Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods

Aritra Dasgupta, Joon-Yong Lee, Ryan Wilson, Robert A. Lafrance, Nick Cramer, Kristin Cook and Samuel H. Payne

Abstract—Combining interactive visualization with automated analytical methods like statistics and data mining facilitates data-driven discovery. These visual analytic methods are beginning to be instantiated within mixed-initiative systems, where humans and machines collaboratively influence evidence-gathering and decision-making. But an open research question is that, when domain experts analyze their data, can they completely trust the outputs and operations on the machine-side? Visualization potentially leads to a transparent analysis process, but do domain experts always trust what they see? To address these questions, we present results from the design and evaluation of a mixed-initiative, visual analytics system for biologists, focusing on analyzing the relationships between familiarity of an analysis medium and domain experts' trust. We propose a trust-augmented design of the visual analytics system, that explicitly takes into account domain-specific tasks, conventions, and preferences. For evaluating the system, we present the results of a controlled user study with 34 biologists where we compare the variation of the level of trust across conventional and visual analytic mediums and explore the influence of familiarity and task complexity on trust. We find that despite being unfamiliar with a visual analytic medium, scientists seem to have an average level of trust that is comparable with the same in conventional analysis medium. In fact, for complex sense-making tasks, we find that the visual analytic system is able to inspire greater trust than other mediums. We summarize the implications of our findings with directions for future research on trustworthiness of visual analytic systems.

Index Terms—trust; transparency; familiarity; uncertainty; biological data analysis.

1 Introduction

Increasing variety and complexity of real-world data has led to a greater adoption of interactive visualization as a medium for accelerating data-driven search and discovery. However, one barrier to such adoption in scientific communities is the perceived lack of trust of domain experts in these cutting-edge visualization tools as opposed to conventional analysis mediums. Broadly, there are two reasons for such a perception. First, many scientists are accustomed to manual data analysis, and might reflexively trust the results from their familiar methods more than the ones based on advanced data-driven analytical methods and visualizations. Second, when integrated with statistical and automated methods, visual analytics processes can be prone to uncertainty and thereby lead to a lack of confidence in domain experts' decision-making and knowledge generation [18].

Currently, there is little empirical research in visual analytics that investigates the role and influence of domain experts' trust in an analytical medium on their knowledge generation process. To fill this gap, in this paper we study the interaction between two factors that can affect the trustworthiness of analytical mediums: familiarity of domain experts with the medium and their perceived confidence in the eventual analysis outcome. Researchers have recently pointed out the need to understand how uncertainty in the analysis process plays a role in the eventual trust that domain experts have in their decision-making [30]. While most of visualization research has focused on representing data-space numeric uncertainty, much less attention has been paid to how visualization tools can better tackle undertainty that stems from the potential outcomes of analysis processes or their implications [22]. The less the analytical uncertainty, the greater is the confidence of domain experts in their analysis outcome. Two key research gaps emerge in this context: 1) *what design criteria should visual analytics systems fulfill for ensuring a high level of domain experts' trust?*, and 2) *can a carefully designed, transparent visual analytic worlflow inspire a higher level-of-trust in analysts as compared to more traditional analysis mediums?*

We address these open questions by reporting on the design and evaluation of a visual analytic workflow in the context of biological data analysis for better understanding the causal factors behind domain experts' trust in an analysis medium. Our methodology consisted of two distinct stages. In the first stage we studied how explicit consideration for trustworthiness of a visual analytic workflow can improve the scientific data analysis process of domains scientists, in terms of developing and exploring alternative hypotheses on the fly, reasoning about the significance of their findings, and the search for evidence to establish the implications their findings. In the second stage, we evaluated this workflow with a larger pool of experts by comparing the levels of trust using the visual analytic tool and covnetional analysis mediums like *R*, *Excel*, etc.

These two stages of design and evalulation of the visual analytic workflow led to three main contributions. As part of our first contribution, we transformed a set of manual biological data analysis processes into a mixed-initiative visual analytic workflow (Figure 1). We collaborated with computational, molecular, and cell biologists for designing and implementing this workflow, which reflects explicit design decisions for maintaining a high level-of-trust in the system. Related to this contribution, we define trust in the context of visual analytics of biological data and describe the associated design criteria for the workflow.

For comparing the scientists' level-of-trust using the visual analytic workflow, we use their conventional analysis methods (Excel, R, etc.) as the baseline. In our second contribution, we describe the design of a quantitative user study with 34 domain scientists for evaluating their level-of-trust. From the broader visual analytic workflow, we

- *Aritra Dasgupta is with Pacific Northwest National Laboratory. E-mail: aritra.dasgupta@pnnl.gov.*
- *Joon-Yong is with Pacific Northwest National Laboratory. E-mail: joonyong.lee@pnnl.gov.*
- *Ryan Wilson is with Pacific Northwest National Laboratory. E-mail: ryan.wilson@pnnl.gov.*
- *Robert A. Lafrance is with Pacific Northwest National Laboratory. E-mail: robert.lafrance@pnnl.gov.*
- *Nick Cramer is with Pacific Northwest National Laboratory. E-mail: nick.cramer@pnnl.gov.*
- *Kristin Cook is with Pacific Northwest National Laboratory. E-mail: kris.cook@pnnl.gov.*
- *Samuel Payne is with Pacific Northwest National Laboratory. E-mail: samuel.payne@pnnl.gov.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

