User Experiences with Data-Intensive Bioinformatics Resources: A Distributed Cognition Perspective

Jongsoon Park

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
Industrial and Systems Engineering

Joseph L Gabbard, Chair
Tonya L. Smith-Jackson
Christopher North
Nathan Ka Ching Lau

May 6, 2015
Blacksburg, VA

Keywords: user experience, bioinformatics, distributed cognition, mixed methods approach

User Experiences with Data-Intensive Bioinformatics Resources: A Distributed Cognition Perspective

Jongsoon Park

ABSTRACT

Advances in science and computing technology have accelerated the development and dissemination of a wide range of big data platforms such as bioinformatics into the biomedical and life sciences environments. Bioinformatics brings the promise of enabling life scientists to easily and effectively access large and complex data sets in new ways, thus promoting scientific discoveries by for example generating, validating, and refining hypotheses based on *in silico* analysis (performed on computer). Meanwhile, life scientists still face challenges in working with big data sets such as difficulties in data extraction and analyses arising from distributed and heterogeneous databases, user interface inconsistencies and discrepancies in results. Moreover, the interdisciplinary nature of modern science adds to significant gaps in scientists' performance caused by limited proficiency levels with bioinformatics resources and a lack of common language across different disciplines.

Although developers of bioinformatics platforms are slowly beginning to move away from function-oriented software engineering approaches and towards to user-centered design approaches, they rarely consider users' value, and expectations that embrace different user contexts. Further, there is an absence of research that specifically aims to support the broad range of users from multiple fields of study, including 'wet' (lab-based) and dry' (computational) research communities.

Therefore, the ultimate goal of this research is to investigate life scientists' user experiences with knowledge resources and derive design implications for delivering consistent user experiences across different user classes in order to better support data-intensive research communities. To achieve this research goal, we used the theory of distributed cognition as a framework for representing the dynamic interactions among end users and knowledge resources within computer-supported and -mediated

environments. To be specific, this research focused on how online bioinformatics resources can be improved in order to both mitigate performance differences among the diverse user classes and better support distributed cognitive activities in data-intensive interdisciplinary research environments. This research consists of three parts: (1) understanding user experience levels with current bioinformatics resources and key determinants to encourage distributed cognitive activities, especially knowledge networking, (2) gaining in-depth understanding of scientists' insight generation behavior and human performance associated with individual differences (i.e., research roles and cognitive styles), and (3) identifying in-context usefulness, and barriers to make better use of bioinformatics resources in real working research contexts and derive design considerations to satisfactorily support positive user experiences. To achieve our research goals, we used a mixed-methods research approach that combines both quantitative (Study 1 and 2) and qualitative (Study 3) methods.

First, as a baseline for subsequent studies, we conducted an empirical survey to examine 1) user experience levels with current bioinformatics resources, 2) important criteria to adequately support user requirements, 3) levels of knowledge networking (i.e., knowledge sharing and use) and relationship to users' larger set of distributed cognitive activities, and, 4) key barriers and enablers of knowledge networking. We collected responses from 179 scientists and our findings revealed that lack of integration, inconsistent results and user interfaces across bioinformatics resources, and perceived steep learning curves are current limitations to productive user experiences. Performance-related factors such as speed and responsiveness of resources and ease of use ranked relatively high as important criteria for bioinformatics resources. Our research also confirmed that source credibility, fear of getting scooped, and certain motivation factors (i.e., reciprocal benefit, reputation, and altruism) have an influence on scientists' intention to engage in distributed cognitive activities.

Second, we conducted a laboratory experiment with a sample of 16 scientists in the broad area of bench and application sciences. We elicited 1) behavior characteristics, 2) insight characteristics, 3) gaze characteristics, and 4) human errors in relation to individual differences (i.e., research roles such as bench and application scientists, cognitive styles

such as field-independent and dependent people) to identify whether human performance gaps exist. Our results (1) confirmed significant differences with respect to insight generation behavior and human performance depending on research roles, and (2) identified some relationships between scientists' cognitive styles and human performance.

Third, we collected a rich set of qualitative data from 6 scientists using a longitudinal diary study and a focus group session. The specific objective of this study was to identify in-context usefulness and barriers to using knowledge resources in a real work context to subsequently derive focused design implications. For this work, we examined 1) the types of distributed cognitive activities participants performed, 2) the challenges and alternative actions they faced, 3) important criteria that influenced tasks, and 4) values to support distributed cognitive activities. Based on the empirical findings of this study, we suggest design considerations to support scientists' distributed cognitive activities from user experience perspectives.

Overall, this research provides insights and implications for user interface design in order to support data-intensive interdisciplinary communities. Given the importance of today's knowledge-based interdisciplinary society, our findings can also serve as an impetus for accelerating a collaborative culture of scientific discovery in online biomedical and life science research communities. The findings can contribute to the design of online bioinformatics resources to support diverse groups of professionals from different disciplinary backgrounds. Consequently, the implications of these findings can help user experience professionals and system developers working in biomedical and life sciences who seek ways to better support research communities from user experience perspectives.

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Dr. Joseph Gabbard. I cannot forget your generous help and continued support provided during the entire course of my doctoral research.

I extend my thanks to committee members, Dr. Tonya L. Smith-Jackson, Dr. Chris North, and Dr. Nathan Lau for their time, insightful questions, and valuable comments.

I would also like to thank my parents and little brother for their love and continued encouragement. Thanks for always believing in me and praying for me.

A special thanks goes to Donna Boyer, Human Factors Engineer, Intel Corporation for her for facilitating me to achieve my goal. She is such a wonderful mentor and role model.

Thanks also are due to the Virginia Bioinformatics Institute for funding a significant portion of this for three years.

I am also thankful to my friends I've made in Blacksburg, Dr. Rahul Soangra, Ari Goldberg, and all others. Thank you for being such wonderful, supportive friends. Thanks for always laughing with me!

Thank you God for sustaining me, assuring me, for blessing me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Chapter 1: Introduction

1.1 Background

1.1.1 Exponential growth of bioinformatics resources

In the past two decades, we have seen an exponential increase in the size and breadth of available scientific data, demanding new integrated solutions to empower scientists to explore and elicit valuable insights quickly and accurately (Kelling et al., 2009; Park & Gabbard, 2013). As such, there are growing interests in building computing infrastructure for data integration, simulation, visualization, and validation in big data environments. The promise of these computing-based solutions has in turn led to a rapid paradigm shift in biological sciences (Bell, Hey, & Szalay, 2009). Indeed, to date, the number of heterogeneous bioinformatics resources has grown in accordance with the deluge in genomic data.

According to the Nucleic Acid Research 2015 Web Server issue, 56 new databases have been released and 115 databases have been updated within the last year (Galperin, Rigden, & Fernández-Suárez, 2015). Meanwhile, a primary concern is that scientists are not aware of available information and online resources in order to address their specific research problems. While scientists wish to find and use appropriate resources and tools to fulfill their needs, it is not straightforward because there are a large number of bioinformatics resources and tools, many of which have similar functions (J. C. Bartlett, Ishimura, & Kloda, 2012). Inconsistent results (e.g., differences of gene naming conventions and different annotations for the same gene), different interfaces of disparate resources and tools also contribute to the time needed to extract meaningful insights (Aniba & Thompson, 2010; McLeod & Burger, 2008; Roos, 2001; Tran, Dubay, Gorman, & Hersh, 2004). To support a better process of data-intensive research, we need to address the challenging issues in bioinformatics such as high technical demands, long learning curves, and lack of awareness of existing resources (Dillon, 1981; Dubay, Gorman, & Hersh, 2004; Palmer, 1991; Rolinson, Meadows, & Smith, 1995).

1.1.2 Iterative and complex research processes

Biology is “an empirical rather than theoretical science” (Stevens, Goble, & Bechhofer, 2000)

requiring scientists to explore, validate, and confirm insights from varying resources while employing different stages of information processing (e.g., stimulus perception, cognition and decision making). Typically, research processes tend not to be linear and formalistic, depending on accessible data, research goals, and prior knowledge (B. Mirel, 2007). As more and more biological research becomes specialized (e.g., intricate, sophisticated, and occur in data-rich contexts); (Lynch, 2009), scientists struggle with a number of iterations of complex processes typically required to prove an initial hypothesis and expect coherent outcomes from different steps along the way (J. Bartlett & Neugebauer, 2005; J. C. Bartlett & Toms, 2005; B. Mirel, 2009; Ouzounis, 2000; Tran et al., 2004). D. De Roure and Goble (2009) illustrate scientists' practice in biomedical and life sciences by stating: "Researchers do not work with just one content type and moreover their data is not in just one place – it is distributed and sometimes quite messy too." Regardless of research disciplines, it is inevitable to expect a high mental demand (and consequently high cognitive load) not only for gathering valuable insights but also for making decisions in complex circumstances. Thus, complex and iterative discovery processes are one of the hindrance factors to scientists' cognitive capability in the context of data-intensive research.

1.1.3 Interdisciplinary research challenges

Bioinformatics is recognized as a prime example of interdisciplinary research in a data-rich context (Aniba & Thompson, 2010; Tarczy-Hornoch & Minie, 2005). *Interdisciplinary collaboration* in biological science domains is defined by Romano, Giugno, and Pulvirenti (2011) as groups of professionals who have different types of backgrounds through the use of methods and technologies from mathematics, statics, computer science, physics and, of course, biology and medicine. As expected, various group members often not only define the problem but also collaboratively determine the approach from different perspectives. Researchers further often have different perspectives, research motivations, interests, and behaviors (Aniba & Thompson, 2010; Letondal & Mackay, 2004; Tadmor & Tidor, 2005). For instance, *biologists* tend to be interested in studying a specific gene, organism or biological process of interest, and are more motivated by publishing their research outcomes in prestigious scientific journals than others in biological sciences (Letondal & Mackay, 2004; Tadmor & Tidor, 2005). *Bioinformations* are interested in both biological questions and the development of new tools

needed to improve their particular analytical processes of interest (Letondal & Mackay, 2004). However, bioinfromaticians sometimes define the problem very narrowly based on a view of personal experiences. This view leads to making tools that just fulfilled system requirements in order to serve system functions, but no consideration of current and potential user requirements. Aniba and Thompson (2010) characterize differences between biologists and computational biologists. *Biologists* generally consider bioinformatics resources as computational methods and tools to easily and effectively handle large amounts of diverse and complex data. However, *computational biologists* view bioinformatics as a direct application area for addressing theoretical and experimental questions. de Matos et al. (2013) point out that 'wet' (lab-based) and dry' (computational) research communities make use of the same software resources to answer very different questions. In sum, it is clear that there are diverse needs and bridges are needed to support common ground between diverse user classes in interdisciplinary research practices (Pickett, Burch, & Grove, 1999).

Currently, however, there is no agreed definition to distinguish scientific communities from 'wet' (lab-based) and dry' (computational) research disciplines, but there is instead a variety of different meanings associated with it. Thus, in this dissertation, we use the term *bench scientist* and *application scientist* to distinguish characteristics of research communities of the established disciplines. The term bench scientist will be used solely when referring to scientists who mainly run *in vivo* (performing an experiment in a controlled environment outside of a living organism) or *in vitro* experiments (experimentation using a living organism) in laboratory settings, and benefit from *in silico* (performed on computer) experimentation. The term application scientist will be used in its broadest sense to refer to all scientists who have backgrounds in life sciences and are, 1) focused on the implementation of developing and validating computational models and methods using empirical data and/or 2) are involved in laboratory experiments.

1.1.4 Technology-driven bioinformatics resources

Advances in technology have become one of central drivers to support data collection, manipulation, organization, and visualization in data-intensive research domains (Kelling et al., 2009). However, advanced technology alone cannot lead to better scientific discovery processes and outcomes (Neumann, 2005). Despite of many new bioinformatics resources and online data integration portals, bioinformatics has been an almost unexplored field of user-centered design.

Thus, usability and user experience issues are often overlooked, and as such these online resources often do not sufficiently support the actual practices of end users (Bolchini, 2009; Javahery, 2004; Barbara Mirel & Wright, 2009). In the past years, a few studies have begun to focus on user-centered design methods such as interviews, persona developments, usability testing and others to improve bioinformatics resources (de Matos et al., 2013; Pavelin et al., 2012). Even with improved interfaces, scientists will still need to learn, understand and master sophisticated analysis processes and procedures much of which can be supported via distributed cognitive activities across internal's internal process, external artifacts, and the environment (Hollan, 2000). Unless end-user groups can be more broadly supported in meaningful ways, diverse bioinformatics resources will continue to demand tremendous cognitive load (Park & Gabbard, 2013, 2014).

1.1.5 The emergence of knowledge networking platforms in biological sciences

As the Internet has become indispensable in daily human life (Wyld, 2008), there is a consensus among scientists that Web 2.0-enabled social networking platforms increasingly have the potential to support biomedical and life science communities (David De Roure, Goble, & Stevens, 2009; Li, 2012; McIntosh et al., 2012; Parnell, 2011). Even though social bonds may not preexist in the real world, scientists can communicate, collaborate, and exchange knowledge with each other in ways that complement current standalone tools and ultimately accelerate scientific discoveries (Fischer & Zigmond, 2010; Gruber, 2008; Tenopir et al., 2011). Knowledge networking, which can be defined as "a special case of social networks in which the links of the network represent shared or related knowledge" (Jones, 2001), can also reduce scientists' cognitive burdens associated with uncertainty and data complexity by affording insights, alternative views, or disproof from other experts with similar research interests (Neumann & Prusak, 2007; Ward, Schmieder, Highnam, & Mittelman, 2013). In this dissertation, we enlarge the notion of knowledge networking and apply it to the study of distributed cognitive activities in a computer-mediated environment: data, information, and knowledge sharing and reuse using Web 2.0 technologies. Several studies have revealed that knowledge networking assists end-user performance, especially in interdisciplinary scientific communities composed of people with different knowledge backgrounds and skills (Aniba & Thompson, 2010; Tarczy-Hornoch & Minie, 2005). In this regard, online knowledge networking can be considered as one

(of the set) of significant activities used in today's scientific practices.

1.1.6 Distributed cognition and implications for bioinformatics

Originally, *human cognition* was narrowly defined as "a single individual's internal processes" (Nilsson, Laere, Susi, & Ziemke, 2012) that occurs only inside the head and is context-independent. However, this classical concept of cognition has led to difficulties in accounting for complexity of behavior and cognitive processes within socio-technical systems (Mansour, 2009; Simon, 1996; Suchman, 1987).

Distributed cognition is a theoretical framework to describe cognitive processes emerging from mutual interactions of humans and artifacts over time (Hutchins, 1995). Namely, an individual's cognitive resources (e.g., the researcher's memory, knowledge, or skills) can be extended beyond the boundary of the individual to involve people and external artifacts (e.g., computer systems and social environment) (Hollan, 2000) (Figure 1).

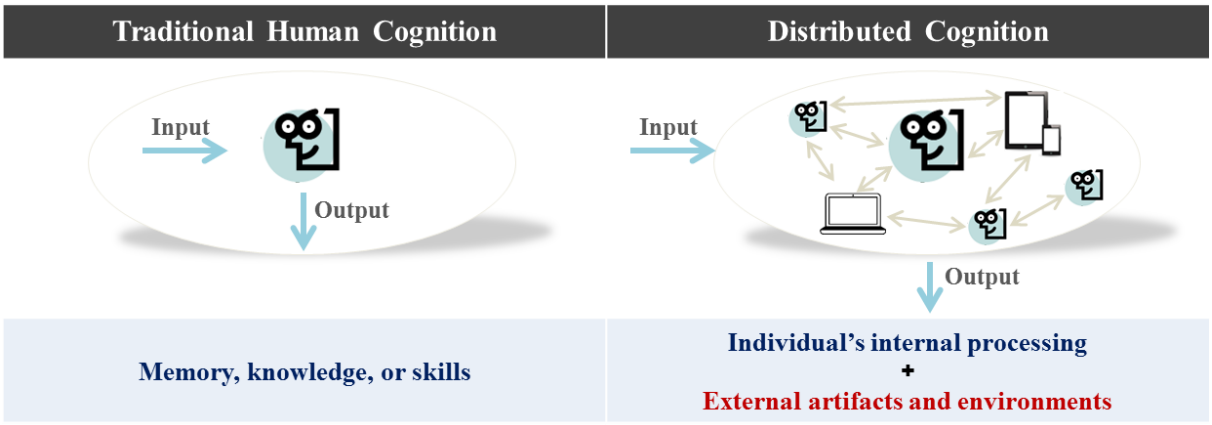


Figure 1. Distributed Cognition.
From a traditional cognitive science view (a) the cognitive system is confined to the inside of the individual's head, while in a distributed cognition view (b) cognitive processes are distributed across people and artifacts within the larger unit of a socio-technical system. Reproduced from "Information fusion in practice: A distributed cognition perspective on the active role of users," by Nilsson, Laere et al, 2012. Information Fusion, 13(1), p.65. Copyright 2011 by Elsevier B.V.

In addition, distributed cognition enables researchers to 1) analyze working environments and 2) adopt different units of analysis to describe a range of cognitive systems (Hutchins, 1995).

As mentioned earlier, bioinformatics experiments demand many sorts of cognitive resources to solve research questions. Cognitive resources could include internal resources (i.e., memory, attention, executive function) as well as external cognitive artifacts provided by bioinformatics tools and technologies. In this sense, the distributed cognition theory can provide a theoretical basis for a deeper understanding how scientists extend their cognitive capabilities to attain research goals under highly complex and ambiguous research circumstances.

Many studies that illustrate distributed cognitive activities refer primarily to interactions with people and external artifacts regardless of whether the human is online or offline (Cohen, Blatter, Almeida, Shortliffe, & Patel, 2006; Furniss & Blandford, 2006; Hansen, Robinson, & Lyytinen, 2012; Hutchins & Klausen, 1996; Nersessian, 2002; Xu & Clarke, 2012). In this research, we did not take up distributed cognitive activities in the offline physical world because it is beyond the scope of our research goals. Instead, we restrict the scope of distributed cognitive activities to include only those that constitute research-related activities towards goals within a computer-mediated environment (Figure 2).

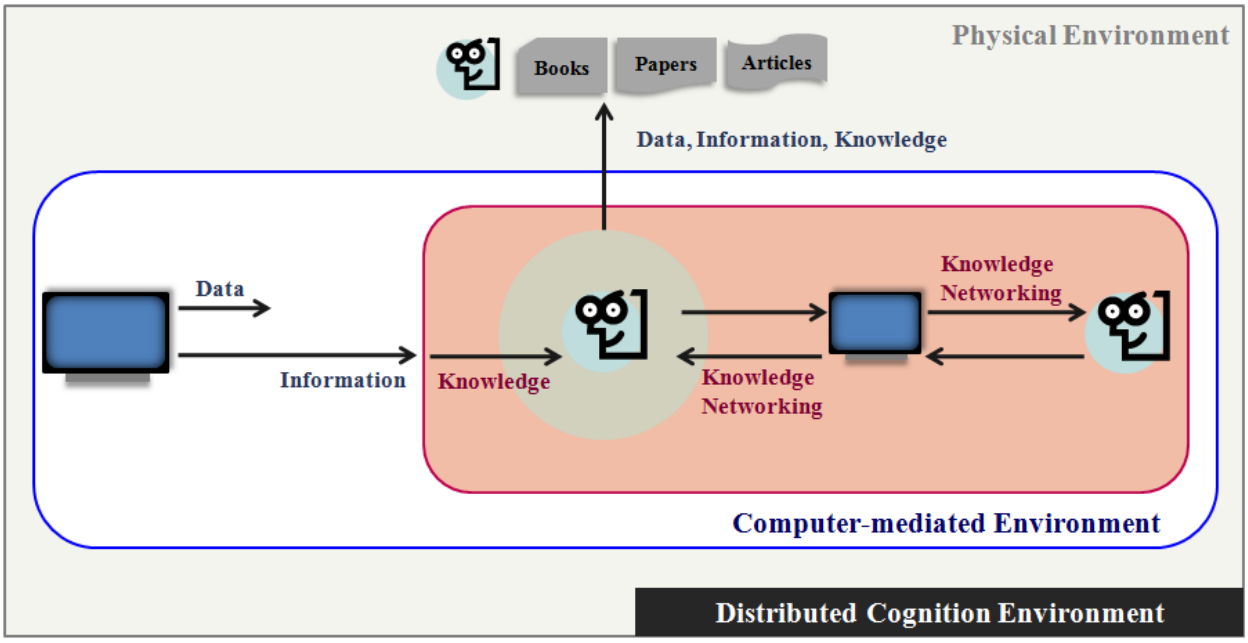


Figure 2. Research Scope.

Within this dissertation, we also use the term *knowledge resource* to refer any types of data, information, and even systems in distributed cognition environments that can be a basis of

scientists' knowledge to be required in order to reach a certain level of research goals. We employ social exchange theory (SET) (Blau, 1964) to scientific communities in order to identify key factors affecting knowledge networking intentions in data-intensive scientific communities.

In terms of methodological approaches, early examples of research into distributed cognition adopted an ethnographic approach as one type of qualitative approach to evaluate the effects of user interface design (Hutchins & Klausen, 1996; Nersessian, 2002; Rinkus et al., 2005). Although ethnographies can account for context of behaviors among multifaceted dimensions of human interactions, there are several challenges with ethnographic approaches, from data collection to data analysis such as intrusiveness, observer bias, or observer drift (Creswell & Clark, 2007). Furthermore, the ethnographic data describing specific events is highly relied on retrospective interpretation so that the value of research is dependent on researchers' ability to demonstrate the credibility of their findings (LeCompte & Goetz, 1982).

This dissertation research differed from traditional approaches to focus mainly on qualitative and ethnographic data as it was oriented toward a mixed methods approach by collecting and analyzing both qualitative and quantitative data (Creswell, 2013). A mixed method approach provides quantitative strengths that mitigate some weaknesses of the qualitative research approach. And furthermore, the mixed methods approach provides a more complete and comprehensive insight than either quantitative or qualitative approaches alone (Creswell & Clark, 2007). Thus, this study employed a mixed-method approach for a better understanding of the research problem in biomedical and life sciences.

1.1.7 Previous user studies for bioinformatics

As bioinformatics databases and tools became more complex, there was a movement to better understand user workflows and usability (Anderson, Ash, & Tarczy-Hornoch, 2007; J. C. Bartlett & Toms, 2005; Bolchini, 2009; Javahery, 2004; B. Mirel, 2007, 2009; Tran et al., 2004). Javahery (2004) addressed usability issues by conducting ethnographic interviews and usability studies with two groups (i.e., expert and novice) and confirmed that "novice users find the learning curve for such sites (e.g., online bioinformatics resources) steeper than they would prefer." Tran et al. (2004) performed task analyses in genomics and proteomics in order to potentially identify common system features and useful bioinformatics tools. Four major

categories emerged such as gene analysis, protein analysis, biostatistical analysis, and literature searching. Their findings also indicated several challenging issues for the bioinformatics community: lack of procedural documentation, use of home-grown strategies to accomplish goals, diverse individual needs and preferences, and lack of awareness of existing bioinformatics tools. J. C. Bartlett and Toms (2005) interviewed 20 bioinformatics experts and modeled biological data analysis processes. This resulted in a standard research protocol which includes a series of 16 steps and each of which specified the type of analysis, how and why each analysis is conducted, the tools used, the data input and output, and the interpretation of the results. Anderson et al. (2007) explored the impact of a commercial bioinformatics software tool on users' empirical workflows. They carried out a 7-month longitudinal qualitative study (including informal semi-structured interviews, in-lab participant observations, and direct observational shadowing) and identified reasons why the tool was underutilized. Specifically, they noted that (1) satisfaction and acceptance of tools tended to be role and goal specific, (2) the system was seen primarily as a measurement system rather than a "total laboratory analysis system", and (3) lab meetings deemphasized the system, preferring more traditional data analysis techniques. To inform the design of specific web-based tools (i.e., MiMI, MiMI-Cytoscape, and SAGA for subgraph matching), B. Mirel (2009) conducted qualitative field studies (i.e., observations and interviews) to create narrative and procedural scenarios and to collect verbatim think-aloud comments and questions. Bolchini (2009) performed usability studies to identify and characterize usability issues associated with online bioinformatics resources. The first study included an overall inspection process to identify obvious usability issues (especially navigation and information architecture) of a bioinformatics website (CATH—Protein Structure Classification—http://cathwww.biochem.ucl.ac.uk). The second study concentrated on analyzing usability problems with a common search task using three repositories (i.e., BioCarta (www.biocarta.com), Swiss Prot (www.expasy.ch/sprot), and NCBI (www.ncbi.nlm.nih.gov).

In recent years, a few authors have begun to focus on user-centered design (UCD) approaches for developing bioinformatics resources (de Matos et al., 2013; Pavelin et al., 2012). Pavelin et al. (2012) carried out two projects which aimed to redesign an online bioinformatics repository (EMBL-EBI) and to develop an online enzyme information portal, respectively. They conducted in-depth interviews and user workshops to determine how end users retrieve information they need from the Internet, user needs about their site, and subsequently iteratively designed and

evaluated user interface prototypes. The most recent study by de Matos et al. (2013) presents a case study to develop new software services at the Enzyme Portal (http://www.ebi.ac.uk/enzymeportal). They applied UCD techniques, including: persona development, interviews, 'canvas sort' card sorting, user workflows, and iterative usability testing using paper and interactive prototypes. Implications include not only key findings from interaction with users, but also the benefits of UCD approaches.

Although many new online bioinformatics resources and data integration portals have been developed recently, the studies presented suggest a slow shift in design philosophy; from feature-oriented to user-oriented. In addition, many published findings have not considered the broad spectrum of life scientists' practical activities and rarely sought to support different user classes. Contemporary research activities consist of a range of interactions with different type of resources, technologies and other professionals from different disciplines. Thus, we argue that more research is needed to enrich user experiences in data-intensive interdisciplinary research communities.

1.2 Objectives

This review of the literature has found that prior studies did not address life scientists' user experiences in terms of two aspects. First, distributed cognitive activities are critical for supporting data-intensive scientific communities, but have not been researched in any comprehensive way. Second, even though bioinformatics is a highly interdisciplinary and rapidly evolving field, the performance aspects (e.g., human errors under high cognitive load conditions) of different user classes have not yet been investigated.

For these important reasons, the overarching goal of this research is to ensure user experiences among different user classes of life scientists within distributed cognitive environments. We aimed to provide actionable insights to user experience professionals in complex interdisciplinary data-rich domains, especially biomedical and life sciences. The overall approach of this work offers a way of investigating user experience in depth, and within real-life contexts, lending to high external validity of the findings. This research deployed a mixed-methods approach that combines both quantitative (Study 1 and 2) and qualitative (Study 3) methods, as shown in Figure 3.

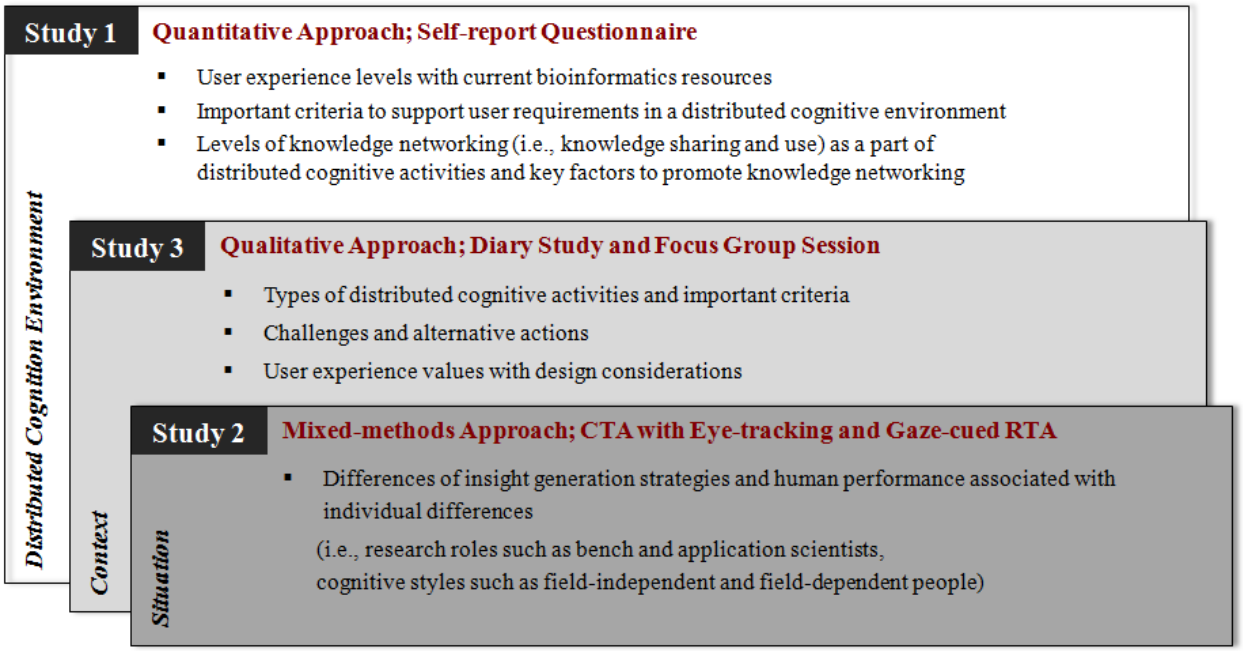


Figure 3. Research Overview.
This research employed three studies that combined both quantitative and qualitative methods to ultimately propose design considerations to support user experiences of data-intensive interdisciplinary research communities in a distributed cognition environment.

We decided to begin with a descriptive and exploratory approach. In Study 1, we administered a questionnaire to 179 users to ascertain aspects of user experience levels with current online bioinformatics resources and to identify key determinants to encourage distributed cognitive activities, especially online knowledge networking, within big data environments.

The aspects of user experience in a data-intensive research environment included: 1) user experience levels with current bioinformatics resources, 2) important criteria to support user requirements, 3) levels of knowledge networking (i.e., knowledge sharing and use) as a part of distributed cognitive activities, and 4) the barriers and enablers of distributed cognitive activities. We developed a quantitative model that describes the relative importance of key barriers and enablers of distributed cognitive activities and tested initial hypotheses using Partial Least Squares (PLSs).

In Study 2, we conducted a laboratory experiment with a sample of 16 scientists in the broad area of basic and applied life sciences. The main purpose was to gain understanding of insight

generation behavior and human performance in an experimental situation associated with individual differences. Specifically, we elicited 1) behavior characteristics, 2) insight characteristics, 3) gaze characteristics, and 4) human errors in relation to individual differences (i.e., research roles such as bench and application scientists, cognitive styles such as field-independent and dependent people).

In Study 3, we collected a rich set of qualitative data from 6 scientists using a longitudinal diary study method and subsequent focus group session. The specific objective of this study was to identify in-context usefulness of knowledge resources, and barriers to using online knowledge resources in real work contexts and derive design considerations to support user experiences within distributed cognitive environments. We examined the 1) types of distributed cognitive activities scientists performed, 2) challenges and alternative actions they faced, 3) important criteria that influence each task, and 4) design considerations for the biomedical and life science communities.

By integrating findings from the three studies, we suggested key design considerations to ensure user experiences to diverse user groups from different disciplinary backgrounds

The following shows the main research questions associated with each study.

Study 1.

- To what extent do current bioinformatics resources meet user requirements and which are the most critical requirements unmet by current systems?
- To what extent do scientists engage in distributed cognition activities?
- What are the key factors for scientists to engage in distributed cognition activities and what are the potential opportunities?

Study 2.

- What are the impacts of individual differences (i.e., research disciplines and cognitive styles) on insight generation behavior?
- When and what kinds of human errors occur, and are there individual differences?

Study 3.

- When and why do scientists use different resources in a real work context?
- Are there limitations of current systems that hinder a user's experience?

- How can online bioinformatics resources be reworked to ensure user experiences for data-intensive interdisciplinary research environments?

1.3 References

Anderson, Nicholas R., Ash, Joan S., & Tarczy-Hornoch, Peter. (2007). A qualitative study of the implementation of a bioinformatics tool in a biological research laboratory. *International Journal of Medical Informatics, 76*(11–12), 821-828. doi: 10.1016/j.ijmedinf.2006.09.022

Aniba, Mohamed Radhouene, & Thompson, Julie D. (2010). *Knowledge Based Expert Systems in Bioinformatics*.

Bartlett, J., & Neugebauer, T. (2005). Supporting information tasks with user-centred system design: The development of an interface supporting bioinformatics analysis. *Canadian journal of information and library science, 29*(4), 486-487.

Bartlett, Joan C., Ishimura, Yusuke, & Kloda, Lorie A. (2012). *Scientists' preferences for bioinformatics tools: the selection of information retrieval systems*. Paper presented at the Proceedings of the 4th Information Interaction in Context Symposium, Nijmegen, The Netherlands.

Bartlett, Joan C., & Toms, Elaine G. (2005). Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach. *Journal of the American Society for Information Science and Technology, 56*(5), 469-482. doi: 10.1002/asi.20136

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science, 323*(5919), 1297-1298.

Blau, Peter Michael. (1964). *Exchange and power in social life*: Transaction Publishers.

Bolchini, D. (2009). Better bioinformatics through usability analysis. *Bioinformatics (Oxford, England), 25*(3), 406-412. doi: 10.1093/bioinformatics/btn633

Cohen, Trevor, Blatter, Brett, Almeida, Carlos, Shortliffe, Edward, & Patel, Vimla. (2006). A cognitive blueprint of collaboration in context: Distributed cognition in the psychiatric emergency department. *Artificial intelligence in medicine, 37*(2), 73-83.

Creswell, John W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*: Sage publications.

Creswell, John W, & Clark, Vicki L Plano. (2007). *Designing and conducting mixed methods research*: Wiley Online Library.

de Matos, Paula, Cham, Jennifer A, Cao, Hong, Alcántara, Rafael, Rowland, Francis, Lopez, Rodrigo, & Steinbeck, Christoph. (2013). The Enzyme Portal: a case study in applying user-centred design methods in bioinformatics. *BMC bioinformatics, 14*(1), 103.

De Roure, D., & Goble, C. (2009). myExperiment: A Web 2.0 Virtual Research Environment for Research using Computation and Services.

De Roure, David, Goble, Carole, & Stevens, Robert. (2009). The design and realisation of the Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems, 25*(5), 561-567. doi: http://dx.doi.org/10.1016/j.future.2008.06.010

Dillon, Martin. (1981). Serving the Information Needs of Scientific Research. *Special Libraries, 72*(3), 215-223.

Dubay, Christopher, Gorman, Paul, & Hersh, William. (2004). *Applying task analysis to describe and facilitate bioinformatics tasks.* Paper presented at the MEDINFO: Proceedings of the... World Conference on Medical Informatics.

Fischer, Beth A, & Zigmond, Michael J. (2010). The essential nature of sharing in science. *Science and engineering ethics, 16*(4), 783-799.

Furniss, Dominic, & Blandford, Ann. (2006). Understanding emergency medical dispatch in terms of distributed cognition: a case study. *Ergonomics, 49*(12-13), 1174-1203.

Galperin, Michael Y, Rigden, Daniel J, & Fernández-Suárez, Xosé M. (2015). The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. *Nucleic acids research, 43*(D1), D1-D5.

Gruber, Tom. (2008). Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web semantics, 6*(1), 4-13. doi: 10.1016/j.websem.2007.11.011

Hansen, Sean W, Robinson, William N, & Lyytinen, Kalle J. (2012). *Computing requirements: Cognitive approaches to distributed requirements engineering.* Paper presented at the System Science (HICSS), 2012 45th Hawaii International Conference on.

Hollan, James. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM transactions on computer-human interaction, 7*(2), 174-196.

Hutchins, Edwin. (1995). *Cognition in the wild*. Cambridge, Mass.: MIT Press.

Hutchins, Edwin, & Klausen, Tove. (1996). Distributed cognition in an airline cockpit. *Cognition and communication at work*, 15-34.

Javahery, Homa. (2004). Beyond power making bioinformatics tools user-centered. *Communications of the ACM, 47*(11), 58.

Jones, Patricia M. (2001). Collaborative knowledge management, social networks, and organizational learning. *Systems, Social and Internationalization Design Aspects of Human-Computer Interaction, 2*, 306-309.

Kelling, S., Hochachka, W.M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: a new paradigm for biodiversity studies. *BioScience, 59*(7), 613-620.

LeCompte, Margaret D, & Goetz, Judith Preissle. (1982). Problems of reliability and validity in ethnographic research. *Review of educational research, 52*(1), 31-60.

Letondal, C., & Mackay, W.E. (2004). *Participatory programming and the scope of mutual responsibility: balancing scientific, design and software commitment.* Paper presented at the Proceedings of the eighth conference on Participatory design: Artful integration: interweaving media, materials and practices-Volume 1.

Li, J. W. (2012). SEQanswers: an open access community for collaboratively decoding genomes. *Bioinformatics (Oxford, England), 28*(9), 1272-1273.

Lynch, C. (2009). Jim Gray's fourth paradigm and the construction of the scientific record. *The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond*, 177-183.

Mansour, Osama. (2009). *Group Intelligence: a distributed cognition perspective.* Paper presented at the Intelligent Networking and Collaborative Systems, 2009. INCOS'09. International Conference on.

