

1           Big Data: New Tricks for Econometrics

2                                   Hal R. Varian\*

3                                   June 2013

4                                   Revised: December 26, 2013

5                                   **Abstract**

6           Nowadays computers are in the middle of almost every economic  
7           transaction. These “computer-mediated transactions” generate huge  
8           amounts of data, and new tools are necessary to manipulate and ana-  
9           lyze this data. This essay offers a brief introduction to some of these  
10          tools and methods.

11          Computers are now involved in many economic transactions and can capture  
12          data associated with these transactions, which can then be manipulated and  
13          analyzed using a variety of techniques. Conventional statistical and econo-  
14          metric techniques such as regression often work well but there are some prob-  
15          lems unique to big data sets that require different tools. There are several  
16          reasons for this.

17          First, the sheer size of the data involved may require more sophisticated  
18          data manipulation tools. Second, we may have more potential predictors than  
19          appropriate for estimation, so we need to do some kind of variable selection.  
20          Third, large data sets may allow for more flexible relationships than simple

---

\*Thanks to Jeffrey Oldham, Tom Zhang, Rob On, Pierre Grinspan, Jerry Friedman, Art Owen, Steve Scott, Bo Cowgill, Brock Noland, Daniel Stonehill, and Gary King for comments on earlier versions of this paper.

21 linear models. Machine learning techniques such as decision trees, support  
22 vector machines, neural nets, deep learning and so on may allow for more  
23 effective ways to model complex relationships.

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

## 32 1 Tools for data manipulation

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

41 However, if you have several gigabytes of data or several million observa-  
42 tions, standard relational databases become unwieldy. Databases to manage  
43 data of this size are generically known as “NoSQL” databases. The term is  
44 used rather loosely, but is sometimes interpreted as meaning “not only SQL.”  
45 NoSQL databases are more primitive than SQL databases in terms of data  
46 manipulation capabilities but can handle larger amounts of data.

47 Due to the rise of computer mediated transactions, many companies have  
48 found it necessary to develop systems to process billions of transactions per

49 day. For example, according to Sullivan [2012], Google has seen 30 trillion  
50 URLs, crawls over 20 billion of those a day, and answers 100 billion search  
51 queries a month. Analyzing even one day’s worth of data of this size is  
52 virtually impossible with conventional databases. The challenge of dealing  
53 with data sets of this size led to the development of several tools to manage  
54 and analyze big data.

55 These tools are proprietary to Google, but have been described in aca-  
56 demic publications in sufficient detail that open-source implementations have  
57 been developed. The list below has both the Google name and the name of  
58 related open source tools. Further details can be found in the Wikipedia  
59 entries associated with the tool names.

60 **Google File System** [Hadoop Distributed File System] This system sup-  
61 ports files of to be distributed across hundreds or even thousands of  
62 computers.

63 **Bigtable** [Cassandra] This is a table of data that lives in the Google File  
64 System. It too can stretch over many computers.

65 **MapReduce** [Hadoop] This is a system for accessing manipulating data  
66 in large data structures such as Bigtables. MapReduce allows you to  
67 access the data in parallel, using hundreds or thousands of machines  
68 to do the particular data extraction you are interested in. The query  
69 is “mapped” to the machines and is then applied in parallel to dif-  
70 ferent shards of the data. The partial calculations are then combined  
71 (“reduced”) to create the summary table you are interested in.

72 **Go** [Pig] Go is an open-source general-purpose computer language that makes  
73 it easier to do parallel data processing.

74 **Dremel** [Hive, Drill, Impala] This is a tool that allows data queries to be  
75 written in a simplified form of SQL. With Dremel it is possible to run  
76 an SQL query on a petabyte of data (1000 terabytes) in a few seconds.

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

## 83   **2   Tools for data analysis**

84   The outcome of the big data processing described above is often a “small”  
85 table of data that may be directly human readable or can be loaded into an  
86 SQL database, a statistics package, or a spreadsheet.

87   If the extracted data is still inconveniently large, it is often possible to  
88 select a subsample for statistical analysis. **At Google, for example, I have**  
89 **found that random samples on the order of 0.1 percent work fine for analysis**  
90 **of economic data.**

91   **Once a dataset has been extracted it is often necessary to do some ex-**  
92 **ploratory data analysis along with consistency and data-cleaning tasks. This**  
93 **is something of an art which can be learned only by practice, but there are**  
94 **data cleaning software tools such as OpenRefine and DataWrangler that can**  
95 **be used to assist in this task.**

96   **Data analysis in statistics and econometrics can be broken down into four**  
97 **categories: 1) prediction, 2) summarization, 3) estimation, and 4) hypothesis**  
98 **testing. Machine learning is concerned primarily with prediction; the closely**  
99 **related field of data mining is also concerned with summarization.** Econo-  
100 metricians, statisticians, and data mining specialists are generally looking  
101 for insights that can be extracted from the data. Machine learning special-  
102 ists are often primarily concerned with developing computers systems that  
103 can provide useful predictions and perform well in the presence of challeng-  
104 ing computational constraints. Data science, a somewhat newer term, is

105 concerned with both prediction and summarization, but also with data ma-  
106 nipulation, visualization, and other similar tasks. The terminology is not  
107 standardized in these areas, so these statements reflect general usage, not  
108 hard-and-fast definitions. Other terms used computer assisted data analysis  
109 include knowledge extraction, information discovery, information harvesting,  
110 data archeology, data pattern processing, and exploratory data analysis.

111 Much of applied econometrics is concerned with detecting and summariz-  
112 ing relationships in the data. The most common tool used to for summariza-  
113 tion is (linear) regression analysis. As we shall see, machine learning offers  
114 a set of tools that can usefully summarize more complex relationships in the  
115 data. We will focus on these regression-like tools since those are the most  
116 natural for economic applications.

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

121 In machine learning, the  $x$ -variables are usually called “predictors” or  
122 “features.” The focus of machine learning is to find some function that pro-  
123 vides a good prediction of  $y$  as a function of  $x$ . Historically, most work in  
124 machine learning has involved cross-section data where it is natural to think  
125 of the data being IID or at least independently distributed. The data may  
126 be “fat,” which means lots of predictors relative to the number of observa-  
127 tions, or “tall” which means lots of observations relative to the number of  
128 predictors.

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

135 When confronted with a prediction problem of this sort an economist  
136 would think immediately of a linear or logistic regression. However, there  
137 may be better choices, particularly if a lot of data is available. These in-  
138 clude nonlinear methods such as 1) neural nets, 2) support vector machines,  
139 3) classification and regression trees, 4) random forests, and 5) penalized  
140 regression such as lasso, lars, and elastic nets.

141 I will focus on the last three methods in the list above, since they seem  
142 to work well on the type of data economists generally use. Neural nets and  
143 support vector machines work well for many sorts of prediction problems, but  
144 they are something of a black box. By contrast it is easy to understand the  
145 relationships that trees and penalized regressions describe. Much more detail  
146 about these methods can be found in machine learning texts; an excellent  
147 treatment is available in Hastie et al. [2009], which can be freely downloaded.  
148 Other suggestions for further reading are given at the end of this article.

