Abstract of "A computational approach to mitigate visualization design barriers"
by Connor C. Gramazio, Ph.D., Brown University, May 2017.

This thesis advances visualization design research by developing and evaluating new theoretical knowledge and computational techniques, which target the rising complexity of data and growing diversity of visualization users. To ground our research, we focus our study on common design limitations that are found in cancer genomics, which is an exemplar of how research at-large is affected by the rising ubiquity and democratization of visualization in analysis.

We first identify four cancer genomics task requirements for visual analysis through interviews and evaluate whether the multiple visualizations in MAGI – a cancer genomics visualization tool – can support such diversity. Second, we evaluate how simple classifiers trained on annotated mouse interaction logs can help designers understand how domain experts use visualizations. Third, we explore the ways in which the size and perceptual grouping of data in visualization can affect visual search performance and visual analysis tasks. Last, we discuss a novel tool for creating categorical color palettes based on user-defined importances of discriminability and aesthetic preference, which can be a common and difficult task in visualization design independent of application area.

These contributions may help mitigate visualization design barriers by providing guidelines and techniques to help visualization creators avoid common pitfalls. For example, our evaluation of Colorgorical demonstrates that the tool can automatically generate color palettes based on user defined balances of discriminability and preference, which are comparably discriminable and typically more preferable compared to industry standards. Colorgorical thus provides an effective alternative to making categorical palettes by hand, which can be time consuming and require design expertise.

While the contributions in this thesis are grounded in cancer genomics, our contributions are not limited in application: Many may generalize to other domains, such as using domain expert interaction log analysis to better understand how visualization is used by different kinds of researchers in the brain sciences. Given the visualization design research similarities between cancer genomics and other domain expert centric applications like brain science, we conclude by hypothesizing how our findings could be used to investigate open research areas.

A computational approach to mitigate visualization design barriers

by

Connor C. Gramazio
B. S., Tufts University, 2012
Sc.M., Brown University, 2014

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

Providence, Rhode Island
May 2017

This dissertation by Connor C. Gramazio is accepted in its present form by the Department of Computer Science as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date ______________ ______________________________
David H. Laidlaw, Advisor

Recommended to the Graduate Council

Date ______________ ______________________________
Jeff Huang, Reader

Date ______________ ______________________________
Benjamin J. Raphael, Reader
(Princeton University, Computer Science)

Date ______________ ______________________________
Karen B. Schloss, Reader
(University of Wisconsin-Madison, Psychology)

Approved by the Graduate Council

Date ______________ ______________________________
Andrew G. Campbell
Dean of the Graduate School

Vita

Connor C. Gramazio was born in Massachusetts, USA in June 1990. He attended Tufts University in Somerville, Massachusetts, graduating *cum laude* with highest thesis honors in 2012 with a B.S. in Computer Science and a minor in Religion. His honors thesis work and other undergraduate research led to a successful NSF Graduate Research Fellowship application, which later funded much of his graduate career. He was co-advised by Professors Ben Hescott and Remco Chang for his major, and was advised by Professor Joseph G. Walser for his minor. While at Tufts he was also a member of Kiniwe for four years, where he performed traditional Ghanaian folk dance-drumming-song compositions under the directorship of Professors Nani Agbeli and David Locke. He earned his Master's degree in Computer Science from Brown University in 2014, where he studied visualization under the supervision of Professor David H. Laidlaw.

Education

PhD, Computer Science, Brown University, 2012–2017.

ScM, Computer Science, Brown University, 2012–2014.

BS, Computer Science, Tufts University, 2008–2012, *cum laude* with highest thesis honors.

Peer-Reviewed Conference and Journal Publications

10. C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *Transactions on Visualization and Computer Graphics (Proc. VIS '16)*, 2017. ISSN 1077-2626. doi: 10.1109/TVCG.2016.2598918.

URL `http://vrl.cs.brown.edu/color`

9. C. C. Gramazio, J. Huang, and D. Laidlaw. An analysis of visual analysis: Modeling the interactive visualization tasks of cancer genomics domain experts. *Transactions on Visualization and Computer Graphics*, In Revision

8. C. C. Gramazio, M. Leiserson, B. Raphael, and D. Laidlaw. A cancer genomics visualization task requirements analysis and design study of magi. *Transactions on Visualization and Computer Graphics*, Under Review

7. M. D. Leiserson, C. C. Gramazio, J. Hu, H.-T. Wu, D. H. Laidlaw, and B. J. Raphael. Magi: visualization and collaborative annotation of genomic aberrations. *Nature Methods*, 12(6): 483–484, 06 2015. URL `http://magi.brown.edu`

6. S. Li, R. J. Crouser, G. Griffin, C. C. Gramazio, H.-J. Schulz, H. Childs, and R. Chang. Exploring hierarchical visualization designs using phylogenetic trees. *Proc. SPIE*, 9397:939709–939709–14, 2015. doi: 10.1117/12.2078857. URL `http://dx.doi.org/10.1117/12.2078857`

5. A. Papoutsaki, H. Guo, D. Metaxa-Kakavouli, C. C. Gramazio, J. Rasley, W. Xie, G. Wang, and J. Huang. Crowdsourcing from scratch: A pragmatic experiment in data collection by novice requesters. *Proceedings of The AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2015

4. C. C. Gramazio, K. B. Schloss, and D. H. Laidlaw. The relation between visualization size, grouping, and user performance. *Transactions on Visualization and Computer Graphics (Proc. VIS '14)*, 20(12):1953–1962, Dec 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346983. URL `http://dx.doi.org/10.1109/TVCG.2014.2346983`

3. S. Kelley, E. Aftandilian, C. Gramazio, N. Ricci, S. L. Su, and S. Z. Guyer. Heapviz: Interactive heap visualization for program understanding and debugging (extended). *Information Visualization*, 12(2):163–177, 2013. doi: 10.1177/1473871612438786. URL `http://ivi.sagepub.com/content/12/2/163.abstract`

2. S. Su, C. Gramazio, D. Extrum-Fernandez, C. Crumm, L. Cowen, M. Menke, and M. Strait. Molli: Interactive visualization for exploratory protein analysis. *Computer Graphics and Applications, IEEE*, 32(5):62–69, Sept 2012. ISSN 0272-1716. doi: 10.1109/MCG.2012.66

1. E. E. Aftandilian, S. Kelley, C. Gramazio, N. Ricci, S. L. Su, and S. Z. Guyer. Heapviz: Interactive heap visualization for program understanding and debugging. *Proceedings of the 5th International Symposium on Software Visualization*, pages 53–62, 2010. doi: 10.1145/1879211.1879222. URL `http://doi.acm.org/10.1145/1879211.1879222`

Peer-Reviewed Conference Abstracts and Posters

9. A. Silverman, C. C. Gramazio, and K. B. Schloss. The dark is more (dark+) bias in colormap data visualizations with legends. In *Journal of Vision/VSS*, 2016. doi: 10.1167/16.12.628. URL `http://dx.doi.org/10.1167/16.12.628`

8. K. B. Schloss, C. C. Gramazio, and C. Walmsley. Which color means more? an investigation of color-quantity mapping in data visualization. In *Journal of Vision/VSS*, volume 15, page 1317, September 2015. doi: 10.1167/15.12.1317. URL `http://dx.doi.org/10.1167/15.12.1317`

7. M. Leiserson, H.-T. Wu, C. Gramazio, and B. Raphael. Magi: A platform for interactive visualization and collaborative annotation of combinations of genetic aberrations. In *The 1st Biological Data Science Meeting*, 2014

6. M. D. Leiserson, H.-T. Wu, C. C. Gramazio, and B. J. Raphael. Cancer genome analysis tool (cgat) for the visualization and exploration of combinations of mutations in cancer. In *The 4th Annual The Cancer Genome Atlas Symposium*, 2014

5. C. Gramazio and R. Chang. Optimizing an spt-tree for visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012

4. G. Griffin, S. Li, C. Gramazio, and R. Chang. An analytical approach for the creative design of new visualizations. In *IEEE Conference on Information Visualization (InfoVis)*, 2011

3. M. Strait, C. Gramazio, J. Park, S. L. Su, and L. Cowen. Moleint: Reducing workload through adaptive interaction. In *VIZBI*, 2012

2. E. E. Aftandilian, S. Kelley, C. Gramazio, N. Ricci, S. L. Su, and S. Z. Guyer. Heapviz: A programmer's tool for data structure visualization. In *IEEE VisWeek Demo*, 2010

1. M. Matt, S. L. Su, C. Gramazio, C. Crumm, D. Extrum-Fernandez, and L. Cowen. Tuftsviewer: An intuitive interface for viewing 3d protein structures. In *3DSIG Workshop at ISMB*, 2010

Select Fellowships, Awards, and Honors

Invited Speaker, Open Visualization Conference, 2017

Brown Dissertation Fellowship, 2016–17

NSF Graduate Research Fellowship, 2012–16

Sheridan Center Teaching Certificate, 2016

Andries van Dam Graduate Fellowship, 2012–13

Runner Up Computing Research Association Outstanding Undergrad Research Award, 2011

Experience

Graduate researcher, Brown University, 2012–2017

Software Engineer Intern, Google, 2012

Undergraduate researcher and teaching assistant, Tufts University, 2009–2012

Scientist/Software Engineer Intern, Charles River Analytics, 2011

Acknowledgements

This work was supported by the National Science Foundation (DGE-1058262) and Brown University's Center for Vision Research. All opinions, findings, conclusions, or recommendations expressed in this document are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

I would like to foremost thank David Laidlaw, my advisor, for encouraging me to tackle important open research questions that not only make strong visualization contributions, but also provide benefit to end-users downstream. I also thank my committee for their support and guidance as I stepped into the unknown: Jeff Huang, Ben Raphael, and Karen Schloss. I would also like to thank my undergraduate research mentors: Remco Chang, Ben Hescott, and Sara Su.

My thesis would not have been possible without the help and support of the Visualization Research Lab. In particular, I want to acknowledge Hua Guo, Ryan Cabeen, and Steve Gomez. I also want to thank Max Leiserson and Hsin-Ta Wu for their help as I ventured into cancer genomics visualization, as well as Lane Harrison who has been a bottomless source of guidance over the past seven years.

I'm thankful to have performed this research in a department that goes above and beyond to create a positive environment for graduate students. Thanks to Lauren Clarke, Dawn Reed, Genie DeGouveia, and Ugur Çentintemel for making this possible, as well as my friends and peers in the department: Thank you Jeff, Layla, Betsy, Michael, John, Hannah, Emmanuel, Philip, and many others.

I'd also like to thank the rest of my friends and family. And, most of all, I'd like to thank Denise for her unceasing help and encouragement throughout this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Problem Statement

Thesis Statement: We hypothesize that visualization design can be broadly empowered and improved through the creation of computational design assistance tools based on new theoretical knowledge of graphical perception and task requirements.

Visualization is a critical component of how humans understand and synthesize information from raw data. Whether it is looking at galaxies light years away or at the results of human genome sequencing, visualization is often the lens through which we make and communicate scientific discovery.

Historically, visualization researchers were the gatekeepers and curators for these lenses of discovery, and could leverage years of expertise to provide scientists with effective tools. Now, creating complex visualizations is no longer restricted to experts with years of tool creation experience, but is instead open to anyone with enough skill to use Tableau or write a few lines of code thanks to libraries like VegaLite [164]. Given the number of powerful visualization authoring tools and prevalence of design inexpertise, it is likely that visualization creators may unintentionally create ineffective or misleading graphics. Similarly, it has become harder to understand the requirements of increasingly diverse tool user populations who may view and use the same visualization for a multitude of purposes.

This dissertation is motivated by these tool design challenges, and provides new knowledge as well as new techniques to help understand and improve the myriad design issues that are present in today's visualization landscape. We ground our research by focusing on design limitations within cancer genomics, which is an exemplar of the kinds of problems that stem from rising ubiquity of visualization in research at-large (e.g., lack of formal design training). By doing so, we consider research through the perspective of *toolsmiths*, which holds that each advancement in our understanding of computer science and visualization should likewise benefit collaborators and tool users [18].

We actualized this research vision through interdisciplinary collaboration such that each of the following chapters are products of partnership with experts from computational biology, vision science, and human-computer interaction. Our interdisciplinary contributions specifically focus on two complementary veins of visualization design research. First, we study how visualization can be used for analysis over heterogeneous and multidimensional data across a variety of research expertise. Grounding our work with an evaluation of a cancer genomics visualization tool called MAGI (Chapter 2.3), we characterize cancer genomics task requirements that span a diversity of research sub-specializations (Chapter 3) and then assess how automated interaction log analysis can help tool designers infer how domain experts naturalistically use tools "in the wild" (Chapter 4). Second, we consider how graphical perception limitations alter tool design effectiveness. We connect our visualization task research to graphical perception limitations with a study of how various types of visualization size can affect search performance and visualization task associations (Chapter 5). Within this vein we then investigate how vision science principles can be applied in a way to automate categorical color palette design for information visualization (Chapter 6). Although these two visualization design research contribution areas are mostly framed in terms of outcomes from our study of MAGI, they are often equally applicable to other visualization tools irrespective of application domain.

Our thesis contributions – encompassing theory, technique, and evaluation alike – provide templates and a foundation for how research into automated visualization design can continue into open areas of visualization design research. In the conclusion, we overview several hypotheses about how continued research in this space could help empower current design experts and lower design personalization barriers to those who would normally have too little expertise (Chapter 7).

1.2 Contributions and Thesis Organization

This thesis is organized into seven chapters including the introduction, background, conclusion, and four chapters that pertain to our primary contributions. We introduce and summarize the contributions for each of these four chapters below. Recall that MAGI is a cancer genomics visualization tool fully described in Chapter 2.3.

Chapter 3: An evaluation of cancer genomics visual analysis with MAGI

In this chapter, we present results from an evaluation that tested to what degree MAGI, a cancer genomics visual analysis tool, supports a range of cancer genomics research needs. Our evaluation of MAGI is grounded on a foundational task requirements analysis derived from interviews with cancer genomics researchers across a diversity of specializations and occupations. We then report how MAGI supports both community-wide and researcher-specific task requirements using results from three in-depth MAGI case studies. We also discuss how MAGI's design helped participants gain new research insights, and how we discovered several shortcoming that led to the continued iterative design of MAGI. Using these findings and existing theoretical frameworks, we suggest that MAGI mostly supported participants' requirements in part because of its multiple-view-based design, which shows many orientations of genomics data simultaneously. We also examine how these findings may apply to other related genomics visualization tools through a design feature comparison We conclude our study with a discussion of how our in-lab results generalize to real-world research settings by examining interaction logs collected from online use of MAGI.

> C.C. Gramazio, M.D.M. Leiserson, B.J. Raphael, D.H. Laidlaw. "An evaluation of cancer genomics visual analysis with MAGI." *In review.*

Chapter 4: An Analysis of Visual Analysis: Modeling the Interactive Visualization Tasks of Cancer Genomics Domain Experts

In this chapter, we show how mouse interaction log classification can help visualization toolsmiths identify visual analysis tasks through an evaluation of MAGI. Our primary contribution is an evaluation of twelve visual analysis task classifiers, which compares predictions to task inferences made by pairs of genomics and visualization experts. Our evaluation uses common models that are accessible to most visualization evaluators: *k*-nearest neighbors, linear support vector machines, and random forests. By comparing classifier predictions to visual analysis task inferences made by experts, we show that simple automated task classification can have up to 73% accuracy and can separate meaningful logs from "junk" logs with up to 91% accuracy. Our second contribution is an exploration of common MAGI interaction trends using the predictive classification results, which expands current knowledge about naturalistic cancer genomics visualization tasks. Our third contribution is a discussion of how automated task classification can inform iterative tool design. As a whole, these contributions suggest that mouse interaction log analysis is a viable method for (1) evaluating task requirements of client-side-focused tools, (2) allowing researchers to study experts on larger scales than is typically possible with in-lab observation, and (3) highlighting potential tool evaluation bias.

> C.C. Gramazio, J. Huang, D.H. Laidlaw. "An Analysis of Visual Analysis: Modeling the Interactive Visualization Tasks of Cancer Genomics Domain Experts." *In review.*

Chapter 5: The relation between visualization size, grouping, and user performance

In this chapter, we make the following contributions: (1) we describe how the grouping, quantity, and size of visual marks affects search time based on the results from two experiments; (2) we report how search performance relates to self-reported difficulty in finding the target for different display types; and (3) we present design guidelines based on our findings to facilitate the design of effective visualizations. Both Experiment 1 and 2 asked participants to search for a unique target in colored visualizations to test how the grouping, quantity, and size of marks affects user performance. In Experiment 1, the target square was embedded in a grid of squares and in Experiment 2 the target was a point in a scatterplot. Search performance was faster when colors were spatially grouped than when they were randomly arranged. The quantity of marks had little effect on search time for grouped displays ("popout"), but increasing the quantity of marks slowed reaction time for random displays. Regardless of color layout (grouped vs. random), response times were slowest for the smallest mark size and decreased as mark size increased to a point, after which response times plateaued. In addition to these two experiments we also include potential application areas, as well as results from a small NASA TLX cognitive workload experiment using a visualization from MAGI where we report preliminary findings that size may affect how users infer how visualizations should be used. We conclude with a list of design guidelines that focus on how to best create visualizations based on grouping, quantity, and size of visual marks.

> C.C. Gramazio, K.B. Schloss, D.H. Laidlaw. "The relation between visualization size, grouping, and user performance." IEEE Transactions on Visualization and Computer Graphics (Proceedings of Information Visualization), 2014.

Chapter 6: Colorgorical:

Creating discriminable and preferable color palettes for information visualization

In this chapter, we present an evaluation of Colorgorical, a web-based tool for creating discriminable and aesthetically preferable categorical color palettes. The motivation for Colorgorical grew out of MAGI development and the difficulty that is posed by visualizing large numbers of cancer datasets that are expressed in MAGI's visualizations through color. Colorgorical uses iterative semi-random sampling to pick colors from CIELAB space based on user-defined discriminability and preference importances. Colors are selected by assigning each a weighted sum score that applies the user-defined importances to Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference scoring functions, which compare a potential sample to already-picked palette colors. After, a color is added to the palette by randomly sampling from the highest scoring palettes. Users can also specify hue ranges or build off their own starting palettes. This procedure differs from previous approaches that do not allow customization (e.g., pre-made ColorBrewer palettes [17]) or do not consider visualization design constraints (e.g., ACE [124] and Adobe Color [132]). In a Palette Score Evaluation, we verified that each scoring function measured different color information. Experiment 1 demonstrated that slider manipulation generates palettes that are consistent with the expected balance of discriminability and aesthetic preference for 3-, 5-, and 8-color palettes, and also shows that the number of colors may change the effectiveness of pair-based discriminability and preference scores. For instance, if the Pair Preference slider were upweighted, users would judge the palettes as more preferable on average. Experiment 2 compared Colorgorical palettes to benchmark palettes (ColorBrewer, Microsoft, Tableau, Random). Colorgorical palettes are as discriminable and are at least as preferable or more preferable than the alternative palette sets. In sum, Colorgorical allows users to make customized color palettes that are, on average, as effective as current industry standards by balancing the importance of discriminability and aesthetic preference.

> C.C. Gramazio, D.H. Laidlaw, K.B. Schloss. "Colorgorical: Creating discriminable and preferable color palettes for information visualization." IEEE Transactions on Visualization and Computer Graphics (Proceedings of Information Visualization), 2016.

1.3 Aim

These four sets of contributions test whether applied visualization design theory can mitigate visualization design barriers. As such, the driving aim of this thesis is two-fold: (1) to expand our present understanding of effective visualization design, and (2) to provide guidelines and techniques that benefit visualization creators. Accordingly, as we discuss the theoretical implications of each contribution in this thesis, we also consider how our findings could be applied into practice. To this end, we conclude with a summary of both primary and secondary contributions, testable hypotheses on how our work could be built upon in the future, as well as the summative thesis takeaways.

Chapter 2

Background and Significance

Here, we provide a foundation for the research contributions detailed in the following chapters. Rather than provide a comprehensive survey of entire fields, this chapter's purpose is to provide an overview of concepts and definitions that are helpful for understanding the significance of our thesis contributions. The first section defines what we mean by "visualization" and related concepts such as "graphical perception." The second section provides a high-level overview of cancer genomics, which serves as a motivation for much of our visualization design research. The third section discusses MAGI, a cancer genomics visualization tool, which also motivates much our thesis contributions. The fourth section is a summary of color science topics that immediately relate to our chapters on graphical perception.

2.1 Visualization

Visualization is the process of rendering data as graphical marks on a computer screen or in other media such as print or as three-dimensional fabrications. It is an inherently interdisciplinary field and incorporates methodologies and techniques from seemingly disparate areas. The focus of this thesis is on visualization design research, which broadly studies how to improve visualization tool usability. In contrast, other areas of visualization research might prioritize the discovery of faster graph drawing algorithms, real-time rendering techniques, or new algorithms to support emerging medical imaging technologies. This thesis approaches visualization design research in two related sub-fields: *graphical perception* and *task requirement analysis.*

Graphical perception research often relates to design by quantifying how visual differences can lead to changes in data comprehension and task performance (e.g., visual search accuracy). In many ways, it can be thought of as applied vision science. Topics can include perceptual evaluation of bar chart vs. pie chart usefulness [25, 176], how to best style visualizations to convey uncertainty [54], how the number of categories in a chart can affect visual search [60], and how animated transitions can help improve viewers' understanding of data [72].

In contrast to graphical perception's focus on vision science, task requirement analysis focuses on understanding how cognition relates to usable design. Task requirement analysis largely approaches this problem by investigating the motivations and intent behind tool users' behavior through ethnography or other similar qualitative methodologies. For example, Brehmer et al. performed longitudinal interviews and observations to characterize how visualization tools can best support investigative journalists' analysis workflows [15]. However, sometimes tasks are also evaluated through quantitative methods, such as whether one type of visualization technique often causes biologists to make more research "*insights*" ("Aha!" moments in analysis) [159].

The following chapters provide novel theoretical contributions to both graphical perception and task requirements analysis sub-fields, and use established theory to build new computational techniques that assist in visualization tool design.

2.2 Cancer Genomics

Although a detailed knowledge of cancer genomics is not strictly necessary to understand the research outcomes of the following chapters, it may be helpful to have familiarity with commonly used genomics terminology to understand our research motivation, evaluation design, and several of our research contributions.

2.2.1 Genomics foundations

The backbone of cancer genomics is the *genome* – a string of `A`, `C`, `G`, and `T` nucleotides that contains the information needed to create life. Parts of the genome describe how to make *proteins*, which are the core building blocks of cells and are one source of what determines cell functionality. The instruction-filled areas of the genome that describe how proteins are constructed from molecular soup are called *genes*. Genetic mutations (hereafter referred to as mutations) are changes to a single or sequences of nucleotides in a gene. Cells decide what proteins to make through a process called *transcription*. To perform transcription, a cell uses instruction sets called *transcripts*, which are in turn created from the information stored in genes. *Protein domains* are areas on transcripts that contain important protein information, such as what the 3D protein structure should look like, which directly affects the protein's purpose and functionality.

2.2.2 Cancer genomics

From a microbiology perspective, cancer is what happens when cellular growth runs out of control and a *tumor* is the physiological region of cells that have grown out of control. Cancerous cell growth is often caused by drastically increased cellular replication or longevity, both of which can lead to unsustainable cell populations. One goal of cancer genomics research is to separate the mutations that drive cancerous growth (*driver mutations*) from the vast majority of harmless mutations. As such, cancer genomicists are interested in understanding genes because even simple mutations, such as changing just one of millions of amino acids, can create malformed proteins that could eventually lead to cancer. Although these *single nucleotide variants* (SNVs) are sometimes to blame for cancer development, other times cancer development can be propagated by larger mutations that encompass thousands of, or more, base pairs. For example, *copy number aberrations* (CNAs) cause entire regions of genetic code to be deleted or amplified (i.e., copied) compared to SNVs, which only change a single letter (e.g., `C` → `A`). Mutations that reduce the functionality of a protein, or eliminate it entirely, are called *inactivating mutations*. Conversely, *activating mutations* amplify protein functionality.

2.3 MAGI: visualization and collaborative annotation of genomic aberrations

MAGI is an online visualization tool that allows cancer genomics researchers to explore genetic mutation data across various cancers [107]. MAGI's primary functionality allows researchers to visualize arbitrary sets of genes across a variety of publicly available cancer datasets. Once a query is loaded by MAGI, it presents the user with five visualizations that show different types of mutation information. An example MAGI query is shown in Figure 2.1.

MAGI was originally developed to meet the research requirements of our immediate collaborators. It was this original development that then led to our follow-up task requirements evaluation (Chapter 3) and interaction log analysis (Chapter 4).

2.3.1 Details about the MAGI query page and its visualizations

The topmost visualization in MAGI's query page is the *aberration matrix*, which shows genetic mutations across different tissue samples from which genetic information was sequenced. Samples typically refer to separate patients. Each cell of the matrix marks whether a particular sample (column) had a mutation in a queried gene (row), and color refers to the type of cancer the sample had. By default, the samples are sorted to emphasize *co-occurrence* and *exclusivity* of genetic mutations across samples, which can be important when interpreting the biological significance for a set of genes. Co-occurring mutations are those that frequently occur together within a single sample and are shown as vertical stripes. Exclusive mutations are those that are mutually exclusive across samples (i.e., only one mutation in a gene query is mutated in each sample) and are shown as "staircase" patterns. Both co-occurrence stripes and exclusivity staircases are shown in Figure 2.1.

The next visualization is the *heatmap* which shows how common the products that each gene (row) makes in each sample (column) (*gene expression*). Users can also upload other continuous data such as methylation information to show instead.

The third visualization that MAGI shows for a query is the *network view.* Network nodes represent each gene in a MAGI query, where redder nodes represent more frequently mutated genes. Network edges mark whether the proteins that two genes (nodes) make are known to interact with one another. For example, two proteins might interact to help a cell decide when to replicate and, if one or both of the proteins' genes are malformed from mutations, researchers might form a prediction that the mutation(s) play a role in the development of cancer.

Next, MAGI renders a *transcript chart*, which shows the physical location of single nucleotide variant mutations (SNVs) on genomic transcripts. In addition to SNVs, the transcript chart also shows protein domains as well as transcript nucleotides if the user zooms in far enough. As in the aberration matrix, color corresponds to cancer type. The different types of glyphs marking SNVs correspond to particular types of SNV mutations. Mutations that are shown below the transcript bar are mutations that are known to reduce or eliminate the functionality of proteins (*inactivating mutations*). Users can navigate between different transcripts associated with a MAGI query using a drop-down menu.

The last visualization included in MAGI's query view is the *copy number aberration browser*, which displays larger amplification or deletion mutations that affect many nucleotides (opposed to SNVs, which only affect single nucleotides). As in the transcript chart, users can navigate between different genes using a drop-down menu. As a user toggles the navigation menu, the copy number aberration browser will display all copy number mutations that affect the part of the genome where a given gene is located. For convenience, the gene encoding region is highlighted with a red vertical bar that bisects each copy number aberration.

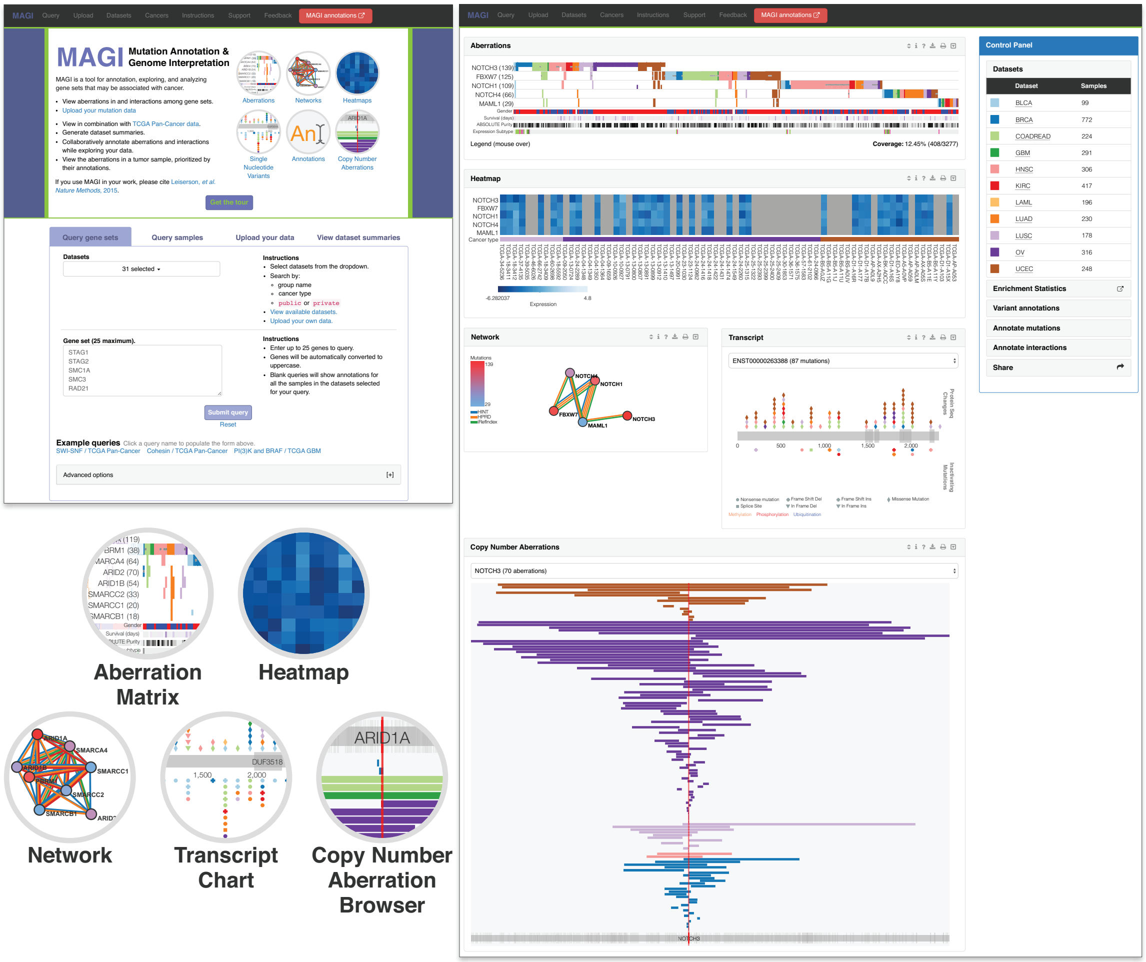
2.3.2 Other features of MAGI

Researchers can also use MAGI for analytical tasks other than visually exploring data. Using the datasets page, users can look at summary information about a particular cancer such as mutation frequency information in each dataset. On the MAGI annotation page, users can browse, add to, or vote on MAGI's annotation database to explore published information about how genes may be implicated in cancer.

2.4 Color as it relates to visualization

Color is integral to visualization design research given that it is one of the most common ways to encode information graphically, and there are many ways it can be manipulated to affect tool use. For example, distinct colors tend to "pop out" in images [193] (e.g., a red dot in a field of light gray dots), which can greatly improve visual search performance [48]. However, color can also be easily misused, such as accidentally picking non-discriminable colors [179] or not considering whether there is sufficient contrast between a color palette and a visualization's background.

Much of the difficulty surrounding color design stems from the fact that computers and monitors encode color differently than how humans perceive it: Monitors display color using red, green, and blue light, and digital color is most often defined using RGB color space. Although human eyes have cone cells that roughly correspond to those three colors, human color perception ultimately relies on different color scales. This process is called opponent color processing, where the visual cortex transforms cone responses into lightness (e.g., black to white), redness-to-greenness, and blueness-to-yellowness scales [80].



MAGI Query Upload Datasets Cancers Instructions Support Feedback MAGI annotations

MAGI Mutation Annotation & Genome Interpretation

MAGI is a tool for annotation, exploring, and analyzing gene sets that may be associated with cancer.

- View aberrations in and interactions among gene sets.
- Upload your mutation data
- View in combination with TCGA Pan-Cancer data.
- Generate dataset summaries.
- Collaboratively annotate aberrations and interactions while exploring your data.
- View the aberrations in a tumor sample, prioritized by their annotations.

If you use MAGI in your work, please cite Leiserson, *et al. Nature Methods*, 2015.

Query gene sets | Query samples | Upload your data | View dataset summaries

Datasets: 31 selected

Instructions
- Select datasets from the dropdown.
- Search by:
  - group name
  - cancer type
  - `public` or `private`
- View available datasets.
- Upload your own data.

Gene set (25 maximum).

STAG1
STAG2
SMC1A
SMC3
RAD21

Instructions
- Enter up to 25 genes to query.
- Genes will be automatically converted to uppercase.
- Blank queries will show annotations for all the samples in the datasets selected for your query.

Example queries Click a query name to populate the form above.
SWI-SNF / TCGA Pan-Cancer Cohesin / TCGA Pan-Cancer PI(3)K and BRAF / TCGA GBM

Advanced options [+]

Control Panel

Datasets

	Dataset	Samples
🟦	BLCA	99
🟦	BRCA	772
🟩	COADREAD	224
🟩	GBM	291
🟥	HNSC	306
🟥	KIRC	417
🟧	LAML	196
🟧	LUAD	230
🟪	LUSC	178
🟪	OV	316
🟫	UCEC	248

Enrichment Statistics
Variant annotations
Annotate mutations
Annotate interactions
Share

Figure 2.1: Screenshots of the MAGI launch page (left) and the query-view page (right). On the launch page, users define what genes they want to query and what cancers they want to look at. On the query view, five visualizations show genetic mutations across the types of cancers selected by the user. The visualizations are ordered so that the aberration matrix is on the top followed by the heatmap. The network view and transcript chart are co-located in the middle and the copy number browser is on the bottom. A control panel floats to the right of the visualizations on its own track, which allows the user to perform various actions such as showing/hiding datasets. Detailed definitions of MAGI's components are in Section 2.3.1.

Because visualization color effectiveness ultimately relies on perceptual accuracy, we typically use a *perceptually modeled color space* instead of RGB color space. This perceptual color space is called CIELAB, and is defined by scales similar to those that the human brain uses: $L*$ (lightness: black = 0, white = 100), $a*$ (redness-to-greenness), and $b*$ (blueness-to-yellowness). Sometimes CIELAB is modeled in CIELCh with polar coordinates: lightness remains the same ($L*$), whereas $a*$ and $b*$ are translated into chroma (C, colorfulness) and $h°$ (hue angle).

The usefulness of perceptual color spaces such as CIELAB stems from the fact that they approximate *perceptual uniformity.* Color spaces that are fully uniform are shaped such that all colors that are equally as distant in color space should also be perceived by humans as equally discriminable. (Euclidean distance in color space is typically referred to as ΔE or DE.) It is important to note that CIELAB is an approximation of uniformity, and while CIELAB is better to use than completely non-uniform RGB, we discuss how approximation error may effect visualization design in Chapter 6.

The difference between perceptually-modeled CIELAB and non-uniform RGB color spaces is shown in Figure 2.2, which renders all displayable colors in the sRGB color gamut in both spaces. One practical implication for design between the two is that color interpolation will result in very different color shifts (Figure 2.3).

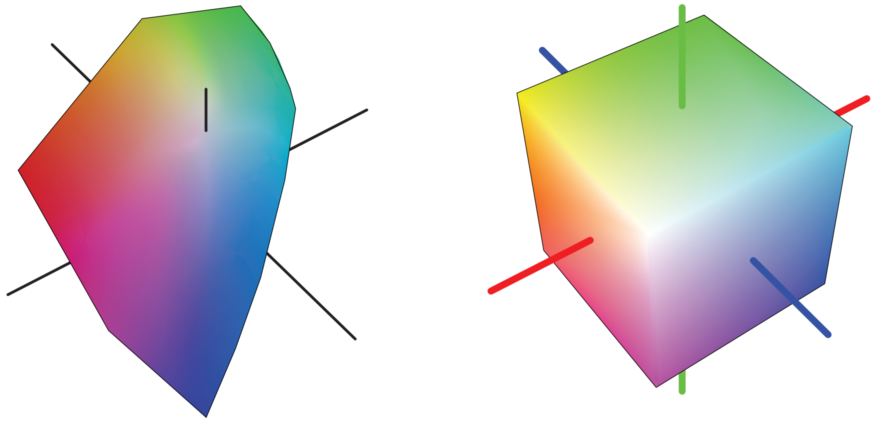


Figure 2.2: The RGB gamut rendered in CIELAB color space characterized with a D65 white point (left) alongside the gamut rendered in RGB space with the same perspective (right).

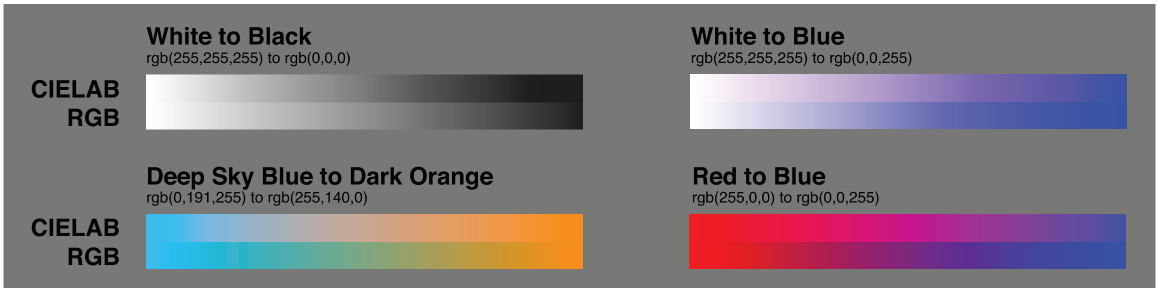


Figure 2.3: Linear interpolations between two colors in CIELAB (top) color space compared to RGB (bottom). Note that the color transition is different even for white-to-black achromatic interpolation.

Chapter 3

An evaluation of cancer genomics visual analysis with MAGI

In this chapter, we present common visual analysis task requirements of the cancer genomics community at-large and evaluate whether MAGI [107], a multiple-view visual analysis tool, supports these requirements. Our study is motivated by the visual analysis design challenges presented by the multidisciplinary nature of cancer research, where researchers with different specializations or backgrounds may use the same visualizations for different purposes. Consequently, cancer genomics visual analysis tools often need to support a range of information foraging and sensemaking strategies. For example, a pharmaceutical industry researcher and a basic science researcher might use the same data for very different purposes. This difficulty is further compounded because many of the visual analysis community's task requirement analyses were performed before the advent of next-gen sequencing [131, 160], which caused a paradigm-shift for genomics analysis due to the lowered cost and increased amount of sequencing data [45].

Cancer genomics background: Cancer is a disease of mutations, from which cell growth becomes unsustainable, often because of "driver" mutations that disrupt critical cellular regulation functionality (e.g., growth-rate [198]). Comparative analysis tools like MAGI [107] enable researchers to compare mutations across different types of mutations and cancers to understand cancer's genomic underpinnings. For example, MAGI allows researchers to study genetic mutation patterns across patients, which can indicate the biological significance of particular genetic mutations (Fig. 3.2) [197].

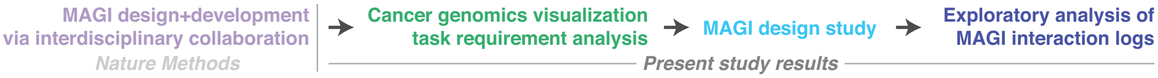


Figure 3.1: A timeline to show the division between initial MAGI development and our present design study contributions.

Contributions: Our primary contribution is a MAGI design study with three cancer researchers, which tested if, and how, MAGI supports cancer genomics visual analysis needs. The study was designed based on our second contribution: a task requirements analysis that resulted in the discovery of four tool-agnostic visual analysis tasks. We synthesized these requirements through interviews with cancer genomics researchers from a variety of specializations (e.g., pharmaceutical industry vs. basic science research) and occupations (e.g., investigator vs. staff programmer). Our third contribution is an exploratory analysis of MAGI interaction logs, which suggests that our case study findings generalize to ecological settings.

Outline: We begin with a brief background of MAGI (Sec. 3.1), which grew organically from a collaboration with cancer genomics researchers (for more information please see Chapter 2.3 or Fig. 3.1). In the related work we highlight potential broader impact of our present study by comparing the design similarities between MAGI and other genomics visual analysis tools (Sec. 6.1). We also draw on related visual analysis task theory to hypothesize that MAGI's multiple-view-based design would support a variety of tasks and information landscape orientations (Sec. 6.1). We use the remainder of the paper to present our MAGI design study contributions (see Fig. 3.1 for a timeline). We first present results from our tool-agnostic task requirements analysis, then provide observations from our MAGI design study, and conclude with an exploratory interaction log analysis.

3.1 MAGI: cancer genomics visual analysis

MAGI is a web-based cancer genomics visual analysis tool, designed and developed through interdisciplinary collaboration, that allows users to query sets of genes and explore various genetic information through five views [107]. Our prior Nature Methods work was a short methods paper detailing MAGI's architecture, whereas our present work is a design study of MAGI. One of the formative design objectives of MAGI was to support comparative analysis tasks such as the example in Figure 3.2.

For more information on MAGI, please consult Chapter 2.3.

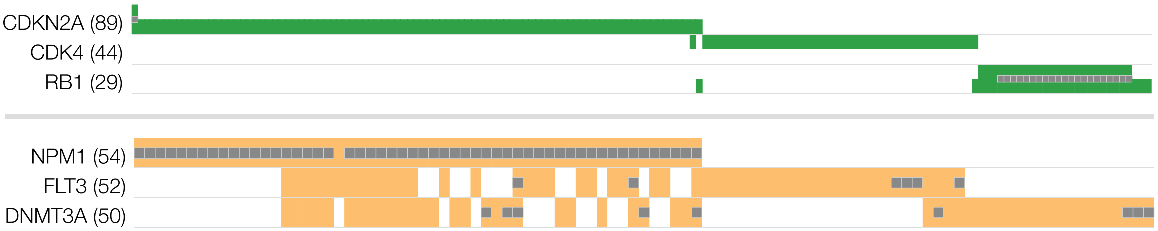


Figure 3.2: Aberration matrices showing gene set mutation mutual exclusivity (top; cancer: GBM) and co-occurrence (bottom; cancer: LAML). Mutated genes that are mutually exclusive do not often appear with each other in a single patient. Co-occurrence is opposite: genes in the same set are often found together. These patterns are one indicator that a set of genes could be biologically significant. Rows are genes and columns are patients in the above matrix, with filled matrix-cells showing mutations. Looking for these patterns is a common comparative visual analysis task for many cancer researchers.

3.2 Related Work

Below we (1) identify potential relevancy that our MAGI evaluation might have to other genomics analysis tools, (2) motivate our prediction that MAGI will support task requirements based on previous sensemaking frameworks, and (3) hypothesize that MAGI's multiple views will support a variety of tasks.

3.2.1 Evaluation contribution relevancy to other visual analysis tools

Genomics is a rapidly growing field, and a number of surveys review the wide array of visual analysis tools used to make sense of sequencing data [39, 136, 141, 168].

While our present evaluation specifically concerns MAGI, our findings may also apply to the many other multiple-view-based visual analysis tools. One example is the Integrative Genomics Viewer, which leverages multiple views to integrate heterogeneous data [151]. Another example is the many tools that are part of the Caleydo Project [181], which rely on similar multiple-view-based designs as MAGI to support analysis. For instance, Lex et al. show that coordinated multiple views help researchers navigate complex multidimensional genomics data [109]. Although our present work does not compare multiple- and single-view interfaces, results from our study establish a foundation on which future comparative evaluations could be built.

We show how our contributions may map onto similar tools with a design comparison in Table 3.1. Three columns contain information about aberration (mutation) matrix, heatmap, and network visualization support, which are taken from Schroeder et al.'s survey of multidimensional cancer

Tool Name	Independent multiple views*	Simultaneous, mult. data types*	Transcript Coding Region*	Heatmaps	Mutation matrices	Networks
MAGI*	×	×	×	×	×	×
cBio	×	×	×		×	×
CircleMap		×		×		
Circos		×				
Caleydo StratomeX	×	×			×	
Cytoscape	×	×				×
Genomica		×		×	×	
GiTools	×	×			×	
IGV		×				
IntOGen		×	×		×	
NAViGaTOR						×
Regulome Explorer		×				×
Savant Genome Browser		×				
CGWB		×		×		
UCSC Cancer Genomics Browser		×		×		

Table 3.1: A comparison of MAGI to other multidimensional cancer genomics visualization tools. The heatmap, mutation matrices, and networks columns are taken from Schroeder et al.'s cancer genomics visualization survey [168]. Fields marked with an asterisk (*) are newly added in our present comparisons to MAGI. "independent multiple views" is a subset of "simultaneous, multiple data types," marking only tools that render different data types across separate viewports. For example, Circos can visualize multiple data types, but as concentric circles in a single figure (i.e., viewport).

genomics visualization tools [168]. We also include new information about MAGI as well as three new columns. One column marks whether tools support transcript coding region visualization. Another marks whether tools support simultaneous visualization of multiple data types. The last, "independent multiple views," is a subset of "simultaneous, multiple data types" and only marks tools that separate the visualization of different data types across viewports. Our multiple-view distinction is motivated by previous findings that showed how small visualization differences, such as circular vs. linear layouts, can lead to different sensemaking procedures and insights [130]. Based on these findings, we thought it beneficial to mark which tools had multiple-view layouts most similar to MAGI. For example, Circos simultaneously visualizes different types of data as concentric circles, but within the same viewport, and could be considered a single figure. In contrast, MAGI renders linked views across five distinct viewports, which could be viewed as separate figures. The many design similarities between MAGI and the survey's tools suggests that findings from our present work may be applicable to the broader collection of multidimensional cancer visualization tools.

3.2.2 Sensemaking models suggest that MAGI's design will support task requirements

O'Day et al. illustrated how differences in sensemaking and information search create many opportunities for tools research by their qualitative study of researchers who regularly used microarray sequencing data [131]. They characterized sequencing-based analysis into two key stages: (1) testing *statistical significance* and (2) identifying *biological significance.* While automated statistical tests and predictive analysis may provide multiple hypotheses, it is ultimately up to researchers to explore statistically significant findings and to filter biologically significant leads. Saraiya et al. also observed the same analytical distinctions through an insight-based evaluation with genomics researchers [161]. Separately, Thébault et al. have shown the advantage that visualization gives in targeting biological significance [189]. MAGI, and our present evaluation, focus on the support of biological significance testing.

Streit et al. follows this separation of analysis along with tool-use observations to define a sensemaking model specifically for biological visualization tools that are designed to compare heterogeneous data [182]. Their model emphasizes how analytical performance can be improved by using different views to support different information landscape orientations and different analytical paths The design of MAGI is consistent with their findings – each MAGI view was designed to support different research foci and analytical paths – and supports our hypothesis that MAGI would support cancer genomics task requirements.

3.2.3 Multiple views may support multiple tasks

Many heterogeneous and comparative genomics analyses are best supported by matching analytical paths to specific data views [182]. Following this logic, tools with multiple views may support a wider range of task requirements and strategies for information foraging and sensemaking than single-view tools. Past task requirements analyses have suggested that “designing only for the ‘average user’ is not realistic” and that coordinated multiple views may not always be useful for all research procedures [209]. However, others have shown that comparative-analysis-support benefits from coordinated multiple views [109]. Similarly, Saraiya et al. suggest that coordinated multiple views may help information landscape orientation in biology-related visual analysis tools [161].

In fact, the potential link between comparative-analysis task affordance and multiple views is a common research theme across domain application areas. Roberts outlined a prospectus that surveyed several ways in which coordinated multiple views could improve interactive visual analysis of complex and varied data [150]. Heer and Agrawala hypothesized that coordinated multiple views might aid in serendipitous discovery [69]. Liu and Stasko speculated that different configurations of coordinated views may afford or constrain different analytical processes [113]. Finally, Gehlenborg and Wong provide design guidelines that emphasize how multiple-view techniques like small multiples can improve multivariate data visualization [38]. From this collection of research, we predicted that MAGI would support cancer genomics task requirements.

3.3 Preliminary Interviews and Task Requirements

We performed a task requirements analysis through a series of interviews to identify cancer researchers’ typical comparative analysis procedures when using visualization tools. Given the diversity of cancer genomics, we interviewed biologists, computer scientists, and multidisciplinary researchers to understand the breadth of requirements necessary to support the cancer genomics community at large. Participants’ research expertise was similarly diverse, covering topics such as understanding the heterogeneity of mutations in tumors, algorithm development to identify mutations that instigate lung cancer, and profiling differences across tumors to define more precise subtypes (e.g., "brain cancer" can be broken down into subtypes such as glioblastoma and anaplastic astrocytoma). Our collection of data was guided by three goals:

1. Identify the types of visualizations used to explore data
2. Compile a list of commonly used visualization tools
3. Understand how visualization is frequently used in analysis

Our intent was to gather information that could generalize to all cancer genomics visual analysis tools, not just MAGI.

3.3.1 Methods

Participants

Our interview participants came from three sources: an annual NIH TCGA meeting, a group interview at a genomics research center, and teleconferencing interviews. In all, we interviewed over twenty cancer genomics researchers covering a breadth of topics in biology, bioinformatics, and computer science. Participants included principal investigators, consultants, postdoctoral researchers, PhD students, and staff programmers.

Procedure

We asked questions based on the following lines of inquiry:

Q1 What does your overall analysis process look like?

Q2 What tasks are time consuming and time wasting?

Q3 What visualizations do you use in your analysis?

Q4 What are your short- and long-term research goals?

Interview formats differed slightly across the three types of interviews based on situation-specific limitations: the TCGA interviews were short and took place during conference breaks; the genomics center interviews took place in a group-format, hour-long interview; and the teleconferencing interviews each took between 30 and 60 minutes. Data were categorized by hand based on the three prior goals and by the four question types.

3.3.2 Interview results

Research goals and analysis process

Researchers' short-term goals typically focused on functional understanding of cancer, whereas long-term goals typically focused on the transfer of functional, biological knowledge into the clinical domain.

Interviewees reported that achieving these goals often required them to spend most of their time visually comparing their own data to related work. For short-term goals, researchers typically programmed their own wrangling and charting pipelines to visualize their results. In contrast, it was only after establishing a theoretical foundation for more specific hypotheses in late-stage research that researchers moved to pre-made visual analysis tools.

We also found differences in visual analysis procedure across research expertise. Computer scientists primarily used visual analysis results to verify the strength of the tools they built, since their focus was on toolsmithing [18] rather than finding novel insight about cancer genomics. In contrast, biologists often used visual analysis tools to better understand sequencing results. Many biologists reported using statistical languages like R to perform visual analysis, though they more commonly relied on visual analysis tools like cBioPortal to understand the biological context of their findings. Both bioinformatician and computational biologist procedures involved using visual analysis to find novel biological contributions and also to evaluate the quality of scripts or tools they had written. Because of their multidisciplinary roles, bioinformaticians and computational biologists often had the most complex analysis pipelines, usually written as a series of bash scripts piping together many charts from predictive algorithms on multiple cancer datasets with prebuilt visual analysis tools to compare their results against existing findings.

Common tasks that were difficult or frustrating

Data wrangling and reimplementation were cited as two of the most difficult and frustrating tasks that researchers commonly performed. One reason why this overhead was so problematic is cancer research's fast growth: using a new dataset often requires data conversion and cleaning alongside writing new visualization scripts. As a result, researchers' work environments were often disorganized collections of data pipelining, quickly reimplemented visualization code, and constellations of statistical or visual analysis tools.

A related issue was that many premade visual analysis tools were too narrow in domain focus or too hard to integrate with the format of data to which researchers had access. This rigidity often caused researchers to instead write their own customized charting scripts and led to the aforementioned data pipelining complexity. While curated data portals reduced the amount of data processing, researchers still often had first to wrangle large amounts of data that were cleaned in different ways. Bioinformatics and computational biology researchers were strongest in their complaints, perhaps in part because their analysis procedures were often the most multidisciplinary.

Nearly all interviewees complained about a lack of interactivity and data migration between visual analysis tools. Although researchers' institutions sometimes had tools to support data wrangling and preliminary data visualization, researchers were often forced to browse hundreds of handmade charts because they either did not want or could not use premade interactive visual analysis tools. One reason for this was that researchers often found pre-made visual analysis tools poorly designed to support their research questions, since the tools were often built with a different set of specific requirements in mind. Installation complexity was also a challenge: staff programmers reported that software often could not be deployed internally, or, if it could, deployment proved to be too complicated to complete. A related issue was difficulties moving data from the intranet of research centers to remotely hosted tools because of size or because of confidentiality.

Another task researchers found difficult was visualization design. Most researchers were concerned about the clarity of their charts, and many reported talking to resident data-visualization experts for charting advice.

Frequently-used visualizations & visual analysis tools for testing biological significance

Researchers often cited visualization as an indispensable part of their workflow because what is of biological significance is often not the same as statistical significance. The difference between significances suggests that visualization plays a critical role in assessing the biological impact of automated statistical analyses.

While researchers used a mix of visualizations to test biological significance, we found that most

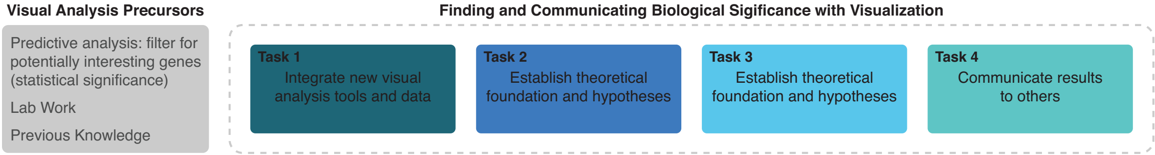

Visual Analysis Precursors
Predictive analysis: filter for potentially interesting genes (statistical significance)
Lab Work
Previous Knowledge
Finding and Communicating Biological Sigificance with Visualization
Task 1
Integrate new visual analysis tools and data
Task 2
Establish theoretical foundation and hypotheses
Task 3
Establish theoretical foundation and hypotheses
Task 4
Communicate results to others

Figure 3.3: A diagram of the four task requirements we identified through interviews that pertain to finding biological significance through visualization. To the left we include other important aspects of cancer genomics research, but that fall out of the scope of our present inquiry.

relied primarily on basic charts (e.g., scatterplots). This may be related to our observation that most researchers relied on handmade analysis pipelines in much of their work. While many researchers created biology-specific visualizations – like signaling pathway diagrams or Circos plots [100] – these specialized visualizations were often too hard or too complicated to make for day-to-day-research. The reliance on basic charts like scatterplots possibly occurs because visualization novices are likely to find them easier to create.

