



Designing for Insight: A Case Study from Tennis Player Analysis

Kim Albrecht and Burcu Yucesoy

Center for Complex Network Research

Can design play a role in the scientific process of generating insights? To answer this question, we need to explore how two often-separated fields, data science and visualization design, can influence each other and create new insights through images.

First off, we need to delve into the meaning of “design.” The origins of the word are based in the Italian word *disegno* and the Latin word *designare*. Both can be translated as “to mark out” or, more literally, “de-sign.”¹ Today the English word design is used as both a noun and a verb and has attained widespread use in contemporary culture. Everything, from nails to spacecrafts, seems to be “designed.”² In this context, we often anticipate some kind of outcome—design as something that suggests an expensive sophistication beyond the actual product, as in designer shoes, designer labels, and designer food.

In this article, rather than connecting the word design with an outcome, we’ll look at it as a process of drawing things together, layer by layer, into more complex arrangements. This notion of design comes from Bruno Latour,³ who has argued that design is always a redesign, connecting elements in a way that their outcome creates something new. Other intellectuals have also thought of it in this way. Vilem Flusser wrote that design does not simply draw art, science, and technology together, but rather it is a synthesis, the combination of two or more elements toward a new unity, connecting separated knowledge domains.⁴ In his book *The Semantic Turn*,⁵ Klaus Krippendorf described a paradigm shift in design from thinking about how artifacts function to how we are influenced

by them and how they connect individuals and create discourses.

Visualization design focuses on one specific aspect of design: the “drawing together” of visual arrangements of information to amplify cognition. Visualization is an important tool, necessary for making sense of vast amounts of data.^{6,7} Many data science projects make use of visualization techniques to illustrate and explain their results. They generally follow this rigid order: analyze and acquire results first; then visualize the findings to facilitate communication.⁸ But complex interactive visualizations can also be excellent exploration tools to help guide the analysis, detect early signs of problems and irregularities, suggest new discoveries, test the effectiveness and efficiency of scientific models, and so forth.⁹ For the same purpose, many data scientists use simple visual methods such as static scatterplots and histograms, but the large amount of data they consider makes this process inefficient. With the expert use of current advanced visualization techniques, the process could be much faster and streamlined.

As part of our research, we apply Hadley Wickham’s notion about the interplay between statistics (in our case, data science) and visualization: visualization is a tool that helps us discover relationships, whereas statistics helps us understand if the relationships really exist.¹⁰ Wickham and his colleagues described statistics as a tool of skepticism and visualization as a tool of curiosity. Both techniques use a feedback loop that can enhance the process of knowledge production—from visual exploration to statistical analysis, model development, and statistical conclusions, and then back

to visual exploration. This loop can pave the way to more efficient exploration and a better understanding of large datasets.

This article provides a case study in the design process of visualizations, sharing how multiple visualizations were drawn together to generate knowledge about a project concerned with measuring success in sports. Specifically, we were interested in how performance and popularity create success in tennis. Popularity is often perceived as only loosely related to performance, but both contribute to success. Our goal is to quantify the relationship between the three (performance, popularity, and success) and, through this, develop a better understanding how they are connected. To achieve this goal, we looked at tennis players and considered their various performance parameters, starting with their current standings in the world rankings and the page-view statistics of their Wikipedia pages (as a proxy for popularity). Visualization became an important tool for discovering patterns, outliers, and features in both performance and popularity and understanding the relationship between them.

Using this case study format, we will describe the process of designing visualizations to analyze data and show the ways they helped reveal novel insights. (The large data science aspect of the project, from analysis to modeling, is outside the scope of this article. See earlier research for more details on this topic.¹¹)

Design Process

Our design process began with one specific purpose: create interactive visualizations to identify and understand outliers. The data scientists developed a model that predicts Wikipedia page views based on a player's performance data. We evaluated the model predictions and investigated why it did not work well for some players. Although we used statistical methods to develop the model and test its fitness (tools of skepticism), visualizations helped us detect anomalies within the predictions (tools of curiosity).

