Modularized Sensemaking Pipeline to Enable Text Analysis Assisted by Crowds and Algorithms

Tianyi Li
PhD Student, GRA

Dr. Chris North
Primary Advisor, Project Co-PI

Dr. Kurt Luther
Co-Advisor, Project PI

1 Abstract

The increasing volume of text datasets is challenging the cognitive capabilities of expert analysts to produce meaningful insights. Large-scale distributed agents like machine learning algorithms and crowd workers present new opportunities to make sense of big data. However, we must first overcome the challenge of modeling and guide the overall process so that many distributed agents can meaningfully contribute to suitable components. Inspired by the sensemaking loop, collaboration models, and investigation techniques used in Intelligence Analysis community, we propose a pipeline as an artifact of cooperation among expert analysts, crowds, and algorithms. We do this by modularizing and clarifying the components in the sensemaking loop so that they are connected via clearly defined inputs and outputs. Different agents can then contribute to different steps with appropriate techniques and pass the intermediate results along the pipeline. We instantiate the pipeline by two commonly used investigation strategies and experimented with crowd workers on Amazon Mechanical Turk. Our results show that the pipeline can successfully guide crowd workers to contribute meaningful insights that are helpful to solve complicated sensemaking challenges. This allows us to imagine broader possibilities for how each component could be executed: with individual experts, crowds, or algorithms, as well as new combinations of these, where each is best suited.

2 Introduction

The information explosion from modern technology has spurred increased interest in sensemaking to help people gain insights and suggest effective actions from the big data. The cognitive capability, time and expertise of individual experts are impressive but fundamentally limited, and domain experts are scarce and expensive resources by nature. Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically [13] by modern computer technology alone, understanding and acting on the information to solve real-world problems requires human computation as well. Making sense of a large amount of text data requires understanding natural languages, which is considered AI-hard [30]. Observations suggest evidence marshaling and synthesis are particularly difficult. To get the big picture by looking at many pages of text, the analyst relies heavily on memory to connect the dots. Crowdsourcing and algorithms present new opportunities for large-scale sensemaking, but we must first understand how sensemaking work can be modularized to allow powerful and diverse techniques to be used where they can contribute best. The existing models and theories of sensemaking process have

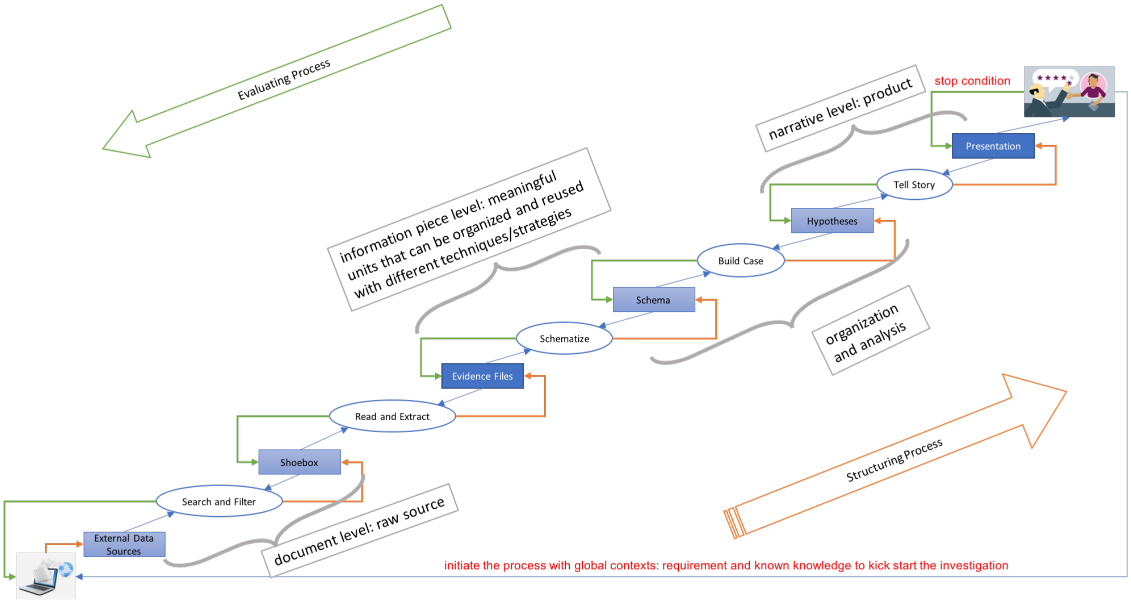

Evaluating Process
information piece level: meaningful units that can be organized and reused with different techniques/strategies
narrative level: product
stop condition
Presentation
Tell Story
Hypotheses
Build Case
Schema
Schematize
Evidence Files
Read and Extract
Shoebox
Search and Filter
External Data Sources
document level: raw source
organization and analysis
Structuring Process
initiate the process with global contexts: requirement and known knowledge to kick start the investigation

Figure 1: Diagram of Modularized Sensemaking Pipeline: The orange-colored lines represent the structuring process when an investigation is initiated; the green-colored lines represents the evaluating process to refine previous results.

been applied to various domains using different types of analysis agents including individual and groups of experts, crowd workers, and machine learning algorithms. Most of the techniques only focus on parts of sensemaking process or provide ideal inputs to non-expert agents, which requires a considerable amount of work from expert analysts. We aim to release this assumption and execute the whole sensemaking loop with non-expert agents, by modularizing different components of the process with more unified inputs and clearly defined outputs that allows combinatorial and flexible usage of them and enabling multiple paths with a mechanism to decide where to proceed next. Such a pipeline can compare or combine different agents on each component to make full use of the complementary strength of human and computation. Having more specific goals for each component enables backtracking and re-evaluation of the analysis progress. Additionally, with multiple agents involved in the analysis process, we can overcome stereotypes, bias, and group-think by re-examining intermediate and overall results.

With this motivation, we aim to answer the following research questions in this study: What are the information needs (inputs) and intermediate results (outputs) at different stages of text analysis? What are the strategies to decide the sequence of analysis tasks? When should the analysis be conducted bottom-up and when top-down? To answer these questions, we combine and adapt the original model by further modularizing the stages in the Representation Construction Model as a micro Data/Frame Model. We define the requirements (inputs and outputs) and decompose each stage into structuring agents and evaluating agents, such that they can be completed by a variety of methods, including individuals, crowds, or automated techniques. As a proof of concept, in this research project, we choose crowdsourcing to demonstrate the utility of the new model, by providing a case study on a simple mystery and a moderately sized dataset (*the Sign of Crescent*). We found that the crowds can provide meaningful insights and analysis at each step, offer multiple perspectives on the same given input, refine previous workers' result to better follow the instructions and address given expert feedback. The crowds successfully solve the simple mystery in one bottom-up execution, provide insightful analysis at each step, and addressed problems pointed out in expert feedback in the moderate dataset. This provides evidence that the more formalized model enables new possibilities for modularized development.