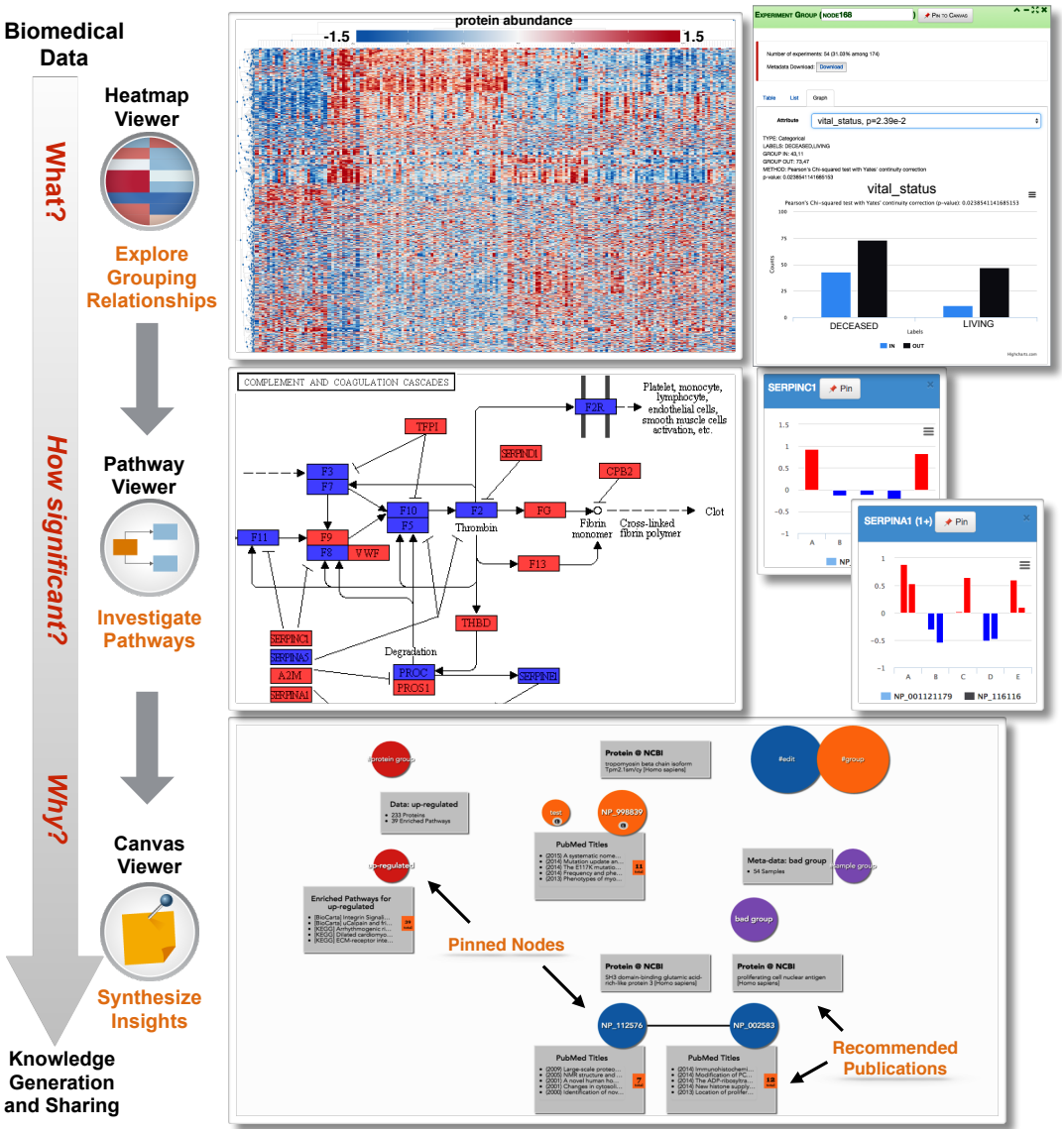

Fig. 1. **Representing the different stages in the visual analytic workflow.** The *Active Data Biology* system provides an end-to-end, transparent visual analytic workflow for biologists to seamlessly shift between verification of alternative hypotheses and generation of scientific knowledge. There are main views (Heatmap, Pathway, and Canvas viewers) to display data in different visual contexts integrated in a workflow that seamlessly connects hypotheses, reasoning, and evidences of findings for inspiring a high level of trust in domain experts. Tutorials online at `https://adbio.pnnl.gov` demonstrate the tool's use.

distilled a set of tasks requiring different degrees of interpretation and domain knowledge, and used these tasks for a between-subjects study design. The tasks were chosen carefully such that they were equally achieveable using both visual analytic tools and conventional tools like R, Excel, etc.

As our final contribution, we present the results of our quantitative study and consider the causes and implications of the variance in the levels-of-trust. We analyze the effect of domain experience, familiarity with manual data analysis medium, and task complexity on the level-of-trust of domain scientists. Finally, we reflect on whether the explicit design criteria for reducing the lack of trust in a new tool can compensate for the lack of familiarity and learning curve associated with that tool.

2 Related Work

We discuss three overlapping threads of research relevant to the work reported here.

2.1 Evaluation of Human-Machine Trust

Using criteria for interpersonal trust among humans [26], researchers have demonstrated similar criteria can be used for qualifying human-machine trust [3, 21]. While there trust is essentially a multidimensional expression based on different factors, in this work look exclusively look into the relationship between familiarity and trust. For different software systems, researchers have shown that people tend to trust familiar mediums more than other mediums, which might be more efficient in solving the tasks [13]. Researchers have also demonstrated the positive effect of transparency of decision support systems on the self-calibration of analysts' trust [10].

McAllister postulated a set of survey questions [23] for defining the different dimensions of trust, which were modified by Takayama et al. [37] to evaluate system administrator's trust in command line interfaces as opposed to software systems. We used these dimensions of trustworthiness for designing a visual analytics system for biologists. Following the principle outlined by Ugiirala et al. [40] we measure the level-of-trust as a self-calibrated measure of analysts' confidence

in the analysis output, and compare the variation with respect to the same using a more familiar analysis medium.

To the best of our knowledge, this is a first attempt at evaluating analysts' trust in visual analytic systems. In a domain science context, these systems need to support high-level sense-making tasks [17] involving the human cognition system. Simply calibrating a system's performance with respect to accuracy and efficiency is not sufficient [4]. This had also been pointed out by Scholtz who argued for moving beyond simple usability metrics and look at development of new metrics based on trust [32].

In a recent classification of existing evaluation practices in visualization, Lam et al. [20] proposed seven scenarios according to which visualizations can be evaluated. Our work fits into both the *understanding environments and work practices* (UWP) and *user experience* (UE) categories and bridges the gap between domain experts' analysis process and the visualization. We differentiate our approach from the work of Saraiya et al. [31], who performed open-ended analysis of user-defined insights generated by a biological data visualization tool. We standardized a set of tasks for comparing perceived trust between two different mediums.

2.2 Relationship Between Trust and Visualization

Compared to other areas of research, the issue of trust has been relatively underexplored in visual analytics. A recent work by Sacha *et al.* [30] describes the inherent link between different forms of uncertainty in visualization and how that can have an effect on analysts' trust. While visual analytics research on uncertainty has been mostly focused on representing and evaluating data-space uncertainty through visualizations, lot less attention has been paid to how different stages of the visual analytic process can introduce uncertainty [22, 36], and how these can interfere with the analytical reasoning process [38]. In one of the few instances of this line of research, visual analytic solutions have been developed for exploring the relationship between trust and interpretability when using analytical models that can act like black-boxes [6]. Injecting transparency by verifying and validating the model outputs has been pointed out as an important factor towards gaining analysts' trust. To this effect, researchers have pointed towards the need for building verifiable visualizations by quantifying and representing error and controlling the uncertainty in the data space [11].

In our work we handled the issue of trust in two phases. First, we followed a participatory design process with a group of expert biologists to understand the causes behind potential lack of trust, and provide support for understanding the analytical uncertainty that can be caused by different stages of data transformation in the visualization pipeline. We use visual representations that biologists are familiar with and allow interactions that capture both low-level goals (e.g., find median, extremum) [1] and high-level sensemaking tasks (e.g., find evidence for supporting hypotheses) [2]. Second, we conducted a user study with a large group of biologists to compare level-of-trust in visual analytic mediums, as opposed to conventional mediums like *Excel* and *R*.

2.3 Visualization Tools for Biomedical Data Analysis

Biomedical data visualization has a long history and several researchers have proposed solutions for analyzing grouping information from the data. The tools most relevant to this work [33, 27, 39] provide web-based interactive heatmaps and help analysts detect interesting patterns. In contrast, our focus in this work was to provide an end-to-end solution through a mixed-initiative visual analytics interface. The existing tools lack several important capabilities for: a) proactively supporting statistical analysis and on-the-fly verification of hypotheses, b) linking the data to external knowledge bases that domain experts often need to find support for their findings, and c) the flexibility to shift between the exploration and verfication stages of analysis, and maintaining analytical provenance of the results. We aimed to address all these needs through a seamless workflow with the ability to switch between the high-level goals of detecting interesting patterns, investigating their significance, and finally synthesizing insights about those patterns for knowledge generation.

We adopt the design principles for a mixed-initiative visual analytics system [15, 7, 12] and implement them in a biological data analysis workflow. Our main motivation was to facilitate a transparent analysis process, where the domain scientists' analytical questions are easily translated to operations within the interface, and the system proactively computes the statistical significance of relationships and searches external knowledge bases for recommending patterns and artifacts that can be of interest to the scientist. In a true spirit of human-in-the-loop visual analytics[19, 38], the scientist is in the control of analysis, while the system supports [25] the evolving goals and questions through analytical computations and visual representations.

