Scientific Analysis by the Crowd: A system for implicit collaboration between experts, algorithms, and novices in distributed work

by

David Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2013

Doctoral Committee:

Professor Thomas A. Finholt, Chair
Assistant Professor Eytan Adar
Professor Brian D. Athey
Professor Mark H. Ellisman, UC San Diego
Assistant Professor Mark W. Newman

ACKNOWLEDGEMENTS

The mission of the School of Information (SI) is “Connecting people, information, & technology in more valuable ways”. My time at the university has taught me that of these elements, people are by far the most important and many people have served invaluable roles in my personal and professional development culminating in this work. These people include my advisor Thomas Finholt, Erik Hofer, the faculty/staff at SI. This group also includes fellow students, cohort, and growing family whose enduring patience inspire me daily. Thank you all.

This work is also made possible through the collaboration and support of the University of Michigan, National Center for Microscopy and Imaging Research, and NIH project 5R01NS075314-03.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Appendices

ABSTRACT

Crowd sourced strategies have the potential to increase the throughput of tasks historically constrained by the performance of individual experts. A critical open question is how to configure crowd-based mechanisms, such as online micro-task markets, to accomplish work normally done by experts. In the context of one kind of expert work, feature extraction from electron microscope images, this thesis describes three experiments conducted with Amazon's Mechanical Turk to explore the feasibility of crowdsourcing for tasks that traditionally rely on experts.

The first experiment combined the output from learning algorithms with judgments made by non-experts to see whether the crowd could efficiently and accurately detect the best algorithmic performance for image segmentation. Image segmentation is an important but rate limiting step in analyzing biological imagery. Current best practice relies on extracting features by hand. Results showed that crowd workers were able to match the results of expert workers in 87.5% of the cases given the same task and that they did so with very little training. The second experiment used crowd responses to progressively refine task instructions. Results showed that crowd workers were able to consistently add information to the instructions and produced results the crowd perceived as more clear by an average of 8.7%. Finally, the third experiment mapped images to abstract representations to see whether the crowd could efficiently and accurately identify target structures. Results showed that crowd workers were able to find 100% of known structures with an 82% decrease in false positives compared to conventional automated image processing.

This thesis makes a number of contributions. First, the work demonstrates that tasks previously performed by highly-trained experts, such as image extraction, can be accomplished by non-experts in less time and with comparable accuracy when organized through a micro-task market. Second, the work shows that engaging crowd workers to reflect on the description of tasks can be used to have them refine tasks to produce increased engagement by subsequent crowd workers. Finally, the work shows that abstract representations perform nearly as well as actual images in terms of using a crowd of non-experts to locate targeted features.

CHAPTER 1

Introduction

There has been much recent excitement about the potential to organize large groups ("the crowd") to perform tasks faster and better than individuals (Westphal, Butterworth et al. 2005; Kittur, Chi et al. 2007; Lintott, Schawinski et al. 2008; Lee, Kapelner et al. 2009; Sullivan, Wood et al. 2009; Bernstein, Little et al. 2010; Lin 2010). Referred to broadly as "crowdsourcing," efforts to harness crowd work have been accelerated by the development and proliferation of online micro-task markets (Surowiecki and Silverman 2007; Kittur, Chi et al. 2008; Sullivan, Wood et al. 2009; Borne and Team 2011; 2013). Amazon's Mechanical Turk (AMT) is the most notable instance of a micro-task market and it has been explored in the context of a wide range of relatively simple tasks (e.g., eliminating duplicate entries from a database). An open question is whether markets like AMT can be harnessed to perform expert tasks. Specifically, preliminary exploration suggests that in a large class of activities throughput is often limited by the pace at which an expert can work. In particular, three primary challenges emerge when trying to transform tasks normally performed by experts into tasks that can be accomplished by the crowd. First, problems need to be converted into forms that can be addressed by non-expert workers. Second, definitions of tasks themselves need to be refinable by the crowd. And finally, data need to be transformed such that potentially sensitive or private information is not exposed.

Scientific data analysis is a domain where crowd-based approaches have been proposed as a solution to many of the difficulties arising from data abundance (Westphal, Butterworth et al. 2005; Lintott, Schawinski et al. 2008; Lin 2010). For example, with modern instrumentation, scientists increasingly have access to more data than can be analyzed. These gains reflect improvements in underlying semiconductors that are at the heart of computers, sensors, and related technologies embedded within instruments. Moore's law, for example, says that processing power doubles every 18 months (Schaller 1997). Therefore, the rate of data acquisition is growing as a function of the capabilities of semiconductor-based systems, such as the charged coupled devices (CCDs), or photonic detectors, used by cameras within instruments to capture images. While methods for collecting data have improved radically, methods for analyzing data have evolved more slowly. After acquiring data from instrumentation, extracting precise models from the data can be a time-consuming and labor-intensive process.

The neurosciences are one domain where scientists struggle to keep pace with the capabilities of instruments, such as CCD-based light and electron microscopes. On the one hand, it is now possible to obtain images that show both cellular and sub-cellular structures over relatively large areas (e.g., millimeters). However, resolution at this scale poses challenges for viewing patterns of organization within biological systems. For example, extracting all of the cell membranes from a moderate dataset of 2k x 2k x 100 pixels could require months of manual labor. As a consequence, scientists modify their analysis protocols to downsize images and reduce the number of structures that need to be identified. This thesis explores an alternative strategy where the crowd is used to find structures, with the potential of using the full extent of available data without increasing analysis overhead. The sections that follow outline this strategy and the subsequent chapters provide detailed descriptions and results. The thesis concludes with an assessment of the feasibility of crowd-based data analysis and an exploration of future directions.

1.1 Organizing the work of the crowd

It may be counterintuitive to solicit work from an unverified and unknown community of users, but crowds can be surprisingly capable. Surowiecki (2005) describes Francis Galton's study (Galton 1907) of county fair attendees who were asked to guess the weight of a butchered and dressed ox. Galton's surprising finding was that the median of the crowd's estimates (1207 lbs) was very close to the actual dressed weight (1198 lbs) – a difference of only 0.8%. Surowiecki draws a number of conclusions from Galton's account and from other descriptions of crowd-based decision making. First, there should be a diversity of opinion (i.e., each person has private information). Second, individual decisions should be independent (i.e., not influenced by the decisions made by others). Third, each decision maker should be able to draw on specialized or local knowledge. And finally, there should be a mechanism to convert private decisions into a collective decision. Evidence from practice suggests that crowds have worked best for independent and iterative tasks that do not require significant specialized training (Little, Chilton et al. 2009). For example, Von Ahn demonstrated several types of work that crowds can do, but are very difficult to accomplish with algorithmic approaches, such as image labeling (Von Ahn 2006). In his tasks, Von Ahn took advantage of the workers' image processing capabilities. People are able to quickly interpret visual imagery, easily outperforming machine equivalents, particularly in tasks with uncertainty, such as when there are ambiguous subjects or situations with high degrees of noise.

1.1.1 The crowd marketplace

A key challenge in performing crowd-based work is recruiting and paying crowd workers. A successful approach to this problem is the creation of online micro-task markets, such as AMT. AMT uses a web interface to match workers and employers while providing mechanisms for authentication, task management, worker history, and information sharing among workers and employers. Demonstrating the utility and cost benefits of AMT, Kittur et al. (2008) conducted user interface studies, a typically expensive and time consuming task, and found that participants produced high-quality evaluations at a much lower cost than with traditional methods. It was also noted that the time required for experimentation was significantly less due to the size of the worker community and the ability to work in parallel. The programmable architecture of AMT is another advantage, allowing the development of tightly integrated applications. For instance, Soylent is a novel word processor application that is linked with AMT, enabling the user to send off portions of text to the crowd to be proofread or otherwise edited (Bernstein, Little et al. 2010). Similarly, GemIdent (Lee, Kapelner et al. 2009) couples automated algorithms and workers via AMT to seed the algorithms. This process increased the throughput and reduced the cost of the relatively labor-intensive process of quantifying and locating immune and cancer cells for increasing the accuracy of breast cancer prognosis.

1.1.2 Outside the crowd markets

Despite AMT's proliferation, there are alternative mechanisms for aggregating the contributions of many users, often oriented around scientific goals. Wikipedia and citizen science projects are two such examples. Wikipedia is a free, online encyclopedia developed by a very large community of volunteers, aggregating their efforts to create a high-quality resource rivaling print alternatives (Giles 2005). Knowledge in Wikipedia is often seeded by a few individuals with a majority of the other contributions being the aggregate of many small edits by a large number of contributors (Kittur, Chi et al. 2007). Citizen science projects are built around communities, often of expert and lay participants, that aggregate work and expertise, typically with no financial compensation. There are a number of different citizen science projects that utilize their participants in different ways to gather data, process data, solve problems, and explore. A citizen science project that uses participants as sensors, Community Collaborative Rain, Hail and Snow Network (CoCoRaHS), collects precipitation data from thousands of volunteers distributed throughout the U.S. By aggregating these precipitation data at a higher spatial resolution than the national weather service grid, scientists and firms can use these data to generate higher resolution forecasts than are typically available, a feature particularly relevant to weather sensitive individuals such as farmers (Cifelli, Doesken et al. 2005). eBird aggregates the observations of bird enthusiasts to discern patterns over large distances, correlating observations and patterns across observations (Sullivan, Wood et al. 2009). Such coordinated efforts from many independent individuals construct a sustainable network of contributors that would be prohibitively expensive for organizations such as the government to implement. GalaxyZoo is a very successful citizen science project that trains participants to classify celestial objects collected from the Sloan Digital Sky Survey (Lintott, Schawinski et al. 2008). To reduce the complexity of the task for the end user, GalazyZoo automatically processes the data to present the user with only a single celestial body, reducing complexity and opportunity for user error. Attracting the attention of amateur astronomers, GalazyZoo boasts over a billion classifications by users. Participants in GalazyZoo that make significant discoveries have been cited in resulting scientific publications or have been mentioned in the acknowledgements. Along similar lines, Stardust@home uses crowd participants to identify the tracks of interstellar dust from microscopic images of an aerogel flown on a spacecraft (Westphal, Butterworth et al. 2005). Finally, a different kind of project where citizen scientists process data is Foldit. In Foldit, participants “solve” protein structures. Participants employ sophisticated collaborator, heuristic, and visual problem solving methods to discover complex attributes of protein structures (Cooper, Khatib et al. 2010). Using a number of different motivational mechanisms such as professional attribution, a game format, and the opportunity to win special prizes, Foldit has gained much recognition in the popular and academic press.

Citizen science projects attract the attention and efforts of a large number of participants for a number of reasons. These reasons include recreation (eBird), possible financial rewards (Foldit), community/peer recognition (Foldit, Stardust@home, GalazyZoo), and by encapsulating work into a game like or competitive format (Foldit, Finding Khan). Citizen projects allow users to pursue altruistic motivations, such as solving key chemical structures in Foldit to further scientific knowledge, while also providing personal benefits, such as improved weather forecasting through contributions to CoCorahs.

The motivations of Mechanical Turk workers are largely different than citizen science projects (Ipeirotis 2010) (Ross, Irani et al. 2010) and largely financial. Ross et al. in a survey of Mechanical Turk workers found that 18% of participants rely on Mechanical Turk to sometimes or always "make basic ends meet" and only 10% of workers reported that financial compensation was "Irrelevant to me". Ipeirotis conducted a similar survey of Mechanical Turk workers and divided the responses between workers in the United States and India. He found that compared to their US counterparts, significantly more workers from India reported Mechanical Turk as a primary source of income and significantly fewer workers from India reported participation as a recreational activity.

1.2 Feature extraction as a setting for crowd work

As noted previously, the neurosciences represent a domain struggling to analyze increased amounts of data due to improvements in instrumentation. The preceding section, on crowd work, presents possible mechanisms for processing increased volumes of data, either through online micro-task markets or through citizen science projects. I am in a unique position to explore the intersection of these approaches with the needs of neuroscientists.

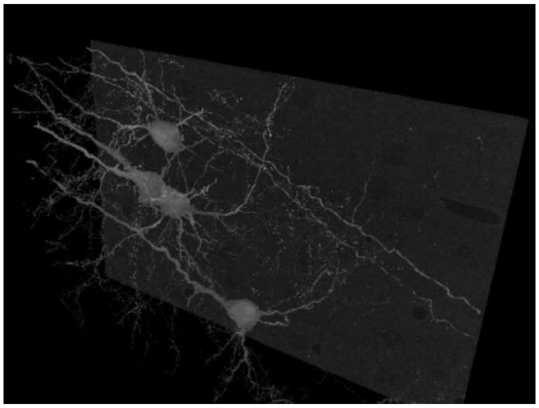


Figure 1: Fully segmented neurons in 3D

Over the period 1999-2012 I spent eight years (including the last three years) in residence at the National Center for Microscopy and Imaging Research (NCMIR) at the University of California, San Diego. Now in its third decade of continuous operation, NCMIR is an international leader in the research and development of technologies for multi-scale, multi-modal 3D and 4D imaging, and correlated light and electron microscopy (Wiseman, Squier et al. 2000; Price, Chow et al. 2006; Milazzo, Lanman et al. 2009; Deerinck 2010; Shu, Lev-Ram et al. 2011). In addition, NCMIR has been a pioneer in the development and use of computational portals for remote instrumentation control (e.g., the Telescience portal) and the deployment of large and high-resolution visualization environments (e.g., the OptIPuter and OptIPortal efforts). My experiences at NCMIR laid the foundations for this thesis work, both through understanding of the underlying scientific work and through relationships formed with NCMIR scientists and support staff. For example, I have observed the work practices of neurophysiologists and other scientists using a number of different instruments, specifically electron microscopes.

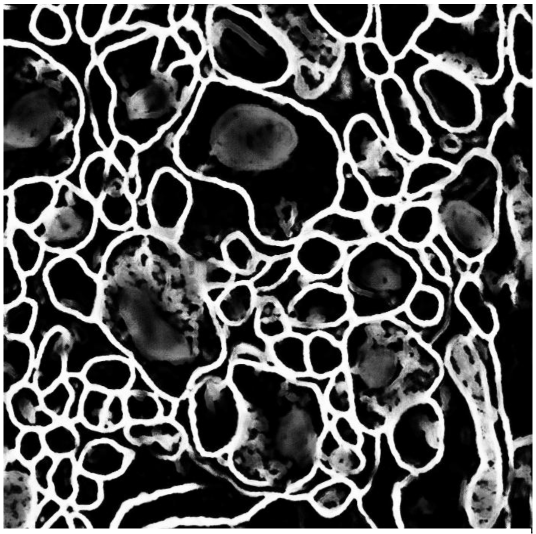


Figure 2: Output from machine learning algorithm detecting cell membranes.

In the neurosciences, instruments such as microscopes are required to see phenomena smaller than what can be resolved by the human eye. These instruments range from multi-photon confocal light microscopes to wide-field transmission and scanning electron microscopes. Because the NCMIR is a center for instrument development, these instruments are at the forefront of available technology, capable of collecting data at higher resolution with improved signal to noise and at increased rates. For example, NCMIR’s scanning electron microscopes routinely acquire close to a trillion pixel 2D images (800k x800k pixels), or about 1.4 TB of data for a single image. It would require 277,778 24-inch monitors to display an image at native resolution. Multiplying the data collected, this instrument also collects data in the Z dimension to create a 3D volume of 100 or more layers. A number of experiments are ongoing with this instrument, and in particular will enable multi-scale electron microscopy, a mode of analysis traditionally reserved for multiple instruments examining the same specimen at different levels of resolution. (Wiseman, Squier et al. 2000; Singh, Schwarz et al. 2006; Perkins, Sun et al. 2009). Multi-scale microscopy allows the researcher to view contexts while preserving the resolution to discriminate between similarly sized but distinct structures within the specimen. This ability is the key to understanding abnormalities, allowing researchers to identify abnormalities at the cellular level, and the details of the abnormality at the subcellular level, providing clues as to the origins of the permutation.

NCMIR’s transmission electron microscopes are fitted with 8k x 8k sensors and have the ability to stitch together multiple images to create very wide field images. Advances such as the 8k x 8k detector allow scientists to minimize the time the microscope is operating, reducing distortion of samples (Milazzo, Lanman et al. 2009). Higher resolution detectors also allow researchers to see context plus focus. Researchers have been able to use the 8k x 8k transmission electron microscope to view entire cells and particles of interest within the cell revealing organization and suggesting function of the particles in context of the broader cellular system (Milazzo, Lanman et al. 2009).

Images direct from the instruments have limited scientific value for rigorous analysis. Often researchers want to "segment" the image, or conduct additional processing to extract features of interest (e.g., neurons as in Figure 1). For example, images from electron microscopes are typically 3D volumes of black and white images. Expert microscopists are able to flip through a stack of images and mentally reconstruct the features within the volume, but sharing and communicating knowledge of objects within these volumes requires additional analysis to create precise models of structures within the data. Extracting features from the volumetric data is the goal of segmentation. Segmentation reveals information about the structure that may not be readily seen when looking at data in 2D such as its three dimensional structure, volume, size, and shape of different components. Specifically, segmentation is the important first step towards understanding the relationship between structure and function. For instance, in 1997, Perkins et al, transformed our understanding of the structure of mitochondria, the energy makers within a cell, through the process of 3D segmentation and computational reconstruction (Perkins, Renken et al. 1997). After 3D reconstruction, Perkins observed that the internal structure, previously thought to be a baffling of endoplasmic reticulum was actually a series of crista that span the cell width. Previous models were based on images where the crista appeared like baffles when projected in a 2D space. Image segmentation and reconstruction also allows for quantitative analysis. As computed models, attributes such as volume, size, and other attributes can be easily counted, compared, and organized. Further developing insight from the segmentation of cells and its components, Perkins et al. used precise volumetric segmentation of mitochondria to suggest unique bioelectric attributes of rods compared to cones in the retina.

Mitochondria are the energy makers of the cell and the precise characterization of its ultrastructure can be used to infer relationships between structure and function, specifically larger mitochondria with more cristae are capable of producing more energy. By reconstructing precise volumetric models of mitochondria associated both with rods and cones, Perkins inferred the energy requirements of the two different cell types and suggested that cones have greater energy consumption compared to light adapted rods (Perkins, Ellisman et al. 2003). As further illustration, Martone et al. have developed a sophisticated database of image and segmentation data useful for cross-correlated quantitative studies of segmented features (Martone, Gupta et al. 2002).

There are currently three methods for image segmentation from electron microscopes. The dominant method involves a manual process where experts trace features within an image with a computer mouse. It is common for users to trace a stack of images that a feature of interest runs through. After tracing one image, the user iterates through the stack of images creating a series of contours. An application compiles the different contours in 3D space and interpolates between the tracings, creating a 3D model of the object of interest (Kremer, Mastronarde et al. 1996). While manual segmentation is very accurate and the current gold standard for scientific analysis, it is slow and does not scale well. A second method involves automated algorithms for the analysis of electron micrographs using machine learning techniques (e.g., Figure 2). The machine-learning routine extracts features from images based on a training dataset (Jurrus, Hardy et al. 2009; Mishchenko 2009). While highly scalable and capable of processing very large amounts of data quickly, these methods do not achieve the accuracy required for scientific analysis. Further, experts believe that machines will not match the performance of experts in the near future due to the noisy nature of electron micrographs (Mishchenko 2009). Finally, a third method involves semi-automated approaches where users seed algorithms, combining human and machine effort. Applications such as Amira or IMOD ask users to establish initial conditions, such as through level-set methods or flood fills, to subsequently assist the user in the segmentation process (Kremer, Mastronarde et al. 1996; Stalling, Westerhoff et al. 2005). Initial interviews with segmentation experts suggest that these algorithms can sometimes increase throughput, but at other times create more work than manual segmentation because of the high number of corrections required. When they work, they can speed up segmentation by as much as ten-fold over manual methods.

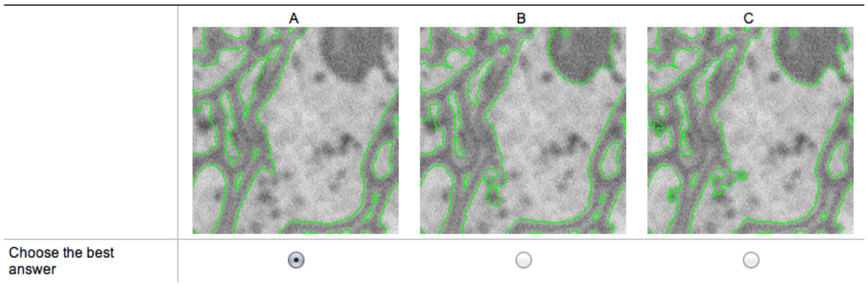
The primary strength of automated segmentation methods is the ability to distribute computation over parallel resources to achieve faster throughput, but this method lacks the accuracy necessary for scientific analysis. The alternative method of experts segmenting images is typically performed alone or in small groups, and while accurate is very slow. Examining the HCI and CSCW literature reveals precedent for applications that are both highly parallel while incorporating the input of human users (Seung 2013). These applications combine the scalability of algorithmic approaches with the human ability to resolve uncertainty, suggesting suitability for crowd-based methods. That is, online micro-task markets, like AMT, provide a mechanism for coordinating the activity of many crowd workers (i.e., achieving parallelism) while exploiting the unique characteristics of human workers (i.e., visual processing). The following section examines existing crowd-based frameworks and evaluates how they might be applied to the task of image segmentation in the neurosciences.

1.3 Three experiments in crowd-based image processing

The thesis is organized around three experiments targeted at the challenges of having a crowd of non-experts achieve results comparable to experts. This section briefly describes each experiment.

1.3.1 Converting expert tasks so they can be performed by the crowd

A key challenge in enabling Turkers to perform image segmentation is eliminating the need for expert knowledge. In the case of membrane tracing, for instance, a worker requires understanding of biological concepts to successfully distinguish between the features of interest within a cell, such as vesicles, the nucleus, endoplasmic reticulum, and mitochondria. Rather than asking workers to draw the outline of specific features, the first experiment in the thesis presents an array of competing segmentation results (i.e., obtained by automatic means) and asks users to choose the best match to a given pattern. For example, in Figure 3, workers see three alternative segmentations and are told to choose the result where the pattern of green lines is most like the description of "outer walls, like the surface of a balloon, excluding any interior features." This re-conceptualization of the segmentation task transforms the work performed from feature identification (requires expertise) to pattern recognition (does not require expertise).



	A	B	C
Choose the best answer	●	○	○

Figure 2: In the task assigned to workers, expertise is embedded in the system, reducing the knowledge required of the worker. In this example, answer A is the best match.

The key insight behind the re-conceptualization of the segmentation task is the recognition that knowledge can be embedded in mechanisms and approaches (in the case of image processing through machine learning algorithms), therefore reducing knowledge required of individual workers (Argote 1999). For instance, Nonaka and Takeuchi (1995) describe the example of expertise built into an automatic bread maker. In the initial product, the manufacturer simply mechanized what were thought to be the complete steps in making a loaf of bread. Bread produced by these machines did not taste right. A member of the design team came up with the idea of shadowing a famous baker to try to learn the secrets of the process. Through this observation they found that a special kind of kneading was essential for high quality bread. The designers were able to achieve the same effect as the special kneading by adding ribbing to the mixing paddles in the bread machine. In this case, Nonaka and Takeuchi document how the tacit knowledge of the baker was externalized in the form of the modified bread maker. A key advantage of this externalization was that an outcome previously achievable only by an experienced practitioner (i.e., the famous baker) could now be achieved through ordinary actions (e.g., measuring and adding ingredients) by the lay public.

1.3.2 Allowing the crowd to refine tasks

In online micro-task markets, communication about tasks is typically one way. Specifically, in AMT, employers create HITs, release these to users, and then harvest results. By contrast, in most settings, workers have modes of varying richness (e.g., email to face-to-face communication) where they can exchange information. For example, rapid feedback can be important in diagnosing and repairing breakdowns. In the case of crowd-based tasks that are outside the usual experience of workers, the lack of two-way channels may be a significant impediment to performance. In particular, experts may formulate requests using language and concepts that are unfamiliar to less expert audiences. The second experiment in the thesis addresses this concern by introducing the possibility that in addition to performing work, Turkers may also play a valuable role in re-formulating requests such that HITs become easier to understand and accomplish.

The key insight behind allowing the crowd to refine tasks is that even though Turkers may lack expertise in the domain of a HIT, they possess the capacity to phrase requests such that workers are more motivated or have a clearer idea of what they are expected to do. That is, Turkers may be untrained with respect to a given task, but they are very experienced (in some instances) at being Turkers. For instance, in a system where requests can be refined, early responders can use their experience of executing a task to re-shape future requests to improve the performance of subsequent workers. If successful, such an approach will show that the benefits of negotiating common ground (Clark 1996) (i.e., the give-and-take between task participants that produces mutual understanding) can be introduced into large-scale crowd activities, where typical approaches to forming common ground (e.g., dyadic conversation) are not practical.

1.3.3 Transforming private or sensitive data

A final challenge related to crowd-based image processing is that making images public may reveal information to competitors, such as in the case of labs racing to make the same discovery, or may compromise privacy, such as in the case of images made from patients. Therefore, before engaging the crowd, underlying data may need to be transformed. The third experiment in the thesis addresses this question by exploring whether transformations exist that allow the crowd to perform work unaware of the true nature or source of the data, while still yielding output that is useful when mapped back to the original data.

The key insight here is that when images have a particular “shape grammar,” (Stiny 1980) or set of rules that can be used to describe objects in the image, abstractions can be produced that preserve important characteristics of an underlying image without revealing the actual image. For example, a common image processing task is to locate occurrences of a specific feature (e.g., a sigmoid body) and then count the frequency of these features. Through transformation of an image, resulting 2D images – when “flipped” back and forth – may show spherical volumes with tubes, similar to the appearance of a cat’s eye marble. If workers, in aggregate, can find the coordinates of these “marbles” the coordinates can be used to show the location of sigmoid bodies in the actual image. This approach has tremendous potential outside of the neurosciences. For instance, if numeric data can be transformed into shapes with shape grammars, crowds could work on a wide array of sensitive data. For example, in the case of financial transaction data, transformed images could be used to detect characteristic shapes for patterns of fraud.

1.3.4 Experiments in the context of popular crowdsourcing projects

Individual contributions to crowdsourcing projects can take a number of different forms. Table 1 categorizes eight well known crowdsourcing applications into three categories: a) the crowd as a sensor; b) the crowd as a computer; and c) the crowd working collaboratively. Projects such as CoCoRaHS and eBird solicit data from the crowd as sensors to create data collection networks that would be difficult if not impossible for a single organization to build. Crowd computing projects such as

Stardust@Home, GalazyZoo, and Finding Khan present the crowd with tasks that are difficult to perform with existing algorithms. Crowd collaborative work in this context relies on the aggregate contribution of crowdworkers such as with community portals like CoCoRaHS and Finding Khan where solutions are the result of aggregate intelligence of the crowd.

	Crowd as a sensor	Crowd computing	Crowd collaborative work
CoCoRaHS(Cifelli, Doesken et al. 2005)	X	X	X
eBird (Sullivan, Wood et al. 2009)	X		
Stardust@Home (Westphal, Butterworth et al. 2005)		X	
GalaxyZoo (Lintott, Schawinski et al. 2008)		X	
Foldit (Cooper, Khatib et al. 2010)		X	X
Finding Khan (Lin 2010)		X	X
Gemident (Lee, Kapelner et al. 2009)		X	
Eyewire (Seung et al. 2012)		X	
Embedding Expertise (Chapter 2)		X	
Dynamic Questions (Chapter 3)			X
Task Transformation (Chapter 4)		X	

Table 1: Popular crowdsourcing applications in terms of worker contribution and the placement of dissertation contributions relative to existing literature.

The three experiments outlined in this dissertation for the most part build on the strategy of employing crowd workers to derive solutions difficult or impossible to accomplish with algorithms (Experiment 1: Embedding expertise and Experiment 3: Task transformation), while the second experiment (Experiment 2: Dynamic Questions) relies on the aggregate response of many users to generate and implement quality task instructions.

CHAPTER 2

Distributing Expertise: Refining cell membrane segmentation by the crowd

2.1 Introduction

This experiment explores whether it is possible to accomplish expert work by orchestrating the efforts of non-experts via Amazon's Mechanical Turk. The approach involves the use of automated algorithms to reduce a complex task, in this case identification of biologically important structures within electron microscope images, from one that requires expert insight (e.g., hand segmentation of target structures) to one that non-experts can perform (e.g., recognizing the best result from a limited set of options). The key hypothesis is that image segmentation can be transformed from a process that requires expert judgment (a scarce commodity) to one that relies on naïve pattern detection (a skill that many workers possess) – and that in doing so the crowd can achieve results similar to an expert.

2.2 Motivation

A challenge to enabling image segmentation via Mechanical Turk is elimination of the need for expert knowledge by the workers. For example, under conventional approaches, successful feature extraction requires the ability to operate the tools for segmentation and the ability to distinguish structures of biological interest, such as the vesicles, nucleus, endoplasmic reticulum, and mitochondria within a cell. Crowd workers can't be assumed to possess this kind of deep biological knowledge. Therefore, the segmentation task must be transformed from one that is difficult for the crowd to accomplish to one that is easier. That is, rather than asking crowd workers to trace the outline of particular features, workers are presented with alternate tracings generated using automated image processing. The crowd workers are then asked to choose the best match to a given pattern**,** such as, "outer walls, like the surface of a balloon, excluding any interior features" – as opposed to the expert task (e.g., find all structures that are mitochondria and trace these structures). Figure 4 shows representative output from an image processing algorithm in green on top of raw image data. Crowd workers would scan the three choices and then select the option that best matches the given pattern description (in this case option A is the best match).

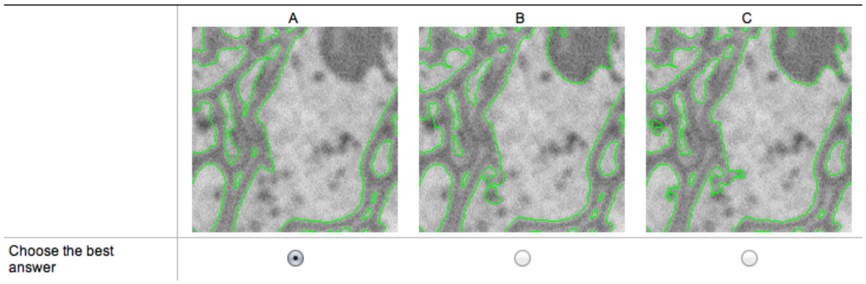


Figure 3: In the task assigned to workers, expertise is embedded in the system, reducing the knowledge required of the workers. In this example, image A shows the best result.

2.3 Methods

2.3.1 Participants

Participants were recruited from Amazon's Mechanical Turk with several worker conditions. These conditions included a requirement that workers had completed 1000 prior hits with a 95% success rate. Workers were given the option of previewing the HIT before accepting it and were provided with a compensation of $0.25. For each HIT, 50 participants were recruited for a total of 250 participants.

2.3.2 Materials

The electron microscope images used in this experiment were processed by both an expert and by machine learning algorithm targeting the cell membranes (see Figure 5). Consisting of 700x700 pixels and 50 slices in Z the selected volume was a fraction of a larger dataset. The dataset was first manually traced by an expert one slice at a time in 2D with the results composited at the end to create a 3D volume. The expert tracings were then used to train a neural network based algorithm to detect the cell membranes (Jurrus, Paiva et al. 2010). Machine learning algorithms in the neurosciences are an increasingly popular solution for segmenting large datasets (Jurrus, Hardy et al. 2009; Mishchenko 2009).

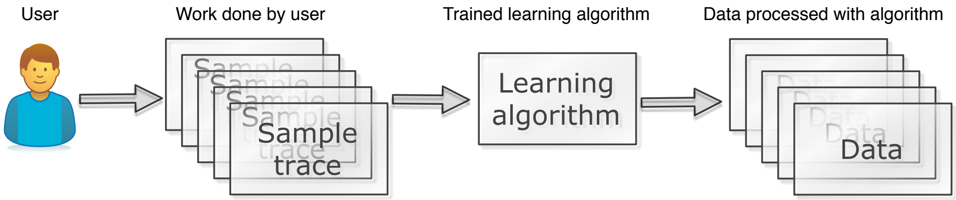


Figure 4: Training of automated segmentation algorithms. Users train algorithms with examples of what the program should do. The program then approximately applies this knowledge to similar data.

2.3.3 Design

Image data were processed using an automated segmentation technique (Jurrus, Paiva et al. 2010). During the automated segmentation process the algorithm goes through a noise reduction process. There are regions in the image where the algorithm has difficulty distinguishing between noise and membrane. Alternative renderings of these difficult regions were produced. A pipeline of codes and scripts was created (see Appendix A) to overlay the generated contours onto electron microscope images, to scale these images to a size viewable on most displays (e.g., in a web browser), and to aggregate results from crowd workers. Figure 6a shows an image before processing. Figure 6b shows the same image with candidate contours determined by the automated segmentation technique and then overlaid on the image using the pipeline of codes and scripts.

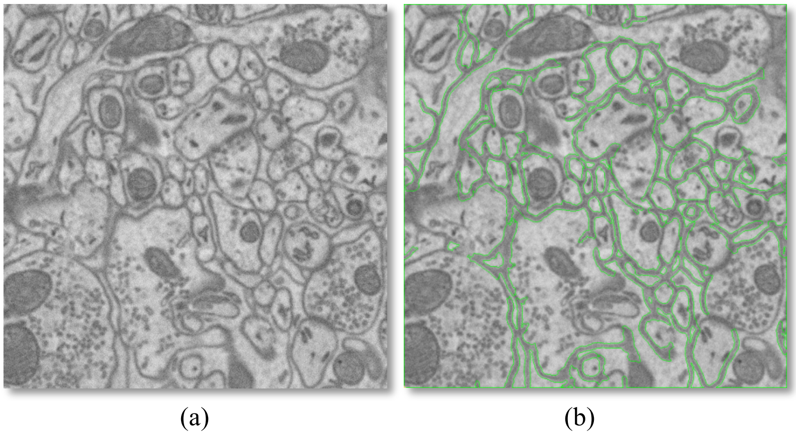


Figure 5 (a): Original unannotated dataset. (b): Original dataset with automated segmentation results overlaid on top.

2.3.4 Procedure

Participants began the experiment by selecting the image segmentation task from within the Mechanical Turk online micro-task marketplace. Once started, participants were shown sixteen instances of three alternative tracings and were asked to identify the option that best matched the description "outer walls, like the surface of a balloon, excluding any interior features." Supporting the distribution of the task to a broad community of workers, a collection of web-based technologies was used including survey software, cloud based file hosting, and Amazon's Mechanical Turk.

Image cropping and assembly

Before the images were distributed to workers, they required processing to decrease the size suitable for distribution on the web. In support of this requirement a script was written by the author included in Appendix A. Three volumes of images 700x700x270 were divided into smaller images. These images were then uploaded to cloud storage easily accessible to Amazon’s Mechanical Turk and Qualtrics software.

Qualtrics

Qualtrics is an online survey suite with a number of features for dynamic content including web services integration. Each worker is tasked with identifying the best performing algorithm given three choices. The task was posed as a multiple choice question within the Qualtrics software, with each decision presented on its own page. Because of the large number of questions, links to data, and surveys that needed to be generated, the creation of the surveys was automated with a BASH shell script written by the author included in Appendix A. The shell script programmatically creates the survey and all of the correct dynamic links to the image data hosted on Amazon’s S3. Each survey also included a random six digit number used to validate participation within the Mechanical Turk environment.

Amazon AWS

Amazon’s web services (AWS) are a collection of cloud technologies including facilities for compute and data storage. Amazon’s S3 (Simple Storage Service) is a cloud storage service. Images of the segmentation data were stored on Amazon’s S3 infrastructure and referenced within the Qualtrics surveys. Being an Amazon hosted technology, the pairing of Amazon S3 and Amazon’s Mechanical Turk ensured compatibility and implicit assurance of performance between the two systems. Data on S3 were uploaded and the permissions were modified to allow public read access to all of the data.

Amazon's Mechanical Turk

As described previously, Mechanical Turk is a micro-labor marketplace where workers are paired with employers for very small tasks. Mechanical Turk offers a number of tools and configurations for managing workers and distributing work. Additional details of the configuration of the tasks are included in Appendix F. Participants completed the task with a median duration of 8 minutes, with some participants taking longer than 30 minutes. A total of 50 workers were assigned to each HIT, with five HITs issued for a total of 250 workers. The data from workers were processed to find consensus from the workers. Consensus was determined by tallying answers generated by workers.