Design Process 1: Data/Model Difference

Our first visualization presented a simple scatterplot that showed one dot per player to correlate Wikipedia page-view counts on one axis with model prediction values on the other. Although outlier dots were visible, all other player dots formed one big area. We then progressed with the ideas shown in Figure 1

To better consider overlapping dots, we adapted ideas from William Playfair's difference chart¹² and floating bar charts. We represented both page views

and the model prediction on the y axis by using a line to represent each player. The length of each line showed the difference between the model and the actual Wikipedia page-count data—that is, the shorter the line is, the better the model fits. Red indicates that the model is over-predicting, and blue that it is under-predicting. The x axis became an ordinal scale where every line had a discrete space. Through a small interface, the x axis was now sortable by different dimensions. This new visualization removed overlap while at the same time showed and compared one additional data dimension than the scatterplot (see Figure 1a).

Still, this newly created visualization showed the data in a highly aggregated form by only assigning

Rather than connecting the word design with an outcome, we'll look at it as a process of drawing things together, layer by layer, into more complex arrangements.

two data points to each player. We were interested in the players' temporal developments, so we created a small multiples visualization to show the development of prediction and page views for each year and every player. Using page-view data from 2009 to 2015, we represented each player with six data points for the page views as well as for the model prediction. Many visual representations would have been possible to show this development. We choose to keep the idea of the line from the previous plot and advance the line to use the area from left to right over time. Height and color depended on the comparison of two data points per year. We ended up with 500 difference charts that showed upcoming players, retiring players, injuries, and other outliers (see Figure 1b). This visualization provided a level of detail that was completely hidden before. We then used these small multiples and arranged them, not only in grid structures that can be arranged by different parameters, but also in multiple scatterplots. Such arrangements become visually complex, but they let the viewer compare two dimensions for each player plus the developments of their Wikipedia careers and the model prediction.

These new highly complex visual arrangements enabled more complex interaction, but the data was still highly aggregated compared with the raw data of daily Wikipedia page views. To allow users to interact more closely with the data, we built a

