Toward a Deeper Understanding of Data Analysis, Sensemaking, and Signature Discovery

Sheriff Jolaoso, Russ Burtner, Alex Endert

Toward a Deeper Understanding of Data Analysis, Sensemaking, and Signature Discovery

Sheriff Jolaoso[1] Russ Burtner[2] Alex Endert[3]

[1]Virginia Tech [2]Pacific Northwest National Laboratory [3]Georgia Tech

sheriff1@vt.edu

Abstract. Data analysts are tasked with the challenge of transforming an abundance of data into knowledge and insights. This complex cognitive process has been studied, and models created to describe how the process works in specific domains. Two popular models used for this generalization are the sensemaking and signature discovery models, which apply a cognitive and computational focus to describe the analytic process, respectively. This work seeks to deepen our understanding of the data analysis process in light of these two models. We present the results of interviews and observations of analysts and scientists in four domains (Biology, Cyber Security, Intelligence Analysis, and Data Science). Our results indicate that specific aspects of both models are exhibited in the analysts from our study, but neither describe the holistic analysis process.

Keywords. Analytic process; sensemaking; signature discovery; visual analytics; data analysis

1 Introduction

Data analysts are tasked with coming about insightful conclusions based on large amounts of unstructured information provided to them. As such, understanding and generalizing the process of gaining such insight has been widely studied. For example, the literature on *sensemaking* depicts the iterative process of developing insight through foraging information and synthesizing knowledge by mapping features from data to a sets of hypotheses, and then performing assessments to understand the resulting complex relationships [1]. Sensemaking is a cognitive activity that entails the iterative development, re-assessment, and refinement of knowledge structures formed by prior experiences and data. Two popular models of sensemaking, Pirolli and Card's sensemaking loop [2] and the data/frame model presented by Klein et al. [3], present frameworks by which this process can be thought of. The emphasis of both of these models is on the cognitive aspects of users performing such tasks.

Another model, called the *signature discovery process* [4], is a model aimed at understanding the data-centric counterpart to such tasks. This approach (shown in Figure

1) takes a more quantitative and computational approach to describing the analysts' processes. Signature discovery describes the process of developing a signature based on an insightful understanding of the data gathered from both computational and cognitive reasoning. The signature discovery process requires analysts to specify a problem, and categorize the necessary data that needs to be inventoried, assessed, and explored. From this data, analysts can find a unique or distinguishing measurement, pattern, or collection of the data that can be used to detect, characterize, or predict a target phenomenon (e.g., state, object, action, or behavior) of interest. As a result, the *signature* represents a mathematical quantification of the insight. For example, a signature can be realized as a parameterized classifier used to generate a cluster of images, where that cluster represents an interesting finding as determined by the user.

In this paper, we present the findings of a user study consisting of semi-structured interviews and observations of data analysts in four domains: Biology, Cyber Security, Intelligence Analysis, Data Science. For this paper, we broadly define *data analyst* and *analyst* to describe a person who is tasked with gaining insights from data, not as a specific job title or formal description of their task. The study seeks to understand how the practice of data analysis across these four domains compares to two current models describing such processes: the sensemaking loop and signature discovery. Our results indicate that while there exists overlap with current models, as the fluidity and personalization of the participants' workflows carries inherent complexity. The primary contributions of our work are:

- Providing a deeper understanding of the data analysis processes in the four domains studied (Cyber Security, Intelligence Analysis, Biology, Data Science)
- Discovering that both sensemaking and signature discovery models depict specific aspects of data analysis, but neither fully describe the holistic process
- Discussion of findings to inform design of future data analysis tools

2 RELATED WORK

2.1 Signature Discovery

Baker et al. define a signature to be a unique or distinguishing measurement, pattern, or collection of data that detects, characterizes, or predicts a target phenomenon (state, object, action, or behavior) of interest as well as the transformation from events to measurements to features to categorical labels and associated uncertainties [4]. The steps of the signature discovery process are diagrammed in Figure 1. The signature discovery process is a linear process that entails problem specification, identification of observable data, specification of measurements for the observable data, feature extraction, and finally signature development with the option to re-assess specific steps. The result of signature discovery, along with a signature, is a signature system; a reusable function, or piece of insight that can be used on similar inputted datasets. Once applied, the process calls for users to test the suitability of the created signature for the task and domain [4].

This work seeks to further understand the definition of signatures, signature systems, and the signature discovery process from the viewpoints of analysts from the four domains chosen. Further, these interviews and observations will help define the relationship between signature discovery and other sensemaking literature.

