1-1-2019

The Paradox of Big Data

Gary N. Smith
Pomona College

The Paradox of Big Data

Gary Smith

Pomona College

Abstract

Data-mining is often used to discover patterns in Big Data. It is tempting believe that because an unearthed pattern is unusual it must be meaningful, but patterns are inevitable in Big Data and usually meaningless. The paradox of Big Data is that data mining is most seductive when there are a large number of variables, but a large number of variables exacerbates the perils of data mining.

Corresponding author:

Gary Smith

Department of Economics

Pomona College

425 N. College Avenue

Claremont CA 91711

gsmith@pomona.edu

running title: Paradox of Big Data

keywords: data mining, big data, machine learning

word count: 6,316

The Paradox of Big Data

It is easy to believe that the availability of vast amounts of data makes it more likely that data mining will discover new, heretofore, unknown relationships. However, the usefulness of data mining is undermined by the reality that coincidental patterns are inevitable in large data sets. The paradox of Big Data is that data mining is most seductive when working with a large number of variables, but a large number of variables exacerbates the perils of data mining. I use several Monte Carlo simulations to illustrate this paradox.

Background

The scientific method begins with a falsifiable theory, followed by the collection of data for a statistical test of the theory. For example, it is known that heart attacks and strokes can be triggered by blood clots, and that aspirin inhibits blood clotting. A natural research hypothesis is whether regular doses of aspirin may reduce the chances of heart attacks and strokes. This conjecture was tested in the 1980s by a five-year, double-blind randomized control trial involving 22,000 doctors, with half taking an aspirin tablet every other day, and half taking a placebo. The doctors taking placebos had more than three times as many fatal heart attacks and nearly twice as many nonfatal heart attacks as the treatment group (Steering Committee 1988).

Data-mining goes in the other direction, analyzing data without being motivated by theories, indeed viewing the use of *a priori* knowledge as an unwelcome constraint that limits the possibilities for knowledge discovery (Piatetsky-Shapiro 1991, Fayyad, Piatetsky-Shapiro, and Smyth 1996; Cios, Pedrycz, Swiniarski, and Kurgan 2007; Begoli and Horsey 2012). Turing winner Jim Gray endorsed "data exploration" as the "fourth paradigm" of science (Gray undated, Bell, Hey, and Szala 2009). Smith and Cordes (2019) give numerous examples of people in

academia and industry using data-mined models with no underlying theory. One business executive repeatedly expressed his disdain for theory with the pithy comment, “Up is up.”

In an article titled, “The End of Theory: The data deluge makes the scientific method obsolete.” Anderson (2008) argued that

> *Petabytes allow us to say: “Correlation is enough.” We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.*

Sagiroglu and Sinanc (2013) describe data mining as a quest “to reveal hidden patterns and secret correlations.” In the opening lines to a forward for a book on using data mining for knowledge discovery (Kecman 2007) wrote, without evident irony,

> *“If you torture the data long enough, Nature will confess,” said 1991 Nobel-winning economist Ronald Coase. The statement is still true. However, achieving this lofty goal is not easy. First, “long enough” may, in practice, be “too long” in many applications and thus unacceptable. Second, to get “confession” from large data sets one needs to use state-of-the-art “torturing” tools. Third, Nature is very stubborn — not yielding easily or unwilling to reveal its secrets at all.*

The reality is that Coase intended his comment not as a “lofty goal,” but as a humorous criticism of the practice of ransacking data in search of patterns (Tullock 2001).

In the heart-attack example, a data-miner might compile a database of, say, 1,000 personal characteristics of 100 people who suffered heart attacks and 10,000 people who did not, and then use a data-mining algorithm to identify distinguishing characteristics. It might turn out that the

heart attack victims were more likely to have a fondness for apples, work for city government, live in a small town, have green eyes, and use the word *great!* on Facebook.

The data-mining researcher might conclude that apples, government jobs, small towns, green eyes, and Facebook *great!*s are unhealthy, and concoct some fanciful stories to explain these correlations. Or the data-miner might argue that the data speak for themselves and that is all that needs to be said. We don't need theories—data are sufficient. Up is up.

The Feynman Trap

Richard Feynman, a Nobel Laureate physicist, once told an audience (Goodstein 1989),

> *You know, the most amazing thing happened to me tonight. I was coming here, on the way to the lecture, and I came in through the parking lot. And you won't believe what happened. I saw a car with the license plate ARW 357. Can you imagine? Of all the millions of license plates in the state, what was the chance that I would see that particular one tonight? Amazing!*

The Feynman trap is the Achilles heal of data mining. A specific pattern (like a specific license plate) may be highly improbable, *a priori*, but the existence of some pattern (like the existence of some license plate), after the fact, is a certainty. Predicting a specific license plate is astonishing. Reporting an observed plate is not. Using theory to predict a pattern before data are collected is impressive. Reporting an observed pattern after looking at the data is not.

