Entropy and Insight: Exploring How Information Theory Can Be Used to Quantify Sensemaking in Visual Analytics

Sidney P. Holman

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science

Christopher North, Chair
Scott McCrickard
Steve Harrison

May 4, 2018
Blacksburg, Virginia

Keywords: Information Theory, Insight, HCI, Sensemaking, Visual Analytics

Entropy and Insight: Exploring How Information Theory Can Be Used to Quantify Sensemaking in Visual Analytics

Sidney P. Holman

(ABSTRACT)

With the dramatic increase and continued growth of digital information, developing Visual Analytic systems that support human cognition and insight generation are more necessary than ever before, but there is currently no content-agnostic method for measuring or comparing how well a system facilitates analysis. Researchers in industry and academia are developing advanced tools that offer automated data analysis combined with support for human sensemaking; tools for a wide variety of sense-making tasks are freely available. Now, the pressing question is: which tool works best, and for what? We show that using Shannon's entropy and self-information measures will provide a measure of the complexity reduction that results from an analyst's actions while sorting the information. Further, we demonstrate that reduced complexity can be linked to the knowledge gained. This is important, because a metric for objectively evaluating the success of current systems in generating insights would establish a standard that future tools could build on. This work could help guide researchers and developers in making the next generation of analytic tools, and in the age of big data the effect of such tools could potentially impact everyone.

Entropy and Insight: Exploring How Information Theory Can Be Used to Quantify Sensemaking in Visual Analytics

Sidney P. Holman

(GENERAL AUDIENCE ABSTRACT)

With the dramatic increase and continued growth of digital information, developing systems that enables humans to make sense of all the data are more necessary than ever before, but there is currently no one-size-fits-all method for measuring or comparing how well a system helps people gain such insight. Rather than trying to pin down a definition of what insight is, we instead look at complexity reduction—with the intuition that, before we can make sense of complex data, we must somehow simplify it in a meaningful way. We show that using Shannon's entropy and self-information values will provide a measure of the complexity reduction that results from an analyst's actions while sorting information, and further demonstrate that reduced complexity can be linked to the knowledge gained. This work is important, because a metric for objectively evaluating the success of current systems in generating insights would establish a standard that future tools could build on. This work could help guide researchers and developers in making the next generation of analytic tools, and in the age of big data the effect of such tools could potentially impact everyone.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A nagging issue in Visual Analytics (VA) has been the lack of a standard or baseline for comparing systems to each other. This has been widely recognized, and attempts have been made to create a standard; however, they are typically time-consuming to use and plagued by potential validity issues [19]. Compounding the problem is the tendency for evaluations that are geared towards specific visualization tools rather than addressing underlying user behaviors. Understanding the problem requires some digging into the history of VA, including various attempts at understanding the workflow, what the goal of VA actually is, and how people have attempted to bridge the gap between metrics and goals. The following chapters provide background and context for our proposed research.

This work contributes in the form of exploration, and an expanded understanding of how we can apply a semantically meaningless engineering approach (information theory) to the semantically-focused task of sensemaking and clustering, in a way that links to the understanding users have of the data they are investigating. Such a connection has potential applications in both evaluation of new systems, and in supporting the sensemaking process by providing feedback on how well a workspace has been organized into semantically meaningful groups. For evaluation, this may look like a comparison of relative improvements in the metric—perhaps over time, or between tools. For supporting the sensemaking process itself, it may take the form of real-time feedback of workspace complexity—maybe in the form of a heat map, or highlighting of related groups during interaction. In this research, we sought to understand what factors in the underlying documents influence the effectiveness of the metrics, and to establish a link between entropy and knowledge gain. In doing so, we hope to work toward a sensemaking baseline standard, that could help in designing and using future VA tools.

Chapter 2

Related Work

2.1 Evaluating Visual Analytics Tools

In her 2006 paper on evaluating VA tools [19], Jean Scholtz expands on the belief that simple usability tests and metrics are not sufficient for what researchers are trying to quantify. There are plenty of new tools being created and shared, but the evaluation of these tools too often turned to usability heuristics that look at a somewhat superficial list of qualities of the system—for example, "assessing the usability of the data in the user's decision-making process." There needs to be a more targeted set of metrics specifically for the VA field, and a baseline should be established for researchers to build upon.

Creating metrics for VA comes with a host of problems. Low-level tasks are specific and easy to measure, but most analytic workflows do not use the same set of tasks, or use them in a way that makes comparison difficult. High-level tasks are, by definition, more generalizable, but more difficult to measure. The process of sense-making may involve many iterations of, as Scholtz puts it, "onion peeling" or "pearl growing" the data, with only small changes made over time. At the same time, the act of analyzing data is a task that spans many higher-level areas of activity.

Other research has tried to advance some alternative approaches to VA evaluation. Kang and Stasko [1], for example, turn to case studies as a means to understand the utility of VA tools, or to try and glean design principles; Using a dataset with a known solution, they give scores to various tools and approaches. These sorts of evaluations require either a known solution or an oracle, and manual scoring of the results—much more informative than usability heuristics, but quite time consuming.

Another approach is to try and measure the qualitative aspects of using a VA tool. Saraiya et al. [17] use a longitudinal study to look at the representations and interaction mechanisms provided by a VA tool, and include some observations about the process of generating insights

in this context. The protocol was to collect think-aloud data from participants using a single tool, and ask participants at regular intervals how much of the total potential insights they gleaned. While this study generated fantastic insights into the tasks and workflow of biologists, the methods are far too time consuming and still don't allow for comparison between VA systems.

2.2 Use of Virtual Space

An added difficulty in establishing baseline metrics is that the field is constantly evolving; our research has roots in understanding workspaces, so the following studies provide both examples and background. In the early 2000s, people were investigating the changes in task performance between single- and multiple-monitor workstations, some with more than 3 million pixels [10]. Recently, the explosion of cheap, high-resolution displays has called into question our understanding of how people use virtual workspaces. In 2010, Andrews et al. [2] looked at how high-resolution displays affected visual analysis tasks. Their belief was that the huge increase in virtual workspace will influence how people perform sense-making tasks, but they wanted to understand what form these changes might take. By providing a massive 10,000 by 3,000 pixel display (30+ million pixels) to users and asking them to solve a scenario with a set of intelligence reports, they hoped to compare the strategies used to those from a group that was only given a standard 17-inch monitor.

They found that being able to simultaneously place documents in space (physical navigation) while being able to see both the layout and details of the documents resulted in very different strategies. The old visualization mantra of "overview first, zoom and pan, details on demand" was no longer as applicable—groups of documents were positioned in clusters on the workspace to show collections of evidence, and their relationships. This suggests that the grouping of documents in space create a semantic layer that adds meaning to the data, and also serves as a medium for creating complex structures (like clusters). As for the actual mechanisms involved, the authors note that the large display became a sort of external, easily-accessible memory, in a "distributed cognition" sense; users with the smaller, regular display were forced to use paper diagrams and notes for the same purpose.

Building on the previous paper, Endert et al. [8] dug deeper into the meaning underlying the user-created clusters in a large digital workspace. They found that users organized the information into clusters (or other structures) based on a combination of items within the data (for example, entities: key people and places) and their own intuition and external/prior knowledge. At the same time, the follow-up interviews with participants found that the clusters created didn't usually have labels that describe entities or specific features contained in the clusters—only 30% of the clusters were described using terms found within the clusters. This suggests that users are pulling primarily from their external knowledge when organizing the workspace. Supporting this is the discovery that there was virtually no clustering on the basis of a single term or entity. Analysis of the clusters revealed that the best way to

understand *why* documents were placed together was through "transitive terms". As the name suggests, these terms occurred in subsets of the documents within a cluster, and act as bridges between seemingly unrelated information. The transitive terms were 20% more likely to occur within their cluster than the non-transitive terms, and they were "far less" likely to occur in documents outside of their cluster.

2.3 Insight

While the clustering studies are interesting, what is the point? What do we hope to do with this information? Ultimately, we want to help people generate insights, as Chris North explains [15]. This paper takes a step back from the position of Scholtz ([19]), who suggests that metrics and benchmarks are required; instead North argues that the ultimate purpose of visualization is to allow insights, and so the ultimate goal of visualization evaluation must be to determine how well a visualization helps users gain insights.

Though there is no concrete definition of "insight", it is possible to create a list of *characteristics* of insight that may be considered and measured individually. North suggests that insights are complex, deep, qualitative, unexpected, and relevant. While echoing Scholtz's views on using usability testing for evaluations, North concludes that creating an array of benchmarks for various properties is only a temporary fix; instead, Information Visualization (InfoVis) researchers should consider how we can evaluate using open-ended, domain relevant tasks, and some sort of qualitative insight analysis.

The definition of "insight" comes back to haunt us, however. Chang et al. [5] discuss the lack of a clear understanding of what insight is—and introduce a seemingly conflicting definition of insight used in cognitive psychology. The InfoVis field, along with VA, has long been saying that the goal of visualization is insight. But it seems as though we've been accidentally using the term "insight" to describe both the gradual event of discovering some new, deep, complex piece of information, and to describe the unit of information that was discovered—leading to some confusion about what exactly we are referring to.

Cognitive psychologists have also been investigating the phenomena of insight, but in the psychology literature it is generally viewed as a more sudden, "Aha!" sort of event. It is distinct from the insight in VA work, in that it isn't a slow development built on understanding of complex data; instead, it is viewed as a sudden realignment of understanding. Chang calls this "spontaneous insight", and suggests that it is actually intertwined with the VA meaning of insight. If, as North suggests [15], insight is the result of deep knowledge about complex data, and spontaneous insight is the realignment of knowledge, then one is built on the other. There is much further discussion on the role of insight in VA and how to measure it [23][18][6], but no further progress is made in defining or measuring it in this context.

2.4 Metacognition

On the psychology side, however, there have been many attempts to understand the processes that lead to insight. Metcalfe and Weibe [13] introduce the 'feeling of warmth' and 'feeling of knowing' scores, and show that we are largely unable to predict the "spontaneous" insights—but there was some predictive power for more procedural proplem-solving. Durso et al. [7] build on this work, and suggest that while the flash of insight often occurs without conscious warning, there is often a long trail of conceptual changes that take place. While there is considerable variance in how successful we are at predicting our insights, analysis done by Flemming and Lau [9] indicates that measuring a simple high/low confidence condition in repeated trials provides a reliable way of characterizing individuals' metacognitive efficiency (how good they are at thinking about their thinking).

2.5 Information Theoretic Measures

With the inclusion of cognitive psychology in our understanding of insight, some history-savvy people might have predicted the return of Shannon's information theory—much of cognitive psychology stems from the information processing model, an attempt to model the human mind as a transmission system. To those few, the work of McNamara et al. may come as no surprise [12]. To everyone else, this paper is worth reading for its overview of the sense-making process alone. In a nutshell, the authors suggest that if we view insight as the result of schematizing information—or "onion peeling"[19], "pearl growing"[19], "deep knowledge" building [15], or semantic structuring [2]—that we should be able to detect these enriched patches of information using information-theoretic metrics.

This paper argues that complexity reduction, while not sufficient, is at least a necessary prerequisite to insight generation. A system that fails to separate the needle from the haystack is unlikely to facilitate insight. As luck (and math) would have it, information theory provides tools to help quantify complexity reduction. In the conclusion, the authors discuss an ongoing study of an all-source intelligence analysis task, with an open-ended goal and qualitative analysis of knowledge. Additionally, they look at the user-generated groups of documents created during the task, measuring the Normalized Compression Distance of the various clusters. While results were not available before the paper was published, it is worth noting that this is incorporating both quantitative metrics *and* the benchmark-free approach advocated by North in 2006 [15].

2.6 Research Context

The selected papers provide a brief narrative tour through the history of evaluating Visual Analytics tools. While it has been a recognized problem for more than a decade [19],[15], it is an issue on which it is difficult to gain traction. In part, this difficulty stems from a lack of clarity about what "insight" actually was; another problem is disagreement over how to measure success. However, the information-theoretic approach advocated by McNamara et al. [12] suggests that it is both possible and reasonable to create metrics that support the deeper goal of VA and InfoVis.

Our research is building on both the clustering studies ([2],[8]), and on the information theory approach to analysis [12]. Possibly the most intuitive way of understanding what we're measuring is through the "Semantics of Clustering" paper [8]. Recall that they found "transitive terms" were the best way to understand why clusters were created; in their study, these terms were manually identified before being analyzed. The terms were much more likely to occur in documents within the cluster, and unlikely to be found at all in documents outside the cluster.