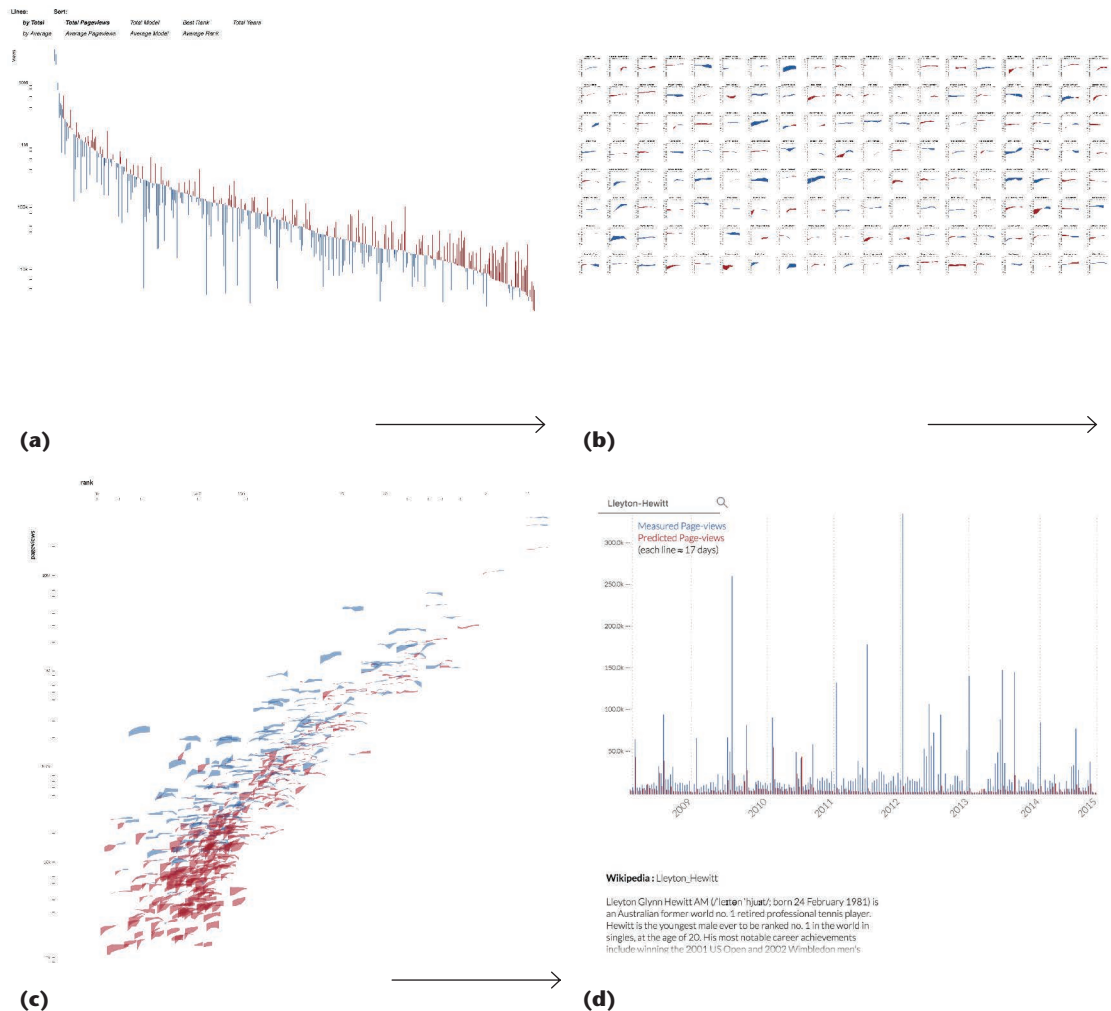


Figure 1. Design process 1. (a) The first visualization compares page views and model prediction for all tennis players. (b) The second interface is based on the same principle as the first, but we show the progression over the years, which results in a grid of small multiples. (c) In the third view, we use the small multiples and arrange them in scatterplots based on different parameters. (d) The last view gives a more detailed view of one player. The user can switch seamlessly between viewing all players and viewing an individual player.

complementary view that became visible when a specific player was selected. This view compared page-count data and model prediction for every 17 days, which is the usual length between tournaments. So from a yearly aggregate, we designed a visualization that lets us view the data even closer. The strong differences between the 17-day intervals suggested a different visualization requirement as area shapes with strong spikes become difficult to read. We decided to use grouped bar charts for a close-up view. In addition to the chart, we embedded the introduction of the corresponding player's Wikipedia article in the interface (see Figure 1d). This was helpful because the strange spikes were often already explained there.

Design Process 2: Turning Models into Graphics

We next asked if we could use visualization to show the inner workings of the model that pre-

dicts the page views? Until this point, the model and its inner workings had been in a black box. To address this, we had to understand all the variables that create the model, rank their influence, and map these dimensions in a meaningful way onto visual variables. The Wikipedia page-view prediction is based on five performance variables: the player ranking, the value of tournaments, the number of matches at the tournament played, the best opponent the player competed against in each tournament, and the total years played. Figure 2 shows the resulting visualization.

We found the most important variable was a player's rank at the start of each tournament. The first chart we created plotted time on the x axis and rank on the y axis. We used a nonlinear timeline based on the number of tournaments played in each year. This made the analysis more meaningful because it more clearly showed how

many tournaments a player had competed in, and it made the tournaments evenly distributed over the page, regardless of whether the player participated in two tournaments or 100. We also used a logarithmic scale for ranking. This emphasized the importance of the high rankings, for which it is much harder to achieve differences than the lower rankings.

The second-most important dimension in the model was the number of rounds a player competed in during a tournament. Seven rounds meant a player reached the final. We mapped lines onto the ranking points and extended them upward. The longer the line, the more rounds a player reached in the tournament and the higher the visual weight on the page.

In a third step, we used color to indicate the tournament's value. Color is always difficult for linear scales like tournament values. Thus, to make the color scale work, we binned tournaments into seven categories. The low-ranked tournaments only changed from yellow to orange, whereas the highly ranked tournaments had more dramatic color changes. We created an interactive legend that only highlighted one set of values when the user hovered over it with a mouse. This helped users more clearly see the patterns of tournament value changes within a career.

In a final step, we added small circles to show tournaments where the player had to play against a much higher ranked opponent.