2.4 Results

The results of this experiment demonstrate that workers can make decisions based on pattern recognition that closely coincide with expert judgments given the same task. Workers were presented with a panel of three alternatives: A, B, and C as shown in Figure 7. Workers were asked to select the image where the algorithm best outlined the cell membrane, as highlighted in Figure 8, while balancing the detection of other objects within the image, such as vesicles seen in Figure 9.

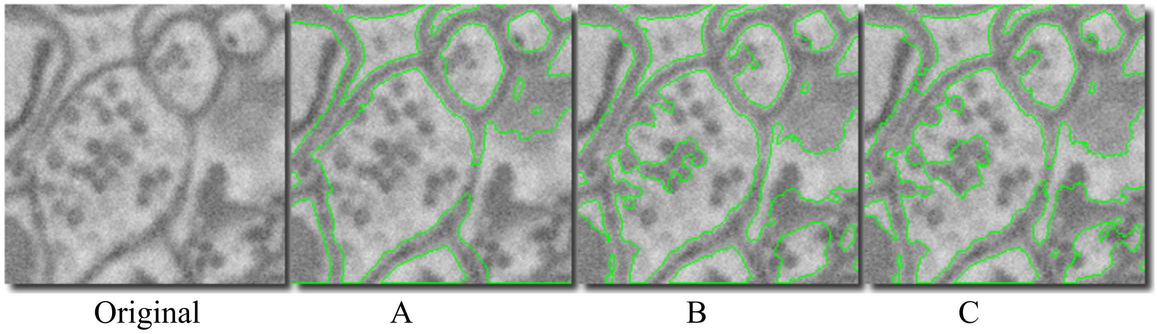


Figure 6: Image choices presented to workers, asked to pick the image where the green labels outline the cell membranes.

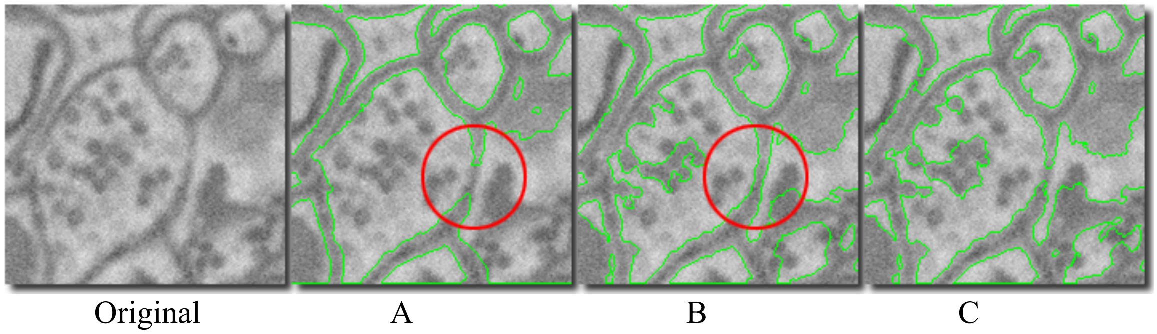


Figure 7: The algorithm for removing noise was overly agressive, missing part of the membrane. The membrane segment was properly detected in panel B.

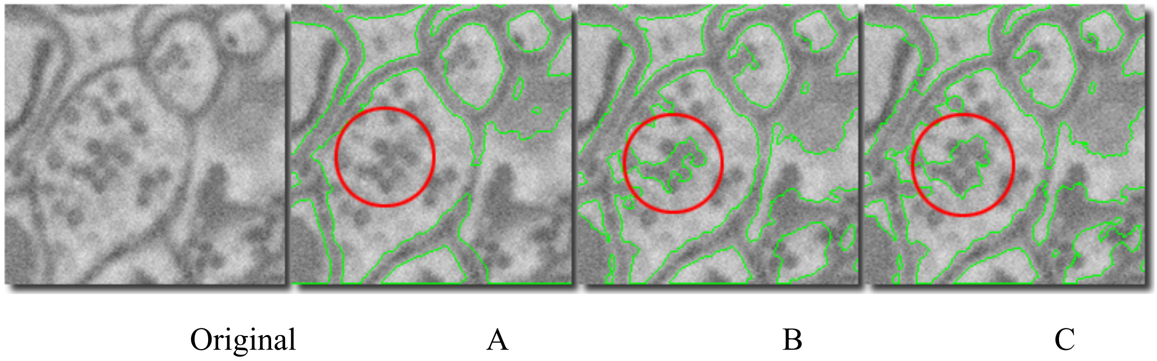


Figure 8: Noise reduction was successful in Panel A, but not in Panels B or C. Panels B and C incorrectly label vesicles as part of the cell membranes.

The gold standard for segmentation in the neurosciences is the expert worker. To better understand the performance of the workers solicited from Mechanical Turk, their results were compared to the results of an expert worker given the same task. Table 2 summarizes the results from both the novices and the expert.

Image number	Novice/Expert agreement	Percentage agreement

1	14/16	87.5%
2	10/16	62.5%
3	12/16	75%
4	15/16	93.8%
5	16/16	100%

Table 2: Agreement between novice and expert of the best segmentation. Novices and expert agree on average 83.8% of the time.

The data available in Appendix A report the performance of the crowd relative to the responses of the "gold standard" or the performance of the human expert. These responses range from a low of 62.5% to a high of 100% with an average of 83.8%.

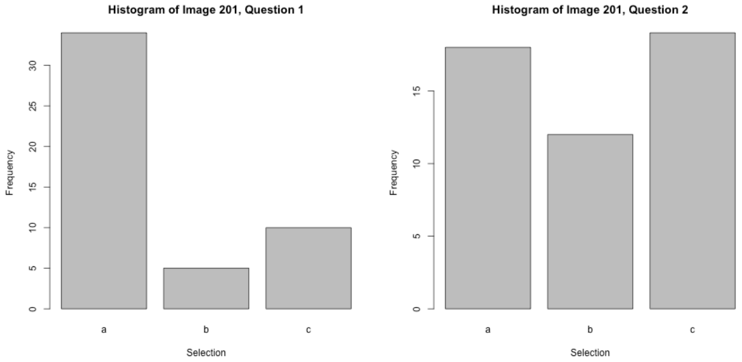
In the neurosciences, segmentation of features from electron micrographs is typically performed by a single person. Even segmentation of whole cells requiring several months or years of labor are performed by a single person (Lenzi, Runyeon et al. 1999; Sosinsky, Deerinck et al. 2005; Noske, Costin et al. 2008). There is no indication that a single expert is insufficient to segment even complex structures from electron tomograms. In fact Martin et al. showed that human performance in segmentation tasks is highly consistent (Martin, Fowlkes et al. 2001). In his work Martin generated performance criteria and measured the performance of several participants tracing objects from natural scenes. He found that in a majority of his measurements, the errors between workers peaked around zero suggesting highly consistent work tracing objects from natural scenes between workers. Furthermore, the computer science image processing community has embraced the concept of a single expert. In pursuing their goal to reproduce human performance in segmentation of objects in natural scenes like electron tomograms, they consistently refer to the effort of a single person as the "gold standard" (Jurrus, Hardy et al. 2009; Mishchenko 2009; Jurrus, Paiva et al. 2010; Giuly, Martone et al. 2012; Giuly, Kim et al. 2013). In these studies, the performances of the algorithms are compared to the results of a single domain expert. Following the example of the computer science community and support that people are highly consistent in tracing objects from natural scenes, a single expert user was recruited to serve as the gold standard in these tests.

To better understand the distribution of responses by the Turkers, I examined the results from the crowd (for Image 1) and compared them to the expert's reasoning as determined from a semi-structured interview.

Question Number	Algorithm A	Algorithm B	Algorithm C	Novice consensus	Expert
1	**40**	9	4	A	A
2	**34**	6	13	A	C
3	**27**	11	15	A	A
4	**22**	15	15	A	A
5	**40**	4	9	A	A
6	**39**	10	4	A	A
7	**28**	13	12	A	A
8	**32**	15	6	A	A
9	**31**	12	10	A	A
10	**31**	15	8	A	A
11	**28**	19	6	A	B
12	**25**	17	11	A	A
13	**37**	6	10	A	A
14	**28**	16	9	A	A
15	15	**26**	12	B	B
16	**25**	18	10	A	A

Table 3: The results from Image 1, a representative dataset, reporting the distribution of selections by Turkers compared to the expert's selections. Randomized presentation of the results yielded no difference in the distribution of responses.

The results shown in Table 3 indicate strong agreement between the Turkers and the expert, when using a plurality rule. That is, the response receiving the most votes from the Turkers agrees with the expert answer for fourteen of the sixteen images (87.5%). Applying a stricter test using a majority rule (i.e., one response receives 50% or more of the Turker selections) still shows agreement on thirteen of the sixteen images (81.3%). Examining the distribution of answers, there are instances where the consensus of the crowd is clear and in agreement with the expert, such as with Questions 1, 5, 6, 8, and 13 where the majority answer collected more than 50% of the votes. Bar charts, such as in Figure 10a are a good way to visualize questions where Turkers were in high agreement. However, there were instances where the Turkers were more divided such as with Questions 3, 4, 7, 9, 10, 12, 14, and 16. Again, bar charts, such as in Figure 10b are a good way to visualize questions where Turkers failed to reach strong consensus. (Additional histograms for the remaining images and questions are included in Appendix A.) Finally, there were two cases, Questions 2 and 11, where the majority answer by the Turkers received more than 50% of the votes – but this answer disagreed with the expert.



(a) (b)

Figure 9: Plot of crowd responses for each question. Additional questions are included in Appendix A. The first image demonstrates clear consensus of the crowd (a) while the second image demonstrates a split in identification by the crowd (b).

The responses from the crowd are categorical, with each participant choosing from three independent choices. To additionally visualize the degree of consensus within the crowd, a convenience assumption was applied where each choice was treated as a scalar value and the differences between each choice were equal. In addition, a lack of response by the crowd was assigned a value of zero. With these assumptions, several boxplots were generated (e.g., Figure 11)showing the response of the crowd (black circles) relative to the expert (red circles). Error bars in Figure 11 represent variance around the crowd's choices. Additional plots visualizing responses of the crowd relative to experts for all images are included in Appendix A.

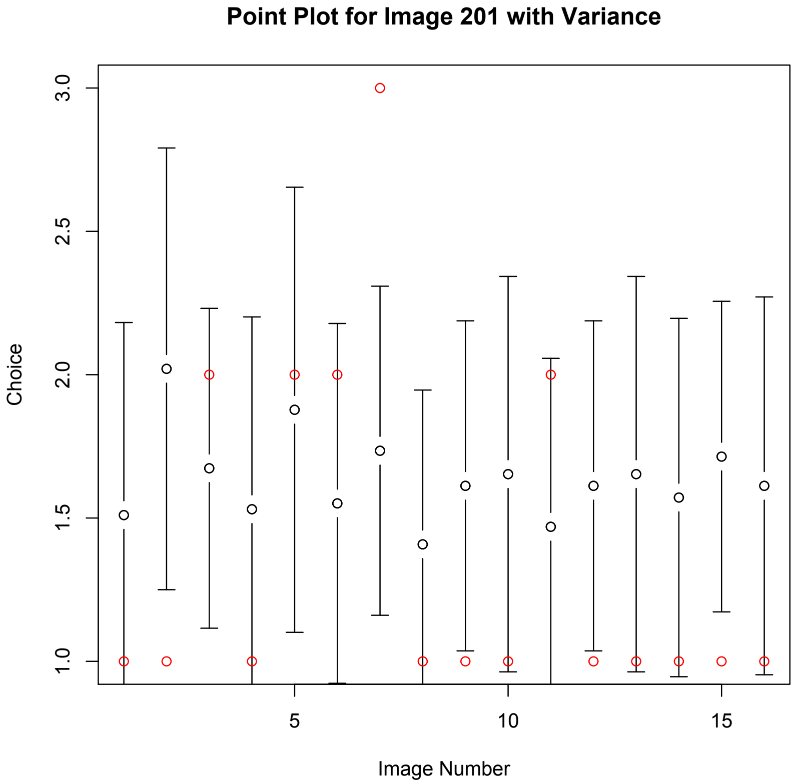


Figure 10: Point plot of expert responses (in red) and novice responses (in black). Error bars report variance calculated in R. Blank responses from the crowd are substituted with zero. Additional plots in Appendix A.

After collecting the responses from the expert, the expert participated in a semi-structured interview discussing the answers of the crowd with particular emphasis on questions where there the Turkers differed or lacked strong consensus. In contrast to a pure pattern recognition process, the expert looked for "elimination criteria" such as "bad joins" and "prioritizing good connections over tracing a few stray vesicles." In a number of cases the expert had trouble making a sure determination of the best selection because of limited contextual information, but made the selections given the available information. Asked about the differences in answers for Questions 2 and 11, the expert indicated “It depends on what you’re prioritizing.” The instructions for the task and the task itself required judgment and interpretation of the instructions and task. In this case it appears that the expert prioritized the exclusion of vesicles over improved detection of the membrane. The expert reviewing the responses of the Turkers also indicated that in the cases where there was no clear majority of the Turker’s responses, “it could have gone either way.”

2.4.1 Cost

The cost of segmentation is mostly the time of the segmentation expert as they go through the stack of images tracing the features of interest with a mouse. Depending on the desired resolution, the analyst may skip a consistent number of slices in Z when tracing objects. They will then mesh the results, combining all of the 2D images into a 3D object. Looking at the 3D model, they will determine if the graphical representation matches their mental representation. If there are differences, experts will then go back and correct contours by adding or subtracting features until the graphical and mental model of the structures match.

The process is different when correcting beginning with the output from learning algorithms. Segmentation experts will process objects using a flood fill tool to fill in the target structure, stopping occasionally to fix the errors made by the learning algorithm. The process is described to be “click, click, click, click, trace” opposed to just tracing. Because of the effort required to detect and correct errors, it is reported to take the same amount of time to correct machine learning results as it is to simply trace manually. In addition to the cost of the time to segment the data by the expert, there is considerable effort required to train and tune automated algorithms for a particular dataset. Each dataset is different and requires tuning of parameters. Each algorithm is generally also optimized for a particular structure.

Crowdsourcing the algorithm to detect some of the errors made by the algorithm requires the additional effort of configuring the task with the techniques outlined in the methods section. In this example, each image was processed by 50 workers at a rate of $0.25 per worker, resulting in a cost of $13.75 per image. The relatively large sample was useful for illustrating those questions where the crowd clearly demonstrated consensus versus those where the crowd was evenly split; distinctions that would not have been apparent with smaller sample methods such as panel or consensus models. However, panel models would reduce costs by only executing additional HITs in the event that preliminary HITs did not reach some threshold of agreement (Little, Chilton et al. 2009). For example, in the case of Questions 1 and 5 from the image comparison experiment, where there was very high consensus, additional HITs beyond the first several did not add much new information (i.e., the consensus was more immediately apparent for these questions). Therefore, one could preclude further HITs if, for instance, five out of five initial responses agreed. Implementing such an algorithm could decrease costs by the difference between the cost of collecting a full sample of HITs and the cost of collecting only those HITs required to be reasonably certain of consensus (i.e., according to some pre-determined level, such as six consecutive agreements, which would cost $1.50) shown in Table 4. Use of still more efficient algorithms, such as the agreement model used by Von Ahn's ESP game, could decrease the cost of analysis of each image to as little as $.50 per image.

Method	Cost per image
Expert	Amortized costs of training the expert (e.g.,

	tuition, stipend) and ongoing costs (e.g., hourly rate, space, equipment)
Mechanical Turk (as implemented)	$13.75
Mechanical Turk (modeled after Little et al.)	$1.50
Mechanical Turk (modeled after Von Ahn et al.)	$0.50

Table 4: Outline of costs associated with soliciting worker participation in Experiment 1 for different agreement methods.

2.4.2 Limitations

The accuracy of the work performed by workers is bound to the accuracy of the best possible result from the collection of image segmentation algorithms used. Because workers are choosing the best answer from a panel of pre-computed options, they cannot deviate from what the computer has already calculated. This may be alleviated by the introduction of simple drawing tools or by using multiple algorithms with greater diversity.

The current implementation is also not optimized for the lowest cost. Iterative improvement schemes (Little, Chilton et al. 2009) (Ipeirotis, Provost et al. 2010) can be implemented to reduce the number of workers required, rather than relying on a consensus model (one of the more inefficient algorithms available, but one of the easiest to implement).

In addition to the training of the automated algorithms, there is a need to create instructions for workers on Mechanical Turk. This step cannot be automated and is specific to the structure being traced and the alternatives being presented to workers. This process is potentially difficult as experts may struggle to effectively communicate with a novice workforce with minimal feedback. This disconnect between experts and novices is potentially alleviated with the use of “Dynamic questions” (described in the next chapter).

2.5 Discussion

Significant infrastructure development is required to attempt the kind of experiment described in this chapter. Specifically, before collecting a single data point, I had to:

- Identify appropriate image-processing algorithms;
- Write code for content distribution, image presentation, compensation for workers, image preparation, image comparison, aggregation of results, and image reconstruction – across multiple software platforms (e.g., AWS, Qualtrics, and MT).

While the automated algorithms used in this experiment are freely available, the content distribution, presentation, worker compensation, and image reconstruction algorithms were custom developed by me for this dissertation (see Appendix A). While not necessarily generalizable to all future applications, the code and systems provide reference implementations where none existed before.

In terms of the findings, crowd workers demonstrated a strong agreement with an expert when performing the same task (83.8% agreement). Further, in cases where the crowd was unable to achieve a clear consensus, the expert concurred that these were more ambiguous situations (i.e., the crowd was legitimately split between responses and the expert thought multiple experts would be similarly split). Additional instruction for these ambiguous cases would be one way to improve performance. For example, a pilot run of an image comparison task may identify those comparisons where workers struggle and these problematic comparisons could be addressed through guidance from an expert (i.e., in the form of modified instructions for these comparisons in subsequent HITs).

Finally, performance could be improved by culling crowd workers over iterative tasks according to accuracy and efficiency, resulting in a smaller set of workers still able to achieve desired results.

CHAPTER 3

Dynamic Questions: Worker refined task instructions in micro-labor tasks

3.1 Iterative refinement of task instructions by the crowd

A critical challenge in distributing scientific work to the crowd is to minimize the expert knowledge required by the worker to engage in the task. Minimizing requirements maximizes the number of eligible workers. The previous chapter examined a technique for embedding expertise into the system by instilling the knowledge of the expert via an automated algorithm. The embedded knowledge was then applied to data where the task of the worker was reduced to a pattern recognition task.

Continuing to seek how to distribute scientific work to the crowd, the next two chapters separates the task issued to workers into two parts: a) modification of task instructions; and b) modification of the underlying data

We see in work by Von Ahn (Von Ahn 2006), that the instructions issued to workers are a powerful framing mechanism that can transform a dull and repetitive task to an entertaining one. Von Ahn et al. created a series of games oriented around tasks ("games with a purpose") difficult for computers to accomplish, but trivial for human users, with one example oriented towards labeling of images. Two users were tasked with labeling images, aiming to agree on the tags used to describe the image, developing a semantic index for the image – a task difficult to accomplish with algorithms. By framing the task as an entertaining game, he was able to solicit the attention of a large number of workers who might otherwise apply the same cognitive effort to less productive games such as Solitaire.

In interactions with crowd workers, workers are only identified by their worker ID, so the human qualities of workers are not as apparent as in traditional work settings. Despite the sparse work context, Turkers do seek ways to make their work more meaningful and enriching. For example, Turkers will sometimes reject poorly formed HITs in favor of other work (Silberman et al. 2010). Additionally, Turkers leverage external information sources such as Turkopticon (Silberman et al. 2010) that provide repositories of Turker-generated feedback. These forums allow Turkers to share information, detailing past experiences with specific employers, and potentially creating consequences for employers with undesirable work or delinquent payment practices.

While clearly important, constructing well-articulated AMT tasks is challenging. With scientific work and the crowd there is the additional challenge of the potential divide between experts and novices based on language and approach. For example, Larkin et al. (Larkin, McDermott et al. 1980) found that experts and novices construct different mental models and solve problems differently. With the case of experts, they may use terms, logical constructs, or background knowledge not accessible to a novice. Compounding this divide, communicating and soliciting feedback in AMT is not always an obvious process. While in face-to-face communication an employer and worker might observe body language, ask and respond to questions, or refine the task, workers on

Amazon’s Mechanical Turk have little motivation and few mechanisms to provide feedback to an employer. The only feedback mechanism typically available to Turkers is through email – resulting in a loss of anonymity and with no guarantee of a timely response.

This chapter examines a methodology for iterative feedback and refinement of task instructions by Turkers termed “Dynamic questions.” With Dynamic questions, Turkers submit alternative task instructions and vote on what they believe to be the best instructions. This chapter demonstrates how the Dynamic question system can result in questions that have more detail, greater clarity, and feature alternative vocabulary for technical terms.

3.1.1 Dynamic Questions

The dynamic questions system solicits feedback from Turkers to aid in the refinement of the task description. Every time a worker accepts a HIT, the Qualtrics software queries a Google Fusion Table and dynamically builds a survey using a set of task instructions that received the most votes from previous Turkers. At the completion of the task, workers are asked to submit their own interpretation of the question and vote from a panel of other user-contributed questions. At the conclusion of the task, the number of votes for each submitted alternate task instruction set is tallied by a PHP script that then updates a Google Fusion Table. Throughout the batch of HITs, the instructions presented to the workers can vary depending on which set of instructions receive the most votes.

3.1.2 One-way communication

The challenge of one person speaking to many in a unidirectional communication stream is not new. In an essay, Norman talks about how the product designer has limited opportunities to speak to the consumer through the design of objects (Norman 2004). Often in product design, designers speak to consumers through affordances and the design of the object where the product can speak to the consumer, a model that Norman coins as the System Image model. This communication between designer and consumer is mostly one way and is slow to iterate as seen in Figure 12 This model of communication is similar to HITs in Mechanical Turk where employers create a single HIT that is distributed to many workers with little feedback.

3.1.3 Boundary Objects

Norman's System Image model also bears a resemblance to models of boundary objects discussed by Star and Greisemer (Star and Griesemer 1989). Unlike the relationship between designers and consumers, there is bi-directional communication with boundary objects where these objects are the center of communication between two or more groups. A boundary object between a novice and an expert can be used to

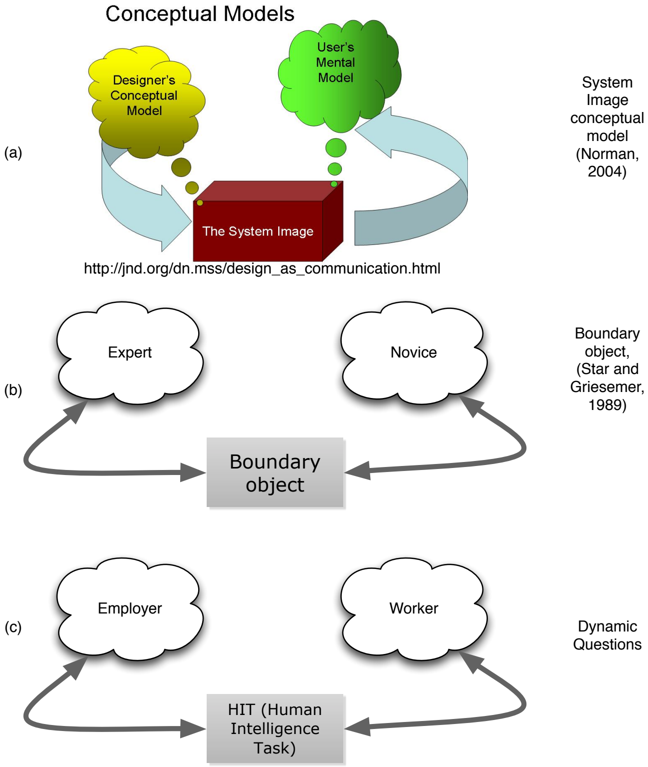

Conceptual Models
Designer's Conceptual Model
User's Mental Model
(a)
The System Image
http://jnd.org/dn.mss/design_as_communication.html
System Image conceptual model (Norman, 2004)
(b)
Expert
Novice
Boundary object
Boundary object, (Star and Griesemer, 1989)
(c)
Employer
Worker
HIT (Human Intelligence Task)
Dynamic Questions

Figure 12: The flow of information between parties in (a) Norman's system Image model (b) boundary objects, and (c) dynamic questions. With Dynamic questions, the HIT changes based on feedback from workers.

negotiate understanding between the two parties.

3.1.4 Dynamic Questions as collaborative information artifact

Blending these different bodies of literature, and looking at the AMT HIT as an information object (Buckland 1991), the HITs created by employers don't necessarily need to follow the model of creating a HIT once and distributing it to many. Orienting the HIT more like a boundary object and taking advantage of modern web technologies, a HIT can be created that sits between the worker and the employer and is dynamically changed to better accommodate the needs of the worker as illustrated in Figure 12 These two parties may have differing perspectives of the same information object based on their expertise. For example a neuroscientist will use specific vocabulary and logical constructs for an image of a tissue sample. A novice will use different vocabulary when interacting with the same image. It's also possible that workers attempting to teach another worker a specific role will be able to provide specific insight that experts may overlook (Rochlin, La Porte et al. 1987).

3.2 Pilot test

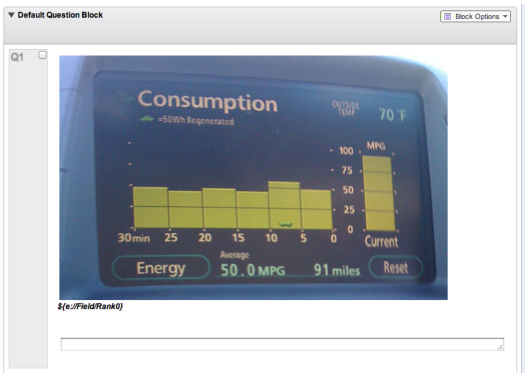
To assess the feasibility of the dynamic questions system, a prototype was implemented using a synthetic task. The pilot test asked workers to interpret a video screen from a car and calculate the fuel consumed for a theoretical trip shown in Figure 15.

3.2.1 Participants

Participants were recruited from Amazon's Mechanical Turk with several worker conditions. These conditions included a requirement that workers had completed 1000 prior hits with a 95% success rate. Workers were given the option of previewing the HIT before accepting it and were provided with a compensation of $0.25. For each trial, a total of 100 participants were recruited for a total of 400 participants.

3.2.2 Materials

The author created a word problem asking workers to calculate the amount of fuel consumed based on an image from a 2008 Prius trip computer displayed in Figure 13. This image was embedded into a Qualtrics survey, distributed to workers using Amazon’s Mechanical Turk described in detail in section 3.2.4.



${e://Field/Rank0}

Figure 13: Image presented to workers in Qualtrics survey software generated by dynamically querying a Google Fusion Table.

3.2.3 Design

The pilot experiment followed a two-group experimental design. The control group had a static set of questions while the experimental group performed tasks dynamically generated at runtime. In the experimental group, input from the crowd was incorporated into the instructions presented to the experimental group. The data collected included vote information, satisfaction data, and user contributed instruction sets from Qualtrics, Amazon’s Mechanical Turk, and Google’s Fusion Tables.

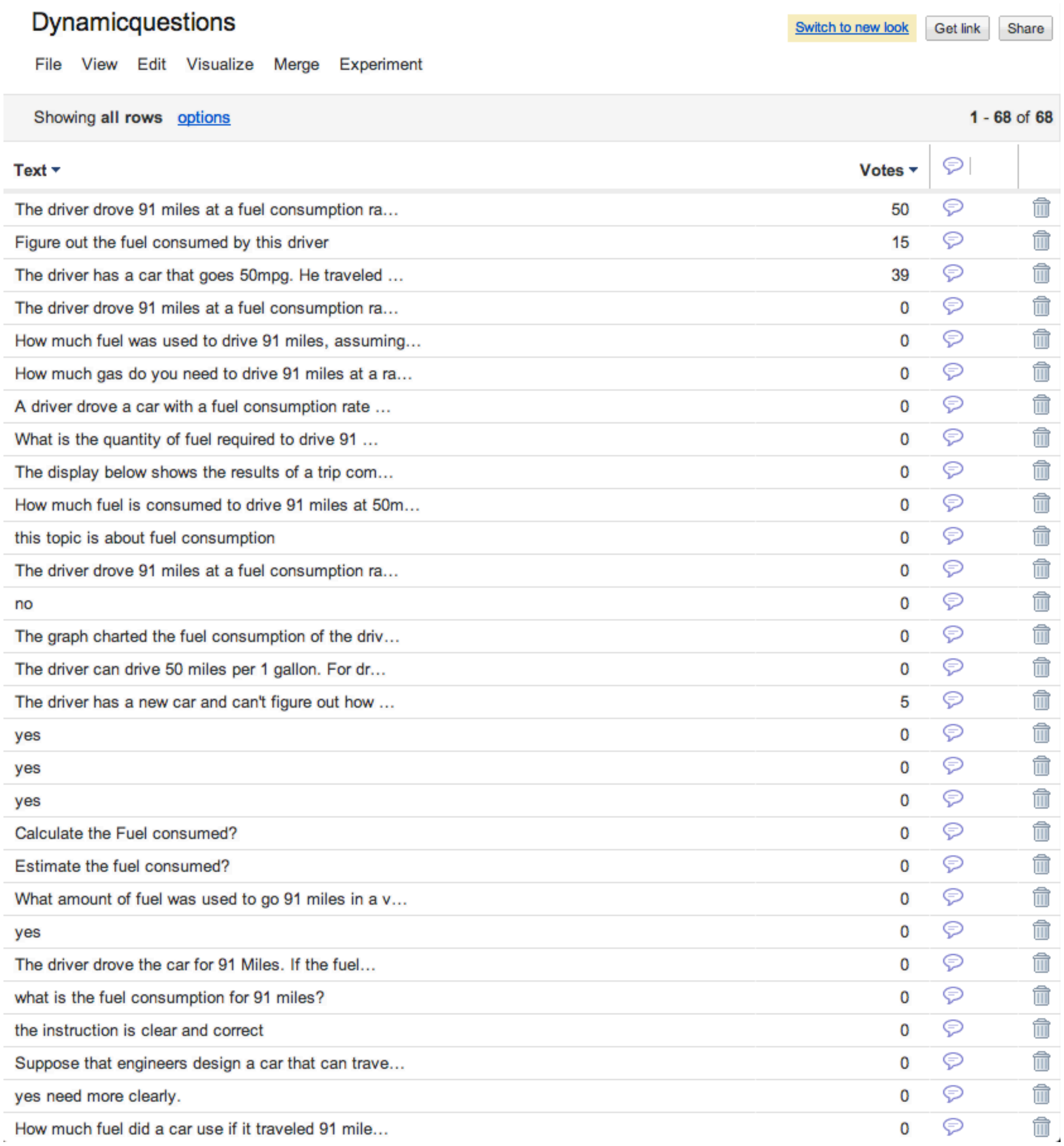
3.2.4 Procedure

To create a dynamic task, the employer needs to create a system that solicits and aggregates feedback. In the second case of dynamic questions, the system needs to dynamically incorporate the feedback into the tasks performed by workers. This system is enabled with a collection of technologies.

Creating a dynamic question in a Mechanical Turk HIT involves combining several technologies including Google's Fusion Tables, Qualtrics, custom PHP scripts, and Amazon's Mechanical Turk. Generally, a survey in Qualtrics is created using an embedded field to populate the question for the HIT. This question is queried from a Google Fusion Table hosted on Amazon's EC2. The following documents in detail the steps required to build dynamic questions in Mechanical Turk.

Google Fusion tables

Google Fusion Tables are SQL style tables with a web GUI and APIs for access through a number of different languages. They are freely hosted and leverage Google's extensive infrastructure. To create a dynamic question, the author created a Google Fusion Table, deleting all of the default columns, and adding a "Votes" and "Text" field. The table is the active repository for questions and recording the number of votes issued to each question. In this case, the table was seeded with four initial values as seen in Figure 14. The Google Fusion Table can be viewed and edited by the investigator at any time during the data collection process.



Text	Votes
The driver drove 91 miles at a fuel consumption ra...	50
Figure out the fuel consumed by this driver	15
The driver has a car that goes 50mpg. He traveled ...	39
The driver drove 91 miles at a fuel consumption ra...	0
How much fuel was used to drive 91 miles, assuming...	0
How much gas do you need to drive 91 miles at a ra...	0
A driver drove a car with a fuel consumption rate ...	0
What is the quantity of fuel required to drive 91 ...	0
The display below shows the results of a trip com...	0
How much fuel is consumed to drive 91 miles at 50m...	0
this topic is about fuel consumption	0
The driver drove 91 miles at a fuel consumption ra...	0
no	0
The graph charted the fuel consumption of the driv...	0
The driver can drive 50 miles per 1 gallon. For dr...	0
The driver has a new car and can't figure out how ...	5
yes	0
yes	0
yes	0
Calculate the Fuel consumed?	0
Estimate the fuel consumed?	0
What amount of fuel was used to go 91 miles in a v...	0
yes	0
The driver drove the car for 91 Miles. If the fuel...	0
what is the fuel consumption for 91 miles?	0
the instruction is clear and correct	0
Suppose that engineers design a car that can trave...	0
yes need more clearly.	0
How much fuel did a car use if it traveled 91 mile...	0

Figure 14: Google Fusion Table, a free SQL like resource accessible by graphical user interface or API. A Google Fusion Table was used to keep track of the worker submitted questions and number of votes each individual question received. This Fusion Table contains the contributions from several workers and the tally of votes for the vaarious contributions.

Qualtrics Integration

Qualtrics is an online survey suite with a number of features for dynamic content including web services integration and the ability to dynamically substitute text within surveys. Figure 15 shows the survey flow used in a dynamic question where the display of instructions is followed by several PHP calls to query a random text entry from the

Google Fusion Table. The survey flow then includes the ability for workers to contribute their own version of the question and ranking of the questions back into the Google Fusion Table. The end of the survey generates a random six digit number used to validate participation within the Mechanical Turk environment.

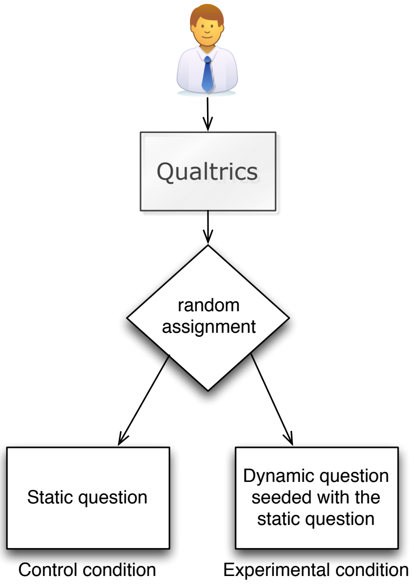
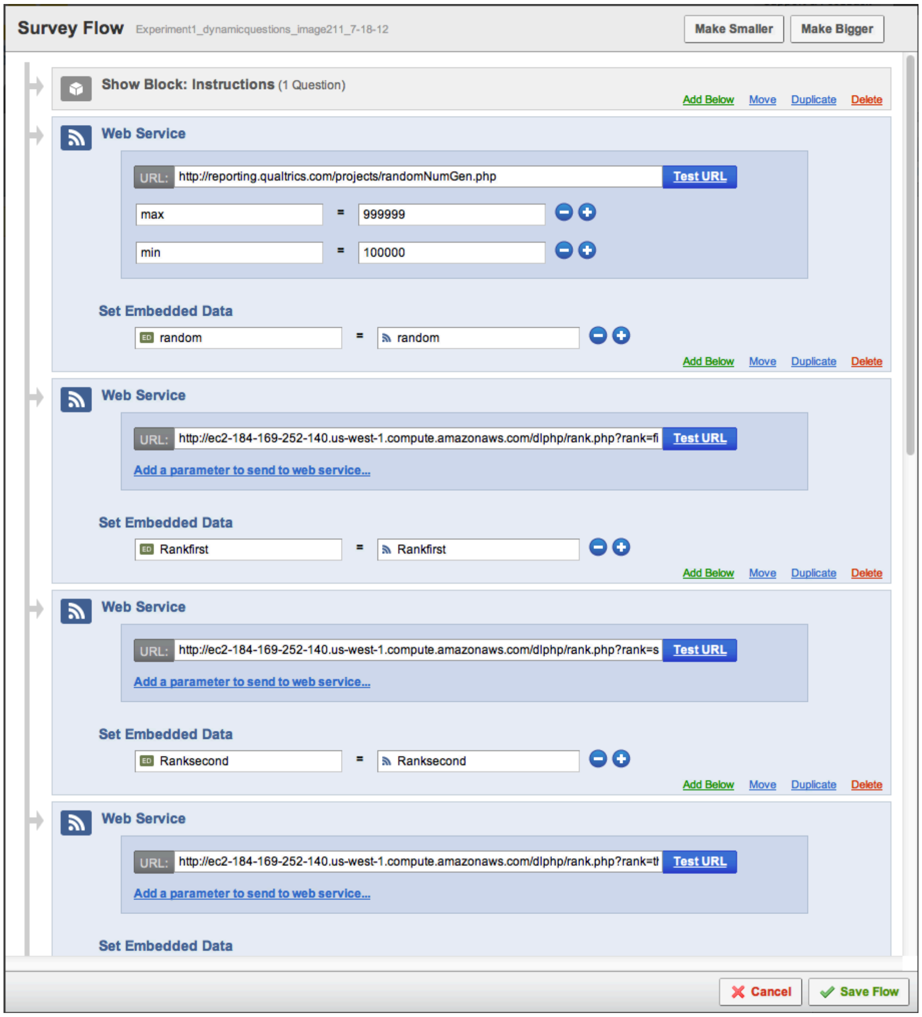


Figure 15: Experimental flow for participant assignment into two groups



Survey Flow Experiment1_dynamicquestions_image211_7-18-12

Show Block: Instructions (1 Question)

Web Service

URL: http://reporting.qualtrics.com/projects/randomNumGen.php

max = 999999

min = 100000

Set Embedded Data

random = random

Web Service

URL: http://ec2-184-169-252-140.us-west-1.compute.amazonaws.com/dlphp/rank.php?rank=fi

Set Embedded Data

Rankfirst = Rankfirst

Web Service

URL: http://ec2-184-169-252-140.us-west-1.compute.amazonaws.com/dlphp/rank.php?rank=s

Set Embedded Data

Ranksecond = Ranksecond

Web Service

URL: http://ec2-184-169-252-140.us-west-1.compute.amazonaws.com/dlphp/rank.php?rank=tl

Set Embedded Data

Figure 16: Qualtrics survey software flow including dynamic querying of questions from the Google Fusion Table.

