

because this road is endless\*.

**machine learning**

**ALL MODELS ARE WRONG BUT SOME ARE USEFUL**



# **A Hierarchy of Limitations in Machine Learning**

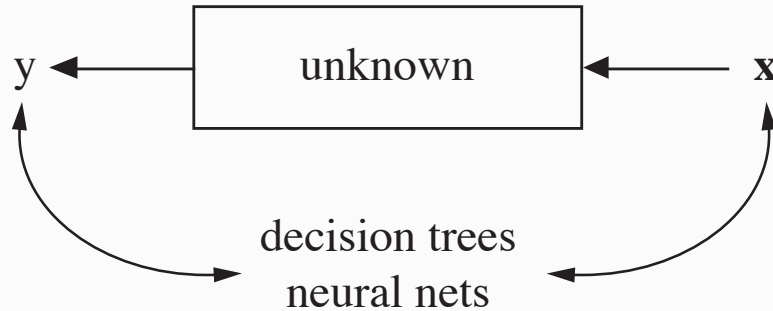
---

Momin M. Malik, Data Science Postdoctoral Fellow, Berkman  
Klein Center for Internet & Society at Harvard University  
> 03 December 2019, Microsoft Research New England

## › Objective

- › Introduction
  - › Meaning and measurement
  - › Central tendency
  - › Causality
  - › Capturing variability
  - › Cross-validation
  - › Reflection
  - › Future steps
  - › References
- › What are all the ways in which machine learning can *fail*\*?
  - › How can we address these failure points?
  - \* Fail = be unreliable, or result in *unanticipated* harm
  - › Failures will form a *hierarchy*

# ➤ The basic bargain of ML

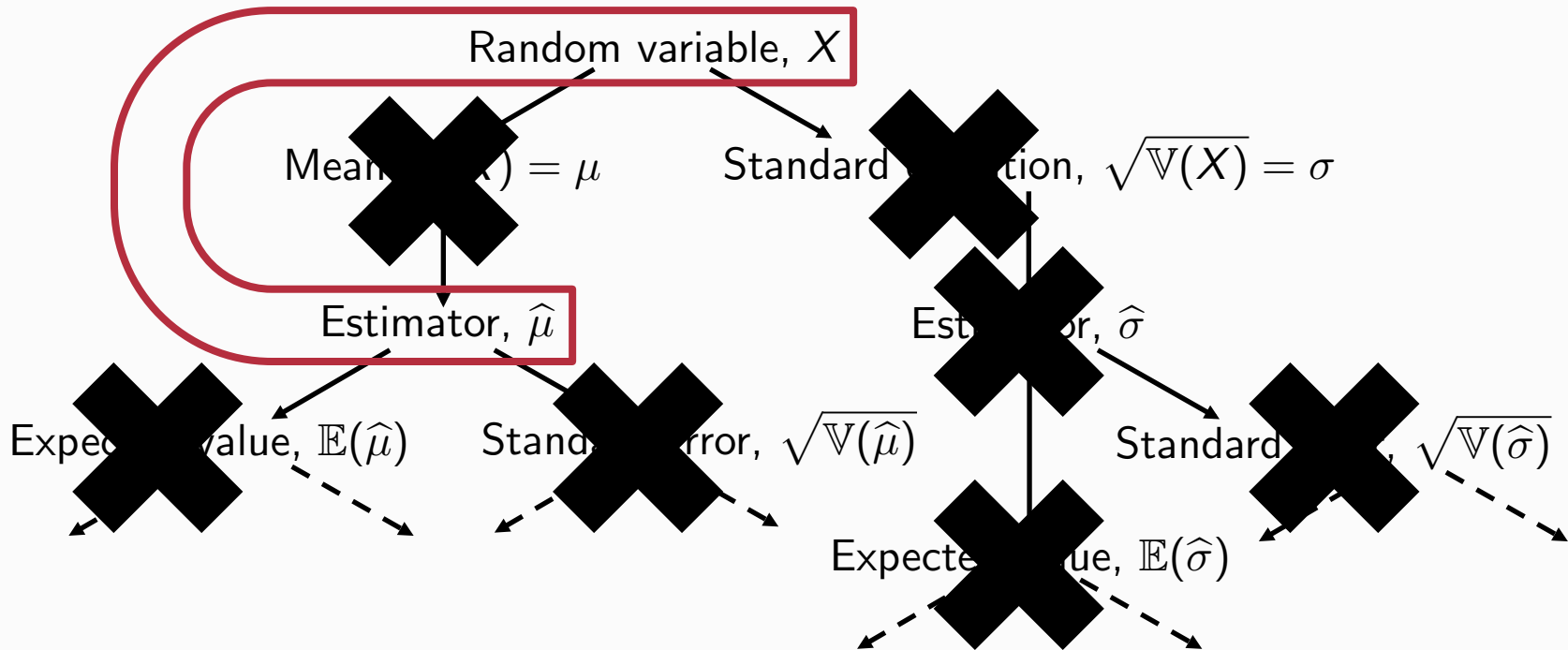


So long as machine learning establishes *external validity* (generalizability) from a given input, it can arguably ignore...

- All other validity questions
- All the problems that plague statistics

Breiman, 2001, *Stat. Sci.* See also Jones, 2018, *Hist. Stud. Nat. Sci.*

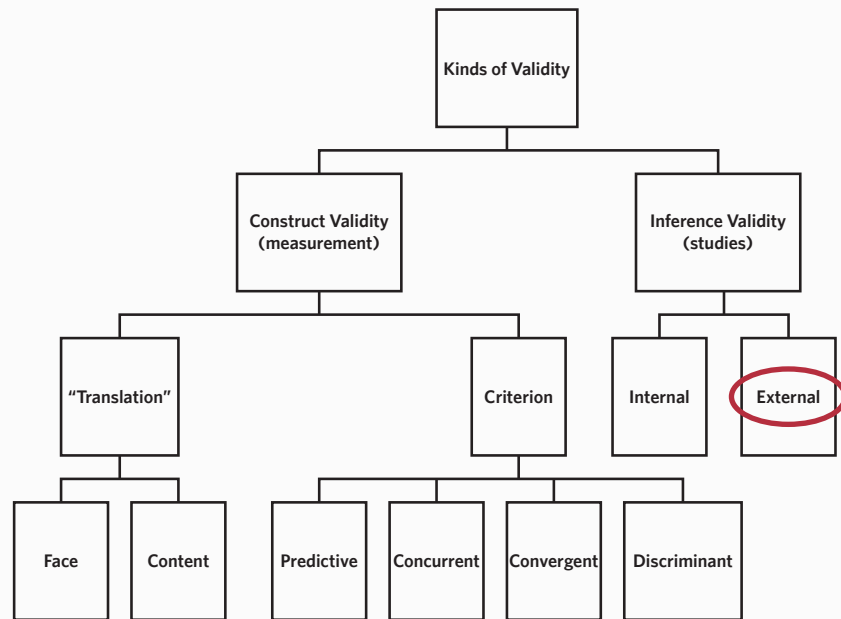
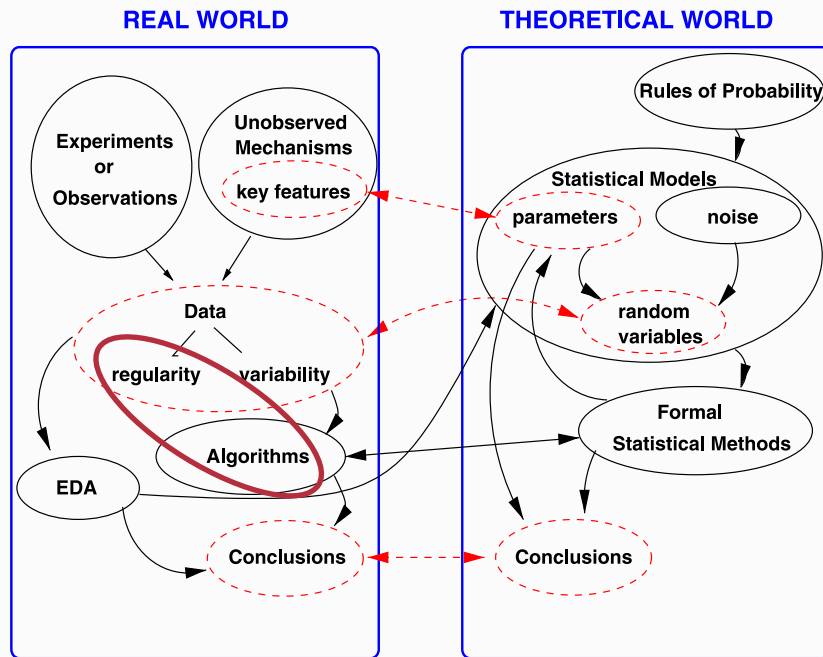
# > Can focus on one goal



See also: Robert Tibshirani, "Recent Advances in Post-Selection Inference" (2015)

# Everything else can be ignored

- > Introduction
- > Meaning and measurement
- > Central tendency
- > Causality
- > Capturing variability
- > Cross-validation
- > Reflection
- > Future steps
- > References



Kass, 2011, *Stat. Sci.*

Adapted from Borgatti, 2012



- › Introduction
- › Meaning and measurement
- › Central tendency
- › Causality
- › Capturing variability
- › Cross-validation
- › Reflection
- › Future steps
- › References

# › ...or can it?

# ➤ Outline

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

Problems with/Failures of:

- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation

Reflection

STS

More general;  
problems  
percolate  
downward in  
this hierarchy

ML

More specific

- › Introduction
- › Meaning and measurement
- › Central tendency
- › Causality
- › Capturing variability
- › Cross-validation
- › Reflection
- › Future steps
- › References

# › Meaning and measurement



# > Meaning-making

“During the writing of this book, my first grandchild was born. The hospital records document her weight, height, health[;] the mother’s condition, length of labor, time of birth, and hospital stay... These are physiological and institutional metrics. When aggregated across many babies and mothers, they provide trend data about the beginning of life—birthing.”

