The Data Revolution and Economic Analysis[1]

Liran Einav, Stanford University and NBER
Jonathan Levin, Stanford University and NBER

Prepared for the NBER Innovation Policy and the Economy Conference, April 2013

Abstract

Many believe that "big data" will transform business, government and other aspects of the economy. In this article we discuss how new data may impact economic policy and economic research. Large-scale administrative datasets and proprietary private sector data can greatly improve the way we measure, track and describe economic activity. They also can enable novel research designs that allow researchers to trace the consequences of different events or policies. We discuss some of the challenges in accessing and making use of these data. We also consider whether the big data predictive modeling tools that have emerged in statistics and computer science may prove useful in economics.

[1] Author disclosures will be available on the NBER website. We thank Erin Scott and Scott Stern for comments on an earlier draft. We are grateful for research support from the NSF and the Alfred P. Sloan Foundation.

"There was five exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days, and the pace is increasing."

- Eric Schmidt, former CEO of Google, 2010.[2]

1. Introduction

The media is full of reports about how big data will transform business, government and other aspects of the economy. The term "data scientist" was hardly heard a few years ago. By last fall, reporters were arguing that the Obama campaign's data scientists had provided an important fundraising and advertising advantage in nothing less than the presidential election.[3] As economists who happen to live and work in the epicenter of the data revolution, Silicon Valley, we have wondered for some time about how these developments might affect economics, especially economic research and policy analysis. In this article, we try to offer some thoughts.

We start by trying to describe what is meant by big data, and what about it is new from the perspective of economists, who have been sophisticated users of data for a long time. We then turn, in Section 3, to the uses of big data that have received the most attention, namely the identification of novel patterns in behavior or activity, and the development of predictive models, that would have been hard or impossible with smaller samples, fewer variables, or more aggregation. Variations on these types of data analytics have had a major impact on many industries, including retailing, finance, advertising and insurance.

Section 4 and Section 5 then engage in a discussion of how new data may affect economic policy and economic research. From an economic policy perspective, we highlight the value of large administrative data sets, the ability to capture and process data in real time, and the potential for

[2] The first number in this quote seems to be drawn from a study led by Peter Lyman and Hal Varian at Berkeley (http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/) which estimated that the worldwide production of original stored information *in 2002* was in the range of 5 exabytes. The Berkeley study estimated that the production of stored information was increasing by around 30% per year. At that rate of increase, we currently would be creating 5 exabytes of information every 18 days.

[3] A Google search for "`Obama campaign' AND `data mining'" returns over 50,000 results (as of Feb 17. 2013).

improving both the efficiency of government operations and informing economic policy-making. From an economic research perspective, we emphasize how large, granular datasets can enable novel research designs, and illustrate with some examples from recent work. We also point to how researchers may be able to observe many more of the potential consequences or particular economic events or policies. Lastly, we consider whether the big data tools being developed in statistics and computer science, such as statistical learning and data mining techniques, will find much application in economics. So far they have not, but we suggest why that might change.

The final section discusses the challenges associated with big data, both in terms of providing access and computing ability to work with it, as well as in terms of training economists to work with large data sets and the various hardware, software, and statistical tools that are commonly required for it.

One point to mention at the start is that our discussion will be in limited in a variety of ways. In particular, we are going to describe some of the potential uses of big data, without discussing the potential abuses, such as threats to privacy, or malfeasance, of the possibility for the state to use detailed data on its citizens in undesirable ways. These are important issues in thinking about creating and managing large datasets on individuals, but not the topic of this paper.

2. What is the Big Deal with "Big Data"?

Even twenty or thirty years ago, data on economic activity was relatively scarce. In just a short period of time, this has changed dramatically. One reason is the growth of the internet. Practically everything on the internet is recorded. When you search on Google or Bing, your queries and subsequent clicks are recorded. When you shop on Amazon or eBay, not only every purchase, but every click is captured and logged. When you read a newspaper online, watch videos, or track your personal finances, your behavior is recorded. The recording of individual behavior does not stop

with the internet: text messaging, cell phones and geo-locations, scanner data, employment records, and electronic health records are all part of the data footprint that we now leave behind us.

A specific example may be illustrative. Consider the data collected by retail stores. A few decades ago, stores might have collected data on daily sales, and it would have been considered high quality if the data was split by products or product categories. Nowadays, scanner data makes it possible to track individual purchases and item sales, capture the exact time at which they occur and the purchase histories of the individuals, and use electronic inventory data to link purchases to specific shelf locations or current inventory levels. Internet retailers observe not just this information, but can trace the consumer's behavior around the sale, including his or her initial search query, items that were viewed and discarded, recommendations or promotions that were shown, and subsequent product or seller reviews. And in principle these data could be linked to demographics, advertising exposure, social media activity, offline spending, or credit history.

There has been a parallel evolution in business activity. As firms have moved their day-to-day operations to computers and then online, it has become possible to compile rich datasets of sales contacts, hiring practices, and physical shipments of goods. Increasingly, there are also electronic records of collaborative work efforts, personnel evaluations and productivity measures. The same story also can be told about the public sector, in terms of the ability to access and analyze tax filings, social insurance programs, government expenditures and regulatory activities.

Obviously, this is a lot of data. But what exactly is new about it? The short answer is that data is now available faster, has greater coverage and scope, and includes new types of observations and measurements that previously were not available. A key aspect of such modern datasets is that they have much less structure, or more complex structure, than the traditional cross-sectional, time-series or panel data models that we teach in our econometrics classes.

