DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-131

# From Vision Science to Data Science: Applying Perception to Problems in Big Data

***Remco Chang, Fumeng Yang, Marianne Procopio***
***Department of Computer Science***
***Tufts University; Medford, MA***

## Abstract

*In the era of big data, along with machine learning and databases, visualization has become critical to managing complex and overwhelming data problems. Vision science has been a foundation of data visualization for decades. As the systems that use visualization become more complex, advances in vision science are needed to provide fundamental theory to visualization researchers and practitioners to address emerging challenges. In this paper, we present our work on modeling the perception of correlation in bivariate visualizations using the Weber's Law. These Weber models can be applied to definitively compare and evaluate the effectiveness of these visualizations. We further demonstrate that the reason for this finding is that people approximate correlation using visual features that are known to follow the Weber's Law. These findings have multiple implications. One practical implication is that results like these can guide practitioners in choosing the appropriate visualization. In the context of big data, this result can lead to perceptually-driven computational techniques. For instance, it could be used for quickly sampling from big data in a way that preserves important data features, which can lead to better computational performance, a less overwhelming user experience, and more fluid interaction.*

## Introduction

The rise of data science, spurred by the growth of data sizes and complexity, has led to new advances in the fields of databases, machine learning, and visualization. These three pillars enable data stakeholders to store, analyze, and make sense of big data.

Of the three areas, visualization represents the last step of the pipeline where automated computation meets the human user. Recent advances in visualization techniques have led to innovative systems that allow the user to interactively and visually explore large amounts of data. Success stories such as Tableau [1, 22], SpotFire [4], SAS Visual Analytics [2] demonstrate the importance of integrating visualization with machine learning and databases to solve big data problems.

However, as the data size and complexity continue to rise, it has become more obvious that the visualization component has become both the critical element as well as the bottleneck in the analysis pipeline. Both the database and machine learning can scale to meet the increased data complexity by adding more storage and more compute nodes in a server farm. The visualization component, on the other hand, is constrained by both the display technology as well as the human user's perceptual and cognitive limitations.

In this paper, we examine the constraints of the visualization component in the context of big data analytics. While these constraints can be considered as limitations to the data analysis pipeline, we propose that they also represent opportunities to develop a new user-centric paradigm that makes use of vision science to design not only new visualization techniques, but also database and machine learning algorithms. The resulting system represents a new approach of big data analytics that puts the human user's needs and limitations first, thereby creating a system that is faster, more fluid, and more intuitive to the user.

## Background

Figure 1 shows a traditional (non-interactive) process of data visualization (adopted from the data state reference model by Chi [7]). First the data is retrieved from the database into the visualization system. The system then maps elements of the data to different perceptual elements (such as color, size, shape, etc.) [5]. Lastly, the human user perceives the image and identifies patterns and trends that might lead to new insights about the data.

Although simplistic, this pipeline serves as the foundation of all visual analytics systems today. Recent advances in this topic can be seen as improving the stages in this pipeline. For example, nanocubes [14] and multivariate data tiles [15] are examples of data storage techniques that make use of compact data structures that aggregate underlying data in a hierarchical way. These data summaries can be precomputed at various levels of abstraction based on the number of pixels available for the visualization and the size of the underlying raw dataset.

Binned aggregation [15], [24] takes this even further by separating the raw data into bins and returning a small set of summary statistics. This technique can show both densities and outliers by varying the bin size. Any issues with variability in the summaries can be resolved with various smoothing methods [24].

Another technique is to provide *approximate incremental answers*. The sampleAction [9] and the VisReduce [12] system incrementally returns partial answers to user queries computed over increasingly larger samples of data. This has the benefit of providing a partial response to an exploratory query quickly and once the user has a good enough answer, they can stop the process and move on.

For exact answers from raw data, systems such as Dremel [16] and MapD [17] take advantage of parallelism and large computing clusters for computational power. Although effective, the cost and proprietary query language can hinder widespread adoption.