Although basic charts could satisfy researchers' analytical requirements, most said that they often had difficulty spotting patterns between images and that parsing results for biological significance was both mentally demanding and time consuming. This largely stems from our previous observation that researchers often had to compare and filter up to hundreds of handmade charts in file system windows. Researchers also used visualization tools like the Integrative Genome Viewer, cBioPortal, and UCSC's Genome Browser [92, 151, 168]; however, these tools were mostly used for testing specific hypotheses that researchers had made with their handmade visualizations. Some researchers also used design programs like Adobe Illustrator to touch up figures or create illustrations from scratch.

3.3.3 Discussion of requirements analysis

Identification of task requirements

We identified four common comparative visual analysis tasks guided by Brehmer et al.'s task typology [14]. Accordingly, we synthesized each task by aggregating and generalizing interviewer responses based on information about why each task was important, and what each task contributed to the larger scope of cancer genomics research. Our task requirement analysis focused on comparative analysis tasks related to finding biological significance across many different datasets and types of data. Therefore, what follows is not an exhaustive list of all possible cancer genomics research tasks, but rather a list of tasks found in our interviews that were relevant to comparative data visual analysis.

Integrate new visual analysis tools and data. Integration was frequently required for cancer genomics visual analysis because researchers first needed to produce refined sequencing data or results from statistical tests to visualize later or to use later in exploring other data. This task covers the majority of data-wrangling-related operations, such as constructing pipelines and cleaning data. But it also includes other operations such as implementing charting scripts and software deployment. The widespread difficulty of pipelining and integrating heterogeneous data replicates the results of many other previous biology visual analysis evaluations [22, 161, 182].

Establish theoretical foundation and hypotheses. While establishing theoretical foundations to understand their data, researchers might generate hypotheses, assemble related work, and look at specific trends in charts of statistical tests. In one example of this task, researchers reported looking through up to hundreds of charts generated through their pipeline to gain an understanding of the information landscape. Researchers might look up specific genes they specialize in, but might also browse through the entire catalog of results to shoebox points of interest to explore further. The output from this task is typically a collection of testable hypotheses about biological significance.

Test hypotheses and explore associations between data. Researchers typically tested hypotheses by looking for common associations across data. For some this meant visually comparing genetic mutations across different variants of cancer. For others, this task meant identifying genetic mutation outliers, such as a specific mutation that might be associated with higher morbidity rates. The outcome of this stage is typically an assessment of biological significance for a particular hypothesis.

Communicate results to others. The final task we identified was using visualization to communicate biologically significant results to others. We include this step because some researchers noted that they had to remake visualizations before sharing their results with others because default visualization output was often too crude. For some, this meant converting basic charts (e.g., scatterplots) into biology-specific visualizations; for others, it meant refining charts with design software (e.g., Adobe Illustrator).

Typical cancer genomics research workflows

Our results provide further support for O'Day et al.'s separation of statistical and biological significance [131] (Sec. 3.2.2): Researchers typically used predictive analysis tools (i.e., statistical significance) to filter large data before testing biological significance through visual exploration (the focus of our analysis).

We also found that while biologists are not first and foremost programmers, it is increasingly more common for biology researchers to adopt programming-based analysis and charting into their workflow to aid in comparative analysis. The increased prevalence of programming across research specializations is reflective not just of genomics, but also of many other analytics-heavy domains [89]. Thus, cancer genomics tools might often need to support research generalists who flexibly draw on skills from multiple areas. This underscores the desirability of comparative analysis tools that support a variety of data views and analysis tasks so as to support research generalists.

3.4 In-depth case studies

The purpose of these case studies was to to evaluate whether MAGI supported the previously identified cancer genomics visual analysis tasks requirements (Sec. 3.3.3). Each case study consisted of a brief interview about their research process followed by a MAGI-use observation. We focused our MAGI observations on hypothesis-generation and hypothesis-testing tasks because participants' institutional policies prevented us from observing participants using private data; however, our interview responses cover all four task requirements.

Our first case study participant (CS1) was a drug investigator at a major oncology pharmaceutical company who focused on identifying viable target sites for drug treatments. Our second case study participant (CS2), a postdoctoral researcher at a major biomedical research institute, focused on understanding the biological reasons how and why a specific type of mutation (copy-number aberration) were associated with certain cancers. Both CS1 and CS2 had graduated from biology-centric PhD programs. Our third case study participant (CS3) was a PhD student in a computational biology program and focused on developing algorithms to identify copy-number aberrations that might be significantly linked with cancer.

	CS1: Pharmaceutical investigator	CS2: Biology postdoctoral researcher	CS3: Bioinformatics PhD student
Education	PhD, Biomedical Sciences	PhD, Molecular and Cell Biology	PhD candidate, Bioinformatics
Focus	Clinical/Pharmaceutical	Biological mechanisms of cancer	Computational analysis of cancer
% of week spent programming	33%	30-50%	> 60%
Programming languages	Perl, Python, R	Python, R	C, Python
Visualization tools/libraries	GeneGo, OmicsSoft, Spotfire	Python, R (ggplot)	D3, Excel, Python
% of time using self-written analysis	< 5%	75%	< 10%

Table 3.2: Background information collected from case study participants.

3.4.1 Methodology

Our methodology was based on contextual design, modified to support time-limited, remote observations [201]. The main difference in our approach was that we provided a high-level, open-ended guiding "task" to jump-start observation sessions, as in previous genomics research observation [131].

3.4.2 Methods

Participants

Participants were drawn from a pool of researchers who had basic familiarity with MAGI and were selected to maximize coverage of cancer expertise, so that we could consider perspectives from basic, clinical and industry research perspectives. The first two participants participated remotely and the third participant was co-located.

Design & Procedure

Before scheduling observations we first e-mailed a background questionnaire to participants asking about their expertise (e.g., clinical, computational research) and research focus. Each participant was also asked to think of a few gene sets to use in the case study that were interesting in their current research.

The case study itself consisted of (1) an interview and (2) a MAGI observation. The interview questions were:

1. What percent of your work hours is spent programming?
2. What programming languages do you use?
3. What visual analysis software do you use in your research?
4. How do you plot/visualize your data?
5. What percent of your time spent on analysis do you use tools made by yourself?
6. What are the most time-consuming parts of your analysis?
7. What parts of your analysis do you find yourself having to do the most often?
8. What parts of analysis are the biggest wastes of time?

After the interview, participants were asked to query gene sets relevant to their research using the TCGA data hosted on MAGI. To retain ecological validity, we verified that participants regularly used TCGA data. The observation's purpose was described as an opportunity to understand how researchers used MAGI to "generate hypotheses about the importance of gene sets." Participants were asked to self-annotate their analytical process by talking out loud.

3.4.3 Results

Pre-observation interview results

Participant's education, research focus, programming languages, visual analysis tools, and visual analysis time allocation are shown in Table 3.2 (Questions 1-5). Below we describe participants' visual analysis tool use (Sec. 3.4.3, Question 6) and summarize their current analysis procedure (Sec. 3.4.3, Question 7). We also include a collection of analysis limitations and usability issues (Sec. 3.4.3, Question 8).

What visual analysis software do you use in your research? When participants were asked what visual analysis tools they use in place of MAGI, they responded:

CS1 Mix of their in-house software's built-in visualizations

CS2 Histograms via R/Python

CS3 Scatterplots via Excel/Python

CS1 reported that their analysis relied almost entirely on prebuilt tools such as OmicSoft and GeneGo, as their role as investigator minimized their programming responsibilities. They emphasized reliance on prebuilt packages in R or Python such as scikit-learn when scripting, and on stencil code written in earlier research projects. This dependency explains why CS1 reported using self-written analysis code only 5% of the time, despite spending 33% of their week programming (Table 3.2).

CS2 reported that their research process relied nearly entirely on their own homemade, pieced-together assembly of analysis scripts. This pipeline required integrating existing predictive analysis tools with their own analysis, charting, and data wrangling code written in R and Python. They did not report a reliance on commercial analysis software packages, like those used by CS1.

CS3 reported a similar strategy to CS2.

Current visual analysis procedure CS1 stated that a large portion of their analysis procedure involved visually browsing and comparing statistical results from predictive analysis to find important new information or hypotheses. They reported that this procedure was often protracted due to the volume of output generated by their predictive algorithms and because the information they often search for cannot be found through statistical tests due to low sample sizes.

CS1's analysis process started by applying predictive analysis techniques on a mixture of in-house and public datasets to create a general body of results. These results were then filtered by scripts and then by hand using background knowledge and visualizations to identify useful data. The resultant data was continually processed until CS1 was left with multiple sets of genes they considered interesting for hypotheses about biological significance. After acquiring gene sets, CS1 next looked at a chain of visualizations to single out mutation information that might indicate that a current drug compound in the company's toolbox might work in a novel application. More rarely, CS1 said that this process occasionally shed light on new drug treatment opportunities. These discoveries would be filed away for future discussion with teammates and with the company more broadly, or used for new gene set queries.

CS2 detailed an analysis procedure similar to that described by CS1: much of their time was spent by stitching predictive analysis results together alongside clinical data before starting visual exploration for biological significance. Unlike CS1, CS2 looked through static images of basic charts (e.g., scatterplots) they created with their own scripts rather than commercial, interactive visualization tools, which rendered data as biology-specific visualizations. To sift through information they would typically go through the entire collection of their catalogue, open up two or more charts concerning different slices of data in an image viewer, and take notes on patterns in A/B comparisons.

CS3's procedure differed from those of both CS1 and CS2 because, while they were interested in cancer genomics, they were first and foremost a tool researcher. Accordingly, the aim of CS3's visual analysis procedure was to evaluate whether results from their predictive analysis tool could help researchers find biological significance. CS3 recently started to incorporate interactive visualizations that they had written themselves using D3 into their workflow; however, in the past they had used Microsoft Excel to parse and visualize data.

Current limitations and self-reported usability issues CS1 reported that their largest time-waste was hypothesis foraging for a "short list of useful hypotheses" by visually foraging through genetic mutation data to follow up statistically significant patterns. They also stated that data wrangling was one of the most painful tasks they regularly needed to perform as part of their analysis, alongside integrating heterogeneous data and tools.

CS2 also reported frustration in spending much of their time piecing together different predictive algorithms together alongside existing datasets, despite their differing procedure. While CS2 did not complain about having to write their own visualization code and felt they had a "good enough" solution, they did comment that there was room for improvement in their process. CS2 stated that the largest barrier to using interactive tools, opposed to writing their own visualization code, was piping data. They said that flexible data importing, like in MAGI, would make them more likely to try out new tools due to lowered startup cost.

Conversely, CS3 was happy to have moved to D3 from Excel, but was dissatisfied with the amount of time visualization programming took. Like CS1 and CS2, CS3 also complained that data wrangling and hypothesis foraging took up a large portion of their time.

MAGI tool-use observations

CS1: Pharmaceutical investigator During the MAGI observation session, CS1 queried three different sets of genes. Most hypothesis testing and data exploration involved comparisons between the aberration matrix and a second view with heavy reliance on linked interactions. For each gene set query, CS1 always used the aberration matrix first because a large amount of their research questions depended on knowing whether genetic mutations co-occur within a given sample (i.e., patient) or whether genetic mutations are mutually exclusive across samples[1]. If CS1 found an interesting pattern, MAGI's other visualizations were then used to add further context. For instance, CS1 mentioned that the transcript annotation view was useful for determining whether an existing drug compound in their company's arsenal might target a certain mutation.

CS1 said that the largest improvement MAGI provided compared to the commercial visual analysis tools they already used was the ability to save vector-formatted visualizations. Although this is a simple feature, they said that they often had to remake their results in Adobe Illustrator when giving research presentations because the tools they typically used did not export in vector format.

As a whole, CS1 remarked that MAGI performed as well as the set of commercial tools they used for their job, which they enjoyed using. CS1 also noted that both the tools they used for their job and MAGI made noticeable improvements in their work compared to the homemade, static visualization scripts they used while a graduate student and postdoctoral researcher.

CS2: Biology postdoctoral researcher In CS2's observation they queried genes that they had identified earlier in the day as interesting and had not yet analyzed. The first visualization CS2 first consulted the aberration matrix for an overview of the data, and focused only on a subset of mutation types based on CS2's research focus on copy number aberrations. To parse the aberration matrix better, CS2 hid all other visualizations except the copy number browser (which would otherwise only be visible through constant scrolling back-and-forth) and rapidly went back and forth between the two visualizations to isolate interesting cases of co-occurring and exclusive mutations. CS2 then reenabled the hidden visualizations to contextualize their collection of interesting mutations. CS2 finished their analysis and added several functional discoveries by using the transcript annotation and MAGI's built-in tooltips that display additional mutation information. Near the end of the observation, CS2 found a potential breakthrough on a problem they had been stuck on for a week by using MAGI's network view, which showed data in a representation not typically included in

[1] Co-occurrence and exclusivity patterns are one approximation of whether or not a gene is implicated in the development of cancer.

their normal workflow.

Afterwards, CS2, who still relied heavily on static visualizations for analysis, said that MAGI felt less taxing and felt faster to use when forming and testing hypotheses.

CS3: Bioinformatics PhD student CS3's use of MAGI resembled CS2's, perhaps in part because CS3 built tools used in one of CS2's research areas (copy-number analysis). Throughout the observation, CS3 iterated through research questions and lines of inquiry faster than CS1 or CS2. While CS1 and CS2 were deliberate in their approach, sticking with one or two visualizations at a time, CS3 quickly cycled through many of the visualizations concurrently. Most of CS3's analysis hinged on observations made using the copy number browser rather than the aberration matrix. Their fast, iterative workflow using MAGI's multiple views was broken by pauses when they checked the copy-number browser for closer examination. One frustration CS3 experienced was the inability to refresh gene queries from the analysis page itself. Because they made constant changes to the gene set they looked at, they often had to return to the query page, enter a new query, and then wait for the results page to load. Like CS2, CS3 also felt that using MAGI was less taxing then traditional static chart analysis.

3.4.4 Discussion

Below, we first discuss the interviews to compare our case study participants' individual, tool-agnostic task requirements to our earlier requirement findings. Then, we discuss how MAGI supports both sets of requirements. After, we discuss how multiple views might be one explanation for why MAGI task requirement support, and why MAGI might also support serendipitous insight.

Tool-agnostic interviews: finding biological significance and current limitations

After interviewing case study participants, we found that their requirements and limitations largely mirrored our task requirement analysis results (e.g, visualization is important for inferring biological significance) as well as previous analyses of genomics research procedure (Sec. 3.2.2).

One common limitation for testing biological significance was the inability to quickly forage biological hypotheses from high volumes of statistically significant results. Each participant faulted integration task bottlenecks that were caused by both data wrangling and visualization.

CS2 and CS3 also cited programming their own visualizations as a limitation. In contrast, CS1 was able to access robust commercial visualization pipelines, which they said helped, but suffered from data integration bottlenecks. The results from CS1's interview are encouraging because they show that interactive visualizations help foraging for biological significance through a large number of statistically-based hypotheses. However, CS2 and CS3's forced reliance on programming their own visualizations suggests that a lack of visualization tools that are easily accessible and that also effectively support researchers' tasks remains a limitation for visual analysis.

The interview results also highlight the importance of task requirement analyses. Although there is an abundance of publicly available interactive genomics visual analysis tools, both CS2 and CS3 felt their specific analysis requirements were underserved. The lack of support led CS2 and CS3 to write their own visualizations, which caused time-consuming A/B comparisons (CS2) and also took time away from research because of the burden of programming.

Although we were unable to observe user data uploading in our case studies, these interview responses reinforce our design decision to support custom user data. A possible alternative to broadening custom data support is to develop novice-friendly declarative interactive visualization toolkits (e.g., Vega [163]) that are specialized for genomics visualizations, which could address these limitations by reducing the workload required to create custom interactive visualizations. To this end, we open-sourced GD3, the visualization library powering MAGI, which makes programming D3-based interactive genomics visualizations more declarative.

Observations: MAGI supports task requirements

Recall that the two tasks we selected to evaluate MAGI's effectiveness were (1) generating and (2) testing hypotheses about biological significance. Our observations suggest that MAGI supports both. For instance, CS1 was able to test hypotheses about a drug target site, CS2 found a potential research breakthrough they had been stuck on, and CS3 was able to gain new knowledge about copy number aberration mutations. Furthermore, participant feedback indicates that MAGI also supported visual communication tasks by providing a vector-graphic export option.

Each observation also suggests that MAGI addressed many of the participant's individual analysis limitations. CS1 was able to sift through hypotheses quickly. CS2 was able to explore existing datasets with interactive visualizations, rather than having to perform many A/B comparisons. Last, CS3 could explore results without the burden of programming.

The participants typically reported that MAGI's limitations were related to tool-workflow integration tasks. CS1 reported that the largest difficulty in using MAGI was deployment difficulty, which led to revisions such that MAGI now supports Docker imaging to reduce deployment overhead. CS2 reported previous difficulty with data upload, which led to revisions of MAGI's uploading to support a wider array of data formats. The only reported limitation in hypothesis generation and testing was CS3's feeling that the lack of results-screen query refinement slowed his workflow, which led to the development of a query menu accessible from any MAGI page.

Task support may stem from multiple views

Our case study observations indicate that MAGI both successfully supports task requirements and supports different specific analytical procedures. These results suggest that MAGI's design successfully addressed complaints from our preliminary and case study interviews about visual analysis tool brittleness.

One potential reason that MAGI supported the range of task requirements and procedures is its multiple-view interface. Our case studies show that participants' procedure often relied on switching between many views to converge on analytical insight. For instance, CS3 was a copy number aberration expert, but frequently compared across all visualizations to refine hypotheses given the broader biological context. Another example is the many ways in which participants interacted with the aberration matrix. It was designed to support the types of tasks that CS1 performed, primarily searching for co-occurrence and exclusivity trends in mutations across patients. However, MAGI's open layout made the visualization useful for CS2 and CS3, who often used the mutation matrices to look only at the frequency and patterns of certain types of mutations shown as glyphs on mutation cells in the matrix. These examples show that MAGI supports a variety of individual differences in analytical procedure.

While our suggestion that MAGI's multiple-view interface supports a variety of analytical procedures is preliminary, given that it was a byproduct of our task requirement evaluation, it is supported by many theoretical frameworks. For example, associative browsing benefits from flat-structured interfaces that can support more various possible analytical paths than more sequential or rigid interfaces [87]. This may be particularly true in cancer genomics, given that past work has shown the field is host to a breadth of research foci, multidisciplinary collaboration, and a correspondingly large variety of analytical procedures [131, 182]. Another benefit of using multiple views is that familiar views of data can help researchers understand nearby, unfamiliar views [16], which our observations also show. Similarly, there is preliminary evidence that multiple views lead to higher usability ratings by case study participants in cancer visualization tool evaluation [146].

Despite these successes, one potential pitfall of multiple view design is the potentially large distances between visualizations. For example, CS2 was unable to view both the copy number browser and aberration matrix at the same time with MAGI's default layout. CS2 was able to solve this by temporarily hiding the other visualizations in MAGI; however, not supporting layout management might have made MAGI unusable for CS2's comparative analysis. As such, supporting a balance of view-completeness versus view-accessibility is an important design decision when constructing multiple-view-based tools.

Taken together, these results suggest that cancer genomics visual analysis interfaces that flexibly allow users to switch among different analysis paths may effectively support a variety of analytical strategies and hence a variety of task requirements.

MAGI task requirement support led to analytical insight

As part of supporting each participant's tasks, MAGI supported small incremental insight discovery. Our interviews also showed that MAGI supported larger, serendipitous insights. While CS2 was being observed, they solved a research problem they had been stuck on for a week by scanning a visualization that would normally be tangentially related to their research. Although it is difficult to understand whether the design of MAGI regularly supports such large leaps in understanding, our observation certainly motivates further study of multiple views and insight generation.

We suggest that MAGI's multiple views may afford serendipitous insights for the same reason that it supports flexible sensemaking: If researchers are exposed to views of data that are outside of their normal analysis perspective, they may well be exposed to a new, more productive orientation to the information landscape. Saraiya et al. made similar observations with respect to multiple data representations and serindiptious insight in a longitudinal study of bioinformatics visualizations [161]. These observations are both consistent with information search research which found links between flexible interfaces and an increase in creativity and inspiration [190]. Dörk et al., combining many of these thoughts, demonstrate that supporting curiosity and flexible sensemaking can greatly enhance the explorative and analytical power of visual analysis tools [27, 28]. Taken together, these case study observations and theoretical frameworks both indicate that there is a high likelihood that

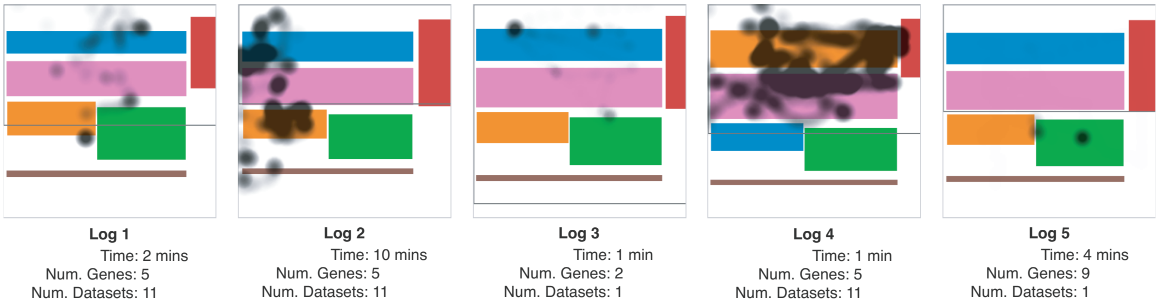


Figure 3.4: Five MAGI mouse interaction logs showing mouse movement over time. Each session suggests widely differing analytical procedures. The black heatmap overlay shows the frequent interaction locations. Each rectangle is a different visualization in MAGI. The copy number browser (bottom) appears short because its height changes based on which gene a user sets the browser to show. Dark gray lines indicate window size.

multiple views may support serendipitous insight and information landscape reorientation.

3.5 MAGI interaction log analysis

We analyzed a random sample of five online MAGI mouse trace interaction logs (i.e., movement over time) to explore the generalizability of our case study observations. The samples were drawn from a dataset with thousands of sessions. The aim of exploring trace samples was to specifically test whether our case study observations that MAGI supported a variety of analytical procedures transferred into ecological settings. Although these traces do not preserve analytical intent or context, they do provide information about how MAGI is typically interacted with. If there were a diversity of procedures in the traces, it would support our suggestion that it is important to design cancer genomics visual analysis tools that are flexible enough to support a variety of individual differences. While it would be possible to examine all mouse traces in the dataset, such an expansive study is beyond the scope of our present work, and random samples are sufficient to test our specific analysis aims.

3.5.1 Methods

To better understand how MAGI is used “in the wild,” we collected thousands of anonymized interaction logs from online use and randomly sampled five for our present evaluation. The logs were selected after removing time outliers, so that session times were between 1 and 20 minutes. Each log contained the window resolution, visualization locations and sizes, the number of genes and datasets queried, and mouse traces containing clicks and moves.

3.5.2 Results

Mouse movement heatmaps for each session, aggregated over time, are shown in Figure 3.4 along with each session's duration and the number of genes and data sets queried.

Interaction log descriptions

The first user (U1) browsed through aberration matrix tooltips, selected a different gene to examine with the transcript annotation chart, and then used the heatmap to look at brushing and linking with the aberration matrix. The mouse activation over the subnetwork came from apparent idle mouse movements.

U2 explored each gene in the subnetwork, pausing to look at tooltips. Then they moved the mouse over the heatmap, possibly looking at an irregularity. Next, they selected a gene in the copy number browser. Last, they switched back and forth between cells in the aberration matrix and the aberration matrix's legend.

U3 looked at several different genes in the transcript annotation chart and then at several different cells in the aberration matrix.

U4 browsed through the heatmap and aberration matrix. Interactions suggest that they used the brushing between both visualizations to keep track of the column (i.e., patient) they were looking at in each.

U5 spent most of their time looking at specific mutations in the transcript annotation chart. However, playback also shows that they briefly consulted the mutation matrix and heatmap between transcript annotation use.

3.5.3 Discussion

While the context of analysis is lost by scraping online tool-use, the trace heatmaps indicate that users exploit MAGI's multiple views in very different ways. Each of the five interaction log samples shows a distinctly different process and different use of visualizations. This task variance supports our earlier case-study observations and interview responses that analytical procedure varies greatly in cancer genomics. The lack of context prevents us from linking analysis procedure differentiation to the breadth of research foci within multidisciplinary application areas. However, it does support our claim that it is important to support a diverse set of procedures, and also suggests that multiple views are an effective design for doing this.

Analytical procedure differentiation persisted regardless of the number of genes or datasets queried. For single-cancer analysis the information traces show both targeted (U5) and comparative (U2) analysis between the different views in MAGI. U5 spent nearly all 4 minutes in the session on cycling through genes in the transcript chart, whereas U2 spent 10 minutes going through a sequence of every visualization but the transcript chart. The multi-cancer logs (U1-3) show similar diversity, even though their total session time tended to be shorter (1 or 2 minutes rather than 4 and 10).

Thus, mouse traces of online use of MAGI supports our claim that the coordinated-multiple-view layout of MAGI supports a robust set of sensemaking procedures. These results also show that many of our case study observations extend into ecological settings. These findings also support other work studying the relation between multiple views and individual differences. For example, Marai found that users' expertise in spatial vs. non-spatial visualizations affected view order-of-use in visual analysis tools [117]. It is possible that our log analysis supports similar expertise based differences with respect to expertise.

3.6 Open research areas

Our case studies were designed to focus on whether MAGI supported a set of common cancer genomics task requirements. As such, while our association between multiple views and robust task requirement support and insight generation are backed by theoretical foundations, both are interesting starts for more detailed and tailored research. Another interesting direction would be to perform a longitudinal study to test whether MAGI can support long-term research gain and task requirements opposed to short-term insight discovery. Last, one potential limitation is the design study population size, which stems from domain expert recruitment difficulty.

Our task requirements analysis highlighted the difficulty that researchers have wrangling data and integrating new software into analysis pipelines. These findings echo Kandel et al.'s findings from interviewing data scientists in many industry sectors [88, 89], and, together, suggests that integration issues in cancer genomics are one instance of larger data processing open research problems.

Another open area of research is testing the extent to which remote interaction log collection can be used to test task requirement analyses. For instance, is it possible to infer task requirements from interaction logs alone?

3.7 Conclusion

We presented a design study of MAGI, a cancer genomics visualization tool. We first performed a task requirements analysis and identified four common tool-agnostic visual analysis tasks after interviewing over twenty researchers and programmers working within cancer genomics (Sec. 3.3). We also found a distinction between biological and statistical significance, which is consistent with previous genomics tools research (Sec. 3.2.2). In our case study observations we found that MAGI supported both these four common tasks and also the specific procedures regularly performed by the case study participants as part of their normal workflow (Sec. 3.4). MAGI provided a useful alternative to their established workflows for a variety of reasons including interactive exploration, providing vector graphics that could be exported for high-quality printing and slideshows, and facilitating serendipitous insight. We suggest that MAGI's success might be traced to its multiple-view-based design. This suggestion is supported by many past theoretical stipulations (Sec. 3.2.3), and is based on observations that demonstrate how using multiple views aided researchers in testing hypotheses and making serendipitous insights. Finally, our exploratory analysis of how MAGI is used online indicates that MAGI is able to support a variety of analytical procedures in ecological research settings (Sec. 3.5).

While our list of task requirements and evaluation focused on cancer genomics and MAGI, we believe that these results are applicable to other tools within cancer genomics, given that MAGI's interface design is similar to other cancer genomics visual analysis tools (Sec. 6.1).

Chapter 4

Evaluating visual analysis task classification to improve understanding of cancer genomics domain expert use of MAGI

In this work we advocate that interaction log classification can serve as a new, effective visualization tool design evaluation methodology, and focus on how it can augment traditional qualitative approaches by providing additional context for previously determined tasks. We also explore how predictive task inferences can be used to improve the iterative design process of interactive visualization tools for domain experts. To accomplish this, we ground our exploration in an analysis of MAGI [107] – a cancer genomics visualization tool. These contributions extend current tool evaluation methodologies, which typically focus on field studies and other similar, typically qualitative, types of observation [103]. Although working side-by-side with domain experts in field research yields high levels of detail about analysis workflows, as Carpendale notes, these types of studies are typically smaller in scale and lack precision [20]. Our contributions could provide an important addition to current evaluation methodologies because interaction logs can be passively collected as part of domain experts' natural workflows and also contain precise, quantitative descriptions of visual analysis. Because of this, interaction log analysis can circumvent several common limitations present in more focused and contextual-rich methodologies (e.g., ethnographies). For example, through interaction log analysis, it is easier to study larger populations of domain experts while retaining naturalistic, ecological validity and without potential interference caused from direct observation. Likewise, analyzing large collections of interaction logs may help thwart bias caused from observing small in-lab populations.

Another motivation of our present work was to understand the degree to which anonymized interaction logs could be used to understand analytic intent given the complete omission of context.

Our evaluations of visual analysis task inference by humans and computers rely on interaction logs that contain the size and location of each visualization in MAGI and the sequence of mouse events caused by user interaction (i.e., clicks, movements, and scrolls).

Contributions: Our first contribution is a discussion that compares the accuracies of twelve automated visual analysis task classification models to hand-coded task inferences made by pairs of genomics and visualization experts. Rather than focusing on sophisticated classification models, our evaluation focuses on classifiers that most visualization researchers could implement themselves: k-nearest neighbors, linear support vector machines (SVMs), and random forests. This way, our findings are more easily applicable to visualization researchers and practitioners at-large. We discuss the potential benefits that might come from evaluating more complex models in Section 4.6. Our second contribution is an exploration of common MAGI interaction trends using the predictions from task classification, which expands our present understanding of how visualization is used in naturalistic settings by cancer genomics domain experts. As part of this investigation, we make our last contribution by exploring how mouse interaction modeling can be used to inform iterative tool design. We also provide design principle hypotheses that can be used to guide future design studies.

Outline: We begin with interaction log mining background and related work (for more information about MAGI and cancer genomics as an application domain, please see Chapter 2). We also explain what types of information we collected in the MAGI interaction logs. Next, we discuss results from a preliminary task inference study in which we worked with two MAGI developers to identify eight common MAGI analysis tasks. We then discuss the results from a task labeling experiment that provided training data to evaluate the performance of MAGI analysis classifiers. Following our in-lab experiments we then move on to our classifier evaluation and explore the potential effect that interaction log mining might have on domain expert tool iterative design. Last, we present open research questions and consider potential broader impact of our contributions.

4.1 Background and Related Work

4.1.1 Understanding users: contribution differences

While our present work is related to previous research "clickstream" interaction analysis, our contributions differ because of our focus on visualization tools for domain experts: we aim to model less deterministic visual analysis behavior of experts instead of modeling typical navigation behavior of the general population through a sequence of URLs (e.g., to optimize search ranking [4] or commerce [56]). (A more thorough discussion of clickstream research is included in the supplemental material.) Such clickstream tasks are more deterministic because a user's goal is to find the most relevant search result and will end with a success (search result click) or a failure (search termination or another query). In contrast, visual analysis is typically driven by deriving "insight," which is subjective and variable across applications [44]. Because of potential empirical differences like these, we test whether clickstream features from the information retrieval community can accurately model visualization interaction. Hence, another contribution of this work is to assess whether features that were advantageous for classifying these simpler, more deterministic interactions in web search apply equally as well to more open-ended visual analysis scenarios. However, further evaluating how visual analysis interaction procedure may differ from better-studied and modeled areas of human-computer interaction remains an important area for future research.

4.1.2 Understanding analytic intent via interaction logs

Our present research compliments and expands on automated analytical task inference techniques within visualization and across the broader human-computer interaction community. Although manual interaction analysis has proven useful in smaller case studies such as studying visual analysis in investigative journalism [15] and in understanding collaborative analysis [81], Guo et al. note that hand-coding users' interactions faces myriad scalability issues [55]. As such, many researchers have investigated the automation of visual analysis interaction log evaluation. These techniques often seek to identify design requirements by leveraging interactions as a record of "analytical provenance," which can be loosely defined as a collection analytical steps undertaken during a visualization's use.

Given the scope of provenance research, rather than survey it here, we recommend Ragan et al.'s survey [145].

Within the context of interaction history mining, much of this research has focused on action log analysis, which relies on basic software interaction sequences (e.g., `filter` → `sort` → `select`). For example, Zgraggen et al. showed how extracting interaction patterns using regular-expression-like queries from large action datasets helped usability researchers at a large technology company identify key issues in their products [208]. Other visual analysis task reconstruction methods draw on techniques such as multiple sequence alignment [8, 34, 203, 204], graphical modeling [75], and human-in-the-loop qualitative exploration [129]. Etemadpour et al.'s investigation into genomics analysis workflows is more similar to our inquiry into domain expert analysis, but also uses an action representation akin to other previous work [33]. Our present work differs from these efforts because we focus on lower-level mouse event analysis (e.g., mouse dwell time) to infer analytic intent, rather than focusing on higher-level action data. In this way, our investigation differs from much of the past work in analytical provenance, which typically models interaction at these higher-level representations (e.g., "undo" in a graph-like structure representing workflows [145].

One benefit to analyzing lower-level mouse events opposed to higher-level representations is the close relationship between mouse movement and gaze, which is a well-studied physiological indicator of intent [77]. Huang et al., as well as Rodden and Fu, explore how the relation between gaze and mouse movement can be used to improve web search [78, 152], and Gomez et al. show that the relation also holds for visualization [43]. We utilize this similarity later in our classification evaluation by creating a new feature set inspired by these similarities (Sec. 4.5.1).

Martín-Albo et al. build on the association between intent and mouse interactions to show that intent can be inferred from mouse movement alone without the aid of eyetracking by testing the geometric and temporal similarity between mouse traces [118]. Others like Edmonds et al. and Matejka et al. developed tools to qualitatively analyze mouse traces and intent through heatmaps of frequently interacted-with interface regions [30, 119]. Blascheck et al. pursued a hybridized in-lab approach and tested how event-level interaction logs can be combined with talk-aloud transcripts and eye-tracking to understand interaction [8]. Noting the potential benefits of using higher-resolution interaction logs, Atterer et al. performed a case study to show how interaction strategies and intent can be reconstructed from low-level event logs [6]. Our present work extends knowledge of user analytic intent by analyzing how interaction log classification can lead to insights about domain experts' naturalistic visual analysis behavior.

4.1.3 Relation to past biology visualization task analyses

Our present contributions extend previous research that also used biology visualization as a test bed for new evaluation methodology and task modeling. For example, Saraiya et al. developed an evaluation methodology to measure visualization effectiveness based on how many analytical insights it may support [159] and then explored how insights could be used to longitudinally understand visual analysis tasks [161]. O'Brien et al. then extended insight-based methodology to improve its precision while also evaluating another biology-visualization-motivated application [130]. Instead of just tallying the total number of insights, they suggested that insights – and the tasks that produced them – could be better understood by also measuring a variety of other information such as hypothesis-driven insights and insight complexity. Unlike these past methodological contributions, which rely on hand-coding data, our present line of inquiry investigates how automated modeling can empower initial human classification. Not only does this continue O'Brien's line of research toward quantifying task analysis, but it also allows these labor-intensive methodologies to scale to much larger collections of data thanks to automated task inference.

Others, like Streit et al., used biology visualization as a way to study visual analysis in areas where there is diverse types and formats of data [182]. Whereas Streit et al. focused on constructing a model for heterogeneous biological data analysis, Murray et al. synthesized common analysis tasks in biological network analysis [128] Although both sought to explain cancer genomics visual analysis, the aims of our present work are distinct. Differences between our present contributions and these past two models might be best understood through Brehmer and Munzner's task typology [14]: Streit et al. primarily focused on "what" each task was operating on, Murray et al. primarily focused on "why" each task was being performed, and our present research primarily focuses on "how" each task was performed.

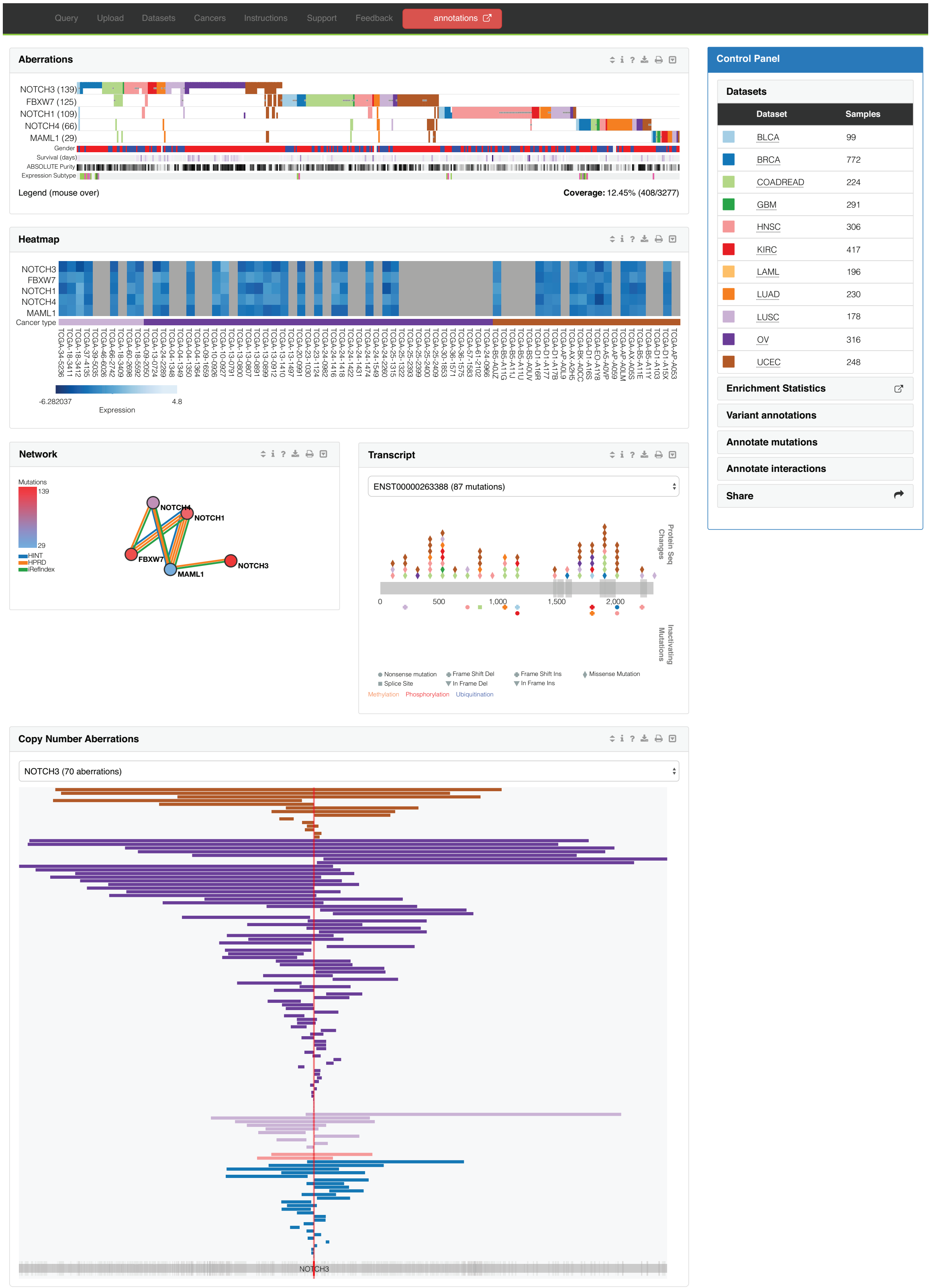
4.2 MAGI and Log Collection

Our present investigation into visual analysis task classification is anchored by studying MAGI mouse interaction logs. MAGI is an online visualization tool that allows cancer genomics researchers to explore a variety of genetic mutation data across many cancers in five visualizations [107]. Given cancer genomics specialization variety, MAGI was designed to support a diversity of expertise through its multiple views (e.g., basic science vs. pharmaceutical research; wet lab biologist vs. bioinformatician). A screenshot of a query in MAGI is shown in Figure 4.1, and Chapter 2.3 provides detailed background information about the tool.

Like with many other visual analysis tools for domain experts, one difficulty in evaluating MAGI is that cancer genomics researchers are geographically distant and are often hard to schedule for observation. This poses a hurdle for user-centered design because these limitations often result in studies that consider only small numbers of tool users. Although small case studies can provide useful information about tool-use, they can be susceptible to sample bias without careful recruitment consideration. This is particularly true in cancer genomics, which has many distinct foci that use the same data (e.g., applied pharmaceutical vs. basic science research). As such, it is possible that relying on small population observations could cause iterative design decisions to overfit a tool to the requirements of a small number of users at the expense of a large, unstudied sub-population. If successful, interaction log classification would provide a way for understanding task requirements of entire populations in naturalistic settings, and would provide a way to help counter sample bias using the smaller scale, in-lab methodologies that tool evaluators already utilize.

4.2.1 Mouse interaction log schema

Our evaluation of interaction log classification focuses on analyzing mouse interaction logs collected on MAGI's gene set query results page. We provide an example query about the Notch pathway, which is implicated in a variety of cancers [175], in Figure 4.1. For each session, we collected all mouse events, information about each visualization's size and location, the window size, and anonymized information about the query. In addition to the five visualizations, we also collected the size and location of MAGI's control panel and tracked when tooltips were activated in each of MAGI's visualizations. Given that users can toggle visualization visibility, we also tracked how size and location of the visualizations might have differed over time. The full collection of log attributes is listed in Table 4.1.



	Dataset	Samples
🟦	BLCA	99
🟦	BRCA	772
🟩	COADREAD	224
🟩	GBM	291
🟥	HNSC	306
🟥	KIRC	417
🟧	LAML	196
🟧	LUAD	230
🟪	LUSC	178
🟪	OV	316
🟫	UCEC	248

Figure 4.1: A screenshot of MAGI showing the aberration matrix (top), heatmap (second top), network view (middle-left), transcript chart (middle-right), copy-number aberration browser (bottom), and control panel (right).

Type of information	Attributes
Mouse events	{click, move, scroll}, time, x, y
Tooltip events	x, y, width, height
MAGI components (×6)	x, y, width, height
Window state	width, height
Query	number of genes and datasets

Table 4.1: Data contained in each MAGI mouse trace interaction log. MAGI components refer to the five visualizations and control panel.

4.2.2 Log culling

We applied a two-step culling process to remove interaction logs that were unlikely to contain important information about visual analysis tasks. The first step in log culling involved the removal of logs without mouse interactions, which were created by web crawlers. This removal resulted in 1,616 logs with mouse event data. Afterwards, we then removed 63 logs that were deemed to have too few events to describe visual analysis tasks. For example, a user might realize that they typed in their query wrong and immediately navigate backward. While this scenario might provide important usability information about tool-use, it does not express information about the analytic intent of what the user hoped to accomplish. We defined "too few events" as any log with a mouse event count under the central 95% interval's lower bound. To compute the central 95% interval, we used an estimated lognormal distribution after visually analyzing the data's distribution with a quantile-quantile plot ($\mu = -71.99, \sigma = 773.38$, threshold=38.5 events).

4.3 Task identification with MAGI creators

Our first analysis of the MAGI interaction log data involved a free-text labeling task with two of the developers of MAGI. The purpose of this was twofold: (1) to pilot the feasibility of labeling analysis tasks from interactions alone, and (2) to derive a shortlist of categories, which could be used as classifier labels and as multiple choice options in our planned follow-up user study.

Here, we use "task" to refer to Gotz and Zhou's interaction characterization for visual analysis tools [46], which defines tasks and sub-tasks as "high-level, logical structures of a user's analytic process, such as the user's cognitive goals and sub-goals." For convenience, and due to their similarity, we refer to both as "task" for the remainder of the manuscript as their distinction is not critical for our present contributions.

4.3.1 Methods

Participants

Two participants remotely completed the free text log labeling task through screen sharing software. Each participant was involved with the development of MAGI and was familiar with MAGI's interface and the full range of ways MAGI could be interacted with.

Design and Displays

Instead of predefining a set number of interaction logs for participants to label, the experimental environment created trials on-demand by randomly sampling as many interaction logs as a participant could label within 45 minutes.

In each trial, an interaction log summary visualization was rendered alongside playback controls (Fig. 4.2). In the visualization, each of MAGI's charts were shown as a differently colored rectangle. A heatmap was overlaid on top of the visualization rectangles, which showed regions that users commonly interacted with. Participants could also watch the mouse move (orange crosshair) and tooltips appear (red rectangles) throughout the log's duration by either clicking a 10×-speed play button, or by dragging one of two sliders that controlled the playback time. The top slider was used to make large changes, and the bottom slider was used to fine-tune time navigation, which was useful for longer logs. Below the sliders we included a small timeline showing click, movement, and scroll events. Additionally, the number of genes and datasets in each MAGI query was shown above the interaction log visualization.

Procedure

Each participant was instructed to work with the experimenter to infer the predominant analytical task for as many interaction logs as possible within 45 minutes. For each log, the participant would brainstorm with the experimenter about what type of task the trial's interaction log depicted. Afterwards, the experimenter would write a 1-2 sentence description of the task and verify with the participant that the description summarized the brainstormed task. If there was no recognizable task, or if the task wasn't considered useful, the log would be labeled as "junk." After entering the log description, participants continued to the next trial.

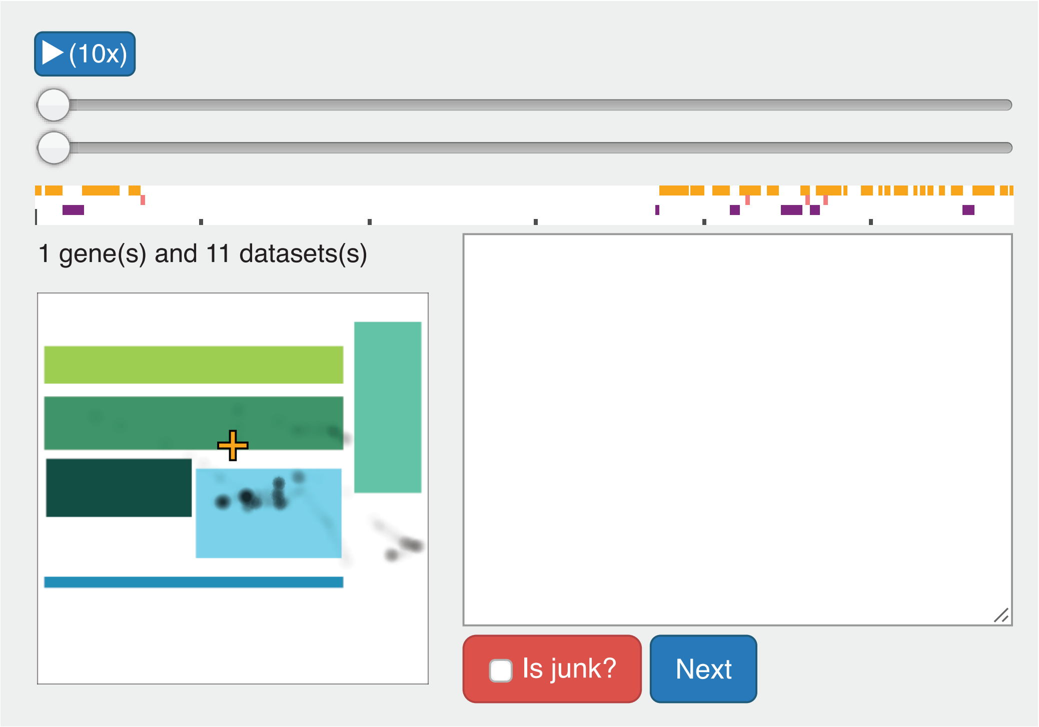


Figure 4.2: Example free-text label trial where participants were asked to provide a 1 to 2 sentence description of what type of task was performed in the visualized interaction log. Interaction logs were summarized in a visualization in each trial, which showed the location for each of MAGI's five visualizations in differently colored rectangles, and mouse activity with a black heatmap overlay. Users could watch the mouse and tooltips appear/disappear by using the playback button and two sliders to change time. The timeline below the sliders showed mouse movement (orange), click (red), and scroll (purple) events. Users could play the log by clicking on a 10× playback button or manually control playback with two sliders (top: whole-log, bottom: small adjustments to top).

4.3.2 Results and Discussion

We collected 50 labels in total (25/participant). Because we were interested in identifying a shortlist of commonly performed analytical tasks we then performed two rounds of manually grouping similar labels. To accomplish this, we printed out cards for each label response that contained the written description and accompanying interaction log visualization, along with a unique ID. Then, referencing the text summary for each card, we grouped similar cards in a manner similar to hierarchal clustering. After, we performed a second round of grouping to consolidate thematically similar groups. The resultant categories were as follows (see Fig. 4.1 for examples of each visualization):

Aberration matrix and transcript chart cross-referencing: Frequent back-and-forth analysis between the transcript chart and aberration matrix. For conciseness, we will refer to this task as "cross-referencing" unless otherwise noted.

All-encompassing or undirected browsing: Interactions with MAGI that appear undirected, that are typically diffuse, and that use many or all of MAGI's visualizations.

Co-occurrence or exclusivity analysis: Interactions that concern the aberration matrix, typically characterized by mousing over columns (co-occurrence) or exclusivity (staircases from column-exclusivity; Fig 4.1).

Copy-Number-Focused Analysis: Analysis characterized by heavy use of the copy-number aberration browser.

Junk: Logs that have no discernible analysis behavior (e.g., immediate page refresh after < 1 second or short, temporally distant bursts of movement).

Targeted gene, mutation, or annotation lookup: Targeted search behavior when a user has a specific piece of information they want to find (e.g., a particular patient-column in the aberration matrix).

Transcript mutation distribution analysis: If users interact with the transcript chart, they typically focus on certain distributional characteristics such as towers of mutations at a single point in the transcript ("hotspots") or at mutations that fall along coding regions.

Other: Behavior that falls outside of what was labeled in this experiment (e.g., use of the network view).

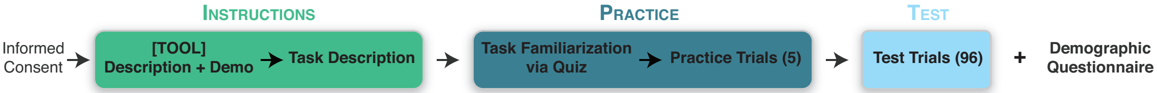


Figure 4.3: The procedure for our pair-participant task labeling study.

This procedure was guided by previous analyses that were part of MAGI's formative iterative design, which identified hypothesis formation and testing tasks targeted on biological significance as two of MAGI's largest use cases.

One question that arises from these results is how consistently these tasks can be inferred using only low-level interaction logging data, which is critical for reliable classification. We test this in the next study.

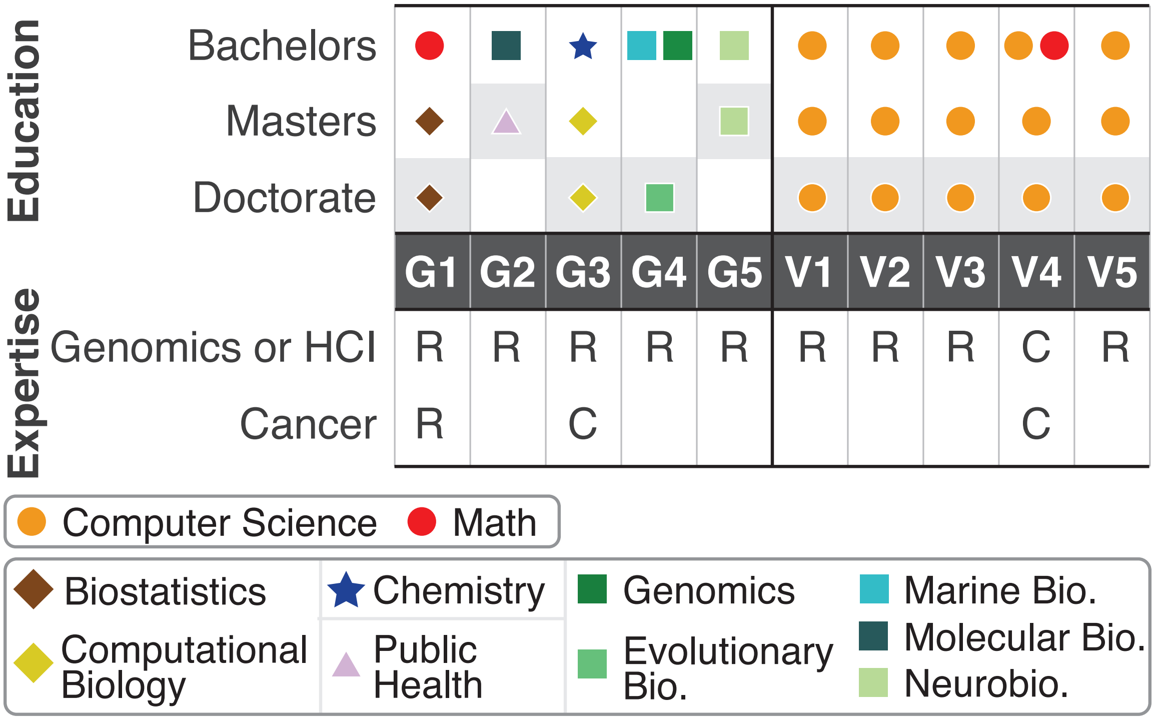
4.4 User Study: Log Task Labeling

The primary goal of this experiment was to collect labels to train, validate, and test interaction log classifiers. We also wanted to test whether humans could reliably infer analytical tasks from mouse interaction logs alone. Our prediction was that interaction-task inference would be reliable between interaction log observers. To these ends, we asked five pairs of visualization and genomics experts (1 of each/pair) to label tasks in a series of MAGI logs using the eight labels from our prior evaluation (Sec. 4.3).

4.4.1 Methods

Participants

10 participants (5 pairs) completed the study. Five subjects were recruited through university mailing lists for graduate students and had formal knowledge of genomics. The remaining five subjects were recruited from human-computer interaction research groups in our institution. Each subject had at least one year of academic or professional experience in either genomics or visualization. The median number of years each participant had spent in their degree program was 2 years (range: 0-5). Figure 4.4 shows participant expertise. There were equal numbers of male and female participants (3 female genomics experts, 2 female computer scientists) and the median age was 27 (range: 22-33).



		G1	G2	G3	G4	G5	V1	V2	V3	V4	V5
Education	Bachelors	●	■	★	■■	■	●	●	●	●●	●
	Masters	◆	▲	◆		■	●	●	●	●	●
	Doctorate	◆		◆	■		●	●	●	●	●
Expertise	Genomics or HCI	R	R	R	R	R	R	R	R	C	R
	Cancer	R		C						C	

● Computer Science
● Math
◆ Biostatistics
◆ Computational Biology
★ Chemistry
▲ Public Health
■ Genomics
■ Evolutionary Bio.
■ Marine Bio.
■ Molecular Bio.
■ Neurobio.