PHP scripts

Interfacing the information contained in the Google Fusion Tables and the Qualtrics survey software are several PHP scripts written by the author specifically for this purpose included in Appendix B. There are three scripts in total that work together. The first script adds worker-contributed text to the Google Fusion Table. The second script queries the Fusion Table and returns a random selection from a specified quartile or rank among the entries in the table. The last script updates the vote count of a specified text entry in the Fusion Table.

Amazon’s AWS

Two technologies are used from Amazon’s Web Services platform, EC2 and S3. Amazon’s EC2 is a virtualized compute architecture that allows a user to launch a computer instance using Amazon’s physical infrastructure. The EC2 instance was created by the author with an installation of Ubuntu server 12.04 LTS. Once configured, the machine was installed and configured as a LAMP (Linux, Apache, MySQL, PHP) server. Amazon’s S3 is a distributed, high reliability cloud storage system. The images distributed to workers in the Qualtrics survey was hosted on Amazon’s S3 as public images.

Amazon’s Mechanical Turk

Another Amazon product, Mechanical Turk is a micro-labor marketplace where workers are paired with employers for very small tasks. Coined as the artificial artificial intelligence, Mechanical Turk offers the programmatic accessibility of artificial intelligence with the cognitive capabilities of people, and is often used to distribute small tasks that are difficult to perform with algorithms but is relatively easy for human workers. Common examples of tasks on Mechanical Turk include tasks such as image tagging, categorizing, and content creation. While identifying all of the objects in a scene would be difficult for a machine to accomplish, a human worker can create image tags of all of the objects in the image. Mechanical Turk offers a number of tools and configurations for managing workers and distributing work. A Qualtrics survey was created by the author to render the question and distribute collect responses from workers. In this task, workers were required to have already completed 1000 hit. Additional details of the configuration of the tasks are included in Appendix F.

3.2.5 Results

The dynamic questions system culled several user contributed contributions. These included re-phrasing the question, nonsense questions, and replication of the original question. One stand-out contribution by workers included reformulation of the task into metric units. A new HIT was constructed, containing the same image, compensation, but with updated task instructions. The new HIT re-phrased the task to the following, "The driver drove 91 miles at a fuel consumption rate of 50mpg. How much fuel has the driver used? 91 miles means 146.45 kilometers. 50 mpg means 21.26 kmpl." The results from rephrasing the question resulted in accuracy of 84% with the same number of 103 participants. Additional detail and the results can be seen in Appendix G. The addition of the metric unit conversion as suggested by a worker made a significant difference in performance.

The results of the pilot test indicated an increase in performance of workers in an experimental group compared to a control group as a result of using the refined instructions. In the pilot test, workers were asked to perform a simple calculation based on the information found in a graphic. Specifically they were given the instructions, "The driver drove 91 miles at a fuel consumption rate of 50mpg. How much fuel has the driver used?" workers completed the task with 57% accuracy.

The results of the pilot test suggested the potential for Turkers to influence or inform task instructions, in this case significantly improving performance. To better understand the implications for a neuroscience-based task, four additional experiments were performed.

3.3 Experiment 1: Counting mitochondria

To better understand the implications of Dynamic questions to a task closer to the driving application, this experiment asked workers to identify the number of mitochondria in an image. They were given a page of instructions with a detailed description and an image with examples highlighted in red boxes, as shown in Figure 15. In the control group, workers saw a static description of the task. In the experimental group, input from the crowd was used to refine the task description over subsequent HITs.

3.3.1 Participants

Participants were recruited from Amazon's Mechanical Turk with several worker conditions. These conditions included a requirement that workers had completed 1000 prior hits with a 95% success rate. Workers were given the option of previewing the HIT before accepting it and were provided with a compensation of $0.25. For the trial, a total of 100 participants were recruited.

3.3.2 Materials

Workers were presented with a page of instructions (shown in Appendix B) including an electron micrograph with mitochondria. In the training image shown in Figure 17, the mitochondria are highlighted in red boxes. The workers were asked to identify all of the mitochondria in Figure 18.

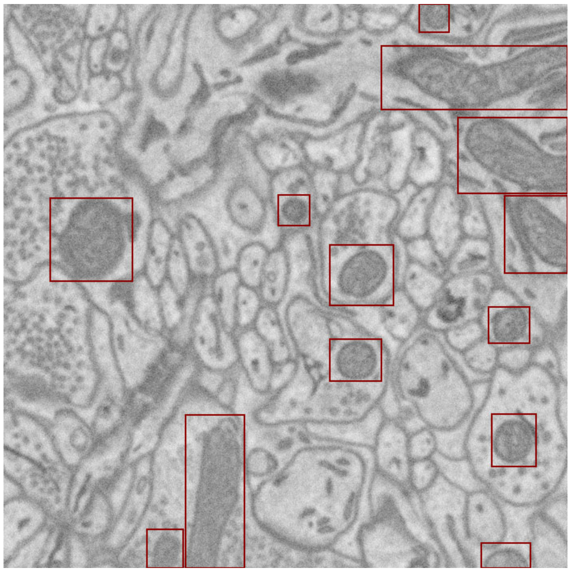


Figure 17: Electron micrograph image presented to workers as a training image. Tasked with finding the number of mitochondria in the image, examples are outlined in red.

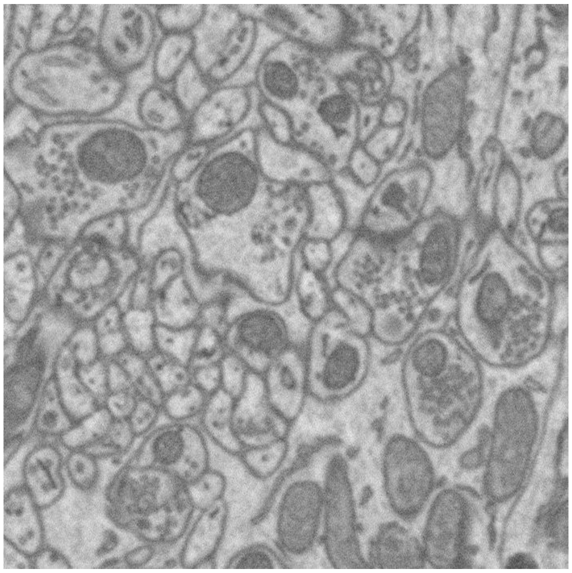


Figure 18: Workers are tasked with finding all instances of mitochondria in this electron micrograph. Both experimental and control groups successfully performed this task.

3.3.3 Design

The experiments followed a two-group experimental design. The control group had a static set of questions while the experimental group performed tasks dynamically generated at runtime. The data collected included vote information, satisfaction data, and user contributed instruction sets from Qualtrics, Amazon's Mechanical Turk, and Google's Fusion Tables.

3.3.4 Procedure

The procedure is identical to the procedure used in the pilot test described in detail in 3.2.4.

3.3.5 Results

In this task, workers in both the experimental and control groups performed very well, approximately matching the performance of an expert. Table 5 shows the performance of the performance of both the control and experimental groups relative to the expert. Both groups performed near the performance of the expert. Table 6 shows the instructions as originally presented to workers and the modified instructions produced by workers through crowd input. The crowd-refined instructions built on the original instructions by adding an accessible analogy for the cristae in mitochondria, referring to them as a "zebra pattern" seen in Table 6 next to the instructions seeded to workers.

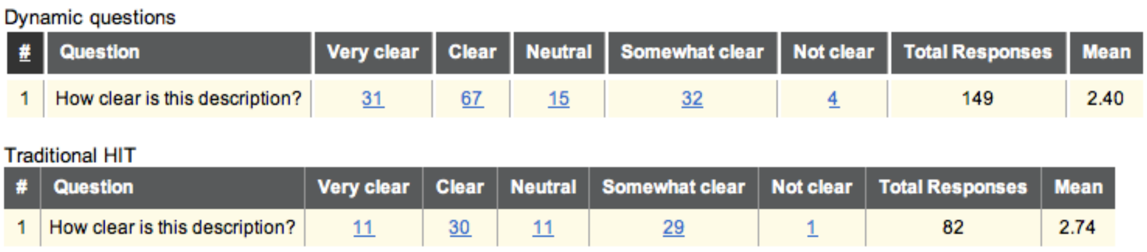
A key attribute of Dynamic questions are the instructions contributed by workers. The top result in Table 6 illustrates the originally seeded instructions, "Mitochondria are dark objects with rubs that cut across them", and the user contributed instructions, "Mitochondria come in shapes that are oblong or circular. They have a thick border with ribs that cut across them, like a zebra pattern." Additionally, Turkers on average thought the Turker contributed questions were clearer, as seen in Table 7 ($\bar{x}_{Control} = 2.74;\ \bar{x}_{Dynamic} = 2.40; t(229) = 1.12; p - value = 0.264$).

	Regular HIT	Dynamic Questions HIT	Expert
Mitochondria		20.5	21

Table 5: Crowd response to the HIT with and without the use of Dynamic questions.

Experiment 1: Find the number of mitochondria	Mitochondria are dark objects with ribs that cut across them.	Mitochondria come in shapes that are oblong or circular. They have a thick border with ribs that cut across them, like a zebra pattern.

Table 6: Task instructions seeded into the system and the result generated by crowd workers.



Dynamic questions

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear is this description?	31	67	15	32	4	149	2.40

Traditional HIT

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear is this description?	11	30	11	29	1	82	2.74

Table 7: Comparison of perceived clarity of the instructions. Very clear is 1 while not clear is 5. Turkers on average thought the Dynamic questions were clearer.

3.4 Experiment 2: Counting whole cells

The first experiment demonstrated a performance difference between the experimental and control groups, possibly because the task was too easy (i.e., as a result of the example image). Therefore, a second experiment was conducted where workers were asked to find the number of whole cells in an image, but this time without providing an example image, as was done in the first experiment.

3.4.1 Participants

Participants were recruited from Amazon's Mechanical Turk with several worker conditions. These conditions included a requirement that workers had completed 1000 prior hits with a 95% success rate. Workers were given the option of previewing the HIT before accepting it and were provided with a compensation of $0.25. For the trial, a total of 100 participants were recruited.

3.4.2 Materials

Workers were not given an example image in this experiment and were asked to count the number of whole cells. The electron micrograph shown in Figure 17 also included a number of other features such as nucleoli, blood vessels, and other structures.

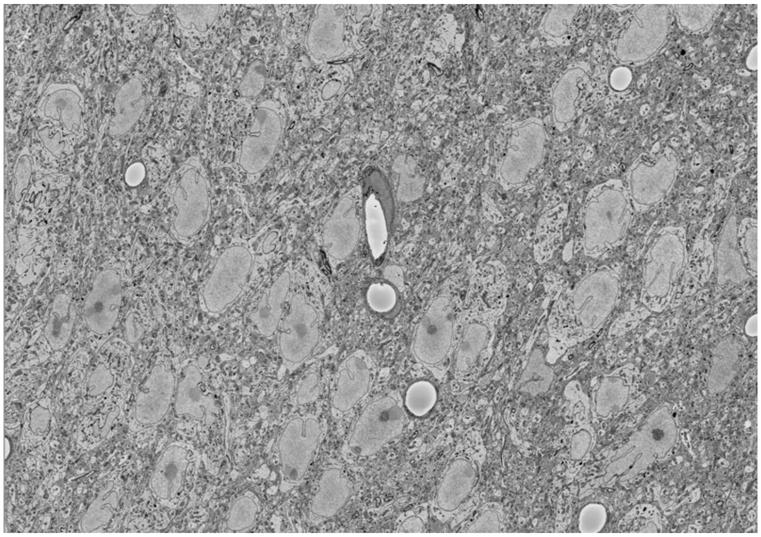


Figure 19: Workers were asked to find the number of whole cells in this electron micrograph.

3.4.3 Design

The experiments followed a two-group experimental design. The control group had a static set of questions while the experimental group performed tasks dynamically generated at runtime. The data collected included vote information, satisfaction data, and user contributed instruction sets from Qualtrics, Amazon's Mechanical Turk, and Google's Fusion Tables.

3.4.4 Procedure

The procedure is identical to the procedure used in the pilot test described in detail in 3.2.4.

3.4.5 Results

In this task, both experimental and control groups performed poorly, not matching the results of the expert. Table 8 shows the originally seeded task instructions and the set of instructions workers voted as the best set of instructions where each Turker has the ability to create or modify their own set of task instructions. This user contributed set of instructions is seen in Table 9. They added the analogy of the cell membrane being a "skin" and reiterated the intent of the task "Thus, you are looking for how many individual cells you are seeing in this picture." It is interesting to note that workers voted for these tasks despite minor grammatical errors suggesting precision of the formulation of the task is less important than accessibility and ability to clearly communicate the intent of the task.

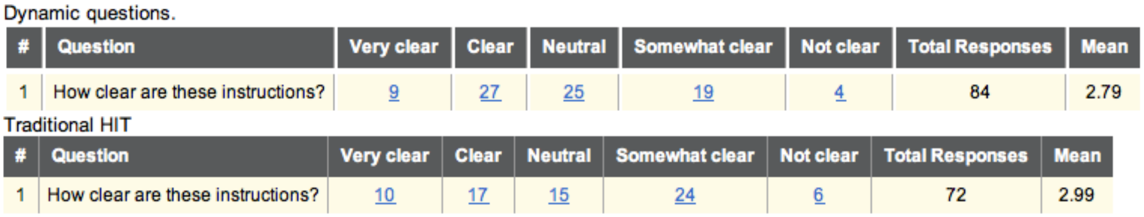
Like the previous experiment, workers in the experimental condition rated the clarity of the task instructions higher than the control group seen in Table 10.

	Regular HIT	Dynamic Questions HIT	Expert
Whole cells	76.2	87.5	70

Table 8: Table showing responses from the crowd asking to count the number of whole cells in the image. The control group performed better. Both groups included extreme answers of greater than 1000.

Experiment 2: Find the number of whole cells	Cells are structures defined by their membranes. You can find more information here: http://en.wikipedia.org/wiki/Cell_(biology)	Cells are defined by their membranes. Membranes are the outter covering, or "skin" on the cell. " For more information on cells please go here: http://en.wikipedia.org/wiki/Cell_(biology)

Table 9: Table of task instructions seeded to Mechanical Turk and the task instructions generated by the crowd.



Dynamic questions.

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear are these instructions?	9	27	25	19	4	84	2.79

Traditional HIT

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear are these instructions?	10	17	15	24	6	72	2.99

Table 10: Comparison of perceived clarity of the instructions. Very clear is 1 while not clear is 5. Turkers on average thought the Dynamic questions were clearer.

3.5 Experiment 3: Finding Nucleoli

The prior experiment failed to yield the expected difference in performance. To further characterize the implications of Dynamic questions, workers were asked to count the number of nuclei in an image. In this task, workers were again not provided with an image or detailed description.

3.5.1 Participants

Participants were recruited from Amazon's Mechanical Turk with several worker conditions. These conditions included a requirement that workers had completed 1000 prior hits with a 95% success rate. Workers were given the option of previewing the HIT before accepting it and were provided with a compensation of $0.25. For the trial, a total of 100 participants were recruited.

3.5.2 Materials

This experiment used the same electron micrograph used in the previous experiment, this time asking workers to count the number of nucleoli as seen in Figure 18.

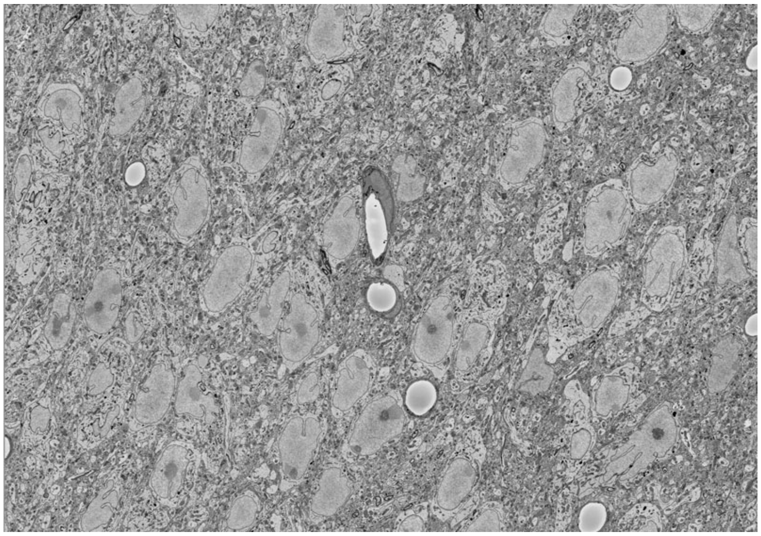


Figure 20: Workers were tasked with finding the number of nucleoli in this electron micrograph.

3.5.3 Design

The experiments followed a two-group experimental design. The control group had a static set of questions while the experimental group performed tasks dynamically generated at runtime. The data collected included vote information, satisfaction data, and user contributed instruction sets from Qualtrics, Amazon's Mechanical Turk, and Google's Fusion Tables.

3.5.4 Procedure

The procedure is identical to the procedure used in the pilot test described in detail in 3.2.4.

3.5.5 Results

In this experiment the community continues to demonstrate their ability to re-cast the instructions into alternative representations. In this example, workers added that each nucleus would include a nucleoli or, “a small, circular, dark mass” seen in Table 12.

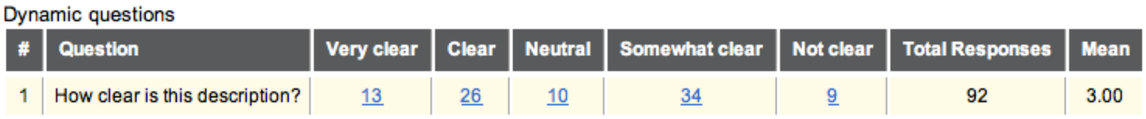
In this test, both groups performed poorly with results significantly different than the expert evaluation as reported in Table 11. The results seem to indicate slightly better performance for the control group. The difficulty of the task was reflected in the top question voted by workers seen in Table 12: “There should be an example picture with the nucleoli circled so people know exactly what they are looking for”. These instructions and the poor performance of workers seem to indicate an overly difficult task.

	Dynamic Questions HIT	Expert
Nucleoli	42.9	6

Table 11: HIT requested that users count the number of nucleoli. The responses from the crowd indicated the HIT instructions were underspecified and included several responses from the crowd to include sample images.

Experiment 3: Find the number of nucleoli	Nucleoli are the dark spots (not the light spots) enclosed by lighter areas and enclosed by cell bodies	(a) There should be an example picture with the nucleoli circled so people know exactly what they are looking for

Table 12: The task instructions seeded to Mechanical Turk and the task instructions generated by the crowd workers.



Dynamic questions

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear is this description?	13	26	10	34	9	92	3.00

Table 13: Perception of clarity by Turkers where 1 is Very clear and Not clear is 5

3.6 Experiment 4: Finding Nuclei

In yet another experiment to differentiate the performance characteristics of dynamic questions compared to traditional HITs, workers were asked to count the number of nuclei in the image. In this task, workers were again not provided with an image or detailed description.

3.6.1 Participants

Participants were recruited from Amazon’s Mechanical Turk with several worker conditions. These conditions included a requirement that workers had completed 1000 prior hits with a 95% success rate. Workers were given the option of previewing the HIT before accepting it and were provided with a compensation of $0.25. For the trial, a total of 100 participants were recruited.

3.6.2 Materials

The same electron micrograph was used for a third time, this time asking workers to identify the nucleus. The image in Figure 19 obscures the target structure with other intracellular and extracellular features.

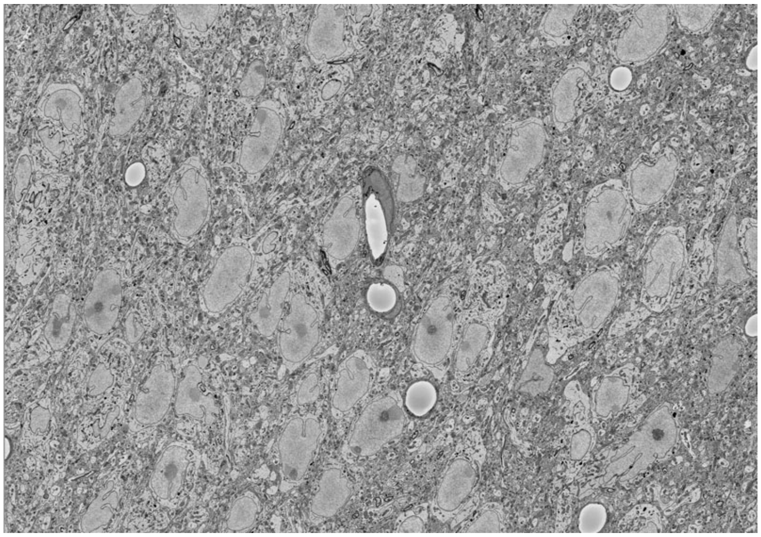


Figure 21: Preferences of crowd workers, (Rankfourth) represents the task instructions with the most votes while the other choices represent randomly selected user contributed selections from the top three quartiles

3.6.3 Design

The experiments followed a two-group experimental design. The control group had a static set of questions while the experimental group performed tasks dynamically generated at runtime. The data collected included vote information, satisfaction data, and user contributed instruction sets from Qualtrics, Amazon's Mechanical Turk, and Google's Fusion Tables.

3.6.4 Procedure

The procedure is identical to the procedure used in the pilot test described in detail in 3.2.4.

3.6.5 Results

In another experiment to differentiate the performance characteristics of dynamic questions compared to traditional HITs, workers were asked to count the number of nuclei in the image. In this task, workers were again not provided with an image or detailed description of the target structure. In terms of performance, the control group outperformed the experimental group seen in Table 14, but as an important secondary result the community continued to demonstrate their ability to re-cast the instructions into alternative representations and on average the experimental group with Dynamic questions rated their task instructions as more clear as shown in Table 15. In this example, workers added that each nucleus would include a nucleoli or, “a small, circular, dark mass”.

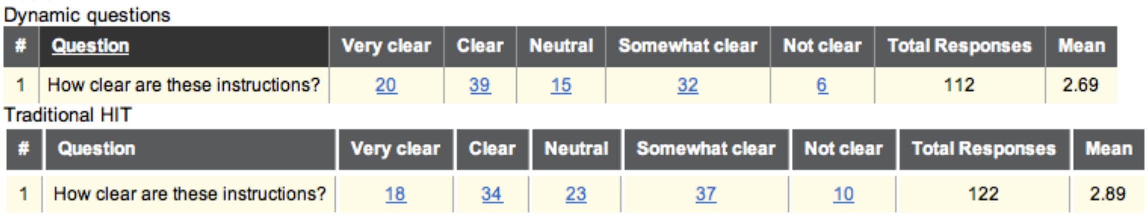
While the performance benefits of dynamic questions may be unclear, Dynamic questions has demonstrated itself to be a tool for generating alternate castings of task instructions generated by the crowd that can provide novice language and analogies that may not be apparent to the expert. On average, workers exposed to the experimental condition rated their instructions as more clear seen in Table 16.

	Regular HIT	Dynamic Questions HIT	Expert
Nuclei	34.2	27.5	59

Table 14: HIT requesting workers count the number of nuclei in the image. Both groups performed poorly.

Trial 4: Find the number of nuclei	In this task you need to count the number of cell nuclei in the image. Cell nuclei are bag-like structures that enclose most of the cell's DNA. http://en.wikipedia.org/wiki/Cell_nucleus	In this task you need to count the number of cell nuclei in the following image. Cell nuclei are bag like structures which will appear as ovular objects in the image. They will contain a small, circular, dark mass.

Table 15: Table with the task instructions seeded to Mechanical Turk and the task instructions generated by crowd workers.



Dynamic questions

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear are these instructions?	20	39	15	32	6	112	2.69

Traditional HIT

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear are these instructions?	18	34	23	37	10	122	2.89

Table 16: Comparison of perceived clarity of the instructions. Very clear is 1 while not clear is 5. Turkers on average thought the Dynamic questions were clearer.

3.6.6 Accuracy

Following the conclusion of the pilot test, a series of additional experiments were performed that better aligned with the nature of the neurosciences driving application. The first explored tasking the crowd with counting the number of mitochondria in an image. Workers were provided with a page of instructions including an image of the target structure. In this experiment, both the experimental and control groups performed well. An additional experiment was planned with the aim of distinguishing performance differences between the two groups. The next experiment increased the complexity and ambiguity of the task. The experiment did not provide an outline example of the structure, instead relied exclusively on the text task instructions. In this experiment, neither group performed particularly well, with the control group results more closely aligned with the expert evaluation. In this experiment as with the other following experiments, several results included grossly incorrect answers that claimed detection of thousands of the targeted structures. Despite the poor performance of the experimental group, workers did increase specificity of the instructions and made contributions to the formation of the task instructions. Workers also rated the clarity of the dynamic questions to be improved. To better differentiate between the performance between control and experimental groups, an additional experiment was performed with a different image that had additional visual interference. The experiment was run without providing an example of the target structure. In this case the workers voted as the most popular formulation of the instructions as request for a visual example of the target structure. The task was determined by the experimenter to be too difficult for the crowd given the conditions. In a further attempt to tease out the performance another experiment was performed with a structure that was considered to be more easily distinguished visually. Workers were asked to identify the number of nuclei in the image. Workers again performed poorly in both groups, both missing the mark set by the expert.

With the exception of the pilot test and the identification of the mitochondria, workers failed with regards to accuracy. Despite this, workers consistently preferred the dynamic questions, suggesting greater satisfaction with the task instructions formulated by other workers.

3.6.7 Throughput

In the pilot test, throughput or how quickly do workers process the task was also different between the two tasks. With the original question, the average time taken to complete the task is 208.9 seconds. In the second task with the revised question, on average it took 133.1 seconds to complete.

In the subsequent experiments, the results were less clear. Several times were measured test results returned in negative seconds reflecting technical malfunction in Qualtrics reporting mechanisms. Potential behavioral complications are reflected by very long task completion times with some participants taking more than 15 minutes to complete the task, likely indicating a pause or external activity during the HIT.

3.6.8 Cost

There are two possible cost models. The first model where the experiment is conducted and the employer reformulates the task based on feedback. The second model where surveys are dynamically built and distributed to workers based on feedback. In the first model, the cost of the experiment will increase to accommodate collecting worker feedback. In the second model, there is no direct additional cost. It could be that the less specific question will result in initially decreased productivity, but this is not clear.

Method	Cost
Expert	Amortized costs of training the expert (e.g., tuition, stipend) and ongoing costs (e.g., hourly rate, space, equipment)
Mechanical Turk large sample	$12.50/structure
Mechanical Turk panel	As low as $1.50
Mechanical Turk consensus model	As low as $0.50

Table 17: Costs associated with dynamic questions.

3.6.9 Limitations

There is clear potential for increasing the satisfaction of the worker by increasing communication between workers and the employer. There are limitations to these results including vandalism, difficulty in identifying useful contributions from the crowd, and possible collusion in some cases.

Workers in this task were told that other workers may see their responses. Browsing contributions from the Fusion Table revealed that one worker added a URL, “sdvisualimages.com” as an alternative question to be presented to workers. This response was a clear attempt to promote their website to other workers on Mechanical Turk.

In scientific applications, the results are limited to the refinement of the question and increasing satisfaction of the test. There is a preference for worker contributed questions, but this experiment was unable to convert this preference to higher performance scores.

In addition, it may be difficult to determine useful contributions contributed by workers. When presenting alternative questions to workers, they are given a list of worker contributed options. The entire table is divided into quartiles. One question is randomly chosen from each quartile and presented to workers. If there are many contributions from workers it is possible that useful contributions may take a long time to accumulate votes that indicate popularity among workers.

It is also possible that questions contributed by workers could defeat the intent of the task. For example, if workers collude and change the task so that workers do nothing and are still compensated.

3.7 Discussion

Mechanical Turk workers have discretion with regards to what work they choose to perform (Silberman, Irani et al. 2010). Mechanical Turk workers have the ability to turn to external forums that track employer traits such as clarity, difficulty of work, and payment history of employers (Irani and Silberman 2013). Poorly formed HITs, including the task instructions, have the potential to turn away potential workers and decrease the likelihood of repeat workers.

Dynamic questions was an approach developed to address the challenge of poorly formed HITs resulting from the language and mental model divide that sometimes exists between novices and experts. Dynamic questions updates task instructions based on changes submitted by workers and subsequent votes as to the most useful instructions from a list of crowd-generated possibilities.

In the pilot test using a simple math problem, there was a clear performance increase resulting from the inclusion of feedback from workers, in this case a translation of the units from Imperial units to metric units. Amazon’s Mechanical Turk employs workers from more than 100 countries, most of which use the metric system. Simple insights such as this can escape employers oriented to their home cultures and systems, and therefore modification of instructions by the crowd can make a significant difference in the performance of workers in terms of accuracy and throughput.

Encouraged by these results, several follow-up experiments explored the utility of dynamic questions in the context of interpreting biological images. In the biologically relevant tasks, the performance benefit was inconclusive. While workers preferred the task instructions provided by other workers, their performance was the same as workers in control conditions (i.e., where instructions were not modified). It may be that the biologically relevant tasks that are heavily visual and potentially ambiguous are not the best tasks for this approach.

Despite the lack of a demonstrable performance difference, dynamic questions potentially lessens the burden on employers laboring to generate a task with the appropriate vocabulary and specificity. Dynamic questions are potentially broadly applicable to a number of different Mechanical Turk tasks where the task instructions do not have to remain consistent for every worker. Additionally, a Mechanical Turk task using the dynamic questions architecture does not add additional cost to the HIT.

While the dynamic questions system shows potential to enhance communication between workers and employers, there is the potential to enhance the performance of the system. Two possible alternations to the existing system might include improved incentives for workers and improved guidance for workers.

The current system rewards workers for performing the task and engages the workers to also participate in refining the task instructions and there is no additional compensation for refining the task instructions. Within the Mechanical Turk system it is possible to reward workers with bonuses after the completion of the HIT. One objective of the dynamic questions system is to generate high-quality instructions to display to subsequent workers. It is possible to directly incentivize workers to produce high quality instruction by promising a bonus if subsequent workers vote their instruction set as the best.

In addition to providing bonuses, it is suggested that additional guidance to workers may enhance their performance. Workers in reiterating the task instructions are asked to generally improve the quality of the instructions. An example of how workers can transform non-descriptive task instructions to informative task instructions can inform workers as to possible ways that instructions can be transformed. For example in the HIT described in this chapter, workers can be provided with sample task instructions and an example high-quality transformation of those instructions. To possibly automate this process, original task instructions and worker improved suggestions from previous HITs can be used to instruct subsequent workers.

CHAPTER 4

Task Transformation: From expert tasks to pattern recognition

4.1 Experiment 3: Indirect work with task transformation

As outlined in Chapter 2, segmentation is an important analysis method in the neurosciences. This experiment looks at a different dataset and scientific objective: classification. Classification is a process of identification where investigators mark the location of target structures in the data. Classification is often used to measure the frequency of occurrence of an object. For example, frequency data are useful for measuring differences between experimental variations and control subjects, such as the frequency of mitochondria in a normal mouse compared to an obese mouse. In this case, variations between the types of mice can provide clues to the correlation between the energy making centers of the cells and fat accumulation.

In this experiment, classification is performed by transforming the original data into an alternative representation and assigning the workers to perform an analogous task using the transformed data. Data are then mapped back to the original data. In response to a scientist interested in classifying specific structure in thousands of images, this experiment looks at the detection of stigmoid bodies in electron micrographs as seen in Figure 20. Instead of asking workers to find stigmoid bodies in the original data, workers are instead asked to find objects in the transformed data that look like marbles. There are two reasons to be concerned with whether Turkers can perform meaningful work using transformed data. First, because raw data can reveal scientific intent, there is value in translating data into more abstract forms. Image data in the neurosciences can reveal the region from which data are collected and the stains used to label the structures of interest. Stains are specifically selected for their ability to bind to particular targets. These stains make some structures visible while passing over others. Second, task transformation work can be simpler. Rather than asking workers to find a biological structure, workers instead search for a target shape (e.g., “find a neuropil” vs. “find a marble-like shape”).

The overall process, as shown in Figure 21, is to train a learning algorithm on a sample structure, run the algorithm on the dataset, cut up and distribute the dataset to distributed workers as an analogous task, and then map the result back to the original data. The amount of data that needs to be analyzed by the expert scientist decreases at key points in this workflow. The task performed by experts changes from scanning the image cell by cell to reviewing automated results to reviewing the refined results by the crowd,

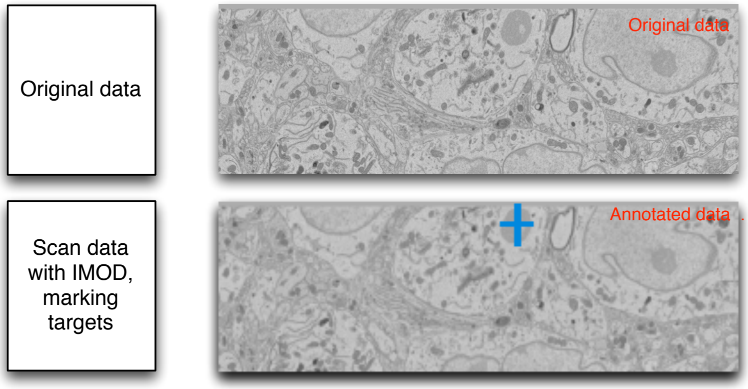


Figure 22: Un-annotated data and annotated data with the target structure marked.

reducing the number of objects to scan to less than a half-dozen objects.

This system demonstrates a methodology for blending work done by algorithms and distributed workers in the process of classification that may be useful for analysis of very large datasets. These results show that even without releasing data to the public, training of a single example yielded significant reduction of false positives, reducing the amount of work required by the expert.

4.2 Methods

4.2.1 Participants

A total of 250 participants were recruited from Amazon's Mechanical Turk with worker conditions. These conditions included a requirement that workers have completed 1000 hits with a 95% success rate. Workers were given the option of previewing the HIT before accepting it and were provided with a compensation of $0.25.

