Leveraging Experts and the Crowd for Visualization Evaluation

PhD Thesis Proposal

Candidate: Steven R. Gomez
Department of Computer Science, Brown University
Providence, Rhode Island, USA

Revision: 2012-4-19

A Vision/Introduction

The general hypothesis in this thesis work is that evaluations for visualization can be made faster and more methodical by leveraging underutilized resources. This work will focus on three areas that have not been well explored: 1) clarifying the design space of current evaluation methods, considering emerging technologies like crowdsourcing and design models; 2) designing and validating new evaluation methods that combine expert and non-expert study participants; 3) extracting general visualization design guidelines from both domain-specific visualizations and other visual artifacts (e.g., PowerPoint presentations, or academic conference posters). Expected deliverables include a taxonomy and interface that facilitate better utilization of the evaluation design space; new paradigms and interface technologies for crowd-powered visual analysis and evaluation, as well as combined crowd+expert evaluations; and methods for extracting design guidelines from handcrafted visualizations. These contributions will be evaluated using both quantitative and qualitative methods using domain scientists and non-experts (e.g., crowd workers on Amazon Mechanical Turk, called “Turkers”). The driving concept behind this work is enabling next-generation evaluation technologies, which I call *magic-button evaluations*, that require minimal effort for the toolsmith to plan and execute.

This work is grounded in human-computer interaction. Human factors directly affect how quickly or accurately visual information can be internalized and reasoned about during hypothesis generation and validation. Some factors, like usability, can be evaluated without knowledge of the visualization itself; an example is checking whether interfaces support essential analysis operations like “undo” or mechanisms for saving or annotating progress. Domain-dependent factors also exist; for instance, a tractography visualization of brain white-matter fiber bundles might be illegible if too many bundles are drawn, or not very useful for sensemaking if too few bundles are drawn. A challenge in this work is producing methods and workflows that decompose domain-dependent evaluation tasks into unit tasks that do not require domain knowledge. If this succeeds, evaluations that typically depend on scarce populations of experts (e.g., the 2 or 3 brain scientists at a given university studying white-matter degeneration) could be scaled up and distributed across many non-experts or crowd workers. The result is stronger tool evaluations that are easier to orchestrate.

Additional contributions will be made in the scientific domains – in particular, brain science – that I will test evaluation tools in. As part of this research, I will perform application development on brain network visualizations with collaborators inside and outside Brown University. The research process will be interdisciplinary, draw on current work in human-computer interaction, cognitive psychology, both scientific visualization (“scivis”) and information visualization (“infovis”), and user interface design. As such, there is a large community that will be excited by advancing the research problems proposed in this thesis. As the “big data” era continues to increase the demand for visualizations and sensemaking tools, I believe systematically validating these tools will have high impact and facilitate broader scientific progress.

B Specific Aims

B.1 Reduce the time and effort of planning appropriate evaluation methods by building a comprehensive, searchable taxonomy of techniques.

This aim is about articulating the design space of evaluation methods for visualization, so that designers can quickly find appropriate methods to validate new tools. Expected contributions of this aim are below.

A taxonomy of visualization evaluation methods. Some design and evaluation guidelines exist for user interfaces; we will synthesize these into one taxonomy that considers combined expert, non-expert, and automated evaluation methods. All papers in the proceedings of the major conferences at Visweek 2011 (Visualization, InfoVis, and VAST) will be coded to refine the taxonomy and build a small dataset of exemplar tool evaluations, and demonstrate the distribution of these methods in practice. The taxonomy will be evaluated with a design space analysis; expressiveness of the taxonomy will be demonstrated by our ability to code all evaluations seen in the proceedings, which describe state-of-the-art systems in scientific and information visualization and visual analytics.

An interactive visual tool for finding relevant visualization evaluation methods. An interactive visual interface will be built to let users explore the taxonomy design space. It will include both a map visualization of the space, showing thumbnails with linked publications that have been coded with the taxonomy. It will also include a recommendation interface that can suggest evaluation methods based on project description text or a bibliography of related tools. The interface will be evaluated anecdotally by participants who will be asked to design evaluations for proposed systems both manually and using the taxonomy interface.

B.2 Improve the speed of visualization prototyping by validating and leveraging non-expert participation in visualization evaluation and design.

This aim explores new methods for visualization design and evaluation that leverage combined expert, crowd, and machine resources; we hypothesize that completing Aim A1 will demonstrate that this part of the evaluation design space is underutilized. Expected contributions of this aim are below.

