Economics in the age of big data

Liran Einav and Jonathan Levin

Science **346**, (2014);
DOI: 10.1126/science.1243089

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

REVIEW SUMMARY

ECONOMICS

Economics in the age of big data

Liran Einav[1,2]* and Jonathan Levin[1,2]

BACKGROUND: Economic science has evolved over several decades toward greater emphasis on empirical work. The data revolution of the past decade is likely to have a further and profound effect on economic research. Increasingly, economists make use of newly available large-scale administrative data or private sector data that often are obtained through collaborations with private firms, giving rise to new opportunities and challenges.

ADVANCES: These new data are affecting economic research along several dimensions. Many fields have shifted from a reliance on relatively small-sample government surveys to administrative data with universal or near-universal population coverage. This shift is transformative, as it allows researchers to rigorously examine variation in wages, health, productivity, education, and other measures across different subpopulations; construct consistent long-run statistical indices; generate new quasi-experimental research designs; and track diverse outcomes from natural and controlled experiments.

Perhaps even more notable is the expansion of private sector data on economic activity. These data, sometimes available from public sources but other times obtained through data-sharing agreements with private firms, can help to create more granular and real-time measurement of aggregate economic statistics. The data also offer researchers a look inside the "black box" of firms and markets by providing meaningful statistics on economic behavior such as search and information gathering, communication, decision-making, and microlevel transactions. Collaborations with data-oriented firms also create new opportunities to conduct and evaluate randomized experiments.

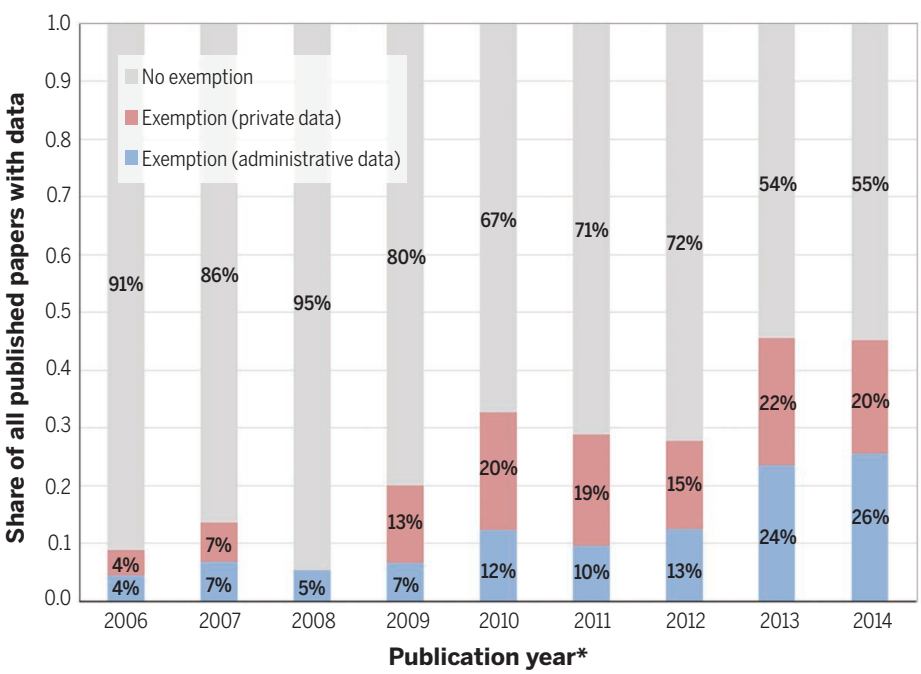
ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.1243089

Economic theory plays an important role in the analysis of large data sets with complex structure. It can be difficult to organize and study this type of data (or even to decide which variables to construct) without a simplifying conceptual framework, which is where economic models become useful. Better data also allow for sharper tests of existing models and tests of theories that had previously been difficult to assess.

OUTLOOK: The advent of big data is already allowing for better measurement of economic effects and outcomes and is enabling novel research designs across a range of topics. Over time, these data are likely to affect the types of questions economists pose, by allowing for more focus on population variation and the analysis of a broader range of economic activities and interactions. We also expect economists to increasingly adopt the large-data statistical methods that have been developed in neighboring fields and that often may complement traditional econometric techniques.

These data opportunities also raise some important challenges. Perhaps the primary one is developing methods for researchers to access and explore data in ways that respect privacy and confidentiality concerns. This is a major issue in working with both government administrative data and private sector firms. Other challenges include developing the appropriate data management and programming capabilities, as well as designing creative and scalable approaches to summarize, describe, and analyze large-scale and relatively unstructured data sets. These challenges notwithstanding, the next few decades are likely to be a very exciting time for economic research. ■



The rising use of non–publicly available data in economic research. Here we show the percentage of papers published in the *American Economic Review* (*AER*) that obtained an exemption from the *AER*'s data availability policy, as a share of all papers published by the *AER* that relied on any form of data (excluding simulations and laboratory experiments). Notes and comments, as well as *AER Papers and Proceedings* issues, are not included in the analysis. We obtained a record of exemptions directly from the *AER* administrative staff and coded each exemption manually to reflect public sector versus private data. Our check of nonexempt papers suggests that the *AER* records may possibly understate the percentage of papers that actually obtained exemptions. The asterisk indicates that data run from when the *AER* started collecting these data (December 2005 issue) to the September 2014 issue. To make full use of the data, we define year 2006 to cover October 2005 through September 2006, year 2007 to cover October 2006 through September 2007, and so on.

[1]Department of Economics, Stanford University, Stanford, CA 94305, USA. [2]National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.
*Corresponding author. E-mail: leinav@stanford.edu
Cite this article as L. Einav, J. Levin, *Science* **346**, 1243089 (2014); DOI: 10.1126/science.1243089

REVIEW

ECONOMICS

Economics in the age of big data

Liran Einav[1,2]* and Jonathan Levin[1,2]

The quality and quantity of data on economic activity are expanding rapidly. Empirical research increasingly relies on newly available large-scale administrative data or private sector data that often is obtained through collaboration with private firms. Here we highlight some challenges in accessing and using these new data. We also discuss how new data sets may change the statistical methods used by economists and the types of questions posed in empirical research.