Calude and Longo (2017) prove that, "Complete disorder is an impossibility. Every large set of numbers, points or objects necessarily contains a highly regular pattern." The larger the data set, the more likely we are to find patterns that are *a priori* highly improbable.

A simple example is a streak of L consecutive heads (or tails) in a sequence of coin flips. The

probability that 10 coin flips will give the same result (all heads or all tails) is 0.002. However, the probability of a streak of at least 10 consecutive heads or 10 consecutive tails is 0.087 in 100 coin flips, and 0.623 in 1,000 flips. Table 1 shows how the probability of a streak of length 10 or more increases as the number of flips increases. Table 1 also shows, for a given number of coin flips, the streak length that is more likely than not; for example, in 100 coin flips, a streak of 7 or more consecutive heads or tails is more likely than not, with a probability of 0.543. As the number of flips increases, so does the length of the most likely longest streak. If we consider all possible patterns (alternating heads and tails, two heads followed by one tail, three tails followed by two heads, and so on), an improbable pattern is inevitable in a large set of coin flips.

The same principal holds in models in which a set of explanatory variables is used to predict the values of another variable, in that the probability that statistically significant coincidental relationships will be found increases with the number of potential explanatory variables that are considered.

Many important scientific theories began as attempts to explain observed patterns; for example, Mendel's principles of inheritance laid the basis for modern genetic theory. However, the growing abundance of data necessarily increases the number of highly regular patterns in the data and, therefore, increases the likelihood that a discovered pattern is coincidental and therefore meaningless.

When a study is done following the scientific method, with theory preceding data, a large number of valid observations on a fixed set of variables (like 22,000 doctors separated into treatment and control groups) can provide a persuasive test of the theory. When data mining is done, however, a large number of variables makes it more likely that the patterns that are

discovered are misleading. This is the paradox of Big Data: data mining is most seductive when there are a large number of variables, but a large number of variables exacerbates the perils of data mining.

Prediction

Some data-mining enthusiasts argue that the goal of machine learning is prediction, not the estimation of casual effects (Mullainathan and Spiess 2017). If there is a correlation between Facebook *great!*s and heart attacks, we do not need to know why these are correlated. It is enough to know that they are correlated since one predicts the other. The problem with this argument is that if there is no logical reason for a pattern, it is likely to be a temporary, spurious correlation—like heart attacks and Facebook *great!*s.

Athey (2018, 10) gives this example:

> *To see the difference between prediction and causal inference, imagine that you have a data set that contains data about prices and occupancy rates of hotels…. Imagine first that a hotel chain wishes to form an estimate of the occupancy rates of competitors, based on publicly available prices. This is a prediction problem… [H]igher posted prices are predictive of higher occupancy rates, since hotels tend to raise their prices as they fill up (using yield management software). In contrast, imagine that a hotel chain wishes to estimate how occupancy would change if the hotel raised prices across the board… This is a question of causal inference. Clearly, even though prices and occupancy are positively correlated in a typical dataset, we would not conclude that raising prices would increase occupancy.*

For predictions to be useful, they must be reliable with fresh data. Predictions with new data

might be temporarily successful by chance alone, but consistently reliable predictions require a causal structure. In the hotel example, the statistical correlation between prices and occupancy rates is not a fluke; it reflects an underlying logical relationship—hotels raise prices when demand surges—and might be useful for predictions. In contrast, a data-mining program found a statistical relationship between U.S. stock prices and the daily high temperature in Curtin, Australia (Smith 2018, 184-185). There was no underlying reason for this relationship and it was useless for predicting stock prices with fresh data.

Many data-miners have tried to predict stock prices by assuming that correlation is enough. In 2008 Michael Kharitonov and Jon McAuliffe, with PhDs in computer science and statistics, respectively, launched an algorithmic trading company they named Voleon. A *Wired* article (Salmon and Stokes 2010) enthusiastically reported that Voleon's trading algorithms operate on their own, with no human supervision:

> *McAuliffe and Kharitonov say that they don't even know what their bots are looking for or how they reach their conclusions. "What we say is 'Here's a bunch of data. Extract the signal from the noise,'" Kharitonov says. "We don't know what that signal is going to be like."*

Voleon's algorithms reportedly sift through all kinds of data, including satellite images, credit-card receipts, and social media language, looking for patterns related to stock prices. Part of their machine-learning involved an analysis of terabytes of data on every price change of every stock over a 15-year period.

