Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 3–28

Big Data: New Tricks for Econometrics[†]

Hal R. Varian

Computers are now involved in many economic transactions and can capture data associated with these transactions, which can then be manipulated and analyzed. Conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools.

First, the sheer size of the data involved may require more powerful data manipulation tools. Second, we may have more potential predictors than appropriate for estimation, so we need to do some kind of variable selection. Third, large datasets may allow for more flexible relationships than simple linear models. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning, and so on may allow for more effective ways to model complex relationships.

In this essay, I will describe a few of these tools for manipulating and analyzing big data. I believe that these methods have a lot to offer and should be more widely known and used by economists. In fact, my standard advice to graduate students these days is go to the computer science department and take a class in machine learning. There have been very fruitful collaborations between computer scientists and statisticians in the last decade or so, and I expect collaborations between computer scientists and econometricians will also be productive in the future.

■ *Hal Varian is Chief Economist, Google Inc., Mountain View, California, and Emeritus Professor of Economics, University of California, Berkeley, California. His email address is hal@ischool.berkeley.edu.*

[†] To access the Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.2.3

doi=10.1257/jep.28.2.3

Tools to Manipulate Big Data

Economists have historically dealt with data that fits in a spreadsheet, but that is changing as new more-detailed data becomes available (see Einav and Levin 2013, for several examples and discussion). If you have more than a million or so rows in a spreadsheet, you probably want to store it in a relational database, such as MySQL. Relational databases offer a flexible way to store, manipulate, and retrieve data using a Structured Query Language (SQL), which is easy to learn and very useful for dealing with medium-sized datasets.

However, if you have several gigabytes of data or several million observations, standard relational databases become unwieldy. Databases to manage data of this size are generically known as "NoSQL" databases. The term is used rather loosely, but is sometimes interpreted as meaning "not only SQL." NoSQL databases are more primitive than SQL databases in terms of data manipulation capabilities but can handle larger amounts of data.

Due to the rise of computer-mediated transactions, many companies have found it necessary to develop systems to process billions of transactions per day. For example, according to Sullivan (2012), Google has seen 30 trillion URLs, crawls over 20 billion of those a day, and answers 100 billion search queries a month. Analyzing even one day's worth of data of this size is virtually impossible with conventional databases. The challenge of dealing with datasets of this size led to the development of several tools to manage and analyze big data.

A number of these tools are proprietary to Google, but have been described in academic publications in sufficient detail that open-source implementations have been developed. Table 1 contains both the Google name and the name of related open-source tools. Further details can be found in the Wikipedia entries associated with the tool names.

Though these tools can be run on a single computer for learning purposes, real applications use large clusters of computers such as those provided by Amazon, Google, Microsoft, and other cloud-computing providers. The ability to rent rather than buy data storage and processing has turned what was previously a fixed cost of computing into a variable cost and has lowered the barriers to entry for working with big data.

Tools to Analyze Data

The outcome of the big-data processing described above is often a "small" table of data that may be directly human readable or can be loaded into an SQL database, a statistics package, or a spreadsheet. If the extracted data is still inconveniently large, it is often possible to select a subsample for statistical analysis. At Google, for example, I have found that random samples on the order of 0.1 percent work fine for analysis of business data.

Once a dataset has been extracted, it is often necessary to do some exploratory data analysis along with consistency and data-cleaning tasks. This is something

Table 1
Tools for Manipulating Big Data

Google name	*Analog*	*Description*
Google File System	Hadoop File System	This system supports files so large that they must be distributed across hundreds or even thousands of computers.
Bigtable	Cassandra	This is a table of data that lives in the Google File System. It too can stretch over many computers.
MapReduce	Hadoop	This is a system for accessing and manipulating data in large data structures such as Bigtables. MapReduce allows you to access the data in parallel, using hundreds or thousands of machines to extract the data you are interested in. The query is "mapped" to the machines and is then applied in parallel to different shards of the data. The partial calculations are then combined ("reduced") to create the summary table you are interested in.
Sawzall	Pig	This is a language for creating MapReduce jobs.
Go	None	Go is flexible open-source, general-purpose computer language that makes it easier to do parallel data processing.
Dremel, BigQuery	Hive, Drill, Impala	This is a tool that allows data queries to be written in a simplified form of of Structured Query Language (SQL). With Dremel it is possible to run an SQL query on a petabtye of data (1,000 terabytes) in a few seconds.

of an art, which can be learned only by practice, but data-cleaning tools such as OpenRefine and DataWrangler can be used to assist in data cleansing.

Data analysis in statistics and econometrics can be broken down into four categories: 1) prediction, 2) summarization, 3) estimation, and 4) hypothesis testing. Machine learning is concerned primarily with prediction; the closely related field of data mining is also concerned with summarization, and particularly with finding interesting patterns in the data. Econometricians, statisticians, and data mining specialists are generally looking for insights that can be extracted from the data. Machine learning specialists are often primarily concerned with developing high-performance computer systems that can provide useful predictions in the presence of challenging computational constraints. Data science, a somewhat newer term, is concerned with both prediction and summarization, but also with data manipulation, visualization, and other similar tasks. Note that terminology is not standardized in these areas, so these descriptions reflect general usage, not hard-and-fast definitions. Other terms used to describe computer-assisted data analysis include knowledge extraction, information discovery, information harvesting, data archaeology, data pattern processing, and exploratory data analysis.

Much of applied econometrics is concerned with detecting and summarizing relationships in the data. The most common tool used for summarization is (linear) regression analysis. As we shall see, machine learning offers a set of tools that can usefully summarize various sorts of nonlinear relationships in the data. We will focus on these regression-like tools because they are the most natural for economic applications.

In the most general formulation of a statistical prediction problem, we are interested in understanding the conditional distribution of some variable y given some other variables $x = (x_1, \ldots, x_P)$. If we want a point prediction, we can use the mean or median of the conditional distribution.

In machine learning, the x-variables are usually called "predictors" or "features." The focus of machine learning is to find some function that provides a good prediction of y as a function of x. Historically, most work in machine learning has involved cross-section data where it is natural to think of the data being independent and identically distributed (IID) or at least independently distributed. The data may be "fat," which means lots of predictors relative to the number of observations, or "tall" which means lots of observations relative to the number of predictors.

We typically have some observed data on y and x, and we want to compute a "good" prediction of y given new values of x. Usually "good" means it minimizes some loss function such as the sum of squared residuals, mean of absolute value of residuals, and so on. Of course, the relevant loss is that associated with *new* out-of-sample observations of x, not the observations used to fit the model.

When confronted with a prediction problem of this sort an economist would think immediately of a linear or logistic regression. However, there may be better choices, particularly if a lot of data is available. These include nonlinear methods such as 1) classification and regression trees (CART); 2) random forests; and 3) penalized regression such as LASSO, LARS, and elastic nets. (There are also other techniques, such as neural nets, deep learning, and support vector machines, which I do not cover in this review.) Much more detail about these methods can be found in machine learning texts; an excellent treatment is available in Hastie, Tibshirani, and Friedman (2009), which can be freely downloaded. Additional suggestions for further reading are given at the end of this article.

General Considerations for Prediction

Our goal with prediction is typically to get good *out-of-sample predictions*. Most of us know from experience that it is all too easy to construct a predictor that works well in-sample but fails miserably out-of-sample. To take a trivial example, n linearly independent regressors will fit n observations perfectly but will usually have poor out-of-sample performance. Machine learning specialists refer to this phenomenon as the "overfitting problem" and have come up with several ways to deal with it.

First, since simpler models tend to work better for out-of-sample forecasts, machine learning experts have come up with various ways to penalize models for excessive complexity. In the machine learning world, this is known as "regularization," and we will describe some examples below. Economists tend to prefer simpler models for the same reason, but have not been as explicit about quantifying complexity costs.

Second, it is conventional to divide the data into separate sets for the purpose of training, testing, and validation. You use the training data to estimate a model, the validation data to choose your model, and the testing data to evaluate how well your chosen model performs. (Often validation and testing sets are combined.)

Third, if we have an explicit numeric measure of model complexity, we can view it as a parameter that can be "tuned" to produce the best out of sample predictions. The standard way to choose a good value for such a tuning parameter is to use *k-fold cross-validation*.

1. Divide the data into k roughly equal subsets (folds) and label them by $s = 1, \ldots, k$. Start with subset $s = 1$.
2. Pick a value for the tuning parameter.
3. Fit your model using the $k - 1$ subsets other than subset s.
4. Predict for subset s and measure the associated loss.
5. Stop if $s = k$, otherwise increment s by 1 and go to step 2.

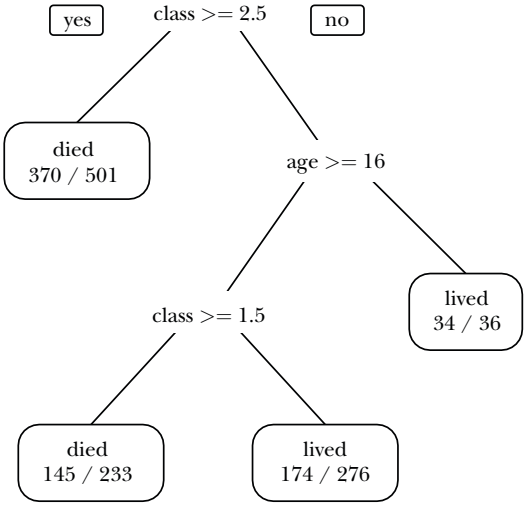
Common choices for k are 10, 5, and the sample size minus 1 ("leave one out"). After cross-validation, you end up with k values of the tuning parameter and the associated loss which you can then examine to choose an appropriate value for the tuning parameter. Even if there is no tuning parameter, it is prudent to use cross-validation to report goodness-of-fit measures since it measures out-of-sample performance, which is generally more meaningful than in-sample performance.

The test-train cycle and cross-validation are very commonly used in machine learning and, in my view, should be used much more in economics, particularly when working with large datasets. For many years, economists have reported in-sample goodness-of-fit measures using the excuse that we had small datasets. But now that larger datasets have become available, there is no reason not to use separate training and testing sets. Cross-validation also turns out to be a very useful technique, particularly when working with reasonably large data. It is also a much more realistic measure of prediction performance than measures commonly used in economics.

Classification and Regression Trees

Let us start by considering a discrete variable regression where our goal is to predict a 0–1 outcome based on some set of features (what economists would call explanatory variables or predictors). In machine learning, this is known as a

Figure 1
A Classification Tree for Survivors of the *Titanic*



Note: See text for interpretation.

classification problem. A common example would be classifying email into "spam" or "not spam" based on characteristics of the email. Economists would typically use a generalized linear model like a logit or probit for a classification problem.

A quite different way to build a classifier is to use a decision tree. Most economists are familiar with decision trees that describe a sequence of decisions that results in some outcome. A tree classifier has the same general form, but the decision at the end of the process is a choice about how to classify the observation. The goal is to construct (or "grow") a decision tree that leads to good out-of-sample predictions.

Ironically, one of the earliest papers on the automatic construction of decision trees (Morgan and Sonquist 1963) was coauthored by an economist. However, the technique did not really gain much traction until 20 years later in the work of Breiman, Friedman, Olshen, and Stone (1984). Nowadays this prediction technique is known as "classification and regression trees," or "CART."

To illustrate the use of tree models, I used the **R** package **rpart** to find a tree that predicts *Titanic* survivors using just two variables: age and class of travel.[1] The resulting tree is shown in Figure 1, and the rules depicted in the tree are shown in Table 2. The rules fit the data reasonably well, misclassifying about 30 percent of the observations in the testing set.

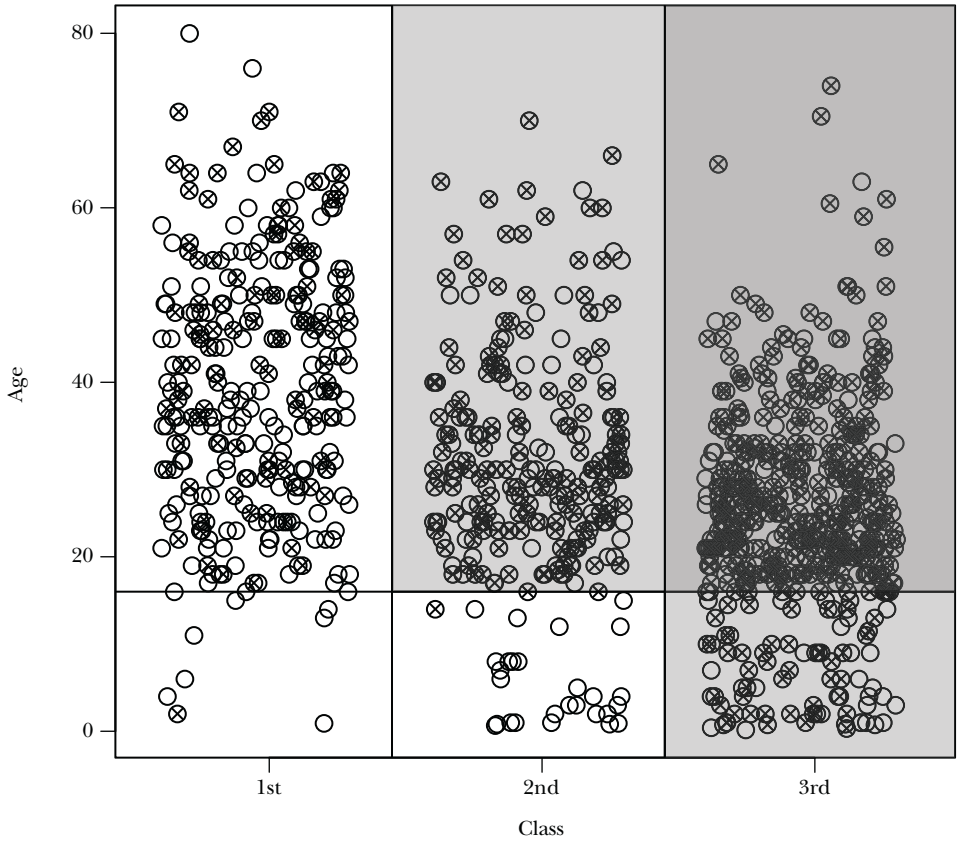
This classification can also be depicted in the "partition plot" (Figure 2), which shows how the tree divides up the space of age and class pairs into rectangular

[1] All data and code used in this paper can be found in the online Appendix available at http://e-jep.org.

Table 2
Tree Model in Rule Form

Features	*Predicted*	*Actual/Total*
Class 3	Died	370/501
Class 1–2, younger than 16	Lived	34/36
Class 2, older than 16	Died	145/233
Class 1, older than 16	Lived	174/276

Figure 2
The Simple Tree Model Predicts Death in Shaded Region
(empty circles indicate survival; circles with x *'s indicate death)*



regions. Of course, the partition plot can only be used for two variables, while a tree representation can handle an arbitrarily large number.

It turns out that there are computationally efficient ways to construct classification trees of this sort. These methods generally are restricted to binary trees (two branches

Table 3
Logistic Regression of Survival versus Age

Coefficient	*Estimate*	*Standard error*	*t value*	*p value*
Intercept	0.465	0.0350	13.291	0.000
Age	−0.002	0.001	−1.796	0.072

Note: Logistic regression relating survival (0 or 1) to age in years.

at each node). They can be used for classification with multiple outcomes ("classification trees") or with continuous dependent variables ("regression trees").

Trees tend to work well for problems where there are important nonlinearities and interactions. As an example, let us continue with the *Titanic* data and create a tree that relates survival to age. In this case, the rule generated by the tree is very simple: predict "survive" if age < 8.5 years. We can examine the same data with a logistic regression to estimate the probability of survival as a function of age, with results reported in Table 3.