The expansion of data being collected on social and economic activity is likely to have profound effects on economic research. In this Review, we describe how newly available public and private sector data sets are being employed in economics. We also discuss how statistical methods in economics may adapt to take advantage of large-scale granular data, as well as some of the challenges and opportunities for future empirical research.

After providing some brief background in the next section, we divide the Review into three parts. We first discuss the shift from relatively small-sample government surveys to administrative data with universal or near-universal population coverage. These data have been used in Europe for some time but are just starting to be explored in the United States. We explain the transformative power of these data to shed light on variation across subpopulations, construct consistent long-run statistical indices, generate new quasi-experimental research designs, and track diverse outcomes from natural and controlled experiments.

The second part of the Review describes the marked expansion of private sector data on economic activity. We outline the potential of these data in creating aggregate economic statistics and some nascent attempts to do this. We then discuss the rise of collaborations between academics and data-rich companies. These relationships have some trade-offs in terms of maintaining data confidentiality and working with samples that have been collected for business rather than research purposes. But as we illustrate with examples from recent work, they also provide researchers with a look inside the "black box" of firms and markets and create new opportunities to conduct and evaluate randomized experiments.

The third part of this Review addresses statistical methods and the role of economic theory in the analysis of large-scale data sets. Today, economists routinely analyze large data sets with the same econometric methods used 15 or 20 years ago. We contrast these methods to some of the newer data mining approaches that have become popular in statistics and computer science. Economists, who tend to place a high premium on statistical inference and the identification of causal effects, have been skeptical about these methods, which put more emphasis on predictive fit and handling model uncertainty and on identifying low-dimensional structure in high-dimensional data. We argue that there are considerable gains from trade. We also stress the usefulness of economic theory in helping to organize complex and unstructured data.

We conclude by discussing a few challenges in making use of new data opportunities, in particular the need to incorporate data management skills into economics training, and the difficulties of data access and research transparency in the presence of privacy and confidentiality concerns.

[1]Department of Economics, Stanford University, Stanford, CA 94305, USA. [2]National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA.
*Corresponding author. E-mail: leinav@stanford.edu

The rise of empirical economics

Hamermesh (*1*) recently reviewed publications from 1963 to 2011 in top economics journals. Until the mid-1980s, the majority of papers were theoretical; the remainder relied mainly on "ready-made" data from government statistics or surveys. Since then, the share of empirical papers in top journals has climbed to more than 70%, and a substantial majority of these papers use data that have been assembled or obtained by the authors or generated through a controlled experiment.

This shift mirrors the expansion of available data. Even 15 or 20 years ago, interesting and unstudied data sets were a scarce resource. Gathering data on a specific industry could involve hunting through the library or manually extracting statistics from trade publications. Collaborations with companies were unusual, as were experiments, both in laboratory settings and in the field. Nowadays the situation is very different along all of these dimensions. Apart from simply having more observations and more recorded data in each observation, several features differentiate modern data sets from many used in earlier research.

The first feature is that data are now often available in real time. Government surveys and statistics are released with a lag of months or years. Of course, many research questions are naturally retrospective, and it is more important for data to be detailed and accurate rather than available immediately. However, administrative and private data that are continuously updated have great value for helping to guide economic policy. Below, we discuss some early attempts to use Internet data to make real-time forecasts of inflation, retail sales, and labor market activity and to create new tracking measures of the economy.

The second feature is that data are available on previously unmeasured activities. Much of the data now being recorded is on activities that were previously difficult to quantify: personal communications, social networks, search and information gathering, and geolocation data. These data may open the door to studying issues that economists have long viewed as important but did not have good ways to study empirically, such as the role of social connections and geographic proximity in shaping preferences, the transmission of information, consumer purchasing behavior, productivity, and job search.

Finally, data come with less structure. Economists are used to working with "rectangular" data, with N observations and $K << N$ variables per observation and a relatively simple dependence structure between the observations. New data sets often have higher dimensionality and less-clear structure. For example, Internet browsing histories contain a great deal of information about a person's interests and beliefs and how they evolve over time. But how can one extract this information? The data record a sequence of events that can be organized in an enormous number of ways, which may or may not be clearly linked and from which an almost unlimited number of variables can be created. Figuring out how to organize and reduce the dimensionality of large-scale, unstructured data is becoming a crucial challenge in empirical economic research.

Public sector data: Administrative records

In the course of administering the tax system, social programs, and regulation, the federal government collects highly detailed data on individuals and corporations. The same is true of state and local governments, albeit with less uniformity, in areas such as education, social insurance, and local government spending. As electronic versions of these data become available, they increasingly are the resource of choice for economists who work in fields such as labor economics, public finance, health, and education.

Administrative data offer several advantages over traditional survey data. Workhorse surveys—such as the Survey of Consumer Finances, the Current Population Survey, the Survey of Income and Program Participation, and the Panel Study on Income Dynamics—can suffer from substantial missing data issues, and the sample size may be limited in ways that preclude natural quasi-experimental research designs (*2*). The rich microlevel administrative data sets maintained by, among others, the Social Security Administration, the Internal Revenue Service, and the Centers for Medicare and Medicaid, often have

high data quality and a long-term panel structure. Sample selection and attrition, a common issue with survey panels, is not a primary concern (*3*).

These "universal" data sets are especially powerful for analyzing population variation. For instance, Piketty and Saez (*4*) have used tax records to calculate income and wealth shares for the very upper portion of the income distribution. These calculations are problematic for traditional surveys because of small sample sizes, underreporting of high incomes or asset levels, and the fact that surveys generally extend back only a few years or, at most, decades. In contrast, tax data allow for the creation of relatively homogeneous time series spanning many decades, or even centuries.

Administrative data have been similarly useful in documenting regional disparities in economic mobility (*5*) (Fig. 1) and health care spending (*6*), in discovering the wide variation in test-score value-added measures across public school teachers (*7*), and in identifying the sizable differences in wages and productivity across otherwise similar firms (*8*, *9*). In each case, researchers have used large-scale administrative data to measure and compare the relevant variable (e.g., income, spending, productivity, or wages) across small subpopulations of individuals or firms. These results have helped to guide policy discussions and define research agendas in multiple subfields of economics.

