Toward Theoretical Techniques for Measuring the Use of Human Effort in Visual Analytic Systems

R. Jordan Crouser, *Member, IEEE*, Lyndsey Franklin, Alex Endert and Kris Cook

Abstract—Visual analytic systems have long relied on user studies and standard datasets to demonstrate advances to the state of the art, as well as to illustrate the efficiency of solutions to domain-specific challenges. This approach has enabled some important comparisons between systems, but unfortunately the narrow scope required to facilitate these comparisons has prevented many of these lessons from being generalized to new areas. At the same time, advanced visual analytic systems have made increasing use of human-machine collaboration to solve problems not tractable by machine computation alone. To continue to make progress in modeling user tasks in these hybrid visual analytic systems, we must strive to gain insight into what makes certain tasks more complex than others. This will require the development of mechanisms for describing the balance to be struck between machine and human strengths with respect to analytical tasks and workload. In this paper, we argue for the necessity of theoretical tools for reasoning about such balance in visual analytic systems and demonstrate the utility of the *Human Oracle Model* for this purpose in the context of sensemaking in visual analytics. Additionally, we make use of the *Human Oracle Model* to guide the development of a new system through a case study in the domain of cybersecurity.

Index Terms—Theoretical models, human oracle, visual analytics, mixed initiative systems, semantic interaction, sensemaking.

1 Introduction

In an age of increasingly complex data, the dynamic interplay between human and machine analysis grows ever more important. By pairing the human analyst with a machine collaborator, we hope to overcome some of the human's limitations such as working memory, bias, and fatigue. Such human-machine collaborative systems rely on the intuition that the domain expertise and perceptual advantage of the human analyst may provide as a critical boost in areas where purely computational analyses fail. We can think of such collaborative systems as distributed cognitive tasks, where information is both internal (the human's mental model) and external (stored explicitly by the machine) [82]. This human-machine hybrid approach has been critical in the support of sensmaking [20, 61] where visual analytic environments are often called on to provide the medium of interaction between human and machine by combining intuitive, interactive interfaces with a strong computational backbone. The result has been a diverse array of contributions across a broad range of topical domains in an effort to advance the art of visual analytics.

The development of tools for streaming analysis has to date been concerned almost entirely with questions of *tractability*. We build systems that capitalize on things we believe humans do well, such as recognizing patterns, in hopes that human input will enable us to make progress in areas where purely computational methods fail. Despite the staggering volume of applications in both the literature and real-world, how do we tell if a new problem would benefit from a similar strategy – and if so, how do we balance the computational workload? While traditional user studies can often help us determine whether our system was useful in solving a particular problem, they fall short of explaining *why* we see the results we do. This makes it challenging to justify the claims we make about a novel system's performance relative to other systems. In addition, this makes it extremely difficult to generalize what we learn from solving one problem to others we might want to solve in the future. For example, it is difficult to explain why established strategies for building visual analytic systems on static data may fall apart when applied to streaming data, even when the systems were designed for use on similar tasks in similar domains [7, 24, 33].

But what if we could describe the computational power of human-computer collaborative systems the same way we describe the complexity of a traditional algorithm? The ability to model the computational resources afforded by humans in a similar fashion to machine resources would begin to unify findings across the research efforts of visual analytics regardless of the specific application domain or data format. To help us do this, we can draw on the language of complexity theory, which takes the study of solvable instances of problems to a deeper level by asking questions that get at the fundamental nature of the problem itself and how we might go about solving it more effectively: *Does randomization help? Can the process be sped up using parallelism? Are approximations easier?* By understanding how systems make use of various **resources**, we can begin to group algorithms and problems according to what it takes to solve them. This abstraction also enables us to investigate the effect of limiting these resources on the classes of problems that can still be solved, which is of extreme importance in the context of streaming analysis.

The remainder of this paper is organized as follows: we begin by reviewing some preliminary techniques for describing analytic tasks in visual analytic systems (Section 2). Next, we argue for the necessity of theoretical tools for reasoning about visual analytic systems (Section 3). We then demonstrate the utility of one simple tool, the *Human Oracle Model*, for evaluating the balance of human and machine work in the context of streaming visual analytics (Sections 4, 5, and 6), as well as for designing new systems through a case study of several examples of streaming data analysis (Section 7). For convenience, we have chosen a single domain (cybersecurity) for exploring the utility of these theoretical models for describing human+machine analysis; in Section 8, we briefly highlight the transferability of these models to other domains. Finally, we discuss the limitations of this model, and propose possible extensions for future study.

2 Background

Substantial work has been done to further our understanding of how analysts develop their expertise, and how this expertise may be exploited by a system in order to perform more nuanced analysis. Additionally, there have been a wealth of successful explorations into computational methods that can assist analysts in performing at their highest levels. Our work is inspired by these previous efforts to leverage human expertise in computation, and in the following sections we will highlight some particularly influential work in these areas.

- *R. Jordan Crouser is at Smith College. Email: jcrouser@smith.edu.*
- *Lyndsey Franklin and Kris Cook are at Pacific Northwest National Laboratory. Email: lyndsey.franklin@pnnl.gov, kris.cook@pnnl.gov.*
- *Alex Endert is at Georgia Tech. Email: endert@gatech.edu.*

Manuscript received 31 Mar. 2016; accepted 1 Aug. 2016. Date of publication 15 Aug. 2016; date of current version 23 Oct. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2598460

2.1 Cognitive Models of User Actions

Models of human cognition have been proposed and studied for decades and have formed the basis of what we know as human-centered design. Early work in the development of the GOMS model [14] and its variations have been applied to describe how people achieve goals through their selections between alternative methods. While these advances provided a critical foundation for understanding human decision-making, a more complete model a human user's task must include not only what their goal is, but also the conditions under which they are trying to meet that goal. This includes the information available to the human, the actions they might take to effect change, what they see and interact with, and even what they might do wrong [42]. Critical design decisions about the trade-offs between human and machine processing must be based on this task analysis if an interactive system has any chance to succeed [42]. Improvements to conditions through such task analysis is often measured in performance duration [26]. That is, if a person can achieve their goals *faster* than they could before, improvements are considered to have been made. While this may be a useful measure under some conditions, we suggest that this may be an overly narrow definition of improvement. To provide us with more nuanced understanding of success, additional measures such as *task complexity* can be formally defined such that they can be evaluated for improvement as well.

Since the development of the GOMS model, there have been many other notable efforts to describe and predict the performance of interface designs. GLEAN [45] offers a computer-based tool for generating quantitative predictions of usability and procedural aspects of an interface design based on a supplied GOMS model. The EPIC Architecture [43, 44, 46] models human multi-modal and multi-task performance, and includes important factors such as sensory-motor processes. The UFuRT framework [81] underscores user analysis as the first phase of design, and emphasizes the importance of accounting for differences in cognitive capacities, limitations, and perceptual variations. UFuRT also emphasizes representational analysis based on the phenomenon of representational effect [82], where different representations of abstract structures can dramatically affect efficiencies, task difficulties, and behavioral outcomes. The visual analytic community has also made attempts to establish cognitive models of analysis, though these efforts are focused primarily on justifying the appropriateness of visualization design choices [38].

2.2 Task Taxonomies for Sensemaking

While human expertise and experience are what enable us to make sense of the information around us, describing the process by which sensemaking happens has not been easy. Early work by Pirolli and Card [61] proposed the prototype to the canonical "sensemaking loop," which has become a central theme in the study of human sensemaking. More recent taxonomies, such as work by Zhang and Soergal [84, 85] and Kang and Stasko [41], extend this work by examining sensemaking *subtasks* in order to form a more detailed picture of the sensemaking process. These prior works establish several common subtasks, such as *task planning and analysis, gap identification, search, building structure, instantiating structure*, and *creating products* [84].