- > Introduction
- > Meaning and measurement
- > Central tendency
- > Causality
- > Capturing variability
- > Cross-validation
- > Reflection
- > Future steps
- > References

# > Meaning-making

“But nowhere in the hospital records will you find anything about what the birth of Calla Quinn *means*. Her existence is documented but not what she means to our family, what decision-making process led up to her birth, the experience and meaning of the pregnancy, the family experience of the birth process, and the familial, social, cultural, political, and economic context...” (Patton, 2015)

- > Introduction
- > Meaning and measurement
- > Central tendency
- > Causality
- > Capturing variability
- > Cross-validation
- > Reflection
- > Future steps
- > References

# > Experience

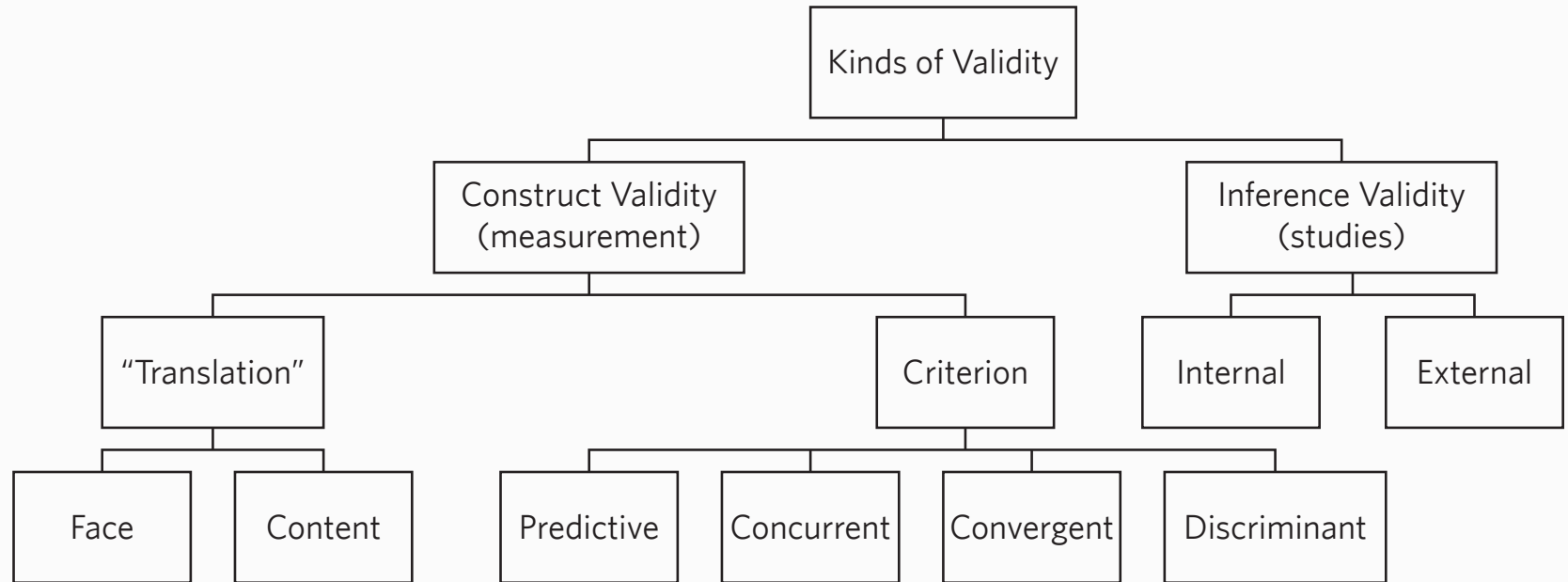
- > Introduction
- > Meaning and measurement
- > Central tendency
- > Causality
- > Capturing variability
- > Cross-validation
- > Reflection
- > Future steps
- > References



- > “A white woman can say that a neighborhood is ‘sketchy’ and most people will smile and nod. She felt unsafe, and we automatically trust her opinion. A black man can tell the world that every day he lives in fear of the police, and suddenly everyone demands statistical evidence to prove that his life experience is real.”

# Validating measurements

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References



Adapted from Borgatti, 2012

## > Solutions

- > See quantitative research as *building* on what is qualitatively known, not replacing it
- > Think about measurement! Work with experts in validation
- > Integrate qualitative research for:
  - Needs assessments prior to ML
  - Annotation for training data
  - Evaluating implementations of ML systems

- › Introduction
- › Meaning and measurement
- › Central tendency
- › Causality
- › Capturing variability
- › Cross-validation
- › Reflection
- › Future steps
- › References

# › Central tendency

# ➤ The world as a data matrix

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

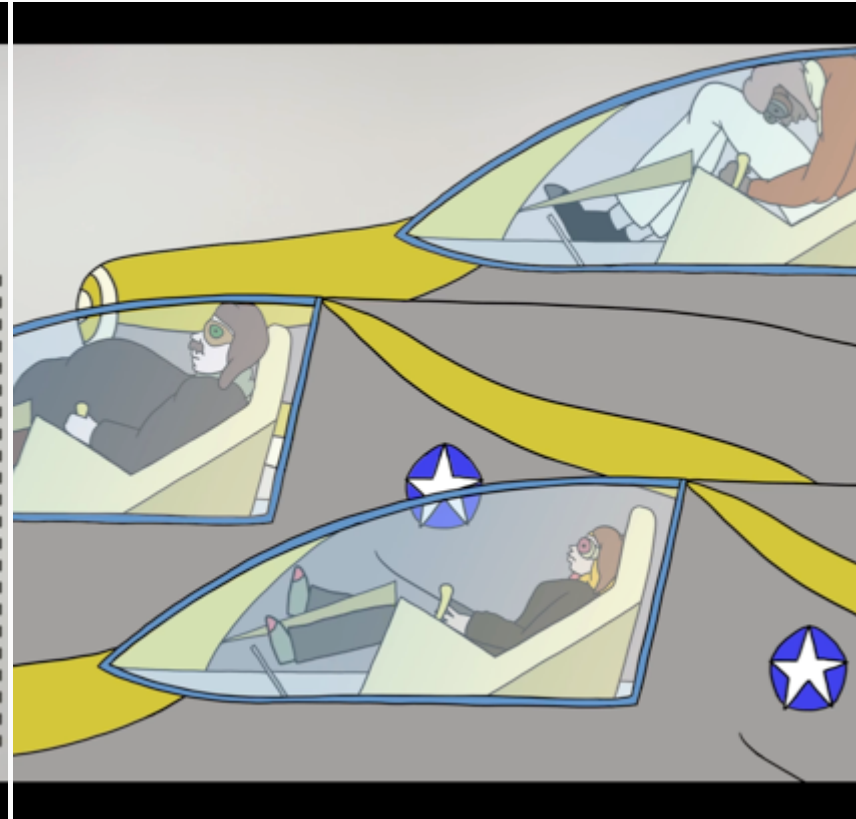
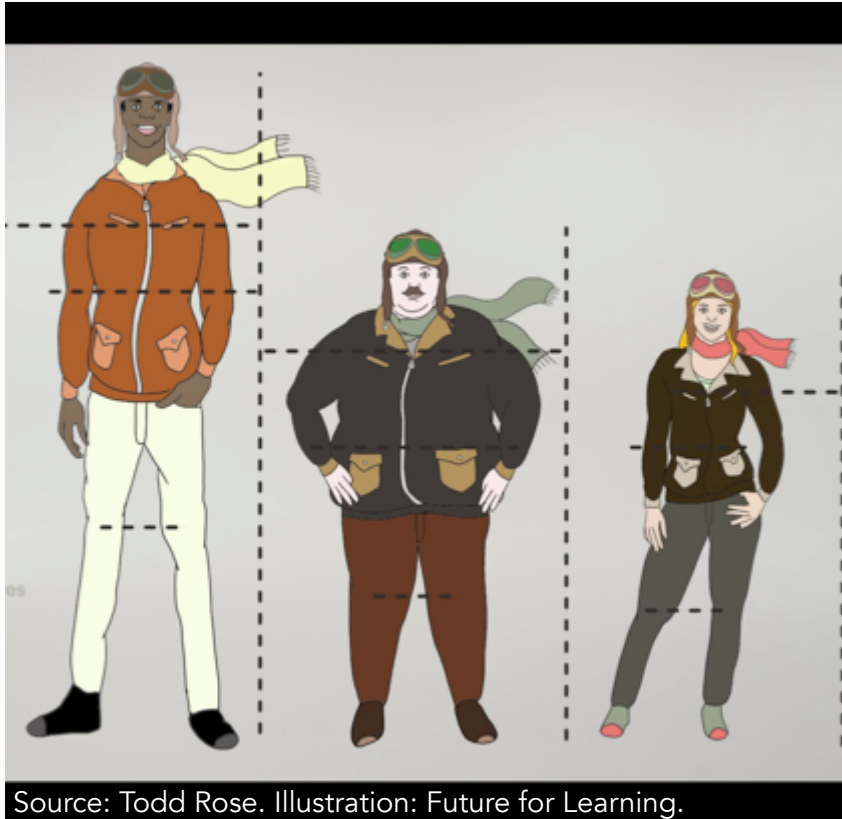
**“it is striking how absolutely these assumptions contradict those of the major theoretical traditions of sociology.**

Symbolic interactionism rejects the assumption of fixed entities and makes the meaning of a given occurrence depend on its location — within an interaction, within an actor's biography, within a sequence of events.

“Both the Marxian and Weberian traditions deny explicitly that a given property of a social actor has one and only one set of causal implications... Marx, Weber, and work deriving from them in historical sociology all approach social causality in terms of stories, rather than in terms of variable attributes.” (Abbott, 1988)

# ➤ “Flaw of averages”

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References



Source: Todd Rose. Illustration: Future for Learning.



## > Solutions

- > No matter how small the bins are, is still a central tendency (i.e.: mean, median, majority class, etc.)
  - No longer a central tendency when the bins have an  $n$  of 1... but then ML and stats can do nothing but restate that datum
- > Recognize that *planning to the central tendency punishes outliers* (Keyes 2018): plan for this!



