Simplifying Data Analysis and Making Sense of Big Data

Wallace Ravven, University of California, Berkeley

Ben Recht is looking for problems. He develops mathematical strategies to help researchers, from urban planners to online retailers, cut through blizzards of data to find what they're after. He resists the "needle in the haystack" metaphor because, he says, the researchers, engineers and business people he has worked with usually don't know enough about their data to reach their goal.

"My work tries to incorporate as much expert knowledge as possible into data analysis. Once you codify this expertise, the needle often turns out to be a pitchfork."

His work is all about simplifying data analysis, whether the payoff is buried deep in mountains of data or stranded by incomplete data.

He is particularly keen on generating ways to solve problems common to a range of investigations, and he has already shown that the same mathematical model can overcome the computational challenges of determining a molecule's structure, predicting traffic flow or an online shopper's habits.

His dual appointment in the departments of statistics [1], as well as in electrical engineering and computer sciences [2], reflects his overlapping fields of expertise. The research is theoretical, but it's driven by the need to unsnarl data analysis in the real world, so Recht seeks out roadblocks wherever they may be.

"My favorite part of the job is the social aspect," he says, looking like most of his grad students in a black t-shirt, and his MacBook Air balanced in his hands, at the ready. "I can't do my job in isolation. If I don't know their problems, I wouldn't have anything to work on."

"Software analysis is a mature field, he says, but with data analysis, "everybody has their favorite technique."

"We're trying to understand if there's a single tool that all can use in data analysis. Is there a way of looking at analysis of huge amounts of data in which you don't have to build a giant new factory from scratch each time?

Even when researchers have amassed a huge amount of data, they know when they're missing vital information that is too computationally demanding to get, Recht says.

He is well known for devising a strategy to handle "noisy" and incomplete data. "We developed an algorithm that allows you to solve a range of different complex computational problems with less data."

Say you're looking for the structure of a protein with a hundred atoms. If you knew all of the distances between all pairs of atoms, classic algorithms from geometry would take these distances and determine the three-dimensional structure of the protein.

"But there are 5000 pairs of distances," Recht says, "and these might be hard to measure all together. Our work roughly showed that if you identify just a small set of a few hundred pairs of distances, you can still reconstruct the 3-D structure exactly."

In molecular reconstruction, then, one can "get away with a few pairs of distances,' he says. And to assess online marketplaces: "You can take advantage of the fact that shoppers can be pigeon-holed into a few types. It turns out that mathematically, both of these examples can be understood in the same framework."

Analysis of Big Data can't be approached from any one discipline. "You can't just use statistics or math or the tools of computer science. But I love to work with multidisciplinary collaborators."

That's why the NSF-funded Algorithms, Machines, and People Lab (AMP) [3] is ideal, he says. The lab consists of eight PI's and their grad students working together in fields from databases and computer architecture to networking, and machine learning.

"We work at the intersection of algorithms, machine learning and collaborative crowd sourcing. We're reaching into lots of different research areas on campus, and way beyond Berkeley.

"Lots of the research is open source. Everybody thinks, 'I'm putting in time and I should share my research 'cuz some other obsessive nerd will put in time and share it with me.' There's a communal altruism."

Data-intensive and data-driven approaches to research span the Berkeley campus, from investigations of artificial intelligence and biology to linguistics and urban planning. The accelerating search for new approaches to aggregating and analyzing data now has a new

focus — the Berkeley Institute for Data Science [4], part of a $38 million collaboration supported by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation.

Recht was recently honored by the White House with a Presidential Early Career Award for Scientists and Engineers, recognizing some of the most promising young researchers. But Big Data science doesn't tap all of his talents. A guitarist and composer, he has recorded 10 CDs with his music partner Isaac Sparks, who plays the turntable. Yes, plays the turntable — not in quotes. "The turntable needs to be formally recognized as an instrument," he insists, and directs any doubter to the Wikipedia page on Turntablism [5].

In some of the pieces, pops and crackles recorded from old LP's take the place of drum beats. To get that sound, "We buy most of our records at thrift stores," Recht says. "Sometimes, the crappier, the better." He and Sparks have been composing and recording for more than 10 years.

Although he briefly considered a career in computer music when he was a grad student, Recht pursues his two passions on separate tracks. "I don't try to apply my research to the music. It's just for pure pleasure. But I love them both."

Still, both his research and music depend on collaborations. As he says, he can't do his work in isolation – or his play.

- Recht's ongoing music forays can be found at his web site, The Fun Years: http://www.thefunyears.com [6]
- Berkeley Institute for Data Science: http://vcresearch.berkeley.edu/datascience [4]
- AMP Lab: https://amplab.cs.berkeley.edu [3]
- Ben's faculty webpage: http://www.eecs.berkeley.edu/~brecht/ [7]

Source URL (retrieved on *09/04/2015 - 10:34am*):
http://www.scientificcomputing.com/news/2014/03/simplifying-data-analysis-and-making-sense-big-data

Links:
[1] http://statistics.berkeley.edu/
[2] http://www.eecs.berkeley.edu/
[3] https://amplab.cs.berkeley.edu/
[4] http://vcresearch.berkeley.edu/datascience
[5] http://en.wikipedia.org/wiki/Turntablism
[6] http://www.thefunyears.com
[7] http://www.eecs.berkeley.edu/%7Ebrecht/