McIntosh, Brenley K., Renfro, Daniel P., Knapp, Gwendowlyn S., Lairikyengbam, Chanchala R., Liles, Nathan M., Niu, Lili, . . . Hu, James C. (2012). EcoliWiki: a wiki-based community resource for Escherichia coli. *Nucleic Acids Research, 40*(D1), D1270-D1277. doi: 10.1093/nar/gkr880

McLeod, Kenneth, & Burger, Albert. (2008). Towards the use of argumentation in bioinformatics: a gene expression case study. *Bioinformatics, 24*(13), i304-i312. doi: 10.1093/bioinformatics/btn157

Mirel, B. (2007, 1-3 Oct. 2007). *Usability and Usefulness in Bioinformatics: Evaluating a Tool for Querying and Analyzing Protein Interactions Based on Scientists' Actual Research Questions.* Paper presented at the Professional Communication Conference, 2007. IPCC 2007. IEEE International.

Mirel, B. (2009). Supporting cognition in systems biology analysis: findings on users' processes and design implications. *J Biomed Discov Collab, 4*, 2. doi: 10.1186/1747-5333-4-2

Mirel, Barbara, & Wright, Zach. (2009). Heuristic evaluations of bioinformatics tools: a development case *Human-Computer Interaction. New Trends* (pp. 329-338): Springer.

Nersessian, Nancy J. (2002). The cognitive basis of model-based reasoning in science. *The cognitive basis of science*, 133-153.

Neumann, Eric. (2005). A Life Science Semantic Web: Are We There Yet? *Sci. STKE, 2005*(283), pe22-. doi: 10.1126/stke.2832005pe22

Neumann, Eric, & Prusak, Larry. (2007). Knowledge networks in the age of the Semantic Web. *Briefings in bioinformatics, 8*(3), 141-149.

Nilsson, Maria, Laere, Joeri van, Susi, Tarja, & Ziemke, Tom. (2012). Information fusion in practice: A distributed cognition perspective on the active role of users. *Information Fusion, 13*(1), 60-78. doi: http://dx.doi.org/10.1016/j.inffus.2011.01.005

Ouzounis, Christos. (2000). Two or three myths about bioinformatics. *Bioinformatics, 16*(3), 187-189. doi: 10.1093/bioinformatics/16.3.187

Palmer, Judith. (1991). Scientists and information: II. Personal factors in information behaviour. *Journal of documentation, 47*(3), 254-275.

Park, Jongsoon, & Gabbard, Joseph L. (2013). An Exploratory Study to Understand Knowledge-Sharing in Data-Intensive Science *Human-Computer Interaction. Users and Contexts of Use* (pp. 217-226): Springer.

Park, Jongsoon, & Gabbard, Joseph L. (2014). *User Experiences with Open Access Knowledge Sharing Platforms Preliminary User-Centered Design Implications for Complex Data-intensive Domains.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Parnell, L. D. (2011). BioStar: An Online Question & Answer Resource for the Bioinformatics Community. *PLoS computational biology, 7*(10), e1002216.

Pavelin, Katrina, Cham, Jennifer A, de Matos, Paula, Brooksbank, Cath, Cameron, Graham, & Steinbeck, Christoph. (2012). Bioinformatics meets user-centred design: a perspective. *PLoS computational biology, 8*(7), e1002554.

Pickett, S. T. A., Burch, William R., Jr., & Grove, J. Morgan. (1999). Interdisciplinary Research: Maintaining the Constructive Impulse in a Culture of Criticism. *Ecosystems, 2*(4), 302-307. doi: 10.2307/3659023

Rinkus, Susan, Walji, Muhammad, Johnson-Throop, Kathy A, Malin, Jane T, Turley, James P, Smith, Jack W, & Zhang, Jiajie. (2005). Human-centered design of a distributed knowledge management system. *Journal of Biomedical Informatics, 38*(1), 4-17.

Rolinson, J, Meadows, AJ, & Smith, H. (1995). Use of information technology by biological researchers. *Journal of information science, 21*(2), 133-139.

Romano, Paolo, Giugno, Rosalba, & Pulvirenti, Alfredo. (2011). Tools and collaborative environments for bioinformatics research. *Briefings in Bioinformatics*. doi: 10.1093/bib/bbr055

Roos, D.S. (2001). Bioinformatics--trying to swim in a sea of data. *Science, 291*(5507), 1260-1261.

Simon, Herbert Alexander. (1996). *The sciences of the artificial*: MIT press.

Stevens, Robert, Goble, Carole A., & Bechhofer, Sean. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics, 1*(4), 398-414. doi: 10.1093/bib/1.4.398

Suchman, Lucille Alice. (1987). *Plans and situated actions: the problem of human-machine communication*: Cambridge university press.

Tadmor, Brigitta, & Tidor, Bruce. (2005). Interdisciplinary research and education at the biology–engineering–computer science interface: a perspective. *Drug Discovery Today, 10*(17), 1183-1189. doi: http://dx.doi.org/10.1016/S1359-6446(05)03540-3

Tarczy-Hornoch, Peter, & Minie, Mark. (2005). Bioinformatics Challenges and Opportunities

Medical Informatics. In H. Chen, S. S. Fuller, C. Friedman & W. Hersh (Eds.), (Vol. 8, pp. 63-94): Springer US.

Tenopir, Carol, Allard, Suzie, Douglass, Kimberly, Aydinoglu, Arsev Umur, Wu, Lei, Read, Eleanor, . . . Frame, Mike. (2011). Data sharing by scientists: practices and perceptions. *PloS one, 6*(6), e21101.

Tran, D., Dubay, C., Gorman, P., & Hersh, W. (2004). Applying task analysis to describe and facilitate bioinformatics tasks. *Stud Health Technol Inform, 107*(Pt 2), 818-822.

Ward, R. Matthew, Schmieder, Robert, Highnam, Gareth, & Mittelman, David. (2013). Big data challenges and opportunities in high-throughput sequencing. *Systems Biomedicine, 1*(1), 23-28.

Wyld, David C. (2008). Management 2.0: a primer on blogging for executives. *Management Research News, 31*(6), 448-483.

Xu, Lihua, & Clarke, David. (2012). What does distributed cognition tell us about student learning of science? *Research in science education, 42*(3), 491-510.

2 Chapter 2: Understanding user experience levels with current bioinformatics resources and key determinants to facilitate distributed cognitive activities in data-intensive research contexts

2.1 Introduction

Technological advancements have increased opportunities for rapid transition from *in vitro* ("within a glass") and/or *in vivo* ("within a living organism") laboratory experiments to an integrated approach within *silico* ("on or via the computer"). As part of the "Big Data Technology" emergence, diverse bioinformatics resources (e.g. online resources that integrate biological data and analysis tools) are being developed to support organization, integration , simulation, visualization, validation of huge and complex genomic data sets (e.g., DNA, RNA, and protein sequences, expressions, structures), as well as literature retrieval (Huerta, Downing, Haseltine, Seto, & Liu, 2000). Bioinformatics opens a wide array of applications and opportunities in multiple research fields including, but not limited to, human health, veterinary medicine, agriculture, and environmental sciences. As a result, scientists are working with a multitude of different types of data resources more than ever before (Ribes & Lee, 2010).

We can see both opportunities and challenges are arising from emerging technologies which allow scientists to access and combine tremendous volumes of biological data (Kelling et al., 2009). On the positive side, bioinformatics resources support scientists with better and faster means to design experiments to generate and validate hypotheses, and analyze complex data sets as compared to traditional methods (Bull, Ward, & Goodfellow, 2000; Katoh, 2002; Yarfitz, 2000). Meanwhile, the bioinformatics field has begun to realize end users' challenges in searching, analyzing and using the growing available data sets as well as selecting proper resources to improve biological research processes. Inconsistent results and different interfaces of disparate resources and tools also lead to burdensome and time-consuming efforts to extract insights (Aniba & Thompson, 2010; McLeod & Burger, 2008; Roos, 2001; Tran, Dubay, Gorman, & Hersh, 2004).

As life sciences become more data-centric and complex, scientists aggregate and synthesize

information from disparate knowledge resources (See Section 1.1.5 for the definition of this term) to extend their scope of knowledge and improve the quality of research outcomes and performance (Goble & Roure, 2007; Li et al., 2012; Parnell, 2011). Especially, there has been a dramatic increase in online knowledge networking platforms which enable scientists to create, re-use, and represent knowledge in new ways, such as open-access electronic journals and libraries (e.g., PLoS, PubMed, or ResearchGate), scientific resource sharing platforms (e.g., myExperiment or Galaxy), and collaborative discussion forums (e.g., SEQanswers or BioStar).

Knowledge networking (See Section 1.1.6 for the definition of this term) is seen as assisting scientists' performance, especially in fields of interdisciplinary scientific communities which is composed of people with different knowledge backgrounds and skills (Aniba & Thompson, 2010; Tarczy-Hornoch & Minie, 2005). Through knowledge networking, scientists are able to learn more from each other by combining their specialized knowledge with new insights from others in a timely manner (Sonnenwald, 2007). For these reasons, funding agencies (e.g., (NSF, NASA, NIH and USDA) have sought to build virtual communities alongside large online bioinformatics projects to establish a collaborative culture for biological research communities at both the individual and organizational levels (Kaye, Heeney, Hawkins, De Vries, & Boddington, 2009).

Given the increasing role of computer-mediated systems in biological research communities, distributed cognition enables us to understand user experiences emerging from interactions where technology acts as a mediator between the users and the activity (Lallemand, Gronier, & Koenig, 2015). As discussed in Section 1.1.5, individual's cognitive resources (such as the scientists' memory, knowledge, or skills) can be metaphorically extended beyond the boundary of the individual to involve people and external artifacts(Hollan, 2000). From this aspect, the distributed cognition theory can provide a basis for a deeper understanding of how scientists may extend their cognitive capabilities to attain research goals. This study focuses on knowledge networking, examining both *knowledge sharing* and *knowledge use* activities in computer-mediated environments.

2.2 Purpose

The goal of study 1 was thus to examine 1) current user experience levels with bioinformatics

resources, 2) important criteria to satisfactorily support user requirements, 3) levels of distributed cognitive activities, and 4) the barriers and enablers of distributed cognitive activities in complex domains, especially biomedical and life sciences. Accordingly, our research questions are as follows:

- Research question 1: To what extent do bioinformatics resources meet user requirements and which are the most critical requirements unmet by current systems?
- Research question 2: To what extent do scientists engage in distributed cognition activities?
- Research question 3: What are the key factors for scientists to engage in distributed cognition activities and what are the potential opportunities?

To address these questions, we administered a self-report questionnaire. As this study was more exploratory and data-driven in nature, we only formulated hypotheses related to key factors of knowledge networking as described above.

2.3 Research Background

This section addresses why and how knowledge networking is important to support data-intensive research communities and discuss factors that may affect scientists' willingness to participate in knowledge networking.

2.3.1 Knowledge networking

Biology is "an empirical rather than theoretical science" (Stevens, Goble, & Bechhofer, 2000). Scientific discovery activities in biological research require more hands on efforts using *in vitro* (controlled experimental environments), *in vivo* (in a living organism or natural setting) as well as *in silico* (performed on computer or via computer simulation) studies repetitively by comparison with other application areas. Once initial hypotheses are generated, scientists (especially bench scientists) are burdened with the number of iterations of complex processes needed to support (or disprove) their hypotheses, expecting coherent outcomes from different steps along the way to inform subsequent iterations (J. Bartlett & Neugebauer, 2005; J. C. Bartlett & Toms, 2005; Mirel, 2009; Ouzounis, 2000; Tran et al., 2004). It is no wonder that

data-intensive research processes inherently demand a high cognitive load not only for exploring scientific findings using incredibly large volumes of heterogeneous data, but also for making decisions under data complexity and uncertainty (Eric Neumann & Prusak, 2007; Ward, Schmieder, Highnam, & Mittelman, 2013).

Data-intensive biological research has relied heavily on interdisciplinary and dynamic research collaboration that consists of experts from different disciplines around the world (Aniba & Thompson, 2010; Kaye et al., 2009; Romano, Giugno, & Pulvirenti, 2011; Tarczy-Hornoch & Minie, 2005). In recent years, an exponential increase of genomic data has demanded new approaches to analyze, store, organize, and visualize tremendous data sets (Kelling et al., 2009), thus requiring even more coordination and collaboration across disciplines both within and outside biology.

Nonetheless, many bioinformatics resources still do not consider the broad spectrum of user classes and goals (Bolchini, 2009; de Matos et al., 2013; Pavelin et al., 2012). Moreover, there are substantial usability and user experience challenges associated with satisfying the diverse user classes and needs not to mention the typical lack of support for different levels of skill and expertise (de Matos et al., 2013).

Given the increasingly complex data and research processes associated with biological sciences, knowledge networking will likely help improvement performance researchers in finding better solutions for complex and repetitive processes (Eric Neumann & Prusak, 2007; Ward et al., 2013). Through knowledge networking, scientists may consider alternative views, and exchange expertise with other competent professionals, thereby promoting scientific discovery and creativity (Fischer & Zigmond, 2010; E. Neumann, 2007). For these reasons, knowledge sharing afford significant contributions to biological science communities beyond the walls of traditional research organizations (De Roure & Goble, 2009; Goecks, Nekrutenko, Taylor, & Team, 2010; Li et al., 2012; Parnell, 2011).

2.3.2 Factors influencing knowledge networking

Trust in scientific information is likely to influence user behavior in a computer-mediated environment— "who produced what, how and where" (Golbeck, 2008). A credible source based on trustworthy and reputable knowledge is likely to promote scientific collaborations, thereby

reducing the costs of research (Assante, Candela, Castelli, Manghi, & Pagano, 2015; Szulanski, 1996). In this study we argue that knowledge shared by individuals perceived as a credible source will be more readily received by others, thereby suggesting that higher source credibility will be correlated with greater knowledge reuse ("include part of the research into another research or for same research, different configuration") among scientists, Here we define *source credibility* as "the extent to which a recipient perceives a source to be trustworthy and reputable" (Joshi, Sarker, & Sarker, 2007) .

Traditionally, the form of reward and acknowledgement for knowledge contribution was through peer-reviewed publication for recognition and respect from peers (Kaye et al., 2009). To date, knowledge sharing is likely to require a voluntary effort in open source platforms, it is probable that knowledge contribution is driven by motivational basis. A major reason being that sharing of knowledge and expertise is a labor-intensive process (Coleman, 1999; Davenport & Pruzak, 2000). Further, the growth potential of data-intensive biological scientific discovery leads to increased competition for scientific achievements on a global scale (Smith et al., 2011). We reasonably assume that a highly competitive research and publishing culture can make some scientists wary to exchange knowledge and expertise on public online spaces; and instead, scientists would likely take on a passive attitude such as just seeking, asking about, or sharing knowledge in virtual communities.

In order to examine individual motivations, we employ the social exchange theory (SET) (Blau, 1964) designed specifically to identify motivation factors in computer-mediated communication. SET starts from the premise that all individuals pursue "benefit maximization and cost minimization". For example, a researcher in a social setting (e.g., online scientific community) might expect actual benefits (e.g., future support, social awareness, professional networking, or self-content) as a result of sharing knowledge. In other cases, an individual might feel intrinsically obligated to help others or enjoy knowledge sharing for the general advancement of science (a different kind of "benefit" to maximize). Thus, knowledge sharing motivations can be examined from both extrinsic and intrinsic perspectives (Hsu, 2008; Hung, Lai, & Chang, 2011; Kankanhalli, Tan, & Wei, 2005; Wasko & Faraj, 2005). *Extrinsic motivation* refers to doing an activity for goal-driven reasons such as external rewards or benefits (Deci & Ryan, 1980). Whereas, *intrinsic motivation* is defined as the doing of an action for inherent satisfaction and enjoyment rather than for incentives or in response to external pressures (Ryan & Deci, 2000).

According to Bourne & Barbour (2011) scientists are not as willing to share their own data into the public domain as they are to eagerly seek open data. In order for knowledge networking to succeed in a specific domain, knowledge contribution needs to be initially perceived and longitudinally regarded as a worthwhile effort. Extrinsic motivations and tangible benefits likely play a critical role in knowledge networks' ability to adopt and retain biologists that contribute knowledge. Indeed, prior research suggests that reciprocal benefit has great implications for knowledge-sharing in online environments (Hsu, 2008; Hung et al., 2011; Kankanhalli et al., 2005). The ability to develop and maintain meaningful relationships and the potential to foster future reciprocal relationships with others are key extrinsic motivations to share knowledge (Bock, Zmud, Kim, & Lee, 2005; He & Wei, 2009). Moreover, individuals are more likely to share their own knowledge when they aim to establish a professional identity in relevant communities (Donath, 1999; Hsu, 2008; Hung et al., 2011). Therefore, we hypothesize that reciprocal benefit, anticipated relationship, and reputation are likely to be extrinsic motivations to stimulate knowledge-sharing in online scientific communities.

Intrinsic motivations inherently encourage people to voluntarily help others (e.g., via knowledge contribution) without the expectation of any tangible, explicit return (Kankanhalli et al., 2005). Specifically, there is a consensus among researchers that altruism enhances knowledge sharing behavior in virtual communities (Chang & Chuang, 2011; Hars & Ou, 2001; Kwok & Gao, 2004). Since an essential requirement of online knowledge networking is spontaneous participation grounded in common interests and objectives (Ardichvili, 2008), we expect that intrinsic motivation leads to increased knowledge contribution in online scientific communities.

Nonetheless, knowledge is highly likely to be perceived as a significant and as a unique individual's asset, suggesting that one's willingness to share knowledge may be tightly coupled to one's perceived risk of "being scooped". In research and other publishing communities, *being scooped* refers to having one's ideas, results, or theory published by another individual. Knowledge networking platforms generally allow guests and registered members to access shared content independent of the individuals' intention to use it. Research shows that the fear associated of being scooped has a negative impact on scientists' knowledge sharing behavior (Waldrop, 2008). Sharing is also likely to expose a risk of being scooped such as copying or exploiting previous works, thereby reducing the uniqueness of the knowledge (Fischer &

Zigmond, 2010). Another reason to be reluctant to share knowledge with others may include increased scientific competitiveness within a field and opportunities for commercial application (Tenopir et al., 2011). As a result, we considered *fear of being scooped* as one of the influential knowledge sharing factors.

2.4 Research model and hypotheses

Based on iterative interviews with domain experts and substantive literature review, we expect that source credibility will affect scientists' willingness to use knowledge, and motivation factors and fear of being scooped simultaneously effect scientists' willingness to share knowledge in open online access spaces.

According to Joshi et al. (2007) the presence of trust is crucial in knowledge networking. In a survey of biologists and other life scientists, (Park & Gabbard, 2013) found that scientists are concerned about the quality of knowledge and want to authenticate knowledge contributors' expertise, to screen and evaluate the value of shared knowledge. Scientists perceive the shared knowledge to be less valuable in the absence of source credibility. *Source credibility* is likely important to promote and maintain knowledge networking and thus we argue that:

H1. 'Source Credibility" will have a positive impact on 'intention to use knowledge'.

According to Borgman (2012), one of major barriers for scientific knowledge sharing is the lack of reciprocating incentives, credits and benefits. Scientists may even expect benefits to be concomitant with the expense of time and effort required to contribute knowledge. We represent this finding by the construct of *reciprocal benefit* which we define as "the degree to which a person believes he or she will obtain mutual benefits through knowledge sharing" (Wasko & Faraj, 2005). We assume that mutual reciprocity is a positive determinant, thus, we hypothesize that:

H2. 'Reciprocal benefit' will have a positive impact on 'intention to share knowledge'.

Beyond merely releasing data or information for use by others, knowledge networking has been shown to be an effective mechanism for social interaction (He & Wei, 2009). In this respect, scientists might expect to make new research connections as well as maintain existing

relationships within the research community to which they contribute knowledge. Indeed, professional connections can be productive resources over the long run. Along these lines, (Bock et al., 2005) empirically show that reciprocal relationships have a positive effect on knowledge-sharing among individuals. These findings we denote as *anticipated relationship* which we define as "the degree to which one believes one can improve mutual relationships with others through one's knowledge sharing" (Bock et al., 2005). Therefore, we hypothesize that:

H3. 'Anticipated relationships' will have a positive impact on 'intention to share knowledge'.

Researchers have empirically confirmed that reputation can be a strong motivation for knowledge-sharing because doing so helps establish and build one's social status, impact, and image within professional networks (He & Wei, 2009; Hung et al., 2011; Kankanhalli et al., 2005; Wasko & Faraj, 2005). Reputation as an intangible asset can render tangible benefits such as acknowledgement by relevant communities (Ba, Stallaert, & Whinston, 2001). According to Cetina (1999), researchers in molecular biology consider reputation to be utmost importance. For example, there is keen competition to make new scientific findings public in different forms of digital documents (data, position papers or journal articles they have published); to essentially "place a stake in the ground" as to claim the concept as uniquely their own (Constant, Sproull, & Kiesler, 1996; Donath, 1999; Wasko & Faraj, 2005). Thus, *reputation* appears to play a key role in motivating knowledge sharing of scientists (Ensign & Hébert, 2010). This leads to the following hypotheses:

H4. 'Reputation' will have a positive impact on 'intention to share knowledge'.

The term *altruism* refers to "the degree to which a person is willing to increase other people's welfare without expecting returns" (Hsu, 2008), and altruism has also been empirically examined as an intrinsic motivation in online knowledge sharing (Chang & Chuang, 2011; Fang & Chiu, 2010; Hsu, 2008). According to Constant et al. (1996), knowledge contributors are satisfied by an intrinsic desire to help others, even when knowledge-sharing is costly in terms of time and effort. In online knowledge networking platforms, some scientists make voluntary efforts to help others by contributing knowledge despite the fact that they don't know each other, nor expect a reciprocal benefit. Thus, we posit:

H5. ‘Altruism’ will have a positive impact on ‘intention to share knowledge’.

A strong publication record has a significant impact for scientists’ careers (Fischer & Zigmond, 2010). Scientists are competitively pursuing unique and novel discoveries and are generally reluctant to make data and knowledge publicly available to be capitalized by competitors and integrated into others' findings (Kansa, Schultz, & Bissell, 2005). As noted earlier, the concern of being scooped whether sharing an idea, concept, or approach, is one of the major barriers in online knowledge sharing in data intensive scientific communities (Park & Gabbard, 2013). That is, the fear of being scooped is a barrier, that likely has a negative effect on one’s willingness to share knowledge (Waldrop, 2008). Therefore, we expect that the fear of being scooped will lead to a decreased intention to share knowledge. Therefore, we propose:

H6. ‘Fear of being scooped’ will have a negative impact on ‘intention to share knowledge’.

It is easier to find and use shared knowledge than to share own knowledge because knowledge sharing requires additional effort to articulate and codify (Wilson, 2002). We assume that commitments emerging from credibility during knowledge use will lead to other types of relational outcomes, such as individuals engaging in longitudinal knowledge sharing. This leads to the following hypothesis:

H7. ‘Intention to use knowledge’ will have a positive impact on ‘intention to share knowledge’.

The proposed research model in Figure 4 depicts the collective hypotheses and relationships.

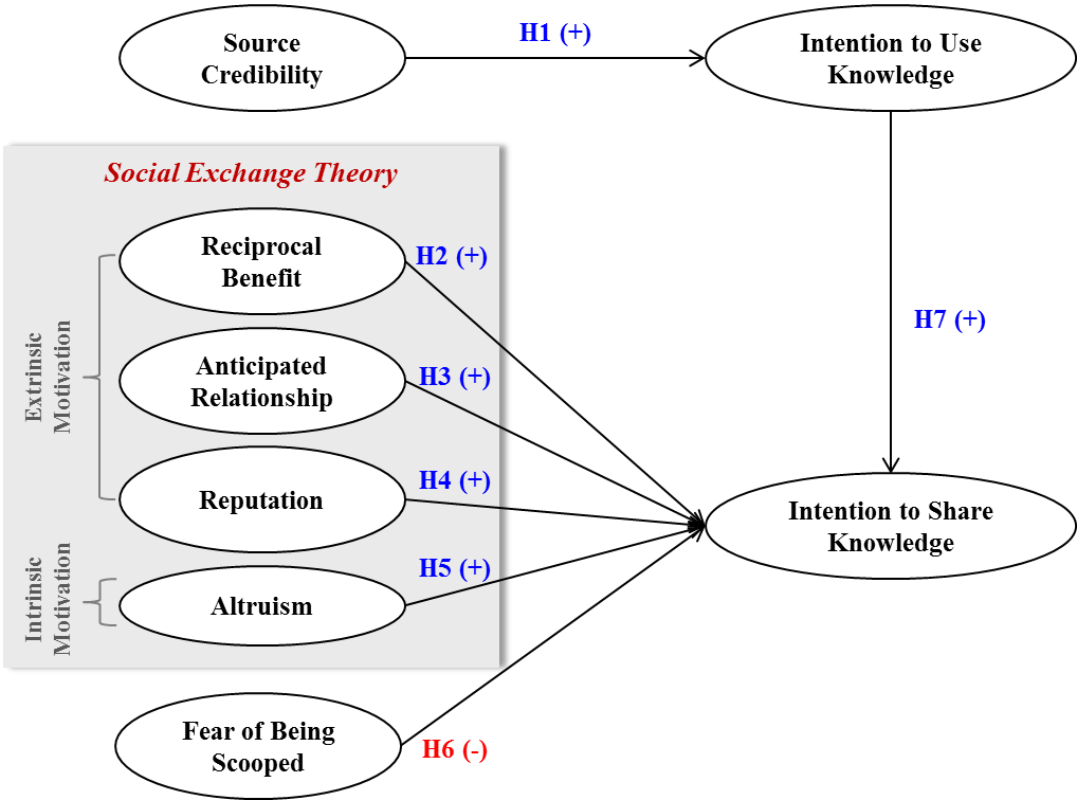


Figure 4. Proposed research model.

The model specification is shown in Table 1.

Table 1. Model specification

Construct		Definition	Adapted from
Motivation Factors	Reciprocal benefits	The degree to which a person believes he or she could obtain mutual benefits through knowledge sharing.	Wasko and Faraj, 2000; Yamagishi and Cook, 1993
	Reputation	The degree to which a person believes that participation could enhance personal reputation through knowledge sharing	Hsu and Lin, 2008
	Anticipated relationships	The degree to which one believes one can improve mutual relationships with others through one's knowledge sharing	Bock et al., 2005
	Altruism	The degree to which a person is willing to increase other people's welfare without expecting returns	Kankanhalli et al., 2005
Trust Factors	Fear of losing one's unique value	The degree to which a person believes that he or she could lose own value (getting scooped) through knowledge sharing	Renzl, 2008
	Source Credibility	An attitude about a source along multiple dimensions, including trustworthiness and expertise	Grewal et al., 1994; McCroskey et al., 1974
Intention to use		The strength one's willingness to reuse knowledge in a virtual community for their research	Bock et al., 2005
Intention to share		The strength one's willingness to contribute knowledge in a virtual community	Bock et al., 2005

2.5 Methods

2.5.1 Participants

Participants were sampled in workshops offered by the Virginia Bioinformatics Institute (VBI) and in cooperation with the Virginia Tech Genomics, Bioinformatics, Computational Biology (GBCB) graduate program. Advertisements also were sent out through email listservs to the Virginia Tech biological sciences graduate program. It was not possible to calculate response rates because we could not know how many people received the emails.

2.5.2 Procedures

We obtained the Virginia Tech Institutional Review Board (IRB) approval (See Appendix A)

prior to any data collection. We then provided study advertisements including a brief overview of the purpose of the questionnaire and directed participants to an online questionnaire available via survey.vt.edu. The online questionnaire included the purpose of the research, instructions for completing the questionnaire, and confidentiality procedures. We assured that participants could freely withdraw from the survey at any time, for any reason.

2.5.3 Survey

We conducted an online survey study with scientists to achieve this goal. Based on pilot testing, the survey required approximately 15–20 minutes to complete. The questionnaire included multiple items to measure the abstract constructs addressed by each research objective. The complete questionnaire is presented in Appendix B. For instance, we solicited research experience, challenges, and expectations with current bioinformatics resources to both understand current user experiences and potential opportunities. We employed open-ended questions to find out more about user situation, opinion, and factual information, and closed-ended questions to gather quantitative data from the response categories. We employed one multiple choice question from our preliminary research, which accounted for a list of criteria of bioinformatics based on scientists' expectations (Park & Gabbard, 2013). We also adopted survey questions as suggested by (Nonnecke & Preece, 2000) which aim to identify reasons why many people lurk, as opposed to actively contribute, on knowledge sharing platforms. In order to build a quantitative model, we employed all construct measurements from established scales that had been validated by previous studies and modified to reflect the data-intensive research context: reciprocal benefit from (Wasko & Faraj, 2005), anticipated relationship from (Bock et al., 2005), reputation from (Hsu, 2008), altruism from (Kankanhalli et al., 2005), fear of being scooped from (Renzl, 2008), and intention to use /share knowledge from (Bock et al., 2005). We designed response categories using a Likert scale format with anchors from 1 (strongly disagree) to 7 (strongly agree). To account for effects of demographics (e.g. age), we presented demographic questions at the bottom of the question list. To ensure the content validity (Messick, 1989), two subject matter experts from Virginia Tech's Virginia Bioinformatics Institute reviewed the entire questionnaire instrument for relevance to the context of biological sciences, content reliability, clarity of instruction, and clarity of survey instrument.

2.6 Analysis and results

We analyze the results from three perspectives. First, we briefly describe participants' demographic information and user experience levels with bioinformatics resources as well as knowledge networking platforms. This is useful in understanding users' overall experience that is situated within the ecology of bioinformatics. Second, we examine why many scientists lurk rather than engage in online knowledge sharing, to inform design of future knowledge-sharing platforms in this domain. Lastly, we present the results from a Partial least squares (PLS) analysis, specifically SmartPLS 2.0 (Ringle, Wende, & Will, 2005), to identify descriptive and predictive relationships (Sellin & Keeves, 1997). PLS requires less restrictive demands on residual distribution than a structural equation model technique (Chin, Marcolin, & Newsted, 2003) and is an efficient statistical prediction tool when working with small to medium- sample size (i.e., less than 200 respondents) (Boulesteix & Strimmer, 2007; Chin, 1998b; Haenlein & Kaplan, 2004).

2.6.1 Descriptive statistics

The demographic information about respondents (*n*=179) is shown in Table 2.

Table 2. Demographic analysis

Variables	Categories	*n*	%
Gender	Female	93	52.0%
	Male	86	48.0%
Research Role (all that apply)	Biologist	137	76.5%
	Bioinformatician	32	17.9%
	Computer Scientist	15	8.4%
	Chemist	12	6.7%
	Mathematician	4	2.2%
	Other	22	12.3%
Experience in biological research fields	One year	15	8.4%
	One to five years	56	31.3%
	Five to ten years	61	34.1%
	More than ten years	47	26.3%
Use of bioinformatics tools	Almost everyday	57	31.8%
	More than once a week	70	39.1%
	More than once a month	28	15.6%
	Around once a month	24	13.4%
Use of online knowledge sharing platforms	I have only seen them.	71	39.7%
	Little experience, I have used them only a few times for my work or research.	63	35.2%
	Experienced, I use them regularly to do my work or research.	40	22.3%
	Very experienced, I use them for almost all my work or research.	5	2.8%

We removed 7 responses with incomplete questions and excluded 2 responses from inexperienced participants who have less than 1 year research experience in biological sciences. A total of 52.0% (*n*=93) of respondents are female, and 48.0% (*n*=86) are male. Approximately 76.5% (*n*=137) of the respondents self-reported their role as biologist. All respondents have experience in conducting biological research and about 60.3% (*n*=108) of participants have been working in the life science domains over 5 years. Around 71.0% (*n*=127) of the participants use bioinformatics resources on a daily basis or several times a week. About 74.3% (*n*=133) of the participants use knowledge networking platforms in a passive manner, that is they consume, but

do not provide or share knowledge.

2.6.2 Current limitations in bioinformatics resources

A commonly identified barrier is that most participants are not satisfied with bioinformatics resources due to the lack of integration and inconsistent results across heterogeneous resources (e.g. different gene naming conventions, different annotations for the same gene). In the same vein, participants repeatedly highlighted current limitations in bioinformatics resources due to the poor quality of genomic sequences and metadata. Some participants noted limited capability of visualizing large and complex data. Inconsistency in user interfaces and general lack of usability were cited as major difficulties for a number of participants, implying a steep learning curve (i.e., long learning times) as a key usability issue. In addition, some participants had trouble accessing data due to complex information architecture and navigation structures. Lastly, data security was noted as an important issue, since many researchers are leveraging these resources to support hypotheses generation, publications or grants.

2.6.3 Important criteria of bioinformatics resources

We asked participants which criteria are the most important or valuable to support their research in a big data environment. Multiple responses were categorized and displayed using the simple frequency table. Table 3 presents the frequencies of important criteria.

Table 3. Distribution of important criteria, including relative frequencies

	Frequency (f)	Relative Frequency (n/f)	Percentage Frequency (%f)
Speed and responsiveness of resource	134	0.140	14.05
Wealth of available data	127	0.133	13.31
Breath of resource tools and functions	96	0.101	10.06
Ease of use	91	0.095	9.54
Degree of data integration	89	0.093	9.33
Advanced visualizations	83	0.087	8.70
Ability to upload my own data	78	0.082	8.18
Ability to ask questions related to my research	75	0.079	7.86
Ability to collect knowledge from others researchers	71	0.074	7.44
Ability to create publication quality images	70	0.073	7.34
Ability to share knowledge with other researchers	40	0.042	4.19
Total	**954**	**1.000**	**100**

Performance-related factors were ranked relatively high in "important criteria of bioinformatics resources". Namely, participants valued "speed and responsiveness" (14.05%), followed by "wealth of available data" (13.31%), and "breadth of resource tools and functions" (10.06%). Next, "ease of use" (9.54%), "degree of data integration" (9.33%), and "ability to upload my own data" (48.1%) were important criteria for supporting their research.

What is interesting in this data is that some participants selected "ability to ask questions related to my research" (7.86%) and "ability to collect knowledge from others" (7.44%) as an important feature. In contrast, a much smaller proportion of participants (4.19%) appear interested in actually sharing their knowledge with others.

2.6.4 Levels of distributed cognitive activities

In order to better understand distributed cognitive activities of scientists in a big data environment, we wanted to know what kinds of online networking platforms scientists use and to what extent scientists get involved in knowledge sharing activities in a distributed cognitive

environment.

Participants reported that they use various types of knowledge resources in their research and the most commonly used were Wiki, NCBI, ResearchGate, Seqanswer, and Biostar. Participants also specified online social networking services most often used as LinkedIn and Twitter. Notably, "Googling" was a common method to find professional forums that provide reliable knowledge resources.

We then asked participants to rate their experience with knowledge networking platforms on a scale of 1 (I have only seen them) to 4 (very experienced, I use them for almost all my work or research). We conducted a linear-by-linear association test to examine the linear association between research experience and knowledge networking experience. The linear by linear association Chi-square is an ordinal measure of significance, which is preferred when testing the significance of linear relationship between ordinal variables (Agresti, 1996, pp. pp 231-236). The result of this analysis presented in Table 4.

The linear by linear association test yielded a p-value of .365 ($\chi^2 = .822$). This value suggests that years of research experience and levels of participation in knowledge networking are not related (i.e., they are independent), supporting the null hypothesis at $\alpha=.05$.

We also tested the relationship between levels of using bioinformatics and levels of participation in knowledge networking. See Table 5.

The linear by linear association test produced a p-value of .000 ($\chi^2 = 16.209$). This result shows a linear trend in levels of using bioinformatics and levels of participation in knowledge networking are related, which indicates that scientists who use bioinformatics resources often were more likely to actively engage in knowledge networking, supporting the null hypothesis at $\alpha=.05$.

Most of participants (88%) reported that their organizations encourage them to engage in knowledge networking. However, only a small minority of participants actively participate in knowledge networking.

Table 4. Crosstabulation of the data sets by years of research experience and levels of participation in knowledge networking

		Levels of participation in knowledge networking				Total
		I have only seen them.	Little experience, I have used them only a few times for my work or research.	Experienced, I use them regularly to do my work or research.	Very experienced, I use them for almost all my work or research.	
Research experience	I have been working 1 year.	6	6	3	0	15
	I have been working 1-5 years.	22	17	15	2	56
	I have been working 5-10 years.	22	20	17	2	6
	I have been working more than 10 years.	21	20	5	1	47
Total		71	63	40	5	179

Table 5. Crosstabulation of the data sets by levels of using bioinformatics and levels of participation in knowledge networking

		Levels of participation in knowledge networking				Total
		I have only seen them.	Little experience, I have used them only a few times for my work or research.	Experienced, I use them regularly to do my work or research.	Very experienced, I use them for almost all my work or research.	
Levels of using bioinformatics	Almost everyday	17	20	16	4	57
	More than once a week	23	25	21	1	70
	More than once a month	16	11	1	0	28
	Around once a month	15	7	2	0	24
Total		71	63	40	5	179

2.6.5 Lurking levels

Table 6 ranked the many reasons why scientists may lurk in terms of the level of participation in knowledge networking, defined as either “no posting at all” or as “some minimal level of posting” (Nonnecke & Preece, 2000). Many participants specified the reason for lurking as "Just reading/browsing is enough" (43.6%). The second ranked reason was “Not enough time to post” (19.6%), followed by "Had no intention to post from the outset "(16.2%) and “Shy about posting” (15.6%). Only two participants (1.1%) indicated that their work does not allow posting and nobody responded to "Wrong group for me".