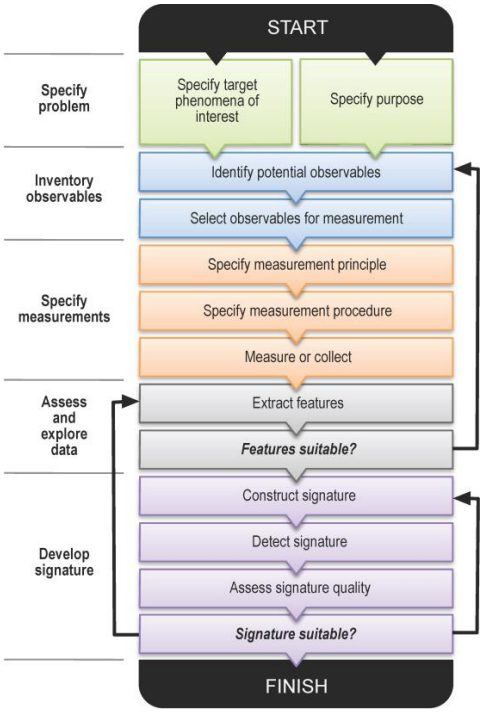

Figure 1. Diagram of the Signature Discovery Process (adapted from [1]).

2.2 Sensemaking

Sensemaking, as defined by Russell et al., is the process of searching for a representation to answer task-specific questions [1]. There are two major loops of activities in sensemaking – a foraging loop that involves processes aimed at seeking information, searching and filtering it, and reading and extracting information and secondly, a sensemaking (or synthesis) loop that involves iterative development of a mental model from the schema that best fits the evidence [1], [2], [5]. A model of the sensemaking (the sensemaking loop) is shown in Figure 2. The foraging loop progresses analysts from external, raw data sources to evidence files that refer to a more case-specific subset of information. The synthesis portion of this loop describes the steps involved in going from an evidence file to a final presentation of the hypothesis developed through schematizing, building a case, searching for evidence and support, and re-evaluating output, making this loop more of a cognitive, synthesizing activity.

The emphasis of sensemaking is on creating a mental model and does not explicitly contain components correlating to scientific tools, such as visualizations [6].

Another model of sensemaking, presented by Klein et al. [3], shows another descriptive model for sensemaking. Their model depicts an exchange of information between the human and the data in terms of frames. That is, a user has an internal "frame" that represents her current understanding of the world. Then, as she continues to explore a particular dataset, her frames of the world are mapped against the information she uncovers. If the information supports a specific frame, that frame is thought to strengthen. However, when evidence is discovered through exploration that refutes the existence of such a mental frame, the frame can either be augmented or a new one created. Their data/frame model consists of reframing data through filtering, making connections, and defining characteristics, which can be equated to the foraging loop from Pirolli and Card's work. Other steps and cycles elaborate frames, refine frames through questioning and preserve frames, and can be said to parallel aspects of the sensemaking loop defined in Pirolli and Card's work.

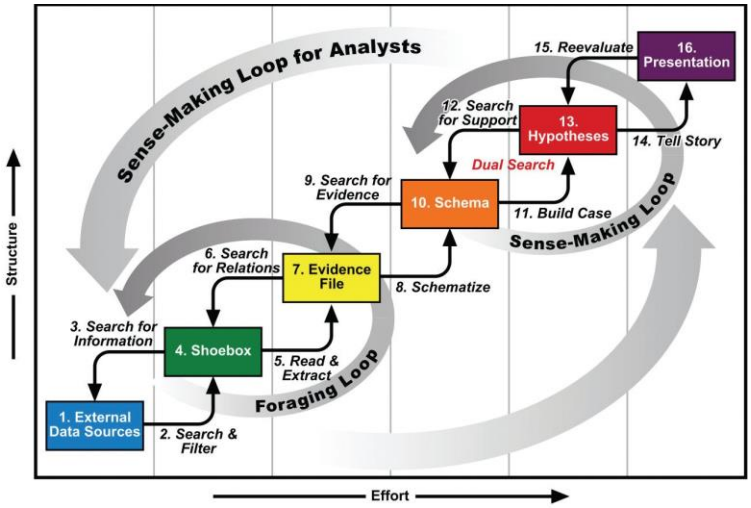

Figure 2. The sensemaking loop, presented by Pirolli and Card [5], depicting cognitive stages of data analysis.

2.3 Workflow Management Tools

It is assumed that there are computational steps in the analytic process. To carry out these computational steps, workflow management tools have been developed. It is expected that during the analytic process, that workflow management tools be used to automate computationally attractive steps. The steps used within these tools can potentially map to steps in the sensemaking and signature discovery processes.

For example, Taverna is a workflow management tool, first used in the area of bioinformatics, to assist in the procedural computation steps of conducting experiments [7]. The tool also provides provenance information, which is information that identi-

fies the source and processing of data by recording the metadata and intermediate results associated with the workflow. The provenance tracking capability enables users to conduct sensemaking activities, specifically those found in the foraging loop and the beginning of the sensemaking loop, through the use of visualizations and a graphical user interface. Steps of the signature discovery process where data is processed, such as feature extraction and measurement specification, can be paralleled to the calculation and data processing steps that a workflow management tool performs.