### 149 **3 General considerations for prediction**

150 Our goal with prediction is typically to get good *out-of-sample predictions*.  
151 Most of us know from experience that it is all too easy to construct a predictor  
152 that works well in-sample, but fails miserably out-of-sample. To take a trivial  
153 example,  $n$  linearly independent regressors will fit  $n$  observations perfectly  
154 but will usually have poor out-of-sample performance. Machine learning  
155 specialists refer to this phenomenon as the “overfitting problem.”

156 There are three major techniques for dealing with the overfitting problem  
157 which are commonly used in machine learning.

158 First, since simpler models tend to work better for out of sample forecasts,  
159 machine learning experts have come up with various ways penalize models for  
160 excessive complexity. In the machine learning world, this is known as “reg-  
161 ularization” and we will encounter a some examples later one. Economists  
162 tend to prefer simpler models for the same reason, but have not been as

163 explicit about quantifying complexity costs.

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

169 Third, in the training stage, it may be necessary to estimate some “tuning  
170 parameters” of the model. The conventional way to do this in machine  
171 learning is to use *k-fold cross validation*.

- 172 1. Divide the data into  $k$  roughly equal subsets and label them by  $s =$   
173  $1, \dots, k$ . Start with subset  $s = 1$ .
- 174 2. Pick a value for the tuning parameter.
- 175 3. Fit your model using the  $k - 1$  subsets other than subset  $s$ .
- 176 4. Predict for subset  $s$  and measure the associated loss.
- 177 5. Stop if  $s = k$ , otherwise increment  $s$  by 1 and go to step 2.

178 Common choices for  $k$  are 10, 5, and the sample size minus 1 (“leave  
179 one out”). After cross validation, you end up with  $k$  values of the tuning  
180 parameter and the associated loss which you can then examine to choose  
181 an appropriate value for the tuning parameter. Even if there is no tuning  
182 parameter, it is useful to use cross validation to report goodness-of-fit mea-  
183 sures since it measures out-of-sample performance which is what is typically  
184 of interest.

185 Test-train and cross validation, are very commonly used in machine learn-  
186 ing and, in my view, should be used much more in economics, particularly  
187 when working with large datasets. For many years, economists have re-  
188 ported in-sample goodness-of-fit measures using the excuse that we had small  
189 datasets. But now that larger datasets have become available, there is no

190 reason not to use separate training and testing sets. Cross-validation also  
191 turns out to be a very useful technique, particularly when working with rea-  
192 sonably large data. It is also a much more realistic measure of prediction  
193 performance than measures commonly used in economics.

## 194 4 Classification and regression trees

195 Let us start by considering a discrete variable regression where our goal is to  
196 predict a 0-1 outcome based on some set of features (what economists would  
197 call explanatory variables or predictors.) In machine learning this is known  
198 as a *classification problem*. Economists would typically use a generalized  
199 linear model like a logit or probit for a classification problem.

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

206 Ironically, one of the earliest papers on the automatic construction of de-  
207 cision trees was co-authored by an economist (Morgan and Sonquist [1963]).  
208 However, the technique did not really gain much traction until 20 years later  
209 in the work of Breiman et al. [1984] and his colleagues. Nowadays this predic-  
210 tion technique is known as “classification and regression trees”, or “CART.”

211 Consider the simple example shown in Figure 1, where we are trying to  
212 predict survivors of the Titanic using just two variables, age and which class  
213 of travel the passenger purchased.

214 Here is a set of rules that can be read off of this tree (more of a bush,  
215 really):

- 216 • class 3: predict died (370 out of 501)
- 217 • class 1 or 2 and younger than 16: predict lived (34 out of 36)



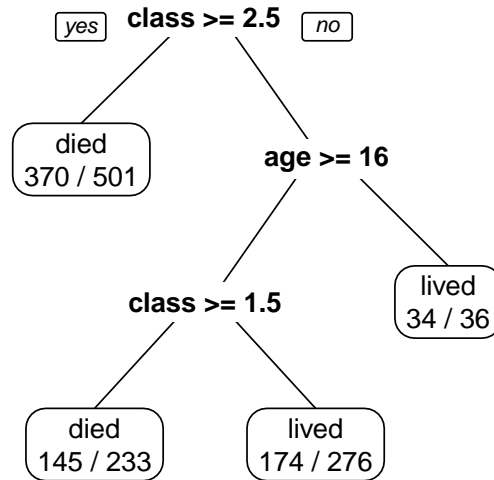


Figure 1: A classification tree for survivors of the Titanic. See text for interpretation.

- 218 • class 2 or 3 and older than 16: predict died (145 out of 233)
- 219 • class 1, older than 16: predict lived: (174 out of 276)

220 The rules fit the data reasonably well, misclassifying about 30% of the  
 221 observations in the testing set.

222 This classification can also be depicted in the “partition plot” shown in  
 223 Figure 2 which shows how the tree divides up the space of (age, class) pairs.  
 224 Of course, the partition plot can only be used for 2 variables while a tree  
 225 representation can handle an arbitrarily large number.

226 It turns out that there are computationally efficient ways to construct  
 227 classification trees of this sort. These methods generally are restricted to  
 228 binary trees (two branches at each node). They can be used for classifi-  
 229 cation with multiple outcomes (“classification trees”) , or with continuous  
 230 dependent variables (“regression trees.”)

231 Trees tend to work well for problems where there are important nonlin-

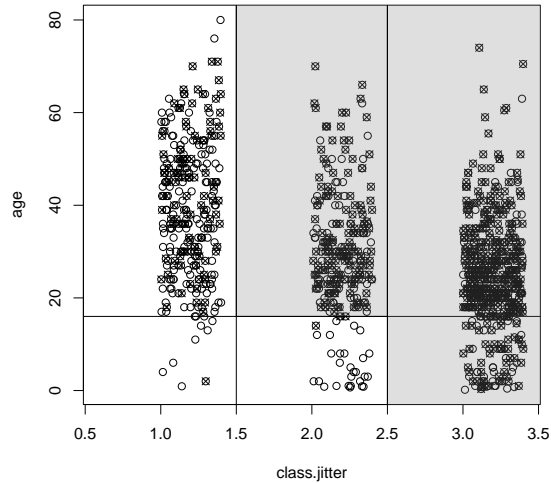


Figure 2: The simple tree model predicts death in shaded region. White circles indicate survival, black crosses indicate death.

232 earities and interactions. As an example, let us continue with the Titanic  
 233 data and create a tree that relates survival to age. In this case, the rule  
 234 generated by the tree is very simple: predict “survive” if age < 8.5 years.  
 235 We can examine the same data with a logistic regression to estimate the  
 236 probability of survival as a function of age:

	Estimate	Std. Error	t value	Pr(> t )
237 (Intercept)	0.464813	0.034973	13.291	<2e-16 ***
238 age	-0.001894	0.001054	-1.796	0.0727 .

