CHAPTER 2

Is Big Data a Big Deal for Applied Microeconomics?

Jesse M. Shapiro

While applications of "big data" methods to social data have exploded, applications to social science have not. I discuss why, and I review some recent applications of big data methods to applied microeconomics. I speculate on opportunities to bring more big data methods into applied economics, and on opportunities to bring more economics to big data.

1 INTRODUCTION

There are $n + 1$ units. The first n units are the training sample. The remaining unit is the target. For each unit i in the training sample, a social scientist measures an r-dimensional response y_i, a b-dimensional vector of background covariates x_i, and a p-dimensional vector of policies z_i. After observing the training sample and the target covariates x_{n+1}, the social scientist chooses the target policy z_{n+1}.

Data are called "big" when scale makes it impossible to use methods we might use easily on a smaller-scale problem. To fix ideas, think of ordinary least squares (OLS) as the canonical "small data" tool. The data may be big in the sense that we cannot load the product of the design matrix into working memory and invert it (large n). Or the data may be big in the sense that there are so many covariates that the OLS solution is ill-defined ($b > n$). Or the data may be big in the sense that the number of covariates grows with the sample (b growing as n grows).

The term "big data" is commonly applied to all of these cases, but especially to large n. The term "high-dimensional" is often applied to large r, b, or p.

Ongoing improvements in information and communication technology are making large-scale data increasingly available to governments and private

Article in preparation for the Eleventh World Congress of the Econometric Society. I benefited enormously from working on the Summer 2013 NBER Methods Lectures on High-Dimensional Data with my co-organizer Matthew Gentzkow and the other lecturers, Victor Chernozhukov, Christian Hansen, and Matt Taddy.

actors, and hence to researchers. Much of this data – on Internet browsing habits, search queries, payment method use, interactions via social media, etc. – is social in nature and of obvious interest to social scientists. Yet most of the methodological advances in big data are occurring outside the social sciences.

In this article, I discuss the strengths and weaknesses of frontier big-data methods as tools for social science. I review recent applications in applied microeconomics. I speculate on some possible opportunities for future applications of big-data methods in applied microeconomics. And I point out how microeconomics may be useful in developing big-data tools.

My goal is not to review statistical or econometric methods for high-dimensional data or computational methods for massive data. These techniques are reviewed in detail elsewhere (Dean and Ghemawat, 2008; Belloni et al., 2013; Friedman et al., 2013). Rather, my goal is to discuss how they can be applied to answer social science questions. I discuss the methods themselves only when this gives useful context for an application.

This article joins, and hopefully complements, a growing set of commentaries on big-data or machine-learning methods in economics, such as Einav and Levin (2014), who emphasize opportunities arising from large-scale data (large n), and Kleinberg et al. (2015), who emphasize the value of predictive methods for policy design. Varian (2014) reviews big-data methods and advocates their adoption in economics. Athey (2015) reviews work using machine-learning methods for causal inference.

2 WHAT MACHINES ARE GOOD AT

The best chess players in the world today are human–machine teams, dubbed "centaurs" by Garry Kasparov (Kelly, 2014). Many of the best empirical social scientists are also centaurs, and nearly all social science today involves a collaboration between at least one human mind and at least one CPU.

There is no question that machines, and the algorithms that they use, are getting better at a fast rate. So why have CPUs not (yet) replaced social scientists, allowing untrained operators to defeat the grandmasters?

It is helpful to flesh out the social scientist's problem. I follow ideas in Marschak (1950). Suppose that the training sample is drawn independently according to an unknown distribution $F(y_i|x_i, z_i)$. The target is similarly governed by an unknown $G(y_{n+1}|x_{n+1}, z_{n+1})$. After choosing z_{n+1}, the social scientist realizes a loss $L(y_{n+1}, z_{n+1})$.

The goal is to use the information in the training sample to minimize some appropriate ex ante representation of the loss, such as its expectation (if the loss is measurable with respect to the social scientist's beliefs) or its upper bound (if the goal is robust control), or its limiting mean or value as $n \to \infty$ at some rate.

What machines are good at is learning F. At some level, this statement is vacuous, as for any learning algorithm there will exist some data that will "trick" it into performing badly. But in practice, many different problems can be approached with related methods, and, as machines get faster and

algorithms improve, it is increasingly possible to use "off-the-shelf" methods to describe relationships in data.

To illustrate, I took a canonical economic dataset, the NLSY79 (Bureau of Labor Statistics, 2015), and applied a machine-learning algorithm to the problem of predicting family income in 2012 using only variables available in 1979. The algorithm that I chose is called a random forest. A random forest is grown as follows. Draw a bootstrap replicate of the data with replacement. Pick a subset of the variables. Grow a tree by splitting the data according to whichever split achieves the maximum reduction in sum of squared residuals within the resulting subsamples, then proceed iteratively through the remaining variables, splitting the data successively. At each resulting "leaf" the predicted value of the dependent variable is its mean in the leaf. Now average the predictions across all the bootstrap replicates to form a final prediction. Figure 1 illustrates a hypothetical tree in such a forest.

I excluded only those variables that had fewer than 5,000 non-missing observations. I did no data cleaning; for example, I did not treat missing data flags (-1, -4) separately from actual coded responses. I used a popular R package for fitting random forests and estimated using "factory" settings. I withheld 500 observations to test the predictions and applied the algorithm to the remaining observations.

The algorithm's output makes some economic sense. Figure 2 provides one view of the algorithm's output: the relative importance of different variables. The algorithm identifies the panelist's schooling expectations, family income in 1978, and father's and mother's education levels as the most important predictors of 2012 family income.