Table 6. Reasons why they didn't post to the online communities in their research field

Categories	No. of responses
Just reading/browsing is enough	78 (43.6%)
Not enough time to post	35 (19.6%)
Had no intention to post from the outset	29 (16.2%)
Shy about posting	28 (15.6%)
Want to remain anonymous	19 (10.6%)
Nothing to offer	18 (10.1%)
Others respond the way I would	16 (8.9%)
Still learning about the group	15 (8.4%)
There are too many messages already	15 (8.4%)
No requirement to post	13 (7.3%)
Poor quality of messages or group/community	9 (5.0%)
If I post, I am making a commitment	7 (3.9%)
Concern about aggressive or hostile responses	6 (3.4%)
Do not know how to post to this group	6 (3.4%)
Long delay in response to postings	5 (2.8%)
My work does not allow posting	2 (1.1%)
Wrong group for me	0 (0%)
Other	11 (2.2%)

Qualitative data also suggests that the fear of getting scooped can discourage participants from

knowledge networking in their research fields. A survey respondent said:

> *"I worried about getting scooped from others in open forums."*

In addition to fear of being scooped, participants suggested that "benchmarking current research trends" is important in order to keep knowledge up-to-date. For instance, participants appear likely to look for current discussion topics to keep abreast of emerging trends outside one's own group or institution, as indicated below:

> *"I like to know who has started a new discussion thread related to my area of interest, because I want to be aware what is going on outside my lab, "and what other researchers are thinking or focusing on."*

Many participants were reluctant to make new accounts in order to participate in knowledge networking.

> *"I don't like registering for too many accounts, and then I can't remember those usernames and passwords."*

2.6.6 Key factors for distributed cognitive activities

2.6.6.1 Measurement validation

We assessed the internal consistency reliabilities and correlation among constructs (Table 7). As recommended by Bagozzi and Yi (1988), all Cronbach's alphas and composite reliabilities (CR) are higher than the required minimum of 0.7. We established convergent validity for cases where the average variance extracted (AVE) for each factor accounted for 0.50 or more of the total variance. We found calculated values of AVE for each one of the seven items (i.e., reciprocity, relationship, reputation, altruism, fear of being scooped, intention to use knowledge, and intention to share knowledge) to be more than 0.50 (Fornell & Larcker, 1981). In addition all factor loading exceeded 0.5 (Nunnally, 2010). Thus, we confirmed convergent validity for the hypothesized research model.

Table 7. Internal consistencies and correlations of constructs

Construct	Indicators	Factor	CR	AVE	Cronbach's α
Source Credibility	SC1	0.7385	0.8782	0.5936	0.8238
	SC2	0.8608			
	SC3	0.8480			
	SC4	0.7622			
	SC5	0.6176			
Reciprocal Benefit	RP1	0.8892	0.9300	0.8158	0.8872
	RP2	0.9082			
	RP3	0.9121			
Anticipated Relationship	AR1	0.9383	0.9508	0.8656	0.9223
	AR2	0.9418			
	AR3	0.9108			
Reputation	RE31	0.7997	0.8740	0.6982	0.7837
	REP2	0.8648			
	REP3	0.8410			
Altruism	ALT1	0.9193	0.9354	0.8787	0.8647
	ALT2	0.9552			
Fear of being scooped	FE1	0.7774	0.9200	0.7425	0.8859
	FE2	0.8755			
	FE3	0.9074			
	FE4	0.8809			
Intention to Use Knowledge	IU1	0.8598	0.9178	0.7367	0.8798
	IU2	0.9057			
	IU3	0.8778			
	IU4	0.7853			
Intention to Share Knowledge	IS1	0.7426	0.9241	0.7096	0.8967
	IS2	0.8179			
	IS3	0.8725			
	IS4	0.8735			
	IS5	0.8963			

We assessed discriminant validity by looking at the square root of each corresponding AVE. Anderson and Gerbing (1988) suggest that if the square root of the AVE for each construct is greater than the levels of correlations between it and other constructs, discriminant validity is confirmed. Table 8 (numbers in diagonal of the cross-loading matrix) shows this criterion is met in all measurements, thus there is discriminant validity.

Table 8. Correlation between constructs

	ALT	SC	FEA	IS	IU	RB	AR	REP
ALT	**.9474**							
SC	. 3713	**.7704**						
FEA	-.3295	-.1690	**.8617**					
IS	. 4796	.3781	-.4482	**.8424**				
IU	.4042	.4744	-.2870	.6942	**.8583**			
RB	. 2674	.3032	-.2639	.4557	. 4189	**.9032**		
AR	. 5713	.2871	-.1776	.4216	. 3767	.4054	**.9304**	
REP	. 3105	.1510	-.2679	.4103	. 3178	.3503	.2516	**.8356**

Note: Values on the shaded diagonal are the square roots of the AVEs.
Altruism (ALT), Source Credibility (SC), Fear of being scooped (FE), Intention to share knowledge (IS), Intention to use knowledge (IU), Reciprocal Benefit (RB), Anticipated Relationship (AR), and Reputation (REP).

2.6.6.2 PLS path model

After confirming reliability and validity of the constructs, we performed the bootstrapping method (with 500 sub-samples, 179 cases) suggested by Chin (1998b) to examine the statistical significance of path coefficients. The path coefficients show the strength and direction of relationships among constructs. Using this method, we tested the significance of each path coefficient using different *t* values ($\rho < 0.1$ when $t > 1.645$; $\rho < 0.05$ when $t > 1.96$; $\rho < 0.001$ when $t > 2.58$). In addition, explanatory power of the model can be explained by R^2. Thus, a larger R^2 provides better explanatory power (Wixom & Watson, 2001). The results of this analysis are described in Figure 5; showing the overall explanatory power, path coefficients, and associated *t-values* of the paths of the research model.

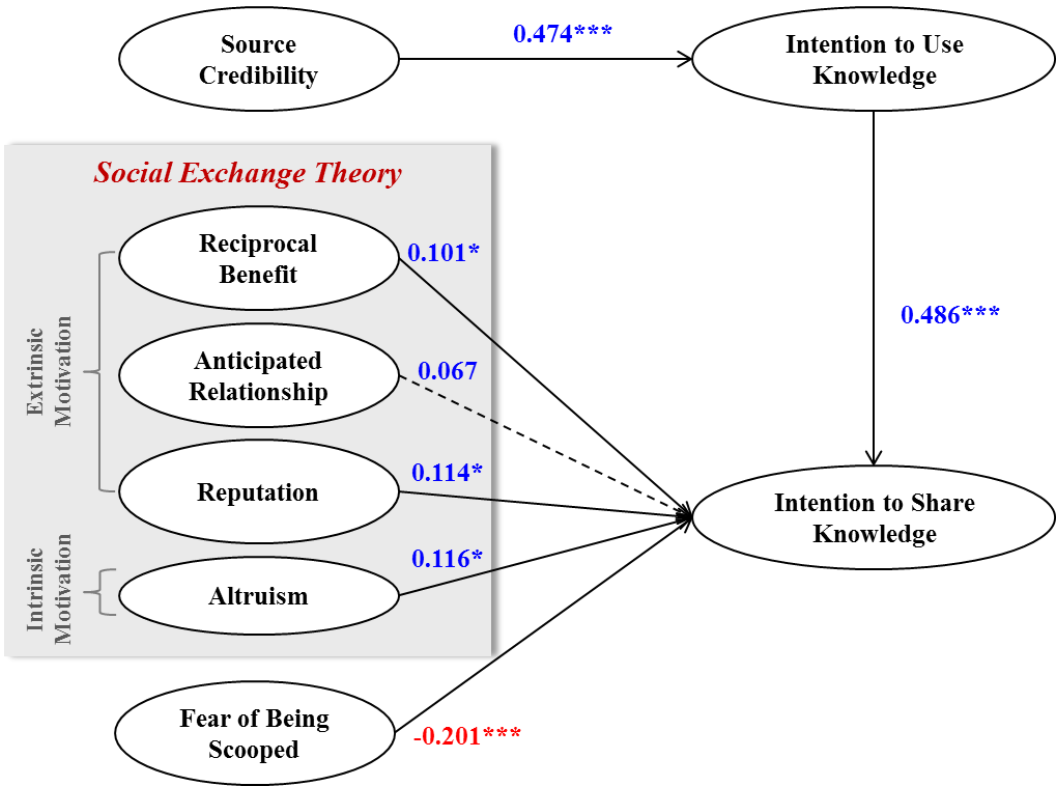


Figure 5. Results of PLS path analysis.
Note: All paths represented by solid lines are statistically significant. * ρ <0.10; ** ρ <0.05; *** ρ <0.01. The paths represented by dotted lines are not statistically significant.

Our results indicate that the target construct 'intention to use knowledge' yields an R^2 value of 23% (Adjusted R^2=0.226) as having moderate explanatory power, and ' intention to share knowledge' an R^2 value of 60% (Adjusted R^2=0.586) as having substantial explanatory power(Chin, 1998a).

Not all structural paths are found to be statistically significant in the research model (see dotted lines in Figure 5). The statistical results support our *H1, H2, H4, H5, H6, and H7* hypotheses. To be specific, our results suggest that **'Source Credibility'** has a positive effect on the intention to use knowledge (β = 0.474, ρ < 0.01), thus supporting *H1*. **'Reciprocal Benefit'** has a positive effect on the intention to share knowledge (β = 0.101, ρ < 0.10), thus supporting *H2*. Also, the path coefficient value between 'Reputation' and 'intention to share knowledge' confirms our assumption (*H3*). *H5* is supported by the data (β = 0.116, ρ < 0.10). It is therefore that '**Altruism**' is positively correlated with 'Intention to share Knowledge'. Supporting *H6*, the value of '**Fear of**

being Scooped' has a negative effect on the intention to share knowledge (β = - 0.201, ρ < 0.01). Moreover, our results explain that '**Intention to use Knowledge**' has a positive effect on '**Intention to share Knowledge**' (β = 0.486, ρ < 0.01) (*H7*).

On the other hand, some of our results are inconsistent with our stated hypotheses. For example '**Anticipated Relationship**' does not appear to have any significant impact on 'intentions to share knowledge'. Hence, we cannot confirm *H3*. Table 9 summarizes the results of path coefficient analysis.

Table 9. Summary of hypothesis testing results

Path (Hypothesis)	*β*	*t*	*p*	Results
Source Credibility → Intention to Use Knowledge (H1)	0.474	5.4350	***	Supported
Reciprocal Benefit → Intention to Share Knowledge (H2)	0.101	1.7258	*	Supported
Anticipated Relationship → Intention to Share Knowledge (H3)	0.067	1.0014		Not Supported
Reputation → Intention to Share Knowledge (H4)	0.114	1.8595	*	Supported
Altruism → Intention to Share Knowledge (H5)	0.116	1.7667	*	Supported
Fear of being Scooped→ Intention to Share Knowledge (H6)	- 0.201	3.9926	***	Supported
Intention to Use knowledge → Intention to Share Knowledge (H7)	0.486	9.3874	***	Supported

2.7 Discussion

Bioinformatics is a vast and complex interdisciplinary research area where numerous online resources have been developed to allow scientists to analyze tremendous amounts of genomic data (Goujon et al., 2010). However, scientific communities in bioinformatics may not be sufficiently supported through the introduction of functions, algorithms, or software alone. Given the increasing roles of computer-mediated systems in biological research communities, it is timely to support user requirements and consider the factors proven to influence user experience.

We first discuss perceptions of the barriers and enablers of bioinformatics influencing user experience. We then summarize the factors affecting distributed cognitive activities, especially the intention to use and share knowledge.

2.7.1 Current user experience

Scientists still struggle to uncover, explore and exploit the capabilities of online bioinformatics resources and further to adapt them to for their purposes and contexts. In terms of user experience, scientists stressed the need for consistent, intuitive and easy bioinformatics resources. It clearly indicates that many online bioinformatics resources have employed a system-oriented software development approach rather than user-centered design approach that aims to better support real user requirements and different levels of domain knowledge. First and foremost, system performance and usability issues should be urgently addressed in order to enhance the overall user experience of online bioinformatics resources.

Although most participants report using 'generic' and well-known knowledge networking resources such as Wiki, ResearchGate, and LinkedIn, the results indicate that active participation by the bioinformatics research community in knowledge networking platforms is rather low. In other words, scientists are more interested in enhancing the overall quality (and performance) of their research by getting answers to their questions and accessing others' shared knowledge rather than sharing their knowledge and skills.

The use of Web 2.0 in scholarly communications is often characterized as being of special interest for a younger generation. However, our result reveals no relationship between years of research experience and levels of participation. That is, even relatively inexperienced scientists (presumably younger) do not participate in knowledge networking actively in comparison to more experienced scientists. This finding is consistent with past studies by Newman (2009) and Procter et al. (2010). An explanation may be found in a Tenopir et al. (2011) study that found younger scientists are more interested in protecting their data from their professional society in hopes of ensuring career advancement.

Our study also reveals that scientists who make use of online bioinformatics resources are often more likely to be active users in knowledge networking platforms (both consuming and sharing). This perhaps suggests that bioinformatics resources user experience design guidelines include prompts to encourage the culture of knowledge networking.

Scientists report many reasons why they lurk in knowledge networking platforms. The leading reasons were "satisfactions of reading/browsing", "insufficient time", and "lack of intentions to

post'. These constraints may be eased by UX initiatives that make searching, consuming and/or contributing data quick and easy and provide explicit perceived benefits.

2.7.2 Key determinants on intention to share and use knowledge

Our results provide support most of our hypotheses: Source credibility, fear of getting scooped, and motivation factors (i.e., extrinsic and intrinsic motivations) directly affect scientists' intent to engage in online knowledge networking.

This study confirms that the **source credibility**, accomplished through trustworthiness of the information source and reputation of a person who shares knowledge, has considerable impact on scientists' intention to use knowledge in online public spaces. This result is consistent with previous studies as well (Assante et al., 2015; Golbeck, 2008; Szulanski, 1996). Our results on source credibility further suggest that scientists may have a perceived benefit from enhanced visibility, credibility associated with sharing knowledge. As such, we encourage bioinformatics resource UX designers to consider provenance mechanisms such as author recognition, citation indices and reviewer comments rankings features.

Our findings also confirm that both extrinsic (i.e., reciprocity benefit, reputation) and intrinsic motivations (i.e., altruism) have significant influence on knowledge sharing behavior. As an extrinsic motivation, **reciprocity** is important in encouraging the intention for knowledge-sharing within scientific communities. If scientists observe constructive examples of mutual reciprocity within a resource, it follows that they would likely be more motivated to engage and participate. This is also in line with previous studies that show mutual benefits have a direct influence on behavioral intention to share knowledge (Bock et al., 2005; Hsu, 2008; McLure Wasko & Faraj, 2000; Wasko & Faraj, 2005). However, there is no evidence that reciprocity has any effect on the *quantity* or *quality* of knowledge contribution (Hung et al., 2011; Wasko & Faraj, 2005). Next, **reputation** appears to significantly influence one's 'intention to share knowledge'. This may explain why many scientists are sharing their own published articles, workflows, or standard operating procedures in different types of online knowledge networking platforms. It is already well known that getting credit through formal citations reflects one's visibility (e.g., recognition level) as a knowledgeable expert in professional communities. Such explanation is supported by Fischer and Zigmond (2010), who found that sharing through

publication influences the welfare of authors. Our results are also consistent with Birnholtz (2007) who found that reputation in molecular biology is attributed to researchers who compete to accumulate first-named author publications. Knowledge networking platforms that employ reputation mechanisms, such as displaying the most useful or "best" knowledge along with authors' identification, will establish a stronger motivation for scientists to share knowledge.

The provision of an intrinsic motivation, **Altruism**, significantly affects our intention to share knowledge. As we stated earlier, altruism is unconditional kindness without the expectation of compensation for that action (Hung et al., 2011). This finding coheres with previous studies that people enjoy helping others without expecting returns (Constant et al., 1996; Hars & Ou, 2001; Hsu, 2008; Hung et al., 2011; Taylor, 2006). Also, this finding is consistent with other domains such as software engineers or R&D employees (Hars & Ou, 2001; Hung et al., 2011; Von Krogh, Spaeth, & Lakhani, 2003). Recently, Tenopir et al. (2011) found that older scientists are more altruistic and more willing to share than younger scientists because older scientists tend to have a greater sense of responsibility to share data. The results in this study therefore need to be further examined in experiments that control for age and/or research experience.

Despite of our initial assumptions, this study shows no evidence of causality between **anticipated relationship** and behavioral intention to share. Scientists were unlikely to expect mutual relationship by sharing knowledge. This finding contradicts previous studies (Donath, 1999; Hsu, 2008; Hung et al., 2011) that suggest scientists are likely to develop and main meaningful relationships with others in professional virtual communities. One possible explanation is that mutual relationship via virtual communities is not a strong enough factor to motivate knowledge sharing among life scientists; hence it is not a critical matter for forming or sustaining stable collaborative relationships in biological scientific communities.

The fear of being scooped is the most obvious barrier for knowledge networking in online scientific communities and it negatively affects the intention to share knowledge. Once data have been released into the public space, it is difficult for contributors to control its usage (Kaye et al., 2009) and potential to foster new research discoveries (Vivien, 2012). In order to offset the fear of being scooped by others, knowledge networking platforms must support ways to alleviate concerns. Although the process of obtaining informed consent could be possible, it is also problematic in knowledge networking (Kaye et al., 2009). Instead, systems may allow

differential access to shared knowledge based on user profiles, preferences and activity histories so that scientists are more likely to trust others and share accumulated expertise as well as formalized knowledge. Also, online bioinformatics resource UX designers should ensure that knowledge contributors are able to protect their interests, receive appropriate credits for sharing their preliminary works, and ensure that trust is maintained.

Lastly, scientists' **'intention to use knowledge'** is likely to be a prerequisite for **'intention to share knowledge'**, because without positive experience in knowledge networking platforms, scientists cannot be encouraged to engage in knowledge networking, and consequently spontaneous participation in knowledge sharing. This result highlights that providing salient and extrinsic benefits for both knowledge reuse and sharing is more likely to facilitate knowledge networking in online biological scientific communities.

2.8 Conclusion

This study examines current user experience levels with bioinformatics resources as well as the barriers and enablers of distributed cognitive activities, especially knowledge networking in big data environments.

Today scientists make use of various knowledge resources to determine if the experiment they wish to run has already been performed or even tangentially explored (Greenwood et al., 2004). Many online bioinformatics resources are designed for highly specialized domain experts and require highly technical skills. To be specific, lack of integration (e.g. different gene naming conventions, different annotations for the same gene), inconsistent results and user interfaces, and steep learning curves repeatedly came up as examples of current limitations. The research has shown that performance-related factors such as speed and responsiveness of resources and ease of use ranked relatively high in importance. These findings are valuable to user experience professionals in complex domains, especially the biomedical and life sciences since they represent real user requirements and can provide general guidance for UX improvements in current bioinformatics resources.

As biological sciences become more data intensive and interdisciplinary (Kaye et al., 2009), the concept of open science is rapidly emerging to support contemporary data-intensive research communities such as bioinformatics (Proctor, 2010). Some claim that open science is becoming

central to scientific progress (Vivien, 2012). Although, knowledge networking hopes to facilitate scientific collaboration and innovation, current knowledge networking practices in bioinformatics have not resulted in widespread adoption and use. One possible explanation, as explored in this work, is the lack of consideration for contextual issues as well as user expectations. However, there is ample opportunity for researchers to make improvements to knowledge networking platforms from user experience perspectives. The larger implication is that the concept of distributed cognition cannot be ignored when designing to support online interdisciplinary research communities.

The work presented herein is intended to empirically address user experience from distributed cognition perspectives and key determinants affecting knowledge networking intentions in data-intensive scientific communities. Given that knowledge networking is commonly anonymous in nature, scientists are strongly influenced by source credibility and fear of being scooped. This perhaps reflects the competitive nature of the scientific communities. A unique contribution to the field, is the fact that this study demonstrated that knowledge networking among scientists is influenced by both extrinsic (e.g., reciprocal benefit, reputation) and intrinsic (e.g., altruism) motivations within professional networks. Whereas anticipated relation is not an influential factor of intention to share. Also, we identified the impacts of intention to use knowledge on one’s intention to subsequently share knowledge.

Given the importance of today's knowledge-based society, our findings further support user experience professionals who seek to uncover unmet user needs and expectations in complex scientific domains. This study, and many others like it, aim to facilitate the emerging paradigm of Science 2.0 that supports the "free and widespread availability of data, the sharing and reuse of methods and tools and the collaborative pursuit of common goals and objectives" (Romano et al., 2011).

Bioinformatics are still at an early stage in user experience sophistication and associated knowledge networking support within data-intensive interdisciplinary research communities. As such, bioinformatics resource designers and developers should seek to balance support for cutting-edge genomic science with support for emerging and creative user requirements.

2.9 Future directions

First, the results of this study do not explain differences across scientists in specialized roles such as bench scientists (those mainly conducting experiments in a laboratory) and application scientists (those designing and developing bioinformatics resources). Thus, it can be further expended to examine distributed cognitive activities in interdisciplinary research communities that contain a broader spectrum of user classes.

In addition, outcomes derived through survey instruments rely on self-reporting. Our findings may not faithfully represent user experiences in terms of other settings, times, or places. It is worth further comprehensive studies with other relevant individual and contextual factors, as well as additional dimensions that may be present in data-intensive research domains.

Finally, the present study highlights which factors must be considered to support user experience using empirical evidence gathered via survey data. However, survey instruments alone may not fully elicit implicit attitudes, beliefs, and values that guide behavior in a distributed cognition environment. Hence, a mixed methods approach should be considered to offset the potential methodological shortcomings associated with surveys and self-reporting behavior.

2.10 References

Agresti, Alan. (1996). *An introduction to categorical data analysis* (Vol. 135): Wiley New York.

Anderson, James C, & Gerbing, David W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411-423.

Aniba, Mohamed Radhouene, & Thompson, Julie D. (2010). *Knowledge Based Expert Systems in Bioinformatics*.

Ardichvili, Alexandre. (2008). Learning and knowledge sharing in virtual communities of practice: Motivators, barriers, and enablers. *Advances in developing human resources*.

Assante, Massimiliano, Candela, Leonardo, Castelli, Donatella, Manghi, Paolo, & Pagano, Pasquale. (2015). Science 2.0 Repositories: Time for a Change in Scholarly Communication. *D-Lib Magazine, 21*(1/2).

Ba, Sulin, Stallaert, Jan, & Whinston, Andrew B. (2001). Research commentary: introducing a third dimension in information systems design—the case for incentive alignment. *Information Systems Research, 12*(3), 225-239.

Bagozzi, Richard P, & Yi, Youjae. (1988). On the evaluation of structural equation models. *Journal of the academy of marketing science, 16*(1), 74-94.

Bartlett, J., & Neugebauer, T. (2005). Supporting information tasks with user-centred system design: The development of an interface supporting bioinformatics analysis. *Canadian journal of information and library science, 29*(4), 486-487.

Bartlett, Joan C., & Toms, Elaine G. (2005). Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach. *Journal of the American Society for Information Science and Technology, 56*(5), 469-482. doi: 10.1002/asi.20136

Birnholtz, Jeremy P. (2007). When do researchers collaborate? Toward a model of collaboration propensity. *Journal of the American Society for Information Science and Technology, 58*(14), 2226-2239.

Blau, Peter Michael. (1964). *Exchange and power in social life*: Transaction Publishers.

Bock, G.W., Zmud, R.W., Kim, Y.G., & Lee, J.N. (2005). Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. *MIS quarterly*, 87-111.

Bolchini, D. (2009). Better bioinformatics through usability analysis. *Bioinformatics (Oxford, England), 25*(3), 406-412. doi: 10.1093/bioinformatics/btn633

Borgman, Christine L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078. doi: 10.1002/asi.22634

Boulesteix, Anne-Laure, & Strimmer, Korbinian. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics, 8*(1), 32-44.

Bull, A.T., Ward, A.C., & Goodfellow, M. (2000). Search and discovery strategies for biotechnology: the paradigm shift. *Microbiology and Molecular Biology Reviews, 64*(3), 573-606.

Cetina, Karin Knorr. (1999). *Epistemic cultures: How the sciences make knowledge*: Harvard University Press.

Chang, Hsin Hsin, & Chuang, Shuang-Shii. (2011). Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. *Information & Management, 48*(1), 9-18. doi: http://dx.doi.org/10.1016/j.im.2010.11.001

Chin, Wynne W. (1998a). Commentary: Issues and opinion on structural equation modeling: JSTOR.

Chin, Wynne W. (1998b). The partial least squares approach for structural equation modeling.

Chin, Wynne W, Marcolin, Barbara L, & Newsted, Peter R. (2003). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study. *Information systems research, 14*(2), 189-217.

Coleman, David. (1999). Groupware: collaboration and knowledge sharing. *Knowledge management handbook, 12*(1), 12-15.

Constant, D., Sproull, L., & Kiesler, S. (1996). The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization science, 7*(2), 119-135.

Davenport, Thomas H, & Pruzak, Laurence. (2000). *Working knowledge: How organizations manage what they know*: Harvard Business Press.

de Matos, Paula, Cham, Jennifer A, Cao, Hong, Alcántara, Rafael, Rowland, Francis, Lopez, Rodrigo, & Steinbeck, Christoph. (2013). The Enzyme Portal: a case study in applying user-centred design methods in bioinformatics. *BMC bioinformatics, 14*(1), 103.

De Roure, D., & Goble, C. (2009). myExperiment: A Web 2.0 Virtual Research Environment for Research using Computation and Services.

Deci, Edward L, & Ryan, Richard M. (1980). The empirical exploration of intrinsic motivational processes. *Advances in experimental social psychology, 13*(2), 39-80.

Donath, Judith S. (1999). Identity and deception in the virtual community. *Communities in cyberspace, 1996*, 29-59.

Ensign, Prescott C, & Hébert, Louis. (2010). How reputation affects knowledge sharing among colleagues. *MIT Sloan Management Review, 51*(2), 79-81.

Fang, Yu-Hui, & Chiu, Chao-Min. (2010). In justice we trust: Exploring knowledge-sharing continuance intentions in virtual communities of practice. *Computers in Human Behavior, 26*(2), 235-246. doi: http://dx.doi.org/10.1016/j.chb.2009.09.005

Fischer, Beth A, & Zigmond, Michael J. (2010). The essential nature of sharing in science. *Science and engineering ethics, 16*(4), 783-799.

Fornell, Claes, & Larcker, David F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, 39-50.

Goble, Carole Anne, & Roure, David Charles De. (2007). *myExperiment: social networking for workflow-using e-scientists*. Paper presented at the Proceedings of the 2nd workshop on Workflows in support of large-scale science, Monterey, California, USA.

Goecks, Jeremy, Nekrutenko, Anton, Taylor, James, & Team, T Galaxy. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol, 11*(8), R86.

Golbeck, Jennifer. (2008). Weaving a web of trust. *Science, 321*(5896), 1640-1641.

Goujon, Mickael, McWilliam, Hamish, Li, Weizhong, Valentin, Franck, Squizzato, Silvano, Paern, Juri, & Lopez, Rodrigo. (2010). A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic acids research, 38*(suppl 2), W695-W699.

Haenlein, Michael, & Kaplan, Andreas M. (2004). A beginner's guide to partial least squares analysis. *Understanding statistics, 3*(4), 283-297.

Hars, Alexander, & Ou, Shaosong. (2001). *Working for free? Motivations of participating in open source projects.* Paper presented at the System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on.

He, W., & Wei, K.K. (2009). What drives continued knowledge sharing? An investigation of knowledge-contribution and-seeking beliefs. *Decision Support Systems, 46*(4), 826-838.

Hollan, James. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM transactions on computer-human interaction, 7*(2), 174-196.

Hsu, C. L. (2008). Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. *Information & management, 45*(1), 65.

Huerta, Michael, Downing, Gregory, Haseltine, Florence, Seto, Belinda, & Liu, Yuan. (2000). NIH working definition of bioinformatics and computational biology. *US National Institute of Health.*

Hung, Shin-Yuan, Lai, Hui-Min, & Chang, Wen-Wen. (2011). Knowledge-sharing motivations affecting R&D employees' acceptance of electronic knowledge repository. *Behaviour & Information Technology, 30*(2), 213-230.

Joshi, Kshiti D, Sarker, Saonee, & Sarker, Suprateek. (2007). Knowledge transfer within information systems development teams: Examining the role of knowledge source attributes. *Decision Support Systems, 43*(2), 322-335.

Kankanhalli, A., Tan, B.C.Y., & Wei, K.K. (2005). Contributing knowledge to electronic knowledge repositories: An empirical investigation. *Mis Quarterly*, 113-143.

Kansa, Eric C, Schultz, Jason, & Bissell, Ahrash N. (2005). Protecting traditional knowledge and expanding access to scientific data: juxtaposing intellectual property agendas via a "some rights reserved" model. *International Journal of Cultural Property, 12*(3), 285-314.

Katoh, M. (2002). Paradigm shift in gene-finding method: From bench-top approach to desk-top approach (review). *Int J Mol Med, 10*(6), 677-682.

Kaye, Jane, Heeney, Catherine, Hawkins, Naomi, De Vries, Jantina, & Boddington, Paula. (2009). Data sharing in genomics—re-shaping scientific practice. *Nature Reviews Genetics, 10*(5), 331-335.

Kelling, Steve, Hochachka, Wesley M., Fink, Daniel, Riedewald, Mirek, Caruana, Rich, Ballard, Grant, & Hooker, Giles. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience, 59*(7), 613-620. doi: 10.1525/bio.2009.59.7.12

Kwok, James SH, & Gao, S. (2004). Knowledge sharing community in P2P network: a study of motivational perspective. *Journal of Knowledge Management, 8*(1), 94-102.

Lallemand, Carine, Gronier, Guillaume, & Koenig, Vincent. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey.

Computers in Human Behavior, 43(0), 35-48. doi: http://dx.doi.org/10.1016/j.chb.2014.10.048

Li, Jing-Woei, Schmieder, Robert, Ward, R. Matthew, Delenick, Joann, Olivares, Eric C., & Mittelman, David. (2012). SEQanswers: an open access community for collaboratively decoding genomes. *Bioinformatics, 28*(9), 1272-1273. doi: 10.1093/bioinformatics/bts128

McLeod, Kenneth, & Burger, Albert. (2008). Towards the use of argumentation in bioinformatics: a gene expression case study. *Bioinformatics, 24*(13), i304-i312. doi: 10.1093/bioinformatics/btn157

McLure Wasko, Molly, & Faraj, Samer. (2000). "It is what one does": why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems, 9*(2), 155-173.

Messick, Samuel. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Mirel, B. (2009). Supporting cognition in systems biology analysis: findings on users' processes and design implications. *J Biomed Discov Collab, 4*, 2. doi: 10.1186/1747-5333-4-2

Neumann, E. (2007). Knowledge networks in the age of the Semantic Web. *Briefings in bioinformatics, 8*(3), 141-149.

Neumann, Eric, & Prusak, Larry. (2007). Knowledge networks in the age of the Semantic Web. *Briefings in bioinformatics, 8*(3), 141-149.

Newman, J. (2009). Researchers of tomorrow. *See http://explorationforchange. net/index. php/current-projects/researchers-of-tomorrow/researchers-of-tomorrow-home. html*.

Nonnecke, Blair, & Preece, Jenny. (2000). *Lurker demographics: Counting the silent.* Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems.

Nunnally, Jum C. (2010). *Psychometric Theory 3E*: Tata McGraw-Hill Education.

Ouzounis, Christos. (2000). Two or three myths about bioinformatics. *Bioinformatics, 16*(3), 187-189. doi: 10.1093/bioinformatics/16.3.187

Park, Jongsoon, & Gabbard, Joseph L. (2013). An exploratory study to understand knowledge-sharing in data-intensive science *Human-Computer Interaction. Users and Contexts of Use* (pp. 217-226): Springer.

Parnell, L. D. (2011). BioStar: An Online Question & Answer Resource for the Bioinformatics Community. *PLoS computational biology*, *7*(10), e1002216.

Pavelin, Katrina, Cham, Jennifer A, de Matos, Paula, Brooksbank, Cath, Cameron, Graham, & Steinbeck, Christoph. (2012). Bioinformatics meets user-centred design: a perspective. *PLoS computational biology*, *8*(7), e1002554.

Procter, Rob, Williams, Robin, Stewart, James, Poschen, Meik, Snee, Helene, Voss, Alex, & Asgari-Targhi, Marzieh. (2010). Adoption and use of Web 2.0 in scholarly communications. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4039-4056.

Renzl, Birgit. (2008). Trust in management and knowledge sharing: the mediating effects of fear and knowledge documentation. *Omega*, *36*(2), 206-220.

Ribes, David, & Lee, Charlotte P. (2010). Sociotechnical studies of cyberinfrastructure and e-Research: Current themes and future trajectories. *Computer Supported Cooperative Work (CSCW)*, *19*(3-4), 231-244.

Ringle, Christian M, Wende, Sven, & Will, Alexander. (2005). SmartPLS 2.0 (beta): Hamburg, Germany.

Romano, Paolo, Giugno, Rosalba, & Pulvirenti, Alfredo. (2011). Tools and collaborative environments for bioinformatics research. *Briefings in Bioinformatics*. doi: 10.1093/bib/bbr055

Roos, D.S. (2001). Bioinformatics--trying to swim in a sea of data. *Science*, *291*(5507), 1260-1261.

Ryan, Richard M, & Deci, Edward L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54-67.

Sellin, Norbert, & Keeves, John P. (1997). Path analysis with latent variables. *Educational research, methodology, and measurement: An international handbook*, 633-640.

Smith, Arnold, Balazinska, Magdalena, Baru, Chaitan, Gomelsky, Mark, McLennan, Michael, Rose, Lynn, . . . Kolker, Eugene. (2011). Biology and data-intensive scientific discovery in the beginning of the 21st century. *Omics: a journal of integrative biology*, *15*(4), 209-212.

Sonnenwald, Diane H. (2007). Scientific collaboration. *Annual review of information science and technology*, *41*(1), 643-681.

Stevens, Robert, Goble, Carole A., & Bechhofer, Sean. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics, 1*(4), 398-414. doi: 10.1093/bib/1.4.398

Szulanski, Gabriel. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic management journal, 17*(S2), 27-43.

Tarczy-Hornoch, Peter, & Minie, Mark. (2005). Bioinformatics Challenges and Opportunities

Medical Informatics. In H. Chen, S. S. Fuller, C. Friedman & W. Hersh (Eds.), (Vol. 8, pp. 63-94): Springer US.

Taylor, Eileen Z. (2006). The effect of incentives on knowledge sharing in computer-mediated communication: An experimental investigation. *Journal of Information Systems, 20*(1), 103-116.

Tenopir, Carol, Allard, Suzie, Douglass, Kimberly, Aydinoglu, Arsev Umur, Wu, Lei, Read, Eleanor, . . . Frame, Mike. (2011). Data sharing by scientists: practices and perceptions. *PloS one, 6*(6), e21101.

Tran, D., Dubay, C., Gorman, P., & Hersh, W. (2004). Applying task analysis to describe and facilitate bioinformatics tasks. *Stud Health Technol Inform, 107*(Pt 2), 818-822.

Vivien, Marx. (2012). My data are your data. *Nature biotechnology, 30*(6), 509-511.

Von Krogh, Georg, Spaeth, Sebastian, & Lakhani, Karim R. (2003). Community, joining, and specialization in open source software innovation: a case study. *Research Policy, 32*(7), 1217-1241.

Waldrop, M Mitchell. (2008). Science 2.0. *Scientific American, 298*(5), 68-73.

Ward, R. Matthew, Schmieder, Robert, Highnam, Gareth, & Mittelman, David. (2013). Big data challenges and opportunities in high-throughput sequencing. *Systems Biomedicine, 1*(1), 23-28.

Wasko, M.M.L., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *Mis Quarterly*, 35-57.

Wilson, Thomas D. (2002). The nonsense of knowledge management. *Information research, 8*(1), 8-1.

Yarfitz, S. (2000). A library-based bioinformatics services program. *Bulletin of the Medical Library Association, 88*(1), 36.

3 Chapter 3: An extended insight-based methodology for scientists' insight generation behavior and human performance: the role of individual differences in bioinformatics research

3.1 Introduction

Many recent advances in genome sequencing-related technology have hastened the development and dissemination of a huge number of online bioinformatics resources within the biomedical and life sciences research communities. For example, the 2015 *Nucleic Acids Research* Database issue alone includes 56 new databases and 115 updates (Galperin, Rigden, & Fernández-Suárez, 2015). The use of bioinformatics allows scientists to gather, store, analyze, and merge heterogeneous datasets, ranging from raw sequences, annotated genomes, protein structures, expression profiles, deep-sequencing data, networks and pathways, ontology relation diagrams, and much, much more (Romano, Giugno, & Pulvirenti, 2011). However, *extracting knowledge* from this mass of bioinformatics data is becoming more complex because of the large number of different repositories with heterogeneous formats, ad hoc information layout, custom-built visualization tools and specialized algorithms – and moreover, these repositories are rarely integrated and intercommunicated.

