Algebraic Visualization Design for Pedagogy

Gordon Kindlmann*
University of Chicago

Carlos Scheidegger†
University of Arizona

ABSTRACT

We report on successes and challenges in using our recently-proposed Algebraic Visualization Design as the basis for teaching visualization. The experiences suggest that the principles in that paper can serve as unifying framework for introducing the fundamentals of data visualization, as long as the overtly mathematical flavor of the framework is not a disincentive. We find that the better we can relate the framework to ideas to which students have already been exposed, the better they understand it.

1 INTRODUCTION

1.1 Summary of Algebraic Visualization Design

In an InfoVis 2014 paper [9], we proposed a abstract mathematical way to describe problems and properties of visualizations, called Algebraic Visualization Design (AVD), reviewed here.

Our model describes relationships between three elements of visualization: the space D of "data" to be visualized, the space R of data representations, and the space V of visualizations. While "data" normally refers to the information input to a visualization algorithm, we distinguish between the underlying object of interest (e.g. a set of numbers), and its actual representation on a computer (e.g. an array in memory). The mappings between these spaces are captured in a single equation, shown with its commutative diagram:

$$v \circ r_2 \circ \alpha = \omega \circ v \circ r_1 \qquad \begin{array}{ccccc} D & \xrightarrow{r_1} & R & \xrightarrow{v} & V \\ \alpha \downarrow & & & & \downarrow \omega \\ D & \xrightarrow{r_2} & R & \xrightarrow{v} & V \end{array} \tag{1}$$

Lowercase r maps from data D to representation R, and lowercase v is the visualization method mapping from R to visual stimulus V. The mappings from D back to D (or mappings on D), are termed *data symmetries*, denoted α. The *visualization symmetries*, denoted ω, are mappings on V. The identity maps on D and V, which send each element to itself, are 1_D and 1_V, respectively.

Using alpha and omega emphasizes that we are interested in the relationship between the very beginning and the very end of the visualization pipeline, while seeking to abstract away as much as possible of the internal details. In a commutative diagram, when two nodes are connected by two paths, the composition of functions along either path must be the same. The equality in (1) is between two possible paths from the upper-left D to the lower-right V. Going down then right is $D \underset{\alpha}{\rightarrow} D \underset{r_2}{\rightarrow} R \underset{v}{\rightarrow} V$, or in terms of function composition (read right to left) $v \circ r_2 \circ \alpha$, and going right then down is $D \underset{r_1}{\rightarrow} R \underset{v}{\rightarrow} V \underset{\omega}{\rightarrow} V$ (or $\omega \circ v \circ r_1$).

AVD says that in successful visualizations, important changes in the data α are well-matched, via (1), with obvious visual changes ω. The α are answers to the question, "If the world had been different in some interesting and important way, how would the data have been different?" If one has data for which visualization may provide insight, the significance of that insight hinges on knowing, at a general level, what phenomena or relationships may be captured by the data. The α embody the low-level tasks that the visualization is designed for. These α describe which structures from the data/operation abstraction layer will be reflected in the encoding/interaction technique layer [12]. The α are not a property of the data itself; they instead capture the kinds of questions about the data that we want the visualization to answer. For the same data, different users may have different α.

Obvious visual changes ω are aligned with elementary perceptual tasks, such as judgments of position along a common scale, or of length and direction [3]. When possible, the ω generated by the chosen α should coincide with perceptually pre-attentive channels [7]. However, applying AVD assumes more than just identifying perceptual channels: it connects to the mathematical structure of those channels. In color perception, for example, opponent color theory [19, 17] endows the space of colors with a kind of negation (green is "negative" red, orange is "negative" blue). Thus, when designing color scales for data in which negation is a meaningful data change α, one can evaluate a color scale by seeing whether the corresponding visual change ω aligns with finding opponent hues.