Our contributions are as follows. First, we disassemble and tailor the process in the context of intelligence analysis on textual evidence documents. More specifically, we define each component with its functionality, input, and output as well as the relationship among them; we also propose two meta-strategies of workflow design. Second, we conducted an empirical case study demonstrating the feasibility of the pipeline. Third, we suggest that the pipeline can be used to open the sensemaking process up for more researchers to contribute.

There are also many benefits as by-products of this work. Applying the pipeline is less demanding in terms of data confidentiality since each component loop only takes part of the data. Furthermore, executing a component requires less expertise and opens the tasks up to more possibilities including novice crowds or algorithms[20] . We can compare or combine different agents on each component to make full use of the complementary strength of human and computation. Having more specific goals for each component enables backtracking and re-evaluation of the analysis progress. Additionally, with multiple agents involved in the analysis process, we can overcome stereotypes, bias, and groupthink by re-examining intermediate and overall results.

3 Related Work

Our proposed pipeline expand on sensemaking models, the human-debugging paradigm, and prior research efforts to apply those theories in supporting expert data analysis in both visual analytics and crowdsourcing community.

3.1 Sensemaking Models and Theories

The notion of sensemaking in the field of human-computer interaction (HCI) was framed in the early 1990s [24] as the process of forming and working with meaningful representations in order to facilitate insight and subsequent intelligent action. The sensemaking process is complicated in that it iterates on multiple intertwined stages, employs combinatorial and individually-variable reasoning heuristics, and differs according to specific goals of problem-solving.

Pirolli and Card [21] proposed the Representation Construction Model (the sensemaking loop) as two iterative processes: information foraging and synthesis. Each process contains smaller loops at different levels of data granularity. Expert analysts generate a theory from raw data by going through a bottom-up process, and trigger re-evaluation on previous intermediate analysis results or re-analysis on previous data resources in a top-down process. The Data-Frame Theory [17] developed in the macrocognition approach of psychology, were also adapted [19] [25] [1] and applied [9][2] in HCI research. The Data-Frame Theory focuses on iteratively developing meaningful representations (frames) that explain external reality (data). Building on the ample literature on frames and similar concepts, Klein et al. synthesized the concepts as a structure for accounting for the data and guiding the search for more data.

However, the models are usually general and do not have clear boundaries between each component, which will lead to vague or varying interpretations and applications in real-world problems.

Empirical user studies [6] have revealed the importance of conceptual connections and domain knowledge. Applying these theories to tackle real-world problems requires experts to decide the data formats and processing methods to use in each step case-by-case, depending on the broad definitions given in the theories, their prior experience, the data under analysis, the goal of investigation and other situational constraints.

Our pipeline modularizes the sensemaking loop and leverages expert guidance to control the flow of executions of different steps by crowds and/or algorithms. For each step, we apply data-frame theory to monitor the execution status, and suggest if the current step is ready to pass to the next step, or waiting for better input from the previous step, or requires improvement in task design. Experts give feedback on final results to trigger re-execution of some steps for improvement. If the pipeline observes repeated execution in a particular sub-loop, experts will be notified and give more guidance. We will introduce more related work on expert-guided crowd work and differentiate our focus in later sections.

3.2 Human Computation and Human-Debugging Paradigm

Quinn et al. examined the focus and distinctions between Human Computation, Crowdsourcing, Social Computing and Collective Intelligence [22], offering a nice taxonomy of on-line human participation in the computational process. Both humans and computers bring strength in information processing, the latter mainly assisting the former with superior working memory and lower-bias environment [8]. Crouser et al. reviewed and identified the patterns in existing affordances representative of the study of human-computer collaborative problem-solving, understanding which forms the basis of a common framework for this domain of problems. Visual analytics community has seen a shift of human in the loop philosophy for visual analytics to a human is the loop viewpoint [10], which supports existing interactive process in situ. In the theories presented in the previous section, researchers should first decide if a problem would benefit from a collaborative technique, then which tasks to delegate to which party, and when. Given these two answers, different systems can then be compared to solve the same problem. Modeling human cognition and sensemaking is essential for developing computer-aided information processing and knowledge visualizations to address today's complex problems.

Human involvement is also important to improve the performance of existing computational AI systems. Parikh et al. proposed the human-debugging paradigm [20] to explore how the human vision process could be decomposed into a pipeline in order to identify computational bottlenecks as well as opportunities where new automated techniques could make the most impact. One of the challenges they have identified is defining inputs and outputs for each component in complicated big problems, that should be equivalent to both human and machine. Following this paradigm, the following research questions need to be answered in the context of text analysis: What are the information needs (inputs) and intermediate results (outputs, also serve as inputs for the following stages) at different stages of text analysis? What are the strategies to decide the sequence of analysis tasks? When should the analysis be conducted bottom-up and when top-down?

3.3 Sensemaking in Visual Analytics

Visual analytics community straddles both foraging and sensemaking loops in its efforts to assist both individual and groups of users in investigating and hypothesizing on complex and dynamically changing information. Individual analysts can receive a rich range of assistance from evidence to hypotheses and with multiple data types with visualization tools like Jigsaw [26]. Multiple visualizations of reports and the entities within them, as well as the connections that exist in between, allows people to interact with the views and explore possible new avenues of examination. Integration of visualization with shared accessibility and discussion enhance collaborative complex problem-solving in pairs and small groups. Timelines are often used in such visual analysis tools to represent temporal relationships within the data being investigated [4]. Bier et al. identified key aspects of the design, featuring flexible, shared information structure and visualization among experts, and a notification system that finds entities of mutual interest to multiple analysts. However, synchronous collaboration among small groups is pretty much restricted to information seeking and organization up to entity-level analysis.

Prior research has also studied collaborative sensemaking and have identified several suggestions for designing collaborative visual analytics tools. Bradel et al. explored how a large, high-resolution display as a workspace in a co-located setting helps to externalize information to the display in meaningful schemas during pairwise collaboration to make sense of large text dataset [5], and addresses problems like common ground, communication, hand-offs, coordination and attention shifting in teamwork, which is shared among most, if not all collaborative work. Dispatching simpler sensemaking tasks to multiple agents may help to solve some of the problems of attention shifting and mental model interfering. By expert-driven task distribution and aggregation, the coordination can be guided thus more efficient. The expert guidance also naturally offers multiple views on the problem, given the rich pool of strategies and methodologies developed in different domains and individual experience of experts.