In many cases, there is a somewhat predictable flow between these subtasks: *information gathering* leads to *evidence identification*, which in turn leads to *confirming or refuting a hypothesis*. While this particular pattern is supported by many independent researchers [41, 61, 84, 85], other sensemaking subtasks are not supported equally in the literature. As Kang and Stasko observe in [41], gaps exist in keeping cohesive, context-preserving mental models synchronized between tools. This is further complicated in the case of collaboration between groups of analysts, which at the time of this writing is largely ad hoc and lacks substantial computational support.

2.3 Leveraging Human Expertise in Computation

The lived experience and tacit knowledge developed over a lifetime is often crucial to being able to solve real-world problems. At the same time, this supplemental information about the larger domain can prove difficult, if not impossible, to encode into a mechanical computation system. Because of this, it is sometimes advantageous to leverage the human's expertise directly rather than invest significant resources in approximating it. For example, in the field of machine learning, this expertise is often used to generate labeled training datasets from which an algorithm may learn an appropriate feature set. These methods have proven highly effective in handwriting recognition [79], classifying text documents [68], learning realistic human motion from video [49], and other areas where predetermining a clear set of classification rules is intractable [86]. Similarly, visual analytics systems rely on human expertise and experience to identify patterns in data that elude purely mechanical detection. These systems have met great success in analyzing medical imagery [9], detecting financial fraud [18], diagnosing network faults [52], and many other applications.

Online marketplaces providing an on-demand workforce for microtasks has resulted in an explosion of systems that apply distributed human processing power to problems previously thought to be intractable. Examples include image labeling [72], optical character recognition [73], and annotating audio clips [50]. Human computation has also been used to develop logical models of mutual exclusion [19], as well as to identify cases where a predictive model is confident but incorrect [5]. Perhaps most intuitively, human computation has also shown great promise in helping refine models of human behavior [8] and natural language [15]. As illustrated by these examples, the term *human computation* spans a wide range of possible applications and distributions of computational workload. Among these, many of the most interesting and successful systems not only leverage the complementary computational strengths of both parties [21], but also make efficient use of both human and machine computational resources.

2.4 Mixed Initiative Systems

Mixed initiative systems have been established as the technical implementation of the balance between human and machine effort. The defining trait of any mixed-initiative system is its ability to act on behalf of the user while at the same time respecting user control [40]. A natural addition to these systems is semantic interaction [29] which attempts to enable co-reasoning between a user and the analytic models of a system without requiring the user to directly control them [30]. This cohesion of user and machine initiatives strengthens the analytic discourse which in turn produces higher quality results. One exemplar system is the Active Data Environment [20] in which system recommendations appear in response to the interactions with a visual canvas metaphor a user has during the course of their analysis. Recommendations, which can provide new data or relationships from a variety of sources and types, appear in context of ongoing analysis and can be encouraged by "pinning" them to parts of the existing canvas or suppressed allowing a user direct control without taking them away from their analysis.

3 Why Study Complexity Measures for VA?

In other areas of the computational sciences, theoretical arguments paved the way for the designs that made provably correct solutions tractable. In contrast, the development of real-world implementations has far outpaced the development of theoretical measures for streaming analytics. Many of these implementations have demonstrated unparalleled success at previously intractable problems. However, in the absence of a rigorous theory in which to ground the development of new algorithms, researchers must rely on intuition and some deeply-rooted assumptions about the differences between human and machine computation in order to design new systems. At a low level, there is significant interest in establishing concrete lower bounds on the complexity of computational problems. That is, what is the *minimum amount of work* that must be done in order to guarantee the solution is correct? At a higher level, complexity theory also explores the connections between different computational problems and processes. This kind of analysis can yield fruitful comparisons that deepen our understanding of the nature of a problem space through canonical or *complete* problems, even if we can't make absolute statements regarding the individual problem instances.

In traditional computational complexity theory, we describe and classify problems in terms of the resources required to solve them. One way to measure a problem's difficulty is with respect to time; we may ask *how long will it take to find an answer*? Alternatively, one might want to measure difficulty in terms of space; here we could ask *how much memory will I need to execute this process*? These questions, which do not rely on the specific details of the implementation, are at the heart of computer science. Theoretical arguments ground our intuitions about the problem space, and pave the way for us to design future systems that make these provably correct solutions a reality. In addition to providing a more precise language for describing the systems we've built, comprehensive models can also help us to reason more effectively about their performance during the design process. As such, theoretical models that capture the complementary roles played by human and machine are an important next step in advancing the science of mixed-initiative systems.

3.1 Preliminary Theoretical Models

Existing theoretical models for systems involving human computation have focused on system-level categorization [63, 64], modeling the social dynamics inherent in human computation systems [16, 80], or optimizing workflows using human computation [23, 77]. In parallel, the machine learning community has developed several models for resource allocation in labeling training data [28, 74], though these models assume a known cost function, which is not generally available in practical applications of human processing power. One major roadblock to developing more sophisticated theoretical models is that our ability to model how the human brain computes is hindered by a limited understanding of the biological mechanisms driving that computation. Until our understanding of the human brain is more fully developed, it seems likely that humans will remain (somewhat finicky) black boxes in the larger system diagram. In the interim, we can begin by characterizing and quantifying the **use** of human processing power as part of an algorithmic process, and later refine these measures once we've developed the tools necessary to directly measure the cost of human computational processes.

Early work in this vein suggests that we might be able to make some progress by modeling human contributions to an algorithm as queries to a Human Oracle [22, 69] – an Oracle with human-level intelligence. In this model, the Human Oracle is able to answer questions to which a human would be able to respond, even if a machine could not. In this work, the authors demonstrate that much of the standard theoretical language holds true when extended to include Human Oracles, including concepts such as *algorithmic complexity*, *problem complexity*, and *complexity classes*. This indicates a complementarity between human and machine contribution the algorithms under study, not unlike the relationship alluded to in previous task taxonomies for mixed-initiative systems [21]. Such complementarity also suggests that the complexity of system involving both human and machine computation could be represented as a pair $\langle \Phi_H, \Phi_M \rangle$, where Φ_H indicates the query complexity (# of questions posed to the Human Oracle) as a function of the input, and Φ_M is the the standard computational complexity of the operations performed by the machine. The minimal complexity of a problem can then be described as the minimization of human and machine cost over all correct algorithms.

4 Human and Machine Effort in Cyber VA Systems

To provide a foundation for later algorithmic exploration, we begin by characterizing the utilization of human and machine effort in existing systems documented in the cybersecurity visualization literature[1]. To begin, we surveyed all 71 papers published in the proceedings of the IEEE Conference on Visualization for Cybersecurity (VizSec) during the years 2009-2015. From this, we identified a subset of 45 papers whose primary contribution was a *system*. This survey identified distinct groups into which the majority of these systems could be classified, with class assignment confirmed by inter-rater agreement. We label these classes **monitor** and **triage**, corresponding to the two high-level patterns of utilizing human computational power that characterize and distinguish each class. We present the results of this characterization here, and formally describe the complexity of these two classes with respect to the *Human Oracle Model* described in 3.1 in the sections to follow.

[1]While these papers represent a convenience sample of works relevant to our future efforts in streaming visual analytics in the context of cybersecurity, the domain-agnostic nature of the theoretical models proposed in this paper analysis suggests that the same analysis could have been performed with similar results had we selected a different domain. We will illustrate the application of these same models to other domains in Section 8.

4.1 Monitor Prototypes

Systems that support *monitoring activities* represent a relatively simplistic level of human-machine collaboration. The human has a single task, which is to either confirm or reject machine recommendations. These machine recommendations may take many forms, but generally appear as highlighted, prioritized, or other visually-distinguished data points. For instance, [54] uses a trust model to summarize signals from sensors to indicate when grid controllers should be suspicious of incoming data. In [65], learned alerts are presented as either highlighted table entries or as part of an interactive incident diagram. The analyst might respond to these distinguished data by querying for more detail or isolating the affected points in order to confirm the machine's recommendation (as in [1, 36, 55, 70, 78, 87]). Some advanced systems take this confirmation as implicit input to refine future recommendations [2, 13, 51, 59, 67], completing the loop of human-machine collaboration and providing explicit options for steering the visualization.