The algorithm does not overfit the data. The R^2 in the training sample of 0.197 is only slightly above the R^2 of 0.196 in the test sample.

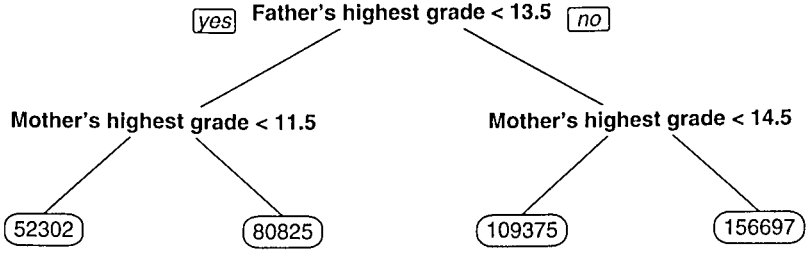

Figure 1 Hypothetical tree for predicting 2012 family income in the NLSY79

Notes: The plot depicts a hypothetical tree for predicting 2012 family income in the NLSY79 using father's and mother's highest grade completed as reported in 1979. The first branch splits the sample according to whether the panelist's father completed more than one year of college. For panelists whose father completed one year of college or less, the second branch splits the sample according to whether the panelist's mother completed high school. For panelists whose father completed more than one year of college, the second branch splits the sample according to whether the panelist's mother completed more than two years of college. Each leaf shows the mean 2012 family income in the relevant segment of the data.

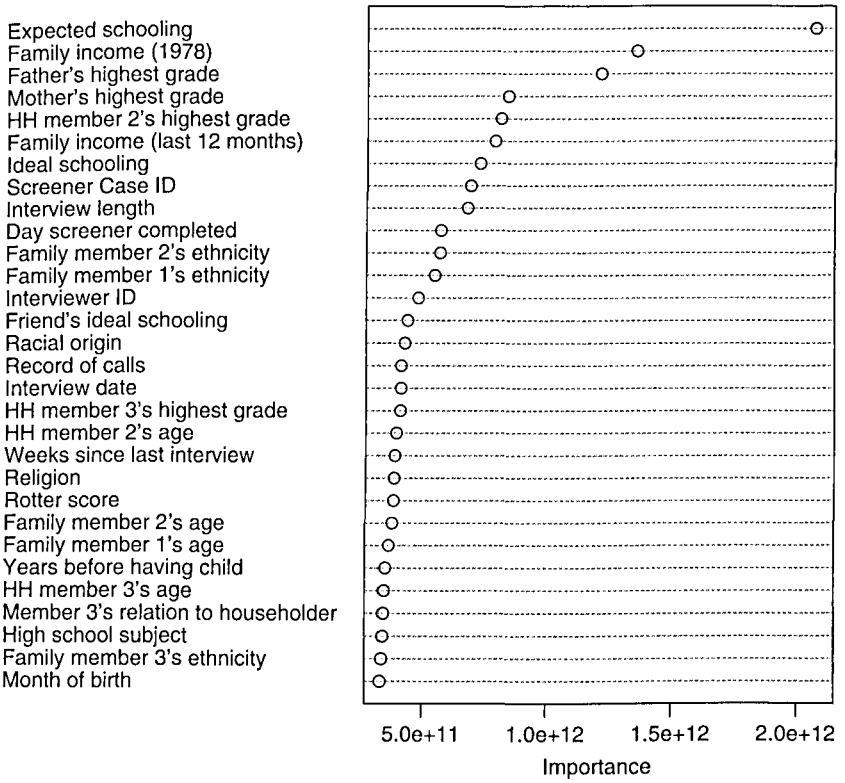

Figure 2 Important variables for predicting 2012 family income in the NLSY79

Notes: The plot depicts the importance of each of the 30 survey variables available in 1979 that are estimated to be most important for predicting 2012 family income in the NLSY79. The sample includes all panelists with non-missing family income in 2012. The variables used in prediction are all non-weight, non-identifier variables with non-missing data for at least 5000 panelists. The model is fit using the R package `randomForest` with default settings. The importance of a variable is measured by the total reduction in sum of squared errors, across all trees, from splitting according to the given variable.

In this sense, the algorithm performs well: it avoids overfitting while identifying meaningful relationships in the data that a social scientist interested in the determination of 2012 family income would likely want to understand.

But finding meaningful, robust relationships in data is only part of the social scientist's task. To translate those relationships into a choice of policy requires another, critical step: that of relating F to G. Suppose that before seeing the data the social scientist knows that $F \in \mathscr{F}$, $G \in \mathscr{G}$, and $(F, G) \in \mathscr{S} \subset \mathscr{F} \times \mathscr{G}$, where $\mathscr{F}$ and $\mathscr{G}$ are sets of possible CDFs and $\mathscr{S}$ is a subset of their Cartesian product. (In a Bayesian setup we would also impose a measure on $\mathscr{F} \times \mathscr{G}$.)

We might call $\mathscr{S}$ our *scientific assumptions*. These assumptions tell us how to translate what we learn about F to what we should believe about G. Without

well-specified scientific assumptions, learning F is not going to do us much good.

Naive defaults for $\mathscr{S}$ are often terrible. A natural one is *stationarity*, i.e., that

$$\mathscr{S} = \{(F, G) \in \mathscr{F} \times \mathscr{G} : F = G\}, \tag{1}$$

or that the target follows the same process as the training sample. In the NLSY case, for example, this means we assume that the relationship between self-reported schooling expectations and future income will remain intact if the policymaker manipulates expectations. While manipulating expectations may well have some effect, it seems likely that a sizable portion of the predictive power of expectations comes from the fact that these expectations summarize the panelist's information about her abilities and circumstances: information that would not change in response to a policy designed to encourage optimism. Put differently, the process that governs the determination of z_i in the training sample is *not* the one that governs it in the target.