Ten years later, the *Wall Street Journal* (Hope and Chung. 2017) reported that Voleon had underperformed the S&P 500. Kharitonov said,

> *Most of the things we've tried have failed…The idea that we could just take the machine-learning techniques in speech recognition and computer vision to generate better forecasts just didn't work.*

The *Journal* attributed Voleon's struggles to the volatile nature of markets:

> *The basic problem they faced was that markets are so chaotic. Machine-learning systems have been best applied so far to situations where patterns are more of a repeating nature, and thus easier to discern, such as in playing the ancient game of Go or even guiding a driverless car. The financial markets are "noisier"—continually being affected by new events, the relationships among which are frequently shifting.*

This explanation is only partly true and misses the bigger point. It *is* easier for algorithms to deal with board games with fixed rules. However, the problem with stock-trading algorithms is not just that markets are buffeted by unexpected events, but that the statistical patterns the algorithms discover are "frequently shifting" because they are coincidental and fleeting. If an algorithm finds a correlation between stock prices and the use of calm words on Twitter, and the pattern disappears when it is used to buy and sell stocks, it is not because the world has changed, but because there never was a real causal relationship—just a temporary statistical correlation.

Causal models are useful precisely because they make trustworthy predictions. For a model to make reliable predictions, we do not need to know the exact reason why two things are related, but there does need to be a reason. If there were no reason for hotel prices and occupancy rates to be related, the model would not make dependable predictions.

Imagine that a thousand small objects of various sizes, shapes, colors, and densities are created by a 3D printer that has been programmed so that the characteristics are independently

determined. A data-mining algorithm looking for patterns in these objects might discover that yellow objects are more likely to be bumpy than are objects with other colors. When the algorithm is used for fresh data, it is likely to fail because there is no systematic relationship between color and bumpiness. Now suppose, instead, that these objects are rocks found at the bottom of a lake and there is some scientific reason why bumpy rocks in this lake tend to be yellow. Now, the correlation between bumpiness and yellowness may be a useful predictor even if we don't completely understand why bumpy rocks are often yellow.

The crucial distinction between these scenarios is whether there is a structural relationship—a reason why bumpy objects are often yellow. We may not know the reason, but it is the existence of a structural relationship that makes predictions useful. Correlation is not enough. Causality is crucial.

Holdout Data

It has been argued that splitting the data into a training set and test set will tell us whether a discovered relationship is merely coincidental (Egami et al. 2016, Athey 2018). In the pebble example, training the algorithm on 500 objects and then testing the algorithm on the holdout data is likely to show the fragility of the yellow/bumpiness statistical relationship if there is no underlying systematic relationship.

If that happens, however, the data mining algorithm can keep looking for other patterns until it finds one that makes successful predictions with both the training data and the holdout data—and it is certain to succeed if there are a large enough number of characteristics. Just as spurious correlations can be discovered with a subset of the data, so spurious correlations can be discovered with the full set of data.

To illustrate this point, I ran 10,000 Monte Carlo simulations using a normal distribution to generate 200 independent observations for 101 random variables, one of which was used as the response variable Y to be predicted. Each simulation used 100 in-sample observations to estimate the simple regression model, $Y = \alpha + \beta X + \varepsilon$, for each of the 100 other variables X. The correlation r between the predicted and actual values of Y is equal to the absolute value of the correlation between X and Y. The p value for a t test of the null hypothesis that the population value of r is zero is equal to the p value for a t test of the null hypothesis that the value of β is zero. Each simulation identified the predictor variable with the smallest (two-sided) p value as long as that p value was less than or equal to 0.05.

With 100 observations, a p value less than or equal to 0.05 requires a correlation r of at least 0.196551. Among the explanatory variables selected by the data mining algorithm in these 10,000 simulations, the smallest correlation was 0.1966; the average was 0.2738, and the maximum was 0.4878—corresponding to p values of 0.0499, 0.0058, and less than 0.000001, respectively. With 10,000 simulations involving 100 predictor variables, it is not surprising that one of the correlations had a one-in-a-million chance of occurring.

The best explanatory variable in each simulation was then used to predict the value of Y using the 100 holdout observations. Figure 1 is a scatterplot of the in-sample and out-of-sample values of r. The average out-of-sample value of r was 0.00015, which is approximately 0 because all the data were generated independently. However, there were 530 simulations in which the out-of-sample correlation was statistically significant at the 5 percent level, and 101 simulations in which the out-of-sample correlation was statistically significant at the 1 percent level, in each case suggesting that the data-mining algorithm had discovered a robust relationship that passed

the holdout test for making useful predictions.