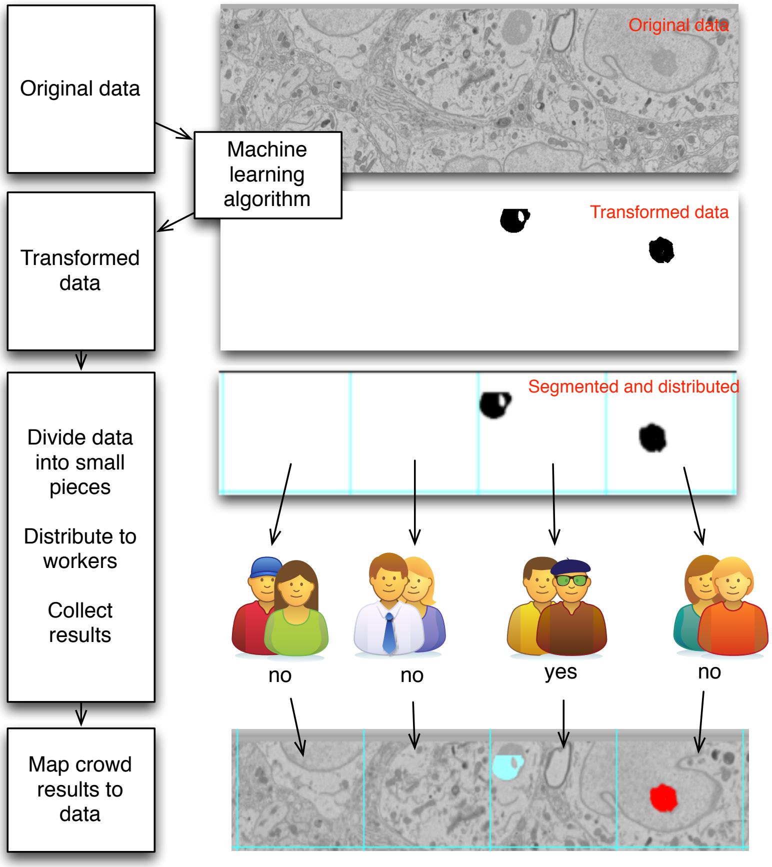

Original data
Original data
Machine learning algorithm
Transformed data
Transformed data
Segmented and distributed
Divide data into small pieces
Distribute to workers
Collect results
no
no
yes
no
Map crowd results to data

Figure 23: Process of transforming original data with automated segmentation routines, distributing to workers, and mapping the results back to the original data - marking all of the target structures.

Participants completed the task with a median duration of 2 minutes, with participants taking as long as 30 minutes.

4.2.2 Materials

The images selected for detection comprised a stack of 50 electron micrographs extending 2666 in X and 2000 in the Y dimension. Each micrograph was processed using the Cytoseg automated segmentation algorithm (Giuly, Martone et al. 2012) trained for the grayscale gradient expressed by the target structure. A stack of every 10 images from the resulting automated segmentation results were then divided into a 4x4 grid and assembled into 3D Z stacks. Each worker was presented with 16 Z stacks in the Qualtrics survey software distributed to workers with Amazon's Mechanical Turk.

4.2.3 Procedure

As depicted in Figure 21, task transformation was a multi-step process. The workflow began with a fully processed dataset from the instrument that had been aligned and stripped of artifacts. In this dataset, data were collected from the SBFSEM. With this instrument data are collected by imaging the top of a specimen block at high resolution. In an iterative process, a very fine knife then scrapes the top of the block and the underlying layer is imaged. Once the data are collected and aligned, the investigator trains the learning algorithm on a feature of interest. The resulting data are processed into smaller subsets, suitable for small displays. Distributing the work with Mechanical Turk and Qualtrics, workers identified a targeted object, which was mapped back to the original data indicating the precise location of the structure of interest.

The following describes these steps from data collection to classification in detail.

The original data in Figure 22 is a grayscale image with the darker portions of the image representing the stained objects. The target structure, a stigmoid body, is a small grey body that exists in intracellular space and looks much like a nucleolus.

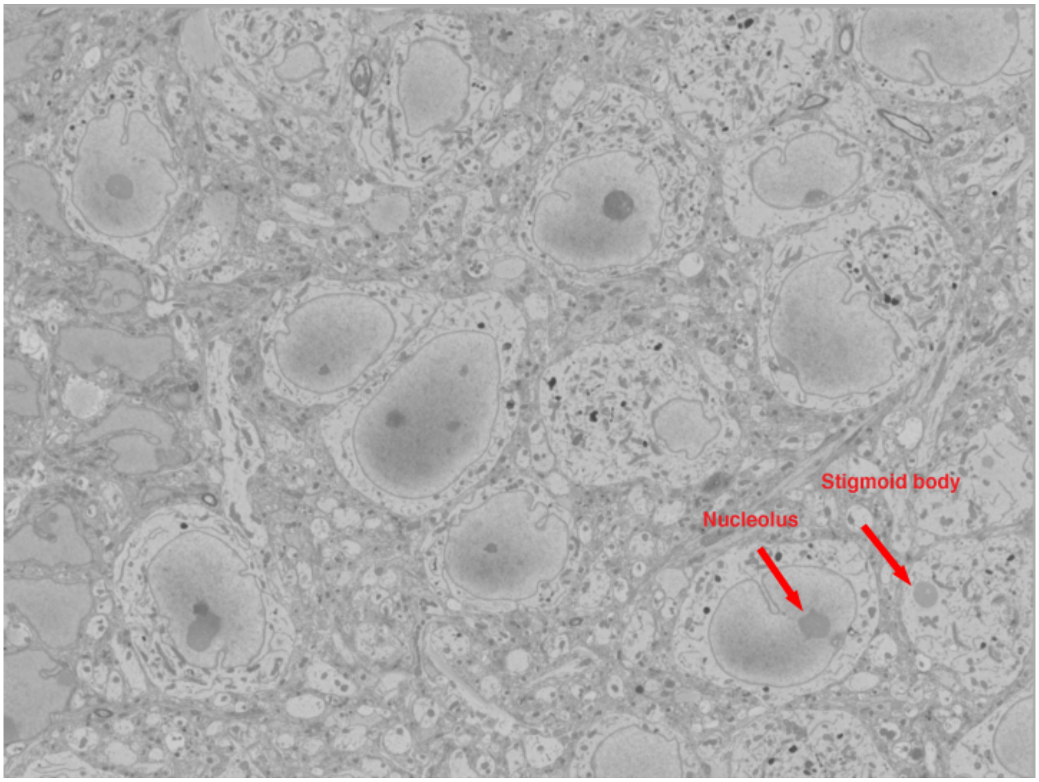


Figure 24: Original image acquired from a SBFSEM (Serial Block Face Scanning Electron Microscope).

The research scientist then trains the learning algorithm by segmenting a sample structure. IMOD (Kremer, Mastronarde et al. 1996) was used to segment a single example of a stigmoid body, this training data was then input into Cytoseg (Giuly, Martone et al. 2012). The output of the algorithm is a set of black and white images where the white regions are potential detections and the black regions are thought not to be stigmoid bodies as seen in Figure 23.

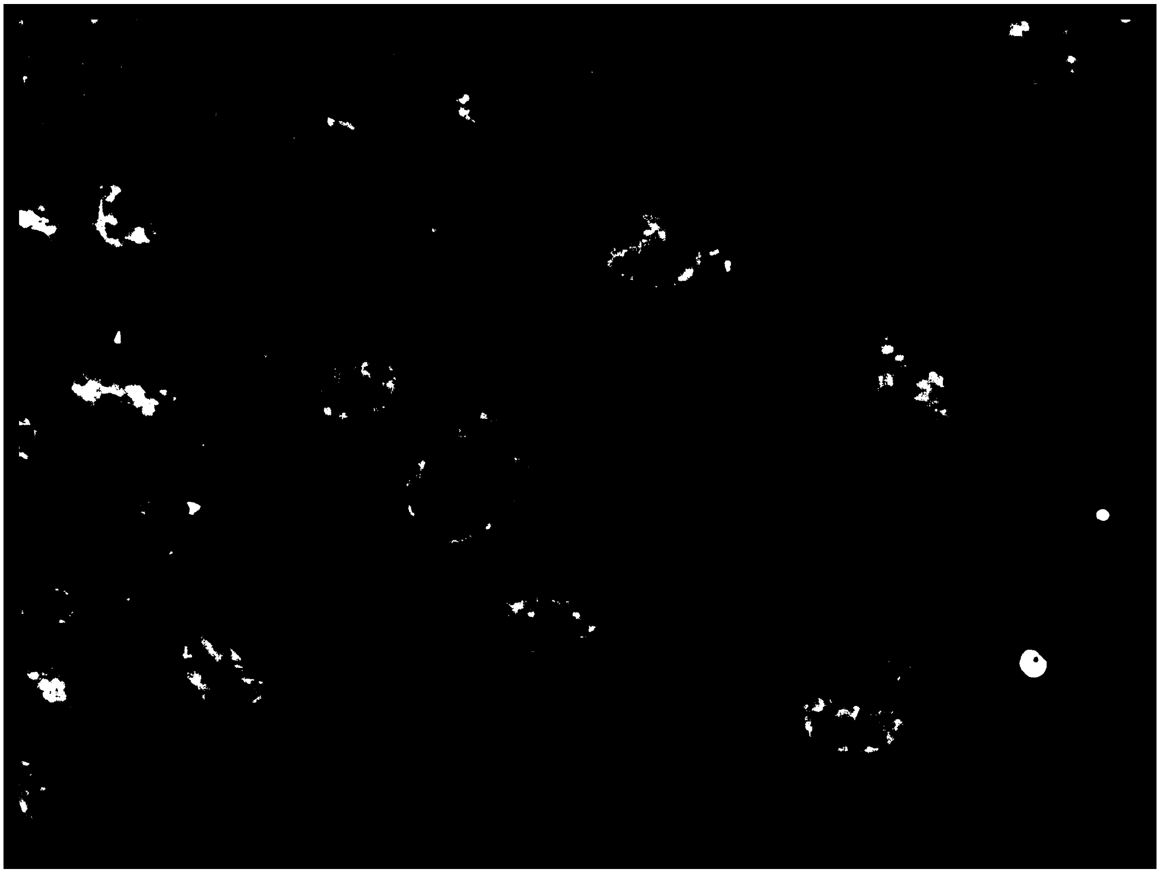


Figure 25: Threshold voxel output from the Cytoseg learning algorithm

A noise filtering step was added at the suggestion of the Cytoseg author using the application ImageJ to remove all of the objects that are too large, too small, and non-circular (Abrmmoff, Magalhaes et al. 2004). The result of this filtering process is shown in Figure 24. Figure 25 overlays the result on top of the original image. We can see in this figure two true positives and ten false positives.

Figure 26: Final mask generated by the automated algorithm Cytoseg.

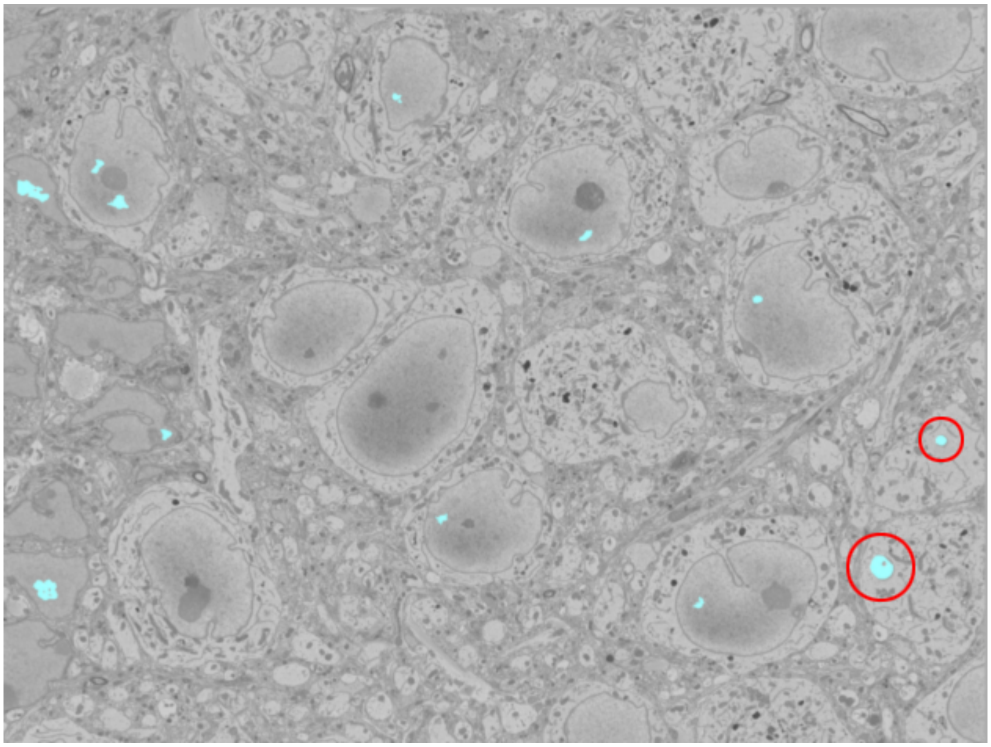


Figure 27: Original data with automated segmentation mask overlaid in light blue. The two true positives are circled in red.

Distributing the task to workers, a stack of 50 images was processed. The data were then divided into sections that can be distributed using Amazon's Mechanical Turk. Figure 26 shows the process of stacking and cutting up images to be distributed to workers. In this case the images were processed into a 4x4 grid with commands located in Appendix C. Once the images are cut up into grid pieces, they are assembled into animated stacks using the generate_animation_alt.sh found in Appendix C. The script written by the author creates a series of bash scripts that are run to create the final files.

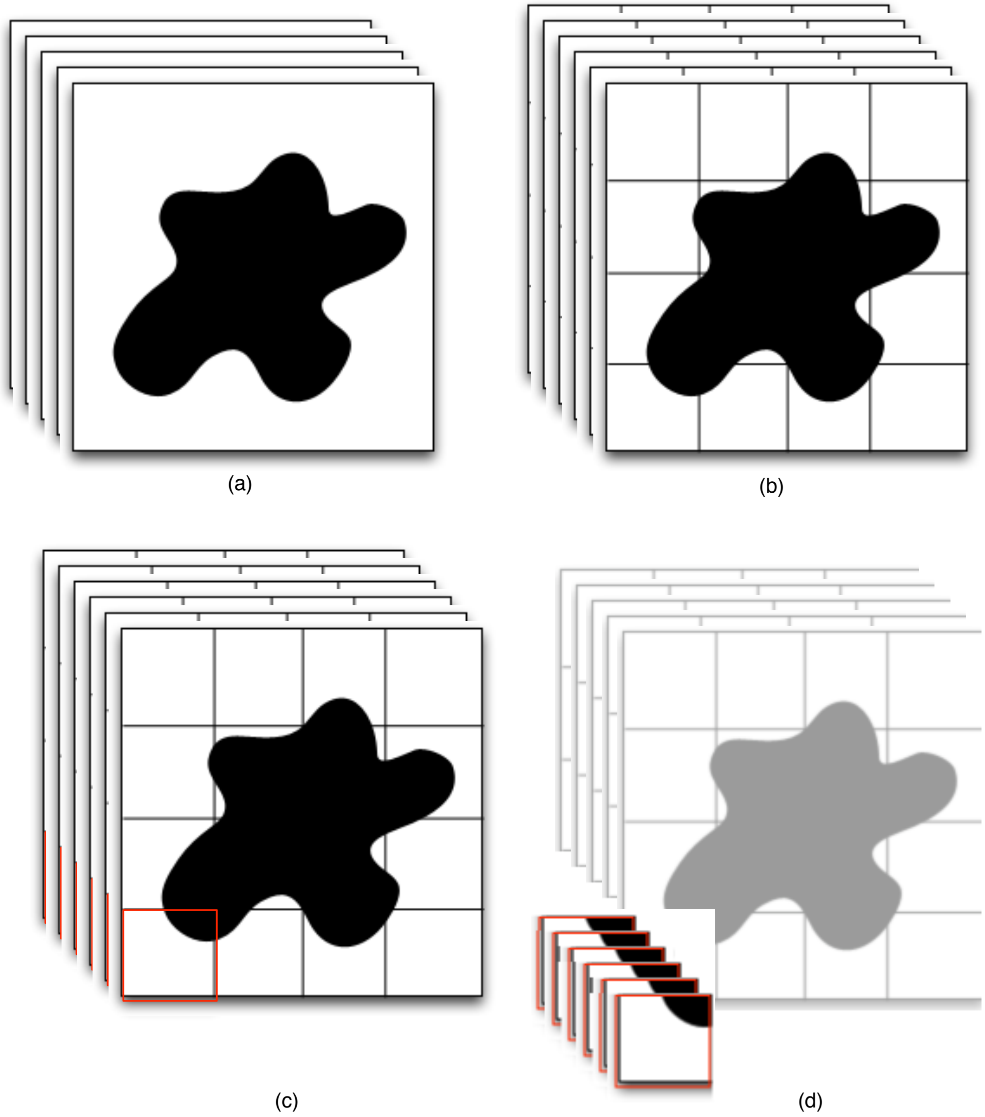


Figure 28: (a) Stack of image masks. (b) Stack of image masks in a 4x4 grid. (c) One of the sixteen stack sub-volumes in a 4x4 grid. (d) Stack sub-section cut away from the rest of the dataset assembled into a flip-book animation.

In these small flip-books, the underlying geometry of the target structure persist through multiple slices in Z whereas errors and some other false positives only persist through single slices or do not follow a predictable structure. In Figure 27, you can see a simple representation of what a stigmoid body may look like through multiple slices. Figure 28 is a small flip-book animation that demonstrates a series of Z images that sections through the target structure. You can see noise in image 5 that does not persist in in the Z dimension.

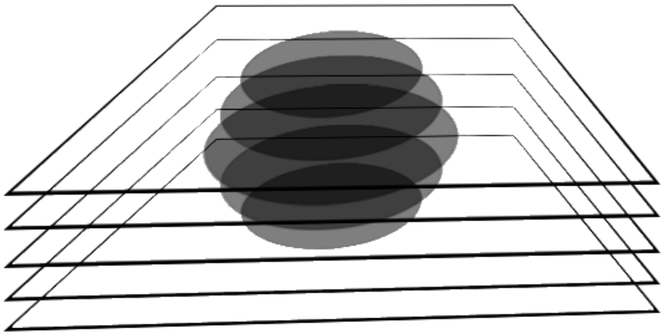


Figure 29: Five layers show in in perspective with each layer showing a circular body. The aggregate of these in 3D represents a 3D sphere.

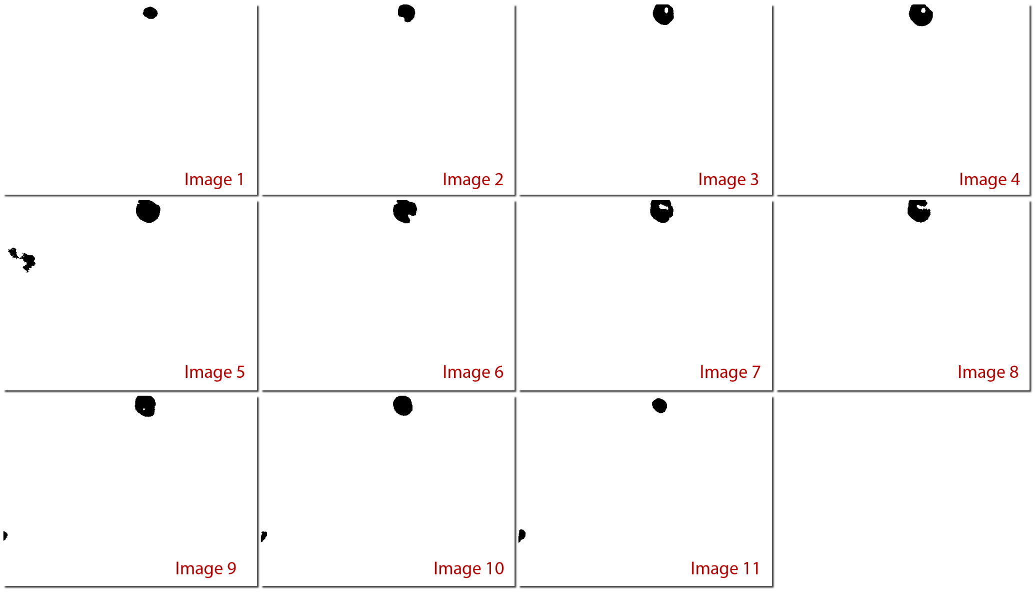
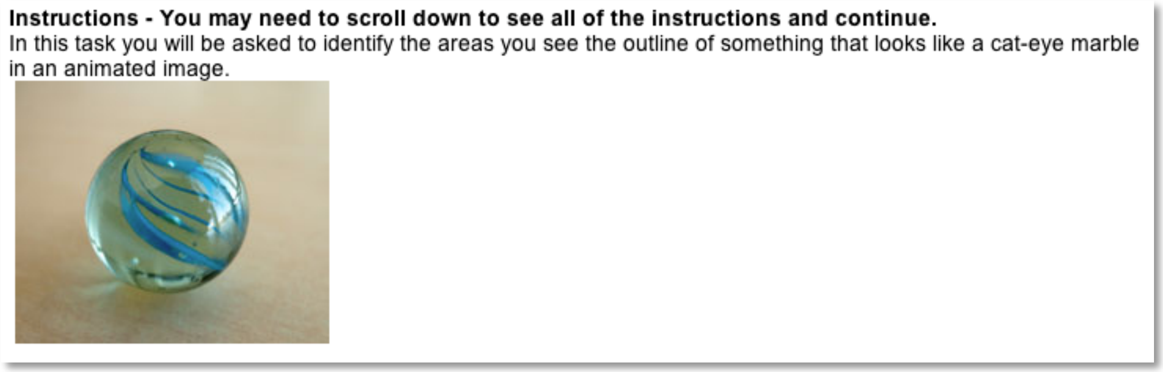


Figure 30: A sample dataset displayed to workers. Each image is one layer in an animated flip-book. Workers are tasked with finding the marble-like sphere in the data.

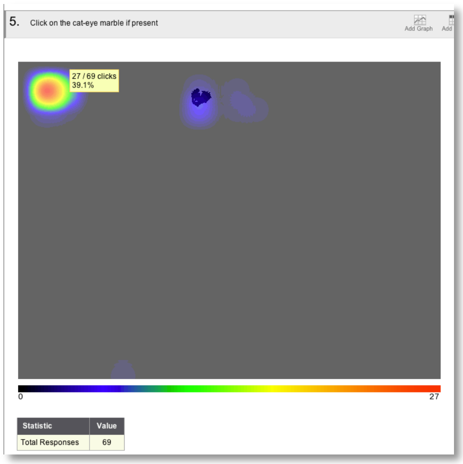
Once the flip-book animations are created, the images are distributed to Mechanical Turk where Turkers were asked to find structures that resembled a cat-eye marble as shown in Figure 29. The cat-eye marble closely resembles the stigmoid body and doesn’t reveal any scientific objectives. These instructions presented to workers made no reference to the original data or to the true underlying target structure.



Instructions - You may need to scroll down to see all of the instructions and continue.
In this task you will be asked to identify the areas you see the outline of something that looks like a cat-eye marble in an animated image.

Figure 31: Instructions presented to Mechanical Turk workers.

Each dataset was distributed to fifty workers, and in the task they were asked to click on the “cat-eye marble.” These clicks were recorded by the Qualtrics heat-map question type and at the end of the task, visualized as seen in Figure 30. The Qualtrics heat map showed points of consensus and how many workers clicked on that region. Once all of the click data were compiled, as shown in Figure 31, the consensus of the workers was mapped back to the original electron micrograph as seen in Figure 22.



5. Click on the cat-eye marble if present

Statistic	Value
Total Responses	69

Figure 32: Heat map from Qualtrics identifying the X/Y coordinates of the corresponding body.

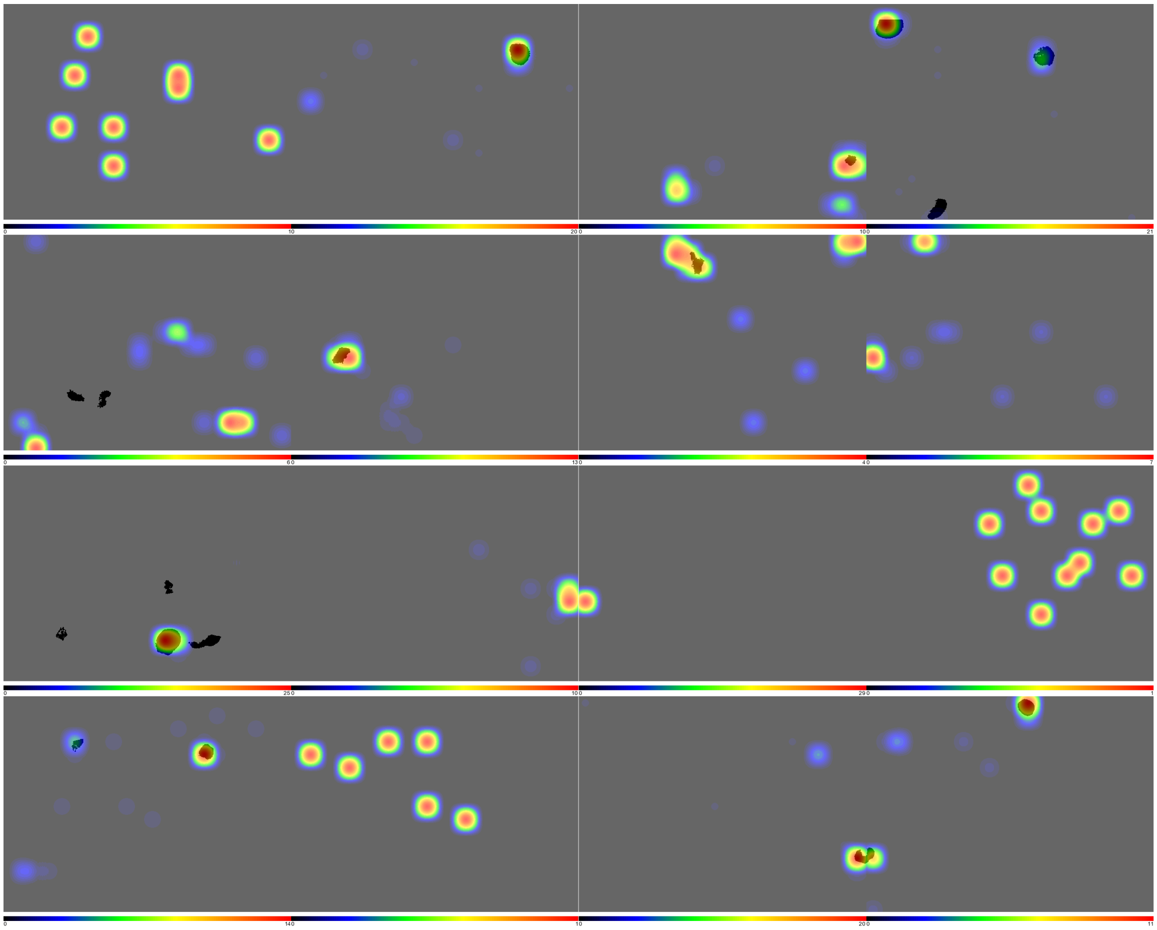


Figure 33: Heat map output from Qualtrics of all images compiled into a single image.

4.3 Results

The gold standard for accuracy of classification in the neurosciences is manual tracking by an expert. In manual classification, workers examine each image, marking the occurrence of each structure. Classification, like segmentation, is a labor-intensive process that varies with the complexity of and number of structures, quality of the data, and the experience of the scientist.

In this experiment 50 images were divided into 5 datasets with the dimensions of 2666×2000x11 pixels, one image overlapping the previous dataset. Workers click on the "cat-eye marble" generating points collected by Qualtrics. These clicks are used to generate heat maps indicating where workers click and how many click in the same region. The resulting heat maps are given threshold points to mark consensus by the workers, in this case agreement of 20 or more participants in the same region (40%) marked a point of consensus. These points were manually correlated with structures in the dataset using print-outs and highlighters to track true and false positives resulting in images such as Figure 31.

4.3.1 Cost

The cost of analysis in image processing is mostly in the human cost of the expert. It is estimated by an expert that it would require several minutes to manually process this dataset finding all instances of the target structure and marking them. This manual process of annotating the image is the gold standard for accuracy.

Automated algorithms have the potential to reduce the burden of experts by reducing the amount of work that needs to be done by the expert. Early interviews indicate that depending on the accuracy of the automated algorithm, particularly false positives, it is easier to manually process the data. With learning algorithms, the expert must train the algorithm with examples of the target structure. Workers must also then go through the dataset to correct mistakes made by the algorithm.

Crowdsourcing the results as with task transformation adds a layer of complexity. With crowdsourcing the data requires additional processing to transform the data into a form that can be distributed and then must be analyzed to collect the intelligence and map the results back to the data. Additionally, systems such as Mechanical Turk and Qualtrics need to be configured. Lastly, Mechanical Turk workers must be compensated for their time. It cost $68.75 for Mechanical Turk workers to process 50 images.

Method	Cost
Expert	Amortized costs of training the expert (e.g., tuition, stipend) and ongoing costs (e.g., hourly rate, space, equipment)
Mechanical Turk large sample for 50 images	$68.75
Mechanical Turk panel per image	As low as $1.50
Mechanical Turk consensus model per image	As low as $0.50

Table 18: Summary of costs for Task transformation.

4.3.2 Accuracy

The gold standard of classification is manual labeling by an expert. An expert familiar with the stigmoid body was asked to find all instances in the stack, a process that required several minutes. The results of the expert were used to determine errors made by the crowd or learning algorithms. Table 15 shows the total number of objects that were detected and thought to be stigmoid bodies. The crowd was able to significantly reduce the number of false positives shown in Table 16. Figure 32 shows the results of the learning algorithm in light blue, false positives from the crowd in red, and true positives identified by the expert in green.

	Expert (number errors)	Crowd (number of errors)	Learning algorithm (number of errors)
Subvolume 1	0	1	4
Subvolume 2	0	2	14
Subvolume 3	0	1	8
Subvolume 4	0	3	6
Subvolume 5	0	1	12

Table 19: The number of errors made by the expert, crowd, and learning algorithms.

	Expert (number of objects detected)	Crowd (number of objects detected)	Learning algorithm (number of objects detected)
Subvolume 1	3	4	7
Subvolume 2	3	5	17
Subvolume 3	2	3	10
Subvolume 4	4	7	10
Subvolume 5	2	3	14

Table 20: The number of objects detected by the expert, crowd, and learning algorithm.

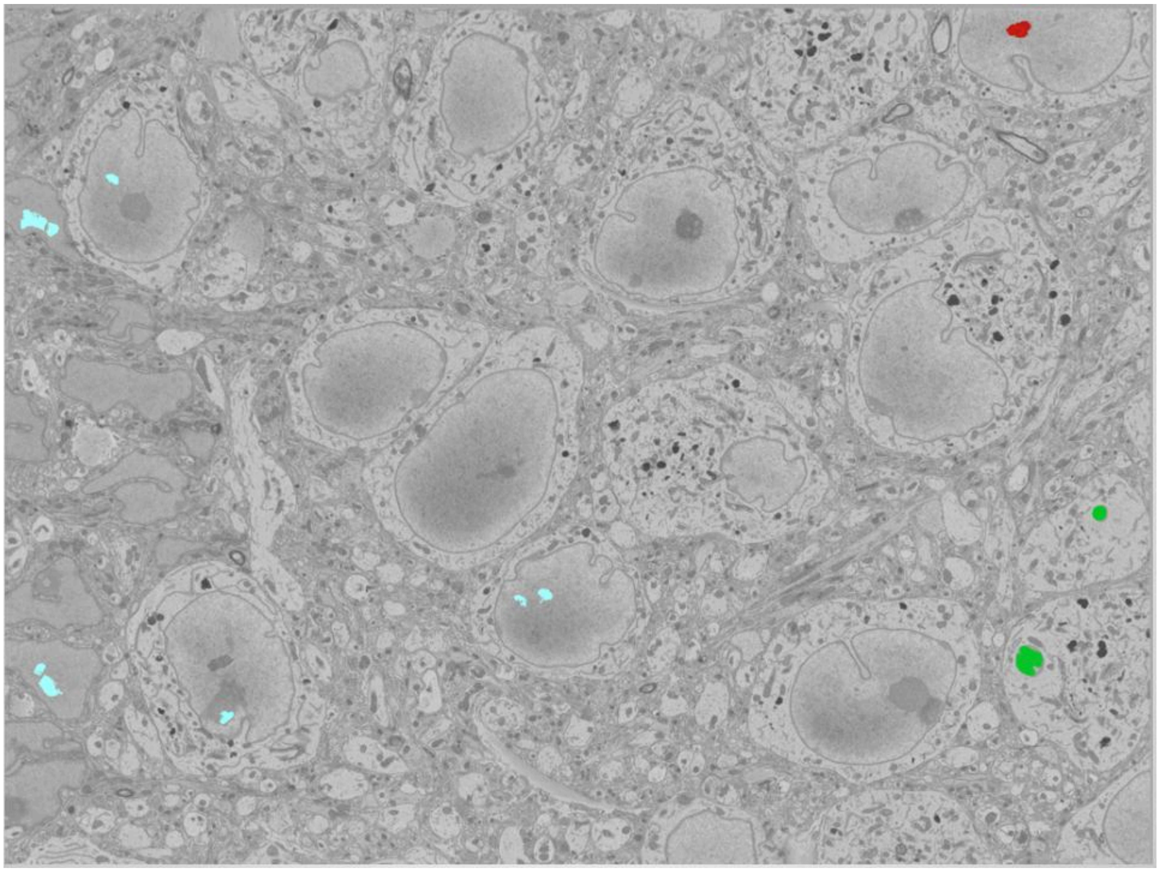
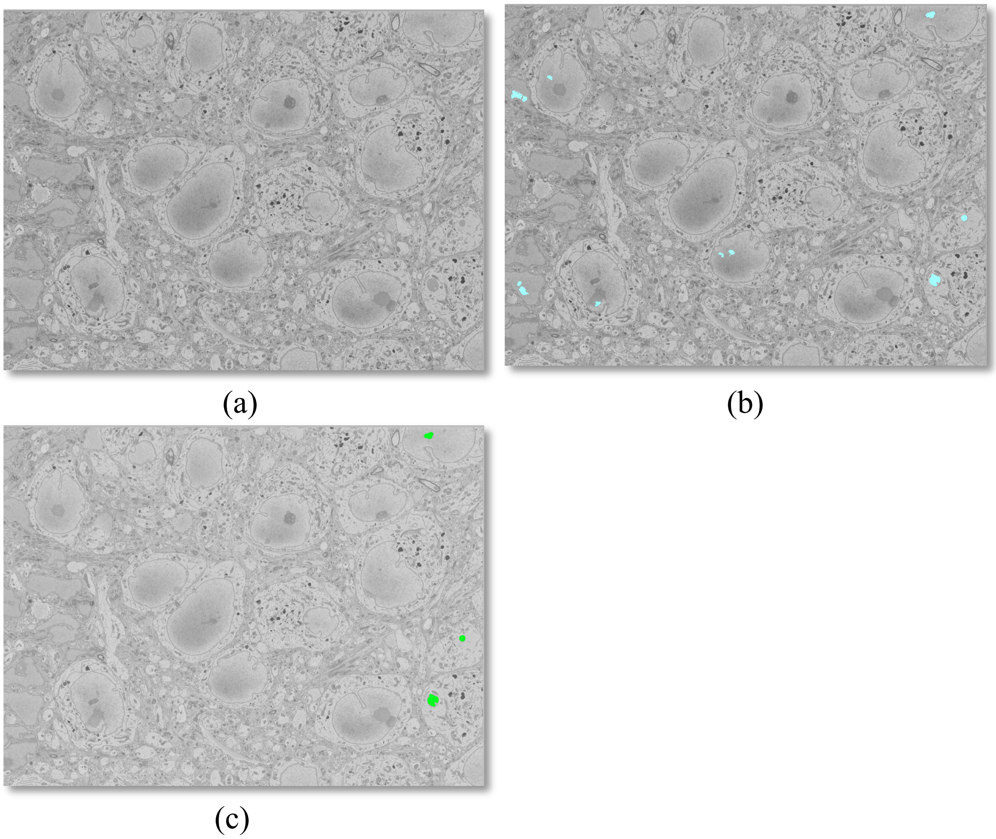


Figure 34: Original image from subvolume 3 with automated images in light blue. The false positives from the crowd in red and true positives in green.

4.3.3 Throughput

The results of the combination of the automated algorithm and Mechanical Turkers didn't produce perfect classification. However, it did result in a significant reduction of data that needed to be evaluated by experts during the process of classification. Figure 33 demonstrates the three variations of the information required to process manually, with automated segmentation, and with task transformation.



(a) (b) (c)

Figure 35: (a) Unannotated electron micrograph where experts look at all objects in the image to find target structures. (b) Machine learning annotated image where regions in blue mark areas where the algorithm detected the target structure. (c) The results of task transformation marked with regions in green representing identification of the target structures.

In the process of classification, experts go through each image and mark the centroid of each instance of the desired structure. With the automated results overlaid on top of the electron micrograph, experts have significantly decreased search space, visually targeting only the annotated regions. Overlaying the automated algorithm results decreases search space from every cell in the image to about a dozen points. In comparison, the crowd-sourced task further reduced the search space in the example shown in Figure 32 from eleven points to three where one was a false positive. There were no false negatives in the five sub-volumes analyzed. Workers significantly reduce the number of false positives generated by the learning algorithm, resulting in annotations with few false positives and no false negatives.