The tree model suggests that age is an important predictor of survival, while the logistic model says it is barely important. This discrepancy is explained in Figure 3 where we plot survival rates by age bins. Here we see that survival rates for the youngest passengers were relatively high, and survival rates for older passengers were relatively low. For passengers between these two extremes, age didn't matter very much. So what mattered for survival is not so much age, but whether the passenger was a child or elderly. It would be difficult to discover this pattern from a logistic regression alone.[2]

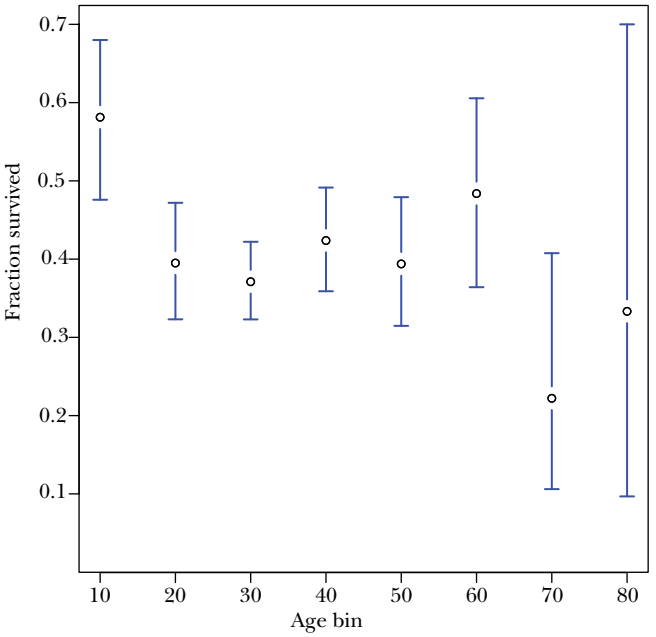
Trees also handle missing data well. Perlich, Provost, and Simonoff (2003) examined several standard datasets and found that "logistic regression is better for smaller data sets and tree induction for larger data sets." Interestingly enough, trees tend *not* to work very well if the underlying relationship really is linear, but there are hybrid models such as RuleFit (Friedman and Popescu 2005) that can incorporate both tree and linear relationships among variables. However, even if trees may not improve on predictive accuracy compared to linear models, the age example shows that they may reveal aspects of the data that are not apparent from a traditional linear modeling approach.

Pruning Trees

One problem with trees is that they tend to overfit the data. Just as a regression with n observations and n variables will give you a good fit in-sample, a tree with many branches will also fit the training data well. In either case, predictions using new data, such as the test set, could be very poor.

[2] It is true that if you *knew* that there was a nonlinearity in age, you could use age dummies in the logit model to capture this effect. However the tree formulation made this nonlinearity immediately apparent.

Figure 3
***Titanic* Survival Rates by Age Group**



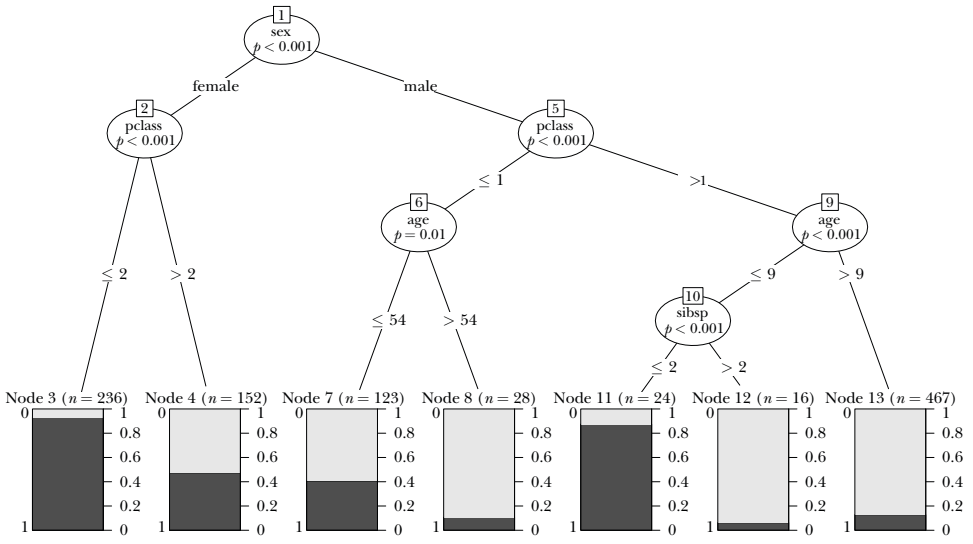
Notes: The figure shows the mean survival rates for different age groups along with confidence intervals. The age bin 10 means "10 and younger," the next age bin is "older than 10 through 20," and so on.

The most common solution to this problem is to "prune" the tree by imposing a cost for complexity. There are various measures of complexity, but a common one is the number of terminal nodes (also known as "leafs"). The cost of complexity is a tuning parameter that is chosen to provide the best out-of-sample predictions, which is typically measured using the 10-fold cross-validation procedure mentioned earlier.

A typical tree estimation session might involve dividing your data into ten folds, using nine of the folds to grow a tree with a particular complexity, and then predict on the excluded fold. Repeat the estimation with different values of the complexity parameter using other folds and choose the value of the complexity parameter that minimizes the out-of-sample classification error. (Some researchers recommend being a bit more aggressive and advocate choosing the complexity parameter that is one standard deviation lower than the loss-minimizing value.)

Of course, in practice, the computer program handles most of these details for you. In the examples in this paper, I mostly use default choices to keep things simple, but in practice these defaults will often be adjusted by the analyst. As with any other statistical procedure, skill, experience, and intuition are helpful in coming up with a good answer. Diagnostics, exploration, and experimentation are just as useful with these methods as with regression techniques.

Figure 4
A ctree for Survivors of the *Titanic*
(black bars indicate fraction of the group that survived)



Note: See text for interpretation.

There are many other approaches to creating trees, including some that are explicitly statistical in nature. For example, a "conditional inference tree," or ctree for short, chooses the structure of the tree using a sequence of hypothesis tests. The resulting trees tend to need very little pruning (Hothorn, Hornik, and Zeileis 2006). An example for the *Titanic* data is shown in Figure 4.

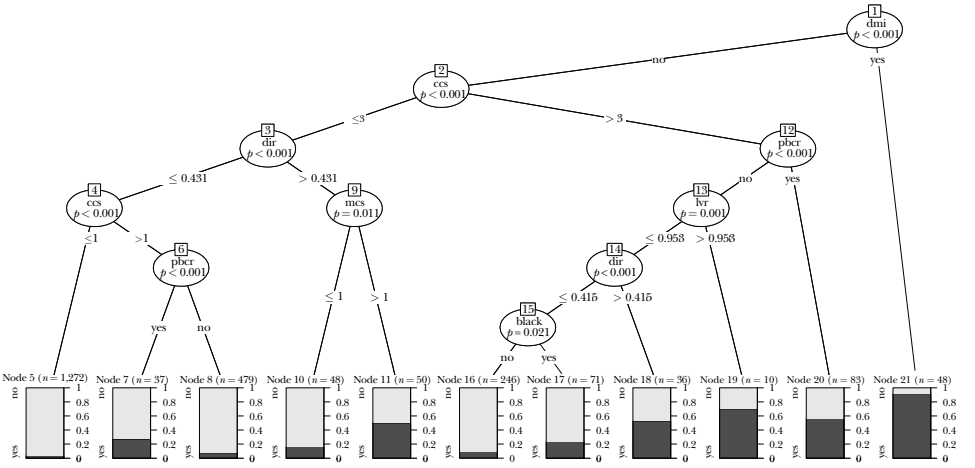
The first node divides by gender. The second node then divides by class. In the right-hand branches, the third node divides by age, and a fourth node divides by the number of siblings plus spouse aboard. The bins at the bottom of the figure show the total number of people in that leaf and a graphical depiction of their survival rate. One might summarize this tree by the following principle: "women and children first . . . particularly if they were traveling first class." This simple example again illustrates that classification trees can be helpful in summarizing relationships in data, as well as predicting outcomes.[3]

An Economic Example Using Home Mortgage Disclosure Act Data

Munnell, Tootell, Browne, and McEneaney (1996) examined mortgage lending in Boston to see if race played a significant role in determining who was approved for a mortgage. The primary econometric technique was a logistic regression where

[3] For two excellent tutorials on tree methods that use the *Titanic* data, see Stephens and Wehrley (2014).

Figure 5
Home Mortgage Disclosure Act (HMDA) Data Tree



Notes: Figure 5 shows a conditional tree estimated using the **R** package **party**. The black bars indicate the fraction of each group who were denied mortgages. The most important determinant of this is the variable "dmi," or "denied mortgage insurance." Other variables are: "dir," debt payments to total income ratio; "hir," housing expenses to income ratio; "lvr," ratio of size of loan to assessed value of property; "ccs," consumer credit score; "mcs," mortgage credit score; "pbcr," public bad credit record; "dmi," denied mortgage insurance; "self," self-employed; "single," applicant is single; "uria," 1989 Massachusetts unemployment rate applicant's industry; "condominium," unit is condominium; "black," race of applicant black; and "deny," mortgage application denied.

race was included as one of the predictors. The coefficient on race showed a statistically significant negative impact on probability of getting a mortgage for black applicants. This finding prompted considerable subsequent debate and discussion; see Ladd (1998) for an overview.

Here I examine this question using the tree-based estimators described in the previous section. The data consists of 2,380 observations of 12 predictors, one of which was race. Figure 5 shows a conditional tree estimated using the **R** package **party**.

The tree fits pretty well, misclassifying 228 of the 2,380 observations for an error rate of 9.6 percent. By comparison, a simple logistic regression does slightly better, misclassifying 225 of the 2,380 observations, leading to an error rate of 9.5 percent. As you can see in Figure 5, the most important variable is "dmi" = "denied mortgage insurance." This variable alone explains much of the variation in the data. The race variable ("black") shows up far down the tree and seems to be relatively unimportant.

One way to gauge whether a variable is important is to exclude it from the prediction and see what happens. When this is done, it turns out that the accuracy of the tree-based model doesn't change at all: exactly the same cases are misclassified. Of course, it is perfectly possible that there was racial discrimination

elsewhere in the mortgage process, or that some of the variables included are highly correlated with race. But it is noteworthy that the tree model produced by standard procedures that omits race fits the observed data just as well as a model that includes race.

Boosting, Bagging, Bootstrap

There are several useful ways to improve classifier performance. Interestingly enough, some of these methods work by *adding* randomness to the data. This seems paradoxical at first, but adding randomness turns out to be a helpful way of dealing with the overfitting problem.

Bootstrap involves choosing (with replacement) a sample of size n from a dataset of size n to estimate the sampling distribution of some statistic. A variation is the "m out of n bootstrap" which draws a sample of size m from a dataset of size $n > m$.

Bagging involves averaging across models estimated with several different bootstrap samples in order to improve the performance of an estimator.

Boosting involves repeated estimation where misclassified observations are given increasing weight in each repetition. The final estimate is then a vote or an average across the repeated estimates.[4]

Econometricians are well-acquainted with the bootstrap but rarely use the other two methods. Bagging is primarily useful for nonlinear models such as trees (Friedman and Hall 2007). Boosting tends to improve predictive performance of an estimator significantly and can be used for pretty much any kind of classifier or regression model, including logits, probits, trees, and so on.

It is also possible to combine these techniques and create a "forest" of trees that can often significantly improve on single-tree methods. Here is a rough description of how such "random forests" work.

Random Forests

Random forests is a technique that uses multiple trees. A typical procedure uses the following steps.

1. Choose a bootstrap sample of the observations and start to grow a tree.
2. At each node of the tree, choose a random sample of the predictors to make the next decision. Do not prune the trees.
3. Repeat this process many times to grow a forest of trees.
4. In order to determine the classification of a new observation, have each tree make a classification and use a majority vote for the final prediction.

This method produces surprisingly good out-of-sample fits, particularly with highly nonlinear data. In fact, Howard and Bowles (2012) claim "ensembles of decision trees (often known as 'Random Forests') have been the most successful general-purpose algorithm in modern times." They go on to indicate that

[4] Boosting is often used with decision trees, where it can dramatically improve their predictive performance.

"the algorithm is very simple to understand, and is fast and easy to apply." See also Caruana and Niculescu-Mitzil (2006) who compare several different machine learning algorithms and find that ensembles of trees perform quite well. There are a number of variations and extensions of the basic "ensemble of trees" model such as Friedman's "Stochastic Gradient Boosting" (Friedman 2002).

One defect of random forests is that they are a bit of a black box—they don't offer simple summaries of relationships in the data. As we have seen earlier, a single tree can offer some insight about how predictors interact. But a forest of a thousand trees cannot be easily interpreted. However, random forests can determine which variables are "important" in predictions in the sense of contributing the biggest improvements in prediction accuracy.

Note that random forests involves quite a bit of randomization; if you want to try them out on some data, I strongly suggest choosing a particular seed for the random number generator so that your results can be reproduced. (See the online supplement for examples.)

I ran the random forest method on the HMDA data and found that it misclassified 223 of the 2,380 cases, a small improvement over the logit and the ctree. I also used the importance option in random forests to see how the predictors compared. It turned out that "dmi" was the most important predictor and race was second from the bottom, which is consistent with the ctree analysis.

Variable Selection

Let us return to the familiar world of linear regression and consider the problem of variable selection. There are many such methods available, including stepwise regression, principal component regression, partial least squares, Akaike information criterion (AIC) and Bayesian information criterion (BIC) complexity measures, and so on. Castle, Qin, and Reed (2009) describe and compare 21 different methods.

LASSO and Friends

Here we consider a class of estimators that involves penalized regression. Consider a standard multivariate regression model where we predict y_t as a linear function of a constant, b_0, and P predictor variables. We suppose that we have standardized all the (nonconstant) predictors so they have mean zero and variance one.

Consider choosing the coefficients $(b_1, \ldots, b_P)$ for these predictor variables by minimizing the sum of squared residuals plus a penalty term of the form

$$\lambda \sum_{p=1}^{P} \left[(1 - \alpha) \, | b_p | + \alpha \, | b_p |^2 \right].$$

This estimation method is called *elastic net regression*; it contains three other methods as special cases. If there is no penalty term ($\lambda = 0$), this is *ordinary least squares*. If $\alpha = 1$, so that there is only the quadratic constraint, this is *ridge regression*.

If $\alpha = 0$, this is called the *LASSO*, an acronym for "least absolute shrinkage and selection operator."

These penalized regressions are classic examples of regularization. In this case, the complexity is the number and size of predictors in the model. All of these methods tend to shrink the least squares regression coefficients towards zero. The LASSO and elastic net typically produces regressions where some of the variables are set to be exactly zero. Hence this is a relatively straightforward way to do variable selection.

It turns out that these estimators can be computed quite efficiently, so doing variable selection on reasonably large problems is computationally feasible. They also seem to provide good predictions in practice.

Spike-and-Slab Regression

Another approach to variable selection that is novel to most economists is spike-and-slab regression, a Bayesian technique. Suppose that you have P possible predictors in some linear model. Let γ be a vector of length P composed of zeros and ones that indicate whether or not a particular variable is included in the regression.

We start with a Bernoulli prior distribution on γ; for example, initially we might think that all variables have an equally likely chance of being in the regression. Conditional on a variable being in the regression, we specify a prior distribution for the regression coefficient associated with that variable. For example, we might use a Normal prior with mean 0 and a large variance. These two priors are the source of the method's name: the "spike'" is the probability of a coefficient being nonzero; the "slab" is the (diffuse) prior describing the values that the coefficient can take on.

Now we take a draw of γ from its prior distribution, which will just be a list of variables in the regression. Conditional on this list of included variables, we take a draw from the prior distribution for the coefficients. We combine these two draws with the likelihood in the usual way, which gives us a draw from posterior distribution on both probability of inclusion and the coefficients. We repeat this process thousands of times using a Markov Chain Monte Carlo (MCMC) technique which gives us a table summarizing the posterior distribution for γ (indicating variable inclusion), β (the coefficients), and the associated prediction of y. We can summarize this table in a variety of ways. For example, we can compute the average value of γ_p which shows the posterior probability that the variable p is included in the regressions.