4.2 Triage Prototypes

Systems that support *triage activities* represent a more sophisticated level of human-machine collaboration. Human operators in these systems are additionally responsible for taking action on valid machine recommendations. In [37], users are given multiple pathways to block system resources from attack and must use their judgement on which to utilize. [4] also supports follow-up quarantine of network resources in response to visual alerts. These follow-on steps require the human to prioritize the results of recommendations or directly refine the recommendations themselves. This may include selecting additional data that the machine should have included in recommendations, or deselecting data which does not fit the analyst's mental map of the world. As with some monitoring systems, advanced triage systems may leverage these refinements to modify future recommendations. For instance, [76] asks human users to co-map data along with its advanced statistical analysis to correctly categorize rating fraud in an iterative process. [39] employs analyst-driven clustering of network traffic patterns to direct alerts. Triage prototypes that include the machine updating its recommendations based on user actions begin to approach the sophistication of mixed-initiative systems, especially when the human is permitted to directly declare types of desired future recommendations that the machine has not yet generated on its own.

5 Characterizing Monitor Systems with the Human Oracle Model

We now employ the *Human Oracle Model* to formally describe these classes. This represents a first attempt to generalize and extend the model to shed light on human contributions in visual analytic contexts. For example, a first-line cyber defender might be asked to monitor network activity for a particular attack signature [7]. This interchange, where an external entity is contracted to perform some challenging subroutine, is exactly the type of system that oracle machines were built to describe [22]. In this case, the "oracle" happens to be the human analyst, who may draw on her prior experience to recognize the signature even if she is unable to articulate the cues she uses.

5.1 Monitoring Example: Human Only

We begin by constructing a simple example designed to emulate the behavior of an unaided human analyst trying to perform the task of monitoring a stream of data to determine if a signal of interest is present (Fig. 1). This resembles many vigilance and signal detection

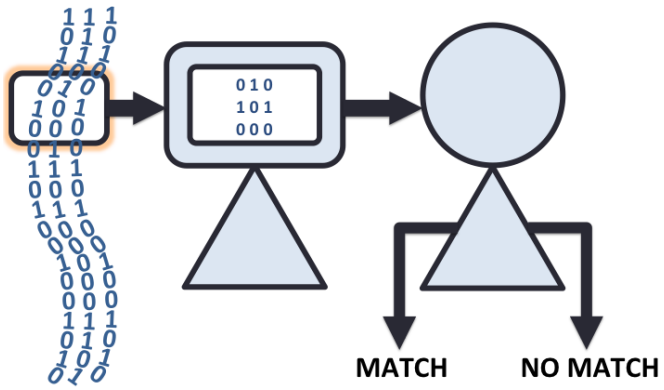

Fig. 1. Illustration of a simple stream monitoring system using a Human Oracle. In this system, the machine performs no function other than sampling the stream (left), passing each sample directly through to the Human Oracle (right) which determines whether or not there is a match.

tasks that have been present in real analytic workflows for decades (for instance, see [71]). Particularly within the realm of vigilance, it is well understood that degradation in performance happens over time [75]. Thus, while this example might appear overly simple, we suggest that the characterization of unaided monitoring tasks is an important baseline to which we will later compare machine-supported analysis.

In this system, streaming data is sampled by the machine as it passes through some data capture mechanism, and each sample is forwarded directly to the Human Oracle for processing. The Human Oracle then performs a single binary operation, determining whether or not the signal is a match. Let us assume for the time being that each of these operations has some **fixed cost** C_{MATCH}. We then imagine that we start this "system" up, use it to process k consecutive samples, and then power it back down. In pseudocode, we could write:

Algorithm 1: Human-Only Approach

```
Let H_match be a Human Oracle that performs MATCH operations;
foreach sample k do
    return H_match(k);
end
```

Using the methodology described in Section 3.1, we can describe the work performed by this algorithm in processing those k samples as $\langle \Phi_H, \Phi_M \rangle = \langle k * C_{MATCH}, 0 \rangle$: the machine passes each sample directly through without performing any computation on it, and the human processes each and every sample at a cost of C_{MATCH}.

5.2 Monitoring Example: Human + Machine

Such a system clearly violates the implicit directive that the use of human computation should be judicious: after all, human effort is expensive (that is, C_{MATCH} is very large with respect to the per-operation machine cost). Additionally, we know that performance suffers over extended periods of boredom induced by repetitive work [60] and eventually there is a point at which human "processors" will simply refuse to perform any more computation. Because of this, we may wish to model the per-query cost as a function that varies over time; we'll discuss this further in Section 9.1. Finally, we have yet to mention the relationship between the rate at which samples are coming in and the rate at which the Human Oracle can process them; if the Human Oracle cannot complete the processing of one sample before the next one arrives, this process will quickly grind to a halt.

Suffice to say that there are many scenarios where requiring a human to conduct an exhaustive search is less than optimal. In such cases, we could imagine a second simple system (Fig. 2) in which the machine performs an initial filtering step on each sample at some cost $C_{FILTER} << C_{MATCH}$. During this filtering step, the machine identifies and screens out any "unambiguous" samples, such as those in which the signal in question is clearly absent. The (now reduced number of) remaining samples are then passed on to the Human Oracle for processing, each at a cost of C_{MATCH}. The pseudocode for running this machine on k samples might look something like this:

Algorithm 2: Human+Machine Approach

```
Let H_match be a Human Oracle that performs MATCH operations;
Let M_screen be a machine that performs SCREEN operations;
foreach sample k do
    if M_screen(k) == false then
        return H_match(k);
end
```

Using the same methodology as before, we can describe the work performed in processing those k samples as $\langle \Phi_H, \Phi_M \rangle = \langle k(1 - E[screened]) * C_{MATCH}, k * C_{FILTER} \rangle$: the machine evaluates each sample at a cost of C_{FILTER}, screening them out at some expected rate $E[screened]$, and the human processes only those that remain.

5.3 Modeling Examples: Comparing Performance

Now that we've characterized these approaches using implementation-agnostic language, comparing between them is straightforward:

	Φ_H	Φ_M
Human Only	$k * C_{MATCH}$	0
Human+Machine	$k(1 - E[screen]) * C_{MATCH}$	$k * C_{FILTER}$

Table 1. Algorithmic comparison of our example Human-Only and Human+Machine approaches to the stream monitoring problem.

Using this framework, we see that our intuition about the role of recommendation systems is nicely articulated: the Human+Machine system trades an increase in (inexpensive) machine effort for a corresponding reduction in (expensive) human effort. Moreover, this reduction is inversely proportional to the expected rate at which recommendations are made.

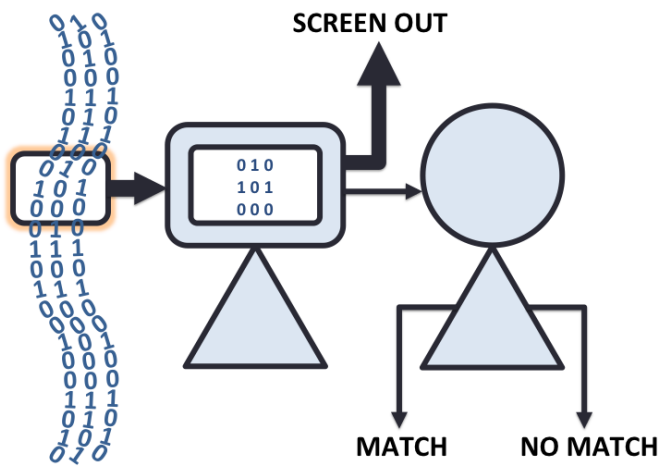

Fig. 2. Illustration of a second stream monitoring system using a Human Oracle. In this system, the machine samples the stream (left), performs an initial filtering step to screen out unambiguous samples (center), and passes each remaining sample through to the Human Oracle (right). Such a system is most effective in scenarios when the number of samples that could be screened out is relatively large.