Data is available in real time. The ability to capture and process large amounts of data in real-time is crucial for many business applications. But at least for now, it has not been used much for economic research and policy. That is perhaps not surprising. Many economic questions are naturally retrospective, so that it is more important for data to be detailed and accurate rather than available immediately. However, we will discuss below some ways in which real-time data may prove to be useful for policy or research.

Data is available at larger scale. A major change for economists is the scale of modern datasets. When we were in graduate school, we worked with data that contained hundreds or thousands of observations. Because datasets often were small, statistical power was an important issue. Nowadays, datasets with tens of millions of distinct observations, and huge numbers of covariates, are quite common. In many cases, the large number of observations can make statistical power much less of a concern. Of course, having a lot of observations is no panacea: even with millions of observations, the relevant variation may be at the state or county level, or it may be desirable to use fixed effects or other methods that control for heterogeneity but also reduce statistical power.

Data come with less structure. The scope of recorded information has also expanded. In the retail example above, the information available about a consumer might include her entire shopping history. With this information, it is possible to create an almost unlimited set of individual-level behavioral characteristics. While this is very powerful, it is also challenging. In econometrics textbooks, data arrives in "rectangular" form, with N observations and K variables, and with K typically a lot smaller than N. When data arrive in its raw form of a digital recording of a sequence of events, with no further structure, there are a huge number of ways to move from that recording into a standard "rectangular" format. Figuring out how to organize unstructured data and reduce its dimensionality, and assessing whether the way we do this matters, is not something that most

empirical economists have been taught or have a lot of experience with, but it is becoming a very common challenge in empirical research.

Data is available on novel types of variables. Much of the data now being recorded is on activities that previously were very difficult to observe. Think about email, or geo-location data that records where people have been, or social network data that captures personal connections. Eventually, these records may well prove to be an amazing boon for social science researchers. For instance, most economists would agree that social connections play an important role in job search, in shaping consumer preferences and in the transmission of information.

The challenge is in figuring out how to make effective use of these data, which may have novel structures. Traditional econometrics methods generally assume that the relationship between observations is relatively simple: the observations are independent, or grouped as in panel data, or linked by time. But individuals in a social network may be interconnected in highly complex ways, and indeed the point of econometric modeling may be to uncover exactly what are the key features of this dependence structure. Developing methods that are well-suited to these settings is an interesting challenge for econometrics research (Imbens et al., 2011).

3. Big Data and Predictive Modeling

The most common uses of "big data" by companies are for tracking business processes and outcomes, and for building a wide array of predictive models. While business analytics is a big deal and surely has improved the efficiency of many organizations, predictive modeling lies behind many of the most striking information products and services introduced in recent years.

Many examples will be familiar to all readers: Amazon and Netflix recommendations, which rely on predictive models of what book or movie an individual might want to purchase, Google's search results and news feed, which rely on algorithms that predict the relevance of particular web pages

or articles, or Apple's auto-complete, which tries to predict the rest of one's text or email. Online advertising and marketing rely heavily on automated predictive models that attempt to target individuals who might be particularly likely to respond to offers.

The application of predictive algorithms extends well beyond the online world. In health care, it is now common for insurers to adjust payments and quality measures based on "risk scores", which are derived from predictive models of individual health costs and outcomes. An individual's risk score is typically a weighted sum of health indicators that identify whether an individual has different chronic conditions, with the weights chosen based on a statistical analysis. Credit card companies use predictive models of default and repayment to guide their underwriting, pricing and marketing activities. One Palo Alto company, Palantir, has become a multi-billion dollar business by developing algorithms that can be used to identify terrorist threats using communications and other data, and to detect fraudulent behavior in health care and financial services.

As put into practice, these applications rely on converting large amounts of unstructured data into "vertical" or predictive scores, often in a fully automated and scalable way, and sometimes in real time. The scores can then be used in various ways. First, they can speed up or automate existing processes: the Amazon recommendation system recommends items that it predicts to be relevant for a given consumer or situation, thus replacing a recommendation one could have obtained previously from, say, a librarian. Second, they can be used to offer a new set of services: Apple's auto-complete takes the word or sentence with the top score, and proposes it as the auto completion. Finally, the scores could be used to support decision making. For example in the context of credit card fraud, the transaction score is reported to the issuing bank, and most banks implement some policy that dictates which transaction scores are approved, which are rejected, and which require further investigation.

There has been a remarkable amount of work on the statistical and machine learning techniques that underlie these applications – classification models, Lasso and Ridge regressions, etc. These methods are now quite common in statistics and computer science, although rarely used in empirical microeconomics. Although a detailed description of the methods would be way beyond the scope of this paper, a short conceptual overview is useful to fix ideas for our later discussion.

The predictive modeling problem can be described by imagining N entries that are associated with N outcome measurements, as well as a set of K potential predictors. In many cases the information about each entry is rich and unstructured, so there are many possible predictors that could be generated. Indeed, the number of potential predictors K may be far larger than the number of observations N. An obvious concern is over-fitting: with K>N it typically will be possible to perfectly explain the observed outcomes, but the out-of-sample performance may be poor.