An Economic Example: Growth Regressions

We illustrate these different methods of variable selection using data from Sala-i-Martín (1997). This exercise involved examining a dataset of 72 counties and 42 variables in order to see which variables appeared to be important predictors of economic growth. Sala-i-Martín (1997) computed at all possible subsets of regressors of manageable size and used the results to construct an importance measure he called CDF(0). Ley and Steel (2009) investigated the same question using Bayesian

Table 4

Comparing Variable Selection Algorithms: Which Variables Appeared as Important Predictors of Economic Growth?

Predictor	*Bayesian model averaging*	*CDF(0)*	*LASSO*	*Spike-and-Slab*
GDP level 1960	1.000	1.000	-	0.9992
Fraction Confucian	0.995	1.000	2	0.9730
Life expectancy	0.946	0.942	-	0.9610
Equipment investment	0.757	0.997	1	0.9532
Sub-Saharan dummy	0.656	1.000	7	0.5834
Fraction Muslim	0.656	1.000	8	0.6590
Rule of law	0.516	1.000	-	0.4532
Open economy	0.502	1.000	6	0.5736
Degree of capitalism	0.471	0.987	9	0.4230
Fraction Protestant	0.461	0.966	5	0.3798

Source: The table is based on that in Ley and Steel (2009); the data analyzed is from Sala-i-Martín (1997).
Notes: We illustrate different methods of variable selection. This exercise involved examining a dataset of 72 counties and 42 variables in order to see which variables appeared to be important predictors of economic growth. The table shows ten predictors that were chosen by Sala-i-Martín (1997) using a CDF(0) measure defined in the 1997 paper; Ley and Steel (2009) using Bayesian model averaging, LASSO, and spike-and-slab regressions. Metrics used are not strictly comparable across the various models. The "Bayesian model averaging" and "Spike-and-Slab" columns are posterior probabilities of inclusion; the "LASSO" column just shows the ordinal importance of the variable or a dash indicating that it was not included in the chosen model; and the CDF(0) measure is defined in Sala-i-Martín (1997).

model averaging, a technique related to, but not identical with, spike-and-slab. Hendry and Krolzig (2004) examined an iterative significance test selection method.

Table 4 shows ten predictors that were chosen by Sala-i-Martín (1997) using his two million regressions, Ley and Steel (2009) using Bayesian model averaging, LASSO, and spike-and-slab. The table is based on that in Ley and Steel (2009) but metrics used are not strictly comparable across the various models. The "Bayesian model averaging" and "spike-slab" columns show posterior probabilities of inclusion; the "LASSO" column just shows the ordinal importance of the variable or a dash indicating that it was not included in the chosen model; and the CDF(0) measure is defined in Sala-i-Martín (1997).

The LASSO and the Bayesian techniques are very computationally efficient and would likely be preferred to exhaustive search. All four of these variable selection methods give similar results for the first four or five variables, after which they diverge. In this particular case, the dataset appears to be too small to resolve the question of what is "important" for economic growth.

Variable Selection in Time Series Applications

The machine learning techniques described up until now are generally applied to cross-sectional data where independently distributed data is a plausible assumption. However, there are also techniques that work with time series. Here we

describe an estimation method that we call Bayesian Structural Time Series (BSTS) that seems to work well for variable selection problems in time series applications.

Our research in this area was motivated by Google Trends data, which provides an index of the volume of Google queries on specific terms. One might expect that queries on "file for unemployment" might be predictive of the actual rate of filings for initial claims, or that queries on "Orlando vacation" might be predictive of actual visits to Orlando. Indeed, in Choi and Varian (2009, 2012), Goel, Hofman, Lahaie, Pennock, and Watts (2010), Carrière-Swallow and Labbé (2011), McLaren and Shanbhoge (2011), Artola and Galan (2012), Hellerstein and Middeldorp (2012), and other papers, many researchers have shown that Google queries do have significant short-term predictive power for various economic metrics.

The challenge is that there are billions of queries so it is hard to determine exactly which queries are the most predictive for a particular purpose. Google Trends classifies the queries into categories, which helps a little, but even then we have hundreds of categories as possible predictors so that overfitting and spurious correlation are a serious concern. Bayesian Structural Time Series is designed to address these issues. We offer a very brief description here; more details are available in Scott and Varian (2013a, 2013b).

Consider a classic time series model with *constant* level, linear time trend, and regressor components:

$$yt = \mu + bt + \beta xt + et.$$

The "local linear trend" is a stochastic generalization of this model where the level and time trend can vary through time.

$$\text{Observation: } y_t = \mu_t + z_t + e_{1t} = \text{level} + \text{regression}$$

$$\text{State variable 1: } \mu_t = \mu_{t-1} + b_{t-1} + e_{2t} = \text{random walk} + \text{trend}$$

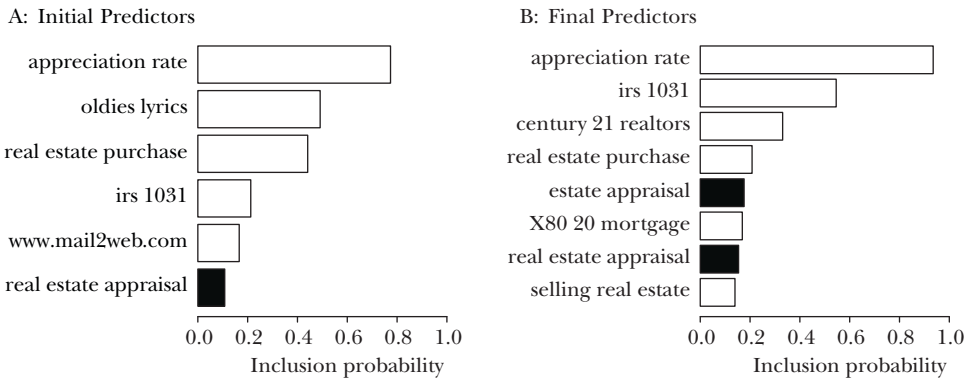
$$\text{State variable 2: } z_t = \beta x_t = \text{regression}$$

$$\text{State variable 3: } b_t = b_{t-1} + e_{3t} = \text{random walk for trend}$$

It is easy to add an additional state variable for seasonality if that is appropriate. The parameters to estimate are the regression coefficients β and the variances of (e_{it}) for $i = 1, \ldots 3$. We can then use these estimates to construct the optimal forecast based on techniques drawn from the literature on Kalman filters.

For the regression, we use the spike-and-slab variable choice mechanism described above. A draw from the posterior distribution now involves a draw of variances of (e_{1t}, e_{2t}, e_{3t}) a draw of the vector γ that indicates which variables are in the regression, and a draw of the regression coefficients β for the included variables. The draws of μ_t, b_t, and β can be used to construct estimates of y_t and forecasts for y_{t+1}. We end up with an (estimated) posterior distribution for each parameter of

Figure 6
An Example Using Bayesian Structural Time Series (BSTS)
(finding Google queries that are predictors of new home sales)



Source: Author using HSN1FNSA data from the St. Louis Federal Reserve Economic Data.
Notes: Consider the nonseasonally adjusted data for new homes sold in the United States, which is (HSN1FNSA) from the St. Louis Federal Reserve Economic Data. This time series can be submitted to Google Correlate, which then returns the 100 queries that are the most highly correlated with the series. We feed that data into the BSTS system, which identifies the predictors with the largest posterior probabilities of appearing in the housing regression; these are shown in Figure 6A. In these figures, black bars indicate a negative relationship, and white bars indicate a positive relationship. Two predictors, "oldies lyrics" and "www.mail2web" appear to be spurious so we remove them and re-estimate, yielding the results in Figure 6B.

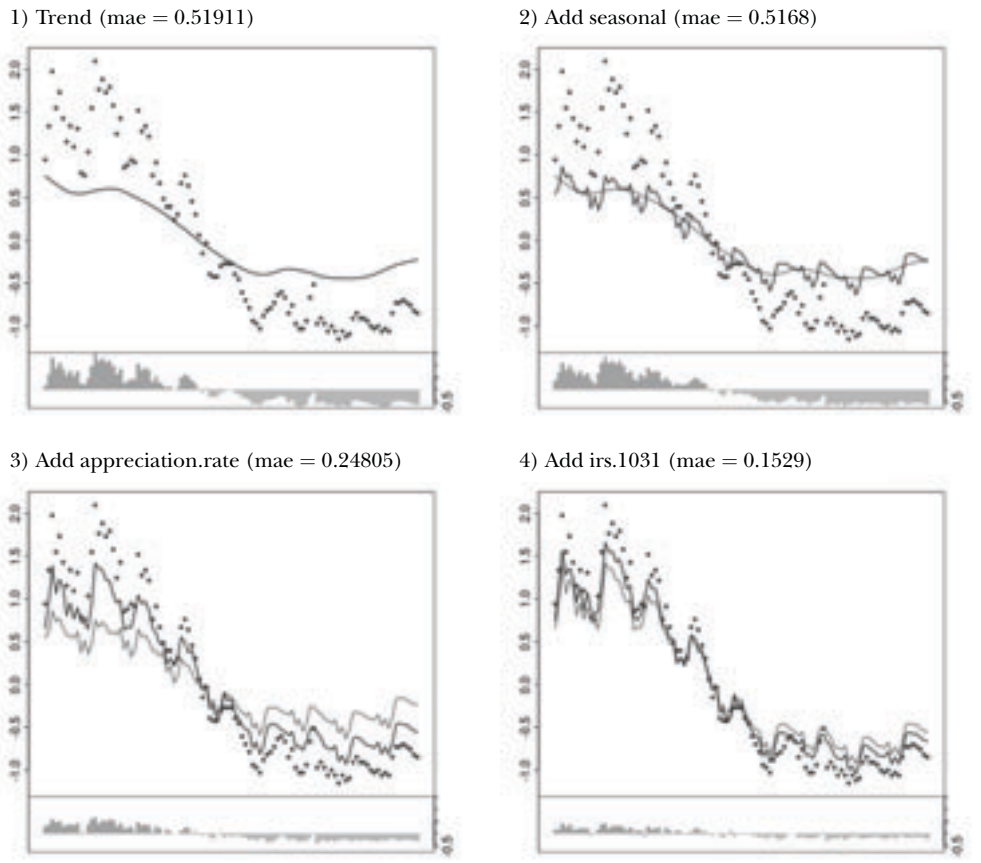
interest. If we seek a point prediction, we can average over these draws, which is essentially a form of Bayesian model averaging.

As an example, consider the nonseasonally adjusted data for new homes sold in the United States, which is (HSN1FNSA) from the St. Louis Federal Reserve Economic Data. This time series can be submitted to Google Correlate, which then returns the 100 queries that are the most highly correlated with the series. We feed that data into the BSTS system, which identifies the predictors with the largest posterior probabilities of appearing in the housing regression; these are shown in Figure 6A.In these figures, black bars indicate a negative relationship and white bars indicate a positive relationship. Two predictors, "oldies lyrics" and "www.mail2web" appear to be spurious so we remove them and re-estimate, yielding the results in Figure 6B.

The fit is shown in Figure 7, which shows the incremental contribution of the trend, seasonal, and two of the regressors. Even with only two predictors, queries on "appreciation rate" and queries on "irs 1031," we get a pretty good fit.[5]

[5] IRS section 1031 has to do with deferring capital gains on certain sorts of property exchange.

Figure 7
Fit for the Housing Regression: Incremental Contribution of Trend, Seasonal, and Two Regressors



Source: Author using (HSN1FNSA) data from the St. Louis Federal Reserve.
Notes: The plots show the impact of the trend, seasonal, and a few individual regressors. Data has been standardized to have mean zero and variance 1. The residuals are shown on the bottom. The abbreviation "mae" stands for "mean absolute error."

Econometrics and Machine Learning

There are a number of areas where there would be opportunities for fruitful collaboration between econometrics and machine learning. I mentioned above that most machine learning uses independent and identically distributed data. However, the Bayesian Structural Time Series model shows that some of these techniques can be adopted for time series models. It is also possible to use machine learning techniques to look at panel data, and there has been some work in this direction.

However, the most important area for collaboration involves causal inference. Econometricians have developed several tools for causal inference such as

instrumental variables, regression discontinuity, difference-in-differences, and various forms of natural and designed experiments (Angrist and Krueger 2001). Machine learning work has, for the most part, dealt with pure prediction. In a way, this is ironic, since theoretical computer scientists, such as Pearl (2009a, b) have made significant contributions to causal modeling. However, it appears that these theoretical advances have not as yet been incorporated into machine learning practice to a significant degree.

Causality and Prediction

As economists know well, there is a big difference between correlation and causation. A classic example: there are often more police in precincts with high crime, but that does not imply that increasing the number of police in a precinct would increase crime.

The machine learning models we have described so far have been entirely about prediction. If our data were generated by policymakers who assigned police to areas with high crime, then the observed relationship between police and crime rates could be highly predictive for the *historical* data but not useful in predicting the causal impact of explicitly *assigning* additional police to a precinct.

To enlarge on this point, let us consider an experiment (natural or designed) that attempts to estimate the impact of some policy, such as adding police to precincts. There are two critical questions.

1) How will police be assigned to precincts in both the experiment and the policy implementation? Possible assignment rules could be 1) random, 2) based on perceived need, 3) based on cost of providing service, 4) based on resident requests, 5) based on a formula or set of rules, 6) based on asking for volunteers, and so on. Ideally the assignment procedure in the experiment will be similar to that used in the policy. Developing accurate predictions about which precincts will receive additional police under the proposed policy based on the experimental data can clearly be helpful in predicting the expected impact of the policy.

2) What will be the impact of these additional police in both the experiment and the policy? As Rubin (1974) and many subsequent authors have emphasized, when we want to estimate the *causal* impact of some treatment we need to compare the outcome with the intervention to what *would have happened* without the intervention. But this counterfactual cannot be observed, so it must be predicted by some model. The better predictive model you have for the counterfactual, the better you will be able to estimate the causal effect, a rule that is true for both pure experiments and natural experiments.

So even though a predictive model will not necessarily allow one to conclude anything about causality by itself, such models may help in estimating the causal impact of an intervention when it occurs.

To state this in a slightly more formal way, consider the identity from Angrist and Pischke (2009, p. 11):

$$\begin{aligned}\text{observed difference in outcome} &= \text{average treatment effect on the treated} \\ &\quad + \text{selection bias.}\end{aligned}$$

If you want to model the average treatment effect as a function of other variables, you will usually need to model both the observed difference in outcome and the selection bias. The better your predictive model for those components, the better your estimate of the average treatment effect will be. Of course, if you have a true randomized treatment–control experiment, selection bias goes away and those treated are an unbiased random sample of the population.

To illustrate these points, let us consider the thorny problem of estimating the causal effect of advertising on sales (Lewis and Rao 2013). The difficulty is that there are many confounding variables, such as seasonality or weather, that cause both increased ad exposures and increased purchases by consumers. For example, consider the (probably apocryphal) story about an advertising manager who was asked why he thought his ads were effective. "Look at this chart," he said. "Every December I increase my ad spend and, sure enough, purchases go up." Of course, in this case, seasonality can be included in the model. However, generally there will be other confounding variables that affect both exposure to ads and the propensity of purchase, which make causal interpretations of observed relationships problematic.