It is important to be clear that this comparison is facilitated by a rather large assumption: we have assumed that C_{MATCH} (the cost of each H_{MATCH} operation) is **constant**. That is, regardless of the specific instance presented to the Human Oracle, the cost to determine whether or not a match is present is the same. Additionally, this presumes that this cost is also fixed with respect to the number of queries passed to the Human Oracle. While practical experience reminds us that such assumptions are overly naive, they enable us to more clearly illustrate the underlying differences in these two approaches at the algorithmic level. We suggest that drawing parallels at the algorithmic level rather than at the implementation level can enable us to compare systems more effectively than using simple A-B testing. As with other branches of computational science, identifying areas where existing algorithms are redundant or inefficient will enable us to design more efficient algorithms in the future. In addition, reporting bounds on the complexity of the algorithms on which our systems are built along with the observed performance of the system would improve study reproducibility, as well as help isolate the effects of interface design and other implementation details. In Section 9.1, we go into further detail on the implications of these assumptions, and suggest directions for their relaxation in pursuit of more accurate models.

6 Human Oracles with Multiple Operation Types

The simple examples described in the previous sections are useful for illustrating how we might model the use of human processing power analytical systems. However, in most real-world situations the human operator will be required to perform a variety of operations rather than just one. To capture this behavior, we extend our notion of the Human Oracle to accommodate several (though finitely many) unique operation types, each with its own corresponding cost. Equivalently, we could define several independent Human Oracles, each performing a single operation type at a fixed cost. These independent oracles could be thought of as representing the various analytical subtasks a human engages in during the sensemaking process. As highlighted in Section 2.2, sensemaking subtasks have been characterized and documented by many independent research efforts. While demonstrating the completeness of these taxonomies is still an area for ongoing research, we do not attempt to replicate these efforts here. Instead, we focus on illustrating how these subtasks are composed to form larger sensemaking strategies, and how to measure the utilization of various subtasks during the execution of an analytical process.

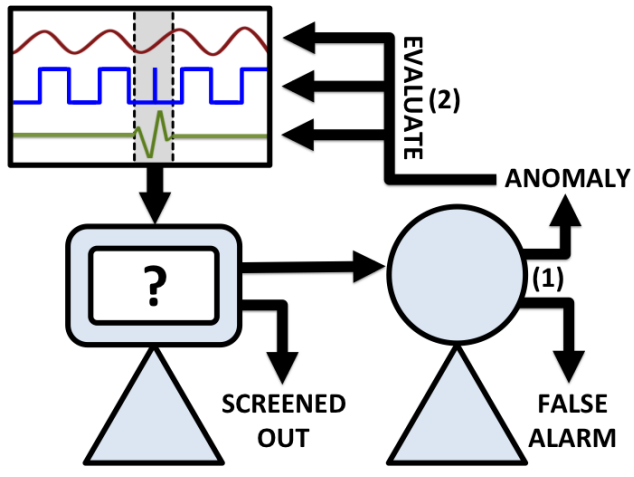

Fig. 3. Illustration of a stream triaging system using a Human Oracle. In this system, the machine samples a collection of streams (left), performs an initial filtering step to screen out unambiguous samples (center), and alerts the Human Oracle of any sample that was not screened out (right). The Human Oracle then confirms whether or not the alert is valid (1), and if so, then evaluates each stream individually to identify the source(s) of the anomalous behavior (2).

6.1 Triage Example

Consider for example an analyst tasked with monitoring and triaging alerts across multiple streams of data that are continuously and simultaneously updating. Due to overwhelming volume, it is not possible to exhaustively monitor all streams at all times. As such, our analyst must leverage computational support to direct her attention when something seems "out of place". However, as with the previous examples, it may be the case that the analyst is unable to articulate precisely what constitutes an anomaly (though we presume she will be able to recognize it when it is presented to her). One the analyst receives an alert she must then apply her domain expertise to determine (1) whether the alert is legitimate or a false alarm, and if it is legitimate (2) scan the data streams to identify the likely cause of the alert and triage the issue to the appropriate mitigation team (see Fig. 3). In pseudocode:

Algorithm 3: Human+Machine Triage

```
Let H_FA be a Human Oracle that identifies FALSE ALARMS;
Let H_TRIAGE be a Human Oracle that performs TRIAGE;
Let M_ALERT be a machine that generates ALERTS;
foreach sample k do
    if M_ALERT(k) == true then
        if H_FA(k) == false then
            foreach stream s in k do
                H_TRIAGE(s)
            end
end
```

We again assume that there is some cost C_{FA} associated with asking the Human Oracle to evaluate an alert. For any alert that is found to be legitimate, the Human Oracle must then triage each of the data streams at some cost C_T for each of the s streams. To determine the complexity of this process, we must also consider the machine's expected alert rate $E[ALERT]$, its expected false alarm rate $E[FA]$, and the cost C_{ALERT} incurred by the machine on each sample to determine whether or not to alert the analyst. Using this information, we can describe the work performed by this algorithm in processing k samples as:

$$\langle \Phi_H, \Phi_M \rangle = \langle k(E[ALERT] * (C_{FA} + (1 - E[FA]) * C_T * s)), k * C_{ALERT} \rangle$$

This captures some of our intuition about the relationship between an analyst's workload and the reliability of an alert system: Φ_H is directly tied to the rate at which the machine generates alerts, and how often those alerts turn out to be false alarms. The relative weighting of the costs C_{FA} and C_T incurred by the Human Oracle can be used to help determine the tolerance to false alarms: if both C_{FA} and $E[FA]$ are relatively high, the system is incurring a large cost without reaping a corresponding benefit.

6.2 Sensemaking Example

We have previously described sensemaking taxonomies as providing insight into how humans make use of the information around them and the mixed-initiative systems which support them in this process. Next we examine searching through the lens of the Human Oracle Model. Of the six general sensemaking subtasks identified by [84], searching represents the best starting point as it is found as a subtask in some form across many taxonomies of sensemaking [25, 41, 61, 62, 66, 84, 85] and may be triggered in any number of ways. Successful searching is required before any gaps can be filled and search results will influence mental maps, reports, and other outcomes. It represents a leverage point [61] where human knowledge begins to interact with machine. It also provides an illustrative example for how mixed-initiative systems might support humans in the sensemaking process.

Though the human-only approach described in Section 5.1 was designed for a monitoring example, it also accurately reflects the process of a human searching for relevant documents in a collection of k documents. Human-only approaches to searching are costly, and

an exhaustive search for relevant information will require the same $\langle \Phi_H, \Phi_M \rangle = \langle k * C_H, 0 \rangle$ work as before with C_H representing the cost of a human to determine if a single given document is relevant. In practice, human sensemakers do not have the luxury of endless search time and so rarely exhaustively search. Expertise and experience teach sensemakers when they have obtained a sufficient q documents to satisfy their current need. This threshold must be determined per search and may be dependent on the quality of sources. Thus, we require a human oracle that can determine when sufficient evidence has been gathered, and so adapt our human-only search algorithm as follows:

Algorithm 4: Human-Only Search

```
Let H_REL be a Human Oracle that performs RELEVANCE
  testing;
Let H_SAT be a Human Oracle that performs SATISFY operations;
foreach sample k do
    if H_REL(k) == true then
        Select sample k into q;
    if H_SAT(q) == true then
        Exit sample k loop;
end
```