Evaluation of which visual analysis tasks non-experts (e.g., Turkers) do relatively well compared to visualization or data-domain experts, and which ones they do poorly. An interactive, online visualization testbed will be created and used to evaluate task performance from expert and non-expert users. A subset of visual analytics tasks will be selected from [17] and fit to specific tasks in the testbed. Finally, I will conduct controlled experiments with both visualization experts and crowd workers on Mechanical Turk, and give quantitative and anecdotal comparisons of task performance (time and accuracy) for these tasks, and concluding guidelines for the types of tasks non-experts can do well and poorly.

Design patterns for evaluating visualization with a crowd or crowd+expert combination. Based on findings from the previous contribution, one or more design patterns will be created for crowdsourcing visualization analysis tasks. These patterns will describe general workflows for using the crowd effectively, similar to Bernstein et al.'s *Find-Fix-Verify* [8]. An interface for experts to help specify crowd evaluation metrics will be created and used to test and iterate on these workflows. We will anecdotally evaluate the cost and effectiveness of expert-guided crowd workflows using domain scientists. The entire crowd+expert visualization evaluation paradigm will be anecdotally evaluated by at least one designer of brain circuit diagrams.

Learn and validate design guidelines for scientific visualization by analyzing visual artifacts from both domain experts and non-experts. We will complete a novel design evaluation of handcrafted visualizations. Manually created visualizations, like slideshows or diagram sketches, will be analyzed to learn visualization design guidelines. Qualitative methods (e.g., interviews, critiques, etc.) will complement quantitative analysis on collected datasets. Expected contributions of this aim are below.

Extracting guidelines from handcrafted visualizations in scientific and general domains. Published handcrafted visualizations of brain circuit diagrams will be collected and posted online to collect scientist and designer critiques. The images and collected critiques will analyzed using qualitative and computational methods to learn layout rules and other domain conventions for these visualizations. Non-scientific or more informal handmade visualizations, like slideshows, whiteboard sketches, and conference posters, will also be analyzed to produce design guidelines. These guidelines will be used to construct models of visualization quality that will be evaluated quantitatively by comparing model-based quality assessment against assessment scores of domain experts.

C Background and Significance

In this section, I discuss the related work and significance for each aim described in Sec. §B.

C.1 Evaluation taxonomy for visualization and design-space recommender tools.

C.1.1 Significance A comprehensive taxonomy of evaluation methods for visualization does not exist, and would add transparency and inform tool developers about how to measure their contributions. It is unclear whether the lack of a standard taxonomy is a result or cause of evaluation challenges, as underscored by Plaisant [29]. Nonetheless, it could lead to the kinds of evaluation benchmarks that are necessary for tool comparisons. Standard benchmarks are common in fields like natural language processing and computer vision but are rare in visualization, possibly due to the difficulty of quantifying usability and insight. Another challenge is that comparable visualization tools sometimes use different evaluation methods. A recommender system that suggests how to evaluate a tool based on its characteristics and on similar tools could lead to more effective visualization tool comparisons and conclusions about feature design.

C.1.2 Background Taxonomies have been used in the visualization community to characterize different visual representations. One of the earliest examples is Bertin's classification of that retinal variables – like glyph color, shape, and size – that can encode data attributes [9]. At a higher level, other work has been important in illustrating subtle differences between visualization constructs when many representation choices exist, as in tree visualizations. Recently, Jurgensmann et al. created an interactive visual map of tree visualizations that illustrates the design space of tree visualization types and lets users locate example applications using these different representations [22, 33] (see Fig. §1). In the proposed work, we will create a visual, searchable survey of evaluation methods similar to the Treevis.net visualization.