From (1), AVD derives three principles, and gives concrete names to ways in which a visualization may fail the principles, summarized in Table 1. Calling these "principles" may suggest that these are tools of judgment, but we intend them to be more descriptive than prescriptive. The **Principle of Representation Invariance** says that from the same dataset, different visualizations will not arise simply by changing the representation of the same underlying data. If this is not the case, there is some mapping on representations, called the *hallucinator*, with a non-trivial effect on the visualization. This notion of invariance is rooted in Stevens's statement of invariantive statistics, which gave specific mathematical meaning to categorical, ordinal, interval, and ratio data scales [13]. Conversely, the **Unambiguous Data Depiction Principle** says that an interesting α applied to the data should induce a non-trivial ω. If this is not the case, then there is an injectivity failure [20], and we give a name to the α for which $\omega = 1_V$: a *confuser*. Visualization design requires trade-offs; designers may accept some confusers (data changes that are invisible) when they do not relate to the intended analysis tasks, to ensure that the α are shown clearly. Finally, assuming both Invariance and Unambiguity hold, the **Correspondence Principle** is satisfied when *neither* α nor ω is the identity, and they solve (1) in a particular way, notated "$\alpha \cong \omega$" to suggest *congruence* [15] between the data and visualization symmetries. The Correspondence Principle says that ω somehow makes sense, given α. The visualization viewer would then be able to infer the α from ω, or, the visualization makes the α *legible*. In the setting of metric spaces on both data and visual perception, Correspondence is closely related to the notion of visual embedding [**?**]. When an important α is not legible because the associated ω is hard to understand, then the visualization as a *jumbler*. Conversely, when the apparent structure of the visualization suggests some obvious visual symmetries ω, those should correspond to important data symmetries α. If not, the visualization has a *misleader*. Diverging color scales for ratio data, for example, tend to satisfy the Correspondence Principle by mapping negation of data to negation of color hue. Considering visualizations not just as functions mapping from data to visual stimuli, but as functors mapping from *changes* in data to *changes* in the visual, and then capturing properties of the visualization that may be problematic in

*e-mail: glk@uchicago.edu
†e-mail:cscheid@email.arizona.edu

Principle Name	Precondition	Requirement	Name for failure	Failure definition
Representation Invariance	$\alpha = 1_D$	$\omega = 1_V$	Hallucinator	$H(v) = \{h \mid r_2 = h \circ r_1 \text{ and } v \circ h \neq v\}$, over all representations r_1, r_2
Unambiguous Data Depiction	$\omega = 1_V$	$\alpha = 1_D$	Confuser	$C(v) = \{\alpha \mid v \circ r \circ \alpha = v \circ r\}$
Visual-Data Correspondence	$\alpha \neq 1_D$, $\omega \neq 1_V$	$\alpha \cong \omega$	Jumblers and Misleaders	(see Sec. 1 text)

Table 1: Algebraic visualization design principles for evaluating a visualization method v, expressed in terms of (1).

terms of those functors (Table 1), is the basis for the "algebraic" in algebraic visualization design.

Our hope, however, is that no prior understanding of functors or algebra is necessary to use AVD for visualization practice or pedagogy. General rules of thumb like Tufte's "show data variation, not design variation" [14] become more actionable when, faced with a new or unsatisfactory visualization, one has a concrete course of action for evaluating or improving the visualization. AVD empowers students to ask pointed questions about what a visualization does or does not show, and AVD supplies a terminology for the problems revealed, which helps focus subsequent discussion and redesign. If the data were different, would the visualization be different (Unambiguous), and, different in an informative way (Correspondence)? If it is ambiguous: what are the data changes we are blind to (Confuser)? If it is not informative: how else can we lay out or encode the data to create a better correspondence (removing Jumblers)? Are there apparent properties in the visualization that are not actually in the data (Misleaders)? Could the visualization have ended up looking different, due only to changes (Hallucinators) in the computational/numerical representation that should be inconsequential, but are not (Invariance)?