3.4 Crowdsourced Sensemaking

The crowdsourcing community has seen success in integrating intelligence power of bigger crowds in complex problem-solving. Starting from low-level data processing tasks like image labeling [14], named entity extraction and merging [28], the crowds have accomplished increasingly complicated and interdependent tasks like article editing as word processor [3], and even writing short stories [15]. Crowdsourcing is different from traditional teamwork in that it aims to collect and aggregate distributed and asynchronous work among strangers. Researchers have strive to address the tension between local micro tasks and the global view of the whole dataset [12] [27], leveraging efforts of previous users [11], balancing structure, flexibility and expert guidance [16] [7], and so on. Nonetheless, asynchronous and higher level analysis is still bottlenecked by communicating insights and reasoning even among few analysts. As is acknowledged in their work, the crowdsourced approach is most valuable where experts generate a lot of valuable information that is unstructured and redundant. Soylent [3] introduce Find-Fix-Verify crowd programming pattern to increase crowd's quality of work; TurKit applied crash-and-rerun programming model [18] to propose iterative tasks on MTurk; IdeaGens [7] explored guiding complex collaborative tasks like brainstorming with Crowds by dividing the crowd into ideation and synthesis tasks. In this research project, we take a step further to explore how crowdsourced approach can be applied to raw textual documents collected on the field that is not pre-processed by expert analysts, and guide the workflow by contextualized feedback.

4 The Pipeline

The pipeline facilitates 5 steps in the sensemaking loops at a different level of details. When a new investigation is started, the pipeline is in *the structuring process* to gain some initial insights in the raw external data. After that, if the final results do not meet the requirement of customers (who initiated the investigation), which is most likely, the pipeline switch to *the evaluating process* to iteratively evaluate and refine the analysis produced in the structuring process until the final results meet the requirement; or more data collection is needed to make any actual improvement.

4.1 Symmetric Structure of Each Step

Each step *acquire* knowledge from two directions: Data Input (from lower steps) and Feedback Input (from upper steps) process the information into Data Output (to upper steps) and Feedback Output (to lower steps). The output is retained in each step and will be retrieved and/or refined to solve problems in later steps.

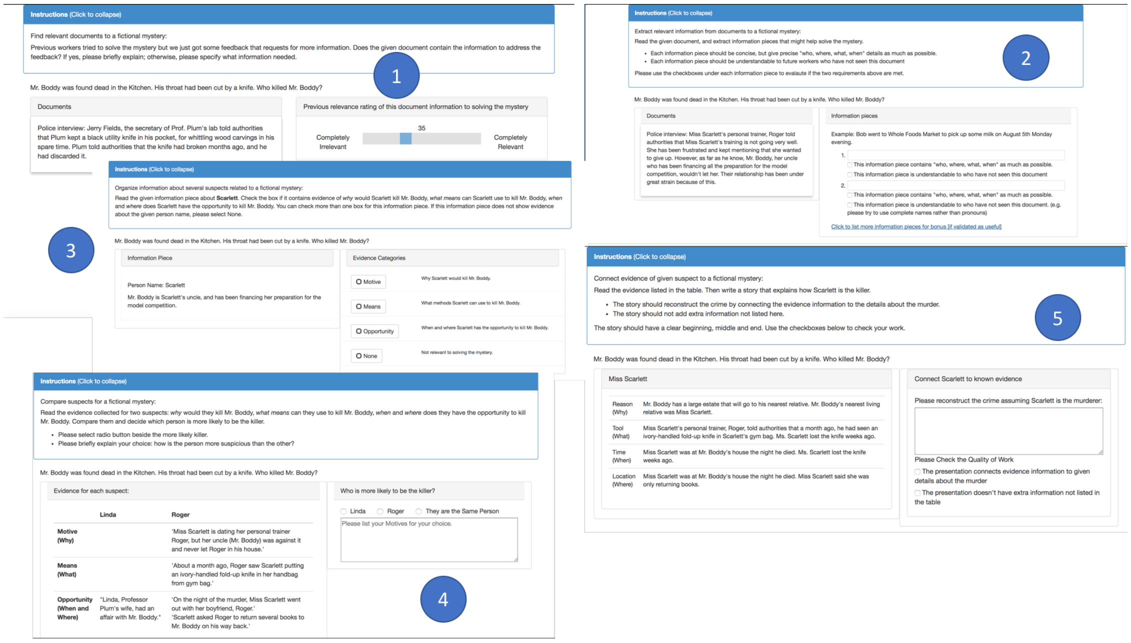
Data Input is the processed information from the lower steps, ready to be further analyzed to produce Data Output. Feedback Input describes the flaws of the previous Data Output of this step. Depending on whether the flaws can be addressed, the step might produce a refined Data Output and move on to the upper steps; or a translated Feedback Output that describes the flaws in the Data Input, which need to be fixed in order to refine the Data Output. Notably, in the structuring process, there's no Feedback Input, and Data Output is being progressively produced. We suggest that Feedback Input can be initiated as to find anything related to the investigation purpose, as we call the *Global Context*.

Each step of the pipeline has its only challenges and has its own body of research as we discussed in related works. Traditionally, analysts deal with those issues by stretching or externalizing their memory. *Our pipeline seeks to address the challenges through iterative feedback and refinement, triggering an additional analysis by demand.* Furthermore, just like any analysis or learning process, we do not expect the algorithm, crowds or even experts to produce the perfect intermediate result in each step at once and for all. *The pipeline correct errors with an iterative execution of each step with updating feedback from upper steps.* Last but not least, each step is outsourced to individual intelligence agent (algorithm, crowds, or expert), which medians careful distribution of input and aggregation of output. *The pipeline deal with the tension between local context and the big picture view by sharing feedback from upper steps to lower steps in evaluating process.*

4.2 Step1: Organize Raw Documents by Relevance

The first step of sensemaking pipeline organizes the documents from external data collection by their relevance to investigation purposes, so that later analysis can situate in a focused information space. If Step1 does not retrieve all the relevant documents, feedback from upper step will trigger a more specific search for documents containing the missing information.

Documents with important information might not be obviously relevant at first, and the definition of relevance can be different as the investigation progresses deeper. Taking fictional murder mysteries as an example, a document about Prof. Plum's handcrafting skills seems irrelevant on its own; however, with a second document revealing a secret love affair between Linda, Prof. Plum's wife with the victim Mr. Boddy, the first document turns out to be relevant and brings Prof. Plum to the list of possible suspects. Expert analysts usually read many documents together, back and



Instructions (Click to collapse)

Find relevant documents to a fictional mystery:

Previous workers tried to solve the mystery but we just got some feedback that requests for more information. Does the given document contain the information to address the feedback? If yes, please briefly explain; otherwise, please specify what information needed.

Mr. Boddy was found dead in the Kitchen. His throat had been cut by a knife. Who killed Mr. Boddy?

Documents

Police interview: Jerry Fields, the secretary of Prof. Plum's lab told authorities that Plum kept a black utility knife in his pocket, for whittling wood carvings in his spare time. Plum told authorities that the knife had broken months ago, and he had discarded it.