Similar, VisTrails visualizes the steps and interactions of the analytic process [8]. The tool captures the provenance of the visualization processes and the data they manipulate as well, which can enable the reevaluation steps within the sensemaking model. The tool also provides the ability to apply a dataflow instance to different data, which can be equated to the feature extraction aspect of signature discovery.

In this work, the importance of these tools and how they coincide with the signature discovery process and sensemaking is investigated.

2.4 Observing Analysts and Domain Experts

Analysis and assessments of the analytic process have been conducted to find out the challenges faced by analysts, design implications for analytic tool development, and future trends of the analytic process. For example, Fink et al. and Endert et al. observed how cyber analysts use visualizations and large high-resolution displays to perform a sensemaking task, producing findings about their processes [9]–[11]. The findings presented enhancements for analyst's performance, such as desire to save the state of their investigation at certain points, a need to provide rich linkages among multiple visualization tools, means of keeping a visual history of steps in the analysis, and deep access to the data provided. Similarly, Kandel et al. interviewed analysts to gain insight on their analytic process within the social and organizational context of companies. They were able to create archetypal patterns to categorize analysts and organizational schemes of workflow processes for enterprise workers [12]. Evaluations of analytic processes were also investigated by Scholtz et al., who (through the use of the VAST challenge) found the importance of assessing the process of the analysis separate from the accuracy of the resulting analysis [13].

The work presented in this paper builds upon the knowledge gained through these studies. First, we study additional domains to gain insights into how data analysis occurs across domains. Second, we map our findings to two existing models for data analysis (signature discovery and sensemaking).

3 Method

The study consisted of a series of semi-structured interviews and observations of participants performing analysis to illuminate domain-specific information from our subject matter experts. The questions asked during the interviews include what type of problems they are tasked with, what types of data they use, which tools they use in their analytic process, and how they measure success. The observations were con-

ducted by having the investigators passively watch as the analysts walked through one of their recent tasks. The sporadic (and at time chaotic and time-sensitive) nature of their actual analysis forced us to observe past data analysis processes.

3.1 Research Questions

We seek to address the following three research questions (RQ1 – RQ3), and their associated hypotheses (H1 – H3):

RQ1: What are the similarities and differences between the signature discovery process from Baker et al., the model of sensemaking from Pirolli and Card, and the analytic process as described by analysts?
H1: Our hypothesis for RQ1 is that the signature discovery process will reflect a more quantitative and sequential process, whereas sensemaking will be shown to be more fluid (as described in the related work). It is likely that the two may complement each other.

RQ2: What are tangible outcomes/products of the analytic process?
H2: We hypothesize that the outcome of the analytic processes is a signature as defined in [1] with little variance between domains and that a combination of the signature and signature system will form the analyst's presentation. That is, analysts seek to find repeatable mathematical descriptions of their findings.

RQ3: How do analysts use workflow management tools in their analytic process?
H3: Our hypothesis for RQ3 is that analysts use workflow management tools within their analytic process for feature extraction and signature construction (i.e., the more computationally-focused aspects of their work).

3.2 Data Analysis

From the semi-structured interviews, the investigators collected answers to the seed questions, captured on paper in natural language. These were later transcribed to digital versions, to ease indexing, searching, and annotating. The observations were conducted by having the investigators passively watch as the analysts walked through one of their recent tasks. The sporadic (and at time chaotic and time-sensitive) nature of their actual analysis forced us to observe past data analysis processes (instead of active analysis tasks). We collected notes on general observations, steps taken, and comments made by the analysts during these sessions. These notes were also transcribed to digital versions, similar to the interview responses.

Each of the three investigators performed an open coding process of binning the digital transcripts to reveal higher-level themes. While there were no formal, predefined topics or themes to code for, the investigators were aware of the previous models being tested (signature discovery and sensemaking). Thus, as each investigator individually analyzed the transcripts, caution was taken to mark instances where the transcripts describe or comment on aspects of either sensemaking or signature

discovery. After each individual investigator saturated the individual themes (or bins), we met collaboratively to compare, contrast, and final set of labels for observed instances. These were then mapped into the stages of the sensemaking and signature discovery models, shown in Figure 3. Note that this set does not consist of all the stages of sensemaking and signature discovery, as some stages or aspects were not observed.

3.3 Participants

Semi-structured interviews were conducted with sixteen analysts (P1-P16) in the areas of Biology (7), Cyber Security (3), Intelligence Analysis (2) and Data Science (4). Of these sixteen, one had to be discarded for analysis due to classification and data sensitivity concerns (P7).

Biology

The biologists that we interviewed covered a wide scope of work from Epidemiology to Genetics. Six of the seven hold a Ph.D. in Biology/Computational Biology, and have five or more years of experience in the field with strong publishing credentials.

Cyber Security

The cyber security analysts had strong IT backgrounds and have been established in the field for 10 or more years (two were only in their current job for 2 years). Their scope of work covered smaller local networks to broader more national investigative cyber networks. Their levels of education are Bachelors of Science to Masters in Computer Science or Computer Engineering.

