Navigating Reductionism and Holism in Evaluation

Michael Correll
University of
Wisconsin-Madison
mcorrell@cs.wisc.edu

Eric Alexander
University of
Wisconsin-Madison
ealexand@cs.wisc.edu

Danielle Albers
University of
Wisconsin-Madison
dalbers@cs.wisc.edu

Alper Sarikaya
University of
Wisconsin-Madison
sarikaya@cs.wisc.edu

Michael Gleicher
University of
Wisconsin-Madison
gleicher@cs.wisc.edu

ABSTRACT

In this position paper, we enumerate two approaches to the evaluation of visualizations which are associated with two approaches to knowledge formation in science: reductionism, which holds that the understanding of complex phenomena is based on the understanding of simpler components; and holism, which states that complex phenomena have characteristics more than the sum of their parts and must be understood as complete, irreducible units. While we believe that each approach has benefits for evaluating visualizations, we claim that strict adherence to one perspective or the other can make it difficult to generate a full evaluative picture of visualization tools and techniques. We argue for movement between and among these perspectives in order to generate knowledge that is both *grounded* (i.e. its constituent parts work) and *validated* (i.e. the whole operates correctly). We conclude with examples of techniques which we believe represent movements of this sort from our own work, highlighting areas where we have both "built up" reductionist techniques into larger contexts, and "broken down" holistic techniques to create generalizable knowledge.

1. INTRODUCTION

There is a tension among practitioners of information visualization. We wish to be good engineers and build things that are solidly constructed and useful: to this end we want to know that a particular visualization *works*. We also wish, as scientists, to advance visualization as a field and to know how to build good tools in the future: we want to know *how* things work. A tightly related tension appears in evaluative methodology as a contrast between reductionist and holistic approaches to evaluation—the former seeking to build up understanding of visualizations by analyzing their component pieces, and the latter examining specific visualizations as indivisible units. We believe that evaluating too close

to an entirely reductionist or entirely holistic methodology makes it difficult to answer both the "what" and "how" of visualizations. Doing so requires that we both ground our design in theoretical principles as well as validate that it performs well in a real setting, with real users and real data. It is our position that visualization practitioners should employ evaluative techniques that allow movement *between* holism and reductionism, melding the benefits of both approaches, either by conducting multiple evaluations from different perspectives, or modifying individual evaluations to take into account new perspectives that would otherwise be overlooked.

In this paper, we will briefly discuss what we mean by "holistic" and "reductionist" evaluation methodologies, their benefits and drawbacks, and techniques for movement between these extremes. Generally, since there are a wide variety of reasons for performing evaluation [11, 13], it is critical to tailor evaluation techniques to generate relevant, extensible knowledge. In this vein, we present examples from our own work in which we have employed evaluation techniques that move both "upwards" (from reductionist to holistic settings) and "downwards" (from holistic to reductionist settings).This movement can take many forms: as "stepping stones" where we build *multiple* related experiments with gradually increasing or decreasing scope; or by designing an *individual* experiment to find a "sweet spot" between the contextualized whole and the controllable parts. We believe both directions of movement have utility. By making reductionist evaluations more holistic, we can give context to our results, and/or test for unexpected interactions between our designs and the real world environments in which they are deployed. By making holistic evaluations more reductionist, we can make reasonable, evidence-driven arguments for why to choose one sort of design over another and diagnose potential confounds before complexity is added.

Other perspectives on evaluation have presented methods which combine low- and high-level evaluations. These different methodologies are usually presented as discrete steps in a pipeline, oriented towards the deployment and assessment of a single tool [14, 17]. We argue that concern with the level of evaluation, and focusing on navigation between reductionist and holistic strategies, is relevant even if the end goal is not the deployment of a specific tool. For instance, we might want to assess a tool that has already been deployed. To do so, we might need to de-construct holistic evaluations to examine the components of the existing system to generate

a more complete picture of performance. Alternatively, we might want to examine individual components of a system (e.g. particular encodings, visual tasks) to inform future designs. To make studies of these sort more widely useful, we ought to make sure our evaluations are generalizable enough to apply to real-world visualization tasks.

In practice, no evaluation is entirely reductionist or holistic in methodology. We are often mindful of an intended whole when assessing even very low-level components of visualizations; likewise, the analysis of visualization tools *in situ* is often structured in modular fashion so that specific subcomponents are under scrutiny. In the following sections, we present our definitions of what holism and reductionism mean for evaluation, not as a complete description of the space, but to present (somewhat idealized) examples of what these terms mean in an idealized setting. We then highlight our experiences with—and describe the benefits of—navigating the space within and between both approaches.