While these new techniques, methods, and systems have led to a faster and more efficient data visualization process, our goal in this paper is fundamentally different. Unlike these advances that seek to improve a component of the pipeline, we propose

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-131

Figure 1: A simplified pipeline of data visualization. The data is first fetched from the database and delivered to a client system to render a visualization. The human user perceives information from the visualization.

that a new paradigm of the data visualization process can lead to advances in vision science, visualization techniques, and closer integration of machine learning, database, and visualization.

## Human Perceptual Limitations

In order to develop a paradigm that focuses on the limitations of the human perceptual and cognitive abilities, we first examine some examples of low-level limitations in the data visualization process.

Consider an example of a visualization display that has a resolution of 1000x1000 pixels resulting in a total of 1 million pixels, each with the capability of displaying three color channels. When used in a visualization, it has been shown that this 1 million pixel (the resolution of the display) is the theoretical upperbound of the maximum amount of information that the human can perceive [6].

This theoretical upper bound is important because it suggests that the first step in visualization pipeline shown in Figure 1 is lossy when displaying a large amount of data. For example, imagine a database that holds 10 million records of data. When the 10 million records are sent to the visualization system, the 10 million records need to be "compressed" into 1 million pixels resulting in a 10:1 ratio of data loss. The "compression" can be performed using a variety of methods. Most commonly the data is aggregated (averaged) into a single value, but other methods such as clustering and sampling are also frequently used [20].

In addition, beyond the theoretical limitation of the display technology, the second step in the visualization pipeline is also

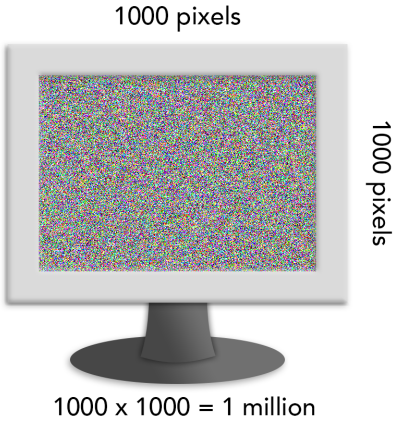



Figure 2: A screen with a resolution of 1000 x 1000 can at most display 1 million pixels. When a visualization reaches this upper bound, however, the resulting image is often unrecognizable.

lossy. While the display resolution constraints what the user is able to perceive, comprehending the visualization is further constrained by the user's cognitive limitations. For example, using the previous example, when each pixel represents 10 data elements in an aggregated fashion, the visualization can result in a colorful "snow" (see Figure 2). Although the data-visual mapping of this visualization may be coherent, accurate, and maximizing of information content, the cognitive limitation of the user makes this visualization less than useful [6].

## Applying the Perceptual Limitations

Based on the user's perceptual and cognitive limitations, we propose that there are two immediate opportunities for optimizing the design of a visualization system.

### *Pixel-Based Constraint*

First, for a display system that can render at most 1 million "pieces" of data, it does not make sense for a database to transfer more than 1 million rows to the visualization system with that display. Since transferring data from the database to the visualization can be costly (especially when the two are connected via network), minimizing the amount of data transferred from the database will improve the performance of the overall system.

It is relevant to note that the 1 million rows of data transferred from the database can be raw or processed data. Using sampling techniques [3], the database can choose the most representative 1 million raw data elements. Alternatively, using aggregation or clustering techniques, each of the 1 million rows can represent the mean of a large number of raw data elements. When combined with the notion that data transfer is costly, this implies that most of the data processing should take place in the database system. Only the resulting computed data should be sent to the visualization system for rendering.

### *Perceptually-Based Constraint*

Second, we consider and leverage the user's perceptual and cognitive limitations when perceiving an image. For example, Figure 3 shows two images that appear very similar. However, the image on the right has a significantly coarser resolution than the one on the left (301 kb vs. 115 kb, a 2.62:1 compression ratio). Many existing image compression techniques (such as JPEG-2000 [23]) are based on this same idea: for as long as the user cannot tell the difference, keep reducing the resolution of the image. The resulting simplified images are smaller in file sizes, which are faster to transit via network and to render on screen.