3 Visual Analytic Workflow Design

Design principles for our visual analytic tool, Active Data Biology, emanate directly from our extensive experience in bioinformatics data analysis. Working for more than 10 years in dozens of collaborations, the bioinformatics co-authors acted as a liaison [35] between visualization designers and biologists at the United States Department of Energy (DOE) operated Pacific Northwest National Laboratory. Together we had several focused discussions for understanding the gaps in the current analysis tools for environmental and biomedical data. The discussions consistently identified the need for novel visual analysis tools, as the existing tools mostly focused on data processing and transformation, and less on data interpretation. But there was also skepticism around the adoption of new tools due to a perceived lack of domain experts' confidence in their outputs. This was consistent with empirical studies that demonstrated that domain experts tend to trust familiar interfaces as part of their daily routine [37].

To address this problem, our goal was to facilitate data interpretation and hypothesis generation in a transparent manner that inspires confidence in the scientists. We engaged the group of experts on biology in a participatory design process for first understanding the analysis goals, producing intermediate prototypes and refining those designs with an exclusive focus on minimizing their lack of trust in the analysis outcomes. The design process was carried out in consultation with a bioinformaticist (one of the co-authors of this paper) and an expert in the area of cancer research. In this section, we report on the different aspects of the design process by discussing the high-level analysis goals which we distilled from our discussions, our definition and considerations for increasing scientists' level-of-trust, and the eventual design of our *Active Data Biology* system that resulted from the mapping between the goals and design heuristics.

3.1 Analysis Goals

In big data experiments across many scientific domains, the first task is often to identify which individual samples are most similar and group them. In a clinical study, this might be finding similar patients; in an environmental study this could be finding similar responses of the ecosystem to a variety of perturbations. Once groups have been established, investigators frequently look to identify which proteins and biological functions distinguish groups. Finally, these characteristic proteins and biological functions are interpreted to understand how they might contribute to the observed response (phenotype) of the sample.

Several biological terms may be unfamiliar yet critical to understand the function of Active Data Biology, and so we define them here:
Proteomics: A *proteomics* experiment measures protein abundance levels in a biological sample. *Co-expression* is used to denote that two or more measurements correlate highly across all the samples. Groups of co-expressed proteins are often used to show patterns in large-scale data.

Pathway: A *pathway* is a logical group of proteins which participate together to achieve a biological function. For example, nine different proteins are used to convert glucose to pyruvate in a process called glycolysis. These proteins and several others are grouped and labeled as members of the Glycolysis pathway. The KEGG (Kyoto Encyclopedia of Genes and Genomes) database categorizes genes into pathways, and serves as a widely used reference for pathway membership [16].

Metadata: *Meta-data* is information collected about experimental design or experimental samples. In a clinical study, meta-data typically

refers to the information about a patient (e.g. age) or the tissue sample (e.g. tumor stage).

Answering an analysis goal typically has three steps:

What?: Scientists are mainly seeking for biologically meaningful information patches from raw datasets and its meta-data.

How significant?: Once scientists have identified interesting patterns, this is followed by the interpretation step which involves high-level cognition and sensemaking. They either confirm, reject, or build a hypothesis and then statistically test it within given datasets.

Why?: To investigate potential cause/effect relationships, analysts place these findings within a context of biological knowledge, by searching through published literature or external knowledge bases for analyzing the novelty of their findings.

Our collaborators informed us that all these three goals are usually pursued together as part of a biological data analysis routine. For example, to understand what molecular changes are associated with cancer outcomes, a researcher would attempt to identify *What* proteins are highly expressed in tumors for patients who die. They run a battery of statistical tests to determine *How significant* this association is. Finally they explore published literature to place these proteins into context and ponder *why* they might be associated with poor prognosis.

3.2 Design Criteria for Increasing Trustworthiness

We define the *level-of-trust of domain scientists as their self-calibrated degree of confidence in their analysis outcome that is produced in course of their interactions with any data analysis medium.* Differing from Chuang et al.'s definition of trust as perceived accuracy [6], we model trust as a confidence measure, as interpretation of patterns from a biological data analysis process can be subjective in athe absence of ground truth and that can vary across domain experts. Therefore, we used a self-calibrated measure that would account for the individual differences. Moreover, from our interactions with the domain scientists, we found that confidence plays a critical role [28] in helping analysts transition between the information seeking or foraging and insight synthesis and knowledge generation steps of visual analytics [29].

Such confidence can be reduced due to the inherent analytical uncertainty at the different stages of the scientific data analysis process, where scientists constantly need to shift perspectives between hypotheses generation and reasoning and evidence gathering for their findings. Amar et al. termed this as a rationale gap [2] in visualizations where sufficient support is not provided for the analyst to explain the significance of the detected relationships. Accounting for the lack of confidence is particularly needed during scientists' interaction with a mixed-initiative system, as a domain expert might hesitate to rely on a decisions taken by the system, thereby reducing its trustworthiness. The design criteria for instantiating a visual analytic workflow was primarily motivated by the need to reduce the rationale gap and minimize the lack of credibility about the detected patterns [36] by inspiring a high level-of-trust in biologists.

In this section we describe the design criteria that were used to ensure a high level of trustworthiness of the Active Data Biology tool. We took inspiration from the trust-centric questionnaire proposed by McAllister [23] and aimed to map the criteria that are relevant for a visual analytic tool to the domain-specific goals. We specifically focused on the following design criteria for minimizing the lack of trust. These criteria were distilled from the discussions we had with experts during the initial phase of the discussions, and were refined during the intermediate prototype development stages.

Intuitiveness: Intuitive data visualization is critical for biomedical data analysis. In course of our interactions with biomedical scientists who are more familiar with traditional non-visual analysis methods, prefer visual representations that are easily interpretable and actionable. We also found that non-computational scientists, who use visualization as an analysis medium, prefer familiar representations over new ones as the latter might entail a learning curve. To enable scientists pursue the *what* question, our goal was to design intuitive graphical representations of clustered data that could trigger the interest for detailed exploration of the data. A second important consideration for intuitiveness is that tasks and capabilities of the visual tool accurately and simply reflect the desired tasks of scientists. From our interaction with domain scientists, we know that looking at a specific visual representation of the data often provokes a question. We designed our tool to answer those exact questions, where there was a perfect match between the scientific questions and the visual cues.

Transparency: Visual analytics processes are prone to uncertainty at different stages. To reduce the amount of uncertainty that can emanate from various levels of data transformation [30, 8, 9], we provide statistical support for confirming or rejecting different hypotheses so that scientists could focus on the how significant question. Moreover, these statistical methods are prominently and clearly described in the visual display. Transparency is essential because it provides the assurance that if the researcher wants to know methodological details, they can easily find them. Without transparency a significant fraction of users (based on our pre-design surveys) get distracted and lost trying to understand exactly what happened. This inhibits their broad exploratory interactions that are an essential feature of the foraging step.

Efficient context-switching: Scientists at any given point-of-time would want to switch between different data perspectives according to the three analysis goals. By converting these goals into a workflow, we enable rapid context-switching according to the evolving mental model about the data. We also let scientists drill-down to various levels of detail through interaction and keep track of their findings, ensuring a provenance-enabled analysis process. Because of the immense dimensionality of the data, it is essential that scientists can switch effortlessly between the overview and the contextual details all the while saving insight for later consideration. The different views in the Active Data Biology tool help integrate analytical and scientific contexts within a single workflow and maintain provenance of the analysis process.