In this setting, the goal is to construct a statistical model that maximizes in-sample predictive power, but at the same time does not "over use" potential predictors in a way that would lead to poor out-of-sample performance. Different methods vary in the way that the model is constructed and how the over-use of predictors is penalized. For example, a Lasso regression chooses coefficients to minimize the sum of squared deviations, subject to a constraint on the sum of the coefficients' absolute values. It also is common to evaluate the trade-off between in-sample predictive power and over-fitting, by splitting the sample into a "training sample" used to estimate the model parameters, and a "test sample" used to evaluate performance.[4] Again, this is an approach rarely taken in empirical microeconomics.

A crucial, but often implicit, assumption in machine learning is that the environment being studied is relatively stable, in the sense that the estimation sample (both training and test samples) is

[4] This type of predictive modeling is sometimes referred to as "supervised learning". There is also a large class of big data techniques that fall under the heading of "unsupervised learning", in which roughly speaking outcome variables are not available and the goal is to describe how a large set of predictor variables relate to one another. This category would include methods such as clustering or principal components.

generated by the same data process that later will generate the sample for which prediction is required. Of course, environments evolve over time, so this is not a perfect assumption. In applications where new data becomes available at high frequency, the algorithm can also "re-train" itself continuously, and adjust the predictive model over time as the environment changes.

When economists consider the utility of machine learning methods, what comes to mind is often a version of the Lucas Critique. If the predictive model is used to decide on a policy intervention, the result may not be what the model predicts, because the policy change may affect the underlying behavior that is generating the relationships in the data. That is, the models are "predictive" rather than "structural". Of course, this does not render a predictive model useless, because the bite of the critique depends a lot on the situation. For example, it is possible that some Amazon shoppers have realized how recommendations are being generated, and changed their shopping behavior to get different recommendations. But probably most haven't. On the other hand, if Amazon started to offer large targeted discounts using a similar model, they might elicit some behavior change.

4. Opportunities for Economic Policy

The potential uses of big data for economic policy roughly parallel the uses in the private sector. In this section, we start by describing the data resources available to the government, and also how private sector data might be used to better track and forecast economic activity. We then describe how big data might be used to inform policy decisions or to improve government services, along the lines of some of the information products and services described in the prior section.

4.1. Making Use of Government Administrative Data

Through its role in administering the tax system, social programs, and regulation, the federal government collects enormous amounts of granular administrative data. Examples include the rich micro-level datasets maintained by the Social Security Administration, the Internal Revenue

Service, and the Centers for Medicare and Medicaid. Although there is less uniformity, state and local governments similarly generate large amounts of administrative data, particularly in areas such as education, social insurance and local government spending.

Government administrative data are almost certainly under-utilized, both by government agencies and, because of limited and restricted access, by researchers and private data vendors who might use this data to uncover new facts. The major datasets also tend to be maintained separately, unlike in many European countries, which may have datasets that merge individual demographic, employment and in some cases health data, for the entire population.

Administrative data is a powerful resource for a number of reasons. First, it typically covers individuals or entities over time, creating a panel structure, and data quality is high (Card et al., 2011). Moreover, because the coverage is "universal", administrative datasets can be linked to other, potentially more selective, data. We elaborate on this in the next section.

In cases where the government has allowed access to administrative datasets, there often have been profound consequences for economic policy discussions. In many cases, this has come not from any clever research design or statistics, but simply from describing basic patterns in the data. For instance, Piketty and Saez (2003) used IRS data to derive a historical series of income shares for the top percentiles of earners among US households. Their paper, and related work, has had a profound influence on recent policy debates, by helping to make the rising income share of top earners a major focus of discussions about economic inequality.

An example which differs in the details, but is similar in spirit, is the work of John Wennberg and colleagues at Dartmouth. Over a period of several decades, they have used large samples of Medicare claims to show that there is a great deal of unexplained variation in Medicare spending per enrollee that cannot be attributed to differences in health status or prices, and that does not appear to correlate with measured health outcomes. This research received an extraordinary

amount of attention during the debate over the Affordable Care Act in 2009, and has become perhaps the leading evidence for inefficiencies in the US healthcare system.

4.2. New Measures of Private Sector Economic Activity

Government agencies also play an important role in tracking and monitoring private sector economic activity. Traditionally much of this has been done using survey methods. To measure price inflation, the Bureau of Labor Statistics sends surveyors out to stores to manually collect information on the posted prices and availability of approximately 80,000 appropriately selected items. These data are aggregated into various inflation indices such as the Consumer Price Index. Measures of unemployment, consumer expenditure, and wages and benefits rely on similar survey-based methodologies.

Alternative approaches to collecting large-scale, and even real-time, data on prices, employment and spending are rapidly becoming available. For instance, the Billion Prices Project (BPP), developed by Alberto Cavallo and Roberto Rigobon, provides an alternative measure of retail price inflation. It relies on data from hundreds of online retail websites in more than fifty countries. The data are used to construct price indices that can be updated in real time. In countries such as the United States, the BPP index seems to track the CPI relatively closely. In other countries where government survey measures may be less reliable or non-existent, the automatically gathered online data already may be preferable. Cavallo (2012) uses the data underlying the BPP index to document patterns of price changes, in the same way that researchers have used the data underlying the CPI (Klenow and Kryvtsov, 2008).

Similar possibilities also exist for augmenting the measurement of consumer spending and employment. MasterCard markets a product called “Spending Pulse” that provides real-time consumer spending data in different retail categories, and Visa generates periodic reports that

successfully predict survey-based outcomes ahead of time. Similarly, Automatic Data Processing (ADP) and Moody's Analytics release a monthly report on private sector employment, based on data from the roughly 500,000 firms for which ADP provides payroll software.