There are important special cases in which stationarity is a reasonable assumption. One of these is *experimentation*. If the machine is allowed to change the policy in the training sample, then it can learn how the policy affects the response at different values of the covariates, and hence (if the target is the same kind of unit as those in the training sample) form a good representation of G. Machines have indeed been used to design and implement experiments (e.g., Gramacy and Lee, 2009), and McKinsey has recommended the establishment of internal "test factories" to manage testing of algorithmic marketing strategies (Goff et al., 2012).

Another important special case is *prediction*. A pure prediction problem has $F(y_i|x_i, z_i) = F(y_i|x_i)$: the weather forecast does not change the weather. This means that it is not important to know how z_i is determined in the training sample. The Netflix Prize (to predict users' movie ratings) is a famous recent example of a pure prediction problem with social data. (Gavin Potter, a retired management consultant, was an early leader in the competition for the Prize; see Ellenberg, 2008.)

Cases of pure prediction are very special within the social sciences: while the weather does not know it is being watched, the units in social scientists' models often do. Self-reported schooling expectations predict later family income in the NLSY. How would this relationship hold up if a predictive model of earnings based on self-reported schooling expectations were used to set tax rates?

That the economic environment responds to the policymaker's presence and choices is one of the most important ideas in the social sciences (Marschak, 1950; Lucas, 1976). It distinguishes social processes from physical ones, and illustrates why human intelligence is often needed to govern the use of machine learning for policy.

The most important role for social sciences occurs when z_i is not observed at all in the training sample, or does not vary, or varies only within a limited

range. In such cases, even under stationarity, some significant structure is needed to make useful statements about G.

The social sciences exist in part to provide that structure (Marschak, 1950). An economist with annual data on prices and quantities of an agricultural commodity, along with world GDP and rainfall, can predict how a never-before-seen tax will affect the market. This is alchemy, and a machine in possession of the key ingredients is unlikely to get it right (Working, 1927).

To summarize, machines increasingly excel at learning relationships among complex data. But social scientists excel (at least by comparison) at knowing how to use those relationships to make choices in a novel environment. I now review some recent applications in which social scientists have leveraged big-data or machine-learning methods, with an emphasis on applied microeconomics.

3 RECENT APPLICATIONS

3.1 High-Dimensional Response (Large r)

Gentzkow and Shapiro (2010) are interested in whether the political orientation of a newspaper's portrayal of the news is affected by the newspaper's owner. In the USA and most developed countries, regulators restrict ownership of news media to maintain diversity of content, so knowing how owners affect content is useful for policy decisions.

Here a unit is a newspaper, z_i is the owner's political ideology, and x_i is a vector of market characteristics such as the voting behavior of the consumers in the newspaper's local market. The high dimension is in the response y_i. A newspaper is a rich object with pictures, layout, text, bylines, headlines, etc. Gentzkow and Shapiro (2010) reduce this complexity by modeling a newspaper as a "bag of words" in which y_i is a vector of counts of all two- and three-word phrases. Even with this simplification, which discards a tremendous amount of information, the dimension r of the response is easily in the millions.

Gentzkow and Shapiro (2010) proceed by envisioning that each newspaper can be characterized by a one-dimensional latent ideology $\tilde{y}_i$ that they call the newspaper's "slant." If $\tilde{y}_i$ can be recovered for each newspaper, then standard methods can be used to relate $\tilde{y}_i$ to z_i and x_i. To recover $\tilde{y}_i$, Gentzkow and Shapiro (2010) borrow an idea from Groseclose and Milyo (2005).

Groseclose and Milyo (2005) are interested in recovering $\tilde{y}_i$ for a sample of news outlets. To do this they assume that

$$y_i \sim MN(q_i) \tag{2}$$

$$q_{ij} = \frac{\exp(\alpha_j + \beta_j \tilde{y}_i)}{\sum_{l=1}^{r} \exp(\alpha_l + \beta_l \tilde{y}_i)}.$$

To estimate $\tilde{y}_i$ they further impose (i) that the same model of speech is applicable to members of congress as to news outlets, (ii) that $\tilde{y}_i$ is known for congresspeople (for example, from their roll-call voting records), and that (iii) y_i consists of the number of citations to each of a set of policy think tanks. Because of (iii), $r = 200$, and the model can be estimated by maximum likelihood.

Gentzkow and Shapiro (2010) adopt (i) and (ii) but not (iii). Because Gentzkow and Shapiro (2010) wish to use the full set of phrases, they cannot simply estimate (2) via maximum likelihood: with millions of phrases, the model would be overfit, and in any case this would not be computationally practical. Instead, Gentzkow and Shapiro (2010) use a statistical test to identify the 10,000 phrases that are most diagnostic of a congressperson's party, and then use a method called *partial least squares*, motivated by a linear analogue of (2), to estimate $\tilde{y}_i$.

Taddy (2013) shows how to estimate (2) without restriction (iii), and without restricting the number of phrases in a first step as in Gentzkow and Shapiro (2010). Taddy (2013) imposes Laplace prior on the β_js. This choice of prior makes the estimated coefficients *sparse* in the sense that many estimated β_js are zero. This avoids the problem of overfitting and has the nice feature that phrase selection and estimation are implemented in a single step. Taddy (2013) shows that his method outperforms Gentzkow and Shapiro's (2010) method in guessing a congressperson's political party from the 10,000 phrases in Gentzkow and Shapiro's (2010) dataset.