Our expectation is that the entropy or self-information measures will be able to detect this increase in similarity within clusters without the need to manually identify the important terms. For example, an intelligence analyst trying to ferret out details of plots might have many ways to group the reports at hand, but during the sensemaking process would group them according to semantically related traits. In one group, perhaps they noticed an unusual trend in toaster purchases; in another, the common thread is the purchase of large plots of land in remote areas. These groups would have some common terms, and may appear related by looking at the entropy. However, as the sensemaking process continues, the analyst may be able to break the groups up into more closely related groups—perhaps linking names, and places, and common events. In this new clustering, these high-entropy terms would likely be grouped together, reducing the entropy and self-information of both the terms and the groups they are part of.

In the McNamara paper, this would show up as improved compressibility; Shannon's source coding theorem suggests a relationship between the amount that data can be compressed, and the entropy of the data. Our hope is that, if we track the change in entropy (or self-information) over time along with qualitative measures of insight over time, we may be able to uncover a relationship between the two. If successful, it would be possible to connect insights to sudden improvements in self-information, or provide investigators with cues to follow up on what prompted the sudden improvement in entropy—thus allowing us to better understand what features facilitate insight.

Chapter 3

Hypothesis

3.1 Main Research Question

The overarching question motivating this research is:

> Can we estimate insight generation by measuring the reduction in complexity of a workspace over time, using information theoretic measures of entropy and self-information?

Our hypothesis is that changes in structure of documents in space reflect insight into the content, which can be seen in reduced entropy measured across time. To reduce the scope of this question to something manageable, we will restrict our investigation to text-based visual analytics. By "structure in space", we refer to the spatial organization schemes that users create during the sensemaking proces, as seen in [8]. Also, although the sensemaking process can be seen in any number of forms, here we are going to focus on a task where users are explicityly organizing documents into meaningful groups in a workspace. In layman's terms, we want to see if we can link complexity reduction to insight, in either the spontaneous or gradual forms described in section 2.3. In an ideal case for our testing, there would be a minimum-complexity condition that corresponded with the "golden solution," or known answer to the analysis task—in other words, we could compare how far from the "golden solution" each users' analysis is to their understanding of the dataset.

As mentioned in section 2.1, however, one of the problems with the current means of evaluating user interfaces is the reliance on "known solutions" as a baseline. So, our approach must also demonstrate some connection between complexity and insight in a more intricate, real-world data set; this set would have no known "golden solution," but we would expect that people who ferret out more interrelated documents would create measurably lower complexity.

3.2 Understanding the Factors

This leads to a natural ordering in what we want to test. First, we must answer the question:

> Do the information-theoretic measures of Entropy or Self-Information change as analysts organize documents in the workspace?

So, we want to demonstrate that there is some detectable effect between the organized workspace and the unstructured workspace. But, to fully answer this, we must also understand what influences the changes in the metrics we observe—so, we must investigate the properties of various documents and groupings. Understanding what factors will affect the information-theoretic measures will form the basis for our research. To this end we've identified several possible factors, shown in Table 3.1. The most obvious factor is the size of the documents we are using; in our early explorations, there was consistently a larger improvement in the self-information of documents that contained many words—for example, whole chapters of books—than in documents that were built from just a few sentences or paragraphs.

"Simulated" refers to the authenticity of the documents—are they synthetic, created artificially for research, or are they documents used in real-world problems? An example of a real document set would be the Litvinenko articles used in [12], or the product information documents used in [10].

The number of "unique symbols" includes each word used in the workspace only once, with no distinction for letter case and all punctuation removed. This is much like a dictionary that just contains the words in the workspace. There may be an interaction between this factor and document size, or with the number of documents in the workspace.

The number of documents in the workspace may impact the results, as moving a document from one cluster to another would cause large changes if there were only a few (or no) documents in either cluster.

The number of clusters we analyze will depend on the natural groupings in the "golden solution" of a synthetic document set, or the user-created clusters in a real-world document

Table 3.1: Possible Message Factors

Factor	**Description**
Document Size	The number of symbols in a document, average.
Simulated	Is this a synthetic or real-world document set?
Unique Symbols	The number of unique words in all the documents.
Number of Documents	The total number of documents in the workspace.
Number of Clusters	The number of clusters generated at the end of an analysis session.

set.

Stating that we expect the information-theoretic measures to detect an improvement begs the question, "improvement with respect to what?" Our explorations suggest that the most meaningful comparison is against the same set of documents, but with the *documents shuffled randomly between clusters*. Our reasoning is that the purpose of human analysis is to add some semantic structure—and randomly grouping the documents is the organization strategy with the least semantic structure. To make sure that we don't compare against an ideal clustering of documents, the "randomized" score we use for comparison will be the average cluster value over several rounds of random shuffling.

3.3 Information Theory Versus Insight

The second question we need to investigate is one of correlation, between user actions and our metrics:

> Do the information-theoretic measures of Entropy or Self-information correspond with increasing levels of insight?

We know, from Flemming and Lau, that simple metacognitive measures can be employed to estimate levels of insight [9]. Additionally, it has been shown by Young and others that qualitative measures such as think-aloud protocols can provide information about a person's insights [24, 14]. Using a realistic dataset and the tool developed in part 1, we will conduct a user study to collect real-world groupings of documents. Though all clusters will be semantically meaningful in some way, we expect the complexity measures will improve as users cluster the documents in a way that also explains the obscured plot. At the same time, we track the change in analysts' insight over time with the metacognitive measure and think-aloud feedback. We expect that insight will increase over time; our goal is to understand if that knowledge gain corresponds with instances of lower complexity (measured with entropy or self-information). At a higher level, testing analysts' understanding of a dataset both during and after a sensemaking session should show similar improvement in complexity measures.

The first part of our research, understanding the factors, helps us verify that there is some change in the measures we're tracking, between the sorted and randomized conditions. The second part, comparing our metrics with insight, will help us link the characterization of the information-theoretic factors to insights and knowledge gain in users. Taken together, they will answer our main research question.

Chapter 4

Methodology

In this chapter, we explain our approach and some of the design decisions that influenced our work. We will start by offering definitions of what, exactly, we mean when we talk of entropy and self information. Following these definitions, we provide a summary and pseudocode for how we are calculating improvement in these measures.

4.1 Definitions

Before continuing to discuss the merits of entropy and self information, some definitions are needed in order to explain how we are using information theory to address the specific research questions detailed in chapter 3. The definitions provided here are geared toward our specific domain, in that we are treating words in documents as the "symbols" in the information-theoretic sense; thus, this paper may use the terms 'words' and 'symbols' interchangeably. For a more complete explanation of Information Theory, see [3] or the original work by Shannon [20].

4.1.1 Information Theory

Information Theory is concerned with the technical problem of transmitting messages through a channel. Messages, in this engineering view of the transmission problem, can be deconstructed (encoded) into a series of bits and sent to the receiver who will then reconstruct (decode) the message. An important thing to understand is that the encoding and decoding of the message is done using the same background reference data—both sender and receiver must have the same knowledge about the underlying data or decoding will fail. Information theory quantifies information in bits, and estimates the complexity of a dataset by how many bits are needed to encode the symbols. For example, if a message is always "X", then the

complexity is very low. If instead the message can be any symbol from "A" to "Z", all with equal likelihood, the complexity of the message is much higher. Information Theory (via the Source Coding Theorem) is also the basis for our understanding of how much we can compress a message before we start to lose information.

4.1.2 Symbol

A symbol is a representation of some piece of knowledge. These symbols can be defined as anything: words, individual characters, sets of pixels, or notes in a tune. Symbols have some predetermined likelihood of appearing, in the discrete context being examined. The set of symbols, then, is tied to some specific probability distribution based on the scope we choose to look at.

$$p_x = \frac{\text{\# of occurences of symbol x}}{\text{\# of occurences of all symbols}} \tag{4.1}$$

In our work, the probability of the word occurring is calculated with respect to all the other words in the document, or group of documents. Changes to scope (adding or removing documents to the group) also changes the probabilities, so it is not meaningful to say something like "Document Q adds 15 bits of information to the group" —because we've changed the underlying probability distribution for all the symbols, we've changed the information value of all the symbols. Put plainly, the probability of the word "the" appearing depends on the documents in the set—and if you add a new document to the set, that probability will likely change.

4.1.3 Self-Information

The self-information (I_s) of a symbol 'x' is calculated from how often the symbol occurs in a set. "I_s refers to a symbol in a set."

I_s for 'x' is calculated as:

$$I_s = -log_2(p_x) = log_2(\frac{1}{p_x}) \tag{4.2}$$

and can be understood as the amount of information, in bits, needed to encode a single instance of this symbol in a message. For example, if a message could be one of five possible symbols, and all the symbols appear the same number of times, the self-information of one symbol is:

$$p_x = 1/5$$

$$I_s = log_2(\tfrac{1}{p_x}) = log_2(\tfrac{1}{1/5}) = log_2(5) = 2.32 \text{ bits}$$

4.1.4 Message

A message is a subset of symbols in a given context. A message may not contain all the words in a set, but an unknown message has a chance of containing any of the symbols. Given that we don't know the contents of a possible message in advance, we need to be able to **estimate** the expected length (in bits) of a message given the number of symbols. We can do this with Shannon entropy.

4.1.5 Entropy

Entropy, denoted "H", is a measure of the **average** expected length, in bits, of the symbols that comprise a message; to estimate the size of the message, we will need to multiply by the number of symbols in the message. "Entropy refers to a set of symbols." It can be understood as an indication of the complexity of the vocabulary of a message–"yes" and "no" are simple, but trying to transmit a chapter of "Alice in Wonderland" would require much more complicated vocabulary.

Entropy is calculated as:

$$H = -\sum_{i=1}^{n} p_i * log_2(p_i) = \sum_{i=1}^{n} p_i * log_2(\frac{1}{p_i}) \tag{4.3}$$

where n is the number of different outcomes possible. To estimate a message with N symbols, simply multiply entropy by N: $H_{message} = H * N$. *In our work, n is the number of unique words in the set of documents in a cluster.*

4.1.6 Self-Information of a Message

The self-information of an entire message is, naturally, the sum of the self-information for all the symbols it contains. The self-information of a message is different from the entropy because it is applied to messages where the contents are known. For example, we would calculate the *self-information* for a message that has 12 symbols from a larger set, and we know what these symbols are; this would give the exact number of bits required to code the message. Keep in mind, both entropy and self-information require that we know the set of symbols and their frequency.

Self-information of a message is calculated as:

$$I_m = -\sum_{i=1}^{n} log_2(p_i) = \sum_{i=1}^{n} log_2(\frac{1}{p_i}) \tag{4.4}$$

where n is the number of words in the message.

As a side note: in these equations, there is the expectation that the probability of the events occurring are independent. While the frequency of words in English are not completely independent, we are treating them as such in our work. Future work could be done to understand how this impacts our metric.

4.2 Analysis Pseudocode

Our code for analyzing the synthetic and user-generated clusters is fairly straightforward, but still most easily explained with some pseudocode. The randomized score used for comparison (described in section 3.2) is determined by randomly placing the documents into empty clusters, calculating the entropy and self-info metrics, and taking the average of each metric over a large number of N runs.

Algorithm 1 Compute random average $MetricValue$ of a document set, over N runs

```
procedure RandomAverage(ClusterCount, DocumentSet)
    MetricSum = 0
    for i = 0; i < N; i++ do
        create (ClusterCount) number of TempCluster
        for each document in DocumentSet do
            put document in randomly selected TempCluster
        for each TempCluster do
            MetricSum = MetricSum + MetricValue(TempCluster)
    TotalClusters = N · ClusterCount
    Return MetricSum/TotalClusters
```

This gives us the average entropy or self-information (depending on the $MetricValue$ function we choose) of a cluster, when the document set is spread randomly between groups. Though it loops through many times, this only needs to be done once for every document set—the average should not change significantly if we choose a large enough number for N. With this, we can calculate the net change in a cluster or set of clusters by simply subtracting the random average from the $MetricValue$ of the cluster/set. Thus, improvement in our metric will be shown as a net negative value for a cluster or cluster set:

$$MetricValue(\textit{user-created}) - \textit{randomized} = \textit{improvement} \tag{4.5}$$

Chapter 5

Evaluation

5.1 Entropy and Self-Information Factors

This research starts by building understanding of what factors influence the entropy and self-information metrics, both so that we can better understand what we are measuring and so that confounding conditions may be identified. We are only able to test document size, unique symbols, number of documents, and number of clusters with the simulated document sets, as controlling each of these factors independently in real-world document sets is impractical. So, we will start by looking at the factorial issues on the synthetic set, then run a simpler test on the real-world document set which will be used in the user study, to see if we can fully answer our first question by detecting an effect in the correctly-sorted arrangement of documents.

