Débat 3

Beyond Big Data[1]

Hal R. Varian
Google

There is now a computer in the middle of most economic transactions. These computer-mediated transactions enable data collection and analysis, personalization and customization, continuous experimentation, and contractual innovation. Taking full advantage of the potential of these new capabilities will require increasing sophistication in knowing what to do with the data that are now available.

Keywords: big data, machine learning, privacy, personalization, experimentation, monitoring.

– A billion hours ago, modern *homo sapiens* emerged.
– A billion minutes ago, Christianity began.
– A billion seconds ago, the IBM PC was released.
– A billion Google searches ago … was this morning.

There is a lot of press about "big data": how important it is, how powerful it is, and how it will change our lives. All that may be true, but I believe that there is something more fundamental going on, which I refer to as "computer-mediated transactions." Due to the dramatic cost decrease in computers and communication, there is a computer in the middle of virtually every transaction. This computer could be as simple as a cash register or as complex as a data center. These computers were mostly put in place for accounting reasons, but now that they are available these computers have several other uses.

I want to talk to you about four such uses:
– Data extraction and analysis.
– Personalization and customization.
– Continuous experiments.
– New kinds of contracts due to better monitoring.

The first one data extraction and analysis is what everyone is talking about when they talk about big data. It is important, but I think that the other three uses go beyond big data and will, in time, become even more important than the first. But let's start at the beginning and look at them in order.

1. Data Extraction and Analysis

According to published reports, Google has seen 30 trillion URLs, crawls over 20 billion of those a day, and answers 100 billion search queries a month [Singhal 2012]. Ordinary databases just can't handle these magnitudes, so we have had to develop new types of databases that can store data in massive tables spread across thousands of machines and can process queries on more than a trillion records in few seconds.

We published descriptions of these tools in the academic literature, and independent developers have created open source tools that have similar functionality. These tools are now widely available and run on cloud computing engines such as Amazon Web Services, Google Computer Engine, and other services.

[1] Presented at the NABE Annual Meeting, September 10, 2013, San Francisco, CA.

From the economic point of view, what was previously a fixed cost (deploying and managing a data center capable of dealing with massive data) is now a variable cost. As any economist knows, if you lower the barriers to entry you will get lots of new entrants and we have seen a number of startups in this area.

But tools for data manipulation are only part of the story. We have also seen significant developments in *methods* for data analysis that have emerged from the machine learning community.

Nowadays we hear a lot about "predictive analytics", "data mining" and "data science". The techniques from these subjects, along with some good old-fashioned statistics and econometrics have allowed for deeper analysis of these vast data sets, enabled by computer-mediated transactions.

There have also been significant developments in open source programs that can be used to apply these tools, such as the R language, Weka (Waikato Environment for Knowledge Analysis), and others. One of most important features of these languages is the thriving communities of users who provide peer support on the web. Given the fact that cloud hardware, database tools, analysis tools, and developer support is now widely available it is not surprising that we have seen many new entrants in the data analysis area.

Back in 2006, NetFlix realized that 75 percent of movie views in its library were driven by recommendations. They created the "NetFlix Prize" of $1 million that would be awarded to the group that developed the best machine learning system for recommendations, as long as it improved the current version by at least 10 percent. They provided training data of about 100M ratings, 500,000 users, and 1,800 movies. A year later, the prize was won by a team that blended 800 statistical models together using model averaging. The success of the NetFlix challenge led to the establishment of a startup named Kaggle, who will set up NetFlixlike challenges. (Note: I am an investor and an adviser to Kaggle.)

Nowadays, there are many organizations that have interesting data but no internal expertise in data analysis. As the same time, there are data analysts all over the world that have expertise but no data (and could use some money). Kaggle puts the two sides of the market together. They now have 114,000 data scientists who tackle the submitted problems. Their motto is "We make data science a sport." Here are some examples of their projects:
– Heritage Health Prize: $3 million to predict hospital readmits.
– Gesture recognition for Microsoft Kinnect: $10,000.
– GE flight optimization: $250,000.
– Belkin energy disaggregation for appliances: $25,000.
– Recognizing Parkinson's disease from smartphone accelerometer data: $10,000.
– ...and many more.

When you combine Data + Tools + Techniques + Expertise, you can solve a lot of hard problems!

2. Personalization and customization

Nowadays, people have come to expect personalized search results and ads. If you ask Google for "pizza near me" you will get back what you expect. When you go to Amazon, they recommend products just for you. The story has it that Jeff Bezos recently signed on to his

Amazon account and saw a message that said "Based on your recent purchases, we recommend The *New York Times*, the *Chicago Tribune*, and the *Los Angeles Times*."

