BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

What Everybody Needs to Know About 'Prediction' in Machine Learning

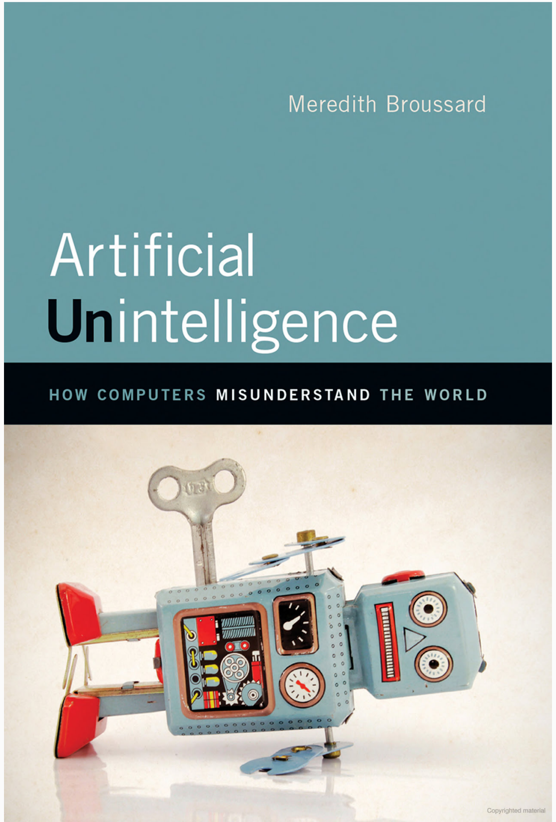
Momin M. Malik, PhD <momin_malik@cyber.harvard.edu>
Data Science Postdoctoral Fellow
Berkman Klein Center for Internet & Society at Harvard University

Leverhulme Centre for the Future of Intelligence, University of Cambridge, 3 December 2018
v1.1, updated 26 December 2018

Slides: https://mominmalik.com/cfi.pdf

Existential threats, or myths?

Solid general resource



- Read Ch. 7, "Machine Learning: The DL on ML"
  - (Two mistakes; see https://mominmalik.com/broussard)
- If you have time, read all of Part II (Ch. 5-9)
- Also, a useful story in Ch. 3, "Hello, AI"
  - "So, it's not real AI?" he asked.
  - "Oh, it's real," I said. "And it's spectacular. But you know, don't you, that there's no simulated person inside the machine? Nothing like that exists. It's computationally impossible."
  - His face fell. "I thought that's what AI meant," he said. "I heard about IBM Watson, and the computer that beat the champion at Go, and self-driving cars. I thought they invented real AI."

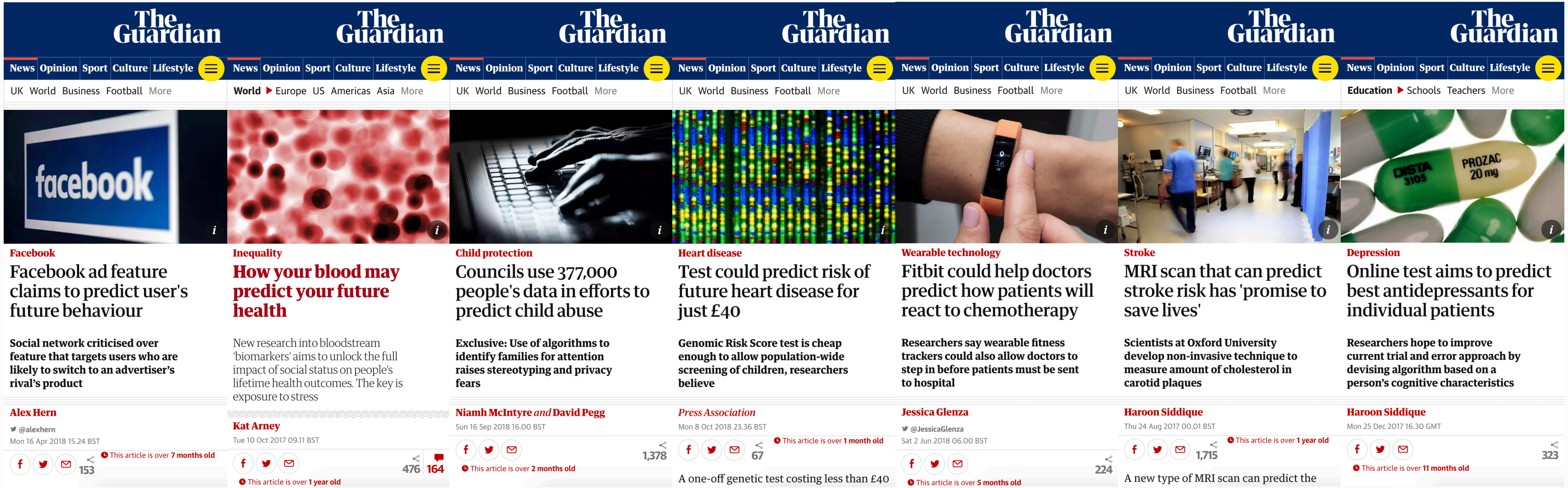
BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The things everybody needs to know

- Language: 'Prediction' (technical term) is not prediction (colloquial term); prediction is prospective, 'prediction' is retrospective.
- Definitions: 'Prediction' is based on correlations
- Validity: Correlations can *overfit*, and cross-validation only partially addresses
- Paradox: The *bias-variance tradeoff* (a consequence of the definition) makes it possible for a 'false' model to predict better than a 'true' one

Language: 'Prediction' is not prediction

Lots of "predict..."