Procedure: Half of the synthetic document set has been generated from public domain literature, found in text format from the Project Gutenberg digital library, by extracting portions of various classic works: Alice in Wonderland, Dracula, Frankenstein, Moby Dick, and The Portrait of Dorian Gray. Excerpts from these will be used as part of the high-unique-symbols condition; for the high-document-size, this means that entire chapters were treated as individual documents, and for the low-document-size condition we extracted paragraphs randomly throughout each text. We consider the natural grouping of the documents to correspond with the book each document originates from.

The other half of the synthetic document set was created with a "Lorum Ipsum" text generator that could produce varying lengths of documents with a fairly low number of unique words. Manually adding some English-language sentences resulted in documents that could be 1000+ words, but with fewer than 150 unique words in total.

With the synthetic set, we conducted a series of runs with each of the factors (defined in

Table 3.1) set to the high or low levels in accordance with the schedule given in Table 5.1. Each run calculates the information-theoretic metrics for the natural grouping (which we call the "sorted" arrangement), and 60 runs that create the randomized average metrics. For document size, we say the (high, low) states are (<300, >1000) words. Unique symbols were adjusted after some exploration, but we settled on using (<150, >300) unique words. Number of documents was (<30, >50), and number of clusters was tested on (3, 5) groups. The number of clusters were chosen based on the "5 ± 2" rule of thumb; our intuition is that most people would create a few broad categories rather than fine-grained sub-divisions in an analysis task. Additionally, if there is a high upper bound on the number of clusters and we are trying to anticipate how people might group documents, we might be creating a situation where one or more clusters would be empty of documents. While this could be dealt with dynamically in a real-world setting, for this analysis we focus on the more human-manageable cluster counts.

A final variable that can be manipulated during this phase of our investigation is our "symbol set", or what words we want to include when calculating entropy and self-information. Many visual analytics projects identify the "named entities" found in documents—things like people, places, dates, and so on—that form the nodes in the narrative schemes people pull together during their sensemaking process. With the basis of our research being built on some notion of "transitive terms"[8], we want to understand how these entities may relate to our metrics. To automatically identify and extract the entities we used the Named Entity Recognizer, part of the Stanford Natural Language Processing software [11]. In our case, we use the default model that comes with the Stanford NER software, which is geared towards English-language person, organization, and location recognition. As a sort of intermediate level between the full text and the extracted entities, we created a third version of all the documents that has all of the "stop words" removed—for example, "an", "while", "because", and so on. While this may seem like an unnecessary step when already looking at entity extraction, it has the advantage of retaining relevant words like "explosive" that might get removed by the default Stanford NER, while still removing some of the noise that might be present in the full text of documents.

5.1.1 Factor Analysis Results

Analysis: The results are analyzed with a two-way repeated measures ANOVA with document length, number of unique words, number of documents, and number of clusters as factors. Results demonstrated that the sorted document set results in lower information-theoretic metrics, with $p \leq 0.01$, which provides support for our hypothesis that semantic structuring of the documents results in lower entropy and self-information. Note, $\alpha = 0.01$ is reported to account for the fairly small sample size; though we calculated the random entropy and self-information 60 times for each cluster, the results of the sorted condition never vary, so there are only 8 observations in each treatment.

Table 5.1: Factor Schedule

Run	**Size**	**Uniques**	**Documents**	**Clusters**
1	Low	Low	Low	Low
2	High	Low	Low	Low
3	Low	High	Low	Low
4	High	High	Low	Low
5	Low	Low	High	Low
6	High	Low	High	Low
7	Low	High	High	Low
8	High	High	High	Low
9	Low	Low	Low	High
10	High	Low	Low	High
11	Low	High	Low	High
12	High	High	Low	High
13	Low	Low	High	High
14	High	Low	High	High
15	Low	High	High	High
16	High	High	High	High

Table 5.2: ANOVA Summary, Full-Text Entropy

Factor	df	MS	f	p
Size	1	0.1271	2.442	0.12473
Uniques	1	1.0556	20.276	$4.28 \cdot 10^{-5}$
Clusters	1	0.0056	0.108	0.74388
Document Count	1	0.5556	10.673	0.00201
Uniques : Doc Count	1	0.2303	4.424	0.04072

Our analysis is run on the net change of the information-theoretic measures. As mentioned, we shuffle all the documents in the set between all the clusters (3 or 5, depending on the factor schedule) and use the average over 60 runs as the "randomized" average. We consider the difference between the user-created clusters and the "randomized" score: the larger the *randomized − user-created = improvement* value, the more improvement we say there is. This is done for both entropy and self-information, on the documents with full text, with "stop words" removed, and with the documents that have been whittled down to just the extracted entities.

Looking at the entropy results, the ANOVA run on the full, unaltered documents indicates that both the number of unique words in a document and the number of documents in the set have a significant impact on the average change in entropy of the clusters. Unique words are

Table 5.3: ANOVA Summary, Full-Text Self-Information

Factor	df	MS	f	p
Size	1	$9.013 \cdot 10^{10}$	11.723	0.001273
Uniques	1	$1.648 \cdot 10^{10}$	2.144	0.149671
Clusters	1	$8.528 \cdot 10^{9}$	1.109	0.297544
Document Count	1	$1.232 \cdot 10^{11}$	16.024	0.000216
Size : Clusters	1	$8.443 \cdot 10^{10}$	10.982	0.001756

significant ($p = 0.0000428$) at a higher level than the number of documents ($p = 0.00201$), which indicates that these are unlikely to be anomalous results. Running the same analysis, but with all the "stop words" removed from the documents, results in similar p-values for unique words and number of documents ($p = 0.0000528$, and 0.00981, respectively). As might be expected, the analysis when we only use the extracted entities is much less impacted by the number of unique words ($p = 0.00228$), likely because the documents have been reduced to virtually nothing *except* unique terms.

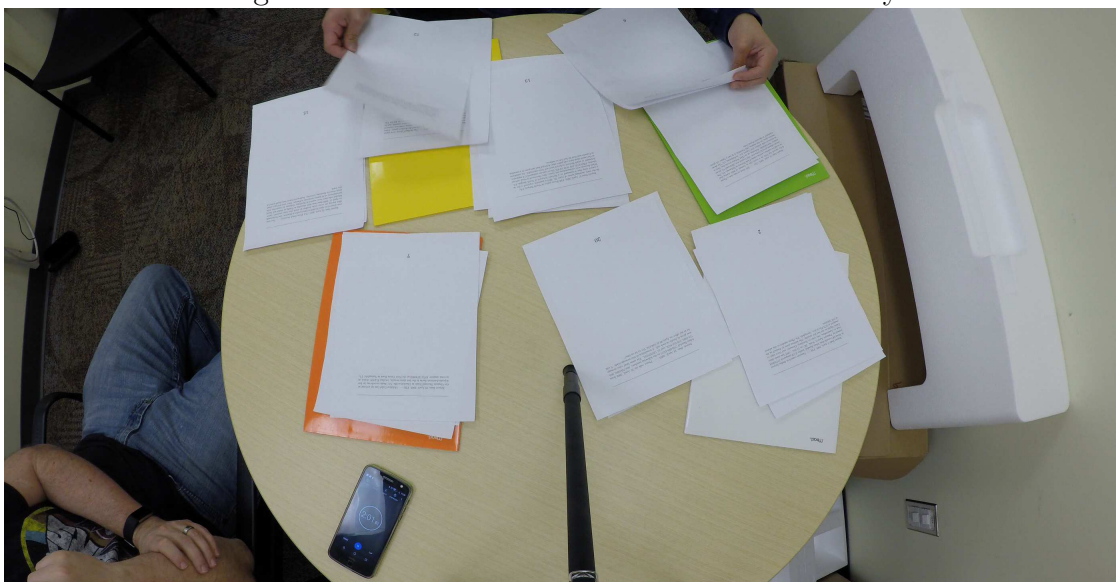
The ANOVA for the self-information results indicates the number of documents, and their size, are the most influential on the metric. For the full text, the number of documents was significant at $p = 0.000216$, and the size of the documents slightly less so at $p = 0.001273$. On the document set with "stop words" removed, there is even higher significance for both the number of documents ($p = 0.00000811$) and their size ($p = 0.000089$). This trend continues with the extracted entities set, with $p = 0.00000329$ for number of documents, and $p = 0.00000142$ for document size. Discussion of these results will be included in chapter 6.

Table 5.4: % vs. Random, Entropy and Self-Information

Symbol Set	Entropy	Self-Info
Full Text	98.8	96.9
No Stop Words	97.8	95.3
Entities Only	94.1	90.6

We close out this part of our investigation by checking the measures on the real-world document set that will be used in our study. The sorted grouping of documents (which best explains the plot) results in improved metrics for all symbol sets. For the full text, the entropy is reduced by 0.33 bits, and self-information by 350 bits; no "stop words" results in improvement of 0.62 bits and 335 bits of entropy and self-information, respectively; and entities only results in 1.17 bits and -106 bits of improvement for the two. The relative improvement is shown in table 5.4 as a percent of the random average value. While the values are not dramatically different, it is worth noting that on some of the synthetic document sets with many documents, the metrics were sometimes dramatically smaller—so the small document set used in the study may not allow the same amount of improvement.

Figure 5.1: Refinement Session of the User Study.



5.2 Evaluating Insight

To investigate insight, we have conducted a user study. Leaning on precedent ([24, 14]), our investigation of insight will use a Think-Aloud approach, and also include a metacognitive measure of participants' confidence in each move of a document. The sessions have been video recorded, for transcribing participants' comments and time-stamping document moves. The document set is a modified version of an intelligence analyst training dataset ("Sign of the Crescent"), with distractors and junk documents removed. It also has a known correct grouping of documents, making later comparisons to confidence and think-aloud comments easier.

Procedure: Each participant was seated at a 4 foot diameter table, with a shuffled stack of 20 modified documents from the training set, and 4 unlabeled folders that served as clusters. Each document is a short intelligence report, listing the date and organization it was gathered from (CIA, FBI, etc), and has a large number at the bottom for tracking group membership from the recording. The documents have been randomly ordered for each participant, and are not readable until the initial session has begun. Participants were asked to sort the documents into groups based on their content, with the goal of explaining the "who, what, when, where, and how" questions surrounding the terrorist plots. While working on this task, participants were asked to both verbalize their thought process (Think Aloud), and provide verbal feedback on their level of confidence (high, or low) on the cluster membership after each move. The cluster memberships were tracked over time by manual review of the recording, and at the end of the session participants were asked to summarize

the clusters and explain why they chose the groupings they did. Each participant was given two sessions to complete the task: an initial session, where they skimmed and decided on initial placement, lasted up to 15 minutes. A refinement session followed, lasting up to 10 minutes.

The relevant events from both sessions were logged with a timestamp after removing any identifying information. We recorded each session as a series of document moves between the groups, with the moves being parsed manually by reviewing the video of the session. When entering the group memberships, documents were logged by recording the following fields in an anonymized Excel spreadsheet (field name, data type):

1. Move Number (move, integer)
2. Document Number (document, integer)
3. Group (group, {o, y, g, w, b})
4. Approximate Time (time, {hh:mm:ss})
5. Confidence (confidence, {h, l})

Both the initial and refinement sessions were scored, using a scheme that requires increasing understanding of the plot details to earn more points. The scoring sheet is included in Appendix B, and when used flexibly, allowed us to award points for plot details mentioned both during the sessions, and during the summaries that followed. As mentioned, this scheme increased the difficulty of each point earned; trivial points were earned for correctly grouping reports on a single plot together, where subsequent points required explanation of general goals within each group—and for the maximum of 20 points, very specific details must have been understood and verbalized. For example, 1 point would be given for grouping all of the reports related to Amsterdam, another point for understanding that there was some plot to bomb something related to shipping, and another point each for recognizing that they were going to use a dirty bomb, and were targeting Boston.

5.2.1 User Study Details

Our user study was conducted in Spring of 2018, with a total of 12 participants. Ten were pulled from undergraduate computer science courses, and the other two were graduate students who were passingly familiar with the dataset used in the study. On average, each session lasted approximately 40 minutes, with 25 minutes spent actually manipulating the groups and the rest spent being introduced to the task or explaining the details of the uncovered plot.