- › Introduction
- › Meaning and measurement
- › Central tendency
- › **Causality**
- › Capturing variability
- › Cross-validation
- › Reflection
- › Future steps
- › References

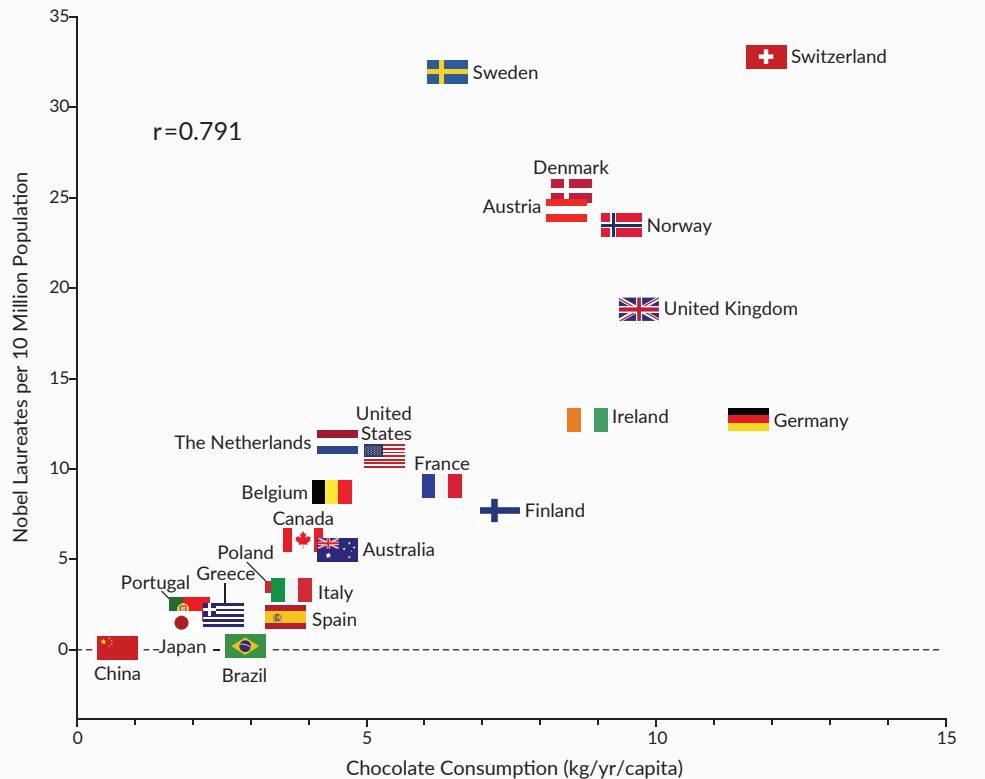
# › Causality

## ➤ Sometimes, people want causality

- "A project I worked on in the late 1970s was the analysis of delay in criminal cases in state court systems... A large decision tree was grown, and I showed it on an overhead and explained it to the assembled Colorado judges. One of the splits was on District N which had a larger delay time than the other districts. I refrained from commenting on this. But as I walked out I heard one judge say to another, 'I knew those guys in District N were dragging their feet.'" (Breiman, 2001)

# ➤ “Predictions” are correlations

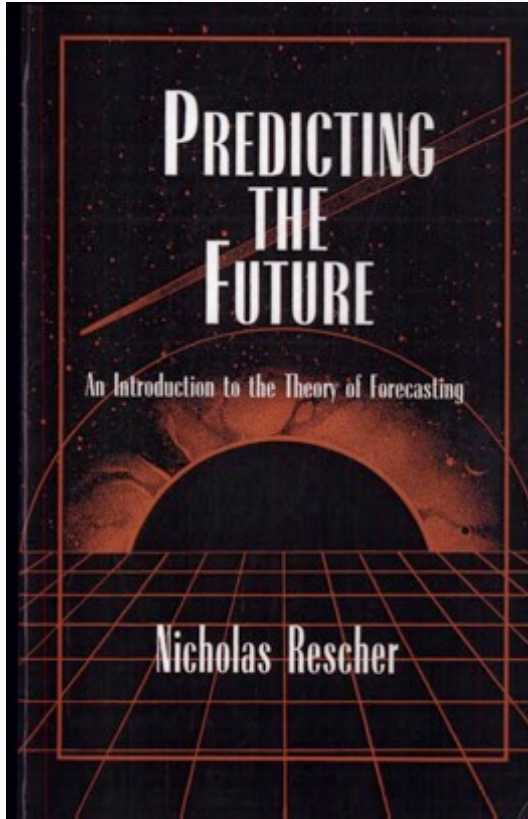
- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References



Messerli, 2012, *NEJM*

# > Not an obvious usage of “predict”

- > Introduction
- > Meaning and measurement
- > Central tendency
- > Causality
- > Capturing variability
- > Cross-validation
- > Reflection
- > Future steps
- > References



A Hierarchy of Limitations of ML

## 88 ■ PREDICTING THE FUTURE

**TABLE 6.1: A SURVEY OF PREDICTIVE APPROACHES**

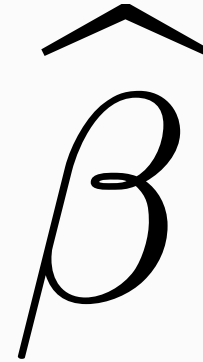
Predictive Approaches	Linking Mechanism	Methodology Of Linkage
<b>UNFORMALIZED/JUDGMENTAL</b>		
judgmental estimation	expert informants	informed judgment
<b>FORMALIZED/INFERENCEAL</b>		
<b>RUDIMENTARY (ELEMENTARY)</b>		
trend projection	prevailing trends	projection of prevailing trends
curve fitting	geometric patterns	subsumption under an established pattern
circumstantial analogy	comparability groupings	assimilation to an analogous situation
<b>SCIENTIFIC (SOPHISTICATED)</b>		
indicator coordination	causal correlations	statistical subsumption into a correlation
law derivation (nomic)	accepted laws (deterministic or statistical)	inference from accepted laws
phenomenological modeling (analogical)	formal models (physical or mathematical)	analogizing of actual ("real-world") processes with presumably isomorphic model process

# > Creates two types of modeling!



Correlations may “predict” well

- > Breiman, 2001: Prediction
- > Shmueli, 2010: Prediction
- > Kleinberg et al., 2015: Umbrella
- > Mullainathan & Spiess, 2017:  $\hat{y}$

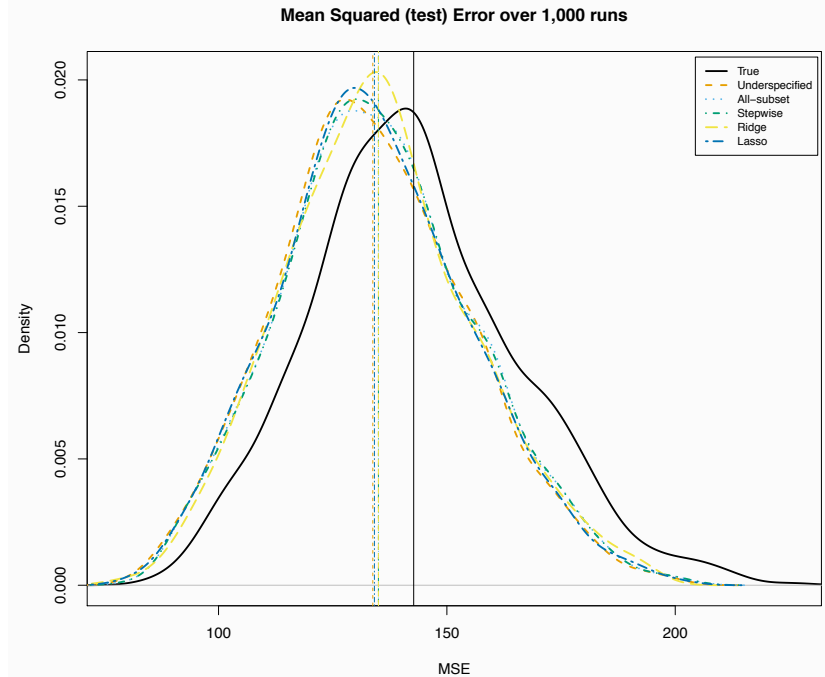
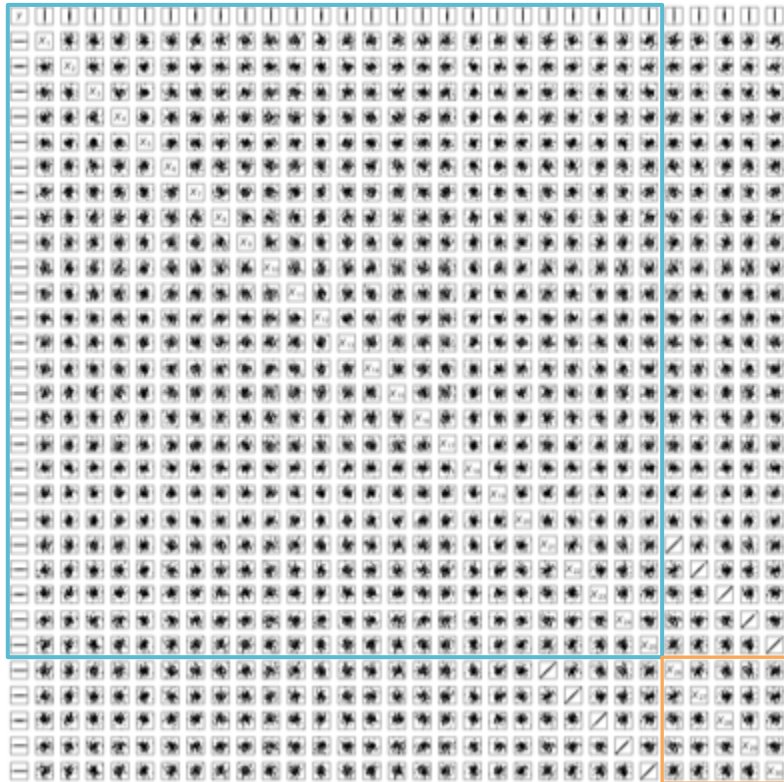


Informative models may not fit well

- > Breiman 2001: Information
- > Shmueli 2010: Explanation
- > Kleinberg et al 2015: Rain dance
- > Mullainathan & Spiess, 2017:  $\hat{\beta}$

# ➤ “True” model can predict worse!

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References



Simulation of Shmueli, 2010, *Stat. Sci*

# › Solution: Determine type of problem

- › By “prediction”, we mean correlation (Caruana et al., 2015; Doshi-Velez & Kim, 2017): communicate this!!
- › If not a “prediction policy problem”, then machine learning may not be appropriate (whether explainable/interpretable or not!)
- › Then: causal modeling, or statistical modeling of data-generating process to get “explanations” (causality lite)
- › Other benefits to causal knowledge:
  - Makes predictions robust to distributions shifts (argument of causal learning literature; Spirtes & Zhang, 2016)
  - Allows us to *intervene*