Although the average out-of-sample correlation might be zero, some spurious correlations can be expected, by chance, to continue out of sample. On average, even with completely independent data, five percent of all tested hypothesis can be expected to yield statistically significant in-sample results, and five percent of these statistically significant in-sample results can be expected to also be statistically significant out of sample. In large data sets with many variables, unrestrained data mining is certain to discover spurious patterns that appear to make successful prediction for both the training data and the test data. Out-of-sample validation is not a guarantee of the value of data-mined patterns discovered in-sample.

Holdout tests are surely valuable, however data mining with a holdout test is still data mining and still subject to the same pitfalls. Using data mining to identify a statistical relationship that holds for 200 observations is more difficult than identifying one for 100 observations, but it is nonetheless inevitable if a large enough data base is data mined, and gives no assurance that the identified relationship will be useful for making predictions with fresh data.

Crowding Out

There is a more subtle problem with wholesale data mining tempered by holdout tests. Suppose that a data-mining algorithm is used to select predictor variables from a data set that includes a relatively small number of "true" variables that are structurally related to the variable being predicted and a large number of "nuisance" variables that are independent of the variable being predicted. One problem is that some nuisance variables are likely to be coincidentally successful both in-sample and out-of-sample, but then flop when the model goes live with new data.

A second problem is that an unintended consequence of a data-mining algorithm selecting nuisance variables is that they may well be chosen in place of true variables that could be used to make reliable predictions. The testing and retesting of a data-mined model may expose the nuisance variables as useless, but will not resuscitate the true variables that were initially crowded out by the nuisance variables. The more nuisance variables that are initially considered, the more likely it is that some true variables will disappear without a trace.

Multiple Regression

I investigated these arguments with Monte Carlo simulations in which some variables are systematically related to the variable being predicted and other variables are random noise. For my data-mining algorithm, I used multiple regression models because these are a very powerful and easily understood data-mining tool for exploring the relationship between a response variable and multiple explanatory variables.

One obstacle to data mining big data in a search for the best-fitting regression model is the sheer number of models that can be considered. A researcher who wants to choose up to 10 out of 100 possible explanatory variables has 19.4 trillion possible combinations to choose from. With 1,000 possible explanatory variables, there are 2.66×10^{23} combinations of up to 10 variables. So, in practice, various work-arounds are used.

More than 50 years ago, Efroymson (1960) proposed choosing the explanatory variables from a group of candidate variables by going through a series of automated steps, called stepwise regression. At every step, each candidate variable is evaluated, typically using the *p* value for its estimated coefficient.

A forward-selection rule starts with no explanatory variables and then adds variables, one by

one, based on which variable is the most statistically significant, until there are no remaining statistically significant variables. This procedure circumvents the computational burden of trying all possible combinations of explanatory variables. The use of forward-selection stepwise regression for identifying the 10 most statistically significant explanatory variables requires only 955 regressions if there are 100 candidate variables, and 9,955 regressions if there are 1,000 candidates.

Stepwise regression was born back when computers were much slower than today, but it has become a popular data-mining tool because it is computationally much less demanding than a full search over all possible combinations of explanatory variables. Several textbooks endorse stepwise regression (Rachev, Mittnik, Fabozzi, Focardi, and Jašić 2006; McDonald 2014), including a handbook explicitly devoted to data mining methods (Hastie, Tibshirani, and Friedman 2016). A survey of papers published in 2004 in three leading ecological and behavioral journals found that 57 percent of the papers that reported multiple regression results used stepwise regression (Whittingham, Stephens, Bradbury, and Freckleton 2006). A survey of four leading epidemiologic journals found that 20 percent of the articles published in 2008 used stepwise regression (Walter and Tiemeier 2009). A study of articles published between 2004 and 2008 in two leading Chinese epidemiology journals found that, of the articles using multiple regression models, 44 percent used stepwise procedures (Liao and Lynn 2010).

Stepwise regression follows an automated rule based on statistical significance, with no regard for whether there is a logical reason for including a potential explanatory variable in the model. It is data mining on steroids. Thus, Cios, Pedrycz, Witold, and Kurgan (2007) recommend stepwise regression as an efficient way of using data mining for knowledge discovery (see also

Hastie, Tibshirani, and Friedman 2009; Varian 2014; and Bruce and Bruce 2017).

Methodology

Monte Carlo simulations were used to explore the perils of data mining. A total of *n* observations for each of *m* candidate explanatory variable were determined by random draws from a normal distribution with mean 0 and standard deviation σ_x:

$$X_{i,j} = \varepsilon_{i,j} \qquad \varepsilon \sim N[0, \sigma_x] \tag{1}$$

The independence of the explanatory variables ensures that there are no structural relationships among the explanatory variables that might cause some variables to be proxies for others.