Each session was filmed in 4K video, from the introduction of the documents to the participants, until the end of the second round of refinements in grouping—where participants

explained, in as much detail as they could, their understanding of the terrorist plot described in the dataset. The setting, shown in Figure 5.1, consisted of a small desk with 4 differently-colored folders on top; the colored folders were chosen both to provide a visual cue to participants while they organized the documents, and to be visibly distinct while reviewing the recorded session at a later time.

As described above, participants were given the stack of documents, and asked to group them in a way that allowed them to answer the "who, what, when, where, and how" of the hidden plot. If pressed, we would say that there are many potential ways to group the documents, but that they (the participant) were allowed to group them in any way that helped them understand and explain the details. In our prompt, we encouraged them to make use of all four folders in developing their groups, and asked that with each move of a document between groups they give some indication of their confidence that the current document belonged with the existing documents in the group. To reduce the subjectivity of "levels of confidence" to something that could be roughly analyzed, participants were directed to use "high" or "low" to indicate their confidence; in cases where users habitually slipped into a percentage description, we considered variations such as "90% confident", or "pretty confident" to also indicate high confidence. The full study prompt can be found in Appendix A.

The sensemaking and grouping process was broken into two time sessions: initially, users were given 15 minutes to read and quickly cluster the documents based on incomplete understanding of the entire dataset. A second, 10 minute period was given to allow them to refine their groupings based on their new understanding of the document contents. In addition to the high/low confidence ratings, these two time periods provide the opportunity to compare our information-theoretic measures to the participants' understanding over time—rather than relying on the high/low scale, we can contrast their first-session answers (score) about the plot to what they understood at the end of the second session.

5.2.2 User Study Results

We calculate entropy and self-information of each user-created cluster with equations 4.3 and 4.4, respectively. Our expectation is that the information-theoretic measures will decrease (show improvement) over time with respect to the average of the randomly shuffled documents, as described in chapter 3. We also expected that the think-aloud feedback and confidence scores will indicate increased understanding of the document set as the users spend more time with understanding the plot. However, initial analysis shows that the metacognitive measure (high/low confidence scores) does not correlate with the points the participants earn by stating their understanding out loud.

Additionally, reviewing the sessions reveals that the metacognitive confidence scores also fail to correspond with meaningful groupings of documents.

Figure 5.2: Users' Score after 15 and 25 Minutes.

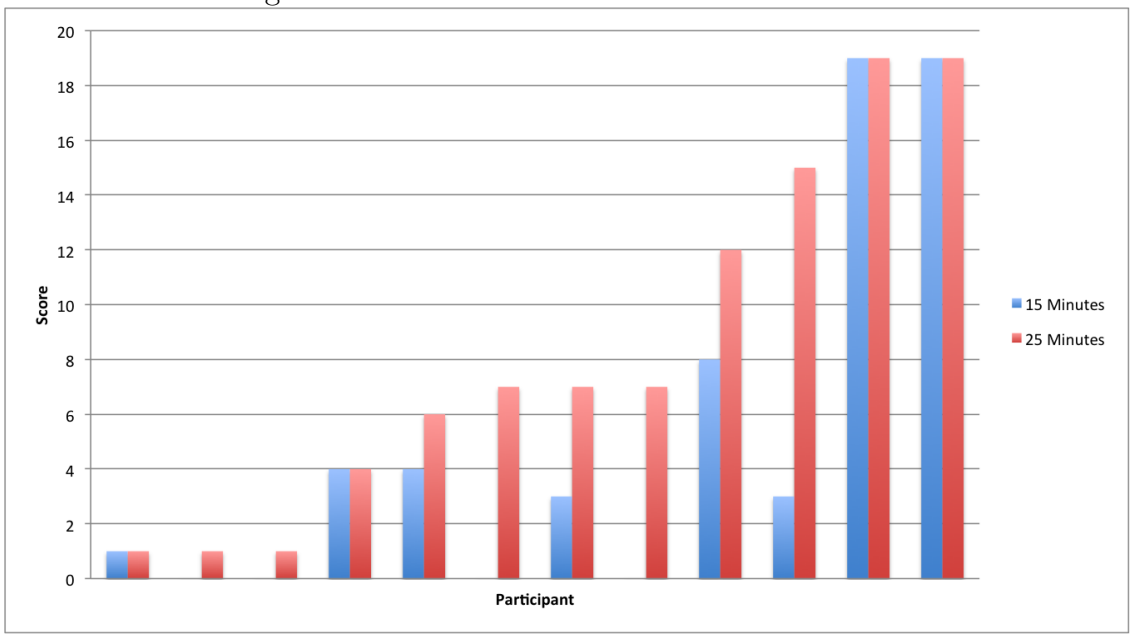
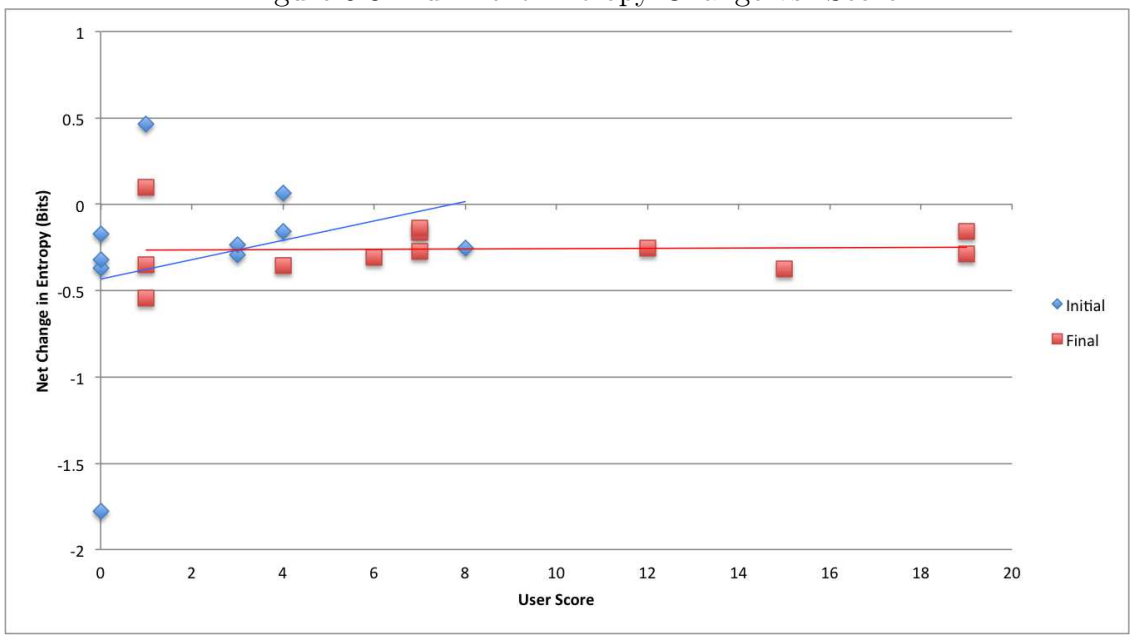


Figure 5.3: Full-Text Entropy Change vs. Score



The goal of the user study is to determine if the complexity reductions that we can detect in the workspace correspond with increases in users' insight, but with the failure of our fine-grained measure of understanding, we will instead rely on the coarse time periods of the initial and refinement sessions for analysis. These coarse time periods are evaluated with

Figure 5.4: No Stop-Words Entropy Change vs. Score

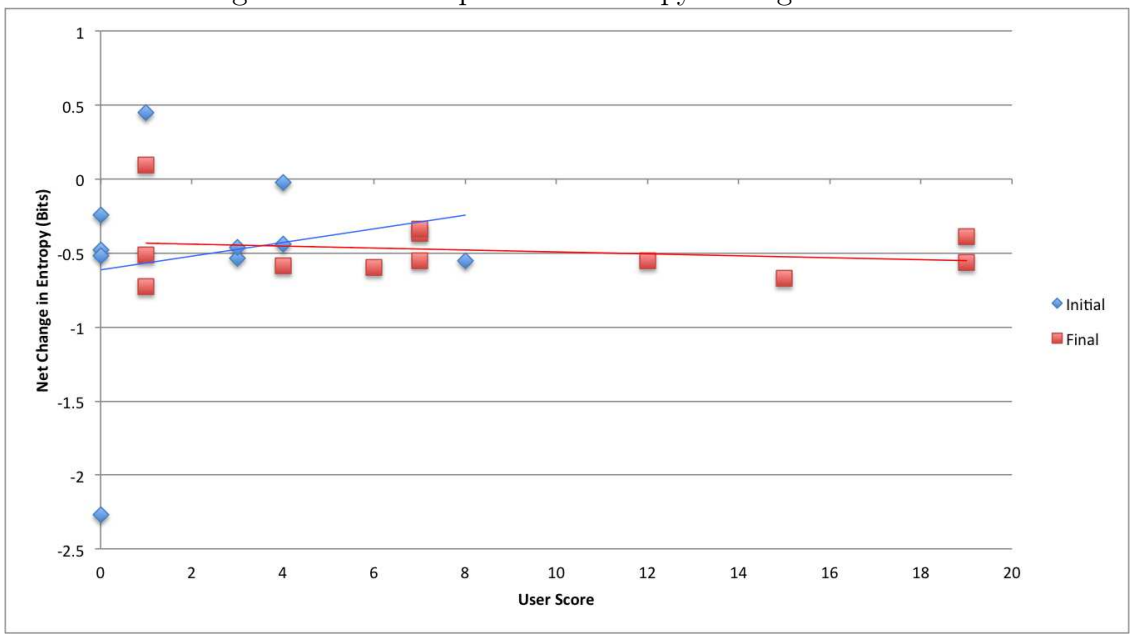
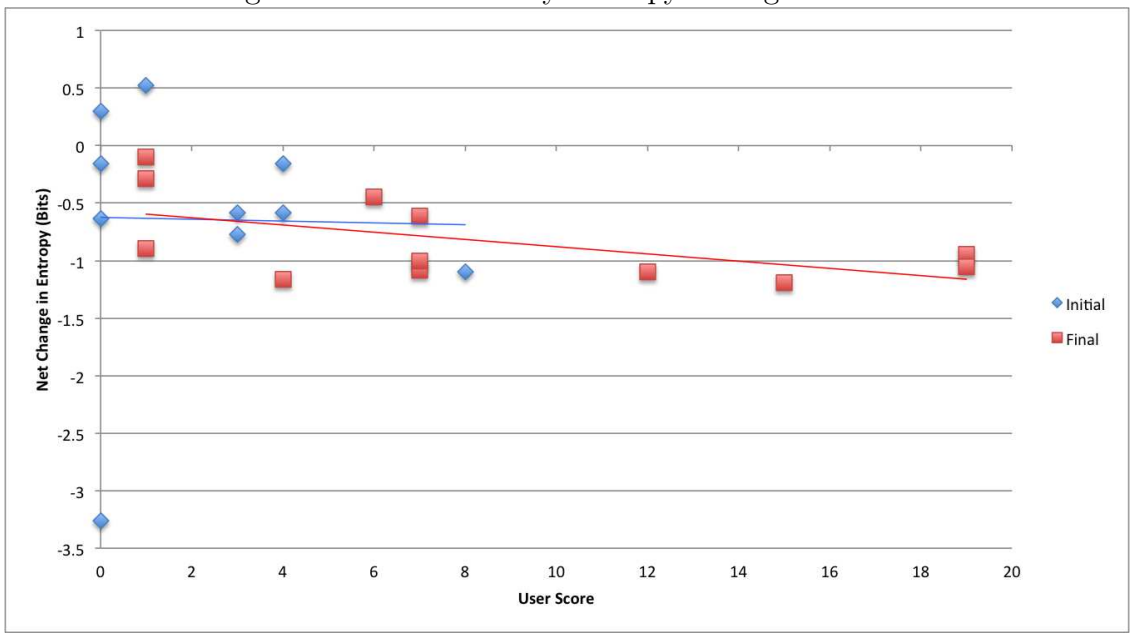


Figure 5.5: Entities-Only Entropy Change vs. Score



a user score, which we will compare to our information theoretic metrics rather than the high:low confidence ratio originally planned; this change will be discussed in more detail in chapter 6. These results *do* show improvement over time: after the initial session, the average score is 2.3 points (out of 20 possible). After the refinement session the average score is 8.25