Figure 4.4: User study participant demographics. Non-circle degree glyphs relate to genomics expertise. Shaded cells mark currently-pursued degrees. “G” columns refer to genomics experts, whereas “V” refers to human-computer interaction (HCI) and/or visualization experts. “R” expertise entries refer to hands-on research experience, whereas “C” refers to coursework exposure.

Each was compensated \$10/hour. The experimental protocol was approved by our university’s IRB.

Design and Displays

The user study was held in pairs such that each session had one genomics expert and one visualization expert. The study was designed for pairs of participants rather than single participants because we believed pair coding would help control labeling variance and because the experiment required expert knowledge of visualization and genomics, which presented single-person recruitment limitations. Another motivation was that fatigue was too prohibitive in an earlier pilot study that tested single participants.

Each pair of participants saw 96 random-order trials, which consisted of 2 replications of a 48-trial design. One replication contained a unique set of interaction logs while the second replication contained logs that were identical between subjects to analyze inter-rater reliability (IRR). We settled on a 48-trial design after performing a power analysis for Fleiss’ kappa [157] ($\kappa_0 = 0.6, \kappa_1 = 0.4, \alpha =$

$0.05, \beta = 0.2$ with 5 raters), which suggested including at least 41 trials.

The 48-trial design consisted of 24 randomly sampled logs and another 24 logs that were sampled based on three feature sets we had planned to use in our eventual classification evaluation (Sec. 4.5.1). To sample the 24 feature-based trials, we first had a MAGI expert create example ground truths for each of the eight previously defined task labels, where we knew the full context of each query (e.g., "the expert was interested in exploring a particular biological pathway"). Then, using each of the three feature sets and eight ground truths, we sampled 24 nearby neighbors.

To create the six unique sets of logs (5 pairs + 1 IRR), we generated all of the feature-set-based trials at the same time by picking the 6-closest logs for each of the 24 {feature set} × {label} combinations. Next, we semi-randomly shuffled the samples so that each pair of participants would be given an unordered, complete collection of the 24 combinations. For example, the first participant would be given one of each 24 combinations, but these 24 logs would not always be the first-closest-neighbors. This procedure was designed to control for potential bias stemming from nearest-neighbor ordering while still including all 24 conditions.

The remaining 144 random-sample logs were then sampled without replacement from the set of remaining logs.

Procedure

Following informed consent, the study took place over three stages: instructions, practice, and test (Fig. 4.3). All participants took between 1.5 and 2 hours to complete the study.

Instructions: In the MAGI overview, each pair read through an overview that detailed each of MAGI's charts and saw a short demo of MAGI. In the study overview, participants were provided text descriptions for each of the task labels and were shown example stimuli.

Practice: The short quiz presented a grid of 8 example ground-truth logs at the same time along with the earlier text description of each task label. Participants were asked to discuss with their partner which label they believed should be assigned to each log. After guessing, participants could reveal the answer by clicking on a "show" button. Following the quiz, participants then completed five practice trials per the test procedure below.

Test: For each practice test trial, participants were provided a single log and were asked to mark which task label they thought was most characteristic of the log. If the participants selected the "other" category they were required to enter a short text description of the task. When satisfied with their response, participants then clicked the “done” button to advance to the next trial. Each trial included label descriptions and earlier quiz examples to the right of the experimental display and in printed handouts as reminders. To help participants finish within two hours, each trial displayed a timer and a beep would play after 45 seconds; however, participants could take as long as they needed to respond.

4.4.2 Results and Discussion

Inter-rater reliability and accuracy: similar strategies

Our planned-analysis of inter-rater reliability (IRR) using Fleiss’ κ was 0.405, which was calculated using the 48 IRR trial responses for each participant-pair. According to Landis and Koch, this maps onto fair-to-good reliability [84]. Fair-to-good reliability suggests that there was a moderate amount of subjectivity between pair responses, but that the individual differences across trials was low enough to be confident in the response reliability. To supplement Fleiss’ κ we also measured the *modal accuracy* of each participant, which defines a correct response as any response that matches the most frequently assigned label(s) for a given interaction log. Participant accuracies, in order of study completion date, were: 69%, 73%, 73%, 65%, and 77%. Both Fleiss’ κ and accuracies suggest that all participants had similar, consistent labeling strategies.

Task label diversity and frequency shows consistency

To understand participant-pair task labeling strategy similarity we analyzed labeling frequencies and labeling consistency across participant-pairs (Fig. 4.5).

To measure similarity we calculated Shannon diversity indexes for each pair-participant using label frequencies. The diversity indexes were 1.90, 1.97, 1.86, 1.76, and 1.91. Values closer to $\ln 8 \approx 2$ refer to more uniform label frequency distributions and values closer to 0 refer to skewed distributions with fewer labels and greater frequencies. Diversity indexes are calculated through Shannon entropy: $H' = -\sum_{i=1}^{L} p_i \ln p_i$. L is the number of labels and p_i is the ith label frequency’s proportion of the 96 total labels for a given participant-pair. Each diversity index fell within the top 15% of the potential range of diversity ($[0, \ln(8)]$), which suggests that all participants applied similarly uniform task labeling strategies. These results also support our initial task selection methodology because our synthesized task labels were used with little favoritism.

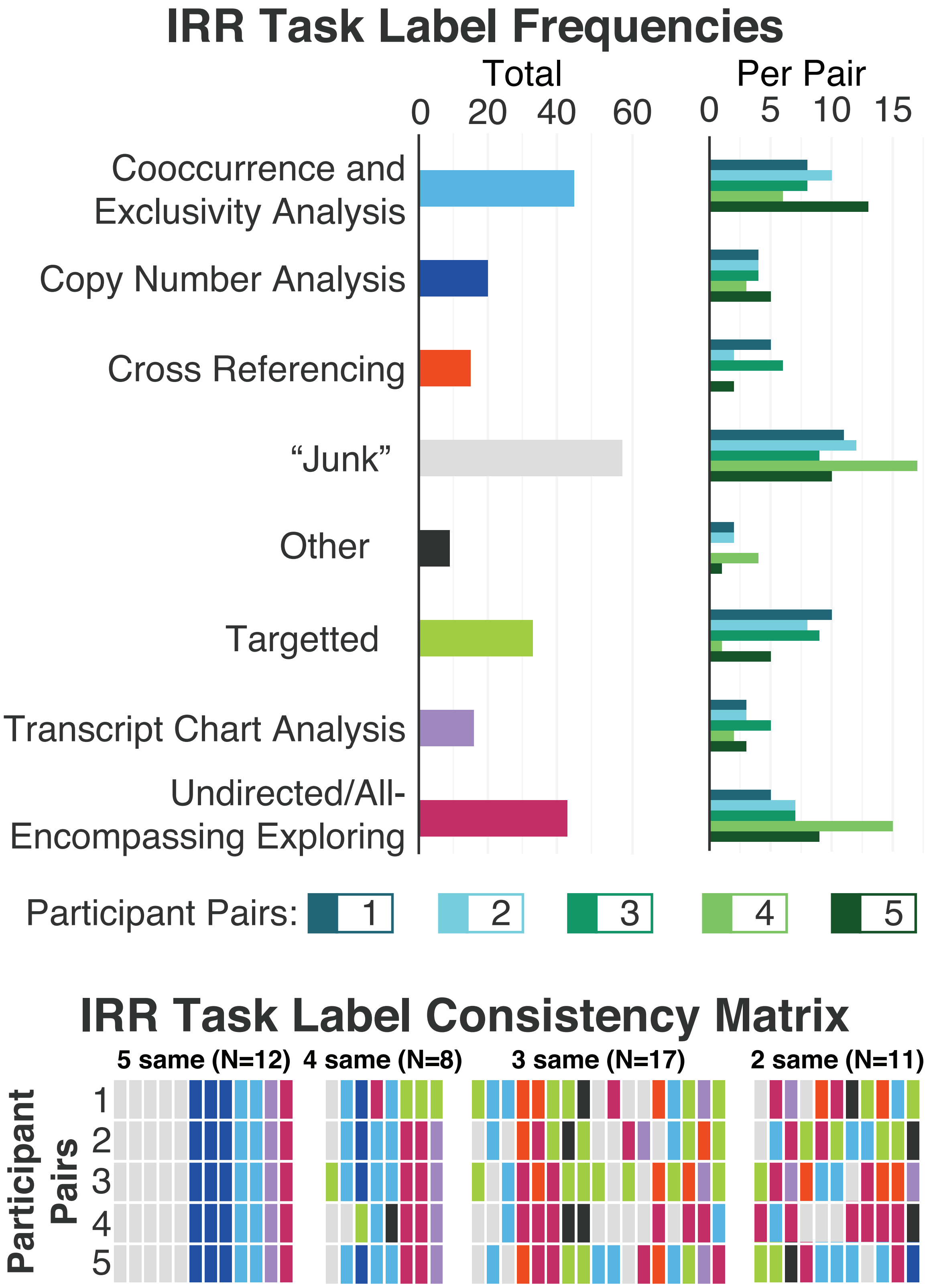


Figure 4.5: Task label frequencies (top) and ordered labeling consistency between participants for each interaction log (bottom; rows: participants, columns: interaction logs).

We also made several qualitative observations based on labeling frequency to drill down beyond reliability summary statistics. First, participant-pair 4's poor accuracy may stem from slightly-deviant labeling proportions: they never provided a cross-referencing task label IRR response, had only one targeted analysis response, and over half of their responses were either "junk" or undirected labels. This skew is the likely source for their comparatively lower accuracy and Shannon diversity index. Another distinction is that participant-pair 3 never provided an "other" response, though this is not necessarily abnormal given the relatively low "other" response rates of the other pairs. Aside from these two deviations, participants' strategies were largely consistent; 20 of the 48 IRR trials had 4 or 5 identical labels out of the 5 labels given by participant-pairs, and 17 IRR trials had 3 identical labels across the participant-pair responses. There were no trials where each pair provided different labels.

We also found no significant difference between modally correct labels between random and feature-based sampling methods through a two-sided Fisher's exact test ($p = 0.57$; feature-set: $73\%, 88/120$; random: $69\%, 83/120$).

"Other" label descriptions

There were 23 "other" labels ($< 5\%$) across the 480 total responses. The most frequent reason for selecting "other" was to report different types of cross-referencing task behavior (9), given that the provided cross-referencing task label only pertained to interactions between the aberration matrix and transcript chart. Other responses pertained to other MAGI features not covered by the 8 labels (e.g., the network visualization and control panel) (11), or to simple page exploration without analytic purpose (2). Only once did participants respond that they were unable to determine what type of task a user was pursuing.

These results support observations from our initial task identification: while it is possible that users will use MAGI for tasks other than the eight we identified, these other tasks are likely to be rare outliers. Similarly, the comparative scarcity of "other" responses suggests that our eight task categories were effective at describing the majority of MAGI interactions.

"Junk" assignment strategies

One concern we had while designing the experiment was whether participants would put potentially meaningful logs in "junk." Our intent was for junk to be a catch-all for logs that slipped past our prefiltering, which eliminated empty or near-empty logs. For example, there was one log that we would have considered to be undirected exploration due to its diffuse interactions; however, the pair of participants could not identify a behavior and marked it as junk (opposed to marking it as "other" as one other participant did). Although we saw some instances of undesirable junk labeling while proctoring the study, we found that participants were overall consistent with our junk-labeling expectations.

Takeaway: reliable, consistent human task inference

Overall, these quantitative and qualitative trends both point to similar conceptual understanding of how each task mapped onto mouse interactions and suggest that participant-pairs used similar labeling strategies. This is an important discovery because it shows that tool evaluators can reconstruct meaningful information about tool use from interaction logs alone. The reliability and presumed reproducibility of these findings establishes a foundation for our next evaluation. Using these results from our human-centered evaluation we can establish a baseline from which automated machine classification can be compared against.

4.5 Log-Task Classification

We evaluated 12 classifiers to test whether automated classification could predict visual analysis tasks with comparable accuracy to domain experts from the previous experiment. Each classifier was built from a selection of three models (k-nearest neighbors, linear support vector machines, random forests) and four feature sets, as described below. Our evaluation predictions focus on identifying a best-performing classifier to use in a follow-up exploratory analysis of the entire MAGI interaction log corpus. To test each model's effectiveness we used the 48 IRR trials from our previous in-lab experiment and used the non-IRR trials for training and cross-validation.

Our model selection was guided by selecting models that would be accessible to typical visualization tool developer/designers. We determined accessibility by how widely classification models were used in-practice and how readily they could be used "out of the box" with well-documented machine learning libraries (e.g., Python's `scikit-learn`). Another selection criterion was to select models that would perform well given few training data, which can be a common-place limitation in domain-expert-focused research. It is important to note that there are a number of important and

ROI Transition [19]	**Dwell** [4]	**Mouse Tracking** [99]
transition count	total time	stationary H
transitioned-to count	μ dwell time	transition H
	σ dwell time	total time $\forall$ ROI
	# datasets	active time $\forall$ ROI
	# genes	dwell time $\forall$ ROI
		μ active time $\forall$ ROI
		μ dwell time $\forall$ ROI

Table 4.2: An overview of three feature sets used in our classification (not shown: "all," the combination of these sets). ROI transition count is short-hand for the complete adjacency matrix of transition features between each ROI. Transitioned-to count sums one dimension of the complete matrix. μ: mean, σ: deviation, H: entropy.

immediately actionable research directions that could be pursued, which might result in more accurate predictions (Sec. 4.6). We opted to pursue simpler models for two reasons. First, we wanted to pursue a systematic approach to studying classifiers' potential use in evaluation, and thought their might be too great a number of unbound decisions given our present knowledge of interaction mining within visualization. Second, we wanted to focus our evaluation on models that would not be too elaborate for much of our target audience to easily use.

4.5.1 Feature Sets

In our present classification evaluation we consider three feature sets: dwell, region-of-interest (ROI) transition, and a novel "mouse tracking" approach. A summary of each feature set is listed in Table 4.2. In our present analysis of MAGI, "region of interest" (ROI) corresponds to MAGI's five visualizations and control panel (Fig. 4.1). For the remainder of the paper we will refer to these feature sets as dwell, ROI transition, and mouse tracking.

Dwell

The features in dwell are: total session time; mean and standard deviation of dwell time; and the number of datasets and genes in a query. Each feature is taken from a subset of Agichtein et al.'s features for modeling web search ranking [4]. We include only a subset due to differences in application areas and in interaction log schemas (multiple-page vs. single-page sessions).

One difficulty raised by the dwell feature set was how to best quantize mouse traces into active and dwell periods. To accomplish this, we chose a dwell threshold (100ms) using the interquartile mean of all contiguous-event time differences across all interaction logs. We operationalized the threshold using the interquartile mean opposed to other methods (e.g., median split) because the distribution of time differences had a long right tail that skewed whole-range averages. A common issue causing the skewed distribution were sessions where a user would leave MAGI open for days, whereas most differences were fractions of a second.

ROI Transition

The ROI transition feature set is comprised of the adjacency matrix describing transition frequencies between ROIs and the total number of transitions to each ROI. The two groups of features are adapted from Brown et al.'s features for modeling visual search task completion time and personality factors such as locus of control [19]. Although Brown et al. tested several predictive models, we use only their state-based feature set, which had the highest predictive accuracy for task completion time (83%).

Mouse Tracking

The mouse tracking feature set includes five types of times for each ROI and two types of entropy that measure how users transitioned between ROIs. The name "mouse tracking" alludes to its adaptation of eye tracking features.

The first three types of time included in mouse tracking are the total cumulative time spent in each ROI, the cumulative active time spent in each ROI, and the cumulative dwell time spent in each ROI. The last two times are the mean active and dwell times for each ROI. These measures are loosely inspired from distance-based region-of-interest analysis in historical scan path clustering analyses [5, 42], and were calculated with the same methods as the dwell feature set.

The other two mouse tracking features describe different kinds of entropy to summarize how users interacted with MAGI at a more global scale. Within the context of MAGI, entropy can be thought of as how deterministic a user's interactions are between ROIs (i.e., targeted vs. diffuse). To calculate entropy, we consider MAGI ROIs ($\Re$), the transition frequency probabilities between each ROI (M), and the stationary distribution of each ROI (π). The stationary distribution (i.e., the limiting probability distribution) represents the probability that the mouse will be over a given ROI at any point in time [188, p. 199]. Both entropies are based on Krejtz et al. scanpath classification methods [99].

The first measurement of entropy uses Shannon entropy to calculate whether the distribution of

Classifier	Feature Set	Parameters
k-nearest	All	$k = 9$, w=distance
k-nearest	Dwell	$k = 10$, w=uniform
k-nearest	ROI Transition	$k = 5$, w=distance
k-nearest	Mouse Tracking	$k = 7$, w=uniform
Linear SVM	All	$c = 69.519$
Linear SVM	Dwell	$c =< 0.001$
Linear SVM	ROI Transition	$c = 0.001$
Linear SVM	Mouse Tracking	$c = 0.004$
Random Forest	All	estimators=75
Random Forest	Dwell	estimators=40
Random Forest	ROI Transition	estimators=40
Random Forest	Mouse Tracking	estimators=40

Table 4.3: Parameter selection for each tested classifier. w: weight

ROI transitions is equal, where entropy values closer to 1 represent equal distributions and values closer to 0 represent focal distributions. Our use of $\log_{10}$ constrains entropy to a unit scale:

$$H_{Shannon} = -\sum_{i \in \Re} \pi_i \log \pi_i \tag{4.1}$$

The second measurement of entropy is similar, but also considers the transition frequency probabilities to understand whether interaction was more random (closer to 1) or more deterministic (closer to 0):

$$H_{Transition} = -\sum_{i \in \Re} \pi_i \sum_{j \in \Re} M_{ij} \log M_{ij} \tag{4.2}$$

All: Dwell + ROI Transition + Mouse Tracking

We also tested a composite “all” feature set, which combined the features from all three aforementioned sets.

4.5.2 Classification Evaluation Methods

Our final experimental design consisted of twelve classification models (3 classifiers × 4 feature sets), all of which were implemented in Python’s `scikit-learn`. For each model we performed an exhaustive search for all parameter combinations using 3-fold cross validation to select parameters. Parameter selections for each model are listed in Table 4.3. To examine predictive variance we evaluated each model fifty times using the same parameters across runs for each model.

4.5.3 Classification Evaluation Predictions

Before conducting the comparative classifier evaluation we made the following predictions:

P1 Random forest models would be more accurate compared to k-nearest neighbor and linear SVM accuracies.

P2 Mouse tracking features would be more accurate compared to dwell and ROI transitions for predicting task labels.

We predicted that random forests would be the most accurate because it was unclear whether our feature sets were linearly separable. Further, random forests provide a way to down-weight less effective features based on how their decision trees are trained, whereas k-nearest neighbors treats all features equally because it uses Euclidean distance. We predicted that mouse tracking would be the most accurate feature because it considered both time and transition, but at multiple levels of detail. In contrast, dwell considers only entire-session times and ignores regions of interest. Similarly, ROI transition focuses only on individual transitions, ignores more global descriptions of behavior, and does not consider interaction times.

4.5.4 Classification Evaluation Results and Discussion

Analysis of classifier performance

Because our test data has five “correct” labels for each interaction log (1 label/participant) we tested P1 and P2 with two types of accuracies: *match-any* and *modal* accuracy.

Match-any accuracy is calculated based on whether a classifier prediction matches any of the five labels provided by any of the participants and is a lower-bound measure of classifier performance.

Modal accuracy is the same accuracy that was used to measure participant accuracy in our previous user study: predictions are correct only if they match the most frequently assigned label(s) for each interaction log.

We used two accuracies — one loose and one strict — due to the qualitative, under-defined nature of what a “reasonably correct” prediction could be. It is important to note that the difference between the two accuracies is also meaningful: if match-any accuracy is 75% and modal accuracy is 50%, then 2/3 of the match-any-correct labels are also modally-correct responses and 1/3 are modally

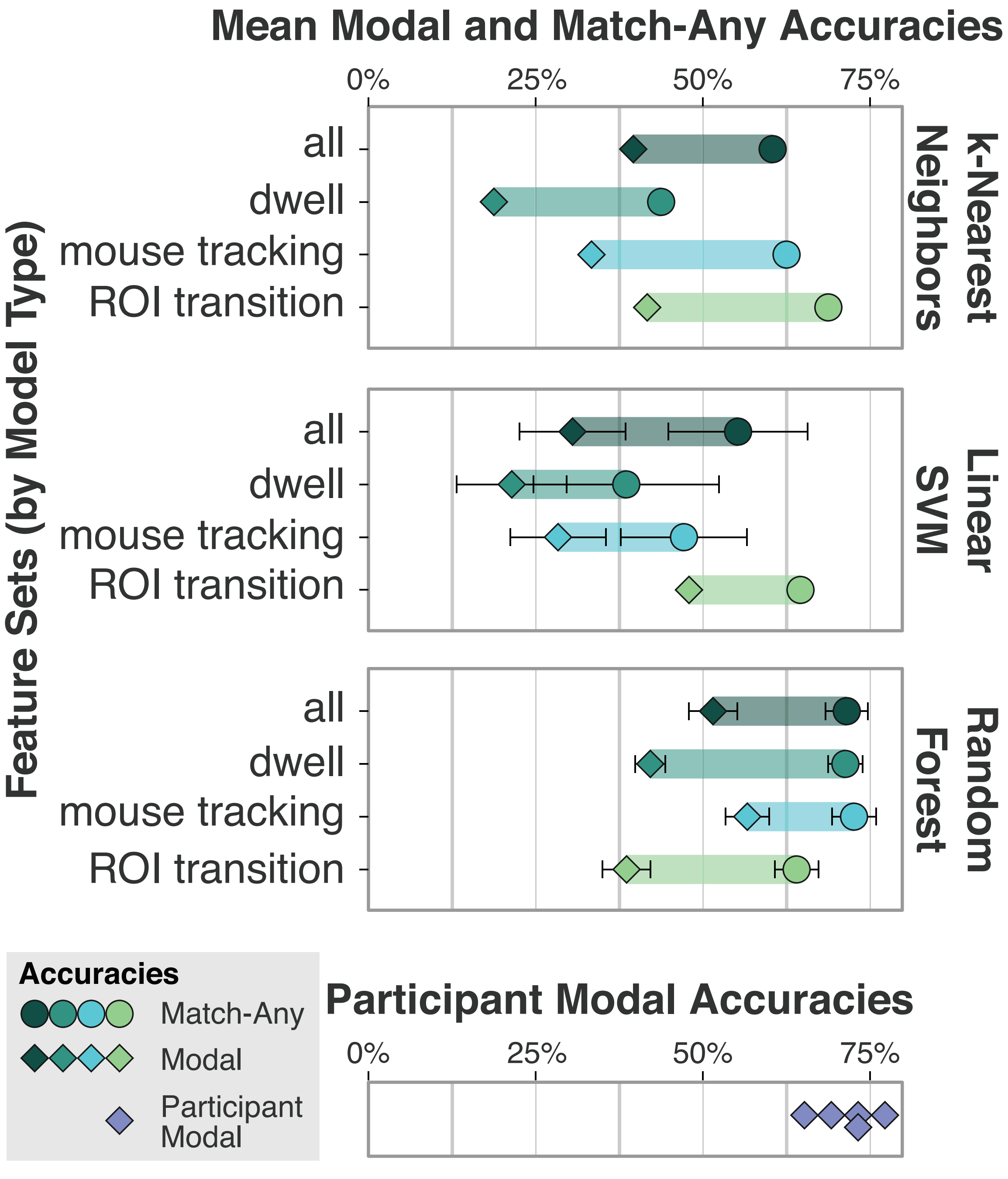


Figure 4.6: Means and standard deviations of classifier accuracies after running each model 50 times. Match-any accuracy is calculated based on whether predictions matched any label assigned to an interaction log by participants. Modal accuracy is calculated based on whether predictions match the most frequently assigned label(s) for an interaction log. Higher accuracies with smaller accuracy intervals are better. *k*-Nearest Neighbors has no standard deviation because successive runs will always select the same *k* shortest Euclidean-distance points. Below the three models we also include modal accuracies for each of the five participant-pairs for easier comparison (stacked glyphs represent multiple participants with the same accuracy).

incorrect responses. For this reason we planned our analyses to first examine match-any accuracy and use modal-accuracy as a mechanism to break match-any accuracy ties.

The twelve models' match-any accuracies ranged from 38% (linear SVM, dwell) to 73% (random forest, mouse tracking) and the modal accuracies ranged from 18% (k-nearest neighbors, dwell) to 56% (random forest, mouse tracking). The full-range of results are shown in Figure 4.6.

Previous visual analysis interaction modeling has achieved similar accuracies. For example, Brown et al.'s task completion time predictive models [19] had between 62% and 83% accuracy and their personality-attribute models had between 61% to 67% accuracy when testing for traits like locus of control and neuroticism. In comparison, our models were similarly accurate, but modeled a more complex and nuanced characterization of interaction (e.g., binary vs. octenary models).

Before testing our predictions, we first analyzed the variance of model type and feature set with respect to match-any accuracy, and found a significant main effect for each (model: $F(2, 588) = 483.74, p < 0.001$; feature: $F(3, 588) = 164.39, p < 0.001$) as well as a significant interaction between the two ($F(6, 588) = 95.53, p < 0.001$).

The significant interaction between model type and feature set likely refers to the dissimilarities in accuracy for k-nearest neighbors and linear SVM models compared to random forest models. Match-any accuracy across model types was largely fixed for ROI transition features and varied for the other three such that ROI transition features were most-accurate for k-nearest neighbors and linear SVM models and were least-accurate for random forests. This suggests that dwell and mouse tracking are not linearly separable and, for similar reasons, are not well-suited for simple Euclidean-distance-based classification models. The lack of separability is supported by close-to-zero SVM margin parameter selections, which suggests that across all feature sets, the data was too noisy to define a hyperplane that cleanly separated data. It would be interesting to test whether certain subsets of data are more easily separated as a way to achieve better performance; however, such analysis falls outside the present comparative model analysis goals.

To better understand the performance differences between model types and feature sets we systematically tested our planned predictions for match-any accuracy using 2-sample Welch's t-tests. We first tested match-any accuracy by model type (P1) and found that random forests were significantly better than both k-nearest neighbor ($t(291.26) = 15.03, p < 0.001$) and linear SVM models ($t(242.83) = 17.99, p < 0.001$), and also that k-nearest neighbor models were better than linear SVM models ($t(348.73) = 6.37, p < 0.001$). Thus, random forests were best, followed by

k-nearest neighbors and then by linear SVM models.

After finding that random forests were the most match-any accurate classifiers, we then tested whether mouse tracking was the most accurate feature set (P2) using only random forest predictions. Our second prediction partially held: mouse tracking was significantly more accurate than dwell ($t(93.01) = 2.17, p = 0.03$) and ROI transition ($t(97.99) = 13.09, p < 0.001$), but was not significantly different compared to all ($t(97.83) = 1.68, p = 0.1$). The non-significant difference between all and mouse tracking may suggest that "all" accuracy primarily stems from mouse tracking and has nearly no benefit from dwell and ROI transition features. Another important result was that the ROI transition feature set performed significantly worse than the three other feature sets (all: $t(97.91) = 11.67, p < 0.001$; dwell: $t(93.43) = 12.35, p = 0$; ROI transition: $t(97.99) = 13.09, p < 0.001$).

While random forest mouse tracking classifiers were significantly more match-all accurate compared to the other random forest classifiers, we also compared modal accuracies due to the small in-practice accuracy differences between all, dwell, and mouse tracking features (Fig. 4.6). As before, mouse tracking was significantly more modally accurate than dwell ($t(86.93) = 25.97, p < 0.001$) and ROI transition ($t(96.99) = 26.29, p < 0.001$), and was also significantly more modally accurate than "all" ($t(96.88) = 7.45, p < 0.001$). Although "all" includes mouse tracking features, mouse tracking may have performed better because the ROI transition and dwell features could have been maladaptive for predicting modal task labels.

Thus, these analyses indicate that random forest mouse tracking classification models were best.

Binary classification: detecting visual analysis

One remaining question after comparing model accuracies was whether certain task labels were more difficult to predict than others. The previous analysis provided overall model accuracies compared to expert-coded "groundtruth," but did not elaborate on why model accuracies differ. Unfortunately, answering "why" is challenging with our present results because of the number of labels. Therefore, we framed our analysis of why accuracies might differ based on how easy it was for the classifiers to detect the presence vs. absence of visual analysis tasks. Rather than consider 8 labels, classifiers that use this simplified task/no-task representation need only consider two. We tested task/no-task classification accuracy by retaining "junk"-label predictions as "no-task" labels and by transforming the rest to "task-present" labels. If accuracies across the twelve binary models were to be universally higher, it would signify that it is easier to distinguish whether there was salient visual analysis

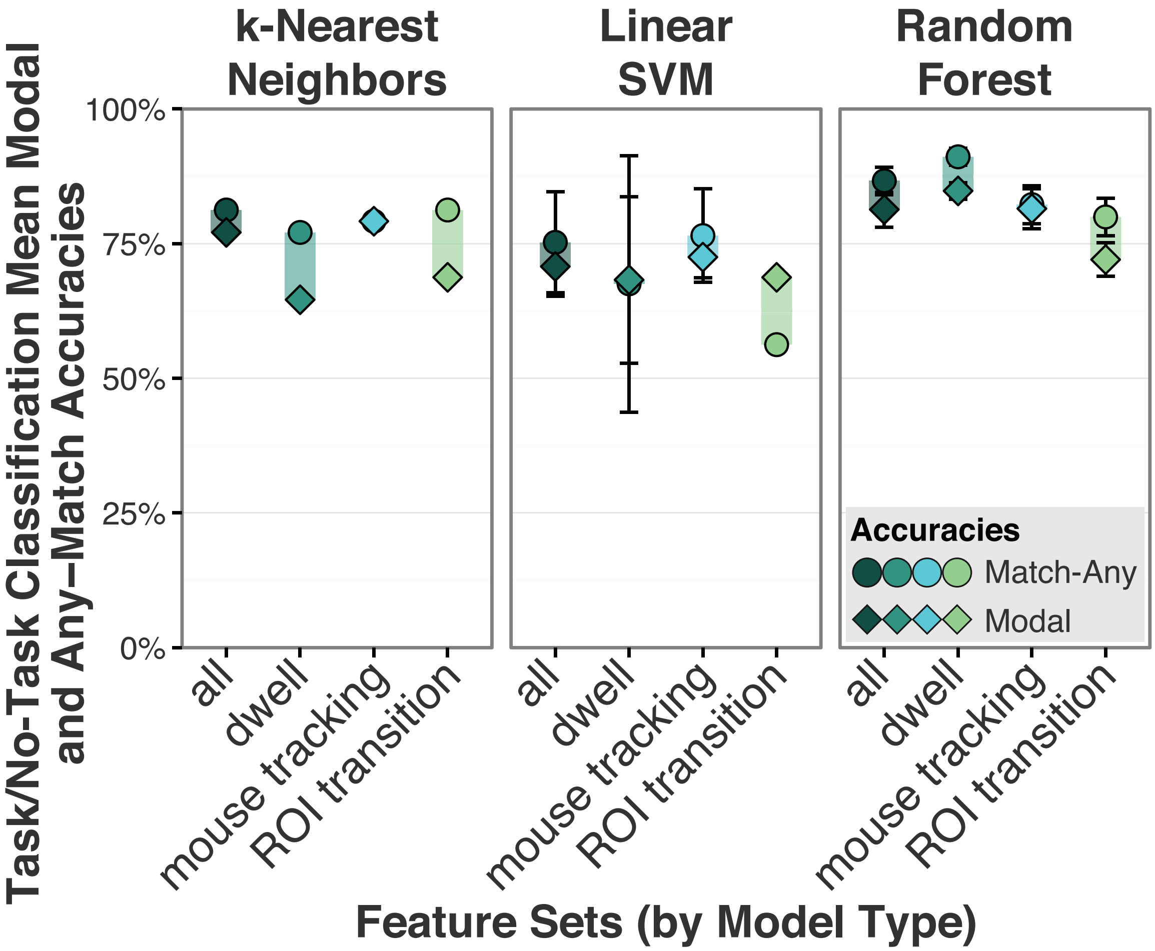


Figure 4.7: Means and standard deviations of task/no-task classifier accuracies after running each model 50 times. Predictions were taken by transforming the earlier multi-class predictions into binary junk vs. non-junk categories.

compared to differentiating what specific visual analysis task a user was undertaking. We predicted that

P3 Binary task-present/no-task classification would result in higher accuracies.

We based P3 on qualitative inferences that “junk”-labeled logs generally have different looking mouse trails compared to the other seven labels. For example, it is easier to differentiate an empty log from one with lengthy interaction sequences, but it may be much harder to identify whether a lengthy interaction sequence depicts undirected exploration or cross-referencing tasks.

We report both match-any and modal binary classification accuracies in Figure 4.7. As predicted (P3), ranges for match-any and modal accuracies were both higher (match-any: 56%–91%; modal: 65%–85%). Random forest mouse tracking classifiers had the same match-any accuracy as modal accuracy (82%). The best performing task-present/no-task classifier was random forest dwell, which had both the best match-any accuracy (91%) and a modal accuracy (85%).

The smaller task-present/no-task accuracy intervals between match-any and modal accuracies compared to octenary classification suggests that most of octenary modal error was due to error between non-junk labels opposed to confusion between the “junk” label vs. other labels (P3). For example, random forest mouse tracking classification had no difference between accuracies in task-present/no-task classification unlike in octenary classification. This difference in labeling confusion between task-present/no-task and octenary classification is an important distinction because it means that both binary and multi-class classifiers can be used as a method for pruning uninteresting interaction logs that lack visual analysis tasks.

The support for P3 also suggest that it is more difficult to differentiate visual analysis tasks from one another opposed to deciding whether an interaction log contains a visual analysis task. We qualitatively validated this by visually exploring predicted “junk” labels and found that most histories showed short or otherwise sparse interactions compared to more lengthy or short, but consecutive, sequences of mouse events. Most often we found that no-task “junk” logs contained interactions indicative of user-error such as “quickbacks:” logs where users immediately navigated backward. In contrast, the other labels were often associated with longer-duration logs with greater numbers of events, which creates a separable boundary between the “junk” and non-junk labels.

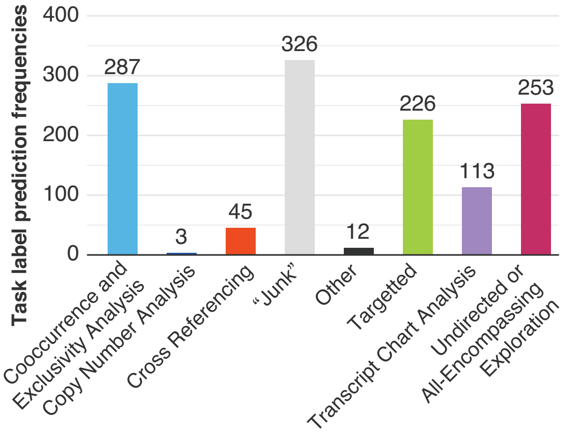


Figure 4.8: Distribution of predicted task labels for the 1,267 logs that were not included in our in-lab labeling study using a random forest classifier and "mouse tracking" feature set.

4.5.5 Exploring possible classification benefits to design

In this section we explore several possible ways that automated visual analysis task classifiers can improve the iterative design process. Our aim is to provide insight about how MAGI is used, to identify how this insight can be incorporated into iterative design, and to enumerate testable hypotheses about cancer genomics visualization interaction, which can be used to inform future design studies. Our discussion is based on exploratory analysis after using random forest mouse tracking classification to predict analysis tasks for the remaining 1,267 logs that were not part of our prior in-lab study (Sec. 4.4). While interpretation of these results is limited by a lack of ground-truth, our previous analyses show that task/no-task separation, and therefore comparison, is reliable. Additionally, we can be sufficiently confident in comparisons where there are large label-count differences given classification error rates.

Prediction results are shown in Figure 4.8. Junk labels were the most common (326) followed by cooccurrence and exclusivity analysis (287), undirected or all-encompassing exploration (253), and targeted analysis (226). The other tasks were assigned smaller label amounts: copy number analysis (2), other (12), cross referencing (45), and transcript chart analysis (113).

Understanding behavior via interaction frequency

Figure 4.8 shows that the aberration matrix was interacted with most frequently compared to the other visualizations. This information provides several testable hypotheses about user behavior that can be used to inform future iterative design decisions. One possibility is that most researchers use MAGI to test co-occurrence and exclusivity predictions and therefore use the aberration matrix more than the other features of MAGI. Another possibility is that the aberration matrix is used most frequently because its spatial positioning at the top of MAGI causes an availability or similar spatial cognitive bias since it is the first chart users see on the query. Or, it could be that the aberration matrix is used most often because of a combination of the two other possibilities. These classification-based hypotheses lend themselves naturally to established iterative design evaluation methodologies such as A/B testing, which could help MAGI designers understand whether the spatial positioning of the aberration matrix is a large factor for its frequent use.

This location proximity effect might also be supported by the comparatively low interaction frequencies associated with the copy number analysis task, which is located at the bottom of the page. One possible explanation for this difference is that it may be partially subsumed by the interactions in targeted look up or undirected or all-encompassing exploration given that the mouse would need to move to access the visualization; however, anecdotally, we did not find that a large number of interaction logs in those two labels displayed copy number exclusive behavior.

Which exploration strategy is more common: Top-down or bottom-up?

Two common visualization design heuristics are to support either top-down or bottom-up exploration. Top-down strategies refers to Ben Schneiderman's popular tool design mantra: "overview first, zoom and filter, then details-on-demand" [173]. In contrast, bottom-up strategies refer to diving into details first: "search, show context, [then] expand on demand" [196]. This is a critical point for tool design because supporting detail-oriented, bottom-up exploration can often be at odds with supporting top-down exploration.

The predictive classification results show it is likely that cancer genomics researchers use MAGI for both top-down analysis tasks (e.g., "undirected exploration") in similar proportion to bottom-up strategies (e.g., "targeted search"). This is of interest because typical visualization design patterns maintain that it is best to focus on dominant tool use patterns (e.g., Ziemkiewicz et al.'s evaluation of immunobiology visualization [209]). Given this pattern, If MAGI were to be iteratively designed using only typical in-lab methods with small samples, and evaluators found similar equivalence, the results might seem suspect. In contrast, if designers found evidence of only one search behavior they might be likely to use a representativeness heuristic and assume MAGI is only used in one way given typical design/evaluation perspectives.

However, through classifying our large collection of online tool logs, we avoid these pitfalls and can see bottom-up and top-down use of MAGI are near-equal in occurrence. As such, both exploration procedures should be supported in future design iterations. This raises an important design question given the previous research by Ziemkiewicz et al., who suggested that visual analysis tools that seek to support all analysis behavior may lead to substandard designs [209]. What, then, is the best strategy for supporting tasks that are equally as common without creating two separate tools?

To address this open research question in the design of MAGI, we implemented and deployed a new resizable and repositionable layout so that researchers can alter MAGI's components to better match their individual requirements.

Can classification counter incorrect generalization?

The predictive classification results were in many ways a surprise to us given past observations of MAGI, which led us to expect that cross-referencing was a common and important task requirement; however, our modeling suggests that this might not be true. The surprise that our prior observations did not generalize to the larger collection of interaction logs is an example of how bias can affect experimental analysis, which we also discuss in Section 4.5.5 with respect to overfitting search task support. In particular, our revelation about cross-referencing task frequency highlights how human tendency to use a representativeness heuristic when generalizing information [86] can be maladaptive in design evaluation. The difficulty lies in the fact that most design studies typically rely on field studies with small populations [20]. Because humans tend to generalize through representativeness, which does not take sample size into consideration, it means that evaluators are likely to overfit task requirements if they do not take extraordinary care. Based on our present findings we hypothesize that interaction log classification can benefit iterative design and task requirement analysis by helping counter sample biases in tool evaluation by showing the distribution of tasks for larger sample sizes than what it typically attainable through in-person observation. By showing an alternative hypothesis to tool evaluators, it is possible that evaluators could not just avoid jumping to incorrect conclusions about tool use, but they might also make new discoveries about qualitative in-person observations.

For this reason, we suggest supplementing in-lab observation with interaction log analysis of how a tool is used remotely by a larger sample of users. By using both, designers can make detailed predictions with in-lab observations and better identify potential sources of bias by consulting the interaction logs of larger tool-use samples. Pursuing this mixed-method evaluation design would preserve the realism of field observations while also affording designers greater generalizability confidence.

We advocate for a mixed evaluation approach given the relative strengths and weaknesses of classification compared to historical qualitative inquiry. Classification can critically serve as a tool to test in-lab ecological validity, and with the right data can paint a comprehensive picture of the types of tasks a tool is most used for across its entire user population. We believe this will be even more true if our simple classification approaches are replaced with more advanced approaches discussed in Section 4.6. However, knowledge acquisition from such automated approaches is inherently limited by the lack of context of interaction logs. Environmental factors, true ground truth, and the cognitive state cannot be known — only inferred. In contrast, these shortcomings are what talk-out-loud qualitative methods excel at collecting. Hence, even with the advent of more powerful task data mining techniques, we believe qualitative evaluation will remain an equally as valuable, rather than replaceable, aspect of tool design. The strength of classification should therefore not be tested by whether such approaches can serve as a replacement to qualitative inquiry, but rather how they can supplement it.

4.6 Broader impact and remaining questions

4.6.1 Generalizability of contributions

Our evaluation findings demonstrate that the usefulness of mouse interaction analysis generalizes from text-dominant applications, like search, to more open-ended and unstructured scenarios that are characteristic of visual analysis applications. We also note that while our evaluation used the study of MAGI as a case study, our contributions are not limited in domain, and could be used irrespective of application area.

4.6.2 More accurate modeling may result in different types of generalizability and implications for design

While classification error is low enough in our evaluation to infer user behavior and possible design implications, the development of more accurate classification could lead to more precise predictions and discussion about the relation between tasks and effective design. One potential way to achieve higher accuracy is to include an explicit feature selection step in future task classification pipelines. An alternative potential approach to increase accuracy is to model tasks as mixtures. For example, mixture models would break away from modeling only the most dominant session interactions, and could provide more robust understanding of likely-heterogeneous tasks such as "all-encompassing exploration."

4.6.3 Can unsupervised learning achieve comparable accuracy?

Our present evaluation only considers supervised learning approaches, which leaves the potential effectiveness of unsupervised approaches an open problem. This open problem can be tested in the future by evaluating whether clustering based on geometric-temporal distances of interaction segments [118] can accurately predict visual analysis tasks. However, one barrier to this approach, which must also be examined, is how to best segment interaction logs into discrete components that accurately represent stages of visual analysis. While it is possible that segmentation could be skipped, it is unlikely that clustering would produce accurate results without it because of the large geometric-temporal variability of entire minutes-long mouse movement between users. One benefit to clustering, opposed to classification, is that the phylogenies produced by hierarchical approaches could be used to test the quality of existing theoretical interaction taxonomies that are either based on literature surveys or qualitative observation.

4.7 Conclusion

Our findings illustrate the potential utility of mouse interaction log analysis as a new method for analyzing typically hard-to-access domain expert populations.

Using 1,553 interaction logs of MAGI, an online cancer genomics visualization tool, we first showed through in-lab evaluation that low-level interaction data alone is sufficient for reliable task inference. We then discussed how accessible classification methods matched our in-lab study inferrences with up to 73% accuracy and could separate interaction logs with visual analysis tasks from those without with up to 91% accuracy. Unlike previous interaction log analysis research, our investigation considered whether interactions could be inferred by humans and machines from mouse event data opposed to higher level representations of interaction that explicitly contain richer semantic information.

We advocate that domain expert tool evaluation can be improved by combining contextually-rich qualitative observation with larger-scale interaction log analysis. By leveraging a mixed-methods approach, tool designers can retain a deep understanding of the environment that their tool is used in and the analytical goals their tool is used to achieve; they can then test specific task-based predictions based on qualitative observation by analyzing interaction logs of larger population samples to assess the ecological validity of their in-lab findings.

Chapter 5

The relation between visualization size, grouping, and user performance

A common goal when designing visualizations to support certain tasks is to consider how stylistic choices might affect the legibility of information. Within the context of MAGI, for example, how might increasing the number of samples in the aberration matrix – which decreases the horizontal width of each column – affect a researcher's ability to interpret information? The main contribution of this chapter is the evaluation of how various types of visualization "size" can affect visual search performance, which in turn is used to inform several guidelines to promote visualization design legibility.

Our work is part of a broader effort to establish a theoretical definition of "effective visualization design" by understanding how principles from visual perception can be applied to visualization [9, 12, 60, 70, 74, 96, 104, 186, 192, 206]. By establishing theoretical foundations such as how chart height affects slope comparisons [185], we pave the way for future automated visualization style design based on human-centered analysis considerations. For example, our later chapter on Colorgorical (Chapter 6) applies previous theoretical foundations in color to create a tool that allows users without

This chapter is an extended version of a work that originally appeared in the Proceedings of Information Visualization 2014 and in IEEE's *Transactions on Visualization and Computer Graphics* as [48].

design expertise to control the balance between color discriminability and preference.

Establishing the theoretical groundwork for such design recommenders is critical given the increasing numbers of visualization design novices with the popularization of easy-to-use charting software like SPSS, Excel, and Tableau. While on one hand these tools can be empowering, on the other, design novices can easily create hard-to-read or misleading charts by making uninformed design decisions (e.g., rainbow color maps [10, 153]).

In our evaluation of visualization "size" and search performance, we were interested in understanding how styling "size" could improve visualization tools that are used for rapid serial viewing, in which search speed is especially important. For example, cancer genomics researchers typically search through many charts using file system windows (Chapter 3) as quickly as possible when forming hypotheses about certain genetic mutations. Improving target detection speed may help these researchers spend less time weeding through data and more time on advancing our understanding of cancer.

We tested this relation through two experiments. In the first experiment, we tested participants' visual search performance when asked to find a uniquely colored square in a field of distractors. In the second experiment, we tested search performance using the same task but with scatterplots.

One way that we apply the empirical relation between size and search performance to design is through the creation of guidelines: Each guideline offers a method for visualization designers to determine informed, scenario-specific balances between how much information is shown in a chart while sustaining an acceptable level of legibility.

To better establish how these guidelines may map on to situation-specific visual analysis tasks in cancer genomics research we also conducted a NASA TLX cognitive workload experiment with three cancer genomics researchers.

Across these experiments, we make the following primary contributions:

- We describe how the grouping, quantity, and size of visual marks affects search time based on the results from two experiments
- We report how search performance relates to self-reported difficulty in finding the target for different display types
- We present design guidelines based on our findings to facilitate the design of effective visualizations

- We present the results of a NASA TLX cognitive workload study that examined how size affected the use of aberration matrix cancer genomics visualizations

In addition, we report the results of a multiple linear regression model constructed from stimulus parameters, which explains 89% of the variance in response times from searching though grids (Experiment 1). This model generalizes to response times from searching through scatterplots in Experiment 2 (86% of the variance explained).

5.1 Related Work

The experiments described in this chapter contribute to the study of *graphical perception* – how visualization usability is affected by visual attributes like grouping by color similarity, shape, and size [25]. This section surveys the prior literature on graphical perception that forms the basis for our research on size and grouping.

According to Eick and Karr, seven categories of scalability issues arise in data visualization: human perception, monitor resolution, visual metaphors, interactivity, data structures and algorithms, and computational infrastructure [31]. Our work lies in their human perception and monitor resolution categories. Within the category of size perception, we define three subcategories: 1) *scale*, the physical size of elements (i.e., zoom level); 2) *quantity*, the number of elements; and 3) *aspect ratio*, scaling one dimension to shrink or expand elements. Each of these subcategories pertains both to individual marks and whole visualizations.

The size of marks in visualizations has substantial effects on performance [178]. Studies of how visual scale (i.e., zooming) influences user performance often focus on tasks involving navigational maps. Work in this area dates back to cartographic research, predating information visualization. For instance, Enoch found that visual search performance had steeper performance declines based on visual angle when map size was 9° or less, compared to a shallower performance difference when map size was greater than 9° [32]. More recently, Jakobsen and Hornbæk compared user navigation performance when maps were displayed on monitors of varying sizes, causing the maps' visual angle to vary across displays [154]. Participants were asked to complete map-based navigation tasks across various zoom levels. Performance was similar for users with medium-sized and large monitors, but was better for those with larger monitors than with small monitors. This was true even after controlling the quantity of information displayed [85]. They report dissimilar findings from Yost and North, who varied the number of elements relative to the monitor size and found no effect on normalized performance time [206]. Jakobsen and Hornbæk suggest that the difference might be due to variance in task difficulty.

A large body of literature in the psychology of attention reports how the quantity of elements in visual displays influences people's ability to find targets. Treisman and Gelade found that the quantity of distractor elements had differential effects on search time, depending on the relation between the visual features of the targets and distractors [194]. If a target (e.g., blue circle) differs from a homogeneous set of distractor elements (e.g., red circles) on a single feature (e.g., color), the number of elements has little to no effect on search performance. Visual search under such conditions is considered to be preattentive, where all the elements are surveyed in parallel and the target "pops out" (i.e., parallel visual search). If the target differs from a heterogeneous distractor set on multiple features (e.g., a blue circle target among red circle and blue square distractors), visual search is serial – people must exhaustively search all elements until they find a target. Parallel search is marked by reaction time functions that have little to no slope as distractor set increases, whereas serial search is marked by reaction times that follow robust positive slopes over set size. This distinction is useful in evaluating users' ability to "automatically" find target information in visualizations, given the display parameters.

Further, visual search is more difficult when: 1) distractors more closely resemble possible targets and 2) distractors have higher variability in visual appearance [29]. This difficulty due to increased distractor variability is consistent with the claim that decreased coherence or order in a visualization impairs performance [65, 98]. Additional evidence from studies using node-link diagrams and adjacency matrices also indicate that response time increases as the set size and data density increases [40]. Our study builds on these results by: 1) looking at a greater range and total number of set sizes, and 2) investigating how set size could interact with grouping by color similarity [200] and the size of marks.

Relating to this work, Haroz and Whitney provide visualization design guidelines based on how color variability (i.e., the number of colors) and grouping affected the ability to find a target in a grid of colored squares [60]. They found that participants were faster at finding targets in displays where marks were grouped by color rather than randomly distributed. Adding additional color variability to displays had little affect for grouped displays. However, the affect of adding color variability for random displays depended on whether the target type was known before the start of each trial.

If the target was unknown ("odd ball" task), performance slowed substantially as color variability increased, whereas the performance decay was minor if the target was known. Unlike Haroz and Whitney who focus on color variety and grouping, we investigated how effects of grouped vs. random layouts influence performance as the size and quantity of marks increased. We predicted that the minor difference in search time for grouped and random layouts found by Haroz and Whitney for grids of 64 elements would increase dramatically as the number of elements increased.

Wolfe provides a survey on many other important visual search considerations when detailing his "Guided Search 2.0" model [202]. Perhaps most relevant to this work, Wolfe discusses how the density of marks influences search performance. For instance, greater density facilitates search performance when the target type is unknown, but has little effect when the target type is known [13]. Related to density, Palmer notes that set size can have a varying effect on performance due to numerous other related factors such as eccentricity [137]. Our first experiment varies total display size with mark size as spacing between marks was kept fixed across all conditions, however we have provided a view of our results that highlights the relation between total display size and response time (Figure 5.4). Our second experiment has a fixed display size for all trials. The present study adds to our knowledge of how search factors such as set size can affect task performance; however, as Wolfe and Palmer have shown, there are many remaining factors that information visualization researchers can use to study performance.

Further research has examined how constraints in the visual system affect how observers interpret scatterplots. Gleicher et al. show that users can effectively compare average values in multiclass scatterplots even with dissimilar number of points between classes, additional distractor classes, and with conflicting cues [41]. Fink et al. take a complementary approach to improving scatterplot efficacy [36]. They found that their method for selecting scatterplot aspect ratio, based on Delaunay triangulation, improved the accuracy of correlation and cluster detection within scatterplots. Where Gleicher et al. examined value comparison in scatterplots and Fink et al. examined aspect ratios, our study examines how the number and size of marks influences visual search performance.

Studies have also revealed that the aspect ratio of graphical elements affects user performance. Looking at individual rectangles, Heer and Bostock found that people were more accurate at comparing the area of two rectangles when they departed from a 1:1 aspect ratio [70], although Kong et al. found that performance was also poor for extreme aspect ratios [96]. Looking at whole graphs, Talbot et al. found that the aspect ratio of line charts influenced people's ability to compare slopes of lines [185]. Participants had more difficulty comparing two large slopes than two shallow slopes; however, reducing chart height to reduce the physical angle of the two lines improved accuracy. In contrast, Heer et al. found that people are better at comparing values in horizon graphs – a type of time series visualization – with taller graphs rather than shorter ones [74]. Heer and Bostock found similar results when looking at bar chart height, and further found that benefits of increasing height plateaued with successively greater height increments [70]. Taken together, these studies suggest that when the goal is to compare angles, visualizations should be shorter, and when the goal is to compare area, visualizations should be taller.

An often challenging part of graphical perception research is designing experiments that capture the complexity of real-world information visualizations. In an effort to improve the ability to capture and account for such complexity in full, Rosenholtz et al. show how they were able to use computational approaches to assess grouping in design and demonstrate how their computational results relate to traditional design rules [156]. We believe work such as this can provide the foundation for creating computational techniques that give designers indicators when it would be useful to apply certain guidelines discovered from graphical perception research.

5.2 Experiment 1: Searching through grids

In this experiment, we studied how visual mark size, the number of visual marks (set size), and the color layout (grouping) influence the time taken to find a known target in a grid of squares. In the experiment, participants were presented with colored grids (Figure 5.2) and were asked to indicate which quadrant contained the purple target. This task is similar to Haroz and Whitney's "Find a Known Target" task, except they varied the target color across trials and their participants indicated whether a known target was present/absent (without reporting its location) [60]. We fixed the target color and used the quadrant localization task because the types of everyday search tasks we are interested in optimizing involve localizing a single target type. For instance, cancer genomics researchers routinely try to localize specific mutations in many types of visualizations such as transcript charts and various distribution plots.

