

machine learning

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

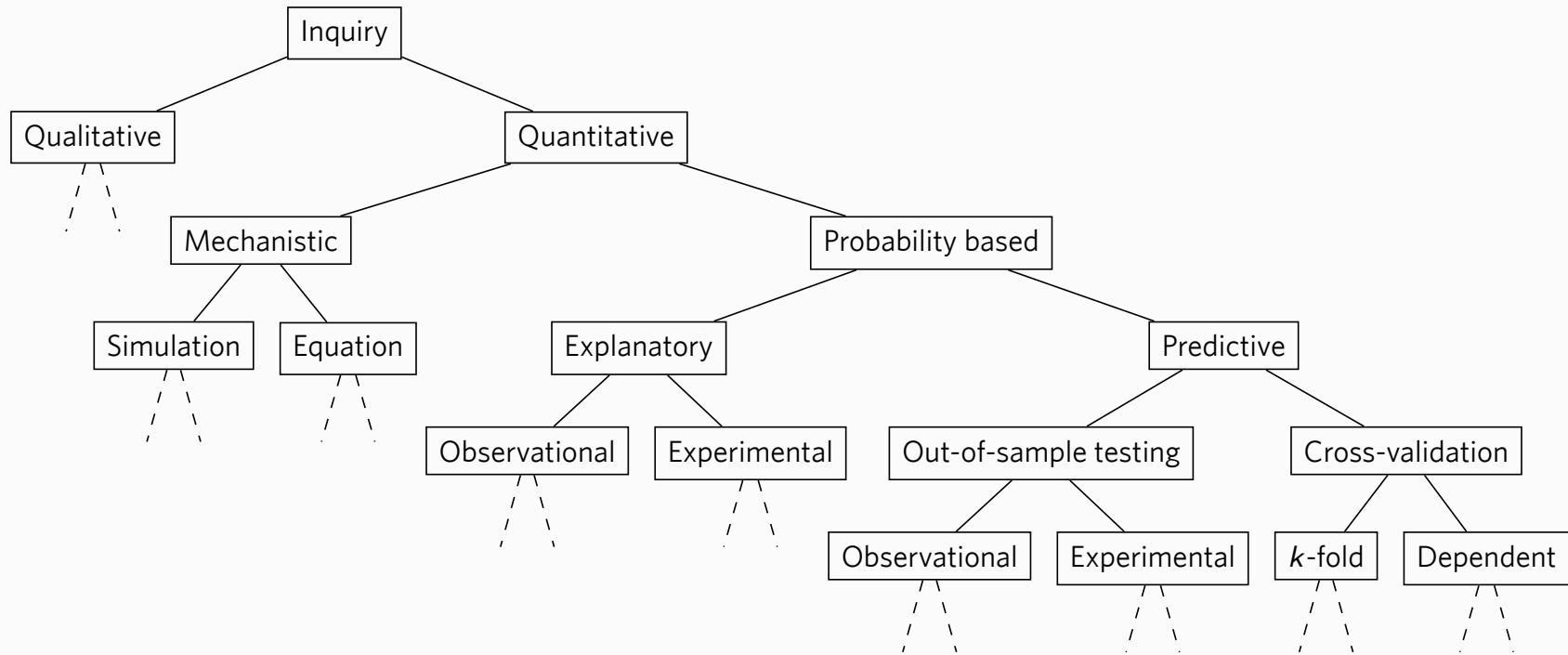
# A Hierarchy of Limitations in Machine Learning

---

**Momin M. Malik**, Data Science Postdoctoral Fellow, Berkman Klein Center for Internet & Society at Harvard University (on leave)

NYU Center for Data Science, Math and Democracy Seminar, 05 Oct 2020

# ➤ The “hierarchy”



➤ Introduction

➤ Quantitative:  
Meanings,  
measurement,  
and constructs

➤ Probability-  
based: Central  
tendency,  
variability

➤ Predictive:  
Correlation vs.  
causation

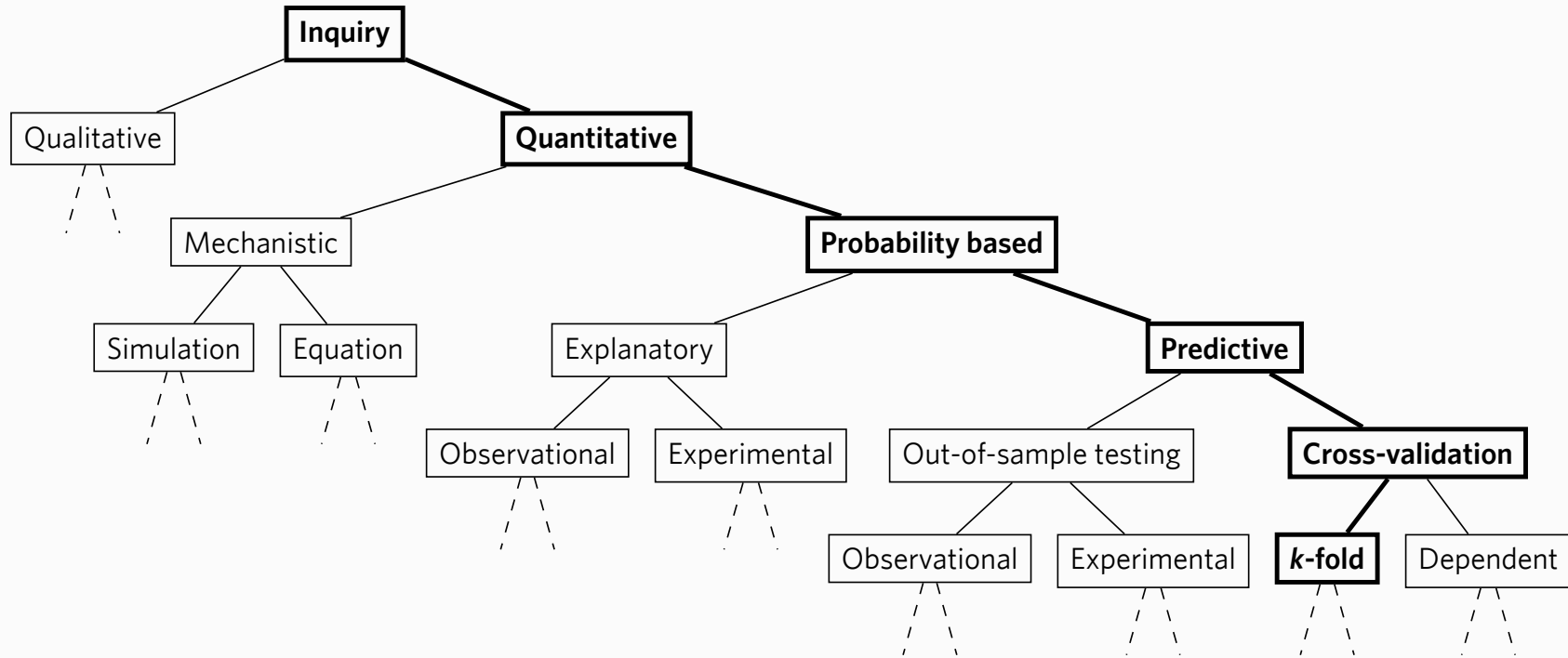
➤ Cross-  
validation:  
Dependencies  
and optimism

➤ Summary

➤ References

# > Typical machine learning

- > Introduction
- > Quantitative: Meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References



# > Objectives

- > Introduction
  - > Quantitative: Meanings, measurement, and constructs
  - > Probability-based: Central tendency, variability
  - > Predictive: Correlation vs. causation
  - > Cross-validation: Dependencies and optimism
  - > Summary
  - > References
- > What are the trade-offs associated with each branch?
  - > When are we justified traveling down the machine learning (“predictive”) branch?
  - > What are the consequences of using machine learning when it is *not* justified?