These approaches still have some disadvantages relative to government survey measures. Although the underlying data samples are large, they are essentially "convenience samples" and may not be entirely representative. They depend on who has a Visa or MasterCard and decides to use it, or on which firms are using ADP to manage their payroll records. On the other hand, the data are available at high frequency and granularity, and their representativeness could be assessed empirically. Plus, it is worth pointing out that many representative surveys are not immune to similar concerns due to selective responses and heterogeneous response quality.

Another intriguing idea is to use indirect measures such as search queries or social media posts to provide contemporaneous forecasts of economic statistics. Choi and Varian (2011) have provided one example, by showing that Google search engine data can provide accurate proxy measures of economic time series such as unemployment claims and consumer confidence. As one application, they consider a short-run "now-cast" of monthly automobile sales, where the underlying series to be predicted comes from a survey conducted each month by the US Census Bureau. Choi and Varian show that relative to a basic autoregressive time-series model for automotive sales, they can improve the mean-squared predictive error by adding two Google Trends measure of contemporaneous search interest: for "Trucks & SUVs" and "Automotive Insurance".[5]

Although Choi and Varian pick a few specific economic time series, the approach is applicable to many data series on consumer spending or sentiments. Of course, one challenge is that there are hundreds or thousands of different search queries that might plausibly predict spending in different categories. Scott and Varian (2013) propose that researchers adopt an automated

[5] See also Goel et al. (2010) for a further discussion of using search queries to predict consumer behavior.

approach employing the tools of statistical learning described earlier. Their paper describes one such approach using Bayesian methods, which in principle can be used to provide short-term forecasts of many narrow categories of consumer goods, or other time series.

We suspect that this type of real-time indices of economic (or other) activity is going to become even more popular. In addition to Google Trends, mentioned earlier, which generates an index that uses information from search queries on Google, Twitter now publishes a daily Twitter index, which is based on the context of Twitter messages. We are not aware of daily employment, or consumer lending, or credit card spending or online shopping index, but one can easily imagine how these types of high-frequency data would complement, and perhaps eventually substitute, more traditional (and lower frequency) data series on economic activity.

4.3. Improving Government Operations and Services

One of the big changes in modern business is that debates and decisions are routinely informed by large amounts of data analytics, and in at least some companies, by extensive experimentation (Varian, 2010). Many government agencies are increasingly smart about using data analytics to improve their operations and services. However, most agencies almost certainly lag behind the best private sector firms, and face challenges of both infrastructure and personnel needs. For example, a 2008 report by the JASON study group described some of these challenges in the context of how the military must try to process and analyze the vast quantities of sensor data that have become available, such as from drone flights and communications monitoring.[6]

In some cases, the government collects a great deal of data that would be useful for guiding policy decisions but has not been utilized very effectively. For example in healthcare, the Center for Medicare and Medicaid Services has a record of every Medicare health claim over the last few

[6] "Data Analysis Challenges" by the JASON study group (JSR-08-142, December 2008), available at http://www.fas.org/irp/agency/dod/jason/data.pdf.

decades, and eventually will have enormous amount of clinical information from electronic health records. It also is routinely criticized for spending money ineffectively. The data it collects almost certainly would allow for detailed cost-benefit analyses of different treatments and procedures, but it is proscribed from using this data-intensive approach by Congress.

One interesting opportunity that some government agencies seem to be exploring is to make datasets accessible and hope that researchers or other individuals will utilize these datasets in ways that end up improving agency functions. For example, New York City provides a huge catalog of datasets available for download at NYC OpenData. The available data include geo-location data on schools, subways, wifi hotspots, information on metropolitan transit and electricity consumption, crime statistics, and hundreds of other types of data. Ho (2012) has used this source to analyze restaurant hygiene data to document that the restaurant hygiene grades in New York have very little consistency across inspection and hence little year-to-year correlation, suggesting serious problems with the grading process.

The federal government has undertaken a similar exercise with the website Data.Gov that has made available several hundreds of thousands of government datasets. One goal appears to be to encourage not just researchers but software developers to develop tools or applications that would be built on the underlying data, although it does not appear that many have been built so far.

4.4. Information Products or Services

The most exciting private sector application of big data that we discussed above was using predictive modeling to automate business processes, or to improve or develop new products or services. While some government agencies probably are engaging in this type of activity, we are not aware of very many salient examples. However, it is easy to think of many examples where

government datasets might be used to create the types of information products that are commonly seen in the private sector.

One area of government activity where we could imagine such products is in the area of consumer protection. A key challenge of consumer protection is to keep individuals from making decisions they will come to regret (and predictably will come to regret) without proscribing individual choices. Behavioral economics has emphasized that one way to strike this balance is through the framing of decisions (e.g. well-chosen defaults), and another way is through the careful presentation of information. For instance, it is frequently pointed out that people can end up making major financial decisions --- buying a house, saving for retirement, planning health care spending --- without very good information about the financial consequences. The types of predictive models discussed above are particularly good for creating personalized summary information: How many consumers who take this type of loan with this type of financial situation ultimately default? What is the range of fees paid by a similar consumer for a particular financial product or service? What is the eventual cost for patients who choose this line of medical treatment? While the government might not be the right entity to create these tools, the information it collects surely would be a useful input.