The overall work of the human in this case depends then on the cost of evaluate the relevance of a single document, C_{REL}, and the additional cost to determine if the selected documents satisfy search goals, C_{SAT}. We also need to consider the order in which documents are presented to the human. In the best case, the first q documents would be satisfactory and the search could terminate before all k documents are seen. In the worst case, the relevant documents will be the last q seen and the search will be exhaustive. Randomly ordered searches will then be dependent on P, the proportion of useful documents in k. Thus, the total work performed by the human in this unsupported search will be

$$\langle \Phi_H, \Phi_M \rangle = \langle (C_{SAT} \div P) * C_{REL}, 0 \rangle$$

The human-only search provides insight into precisely why supported searching in mixed-initiative systems is so useful. System-generated recommendations can make it more likely that relevant documents will be found early in a search. Moreover, as the available k grows beyond what can possibly be exhaustively searched, system recommendations may be the only way to reach a satisfied search. This is easily the case in streaming environments where analysts look for sufficient signals to support their hypothesis regarding the occurrence of particular activities. Utilizing machine effort as well makes it more likely that the search will be satisfied without requiring the human to determine the relevance of all k documents:

Algorithm 5: Human+Machine Recommended Search

```
Let H_REL be a Human Oracle that performs RELEVANCE
  testing;
Let H_SAT be a Human Oracle that performs SATISFY operations;
Let M_REC be a machine that makes RECOMMENDATIONS;
foreach sample k do
    if M_REC(k) == true then
        if H_REL(k) == true then
            Select sample k into q;
        if H_SAT(q) == true then
            Exit sample k loop;
end
```

We must now account for the expected number of recommendations as the limiting factor in the work performed by the human. We still have some P proportion of truly relevant documents in k, but this proportion now affects the machine instead of the human. The better the recommendations, the more it will appear to a human that *all* documents are useful. We can replace this now with the expected number of recommendations generated by a system (no system has been declared perfect yet, we can still expect some spurious recommendations). There will now be a cost of C_{REC} to the machine to recommend documents. The total work of the mixed-initiative search then becomes

$$\langle \Phi_H, \Phi_M \rangle = \langle E[REC] * (C_{REL} + C_{SAT}), k * C_{REC} \rangle$$

7 Design Demonstration: Cybersecurity Scenario

Next we examine a hypothetical sensemaking environment designed to leverage the insights we've gained by examining sensemaking tasks in streaming environments. Our hypothetical scenario is sensitive to the needs of cybersecurity analysts, who must assemble clues from a multitude of sources on-the-fly. This has traditionally been accomplished through the use of tools such as Splunk[2] which supports complex querying for patterns. Other technologies such as OpenStack[3] are used to set up and manage networks and clouds while providing a wealth of telemetry information including firewall creation and updates, vpn connections, load balancing metrics, etc. The number of data streams k that a cyber analyst must monitor in real time quickly grows to the hundreds and thousands. For our current scenario, we will limit the focus of our cyber analyst to spotting suspicious firewall activity on their network, particularly from users with escalated privileges.

7.1 User Tasks

The analyst in this scenario is engaged in several goal-oriented tasks at once. They are actively monitoring for indications of threats that they have seen before (and therefore know how to respond). They are also concerned about the appearance of new threats they have not encountered before. When such a threat is encountered, they must then determine how to isolate the threat, remediate the affected situations, and then prevent such a threat from threatening their system again. In our focused case of suspicious firewall activity, our cyber analyst is looking for indications that a set of valid network credentials is modifying the network's firewalls in ways that indicate they've somehow escalated their privileges beyond what they should normally be allowed to do. For our scenario, the cyber analyst is monitoring firewall changes and selecting suspicious activity for further investigation.

7.2 Visualizations and Interactions

As noted in [7], visualizations have been a difficult problem for cybersecurity research to effectively solve. To crudely summarize, visualizations appear to impose an arbitrary "middle layer" between an analyst's ability to monitor a system and navigate to specific data that need investigating. Analysts prefer to work directly with their data to be certain of their conclusions. Unfortunately, humans are better suited to pick out graphical anomalies compared with textual ones. Anscombe's quartet [3] is a simple, well known demonstration of the power of visualization over raw data.

In our scenario of firewalls and escalated privileges, our interface will defer to analyst preference for tabular, raw data and the machine/interface will be challenged with presenting these rows of data in such a way that the analysts using them don't need to search. We will attempt to design an interface which balances tasks similar to the searching with recommendations of Section 5.2. The system will present firewall meters in a particular order determined by the analyst. Selection of a row will expand the row to include more details about the network user responsible for the entry. The system will also then visually flag other rows the same user is responsible for as well as similar rows regardless of the network user responsible. The analyst driving the investigation should be able to select rows or network users in this manner an mark them as important enough for continued observation. In turn, the system would give such marked or followed meters and network users priority in the interface and reduce the time the analyst spends locating them among other entries. Additional embellishments to the interface might include system generated metadeta

[2] http://www.splunk.com

[3] http://www.openstack.org/

such as trends in the amount of activity associated with a specific network user or network traffic affected by a given firewall. These additions would help an analyst determine whether or not a network user or firewall merits continued investigation.

7.3 Scenario Complexity

Our scenario closely parallels a triage-type system. The analyst is responsible for verifying that the machine-recommended data is important, and then prioritizing which data they wish to handle first. Data in which the human has indicated interest is kept elevated by the machine in order to reduce the amount of effort the human is required to spend following it for updates. Additionally, whenever the human interacts with a given piece of data, similar data is flagged as relevant in the event that the human decides to expand or alter her investigation. This gives us a mixed-initiative system where a cyber analyst drives the investigation of firewall rules and activity by network users and is supported by machine abilities to process large volumes of data for similarities and remember previously important actions. In the language of the *Human Oracle Model*, this system is following the same approach that was outlined in the Human+Machine Triage algorithm of 6.1. as before, we have a machine that generates alerts/recommendations (M_{ALERT}). The analyst determines if the alert is legitimate (H_{FA}), and then marks important alerts for action and follow-up (H_{TRIAGE}). Thus, we would expect the overall performance to again be:

$$\langle \Phi_H, \Phi_M \rangle = \langle k(E[ALERT] * (C_{FA} + (1 - E[FA]) * C_T * s)), k * C_{ALERT} \rangle$$

with limiting factors being the rate at which alerts are generated $E[ALERT]$, as well as the false alarm rate $E[FA]$.

8 Additional Domain Applications

In this paper, we demonstrated how the Human Oracle Model can aid in the design and evaluation of tools for streaming data analysis through example problems in the domain of cybersecurity. While the rich canon of human-machine interactive systems in this domain makes this a useful exemplar, we suggest that the benefits afforded by this model for improving our understanding of streaming analysis reach far beyond cybersecurity. For example, the theoretical models of streaming analysis tasks in cybersecurity that were illustrated in this paper could also apply directly to other important application areas. Indeed, one of the primary strengths of these models is that they provide a level of abstraction that is useful not only in characterizing common approaches *within* a domain, but which may facilitate the transfer of ideas *across* domains.

8.1 Healthcare

Interactive health information technology systems are providing critical support to medical professionals, but as with the examples presented in this work, providing appropriate computational support for triage and monitoring is a nontrivial task. At present, workflow models and model-based design methods are being used to provide designers and developers with fundamental system requirements [6]. Along with model checking [12], our work compliments these efforts by providing a formal description of human and machine balance so that areas of inefficiency or inaccuracy can be targeted and improved. Frameworks such as TURF [83] consider measure important factors such as learnability, efficiency, and error prevention. Our work can add to this discussion by illuminating how the balance between human and machine may influence efficiency and error rate, specifically when it comes to describing the human effort involved.