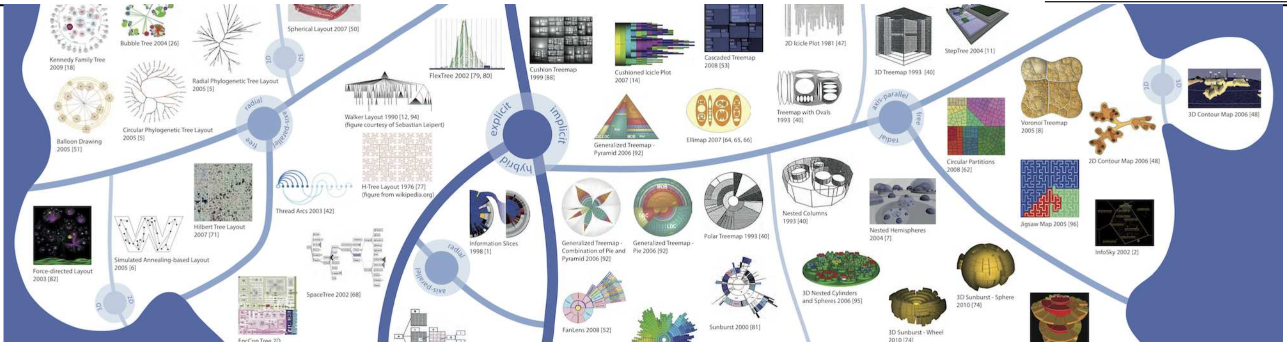
Kennedy Family Tree 2009 [18]
Bubble Tree 2004 [26]
Radial Phylogenetic Tree Layout 2005 [5]
Spherical Layout 2007 [50]
2D
3D
Balloon Drawing 2005 [51]
Circular Phylogenetic Tree Layout 2005 [5]
radial
axis-parallel
tree
Walker Layout 1990 [12, 94] (figure courtesy of Sebastian Leipert)
H-Tree Layout 1976 [77] (figure from wikipedia.org)
Thread Arcs 2003 [42]
Hilbert Tree Layout 2007 [71]
Force-directed Layout 2003 [82]
Simulated Annealing-based Layout 2005 [6]
SpaceTree 2002 [68]
EncCon Tree 2D
FlexTree 2002 [79, 80]
Cushion Treemap 1999 [88]
Cushioned Icicle Plot 2007 [14]
Cascaded Treemap 2008 [53]
2D Icicle Plot 1981 [47]
explicit
implicit
hybrid
Generalized Treemap - Pyramid 2006 [92]
Ellimap 2007 [64, 65, 66]
Treemap with Ovals 1993 [40]
Information Slices 1998 [1]
Generalized Treemap - Combination of Pie and Pyramid 2006 [92]
Generalized Treemap - Pie 2006 [92]
Polar Treemap 1993 [40]
Nested Columns 1993 [40]
FanLens 2008 [52]
Sunburst 2000 [81]
3D Nested Cylinders and Spheres 2006 [95]
3D Treemap 1993 [40]
StepTree 2004 [11]
Voronoi Treemap 2005 [8]
Circular Partitions 2008 [62]
2D Contour Map 2006 [48]
3D Contour Map 2006 [48]
Nested Hemispheres 2004 [7]
Jigsaw Map 2005 [96]
InfoSky 2002 [2]
3D Sunburst - Sphere 2010 [74]
3D Sunburst - Wheel 2010 [74]

Figure 1: Some tree visualization methods surveyed in Treevis.net.

Recommender systems have also been used previously in visualization to select visual representations for a data selection. One of the earliest works in this area was Mackinlay's ATP [25], which automatically designed charts for relational data of with several types. This has been extended recently as part of Tableau's "Show Me!" feature [26], and is related to the aim of ReVision [32], which extracts data and re-visualizes chart bitmaps using prescribed design rules. Rather than recommending visual representations, our aim is to recommend evaluation methods for a given visualization tool.

C.2 Combined expert+crowd evaluation patterns.

C.2.1 Significance The proposed research will help prescribe new workflows that leverage combined resources (expert, non-experts, and machines) to make visualization evaluation faster for developers to plan and execute. Crowdsourcing has become a popular tool for solving large-scale sets of unit tasks that humans are good at but are difficult for machines, e.g., object detection or recognition in images. As such, crowdsourcing presents an exciting opportunity for new evaluation technologies in scientific visualization, which is often forced to rely on small sample sizes of domain experts.

The focus of this work is on design patterns and workflows for crowdsourcing evaluation, rather than simply new applications. Inventing these workflows is non-trivial [1] and will have impact for applications beyond the tools I will develop.

C.2.2 Background Understanding which kinds of tasks crowds do well, and why, provides some direction for work in this area. Heer et al. demonstrated that a basic visual analysis task – interpreting numeric values from chart types – can be done reliably with Turkers [16]. Going beyond simple chart perception into evaluations of spatial visualizations, or node-link diagrams, will require more kind of analysis tasks to be tested. Visual analytics research has provided some study of what scientists and analysts actually do with exploratory visualization. Amar et al. gave a set of low-level analysis tasks [4], including identifying extrema or outliers and filtering in data. More recently, Heer and Shneiderman presented a three-tier set of analysis and manipulation activities, including higher level actions like annotation [17]. One focus of these efforts is to provide a necessary set of criteria for visual analytics systems; i.e., defining which analyst activities a visualization environment must support. They also suggest benchmark tasks to evaluate the abilities of analysts, including crowd workers like Turkers.

Important work has recently focused on design patterns for crowd-powered interfaces, like Bernstein et al.'s *Find-Fix-Verify* (FFV) pattern [8]. FFV was demonstrated in a word processor that uses Turkers to identify problematic text (for instance, when text is too long or could otherwise be edited), suggest edits, and verify those edits. That pattern may not be directly transferable to visual analysis tasks; the "Fix" step, for example, does not make sense for content that cannot be edited, like visualizations.