The central question is how effective the estimated model is at making reliable predictions with fresh data. So, in each simulation, half the observations were used to estimate the model's coefficients, and the remaining half were used to test the model's reliability.

All the data were centered by subtracting the sample means. The in-sample data were centered on the in-sample means and the out-of-sample data were centered on the out-of-sample means so that the out-of-sample predictions would not be inflated if the in-sample and out-of-sample means differed.

Five randomly selected explanatory variables (the *true* variables) were used to determine the value of a dependent variable *Y*,

$$Y_j = \sum_{i=1}^{5} \beta_i X_{i,j} + \nu_j, \qquad \upsilon \sim N[0, \sigma_y] \tag{2}$$

where the value of each β coefficient was randomly determined from a uniform distribution ranging from 2 to 4, and υ is normally distributed with mean 0 and standard deviation σ_y. The range 0 to 2 was excluded so that the true variables would have substantial effects on the

dependent variable. The other candidate variables are *nuisance* variables that have no effect on Y, but might be coincidentally correlated with Y.

The base case was $\sigma_x = 5$, $\sigma_y = 20$, $m = 100$ candidate variables; and $n = 200$ observations, but I also considered all combinations of $\sigma_x = 5$, 10, or 20; $\sigma_y = 10$, 20, or 30; $m = 5$, 10, 50, 100, 500, or 1000; and $n = 100$, 200, 500, or 1,000. For the range of values considered here, the results were robust with respect to the values of σ_x and σ_y, so I only report results for the base case, $\sigma_x = 5$ and $\sigma_y = 20$.

One hundred thousand simulations were done for each parameterization of the model. Every step of the stepwise regression procedure added the candidate explanatory variable with the lowest two-sided p-value if this p value was less than 0.05. The results were not due to the use of stepwise regression to select the explanatory variables, but rather to data mining. Stepwise regression is simply a practical data-mining tool for identifying explanatory variables that are statistically correlated with the variable being predicted.

In one set of simulations, all of the candidate dependent variables were nuisance variables. In the second set of simulations, five of the candidate variables were the five true variables that were used to generate the values of the dependent variable. The first set of simulations, with entirely spurious variables, explores the extent to which coincidental correlations with the dependent variable can create an illusion of a successful prediction model. The second set of simulations, in which all five of the true explanatory variables are among the candidate variables, explores how well a data mining procedure is able to distinguish between meaningful and meaningless variables.

The predictive success of the model was gauged by the correlation between the actual values

of the dependent variable and the model's predicted values. The square of the in-sample correlation is the coefficient of multiple determination, R^2 for the estimated model. The out-of-sample correlation is the corresponding statistic using the out-of-sample data with the in-sample estimated coefficients.

Results With No True Variables

Tables 2 and 3 report the results of the Monte Carlo simulations with no true variables. Table 2 shows that, even though all of the potential explanatory variables are nuisance variables, a data mining procedure that includes variables based on statistical significance can include a substantial number of nuisance variables. With a small number of observations, the number of selected variables is constrained by the number of observations but, otherwise, the number of included variables increases as the number of candidate variables increases and declines as the number of observations increases—presumably because the more accurate estimates are more likely to be close to zero. Yet, even with 1,000 observations on each variable, many nuisance variables are included in the model when a large number of candidate variables are considered, confirming the argument of Calude and Longo.

Table 3 shows that even when the data-mined models only considered nuisance explanatory variables, they were still likely to find coincidental patterns that create a false impression of success. For example, with 100 nuisance variables, the average in-sample correlation between the predicted and actual value of the response variable was 0.549 with 100 observations and 1.000 with 1,000 observations.

Table 3 also shows that the in-sample correlations increase with the number of candidate variables and decline as the number of observations increases. A large number of candidate

variables can create a false illusion that the model is capable of making accurate predictions. Indeed, the model can overfit the data perfectly, with a correlation of 1.000. On the other hand, for a given number of candidate variables, an increase in the number of observations worsens the in-sample fit because it is harder to overfit a model with nuisance variables when there are a large number of observations relative to the number of candidate variables.

No matter what the in-sample correlation, it is misleading large because these are, after all, nuisance variables that are independent of the dependent variable. The out-of-sample correlations between the dependent variable and the selected explanatory variables averaged zero. Although the out-of-sample correlations were, on average, zero, there were many cases in which the out-of-sample correlations were as high or higher than the in-sample correlations suggesting that the model is useful for predictions when it is, in fact, useless.