8.2 Finance

Each day, billions of financial transactions impact the lives of people all over the world. Some (hopefully small) portion of these transactions are fraudulent, or in some way connected to criminal activity. Detecting and remediating such financial crime is a task that has historically fallen on the shoulders of human analysts, who could easily become overwhelmed by the sheer volume of transactions. However, because the individuals and organizations perpetrating financial fraud are constantly adapting their strategies to elude detection, the detection of suspicious transactions cannot be fully automated. Systems such as *WireVis* [17, 18] have been presented to support hybrid human-machine analysis in this domain. Coordinated views including time-series, networks of entities or keywords, and heatmaps have all been implemented and prototypes evaluated in terms of the hardware requirements to support the analysis of at-scale data [18]. More recent contributions have applied visual analytics techniques to financial stability monitoring [35]. As with many other domains, one of the largest barriers to the adoption of these systems has been in justifying the cost (both in terms of time and potential downtime of detection) of retraining analysts. The Human Oracle Model could support work such as these by describing the complexity of user tasks, and for comparing the potential benefits of these novel systems and workflows over existing best practices.

8.3 Power Grid Management

Large infrastructure systems such as the power grid produce an enormous volume of data, and real-time situational awareness is critical for maintaining safe and efficient operations [27]. Intensive observations and task analyses have produced an understanding of the complex and collaborative behavior of grid operators [56]. The work we present here complements this analysis by providing insight into the complexity of observed tasks, and could help to prioritize how improvements based on such analysis are made. In addition to these analyses, highly-specialized decision support systems such as M-DART [53] have been introduced to support grid operators in managing high-volume data from multiple streams, assist in anomaly detection, and even facilitate causal inference. The importance of small improvements in reaction time cannot be understated: the amount of time elapsed in responding to power grid events can mean the difference between the nuisance of a brief flicker of the lights and the potentially life-threatening effects of a full scale blackout. Work such as [47] describes the cascading effects of losing multiple substations, and provides evidence that preventing substation loss early has a profound impact on the resilience of the grid. The work presented here contributes to this discussion by helping us to understand where obstacles in the human operator's task load could exacerbate these conditions, and where appropriate automation would improve grid reliability.

9 Discussion and Future Extensions

The model presented in this paper provides a critical first step in quantifying the use of humans as computational resources, and helps us to better understand the intricate relationships among different problems and problem families when viewed through this simple lens. That said, this model only just scratches the surface of this potentially rich area for future research, and conveniently ignores some very real factors present in any system involving the variability of computing using humans as part of the computational engine. In the following sections, we will discuss some of the limitations of this model, as well as motivate our continued research in this area.

9.1 Resource Constraints and Imperfect Oracles

Under the straightforward Human Oracle Model, there is an implicit assumption a Human Oracle is always be able to provide the correct answer at a fixed cost. Intuition and experience indicates that in reality, humans don't work this way: they eventually get tired or bored, and as a consequence their speed and accuracy suffer. In addition to variation over time, individual differences in ability and cost are absent, despite their demonstrated impact on visualization performance [57]. In the real world, not all humans are equal in their capacity to answer the questions we ask. Some are more skilled or have better training, and their expertise comes (we presume) at a higher cost.

Similar issues have arisen in the area of active learning, which has historically assumed a single tireless, flawless, benevolent oracle was always available to provide labels for its training data. *Proactive learning* relaxes these assumptions, adopting a decision-theoretic approach to match one of a collection of (possibly imperfect) oracles to each instance [28]. More recent work in the area of *multiple expert active*

learning (MEAL) improves upon this model by incorporating load balancing to ensure that no worker has to shoulder an inequitable share of the burden [74]. These methods assume there exists some mechanism to model both how hard any single problem instance is, as well as the cost and effectiveness of a given worker.

9.1.1 Future Work: Empirical Validation

It is our claim that theoretical arguments are useful in helping us to unravel the complex phenomena involved in the integration of human and machine effort. In order for this claim to hold true, it is critical that the models we build actually parallel the observations that we make of these phenomena in the wild. While perhaps ironic, we believe that the next step in advancing the utility of the theoretical models presented here is empirical validation. As of this writing, we have therefore deployed a web-based human subjects experiment via Amazon Mechanical Turk [11, 48, 58] in which observe human analytical behavior and performance as we manipulate various components of a streaming analysis environment (stream volume, sampling rate, etc.).

Fig. 4. The training module of our ongoing large-scale human subjects experiment in which we manipulate various components of a simple streaming analysis environment in order to provide empirical validation for the predictions made by the Human Oracle Model. Participants are asked to perform analysis of a synthetic data stream under conditions that simulate the *Triage*, *Monitor*, and *Sensemaking* tasks modeled in Sections 5 and 6.

Through this experiment, we will collect interaction data from a large number of participants (target $n = 1,000$) as they perform various streaming analysis tasks. Participants are first trained to recognize patterns in the streaming data, and then to identify deviations from those patterns (see Fig. 4). They will then be assigned to one of several analytical conditions that simulate the *Triage*, *Monitor*, and *Sensemaking* tasks modeled in Sections 5 and 6. Upon collecting sufficient data, we will then compare participants' performance with that predicted by the models described in this paper. This data will help us to better understand the relative cost of various sensemaking subtasks, as well as characterize how that cost varies between individuals and how it changes throughout the course of the analytical process. By comparing the performance of predicted by the Human Oracle Model with that of actual human-machine systems, we hope to identify areas for tuning and refinement of these models overall, as well as identify values for the constants and coefficients that are at present conveniently ignored. As the empirical validation of physiological models such as Fitts' Law [34] as they apply to human-computer interaction has enabled the development of more effective user interfaces, we hope that this line of inquiry will facilitate the development of more effective analytical tools.

9.2 Quantifying the Human Brain

This highlights another problem: at present, there does not exist any reliable method for quantifying how hard a human has to work in order to accomplish a given task. While cognitive modeling techniques can help us to understand the interplay between stimulus and response, existing architectures are not designed to determine the "complexity" of the model itself. As such, at present this model cannot actually tell us *how much work* the human is doing; it only tells us **how many times** the human is working. When the task is comparable, such as when we are comparing various monitoring algorithms, this does not pose a significant problem. However, because we don't fully understand the fundamental operations of the human brain or how they assemble to perform computation, it is not yet possible to calculate a precise per-operation cost. This leaves us unfortunately stuck when we try to make comparisons between systems that ask the human to perform different kinds of actions.

9.2.1 Future Work: Semantic Interaction

One approach to understanding how much work a human is doing comes from the concepts of semantic interaction [30] which emphasizes co-reasoning between human analysts and analytic models. The critical task is for tacit user knowledge to be captured via direct manipulation of data in visualizations. By directly binding model-steering to the interactive elements of a visualization, we reduce the confounding influence of separate model parameters and their impacts on interface usability. Human users needed put forth effort to understand any given model parameter, they can focus on their own thoughts and reasoning process while underlying system models are steered by the feedback from the visualization. Such semantic interaction has been widely studied in spatial visual metaphors, particularly in text processing domains where relationships and similarities between data objects can be easily captured with proximity [32, 10, 31]. Future work is needed to establish other visual metaphors and interactions which elicit the same reflection of human cognition. Additional benefits of a wider adoption of semantic interaction include a reduction in the number of tasks a human user may be asked to complete for a given analysis: if the human is able to steer supporting models during the course of their analysis, they do not need to pause their analysis for the additional work of model tuning which will reduce the overall complexity of any analysis process.

10 Conclusions

Human-machine collaborative systems are becoming increasingly important to visual analytics as the complexity and velocity of data increases. We have made brief allusions to streaming data throughout this work. Streaming data has made human-machine collaborative systems even more critical to the analytical tasks required for sensemaking. Compared with static data, streaming data presents several additional challenges to sensemaking: it arrives from a multitude of sources both human and machine generated and at such speeds and volumes that it cannot be collected, stored, or processed fast enough for complete samplings. To understand the full impact of streaming data on visual analytic systems, we must have methods for describing the expected effort of humans and machines as they work together. In this work, we have argued for the need for a theoretical framework through which to understanding the complexity of these human-machine hybrid systems. We have demonstrated the use of the *Human Oracle Model* for classifying the task complexity of existing systems, as well as its use in understanding yet-to-be-implemented systems. By making use of tools like the *Human Oracle Model*, we can begin to understand how human tasks must me modified in order to cope with reduced time and increased data volume. A better understanding of task complexity means that we can better understand where small modifications to workflows will improve collaborative results. The theoretical underpinnings of these complexity models will provide a powerful mechanism to proactively select solutions from across the visual analytics domain and generalize future findings to new areas.