Bioinformatics has been the driving force behind the explosive demand for more specialized and widely distributed expertise and skills (Romano et al., 2011). Bioinformatics naturally became an interdisciplinary research field of science that integrates knowledge from various groups of professionals in the biological, computational, and mathematical disciplines to analyze and interpret large-scale biological data sets through the use of methods and technologies (Aniba & Thompson, 2010; Romano et al., 2011; Tarczy-Hornoch & Minie, 2005). As mentioned above, bioinformatics has a broad range of users from multiple fields of study, including 'wet' (lab-based) and dry' (computational) research communities (de Matos et al., 2013). However, current online bioinformatics resources are not adequate to support these diverse user populations. Although, previous studies have conceptually discussed the need to induce close interaction between bench scientists and application scientists (Aniba & Thompson, 2010; Letondal & Mackay, 2004; Marcus, 2008; Tadmor & Tidor, 2005), no previous studies have tried to

understand different user groups with the ultimate goal of informing consistent user experiences across biomedical and life sciences research communities.

The aim of this study was to examine scientists' behavioral and cognition characteristics, as well as how they work with various kinds of knowledge resources to better understand scientists' insight generation behavior and human performance associated with individual differences (i.e., research roles and cognitive styles). The results from this study can help us to better support different user groups in data-intensive cross-disciplinary research circumstances.

We extended an insight-based methodology proposed by P. Saraiya (2005) by integrating complementary eye-tracking methods as well as a gaze-cued retrospective think-aloud protocols. This approach offered a deep understanding of cognitive decision points by examining an individual's behavioral characteristics, insight characteristics, gaze characteristics, and human performance while interacting with different types of knowledge resources. Detailed information about the insight-based method is provided below in Section 3.3.1.

3.2 Research Background

3.2.1 Extended insight-based method

Biology has been transformed into a digitalized and computer-centric science (Hunter, Apweiler, & Martin, 2010), resulting in vast amount of biological data currently available and a concomitant suite of bioinformatics tools (e.g., sequence data, gene expression data, protein-protein interaction data, pathway data) (Karasavvas, Baldock, & Burger, 2004). Along with increasing complexity of new datasets and tools, several user studies have tried to uncover what scientists' actual analytical processes and strategies are and how they need to be better supported (Mirel, 2009; P. Saraiya, 2005; Tran, Dubay, Gorman, & Hersh, 2004). Of specific interest is an insight-based method by P. Saraiya (2005) that assesses how "visualization tools" are used to extract biological insights and generate hypotheses in bioinformatics. This approach has been widely adopted to quantify insights gained from actual exploratory use of visualizations (Boyandin, Bertini, & Lalanne, 2012; Prabhat, Forsberg, Katzourin, Wharton, & Slater, 2008; Purvi Saraiya, North, Lam, & Duca, 2006; Smuc et al., 2009).

P. Saraiya (2005) defines an insight as "an individual observation about the data by the

participant, a unit of discovery." and examines the insights available in specific resources (e.g., the insights available in a specific visualization tool). It should be noted, however, that data-intensive research processes are rarely isolated to single tools (Plaisant, Fekete, & Grinstein, 2008). Instead, scientists use a plurality of resources sporadically and expect coherent outcomes from different processes employing multiple resources (Bartlett & Neugebauer, 2008; Bartlett & Toms, 2005).

Our approach is different from the previous insight-based method works in two respects. First, P. Saraiya (2005) measured insights gained from visualization tools. However, scientists can perceive information representations or use available interaction mechanisms in different ways, suggesting that there exists a wide range of combinations of knowledge resource presentation formats and mechanisms of user interaction that may be used to generate insights. Thus, our study extended the scope of insight drivers from visualization tools to applicable knowledge resources and interaction in computer-mediated environments. Second, we applied eye-tracking to examine insight generation behavior and identify most often and/or preferred information representation (i.e., data formats/presentation of insight drivers) derived from different kinds of knowledge resources in bioinformatics.

3.2.2 Individual differences

Understanding individual differences allows us to gain insights to predict human performance with complex data sets and huge numbers of heterogeneous resources in a big data environment. (Cegarra & Hoc, 2006; Dillon & Watson, 1996), which in turn may better inform interface designs to accommodate different user groups. Traditional studies on individual differences primarily focus on different levels of expertise, mainly comparing task performance between novices and experts (Cellier, 1997; Chase & Simon, 1973) and how this performance gap can be reduced by training (Chase & Simon, 1973). However, Cegarra and Hoc (2006) argue that individual differences are not exclusively restricted to expertise effects and may include for instance, diagnosis strategies (Duncan, 1985; Moran, 1986) and design strategies (Visserl, Hocz, & Chesnay, 1990).

3.2.2.1 Research discipline

As discussed in Chapter 1, bioinformatics is an interdisciplinary research area that consists of

various groups of professionals who have different types of research backgrounds and represent different perspectives of research motivations, interests, and behaviors (Aniba & Thompson, 2010; Letondal & Mackay, 2004; Tadmor & Tidor, 2005). Notably, there is an increasing demand for accommodating multiple diverse user population such as, consistent user experience for both 'wet' (lab-based) and 'dry' (computational) research communities. In light of this, online bioinformatics resources should allow interdisciplinary research communities to effectively utilize and coordinate information available to them. Thus, we considered research discipline as an indicator to address distinct characteristics among scientists.

3.2.2.2 Cognitive style

Cognitive styles can be used to explain individual differences, complementary to expertise (Cegarra & Hoc, 2006), examining *how* or *in what way* individuals process information (Runco & Pritzker, 1999). Thus, understanding cognitive styles could help not only predict scientists' performance but also to explain variability in performance between scientists' at a similar expertise level (Cegarra & Hoc, 2006).

Among different cognitive styles, field dependency has emerged as one of the most widely studied because it addresses how people perceive information and solve problems (Keen & Morton, 1978; Herman A Witkin, Moore, Goodenough, & Cox, 1977). *Field dependency* can be defined as "the degree to which an individual's perception or comprehension of information is influenced by the surrounding perceptual or contextual fields" (Jonassen & Grabowski, 2012). The degree of field dependency is described as a continuum running from extreme field-dependence (FD) to extreme field-independence (FI). FD individuals are externally directed and easily influenced by salient features. They are more likely to perceive surroundings in a relatively global fashion. Also they are attentive to social cues and, in general, are oriented towards other people. FI individuals, on the other hand, are internally directed and process information with their own structure. They experience surroundings analytically, and are able to perceive objects as separate and discrete from their background. They are not very interested in other's opinions, and show a preference for nonsocial situations (Herman A Witkin, 1973).

The Group Embedded Figure test (GEFT) has been theoretically and empirically studied to assess the cognitive styles of field dependence/independence and it has satisfactory reliability

(.89 on test-retest over a three year period) and validity of the test (a correlation of .82 between the two major sub-sections) (Herman A Witkin, 1971).

Previous studies point out that cognitive style differences are important when individuals face new and unfamiliar tasks that demand a specialized response (H. A. Witkin, 1981; Herman A Witkin et al., 1977). For example, field dependent individuals make more use of information provided by other resources when the situation is ambiguous whereas field dependent individuals tend to be less influenced by others individuals or information under such conditions. In other contexts there may be no significant differences between field-dependent and field-independent individuals external social responses, such as in well-structured situations (Herman A. Witkin & Goodenough, 1977).

Arguably, many data-intensive bioinformatics practices represent examples of uninvestigated areas of biological research. Thus, understanding cognitive style will be worthwhile to help understand performance differences in cases where scientists have both similar backgrounds and levels of expertise.

3.2.3 Mixed methods approach

We combined an eye-tracking methodology with concurrent think-aloud to generate objective, quantitative data to confirm validity and reliability issues in concurrently gathered verbal data. Eye movement data is a rich resource for understanding cognitive processes employed during insight generation such as what specific types of insight drivers are attended to and for how long (Rayner, 1998). Also, we could identify participants strategies, decisions and actions when they do not verbalize their thoughts (Elling, Lentz, & De Jong, 2012).

Since eye-tracking data offers answers to questions such as *what* and *when* users visually attend to certain user interface elements, but not *why* they attend them, we also applied a gaze-cued think-aloud protocol. Previous studies have revealed that cued verbalizations produce qualitatively rich data in the form of detailed descriptions of how steps are performed, as well as metacognitive comments on knowledge, actions, and strategies. (Guan, Lee, Cuddihy, & Ramey, 2006; Hansen, 1991; Kuusela & Paul, 2000; van Gog, Paas, van Merriënboer, & Witte, 2005). Though retrospective protocols reflect the memory traces of task performance processes that are retrieved from short-term memory (in tasks of very short duration) or long-term memory directly

after a task is completed, participants are likely to infer and generate retrospective reports based on information provided in questions. In this regard, retrospective data is less accurate than concurrent data (Camps, 2003; Ericsson & Simon, 1993). Conversely, concurrent protocols, by their nature, reflect the information available in short-term memory during task performance. According to Taylor and Dionne (2000), verbal data from both the concurrent and retrospective protocols improve the completeness, reliability and validity of the data (Taylor & Dionne, 2000). Thus, we adopted both the concurrent and retrospective gaze-cued think-aloud to compensate challenges of each method, thereby enhancing the quality of verbalization data.

To the best of our knowledge, very few studies have used a mixed methods approach of qualitative and quantitative methods in the context of biomedical and life sciences and none have addressed the links between cognitive style and insight drivers as detailed in this present research. Studies to date have focused on usability inspections (Anderson, Ash, & Tarczy-Hornoch, 2007; Bartlett & Toms, 2005; Bolchini, 2009; Javahery, 2004; Mirel, 2007, 2009; Tran et al., 2004) – not cognitive perspectives of users captured during interactions with online bioinformatics resources.

3.3 Methods

3.3.1 Experimental design

The study was designed as a single-factor, between-subject experimental design. The independent variable was main research field, which had two levels: *bench scientists* (those mainly conduct experiments in a laboratory) and *application scientists* (those implementing computational algorithms and building/using software to analyze big data sets) (see section 1.1.3. for our working definitions of these terms). There were several dependent variables representing scientists' behavioral characteristics including: average time to first insight, number of resources and pages that participants visit, number of resources and pages where participants find insights, and information representation of insight drivers), insight-generation behaviors (characteristics of insights, a set of gaze behaviors, and information representation of insight drivers), and human errors (slips, lapses, and mistakes). Details of the dependent variables are described in Section 3.4.5. We also considered the participants' cognitive styles as a dependent variable.

3.3.2 Participants

We used purposive sampling (Cohen, Manion, & Morrison, 2000; Patton, 2005; Silverman, 2009) that involves the conscious selection by researchers to serve a very specific need or purpose in the study (Davies & Crookes, 1998). Using email solicitations and fliers posted on campus, we recruited 18 participants (5 males and 13 females) to include ten bench scientists and eight application scientists who are currently involved in life science research at Virginia Tech,. The study population was purposefully limited to biological scientists, have more than 3 years of research experience, and use online bioinformatics resources regularly as part of their research. We asked potential participants a set of pre-screening questions via email to determine their eligibility for, and interest in our study. To be specific, we asked participants to identify their experience levels (years of research experience) in biological sciences, and to specify all the research role options (from a list) that currently applied. We considered self-identified research roles as a key indicator to distinguish between bench scientists and application scientists. If a participant only marked their research role as a biologist, we considered them a bench scientist who mainly conducts *in vivo* or *in vitro* experiments. If a participant marked several research roles related to biological sciences, we regarded them as an application scientist who is more focused on the design and implementation of processing methodologies for big data sets. Since the eye tracker failed to capture eye movements, we lost the data of one of the application scientists. Also, one application scientist was eliminated because of difficulties with his verbalization. Therefore, we report results from nine bench scientists and seven application scientists. Participants self-reported normal or corrected-to-normal vision and normal color vision to ensure consistent eye-tracking data collection. Participants were proficient in the use of bioinformatics resources and had used online bioinformatics resources regularly as part of their work. We collected participants' demographic information using a pre-experiment questionnaire. The mean ages (SD) of participants was 29.4 (4.08) years. All participants had postgraduate degrees, were currently involved in biological science research, and represented a relatively core group of bioinformatics users. Over three-quarters of participants (77.8%) have used online bioinformatics resources more than once a week. A summary of participant demographics is presented in Section 3.5.1. Apparatus

We set up the experiment on a desktop computer with a 22" LCD monitor (resolution of 1600 × 1024). Participants used a mouse and keyboard during the experiment to navigate online bioinformatics resources via the Internet Explorer web browser. We used an unobtrusive eye tracking device, the Tobii X2-60, to capture gaze behaviors for subsequent insight generation analysis. The sampling rate of the Tobii X2-60 eye tracker is 60 Hz.

To conduct the gaze-cued retrospective think aloud session, we used the Tobii Studio 3.2 application software. The moderator observed participants' gaze paths on another monitor.

3.3.3 Procedures

We obtained Virginia Tech Institutional Review Board (IRB) approval prior to any data collection. The experiment had four sessions: pre-questionnaires, a task performance session with a concurrent think aloud (CTA), a gaze-cued retrospective think aloud (RTA), and a post questionnaire. We compensated participants $10 per hour for a maximum of 2 hours. Participants completed an informed consent procedure approved by the Virginia Tech Institutional review Board (See Appendix C and Appendix D).
Participants completed the pre-questionnaire (See Appendix E) on demographics such as age, gender, ethnicity, research background and experience, and previous bioinformatics experience. Also participants took the GEFT that designed to identify the field-dependence-independence level. The GEFT consists of three sections. Section 1 is intended for practice purpose. Section 1 contains seven relatively easy forms, and requested participants to finish items within 2 minutes. Both sections 2 and 3 consisted of nine questions and required participants to complete each of them within 5 minutes.

After administering the pre-questionnaire and GEFT, we tested participants to determine whether our eye tracking equipment could accurately measure gaze behaviors (If not, we ended the study). If the eye calibration succeeded, we explained to participants the basic concept of concurrent think aloud (CTA) and asked them to apply these concepts in a training session. Before working on main tasks, we asked participants to perform training tasks for approximately 10 minutes to get acquainted with the concurrent think aloud protocol. We used the Pathosystems Resource Integration Center (PATRIC) online bioinformatics resource as a starting

point of the main task and a reference site for participants. PATRIC is a Virginia Tech-build portal of biological information designed to support bacterial research communities.

During this training session, participants were asked to explore functions of resources and to verbalize their thoughts freely. When participants felt comfortable enough with the procedure, we finished the training session and gave them task instructions for the main task.

Participants' main task was to characterize a given gene by using the PATRIC website and any other available online resources within the time constraints of the study. At the same time, we encouraged participants to think aloud so that their concurrent verbalization could be recorded during the task performance. The Tobii X2-60 eye tracker recorded gaze behaviors on the computer screen. We asked participant to press the Escape key when they found enough insights to end the main task session.

Following the main task with concurrent think-aloud session, we played back the entire video of their task completion, with gaze data superimposed on top of the video (See Figure 6).

Figure 6. Screen-shot of gaze behaviors from a sample participant.
(screen-captured by the researcher, used with permission of the participant)

During playback, we instructed participants to verbalize what they found and why they engaged with specific information elements while conducting the task. Participants' verbalizations during the retrospective think-aloud sessions were recorded using an audio recorder. Frequently, we asked participants to prove additional detail in order to clarify a verbal comments captured during main task performance. After they completed the gaze-cued retrospective think-aloud, participants filled out a post-questionnaire to capture overall experience levels of participants with other resources.

3.3.3.1 Simulated task

We designed the main task (i.e., characterize a specific gene) so that it 1) embodied the open-ended nature of scientists' real-world practices but also 2) was abstracted away from any particular domain to increase generalizability of the results (Huang, 2013).

The main task allowed scientists to generate new findings/insights and to make interactions with different types of knowledge representations. To be specific, we asked participants to characterize functions of a specific gene (the ADD gene in Escherichia coli str. K-12 substr. MG 1655) to do best of their ability using all available resources. The instructions were:

> *I'm interested in understanding the process you follow to try to find out what the functions of a gene are. Please characterize functions of the ADD gene in Escherichia coli str. K-12 substr. MG1655 in as much detail as possible.*
>
> *Please look for as many different types of relevant data as possible using any available online resources using a computer. Please keep constantly talking about 1) what steps you follow, 2) why you choose specific information and 3) what insights you get until the end of the task.*

3.3.4 Measures

Dependent variables included a set of cognitive styles, behavioral characteristics, insight characteristics, gaze behaviors, and types of human error. The majority of the dependent measures were collected during the main task. However, cognitive styles were collected prior to the task via the GEFT.

Cognitive styles:

We identified cognitive styles of participants using the Group Embedded Figure test (GEFT) to assess the degree to which participants were field-independent or field-dependent thinkers. Since the section 1 of the Group Embedded Figures Test (GEFT) was a practice section, the primary researcher scored only section 2 and 3 of the GEFT to determine the degree of field dependence or independence.

Behavioral characteristics:

We logged the web addresses (uniform resource locators; URLs) of online resources participants visited during the session along with timestamps. Measures of behavioral characteristics included: 1) average time to first insight, 2) the number of resources visited (e.g., databases, scholarly journals, search engines), 3) the number of pages viewed (on resources), 4) the number of relevant resources retrieved by each participant, and 5) the number of relevant pages retrieved by each participant and information representation of insight drivers (e.g., text, visualization).

Relevant resources and pages were determined by participants during the main task; we did not restrict or dictate which online resources should be used to complete the main task.

Insight characteristics:

We defined an insight as an individual observation about the gene by the participant, a unit of discovery (Purvi Saraiya et al., 2006). We considered any domain-specific observation that participants' mention during the concurrent think aloud as an insight occurrence.

The insight measures included a set of four dimensions: number of insights (insight count), domain value of insight, degree of insight depth, and correctness of insight (P. Saraiya, 2005). More specifically, we defined the *number of insights* as the count of the actual findings by each participant. The *domain value* referred to "the value, importance, or significance of the insights" and was coded using 5-point scale (ranging from 1 or 2 = trivial observations, 3 = insights about a particular processes, to 4 or 5 = confirm, deny, or create a hypotheses). The *degree of insight depth* was obtained using a 5-point coding scale ranging from an overview of biological processes (1 = breath insight) to focused and detailed insight (5 = depth insight) for the given

specific gene. The *correctness of insight* metric quantified whether the observed insight is correct, partially correct or incorrect.

Before characterizing each insight, two independent coders transcribed participants' verbalizations and identified insights that participants discovered. Two *domain experts* then evaluated each insight based on the decision rules developed by P. Saraiya (2005).

Gaze characteristics:

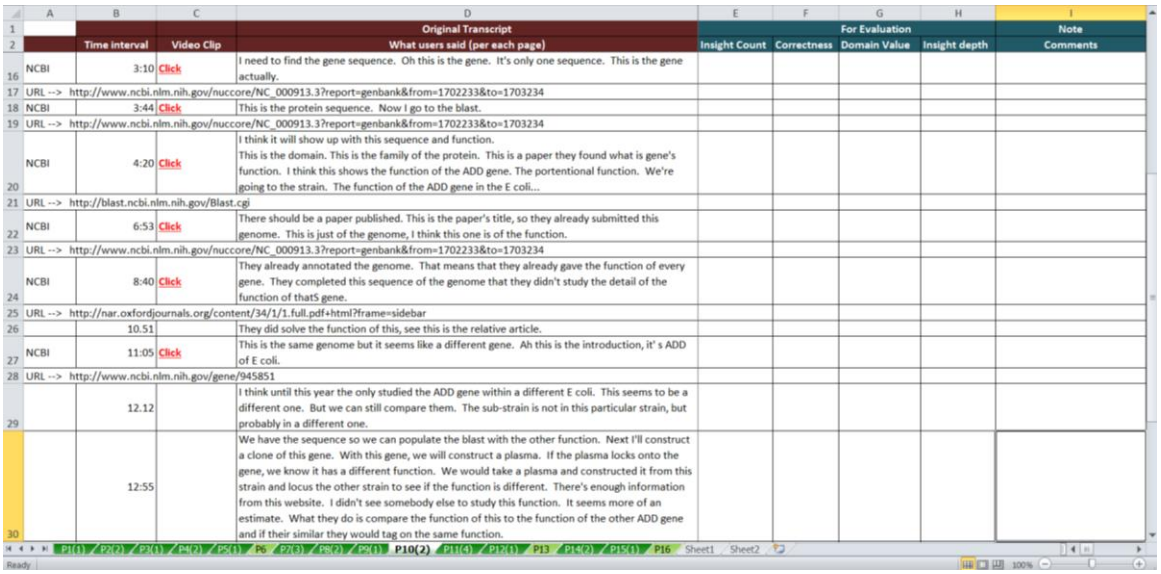
Fixations are the most relevant metric for identify information processing (Marcel A Just & Carpenter, 1980; Rayner, 1978). Thus, we relied on gaze fixations with a minimum threshold of 100 ms in areas of interest during the main task (Cutrell & Guan, 2007). Gaze characteristics included the number of fixation and fixation duration on area of interests (AOIs) related to the insights. According to wickens' multiple resource theory (Wickens & Liu, 1988), visual component can be coded as either a verbal or a spatial representation. Hence, we considered the text-based data and visually-based data as area of interests.

Human Errors:

In this study, human errors were classified according to the widely accepted Reason's error type: slips, lapses, and mistakes. *Slips* can be observable but unintended error. Reason (1995) uses the term 'slip' to refer to an observable but unintended error where there is no conscious control, even though the initial plan was correct. To be specific, some participants failed to find insights during the CTA. But they then realized later during the gaze-cued RTA that they had failed to observe meaningful information during the main task. A *lapse* is simply forgetting to do something. An example of a common lapse was that some participants dropped an activity for a moment to make judgments and assessments on data and strategies, but then failed to remember to pick the activity back up. Then they remembered what they intended to do, but did not, during the RTA. *Mistakes* are planning failures. For example, when a participant chose an online resource but it did not lead to the desired results. Or participants misinterpreted information and/or made poor decisions. Consequently, the insights they found were incorrect. In this research, mistakes were observed by domain experts, but slips and lapses were identified by participants. Slips are more likely to be found among scientists who are more experienced (Norman, 1981).

3.3.5 Analysis

Two independent coders transcribed participants' verbalizations during the CTA (Figure 7) and the gaze-cued RTA (Figure 8), and identified insights that participants discovered.



		Time interval	Video Clip	Original Transcript What users said (per each page)	For Evaluation Insight Count	Correctness	Domain Value	Insight depth	Note Comments
16	NCBI	3:10	Click	I need to find the gene sequence. Oh this is the gene. It's only one sequence. This is the gene actually.					
17	URL --> http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3?report=genbank&from=1702233&to=1703234								
18	NCBI	3:44	Click	This is the protein sequence. Now I go to the blast.					
19	URL --> http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3?report=genbank&from=1702233&to=1703234								
20	NCBI	4:20	Click	I think it will show up with this sequence and function. This is the domain. This is the family of the protein. This is a paper they found what is gene's function. I think this shows the function of the ADD gene. The portentional function. We're going to the strain. The function of the ADD gene in the E coli...					
21	URL --> http://blast.ncbi.nlm.nih.gov/Blast.cgi								
22	NCBI	6:53	Click	There should be a paper published. This is the paper's title, so they already submitted this genome. This is just of the genome, I think this one is of the function.					
23	URL --> http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3?report=genbank&from=1702233&to=1703234								
24	NCBI	8:40	Click	They already annotated the genome. That means that they already gave the function of every gene. They completed this sequence of the genome that they didn't study the detail of the function of thatS gene.					
25	URL --> http://nar.oxfordjournals.org/content/34/1/1.full.pdf+html?frame=sidebar								
26		10.51		They did solve the function of this, see this is the relative article.					
27	NCBI	11:05	Click	This is the same genome but it seems like a different gene. Ah this is the introduction, it' s ADD of E coli.					
28	URL --> http://www.ncbi.nlm.nih.gov/gene/945851								
29		12.12		I think until this year the only studied the ADD gene within a different E coli. This seems to be a different one. But we can still compare them. The sub-strain is not in this particular strain, but probably in a different one.					
30		12:55		We have the sequence so we can populate the blast with the other function. Next I'll construct a clone of this gene. With this gene, we will construct a plasma. If the plasma locks onto the gene, we know it has a different function. We would take a plasma and constructed it from this strain and locus the other strain to see if the function is different. There's enough information from this website. I didn't see somebody else to study this function. It seems more of an estimate. What they do is compare the function of this to the function of the other ADD gene and if their similar they would tag on the same function.					

Figure 7. A sample of transcriptions from the CTA.

		Time interval	Original Transcript What users said	Insight Evaluation Correctness	Human Error Slip, Lapse, Mistake	Note Comments
27	PATRIC Pathways	13:42	So I found it. I'm quite proud of myself. Now what can I do with it? So I'm already on the gene tab and looking for more specific information such as publications. Maybe if I go to Locus Tab it will give me such.	Incorrect	Mistake	
28	PATRIC Overview	14:21	Surely they made locus tab when they found it which should be publicized. Basic information about it and don't remember these being locations that NCBI gave me. I'm not overly concerned because my focus is the nucleotide length. Where it is. I'm deciding what else I can go after and look for. Don't know what that is, Fig Fam. Here I'm very happy about that request thing. So I decided genotology which isn't something I go to and haven't been there in a long time.			
29	Geneontology	15:15	So I'm not sure what I'm going to get when I go here. So now I'm just reading what the website has. Prokaryotic, that matches up, very good. I'm trying to recall the relationships that are on gerontology. Not sure where these two are. Like these are the children. One makes these children or is it referring to what it does? Is it messing with RNA? I don't think so. So maybe Kegg pathways have something about that.	Correct	Lapse	
30						
			So here I am at pathways and there it is. So now I try and get a feel for what its purpose is of the organism by looking at the pathway around it. Further upstream and downstream			

Figure 8. A sample of transcriptions from the gaze-cued RTA.

Cognitive Styles:

If the simple figure is correctly outlined within the complex figure, we considered it as "correct". Figure 9 shows a sample image from the GEFT.

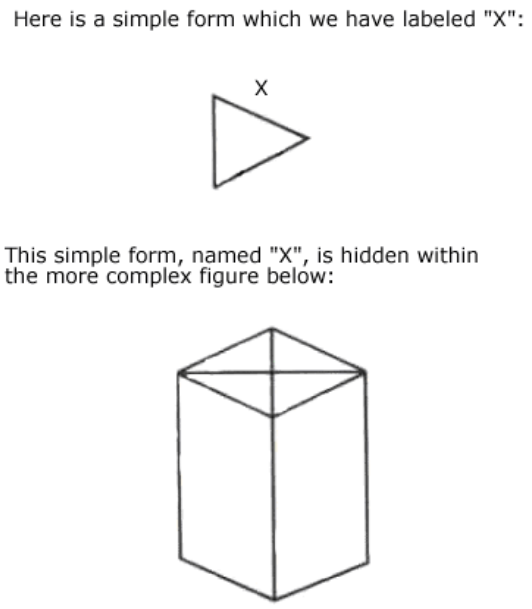


Figure 9. Sample image from the GEFT.

The GEFT is scored from 0 (low) to 18 (high). In this study, the degree of field dependence or independence was described as a continuum between 0 and 18. Higher scores suggest higher degrees of field-dependence while lower scores suggest relatively field dependence. Cognitive style is scored along a continuum; we refer to participants as either field-independent researchers or field-dependent scientists as a matter of convenience. In reality, individuals are, for example "more field dependent than field-independent" or similarly "more field-independent than field-dependent".

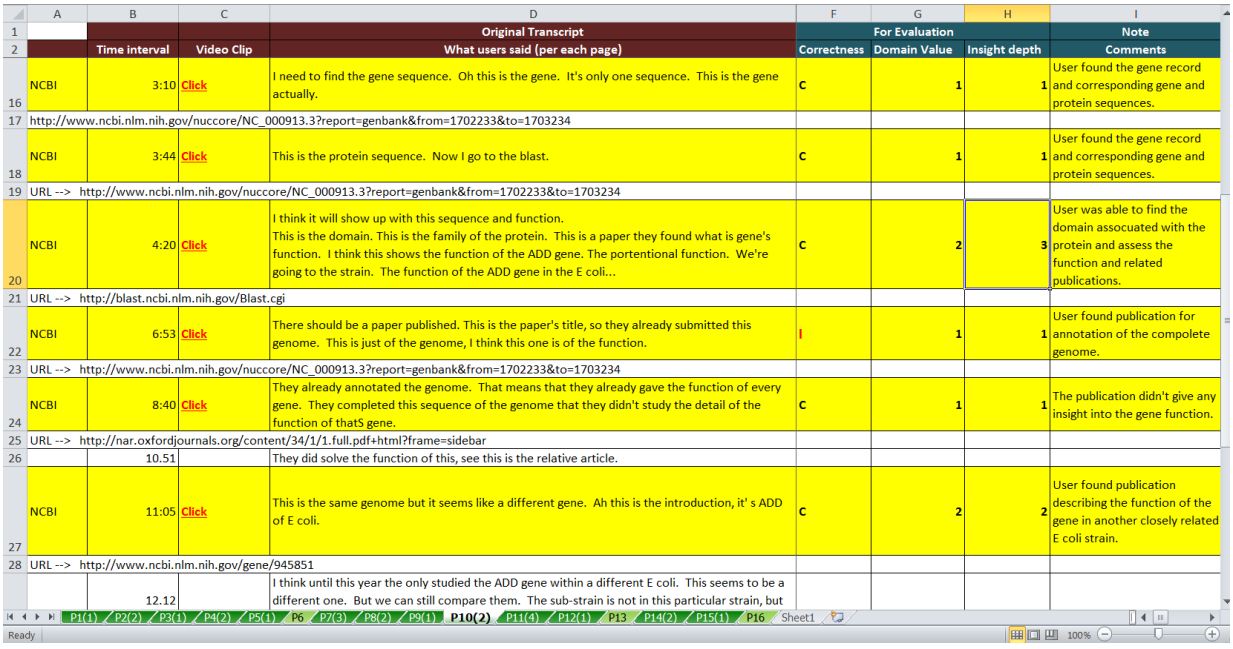
Behavioral characteristics:

In order to understand participant's behavior during the insight generation process, we identified the average time to first insight, the number of resources visited, the number of pages viewed, the number of relevant resources and pages retrieved by each participant, and information representation of insight drivers.

Insight characteristics:

In order to evaluate insights, we adopted the decision rules developed by P. Saraiya (2005). To be specific, two domain experts carefully reviewed the video clips (from the CTA eye-tracked main task trials) and the transcribing notes to establish a global, overall understanding of the breadth and depth of all participants' responses. Then, the domain experts with over 10 years research experiences in biological sciences evaluated the domain value, the degree of insight depth, and correctness independently based on pre-defined decision rules, compared the evaluation results, and discussed/reconciled any disagreements in scoring (see Appendix F for the complete insight evaluation manual).

Inter-rater reliability was assessed by the Kappa. The initial Kappa value was 0.73. Thus, we refined and clarified the decision rules, and evaluated the insights again through a reconciliation meeting. Finally, we achieved 96.7% agreement, which is above the accepted standard of 75% suggested by (Fleiss, Levin, & Paik, 1981).



	A	B	C	D	F	G	H	I
1				Original Transcript	For Evaluation			Note
2		Time interval	Video Clip	What users said (per each page)	Correctness	Domain Value	Insight depth	Comments
16	NCBI	3:10	Click	I need to find the gene sequence. Oh this is the gene. It's only one sequence. This is the gene actually.	C	1	1	User found the gene record and corresponding gene and protein sequences.
17	http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3?report=genbank&from=1702233&to=1703234							
18	NCBI	3:44	Click	This is the protein sequence. Now I go to the blast.	C	1	1	User found the gene record and corresponding gene and protein sequences.
19	URL --> http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3?report=genbank&from=1702233&to=1703234							
20	NCBI	4:20	Click	I think it will show up with this sequence and function. This is the domain. This is the family of the protein. This is a paper they found what is gene's function. I think this shows the function of the ADD gene. The portentional function. We're going to the strain. The function of the ADD gene in the E coli...	C	2	3	User was able to find the domain assocuated with the protein and assess the function and related publications.
21	URL --> http://blast.ncbi.nlm.nih.gov/Blast.cgi							
22	NCBI	6:53	Click	There should be a paper published. This is the paper's title, so they already submitted this genome. This is just of the genome, I think this one is of the function.	I	1	1	User found publication for annotation of the compolete genome.
23	URL --> http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3?report=genbank&from=1702233&to=1703234							
24	NCBI	8:40	Click	They already annotated the genome. That means that they already gave the function of every gene. They completed this sequence of the genome that they didn't study the detail of the function of thatS gene.	C	1	1	The publication didn't give any insight into the gene function.
25	URL --> http://nar.oxfordjournals.org/content/34/1/1.full.pdf+html?frame=sidebar							
26		10.51		They did solve the function of this, see this is the relative article.				
27	NCBI	11:05	Click	This is the same genome but it seems like a different gene. Ah this is the introduction, it' s ADD of E coli.	C	2	2	User found publication describing the function of the gene in another closely related E coli strain.
28	URL --> http://www.ncbi.nlm.nih.gov/gene/945851							
		12.12		I think until this year the only studied the ADD gene within a different E coli. This seems to be a different one. But we can still compare them. The sub-strain is not in this particular strain, but				

P1(1) P2(2) P3(1) P4(2) P5(1) P6 P7(3) P8(2) P9(1) P10(2) P11(4) P12(1) P13 P14(2) P15(1) P16 Sheet1

Ready 100%

Figure 10. A sample of a domain expert's evaluation and scoring form.

Gaze characteristics:

One of the valuable aspects of gaze behavior is that we can determine areas on the screen from which insights are most likely derived. Using the CTA eye-tracking video, we identified areas of interests (AOIs) (Hyrskykari, Ovaska, Majaranta, Räihä, & Lehtinen, 2008) based on information representation, specifically separating areas that were heavily text-based (e.g., reference papers, gene sequence, taxonomy) (Figure 11) from visually-represented data (e.g., pathway maps, genome browsers, or heat maps of protein families) (Figure 12). For each AOI (across scores of online bioinformatics resources and pages) we calculated common eye-tracking measures such as number of fixations and fixation durations (i.e., the amount of time associated with each fixation)

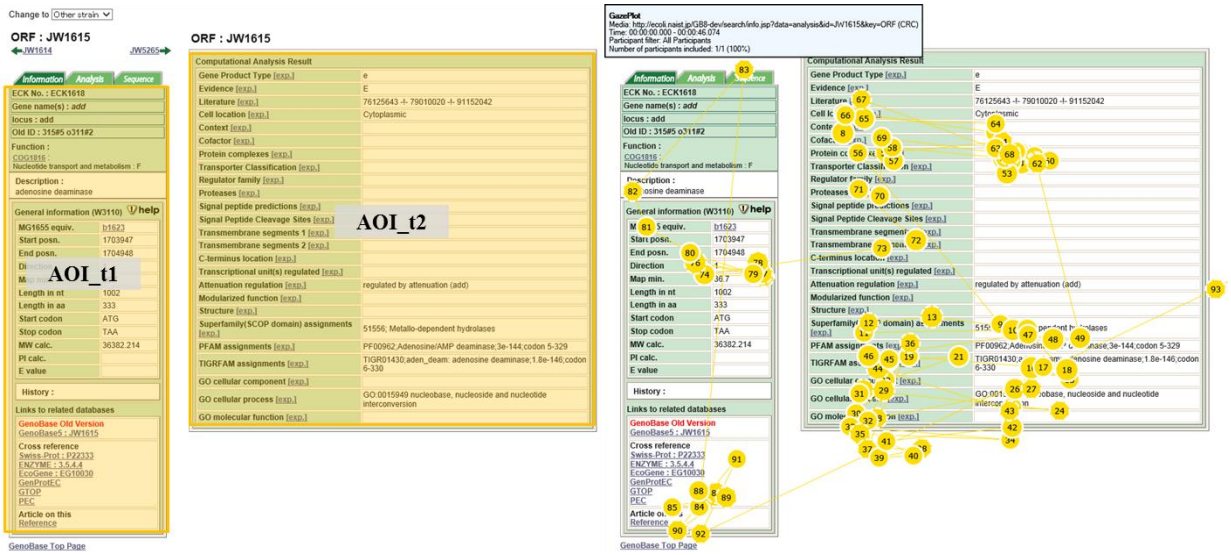


Figure 11. Area of interests on exemplar text-based data
with information relevant for the target gene (left) and individual gaze plots for this text-based data example (right). Note that the size of circles indicates fixation duration with larger circles representing longer fixations; and the numbers in the circles represent the order of the fixations.

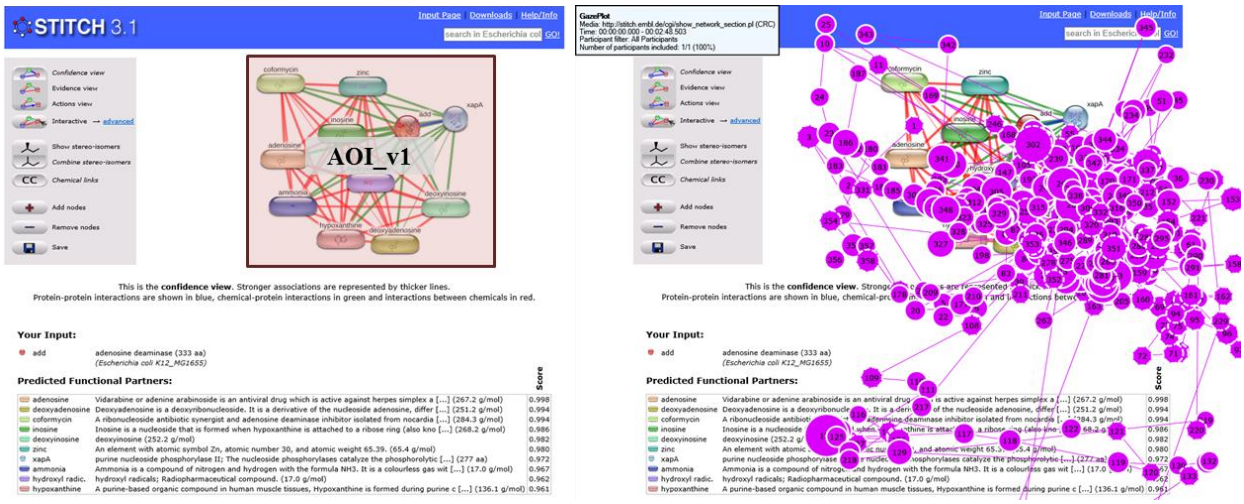


Figure 12. Area of Interest on exemplar visually-represented data.
with information relevant for the target gene (left) and
individual gaze plots for this visualization example (right).