Results With Five True Variables

Tables 4 - 7 report the results of the Monte Carlo simulations with the five true variables included among the candidate variables. Table 4 shows that the inclusion of five true variables did not eliminate the selection of nuisance variables. More often, the total number of nuisance variables that were included increased. An increase in the number of observations had conflicting effects—making it more likely that true variables were included and somewhat less likely that nuisance variables were included. Thus, for a given number of candidate variables, the number of included variables may initially increase and then decline as the number of observations is increased.

Table 5 confirms that an expansion in the number of candidate variables increased the chances that an included variable is a nuisance variable. For example, with 100 observations for

100 candidate variables (5 true and 95 nuisance), the probability that an included variable is a nuisance variable was 0.599. This probability approached 1 as the number of candidate variables increased. With 100 observations on 500 variables, the average number of included variables was 96.82 and, on average, 0.963(96.82) = 93.23 were nuisance variables and 3.59 were true variables.

These simulations also documented how a plethora of nuisance variables can crowd out true variables. With 100 observations and 100 candidate variables, for example, one or more true variables were crowded out 50.2 percent of the time, and two or more variables were crowded out 16.0 percent of the time. There were even occasions when all five true variables were crowded out.

Tables 6 and 7 show the in-sample and out-of-sample correlations. Comparing Tables 3 and 6, the in-sample correlations were substantially higher with the five true variables among the candidate variables if there were few candidate explanatory variables or a large number of observations. The out-of-sample correlations (Table 7) with five true variables were substantially lower than the in-sample correlations (Table 6) when the number of candidate variables was large relative to the number of observations. For example, with 50 observations for 1,000 candidate variables, the average correlation was 1.000 in-sample and 0.097 out-of-sample.

Discussion

Data-mined prediction models can make statistically persuasive in-sample predictions even they only consider irrelevant explanatory variables. Out-of-sample validation is unreliable because some data-mined models using nothing but nuisance variables will be statistically impressive both in-sample and out-of-sample.

The inclusion of true explanatory variables among the candidate explanatory variables makes it more likely that a model will be useful for out-of-sample predictions, but this usefulness is undermined by the inclusion of nuisance variables that crowd out true variables and improve the in-sample predictions but worsen the out-of-sample predictions. Tests with holdout data can help assess the overall effectiveness of a model and weed out nuisance variables, but cannot resuscitate true explanatory variables that have been crowded out.

Data-mining algorithms cannot effectively distinguish between true variables and nuisance variables because computer algorithms do not know what variables are in any meaningful way and cannot assess how or why variables might be related (Smith 2018). Currently, only human expertise can do that.

The increasing reliance on data mining algorithms to build models unguided by human expertise may be partly responsible for the reproducibility crisis, in which attempts to replicate published research findings often fail (Ioannidis 2005, Pashler and Wagenmakers 2012, Baker 2017). Results reported with data-mined models are inherently not reproducible, since they will almost certainly include nuisance variables that increase the model's in-sample fit, while reducing the out-of-sample success.

Conclusion

The Monte Carlo simulations reported here demonstrate that the performance of data-mined prediction models can be misleading, both in-sample and for out-of-sample validation data.

These empirical simulations illustrate the validity of the theoretical argument of Calude and Longo that spurious correlations are endemic in large data sets and, consequently, "Too much information tends to behave like very little information (Calude and Longo 2017, 595)."

Prediction models are more likely to be useful and the results are more likely to be reproducible if expert opinion is used to select a plausible list of explanatory variables, instead of viewing human expertise as an unhelpful constraint on knowledge discovery. This is a corollary of the paradox of big data: the larger the number of possible explanatory variables, the more important is human expertise.

The reliance on computer algorithms to select useful explanatory variables will continue to be problematic until computers acquire the common sense, wisdom, and expertise needed to distinguish between meaningful and meaningless patterns.

References

Anderson, Chris. 2008. The end of theory, will the data deluge make the scientific method obsolete?, *Wired*, June 23.

Athey, Susan. 2018. The Impact of Machine Learning on Economics, In: Agrawal, Ajay, Gans, Joshua, and Avi Goldfarb, Eds, *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press.

Baker, Monya, 2017. 1,500 scientists lift the lid on reproducibility, *Nature*, 533(7604): 452-4.

Begoli E, Horsey, J. 2012. Design principles for effective knowledge discovery from big data, Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference.

Bell G, Hey T, Szalay A. 2009. Beyond the Data Deluge, *Science*. 323 (5919): 1297-1298.

Bruce P, and Bruce A. 2017. *Practical Statistics for Data Scientists: 50 Essential Concepts*, O'Reilly Media.