240 The tree model suggests that age is an important predictor of survival impor-  
 241 tant, while the logistic model says it is barely important. This discrepancy is  
 242 explained in Figure 3 where we plot survival rates by bins. Here we see that  
 243 survival rates for those under 10 years old were elevated compared to older  
 244 passengers, except for the very oldest group. So what mattered for survival  
 245 is not so much age, but whether the passenger was a child or a senior. It

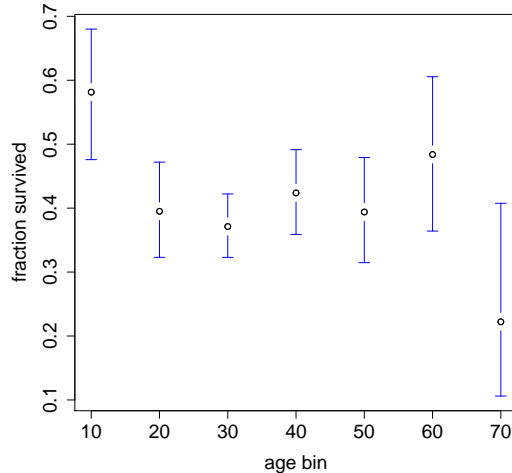


Figure 3: The figure shows the fraction of the population that survived for different age groups (0-10,10-20, and so on). The error bars are computed using the Wilson method.

246 would be difficult to discover this fact from a logistic regression alone.<sup>1</sup>

247 Trees also handle missing data well. Perlich et al. [2003] examined several  
 248 standard data sets and found that “logistic regression is better for smaller  
 249 data sets and tree induction for larger data sets.” Interestingly enough, trees  
 250 tend *not* to work very well if the underlying relationship really is linear,  
 251 but there are hybrid models such as RuleFit (Friedman and Popescu [2005])  
 252 which can incorporate both tree and linear relationships among variables.

253 **However, even if trees may not improve on predictive accuracy compared**  
 254 **to linear models, the age example shows that they may reveal aspects of the**  
 255 **data that are not apparent from a traditional linear modeling approach.**

<sup>1</sup>It is true that if you knew that there was a nonlinearity in age, you use age dummies in the logit model to capture this effect. However the tree formulation made this nonlinearity quite apparent.

## 256 4.1 Pruning trees

257 One problem with trees is that they tend to overfit the data. The most  
258 widely-used solution to this problem is to “prune” the tree by imposing some  
259 complexity cost for having too many branches. This penalty for complexity  
260 is a form of regularization, which was mentioned earlier.

261 So, a typical tree estimation session might involve dividing your data  
262 into 10 folds, using 9 of the folds to grow a tree with a particular complexity,  
263 and then predict on the excluded fold. Repeat the estimation with different  
264 values of the complexity parameter using other folds and choose the value  
265 of the complexity parameter that minimizes the out-of-sample classification  
266 error. (Some researchers recommend being a bit more aggressive than that  
267 and choosing the complexity parameter that is one standard deviation lower  
268 than the loss-minimizing value.)

269 Of course, in practice, the computer program handles most of these details  
270 for you. In the examples in this paper I mostly use default choices, but in  
271 practices these default will often be tuned. As with any other statistical  
272 procedure, skill, experience and intuition are helpful in coming up with a  
273 good answer and diagnostics, exploration, and experimentation are just as  
274 useful with these methods as with regression techniques.

275 There are many other approaches to creating trees, including some that  
276 are explicitly statistical in nature. For example, a “conditional inference  
277 tree,” or ctree for short, chooses the structure of the tree using a sequence  
278 of hypothesis tests. The resulting trees tend to need very little pruning.  
279 (Hothorn et al. [2006]) An example for the Titanic data is shown in Figure 4.

280 One might summarize this tree by the following principle: “women and  
281 children first . . . particularly if they were traveling first class.” This simple  
282 example again illustrates that classification trees can be helpful in summa-  
283 rizing relationships in data, as well as predicting outcomes.

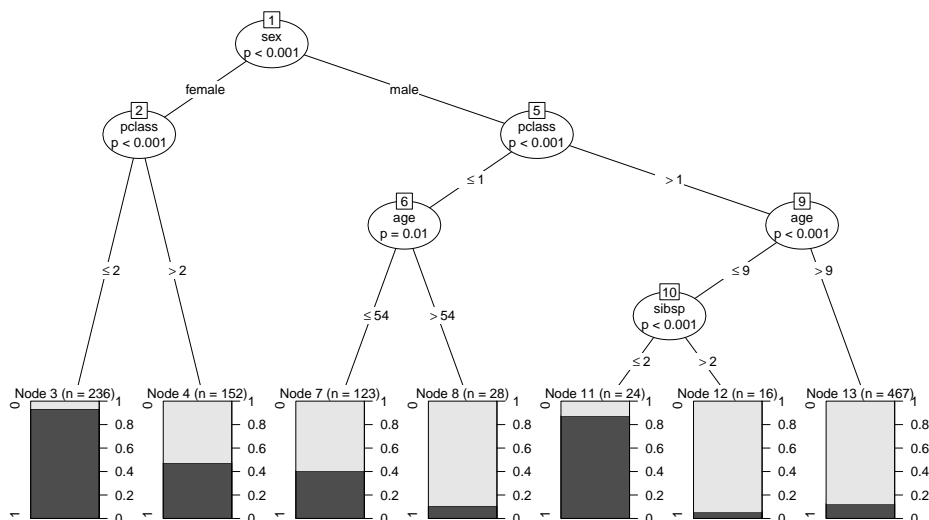


Figure 4: A ctree for survivors of the Titanic. The black bars indicate fraction of the group that survival.

## 284 4.2 Economic example: HMDA data

285 Munnell et al. [1996] examined mortgage lending in Boston to see if race  
 286 played a significant role in determining who was approved for a mortgage.  
 287 The primary econometric technique was a logistic regression where race was  
 288 included as one of the predictors. The race effect indicated a statistically  
 289 significant negative impact on probability of getting a mortgage for black  
 290 applicants. This finding prompted lively subsequent debate and discussion,  
 291 with 725 citations on Google Scholar as of July 2013.

292 Here I examine this question using the tree-based estimators described in  
 293 the previous section. The data consists of 2380 observations of 12 predictors,  
 294 one of which was race. Figure 5 shows a conditional tree estimated using the  
 295 R package `party`. (For reasons of space, I have omitted variable descriptions  
 296 which are readily available on the web site.)

297 The tree fits pretty well, misclassifying 228 of the 2380 observations for an  
 298 error rate of 9.6%. By comparison, a simple logistic regression does slightly

299 better, misclassifying 225 of the 2380 observations, leading to an error rate  
 300 of 9.5%. As you can see in Figure 5, the most important variable is `dmi`  
 301 = “denied mortgage insurance”. This variable alone explains much of the  
 302 variation in the data. The race variable (`black`) shows up far down the tree  
 303 and seems to be relatively unimportant.