We note that although the use of response time as a dependent measure in visualization research is controversial [79], it is an appropriate measure for our present objectives. We are most concerned

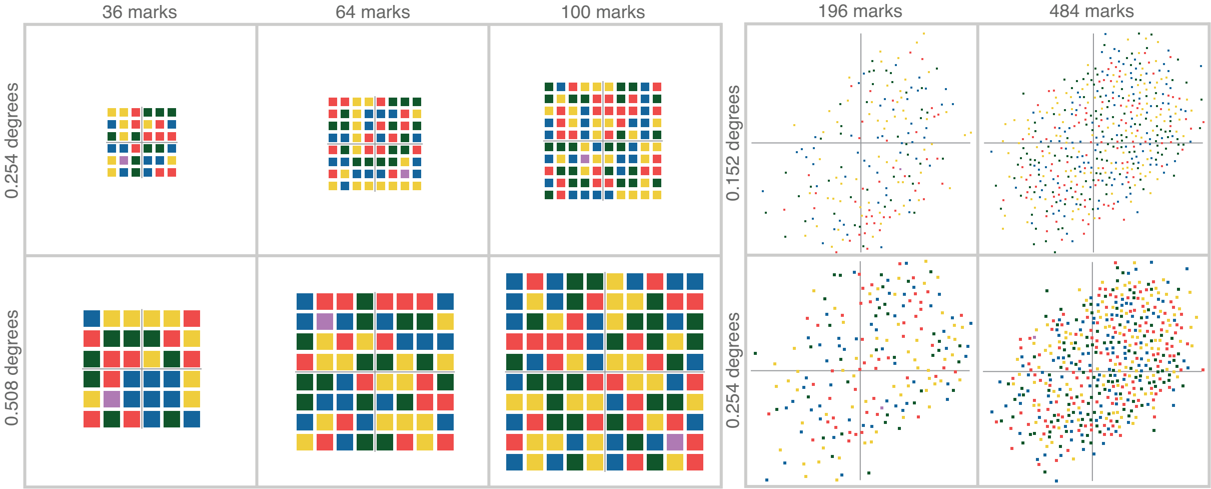


Figure 5.1: Examples of experimental displays. Participants were asked to find a target (purple square) in visualizations with varying mark sizes, set sizes, and color configurations. Figures not drawn to scale.

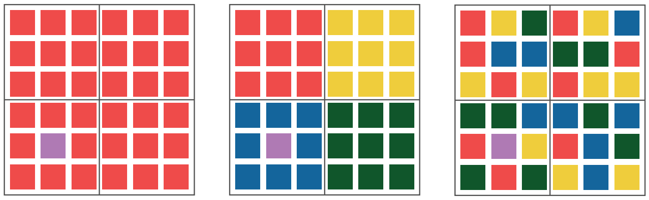


Figure 5.2: Three example grids that were presented in Experiment 1. The left shows the single-colored layout, the middle shows the group-colored layout, and the right shows the random-colored layout.

with understanding and optimizing tasks where users need to look through many series of visualizations and find a target as quickly as possible. We acknowledge that other measures (e.g., long-term memory) are important in improving our knowledge of visualization usability, and that the goal of a given study is paramount in choosing a dependent measure.

We predicted the following:

P1 Participants take longer to find targets in random-colored grids than in grouped- and single-colored grids

P2 Set size and mark size influence responses to random-colored grids to a larger degree than to grouped- and single-colored grids

P3 Responses are slowest when grids have large quantities of visual marks of very small or very large mark sizes (e.g., a 14×14, 50 px condition)

We derive P1 and P2 from our prediction that visual search response time trends, due to pop-out, will be uniform and parallel for single and grouped colored grids independent of changes to visual mark size and set size. Related, we believe that random grids – which we predict do not afford pop-out effects – will be influenced by changes in visual mark size and set size. We derive P3 from our belief that processing many small marks requires effort to differentiate and parse and that processing many large marks requires effort from gaze shifting during search.

5.2.1 Methods

Participants

There were 15 participants (mean age 24.2 years, range 19-30 years) recruited from on-campus fliers and university mailing lists. All had normal color vision (assessed with *H.R.R. Pseudoisochromatic Plates* [59]). All gave informed consent and were compensated for their participation. The Brown University Institutional Review Board approved the experiment protocol.

Design and Displays

Experiment 1 included two size factors: visual mark size (length of one edge of the square marks) and mark set size (the total number of visual marks). The levels for mark sizes and set size were:

Mark size: {.254 ° (10px), .508 ° (20px), .762 ° (30px), 1.016 ° (40px), 1.271 ° (50px)}

Set size: $\{6 \times 6, 8 \times 8, 10 \times 10, 12 \times 12, 14 \times 14\}$

Mark size is given in terms of visual angle, where 1° is roughly equivalent to 1.064cm. We limited the maximum set size to 14×14 due to the resolution constraints of the testing environment's monitor while trying to maintain a diversity of set size and mark size conditions. We also tested three color layout variations (Figure 5.2): 1) single-color, 2) 4-color grouped, and 3) 4-color random layouts. In the single-color layout (Figure 5.1, left), the distractor marks were all the same color (see below for color details). In the 4-color grouped layout, the distractor marks were spatially grouped by color into four quadrants (Figure 5.1, center). In the 4-color random layout (Figure 5.1, right), the distractor marks were randomly colored (equal numbers of each color except one color in which one square became the target). The three color layouts crossed with the 25 combinations of set size and mark size created the 75 main conditions. Henceforth, the 4-color grouped condition is referred to as "grouped" and the 4-color random condition is referred to as "random."

Within each color layout there were four variants. In the single-color layout condition the variants were four distractor colors (red, yellow, green or blue). In the grouped layout the variants were four different permutations of color group placement (e.g., in one condition red was in the top-left quadrant but in another it was in the top-right). In the random layout the variants were for random assignment of color positions. These variants were treated as replications because they were not central to the aims of this study. We had an equal number of colored squares in the grouped and random conditions (e.g., 10×10 grids had 25 squares of each color). This constraint guaranteed that each quadrant in the grouped condition corresponded to a unique color. We placed the same constraint on the random condition for comparability.

Each display type described above was presented four times so the target would appear an equal number of times in each quadrant for each display type. The four target locations were treated as replications.

The full experiment design included 1200 displays (5 mark sizes × 5 set sizes × 3 color layouts × 4 color variants × 4 target locations). There was one replication of the full design (total of 2400 trials) to ensure that there were enough data to analyze participant reaction times. When averaging all replications, there were 32 trials for each of the 75 main conditions for each participant.

Grid creation

The grids were generated individually for each participant using a Python script to create grid data, which were rendered with a D3/Node.js script [11, 191]. The rendered squares were always separated by a .127° (5px) gap, regardless of the other size conditions. The target location within each quadrant was randomly assigned for each trial. However, we added a constraint that targets could not exist on any of the four edges of the quadrant because targets falling on a border elicit different results from those one or more marks away [193].

Color selection

Many researchers have shown how color selection is an important consideration when designing visualizations. For instance, Healey et al. found that encoding search targets with a differing hue can lead to more accurate responses [68]. Because of the relation between performance and color selection, many have suggested color selection techniques to improve the usability of visualizations [65, 73, 104, 111]. Healey suggests a method to pick colors using the Munsell color model [67], which closely resembles our color selection process. The colors we selected were: red, yellow, green (Healey used green-yellow in his method), blue, and purple. We arrived at our similar colors independently. We achieved this by choosing a purple that had the most intermediate luminance, chroma, and hue arc values of the chosen palette.

In Experiment 1, the target color was always purple and the distractors were blue, green, yellow, and red (see Table 5.1 for CIE xyY and LCH coordinates). The reason for using only one target color was described above (Section A.5), and the choice to make the target hue purple was arbitrary. All of the colors were nameable and categorically different. The colors were all assigned different luminance values, given that incorporating luminance contrast between elements facilitates legibility [178].

The purple target was set to have a mid-level luminance (30 cd/m^2) with respect to the distractor colors. The Michelson contrasts between the purple target and the blue and red distractors was +/- 16.5%, and the contrast between the purple targets and the yellow and green distractors was +/- 33% (see Table 5.1 for luminance values). The purple had a mid-level chroma, situated halfway between the higher chroma red and yellow and the lower chroma blue and green (see Table 5.1).

The CIE L*u*v* coordinates were translated into CIE 1931 xyY space using an Illuminant D65 white point (x = .3127, y = .3290, Y = 100). These device-independent coordinates were translated

Color	x	y	Y	Lightness	Chroma	Hue
Yellow	0.4393	0.4769	60.0	81.838	95	75
Red	0.4335	0.2982	42.0	70.871	95	5
Purple	0.2899	0.1933	30.0	61.654	83	295
Blue	0.1768	0.2373	21.5	53.492	71	225
Green	0.1903	0.4681	15.0	45.634	71	155

Table 5.1: Colors used in the study expressed in xyY color space and each color's corresponding lightness, hue angle, and chroma (LCH)

to monitor-specific RGB values so they could be accurately rendered on our calibrated monitor.

Each grid was displayed on a black background. Dark gray lines delineated the borders between the four quadrants (CIE x=.3021, y=.3121, Y=12.43).

Procedure

The monitor was warmed up for 30 minutes before each test session to prevent color shifting during the experiment. Participants first gave consent, completed the *H.R.R. Pseudoisochromatic Plates* [59] color vision test, and filled out demographic information. The lights were then turned off in the testing booth. The participants were told that they would be presented with a series of grids, each containing a purple target, and their task was to indicate which quadrant contained the target (i.e., top-left, top-right, bottom-left, bottom-right). To respond they used four labeled keys on the keyboard numpad (one for each quadrant). The experimenter remained in the room while participants completed 10 practice trials to answer questions, after which the experimenter left the room. During the experiment participants were shown each of the 2400 grids one at a time in a random order. Each grid remained on the screen until participants made their response. Each trial was separated by a 500ms intertrial interval during which the screen was black except for a fixation cross of the same color as quadrant grid lines. Short breaks were given after every set of 15 displays and long breaks were given 25%, 50%, and 75% of the way through the study. Participants were seated approximately 60 cm away from the screen and were asked to reduce any movement towards or away from the screen; this was reinforced throughout the practice trials.

Equipment

We used an ASUS ProArt Series PA246Q Black 24.1" monitor (1920 × 1200 pixel resolution). The monitor was characterized with a Konica Minolta CS-200 Luminance and Color Meter. The experiment was conducted through a locally hosted instance of Experimentr [62].

5.2.2 Results and Discussion

Before analyzing results we filtered the data using standard procedures for reaction-time datasets [148]. We first removed all trials where participants made incorrect responses because we were interested in participants' reaction times when they were successful in finding the target. The mean accuracy across participants was 92% (range: 90%-93%). Upon inspection, the errors appeared evenly divided across conditions, but there were too few errors for systematic statistical analysis.

We next removed outlier trials for each participant, defined as response times more than two standard deviations away from the mean of all trials for that participant. The mean number of outlier trials across participants was 89 trials (range 32-103 trials). Given that participants completed 32 trials for each critical condition, ample data remained after outliers and incorrect responses were removed. Across all subjects and conditions 28 out of 34 trials were considered on average (range: 9-32).

Interaction between mark-set size and color layout

Figure 5.3 (left) shows the effect of set size on response time for each color layout condition, averaged over mark size. For each color layout condition, we tested whether set size influenced performance by first calculating the best-fit line for each subject and then using *t*-tests to compare the mean slope of the best fit lines with zero. There was a robust effect of set size for the random color condition ($t(14) = 7.17$, $p < .001$): participants took longer to find the target as the set size increased. The positive slope indicates that participants used serial search until they found the target. In contrast, the slope for the grouped- and single-color conditions did not differ from zero ($t(14) = 1.49, 1.76, ps > .05$, respectively), indicating that participants used parallel search and the target "popped out," regardless of the number of distractors.

We next compared the random and grouped conditions to look at effects of grouping and set size when the number of distractor colors was held constant at four. There was a robust effect of color layout: response times were significantly faster in the grouped than in the random condition ($F(1, 14) = 403.96, p < .001$). The magnitude of this difference varied with set size, as indicated by a layout × set size interaction ($F(4, 56) = 98.75, p < .001$). The extent to which the random layout slowed performance increased as the number of elements increased. Recall Haroz and Whitney's

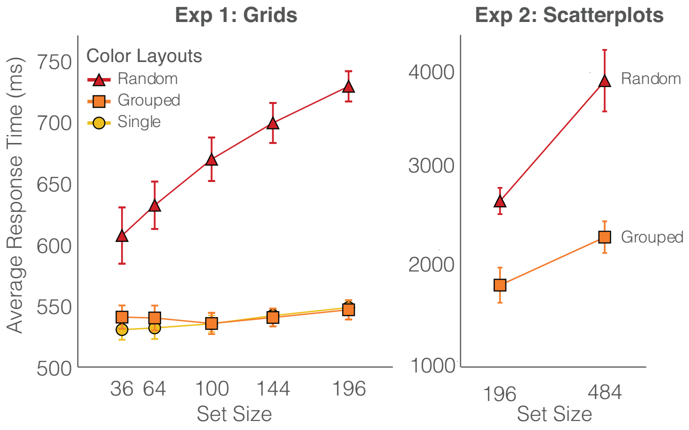


Figure 5.3: Averaged response times for all color layouts for each set size. Bars show standard error.

report that grouping had a minor effect in their known-target condition (difference of roughly 100 ms) for their displays of 64 elements. We found a comparable difference for our displays containing 64 elements (91.2ms averaged over mark size), but the difference increased to 180.4ms for our largest set size of 196 elements. Thus, color layout has a larger impact on displays with more data, complementing Haroz and Whitney's finding that layout has a larger impact on displays with higher color variety.

This difference between the random and grouped conditions can be understood by considering how the "number of distractors" is defined by the visual system. In the grouped condition, the same-colored elements are grouped by color similarity and by common region (due to grid lines), causing them to form four global "objects." In this interpretation, the small squares can be considered texture elements that comprise the four global objects [94]. Adding more texture elements (which we have been describing as increasing set size) does not change the perceived number of distractors, which is still four – one for each color group. If the number of global objects remains constant, previous work on texture predicts little to no increase in response time with the addition of more elements. Consistent with this interpretation, the average response time in the single-color layout condition was faster than in the 4-color grouped condition ($F(1, 14) = 6.09, p < .05$).

Effects of mark size and its interaction with set size

Figures 5.4A,D, and G show the data from Figure 5.3 separated by mark size, with an individual chart for each color layout condition. We see two main patterns in these data. The first is that the lines within each color layout condition are roughly parallel, indicating that the effect of set size is similar for the different mark sizes. We tested this observation by first calculating the best-fit line of each participant's response time as a function of set size for each square length in each color layout condition (the 15 lines in Figure 5.4A,D,G). We then conducted a one-way repeated-measures ANOVA for the five slopes within each color-layout condition. The slopes for the different mark sizes did not differ significantly within the single, grouped, or random color layout conditions ($F(4, 56) = 2.10, 1.64, .74, ps > .05$, respectively). This analysis suggests that the effects of mark size are independent of the effects of set size within each color-layout condition.

The second pattern is that response times for the smallest mark size were the greatest for all color layout conditions, and that the effect of mark size on response time plateaus as mark size increases. This pattern is clearer in Figure 5.4B,E,H where the rate of decline between adjacent mark sizes decreases as mark size increases. For all color layouts we see the sharpest decline in response time between .254 ° and .508 ° mark sizes, with subsequent slopes between other neighboring mark sizes about half or less. This observation is supported by robust linear and quadratic contrasts in the mark size factor for all three color layout conditions: random-linear $F(1, 14) = 76.87, p < .001$; random-quadratic $F(1, 14) = 56.10, p < .001$); grouped-linear $F(1, 14) = 11.54, p < .001$; grouped quadratic $F(1, 14) = 83.71, p < .001$; single-linear $F(1, 14) = 95.21, p < .001$; single-quadratic $F(1, 14) = 86.06, p < .001$. The finding that performance is worst when marks are small and that performance improvement plateaus as mark size increases is consistent with prior results. Heer and Bostock found that comparing bar-chart values had similar plateauing advantages when increasing chart height [70], and Jakobsen et al. found similar plateaus when increasing physical displays for map navigation tasks [85, 154]. These findings only partially fulfill P3, as .254 ° lengths do have the highest response time; however, 1.271 ° mark sizes have roughly the same response time as 1.016 ° mark sizes for all conditions. It is possible that P3 may still be supported by examining larger set sizes.

We also plotted response time as a function of total grid length to examine the impact of adding more data points given a fixed frame (Figures 5.4C,F,I). If a designer is working with a small amount of screen real estate and with ungrouped data, our results show that while you can fit 196, .254 ° elements in a slightly greater space as 36, .508 ° elements, doing so instills a large penalty to performance.

The difference in search type (parallel vs. serial) shows that increasing set size is a barrier to efficiency in noisily colored visualizations but a negligible influence in ordered or simply colored visualizations. There is also a significant interaction between mark size and color layout ($F(8, 112) = 17.773, p < .001$), where participants perform significantly worse in the random condition. These two results taken together support P2, as increasing set size and mark size will slow response time for randomly colored grids at a faster rate in comparison to grouped and single colored grids. As seen in Figure 5.4, random layouts as a whole elicit slower response times in comparison to grouped and single colored grids thus supporting P1.

It is possible that the response time trend in our results could be in part due to interactions that Stone notes between color discriminability and size [178], however further testing is required to determine such an interaction. The results of this experiment indicate that if data can be grouped (e.g., by color) then search performance is not affected by the quantity of data marks. However, it is not always possible to group data, such as in scatterplots where ordering cannot be altered. We will will investigate the effects of mark and set size in less ordered displays in Experiment 2.

Predicting Search Time: Experiment 1

We used multiple linear regression analysis to better understand the relative importance of the main factors in our study. The factors we used were grouping (1 or 0), set size (total number of marks), log of mark size, and the number of colors (1 or 4). We chose to take the log of mark size for our model because of the decreasing response time trend seen in Figure 5.4.

The model accounted for 89% of the variance in the data from Experiment 1. Grouping accounted for the most variance (75%), log mark size accounted for an additional 7%, set size an additional 6%, and the number of colors did not account for additional variance. From this model we obtained a regression equation, where RT is response time, g is grouping, l is the log mark size, and s is set size: $RT = 62.31 - 127.36g - 83.17l + .30s$.

Results in context

Haroz and Whitney showed that grouping counteracts large increases in response time for increasing color and motion complexity; we corroborated this and add that grouping negates large changes in performance for mark size and set size variation. Our results show that random grids are affected by mark size and set size manipulation whereas single and grouped grids are not. We disagree with Haroz and Whitney's statement that the variety of visual features has a weak effect on response time when people know what they are looking for. We think it more appropriate to say that prior

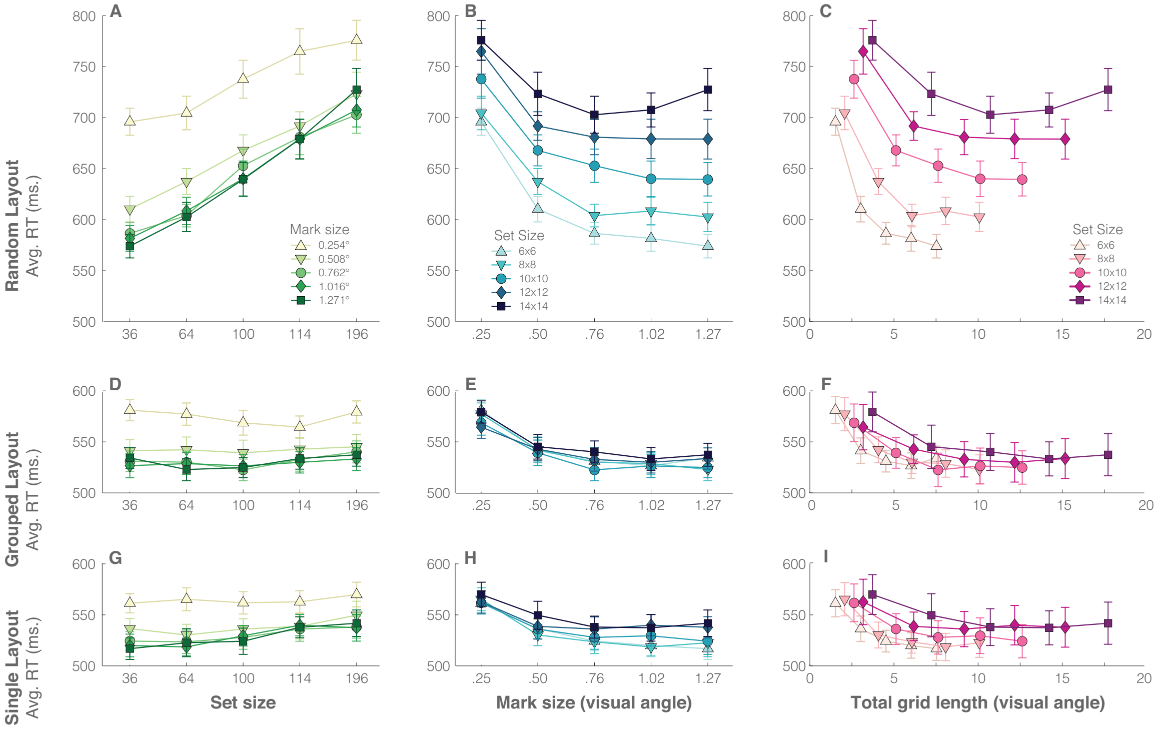


Figure 5.4: Charts showing response times for each color layout in relation to set size (column 1), mark length (column 2), and total grid length (column 3). The first row is random-, the second row is grouped-, and the last is single-colored grids. Bars show standard error.

knowledge can reduce the magnitude of the difference created by pop-out, rather than that prior knowledge eliminates pop-out and thus eliminates the differences between random and grouped layouts.

5.3 Post-Experiment-1 Survey Results

We were also interested if participants' perception of search difficulty mirrored their response times. In particular we asked if, after completing the experiment, participants could intuit which grid configurations were easier to use. To investigate this question, we gave participants a post-test survey asking them to rate how difficult it was to search for the target in each grid. We tested the orthogonal combination of all color layouts, mark sizes, and set sizes ($3 \times 5 \times 5 = 75$ trials). Grids were rated from 1 (very easy) to 7 (very difficult). Grids were presented in a random order and had randomized target location.

Results show that participants thought that grouped- and single-colored grids were always easier than random-colored grids (group vs. random: $t(14) = 54.93, p < .001$; single vs. random: $t(14) = 73.93, p < .001$). The most difficult grids were those with small marks or with large set sizes.

We evaluated how accurately participants could gauge visualization difficulty by correlating each participant's mean response time for the 75 grid types with their ratings of perceived difficulty. The average correlation was .80 (range: .42-.92). We then tested whether the mean correlation was different from zero by first calculating the arc-hyperbolic tangent transformation on each participant's correlation coefficients to unconstrain their limits and then conducting a one-sample *t*-test. The participants' correlations were significantly greater than zero ($t(14) = 12.17, p < .001$), indicating that visualization users can provide accurate feedback on difficulty relating to scale even if they are not necessarily visualization designers. We believe that this means that asking novice visualization creators – even those without design expertise – about usability issues relating to size can provide accurate design suggestions. For instance, even if cancer genomicists might have difficulty designing visualizations from scratch, if they are familiar with using the visualizations their assessment of what is too small to use will be accurate. Researchers, such as Levin [108], have shown that people are often poor at self-assessment. It is possible that this discrepancy could be due to the perception of visual clutter (e.g., Rosenholtz et al. [155]) or graph complexity (e.g., Carpenter and Shah [21]). However more research is required to deduce any such relations.

5.4 Experiment 2: Searching through scatterplots

In Experiment 2 we studied how search for a target data point in scatter plots is affected by variations in the same factors from Experiment 1: 1) visual mark size, 2) the number of visual marks (set size), and 3) color grouping. We designed this experiment to investigate smaller mark size and larger set size combinations we thought might have greater performance differences based on our findings from Experiment 1. While most of the set and mark size combinations in Experiment 2 are distinct from those in Experiment 1, we included one overlapping condition to serve as a reference point. To test a greater number of set size and mark size combinations, we omitted the single color condition tested in Experiment 1 because the results from the single and grouped conditions were similar. We also changed the grouped condition tested in Experiment 1 to a "semi-grouped" condition where there is partial overlap between groups to make the data look more like natural scatterplots (rather than distinct clusters). The colors in Experiment 2 are the same used in Experiment 1. Examples of the scatterplots used are shown in Figure 5.5. Our predictions for Experiment 2, based on the results of Experiment 1, include:

P1 Random conditions will yield slower response times compared to semi-grouped conditions

P2 Response time will increase as set size increases

P3 Response time will decay as mark size shrinks

Although finding unique targets might be only a subset of analysis tasks in scatterplot use (e.g., brushing and linking), using scatterplots as stimuli has several advantages. The scatterplots we created have high visual similarity to the grids used in Experiment 1. Ignoring the data that fuels each type of visualization, if you eliminate row and column alignment of a grid and then vary mark spacing, you get a scatterplot. This similarity is beneficial as it gives us a glimpse into how the spatial ordering of marks might affect performance.

5.4.1 Methods

Participants

There were 16 participants (mean age 25, range 20-31 years) recruited from on-campus fliers and a university mailing list. All participants had normal color vision as assessed using *H.R.R. Pseudoisochromatic Plates* [59]. All gave informed consent and were compensated for participation. The experimental protocol was approved by the Brown University Institutional Review Board. One participant was excluded from analysis because s/he took over two hours to complete the experiment whereas other participants needed only 30-50 minutes.

Design

As in Experiment 1, we varied mark size and set size, but the values were different:

Length: {.102 ° (4px), .152 ° (6px), .203 ° (8px), .254 ° (10px)}

Set Size: $\{14 \times 14, 22 \times 22\}$

There were two color layouts, one in which the colors were semi-grouped and one in which they were random. In the semi-grouped condition, the distractors that were the same color were clustered together (see Plot Creation, Section 5.4.1, below and Figure 5.5) but were not perfectly grouped and separate from one another as in the grouped condition of Experiment 1 (see Figure 5.2). As in Experiment 1, there were equal amounts of marks assigned to each color. The orthogonal combination of these three factors created the 16 main conditions of interest. Other factors included slope (positive, negative) and, as in Experiment 1, target quadrant location. Those factors were included to provide additional control but were treated as replications because they were not of central interest.

The full design included 128 conditions (4 mark sizes × 2 set sizes × 2 color layouts × 2 slopes × 4 target locations). We included a 4x replication of the full design so that each of the main conditions in Experiment 2 had 32 trials – the same number as for the main conditions in Experiment 1. With replications the experiment had 512 trials. We chose a reduced number of trials after discovering that a 1000-trial pilot study took prohibitively long. The 512 trial variant took participants up to an hour to complete.

Plot Creation

The plots described above were generated individually for each participant using the same Python and D3/Node.js pipeline as in Experiment 1. All data were generated from sampling a multivariate normal distribution with four clusters. The data were then rotated to have a slope of $y = x$ or $y = -x$. After rotation we also imposed the constraint that no data point may overlap or touch. This constraint ensured that each square corresponded to a distinct perceived object and that set size remained constant within a given condition. Any points violating the constraint were removed and new marks were generated until the desired condition for the grid was met. The target location within each plot was randomized, and target placement was less restricted than in Experiment 1 for greater ecological applicability. In Experiment 1 targets could only be placed in non-quadrant-edge locations, whereas in Experiment 2 a target could be placed at any location. Color assignment for grouped conditions was randomly decided for each group-colored plot. Frame size was fixed for all plots at 20.612 °.

Color Selection and Equipment

The same colors and equipment were used as in Experiment 1.

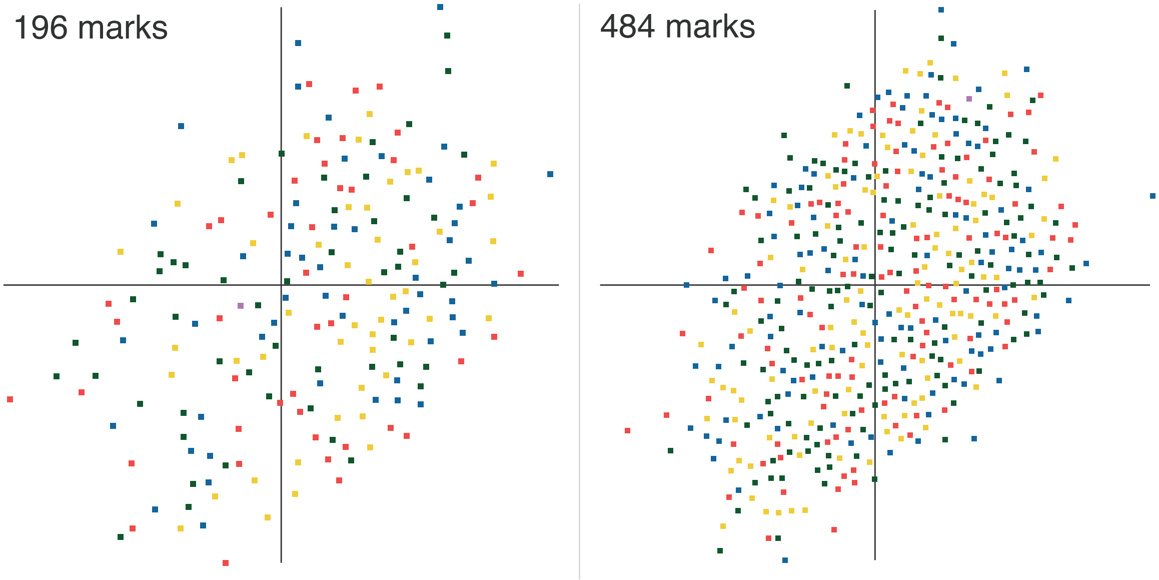


Figure 5.5: Two example scatterplots that were presented in Experiment 2. The left has 196 marks and the right has 484.

Procedure

The procedure was identical to that in Experiment 1 except short breaks were given after every set of 10 displays. We reduced the number of trials between breaks from Experiment 1 to account for longer task completion time.

5.4.2 Results and Discussion

Before analyzing results we applied the same data filtering procedure as in Experiment 1. Accuracy was lower in Experiment 2 (mean: 86%, range: 84%-87%), but still acceptable. The average number of outlier trials across participants was 17 trials, with a range 10-26 trials. As in Experiment 1, ample data remained after removing outliers and incorrect responses.

Interaction between set size and color layout

Figure 5.6A,C shows the effect of set size on response time for each mark size separately for the random (A) and grouped (C) layouts. Like in Experiment 1, we tested whether set size influenced performance by first calculating the best-fit line for each subject and then using t-tests to compare the mean slope of the best fit lines with zero. Results match those from Experiment 1: there was an effect for grouped color layouts ($t(14) = 4.842, p < .001$) and also for random color layouts ($t(14) =$

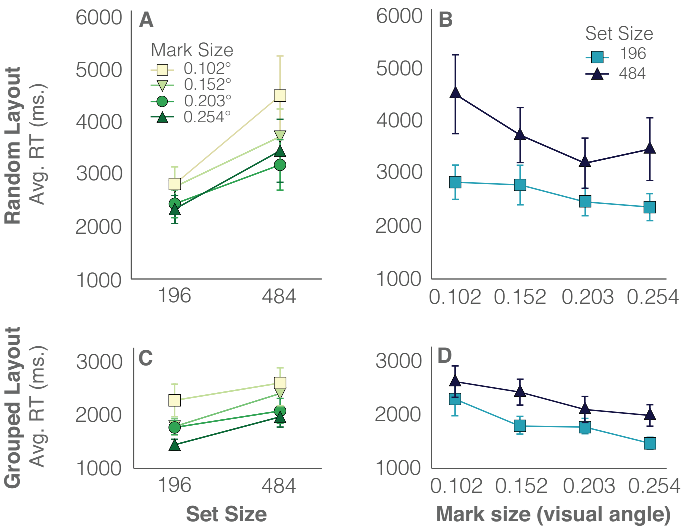


Figure 5.6: Results from experiment 2 showing random layouts (top) and grouped layouts (bottom). Bars show standard error.

$3.813, p = .002$). We also found a difference between the two color layouts ($t(14) = 2.630, p = .020$), where random color layouts took longer. This supports P1.

We next compared random and grouped conditions to look at effects of grouping and set size. As in Experiment 1 there was a robust effect of color layout, where response times were significantly lower in the grouped than in the random condition ($F(1, 14) = 19.803, p = .001$). There was a layout × set size interaction ($F(1, 14) = 6.917, p = .02$), in which the difference in response time as set size increased was greater for the random condition than for the semi-grouped condition (see Figure 5.6A,C). These findings support P2.

Effects of mark size and its interaction with set size

Figure 5.6B,D shows the effect of mark size on response time for each set size separately for the random (B) and grouped (D) layouts. In Figure 5.6, we see the same main patterns in Experiment 1. First, lines within each color layout are roughly parallel. Second, the response times for the smallest mark sizes were the longest in both color layout conditions, and the effect of mark size plateaus as mark size increases.

To test our first observation we calculated the best-fit line for each participant's response times as a function of set size for each square length in each color layout condition. We then conducted a one-way repeated-measures ANOVA for the four slopes within each color-layout condition. The slopes for different mark sizes did not differ significantly within either layout condition ($F(3, 42) < 1, p > .05$, for both layouts). As in Experiment 1, this analysis suggests that the effects of mark size are independent of the effects of set size within each color-layout condition.

To examine our second observation we tested for linear and quadratic contrasts as a function of mark size for each color layout condition. There were robust linear contrasts for both layouts (grouped:$F(1, 14) = 22.224, p < .001$; random:$F(1, 14) = 23.796, p < .001$). There was also a quadratic contrast for the random layout ($F(1, 14) = 7.423, p = .016$), and a marginal effect for the grouped layout ($F(1, 14) = 4.348, p = .056$). These two observations support P3.

Self-Reported Experiment 2 Feedback

Many participants said that it was harder to find the purple dot when: (1) it was close to the axis, (2) it was surrounded by various different colors (as opposed to within a cluster), and (3) it was not an outlier. The comment that some found it easier to find a target when it was surrounded by different colors was surprising given the results in Haroz and Whitney [60]; however, the other feedback supports existing visual search knowledge proposed by Treisman [193].

Predicting search time

We applied the equation generated from the regression model in Experiment 1 (Section 5.2.2) to test whether it generalized to predict response times in Experiment 2. We did not include the number of colors as a factor because color was not varied in Experiment 2. The Experiment 1 model fit the Experiment 2 data well, accounting for 86% of the variance. Despite changing the type of visualization, the set sizes, and mark sizes, both datasets reveal similar patters, in that both show higher relative response times for random visualizations compared to grouped visualizations. Although the exact equation generated from Experiment 1 might not be applicable to more complex visualizations, we believe that the relative ordering of the factors (grouping, mark size, and set size) suggested by our analysis will generalize. In future work, it would be beneficial to determine how our findings might be incorporated into more robust predictive modeling, such as Rosenholtz et al.'s model that detects groups in visualizations [156]. Other interesting directions related to a more robust measure of grouping are to look at color surround of targets and to investigate how semantic ramifications of groupings may influence target search in information visualizations.

5.5 Application Area Evaluation: Cancer Genomics Visualization

One relevant domain is cancer genomics analysis. cancer genomics To understand the role of mark size in this domain, we conducted a case study with two pairs of cancer genomic researchers

These particular scientists work with genome mutation matrices, which are tabular visualizations where rows indicate genes of interest, columns indicate patients, and each cell indicates whether there is a mutation (filled color indicates mutation present and the particular color codes cancer type) (see Chapter 2.3). The data are from the TCGA Pan-Cancer initiative. To increase ecological validity we used MAGI to present the mutation matrices, which all participants used in their everyday research activities. During the study we varied the size of the visualizations and observed how they affected perceptions of usability.

5.5.1 Methods

Participants

Four cancer researchers participated in the case study. All were either graduate students or postdoctoral researchers in a computational biology program. We worked with participants in pairs (two groups of two people) because that procedure has been shown to encourage dialogue and generate more information from user studies [1]. The experimental protocol was approved by the Brown University institutional review board.

To test how changes in size affected usability perceptions we used NASA's TLX evaluation, which measures an estimate of workload by participants to rate workload via six different factors [66]. We also asked participants to rate the difficulty of each condition using the same 7-point Likert difficulty rating scale from Experiment 1.

Design

We tested two mutation matrix cell sizes (.102 ° and .254 ° visual angle). The size order was counterbalanced across participants. We selected these particular visual angles because (a) they corresponded to the smallest and largest mark sizes in Experiment 2 and (b) they were the zoom limits of the analysis tool. The mutation matrices are grouped by color (cancer type) in the tool we modified.

Equipment and Materials

We used the same monitor as in Experiments 1 and 2 to present the visualizations and used printed copies of NASA's TLX evaluation and Likert scales. We additionally used Camtasia screen and audio recording software to log tool use.

Procedure

After participants gave consent, we turned on the screen and audio recording software. We began by interviewing the researchers about the type of tasks for which they used the analysis tool. We then asked them to list sets of genes they found interesting and/or worked with frequently. Before looking at their tool usage we asked each pair to go through a TLX familiarization task using displays from Experiments 1 and 2, so that participants could ask questions about and become familiar with the questionnaire. We then asked each pair to perform some of the tasks they listed earlier for each size condition. After participants used both versions of the tool, we asked them to complete two TLX evaluation sheets – one for each condition – and our Likert rating sheet. At the end of the study we asked if the participants had any additional feedback.

5.5.2 Results & Discussion

All participants reported that they used mutation matrices for different analytical purposes based on the size of the matrices' cells. The smallest mark was always associated with global, overview tasks (e.g., find exclusively occurring mutations), whereas a 6 px difference on our screen (.152 degrees) was always associated with detail tasks (e.g., what mutations a specific patient had). All participants reported that it was undesirable to search through small marks for detailed information. One participant even said that to avoid this issue in publications they use overview+detail views of mutation matrices in figures. Similarly, it was equally undesirable to try to identify global patterns like mutation exclusivity with large marks because participants would have to pan the viewport. Nearly all of the tasks listed involved using multiple visualizations in the tool.

Further, all participants had difficulty assigning workload ratings given the dissimilarity of their experience in the two conditions. On an evaluation, one participant wrote, "the two [sizes] provide different views of the same data. I find that they are equally useful when applied appropriately to relevant questions." Because of this, we could not draw conclusions from the data. Likert scales showed little difference between conditions – the small condition mean was 1.75, the large condition mean was 1.5, and all ratings were either 1 or 2. The average rating for grouped displays was 2.27 in the post-Experiment 1 survey, which is similar to our findings given that the mutation matrices were grouped by color. After the experiment, one participant said that in his/her research, the ordering of mutation matrices had a large affect on how quickly he/she was able to find information in the visualization.

The participants' reports about task-switching based on the visualization size is consistent with what is known about global vs. local processing. For example, when people are presented with Navon letters (e.g., a global "H" constructed from small local "S"s) they are faster at reporting the global letter when the display occupies a small visual angle and faster at the local letter when the display occupies a large visual angle [93, 95]. This idea predicts that people should be faster at finding local targets when the marks are large (as reported in Experiments 1 and 2 and described by the case study participants), but conversely they should be faster at detecting global trends when the marks are small. This also suggests that inappropriately sized marks create usability issues in software and that size is an important design consideration for visualizations.

5.6 Other Potential Application Areas

Although the present study involved the evaluation of simple visualizations in a lab setting, we believe our results can be applied to various tools currently used by analysts. One application area of our results is in complex analysis environments such as those provided in Bloomberg Professional. In Bloomberg Professional analysts often perform tasks involving multiple displays, multiple types of charts and data, and must make decisions with time sensitive data. We hypothesize that in complex analysis environments, fast search time can improve the analysis process by reducing the time analysts spend weeding through data in favor of time spent on using located information to generate hypotheses. It is possible that other financial software packages that do not rely on as complex monitor configurations (e.g., Palantir Metropolis) can still benefit equally as much from our findings. Other application areas include, but are not limited to, network security applications (e.g., Traffic Circle [7]), financial security monitoring (e.g., WireVis [23]), and intelligence analysis environments (e.g., Palantir Gotham).

5.7 Design Guidelines

5.7.1 Group similar marks

People are faster at search through visualizations in which similar marks are grouped together (e.g., grouping by color similarity) compared to visualizations with little grouping. In some types of visualizations it is impossible to group similar marks (e.g., one cannot decide where data are placed in scatterplots). However, if the ordering does not matter in the visualization, such as in cancer mutation matrices, treemaps, and even bar graphs, it is beneficial to group marks by similarity. Designers should, however, be careful when applying this guideline that the ordering does not cause other detrimental effects. For example, given a set of colors that encode data in a categorical heatmap, the ordering of color may give the illusion of a continuous gradient even though the data is categorical. It would be interesting to study if the benefits of ordering outweigh such illusions.

Another effect of ordering can be seen in our comparison of Experiments 1 and 2. The scatterplot stimuli we tested in Experiment 2 were very similar to the grid stimuli from Experiment 1, with the main difference being spatial location: rather than being arrayed in a tight grid, our scatter plots had squares with varying distances and alignments to one another. Our preliminary comparisons between types of visualizations suggest that spatial ordering of marks magnifies usability issues related to the number and size of visual marks.

5.7.2 Avoid large mark quantities when data cannot be grouped

When marks are strongly grouped (e.g., by color similarity), search time is not affected by the quantity of data. However, as visualizations become less ordered, the quantity of data marks becomes scalar for search response time. Visualization summarization is often used to compensate for the impossibility of showing all data in a visualization at once. Such scenarios can occur when there is more data than pixels or when node-link diagrams become “hairballs” from a large number of nodes and high connectivity. Our results suggest that even if all data can be shown at once, such data reduction methods can be beneficial if the marks cannot be grouped. While summarizing data might not make sense in every scenario – as summarizing the data could limit tasks other than visual search that require a fuller representation of data – this guideline nonetheless gives designers another tool.

5.7.3 Use large (enough) mark sizes

In tasks that involve finding a target, avoid using small mark sizes (i.e., $\leq .508\,^{\circ}$ visual angle) because of the slow performance. The range of mark sizes most susceptible to slowing performance happen to be the range of mark sizes used in typical scatterplots and marked line graphs. The importance of choosing size is even greater when considering Stone's findings that perceived color can differ based on mark size [178]. However, the usefulness of increasing mark size plateaus with increasingly larger mark sizes (Figure 5.4). Performance was roughly equivalent for marks whose visual angle ranged from $.762\,^{\circ}$ to $1.271\,^{\circ}$.

It is unclear what the effect of increasing mark size is beyond that tested in Experiment 1. One possibility is that as sizes become larger there is a point at which response time increases due to the need for users to move their head to view different parts of the display. This can become an issue for large format visualizations, such as those that can be found in virtual reality.

We note that this design recommendation pertains to finding a single target within a visualization. User reports from our case study suggest that if the goal for the visualization is to discern a global pattern, then smaller marks can be better.

5.8 Limitations

Although our study examines the relation between grouping, mark size, and set size in depth, there are numerous other factors that are involved in visual search performance for information visualization. For instance, Stone has claimed that color can interact with size to affect legibility [178], and it is unclear from the present results to what degree size was a problem due to its affect on discriminability. One way to test this potential interaction is to control for color discriminability at different sizes and see if the response times are similar to those found in Experiments 1 and 2. There is also the question of target saliency in search. If the target in a grid were encoded with a bright white, another salient color (e.g., pink [102]), or were blinking, then it is possible that current effects of set size, mark size, and grouping would be diminished. Other potentially relevant factors include density [13] or the amount of marks assigned to each color category.

Another concern is that we tested only a subset of sizes given our monitor, and it is unclear how our results extend to larger visualizations (e.g., virtual reality). It could be that the observed performance plateau extends into larger display configurations. If the curve is only due to color discriminability then the plateau should remain. However, it is possible that for sufficiently large sizes there could be another factor that causes a dip in performance (e.g., head movement in virtual reality). The slight upturn in the data for large set sizes in random displays (Figure 5.4B,C) hint to this phenomena. Related, although we tested both fixed and varied spacing of marks (Experiment 1 and 2, respectively) the effects of spacing warrants further investigation.

We also note that the guidelines here apply directly to target search tasks, and further study is necessary to determine whether they generalize to other types of tasks (e.g., average comparison tasks [41] and correlation and cluster detection [36]). For tasks that consider global pattern understanding, it is possible that ideas from the ensemble statistics literature may prove useful (e.g., Haberman and Whitney [57]), as it is possible that ensemble parameters could influence global pattern task type performance. Finally, how might our results transfer to continuous, not categorical, data? We believe all of these points provide interesting future lines of research.

5.9 Conclusion

We explored how the color layout, quantity, and size of marks in a visualization can impact visual search time based on the results of two experiments. Each experiment asked participants to search for a unique target in colored visualization, where the first experiment tested various colored grids and the second tested various scatterplots. We found that search performance was faster when colors were spatially grouped. We also found that the number of marks had little effect on search time when colors were grouped, but had a robust effect when colors were laid out randomly. Finally, we found that the smallest mark size was always slower and that increasing mark size led to plateauing response times. We assessed the difficulty associated with mark size beyond our quantitative experiments through a post-experiment survey, finding that participants were accurate in rating how difficult visualizations were. We also conducted a small NASA TLX cognitive workload study with cancer researchers and found that even small design changes in size can have notable effects on usability by altering what task associations users have with a visualization. These results led to several design guidelines for improving visualization search performance. In full these contributions expand our present knowledge of effective visualization design practices with respect to altering data density and visualization size so as to support search tasks.

Chapter 6

Colorgorical: Creating discriminable and preferable color palettes for information visualization

As described in our study of how visualization size can affect performance, visualization creators often lack design expertise to make informed style adjustments to charts (Chapter 5). Rainbow colormaps are perhaps the most extreme example of this, where people's preference for creating visualizations that are aesthetically pleasing to them without design expertise can lead to misleading and hard-to-interpret charts [10, 153]. For example, Borkin et al. evaluated doctors' accuracies with different color maps when looking for a symptom of heart disease called endothelial shear stress sites [9]. Using conventional rainbow-colored 3D imaging charts, doctors had 39% accuracy and had 62% accuracy when using 2D charts. In contrast, doctors' accuracies were 71% (+32%) and 91% (+29%), respectively, when they performed the same search task using better designed, non-rainbow diverging color maps.

In this chapter we tackle a related, but separate, open research problem: automated categorical

This chapter is an extended version of a work that originally appeared in the Proceedings of Information Visualization 2016 and in IEEE's *Transactions on Visualization and Computer Graphics* as [49].

palette design. As in other areas of design, it is important that a visualization color palette is aesthetically pleasing; but, unlike many other areas of design, visualization color palettes must also be highly discriminable. For example, it is important in MAGI that each cancer category that is encoded with color is distinct from one another. Balancing discriminability and aesthetic preference is challenging because they can be inversely related (i.e., preference increases with hue similarity [165], whereas discriminability decreases). Navigating this tradeoff requires design skill and experience, both beyond those of many visualization creators. Our primary contribution is twofold. First, we discuss Colorgorical (Fig. 5.1) – a novel tool that can create arbitrarily sized color palettes designed for visualization based on how important discriminability vs. preference is for any given user. Second, we evaluate various Colorgorical settings and compare to tool to existing industry standards.

Colorgorical operationalizes effective color palette selection with three color-scoring functions to balance discriminability and aesthetic preference: Perceptual Distance (CIEDE2000) [172], Name Difference [73], and a quantified model of color Pair Preference [165] (Sec. 6.2). A fourth, Name Uniqueness, was originally included, but was later removed because it had little effect on behavior (Sec. 6.5). With Colorgorical, color palette creation is simplified so that users need only specify the number of desired colors and drag sliders controlling color-scoring function importance to (1) create custom palettes that the average individual would find preferable while maintaining discriminability, and (2) explore how relative weights on discriminability vs. preference affect palette appearance. Users can further customize palettes by specifying desired hues and by building onto existing palettes (Sec. 6.3).

We evaluated Colorgorical's effectiveness in four ways: (1) runtime benchmarks (Sec. 6.3), (2) discriminability and preference score analysis (Sec. A.5), (3) human-subject evaluation of different model settings (Sec. 6.5), and (4) human-subject evaluation of Colorgorical compared to industry standards (Sec. 6.6). We make the following contributions:

- We provide a technique to generate custom color palettes via user-defined importance of discriminability and preference
- We detail the relations between Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference scoring functions
- We show how varying the relative weights of discriminability and preference sliders affects

human discrimination performance and preference ratings

- We present evidence that Colorgorical palettes are as discriminable and often more preferable than industry standard, professionally hand-made color palettes

Colorgorical combines three features, making it a novel approach to palette design. First, it is designed specifically for visualization rather than for general art and design applications. Second, it uses empirically derived color preference data to inform categorical palette generation [165]. Third, it approaches visualization palette design by balancing categorical palette discriminability and preference.

6.1 Related Work

Current color palette tools are typically designed based on three types of strategies: discriminability optimization, color-term association mapping, or harmonic template application. We describe these approaches and discuss how Colorgorical targets limitations of past research.

6.1.1 Palette discriminability methods

A key issue in palette discriminability is whether a graphical mark can be quickly and accurately identified. Healey demonstrated that this problem can be addressed by using palettes whose colors are named with the 10 Munsell hues and that maximize perceptual distance between colors (CIEDE1976, Sec. 6.2) [67]. Maxwell also developed a discriminability-based technique to create categorical color palettes for multidimensional datasets based on classification dissimilarity of categories [122]. These approaches created discriminable palettes, but each has multiple limitations for design more broadly: (1) they do not address aesthetics, (2) Healey's technique is constrained to 10 or fewer color terms, and (3) they define perceptual distance using Euclidean distance (Healey) or maximum scaled difference (Maxwell) in CIELAB color space, which can be problematic due to perceptual uniformity limitations [116] (i.e., the same distance can have different perceptual consequences depending on the sampled region).

Colorgorical addresses these issues by (1) considering aesthetics in addition to discriminability (Sec. 6.2) [165], (2) using 153 crowdsourced color terms compared to the 10 Munsell hues in Healey's method, and (3) using an updated perceptual distance function (CIEDE2000) that improves perceptual uniformity in the distance metric [172].

The color-name associations in Colorgorical are based on Heer and Stone's color-name statistics (Sec. 6.2) [73], which are derived from color-name association frequencies from the 153 most commonly-used names from the XKCD color-name crowdsourcing survey [126]. Name Difference measures the difference in color-name association frequency distributions between two colors. For example, green and red colors have large name differences because green colors have few associations with red names and vice versa. Presumably Name Difference is related to Perceptual Distance, but it is possible that they differ systematically, which we test in Sections A.5 and 6.5. Name Salience, which we call Name Uniqueness to avoid confusion with color salience, captures the degree to which a color is specifically named (highly associated with only a few colors) vs. broadly named (moderately associated with many colors) (Fig. 1 in Supp. Mat.).

Another approach to designing discriminable palettes is for color experts to make pre-defined palettes (e.g., ColorBrewer [64]). Typically made through iterative design, experts construct these palettes by selecting colors that are discriminable under a variety of viewing conditions (e.g., after photocopying) and that support specialized tasks (e.g., ColorBrewer's "Accent" palettes emphasize certain colors). Although pre-made palettes are easy to use, they do not give visualization creators design flexibility or customizability. And although guidelines for hand-designing palettes exist [207], a visualization creator might not want to spend time or effort to learn about palette design. Colorgorical addresses this problem by allowing customization while building in constraints on aesthetics and discriminability; however, we leave support for specialized palettes (e.g., accent colors) for future research.

6.1.2 Color-term tools

Another way to create categorical palettes is through color-term associations. Crowdsourcing and linguistics-based approaches can produce color-term associations that create semantically meaningful palettes (e.g., a "mango ice cream" category might produce a light orange) [111, 171]. Setlur and Stone show that various natural language processing techniques can be used to mine color-semantic pairings from large text datasets [171]. Colorgorical does not currently support semantic mappings, but it is an exciting future direction.

6.1.3 Harmonic template tools

Many harmony-based categorical color palette tools are targeted for general-purpose design and do not focus on visualization design constraints (e.g., discriminability). These tools create palettes based on harmony principles in color theory [127, 133]. A common implementation of harmony is through *harmonic templates* based on hue relations [120], such as the two-color complementary relation that stems from Itten's version of harmony (e.g., blue and orange) [82]. For example, Adobe Color creates 5-color palettes based on harmonic templates and optional image color analysis [132]. Similarly, Dial-a-color, uses harmonic templates as a starting point and allows users to alter color properties like lightness and saturation [123]. ACE lets users manipulate discrimination and harmony importance for interface design by answering a series of questions in a text interface about each colored interface component [124] (unlike ACE, Colorgorical is not limited to interface coloration and uses sliders to balance discrimination and aesthetic preference rather than a text interface). Finally, the Harmonious Color Scheme Generator constructs color palettes through *familial factors* (promoting similarity along hue, saturation, or lightness dimensions) and *rhythmic spans* (sampling colors using a fixed uniform interval along a color dimension) [76].

Harmonic templates were generated from color theory in art without empirical validation [82], and do not necessarily correspond to human judgments of harmony. For example, the notion that complementary colors are harmonious is key to the notion of harmonic templates. Yet humans judge complementary hues as among the least harmonious and instead judge more similar hues as more harmonious [134, 135, 165, 184].

Although the term "harmony" is often used interchangeably with aesthetic preference [24], the two are not the same [134, 165]. Schloss and Palmer demonstrated how they differ, where harmony was defined as "how strongly an observer experiences the colors in the combination as going or belonging together, regardless of whether the observer likes the combination or not," and preference is "how much an observer likes a given pair of colors as a Gestalt, or whole" [165]. Although both increased with hue similarity, pair preference relied more on preference ratings for individual colors and on lightness contrast, whereas harmony relied more on desaturation (i.e., pairs with less saturated colors were more harmonious).

Colorgorical uses Schloss and Palmer's pair preference model (Sec. 6.2) [165] rather than harmony because we reasoned that how much people like visualization palette colors is more central to the present aims than how well they feel the colors go together.

6.2 Background: model scoring functions

Colorgorical iteratively samples colors using three color discriminability scores (*Perceptual Distance*, *Name Difference*, *Name Uniqueness*) and a color preference score (*Pair Preference*). Colorgorical assumes that discriminability and preference for large combinations of colors can be predicted by these lower-order scores. Name Uniqueness was ultimately removed from the model because it had little effect on discriminability performance or preference (Sec. 6.5).