Recent work also highlights the value of using administrative data for causal inference and policy evaluation. For these purposes, administrative data can be valuable both because its coverage and detail allow for novel research designs and because of the possibility of linking records to track outcomes from an existing experiment or quasi-experiment. The last point is an important one. Matching a data set with a random survey of 1 million U.S. households will reduce the original sample to just 1% of its original size. Merging with administrative data may leave the sample virtually unchanged.

Akerman *et al.*'s (*10*) recent study of the effects of broadband Internet access is illustrative of how administrative data sets can be combined to perform a successful evaluation study. Their research design relies on the gradual expansion of broadband access in Norway into different geographic regions. The authors link this staggered rollout to administrative tax records to estimate how broadband adoption affected firm wages and productivity. By linking individual and firm-level administrative data sets, the authors can observe multiple outcome measures and assess the effect broadband access has on specific subpopulations—for example, broadband access turns out to have very different effects on workers of different education levels.

The same advantages of universal coverage apply when the experiment or quasi-experiment that forms the basis for the study's research design affects only a relatively small population. A recent example is Chetty *et al.*'s (*11*, *12*) study of the long-term effects of teacher quality. The authors use student-level test-score data from a specific city and identify a quasi-experiment in the way students are assigned to teachers that creates variation in teacher quality. The notable step comes when the authors link the student records to administrative tax data and are able to trace the effect of teacher quality on the students' subsequent wages, two decades later.

Several recent studies have also used administrative records in powerful fashion to track outcomes from truly randomized experiments. Chetty *et al.* (*13*) track the future earnings of students who were randomly assigned to classrooms during the Tennessee STAR (Student-Teacher Achievement Ratio) experiment conducted in the late 1980s. Taubman *et al.*'s (*14*) evaluation of the Oregon Medicaid expansion similarly uses a range of administrative data to track outcomes after an episode in which Oregon expanded its Medicaid program to a randomly selected subset of newly eligible individuals. The latter study links state administrative data, hospital admission records, private sector credit bureau records, and more targeted survey data to estimate the impact of Medicaid on health and financial measures.

The potential of administrative data for academic research is just starting to be realized, and

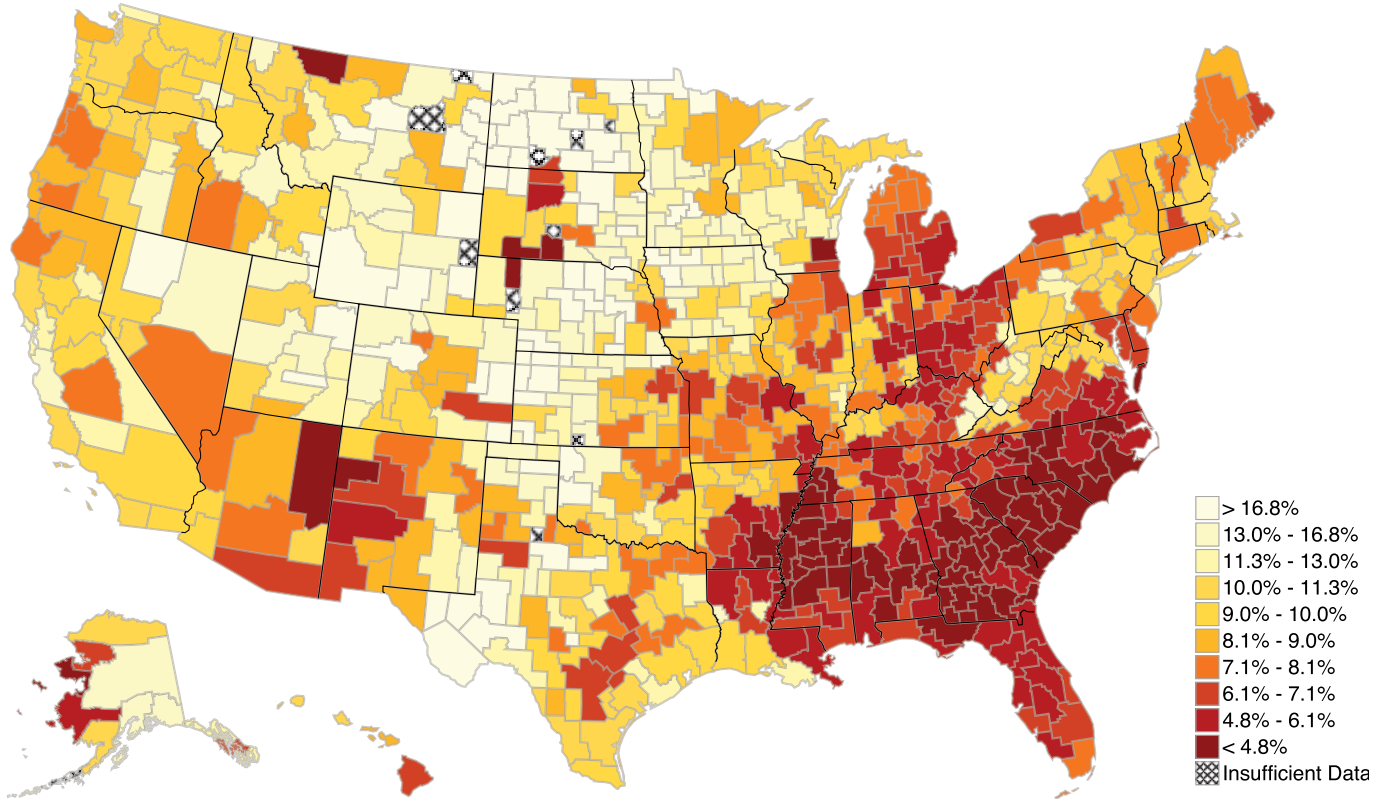


Fig. 1. Economic mobility across U.S. commuting zones. Heat map of upward income mobility using anonymous earnings records on all children in the 1980–1985 birth cohorts. Upward income mobility is measured by the probability that a child reaches the top quintile of the national family income distribution for children, conditional on having parents in the bottom quintile of the family income distribution for parents. Children are assigned to commuting zones based on the location of their parents (when the child was claimed as a dependent), irrespective of where they live as adults. [Reprint of appendix figure VIb in (*5*)]

substantial challenges remain (*15*, *16*). This is particularly true in the United States, where confidentiality and privacy concerns, as well as bureaucratic hurdles, have made accessing administrative data sets and linking records between these data sets relatively cumbersome. European countries such as Norway, Sweden, and Denmark have gone much farther to merge distinct administrative records and facilitate research. Card *et al.* (*3*) have articulated a set of principles for expanding access to administrative data, including competition for data access, transparency, and prevention of disclosure of individual records. We view these as useful guideposts. However, even with today's somewhat piecemeal access to administrative records, it seems clear that these data will play a defining role in economic research over the coming years.