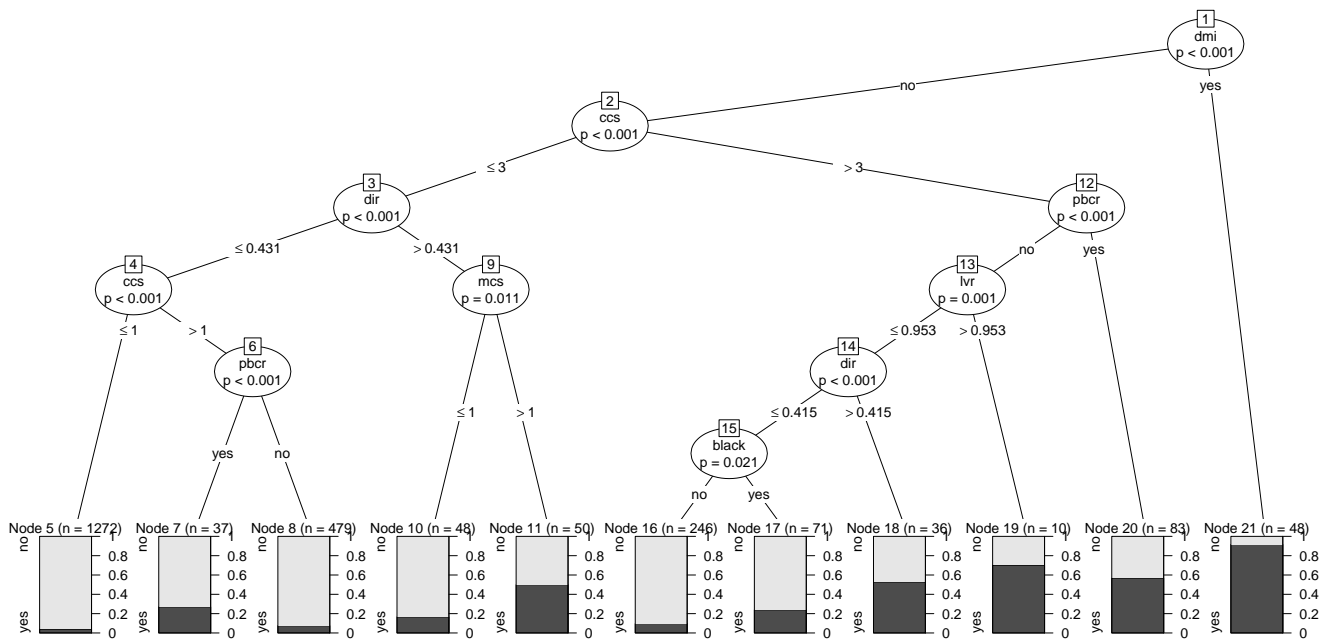


Figure 5: HMDA tree. The black bars indicate the fraction of each group that were denied mortgages. The most important determinant of this is the variable `dmi`, “denied mortgage insurance.”

304 One way to gauge whether a variable is important is to exclude it from  
 305 the prediction and see what happens. When this is done, it turns out that  
 306 the accuracy of the tree based model doesn’t change at all: exactly the same  
 307 cases are misclassified. So there is a plausible decision tree model that ignores  
 308 race that fits the observed data just as well as a model that includes race.

## 309 5 Boosting, bagging and bootstrap

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

314 **Bootstrap** involves choosing (with replacement) a sample of size  $n$  from a  
315 data set of size  $n$  to estimate the sampling distribution of some statistic.  
316 A variation is the “ $m$  out of  $n$  bootstrap” which draws a sample of size  
317  $m$  from a dataset of size  $n > m$ .

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

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

323 Econometricians are well-acquainted with the bootstrap rarely use the  
324 other two methods. Bagging is primarily useful for nonlinear models such  
325 as trees. (Friedman and Hall [2005].) Boosting tend to improve predictive  
326 performance of an estimator significantly and can be used for pretty much  
327 any kind of classifier or regression model, including logits, probits, trees, and  
328 so on.

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

332 **Random forests** refers to a technique that uses multiple trees. A typical  
333 procedure uses the following steps.

- 334 1. Choose a bootstrap sample of the observations and start to grow  
335 a tree.

- 336           2. At each node of the tree, choose a random sample of the predictors  
337           to make the next decision. Do not prune the trees.
- 338           3. Repeat this process many times to grow a forest of trees
- 339           4. The final classification is then determined by majority vote among  
340           all the trees in the forest

341       This method produces surprisingly good out-of-sample fits, particularly  
342 with highly nonlinear data. In fact, Howard [2013] claims “ensembles of  
343 decision trees (often known as Random Forests) have been the most successful  
344 general-purpose algorithm in modern times.” He goes on to indicate that  
345 “the algorithm is very simple to understand, and is fast and easy to apply.”  
346 See also Caruana and Niculescu-Mizil [2006] who compare several different  
347 machine learning algorithms and find that ensembles of trees perform quite  
348 well. There are a number variations and extensions of the basic “ensemble of  
349 trees” model such as Friedman’s “Stochastic Gradient Boosting” (Friedman  
350 [1999]).

351       One defect of random forests is that they are a bit of a black box—  
352 they don’t offer simple summaries of the data. However, they can be used  
353 to determine which variables are “important” in predictions in the sense of  
354 contributing the biggest improvements in prediction accuracy.

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

358       I ran the random forest method on the HMDA data and found that it  
359 misclassified 223 of the 2380 cases, a small improvement over the logit and  
360 the ctree. I also used the importance option in random forests to see how  
361 the predictors compared. It turned out that `dmi` was the most important  
362 predictor and race was second from the bottom which is consistent with the  
363 ctree analysis.



## 364 6 Variable selection

365 Let us return to the familiar world of linear regression and consider the prob-  
366 lem of variable selection. There are many such methods available, including  
367 stepwise regression, principal component regression, partial least squares,  
368 AIC and BIC complexity measures and so on. Castle et al. [2009] describes  
369 and compares 21 different methods.

### 370 6.1 Lasso and friends

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

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

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

376 This estimation method is called *elastic net regression*; it contains three other  
377 methods as special cases. If there is no penalty term ( $\lambda = 0$ ), this is *ordinary*  
378 *least squares*. If  $\alpha = 1$  so that there is only the quadratic constraint, this  
379 is *ridge regression*. If  $\alpha = 0$  this is called the *lasso*, an acronym for “least  
380 absolute shrinkage and selection operator.”

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

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

## 390 **6.2 Spike and slab regression**

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

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

404 Now we take a draw of  $\gamma$  from its prior distribution, which will just  
405 be a list of variables in the regression. Conditional on this list of included  
406 variables, we take a draw from the prior distribution for the coefficients. We  
407 combine these two draws with the likelihood in the usual way which gives us  
408 a draw from posterior distribution on both  $\gamma$  and the coefficients. We repeat  
409 this process thousands of times using a Markov Chain Monte Carlo (MCMC)  
410 technique which give us a table summarizing the posterior distribution for  $\gamma$   
411 and the coefficients and the associated prediction of  $y$ .