4.3.4 Limitations

Despite the advantages in reducing information that the investigator needs to interact with, the process of task transformation has limitations. Task transformation in this example utilizes the naturally occurring geometry of the target structure, a simple sphere. Many structures of interest are much more complex in shape and may not be as conducive for transformation. There are also a number of steps that are done manually and could negate the performance gains by experts, including the generation of instructions for Mechanical Turk workers, analysis of data, and mapping the results back to the original data.

Additionally false positives are an area for concern. While there were no false positives in this example, if the automated algorithm does not detect the structure, workers will never be given the opportunity to find the structure. Learning algorithms may not detect unexpected mutations of the targeted structure. It is these unexpected results that are often the most scientifically interesting.

4.4 Discussion

There are other examples of data reduction of large volumes of image data. Finding Khan reduced the number of possible sites from an intractable number of structures to investigate to a number of sites that could be prioritized and accomplished given limited resources for exploration (Lin 2010).

The scale of the problems in the neurosciences is no less than in applications like Finding Khan. The increases in efficiency in both instrumentation and methods for data collection are creating vast expanses of tissue that need to be analyzed, more than individual investigators can do themselves. Task transformation in this example steps in that direction, and does so while addressing a key practical problem of obscuring the underlying and confidential data.

Distributing work to the crowd again multiplies the effort of the expert and solicits the effort of the crowd to perform tasks often less desirable to the expert worker, but necessary for publishable results. In this example, reducing the number of false positives by a factor of 5.5.

Existing methods for extracting data with semi-automated methods are increasing throughput, but the time consuming and repetitive nature of the work creates barriers to participation greater than the effort required to perform them.

Production of this example required building the infrastructure and assembling a suite of infrastructures for content delivery, image processing by automated algorithms, image processing to send to workers, content delivery network, answer preparation and etc.

The technologies assembled in this experiment along with the new code in Appendix C were applied to a small set of data. Experiments suggest that the time required for crowd work doesn't increase with the number of jobs, but workers act in parallel.

There are limited practical applications for task transformation in the neurosciences; the overhead and possibility for false positives narrows utility beyond data reduction. Still, the results are promising. For example, this approach begins to address the tension between recruiting a large and diverse workforce while retaining some data security. Workers are never presented with the raw data and it would be difficult to map the task data back to the original data. Future directions for this approach could include use for estimating the occurrence of objects in data. this approach may result in faster tuning of learning algorithms. Learning algorithms often require months of developer support to tune the algorithm to a specific structure or even dataset. Crowdsourcing may be a way to cheaply identify where algorithms require additional tuning.

CHAPTER 5

Theoretical Contributions and Conclusions

5.1 Contributions

The techniques described in this dissertation demonstrate the potential for applying crowdsourcing to expert work in the sciences. Current methods for data analysis lack the throughput for processing large datasets made possible by modern instrumentation. Existing crowdsourcing methods are not designed for applications that require expert knowledge or require privacy of underlying data. This dissertation demonstrates a system that can be applied to crowdsourcing the analysis of electron micrographs. While these techniques are not intended to replace the expert, they can multiply the effort of the expert worker, reducing the number of false positives that experts need to analyze, and use self-refining mechanisms to potentially guide the expert in the development of more effective crowdsourcing applications. The following describe the contributions of each individual experiment, limitations, and future directions.

5.1.1 Experiment 1: Distributing expertise

Modern instrumentation such as the SBFSEM produces large amounts of data, too much for an individual investigator to process completely. The method outlined in the first experiment distributes the burden of processing to a combination of automated algorithms and the crowd, leveraging the scalability of automated algorithms with the visual processing capabilities of crowd workers. In this system, experts train the algorithms that perform the initial work and crowd workers refine the results to determine the best match of the data to a given pattern. The contributions of the crowd add dynamic organizational capacity to the workplace, but without the recurring costs of employing workers. Crowd workers have cost advantages compared to traditional expert labor, they are transient and do not require continuous mentoring, employment, space accommodations, or benefits. In addition, crowd workers are numerous and can be quickly recruited, are open to small quantities of work, and require less initial investment compared to developing experts and recurring costs of expert mentoring.

This experiment provides two theoretical contributions. First, it extends the organizational literature regarding how expertise is externalized and reused within crowdsourcing. While the organizational literature recognizes machine learning as an externalized form of expertise, there is no example application to broad online communities like Amazon's Mechanical Turk where the electronic knowledge is highly portable with very low incremental costs for replication. Second, it extends the existing literature of crowdsourcing applications to include tasks that require expertise in the neurosciences for the application of segmentation. Most current crowdsourcing applications are typically simple tasks with similarly simple results. The first experiment demonstrated knowledge embedded into a system that aggregated the effort of crowd workers who were not experts, to produce outcomes that were similar to outcomes normally obtained from experts. To some degree this work also extends image processing literature to automated and semi-automated methodologies by providing a method for refining the results from automated segmentation through programmatic access of an online worker community.

5.1.2 Experiment 2: Dynamic questions

Experts and novices approach problems differently (Larkin, McDermott et al. 1980). Experts and novices differ in approach, vocabulary, and mental models when solving problems. These differences can be invisible to the expert, but can be impassable barriers to novices.

Dynamic questions bridge expert formulation of language and logical constructs presented in the task by re-formulating descriptions in the language of naïve workers. Workers progressively refine task instructions provided by an expert. In this model, results demonstrate that even crude formulations of task instructions gain specificity and remodel the language used into more accessible terms. Experiment 2 also demonstrated that workers can provide direct feedback to employers through this mechanism. For example, such as when workers provided instructions on how the employer could enhance the task instructions by including example images.

This work contributes to the literature of crowdsourcing. It re-imagined the HIT in mTurk work as an object that is dynamic, capable of integrating feedback and contributions from workers, and that opens limited two-way communication between employer and workers.

5.1.3 Experiment 3: Task transformation

Distributing unpublished work to public workforces can inadvertently reveal unpublished hypothesis (Wright 2010). In crowdsourcing data, there is a tension between enlisting large communities to participate with the need to keep unpublished data confidential.

In the neurosciences, electron micrographs reflect several steps of the scientific process including region where the tissue is collected and stains used to reveal specific structures. Correlating the region where tissue was collected and the stains used may reveal an investigator's hypothesis or interests. To reduce the chance of revealing confidential information, this experiment obscures the data, revealing only information required to accomplish the task and presenting the task in an alternate framing.

The third experiment continues to demonstrate the potential of pairing automated algorithms with crowd workers. It builds on the work of the first experiment by also concealing the underlying data to prevent the unintended distribution of information revealing of an employers' scientific methods, tools, or hypotheses. Results demonstrated the ability of crowd workers to filter noise and reduce the number of false positives detected by automated algorithms.

This work contributes to the literature of crowdsourcing scientific work by demonstrating a technique for engaging crowd workers without exposing sensitive underlying data. This was accomplished by transforming data into another form, asking workers to complete an analogous task on the transformed data, and mapping the results back to the original data.

5.2 Limitations and Generalizability

With the broad goal of applying crowdsourcing to scientific applications, this dissertation focused its scope using a driving application, specifically image segmentation in the neurosciences. As a result of this driving application, the development of the infrastructure is specific to the requirements and challenges of image segmentation. These limitations to generalization are specific to each experiment and detailed in the following sections.

5.2.1 Experiment 1: Distributing expertise

Experiment 1 is directed specifically to membrane detection of cells in electron micrographs. The experiment implements existing automatic segmentation algorithms that are not explicitly designed to generate a panel of alternate segmentations such as those presented in Experiment 1. Future algorithms can be architected to provide purposefully unique segmentations for feedback from the crowd and feedback the responses of the crowd into the training of the algorithm.

Within the neurosciences there are other types of neuropil commonly segmented such as mitochondria, nucleoli, and endoplasmic reticulum. The methods presented here could be applied to those structures as well but may be more difficult to apply if there isn't a clear shape geometry the crowd workers can be tasked with recognizing. For example, the membrane or cell walls may be easier to describe than the filament like structure of endoplasmic reticulum.

Application to disciplines other than the neurosciences may be possible if the judgments made by crowd workers can be broken down into a panel of possible choices. One such application may be possible in the atmospheric sciences where crowd workers are asked to provide local expertise and intuition to an ensemble prediction of a hurricane path. Workers familiar with the terrain the hurricane is expected to travel could provide intuitive or informed decisions as to the path of the storm.

5.2.2 Experiment 2: Dynamic questions

The dynamic questions system demonstrated in Experiment 2 is possibly the most generalizable of all the experiments. The infrastructure leverages several scalable cloud systems such as Amazon's EC2, Google's Fusion Tables, and Qualtrics. Each of these systems are capable of serving many simultaneous users. Intended to bridge experts with novices, dynamic questions can be applied to the refinement of Mechanical Turk tasks instructions where experts may have trouble communicating with novices.

Despite best efforts, dynamic questions did not demonstrate improved performance of workers when applied to applications in the neurosciences, but did show performance increases in a pilot test. It is possible that these methods can increase performance in other tasks that are less qualitative or with less uncertainty.

5.2.3 Experiment 3: Task transformation

Similar to Experiment 1, task transformation combines automated algorithms with the crowd. Within the neurosciences, these same methods can be applied to other structures that have a contiguous 3D shape or pattern.

These methods can possibly be applied to other domains such as the earth sciences, for example, one could imagine recasting the Finding Khan project with a combination of automated learning algorithms that can produce an image mask of possible man-made structures. Crowd workers would then annotate data marking possible ancient architectural structures that do not conform to naturally occurring structures. It may be that workers can distinguish between man-made objects and naturally occurring objects in vast expanses of land, reducing the number of possible sites that the team targets their investigation.

5.3 Costs of analyzing data

Many examples of very successful crowdsourcing efforts do not compensate their workers, but rather rely on the altruistic efforts of the crowd (Sullivan, Wood et al. 2009), or entertain them (Von Ahn 2006; Cooper, Khatib et al. 2010). While the driving application in this dissertation have clinical implications that may attract volunteers (e.g., by contributing to understanding of Alzheimer's disease, Parkinson's disease and etc.), building communities of workers remains a difficult task. It is unknown how many failed crowdsourcing projects there are. Separating motivation and work mechanisms, this dissertation work focuses on the mechanisms for work. Nonetheless, there may be members of the community that are less interested in building communities and would like to multiply the effort of the workers, in which case, costs are of significant concern. The following is an analysis of the associated costs in implementing crowdsourcing of image processing in the neurosciences using the methods outlined in this dissertation.

However, when calculating the costs of crowdsourcing work, the employer must examine not just the recurring costs (Amazon fees), but also initial costs. An example of these costs are broken down in the following table:

	Initial costs	Recurring costs
Expert work	Tuition, stipend, mentoring, physical space, computer resources, education materials, and software licenses	Amortized costs of tuition, stipend, mentoring, physical space, educational materials, and software licenses.
Machine learning	Hardware installation, systems administration, software maintenance, and energy costs.	Hardware maintenance, energy costs, systems administration, and facilities.
Experiment 1 : Embedding expertise	Leverage costs of expert and machine hardware in addition to setup with Qualtrics, AWS, AMT, and	None

	worker compensation.	
Experiment 2: Dynamic Questions	System setup with Qualtrics, AWS, AMT, and worker compensation.	None
Experiment 3: Task Transformation	Leverage costs of expert and machine hardware in addition to setup with Qualtrics, AWS, AMT, and worker compensation.	None

Table 21: Summary of costs for experiments in comparison to existing methods of work

To further illustrate the costs of crowdsourcing feature extraction, it is possible to estimate the costs of segmenting a full dataset such as the one used in Experiment 1. The volume from which images were selected in Experiment 1 was 700x700x269. From a semi-structured interview with an expert analyst, extracting a single structure including the cell membranes would require about 8 hours of continuous labor. Graduate students capable of this type of work such as those employed by the University of Michigan costs about $25,255.38 for a candidate graduate student for a nine-month term as of the year 2012. Given 20-23 working days per month without holidays and a nine-month term, it would cost approximately $127.55 of a graduate student candidate’s time (8 hours) to segment a cell from this volume. To extract all of the structures, similar to the work performed by the Turkers, within the volume (assuming an estimated 57 partial or whole cells within the volume), it would cost as much as $7,270.35 with an assumed rate of $15.90 per hour and $127.55 per structure. To approximately segment all of the cell membranes using the techniques outlined in Experiment 1 as it is implemented would cost $9,625.00. If the experiment were to be modified to only require two workers to agree on the same answer, costs could be as low as $350.00 for all of the membranes in the volume assuming all of the workers agreed and there were minimal set-up costs and fees.

In addition to costs, there is a question as to the availability of sufficient workers to complete the segmentation of a large volume. In the experiments conducted here with the segmentation tasks, it was common to receive all of the results within three days. It is suggested by others (Giuly, Kim et al. 2013) that it is possible to cull a group of workers familiar with your task and reliably receive results for thousands of hits per day.

While the recurring costs for crowdsourcing with systems such as Mechanical Turk are significant, they are typically less than the initial costs of acquiring new expert work or developing new machine learning algorithms. These experiments build on existing experts and algorithms, multiplying the existing investment in these resources without long term commitment to the infrastructure or people required to do the work.

5.3.1 Future Work

Looking towards a vision where scientific work is routinely accomplished in tandem between algorithms and crowd workers, these methods could be applicable to workforces that are unable to perform manual labor, such as those with disability due to back or other injuries. The methods presented in this dissertation make use of innate human ability and require little to no training. The combination of low threshold for participation and the use of innate ability extends the reach of these methods to a large community of users, and could be hosted on platforms other than Amazon’s Mechanical Turk.

To extend these techniques to a broader audience such as the one outlined above, a number of changes would need to be made including improving the ease of publishing work to AMT, compatibility with mobile devices, further development of algorithms tuned for these methods, and optimizing the costs of human participation.

The data described in this dissertation required pre-processing, or manipulation of the data into a format that could be distributed to workers. In the first experiment, the output of the algorithm was separated at three stages, superimposed with the original data, and cut into smaller pieces. Although the code are included to accomplish each of these steps, each process requires parameters to be calculated to determine the correct grid size for the given dataset. To automate these steps, it is possible to chain together applications using a scientific workflow system such as Kepler (Ludascher, Altintas et al. 2006). Encapsulating these steps into workflows will provide users with visual workflows that can be easily edited and adapted to new datasets or applications, and create an information artifact that can be shared with others interested in performing similar work.

Further extending the reach of the experiments detailed in this work, extending participation to other devices such as tablets and phones would increase the reach of Experiments 1 and 3. As the number of mobile devices capable of rendering high-resolution images increases, they become a compelling platform for distributing visual work, particularly with touch displays. The touch displays on these devices enable direct manipulation of data and may be a good platform for editing segmentations or marking features in data.

While the experiments described in this dissertation demonstrate new techniques, they are not necessarily optimized for cost and efficiency. Further work needs to be directed towards building online communities of volunteers and optimizing the costs associated with labor marketplaces such as Amazon's Mechanical Turk. The optimization of costs may include decreasing the number of decisions made by workers.

5.3.2 Extensions to other applications

Beyond the neurosciences, there are several potential extensions to the techniques described in this dissertation. For example, this work could be extended to capture the intuition of the crowd, refining advertisements, and distributing work with security in mind.

In the first experiment, distributing expertise, the emphasis of the task was to choose the best segmentation rendered by the automated algorithm from a panel of choices. In this decision workers had to interpret the data and judge the output of the algorithms. The results demonstrated the selection of answers by workers can reflect the varying degrees of consensus between workers. The expert indicated that in several instances where the expert and the crowd sourced answers differed, the answer could have gone “either way”. It is possible to use the distribution of answers by the crowd to assign probabilities to each answer where a clear consensus of the crowd indicates high degree of confidence and a split in the crowd’s answers indicates a low degree of confidence. Such a system could potentially be used to extract features from other volumetric data such as in the earth sciences.

In the second experiment, Dynamic questions, workers refined task instructions provided by the employers adding specificity and information. Applied to an application other than the neurosciences, it is possible to imagine the same technique used for feature extraction in the earth sciences. The organization could recruit the crowd to refine task instructions to include alternate wording, lay language, or even sentence structures that are familiar to the target audience.

The third experiment, Task transformation, addressed the tension between distributing work to a large and public workforce and retaining control of scientific data. The same method for obscuring the underlying data could be applied to other applications where the underlying data are sensitive, such as with national security applications. In these applications, like in the neurosciences, there are large amounts of data where the expert might want to identify sparse objects of interest such as specific facilities. The attention of security experts is limited, but if the data could be distributed in a less recognizable In this case, the effort of the crowd could focus the limited attention of high-value security experts to likely targets of interest.

5.4 Conclusions

Technology has a long history as a factor that increases efficiency, reduces effort, and enables new kinds of work. In agriculture, the plow and beast of burden were used to till fields larger than could be done by farmers alone, early computers were used to break codes in war, and calculators simplify tedious mathematical calculations. As our relationship with technology matures, we continue to pair people and technology.

As demonstrated in the sciences, there are domains of inquiry where technology alone is insufficient. This dissertation demonstrates the synergy between advanced algorithms and a micro-labor marketplace, working together to benefit from the scalability of automated algorithms and the image processing capabilities of people. This dissertation makes a step toward enabling the crowd to participate in new applications while solving important problems in the neurosciences.

APPENDICIES

APPENDIX A

Experiment 1 materials

A series of BASH shell scripts are required for processing images to be distributed to workers. There are two primary scripts. The first automatically generates a series of surveys that are loaded into Qualtrics. The second script creates These scripts were performed on a Dell workstation with 16GB of RAM and 1TB of hard drive using Ubuntu 12.04LTS.

Generating surveys to load into Qualtrics

This script generates text files that are imported into Qualtrics. The generated survey points to a set of images uploaded to Amazon's S3 distributed storage infrastructure.

```
#!/bin/bash

imagenum=265;

echo [[AdvancedFormat]]
for i in `seq 0 15`;
do
  echo [[Question:Matrix]]
  echo '${e://Field/Rank0z'
  echo '<br>'
  echo '<br>'
  echo [[Choices]]
  echo '<br>'
  echo [[Answers]]
  echo A '<img alt="microsopy image"
```

```
src="https://s3.amazonaws.com/e1data/'$imagenum'/cv8squares/'$ima
genum'_cv8_4x4_'$i'.png" />'
  echo B '<img alt="microsopy image"
src="https://s3.amazonaws.com/e1data/'$imagenum'/cv7squares/'$ima
genum'_cv7_4x4_'$i'.png" />'
  echo C '<img alt="microsopy image"
src="https://s3.amazonaws.com/e1data/'$imagenum'/cv6squares/'$ima
genum'_cv6_4x4_'$i'.png" />'
  echo [[PageBreak]]
done
```

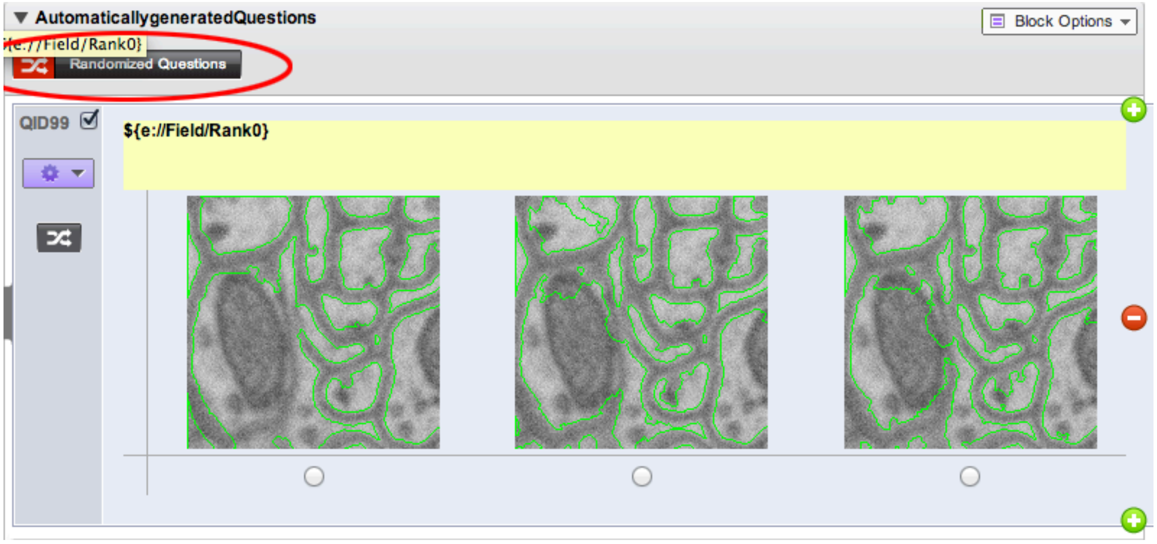
Generate Bash scripts to combine images

The following creates a series of Bash shell scripts to create a combined image of the variations generated by the automated algorithm.

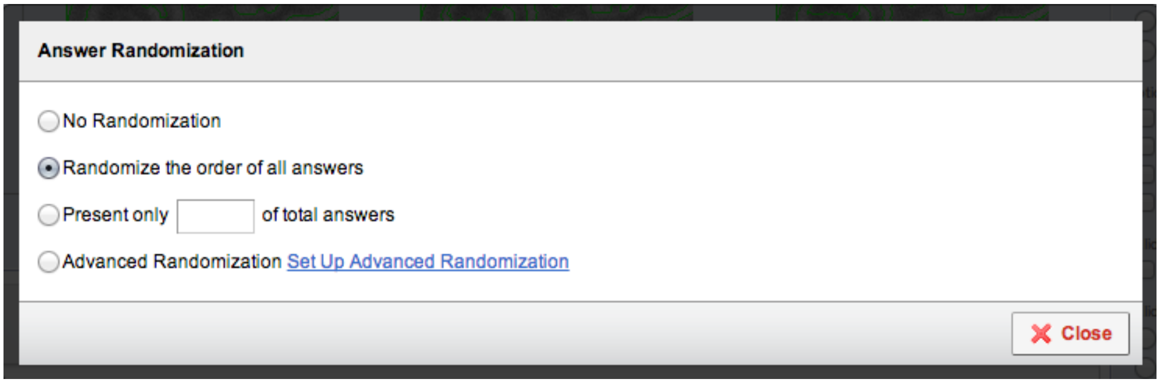
```
#!/bin/bash


echo "#!/bin/bash"
for i in 201 211 212 218 221 225 235 258 260 265; do
    echo "cd" $i;
//     echo "convert zap"$i"_cv6.tif -crop 4x4@ +repage +adjoin
cv6squares/"$i"_cv6_4x4_%d.png";
//     echo "convert zap"$i"_cv7.tif -crop 4x4@ +repage +adjoin
cv7squares/"$i"_cv7_4x4_%d.png";
//     echo "convert zap"$i"_cv8.tif -crop 4x4@ +repage +adjoin
cv8squares/"$i"_cv8_4x4_%d.png";
    echo "mkdir 4panelimages"
    for j in {0..15}; do
    echo "montage ./raw/"$i"_raw_4x4_"$j".png
./cv8squares/"$i"_cv8_4x4_"$j".png
./cv7squares/"$i"_cv7_4x4_"$j".png
./cv6squares/"$i"_cv6_4x4_"$j".png -geometry 175x175+2+2 -shadow
-tile x1 ./4panelimages/"$i"_combined_"$j".png";
    done;
    echo "cd ../";
done;
```

Experiment re-run with random presentation

Experiment1 was performed with randomization of the 16 questions of the HIT but with a fixed presentation of the three individual choices. There was some concern that workers simply picked “A” for all choices. The experiment was re-run on two images with 100 additional participants randomizing the presentation of the 16 questions in the HIT and randomizing the presentation of the three choices.



All of the 16 questions are randomized in presentation by Qualtrics



Answer Randomization

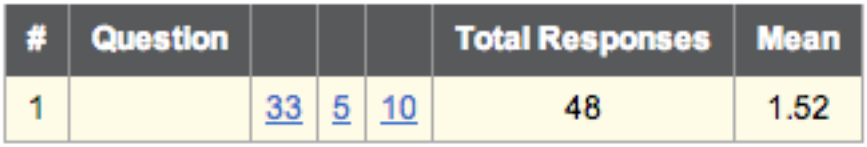
○ No Randomization

● Randomize the order of all answers

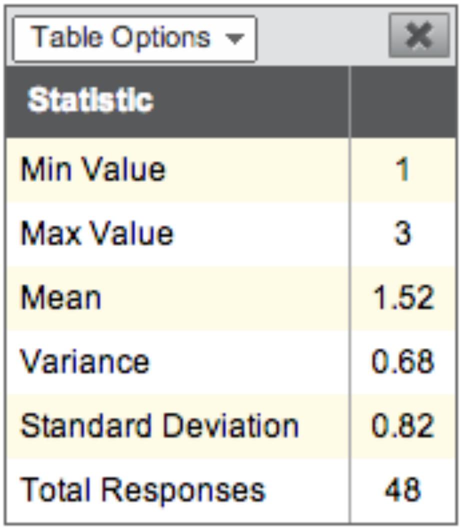
○ Present only ☐ of total answers

○ Advanced Randomization Set Up Advanced Randomization

The presentation of the answers are also randomized by Qualtrics

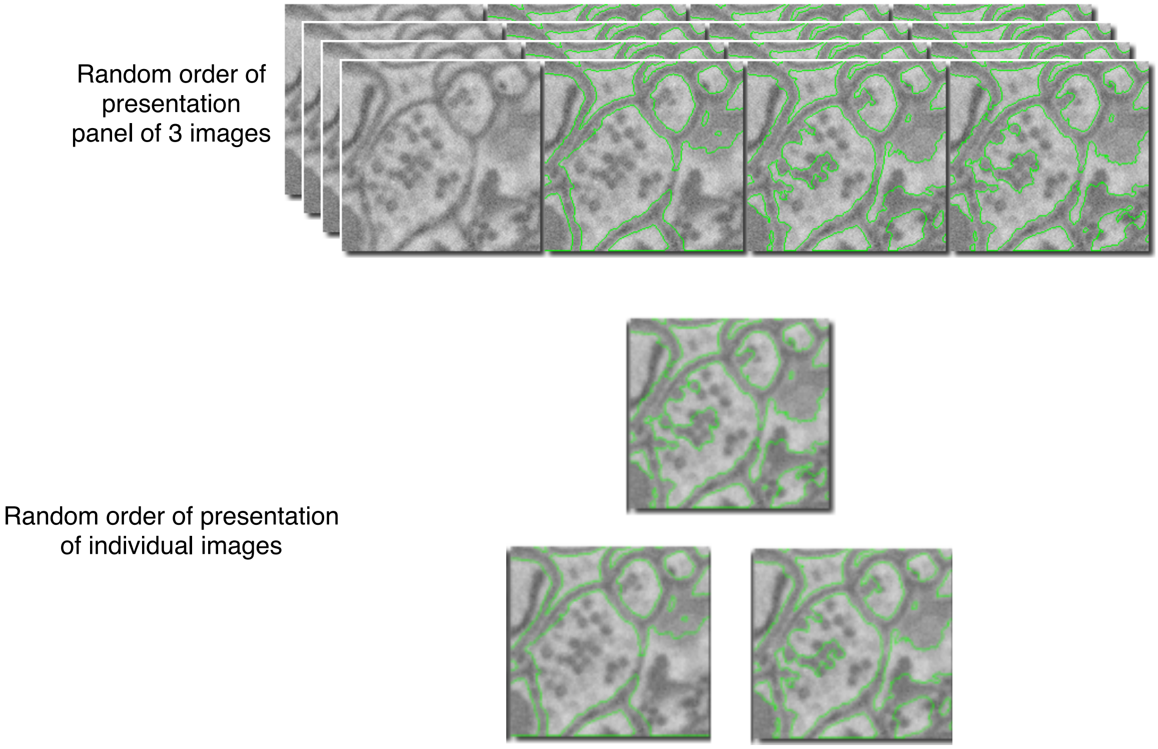


#	Question				Total Responses	Mean
1		33	5	10	48	1.52



Statistic	
Min Value	1
Max Value	3
Mean	1.52
Variance	0.68
Standard Deviation	0.82
Total Responses	48

Qualtrics maps the randomized responses back to the original placement of the data. This figure shows 33 responses map back to Image 1, 5 responses map back to Image 2, and 10 responses map back to Image 3.


Random order of presentation panel of 3 images
Random order of presentation of individual images

Experiment1 image 201
Crowd consensus in bold

	Image 1 (original)	Image 2 (original)	Image 3 (original)	Image 1 (randomized answer presentation)	Image 2 (randomized answer presentation)	Image 3 (randomized answer presentation)	Expert answer
1	**35**	6	10	**33**	5	10	Image 1
2	20	12	**21**	**23**	10	13	Image 1
3	**26**	18	8	**28**	10	9	Image 1
4	**34**	7	10	**28**	9	10	Image 1
5	**24**	12	18	**23**	13	12	Image 1
6	**32**	11	10	**31**	10	7	Image 1
7	**22**	21	9	**21**	14	12	Image 1
8	**39**	6	7	**29**	8	9	Image 1
9	**25**	19	7	**24**	10	13	Image 1
10	**30**	12	12	**26**	10	12	Image 1
11	**34**	11	8	**26**	14	7	Image 1
12	**30**	15	9	**28**	11	10	Image 1
13	**29**	13	11	**26**	14	7	Image 1
14	**32**	13	10	**33**	9	5	Image 1
15	**23**	21	9	**18**	**18**	11	Image 2
16	**30**	11	10	**30**	8	8	Image 1

Experiment1 image 212
Crowd consensus in bold

	Image 1 (original)	Image 2 (original)	Image 3 (original)	Image 1 (randomized image presentation)	Image 2 (randomized image presentation)	Image 3 (randomized image presentation)	Expert Answer
1	**26**	19	10	**19**	10	14	Image 2
2	**41**	8	6	**30**	7	8	Image 1
3	**24**	17	16	16	9	**21**	
4	23	**28**	5	**22**	13	10	Image 1
5	**23**	21	12	**20**	12	13	Image 1
6	**31**	11	15	**17**	13	13	Image 1

7	**31**	9	15	**24**	9	11	Image 1
8	**36**	14	4	**23**	11	11	Image 1
9	**32**	11	13	**24**	7	14	Image 1
10	**26**	23	5	**18**	17	9	Image 2
11	**26**	15	13	**20**	11	14	Image 1
12	**24**	**24**	8	**19**	12	13	Image 1
13	17	**25**	12	14	13	**19**	5. m a g e 2
14	**26**	22	7	**21**	12	12	Image 1
15	19	15	**21**	17	7	**22**	
16	**30**	16	9	**24**	13	9	Image 2

The results of the experiment do not indicate the fixed order of the options led to excessive section of choice “A”. In image 201, both the randomized and the fixed conditions resulted in the selection of the first image panel as the choice of the crowd. In image 212, 12/16 questions were selected as choice “A” in the fixed presentation. Workers selected the same panel 13/16 times in the randomized presentation.

Agreement between experts and novices rating the quality of the algorithm output

201		211	
survey 1 novice	survey 1 expert	survey 2 novice	survey 2 expert

a	a	a	a
a	c	a	a
a	a	a	c
a	a	a	b
a	a	a	b
a	a	a	a
a	a	a	a
a	a	a	a
a	a	a	c
a	a	a	b
a	b	a	a
a	a	a	a
a	a	a	a
a	a	a	a
b	b	b	c
a	a	a	a

212		218	
survey 3 novice	survey 3 expert	survey 4 novice	survey 4 expert
a	b	a	a
a	a	a	a
a	b	b	b
b	a	a	a
a	a	a	a
a	a	a	a
a	a	a	a
a	a	a	a
a	a	a	a
a	b	a	a
a	a	a	b
a	a	a	a
b	b	a	a
a	a	a	a
c	c	a	a
a	b	a	a

	221	
	survey 5	
1	novice	survey 5 expert
2	a	a
3	a	a
4	a	a

5	a	a
6	a	a
7	a	a
8	a	a
9	a	a
10	a	a
11	a	a
12	a	a
13	a	a
14	a	a
15	a	a
16	a	a
	a	a

Kappa Values

Image 201

```
> kappa2(cbind(a, b))
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 16
   Raters = 2
    Kappa = 0.458

        z = 2.56
  p-value = 0.0106
```

Image 211

```
> kappa2(cbind(a, b))
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 16
   Raters = 2
    Kappa = 0.068

        z = 0.573
  p-value = 0.566
```

Image 212

```
> kappa2(cbind(d, e))
```

```
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 16
   Raters = 2
    Kappa = 0.385

        z = 2.02
  p-value = 0.0439
```

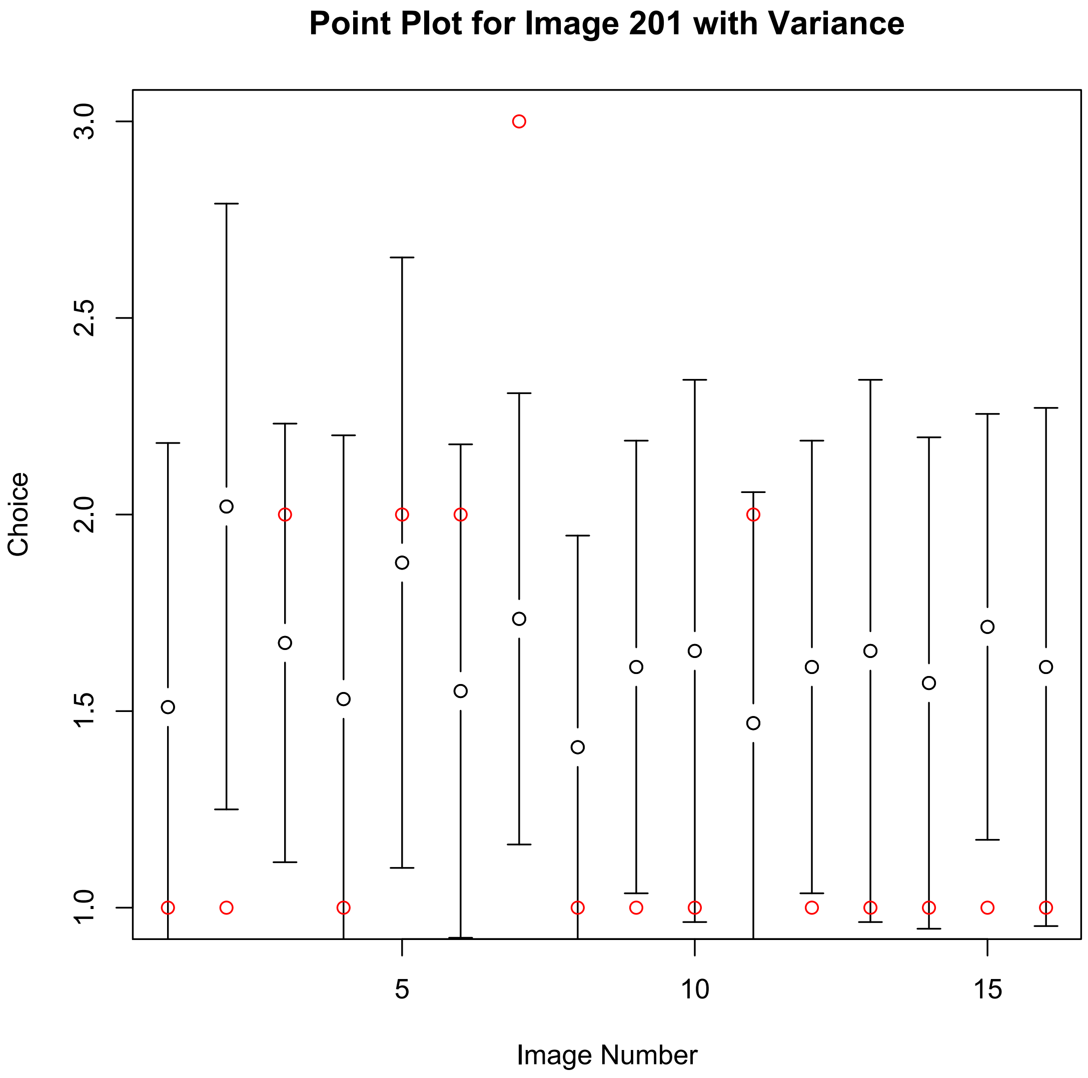
Image 218

```
> kappa2(cbind(f, g))
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 16
   Raters = 2
    Kappa = 0.636

        z = 2.73
  p-value = 0.00629
```