2. HOLISM

Holism as a philosophy is the notion that the whole is greater than the sum of its parts. For evaluation it is embodied by belief that visualizations are best evaluated as complete, deployed systems, which are given to a specific audience with a specific type of data (even a specific data set). Many questions can be answered using holistic evaluation. While true holistic evaluation at its extreme is an ideal—evaluating all scenarios addressed by a system in a meaningful way is infeasible—the question of utility is well-supported in holistic evaluation in practice [18]: is the value of the visualization higher than the cost? Did we profit (e.g. monetarily, in terms of utility, or number of citations) from the design and deployment of the tool? Did the viewers of the tool gain significant insights [15] while exploring the visualization? Were viewers engaged with the visualization through its aesthetics, and did our design decisions really work?

Holistic evaluations are therefore an excellent tool for the *validation* of visualizations. This frequently means that the knowledge from whole-tool evaluations in real environments is summative, rather than formative; we can know that a tool has succeeded or failed, and perhaps even what has gone wrong (and so what to avoid in the future), but by conducting evaluations by presenting entire tools to real populations we make it difficult to gain formative knowledge about mistakes that have been made at steps prior to the prototyping stage, or at granularities lower than entire prototype. It is also possible to conduct holistic evaluations that are rigorous and complete but still somewhat unsatisfying. If the intended audience for our visualization is small (perhaps the domain is esoteric enough that our collaborators are the entire population of people with the skills to interpret the data, or perhaps the data is confidential and so cannot be shared with a wider pool), it is difficult to claim that our small sample size of collected data describes generalizable facts about visualization practice. We believe that this is true of holistic visualization generally: as we add more context to the evaluations of the visualizations we create (both in terms of the problem context, and the context of visual encodings within an environment) we gain knowledge about how we believe our tool will perform in a desired setting, but at the expense of creating generalizable knowledge for designers in different domains, or designers seeking to improve on other tools.

3. REDUCTIONISM

Reductionism underpins much of philosophy of science, and seeks to explain complex phenomena by examining their individual components. In visualization evaluations, this mindset can be seen as a desire to examine individual components of viewer behavior and use those to make inferences about the whole. Many "time and error" evaluations are at least partly reductionist in construction. For instance, an evaluation may seek to identify which particular encoding allows a viewer to complete a particular task in the fastest time, with highest accuracy. By conducting a sufficient number of these studies, the experimenter hopes to formulate a general, predictive model of the behavior of a particular facet of their design. While reductionism in its extreme is embodied in traditional psychophysics evaluations that are devoid of data and problem contexts, reductionist evaluations have been used in visualization to aid in understanding in many common scenarios: we can compare different encoding channels for information across common tasks (as in the canonical studies of Cleveland & McGill [7, 8]), we can build predictive models of performance, and we can collect information that is not contingent on the operation of an entire, interactive system.

Reductionist evaluations are therefore good for *grounding* visualizations. This grounding is frequently formative, rather than summative; we know which components of a visualization we except to succeed or fail, but this might not be predictive of the tool as a whole. We can examine individual design decisions (e.g. using shape to encode a value rather than position), compare them to alternatives, collect data, and make an empirically grounded decision (or explanation) for one choice over another. However, this desire to assess at the component level, and thus control for irrelevant factors, means that evaluative results can be devoid of context. A design design that seems reasonable from a reductionist standpoint might be a clear point of failure in a deployed tool: perhaps the problem domain is eccentric, and accustomed to data presented in a particular way that you did not anticipate; perhaps each component part of your design functions as expected, but they have unexpected interactions when combined into a deployed system. In general, we maintain that as reductionist techniques attempt to control for more and more components, they become less and less adept at validating particular real-world systems.

4. BRIDGING THE GAP

We believe that, despite the dichotomous appearance of the two perspectives presented above, visualization evaluation in practice has elements of both reductionist and holistic reasoning—evaluations *break down* design into component techniques and focus on more general tasks as they move towards reductionism and *build up* knowledge about more complex design techniques and high-level domain tasks as they move towards holism. We believe that restructuring, adapting, and modifying evaluative designs to move around in this space can better balance the trade-offs of holism and reductionism and will result in evaluations which are better at both grounding and validating designs.