# › Outline

› Introduction

› Quantitative:  
Meanings,  
measurement,  
and constructs

› Probability-  
based: Central  
tendency,  
variability

› Predictive:  
Correlation vs.  
causation

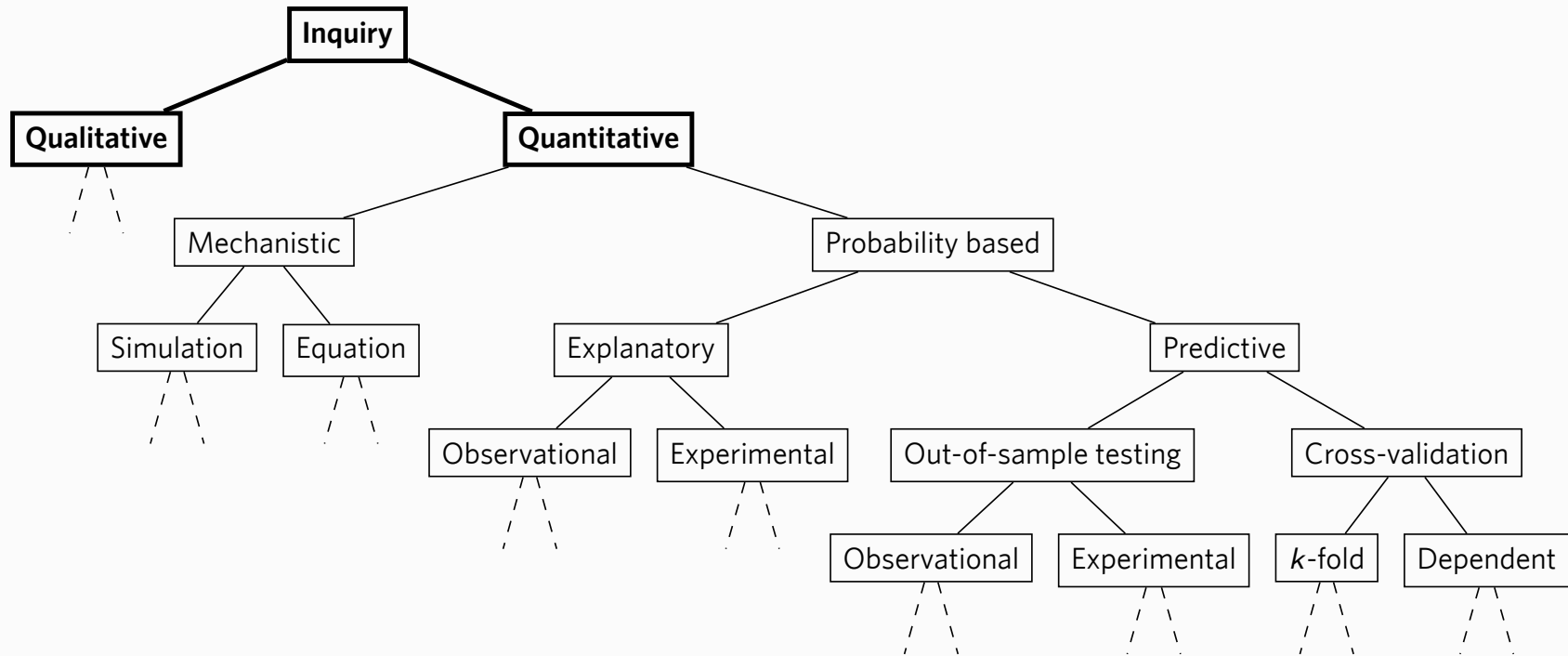
› Cross-  
validation:  
Dependencies  
and optimism

› Summary

› References

1. Quantitative: Problems of meanings, measurement, and constructs
2. Probability-based: Problems of central tendencies, variability as nuisance
3. Predictive: Problems of correlation over causation
4. Cross-validation: Problems of dependencies and optimism

# 1. Quantitative



- > Introduction
- > Quantitative: Meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References

# > 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
- > Quantitative: Meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References

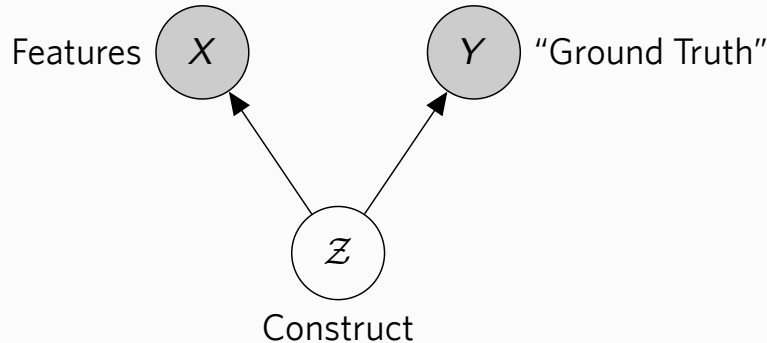
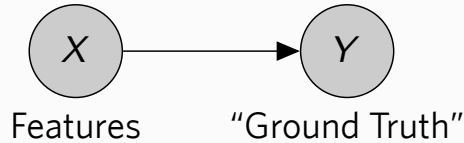
# > Meaning-making

- > Introduction
- > Quantitative: Meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References

“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)



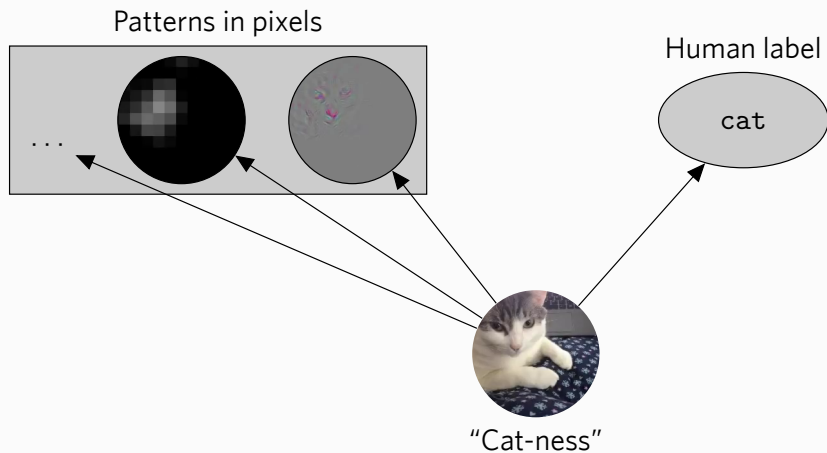
# ➤ Measurement and constructs



- *Constructs*: primitives of social science
  - What we care about
  - Often unobservable (and hypothetical/subjective, e.g. friendship)
  - Proxies always give errors (for binary-valued constructs: false negatives and false positives)

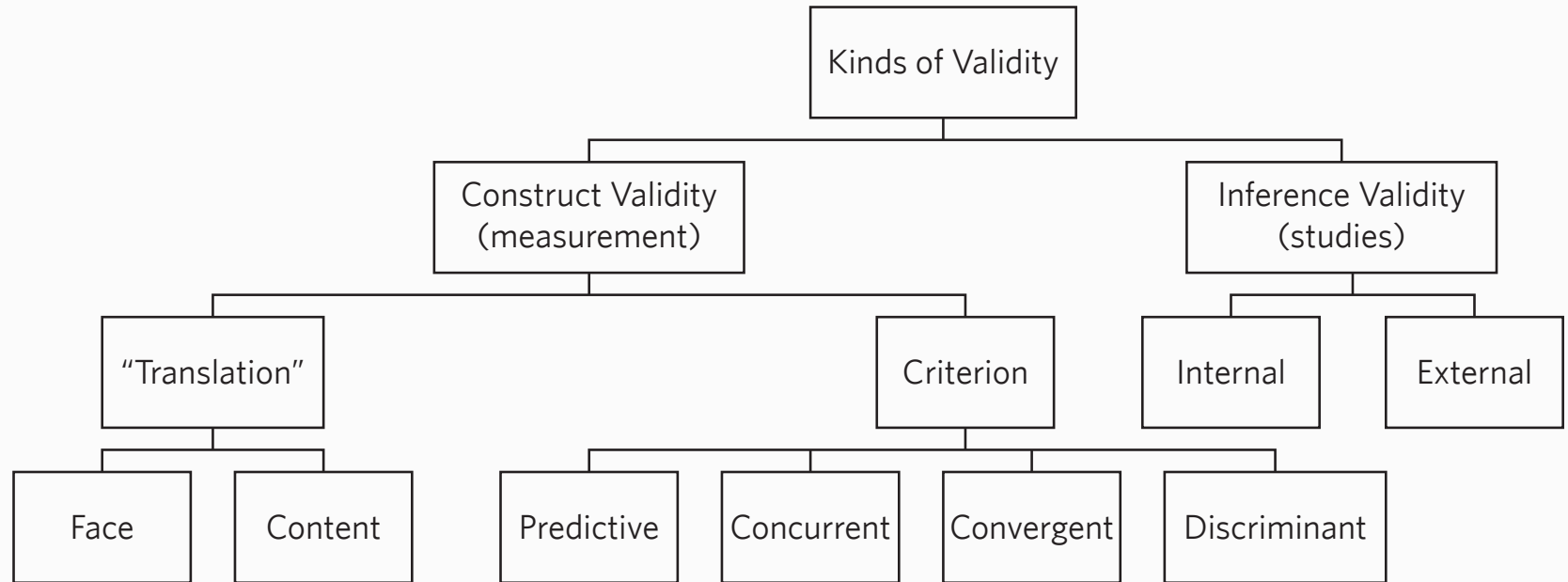
# Constructs: Subjective, multifaceted

- Introduction
- Quantitative: Meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