Guiding crowd workers has been explored as a way to improve crowd task quality. Bernstein et al. used dynamic refinement of the search space for a crowd task that involved selecting important frames in a video sequence [7]. This refinement works because crowd workers can find approximate solutions for this task quickly, and can be used to shepherd multiple workers toward a consensus solution. The task also recruited workers on a *retainer* model to get realtime responses. Another tool that used this approach was VizWiz [10], which leveraged realtime crowd responses to identify objects and provide descriptions of phone-captured photographs. Retained workers could be used as part of a pattern for guiding crowd visualization evaluations using recruited and retained domain experts. Another crowd-guiding tool is Turkomatic [23], which uses Turkers to either solve a task as-written or refine it for other workers. In this way, the original task requester needs only to provide a high-level description of tasks. In our work, we will consider how domain experts can help design or refine tasks when crowd workers struggle.

C.3 Handcrafted visualization and language analysis for design guidelines.

C.3.1 Significance In leveraging all resources for improved visualization and evaluation, handcrafted visualizations and diagrams might contain important visual conventions for data domains. For instance, biologists studying protein interactions have used stylized node-link diagrams to represent interaction pathways. It makes sense that new computer-generated visualizations that adhere to accepted handcrafted conventions might have better comprehension or recall for these biologists than other visualizations. While design studies have been used to learn some rules from handcrafted visualizations, visualization research will benefit from easy to replicate, systematic methods that use machine learning or other techniques to extract design conventions from existing visualizations. In addition to using these extracted rules to make better visualizations, they can also be used as evaluation criteria.

C.3.2 Background Effective visualization design can be subtle or even counterintuitive. For instance, recent research has shown that "junk ink", which includes visual details that do not explicit encode information, can improve chart memorability and comprehension [5]; systematically adding visual difficulties can improve information visualizations [19]; and designing discomfort into interactions can enable more enlightening experiences [6]. Design studies that identify rules from handcrafted visualizations have been used to improve route maps [2, 3] and exploded-view assembly diagrams [24]. In these cases, software was built to generate these kinds of visualizations automatically. Another recent project called ReVision presents a method to interpret and redesign simple information visualizations, like bar charts, automatically using basic design rules [32]. Similarly, we plan to use of computer vision and machine learning to learn and apply design rules.

Non-traditional visualizations have also been studied for design. Whiteboards were examined as *spontaneous* visualizations and coded for recognizable information visualization constructs [34]. We extended the ideas in this work to characterize marks in electronic slideshows and whiteboard talks from graduate students and found discipline-specific design patterns in visual presentations [14]. Metoyer et al. recently studied descriptions of information visualizations [27]. The resulting guidelines from these works are fairly general, though focusing on scientific visualizations might reveal domain-specific design rules.

D Preliminary Work

In this section, I discuss preliminary work toward the Aims 2 and 3 described in Sec. §B. I also give an overview of resources and established collaborations that will be leveraged in the proposed research.

D.1 Combined expert+crowd evaluation patterns. In two previous projects, Tome and Crowd Control, we have established that some expert tasks related to user interface evaluation and visual analysis can be replicated or approximated by crowds. In both cases, traditionally 'expert' tasks are completed by non-experts, reducing expert involvement without significant loss of task quality.

Tome The Tome project is aimed at making quantitative user-interface evaluations faster to execute. Tome is a framework that automatically compiles and exports interaction histories as end users interact with instrumented user interfaces; after collecting histories from a crowd of end users, Tome builds predictive models that characterize how the crowd completes specific tasks [15]. The output of Tome is project files for CogTool [21], a popular mockup-based performance modeling tool that lets visualization developers modify, explore, and compute time predictions for the tasks models built by Tome.

The main contribution of Tome is a novel approach to evaluation: applications are instrumented during development, then deployed like other applications to end users. As end users complete tasks with the application, they are passively providing data for evaluation that can be collected and aggregated by the UI developer. Tome then generates baseline performance models for the deployed UI, and these models can be modified with CogTool to test new UI features, as demonstrated in [15]. Using aggregated histories from a crowd of end users is a novel extension of Hudson et al.'s CRITIQUE instrumentation system [18]; Tome uses filtering to model canonical task strategies without knowing them a priori, while CRITIQUE requires an expert user to demonstrate task executions.

Crowd Control We evaluated a workflow we called Crowd Control that asks crowd workers to assess image quality features. The goal of Crowd Control is to provide a crowd-powered quality control (QC) system for MRI scans that lets expert scientists offload some time-consuming visual analysis tasks. QC is an important step for neuroscientists using brain imaging for hypothesis generation, but as one brain science collaborator commented, there sometimes isn't enough "RA [research assistant] power" to do QC on each scan. If the visual analysis tasks involved in QC can be done by a scalable, non-expert workforce like Turkers, crowdsourcing may be a viable approach to assessing quality for large imaging datasets.