Facebook
Facebook ad feature claims to predict user's future behaviour
Social network criticised over feature that targets users who are likely to switch to an advertiser's rival's product
Alex Hern
@alexhern
Mon 16 Apr 2018 15.24 BST
This article is over 7 months old

Inequality
How your blood may predict your future health
New research into bloodstream 'biomarkers' aims to unlock the full impact of social status on people's lifetime health outcomes. The key is exposure to stress
Kat Arney
Tue 10 Oct 2017 09.11 BST
This article is over 1 year old

Child protection
Councils use 377,000 people's data in efforts to predict child abuse
Exclusive: Use of algorithms to identify families for attention raises stereotyping and privacy fears
Niamh McIntyre and David Pegg
Sun 16 Sep 2018 16.00 BST
This article is over 2 months old

Heart disease
Test could predict risk of future heart disease for just £40
Genomic Risk Score test is cheap enough to allow population-wide screening of children, researchers believe
Press Association
Mon 8 Oct 2018 23.36 BST
This article is over 1 month old
A one-off genetic test costing less than £40

Wearable technology
Fitbit could help doctors predict how patients will react to chemotherapy
Researchers say wearable fitness trackers could also allow doctors to step in before patients must be sent to hospital
Jessica Glenza
@JessicaGlenza
Sat 2 Jun 2018 06.00 BST
This article is over 5 months old

Stroke
MRI scan that can predict stroke risk has 'promise to save lives'
Scientists at Oxford University develop non-invasive technique to measure amount of cholesterol in carotid plaques
Haroon Siddique
Thu 24 Aug 2017 00.01 BST
This article is over 1 year old
A new type of MRI scan can predict the

Depression
Online test aims to predict best antidepressants for individual patients
Researchers hope to improve current trial and error approach by devising algorithm based on a person's cognitive characteristics
Haroon Siddique
Mon 25 Dec 2017 16.30 GMT
This article is over 11 months old

- If you relied on *The Guardian*, what sort of picture might you get?

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

- Introduction
- Language: 'Prediction' is retrospective
- Definitions: 'Prediction' is correlation
- Validity: Correlations can overfit
- Paradox: 'Truth' may not predict
- Summary
- References

Predict... the future?

Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

Mar 2010

***Abstract*—In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from**

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter [1], a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of

Predicting the Future — Big Data, Machine Learning, and Clinical Medicine

Ziad Obermeyer, M.D., and Ezekiel J. Emanuel, M.D., Ph.D.

By now, it's almost old news: big data will transform medicine. It's essential to remember, however, that data by themselves are useless. To be useful, data must be analyzed, interpreted, and acted on. Thus, it is algorithms — not data sets — that will prove transformative. We believe, therefore, that attention has to shift to new statistical tools from the field of machine learning that will be critical for anyone practicing medicine in the 21st century.

First, it's important to understand what machine learning is not. Most computer-based algorithms in medicine are "expert systems" — rule sets encoding knowledge on a given topic, which are applied to draw conclusions

1216 N ENGL J MED 375;13 NEJM.ORG SEPTEMBER 29, 2016

OED Oxford English Dictionary
The definitive record of the English language

predict, *v.*

Pronunciation: Brit. /prɨˈdɪkt/, U.S. /priˈdɪk(t)/, /prəˈdɪk(t)/

Forms: 15–16 **praedict**, 16– **predict**.

Frequency (in current use): ●●●●●●○○

Origin: A borrowing from Latin. **Etymon:** Latin *praedict-*.

Etymology: < classical Latin *praedict-*, past participial stem of *praedīcere* to say beforehand, to give warning of, to foretell, prophesy, to appoint beforehand, to prescribe, recommend, to advise < *prae-* PRE- *prefix* + *dīcere* to say, tell (see DICTUM *n.*). Compare Middle French, French *prédire* (c1170 in Old French in sense 'to ordain', c1430 in sense 'to foretell'). Compare earlier PREDICTED *adj.*

1. *transitive.*

a. To state or estimate, esp. on the basis of knowledge or reasoning, that (an action, event, etc.) will happen in $\boxed{\text{the future}}$ or will be a consequence of something; to forecast, foretell, prophesy. Also with clause as object.

"the future" is already in the definition!

1590—2003

b. Of a theory, observation, scientific law, etc.: to have as a deducible or inferable consequence; to imply.