Private sector data: Collection and collaborations

An even more dramatic change in data collection is occurring in the private sector. Whereas the popular press has focused on the vast amount of information collected by Internet companies such as Google, Amazon, and Facebook, firms in every sector of the economy now routinely collect and aggregate data on their customers and their internal businesses. Banks, credit card companies, and insurers collect detailed data on household and business financial interactions. Retailers such as Walmart and Target collect data on consumer spending, wholesale prices, and inventories. Private companies that specialize in data aggregation, such as credit bureaus or marketing companies such as Acxiom, are assembling rich individual-level data on virtually every household.

Although the primary purpose of all this data collection is for business use, there are also potential research applications in economics and other fields. These applications are just starting to be identified and explored, but recent research already provides some useful signals of value.

One potential application of private sector data is to create statistics on aggregate economic activity that can be used to track the economy or as inputs to other research. Already the payroll service company ADP publishes monthly employment statistics in advance of the Bureau of Labor Statistics, MasterCard makes available retail sales numbers, and Zillow generates house price indices at the county level. These data may be less definitive than the eventual government statistics, but in principle they can be provided faster and perhaps at a more granular level, making them useful complements to traditional economic statistics.

The Billion Prices Project (BPP) at the Massachusetts Institute of Technology is a related researcher-driven initiative. The BPP researchers coordinate with Internet retailers to download daily prices and detailed product attributes on hundreds of thousands of products (*17*). These data are used to produce a daily price index. Although the sample of products is, by design, skewed toward products stocked by online retailers, it can replicate quite closely the consumer price index (CPI) series generated by the Bureau of Labor Statistics, with the advantage that the standard consumer series is published monthly, with a lag of several weeks. More interestingly, the project generates price indices for countries in which government statistics are not regularly available or countries in which the published government statistics may be suspect for misreporting, as in Argentina (*18*) (Fig. 2).

Baker *et al.* (*19*) have adopted a similar data aggregation strategy by assembling the full texts of 10 leading newspapers to construct a daily index of economic policy uncertainty. In contrast to the BPP indices, their Economic Policy Uncertainty Index is a new measure of economic activity that does not have a parallel in any formal government report. However, it captures a concept that economists have argued may be important for understanding firm investment decisions and macroeconomic activity.

Recent work suggests that publicly available search query data or tweets on Twitter might be used to provide similar statistics on aggregate activity (*20*, *21*). As an example, Varian and coauthors (*22*, *23*) use Google search data to provide short-run forecasts of unemployment, consumer confidence, and retail sales. Their analysis has parallels to the well-known Google Flu Trends index, which used search query data to predict the Center for Disease Control's measure of flu infections. There is a cautionary note here as well, given that the Google Flu Trends index model broke down as Google changed its underlying search algorithm (*24*). It is likely that successful economic indices using private data will have to be maintained and updated carefully.

A second application of private data is to allow researchers to look "inside" specific firms or markets to study employee or consumer behavior or the operation of different industries. Recent work in this vein often relies on proprietary data obtained through collaborations with private firms. These agreements may take various forms, depending on the sensitivity of the data

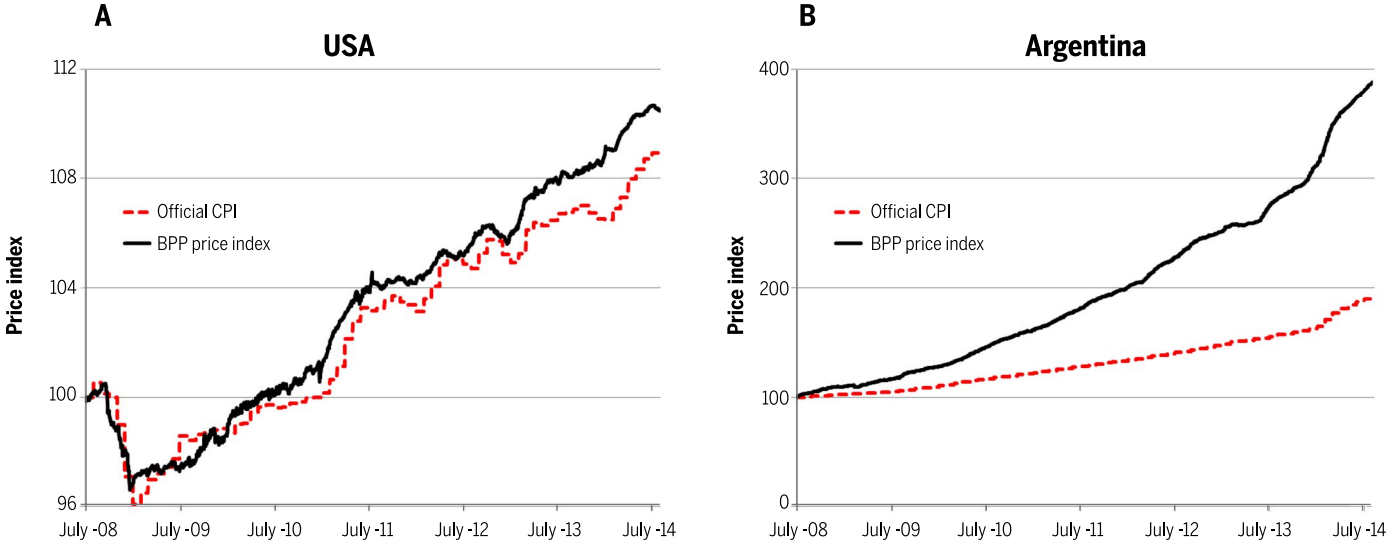


Fig. 2. BPP price index. Dashed red lines show the monthly series for the CPI in the United States (**A**) and Argentina (**B**), as published by the formal government statistics agencies. Solid black lines show the daily price index series, the "State Street's PriceStats Series" produced by the BPP, which uses scraped Internet data on thousands of retail items. All indices are normalized to 100 as of 1 July 2008. In the U.S. context, the two series track each other quite closely, although the BPP index is available in real time and at a more granular level (daily instead of monthly). In the plot for Argentina, the indices diverge considerably, with the BPP index growing at about twice the rate of the official CPI. [Updated version of figure 5 in (*18*), provided courtesy of Alberto Cavallo and Roberto Rigobon, principal investigators of the BPP]

from a privacy and business perspective. Researchers may have to agree to keep the underlying data confidential. In exchange, however, they often get to work with granular employee- or customer-level data that provide a window into the detailed operations of specific businesses or markets.

