JIT

Invited Article

Data mining fool's gold

Gary Smith

Journal of Information Technology
1–13
© Association for Information Technology Trust 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0268396220915600
Journals.sagepub.com/jinf
SAGE

Abstract

The scientific method is based on the rigorous testing of falsifiable conjectures. Data mining, in contrast, puts data before theory by searching for statistical patterns without being constrained by prespecified hypotheses. Artificial intelligence and machine learning systems, for example, often rely on data-mining algorithms to construct models with little or no human guidance. However, a plethora of patterns are inevitable in large data sets, and computer algorithms have no effective way of assessing whether the patterns they unearth are truly useful or meaningless coincidences. While data mining sometimes discovers useful relationships, the data deluge has caused the number of possible patterns that can be discovered relative to the number that are genuinely useful to grow exponentially—which makes it increasingly likely that what data mining unearths is likely to be fool's gold.

Keywords

Data mining, data exploration, knowledge discovery, artificial intelligence, machine learning, HARKing

Introduction

The scientific revolution was fueled by what has come to be known as the scientific method: specify a falsifiable conjecture and then collect data, ideally through a controlled experiment, to test this hypothesis. The modern availability of powerful computers and vast amounts of data makes it tempting to reverse the process by using data "to reveal hidden patterns and secret correlations" (Sagiroglu and Sinanc, 2013). When a pattern is found, a theory can be conceived after the fact to explain the pattern, or it might be argued that theories are unnecessary (Begoli and Horsey, 2012; Cios et al., 2007; Fayyad et al., 1996). Some assert that using a priori knowledge before looking at the data is not only unnecessary, but limiting (Piatetsky-Shapiro, 1991).

This reversal of the scientific method goes by many names, including data mining, data exploration, knowledge discovery, and information harvesting. What they have in common is the belief that data come before theory. This is known as HARKing: Hypothesizing After the Results are Known. The harsh sound of the word reflects the dangers of HARKing: It is tempting to believe that patterns are unusual and their discovery meaningful; in large data sets, patterns are inevitable and generally meaningless.

Calude and Longo (2017) prove that large amounts of data necessarily contain a large number of patterns and correlations waiting to be discovered:

> the more data, the more arbitrary, meaningless and useless (for future action) correlations will be found in them. Thus, paradoxically, the more information we have, the more difficult is to extract meaning from it. Too much information tends to behave like very little information.

> If there is a fixed set of true statistical relationships that are useful for making predictions, the data deluge necessarily increases the ratio of meaningless statistical relationships to true relationships.

From a Bayesian perspective, suppose that 1 out of every 1000 patterns that might be discovered *is* useful and the other 999 are useless, and that we use a reliable statistical test that will correctly identify a real pattern as truly useful and a coincidental pattern as truly useless 95% of the time. Our prior probability that we find a useful pattern by searching randomly is 1 in 1000. After we have found a pattern and determined that it is statistically significant at the 5% level, the posterior probability that it is useful is less than 1 in 50. This is higher than 1 in 1000, but it is hardly persuasive. We are far more likely than not to have discovered a pattern that is genuinely useless.

Table 1 shows the posterior probabilities for other values of the prior probability.

Pomona College, USA

Corresponding author:
Gary Smith, Department of Economics, Pomona College, 425 N. College Avenue, Claremont, CA 91711, USA.
Email: gsmith@pomona.edu

Table 1. Probability that a discovered pattern is useful.

Prior probability	Posterior probability
0.001	0.018664
0.0001	0.001897
0.00001	0.000190
0.000001	0.000019

We do not know precisely how many useless patterns are out there waiting to be discovered, but we do know that with big data and powerful computers, it is a very large number that is getting larger every day, which means that the probably that a randomly discovered pattern is useful is getting ever closer to 0.

Data mining

Decades ago, being accused of data mining, fishing expeditions, and data dredging were insults comparable to being accused of plagiarism. James Tobin (1972), a Nobel laureate in economics, wryly observed that when researchers did calculations by hand, they thought hard before calculating. With terabytes of data and lightning-fast computers, it is too easy to calculate first, think later.

Ronald Coase (1988), another economics Nobel laureate, famously remarked, "If you torture the data long enough, it will confess." Since those who ransack data looking for statistical patterns will surely find some, their discoveries demonstrate nothing more than that data were ransacked.

Today, the combination of the powerful computers and the data explosion has made data mining irresistible for some. Anderson (2008), at the time the editor-in-chief of *Wired*, wrote an article with the provocative title, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." Anderson argued:

> With enough data, the numbers speak for themselves. . . . Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

In the opening lines in a forward for a book on using data mining for knowledge discovery, a computer scientist (Kecman, 2007) cited Coase while arguing that state-of-the-art data-torturing tools may sometimes be needed to reveal nature's secrets:

> "If you torture the data long enough, [it] 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.

Coase did not intend his comment to be a lofty goal worth seeking, but a succinct criticism of the practice of pillaging data in search of statistical significance (Tullock, 2001).

Research published in highly respected Information Systems (IS) journals has traditionally followed the scientific method, with accepted theories suggesting hypotheses that are tested with high-quality data. Grover and Lyytinen (2015) make an intentionally provocative argument for less reliance on established theories and more openness to data-driven research, and such papers are now being published in top-tier IS journals (e.g. Brynjolfsson et al., 2016; Guo et al., 2017; Martens et al., 2016; Shi et al., 2016). Jafar et al. (2017) report that undergraduate and graduate IS curricula increasingly include data-mining courses offered under a variety of names.

Data mining sometimes discovers useful models that are later confirmed by the scientific method. What is problematic is the widespread uncritical acceptance of data-mined results as equivalent to the scientific method. If an explanation is desired, it is easy for creative humans to think up fanciful theories after the fact. If the data mining is hidden inside a black box algorithm, no explanation is possible. If the researcher believes that correlation supersedes causation, no explanation is needed.

Prediction

Some data-mining enthusiasts argue that the goal is prediction, not the confirmation of causal effects (Mullainathan and Spiess, 2017). If there is a correlation between the number of Google searches for the word *Scorpio* and the price of avocados in San Francisco, we do not need to know why these are correlated. It is enough to know that they are correlated since one predicts the other. Smith and Cordes (2019) quote a business executive who repeatedly embraced data mining with the pithy comment, "Up is up."

Athey (2018: 10) argued that prediction does not require causation:

> 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.

However, for a statistical relationship to be useful for making predictions, there must be a reason for the relationship. For example, the ancient Egyptians noticed that the annual flooding of the Nile was regularly preceded by seeing Sirius—the brightest star visible from earth—appear to rise in the eastern horizon just before the sun rose. Sirius did not cause the flooding, but it was a useful predictor because there was an underlying reason: Sirius rose before dawn every year in mid-July and heavy rains that began in May in the Ethiopian Highlands caused the flooding of the Nile in late July.