# Validating measurements

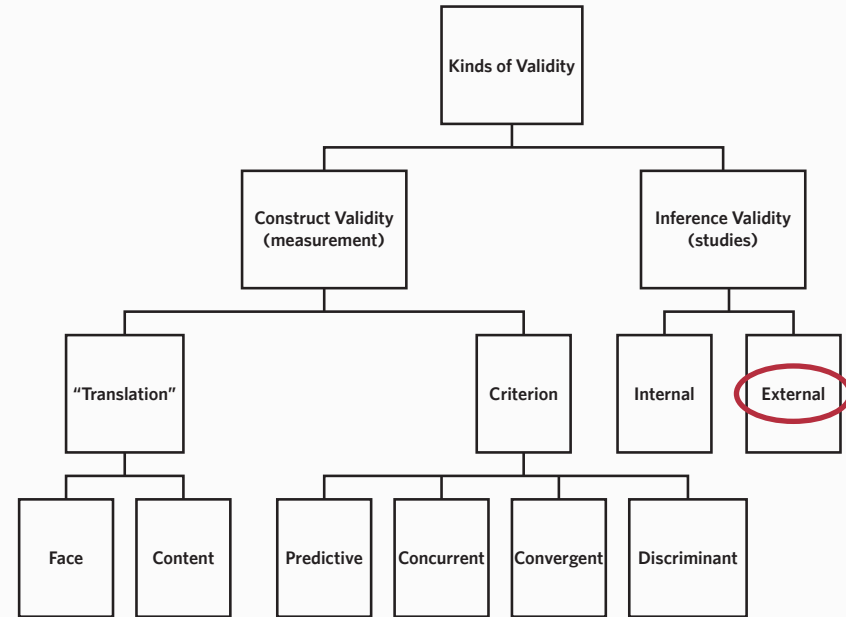
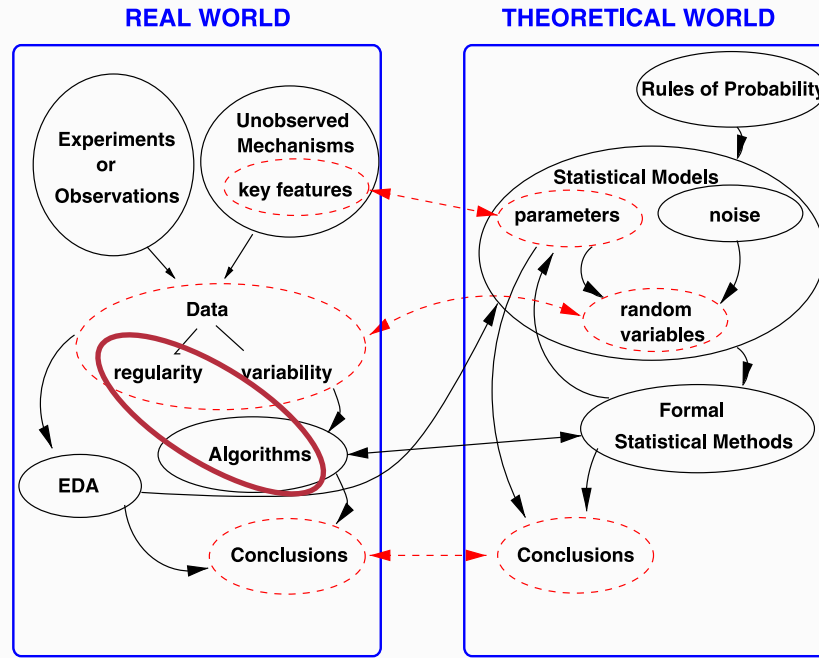
- Introduction
- Quantitative: Meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



Adapted from Borgatti, 2012  
A Hierarchy of Limitations in ML

# ➤ (ML: Only external validity)

- Introduction
- Quantitative: Meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



Kass, 2011, *Stat. Sci.*

Adapted from Borgatti, 2012

# › “Thin description”

› Introduction

› Quantitative:  
Meanings,  
measurement,  
and constructs

› Probability-  
based: Central  
tendency,  
variability

› Predictive:  
Correlation vs.  
causation

› Cross-  
validation:  
Dependencies  
and optimism

› Summary

› References

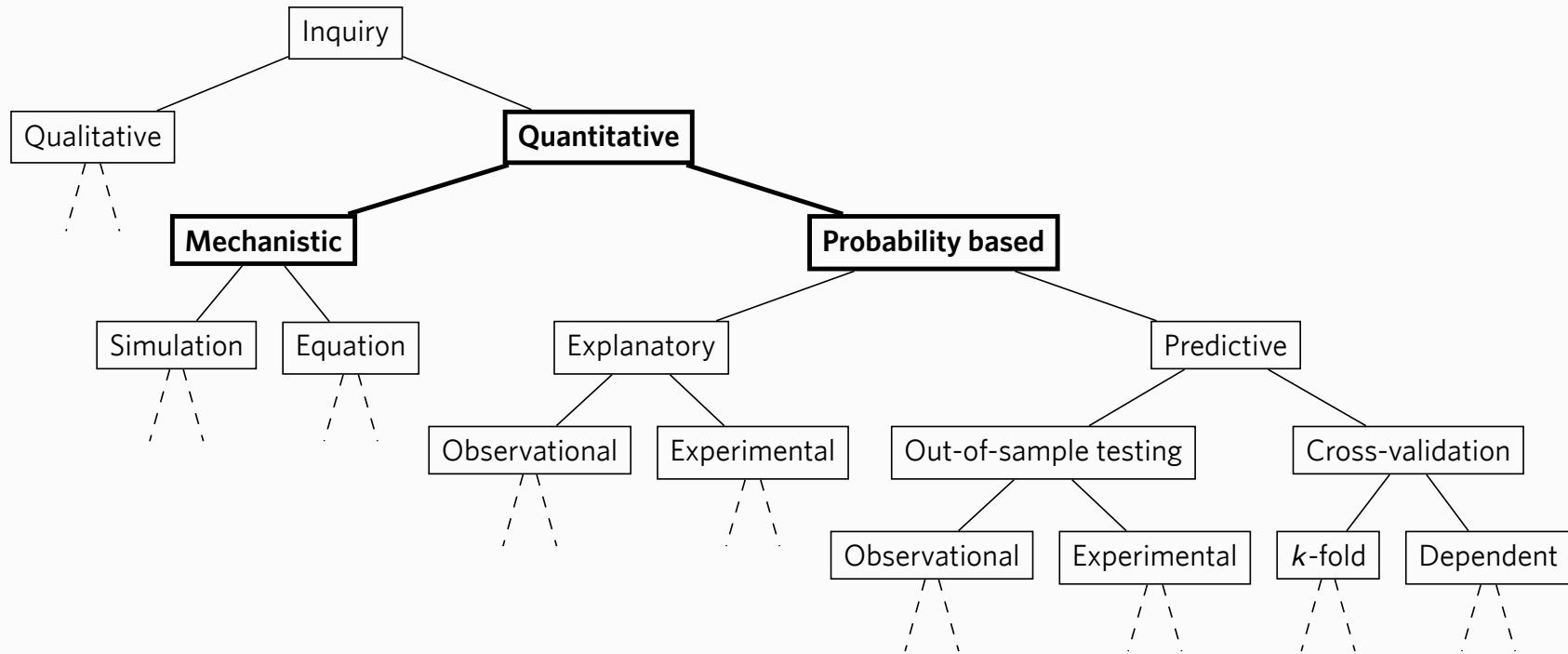
“what exactly is thin description? In a thin world, surfaces should be valid and deep meanings superfluous...

“The [quantitative social science] focus on behavior was a strategy to make social science into real science, something more like physics and less subject to values and prejudices because restricted to observable phenomena of the sort that could be registered by instruments. The **sacrifice of human meaning** seemed **not just a price worth paying for solid results, but the liberating essence of a proper objective methodology that now would rise above stubborn tradition and invisible culture.**” (Porter, 2012)

# > Consequences

- > Introduction
  - > Quantitative: Meanings, measurement, and constructs
  - > Probability-based: Central tendency, variability
  - > Predictive: Correlation vs. causation
  - > Cross-validation: Dependencies and optimism
  - > Summary
  - > References
- > The world is, ultimately, “thick”: the same behavior can have infinitely many different meanings
  - > What is it we ultimately care about? Relating to human experience? Or “solid results”?
  - > Quantification requires choosing one set of meanings; nothing subsequent can “unpack” this (it has to be done again), and there is never one “best” meaning
  - > Quantification solidifies that meaning, which lets us build upwards
  - > Quantification can serve to de-legitimize other meanings

# 2. Probability-based



- > Introduction
- > Quantitative: meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References

# > Probability: signal and noise

- > “Probability is used in two distinct, although interrelated, ways in statistics, phenomenologically to describe haphazard variability arising in the real world and epistemologically to represent uncertainty of knowledge.” (Cox, 1990)
- > Implies a philosophical commitment: the world is made up of entities that are interchangeable, where the important thing is *central tendency amidst variability*