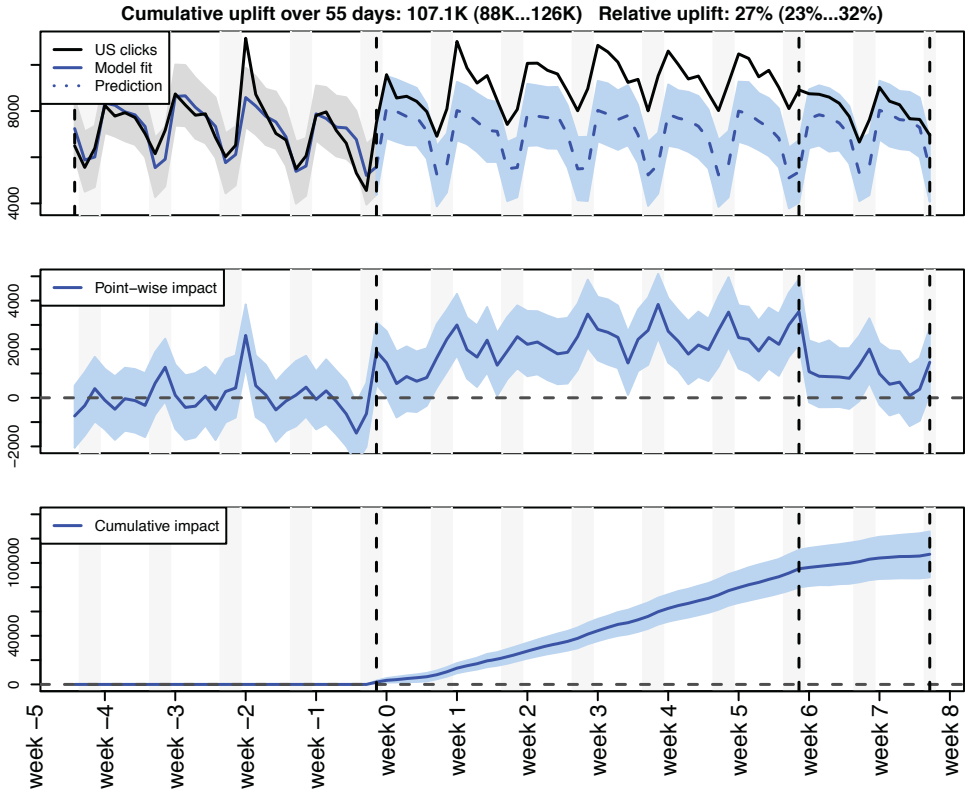
The ideal way to estimate advertising effectiveness is, of course, to run a controlled experiment. In this case the control group provides an estimate of the counterfactual: what would have happened without ad exposures. But this ideal approach can be quite expensive, so it is worth looking for alternative ways to predict the counterfactual. One way to do this is to use the Bayesian Structural Time Series (BSTS) method described earlier.

Suppose a given company wants to determine the impact of an advertising campaign on visits to its website. It first uses BSTS (or some other technique) to build a model predicting the time series of visits as a function of its past history, seasonal effects, and other possible predictors such as Google queries on its company name, its competitors' names, or products that it produces. Since there are many possible choices for predictors, it is important to use some variable selection mechanism such as those described earlier.

It next runs an ad campaign for a few weeks and records visits during this period. Finally, it makes a forecast of what visits *would have been* in the absence of the ad campaign using the model developed in the first stage. Comparing the actual visits to the counterfactual visits gives us an estimate of the causal effect of advertising.

Figure 8, shows the outcome of such a procedure. It is based on the approach proposed in Brodersen, Gallusser, Koehler, Remy, and Scott (2013), but the covariates are chosen automatically from Google Trends categories using Bayesian Structural Time Series (BSTS). Panel A shows the actual visits and the prediction

Figure 8
Actual and Predicted Website Visits



Source: This example is based on the approach proposed in Brodersen, Gallusser, Koehler, Remy, and Scott (2013), but the covariates are chosen automatically from Google Trends categories using Bayesian Structural Time Series (BSTS).
Notes: Suppose a given company wants to determine the impact of an advertising campaign on its website visits. Panel A shows the actual visits and the prediction of what the visits would have been without the campaign based on the BSTS forecasting model. Panel B shows the difference between actual and predicted visits, and Panel C shows the cumulative difference.

of what the visits would have been without the campaign based on the BSTS forecasting model. Panel B shows the difference between actual and predicted visits, and Panel C shows the cumulative difference. It is clear from this figure that there was a significant causal impact of advertising, which can then be compared to the cost of the advertising to evaluate the campaign.

This procedure does not use a control group in the conventional sense. Rather it uses a general time series model based on trend extrapolation, seasonal effects, and relevant covariates to forecast what would have happened without the ad campaign.

A good predictive model can be better than a randomly chosen control group, which is usually thought to be the gold standard. To see this, suppose that you run

an ad campaign in 100 cities and retain 100 cities as a control. After the experiment is over, you discover the weather was dramatically different across the cities in the study. Should you add weather as a predictor of the counterfactual? Of course! If weather affects sales (which it does), then you will get a more accurate prediction of the counterfactual and thus a better estimate of the causal effect of advertising.

Model Uncertainty

An important insight from machine learning is that averaging over many small models tends to give better out-of-sample prediction than choosing a single model.

In 2006, Netflix offered a million dollar prize to researchers who could provide the largest improvement to their existing movie recommendation system. The winning submission involved a "complex blending of no fewer than 800 models," though they also point out that "predictions of good quality can usually be obtained by combining a small number of judiciously chosen methods" (Feuerverger, He, and Khatri 2012). It also turned out that a blend of the best- and second-best submissions outperformed either of them.

Ironically, it was recognized many years ago that averages of macroeconomic model forecasts outperformed individual models, but somehow this idea was rarely exploited in traditional econometrics. The exception is the literature on Bayesian model averaging, which has seen a steady flow of work; see Steel (2011) for a survey.

However, I think that model uncertainty has crept into applied econometrics through the back door. Many papers in applied econometrics present regression results in a table with several different specifications: which variables are included in the controls, which variables are used as instruments, and so on. The goal is usually to show that the estimate of some interesting parameter is not very sensitive to the exact specification used.

One way to think about it is that these tables illustrate a simple form of model uncertainty: how an estimated parameter varies as different models are used. In these papers, the authors tend to examine only a few representative specifications, but there is no reason why they couldn't examine many more if the data were available.

In this period of "big data," it seems strange to focus on *sampling uncertainty,* which tends to be small with large datasets, while completely ignoring *model uncertainty,* which may be quite large. One way to address this is to be explicit about examining how parameter estimates vary with respect to choices of control variables and instruments.

Summary and Further Reading

Since computers are now involved in many economic transactions, big data will only get bigger. Data manipulation tools and techniques developed for small datasets will become increasingly inadequate to deal with new problems. Researchers in machine learning have developed ways to deal with large datasets and economists

interested in dealing with such data would be well advised to invest in learning these techniques.

I have already mentioned Hastie, Tibshirani, and Friedman (2009), who provide detailed descriptions of all the methods discussed here but at a relatively advanced level. James, Witten, Hastie, and Tibshirani (2013) describe many of the same topics at an undergraduate-level, along with **R** code and many examples. (There are several economic examples in the book where the tension between predictive modeling and causal inference is apparent.) Murphy (2012) examines machine learning from a Bayesian point of view.

Venables and Ripley (2002) offer good discussions of these topics with emphasis on applied examples. Leek (2013) presents a number of YouTube videos with gentle and accessible introductions to several tools of data analysis. Howe (2013) provides a somewhat more advanced introduction to data science that also includes discussions of SQL and NoSQL databases. Wu and Kumar (2009) give detailed descriptions and examples of the major algorithms in data mining, while Williams (2011) provides a unified toolkit. Domingos (2012) summarizes some important lessons including "pitfalls to avoid, important issues to focus on and answers to common questions."

■ *Thanks to Jeffrey Oldham, Tom Zhang, Rob On, Pierre Grinspan, Jerry Friedman, Art Owen, Steve Scott, Bo Cowgill, Brock Noland, Daniel Stonehill, Robert Snedegar, Gary King, Fabien Curto-Millet, and the editors of this journal for helpful comments on earlier versions of this paper. The author works for Google, and Google had the right to review this paper before publication.*

References

▶**Angrist, Joshua D., and Alan B. Krueger.** 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 5(4): 69–85.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Artola, Concha, and Enrique Galan. 2012. "Tracking the Future on the Web: Construction of Leading Indicators Using Internet Searches." Documentos Ocasionales 1203T, Bank of Spain. http://www.bde.es/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf.

Breiman, Leo, Jerome H. Friedman, R. A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey.

Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2013. "Inferring Causal Impact Using Bayesian Structural Time-Series Models." http://research.google.com/pubs/pub41854.html.

▶**Carrière-Swallow, Yan, and Felipe Labbé.** 2011. "Nowcasting with Google Trends in an Emerging Market." *Journal of Forecasting* 32(4): 289–98.

Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. "An Empirical Comparison of Supervised Learning Algorithms." In *Proceedings of the 23rd*

International Conference on Machine Learning, Pittsburgh, PA. Available at: http://www.autonlab.org/icml2006/technical/accepted.html.

Castle, Jennifer L., Xiaochuan Qin, and W. Robert Reed. 2009. "How to Pick the Best Regression Equation: A Review and Comparison of Model Selection Algorithms." Working Paper 13/2009, Department of Economics, University of Canterbury. http://www.econ.canterbury.ac.nz/RePEc/cbt/econwp/0913.pdf.

Choi, Hyunyoung, and Hal Varian. 2009. "Predicting the Present with Google Trends." http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Choi, Hyunyoung, and Hal Varian. 2012. "Predicting the Present with Google Trends." *Economic Record* 88(1): 2–9.

Domingos, Pedro. 2012. "A Few Useful Things to Know about Machine Learning." *Communications of the ACM* 55(10): 78–87.

Einav, Liran, and Jonathan D. Levin. 2013. "The Data Revolution and Economic Analysis." Technical report, NBER Innovation Policy and the Economy Conference, 2013. NBER Working Paper 19035.

Feuerverger, Andrey, Yu He, and Shashi Khatri. 2012. "Statistical Significance of the Netflix Challenge." *Statistical Science* 27(2): 202–231.

Friedman, Jerome. 2002. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis* 38(4): 367–78.

Friedman, Jerome, and Peter Hall. 2007. "On Bagging and Nonlinear Estimation." *Journal of Statistical Planning and Inference* 137(3): 669–83.

Friedman, Jerome H., and Bogdan E. Popescu. 2005. "Predictive Learning via Rule Ensembles." Technical report, Stanford University. http://www-stat.stanford.edu/~jhf/ftp/RuleFit.pdf

Goel, Sharad, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. 2010. "Predicting Consumer Behavior with Web Search." *PNAS* 107(41).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer-Verlag.

Hellerstein, Rebecca, and Menno Middeldorp. 2012. "Forecasting with Internet Search Data." *Liberty Street Economics* Blog of the Federal Reserve Bank of New York, January 4. http://libertystreeteconomics.newyorkfed.org/2012/01/forecasting-with-internet-search-data.html.

Hendry, David F., and Hans-Martin Krolzig. 2004. "We Ran One Regression." *Oxford Bulletin of Economics and Statistics* 66(5): 799–810.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15(3): 651–74.

Howard, Jeremy, and Mike Bowles. 2012. "The Two Most Important Algorithms in Predictive Modeling Today." Strata Conference presentation, February 28. http://strataconf.com/strata2012/public/schedule/detail/22658.

Howe, Bill. 2013. Introduction to Data Science. A course from the University of Washington. https://class.coursera.org/datasci-001/lecture/index.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Ladd, Helen F. 1998. "Evidence on Discrimination in Mortgage Lending." *Journal of Economic Perspectives* 12(2): 41–62.

Leek, Jeff. 2013. Data Analysis. Videos from the course. http://blog.revolutionanalytics.com/2013/04/coursera-data-analysis-course-videos.html.

Lewis, Randall A., and Justin M. Rao. 2013. "On the Near Impossibility of Measuring the Returns to Advertising." Unpublished paper, Google, Inc. and Microsoft Research. http://justinmrao.com/lewis_rao_nearimpossibility.pdf.

Ley, Eduardo, and Mark F. J. Steel. 2009. "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression." *Journal of Applied Econometrics* 24(4): 651–74.

McLaren, Nick, and Rachana Shanbhoge. 2011. "Using Internet Search Data as Economic Indicators." *Bank of England Quarterly Bulletin* 51(2): 134–40.

Morgan, James N., and John A. Sonquist. 1963. "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association* 58(302): 415–34.

Munnell, Alicia H., Geoffrey M. B. Tootell, Lynne E. Browne, and James McEneaney. 1996. "Mortgage Lending in Boston: Interpreting HMDA Data." *American Economic Review* 86(1): 25–53.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Pearl, Judea. 2009a. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.

Pearl, Judea. 2009b. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3: 96–146.

Perlich, Claudia, Foster Provost, and Jeffrey S. Simonoff. 2003. "Tree Induction vs. Logistic Regression: A Learning-Curve Analysis." *Journal of Machine Learning Research* 4: 211–55.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5): 689–701.

Sala-i-Martín, Xavier. 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87(2): 178–83.

Scott, Steve, and Hal Varian. 2013a. "Bayesian Variable Selection for Nowcasting Economic Time Series." NBER Working Paper 19567.

Scott, Steve, and Hal Varian. 2013b. "Predicting the Present with Bayesian Structural Time Series." NBER Working Paper 19567.

Steel, Mark F. J. 2011. "Bayesian Model Averaging and Forecasting." *Bulletin of E.U. and U.S. Inflation and Macroeconomic Analysis*, 200: 30–41.

Stephens, Revor, and Curt Wehrley. 2014. "Getting Started with R." *Kaggle*, September 28. https://www.kaggle.com/c/titanic-gettingStarted/details/new-getting-started-with-r.

Sullivan, Danny. 2012. "Google: 100 Billion Searches per Month, Search to Integrate Gmail, Launching Enhanced Search App for iOS." Search Engine Land. http://searchengineland.com/google-search-press-129925.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*, 4th edition. New York: Springer.

Williams, Graham. 2011. *Data Mining with Rattle and R*. New York: Springer.

Wu, Xindong, and Vipin Kumar, eds. 2009. *The Top Ten Algorithms in Data Mining*. CRC Press.

This article has been cited by:

1. Lulin Xu, Zhongwu Li. 2020. A New Appraisal Model of Second-Hand Housing Prices in China's First-Tier Cities Based on Machine Learning Algorithms. *Computational Economics* **71**. . [Crossref]
2. Hailong Cui, Sampath Rajagopalan, Amy R. Ward. 2020. Predicting product return volume using machine learning methods. *European Journal of Operational Research* **281**:3, 612-627. [Crossref]
3. Chinmay Kakatkar, Volker Bilgram, Johann Füller. 2020. Innovation analytics: Leveraging artificial intelligence in the innovation process. *Business Horizons* **63**:2, 171-181. [Crossref]
4. Kenneth David Strang. 2020. Problems with research methods in medical device big data analytics. *International Journal of Data Science and Analytics* **9**:2, 229-240. [Crossref]
5. SeyedSoroosh Azizi, Kiana Yektansani. 2020. Artificial Intelligence and Predicting Illegal Immigration to the USA. *International Migration* . [Crossref]
6. Silvia Emili, Attilio Gardini, Enrico Foscolo. 2020. High spatial and temporal detail in timely prediction of tourism demand. *International Journal of Tourism Research* **29**. . [Crossref]
7. David Easley, Eleonora Patacchini, Christopher Rojas. 2020. Multidimensional diffusion processes in dynamic online networks. *PLOS ONE* **15**:2, e0228421. [Crossref]
8. Carlos Poza, Manuel Monge. 2020. A real time leading economic indicator based on text mining for the Spanish economy. Fractional cointegration VAR and Continuous Wavelet Transform analysis. *International Economics* . [Crossref]
9. David Lenz, Peter Winker. 2020. Measuring the diffusion of innovations with paragraph vector topic models. *PLOS ONE* **15**:1, e0226685. [Crossref]
10. Abigail Devereaux, Linan Peng. 2020. Give us a little social credit: to design or to discover personal ratings in the era of Big Data. *Journal of Institutional Economics* **66**, 1-19. [Crossref]
11. Gang Xie, Xin Li, Yatong Qian, Shouyang Wang. 2020. Forecasting tourism demand with KPCA-based web search indexes. *Tourism Economics* 135481661989857. [Crossref]
12. Marios Poulos, Nikolaos Korfiatis, Sozon Papavlassopoulos. 2020. Assessing stationarity in web analytics: A study of bounce rates. *Expert Systems* **40**. . [Crossref]
13. Kong Lu. Computer Performance Determination System Based on Big Data Distributed File 877-884. [Crossref]
14. Giorgio Gnecco, Federico Nutarelli. On the Trade-Off Between Number of Examples and Precision of Supervision in Regression 1-6. [Crossref]
15. W. O. K. I. S. Wijesinghe, C. U. Kumarasinghe, J. Mannapperuma, K. L. D. U. Liyanage. Socioeconomic Status Classification of Geographic Regions in Sri Lanka Through Anonymized Call Detail Records 299-311. [Crossref]
16. Federico Bassetti, Roberto Casarin, Francesco Ravazzolo. Density Forecasting 465-494. [Crossref]
17. Marian Socoliuc, Cristina-Gabriela Cosmulese, Marius-Sorin Ciubotariu, Svetlana Mihaila, Iulia-Diana Arion, Veronica Grosu. 2020. Sustainability Reporting as a Mixture of CSR and Sustainable Development. A Model for Micro-Enterprises within the Romanian Forestry Sector. *Sustainability* **12**:2, 603. [Crossref]
18. Marcus H. Böhme, André Gröger, Tobias Stöhr. 2020. Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics* **142**, 102347. [Crossref]
19. Vikram Dayal. Introduction 3-8. [Crossref]
20. Vikram Dayal. From Trees to Random Forests 315-326. [Crossref]
21. Laurie A. Schintler. Regional Policy Analysis in the Era of Spatial Big Data 93-109. [Crossref]