Evidence presentation: For synthesizing their findings, scientists have to reconcile multiple sources of information like KEGG databases, publication record, etc. This can be time-consuming and not getting the right evidence can reduce their confidence in the findings. We address this by following a mixed-initiative design approach [7]. Active Data Biology gathers evidence from external knowledge bases to assist users and allow them more time to pursue scientific questions and investigate the *why* aspects of their findings.

3.3 System Components

We designed Active Data Biology tool as a visual analytic tool based on Active Data Environment [7], which carefully considers design guidelines to support the complex cognitive process of data analysis. From our intimate experience collaborating on many biomedical projects, we distilled a set of exploratory data analysis tasks that can support scientific discovery. This includes: identifying enriched pathways in a set of proteins, identifying meta-data that distinguishes one group of samples from another, viewing data directly on a pathway of choice, and finding literature that is relevant to a hypothesis as it emerges from the data. Each of these foundational tasks along with the design criteria described earlier were used to guide the design of Active Data Biology. Explicitly, we wanted these tasks to be single click or easier, resulting in fluid navigation and rapid context-switching across perspectives.

Active Data Biology (http://adbio.pnnl.gov) is a web-based visual analytic tool (Figure 1) suite that allows data exploration within familiar biological contexts such as heatmaps and metabolic pathways. This tool provides fluid navigation and easy collaboration with three main views. Each view provides a unique perspective on the data and lets domain experts explore data within a visual context that is familiar and productive.

HeatMap: Heatmaps show a quick overview of the entire dataset of quantitative measurements. Rows represent protein measurements; columns show different samples in the experiment. The rows and columns have been grouped according to similarity, and this hierarchical clustering is shown in a dendrogram adjacent to the heatmap. The color gradient ordered from blue to red shows protein abundance values ordered from low to high. The heatmap is a ubiquitous vi-

sual metaphor for biological data presentation and provokes several questions about an experiment. While alternative visualizations like multidimensional projections could be used for displaying similarity based information, they were not considered in the design phase as our aim was to use visual representations that biologists are most familiar with, and minimize the learning curve for adopting new representations. A heatmap triggers several questions about the grouped samples. Most investigators would want to know, "Does this group have a different phenotype? "What biological functions are up-regulated in this group? These mirror analysis tasks discussed earlier, and we followed the intuitiveness design principle to assist analysts in quickly answering these questions. Any selection on the clustering dendrogram automatically searches for distinguishing characteristics of the group (Fig. 1), thereby presenting evidence about potentially interesting findings to the scientist. Groups of proteins (rows) are searched for enriched pathways. Groups of samples (columns) are searched for distinguishing meta-data. The transparency design principle led us to provide explicit statistical details for each test in an obvious yet non-intrusive place of the visualization.

Pathway: The pathway view allows users to view data within the context of biological functions. Rapid context-switching across heatmaps and pathways allows users to see the patterns in a biological context and trigger their sensemaking process. It layers data directly onto images provided by KEGG, representing molecular pathways for metabolism, genetic and environmental information processing, etc. Initially entities in the KEGG diagram are colored red to indicate that the project contains quantitative data for this gene/metabolite. Users can expand the view of any entity, which displays the quantitative data across the cohort, thus providing support for statistical analysis and resulting in a transparent exploration process (Figure 1).

Canvas: The canvas view allows users to investigate and interact with the data they have pinned from all across the other views. Users can track emerging hypotheses by aggregating and assimilating their thoughts. As entities are pinned to the canvas, automated software assistants search external knowledge sources to find relevant information, surfaced to the users in a recommendation card. Each type of data (e.g. protein, group of proteins, etc.) has different recommendation cards, reflecting the distinct information sources that curate biological information. Cards serve as a link to the external resource and can be saved or dismissed as an analyst develops their hypothesis. (Figure 1). Active Data Biology is designed to be a knowledge gathering and sharing space. All data and analyses are versioned and backed at GitHub, making sharing and collaboration natural. Each user in a collaborative project has their own canvas, so a user can customize it and make it their own. Users can see the canvas of others users and copy data from one to the other.

4 Study Design

We conducted a quantitative user study for evaluating how well our design rationales translated into practical benefits, and consequently investigate the causes and implications of variation in domain scientists' level-of-trust in the analysis medium. The main goal was to compare the level-of-trust using Active Data Biology and with that using manual analysis methods, such as Excel, R, Matlab, etc., that are more commonly used in bioinformatics. We used a between-subjects design with 34 scientists, where the level-of-trust was the dependent variable while the task type and analysis mediums were the two conditions.

We divided the participants into two groups: group **G1** performed all the tasks using the Active Data Biology tool, while group **G2** performed all the tasks using a manual analysis medium. We did carefully consider a within-subjects design. However, since we were using a single dataset, a within-subjects design had the potential to introduce a learning effect, that would have been very hard to detect across the different analysis mediums. It is also difficult to replicate similar tasks with different datasets across different analysis mediums. In a between-subjects design even if there was a learning effect, we could track its progression for the different tasks and analyze the effect across analysis mediums. In this section we describe the mapping between analysis goals and concrete tasks for the experiment, and our hypotheses and study conditions.

Task	Type	Analytic Workflow		
				
T1	Retrieval	✓		
T2	Interpretation	✓	✓	
T3	Retrieval			✓
T4	Interpretation	✓	✓	
T5	Retrieval	✓		
T6	Interpretation	✓		

Fig. 2. **Mapping the tasks to the Active Data Biology workflow**: Interpretation tasks typically need a combination of multiple visual components and are more complex than retrieval tasks.

4.1 Proteomic Data Generation

Although numerous collaborative projects were influential in the design and implementation of Active Data Biology, we chose one to use during the user study: a proteomic investigation of ovarian cancer tumors from a well-defined clinical cohort of 174 women. In addition to protein quantitation, each tumor sample has associated meta-data describing the patient and their tumor. An important goal of the ovarian proteomics study was to identify subtypes within the cohort and then to identify which proteins and biological functions correlated with specific subtypes. Finally, these characteristic proteins and biological functions were interpreted to understand how they might contribute to clinical outcome, e.g. survival.

4.2 Choice of Tasks

The most important consideration for the choice of tasks was that they had to be substantive and typical tasks when analyzing biological datasets. For mapping high-level analysis goals into concrete visualization tasks, we used the task classification schemes proposed by various researchers [1, 34, 41]. The first class of tasks belongs to the analysis goal of detecting *what* is interesting in the data and is related to the low-level tasks classification scheme proposed by Amar and Stasko [1]. To generalize these tasks into one group, we call them Retrieval Tasks. The second class of tasks of reasoning about how significant and why, overlap with the *why* class of tasks proposed by Brehmer et al [5]. We call them Interpretation Tasks and have a greater level of difficulty than the retrieval tasks. These tasks are explained as follows:

Retrieval Tasks: In retrieval tasks, we asked subjects to retrieve a part of the information from the dataset that answers a specific question. These often include low-level tasks like identifying meta-data, expression patterns, etc. Retrieval tasks should be performed by only browsing/exploring the data in spite of lack of prior domain knowledge.

Interpretation Tasks: In interpretation tasks, we asked subjects to interpret data gathered in retrieval tasks through placing it in an appropriate experimental or biological context. Interpretation tasks require a modicum of domain knowledge.