Intelligence Analysts

The intelligence analysts we interviewed worked for multiple government agencies including the Department of Energy. Each had a diverse background in either a nuclear science related field (thus the "Nuke" label in figures in this paper) or computational science. Both have been in intelligence for more than ten years.

Data Science

The data scientist we talked to are all traditional computer science or formal statisticians. Their level of expertise varied from over 10+ years of experience to only two years out of school. Their level of education also varied from Master's degree to Bachelors of Science.

4 Findings

Through the thematic coding approach used to analyze the data collected from the interviews and data analysis observation sessions, we present the following findings.

First, we present details regarding the characteristics of the analytic process based on the data sources, tools used, and methods for recording their analytic provenance (or process). Second, we compare the self-reported and observed processes to existing models of sensemaking and signature discovery. Third, we analyze the analysts' methods for reporting the findings and insights.

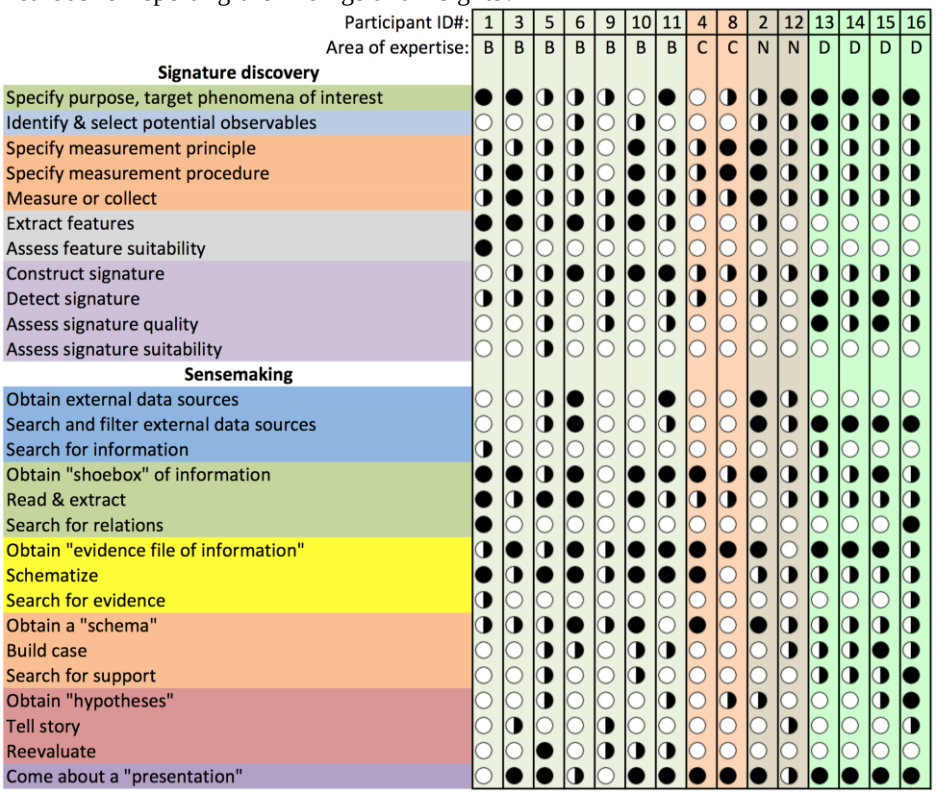

Figure 3. Summary of how the analytic processes of analysts interviewed relate to the signature discovery and sensemaking processes based on our data analysis. Full circles indicate steps explicitly mentioned or observed. Half circles indicate steps that were generally mentioned, but not explicitly performed. Empty circles indicate steps that were not observed or mentioned during the interviews. The areas of expertise are N=Intelligence Analyst, B=Biology, C=Cyber Security, and D=Data Science. P7 was removed for data classification and sensitivity concerns.

4.1 Analytic Process Characteristics

Data sources

All of the domains we studied utilized databases as a primary method for storing and retrieving their data. Among the intelligence analysts, data sources commonly used were media such as images and text. Our biology users accessed many scientific

databases shared among other scientists in their discipline or clinical study. The Data Science analysts were mostly users of a variety of observational sensor data gathered from databases, log files, or data streamed directly from a sensor or system. These datasets were described to be rather large and high dimensional. The cyber security analysts primarily accessed locally-hosted databases of log files. One challenge faced by all of our analysts we spoke with was how to subset the large quantity of data, so that a smaller amount of it can be worked on locally.

Tools used

Most of the analysts across the four domains used a data analytic tool, like R or Matlab, as well as a variety of visualization tools. A commonality among the analysts across domains was the use of spatialization tools, which are visualizations that cluster like data together based on specific attributes of the data. Spatialization tools such as Starlight [14], IN-SPIRE [15], Clique and Traffic Circle [16], and ARGIS were used by the analysts that we interviewed. These spatialization tools were semi-autonomous processes that allowed manipulations by the analysts to occur in order to perform feature extraction. A myriad of custom scripts were also used by analysts as well as other feature extracting processes, which were used in a more batch/sequential manner in comparison to the spatialization tools focused on exploration.