22. Thiago Christiano Silva, Benjamin Miranda Tabak, Idamar Magalhães Ferreira. 2019. Modeling Investor Behavior Using Machine Learning: Mean-Reversion and Momentum Trading Strategies. *Complexity* **2019**, 1-14. [Crossref]
23. Jiafu An, Raghavendra Rau. 2019. Finance, technology and disruption. *The European Journal of Finance* **12**, 1-12. [Crossref]
24. Huy Duc Dang, Au Hai Thi Dam, Thuyen Thi Pham, Tra My Thi Nguyen. 2019. Determinants of credit demand of farmers in Lam Dong, Vietnam. *Agricultural Finance Review* **ahead-of-print**:ahead-of-print. . [Crossref]
25. Ranjith Vijayakumar, Mike W.-L. Cheung. 2019. Assessing Replicability of Machine Learning Results: An Introduction to Methods on Predictive Accuracy in Social Sciences. *Social Science Computer Review* **13**, 089443931988844. [Crossref]
26. Celso Martínez Musiño. 2019. Big Data-Análisis informétrico de documentos indexados en Scopus y Web of Science. *Investigación Bibliotecológica: archivonomía, bibliotecología e información* **34**:82, 87. [Crossref]
27. Vivek Anand Asokan, Masaru Yarime, Motoharu Onuki. 2019. A review of data-intensive approaches for sustainability: methodology, epistemology, normativity, and ontology. *Sustainability Science* **26**. . [Crossref]
28. Kurt Stockinger, Nils Bundi, Jonas Heitz, Wolfgang Breymann. 2019. Scalable architecture for Big Data financial analytics: user-defined functions vs. SQL. *Journal of Big Data* **6**:1. . [Crossref]
29. Shengying Zhai, Qihui Chen, Wenxin Wang. 2019. What Drives Green Fodder Supply in China?—A Nerlovian Analysis with LASSO Variable Selection. *Sustainability* **11**:23, 6692. [Crossref]
30. Arthur Lewbel. 2019. The Identification Zoo: Meanings of Identification in Econometrics. *Journal of Economic Literature* **57**:4, 835-903. [Abstract] [View PDF article] [PDF with links]
31. Juan D. Montoro-Pons, Manuel Cuadrado-García. 2019. Music festivals as mediators and their influence on consumer awareness. *Poetics* 101424. [Crossref]
32. Ron Adner, Phanish Puranam, Feng Zhu. 2019. What Is Different About Digital Strategy? From Quantitative to Qualitative Change. *Strategy Science* **4**:4, 253-261. [Crossref]
33. Jorge Mejia, Shawn Mankad, Anandasivam Gopal. 2019. A for Effort? Using the Crowd to Identify Moral Hazard in New York City Restaurant Hygiene Inspections. *Information Systems Research* **30**:4, 1363-1386. [Crossref]
34. Wenbo Wu, Jiaqi Chen, Liang Xu, Qingyun He, Michael L. Tindall. 2019. A statistical learning approach for stock selection in the Chinese stock market. *Financial Innovation* **5**:1. . [Crossref]
35. Horacio E. Rousseau, Pascual Berrone, Liliana Gelabert. 2019. Localizing Sustainable Development Goals: Nonprofit Density and City Sustainability. *Academy of Management Discoveries* **5**:4, 487-513. [Crossref]
36. Stefan P. Penczynski. 2019. Using machine learning for communication classification. *Experimental Economics* **22**:4, 1002-1029. [Crossref]
37. David McKenzie, Dario Sansone. 2019. Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria. *Journal of Development Economics* **141**, 102369. [Crossref]
38. L. Maria Michael Visuwasam, D. Paul Raj. 2019. NMA: integrating big data into a novel mobile application using knowledge extraction for big data analytics. *Cluster Computing* **22**:S6, 14287-14298. [Crossref]
39. Cathy W.S. Chen, Manh Cuong Dong, Nathan Liu, Songsak Sriboonchitta. 2019. Inferences of default risk and borrower characteristics on P2P lending. *The North American Journal of Economics and Finance* **50**, 101013. [Crossref]

40. Ashwin Madhou, Tayushma Sewak, Imad Moosa, Vikash Ramiah. 2019. Forecasting the GDP of a small open developing economy: an application of FAVAR models. *Applied Economics* **1**, 1-12. [Crossref]
41. Nicolas Huck. 2019. Large data sets and machine learning: Applications to statistical arbitrage. *European Journal of Operational Research* **278**:1, 330-342. [Crossref]
42. Katsuyuki Tanaka, Takuo Higashide, Takuji Kinkyo, Shigeyuki Hamori. 2019. ANALYZING INDUSTRY-LEVEL VULNERABILITY BY PREDICTING FINANCIAL BANKRUPTCY. *Economic Inquiry* **57**:4, 2017-2034. [Crossref]
43. Giorgio Gnecco, Federico Nutarelli. 2019. On the trade-off between number of examples and precision of supervision in machine learning problems. *Optimization Letters* **113**. . [Crossref]
44. Christian Lessmann, Arne Steinkraus. 2019. The geography of natural resources, ethnic inequality and civil conflicts. *European Journal of Political Economy* **59**, 33-51. [Crossref]
45. Clint L.P. Pennings, Jan van Dalen, Laurens Rook. 2019. Coordinating judgmental forecasting: Coping with intentional biases. *Omega* **87**, 46-56. [Crossref]
46. Liqian Cai, Arnab Bhattacharjee, Roger Calantone, Taps Maiti. 2019. Variable Selection with Spatially Autoregressive Errors: A Generalized Moments LASSO Estimator. *Sankhya B* **81**:S1, 146-200. [Crossref]
47. Willem Boshoff, Rossouw Jaarsveld. 2019. Market Definition Using Consumer Characteristics and Cluster Analysis. *South African Journal of Economics* **87**:3, 302-325. [Crossref]
48. Hugo Storm, Kathy Baylis, Thomas Heckelei. 2019. Machine learning in agricultural and applied economics. *European Review of Agricultural Economics* **105**. . [Crossref]
49. Xia Li, Ruibin Bai, Peer-Olaf Siebers, Christian Wagner. 2019. Travel time prediction in transport and logistics. *VINE Journal of Information and Knowledge Management Systems* **49**:3, 277-306. [Crossref]
50. Susan Athey, Guido W. Imbens. 2019. Machine Learning Methods That Economists Should Know About. *Annual Review of Economics* **11**:1, 685-725. [Crossref]
51. Qiuqin He, Bing Xu. 2019. Determinants of economic growth: A varying-coefficient path identification approach. *Journal of Business Research* **101**, 811-818. [Crossref]
52. Lucy C. Sorensen. 2019. "Big Data" in Educational Administration: An Application for Predicting School Dropout Risk. *Educational Administration Quarterly* **55**:3, 404-446. [Crossref]
53. Mario Molina, Filiz Garip. 2019. Machine Learning for Sociology. *Annual Review of Sociology* **45**:1, 27-45. [Crossref]
54. Yi Ren, Tong Xia, Yong Li, Xiang Chen. 2019. Predicting socio-economic levels of urban regions via offline and online indicators. *PLOS ONE* **14**:7, e0219058. [Crossref]
55. Jermain C. Kaminski, Christian Hopp. 2019. Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals. *Small Business Economics* **39**. . [Crossref]
56. Kenneth David Strang, Zhaohao Sun. 2019. Hidden big data analytics issues in the healthcare industry. *Health Informatics Journal* **1**, 146045821985460. [Crossref]
57. Desamparados Blazquez, Josep Domenech, Jose A. Gil, Ana Pont. 2019. Monitoring e-commerce adoption from online data. *Knowledge and Information Systems* **60**:1, 227-245. [Crossref]
58. Yu-Chien Ko, Yang-Yin Ting, Hamido Fujita. 2019. A visual analytics with evidential inference for big data: case study of chemical vapor deposition in solar company. *Granular Computing* **4**:3, 531-544. [Crossref]
59. Marco Castellani. 2019. Does culture matter for the economic performance of countries? An overview of the literature. *Journal of Policy Modeling* **41**:4, 700-717. [Crossref]

60. Jens Prüfer, Patricia Prüfer. 2019. Data science for entrepreneurship research: studying demand dynamics for entrepreneurial skills in the Netherlands. *Small Business Economics* **54**. . [Crossref]
61. Abigail N. Devereaux. 2019. The nudge wars: A modern socialist calculation debate. *The Review of Austrian Economics* **32**:2, 139-158. [Crossref]
62. Rui Gonçalves, Vitor Miguel Ribeiro, Fernando Lobo Pereira, Ana Paula Rocha. 2019. Deep learning in exchange markets. *Information Economics and Policy* **47**, 38-51. [Crossref]
63. Jeffrey T. Prince. 2019. A paradigm for assessing the scope and performance of predictive analytics. *Information Economics and Policy* **47**, 7-13. [Crossref]
64. Feiyu Hu, Jim Warren, Daniel J. Exeter. 2019. Geography and patient history in long-term lipid lowering medication adherence for primary prevention of cardiovascular disease. *Spatial and Spatio-temporal Epidemiology* **29**, 13-29. [Crossref]
65. Deepak Gupta, Rinkle Rani. 2019. A study of big data evolution and research challenges. *Journal of Information Science* **45**:3, 322-340. [Crossref]
66. Henry E. Brady. 2019. The Challenge of Big Data and Data Science. *Annual Review of Political Science* **22**:1, 297-323. [Crossref]
67. Michael Mayer, Steven C. Bourassa, Martin Hoesli, Donato Scognamiglio. 2019. Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research* **12**:1, 134-150. [Crossref]
68. Ajay Agrawal, Joshua S. Gans, Avi Goldfarb. 2019. Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction. *Journal of Economic Perspectives* **33**:2, 31-50. [Abstract] [View PDF article] [PDF with links]
69. Susan Athey, Mohsen Bayati, Guido Imbens, Zhaonan Qu. 2019. Ensemble Methods for Causal Effects in Panel Data Settings. *AEA Papers and Proceedings* **109**, 65-70. [Abstract] [View PDF article] [PDF with links]
70. Cetin Ciner. 2019. Do industry returns predict the stock market? A reprise using the random forest. *The Quarterly Review of Economics and Finance* **72**, 152-158. [Crossref]
71. Heiko Kirchhain, Jan Mutl, Joachim Zietz. 2019. The Impact of Exogenous Shocks on House Prices: the Case of the Volkswagen Emissions Scandal. *The Journal of Real Estate Finance and Economics* **177**. . [Crossref]
72. Jong-Min Kim, Hojin Jung. 2019. Predicting bid prices by using machine learning methods. *Applied Economics* **51**:19, 2011-2018. [Crossref]
73. Alex Singleton, Daniel Arribas-Bel. 2019. Geographic Data Science. *Geographical Analysis* **1**. . [Crossref]
74. Patrick Dunleavy, Mark Evans. 2019. Australian administrative elites and the challenges of digital-era change. *Journal of Chinese Governance* **4**:2, 181-200. [Crossref]
75. Pelin Demirel, Qian Cher Li, Francesco Rentocchini, J. Pawan Tamvada. 2019. Born to be green: new insights into the economics and management of green entrepreneurship. *Small Business Economics* **52**:4, 759-771. [Crossref]
76. O. Rampado, L. Gianusso, C.R. Nava, R. Ropolo. 2019. Analysis of a CT patient dose database with an unsupervised clustering approach. *Physica Medica* **60**, 91-99. [Crossref]
77. Dario Sansone. 2019. Beyond Early Warning Indicators: High School Dropout and Machine Learning. *Oxford Bulletin of Economics and Statistics* **81**:2, 456-485. [Crossref]
78. Colin F. Camerer, Gideon Nave, Alec Smith. 2019. Dynamic Unstructured Bargaining with Private Information: Theory, Experiment, and Outcome Prediction via Machine Learning. *Management Science* **65**:4, 1867-1890. [Crossref]

79. Emmanuel Silva, Hossein Hassani, Dag Madsen, Liz Gee. 2019. Googling Fashion: Forecasting Fashion Consumer Behaviour Using Google Trends. *Social Sciences* **8**:4, 111. [Crossref]
80. Paolo Brunori, Vito Peragine, Laura Serlenga. 2019. Upward and downward bias when measuring inequality of opportunity. *Social Choice and Welfare* **52**:4, 635-661. [Crossref]
81. Ilias Pasidis. 2019. Congestion by accident? A two-way relationship for highways in England. *Journal of Transport Geography* **76**, 301-314. [Crossref]
82. Mustafa Yahşi, Ethem Çanakoğlu, Semra Ağralı. 2019. Carbon price forecasting models based on big data analytics. *Carbon Management* **10**:2, 175-187. [Crossref]
83. Chinmay Kakatkar, Martin Spann. 2019. Marketing analytics using anonymized and fragmented tracking data. *International Journal of Research in Marketing* **36**:1, 117-136. [Crossref]
84. Xia Liu. 2019. Analyzing the impact of user-generated content on B2B Firms' stock performance: Big data analysis with machine learning methods. *Industrial Marketing Management* . [Crossref]
85. Kohei Kawamura, Yohei Kobashi, Masato Shizume, Kozo Ueda. 2019. Strategic central bank communication: Discourse analysis of the Bank of Japan's Monthly Report. *Journal of Economic Dynamics and Control* **100**, 230-250. [Crossref]
86. Nicholas Berente, Stefan Seidel, Hani Safadi. 2019. Research Commentary—Data-Driven Computationally Intensive Theory Development. *Information Systems Research* **30**:1, 50-64. [Crossref]
87. Koffi Dumor, Li Yao. 2019. Estimating China's Trade with Its Partner Countries within the Belt and Road Initiative Using Neural Network Analysis. *Sustainability* **11**:5, 1449. [Crossref]
88. Wolfram Höpken, Tobias Eberle, Matthias Fuchs, Maria Lexhagen. 2019. Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden. *Information Technology & Tourism* **21**:1, 45-62. [Crossref]
89. Jinu Lee. 2019. A Neural Network Method for Nonlinear Time Series Analysis. *Journal of Time Series Econometrics* **11**:1. . [Crossref]
90. Krista L. Uggerslev, Frank Bosco. Raising the Ante 745-760. [Crossref]
91. Erik Nelson, John Fitzgerald, Nathan Tefft. 2019. The distributional impact of a green payment policy for organic fruit. *PLOS ONE* **14**:2, e0211199. [Crossref]
92. Ron S. Jarmin. 2019. Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics. *Journal of Economic Perspectives* **33**:1, 165-184. [Abstract] [View PDF article] [PDF with links]
93. Michael Friendly, Jürgen Symanzik, Ortac Onder. 2019. Visualising the Titanic disaster. *Significance* **16**:1, 14-19. [Crossref]
94. Yan Liu, Tian Xie. 2019. Machine learning versus econometrics: prediction of box office. *Applied Economics Letters* **26**:2, 124-130. [Crossref]
95. Eli P Fenichel, Yukiko Hashida. 2019. Choices and the value of natural capital. *Oxford Review of Economic Policy* **35**:1, 120-137. [Crossref]
96. Jorge Iván Pérez-Rave, Juan Carlos Correa-Morales, Favián González-Echavarría. 2019. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research* **36**:1, 59-96. [Crossref]
97. Cinzia Daraio. Econometric Approaches to the Measurement of Research Productivity 633-666. [Crossref]
98. Evgeniy M. Ozhegov, Daria Teterina. Methods of Machine Learning for Censored Demand Prediction 441-446. [Crossref]