T1 *Identify the patient subgroup with the worst prognosis.*

T2 *Identify over-represented biological functions (e.g. pathways) in proteins up-regulated in this subgroup.*

T3 *Find relevant literature publications relating to this observation.*

T4 *Is there systematic bias in a subtype for tumors collected at a specific location?*

T5 *What is the median age of good prognosis subtype?*

T6 *Is it significantly different from the poor prognosis subtype?*

For all the tasks, we asked subjects to self-calibrate their level-of-trust on a 5-point Likert scale (1 = Very Uncertain and 5 = Very Certain) and also mention if there were some particular reasons about their lack of trust, or concern regarding complexity of the tasks.

The tasks were ordered based on the scientific workflow and were grouped based on their similarity and dependence: the groups were T1, T2, T3, T4, and T5, T6. Each task could be readily accomplished using the views within Active Data Biology as illustrated in Figure 2.

In **T1**, we asked participants to find an answer from clinical meta-data related to prognosis such as vital status, days to death, and so on using the heat map viewer. Based on the this answer, we asked participants to find a list of proteins having higher abundance levels in the worst subgroup than in other subgroups in **T2**, and to identify biological functions represented by these proteins using both heat map and pathway viewers. In **T3**, participants identified the relevant publications about biological functions from **T2** and whether these affect to clinical prognosis of ovarian cancer patients using the canvas view. Given the meta-data describing where these tumor samples were collected (tissue source site or TSS), participants were asked to investigate whether the tumor subtypes were an artifact caused by biased collection in **T4** using a combination of heat map and pathway viewers. **T5** was similar to **T1** but participants had to coordinate multiple meta-data values to retrieve the answer from the heatmap viewer. After retrieving the median age in **T5**, users had to determine whether there was a difference between two prognosis groups in **T6**.

4.3 Hypotheses

H1 The group using the new Active Data Biology tool will report a level-of-trust that is comparable with the group using more familiar and conventional analysis methods.

H2 For retrieval tasks, the group using the conventional analysis methods will report higher level-of-trust. For more complex tasks interpretation tasks, the level-of-trust using the Active Data Biology tool will be comparable with that using conventional analysis methods,

H3 Experience will play a role in the level-of-trust ratings. More experienced scientists will report a higher level-of-trust in conventional analysis methods as opposed to Active Data Biology.

H4 More experienced scientists will have a higher level-of-trust in interpretation tasks due to their domain knowledge, than less experienced scientists.

For our hypotheses we took into consideration the learning curve for understanding the functionalities of a new tool and expected the associated lack of familiarity to play a role in the self-calibrated levels-of-trust. However, we also expected that our explicit design decisions to minimize the lack of trust would compensate for some of the disadvantages involving the use of a new visual analytic tool. This resulted in **H1**. Similarly, in **H2** we speculate that for more complex tasks, the advantage of visualization-based methods would somewhat compensate for the lack of familiarity. Therefore, we expected to see a difference in reporting of level-of-trust between the interpretation and retrieval tasks. For **H3** and **H4**, we expected domain experience to have a significant effect on the level-of-trust. With greater domain experience, we expected participants to report a higher level-of-trust in conventional analysis mediums and with complex interpretation tasks, as they would perceive the familiar tools to be more reliable.

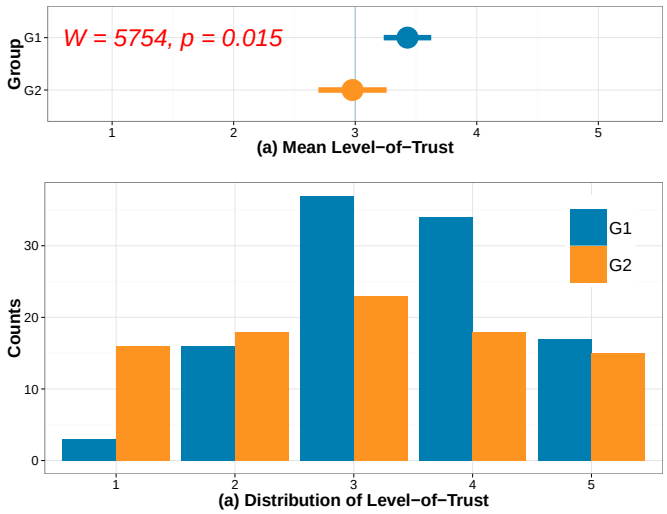

Fig. 3. **Mean level-of-trust across all tasks** with Active Data Biology user group (G1) and conventional tool user group (G2). Error bars represent 95% confidence intervals. We found statistically significant differences in the level-of-trust in the visual analytic tool, as opposed to that in familiar analysis methods.

4.4 Participants

We recruited 34 staff and interns working at the U.S. Department of Energy operated Pacific Northwest National Laboratory, who have a background in biology and computation. We recruited only those researchers who would be using the Active Data Biology workflow for the first time, and excluded the people involved in the participatory design phase to ensure all participants had equal standing. Assignment of participants into groups were mostly random. Participants were assigned s numerical *ID* and based on if that was odd or even, they were assigned to the groups: participants with an odd-numbered ID were assigned to G1, whereas those with even-numbered IDs were assigned to G2. Since our initial pilot tests revealed that a high degree of domain experience is needed to solve the tasks, especially using manual analysis methods, we aimed to recruit participants with experience with performing these tasks.

With two control groups and 34 participants we ended up with 194 trials:

$$\begin{aligned}\text{Number of trials} &= 2 \text{ task types} \\ &\times 3 \text{ tasks} \\ &\times 34 \text{ participants} \\ &= 194\end{aligned}$$

4.5 Study Settings and Procedure

At the beginning, we introduced the study to the participants and let them fill out a background survey, where they had to answer questions about their demographic information, and their experience in biology, statistical analysis, visual analytics, etc. For the group which performed the tasks using the Active Data Biology tool (G1), we first trained them on using the tool by showing them a video where different interactions and operations that were needed during the study were demonstrated. We refined the training video based on our initial pilots, and in most cases during the study, participants were confident about using the tool after looking at the vide.

For the other group we asked them to begin the study after answering the initial survey. Paricipants in group G2 were free to choose any data analysis tool (R, Excel) of their choice. Both groups were instructed that the questions in this study will ask about their level-of-trust for all the tasks, and were asked to answer all the questions as efficiently as possible without compromising on accuracy. Participants were reminded that the goal of the study is not to assess abilities

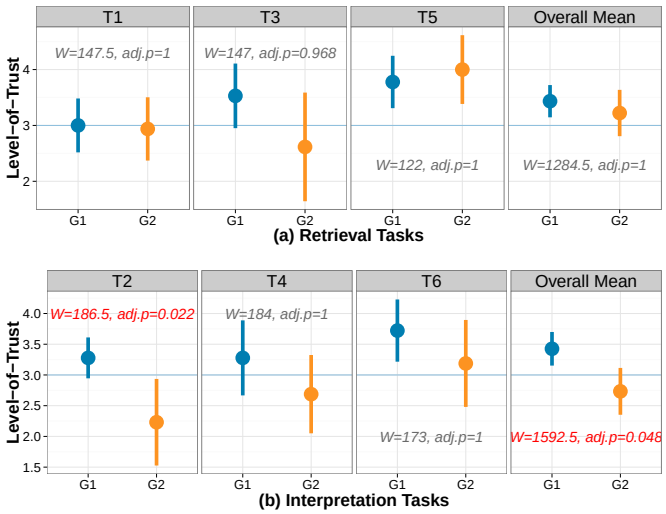

Fig. 4. **Mean Level-of-trust for two types** (retrieval and interpretation) of tasks for G1 and G2. Error bars represent 95% confidence intervals. *p*-values have been adjusted for multiple comparisons using Bonferroni correction [14].