Another interesting, but far more controversial idea, would be to use predictive modeling to improve the targeting of government services. For instance, it is possible to imagine a utilitarian argument that Medicare should score individuals based on their likely response to a treatment and cover the treatment only if the score exceeded a particular level. Similarly, a tax rebate program that aimed to provide economic "stimulus" might be most effective if it were targeted specifically to those households who were predicted to have a particularly high marginal propensity to consume.

These examples are useful because they correspond roughly to the sorts of things that private sectors companies are now doing all the time --- targeting discounts or rebates to particular

consumers, or approving individuals for insurance or credit only if they meet certain scoring criteria. Of course, we tolerate this in the private sector, but many people's reaction to parallel approaches taken by the government would be horror. In this sense, it seems clear that there are constraints on the way that the government can target services that probably would rule out a range of "private sector like" uses of predictive modeling.

5. Opportunities for Economic Research

We now take up the question of how the data revolution might affect economic research, in terms of the scope and quality of the results, the methods used, and the required training of empirical economists. Since economic research is primarily retrospective analysis, some of the most obvious impact of big and granular data – providing a detailed snapshot of economic activity (almost) in real time – may not be as important for research. Yet, other aspects may.

The first, and most obvious, effect of big data on economic research will be to allow better measurements of economic effects and outcomes. More granular and comprehensive data also can help to pose new sorts of questions and enable novel research designs that can inform us about the consequences of different economic policies and events. We will provide some examples below, all of which are in the spirit of empirical economics continuing to look more or less the same as it does now, only with more and better data.

A less obvious possibility is that new data may end up changing the way economists approach empirical questions and the tool they use to answer them. As one example, we consider whether economists might end up embracing some of the statistical data-mining tools described earlier. Why is this less obvious? To begin, it would mean something of a shift away from the single covariate causal effects framework that has dominated much of empirical research over the last few decades. In the minds of many economists, there is a sharp distinction between predictive modeling

and causal inference, and as a result statistical learning approaches have little to contribute. Our view is that such a distinction is not always so sharp, and we think that this type of work will be increasingly used in economics as big data sets become available for researchers and as empirical economists gain greater familiarity and comfort with machine learning statistical tools.

5.1. Novel Measurement and Research Designs

Both large-scale administrative datasets and new private sector data have the potential to enable a variety of novel research designs. We illustrate with a few examples.

One salient example is the study by Chetty et al. (2012) on the long-term effects of better teaching. The study combines administrative data on 2.5 million New York City schoolchildren with their earnings as adults twenty years later. The main question is whether the students of teachers who have higher "value-added" in the short run subsequently have higher earnings as adults, where teachers' valued added is measured by the amount that test scores are improved. The results are striking. The authors find that replacing a teacher in the bottom 5% with an average teacher raises the lifetime earnings of students by a quarter of a million dollars in present value terms.

The study demonstrates the value of large-scale administrative data in several ways. First, the authors are able to link school test score data and subsequent tax records for a large number of students. Such an exercise would be difficult or impossible with aggregate data or a small random sample. Second, the long-term nature of the tax data makes it possible to capture both the adult earnings of the students as well as information about their parents at the time they were dependents. Finally, the granular nature of the test score data allows the authors to examine a key assumption needed for identification, namely that students are not sorted to teachers on the basis of their underlying ability. While this cannot be ruled out definitely, the authors are able to present a variety of checks that make the assumption convincing.

A second recent example, that uses both large-scale administrative data as well as proprietary private sector data, is the evaluation of Oregon's Medicaid expansion conducted by Finkelstein et al. (2012). In 2008, Oregon expanded the eligible population for its Medicaid program and used a lottery to determine which individuals from a larger underlying set of potential eligibles would be allowed to enroll. This expansion and the associated lottery created a large natural experiment and an opportunity to study the effects of providing people with relatively generous health insurance. The researchers combined the Oregon Medicaid lottery and enrollment data with administrative records on hospital discharges and mortality, with credit records obtained from TransUnion, and with detailed survey data.

Again, the results are quite striking. After the first year, the population that received Medicaid coverage had substantially higher health care utilization, lower medical debt, fewer delinquencies, and better self-reported health (although a follow-up study by the same authors found little evidence of improvement on a variety of objective biometric measures). The study illustrates some of the same benefits of large-scale universal datasets: the authors are able to take a given subset of the Oregon population and locate not just their subsequent hospital records, but also their credit histories in comprehensive datasets, allowing them to trace out the consequences of the Medicaid experiment for a large number of outcome measures.

A third example comes from some of our own recent work on internet commerce, which is rather different in that we have used large-scale proprietary data, obtained through a collaboration with eBay. In one paper (Einav et al., 2013c), we use detailed browsing and purchase data on the universe of eBay customers (more than 100 million in the United States), to study the effect of sales taxes on internet commerce. Currently retailers must collect sales taxes on online purchases only if the buyer is a resident of the same state; retailers do not collect sales tax on interstate purchases, which account for a large fraction of internet commerce.

Aggregated data on state-to-state trade flows provide relatively standard estimates of tax elasticities, but we also use the detailed browsing data to obtain more micro-level evidence on tax responsiveness. Specifically, we find groups of individuals who clicked to view the same item, some of whom were located in the same state as the seller, and hence taxed, and some of whom were not, and hence went untaxed. We then compare the purchasing propensities of the two groups, doing this for many thousands of items and millions of browsing sessions. We find significant tax responsiveness, and evidence of substitution to similar (but untaxed) alternative products, but much lower responsiveness than one would expect for retail price changes, suggesting a wedge between internet price elasticities and internet tax elasticities.