To perform any of their scripts and queries, it was assumed that analysts utilized workflow management tools, such as Taverna, but that did not prove to be true among the majority of analysts (only two stated they use such tools). For example, P11 stated that Taverna "provided too many options I do not need." This analyst preferred to maintain a detailed record of inputs and outputs of scripts and queries to recall the process and execute the sequence of scripts for the task. P9 detailed that the purpose of using the Galaxy workflow management tools was to "replay workflows" as well as "share and collaborate on workflows", something that was not desired and thus added un-needed overhead to using the tool. Other participants detailed natural progressions in their work, which may have been complimented by a workflow management tools. While they commented that these sequences of scripts could be described as a workflow, they are created in-situ, and thus they prefer the flexibility of performing these computational steps as scripts, often executed in standard command line interfaces. These findings were particularly interesting with regards to our third research questions and hypothesis (R3 and H3), in that workflow management software, in the computational sense, was not used by our participants.

Analytic Provenance Recording and Documentation.

Upon coming across phenomena of interest, analysts look to capture/document it in some way in order to recall it later. These moments are captured in two ways: the process and the finding. P11 captures the process by using a "configuration file to specify what to run", so that it can be repeated when needed. P2, an analyst that performed most work manually, filed pages (printed hard copies) to capture findings. P3 stated that findings are best captured and described through publications. From the interviews, it was seen that there is an importance set on documenting and filing phe-

nomenon for sharing and revisiting findings. P15 stated that he and any collaborators kept an online journal that was used to keep track of important findings throughout their processes for easy reference later on and to provide an opportunity for peer reviewing. In general, we observed that the provenance or process of our analysts was maintained primarily via inputs and outputs of data to and from scripts.

4.2 Comparisons to Existing Models

Our hypothesis (H1) for comparing the observed processes with both sensemaking and signature discovery is that signature discovery would be more structured (even, sequential), compared to a more fluid sensemaking process. Below, we describe our findings in more detail.

Comparison to the Signature Discovery Process

Figure 1 describes the signature discovery process presented by Baker et al. [4]. Through the interviews, we found that analysts did not follow the linear design of this model. One noticeable difference between the analytic process described by participants and the model of the signature discovery process was the exclusion of certain steps that are present within the signature discovery model. For example, the majority of analysts interviewed did not take inventory of observables; typically the step where data intended for use is defined. In the case of most analysts, the data provided to them for analysis did not require a definition or schema to define the atomic units in the data. Another difference was the lack of assessment and exploration of data, mainly seen amongst the intelligence analysts interviewed. P11 stated that there were "no feature detection models" needed, as the data was structured upon reception by the analyst. P2 and P12 noted how the feature extraction process was not computational and expressed the desire to have "entity extraction in the future." We also saw that there was a lack of explicit description of assessments, detail in the discussion section.

The participants also emphasized the cyclic nature of the analytic process. P5 described the process as the "refinement of a filter to understand the tradeoffs in the data". This cyclic process follows the loop between signature suitability and extracting features within the signature discovery process.

Within the Biology domain, we saw that all analysts performed most of the data-intensive tasks within the signature discovery process, mainly feature extraction, and measuring, and collecting data. We found that the Cyber and Data Science analysts did not perform feature extraction for the most part because the data provided to them was typically already structured and mostly quantitative data, so their main concern was to gain insight from what was provided. This is a similar case for the Intelligence analysts. However, Intelligence analysts differ from the other domains in that they identified observables as a critical part in their process. Analysts in all domains came to an endpoint or what could be called a signature, based on the data provided to them. There were a variety of endpoints that came about, such as patterns to be compared to databases, feature sets, and points of interest.

There are two forms of outcomes described by the analysts we interviewed that can be described in terms of the signature discovery process. One is a signature, which is a unique or distinguishing measurement, pattern, or collection of data that detects, characterizes, or predicts a target phenomenon of interest. In the case of the Cyber domain, this echoes prior work by Fink et al. that found the final produce of cyber security analysts to be a complex query [10]. Another form of result is a signature system, which describes a sequence of steps taken to attain the insight. For example, P6 stated how the findings attained through the use of scripts, filters, and other data transformations are "not valuable without remembering how all of these steps fit together".

Comparison to the Sensemaking Loop

There are differences based on the sensemaking loop presented by Pirolli and Card [5], and the processes observed. One difference that was persistent throughout most of the participants was the lack of top-down reassessment steps performed.

Another component of the sensemaking process that was not present in the analytic process of most of our analysts was the steps prior to obtaining a shoebox of evidence. Most analysts were provided scoped, domain-specific data so there was no need for searching and filtering external, unrelated information. Most analysts did work to approach a presentation as a deliverable in order to find if the analytic process should be continued or if the final result had been sufficiently reached. Kang and Stasko's findings are similar to these findings, in that they identified different categorizations of analyses: structured and intuitive [17].