Previous relevance rating of this document information to solving the mystery

35

Completely Irrelevant — Completely Relevant

Instructions (Click to collapse)

Extract relevant information from documents to a fictional mystery:

Read the given document, and extract information pieces that might help solve the mystery.

- Each information piece should be concise, but give precise "who, where, what, when" details as much as possible.
- Each information piece should be understandable to future workers who have not seen this document

Please use the checkboxes under each information piece to evalaute if the two requirements above are met.

Mr. Boddy was found dead in the Kitchen. His throat had been cut by a knife. Who killed Mr. Boddy?

Documents

Police interview: Miss Scarlett's personal trainer, Roger told authorities that Miss Scarlett's training is not going very well. She has been frustrated and kept mentioning that she wanted to give up. However, as far as he know, Mr. Boddy, her uncle who has been financing all the preparation for the model competition, wouldn't let her. Their relationship has been under great strain because of this.

Information pieces

Example: Bob went to Whole Foods Market to pick up some milk on August 5th Monday evening.

1. ______
□ This information piece contains "who, where, what, when" as much as possible.
□ This information piece is understandable to who have not seen this document
2. ______
□ This information piece contains "who, where, what, when" as much as possible.
□ This information piece is understandable to who have not seen this document. (e.g. please try to use complete names rather than pronouns)

Click to list more information pieces for bonus [if validated as useful]

Instructions (Click to collapse)

Organize information about several suspects related to a fictional mystery:

Read the given information piece about **Scarlett**. Check the box if it contains evidence of *why* would Scarlett kill Mr. Boddy, *what means* can Scarlett use to kill Mr. Boddy, *when* and *where* does Scarlett have the opportunity to kill Mr. Boddy. You can check more than one box for this information piece. If this information piece does not show evidence about the given person name, please select None.

Mr. Boddy was found dead in the Kitchen. His throat had been cut by a knife. Who killed Mr. Boddy?

Information Piece

Person Name: Scarlett

Mr. Boddy is Scarlett's uncle, and has been financing her preparation for the model competition.

Evidence Categories

□ Motive	Why Scarlett would kill Mr. Boddy.
□ Means	What methods Scarlett can use to kill Mr. Boddy.
□ Opportunity	When and where Scarlett has the opportunity to kill Mr. Boddy.
□ None	Not relevant to solving the mystery.

Instructions (Click to collapse)

Compare suspects for a fictional mystery:

Read the evidence collected for two suspects: *why* would they kill Mr. Boddy, *what means* can they use to kill Mr. Boddy, *when* and *where* does they have the opportunity to kill Mr. Boddy. Compare them and decide which person is more likely to be the killer.

- Please select radio button beside the more likely killer.
- Please briefly explain your choice: how is the person more suspicious than the other?

Mr. Boddy was found dead in the Kitchen. His throat had been cut by a knife. Who killed Mr. Boddy?

Evidence for each suspect:

	Linda	Roger
Motive (Why)		'Miss Scarlett is dating her personal trainer Roger, but her uncle (Mr. Boddy) was against it and never let Roger in his house.'
Means (What)		'About a month ago, Roger saw Scarlett putting an ivory-handled fold-up knife in her handbag from gym bag.'
Opportunity (When and Where)	"Linda, Professor Plum's wife, had an affair with Mr. Boddy."	'On the night of the murder, Miss Scarlett went out with her boyfriend, Roger.' 'Scarlett asked Roger to return several books to Mr. Boddy on his way back.'

Who is more likely to be the killer?

○ Linda ○ Roger ○ They are the Same Person

Please list your Motives for your choice.

Instructions (Click to collapse)

Connect evidence of given suspect to a fictional mystery:

Read the evidence listed in the table. Then write a story that explains how Scarlett is the killer.

- The story should reconstruct the crime by connecting the evidence information to the details about the murder.
- The story should not add extra information not listed here.

The story should have a clear beginning, middle and end. Use the checkboxes below to check your work.

Mr. Boddy was found dead in the Kitchen. His throat had been cut by a knife. Who killed Mr. Boddy?

Miss Scarlett

Reason (Why)	Mr. Boddy has a large estate that will go to his nearest relative. Mr. Boddy's nearest living relative was Miss Scarlett.
Tool (What)	Miss Scarlett's personal trainer, Roger, told authorities that a month ago, he had seen an ivory-handled fold-up knife in Scarlett's gym bag. Ms. Scarlett lost the knife weeks ago.
Time (When)	Miss Scarlett was at Mr. Boddy's house the night he died. Ms. Scarlett lost the knife weeks ago.
Location (Where)	Miss Scarlett was at Mr. Boddy's house the night he died. Miss Scarlett said she was only returning books.

Connect Scarlett to known evidence

Please reconstruct the crime assuming Scarlett is the murderer:

Please Check the Quality of Work

□ The presentation connects evidence information to given details about the murder

□ The presentation doesn't have extra information not listed in the table

Figure 2: Task interfaces of Step1-5, numbered by step number

forth, to include all relevant documents [29]. The pipeline will start with directly relevant information and reveals the motive of Prof. Plum, then request more information about him through feedback.

Data Input from external data sources: all collected documents that haven't been investigated yet;

Feedback Input from Step2: missing information needed in the current relevant documents under investigation.

Data Output to Step2: documents pertinent to the investigation purposes are pushed to the next step;

Feedback Output to external: if no document contains the missing information, a translated information request will be sent to the external data collectors.

Example Task Design: Following the fictional murder example, the task of the first step is to find documents relevant to the known facts of the murder (first execution) or missing information requested from upper steps (second and later executions). Figure 3 is a task interface for crowd workers on Amazon Mechanical Turk. To better use the human intelligence of crowds, we first use an algorithm to find the documents that directly mentions the key elements in the global contexts (victim name: Mr. Boddy, murder weapon: knife, murder location: kitchen), and ask the crowds to search for indirectly relevant documents.

4.3 Step2: Collect information pieces

The second step collects medianingful information pieces from the relevant documents, so that analysts do not have to read through all the words for a piece of information. If Step2 did not extract all needed information, feedback from upper step will trigger another execution of Step2 to fill the information hole.

Important parts of information are sometimes scattered in different sentences of one document. The document might describe a person's location in the first sentence, then talks about the date and activity of that person in the end, with other information in between (3). A meaningful information piece should aggregate these scattered parts into a complete sentence describing the who, what, where, when elements. However, the information in documents is not always complete. For example, a document might describe a series phone calls from several numbers, but the person names to whom the phone numbers belong is not mentioned. By reading this document alone, we cannot create a complete piece of evidence, since the "who" element is not available. In other documents, the (phone number - person name) relationship might not draw enough attention to be extracted and pushed to upper steps. If the pipeline later detects that the phone calls are critical clues to solve the mystery and information of the person involved are needed, a refining execution of Step2 will revisit documents with both person names and those phone numbers to fill this information hole.