Relative to government surveys or administrative data, company data have some important differences. Sampling usually is not representative, and how well findings generalize must be evaluated case by case. Data collection emphasizes recency and relevance for business use, so variables and data collection may not be comparable and uniform over long periods. In short, the data are best viewed as "convenience" samples, albeit with potentially enormous scale. At the same time, private entities are not bound by some of the bureaucratic constraints that limit public agencies. The detail of private data can be much greater, the computing resources can be more powerful, and private companies can have far more flexibility to run experiments.

The detail and granularity of private data can offer novel opportunities to study a range of markets. For example, as part of collaboration with researchers at eBay, we recently used their marketplace data to study the effect of sales taxes on Internet shopping (*25*). One of our empirical strategies was to find instances in which multiple consumers clicked on a particular item and then compare consumers located in the same state as the seller (in which case the seller collected sales tax) to consumers located at a similar distance but across state lines (so that no sales tax was collected). The idea of the research design is to assess the sensitivity to sales taxes for otherwise similar consumers looking at the exact same product listing. This sort of analysis would not have been feasible without access to underlying browsing data that allowed us to sift through billions of browsing events to identify the right ones for our empirical strategy.

In two other recent studies (*26*, *27*), also undertaken in collaboration with eBay, we studied the effectiveness of different Internet pricing and sales strategies. To do this, we identified millions of instances in which an online seller listed the same item for sale multiple times with different pricing or shipping fees or using alternative sales mechanisms (e.g., by auction or by posted price) (Fig. 3). We then used the matched listings to estimate the demand response to different item prices and shipping fees, compare auctions with posted price selling, and study alternative sales mechanisms such as auctions with a "buy-now" option. This type of large-scale, microlevel study of market behavior is likely to become more and more common in coming years.

Similar to some of the research described above, a central theme in these papers is the use of highly granular data to find targeted variation that plausibly allows for causal estimates (in these examples, estimates of the effects of sales tax collection, pricing changes, and so forth). In the Internet case, this comes in moving from aggregated data on market prices and quantities to individual browsing data or seller listing data. Having granular data on a market with billions of transactions also provides a chance to analyze specific consumer or market segments: geographic variation, new and used goods, or experienced versus inexperienced sellers. In addition, having richer data can be useful in constructing more nuanced outcome measures. As an example, in studying the effects of sales taxes, we were able to examine not only whether facing a sales tax deterred buyers from purchasing but also whether they continued browsing and then purchased a similar untaxed item.

Large-scale granular data can also be particularly useful for assessing the robustness of identifying assumptions. Virtually every observational study in economics must deal with the critique that even after controlling for sources of confounding, the data do not approximate a controlled experiment. For example, in our work on Internet selling strategies, we aggregated many matched-listing episodes, hoping that each episode might approximate a pricing experiment conducted by the seller. But sometimes sellers may make pricing changes in response to consumer demand, complicating what one can infer from the price change. One way to check if this contaminates the results is to use narrower matching strategies that remove potential sources of confounding—for instance, focusing on cases in which sellers post two offers at the exact same time. This type of extra detective work is much easier with plentiful data.

Collaborations with private sector firms can also give rise to structured economic experiments. This type of research has accelerated