For the Cyber and Data science domain, we found the synthesis stages often lacking. In the Cyber domain, based on their described analytic techniques, there is less information synthesis. Their analytic processes were more focused on foraging. P8 commented that the "findings of our queries are sent off and evaluated" by other collaborators. Similarly, the participants from the data science domain commented that they were sometimes required to simply forage through data. That is, to transform data from one (or a set of) structured or unstructured sources to a single, clean, unified spreadsheet or database.

The biology and intelligence analysts, on the other hand, performed more of their work in the synthesis part of the sensemaking loop, specifically schematizing. The biologist exhibited this in the form of generating research questions and hypotheses. The intelligence analysts often created hypothetical scenarios and observe how they panned out with respect to the current and new data.

4.3 Products and Results of the Analytic Process

We analyzed the data collected from the interviews and observations to determine what constituted as "products" from the analyses of our participants. These are primarily the *insights* our participants found, and also the *repeatable processes* that allowed them to acquire the insights. The details below help address our second hypothesis (H2), regarding the outcomes and products.

Insights

A majority of the analysts consider a valuable outcome of their process to be broadly categorized as novel *insight*. P10 stated that a goal is to "represent data in a meaningful way" and that this is often through the detection of a "pattern found in the data." P8 stated that a valuable finding was the "unique identification for features," which can be considered a derived facet from the original data to gain insight. P4 stated that findings typically come in the form of behavior of extracted features (biomarkers). P9 defined a the valuable insights as a "piece of info that stands out from background" and "that gives you meaning."

We found that the insights of the analysts in our study were also focused on features or characteristics in the data, rather than the data objects themselves. For example, P3 states that characteristics of "a panel of genes"', such as "how it behaves", "a set of features/markers", and "abundance level[s]", are valuable results of analysis. This form of an analytic product maps well onto the prior work of defining what a signature is in different contexts. For example, signatures have been described as nucleotide sequences [18][19], features of network communication packets [20], or a set of proteins and their measured abundance levels [21].

Repeatable Processes

Analysts also described their findings in terms of the process (rather than the outcome). P5 stated that findings are typically a "sequence of up-front data processing, followed by decision making unique to the data" to come about a set of answers. These steps and decisions made are the critical findings, as they help inform collaborators of how the findings were generated. P5 also described the process as "hypothesis generation, not hypothesis answering." P11 stated that 75 percent of the process was processing data through mass spectrometry and matching it to a database, and "the other 25 percent is generating p-values".

4.4 Collaboration

The analysts exhibited variations of informal collaboration during their analysis. P6 often performs "pair analytics" [22], in which P6 and a domain expert work alongside an expert in the area they're performing analysis for. P3 leverages expert knowledge in combination with "multiple sensor sources to find patterns" with the goal of providing "trained models to get experts further along … to make better distinctions in features." However, there was no explicit collaboration between analysts observed or reported (with the exception of handing off results and steps in a sequential manner to collaborators). For example, the "identify and select potential observables" step of the signature discovery process was not performed because another analyst had previously completed this step and communicated the resulting data to the next analyst.

5 Discussion

5.1 Sensemaking and Signature Discovery

The entirety of the analytic process can be said to be composed of aspects of both sensemaking and signature discovery. Figure 3 shows a summary of how the analytic processes of analysts interviewed relate to the signature discovery and sensemaking processes based on our data analysis. Full circles indicate steps explicitly mentioned or observed. Half circles indicate steps that were generally mentioned, but not explicitly performed. Empty circles indicate steps that were not observed or mentioned during the interviews. As can be seen by these figures, it is rare for individual analysts to conduct every step explicitly.

The signature discovery process is computationally based procedure. From the interviews, it was shown that data processing was a necessary part of the analytic process. The goal of signature discovery is quantitative and well defined, where a signature and signature system are the ideal results of the process. In contrast, sensemaking is primarily a cognitive process and is not as rigidly structured as the signature discovery process. There are more avenues for assessing previously performed tasks in the process compared to signature discovery. As a result, we saw similarities with the sensemaking process in the more synthesis-focused analysis tasks, and more similarities to signature discovery when data-centric methods were used.

Further, while specific steps in the foraging loop map directly to some of the steps in the signature discovery process, a link between the synthesis stages from sensemaking to signature discovery are not as apparent. For example, the task of transitioning from external data sources to an evidence file maps to the steps in the signature discovery process steps of taking inventory of observables to feature extraction. From there, the two processes appear to branch off into independently.

As a result, we found that both models do not encompass the process of analysis across all of the domains fully. The steps they performed were mostly based on the data sources and tasks given. Signature discovery and sensemaking have been identified as independent processes, focused on the computational and cognitive aspects of data analysis, respectively. The differences are also illuminated in observing the result of the models. Signature discovery has the signature as a result, while sensemaking suggest a presentation to share or present findings, processes, and knowledge to someone else. From our study, we find that the distinction between these two is dependent primarily on the domain. In Cyber Security, for example, a repeatable query or classifier is often desired, whereas intelligence analysis emphasizes the shared knowledge (i.e., the intelligence being discovered) to give to someone else.