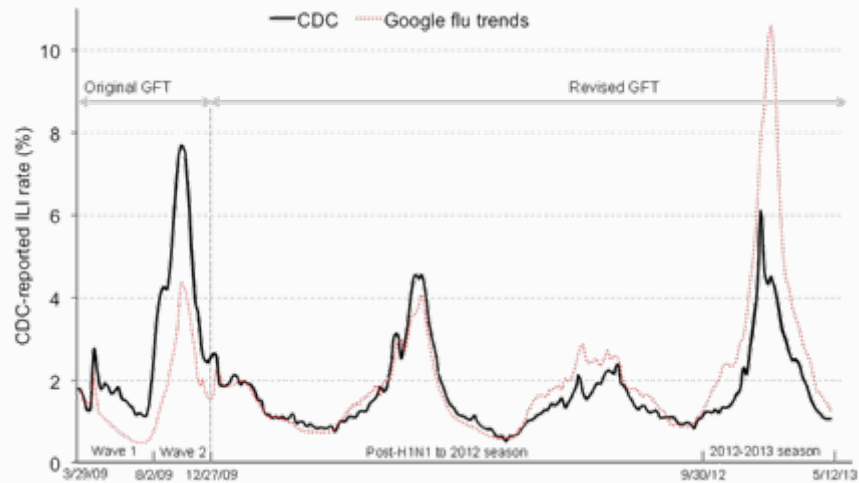
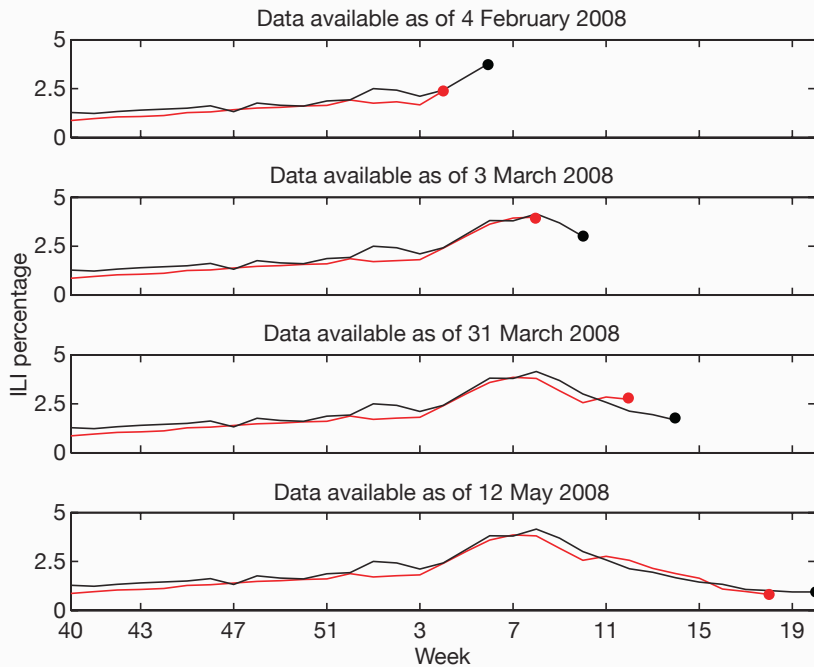


- › Introduction
- › Meaning and measurement
- › Central tendency
- › Causality
- › **Capturing variability**
- › Cross-validation
- › Reflection
- › Future steps
- › References

# › Capturing variability

# ➤ When data don't capture key variability (Google Flu Trends)

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

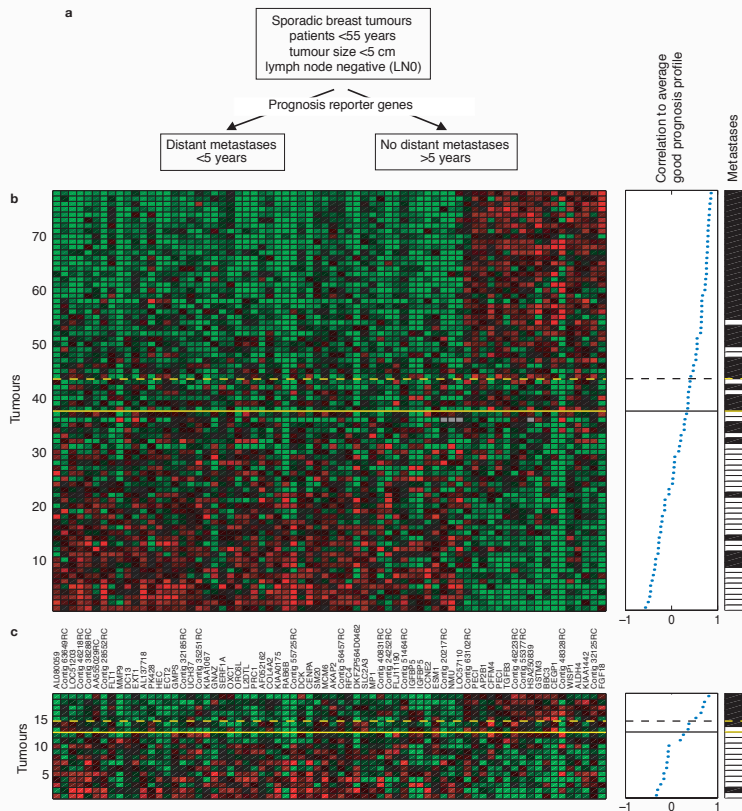


Ginsberg et al., 2012, *Nature*

Santillana et al., 2014, *Am. J. Prev. Med.*

# ➤ Real-world testing of ML results

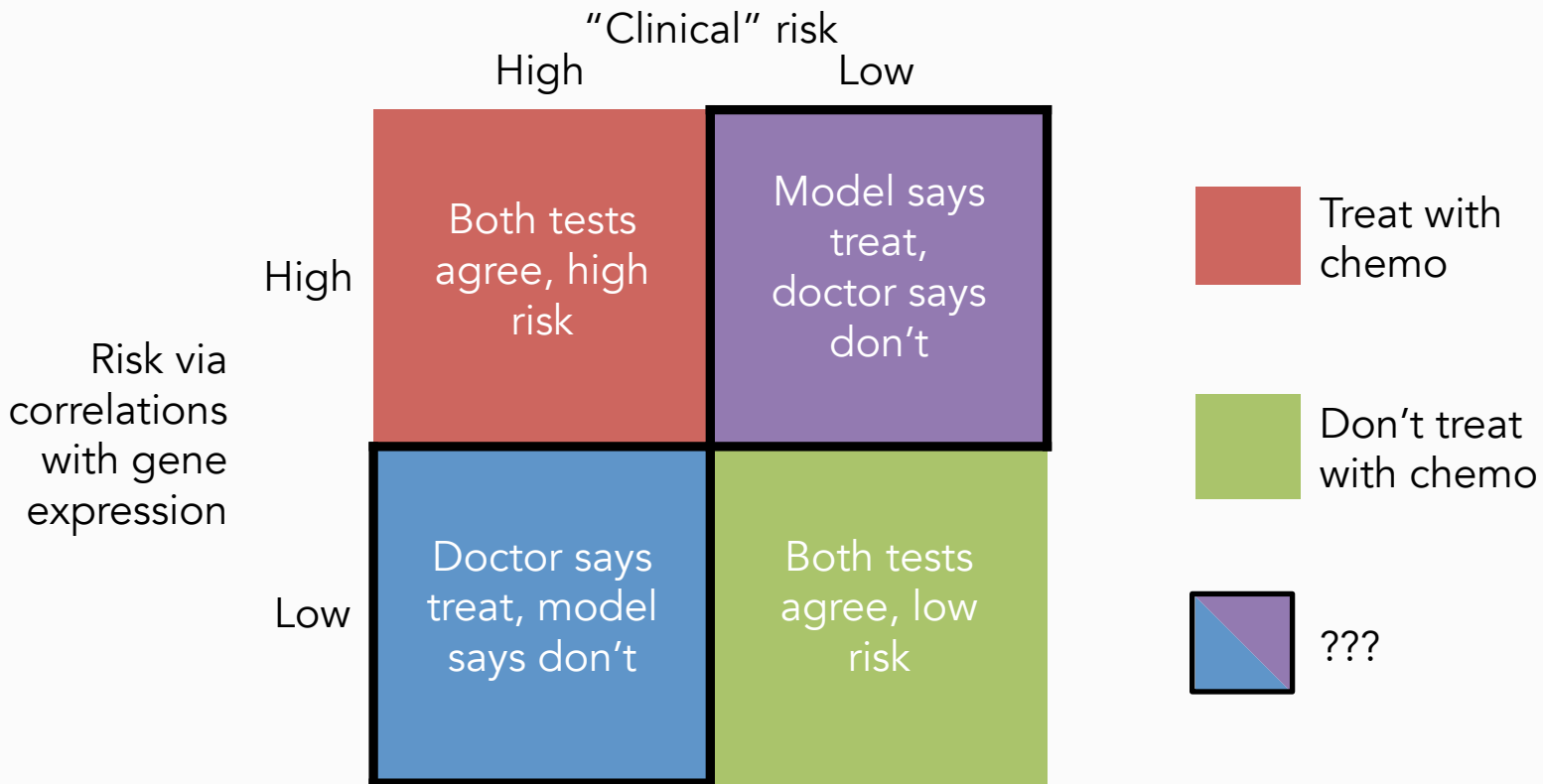
- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References



- van't Veer et al. (2002) found 70 genes correlated with developing breast cancer
- Of course the correlations were optimal, post-hoc. But did it generalize?