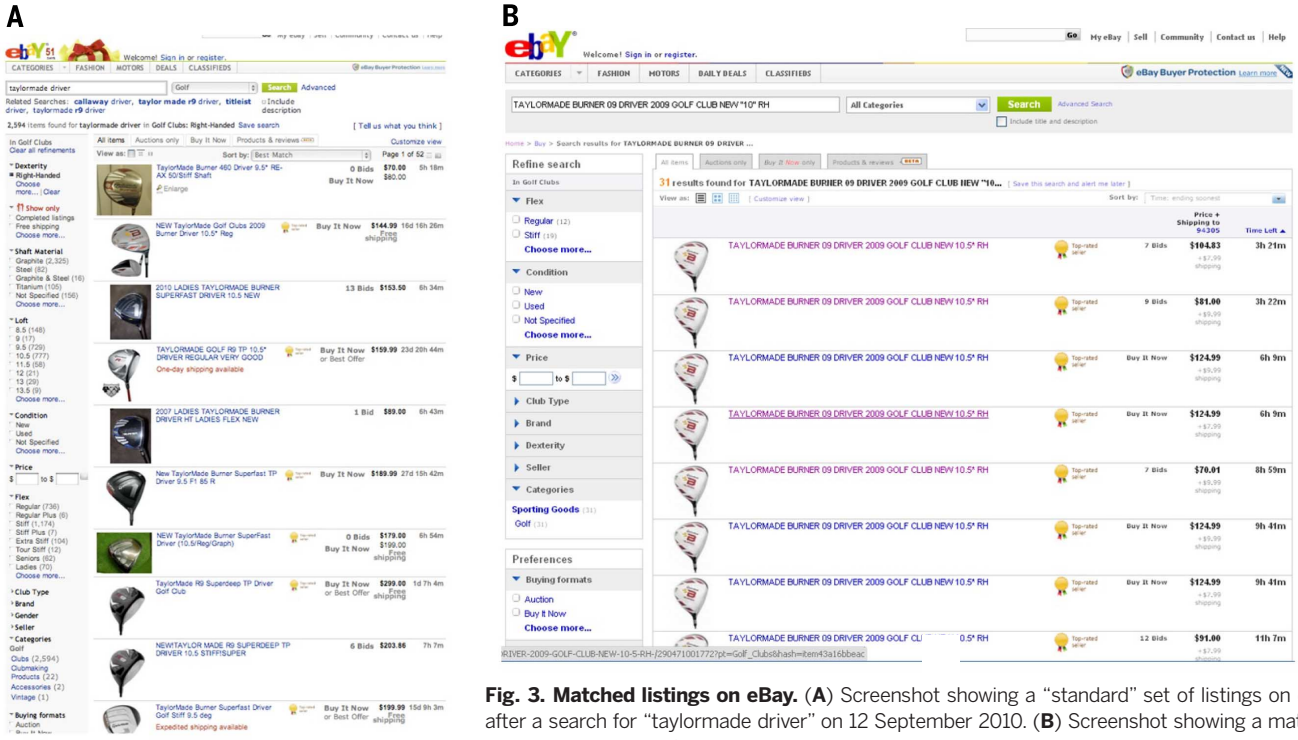


Fig. 3. Matched listings on eBay. (**A**) Screenshot showing a "standard" set of listings on eBay, after a search for "taylormade driver" on 12 September 2010. (**B**) Screenshot showing a matched set. It shows the first 8 out of 31 listings for the same golf driver by the same seller. All of the listings were active on 12 September 2010. Of the eight listings shown, four are offered at a fixed price of $124.99. The other four listings are auctions with slightly varying end times. The listings have different shipping fees (either $7.99 or $9.99). Such matched sets are ubiquitous on eBay and are useful as natural experiments in assessing the effects of changes to sale format and parameters. [Reprint of figure 1 in (*26*)]

and is particularly low-cost and scalable on the Internet, where experimentation is already a standard business practice (*28*, *29*). Recent examples include Ostrovsky and Schwarz (*30*), who worked with Yahoo! to test the use of different reserve prices in advertising auctions; Blake *et al.* (*31*), who worked with eBay to selectively shut down its Google search advertising and track the effect on eBay site visits and sales; and Horton (*32*), who worked with oDesk to provide recommendations to employers about who to hire (*33*).

As in the case of administrative data, economists working with private companies face some challenges, particularly regarding data access. Although companies may be willing to make small, nonsensitive data sets public, researchers usually have to agree to keep data confidential if they want to work directly with company records. As a result, opportunities for other researchers to replicate or extend studies may be limited. In addition, some collaborative research projects are part of broader consulting or employment relationships, raising issues regarding conflict of interest and selectivity in what results are pursued or submitted for publication.

These issues have only recently become a major topic of discussion in economics, as journals and research organizations have begun to adopt policies on transparency and disclosure. As companies capture increasing amounts of economic data, however, it seems almost certain that collaborations between academics and private sector firms will expand, so we hope that disclosure policies will prove effective and that companies will begin to establish open processes for allowing researchers access to data in ways that reasonably maintain privacy and confidentiality. The underlying issues around data privacy and acceptable types of research experiments are clearly sensitive ones that need to be handled with care and thoughtfulness (*34*).

Econometrics, machine learning, and economic theory

Recent economic research using large data sets has relied primarily on traditional econometric techniques. The estimated models usually focus on one or a few coefficients of interest, which often represent the causal effect of a particular policy or policies. Researchers put considerable thought and effort into controlling for heterogeneity or other confounding factors, often using a large set of fixed effects, and into obtaining carefully constructed standard errors for the main parameters of interest. Though studies often focus on a single preferred specification, frequently linear, it is typical to assess the robustness of the results by estimating a variety of alternative specifications and running placebo regressions to see if the preferred model generates false-positive findings.

This approach, both in conception and execution, stands in contrast to some of the data mining methods that have become popular for large-data applications in statistics and computer science [e.g., (*35*, *36*)]. These latter approaches put more emphasis on predictive fit, especially out-of-sample fit, and on the use of data-driven model selection to identify the most meaningful predictive variables (*37*). There often is less attention paid to statistical uncertainty and standard errors and considerably more to model uncertainty. The common techniques in this sort of data mining—classification and regression trees, lasso and methods to estimate sparse models, boosting, model averaging, and cross-validation—have not seen much use in economics (*38*).

There are some good reasons why empirical methods in economics look the way they do. Economists are often interested in assessing the results of a specific policy or testing theories that predict a particular causal relationship. So empirical research tends to place a high degree of importance on the identification of causal effects and on statistical inference to assess the significance of these effects. Having a model with an overall high degree of predictive fit is often viewed as secondary to finding a specification that cleanly identifies a causal effect.

Consider a concrete example: Suppose we set out to measure whether taking online classes improves a worker's earnings. An economist might hope to design an experiment or to find a natural experiment that induced some workers to take online classes for reasons unrelated to their productivity or current earnings (e.g., a change in the advertising or pricing of online classes). Absent an experimental design, however, she might consider estimating a model such as

$$y_i = \alpha + \beta x_i + \boldsymbol{z}_i'\gamma + \varepsilon_i \qquad (1)$$

where y_i is the outcome (an individual's earnings in a given year), x_i is the policy of interest (whether the worker has taken online classes before that year), β is the key parameter of interest (the effect of online education on earnings), α and γ are other parameters, $\boldsymbol{z}_i$ is a set of control variables, and ε_i is an error term.

The hope is that in a group of individuals with the same $\boldsymbol{z}_i$, whether or not an individual decides to take online classes is not related in a meaningful way to their earnings. Better data obviously help. With detailed individual data over time, the control variables might include a dummy variable for every individual in the sample and perhaps for every employer. Then the effect of online education would be estimated by comparing increases in worker earnings for those who take online classes to increases in earnings for those who do not, perhaps even making the comparison within a given firm. The focus of the analysis would be on the estimate of β, its precision, and on whether there were important omitted variables (e.g., a worker becoming more ambitious and deciding to take classes and work harder at the same time) that might confound a causal interpretation.

Given the same data, a machine learning approach might start with the question of exactly what variables predict earnings, given the vast set of possible predictors in the data, and the potential for building a model that predicts earnings well, both in-sample and out-of-sample. Ultimately, a researcher might estimate a model that provides a way to predict earnings for individuals who have and have not taken online classes, but the exact source of variation identifying this effect—in particular, whether it was appropriate to view the effect as causal—and inference on its statistical significance might be more difficult to assess.

This example may help to illustrate a few reasons economists have not immediately shifted to new statistical approaches, despite changes in data availability. An economist might argue that, short of an experimental approach, the first observational approach has the virtue of being transparent or interpretable in how the parameter of interest is identified, as well as conducive to statistical inference on that parameter. Yet a researcher who wanted to predict earnings accurately might view the first model as rather hopeless, particularly if it included a dummy variable for every individual and the researcher wanted to predict out-of-sample.