1.2 "It was my understanding there would be no math"

Often, data visualization students will not be computer science majors, and even CS students will have relatively little mathematical training. We personally value the rigor that the underlying theory provides, but it is natural to wonder whether the notation behind AVD is worth the price. In our experience, the most important consequence of using AVD in the classroom is that it forces students to think about the interplay between α and ω. The central insight is the notion that every change in the data — every possible different world — corresponds to a possible change in the visualization. If the change in the data does not change the visualization, this is also important information.

In other words, the mathematics are there to ensure us that we can make the notions precise. But no math is needed to use AVD while thinking about design: we imagine possible different worlds where the data were different in an interesting way, and think about what that change *does* to the visualization.

2 Successes and Challenges

We give some examples of how we have used AVD to successfully explain some principles of visualization design in our classes, and highlight things that have repeatedly arisen as points of confusion.

2.1 Banking to 45

The "Banking to 45" rule of thumb [18] for determining the aspect ratio of a 2D function plot is a simple introductory example of what can be directly derived from the Correspondence Principle. We start with the recognition that the purpose of the plot may not be just to show values of a function $y = f(x)$, but also its derivative $m = f' = dy/dx$. This leads to considering a particular α at some point on the graph: $\alpha(m) = m \pm \Delta m$. This asserts that small changes in slope are interesting and important to see. The corresponding ω will be some local change in the slope of the plot. The slope (rise over run) of a line is not an elementary perceptual channel, but line orientation is [3]. "Banking to 45" answers how to map the chosen α to a ω that on average is as clear as possible.

Prior to any visualization, the tangent to $y = f(x)$ has slope $m = \Delta y/\Delta x$. Once plotted on the page with vertical v and horizontal h coordinates, the slope is $\Delta v/\Delta h = Rm$, where R is determined by the aspect ratio of the plot. Viewers perceive the line orientation $\theta = \tan^{-1}(Rm)$. The Correspondence Principle tells us to maximize the change ω in the orientation for a given change α in slope, i.e. find the R that maximizes $\frac{d\theta}{dm} = \frac{d}{dm}\tan^{-1}(Rm) = \frac{r}{1+m^2r^2}$. Using calculus, students solve $\frac{d}{dR}\frac{d\theta}{dm} = 0$ and find that $R = \pm\frac{1}{m} \Rightarrow \theta = \tan^{-1}(\pm 1) = \pm 45°$. While this solves Correspondence for a single point, it gives students a starting point for learning about considering average plot slope, or even multi-scale slope [8].

2.2 Including Zero in plot scale

AVD may help reduce confusion about the necessity of including zero in the vertical scale of a 2D function plot. A blog post by Andy Cotgreave[1] shows an example of when including zero is unhelpful, summarized in Fig. 1, which plots the world record times

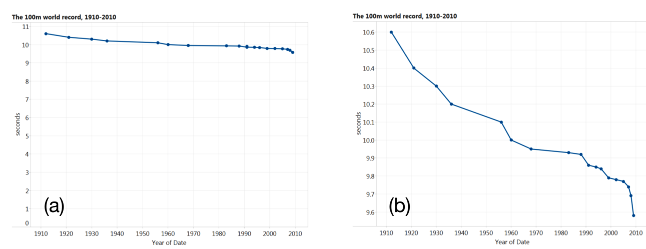

Figure 1: History of world record times in men's 100m dash, plotted with (a) or without (b) including zero in the vertical scale.

for men's 100m dash over the last century. Cotgreave discounts Fig. 1(a), which includes zero in the vertical axis as having four problems: "It doesn't really expose the change of the record over time", "It especially doesn't highlight the impact Usain Bolt had on the record", "It doesn't make great use of the space – there's lots of dead space.", and "It's boring.". Through a class discussion, students were encouraged to give mathematical "teeth" to these statements, using AVD. This in turn requires students to isolate the data changes α that are important to show: the changes in the data that would arise in some interesting but plausible alternative universe.