Data Input from Step1: documents pertinent to the investigation purposes are pushed to the next step;

Feedback Input from Step3: missing information needed in the current list of information pieces.

Data Output to Step3: list of all information pieces extracted from the document;

Feedback Output to Step1: if no document contains the missing information, a translated information request will be sent to Step1 to retrieve missing relevant documents.

Example Task Design: Following the fictional murder example, the task of the second step is to extract complete pieces of evidence relevant to the known facts of the murder (first execution) or missing information requested from upper steps (second and later executions). Figure 3 is a task interface for crowd workers on Amazon Mechanical Turk. Since the information density and topics covered vary by documents, we ask crowd workers to find the three most important pieces of information in each document. This design choice is drawn both from experience of pilot studies and the self-refining nature of the pipeline. In pilot experiments, we tried requesting for a different number of information pieces according to document length, number of entities and so on, which might result in a long list of input boxes that overwhelms workers. Nonetheless, we do not expect perfect nuggets given the human factors and imperfect data input, and the pipeline enables refinement on demand.

4.4 Step3: Orchestrate Information Pieces

The third step of sensemaking pipeline orchestrate the information pieces from Step2 so as to draw an inference from multiple perspectives. If Step3 fail to include crucial information pieces, feedback from upper steps will request for the missing ones.

The success of analysis heavily relies on the insightful orchestration of evidence that leads to the discovery of inconspicuous yet crucial clues. Expert analysts code and tag information pieces put them into appropriate structured groups (sort by time, geo-location, etc.), and "connect the dots". For example, when investigating several suspects of a murder, the evidence is usually categorized as "medians, motive, opportunity" for each person. However, the connection between dots is not always obvious. Let's say we have four information pieces: *1. Scarlett treated the victim Mr. Boddy a bottle of lemon juice on Wednesday afternoon, 2. Scarlett knew that Mr. Boddy was going to a seafood party on Wednesday night, 3. Mr. Boddy was found dead in his bedroom, the cause of death is a poison called DS. 4. It is found that mixing vitamin-c and seafood will produce DS within 24* *hours.* The four information pieces combined to reveal ”medians” for Scarlett, which suggests she controlled the timing and used a combination of chemicals to kill Mr. Boddy. However, only the third piece is directly connected to the murder case. It is challenging to identify and put the four aside from all other information pieces collected. In the lack of this category of evidence, the later steps will have trouble developing convincing hypotheses and conclusions on the real murderer. The pipeline will draw attention to ”DS” first, then found the fourth piece when requesting for more information about ”DS”. With more iterations, the pipeline will continue to request more information about the two ingredients and finally connect all four pieces and uncover the ”medians” of Scarlett.

Data Input from Step2: list of all information pieces extracted from the document.

Feedback Input from Step4: missing information in a small-scale story.

Data Output to Step4: tagged information pieces grouped in multiple ways to tell small-scale stories.

Feedback Input to Step2: if no information pieces contain the missing information, a translated information request will be sent to Step2 to retrieve missing relevant documents.

Example Task Design: Following the fictional murder example, the task of the third step comes down to tag the information pieces by categories. Figure 3 is a task interface for crowd workers on Amazon Mechanical Turk to tag the medians, motive, opportunity evidence for a person. For the sake of scalability, we ask one worker to tag one information piece and take a majority vote out of three workers to decide the tag.

4.5 Step4: Develop Hypotheses

The fourth step of sensemaking pipeline draws inference from different structures of evidence. Feedback from upper steps will challenge the reasoning logic and the weight of different perspectives, directing the creativity to the right direction.

On the one hand, people are biased with different experience and perspective, which is an actually different analytical advantage to draw creative and deep insight on the same evidence network. With the power of crowds, the ”biases” can be leveraged to achieve a comprehensive and balanced big picture of the evidence. On the other hand, different organization of evidence could lead to different inference and conclusions. With the rich structures passed from Step3, competing hypotheses can be developed to finally uncover the hidden plot. In the murder mystery example, when comparing two suspects: Prof. Plum and Scarlett, each has all three types of evidence against them. Prof. Plum has stronger motive to commit the crime of passion, and a chemistry expertise as medians (poison) of murder; whereas Scarlett has stronger medians that have the right timing and ingredients to poison the victim, and a potential motive to kill Mr. Boddy for wealth. Analysts would attribute different weights of that evidence and have competing hypotheses over who is more likely to be the murderer. If there are hypotheses from other aspects, say, Scarlett is already rich and maintain a good relationship with Mr. Boddy, then Prof. Plum might end up being more likely; otherwise, the pipeline will rely on expert feedback to incorporate professional suggestions.

Data Input from Step3: information pieces organized in multiple ways to tell small-scale stories.

Feedback Input from Step5: missing information or weakness of previous hypotheses.

Data Output to Step5: best hypotheses are drawn from the evidence available

Feedback Input to Step3: if no information piece groups contain the missing information to strengthen the previous hypotheses, a translated information request will be sent to Step3 to retrieve missing information pieces.

Example Task Design: Figure 3 is a task interface for crowd workers on Amazon Mechanical Turk to compare potential suspects and pick the most likely one. For the sake of scalability, we adopt the single elimination tournament structure and use majority vote for each pair of comparison.

4.6 Step5: State Conclusion

The final step of sensemaking pipeline combine the best hypotheses to state clear and actionable analysis results to expert analysts. Experts will evaluate and request missing information needed to fix the weakness or flaws of the current result.

The ultimate product of sensemaking process is a piece of knowledge applicable to real-world situations. This requires both a clear answer to the investigation question (e.g. who is the murderer?) and sufficient backup evidence (e.g. why he or she? what is his or her medians, motive, opportunity?). Still with the murder mystery example, when stating evidence for Scarlett to be the murderer, the fact that she is the closest relative of Mr. Boddy might be interpreted as ”opportunity” because Mr. Boddy trusts her and eat the food she gives him. There is nothing wrong with this logic, but a more concrete piece of evidence that Scarlett was with Mr. Boddy earlier the day he was murdered is not mentioned. Thus the conclusion makes sense but not enough to win the lawsuit against her. Experts understand this and will request more specific information about time and location, which might get left off when stating the conclusion (refine Step5) or is missing from the given hypotheses (pass feedback to refine Step4).

Data Input from Step4: best hypotheses are drawn from the evidence available.

Feedback Input from experts: missing information needed to strengthen the conclusion.

Data Output to Step5: a narrative of the analysis results

Feedback Input to Step4: if no hypotheses contain the missing information to strengthen the conclusion, a translated information request will be sent to Step4 to develop hypotheses from that perspective.