> Introduction

> Quantitative: meanings, measurement, and constructs

> Probability-based: Central tendency, variability

> Predictive: Correlation vs. causation

> Cross-validation: Dependencies and optimism

> Summary

> References



# › The world as a data matrix

› Introduction

› Quantitative:  
meanings,  
measurement,  
and constructs

› Probability-  
based: Central  
tendency,  
variability

› Predictive:  
Correlation vs.  
causation

› Cross-  
validation:  
Dependencies  
and optimism

› Summary

› References

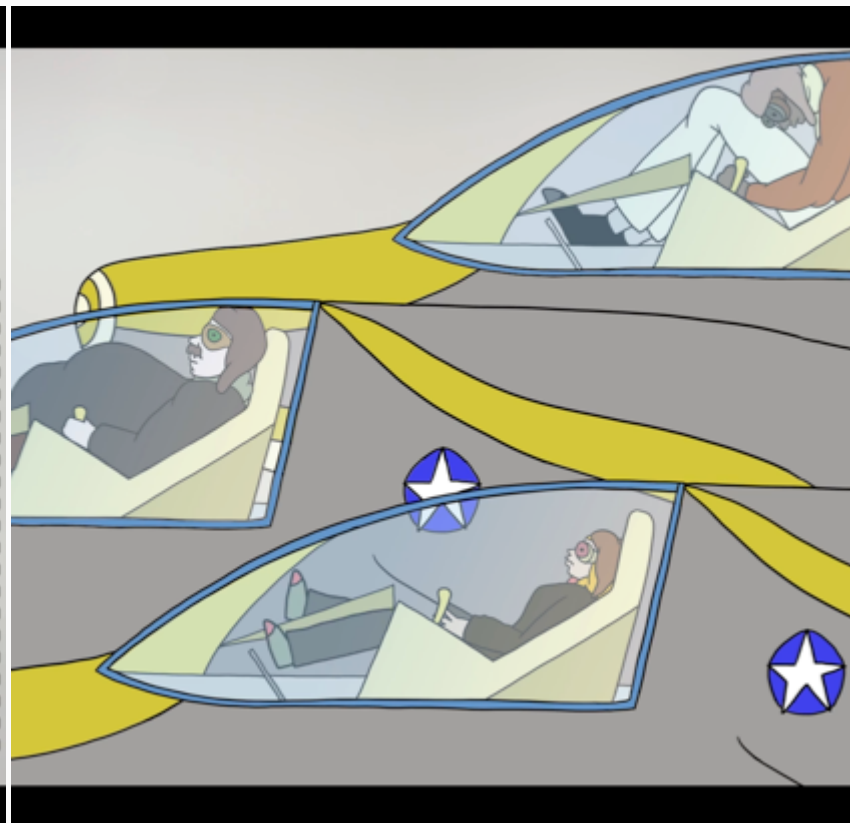
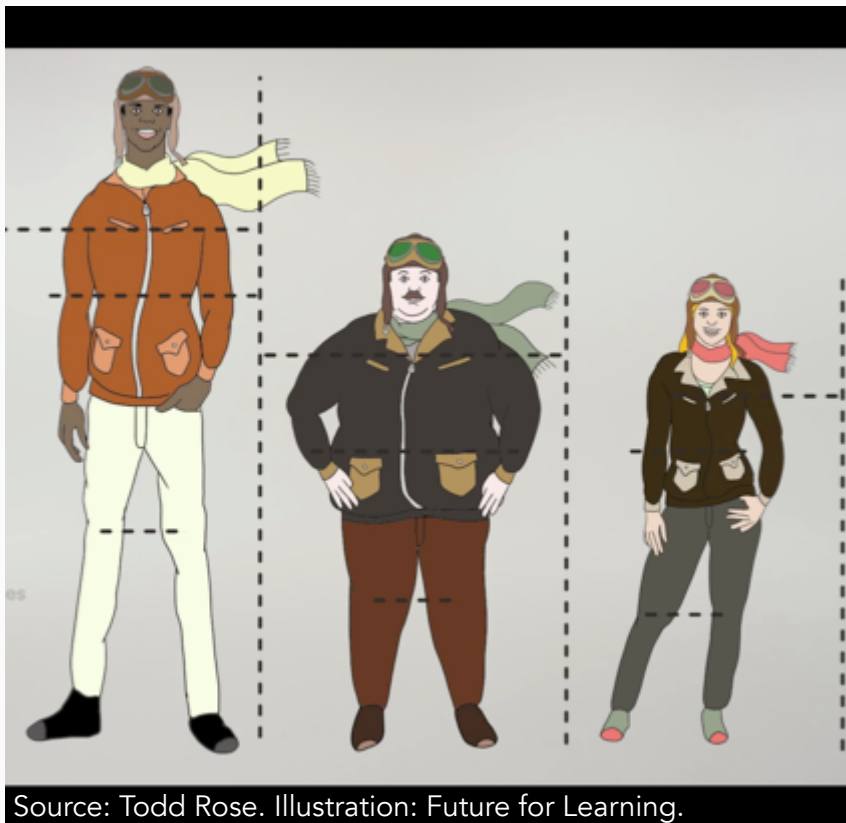
**“it is striking how absolutely these assumptions contradict those of the major theoretical traditions of sociology.**

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)

# ➤ Concretely: “Flaw of averages”

- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



Source: Todd Rose. Illustration: Future for Learning.

# › Consequences

- › These problems are not necessarily unique to probability-based modeling
  - SIR equations are “equilibrium” solution
  - Agent-based modeling often has interchangeable agents, and summarizes outcomes over multiple simulations with summary statistics
- › Neither statistics nor machine learning can do anything with an  $n$  of 1: cannot account for individuality, nor do anything with it
- › Planning to the central tendency punishes outliers (Keyes 2018)

› Introduction

› Quantitative: meanings, measurement, and constructs

› Probability-based: Central tendency, variability

› Predictive: Correlation vs. causation

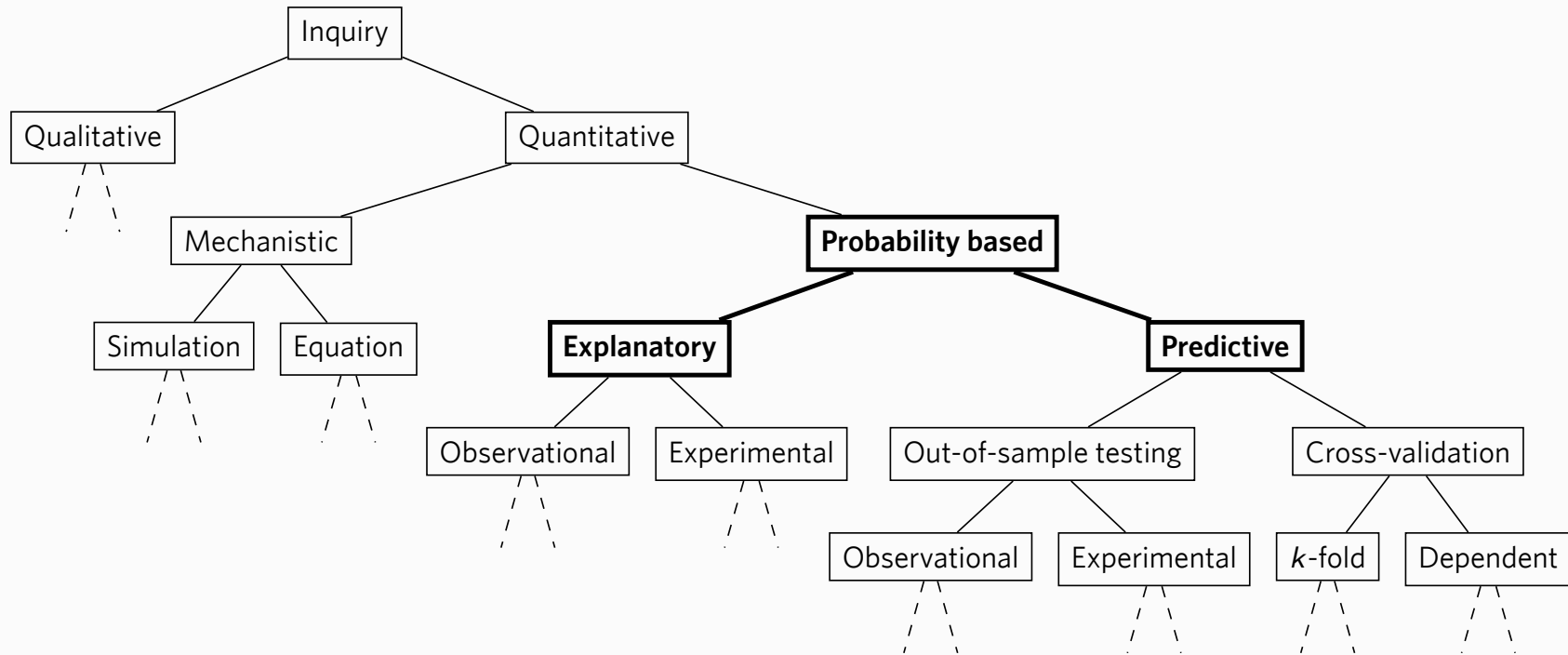
› Cross-validation: Dependencies and optimism

› Summary

› References

# 3. Predictive

- > Introduction
- > Quantitative: meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References