3.2 High-Dimensional Covariates (Large b)

3.2.1 Prediction

Kelly and Pruitt (2013) are interested in whether the US annual equity market return is predictable. Different theories of asset pricing have different implications regarding how the expected return on risky assets varies over time. Forecasting returns based on information available to the market is one way to measure expected returns.

Here a unit is a year, z_i is a forecast, y_i is the (future) market return, and x_i is a vector of book-to-market ratios of a set of equity portfolios.

The high dimension is in the covariates x_i. Kelly and Pruitt (2013) study annual equity returns from 1930 to 2010 ($n = 81$), and use as many as 100 portfolios ($b = 100$). Because $b > n$, OLS is inappropriate: there are infinitely many coefficient vectors that give $R^2 = 1$ in sample, but these will perform poorly out of sample (Goyal and Welch, 2008).

One approach is to lower b by, say, focusing only on the book-to-market ratio of the aggregate equity market. This would permit use of OLS as in Pontiff and Schall (1998). Such an approach clearly discards a lot of information, as it may be that some equity portfolios are more predictive of aggregate market returns than others.

Kelly and Pruitt (2013) instead use partial least squares to aggregate the predictive information across portfolios, much as Gentzkow and Shapiro (2010) used it to aggregate information across phrases. The first step is to compute the direction d_i:

$$d_i = \sum_j \text{corr}\left(x_j, y\right) x_{ij}, \tag{3}$$

where $\text{corr}\left(x_j, y\right)$ is the correlation between future market returns and the book-to-market ratio of portfolio j. The forecast z_i is found as the predicted value from an OLS regression of y_i on d_i.

Kelly and Pruitt (2013) find that moving from $b = 6$ to $b = 100$ more than doubles the out-of-sample R^2 of the prediction, illustrating the value of increasing the dimension of the covariates. In a companion paper, Kelly and Pruitt (2015) generalize their method, develop its asymptotic properties, and show that it outperforms some common alternatives such as principal components regression.

Intuitively, partial least squares "works" because the index d_i depends the most on the portfolios that are most predictive of the aggregate return. Unlike, say, principal components regression, which begins by reducing the covariates to a lower-dimensional set of underlying factors, partial least squares is designed to "find" the factors that best predict the target y_i.

3.2.2 Instruments

Estimating market expectations is a natural social science application of methods designed to maximize fit. Another is instrumental variables estimation. Often, more instruments are available than can be practically employed by researchers, so some dimension-reduction is needed. As in prediction contexts, it is reasonable to think of the goal as maximizing the (out-of-sample) fit of the "first stage" that predicts endogenous variables from exogenous ones.

Belloni et al. (2012) want to know how eminent domain policies affect the housing market. Eminent domain policies may be endogenous to housing market characteristics, so Belloni et al. (2012) exploit the fact that appellate court judges are randomly assigned to three-member panels. Because an appellate court decision makes legal precedent for the court's circuit, the randomization of judges provides a set of suitable instruments for local eminent domain policy.

Here a unit is a circuit (actually a circuit-year, but the simplification is innocent), z_i is an index of eminent domain policy, y_i is the growth in housing prices, and x_i are characteristics of judges assigned to appellate cases in the circuit.

The high dimension is in the instruments x_i. Judges are characterized by gender, race, religion, political affiliation, and various indicators of education and career history, along with some interactions of these, leading to $b > 130$ possible instruments and $n < b$ for some models. Belloni et al. (2012) propose

to fit the first stage using the lasso (Tibshirani, 1996), i.e., by solving the program

$$\min_{\beta} \left[\sum_{i} (z_i - \beta x_i)^2 + \lambda \|\beta\|_1 \right] \tag{4}$$

where $\|\cdot\|_1$ denotes the L_1 norm and $\lambda > 0$ is a penalty parameter. Relative to the two-stage least squares approach, which has $\lambda = 0$, the program in (4) differs in that it shrinks estimates of β toward zero and also (in general) sets some estimates to exactly zero. In some variants (the post-lasso), Belloni et al. (2012) re-fit the first stage with no penalty but using only the lasso-selected instruments (i.e., those with nonzero estimated coefficients).

Lasso-selected instruments outperform hand-selected ones in first-stage predictive power and tend to deliver more precise estimates of the effect of eminent domain law z_i on house prices y_i.

The problem of high-dimensional instruments is more general than it may seem at first because a given exogenous process may be measured in many ways. A good example is the weather. Weather shocks have been used diversely as instruments to estimate the demand for fish at the Fulton fish market (Graddy, 1995), the effect of economic growth on civil war (Miguel et al., 2004), and the persistence of local criminal activity (Jacob et al., 2007).

The weather is multidimensional and different dimensions matter for different economic variables. Researchers typically solve this problem by using prior knowledge of the setting, for example, that wave height is important for fishing or that rainfall is important for agriculture. Gilchrist and Sands (2016) adopt a data-driven approach that uses Belloni et al.'s (2012) post-lasso method. Selecting from a set of 52 possible weather features, the lasso-fit first stage of their estimator reaches the intuitive conclusion that people do not like to be indoors at the movies when it is 75 degrees outside. Gilchrist and Sands (2016) find that weather-driven variation in opening-weekend movie sales has large and persistent effects on demand in subsequent weeks, implying social spillovers in demand.