Example Task Design: Given that we are using a simpler dataset and only ask for one murderer, the major task is to back up the choice of a suspect with the evidence. Figure 3 is a task interface for crowd workers on Amazon Mechanical Turk to connect evidence in the profile of the most likely murderer to the known facts of the murder.

5 Pilot Study

In order to validate the pipeline concept and experiment with the pipeline components, we conducted a case study on one simple dataset and some adapted versions of it with deliberate challenges (explained in previous section). We recruited crowd agents on Amazon Mechanical Turk and performed as expert agents by ourselves (authors of this paper).

5.1 Dataset Preparation

One challenge we face to experiment the performance of the pipeline is that the available datasets are usually too complicated and ill-structured to control experiment variables. In order to gain a more accurate understanding of analytical performance at different levels, we adopt the story generation theories and frameworks [23] and designed a simple murder mystery with one victim, three main characters (suspects) and four other characters. The mystery comes with a global context (the crime scene specifying the victim, murder weapon, and location). Each main character has documents establishing his or her medians, motive and opportunity to kill the victim. The real killer will have all three types of evidence solid and concrete, whereas the other two will have one or two subtly mismatch the global context. Each other character is somehow related to one of the main characters. The ultimate plan is to experiment with a moderately-sized dataset *the Sign of Crescent* (41 documents), one of the professional intelligence analysts training materials widely used in visual analytics community.

5.2 Procedure

We deploy the pipeline with the same bottom-up workflow of both datasets on Amazon Mechanical Turk (AMT). We used majority vote mechanism for rating and voting tasks (Step1, Step3, and Step4), peer review mechanism for narrative creation tasks (Step2 and Step5). We implemented task interfaces for each task of each step with consistent design and layout. After accepting the HIT, workers would see instructions explaining the background context, task requirements, and purpose of the task. All tasks have one data input on the left side. For first-time structuring tasks, the workspace to create data output is given on the right; for refining tasks, the previous output, expert feedback, and workspace for decision making and creating corresponding output is laid out on the right side. Once they have finished the task and click the Submit button, there will be a brief validation on their output depending on the task. If the output past validation, the result will be submitted to the authors to review and approve/reject.

5.3 Participants

We recruited crowd workers from AMT, not require that workers are Masters nor did we set additional qualifications Workers must meet to work on our HITs. When workers accept a HIT, they are randomly assigned to a piece of input for a given step. For majority vote tasks (rating, tagging, comparing), we assign 3 workers for each task; for create-review tasks (extract, narrate), we assign 2 workers as needed. Each worker was unique and assigned to only one HIT to mitigate learning effects or collusion. Crowd workers who quit an accepted HIT without submitting it were not allowed to resume the unfinished work or take a new HIT.

5.4 Implementation Details

On Amazon Mechanical Turk, the unit of tasks on AMT is called Human Intelligence Task (HIT). Each HIT can have multiple ”assignment” that let multiple people working on the same task. To create a HIT, requesters on AMT create a new project under the ”Create” tab, specifying project name, contents to be displayed to workers, including the HIT title, description, keywords, reward per assignment, number of assignments per HIT, expiration date, auto-approval time (in case requesters forgot to review the results), and worker requirements. Each HIT can take a CSV file as data input and will produce a CSV file that aggregates crowd's result for all assignments. The task interface was implemented using HTML, CSS, Bootstrap, JQuery and related library for managing layout and aesthetic presentation, interactive interface, and result validation. The tasks are embedded in the MTurk system and the results are saved in CSV files on MTurk systems. Crowd workers with a web browser will be able to log in to Amazon Mechanical Turk and take on tasks.

First-time execution given new dataset

HIT Assignment	Step 1	Step 2		Step 3	Step 4	Step 5	
worker 1	rate	create		tag	select	create	
worker 2	rate		review	tag	select		review
worker 3	rate			tag	select		
#HIT	11	14	14	30	6	1	1
#assignment	33	14	14	90	18	1	3
Avg Time	6.83	9.75	4.32	4.66	21.23	12.61	3.37
median	3.84	1.56	3.23	1.67	18.65	14.60	2.30
SD	8.49	16.87	3.33	7.76	15.09	4.94	2.57

Table 1: Structuring Process: HIT task type, number of assignment, average, median time spent and standard deviation

Refine previous output to address feedback

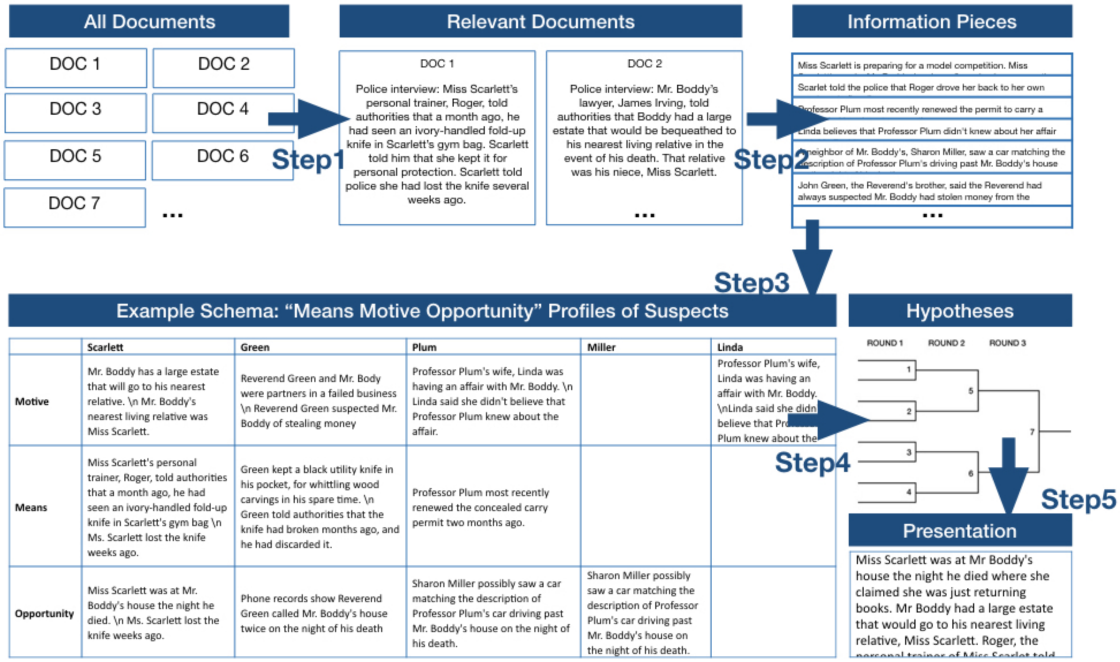
for each HIT	Step 1	Step 2	Step 3	Step 4	Step 5
worker 4	re-rate	refine	re-tag	re-select	refine
worker 5	re-rate		re-tag	re-select	
worker 6	re-rate		re-tag	re-select	
# HIT	4	4	4	3	1
# assignment	12	4	12	9	3
Avg Time	6.26	2.44	7.08	21.61	2.54
median	2.31	2.44	2.03	10.60	1.68
SD	13.0	0.68	13.59	23.52	0.97

Table 2: Evaluating Process: HIT task type, number of assignment, average, median time spent and standard deviation

5.5 Result

A total of 211 crowd workers were hired to analyze the dataset. In the first pilot with simple mystery dataset, crowd workers at each step performed well enough to generate the right solution to the mystery. We found that despite some minor flaws in intermediate output (mostly false positive), the extra useless information will stay isolated and not influence the final result as the data get passed along the pipeline. In second pilot (deliberate problem fixing), we found that the crowds are able to address the feedback and provide constructive analysis that helps direct the pipeline.