Figure 5.6: Full-Text Self-Info Change vs. Score

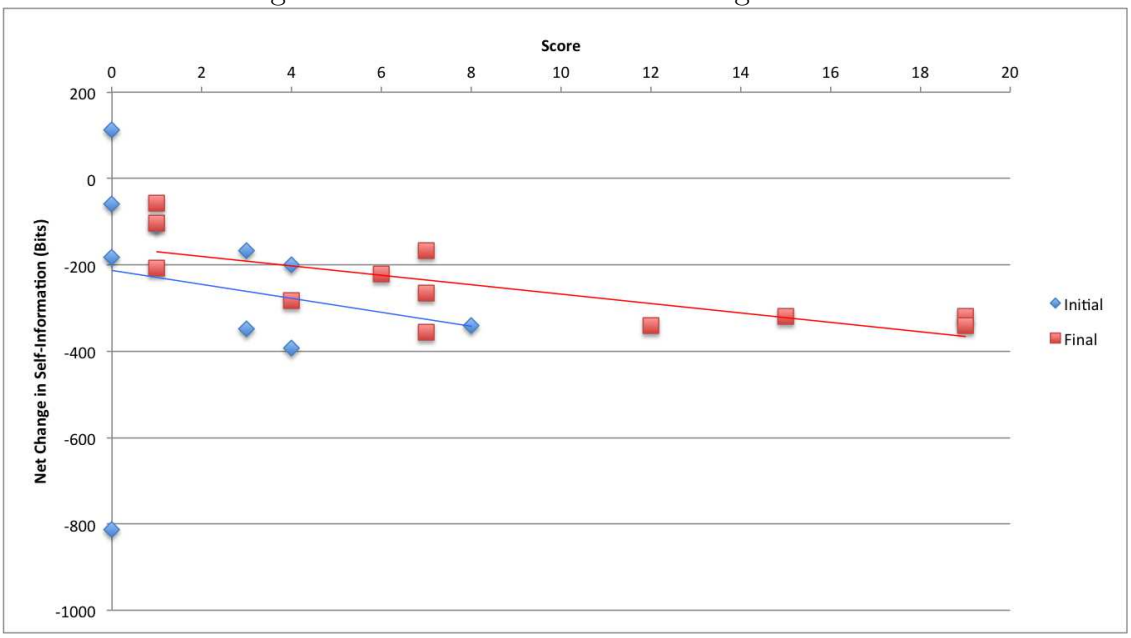
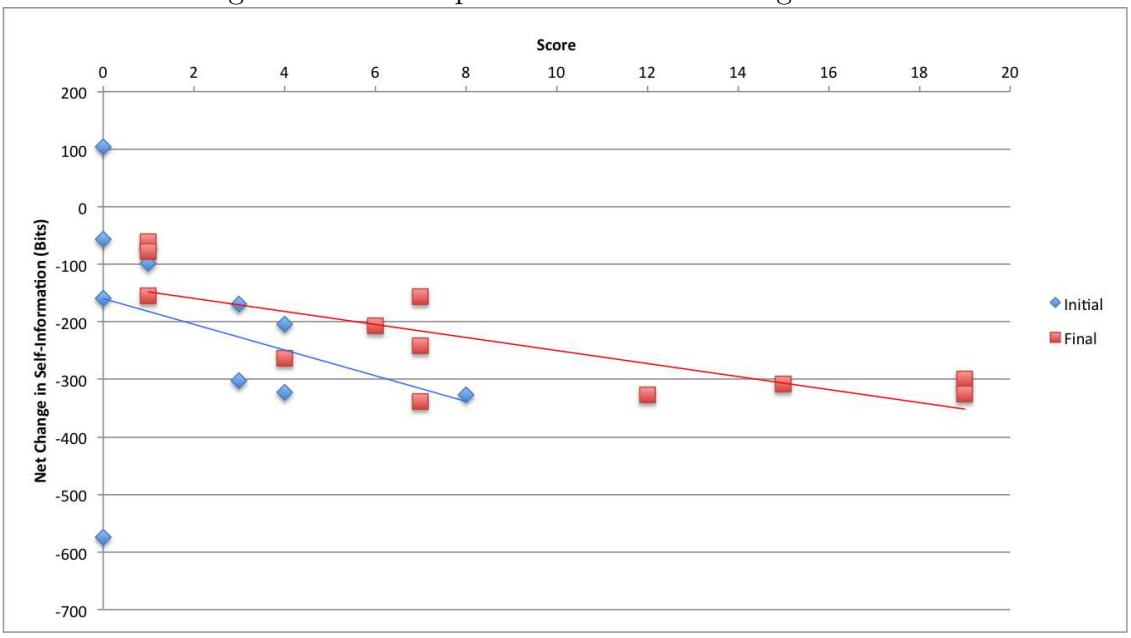


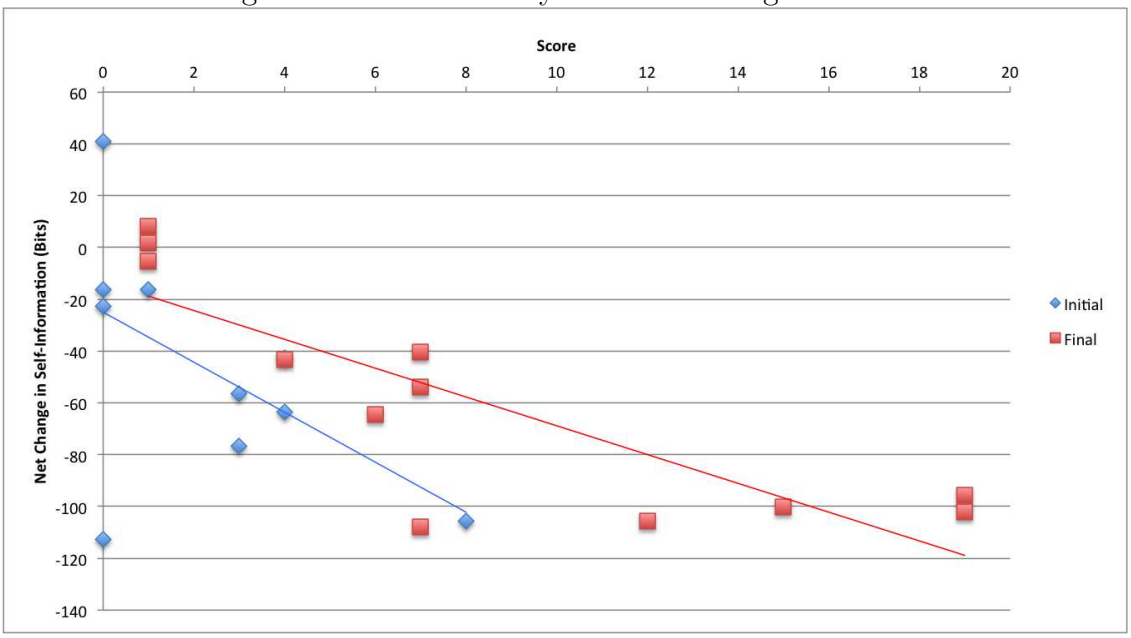
Figure 5.7: No Stop-Words Self-Info Change vs. Score



points. Figure 5.2 shows the scores for each participant, between the two sessions. Two participants completed their analysis of the documents in only one round, so their sessions are considered part of the final session group.

As described in section 5.1, there are three ways to treat the document set—we can use

Figure 5.8: Entities-Only Self-Info Change vs. Score



the full text, or remove the "stop words", or use only the extracted entities. Here, we use all three approaches on both the entropy and self-information scores, and plot the net change in our metric to the score earned with that configuration of groups. In Figure 5.3, we can see that despite uncovering much more of the plot, the higher-scoring participants didn't create clusters that were lower-entropy; this is true for the "stop words" approach as well, shown in Figure 5.4. However, when we consider only the entities present in the documents (Figure 5.5), a gradual trend towards lower entropy appears as scores increase. The counterintuitive results for the full-text and "stop words" approaches may be related to the factors investigated in section 5.1, and will be discussed further in section 5.3.

Looking at the self-information measures, we can see that our expectations were correct. The full-text approach shown in 5.6 demonstrates the decreasing trend in the metric as users' scores increase. This trend continues in the "stop words" condition, Figure 5.7, and is even more dramatic in the entities-only approach, as shown in Figure 5.8.

5.3 Alternate Calculation of the Metric

One thing we noticed during the user study was that users sometimes turned to logical-but-semantically-useless schemes for organizing the documents. If a user puts 15 of the 20 documents into one group, as happened more than once during the study, there are only a few documents left to be spread between the other three groups. The key factors in calculating entropy are the number of documents, and the number of unique words—so

packing a cluster with a high number of documents, simultaneously skewing the number of unique words, will show one cluster with relatively high entropy and the others relatively low. This would create the appearance of a net decrease in entropy, while being useless from a sensemaking perspective. As we can see in Figure 5.3, at a score of zero there are multiple net-improvement values that meet or exceed the improvement shown by the high-scoring participants. This is true even in Figure 5.5, though the trend is what we expected.

Still, we feel there is potential for the entropy metric. As described in sections 5.1 and 4.2, we have calculated the "randomized average" entropy by placing all the documents in the set into the clusters randomly, and recording the average entropy over repeated runs. This means that, over 60 runs, if we have 20 documents between 4 clusters our "randomized average" will be based on clusters with roughly 5 documents. However, if we change our method slightly, so that we calculate *each cluster's* "randomized average" based on the number of documents in the *user-created cluster*, we may be able to control for the case where a user creates unusually large groupings. This is, in a sense, normalizing the "randomized average" based on cluster size. Put another way, we would be comparing users' clusters to randomized clusters with the same number of documents—if a user has one cluster with 14 documents and three more clusters with only 2 documents in each, the improvement of each cluster would be based off of random clusters with 14, 2, 2 and 2 documents respectively.

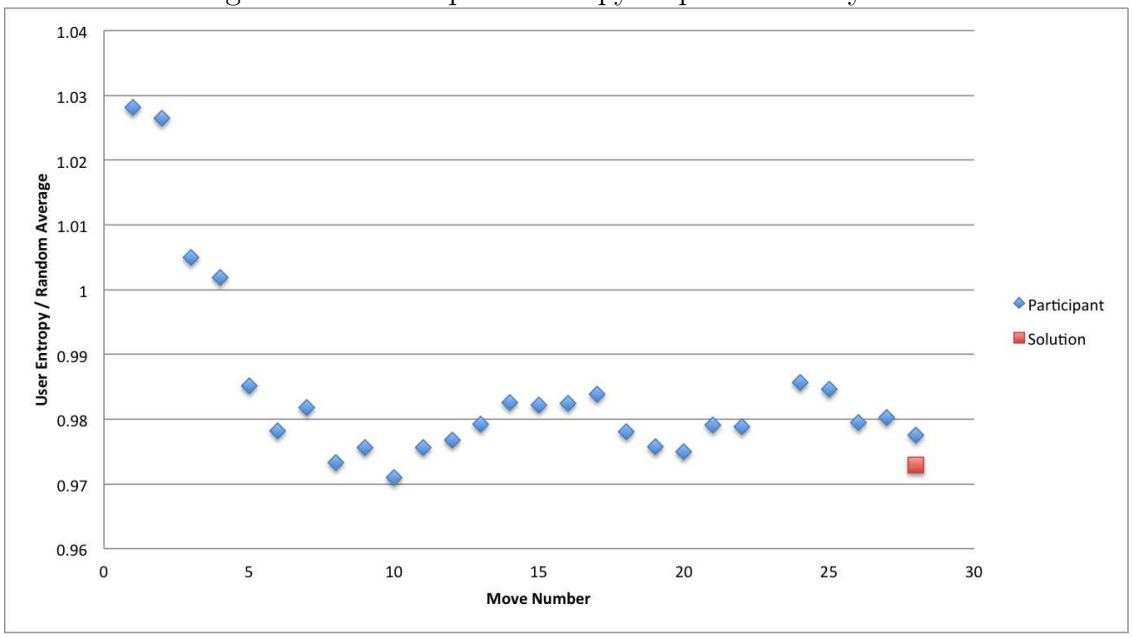
We implemented and tested this "Per-Document" method, with the results shown below. We used the same user groupings from section 5.2.2, again considering the net change across all of a user's clusters—so direct comparisons can be drawn between the two methods. Unlike our original method, the per-document values show a strong link between score and entropy (Figures 5.10, 5.11, and 5.12). This is especially noticeable for the results of the final session, where we can see the full extent of user's understanding reflected in their score. There was also a dramatic reduction in the range of values, from the (0.5, -1.75) seen in Figure 5.3 to (0, -0.75) seen in Figure 5.12. We take this as an indication that this new method is better able to handle unbalanced clusters.