These personalized searches, services and ads have the potential of revolutionizing marketing. The average marketing cost per car sold in the US is about $650. But much of that marketing expenditure is wasted. Why show ads to someone who just bought a car? Larry Page used to say that the trouble with Google was that you had to ask it questions: he thought Google should know what you want and tell it to you before you ask the question. We all thought he was joking but Larry's vision has been realized by Google Now, an application that runs on Android phones. One day my phone buzzed and I looked at a message from Google Now. It said: "Your meeting at Stanford starts in 45 minutes and the traffic is heavy, so you better leave now." The kicker is that I had never told Google Now about my meeting. It just looked at my Google Calendar, saw where I was going, sent my current location and destination to Google Maps, and figured out how long it would take me to get to my appointment given current traffic conditions.

Some people think that's the coolest thing in the world, and others are just completely freaked out by it. The issue is that Google Now has to know a lot about you and your environment to provide these services. This worries some people. But, of course, I share highly private information with my doctor, lawyer, accountant, trainer, and others because I receive identifiable benefits and I trust them to act in my interest. If I want to get a mortgage, I have to send the bank two years of income tax returns, a month of paychecks, a printout of my net worth, and dozens of other documents. Why am I willing to share all this private information? Because I get something in return: the mortgage.

One easy way to forecast the future is to predict that what rich people have now, middle-class people will have in five years, and poor people will have in ten years. It worked for radio, TV, dishwashers, mobile phones, flat screen TV, and many other pieces of technology.

What do rich people have now? Chauffeurs? In a few more years, we'll all have access to driverless cars. Maids? We will soon be able to get housecleaning robots. Personal assistants? That's Google Now. This area will be an intensely competitive environment: Apple already has SIri and Microsoft is hard at work at developing their own digital assistant. And don't forget IBM's Watson.

Of course there will be challenges. But these digital assistants will be so useful that everyone will want one, and the scare stories you read today about privacy concerns will just seem quaint and old-fashioned.

3. Experiments

As all economists are aware, correlation is not the same thing as causation. Observational data – no matter how big it is– can usually only measure correlation, not causality. To take a trivial example, suppose we observe that there are more police in areas with higher crime rates, can we conclude that police cause crime? More specifically, does the correlation mean that if you assign more police to an area will you get more crime? You may have a very good model that can predict the crime rate by precinct depending on how many police have been assigned to that precinct, but that model could fail miserably in estimating how the crime rate changes when you add more policemen to a precinct. It is quite possible to see a positive relationship based on the observational data and a negative relation based on an actual experiment.

Likewise, bigger fires have more firemen; do more firemen cause bigger fires? If you assign more firemen will the fire get bigger? Or, to take an example closer to home, what about the impact of advertising on sales? Consider this probably apocryphal question asked to a marketing manager: "How do you know increased advertising will generate more sales?" "Look at this chart," he responded, "Every December I increase ad spend, and every December I get more sales." Companies are very interested in using their big data to estimate demand functions: "if I cut my price, how will the amount sold change?" This is actually one of the most common requests for big data analysts. Alas, usually observational data cannot answer this question. Suppose my sales fall when disposable income is low, so I respond by cutting price. Conversely, sales rise when disposable income is high, so I raise price. Over time, we see high prices associated with high sales and low prices with low sales; so we get an upward sloping demand function! The historical relationship between price and quantity would not be a good guide for future pricing decisions.

This is where econometrics comes in: our goal is to estimate the causal response from changing price, which will often be different from the historical relationship between price and sales. In these simple examples the problem is obvious, but in other cases it can be tricky to distinguish causal effects from mere correlation.

What's the solution? Experiments. These are the gold standard for causality. More specifically, what you want are randomly assigned treatment control experiments. Ideally these would be carried on continuously. This is pretty easy to do on the web. You can assign treatment and control groups based on traffic, cookies, usernames, geographic areas, and so on. Google runs about 10,000 experiments a year in search and ads. There are about 1,000 running at any one time, and when you access Google you are in dozens of experiments.
What types of experiments? There are many:
– user interface experiments;
– ranking algorithms for search and ads;
– feature experiments;
– product design;
– tuning experiments.

All of these can be ongoing experiments or special purpose experiments. Google has been so successful with our own experiments that we have made them available to our advertisers and publishers in two programs. The first, Advertiser Campaign Experiments (ACE), allows advertisers to experiment with bids, budgets, creatives, and so on, in order to find the optimal settings for their ads.

The second, Contents Experiment Platform, is part of Google Analytics; it allows publishers to experiment with different web page designs to find the one(s) that perform best for them.
So big data is only part of the story. If you really want to understand causality, you have to run experiments. And if you run experiments continuously, you can continuously improve your system.

In 1910, Henry Ford and his colleagues were on the factory floor every day, fine-tuning the assembly line –a method of production that revolutionized manufacturing. In the 1970s, the buzz in manufacturing was "kaizen" the Japanese term for continuous improvement. Now we have "computer kaizen," where the experimentation can be entirely automated. Just as mass production changed the way products were assembled and continuous improvement changed how manufacturing was done, continuous experimentation will improve the way we optimize business processes in our organizations.