We first process the simple dataset with bottom-up structuring process. The crowds successfully rate all 11 relevant documents as relevant (100%), and extracted all useful information pieces (*”Professor Plum’s wife, Linda was having an affair with Mr. Boddy.”*), along with non-perfect or less meaningful ones (*”Police confirm the books were found on the premises of Mr. Boddy’s house.” is not directly contributing to the conclusion*). Despite this ”Evidence File” with minor false positive flaws, the information pieces are categorized correctly. This median that not only the useful information is tagged with the right labels, but the false positive flaws are also isolated,



DOC 1

Police interview: Miss Scarlett's personal trainer, Roger, told authorities that a month ago, he had seen an ivory-handled fold-up knife in Scarlett's gym bag. Scarlett told him that she kept it for personal protection. Scarlett told police she had lost the knife several weeks ago.

DOC 2

Police interview: Mr. Boddy's lawyer, James Irving, told authorities that Boddy had a large estate that would be bequeathed to his nearest living relative in the event of his death. That relative was his niece, Miss Scarlett.

...

Miss Scarlett is preparing for a model competition. Miss

Scarlet told the police that Roger drove her back to her own

Professor Plum most recently renewed the permit to carry a

Linda believes that Professor Plum didn't knew about her affair

neighbor of Mr. Boddy's, Sharon Miller, saw a car matching the description of Professor Plum's driving past Mr. Boddy's house

John Green, the Reverend's brother, said the Reverend had always suspected Mr. Boddy had stolen money from the

...

	Scarlett	Green	Plum	Miller	Linda
Motive	Mr. Boddy has a large estate that will go to his nearest relative. \n Mr. Boddy's nearest living relative was Miss Scarlett.	Reverend Green and Mr. Body were partners in a failed business \n Reverend Green suspected Mr. Boddy of stealing money	Professor Plum's wife, Linda was having an affair with Mr. Boddy. \n Linda said she didn't believe that Professor Plum knew about the affair.		Professor Plum's wife, Linda was having an affair with Mr. Boddy. \nLinda said she didn believe that Pr Plum knew about the
Means	Miss Scarlett's personal trainer, Roger, told authorities that a month ago, he had seen an ivory-handled fold-up knife in Scarlett's gym bag \n Ms. Scarlett lost the knife weeks ago.	Green kept a black utility knife in his pocket, for whittling wood carvings in his spare time. \n Green told authorities that the knife had broken months ago, and he had discarded it.	Professor Plum most recently renewed the concealed carry permit two months ago.		
Opportunity	Miss Scarlett was at Mr. Boddy's house the night he died. \n Ms. Scarlett lost the knife weeks ago.	Phone records show Reverend Green called Mr. Boddy's house twice on the night of his death	Sharon Miller possibly saw a car matching the description of Professor Plum's car driving past Mr. Boddy's house on the night of his death.	Sharon Miller possibly saw a car matching the description of Professor Plum's car driving past Mr. Boddy's house on the night of his death.	

Miss Scarlett was at Mr Boddy's house the night he died where she claimed she was just returning books. Mr Boddy had a large estate that would go to his nearest living relative, Miss Scarlett. Roger, the personal trainer of Miss Scarlet told

Figure 3: Data flow in the first pilot study using crowdsourcing

either because the information piece does not provide clear evidence about any suspect, or of poor quality, thus receiving three different tags or ”None” tag, either case that information piece won’t be tagged by majority vote thus cannot be pushed to the next step. However, Step3 saw new false positive tags. Crowds are observed to give more tags than expected by the authors. For example, the information piece ”About a month ago, Roger saw Scarlett putting an ivory-handled fold-up knife in her handbag from the gym bag.” received both medians and motives tags of Scarlett by two out of three crowd workers (different two workers). Although we do not see how this information piece provides evidence for Scarlett’s motive, having an extra tag will not influence the number of information pieces under the profile of Scarlett. In Step4, we rank the profiles by the number of categories, break the tie with the number of information pieces. Since we had only five suspects with supporting evidence to form a profile, the suspect ranked at the third place (Plum) are waived from the first round competition. With the two winners of the first round competition (between two pairs), the second round compares each pair of the three remaining suspects to find the winner. In the final step, the crowds connect the winner profile (Scarlett) to known global contexts and present a narrative reconstruction of crime assuming Scarlett is the killer.

The crowds found the right murderer and present logical narrative as a final product. In order to experiment with the crowd’s ability to fix imperfect intermediate results and address expert feedback, we run a second pilot study. We learned from the first pilot that the pipeline is robust and tolerates false positive errors. The second pilot will focus on the false negative flaws, and control the type of errors by providing a deliberately flawed input that might be possible with the more complicated dataset, following the challenges listed in previous sections.

In the second pilot, we found that the crowds are able to retrieve only the missing documents ”left out” in the previous analysis, as well as give a constructive explanation of their choice. One of the crowd workers rates an irrelevant document as 0, saying that ”It is a permit for a concealed handgun and issued to Reverend Green. This has nothing to do with Professor Plum and is certainly not related to a knife.” Another crowd worker rate a relevant document as 50, explaining the rationale as ”The document provided is relevant towards solving the mystery as it provides possible medians to kill Mr. Bobby.” These example explanations provide rich perspectives and reasoning details that are helpful for expert analysts and exhibit the crowds great potential for complicated analytical tasks.

6 Discussion and Upcoming Work

Our pilot studies show great potential of the crowds to accomplish complicated analytical tasks in a short amount of time. On the other hand, quality of work varies among steps and crowd workers. We discuss the lessons learned in the pilot studies and plans for upcoming work.