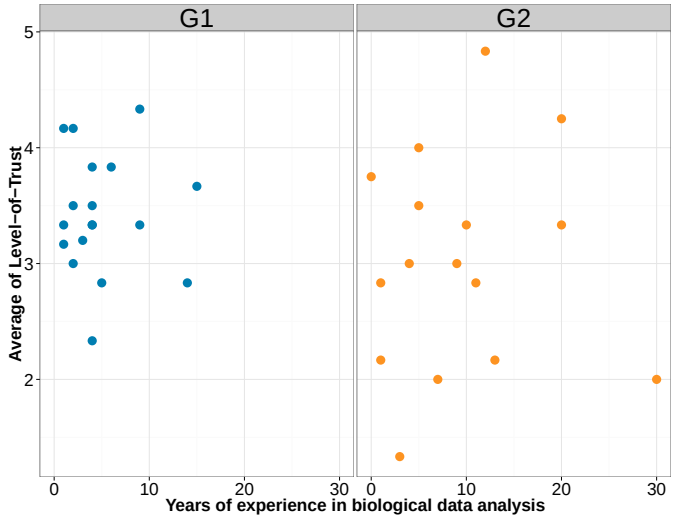

Fig. 5. **Average of level-of-trust vs. years of experience in biological data analysys for G1 and G2.** Within each group, we did not find any effect of experience on the the trust ratings.

as a data analyst, but instead find causes for variation in level of trust, which will enable us design better tools in the future.

Participants sat at a lab machine with one of the investigators sitting adjacent. The machine was a dual-monitor computer manufactured by Dell with Windows 10 operating system and Intel Pentium Quad Core processor. A dual-monitor set up was specifically used such that the participants could see the questions on one screen and use the other screen for working on the software.

The study was proctored by at least one of the co-authors. Think-aloud protocol was encouraged by the proctors: participants were encouraged to talk about any significant problems or experiences they had while performing the tasks. Due to the lengthy nature of the study, participants were encouraged to take breaks if necessary. Both groups could record their responses, indicate the level-of-trust in their response and also comment about any doubt or surprises they had during the analysis. For all the tasks, we used the same dataset across both the groups. Even if there was a learning effect due to familiarity with the data, we could measure that in terms of progressing of their level-of-trust, and assess the effect of the analysis mediums on that effect.

After the study, the participants completed a post-study questionnaire to provide their subjective assessment of their experience performing the tasks. For group G2 that performed the tasks without the Active Data Biology tool, we explained the functionality of the tool through the demonstration video once they were finished with the tasks and encouraged them to spend about 15 minutes for answering some of the questions using that tool. They indicated their feedback about the tool in the post-study questionnaire. The whole study took approximately 1.5 hours per participant.

5 Results

In this section, we describe and highlight the significant results from our tudy. First, we investigate the variation of overall level-of-trust across two different analysis mediums and compare level-of-trust based on tasks and its types. Then, we also investigate the effect of domain experience on the level-of-trust according to complexity of tasks. We have annotated the results of statistical tests in Figures (3-7), where the red-colored text indicates statistical significance.

Overall Level-of-Trust between G1 and G2: Figure 3 shows the overall level-of-trust of all the tasks for each user group (G1 and G2). Wilcoxon-Mann-Whitney test was performed to compare the distributions of ordinal level-of-trust of each user group and it shows a significant difference between G1 and G2 (W=5754, adjusted p-value=0.015). Overall level-of-trust of G1 (M=3.43, SE=1.02) was higher than level-of-trust of G2 (M=2.98, SE=1.34) in general. From the frequencies of level-of-trust, we found that G1 answered 'Uncertain' (1 or 2 for level-of-trust rating) or skipped tasks in 18.5% cases, whereas G2 did in 41.7% cases. On the other hand, G1 selected 'Certain' (4 or 5 for level-of-trust rating) for tasks in 47.2% cases, whereas G2 did in 34.4% cases. We thus found evidence to support *H1* and we believe that the trustworthiness-aware design of Active Data Biology workflow somewhat compensated for the lack of familiarity that could cause a lack of trust.

Level-of-Trust Vs Task Type: Figure 4(a) and (b) show the averages and 95% confidence intervals (CIs) of level-of-trust in retrieval and interpretation tasks, respectively. We employed Wilcoxon-Mann-Whitney test to compare the level-of-trust between G1 and G2 of each type and also adjusted p-values by Bonferroni correction for multiple comparison. As labeled in the figure, we found no significant difference between level-of-trust of G1 and G2 for a retrieval type. However we found that overall level-of-trust of G1 for an interpretation type was significantly higher than of G2 (W=1592.5, p=0.048). This finding partially supported **H2** and showed that Active Data Workflow was comparably trustworthy for retrieval analysis and more reliable when performing complex tasks. As shown in Fig. 4, we found a significant difference between G1 and G2 (W=186.5, p=0.003) in **T2**, which was the most complex interpretation task.

Level-of-Trust Vs Experience: For understanding the effect of experience, we chose a threshold for distinguishing more experienced participants from less experienced ones for each group. From the experience profiles across G1 and G2, we found that the median year of experience in biology of G1 and G2 was 4.5 years (Figure 5). We therefore made subgroups for each group as follows: we had a low-experienced subgroup and a high-experienced subgroup with participants having less than median years of experience and having equal to or greater than median years. Low and high subgroups in G1 had 12 and 6 people respectively, while low and high subgroups in G2 had 5 and 11 people repsectively.

First, we compared level-of-trust of high experienced (longer than 4.5 years) scientists with low experienced (less than 4.5 years) in G1. On average, level-of-trust of low experienced participants was slightly greater (M=3.41 SE=1.05) than level-of-trust of high experienced participants in G1. However, we found that there was no significant difference (W=1237.5, p=1.0). In the same manner, we compared level-of-trust of high experienced scientists with low experienced in G2. Also,

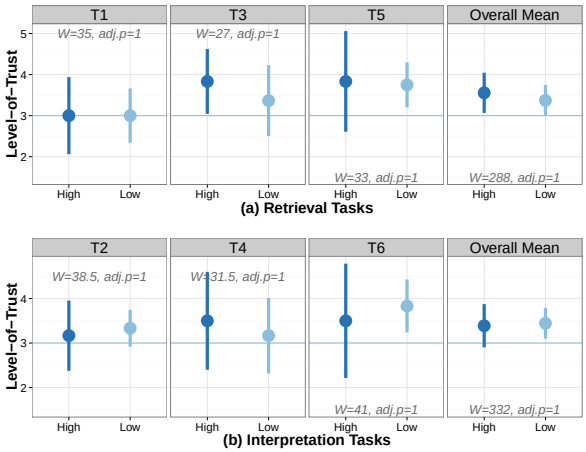

Fig. 6. **Mean Level-of-trust for two types (retrieval and interpretation) of tasks for low experienced participants and high experienced participants in G1**. Error bars represent 95% confidence intervals. We employed Bonferroni correction for multiple comparisons. The low group exhibited less variability in their trust ratings compared to the experienced group, although there were no significant differences.