In two other recent studies (Einav et al., 2013a, 2013d), we studied online pricing and sales strategies using a different research design that takes advantage of the granular nature of internet data. In these studies, we took the full set of listings posted each year on eBay and identified hundreds of thousands of items that had been listed for sale multiple times by the same seller, either simultaneously or sequentially, with different pricing or fees or sales mechanisms. We then used these "seller experiments" to estimate the degree of price dispersion, residual demand curves, and how consumers respond to potentially non-transparent charges such as shipping fees.

One lesson we have drawn from our internet commerce research is that highly granular data can be particularly useful for finding natural experiments. For example, moving from weekly data to minute-by-minute data, or to data on individual consumers and units being sold, one can take advantage of very specific institutional details or micro-level variation that would be difficult to isolate and exploit with more aggregated data. As with the studies above that rely on administrative data, there are also opportunities to obtain rich data on the individuals being studied (e.g. to segment consumers by their purchase histories), or to explore a variety of consequences from a given experiment – e.g. substitution to different items in the event of a price change.

A second, related feature is that in such types of research, when the estimates are based on many small experiments, it is almost certain that some (hopefully small) fraction of the experiments could suffer from various problems, or that it would be difficult to establish the credibility of each single little experiment separately. However, one advantage of big data is that the size and scope of the data allow for many strategies by which one could assess the robustness of the results and the validity of the key assumptions. For example, in Einav et al. (2013d) described earlier we use many alternative definitions to group items in order to make sure that our primary definition is not too broad. In our study of taxes described above, we go on and examine subsequent activity of the user within the same browsing session, after he bought or decided not to buy the clicked item. Such additional "detective work", which cannot be carried out with more traditional data, can provide further reassurance about the causal interpretation of the results.

A related observation is that as companies rely more heavily on data for their day-to-day operations, it has become easier and more cost-effective for companies to experiment. First, it is much easier to run an experiment when pricing or other instruments are automated. Furthermore, when firms have more customized and granular pricing strategies, running an experiment is easier, less observable, and not as risky. Indeed, many online platforms – as part of their regular operations – constantly use a small share of their operations as an experimentation platform. Once data is captured quickly, it is easier and cheaper to capture the results of an experiment, and (if successful) implement on it. Finally, with automated strategies, it is viable for firms to use multiple strategies at the same time, and to sometimes even randomize the set of customers that are offered one option or another, so there is even potential for some explicit randomization.

5.2. Statistical Learning and Economic Research

The studies described in the prior section make use of big data, but the conceptual approaches and statistical methods are familiar ones. In particular, the object being studied is the relationship

between a particular treatment (having a better teacher, getting health insurance, being charged sales tax), and an outcome variable (adult earnings, health utilization, purchasing). Many, if not most, studies in empirical microeconomics have this structure, where the goal is to study a particular bivariate relationship – often, but not always, a causal one – holding "all else equal", where the "all else equal" part is often implemented by controlling for other predictive variables.

In contrast, the predictive modeling approaches described in Section 3 are inherently multivariate. The focus is not on how a single variable affects a given outcome measure, but on how the outcome varies with a large number of potential predictors, and where indeed the analyst does not use any prior theory (if she has any) as to which predictors are relevant. This conceptual difference raises the question of whether any of the "big data" techniques common in statistics and computer science will turn out to be useful in economic research.

We think the answer is likely to be affirmative. One application that already has been explored (Belloni et al., 2012a, 2012b) is to use machine learning techniques to improve the efficiency of treatment effects studies when a research has either a large number of potentially confounding variables, or alternatively a large number of potential instruments. Here the goal is still to estimate a particular bivariate relationship, but to use penalized regressions either to identify an optimal set of controls, or an optimal set of instruments given a large potential number.

Another potentially powerful use of predictive modeling is to incorporate heterogeneity into econometric models and analyses. In our own research on credit and insurance markets (Bundorf et al., 2012; Einav et al., 2012; Einav et al., 2013b), we have used "off-the-shelf" credit and health risk scores to account for the default propensities or likely health expenditures of individual consumers. For example, in Einav et al. (2012) we were interested in understanding consumer borrowing behavior and how lenders should set loan prices and credit limits for different segments of borrowers as stratified by their default risk. Predictive modeling provides a natural way to

achieve this stratification, although the particular choice of predictive model was made by statisticians whose predictive scores derived from credit bureau records that were used as data.

Similarly, in theoretical models of insurance markets, it is common to associate individuals with a "risk type" that summarizes their probability of accident of loss. In recent empirical work that looks at consumer choice of insurance or health plans (Bundorf et al., 2012; Einav et al., 2013b), it has been common to again use off-the-shelf risk scores to summarize individual heterogeneity in a parsimonious way. These scores provide a useful way of assessing, for instance, whether riskier individuals systematically choose more generous insurance coverage, and whether prices in a market accurately adjust for the likely cost of different individuals to insurers that underwrite them.

In these examples, economic researchers are consumers of machine learning models, but not the producers of them. However, it is easy to imagine future applications where economists will be interested in characterizing the heterogeneity of individuals or products or firms in order to analyze differences in decisions or treatment effects. In such cases, machine learning techniques can provide a useful way to obtain a one-dimensional "score" that summarizes a large amount of information about the entities being studies, just as a consumer's credit score summarize a rich unstructured history of borrowing and repayments into a scalar summary of default risk.