Human Errors:

To assess human performance, we adopted the decision rules that were developed by (Reason, 1990). We observed slips and lapses when participants were asked to verbalize their thought during playback (the gaze-cued RTA). As mentioned earlier, some participants recognized missing insights and acknowledged missed information during the gaze-cued RTA. The following statement is an example of a slip:

> (During the gaze-cued RTA)
> *I didn't see the blue marker for this gene. I think it is meaningful.*

In terms of lapses, participants indicated that they forgot to do something to achieve the goal of the given task. For instance,

> (During the CTA session) *This is a pathway. I will look through it because…..*

> (During the gaze-cued RTA session) *Oh. I missed it (the pathway), the gene name here.*

Also, we identified mistakes through an insight evaluation process. To be specific, participants indicated information about a wrong gene name, incorrect strains, and/or incorrect substrains during both CTA and gaze-cued RTA sessions. For instance,

(During the CTA session) *It has a function here. Here it characterizes this gene.*

(During the gaze-cued RTA session) *I think this was useful. It talks about the inducer in this particular protein.*

Some participants misinterpreted information or were confused about the name of the target gene on specific resources. Cases like this also resulted in mistake. For example,

(During the CTA session) *I didn't think this paper was useful. It gives the function of the other E coli.*

The decision rules are shown in Table 10.

Table 10. Decision rules of human performance.

				Observed insights by participants	
Category		**Process**	**Correctness**	**CTA**	**Gaze-cued RTA**
		Insight	Correct	Correct	Correct
Human Error	**Slip**	e.g., Failure to observe, Inattention, reversal of actions	N/A	N/A	Correct
	Lapse	e.g., Omitting of planned actions, forgetting intended actions	N/A	N/A	Correct
	Mistake	e.g., Making wrong decisions	Incorrect	Correct	Correct

Figure 13 shows our data analysis process.

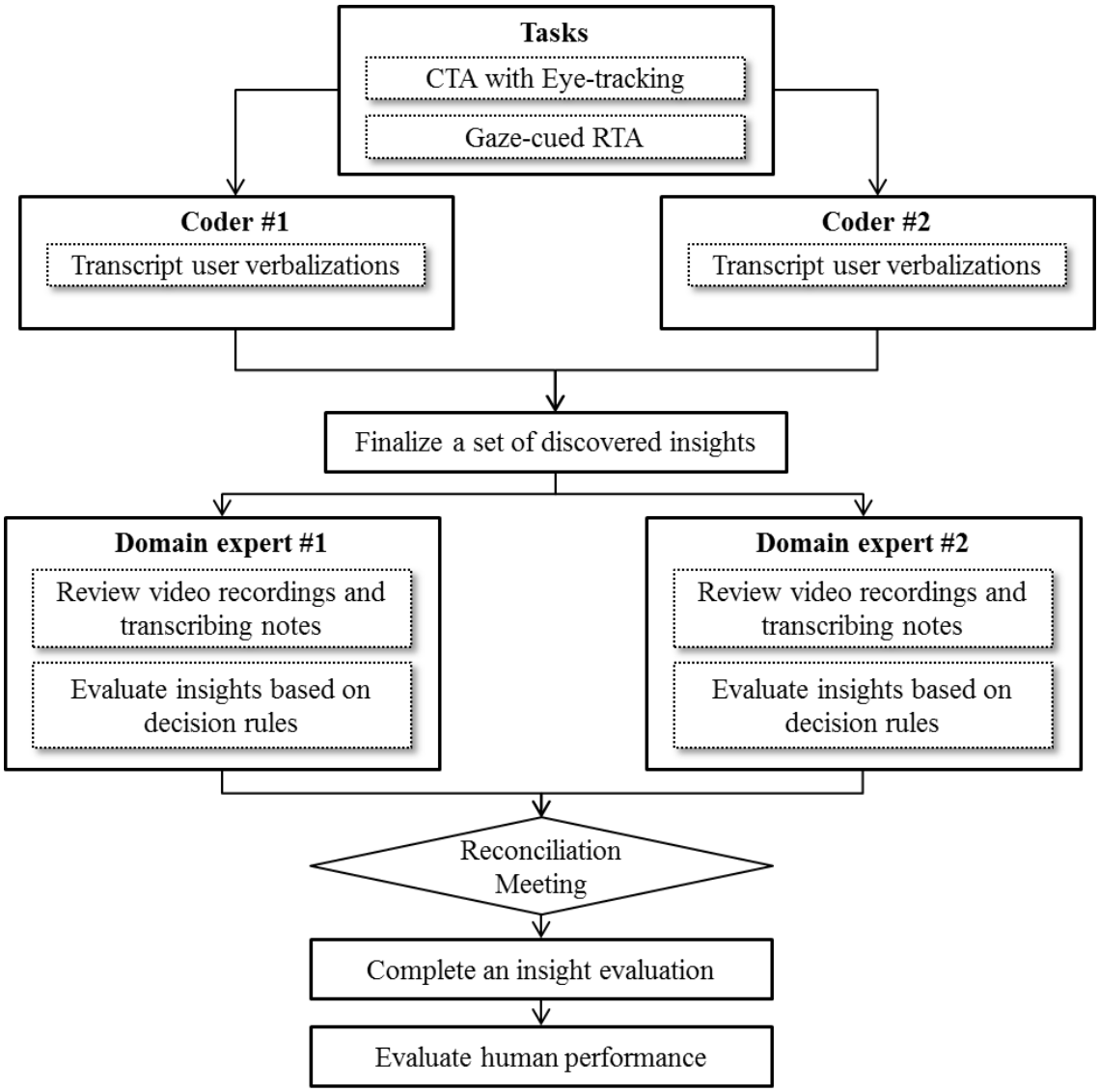


Figure 13. Data analysis process

3.4 Results

Prior to any subsequent statistical tests, we conducted a Shapiro-Wilk test in order to check the normality for all measures. The results revealed that field dependency scores, number of resources that participants used, number of pages that participants viewed, domain value, insight depth, number of fixation and fixation duration on text-based data grouped by research roles were normally distributed. Consequently, we applied parametric data analyses.

However, number of insights that participant reported, first time to insight, number of relevant resources for insights, number of relevant pages for insights, number of correct insights, data formats of insight drivers, number of fixation and fixation duration on visually-represented data, and human errors were not normally distributed. For all variables that did not satisfy the normality assumption, we performed non-parametric data analyses.

We were interested in examining how individual differences affect scientists' insight generation behavior under highly complex and ambiguous research situations. We present results regarding insight generation behavior derived from the eye-tracking with concurrent think aloud, i.e., behavioral characteristics, insight characteristics, and gaze characteristics. Subsequently, we examined human errors participants generated using the gaze-cued retrospective think aloud data.

3.4.1 Descriptive statistics

Participants' demographic information (n=16) is shown in Table 11. Two participants who did not calibrate sufficiently well were excluded from the data analysis. Seventy five percent (n=12) of respondents are female, and 25.0% (n=4) are male, the age of the participants ranged from 24 to 39 (M = 29.4, SD = 4.08). All respondents have experience in conducting biological research and 50% (n=8) of participants have been working in the domain over 5 years (M = 6 years 1 month). Around 77.8% (n=14) of the participants use bioinformatics resources on a daily basis or several times a week. About 87.5% (n=14) of the participants use knowledge networking platforms more than once a week (i.e., SEQanswers, ResearchGate), and 68.8% (n=11) of the participants mainly consume knowledge from the virtual repository, not provide or share knowledge. This passive behavior towards knowledge sharing is consistent with our previous study (See Section 2.7.4).

About 87.5% of participants (*n*=14) mentioned that they have been encouraged to share knowledge with others in their networks. The range of the GEFT scores in this study was 0 to 18, the mean was 10.88, and the standard deviation was 3.44.

Table 11. Demographic analysis

Variables	Categories	*n*	%
Gender	Female	12	75.0%
	Male	4	25.0%
Research Role (all that apply)	Biologist	11	68.8%
	Bioinformatician	4	25.0%
	Computer Scientist	2	12.5%
	Chemist	2	12.5%
	Animal Scientist	1	6.3%
	Biostatistician	1	6.3%
Experience in biological research fields	One year	0	0.0%
	One to five years	8	50.0%
	Five to ten years	4	25.0%
	More than ten years	4	25.0%
Use of bioinformatics tools	Almost everyday	7	43.8%
	More than once a week	9	56.3%
	More than once a month	0	0.0%
	Around once a month	2	12.5%
Use of online knowledge sharing platforms	I have only seen them.	3	18.8%
	Little experience, I have used them only a few times for my work or research.	6	37.5%
	Experienced, I use them regularly to do my work or research.	6	37.5%
	Very experienced, I use them for almost all my work or research.	1	6.3%

3.4.2 Correlations between research roles and cognitive styles

We tried to determine if a relationship exists between research roles (i.e., bench scientists and application scientists) and field dependency. We conducted the point-biserial correlation analysis in order to measure the strength of association between a binary variable (research roles) and a continuous-level variable (cognitive style). The point-biserial correlation coefficient was

significant ($r_{pb(14)}$ = -0.610, $p < .05$). As such, we can state that we observed a strong relationship between research role and field dependency; where bench scientists are significantly more likely to be field independent than application scientists.

3.4.3 The impact of research roles on insight generation

3.4.3.1 Behavioral characteristics

Participants used a wide range of resources including both generic websites (e.g., Google, Wikipedia) and domain-specific online resources (e.g., NCBI PuBmed, Blast, PATRIC, Gene Wiki, Ecoli Wiki) in order to characterize the functions of the ADD gene in Escherichia coli (str. K-12 substr. MG 1655). Nearly all participants (*n*=15) used Google to do a preliminary search.

We were interested in whether bench scientists and application scientists differ in terms of behavioral characteristics. With regard to time to first insight, a Mann-Whitney U-test indicated that there were significant differences between bench scientists and application scientists, *U*= 12.00*; p*= .04. Application scientists took longer to find the first insight as compared to bench scientists (Figure 14).

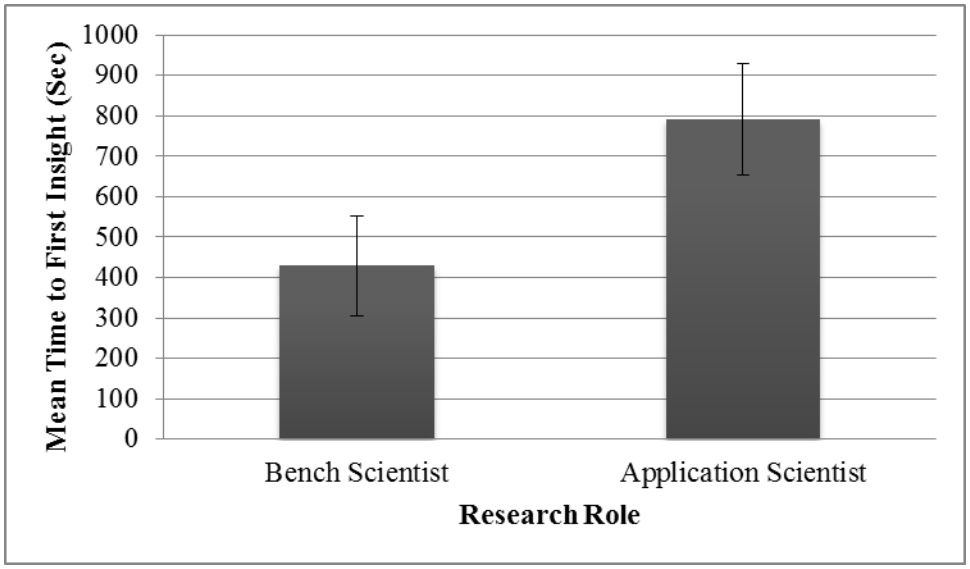


Figure 14. Mean time to first insight by research role.

In terms of information representation of insight drivers, a Mann-Whitney U-test revealed that the use of text-based data format as insight drivers was similar, *U*= 24.50; *p*= .45. But we observed significant differences of research roles on the use of visually represented data (Mann-

Whitney U-test: U= 10.50; p= .02). Specifically, application scientists made more use of visualization information than bench scientists during insight generation processes (Figure 15).

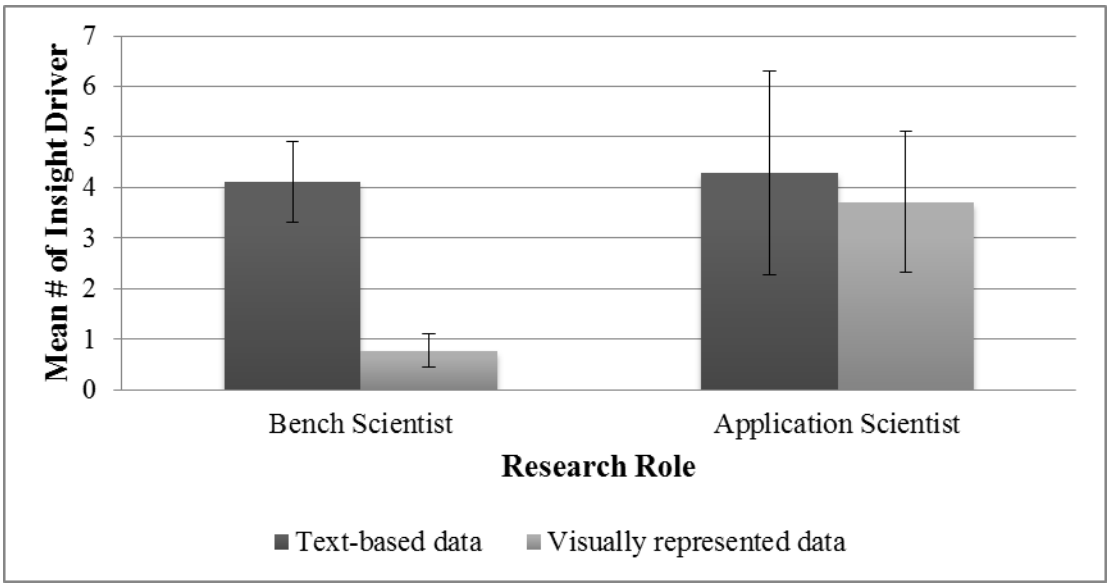


Figure 15. Information representations of insight drivers by research role.

However, the independent samples t-test revealed that there was no significant difference of research roles on the number of resources (bench scientists, M= 5.89, SD= 2.57; application scientists, M= 7.14, SD= 2.12), $t(14) = -1.04$, $p > .05$); (Figure 16) and the number of pages (bench scientists, M= 58.78, SD= 23.13; application scientists, M= 73.00, SD= 23.66), $t(14) = -1.21$, $p > .05$) that were used to find insights; (Figure 17).

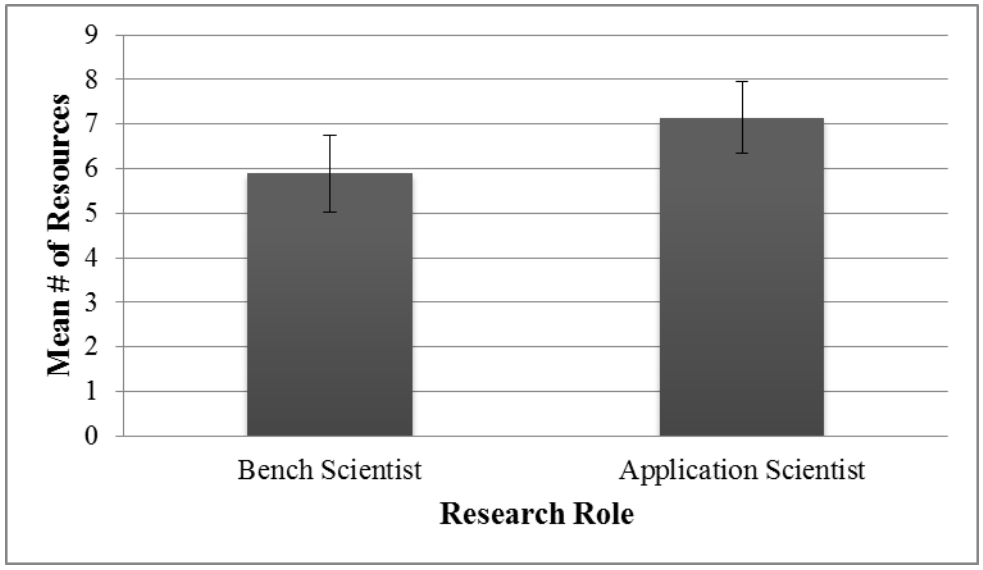


Figure 16. Mean number of resources that were used to explore insights by research role.

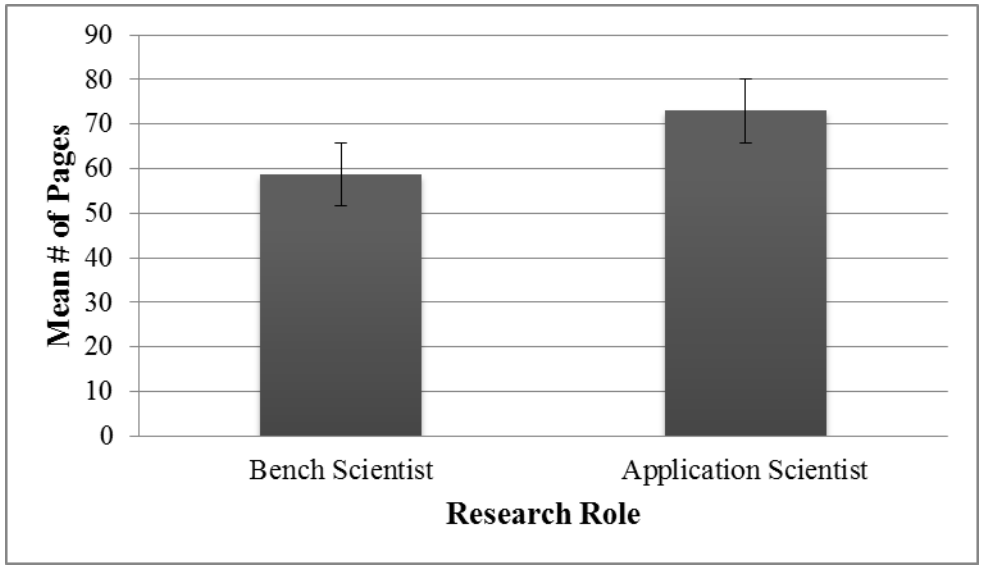


Figure 17. Mean number of pages that were used to explore insights by research role.

With regard to the number of insights that participants reported, there was a marginally significant difference between bench scientists and application scientists (Mann-Whitney U-test: U= 30.00; p= .87), which indicates more observed insights from application scientists regardless of the correctness of the insights (Figure 18).

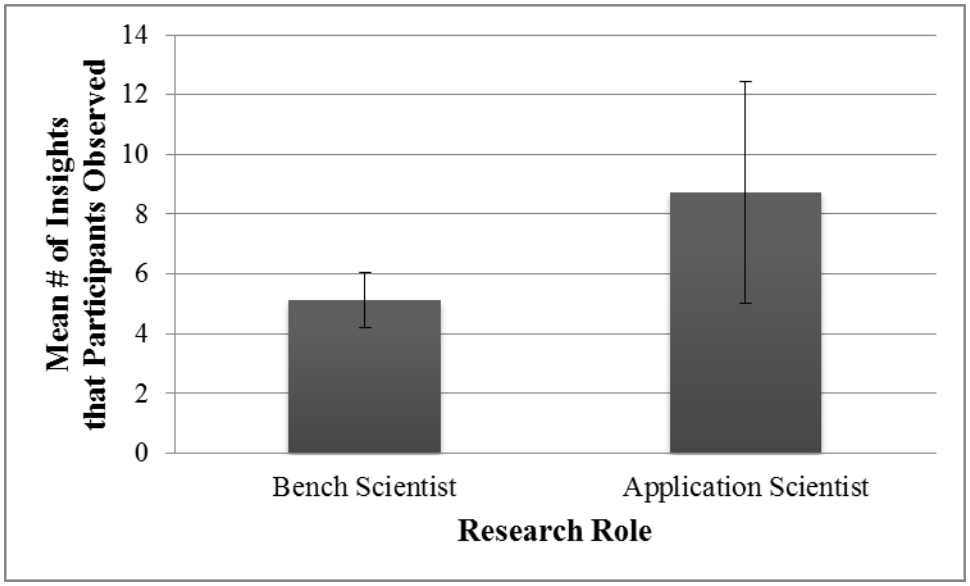


Figure 18. Mean number of insights that were observed by research role.

However, there was no significant difference of research roles on the number of resources (Mann-Whitney U-test: U= 29.50; p= .81) and pages (Mann-Whitney U-test: U= 31.50; p= 1.00) (Figure 19) used related to insights.

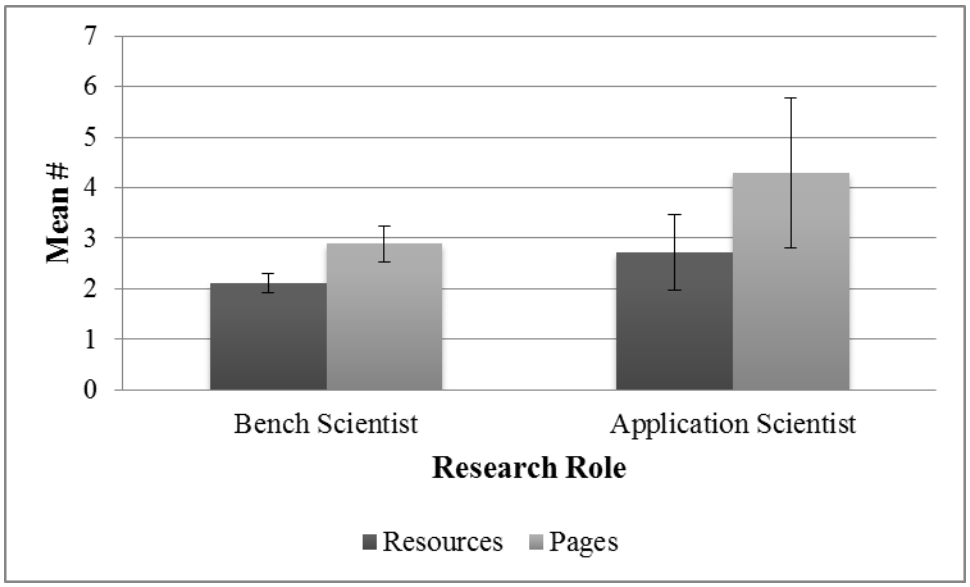


Figure 19. Mean number of resources and pages in respect to insights by research role.

3.4.3.2 Insight characteristics

A Mann-Whitney U test shows a marginally significant difference between bench scientists and application scientists on insight count (U= 29.50; p= .83) (Figure 20).

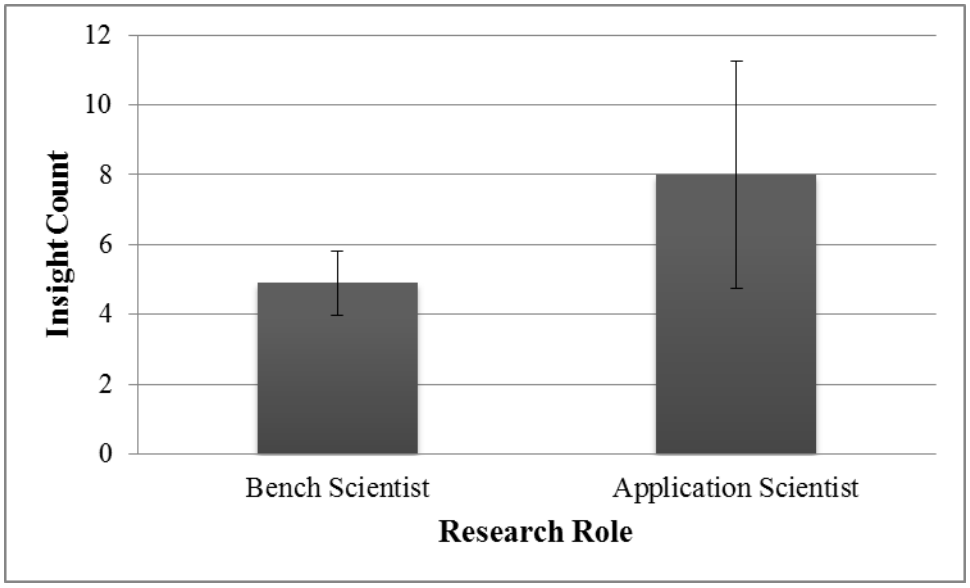


Figure 20. Mean number of correct insights by research role.

However, independent samples t-tests assuming equal variances indicate that there were no significant difference of research roles on the insight domain value (bench scientists, M= 1.23, SD= 0.27; application scientists, M= 1.26, SD= 0.29), $t(14) = -0.17$, $p > .05$) and the degree of depth (bench scientists, M= 1.00, SD= 0.12; application scientists, M= 0.97, SD= 0.10), $t(14) = 0.51$, $p > .05$) of observed insights (Figure 21). Thus, research roles do not affect insight values and depth.

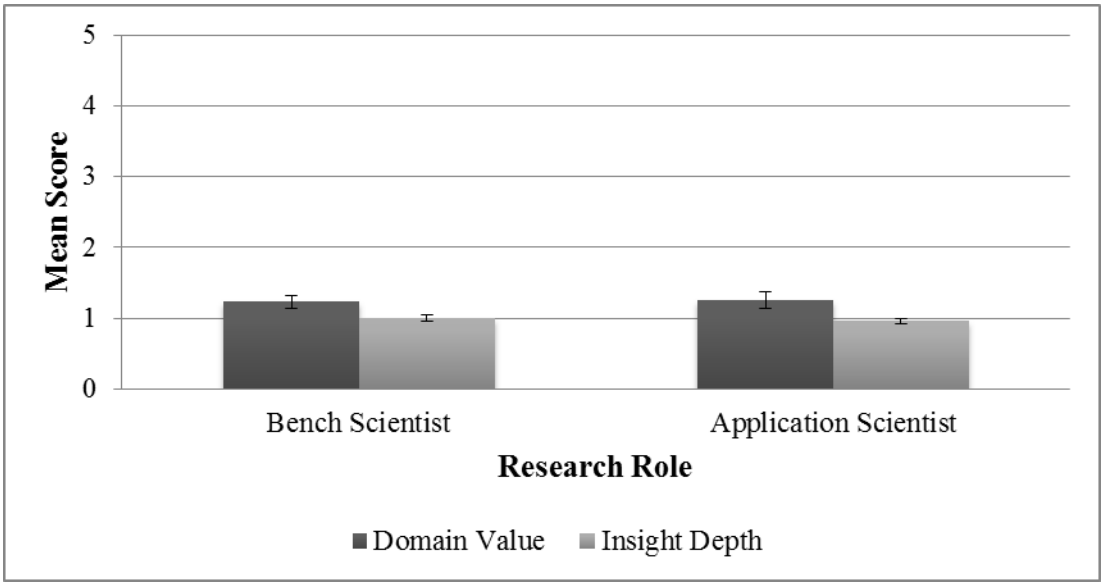


Figure 21. Mean score of domain value and insight depth by research role.

(Domain value: 5=very significant and important, 1= simple and trivial;
Insight depth: 5=more focused and detailed, 1=overview).

3.4.3.3 Gaze characteristics

We analyze the number of fixations and the fixation duration mean on each insight drivers in order to identify relationships between research roles and gaze characteristics on insight drivers. An independent samples t-test found that bench scientists produce more fixations on text-based data (bench scientists, M= 60.24, SD= 12.09; application scientists, M= 33.42, SD= 8.20), $t(14) = 5.02$, $p < .05$). Also, a Mann-Whitney U-test showed that bench scientists generate more fixations on visually-represented data ($U= 0.00$; $p=.00$) (Figure 22).

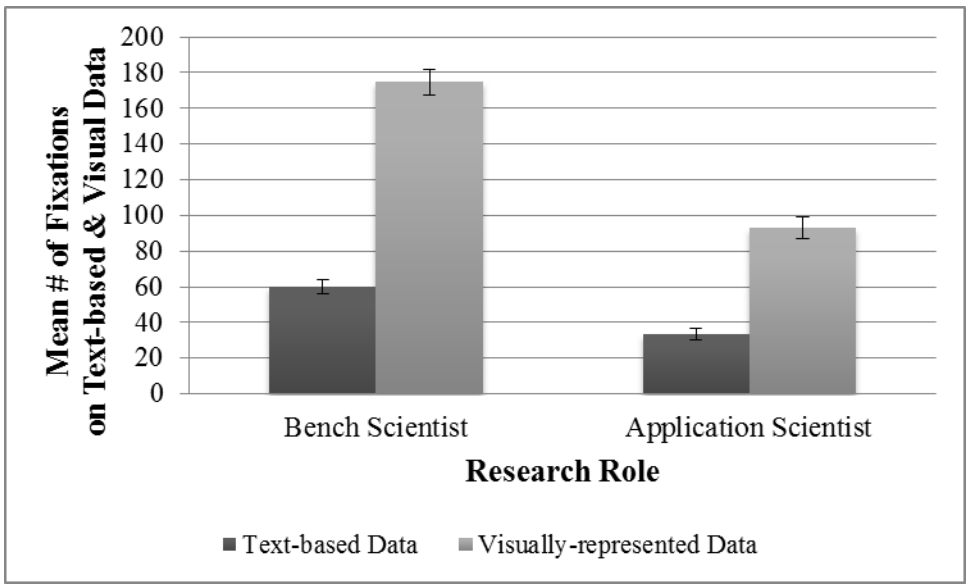


Figure 22. Mean number of fixation on insight drivers by research role.

For fixation durations, an independent samples t-test revealed that bench scientists had longer gaze duration on text-based data than application scientists (bench scientists, M= 17.62, SD= 2.81; application scientists, M= 11.38, SD= 3.56), $t(14) = 3.93$, $p < .00$). Also, a Mann-Whitney U-test indicated that bench scientists had longer gaze duration on visually-represented data (U= 3.00; p= 0.02) than application scientists (Figure 23).

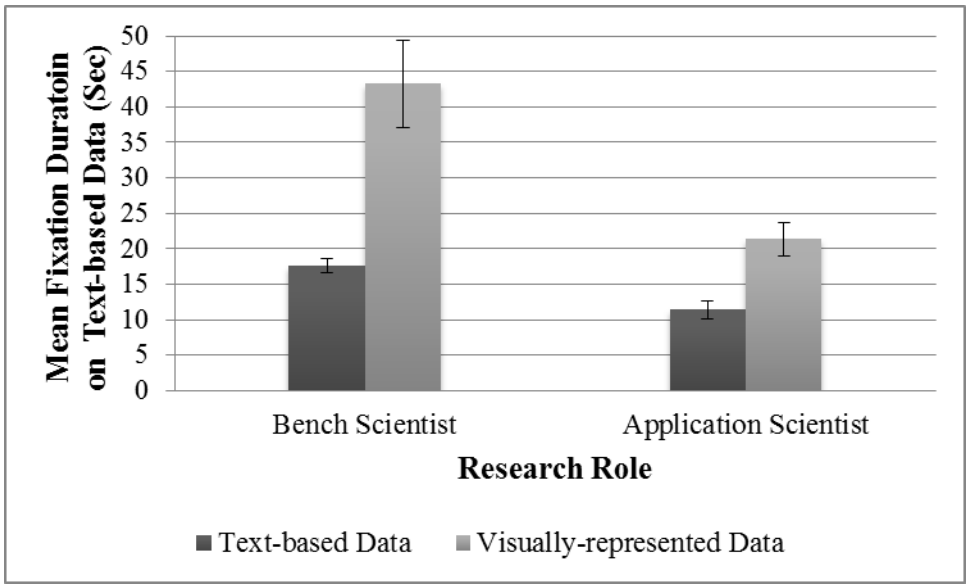


Figure 23. Mean fixation duration on insight drivers by research role.

Since eye-tracking data can't explain why participants attended specific data, we confirmed the reasons during the gaze-cued retrospective think aloud. For bench scientists, text-based data was considered as an important source of insights so they generated large number of fixations on text-based data (Marcel Adam Just & Carpenter, 1976). However, they experienced difficulties in interpreting information and extracting insights from the visually-represented data. Consequently, they made more fixations and had longer duration on visually-represented data (J. Goldberg & Kotval, 1998).

3.4.3.4 Human errors

We applied a Man-Whitney test in order to identify whether bench scientists and application scientists differ in terms of human errors. A Mann-Whitney test suggests that application scientists made more slips than bench scientists during insight generation (U= 9.00; p= 0.01). However, there was no statistically significant difference between bench scientists and application scientists on lapses (Mann-Whitney U-test: U= 30.00; p= 0.84) and mistakes (Mann-Whitney U-test: U= 23.00; p= 0.27) (Figure 24).

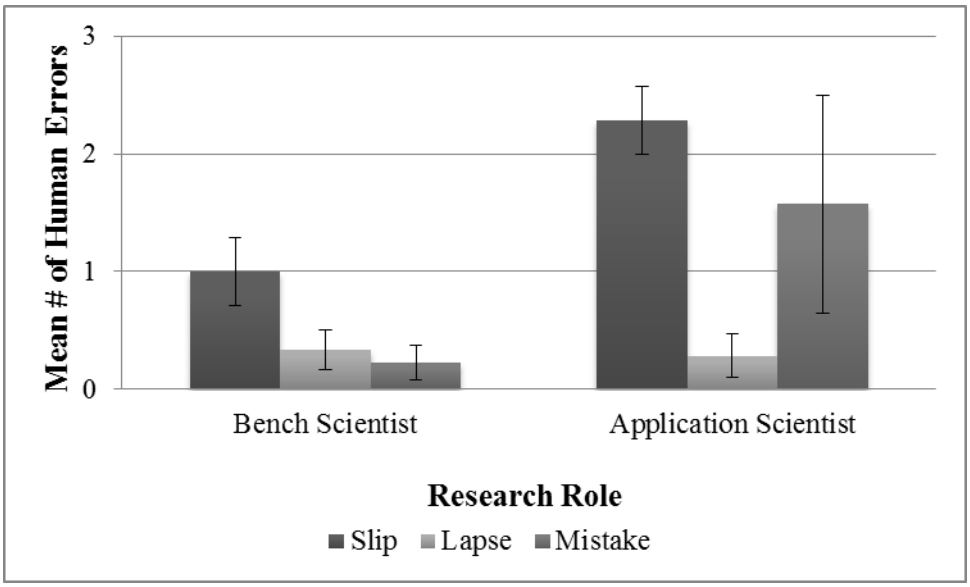


Figure 24. Mean number of human errors by research role.

3.4.4 The impact of cognitive styles on insight generation

To test the effect of cognitive style on 1) behavioral characteristics, 2) insight characteristics, 30 gaze-related behavioral measures, and 4) types of human error, we performed a Spearman's Rank-Order correlation.

Field dependency had a positive correlation with number of fixation on text-based data related to the target gene, which suggests that field independent scientists are more likely to fixate on text-based data (r_s = 0.707, p = .00). Also, number of fixation on visually-represented data was marginally correlated with field dependency, which suggests that field independent individuals generate higher fixation counts on visually-represented data to generate insights as compared to field dependent individuals (r_s = 0.488, p = .06). However, field dependency showed a negative correlation with slips, which suggests that field dependent individuals made more slips as compared to field independent individuals (r_s = -0.533, p = .03). However, no correlations were found between the field dependencies and other measures.

3.5 Discussion

Even though all participants were well-trained in their domain, their behaviors have been represented in different ways in order to cope with complex and unfamiliar situations. . Our results (1) confirmed significant differences with respect to insight generation behavior and human performance depending on research roles, and (2) identified some relationships between scientists' cognitive styles and human performance.

3.5.1 The impact of research roles

The impact of research roles on insight generation behavior and human performance has significant implications. In terms of behavioral characteristics, bench scientists' insight generation process was primarily driven by text-based data (e.g., sequence data, published references) and took less time to find the first insight as compared to application scientists. Unlike bench scientists, application scientists took longer time to find the first insight, made more use of visualization information, and reported more insights from the available resources. We speculate that the differences of behavior are the nature of the tasks they commonly perform as well as the level of training about bioinformatics systems. For instance, biologists typically

have been exposed to text-based data representations such as a text, cvs, or excel (Peakall & Smouse, 2012) and extensive literature reviews are frequent used to investigate defined hypotheses (Tran et al., 2004) .Whereas, application scientists have been well-trained in the use of more interactive and advanced computational and visualization tools so that they tend to report possible insights more quickly and easily from any types of resources than bench scientists.