1886—2002

2. *intransitive.* To make a prediction or predictions; to prophesy.

1652—2005

OXFORD UNIVERSITY PRESS

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

'Prediction' is not prediction!

"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"

A Balanced Survey on Election Prediction using Twitter Data

Daniel Gayo-Avello
dani@uniovi.es
@PFCdgayo
Department of Computer Science - University of Oviedo (Spain)

May 1, 2012

Abstract

Predicting X from Twitter is a popular fad within the Twitter research subculture. It seems both appealing and relatively easy. Among such kind of studies, electoral prediction is maybe the most attractive, and at this moment there is a growing body of literature on such a topic.

This is not only an interesting research problem but, above all, it is extremely difficult. However, most of the authors seem to be more interested in claiming positive results than in providing sound and reproducible methods.

"*It's not prediction at all!* I have not found a single paper predicting a future result. All of them claim that a prediction could have been made; i.e. they are *post-hoc* analysis and, needless to say, negative results are rare to find."

"Wishful mnemonics" of AI

ARTIFICIAL INTELLIGENCE MEETS NATURAL STUPIDITY

Drew McDermott
MIT AI Lab Cambridge, Mass 02139

As a field, artificial intelligence has always been on the border of respectability, and therefore on the border of crackpottery. Many critics <Dreyfus, 1972>, <Lighthill, 1973> have urged that we are over the border. We have been very defensive toward this charge, drawing ourselves up with dignity when it is made and folding the cloak of Science about us. On the other hand, in private, we have been justifiably proud of our ideas, because pursuing them is the onl

Unfortunately, the necessity for s the culture of the hacker in computer to cripple our self-discipline. In a young field, self-discipline is not necessarily a virtue, but we are not getting any younger. In the past few years, our tolerance of sloppy thinking has led us to repeat many mistakes over and over. If we are to retain any credibility, this should stop.

This paper is an effort to ridicule some of these mistakes. Almost everyone I know should find himself the target at some point or other; if you don't, you are encouraged to write up your own favorite fault. The three described here I suffer from myself. I hope self-ridicule will be a complete catharsis, but I doubt it. Bad ... though, if we can't ...le. ...m crowd for

Wishful Mnemonics

Wishful Mnemonics

A major source of simple-mindedness in AI programs is the use of mnemonics like "UNDERSTAND" or "GOAL" to refer to programs and data structures. This practice has been inherited from more

Compare the mnemonics in Planner <Hewitt,1972> with those in Conniver <Sussman and McDermott, 1972>:

Planner	Conniver
GOAL	FETCH & TRY-NEXT
CONSEQUENT	IF-NEEDED
ANTECEDENT	IF-ADDED
THEOREM	METHOD
ASSERT	ADD

It is so much harder to write programs using the terms on the right! When you say (GOAL . . .), you can just feel the enormous power at your fingertips. It is, of course, an illusion.

When you say (GOAL . . .), you can just feel the enormous power at your fingertips. It is, of course, an illusion.

1965> What if atomic symbols had been called "concepts", or CONS had been called ASSOCIATE? As it is, the programmer has no debts to pay to the system. He can build whatever he likes. There are some minor faults; "property lists" are a little risky; but by now the term is sanitized.

Resolution theorists have been pretty good about wishful mnemonics. They thrive on hitherto meaningless words like RESOLVE and PARAMODULATE, which can only have their humble, technical meaning. There are actually quite few pretensions in the resolution literature. <Robinson, 1965> Unfortunately, at the top of their intellectual edifice stand the word "deduction". This is *very* wishful, but not entirely their fault. The logicians who first misused the term (e.g., in the "deduction" theorem) didn't have our problems; pure resolution theorists don't either. Unfortunately, too many AI researchers took them at their word and assumed that deduction, like payroll processing, had been tamed.

Of course, as in many such cases, the only consequence in the long run was that "deduction" changed in meaning, to become something narrow, technical, and not a little sordid.

Page 4

SIGART Newsletter No. 57 April 1976

Proposal: More precise language

- ~~Predict the likelihood~~: Calculate the likelihood
- ~~Predict the risk~~, ~~predict the probability~~: Estimate the risk, estimate the probability
- ~~Prediction~~, ~~predicted~~: Fitted value, fitted
- ~~We predict~~: We detect, we classify, we model
- ~~X predicts Y~~: X is correlated with Y
- ~~X predicts Y, *ceteris paribus*~~ (partial correlation): X is associated with Y

Proposal: Use alternatives

- Retrodiction
- Backtesting (retrodiction for testing)
- Hindcasting (backtesting for forecasting)

In-sample	vs.	Out of-sample
Interpolation	vs.	Extrapolation
Diagnosis	vs.	Prognosis
Retrospective	vs.	Prospective

(Language not enough: *mechanics* matter)

Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu

Another thing I must point out is that you cannot prove a vague theory wrong. [...] Also, if the process of computing the consequences is indefinite, then with a little skill any experimental result can be made to look like the expected consequences.