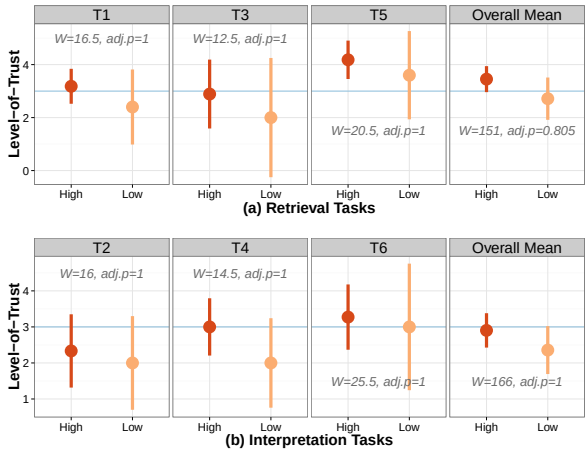

Fig. 7. **Mean Level-of-trust for two types (retrieval and interpretation) of tasks for low experienced participants and high experienced participants in G2**. Error bars represent 95% confidence intervals. We employed Bonferroni correction for multiple comparisons. Contrary to the results for G1, the high group exhibited much less variability in their trust ratings. This can potentially be attirbuted to their greater familiarity with the analysis mediums like Excel, R, etc.

we found that there was no significant difference between level-of-trust of low experienced participants and high experienced participants in G2 (W=632, p=0.144).

Interestingly, when we compared level-of-trust of low experienced scientists in G1 with low experienced in G2, we found a significant difference between low experienced scientists in G1 and G2 (W=1389.5, p=0.006). Level-of-trust of low experienced participants in G1 was significantly greater than low experienced in G2 (M=2.54 SE=1.26). On the other hand, we found no significant difference between high experienced participants in G1 and those in G2 (W=1245.5, p=1.0). In addi- tion, we found that level-of-trust of high experienced participants in G2 (M=3.18 SE=1.34) was slightly lower than overall level-of-trust of low experienced G1 on average without any significance (W = 2398, p = 1.0). Therefore we could not find evidence for supporting **H3**, which was surprising as we thought that participants with high domain experience will report significantly higher level-of-trust in the familiar manual data analysis tools.

Level-of-Trust Vs Experience and Task Type: Next we drilled down to the task types for detecting the effect of experience in each group with respect to retrieval (Figure 6(a) for G1 and Figure 7(a) for G2) and interpretation tasks (Figure 6(b) for G1 and Figure 7(b) for G2). For retrieval tasks, we found that high experienced participants had pretty equal level-of-trust to or insignificantly higher than less experienced participants in either G1 or G2. For interpretation tasks, we found that there was no significant difference between low experienced and high experienced participants in both G1 and G2. Our findings therefore did not support **H4** and showed that even with greater domain experience and familiarity with an analysis medium, for complex interpretations tasks the average level-of-trust in the visual analytics system was comparable to that in manual data analysis methods .

Post-Study Questionnaire: After the study, we asked participants several questions regarding their experience and the usability of Active Data Biology tool to which the participants responded on a 5-point Likert scale (1= Not Useful and 5=Useful) in post-study questionnaire. Overall usability as rated by G2 (M=4.81 SE=0.54) was significantly higher than that rated by G1 (W=53.5, p=0.0005). 88.2% participants gave a 4 or 5 rating for usability of overall tool. 41.2% of participants strongly agreed that '*I would incorporate the Active Data Biology tool in my daily analysis routine*' and 52.9% strongly agreed that '*I would recommend this tool to my colleagues for their work*'.

6 Analysis and Discussion

In this section, we present our analysis involving the results and subjective feedback collected as part of the post-study questionnaire and think-aloud protocol during the study.

Effect of Analysis Medium: While designing Active Data Biology, we aimed for a level-of-trust that is comparable with the same when scientists use conventional analysis methods. As we showed in Figure 3 the level-of-trust of G1 was not only comparable with G2, but the variance expressed by the lower inter-quartile range was also less in case of G1. This demonstrates that although participants were exposed to a new visual analytic tool and there was a learning curve associated with it, they had a greater agreement about their level-of-trust than those participants using a conventional analysis medium.

The advantage of Active Data Biology in terms of the greater efficiency and transparency was also widely acknowledged by most of the participants and is captured in this statement by one of them: *"This tool is much much better than trial and error in Excel. I am a biologist and not a statistician so having a visual tool like this that does the statistics under the hood would be a boon to productivity!"*

Effect of Task Complexity: The abiding design decision behind Active Data Biology was to reduce the rationale gap by letting scientists perform the high-level sensemaking and interpretation tasks efficiently with help from the mixed-initiative design of the system. In keeping with H2, we noted that for interpretation tasks (Figures 4), there was a significant difference in the level-of-trust between the two groups. We posit there is a direct effect of the reduction in rationale gap that we can observe in this result. Participants had to complete the tasks within a stipulated amount of time.

When using excel or R, they had to do the transformation/computation of the results themselves for investigating significance, which was time-consuming. In Active Data Biology, these were supported within the heatmap interface and there was efficient context-switching between detection of patterns and exploration of the evidence presented to the scientist, which potentially led to a higher level-of-trust. This was a validation for our design criteria and was also reflected in the following comment by one of the participants: *"It organized data by abundance and that helped me focus into why those abundance patterns were either high or low. It allows for more investigation that a regular heatmap would allow."*

Effect of Experience: The most surprising finding was that we failed to detect any significant effect of experience on the level-of-trust ratings. Our assumption was that experience and familiarity with

a particular analysis medium might compensate for the difficulty in task-solving using that medium. However, we demonstrated that irrespective of experience, especially for complex tasks like those involving interpretation and sensemaking, visual analytic tools due to their transparency and flexibility, are able to inspire a greater sense of confidence in the scientists.

While we do leave some room for skepticism around over-trust and bias that can happen due to inexperience, we cannot experimentally verify that in this case, and should be investigated in future studies. The enthusiasm about the comfort level with the tool was shared by most participants and is captured by the following comment by a participant from the less experienced group: *"THIS IS AMAZING! The tool will enable biologists with limited statistical knowledge and data analysis skills to begin interrogating large data sets without needing to know how to code."*

For the more experienced group, understandably there was some skepticism about the underlying analytical methods used in Active Data Biology, which captured by the following comment by a participant from the more experienced group who solved the tasks using Excel and explored the tool once the study was completed: *"Again, since you're asking about trust: To fully trust the results from your tool, I would need to spend enough time to understand what it is doing, to verify that the statistical tests it's using are the appropriate ones for the task, etc. That said, the speed and efficiency of using this are amazing. I almost couldn't afford not to use it."*

Efficiency and Usability: We received unanimously good feedback from both G1 and G2 participants about the usability. For example, one of the participants from G2 remarked *"My methods incorporated a wide variety of disparate tools that could be easily combined into a single interface. I think that the Active tool performed all of those actions in a single space."* Participants also widely appreciated the efficiency in the analysis process due to Active Data Biology, as is reflected in the following remark by a participant from G1: *"Enables immediate analysis without sorting or writing code. Huge time saver."* Compared to participants in G2, who spent a lot of time in getting familiar with the dataset and the questions, we found that participants in G1 spent lot less time in either understanding the data, or the tool. We had intentionally chosen intuitive visual representations that scientists often use with the goal of minimizing the learning curve. This was validated by informal feedback from many of the paricipants, and this remark from one of the G1 participants: *The active data tool allows you to easily zoom in on the areas that intuitively look different and this is the most frequent approach I use when I scan datasets. But it has the advantage of actually telling you something about the similarity/pathways that are conserved in that data (not just what the protein names are). Of course there is always more complexity, but this would be a great place to start.* Given the feedback about usability, we believe that cases where scientists did not have a high level-of-trust in Active Data Biology was a function of the complexity of the patterns and the associated domain knowledge that can cause skepticisim, and not a function of the lack of the usability of the tool.