99. Pier Francesco De Maria, Leonardo Tomazeli Duarte, Álvaro de Oliveira D'Antona, Cristiano Torezzan. Digital Humanities and Big Microdata: New Approaches for Demographic Research 217-231. [Crossref]
100. Raffaele Dell'Aversana, Edgardo Bucciarelli. Towards a Natural Experiment Leveraging Big Data to Analyse and Predict Users' Behavioural Patterns Within an Online Consumption Setting 103-113. [Crossref]
101. Andrew Haughwout, Benjamin R. Mandel. Empirical analysis of the US consumer 1-21. [Crossref]
102. Thomas B. Götz, Thomas A. Knetsch. 2019. Google data in bridge equation models for German GDP. *International Journal of Forecasting* **35**:1, 45-66. [Crossref]
103. Yang Xiao, De Wang, Jia Fang. 2019. Exploring the disparities in park access through mobile phone data: Evidence from Shanghai, China. *Landscape and Urban Planning* **181**, 80-91. [Crossref]
104. Jessica Lichy, Maher Kachour. Big Data Perception & Usage 89-94. [Crossref]
105. Jens Prufer, Patricia Prufer. 2019. Data Science for Entrepreneurship Research: Studying Demand Dynamics for Entrepreneurial Skills in the Netherlands. *SSRN Electronic Journal* . [Crossref]
106. Ajay Agrawal, Joshua S. Gans, Avi Goldfarb. 2019. Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction. *SSRN Electronic Journal* . [Crossref]
107. Bo Cowgill, Catherine E. Tucker. 2019. Economics, Fairness and Algorithmic Bias. *SSRN Electronic Journal* . [Crossref]
108. Laurent Ferrara, Anna Simoni. 2019. When Are Google Data Useful to Nowcast Gdp? An Approach Via Pre-Selection and Shrinkage. *SSRN Electronic Journal* . [Crossref]
109. Jan Abrell, Mirjam Kosch, Sebastian Rausch. 2019. How Effective Was the UK Carbon Tax?—A Machine Learning Approach to Policy Evaluation. *SSRN Electronic Journal* . [Crossref]
110. George G. Judge. 2019. Combining the Information From Econometrics Learning (EL) and Machine Learning (ML). *SSRN Electronic Journal* . [Crossref]
111. John A. Clithero, Jae Joon Lee, Joshua Tasoff. 2019. Supervised Machine Learning for Eliciting Individual Reservation Values. *SSRN Electronic Journal* . [Crossref]
112. Muhammad Zia Hydari, Idris Adjerid, Aaron Striegel. 2019. Health Wearables, Gamification, and Healthful Activity. *SSRN Electronic Journal* . [Crossref]
113. Kenneth David Strang, Zhaohao Sun. Managerial Controversies in Artificial Intelligence and Big Data Analytics 55-74. [Crossref]
114. Marvin N. Wright, Inke R. König. 2019. Splitting on categorical predictors in random forests. *PeerJ* **7**, e6339. [Crossref]
115. Manuel J. García Rodríguez, Vicente Rodríguez Montequín, Francisco Ortega Fernández, Joaquín M. Villanueva Balsera. 2019. Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning. *Complexity* **2019**, 1. [Crossref]
116. Alessandro Roncaglia. . [Crossref]
117. Alexandre Rubesam. 2019. Machine Learning Portfolios with Equal Risk Contributions. *SSRN Electronic Journal* . [Crossref]
118. Hossein Hassani, Xu Huang, Emmanuel Sirimal Silva. Big Data and Blockchain 7-48. [Crossref]
119. Giorgio Gnecco, Federico Nutarelli. Optimal Trade-Off Between Sample Size and Precision of Supervision for the Fixed Effects Panel Data Model 531-542. [Crossref]
120. Danxia Xie, Longtian Zhang, Ke Tang, Zhen Sun. 2019. Data in Growth Model. *SSRN Electronic Journal* . [Crossref]
121. Kenny Ching, Enrico Forti, Evan Rawley. 2019. Extemporaneous Coordination in Specialist Teams: The Familiarity Complementarity. *SSRN Electronic Journal* . [Crossref]

122. Andres Algaba, David Ardia, Keven Bluteau, Samuel Borms, Kris Boudt. 2019. Econometrics Meets Sentiment: An Overview of Methodology and Applications. *SSRN Electronic Journal* . [Crossref]

123. Nicolas Pröllochs, Stefan Feuerriegel, Dirk Neumann. 2018. Statistical inferences for polarity identification in natural language. *PLOS ONE* **13**:12, e0209323. [Crossref]

124. Atin Basuchoudhary, James T. Bang. 2018. Predicting Terrorism with Machine Learning: Lessons from "Predicting Terrorism: A Machine Learning Approach". *Peace Economics, Peace Science and Public Policy* **24**:4. . [Crossref]

125. Fritz Schiltz, Chiara Masci, Tommaso Agasisti, Daniel Horn. 2018. Using regression tree ensembles to model interaction effects: a graphical approach. *Applied Economics* **50**:58, 6341-6354. [Crossref]

126. Fritz Schiltz, Paolo Sestito, Tommaso Agasisti, Kristof De Witte. 2018. The added value of more accurate predictions for school rankings. *Economics of Education Review* **67**, 207-215. [Crossref]

127. Arthur Dyevre, Nicolas Lampach. 2018. The origins of regional integration: Untangling the effect of trade on judicial cooperation. *International Review of Law and Economics* **56**, 122-133. [Crossref]

128. Monica Andini, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, Viola Salvestrini. 2018. Targeting with machine learning: An application to a tax rebate program in Italy. *Journal of Economic Behavior & Organization* **156**, 86-102. [Crossref]

129. Eder Johnson de Area Leão Pereira, Marcus Fernandes da Silva, I.C. da Cunha Lima, H.B.B. Pereira. 2018. Trump's Effect on stock markets: A multiscale approach. *Physica A: Statistical Mechanics and its Applications* **512**, 241-247. [Crossref]

130. Otto Kässi, Vili Lehdonvirta. 2018. Online labour index: Measuring the online gig economy for policy and research. *Technological Forecasting and Social Change* **137**, 241-248. [Crossref]

131. Marco Pangallo, Michele Loberto. 2018. Home is where the ad is: online interest proxies housing demand. *EPJ Data Science* **7**:1. . [Crossref]

132. Gary Smith. 2018. Step away from stepwise. *Journal of Big Data* **5**:1. . [Crossref]

133. Alex Coad, Dominik Janzing, Paul Nightingale. 2018. Tools for causal inference from cross-sectional innovation surveys with continuous or discrete variables: Theory and applications. *Cuadernos de Economía* **37**:75, 779-808. [Crossref]

134. Misheck Mutize, Sean Joss Gossel. 2018. Do sovereign credit rating announcements influence excess bond and equity returns in Africa?. *International Journal of Emerging Markets* **13**:6, 1522-1537. [Crossref]

135. Maddalena Cavicchioli, Angeliki Papana, Ariadni Papana Dagiasis, Barbara Pistoresi. 2018. Maximum Likelihood Estimation for the Generalized Pareto Distribution and Goodness-of-Fit Test with Censored Data. *Journal of Modern Applied Statistical Methods* **17**:2. . [Crossref]

136. Bo Xiong, Yuhe Song. 2018. Big Data and Dietary Trend: The Case of Avocado Imports in China. *Journal of International Food & Agribusiness Marketing* **30**:4, 343-354. [Crossref]

137. Ranjith Vijayakumar, Mike W.-L. Cheung. 2018. Replicability of Machine Learning Models in the Social Sciences. *Zeitschrift für Psychologie* **226**:4, 259-273. [Crossref]

138. Chiara Masci, Geraint Johnes, Tommaso Agasisti. 2018. Student and school performance across countries: A machine learning approach. *European Journal of Operational Research* **269**:3, 1072-1085. [Crossref]

139. Stelios Michalopoulos, Elias Papaioannou. 2018. Spatial Patterns of Development: A Meso Approach. *Annual Review of Economics* **10**:1, 383-410. [Crossref]

140. Matheus Albergaria, Maria Sylvia Saes. 2018. Measuring externalities in an information commons: the case of libraries. *Journal of Cleaner Production* **192**, 855-863. [Crossref]

141. Carsten Fink, Christian Helmers, Carlos J. Ponce. 2018. Trademark squatters: Theory and evidence from Chile. *International Journal of Industrial Organization* **59**, 340-371. [Crossref]

142. Ozalp Babaoglu, Alina Sirbu. Cognified Distributed Computing 1180-1191. [Crossref]

143. Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science. *ACM Transactions on Interactive Intelligent Systems* **8**:2, 1-20. [Crossref]

144. Diego Aparicio, Marcos López de Prado. 2018. How hard is it to pick the right model? MCS and backtest overfitting. *Algorithmic Finance* **7**:1-2, 53-61. [Crossref]

145. Olga Takács, János Vincze. 2018. Bérelőrejelzések – prediktorok és tanulságok. *Közgazdasági Szemle* **65**:6, 592-618. [Crossref]

146. Vincenzo Butticè, Carlotta Orsenigo, Mike Wright. 2018. The effect of information asymmetries on serial crowdfunding and campaign success. *Economia e Politica Industriale* **45**:2, 143-173. [Crossref]

147. Guy David, Philip A. Saynisch, Aaron Smith-McLallen. 2018. The economics of patient-centered care. *Journal of Health Economics* **59**, 60-77. [Crossref]

148. Desamparados Blazquez, Josep Domenech. 2018. Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change* **130**, 99-113. [Crossref]

149. Katsuyuki Tanaka, Takuji Kinkyo, Shigeyuki Hamori. 2018. Financial Hazard Map: Financial Vulnerability Predicted by a Random Forests Classification Model. *Sustainability* **10**:5, 1530. [Crossref]

150. Baban Hasnat. 2018. Big Data: An Institutional Perspective on Opportunities and Challenges. *Journal of Economic Issues* **52**:2, 580-588. [Crossref]

151. Benjamin Seligman, Shripad Tuljapurkar, David Rehkopf. 2018. Machine learning approaches to the social determinants of health in the health and retirement study. *SSM - Population Health* **4**, 95-99. [Crossref]

152. Patrick Mikalef, Michail N. Giannakos, Ilias O. Pappas, John Krogstie. The human side of big data: Understanding the skills of the data scientist in education and industry 503-512. [Crossref]

153. Jessica M. Franklin, Chandrasekar Gopalakrishnan, Alexis A. Krumme, Karandeep Singh, James R. Rogers, Joe Kimura, Caroline McKay, Newell E. McElwee, Niteesh K. Choudhry. 2018. The relative benefits of claims and electronic health record data for predicting medication adherence trajectory. *American Heart Journal* **197**, 153-162. [Crossref]

154. Myron P. Gutmann, Emily Klancher Merchant, Evan Roberts. 2018. "Big Data" in Economic History. *The Journal of Economic History* **78**:1, 268-299. [Crossref]

155. Mochen Yang, Gediminas Adomavicius, Gordon Burtch, Yuqing Ren. 2018. Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining. *Information Systems Research* **29**:1, 4-24. [Crossref]

156. Gregorio Caetano, Vikram Maheshri. 2018. Identifying dynamic spillovers of crime with a causal approach to model selection. *Quantitative Economics* **9**:1, 343-394. [Crossref]

157. Keith H Coble, Ashok K Mishra, Shannon Ferrell, Terry Griffin. 2018. Big Data in Agriculture: A Challenge for the Future. *Applied Economic Perspectives and Policy* **40**:1, 79-96. [Crossref]

158. Stephan D. Whitaker. 2018. Big Data versus a survey. *The Quarterly Review of Economics and Finance* **67**, 285-296. [Crossref]

159. Patrick Zschech, Vera Fleißner, Nicole Baumgärtel, Andreas Hilbert. 2018. Data Science Skills and Enabling Enterprise Systems. *HMD Praxis der Wirtschaftsinformatik* **55**:1, 163-181. [Crossref]

160. Gilbert Saporta. From Conventional Data Analysis Methods to Big Data Analytics 27-41. [Crossref]

161. Desamparados BLAZQUEZ, Josep DOMENECH. 2018. WEB DATA MINING FOR MONITORING BUSINESS EXPORT ORIENTATION. *Technological and Economic Development of Economy* **24**:2, 406-428. [Crossref]
162. Rimvydas Skyrius, Gintarė Giriūnienė, Igor Katin, Michail Kazimianec, Raimundas Žilinskas. The Potential of Big Data in Banking 451-486. [Crossref]
163. Chaitanya Baru. Data in the 21st Century 3-17. [Crossref]
164. Yong Yoon. Spatial Choice Modeling Using the Support Vector Machine (SVM): Characterization and Prediction 767-778. [Crossref]
165. Wolfram Höpken, Tobias Eberle, Matthias Fuchs, Maria Lexhagen. Search Engine Traffic as Input for Predicting Tourist Arrivals 381-393. [Crossref]
166. Shu-Heng Chen, Ye-Rong Du, Ying-Fang Kao, Ragupathy Venkatachalam, Tina Yu. On Complex Economic Dynamics: Agent-Based Computational Modeling and Beyond 1-14. [Crossref]
167. Thomas K. Bauer, Phillip Breidenbach, Sandra Schaffner. Big Data in der wirtschaftswissenschaftlichen Forschung 129-148. [Crossref]
168. Cinzia Daraio. Nonparametric Methods and Higher Education 1-7. [Crossref]
169. Peng Ye, Julian Qian, Jieying Chen, Chen-hung Wu, Yitong Zhou, Spencer De Mars, Frank Yang, Li Zhang. Customized Regression Model for Airbnb Dynamic Pricing 932-940. [Crossref]
170. Thiago Gonçalves dos Santos Martins, Ana Luiza Fontes de Azevedo Costa, Thomaz Gonçalves dos Santos Martins. 2018. Big Data use in medical research. *Einstein (São Paulo)* **16**:3. . [Crossref]
171. Andre Boik. 2018. Prediction and Identification in Two-Sided Markets. *SSRN Electronic Journal* . [Crossref]
172. Lucie Martin-Bonnel de Longchamp, Nicolas Lampach, Ludovic Parisot. 2018. How Cognitive Biases Affect Energy Savings in Low Energy Buildings. *SSRN Electronic Journal* . [Crossref]
173. Jens Prufer, Patricia Prufer. 2018. Data Science for Institutional and Organizational Economics. *SSRN Electronic Journal* . [Crossref]
174. Phanish Puranam, Yash Raj Shrestha, Vivianna Fang He, Georg von Krogh. 2018. Algorithmic Induction Through Machine Learning: Opportunities for Management and Organization Research. *SSRN Electronic Journal* . [Crossref]
175. Oz Shy. 2018. Alternative Methods for Studying Consumer Payment Choice. *SSRN Electronic Journal* . [Crossref]
176. Akos Lada, Diego Aparicio, Michael Bailey. 2018. Predicting Heterogeneous Treatment Effects in Ranking Systems. *SSRN Electronic Journal* . [Crossref]
177. Silvia Emili, Attilio Gardini. 2018. High Spatial and Temporal Detail in Timely Prediction of Tourism Demand. *SSRN Electronic Journal* . [Crossref]
178. Santiago Carbo-Valverde, Pedro Cuadros-Solas, Francisco Rodriguez-Fernandez. 2018. How Do Bank Customers Go Digital? A Random Forest Approach. *SSRN Electronic Journal* . [Crossref]
179. Jeffrey Prince. 2018. A Paradigm for Assessing the Scope and Performance of Predictive Analytics. *SSRN Electronic Journal* . [Crossref]
180. Matthew Grennan, Kyle Myers, Ashley Teres Swanson, Aaron Chatterji. 2018. Physician-Industry Interactions: Persuasion and Welfare. *SSRN Electronic Journal* . [Crossref]
181. Roberto Casarin, Fausto Corradin, Francesco Ravazzolo, Domenico Sartore. 2018. A Scoring Rule for Factor and Autoregressive Models Under Misspecification. *SSRN Electronic Journal* . [Crossref]
182. Marco Pangallo, Michele Loberto. 2018. Home Is Where the Ad Is: Online Interest Proxies Housing Demand. *SSRN Electronic Journal* . [Crossref]