However, the two approaches are not necessarily in competition. For instance, if only a subset of control variables is truly predictive, an automated model-selection approach may be helpful to identify the relevant ones (*39*, *40*). Data mining methods may also be useful if there are important interaction effects (*41*) so that one cares about predicting effects for specific individuals rather than an average effect for the population. A potential benefit of large data sets is that they allow for more tailored predictions and estimates (e.g., a separate β depending on many specifics of the environment). Rather than estimate only average policy treatment effects, it is possible to build models that map individual characteristics into individual treatment effects and allow for an analysis of more tailored or customized policies.

The potential gains from trade go in the other direction as well. To the extent that machine learning approaches are used to assess the effect of specific policy variables and the estimates are given a causal interpretation, the economists' focus on causal identification is likely to be useful.

Economic theory also plays a crucial role in the analysis of large data sets, in large part because the complexity of many new data sets calls for simpler organizing frameworks. Economic models are useful for this purpose.

The connection between big data and economic theory can already be seen in some applied settings. Consider the design of online advertising auctions and exchanges. These markets—run by companies such as Google, Yahoo!, Facebook, and Microsoft—combine big data predictive models with sophisticated economic market mechanisms. The predictive models are used to assess the likelihood that a given user will click on a given ad. This might be enough for a company such as Google or Facebook, with enormous amounts of data, to figure out which ads to show. However, it does not necessarily tell them how much to charge, and given that each ad impression is arguably distinct, trying to experimentally set hundreds of millions of prices could be a challenge. Instead, these companies use (quite sophisticated) auction mechanisms to set prices.

The operation of the auction market depends on the interplay between the predictive modeling and the incentive properties of the auction. Therefore, making decisions about how to run this type of market requires a sophisticated understanding of both big data predictive modeling and economic theory. In this sense, it is no surprise that over the past several years many of the large e-commerce companies have built economics teams (in some cases, headed by high-profile academic researchers) or combined economists with statisticians and computer scientists or that computer science researchers interested in online marketplaces draw increasingly on economic theory.

More generally, we see some of the main contributions that economists can make in data-rich environments as coming from the organizing framework provided by economic theory. In the past century, most of the major advances in economics came in developing conceptual or mathematical models to study individual decisions, market interactions, or the macroeconomy. Frequently, the key step in successful modeling has been simplification: taking a complex environment and reducing it down to relationships between a few key variables. As data sets become richer and more complex and it is difficult to simply look at the data and visually identify patterns, it becomes increasingly valuable to have stripped-down models to organize one's thinking about what variables to create, what the relationships between them might be, and what hypotheses to test and experiments to run. Although the point is not usually emphasized, there is a sense that the richer the data, the more important it becomes to have an organizing theory to make any progress.

Outlook

This review has discussed the ways in which the data revolution is affecting economic and broader social science research. More granular and comprehensive data surely allow improved measurements of economic effects and outcomes, better answers to old questions, and help in posing new questions and enabling novel research designs. We also believe that new data may change the way economists approach empirical research, as well as the statistical tools they employ.

Several challenges confront economists wishing to take advantage of these large new data sets. These include gaining access to data; developing the data management and programming capabilities needed to work with large-scale data sets (*42*); and, most importantly, thinking of creative approaches to summarize, describe, and analyze the information contained in these data (*29*). Big data is not a substitute for common sense, economic theory, or the need for careful research designs. Nonetheless, there is little doubt in our own minds that it will change the landscape of economic research. Here we have outlined some of the vast opportunities. We look forward to seeing how they will be realized.