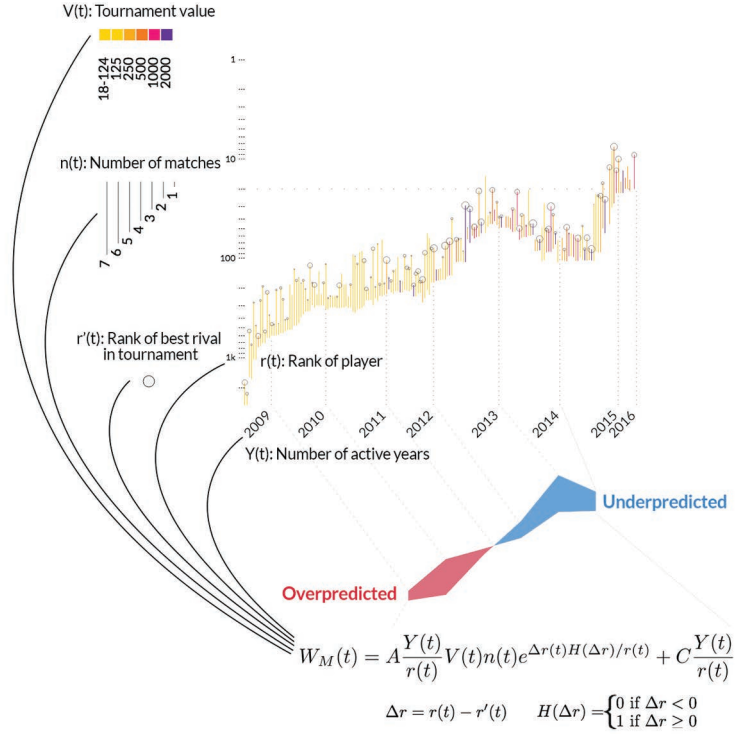


Figure 2. Design process 2 model/graphic connection. A visual explanation about how the model is connected to the two resulting graphics. The performance visualization shows all parameters from the model. The page view and model comparison shows the outcome of the prediction and reality.

The visualizations in Figure 3a show both a tennis player's entire career and all the variables used in the model to predict Wikipedia page views

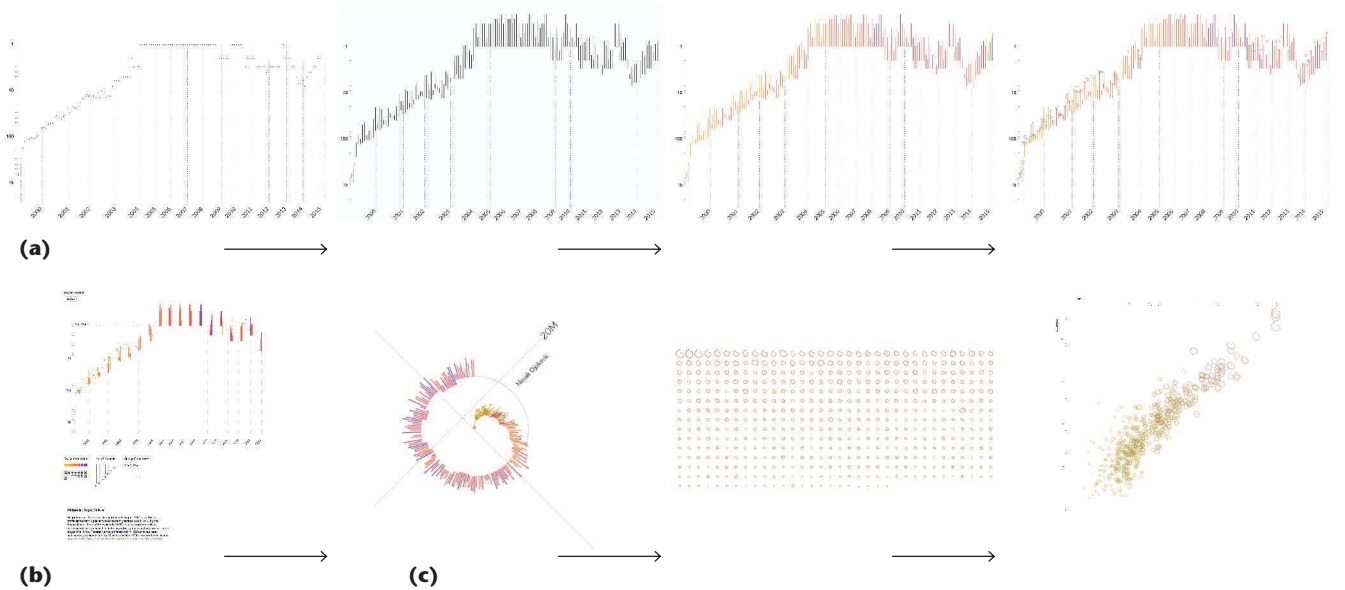


Figure 3. Design process 2 process image. In these eight steps, we see the evolution of creating a visual description of the data underlying the model prediction. (a) The first four steps show the process of the graphic of a single player. (b) The fifth image shows the resulting interface. (c) Steps six to eight show the evolution from one player to a view that compares all players.