"training set" in the machine-learning literature). The OOS performance is simulated over a sample not used in the design of the strategy (a.k.a. "testing set"). A backtest is *realistic* when the IS performance

(I.e., using "backtest" in place of "predict" has not prevented financial analysts from unwitttting overfitting)

Definitions: 'Prediction' is correlation, not causation

Prediction is correlation

- Prediction = "Fitted value" minimizing *loss*
- $L(y, f(x)) = (y - f(x))^2$
- Spurious (non-causal) correlations can *fit* really well!
- But such fits fall apart if the context changes (Google Flu Trends)

POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional

ability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers,

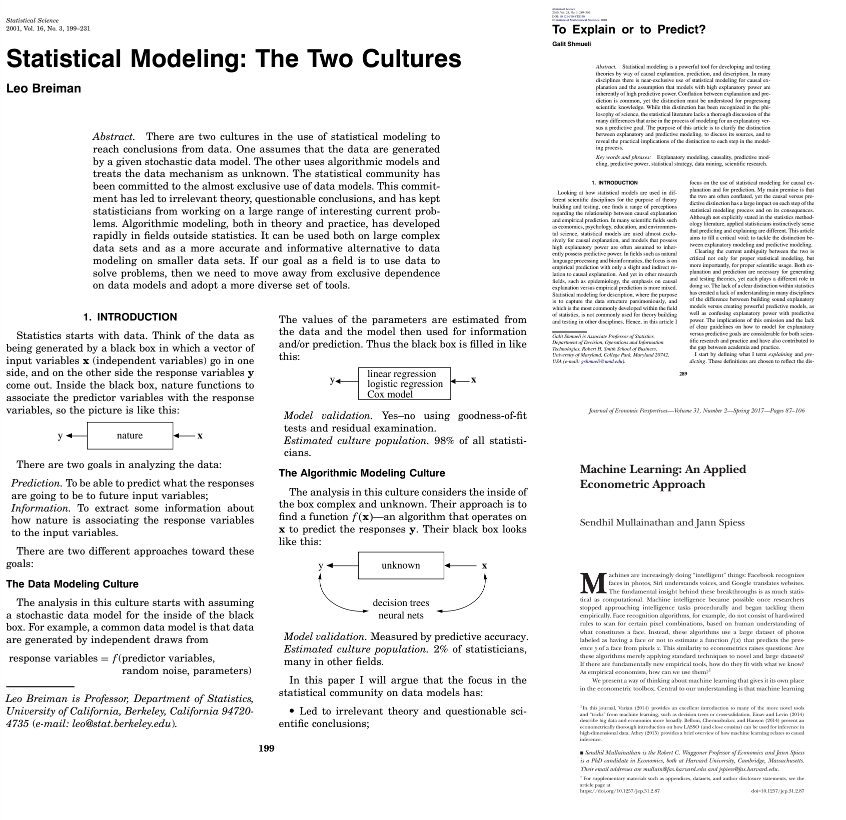
run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influ-

"To explain or to predict?"



Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

y ← nature ← x

There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-4735 (e-mail: leo@stat.berkeley.edu).

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

y ← linear regression / logistic regression / Cox model ← x

Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:

y ← unknown ← x
decision trees
neural nets

Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;

199

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330
© Institute of Mathematical Statistics, 2010

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

Key words and phrases: Explanatory modeling, causality, predictive modeling, predictive power, statistical strategy, data mining, scientific research.

1. INTRODUCTION

Looking at how statistical models are used in different scientific disciplines for the purpose of theory building and testing, one finds a range of perceptions regarding the relationship between causal explanation and empirical prediction. In many scientific fields such …

focus on the use of statistical modeling for causal explanation and for prediction. My main premise is that the two are often conflated, yet the causal versus predictive distinction has a large impact on each step of the statistical modeling process and on its consequences. Although not explicitly stated in the statistics method- …

289

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 87–106

Machine Learning: An Applied Econometric Approach

Sendhil Mullainathan and Jann Spiess

Machines are increasingly doing "intelligent" things: Facebook recognizes faces in photos, Siri understands voices, and Google translates websites. The fundamental insight behind these breakthroughs is as much statistical as computational. Machine intelligence became possible once researchers stopped approaching intelligence tasks procedurally and began tackling them empirically. Face recognition algorithms, for example, do not consist of hard-wired rules to scan for certain pixel combinations, based on human understanding of what constitutes a face. Instead, these algorithms use a large dataset of photos labeled as having a face or not to estimate a function $f(x)$ that predicts the presence y of a face from pixels x. This similarity to econometrics raises questions: Are these algorithms merely applying standard techniques to novel and large datasets? If there are fundamentally new empirical tools, how do they fit with what we know? As empirical economists, how can we use them?[1]

