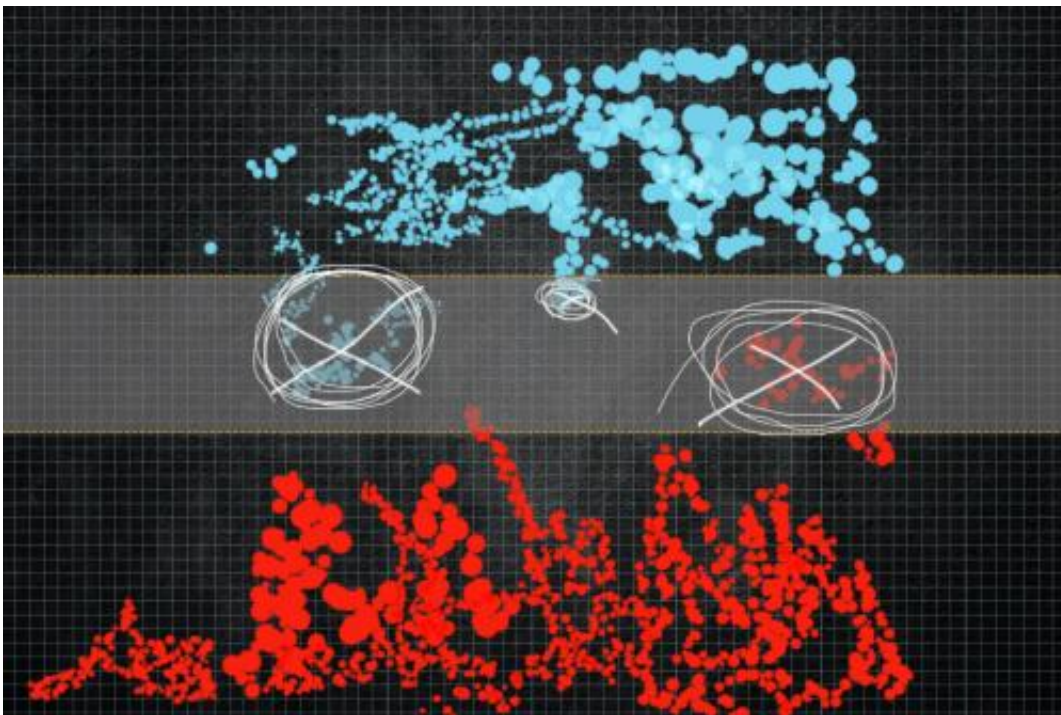


Data-visualization tool identifies sources of aberrant results and recomputes visualizations without them

August 15 2014, by Larry Hardesty



Credit: Christine Daniloff/MIT

In the age of big data, visualization tools are vital. With a single glance at a graphic display, a human being can recognize patterns that a computer might fail to find even after hours of analysis.

But what if there are aberrations in the patterns? Or what if there's just a

suggestion of a visual pattern that's not distinct enough to justify any strong inferences? Or what if the pattern is clear, but not what was to be expected?

The Database Group at MIT's Computer Science and Artificial Intelligence Laboratory has released a [data-visualization](#) tool that lets users highlight aberrations and possible patterns in the graphical display; the tool then automatically determines which data sources are responsible for which.

It could be, for instance, that just a couple of faulty sensors among dozens are corrupting a very regular pattern of readings, or that a few underperforming agents are dragging down a company's sales figures, or that a clogged vent in a hospital is dramatically increasing a few patients' risk of infection.

Big data is big business

Visualizing [big data](#) is big business: Tableau Software, which sells a suite of [visualization tools](#), is a \$4 billion company. But in creating attractive, informative graphics, most visualization software discards a good deal of useful data.

"If you look at the way people traditionally produce visualizations of any sort, they would have some big, rich data set—that has maybe hundreds of millions of data points, or records—and they would do some reduction of the set to a few hundred or thousands of records at most," says Samuel Madden, a professor of [computer science](#) and engineering and one of the Database Group's leaders. "The problem with doing that sort of reduction is that you lose information about where those output data points came from relative to the input data set. If one of these data points is crazy—is an outlier, for example—you don't have any real ability to go back to the data set and ask, 'Where did this come from and

what were its properties?"

That's one of the problems solved by the new visualization tool, dubbed DBWipes. For his thesis work, Eugene Wu, a graduate student in electrical engineering and computer science who developed DBWipes with Madden and adjunct professor Michael Stonebraker, designed a novel "provenance tracking" system for large data sets.

If a visualization system summarizes 100 million data entries into 100 points to render on the screen, then each of the 100 points will in some way summarize—perhaps by averaging—1 million data points. Wu's provenance-tracking system provides a compact representation of the source of the summarized data so that users can easily trace visualized data back to the source—and conversely, track source data to the pixels that are rendered by it.

The idea of provenance tracking is not new, but Wu's system is particularly well suited to the task of tracking down outliers in data visualizations. Rather than simply telling the user the million data entries that were used to compute the outliers, it first identifies those that most influenced the outlier values, and summarizes those data entries in human readable terms.

Best paper

Wu and Madden's work on their "Scorpion" algorithm was selected as one of the best papers of the Very Large Database conference last year. The algorithm tracks down the records responsible for particular aspects of a DBWipes visualization and then efficiently recalculates the visualization to either exclude or emphasize the data they contain.

If some of the points in the visualization suggest a regular pattern, the user can highlight them and mark them as "normal data"; if some of the

points disrupt that pattern, the user can highlight them and mark them as "outlier data"; and if the pattern is surprising, the user can draw the anticipated pattern on-screen.

Scorpion then tracks down the provenance of the highlighted points, and filters the provenance down to the subset that most influenced the outliers. Their paper introduces several properties about the specific computation that can be used to develop more efficient algorithms for finding these subsets.

Scorpion, Madden says, was partly motivated by a study conducted by a researcher at a Boston hospital, who noticed that a subset of patients in one of the hospital's wards was incurring much higher treatment costs than the rest. Any number of factors could have been responsible: the patients' age and fitness, the severity of their conditions, their particular constellations of symptoms, their health plans, or perhaps something as banal as their proximity to the hospital—nothing could be ruled out.

After six months of work, the researcher concluded that most of the variance in patients' treatment costs could be explained by a single variable: their doctors. It turned out that three doctors on the hospital staff, in an effort to leave no stone unturned, simply prescribed more interventions than their peers.

As an experiment, Wu and Madden turned Scorpion loose on the researcher's data. Within five minutes, it had concluded that the data point most strongly correlated with the increase in patients' treatment costs was the names of their doctors. Because it was combing through a massive data set and, like all big-data search algorithms, had to sacrifice some precision for efficiency, it couldn't pinpoint just the three doctors identified by the six-month study. But it did produce a list of 10 doctors most likely to be responsible for cost variance, and those three were among them. "You would at least know where to begin looking,"

Madden says.

More information: Paper: "Scorpion: Explaining Away Outliers in Aggregate Queries" [www.mit.edu/~eugenewu/files/pa ... /scorpion-vldb13.pdf](http://www.mit.edu/~eugenewu/files/papers/scorpion-vldb13.pdf)

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Data-visualization tool identifies sources of aberrant results and recomputes visualizations without them (2014, August 15) retrieved 20 June 2024 from <https://phys.org/news/2014-08-data-visualization-tool-sources-aberrant-results.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.