3.3 High-Dimensional Policies (Large p)

In some cases the policymaker controls a high-dimensional policy z_i and desires a representation of the effect of z_i on the outcome y_i that permits a good choice of z_{n+1}, say one that maximizes the expected value of y_{n+1}. If p is large relative to n, methods such as OLS are inappropriate.

In the context of cross-country growth, Mankiw (1995) calls this the "degrees of freedom problem," noting that there are far more factors that may reasonably influence growth than there are countries (or even country-years). Levine and Renelt (1992) and Sala-i-Martin (1997) address this problem by searching over all possible specifications of cross-country growth models and identifying variables that are robustly important in a particular sense. More

recent approaches, surveyed in Durlauf et al. (2005), use Bayesian model averaging techniques that impose a prior over the space of possible growth models.

I am not aware of many microeconometric studies that use similar methods, but similar problems do arise in microeconomic settings. For example, Bertrand et al. (2010) randomized the content of direct-mail advertising for a loan product in South Africa along eight feature dimensions. Although Bertrand et al. (2010) are interested in testing substantive theories and not in optimal feature design, the design problem is a natural one. That problem is high-dimensional because the independent randomization across eight dimensions generates thousands of treatment conditions for the $n = 53,194$ households in the study. With no restrictions on the combinations of interest, then $p > 1000$ and the number of observations per policy option is small even though the data are large.

Using data from Bertrand et al. (2010), I computed the mailer design that maximizes the probability of loan application using three models: a probit with no interactions (similar to the models estimated in Bertrand et al. 2010), a fully nonparametric model that computes the mean for each treatment cell, and a binomial lasso that includes all first-order interactions among the mailer attributes. Each model focuses on design elements that (I presume) can be varied at no cost to the lender. I estimated each model on 48,194 observations, holding out the remaining 5,000 observations to serve as a test sample.

Table 1 presents the results. The three models agree on some things. For example, all three think that the race of the advertising photo shown in the mailer should match the client's race. However, the models differ in the number of elements they recommend including in the mailer, with the naive model recommending more than the probit model, and the lasso recommending the most of all. Not surprisingly, the out-of-sample fit of the naive model is worse than that of the probit model. The lasso performs the best out of sample of the three models, even though it is fit using many more covariates.

As another example, Dobbie and Fryer (2013) study the effect of $p = 31$ charter school policies z_i on a measure of effectiveness y_i for $n = 39$ schools. Here again, p is large relative to n, making it difficult to learn the independent effect of each policy, even if we stipulate that unobservable school policies are not correlated with the observable ones. Dobbie and Fryer (2013) reduce the dimension of the problem by combining related measures into scalar indices. To illustrate an alternative approach, Figure 3 plots the coefficient path from a lasso analysis of Dobbie and Fryer's (2013) replication data[1]. When the penalty λ is large, few coefficients are nonzero; as λ approaches zero, the coefficients approach their OLS values.

The coefficient path shows that the three variables that the lasso selects first – setting high expectations, frequent teacher feedback, and extra instructional

[1] Due to confidentiality concerns only a subset of variables are available.

Table 1 *Optimal mailer design for Bertrand et al. (2010)*

Model	Probit	Naive	Lasso
Feature included in optimal mailer?			
No photo			
Photo gender matches client's		X	
Photo race matches client's	X	X	X
One example loan shown			
Interest rate shown			
Cell phone raffle mentioned			X
No specific loan use mentioned			X
Comparison to competitor rate		X	X
Loss frame comparison		X	
We speak your language			X
A low or special rate for you	X		X
Mean likelihood in test sample:			
Successes	0.0872	0.0894	0.0889
Failures	0.0847	0.0874	0.0846
Log ratio of likelihoods	0.0297	0.0228	0.0494
Size of training sample	48194	48194	48194
Size of test sample	5000	5000	5000

Note: All models have as their outcome variable an indicator for whether the household applied for a loan before the mailer deadline. All models have as their explanatory variables the indicators listed as well as indicators for gender and race of photo. The choice of explanatory variables follows column (1) of Table III of Bertrand et al. (2010) but excludes randomization controls and the interest rate. The probit model is a binary probit with no interactions. The naive model is a fully saturated model with all interactions. Interactions present in the test sample but not in the training sample have likelihood equal to the training sample mean. The lasso model is a binomial-family lasso model with a logit link whose penalty parameter is chosen based on tenfold cross-validation.

time – are among the five variables highlighted in Dobbie and Fryer's (2013) abstract as important for charter school success. Although the lasso model cannot substitute for Dobbie and Fryer's (2013) domain knowledge, that the plot agrees with many of their findings suggests that tools like the lasso might serve as a nice complement to other forms of analysis when there is a need to explore or visualize relationships in high-dimensional data.

An important caveat is that learning the effect of a high-dimensional policy is econometrically difficult. Nickl and van de Geer (2013) show that uniform inference on the effect of high-dimensional policies can be impossible even when attractive estimators exist. In a sense this result contrasts with the encouraging findings of Chernozhukov et al. (2015) for the case where the object of interest is low-dimensional but there is a high-dimensional nuisance parameter (as in the instrumental variables applications in section 3.2.2).

High-dimensional policies do arise in important microeconomic settings, so researchers will need to confront the resulting challenges. Signs of progress include Athey and Imbens' (2015) model-selection methods for estimation and inference on heterogeneous treatment effects.