Being a time interval, the data contains ratio values (for which zero is an intrinsically meaningful), but the interesting α are *not* ratio changes $\alpha(x) = bx$. The noteworthy changes have the form $\alpha(x) = x - d$, an example of how ratio-valued data may be effectively interval-valued (like degrees F or C). AVD answers the question of "should I include zero?" with another question requiring some awareness of the purpose of the visualization: "do the relevant α depend on zero?". For the race times, $\alpha(x) = x - d$ do not involve zero

[1] `http://gravyanecdote.com/uncategorized/mythbusters-should-you-start-your-axes-at-zero`

because no plausible record time will ever be close to zero, and the importance of Usain Bolt's record is from d, not $\frac{x-d}{x}$. Correspondence suggests using available space to show all such ω as clearly as possible, disregarding zero as a special value. Cotgreave suggests that not including zero is "bending or breaking" a "rule". We suggest AVD provides a simpler consistent framework.

2.3 Misleading Plots

Even when a visualization is widely recognized as being misleading, AVD provides a mathematical way of describing what exactly is misleading about it. Representative Jason Chaffetz (R Utah) used Fig. 2(a) to question Planned Parenthood funding. The graph title "Abortions up – Life-saving procedures down" suggests a negative correlation between the two. However, the red and pink lines have different vertical scales (a dual-axis plot), preventing visual comparison of their slopes. This plot has been called misleading[2], but students were pressed to articulate precisely why. The pedagogical value of discovering such a mathematical basis is that it facilitates generalization and application to other examples.

With some guided class discussion, students considered an ω that mathematically expresses the relationship implied by the "Abortions up – Life-saving procedures down" title: what if the plots are reflected across a horizontal axis? This negates the trends by effectively swapping the two plots, while still interpreting each relative to its original vertical scale. The simple X-shaped visual structure of plot supports this clear ω. The Correspondence Principle then asks whether the α corresponding to ω is an interesting or important data change. Figure 2(c) shows the α by a 2D parametric plot over time. Because the vertical scale for abortions was so exaggerated in the original plot, there is very little change along the abortion axis. This is not an especially interesting or significant α. On the other hand, with a single vertical axis and the X-shaped plot structure, the same ω would lead to a very different α, shown in Fig. 2(d). Here α preserves a significant negative correlation. The the absence of a significant α corresponding to a clear ω is one way of articulating why Fig. 2(a) is misleading.

As a teaching example, the challenge is getting students to recognize and state the visibly apparent ω. This requires a kind of thinking that may arise from modern interactive visualization tools, but which we find to be nonetheless uncommon. Students must shift their thinking from "from this visualization, what can I learn about the (single) dataset it came from?" to "how could I manipulate the visualization itself, and what changes in the data could give rise to that?" (excluding parameter changes affecting appearance). This approach may be fostered with a human-computer interface perspective that analyzes the visualizations in terms of their *affordances* [6, 5, 16]: apparent opportunities for direct manipulation.

2.4 Kernel Choice by Representation Invariance

When teaching "scientific visualization" methods such as volume rendering or streamlines, students must reconstruct a continuous field from a discrete grid of regular samples, by convolving the data with a reconstruction kernel. The literature on kernel choice can be somewhat daunting, but AVD provides a way of introducing it in a concrete way. Möller et al. (and others previously) suggest that kernel accuracy can be understood not just in the traditional terms of pass-bands in frequency space, but in terms of Taylor series, which expands a function according to a series of derivatives [11]. More accurate kernels are those that push the introduction of error terms to higher-order terms in the Taylor series. This means that for some low-order polynomial, a kernel may *exactly* reconstruction the function from discrete samples.

[2] `http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/`

Figure 3 shows this with two different samplings of the same function $f(x) = -1.5x + x^2$. In this case, the "data" is the underlying function $f(x)$, and the "representations" are different possible regular samplings of the function, producing some array of values. The visualization v_{tent} with linear interpolation creates a plot (Fig.3(a)) that reveals the sampling specifics. A change in representation h (a change in how the function is sampled) is a hallucinator because it caused a visible change $\omega \neq 1_V$ in the visualization. Visualizing v_{ctmr} with more accurate kernel, such as Catmull-Rom, the quadratic function in question can be reconstructed exactly, removing the hallucinator. The plot is the same regardless of sampling. Implementing 1D convolution, and testing it by plotting data for which the correct answer is known a priori (similar to Fig. 3) was the first assignment in a scientific data visualization class.