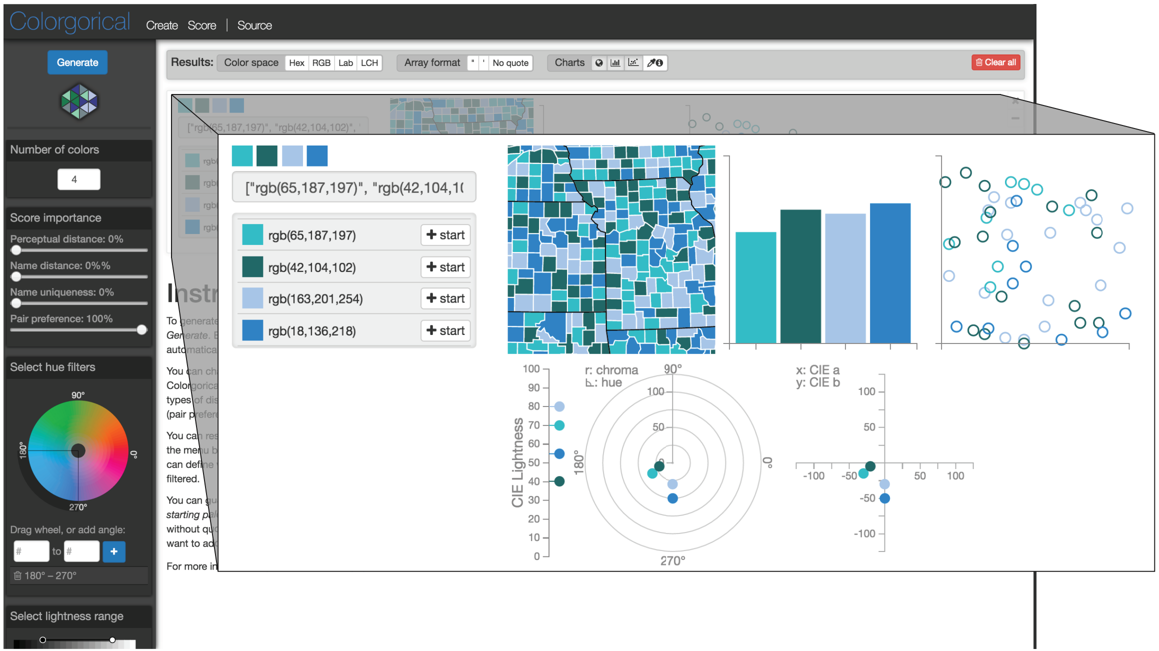
Each score operates in CIELAB. The $L*$ axis of CIELAB approximates a color's lightness, the $a*$ axis approximates its redness-to-greenness, and the $b*$ axis approximates its blueness-to-yellowness. To support Name Difference and Name Uniqueness, we use a modification of CIELAB that quantizes the space into 8,325 discrete colors by sampling every 5 units along each axis starting at the origin [73]. Some scores also depend on CIE LCh, which is a polar representation of the Euclidean CIELAB space. In CIE LCh, $L*$ is the same as in CIELAB, but the $a*$ and $b*$ axis are converted to chroma (C, radius) and hue (h, angle).

6.2.1 Color discriminability scores

We used multiple discriminability scores because perceptual difference might differ from name difference. For instance, a chartreuse (yellow-green) might be perceptually distinct from a green or yellow but might be called green or yellow, making it easy to confuse with other greens or yellows in a visualization when referenced by name.

CIEDE2000: Perceptual Distance

To calculate Perceptual Distance between two colors we use CIEDE2000 (DE_{00}) [172]. It is similar to the original CIEDE, DE_{76} (Euclidean CIELAB), but DE_{00} calculates distance in CIE LCh with a hue rotation term (R_T) and corrections for lightness (S_L), chroma (S_C), and hue (S_h) to improve perceptually uniformity [115].



Number of colors: 4

Score importance
- Perceptual distance: 0%
- Name distance: 0%%
- Name uniqueness: 0%
- Pair preference: 100%

Hue filter: 180° – 270°

["rgb(65,187,197)", "rgb(42,104,1(

- rgb(65,187,197) + start
- rgb(42,104,102) + start
- rgb(163,201,254) + start
- rgb(18,136,218) + start

Figure 6.1: A Colorgorical screenshot. Here a user has specified a hue filter (left) and has generated a 4-color palette (detail). Users can list colors in many color spaces and render colors in a variety of charts.

$$\mathrm{DE}_{76} = \sqrt{\Delta L^2 + \Delta a^2 + \Delta b^2} \tag{6.1}$$

$$\mathrm{DE}_{00} = \sqrt{\left(\frac{\Delta L}{S_L}\right)^2 + \left(\frac{\Delta C}{S_C}\right)^2 + \left(\frac{\Delta H}{S_H}\right)^2 + R_T \frac{\Delta C}{S_C}\frac{\Delta H}{S_H}} \tag{6.2}$$

Name Difference

Name Difference (ND) captures the degree to which two colors have distinct color-name association frequency distributions [73]. Color-name associations are mappings between colors and names (e.g., `rgb(255,0,0)` → "bright red"). The name data are composed of the discretized CIELAB color space (C) described earlier, a list of 153 popular color names (W), and a color-name association frequency matrix (T) that has C rows and W columns. The scores also rely on the conditional probability of a color name w given any color in C:

$$p(w|c) = T_{c,w} / \sum_{w} T_{c,w} \tag{6.3}$$

We calculate Name Difference using Hellinger distance [73]:

$$\mathrm{ND}(c_1, c_2) = \sqrt{1 - \sum_{w \in W} \sqrt{p(w|c_1)p(w|c_2)}} \tag{6.4}$$

Name Uniqueness

Name Uniqueness (NU) captures the degree to which colors have uniform distributions of color-name association frequencies. Colors that have few strongly associated names (i.e., a focal distribution) result in lower scores, whereas colors that have many weakly associated names (i.e., a more-uniform distribution) result in higher scores. Name Uniqueness is calculated by using the negative entropy of a color's name-association frequency distribution from the color-name-association matrix (T) and the list of color names (W):

$$\mathrm{NU}(c) = -\mathrm{H}(p(W|c)) = \sum_{w \in W} p(w|c) \mathrm{log} p(w|c) \tag{6.5}$$

Unlike the other two discriminability measures, Name Uniqueness relies on individual colors rather than relations between other colors within the palette. We believe this is a key reason why it was not useful in the Colorgorcial model (Sec. 6.5).

6.2.2 Aesthetic preference score: Pair Preference

Pair Preference (PP) is based on a linear regression model used to predict pair preferences from three color-appearance and color-relation factors, which was previously operationalized in Munsell space [165]. The best-fit model explained 53.5% of the variance in pair preference judgments with three factors: coolness (κ), hue similarity ΔH, and lightness contrast ΔL. We have altered the original equation to use CIE LCh rather than Munsell color space coordinates, as is reflected in the hue similarity and lightness contrast terms[1]. Coolness scores are calculated in CIE LCh using a linear interpolation of the original 32 color-coolness mappings, which approximates the number of hue-steps a color is from Munsell 10R, such that greenish blues are cool and orangish reds are not cool (Supp. Mat.). The Pair Preference scoring function reflects people's preference for color

[1] The CIE LCh model explains 51.8% of the variance in Schloss and Palmer's preference data (their Munsell-based model explains 53.5%).

combinations that contain cool colors that differ in lightness and are similar in hue.

$$\mathrm{PP}(c_1, c_2) = 75.15(\kappa_1 + \kappa_2) + 47.61|\Delta L| - 46.42|\Delta H| \tag{6.6}$$

6.3 Colorgorical model

Colorgorical generates color palettes using iterative semi-random sampling. Users specify the number of desired colors and use sliders to set the relative balance of aesthetic preference and discriminability (Sec. 6.2). Generated palettes are displayed to the user as a swatch, map, bar chart, and scatterplot, which highlights how the discriminability may shift with different types and sizes of graphical marks [48, 179].

6.3.1 Minimum discriminability & preference assertions

Each palette is built from an 8,325-color discretized D65 CIELAB space (Sec. 6.2) and is additionally filtered in three ways to help increase discriminability and preference, which we describe below: (1) noticeable difference; (2) lightness clamping (from $L* = 25$ to $L* = 85$) and (3) filtering the dark yellow (generally disliked) region of color space. Although the same RGB coordinates can result in different CIELAB colors on different monitors if monitors are uncalibrated, Stone et al. show that using a fixed correspondence between D65 CIELAB and RGB can be used effectively for online tools in practice [179].

Discriminability The model enforces a lower discriminability bound by sampling *noticeably different colors* using Stone et al.'s noticeable difference function, which provides a minimum CIELAB interval required to discriminate the colors of two graphical marks more than 50% of the time (based on their physical size) [179]. We use a small, conservative visual angle in our calculations ($1/3^\circ$) and multiply the function's suggested interval by three for extra caution.

To ensure discriminability we also exclude colors that are lighter than $L = 85$ and darker than $L = 25$ so that all colors are visible on black or white backgrounds ($L_{\text{black}} = 0$, $L_{\text{white}} = 100$). Colorgorical only includes RGB-valid colors.

Preference Colorgorical excludes the dark yellowish-green region of CIE LCh, which has strongly disliked colors, on average, across many cultures [139, 187, 205]. We define this region as $\mathrm{L} \in [35, 75]$

and H $\in [85°, 114°]$. While there are individual differences in preference [138, 165] and some observers may like these colors [167], the goal is to cater to the average observer. This filter was especially important for generating aesthetically preferable discriminable palettes because of the way Pair Preference and discriminability functions interact. In the Pair Preference equation, the coolness term biases selection toward bluish hues and the lightness term biases selection of contrasting lightness. The discriminability functions bias selection for colors that are far apart in CIELAB color space (i.e., contrasting hue and lightness). Once bluish hues are selected, discriminability would be promoted in subsequent color selections by selecting opposite, yellowish hues of a different lightness level (opposite ends of the $b*$ and $L*$ axes). If the blues are remotely light, then selected yellows will be the dark yellows that people generally dislike. The removal of this region still retained a large region of color space that was sufficiently discriminable to pair with blues, while increasing typical aesthetic palette preference.

To maximize preference within a defined balance, the model generates 10 palettes and returns the palette with the highest minimum-Pair-Preference given all color pairings in each palette.

6.3.2 User-defined model parameters

In addition to specifying the number of colors and manipulating discriminability and preference sliders, users can also configure two optional parameters. First, they can limit color sampling to certain hue ranges (e.g., reds only, or reds and blues), which supports tasks such as designing around brand colors. Second, users can supply an existing palette for Colorgorical to build on. If users provide a palette, Colorgorical rounds the input to the nearest quantized CIELAB color and adds new colors until the palette reaches the desired size.

6.3.3 Palette construction process

Palettes are generated in three steps: (1) initialize, (2) start a palette with the first color, and (3) iteratively add new colors (Fig. 6.2). Colorgorical can typically generate palettes with up to 22 colors before exhausting color space. However, it is inadvisable to use that many colors due to perceptual limitations [60]. If no more colors can be sampled, Colorgorical returns a partial palette and an error message.

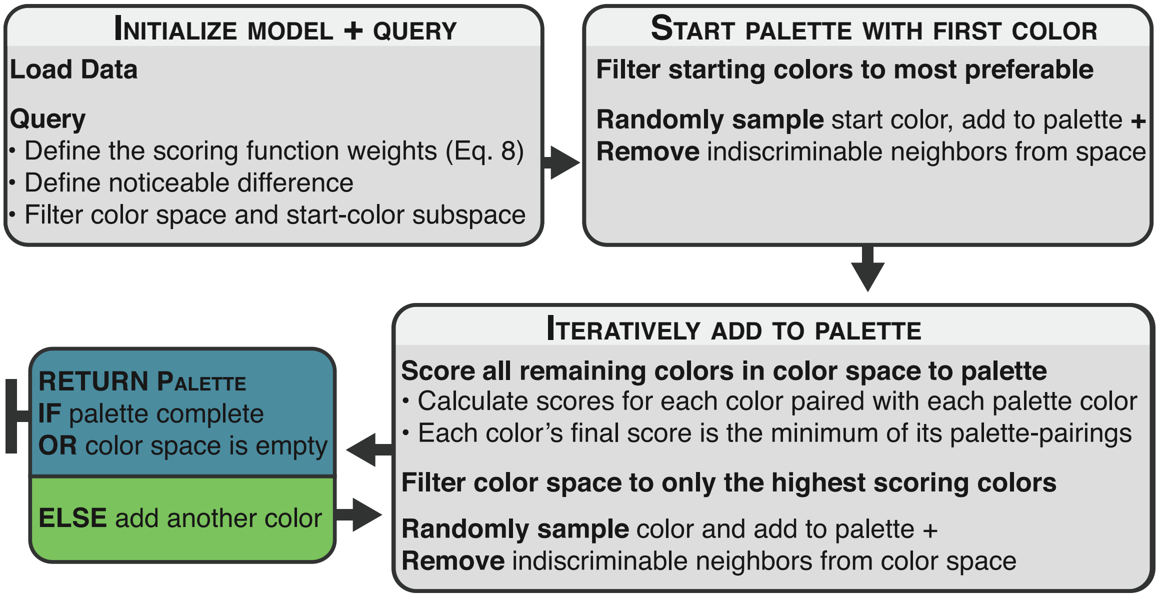

INITIALIZE MODEL + QUERY
Load Data
Query
• Define the scoring function weights (Eq. 8)
• Define noticeable difference
• Filter color space and start-color subspace
START PALETTE WITH FIRST COLOR
Filter starting colors to most preferable
Randomly sample start color, add to palette +
Remove indiscriminable neighbors from space
ITERATIVELY ADD TO PALETTE
Score all remaining colors in color space to palette
• Calculate scores for each color paired with each palette color
• Each color's final score is the minimum of its palette-pairings
Filter color space to only the highest scoring colors
Randomly sample color and add to palette +
Remove indiscriminable neighbors from color space
RETURN PALETTE
IF palette complete
OR color space is empty
ELSE add another color

Figure 6.2: Diagram of Colorgorical palette construction procedure.

Step 1: Initialize

Initialization starts by loading CIELAB space, color coolness scores, and color-name associations into memory. A CIELAB subspace is also loaded into memory, which samples every 15 units along each CIELAB axis and is used along with a precomputed Pair Preference score matrix to pick the first palette color. We use a coarser subspace to select the first color because using precomputed Pair-Preference scores for all pairs of 8,325-colors takes too long for interactivity due to combinatorial explosion. Color space can be filtered based on parameters provided by the user (e.g., hue filters). After applying optional filters, the model limits the subsampled space colors (c) and the color pair preference matrix (Φ) to highly preferable colors (i.e., no dark yellows) using a standard deviation (SD) preference threshold (Eq. 6.7). The threshold removes any color-pair row from Φ whose pair preference score is less than the standard deviation-based limit. Then, the starting color is sampled from the unique colors remaining in Φ's color-pair rows.

$$\text{threshold}(c) = \Phi_c > \max(\Phi) - 0.75 * \text{SD}(\Phi) \tag{6.7}$$

The last initialization step also defines a noticeable difference with Stone et al.'s CIELAB intervals described above, which removes colors that are too similar to each sampled color. Sampled color differences must have at least one axis above the following intervals: $\Delta_L = 22.747, \Delta_a = 31.427, \Delta_b =$

44.757.

Step 2: Start palette

The first color of a palette is selected by randomly sampling a seed color from the remaining colors after Step 1. Next, all colors that are not noticeably different from the seed are removed from color space using the CIELAB intervals defined in Step 1. Sampling is skipped if users provide their own seed color(s), but indiscriminable neighboring colors are still eliminated.

Step 3: Add to palette

To add a new color, the model computes scores for all remaining colors using a weighted sum (Eq. 6.8). This function sums each of the four minimum palette scores ($\vec{\Psi}$) with user-defined weights ($\vec{w}$), given all possible scores between a potential new color (c) and the already picked colors (P). The model uses minimum palette scores assuming that a palette is only as discriminable or preferable as its lowest score. There is also a hue-dependent penalty term (τ) to reduce the likelihood of sampling a color bordering the dark yellow filter region. The new color is then randomly sampled from colors that fall above a score threshold (Eq. 6.7, where Φ is now weighted-sum scores). Non-discriminable colors are removed after sampling.

$$\text{score}(c, P) = \tau(\vec{w} \cdot \vec{\Psi})$$

$$\tau = \begin{cases} 0.75, & \text{if } 115^\circ < c_{\text{hue}} < 138^\circ \wedge c_{\text{L}} \leq 45 \\ 0.8, & \text{if } 70^\circ \leq c_{\text{hue}} \leq 115^\circ \wedge 45 < c_{\text{L}} \leq 75 \\ 0.85, & \text{if } 70^\circ \leq c_{\text{hue}} \leq 115^\circ \wedge c_{\text{L}} > 75 \\ 1, & \text{otherwise} \end{cases} \tag{6.8}$$

6.3.4 Implementation and Performance

Colorgorical is implemented in C-accelerated Python. To evaluate average model runtime (50 runs) as a function of palette size (1 to 20 colors), we profiled single-palette generation on a Mid 2012 MacBook Pro Retina with a 2.6 GHz Intel Core i7 CPU and 16GB 1600MHz DDR3 RAM. Average initialization time was 140ms (SEM = 0.004). If a palette was returned before reaching the required number of colors, it was discarded and the test was run again. Runtime performance increased linearly in the number of colors such that adding a color increased runtime by 17.6ms on average (Supp. Mat.).

6.4 Palette Score Evaluation

Before conducting human-subject testing, we first tested whether any of Colorgorical's scoring functions (i.e., Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference) could be removed from the model to simplify its design without significantly affecting palette output. For instance, if Perceptual Distance were to explain most of the variance in Name Difference scores, then the Name Difference scoring function could be removed from the model with little effect on palette output.

We examined the similarity among the four Colorgorical scoring functions using multiple linear regressions to predict 39,600 *palette scores* for each *palette scoring function* (Sec. 6.2) from the three remaining functions (e.g., predicting Perceptual Distance from Name Difference, Name Uniqueness, and Pair Preference). *Palette scores* are the minimum palette scoring function output given all color pairs in a palette. We use the minimum score because we assumed that a palette is only as preferable or discriminable as its lowest pair. The number 39,600 stems from the full range of possible Colorgorical slider settings and 3 palette sizes (66 settings, {3,5,8}-colors, 200 repeats). The 66 settings were made from the different unique combinations from dragging each of the four sliders to 0%, 50%, or 100%, which ignore duplicate settings encountered when moving one or more sliders to 0%.

We also examined how the four palette scores changed with palette size. Below we highlight results and implications from our analyses, and the methods and full results are in Supplementary Material.

Both Perceptual Distance and Name Difference were strong positive predictors of one another. Name Uniqueness was always a weak negative predictor of the other scores. Pair Preference was always a strong negative predictor of Perceptual Distance and Name Difference. Further, Pair Preference was more strongly related to Name Difference than to Perceptual Distance. Given that palette scores in each palette scoring function were significantly predicted by all three of the other palette scoring functions, we concluded that each scoring function measured sufficiently different color information to justify keeping them all in the model for Experiment 1.

6.5 Exp. 1: Model human-subject evaluation

Experiment 1 tested how palette discriminability performance and preference ratings varied as the relative weights on the Colorgorical sliders varied (i.e., the relative importance of each scoring function; Sec. 6.3). We also identified which slider settings produced the most discriminable or preferable palettes to prepare for a comparison between Colorgorical and current industry standards in Experiment 2 (Sec. 6.6).

Experiment 1 used the same representative palettes as in Section A.5, which were analogous to the slider settings produced by moving each to either 0%, 50%, or 100% for 3-, 5-, and 8-color palettes.

Discrimination performance and preference were assessed using two difference tasks (Fig. 6.3). In the discrimination task, participants reported which side of a map had more counties of a target color, providing data on number of errors and response time (RT). In the aesthetic preference task, participants rated how much they liked the color combinations in each palette. We predicted that:

P1 Palettes with fewer color would be more discriminable

P2 Discrimination RT and error would correlate in a strong negative direction with Perceptual Distance and in a strong positive direction with Pair Preference, whereas preference ratings would show the opposite pattern

P3 Palette size would modulate the discriminability and preference ratings associated with each slider setting.

P4 Slider settings would significantly predict discrimination performance and preference ratings

P1 is based on previous evidence that visualizations with more colors are harder to process [60]. **P2** extends Palette Score Evaluation findings that Perceptual Distance and Name Difference negatively predicted Pair Preference. **P3** builds on the first two predictions: based on **P1** we expect that palette size will modulate the discriminability of slider settings, and based on **P2** we expect that preference will be negatively correlated with discriminability. **P4** makes two strings of assumptions based on the Palette Score Evaluation: (1) the trade-off between discrimination and preference palette scores will extend to behavior (**P2**) and (2) the relative importance of scoring functions (i.e., slider settings) would affect behavior in the same manner as palette scores (e.g., a higher relative importance of Pair Preference will produce higher Pair Preference palette scores). By transitivity, we predict that slider settings will be indicative of behavior.

6.5.1 Methods

Participants

77 participants completed the discrimination task and 60 completed the preference rating task (recruited through Amazon Mechanical Turk, \$3 compensation). Palette size (3-, 5-, 8-colors) was a between-subjects factor. For quality control, we determined *a priori* to discard participants who were $< 60\%$ accurate across all trials in the discriminability task (3-color: $n = 3$; 5-color: $n = 6$; 8-color: $n = 8$). No participants were discarded in the preference task. In the final datasets there were 20 participants per palette size in each task, and discard frequency did not significantly differ between palette size conditions ($\chi^2(2) = 1.793, p = 0.408$). All self-reported having normal color vision and gave informed consent. The Brown University IRB approved the experiment protocol.

Design & Displays

The experimental designs for the discrimination and preference tasks were similar. In both, each participant saw 660 palettes from 66 slider settings (see Sec. A.5.1 for setting information) with 10 different color palettes within each slider setting (treated as repetitions). The specific colors in each palette varied across participants (simulating different runs of Colorgorical), but were generated with the same experimental design. Palette size varied between-subjects (3, 5, or 8 colors).

The palettes that comprised the displays for the discrimination task were also used for the preference task, such that each discrimination participant was yoked to a preference participant (i.e., both saw the same palettes). Palettes were displayed on a predefined map of 554 counties in the U.S. (300×300 pixels). The map itself differed slightly based on the task (Fig. 6.3).

For the discrimination task, a 5-pixel-wide contour bisected the map (adhering to county borders). The contour was black and the county borders were white so that both would fall outside of Colorlogical's default lightness sampling range ($\mathrm{L} \in [25, 85]$; $\mathrm{L}_{\mathrm{Black}} = 0; \mathrm{L}_{\mathrm{White}} = 100$). The size of the counties on each side were slightly altered so they were approximately equal (left: 165 px; right: 163 px). A legend rendered to the right of each map assigned each palette color to a nonsense word category. The target "Neek" color was always at the top of the legend to prevent participants from

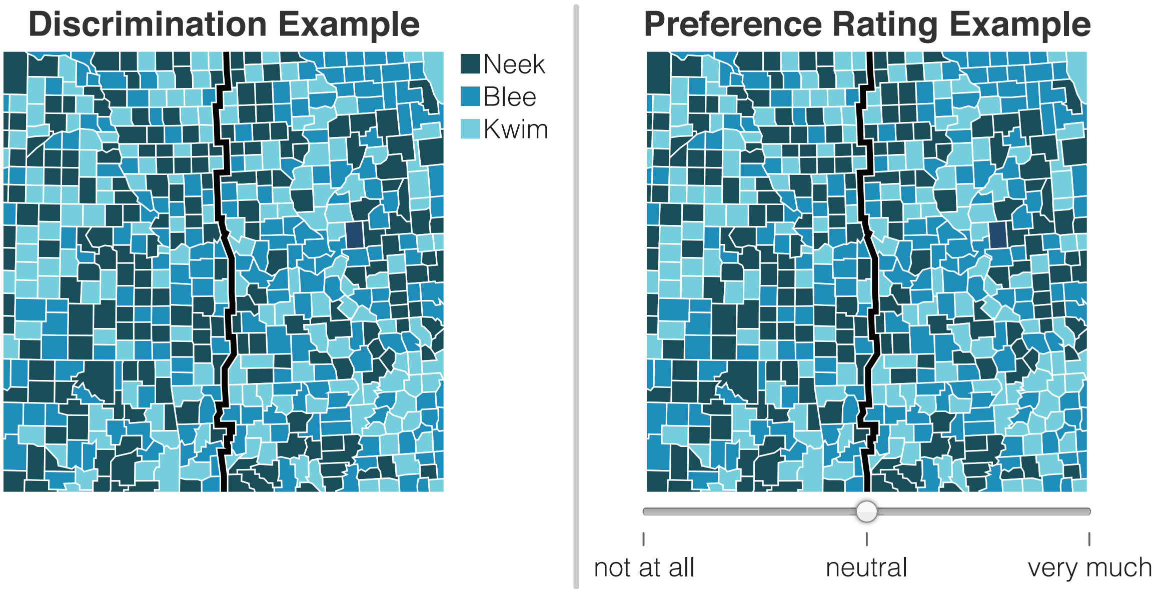


Figure 6.3: Discrimination and preference rating task stimuli. The discrimination task asked users which side had more “Neek” counties (← and → keys). The preference rating task asked users to click on the slider.

having to search for the target color. One side of the map had over-represented target color (“Neek”; 1.5× more frequent on one side than the base rate) and the opposite side had an over-represented distractor color (1.3× more frequent). The target side was left/right balanced across trials. Based on our assumption that a palette is only as effective as its least discriminable pair of colors, the target and distractor colors were always the palette colors with the lowest and second-lowest Perceptual Distance scores compared with all other colors in the palette, respectively.

In the preference task, there was no dividing contour and no legend, the colors were roughly equal in proportion, and they were randomly assigned to positions across the map (no left/right asymmetry). Below the map there was a 300-pixel-wide continuous response slider scale ranging from -100 to 100 with labeled extrema and midpoint (left: “not at all”; right: “very much”; midpoint: “neutral”) [165]. The scale was initialized with the slider set to “neutral” to avoid biasing participants.

Procedure

Discrimination Task. Participants were first presented with an example display and were told that their task would be to indicate which half (left/right) of a map had more Neek counties using the left/right arrow keys. They were also told that the target Neek color would always be shown at the top of a legend and that answers would be marked incorrect if they did not respond within 3.5 seconds. Participants completed five practice trials using distinct displays from the 660 test maps they would see in the experiment, followed by 660 test trials. Maps were shown in random order in the center of the window. Trials were separated by a 500-ms inter-trial interval with a fixation cross displayed at the center of the screen. Optional breaks were given every 20 trials. This task took ~30 minutes to complete.

Preference Task. Participants were asked to rate their aesthetic preference for the color combination in each palette by clicking a point on a slider between the left ("not at all" preferable) and the right ("very much" preferable) ends (Fig. 6.3). To help them gauge what liking "not at all" and "very much" meant to them in the context of these color combinations, participants were shown an anchoring page containing 66 representative maps. They scrolled through the maps and considered how they would rate each map while using the full range of the scale. During the experiment, each map was presented one at a time in a random order (separated by a 250-ms blank pause screen). The preference slider appeared 1 second after the map appeared to encourage participants to consider their preference carefully before responding. This task took ~40 minutes to complete.

6.5.2 Results and Discussion

Before analysis, we pruned response time (RT) data by removing incorrect trials and then eliminating trials for each subject that were more than ±2.5 standard deviations away from their mean RT [147]. On average, 129 errors (19.5%) and 24 outliers (3.6%) were removed.

Overall participant accuracy decreased as palette size increased (3-color average error: 79/660; 5-color: 119/660; 8-color: 190/660), indicating that displays with smaller color palettes were more discriminable (**P1**, **P3**). This result mirrored the increased participant discard rate for larger palette size conditions due to high error rates (Sec. 6.5.1) and is consistent with previous findings that showed visualizations with fewer color categories are more effective [60]. Between-subjects one-way ANOVAs testing for effects of palette size (3, 5, 8) within each measure indicated significant effects for number of errors ($F(2, 57) = 30.801, p < .001$) but not for RT or preference ($F(2, 57) = 1.574, 1.035; p = 0.216, 0.362$, respectively).

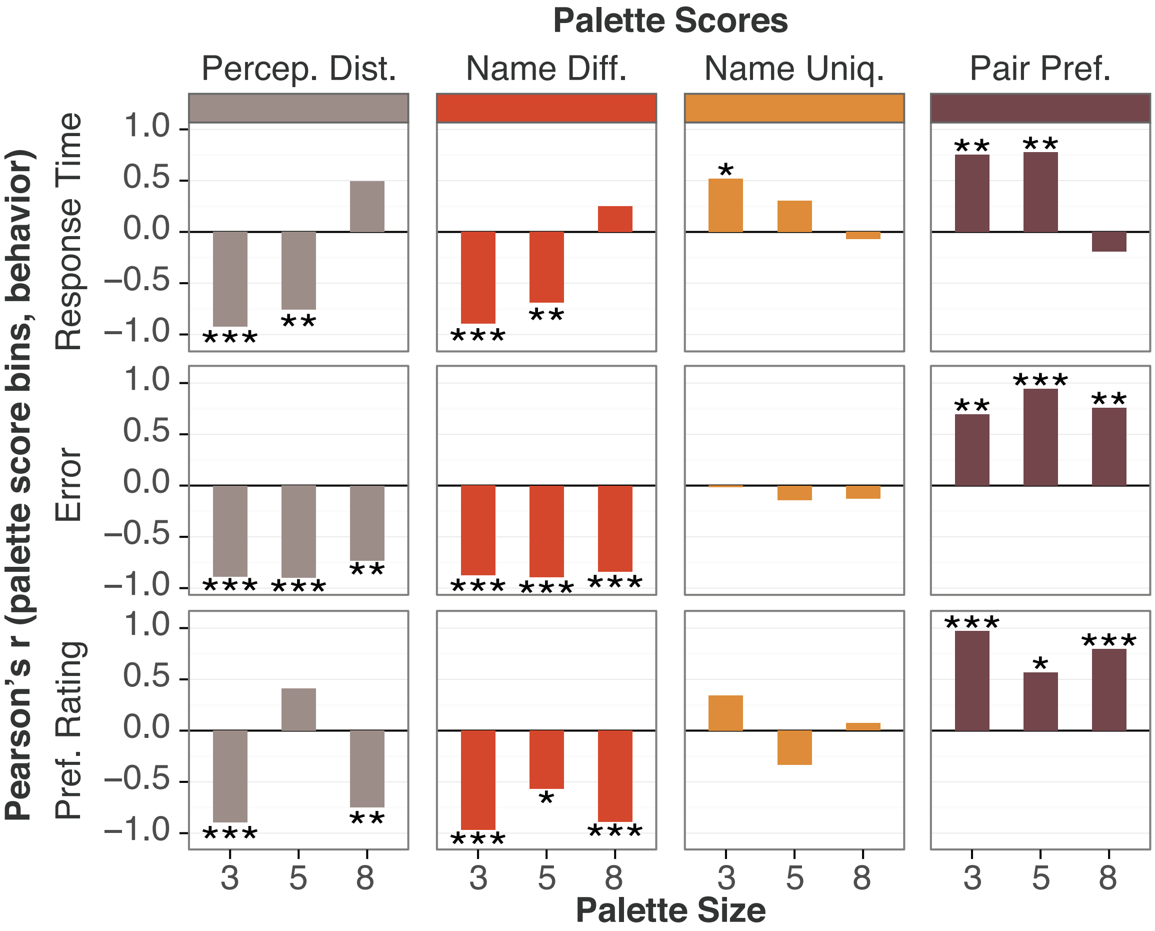


Figure 6.4: Correlations between binned palette scores and responses on each measure for each palette size ($* : p < 0.05; ** : p < 0.01; *** : p < 0.001$).

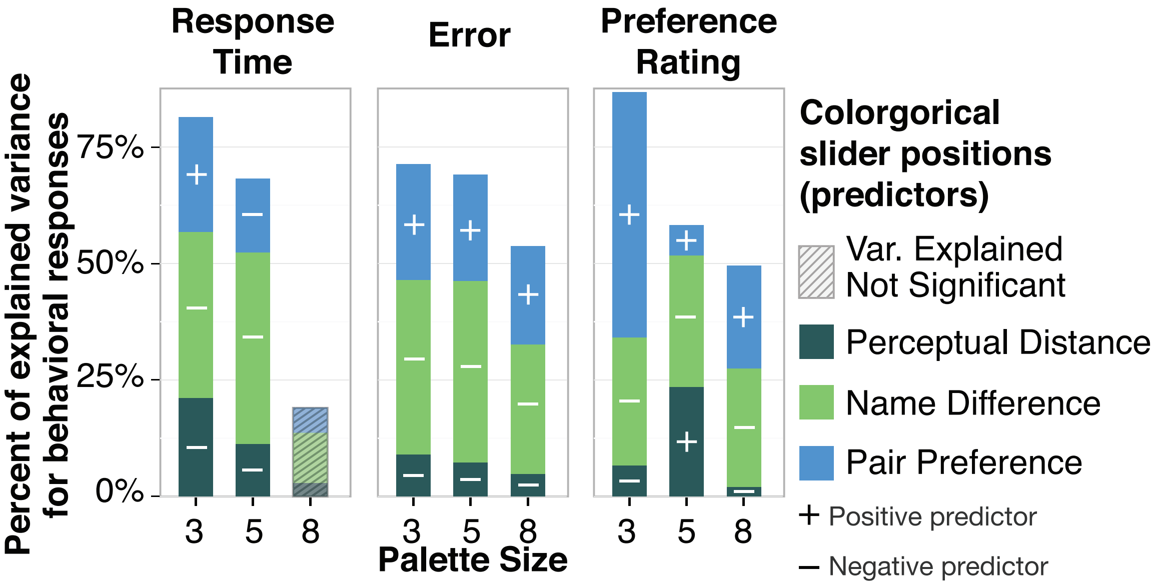


Figure 6.5: Variance explained (R^2) for the 9 models decomposed to look at the variance explained of behavioral data in terms of slider settings (i.e., palette score relative importance).

Palette score and behavioral measure correlations

Figure 6.4 shows the correlations between each type of palette score (Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference) and the three behavioral measures (RT, error rate, and preference ratings), averaged over participants. *Palette score* refers to the lowest *palette scoring function* value (Sec. 6.2) given all color pairs in a palette. To conduct these analyses, we first binned the behavioral data for each measure according to palette scores (15 equally-spaced bins) for each subject[2]. After, we averaged the data for all palettes that scored in the same bin and then averaged those values across participants. This binning was necessary prior to averaging across participants because each participant saw different palettes with slightly different scores (Supp. Mat.). For example, RT for palettes with a Pair Preference scores of 30.03 and 30.05 would be binned together.

We cross-checked the binned-score correlations by calculating the within-subject correlations for each behavioral measure and palette score and then used Fisher's Z transform prior to calculating the between-subject average Pearson's r for each measure and score combination. For the most part, these analyses showed the same pattern of results as the binned correlation statistics (Supp. Mat.).

The binned-score correlations are presented below (see Supp. Mat. for individual correlations

[2] The degrees of freedom for 8-color perceptual distance correlations is one less due to an empty bin, which is shown in the Supplementary Material.

on non-binned data). In summary, the Perceptual Distance, Name Difference, and Pair Preference scores had the predicted effects: RT and error rates decreased (i.e., better performance) as Perceptual Distance and Name Difference increased, but they increased (i.e., worse performance) as Pair Preference increased (**P2**). In contrast, preference decreased as Perceptual Distance and Name Difference increased and they increased as Pair Preference increased. Name Uniqueness had little to no effect.

RT. RT decreased as Perceptual Distance and Name Difference increased for 3- and 5-color palettes (Perceptual Distance: $r(13) = -0.926, r(13) = -0.757; p \leqslant 0.001$ respectively; Name Difference: $r(13) = -0.893, r(13) = -0.689; p \leqslant 0.005$ respectively). Similarly, Pair Preference followed **P2** for 3- and 5-color palettes with strong positive correlations with RT ($r(13) = 0.755, 0.776; p = 0.001$). Name Uniqueness was significantly correlated with RT for 3-colors ($r(13) = 0.521; p = 0.046$), but not for 5-colors ($r(13) = 0.306; p = 0.268$). No scores were significantly correlated with RT for 8-color conditions ($r(12) = 0.496, r(13) = 0.251, -0.071, -0.193; p \geqslant 0.071$ for Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference respectively). These findings largely support **P2** for 3 and 5 colors; however, 8-color palette correlations were not significant.

Error. Error rate correlations were significant for all sizes with Perceptual Distance ($r(13) = -0.887, -0.898, r(12) = -0.731; p \leqslant 0.003$, for 3-, 5-, and 8-colors respectively), Name Difference ($r(13) = -0.874, -0.892, -0.838; p < 0.001$), and Pair Preference ($r(13) = 0.697, 0.945, 0.761; p \leqslant 0.004$). Similar to RT correlations, Name Uniqueness was not significantly related to error measures ($r(13) = -0.016, -0.141, -0.126; p \geqslant 0.616$).

Preference Rating. Preference rating trends were the opposite of error and RT, and consistent with **P2**. Increasing 3- and 8-color Perceptual Distance reduced preference ratings, ($r(13) = -0.897, r(12) = -0.751; p \leqslant 0.002$) but not significantly so for 5-color palette ($r(13) = 0.412; p = 0.127$). Preference ratings also decreased as Name Difference increased for 3-, 5-, and 8-colors ($r(13) = -0.969, -0.57, -0.891; p \leqslant 0.026$, respectively). Increasing Pair Preference increased preference ratings ($r(13) = 0.971, 0.57, 0.796; p \leqslant 0.026$). Again, Name Uniqueness was not significantly related ($r(13) = 0.346, -0.333, 0.073; p \geqslant 0.207$).

Predicting behavioral measures from slider settings

To test whether slider settings (i.e., relative importance of the *palette scoring functions*) significantly predict behavior (**P4**), we performed a series of multiple linear regressions that predicted behavioral measures as a function of changing sliders to 0%, 50%, or 100% (Fig. 6.5). Given that the correlational analyses above suggested that Name Uniqueness had little effect on behavior, we averaged slider configurations that would be equivalent if Name Uniqueness were ignored. For example, if Perceptual Distance and Name Uniqueness were both set to 50%, the new setting would be Perceptual Distance as 100% and would be averaged with other palettes where Perceptual Distance is 100%. This reduced the regression analysis to predict 20 unique slider settings rather than the previous 66. The data that were input to the correlations are graphed in the Supplementary Material.

Below we detail the results of the multiple linear regressions using slider relative importance to predict the three behavioral measures. More information about the relation between sliders, size, and behavioral measures is provided in the Supplementary Material. In summary, the slider settings were typically able to significantly predict the behavioral measures (**P4**).

RT. RT decreased (improved) as Perceptual Distance and Name Difference slider weights increased and RT increased (got worse) as Pair Preference slider weights increased (**P4**; Supp. Mat.). Name Difference was always the most predictive and Perceptual Difference and Pair Preference were similarly less predictive (Fig. 6.5). Although this pattern was present for all three palette sizes, the models were significant for the 3- and 5-color palettes ($F(3,16) = 23.442, 11.447; R^2 = 0.815, 0.682; p < 0.001$, respectively), but not the 8-color palettes ($F(3,16) = 1.267, R^2 = 0.192, p = 0.319$). The lack of significance for 8-color palettes coincides with the oddity that response time was typically faster for 8-color palettes than 5-color ones; this is unexpected, given (1) past visual search research finding that more colors take longer to discriminate [60] and (2) the previously discussed palette size relation with accuracy. We suspect that this difference may be because participants tried less hard or the task became too difficult in the 8-color condition because they had higher overall error rates. Another possibility is that pair-based color discriminability scores (e.g., Perceptual Distance) may break down as the number of colors increases, which would create a need for higher-order combination discriminability scores. Each of these possibilities raise interesting future directions for studying the relation between palette effectiveness and number of colors.

Error. Slider relative importance analysis mirrored RT (**P4**; Supp. Mat.), except that Pair Preference was more important than Perceptual Distance (Fig. 6.5 and Supp. Mat.). The reason for this difference is unknown. The multiple linear regressions for all 3-, 5-, and 8-colors were all significant ($F(3, 16) = 13.186, 11.964, 6.192; R^2 = 0.712, 0.692, 0.537; p \leqslant 0.005$, respectively).

Preference Rating. Preference ratings increased with weights on the Pair Preference slider and decreased with weights on the Perceptual Distance and Name Difference sliders (**P4**; Supp. Mat. slider-behavior figure). Pair Preference was the most predictive slider for 3-colors, but not for 5- and 8-colors (Fig. 6.5 and Supp. Mat.); instead, Name Difference was most predictive. Perceptual Distance was more important than Pair Preference for 5-colors, but was otherwise the least important slider. The multiple linear regressions for 3-, 5-, and 8-colors were all significant ($F(3, 16) = 35.089, 7.396, 5.228; R^2 = 0.868, 0.581, 0.495; p \leqslant 0.01$, respectively).

It is noteworthy that the model's ability to predict preference ratings decreased for 5-colors relative to 3-colors, suggesting that the mechanism behind human aesthetic preference ratings may deviate from pair-based preference predictions as the number of palette colors changes. Another difference for 5-color palettes, compared to 3- and 8-colors, was that all settings were rated either neutral or slightly negative. These results suggest that the assumption that pair-wise based preference models generalize to palettes of three colors might break down for larger palettes. The differences in preference ratings over palette sizes motivates the need for further research on the aesthetics of higher-order color combinations.

We also found that preference ratings decreased faster as Name Difference relative importance was increased compared to increases in Perceptual Distance relative importance (see Supp. Mat.). This asymmetry might be caused by differences in how Perceptual Distance and Name Difference measure distances in color space. It could be that Perceptual Distance is more supportive because it can generate color pairs that differ primarily in lightness (which is one of the terms in Pair Preference), whereas Name Difference might be more likely to favor differences in hue, which would be in opposition to Pair Preference's hue similarity term.

Lowest-Error and Highest-Preference settings

A main goal of Experiment 1 was to determine which Colorgorical settings to use to generate color palettes for comparison against current standards (Experiment 2). The combinatorial explosion of conditions prevented comparing all slider combinations to current standards. Therefore, we chose to select slider settings that either produced highly discriminable or highly preferable palettes (i.e., at either end of the previously-discussed discriminability-preference trade-off). Figure 6.7 shows the lowest discrimination error setting (subsequently called "Low-Error" palettes) and the highest preference rating setting ("Preferable" palettes) for each palette size. There were significantly fewer errors for Low-Error palettes than for Preferable palettes ($t(19) = 3.322, 7.589, 3.15; p \leqslant 0.005$, 3-,5-,8-colors). Preference ratings were significantly greater for Preferable palettes than for Low-Error palettes for 3- and 8-colors ($t(19) = 4.610, 2.841, p \leqslant 0.01$), but not for 5-colors ($t(19) = 0.499, p = 0.623$) (consistent issues about 5-color palettes discussed above).

Summary

Experiment 1's results largely support each of our four predictions and suggest that Colorgorical's sliders are effective at controlling the discriminability and preference of color palettes, although some 5- and 8-color conditions led to unexpected behavioral results. Discriminability performance typically improved (faster RT, fewer errors) as the Perceptual Distance and Name Difference palette scores increased (and with greater weight on their corresponding sliders) and Preference judgments typically increased as Pair Preference palette scores increased (with greater weight on its slider) (**P2**). There was also evidence for a tradeoff – discriminability decreased as both Pair Preference scores and scoring function weights increased, and preference judgments decreased as Perceptual Difference and Name Difference increased. This finding supports our earlier claim that care must be taken to design palettes that balance both discriminability and aesthetic preference. We also found that Name Difference, not Perceptual Distance, might better predict discriminability. This would also support Demiralp et al.'s previous findings that suggested Name Difference is a better measure of color distance than Perceptual Difference [26].

Additionally, our results suggest that smaller palettes are more discriminable (**P1**), that palette size modulates discriminability and preference ratings (**P3**), and that slider configurations significantly predict behavior (**P4**). We provide additional analysis and discussion for each prediction in the Supplementary Material.

Last, differences in discriminability and aesthetic preference trends across palette sizes motivate additional research beyond pairwise theoretical models of color discrimination and preference rating.

6.6 Exp. 2: Colorgorical-Others benchmark

Experiment 2 compares palettes generated by Colorgorical Low-Error and Preferable slider settings to commonly used "benchmark" palettes (ColorBrewer, Microsoft Excel, and Tableau; Fig. 6.6). We also included randomly sampled palettes with noticeably different colors to simulate palettes made by someone without design expertise who tried to choose colors that were not confusable. We predicted that:

P Colorgorical Low-Error and Preferable settings would produce palettes that are at least as discriminable and typically more preferable compared to the majority of benchmarks

We based this prediction on expected outcomes of Colorgorical and benchmark palettes by applying regressions modeled on Experiment 1 palette scores and behavioral responses to the palette scores of Experiment 2 palette sets. As shown in Figure 6.7, Colorgorical palettes were expected to create more preferable palettes, with the exception of Microsoft 5- and 8-color palettes, which were predicted to outperform both Colorgorical settings. We also expected that Colorgorical would produce palettes with error rates similar to Tableau across all three sizes. We specified planned comparisons to test these predictions with the human-subject data from Experiment 1.

6.6.1 Methods

Participants

75 participants (recruited through Amazon Mechanical Turk; paid $1) completed the discrimination task and 60 completed the preference task. All gave informed consent, and the Brown University IRB approved the experiment protocol. All self-reported having normal color vision. 15 discrimination participants were less than 60% accurate and were discarded, per Experiment 1 procedure (3-colors: $n = 0$, 5-colors: $n = 7$, 8-colors: $n = 8$). Participants were divided equally across size conditions ($n = 20$ per size), and there was a significant effect between discard rate and size ($\chi^2(2) = 6.878, p = 0.032$).

Design, Displays, & Procedure

Palette size (3,5,8) varied between subjects and the rest of the factors varied within-subject. Participants in the discriminability task completed 96 trials (6 palette sets {Colorgorical Low-Error

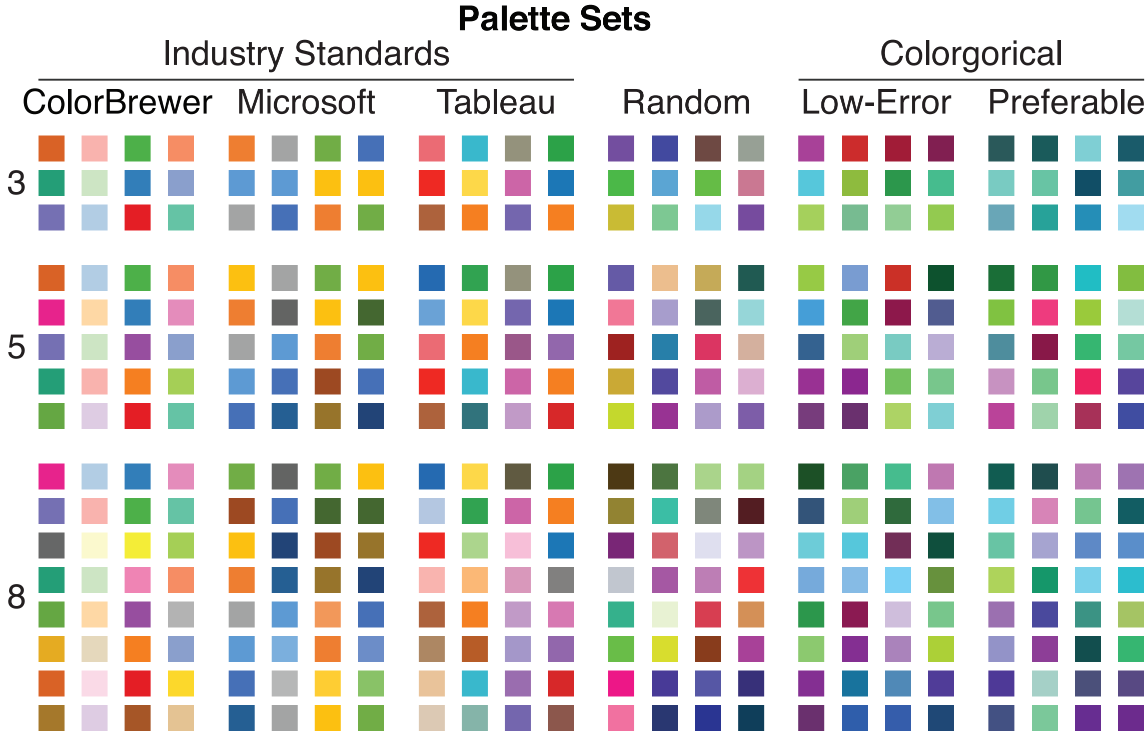
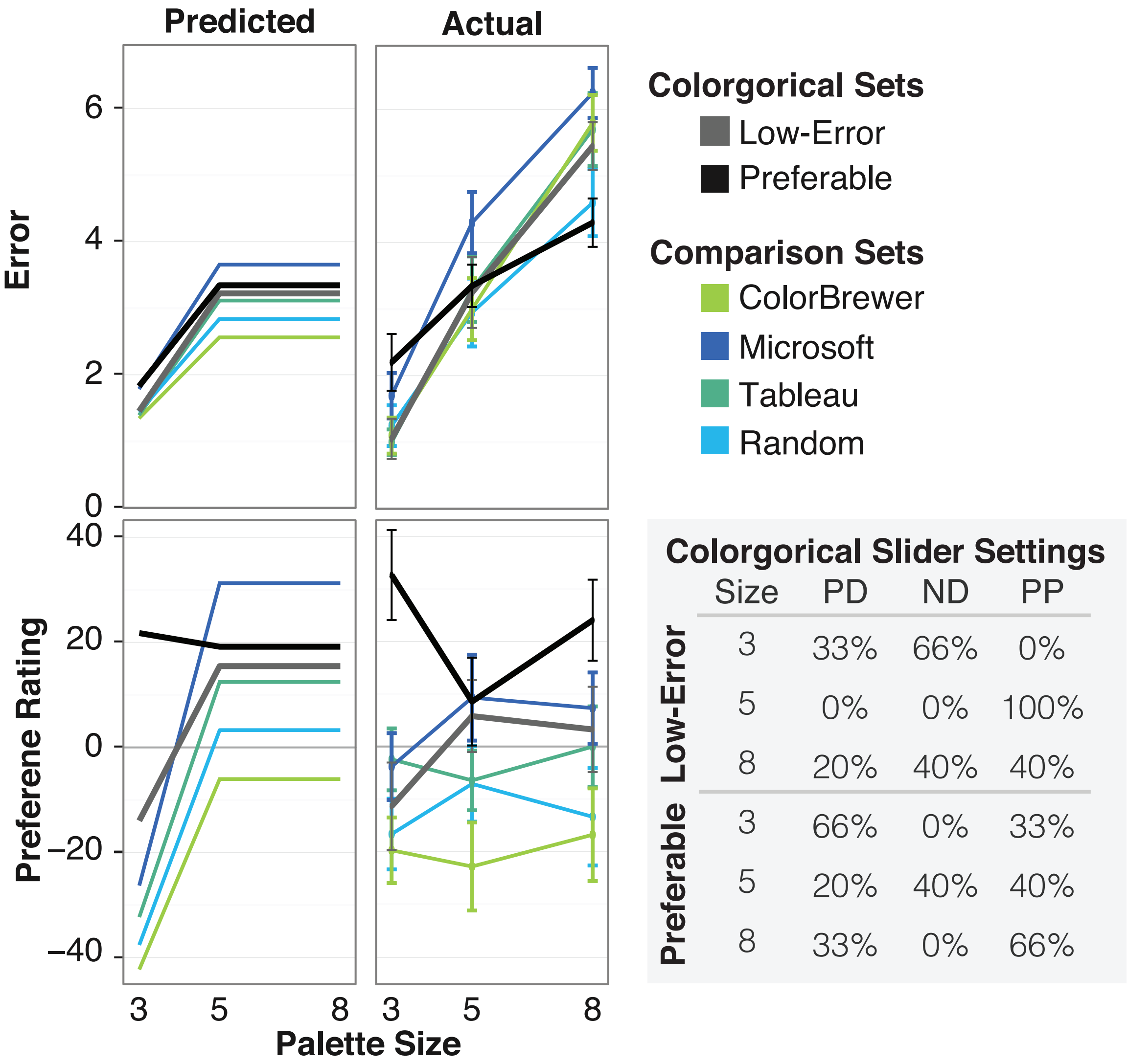


Figure 6.6: Exp. 3 palettes: ColorBrewer (Dark2, Pastel1, Set1, Set2); Microsoft (all); Tableau (10, Blue-Red, Green-Orange, Purple-Gray); Colorgorical and Random palettes varied across participants.



Colorgorical Slider Settings

	Size	PD	ND	PP
Low-Error	3	33%	66%	0%
	5	0%	0%	100%
	8	20%	40%	40%
Preferable	3	66%	0%	33%
	5	20%	40%	40%
	8	33%	0%	66%

Figure 6.7: Palette-behavior predictions and actual results for Experiment 2 (e.g., 4 errors = 25% error rate). Prediction models were trained on Experiment 1 palette scores and behavioral responses. Error bars show SEM. The table shows the Exp.2 Colorgorical slider settings (PD: Perceptual Distance; ND: Name Difference; PP: Pair Preference).

and Preferable, ColorBrewer, Microsoft, Tableau, Random} × 4 palettes taken from each set × 4 repetitions). Participants in the preference rating task were presented with 24 trials (6 palette sets × 4 palettes, no repetition).

The benchmark palette sets included four palettes from each palette group's larger collection (Fig. 6.6). Microsoft palettes included all four available palettes in Microsoft Excel for Mac (v.15.8). ColorBrewer palettes included four of the eight available palettes, including those with the greatest minimum Perceptual Distance and excluding palettes with niche purposes (e.g., "Paired") [64]. Tableau palettes included the default Tableau 10 and the three palettes that were not designed for niche applications. We created random palettes by randomly sampling discriminable colors in RGB space for each participant (Sec. 6.3.3). All participants saw the same benchmark palettes aside from random. Each participant was given different random and Colorgorical palettes to test each palette type's full potential variance. The Low-Error and Preferable palettes were made with settings described at the end of Section 6.5 (Fig. 6.7). Otherwise, the design, stimuli, and procedure were the same as Experiment 1. The discrimination task took ~5 minutes to complete and the preference rating task took ~10 minutes to complete.

6.6.2 Results and Discussion

We focused only on error and preference rating data (not RT) because error and RT results in Experiment 1 were similar and because we chose the Colorgorical palettes based on error rates and preference ratings. All reported *t*-tests were paired sample and two-tailed.

We first conducted two 6 palette set (within-subject) × 3 palette size (between-subject) mixed-design ANOVAs: one for error rates (averaged over replications) and a second for preference ratings. For error, there were main effects of palette set ($F(5, 285) = 3.538, p = 0.004$), palette size ($F(2, 57) = 59.34, p < 0.001$), and a 2-way interaction between them ($F(10, 285) = 5.896, p = 0.01$). For preference ratings, there was a main effect of palette set ($F(5, 285) = 13.235, p < 0.001$) with no effect of palette size ($F(2, 57) = 0.258, p = 0.773$) and no interaction ($F(10, 285) = 1.283, p = 0.239$). As shown in Figure 6.7, error increased with size, but preference ratings were more stable as size increased. Palette set differences are shown through the vertical separation of behavioral responses across palette sets. Our planned comparisons below delve into these effects, and they largely support the trends in our predictive models based on palette score with (although size does not show the predicted effect for preference ratings).