We interviewed a brain scientist whose group numerically scores image quality (on a scale of 0 to 3) in MRI acquisitions. This scoring task was converted into Mechanical Turk tasks to evaluate how well internet workers could perform *image contrast assessment* using only a couple scoring examples [12]. A 3x2 study asked Turkers to score MRI scans with 3 distinct contrast levels for 2 projection views, axial and sagittal, and 20 scores were collected for each condition for a total of 120 scores (120 = 3 x 2 x 20). We found that Turkers gave scores close to an expert's scores for the contrast conditions in both view types.

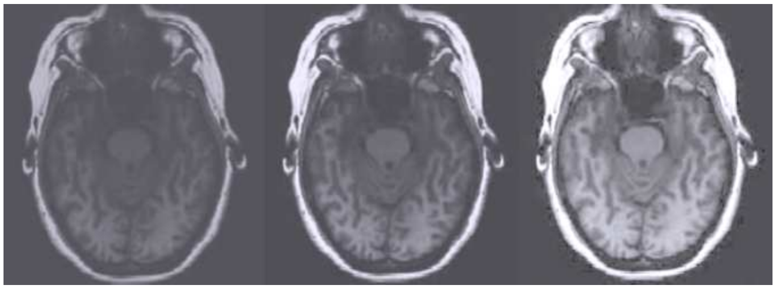

Figure 2: Contrast variations in a T1-weighted MRI. Turkers numerically scored the contrast quality of MRI slice-sequence videos.

D.2 Handcrafted visualization and language analysis for design guidelines. We have also demonstrated the feasibility of extracting some design guidelines from handcrafted visualizations.

Slideshows and whiteboard talks We studied visual presentation design between academic disciplines at Brown, Harvard, and RISD, and found insights about discipline-specific conventions in these presentations [14]. Disciplines were compared at a coarse scale between four groups of fields: social, natural, and formal sciences, and the humanities. Several experiments were performed on ubiquitous narrative presentations, including electronic slideshows (e.g., PowerPoint and Keynote presentations) and whiteboard "chalk talks".

Principal components of slide images, called *eigenslides*, and manually-selected features inspired by Walny et al.'s whiteboard taxonomy [34], like the use of bullet points and visualization constructs, illustrate presentation design differences between disciplines. We followed up the slideshow analysis with a live study in which participants were asked to design and videotape whiteboard talks for controlled presentation topics. We found some study participants used representations and levels of argument formality that are characteristic of their own fields of study. Based on these findings, we gave implications for designing information visualizations in different user domains, including support for domain-tailored recommender systems and crowd-powered presentation evaluation.

Handcrafted brain diagrams Static brain circuit diagrams found on the Web and in textbooks are a resource for developing our own visualization tools. We made novel use of static, handcrafted diagrams as part of a prototype input mechanism that lets users re-visualize the relationships drawn. A user sketches over the diagram elements to "select" information from a linked database and visualize it in a new environment. A similar method was used by Jianu et al. to layout large protein networks around a user-drawn diagram 'skeleton' [20], but our method is not limited to laying out diagrams in place. With our interface, the diagram can be used as scaffolding for general user input to filter or further explore components of the diagram in a smart, multi-view setting, which could include features like link-outs to extra information; visualizing related information or suggested extensions to the diagram; or dynamically changing visual mappings or laying out a diagram differently.

This work is still preliminary, though the interaction metaphor is a potential contribution. Another direction I plan to continue is extracting aesthetic features from handcrafted network diagrams for use in automatic domain-tailored visualization layouts.

D.3 Research Environment In my preliminary work, I have demonstrated successful research contributions in user interface research and HCI research. Aside from studying related problems, I have exposed myself to the methods necessary to carry out the proposed work, including user studies with domain experts [13]

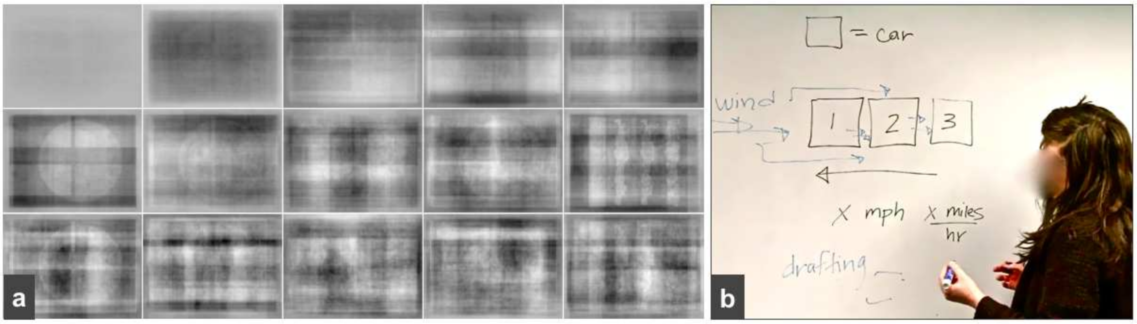

Figure 3: Principal components of electronic slides, called *eigenslides*, are shown in (a). Contrast patterns in eigenslides and semantic features, like visualizations and bullet points, distinguish these visual artifacts between different academic disciplines. Some discipline-related design conventions, like building on representations over time and using visualization constructs, are used by authors even when we controlled for topics in video-recorded whiteboard presentations (b).

and non-experts [12]; coding features in video and other visual artifacts [14, 36]; software development; and visualization prototyping [15].