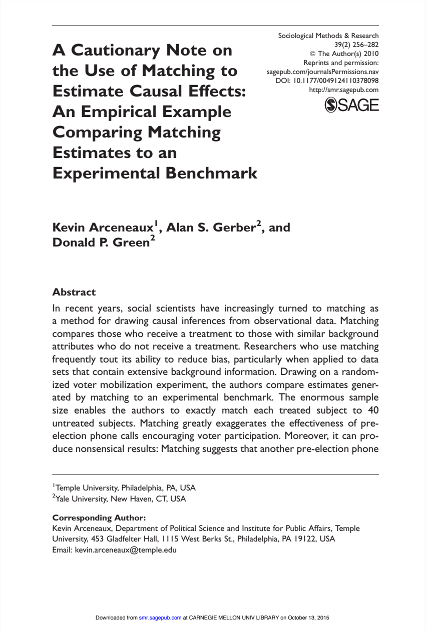
We present a way of thinking about machine learning that gives it its own place in the econometric toolbox. Central to our understanding is that machine learning

[1] In this journal, Varian (2014) provides an excellent introduction to many of the more novel tools and "tricks" from machine learning, such as decision trees or cross-validation. Einav and Levin (2014) describe big data and economics more broadly. Belloni, Chernozhukov, and Hanson (2014) present an econometrically thorough introduction on how LASSO (and close cousins) can be used for inference in high-dimensional data. Athey (2015) provides a brief overview of how machine learning relates to causal inference.

- Under this notion of prediction, "prediction" becomes its own task
- The traditional task is information, or *explanation*
- "y-hat" versus "beta-hat" problems

(Caution: "Causation" is itself limited)

- Critique 1: Causal inference (econometrics) can fail hopelessly
- Critique 2: Automated methods (from "causal learning") have strong, unrealistic, and untestable assumptions
- Critique 3: Statistical expression of causation is short-range (Gene Richardson)



Sociological Methods & Research
39(2) 256–282
© The Author(s) 2010
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0049124110378098
http://smr.sagepub.com
SAGE

A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark

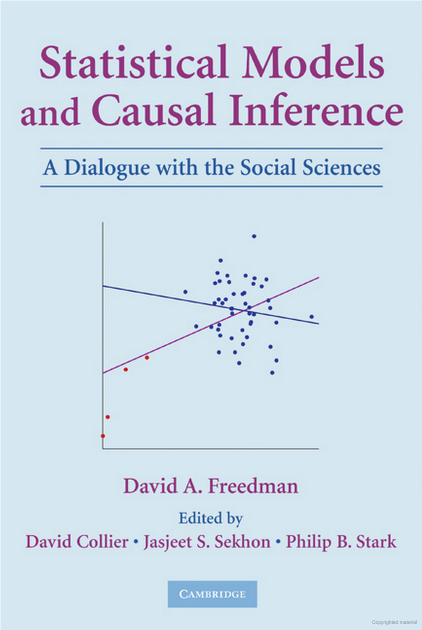
Kevin Arceneaux[1], Alan S. Gerber[2], and Donald P. Green[2]

Abstract

In recent years, social scientists have increasingly turned to matching as a method for drawing causal inferences from observational data. Matching compares those who receive a treatment to those with similar background attributes who do not receive a treatment. Researchers who use matching frequently tout its ability to reduce bias, particularly when applied to data sets that contain extensive background information. Drawing on a randomized voter mobilization experiment, the authors compare estimates generated by matching to an experimental benchmark. The enormous sample size enables the authors to exactly match each treated subject to 40 untreated subjects. Matching greatly exaggerates the effectiveness of pre-election phone calls encouraging voter participation. Moreover, it can produce nonsensical results: Matching suggests that another pre-election phone

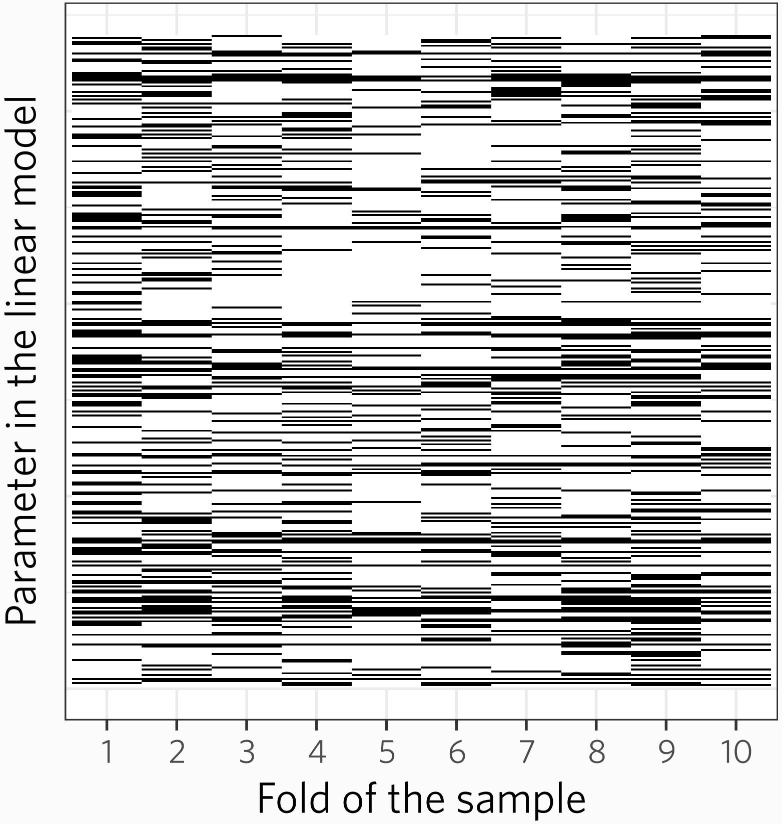
[1]Temple University, Philadelphia, PA, USA
[2]Yale University, New Haven, CT, USA