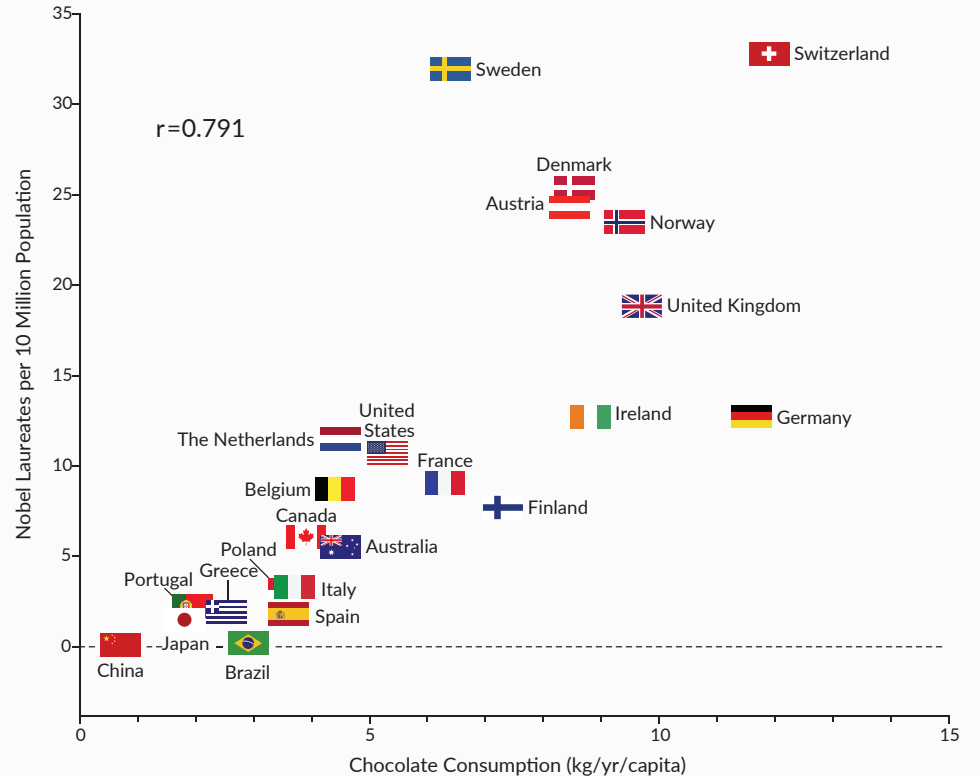


# > “Prediction” is not prediction!

- > *“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.” (Gayo-Avello, “I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper”, 2012)

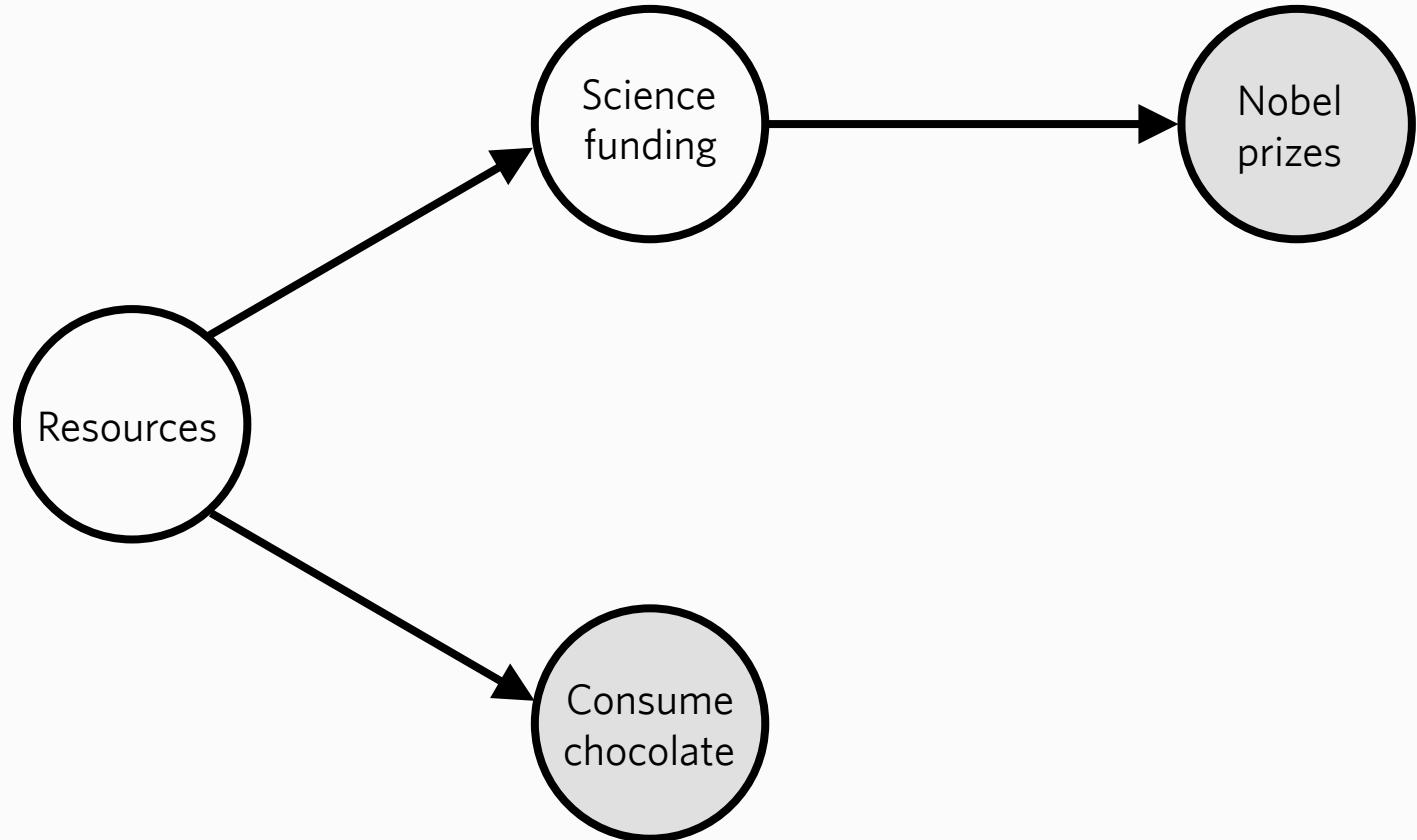
# ➤ “Predictions” are correlations

- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



Messerli, 2012, *NEJM*

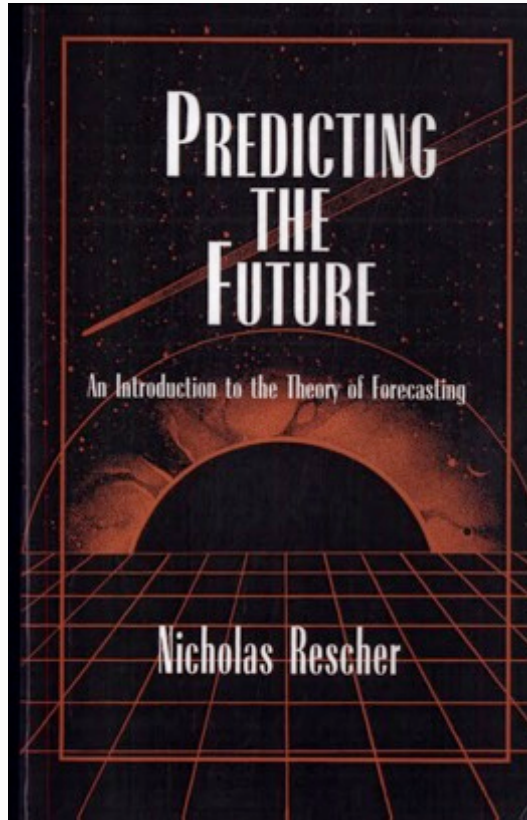
# > Cause is resources



- > Introduction
- > Quantitative: meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References

# ➤ Not an obvious usage of “predict”

- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



A Hierarchy of Limitations in 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/INFERENTIAL</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