Here, we present examples from our own work that begins to navigate the space between holism and reductionism. These examples are drawn from three domain areas. Our work with **sequences** represents an initial collaboration with geneticists who wanted to conduct comparisons between large numbers of genetic sequences aligned at the level of genes. Our work with **text** was a collaboration with humanities researchers interested in visualizing trends across a corpus of early-modern literature. We also present work with **proteins**, a collaboration with structural biologists interested in visualizing protein-ligand interactions across molecular surfaces. In our attempts to both validate the tools we built in these collaborations as well as ground them in theory that would inform future design, we have navigated within holistic and reductionist perspectives—both by making holistic techniques more reductionist as well as vice versa.

4.1 Sequence Visualization

Our collaborators in genetics wanted to conduct comparisons between large numbers of genetic sequences aligned at the level of genes. Existing solutions supported comparison across a handful of unique sequences, but their datasets had dramatically outgrown what their existing solutions supported. We analyzed existing designs based on grounded reductionist findings from perceptual psychology to hypothesize how to overcome limits in these designs. We used these findings to build up novel visualization solutions that addressed our collaborators specific needs, culminating in the aggregate colorfield designs used in Sequence Surveyor [2].

The resulting design supported our collaborators' analysis goals, but blended design elements motivated by isolated reductionist findings in novel ways. The specificity of the underlying domain science made a formal quantitative evaluation of these new design components infeasible: our collaborators were only able to reason about their own datasets, leaving the effectiveness of our techniques for other visualization problems unevaluated. We thus had two sorts of problems relating to holism and reductionism: we needed to *break down* our completed design into design elements we believe would transfer to other visualization problems, and we needed to *build up* from our low-level design assumptions drawn from psychology to confirm that our assumptions would hold in complex visualization systems.

To address these issues, we reduced the system design to its underlying techniques and applied them to model problems from a more common domain (time series data) to better understand their effectiveness for different visualization tasks [1, 9]. This new domain provided tasks that were roughly analogous to our original domain (e.g. identifying regions with high average values versus regions with lots of frequently occurring genes), and also allowed us to evaluate the component techniques of our blended design under a realistic visualization scenario both across a broader population and for a variety of more general tasks.

These studies represent some of our early efforts at moving around the holism-reductionism space: we began by evaluating how well reductionist findings from psychology could inform encoding design for a single aggregate task [9], and then took a second step towards holism by exploring interactions between these designs across a larger space of tasks.

Through these initial efforts, we gained a better understanding of our high-level system design, as well as the relationship between these novel component techniques and low-level visualization tasks. We have used this knowledge to make informed decisions about how these techniques can be used to address other problems (§4.3). While these findings help to contextualize the highly reductionist perceptual results that motivated the original design, we have yet to fully bridge the gap between these evaluations and the high-level design.

4.2 Text Visualization

In a different collaboration, we worked with humanities researchers interested in visualizing trends across a corpus of early-modern literature. Through ethnographic observations of their research and workflow, we identified a number of design principles, including the necessity of supporting analysis down to individual passages—bridging the unfamiliar task of analyzing high-level statistical overviews with the more familiar one of performing close reading. We turned to tagged text (using colored highlighting to indicate information about individual words) as a useful encoding for augmenting their reading with information from the model.

To evaluate how well tagged text would address our collaborators' needs at this scale, we turned to previous experiments in perceptual psychology to provide us with principles to build on, specifically to confirm that people are capable of making efficient numerosity estimations on small number of objects [12]. However, it was unclear how well these findings would generalize to tagged text, as words are different in shape and semantic content than more simple stimuli like dots on a scatterplot. We performed a set of experiments to confirm that numerosity estimates of tagged text was feasible for human readers, indicate areas of difficulty, and provide encodings to address these difficulties [10].

We have since used tagged text encodings in multiple tools by generalizing these low-level findings to build up more complex systems [4, 10]. Each of these tools has required independent holistic evaluation to show that there is worthwhile insight to be derived from our tagged text encodings in context, and we are still developing a more holistic understanding of our current solutions. However, our experiments helped ground the design of these high-level systems with principles that we have already been able to generalize.

4.3 Protein and Surface Visualization

Collaboration with structural biologists led us to the domain of molecular visualization, specifically visualizing protein-ligand interactions [5, 6]. We made several rendering decisions for visualizing molecular surfaces that utilized surface shading to convey molecular shape, but soon realized that this shading could confound judgments of data if the data was conveyed via color on the surface—colors in shadow had a darker pixel value than the actual encoding color.