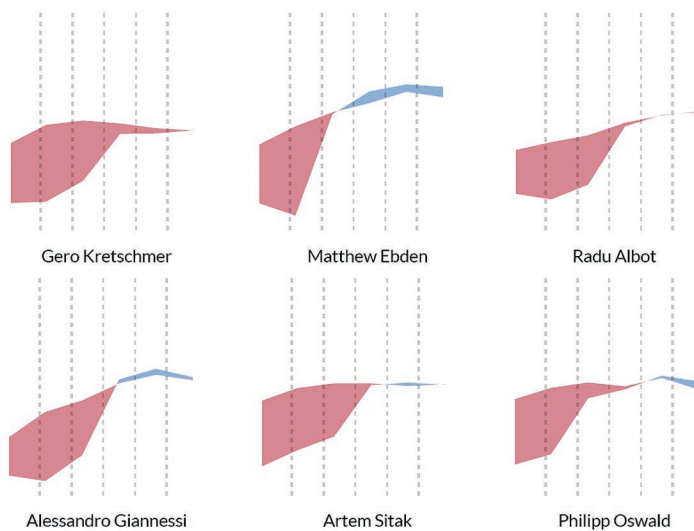


Figure 4. Data/model difference charts over time showing late catch-ups. Here we see how page views and model predictions synchronize after the Wikipedia page is completed.

for that player. However, they do not allow a comparison between different players. To address this, we built an interface similar to the scatterplot in the data/model difference chart for the players' careers. Because the information density per player is high, plotting the individual charts in one scatterplot became unreadable. To make it possible to compare all players, we made a seemingly bad design decision. We switched the time axis from being linear to radial. Radial timelines are much harder to read but they create iconic symbols, offering a unity for each that the linear version does not provide. This switch let us see the data for all 500 players together in one plot and compare their careers using different dimensions (see Figure 3c).

Findings

The resulting visualizations from the two design processes ended up being complementary. Although emerging from different starting points, both have views that compare all players as well as close-up views that show more detailed views of single players. We integrated the two together into one interface and added another visual step between the two overviews. Fluid transitions between all the views enabled considerations beyond those from the single visualizations. The transition between overviews can also be interpreted and analyzed to add another dimension of comprehension.

Our two design processes have some underlying similarities. Both add more information into the visualization at each stage of the process and,

by doing so, create higher visual complexity. However, they also show more data dimensions and more data points for each dimension. This process creates graphics that are harder to encode but, at the same time, allow us to see more relations in one arrangement. The additional layers of information not only allowed us to find outliers within the data but introduced questions we did not think to ask before the visualizations existed. Most findings can be seen in multiple views, but we focused on the visualization in which the pattern became most visible.

The data/model difference charts over time helped us discover a large number of players whose popularity seemed to be overestimated by our model for years until the actual page views caught up (see Figure 4). We discovered that unstable or not-yet-existing Wikipedia pages caused these mismatches. The model was predicting Wikipedia pages that were not set up or consisted only of the player's name. Without seeing this pattern in the visualization, we would not even have thought about such anomalies.

In the close-up data/model difference visualization, variations between the model and actual page views became even more prominent. This led us to all kinds of stories in which players attained popularity without actual success in tennis. For example, in mid-2013 Ryan Sweeting married the actress Kaley Couco. At the same time, there was an unexpected burst in his popularity on Wikipedia (see Figure 5a). In another example, Lukas Rosol's Wikipedia page had an increased number of views in mid-2012 when he defeated the then number two ranked player in the world, Rafael Nadal, in the second round of Wimbledon (see Figure 5b). The interactive visualization helped us identify these and other short bursts of popularity not accounted for by the model.

The performance visualization gives us insights about entire tennis careers as well as insights about how the model measured the number of page views. Abrupt career ends, periods of injuries, and paths of success became visible. One of the most outstanding players in this respect was Tommy Haas, who managed to recover from several serious injuries, each time bouncing back to his previous performance levels even after a prolonged absence from the sport (see Figure 6).