Respecting insight characteristics, application scientists generate somewhat more correct insights than bench scientists in the limited time available. However, there were no significant differences of research roles on both the domain value and the degree of depth. In general, broad and trivial insights were frequently found by most participants. A possible explanation for this might be that since the participants conducted the main task within a limited time period, increasing levels of mental demands may lead to no significantly valuable insights by both bench scientists and application scientists.

Eye movement data provided useful information about how scientists generate insights while using different kinds of insight drivers. To be specific, bench scientists produce more fixations on both text-based data and visually-represented data than application scientists but found fewer insights than application scientists in general. It can therefore be assumed that text-based data is considered an important source of knowledge for bench scientists (Fitts, Jones, & Milton, 2005) and similarly bench scientists might experience difficulty in extracting or interpreting information from the visually-represented data (Fitts et al., 2005; Marcel Adam Just & Carpenter, 1976). These findings imply that bioinformatics resources should clearly support successful interactions with various information representations, especially visually-represented data for bench scientists (J. H. Goldberg & Kotval, 1999).

With regard to human performance, one striking difference was the number of slips between bench and application scientists. Application scientists made more slips than bench scientists while seeking insights across available knowledge resources. For example, an application scientist originally intended to browse a pathways analysis tool to identify regulation of gene expression in Escherichia, and once the analysis tool was presented s/he couldn't find the target gene. Another example was omitting a step to sort gene names without much conscious attention. Previous studies pointed out that errors can be affected by skill, experience and familiarity with

the situation encountered (Stanton & Salmon, 2009), especially slips and lapses occur at the skill-based level (Reason, 1990). As much in biological research, scientists require the three levels of behavior (skill-, rule-, and knowledge-based behavior). However, our result implies that application scientists can be more frequently influenced by *'skill-based behavior'* (Rasmussen, 1983) during insight generation as compared to bench scientists. This finding stresses the need for designs that reduce slips (e.g., failure to observe, inattention) resulting from skill-based behavior in biomedical and life sciences.

3.5.2 The impact of cognitive styles

Our research also indicated that cognitive styles have an impact on insight generation behavior and human performance. Specifically, more field independent individuals attended to both text-based data and visually-represented data than field dependent individuals. The number of fixations on specific components indicate the importance of a particular information elements (Fitts et al., 2005). Thus, the visual attention strategies they exhibit suggest that field independent people are likely to attend to the stimulus field in a more active manner, which is consistent with results of Herman A Witkin and Goodenough (1981). However we observed no significant correlation between fixation durations on types of insight drivers and cognitive styles. Although, previous studies revealed different tendencies to attend to a stimulus field (for example, field dependent individuals prefer to have lots of options available and tend to be greatly influenced by the dominant visual field (Herman A Witkin & Goodenough, 1981; Herman A Witkin et al., 1977)), we have found no evidence to this end. This (non) result might be attributable to the fact that interpreting visually-represented data under time pressure may negatively affect field dependent people during cognitively-demanding insight generation processes.

With regard to three human errors (i.e., slips, lapses, and mistakes), only slip correlated with cognitive style. More specifically, field dependent scientists generated more slips when they interacted with multiple information representations. This finding is consistent with previous studies that field independent individuals tend to solve problems analytically while dealing with ambiguous and demanding problems in comparison to field dependent individuals (Antonietti & Gioletta, 1995; H. A. Witkin, 1981; Herman A Witkin, Oltman, Raskin, & Karp, 1971). However, field dependency was not a significant predictor of lapses and mistakes. In other words, field dependent participants performed similarly to those who were field independent.

Furthermore, our findings highlight some interesting aspects for the use of knowledge resources. Scientists often employed generic search engines, especially Google (Pavelin et al., 2012) and generally clicked on links ranked highly by search engines regardless of individual differences. This is an obvious example of a "trust bias" where people tend to trust links higher in position and skip over others (Craswell, Zoeter, Taylor, & Ramsey, 2008; Joachims et al., 2007). However, no evidence of the trust bias was found when scientists use domain-specific search engines or databases of references on biomedical and life sciences topics such as PubMed, EcoCyc, and PATRIC. From a methodological perspective, the gaze-cued retrospective think aloud was useful in identifying human errors by inspecting participants' cognitive processes (van Gog et al., 2005). Participants were more motivated to explain their thoughts and reasons for looking at various parts of the screen as compared to during the concurrent think aloud. Sometimes they mentioned their experiences about online resources that they used during the main task. Previous studies suggest the gaze-augmented playback is especially useful in evaluating more complex environment (Eger, Ball, Stevens, & Dodd, 2007). Hyrskykari et al. (2008) found that participants seeing their eye movements in the RTA replay of the recoding may have helped in avoiding omissions (for instance, participants did not report all the thoughts, actions, reactions or feelings they experienced, Barkaoui (2011)). This finding is also consist with Guan et al. (2006) who indicated that when the participant has worked on difficult tasks such omission occur more often than with simple task. However, most participants in this study referred to the gaze path when they could not recall the steps they had just performed or the reasons why they had selected specific data in order to conduct the task. The participants might be influenced by the nature of the task and time pressures (Taylor & Dionne, 2000). Thus, further study such as the impact of task difficulty and time pressure on the gaze-cued retrospective think aloud would be interesting.

3.6 Conclusions

Bioinformatics practitioners have often by-passed the process of understanding users' tasks, contexts, and characteristics. A major challenge for the design of online bioinformatics resources is to support diverse users under conditions of data-intensive and interdisciplinary collaboration.

Maxion and Reeder (2005) insisted that "one reason why some user interfaces fail to meet their speed and accuracy specifications is human error". Research into impacts of individual

differences on insight generation behavior and human performance help us gain insight to, and possibly predict human performances (Cegarra & Hoc, 2006; Dillon & Watson, 1996). Such insights could lead to design guidelines to accommodate different user groups and reduce disparities in performance among interdisciplinary research communities.

The methodological approach we adopted for this study was meaningful as a new insight-based method, extended not only to understand how scientists generate insights with different types of knowledge resources but also to examine human performance associated with individual differences. The results produces by this method can be used to improve online bioinformatics resources user interface design to support higher cognitive activities.

3.7 Future directions

Our study raises several issues that can be addressed by future research.

First, even though this study demonstrated that individual differences play a substantial role on the interaction between scientists and various resources, this study did not address specific information to explain why errors occurs as they do and how to cope with errors. Such future research would contribute to a more in-depth understanding of the 1) impact of individual differences on human errors and 2) error handling process which is defined as "the process from error detection to recovery" (Zapf & Reason, 1994) occurring in the context of bioinformatics experiments.

Furthermore, this study considered recognizable human errors from user behaviors and verbalization while employing concurrent think aloud and retrospective think aloud protocols. Future research could examine the impact of individual differences on heuristics and biases in order to explain a wider range of human performance under scientific uncertainty from cognitive-oriented perspectives.

Lastly, it should be noted that this study asked participants to perform the main task (i.e., eye-tracking with concurrent think aloud protocol) for about 30 minutes; much shorter in time as compared to actual work practices. Future research should examine the effects of interrelationships among different knowledge resources over longer periods of time. To address

this issue, we propose a longitudinal study (see Study 3) that is more representative of scientists' real world knowledge resource usage from a distributed cognition perspective.

3.8 References

Anderson, Nicholas R., Ash, Joan S., & Tarczy-Hornoch, Peter. (2007). A qualitative study of the implementation of a bioinformatics tool in a biological research laboratory. *International Journal of Medical Informatics, 76*(11–12), 821-828. doi: 10.1016/j.ijmedinf.2006.09.022

Aniba, Mohamed Radhouene, & Thompson, Julie D. (2010). *Knowledge Based Expert Systems in Bioinformatics*.

Antonietti, Alessandro, & Gioletta, Maria Alfonsa. (1995). Individual differences in analogical problem solving. *Personality and Individual Differences, 18*(5), 611-619.

Barkaoui, Khaled. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75.

Bartlett, Joan C., & Neugebauer, Tomasz. (2008). *A task-based information retrieval interface to support bioinformatics analysis*. Paper presented at the Proceedings of the second international symposium on Information interaction in context, London, United Kingdom.

Bartlett, Joan C., & Toms, Elaine G. (2005). Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach. *Journal of the American Society for Information Science and Technology, 56*(5), 469-482. doi: 10.1002/asi.20136

Bolchini, D. (2009). Better bioinformatics through usability analysis. *Bioinformatics (Oxford, England), 25*(3), 406-412. doi: 10.1093/bioinformatics/btn633

Boyandin, Ilya, Bertini, Enrico, & Lalanne, Denis. (2012). A Qualitative Study on the Exploration of Temporal Changes in Flow Maps with Animation and Small-Multiples. *Computer Graphics Forum, 31*(3pt2), 1005-1014. doi: 10.1111/j.1467-8659.2012.03093.x

Camps, Joaquim. (2003). Concurrent and retrospective verbal reports as tools to better understand the role of attention in second language tasks. *International Journal of Applied Linguistics, 13*(2), 201-221.

Cegarra, Julien, & Hoc, Jean-Michel. (2006). Cognitive styles as an explanation of experts' individual differences: A case study in computer-assisted troubleshooting diagnosis.

International Journal of Human-Computer Studies, 64(2), 123-136. doi: http://dx.doi.org/10.1016/j.ijhcs.2005.06.003

Cellier, J. M. (1997). Expertise in dynamic environments. *Ergonomics, 40*(1), 28-50. doi: 10.1080/001401397188350

Chase, William G, & Simon, Herbert A. (1973). The mind's eye in chess.

Cohen, Louis, Manion, Lawrence, & Morrison, Keith. (2000). *Research methods in education*: Routledge.

Craswell, Nick, Zoeter, Onno, Taylor, Michael, & Ramsey, Bill. (2008). *An experimental comparison of click position-bias models.* Paper presented at the Proceedings of the 2008 International Conference on Web Search and Data Mining.

Cutrell, Edward, & Guan, Zhiwei. (2007). *What are you looking for?: an eye-tracking study of information usage in web search.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Davies, Sue, & Crookes, Patrick. (1998). *Research into practice: essential skills for reading and applying research in nursing and health care*: Bailliere Tindall, published.

de Matos, Paula, Cham, Jennifer A, Cao, Hong, Alcántara, Rafael, Rowland, Francis, Lopez, Rodrigo, & Steinbeck, Christoph. (2013). The Enzyme Portal: a case study in applying user-centred design methods in bioinformatics. *BMC bioinformatics, 14*(1), 103.

Dillon, Andrew, & Watson, Charles. (1996). User analysis in HCI—the historical lessons from individual differences research. *International journal of human-computer studies, 45*(6), 619-637.

Duncan, K. D. (1985). Representation of Fault-Finding Problems and Development of Fault-Finding Strategies. *Innovations in Education & Training International, 22*(2), 125-131. doi: 10.1080/1355800850220204

Eger, Nicola, Ball, Linden J, Stevens, Robert, & Dodd, Jon. (2007). *Cueing retrospective verbal reports in usability testing through eye-movement replay.* Paper presented at the Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1.

Elling, Sanne, Lentz, Leo, & De Jong, Menno. (2012). Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *Professional Communication, IEEE Transactions on, 55*(3), 206-220.

Ericsson, KA, & Simon, HA. (1993). Protocol analysis: Verbal reports as data (rev. ed.) MIT Press. *Cambridge, MA*.

Fitts, Paul M, Jones, Richard E, & Milton, John L. (2005). Eye movements of aircraft pilots during instrument-landing approaches. *Ergonomics: Psychological mechanisms and models in ergonomics, 3*, 56.

Fleiss, L, Levin, Bruce, & Paik, Myunghee Cho. (1981). *The measurement of interrater agreement.* Paper presented at the In Statistical methods for rates and proportions (2nd ed.

Galperin, Michael Y, Rigden, Daniel J, & Fernández-Suárez, Xosé M. (2015). The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. *Nucleic acids research, 43*(D1), D1-D5.

Goldberg, JH, & Kotval, XP. (1998). Eye movement-based evaluation of the computer interface. *Advances in occupational ergonomics and safety*, 529-532.

Goldberg, Joseph H, & Kotval, Xerxes P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics, 24*(6), 631-645.

Guan, Zhiwei, Lee, Shirley, Cuddihy, Elisabeth, & Ramey, Judith. (2006). *The validity of the stimulated retrospective think-aloud method as measured by eye tracking.* Paper presented at the Proceedings of the SIGCHI conference on Human Factors in computing systems.

Hansen, John Paulin. (1991). The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica, 76*(1), 31-49.

Huang, Weidong. (2013). Handbook of Human Centric Visualization: Springer.

Hunter, Sarah, Apweiler, Rolf, & Martin, Maria Jesus. (2010). Design, Implementation and Updating of Knowledge Bases *Knowledge-Based Bioinformatics* (pp. 87-105): John Wiley & Sons, Ltd.

Hyrskykari, Aulikki, Ovaska, Saila, Majaranta, Päivi, Räihä, Kari-Jouko, & Lehtinen, Merja. (2008). Gaze path stimulation in retrospective think-aloud. *Journal of Eye Movement Research, 2*(4), 1-18.

Javahery, Homa. (2004). Beyond power making bioinformatics tools user-centered. *Communications of the ACM, 47*(11), 58.

Joachims, Thorsten, Granka, Laura, Pan, Bing, Hembrooke, Helene, Radlinski, Filip, & Gay, Geri. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS), 25*(2), 7.

Jonassen, David H, & Grabowski, Barbara L. (2012). *Handbook of individual differences learning and instruction*: Routledge.

Just, Marcel A, & Carpenter, Patricia A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review, 87*(4), 329.

Just, Marcel Adam, & Carpenter, Patricia A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8*(4), 441-480.

Karasavvas, KA, Baldock, R, & Burger, A. (2004). Bioinformatics integration and agent technology. *Journal of biomedical informatics, 37*, 205-219.

Keen, Peter GW, & Morton, Michael S Scott. (1978). *Decision support systems: an organizational perspective* (Vol. 35): Addison-Wesley Reading, MA.

Kuusela, Hannu, & Paul, Pallab. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American journal of psychology*.

Letondal, C., & Mackay, W.E. (2004). *Participatory programming and the scope of mutual responsibility: balancing scientific, design and software commitment.* Paper presented at the Proceedings of the eighth conference on Participatory design: Artful integration: interweaving media, materials and practices-Volume 1.

Marcus, Frederick. (2008). *Bioinformatics and systems biology: collaborative research and resources*: Springer Science & Business Media.

Maxion, Roy A, & Reeder, Robert W. (2005). Improving user-interface dependability through mitigation of human error. *International Journal of Human-Computer Studies, 63*(1), 25-50.

Mirel, B. (2007, 1-3 Oct. 2007). *Usability and Usefulness in Bioinformatics: Evaluating a Tool for Querying and Analyzing Protein Interactions Based on Scientists' Actual Research Questions.* Paper presented at the Professional Communication Conference, 2007. IPCC 2007. IEEE International.

Mirel, B. (2009). Supporting cognition in systems biology analysis: findings on users' processes and design implications. *J Biomed Discov Collab, 4*, 2. doi: 10.1186/1747-5333-4-2

Moran, A. P. (1986). Field Independence and Proficiency in Electrical Fault Diagnosis. *Systems, Man and Cybernetics, IEEE Transactions on, 16*(1), 162-165. doi: 10.1109/TSMC.1986.289294

Norman, Donald A. (1981). Categorization of action slips. *Psychological review, 88*(1), 1.

Patton, Michael Quinn. (2005). Qualitative Research *Encyclopedia of Statistics in Behavioral Science*: John Wiley & Sons, Ltd.

Pavelin, Katrina, Cham, Jennifer A, de Matos, Paula, Brooksbank, Cath, Cameron, Graham, & Steinbeck, Christoph. (2012). Bioinformatics meets user-centred design: a perspective. *PLoS computational biology, 8*(7), e1002554.

Peakall, Rod, & Smouse, Peter E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics, 28*(19), 2537-2539.

Plaisant, C., Fekete, J. D., & Grinstein, G. (2008). Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository. *Visualization and Computer Graphics, IEEE Transactions on, 14*(1), 120-134. doi: 10.1109/TVCG.2007.70412

Prabhat, Forsberg, A., Katzourin, M., Wharton, K., & Slater, M. (2008). A Comparative Study of Desktop, Fishtank, and Cave Systems for the Exploration of Volume Rendered Confocal Data Sets. *Visualization and Computer Graphics, IEEE Transactions on, 14*(3), 551-563. doi: 10.1109/TVCG.2007.70433

Rasmussen, Jens. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *Systems, Man and Cybernetics, IEEE Transactions on*(3), 257-266.

Rayner, Keith. (1978). Eye movements in reading and information processing. *Psychological bulletin, 85*(3), 618.

Rayner, Keith. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124*(3), 372.

Reason, James. (1990). *Human error*: Cambridge university press.

Reason, James. (1995). Understanding adverse events: human factors. *Quality in health care, 4*(2), 80-89.

Romano, Paolo, Giugno, Rosalba, & Pulvirenti, Alfredo. (2011). Tools and collaborative environments for bioinformatics research. *Briefings in Bioinformatics*. doi: 10.1093/bib/bbr055

Runco, Mark A, & Pritzker, Steven R. (1999). *Encyclopedia of creativity. 2. I-Z; Indexes* (Vol. 2): Access Online via Elsevier.

Saraiya, P. (2005). An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. *IEEE transactions on visualization and computer graphics, 11*(4), 443-456. doi: 10.1109/tvcg.2005.53

Saraiya, Purvi, North, Chris, Lam, Vy, & Duca, Karen A. (2006). An insight-based longitudinal study of visual analytics. *Visualization and Computer Graphics, IEEE Transactions on, 12*(6), 1511-1522.

Silverman, David. (2009). *Doing qualitative research*: SAGE Publications Limited.

Smuc, Michael, Mayr, Eva, Lammarsch, Tim, Aigner, Wolfgang, Miksch, Silvia, & Gartner, Johannes. (2009). To score or not to score? Tripling insights for participatory design. *Computer Graphics and Applications, IEEE, 29*(3), 29-38.

Stanton, Neville A, & Salmon, Paul M. (2009). Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science, 47*(2), 227-237.

Tadmor, Brigitta, & Tidor, Bruce. (2005). Interdisciplinary research and education at the biology–engineering–computer science interface: a perspective. *Drug Discovery Today, 10*(17), 1183-1189. doi: http://dx.doi.org/10.1016/S1359-6446(05)03540-3

Tarczy-Hornoch, Peter, & Minie, Mark. (2005). Bioinformatics Challenges and Opportunities

Medical Informatics. In H. Chen, S. S. Fuller, C. Friedman & W. Hersh (Eds.), (Vol. 8, pp. 63-94): Springer US.

Taylor, K Lynn, & Dionne, Jean-Paul. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*(3), 413.

Tran, D., Dubay, C., Gorman, P., & Hersh, W. (2004). Applying task analysis to describe and facilitate bioinformatics tasks. *Stud Health Technol Inform, 107*(Pt 2), 818-822.

van Gog, Tamara, Paas, Fred, van Merriënboer, Jeroen J. G., & Witte, Puk. (2005). Uncovering the Problem-Solving Process: Cued Retrospective Reporting Versus Concurrent and

Retrospective Reporting. *Journal of Experimental Psychology: Applied, 11*(4), 237-244. doi: 10.1037/1076-898x.11.4.237

Visserl, Willemien, Hocz, Jean-Michel, & Chesnay, France. (1990). Expert software design strategies.

Wickens, Christopher D, & Liu, Yili. (1988). Codes and modalities in multiple resources: A success and a qualification. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 30*(5), 599-616.

Witkin, H. A. (1981). Cognitive styles: essence and origins. Field dependence and field independence. *Psychological issues*(51), 1-141.

Witkin, Herman A. (1971). *A manual for the embedded figures tests*: Consulting Psychologists Press.

Witkin, Herman A. (1973). THE ROLE OF COGNITIVE STYLE IN ACADEMIC PERFORMANCE AND IN TEACHER-STUDENT RELATIONS12. *ETS Research Bulletin Series, 1973*(1), i-58.

Witkin, Herman A, & Goodenough, Donald R. (1981). *Cognitive styles, essence and origins: Field dependence and field independence*: Intl Universities Pr Inc.

Witkin, Herman A, Moore, Carol Ann, Goodenough, Donald R, & Cox, Patricia W. (1977). Field-dependent and field-independent cognitive styles and their educational implications. *Review of educational research, 47*(1), 1-64.

Witkin, Herman A, Oltman, PK, Raskin, Evelyn, & Karp, Stephen A. (1971). A manual for the group embedded figures test. *Palo Alto, California*.

Witkin, Herman A., & Goodenough, Donald R. (1977). Field dependence and interpersonal behavior. *Psychological Bulletin, 84*(4), 661-689. doi: 10.1037/0033-2909.84.4.661

Zapf, Dieter, & Reason, James T. (1994). Introduction: Human errors and error handling. *Applied Psychology, 43*(4), 427-432.

4 Chapter 4: Empirically-driven design considerations for supporting distributed cognitive activities in bioinformatics

4.1 Introduction

The huge demand for analysis and interpretation of biological data is accelerating both the need for, and growth of, data-rich computational and informatics methods (Bayat, 2002). By accessing disparate biological databases and analysis tools, technically-savvy life scientists are able to utilize a wide range of resources with the minimum of effort (Goble et al., 2001). With an increased awareness of user-centered system design, a growing body of research ushered in a paradigm shift to elicit user needs from both "wet" bench scientists and computer-based "dry" bench scientists through unstructured interviews and observations (B. Mirel, 2009; Tran, Dubay, Gorman, & Hersh, 2004), task analysis (J. Bartlett & Neugebauer, 2005), and survey approaches (J. C. Bartlett, Ishimura, & Kloda, 2012). Other UCD works in this domain include approaches where participants are given a range of predefined tasks to conduct in short-term controlled studies, such as typical formative usability studies (Bolchini, 2009; Javahery, 2004). A common limitation of formative evaluations are that they are conducted over the course of few hours; obviously a much shorter timeframe as compared to actual scientists'' work practices. Moreover, these approaches shed light onto a few usability issues (the most critical incidents) that are encountered during a brief moment of the user experience.

However, while good usability is necessary, but not sufficient to ensure a resources' ability to enrich user experiences of data-intensive research communities. Moreover, a variety of data-intensive research activities cannot be explored in a one or two-hour session nor by investigation of the partial set of possible user tasks and experiences during short-term observation.

To date, scientists are governed by various means to deliver enhanced performance and quality of research by leveraging different types of resources, technologies and other people (J. Bartlett & Neugebauer, 2005; B. Mirel, 2009; Tran et al., 2004). No study has examined how external artifacts and contexts combine to influence data-intensive research processes over long periods of time. Distributed cognition (See Section 1.1.5 for the definition of this term) offers a comprehensive approach to understanding user experiences emerging from interactions where

"technology acts as a mediator between the user and the activity" (Lallemand, Gronier, & Koenig, 2015). We believe that there are opportunities here that are currently underexplored. To be specific, having temporal and longitudinal information from distributed cognitive activities in a natural working context of human-computer interaction might allow us to 1) capture cumulative user experiences, shedding light on how user experiences are formed and 2) lead to more persuasive information to inform design to better support data-intensive research communities.

We performed a longitudinal diary study over the course of two weeks followed by a focus group session. A goal of the diary study was to gain basic understanding into distributed cognitive activities of life scientists. Specifically, the research questions of the diary study were:

- What kind of distributed cognitive activities do scientists perform in real-world contexts?
- Are there limitations of current online resources that hinder user experiences and what are alternative design strategies to overcome these limitations?

We then carried out a focus group session to identify and determine UX/UI design considerations and priorities. During the subsequent focus group sessions we used diary entries as memory cues to drive discussion. From the focus group session, we were interested in identifying:

- How can online bioinformatics resources be improved to enrich user experiences in distributed cognition environments?

Our findings revealed six broad categories of distributed cognitive activities, current user experience levels (e.g., perceived helpfulness, challenges, and alternative actions), and important criteria associate with each category. Based on these findings, we suggest design considerations for building online bioinformatics resources to mitigate the current barriers of use and to enrich user experiences. Distinct from prior research mainly aiming at addressing user workflows or usability issues, our research outcomes are more comprehensive to embrace, leverage, and support distributed cognitive activities that are based on user experiences in real contexts of use.

4.2 Research background

There are several ways for capturing *in situ* data from users. We employed a diary study which was less intrusive than experiments or protocol studies (Wild, McMahon, Darlington, Liu, & Culley, 2010). Diary techniques are a form of self-logging and -reporting of activities (ZIMMERMAN & Wieder, 1976), in which participants record their activities and events into a diary about how they are interact with online knowledge resources as well as to processes by which they conduct research. Diary techniques provide high ecological value (Czerwinski, Horvitz, & Wilhite, 2004) (Brunswik, 1941), and capture particular events and experiences in natural and spontaneous contexts (Reis, 1994). Thus, diary studies are a useful technique for data collection in scenarios when no automated logging can be employed (Möller, Kranz, Schmid, Roalter, & Diewald, 2013).

What we know about distributed cognition is largely based upon ethnographic approaches (e.g., naturalistic observation with interview sessions). However, researchers often face difficulties in getting physical access to real experts in real-world settings (Chilana, Wobbrock, & Ko, 2010; Gabbard et al., 2003; Redish, 2007). In life sciences, for instance, it is a common that access to laboratories is commonly restricted when research is in-progress and data is being collected (B. Mirel, 2009). Another limitation is that the presence of the researcher during an observation or an interview may bias the results (e.g., Heisenberg's Uncertainty Principle), leading to responses in a positively biased manner (Andrews, Robinson, & Wrightsman, 1991; Dell, Vaidyanathan, Medhi, Cutrell, & Thies, 2012).

From a methodological perspective, ethnographic approaches have several other challenges concerning data collection such as intrusiveness, observer bias, and observer drift (Bogdan & Biklen, 1982). Furthermore, the ethnographic data tend to highly rely on researchers' retrospective interpretation (LeCompte & Goetz, 1982). A common alternative is to employ retrospective interviews, however these are known to be prone to bias by heuristic and participants' current affective state (Jones & Johnston, 2011). Given that, diary studies may be helpful to minimize retrospective bias (Bolger, Davis, & Rafaeli, 2003). Therefore, we undertook a diary method as a substitutive measure to identify the implicit needs and desires of users by capturing real-world online knowledge resource usage (Wild et al., 2010).

To identify specific classes of distributed cognitive activities (i.e., interactions between people, artifacts and technological systems), we took an event-contingent approach that requires participants to report each time a specific event occurs (Bolger et al., 2003). In general, the event-contingent approach is desirable when researchers are interested in capturing pre-defined events (Wheeler & Reis, 1991). This approach requires data recordings that are very close in time to the actual event to help reduce the subjects' likelihood of forgetting or reappraising (Wheeler & Reis, 1991).

To obtain reliable and valid data, a researcher or moderator provides participants a clear definition of triggering events and simple data entry method (Bolger et al., 2003). Accordingly, we define "*distributed cognitive activities*" as a range of interactions with any online knowledge resources (e.g., visual analytics tools, data and knowledge repositories, and online communities) that extend individuals' cognitive capacities. The combined effects of interactions with all possible physical artifacts and/or people are too complicated to be examined in detail here. As such, we limited the scope of this study to online computer-mediated distributed cognitive activities at the individual level.

As a research technique, the focus group session employs interaction-based discussion as a means of generating " rich details of complex experiences and the reasoning behind [an individual's] actions, beliefs, perceptions and attitudes" (Powell & Single, 1996). Further, researchers can apply a focus group interview "prior to, concurrently with, or after a quantitative study, or separately" (Powell & Single, 1996). Focus groups explicitly adopt group interactions, where people are encouraged to exchange experiences, points of view and follow-up comments regarding others' opinions (Kitzinger, 1995). Such interaction in focus groups offers valuable data on the scope of consensus and/or diversity among the participants (Morgan, 1996). A focus group approach can be used with other qualitative approaches in multi-method designs to specify both breadth and depth of data (Hesse-Biber & Leavy, 2010). In this study, we employed a focus group session as a complementary means to 1) investigate unexplored areas of user experiences via the self-reporting diary approach and 2) identify, organize and prioritize UX/UI design values to support distributed cognitive activities in data-intensive interdisciplinary research communities.

Previous studies recommend that a focus groups recruit four and six participants per group (Greenbaum, 1997; Kitzinger, 1995; Powell & Single, 1996). Morgan (1992) suggests that smaller groups result in a higher level of involvement (as compared to larger groups) because each participant has more opportunity to discuss their views and experiences on a topic. In addition, smaller groups can lead to active discussions as guided by moderators as compared to large groups (Morgan, 1996). The number of group sessions depends on the nature and complexity of research objectives. Powell and Single (1996) recommend that anywhere between one to ten sessions are generally sufficient for most focus group studies because group discussions will eventually converge and essentially repeat existing data. Given the recruitment, scheduling and budget issues, we accordingly conducted one focus group session with six biological scientists.

4.3 Methods

4.3.1 Participants

We recruited a convenience sample of 7 participants (3 males and 4 females), to include three bench scientists and four application scientists (See Section 1.1.4 for definitions of these terms). We used the following sampling criteria: 1) participated in Study 2 and were interested in the subsequent follow on study, and 2) were representative of each group (bench scientists and application scientists). All participants had at least a post-graduate degree in biomedical and life sciences. Due to difficulties of running the diary study, we lost the data of one of the application scientists. Therefore, we report results from the diary study of three bench scientists and three application scientists.

For the focus group session, we selected four scientists from the diary study and recruited two domain experts with over 10 years research experiences in biological sciences.

4.3.2 Procedures

The experiment had four types of activities: a training session, self-reporting, weekly individual interviews, and a focus group interview. The self-reporting and weekly interviews were recurring activities.

Once we recruited participants, we scheduled training sessions to ensure participants fully understood the protocol as well as their responsibilities as diary study participants (Bolger et al., 2003). We first conducted the informed consent process approved by the Virginia Tech Institutional review Board (IRB) prior to any data collection (See Appendix G and Appendix H). We then asked participants about their main research/job responsibilities, online bioinformatics resources commonly used, and general experiences to date with online knowledge-sharing platforms (See Section 1.1.6 for the definition of this term). Next, we provided instructions on how to use the electronic diary application to record their tasks and experiences while using external resources (Appendix I).

After administering the training session, we asked participants to keep a detailed diary for two weeks as they conducted their research. As mentioned, one of objectives in this study was to propose design considerations for bioinformatics resources to satisfactorily support user experiences in biomedical and life sciences. Thus we asked participants to include distributed cognitive activities related to conducting their research, that occurred when they were at work in computer-mediated environments.

We sent participants one email per day to remind them to diligently fill in their diary entries. Detailed questions for the diary are presented in Appendix J. We provided compensation for the portion of day for which they participated, up to $80 at the end of whole process of the diary study. To be specific, if a participant successfully completed 3 diary entries, we compensated the participant at a rate of $8 per day. If a participant successfully completed 2 diary entries, we compensated the participant at a rate of $5 per day. If a participant successfully completed 1 diary entry, we compensated the participant at a rate of $2 per day.

When needed, we interviewed participants using online communication tools (e.g., WebEx, Skype, etc) to clarify missing or ambiguous diary entries,

Following the two-week diary period, we scheduled a focus group session at time that was convenient for all participants. The focus group interview session took about 2 hours. We provided compensation for participants at a rate of $10 per hour for a maximum of 2 hours, or a maximum of $20 total. We employed the affinity diagramming process as a tool for elicitation, organization and prioritization of user requirements. At the beginning of the session, we

introduced the objectives and summaries of the research and then asked participants to share their experiences and challenges related to research activities. Next, we asked all participants to identify, how good user experiences provide value based on their experiences and expectations. Participants wrote these potential values on separate post-it™ notes for about 30 minutes. The participants placed these notes on a wall, physically sorting notes into logical groupings of value. They created and placed group headers at the top of related notes. Also they moved any values from their initial position as needed. This process was repeated until all possible values were identified, grouped and prioritized. After capturing all participants' values, participants prioritized them through a voting protocol and discuss details about specific needs and desires. The focus group session lasted approximately 90 minutes.

4.3.3 Measures

We collected diary entries from scientists who are active daily users of online bioinformatics resources and have a strong commitment to spontaneously self-reporting online activities. For each diary entry, we asked a number of different questions: 1) what was the action or activity; 2) what the purpose of action; 3) what external resources/tools have been accessed; 4) based on the resources accessed, which perceived/anticipated benefits of each resource influenced their decision to use the resources; 5) which part(s) of the resource was helpful; 6) which part(s) of the resource was challenging; 7) if so, what your alternative action was; and, 8) general recommendations to better support their research.

The focus group session of this study was to elicit, organize, and prioritize values associated with user experience design considerations so that future bioinformatics platforms might better address user needs in data-intensive research circumstances.

4.3.4 Analysis

Two independent coders reviewed and categorized diary entries iteratively based on pre-defined decision rules (Tran et al., 2004), compared the categories, and discussed any disagreements. Then, we calculated the Kappa values of inter-rater agreement to assess inter-rater reliability. The initial Kappa value was 0.73. After the reconciliation meeting, we achieved 92.0% agreement, which is above the accepted standard of 75% suggested by (Fleiss, Levin, & Paik, 1981).

4.4 Results

4.4.1 Descriptive statistics

The demographic information about participants (*n*=6) is shown in Table 12. A total of 66.6% (*n*=4) of respondents are female, and 33.3% (*n*=2) are male, the age of the participants ranged from 24 to 33 (M = 28.00, SD = 2.97). All participants use bioinformatics resources on a daily basis or several times a week. 83.3% of participants (*n*=5) mentioned that their organization culture tends to promote knowledge networking amongst and between each other.

Table 12. Participant descriptions

Subject	Cluster	Age	Gender	Research Experience	Online bioinformatics resource usage
1	Bench scientist	24	F	2 year 7 months	Almost every day
2	Application scientist	27	M	4 years	More than once a week
3	Bench scientist	29	F	4 years 3 months	More than once a week
4	Application scientist	27	F	3 years	More than once a week
5	Bench scientist	33	M	10 years	More than once a week
6	Application scientist	28	F	5 years	More than once a week

4.4.2 Classification of distributed cognitive activities

Our study generated 98 diary entries, with an average of 16.3 entries per person (min: 8 max: 28 SD: 7.17). We sorted the diary entries into six broad categories of distributed cognitive activities based on the content of participants' diary entries. In total, about half of the reported activities in participants' diaries were described as "information seeking" that includes "recognizing and interpreting the information problem, establishing a plan of search, conducting the search, evaluating the results" (Marchionini, 1989).

Since the Internet has become the main source of information access today, it is not a surprise that our scientists relied heavily on it. We consider "information search" and "literature search" as two separate categories because information search usually involves comprehensive

information seeking, grounded in a research questions such as finding images, searching tutorials, querying databases, and foraging for available tools. Whereas literature search is uniquely associated with seeking published literature in order to identify a breath of good quality references relevant to specific research topics.

Overwhelmingly, 'information search' (34.7%) was a main activity that participants performed in order to understand protocols, reaction mechanisms related their research, or analyze specific datasets regardless of their research roles. We found that information search could best be described as using Google, Wikipedia, various university and/or faculty-run websites, collaborative discussion forums such as Stack Overflow, Seqanswers. ResearchGate, and generic online social network services such as Facebook. The following responses suggest that public community spaces can enable scientists to connect with other scientists with similar methods and/or goals.

> *"They provide examples that are closely related to my questions in most cases. So it's easy to get the solution to my questions."*

Also, several different options of searching such as filtering, keyword match were noted as very helpful.

> *"Peer-reviewed answers are helpful and trust-worthy."*

> *"Ability to get quick answers from experts was useful."*

However, many participants had difficult user experiences due to overuse of jargon and acronyms.

> *"I don't understand some terminologies and acronyms used on the website."*

Also, inconsistent results (e.g., different versions of gene sequences) across different online resources led to confusing and loss of time, since participants needed to identify the "best" online resource among the many that purport similar capabilities, data, and performance.

> *"The way in which website leads to different versions of gene sequences is sometimes very time consuming and also leads to too many unwanted data pages."*

There appeared to be no well-documented, comprehensive comparative studies on different datasets. In addition, there are still some usability issues to be solved.

> *"The forums are hard to navigate/search to find the thread that can answer my question. There are no results displayed for the demo data, choices of certain values for some parameters not clearly explained."*

As an alternative method, a number of participants tended to use more specific keywords in order to leverage common resources such as Google. For example,

> *"The forums are hard to navigate/search to find the thread that can answer my question. There are no results displayed for the demo data, choices of certain values for some parameters not clearly explained."*

> *"Google search is the best options. If search important keywords that best describe your problem, you can get links to very good discussion forums."*

> *"I searched my question on Google to find the most suitable thread."*

Participants tended to use online communities when they couldn't find appropriate information. For example,

> *"I usually use Facebook, because other people with the same issues could answer the questions more realistically."*

> *"I posted my question on forums."*

Some participants found acquaintances and asked them. For example, one participant contacted lab members, or sends emails to colleagues. For example, participants reported the following:

> *"If I have a question I always use an expert email address because I worried about getting scooped."*

> *"Using Facebook in this case was good because I know the person who gave me the answer and this seems more reliable to me that getting info from random websites."*

Interestingly, one participant remarked:

> *"Forget about this tool. I have to find some other option. There are many other primer design tools."*

Better ways of tagging and indexing question threads are often cited as highly desired functions. Also, the need for "detailed comparison and explanations of different datasets" occurred with high frequency.