The notion of perceptible differences is often measured in

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-131

Figure 3: Examples of JPEG2000 compression. The picture on the right has less than half the resolution as the one on the right but results in little difference in perception.

terms of *just noticeable difference*, or *JND*. One unit of JND is defined as the minimum change in the stimulus (e.g. the image) before a human can perceive that change has occurred. Different stimuli, such as color, brightness, smell, will have different values of JND, but the notion of one JND can be universally applied to all aspects of human sensory perception. For a more detailed definition of JND, see the following section on *Just Noticeable Difference*.

Our second observation can therefore be specifically defined as: a visualization can continue to be simplified for as long as the change does not exceed one JND. However, in data visualization, the measurement of JND is less well understood. For example, consider the two boxplot visualizations (also known as box-and-whisker plots) in Figure 4a. Although these two visualizations appear very similar, there is in fact a difference between the two. Figure 4b uses a red line to highlight that difference.

Now consider the two barchart visualizations in Figure 5a. The difference between these two barcharts should be much easier to detect than the previous boxplot example (refer to Figure 5b to see the difference). Although the *magnitude* of the difference between the two boxplots in Figure 4a and the two barcharts in Figure 5a are the same, the *perceptual differences* are different. This suggests that the JND of these two visualizations are in fact different.

Similar to the JPEG-2000 example (Figure 3), a boxplot or barchart visualization can be "simplified" for as long as the user cannot detect the differences. In the boxplot example (Figure 4a), the amount of simplification can be higher (because the JND is larger), whereas less simplification can be done in the barcharts (because the JND is smaller).

An interesting open question is how visualizations can be "simplified". As noted earlier, existing approaches have utilized sampling, streaming, and aggregation techniques [3, 9, 10]. Regardless of the applied technique, the value of using a simplified visualization that is perceptually the same (i.e. within one JND) as the original is that the data size can be reduced, which leads to faster data transfer, processing, and rendering.

## Perceptually-Driven Visualizations

The fact that differences in barcharts are easier to detect than in boxplots is well-known. However, the visualization community does not yet have clear answers to "by how much are barcharts easier detect than boxplots and why?" In the seminal work by Cleveland and McGill [8], the authors examined various visualizations and compared and documented differences in their effectiveness. This study, and many others that have extended this work, largely focus on "ranking" the effectiveness of the visualizations, but do not directly model the relationship between the amount of change in visualization versus the amount of change in perception.

Since JND is a measure of change of perception (one unit of JND is the minimum amount of change in stimulus that can be perceived), modeling the relationship between changes in visualization and perception is the first step towards quantitatively comparing visualizations and developing perceptually-driven visualizations. Changes in the visualization can be measured in changes in pixels. For example, for images such as the ones in Figure 3, one can compute the sum of changes for all pixels. For information visualizations such as boxplots and barcharts, the difference in length between bars can also be measured in number of pixels.

In the sections below, we introduce the formal definition of JND and give an example of modeling JND in information visualization.

### *Just Noticeable Difference*

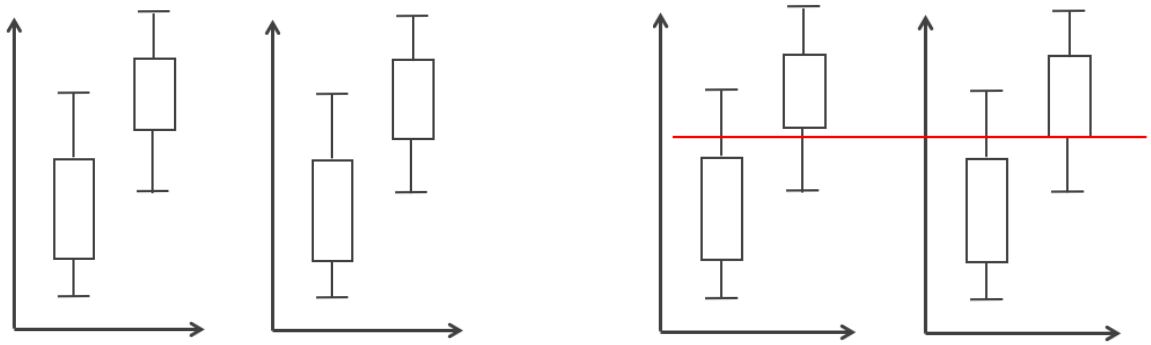
Just noticeable difference, or JND, was first defined by Ernest Heinrich Weber in the 18th century as:

$$\frac{\Delta I}{I} = k \tag{1}$$

Known as the *Weber's Law*, in this equation $I$ represents the intensity of the original stimulus and $\Delta I$ represents the smallest amount of change that is required for a human to perceive the difference (the JND). This simple relationship elegantly captures the relationship between sensory stimulus and human perception across a large number of stimuli and became the foundation of the field of psychophysics.

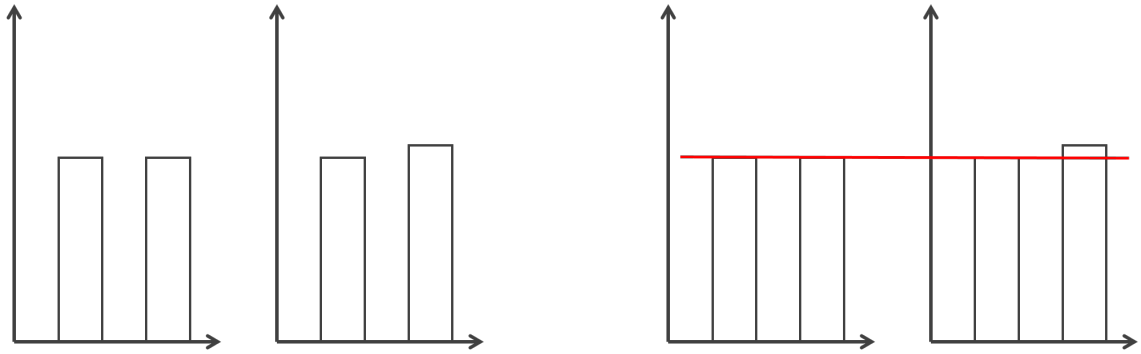
It is relevant to note that not all JNDs have the linear relationship with the intensity of the stimulus. Other perceptual laws, such as Fechner's Law and Stevens' Power Law [21] have been used to model more complex sensory perception, such as

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-131



(a) (b)

Figure 4: (a)Two similar boxplots. The difference between them is very hard to perceive. (b) The red line highlights the difference between two boxplots. The length of the box on the right is different.



(a) (b)

Figure 5: (a) Two barcharts with the same difference as the above two boxplots. However, the difference is easier to detect. (b) The red line highlights the difference between two barcharts.

wavelength of light. For example, Stevens' Power Law models perception and the intensity of the stimulus using an exponential relationship. A full treatment of perceptual modeling and psychophysics is out of the scope for this paper. Our goal in this paper is to demonstrate how perceptual modeling in information visualization can be used in connection to data science and big data computation. Below we give an example on how such perceptual models can be developed and suggest how these models can be used in data science.

### *Modeling JND in Perception of Correlations using Bivariate Visualizations*

In a 2010 paper by Rensink and Baldridge, the authors showed that the perception of correlation in a scatterplot can be modeled using the Weber's Law [19]. The study utilized vision science experimental techniques to establish that the participants' ability to discern correlation in the data decreases linearly as the data becomes less correlated (see Figure 7). Using a side-by-side stimuli to estimate the discriminability between scatterplots of two correlation values and a staircase method to systematically and dynamically adjust the difference between the two, the authors were able to establish that the JND in the perception of correlation in scatterplots (i.e. minimal perceptible difference between the side-by-side stimuli, or $\Delta I$ in Equation (1) is linearly correlated with the value of the correlation (i.e. intensity of stimuli, $I$).