412 We end up with a table of thousands of draws from the posterior distri-  
413 butions of  $\gamma$ ,  $\beta$ , and  $y$  which we can summarize in a variety of ways. For  
414 example, we can compute the average value of  $\gamma_p$  which shows the posterior  
415 probability variable  $p$  is included in the regressions.

predictor	BMA	CDF(0)	lasso	spike-slab
GDP level 1960	1.000	1.000	-	0.9992
Fraction Confucian	0.995	1.000	6	0.9730
Life expectancy	0.946	0.942	5	0.9610
Equipment investment	0.757	0.997	1	0.9532
Sub-Saharan dummy	0.656	1.000	-	0.5834
Fraction Muslim	0.656	1.000	-	0.6590
Rule of law	0.516	1.000	-	0.4532
Open economy	0.502	1.000	3	0.5736
Degree of Capitalism	0.471	0.987	-	0.4230
Fraction Protestant	0.461	0.966	-	0.3798

Table 1: Comparing variable selection algorithms. See text for discussion.

### 416 6.3 Economic example: growth regressions

417 We illustrate the lasso and spike and slab regression with an example from  
418 Sala-i-Martin [1997]. This involves examining a multi-country set of pre-  
419 dictors of economic growth in order to see which variables appeared to be  
420 the most important. Sala-i-Martin [1997] looked at all possible subsets of  
421 regressors of manageable size. Ley and Steel [2009] investigated the same  
422 question using Bayesian techniques related to, but not identical with, spike-  
423 and-slab, while Hendry and Krolzig [2004] examined an iterative significance  
424 test selection method.

425 Table 1 shows 10 predictors that were chosen by Sala-i-Martin [1997], Ley  
426 and Steel [2009], `lasso`, and `spike-and-slab`. The table is based on that  
427 in Ley and Steel [2009] but metrics used are not strictly comparable across  
428 models. The “BMA” and “spike-slab” columns are posterior probabilities of  
429 inclusion; the “lasso” column is just the ordinal importance of the variable  
430 with a dash indicating that it was not included in the chosen model; and the  
431 CDF(0) measure is defined in Sala-i-Martin [1997].

432 The `lasso` and the Bayesian techniques are very computationally efficient  
433 and on this ground would likely be preferred to exhaustive search. **All 4**  
434 **of these variable selection methods give similar results for the first 4 or 5**

435 variables, after which they diverge. In this particular case, the data set  
436 appears to be too small to resolve the question of what is “important” for  
437 economic growth.

## 438 7 Time series

439 The machine learning techniques described up until now are generally applied  
440 to cross-sectional data where independently distributed data is a plausible  
441 assumption. However, there are also techniques that work with time series.  
442 Here we describe an estimation method which we call Bayesian Structural  
443 Time Series (BSTS) that seems to work well for variable selection problems  
444 in time series applications.

445 Our research in this area was motivated by Google Trends data which  
446 provides an index of the volume of Google queries on specific terms. One  
447 might expect that queries on [file for unemployment] might be predictive  
448 of the actual rate of filings for initial claims, or that queries on [Orlando  
449 vacation] might be predictive of actual visits to Orlando. Indeed, Choi and  
450 Varian [2009, 2012], Goel et al. [2010], Carrière-Swallow and Labbé [2011],  
451 McLaren and Shanbhoge [2011], Arola and Galan [2012], Hellerstein and  
452 Middeldorp [2012] and many others have shown that Google queries do have  
453 significant short-term predictive power for various economic metrics.

454 The challenge is that there are billions of queries so it is hard to determine  
455 exactly which queries are the most predictive for a particular purpose. Google  
456 Trends classifies the queries into categories, which helps a little, but even then  
457 we have hundreds of categories as possible predictors so that overfitting and  
458 spurious correlation are a serious concern. BSTS is designed to address these  
459 issues. We offer a very brief description here; more details are available in  
460 Scott and Varian [2012a,b].

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

463     •  $y_t = \mu + bt + \beta x_t + e_t$ .

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

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

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

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

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

470 It is easy to add an additional state variable for seasonality if that is ap-  
471 propriate. The parameters to estimate are the regression coefficients  $\beta$  and  
472 the variances of  $(e_{it})$  for  $i = 1, \dots, 3$ . We can then use these estimates to  
473 construct the optimal Kalman forecast.

474 For the regression we use the spike-and-slab variable choice mechanism  
475 described above. A draw from the posterior distribution now involves a draw  
476 of variances of  $(e_{1t}, e_{2t})$ , a draw of the vector  $\gamma$  that indicates which vari-  
477 ables are in the regression, and a draw of the regression coefficients  $\beta$  for  
478 the included variables. The draws of  $\mu_t$ ,  $b_t$ , and  $\beta$  can be used to construct  
479 estimates of  $y_t$  and forecasts for  $y_{t+1}$ . We end up with an (estimated) pos-  
480 terior distribution for the metric of interest. If we seek a point prediction,  
481 we could average over these draws, which is essentially a form of Bayesian  
482 model averaging.

483     As an example, consider the non-seasonally adjusted data for new homes  
484 sold in the U.S. (HSN1FNSA) from the St. Louis Federal Reserve Economic  
485 Data. This time series can be submitted to Google Correlate, which then  
486 returns the 100 queries that are the most highly correlated with the series.  
487 We feed that data into the BSTS system which identifies the predictors with  
488 the largest posterior probabilities of appearing in the housing regression are  
489 shown in Figure 6. Two predictors, [oldies lyrics] and [www.mail2web] ap-  
490 pear to be spurious so we remove them and re-estimate, yielding the results

491 in Figure 7. The fit is shown in Figure 8 which shows the incremental con-  
 492 tribution of the trend, seasonal, and individual regressors components.

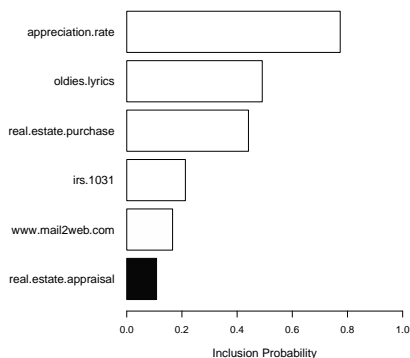


Figure 6: Initial predictors.

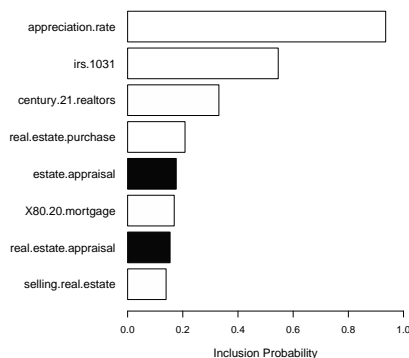


Figure 7: Final predictors.

## 493 8 Econometrics and machine learning

494 There are a number of areas where there would be opportunities for fruitful  
 495 collaboration between econometrics and machine learning. I mentioned above  
 496 that most machine learning uses IID data. However, the BSTS model shows  
 497 that some of these techniques can be adopted for time series models. It is  
 498 also possible to use machine learning techniques to look at panel data and  
 499 there has been some work in this direction.