4. Monitoring and contracts

My last example of how computer mediated transactions affects economic activity has to do with contracts. Contracts can be very simple: “You give me a *latte* and I will give you $2.” This contract is easily verified: I can see that I got my *latte*, you can see that you got your $2. However, there are other contracts that are not so easy to verify. When I rent a car, somewhere in the fine print there is a statement to the effect that I will operate the car in a safe manner. But how can they verify that? There used to be no way, but now insurance companies can put vehicular monitoring systems in the car [Scism 2013 and Stross 2012]. They can use these systems to verify whether or not you are fulfilling your part of the contract. They get lower accident rates and I get lower prices.

Here is another car example: “I will lease a car to you, if you send in your monthly payments on time.” What happens if you stop sending in the monthly payments? The lender will likely send out a “repro man” to repossess the car. Nowadays it’s a lot easier just to instruct the vehicular monitoring system not to allow the car to be started and to send a message to the leasing telling it where to go to pick it up.

The examples of contractual innovation aren’t all about misbehavior. Suppose an advertiser went to a newspaper and said “I will buy an ad in your publication, as long as I only have to pay for the people that saw my ad and then came to my store.” If would be nearly impossible to verify performance in that contract. But now, in pay per click advertising, it is common for advertisers to pay only for those advertisers that click through to their web site.

Because transactions are now computer-mediated, we can observe behavior that was previously unobservable and write contracts on it. This enables transactions that were simply not feasible before. Let me give you an example: Premise Data Corporation is a small startup funded by Google Ventures that collects economic data. (Note: I am an adviser to Premise.) Suppose you want to track the price of pork in Shanghai. Premise will send 20 college students out with mobile phones to photograph the price of pork in hundreds of stores in Shanghai. The nice thing is that they can prove that they actually went to the store and weren’t sitting a coffee shop typing prices into a spreadsheet. In fact, their mobile phones provide the proof: a geo-located, time-stamped photograph of the pork price in a store. You can do the same thing with other data collections: counting cars in parking lots, people in shopping centers, or traffic through an intersection. Premise uses crowdsourcing + smartphone + camera + timestamp + geolocation, all of which can be used to verify data integrity.

Premise is using the latest technology, but there are historical examples of computer-mediated transactions that date back more than a century. The classic cash register was invented by saloon-keeper James Ritty in 1883, who used the name the “incorruptible cashier” for his invention. What kind of monitoring did it provide? First, it went ka-ching when the cash drawer was opened, so the cashier knew to pay attention to what was going on. Secondly, it used a paper tape to record all transactions, so they could be reconciled with cash flow and product flow. Some historians claim that credit cash registers were a critical invention for economic growth since they with allowed retail businesses to hire clerks and cashiers from outside the trusted family.

Computer mediated transactions have enabled new business models that were simply not feasible before. Here are two more examples of contractual innovation.

– **Uber.** This service allows you to load an app on your phone and use it to call a black car sedan when you need it. You see the car on a map on your phone and literally watch it come to you. You get in the car, go to where you're going, and automatically pay by using your mobile phone. The entire transaction is monitored. If something goes wrong with the transaction, you can use the computerized record to find what went wrong. Uber is giving both the driver and the passenger a "no surprise" experience via the identity verification due to the computer mediation.

– **Air BnB.** This company allows you to rent out a spare room, an apartment, or an in-law unit. Just as with Uber, computers verify identity on both sides of the transaction and each side of the transaction can rate the performance of the other side. What was previously hard to find information about reputation of the renter and the rentee has now become easily accessible, allowing people to trust more because verification has been automated.

Summary

Computer mediated transactions can make a big difference to economic performance. They allow for analysis, personalization, experimentation, and monitoring. These capabilities allow for new transactions that were not feasible before. Kaggle, Premise, Uber, Air BnB are all examples, but there are many more.

Many companies have the data: they just don't know what to do with it. Missing ingredients: data tools (easy), knowledge (hard), and experience (very hard). Three years ago, I said "statistics will be the sexy job of the next decade", which endeared me to statisticians everywhere. Since then, applications to the Stanford Statistics Department have tripled. Is this causation or correlation? I don't know, and they don't know; but I'm willing to take credit anyway.

References

Lohr, Steve. 2013. "How Surveillance Changes Behavior: a Restaurant Worker's Case Study", *New York Times*. August 26.

Scism, Leslie. 2013. "State Farm is There: As you Drive", *Wall Street Journal*. August 15.

Singhal, Amit 2012. "Breakfast with Google's Search Team", http://www.youtube.com/watch?v=8a2VmxqFg8A#!

Stross, Randall. 2012. "So You're a Good Driver? Let's Go to the Monitor", *New York Times,* November 24

Varian, Hal R. 2010. "Computer Mediated Transactions," *American Economic Review,* 100(2) 110.