In addition to stating clear insights, this view also allowed our data scientist to see the entire underlying structure of the model in one view. Even when the visual tool offered no distinct declaration, it provided insightful views for the scientist to continue considering.

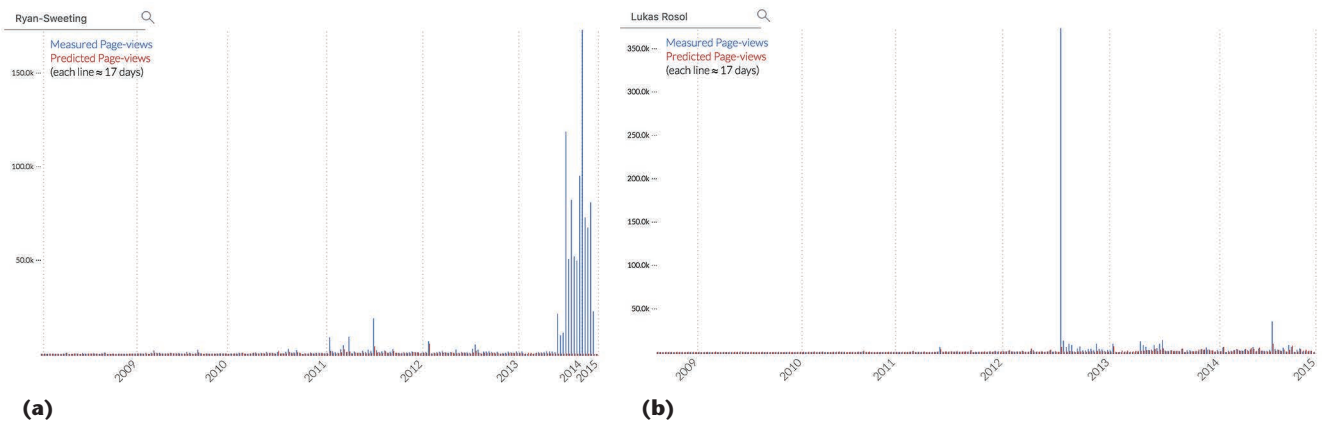


Figure 5. Non-tennis related changes in player popularity. (a) Ryan Sweetings married the actress Kaley Couco in 2013. (b) Lukas Rosol earned a burst of popularity when he defeated the then number two ranked player in the world, Rafael Nadal.

Thus far, we have described a combinatory design process that uses a method of incremental addition to create increasingly complex arrangements. Through this process, we can create new ways to see data and discover new insights. Interestingly, the scientists in our laboratory often work the other way around. They start with large data collections and try to find the smallest number of variables that explain the studied phenomena. That process reduces dimensions to the smallest set of variables for the most insightful explanations of a given phenomena. This problem-solving principal was first articulated as Occam’s razor (William of Ockham, 1287–1347), which states that among the competing hypothesis the one with the fewest assumptions should be selected.

While science searches for meaningful signals with the fewest assumptions, visualization can reconstruct the complexity of the investigated phenomena by drawing them together. The combination of these two inverse processes of drawing together and reduction to the smallest possible assumptions can generate meaningful syntheses that have yet to be fully considered.

In this sense, the design processes presented here can, in a scientific environment, help analysis along by finding outliers, data errors, or strange characteristics within the data. We hope they will be used as a tool to form questions about an undiscovered, or newly mapped, territory.

Similarly, Charles Sander Peirce introduced the idea of *abduction*.¹³ Epistemology differentiates between deduction, which draws a conclusion from the general to the specific, and induction, which draws a conclusion from specifics to the general. Abduction, on the other hand, goes from an observation to a hypothesis. Peirce first proposed the idea under the term “guessing.” As a method of drawing conclusions from observation of rela-