In the hotel example, the statistical correlation between prices and occupancy rates is not a fluke; it reflects a real underlying structural relationship. In contrast, a discovered statistical relationship between hotel occupancy rates in Denver and the price of tea in China would be useless for predicting either.

Out-of-sample data

The perils of data mining are often exposed when a pattern that has been discovered by rummaging through data disappears when it is applied to fresh data. So, it would seem that an effective way of determining whether a statistical pattern is meaningful or meaningless is to divide the original data into two parts—*in-sample data* that can be used to discover models, and *out-of-sample* data that can be used to test the models that were discovered with the in-sample data (Athey, 2018; Egami et al., 2018). This procedure is sensible but, unfortunately, provides no guarantees.

Suppose that we are trying to figure out a way to predict Liverpool's margin of victory in the English Premier League, and we divide the 2018 season into the first half (19 in-sample games) and the second half (19 out-of-sample games). If a data-mining algorithm looks at temperature data in hundreds of California cities on the day before Liverpool matches, it might discover that the difference between the high and low temperatures in Claremont, California, is a good predictor of the Liverpool score.

If this statistical pattern is purely coincidental (as it surely is), then testing the relationship on the out-of-sample data is likely to show that it is useless for predicting Liverpool scores. If that happens, however, the data-mining algorithm can keep looking for other weather patterns (there are lots of cities in California, and other places, if needed) until it finds one that makes successful predictions with both the in-sample data and the out-of-sample data—and it is certain to succeed if a sufficiently large number of cities are considered. Just as spurious correlations can be discovered for the first 19 games of the Premiere League season, so spurious correlations can be discovered for all 38 games.

A pattern is generally considered to be statistically significant if there is less than a 5% chance that it would occur by luck alone. This means that if we are so misguided as to only consider correlations between pairs of independently generated random numbers, we can expect 1 out of every 20 spurious correlations to pass the in-sample test, and 1 out of 400 to pass both the in-sample and out-of-sample tests. A determined head-in-the-sand researcher who analyzes 10,000 pairs of unrelated data can expect to find 25 correlations that are statistically significant both in-sample and out-of-sample. In the age of the data deluge, there are a *lot* more than 10,000 pairs that can be analyzed and a lot more than 25 spurious correlations that will survive in-sample and out-of-sample tests.

Out-of-sample tests are surely valuable; however, data mining with out-of-sample data is still data mining and still subject to the same pitfalls.

Crowding out

There is a more subtle problem with wholesale data mining tempered by out-of-sample tests. Suppose that a data-mining algorithm is used to select several predictor variables from a data set that includes a relatively small number of "true" variables that are causally related to the variable being predicted and a large number of "nuisance" variables that are independent of the variable being predicted. One problem, as we have seen, 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.

An additional problem is that a data-mining algorithm may select nuisance variables in place of true variables that would be useful for making reliable predictions. Testing and retesting a data-mined model may eventually expose the nuisance variables as useless, but it cannot bring back the true variables that were 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.

Not enough good data?

In fields where laboratory experiments are possible, theories can be tested endlessly and coincidental patterns will eventually be exposed as such. However, in many fields, there are not enough data for a large number of out-of-sample tests (Arnott et al., 2018).

An extreme example is that shortly after the 2016 U.S. presidential election, it was widely reported that a history professor had correctly predicted that Trump would win the popular vote based on 13 key variables (Stevenson, 2016). Overfitting was an obvious concern. As it turned out, Trump lost the popular vote, contrary to the model's prediction, but the bigger point is that one observation every 4 years does not allow for much out-of-sample testing. Other cases are less extreme, but still limiting.

Another problem with observational data collected outside of controlled experiments is self-selection bias. People who make different choices may experience different outcomes not because of their choices but because of the types of people who make such choices. Data-mining algorithms are ill-equipped to recognize such biases because they do not consider the nature of the data being mined.

Social media data are currently fashionable because they provide vast amounts of data, but their usefulness is questionable. Are the people who use social media representative of the population? Are the messages they send representative of their feelings? Are they even people? A 2018 Pew Research Center study (Wojcik et al., 2018) estimated that two-thirds of the tweeted links to popular web sites were made by suspected bots. Again, data-mining algorithms are hard-pressed to consider the relevance and quality of the data they mine.

Monte Carlo simulations

Monte Carlo simulations, named after the gambling mecca, were first used in the Manhattan Project in the 1940s (Metropolis, 1987). Today, they are widely employed in many probabilistic situations where an exact numerical solution cannot be derived mathematically.

For example, a financial planner might specify a person's initial wealth, age, spending, and investment strategies, and the model's parameters. In a Monte Carlo simulation of annual outcomes, a computer random number generator determines whether the person lives another year, how much she spends, and the return on her investments. The simulations might be run one million times in order to provide an estimate of the probability that she outlives her wealth and the probability distribution of her bequest. After the simulations are run with a variety of behavioral assumptions, the planner and the client can choose a strategy.

Here, I used Monte Carlo simulations to explore the perils of data mining. A total of 200 observations for each of *m* candidate explanatory variable were determined by computer-generated random draws from a normal distribution with mean 0 and standard deviation σ_x

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

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.

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

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

where the value of each β coefficient was randomly determined from a uniform distribution ranging from 2 to 4, with the range 0 to 2 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 $\sigma_x = 5$, $\sigma_y = 20$, but I also considered all combinations of $\sigma_x = 5$, 10, or 20, and $\sigma_y = 10$, 20, or 30. 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. One hundred thousand simulations were done for each parameterization of the model.

The central question is how effective the estimated model is at making reliable predictions with fresh data. So, in each simulation, 100 observations were used to estimate the model's coefficients, and the remaining 100 observations were used to test the model's reliability.

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.

A stepwise regression procedure was used to select the explanatory variables, one by one, with the lowest two-sided *p* values if the *p* value was less than 0.05. The results were not due to the use of stepwise regression, 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 when there are a large number of candidate explanatory variables (Bruce and Bruce, 2017; Cios et al., 2007; Hastie et al., 2016; Varian, 2014).

In one set of simulations, all of the candidate explanatory variables were nuisance variables. In the second set of simulations, five true variables were included among the candidate variables. The first set of simulations, with entirely spurious variables, considers the extent to which coincidental correlations with the dependent variable can create an illusion of a successful prediction model. The second set of simulations considers how well data mining 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 coefficient estimates.

Table 2 reports the results of simulations in which all of the candidate explanatory variables were nuisance variables. Every variable selected by the data-mining algorithm as being useful was actually useless, yet data mining consistently discovered a substantial number of variables that were highly correlated with the target variable. For example, with 100 candidate variables, the data-mining algorithm picked out, on average, 6.63 useless variables for making predictions.

Table 2. Simulations with no true variables.