The second highest category among both bench scientists and application scientists was 'literature search' (21.4%) via Google search engine, NCBI Pubmed, Google scholar, Wikipedia, and SciFinder Scholar. Common objectives involved finding and evaluating papers related to research interests, learning how to explain unexpected results from other people's perspectives, and determining and synthesizing information in order to achieve their research goals. Most participants were satisfied with advanced search options such as by author, publisher, published year, area of research. For example, Pubmed allowed participants to click on an author's name and bring up all the publications associated with that same name.

> *"They provide advanced search options such as by author, publisher, published year, area of research and so on."*

> *"I was quite impressed that they had a 'search by structure' option that allowed you to try the chemical you were looking for."*

> *"It's easy to get useful information I need by inputting any related key words. Especially when I am uncertain about the paper I really want, by inputting a few key words it would automatically return several highly related papers."*

Participants often had difficulties finding papers without knowledge the titles.

> *"Sometimes it's hard to get the specific one or two desired papers without knowing the titles. It really needs the specific title."*

Another difficulty that participants addressed was:

> *"[it] links publications about the chemical to the product, but what I really need are*patents* about the chemicals"*

When participants were faced with a difficult problem, most kept looking across different online resources or accessed social networking resources in order to look up the author's information.

> *"Use public social network resources, such as LinkedIn, to search the authors' information. Based on the authors, I can search the specific paper he published in certain year."*

Most participants expect that provenance of the references will be complemented by authors' profile, citation rates, and cross references.

The third category of the distributed cognitive activity was 'protein analysis' (10.2%) and was mainly associated with bench scientists' research. Some purposes included 1) collecting information about protein-protein interactions for specific protein samples, 2) visualizing the protein interaction network based on expression level data to identify trends in protein expression profiles.

Participants stated that *data integration from multiple sources* and *data visualization* was important to thorough analyses of protein-protein interactions. Also, bench scientists remarked that detailed information of protein localization generated by reference search is useful and trustworthy**:**

> *"...if I click any part of the PTM location diagram, it will lead me to the related publication. Therefore, [the online resource] provides curated, and highly reliable level of PTM information. And it covers various species."*

As the amount of tools available, it was often hard for participants to find the "right one" for their task. In general, participants were not entirely sure what information or tools they needed:

> *"I have to look it up individually. It's a tedious task since I need information for several hundred proteins."*

Inconsistent information or tool location within a resource was also a barrier for some participants:

> *"Protein localization information is not always located at the same position of web site."*

When working with large amounts of data from protein sequence databases, scientists tended to open two windows and adjust the size of each window to create customized concurrent views. For example, one participant noted the following:

> *"I opened two windows in my browser that shows identical web pages. And decrease the size of the window. And adjust the size of first window to upper half of my computer screen. And fix the first window for diagram. Second window is just for the lower table.*
> *By doing so, I can see the diagram the table together."*

All bench scientists expected the availability of upload and download data tools, version management of sequence data, and more flexible controls for page view navigation (especially for huge tabular data sets).

The forth category was 'computational analysis' (9.2%) was specifically related to application scientists' research. Some examples included simulating several large data sets, and applying the same two methods on both simulation data and real data in order to help understand the real data. *All application scientists* agreed on the need to import their own data or data from external sources, easily filter and integrate data, and access state-of-the-art visualization capabilities. The following two quotes illustrate examples:

> *"…. easy to upload my own data, easy to filter and integrate data, easy functions for matrix and vector operations."*

> *"It provides simple commands to calculate summary statistics, and easy to make plots (especially plotting different results together for comparison using different colors and line types)."*

With regard to the perceived usefulness of online resources and tools while conducting computational analysis, diary entries revealed that the ability to easily upload their own data, use functions (within packages) from public resources by importing them into specific software, apply simple commands to calculate summary statistics, and easily make plots were helpful features. The following quote illustrates these findings:

> *"Easy to upload my own data, easy to filter and integrate data, easy functions for matrix and vector operations."*

However, for much larger databases it was difficult to handle the data because some tools were not particularly well-suited to manipulations of large-scale data:

> *"It was impossible to import such a large data into [online resource]."*

As alternative ways to overcome common challenges, we identified that application scientists chose one of two common strategies; narrow down the candidate parameters and/or use computational server that has much more RAM and CPU to handle high-dimensional data sets.

The fifth category of distributed cognitive activities participants reported was 'creation of data algorithms' (7.1%). All application scientists tried to develop simulations and verify if their computer programs could identify true effects in the data, and determine how to fix errors in the code they made. User forums were reported to be helpful for tracking down errors and seeing how other people resolved them. The following quote shows this finding:

> *"The answers in the users group usually provide hints at best- they make assumptions about the base knowledge of those reading the forum."*

Also, all participants used online manuals to identify what options there are for various commands. Two participants, who are application scientists, explained their negative experiences due to technical terminologies:

"Though I can mostly understand the technical jargon now, I still learn better from the conversational tone."

"The manual is very technical, and seems more like it was written for the people who wrote the program in order to remind themselves of things they've forgotten, rather than for users who may not know the technical jargon."

To overcome several issues, participants made use of discussion forums and search engines. Another strategy was to assess new functions on small data sets before applying them to large data. For example, one participant remarked:

"I have to try the functions on small data sets first, before using it on large data… break down the whole genome into several segments."

All participants stressed the importance of collaboration and networking using a range of Web 2.0 resources such as discussion forums that specialize in openly posting ideas, tips, or best practices.

The sixth category of distributed cognitive activities participants reported was 'primer design' (5.1%). Particularly, bench scientists used various tools (e.g., NCBI, Primer 3.0, Soybase, QuantPrime) in order to design species-specific primers. Participants expressed that tools with in-built genome sequence selection functionality is helpful because it provides more specific results.

"It allows me to set different parameters for primer search. It gives me many primer pairs for one template."

However, many participants struggled with this task, since some tools give many primer pairs, and selecting few from these pair of primers is a critical and challenging task. Another challenge of primer design was to find a good sequence to work with:

"Sometimes they are predicted sequences or are not well annotated."

Participants also indicated that it is difficult to generate primers in practice:

> *"Generating the primers "by hand" can become confusing and increase the chance of making mistakes."*

Furthermore, understanding a protocol of a commercial kit with a lot of options hindered application scientists' research. Most participants tried to find as many files as possible on specific resources that describe the same gene, and from there select the best one. Instead of attempting to look up one tool, they tested results of the tool with other primer designing software.

> *"[I] just try to find as many files as possible on NCBI that describe the same gene, and from there select the best one."*

All application scientists were looking for ways to cross check results so that they can be more confident their outcomes. Also they emphasized that published protocols should be straightforward, usable processes.

Several of the documented activities did not naturally fall into one of the distributed cognitive categories listed above, For example, some bench scientists made use of cloud storage services (e.g., Dropbox) to access shared information or share information/results with lab members so that they can see the files at home and/or at the same time. However, there are compatibility issues that may hinder user experiences; for example, some programs did not recognize files in the Dropbox. Alternatively, s/he transferred the files in the Dropbox to their computer in order to import the file to the software they need to use. Another activity was annotation (e.g., characterization of a plasmid insert) because participants needed to clone one particular segment of the insert, but it was not well annotated. In addition, bench scientists tended to manage laboratory reagents using an online tool thereby they can share the current status with other lab personnel in real time. One bench scientist regularly makes their own protein data bank (including .pdb files) for different ligands so that they can incorporate them into simulations.

Figure 25 represented distributed cognitive activities by categories, the % of diary entries, and the # of diary entries for each category.

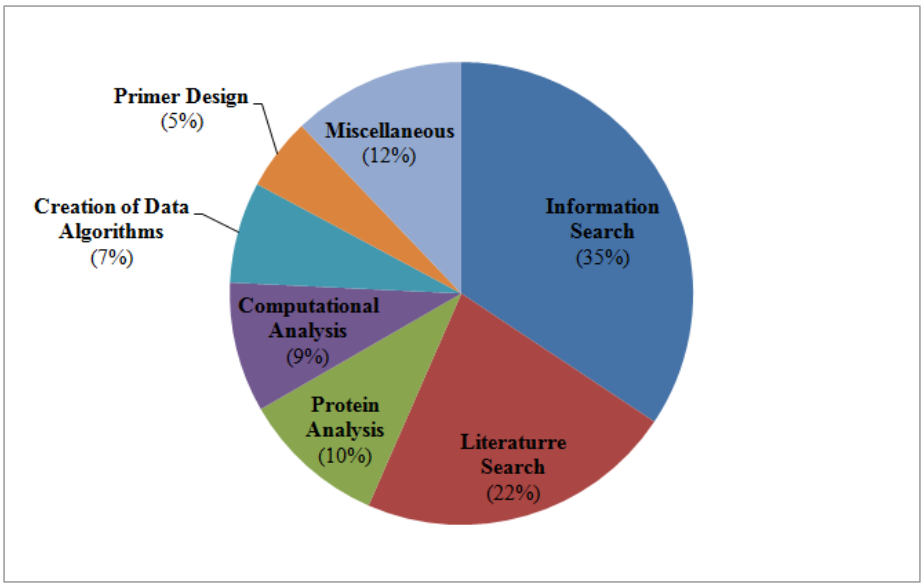


Figure 25. Percentage of distributed cognitive activities that participants performed.
Note. The numbers above each bar denote the number of diary entries.

4.4.3 Important Criteria that influence distributed cognitive activities

The most important criterion for all distributed cognitive activities was *ease of use*. Other highly valued criteria that influence tasks were *speed* and *responsiveness of resources*. The majority of participants consider the ability to ask questions and collect knowledge from blogs, academic portals, social networks, websites, or even personal contact via email just as important as the ability to search information and literature.

We also found several trends related to important criteria associated with different research roles. Application scientists in this study expect to be able to ask questions related to their work while developing and validating models and methodologies. In contrast, the need to ask questions or collect knowledge or information from other researchers was less attractive to bench scientists in a computer-mediated environment. Another finding was that *all bench scientists* noted advanced visualizations as the most suitable for conducting protein analysis and primer design. With respect to computational analyses, the ability to upload one's own data was universally highly rated.

However, our study shows that the ability to share knowledge with others was given a very low priority overall. Table 13 shows the important criteria that influence distributed cognitive activities.

Table 13. Important Criteria that influence each distributed cognitive activity
We asked participants to select their highest criteria priorities for each activity during the diary study.

Distributed Cognitive Activity / Criteria	Information Search	Literature Search	Computational Analysis	Protein Analysis	Creation of Data Algorithm	Primer Design
Speed and responsiveness of resource	**10.30%**	**15.20%**	**15.40%**	**24%**	5.90%	**13.60%**
Breadth of resource tools and functions	6%	9.10%	**19.20%**	4%	5.90%	9.10%
Wealth of available data	**15.50%**	**21.20%**	0%	**20%**	5.90%	**13.60%**
Degree of data integration	3.40%	3%	**15.40%**	4%	5.90%	**13.60%**
Advanced visualizations	3.40%	3%	7.70%	**12%**	0%	**18.20%**
Ability to upload my own data	0.90%	0%	**11.50%**	0%	5.90%	4.50%
Ability to share knowledge with others	5.20%	0%	0%	0%	0%	0%
Ability to ask questions related to my research	**11.20%**	6.10%	0%	0%	**29.40%**	0%
Ability to collect knowledge or information from other researchers	**19%**	**21.20%**	0%	4%	**11.80%**	9.10%
Ease of use	**16.40%**	**15.20%**	**23.10%**	**32%**	**29.40%**	**18.20%**
Other	8.60%	3%	0%	0%	0%	0%

4.4.4 Values to support distributed cognitive activities

We wanted to know what kinds of online distributed cognitive activities scientists would like to see supported. During a focus group session, participants identified where user experience brings values based on their experience and expectations (See Figure 26).

Figure 26. A focus group session.
(Photographed by the researcher, used with permission of all participants)

Participants wrote down about 30 opportunity areas on post-it™ notes. Then, they grouped and prioritized the notes into five opportunity areas to be considered: (1) ensuring reliability of data & tools, (2) providing all details and supporting information about data, (3) functions of tools, (4) performance in data processing, integrated data sources and tools, and (5) data accessibility. Priority was determined by consensus among participants (Figure 27). All participants preferred ensuring reliability of data and tools in various ways, for example, through the availability of features to compute different resources on published datasets, to choose ideal tools that meet research objectives, and to access appropriate and relevant information such as reliable scientists review comments or citation index.

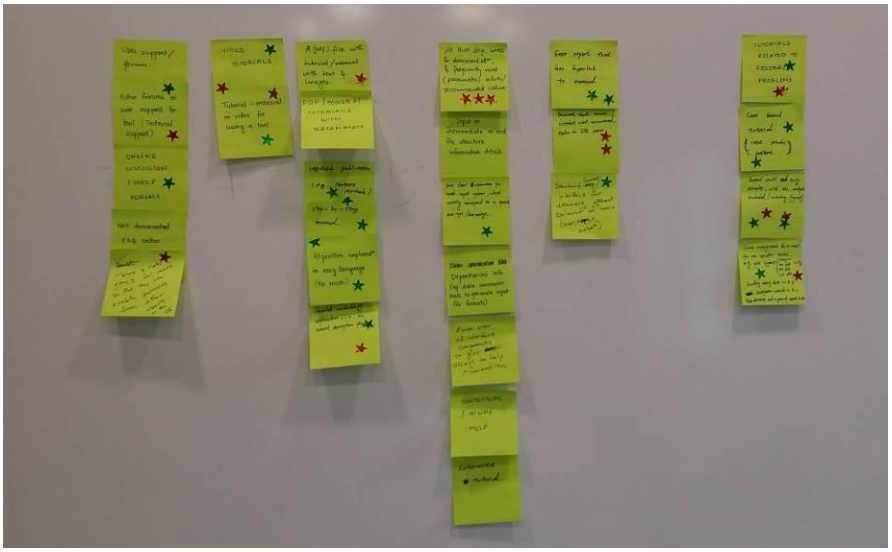


Figure 27. Example of opportunity areas.
Participants prioritized opportunity areas by affixing stickers on post-it™ notes

Given the volume and exponential growth in the number of tools with large datasets, *all participants* emphasized the importance of supporting information to easily understand data and functions in many ways, such as best practices with a set of cases, contextual help, and instructional information. *All participants* expected that the total processing time for handling and processing the highly amount of data such as importing, analyzing, visualizing, and even exporting datasets should be tolerable. Integration of heterogeneous data sources and tools has a value; this would dramatically improve research productivity and efficiency by allowing scientists to employ datasets from different databases in integrative analyses. Data accessibility was rated as a relatively lower priority compared to other values. Table 14 summarizes each opportunity with detailed descriptions and relative priories.

Table 14. Summaries of values and detailed opportunities

Values	Priority	Description
Ensuring reliability of data & tools	High	Help selecting the best tools to analyze previously published datasets (that are similar to mine).
		Accessible & reliable feedbacks/comments from other scientists
		Provenance for data and processes
		Ability to search (all) citations for specific protocols and pipelines
Providing all details and supporting information about data and functions of tools	High	More details in the manual of packages, such as the special type of data, standardization etc.
		Tools that are easy to understand and learn.
		Show examples of 'best practices'
		Contextual help information
Integrated data sources and tools	High	Integration of diverse tools
		Summarize results from different DBs
Performance in data processing	Medium	Acceptable processing time
		Ability to work with very large data sets (e.g., importing, analysis, visualizing, and exporting)
Data accessibility	Low	Availability + Accessibility of data in structured formats
		Easy to use open source data and visualization tools

4.5 Discussion

Previous studies have analyzed and classified the various types of tasks in bioinformatics (MacMullen & Denn, 2005; R. Stevens, Goble, Baker, & Brass, 2001; Tran et al., 2004), identified the need to integrate resources in support of a specific tasks (J. C. Bartlett & Toms, 2005; Kumpulainen & Järvelin, 2010), or reported usability issues (J. C. Bartlett et al., 2012). However, there appears to be scarce empirical knowledge published about user experiences in biological scientific research communities. The lack of such research creates a critical challenge for online system developers who seek ways to design systems that are accessible and usable to scientists.

For this work, we used a self-reporting technique to identify and capture user experiences during the context of use, and subsequently conducted a focus group session to draw out broader implications. The following discussion is structured around these two issues: 1) distributed cognitive activities derived from documented user experiences with knowledge resources and 2) key design considerations derived from the focus group session to support distributed cognition in biomedical and life sciences communities.

4.5.1 Distributed cognitive activities in biomedical and life sciences

The most obvious finding to emerge from this study is that *information search* and *literature search* tasks comprised the primary activities within distributed cognition environments regardless of scientists' research roles. While there are a wide range of systems to support information search and literature search, there is also the challenge of choosing which tools will provide the required information, as well as meet scientists' preferences. The vast numbers of online heterogeneous tools with associated complex data sets led to inconsistent results and inconsistent user interfaces/experiences (Bar-Ilan & Fink, 2005; Curtis & Weller, 1993; Hemminger, Lu, Vaughan, & Adams, 2007). As mentioned earlier (See Section 1.1.2), most scientists struggle with the iterative nature of complex research processes typically required to prove an initial hypothesis and further expect coherent outcomes from different steps along the way (J. Bartlett & Neugebauer, 2005; J. C. Bartlett & Toms, 2005; B. Mirel, 2009; Ouzounis, 2000; Tran et al., 2004). Obviously, a lack of consistency in results and user interfaces will

significantly affect user performance and cognitive load during iterative problem solving under these complex and ambiguous research circumstances (de Matos et al., 2013).

Overall, scientists expressed frustration in understanding technical jargon and acronyms irrespective of types of research activities. In addition, about two-third of our diary entries indicated that descriptions about tools are perceived to be written for experienced users (as opposed to novice users). In other cases, there is a lack of formal documentation, further hindering their ability to fully utilize a set of functions. As participants stressed in the focus group session, bioinformatics developers should take into account various user groups with different levels of domain knowledge and expertise. Further, developers should include supporting help and tutorial information as well as specific implementation details of tools and databases to increase understanding and increase efficiency (J. C. Bartlett et al., 2012; de Matos et al., 2013).

Lack of provenance- related information on data/tools also influences the perceived difficulty of searching information and searching literature across different resources (Simmhan, Plale, & Gannon, 2005; R. D. Stevens, Robinson, & Goble, 2003). It appears from our study, that scientists are discouraged from making use of many tools or datasets because they are unable to put their trust in these resources. In addition, integrative experiments, which inherently require use of multiple databases and multiple tools, further complicate research processes (Goecks, Nekrutenko, Taylor, & Team, 2010). As remarked by all participants, many tools across different online resources have similar functions but vary in complexity, user interaction and/or visualization capabilities (e.g., genome browsing), so the choice is not as straightforward as simply selecting one tool to conduct their research. Participants reported that a major obstacle is getting access to the right tools at the right time (in part, because many tools have similar yet slightly different functions). Even if participants found the right tool, they ultimately still struggled to understand how to use the tool to it fullest extent. Thus, scientists must not only navigate the complexity of the tools themselves, and the information they provide, but also overcome the challenge of determining which tools(s) to select to accomplish their research goals (J. C. Bartlett et al., 2012). These findings are congruent with our focus group session results stating the need for better integrated resources and databases. In this sense, system developers should consider compatibility with other tools and datasets, so that scientists are able

to move between resources easily and obtain results quickly with minimal effort. Regardless of what distributed cognitive activities occur, online bioinformatics resources should take into account a scientist's current task and provide intuitive interface mechanisms to address user needs.

As evident in data entries across many of the diaries, scientists suggest that the availability of open online spaces to share and discuss their work would help them when conducting research and exploring potential solutions. Scholarly publications in this area of science defines this concept as "Science 2.0"; which generally refers to "new practices of scientists who post raw experimental results, nascent theories, claims of discovery and draft papers on the Web for others to see and comment on" (Waldrop, 2008). In examining the dairy data, there were many entries related to key aspects of Science 2.0, such as sharing and asking about research experiences, useful resources, and/or interesting results amongst people with similar domain interests. Some scientists who were not comfortable with open online resources preferred to contact authors or senior researchers via email, blogs, academic portals, or social networks. This finding suggests the need for more UX support to allow scientists to follow and link up with researchers of interested. A perceived benefit of this approach is that connecting to others in the field will result in more collaborative and productive scientific progress (Waldrop, 2008).

One interesting finding from the diary entries was differences in distributed cognitive activities identified between bench scientists and application scientists. To be specific, bench scientists reported that they have more interest in, and are more eager to use, advanced visualization tools that not only make it easier to work with large datasets (collected from disparate online resources) but also offer powerful yet easy to use workflows.

Application scientists on the other hand, were more apt to engage in open community platforms than bench scientists, suggesting that bench scientists tend to share resources and communicate mainly with known acquaintances in their established social networks. These results are consistent with previous studies that reported different attitudes toward data sharing in different scientific disciplines (Birnholtz & Bietz, 2003). According to Birnholtz and Bietz (2003)'s finding, HIV/AIDS researchers, who are mainly involved in laboratory experiments, tend to talk directly with colleagues to figure out how to produce data and are quite open to share laboratory techniques between labs. But they are prone to be anxious about getting 'scooped', so they less

willing to share unpublished data. On the other hands, theoretical modelers, who utilize empirical data to develop and validate their models in earthquake engineering community, showed more interests in the data-sharing because they consider data as an external benchmark to develop models. These findings have significant implications for considering the requirements of different user groups in the design and development of bioinformatics, rather than assuming that one size fits all (J. C. Bartlett et al., 2012).

In addition, our results suggest that some online resources fail to adhere to basic HCI principles such as consistency in user interface, structure, and even information. Thus, lack of usability was identified as a major limitation, especially when processing huge amounts of ambiguous and uncertain biological data.

4.5.2 Key design considerations

Previous research indicates some important characteristics of online tools needed to support bioinformatics research communities (J. C. Bartlett et al., 2012; Javahery, 2004; Barbara Mirel & Wright, 2009), but most studies rely on indirect measurement approaches such as questionnaires or heuristics examining very short time frames of use, and typically examining one online resource at a time. As such, we believe that the resulting implications are potentially limited. This study contributes to the body of knowledge by extending previous basic findings related to online bioinformatics resource UI design and identifies specific UX issues that scientists encountered across various online resources over the course of two weeks. While we identified some differences in distributed cognitive activities among participants, other findings common across participants will help improve the efficacy of bioinformatics UI designs to accommodate long-term user experiences in real contexts of use, and thus will be effective for the general population of scientists in the field. Based on several empirical findings of the current study, we suggest design considerations to support scientists' distributed cognitive activities. These design considerations are intended for user experience professionals and system developers working in biomedical and life sciences who seek ways to better support research communities from user experience perspectives.

The presented design considerations are intended to be applied in the context of an online bioinformatics resource:

- **Avoid technical jargons, acronyms, and abbreviations**

  Regardless of knowledge levels or disciplines, users should be able to understand any terminology and work with the tools without any difficulties.

  - Take into account users with different levels of domain knowledge and experience
    - E.g. display glossary definitions using tooltips, popup windows or links, if necessary

- **Enable scientists to make use of tools easily**

  Given increasingly complex tools, interested scientists should be able to easily view and make sense of basic implementation details.

  - Give users a clear and detailed instructions in various such forms as:
    - Step-by-step user guides
    - FAQs
    - Related information (related keywords, link to related internal and external resources)
    - Keyword suggestions
  - Provide the rationale for including/requiring a step in a workflow or process
  - Use easy to understand input parameters and easy to read and/or use output data
  - Provide support on how to interpret or further utilize results, by for example,
    - Case-based tutorials with clear explanations for input/out options
    - Demonstrations on how to utilize tools
    - Interactive graphical features

- Online discussion/help forums, with “most popular” questions & answers
- Keyword suggestions

Error messages should be clear so that users do not repeat the same errors.

- Enable users to handle errors by providing, for example:
  - Contextual information to prevent potential errors
  - Documentation about common mistakes users encountered and how to recover
  - Error reports with includes a hyperlink to a user guide, that should further include what step(s) to follow next
    - Technical assistance forum that includes features such as “Most popular” keywords, and Frequently asked questions
- Use consistent, concise, and clear labels and text for menu labels, dialog messages, help messages, and tooltip text.

- **Ensure consistent user experience**

To ensure consistent user experiences, it is important to maintain consistency within and across tools.

- Ensure a consistent look and feel
- Strive for consistency of results (e.g., terminology, information, annotation)

- **Consider integrated data sources and tools**

Support compatibility between users’ tools/data sets and equivalent tools/data sets so that users can determine for themselves if a given dataset or tool is the right one for the job. This can be accomplished by for example:

- Support ability to access published databases

    - Ensure compatible with other tools or data sets, by for example, supporting a large set of import and export data formats

- **Reliable data sources and tools**

Exceedingly large dataset sizes and the deluge of tools and protocols generated in recent years have placed additional burdens on users to select appropriate and reliable data sources and tools. A well designed online bioinformatics resource will ensure reliability of data & tools in a various ways, including:

- Show provenance information such as origin and history of sources or tools and information about authors/publishers
- Provide reliability indicators such as:
    - Citation indices
    - Scope of current grants or funding
    - Community rating systems including review scores and recommenders' comments
    - Version update information
    - Most popular
- Enable users to do comparisons on different datasets and generate easily exported visual and text-based explanations of differences/similarities.
- Help users link to people working in similar fields.

- **Ensure system responsiveness**

Due to long processing times, scientists using online resources with large data sets were often not sure whether or not the tools were stuck, broken or simply taking a while to run. The length of processing time should be acceptable and tolerable.

If systems cannot support expected responsiveness, an activity indicator, along with an informative label should be used to help to reduce users' uncertainty and perceived waiting duration. For example:

- Show an activity status bar, that for example, displays estimated time to completion

Another way to ensure system responsiveness is to provide advanced options to refine or filter results prior to executing queries or analysis tools. This approach further helps users narrow down their search results a priori and subsequently quickly find what they are looking for in the result set. For example:

- Provide advanced search options such as:
    - Filter by available metadata (e.g., type of data format)
    - Sort by relevance, date, etc.
- Allow users to see preview results

- **Support internal and external communication and collaboration**

Open sharing of datasets, workflows, protocols, and experimental results through internal and external communication will promote accountability and collaboration in scientific communities. This could be accomplished by, for example:

- Open repositories or archives to share datasets, histories, workflows, and protocols that include:
    - User publishing mechanisms for datasets, analyses, and workflows
    - Import features that allow users to upload their own datasets into workspaces
- Supported communication between members of a network
- Features that enable users to post comments and review the feedback of others
- Features that enable users to search threads on the repository

- **Support flexibility to accommodate multiple diverse user populations**

It is good UX practice to grant users control of how their information is presented to themselves, to private collaborators and to the general public. This is particularly needed when multiple tasks are conducted across multiple collaborating groups, or when customized (e.g., proprietary) analysis are needed. Thus, online bioinformatics resources should:

- Enable users to customize
    - Organization or visualization of datasets with both embedded components and custom parameters
    - The structure or format of results
    - The size of the frame on the layout (for publication quality image export)
    - The set of user classes allowed to access various views of data and/or visualizations.

Our results also suggest that scientists value flexibility in a tool that allows them to specify their own parameters for their analysis. Examples of such flexibility include:

- Ability to specify own parameters
- Ability to download results in other formats
- Ability to search, filter or group large quantitative of data by various parameters (e.g., keyword, author, tag, and annotation)

- **Promote reproducibility**

In consideration of scientific research process characteristics (e.g., iterative and complex), our results suggest that reproducibility is desired since it will enhance the productivity of scientists under otherwise complex and ambiguous research circumstances. Reproducibility can be supported through:

- Reuse of outcomes such as the ability to import datasets into personal workspaces, upload custom datasets, and import workflows and datasets to run said workflows (to name a few).

- Related threads in accordance with user input
- Ability for scientist to add descriptions or notes about datasets, histories or analyses steps to explain why a particular step is needed or important.
    - Inclusion of user metadata such as annotations and tags to assist in future searching and facilitate reuse, and annotation to aid in understanding analysis at later dates or when sharing.

4.6 Conclusions

Although many studies tout the importance of naturalistic observation and investigation within biomedical and life science communities (Cohen, Blatter, Almeida, Shortliffe, & Patel, 2006; B. Mirel, 2009), these studies rarely consider user experiences in actual long term usage contexts (J. Bartlett & Neugebauer, 2005; J. C. Bartlett et al., 2012; Bolchini, 2009; de Matos et al., 2013; Javahery, 2004; Barbara Mirel & Wright, 2009; Pavelin et al., 2012; Tran et al., 2004).

This research extends our knowledge of distributed cognitive activities of scientists within big data environments. Although we observe important areas of overlap in fundamental aspects of usability (e.g., ease of use, accessibility, and consistency in user interface), our findings make the compelling case that the consideration of fundamental usability issues confronting users was not enough to support distributed cognitive activities under highly complex research circumstances. Rather, the user experiences explained in Section 4.5.2 allowed us to learn the complexities of *when* and *why* different knowledge resources work and do not work within a distributed cognitive environment. Through this lens, we gain a holistic understanding of current practices, for example, understanding how scientists make use of high volumes of knowledge resources and cope with barriers over time. The additional information provided in the focus group session was also valuable as it helped to identify the detailed design considerations to enrich user experiences in data-intensive and interdisciplinary research communities. These findings can be valuable for UX professionals who seek UX solutions to support big data research communities.

The empirical findings in this study revealed in-depth understanding of value perspectives of knowledge-sharing and -reuse activities that are recognized from end user perspectives in real-world practices. These can be used to identify potential design opportunities to overcome

knowledge barriers and enhance productivity among scientists in biomedical and life sciences communities.

With regard to methodological contributions, this research demonstrated synergistic effects of combining the diary study with focus group interviews. For instance, the diary study allows us to understand particular events and experiences in natural usage contexts and in less intrusive ways (as compared to classic approaches such as ethnography). Moreover, diary entries can offer the space for intimate discussions about specific events that could not be fully outlined in general focus group interviews. Thus, this multifaceted approach could motivate other UX/UI practitioners in designing studies to better examine distributed cognitive activities in complex and dynamic environments.

4.7 Future directions

We conducted a two-week naturalistic research study to understand scientists' distributed cognitive activities using qualitative approaches. In this way, this research contributed to understanding scientists' experience in their research contexts, and informs a set of design consideration we put forth in section 4.6.2.

We acknowledge that a relatively short timeframe of self-reporting (i.e., two weeks) may be weakness of this study. It is possible that there were some unknown patterns of distributed cognitive activities. Nevertheless, the amount of data from diary entries were enough to understand current user experience levels (e.g., perceived helpfulness, challenges, and alternative actions), as well as important criteria associated with each research activity in a distributed cognitive environment.

Since we have only targeted the experiences of a very special population (all participants had at least a post graduate degree in biomedical and life sciences and made use of bioinformatics tools almost every day), our findings are worthy of further study with different user groups in this domain such as a beginning college students in life science or scientists who have not routinely performed bioinformatics analysis. Also, we did not investigate and control for organizational culture. Hence, more emphasis on considering perspectives of research community culture could also be explored.

In light of our results, we believe that continuing research will yield new kinds of insights for bioinformatics tools that promise to enhance the productivity and satisfaction of scientists in distributed cognition environments.

4.8 References

Andrews, Frank M, Robinson, John P, & Wrightsman, Lawrence S. (1991). *Measures of personality and social psychological attitudes* (Vol. 1): Gulf Professional Publishing.

Bar-Ilan, Judit, & Fink, Noa. (2005). Preference for electronic format of scientific journals—A case study of the Science Library users at the Hebrew University. *Library & Information Science Research, 27*(3), 363-376.

Bartlett, J., & Neugebauer, T. (2005). Supporting information tasks with user-centred system design: The development of an interface supporting bioinformatics analysis. *Canadian journal of information and library science, 29*(4), 486-487.

Bartlett, Joan C., Ishimura, Yusuke, & Kloda, Lorie A. (2012). *Scientists' preferences for bioinformatics tools: the selection of information retrieval systems*. Paper presented at the Proceedings of the 4th Information Interaction in Context Symposium, Nijmegen, The Netherlands.

Bartlett, Joan C., & Toms, Elaine G. (2005). Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach. *Journal of the American Society for Information Science and Technology, 56*(5), 469-482. doi: 10.1002/asi.20136

Bayat, Ardeshir. (2002). Bioinformatics: Science, medicine, and the future. *BMJ, 324*, 1018-1022.

Birnholtz, Jeremy P, & Bietz, Matthew J. (2003). *Data at work: supporting sharing in science and engineering.* Paper presented at the Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work.

Bogdan, Robert Charles, & Biklen, Sari Knopp. (1982). *Qualitative research for education*: Allyn and Bacon Boston.

Bolchini, D. (2009). Better bioinformatics through usability analysis. *Bioinformatics (Oxford, England), 25*(3), 406-412. doi: 10.1093/bioinformatics/btn633

Bolger, Niall, Davis, Angelina, & Rafaeli, Eshkol. (2003). Diary methods: Capturing life as it is lived. *Annual review of psychology, 54*(1), 579-616.

Brunswik, Egon. (1941). *Systematic and representative design of psychological experiments.* Paper presented at the Proc. Berkeley Symp. Math. Stat. Probab.

Chilana, Parmit K., Wobbrock, Jacob O., & Ko, Andrew J. (2010). *Understanding usability practices in complex domains*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA.

Cohen, Trevor, Blatter, Brett, Almeida, Carlos, Shortliffe, Edward, & Patel, Vimla. (2006). A cognitive blueprint of collaboration in context: Distributed cognition in the psychiatric emergency department. *Artificial intelligence in medicine, 37*(2), 73-83.

Curtis, Karen L, & Weller, Ann C. (1993). Information-seeking behavior: a survey of health sciences faculty use of indexes and databases. *Bulletin of the Medical Library Association, 81*(4), 383.

Czerwinski, Mary, Horvitz, Eric, & Wilhite, Susan. (2004). *A diary study of task switching and interruptions.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

de Matos, Paula, Cham, Jennifer A, Cao, Hong, Alcántara, Rafael, Rowland, Francis, Lopez, Rodrigo, & Steinbeck, Christoph. (2013). The Enzyme Portal: a case study in applying user-centred design methods in bioinformatics. *BMC bioinformatics, 14*(1), 103.

Dell, Nicola, Vaidyanathan, Vidya, Medhi, Indrani, Cutrell, Edward, & Thies, William. (2012). *Yours is better!: participant response bias in HCI.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Fleiss, L, Levin, Bruce, & Paik, Myunghee Cho. (1981). *The measurement of interrater agreement.* Paper presented at the In Statistical methods for rates and proportions (2nd ed.

Gabbard, Joseph L, Hix, Deborah, Swan, II, Livingston, Mark A, Hollerer, Tobias H, Julier, Simon J, . . . Baillot, Yohan. (2003). Usability engineering for complex interactive systems development: DTIC Document.

Goble, Carole A., Stevens, Robert, Ng, Gary, Bechhofer, Sean, Paton, Norman W., Baker, Patricia G., . . . Brass, Andy. (2001). Transparent access to multiple bioinformatics information sources. *IBM Systems Journal, 40*(2), 532-551.

Goecks, Jeremy, Nekrutenko, Anton, Taylor, James, & Team, T Galaxy. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol, 11*(8), R86.

Greenbaum, Thomas L. (1997). *The handbook for focus group research*: SAGE Publications, Incorporated.

Hemminger, Bradley M, Lu, Dihui, Vaughan, KTL, & Adams, Stephanie J. (2007). Information seeking behavior of academic scientists. *Journal of the American society for information science and technology, 58*(14), 2205-2225.

Hesse-Biber, Sharlene Nagy, & Leavy, Patricia. (2010). *The practice of qualitative research*: Sage.

Javahery, Homa. (2004). Beyond power making bioinformatics tools user-centered. *Communications of the ACM, 47*(11), 58.

Jones, Martyn, & Johnston, Derek. (2011). Understanding phenomena in the real world: the case for real time data collection in health services research. *Journal of Health Services Research & Policy, 16*(3), 172-176.

Kitzinger, Jenny. (1995). Qualitative research. Introducing focus groups. *BMJ: British medical journal, 311*(7000), 299.

Kumpulainen, Sanna, & Järvelin, Kalervo. (2010). *Information interaction in molecular medicine: integrated use of multiple channels.* Paper presented at the Proceedings of the third symposium on Information interaction in context.

Lallemand, Carine, Gronier, Guillaume, & Koenig, Vincent. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior, 43*(0), 35-48. doi: http://dx.doi.org/10.1016/j.chb.2014.10.048

LeCompte, Margaret D, & Goetz, Judith Preissle. (1982). Problems of reliability and validity in ethnographic research. *Review of educational research, 52*(1), 31-60.

MacMullen, W John, & Denn, Sheila O. (2005). Information problems in molecular biology and bioinformatics. *Journal of the American Society for Information Science and Technology, 56*(5), 447-456.

Marchionini, Gary. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science, 40*(1), 54-66.

Mirel, B. (2009). Supporting cognition in systems biology analysis: findings on users' processes and design implications. *J Biomed Discov Collab, 4*, 2. doi: 10.1186/1747-5333-4-2

Mirel, Barbara, & Wright, Zach. (2009). Heuristic evaluations of bioinformatics tools: a development case *Human-Computer Interaction. New Trends* (pp. 329-338): Springer.

Möller, Andreas, Kranz, Matthias, Schmid, Barbara, Roalter, Luis, & Diewald, Stefan. (2013). *Investigating self-reporting behavior in long-term studies.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Morgan, David L. (1992). Designing focus group research. *Tools for primary care research, 2*, 177-193.