A related point is that predictive scores themselves can be interesting objects to study. For instance, in health insurance, risk scores provide a mapping from an individual's demographics and past health care utilization into a one-dimensional prediction of future health care utilization. An interesting question may be whether these relationships are stable when there are various changes in the environment. For example, if insurers begin to manage utilization or charge higher copayments, the prior relationships between demographics and past utilization and current utilization may not hold.

Here, a key aspect, which is important to evaluate on a case by case, is how far out of sample the predictions need to go in order to answer a specific question. It is plausible to imagine cases where all or some of the endogenous (not causal) relationships identified in sample may remain valid, and rather than using quasi experiments to estimate a specific relationship, one can ask how the whole correlation structure in the data respond to changes in the environments, and perhaps identify subsets that are more stable and then use them as more structural relationships despite the various a-priori concerns.

5.3 New Questions and Objects of Interest

Perhaps the most dramatic way by which big and rich data may shift economic research is in introducing new objects of interest. Much of current research in applied microeconomics is focused on obtaining clear and well measured estimates, which could translate to specific policy recommendations. But when the units of analysis are heterogeneous, and the data is rich enough to capture this heterogeneity, it is plausible that the parameter of interest is heterogeneous as well, and so does the optimal policy. In such situations, one can imagine some economic research shifting from pure measurement toward "tool building", perhaps making the output of some economic research resemble somewhat current research in computer science.

As an example, one can consider a textbook problem of monopoly pricing. A standard analysis in industrial organization would obtain data on demand, will try to isolate identifying price variation, and measure the price elasticity, which would then translate to the monopolist's optimal price. But suppose now that the data on individual consumers is sufficiently large and sufficiently rich, that they can be classified to small groups, each with its own idiosyncratic price elasticity. In such a case, the required output from the analyst would not be "the" price elasticity, but rather a classification algorithm that would classify individual consumers to types, and a measure of price elasticity (or an optimal price) that is customized to each type.

Indeed, this type of empirical output has been quite common for a while in more quantitative sectors of the economy, where insurance companies and lenders customized their offer terms to individual customers. The advent of big data makes such analysis feasible and common in other sectors, for example in grocery stores, such as Safeway, which now offers customized individual-specific discounts as a function of individual price elasticities.

But while the examples above all use research associated with firms' pricing, similar points could apply to many other policy instruments. The optimal amount of insurance coverage or physician incentives could depend on the healthcare environment and the physician and patient characteristics, the optimal class size could depend on the grade, the school, the teacher, or the student mix, and the optimal amount of infrastructure could depend on the location. As the combination of big data and machine learning and other techniques allow empirical researchers to capture large amount of heterogeneity, research might shift toward studying the heterogeneous impact of certain policies on different economic units, and the mapping from measurable heterogeneity of such units to (customized) policy instruments.

6. Challenges

Several challenges confront economists wishing to take advantage of large new datasets. These include gaining access to data, developing the data management and programming capabilities needed to work with large-scale datasets, and finally (and most importantly!) thinking of creative approaches to summarize, describe and analyze the information contained in these data.

Data Access. Research on topics such as labor economics, productivity and household consumption traditionally have relied on government survey data such as the U.S. Census, the Panel Study of Income Dynamics (PSID) and the National Longitudinal Survey of Youth (NLSY). For many of these data, there are well-established protocols for accessing and making use of the data. In some cases,

such as the U.S. Census Data Research Centers, these protocols are cumbersome and probably discourage a fair number of researchers, but at least they reflect a conscious effort to trade off between research access and confidentiality concerns.

These systems are still being worked out for the large-scale administrative data that recently has been used for economic research: from the IRS, Medicare, or Social Security Administration. The privacy issues associated with the increased amount of data are important, and have been already discussed in this publication just a year ago (Goldfarb and Tucker, 2012). But as Card et al. (2010) point out, many European countries, such as Norway, Sweden and Denmark, have gone much farther to facilitate research. The experience in these countries suggests that broader access is possible, and that as may be expected, reducing the barriers to accessing such data sets a profound effect on the amount of researchers using it.

Many of the novel data we have discussed above belongs to private companies. Accessing private company data creates several issues for researchers. First and most obviously, not every company wants to work with researchers. While many view it as potentially beneficial, and a useful way to learn from outsiders, others may view it as a distraction or focus on the publicity risks. Researchers who collaborate with companies generally need to enter into contracts to prevent disclosure of confidential information, and may face some limits on the questions they can study. Our experience has been that the benefits of working with company data generally far outweigh the costs, but that a fair amount of effort on both sides is required to develop successful collaborations.

Private sector datasets also can be limited in certain ways. They generally contain information only on a firm's customers, who may not be representative even within a particular industry. Also, many private sector datasets are collected for transactional purposes, and as a result may contain a very specific set of information that is ideal for some purposes but not for others. For example, a computerized record exists for practically every physician visit in the United States, but it is

generally an insurance claim record that records a very specific set of information necessary for payment. The record may not reveal any type of actual health information such as the patient's biometrics or how they feel. Nor is the data easily linked to employment records (at least in the United States, although it is in some European countries), household financial information, or social network indicators, even though electronic records on the same population may be available. It seems conceivable that what can be learned from any specific type of data now being captured will be far less than what might be learned from linking currently distinct types of information.