We conducted a series of reductionist studies to understand this issue by exploring lightness constancy in molecular visualization [3], specifically, evaluating whether a viewer can recover the 'actual' color value of data encoded on a surface even if the surface as been attenuated by a simulated shadow. In these studies, we also took intiial steps toward holism by exploring how different rendering decisions (us-

ing stylized contours, [non-]existence of diffuse light cues, etc.) and combinations of these decisions affect the viewer's ability to perceive the 'original' encoded color. This served to *break down* our previous design decisions into general, verifiable rules for promoting lightness constancy in visualization of three-dimensional objects. This lower-level study allows for incremental development of a visualization technique that is decoupled from the data domain, ready to be applied to related contexts.

We used these results in conjunction with our studies from §4.1 to *build up* a system for exploring corpra of machine learning data about these protein-ligand interactions. We worked with our collaborators to understand their task requirements and used both this work and our summary encoding work [9] to construct a system to address their specific needs. We used the constancy studies to inform our decisions of how to encode data using color on a surface and the summary encoding work as a basis for the design of an overview for hundreds of molecules simultaneously in our subsequent protein-ligand machine learning visualization [16]. This approach blended results from our initial reductionist evaluations to create a holistic design aimed at resolving a specific domain problem in context.

5. CONCLUSION

In this position paper, we have shared our thoughts and experiences on adapting evaluations to take advantage of both holistic and reductionist reasoning, techniques, and epistemologies. We believe that both perspectives offer utility in different areas of knowledge generation, but that we typically want knowledge that is most easily gained by incorporating *both* aspects. For instance, holistic evaluation makes it easier for us to confirm the overall utility of our visualization tool in real-world settings, but reductionist evaluation makes it easier for us to guide the design process with respect to specific design components.

We believe that there is further promising work to be done in the enumeration of other techniques for movement in this space, both upwards (towards holism) and downwards (towards reductionism). This work is especially important as we believe that in practice, designers often combine aspects of both approaches, consciously or not, and understanding this movement can help to elucidate the trade-offs between different evaluative approaches. We believe that future productive work can be done by shelving the false dichotomy of reductionism and holism and mapping out the territory between them descriptively rather than prescriptively, tailoring evaluation strategies toward what we hope to know and how we hope to share this knowledge with the field.

Acknowledgements

6. REFERENCES

[1] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proc. ACM Conference on Human Factors in Computing Systems*, pages 551–560. ACM, 2014.

[2] D. Albers, C. Dewey, and M. Gleicher. Sequence Surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE TVCG*, 17(12):2392 – 2401, Dec 2011.

[3] D. Albers, A. Sarikaya, and M. Gleicher. Lightness constancy in surface visualization. In *IEEE Visualization Poster Proceedings*. IEEE, Oct 2013.

[4] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *Proc. IEEE Conference on Visual Analytics Science and Technology*, 2014.

[5] G. Cipriano and M. Gleicher. Molecular surface abstraction. *IEEE TVCG*, 13(5):1608–1615, Oct 2007.

[6] G. Cipriano, G. Phillips Jr., and M. Gleicher. Local functional descriptors for surface comparison based binding prediction. *BMC Bioinf.*, 13(1):314–336, 2012.

[7] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Amer. Statist. Assoc.*, 79(387):531–554, 1984.

[8] W. S. Cleveland and R. McGill. An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25(5):491–500, 1986.

[9] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proc. ACM Conference on Human Factors in Computing Systems*, pages 1095–1104. ACM, 2012.

[10] M. Correll, E. Alexander, and M. Gleicher. Quantity estimation in visualizations of tagged text. In *Proc. ACM Conference on Human Factors in Computing Systems*, pages 2697–2706. ACM, 2013.

[11] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Moller. A systematic review on the practice of evaluating visualization. *IEEE TVCG*, 19(12):2818–2827, 2013.

[12] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann. The discrimination of visual number. *Am. J. Commun. Psychol.*, 62(4):pp. 498–525, 1949.

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

[14] M. Meyer, M. Sedlmair, and T. Munzner. The four-level nested model revisited: blocks and guidelines. In *Proc. BELIV Workshop: Beyond Time and Errors—Novel Evaluation Methods for Visualization*, pages 11–17. ACM, 2012.

[15] C. North. Toward measuring visualization insight. *Computer Graphics and Applications, IEEE*, 26(3):6–9, 2006.

[16] A. Sarikaya, D. Albers, J. C. Mitchell, and M. Gleicher. Visualizing validation of protein surface classifiers. *Computer Graphics Forum*, 33(3):171–180, Jun 2014.

[17] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proc. AVI workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, number 8. ACM, 2006.

[18] J. J. Van Wijk. The value of visualization. In *IEEE Visualization*, pages 79–86. IEEE, 2005.