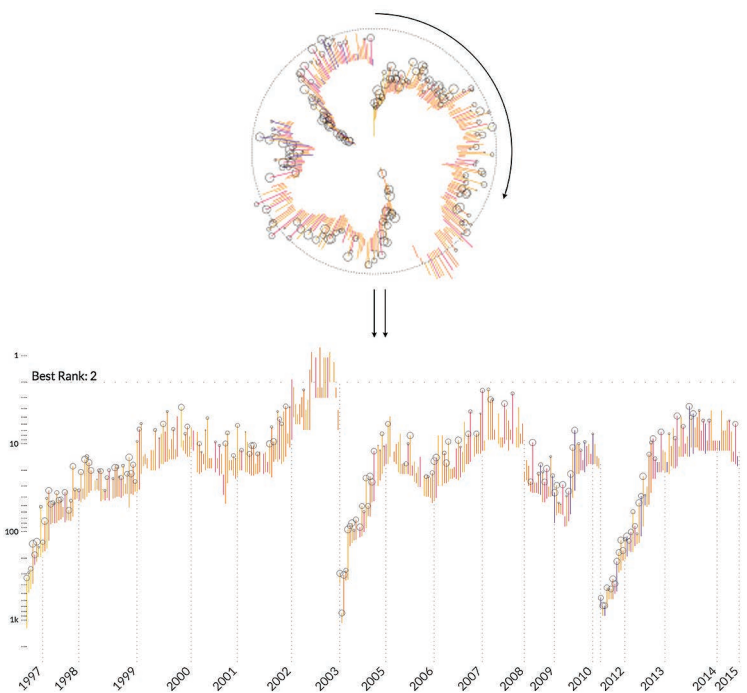


Figure 6. Performance visualization of Tommy Haas. This visual representation shows the course of Haas’ career and that he recovered from serious injuries three times.

tions, visualization facilitates reasoning through abduction. By visualizing the relationships between objects, we can create hypotheses that can then be tested by scientific methods.

In this sense, visualization suggests tools as hypothesis generators via a method of curiosity. This case study should be seen as a first step toward linking design and science. The scientific process of Occam’s razor in connection with the notion of design as drawing things together can create new insights and inspire new paths in knowledge production. ■


References

1. V. Flusser, *The Shape of Things: A Philosophy of Design*, C. Hanser, trans., Reaktion Books, 1999.
2. C. Mareis, *Theorien des Designs* [Theories of Design], Junius, 2014.
3. B. Latour, "A Cautious Prometheus? A Few Steps toward a Philosophy of Design (with Special Attention to Peter Sloterdijk)," *Proc. Ann. Int'l Conf. Design History Soc.*, 2008, pp. 2–10.
4. V. Flusser, *Vom Stand der Dinge: Eine kleine Philosophie des Design* [State of Affairs: A Small Philosophy of Design], Steidl Verlag, 1993.
5. K. Krippendorff, *The Semantic Turn*, CRC Press, 2006.
6. F.J. Anscombe, "Graphs in Statistical Analysis," *The American Statistician*, vol. 27, no. 1, 1973, pp. 17–21.
7. R. Borgo et. al., "Glyph-Based Visualization: Foundations, Design Guidelines, Techniques and Applications," *Eurographics State of the Art Reports*, 2013, pp. 39–63.
8. P. Fox and J. Hendler, "Changing the Equation on Scientific Data Visualization," *Science*, vol. 331, no. 6018, 2011, pp. 705–708.
9. T. Ropinski, S. Oeltze, and B. Preim, "Survey of Glyph-Based Visualization Techniques for Spatial Multivariate Medical Data," *Computers & Graphics*, vol. 35, no. 2, 2011, pp. 392–401.
10. H. Wickham et al., "Graphical Inference for Infovis," *IEEE Trans. Visualization and Computer Graphics*, vol. 16, no. 6, 2010, pp. 973–979.
11. B. Yucesoy and A.-L. Barabási, "Untangling Performance from Success," *EPJ Data Science*, vol. 5, no. 1, 2016; <http://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0079-z>.
12. W. Playfair, *The Commercial and Political Atlas and Statistical Breviary*, reprinted ed., Cambridge Univ. Press, 2005.
13. C.S. Peirce, *Selected Writings (Values in a Universe of Chance)*, Courier Corporation, 1958.

Kim Albrecht is a visualization researcher at the Center for Complex Network Research. Contact him at me@kimalbrecht.com.

Burcu Yucesoy is a postdoctoral research associate at the Center for Complex Network Research. Contact her at yucesoyb@gmail.com.

Contact department editors Bruce Campbell at bcampbel01@risd.edu and Francesca Samsel at figs@cat.utexas.edu.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.