2.5 Are Confusers Bad?

Students have frequently noted that a wide range of visualization tools have confusers, from dimensionality reduction methods that necessarily project out many degrees of freedom in the data [4], to things as simple as summary statistics (e.g. mean, covariance), which may still have a role in a visualization, even though Anscombe's quartet famously illustrates their confusers [1]. Students then ask, "So how can you possibly avoid all confusers?" We believe this highlights two things. First, we have to date failed to sufficiently emphasize that these principles, and the descriptive terms for their failures (like confusers) are only descriptive. The point is *not* to simply eliminate confusers: any sufficiently interesting visualization must have confusers. The point is to encourage students to think critically and concretely about the space of possible data changes α, and to then make an informed distinction between the important α that should have clearly visible ω, and the non-essential α that may be confusers (with faint or invisible or ω).

Second, we sense that the culture around data visualization, at least as represented by students' ideas about why they should learn visualization, is oriented around an ideal of "quality" or "excellence" that is an objective and intrinsic property of a visualization, which implies that a role of design is to judge a bad visualization from a good one. The quality of a visualization, however, is much more dependent on context, purpose, and audience: an infographic that effectively engages a reader in a journalistic setting satisfies very different needs than a figure in a scientific research paper. The reflexive assertions about the necessity of including zero in the vertical scale of a graph is a specific example of a judgment that can be made more helpful with consideration of the α that actually matter, given the context, as noted above. AVD provides a vocabulary to teach visualization design as a certain kind of informed comparison and compromise, but it is not a deterministic machinery to either create visualizations or increase their quality.

2.6 Nomenclature

The nomenclature of AVD has benefits and drawbacks. Students seem to like the terms "hallucinators" and "confusers", because they name specific things that can be tested and compared. The distinction between "misleaders" and "jumblers" is harder to convey, consistent with our original concern [9] about slippage between these terms. Still, having some way of distinguishing Invariance, Unambiguity and Correspondence failures is appreciated because it is more fine-grained than existing terms like "expressive" and "effective" [10]. A visualization can fail to be expressive, for example, either by suggesting non-existent facts about the data (against Invariance, with a hallucinator) or by failing to show something important about the data, which in turn can either be because some important α is invisible (against Unambiguity, with a confuser) or because it is shown in a confusing way (against Correspondence). For future classes, we may adopt the term "non-correspondence" or "incongruity" to refer collectively to jumblers and misleaders.