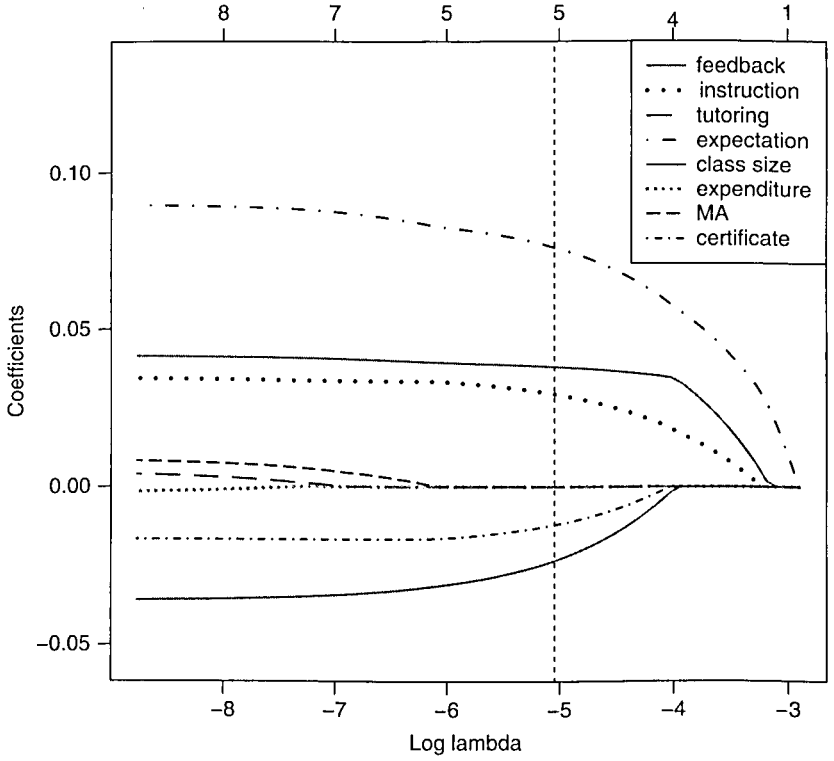

Figure 3 Coefficient path for Dobbie and Fryer (2013)

Notes: The plot shows estimated coefficients from the lasso (Tibshirani, 1996) at various levels of the L_1 penalty (λ). The vertical dashed line shows the value of log (λ) that minimizes deviance in tenfold cross-validation. The unit of analysis is a charter school. Two charter schools are omitted due to missing values in the included school characteristics ($n = 37$). The dependent variable is Dobbie and Fryer's (2013) nonexperimental measure of the improvement in math score from attending the school. The independent variables are indicators for the following characteristics, with bold words corresponding to the short labels in the plot: frequent teacher **feedback**, extra **instruction** time, high quality **tutoring**, high **expectations**, small **class size**, high per-pupil **expenditures**, high fraction of teachers with **MA** degrees, low fraction of teachers without **certification**.

3.4 Large-Scale Data (Large n)

3.4.1 Subsampling

Einav et al. (2015) are interested in how the start price of an eBay auction affects the sale price and the likelihood of sale, which together determine the seller's expected revenue. Estimates of sellers' marginal revenue curves permit making and testing predictions about the start prices that would be chosen by optimizing sellers.

Here a unit is an auction, y_i is the seller's revenue, z_i is the start price, and x_i is a set of characteristics of the seller or item, such as the seller's identity or

item title. The universe is the set of all non-real estate, non-auto eBay listings in 2009.

The data here are "big" in two respects. First, n is very large: well into the hundreds of millions. Second, x_i is very rich, as there are millions of unique items and sellers.

Einav et al. (2015) address both the scale and the high dimension in the same way: by identifying matched sets of auctions for which z_i varies while x_i does not. This avoids the need to model the role of the characteristics in x_i, and it results in a manageable sample of $n = 7.7m$ listings. Einav et al. (2015) also experiment with imposing tighter criteria on matched sets to confirm that remaining variation in x_i does not confound estimates of the effect of z_i.

Although the data that Einav et al. (2015) ultimately work with is small enough to fit into working memory (their analysis code was written in *Stata*), scale is crucial to their research design. Many of the interesting tests that they present are based on several thousand listings. Were eBay data one order of magnitude smaller, many of these tests would not be informative.

Exploiting scale to limit confounds or to suit a particular research design is a common strategy. For example, Bronnenberg et al. (2015) study whether more informed shoppers are more likely to buy store-brand products at the grocery store. The underlying purchase data includes hundreds of millions of transactions. Bronnenberg et al. (2015) focus on the subset of products for which a store-brand alternative exists that shares all of its measured attributes in common with the national brand. As in Einav et al. (2015), focusing on this subset of the data makes the data easier to work with and simplifies the analysis by limiting the amount of variation in product characteristics.

As another example, Hastings and Shapiro (2013) study the choice of octane level in a panel of households tracked by a retailer. From a universe of 1.3m households they extract data on 61,000 households who buy gasoline frequently at the retailer's stations. Focusing on this subset of households makes it possible to allow for rich heterogeneity in preferences.

An obvious drawback of focusing on a subset of the data for analysis is a lack of representativeness. Often researchers document that important patterns hold up in representative data. For example, Hastings and Shapiro (2013) show that the patterns they document are present in representative survey data collected by the US government. Einav et al. (2014) study the effect of sales taxes on eBay purchases, using both a detailed analysis of a subset of transactions, and an aggregate analysis of state-to-state purchase flows.

An alternative way to gauge representativeness would of course be to apply the same micro-econometric techniques to the universe of data that are applied to the subset of interest. For the applications I have discussed, that would involve a change in computing methods. I expect that in the future we will see more researchers use parallelization methods such as MapReduce (Dean and Ghemawat, 2008) to scale their analysis.