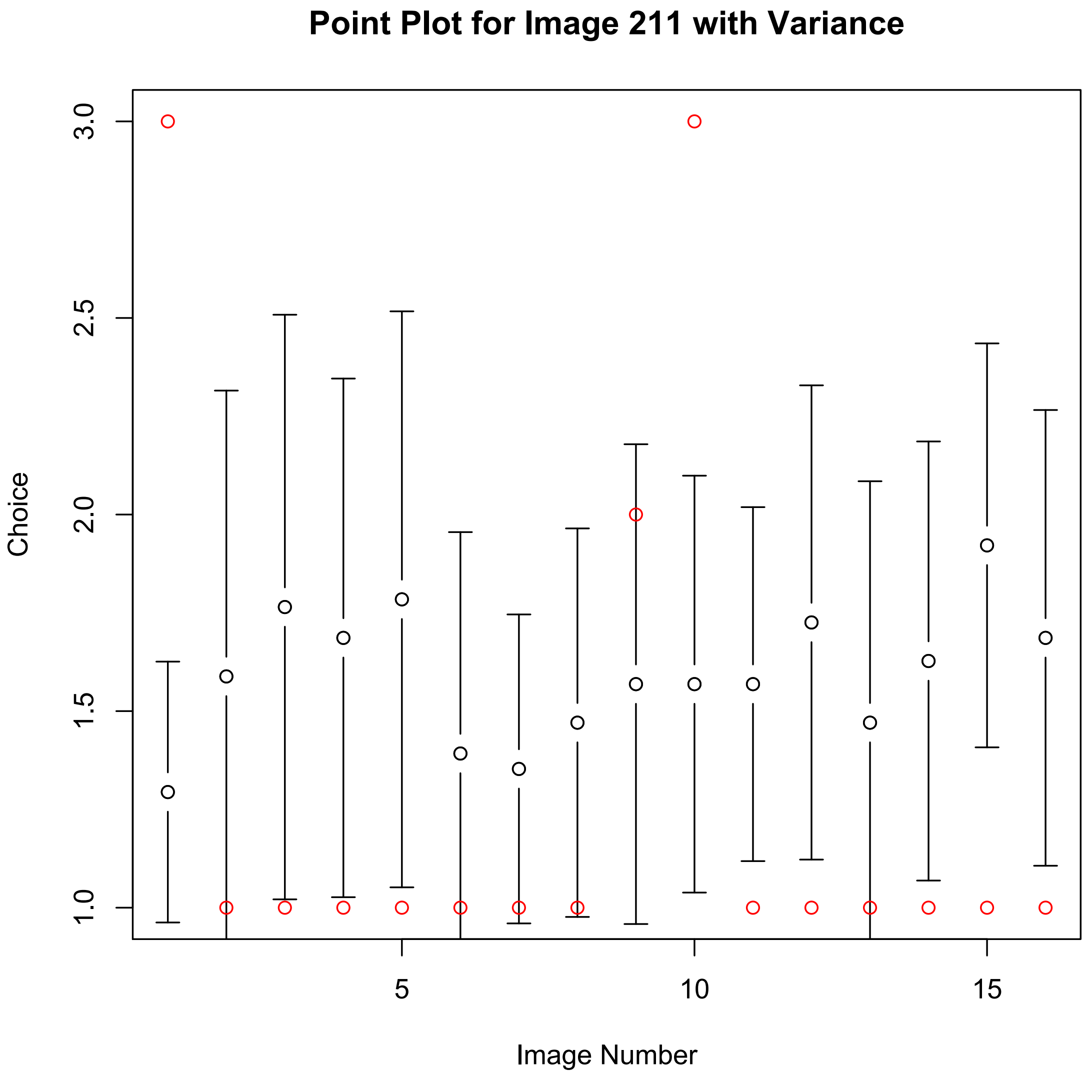
Image 221

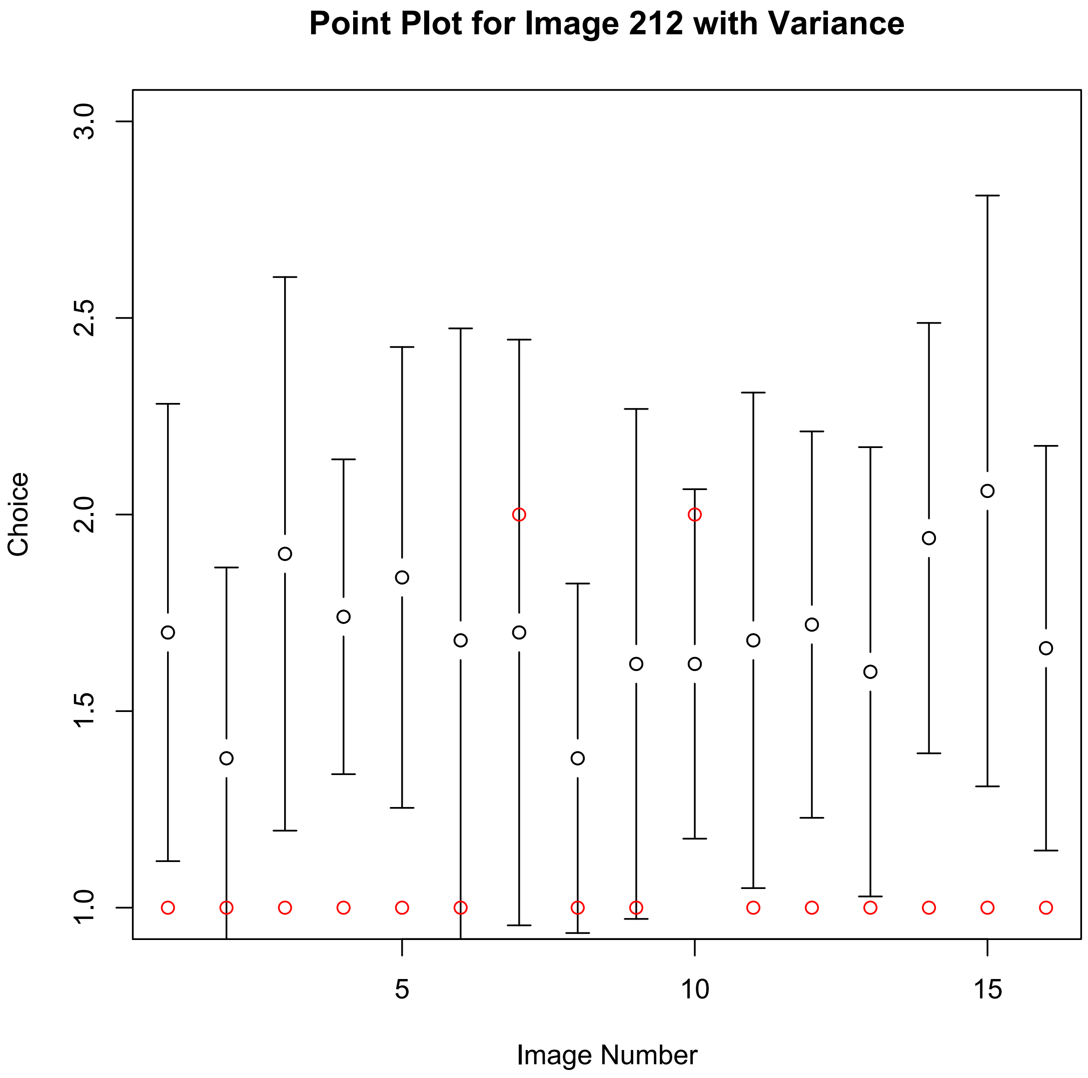
(Answers are the same between the two groups)

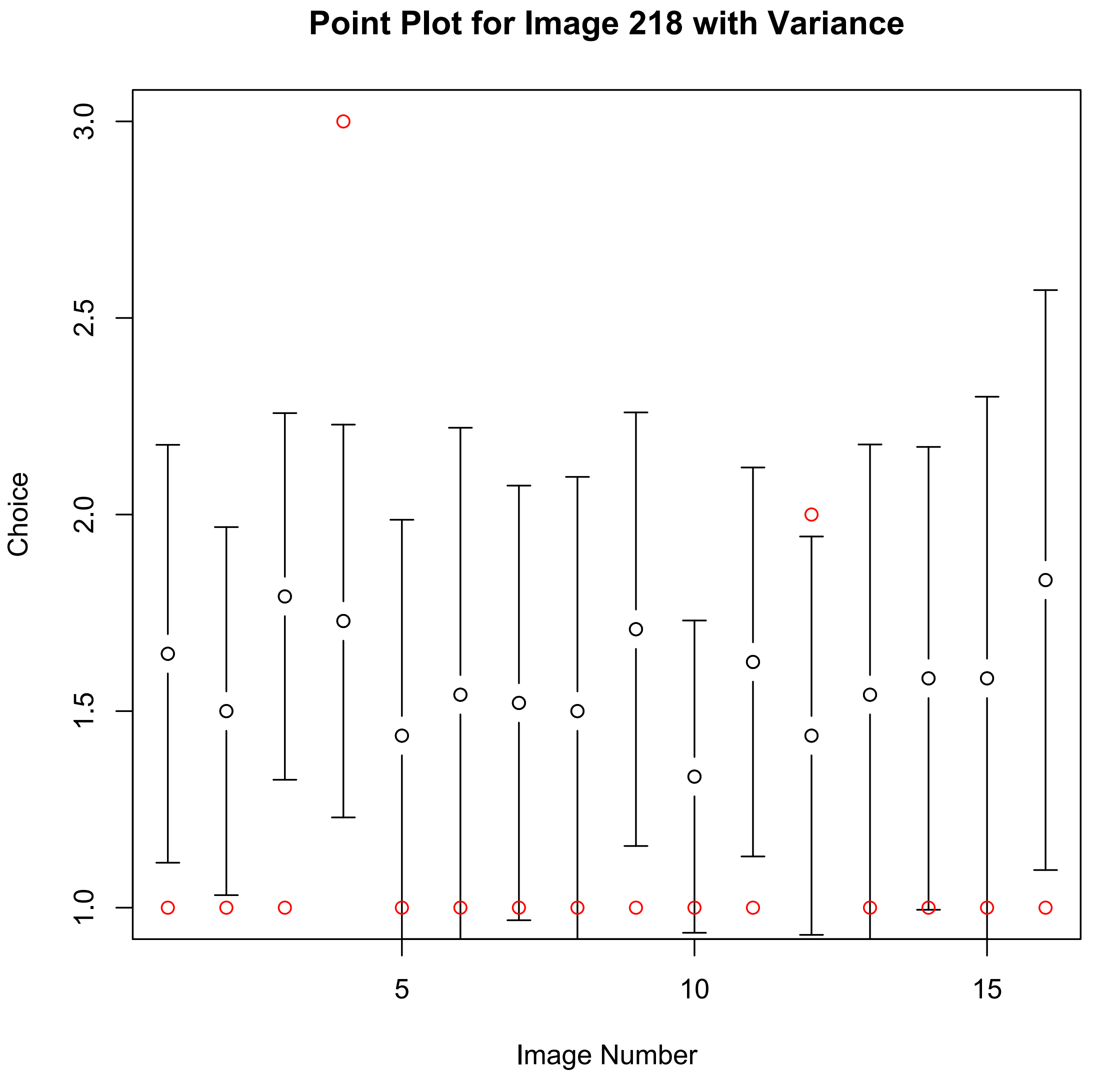
```
> kappa2(cbind(h, i))
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 16
   Raters = 2
    Kappa = NaN

        z = NaN
  p-value = NaN
```

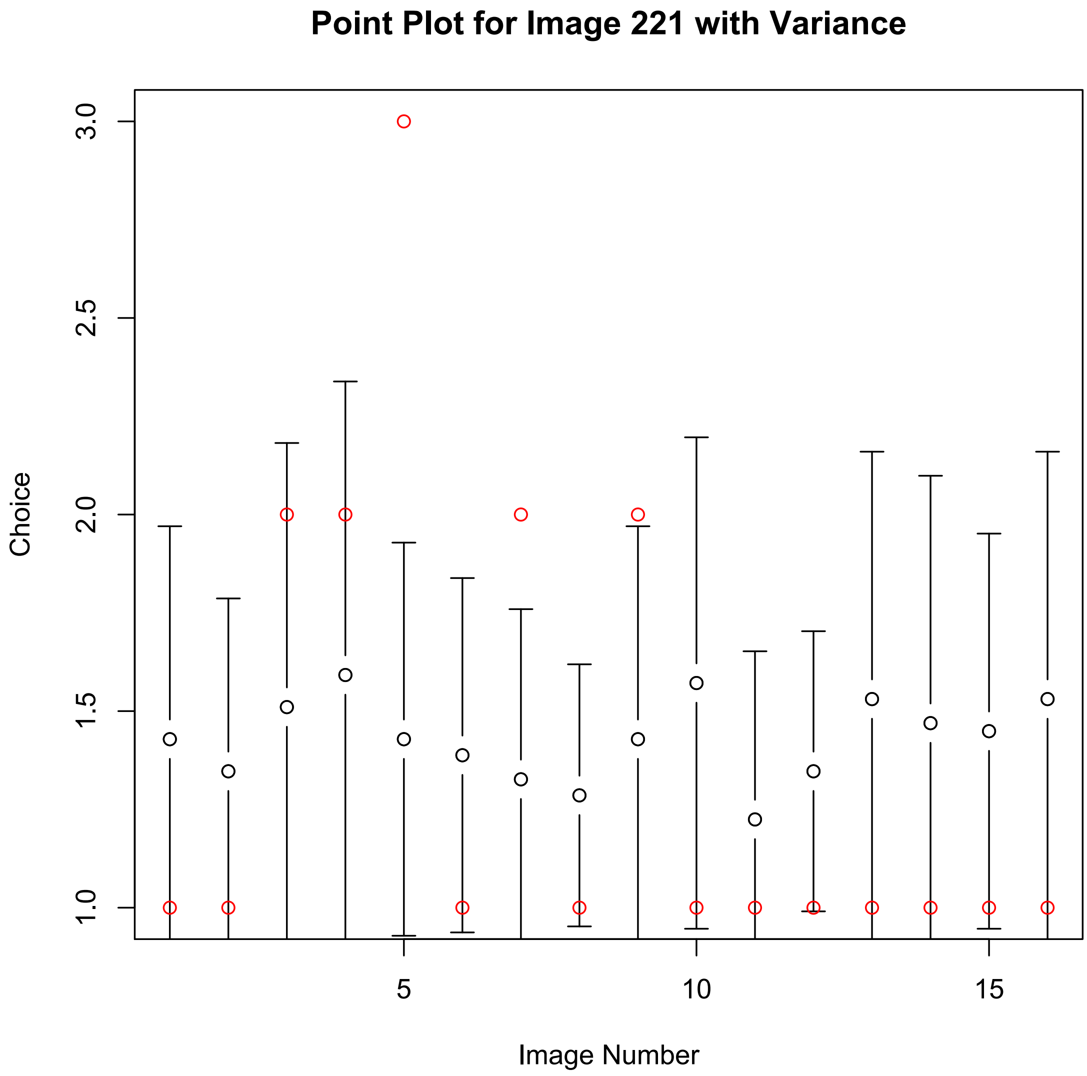


Point Plot for Image 201 with Variance
Choice
1.0
1.5
2.0
2.5
3.0
5
10
15
Image Number

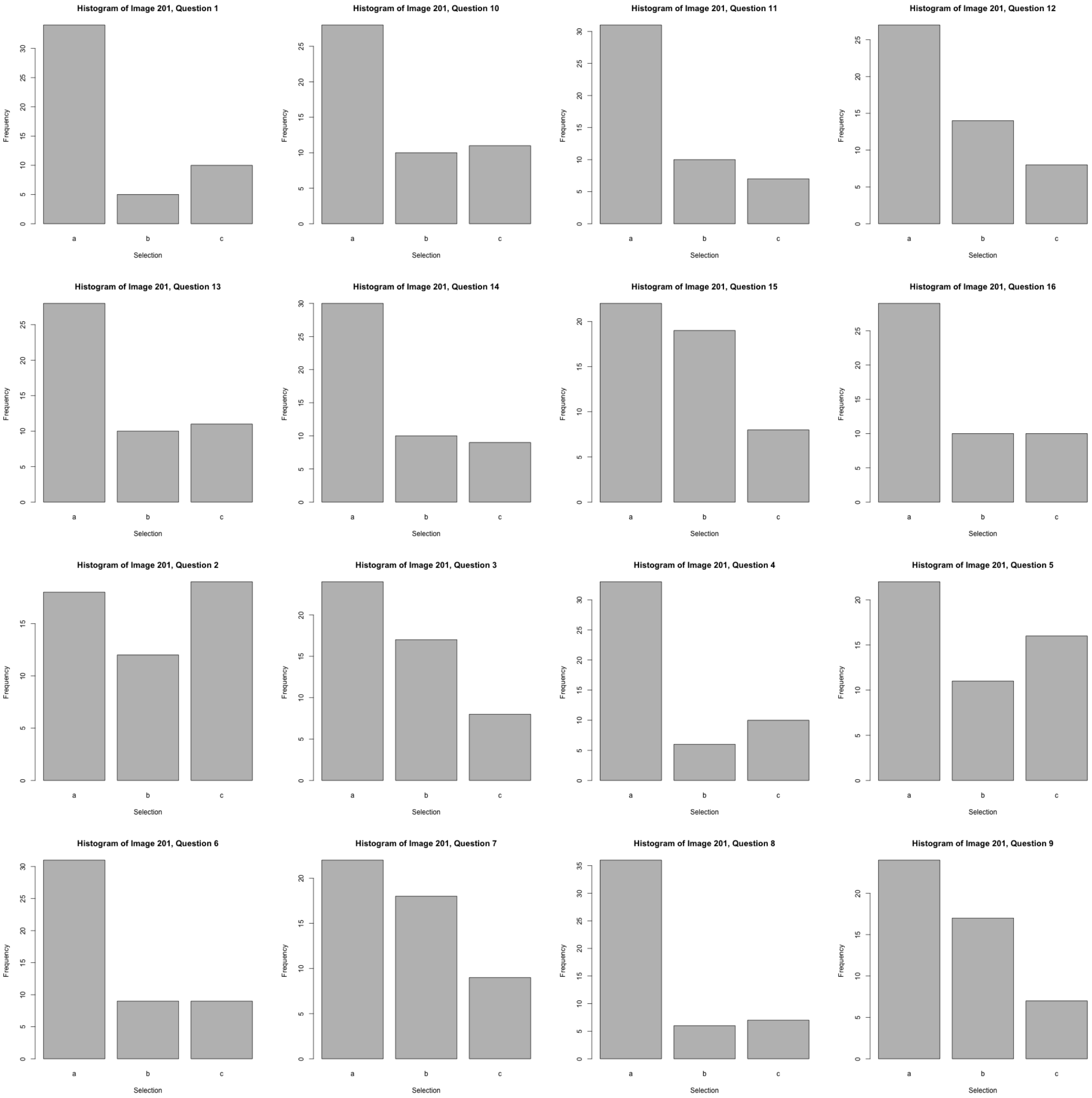

Point Plot for Image 211 with Variance
Choice
1.0
1.5
2.0
2.5
3.0
5
10
15
Image Number


Point Plot for Image 212 with Variance
Choice
1.0
1.5
2.0
2.5
3.0
5
10
15
Image Number


Point Plot for Image 218 with Variance
Choice
1.0
1.5
2.0
2.5
3.0
Image Number
5
10
15


Point Plot for Image 221 with Variance
Choice
3.0
2.5
2.0
1.5
1.0
5
10
15
Image Number

Series of plots visualizing user responses for Image 201



Histogram of Image 201, Question 1

Histogram of Image 201, Question 10

Histogram of Image 201, Question 11

Histogram of Image 201, Question 12

Histogram of Image 201, Question 13

Histogram of Image 201, Question 14

Histogram of Image 201, Question 15

Histogram of Image 201, Question 16

Histogram of Image 201, Question 2

Histogram of Image 201, Question 3

Histogram of Image 201, Question 4

Histogram of Image 201, Question 5

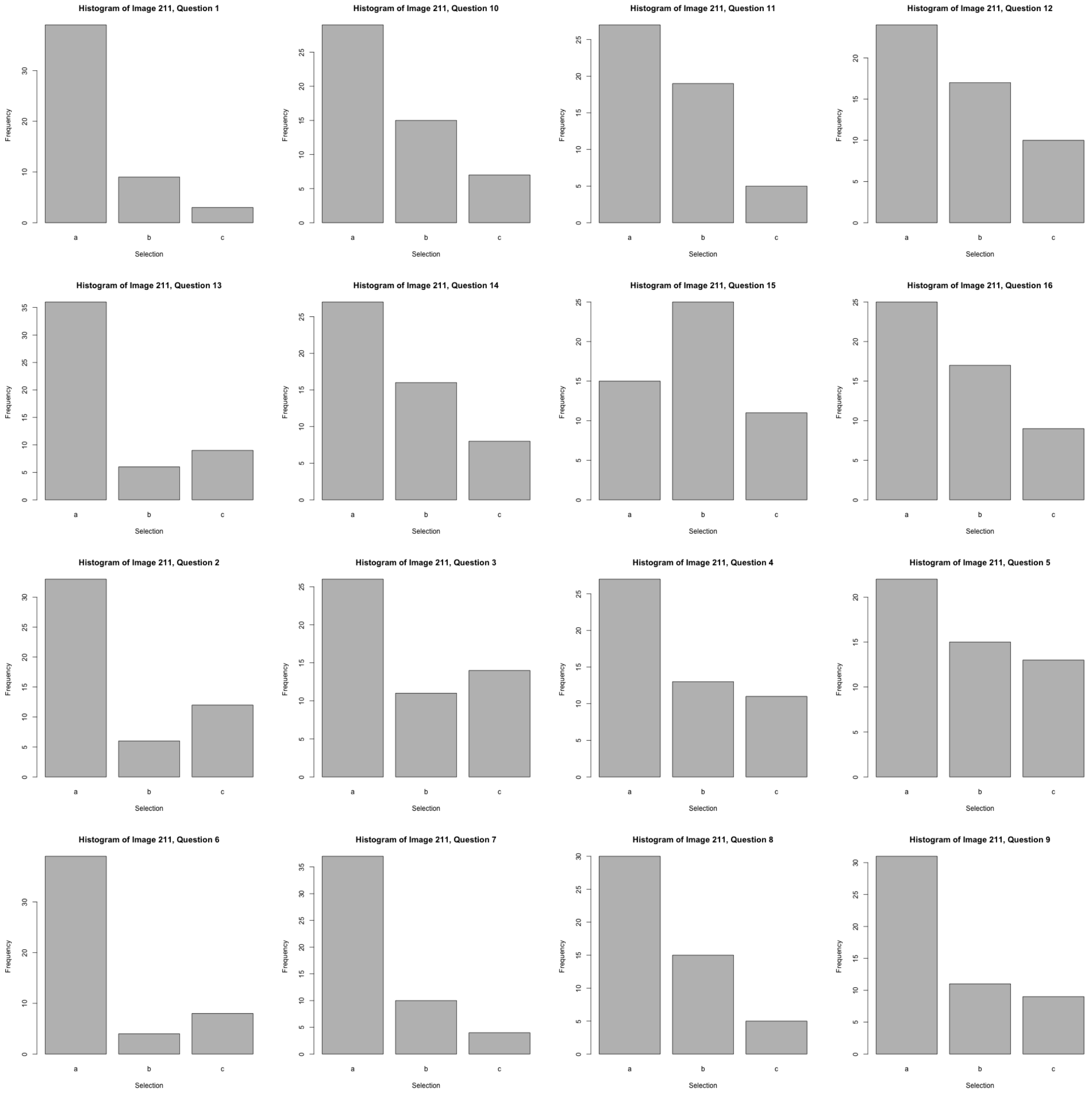
Histogram of Image 201, Question 6

Histogram of Image 201, Question 7

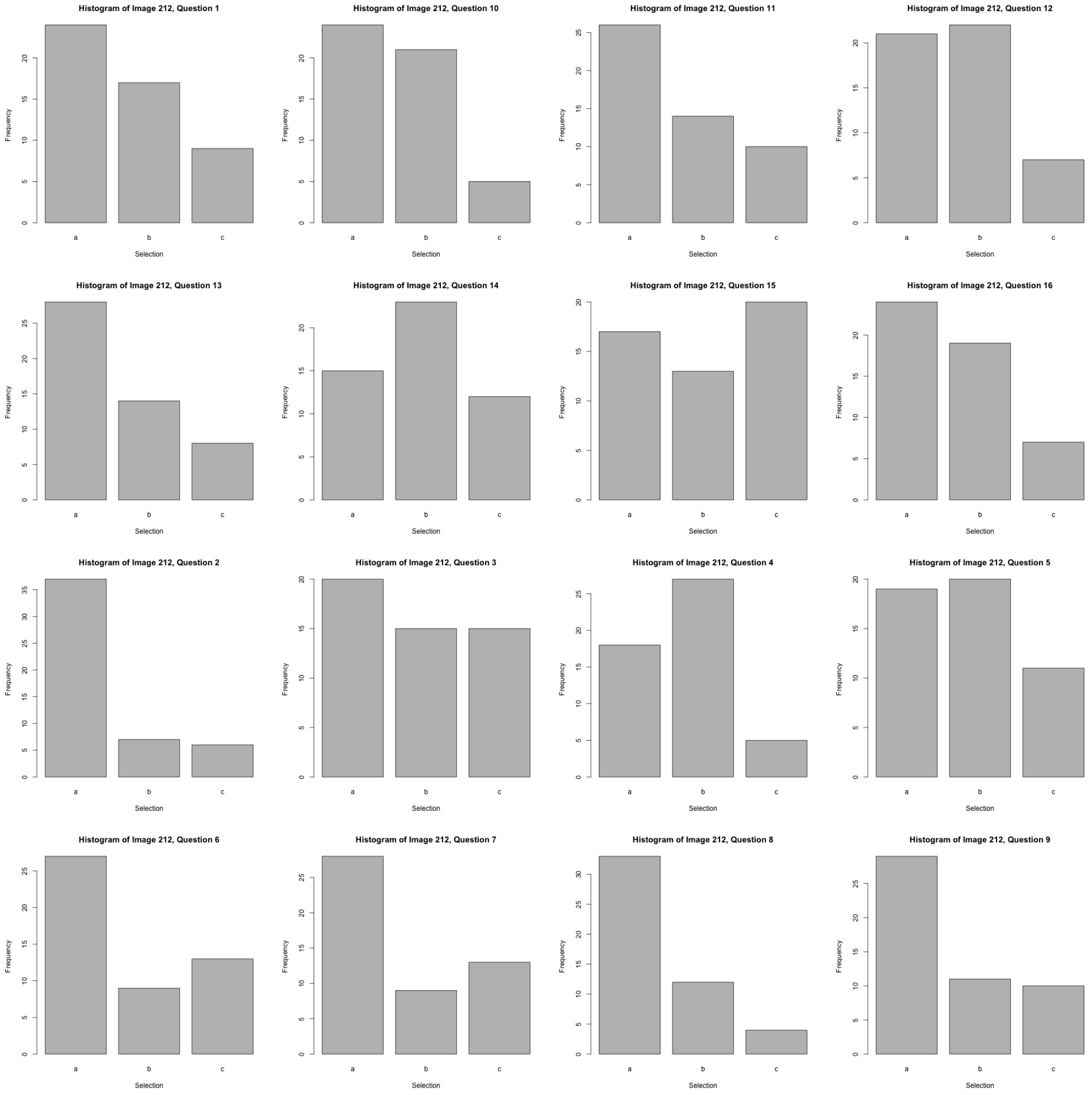
Histogram of Image 201, Question 8

Histogram of Image 201, Question 9

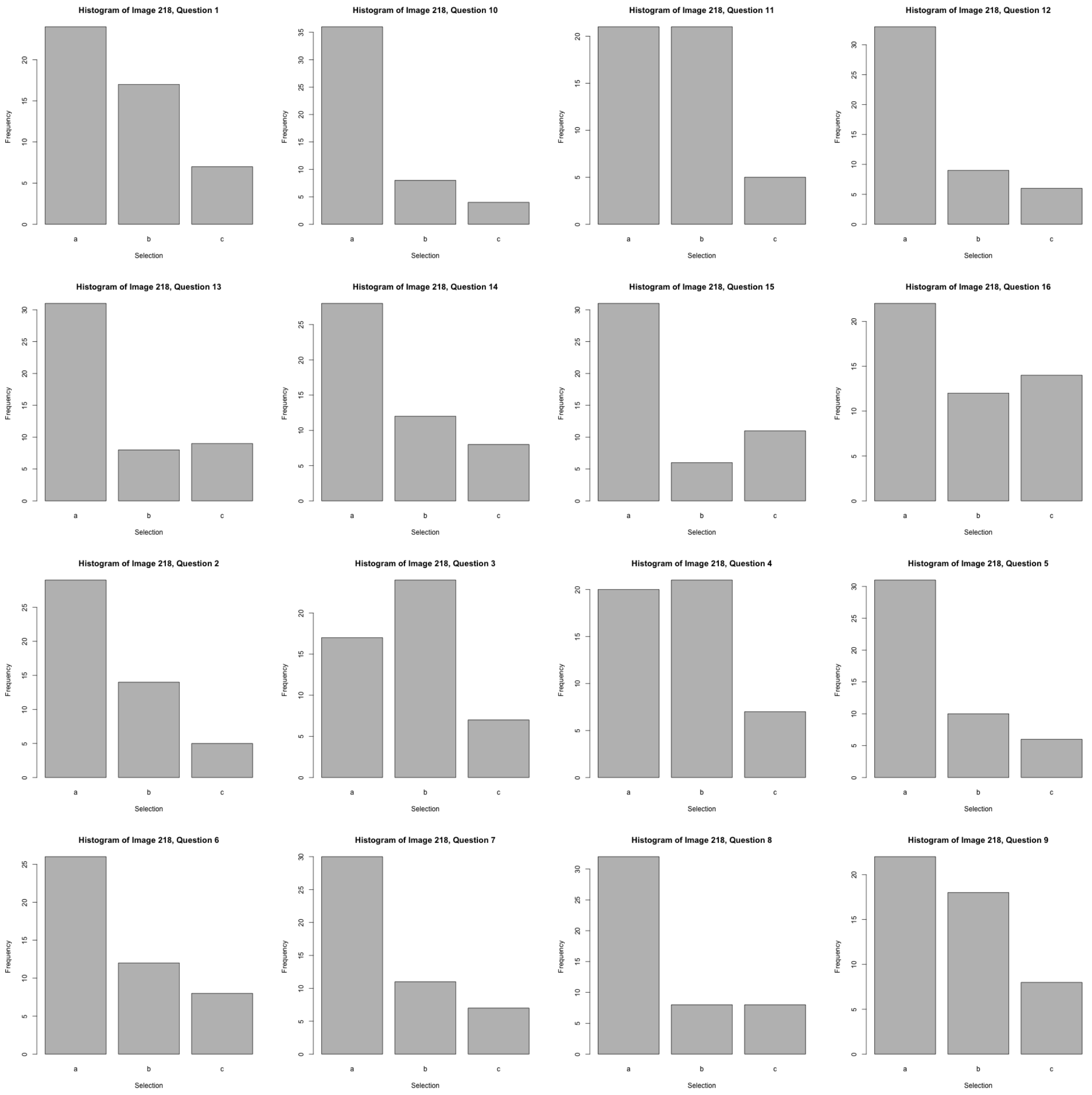
Series of plots visualizing user responses for Image 211



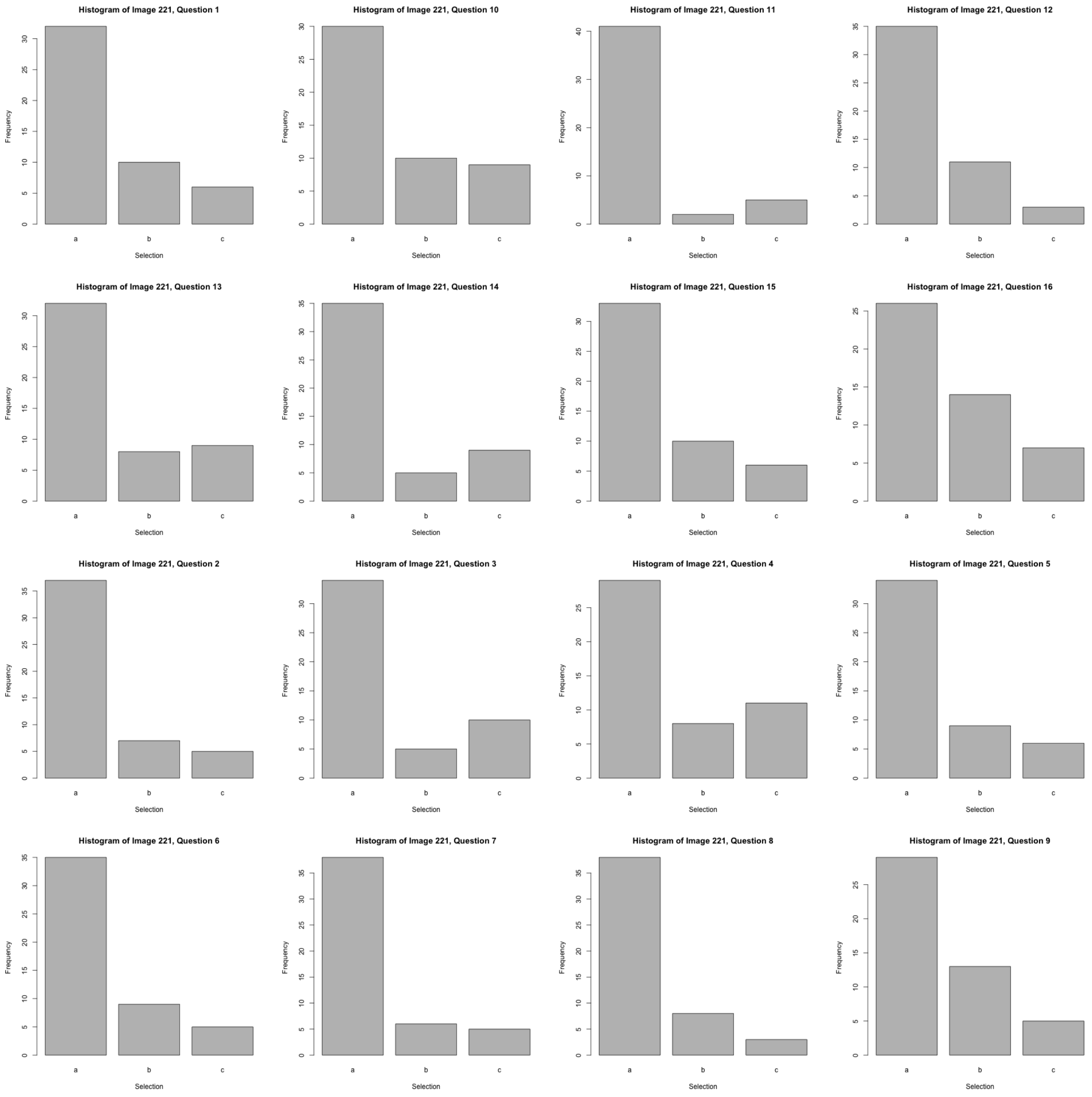
Series of plots visualizing user responses for Image 212



Series of plots visualizing user responses for Image 218



Series of plots visualizing user responses for Image 221



APPENDIX B

Experiment 2 materials

PHP Script for adding a new entry to the Google Fusion Table

```
<?php

include('../phplibs/clientlogin.php');
include('../phplibs/sql.php');
include('../phplibs/file.php');

$question = $_GET['q'];
$ftable= $_GET['table'];

//if question is not null then....

if ($question != NULL) {
   //Login to fusiontables
   $token                                                        =
ClientLogin::getAuthToken('davidcalit2computer@gmail.com',
'Ki$$my@ss!');
   $ftclient = new FTClientLogin($token);

  //Insert the text into table with Votes = 0
  echo               $ftclient->query(SQLBuilder::insert($ftable,
array('Text'=> $question, 'Votes' => 0)));

}

      ?>
```