This finding is significant to the information visualization community because it establishes that techniques and principles in vision science used to model low-level sensory stimuli can be applied to model higher-level perception (such as perceiving correlation) in abstract information visualization. Further, with the Weber equation, it is possible to quantitatively measure how other design channels can affect perception, thus leading to a path towards developing a foundation of vision science for information visualization [18].

Extending the work by Rensink and Baldridge, Harrison *et al.* applied the same experimental techniques and tested whether the perception of correlation follows the Weber's Law in nine bivariate visualizations [11]. First, the authors established that using crowdsourcing (on Amazon's Mechanical Turk), they were able to replicate the findings of Rensink and Baldridge which were established using traditional in-person studies. The authors then tested a total of nine common bivariate visualizations used in commercial office tools like Excel (see Figure 6). The result of this work shows that in all tested visualizations, the perception of correlation follows the Weber's Law similar to Rensink and Baldridge's finding for scatterplots. Further, since each of these visualizations have a different Weber fraction ($k$ in Equation (1)), the authors were able to rank the visualizations based on their effectiveness in representing correlations in the data (see Figure 8).

More recently, Kay and Heer extended the work by Harri-

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-131

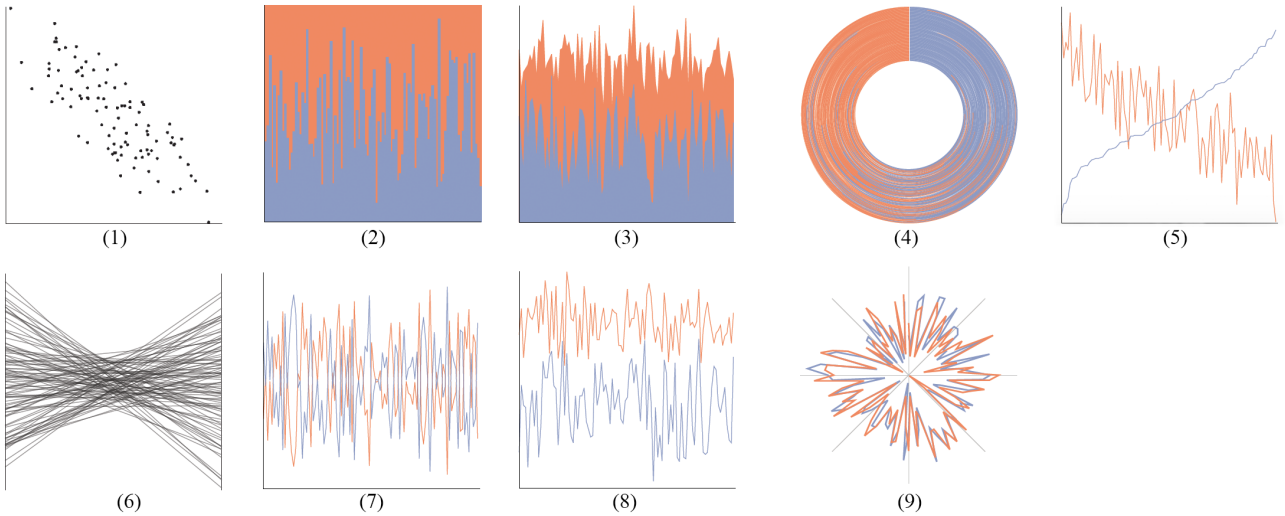



Figure 6: All the tested visualizations in the stud by Harrison *et al.*

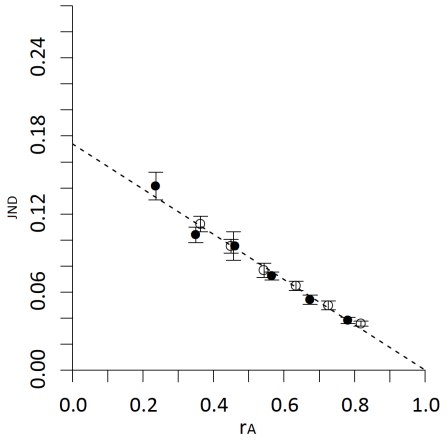



Figure 7: The result of study by Rensink and Baldridge showing that the relationship between stimulus (correlation, *r*) and perception (JND) is linear.