3.4.2 Designing for Parallelization

Gentzkow et al. (2015) are interested in estimating trends in the partisanship of Congressional speech; more specifically, they wish to know whether Republicans and Democrats in Congress speak more differently from one another today than in the past. They measure the frequency with which each speaker in Congress uses each two-word phrase in each session from 1872 to 2009. Their data include on the order of 10^6 unique phrases spoken a total of 10^8 times in 10^4 speaker-sessions. Because of the high dimension of the response (large r), naive adaptations of standard segregation metrics perform poorly on these data.

Gentzkow et al. (2015) propose to estimate trends in segregation by adapting Taddy's (2013) multinomial logit model. To the basic model in equation (2), Gentzkow et al. (2015) add a rich set of covariates, a lasso-type penalty that imposes sparsity on the loadings (β_js), and a further penalty that regularizes the evolution of the loadings over time.

Direct estimation of Gentzkow et al.'s (2015) model on their data is impossible. The authors adopt a suggestion of Taddy (2015) to approximate the multinomial logit by a Poisson count model. The advantage of the Poisson approximation is that, up to a nuisance parameter, its likelihood factors across phrases. This means that computation can be separated across phrases, making distributed computing possible. Gentzkow et al. (2015) show that the Poisson approximation performs well in sampling experiments for their problem.

Parallel computing is by no means new to economics (Aldrich et al., 2011). I expect, however, that as more applied microeconomists encounter large-scale data, parallelizability will become an increasingly important consideration in designing econometric models.

4 FUTURE APPLICATIONS

I have already discussed some possible future directions for applying big-data methods to microeconomic problems. Here I speculate on several other possible applications, and on some ways that economics may contribute to the development of robust methods.

4.1 Measuring Agents' Beliefs

Many estimators for dynamic microeconomic models (e.g., Pakes et al., 2007) begin by approximating agents' beliefs about the future as a function of a set of state variables. Under rational expectations this can be done by relating realized outcomes to information known to agents at the time of decision-making. Estimates of model parameters are sensitive to assumptions about which state variables are known to the agents (Dickstein and Morales, 2015). This fact puts a premium on good selection of state variables. Forming a good statistical model of the future based on a (possibly large) set of variables is the sort

of problem machines are now very good at. Applying modern machine learning tools might allow researchers to better approximate agents' beliefs and to include more state variables without overfitting. Along these lines, Bajari et al. (2015) apply machine-learning methods to compute policy improvements for an agent in a dynamic store entry setting.

4.2 Designing Dynamic Experiments

Economists are increasingly using randomized experiments to test social science hypotheses. Typically, these experiments are static in the sense that the design and hypotheses are specified in advance and then estimated. However, nothing prevents a researcher from changing the sampling scheme for an experiment in response to initial data. For example, a researcher studying direct-mail for consumer credit as in Bertrand et al. (2010) could stop using a content feature after its effect has already been learned, or could design an experiment to dynamically pursue the most effective content. Such methods have a long pedigree (Wald, 1945), and flexible modern tools are available for active learning and for optimization based on statistical evidence (e.g., Taddy et al., 2011). I predict that economists will soon begin to use these methods to design experiments whose designs adapt to data in real time.

4.3 Incentive-Compatible Machine Learning

Many canonical machine-learning problems, such as credit-scoring, involve incentives in a fundamental way. Which items should be included in a credit score is not solely a matter of predictive fit; it also depends on how the households being scored will respond to the scoring system (Frankel and Kartik, 2014). For example, the color of the family car might be a bad input into a credit score even if it is a good predictor of default, because it is so easily changed. Building tools for statistical learning that impose parameter constraints based on incentive as well as statistical or computational considerations seems an interesting direction for future work.

References

Aldrich, Eric M., Jesús Fernández-Villaverde, A. Ronald Gallant, and Juan F. Rubio-Ramírez. 2011. "Tapping the Supercomputer Under Your Desk: Solving Dynamic Equilibrium Models with Graphics Processors." *Journal of Economic Dynamics and Control* 35(3): 386–393.

Athey, Susan. 2015. "Machine Learning and Causal Inference for Policy Evaluation." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 5–6.

Athey, Susan and Guido W. Imbens. 2015. "Machine Learning Methods for Estimating Heterogeneous Causal Effects." Stanford University Working Paper.

Bajari, Patrick, Ying Jiang, and Carlos A. Manzanares. 2015. "Improving Policy Functions in High-Dimensional Dynamic Games: An Entry Game Example." University of Washington Working Paper.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian B. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica* 80(6): 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian B. Hansen. 2013. "Inference for High-Dimensional Sparse Econometric Models." In Part III of Daron Acemoglu, Manuel Arellano, and Eddie Dekel (eds.), *Advances in Economics and Econometrics: Tenth World Congress, Volume III, Econometrics*. New York: Cambridge University Press.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment." *Quarterly Journal of Economics* 125(1): 263–306.

Bronnenberg, Bart J., Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M. Shapiro. 2015. "Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium." *Quarterly Journal of Economics* 130(4): 1669–1726.

Bureau of Labor Statistics, US Department of Labor. "National Longitudinal Survey of Youth 1979 cohort, 1979–2012." 2015. Accessed at www.nlsinfo.org/investigator/pages/login.jsp on June 28, 2015.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach." *Annual Review of Economics* 7: 649–688.

Dean, Jeffrey and Sanjay Ghemawat. 2008. "MapReduce: Simplified Data Processing on Large Clusters." *Communications of the ACM* 51(1): 107–113.

Dickstein, Michael J. and Eduardo Morales. 2015. "What do Exporters Know?" Working paper.