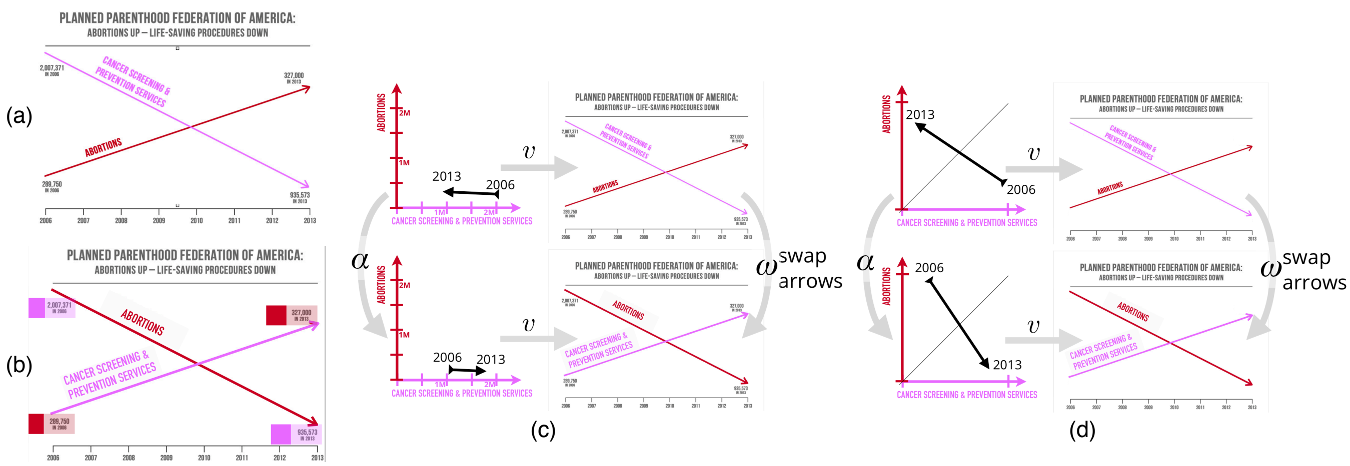

Figure 2: Original plot (a) of abortions (red) and cancer screenings and prevention services (pink) with dual-axis plot. Possible ω is reflecting graph across horizontal axis, swapping the two lines (b). Consideration of corresponding α with misleading dual scales (c), versus a single vertical scale (d).

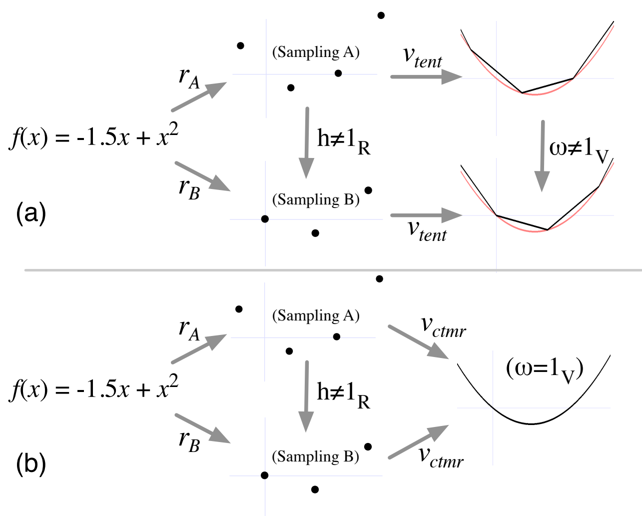

Figure 3: Illustration of representation invariance for choice of convolution-based reconstruction kernel, with linear interpolation (a) versus Catmull-Rom interpolation (b).

A more fundamental point of confusion arises with "data" versus "representation". Distinguishing them risks of a tautology relative to the hallucinator, but we could say that representation is where changes should not have an effect (e.g. representing one set with two different orderings of elements in an array), whereas data is where changes should have an effect (e.g. two consequentially different orderings of elements in a list). Ambiguity between hallucinators and confusers is possible. It is also disorienting for students to hear that the input to the visualization method is not data, but a representation, given the ubiquitous use of "data" to mean whatever is stored in files and processed by programs. Noting that a tool to visual a spreadsheet should give the same result if the rows were re-ordered is one possibly helpful example. Appealing to standard computer science pedagogy, we can also say that the difference between data and representation is analogous to the difference between an abstract data structure and its implementation, respectively: the interface to a data structure should be independent of the implementation, and it is a bug to have the interface expose specifics of the implementation. More philosophically, especially with a statistical consideration of "data" as underlying distributions or functions (e.g. Fig. 3) and "representation" as some specific sampling of the distributions, students can reasonably ask, "So is the purpose of visualization to look at the numbers we have, or is to infer something about whatever generated the numbers?" The answer is of course "yes; it depends", which can lead into a discussion about the range of possible tasks in visualization, from mundane data quality checks, to discovering general data trends.

3 Discussion

Our experience with teaching visualization via AVD is mainly limited to young computer science undergraduates, who typically imagine that assigned problems have solutions, and that solutions are either correct or incorrect (though those with some software engineering experience may recognize a more flexible space of possible answers). Using AVD, students must answer two essential questions that do not have obviously correct answers (from Sec. 1): (1) "If the world had been different in some interesting way, how would the data have been different?" (what are the important α), and (2) "How then would the visualization look different, and would that make visual sense?" (what are the resulting ω, and does $\alpha \cong \omega$).