Morgan, David L. (1996). Focus groups. *Annual review of sociology*, 129-152.

Ouzounis, Christos. (2000). Two or three myths about bioinformatics. *Bioinformatics, 16*(3), 187-189. doi: 10.1093/bioinformatics/16.3.187

Pavelin, Katrina, Cham, Jennifer A, de Matos, Paula, Brooksbank, Cath, Cameron, Graham, & Steinbeck, Christoph. (2012). Bioinformatics meets user-centred design: a perspective. *PLoS computational biology, 8*(7), e1002554.

Powell, Richard A, & Single, Helen M. (1996). Focus groups. *International journal for quality in health care, 8*(5), 499-504.

Redish, Janice. (2007). Expanding usability testing to evaluate complex systems. *Journal of Usability Studies, 2*(3), 102-111.

Reis, Harry T. (1994). Domains of experience: Investigating relationship processes from three perspectives.

Simmhan, Yogesh L, Plale, Beth, & Gannon, Dennis. (2005). A survey of data provenance in e-science. *ACM Sigmod Record, 34*(3), 31-36.

Stevens, Robert D, Robinson, Alan J, & Goble, Carole A. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics, 19*(suppl 1), i302-i304.

Stevens, Robert, Goble, Carole, Baker, Patricia, & Brass, Andy. (2001). A classification of tasks in bioinformatics. *Bioinformatics, 17*(2), 180-188.

Tran, D., Dubay, C., Gorman, P., & Hersh, W. (2004). Applying task analysis to describe and facilitate bioinformatics tasks. *Stud Health Technol Inform, 107*(Pt 2), 818-822.

Waldrop, M Mitchell. (2008). Science 2.0. *Scientific American, 298*(5), 68-73.

Wheeler, Ladd, & Reis, Harry T. (1991). Self-Recording of Everyday Life Events: Origins, Types, and Uses. *Journal of personality, 59*(3), 339-354.

Wild, Peter J, McMahon, Chris, Darlington, Mansur, Liu, Shaofeng, & Culley, Steve. (2010). A diary study of information needs and document usage in the engineering domain. *Design Studies, 31*(1), 46-73.

ZIMMERMAN, DONH, & Wieder, D Lawrence. (1976). The diary. *American Sociologist*.

5 Chapter 5: Conclusions and recommendations

The overall objective of this dissertation was to examine user experiences with data-intensive online bioinformatics resources, from a number of different perspectives, and propose a set of UX design considerations for bioinformatics resources to support biomedical and life sciences research communities. Specifically, the current research has addressed perspectives of distributed cognitive activities and the impacts of individual differences (research disciplines and cognitive styles) on insight generation behavior and human performance to satisfactorily support user experiences in data-intensive interdisciplinary research communities.

5.1 Summary

Study 1 empirically examined scientists' user experience levels with current online bioinformatics resources and key determinants that influence their distributed cognitive activities (or lack thereof), especially knowledge networking (Chapter 2). An empirical survey revealed that many online bioinformatics resources are designed for highly specialized domain experts and require highly technical skills. With regard to distributed cognitive activities (especially knowledge networking), most scientists have been making use of different types of online knowledge networking platforms (scientific resource sharing platforms and collaborative discussion forums) in order to combine their specialized knowledge with new insights from others, but nonetheless active participation (i.e., sharing and contributing knowledge) by the bioinformatics research community in knowledge networking platforms is rather low. The results from a partial least squares analysis revealed that 1) scientists were strongly influenced by source credibility and fear of being scooped in the competitive nature of the scientific communities, and 2) knowledge networking activities among scientists were influenced by both extrinsic (e.g., reciprocal benefit, reputation) and intrinsic (e.g., altruism) motivations within professional networks.

Our second study investigated the impacts of individual differences (i.e., research role and cognitive style) on insight generation behavior (i.e., behavioral characteristics, insight characteristics, and gaze characteristics) and human performance (i.e., slip, lapse, and mistake), especially in the bioinformatics experiment settings with high cognitive demands (Chapter 3). A

laboratory experiment confirmed significant differences with respect to insight generation behavior and human performance depending on research roles (bench and application scientists). More specifically, bench scientists 1) took shorter time to find the first insight, 2) reported fewer insights, 3) made greater use of text-based content forms, and 4) made less slips as compared to application scientists. This study also confirmed relationships between scientists' cognitive styles and human performance suggesting that field dependent individuals 1) generated more fixations on visually-represented data and 2) made more slips than field independent individuals.

The third study identified in-context usefulness and barriers of knowledge resources and the challenges and alternative actions scientists faced in real work contexts to derive focused design considerations (Chapter 4). A longitudinal diary study revealed that information search and literature search tasks comprised scientists' primary activities within distributed cognition environments. We identified a handful of common barriers independent of any specific distributed cognitive activity including technical jargon and acronyms, lack of provenance-related information on data/tools, inconsistent results (e.g., different versions of gene sequences) across different online resources, and low-speed data processing. We also found differences between preferences of the user groups with bench scientists more interested in, and more eager to use, advanced visualization tools that make it easier (for them) to work with large datasets. We also revealed that application scientists are more apt to engage in open community platforms, while bench scientists tend to share resources and communicate mainly with known acquaintances in their established social networks. However, all scientists commonly emphasized the importance of open online spaces to share and discuss their work in computer-mediated environments. By synthesizing complementary results from Study 1 and Study 2, we suggest that bioinformatics resources:

- Avoid technical jargons, acronyms, and abbreviations.
- Enable scientists to make use of tools easily.
- Ensure consistent user experiences within and across tools.
- Consider integrated data sources and tools.
- Ensure reliable data sources and tools.
- Ensure system responsiveness.
- Support internal and external communication and collaboration.

- Support flexibility to accommodate multiple diverse user populations.
- Promote reproducibility.

5.2 Contribution

This dissertation introduced distributed cognition as a theoretical framework for capturing holistic perspectives (ranging from fundamental requirements such as usability issues to in-depth insights from longitudinal contexts of use) to enrich user experiences in data-intensive research communities. Given the lack of understanding of users' values, and expectations inherent in different user contexts in biomedical and life sciences, the distributed cognition framework provided deeper insights into how scientists extend their cognitive capabilities to attain research goals under uncertain, complex and ambiguous environments. In order to examine distributed cognitive activities, the integration of a mixed-methods approach into bioinformatics will also serve as a practical example to both the UX research and professional communities.

This research initially identified distinct preferences and performance gaps in a population of scientists. Understanding differences in research expertise allowed us to address system considerations arising from interdisciplinary perspectives, thereby supporting consistent user experiences among interdisciplinary research communities composed of professionals with different knowledge backgrounds and skills. Cognitive style differences can also provide a lens that facilitates the exploration of more persuasive systems to support the broad spectrum of user classes. For example, results of this work can help bioinformatics designers anticipant how specific knowledge representations and interaction techniques may be perceived by different user groups and how those representations and interaction techniques may lead to meaningful insights.

It is important to note that there are still fundamental challenges for supporting end-users who engage in complex domains. Several studies have identified usability professionals' challenging issues due to lack of expert knowledge about the subject domain, limited access to target users, natural empathy towards users, and uniqueness of each situation (planning studies, conducting studies, and interpreting results) (Chilana, Wobbrock, & Ko, 2010; Pavelin et al., 2012). In light of these matters, user experience practitioners in biomedical and life sciences can apply our research findings as fundamental drivers for online bioinformatics resource user interface design.

5.3 Recommendations for future research

This research has consistently proved the importance of effectively supporting knowledge networking in data-intensive interdisciplinary research communities. All scientists commonly stress the importance of open spaces to share and discuss in a computer-mediated environment. Nonetheless, only a small number of scientists actively have participated in knowledge contribution. Ideal knowledge networking would not be a unilateral mechanism to only take knowledge from others. In order to promote a constructive culture of knowledge networking, a possible venue for future research is to focus more on determinants of scientists' intention to share knowledge in biomedical and life sciences research communities.

It is noteworthy that previous literature has mainly focused on sharing data (Birnholtz, 2007; Field et al., 2009; Kaye, Heeney, Hawkins, De Vries, & Boddington, 2009). As discussed in Chapter 2 and Chapter 4, scientists do indeed utilize both data and acquired knowledge from educational training and practical experience. Thus, identifying the impacts of type of knowledge (e.g., explicit and implicit knowledge) on scientists' intention to share knowledge also warrants further attention.

It is obvious that scientific discoveries in biomedical and life sciences requires a high cognitive load not only for gaining valuable insights but also for making decisions, often under highly ambiguous and complex conditions. By leveraging both concurrent and retrospective verbal protocols, our studies revealed some evidence (i.e., human errors) of moderate to high cognitive loads while generating insights. Tversky and Kahneman (1974) noted that "People rely on a limited number of heuristics principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations." It seems clear, then, that investigating biases in insight generation processes could be a useful research endeavour because there is an opportunity to identify strategies to counter biases or support debiasing judgement to ultimately help scientists make better decisions.

Finally, the ultimate goal of this research was to suggest key design considerations to enrich user experiences to diverse groups of professionals from different disciplinary backgrounds. We acknowledge that our findings may or may not be applicable to beginner bioinformatics researchers. Thus, comparisons among beginner scientists from different disciplines may also

deserve future attention in order to address interdisciplinary nature and extensive technological changes in biomedical and life sciences.

5.4 References

Birnholtz, Jeremy P. (2007). When do researchers collaborate? Toward a model of collaboration propensity. *Journal of the American Society for Information Science and Technology, 58*(14), 2226-2239.

Chilana, Parmit K., Wobbrock, Jacob O., & Ko, Andrew J. (2010). *Understanding usability practices in complex domains*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA.

Field, Dawn, Sansone, Susanna-Assunta, Collis, Amanda, Booth, Tim, Dukes, Peter, Gregurick, Susan K, . . . Maxon, Mary. (2009). 'Omics data sharing. *Science (New York, NY), 326*(5950), 234.

Kaye, Jane, Heeney, Catherine, Hawkins, Naomi, De Vries, Jantina, & Boddington, Paula. (2009). Data sharing in genomics—re-shaping scientific practice. *Nature Reviews Genetics, 10*(5), 331-335.

Pavelin, Katrina, Cham, Jennifer A, de Matos, Paula, Brooksbank, Cath, Cameron, Graham, & Steinbeck, Christoph. (2012). Bioinformatics meets user-centred design: a perspective. *PLoS computational biology, 8*(7), e1002554.

Tversky, Amos, & Kahneman, Daniel. (1974). Judgment under uncertainty: Heuristics and biases. *science, 185*(4157), 1124-1131.

Appendices

Appendix A - IRB approval (Study 1)

VirginiaTech

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: August 18, 2014

TO: Joseph L Gabbard Jr, Jongsoon Park

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)

PROTOCOL TITLE: Knowledge sharing motivations in the context of bioinformatics

IRB NUMBER: **12-754**

Effective August 15, 2014, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the Continuing Review request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 7**
Protocol Approval Date:	**August 31, 2014**
Protocol Expiration Date:	**August 30, 2015**
Continuing Review Due Date*:	**August 16, 2015**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Appendix B - Survey sample (Study 1)

Our designers are currently working to define how best to support user experience in Biomedical and Life Science Communities; from sharing data within a small group of colleagues to sharing expertise and opinions across the broader bacterial community.

Your input is an invaluable part of this design process. Help us learn about needs and interests by taking our short survey. Your responses will be anonymous and you may participate in a lottery for one of twelve $25 Amazon gift cards. If you want to participate in the lottery, please make sure to leave your email at the end of the survey. We estimate that approximately 150 individuals will participate in this survey, thus your odds of winning can be estimated at 12/150. Each participant has an equal opportunity for winning.

It will take about 20 minutes; Your time and opinions are greatly appreciated!

** This research is happening at Virginia Tech*

I am a:

☐ Biologist ☐ Bioinformatician ☐ Chemist ☐ Computer Scientist ☐ Mathmematician

Other: _____

How many years have you worked in this role?
O I have no work experience at all.
O I only did internships.
O I have been working since 1 year.
O I have been working since 1-5 years.
O I have been working since 5-10 years.
O I have been working since more than 10 years.

Which online bioinformatics resources (e.g. tools and/or websites) do you use the most?

_____

During last six months, how often did you use bioinformatics tools & web sites?
O Almost everyday O More than once a week O More than once a month
O Around once a month

How did these online bioinformatics resources help you in your research/work?

_____

What are the major shortcomings or limitations in online bioinformatics resources you currently use?

_____

Which factors are more important when you use online bioinformatics resources (check all that apply)?

☐ Speed and responsiveness of resource
☐ Breath of resource tools and functions
☐ Wealth of available data
☐ Degree of data integration
☐ Advanced visualizations
☐ Ability to upload my own data
☐ Ability to share knowledge or information with other researchers
☐ Ability to ask questions related to my research
☐ Ability to collect knowledge or information from other researchers
☐ Ability to create publication quality images
☐ Ease of use
Other:_____

"Knowledge-sharing is an activity through which knowledge (i.e. information, skills, or expertise) is exchanged among people, friends, or members of a family, a community (e.g. Wikipedia) or an organization." – Wikipedia
In the online context, knowledge-sharing activities can involve knowledge contribution and knowledge seeking via the internet such as consuming useful information, asking questions, answering comments and/or questions, and sharing one's own knowledge or valuable information with others.

Rate your experience with online resources that support knowledge-sharing activities to share similar interests and values with others in your research field.

O 1) I have only seen them.
O 2) Little experience, I have used them only a few times for my work or research.
O 3) Experienced, I use them regularly to do my work or research.
O 4) Very experienced, I use them for almost all work or research.

If you choose 2), 3), or 4) in the above question, Name or URL of online resources:

_____

If you wrote URLs or names in the above question, which of the following are your main activities (check all that apply)?

☐ I browse social network services and consume useful information.
☐ I ask questions to others.
☐ I reply to simple questions from others.
☐ I try to form strong ties to the community.
☐ I make more elaborate comments.
☐ I answer complex questions.
other: _____

To be specific, what kinds of information did you post/write in social networking services (check all that apply)?
☐ None
☐ Accumulative working experiences
☐ Unique opinions
☐ General ideas
☐ Articles published in books, periodicals, magazines, websites, documents, manuals, handout materials and so on.
☐ Products, patents, software code, computer databases, technical drawings, tools, prototypes, photographs, voice recordings, films and so on.
☐ Rules, routines, or operating procedures
other: _____

If you choose several options in the above question, on average, how many times per week did you engage in online knowledge-sharing activities for your research?
O Not at all
O Less than once a week
O About once a week
O 2 or 3 times a week
O Several times a week
O About once a day
O Several times each day

To be specific, what kinds of information did you find from others there (check all that apply)?
☐ None
☐ Accumulative working experiences
☐ Unique opinions
☐ General ideas
☐ Articles published in books, periodicals, magazines, websites, documents, manuals, handout materials and so on.
☐ Products, patents, software code, computer databases, technical drawings, tools, prototypes, photographs, voice recordings, films and so on.
☐ Rules, routines, or operating procedures
other: _____

If you choose several options in the above question, on average, how many times per week did you engage in online knowledge-sharing activities for your research?
O Not at all
O Less than once a week
O About once a week
O 2 or 3 times a week
O Several times a week
O About once a day
O Several times each days

If you never post to the online group/community in your research field, what were your

reasons? (check all that apply)
☐ Just reading/browsing is enough
☐ Want to remain anonymous
☐ Shy about posting
☐ Others respond the way I would
☐ Had no intention to post from the outset
☐ If I post, I am making a commitment
☐ Nothing to offer
☐ Wrong group for me
☐ Do not know how to post to this group
☐ Still learning about the group
☐ There are too many messages already
☐ Poor quality of messages or group/community
☐ No requirement to post
☐ Group treats new members badly
☐ Concern about aggressive or hostile responses
☐ Long delay in response to postings
☐ Of no value to me
☐ My work does not allow posting
☐ Not enough time to post
other:_____

When sharing my knowledge to online resources in my research field,
...I believe that I (would) get an answer when I give an answer.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I expect to get back knowledge when I need it.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I expect somebody to respond when I'm in need.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

When sharing my knowledge to online resources in my research field,
...I (would) enhance my personal reputation in the context of bioinformatics.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I believe that knowledge sharing improves my status in the profession.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) improve my status.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

When engaging in knowledge sharing activities (i.e., knowledge-contribution or -seeking) in my research field,
...I (would) strengthen ties between me and others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) expand the scope of my association with other users in the context of bioinformatics.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) create strong relationship with others who have common interests in the context of bioinformatics.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

When engaging in knowledge sharing activities (i.e., knowledge-contribution or -seeking) in my research field,
...I believe that writing and commenting can help others with similar problems.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) enjoy helping others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

When engaging in knowledge sharing activities (i.e., knowledge-contribution or -seeking) in my research field,
...if I provide everybody with my knowledge I am afraid of being replaceable.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I don’t gain anything if I (would) share my knowledge.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...if I share my knowledge I (would) lose my knowledge advantage.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...knowledge-sharing means losing power.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

In online knowledge-sharing activities (i.e., knowledge-contribution or -seeking),
...I believe information/knowledge providers (would be) are trustworthy people.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I believe information/knowledge providers (would be) are experienced.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I believe information/knowledge providers (would be) are well-trained.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I most likely use knowledge created by people I consider experts on the topic.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I most likely use knowledge artifacts created by people I consider to be knowledgeable on the topic.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

Through online knowledge-sharing activities in my research field,
...I (would) share copies from articles published in books, periodicals, magazines, websites, or documents with others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) provide official manuals, methodologies and models for others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) share my experience or unique opinions from the accumulative experience with others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) provide my know-how at the requests of others.

O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) try to share my expertise from my education or training with others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

Through online knowledge-sharing activities in my research field,
...I (would) utilize articles published in books, periodicals, magazines, websites, or documents which are shared by others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) utilize official manuals, methodologies and models which are shared by others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) utilize experience or unique opinions from the accumulative experience from others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

...I (would) request know-how from others.
O Strongly Disagree O Moderately Disagree O Slightly Disagree O Neutral
O Slightly Agree O Moderately Agree O Strongly Agree

Etc
What do you think may be researchers' biggest concerns or fears about knowledge sharing activities (i.e. Knowledge- contribution or -seeking) in the context of bioinformatics?

_____

What specific knowledge (gleaned via online knowledge- sharing activities) would be helpful for your research?

_____

Gender: O Male O Female

Age: _____

Nationality: _____

E-mail Address: _____

Thank you for participation!

Appendix C - IRB approval (Study 2)

VirginiaTech

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: January 14, 2015

TO: Joseph L Gabbard Jr, Jongsoon Park

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)

PROTOCOL TITLE: An extended approach of an insight method: Evidence from visual attention and self- reporting

IRB NUMBER: **14-055**

Effective January 14, 2015, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the Continuing Review request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 6,7**
Protocol Approval Date:	**January 22, 2015**
Protocol Expiration Date:	**January 21, 2016**
Continuing Review Due Date*:	**January 7, 2016**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Appendix D - Informed consent form (Study 2)

Laboratory Study Informed Consent Form

Grado Department Industrial & Systems Engineering, Virginia Tech

Participant number:_____

TITLE of PROJECT: An extended approach of an insight method: Evidence from visual attention and self- reporting

INVESTIGATOR(s): Jongsoon Park and Dr. Joseph L.Gabbard

1. PURPOSE

You are invited to participate in a study on insight-generation processes in data-intensive research practices. The aim is of this study is to identify underlying processes of insight generation by understanding how, when and why a range of different resources is called upon to generate insights. We developed several tasks to achieve our research goals. It is important for you to understand that we are not evaluating you or your abilities. There is no right or wrong answer. We are simply interested in task processes about: (1) what kinds of knowledge resources in an online environment you assemble to solve tasks, and (2) what different types of resources influence information processing. This experiment will last no more than 120 minutes to complete.

2. PROCEDURE

The study will begin with you reading and signing the informed consent document. After you are implying consent to participate in this research, we will ask you conduct four main sessions: a pre-questionnaire with a group embedded figure test, a task performance with think-aloud, a gaze-cued retrospective think-aloud, and a post questionnaire.

1) Pre-questionnaires with training

   We will ask you a pre-questionnaire to get your demographic data (i.e., age, gender, research background, and experience) and a GEFT (Group Embedded Figure Test) to identify cognitive styles of you. After administering the pre-questionnaire, we will conduct an eye tracker calibration to determine whether your eyes can be accurately calibrated (if not, I will terminate the experiment). If the eye calibration succeeded, we will provide a short description about tasks. Before working on main tasks, we will overview functions in the bioinformatics resource. You will then perform training tasks for approximately 10 minutes to get acquainted with a specific bioinformatics resource. During the training task, we will require you to verbalize their thoughts constantly. When you finish the training session, we will give you task instructions for the main tasks.

2) Task performance with concurrent think aloud

Next, we will request you to conduct tasks by the combined use of the PATRIC site and other available resources to the extent to a point after which they will not gain anything more from available resources. At the same time, we will encourage you to think aloud so that your verbalization and inferences will be recorded while you conduct tasks. We will record a computer screen, eye movements, and your verbalizations using an eye tracker and video recorder during entire sessions. After each task, you will have a short break to read the directions for the next task.

3) Gaze-cued retrospective think aloud

Following the insight generation with think-aloud session, we will briefly explain the basic instruction to you for processes of gaze-cued retrospective think-aloud. During this session, we will play the video of screen captures that include overlaid graphics depicting eye movements and fixations. We will ask you to verbalize what you were doing and what you thought on tasks while watching a replay of a gaze recording. Frequently, we ask you for proving questions in order to clarify a verbal report.

4) Post-questionnaire

After finishing a gaze-cued retrospective think-aloud session, we will ask for a post-questionnaire about their perceptions of task complexity and experiences. If you are interested in the results of this study, please send an e-mail to Miss. Jongsoon Park on or after March, 30th, 2014.

3. RISKS OF PARTICIPATION

There are no known risks associated with Tobii X-60 Eye Tracker.

4. BENEFITS

You will receive a copy of the group embedded figure test results by contacting the co-principal investigator. We will provide compensation for your participation, and you will also benefit from knowing that you have participated in worthwhile research that has immediate and positive applications. To get the results of the test, please send an email to Jongsoon Park.

5. COMPENSATION

If you successfully complete given tasks, we will compensate you at the rate $10 per hour for a maximum of 2 hours (a maximum of $20 total) at the end of this study. Should you choose to withdraw from the study before completing all experimental sessions, we will compensate you for the portion of time of the study for which you participated.

6. ANONYMITY AND CONFIDENTIALITY

We will store the data on a computer with password-protected access, and we will keep hard copies of data in a locked filing cabinet in the lab or in the PI's office. Data will be identified by your participation number.

7. APPROVAL OF RESEARCH

This study has been approved, as required, by the Institutional Review Board for research involving human subjects at Virginia Tech and by the department of Industrial & Systems Engineering.

8. FREEDOM TO WITHDRAW

You are free to withdraw at any time from the study at any time for any reason. There is no penalty if you choose to withdraw from this study. If you choose to withdraw, we will compensate you for the hours you have participated in the study up to the point of your withdrawal.

9. PARTICIPANT'S RESPONSIBILITIES

Upon signing this form below, I voluntarily agree to participate in this study. I have the responsibilities to 1) work under the conditions specified by the experimenter to the best of my ability and 2) answer questions, survey, etc. honestly and to the best of my ability. I have no restrictions to my participation in this study.

10. PARTICIPANT'S PERMISSION

I consent to have my photos and video images publicly displayed in conference presentations and journal papers (except for my face). (Yes / No)

I have read and understood the Informed Consent and conditions of this study. Upon signing this form below, I voluntarily agree to participate in this study. I have the responsibilities to 1) work under the conditions specified by the experimenter to the best of my ability and 2) answer questions, survey, etc. honestly and to the best of my ability. I have no restrictions to my participation in this study.

_____ _____

Participant's Signature Date

Should I have any questions or concerns about this research or its conduct, I may contact:

Dr. Joseph L.Gabbard Email: jgabbard@vt.edu Phone: (540) 449-1222

Dr. David M. Moore, Chair, IRB Email: moored@vt.edu Phone: (540) 231-4991

Appendix E - Pre-questionnaire (Study 2)

Pre Questionnaire (Laboratory Study)

Participant number: _____

Gender: _____ **Age:** _____

I am a (check all that apply):

□ Biologist □ Bioinformatician □ Chemist □ Computer Scientist

□ Mathematician □ Other: _____

How many years have you worked in this role? _____ year(s), _____ month(s).

Which online bioinformatics resources (e.g. tools and/or websites) do you use the most?

_____

During last six months, how often did you use bioinformatics tools & web sites?

○ Almost every day ○ More than once a week

○ More than once a month ○ Around once a month

Which online knowledge-sharing platforms (e.g. Seqanswer, Pubmed) do you use the most?

_____

During last six months, how often did you use online knowledge-sharing platforms?

○ Almost every day ○ More than once a week

○ More than once a month ○ Around once a month

What are the major shortcomings or limitations in online bioinformatics resources you currently use?

_____

Which factors are more important when you use online bioinformatics resources (check all that apply)?

□ Speed and responsiveness of resource □ Breath of resource tools and functions

□ Wealth of available data □ Degree of data integration

□ Advanced visualizations □ Ability to upload my own data

□ Ability to share knowledge or information with other researchers

□ Ability to ask questions related to my research

□ Ability to collect knowledge or information from other researchers

□ Ability to create publication quality images □ Ease of use

Other: _____

Appendix F - Insight evaluation manual (Study 2)

Insight Evaluation Manual

1. Objectives

Gaining in-depth understanding of roles of different knowledge resources when scientists implement insight generation progresses

2. User Task

Please characterize functions of the ADD gene in Escherichia coli str. K-12 substr. MG1655 as much detail as possible. Please look for many different types of relevant data as possible using any available resources.

3. Process

Your evaluation process consists of following steps:

1) Data evaluation

Two domain experts review the video (from concurrent think aloud with eye-tracking experiments) and code the data independently on the attached excel sheet (filename: Coding sheet-final.xlsx).

i. Domain experts carefully review the video recording (Each YouTube URL is provided on the excel sheet) and the transcribing notes (highlighted yellow cells on the excel sheet) to obtain the overall flavor of the participants' response.

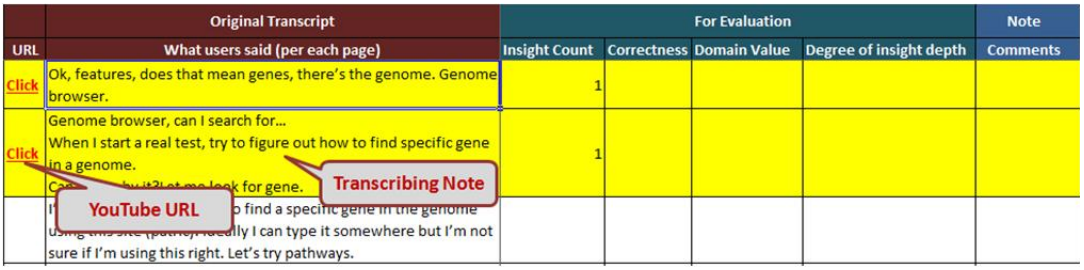
→ Please note that:

- No need to watch content which is not related to the specific insight they are scoring at that time.
- If it is necessary, insights can be divided or added by domain experts. If so, please make comments.
- I recommended the Theater mode instead of the Full mode due to the resolution issue (video size limit)
- You can select the Theater mode icon in the bottom corner of the video.

i. During this process, each domain expert quantifies insights (i.e., Domain value, Breath and depth of insights, and Correctness) based on a pre-defined code (See 4. The definition and decision rule for each code).

→ Please note that:

- Please listen to participants and observe their behavior carefully (e.g., what information does the participant use to describe the functions of the ADD gene? does the participant use the right information?). As of now, the transcribing note is just for supplement and not all is correct.



[Example of the excel sheet]

ii. Please make your comments for each insight (e.g., reason why this insight is incorrect/partially correct).

1) Reconciliation meeting

The researcher (Jongsoon Park) and two experts will have a reconciliation meeting to compare quantification value, and reconcile any differences (if necessary).

2) Reliability

The researcher (Jongsoon Park) will check the reliability of the coding. If the level of reliability is not acceptable, the researcher and domain experts will repeat the previous steps.

1. The definition and decision rule for each code

(1) Number of insights (insight count)
• Definition: The actual finding from knowledge resources • Decision Rule   • Verbal reports: the user reports insights as conducting tasks using different resources. • For example, 1, 2, 3.... • **Researcher has done this part. Please refer to yellow cells on the excel sheet.** • **If necessary, insights can be divided or added by domain experts. If so, please make comments.**
(2) Correctness
• Definition: Some insights are incorrect that result from misinterpreting knowledge resources • Decision Rule with examples **- C: correct** **- P: partially correct** **- I: incorrect**
(3) Domain value
• Definition: The value, importance, or significance of the insight • Decision Rule with examples **- 1 or 2 points:** Simple observations such as "Gene A is high in experiment B" are fairly trivial, and trivial observations **- 3 point:** Insights about a particular process earn an intermediate value of 3. **- 4 or 5 points:** More global observations of a biological pattern such as "deletion of the viral NS1 gene causes a major change in genes relating to cytokine expression" are more valuable. Insights that confirm, deny, or create a hypothesis.
(4) Degree of insight depth
• Definition: Breadth insights present an overview of biological processes, but not much detail. Depth insights are more focused and detailed. • Decision Rule with examples The degree of insight depth is coded using a five-point scale ranging from 1 (Breadth insights) to 5 (Depth insight). **- 1 point:** Breadth insight - "there is a general trend of increasing variation in the gene expression patterns." **- 5 point:** Depth insight - "gene A mirrors the up-down pattern of gene B, but is shifted in time."

Appendix G - IRB approval (Study 3)

VirginiaTech

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: January 12, 2015

TO: Joseph L Gabbard Jr, Jongsoon Park

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)

PROTOCOL TITLE: Understanding scientists’ distributed cognitive activities in actual contexts and suggesting UX/UI implications to inform a future bioinformatics system design

IRB NUMBER: **14-058**

Effective January 9, 2015, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the Continuing Review request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 6,7**
Protocol Approval Date:	**January 23, 2015**
Protocol Expiration Date:	**January 22, 2016**
Continuing Review Due Date*:	**January 8, 2016**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Appendix H - Informed consent form (Study 3)

Longitudinal Study Informed Consent Form

Grado Department Industrial & Systems Engineering, Virginia Tech

Participant number:_____

TITLE of PROJECT: Understanding scientists' distributed cognitive activities in actual contexts and suggesting UX/UI implications to inform a future bioinformatics system design

INVESTIGATOR(s): Jongsoon Park and Dr. Joseph L.Gabbard

1. PURPOSE

You are invited to participate in a study on a qualitative study to investigate how different types of knowledge and insights from external resources is being adopted and incorporated to conduct data-intensive researches in a natural context and identify UI/UX design opportunities. We are simply interested in research processes in actual practices about: (1) how you adopt and incorporate different types of knowledge from external resources to achieve research goals, (2) why you engage in knowledge-sharing and/or reuse activities, and (3) how to support identified processes in computer-mediated environments. This experiment will be divided into two parts to complete.

2. PROCEDURE

The study will begin with you reading and signing the informed consent document. After you provide you consent to the study, we will ask you to conduct four main sessions: a pre-questionnaire, a diary study with interview, a focus group interview, and a post-questionnaire.

1) Pre-Questionnaire

   At first, we will ask you a pre-questionnaire to get your demographic data (i.e., age, gender, research background, and experience).

2) Diary study with interview

   We will then conduct a training session with you to ensure fully understanding of the self-reporting protocol. Next, we will give you guidance on how to use the diary application and required to record insights obtained from external resources and insight drivers using the electronic diary.

   The data entry is required when events of interest (i.e., they undertook, the insights gained from distributed cognitive activities that led to the insights, and the successes and frustrations they experienced knowledge resources) take place during two weeks (10 days). All diary entries will be recorded using the camera for data analysis.

   We will conduct regular interviews with you through an online communication tool, once every week during two weeks, to discuss insights, your experience, and concepts of desired UI/UX design obtained from insights of your ongoing experiences. An individual

interview will last no more than 15 minutes to complete and all interviews will be recorded to support data analysis and our conversation will be recorded using the audio recorder to support data analysis.

3) Focus group interview

   Following the self-reporting with interview sessions, we will schedule a focus group meeting at time that is convenient for all participants. The procedures for the focus group interview are as follows. The session will begin with a short introduction of the focus group interview. You will then review a set of themes of bioinformatics usage information based on diary entries. You will discuss UX/UI implications of bioinformatics resources to support your research context. You will then participate in brain storming to generate, organized, and prioritize ideas for supporting cross-disciplinary data-intensive research. The video camera will be used mainly to record the entire session.

4) Post-Questionnaire

   After finishing the focus group session, we will ask for a post-questionnaire about their perceptions of task complexity and experiences.

3. RISKS OF PARTICIPATION

There are no known risks associated with this project.

4. BENEFITS

We will compensate you for your participation, and you will also benefit from knowing that you have participated in worthwhile research that has positive effect with respect to bioinformatics experiments.

5. COMPENSATION

We will offer you up to $100 at the end of whole process of this study (Diary and focus group session) if you report three diary entries every day for two weeks (10 days).

Diary Session: We will compensate you for the portion of diary entries.

Diary entries/day	Compensation/day
3	$8
2	$5
1	$2

For example, if you successfully complete 3 diary entries, we will compensate you at a rate of $8 per day. If you successfully complete 2 diary entries, we will compensate you at a rate of $5 per day. If you successfully complete 1 diary entries, we will compensate you at a rate of $2 per day. Should you choose to withdraw from the study before completing all diary sessions, we will compensate you for the portion of day of the diary session for which you participated.

Focus Group Session: If you successfully complete given tasks, we will compensate you at a rate of $10 per hour for a maximum of 2 hours, or a maximum of $20 total.

6. ANONYMITY AND CONFIDENTIALITY

We will store the data on a computer with password-protected access, and we will keep hard copies of data in a locked filing cabinet in the lab or in the PI's office. Data will be identified by your participation number.

7. APPROVAL OF RESEARCH

This study has been approved, as required, by the Institutional Review Board for research involving human subjects at Virginia Tech and by the department of Industrial & Systems Engineering.

8. FREEDOM TO WITHDRAW

You are free to withdraw at any time from the study at any time for any reason. There is no penalty if you choose to withdraw from this study. If you choose to withdraw, we will compensate you for the hours you have participated in the study up to the point of your withdrawal.

9. PARTICIPANT'S RESPONSIBILITIES

Upon signing this form below, I voluntarily agree to participate in this study. I have the responsibilities to 1) work under the conditions specified by the experimenter to the best of my ability and 2) answer questions, survey, etc. honestly and to the best of my ability. I have no restrictions to my participation in this study.

10. PARTICIPANT'S PERMISSION

I consent to have my photos and video images publicly displayed in conference presentations and journal papers (except for my face). (Yes / No)

I have read and understood the Informed Consent and conditions of this study. Upon signing this form below, I voluntarily agree to participate in this study. I have the responsibilities to 1) work under the conditions specified by the experimenter to the best of my ability and 2) answer questions, survey, etc. honestly and to the best of my ability. I have no restrictions to my participation in this study.

_____ _____

Participant's Signature Date

Should I have any questions or concerns about this research or its conduct, I may contact:

Dr. Joseph L.Gabbard	Email: jgabbard@vt.edu	Phone: (540) 449-1222
Dr. David M. Moore, Chair, IRB	Email: moored@vt.edu	Phone: (540) 231-4991

Appendix I - Diary instruction (Study 3)

Diary Study Instruction

The main objective of this diary study is to understand investigate how different types of knowledge and insights from external resources is being adopted and incorporated to conduct data-intensive researches in a natural context and the role of knowledge-sharing and reuse activities. Especially, we are interested in computer-mediated interactions at the individual level.

We will conduct regular interview sessions (i.e., once a week during two weeks) through an online communication tool.

As mentioned in the informed consent form, we will then schedule a focus group interview at time that is convenient for all participants. In order for us to conduct more effective focus group interview in the future, please share your experience and opinions regarding data-intensive scientific research you have conducted.

As I said before, we will compensate you for the portion of diary entries.

Several domain experts can examine the quality of diary entries.

- Format

  We will provide a diary form via https://virginiatech.qualtrics.com, please check your e-mail.
- Items

  We will provide a set of pre-established events and you are required recording that is very close in time to the event, reducing the likelihood of forgetting.

Appendix J - Diary sample (Study 3)

Please indicate your experience and opinions about each of the following statements.

Question		
Name		free text
Date/Time		free text
1. Activity		Checkbox & optional free text
2. Purpose		Make brief notes of why you did
Please describe your experience per each resource	3. What external resource has been accessed	Please write the name and the URL
	3-1. Based on the resources above, which of the benefits influenced your decision to use the resources (check all that apply)?	□ Speed and responsiveness of resource □ Breath of resource tools and functions □ Wealth of available data □ Degree of data integration □ Advanced visualizations □ Ability to upload my own data □ Ability to share knowledge with others □ Ability to ask questions related to my research □ Ability to collect knowledge or information from other researchers □ Ability to create publication quality images □ Ease of use Other:_____
	4. Which part of the resource is it helping you with? Please be as specific as possible.	free text
	5. Which part of the resource is it challenging you with? Please be as specific as possible.	free text
	5-1. Alternative action (if applicable)	free text
6. Do you have any other recommendations based on your experience?		free text