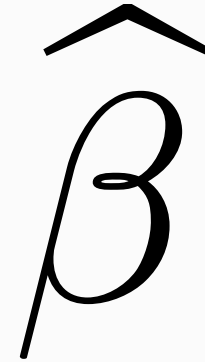


# > Creates two types of modeling!



Correlations may “predict” well

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

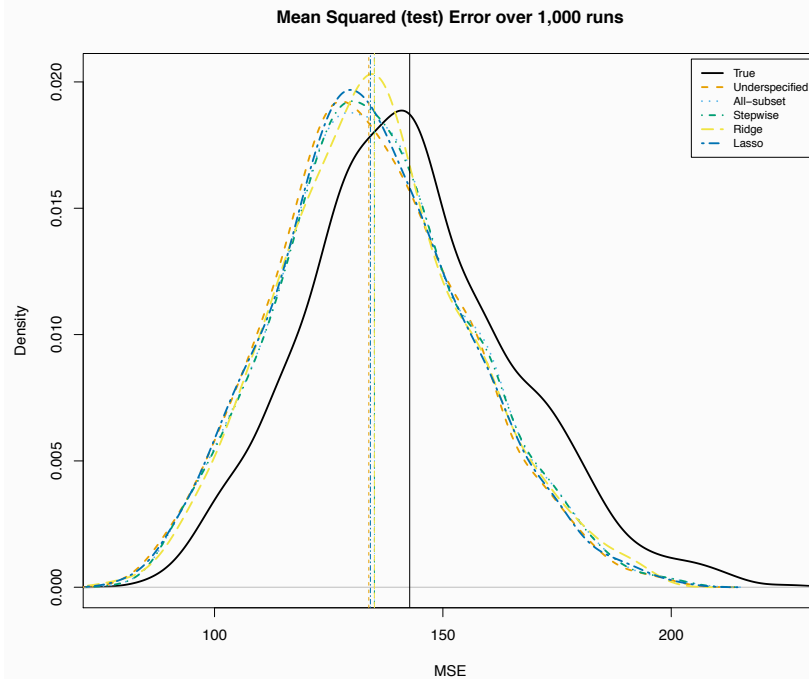
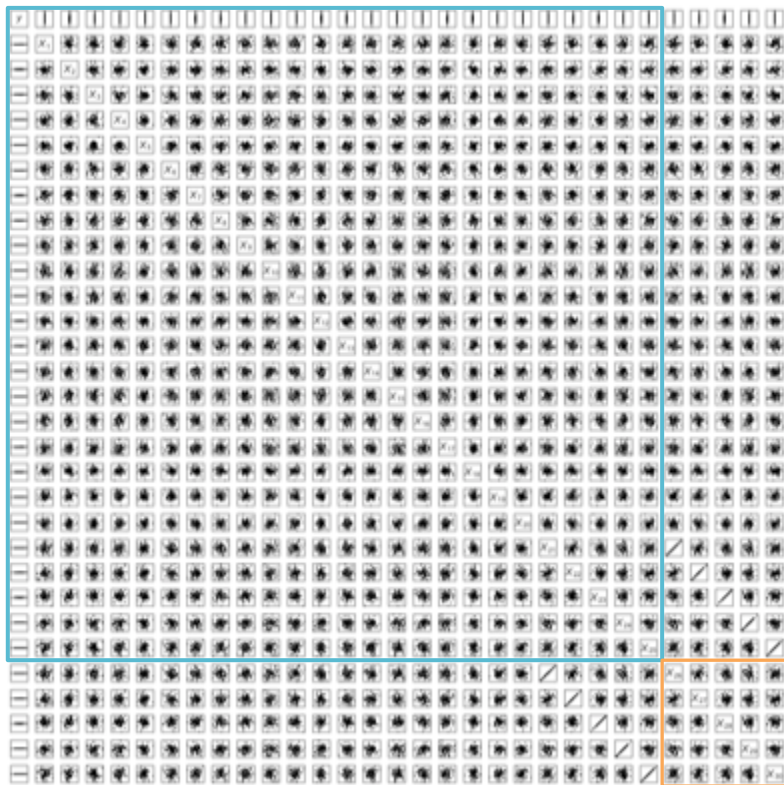


Informative models may not fit well

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

# ➤ “True” model can predict worse!

- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



Simulation of Shmueli, 2010, *Stat. Sci*

## ➤ Sometimes, people want causality

- Introduction
  - Quantitative: meanings, measurement, and constructs
  - Probability-based: Central tendency, variability
  - Predictive: Correlation vs. causation
  - Cross-validation: Dependencies and optimism
  - Summary
  - References
- “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)

# > Correlations and injustice

- > Introduction
- > Quantitative: meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References



Julius C. Chappelle proposed a bill in Massachusetts to ban charging Black people more for life insurance

A lawyer opposing the bill "cited statistics from around the nation showing shorter life spans for blacks, including 1870 census figures showing a 17.28 death rate for 'colored people' against 14.74 for whites. These numbers, Williams argued, and not any 'discrimination on the ground of color' motivated insurers' rates. It was a 'matter of business,' and any interference, he warned ominously and presciently, 'would probably cut off insurance entirely from the colored race.'"

# > Correlations and injustice

- > Introduction
- > Quantitative: meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References



“Chappelle’s allies noted that Williams’s statistics, while bleak enough, answered the wrong question. The question was not whether blacks in slavery or adjusting to freedom were poor insurance risks, or even whether southern blacks were poor risks. The question was African Americans’ potential for equality and specifically the present and future state of Massachusetts’ African Americans—about whom no statistics had been offered by either side.” (Bouk, 2015)

# ➤ Sometimes, causality affects prediction

➤ Introduction

➤ Quantitative: meanings, measurement, and constructs

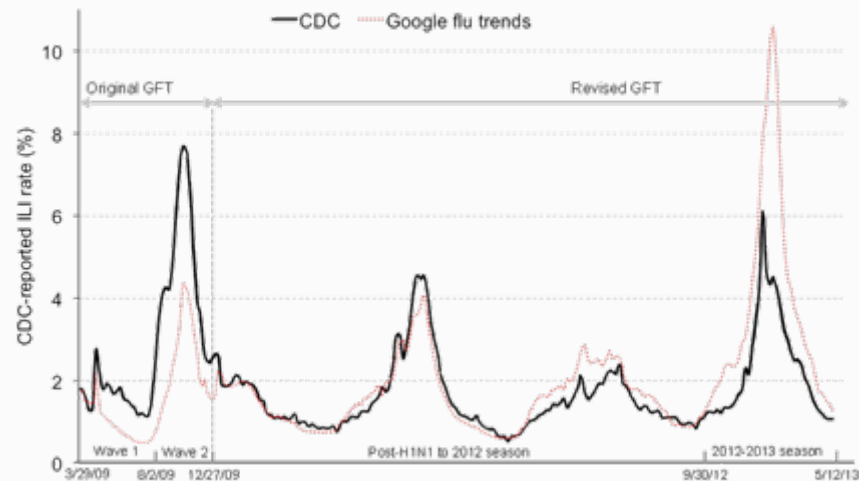
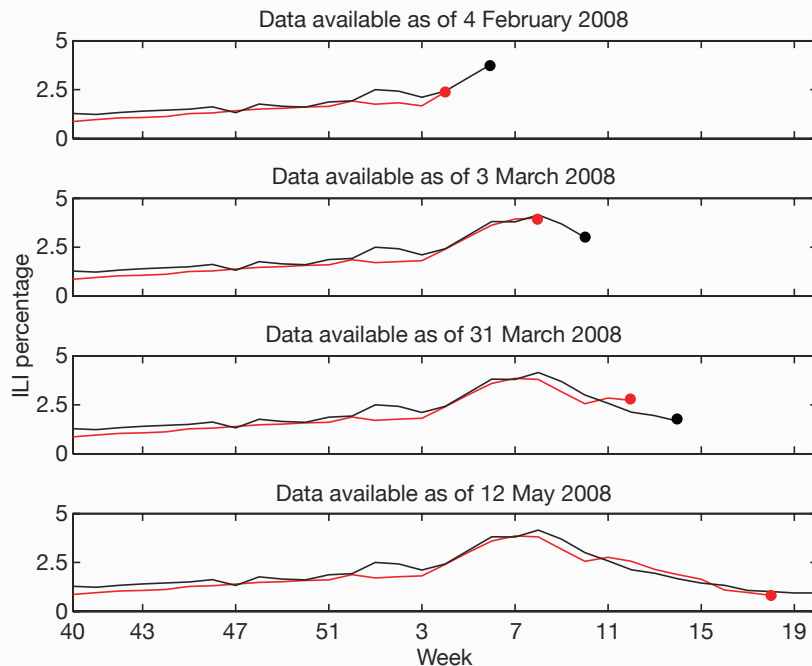
➤ Probability-based: Central tendency, variability