500 Econometricians have developed several tools for causal modeling such  
 501 as instrumental variables, regression discontinuity, and various forms of ex-  
 502 periments. (Angrist and Krueger [2001].) Machine learning work has, for  
 503 the most part, dealt with pure prediction. In a way this is ironic, since the-  
 504 oretical computer scientists, such as Pearl [2009a,b] have made significant  
 505 contributions to causal modeling. However, it appears that these theoretical  
 506 advances have not as yet been incorporated into machine learning practice  
 507 to a significant degree.

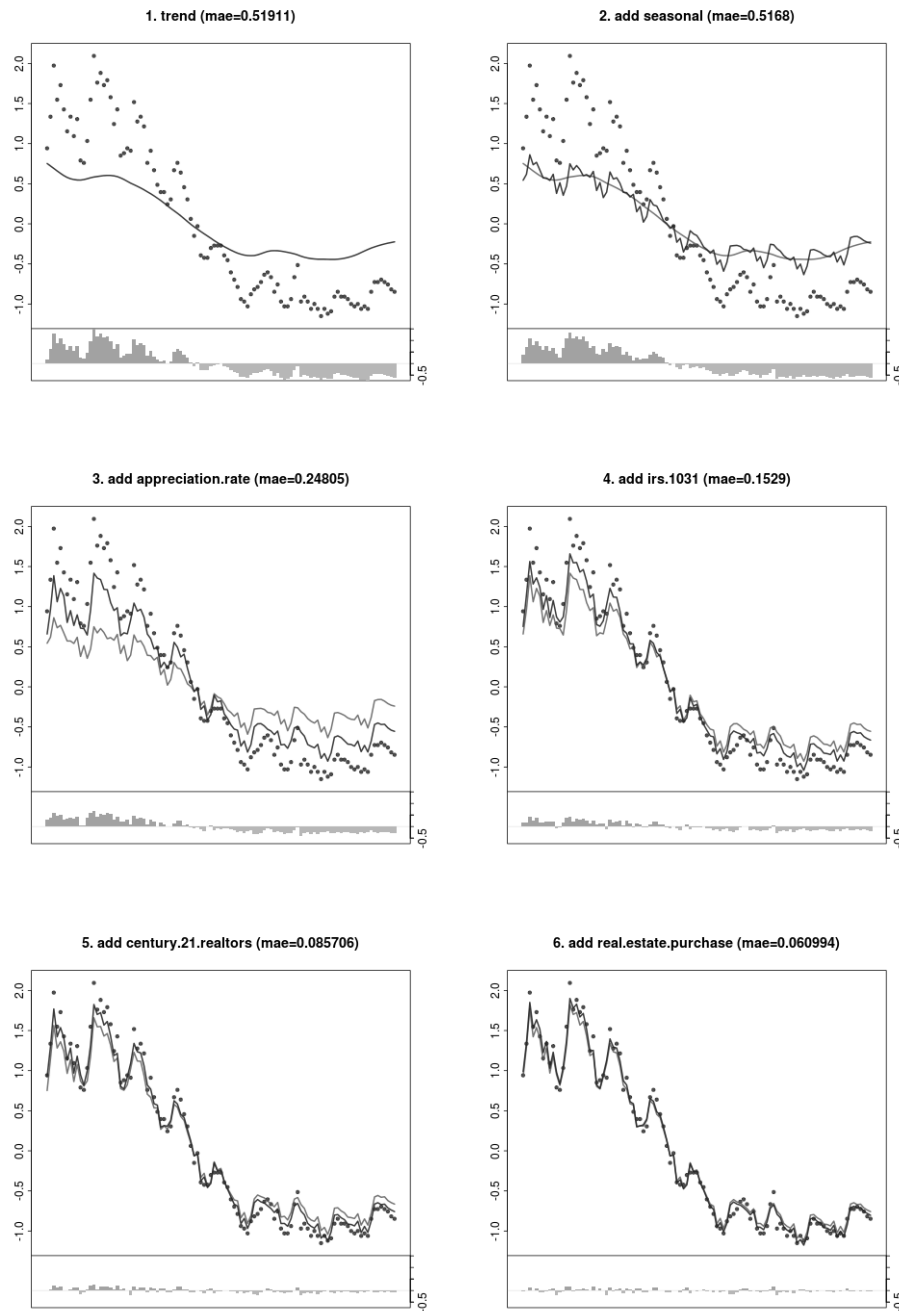


Figure 8: Incremental plots. The plots show the impact of the trend, seasonal, and a few individual regressors. The residuals are shown on the bottom.

## 508 8.1 Causality and prediction

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

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

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

- 522 • Which precincts will receive additional police in the experiment and  
523 policy implementation and how will this be determined? Possible as-  
524 signment rules could be 1) random, 2) based on perceived need, 3)  
525 based on cost of providing service, 4) based on resident requests, 5)  
526 based on a formula or set of rules, 6) based on asking for volunteers,  
527 and so on. Ideally the assignment procedure in the experiment will be  
528 similar to that used in the policy. A good model for predicting which  
529 precincts will receive additional police under the proposed policy can  
530 clearly be helpful in estimating the impact of the policy.
- 531 • What will be the impact of these additional police in both the exper-  
532 iment and the policy? As Rubin [1974] and many subsequent authors  
533 have emphasized, when we consider the causal impact of some treat-  
534 ment we need to compare the outcome with the intervention to what  
535 *would have happened* without the intervention. But this counterfactual  
536 cannot be observed, so it must be predicted by some model. The better



537 predictive model you have for the counterfactual, the better you will be  
538 able to estimate the causal effect, an observation that is true for both  
539 pure experiments and natural experiments.

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

543 To state this in a slightly more formal way, consider the identity from  
544 Angrist and Pischke [2008], page 11:

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

545 If you want to model the average treatment effect as a function of other  
546 variables, you will usually need to model both the observed difference and  
547 the selection bias. The better your predictive model for those components,  
548 the better predictions you can make about the average treatment effect. Of  
549 course, if you have a true randomized treatment-control experiment, selection  
550 bias goes away and those treated are an unbiased random sample of the  
551 population.

552 To illustrate these points, let us consider the thorny problem of estimat-  
553 ing the causal effect of advertising on sales. (Lewis and Rao [2013].) The  
554 difficulty is that there are many confounding variables, such as seasonality or  
555 weather, that cause both increased ad exposures and increased purchases by  
556 consumers. Consider the (probably apocryphal) story about an advertising  
557 manager who was asked why he thought his ads were effective. “Look at this  
558 chart,” he said. “Every December I increase my ad spend and, sure enough,  
559 purchases go up.” Of course, seasonality can be observed and included in  
560 the model. However, generally there will be other confounding variables that  
561 affect both exposure to ads and the propensity of purchase, which makes  
562 causal interpretations of relationships problematic.

563 The ideal way to estimate advertising effectiveness is, of course, to run a  
564 controlled experiment. In this case the control group provides an estimate of  
565 what would have happened without ad exposures. But this ideal approach  
566 can be quite expensive, so it is worth looking for alternative ways to predict  
567 the counterfactual. One way to do this is to use the Bayesian Structural Time  
568 Series method described earlier. In this case, a model based on historical time  
569 series data can, in some cases, be used to estimate what *would have happened*  
570 in the absence of the advertising intervention. See Brodersen et al. [2013] for  
571 an example of this approach.

