Stanford eCorner

Data Will Help Solve Major Problems

Mike Olson, *Cloudera*

November 13, 2013

Video URL: http://ecorner.stanford.edu/videos/3226/Data-Will-Help-Solve-Major-Problems

Cloudera Co-Founder Mike Olson offers real-world examples of how big data will play a role in solving business and societal challenges, from battling credit fraud to working to improve clean water access and more effective medical care.

Transcript

We make a very good living at Cloudera, helping companies run workloads. And I will describe some of those workloads in a minute. We sell to banks, we sell to insurance companies, we sell to hospitals, right. But it is my deep conviction that as a society, we're going to attack really important problems in the next decade. We've got to figure out how to produce and distribute clean water to people in a warming world, and we're going to figure out ways to do that by looking at data in new ways. 7 billion people on the planet today, forecast to be 9 billion people in 2050, that's 2 billion new bodies with no available planet to find new farmland on. We have to figure out a way to produce food more efficiently. And we're going to do that with data. Meaningful cancers will become manageable chronic diseases in our lifetime, will no longer be a death sentence, but will be diseases that, like AIDS in rich countries now, you're going to actually manage and live with for a long period of time. And we're going to do that by understanding the disease by using data to understand its progress better than ever before.

Data really is going to matter to us as a society in ways that it hasn't. IT is going to be a social good. I believe this deeply and we actually see this happening in our installed base. So while nobody woke up feeling that way about relational systems. Look, I know we're not going to cure cancer at Cloudera, we're not going to feed 2 billion people, but we're going to make software that lets the planet attack those problems in new ways. I'll give you some concrete examples of use cases we see, some fairly pedestrian, some a little more inspirational. If you are a credit card provider, you care about fraud. There are bad guys all the time stealing credit cards, using them on the web, trying to get goods and not get caught, that's always been a problem and those firms have been looking for ways to detect and prevent fraud for a long time. They've been using techniques by the way like machine learning, at small scale, looking at the last week's or month's worth of transactional activity. It turns out if you can feed those models a decade's worth of data, you can see patterns, you can learn behaviors that were invisible in small amounts of data.

And it is glib example, but flip a coin three times, you know nothing about that coin. Flip a coin a thousand times; you get an idea whether or not that's a fair coin. It turns out that lots of data yields value, disproportionally, right. Smarter algorithms are good, but more data to train your existing algorithms even better. One of our customers a global credit card - basically a global card processing company, had been doing fraud detection at a small scale for a long time. They brought us in just because they wanted to store way more data, way more cheaply than ever before, LAN the last 10 years' worth of transaction data and that was for cost reasons, that's all. Save a few bucks. Once they did that, they had all that data on spinning disks, some of their analysts said you know what let's turn our fraud models loose on this data, tweak them a little bit, but let's feed them 10 years' worth of the data. They discovered the single largest instance of fraud in the company's history as a side effect of saving money on storage. So just aggregating data has allowed them to do an old thing in a very powerful new way.

One other example I will hit and then I will let you either direct with more or go on to another question. So one of our customers is a company named Explorys Medical and they're based in Cincinnati, they're - I'm sorry they're based in Cleveland, they are spinout of the Cleveland Clinic. Much in the news these days of course are Obamacare and healthcare.gov and the challenges we have around that. And whatever your politics are, I think you will agree that better patient outcomes more reliably at lower costs that will be a good thing. So you get sick, you go see your doc, you're filling a form when you get in, maybe whacks on your knees, takes your blood pressure, talks to you about your symptoms, writes up a script, sends you home. You fill that script, you take the pills that you're prescribed for a week or two, you don't feel much better, right. You go back, see your doctor again, maybe some more invasive testing, blood draw or other body samples happen, maybe some imaging, maybe calling a specialist, look at another doctor to give you some more feedback. This time maybe you get some physical therapy, you get a different prescription, you follow that course of action, I'm rooting for you, so I'm going to say you get better in this one. If you could string that whole story together, you'd have a pretty good picture of your illness, right. The progress over multiple weeks, lots of different data types in there, lots of interactions, there is interstitial delays, the amount of time between your doctor visits, that's significant.

Now imagine you could do that for an entire hospital network's worth of patients. You could use machine learning to realize that Dr. Jones was getting disproportionately good outcome from women between 35 years and 50 years old, by following a specific course of treatment when they present with similar symptoms and you could use Dr. Jones' outstanding performance to make recommendations to Dr. Smith and others to adapt their treatment. That is exactly what Explorys is doing. They're looking at the course of the care that aggregate - basically large aggregations of patients are getting and the outcomes, and thereby directing healthcare providers to get better outcomes faster. Healthcare actually is going to be a big, big area for exploitation of data in this way.