183. Evgeniy Ozhegov, Daria Teterina. 2018. The Ensemble Method for Censored Demand Prediction. *SSRN Electronic Journal* . [Crossref]

184. Julian TszKin Chan, Weifeng Zhong. 2018. Reading China: Predicting Policy Change with Machine Learning. *SSRN Electronic Journal* . [Crossref]

185. Chinmay Kakatkar, Volker Bilgram, Johann Füller. 2018. Innovation Analytics: Leveraging Artificial Intelligence in the Innovation Process. *SSRN Electronic Journal* . [Crossref]

186. Michael Mayer, Steven C. Bourassa, Martin Edward Ralph Hoesli, Donato Flavio Scognamiglio. 2018. Estimation and Updating Methods for Hedonic Valuation. *SSRN Electronic Journal* . [Crossref]

187. Tommaso Tani. L'incidenza dei big data e del machine learning sui principi alla base del Regolamento Europeo per la tutela dei dati personali (2016/679/UE) e proposte per una nuova normativa in tema di privacy 35-65. [Crossref]

188. Pilsun Choi, Insik Min. 2018. A Predictive Model for the Employment of College Graduates Using a Machine Learning Approach. *Journal of Vocational Education & Training* **21**:1, 31. [Crossref]

189. Monika Glavina. 2018. 'To Submit or Not to Submit – That Is the (Preliminary) Question': Explaining National Judges' Reluctance to Participate in the Preliminary Ruling Procedure. *SSRN Electronic Journal* . [Crossref]

190. Roberto Moro Visconti, Giuseppe Montesi, Giovanni Papiro. 2018. Big data-driven stochastic business planning and corporate valuation. *Corporate Ownership and Control* **15**:3-1, 189-204. [Crossref]

191. Benjamin Bluhm. 2018. Time Series Econometrics at Scale: A Practical Guide to Parallel Computing in (Py)Spark. *SSRN Electronic Journal* . [Crossref]

192. Thomas Renault. 2018. 2. Données massives et recherche en économie : une (r)évolution ?. *Regards croisés sur l'économie* **n°23**:2, 32. [Crossref]

193. Gilles Bastin, Paola Tubaro. 2018. Le moment big data des sciences sociales. *Revue française de sociologie* **59**:3, 375. [Crossref]

194. Sumit Agarwal, Long Wang, Yang Yang. 2018. Blessing in Disguise? Environmental Shocks and Performance Enhancement. *SSRN Electronic Journal* . [Crossref]

195. Raghavendra Rau. 2017. Social networks and financial outcomes. *Current Opinion in Behavioral Sciences* **18**, 75-78. [Crossref]

196. Chris Schilling, Josh Knight, Duncan Mortimer, Dennis Petrie, Philip Clarke, John Chalmers, Andrew Kerr, Rod Jackson. 2017. Australian general practitioners initiate statin therapy primarily on the basis of lipid levels; New Zealand general practitioners use absolute risk. *Health Policy* **121**:12, 1233-1239. [Crossref]

197. Lei Dong, Sicong Chen, Yunsheng Cheng, Zhengwei Wu, Chao Li, Haishan Wu. 2017. Measuring economic activity in China with mobile big data. *EPJ Data Science* **6**:1. . [Crossref]

198. Thiago Gonçalves dos Santos Martins, Ana Luiza Fontes de Azevedo Costa. 2017. A new way to communicate science in the era of Big Data and citizen science. *Einstein (São Paulo)* **15**:4, 523-523. [Crossref]

199. János Vincze. 2017. Információ és tudás. A big data egyes hatásai a közgazdaságtanra. *Közgazdasági Szemle* **64**:11, 1148-1159. [Crossref]

200. Paola D'Orazio. 2017. Big data and complexity: Is macroeconomics heading toward a new paradigm?. *Journal of Economic Methodology* **24**:4, 410-429. [Crossref]

201. Shu-Heng Chen, Ragupathy Venkatachalam. 2017. Agent-based modelling as a foundation for big data. *Journal of Economic Methodology* **24**:4, 362-383. [Crossref]

202. Teck-Hua Ho, Noah Lim, Sadat Reza, Xiaoyu Xia. 2017. OM Forum—Causal Inference Models in Operations Management. *Manufacturing & Service Operations Management* **19**:4, 509-525. [Crossref]

203. Ernesto D'Avanzo, Giovanni Pilato, Miltiadis Lytras. 2017. Using Twitter sentiment and emotions analysis of Google Trends for decisions making. *Program* **51**:3, 322-350. [Crossref]
204. Takayuki Morimoto, Yoshinori Kawasaki. 2017. Forecasting Financial Market Volatility Using a Dynamic Topic Model. *Asia-Pacific Financial Markets* **24**:3, 149-167. [Crossref]
205. Rajesh Chandy, Magda Hassan, Prokriti Mukherji. 2017. Big Data for Good: Insights from Emerging Markets*. *Journal of Product Innovation Management* **34**:5, 703-713. [Crossref]
206. Silvia Mendolia, Peter Siminski. 2017. Is education the mechanism through which family background affects economic outcomes? A generalised approach to mediation analysis. *Economics of Education Review* **59**, 1-12. [Crossref]
207. Yuh-Jong Hu, Shu-Wei Huang. Challenges of automated machine learning on causal impact analytics for policy evaluation 1-6. [Crossref]
208. Carlos Tapia, Beñat Abajo, Efren Feliu, Maddalen Mendizabal, José Antonio Martinez, J. German Fernández, Txomin Laburu, Adelaida Lejarazu. 2017. Profiling urban vulnerabilities to climate change: An indicator-based vulnerability assessment for European cities. *Ecological Indicators* **78**, 142-155. [Crossref]
209. Jessica Lichy, Maher Kachour, Tatiana Khvatova. 2017. Big Data is watching YOU: opportunities and challenges from the perspective of young adult consumers in Russia. *Journal of Marketing Management* **33**:9-10, 719-741. [Crossref]
210. Adam Nowak, Patrick Smith. 2017. Textual Analysis in Real Estate. *Journal of Applied Econometrics* **32**:4, 896-918. [Crossref]
211. Christopher Krauss, Xuan Anh Do, Nicolas Huck. 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research* **259**:2, 689-702. [Crossref]
212. Sendhil Mullainathan, Jann Spiess. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* **31**:2, 87-106. [Abstract] [View PDF article] [PDF with links]
213. Aloisio Dourado, Rommel N. Carvalho, Gustavo C. G. van Erven. Brazil's Bolsa Familia and young adult workers: A parallel RDD approach to large datasets 17-24. [Crossref]
214. Michael Creel. 2017. Neural nets for indirect inference. *Econometrics and Statistics* **2**, 36-49. [Crossref]
215. Xin Li, Bing Pan, Rob Law, Xiankai Huang. 2017. Forecasting tourism demand with composite search index. *Tourism Management* **59**, 57-66. [Crossref]
216. Yael Grushka-Cockayne, Victor Richmond R. Jose, Kenneth C. Lichtendahl. 2017. Ensembles of Overfit and Overconfident Forecasts. *Management Science* **63**:4, 1110-1130. [Crossref]
217. Patricia Kuzmenko FURLAN, Fernando José Barbin LAURINDO. 2017. Agrupamentos epistemológicos de artigos publicados sobre big data analytics. *Transinformação* **29**:1, 91-100. [Crossref]
218. Felix Ward. 2017. Spotting the Danger Zone: Forecasting Financial Crises With Classification Tree Ensembles and Many Predictors. *Journal of Applied Econometrics* **32**:2, 359-378. [Crossref]
219. Omar A. Guerrero, Eduardo López. 2017. Understanding Unemployment in the Era of Big Data: Policy Informed by Data-Driven Theory. *Policy & Internet* **9**:1, 28-54. [Crossref]
220. Thomas Pave Sohnesen, Niels Stender. 2017. Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment. *Poverty & Public Policy* **9**:1, 118-133. [Crossref]
221. Benjamin David. 2017. Computer technology and probable job destructions in Japan: An evaluation. *Journal of the Japanese and International Economies* **43**, 77-87. [Crossref]

222. Greg Distelhorst, Jens Hainmueller, Richard M. Locke. 2017. Does Lean Improve Labor Standards? Management and Social Performance in the Nike Supply Chain. *Management Science* **63**:3, 707-728. [Crossref]

223. Sebastian Tillmanns, Frenkel Ter Hofstede, Manfred Krafft, Oliver Goetz. 2017. How to Separate the Wheat from the Chaff: Improved Variable Selection for New Customer Acquisition. *Journal of Marketing* **81**:2, 99-113. [Crossref]

224. Johan L. Perols, Robert M. Bowen, Carsten Zimmermann, Basamba Samba. 2017. Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction. *The Accounting Review* **92**:2, 221-245. [Crossref]

225. Jayson L. Lusk. 2017. Consumer Research with Big Data: Applications from the Food Demand Survey (FooDS). *American Journal of Agricultural Economics* **99**:2, 303-320. [Crossref]

226. Zaheer Khan, Tim Vorley. 2017. Big data text analytics: an enabler of knowledge management. *Journal of Knowledge Management* **21**:1, 18-34. [Crossref]

227. Chris Schilling, Duncan Mortimer, Kim Dalziel. 2017. Using CART to Identify Thresholds and Hierarchies in the Determinants of Funding Decisions. *Medical Decision Making* **37**:2, 173-182. [Crossref]

228. Kenneth David Strang, Zhaohao Sun. 2017. Analyzing Relationships in Terrorism Big Data Using Hadoop and Statistics. *Journal of Computer Information Systems* **57**:1, 67-75. [Crossref]

229. Haiyan Song, Han Liu. Predicting Tourist Demand Using Big Data 13-29. [Crossref]

230. Carlianne Patrick, Amanda Ross, Heather Stephens. Designing Policies to Spur Economic Growth: How Regional Scientists Can Contribute to Future Policy Development and Evaluation 119-133. [Crossref]

231. Neelam Younas, Zahid Asghar, Muhammad Qayyum, Fazlullah Khan. Education and Socio Economic Factors Impact on Earning for Pakistan - A Bigdata Analysis 215-223. [Crossref]

232. Anne Fleur van Veenstra, Bas Kotterink. Data-Driven Policy Making: The Policy Lab Approach 100-111. [Crossref]

233. Atin Basuchoudhary, James T. Bang, Tinni Sen. Why This Book? 1-6. [Crossref]

234. Khyati Ahlawat, Amit Prakash Singh. A Novel Hybrid Technique for Big Data Classification Using Decision Tree Learning 118-128. [Crossref]

235. Xiangjun Meng, Liang Chen, Yidong Li. A Parallel Clustering Algorithm for Power Big Data Analysis 533-540. [Crossref]

236. Alexander Peysakhovich, Jeffrey Naecker. 2017. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization* **133**, 373-384. [Crossref]

237. Yu Hou, Artur Hugon, Matthew R. Lyle, Seth Pruitt. 2017. Macroeconomic News in the Cross Section of Asset Growth. *SSRN Electronic Journal* . [Crossref]

238. Scott Kostyshak. 2017. Non-Parametric Testing of U-Shapes, with an Application to the Midlife Satisfaction Dip. *SSRN Electronic Journal* . [Crossref]

239. Kweku A. Opoku-Agyemang. 2017. Priming Human-Computer Interactions: Experimental Evidence from Economic Development Mobile Surveys. *SSRN Electronic Journal* . [Crossref]

240. Max Biggs, Rim Hariss. 2017. Optimizing Objective Functions Determined from Random Forests. *SSRN Electronic Journal* . [Crossref]

241. Dong-Jin Pyo. 2017. Can Big Data Help Predict Financial Market Dynamics?: Evidence from the Korean Stock Market. *SSRN Electronic Journal* . [Crossref]

242. Mike Horia Teodorescu. 2017. Machine Learning Methods for Strategy Research. *SSRN Electronic Journal* . [Crossref]
243. Chiranjit Chakraborty, Andreas Joseph. 2017. Machine Learning at Central Banks. *SSRN Electronic Journal* . [Crossref]
244. Nicolas Lampach, Arthur Dyevre. 2017. The Origins of Regional Integration: Untangling the Effect of Trade on Judicial Cooperation. *SSRN Electronic Journal* . [Crossref]
245. Guy David, Phil Saynisch, Aaron Smith-McLallen. 2017. The Economics of Patient-Centered Care. *SSRN Electronic Journal* . [Crossref]
246. Diego Aparicio, Marcos Lopez de Prado. 2017. How Hard Is It to Pick the Right Model?. *SSRN Electronic Journal* . [Crossref]
247. Daniel Fricke. 2017. Financial Crisis Prediction: A Model Comparison. *SSRN Electronic Journal* . [Crossref]
248. Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, Ansgar Walther. 2017. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *SSRN Electronic Journal* . [Crossref]
249. Monica Andini, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, Viola Salvestrini. 2017. Targeting Policy-Compliers with Machine Learning: An Application to a Tax Rebate Programme in Italy. *SSRN Electronic Journal* . [Crossref]
250. Javier Vidal-García, Marta Vidal, Rafael Hernandez Barros. Computational Business Intelligence, Big Data, and Their Role in Business Decisions in the Age of the Internet of Things 249-268. [Crossref]
251. Amy K. Johnson, Tarek Mikati, Supriya D. Mehta. 2016. Examining the themes of STD-related Internet searches to increase specificity of disease forecasting using Internet search terms. *Scientific Reports* **6**:1. . [Crossref]
252. Jacques Bughin. 2016. Reaping the benefits of big data in telecom. *Journal of Big Data* **3**:1. . [Crossref]
253. Dave Donaldson, Adam Storeygard. 2016. The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives* **30**:4, 171-198. [Abstract] [View PDF article] [PDF with links]
254. R. Rajesh. 2016. Forecasting supply chain resilience performance using grey prediction. *Electronic Commerce Research and Applications* **20**, 42-58. [Crossref]
255. Christian Pierdzioch, Marian Risse, Sebastian Rohloff. 2016. Are precious metals a hedge against exchange-rate movements? An empirical exploration using bayesian additive regression trees. *The North American Journal of Economics and Finance* **38**, 27-38. [Crossref]
256. Nuha Almoqren, Mohammed Altayar. The motivations for big data mining technologies adoption in saudi banks 1-8. [Crossref]
257. Uwe Deichmann, Aparajita Goyal, Deepak Mishra. 2016. Will digital technologies transform agriculture in developing countries?. *Agricultural Economics* **47**:S1, 21-33. [Crossref]
258. Michel Wedel, P.K. Kannan. 2016. Marketing Analytics for Data-Rich Environments. *Journal of Marketing* **80**:6, 97-121. [Crossref]
259. Linden McBride, Austin Nichols. 2016. Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning. *The World Bank Economic Review* **2**, lhw056. [Crossref]
260. Ben Vinod. 2016. Big Data in the travel marketplace. *Journal of Revenue and Pricing Management* **15**:5, 352-359. [Crossref]
261. Jaideep Ghosh. 2016. Big Data Analytics: A Field of Opportunities for Information Systems and Technology Researchers. *Journal of Global Information Technology Management* **19**:4, 217-222. [Crossref]