The first question forces students to expand their thinking from data visualization to the larger context of the system or investigation that generated the data. The importance of a given α is not just a property of the data, but is derived from the original motivations to measure, organize, and understand the data. The second question forms a linkage into perceptual psychology and human-computer interface design. Especially when teaching visualization with tools like D3 [2] that facilitate interaction, students can use buttons to trigger specific α, and observe the resulting ω. Based on our limited experience, guiding students to consider specific (α, ω) pairs in terms of Unambiguity and Correspondence is more actionable than trying to adhere to general guidelines (e.g. "maximize the data-ink ratio" [14]), and it provides a concrete path towards understanding visualization design as a mode of critical thinking.

References

[1] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, Feb. 1973.
[2] M. Bostock, V. Ogievetsky, and J. Heer. $\mathbb{D}^3$: Data-driven documents. *IEEE T. Vis. Comp. Graph. (Proc. InfoVis)*, 17(12):2301–2309, 2011.
[3] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. American Statistical Association*, 79(387):531–554, 1984.
[4] D. Engel, L. Hüttenberger, and B. Hamann. A survey of dimension reduction methods for high-dimensional data analysis and visualization. In C. Garth, A. Middel, and H. Hagen, editors, *Visualization of Large*

and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering - Proceedings of IRTG 1131 Workshop 2011, volume 27, pages 135–149. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2012.

[5] W. W. Gaver. Technology affordances. In *Proceedings SIGCHI Conference on Human Factors in Computing Systems*, pages 79–84. ACM, Apr.–May 1991.

[6] J. J. Gibson. *The Ecological Approach To Visual Perception*, chapter 8: The Theory of Affordances. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1986.

[7] C. G. Healey, K. S. Booth, and J. T. Enns. High-speed visual estimation using preattentive processing. *ACM T. Comp.-Hum. Int.*, 3(2):107–135, 1996.

[8] J. Heer and M. Agrawala. Multi-scale banking to 45 degrees. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):701–708, Sept. 2006.

[9] G. Kindlmann and C. Scheidegger. An algebraic process for visualization design. *IEEE Transactions on Visualization and Computer Graphics (Proceedings VIS 2014)*, 20(12):2181–2190, Nov. 2014.

[10] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM T. Graph.*, 5(2):110–141, 1986.

[11] T. Möller, K. Müller, Y. Kurzion, R. Machiraju, and R. Yagel. Design of accurate and smooth filters for function and derivative reconstruction. In *Proc. 1998 IEEE Symposium on Volume Visualization*, pages 143–151. ACM, 1998.

[12] T. Munzner. A nested model for visualization design and validation. *IEEE T. Vis. Comp. Graph.*, 15(6):921–928, 2009.

[13] S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

[14] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 2nd edition, 2001.

[15] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *Intl. J. Hum.-Comp. Stud.*, 57(4):247–262, 2002.

[16] C. Ware. *Information Visualization*, chapter 1: Foundations for an Applied Science of Data Visualization, pages 1–30. Elsevier, third edition, 2012.

[17] C. Ware. *Information visualization: perception for design*. Elsevier, 2012.

[18] R. M. William S. Cleveland, Marylyn E. McGill. The shape parameter of a two-variable graph. *Journal of the American Statistical Association*, 83(402):289–300, 1988.

[19] G. Wyszecki and W. S. Stiles. *Color science: Concepts and methods, quantitative data and formulae*. Wiley New York, 1982.

[20] C. Ziemkiewicz and R. Kosara. Embedding information visualization within visual representation. In Z. W. Ras and W. Ribarsky, editors, *Advances in Information and Intelligent Systems*, volume 251 of *Studies in Computational Intelligence*, pages 307–326. Springer, 2009.