5.2 Hypotheses and Assessments

The formation and assessment of hypotheses played a central role in the analytic processes of our participants. The position of the hypothesis in the analytic process can vary, depending on the initial starting point of the analysis. Proving or disproving a hypothesis can be the driving force behind the analytic process, or developing a

hypothesis can be the purpose of the analysis. It was found that some of the analysts' processes were based on either hypothesis generation or hypothesis verification that has an impact on how the analytic process is conducted. P1 stated the measure of success for their work was "hypothesis confirmation and verifying a prediction," which correlates to the scientific method. P5 stated that analysis was about "hypothesis generation" and P2 stated that the goal for their work is a hypothesis. Among the data scientists interviewed, the placement of hypothesis formation within their analytic processes significantly varied depending on the question that needed to be asked of the data and the type data being dealt with. Typically, when dealing with cyber data, the Data analysts had a hypothesis near the earlier steps of their process with an end goal of verifying their hypothesis. Hypothesis generation can also arise when there is no specific initial problem. In this case, the analytic process is used to develop more refined problems as opposed to solving a succinct problem or question.

Our participants also exhibited tactics of hypothesis assessment. As seen in Figure 3, many analysts do not explicitly state that they conduct assessments of suitability in signature discovery, or any of the top-down analytical steps within the sensemaking loop. However, during the observation of analysis, it appears these steps are inherent to the analytic process that analysts do not consider them to independently coincide with the signature discovery or sensemaking process. This is reflected in Pirolli and Card's "dual search" activity in the sensemaking model [5], and may evidence the importance of that aspect of the sensemaking process. Similarly, the analytic processes observed can be generally fit either the "top-down" or "bottom-up" terminology [2], [5], as hypothesis-driven analysis, or the process of forming hypotheses.

5.3 Implications for Design

The results of this study can help inform the design and implementation of systems created to support data analysis. Through gaining a deeper understanding of the cognitive and computational processes involved in gaining insights in data-driven domains, we can develop design guidelines and considerations.

There is an inherent importance of **capturing and tracking the analytic provenance** [23]. Analytic provenance encompasses the interactive steps, intermediate results, and hypotheses considered throughout analysis. The participants of our study commented that recalling their process after a focused data analysis session is difficult. Analytic provenance support in such tools needs to strike a fine balance between allowing users to maintain this "cognitive zone" [24] during analysis, and encouraging users to record and annotate their process. Our participants commented that such provenance support should not interfere with the actual data analysis process. Thus, there is an open challenge for the field to determine how formal and explicit such capturing and tracking techniques should be. In comparison, passive capture and interpretation of user interaction may be a more well-suited approach [25], [26].

Further, the analysts from the domains we studied emphasized the **importance of hypotheses during their analytic processes**. This included generation, testing, and validation of hypotheses. However, hypotheses were not typically formal (or mathematically grounded). For example, they came in the form of "hunches" or "moments

of interest" for following one's curiosity about the data. This also came through in the observations, where the investigators often asked how the analyst decided to try one approach over another. Often, the analysts stated that they did not know, or that they had seen something like it before. This finding echoes the importance of researching and developing more formal methods of tracking hypotheses, but also of evaluating exploratory data analysis tools.

6 Conclusion

In this paper, we present the results of a user study of professional analysis in four domains (Biology, Cyber Security, Intelligence Analysis, and Data Science). We analyzed the tasks and processes of analysts to get a deeper understanding of the cognitive and computational aspects of data analysis. In order to do this, we conducted a series of semi-structured interviews and analysis observations.

Our results indicate that the data analysis processes exhibited by our study participants was a combination of the known sensemaking and signature discovery models. Generally, the cognitively-focused aspects of synthesizing the information into knowledge was represented in the sensemaking models. In contrast, the computationally-focused data foraging and features extraction stages were better represented in the signature discovery model. The diversity of analytic processes, among those we interviewed suggests a combination of sensemaking and signature discovery as a viable model for data-driven analytic processes. These findings are further discussed, including implications for design, which can inform the design and implementation of future data analysis systems.

7 Acknowledgements

This research is part of the Signature Discovery Initiative at Pacific Northwest National Laboratory, conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

8 References

[1] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card, "The cost structure of sensemaking," in *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, New York, NY, USA, 1993, pp. 269–276.

[2] P. Pirolli and S. Card, "Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Though Cognitive Task Analysis," *Proc. 2005 Int. Conf. Intell. Anal. McLean Va.*, p. 6, 2005.

[3] G. Klein, B. Moon, and R. R. Hoffman, "Making Sense of Sensemaking 2: A Macrocognitive Model," *IEEE Intell. Syst.*, vol. 21, no. 5, pp. 88–92, 2006.