Number of candidate variables	Average number of variables selected	Average in-sample correlation	Average out-of-sample correlation
5	1.11	0.244	0.000
10	1.27	0.258	0.000
50	3.05	0.385	0.000
100	6.63	0.549	0.000
500	97.79	1.000	0.000

Table 3. Simulations with five true variables.

Number of candidate variables	Average number of variables selected	Average in-sample correlation	Average out-of-sample correlation
5	4.50	0.657	0.606
10	4.74	0.663	0.600
50	6.99	0.714	0.543
100	10.71	0.780	0.478
500	97.84	1.000	0.266

Table 2 also shows that, as the number of candidate explanatory variables increases, so do the average number of nuisance variables selected. Regardless of how highly correlated these variables are with the target variable, they are, on average, completely useless for future predictions. The out-of-sample correlations average zero. However, some models, by luck alone, survived out-of-sample tests. In every case, approximately 5% of the models that were statistically significant in-sample were also statistically significant out-of-sample.

Table 3 shows the simulation results when the five true variables were among the candidate variables considered by the data-mining algorithm. The inclusion of five true variables did not eliminate the selection of nuisance variables; instead, it increased the number of variables selected. The larger the number of candidate variables, the more nuisance variables are included and the worse are the out-of-sample predictions. This is empirical confirmation of what might be called the paradox of big data:

> It would seem that having data for a large number of variables will help us find more reliable patterns; however, the more variables we consider, the less likely it is that what we find will be useful.

Notice, too, that when fewer nuisance variables are considered, fewer nuisance variables are selected. Instead of unleashing a data-mining algorithm on hundreds or thousands or hundreds of thousands of unfiltered variables, it would be better to use human expertise to exclude as many nuisance variables as possible. This is a corollary of the paradox of big data:

> The larger the number of possible explanatory variables, the more important is human expertise.

These simulations also document how a plethora of nuisance variables can crowd out true variables. With 100 candidate variables, for example, one or more true variables were crowded out 50% of the time, and two or more true variables were crowded out 16% of the time. There were even occasions when all five true variables were crowded out.

The bottom line is straightforward. Variables discovered through data mining can appear to be useful even when they're irrelevant, and can cause true variables to be overlooked and discarded. Both flaws undermine the usefulness of data mining.

Trump's tweets

A simple data-mining example is an exploration of Donald Trump's tweets (Trump Twitter Archive, 2019) during the 3-year period beginning on 9 November 2016, the day after his election as President of the United States.

Trump has 66 million Twitter followers and averaged 10.64 tweets a day during this 3-year period. He holds the most powerful office in the world, so perhaps his tweets have real consequences. I restricted my analysis to words that Trump used at least 100 times and also appeared in his tweets on at least 50 different days, and I ignored simple filler words like *a, an, it*, and *to*. I calculated the mean and standard deviation of his daily usage of each word and, from these, calculated the daily *Z*-value for each word.

I then used a 10-fold cross-validation data-mining algorithm to identify words whose daily *Z*-values could be used to predict various variables 1 to 5 days later. It turned out that the S&P 500 index of stock prices is predicted to be 97 points higher 2 days after a one-standard-deviation increase in the Trump's use of the word *president*, Table 4 shows

Table 4. Data-mining Donald Trump's tweets.

Dependent variable	Explanatory word	(Test MSE)/(training MSE)	Reduction in MSE (%)	Full-sample correlation
S&P 500 (+ 2)	President	1.0017	22.69	0.43
Moscow Low (+ 4)	Ever	1.0006	3.97	0.20
Pyongyang High (+ 5)	Wall	1.0021	5.04	−0.22
Urban Tea (+ 4)	With	1.0000	13.64	−0.32
RV (+ 5)	Democrat	1.0043	28.87	0.47

MSE: mean square error; RV: random variable.

that the ratio of the average test-period mean square error (MSE) to the training-period MSE was barely above 1 and that the use of the model reduced the MSE by 22.69% relative to simply using the average value of S&P 500 as a predictor. The correlation between the daily Z-values for *president* and the S&P 500 2 days later was a remarkable 0.43. The two-sided *p* value was essentially zero, though statistical tests with data-mined models are problematic.

We can surely concoct a plausible explanation for why Trump tweeting the word *president* affects the stock market a few days later, so I predicted a few other variables. Trump seems to admire Russian President Vladimir Putin (Somerlan, 2019; Yeung et al., 2019) and North Korean Chairman Kim Jong-un (Bierman and Stokols, 2018; Haltiwanger, 2019). Perhaps his tweets reverberate in these countries.

Using Weather Underground (2019) data, my 10-fold cross-validation data-mining algorithm discovered that the low temperature in Moscow is predicted to be 3.30°F higher 4 days after a one-standard-deviation increase in Trump's use of the word *ever*, and that the low temperature in Pyongyang is predicted to be 4.65°F lower 5 days after a one-standard-deviation increase in the use of the word *wall*.

We might concoct an explanation for these correlations too, perhaps related to the fact that temperatures in Moscow, Pyongyang, and the Eastern United States are related and that Trump's choice of words is influenced by the weather. Going farther afield, I considered the proverbial price of tea in China. I could not find daily data on tea prices in China, so I used the daily stock prices of Urban Tea, a tea product distributer headquartered in Changsha City, Hunan Province, China, with retail stores in Changsha and Shaoyang that sell tea and tea-based beverages. The data-mining algorithm found that Urban Tea's stock price is predicted to fall 4 days after Trump used the word *with* more frequently.

This data-mined correlation between *with* and Urban Tea's stock price might inspire a creative explanation, so I generated something even more difficult to explain after the fact—a random-walk random variable with the daily change in the value of the variable determined by random draws from a normal distribution with mean zero. The data-mining algorithm found that a one-standard deviation increase in Trump's use of the word *democrat* had a strong positive correlation with the value of this random variable 5 days later.

The intended lessons are how easy it is for data-mining algorithms to find transitory patterns and how tempting it is to think up explanations after the fact.

Here, I considered thousands of tweeted words, 19 dependent variables (the S&P 500 and the Dow Jones Industrial Average, Moscow daily high and low temperatures, Pyongyang daily high and low temperatures, Urban Tea stock returns and Jay Shree Tea stock returns, the number of runs scored by the Washington Nationals baseball team, and 10 random-walk variables), and lags of 1 to 5 days, and I only reported the most striking relationships. That is the nature of the beast we call data mining: seek and ye shall find.

Data-mined investment strategies

My data mining of Trump's tweets is not far-fetched. A Bank of America study (Franck, 2019) reported that the stock market does better on days when Trump tweets less. A JP Morgan study (Alloway, 2019) concluded that Trump tweets containing the words *China, billion, products, Democrats*, or *great* have statistically significant effects on interest rates.