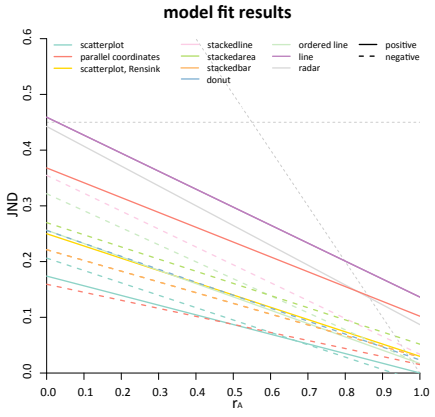



Figure 8: The result of the study by Harrison *et al.* showing that all nine tested visualizations follow the Weber Law for the perception of correlation.

son et al. to improve the comparison between the visualizations for the task of perception of correlation [13]. Using a log-linear model, the authors find a better fit to account for residuals and variance in the same data. The result suggests that the tested visualizations can be grouped into four categories ranging from high precision to indistinguishable from chance. Although the log-linear model no longer fits the Weber's assumption (that the intensity of the stimulus and the JND are linear), the research is a step forward in developing a vision science for information visualization.

## Discussion

The research above in modeling the perception of correlation using visualizations are examples of how vision science can be used to better understand the effectiveness of information visualization. Differing from most research in information visualization that focuses on "which visualization is better", these examples that produce models can go a step beyond to answer "by *how much* is the visualization better."

Beyond using these models to compare visualizations, another value for developing perception models is to use them for data computation because of the predictive capability of models. For example, knowing that the JND of correlation using a scatterplot when the correlation value is high means that the data cannot be aggressively sampled or aggregated. Conversely, when the correlation value is low, the user cannot easily distinguish two scatterplots of similar correlations, which leaves more room for data approximation that could save computation time without affecting the user's ability to perceive information and make decisions.

However, since a user can use a visualization for a range of purposes (e.g. reading a specific value, comparing subsets of data, identifying outliers, etc.), having a model for a single task is insufficient. As these examples only focus on one particular task (perception of correlation), the resulting models are limited and not yet applicable for general use. What is needed then is a continued collaboration between the vision science and the information visualization communities to research and develop similar models of visualizations and tasks. The outcome of such founda-

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-131

tional work will lead to improved understanding of the effectiveness of visualizations, but also applicable models that can inform data computation and representation.

## Conclusion

In this paper, we introduced the concept of perceptually-driven visualizations for large scale data computation and visualization. We demonstrate that a closer examination of the traditional data visualization pipeline in the context of big data leads to the identification of two bottlenecks in the data visualization process where the amount of the transferred data is limited. While these bottlenecks can be perceived as constraints in the process, we propose that they can also be viewed as opportunities to improve data computation and information flow.

First, the resolution of the display pose as the initial bottleneck. Given any display, its resolution serves as a theoretical upper bound of how much information can be shown to the user. This upper bound dictate the maximum amount of data that should be transferred from the back-end data source to the visualization system. By adhering to this upper bound, a visualization system can render data at a faster rate by reducing the latency caused by transferring unnecessary data.

Second, the human perceptual system serves as another bottleneck. Following the notion of just noticeable difference (JND), we demonstrate that taking advantage of two visualizations that are indistinguishable can lead to opportunities for applying data sampling and aggregation. Further, we suggest that models of perception of abstract information visualizations developed by applying experimental techniques in vision science can be used to measure the tradeoff between perceptibility versus data accuracy.