Data Management and Computation. One way that some commentators have defined "big data" is that datasets are "big" when they require a significant investment of time and resources simply to manage. Virtually all of the successful internet companies, and more and more data-intensive companies have invested substantial resources into data storage and distributed data processing, also into hiring skilled computer scientists and engineers. Indeed, even when these companies hire "data scientists" whose job is to analyze data to look for empirical patterns, they look for people trained in computer science, rather than econometrics. Our expectation is that future economists who want to work with large datasets will have to acquire at least some of the new tools of the computer scientists, so they can combine the conceptual framework of economics with the ability to actually implement ideas quickly and efficiently on large-scale data.

Asking the Right Questions. One additional observation is that in working with very large, rich datasets, it can be non-trivial just to figure out what questions the data might be able to answer convincingly. While in the past a researcher could simply open up her data on the screen and visually get a sense of the key features, large data sets require time and effort for conceptually trivial tasks, such as extracting and summarizing different variables, and exploring relationships between them. Just looking within the last several years of our own program at Stanford, we see dissertations that use data from retail platforms (eBay), from job matching platforms (oDesk.com

and freelancer.com), lending platform (prosper.com), accommodation platform (airbnb.com), and financial management sites (several, names not circulated). Many of these projects have turned out very successfully, but almost all of them started with a long and slow process of just figuring out what exactly was in the data, and how to manage it.

Of course, the situation may turn out to be a bit different with large administrative datasets, to the extent that they end up being used by many researchers, because over time there will be common learning about what are the advantages and drawbacks of the data, and about various methods and strategies that are useful for organizing the data, and exploring different questions. So this may be one further difference between future research with large government datasets, which if access is increased may occupy many economics scholars, relative to research with proprietary datasets that are likely to allow much more limited access.

7. Final Thoughts

There is little doubt, at least in our own minds, that over the next decades "big data" will change the landscape of economic policy and economic research. As we emphasized throughout, we don't think that big data will substitute for common sense, economic theory, or the need for careful research designs. Rather, it will complement them. How exactly remains to be seen. In this article we tried to lay out what we see as the vast opportunities, as well as challenges, that come with the ongoing data revolution. We look forward to seeing how it will play out.

References

Belloni, Alexandre, D. Chen, Victor Chernozhukov, and Christian Hansen (2012a). "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain." *Econometrica* 80(6), 2369-2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2012b). "Inference on treatment effects after selection amongst high-dimensional controls." Cemmap Working Paper # CWP10/12.

Bundorf, Kate, Jonathan Levin, and Neale Mahoney (2012). "Pricing and Welfare in Health Plan Choice." *American Economic Review* 102(7), 3214-3248.

Card, David, Raj Chetty, Martin Feldstein, and Emmanuel Saez (2010). "Expanding Access to Administrative Data for Research in the United States." NSF SBE 2020 White Paper.

Cavallo, Alberto (2012). "Scraped Data and Sticky Prices," MIT Sloan Working Paper.

Chetty, Raj, John Friedman, and Jonah Rockoff (2011). "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." NBER Working Paper No. 17699.

Choi, Hyunyoung, and Hal Varian (2012). "Predicting the Present with Google Trends." *Economic Record* 88, 2–9.

Einav, Liran, Chiara Farronato, Jonathan Levin, and Neel Sundaresan (2013a). "What Happened to Online Auctions?" Mimeo, Stanford University.

Einav, Liran, Amy Finkelstein, Stephen Ryan, Paul Schrimpf, and Mark Cullen (2013b). "Selection on Moral Hazard in Health Insurance." *American Economic Review* 103(1), 178-219.

Einav, Liran, Mark Jenkins, and Jonathan Levin (2012b). "Contract Pricing in Consumer Credit Markets." *Econometrica* 80(4), 1387-1432.

Einav, Liran, Dan Knoepfle, Jonathan Levin, and Neel Sundaresan (2013c). "Sales Taxes and Internet Commerce," NBER Working Paper No. 18018.

Einav, Liran, Theresa Kuchler, Jonathan Levin and Neel Sundaresan (2013d). "Learning from Seller Experiments in Online Markets," NBER Working Paper No. 17385.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group (2012). "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics* 127(3), 1057-1106.

Goel, Sharad, Jake Hofman, Sebastien Lahaie, David Pennock and Duncan Watts (2010). "Predicting Consumer Behavior with Web Search." *Proceedings of the National Academy of Sciences* 107(41), 17486-17490.

Goldfarb, Avi, and Catherine Tucker (2012). "Privacy and Innovation." *Innovation Policy and The Economy* 12, 65-90.

Ho, Daniel E. (2012). "Fudging the Nudge: Information Disclosure and Restaurant Grading." *Yale Law Journal* 122.

Imbens, Guido, T. Barrios, Rebecca Diamond, and M. Kolesar (2011). "Clustering, Spatial Correlations and Randomization Inference." Mimeo, Harvard University.

Klenow, Peter J., and Oleksiy Kryvtsov (2008). "State-Dependent or Time-Dependent Pricing: Does It Matter for Recent U.S. Inflation?" *Quarterly Journal of Economics* 123, 863-904.

Piketty, Thomas, and Emmanuel Saez (2003). "Income Inequality in the United States, 1913-1998." *Quarterly Journal of Economics* 118(1), 1-39.

Scott, Steve, and Hal Varian (2013). "Bayesian Variable Selection for Nowcasting Economic Time Series." ASSA Annual Meeting, Presentation overheads.

Varian, Hal (2010). "Computer-Mediated Transactions." *American Economic Review Papers and Proceedings* 100(2), 1-10.