## 572 9 Model uncertainty

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

576 In 2006, Netflix offered a million dollar prize to researchers who could  
577 provide the largest improvement to their existing movie recommendation  
578 system. The winning submission involved a “complex blending of no fewer  
579 than 800 models” though they also point out that “predictions of good quality  
580 can usually be obtained by combining a small number of judiciously chosen  
581 methods.” (Feuerverger et al. [2012].) It also turned out that a blend of the  
582 best and second-best model outperformed both of them.

583 Ironically, it was recognized many years ago that averages of macroeco-  
584 nomic model forecasts outperformed individual models, but somehow this  
585 idea was rarely exploited in traditional econometrics. The exception is the  
586 literature on Bayesian model averaging which has seen a steady flow of work;  
587 see Steel [2011] for a survey.

588 However, I think that model uncertainty has crept in to applied econo-  
589 metrics through the back door. Many papers in applied econometrics present  
590 regression results in a table with several different specifications: which vari-

591 ables are included in the controls, which variables are used as instruments,  
592 and so on. The goal is usually to show that the estimate of some interesting  
593 parameter is not very sensitive to the exact specification used.

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

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

## 604 10 Summary and further reading

605 Since computers are now involved in many economic transactions, big data  
606 will only get bigger. Data manipulation tools and techniques developed for  
607 small datasets will become increasingly inadequate to deal with new prob-  
608 lems. Researchers in machine learning have developed ways to deal with  
609 large data sets and economists interested in dealing with such data would be  
610 well advised to invest in learning these techniques.

611 I have already mentioned Hastie et al. [2009] which has detailed descrip-  
612 tions of all the methods discussed here but at a relatively advanced level.  
613 James et al. [2013] describes many of the same topics at an undergraduate-  
614 level, along with R code and many examples.<sup>2</sup>

615 Venables and Ripley [2002] contains good discussions of these topics with  
616 emphasis on applied examples. Leek [2013] presents a number of YouTube

---

<sup>2</sup>There are several economic examples in the book where the tension between predictive modeling and causal modeling is apparent.

617 videos with gentle and accessible introductions to several tools of data anal-  
618 ysis. Howe [2013] provides a somewhat more advanced introduction to data  
619 science that also includes discussions of SQL and NoSQL databases. Wu  
620 and Kumar [2009] gives detailed descriptions and examples of the major al-  
621 gorithms in data mining, while Williams [2011] provides a unified toolkit.  
622 Domingos [2012] summarizes some important lessons which include “pitfalls  
623 to avoid, important issues to focus on and answers to common questions.”

## 624 **References**

625 Joshua D. Angrist and Alan B. Krueger. Instrumental variables and the  
626 search for identification: From supply and demand to natural experiments.  
627 *Journal of Economic Perspectives*, 15(4):69–85, 2001. URL [http://www.  
628 aeaweb.org/articles.php?doi=10.1257/jep.15.4.69](http://www.aeaweb.org/articles.php?doi=10.1257/jep.15.4.69).

629 Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics*.  
630 Princeton University Press, 2008.

631 Concha Arola and Enrique Galan. Tracking the future on the web: Con-  
632 struction of leading indicators using internet searches. Technical report,  
633 Bank of Spain, 2012. URL [http://www.bde.es/webbde/SES/Secciones/  
634 Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/  
635 Fich/do1203e.pdf](http://www.bde.es/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf).

636 L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification  
637 and Regression Trees*. Wadsworth and Brooks/Cole, Monterey, 1984.

638 Kay H. Brodersen, Nicolas Remy, Fabian Gallusser, Steven L. Scott, Jim  
639 Koehler, and Penny Chu. Inferring causal impact using Bayesian structural  
640 time series models. Technical report, Google, Inc., 2013. URL [http:  
641 //research.google.com/pubs/pub41854.html](http://research.google.com/pubs/pub41854.html).

- 642 Yan Carrière-Swallow and Felipe Labbé. Nowcasting with Google Trends in  
643 an emerging market. *Journal of Forecasting*, 2011. doi: 10.1002/for.1252.  
644 URL <http://ideas.repec.org/p/chb/bcchwp/588.html>. Working Pa-  
645 pers Central Bank of Chile 588.
- 646 Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of  
647 supervised learning algorithms. In *Proceedings of the 23rd International*  
648 *Conference on Machine Learning*, Pittsburgh, PA, 2006.
- 649 Jennifer L. Castle, Xiaochuan Qin, and W. Robert Reed. How to pick the  
650 best regression equation: A review and comparison of model selection algo-  
651 rithms. Technical Report 13/2009, Department of Economics, University  
652 of Canterbury, 2009. URL [http://www.econ.canterbury.ac.nz/RePEc/  
653 cbt/econwp/0913.pdf](http://www.econ.canterbury.ac.nz/RePEc/cbt/econwp/0913.pdf).
- 654 Hyunyoung Choi and Hal Varian. Predicting the present with Google Trends.  
655 Technical report, Google, 2009. URL [http://google.com/googleblogs/  
656 pdfs/google\\_predicting\\_the\\_present.pdf](http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf).
- 657 Hyunyoung Choi and Hal Varian. Predicting the present with Google Trends.  
658 *Economic Record*, 2012. URL [http://people.ischool.berkeley.edu/  
659 ~hal/Papers/2011/ptp.pdf](http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf).
- 660 Pedro Domingos. A few useful things to know about machine learning. *Com-  
661 munications of the ACM*, 55(10), October 2012. URL [http://homes.cs.  
662 washington.edu/~pedrod/papers/cacm12.pdf](http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf).
- 663 Liran Einav and Jonathan Levin. The data revolution and economic analysis.  
664 Technical report, NBER Innovation Policy and the Economy Conference,  
665 2013.
- 666 Andrey Feuerverger, Yu He, and Shashi Khatri. Statistical significance of  
667 the Netflix challenge. *Statistical Science*, 27(2):202–231, 2012. URL [http:  
668 //arxiv.org/abs/1207.5649](http://arxiv.org/abs/1207.5649).