Dobbie, Will and Roland G. Fryer. 2013. "Getting Beneath the Veil of Effective Schools: Evidence from New York City." *American Economic Journal: Applied Economics* 5(4): 28–60.

Durlauf, Steven N., Paul A. Johnson, and Jonathan R.W. Temple. 2005. "Growth Econometrics." In Chapter 8 of Philippe Aghion and Steven N. Durlauf (eds.), *Handbook of Economic Growth, Volume 1A*. Amsterdam: Elsevier.

Einav, Liran, Dan Knoepfle, Jonathan Levin, and Neel Sundaresan. 2014. "Sales Taxes and Internet Commerce." *American Economic Review* 104(1): 1–26.

Einav, Liran, Theresa Kuchler, Jonathan Levin, and Neel Sundaresan. 2015. "Assessing Sale Strategies in Online Markets Using Matched Listings." *American Economic Journal: Microeconomics* 7(2): 215–247.

Einav, Liran and Jonathan Levin. 2014. "Economics in the Age of Big Data." *Science* 346(6210).

Ellenberg, Jordan. 2008. "This Psychologist Might Outsmart the Math Brains Competing for the Netflix Prize." *Wired Magazine*, February 25, 2008.

Frankel, Alex and Navin Kartik. 2014. "Muddled Information." Working paper.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2013. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78(1): 35–71.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2015. "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech." Working paper.

Gilchrist, Duncan S. and Emily G. Sands. 2016. "Something to Talk About: Social Spillovers in Movie Consumption." *Journal of Political Economy* 124(5): 1339–1382.

Goff, Joshua, Paul McInerney, and Gunjan Soni. 2012. "Need for Speed: Algorithmic Marketing and Customer Data Overload." *McKinsey & Company*, May 2012.

Goyal, Amit and Ivo Welch. 2008. "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction." *The Review of Financial Studies* 21(4): 1455–1508.

Graddy, Kathryn. 1995. "Testing for Imperfect Competition at the Fulton Fish Market." *The Rand Journal of Economics* 26(1): 75–92.

Gramacy, Robert B. and Herbert K. H. Lee. 2009. "Adaptive Design and Analysis of Supercomputer Experiments." *Technometrics* 51(2): 130–145.

Groseclose, Tim and Jeffrey Milyo. 2005. "A Measure of Media Bias." *Quarterly Journal of Economics* 120(4): 1191–1237.

Hastings, Justine S. and Jesse M. Shapiro. 2013. "Fungibility and Consumer Choice: Evidence from Commodity Price Shocks." *Quarterly Journal of Economics* 128(4): 1449–1498.

Jacob, Brian, Lars Lefgren, and Enrico Moretti. 2007. "The Dynamics of Criminal Behavior: Evidence from Weather Shocks." *The Journal of Human Resources* 42(3): 489–527.

Kelly, Bryan and Seth Pruitt. 2013. "Market Expectations in the Cross-Section of Present Values." *The Journal of Finance* 68(5): 1721–1756.

Kelly, Bryan and Seth Pruitt. 2015. "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors." *Journal of Econometrics* 186(2): 294–316.

Kelly, Kevin. 2014. "The Future of AI? Helping Human Beings Think Smarter." *Wired Magazine*, December 3, 2014.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review Papers and Proceedings* 105(5): 491–495.

Levine, Ross and David Renelt. 1992. "A Sensitivity Analysis of Cross-Country Growth Regressions." *American Economic Review* 82(4): 942–963.

Lucas, Robert. 1976. "Econometric Policy Evaluation: A Critique." In Karl Brunner and Allan H. Meltzer (eds.), *The Phillips Curve and Labor Markets, Volume 1 of Carnegie-Rochester Conference Series on Public Policy*. New York: American Elsevier.

Mankiw, N. Gregory. 1995. "The Growth of Nations." *Brookings Papers on Economic Activity* 26(1): 275–326.

Marschak, Jacob. 1950. "Statistical Inference in Economics: An Introduction." In Chapter 1 of Tjalling C. Koopmans (ed.), *Statistical Inference in Dynamic Economic Models*. New York: John Wiley & Sons, Inc.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 112(4): 725–753.

Nickl, Richard and Sara van de Geer. 2013. "Confidence Sets in Sparse Regression." *Annals of Statistics* 41(6): 2852–2876.

Pakes, Ariel, Steven Berry, and Michael Ostrovsky. 2007. "Simple Estimators for the Parameters of Discrete Dynamic Games (with Entry/Exit Examples)." *The Rand Journal of Economics* 38(2): 373–399.

Pontiff, Jeffrey and Lawrence D. Schall. 1998. "Book-to-market Ratios as Predictors of Market Returns." *Journal of Financial Economics* 49: 141–160.

Sala-i-Martin, Xavier X. 1997. "I Just Ran Two Million Regressions." *American Economic Review Papers and Proceedings* 87(2): 178–183.

Taddy, Matthew A., Robert B. Gramacy, and Nicholas G. Polson. 2011. "Dynamic Trees for Learning and Design." *Journal of the American Statistical Association* 106(493): 109–123.

Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108(503): 755–770.

Taddy, Matt. 2015. "Distributed Multinomial Regression." *Annals of Applied Statistics* 9(3): 1394–1414.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B (Methodological)* 58(1): 267–288.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28(2): 3–28.

Wald, Abraham. 1945. "Sequential Tests of Statistical Hypotheses." *The Annals of Mathematical Statistics* 16(2): 117–186.

Working, Elmer J. 1927. "What Do Statistical 'Demand Curves' Show?" *Quarterly Journal of Economics* 41(2): 212–235.