ACKNOWLEDGMENTS

The authors wish to thank Samantha Behrens, Zheng "Alice" Mu, and the rest of the Human Computation and Visualization Laboratory at Smith College for their implementation and design efforts in support of this work. The research described in this paper is part of the Analysis In Motion Initiative at Pacific Northwest National Laboratory. It was 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.

REFERENCES

[1] M. Alsaleh, A. Alqahtani, A. Alarifi, and A. Al-Salman. Visualizing phpids log files for better understanding of web server attacks. In *Proceedings of the Tenth Workshop on Visualization for Security*. ACM, 2013.

[2] M. Angelini, N. Prigent, and G. Santucci. Percival: proactive and reactive attack and response assessment for cyber incidents using visual analytics. In *Proceedings of Visualization for Security, IEEE Symposium on*, 2015.

[3] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.

[4] D. L. Arendt, R. Burtner, D. M. Best, N. D. Bos, J. R. Gersh, C. D. Piatko, and C. L. Paul. Ocelot: user-centered design of a decision support visualization for network quarantine. In *Proceedings of Visualization for Security, IEEE Symposium on*, Oct 2015.

[5] J. Attenberg, P. Ipeirotis, and F. Provost. Beat the machine: Challenging workers to find the unknown unknowns. In *Workshop on Human Computation, AAAI Conference on Artificial Intelligence*, 2011.

[6] A. B. Berry, K. A. Butler, C. Harrington, M. O. Braxton, A. J. Walker, N. Pete, T. Johnson, M. W. Oberle, J. Haselkorn, W. P. Nichol, et al. Using conceptual work products of health care to design health it. *Journal of Biomedical Informatics*, 59:15–30, 2016.

[7] D. M. Best, A. Endert, and D. Kidwell. 7 key challenges for visualization in cyber network defense. In *Proceedings of the Eleventh Workshop on Visualization for Security*, pages 33–40. ACM, 2014.

[8] F. Boujarwah, J. Kim, G. Abowd, and R. Arriaga. Developing scripts to teach social skills: can the crowd assist the author? In *Workshop on Human Computation, AAAI Conference on Artificial Intelligence*, 2011.

[9] I. Bowman, S. Joshi, and J. Van Horn. Query-based coordinated multiple views with feature similarity space for visual analysis of mri repositories. In *Proceedings of the Conference on Visual Analytics Science and Technology (VAST)*, pages 267–268. IEEE, 2011.

[10] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *Proceedings of the Conference on Visual Analytics Science and Technology (VAST)*, pages 83–92. IEEE, 2012.

[11] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[12] K. A. Butler, E. Mercer, A. Bahrami, and C. Tao. Model checking for verification of interactive health it systems. In *Proceedings of the AMIA Annual Symposium*, volume 2015, page 349. American Medical Informatics Association, 2015.

[13] B. C. M. Cappers and J. J. van Wijk. Snaps: Semantic network traffic analysis through projection and selection. In *Proceedings of Visualization for Security, IEEE Symposium on*, 2015.

[14] S. K. Card, A. Newell, and T. P. Moran. *The psychology of human-computer interaction*. L. Erlbaum Associates Inc., 1983.

[15] J. Chamberlain, M. Poesio, and U. Kruschwitz. A demonstration of human computation using the phrase detectives annotation game. In *SIGKDD Workshop on Human Computation*, pages 23–24. ACM, 2009.

[16] K. Chan, I. King, and M. Yuen. Mathematical modeling of social games. In *Proceedings of the Computational Science and Engineering, International Conference on*, volume 4, pages 1205–1210. IEEE, 2009.

[17] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *Proceedings of the Symposium on Visual Analytics Science and Technology (VAST)*, pages 155–162. IEEE, 2007.

[18] R. Chang, A. Lee, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. Scalable and interactive visual analysis of financial wire trans. for fraud detection. *Journal of Information Visualization*, 7(1):63–76, 2008.

[19] T. Chang, C. Chan, and J. Hsu. Musweeper: An extensive game for collecting mutual exclusions. In *Workshop on Human Computation, AAAI Conference on Artificial Intelligence*, 2011.

[20] K. Cook, N. Cramer, D. Israel, M. Wolverton, J. Bruce, R. Burtner, and A. Endert. Mixed-initiative visual analytics using task-driven recommendations. In *Proceedings of the Conference on Visual Analytics Science and Technology (VAST)*, pages 9–16. IEEE, 2015.

[21] R. Crouser and R. Chang. An affordance-based framework for human computation and human-computer collaboration. *Visualization and Computer Graphics, IEEE Trans. on*, 18(12):2859–2868, 2012.

[22] R. J. Crouser, R. Chang, and B. Hescott. Toward complexity measures for systems involving human computation. *Journal of Human Computation*, 1(1):45–65, 2014.

[23] P. Dai, D. Weld, and Mausam. Human intelligence needs ai. In *Workshop on Human Computation, AAAI Conference on Artificial Intelligence*, 2011.

[24] E. Della Valle, S. Ceri, F. van Harmelen, and D. Fensel. It's a streaming world! reasoning upon rapidly changing information. *Journal of Intelligent Systems*, 24(6):83–89, 2009.

[25] B. Dervin. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Knowledge Management*, 2(2):36–46, 1998.

[26] D. Diaper and N. Stanton. *The handbook of task analysis for human-computer interaction*. CRC Press, 2003.

[27] U. DOE. Grid 2030: A national vision for electricity's second 100 years. *US DOE Report*, 2003.

[28] P. Donmez and J. Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the Seventeenth Conference on Information and Knowledge Management*, pages 619–628. ACM, 2008.

[29] A. Endert. *Semantic Interaction for Visual Analytics: Inferring Analytical Reasoning for Model Steering*. PhD thesis, Virginia Polytechnic Institute and State University, 2012.

[30] A. Endert. Semantic interaction for visual analytics: Toward coupling cognition and computation. *Computer Graphics and Applications*, 34(4):8–15, 2014.

[31] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 473–482. ACM, 2012.

[32] A. Endert, C. Han, D. Maiti, L. House, and C. North. Observation-level interaction with statistical models for visual analytics. In *Proceedings of the Conference on Visual Analytics Science and Technology (VAST)*, pages 121–130. IEEE, 2011.

[33] R. F. Erbacher. Visualization design for immediate high-level situational assessment. In *Proceedings of the Ninth International Symposium on Visualization for Security*, pages 17–24. ACM, 2012.

[34] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6):381, 1954.

[35] M. D. Flood, V. L. Lemieux, M. Varga, and B. W. Wong. The application of visual analytics to financial stability monitoring. *Journal of Financial Stability*, 2016.

[36] J. J. Fowler, T. Johnson, P. Simonetto, M. Schneider, C. Acedo, S. Kobourov, and L. Lazos. Imap: Visualizing network activity over internet maps. In *Proceedings of the Eleventh Workshop on Visualization for Security*, pages 80–87. ACM, 2014.

[37] C. C. Gray, P. D. Ritsos, and J. C. Roberts. Contextual network navigation to provide situational awareness for network administrators. In *Proceedings of Visualization for Security, IEEE Symposium on*, 2015.

[38] T. M. Green, W. Ribarsky, and B. Fisher. Building and applying a human cognition model for visual analytics. *Journal of Information Visualization*, 8(1):1–13, 2009.

[39] L. Hao, C. G. Healey, and S. E. Hutchinson. Ensemble visualization for cyber situation awareness of network security data. In *Proceedings of Visualization for Security, IEEE Symposium on*, 2015.