In coursework and research, I have investigated visualization problems in protein interaction pathways, brain imaging, and brain networks. The aims of this proposal will be developed and tested with respect to ongoing visualization challenges from collaborations at Brown, including tractography user interfaces with Dr. Win Gongvatana and Dr. Stephen Correia (neuroscience), and outside Brown, including brain circuit analysis with Dr. Mark Schnitzer's lab at Stanford (neuroscience).

E Research Plan

In this section, I discuss detailed plans for achieving the contributions in each aim. Aim 1, which will help survey the state of evaluation and will support later contributions, will be completed first; Aims 2 and 3 both involve analyzing insights and abilities of non-experts and will be approached in parallel. Expert guidance and guidance from machine learning technologies will be used in both crowd-powered evaluation tools and design analysis tools. The expected completion date for the proposed work is December 2014.

E.1 Evaluation taxonomy and dataset To achieve Aim 1, I plan to complete the following steps: understand and aggregate UI and visualization evaluation methods into a single taxonomy; categorize recent visualization tools using the taxonomy; and build an exploratory visualization and recommender system for the evaluation design space.

First, I will create the taxonomy by first reviewing research on evaluation techniques from the user interface community and techniques at the intersection of visualization, perception, and cognitive psychology. Some of these have been developed specifically with visualization in mind, like insight-based methods [31, 35] and perceptual modeling [28], while others like cognitive walkthrough and heuristic evaluation have been used to evaluate user interfaces generally. After forming an initial taxonomy, I will use open-coding to characterize the evaluation methods used in a dataset composed of all papers from IEEE Visualization 2011, IEEE Information Visualization 2011, and IEEE VAST 2011. During open coding [11], the taxonomy may be refined until it satisfactorily describes the evaluation design space seen in the dataset. An extensible database of categorized papers and evaluations will be made public online.

Next, I will leverage this database by creating an exploratory visualization of the design space, using coded systems in the database as data points. The visualization will be open source and made public online. I will also develop a recommendation feature for visualization developers who can provide metadata about their own projects and receive evaluation hints based on similar projects in the database. Initially, users will be able to upload a bibtex bibliography file with references, see highlighting of any coded references in the visualization, and receive a summary of methods used by those references. I will evaluate this tool anecdotally with visualization developers and incorporate feedback into the tool's design.

E.2 Crowd+expert evaluation patterns and systems To achieve Aim 2, I plan to: implement a testbed for posting visualization tasks on Amazon Mechanical Turk; evaluate both expert and Turker performance on a set visual analytics tasks with a scientific visualization; design crowd+expert workflows for tasks that Turkers cannot do well independently. I expect the database and visualization produced in Sec. §E.1 to reveal that crowds have not been used extensively to evaluate visualizations, despite some early proofs of concept [16]; my next research activities are aimed at validating whether crowds can meaningfully contribute to these evaluations and to establish workflows that utilitize crowds effectively.

First, I will implement a testbed that allows lightweight visualizations to be hosted online and analyzed or evaluated by workers on Mechanical Turk. The visualization itself will either support interactivity, or the webpage that embeds the visualization will contain interactive features. This is important for collecting information from users and for letting them do realistic analysis activities that might require interaction (e.g., hypothesis formation, annotation, etc.). The precise tasks users will be asked to complete will be determined by finding task instances for a brain circuit visualization that are adapted from the general analysis types in Heer's survey [17]. The brain circuit visualization will be developed in conjunction with other members of the visualization lab, and with collaborators at Stanford. As a contingency in case the visualization process is behind schedule, static, handcrafted visualizations of brain circuits will be collected online and used for the analysis tasks.

Task performance and feedback will be collected both from a small set (3-5) of brain experts and from a larger set of Turkers (20-30). I will treat task performance from experts as ground truth and analyze the successes and failures of the crowd to complete each task. From this meta-analysis, I will then propose workflows that allow experts to guide workers or refine their analysis tasks. This will be an iterative process: Each proposed workflow will be evaluated again using Turkers. This end result of this hypothesize-develop-test cycle will a conclusion about which evaluation tasks can be done without expert intervention, and which can be done using some systematic expert guidance.