The self-information metric does demonstrate the connection we were looking for, and the relationship seems to grow stronger with the better-targeted sets of symbols (no stop words, or just entities). This is true for both our initial results in section 5.2.2, and the results of our new method (Figures 5.13—5.15). While the small sample and qualitative scoring system doesn't lend to a meaningful statistical analysis, we do consider these results to support our

Table 5.5: % vs. Random, Entropy and Self-Information, Per-Document

Symbol Set	Entropy	Self-Info
Full Text	97.6	97.3
No Stop Words	96.4	96.1
Entities Only	93.6	94.9

Figure 5.9: Participant Entropy Improvement by Move



hypothesis that decreased information-theoretic metrics correspond to improved insight.

We also found that this revised method resulted in mutual improvements between the entropy and self-information metrics. Comparing the net improvement across clusters, as we did in table 5.4, there is between a 2.5% to 6.4% reduction in entropy and self-information in the sorted document groups (shown in table 5.5). Unlike our initial method, the entropy and self-information improve by approximately the same amount for each symbol set used.

Finally, because we control for the number of documents, we can use this per-document method to plot the net change in a user's groups, move-by-move. Starting at the first instance of a 4-cluster workspace (for consistency of comparison), we calculate the size of the clusters and show the net reduction as a percentage of the randomized average. Our results, shown in Figure 5.9, indicate that as participants spend time making sense of the documents, the entropy of their clusters decreased further with respect to the randomized average. The large decreases at the beginning of the process likely stem from the inclusion of more documents to the workspace—as the users read and place reports, there is both larger "random average" entropy AND more similarity in symbols that could drive entropy reduction. Review of the session video shows that, although the participant did not verbally indicate strong connections, there are some non-verbal cues that suggest moments of insight (excited finger tapping, rapid scanning through other documents) that happen in conjunction with the lower-entropy moves. These also often move documents together in a way that matches the known solution, so it is tempting to call these "good" moves. This reinforces the results of the large-time-step, initial-versus-refinement session analysis described by Figures 5.10—5.15.

Figure 5.10: Full-Text Entropy Change vs. Score, Per-Document Implementation

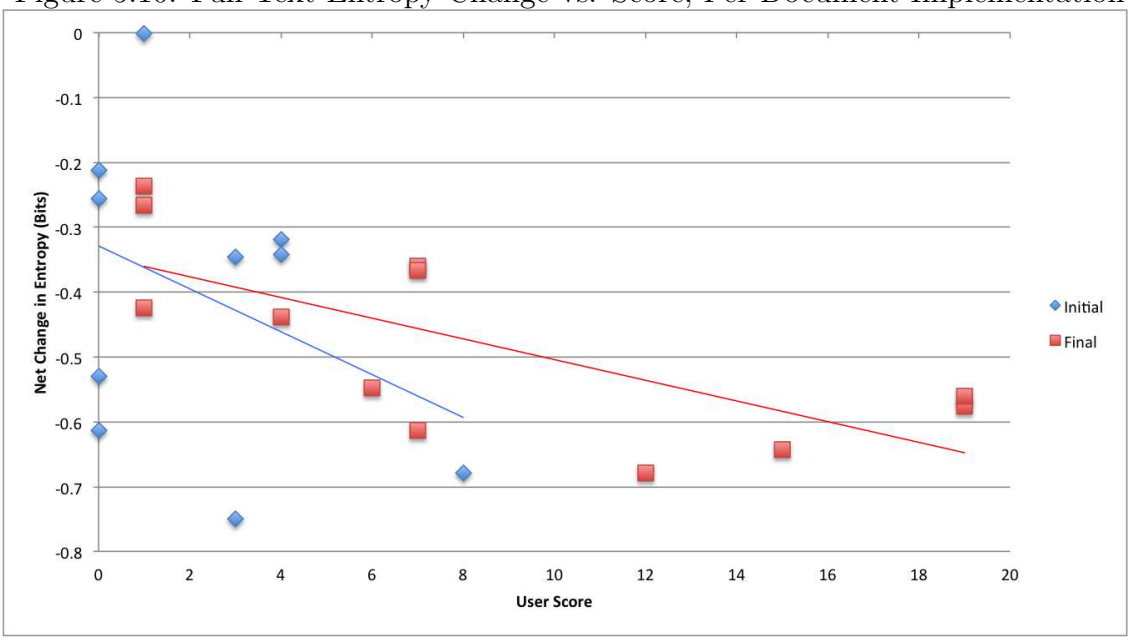


Figure 5.11: No Stop-Words Entropy Change vs. Score, Per-Document Implementation

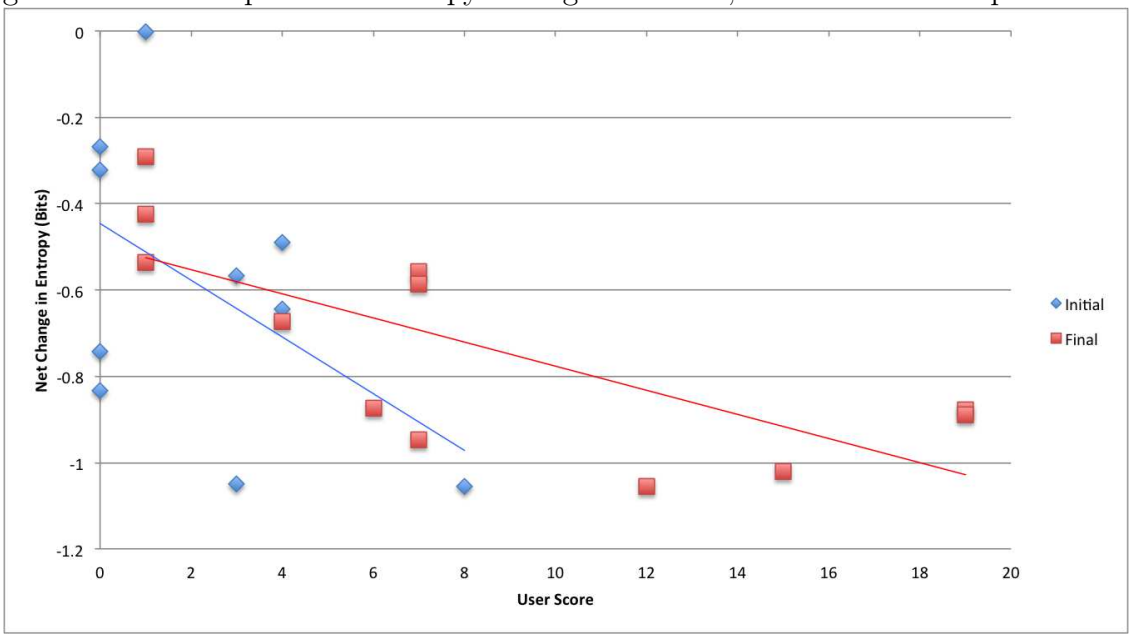


Figure 5.12: Entities-Only Entropy Change vs. Score, Per-Document Implementation

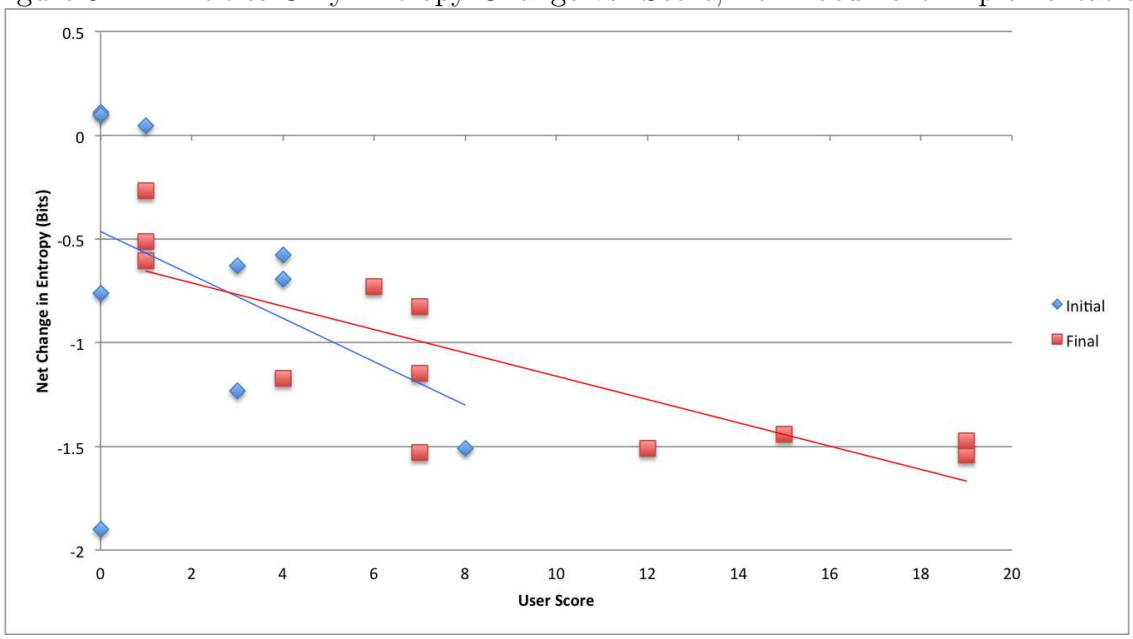


Figure 5.13: Full-Text Self-Info Change vs. Score, Per-Document Implementation

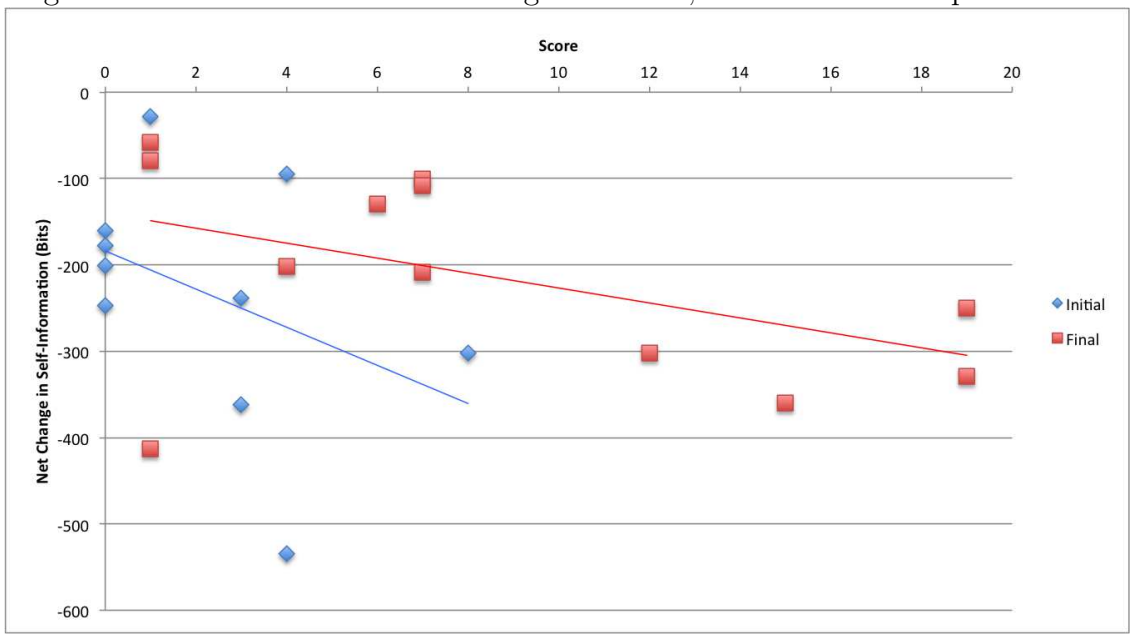


Figure 5.14: No Stop-Words Self-Info Change vs. Score, Per-Document Implementation