Bolen et al. (2011) reported that a data-mining analysis of nearly 10 million Twitter tweets during the period February to December 2008 found that an upswing in "calm" words was often followed an increase in the Dow Jones average 6 days later. They looked at seven different Dow predictors: an assessment of positive versus negative moods and 6 mood states (calm, alert, sure, vital, kind, and happy). There is, no doubt, considerable noise in assigning mood states to various tweets. Is *nice* a calm, kind, or happy word? Is *yes!* an alert, sure, or vital word? The researchers also considered several different days into the future for correlating with the Dow. Finally, why did they use data from February to December 2008? What happened to January? With so much flexibility, data mining was bound to discover some coincidental patterns.

Preis et al. (2013) reported that they had found a novel way to time the stock market by using Google search data. They considered weekly data on the frequency with which users searched for 98 different keywords:

> We included terms related to the concept of stock markets, with some terms suggested by the Google Sets service, a tool which identifies semantically related keywords. The set of terms used was therefore not arbitrarily chosen, as we intentionally introduced some financial bias.

The use of the pejorative word *bias* is unfortunate since it suggests that there is something wrong with using search terms that are related to the stock market. The belief that correlation supersedes causation assumes that the way to discover new insights is to look for patterns unfettered by expert opinion—here, to discover ways to beat the stock market by looking for "unbiased" words that have nothing to do with stocks. The fatal flaw in such a blind strategy is that coincidental patterns will almost certainly be found, and data alone cannot distinguish between meaningful and meaningless patterns. If we have wisdom about something, it is generally better to use it—to introduce some financial expertise.

The researchers considered moving averages of 1 to 6 weeks for each of their 98 keywords and reported that the most successful stock trading strategy was based on the keyword *debt*, using a 3-week moving average and this decision rule:

> Buy the Dow if the momentum indicator is negative.
>
> Sell the Dow if the momentum indicator is positive.

Using data for the 7-year period 1 January 2004, through 22 February 2011, they reported that this strategy had an astounding 23.0% annual return, compared with 2.2% for a buy-and-hold strategy. Their conclusion:

> Our results suggest that these warning signs in search volume data could have been exploited in the construction of profitable trading strategies.

They offer no reasons:

> Future work will be needed to provide a thorough explanation of the underlying psychological mechanisms which lead people to search for terms like debt before selling stocks at a lower price.

The researchers considered 98 different keywords and 6 different moving averages (a total of 588 strategies). If they considered two trading rules (buying when the momentum indicator was positive *or* selling when the momentum indicator was positive), then 1,176 strategies were explored. With so many possibilities, some chance patterns would surely to be discovered—which undermines the credibility of those that were reported.

I tested their *debt* strategy for predicting the Dow over the next 7 years, from 22 February 2011, through 31 December 2018. Figure 1 shows the results. Their *debt* strategy had an annual return of 2.81%, compared with 8.60% for buy-and-hold.

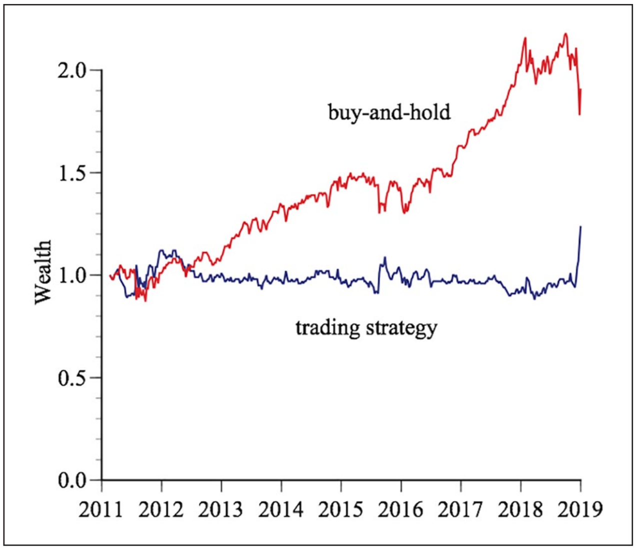

Figure 1. Clumsily staggering in and out of the market.

Equbot

In 2017, a company named Equbot launched AIEQ, which claimed to be the first exchange-traded fund (ETF) run by artificial intelligence. Equbot boasted that AIEQ removes "human error and bias from the process" by using IBM's Watson and genetic algorithms, fuzzy logic, and adaptive tuning (Equbot, 2019). How well did it perform? Figure 2 shows that AIEQ seems to be a "closet indexer," tracking the S&P 500 while underperforming it. From inception through 1 November 2019, AIEQ had a cumulative return of 18%, compared with 23% for the S&P 500.

Figure 3 compares the volume of trading in AIEQ to the volume of trading in the S&P 500, both scaled to equal 1 when AIEQ was launched. Once the disappointing results became apparent, customers lost interest.

The Voleon group

In 2008, Michael Kharitonov and Jon McAuliffe, with PhDs in computer science and statistics, respectively, started Voleon, an investment management firm that picked stocks based on a data-mining algorithm:

> 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." (Salmon and Stokes, 2010)

Voleon's algorithms reportedly sifted through massive amounts of data, including satellite images, credit card

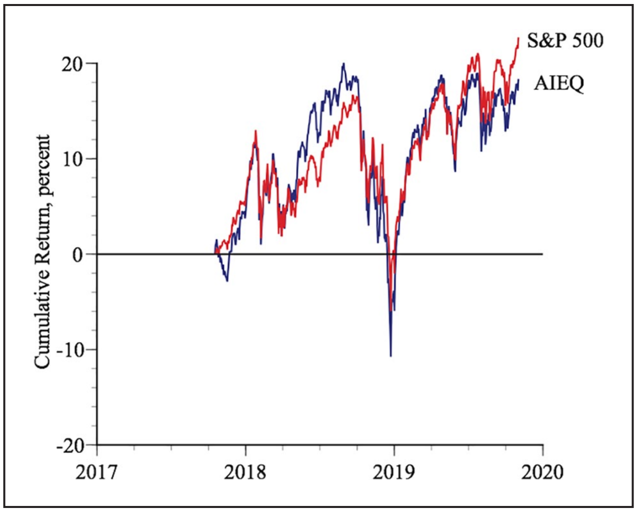

Figure 2. Underwhelming performance.