262. Stefan Feuerriegel. 2016. Decision support in healthcare: determining provider influence on treatment outcomes with robust risk adjustment. *Journal of Decision Systems* **25**:4, 371-390. [Crossref]
263. Matthias Duschl. 2016. Firm dynamics and regional resilience: an empirical evolutionary perspective. *Industrial and Corporate Change* **25**:5, 867-883. [Crossref]
264. Reinout Heijungs, Patrik Henriksson, Jeroen Guinée. 2016. Measures of Difference and Significance in the Era of Computer Simulations, Meta-Analysis, and Big Data. *Entropy* **18**:10, 361. [Crossref]
265. Gerard George, Ernst C. Osinga, Dovev Lavie, Brent A. Scott. 2016. Big Data and Data Science Methods for Management Research. *Academy of Management Journal* **59**:5, 1493-1507. [Crossref]
266. Alison L. Bailey, Anne Blackstock-Bernstein, Eve Ryan, Despina Pitsoulakis. DATA MINING WITH NATURAL LANGUAGE PROCESSING AND CORPUS LINGUISTICS 255-275. [Crossref]
267. P. Racca, R. Casarin, F. Squazzoni, P. Dondio. 2016. Resilience of an online financial community to market uncertainty shocks during the recent financial crisis. *Journal of Computational Science* **16**, 190-199. [Crossref]
268. Michael Mann, Eli Melaas, Arun Malik. 2016. Using VIIRS Day/Night Band to Measure Electricity Supply Reliability: Preliminary Results from Maharashtra, India. *Remote Sensing* **8**:9, 711. [Crossref]
269. Benjamin F. Mundell, Hilal Maradit Kremers, Sue Visscher, Kurtis M. Hoppe, Kenton R. Kaufman. 2016. Predictors of Receiving a Prosthesis for Adults With Above-Knee Amputations in a Well-Defined Population. *PM&R* **8**:8, 730-737. [Crossref]
270. Hu Shuijing. Big Data Analytics: Key Technologies and Challenges 141-145. [Crossref]
271. Michael Peneder. 2016. Competitiveness and industrial policy: from rationalities of failure towards the ability to evolve. *Cambridge Journal of Economics* **11**, bew025. [Crossref]
272. Julia Lane. 2016. BIG DATA FOR PUBLIC POLICY: THE QUADRUPLE HELIX. *Journal of Policy Analysis and Management* **35**:3, 708-715. [Crossref]
273. Gérard Biau, Erwan Scornet. 2016. A random forest guided tour. *TEST* **25**:2, 197-227. [Crossref]
274. Joyce P Jacobsen, Laurence M Levin, Zachary Tausanovitch. 2016. Comparing Standard Regression Modeling to Ensemble Modeling: How Data Mining Software Can Improve Economists' Predictions. *Eastern Economic Journal* **42**:3, 387-398. [Crossref]
275. Dror Etzion, J. Alberto Aragon-Correa. 2016. Big Data, Management, and Sustainability. *Organization & Environment* **29**:2, 147-155. [Crossref]
276. William G. Bostic Jr., Ron S. Jarmin, Brian Moyer. 2016. Modernizing Federal Economic Statistics. *American Economic Review* **106**:5, 161-164. [Abstract] [View PDF article] [PDF with links]
277. Alberto Cavallo, Roberto Rigobon. 2016. The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives* **30**:2, 151-178. [Abstract] [View PDF article] [PDF with links]
278. Paul Smith. 2016. Google's MIDAS Touch: Predicting UK Unemployment with Internet Search Data. *Journal of Forecasting* **35**:3, 263-284. [Crossref]
279. Nalan Baştürk, Roberto Casarin, Francesco Ravazzolo, Herman van Dijk. 2016. Computational Complexity and Parallelization in Bayesian Econometric Analysis. *Econometrics* **4**:1, 9. [Crossref]
280. Chris Schilling, Duncan Mortimer, Kim Dalziel, Emma Heeley, John Chalmers, Philip Clarke. 2016. Using Classification and Regression Trees (CART) to Identify Prescribing Thresholds for Cardiovascular Disease. *PharmacoEconomics* **34**:2, 195-205. [Crossref]
281. Alexander T. Janke, Daniel L. Overbeek, Keith E. Kocher, Phillip D. Levy. 2016. Exploring the Potential of Predictive Analytics and Big Data in Emergency Care. *Annals of Emergency Medicine* **67**:2, 227-236. [Crossref]

282. Jessica M. Franklin, William H. Shrank, Joyce Lii, Alexis K. Krumme, Olga S. Matlin, Troyen A. Brennan, Niteesh K. Choudhry. 2016. Observing versus Predicting: Initial Patterns of Filling Predict Long-Term Adherence More Accurately Than High-Dimensional Modeling Techniques. *Health Services Research* **51**:1, 220-239. [Crossref]
283. . Spotlight 1: How the internet promotes development 42-46. [Crossref]
284. Karsten Luebke, Joachim Rojahn. Firm-Specific Determinants on Dividend Changes: Insights from Data Mining 335-344. [Crossref]
285. Ali Emrouznejad, Marianna Marra. Big Data: Who, What and Where? Social, Cognitive and Journals Map of Big Data Publications with Focus on Optimization 1-16. [Crossref]
286. Richard W. Evans, Kenneth L. Judd, Kramer Quist. Big Data Techniques as a Solution to Theory Problems 219-231. [Crossref]
287. Luca Onorante, Adrian E. Raftery. 2016. Dynamic model averaging in large model spaces using dynamic Occam#s window. *European Economic Review* **81**, 2-14. [Crossref]
288. Michael S. Hand, Matthew P. Thompson, David E. Calkin. 2016. Examining heterogeneity and wildfire management expenditures using spatially and temporally descriptive data. *Journal of Forest Economics* **22**, 80-102. [Crossref]
289. Scott McQuade, Claire Monteleoni. Online Learning of Volatility from Multiple Option Term Lengths 1-3. [Crossref]
290. Erik Nelson, Clare Bates Congdon. 2016. Measuring the relative importance of different agricultural inputs to global and regional crop yield growth since 1975. *F1000Research* **5**, 2930. [Crossref]
291. Leroi Raputsoane. 2016. Real Effective Exchange Rates Comovements, Common Factors and the South African Currency. *SSRN Electronic Journal* . [Crossref]
292. Omar A. Guerrero, Eduardo Lopez. 2016. Understanding Unemployment in the Era of Big Data: Policy Informed by Data-Driven Theory. *SSRN Electronic Journal* . [Crossref]
293. Kohei Kawamura, Yohei Kobashi, Masato Shizume. 2016. Strategic Central Bank Communication: Discourse and Game-Theoretic Analyses of the Bank of Japan's Monthly Report. *SSRN Electronic Journal* . [Crossref]
294. Georg von Graevenitz, Christian Helmers, Valentine Millot, Oliver Turnbull. 2016. Does Online Search Predict Sales? Evidence from Big Data for Car Markets in Germany and the UK. *SSRN Electronic Journal* . [Crossref]
295. Inna Grinis. 2016. The STEM Requirements of 'Non-STEM' Jobs: Evidence from UK Online Vacancy Postings and Implications for Skills & Knowledge Shortages. *SSRN Electronic Journal* . [Crossref]
296. Serena Ng. 2016. Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data. *SSRN Electronic Journal* . [Crossref]
297. Leif Anders Thorsrud. 2016. Nowcasting Using News Topics. Big Data versus Big Bank. *SSRN Electronic Journal* . [Crossref]
298. Andrew Tiffin. 2016. Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon. *IMF Working Papers* **16**:56, 1. [Crossref]
299. Jurgen A. Doornik, David F. Hendry. 2015. Statistical model selection with ���Big Data���. *Cogent Economics & Finance* **3**:1. . [Crossref]
300. Qing-Ting Zhang, Yuan Liu, Wen Zhou, Zhou-Wang Yang. 2015. A Sequential Regression Model for Big Data with Attributive Explanatory Variables. *Journal of the Operations Research Society of China* **3**:4, 475-488. [Crossref]

301. robert neumann, peter graeff. 2015. quantitative approaches to comparative analyses: data properties and their implications for theory, measurement and modelling. *European Political Science* **14**:4, 385-393. [Crossref]

302. Ben Vinod. 2015. The expanding role of revenue management in the airline industry. *Journal of Revenue and Pricing Management* **14**:6, 391-399. [Crossref]

303. Alan Schwartz, Robert E. Scott. 2015. Third-Party Beneficiaries and Contractual Networks. *Journal of Legal Analysis* **7**:2, 325-361. [Crossref]

304. Julia I. Lane, Jason Owen-Smith, Rebecca F. Rosen, Bruce A. Weinberg. 2015. New linked data on research investments: Scientific workforce, productivity, and public value. *Research Policy* **44**:9, 1659-1671. [Crossref]

305. Max Nathan, Anna Rosso. 2015. Mapping digital businesses with big data: Some early findings from the UK. *Research Policy* **44**:9, 1714-1733. [Crossref]

306. Maryann Feldman, Martin Kenney, Francesco Lissoni. 2015. The New Data Frontier. *Research Policy* **44**:9, 1629-1632. [Crossref]

307. Imanol Arrieta-ibarra, Ignacio N. Lobato. 2015. Testing for Predictability in Financial Returns Using Statistical Learning Procedures. *Journal of Time Series Analysis* **36**:5, 672-686. [Crossref]

308. David H. Autor. 2015. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives* **29**:3, 3-30. [Abstract] [View PDF article] [PDF with links]

309. Yan Chen, Joseph Konstan. 2015. Online field experiments: a selective survey of methods. *Journal of the Economic Science Association* **1**:1, 29-42. [Crossref]

310. David Bholat. 2015. Big Data and central banks. *Big Data & Society* **2**:1, 205395171557946. [Crossref]

311. Levi Boxell. 2015. K-fold Cross-Validation and the Gravity Model of Bilateral Trade. *Atlantic Economic Journal* **43**:2, 289-300. [Crossref]

312. Hallie Eakin, Kirsten Appendini, Stuart Sweeney, Hugo Perales. 2015. Correlates of Maize Land and Livelihood Change Among Maize Farming Households in Mexico. *World Development* **70**, 78-91. [Crossref]

313. Patrick Bajari, Denis Nekipelov, Stephen P. Ryan, Miaoyu Yang. 2015. Machine Learning Methods for Demand Estimation. *American Economic Review* **105**:5, 481-485. [Abstract] [View PDF article] [PDF with links]

314. Hossein Hassani, Emmanuel Sirimal Silva. 2015. Forecasting with Big Data: A Review. *Annals of Data Science* **2**:1, 5-19. [Crossref]

315. Jorge Guzman, Scott Stern. 2015. Where is Silicon Valley?. *Science* **347**:6222, 606-609. [Crossref]

316. Nicole Ludwig, Stefan Feuerriegel, Dirk Neumann. 2015. Putting Big Data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests. *Journal of Decision Systems* **24**:1, 19-36. [Crossref]

317. Thach V. Bui, Thuc D. Nguyen, Noboru Sonehara, Isao Echizen. Tradeoff Between the Price of Distributing a Database and Its Collusion Resistance Based on Concatenated Codes 163-182. [Crossref]

318. Vlad Diaconita. 2015. Processing unstructured documents and social media using Big Data techniques. *Economic Research-Ekonomska Istraživanja* **28**:1, 981-993. [Crossref]

319. Alex Street, Thomas A. Murray, John Blitzer, Rajan S. Patel. 2015. Estimating Voter Registration Deadline Effects with Web Search Data. *Political Analysis* **23**:2, 225-241. [Crossref]

320. Lilli Japec, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, Abe Usher. 2015. Big Data in Survey Research. *Public Opinion Quarterly* **79**:4, 839-880. [Crossref]

321. Kaushik Basu, Andrew Foster. 2015. Development Economics and Method: A Quarter Century of ABCDE. *The World Bank Economic Review* **29**:suppl 1, S2-S8. [Crossref]

322. Alexander Peysakhovich, Jeffrey Naecker. 2015. Machine Learning and Behavioral Economics: Evaluating Models of Choice Under Risk and Ambiguity. *SSRN Electronic Journal* . [Crossref]

323. Johan Perols, Robert M. Bowen, Carsten Zimmermann, Basamba Samba. 2015. Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction. *SSRN Electronic Journal* . [Crossref]

324. Paul Smith. 2015. Predicting UK Unemployment with Internet Search and Survey Data. *SSRN Electronic Journal* . [Crossref]

325. Ananya Sen, Pinar Yildirim. 2015. Clicks and Editorial Decisions: How Does Popularity Shape Online News Coverage?. *SSRN Electronic Journal* . [Crossref]

326. Roberto Casarin, Stefano Grassi, Francesco Ravazzolo, H. K. van Dijk. 2015. Dynamic Predictive Density Combinations for Large Data Sets in Economics and Finance. *SSRN Electronic Journal* . [Crossref]

327. Christian Pierdzioch, Marian Risse, Sebastian Rohloff. 2015. Are Precious Metals a Hedge Against Exchange-Rate Movements? An Empirical Exploration Using Bayesian Additive Regression Trees. *SSRN Electronic Journal* . [Crossref]

328. Anja Lambrecht, Catherine Tucker. 2015. Can Big Data Protect a Firm from Competition?. *SSRN Electronic Journal* . [Crossref]

329. Allison Baker, Timothy Brennan, Jack Erb, Omar Nayeem, Aleksandr Yankelevich. 2014. Economics at the FCC, 2013–2014. *Review of Industrial Organization* **45**:4, 345-378. [Crossref]

330. Sunny L Jardine, Juha V Siikamäki. 2014. A global predictive model of carbon in mangrove soils. *Environmental Research Letters* **9**:10, 104013. [Crossref]

331. Yael Grushka-Cockayne, Victor Richmond R. Jose, Kenneth C. Lichtendahl. 2014. Ensembles of Overfit and Overconfident Forecasts. *SSRN Electronic Journal* . [Crossref]

332. Tadas Bruzikas, Adriaan R. Soetevent. 2014. Detailed Data and Changes in Market Structure: The Move to Unmanned Gasoline Service Stations. *SSRN Electronic Journal* . [Crossref]

333. Sriganesh Lokanathan, Roshanthi Lucas Gunaratne. 2014. Behavioral Insights for Development from Mobile Network Big Data: Enlightening Policy Makers on the State of the Art. *SSRN Electronic Journal* . [Crossref]

334. Greg Distelhorst, Jens Hainmueller, Richard M. Locke. 2013. Does Lean Capability Building Improve Labor Standards? Evidence from the Nike Supply Chain. *SSRN Electronic Journal* . [Crossref]

335. Kaito Yamauchi, Takayuki Morimoto. 2013. Forecasting Financial Market Volatility Using a Dynamic Topic Model. *SSRN Electronic Journal* . [Crossref]

336. Hui Chen, Winston Wei Dou, Leonid Kogan. 2013. Measuring the 'Dark Matter' in Asset Pricing Models. *SSRN Electronic Journal* . [Crossref]

337. José Luis Gómez-Barroso, Juan Ángel Ruiz. Behavioural Targeting in the Mobile Ecosystem 44-57. [Crossref]

338. Marta Vidal, Javier Vidal-García, Rafael Hernandez Barros. Big Data and Business Decision Making 140-157. [Crossref]

339. Javier Vidal-García, Marta Vidal. Big Data Management in Financial Services 217-230. [Crossref]

340. José Luis Gómez-Barroso, Juan Ángel Ruiz. Behavioural Targeting in the Mobile Ecosystem 141-154. [Crossref]

341. Kees Zeelenberg, Barteld Braaksma. Big Data in Official Statistics 274-296. [Crossref]
342. Javier Vidal-García, Marta Vidal, Rafael Hernández Barros. Business Applications of Big Data 104-125. [Crossref]
343. Ruben Xing, Jinluan Ren, Jianghua Sun, Lihua Liu. A Critical Review of the Big-Data Paradigm 75-88. [Crossref]
344. Javier Vidal-García, Marta Vidal, Rafael Hernández Barros. Computational Business Intelligence, Big Data, and Their Role in Business Decisions in the Age of the Internet of Things 1048-1067. [Crossref]
345. Javier Vidal-García, Marta Vidal, Rafael Hernández Barros. Business Applications of Big Data 1346-1367. [Crossref]