These two approaches represent two starting points for developing a framework of perceptually-driven visualization. While these approaches remain mostly theoretical in nature, early examples suggest that they are promising in scaling visualizations to rendering large amounts of data. By further close collaborations between the vision science and the information visualization communities, these approaches can be realized that can have foundational and applied impact to the future of information visualization research and practice.

## References


[1] *Tableau*, 2003 (accessed February 29, 2016).

[2] *SAS*, 2013 (accessed February 29, 2016).

[3] Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, and Ion Stoica. Blinkdb: queries with bounded errors and bounded response times on very large data. In *Proceedings of the 8th ACM European Conference on Computer Systems*, pages 29–42. ACM, 2013.

[4] Christopher Ahlberg. Spotfire: an information exploration environment. *ACM SIGMOD Record*, 25(4):25–29, 1996.

[5] Jacques Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[6] Min Chen and Heike Jaenicke. An information-theoretic framework for visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1206–1215, 2010.

[7] Ed H Chi. A taxonomy of visualization techniques using the data state reference model. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pages 69–75. IEEE, 2000.

[8] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[9] Danyel Fisher, Igor Popov, Steven Drucker, et al. Trust me, i'm partially right: incremental visualization lets analysts explore large datasets faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1673–1682. ACM, 2012.

[10] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data mining and knowledge discovery*, 1(1):29–53, 1997.

[11] Lane Harrison, Fumeng Yang, Steven Franconeri, and Ronald Chang. Ranking visualizations of correlation using weber's law. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1943–1952, 2014.

[12] Jean-François Im, Felix Giguere Villegas, and Michael J McGuffin. Visreduce: Fast and responsive incremental information visualization of large datasets. In *Big Data, 2013 IEEE International Conference on*, pages 25–32. IEEE, 2013.

[13] Matthew Kay and Jeffrey Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):469–478, 2016.

[14] Lauro Lins, James T Klosowski, and Carlos Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2456–2465, 2013.

[15] Zhicheng Liu, Biye Jiang, and Jeffrey Heer. immens: Real-time visual querying of big data. In *Computer Graphics Forum*, volume 32, pages 421–430. Wiley Online Library, 2013.

[16] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. Dremel: interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment*, 3(1-2):330–339, 2010.

[17] Todd Mostak. An overview of mapd (massively parallel database). *White paper, Massachusetts Institute of Technology, Cambridge, MA*, 2013.

[18] Ronald A Rensink. On the prospects for a science of visualization. In *Handbook of human centric visualization*, pages 147–175. Springer, 2014.

[19] Ronald A Rensink and Gideon Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, volume 29, pages 1203–1210. Wiley Online Library, 2010.

[20] Ben Shneiderman. Extreme visualization: squeezing a billion records into a million pixels. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 3–12. ACM, 2008.

[21] Stanley S Stevens. On the psychophysical law. *Psychological review*, 64(3):153–181, 1957.

[22] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):52–65, 2002.

[23] David Taubman and Michael Marcellin. *JPEG2000 Image Compression Fundamentals, Standards and Practice: Image Compression Fundamentals, Standards and Practice*, volume 642. Springer Science & Business Media, 2012.

[24] Hadley Wickham. Bin-summarise-smooth: a framework for visualising large data, 2013.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-131

## Author Biography

*Remco Chang received his BA in Computer Science and Economics from the Johns Hopkins University (1997), his MSc in Computer Science from Brown University (2000), and his PhD in Computer Science from the University of North Carolina Charlotte (2009). Since then he has worked as a software engineer at Boeing, a research scientist at the University of North Carolina Charlotte. He is currently an Assistant Professor in Computer Science at Tufts University. His research interests include visual analytics, information visualization, human computer interaction, and databases.*

*Fumeng Yang is currently a graduate student in the Department of Computer Science at Tufts University. She received her B.E. in Computer Science and Technology from Shandong University, China (2013).*

*Marianne Procopio is currently a graduate student in the Department of Computer Science at Tufts University. She received her BA and MA in Computer Science from Boston University (2007).*