[4] N. Baker, J. Barr, G. Bonheyo, C. Joslyn, K. Krishnaswam, M. Oxley, R. Quadrel, L. Sego, M. Tardiff, and A. Wynee, "Research towards a systematic signature discovery process," in *IEEE Intelligence and Security Informatics Signature Discovery Workshop, 2013*.

[5] P. Pirolli and S. Card, "Information foraging in information access environments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1995, pp. 51–58.

[6] M. Pohl, M. Smuc, and E. Mayr, "The User Puzzle: Explaining the Interaction with Visual Analytics Systems," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2908–2916, 2012.

[7] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, no. 17, pp. 3045–3054, Nov. 2004.

[8] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "VisTrails: visualization meets data management," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, New York, NY, USA, 2006, pp. 745–747.

[9] A. Endert, C. Andrews, G. A. Fink, and C. North, "Professional Analysts using a Large, High-Resolution Display," presented at the IEEE VAST Extended Abstract, 2009.

[10] G. Fink, C. North, A. Endert, and S. Rose, "Visualizing Cyber Security: Usable Workspaces," *VizSec*, 2009.

[11] C. Andrews, A. Endert, and C. North, "Space to think: large high-resolution displays for sensemaking," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 55–64.

[12] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, "Enterprise Data Analysis and Visualization: An Interview Study," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2917–2926, 2012.

[13] J. Scholtz, C. Plaisant, M. Whiting, and G. Grinstein, "Evaluation of visual analytics environments: The road to the Visual Analytics Science and Technology challenge evaluation methodology," *Inf. Vis.*, Jun. 2013.

[14] J. S. Risch, D. B. Rex, S. T. Dowson, T. B. Walters, R. A. May, and B. D. Moon, "The STARLIGHT information visualization system," in *Proceedings of the IEEE Conference on Information Visualisation*, Washington, DC, USA, 1997, p. 42–.

[15] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: spatial analysis and interaction with information for text documents," presented at the Readings in information visualization: using vision to think, 300791, 1999, pp. 442–450.

[16] K. Abdullah, C. P. Lee, G. Conti, and J. A. Copeland, "Visualizing network data for intrusion detection," in *Information Assurance Workshop, 2005. IAW '05. Proceedings from the Sixth Annual IEEE SMC*, 2005, pp. 100–108.

[17] Y. Kang and J. Stasko, “Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study,” 2011, pp. 21–30.

[18] J. Vieira, M. V. Mendes, P. Albuquerque, P. Moradas-Ferreira, and F. Tavares, “A novel approach for the identification of bacterial taxa-specific molecular markers,” *Lett. Appl. Microbiol.*, vol. 44, no. 5, pp. 506–512, May 2007.

[19] A. M. Phillippy, J. A. Mason, K. Ayanbule, D. D. Sommer, E. Taviani, A. Huq, R. R. Colwell, I. T. Knight, and S. L. Salzberg, “Comprehensive DNA Signature Discovery and Validation,” *PLoS Comput Biol*, vol. 3, no. 5, p. e98, May 2007.

[20] H. Han, X. L. Lu, J. Lu, C. Bo, and R. L. Yong, “Data mining aided signature discovery in network-based intrusion detection system,” *SIGOPS Oper Syst Rev*, vol. 36, no. 4, pp. 7–13, Oct. 2002.

[21] C. Bock, M. Coleman, B. Collins, J. Davis, G. Foulds, L. Gold, C. Greef, J. Heil, J. S. Heilig, B. Hicke, M. Nelson Hurst, G. M. Husar, D. Miller, R. Ostroff, H. Petach, D. Schneider, B. Vant-Hull, S. Waugh, A. Weiss, S. K. Wilcox, and D. Zichi, “Photoaptamer arrays applied to multiplexed proteomic analysis,” *PROTEOMICS*, vol. 4, no. 3, pp. 609–618, 2004.

[22] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher, “Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics,” presented at the Hawaii International Conference on System Sciences, 2011, vol. 0, pp. 1–10.

[23] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, and G. Fink, “Analytic provenance: process+ interaction+ insight,” in *CHI’11 Extended Abstracts on Human Factors in Computing Systems*, 2011, pp. 33–36.

[24] T. M. Green, W. Ribarsky, and B. Fisher, “Building and applying a human cognition model for visual analytics,” *Inf. Vis.*, vol. 8, pp. 1–13, 2009.

[25] A. Endert, “Semantic Interaction for Visual Analytics: Toward Coupling Cognition and Computation,” *IEEE Comput. Graph. Appl.*, vol. 34, no. 4, pp. 8–15, Jul. 2014.

[26] E. T. Brown, A. Ottley, J. Zhao, Q. Lin, A. Endert, R. Souvenir, and R. Chang, “Finding Waldo: Learning about Users from their Interactions,” *Trans Vis Comput Graph*, 2014.