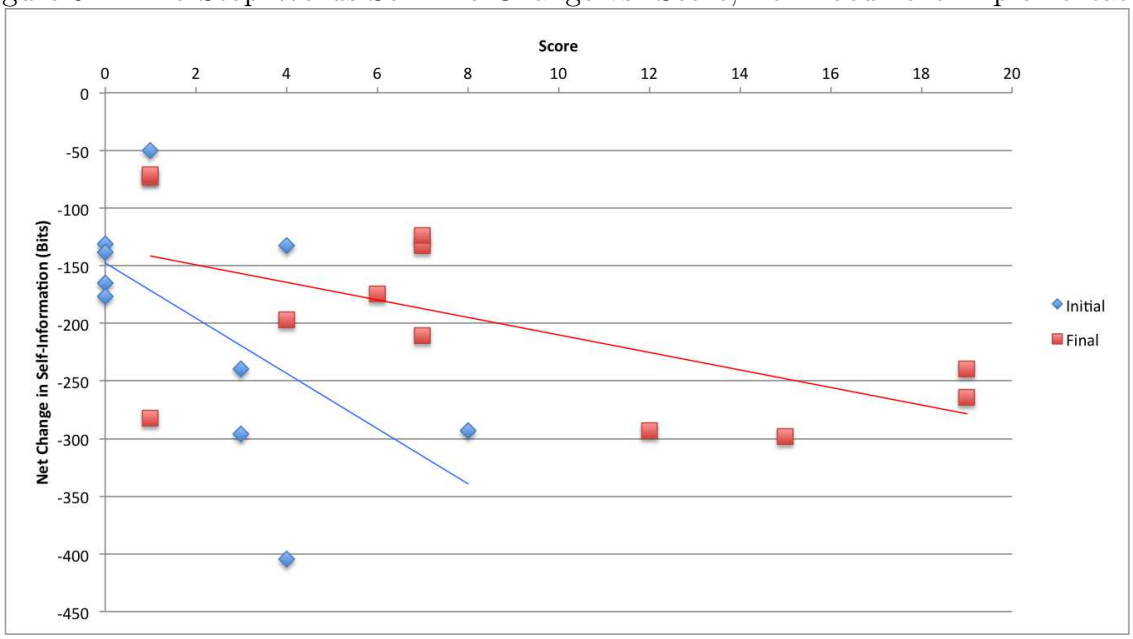
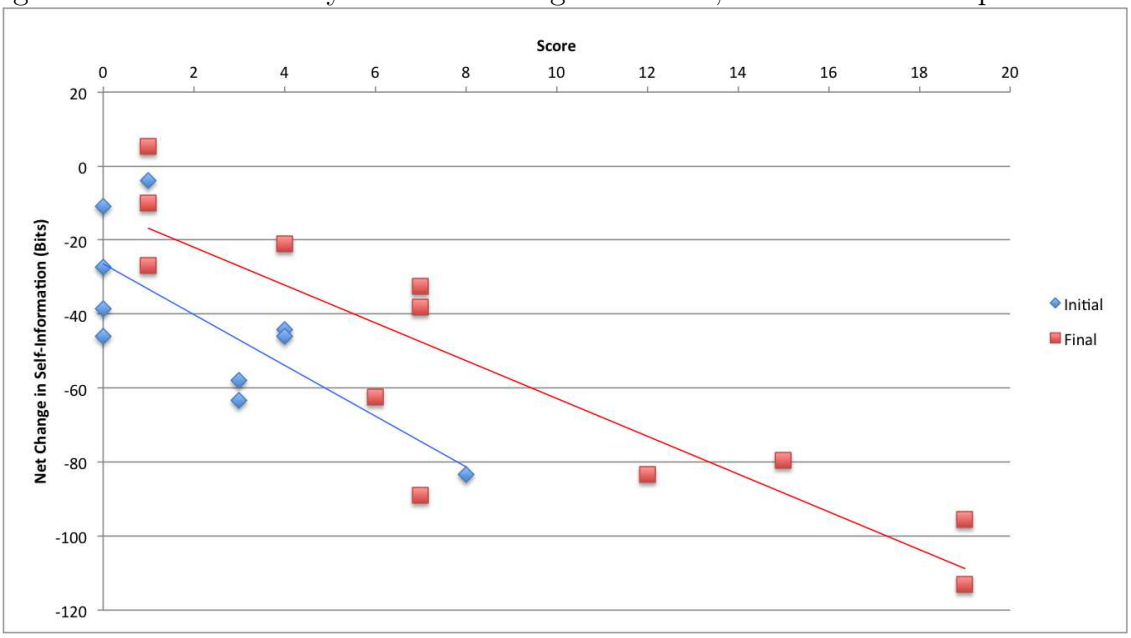


Figure 5.15: Entities-Only Self-Info Change vs. Score, Per-Document Implementation



Chapter 6

Discussion

The task now is to synthesize the two parts of our research, and see if our hypothesis is supported. To start, let's review the findings of the interrelated parts of this project. In section 5.1, we uncover factors that have the largest impact on the entropy and self-information metrics we are investigating; our result shows that entropy is affected by the number of unique words, and by the number of documents. Self-information is affected by the number of documents in the set, and by the length of those documents. It also lets us answer our first question: do our measures change as documents are structured in groups? Yes, they do, in measurable and statistically significant ways.

In section 5.2, we wonder if we can see a correlation between our measures and users' insight. Here, the results were mixed in our initial analysis—self-information seems to be strongly linked to the number of plot details users uncovered, but entropy seems only tentatively connected after enough preprocessing of the data. In section 5.3 we refine our method for calculating the randomized average, demonstrating that the entropy of a user's clusters decreases as they group the documents, and that improved sensemaking scores correspond to lower entropy. This confirms our hypothesis that the measures are linked to understanding, but it comes with some qualifications. First, we focus on a text-based sensemaking task, in a workspace with some inherent affordance of spatial grouping; applications to other domains need further investigation. More critically, our initial goal was to link *insight*, not simply understanding. There is doubtless a great amount of overlap between these two ideas, but to actually pinpoint instances of insight is more difficult with the failure of our metacognitive measure—we must use the think-aloud results, and hope that participants verbalized their thought processes well enough that we can identify those 'Aha!' moments. Reviewing the sessions, it seems that the weakness of our metacognitive measure was in the open-ended nature of our task; users were creating a semantic structure for the documents, then assigning confidence based on how well each document fit into *their own* structure, rather than how well the grouping helps understand the plot. This means that, for example, someone may put 15 of the 20 documents into one pile with high confidence, because they are all FBI

reports.

There is still much more that can be investigated or implemented. Our intent, when starting this exploration, was to uncover a method for evaluating visual analytics tools that could be compared across systems; our research hints at a solution, but much more implementation and testing is required before we could claim progress in this area—most notably, a user study with a much larger set of documents. A potential application is using cluster compressiblity (entropy) indicators as a feedback tool for analysts in real time, perhaps encoded as a visual indication of how well a document might relate to the existing contents of a group—added to systems like those by Wenskovitch and North [22] or Endert et al. [8]. For example, in an interactive VA system like StarSPIRE [4] or Jigsaw [21], some color coding might be shown while interacting with objects in the spatial/graph layouts; moving an object near others could provide an indication of the relative improvement in complexity achieved by grouping the objects together.

A related application could be use as a high-level overview of the relative improvements in the user's clusters—in this application the determination of clusters could be explicit or algorithmic, and would help indicate patches of complex information that could use more attention. Indeed, during the course of this research we found ourselves using our analysis tool to investigate *why* some clusters where higher entropy than others—so we have already seen that it has some applications in text-based visual analytics.

The fact that unusually low net entropy scores can be achieved by very low-scoring participants (as seen in Figure 5.10, for example) suggests that our approach must be interpreted with respect to the semantic meaning created during the process; an automated system may manage to minimize the net entropy, but without a human in the loop there may be no meaning to be gleaned from the groups. This also touches on the issue of useful truths versus useless truths: a person might group documents together in a way that is technically correct, according to a scheme that does nothing to help them make sense of the information. An example would be grouping documents by month—it's correct, but possibly not useful.

While on the topic of machine learning systems like StarSPIRE, more work needs to be done to understand the interaction between our metrics and automated systems for selecting and displaying documents to analysts. On the surface, it seems as though measures of entropy and self-information might be confounded by the potentially homogeneous documents such systems might arrange in the users' workspace. However, it's possible that interesting and unexpected similarities might be detected that were previously obscured by more overt traits.

Another avenue for further investigation is the move-by-move changes. In Figure 5.9 we can see that the sensemaking progressed in a series of hops, producing local minimums—though in the end, this particular participant was able to discern most of the plot. But what does this move-by-move offer some of the other users who became mired in organization schemes that were much less informative? Can this be used to identify such local minimums, in the hope of guiding the users to a better solution?

Finally, it is worth emphasizing that this work focused on a fairly narrow range of the visual analytics problem space. Text-based intelligence analysis is far from the only task that VA is called on to support, and we feel that there are potential applications of information theory beyond this task. As we know, almost anything that can be represented in a digital format can be compressed and transmitted—to apply this metric to a domain other than text-based analytics, we will need to give some thought to what the meaningful atoms are in a particular representation. Where we use words as symbols in this exploration, perhaps genes would work as symbols in another application; in yet another situation, maybe various color palettes would be appropriate as symbols. Where we use explicit groupings in folders to define what constitutes a cluster, maybe the pattern of access would be a more appropriate delineation in another application. It seems like a topic full of potential for further exploration.

Chapter 7

Conclusion

To answer the question, "Can we estimate insight generation by measuring the reduction in complexity of a workspace over time, using information theoretic measures of entropy and self-information?", we start by looking at the factors that influence the measures (section 5.1). We found that the number of documents in the set, the number of unique words in those documents, and the average lengths of the documents all influence our metric; at the same time, we see that our metric is independent of the number of clusters calculated over. We also show that entropy and self-information improve when the documents are grouped correctly. In section 5.2 we use results from our user study to compare how our metrics relate to the understanding of the documents demonstrated by users. While we found mixed results initially (section 5.2.2), our understanding of the factors allowed us to improve the way our metric was calculated—producing results that demonstrate a consistent improvement in entropy and self-information, as users' scores improve.

Put concisely, we demonstrate that the entropy and self-information improves, as documents are put into semantically meaningful groups by users.

Our contributions are:

1. Development of a metric that demonstrates change in entropy and self-information, in groups of documents
2. Characterization of the metric, showing significant results for various factors
3. Application of the metric to user-generated groups, demonstrating a link between the metric and users' understanding.

Measuring complexity means analyzing the documents that an analyst is working to understand (in other words, the analyst's workspace), and assigning value to the contents in such a way that grouping related information together produces a detectable change. The

information-theoretic notions of entropy and self-information provide just such measures of complexity. We've shown that these measures do change as users sort and semantically structure their workspace, and we've shown a link between the measures and the users' performance on a sensemaking task. Complications with our approach to tracking insight means we were not able to get the fine-grained look at how a user's "aha!" moments (described in [5]) correspond with our information-theoretic measures. However, if we consider insight to be the more gradual, knowledge-building process described by North [15] we show that increases in insight *do* correspond with improvements in our measures. Taken collectively, our work suggests that we can potentially quantify insight generation via Shannon entropy and self-information measures.

Chapter 8

Bibliography

[1] Y. a. Kang, C. Gorg, and J. Stasko. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 139–146, 2009.

[2] Christopher Andrews, Alex Endert, and Chris North. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 55–64. ACM, 2010.

[3] T. L. Bauer. Information and meaning: Revisiting shannon's theory of communication and extending it to address today's technical problems. Technical report, Sandia National Laboratories, 2009.

[4] Lauren Bradel, Chris North, and Leanna House. Multi-model semantic interaction for text analytics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 163–172. IEEE, 2014.

[5] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky. Defining insight for visual analytics. *IEEE Computer Graphics and Applications*, 29(2):14–17, March 2009.

[6] Yang Chen, Jing Yang, and William Ribarsky. Toward effective insight management in visual analytics systems. In *Visualization Symposium, 2009. PacificVis' 09. IEEE Pacific*, pages 49–56. IEEE, 2009.

[7] Francis T Durso, Cornelia B Rea, and Tom Dayton. Graph-theoretic confirmation of restructuring during insight. *Psychological Science*, 5(2):94–98, 1994.

[8] Alex Endert, Seth Fox, Dipayan Maiti, Scotland Leman, and Chris North. The semantics of clustering: Analysis of user-generated spatializations of text documents. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 555–562, New York, NY, USA, 2012. ACM.

[9] Stephen M Fleming and Hakwan C Lau. How to measure metacognition. *Frontiers in human neuroscience*, 8, 2014.

[10] Dugald Ralph Hutchings, Greg Smith, Brian Meyers, Mary Czerwinski, and George Robertson. Display space usage and window management operation comparisons between single monitor and multiple monitor users. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pages 32–39, New York, NY, USA, 2004. ACM.