# ➤ Implementation testing

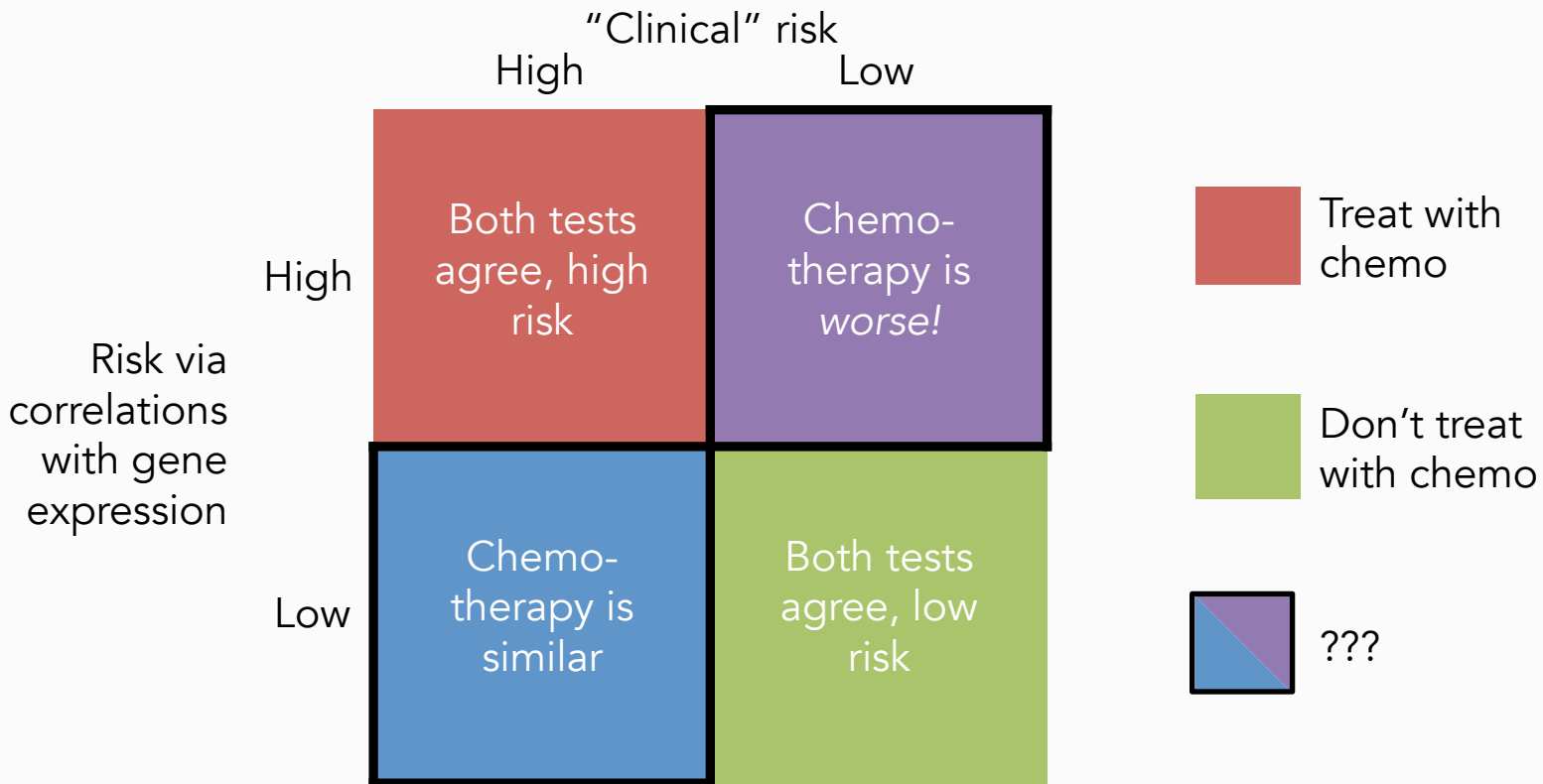
- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References



Cardoso et al., 2016, *NEJM*

# Implementation testing

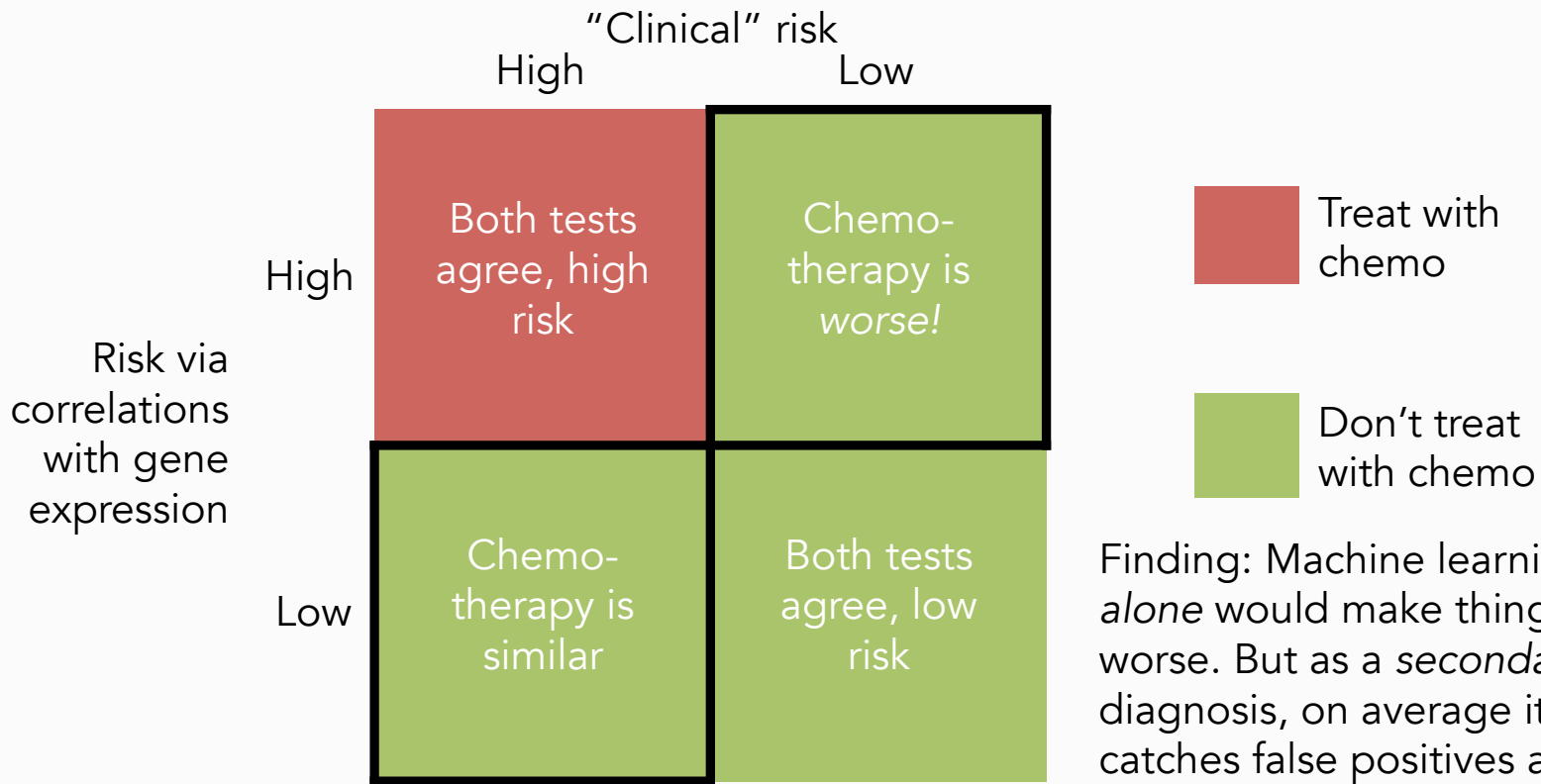
- > Introduction
- > Meaning and measurement
- > Central tendency
- > Causality
- > Capturing variability
- > Cross-validation
- > Reflection
- > Future steps
- > References



Cardoso et al., 2016, *NEJM*

# > Implementation testing

- > Introduction
- > Meaning and measurement
- > Central tendency
- > Causality
- > Capturing variability
- > Cross-validation
- > Reflection
- > Future steps
- > References

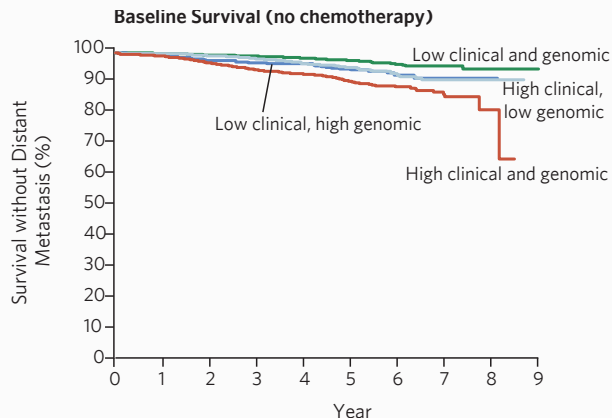


Finding: Machine learning *alone* would make things worse. But as a *secondary* diagnosis, on average it catches false positives and avoids unhelpful chemo!

Cardoso et al., 2016, *NEJM*

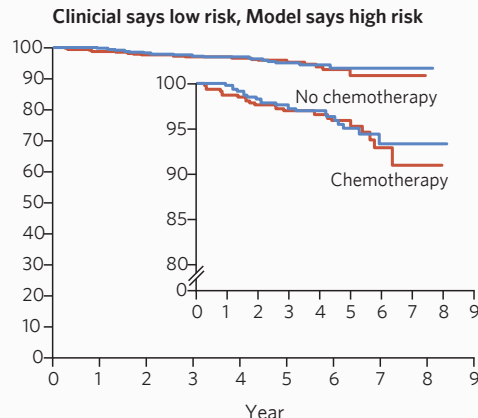
# Implementation testing: Details

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

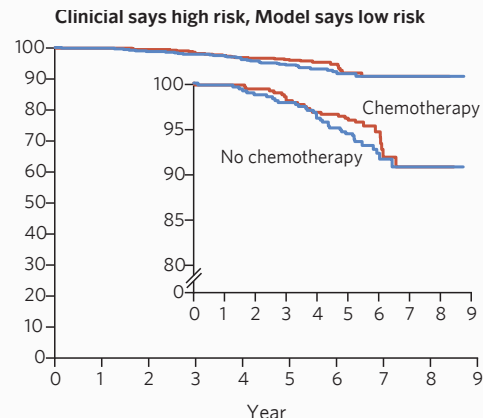


➤ Before experiment (training data)

(Note: still limitations in how experimental subjects may be unrepresentative.)



➤ High model risk, low clinical risk: randomize. Chemo worse!