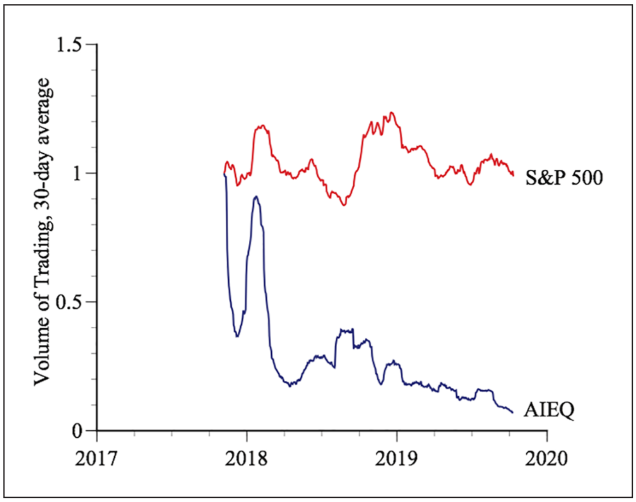

Figure 3. Overwhelming disinterest.

receipts, and social media language, looking for patterns related to stock prices.

Ten years after its launch, the *Wall Street Journal* reported that Voleon's annual return had been slightly worse than the S&P 500, with Kharitonov admitting, "Most of the things we've tried have failed" (Hope and Chung, 2017).

The *Journal* attributed Voleon's struggles to the fact that financial markets are "continually being affected by new events, the relationships among which are frequently shifting." This is a common excuse when data-mined patterns vanish—the world has changed. Perhaps, but an alternative explanation is that the statistical patterns that data-mining algorithms discover were fleeting because they were fortuitous. If an algorithm finds a correlation between stock prices and Google searches for the word *debt*, 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 relationship—just a transitory statistical correlation.

Bitcoin

Bitcoin is the most well-known cryptocurrency, a digital medium of exchange that operates independently of the central banking system. As an investment, bitcoins are pure speculation. Investors who buy bonds receive interest. Investors who buy stocks receive dividends. Investors who buy apartment buildings receive rent. The only way people who invest in bitcoins can make a profit is if they sell their bitcoins for more than they paid for them—and there is little reason to think that bitcoin prices are truly related to anything other than what Keynes called "animal spirits."

Nonetheless, using daily data for 1 January 2011 through 31 May 2018 Liu and Tsyvinski (2018) estimated 810 associations between bitcoin returns and various variables, such as the effect of bitcoin returns on stock returns in the beer, book, and automobile industries. (They also estimated hundreds of additional equations for two other cryptocurrencies, Ethereum and Ripple, and for various sub-periods of their data set.) The most sensible relationship of the many they consider is the effect of bitcoin returns on the number of Google searches for *bitcoin*. This relationship is not particularly useful, but at least there is a plausible explanation.

Perhaps the most unusual thing about this study is that the authors reported thousands of estimated relationships, not just those that were statistically significant. Overall, for the full-sample bitcoin data, they found 63 of the 810 estimated relationships (7.8%) to be statistically significant at the 10% level, somewhat fewer than would be expected if they had just correlated bitcoin returns with random numbers. They did not attempt to explain the correlations they found: "We don't give explanations, we just document this behavior."

Patterns without explanations are treacherous. A search for patterns in large databases will almost certainly discover some, and the coincidental patterns that are discovered are likely to vanish when the results are used to make predictions. What is the point of documenting temporary patterns that are likely to vanish?

Rosebeck and Smith (2019) used out-of-sample data from 1 June 2018 through 31 July 2019 to try to replicate 59 of the 63 statistically significant relationships that Liu and Tsyvinski reported (there were no out-of-sample data for 4 of the relationships) and found that 13 continued to be statistically significant out of sample. Five of these 13 persistently significant coefficients were the coefficients in five related equations that used bitcoin weekly returns to predict *bitcoin* searches in the same week, which is one of the few models that has a logical explanation. These five coefficients had the same signs in-sample and out-of-sample.

For the other eight persistently significant coefficients, two had the same sign in both periods and six had opposite signs. Should we conclude that, because bitcoin returns happened to have had a statistically significant negative effect on stock returns in the paperboard-containers-and-boxes industry that was confirmed with out-of-sample data, a useful, meaningful relationship has been discovered?

There are three lessons here: First, energetic data mining is certain to discover coincidental patterns. Second, relationships that have a logical foundation are more likely to be confirmed out-of-sample. Third, with a sufficient amount of data mining, some coincidental patterns will, by luck alone, persist out-of-sample.

Google flu trends

Google researchers created a data-mining program called Google Flu Trends that analyzed 50 million search queries and identified 45 key words that "can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day" (Ginsberg et al., 2009). An MIT professor praised the model:

> This seems like a really clever way of using data that is created unintentionally by the users of Google to see patterns in the world that would otherwise be invisible. I think we are just scratching the surface of what's possible with collective intelligence. (Helft, 2008)

However, after issuing its report, Google Flu Trends over-estimated the number of flu cases for 100 of the next 108 weeks, by an average of nearly 100% (Lazer et al., 2014). Google Flu Trends no longer makes flu predictions.

Messing with our lives

Many data-mining algorithms for screening job applicants, pricing car insurance, approving loan applications, and determining prison sentences have significant errors and biases that are not due to programmer mistakes and biases, but to data mining.

Gild developed data-mining software for evaluating applicants for software engineering jobs by monitoring their online activities (Peck, 2013). The chief scientist acknowledged that some of the factors chosen by its data-mining software do not make sense. For example, the software found that several good programmers in its database visited a particular Japanese manga site frequently, so it decided that people who visit this site are likely to be good programmers. The chief scientist said, "Obviously, it's not a causal relationship," but argued that it was still useful because Gild has 6 million software engineers in its database and there was a strong statistical correlation.

The chief scientist said that the company's algorithm selects dozens of variables, and constantly changes the variables selected as correlations come and go. She believes that the ever-changing list of variables demonstrates the model's power and flexibility. A more compelling interpretation is that the data-mining algorithm captures transitory coincidental correlations that are of little value. If these were causal relationships, they would not come and go. They would persist and be useful.

Was this firm's software successful in identifying good job candidates? A person who worked for the company for 3 years wrote that (Anonymous, 2016):

> Customers really, really hate the product. There are almost no customers that have an overall positive experience. This has been true for years, and management is not able to reimagine the company in a way that would let them fix that core problem.

Evaluating job applicants based on whether they visit certain web sites is potentially discriminatory. Similarly, an Amazon algorithm for evaluating job applicants discriminated against women who had gone to women's colleges or belonged to women's organizations because there were few women in the algorithm's data base of current employees (Dastin, 2018; Reuters, 2018).

In 2016, Admiral Insurance, Britain's largest car insurance company, planned to launch firstcarquote, which would base its car insurance rates on a data-mined analysis of an applicant's Facebook posts (Ruddick, 2016; Rudgard, 2016). One example the company cited was whether a person liked Michael Jordan or Leonard Cohen—which humans would recognize as ripe with errors and biases.

The Admiral advisor who designed the algorithm said:

> Our analysis is not based on any one specific model, but rather on thousands of different combinations of likes, words and phrases and is constantly changing with new evidence that we obtain from the data. As such our calculations reflect how drivers generally behave on social media, and how predictive that is, as opposed to fixed assumptions about what a safe driver may look like.