[11] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[12] Laura A. McNamara, Travis L. Bauer, Michael Haass, and Laura Matzen. Information theoretic measures for visual analytics: The silver ticket? In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, pages 53–61, New York, NY, USA, 2016. ACM.

[13] Janet Metcalfe and David Wiebe. Intuition in insight and noninsight problem solving. *Memory & cognition*, 15(3):238–246, 1987.

[14] Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. Getting access to what goes on in people's heads?: Reflections on the think-aloud technique. In *Proceedings of the Second Nordic Conference on Human-computer Interaction*, NordiCHI '02, pages 101–110, New York, NY, USA, 2002. ACM.

[15] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, May 2006.

[16] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4, 2005.

[17] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, July 2005.

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

[19] Jean Scholtz. Beyond usability: Evaluation aspects of visual analytic environments. In *Visual Analytics Science and Technology, 2006 IEEE Symposium On*, pages 145–150. IEEE, 2006.

[20] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

[21] John Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[22] John Wenskovitch and Chris North. Observation-level interaction with clustering and dimension reduction algorithms. In *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*, HILDA'17, pages 14:1–14:6, New York, NY, USA, 2017. ACM.

[23] Ji Soo Yi, Youn-ah Kang, John T. Stasko, and Julie A. Jacko. Understanding and characterizing insights: How do people gain insights using information visualization? In *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization*, BELIV '08, pages 4:1–4:6, New York, NY, USA, 2008. ACM.

[24] Kirsty A Young. Direct from the source: the value of 'think aloud' data in understanding learning. *The Journal of Educational Enquiry*, 2005.

Appendix A

Study Prompt

Our goal is to measure the compressibility of information changes as humans sort and cluster documents in some way, and see if it correlates with insight. In order to do that, we need to have some people sort some reports so we can look at the result. The task for you today is to quickly look at these documents, and sort them into groups in a way that allows you to explain the "who, what, when, where, and how" of the plot. Also, I want to encourage you to use all four of the available folders to form your groups.

While you go through these reports, I'd like you to try and talk out loud about what your thoughts are on the relationships between the reports, and why you think the various reports do or don't belong together. While you're doing this, I may ask you for more information or to clarify a comment.

The procedure is as follows: The randomly ordered intelligence reports are, initially, all stacked together outside of the folders. One at a time, I would like you to skim a report, and place it in a folder you feel it best fits with. You can move documents between folders as often as you like, documents from the stack must be placed after reading. Also, each time you place a report in a group, I'd like you to state your level of confidence in that choice—saying either that you're confident they belong together, or not very confident. This can be done by saying either "High" or "Low" as you place the document. For example, having no idea—placing a document at random—would be low. You don't have to be 100% positive for the "high" condition—a feeling that these documents belong together would be enough.

You will have 15 minutes to make an initial placement, then another 10 minutes to refine your groups. At the end of this, we will have you describe the plots you have uncovered.

Appendix B

User Study Scoring

1 point for each cluster that roughly describes one thread of the plot:

- Adam Randall/PA
- Mark Davis/Queens
- Charlottesville/ Students /Faysal Goba
- Amsterdam/Hans Pakes/Europe

1 point for each for being able to describe the general theme of the plot, beyond terrorism or explosives.

- Randall/Aykroyd as coordinators
- Charlottesville planning an attack with C-4
- NYSE Attack with C-4
- Holland Queen

1 point for each specific plot detail:

- All people correctly grouped
- C-4 stolen from PA was distributed by Aykroyd
- Randall provided money and phone coordination of cells
- Goba and students are working together

- Target is Atlanta, or Amtrak train itself
- Storage Unit was used to store at least some of their explosives
- Mark Davis and roommate are targeting NYSE
- Davis is known Al Qaeda
- Hans Pakes is transporting something on Holland Queen
- Dirty Bomb
- Target is Boston
- Specific date of attacks

20 points total

Appendix C

Study Documents

Report Date 1 April, 2003. FBI: — Adam Randall is the owner of the Select Gourmet Foods shop in Springfield Mall, Springfield, PA. [Phone number 703-659-2317]. First Union National Bank lists Select Gourmet Foods as holding account number 1070173749003. Six checks totaling $35,000 have been deposited in this account in the past four months and are recorded as having been drawn on accounts at the Pyramid Bank of Cairo, Egypt and the Central Bank of Dubai, United Arab Emirates. Both of these banks have just been listed as possible conduits in money laundering schemes.

Report Date 5 April, 2003. FBI: — Passport control at Dulles Airport in Wash DC records that Adam Randall, holder of US passport# 177183634 [issued by Passport Agency, Wash. DC on 12 Feb. 1997] has made three trips to Amsterdam, two trips to Hamburg, Germany, and three trips to Cairo, Egypt in the last five months. The address given by Randall on his passport is 1176 Floyd Ave., Springfield, PA. Phone number at this address is 703-734-0104.

Report Date 15 April, 2003. FBI: — The Powhatan Company is a manufacturer of military explosives of various types, including C-4. It is located just outside of Springfield, PA. On 11 April, 2003 this company reported that 200 pounds of C-4 could not be accounted for during a recent inventory.

Report Date 18 April, 2003. FBI: — A routine check of security at the New York Stock Exchange (NYSE) reveals some anomalies in background checks of several persons who now hold vendor's IDs that allow them access to the NYSE provided that they are accompanied by security guards. A man named Mark Davis, reported age 32 years, obtained a social security card and a New York State Driver's license in 1999 using a birth certificate now believed to have been forged. He is employed by Empire State Vending Services in Manhattan and he services vending machines such as coffee, soft drink, and candy machines. He lists his home address as: 2462 Myrtle Ave. Apt. 307, Queens, NYC.

Report Date 18 April, 2003. FBI: — An INS check of expired student visas reveals the names Mukhtar Galab and Yasein Mosed. They have been enrolled at the University of

Virginia in Charlottesville, Virginia. Checks with the University of Virginia reveal that these two persons have not attended any classes for the past two semesters. The address they both gave to the University of Virginia is 2932 University Drive, Charlottesville, VA. There is presently no one living at this address. A check with mobile phone providers shows that a Sprint cell phone #804-774-8920 is registered in the name Mukhtar Galab.

Report Date 20 April, 2003. FBI: — Investigating Mark Davis reveals that Bagwant Dhaliwal also lists 2462 Myrtle Ave. Apt. 307, Queens, NYC as his home address, and is employed by Empire State Vending Services in Manhattan.

Report Date 20 April, 2003: FBI: — Mukhtar Galab has an account at the Virginia National Bank in Charlottesville, VA. Bank records say he has deposited several checks in the last three months, totaling $13,000, drawn on account number 1070173749003 at the First Union Bank in Springfield, PA.

Report Date 22 April, 2003. FBI: — Herbert Aykroyd , of North Bergen, PA, has deposited checks in his bank account that were drawn on First Union Bank account number 1070173749003 in Springfield PA in the name Adam Randall. The latest check is dated 16 April, 2003 and was in the amount of $8500. Examining payment records shows Aykroyd has sent several packages to Charlottesville, VA and Queens, NYC.

Report Date 25 April, 2003. FBI: — A report from AMTRAK reveals a reservation, paid in cash in Charlottesville, and made by Faysal Goba on 23 April, 2003. Reservation is for three one-way first class tickets and one sleeping compartment from Charlottesville, VA to Atlanta, GA on 29 April, 2003. Reservation is on AMTRAK Train #19. Reservations are in the names: Faysal Goba, Mukhtar Galab and Yasein Mosed.

Report Date 26 April, 2003. FBI: — A check of rented storage facilities in the Richmond and Charlottesville areas reveals that a man giving his name as Faysal Goba rented storage unit #174 on 10 April, 2003 at the Budget Storage Units in Richmond, VA. Goba gave his address as 2932 University Drive, Charlottesville, VA. Goba paid in cash for a month's rental. In an examination of storage unit #I74 at the Budget Storage Units in Richmond, VA, fifty pounds of C-4 plastic explosive were found along with some fusing devices.

Report Date 26 April, 2003. Coast Guard intelligence to FBI and CIA: — A routine reporting of ships bound for the USA. This report contains departure date, place, destination, cargo manifest and crew roster and is required 96 hours before arrival in ports in the USA. Report from Amsterdam on 25 April, 2003 lists one container ship bound for the US, the Holland Queen, which is bound for Boston. Arrival time in Boston of the Holland Queen is 29 April, 1930 hrs.

Report Date 27 April, 2003. FBI: — A photo of the man using the name Mark Davis was examined by a representative of the Canadian police in NYC. The Canadian police investigator identified the man in the photo to be a Saudi who overstayed a travel visa and is wanted by the police in Canada. It is now known that Davis received explosives training in the Sudan and in Afghanistan.

Report Date 27 April, 2003. FBI [From police in North Bergen, PA]: — In the early morning hours of April 26, 2003 a passerby reported a fire in a carpet shop that is managed by a Herbert Aykroyd of North Bergen, PA. While firemen were extinguishing the blaze, they discovered several cartons labeled: PRIVATE: DO NOT OPEN.These cartons contained C-4 explosive. Attempts to reach Herbert Aykroyd have not been successful. An employee at the carpet shop later told police that Aykroyd had just gone on a vacation in Canada and that he had left no address.

Report Date 14 April, 2003. CIA: — INS check reveals that a Faysal Goba entered the USA on a travel visa in January of 2003 stating that he would be visiting a person named Clark Webster in Richmond, VA. The contact address given by Goba was: 1631 Capitol Ave., Richmond VA; phone number: 804-759-6302. Following up on Clark Webster reveals that there is no residence with address 1631 Capitol Ave. in Richmond, VA. The phone number given for this address (804-759-6302) is in fact a Sprint cell phone registered in the name Faysal Goba.

Report Date 22 April, 2003. CIA [From Dutch Security]: — Two men were arrested on April 20,2003 by Dutch police in Haarlem, The Netherlands after leaving the scene of an accident in which they were involved. Later, on April 21, 2003 these men were identified as Tawfiq al Ahdal and Saeed Khaliad who have been wanted in Hamburg, Germany in connection with investigations of Al Qaeda operations in Europe. They were driving a panel truck rented in Hamburg, Germany. Radioactive traces were found in bed of this truck.

Report Date 24 April, 2003. CIA [From Dutch Security]: — An address in Haarlem was found in the truck driven by Tawfiq al Ahdal and Saeed Khallad. This address is for an apartment rented in Haarlem by two Saudi's who hold Dutch passports, and one Hans Pakes. The apartment owner says they also rent storage unit #206. The two men al Ahdal and Khallad are employed as container loaders at Sealink Container Corp. in Amsterdam. This corporation owns a large fleet of containers that are shipped all over the world and subsequently hauled on trains or trucks. Inspection of unit #206 on 23 April, 2003 reveals radiation traces, one pair of lead gloves, and two sticks of TNT. Also found are pieces of a wooden crate. Pieces of this crate, also revealing radioactive traces, are labeled and identify a company in Lublin, Poland.

Report Date 25 April, 2003. CIA [From Dutch security]: — Analysis of radioactive particles found in truck driven by Tawfiq al Ahdal and Saeed Khallad, and in the storage unit rented by them and Hans Pakes, reveals them to be powdered cesium 137. One ounce of this substance could spread radioactive fallout over 60 city blocks.

Report Date 27 April, 2003. CIA: — The Holland Queen crew roster lists a person named Hans Pakes as being aboard the Holland Queen. Pakes works in the ship's galley. From an Al Qaeda laptop computer captured in Afghanistan there is a roster of trainees in an explosives training unit in Kandahar, 1996. A Yemeni is listed on this roster, and says he left for The Netherlands in January 1996. He is using the alias: Hans Pakes.

Report Date 27 April, 2003. CIA: —— From a laptop computer captured in Afghanistan it was learned that a Pakistani named Sahim Albakri, who fought with the Taliban in 1990 - 1992, travels using an Indian passport in the name Bagwant Dhaliwal. Intercept of phone calls made from Bagwant Dhaliwal at 2462 Myrtle Ave. Apt. 307, Queens. NYC revealed several calls to a phone 732-455-6392. In the latest call, the caller from 2462 Myrtle Ave. Apt. 307, Queens, NYC confirmed that he received the carpet he ordered on April 25,2003.

Report Date 24 April, 2003: —— Phone calls on 22 April, 2003 from Adam Randall (703-659-2317) to the following numbers: 804-759-6302, 804-774-8920, 718-352-8479, and 01 1207670734. The same brief voice message was given in Arabic in each call. A translation of this message reads: "I will be in my office on April 30 at 9:00AM. Try to be on time".