Ranking Entries

```
<?php

include('../phplibs/clientlogin.php');
include('../phplibs/sql.php');
include('../phplibs/file.php');

require_once("../phplibs/Stat.class.php");

$rank = $_GET['rank'];
$debug= FALSE;
$ftable= $_GET['table'];

//get token
$token                                                           =
ClientLogin::getAuthToken('davidcalit2computer@gmail.com',
'PASSWORD_REMOVED');
$ftclient = new FTClientLogin($token);

$av=explode("\n",$ftclient->query(SQLBuilder::select($ftable,
array('rowid', 'Votes'))));

//echo "length of array". sizeof($av);
//echo "<br />";

  unset($av[0]);

  $av=(array_filter($av));

  if ($debug) print_r($av);

$counter=0;
foreach($av as $cliff_av){
  $newArray = explode(',',$cliff_av);
  $rowId[] = $newArray['0'];
  $voteCount[] = $newArray['1'];
  $counterarray[]= $counter++;
}

$counter=0;
arsort($voteCount);
foreach($voteCount as $key=>$value){
  $counterarray[$key]=$counter++;

  if ($counterarray[$key] == $rank){
     $myrank=$rowId[$key];
     }
}

  if (!(is_numeric($rank))){
    //nab array of votes from fusiontables
    $getvalues=$ftclient-
>query(SQLBuilder::select($ftable,array('Votes')));
    $valuearray=explode("\n",$getvalues);
    unset($valuearray[0]);
    $count=count($valuearray);
    $valuearray=(array_filter($valuearray));
    if ($debug) print_r($valuearray);

    unset($value);

    if ($debug) print_r($valuearray);

    $stat= new Stat();

    $first = $stat->percentile($valuearray,25);
    $second = $stat->median($valuearray);
    $third = $stat->percentile($valuearray,75);

    if ($debug==TRUE){
      echo "<br />";
      echo "<br />";
      echo "Quartiles:  ";
      echo "<br />";
      echo "first quartile: " . $first;
      echo "<br />";
      echo "second quartile: " . $second;
      echo "<br />";
      echo "third quartile: " . $third;
      echo "<br />";
    }

    switch ($rank) {
      case "first":
        $tempstring=$ftclient-
>query(SQLBuilder::select($ftable,array('rowid'),"'Votes'      <=
'$first'"));
        break;
      case "second":
        $tempstring=$ftclient-
>query(SQLBuilder::select($ftable,array('rowid'),"'Votes'      >=
'$first' AND 'Votes' <= '$second'"));
        break;
      case "third":
        $tempstring=$ftclient-
>query(SQLBuilder::select($ftable,array('rowid'),"'Votes'      >=
'$second' AND 'Votes' <= '$third'"));
    break;
      case "fourth":
        $tempstring=$ftclient-
>query(SQLBuilder::select($ftable,array('rowid'),"'Votes'      >=
'$third'"));
        break;
    }

    $temparray = explode("\n",$tempstring);
    unset($temparray[0]);

    array_pop($temparray);

    if ($debug==TRUE) {
      echo "<br />";
      print_r($temparray);
      echo "<br />";
    }

    $tempindex=array_rand($temparray);
    $myrank=$temparray[$tempindex];
    $string=$ftclient-
>query(SQLBuilder::select($ftable,array('Text'),"'rowid'        =
'$myrank'"));
  }


  if (is_numeric($rank)){
    $string=$ftclient-
>query(SQLBuilder::select($ftable,array('Text'),"'rowid'        =
'$myrank'"));
  }

  $replacements='Rank' . $rank . '=';
  $pattern='/Text/';
  echo preg_replace($pattern, $replacements, $string);

  if ($debug){
    echo "<br /> rowid: " . $myrank;
    echo       "<br       />"        .       $string=$ftclient-
>query(SQLBuilder::select($ftable,array('Votes'),"'rowid'       =
'$myrank'"));
  }
?>
```

Updating Votes

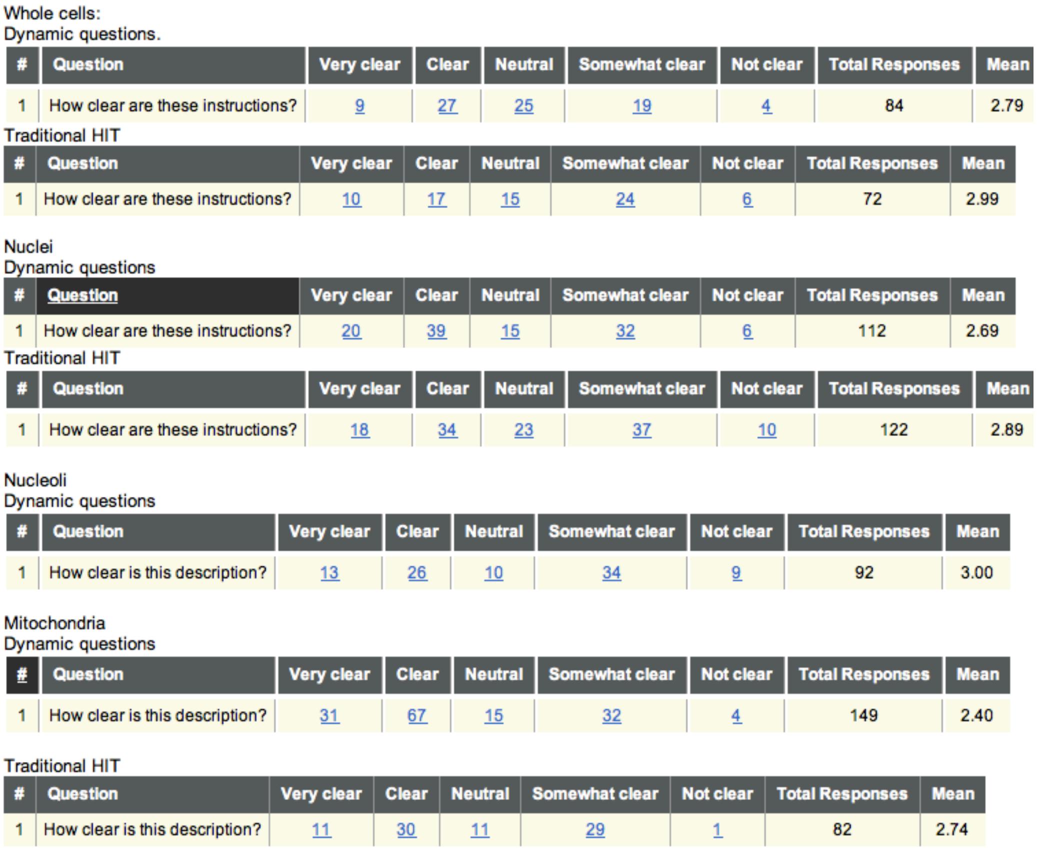
```
<?php
//ini_set('error_reporting', E_ALL);

include('../phplibs/clientlogin.php');
include('../phplibs/sql.php');
include('../phplibs/file.php');
```

```
$topvote = trim($_GET['v']);
$ftable= $_GET['table'];
$myrow=0;

$debug=FALSE;

//if question is not null then....

if ($topvote != NULL) {
   //Login to fusiontables
   $token                                                          =
ClientLogin::getAuthToken('davidcalit2computer@gmail.com',
'Ki$$my@ss!');
   $ftclient = new FTClientLogin($token);

   if ($debug) echo "query string:  ".$topvote;
   echo "<br />";
   echo "table: " . $ftable;
   echo "<br />";

     $rawrows=$ftclient->query(SQLBuilder::select($ftable,
array('rowid')));
     $rowarray=explode("\n",$rawrows);

     unset($rowarray[0]);
     print_r($rowarray);

     for ($i=1; $i < sizeof($rowarray); $i++){

$textarray[$i]=preg_replace('/[\n\r]/','',substr(trim($ftclient-
>query(SQLBuilder::select($ftable,                 array('Text'),
"'rowid'='$rowarray[$i]'")),4));
     if ($textarray[$i] == $topvote){
        echo "****BAZINGA <br />";
        echo   "rowid   of   what   I'm   looking   for   is   "   .
$myrow=$rowarray[$i]."<br />";
        }
     }

if ($debug)     print_r($textarray);

     echo "<br />";
     $myrow=intval($myrow);

     preg_match("/([\d]+)/",$ftclient-
>query(SQLBuilder::select($ftable,                array('Votes'),
"'rowid'='$myrow'")),$newvote);

   //increment vote count
```

```
    $newvote[0]++;
    echo "<br />new vote: " . $newvote[0] . "<br />";
    echo               $ftclient->query(SQLBuilder::update($ftable,
array('Votes'=>$newvote[0]), $myrow));
}

?>
```

Summary of satisfaction data

The following is a table of the satisfaction data reported for both the control and experimental groups for each of the four experiments as reported by Qualtrics. Lower scores are better.



Whole cells:
Dynamic questions.

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear are these instructions?	9	27	25	19	4	84	2.79

Traditional HIT

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear are these instructions?	10	17	15	24	6	72	2.99

Nuclei
Dynamic questions

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear are these instructions?	20	39	15	32	6	112	2.69

Traditional HIT

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear are these instructions?	18	34	23	37	10	122	2.89

Nucleoli
Dynamic questions

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear is this description?	13	26	10	34	9	92	3.00

Mitochondria
Dynamic questions

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear is this description?	31	67	15	32	4	149	2.40

Traditional HIT

#	Question	Very clear	Clear	Neutral	Somewhat clear	Not clear	Total Responses	Mean
1	How clear is this description?	11	30	11	29	1	82	2.74

APPENDIX C

Experiment 3 materials

Two methods for generating animations. The first method attempts to create all of the images while the second method creates a script that creates the images. The second method is thought to be more reliable.

Generating all animations

Automated methods for generating animations from all of the images files

```
#!/bin/bash

NUMOFTILES=16
#NUMOFTILES - how many chunks the image is divided into.
NUMOFZ=10
#NUMOFZ - how many slices in Z you want the stack to go.
TOTALIMAGES=184
#NUMOFIMAGES - Number of images that need to be sliced and diced.
DELAY=50
#DELAY - the gap between displaying images
COMMAND="convert -delay $DELAY "

for i in `seq 0 $NUMOFTILES`;
 do
  for j in `seq 0 $(($TOTALIMAGES/$NUMOFZ))`;
   do
    BASE=$(($j*$NUMOFTILES*$NUMOFZ))
    for k in `seq 0 $(($NUMOFZ-1))`;
     do
     COMMAND="$COMMAND
image$(($k*$NUMOFTILES+$BASE+$i)).png"
     done
  COMMAND="$COMMAND ./testgifs/animation$(($j+$i)).gif"
```

```
    echo $(($j+$i))
    COMMAND="convert -delay $DELAY "
      done
  done
```

Alternate method for generating animations one stack at a time

```
#!/bin/bash

counter=0;

echo "#!/bin/bash"
for i in `seq 50 59`;
  do
    echo "convert dlee_mask0"$i".png -crop 4x4@ +repage +adjoin
mask"$counter"-%d.png";
    let counter++;
  done

for i in {0..15}; do convert -delay 50 \
mask0-$i.png \
mask1-$i.png \
mask2-$i.png \
mask3-$i.png \
mask4-$i.png \
mask5-$i.png \
mask6-$i.png \
mask7-$i.png \
mask8-$i.png \
mask9-$i.png animation$i.gif; done
```

5.4.3 Sample output from the script

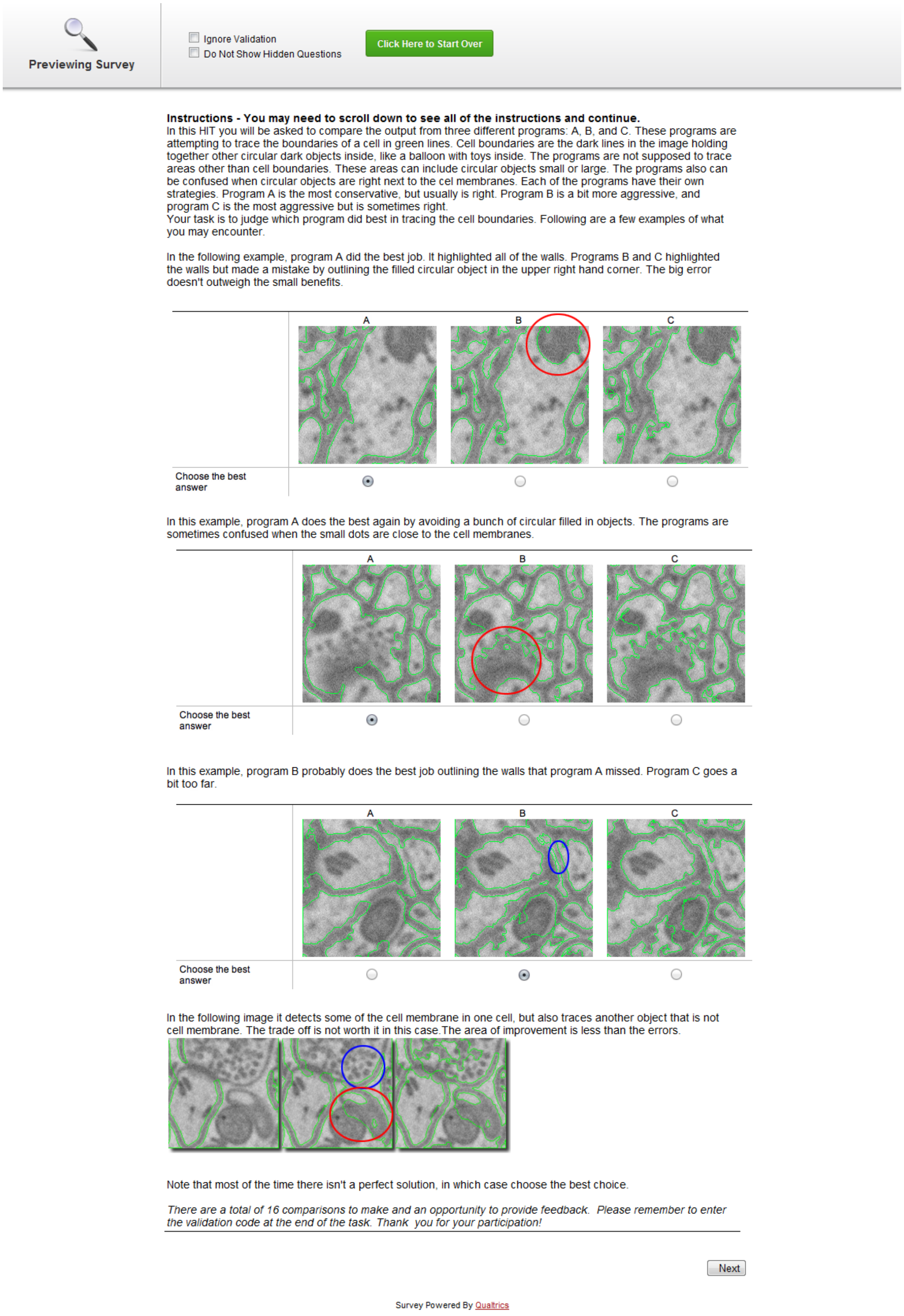
```
#!/bin/bash
convert dlee_mask010.png -crop 4x4@ +repage +adjoin mask0-%d.png
convert dlee_mask011.png -crop 4x4@ +repage +adjoin mask1-%d.png
convert dlee_mask012.png -crop 4x4@ +repage +adjoin mask2-%d.png
convert dlee_mask013.png -crop 4x4@ +repage +adjoin mask3-%d.png
convert dlee_mask014.png -crop 4x4@ +repage +adjoin mask4-%d.png
convert dlee_mask015.png -crop 4x4@ +repage +adjoin mask5-%d.png
convert dlee_mask016.png -crop 4x4@ +repage +adjoin mask6-%d.png
convert dlee_mask017.png -crop 4x4@ +repage +adjoin mask7-%d.png
convert dlee_mask018.png -crop 4x4@ +repage +adjoin mask8-%d.png
convert dlee_mask019.png -crop 4x4@ +repage +adjoin mask9-%d.png
```

APPENDIX D

Qualtrics Questions

Instructions for distributing expertise

PHP Script for adding a new entry to the Google Fusion Table



Instructions - You may need to scroll down to see all of the instructions and continue.
In this HIT you will be asked to compare the output from three different programs: A, B, and C. These programs are attempting to trace the boundaries of a cell in green lines. Cell boundaries are the dark lines in the image holding together other circular dark objects inside, like a balloon with toys inside. The programs are not supposed to trace areas other than cell boundaries. These areas can include circular objects small or large. The programs also can be confused when circular objects are right next to the cel membranes. Each of the programs have their own strategies. Program A is the most conservative, but usually is right. Program B is a bit more aggressive, and program C is the most aggressive but is sometimes right.
Your task is to judge which program did best in tracing the cell boundaries. Following are a few examples of what you may encounter.

In the following example, program A did the best job. It highlighted all of the walls. Programs B and C highlighted the walls but made a mistake by outlining the filled circular object in the upper right hand corner. The big error doesn't outweigh the small benefits.

	A	B	C
Choose the best answer	●	○	○

In this example, program A does the best again by avoiding a bunch of circular filled in objects. The programs are sometimes confused when the small dots are close to the cell membranes.

	A	B	C
Choose the best answer	●	○	○

In this example, program B probably does the best job outlining the walls that program A missed. Program C goes a bit too far.

	A	B	C
Choose the best answer	○	●	○

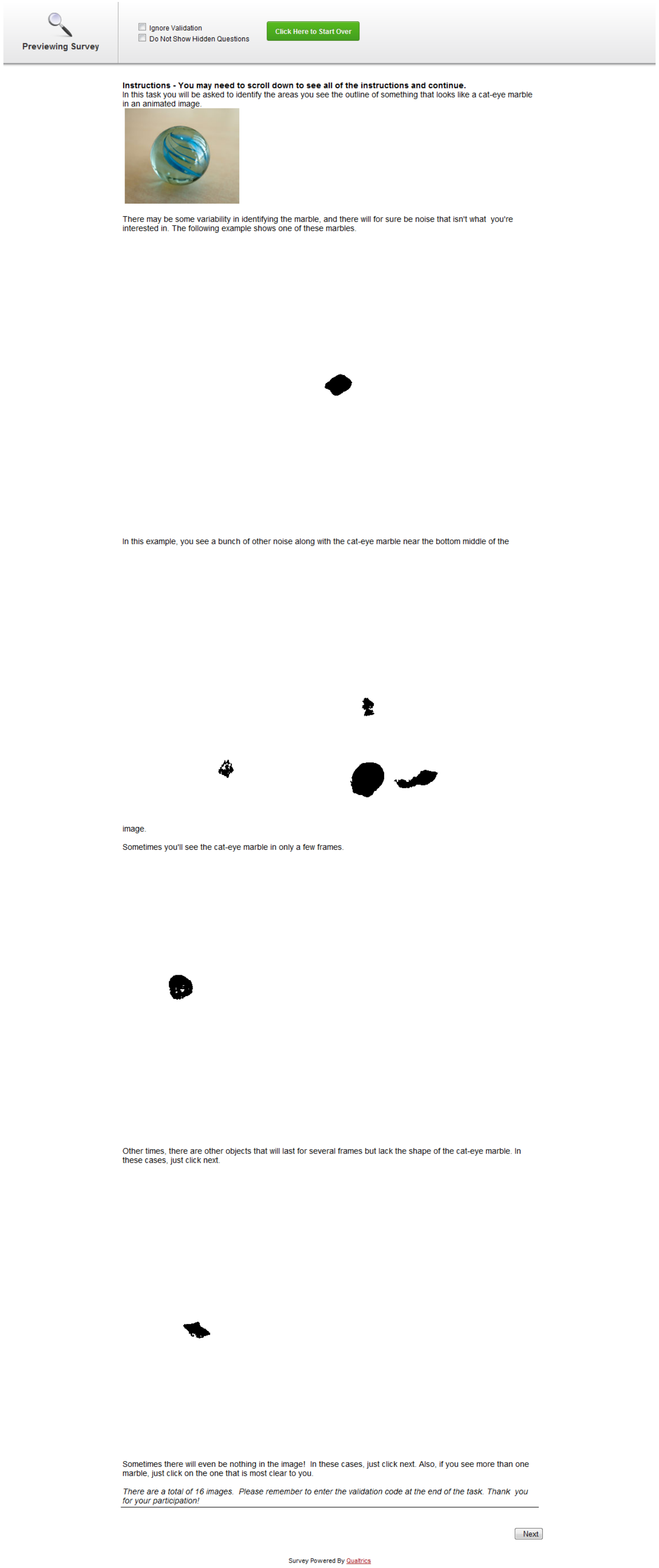
In the following image it detects some of the cell membrane in one cell, but also traces another object that is not cell membrane. The trade off is not worth it in this case.The area of improvement is less than the errors.

Note that most of the time there isn't a perfect solution, in which case choose the best choice.

There are a total of 16 comparisons to make and an opportunity to provide feedback. Please remember to enter the validation code at the end of the task. Thank you for your participation!

Survey Powered By Qualtrics

Instructions for task transformation



Instructions - You may need to scroll down to see all of the instructions and continue.
In this task you will be asked to identify the areas you see the outline of something that looks like a cat-eye marble in an animated image.

There may be some variability in identifying the marble, and there will for sure be noise that isn't what you're interested in. The following example shows one of these marbles.

In this example, you see a bunch of other noise along with the cat-eye marble near the bottom middle of the image.

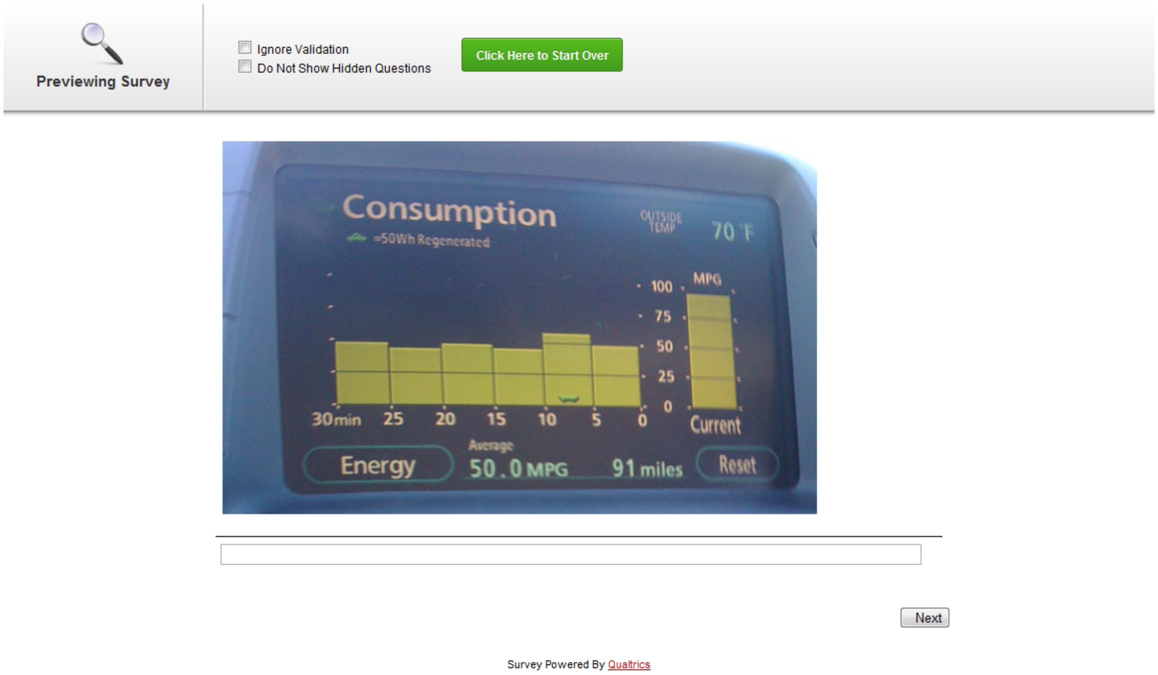
Sometimes you'll see the cat-eye marble in only a few frames.

Other times, there are other objects that will last for several frames but lack the shape of the cat-eye marble. In these cases, just click next.

Sometimes there will even be nothing in the image! In these cases, just click next. Also, if you see more than one marble, just click on the one that is most clear to you.

There are a total of 16 images. Please remember to enter the validation code at the end of the task. Thank you for your participation!

Instructions for Dynamic Questions (Pilot Test)

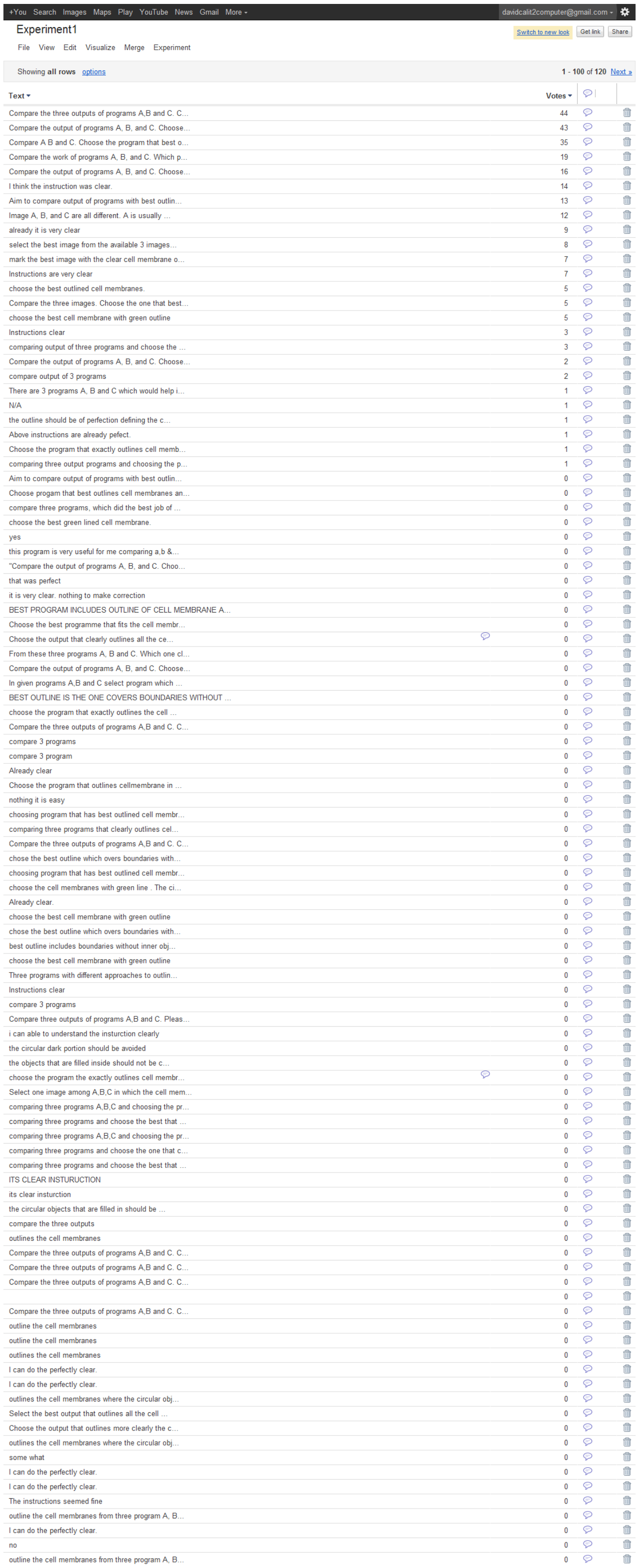


APPENDIX E

Google Fusion Tables

Results from Experiment 1 – Distributing expertise

The results stored in Google Fusion Tables are accessed through a GUI. Screenshots of the user interface and data are included below.

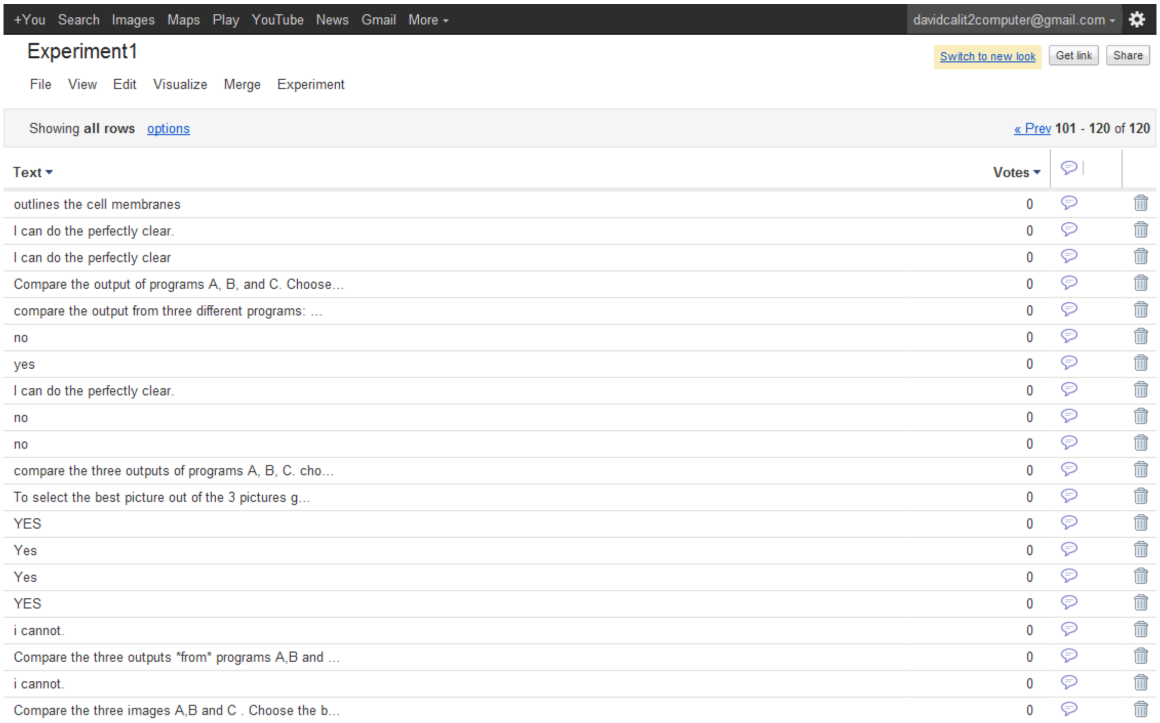

+You Search Images Maps Play YouTube News Gmail More
davidcalit2computer@gmail.com
Switch to new look
Get link
Share
File View Edit Visualize Merge Experiment

Experiment1

Showing **all rows** options

1 - **100** of **120** Next »

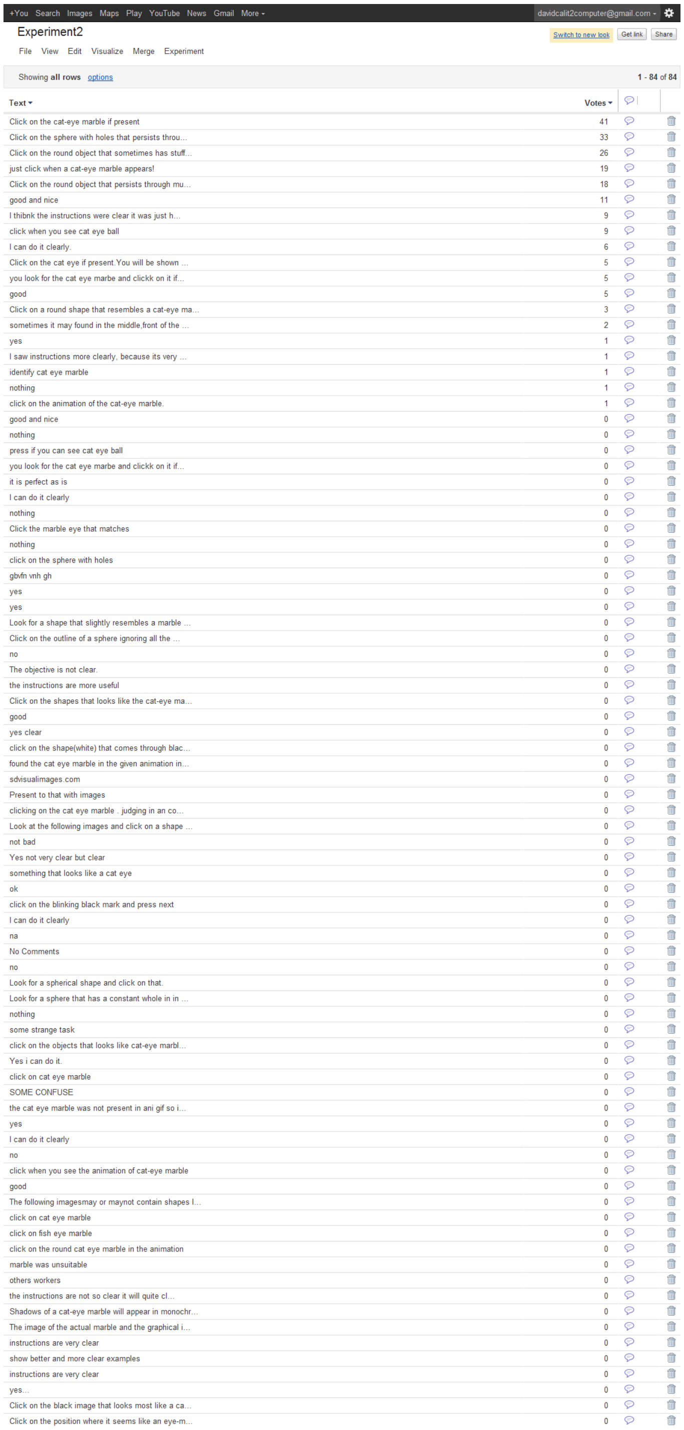
Text	Votes
Compare the three outputs of programs A,B and C. C...	44
Compare the output of programs A, B, and C. Choose...	43
Compare A B and C. Choose the program that best o...	35
Compare the work of programs A, B, and C. Which p...	19
Compare the output of programs A, B, and C. Choose...	16
I think the instruction was clear.	14
Aim to compare output of programs with best outlin...	13
Image A, B, and C are all different. A is usually ...	12
already it is very clear	9
select the best image from the available 3 images...	8
mark the best image with the clear cell membrane o...	7
Instructions are very clear	7
choose the best outlined cell membranes.	5
Compare the three images. Choose the one that best...	5
choose the best cell membrane with green outline	5
Instructions clear	3
comparing output of three programs and choose the ...	3
Compare the output of programs A, B, and C. Choose...	2
compare output of 3 programs	2
There are 3 programs A, B and C which would help i...	1
N/A	1
the outline should be of perfection defining the c...	1
Above instructions are already pefect.	1
Choose the program that exactly outlines cell memb...	1
comparing three output programs and choosing the p...	1
Aim to compare output of programs with best outlin...	0
Choose progam that best outlines cell membranes an...	0
compare three programs, which did the best job of ...	0
choose the best green lined cell membrane.	0
yes	0
this program is very useful for me comparing a,b &...	0
"Compare the output of programs A, B, and C. Choo...	0
that was perfect	0
it is very clear. nothing to make correction	0
BEST PROGRAM INCLUDES OUTLINE OF CELL MEMBRANE A...	0
Choose the best programme that fits the cell membr...	0
Choose the output that clearly outlines all the ce...	0
From these three programs A, B and C. Which one cl...	0
Compare the output of programs A, B, and C. Choose...	0
In given programs A,B and C select program which ...	0
BEST OUTLINE IS THE ONE COVERS BOUNDARIES WITHOUT ...	0
choose the program that exactly outlines the cell ...	0
Compare the three outputs of programs A,B and C. C...	0
compare 3 programs	0
compare 3 program	0
Already clear	0
Choose the program that outlines cellmembrane in ...	0
nothing it is easy	0
choosing program that has best outlined cell membr...	0
comparing three programs that clearly outlines cel...	0
Compare the three outputs of programs A,B and C. C...	0
chose the best outline which overs boundaries with...	0
choosing program that has best outlined cell membr...	0
choose the cell membranes with green line . The ci...	0
Already clear.	0
choose the best cell membrane with green outline	0
chose the best outline which overs boundaries with...	0
best outline includes boundaries without inner obj...	0
choose the best cell membrane with green outline	0
Three programs with different approaches to outlin...	0
Instructions clear	0
compare 3 programs	0
Compare three outputs of programs A,B and C. Pleas...	0
i can able to understand the insturction clearly	0
the circular dark portion should be avoided	0
the objects that are filled inside should not be c...	0
choose the program the exactly outlines cell membr...	0
Select one image among A,B,C in which the cell mem...	0
comparing three programs A,B,C and choosing the pr...	0
comparing three programs and choose the best that ...	0
comparing three programs A,B,C and choosing the pr...	0
comparing three programs and choose the one that c...	0
comparing three programs and choose the best that ...	0
ITS CLEAR INSTURUCTION	0
its clear insturction	0
the circular objects that are filled in should be ...	0
compare the three outputs	0
outlines the cell membranes	0
Compare the three outputs of programs A,B and C. C...	0
Compare the three outputs of programs A,B and C. C...	0
Compare the three outputs of programs A,B and C. C...	0
	0
Compare the three outputs of programs A,B and C. C...	0
outline the cell membranes	0
outline the cell membranes	0
outlines the cell membranes	0
I can do the perfectly clear.	0
I can do the perfectly clear.	0
outlines the cell membranes where the circular obj...	0
Select the best output that outlines all the cell ...	0
Choose the output that outlines more clearly the c...	0
outlines the cell membranes where the circular obj...	0
some what	0
I can do the perfectly clear.	0
I can do the perfectly clear.	0
The instructions seemed fine	0
outline the cell membranes from three program A, B...	0
I can do the perfectly clear.	0
no	0
outline the cell membranes from three program A, B...	0


+You Search Images Maps Play YouTube News Gmail More
davidcalit2computer@gmail.com
Experiment1
Switch to new look
Get link
Share
File View Edit Visualize Merge Experiment
Showing all rows options
« Prev 101 - 120 of 120

Text	Votes
outlines the cell membranes	0
I can do the perfectly clear.	0
I can do the perfectly clear	0
Compare the output of programs A, B, and C. Choose…	0
compare the output from three different programs: …	0
no	0
yes	0
I can do the perfectly clear.	0
no	0
no	0
compare the three outputs of programs A, B, C. cho…	0
To select the best picture out of the 3 pictures g…	0
YES	0
Yes	0
Yes	0
YES	0
i cannot.	0
Compare the three outputs "from" programs A,B and …	0
i cannot.	0
Compare the three images A,B and C . Choose the b…	0

Additional data are available in the attached archive.