This claim was intended to show that the algorithm is flexible and innovative. What it actually reveals is that their data-mining algorithm identifies historical patterns, not useful predictors. The algorithm changes constantly because it has no logical basis and is continuously buffeted by short-lived correlations.

We never had the opportunity to see how this algorithm would have fared because, a few hours before the scheduled launch, Facebook announced that it would not allow Admiral to access Facebook data, citing its policy that "prohibits the use of data obtained from Facebook to make decisions about eligibility, including whether to approve or reject an application or how much interest to charge on a loan" (Cohn, 2016).

In 2017, the founder and CEO of a Chinese tech company reported that they had developed a data-mining algorithm that evaluates loan applications based on an analysis of the usage of hundreds of millions of smartphones in China (Yuan, 2017):

> We don't need human beings to tell us who's a good customer and who's bad. Technology is our risk control.

Among the data that show up as evidence of a person being a good credit risk: using an Android phone instead of an iPhone; not answering incoming calls; having outgoing calls not answered, and not keeping the phone fully charged.

We could invent plausible theories to explain the discovered statistical patterns. Or, if these patterns had been reversed, indicators of being a bad credit risk, we could invent reasonable explanations for that too. That's the thing about making up theories after the fact—we are clever enough to invent plausible stories for whatever statistical patterns are found, even if the statistical patterns are random noise discovered by data-mining software. Finding patterns proves nothing. Making up stories to fit the patterns proves nothing. (If you were wondering, the data-mining software found these particular patterns to be correlated with being a bad credit risk.)

Richard Berk has appointments in the Department of Criminology and the Department of Statistics at the University of Pennsylvania. One of his specialties is algorithmic criminology, which is becoming increasingly common in pre-trial bail determination, post-trial sentencing, and post-conviction parole decisions. Berk (2013) writes, "The approach is 'black box,' for which no apologies are made." In an article in *The Atlantic*, Berk is more explicit: "If I could use sun spots or shoe size or the size of the wristband on their wrist, I would. If I give the algorithm enough predictors to get it started, it finds things that you wouldn't anticipate" (Labi, 2012).

Most "things that you wouldn't anticipate" are things that do not make sense, like sun spots, shoe sizes, and wristband sizes. They reflect temporary coincidences that are useless predictors of criminal behavior. Angwin et al. (2016) analyzed one of the most popular risk-assessment algorithms and found that only 20% of the people predicted to commit violent crimes within 2 years actually did so—and that the predictions discriminate against black defendants.

Berk no doubt has good intentions, but it is unsettling that he thinks people should be paroled or remain incarcerated based on sunspots, shoes, and wristbands. That's what happens when you trust data-mining algorithms too much.

If bail, sentencing, and parole decisions are based on data mining, it is just a short step to using data-mined models to decide who should be arrested and imprisoned. Sure enough, Wu and Zhang (2016, 2017) reported that they could predict with 89.5% accuracy whether a person is a criminal by applying their AI algorithm to scanned facial photos.

They argued:

> Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggages, having no emotions, no biases whatsoever due to past experience, race, religion, political doctrine, gender, age, etc., no mental fatigue, no preconditioning of a bad sleep or meal.

They scanned 1856 male ID photos—730 criminals and 1126 non-criminals—and their data-mining program found "some discriminating structural features for predicting criminality, such as lip curvature, eye inner corner distance, and the so-called nose-mouth angle."

An article in the *MIT Technology Review* (Emerging Technology from the arXiv, 2016) was optimistic: "All this heralds a new era of anthropometry, criminal or otherwise. . . . And there is room for more research as machines become more capable." Vorhees (2016) wrote, "the study has been conducted with rigor. The results are what they are." Spoken like a true data miner. Who needs theories? If a data-mining algorithm finds statistical patterns, that's proof enough. The results are what they are. Up is up.

To their great credit, data scientists, as a whole, dismissed this study as perilous pseudoscience—unreliable and misleading, with potentially dangerous consequences if taken seriously. However, a blogger (cageymaru, 2016) argued:

> What if they just placed the people that look like criminals into an internment camp? What harm would that do? They would just have to stay there until they went through an extensive rehabilitation program. Even if some went that were innocent; how could this adversely affect them in the long run?

If such blind faith in data mining becomes the norm, governments may well start imprisoning people based on data-mined facial analyses.

Discussion

Artificial intelligence (AI) systems often rely on data-mining algorithms to specify and parameterize models. Such algorithms have no effective way of assessing the plausibility of what they discover because computers are not intelligent in any meaningful sense of the word (Smith, 2018). Consider, for example, the challenges identified by Terry Winograd that have come to be known as Winograd schemas (Davis, 2018). What does the word *it* refer to in this sentence?

> I can't cut that tree down with that axe; it is too [thick/small].

Humans know that if the bracketed word is *thick*, then *it* refers to the tree and, if the bracketed word is *small*, then *it* refers to the axe. Winograd schemas are very difficult for computers because they do not understand what words mean. They do not know what *tree, axe, cut down, thick*, or *small* mean, or how they might be related.

There is a Winograd Schema Challenge with a $25,000 prize for a computer program that is 90% accurate in interpreting Winograd schemas (Levesque et al., 2012). In the 2016 competition, the expected value of the score for guessing was 44% correct (some schemas had more than two possible answers). The highest computer score was 58% correct, the lowest 32%, a variation that may have been due more to luck than to differences in the competing programs' abilities. Computers are like New-Zealand-born Nigel Richards who has won the French-language Scrabble World Championship twice without knowing the meaning of the words he spells.

How could a data-mining algorithm interpret a correlation between Trump tweeting the word *with* and the price of Urban Tea stock 4 days later when computer algorithms do not know what any of the words mean and have no understanding of what might cause stock prices to go up or down?

Computer image-recognition software is similarly brittle because it identifies and matches pixel patterns without any understanding of the image formed by the pixels. Putting graffiti on a photograph of a stop sign or even changing a few pixels in a picture of a stop sign—alterations that would not be noticed by humans—can cause state-of-the-art deep neural networks to fail miserably (Evtimov et al., 2017; Su et al., 2017). Data-mining pixels is not the same as knowing what a stop sign is.

Nguyen et al. (2015) demonstrated something even more surprising. In addition to making nothing out of something (like a computer not recognizing a stop sign), computers can make something out of nothing by misinterpreting meaningless images as real objects. For example, a powerful image-recognition program was 99% certain that a horizontal sequence of black and yellow lines was a school bus, completely ignoring the fact that there were no wheels, door, or windows in the picture.

