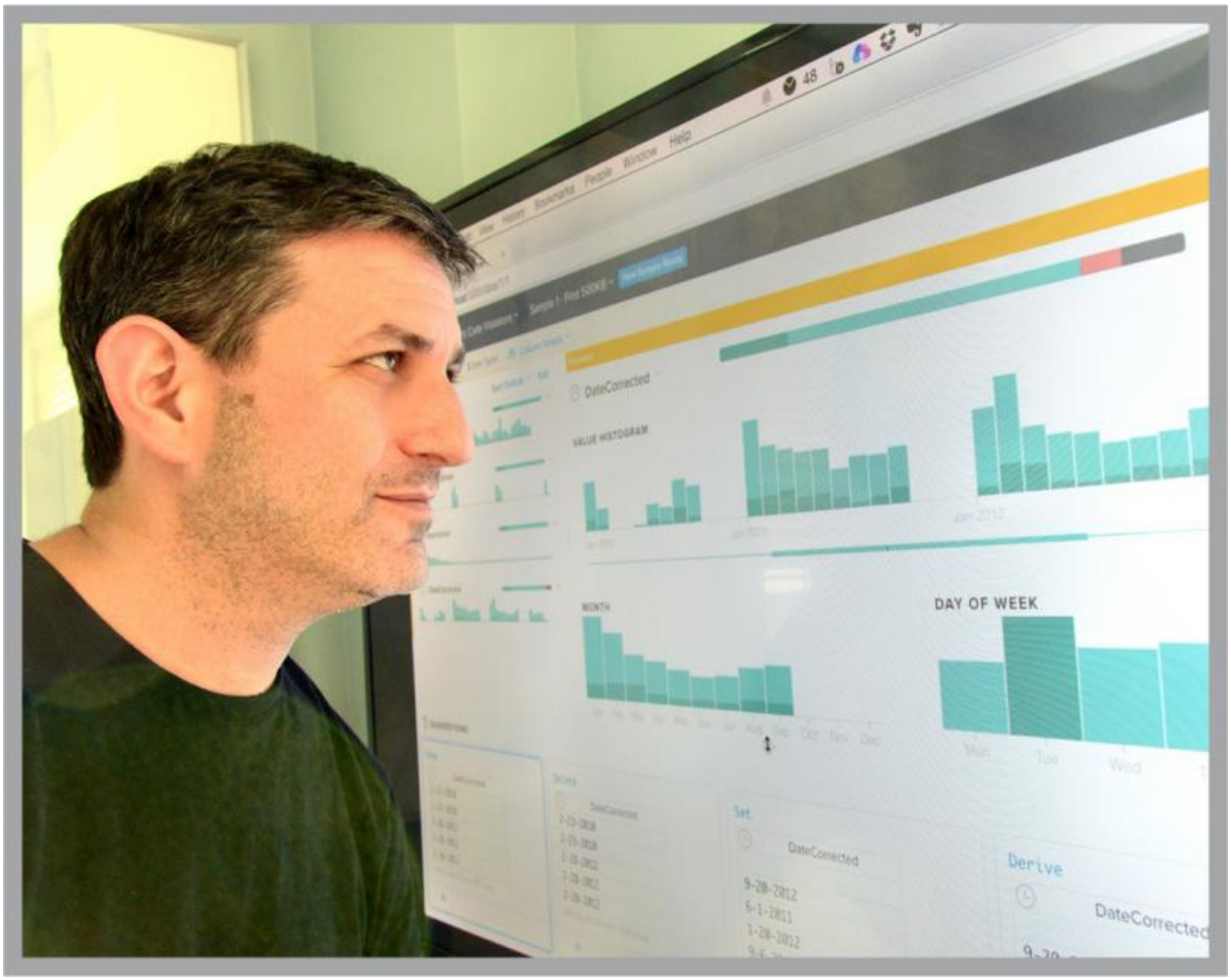
Seeing through the big data fog

December 17 2015, by Wallace Ravven



Joe Hellerstein and his students developed a new programming model for distributed computing which MIT Technology Review named one of the 10 technologies “most likely to change our world”.