[40] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 159–166. ACM, 1999.

[41] Y.-a. Kang and J. Stasko. Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *Proceedings of the Conference on Visual Analytics Science and Technology (VAST)*, pages 21–30. IEEE, 2011.

[42] D. Kieras and K. A. Butler. Task analysis and the design of functionality.

In H. Topi and A. Tucker, editors, *Computing Handbook, Third Edition: Information Systems and Information Technology*, pages 33–1. Chapman and Hall/CRC, 2014.

[43] D. Kieras and S. P. Marshall. Visual availability and fixation memory in modeling visual search using the epic architecture. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 423–428. Citeseer, 2006.

[44] D. E. Kieras and D. E. Meyer. An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, 12(4):391–438, 1997.

[45] D. E. Kieras, S. D. Wood, K. Abotel, and A. Hornof. Glean: A computer-based tool for rapid goms model usability evaluation of user interface designs. In *Proceedings of the 8th annual ACM symposium on User interface and software technology*, pages 91–100. ACM, 1995.

[46] D. E. Kieras, S. D. Wood, and D. E. Meyer. Predictive engineering models based on the epic architecture for a multimodal high-performance human-computer interaction task. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 4(3):230–275, 1997.

[47] R. Kinney, P. Crucitti, R. Albert, and V. Latora. Modeling cascading failures in the north american power grid. *The European Physical Journal B-Condensed Matter and Complex Systems*, 46(1):101–107, 2005.

[48] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456. ACM, 2008.

[49] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 1–8. IEEE, 2008.

[50] E. Law, L. von Ahn, R. Dannenberg, and M. Crawford. Tagatune: A game for music and sound annotation. *of ISMIR (Vienna, Austria)*, 2007.

[51] P. A. Legg. Visualizing the insider threat: challenges and tools for identifying malicious user activity. In *Proceedings of Visualization for Security, IEEE Symposium on*, 2015.

[52] Z. Liu, B. Lee, S. Kandula, and R. Mahajan. Netclinic: Interactive visualization to enhance automated fault diagnosis in enterprise networks. In *Proceedings of the Symposium on Visual Analytics Science and Technology (VAST)*, pages 131–138. IEEE, 2010.

[53] N. Lu, P. Du, F. L. Greitzer, X. Guo, R. E. Hohimer, and Y. G. Pomiak. A multi-layer, data-driven advanced reasoning tool for intelligent data mining and analysis for smart grids. In *Power and Energy Society General Meeting*, pages 1–7. IEEE, 2012.

[54] W. J. Matuszak, L. DiPippo, and Y. L. Sun. Save: Situational awareness visualization for security of smart grid systems. In *Proceedings of the Tenth Workshop on Visualization for Security*, pages 25–32. ACM, 2013.

[55] T. Nunnally, K. Abdullah, A. S. Uluagac, J. A. Copeland, and R. Beyah. Navsec: A recommender system for 3d network security visualizations. In *Proceedings of the Tenth Workshop on Visualization for Security*, pages 41–48. ACM, 2013.

[56] J. H. Obradovich. Understanding cognitive and collaborative work observations in an electric transmission operations control center. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 55, pages 247–251. SAGE Publications, 2011.

[57] A. Ottley, R. J. Crouser, C. Ziemkiewicz, and R. Chang. Manipulating and controlling for personality effects on visualization tasks. *Journal of Information Visualization*, page 1473871613513227, 2013.

[58] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[59] S. Papadopoulos, G. Theodoridis, and D. Tzovaras. Bgpfuse: Using visual feature fusion for the detection and attribution of bgp anomalies. In *Proceedings of the Tenth Workshop on Visualization for Security*, pages 57–64, New York, NY, USA, 2013. ACM.

[60] R. Pekrun, T. Goetz, L. M. Daniels, R. H. Stupnisky, and R. P. Perry. Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, 102(3):531, 2010.

[61] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of the International Conference on Intelligence Analysis*, volume 5, pages 2–4, 2005.

[62] Y. Qu and G. W. Furnas. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *Journal of Information Processing & Management*, 44(2):534–555, 2008.

[63] A. Quinn and B. Bederson. A taxonomy of distributed human computation. *Human-Computer Interaction Lab Tech Report, University of Maryland*, 2009.

[64] A. Quinn and B. Bederson. Human computation: a survey and taxonomy of a growing field. In *29th SIGCHI Conference on Human Factors in Computing Systems*, pages 1403–1412. ACM, 2011.

[65] J. Rasmussen, K. Ehrlich, S. Ross, S. Kirk, D. Gruen, and J. Patterson. Nimble security incident management through visualization and defensible recommendations. In *Proceedings of the Seventh International Symposium on Visualization for Security*, pages 102–113. ACM, 2010.

[66] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 269–276. ACM, 1993.

[67] J. Saxe, D. Mentis, and C. Greamo. Visualization of shared system call sequence relationships in large malware corpora. In *Proceedings of the Ninth International Symposium on Visualization for Security*, pages 33–40. ACM, 2012.

[68] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

[69] D. Shahaf and E. Amir. Towards a theory of AI completeness. In *Proceedings of the AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 150–155, 2007.

[70] F. Stoffel, F. Fischer, and D. A. Keim. Finding anomalies in time-series using visual correlation for interactive root cause analysis. In *Proceedings of the Tenth Workshop on Visualization for Security*, pages 65–72. ACM, 2013.

[71] J. A. Swets, editor. *Signal Detection and Recognition by Human Observers*. John Wiley & Sons Inc., NY, 1964.

[72] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326. ACM, 2004.

[73] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. recaptcha: human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[74] B. Wallace, K. Small, C. Brodley, and T. Trikalinos. Who should label what? instance allocation in multiple expert active learning. In *SDM*, pages 176–187, 2011.

[75] J. S. Warm, R. Parasuraman, and G. Matthews. Vigilance requires hard mental work and is stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):433–441, 2008.

[76] K. Webga and A. Lu. Discovery of rating fraud with real-time streaming visual analytics. In *Proceedings of Visualization for Security, IEEE Symposium on*, 2015.

[77] D. Weld, P. Dai, et al. Execution control for crowdsourcing. In *24th Annual Symposium on User Interface Software and Technology*, pages 57–58. ACM, 2011.

[78] T. Wüchner, A. Pretschner, and M. Ochoa. Davast: Data-centric system level activity visualization. In *Proceedings of the Eleventh Workshop on Visualization for Security*, pages 25–32, New York, NY, USA, 2014. ACM.

[79] L. Xu, A. Krzyzak, and C. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, Man and netics, IEEE Trans. on*, 22(3):418–435, 1992.

[80] Y. Yang, B. Zhu, T. Guo, L. Yang, S. Li, and N. Yu. A comprehensive human computation framework: with application to image labeling. In *16th International Conference on Multimedia*, pages 479–488. ACM, 2008.

[81] J. Zhang and K. Butler. Ufurt: A work-centered framework and process for design and evaluation of information systems. In *Proceedings of HCI international*, pages 1–5, 2007.

[82] J. Zhang and D. A. Norman. Representations in distributed cognitive tasks. *Cognitive science*, 18(1):87–122, 1994.

[83] J. Zhang and M. F. Walji. Turf: Toward a unified framework of ehr usability. *Journal of biomedical informatics*, 44(6):1056–1067, 2011.

[84] P. Zhang. Sensemaking: Conceptual changes, cognitive mechanisms, and structural representations. a qualitative user study. *ProQuest LLC*, 2010.

[85] P. Zhang and D. Soergel. Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. *Journal of the Association for Information Science and Technology*, 65(9):1733–1756, 2014.

[86] X. Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2:3, 2006.

[87] W. Zhuo and Y. Nadjin. Malwarevis: Entity-based visualization of malware network traces. In *Proceedings of the Ninth International Symposium on Visualization for Security*, pages 41–47. ACM, 2012.