Results from Experiment 2 – Task transformation

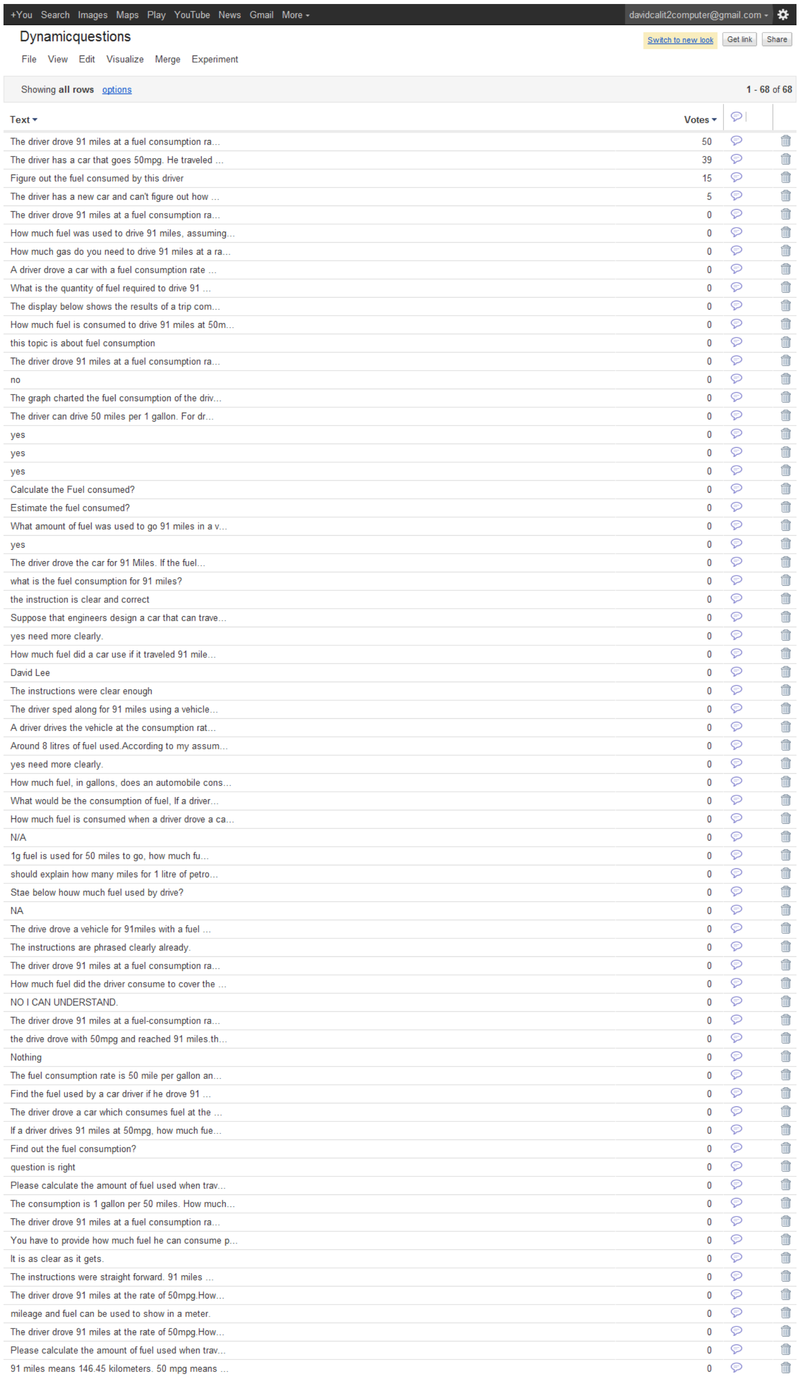


Experiment2

Showing **all rows** options **1 - 84** of **84**

Text ▾	Votes ▾
Click on the cat-eye marble if present	41
Click on the sphere with holes that persists throu...	33
Click on the round object that sometimes has stuff...	26
just click when a cat-eye marble appears!	19
Click on the round object that persists through mu...	18
good and nice	11
I thibnk the instructions were clear it was just h...	9
click when you see cat eye ball	9
I can do it clearly.	6
Click on the cat eye if present.You will be shown ...	5
you look for the cat eye marbe and clickk on it if...	5
good	5
Click on a round shape that resembles a cat-eye ma...	3
sometimes it may found in the middle,front of the ...	2
yes	1
I saw instructions more clearly, because its very ...	1
identify cat eye marble	1
nothing	1
click on the animation of the cat-eye marble.	1
good and nice	0
nothing	0
press if you can see cat eye ball	0
you look for the cat eye marbe and clickk on it if...	0
it is perfect as is	0
I can do it clearly	0
nothing	0
Click the marble eye that matches	0
nothing	0
click on the sphere with holes	0
gbvfn vnh gh	0
yes	0
yes	0
Look for a shape that slightly resembles a marble ...	0
Click on the outline of a sphere ignoring all the ...	0
no	0
The objective is not clear.	0
the instructions are more useful	0
Click on the shapes that looks like the cat-eye ma...	0
good	0
yes clear	0
click on the shape(white) that comes through blac...	0
found the cat eye marble in the given animation in...	0
sdvisualimages.com	0
Present to that with images	0
clicking on the cat eye marble . judging in an co...	0
Look at the following images and click on a shape ...	0
not bad	0
Yes not very clear but clear	0
something that looks like a cat eye	0
ok	0
click on the blinking black mark and press next	0
I can do it clearly	0
na	0
No Comments	0
no	0
Look for a spherical shape and click on that.	0
Look for a sphere that has a constant whole in in ...	0
nothing	0
some strange task	0
click on the objects that looks like cat-eye marbl...	0
Yes i can do it.	0
click on cat eye marble	0
SOME CONFUSE	0
the cat eye marble was not present in ani gif so i...	0
yes	0
I can do it clearly	0
no	0
click when you see the animation of cat-eye marble	0
good	0
The following imagesmay or maynot contain shapes I...	0
click on cat eye marble	0
click on fish eye marble	0
click on the round cat eye marble in the animation	0
marble was unsuitable	0
others workers	0
the instructions are not so clear it will quite cl...	0
Shadows of a cat-eye marble will appear in monochr...	0
The image of the actual marble and the graphical i...	0
instructions are very clear	0
show better and more clear examples	0
instructions are very clear	0
yes...	0
Click on the black image that looks most like a ca...	0
Click on the position where it seems like an eye-m...	0

Results from Experiment 3 – Dynamic Questions


+You Search Images Maps Play YouTube News Gmail More ▾
davidcalit2computer@gmail.com ▾
Switch to new look
Get link
Share
File View Edit Visualize Merge Experiment
options

Dynamicquestions

Showing **all rows** options

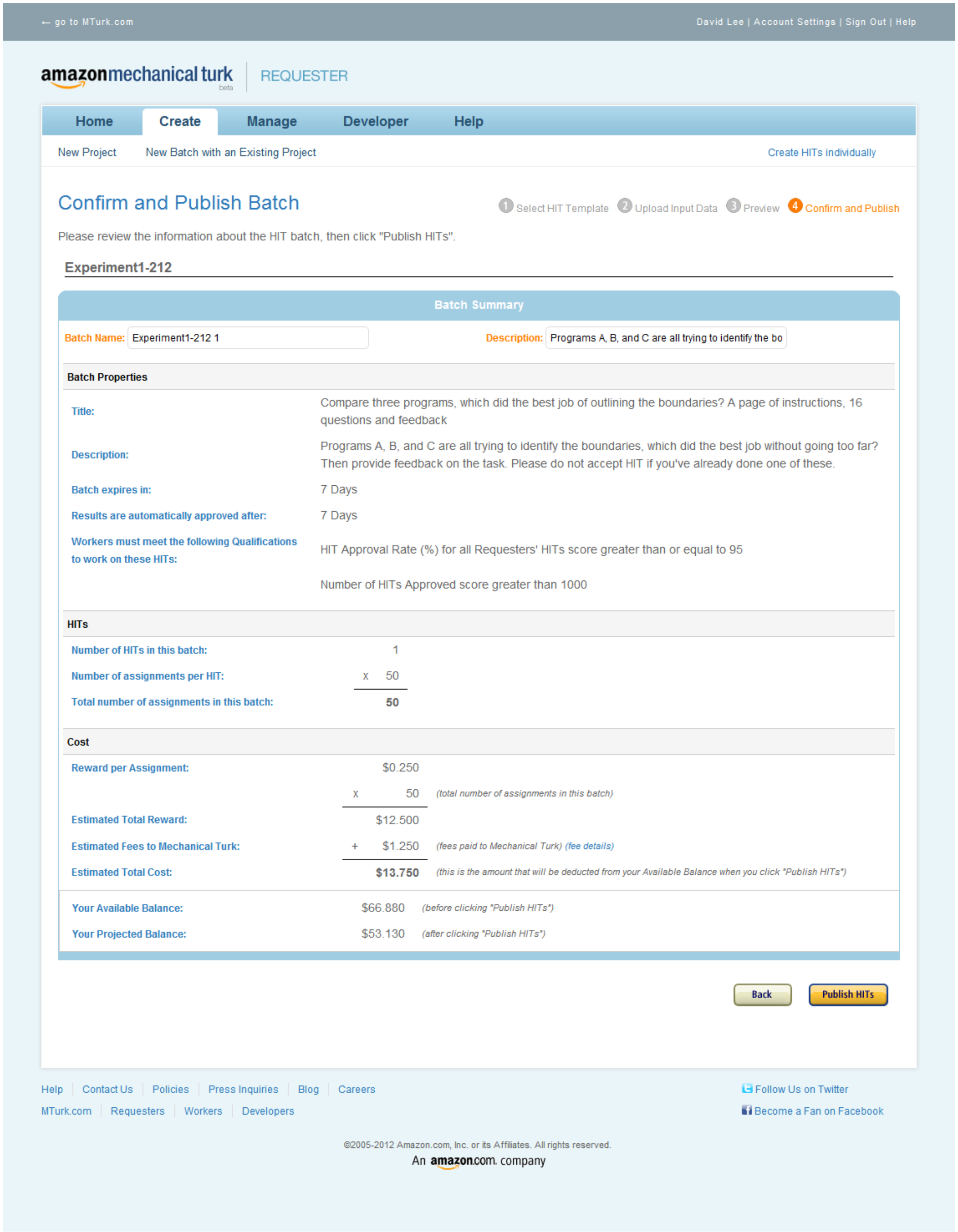
1 - 68 of 68

Text ▾	Votes ▾
The driver drove 91 miles at a fuel consumption ra...	50
The driver has a car that goes 50mpg. He traveled ...	39
Figure out the fuel consumed by this driver	15
The driver has a new car and can't figure out how ...	5
The driver drove 91 miles at a fuel consumption ra...	0
How much fuel was used to drive 91 miles, assuming...	0
How much gas do you need to drive 91 miles at a ra...	0
A driver drove a car with a fuel consumption rate ...	0
What is the quantity of fuel required to drive 91 ...	0
The display below shows the results of a trip com...	0
How much fuel is consumed to drive 91 miles at 50m...	0
this topic is about fuel consumption	0
The driver drove 91 miles at a fuel consumption ra...	0
no	0
The graph charted the fuel consumption of the driv...	0
The driver can drive 50 miles per 1 gallon. For dr...	0
yes	0
yes	0
yes	0
Calculate the Fuel consumed?	0
Estimate the fuel consumed?	0
What amount of fuel was used to go 91 miles in a v...	0
yes	0
The driver drove the car for 91 Miles. If the fuel...	0
what is the fuel consumption for 91 miles?	0
the instruction is clear and correct	0
Suppose that engineers design a car that can trave...	0
yes need more clearly.	0
How much fuel did a car use if it traveled 91 mile...	0
David Lee	0
The instructions were clear enough	0
The driver sped along for 91 miles using a vehicle...	0
A driver drives the vehicle at the consumption rat...	0
Around 8 litres of fuel used.According to my assum...	0
yes need more clearly.	0
How much fuel, in gallons, does an automobile cons...	0
What would be the consumption of fuel, If a driver...	0
How much fuel is consumed when a driver drove a ca...	0
N/A	0
1g fuel is used for 50 miles to go, how much fu...	0
should explain how many miles for 1 litre of petro...	0
Stae below houw much fuel used by drive?	0
NA	0
The drive drove a vehicle for 91miles with a fuel ...	0
The instructions are phrased clearly already.	0
The driver drove 91 miles at a fuel consumption ra...	0
How much fuel did the driver consume to cover the ...	0
NO I CAN UNDERSTAND.	0
The driver drove 91 miles at a fuel-consumption ra...	0
the drive drove with 50mpg and reached 91 miles.th...	0
Nothing	0
The fuel consumption rate is 50 mile per gallon an...	0
Find the fuel used by a car driver if he drove 91 ...	0
The driver drove a car which consumes fuel at the ...	0
If a driver drives 91 miles at 50mpg, how much fue...	0
Find out the fuel consumption?	0
question is right	0
Please calculate the amount of fuel used when trav...	0
The consumption is 1 gallon per 50 miles. How much...	0
The driver drove 91 miles at a fuel consumption ra...	0
You have to provide how much fuel he can consume p...	0
It is as clear as it gets.	0
The instructions were straight forward. 91 miles ...	0
The driver drove 91 miles at the rate of 50mpg.How...	0
mileage and fuel can be used to show in a meter.	0
The driver drove 91 miles at the rate of 50mpg.How...	0
Please calculate the amount of fuel used when trav...	0
91 miles means 146.45 kilometers. 50 mpg means ...	0

APPENDIX F

mTurk Configurations

Configuration from Experiment 1 – Distributing expertise



Confirm and Publish Batch

Please review the information about the HIT batch, then click "Publish HITs".

Experiment1-212

Batch Summary

Batch Name: Experiment1-212 1

Description: Programs A, B, and C are all trying to identify the bo

Batch Properties

Title:	Compare three programs, which did the best job of outlining the boundaries? A page of instructions, 16 questions and feedback
Description:	Programs A, B, and C are all trying to identify the boundaries, which did the best job without going too far? Then provide feedback on the task. Please do not accept HIT if you've already done one of these.
Batch expires in:	7 Days
Results are automatically approved after:	7 Days
Workers must meet the following Qualifications to work on these HITs:	HIT Approval Rate (%) for all Requesters' HITs score greater than or equal to 95
	Number of HITs Approved score greater than 1000

HITs

Number of HITs in this batch:	1
Number of assignments per HIT:	x 50
Total number of assignments in this batch:	**50**

Cost

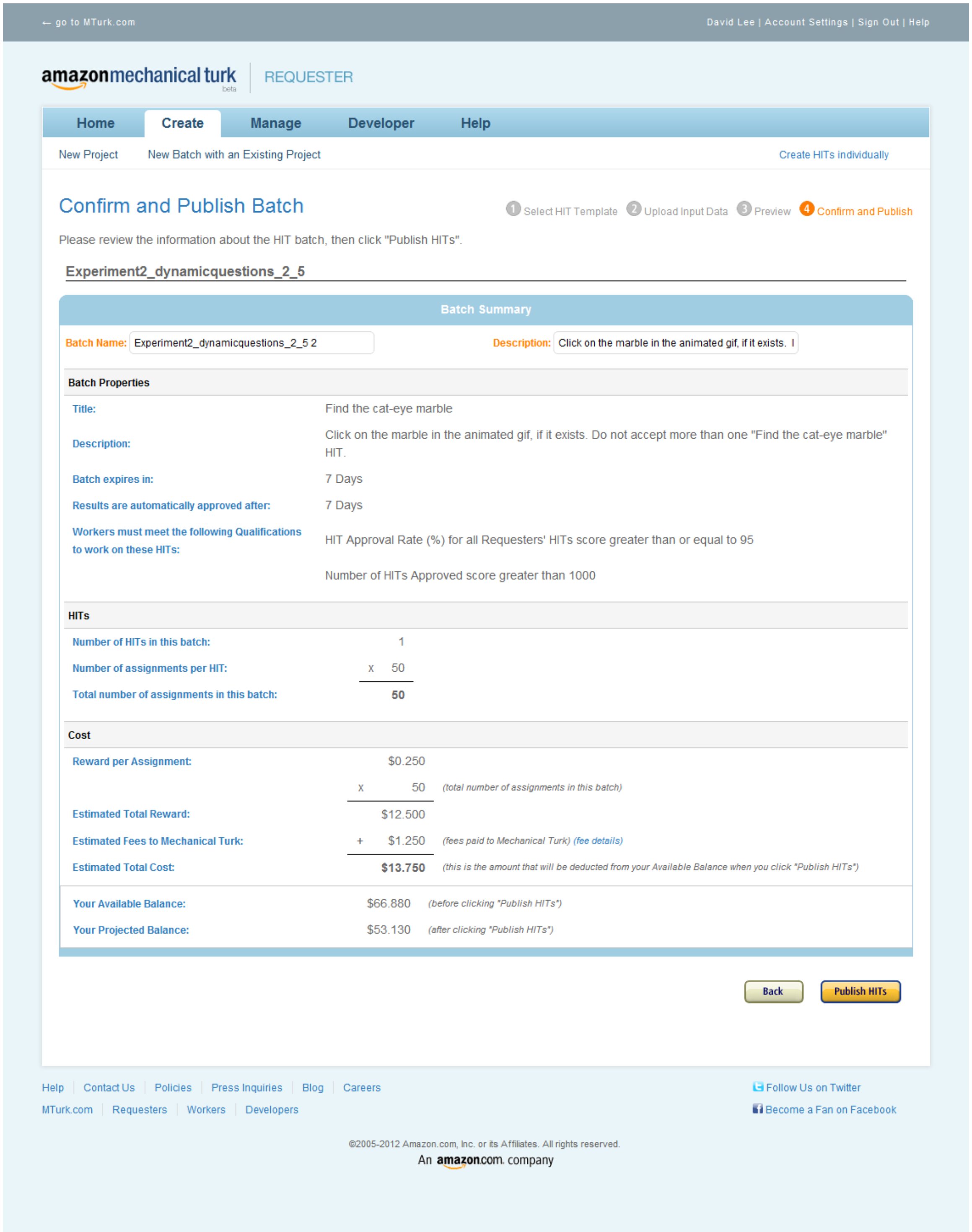
Reward per Assignment:	$0.250	
	x 50	*(total number of assignments in this batch)*
Estimated Total Reward:	$12.500	
Estimated Fees to Mechanical Turk:	+ $1.250	*(fees paid to Mechanical Turk) (fee details)*
Estimated Total Cost:	**$13.750**	*(this is the amount that will be deducted from your Available Balance when you click "Publish HITs")*

Your Available Balance:	$66.880	*(before clicking "Publish HITs")*
Your Projected Balance:	$53.130	*(after clicking "Publish HITs")*

Back
Publish HITs
Help | Contact Us | Policies | Press Inquiries | Blog | Careers
MTurk.com | Requesters | Workers | Developers
Follow Us on Twitter
Become a Fan on Facebook

©2005-2012 Amazon.com, Inc. or its Affiliates. All rights reserved.
An amazon.com. company

Configuration from experiment 2 – Task transformation


← go to MTurk.com
David Lee | Account Settings | Sign Out | Help
amazonmechanical turk beta
REQUESTER
Home
Create
Manage
Developer
Help
New Project
New Batch with an Existing Project
Create HITs individually

Confirm and Publish Batch

Please review the information about the HIT batch, then click "Publish HITs".

Experiment2_dynamicquestions_2_5

Batch Summary

Batch Name: Experiment2_dynamicquestions_2_5 2

Description: Click on the marble in the animated gif, if it exists. I

Batch Properties

Title:	Find the cat-eye marble
Description:	Click on the marble in the animated gif, if it exists. Do not accept more than one "Find the cat-eye marble" HIT.
Batch expires in:	7 Days
Results are automatically approved after:	7 Days
Workers must meet the following Qualifications to work on these HITs:	HIT Approval Rate (%) for all Requesters' HITs score greater than or equal to 95 Number of HITs Approved score greater than 1000

HITs

Number of HITs in this batch:	1
Number of assignments per HIT:	x 50
Total number of assignments in this batch:	**50**

Cost

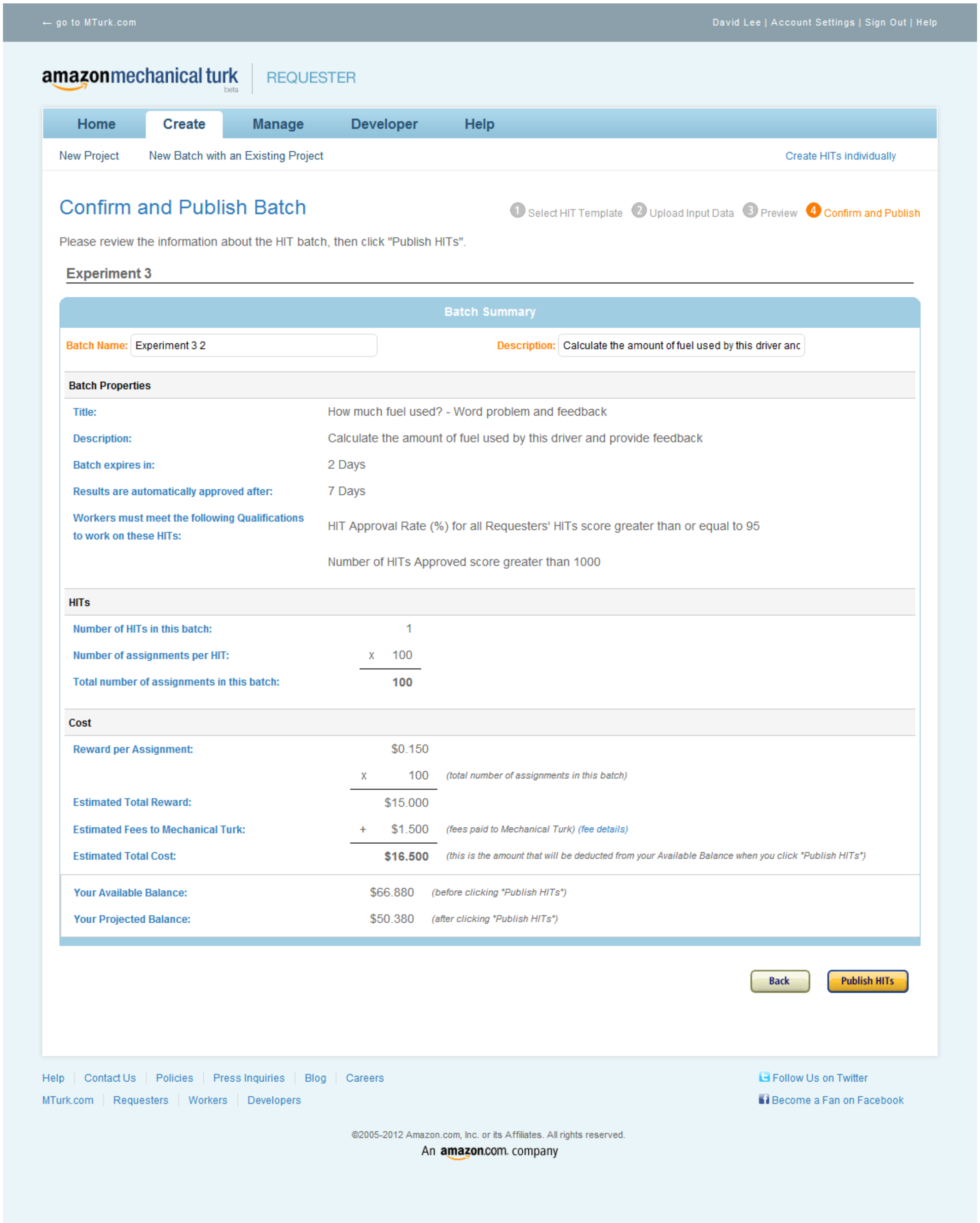
Reward per Assignment:	$0.250	
	x 50	*(total number of assignments in this batch)*
Estimated Total Reward:	$12.500	
Estimated Fees to Mechanical Turk:	+ $1.250	*(fees paid to Mechanical Turk) (fee details)*
Estimated Total Cost:	**$13.750**	*(this is the amount that will be deducted from your Available Balance when you click "Publish HITs")*
Your Available Balance:	$66.880	*(before clicking "Publish HITs")*
Your Projected Balance:	$53.130	*(after clicking "Publish HITs")*

Back
Publish HITs
Help | Contact Us | Policies | Press Inquiries | Blog | Careers
MTurk.com | Requesters | Workers | Developers
Follow Us on Twitter
Become a Fan on Facebook

©2005-2012 Amazon.com, Inc. or its Affiliates. All rights reserved.

An amazon.com. company

Configuration from experiment 3 – Dynamic questions



Confirm and Publish Batch

Please review the information about the HIT batch, then click "Publish HITs".

Experiment 3

Batch Summary

Batch Name: Experiment 3 2

Description: Calculate the amount of fuel used by this driver anc

Batch Properties

Title:	How much fuel used? - Word problem and feedback
Description:	Calculate the amount of fuel used by this driver and provide feedback
Batch expires in:	2 Days
Results are automatically approved after:	7 Days
Workers must meet the following Qualifications to work on these HITs:	HIT Approval Rate (%) for all Requesters' HITs score greater than or equal to 95 Number of HITs Approved score greater than 1000

HITs

Number of HITs in this batch:		1
Number of assignments per HIT:	x	100
Total number of assignments in this batch:		**100**

Cost

Reward per Assignment:		$0.150	
	x	100	*(total number of assignments in this batch)*
Estimated Total Reward:		$15.000	
Estimated Fees to Mechanical Turk:	+	$1.500	*(fees paid to Mechanical Turk) (fee details)*
Estimated Total Cost:		**$16.500**	*(this is the amount that will be deducted from your Available Balance when you click "Publish HITs")*

Your Available Balance:	$66.880	*(before clicking "Publish HITs")*
Your Projected Balance:	$50.380	*(after clicking "Publish HITs")*

Back
Publish HITs
Help | Contact Us | Policies | Press Inquiries | Blog | Careers
MTurk.com | Requesters | Workers | Developers
Follow Us on Twitter
Become a Fan on Facebook
©2005-2012 Amazon.com, Inc. or its Affiliates. All rights reserved.
An amazon.com. company

Appendix G

Mechanical Turk data

Online hosting of results from Mechanical Turk

Amazon’s Mechanical Turk generates significant amount of information from every HIT result including WorkerIDs, assignementIDs, IP addresses, and answers to all of the individual questions. The results from all of the experiments conducted on Mechanical Turk are included in an online archive found at the following web address:

https://s3.amazonaws.com/dlee_dissertation/Mechanical_Turk_results_dlee_dissertation.zip

These results are stored in CSV or comma-separated values. This format can be easily imported into word processing, spreadsheet, and statistical applications.

Bibliography

(2013). "Crowdflower: The World's Largest Workforce." Retrieved 2-22, 2013, from http://crowdflower.com/.

Abrmmoff, M. D., P. J. Magalhaes, et al. (2004). "Image processing with ImageJ." Biophotonics international **11**(7): 36-42.

Argote, L. (1999). Organizational learning: Creating, retaining, and transferring knowledge, Springer Netherlands.

Bernstein, M. S., G. Little, et al. (2010). Soylent: a word processor with a crowd inside. UIST '10 Proceedings of the 23nd annual ACM symposium on User interface software and technology, ACM.

Borne, K. and Z. Team (2011). The Zooniverse: A Framework for Knowledge Discovery from Citizen Science Data. AGU Fall Meeting Abstracts.

Buckland, M. K. (1991). "Information as thing." Journal of the American Society for information science **42**(5): 351-360.

Cifelli, R., N. Doesken, et al. (2005). "The community collaborative rain, hail, and snow network." Bulletin on the American Meteorological Society **86**: 1069-1077.

Clark, H. H. (1996). Using language, Cambridge University Press Cambridge.

Cooper, S., F. Khatib, et al. (2010). "Predicting protein structures with a multiplayer online game." Nature **466**(7307): 756-760.

Deerinck, T., Bushong, E., Lev-Ram, V., Shu, X., Tsien, R., and M. H. Ellisman (2010). Enhancing Serial Block-Face Scanning Electron Microscopy to Enable High Resolution 3-D Nanohistology of Cells and Tissues. Microscopy and Microanalysis.

Galton, F. (1907). "Vox populi." Nature **75**: 450-451.

Giles, J. (2005). "Internet encyclopaedias go head to head." Nature **438**(7070): 900-901.

Giuly, R. J., K.-Y. Kim, et al. (2013). "DP2: Distributed 3D image segmentation using micro-labor workforce." Bioinformatics **29**(10): 1359-1360.

Giuly, R. J., M. E. Martone, et al. (2012). "Method: Automatic segmentation of mitochondria utilizing patch classification, contour pair classification, and automatically seeded level sets." BMC bioinformatics **13**(1): 29.

Ipeirotis, P. (2010). "Demographics of mechanical turk."

Ipeirotis, P. G., F. Provost, et al. (2010). Quality management on Amazon Mechanical Turk. Proceedings of the ACM SIGKDD Workshop on Human Computation. Washington DC, ACM**:** 64-67.

Irani, L. and M. Silberman (2013). Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. Proceeding of the Annual ACM SIGCHI Conference on Human Factors in Computing Systems.

Jurrus, E., M. Hardy, et al. (2009). "Axon tracking in serial block-face scanning electron microscopy." Medical image analysis **13**(1): 180-188.

Jurrus, E., A. R. Paiva, et al. (2010). "Detection of neuron membranes in electron microscopy images using a serial neural network architecture." Medical image analysis **14**(6): 770-783.

Kittur, A., E. Chi, et al. (2007). "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie." World Wide Web **1**(2): 19.

Kittur, A., E. H. Chi, et al. (2008). Crowdsourcing user studies with Mechanical Turk. CHI '08 Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM.

Kremer, J. R., D. N. Mastronarde, et al. (1996). "Computer visualization of three-dimensional image data using IMOD." Journal of Structural Biology **116**(1): 71-76.

Larkin, J., J. McDermott, et al. (1980). "Expert and novice performance in solving physics problems." Science **208**(4450): 1335-1342.

Lee, P., A. Kapelner, et al. (2009). "An Interactive Java Statistical Image Segmentation System: GemIdent." Journal of Statistical Software **30**(10).

Lenzi, D., J. W. Runyeon, et al. (1999). "Synaptic vesicle populations in saccular hair cells reconstructed by electron tomography." The Journal of neuroscience **19**(1): 119-132.

Lin, A. Y. M. (2010). The search for Genghis Khan: Using modern tools to hunt for an ancient past. Aerospace Conference, 2010 IEEE.

Lintott, C., K. Schawinski, et al. (2008). "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey." Monthly Notices of the Royal Astronomical Society **389**(3): 1179-1189.

Little, G., L. B. Chilton, et al. (2009). Turkit: Tools for iterative tasks on mechanical turk. HCOMP '09 Proceedings of the ACM SIGKDD Workshop on Human Computation, ACM.

Ludascher, B., I. Altintas, et al. (2006). "Scientific workflow management and the Kepler system." Concurrency and Computation: Practice and Experience **18**(10): 1039-1065.

Martin, D., C. Fowlkes, et al. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, IEEE.

Martone, M. E., A. Gupta, et al. (2002). "A cell-centered database for electron tomographic data." Journal of Structural Biology **138**(1-2): 145-155.

Milazzo, A. C., J. Lanman, et al. (2009). "Advanced Detector Development for Electron Microscopy Enables New Insight into the Study of the Virus Life Cycle in Cells and Alzheimers Disease." Microscopy and Microanalysis **15**(S2): 8-9.

Mishchenko, Y. (2009). "Automation of 3D reconstruction of neural tissue from large volume of conventional serial section transmission electron micrographs." Journal of Neuroscience Methods **176**(2): 276-289.

Norman, D. A. (2004). "Design as Communication." Retrieved 9/3/2012, 2012, from http://www.jnd.org/dn.mss/design_as_communicat.html.

Noske, A. B., A. J. Costin, et al. (2008). "Expedited approaches to whole cell electron tomography and organelle mark-up in situ in high-pressure frozen pancreatic islets." Journal of structural biology **161**(3): 298-313.

Perkins, G., C. Renken, et al. (1997). "Electron Tomography of Neuronal Mitochondria: Three-Dimensional Structure and Organization of Cristae and Membrane Contacts." Journal of Structural Biology **119**(3): 260-272.

Perkins, G. A., M. H. Ellisman, et al. (2003). "Three-dimensional analysis of mouse rod and cone mitochondrial cristae architecture: bioenergetic and functional implications." Mol Vis **9**: 60-73.

Perkins, G. A., M. G. Sun, et al. (2009). "Correlated light and electron microscopy/electron tomography of mitochondria in situ." Methods in enzymology **456**: 29-52.

Price, D. L., S. K. Chow, et al. (2006). "High-resolution large-scale mosaic imaging using multiphoton microscopy to characterize transgenic mouse models of human neurological disorders." Neuroinformatics **4**(1): 65-80.

Rochlin, G. I., T. R. La Porte, et al. (1987). "The self-designing high-reliability organization: Aircraft carrier flight operations at sea." Naval War College Review **40**(4): 76-90.

Ross, J., L. Irani, et al. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. CHI'10 Extended Abstracts on Human Factors in Computing Systems, ACM.

Schaller, R. R. (1997). "Moore's law: past, present and future." Spectrum, IEEE **34**(6): 52-59.

Seung, S. (2013). "Eyewire." from http://www.eyewire.org.

Shu, X., V. Lev-Ram, et al. (2011). "A Genetically Encoded Tag for Correlated Light and Electron Microscopy of Intact Cells, Tissues, and Organisms." PLoS Biology **9**(4): e1001041.

Silberman, M., L. Irani, et al. (2010). "Ethics and tactics of professional crowdwork." XRDS: Crossroads, The ACM Magazine for Students **17**(2): 39-43.

Singh, R., N. Schwarz, et al. (2006). "Real-time multi-scale brain data acquisition, assembly, and analysis using an end-to-end OptIPuter." Future Generation Computer Systems **22**(8): 1032-1039.

Sosinsky, G. E., T. J. Deerinck, et al. (2005). "Development of a model for microphysiological simulations." Neuroinformatics **3**(2): 133-162.

Stalling, D., M. Westerhoff, et al. (2005). "Amira: A highly interactive system for visual data analysis." The Visualization Handbook **38**: 749-767.

Star, S. L. and J. R. Griesemer (1989). "Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." Social studies of science **19**(3): 387-420.

Stiny, G. (1980). "Introduction to shape and shape grammars." Environment and planning B **7**(3): 343-351.

Sullivan, B. L., C. L. Wood, et al. (2009). "eBird: A citizen-based bird observation network in the biological sciences." Biological Conservation **142**(10): 2282-2292.

Surowiecki, J. and M. P. Silverman (2007). "The wisdom of crowds." American Journal of Physics **75**: 190.

Von Ahn, L. (2006). "Games with a purpose." Computer **39**(6): 92-94.

Westphal, A. J., A. L. Butterworth, et al. (2005). Stardust@ home: A Massively Distributed Public Search for Interstellar Dust in the Stardust Interstellar Dust Collector. 36sth Annual Lunar and Planetary Science Conference, League City, Texas.

Wiseman, P., J. Squier, et al. (2000). "Two photon image correlation spectroscopy and image cross correlation spectroscopy." Journal of microscopy **200**(1): 14-25.

Wright, J. (2010). "Did Spanish Astronomers Steal a Dwarf Planet?". Retrieved 8/7/12, 2012, from http://jameswight.wordpress.com/2010/03/04/did-spanish-astronomers-steal-planet/.