Colorgorical Low-Error vs. Preferable Palettes

We first tested whether the error and preference differences between Colorgorical-Low-Error and -Preferable palettes replicated the results of Experiment 1. As in Experiment 1, the Preferable palettes were preferred to the Low-Error palettes for 3- and 8-color palettes ($t(19) = 3.573, -3.79; p = 0.002, 0.001;$), but not for 5-color palettes ($t(19) = -0.405, p = 0.690$). There were fewer errors for the 3-color Low-Error palettes than for the Preferable palettes ($t(19) = 3.286, p = 0.004$), but there was no difference for the 5-color palettes ($t(19) = 0.195, p = 0.847$). The only test that was inconsistent with our previous findings was that error rates for 8-colors were lower for Preferable palettes than for Low-Error palettes ($t(19) = 2.113, p = 0.048$). The reason for this result is unknown.

Comparing Colorgorical to industry standard palettes

We next tested our prediction that Colorgorical palettes would be as discriminable and typically more preferable than the benchmark palettes. The tests were planned *a priori* based on predictions from Colorgorical and benchmark palette scores described below (Fig. 6.7). We conducted 48 paired two-sample *t*-tests comparing participants' discrimination error and preference ratings within the Colorgorical palettes and between the Colorgorical palettes and the four benchmark palette sets within each palette size (Fig. 6.7).

Error rate. Based on the model predictions (Fig. 6.7), we expected that error would not significantly differ between Colorgorical Low-Error palettes and all benchmarks except for Microsoft, where we predicted that Low Error palettes would elicit fewer errors. For 5- and 8-colors we predicted that Low-Error errors would be similar to Tableau, worse than ColorBrewer and Random, and better than Microsoft. We made the same predictions for Preferable palettes, except that 3-color error might only be as good as Microsoft, and 5- and 8-color error might be worse than Tableau.

Performance for Low-Error palettes was slightly better than expected. There were significantly fewer errors for 5-color Low Error than for 5-color Microsoft ($t(19) = 2.396, p = 0.027$) and no significant difference from the other benchmarks ($t(19) < 1.628, p \geqslant 0.12$).

Colorgorical-Preferable error also matched our predictions because there was always at least one benchmark that had non-significantly different error rates compared to the setting ($t(19) \leqslant 1.898, p \geqslant 0.073$). Unexpectedly, Colorgorial-Preferable palettes led to significantly lower error than 8-color ColorBrewer and Microsoft palettes ($t(19) \geqslant 2.910, p \leqslant 0.009$). However, consistent with our predictions, 3-color Colorgorial-Preferable led to significantly more errors than ColorBrewer, Tableau, and Random benchmarks ($t(19) = 2.531, 3.644, 3.047; p = 0.020, 0.002, 0.007$, respectively).

The fewer errors for random than for Colorgorical preferable may be surprising, but it is consistent with our earlier observations. There is a high likelihood that three randomly sampled colors will be far apart in our quantized CIELAB space, leading to very high discriminability but also low preference. As the number of randomly sampled colors increases, discriminability decreases, as shown in the non-significant comparisons to 5- and 8-color Colorgorical Preferable. Although the Colorgorial-Preferable settings produced less discriminable results in some conditions (e.g., 3-color error), there was always at least one benchmark that lacked significantly different error rates.

Preference ratings. We predicted that both Low-Error and Preferable palettes would be more preferable in all comparisons except to 5- and 8-color Microsoft palettes (Fig. 6.7).

Low-Error was significantly more preferred than 5-color ColorBrewer and 8-color Random ($t(19) = 2.784, 2.279, p = 0.012, 0.034$, respectively) and was never significantly less preferred than the other benchmarks ($t(19) \leqslant 1.781, p \geqslant 0.091$). Colorgorial-Preferable palettes often led to significantly more preferable palettes (8 of 12, all but 5- and 8-color Microsoft, 5-color Random, and 8-color Tableau; $t(19) \geqslant 2.105, p < 0.05$).

Summary. Colorgorical Low-Error and Preferable palettes are almost always as discriminable and often more preferable than the current standard visualization-specific categorical color palettes (**P**). Low-Error palettes were sometimes more discriminable and more preferable or otherwise not significantly different than the benchmark palettes. Similarly, Preferable palettes often led to significantly higher preference ratings, and discriminability was not significantly different compared to at least one industry standard for all sizes. Thus, Colorgorical allows users without design expertise to create discriminable and preferable palettes that often do not have significantly different discriminability and that sometimes are more preferable than current pre-made standards.

6.7 Open research areas

We found that Colorgorical palettes, based on models of aesthetics and discriminability, can be as effective as expert-made visualization palettes and even more aesthetically preferable. These findings lead to several future research directions. First, given that color combination discriminability and preference can be inversely related, how can discriminability and preference be automatically optimized? Second, what alternatives to the current pairwise theoretical models might better predict discriminability and aesthetic preference for higher-order combinations (e.g., 5- or 8-colors)? Third, how would color preference models that diverge from figure/ground preference alter palette construction? For instance, how might Lin et al.'s preferable palette generation technique that learns from artist-generated training palettes [112] compare to palettes made with Pair Preference? Fourth, would the same results hold if hue filters are applied when constructing Colorgorical palettes? Fifth, how might Colorgorical help designers foresee palettes that might be indiscriminable given color deficiencies [149]?

6.8 Conclusion

We presented Colorgorical, a model-driven approach to generating categorical color palettes for information visualizations by configuring palette discriminability and preference. Colorgorical uses an iterative, semi-random-sampling procedure to generate palettes of a specified size. User-defined configurations work by changing the relative importance of Perceptual Distance, Name Difference, and Pair Preference scoring functions. Users can further customize palette creation by modifying the number of colors, by defining which hues to sample from, and by providing an existing palette to build upon.

The novelty of our approach stems from our departure from previous palette creation strategies. Whereas previous palette creation tools focused primarily on discriminability or favored color relations in harmonic templates whose empirical validity is questionable (e.g., Adobe Color [132]), Colorgorical generates palettes with user-defined relative importances for discriminability and aesthetic preference (Sec. 6.2). Our color sampling approach also differs in strategy from pre-made palette sets such as ColorBrewer, in which categorical palettes are generated by first choosing colors representing different names and then varying each palette color's value [17].

Empirical tests show that each of Colorgorical's sliders, which are used to balance palette discrimination and preference, measure different aspects of color (Sec. A.5) and modulate behavior as they were designed to do (e.g., weighting discriminability sliders increases discriminability performance) (Sec. 6.5).

Empirical tests that compare Colorgorical palettes and industry standards revealed that our model-derived palettes are as effective as, and sometimes better than, current categorical color palette standards. Our findings also indicate that the number of colors may alter the effectiveness of pair-based discriminability and preference scores. Colorgorical also improves upon industry standards by giving users the flexibility to create their own discriminable and preferable palettes while enforcing visualization design constraints. These results indicate that Colorgorical provides an effective way to create categorical visualization color palettes. Colorgorical is open-sourced at `http://vrl.cs.brown.edu/color`.

Chapter 7

Conclusion

The purpose of this chapter is to integrate each thesis contribution into the larger context of visualization design research posed by our thesis statement, and to identify related future research opportunities. We begin with brief closing remarks that discuss the potential broader impact of this thesis. Afterwards, we explore how this thesis provides a platform to investigate topics such as automating visualization stylization, expanding common definitions of "effective design," and supporting adaptive visualization design. We end with a discussion of the summative contributions of this thesis, and how the contributions in this thesis support our thesis statement: "that visualization design can be broadly empowered and improved through the creation of computational design assistance tools based on new theoretical knowledge of graphical perception and task requirements."

7.1 Closing Remarks

This thesis expands our knowledge of visualization design research and explains how graphical perception and task requirements theory can be used to build effective design assistance techniques. While expanding this knowledge, we also increased our present understanding of how visualization is used in cancer genomics research and how to better design visualization tools to support cancer researchers' analytical workflows. Rather than replace the role of designers, we believe the advances described in this thesis instead augment and enhance designers' abilities to do their jobs by reducing the need to perform under-defined, time-intensive, and often menial tasks. For example, designers can quickly iterate through discriminable and preferable palettes with Colorgorical rather than spending a day trying to manipulate many colors in CIELAB color space. Although this thesis is only a step towards realizing more complete visualization design automation, we believe it provides a solid foundation from which future endeavors within and outside visualization design research can be built from.

7.2 Looking forward: research opportunities and directions

In the closing of each previous chapter we briefly outlined testable predictions and other actionable open research questions. Here, we hypothesize about additional, long-term open research problems that extend beyond the smaller-scale possibilities we previously discussed and that immediately build off of our thesis contributions.

7.2.1 Visualization tool design automation

Hypothesis: evaluating vision science principles in the context of visualization will improve our understanding of how effective visualization design can be quantified. Psychophysical phenomena like Weber's law [63, 90] and Treisman's "pop out" effect [48, 60] alongside many other vision science principles have made it possible to begin quantifying what makes visualization effective. We believe that these inquiries are only just scraping the surface of how we can leverage psychophysical explanations to systematically understand what does and does not promote effective visualization design. Future directions could cover topics such as potential interactions between attributes of visual appearance or into perception of changing visualizations. One possibility might be further exploring the relationship between the size of visualization and various effects it has on design effectiveness. In addition to the visual search work explained in Chapter 5, Stone et al. investigated visual angle's effect on color discriminability with similar "diminishing returns" benefits as size is increased [177]. It would be interesting to further investigate this phenomena to see if this is a more widespread psychophysical phenomena in graphical perception.

Hypothesis: graphical perception and scene decomposition can be leveraged to create helpful style checking and recommendation techniques. Is is now possible to decompose visualizations into their graphical primitives [61], use these primitives to power novice-friendly design tools [162], and to automatically recommend categorical color palettes that satisfy users' own definitions of effectiveness (Chapter 6) [49]. Based on preliminary experimental data [169], we believe that this line of research can be further extended to more robustly recommend visualization design. Pursuing design automation along this vein could result in new scene-aware usability predictions (e.g., accuracy), "style checking" warnings, and perhaps even provide alternative designs akin to what Bricolage did for web design by automatically redesigning website themes [101]. While fully automated redesign might be far away, it is possible that learning algorithms could decompose vector formatted visualizations to predict visual analysis task accuracy, similar to how we predicted color discriminability in Chapter 6. Similar techniques might be also used in conjunction with empirically derived functions, such as Stone et al.'s noticeable difference work [179], to implement legibility warnings or to identify designs that might not be viewable by those with color vision deficiencies.

7.2.2 Expand the usefulness of design principle contributions

Hypothesis: inclusive perspectives on "effective design" will better characterize the growing diaspora of visualization audiences. Given that the field of visualization was largely founded to support medicine, science, and intelligence operations, the bulk of design studies [170] and other visualization tool evaluations have led us to formalize design recommendations skewed heavily for domain expert populations. For example, in Chapters 3 and 4 we evaluated how domain expertise diversity can affect cancer genomics research tool design. While important, it is unclear how domain expert centered contributions apply to the growing prevalence of visualization in the day-to-day lives of the general public (e.g., in The New York Times or at the doctor's office). For example, Hakone et al. found that although visualization interaction is often viewed as a safe way to improve data comprehension, it instead limited elderly cancer patients' understanding of personalized predictive mortality models [58]. Similarly, visualization accessibility remains a significant issue: individual differences such as color vision deficiencies or visual acuity are rarely studied despite the fact that these differences can profoundly affect how data is understood, or if data can be understood at all. Evaluating how effective design may differ for non-domain-expert, non-WEIRD (Western, educated, industrialized, rich, and democrat), or otherwise atypical populations is both largely under-explored and a critical area of study.

Hypothesis: the utility of interaction task taxonomies can be evaluated through computational interaction log classification. Organizing visualization interaction into taxonomies and other structures is a perennially topical visualization design research area [14]. However, a growing point of discussion within the visualization research community is how to measure the utility of these structures. These concerns largely pertain to how often taxonomies are predictively applied and whether they are useful research contributions [97]. We believe it is possible to answer open questions like these and work towards better and more effective interaction characterizations by implementing established taxonomies into classification models or other machine learning techniques, which in turn can be systematically evaluated. Such attempts could resemble our interaction log classification research and might also help address the aforementioned design generalizability concerns by modeling requirements on larger, more representative tool user populations.

7.2.3 Dynamic and adaptive visualization design.

Hypothesis: combining design automation with task modeling will allow evaluators to systematically test perceptual-behavioral nudging for visual analysis. Human decision making is feathered with different kinds of heuristics that enable us to navigate uncertainty in the world around us [195]. Finding ways to counter potentially maladaptive subconscious use of these heuristics (*cognitive biases*) could improve the effectiveness of visual analysis and of visualization design evaluation. For example, humans often use representativeness assumptions when looking at data, which can lead to judgements affected by sample bias [86]. Chapter 4 provides preliminary work in this direction within the context of performing more representative evaluation. Working at a psychophysical level, Feng et al. also studied how visual cues could serve as nudges that increase information space exploration during visualization use [35]. It would be interesting to incorporate nudging with A/B testing to test whether cognitive biases can be avoided throughout the visual analysis workflow by leveraging interaction log analysis.

Hypothesis: design automation and task modeling techniques can provide adaptive support for situation-dependent task demands. Although computational design techniques, such as those outlined in Chapters 4 and 6, help tool creators improve visualization design, this iterative design process is still bottlenecked by revision and deployment speed. One interesting direction would be to test whether this knowledge could be integrated into intelligent, adaptive visualization applications, such that interfaces could change along with a user's environment. These algorithms could build on existing augmented reality platforms, where visualization appearance could dynamically updated as viewing conditions change. Another possibility would be to use measures of cognitive workload, as explored in Peck et al.'s brain-visualization interface studies [142, 143], to dynamically update visualizations.

7.3 Research Contributions

7.3.1 Overview of primary contributions

We provide a "Table of Contributions" in Table 7.1 to highlight each preceding chapter's contributions and to highlight how each chapter supports our thesis statement. These research contributions are founded on both qualitative (C1, C2, C9) and quantitative evaluation results (C4, C7, C11, C12). The remaining primary contributions focus on how our improved theoretical understanding of visualization design can be applied. In full, these primary contributions improve visualization design by examining and developing new techniques to analyze visual analysis tasks within cancer genomics research (C1–C6) and by evaluating graphical perception phenomena to discover effective visual design practices (C7–C13).

In Chapter 3, our contributions extend the theoretical understanding of visualization tool design challenges posed by the diverse research backgrounds and research goals that exist within cancer research. This advancement is a product of a design study and task requirement analysis evaluation that characterized how visualization tool design can better support cancer genomics researchers' exploratory visual analysis workflows. Using this new understanding we also discuss how the design of visualization tools might robustly support such diverse sets of requirements by analyzing direct observations of how several cancer genomics researchers interacted with MAGI.

Chapter 4 expands Chapter 3 by evaluating how task requirements can be analyzed using simple classifiers trained on annotated mouse interaction logs. Our findings validate that classification is a reliable methodology for task requirement analysis and likewise show that humans are also reliable at task inference from mouse interaction logs even without full analytical context. Importantly, the classifier evaluation used MAGI interactions collected over a year, which simultaneously enhances our knowledge of naturalistic cancer genomics visual analysis. Another key finding was that our newly proposed "mouse tracking" features based on eye tracking methodologies outperformed established

Chapter 3:	**A cancer genomics visualization task requirements analysis and design study of MAGI.**
C1	A MAGI design study that explains how MAGI supports cancer genomics visualization needs.
C2	A task requirements analysis that identifies cancer genomics researchers' visual analysis needs across a variety of specializations (e.g., pharmaceutical industry vs. basic science research).
C3	An exploratory analysis of MAGI interaction logs that suggest the design study findings generalize to ecological settings.
Chapter 4:	**Evaluating visual analysis task classification to improve understanding of cancer genomics domain expert use of MAGI.**
C4	An evaluation of twelve automated visual analysis task classification accuracies that found similar accuracy when compared to hand-coded task inferences made by pairs of genomics and visualization experts.
C5	An exploration of common MAGI interaction trends using the predictions from task classification, which supplements C3 to further our understanding of how visualization is used in naturalistic settings by domain experts.
C6	A discussion of how iterative visualization tool design can be improved by our mouse interaction log analysis contributions.
Chapter 5:	**The relation between visualization size, grouping, and user performance.**
C7	An evaluation of how human perception of grouping, quantity, and size affects visualization search performance (i.e., one measure of visualization design effectiveness).
C8	An analysis of how search performance relates to self-reported difficulty for different types of visualization.
C9	A discussion of a NASA TLX cognitive workload study that found size can modulate perceived visualization-task associations.
Chapter 6:	**Colorgorical: Creating discriminable and preferable color palettes for information visualization.**
C10	An automated visualization design technique that creates categorical color palettes based on user-defined balances of discriminability and preference.
C11	An analysis of how Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference color appearance functions systematically differ.
C12	A first evaluation of how varying user-defined balances of discriminability and preference affects palette creation and also human discrimination performance and preference ratings.
C13	A second evaluation that found Colorgorical palettes are as discriminable and typically more preferable compared to ColorBrewer, Microsoft, and Tableau color palettes.

Table 7.1: A table of primary contributions for all preceding chapters, which highlights the ways in which each chapter expands our present knowledge of information visualization.

interaction models adapted from previous work in information retrieval and visual analytics.

In Chapter 5, we established several design principles to promote visualization legibility while maintaining a high level of data density. These principles are founded on in-lab evaluations that tested how the physical size of visualizations and how perceptual grouping of rendered data can affect visualization search performance. To investigate how these low-level psychophysical findings might effect analysis on a broader scale, we designed a study using NASA's TLX methodology [66] to measure cognitive workload. Our findings show that size can affect the types of tasks that users might associate with visualization.

Last, in Chapter 6, we described a new technique for generating categorical color palettes based on user-defined balances of color discriminability and aesthetic preference. The goal of this technique was to reduce design barriers for visualization creators without design expertise who might struggle to select custom color palettes when coloring many categories (e.g., cancer types or genes). We found that this new tool and technique – Colorgorical – could make palettes that were as discriminable and often more preferable compared to current industry standards (Microsoft, ColorBrewer, and Tableau). As such, Colorgorical provides immediate benefits to visualization creators and expands our conceptual understanding of visualization color palette effectiveness. It also establishes that principles from vision science can be directly applied in computational tools in ways that reduce barriers for novices to create their own custom visual stylizations while also minimizing the risk of making poor visualization design decisions.

7.3.2 Supplementary thesis contributions

In addition to publication-focused contributions, we made a number of additional supporting contributions that helped encourage the practice of effective visualization design. Although these efforts did not necessarily expand human understanding of information visualization, they did improve practitioner access to the frontiers of visualization design research. These types of technology transfers are essential if visualization research discoveries are to improve tool development [18].

The first supplementary contribution is the creation of GD3, a genomics visualization library that reduces implementation barriers for genomics researchers to create complex and interactive genomics-specific charts. This library explored declarative visualization design patterns similar to those that were later published as part of the Vega visualization programming language research projects [164]. Related, we also aided in the development of MAGI (Chapter 2.3) by iteratively designing various interfaces and visualizations. Last, we released Colorgorical as an open-sourced tool, which has seen widespread use by practitioners since its publication.

7.3.3 Summative thesis contributions

The overarching contribution of this thesis work is support for our thesis statement. Each primary and secondary contribution above shows by example how research into visualization theory can inspire and inform visualization design assistance technique development. Specifically, we illustrate this relation between discovering new knowledge and developing novel techniques for task requirement analyses and graphical-perception-inspired visual design while working within a cancer genomics visualization application area. Our primary contributions expand the frontiers of human knowledge about visualization design research, and our secondary contributions help make these expansions accessible to visualization practitioners by providing open sourced implementations. By adopting a multidisciplinary and collaborative approach to research, the outcomes of this thesis provide new tools and knowledge for fellow toolsmiths to improve their users' visual analysis workflows.

Appendices

Appendix A

Supplementary Material, Colorgorical: Creating discriminable and preferable color palettes for information visualization

A.1 Overview

We present (1) additional, more thorough explanations of how each of Colorgorical's palette scores operate; (2) an analysis of how the scores are related to one another; (3) extended analysis of Experiments 1 and 2; and (4) example palettes made with 20 representative Colorgorical slider settings. We include supplementary figures and the tables presenting the statistics from our analyses.

This chapter is the supplementary material for an extended version of a work that originally appeared in the proceedings of InfoVis'16 as [**?**].

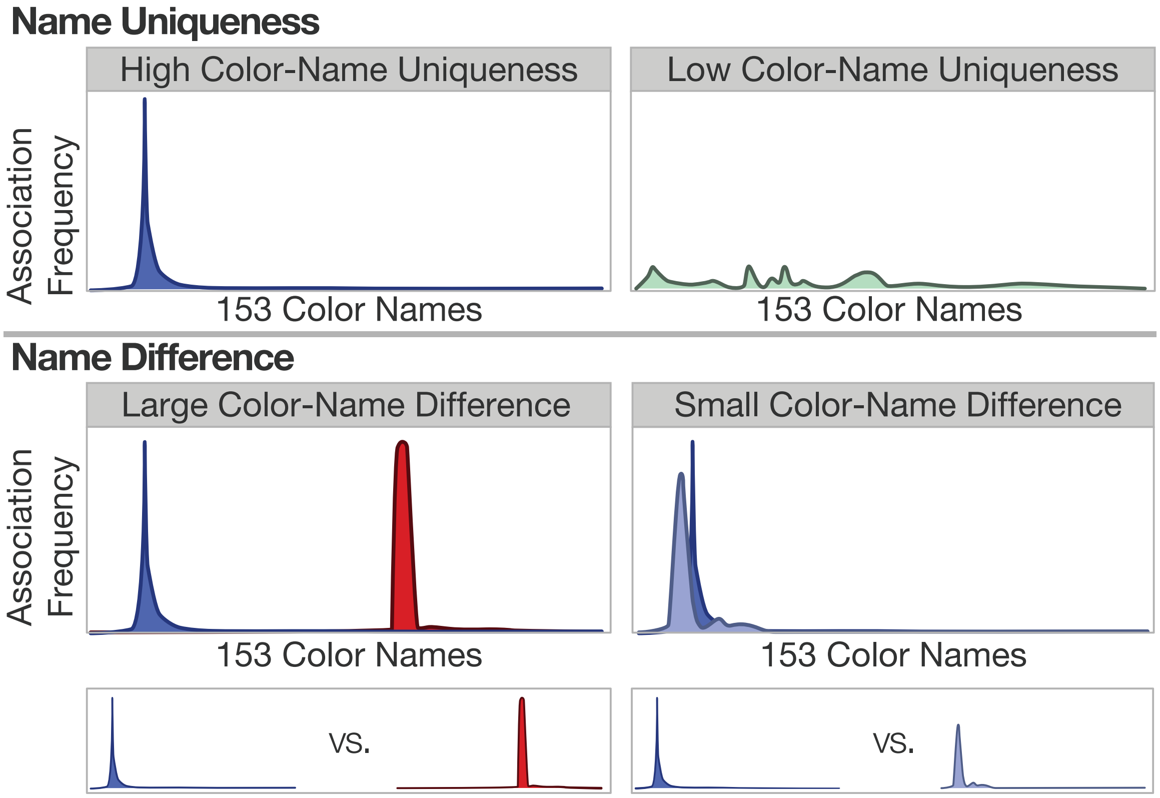


Figure A.1: Name Uniqueness and Name Difference. High Name Uniqueness scores are focally distributed color-name associations, whereas low Name Uniqueness scores are more uniform distributions. Name Difference scores are proportional to the difference between color-name association distributions. The blue and red example is large because there is little color-name association overlap.

A.2 Name Uniqueness and Difference score explanations

Both Name Uniqueness and Name Difference are color-term association statistics that were originally created by Heer and Stone [73]. The color-name associations map every color in a quantized 8,325-color CIELAB space to 153 popular color names, which was based on data from an XKCD crowdsourcing experiment. Name Uniqueness refers to their "name saliency" statistic, which we renamed to avoid confusion with color saliency.

Name Difference can be thought of as how much two colors' association mappings overlap, whereas Name Uniqueness can be thought of as how uniformly distributed a colors' associations are to the 153 names. Each scoring function is illustrated in Supplementary Figure A.1.

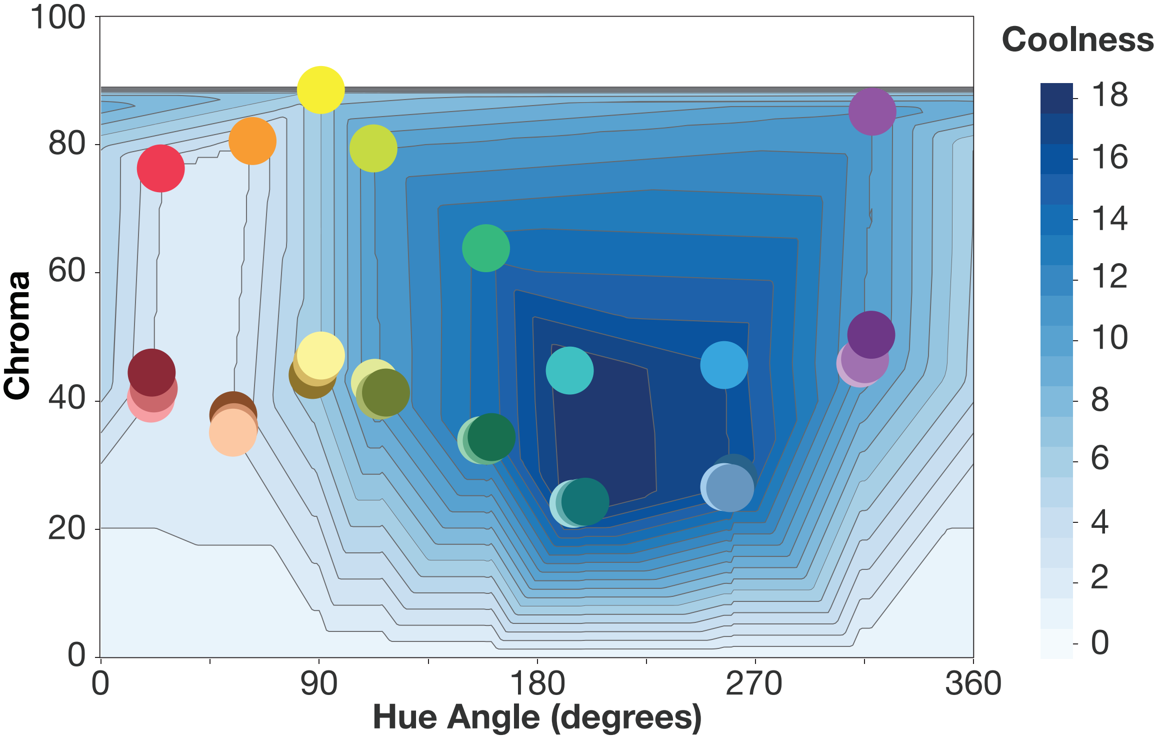


Figure A.2: Interpolated coolness layered on chroma and hue from CIE LCh. The original Schloss & Palmer coolness values are derived from how many steps each color used in the experiment used to derive Pair Preference is from the color 10R in Munsell color space.

A.3 Interpolating the Pair Preference *Coolness* term

In Supplementary Figure A.2 we show the linear interpolation results that calculate coolness values for CIE LCh space. The interpolated values are derived from the coolness values of the 32 Munsell colors used in the original Schloss and Palmer pair preference in-lab experiment [165]. These values were calculated by counting the number of steps each of the colors was from the color 10R in Munsell color space.

A.4 Runtime performance

To evaluate model runtime with respect to palette size, we profiled single-palette generation with 1 to 20 colors 50 times each on a Mid 2012 MacBook Pro Retina with a 2.6 GHz Intel Core i7 CPU and 16GB 1600MHz DDR3 RAM. Average initialization time was 0.14 seconds (SEM = 0.004). If a palette was returned before reaching the required number of colors, it was discarded and the test

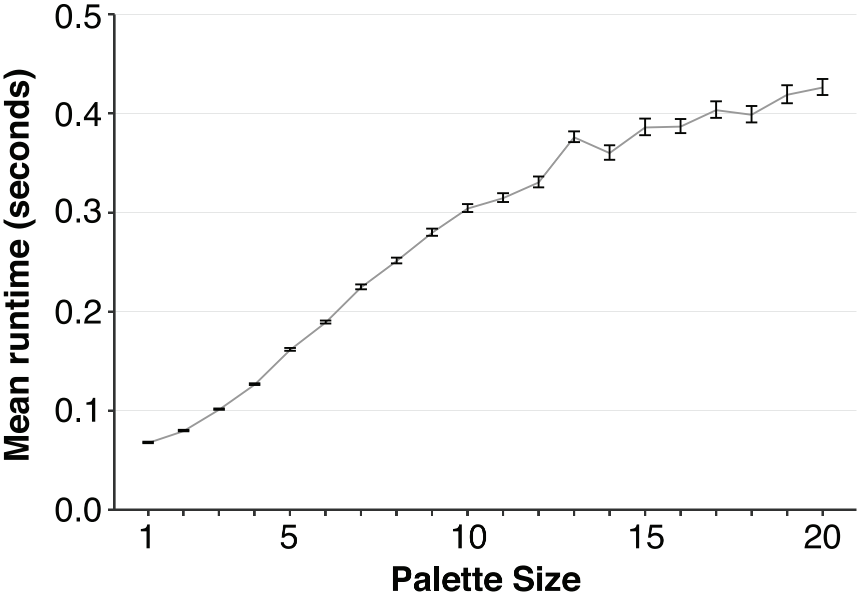


Figure A.3: Runtime performance of Colorgorical for 20 palette sizes. Error bars show standard error for each number of colors' 50 tests.

was run again. Runtime performance increased linearly in the number of colors (S.Fig. A.3).

A.5 Palette Score Evaluation

The aim of this experiment was to assess whether the model could be simplified by removing redundant scoring functions. For instance, if Perceptual Distance were to explain most of the variance in Name Difference scores, then the Name Difference scoring function could be removed from the model with little effect on palette output.

To examine how similar the four Colorgorical scoring functions were to one another, we tested the degree of independence between *palette scores*. Palette scores are derived by taking the minimum scoring function output given all color pairs in a palette. We use the minimum score based on our model's assumption that a palette is only as preferable or discriminable as its lowest pair.

We also tested (1) how the three discriminability palette scores compare to Pair Preference, and (2) how each of the four palette scores changes with the number of colors in a palette. We predicted that:

P1 All palette scores measure different color-relation information and are not redundant

P2 Pair Preference would be a negative predictor of the Perceptual Distance and Name Difference

A.5.1 Methods

To test the full range of Colorgorical output, we created a representative set of 39,600 palettes. This collection was made using 66 unique slider settings, which tested different relative importance of scoring functions, and palettes of 3, 5, and 8 colors. The 66 settings are the different unique combinations a user could make by dragging each slider to 0%, 50%, or 100%. The combinations ignore duplicate settings encountered when moving one or more sliders to 0%. For instance, if three sliders are turned to 0%, any non-0% position of the fourth slider would give it a relative importance of 100%.

A.5.2 Results & Discussion

Before testing the relation between each of the four palette scores, we first plotted the distribution for each of the palette scores across the different palette sizes (S.Fig. A.4). One noticeable trend is that the palette scores decrease as a whole with respect to palette size. To test whether this trend was significant we correlated each collection of palette scores with palette size ({3,5,8}-colors) and found that each trend was significant (Pearson's $r(39598) = -0.667$ (PD), -0.429 (ND), -0.267 (NU), -0.727 (PP); $p < 0.001$). These trends might originate from a combination of two sources. First, increasing the number of colors leaves successively fewer regions of available color pace to sample from. Second, using wider swaths of color space increases the likelihood that there is a low score in the exponentially growing number of color pairs in a palette. Using different aggregation techniques (e.g., leave-lowest-out or averaging) might result in higher palette scores, but would also go against our assumption that a palette is only as discriminable or preferable as the worst-performing pair of colors in a palette. However, we believe that testing this assumption would be an interesting direction for future research.

After, we tested whether Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference *palette scores* were independent (P1). To evaluate this prediction we used multiple linear regression analyses to predict palette scores as a function of the other three (e.g., predicting Perceptual Distance with Name Difference, Name Uniqueness, and Pair Preference). We conducted separate regressions for each palette size, resulting in 12 regressions (4 palette scores × 3 sizes)

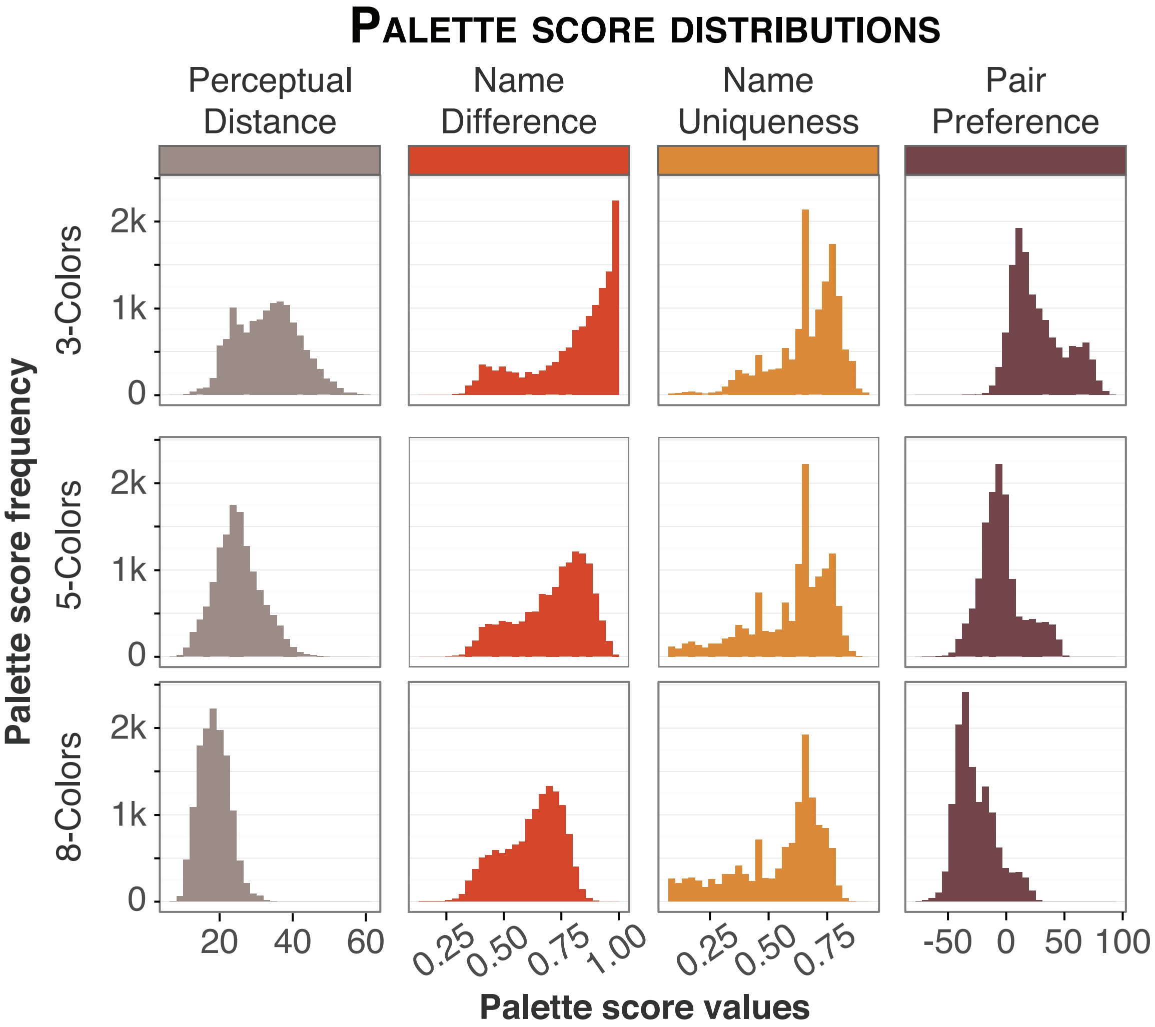


Figure A.4: Palette score distributions for each of the 39,600 palettes used in the palette score verification and Experiment 1.

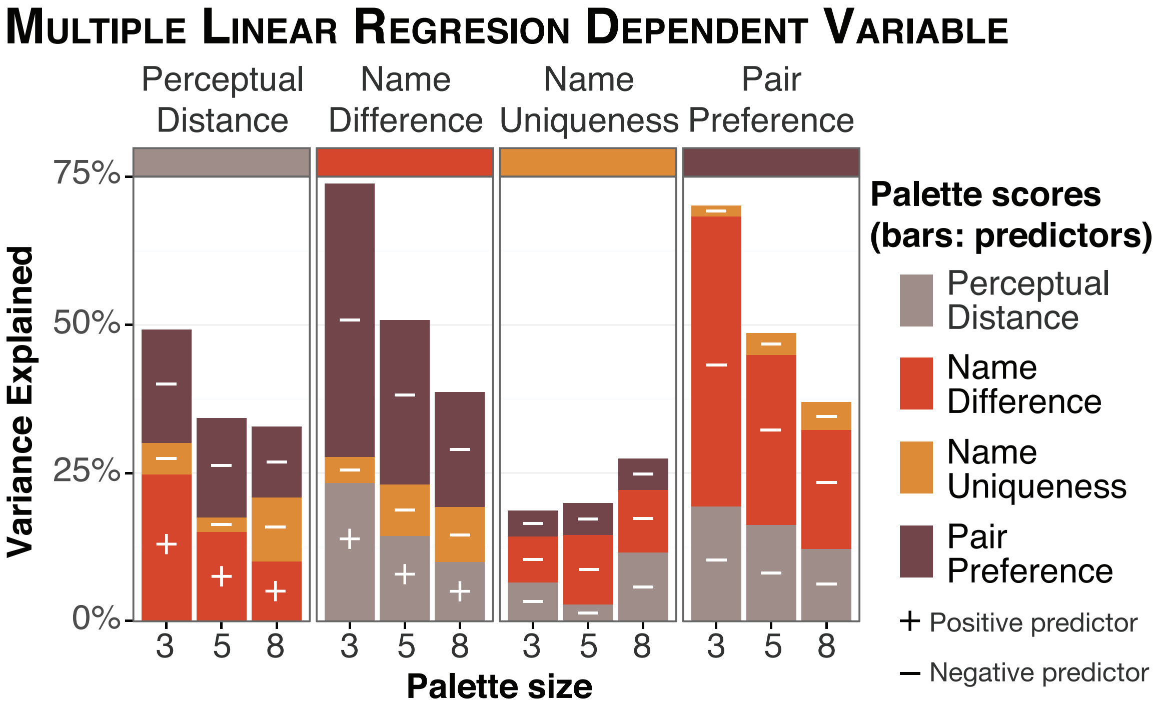


Figure A.5: Percent of explained variance of 12 linear regressions that test whether palette scores predict one another (stack height). Each bar shows the percent of explained variance for each palette score. All regressions and predictors were significant.

used to predict 13,200 palettes. For each regression, all predictors explained a significant amount of variance (all $F(3, 13196) \geqslant 1007.676$, all $p < 0.001$; all $t(13196) \geqslant 16.26$, all $p < 0.001$).

Supplementary Figure A.5 shows the relative importance of the predictors in each regression model [53]. Perceptual Distance and Name Difference showed similar trends in that both were positive predictors of one another, Name Uniqueness was a small negative predictor, and Pair Preference was a large negative predictor. The largest difference between Perceptual Distance and Name Difference was that Pair Preference explained a much larger portion of Name Difference's variance (and vice versa; 3-Color ND predicted by PP $= 46.1\%$; 3-Color PP predicted by ND $= 48.9\%$). This strong negative association could be linked to the hue similarity term in Pair Preference, which might sometimes create a discrimination-preference trade-off (P2). These findings suggest that Pair Preference is more strongly related to the Name Difference of colors than to Perceptual Distance.

We concluded from these results that each function measured sufficiently different color information (P1) because a single palette score never predicted greater than 50% of variance in another. Therefore, we kept all of the four scores in the model while generating the stimuli for Experiment 1.

A.6 Exp. 1: palette score correlation with behavior

A.6.1 Selecting palette score binning widths

A large problem when correlating palette scores with behavior is the individual differences that occur between subjects. Another problem is that near-scores are treated as separate values, leading to uninformative correlations. To avoid both problems, we quantized each palette score into 15 bins. Bin widths were calculated using the full range values over 3-, 5-, and 8-colors for each score. Our goal when selecting the number of bins to use was to maximize the number of subjects who were shown all bins. In other words, we wanted to avoid having bins with few subjects in them to improve the consistency of analyses.

We first attempted binning using the Freedman-Dianconis rule, which resulted in 29 bins:

$$\text{number of bins} = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil \tag{A.1}$$

$$h_{\text{F-D}} = 2\frac{\text{IQR}(x)}{n^{1/3}} \tag{A.2}$$

We also attempted binning using Sturge's formula, which resulted in 13 bins:

$$h_{\text{Sturge}} = \lceil \log_2 n + 1 \rceil \tag{A.3}$$

Our ultimate selection method relied on picking bins after charting different widths because the Freedman-Diaconis rule resulted in many bins with few subjects for some sizes, and Sturge's formula was too coarse of a score description. We tested 10 to 30 bins in increments of 5, reflecting the range between rounded Sturge and Freedman-Dianconis bin suggestions. The chart is shown in Supplementary Figure A.6.

A.6.2 Correlation results

Binned palette score correlation results are shown in Supplementary Figure A.8 (Pearson's r). An alternative to our binning approach is to correlate within-subject unbinned palette scores with behavioral measures, and then apply Fisher's Z transform to average Pearson's r between subjects[1].

[1] Fisher's Z transform converts correlation coefficient distributions to be more normal-like [125]

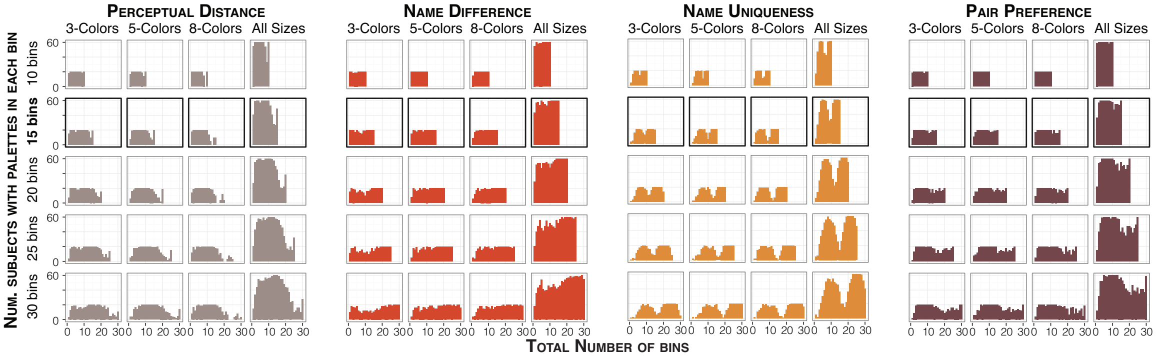


Figure A.6: Plots of different bin amounts for Experiment 1, where each bar represents the number of subjects who were shown palettes within a given bin. An ideal setting first minimizes the number of low-subject bins, and then favors a larger number of bins to better describe the distribution.

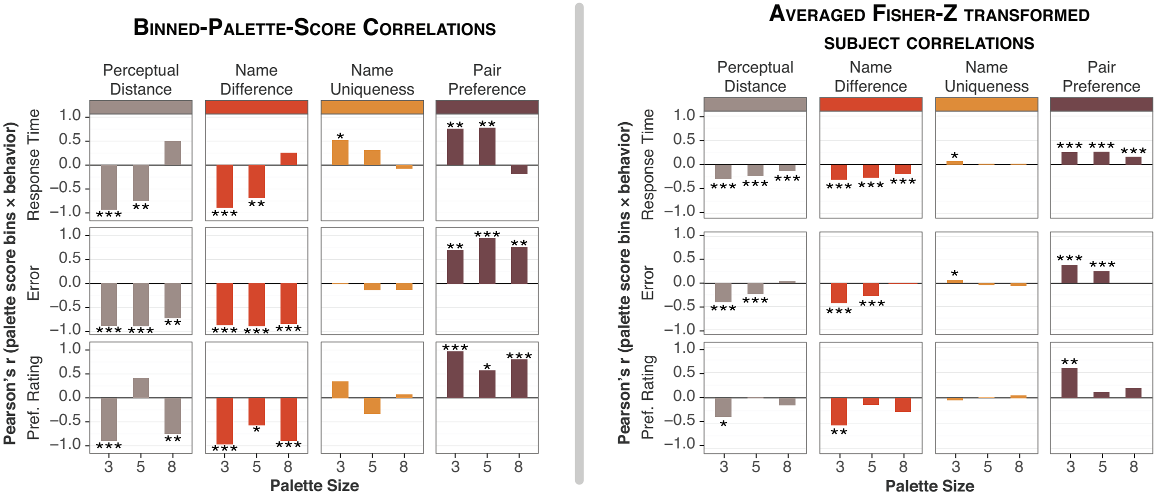


Figure A.7: Side-by-side comparison of correlation results. The left chart is reproduced from the primary manuscript and shows the correlations for each palette score × size condition using 15 bins. The right chart shows the same conditions, but the correlations are the average correlations between subjects using Fisher's Z transform.

A side-by-side comparison of the correlation results for each method is shown in Supplementary Figure A.7. To test for significance in the individual correlations, we conducted one-sample t-tests for each palette score within each size to compare the mean of the individual subjects' correlations against zero. The magnitude of Pearson's r was smaller for the mean of the individual correlations (S.Fig. A.7, left) compared to the correlations across means between the mean data (S.Fig. A.7, right). The difference in magnitude is expected, given that averaging between subjects reduces the noisy variance stemming from individual differences. Nonetheless, the pattern of results is similar: 8 bars (of 36) are significant with binning that are not with Fisher's transform, and 4 bars are significant with Fisher's transform that are not with binning (12 of 36 total). Both methods show few significant Name Uniqueness correlations despite the number of false negatives and false positives. Therefore, both correlation methods support our decision to remove Name Uniqueness from further analysis.

A.6.3 Slider settings' mapping onto behavior

To capture how manipulating sliders' relative importances mapped onto the behavioral data captured in Experiment 1, we created a set of Barycentric plots (S.Fig. A.9). Each facet of the plot shows a different size × behavioral measure condition, and each circle is one of the 20 tested slider settings. The triangle fill colors are the Barycentric interpolated values between each tested setting and a thicker stroke indicates the Experiment 2 Low-Error and Preferable palette settings. Of note, Perceptual Distance was more amenable to preserving preference ratings when increased. The neutral preference rating trend across 5-color slider settings discussed in the primary manuscript is also reflected, as is the 8-color response time drop off compared to 3- and 5-color response times.

A.7 Exp. 2: Supplementary Material

The full list of palette set means and standard errors for the Experiment 2 palettes are listed in Table A.1.

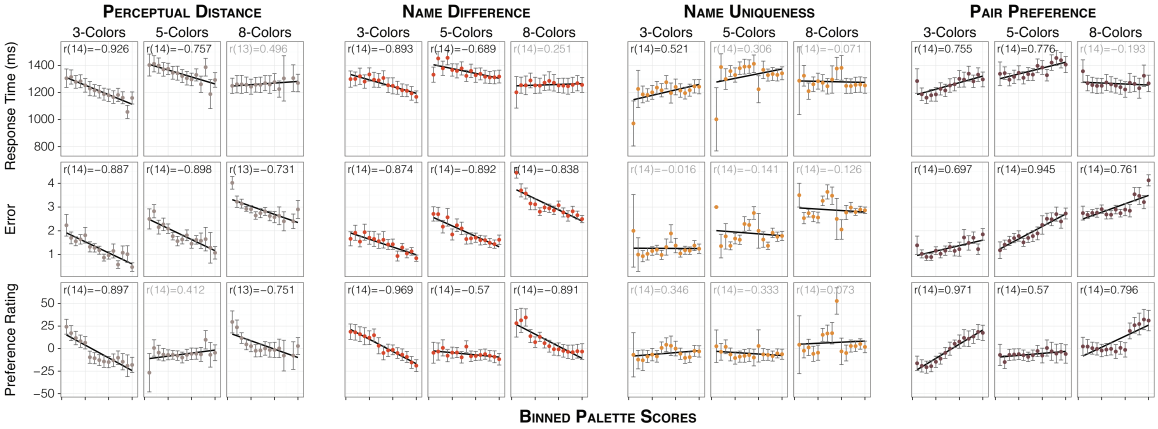


Figure A.8: Results from Exp.1 correlations between palette scores and behavioral measures for 3-, 5-, and 8-color palettes. To compare similar palette scores, each score was quantized into 15 bins. Error bars show standard error within each bin. Pearson's correlation coefficients are shown in black ($p < 0.05$) or grayed out ($p \geqslant 0.05$) text. Lines show a linear regression fit between palette score and behavior.

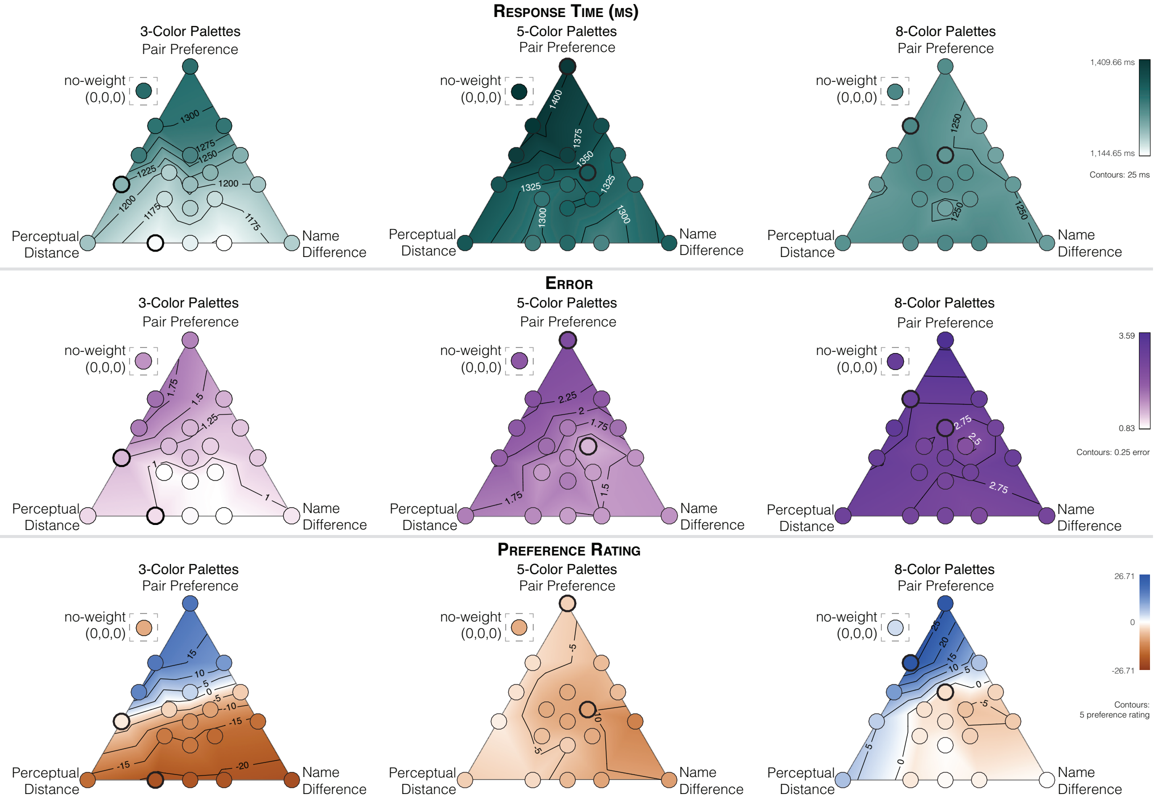


Figure A.9: Results from Experiment 1 investigating how slider settings for each scoring function (i.e., relative importance) change discrimination performance and preference ratings. Colored circles represent the settings tested. Error refers to the total number of errors made with a particular slider setting, where each of the original 66 slider settings had 10 repetitions within each subject (i.e., 1 error = 10%).

Size	Palette Set	Measure	Mean	Std. Error
3	ColorBrewer	Error	1.1	0.27
3	Low-Error	Error	1.05	0.303
3	Preferable	Error	2.2	0.427
3	Microsoft	Error	1.7	0.341
3	Random	Error	1.25	0.307
3	Tableau	Error	1	0.192
5	ColorBrewer	Error	3	0.465
5	Low-Error	Error	3.25	0.532
5	Preferable	Error	3.35	0.319
5	Microsoft	Error	4.3	0.459
5	Random	Error	2.95	0.51
5	Tableau	Error	3.3	0.493
8	ColorBrewer	Error	5.8	0.421
8	Low-Error	Error	5.45	0.359
8	Preferable	Error	4.3	0.363
8	Microsoft	Error	6.25	0.376
8	Random	Error	4.6	0.505
8	Tableau	Error	5.7	0.548
3	ColorBrewer	RT	1474.66	78.458
3	Low-Error	RT	1479.618	79.75
3	Preferable	RT	1542.629	87.427
3	Microsoft	RT	1568.527	94.217
3	Random	RT	1454.229	79.477
3	Tableau	RT	1453.156	81.904
5	ColorBrewer	RT	1824.114	81.677
5	Low-Error	RT	1875.154	83.546
5	Preferable	RT	1858.203	79.526
5	Microsoft	RT	1955.883	88.54
5	Random	RT	1837.006	70.413
5	Tableau	RT	1865.575	75.587
8	ColorBrewer	RT	1769.46	78.733
8	Low-Error	RT	1780.639	73.004
8	Preferable	RT	1772.981	68.538
8	Microsoft	RT	1755.502	96.269
8	Random	RT	1719.034	77.138
8	Tableau	RT	1770.159	87.515
3	ColorBrewer	Pref. Rating	-19.744	6.249
3	Low-Error	Pref. Rating	-11.369	8.321
3	Preferable	Pref. Rating	32.644	8.56
3	Microsoft	Pref. Rating	-3.806	6.352
3	Random	Pref. Rating	-16.663	6.724
3	Tableau	Pref. Rating	-2.444	5.897
5	ColorBrewer	Pref. Rating	-22.837	8.374
5	Low-Error	Pref. Rating	5.787	6.845
5	Preferable	Pref. Rating	8.55	8.364
5	Microsoft	Pref. Rating	9.312	8.181
5	Random	Pref. Rating	-7.1	7.205
5	Tableau	Pref. Rating	-6.45	5.692
8	ColorBrewer	Pref. Rating	-16.788	8.818
8	Low-Error	Pref. Rating	3.25	8.119
8	Preferable	Pref. Rating	24.038	7.715
8	Microsoft	Pref. Rating	7.312	6.776
8	Random	Pref. Rating	-13.375	9.262
8	Tableau	Pref. Rating	-0.012	7.65

Table A.1: The mean and standard error responses for each palette set × size combination. Low-Error and Preferable palette sets are the two Colorgorical settings included in Experiment 2.