Sharif et al. (2016) reported that the state-of-the-art deep neural network programs used in facial biometric systems can be fooled by people wearing colorful eyeglass frames. One of the authors, a white male, was misidentified as Milla Jovovich, a white female, 88% of the time, and another author, a 24-year-old Middle Eastern male, was misidentified as Carson Daly, a 43-year-old white male, 100% of the time—all because the eyeglass frame colors led the computer program astray. Humans do not make such mistakes because we know what eyeglass frames are, and we know that we should look past the frames to identify the person we see. Computers know none of this; they just match pixels as best they can.

The reproducibility crisis (Baker, 2017; Ioannidis, 2005; Pashler and Wagenmakers, 2012), in which attempts to replicate research findings often fail, may be partly due to the uncritical acceptance of data-mining discoveries as real phenomena. The crisis might be partly abated by recognizing the fact that data-mined coincidences are inevitably temporary.

Conclusion

Data-mining algorithms—often operating under the label artificial intelligence—are now widely used to discover statistical patterns. However, in large data sets streaks, clusters, correlations, and other patterns are the norm, not the exception. While data mining might discover a useful relationship, the number of possible patterns that can be spotted relative to the number that are genuinely useful has grown exponentially—which means that the chances that a discovered pattern is useful is rapidly approaching zero. This is the paradox of big data:

> It would seem that having data for a large number of variables will help us find more reliable patterns; however, the more variables we consider, the less likely it is that what we find will be useful.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Gary Smith https://orcid.org/0000-0002-5173-2741

References

Alloway T (2019) JPMorgan creates "Volfefe" index to track Trump tweet impact. *Bloomberg.Com*. Available at: https://www.bloomberg.com/news/articles/2019-09-09/jpmorgan-creates-volfefe-index-to-track-trump-tweet-impact (accessed 1 December 2019).

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete? *Wired*, 23 June. Available at: https://www.wired.com/2008/06/pb-theory/ (accessed 21 April 2020).

Angwin J, Larson J, Mattu S, et al. (2016) Machine bias. *ProPublica*, 23 May. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 21 April 2020).

Anonymous (2016) Gild review. *Glassdoor*, 20 March. Available at: https://www.glassdoor.com/Reviews/Gild-Reviews-E459358.htm (accessed 21 April 2020).

Arnott RD, Harvey CR and Markowitz H (2018) A backtesting protocol in the era of machine learning, 21 November, p. 18.

Athey S (2018) The impact of machine learning on economics. In: Agrawal A, Gans J and Avi G (eds) *The Economics of Artificial Intelligence: An Agenda*. Chicago, IL: University of Chicago Press, pp. 507–547.

Baker M (2017) 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604): 452–454.

Begoli E and Horsey J (2012) Design principles for effective knowledge discovery from big data. In: *2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture*, Helsinki, Finland, 20–24 August 2012. New York: IEEE, pp. 215–218.

Berk R (2013) Algorithmic criminology. *Security Informatics* 2: 5.

Bierman N and Stokols E (2018) Trump voices admiration and envy of Kim Jong Un, underscoring his respect for autocrats. *Los Angeles Times*, 15 June. Available at: https://www.latimes.com/politics/la-na-pol-trump-kim-values-20180615-story.html

Bolen J, Mao H and Zeng X (2011) Twitter mood predicts the stock market. *Journal of Computational Science* 2(1): 1–8.

Bruce P and Bruce A (2017) *Practical Statistics for Data Scientists: 50 Essential Concepts*. Newton, MA: O'Reilly Media. p. 250.

Brynjolfsson E, Geva T and Reichman S (2016) Crowd-squared: Amplifying the predictive power of search trend data. *MIS Quarterly* 40(4): 941–961.

cageymaru (2016) post, HardForum Tech News, 21 November. Available at: https://hardforum.com/threads/new-program-judges-if-youre-a-criminal-from-your-facial-features.1917912/ (accessed 21 April 2020).

Calude CS and Longo G (2017) The deluge of spurious correlations in big data. *Foundations of Science* 22(3): 595–612.

Cios KJ, Pedrycz W, Swiniarski RW, et al. (2007) *Data Mining: A Knowledge Discovery Approach*. New York: Springer.

Cohn C (2016) Facebook stymies Admiral's plans to use social media data to price insurance premiums. *Reuters*, 2 November. Available at: https://www.reuters.com/article/us-insurance-admiral-facebook/facebook-stymies-admirals-plans-to-use-social-media-data-to-price-insurance-premiums-idUSKBN12X1WP (accessed 21 April 2020).

Coase R (1988) How should economists choose? In: *Ideas, Their Origins and Their Consequences: Lectures to Commemorate the Life and Work of G. Warren Nutter*. Thomas Jefferson Center Foundation. Washington, DC: American Enterprise: Institute for Public Policy Research, pp. 63–79.

Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 9 October. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (accessed 21 April 2020).

Davis E (2018) Collection of Winograd schemas. Available at: https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.html (accessed 21 April 2020).

Egami N, Fong CJ, Grimmers J, et al. (2018) How to make causal inferences using text. 15 October. Available at: https://arxiv.org/pdf/1802.02163.pdf (accessed 21 April 2020).

Emerging Technology from the arXiv (2016) Neural network learns to identify criminals by their faces. *MIT Technology Review*, 22 November. Available at: https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/ (accessed 21 April 2020).

Equbot (2019) Available at: https://equbot.com (accessed 21 April 2020).

Evtimov I, Eykholt K, Fernandes E, et al. (2017) Robust physical-world attacks on deep learning models. Available at: https://arxiv.org/abs/1707.08945 (accessed 21 April 2020).

Fayyad U, Piatetsky-Shapiro G and Smyth P (1996) From data mining to knowledge discovery in databases. *AI Magazine* 17(3): 37–54.

Franck T (2019) On days when President Trump tweets a lot, the stock market falls, investment bank finds. *CNBC*. Available at: https://www.cnbc.com/2019/09/03/on-days-when-president-trump-tweets-a-lot-the-stock-market-falls-investment-bank-finds.html (accessed 21 April 2020).

Ginsberg J, Mohebbi MH, Patel RS, et al. (2009) Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–1014.

Grover V and Lyytinen K (2015) New state of play in information systems research: The push to the edges. *MIS Quarterly* 39(2): 271–296.

Guo X, Wei Q, Chen G, et al. (2017) Extracting representative information on intra-organizational blogging platforms. *MIS Quarterly* 41(4): 1105–1127.

Haltiwanger J (2019) Trump calls North Korea's Kim Jong Un, who's threatened the US with nuclear war, a "great leader." *Business Insider*, 27 February. Available at: https://www.businessinsider.my/page/5769?m&jwsource=cl (accessed 21 April 2020).

Hastie T, Tibshirani R and Friedman J (2016) *The Elements of Statistical Learning* (2nd edn). New York: Springer.

Helft M (2008) Google uses searches to track flu's spread. *The New York Times*, 11 November. Available at: https://www.nytimes.com/2008/11/12/technology/internet/12flu.html (accessed 21 April 2020).