6.1 Time Efficiency

In table 1 and table 2, we observe several patterns. First, the time spent on each step is relatively stable, in both first-time creation process and later refining the process. Step 4 is most time-expensive, which is natural given that it deals with the most amount of information (two profiles with three categories of evidence, whereas the previous three steps only deal with one or two pieces of document or information). Creation tasks for Step5 are second most time-expensive, where crowds aggregate evidence for the winner suspect profile and create a narrative explaining the connections. Reviewing/refining tasks took less time than first-time creation, which aligns with previous findings in crowdsourcing research. Furthermore, we observe for each HIT, the amount time each crowd worker took to finish can be diverse. This might be because some crowd workers were idling when they work on the task, or they were thinking deeper than others. Thus, the median time spent for each HIT is more representative.

6.2 In-depth Qualitative Analysis on Explanation

In Step 1 and Step 4, we asked the crowds to provide explanations for their rating/choice. In the refining process, we also ask them to explain their refinement or reason for not being able to refine. We found that most of the crowd workers provided a logical and in-depth explanation to justify their results, and there are some inspiring points as well. One worker in Step 4 provided a long, detailed explanation *”I think that Miss Scarlet had a more convincing incentive to kill Mr. Boddy: inheriting a large estate. By contrast, Reverend Green’s possible motive would have been anger over suspicion that Mr. Boddy stole money from their business - it doesn’t look as serious an incentive to kill someone. It is also established Miss Scarlett was at the house of Mr. Boddy the night of the murder.”* that justified why this worker weighed the evidence differently, as well as a comprehensive overview he or she had for the two suspects.

6.3 Threshold Tuning

In Step 1, we asked the crowds to rate relevance in a 0-100. This allows a wider spectrum of document relevance and flexibility of further tuning on the threshold. When experimenting with a more challenging dataset, we will be able to test the precision-recall value with a different threshold on rating score.

6.4 Expert User Interface

In our pilot studies, we manually import and export input and result in CSV document, and embed expert feedback through directly coding on HTML files. We will develop an expert UI that will: a. read raw documents uploaded by expert/customer b. allow expert/customer to set global context c. automatically transform output from each step to input for next step d. control the next step to execute by expert e. automatically detect the appropriate next step if no experts provided guidance f. tuning the tags employed to organize evidence g. visualize the intermediate and final output by crowds.

References

[1] Attfield, S., and Blandford, A. Improving the cost structure of sensemaking tasks: Analysing user concepts to inform information system design. *Human-Computer Interaction–Interact 2009* (2009), 532–545.

[2] Attfield, S., and Blandford, A. Making sense of digital footprints in team-based legal investigations: The acquisition of focus. *Human–Computer Interaction 26*, 1-2 (2011), 38–71.

[3] Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. *Communications of the ACM 58*, 8 (2015), 85–94.

[4] Bier, E. A., Card, S. K., and Bodnar, J. W. Principles and tools for collaborative entity-based intelligence analysis. *IEEE Transactions on Visualization and Computer Graphics 16*, 2 (Mar. 2010), 178–191.

[5] Bradel, L., Endert, A., Koch, K., Andrews, C., and North, C. Large high resolution displays for co-located collaborative sensemaking: Display usage and territoriality. *Int. J. Hum.-Comput. Stud. 71*, 11 (Nov. 2013), 1078–1088.

[6] Bradel, L., Self, J. Z., Endert, A., Hossain, M. S., North, C., and Ramakrishnan, N. How analysts cognitively connect the dots. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on* (2013), IEEE, pp. 24–26.

[7] Chan, J., Dang, S., and Dow, S. P. Ideagens: Enabling expert facilitation of crowd brainstorming. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion* (New York, NY, USA, 2016), CSCW '16 Companion, ACM, pp. 13–16.

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

[9] Duarte, E. F., Oliveira, E., Côgo, F. R., and Pereira, R. Dico: a conceptual model to support the design and evaluation of advanced search features for exploratory search. In *Human-Computer Interaction* (2015), Springer, pp. 87–104.

[10] Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P., and Andrews, C. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems 43*, 3 (2014), 411–435.

[11] Fisher, K., Counts, S., and Kittur, A. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 247–256.

[12] Hahn, N., Chang, J., Kim, J. E., and Kittur, A. The knowledge accelerator: Big picture thinking in small pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016), CHI '16, ACM, pp. 2258–2270.

[13] Hearst, M. A. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (Stroudsburg, PA, USA, 1999), ACL '99, Association for Computational Linguistics, pp. 3–10.

[14] Karger, D. R., Oh, S., and Shah, D. Efficient crowdsourcing for multi-class labeling. *ACM SIGMETRICS Performance Evaluation Review 41*, 1 (2013), 81–92.

[15] Kim, J., Sterman, S., Cohen, A. A. B., and Bernstein, M. S. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Design Thinking Research*. Springer, 2018, pp. 79–104.

[16] Kittur, A., Peters, A. M., Diriye, A., Telang, T., and Bove, M. R. Costs and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), ACM, pp. 2989–2998.

[17] Klein, G., K Phillips, J., L Rall, E., and A Peluso, D. A data-frame theory of sensemaking, 01 2007.

[18] Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. Turkit: tools for iterative tasks on mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation* (2009), ACM, pp. 29–30.

[19] Liu, Z., and Stasko, J. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE transactions on visualization and computer graphics 16*, 6 (2010), 999–1008.

[20] Parikh, D., and Zitnick, C. Human-debugging of machines. *NIPS WCSSWC 2*, 7 (2011), 3.

[21] Pirolli, P., and Card, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. 2–4.

[22] Quinn, A. J., and Bederson, B. B. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2011), CHI '11, ACM, pp. 1403–1412.

[23] Riedl, M. O., and Young, R. M. From linear story generation to branching story graphs. *IEEE Computer Graphics and Applications 26*, 3 (May 2006), 23–31.

[24] Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. 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), CHI '93, ACM, pp. 269–276.

[25] Sedig, K., and Parsons, P. Interaction design for complex cognitive activities with visual representations: A pattern-based approach. *AIS Transactions on Human-Computer Interaction 5*, 2 (2013), 84–133.

[26] Stasko, J., Görg, C., and Liu, Z. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization 7*, 2 (Apr. 2008), 118–132.

[27] Verroios, V., and Bernstein, M. S. Context trees: Crowdsourcing global understanding from local views. In *Second AAAI Conference on Human Computation and Crowdsourcing* (2014).

[28] Wang, J., Kraska, T., Franklin, M. J., and Feng, J. Crowder: Crowdsourcing entity resolution. *Proc. VLDB Endow. 5*, 11 (July 2012), 1483–1494.

[29] Wright, W., Schroh, D., Proulx, P., Skaburskis, A., and Cort, B. The sandbox for analysis: Concepts and methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2006), CHI '06, ACM, pp. 801–810.

[30] Yampolskiy, R. Turing test as a defining feature of ai-completeness. *Artificial intelligence, evolutionary computing and metaheuristics* (2013), 3–17.