Size	Measure	$t(19)$	p
3	Pref. Rating	-3.573	0.002
5	Pref. Rating	-0.405	0.69
8	Pref. Rating	-3.79	0.001
3	Error	-3.286	0.004
5	Error	-0.195	0.847
8	Error	2.113	0.048
3	RT	-1.513	0.147
5	RT	0.309	0.761
8	RT	0.221	0.828

Table A.2: Experiment 2 t-tests between Colorgorical Low-Error and Preferable settings. Negative t-values favor Preferable.

A.8 Predictive comparison: Colorgorical vs. Tableau v.10 palettes

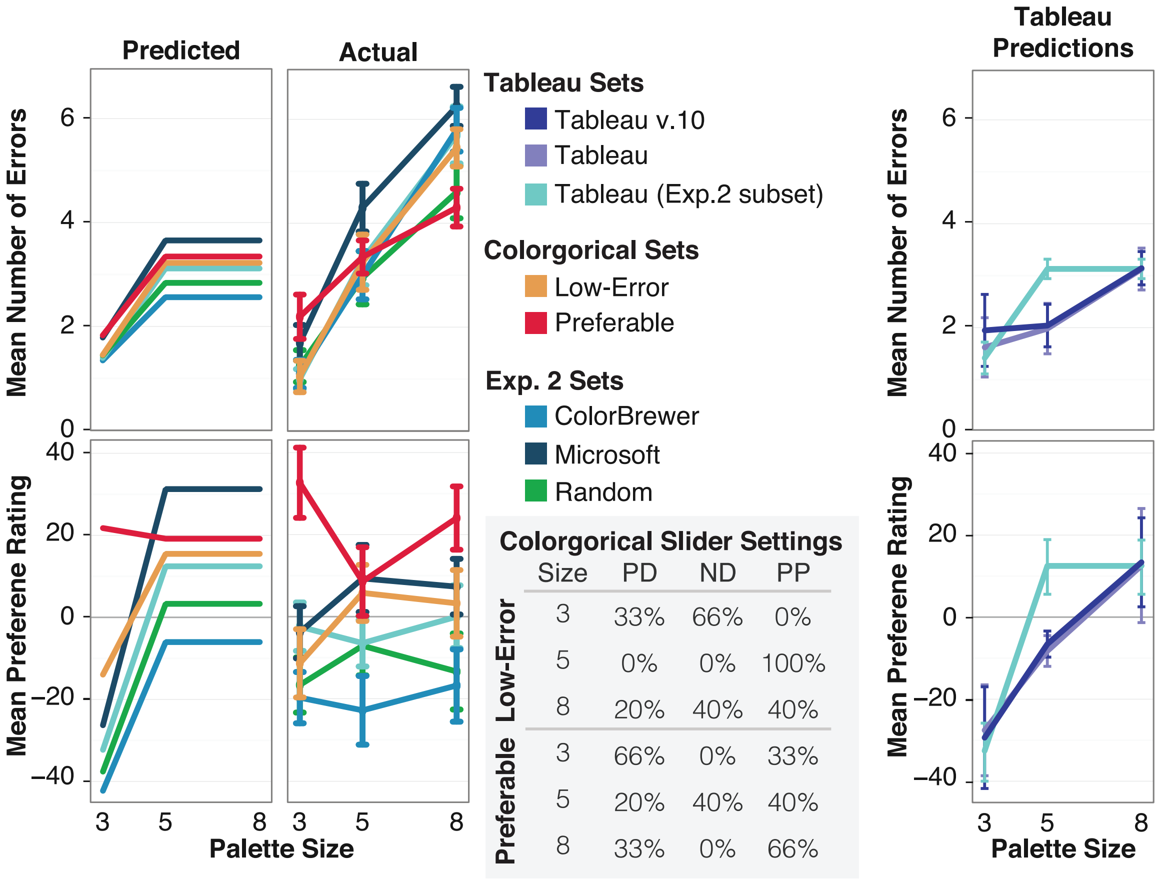
A year after running our Colorgorical evaluation, Tableau released an entirely redesigned collection of palettes. To predict how these palettes might perform compared to Colorgorical, we applied our previously described palette score linear regression models to these new palettes. We show results in Figure A.10 alongside Experiment 2 data. Recall that in Experiment 2 we tested only a subset of Tableau color palettes. As such, the predictive modeling results in Figure A.10 show predictions for the tested subset, the entirety of old Tableau color palettes, and the entirety of new Tableau color palettes. These results suggest that the new Tableau color palettes are predicted to, overall, have similar error rates and preference ratings as the old Tableau color palettes.

A.9 Linear regression and t-test analysis results

The tables for Palette Verification and Experiments 1 and 2 linear regressions and t-tests are shown in the tables below. All t-tests were paired and two-tailed.

A.10 Colorgorical output examples

Below are examples of the 20 slider settings that can be made by changing Perceptual Distance (PD), Name Difference (ND), and Pair Preference (PP) sliders to 0%, 50%, and 100%. We left Name Uniqueness at 0% given Experiment 1 results. Note that color appearance is slightly off,



Colorgorical Slider Settings

	Size	PD	ND	PP
Low-Error	3	33%	66%	0%
	5	0%	0%	100%
	8	20%	40%	40%
Preferable	3	66%	0%	33%
	5	20%	40%	40%
	8	33%	0%	66%

Figure A.10: A comparison of predicted error rate and preference ratings based on the new Tableau color palette scores. The two left columns are the previously reported Experiment 2 results. The right-most column are predicted results based on palette score. These predictions were generated with the same linear regressions that we previously described in Experiment 2. Error bars show standard error.

Measure	Colors	Setting	Benchmark	$t(19)$	p
Error	3	Low-Error	ColorBrewer	-0.139	0.891
Error	3	Low-Error	Microsoft	-1.628	0.12
Error	3	Low-Error	Tableau	0.181	0.858
Error	3	Low-Error	Random	-0.525	0.606
Error	5	Low-Error	ColorBrewer	0.665	0.514
Error	5	Low-Error	Microsoft	-2.396	0.027
Error	5	Low-Error	Tableau	-0.103	0.919
Error	5	Low-Error	Random	0.603	0.554
Error	8	Low-Error	ColorBrewer	-0.649	0.524
Error	8	Low-Error	Microsoft	-1.417	0.173
Error	8	Low-Error	Tableau	-0.366	0.719
Error	8	Low-Error	Random	1.342	0.196
Error	3	Preferable	ColorBrewer	2.531	0.02
Error	3	Preferable	Microsoft	1.097	0.287
Error	3	Preferable	Tableau	3.644	0.002
Error	3	Preferable	Random	3.047	0.007
Error	5	Preferable	ColorBrewer	0.78	0.445
Error	5	Preferable	Microsoft	-1.727	0.1
Error	5	Preferable	Tableau	0.101	0.921
Error	5	Preferable	Random	0.867	0.397
Error	8	Preferable	ColorBrewer	-2.91	0.009
Error	8	Preferable	Microsoft	-3.456	0.003
Error	8	Preferable	Tableau	-1.898	0.073
Error	8	Preferable	Random	-0.501	0.622
RT	3	Low-Error	ColorBrewer	0.163	0.872
RT	3	Low-Error	Microsoft	-2.935	0.008
RT	3	Low-Error	Tableau	1.004	0.328
RT	3	Low-Error	Random	0.825	0.42
RT	5	Low-Error	ColorBrewer	1.311	0.205
RT	5	Low-Error	Microsoft	-2.147	0.045
RT	5	Low-Error	Tableau	0.27	0.79
RT	5	Low-Error	Random	0.594	0.559
RT	8	Low-Error	ColorBrewer	0.309	0.761
RT	8	Low-Error	Microsoft	0.457	0.653
RT	8	Low-Error	Tableau	0.223	0.826
RT	8	Low-Error	Random	1.639	0.118
RT	3	Preferable	ColorBrewer	2.008	0.059
RT	3	Preferable	Microsoft	-0.536	0.598
RT	3	Preferable	Tableau	2.171	0.043
RT	3	Preferable	Random	2.009	0.059
RT	5	Preferable	ColorBrewer	0.779	0.446
RT	5	Preferable	Microsoft	-1.804	0.087
RT	5	Preferable	Tableau	-0.156	0.878
RT	5	Preferable	Random	0.312	0.758
RT	8	Preferable	ColorBrewer	0.105	0.918
RT	8	Preferable	Microsoft	0.349	0.731
RT	8	Preferable	Tableau	0.063	0.95
RT	8	Preferable	Random	1.761	0.094
Pref. Rating	3	Low-Error	ColorBrewer	1.075	0.296
Pref. Rating	3	Low-Error	Microsoft	-0.697	0.494
Pref. Rating	3	Low-Error	Tableau	-1.193	0.248
Pref. Rating	3	Low-Error	Random	0.629	0.537
Pref. Rating	5	Low-Error	ColorBrewer	2.784	0.012
Pref. Rating	5	Low-Error	Microsoft	-0.385	0.705
Pref. Rating	5	Low-Error	Tableau	1.687	0.108
Pref. Rating	5	Low-Error	Random	1.781	0.091
Pref. Rating	8	Low-Error	ColorBrewer	1.768	0.093
Pref. Rating	8	Low-Error	Microsoft	-0.372	0.714
Pref. Rating	8	Low-Error	Tableau	0.301	0.767
Pref. Rating	8	Low-Error	Random	2.279	0.034
Pref. Rating	3	Preferable	ColorBrewer	4.439	< 0.001
Pref. Rating	3	Preferable	Microsoft	3.375	0.003
Pref. Rating	3	Preferable	Tableau	3.252	0.004
Pref. Rating	3	Preferable	Random	4.416	< 0.001
Pref. Rating	5	Preferable	ColorBrewer	3.434	0.003
Pref. Rating	5	Preferable	Microsoft	-0.064	0.95
Pref. Rating	5	Preferable	Tableau	2.105	0.049
Pref. Rating	5	Preferable	Random	1.821	0.084
Pref. Rating	8	Preferable	ColorBrewer	3.072	0.006
Pref. Rating	8	Preferable	Microsoft	1.581	0.13
Pref. Rating	8	Preferable	Tableau	2.04	0.056
Pref. Rating	8	Preferable	Random	5.2	< 0.001

Table A.3: Experiment 2 *t*-tests between Colorgorical and benchmarks.

Measure	Colors	R^2	$F(3, 13196)$	p	β_{PD}	t_{PD}	p_{PD}	β_{ND}	t_{ND}	p_{ND}	β_{NU}	t_{NU}	p_{NU}	β_{PP}	t_{PP}	p_{PP}	$\mathrm{RI}_{\mathrm{PD}}$	$\mathrm{RI}_{\mathrm{ND}}$	$\mathrm{RI}_{\mathrm{NU}}$	$\mathrm{RI}_{\mathrm{PP}}$
PD	3	0.492	4262.806	*	NA	NA	NA	0.16	36.358	*	-0.083	-24.822	*	-0.165	-23.019	*	NA	0.248	0.052	0.191
ND	3	0.738	12411.646	*	0.571	36.358	*	NA	NA	NA	-0.215	-34.622	*	-1.138	-117.653	*	0.233	NA	0.043	0.461
NU	3	0.186	1007.676	*	-0.537	-24.822	*	-0.388	-34.622	*	NA	NA	NA	-0.646	-36.408	*	0.065	0.076	NA	0.043
PP	3	0.701	10310.915	*	-0.234	-23.019	*	-0.45	-117.653	*	-0.141	-36.408	*	NA	NA	NA	0.193	0.489	0.018	NA
PD	5	0.343	2295.15	*	NA	NA	NA	0.087	26.478	*	-0.038	-16.259	*	-0.206	-38.652	*	NA	0.15	0.024	0.167
ND	5	0.508	4537.629	*	0.579	26.478	*	NA	NA	NA	-0.263	-47.172	*	-0.884	-72.06	*	0.143	NA	0.086	0.277
NU	5	0.2	1096.931	*	-0.521	-16.259	*	-0.549	-47.172	*	NA	NA	NA	-0.808	-41.001	*	0.028	0.117	NA	0.053
PP	5	0.486	4155.721	*	-0.494	-38.652	*	-0.319	-72.06	*	-0.14	-41.001	*	NA	NA	NA	0.162	0.286	0.036	NA
PD	8	0.328	2148.682	*	NA	NA	NA	0.042	17.753	*	-0.058	-41.791	*	-0.14	-40.325	*	NA	0.1	0.108	0.119
ND	8	0.387	2775.101	*	0.561	17.753	*	NA	NA	NA	-0.211	-41.174	*	-0.701	-58.12	*	0.099	NA	0.093	0.193
NU	8	0.274	1662.23	*	-2.011	-41.791	*	-0.54	-41.174	*	NA	NA	NA	-0.866	-42.72	*	0.115	0.105	NA	0.052
PP	8	0.369	2577.387	*	-0.785	-40.325	*	-0.291	-58.12	*	-0.14	-42.72	*	NA	NA	NA	0.121	0.201	0.046	NA

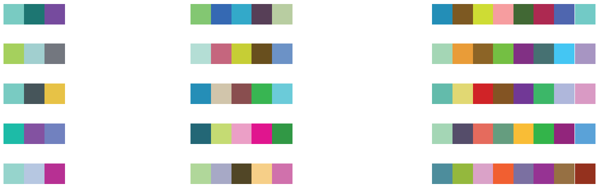
Table A.4: Linear regression tables for the 12 Palette Verification regressions that predicted palette scores in terms of the other three. Relative importance is calculated with `lmg` in the "relaimpo" R package [53]. $* : p < 0.001$

Measure	Colors	R^2	$F(3,16)$	p	β_{PD}	t_{PD}	p_{PD}	β_{ND}	t_{ND}	p_{ND}	β_{PP}	t_{PP}	p_{PP}	RI_{PD}	RI_{ND}	RI_{PP}
RT	3	0.815	23.442	< 0.001	-137.624	-4.485	< 0.001	-165.412	-5.39	< 0.001	17.627	0.574	0.574	0.211	0.357	0.246
RT	5	0.682	11.447	< 0.001	-81.807	-2.84	0.012	-126.571	-4.394	< 0.001	-2.646	-0.092	0.928	0.113	0.411	0.158
RT	8	0.192	1.267	0.319	-5.719	-0.411	0.687	-22.447	-1.613	0.126	-18.198	-1.308	0.209	0.029	0.108	0.054
Error	3	0.712	13.186	< 0.001	-0.488	-2.251	0.039	-0.85	-3.921	0.001	0.227	1.049	0.31	0.09	0.375	0.248
Error	5	0.692	11.964	< 0.001	-0.486	-2.001	0.063	-0.934	-3.85	0.001	0.224	0.924	0.369	0.073	0.39	0.228
Error	8	0.537	6.192	0.005	-0.267	-1.157	0.264	-0.568	-2.466	0.025	0.241	1.048	0.31	0.049	0.277	0.211
Pref. Rating	3	0.868	35.089	< 0.001	-6.754	-1.068	0.301	-20.665	-3.268	0.005	32.271	5.103	< 0.001	0.067	0.275	0.526
Pref. Rating	5	0.581	7.396	0.003	5.527	2.208	0.042	-4.013	-1.603	0.128	2.959	1.182	0.254	0.235	0.282	0.065
Pref. Rating	8	0.495	5.228	0.01	-0.876	-0.103	0.919	-15.615	-1.837	0.085	12.843	1.511	0.15	0.021	0.254	0.22

Table A.5: Linear regression tables for Experiment 1 that predicted participants' Response Time, Error, and Preference Rating as a function of Perceptual Distance (PD), Name Difference (ND), and Pair Preference (PP) slider settings. Relative importance is calculated with `lmg` in the "relaimpo" R package [53].

given that Colorgorical designs RGB palettes (figures in both this document and in the primary manuscript are rendered in CYMK). As such, we include palette color D65 CIELAB coordinates to the right of all palettes.

Slider settings:: PD:0.0 ND:0.0 NU:0.0 PP:0.0

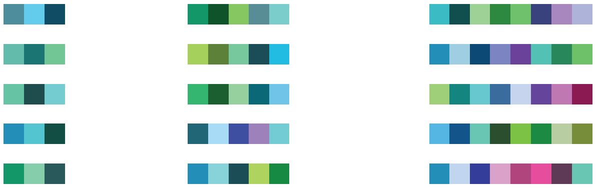


[85,-45,0]; [40,40,-45]; [45,-25,-5]
[85,-45,60]; [80,-15,-5]; [50,0,-5]
[85,-45,0]; [35,-5,-5]; [80,0,65]
[70,-45,0]; [45,60,-60]; [55,10,-35]
[85,-30,0]; [80,0,-15]; [45,70,-30]

[85,-60,45]; [45,15,-55]; [65,-20,-25]; [30,15,-10]; [80,-15,20]
[85,-15,0]; [55,40,5]; [80,-20,70]; [35,5,35]; [60,0,-30]
[55,-15,-30]; [80,0,15]; [40,25,10]; [65,-55,40]; [85,-40,-15]
[40,-15,-15]; [85,-25,50]; [75,40,-15]; [50,80,-15]; [55,-50,50]
[85,-30,30]; [70,5,-15]; [30,0,20]; [85,5,40]; [60,45,-15]

[55,-15,-30]; [40,10,45]; [85,-25,75]; [75,35,15]; [40,-25,25]; [40,55,10]; [45,15,-45]; [80,-35,-5]
[85,-30,15]; [70,20,60]; [45,10,50]; [70,-45,60]; [35,45,-30]; [45,-15,-5]; [75,-20,-35]; [65,15,-20]
[70,-30,0]; [85,-10,50]; [45,65,50]; [40,15,35]; [35,45,-45]; [70,-60,35]; [75,5,-20]; [75,55,-35]
[85,-30,15]; [35,10,-15]; [60,45,30]; [60,-25,10]; [80,10,70]; [65,-60,50]; [35,55,-25]; [65,-5,-40]
[55,-15,-15]; [70,-30,60]; [75,40,-25]; [60,55,55]; [50,15,-25]; [40,55,-35]; [50,10,30]; [35,40,40]

Slider settings:: PD:0.0 ND:0.0 NU:0.0 PP:1.0

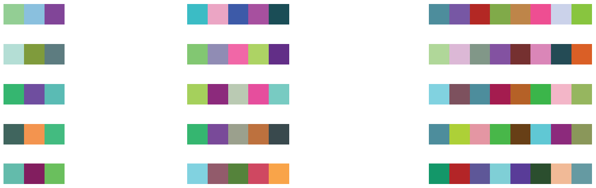


[55,-15,-15]; [80,-25,-25]; [30,-10,-20]
[70,-30,0]; [45,-25,-5]; [85,-60,25]
[85,-60,15]; [30,-15,-5]; [80,-30,-10]
[55,-15,-30]; [80,-40,-15]; [30,-20,0]
[55,-45,15]; [85,-45,15]; [35,-15,-5]

[55,-45,15]; [30,-35,25]; [80,-50,50]; [55,-15,-10]; [85,-40,-5]
[85,-45,60]; [50,-25,35]; [75,-35,15]; [30,-15,-10]; [70,-25,-30]
[70,-60,30]; [35,-35,30]; [80,-30,20]; [40,-20,-15]; [75,-15,-25]
[40,-15,-15]; [85,-10,-20]; [35,30,-60]; [60,30,-40]; [85,-40,-10]
[55,-15,-30]; [85,-30,-10]; [30,-15,-10]; [85,-40,60]; [50,-50,40]

[70,-30,-15]; [30,-20,-5]; [85,-40,30]; [50,-45,45]; [85,-70,45]; [30,15,-35]; [65,40,-45]; [75,10,-30]
[55,-15,-30]; [80,-10,-15]; [30,0,-30]; [60,20,-50]; [35,55,-60]; [80,-50,0]; [50,-35,15]; [80,-60,45]
[85,-45,45]; [50,-30,-5]; [85,-45,-10]; [45,0,-30]; [85,0,-15]; [35,50,-60]; [60,40,-25]; [30,55,-5]
[70,-15,-30]; [35,0,-35]; [80,-45,5]; [30,-20,15]; [75,-50,60]; [50,-50,45]; [80,-15,20]; [55,-20,40]
[55,-15,-30]; [85,0,-20]; [30,30,-55]; [75,40,-25]; [45,50,-10]; [60,80,-20]; [30,20,-10]; [80,-45,5]

Slider settings:: PD:0.0 ND:1.0 NU:0.0 PP:0.0

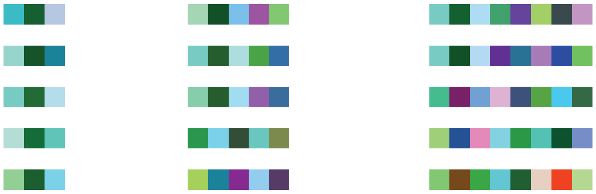


[85,-45,30]; [75,-10,-20]; [40,40,-35]
[85,-15,0]; [60,-25,50]; [50,-10,-5]
[70,-60,30]; [40,50,-60]; [70,-30,-5]
[40,-15,0]; [70,30,50]; [75,-60,25]
[70,-30,0]; [30,50,-15]; [80,-65,50]

[70,-30,-15]; [75,30,-5]; [40,15,-45]; [55,85,-60]; [30,-15,-10]
[85,-60,45]; [60,10,-20]; [65,65,-10]; [80,-30,50]; [30,40,-40]
[85,-45,60]; [35,50,-25]; [80,-10,10]; [60,80,-20]; [85,-45,0]
[70,-60,30]; [40,35,-35]; [65,-5,10]; [55,25,40]; [30,-5,-5]
[85,-30,-15]; [45,25,0]; [50,-30,40]; [50,55,15]; [75,25,60]

[55,-15,-15]; [45,50,-60]; [40,55,50]; [65,-30,45]; [60,15,40]; [60,70,0]; [85,5,-20]; [85,-65,80]
[85,-30,30]; [80,25,-20]; [60,-10,5]; [45,50,-50]; [30,30,15]; [70,60,-30]; [30,-10,-10]; [55,45,55]
[85,-30,-15]; [40,20,0]; [55,-15,-15]; [35,60,5]; [50,30,55]; [65,-60,65]; [80,25,0]; [70,-25,40]
[55,-15,-15]; [80,-35,75]; [70,30,5]; [70,-65,60]; [30,15,40]; [80,-35,-15]; [35,50,-25]; [60,-15,30]
[55,-45,15]; [40,55,35]; [40,20,-35]; [85,-35,-10]; [30,55,-65]; [30,-20,15]; [80,15,25]; [60,-15,-10]

Slider settings:: PD:0.0 ND:1.0 NU:0.0 PP:1.0

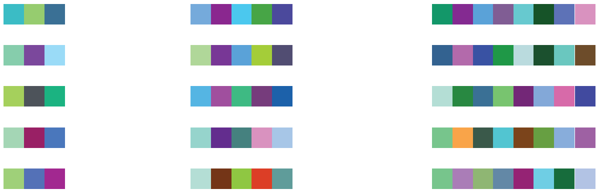


[70,-30,-15]; [35,-35,25]; [80,0,-15]
[85,-30,0]; [30,-35,30]; [50,-20,-20]
[85,-45,0]; [40,-35,30]; [85,-10,-10]
[85,-15,0]; [40,-40,30]; [80,-45,0]
[85,-45,30]; [35,-35,30]; [85,-30,-20]

[85,-30,15]; [30,-30,20]; [75,-10,-30]; [50,60,-45]; [85,-60,45]
[85,-45,0]; [35,-30,25]; [85,-15,-5]; [60,-45,40]; [45,0,-35]
[85,-45,15]; [35,-30,25]; [85,-15,-15]; [50,40,-40]; [45,0,-30]
[55,-45,30]; [80,-20,-20]; [30,-15,10]; [85,-50,0]; [55,-15,30]
[85,-45,60]; [50,-20,-20]; [35,60,-45]; [80,-10,-25]; [30,20,-20]

[85,-45,0]; [35,-40,35]; [85,-10,-15]; [60,-40,20]; [35,50,-60]; [85,-45,55]; [30,-5,-5]; [70,40,-35]
[85,-45,0]; [30,-30,20]; [85,-5,-20]; [30,50,-55]; [45,-10,-25]; [60,40,-40]; [35,20,-50]; [85,-70,50]
[70,-45,15]; [30,45,-20]; [65,0,-35]; [80,35,-25]; [35,5,-25]; [60,-45,60]; [80,-30,-30]; [40,-25,15]
[85,-45,45]; [35,10,-40]; [70,50,-20]; [80,-20,-15]; [55,-50,45]; [80,-50,0]; [30,-30,15]; [60,10,-40]
[85,-60,45]; [35,15,35]; [60,-50,40]; [75,-25,-15]; [35,-30,20]; [85,5,10]; [55,65,60]; [85,-30,35]

Slider settings:: PD:0.0 ND:1.0 NU:0.0 PP:0.5

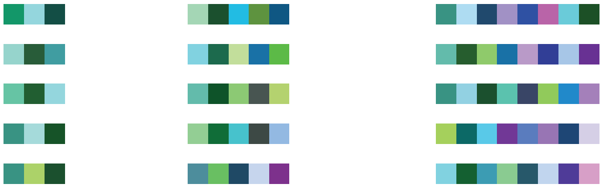


[70,-30,-15]; [80,-40,45]; [45,-5,-25]
[85,-45,15]; [40,55,-55]; [85,-15,-20]
[85,-45,60]; [35,0,-5]; [65,-50,15]
[85,-30,15]; [35,55,-10]; [55,20,-70]
[85,-45,45]; [50,15,-50]; [40,60,-30]

[70,0,-45]; [35,65,-45]; [80,-30,-30]; [60,-50,55]; [35,25,-45]
[85,-30,30]; [35,55,-50]; [65,-5,-40]; [80,-40,75]; [35,10,-20]
[70,-15,-30]; [50,70,-50]; [70,-50,20]; [35,35,-25]; [40,5,-45]
[85,-30,0]; [30,45,-45]; [50,-20,-5]; [70,40,-20]; [80,0,-30]
[85,-15,0]; [30,25,30]; [75,-40,60]; [50,60,60]; [60,-20,-5]

[55,-45,15]; [35,60,-45]; [65,-5,-40]; [45,25,-25]; [85,-45,-10]; [30,-35,30]; [55,30,-70]; [70,40,-20]
[40,0,-30]; [55,40,-25]; [35,40,-75]; [55,-50,35]; [85,-10,-5]; [30,-25,15]; [85,-50,0]; [35,10,25]
[85,-15,0]; [50,-45,40]; [45,-5,-25]; [85,-65,45]; [30,45,-30]; [70,5,-45]; [60,50,-15]; [30,60,-85]
[85,-60,30]; [75,25,60]; [35,-15,5]; [80,-40,-15]; [35,20,35]; [60,-35,45]; [70,0,-30]; [50,35,-25]
[85,-60,30]; [60,35,-35]; [70,-25,30]; [55,-5,-20]; [35,55,-20]; [85,-35,-20]; [40,-35,20]; [80,5,-30]

Slider settings:: PD:0.0 ND:0.5 NU:0.0 PP:1.0

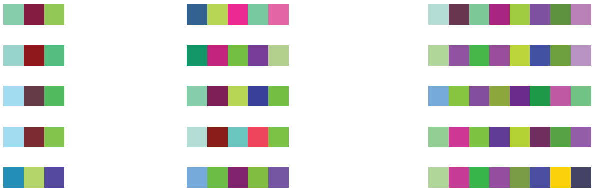


[55,-45,15]; [85,-25,-10]; [30,-20,0]
[85,-30,0]; [35,-25,15]; [60,-25,-10]
[85,-60,15]; [35,-30,20]; [85,-25,-10]
[55,-30,0]; [85,-20,-5]; [30,-35,30]
[55,-30,0]; [85,-40,55]; [30,-25,15]

[85,-30,15]; [30,-25,15]; [70,-25,-30]; [55,-35,50]; [35,-5,-30]
[85,-30,-15]; [40,-30,10]; [85,-20,30]; [45,-5,-35]; [70,-55,55]
[70,-30,0]; [30,-35,25]; [85,-55,45]; [35,-5,0]; [80,-25,45]
[85,-45,30]; [40,-40,25]; [75,-35,-15]; [30,-5,0]; [75,0,-35]
[55,-15,-15]; [75,-55,45]; [30,-5,-20]; [85,0,-15]; [35,45,-35]

[55,-30,0]; [85,-10,-15]; [30,0,-25]; [65,25,-40]; [35,30,-65]; [60,70,-45]; [85,-40,-15]; [30,-30,25]
[70,-30,0]; [35,-30,25]; [85,-55,50]; [45,-5,-35]; [70,30,-35]; [30,25,-50]; [80,0,-30]; [30,60,-60]
[55,-30,0]; [80,-15,-15]; [30,-25,15]; [80,-50,5]; [30,5,-20]; [80,-45,55]; [55,-5,-45]; [60,30,-35]
[85,-45,60]; [40,-25,-5]; [80,-30,-25]; [35,45,-45]; [55,15,-55]; [60,50,-55]; [30,5,-30]; [85,10,-20]
[85,-30,-15]; [35,-40,35]; [60,-20,-20]; [85,-50,30]; [35,-10,-15]; [85,0,-20]; [30,45,-60]; [75,45,-30]

Slider settings:: PD:1.0 ND:0.0 NU:0.0 PP:0.0

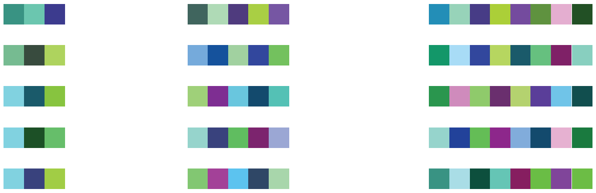


[85,-45,15]; [30,45,5]; [85,-55,60]
[85,-30,0]; [30,50,35]; [80,-65,30]
[85,-15,-15]; [30,20,0]; [75,-65,45]
[85,-15,-15]; [30,35,15]; [80,-55,60]
[55,-15,-30]; [85,-35,55]; [35,40,-60]

[40,0,-30]; [85,-35,65]; [55,85,-15]; [85,-55,20]; [60,55,-10]
[55,-45,15]; [45,65,-10]; [75,-55,60]; [35,70,-65]; [80,-20,30]
[85,-45,15]; [30,45,-10]; [85,-35,65]; [30,35,-60]; [85,-75,80]
[85,-15,0]; [30,45,35]; [85,-50,0]; [55,65,25]; [75,-50,60]
[70,0,-45]; [85,-80,80]; [30,55,-25]; [70,-40,60]; [45,60,-70]

[85,-15,0]; [30,25,-5]; [85,-55,25]; [40,60,-20]; [85,-50,70]; [45,80,-80]; [55,-35,55]; [65,50,-40]
[85,-30,30]; [50,75,-65]; [70,-65,60]; [45,45,-30]; [85,-35,75]; [35,45,-75]; [60,-35,60]; [70,45,-45]
[70,0,-45]; [80,-55,75]; [45,80,-75]; [65,-25,50]; [30,50,-45]; [55,-55,50]; [55,60,-30]; [80,-55,30]
[85,-45,30]; [50,65,-15]; [85,-70,80]; [30,65,-70]; [85,-40,80]; [30,35,-15]; [60,-40,40]; [55,70,-65]
[85,-30,30]; [50,70,-25]; [65,-60,55]; [50,85,-70]; [60,-25,40]; [35,45,-70]; [85,0,85]; [30,10,-20]

Slider settings:: PD:1.0 ND:0.0 NU:0.0 PP:1.0

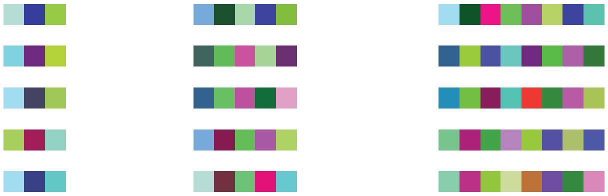


[55,-30,0]; [85,-55,10]; [30,25,-45]
[70,-30,15]; [30,-10,5]; [85,-40,60]
[85,-30,-15]; [35,-15,-15]; [75,-45,65]
[85,-30,-15]; [30,-30,25]; [75,-55,40]
[85,-30,-15]; [30,15,-35]; [80,-40,65]

[40,-15,0]; [85,-25,15]; [30,25,-35]; [80,-35,65]; [45,55,-65]
[70,0,-45]; [35,10,-45]; [80,-25,20]; [30,40,-70]; [80,-60,50]
[85,-45,45]; [35,50,-40]; [75,-20,-20]; [30,-5,-25]; [80,-50,0]
[85,-30,0]; [30,15,-35]; [75,-60,45]; [30,50,-25]; [70,10,-35]
[85,-60,45]; [45,65,-40]; [75,-15,-35]; [30,0,-20]; [85,-30,20]

[55,-15,-30]; [85,-35,10]; [30,25,-40]; [85,-45,75]; [40,60,-65]; [55,-35,55]; [80,40,-25]; [30,-25,20]
[55,-45,15]; [85,-10,-20]; [30,40,-70]; [85,-35,60]; [35,-15,-15]; [85,-70,35]; [30,50,-20]; [85,-40,5]
[55,-45,30]; [70,55,-35]; [85,-55,50]; [30,35,-25]; [80,-25,45]; [30,60,-70]; [75,-15,-25]; [30,-20,-5]
[85,-30,0]; [30,25,-55]; [75,-60,50]; [35,60,-35]; [70,0,-35]; [30,-5,-25]; [80,35,-20]; [45,-40,25]
[55,-30,0]; [85,-15,-10]; [30,-25,5]; [85,-55,5]; [30,55,-15]; [70,-50,60]; [40,65,-60]; [85,-80,75]

Slider settings:: PD:1.0 ND:0.0 NU:0.0 PP:0.5

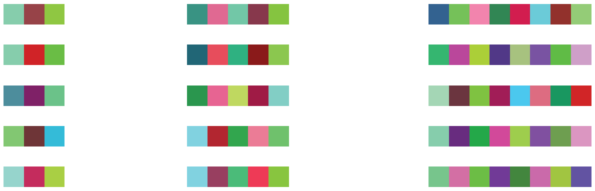


[85,-15,0]; [30,30,-55]; [80,-45,65
[85,-30,-15]; [30,45,-35]; [85,-40,75
[85,-15,-15]; [30,10,-20]; [75,-30,50
[85,-45,60]; [35,60,0]; [85,-35,5
[85,-15,-15]; [30,20,-40]; [85,-50,-5

[70,0,-45]; [30,-25,15]; [85,-30,20]; [30,40,-65]; [70,-40,70
[40,-15,0]; [80,-70,50]; [55,65,-25]; [85,-35,30]; [30,35,-25
[40,0,-30]; [75,-55,45]; [55,70,-35]; [40,-40,30]; [75,40,-20
[70,0,-45]; [30,50,-5]; [75,-60,50]; [55,70,-50]; [80,-30,50
[85,-15,0]; [30,30,5]; [85,-70,40]; [50,75,0]; [85,-45,-10

[85,-15,-15]; [30,-35,25]; [55,85,0]; [75,-55,50]; [50,70,-50]; [80,-25,50]; [30,40,-65]; [80,-50,5
[40,0,-30]; [80,-45,70]; [35,65,-90]; [85,-50,0]; [30,45,-35]; [80,-75,60]; [60,80,-60]; [45,-35,30
[55,-15,-30]; [75,-55,60]; [30,55,-10]; [80,-50,0]; [55,70,50]; [50,-40,35]; [55,60,-35]; [75,-25,50
[85,-60,30]; [40,60,-15]; [60,-50,45]; [65,45,-40]; [85,-55,75]; [35,70,-90]; [75,-20,40]; [40,35,-65
[85,-45,15]; [45,60,-15]; [80,-50,75]; [35,50,-60]; [85,-15,30]; [40,80,-85]; [50,-45,50]; [70,60,-30

Slider settings:: PD:1.0 ND:1.0 NU:0.0 PP:0.0

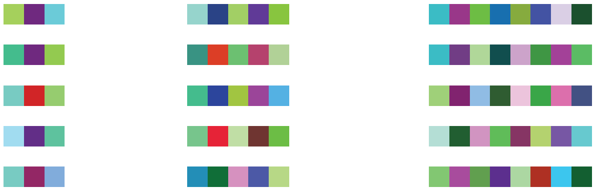


[85,-45,15]; [40,35,15]; [75,-40,60]
[85,-45,15]; [45,65,50]; [70,-50,60]
[55,-15,-15]; [30,50,-20]; [85,-65,30]
[85,-60,45]; [30,25,10]; [70,-25,-25]
[85,-30,0]; [45,60,10]; [80,-35,65]

[55,-30,0]; [60,50,0]; [85,-55,15]; [35,35,5]; [80,-55,65]
[40,-15,-15]; [55,60,25]; [65,-45,15]; [30,45,30]; [80,-50,60]
[55,-45,30]; [60,55,0]; [85,-30,60]; [35,55,10]; [85,-40,0]
[85,-30,-15]; [40,55,30]; [60,-50,35]; [65,45,5]; [85,-70,45]
[85,-30,-15]; [40,40,0]; [75,-60,30]; [55,70,30]; [85,-65,80]

[40,0,-30]; [85,-70,55]; [70,50,0]; [50,-35,20]; [45,70,20]; [80,-30,-15]; [35,40,25]; [80,-40,40]
[70,-60,30]; [50,65,-30]; [85,-45,75]; [30,30,-40]; [75,-20,30]; [45,75,-80]; [80,-75,75]; [75,45,-35]
[85,-30,15]; [30,25,5]; [75,-50,65]; [35,60,0]; [80,-30,-30]; [60,45,10]; [55,-45,20]; [45,65,55]
[85,-45,15]; [30,40,-35]; [60,-60,60]; [55,75,-25]; [85,-50,65]; [45,70,-70]; [60,-30,35]; [70,35,-15]
[85,-60,30]; [60,45,-10]; [85,-80,80]; [35,50,-50]; [50,-40,50]; [65,80,-45]; [75,-30,60]; [40,80,-95]

Slider settings:: PD:1.0 ND:1.0 NU:0.0 PP:1.0

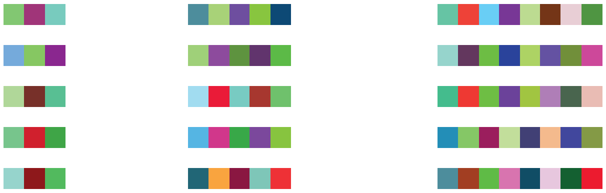


[85,-45,60]; [30,45,-35]; [80,-30,-15]
[70,-45,15]; [30,45,-35]; [80,-45,60]
[85,-45,0]; [45,65,55]; [80,-40,45]
[85,-15,-15]; [30,40,-40]; [80,-55,15]
[85,-45,0]; [35,50,-10]; [70,0,-35]

[85,-30,0]; [30,15,-40]; [80,-35,50]; [30,60,-65]; [85,-65,80]
[55,-30,0]; [50,60,60]; [80,-60,40]; [45,50,0]; [80,-20,25]
[70,-45,15]; [30,35,-65]; [75,-30,60]; [45,65,-45]; [70,-10,-45]
[85,-60,30]; [50,70,40]; [85,-20,25]; [30,25,15]; [85,-80,80]
[55,-15,-30]; [40,-45,40]; [70,45,-25]; [40,35,-65]; [85,-30,40]

[70,-30,-15]; [40,50,-25]; [75,-60,65]; [45,0,-40]; [65,-30,60]; [35,60,-90]; [85,10,-15]; [30,-25,15]
[70,-30,-15]; [35,35,-30]; [85,-30,30]; [30,-20,-5]; [75,40,-35]; [55,-45,45]; [45,65,-40]; [80,-70,45]
[85,-45,45]; [30,55,-25]; [75,-5,-25]; [35,-25,20]; [85,30,-20]; [60,-50,40]; [65,65,-25]; [35,10,-30]
[85,-15,0]; [35,-30,20]; [70,40,-25]; [80,-70,50]; [35,40,-10]; [80,-25,45]; [45,50,-60]; [85,-45,-10]
[85,-60,45]; [50,70,-45]; [60,-35,35]; [30,40,-45]; [85,-30,25]; [40,50,40]; [75,-25,-30]; [35,-35,20]

Slider settings:: PD:1.0 ND:1.0 NU:0.0 PP:0.5

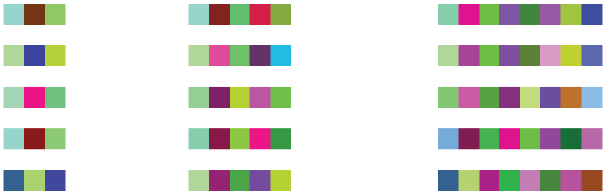


[85,-60,45]; [40,50,-15]; [80,-35,0]
[70,0,-45]; [80,-50,50]; [35,65,-45]
[85,-30,30]; [30,30,20]; [80,-60,20]
[85,-60,30]; [45,65,40]; [60,-50,45]
[85,-30,0]; [30,50,35]; [75,-65,45]

[55,-15,-15]; [80,-30,40]; [40,45,-55]; [85,-65,80]; [30,0,-30]
[85,-45,45]; [45,65,-55]; [55,-35,50]; [30,30,-25]; [80,-75,65]
[85,-15,-15]; [50,75,40]; [85,-45,0]; [40,45,30]; [85,-70,45]
[70,-15,-30]; [50,65,-10]; [60,-55,60]; [40,60,-60]; [85,-65,75]
[40,-15,-15]; [75,25,65]; [30,50,5]; [75,-25,0]; [55,75,50]

[85,-60,15]; [55,65,45]; [80,-20,-30]; [35,55,-50]; [85,-25,35]; [30,25,30]; [85,10,0]; [55,-40,45]
[85,-30,0]; [30,25,-15]; [75,-60,65]; [30,30,-60]; [80,-30,50]; [40,70,-85]; [55,-25,50]; [55,80,-30]
[70,-45,15]; [55,70,50]; [75,-60,70]; [35,55,-60]; [75,-30,60]; [65,55,-50]; [40,-15,10]; [80,15,10]
[55,-15,-30]; [85,-60,50]; [35,55,-5]; [85,-20,30]; [30,15,-30]; [80,15,30]; [30,50,-75]; [60,-20,40]
[55,-15,-15]; [40,40,40]; [70,-55,65]; [65,60,-25]; [30,-10,-20]; [85,25,-20]; [35,-40,35]; [50,75,45]

Slider settings:: PD:1.0 ND:0.5 NU:0.0 PP:0.0

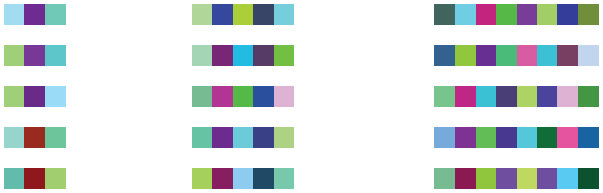


[85,-30,0]; [30,25,30]; [80,-45,50]
[85,-30,30]; [30,40,-65]; [85,-40,75]
[85,-30,15]; [55,85,0]; [85,-65,35]
[85,-30,0]; [30,45,30]; [85,-55,45]
[40,0,-30]; [80,-30,45]; [30,50,-75]

[85,-30,0]; [30,40,25]; [80,-65,40]; [45,70,25]; [65,-30,55]
[85,-30,30]; [55,65,-10]; [80,-60,45]; [30,30,-20]; [70,-25,-30]
[85,-45,30]; [30,50,-20]; [85,-40,80]; [55,65,-35]; [85,-75,60]
[85,-45,15]; [30,50,0]; [80,-50,65]; [55,85,0]; [55,-50,55]
[85,-30,30]; [35,55,-20]; [60,-50,60]; [40,60,-65]; [85,-40,80]

[85,-45,15]; [50,80,-15]; [75,-60,70]; [45,45,-50]; [50,-40,45]; [55,80,-70]; [75,-30,60]; [30,70,-95]
[85,-30,30]; [45,50,-25]; [75,-60,70]; [45,70,-70]; [50,-25,35]; [75,55,-35]; [80,-25,75]; [50,40,-75]
[85,-60,45]; [60,75,-35]; [60,-40,40]; [35,45,-25]; [85,-25,45]; [40,70,-80]; [55,25,50]; [75,-5,-30]
[70,0,-45]; [30,45,-5]; [65,-50,40]; [50,80,-15]; [85,-80,80]; [45,75,-60]; [40,-40,35]; [55,40,-20]
[40,0,-30]; [80,-25,45]; [40,65,-25]; [65,-65,65]; [65,55,-35]; [50,-35,35]; [60,95,-60]; [40,30,45]

Slider settings:: PD:1.0 ND:0.5 NU:0.0 PP:1.0

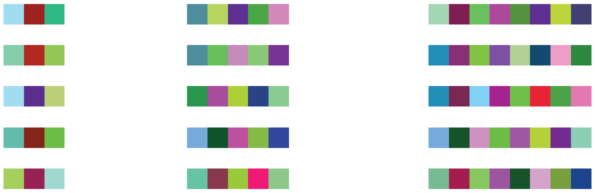


[85,-15,-15]; [30,60,-55]; [85,-50,5]
[85,-45,45]; [35,60,-55]; [80,-40,-10]
[85,-45,45]; [30,45,-40]; [85,-15,-20]
[85,-30,0]; [35,45,35]; [85,-60,20]
[70,-30,0]; [30,50,35]; [80,-35,45]

[85,-30,30]; [30,45,-75]; [85,-45,75]; [30,5,-20]; [85,-35,-15]
[85,-30,15]; [30,50,-30]; [70,-25,-30]; [30,20,-20]; [75,-55,60]
[70,-30,15]; [45,65,-30]; [70,-60,60]; [35,15,-45]; [80,35,-25]
[85,-60,15]; [30,55,-50]; [80,-30,-15]; [30,20,-40]; [80,-25,35]
[85,-45,60]; [30,55,-15]; [80,-10,-30]; [30,-5,-20]; [80,-40,10]

[40,-15,0]; [85,-35,-20]; [45,65,-10]; [75,-70,65]; [35,70,-65]; [80,-35,50]; [30,30,-55]; [55,-25,50]
[40,0,-30]; [85,-60,75]; [30,55,-55]; [75,-60,30]; [60,70,-25]; [75,-35,-20]; [35,30,-10]; [85,0,-20]
[85,-60,30]; [45,65,-15]; [75,-35,-20]; [30,20,-30]; [80,-30,50]; [30,55,-75]; [80,35,-25]; [55,-45,50]
[70,0,-45]; [35,60,-50]; [70,-50,45]; [30,30,-45]; [80,-35,-20]; [40,-40,25]; [60,75,-20]; [40,0,-40]
[70,-30,15]; [30,55,-5]; [85,-60,70]; [40,80,-85]; [85,-30,60]; [40,45,-55]; [80,-25,-30]; [30,-30,15]

Slider settings:: PD:1.0 ND:0.5 NU:0.0 PP:0.5

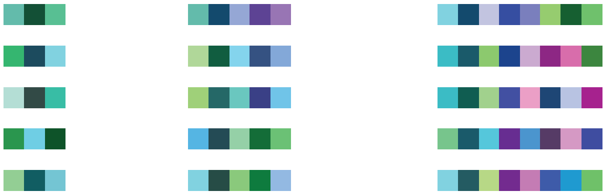


[85,-15,-15]; [35,50,45]; [70,-55,20]
[85,-45,15]; [40,55,45]; [85,-55,60]
[85,-15,-15]; [30,40,-45]; [80,-20,40]
[70,-30,0]; [30,40,40]; [85,-80,80]
[85,-45,60]; [35,50,0]; [85,-25,0]

[55,-15,-15]; [85,-35,60]; [30,45,-50]; [60,-50,60]; [70,60,-35]
[55,-15,-15]; [80,-65,50]; [70,60,-45]; [75,-35,35]; [35,50,-45]
[55,-45,30]; [50,70,-45]; [80,-35,75]; [30,15,-40]; [85,-50,30]
[70,0,-45]; [30,-35,25]; [55,70,-35]; [70,-35,50]; [30,40,-70]
[85,-60,15]; [35,35,5]; [80,-45,75]; [55,80,10]; [75,-30,25]

[85,-30,15]; [30,45,-5]; [85,-75,50]; [50,75,-45]; [55,-40,55]; [30,45,-50]; [85,-35,75]; [30,15,-30]
[55,-15,-30]; [35,45,-20]; [75,-50,70]; [45,85,-85]; [80,-20,25]; [30,-5,-25]; [75,40,-15]; [50,-45,45]
[55,-15,-30]; [30,40,-5]; [80,-15,-25]; [40,70,-35]; [85,-75,60]; [50,70,40]; [60,-45,40]; [65,50,-15]
[70,0,-45]; [30,-35,30]; [70,45,-30]; [80,-70,70]; [55,80,-65]; [85,-40,75]; [30,60,-50]; [85,-40,10]
[70,-30,15]; [35,60,5]; [85,-60,55]; [50,60,-45]; [30,-35,30]; [75,35,-25]; [60,-30,60]; [30,15,-45]

Slider settings:: PD:0.5 ND:0.0 NU:0.0 PP:1.0

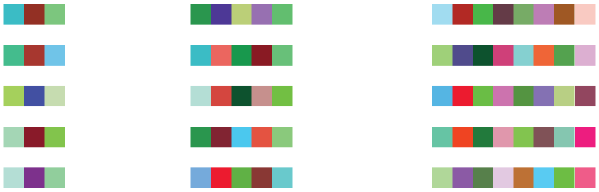


[70,-30,0]; [30,-25,10]; [80,-60,20]
[70,-60,30]; [30,-10,-15]; [85,-30,-15]
[85,-15,0]; [30,-10,0]; [75,-50,5]
[55,-45,30]; [85,-35,-20]; [30,-35,25]
[85,-45,30]; [35,-20,-10]; [75,-20,-15]

[70,-30,0]; [30,-5,-25]; [70,10,-40]; [35,30,-40]; [60,50,-55]
[85,-30,30]; [35,-30,10]; [85,-25,-20]; [35,5,-30]; [70,5,-45]
[85,-45,45]; [40,-20,-5]; [85,-50,0]; [30,20,-40]; [75,-15,-25]
[70,-15,-30]; [30,-10,-10]; [85,-40,20]; [40,-40,30]; [85,-70,40]
[85,-30,-15]; [30,-15,0]; [85,-55,40]; [45,-45,30]; [75,0,-35]

[85,-30,-15]; [30,-5,-25]; [80,5,-15]; [35,25,-55]; [60,30,-60]; [80,-40,45]; [35,-35,25]; [85,-70,45]
[70,-30,-15]; [35,-15,-15]; [80,-45,45]; [30,15,-45]; [75,25,-25]; [35,55,-30]; [65,70,-30]; [50,-40,40]
[70,-30,-15]; [35,-25,0]; [85,-40,35]; [35,40,-70]; [75,40,-15]; [30,5,-30]; [80,5,-25]; [40,70,-35]
[85,-60,30]; [35,-15,-15]; [80,-35,-20]; [30,50,-50]; [60,-5,-35]; [30,20,-20]; [70,30,-15]; [30,70,-95]
[85,-30,-15]; [35,-15,-10]; [85,-30,40]; [30,60,-50]; [65,55,-35]; [40,30,-65]; [60,-15,-35]; [80,-60,45]

Slider settings:: PD:0.5 ND:1.0 NU:0.0 PP:0.0

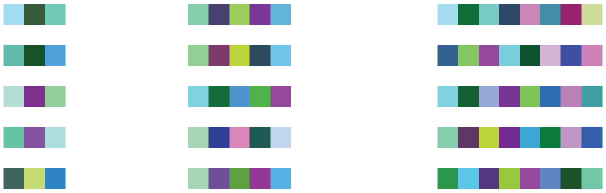


[70,-30,-15]; [35,40,30]; [80,-50,35]
[70,-45,15]; [40,45,30]; [75,-15,-25]
[85,-45,60]; [35,40,-70]; [85,-15,20]
[85,-30,15]; [30,45,20]; [80,-55,60]
[85,-15,0]; [35,45,-35]; [85,-45,25]

[55,-45,30]; [30,40,-55]; [80,-20,40]; [55,35,-40]; [80,-65,40]
[70,-30,-15]; [60,50,30]; [55,-50,30]; [30,45,25]; [80,-60,35]
[85,-15,0]; [50,55,35]; [30,-30,15]; [65,20,10]; [70,-45,55]
[55,-45,30]; [30,40,15]; [80,-30,-30]; [55,55,40]; [85,-55,40]
[70,0,-45]; [50,75,45]; [65,-45,50]; [35,35,20]; [80,-35,-10]

[85,-15,-15]; [40,55,45]; [70,-65,60]; [30,20,0]; [65,-30,30]; [65,55,-40]; [45,25,50]; [85,15,10]
[85,-45,45]; [35,20,-35]; [30,-30,15]; [50,60,0]; [85,-35,-5]; [60,50,50]; [60,-40,35]; [80,40,-30]
[70,-15,-30]; [50,75,45]; [85,-80,70]; [65,65,-35]; [55,-40,55]; [55,35,-50]; [80,-20,35]; [40,35,0]
[85,-60,15]; [55,65,65]; [45,-40,30]; [70,30,0]; [85,-65,60]; [40,20,5]; [75,-25,5]; [55,80,5]
[85,-30,30]; [50,55,-55]; [50,-25,25]; [85,20,-20]; [55,25,45]; [80,-25,-30]; [80,-70,75]; [60,60,5]

Slider settings:: PD:0.5 ND:1.0 NU:0.0 PP:1.0

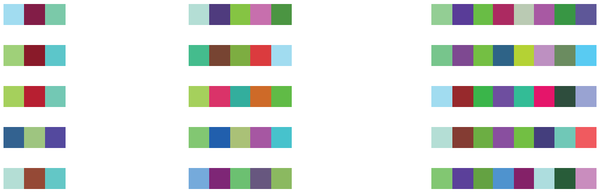


[85,-15,-15]; [35,-20,15]; [85,-50,5]
[70,-30,0]; [30,-35,30]; [65,-5,-45]
[85,-15,-15]; [40,-35,20]; [85,-45,-5]
[85,-60,15]; [45,50,-50]; [85,-15,-5]
[40,-15,0]; [85,-25,50]; [55,0,-50]

[85,-45,15]; [30,15,-25]; [85,-50,60]; [35,60,-55]; [70,-15,-25]
[85,-45,30]; [35,35,-15]; [85,-35,75]; [30,-5,-15]; [75,-15,-25]
[85,-30,-15]; [40,-40,30]; [60,0,-45]; [65,-50,50]; [45,60,-45]
[85,-30,15]; [30,25,-50]; [70,60,-30]; [35,-20,0]; [85,0,-20]
[85,-30,15]; [40,30,-35]; [60,-40,50]; [40,65,-45]; [70,-10,-45]

[85,-15,-15]; [40,-40,25]; [85,-45,0]; [30,0,-20]; [70,65,-45]; [55,-15,-20]; [35,55,-15]; [85,-15,30]
[40,0,-30]; [75,-40,45]; [45,60,-45]; [80,-25,-15]; [30,-30,15]; [80,30,-30]; [35,30,-60]; [70,70,-45]
[85,-30,-15]; [35,-35,20]; [70,10,-40]; [35,45,-40]; [75,-45,50]; [45,5,-40]; [65,50,-40]; [60,-25,-10]
[85,-45,15]; [30,25,-20]; [85,-35,75]; [30,60,-50]; [65,-15,-30]; [45,-45,30]; [70,35,-35]; [40,30,-70]
[55,-45,30]; [75,-20,-25]; [30,30,-35]; [85,-55,75]; [45,60,-45]; [60,15,-60]; [30,-30,25]; [85,-55,15]

Slider settings:: PD:0.5 ND:1.0 NU:0.0 PP:0.5

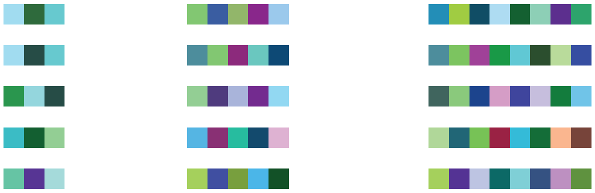


[85,-15,-15]; [30,45,0]; [85,-50,15]
[85,-45,45]; [30,45,20]; [80,-40,-10]
[85,-45,60]; [40,60,30]; [80,-40,5]
[40,0,-30]; [75,-25,30]; [35,40,-60]
[85,-15,0]; [40,30,25]; [85,-50,-5]

[85,-15,0]; [30,25,-35]; [85,-65,65]; [60,50,-25]; [55,-40,40]
[70,-45,15]; [35,20,20]; [65,-35,60]; [50,60,35]; [85,-15,-15]
[85,-45,60]; [50,65,10]; [65,-35,0]; [55,35,55]; [80,-75,75]
[85,-60,45]; [40,10,-50]; [75,-20,35]; [50,50,-35]; [75,-35,-15]
[70,0,-45]; [30,55,-30]; [80,-60,40]; [40,15,-20]; [70,-30,40]

[85,-45,30]; [30,55,-65]; [85,-80,65]; [40,55,0]; [80,-10,10]; [55,70,-50]; [55,-45,40]; [40,20,-35]
[85,-60,30]; [40,35,-30]; [85,-75,80]; [40,-5,-25]; [85,-40,80]; [70,50,-45]; [55,-20,20]; [80,-25,-30]
[85,-15,-15]; [35,45,25]; [65,-60,60]; [40,40,-50]; [70,-45,10]; [50,75,10]; [30,-15,5]; [70,15,-45]
[85,-15,0]; [35,30,20]; [65,-40,50]; [45,55,-50]; [85,-75,70]; [30,20,-35]; [85,-50,5]; [60,60,30]
[85,-60,45]; [30,65,-75]; [60,-40,60]; [60,0,-45]; [30,55,-20]; [85,-15,-5]; [35,-25,15]; [70,55,-40]

Slider settings:: PD:0.5 ND:0.5 NU:0.0 PP:1.0



[85,-15,-15]; [40,-30,20]; [85,-45,-10]
[85,-15,-15]; [30,-15,0]; [85,-45,-10]
[55,-45,30]; [85,-25,-10]; [30,-15,0]
[70,-30,-15]; [35,-35,20]; [85,-45,30]
[85,-60,15]; [30,45,-55]; [85,-20,-5]

[85,-60,45]; [40,10,-40]; [70,-25,35]; [35,55,-35]; [80,-5,-30]
[55,-15,-15]; [85,-60,45]; [35,50,-25]; [85,-50,0]; [30,0,-30]
[85,-45,30]; [30,25,-35]; [75,10,-35]; [30,60,-50]; [85,-20,-20]
[70,-15,-30]; [35,45,-20]; [70,-45,5]; [30,-5,-25]; [80,35,-25]
[85,-45,60]; [35,35,-65]; [60,-30,60]; [70,-15,-35]; [30,-30,20]

[55,-15,-30]; [85,-50,70]; [30,-10,-20]; [85,-10,-15]; [35,-40,35]; [85,-40,10]; [30,40,-45]; [60,-45,20]
[55,-15,-15]; [80,-55,50]; [45,70,-45]; [55,-55,45]; [80,-35,-15]; [30,-20,15]; [85,-25,30]; [35,25,-55]
[40,-15,0]; [85,-55,40]; [30,15,-45]; [75,50,-35]; [30,45,-70]; [80,15,-30]; [45,-45,35]; [75,-15,-25]
[85,-30,30]; [40,-15,-15]; [80,-60,55]; [35,50,10]; [70,-25,-25]; [40,-40,30]; [80,20,30]; [35,20,15]
[85,-45,60]; [30,40,-50]; [80,5,-20]; [40,-25,-5]; [85,-35,-10]; [35,5,-30]; [70,50,-45]; [55,-35,55]

Appendix B

Supplementary Material, Evaluating visual analysis interaction classification to improve understanding of cancer genomics domain experts and their tasks

B.1 Overview

In this supplemental material we provide additional context for related background material and our log culling procedure.