Hope B and Chung J (2017) The future is bumpy: High-tech hedge fund hits limits of robot stock picking, wall street journal. 17 December. Available at: https://www.wsj.com/articles/the-future-is-bumpy-high-tech-hedge-fund-hits-limits-of-robot-stock-picking-1513007557 (accessed 21 April 2020).

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

Jafar M, Babb J and Abdullat A (2017) Emergence of data analytics in the information systems curriculum. *Information Systems Education Journal* 15: 22–36.

Kecman V (2007) Forward. In: Cios KJ, Pedrycz W, Swiniarski RW, et al. (eds) *Data Mining: A Knowledge Discovery Approach*. New York: Springer, p. xi.

Labi N (2012) Misfortune teller. *The Atlantic*, January–February. Available at: https://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846/ (accessed 21 April 2020).

Lazer D, Kennedy R, King G, et al. (2014) The parable of Google flu: Traps in big data analysis. *Science* 343(6176): 1203–1205.

Levesque HJ, Davis E and Morgenstern L (2012) The Winograd schema challenge. In: *KR 2012: 13th International Conference on the Principles of Knowledge Representation and Reasoning*, Rome, 10–14 June 2012, pp. 552–561. Ultimo NSW, Australia: UT Sydney.

Liu Y and Tsyvinski A (2018) *Risks and returns of cryptocurrency*. NBER working paper no. 24877, 13 August. Available at: https://ssrn.com/abstract=3226952 (accessed 21 April 2020).

Martens D, Provost F, Clark J, et al. (2016) Mining massive fine-grained behavior data to improve predictive analytics. *MIS Quarterly* 40(4): 869–888.

Metropolis N (1987) The beginning of the Monte Carlo method. *Los Alamos Science* 15: 125–130.

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

Nguyen A, Yosinski J and Clune J (2015) Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8-10 June 2015, Boston, MA.

Pashler H and Wagenmakers EJ (2012) Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* 7(6): 528–530.

Peck D (2013) They're watching you at work. *Atlantic*, December. Available at: https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/

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

Preis T, Moat HS and Stanley HE (2013) Quantifying trading behavior in financial markets using Google trends. *Scientific Reports* 3: 1684.

Reuters (2018) Amazon ditched AI recruiting tool that favored men for technical jobs. *The Guardian*, 10 October. Available at: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine (accessed 21 April 2020).

Rosebeck O and Smith G (2019) The reproducibility crisis: A case study. Working paper, Pomona College, Claremont, CA, July.

Ruddick G (2016) Admiral to price car insurance based on Facebook posts. *The Guardian*, 1 November. Available at: https://www.theguardian.com/technology/2016/nov/02/admiral-to-price-car-insurance-based-on-facebook-posts (accessed 21 April 2020).

Rudgard O (2016) Admiral to use Facebook profile to determine insurance premium. *The Telegraph*, 2 November. Available at: https://www.telegraph.co.uk/insurance/car/insurer-trawls-your-facebook-profile-to-see-how-well-you-drive/ (accessed 21 April 2020).

Sagiroglu S and Sinanc D (2013) Big data: A review. In: *Proceedings of 2013 international conference on collaboration technologies and systems (CTS)*, San Diego, CA, 20–24 May 2013.

Salmon F and Stokes J (2010) Algorithms take control of wall street. *Wired*, 27 December. Available at: https://www.wired.com/2010/12/ff-ai-flashtrading/ (accessed 21 April 2020).

Sharif M, Bhagavatula S, Bauer L, et al. (2016) Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, 24–28 October 2016, pp. 1528–1540. New York: ACM.

Shi Z, Lee GM and Whinston AB (2016) Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quarterly* 4(4): 1035–1056.

Smith G (2018) *The AI Delusion*. Oxford: Oxford University Press.

Smith G and Cordes J (2019) *The 9 Pitfalls of Data Science*. Oxford: Oxford University Press.

Somerlan J (2019) Donald Trump's gushing praise of Vladimir Putin under fresh scrutiny after Michael Cohen allegations. *Independent*, 18 January. Available at: https://www.independent.co.uk/news/world/americas/us-politics/trump-cohen-putin-russia-investigation-mueller-congress-fbi-a8734231.html (accessed 21 April 2020).

Stevenson PW (2016) Professor who predicted 30 years of presidential elections correctly called a Trump win in September. *The Washington Post*, 8 November. Available at: https://www.washingtonpost.com/news/the-fix/wp/2016/10/28/professor-whos-predicted-30-years-of-presidential-elections-correctly-is-doubling-down-on-a-trump-win/ (accessed 21 April 2020).

Su J, Vargas DV and Kouichi S (2017) One pixel attack for fooling deep neural networks. November. Available at: https://arxiv.org/abs/1710.08864 (accessed 21 April 2020).

Tobin J (1972) Personal communication.

Trump Twitter Archive (2019) Available at: http://www.trumptwitterarchive.com (accessed 21 April 2020).

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.

Vorhees W (2016) Has AI gone too far? Automated inference of criminality using face images. *Data Science Central*, 29 November. Available at: https://www.datasciencecentral.com/profiles/blogs/has-ai-gone-too-far-automated-inference-of-criminality-using-face (accessed 21 April 2020).

Weather Underground (2019) Available at: https://www.wunderground.com (accessed 21 April 2020).

Wojcik S, Messing S, Smith A, et al. (2018) Bots in the Twittersphere. *Pew Research Center*, 18 April. Available online: https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/ (accessed 21 April 2020).

Wu X and Zhang X (2016) Automated inference on criminality using face images. *Shanghai Jiao Tong University*, 21 November. Available at: https://arxiv.org/abs/1611.04135v1 (accessed 21 April 2020).

Wu X and Zhang X (2017) Responses to critiques on machine learning of criminality perceptions. *Shanghai Jiao Tong University*, 26 May. Available at: https://arxiv.org/abs/1611.04135v3 (accessed 21 April 2020).

Yeung J, Westcott B, Liptak K, et al. (2019) G20 summit 2019: Trump meets leaders in Osaka. *CNN*, 29 June. Available online: https://www.cnn.com/politics/live-news/g20-june-2019-intl-hnk/index.html (accessed 21 April 2020).

Yuan L (2017) Want a loan in China? Keep your phone charged. *The Wall Street Journal*, 6 April. Available online: https://www.wsj.com/articles/want-a-loan-in-china-keep-your-phone-charged-1491474250 (accessed 21 April 2020).

Author biography

Gary Smith is the author of more than 90 papers and 15 books, most recently *The AI Delusion* (Oxford 2018), *The 9 Pitfalls of Data Science* (Oxford 2019, co-authored with Jay Cordes and winner of the PROSE award for Excellence in Popular Science & Popular Mathematics), and *The Phantom Pattern Problem: The Mirage of Big Data*.