Corresponding Author:
Kevin Arceneaux, Department of Political Science and Institute for Public Affairs, Temple University, 453 Gladfelter Hall, 1115 West Berks St., Philadelphia, PA 19122, USA
Email: kevin.arceneaux@temple.edu



The problem with correlation

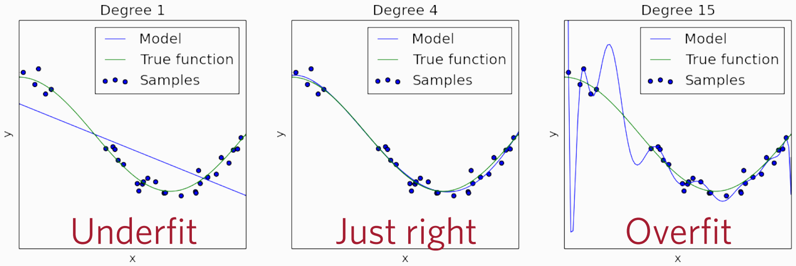
- Very different models will 'predict' equally well, and often better than any theory-driven model (Mullainathan & Spiess, 2017)
- For *intervention*, we need causality (or at least associations)
- Another problem: correlations can *overfit*



Validity: Correlations can overfit, cross-validation doesn't fully address

Overfitting and cross validation

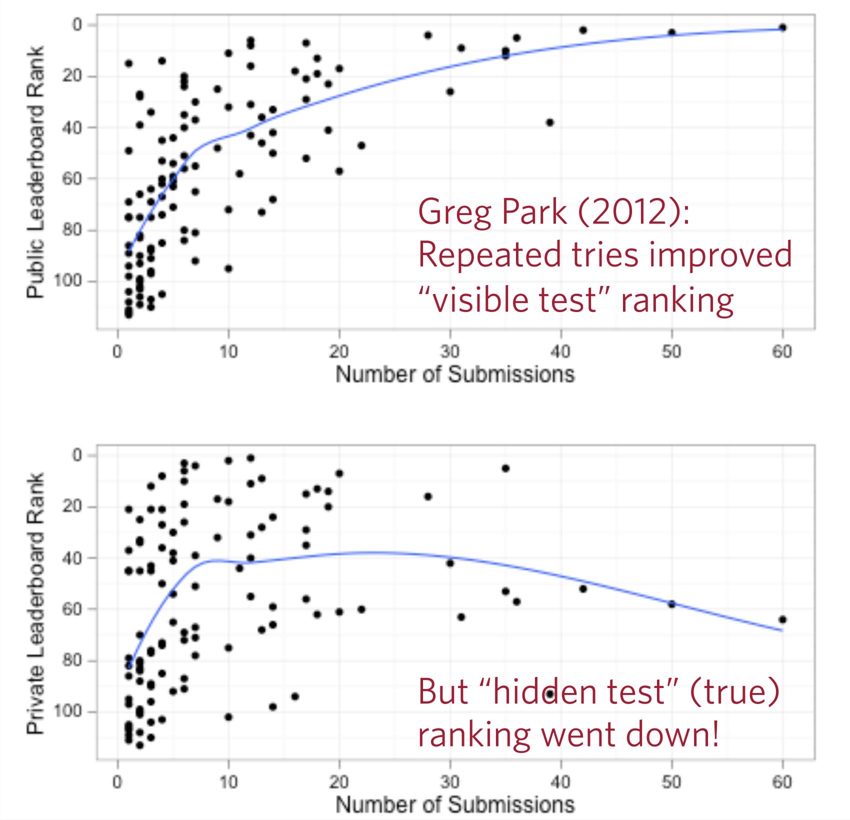
- *Overfitting*: Model fits to 'noise' rather than the cause/signal/ *data-generating process*. Machine learning metaphor: "memorize the data."



- (p-hacking relates to both fit and *variability*; overfitting is related but simpler)
- Cross validation: split the data into two parts (e.g., 1:1, 4:1, 9:1). *The signal should be the same, but not the noise.* Error rate on the held-out "test" set should say how well correlations generalize.

But cross-validation can fail

- Re-using a test set can overfit to the test set! Happens in Kaggle
- Or, if there are dependencies (temporal, network, group) between data splits, it "shares" information
- E.g., temporal: Fitting on values that come after test values is "time traveling"!