➤ Low model risk, high clinical risk: chemo makes no difference

# ➤ Solutions to capturing variability

- Introduction
  - Meaning and measurement
  - Central tendency
  - Causality
  - Capturing variability
  - Cross-validation
  - Reflection
  - Future steps
  - References
- GFT is a prediction policy problem: problem wasn't overfitting, or causality (although causality may have helped), but that there was key variability (non-winter flu) that had not yet been observed
  - Rhetoric: Emphasize that cross-validation does not give performance guarantees, only suggestive of performance until testing is done
  - Real-world testing reveals how system can be used!
  - (Also: experimental data can introduce full amount of variability, even if not doing causal inference)



- › Introduction
- › Meaning and measurement
- › Central tendency
- › Causality
- › Capturing variability
- › **Cross-validation**
- › Reflection
- › Future steps
- › References

# › Cross validation

# ➤ Generalizability through CV

- Generalizability is shown (at the very least) *through cross validation*
- CV can go wrong in known ways:
  - improper splitting
  - publication bias (Gayo-Avello, 2012)
  - overfitting to the test set (Dwork et al. 2015, Park 2012)
- Not systematically acknowledged: *dependencies among observations*

# > Classic argument for CV

Training:

$$\begin{aligned}
 \text{err}(\hat{\mu}) &= \frac{1}{n} \mathbb{E}_f \|Y - \hat{Y}\|_2^2 \\
 &= \frac{1}{n} \left[ \text{tr} \Sigma + \|\mu - \mathbb{E}(\hat{Y})\|_2^2 + \text{tr} \text{Var}_f(\hat{Y}) - 2 \text{tr} \text{Cov}_f(Y, \hat{Y}) \right]
 \end{aligned}$$

Testing:

$$\begin{aligned}
 \text{Err}(\hat{\mu}) &= \frac{1}{n} \mathbb{E}_f \|Y^* - \hat{Y}\|_2^2 \\
 &= \frac{1}{n} \left[ \text{tr} \Sigma + \|\mu - \mathbb{E}(\hat{Y})\|_2^2 + \text{tr} \text{Var}_f(\hat{Y}) - \cancel{2 \text{tr} \text{Cov}_f(Y^*, \hat{Y})} \right]
 \end{aligned}$$

The difference is the *optimism* (Efron, 2004; Rosset & Tibshirani, 2018):

$$\text{Opt}(\hat{\mu}) = \text{Err}(\hat{\mu}) - \text{err}(\hat{\mu}) = \frac{2}{n} \text{tr} \text{Cov}_f(Y, \hat{Y})$$

## ➤ Apply this to non-iid data

➤ Imagine we have, for  $\Sigma_{ii} = \sigma^2$  and  $\Sigma_{ij} = \rho\sigma^2$ ,  $i \neq j$

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \mathbf{X} \\ \mathbf{X} \end{bmatrix} \beta, \begin{bmatrix} \Sigma & \rho\sigma^2 \mathbf{1}\mathbf{1}^T \\ \rho\sigma^2 \mathbf{1}\mathbf{1}^T & \Sigma \end{bmatrix} \right)$$

➤ Then, optimism in the training set is:

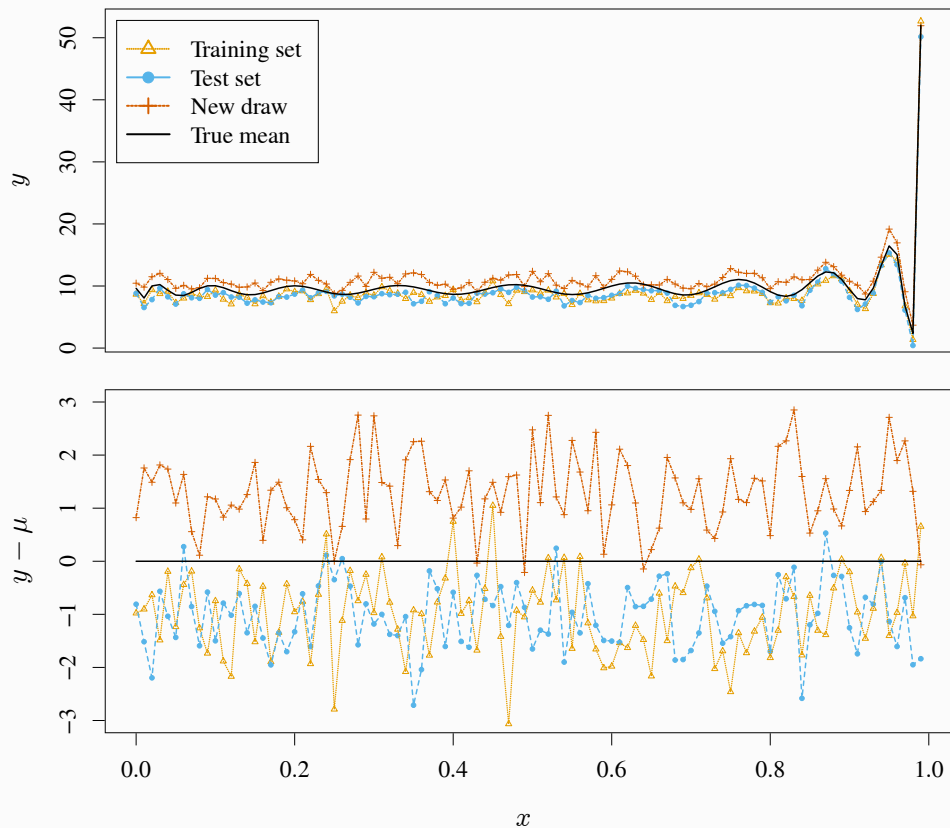
$$\frac{2}{n} \text{tr Cov}_f(Y_1, \hat{Y}_1) = \frac{2}{n} \text{tr Cov}_f(Y_1, \mathbf{H}Y_1) = \frac{2}{n} \text{tr } \mathbf{H} \text{Var}_f(Y_1) = \frac{2}{n} \text{tr } \mathbf{H}\Sigma$$

➤ But test set also has nonzero optimism!

$$\frac{2}{n} \text{tr Cov}_f(Y_2, \hat{Y}_1) = \frac{2}{n} \text{tr Cov}_f(Y_2, \mathbf{H}Y_1) = \frac{2\rho\sigma^2}{n} \text{tr } \mathbf{H}\mathbf{1}\mathbf{1}^T = 2\rho\sigma^2$$

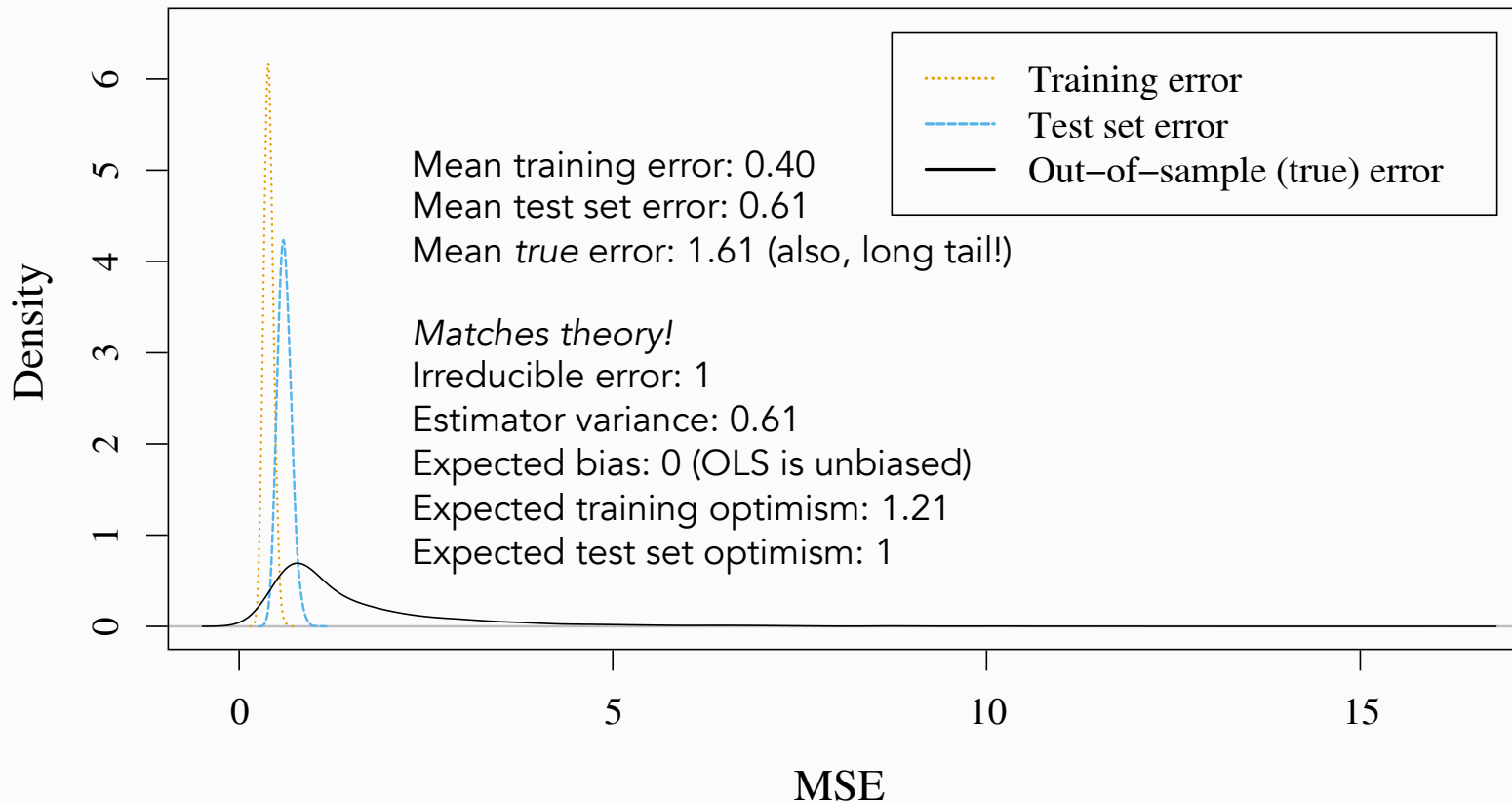
# > One draw as an example

Correlation  
 between  
 observations can  
 pull training and  
 test  
 observations  
 close to one  
 another, but  
 potentially far  
 from an  
 independent  
 draw



# Simulated MSE

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References



# ➤ Solution: Split by dependencies

- Study covariance, and design data splitting around it
  - But can't estimate both mean and the covariance structure, have to assume one (Opsomer et al., 2001)
  - (For covariance, *no amount of data is ever enough!*)
- Examples in literature:
  - Temporal block cross-validation (Bergmeir et al., 2018)
  - Leave-one-subject-out cross-validation (Hammerla & Plötz, 2015)
  - Network cross-validation

- › Introduction
- › Meaning and measurement
- › Central tendency
- › Causality
- › Capturing variability
- › Cross-validation
- › Reflection
- › Future steps
- › References

# › Reflection



# » About me

- » Introduction
- » Meaning and measurement
- » Central tendency
- » Causality
- » Capturing variability
- » Cross-validation
- » Reflection
- » Future steps
- » References

- »  **DEPARTMENT OF THE HISTORY OF SCIENCE**  
HARVARD UNIVERSITY
- »   **Berkman**  
The Berkman Center for Internet & Society at Harvard University
- »  
- » **Carnegie Mellon University**  
School of Computer Science
- » **Data Science For Social Good**  
———— Summer Fellowship ————
- »  **BERKMAN KLEIN CENTER**  
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY



Carnegie Mellon University  
**Societal  
Computing**



## > Background

- > “Methods are like people: if you focus only on what they can’t do, you will always be disappointed.” (Shapiro, 2014)
- > Trivially, all models are wrong because they aren’t the thing itself. But *specifically*, when, why, and how does it matter that ML is “wrong”?