- 669 Jerome Friedman. Stochastic gradient boosting. Technical report, Stan-  
670 ford University, 1999. URL [http://www-stat.stanford.edu/~jhf/ftp/  
671 stobst.pdf](http://www-stat.stanford.edu/~jhf/ftp/stobst.pdf).
- 672 Jerome Friedman and Peter Hall. On bagging and nonlinear estimation.  
673 Technical report, Stanford University, 2005. URL [http://www-stat.  
674 stanford.edu/~jhf/ftp/bag.pdf](http://www-stat.stanford.edu/~jhf/ftp/bag.pdf).
- 675 Jerome Friedman and Bogdan E. Popescu. Predictive learning via rule  
676 ensembles. Technical report, Stanford University, 2005. URL [http:  
677 //www-stat.stanford.edu/~jhf/R-RuleFit.html](http://www-stat.stanford.edu/~jhf/R-RuleFit.html).
- 678 Sharad Goel, Jake M. Hofman, Sbastien Lahaie, David M. Pennock, and  
679 Duncan J. Watts. Predicting consumer behavior with web search. *Pro-  
680 ceedings of the National Academy of Sciences*, 2010. URL [http://www.  
681 pnas.org/content/107/41/17486.full](http://www.pnas.org/content/107/41/17486.full).
- 682 Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of  
683 Statistical Learning: Data Mining, Inference, and Prediction*. Springer-  
684 Verlag, 2 edition, 2009. URL [http://www-stat.stanford.edu/~tibs/  
685 ElemStatLearn/download.html](http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html).
- 686 Rebecca Hellerstein and Menno Middelorp. Forecasting with  
687 internet search data. *Liberty Street Economics Blog of the  
688 Federal Reserve Bank of New York*, January 2012. URL  
689 [http://libertystreeteconomics.newyorkfed.org/2012/01/  
690 forecasting-with-internet-search-data.html](http://libertystreeteconomics.newyorkfed.org/2012/01/forecasting-with-internet-search-data.html).
- 691 David F. Hendry and Hans-Martin Krolzig. We ran one regression. *Oxford  
692 Bulletin of Economics and Statistics*, 66(5):799–810, 2004.
- 693 Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive par-  
694 titioning: A conditional inference framework. *Journal of Computational  
695 and Graphical Statistics*, 15(3):651–674, 2006.

696 Jeremy Howard. The two most important algorithms in predictive mod-  
697 eling today. Conference presentation, February 2013. URL [http://](http://strataconf.com/strata2012/public/schedule/detail/22658)  
698 [strataconf.com/strata2012/public/schedule/detail/22658](http://strataconf.com/strata2012/public/schedule/detail/22658).

699 Bill Howe. Introduction to data science. Technical report, University of  
700 Washington, 2013. URL [https://class.coursera.org/datasci-001/](https://class.coursera.org/datasci-001/lecture/index)  
701 [lecture/index](https://class.coursera.org/datasci-001/lecture/index).

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

705 Jeff Leek. Data analysis, 2013. URL [http://blog.revolutionanalytics.](http://blog.revolutionanalytics.com/2013/04/coursera-data-analysis-course-videos.html)  
706 [com/2013/04/coursera-data-analysis-course-videos.html](http://blog.revolutionanalytics.com/2013/04/coursera-data-analysis-course-videos.html).

707 Randall A. Lewis and Justin M. Rao. On the near impossibility of mea-  
708 suring the returns to advertising. Technical report, Google, Inc. and  
709 Microsoft Research, 2013. URL [http://justinmrao.com/lewis\\_rao\\_](http://justinmrao.com/lewis_rao_nearimpossibility.pdf)  
710 [nearimpossibility.pdf](http://justinmrao.com/lewis_rao_nearimpossibility.pdf).

711 Eduardo Ley and Mark F. J. Steel. On the effect of prior assumptions in  
712 Bayesian model averaging with applications to growth regression. *Jour-*  
713 *nal of Applied Econometrics*, 24(4):651–674, 2009. URL [http://ideas.](http://ideas.repec.org/a/jae/japmet/v24y2009i4p651-674.html)  
714 [repec.org/a/jae/japmet/v24y2009i4p651-674.html](http://ideas.repec.org/a/jae/japmet/v24y2009i4p651-674.html).

715 Nick McLaren and Rachana Shanbhoge. Using internet search data  
716 as economic indicators. *Bank of England Quarterly Bulletin*,  
717 June 2011. URL [http://www.bankofengland.co.uk/publications/](http://www.bankofengland.co.uk/publications/quarterlybulletin/qb110206.pdf)  
718 [quarterlybulletin/qb110206.pdf](http://www.bankofengland.co.uk/publications/quarterlybulletin/qb110206.pdf).

719 James N. Morgan and John A. Sonquist. Problems in the analysis of survey  
720 data, and a proposal. *Journal of the American Statistical Association*, 58  
721 (302):415–434, 1963. URL <http://www.jstor.org/stable/2283276>.

- 722 Alicia H. Munnell, Geoffrey M. B. Tootell, Lynne E. Browne, and James  
723 McEneaney. Mortgage lending in Boston: Interpreting HDMA data. *Amer-*  
724 *ican Economic Review*, pages 25–53, 1996.
- 725 Judea Pearl. *Causality*. Cambridge University Press, 2009a.
- 726 Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*,  
727 4:96–146, 2009b.
- 728 Claudia Perlich, Foster Provost, and Jeffrey S. Simonoff. Tree induction vs.  
729 logistic regression: A learning-curve analysis. *Journal of Machine Learning*  
730 *Research*, 4:211–255, 2003. URL [http://machinelearning.wustl.edu/  
731 mlpapers/paper\\_files/PerlichPS03.pdf](http://machinelearning.wustl.edu/mlpapers/paper_files/PerlichPS03.pdf).
- 732 Donald Rubin. Estimating causal effects of treatment in randomized and non-  
733 randomized studies. *Journal of Educational Psychology*, 66(5):689, 1974.
- 734 Xavier Sala-i-Martin. I just ran two million regressions. *American Economic*  
735 *Review*, 87(2):178–83, 1997.
- 736 Steve Scott and Hal Varian. Bayesian variable selection for nowcasting  
737 economic time series. Technical report, Google, 2012a. URL [http://  
738 www.ischool.berkeley.edu/~hal/Papers/2012/fat.pdf](http://www.ischool.berkeley.edu/~hal/Papers/2012/fat.pdf). Presented  
739 at JSM, San Diego.
- 740 Steve Scott and Hal Varian. Predicting the present with Bayesian structural  
741 time series. Technical report, Google, 2012b. URL [http://www.ischool.  
742 berkeley.edu/~hal/Papers/2013/pred-present-with-bsts.pdf](http://www.ischool.berkeley.edu/~hal/Papers/2013/pred-present-with-bsts.pdf).
- 743 Mark F. J. Steel. Bayesian model averaging and forecasting. *Bulletin*  
744 *of E.U. and U.S. Inflation and Macroeconomic Analysis*, 200:30–41,  
745 2011. URL [http://www2.warwick.ac.uk/fac/sci/statistics/staff/  
746 academic-research/steel/steel\\_homepage/publ/bma\\_forecast.pdf](http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/steel/steel_homepage/publ/bma_forecast.pdf).



747 Danny Sullivan. Google: 100 billion searches per month, search to integrate  
748 gmail, launching enhanced search app for iOS. *Search Engine Land*, 2012.  
749 URL <http://searchengineland.com/google-search-press-129925>.

750 W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer-  
751 Verlag, New York, 4 edition, 2002.

752 Graham Williams. *Data Mining with Rattle and R*. Springer, New York,  
753 2011.

754 Xindong Wu and Vipin Kumar, editors. *The Top Ten Algorithms in*  
755 *Data Mining*. CRC Press, 2009. URL [http://www.cs.uvm.edu/~icdm/  
756 algorithms/index.shtml](http://www.cs.uvm.edu/~icdm/algorithms/index.shtml).