A 'false' model may predict better than a 'true' one

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The bias-variance tradeoff

- The bias-variance 'decomposition', a foundational result for machine learning and modern statistics:

$$\begin{aligned}\mathrm{EPE}(x) &= \mathbb{E}\big[\big(Y - \widehat{f}(x)\big)^2 \,\big|\, X = x\big] \\ &= \mathrm{Var}(Y) + \mathbb{E}\big[\big(\widehat{f}(x) - f(x)\big)^2 \,\big|\, X = x\big] + \mathbb{E}\big[\big(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)]\big)^2 \,\big|\, X = x\big] \\ &= \sigma^2 + \mathrm{bias}^2\big(\widehat{f}(x)\big) + \mathrm{Var}\big(\widehat{f}(x)\big)\end{aligned}$$

- Leads to a 'tradeoff': *Even if we have all the "right" variables, a biased model may be better*
- This is very strange!

Simulation illustration: Setup

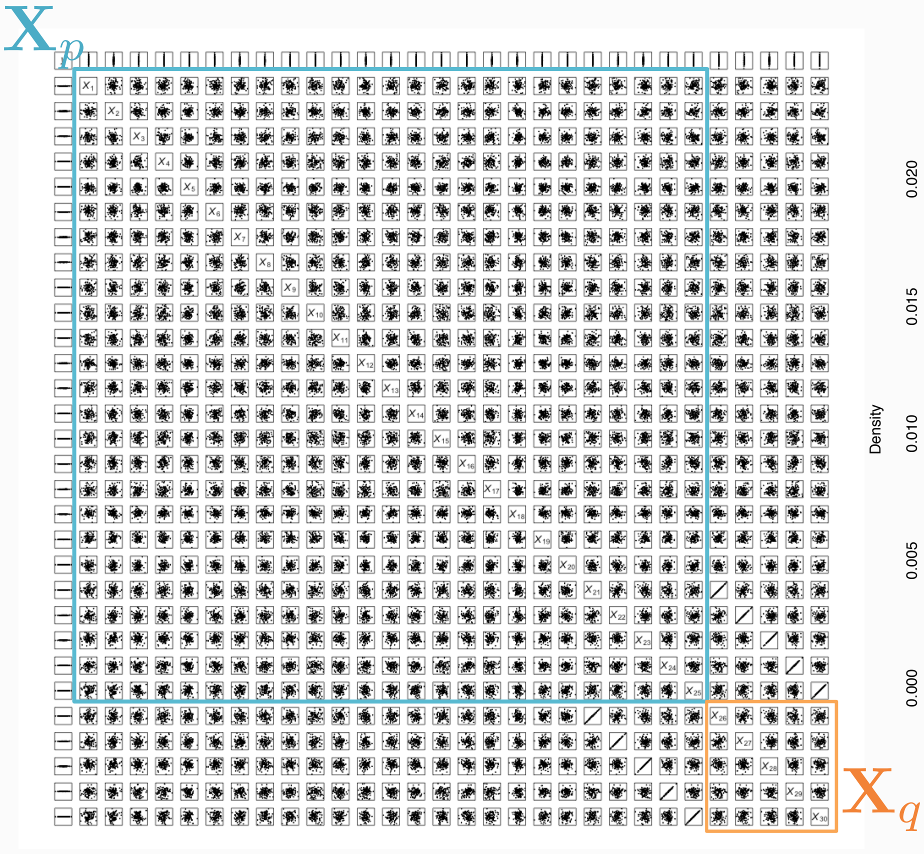
- A linear data-generating process.

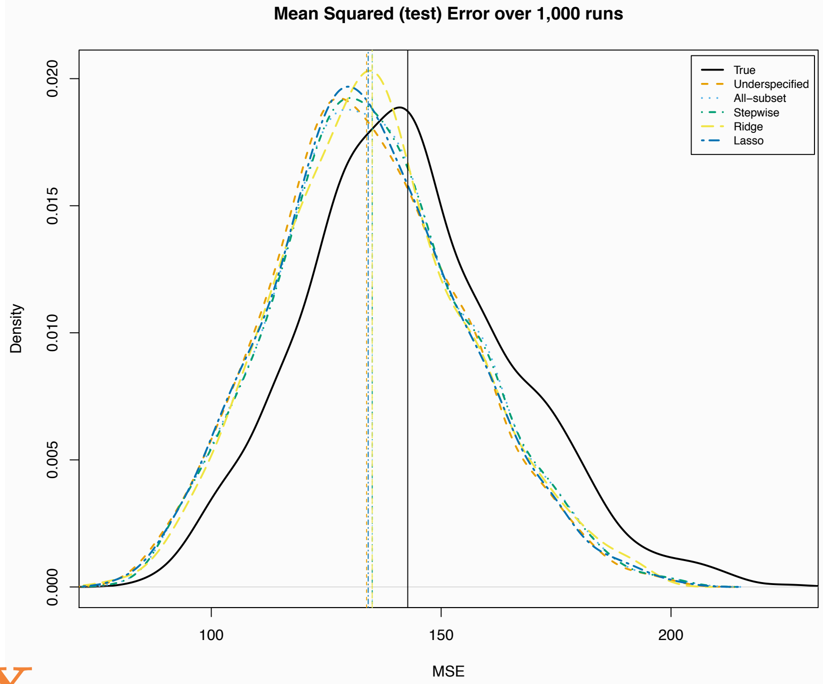
$$\mathbf{y} \sim \mathcal{N}\left(\beta_p \mathbf{X}_p + \beta_q \mathbf{X}_q,\ \sigma^2 \mathbf{I}\right)$$