➤ Predictive: Correlation vs. causation

➤ Cross-validation: Dependencies and optimism

➤ Summary

➤ References

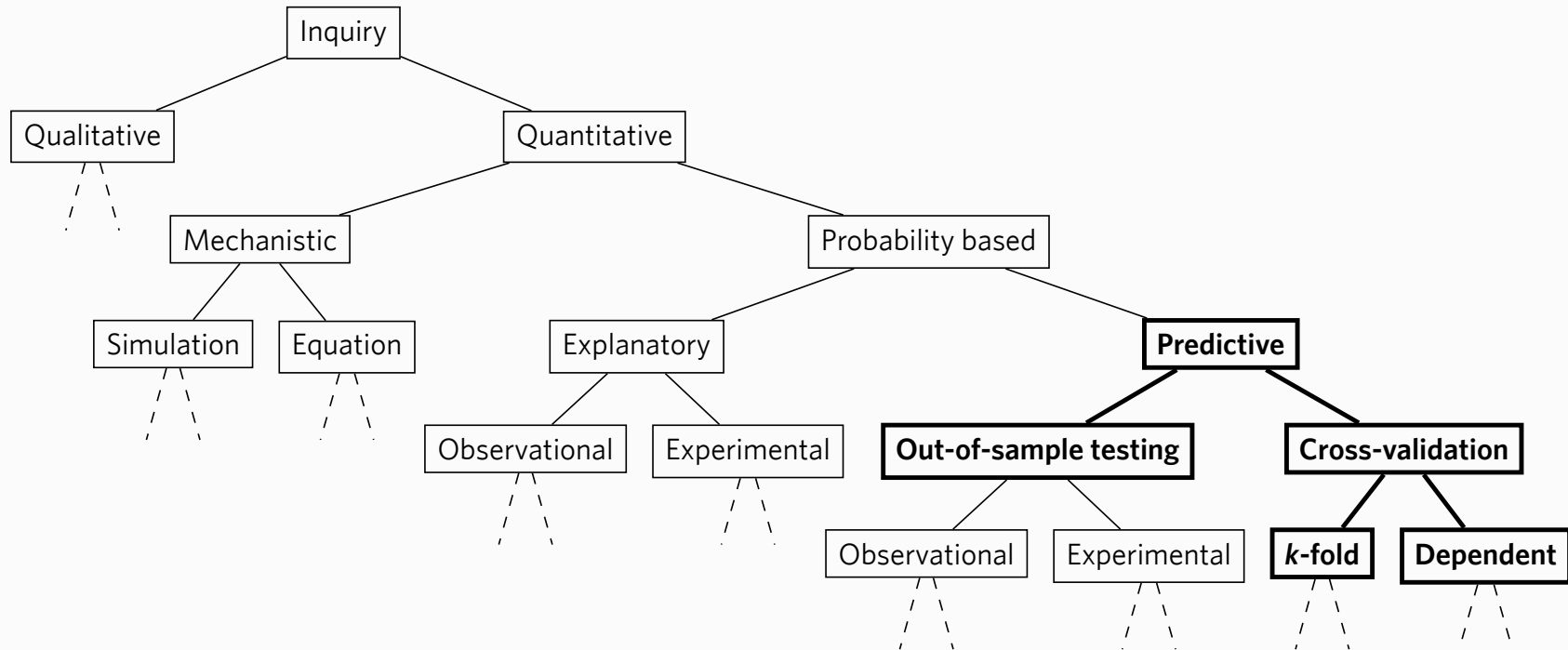


Ginsberg et al., 2012, *Nature*

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

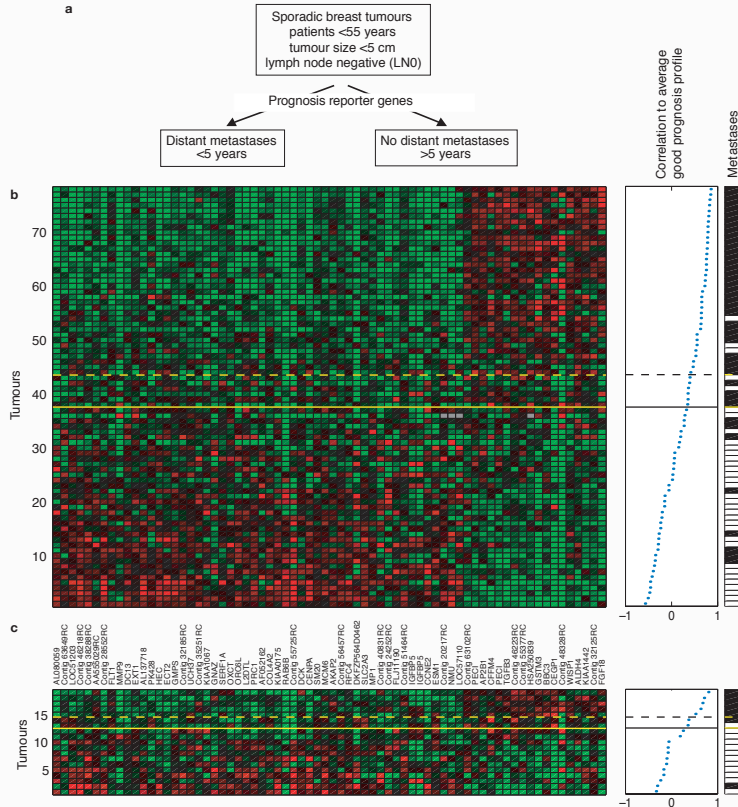
# 4. Cross-validation

- > Introduction
- > Quantitative: meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References



# ➤ Real-world testing of ML results

- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References

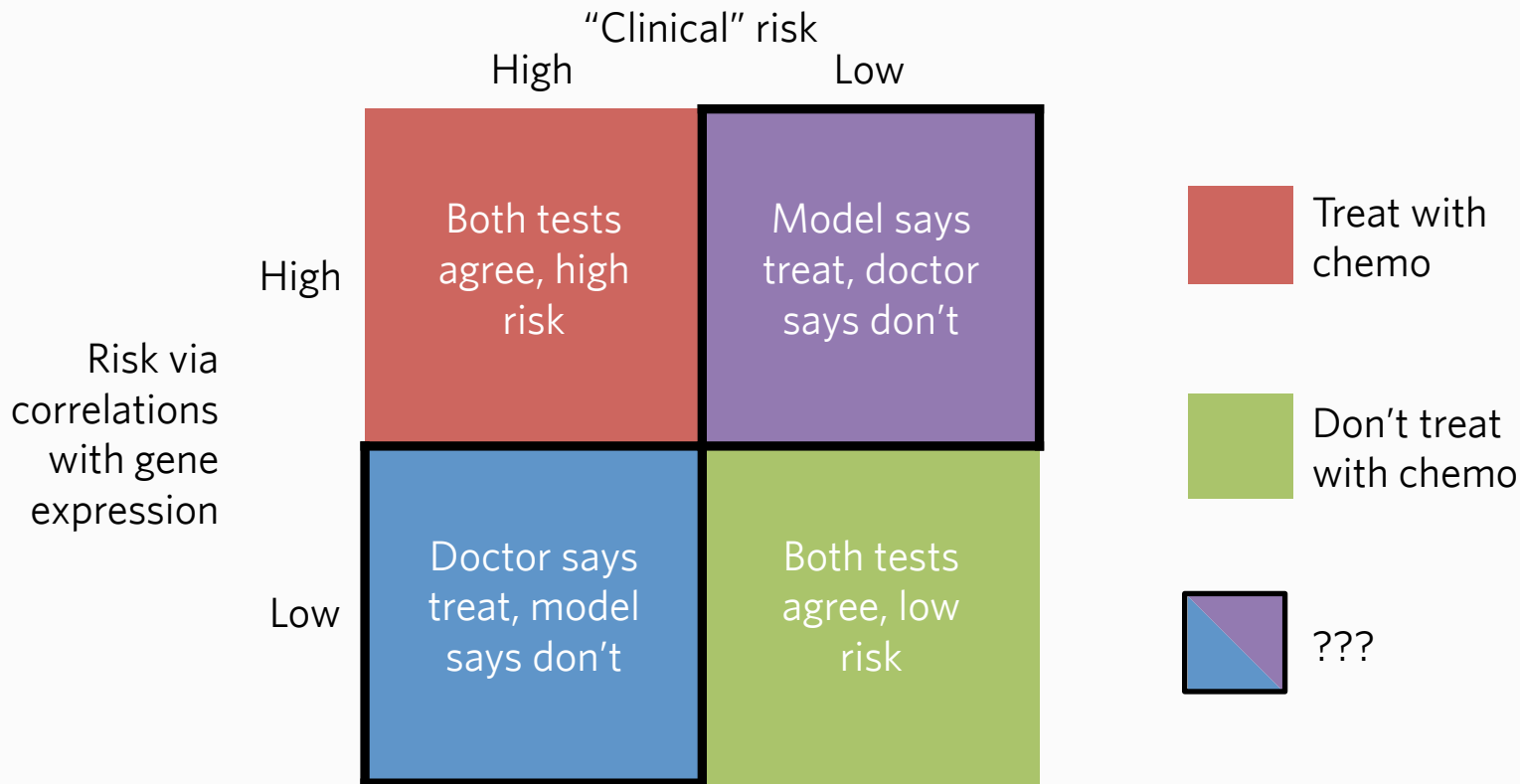


- van't Veer et al. (2002) found 70 genes correlated with developing breast cancer
- Of course the correlations were optimal, post-hoc. But did it generalize?