7 Conclusion and Future Work

In this paper we have presented a comparative study of the level-of-trust of domain scientists in visual analytics systems as opposed to more familiar manual analysis methods. We took explicit design decisions accounting for the trustworthiness of Active Data Biology and demonstrated that these decisions had an effect on the eventual calibration of trust by the domain scientists. The most significant finding of this study is that even with new visual analytic tools, domain scientists can have a high-level of trust in those systems when the system is intuitive, transparent and lets them seamlessly switch between hypotheses generation and evidence gathering. We also demonstrated that for complex tasks, irrespective of experience and familiarity, the average level-of-trust in visual analytic systems exceeds the same in manual data analysis methods. We consider our work as a first step towards understanding the relationship between domain expert's trust and visual analytics systems, and can lead to new research opportunities.

Limitations: The findings from our work are not prescriptive of design guidelines, but descriptive in nature. We believe, while these design criteria, study methodology, and the findings have a great potential for generalizability, we need to conduct more studies with domain scientists under various conditions to understand the issue of trust from different perspectives. Second, we have modeled level-of-trust as only a degree of perceived confidence, whereas there can be many other definitions, like those suggested in the software systems and web technology domain [3]. It would be worthwhile to explore those definitions and apply them in the context of visual analytics. Third, we have not handled the scenario where there can be uncertainty propagation from the data source itself. In that case, lack of uncertainty representation can reduce domain experts' trust [30].

Research Directions: There are different research directions we can pursue for further exploring the issue of trust. In this study we have not investigated what role bias and over-trust play in domain experts' judgment [24]. Visualization is thought of a medium that can reduce bias through transparent representations. But in case of mixed-initiative systems, where the machine is making judgments on behalf of the user (for example, finding the relevant set of publications in Active Data Biology), that can lead to over-reliance on the machine. This is closely related to the balance of decision-making and allocation of functions [21] in mixed initiative systems. Conducting controlled user studies can help model the relationship between trust and functionality of such mixed-initiative systems. Motivated by our results, we now plan to conduct similar experiments in other scientific disciplines such as climate science, for investigating how adoption of visual analytics tools can be accelerated by reducing the perceived lack of trust and inspiring more confidence in the domain experts' visual analysis and exploration processes. Overall, We believe, by attacking the deep-rooted perception about potential lack of trust, the insights gained from this study will be impactful in designing future visual analytic tools for domain scientists.

8 Acknowledgement

The research described in this paper is part of the Analysis in Motion Initiative at Pacific Northwest National Laboratory (PNNL). This work was funded by the Laboratory Directed Research and Development Program (LDRD) at PNNL. Battelle operates PNNL for the U.S. Department of Energy (DOE) under contract DE-AC05-76RLO01830.

References

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, pages 111–117, 2005.

[2] R. Amar, J. T. Stasko, et al. Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):432–442, 2005.

[3] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58–71, 2007.

[4] E. Bertini, A. Perer, C. Plaisant, and G. Santucci. *BELIV: Beyond time and errors: novel evaluation methods for information visualization*. ACM, 2008.

[5] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[6] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM, 2012.

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

[8] A. Dasgupta, M. Chen, and R. Kosara. Conceptualizing visual uncertainty in parallel coordinates. In *Computer Graphics Forum*, volume 31, pages 1015–1024. Wiley Online Library, 2012.

[9] A. Dasgupta and R. Kosara. The importance of tracing data through the visualization pipeline. In *Proceedings of the BELIV Workshop: Beyond*

Time and Errors-Novel Evaluation Methods for Visualization, page 9. ACM, 2012.

[10] E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman. A design methodology for trust cue calibration in cognitive agents. In *International Conference on Virtual, Augmented and Mixed Reality*, pages 251–262. Springer, 2014.

[11] D. Fisher, I. Popov, S. Drucker, et al. Trust me, i'm partially right: incremental visualization lets analysts explore large datasets faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1673–1682. ACM, 2012.

[12] S. Gardiner, A. Tomasic, J. Zimmerman, R. Aziz, and K. Rivard. Mixer: mixed-initiative data retrieval and integration by example. In *Human-Computer Interaction–INTERACT 2011*, pages 426–443. Springer, 2011.

[13] D. Gefen. E-commerce: the role of familiarity and trust. *Omega*, 28(6):725–737, 2000.

[14] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[15] 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.

[16] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[17] Y.-a. Kang and J. Stasko. Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2869–2878, 2012.

[18] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. *Visual analytics: Definition, process, and challenges*. Springer, 2008.

[19] D. A. Keim, F. Mansmann, and J. Thomas. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, 11(2):5–8, 2010.

[20] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[21] J. Lee and N. Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.

[22] A. M. MacEachren. Visual analytics and uncertainty: Its not about the data. In *In Proceedings, EuroVA Workshop*. The Eurographics Association, 2015.

[23] D. J. McAllister. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, 38(1):24–59, 1995.

[24] G. J. McInerny, M. Chen, et al. Information visualisation for science and policy: engaging users and avoiding bias. *Trends in Ecology & Evolution*, 29(3):148–157, 2014.

[25] K. Morton, M. Balazinska, D. Grossman, and J. Mackinlay. Support the data enthusiast: Challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment*, 7(6):453–456, 2014.

[26] B. M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5-6):527–539, 1987.

[27] C. Perez-Llamas and N. Lopez-Bigas. Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS One*, 6(5):e19541, 2011.

[28] B. Phillips, V. R. Prybutok, and D. A. Peak. Decision confidence, information usefulness, and information seeking intention in the presence of disconfirming information. *Informing Science*, 17:1–24, 2014.

[29] P. Pirolli and S. Card. Information foraging. *Psychological review*, 106(4):643, 1999.

[30] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1):240–249, 2016.

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

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

[33] M. P. Schroeder, A. Gonzalez-Perez, N. Lopez-Bigas, et al. Visualizing multidimensional cancer genomics data. *Genome Med*, 5(9):10–1186, 2013.

[34] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375, 2013.

[35] S. Simon, S. Mittelstädt, D. A. Keim, and M. Sedlmair. Bridging the gap of domain and visualization experts with a liaison. In *Eurographics Conference on Visualization (EuroVis) 2015*, pages 127–131, 2015.

[36] M. Skeels, B. Lee, G. Smith, and G. G. Robertson. Revealing uncertainty for information visualization. *Information Visualization*, 9(1):70–81, 2010.

[37] L. Takayama and E. Kandogan. Trust as an underlying factor of system administrator interface choice. In *Extended abstracts on Human factors in computing systems*, pages 1391–1396. ACM, 2006.

[38] J. J. Thomas and K. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[39] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, pages 178–192, 2012.

[40] A. Uggirala, A. K. Gramopadhye, B. J. Melloy, and J. E. Toler. Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics*, 34(3):175–186, 2004.

[41] M. X. Zhou and S. K. Feiner. Visual task characterization for automated visual discourse synthesis. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 392–399. ACM Press/Addison-Wesley Publishing Co., 1998.