B.2 Additional background on historical log analysis techniques

In Chapter 4.1 we briefly mentioned previous clickstream research. Here, we specifically enumerate previous contributions given their importance in the development of interaction analysis and their important historical context.

The rise of internet search engines caused a large effort in understanding users online behavior as a way to optimize search ranking results for individual users. Although clickstreams are fundamentally different than our mouse-based approach, it is important to mention this past work given their historical importance. Clickstream analysis typically models user navigation as a graphical representation, similar to our region of interest feature set, and has modeled web page browsing through a variety of methods. Researchers have modeled web browsing behavior with clickstream data using techniques such as longest repeating subsequence analysis [144], decision trees [37], pattern extraction [37], various types of Markov and other graphical models [75, 158], neural nets [4], and clustering [71, 199]. Some, like Liu et al., have also explored how these pattern extraction techniques can be paired with visual analysis to empower analysts' abilities to uncover important user behavior [114]. Given the breadth of clickstream research, and its importance to search, these are but a few articles from a much larger literature. For an overview of other historical clickstream and related server-based log analysis contributions, we recommend Ivory and Hearst's survey [83].

B.3 Additional interaction log culling information

One caveat with our MAGI interaction log collection was that we also collected a large amount of logs with few or no interaction events over our year-long data collection period. While it was easy to remove logs without any events, identifying a "too few" threshold was not obvious.

As noted in the main text, our ultimate strategy was to remove any logs with fewer than 39 events. We derived this threshold in a three-step process after discovering that typical, simple thresholding methods (e.g., standard deviation) did not give satisfactory results. First, we generated quantile-quantile plots to examine the distribution of event totals across all logs in an effort to discover why typical approaches were failing. Second, after initially plotting against a normal distribution we believed the skew was so severe that it would be better to derive the threshold through a

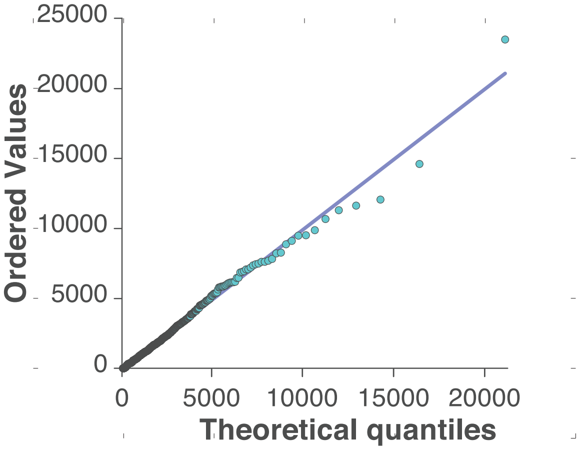


Figure B.1: A quantile-quantile plot shows that number of mouse events has a strong lognormal fit (lognormal parameters: $\mu = -71.99$, $\sigma = 774.48$).

lognormal distribution (Figure B.1). Last, while the lognormal quantile-quantile plot supported our prediction, we also ran a follow up Kolmogorov–Smirnov for further validation. Given the close fit ($D = 0.04, p = 0.006$) we then settled on using the lower boundary of the central 95% interval as the threshold point. We did not remove outliers in the top 5% given that they were likely to contain analytical tasks.

Because only 63 logs were removed, and our ultimate classification results showed a high proportion of junk logs that were too short, it is likely that we could have used a less conservative bound. However, less conservative filtering comes with added risk of removing useful logs. While future applications of our proposed interaction analysis methods could take advantage of less restrictive filtering, we did not feel comfortable doing so in this work given that we were investigating a new methodology rather than applying an established one.

Bibliography

[1] D. Acevedo, E. Vote, D. H. Laidlaw, and M. S. Joukowsky. Archaeological data visualization in vr: Analysis of lamp finds at the great temple of petra, a case study. In *Proceedings of the Conference on Visualization '01*, VIS '01, pages 493–496, Washington, DC, USA, 2001. IEEE Computer Society. ISBN 0-7803-7200-X. URL `http://dl.acm.org/citation.cfm?id=601671.601760`.

[2] E. E. Aftandilian, S. Kelley, C. Gramazio, N. Ricci, S. L. Su, and S. Z. Guyer. Heapviz: A programmer's tool for data structure visualization. In *IEEE VisWeek Demo*, 2010.

[3] E. E. Aftandilian, S. Kelley, C. Gramazio, N. Ricci, S. L. Su, and S. Z. Guyer. Heapviz: Interactive heap visualization for program understanding and debugging. *Proceedings of the 5th International Symposium on Software Visualization*, pages 53–62, 2010. doi: 10.1145/1879211.1879222. URL `http://doi.acm.org/10.1145/1879211.1879222`.

[4] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proc. of Research and development in information retrieval (SIGIR)*, pages 19–26, 2006. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148177. URL `http://doi.acm.org/10.1145/1148170.1148177`.

[5] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof. A comparison of scanpath comparison methods. *Behavior Research Methods*, 47(4):1377–1392, 2015. ISSN 1554-3528. doi: 10.3758/s13428-014-0550-3. URL `http://dx.doi.org/10.3758/s13428-014-0550-3`.

[6] R. Atterer, M. Wnuk, and A. Schmidt. Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction. In *Proc. of World Wide Web (WWW)*,

pages 203–212, 2006. ISBN 1-59593-323-9. doi: 10.1145/1135777.1135811. URL http://doi.acm.org/10.1145/1135777.1135811.

[7] D. M. Best, S. Bohn, D. Love, A. Wynne, and W. A. Pike. Real-time visualization of network behaviors for situational awareness. In *Proceedings of the Seventh International Symposium on Visualization for Cyber Security*, VizSec, pages 79–90, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0013-1. doi: 10.1145/1850795.1850805. URL http://doi.acm.org/10.1145/1850795.1850805.

[8] T. Blascheck, M. John, K. Kurzhals, S. Koch, and T. Ertl. VA^2: A visual analytics approach for evaluating visual analytics applications. *Trans. Vis. Comput. Graphics*, 22(1):61–70, Jan 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467871. URL http://dx.doi.org/10.1109/TVCG.2015.2467871.

[9] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013. ISSN 1077-2626. doi: http://doi.ieeecomputersociety.org/10.1109/TVCG.2013.234.

[10] D. Borland and R. Taylor. Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications*, 27(2):14–17, March 2007. ISSN 0272-1716. doi: 10.1109/MCG.2007.323435.

[11] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011. ISSN 1077-2626. doi: 10.1109/TVCG.2011.185. URL http://dx.doi.org/10.1109/TVCG.2011.185.

[12] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J. Fekete. Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2769–2778, Dec 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2012.220.

[13] M. J. Bravo and K. Nakayama. The role of attention in different visual-search tasks. *Perception & Psychophysics*, 51(5):465–472, 1992. ISSN 0031-5117. doi: 10.3758/BF03211642. URL http://dx.doi.org/10.3758/BF03211642.

[14] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, Dec 2013. ISSN 1077-2626. doi: 10.1109/TVCG.2013.124.

[15] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *Trans. Vis. Comput. Graphics*, 20(12):2271–2280, Dec 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346431. URL `http://dx.doi.org/10.1109/TVCG.2014.2346431`.

[16] M. Brehmer, J. Ng, K. Tate, and T. Munzner. Matches, mismatches, and methods: Multiple-view workflows for energy portfolio analysis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):449–458, Jan 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2466971.

[17] C. A. Brewer, G. W. Hatchard, and M. A. Harrower. Colorbrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30(1):5–32, 2003. doi: 10.1559/152304003100010929. URL `http://dx.doi.org/10.1559/152304003100010929`.

[18] F. P. Brooks, Jr. The computer scientist as toolsmith ii. *Commun. ACM*, 39(3):61–68, Mar. 1996. ISSN 0001-0782. doi: 10.1145/227234.227243. URL `http://doi.acm.org/10.1145/227234.227243`.

[19] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *Trans. Vis. Comput. Graphics*, 20(12):1663–1672, Dec 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346575. URL `http://dx.doi.org/10.1109/TVCG.2014.2346575`.

[20] S. Carpendale. *Evaluating Information Visualizations*, pages 19–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[21] P. A. Carpenter and P. Shah. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4(2):75–100, 1998.

[22] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012. doi: 10.1158/2159-8290.

CD-12-0095. URL http://cancerdiscovery.aacrjournals.org/content/2/5/401.abstract.

[23] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *IEEE Symposium on Visual Analytics Science and Technology.*, pages 155–162, Oct 2007. doi: 10.1109/VAST.2007.4389009.

[24] M. E. Chevreul. *The principles of harmony and contrast of colours, and their applications to the arts*. Van Nostrand Reinhold, New York, NY, USA, 1987 (1839).

[25] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[26] Çağatay. Demiralp, M. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1933–1942, Dec 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346978. URL http://dx.doi.org/10.1109/tvcg.2014.2346978.

[27] M. Dörk, S. Carpendale, and C. Williamson. The information flaneur: A fresh look at information seeking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1215–1224, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979124. URL http://doi.acm.org/10.1145/1978942.1979124.

[28] M. Dörk, N. H. Riche, G. Ramos, and S. Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, Dec 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2012.252.

[29] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433–458, 1989.

[30] A. Edmonds, R. W. White, D. Morris, and S. M. Drucker. Instrumenting the dynamic web. *Journal of Web Engineering*, 6(3):243, 2007.

[31] S. G. Eick and A. F. Karr. Visual scalability. *Journal of Computational and Graphical Statistics*, 11(1):22–43, 2002.

[32] J. M. Enoch. Effect of the size of a complex display upon visual search. *Journal of the Optical Society of America*, 49(3):280–285, Mar 1959. doi: 10.1364/JOSA.49.000280. URL `http://www.opticsinfobase.org/abstract.cfm?URI=josa-49-3-280`.

[33] R. Etemadpour, M. Bomhoff, E. Lyons, P. Murray, and A. Forbes. Designing and evaluating scientific workflows for big data interactions. In *2015 Big Data Visual Analytics (BDVA)*, Sept 2015. doi: 10.1109/BDVA.2015.7314290.

[34] J. A. Fails, A. Karlson, L. Shahamat, and B. Shneiderman. A visual interface for multivariate temporal data: Finding patterns of events across multiple histories. In *Proc. of Visual analytics, science, and technology (VAST)*, pages 167–174, Oct 2006. doi: 10.1109/VAST.2006.261421. URL `http://dx.doi.org/10.1109/VAST.2006.261421`.

[35] M. Feng, C. Deng, E. Peck, and L. Harrison. Hindsight: Encouraging exploration through direct encoding of personal interaction history. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2016.2599058.

[36] M. Fink, J.-H. Haunert, J. Spoerhase, and A. Wolff. Selecting the aspect ratio of a scatter plot based on its delaunay triangulation. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2326–2335, 2013. ISSN 1077-2626. doi: http://doi.ieeecomputersociety.org/10.1109/TVCG.2013.187.

[37] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, Apr. 2005. ISSN 1046-8188. doi: 10.1145/1059981.1059982. URL `http://doi.acm.org/10.1145/1059981.1059982`.

[38] N. Gehlenborg and B. Wong. Points of view: Power of the plane. *Nat Meth*, 9(10):935–935, 10 2012. URL `http://dx.doi.org/10.1038/nmeth.2186`.

[39] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, et al. Visualization of omics data for systems biology. *Nature methods*, 7:S56–S68, 2010.

[40] M. Ghoniem, J.-D. Fekete, and P. Castagliola. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135, July 2005. ISSN 1473-8716. doi: 10.1057/palgrave.ivs.9500092. URL `http://dx.doi.org/10.1057/palgrave.ivs.9500092`.

[41] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 19(12): 2316–2325, Dec 2013. ISSN 1077-2626. doi: 10.1109/TVCG.2013.183.

[42] J. H. Goldberg and J. I. Helfman. Scanpath clustering and aggregation. In *Proc. of Eye-Tracking Research & Applications*, pages 227–234, 2010. ISBN 978-1-60558-994-7. doi: 10.1145/1743666.1743721. URL `http://doi.acm.org/10.1145/1743666.1743721`.

[43] S. Gomez, R. Jianu, R. Cabeen, H. Guo, and D. Laidlaw. Fauxvea: Crowdsourcing gaze location estimates for visualization analysis tasks. *Trans. Vis. Comput. Graphics*, PP(99):1–1, 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2016.2532331. URL `http://dx.doi.org/10.1109/TVCG.2016.2532331`.

[44] S. R. Gomez, H. Guo, C. Ziemkiewicz, and D. H. Laidlaw. An insight- and task-based methodology for evaluating spatiotemporal visual analytics. In *Proc. of Visual analytics, science, and technology (VAST)*, pages 63–72, Oct 2014. doi: 10.1109/VAST.2014.7042482. URL `http://dx.doi.org/10.1109/VAST.2014.7042482`.

[45] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6):333–351, 06 2016. URL `http://dx.doi.org/10.1038/nrg.2016.49`.

[46] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. In *Proc. of Visual analytics, science, and technology (VAST)*, pages 123–130, Oct 2008. doi: 10.1109/VAST.2008.4677365. URL `http://dx.doi.org/10.1109/VAST.2008.4677365`.

[47] C. Gramazio and R. Chang. Optimizing an spt-tree for visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012.

[48] C. C. Gramazio, K. B. Schloss, and D. H. Laidlaw. The relation between visualization size, grouping, and user performance. *Transactions on Visualization and Computer Graphics (Proc.*

VIS ‘14), 20(12):1953–1962, Dec 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346983. URL `http://dx.doi.org/10.1109/TVCG.2014.2346983`.

[49] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *Transactions on Visualization and Computer Graphics (Proc. VIS ’16)*, 2017. ISSN 1077-2626. doi: 10.1109/TVCG.2016.2598918. URL `http://vrl.cs.brown.edu/color`.

[50] C. C. Gramazio, J. Huang, and D. Laidlaw. An analysis of visual analysis: Modeling the interactive visualization tasks of cancer genomics domain experts. *Transactions on Visualization and Computer Graphics*, In Revision.

[51] C. C. Gramazio, M. Leiserson, B. Raphael, and D. Laidlaw. A cancer genomics visualization task requirements analysis and design study of magi. *Transactions on Visualization and Computer Graphics*, Under Review.

[52] G. Griffin, S. Li, C. Gramazio, and R. Chang. An analytical approach for the creative design of new visualizations. In *IEEE Conference on Information Visualization (InfoVis)*, 2011.

[53] U. Groemping. Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software*, 17(1):1–27, 2006. ISSN 1548-7660. doi: 10.18637/jss.v017.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v017i01`.

[54] H. Guo, J. Huang, and D. H. Laidlaw. Representing uncertainty in graph edges: An evaluation of paired visual variables. *IEEE Transactions on Visualization and Computer Graphics*, 21 (10):1173–1186, Oct 2015. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2424872. URL `http://dx.doi.org/10.1109/TVCG.2015.2424872`.

[55] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *Trans. Vis. Comput. Graphics*, 22(1):51–60, Jan 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467613. URL `http://dx.doi.org/10.1109/TVCG.2015.2467613`.

[56] Q. Guo and E. Agichtein. Ready to buy or just browsing?: Detecting web searcher goals from interaction data. In *Proc. of Research and development in information retrieval (SIGIR)*,

pages 130–137, 2010. ISBN 978-1-4503-0153-4. doi: 10.1145/1835449.1835473. URL http://doi.acm.org/10.1145/1835449.1835473.

[57] J. Haberman and D. Whitney. Ensemble perception: summarizing the scene and broadening the limits of visual processing. In J. Wolfe and L. Robertson, editors, *From Perception to Consciousness: Searching with Anne Treisman*, pages 339–349. Oxford University Press, 2012.

[58] A. Hakone, L. Harrison, A. Ottley, N. Winters, C. Gutheil, P. K. J. Han, and R. Chang. Proact: Iterative design of a patient-centered visualization for effective prostate cancer health risk communication. *IEEE Transactions on Visualization and Computer Graphics*, PP(99): 1–1, 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2016.2598588.

[59] L. H. Hardy, G. Rand, M. C. Rittler, J. Neitz, and J. Bailey. *HRR pseudoisochromatic plates*. Richmond Products, 2002.

[60] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2402–2410, Dec 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2012.233.

[61] J. Harper and M. Agrawala. Deconstructing and restyling d3 visualizations. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 253–262, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3069-5. doi: 10.1145/2642918.2647411. URL http://doi.acm.org/10.1145/2642918.2647411.

[62] L. Harrison. Experimentr. https://github.com/codementum/experimentr, 2014.

[63] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, Dec 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346979.

[64] M. Harrower and C. A. Brewer. Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. doi: 10.1179/000870403235002042. URL http://dx.doi.org/10.1179/000870403235002042.

[65] M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. doi: 10.1179/000870403235002042. URL http://www.maneyonline.com/doi/abs/10.1179/000870403235002042.

[66] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139 – 183. North-Holland, 1988. doi: http://dx.doi.org/10.1016/S0166-4115(08)62386-9. URL `http://www.sciencedirect.com/science/article/pii/S0166411508623869`.

[67] C. Healey. Choosing effective colours for data visualization. In *Proceedings of Visualization*, pages 263–270, Oct 1996. doi: 10.1109/VISUAL.1996.568118.

[68] C. G. Healey, K. S. Booth, and J. T. Enns. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction*, 3(2):107–135, June 1996. ISSN 1073-0516. doi: 10.1145/230562.230563. URL `http://doi.acm.org/10.1145/230562.230563`.

[69] J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. *Information visualization*, 7(1):49–62, 2008.

[70] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 203–212, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9. doi: 10.1145/1753326.1753357. URL `http://doi.acm.org/10.1145/1753326.1753357`.

[71] J. Heer and E. H. Chi. Separating the swarm: Categorization methods for user sessions on the web. In *Proc. of Human Factors in Computing Systems (CHI)*, pages 243–250, 2002. ISBN 1-58113-453-3. doi: 10.1145/503376.503420. URL `http://doi.acm.org/10.1145/503376.503420`.

[72] J. Heer and G. Robertson. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247, Nov 2007. ISSN 1077-2626. doi: 10.1109/TVCG.2007.70539. URL `http://dx.doi.org/10.1109/TVCG.2007.70539`.

[73] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1007–1016, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2208547. URL `http://doi.acm.org/10.1145/2207676.2208547`.

[74] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1303–1312, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518897. URL `http://doi.acm.org/10.1145/1518701.1518897`.

[75] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse. The lumière project: Bayesian user modeling for inferring the goals and needs of software users. In *Proc. of Uncertainty in Artificial Intelligence*, pages 256–265, 1998. URL `http://dl.acm.org/citation.cfm?id=2074094.2074124`.

[76] G. Hu, Z. Pan, M. Zhang, D. Chen, W. Yang, and J. Chen. An interactive method for generating harmonious color schemes. *Color Research and Application*, 39(1):70–78, 2014. ISSN 1520-6378. doi: 10.1002/col.21762. URL `http://dx.doi.org/10.1002/col.21762`.

[77] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6:1049, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.01049. URL `http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01049`.

[78] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1225–1234, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979125. URL `http://doi.acm.org/10.1145/1978942.1979125`.

[79] J. Hullman, E. Adar, and P. Shah. Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2213–2222, Dec 2011. ISSN 1077-2626. doi: 10.1109/TVCG.2011.175.

[80] L. M. Hurvich and D. Jameson. An opponent-process theory of color vision. *Psychological Review*, 64(6):384–404, November 1957. doi: 10.1037/h0041403. URL `http://dx.doi.org/10.1037/h0041403`.

[81] P. Isenberg, A. Tang, and S. Carpendale. An exploratory study of visual information analysis.

In *Proc. of Human Factors in Computing Systems (CHI)*, pages 1217–1226, 2008. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357245. URL `http://doi.acm.org/10.1145/1357054.1357245`.

[82] J. Itten. *The art of color.* Van Nostrand Reinhold, New York, NY, USA, 1961.

[83] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR)*, 33(4):470–516, 2001. URL `http://dx.doi.org/10.1145/503112.503114`.

[84] G. G. K. J. Richard Landis. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2529310`.

[85] M. Jakobsen and K. Hornbaek. Interactive visualizations on large and small displays: The interrelation of display size, information space, and scale. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2336–2345, Dec 2013. ISSN 1077-2626. doi: 10.1109/TVCG.2013.170.

[86] D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972. URL `http://dx.doi.org/10.1016/0010-0285(72)90016-3`.

[87] S. Kairam, N. H. Riche, S. Drucker, R. Fernandez, and J. Heer. Refinery: Visual exploration of large, heterogeneous networks through associative browsing. *Computer Graphics Forum*, 34(3):301–310, 2015. ISSN 1467-8659. doi: 10.1111/cgf.12642. URL `http://dx.doi.org/10.1111/cgf.12642`.

[88] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011. doi: 10.1177/1473871611415994. URL `http://ivi.sagepub.com/content/10/4/271.abstract`.

[89] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2917–2926, Dec 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2012.219.

[90] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):469–478, Jan 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467671.

[91] S. Kelley, E. Aftandilian, C. Gramazio, N. Ricci, S. L. Su, and S. Z. Guyer. Heapviz: Interactive heap visualization for program understanding and debugging (extended). *Information Visualization*, 12(2):163–177, 2013. doi: 10.1177/1473871612438786. URL `http://ivi.sagepub.com/content/12/2/163.abstract`.

[92] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002. doi: 10.1101/gr.229102. URL `http://genome.cshlp.org/content/12/6/996.abstract`.

[93] R. Kimchi. Primacy of wholistic processing and global/local paradigm: a critical review. *Psychological bulletin*, 112(1):24–38, 1992.

[94] R. Kimchi and S. E. Palmer. Form and texture in hierarchically constructed patterns. *Journal of Experimental Psychology: Human Perception and Performance*, 8(4):521–535, 1982.

[95] R. Kinchla and J. Wolfe. The order of visual processing:"top-down," "bottom-up," or "middle-out". *Perception & Psychophysics*, 25(3):225–231, 1979. ISSN 0031-5117. doi: 10.3758/BF03202991. URL `http://dx.doi.org/10.3758/BF03202991`.

[96] N. Kong, J. Heer, and M. Agrawala. Perceptual guidelines for creating rectangular treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):990–998, Nov 2010. ISSN 1077-2626. doi: 10.1109/TVCG.2010.186.

[97] R. Kosara. The problem with vis taxonomies. `https://eagereyes.org/blog/2016/the-problem-with-vis-taxonomies`. Accessed: 2016-11-28.

[98] S. M. Kosslyn. *Graph design for the eye and mind*. Oxford University Press, 2006.

[99] K. Krejtz, T. Szmidt, A. T. Duchowski, and I. Krejtz. Entropy-based statistical analysis of eye movement transitions. In *Proc. of Eye Tracking Research and Applications*, pages 159–166, 2014. ISBN 978-1-4503-2751-0. doi: 10.1145/2578153.2578176. URL `http://doi.acm.org/10.1145/2578153.2578176`.

[100] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009. doi: 10.1101/gr.092759.109. URL `http://genome.cshlp.org/content/19/9/1639.abstract`.

[101] R. Kumar, J. O. Talton, S. Ahmad, and S. R. Klemmer. Bricolage: Example-based retargeting for web design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2197–2206, 2011. doi: 10.1145/1978942.1979262. URL `http://doi.acm.org/10.1145/1978942.1979262`.

[102] Y. Kuzmova, J. Wolfe, A. Rich, A. Brown, D. Lindsey, and E. Reijnen. Pink: the most colorful mystery in visual search. *Journal of Vision*, 8(6):382, 2008. doi: 10.1167/8.6.382. URL `http://www.journalofvision.org/content/8/6/382.abstract`.

[103] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *Trans. Vis. Comput. Graphics*, 18(9):1520–1536, Sept 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2011.279. URL `http://dx.doi.org/10.1109/TVCG.2011.279`.

[104] S. Lee, M. Sips, and H.-P. Seidel. Perceptually driven visibility optimization for categorical data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1746–1757, Oct 2013. ISSN 1077-2626. doi: 10.1109/TVCG.2012.315.

[105] M. Leiserson, H.-T. Wu, C. Gramazio, and B. Raphael. Magi: A platform for interactive visualization and collaborative annotation of combinations of genetic aberrations. In *The 1st Biological Data Science Meeting*, 2014.

[106] M. D. Leiserson, H.-T. Wu, C. C. Gramazio, and B. J. Raphael. Cancer genome analysis tool (cgat) for the visualization and exploration of combinations of mutations in cancer. In *The 4th Annual The Cancer Genome Atlas Symposium*, 2014.

[107] M. D. Leiserson, C. C. Gramazio, J. Hu, H.-T. Wu, D. H. Laidlaw, and B. J. Raphael. Magi: visualization and collaborative annotation of genomic aberrations. *Nature Methods*, 12(6): 483–484, 06 2015. URL `http://magi.brown.edu`.

[108] D. T. Levin, N. Momen, S. B. Drivdahl, and D. J. Simons. Change blindness blindness: The metacognitive error of overestimating change-detection ability. *Visual Cognition*, 7(1-3): 397–412, 2000. URL http://dx.doi.org/10.1080/135062800394865.

[109] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg. Comparative analysis of multidimensional, quantitative data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1027–1035, Nov 2010. ISSN 1077-2626. doi: 10.1109/TVCG.2010.138.

[110] S. Li, R. J. Crouser, G. Griffin, C. C. Gramazio, H.-J. Schulz, H. Childs, and R. Chang. Exploring hierarchical visualization designs using phylogenetic trees. *Proc. SPIE*, 9397:939709–939709–14, 2015. doi: 10.1117/12.2078857. URL http://dx.doi.org/10.1117/12.2078857.

[111] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. volume 32, pages 401–410. Blackwell Publishing Ltd, 2013. doi: 10.1111/cgf.12127. URL http://dx.doi.org/10.1111/cgf.12127.

[112] S. Lin, D. Ritchie, M. Fisher, and P. Hanrahan. Probabilistic color-by-numbers: Suggesting pattern colorizations using factor graphs. *ACM Trans. Graph.*, 32(4):37:1–37:12, July 2013. ISSN 0730-0301. doi: 10.1145/2461912.2461988. URL http://doi.acm.org/10.1145/2461912.2461988.

[113] Z. Liu and J. Stasko. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):999–1008, Nov 2010. ISSN 1077-2626. doi: 10.1109/TVCG.2010.177.

[114] Z. Liu, Y. Wang, M. Dontcheva, M. Hoffman, S. Walker, and A. Wilson. Patterns and sequences: Interactive exploration of clickstreamsto understand common visitor paths. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2016.2598797. URL http://dx.doi.org/10.1109/TVCG.2016.2598797.

[115] M. R. Luo, G. Cui, and B. Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26(5):340–350, 2001. ISSN 1520-6378. doi: 10.1002/col.1049. URL http://dx.doi.org/10.1002/col.1049.

[116] M. R. Luo, G. Cui, and C. Li. Uniform colour spaces based on ciecam02 colour appearance model. *Color Research & Application*, 31(4):320–330, 2006. ISSN 1520-6378. doi: 10.1002/col.20227. URL `http://dx.doi.org/10.1002/col.20227`.

[117] G. E. Marai. Visual Scaffolding in Integrated Spatial and Nonspatial Analysis. In E. Bertini and J. C. Roberts, editors, *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2015. doi: 10.2312/eurova.20151097.

[118] D. Martín-Albo, L. A. Leiva, J. Huang, and R. Plamondon. Strokes of insight: User intent detection and kinematic compression of mouse cursor trails. *Information Processing & Management*, 2016. doi: http://dx.doi.org/10.1016/j.ipm.2016.04.005. URL `http://dx.doi.org/10.1016/j.ipm.2016.04.005`.

[119] J. Matejka, T. Grossman, and G. Fitzmaurice. Patina: Dynamic heatmaps for visualizing application usage. In *Proc. of Human Factors in Computing Systems (CHI)*, pages 3227–3236, 2013. ISBN 978-1-4503-1899-0. doi: 10.1145/2470654.2466442. URL `http://doi.acm.org/10.1145/2470654.2466442`.

[120] Y. Matsuda. *Color Design*. Asakura Shoten, 1995.

[121] M. Matt, S. L. Su, C. Gramazio, C. Crumm, D. Extrum-Fernandez, and L. Cowen. Tuftsviewer: An intuitive interface for viewing 3d protein structures. In *3DSIG Workshop at ISMB*, 2010.

[122] B. A. Maxwell. Visualizing geographic classifications using color. *The Cartographic Journal*, 37(2):93–99, 2000. doi: 10.1179/0008704.37.2.p93. URL `http://www.tandfonline.com/doi/abs/10.1179/0008704.37.2.p93`.

[123] B. Meier, A. Spalter, and D. Karelitz. Interactive color palette tools. *IEEE Computer Graphics and Applications*, 24(3):64–72, May 2004. ISSN 0272-1716. doi: 10.1109/MCG.2004.1297012. URL `http://dx.doi.org/10.1109/MCG.2004.1297012`.

[124] B. J. Meier. Ace: A color expert system for user interface design. In *Proceedings of the ACM SIGGRAPH Symposium on User Interface Software*, UIST '88, pages 117–128, New York, NY, USA, 1988. ISBN 0-89791-283-7. doi: 10.1145/62402.62424. URL `http://doi.acm.org/10.1145/62402.62424`.

[125] X.-L. Meng, R. Rosenthal, and D. B. Rubin. Comparing correlated correlation coefficients. *Psychological bulletin*, 111(1):172, 1992.

[126] R. Munroe. Color survey results, May 2010. URL `http://blog.xkcd.com/2010/05/03/color-survey-results/`.

[127] A. H. Munsell. Van Nostrand Reinhold, New York, NY, USA, 1969 (1921).

[128] P. Murray, F. McGee, and A. Forbes. A taxonomy of visualization tasks for the analysis of biological pathway data. In *Proceedings of BioVis*, 2016.

[129] P. H. Nguyen, K. Xu, A. Wheat, B. W. Wong, S. Attfield, and B. Fields. Sensepath: Understanding the sensemaking process through analytic provenance. *Trans. Vis. Comput. Graphics*, 22(1):41–50, Jan 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467611. URL `http://dx.doi.org/10.1109/TVCG.2015.2467611`.

[130] T. O'Brien, A. Ritz, B. Raphael, and D. Laidlaw. Gremlin: An interactive visualization model for analyzing genomic rearrangements. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):918–926, Nov 2010. ISSN 1077-2626. doi: 10.1109/TVCG.2010.163.

[131] V. O'Day, A. Adler, A. Kuchinsky, and A. Bouch. When worlds collide: Molecular biology as interdisciplinary collaboration. In *Proceedings of the Seventh Conference on European Conference on Computer Supported Cooperative Work*, ECSCW'01, pages 399–418, Norwell, MA, USA, 2001. Kluwer Academic Publishers. ISBN 0-7923-7162-3. URL `http://dl.acm.org/citation.cfm?id=1241867.1241888`.

[132] P. O'Donovan, A. Agarwala, and A. Hertzmann. Color compatibility from large datasets. *ACM Trans. Graph.*, 30(4):63:1–63:12, 2011. ISSN 0730-0301. doi: 10.1145/2010324.1964958. URL `http://doi.acm.org/10.1145/2010324.1964958`.

[133] W. Ostwald. *Colour Science (Vol. II)*. Winsor and Newton, Ltd., London, UK, 1933.

[134] L.-C. Ou and M. R. Luo. A colour harmony model for two-colour combinations. *Color Research & Application*, 31(3):191–204, 2006. ISSN 1520-6378. doi: 10.1002/col.20208. URL `http://dx.doi.org/10.1002/col.20208`.

[135] L.-C. Ou, P. Chong, M. R. Luo, and C. Minchew. Additivity of colour harmony. *Color Research & Application*, 36(5):355–372, 2011. ISSN 1520-6378. doi: 10.1002/col.20624. URL http://dx.doi.org/10.1002/col.20624.

[136] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M. R. Speicher, J. Zschocke, and Z. Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 2013. doi: 10.1093/bib/bbs086. URL http://dx.doi.org/10.1093/bib/bbs086.

[137] J. Palmer. Attention in visual search: Distinguishing four causes of a set-size effect. *Current Directions in Psychological Science*, 4(4):118–123, 1995.

[138] S. E. Palmer and W. S. Griscom. Accounting for taste: Individual differences in preference for harmony. *Psychonomic Bulletin & Review*, 20(3):453–461, 2012. ISSN 1531-5320. doi: 10.3758/s13423-012-0355-2. URL http://dx.doi.org/10.3758/s13423-012-0355-2.

[139] S. E. Palmer and K. B. Schloss. An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, 107(19):8877–8882, 2010. doi: 10.1073/pnas.0906172107. URL http://dx.doi.org/10.1073/pnas.0906172107.

[140] A. Papoutsaki, H. Guo, D. Metaxa-Kakavouli, C. C. Gramazio, J. Rasley, W. Xie, G. Wang, and J. Huang. Crowdsourcing from scratch: A pragmatic experiment in data collection by novice requesters. *Proceedings of The AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2015.

[141] G. A. Pavlopoulos, D. Malliarakis, N. Papanikolaou, T. Theodosiou, A. J. Enright, and I. Iliopoulos. Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *GigaScience*, 4(1):1–27, 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0077-2. URL http://dx.doi.org/10.1186/s13742-015-0077-2.

[142] E. M. Peck, D. Afergan, and R. J. K. Jacob. Investigation of fnirs brain sensing as input to information filtering systems. In *Proceedings of the 4th Augmented Human International Conference*, AH '13, pages 142–149, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1904-1. doi: 10.1145/2459236.2459261. URL http://doi.acm.org/10.1145/2459236.2459261.

[143] E. M. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 473–482, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1899-0. doi: 10.1145/2470654.2470723. URL `http://doi.acm.org/10.1145/2470654.2470723`.

[144] J. Pitkow and P. Pirolli. Mining longest repeating subsequences to predict world wide web surfing. In *Proc. of the USENIX Symposium on Internet Technologies and Systems*, USITS'99, pages 13–13, 1999. URL `http://dl.acm.org/citation.cfm?id=1251480.1251493`.

[145] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, Jan 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467551.

[146] R. Raidou, U. van der Heide, C. Dinh, G. Ghobadi, J. Kallehauge, M. Breeuwer, and A. Vilanova. Visual analytics for the exploration of tumor tissue characterization. *Computer Graphics Forum*, 34(3):11–20, 2015. ISSN 1467-8659. doi: 10.1111/cgf.12613. URL `http://dx.doi.org/10.1111/cgf.12613`.

[147] R. Ratcliff. Methods for dealing with reaction time outliers. *Psychological bulletin*, 114(3): 510, 1993.

[148] R. Ratcliff. Methods for dealing with reaction time outliers. *Psychological bulletin*, 114(3): 510–532, 1993.

[149] K. Reinecke, D. R. Flatla, and C. Brooks. Enabling designers to foresee which colors users cannot see. In *ACM Human Factors in Computing Systems (CHI)*, pages 2693–2704, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858077. URL `http://doi.acm.org/10.1145/2858036.2858077`.

[150] J. C. Roberts. State of the art: Coordinated multiple views in exploratory visualization. In *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV '07. Fifth International Conference on*, pages 61–71, July 2007. doi: 10.1109/CMV.2007.20.

[151] J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nat Biotech*, 29(1):24–26, 01 2011. URL `http://dx.doi.org/10.1038/nbt.1754`.

[152] K. Rodden and X. Fu. Exploring how mouse movements relate to eye movements on web search results pages. *SIGIR Workshop on Web Information Seeking and Interaction*, pages 29–32.

[153] B. E. Rogowitz, L. A. Treinish, and S. Bryson. How not to lie with visualization. *Computers in Physics*, 10(3):268–273, 1996.

[154] M. Rønne Jakobsen and K. Hornbæk. Sizing up visualizations: Effects of display size in focus+context, overview+detail, and zooming interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1451–1460, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979156. URL `http://doi.acm.org/10.1145/1978942.1979156`.

[155] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *Journal of Vision*, 7(2), 2007. doi: 10.1167/7.2.17. URL `http://www.journalofvision.org/content/7/2/17.abstract`.

[156] R. Rosenholtz, N. R. Twarog, N. Schinkel-Bielefeld, and M. Wattenberg. An intuitive model of perceptual grouping for hci design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1331–1340, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518903. URL `http://doi.acm.org/10.1145/1518701.1518903`.

[157] M. A. Rotondi. *kappaSize: Sample Size Estimation Functions for Studies of Interobserver Agreement*, 2013. URL `http://CRAN.R-project.org/package=kappaSize`.

[158] N. Sadagopan and J. Li. Characterizing typical and atypical user sessions in clickstreams. In *Proc. of World Wide Web (WWW)*, pages 885–894, 2008. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367617. URL `http://doi.acm.org/10.1145/1367497.1367617`.

[159] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *Trans. Vis. Comput. Graphics*, 11(4):443–456, July 2005. ISSN 1077-2626. doi: 10.1109/TVCG.2005.53. URL `http://dx.doi.org/10.1109/TVCG.2005.53`.

[160] P. Saraiya, C. North, and K. Duca. Visualizing biological pathways: Requirements analysis, systems evaluation and research agenda. *Information Visualization*, 4(3):191–205, 2005. doi: 10.1057/palgrave.ivs.9500102. URL `http://ivi.sagepub.com/content/4/3/191.abstract`.

[161] P. Saraiya, C. North, V. Lam, and K. A. Duca. An insight-based longitudinal study of visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1511–1522, Nov 2006. ISSN 1077-2626. doi: 10.1109/TVCG.2006.85.

[162] A. Satyanarayan and J. Heer. Lyra: An interactive visualization design environment. *Computer Graphics Forum*, 33(3):351–360, 2014. ISSN 1467-8659. doi: 10.1111/cgf.12391. URL `http://dx.doi.org/10.1111/cgf.12391`.

[163] A. Satyanarayan, K. Wongsuphasawat, and J. Heer. Declarative interaction design for data visualization. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 669–678, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3069-5. doi: 10.1145/2642918.2647360. URL `http://doi.acm.org/10.1145/2642918.2647360`.

[164] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2017. URL `http://idl.cs.washington.edu/papers/vega-lite`.

[165] K. B. Schloss and S. E. Palmer. Aesthetic response to color combinations: preference, harmony, and similarity. *Attention, Perception, & Psychophysics*, 73(2):551–571, 2011. ISSN 1943-3921. doi: 10.3758/s13414-010-0027-0. URL `http://dx.doi.org/10.3758/s13414-010-0027-0`.

[166] K. B. Schloss, C. C. Gramazio, and C. Walmsley. Which color means more? an investigation of color-quantity mapping in data visualization. In *Journal of Vision/VSS*, volume 15, page 1317, September 2015. doi: 10.1167/15.12.1317. URL `http://dx.doi.org/10.1167/15.12.1317`.

[167] K. B. Schloss, D. Hawthorne-Madell, and S. E. Palmer. Ecological influences on individual differences in color preference. *Attention, Perception, & Psychophysics*, 77(8):2803–2816, 2015. ISSN 1943-393X. doi: 10.3758/s13414-015-0954-x. URL `http://dx.doi.org/10.3758/s13414-015-0954-x`.

[168] M. P. Schroeder, A. Gonzalez-Perez, and N. Lopez-Bigas. Visualizing multidimensional cancer genomics data. *Genome Medicine*, 5(1):1–13, 2013. ISSN 1756-994X. doi: 10.1186/gm413. URL `http://dx.doi.org/10.1186/gm413`.

[169] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Computer Graphics Forum*, 34(3):201–210, 2015. ISSN 1467-8659. doi: 10.1111/cgf.12632. URL `http://dx.doi.org/10.1111/cgf.12632`.

[170] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12): 2431–2440, Dec 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2012.213.

[171] V. Setlur and M. Stone. A linguistic approach to categorical color assignment for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):698–707, Jan 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467471.

[172] G. Sharma, W. Wu, and E. N. Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research and Applications*, 30(1), 2005. doi: 10.1002/col.20070. URL `http://dx.doi.org/10.1002/col.20070`.

[173] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. of Symposium on Visual Languages*, pages 336–343, Sep 1996. doi: 10.1109/VL.1996.545307. URL `http://dx.doi.org/10.1109/VL.1996.545307`.

[174] A. Silverman, C. C. Gramazio, and K. B. Schloss. The dark is more (dark+) bias in colormap data visualizations with legends. In *Journal of Vision/VSS*, 2016. doi: 10.1167/16.12.628. URL `http://dx.doi.org/10.1167/16.12.628`.

[175] J. Sjölund, C. Manetopoulos, M.-T. Stockhausen, and H. Axelson. The notch pathway in cancer: Differentiation gone awry. *European Journal of Cancer*, 41(17):2620 – 2629, 2005. ISSN 0959-8049. doi: http://dx.doi.org/10.1016/j.ejca.2005.06.025. URL `http://www.sciencedirect.com/science/article/pii/S0959804905007276`.

[176] D. Skau and R. Kosara. Arcs, angles, or areas: Individual data encodings in pie and donut

charts. *Computer Graphics Forum*, 35(3):121–130, 2016. ISSN 1467-8659. doi: 10.1111/cgf.12888. URL `http://dx.doi.org/10.1111/cgf.12888`.

[177] M. Stone. In color perception, size matters. *IEEE Computer Graphics and Applications*, 32 (2):8–13, March 2012. ISSN 0272-1716. doi: 10.1109/MCG.2012.37.

[178] M. Stone. In color perception, size matters. *IEEE Computer Graphics and Applications*, 32(2): 8–13, 2012. ISSN 0272-1716. doi: http://doi.ieeecomputersociety.org/10.1109/MCG.2012.37.

[179] M. Stone, D. A. Szafir, and V. Setlur. An engineering model for color difference as a function of size. In *Proceedings of the Color and Imaging Conference*, pages 228–233, 2014.

[180] M. Strait, C. Gramazio, J. Park, S. L. Su, and L. Cowen. Moleint: Reducing workload through adaptive interaction. In *VIZBI*, 2012.

[181] M. Streit, A. Lex, M. Kalkusch, K. Zatloukal, and D. Schmalstieg. Caleydo: Connecting pathways and gene expression. *Bioinformatics*, 25(20):2760–2761, 2009. doi: 10.1093/bioinformatics/btp432.

[182] M. Streit, H. J. Schulz, A. Lex, D. Schmalstieg, and H. Schumann. Model-driven design for the visual analysis of heterogeneous data. *IEEE Transactions on Visualization and Computer Graphics*, 18(6):998–1010, June 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2011.108.

[183] S. Su, C. Gramazio, D. Extrum-Fernandez, C. Crumm, L. Cowen, M. Menke, and M. Strait. Molli: Interactive visualization for exploratory protein analysis. *Computer Graphics and Applications, IEEE*, 32(5):62–69, Sept 2012. ISSN 0272-1716. doi: 10.1109/MCG.2012.66.

[184] F. Szabó, P. Bodrogi, and J. Schanda. Experimental modeling of colour harmony. *Color Research & Application*, 35(1):34–49, 2010. ISSN 1520-6378. doi: 10.1002/col.20558. URL `http://dx.doi.org/10.1002/col.20558`.

[185] J. Talbot, J. Gerth, and P. Hanrahan. An empirical model of slope ratio comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2613–2620, Dec 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2012.196.

[186] A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind. Visual quality metrics and human perception: An initial study on 2d projections of large multidimensional data. In *Proceedings*

of the International Conference on Advanced Visual Interfaces, AVI '10, pages 49–56, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0076-6. doi: 10.1145/1842993.1843002. URL `http://doi.acm.org/10.1145/1842993.1843002`.

[187] C. Taylor and A. Franklin. The relationship between color–object associations and color preference: Further investigation of ecological valence theory. *Psychonomic Bulletin & Review*, 19(2):190–197, 2012. ISSN 1531-5320. doi: 10.3758/s13423-012-0222-1. URL `http://dx.doi.org/10.3758/s13423-012-0222-1`.

[188] H. M. Taylor and S. Karlin. *An introduction to stochastic modeling*. Academic press, 2014.

[189] P. Thébault, R. Bourqui, W. Benchimol, C. Gaspin, P. Sirand-Pugnet, R. Uricaru, and I. Dutour. Advantages of mixing bioinformatics and visualization approaches for analyzing srna-mediated regulatory bacterial networks. *Briefings in Bioinformatics*, 16(5):795–805, 2015. doi: 10.1093/bib/bbu045. URL `http://bib.oxfordjournals.org/content/16/5/795.abstract`.

[190] A. Thudt, U. Hinrichs, and S. Carpendale. A modular approach to promote creativity and inspiration in search. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, C&#38;C '15, pages 245–254, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3598-0. doi: 10.1145/2757226.2757253. URL `http://doi.acm.org/10.1145/2757226.2757253`.

[191] S. Tilkov and S. Vinoski. Node.js: Using javascript to build high-performance network programs. *IEEE Internet Computing*, 14(6):80–83, 2010. ISSN 1089-7801. doi: http://doi.ieeecomputersociety.org/10.1109/MIC.2010.145.

[192] C. Tominski, G. Fuchs, and H. Schumann. Task-driven color coding. In *Information Visualisation, 2008. IV '08. 12th International Conference*, pages 373–380, July 2008. doi: 10.1109/IV.2008.24.

[193] A. Treisman. Features and objects in visual processing. *Scientific American*, 255(5):114–125, 1986.

[194] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[195] A. Tversky and D. Kahneman. *Judgment under Uncertainty: Heuristics and Biases*, pages 141–162. Springer Netherlands, Dordrecht, 1975. ISBN 978-94-010-1834-0. doi: 10.1007/978-94-010-1834-0_8. URL `http://dx.doi.org/10.1007/978-94-010-1834-0_8`.

[196] F. van Ham and A. Perer. “search, show context, expand on demand”: Supporting large graph exploration with degree-of-interest. *Trans. Vis. Comput. Graphics*, 15(6):953–960, Nov 2009. ISSN 1077-2626. doi: 10.1109/TVCG.2009.108. URL `http://dx.doi.org/10.1109/TVCG.2009.108`.

[197] F. Vandin, E. Upfal, and B. J. Raphael. De novo discovery of mutated driver pathways in cancer. *Genome Research*, 22(2):375–385, 2012. doi: 10.1101/gr.120477.111. URL `http://genome.cshlp.org/content/22/2/375.abstract`.

[198] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013. ISSN 0036-8075. doi: 10.1126/science.1235122. URL `http://science.sciencemag.org/content/339/6127/1546`.

[199] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao. Unsupervised clickstream clustering for user behavior analysis. In *Proc. of Human Factors in Computing Systems (CHI)*, pages 225–236, 2016. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858107. URL `http://doi.acm.org/10.1145/2858036.2858107`.

[200] M. Wertheimer. Laws of organization in perceptual forms. *A source book of Gestalt psychology*, pages 71–88, 1938.

[201] D. Wixon, K. Holtzblatt, and S. Knox. Contextual design: An emergent view of system design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’90, pages 329–336, New York, NY, USA, 1990. ACM. ISBN 0-201-50932-6. doi: 10.1145/97243.97304. URL `http://doi.acm.org/10.1145/97243.97304`.

[202] J. M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994. ISSN 1069-9384. doi: 10.3758/BF03200774. URL `http://dx.doi.org/10.3758/BF03200774`.

[203] K. Wongsuphasawat and D. Gotz. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *Trans. Vis. Comput. Graphics*, 18(12):2659–2668,

Dec 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2012.225. URL http://dx.doi.org/10.1109/TVCG.2012.225.

[204] K. Wongsuphasawat and J. Lin. Using visualizations to monitor changes and harvest insights from a global-scale logging infrastructure at twitter. In *Proc. of Visual analytics, science, and technology (VAST)*, pages 113–122, Oct 2014. doi: 10.1109/VAST.2014.7042487. URL http://dx.doi.org/10.1109/VAST.2014.7042487.

[205] K. Yokosawa, K. B. Schloss, M. Asano, and S. E. Palmer. Cross-cultural studies of color preferences: Us and japan. *Cognitive Science*, 2015.

[206] B. Yost and C. North. The perceptual scalability of visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):837–844, Sept 2006. ISSN 1077-2626. doi: 10.1109/TVCG.2006.184.

[207] A. Zeileis, K. Hornik, and P. Murrell. Escaping rgbland: Selecting colors for statistical graphics. *Computational Statistics and Data Analysis*, 53(9):3259 – 3270, 2009. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2008.11.033. URL http://dx.doi.org/10.1016/j.csda.2008.11.033.

[208] E. Zgraggen, S. M. Drucker, D. Fisher, and R. DeLine. (s|qu)eries: Visual regular expressions for querying and exploring event sequences. In *Proc. of Human Factors in Computing Systems (CHI)*, pages 2683–2692, 2015. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702262. URL http://doi.acm.org/10.1145/2702123.2702262.

[209] C. Ziemkiewicz, S. Gomez, and D. Laidlaw. Analysis within and between graphs: Observed user strategies in immunobiology visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1655–1658, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2208291. URL http://doi.acm.org/10.1145/2207676.2208291.