can help empower marginalized communities.
Chapter 8

Conclusion

In this chapter, I provide a conclusion for my dissertation. I first return to the research questions that I posted in Chapter 1, showing how the studies presented in my dissertation addresses each research question. I then state the major contributions offered by these studies.

8.1 Contributions

The contribution of this work is the introduction and explication of expert-led crowdsourcing as a new paradigm for conducting high-stakes investigations with professional investigators and novice crowds. This dissertation describes how to manage the challenges in ELC investigations through regulation mechanisms, as well as how to augment expert-crowd collaboration through the design and development of two ELC systems. This dissertation then examines how ELC works across the three studies presented here (Ch. 3, 4, 5), comparing and contrasting them across four factors, each comprised of several dimensions.

Throughout, this research has focused on both the social and technical elements that comprise ELC investigations, as well as the interplay between the two. This dissertation makes the following contributions:
8.1.1 Introduced the Concept of Expert-Led Crowdsourcing

In Chapter 3, I introduced the concept of expert-led crowdsourcing for high-stakes investigations. This work examines how ELC differs from existing crowdsourcing approaches focused on a traditional framing of requester-crowd interaction, expanding the scope and type of work that crowdsourcing can address. Identifying areas where traditional crowdsourcing falls short, ELC shows how the complementary skills of experts and crowds can be synergistically leveraged to do more than either could alone.

This work also uncovered the significant human infrastructure and articulation work required to make ELC investigations not only possible, but also successful. Organizers brought together the various stakeholder groups and set up the crowdsourcing initiative or developed the crowdsourcing system. Experts provided real-time training and guidance to the crowd, while also specifying tasks for crowd workers to complete, reviewing their work, and utilizing it in their investigations. The crowd sped up and scaled up experts’ work practice, generated novel and useful solutions, without engaging in harmful or vigilante behavior found in typical crowdsourcing initiatives. Finally, affected stakeholders motivated the crowd to participate and brought a personal element to the investigation.

8.1.2 Identified How to Manage the Sociotechnical Challenges in ELC Investigations

This work examined the varying motivations and definitions of success for each stakeholder group in Chapters 3 and 5. This work also highlighted the tensions in the design space, between experts and crowds, security and privacy versus openness, and entertainment versus reality. By highlighting these tensions, as well as how they are managed through four regulatory mechanisms — laws, norms, markets, and architecture — this dissertation points to
how and why ELC investigations succeed while minimizing harm to others and vigilantism.

8.1.3 Developed ELC Systems to Support a Range of Complex Investigative Tasks

GroundTruth (Ch. 4) demonstrated how a crowd could successfully support an expert with one complex investigative task: image geolocation. Designing GroundTruth required a careful consideration of not only what types of tasks crowds could support, but also how to represent expert work practice in a way that is understandable to novices. Shared representations — in the form of an aerial diagram, gridded satellite imagery, and heat map — enabled experts and crowds to collaborate in real-time, while minimizing coordination and communication challenges.

CuriOSINTy (Ch. 5) extended GroundTruth’s capabilities, allowing a crowd to support a range of complex investigative tasks, from discovery and archival of content, to verification and reporting. CuriOSINTy also leveraged a different motivation mechanism — gamification, compared to monetary incentives in GroundTruth and social motivation in CrowdSolve. It also showed that capture the flag competitions, that are traditionally theoretical, could be applied to a real-world setting: combating social media misinformation. Further, CuriOSINTy demonstrated that beneficial elements of collaboration could be introduced into a CTF competition to minimize the challenges of competition. Namely, CuriOSINTy minimized duplication of effort and information silos between competing teams.
8.1.4 Examined How Expert-Led Crowdsourcing Works to Develop an ELC Framework

Finally, in Chapter 6, I examined how expert-led crowdsourcing works and developed a framework to highlight its five constituent components, each with several dimensions of their own. This chapter highlights how existing crowdsourcing frameworks are insufficient to explain how ELC investigations work. Then, the chapter explicates its five components: the domain, the stakeholder groups, the sociotechnical infrastructure, the nature of coordinated action, and the articulation work involved.

I find that ELC investigations may require considerable effort to set up and conduct. However, this dissertation posits that by involving diverse stakeholder groups, from experts and crowds, to organizers and affected stakeholders, as well as by leveraging their complementary skills and motivations, the whole can be greater than the sum of its parts.

I then introduce the ELC framework that not only explains how the three ELC investigations studied in this dissertation work, but posit how other investigations may work, and suggests design recommendations for those who seek to conduct future ELC investigations.

Recommendations include determining the type of tasks, choosing and recruiting stakeholder groups, motivating stakeholders and leveraging their complementary skills, considering existing and developing new sociotechnical infrastructure, and fine-tuning every element of the coordinated action itself.

Some elements that should be fine-tuned are: determining whether the ELC investigation should be a traditional crowdsourcing setup, a competition, or a collaboration; as well as considering control and agency depending upon security and privacy requirements; and determining the duration, synchronicity, and co-locatedness of the investigation.
The ELC framework also introduces three under-explored and beneficial dimensions: 1) training, 2) control and agency, and 3) security and privacy. By focusing on these three dimensions, expert-led crowdsourcing allows experts and crowds to do more, and to do it more ethically, than either could alone.
Bibliography


[18] WhatsApp can be a black box of misinformation, but Maldita may have opened a window, June 2021. URL https://www.poynter.org/fact-checking/2021/whatsapp-can-be-a-black-box-of-misinformation-but-maldita-may-have-opened-a-window.


[27] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Colisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan,


[46] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom


[53] Jonathan Bragg, Mausam, and Daniel S. Weld. Sprout: Crowd-powered task design for


[57] Benjamin Carlisle, Michael Reininger, Dylan Fox, and Daniel Votipka. On the Other Side of the Table: Hosting Capture-the-Flag (CTF) Competitions An Investigation from the CTF Organizer’s Perspective. page 8.


[68] Peter Chapman and David Brumley. picoctf: Teaching 10,000 high school students to hack, 2013.


[86] Thomas Debeauvais, Bonnie A Nardi, Cristina V Lopes, Nicholas Yee, and Nicolas Ducheneaut. 10,000 gold for 20 dollars: An exploratory study of world of warcraft


[100] Periwinkle Doerfler, Andrea Forte, Emilio De Cristofaro, Gianluca Stringhini, Jeremy Blackburn, and Damon McCoy. "i’m a professor, which isn’t usually a dangerous job": Internet-facilitated harassment and its impact on researchers. Proc.


[174] Sangmi Kim and Lionel P Robert Jr. Crowdsourcing coordination: A review and re-


[192] Edith Law, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams.


[197] Sang Won Lee, Yujin Zhang, Isabelle Wong, Yiwei Yang, Stephanie D. O’Keefe, and Walter S. Lasecki. Sketchexpress: Remixing animations for more effective crowd-


Publisher: National Academy of Sciences Section: Social Sciences.


[291] Susan Leigh Star and Karen Ruhleder. Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In *Pro-


[318] Melissa A Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S Bernstein. Flash organizations: Crowdsourcing complex work by struc-