E.3 Extracting design guidelines from handcrafted visualizations To achieve Aim 3, I plan to: collect datasets of visualizations and transcripts of think-aloud visual analysis activities; evaluate these datasets quantitatively and qualitatively; and synthesize design guidelines from these findings. These guidelines will serve as criteria for domain-specific heuristic-based evaluations. The components of this Aim that related to non-scientific, everyday visualizations (e.g., whiteboards, slideshows) are mostly completed and are in review.

First, I will collect a dataset of handcrafted brain-circuit visualizations (bitmaps). These will be gathered from the Web and from textbooks on brain science. I will post these online with an interface to collect and show comments and labels about these images. The implementation of this interface will be extended from the experiment testbed interface described in Sec. §E.2. The vision for this tool is similar to a visualization-specific version of LabelMe [30], which provides an image dataset and annotation interface that have been widely used by computer vision researchers. In addition to allowing feedback from Web contributors, I will conduct a think-aloud study with collaborators in brain science who will use the interface and evaluate legibility and comprehension in these diagrams, and collect feedback about these diagrams.

Next, I will synthesis design guidelines that relate visualization features in these diagrams (e.g., edge/node density, layout type) to evaluation metrics. These will be useful as heuristics for evaluation. Afterwards, I will use vision techniques that detect and quantify these visualization features. In [14], we demonstrated some simple approaches in this direction by computing principal components of bitmap slides (e.g., PowerPoints). We will use methods like *bag of image patches* models, which was used to classify chart types in ReVision [32], and OCR to detect and classify diagram elements. Using features that can be extracted from diagrams automatically, I will construct and fit a predictive model of diagram legibility and comprehension. The model will be evaluated with a held-out dataset against expert quality assessments of these diagrams.

E.4 Potential challenges and solutions To the best of my knowledge, there does not exist a unified taxonomy like the one described in Aim 1. If another similar taxonomy is found, I will use that to code the Visweek 2011 proceedings, as described in Sec. §E.1. The method recommendation contribution in Aim 1 would also be unaffected by the choice of taxonomy. There will still be a contribution in relating this work to the visualization community by providing a dataset of evaluation method exemplars. If coding the papers from the Visweek proceedings suggests that the papers do not involve enough examples or diverse enough examples, I will code visualization papers published at other top venues, including IEEE TVCG, ACM CHI, and ACM UIST.

Aim 2 will involve creating a set of analysis tasks with scientific visualizations. It is possible that not all task types will be easy to fit to our expected visualization domain of brain circuit visualizations. In this case, I will use other scientific diagrams of protein interaction networks, or other domains, to test crowds and experts. The software infrastructure I will build to post tasks will be domain-independent, so altering the tasks or image domains will not affect its development and deployment.

Both Aims 2 and 3 will result in generalizations about crowd workers and learning design visualization heuristics. It is possible that results will be too domain-specific to generalize beyond the visualization domains that are tested. If results during early piloting of the experiments seem too domain-specific, I will expand the experiments to new visualization domains. This will not affect the software developed as part of Aims 2 or 3. An analysis of why domains do not generalize will also be produced to give insight about the design space of analytics in these fields.

Date	*Contribution*	*Description*
July 1, 2012	[Aim 1] Taxonomy prototype	Taxonomy refined, IEEE Visualization proceedings coded, and prototype visualization of dataset.
Sept 1, 2012	[Aim 3] Brain diagram online tool	Brain diagram dataset is online and annotation tool implemented, and anecdotally evaluated.
Sept 20, 2012	[Aim 3] Submission: CHI paper	Evaluation of scientific diagram Web viewer/annotation portal.
Nov 1, 2012	[Aim 2] Testbed for crowd evaluation	Prototype interface implemented that lets end user post visualization tasks to Mechanical Turk.
Dec 1, 2012	[Aim 2] Expert and crowd task evaluation	Evaluation of experts and crowd on visual analysis tasks.
March 15, 2013	[Aim 1] Submission: InfoVis/TVCG paper	Taxonomy, dataset, and browser tool for evaluation methods.
April 15, 2013	[Aim 2] Submission: UIST paper	Patterns and UI for crowd-powered visualization evaluation.
June 30, 2013	[Aim 3] Vision analysis of brain diagrams	Models of diagram legibility/comprehension.
September 20, 2013	[Aim 2] Submission: CHI paper	Pattern and UI for expert-guided crowd-powered visualization evaluation.
March 2014	Thesis draft completed	
March 15, 2014	[Aim 3] Submission: InfoVis paper	Automatic heuristic-based evaluation of diagrams. Tying together suite of automatic evaluation methods; models, crowds, experts.
May 2014	Thesis defense	

Table 1: Timeline for research plan deliverables.