# ➤ Encountering social science

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

Annu. Rev. Sociol. 2004. 30:243–70  
doi: 10.1146/annurev.soc.30.020404.104342  
Copyright © 2004 by Annual Reviews. All rights reserved  
First published online as a Review in Advance on March 9, 2004

## THE “NEW” SCIENCE OF NETWORKS

Duncan J. Watts

*Department of Sociology, Columbia University, New York, NY 10027; Santa Fe Institute, Santa Fe, New Mexico 97501; email: djw24@columbia.edu*

**Key Words** graph theory, mathematical models, network data, dynamical systems

■ **Abstract** In recent years, the analysis and modeling of networks, and also networked dynamical systems, have been the subject of considerable interdisciplinary interest, yielding several hundred papers in physics, mathematics, computer science, biology, economics, and sociology journals (Newman 2003c), as well as a number of books (Barabasi 2002, Buchanan 2002, Watts 2003). Here I review the major findings of this emerging field and discuss briefly their relationship with previous work in social and mathematical sciences.

### INTRODUCTION

Building on a long tradition of network analysis in sociology and anthropology (Degenne & Forse 1994, Scott 2000, Wasserman & Faust 1994) and an even longer history of graph theory in discrete mathematics (Ahuja et al. 1993, Bollobas 1998, West 1996), the study of networks and networked systems has exploded across the academic spectrum in the past five years. Spurred by the rapidly growing availability of cheap yet powerful computers and large-scale electronic datasets,

Viewpoints

DOI:10.1145/3132698

Hanna Wallach

## Viewpoint Computational Social Science ≠ Computer Science + Social Data

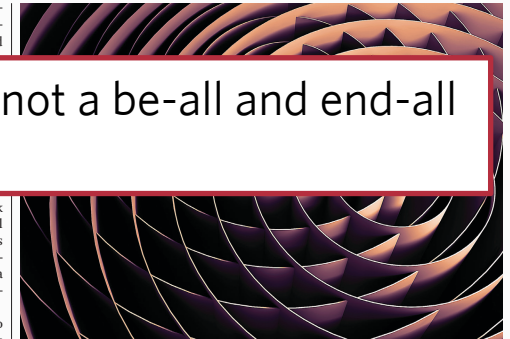
*The important intersection of computer science and social science.*

THIS VIEWPOINT IS ABOUT DIFFERENCES BETWEEN COMPUTER SCIENCE AND SOCIAL SCIENCE, AND

“machine learning is not a be-all and end-all solution.”

by training. That said, my recent work has been pretty far from traditional machine learning. Instead, my focus has been on computational social science—the study of social phenomena using digitized information and computational and statistical methods.

For example, imagine you want to know how much activity on websites such as Amazon or Netflix is caused by



based adjustments to each senator's | tional social science sits at the inter-

## > *Which social science?*

- > Is a “hard” version of social science compatible with modeling (econometrics, political science, quantitative sociology, biological anthropology, etc.)
- > Enriches ML and avoid pitfalls, but is not the most valuable and profound offering
- > Critical sociology, cultural anthropology, critical race studies, feminist studies, STS: not just about the world, but *about the very ways we see and interact with the world*

# > The original

6

---

## *Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI*

Philip E. Agre  
*University of California, San Diego*

- > Introduction
- > Meaning and measurement
- > Central tendency
- > Causality
- > Capturing variability
- > Cross-validation
- > Reflection
- > Future steps
- > References

## > What is “critical”?

- > “I finally comprehended the difference between critical thinking and its opposite. Technical people are not dumb, quite the contrary, but technical curricula rarely include critical thinking in the sense I have in mind. **Critical thinking means that you can, so to speak, see your glasses. You can look at the world, or you can back up and look at the framework of concepts and assumptions and practices through which you look at the world.**” (Agre, 2000)

## > Backing up in ML

- > Back up and ask, *what is the framework of concepts and assumptions and practices through which ML looks at the world?*
- > Identify this to understand the limits, and proper use, of ML
- > Start from the very beginning: quantification. End at ML-specific model validation.

# > Larger goals of this work

- > Communication oriented to *solutions*
  - “Burn it all down” on one side; totalitarian quantification + optimization on the other
- > Critics can be specific about what they object to, not just “ML is quantitative therefore evil”
- > ML can see how critiques show points of potential failure *that can be addressed*, rather than dismissing critiques



## > Pitfalls

- > The *fallacy of alternatives* (Agre, 1997): “Their stance was: if your alternative is so good then you will use it to write programs that solve problems better than anybody else's, and then everybody will believe you.”
- > Too limited to say, “social science can help solve a problem better”
- > Or even: “social science can help you identify the right problems”
- > Critical social science will reframe what even *counts* as a problem!

- › Introduction
- › Meaning and measurement
- › Central tendency
- › Causality
- › Capturing variability
- › Cross-validation
- › Reflection
- › Future steps
- › References

## › Future steps

## > Future steps

- > Many (but not all) of the problems can be alleviated (but not solved) through *mixed methods*: incorporating alternatives (qualitative research, statistical modeling, experimental design, etc.)
- > But mixed methods are hard!
- > We will need to develop both intellectual frameworks for mixed methods ML (e.g., purposeful division of labor), as well as *cultural practices* within ML for actually doing it. Future work, guided by this attempt to comprehensively map key points of failure!

## > Limitations

- > Focused on *research*, not industry; some things (especially implementation, which industry may know lots about) may not be meaningful critiques
- > Still developing what the intellectual framework would be, and what the cultural practices would be!

**➤ Thank you!**

# References (1/8)

Abbott, Andrew. "Transcending General Linear Reality." *Sociological Theory* 6, no. 2 (1988): 169-186. <https://dx.doi.org/10.2307/202114>.

Agre, Philip E. "Towards a Critical Technical Practice: Lessons Learned from Trying to Reform AI." In *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*, edited by Geoffrey C. Bowker, Susan Leigh Star, Will Turner, and Les Gasser, 131-158. Lawrence Erlbaum Associates, 1997.

<https://web.archive.org/web/20040203070641/http://polaris.gseis.ucla.edu/pagre/critical.html>.

Agre, Philip E. "Notes on critical thinking, Microsoft, and eBay, along with a bunch of recommendations and some URL's." *Red Rock Eater Newsletter*, 12 July 2000.

<https://pages.gseis.ucla.edu/faculty/agre/notes/00-7-12.html>.

Bailey, David H., 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>.

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo. "A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction." *Computational Statistics & Data Analysis* 120 (2018): 70-83.

<https://dx.doi.org/10.1016/j.csda.2017.11.003>.

# ➤ References (2/8)

Borgatti, Steve. "Types of Validity." BA 762: Research Methods. Gatton College of Business & Engineering, University of Kentucky, 2019.

<https://sites.google.com/site/ba762researchmethods/materials/handouts/typesofvalidity>.

Box, George E. P. "Robustness in the Strategy of Scientific Model Building." Technical Report #1954. Mathematics Research Center, University of Wisconsin-Madison, 1979.

Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16, no. 3 (2001): 199–231.

<https://dx.doi.org/10.1214/ss/1009213726>

Cardoso, Fatima, Laura J. van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delalogue, Jean-Yves Pierga, Etienne Brain, Sylvain

Causeret, Mauro DeLorenzi, Annuska M. Glas, Vassilis Golfinopoulos, Theodora Goulioti, Susan Knox, Erika Matos, Bart Meulemans, Peter A. Neijenhuis, Ulrike Nitz, Rodolfo Passalacqua, Peter Ravdin, Isabel T. Rubio, Mahasti Saghatchian, Tineke J. Smilde, Christos Sotiriou, Lisette Stork, Carolyn Straehle, Geraldine Thomas, Alastair M. Thompson, Jacobus M. van der Hoeven, Peter Vuylsteke, René Bernards, Konstantinos Tryfonidis, Emiel Rutgers, and Martine Piccart. "70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer." *New England Journal of Medicine* 375, no. 8 (2016): 717–729.

<https://dx.doi.org/10.1056/NEJMoa1602253>.

# ➤ References (3/8)

Chatfield, Chris. "Model Uncertainty, Data Mining and Statistical Inference." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158, no. 3 (1995): 419–466.

<https://dx.doi.org/10.2307/2983440>.

Cox, David R. "Role of Models in Statistical Analysis." *Statistical Science* 5, no. 2 (May 1990): 169–174.

<https://dx.doi.org/10.1214/ss/1177012165>

Doshi-Velez, Finale and Been Kim. *Towards a Rigorous Science of Interpretable Machine Learning*. 2017.

<https://arxiv.org/abs/1702.08608>.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth.

"The Reusable Holdout: Preserving Validity in Adaptive Data Analysis." *Science* 349, no. 6248 (2015): 636–638.

<https://dx.doi.org/10.1126/science.aaa9375>.

Efron, Bradley. "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation." *Journal of the American Statistical Association* 99, no. 467 (2004): 619–632.

<https://dx.doi.org/>

[10.1198/016214504000000692](https://doi.org/10.1198/016214504000000692).

Fisher, R. A. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922): 309–368.

<https://dx.doi.org/10.1098/rsta.1922.0009>



# References (4/8)

Gayo-Avello, Daniel. "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper": A Balanced Survey on Election Prediction using Twitter Data." 2012.

<https://arxiv.org/abs/1204.6441>.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457 (2009): 1012-1015.

<https://dx.doi.org/10.1038/nature07634>.

Hammerla, Nils Y., and Thomas Plötz. "Let's (Not) Stick Together: Pairwise Similarity Biases Cross-Validation in Activity Recognition." In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*, 1041-1051.

<https://dx.doi.org/10.1145/2750858.2807551>.

Jones, Matthew L. "How We Became Instrumentalists (Again): Data Positivism since World War II." *Historical Studies in the Natural Sciences* 48, no. 5 (2018): 673-684.

<https://dx.doi.org/10.1525/hsns.2018.48.5.673>.

Kass, Robert E. "Statistical Inference: The Big Picture." *Statistical Science* 26, no. 1 (2011): 1-9.

<https://dx.doi.org/10.1214/10-STS337>.

Keys, Os. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2, 88:1-88:22, 2018.

# > References (5/8)

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. "Prediction Policy Problems." *American Economic Review* 105, no. 5 (2015): 491-495.

<https://dx.doi.org/10.1257/aer.p20151023>.

Lanius, Candice. "Fact Check: Your Demand for Statistical Proof is Racist." Cyborgology blog, January 15, 2015.

<https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/>.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343, no. 6176 (2014): 1203-1205.

<https://dx.doi.org/10.1126/science.1248506>.

Lipton, Zachary C. "The Myth of Model Interpretability." *KDnuggets* 15, no. 13 (April 2015).

<https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html>.