Calude, Cristian S, Longo, Giuseppe. 2017. The deluge of spurious correlations in big data, *Foundations of Science,* 22(3), 595–612 https://doi.org/10.1007/s10699-016-9489-4.

Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA. 2007. *Data Mining: A Knowledge Discovery Approach*, New York, Springer.

Efroymson MA., 1960. Multiple Regression Analysis, In: A. Ralston and H. S. Wilf, Eds., *Mathematical Methods for Digital Computers*, John Wiley, New York.

Egami, Naoki, Fong, Christian J., Grimmers, Justin, Roberts, Margaret E., and Brandon M. Stewart. 2018. How to Make Causal Inferences Using Text. arXiv:1802.02163v12016.

Fayyad U, Piatetsky-Shapiro G, Smyth P. 1996. From data mining to knowledge discovery in

databases, *AI Magazine*, 17 (3): 37-54.

Goodstein, David L. 1989. Richard P. Feynman, Teacher, Physics Today, 42 (2): 70 - 75.

Gray, Jim. Undated. eScience—A Transformed Scientific Method, http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2nd edition. URL http://www-stat.stanford.edu/~tibs/ ElemStatLearn/download.html.

Hastie T, Tibshirani R, Friedman J. 2016. *The Elements of Statistical Learning*, 2nd Edition, New York, Springer.

Hope, Bradley, and Juliet Chung. 2017. The Future Is Bumpy: High-Tech Hedge Fund Hits Limits of Robot Stock Picking, *Wall Street Journal*, December 17.

Ioannidis, John A. 2005. Contradicted and initially stronger effects in highly cited clinical research, *Journal of the American Medical Association*, 294 (2): 218–228.

Kecman V, 2007. Forward, in Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA. *Data Mining: A Knowledge Discovery Approach*, New York, Springer, xi.

Liao H, Lynn HS, 2010. A survey of variable selection methods in two Chinese epidemiology journals. *BMC Medical Research Methodology*, 10: 87. http://doi.org/10.1186/1471-2288-10-87.

McDonald JH, 2014. *Handbook of Biological Statistics*, 3rd ed.. Baltimore, Maryland: Sparky House Publishing.

Mullainathan, S. and J. Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Pashler, Harold, Wagenmakers, Eric Jan. 2012. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?. *Perspectives on Psychological Science*. 7 (6): 528–530.

Piatetsky-Shapiro, G. 1991. Knowledge discovery in real databases: A report on the IJCAI-89 workshop. *AI Magazine*. 11(5): 68–70.

Rachev ST, Mittnik S, Fabozzi FJ, Focardi SM, Jašić T, 2006. *Financial Econometrics: From Basics to Advanced Modeling Techniques*, New York: Wiley.

Sagiroglu S, Sinanc D. 2013. Big data: A review, collaboration Technologies and Systems (CTS), 2013 International Conference.

Salmon, Felix, and Jon Stokes. 2010. Algorithms Take Control of Wall Street, *Wired*, December 27.

Smith, Gary, 2018. *The AI Delusion*, Oxford: Oxford University Press.

Smith, Gary, and Jay Cordes. 2019. *The 9 Pitfalls of Data Science*, Oxford: Oxford University Press.

Steering Committee of the Physicians' Health Study Research Group. 1988. Preliminary Report: Findings From the Aspirin Component of the Ongoing Physicians' Health Study, *New England Journal of Medicine*, January 28, 262-264.

Tullock G, 2001. A comment on Daniel Klein's "A plea to economists who favor liberty," *Eastern Economic Journal*. 27 (2): 203-207.

Varian HR, 2014. Big data: New tricks for econometrics, *The Journal of Economic Perspectives*. 28 (2): 3-27.

Walter S, Tiemeier H. 2009. Variable selection: current practice in epidemiological studies,

European Journal of Epidemiology. 24 (12): 733-736.

Whittingham, MJ, Stephens PA, Bradbury RB, Freckleton RP. 2006. Why do we still use stepwise modelling in ecology and behaviour?, *Journal of Animal Ecology*, 75 (5): 1182-1189.