Literature Cited

[1] Eytan Adar. Why I Hate Mechanical Turk Research (and Workshops). In *CHI 2011 Workshop on Crowdsourcing and Human Computation*, 2011.

[2] M. Agrawala and C. Stolte. Rendering effective route maps: Improving usability through generalization. In *Proceedings of SIGGRAPH*, pages 241–250, 2001.

[3] Maneesh Agrawala, Wilmot Li, and Floraine Berthouzoz. Design principles for visual communication. *Commun. ACM*, 54:60–69, April 2011.

[4] Robert Amar, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, INFOVIS '05, pages 15–, Washington, DC, USA, 2005. IEEE Computer Society.

[5] Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 2573–2582, New York, NY, USA, 2010. ACM.

[6] Steve Benford, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, and Tom Rodden. Uncomfortable interactions. In *ACM CHI*, 2012.

[7] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. Crowds in two seconds: enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 33–42, New York, NY, USA, 2011. ACM.

[8] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 313–322, New York, NY, USA, 2010. ACM.

[9] Jacques Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983.

[10] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM.

[11] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, 2007.

[12] Steven R. Gomez. Crowd Control: Outsourcing quality assessment in brain-imaging workflows. 2012.

[13] Steven R. Gomez, Radu Jianu, and David H. Laidlaw. A fiducial-based tangible user interface for white matter tractography. In *Proceedings of ISVC*, 2010.

[14] Steven R. Gomez, Radu Jianu, Caroline Ziemkiewicz, Hua Guo, and David H. Laidlaw. Different strokes for different folks: Visual presentation design between disciplines. 2012.

[15] Steven R. Gomez and David H. Laidlaw. Modeling task performance for a crowd of users from interaction histories. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2012. In Press.

[16] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 203–212, New York, NY, USA, 2010. ACM.

[17] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Commun. ACM*, 55(4):45–54, April 2012.

[18] Scott E. Hudson, Bonnie E. John, Keith Knudsen, and Michael D. Byrne. A tool for creating predictive performance models from user interface demonstrations. pages 93–102, 1999.

[19] Jessica Hullman, Eytan Adar, and Priti Shah. Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2213–2222, December 2011.

[20] Radu Jianu, Kebing Yu, Vinh Nguyen, Lulu Cao, Arthur Salomon, and David H. Laidlaw. Visual integration of quantitative proteomic data, pathways and protein interactions. *IEEE Trans. on Visualization and Computer Graphics*, September 2009.

[21] Bonnie E. John, Konstantine Prevas, Dario D. Salvucci, and Ken Koedinger. Predictive human performance modeling made easy. In *ACM CHI*, pages 455–462, 2004.

[22] Susanne Jürgensmann and Hans-Jörg Schulz. A visual survey of tree visualization, 2010.

[23] Anand Kulkarni, Matthew Can, and Bjorn Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *CSCW*, 2012.

[24] Wilmot Li, Maneesh Agrawala, Brian Curless, and David Salesin. Automated generation of interactive 3d exploded view diagrams. *ACM Trans. Graph.*, 27(3):101:1–101:7, August 2008.

[25] Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, April 1986.

[26] Jock Mackinlay, Pat Hanrahan, and Chris Stolte. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, November 2007.

[27] Ronald Metoyer, Bongshin Lee, Nathalie Henry Riche, and Mary Czerwinski. Understanding the verbal language and structure of end-user descriptions of data visualizations. In *ACM CHI*, 2012.

[28] Daniel Pineo and Colin Ware. Data visualization optimization via computational modeling of perception. *IEEE Transactions on Visualization and Computer Graphics*, 18(2):309–320, 2012.

[29] Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '04, pages 109–116, New York, NY, USA, 2004. ACM.

[30] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, May 2008.

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

[32] Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. Revision: automated classification, analysis and redesign of chart images. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 393–402, New York, NY, USA, 2011. ACM.

[33] Hans-Jörg Schulz. Treevis.net: A tree visualization reference. *IEEE Computer Graphics and Applications*, 31(6):11–15, 2011.

[34] Jagoda Walny, Sheelagh Carpendale, Nathalie Henry Riche, Gina Venolia, and Philip Fawcett. Visual thinking in action: Visualizations as used on whiteboards. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2508–2517, December 2011.

[35] Ji Soo Yi, Youn-ah Kang, John T. Stasko, and Julie A. Jacko. Understanding and characterizing insights: how do people gain insights using information visualization? In *Proceedings of the 2008 conference on BEyond time and errors: novel evaLuation methods for Information Visualization*, BELIV '08, pages 4:1–4:6, New York, NY, USA, 2008. ACM.

[36] Caroline Ziemkiewicz, Steven R. Gomez, and David H. Laidlaw. Analysis within and between graphs: Observed user strategies in immunobiology visualization. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2012. In Press.