REFERENCES AND NOTES

1. D. S. Hamermesh, Six decades of top economics publishing: Who and how? *J. Econ. Lit.* **51**, 162–172 (2013). doi: 10.1257/jel.51.1.162
2. R. Chetty, "Time trends in the use of administrative data for empirical research," presentation slides (2012); http://obs.rc.fas.harvard.edu/chetty/admin_data_trends.pdf.
3. D. Card, R. Chetty, M. Feldstein, E. Saez, "Expanding access to administrative data for research in the United States," NSF SBE 2020 white paper ID 112 (2010); www.nsf.gov/sbe/sbe_2020/submission_detail.cfm?upld_id=112.
4. T. Piketty, E. Saez, Inequality in the long run. *Science* **344**, 838–843 (2014). doi: 10.1126/science.1251936; pmid: 24855258
5. R. Chetty, N. Hendren, P. Kline, E. Saez, Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Q. J. Econ.* 10.1093/qje/qju022 (2014). doi: 10.1093/qje/qju022
6. Dartmouth Atlas of Health Care, www.dartmouthatlas.org.
7. S. G. Rivkin, E. A. Hanushek, J. F. Kain, Teachers, schools and academic achievement. *Econometrica* **73**, 417–458 (2005). doi: 10.1111/j.1468-0262.2005.00584.x
8. C. Syverson, What determines productivity? *J. Econ. Lit.* **49**, 326–365 (2011). doi: 10.1257/jel.49.2.326
9. J. M. Abowd, F. Kramarz, D. N. Margolis, High wage workers and high wage firms. *Econometrica* **67**, 251–333 (1999). doi: 10.1111/1468-0262.00020
10. A. Akerman, I. Gaarder, M. Mogstad, "The skill complementarity of broadband Internet," Institute for the Study of Labor (IZA) discussion paper no. 7762 (2013); http://ftp.iza.org/dp7762.pdf.
11. R. Chetty, J. Friedman, J. Rockoff, Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* **104**, 2593–2632 (2014). doi: 10.1257/aer.104.9.2593
12. R. Chetty, J. Friedman, J. Rockoff, Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *Am. Econ. Rev.* **104**, 2633–2679 (2014). doi: 10.1257/aer.104.9.2633
13. R. Chetty *et al*., How does your kindergarten classroom affect your earnings? Evidence from Project Star. *Q. J. Econ.* **126**, 1593–1660 (2011). doi: 10.1093/qje/qjr041; pmid: 22256342
14. S. L. Taubman, H. L. Allen, B. J. Wright, K. Baicker, A. N. Finkelstein, Medicaid increases emergency-department use: Evidence from Oregon's health insurance experiment. *Science* **343**, 263–268 (2014). doi: 10.1126/science.1246183; pmid: 24385603
15. G. King, Ensuring the data-rich future of the social sciences. *Science* **331**, 719–721 (2011). doi: 10.1126/science.1197872; pmid: 21311013
16. H. C. Kum, S. Ahalt, T. M. Carsey, Dealing with data: Governments records. *Science* **332**, 1263 (2011). doi: 10.1126/science.332.6035.1263-a; pmid: 21659589
17. A. Cavallo, "Scraped data and sticky prices," Massachusetts Institute of Technology Sloan working paper no. 4976-12 (2012); www.mit.edu/~afc/papers/Cavallo-Scraped.pdf.
18. A. Cavallo, Online and official price indexes: Measuring Argentina's inflation. *J. Monet. Econ.* **60**, 152–165 (2012). doi: 10.1016/j.jmoneco.2012.10.002
19. S. Baker, N. Bloom, S. Davis, "Measuring economic policy uncertainty," Chicago Booth research paper no. 13-02 (2013); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2198490.
20. S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, D. J. Watts, Predicting consumer behavior with Web search. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17486–17490 (2010). doi: 10.1073/pnas.1005962107; pmid: 20876140
21. D. Antenucci, M. Cafarella, M. Levenstein, C. Re, M. Shapiro, "Using social media to measure labor market flows," National Bureau of Economic Research (NBER) working paper no. 20010 (2014); www.nber.org/papers/w20010.
22. H. Choi, H. Varian, Predicting the present with Google trends. *Econ. Rec.* **88**, 2–9 (2012). doi: 10.1111/j.1475-4932.2012.00809.x
23. H. Varian, S. Scott, Predicting the present with Bayesian structural time series. *Int. J. Math. Model. Numer. Optim.* **5**, 4–23 (2014). doi: 10.1504/IJMMNO.2014.059942
24. D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google flu: Traps in big data analysis. *Science* **343**, 1203–1205 (2014). pmid: 24626916
25. L. Einav, D. Knoepfle, J. Levin, N. Sundaresan, Sales taxes and Internet commerce. *Am. Econ. Rev.* **104**, 1–26 (2014). doi: 10.1257/aer.104.1.1
26. L. Einav, T. Kuchler, J. Levin, N. Sundaresan, "Learning from seller experiments in online markets," NBER working paper no. 17385; www.nber.org/papers/w17385.
27. L. Einav, C. Farronato, J. Levin, N. Sundaresan, "Sales mechanisms in online markets: What happened to Internet auctions?" NBER working paper no. 19021 (2013); www.nber.org/papers/w19021.
28. R. Kohavi, R. Longbotham, D. Sommerfield, R. Henne, Controlled experiments on the Web: Survey and practical guide. *Data Min. Knowl. Discov.* **18**, 140–181 (2009). doi: 10.1007/s10618-008-0114-1
29. H. Varian, "Beyond big data," presented at the National Associate for Business Economics Annual Meeting, San Francisco, CA, 7 to 10 September 2013; http://people.ischool.berkeley.edu/~hal/Papers/2013/BeyondBigDataPaperFINAL.pdf.
30. M. Ostrovsky, M. Schwarz, "Reserve prices in Internet advertising auctions: A field experiment," Stanford University Graduate School of Business research paper no. 2054 (2009); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1573947.
31. T. Blake, C. Nosko, S. Tadelis, "Consumer heterogeneity and paid search effectiveness: A large scale field experiment." NBER working paper no. 20171; www.nber.org/papers/w20171.
32. J. J. Horton, "The effects of subsidizing employer search," New York University working paper (2013); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2346486.
33. Another example is Lewis and Reiley (*43*), who report on consumer advertising experiments done in conjunction with Yahoo!. These experiments have become common, although Lewis and Rao (*44*) have recently argued that extracting useful information from them may be more challenging than one might have hoped or expected.
34. The recent episode involving an experiment that manipulated Facebook's newsfeed (*45*) is a case in point.
35. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer, New York, 2009).
36. A. Rajaraman, J. D. Ullman, *Mining of Massive Datasets* (Cambridge Univ. Press, New York, 2011).
37. A. Vespignani, Predicting the behavior of techno-social systems. *Science* **325**, 425–428 (2009). doi: 10.1126/science.1171990; pmid: 19628859
38. This statement mainly applies to microeconomics; there is more work in time-series macroeconomics that uses such methods.
39. A. Belloni, D. Chen, V. Chernozhukov, C. Hansen, Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–2429 (2012). doi: 10.3982/ECTA9626
40. A. Belloni, V. Chernozhukov, C. Hansen, "Inference on treatment effects after selection amongst high-dimensional controls," Cemmap working paper no. CWP10/12 (2012); http://arxiv.org/abs/1201.0224.
41. H. Varian, Machine learning: New tricks for econometrics. *J. Econ. Perspect.* **28**, 3–28 (2014). doi: 10.1257/jep.28.2.3
42. D. Lazer *et al*., Computational social science. *Science* **323**, 721–723 (2009). doi: 10.1126/science.1167742 pmid: 19197046
43. R. Lewis, D. Reiley, Online ads and offline sales: Measuring the effects of retail advertising via a controlled experiment on Yahoo! *Quant. Mark. Econ.* **12**, 235–266 (2014). doi: 10.1007/s11129-014-9146-6
44. R. A. Lewis, J. M. Rao, "The unfavorable economics of measuring the returns to advertising," working paper (2014); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2367103.
45. R. Albergotti, "Facebook experiments had few limits," *Wall Street Journal*, 2 July 2014, http://online.wsj.com/articles/facebook-experiments-had-few-limits-1404344378.
46. L. Einav, J. Levin, "The data revolution and economic analysis," in *Innovation Policy and the Economy*, J. Lerner, S. Stern, Eds. (Univ. of Chicago Press, Chicago, 2014), vol. 14, pp. 1–24.

ACKNOWLEDGMENTS

Parts of this Review draw on an earlier article (*46*). We have benefited from discussions with S. Athey, P. McAfee, and H. Varian. We acknowledge research support from the NSF, the Alfred P. Sloan Foundation, and the Toulouse Network on Information Technology.

10.1126/science.1243089