Lipton, Zachary C. and Jacob Steinhardt. "Troubling trends in machine learning scholarship." 2018.

<https://arxiv.org/abs/1807.03341>.

Messerli, Franz H. "Chocolate Consumption, Cognitive Function, and Nobel Laureates." *The New England Journal of Medicine*, 367 (2012): 1562-1564. [doi:10.1056/NEJMon1211064](https://doi.org/10.1056/NEJMon1211064).

Mullainathan, Sendhil and Jann 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>.

# References (6/8)

Opsomer, Jean, Yuedong Wang, and Yuhong Yang. "Nonparametric Regression with Correlated Errors." *Statistical Science* 16, no. 2 (2001): 134–153. <https://dx.doi.org/10.1214/ss/1009213287>.

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

Patton, Michael Quinn. "The Nature, Niche, Value, and Fruit of Qualitative Inquiry." In *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*, 4th edition, 2–44. SAGE Publications, Inc., 2014. [https://uk.sagepub.com/sites/default/files/upm-binaries/64990\\_Patton\\_Ch\\_01.pdf](https://uk.sagepub.com/sites/default/files/upm-binaries/64990_Patton_Ch_01.pdf).

Rescher, Nicholas. *Predicting the Future: An Introduction to the Theory of Forecasting*. State University of New York Press, 1998.

Rose, Todd. *The End of Average: How We Succeed in a World That Values Sameness*. New York: HarperOne, 2016. See excerpt at <https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html>. Animated video: <https://vimeo.com/237632676>.

Rosset, Saharon, and Ryan J. Tibshirani. "From Fixed-X to Random-X Regression: Bias-Variance Decompositions, Covariance Penalties, and Prediction Error Estimation." *Journal of the American Statistical Association* (2019). <https://dx.doi.org/10.1080/01621459.2018.1424632>.

# ➤ References (7/8)

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References

Santillana, Mauricio, Wendong Zhang, Benjamin M. Althouse, and John W. Ayers. "What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?" *American Journal of Preventive Medicine* 47, no. 3 (2014): 341–347.  
<http://dx.doi.org/10.1016/j.amepre.2014.05.020>.

Shapiro, Ian. "Methods are like people: If you focus only on what they can't do, you will always be disappointed." In *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, edited by Dawn Langan Teele, 228–241. Yale University Press, 2014.

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

Spirtes, Peter and Kun Zhang. "Causal Discovery and Inference: Concepts and Recent Methodological Advances." *Applied Informatics* 3, no. 3 (2016): 1–28.  
<https://dx.doi.org/10.1186/s40535-016-0018-x>.

Tibshirani, Robert. "Recent Advances in Post-Selection Inference." Breiman Lecture, NIPS 2015 (9 December 2015)  
<http://statweb.stanford.edu/~tibs/ftp/nips2015.pdf>

# ➤ References (8/8)

van't Veer, Laura J., Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer." *Nature* 415, no. 6871 (2002): 530–536.  
<https://dx.doi.org/10.1038/415530a>.

Wallach, Hanna. "Computational Social Science ≠ Computer Science + Social Data." *Communications of the ACM* 61, no. 3 (2018): 42–44. <https://dx.doi.org/10.1145/3132698>.

Wasserman, Larry A. "Rise of the Machines." In *Past, Present, and Future of Statistical Science*, 525–536. Boca Raton, FL: Chapman and Hall/CRC, 2013.  
<http://www.stat.cmu.edu/~larry/Wasserman.pdf>

Watts, Duncan J. "The 'New' Science of Networks." *Annual Review of Sociology* 30 (2004): 243–270.  
<https://dx.doi.org/10.1146/annurev.soc.30.020404.104342>.

Wu, Shaohua, 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>.

- › Introduction
- › Meaning and measurement
- › Central tendency
- › Causality
- › Capturing variability
- › Cross-validation
- › Reflection
- › Future steps
- › References

# › Backup slides

# ➤ “True” models predict worse

➤ A linear data-generating process.

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

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

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

# › Proposal: 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: Alternative language

- › Retrodiction
- › Backtesting (retrodiction for testing)
- › Hindcasting (backtesting for forecasting)
- › In-sample vs. Out of-sample
- › Interpolation vs. Extrapolation
- › Diagnosis vs. Prognosis
- › Retrospective vs. Prospective

# ➤ But language not enough

- Introduction
- Meaning and measurement
- Central tendency
- Causality
- Capturing variability
- Cross-validation
- Reflection
- Future steps
- References
- Backup slides

## 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*

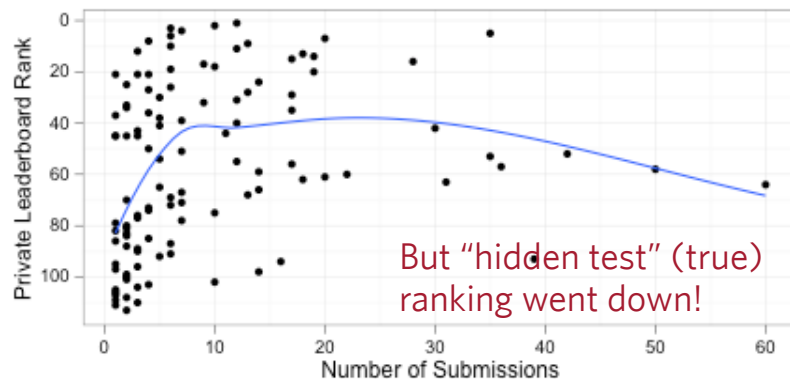
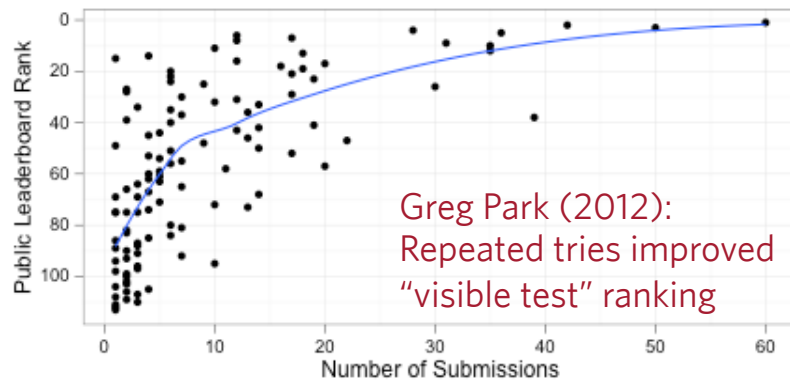
(I.e., using “backtest” in place of “predict” has not prevented financial analysts from unwitting overfitting)

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

# ➤ Overfitting on the test set

- Re-using a test set can overfit to the test set! (Dwork et al., 2015)
- Happens in Kaggle, which has public leaderboard (visible throughout) and private leaderboard (revealed only at end of competition)



# > Matrix bias-variance decomposition

$$\begin{aligned}
 \text{err}(\hat{\mu}) &= \frac{1}{n} \mathbb{E}_f \|Y - \hat{Y}\|_2^2 \\
 &= \frac{1}{n} \left[ \mathbb{E}_f \|Y\|_2^2 + \mathbb{E}_f \|\hat{Y}\|_2^2 - 2\mathbb{E}_f(Y^T \hat{Y}) \right] \\
 &= \frac{1}{n} \left[ \mathbb{E}_f \|Y\|_2^2 + \mathbb{E}_f \|\hat{Y}\|_2^2 - 2\text{tr} \mathbb{E}_f(Y \hat{Y}^T) \right] \\
 &\quad + \frac{1}{n} \left[ \mu^T \mu + \mathbb{E}_f(\hat{Y})^T \mathbb{E}_f(\hat{Y}) + 2\text{tr} \mu \mathbb{E}_f(\hat{Y})^T \right] \\
 &\quad + \frac{1}{n} \left[ -\mu^T \mu - \mathbb{E}_f(\hat{Y}) \mathbb{E}_f(\hat{Y})^T - 2\mu^T \mathbb{E}_f(\hat{Y}) \right] \\
 &= \frac{1}{n} \left[ \text{tr} \Sigma + \|\mu - \mathbb{E}(\hat{Y})\|_2^2 + \text{tr} \text{Var}_f(\hat{Y}) - 2\text{tr} \text{Cov}_f(Y, \hat{Y}) \right]
 \end{aligned}$$

# › Critical technical practice (1)

- › Agre (1997) describes “mov[ing] intellectually from AI to the social sciences — that is, to stop thinking the way that AI people think, and to start thinking the way that social scientists think...”
- › **“Criticisms of [AI], no matter how sophisticated and scholarly they might be, are certain to be met with the assertion that the author simply fails to understand a basic point... even though I was convinced that the field was misguided and stuck, it took tremendous effort and good fortune to understand how and why... I spent several years attempting to reform the field by providing it with the critical methods it needed — a critical technical practice.”**

## ➤ Critical technical practice (2)

- Introduction
  - Meaning and measurement
  - Central tendency
  - Causality
  - Capturing variability
  - Cross-validation
  - Reflection
  - Future steps
  - References
- “As an AI practitioner already well immersed in the literature, I had incorporated the field's taste for technical formalization so thoroughly into my own cognitive style that I literally could not read the literatures of nontechnical fields at anything beyond a popular level. The problem was not exactly that I could not understand the vocabulary, but that **I insisted on trying to read everything as a narration of the workings of a mechanism.**”
  - “At first I found [nontechnical] texts impenetrable, not only because of their irreducible difficulty but also because I was still tacitly attempting to read everything as a specification for a technical mechanism... My first intellectual breakthrough came when, for reasons I do not recall, **it finally occurred to me to stop translating these strange disciplinary languages into technical schemata, and instead simply to learn them on their own terms.**”

## > Critical technical practice (3)

- > “I still remember the **vertigo** I felt during this period; I was speaking these strange disciplinary languages, in a wobbly fashion at first, without knowing what they meant -- without knowing what *sort* of meaning they had.”
- > “in retrospect this was the period during which **I began to ‘wake up’, breaking out of a technical cognitive style that I now regard as extremely constricting.**”

## ➤ Critical technical practice (4)

- Introduction
  - Meaning and measurement
  - Central tendency
  - Causality
  - Capturing variability
  - Cross-validation
  - Reflection
  - Future steps
  - References
- “Without the idea that ideologies and social structures can be reproduced through a myriad of unconscious mechanisms such as linguistic forms and bodily habits, all critical analysis may seem like accusations of conscious malfeasance. **Even sociological descriptions that seem perfectly neutral to their authors can seem like personal insults to their subjects if they presuppose forms of social order that exist below the level of conscious strategy and choice.**”