- Wu et al. (2007): Fitting only $\mathbf{X}_p$ has lower expected MSE than fitting the model that generated the data when:

$$\beta_q^T \mathbf{X}_q^T (\mathbf{I}_n - \mathbf{H}_p) \mathbf{X}_q \beta_q < q\sigma^2$$

The 'true' model predicts worse!





Summary

- 'Prediction' is a metaphor used for fitted values, not (necessarily) actual prediction
- Spurious correlations count as 'prediction' and can do quite well in narrow terms, but are fragile and don't help us intervene
- Correlations can overfit, and cross-validation doesn't fully solve
- The bias-variance tradeoff means things are even more strange
- *I would argue: These are the pertinent issues*

Thank you!


WHAT WILL BECOME OF US
↳ HOW TECHNOLOGY IS
CHANGING WHAT IT MEANS TO BE HUMAN.
NOVEMBER 18, 2018

Citations/Credits by slide number

2 Robot holding skull: Cover image of "What Will Become of Us?", *New York Times Magazine* (The Tech & Design issue), 14 November 2018. Concept by delcan & company. Photo illustration by Jamie Chung. Prop styling by Pink Sparrow. C.G. work by Justin Metz. https://www.nytimes.com/2018/11/14/magazine/behind-the-cover-what-will-become-of-us.html.

2 Terminator skull: Nemesis Now Ltd, Terminator Skull Box T-800 (18CM). https://www.menkind.co.uk/terminator-t800-skull-box.

2 Hand: Hayati Kayhan, Holding human skull in hand, Conceptual image (Shakespeare's Hamlet scene concept). 19 October 2014.

3 Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.

7 Sitaram Asur and Bernando A. Huberman, "Predicting the Future With Social Media." In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT '10), 492–499. 2010. https://dx.doi.org10.1109/WI-IAT.2010.63.

7 Ziad Obermeyer and Ezekiel J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine." *New England Journal of Medicine* 375, no. 13 (2016): 1216–1219. https://dx.doi.org/10.1056/NEJMp1606181.

7 OED Online, "predict, v." Oxford University Press, July 2018. http://www.oed.com/view/Entry/149856.

8 Daniel Gayo-Avello, "'I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper': A Balanced Survey on Election Prediction using Twitter Data." *arXiv*, 28 April 2012. https://arxiv.org/abs/1204.6441.

8 Daniel Gayo-Avello, "No, You Cannot Predict Elections with Twitter." *IEEE Internet Computing* 16 (2012): 91–94. https://dx.doi.org/10.1109/MIC.2012.137.

9 Drew McDermott, "Artificial Intelligence meets Natural Stupidity." *SIGART Bulletin* 57 (April 1976): 4–9.

12 David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu, "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance." *Notices of the AMS* 61, no. 5 (2014): 458–471. https://dx.doi.org/10.1090/noti1105.

14 David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, "The Parable of Google Flu Trends: Traps in Big Data Analysis." *Science* 343 (14 March 2014): 1203–1205. https://dx.doi.org/10.1126/science.1248506.

15 Leo Breiman, "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16, no. 3 (2001): 199–231. https://dx.doi.org/10.1214/ss/1009213726.

15 Galit Shmueli, "To Explain or to Predict?" *Statistical Science* 25, no. 3 (2010): 289–310. https://dx.doi.org/10.1214/10-STS330.

15 Sendhil Mullainathan and Jan Spiess, "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31, no. 2 (2017): 87–106. https://dx.doi.org/10.1257/jep.31.2.87.

16 Kevin Arceneaux, Alan S. Gerber, and Donald P. Green, "A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark." *Sociological Methods & Research* 39, no. 2 (2010): 256–282. https://dx.doi.org/10.1177/0049124110378098.

16 David A. Freedman, *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge University Press, 2009.

19 scikit-learn developers, "Underfitting vs. Overfitting." 2014. https://scikit-learn.org/0.15/auto_examples/plot_underfitting_overfitting.html.

20 Greg Park, "The Dangers of Overfitting: A Kaggle Postmortem." 6 July 2012. http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem.

23 Shaohua Wu, T. J. Harris, and K. B. McAuley, "The Use of Simplified or Misspecified Models: Linear Case." *The Canadian Journal of Chemical Engineering* 85, no. 4 (2007): 386–398. https://dx.doi.org/10.1002/cjce.5450850401.