A neuroscientist studies how stress affects the brain's ability to form new memories. Across the campus, another researcher looks for telltale signs of distant planets in a sliver of sky. What each of them seeks may lie hidden in an avalanche of data.

The same is true in industry, where data must be diced, sliced and analyzed to identify changes in customer behavior or the promise of new fabrication techniques.

Working the data so that it can yield to analysis regularly runs into a bottleneck—a human bottleneck, says Berkeley computer science professor Joe Hellerstein.

In 2011, Sean Kandel, a grad student working with Hellerstein and Stanford computer scientist Jeffrey Heer, interviewed three dozen analysts at 25 companies in different industries to ask them how they spend their time, what their "pain points" were, as Hellerstein says.

"It became very clear that the task of wrangling data takes up the lion's share of their time," Hellerstein says. "People come at data differently. They name data differently, or it may be incomplete. You have to sort this out. You find oddball data, and you don't know if it was input incorrectly or if it's a meaningful outlier. All this precedes analysis. It's very tedious."

Hellerstein, Heer and Kandel devised a software program to refine and speed the process. They called it, reasonably enough, Data Wrangler, and made it freely available online. Data Wrangler became the core of Trifacta, a startup they founded in 2012.

Trifacta provides a platform to efficiently convert raw data into more structured formats for analysis. Its flagship product for data wrangling enables data analysts to easily transform data from messy traces of the

real world into structured tables and charts that can reveal unsuspected patterns, or suggest new directions for analysis.

Trifacta was quickly adopted by dozens of companies, from Linkedin to Lockheed Martin, and typically provides a major productivity gain.

Joe Hellerstein and his postdoc Eugene Wu worked on designing a high-level language for crafting interactive visualizations of data. Wu is now a professor at Columbia University. Credit: Peg Skorpinski

"What used to take weeks suddenly takes minutes", Hellerstein says. "So you can experiment a great deal more with the data. This was far and away the most useful piece of research that I have been involved in."

In 2014, CRN, a high-profile communications technology magazine, placed Trifacta on its short list of The 10 Coolest Big Data Products.

GoPro, the company that makes wearable video recorders, was an early Trifacta client. On YouTube, GoPro videos run the gamut from a sky diver's death-defying leap to Kama, the surfing pig. (He prefers three-to four-foot waves.)

After sales of its recorders took off, GoPro moved into developing media software and other online services for customers. The company was soon inundated with coveted consumer data from devices, retail sales, social media and other sources.

GoPro built a data science team, which brought in Trifacta to clean up the data and present it in an intuitive and accessible format, so the less techy business people could use it to tailor services to customers and offer new products.

Hellerstein's research also targets software

developers who build Big Data systems—systems that may harness hundreds or thousands of computers to do their work. These "distributed computing" platforms, which also form the foundation of Cloud Computing, create major new hurdles for software engineering.

Code for a single computer is an ordered list of instructions, and most programming languages were designed for simple, orderly computing on a single machine.

With a distributed system, Hellerstein says, "If you force order, the machines spend all their time coordinating, and progress is limited by the slowest machine. Working around this with a traditional programming language is incredibly hard, and typically leads to all kinds of tricky bugs

and design flaws."

With his students, he launched the BOOM (Berkeley Orders of Magnitude) project to develop a new programming model for distributed computers that helps programmers avoid specifying the steps of a computation in a particular order. Instead, it focuses on the information that the program must manage, and the way that information flows through machines and tasks.

"The main result of the BOOM project is a 'disorderly' programming language called Bloom, which has enabled us to write complex distributed programs in simple, intuitive ways —with tens or hundreds of times less code than traditional languages," Hellerstein says.

In 2010, Bloom was recognized by MIT Technology Review as one of the 10 technologies "most likely to change our world."

Hellerstein has since used it in his courses on "Programming the Cloud" at Berkeley. It has been adopted by a number of research groups and forms the basis of a startup company in the Bay Area called Eve that Hellerstein advises.

As he describes his work to ease data wrangling and speed cloud programming, Hellerstein turns a small metal hammer in his hands. "It's true," he says. "I do like to tinker." Of course, he does way more than tinker. He's developing better tools for the trade.

Provided by University of California - Berkeley