Table 1 Longest Streak, L

Number of Flips	P[L ≥ 10]	*n* such that P[L ≥ *n*] > 0.5	P[*L* ≥ *n*]
10	0.002	3	0.826
100	0.087	7	0.543
250	0.212	8	0.625
500	0.385	9	0.625
1,000	0.623	10	0.623
10,000	1.000	13	0.705
100,000	1.000	17	0.531
1,000,000	1.000	20	0.623

Table 2 Average Number of Variables Per Equation, No True Variables

Observations Used	Number of Candidate Variables					
to Estimate Model	5	10	50	100	500	1,000
50	1.12	1.29	3.40	9.32	47.58	47.99
100	1.11	1.27	3.05	6.63	97.79	96.79
500	1.11	1.25	2.78	5.30	36.38	98.70
1,000	1.11	1.25	2.74	5.18	29.76	73.92

Table 3 In-Sample Correlation, No True Variables

Observations Used	Number of Candidate Variables					
to Estimate Model	5	10	50	100	500	1,000
50	0.344	0.365	0.551	0.784	1.000	1.000
100	0.244	0.258	0.385	0.549	1.000	1.000
500	0.109	0.115	0.169	0.234	0.582	0.852
1,000	0.078	0.081	0.119	0.164	0.392	0.586

Table 4 Average Number of Variables Per Equation, Five True Variables

Observations Used	Number of Candidate Variables					
to Estimate Model	5	10	50	100	500	1,000
50	3.15	3.39	5.90	12.06	47.66	47.99
100	4.50	4.74	6.99	10.71	97.84	97.88
500	5.00	5.25	7.30	9.96	40.89	97.68
1,000	5.00	5.25	7.28	9.83	34.66	80.54

Table 5 Chances of Being a Nuisance Variable, Five True Variables

Observations Used	Number of Candidate Variables					
to Estimate Model	5	10	50	100	500	1,000
50	0.000	0.084	0.520	0.783	0.968	0.979
100	0.000	0.055	0.371	0.599	0.962	0.969
500	0.000	0.047	0.315	0.498	0.878	0.949
1,000	0.000	0.047	0.313	0.491	0.856	0.938

Table 6 In-Sample Correlation, Five True Variables

Observations Used	Number of Candidate Variables					
to Estimate Model	5	10	50	100	500	1,000
50	0.639	0.652	0.758	0.884	1.000	1.000
100	0.657	0.663	0.714	0.780	1.000	1.000
500	0.650	0.651	0.661	0.674	0.788	0.911
1,000	0.648	0.649	0.654	0.660	0.715	0.791

Table 7 Out-of-Sample Correlation, Five True Variables

Observations Used	Number of Candidate Variables					
to Estimate Model	5	10	50	100	500	1,000
50	0.509	0.491	0.380	0.280	0.131	0.097
100	0.606	0.600	0.543	0.478	0.266	0.245
500	0.642	0.641	0.631	0.618	0.505	0.400
1,000	0.645	0.644	0.639	0.632	0.577	0.503

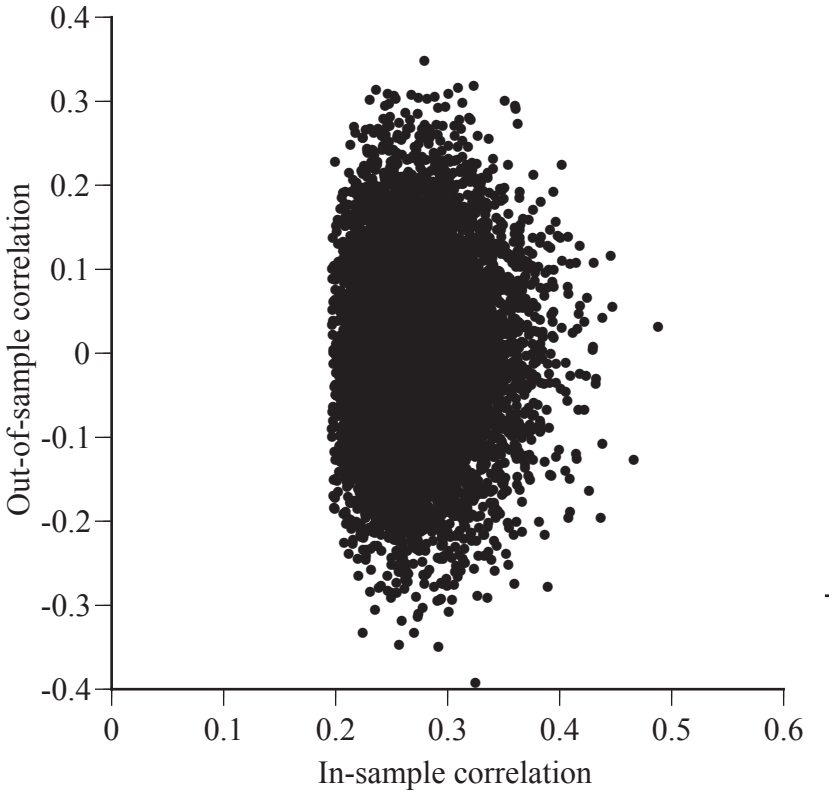

Figure 1 In-Sample and Out-of-Sample Correlations