# Implementation testing

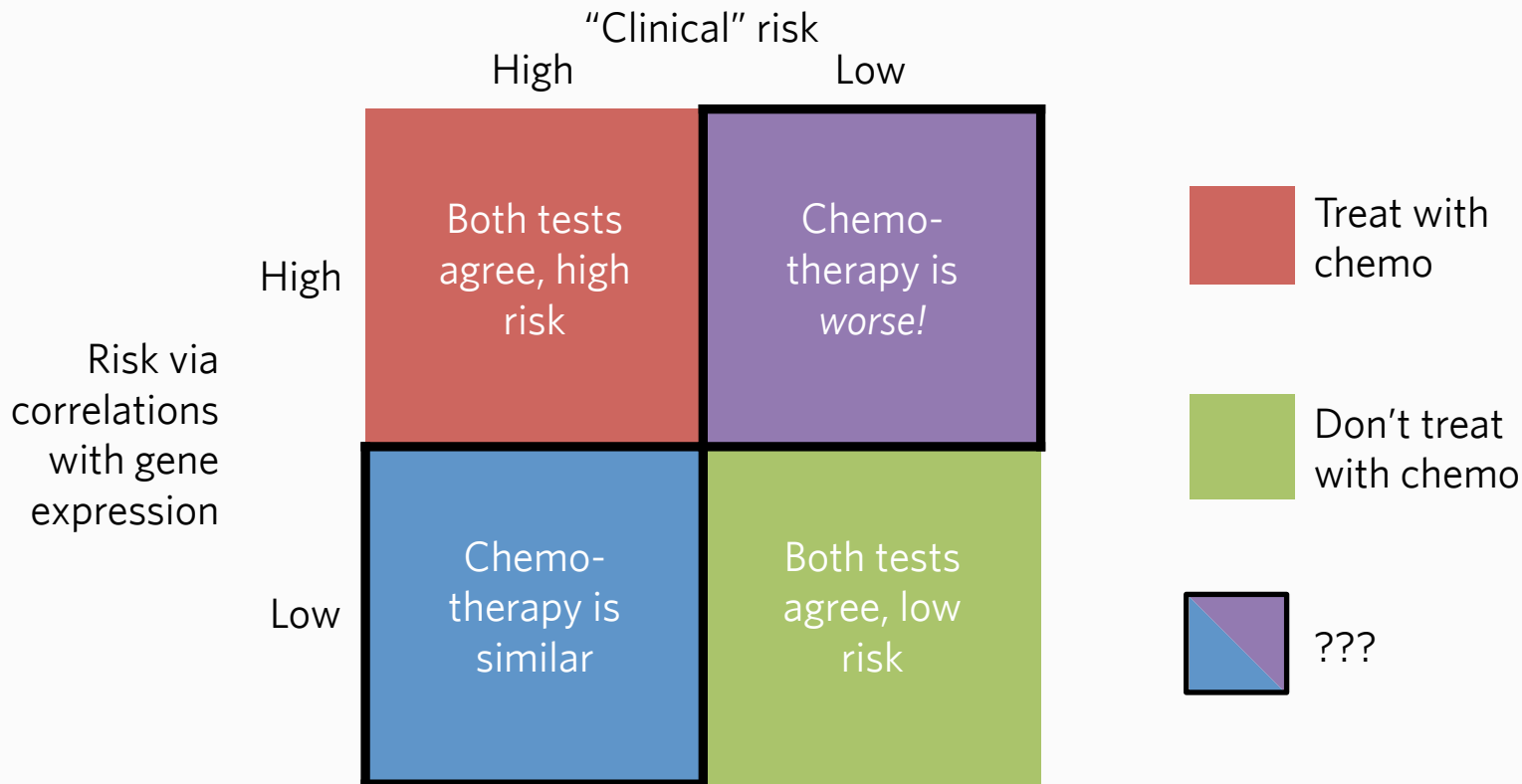
- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



Cardoso et al., 2016, *NEJM*

# ➤ Implementation testing

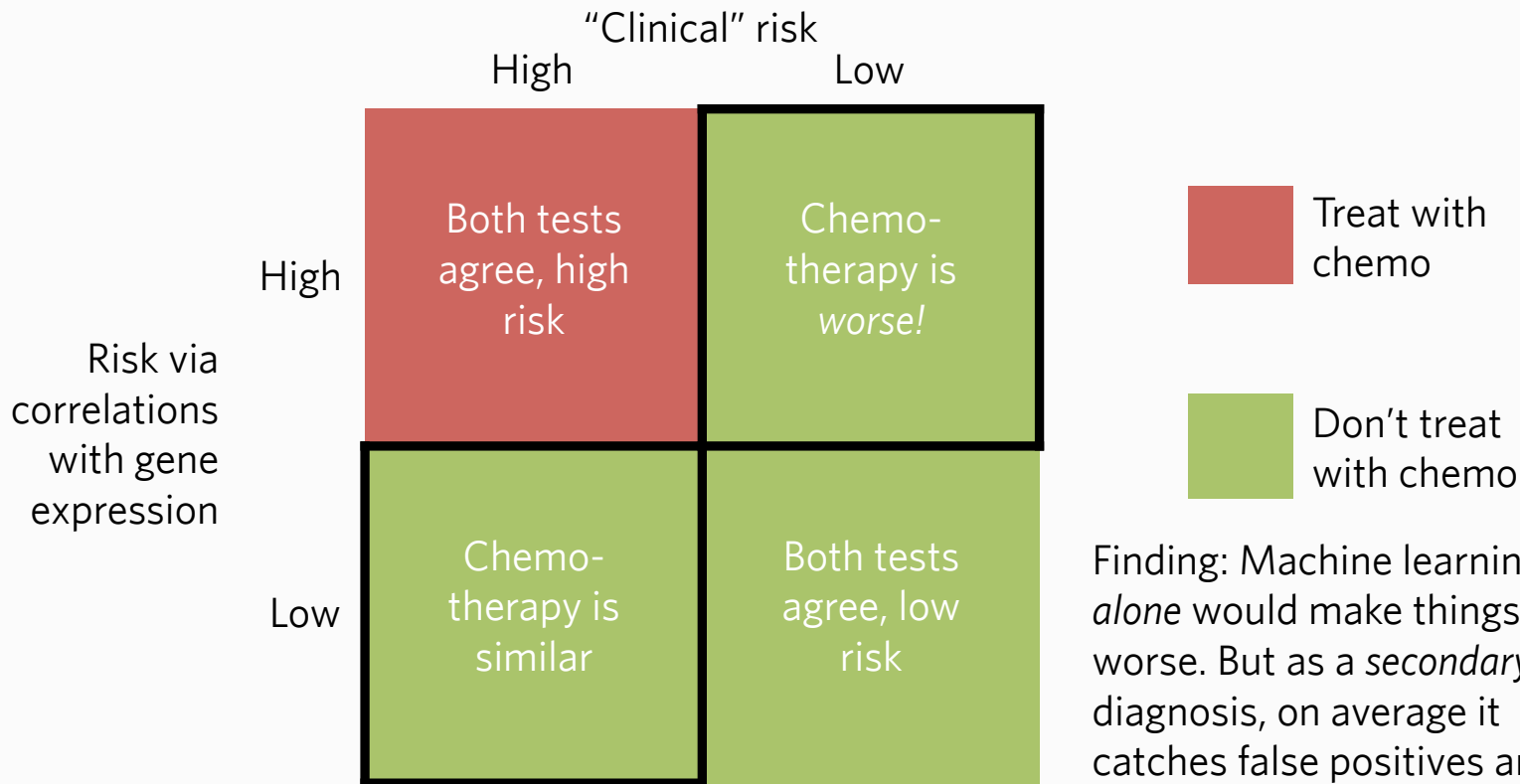
- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



Cardoso et al., 2016, *NEJM*

# > Implementation testing

- > Introduction
- > Quantitative: meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References

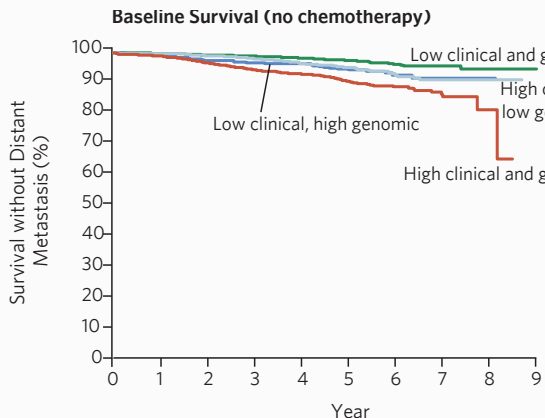


Finding: Machine learning *alone* would make things worse. But as a *secondary* diagnosis, on average it catches false positives and avoids unhelpful chemo!

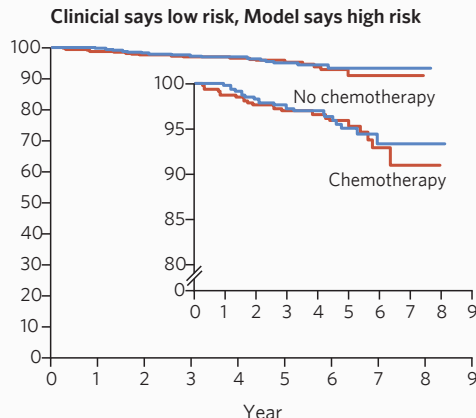
Cardoso et al., 2016, *NEJM*

# Implementation testing: Details

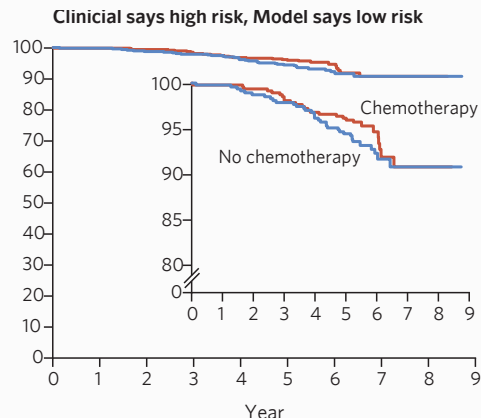
- ▶ Introduction
- ▶ Quantitative: meanings, measurement, and constructs
- ▶ Probability-based: Central tendency, variability
- ▶ Predictive: Correlation vs. causation
- ▶ Cross-validation: Dependencies and optimism
- ▶ Summary
- ▶ References



> 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

# › Generalizability through CV

- › Non-experimentally, generalizability is shown *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*

› Introduction

› Quantitative:  
meanings,  
measurement,  
and constructs

› Probability-  
based: Central  
tendency,  
variability

› Predictive:  
Correlation vs.  
causation

› Cross-  
validation:  
Dependencies  
and optimism

› Summary

› References

# > 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):

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

> Introduction

> Quantitative:  
meanings,  
measurement,  
and constructs

> Probability-  
based: Central  
tendency,  
variability

> Predictive:  
Correlation vs.  
causation

> Cross-  
validation:  
Dependencies  
and optimism

> Summary

> References

# > 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:

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

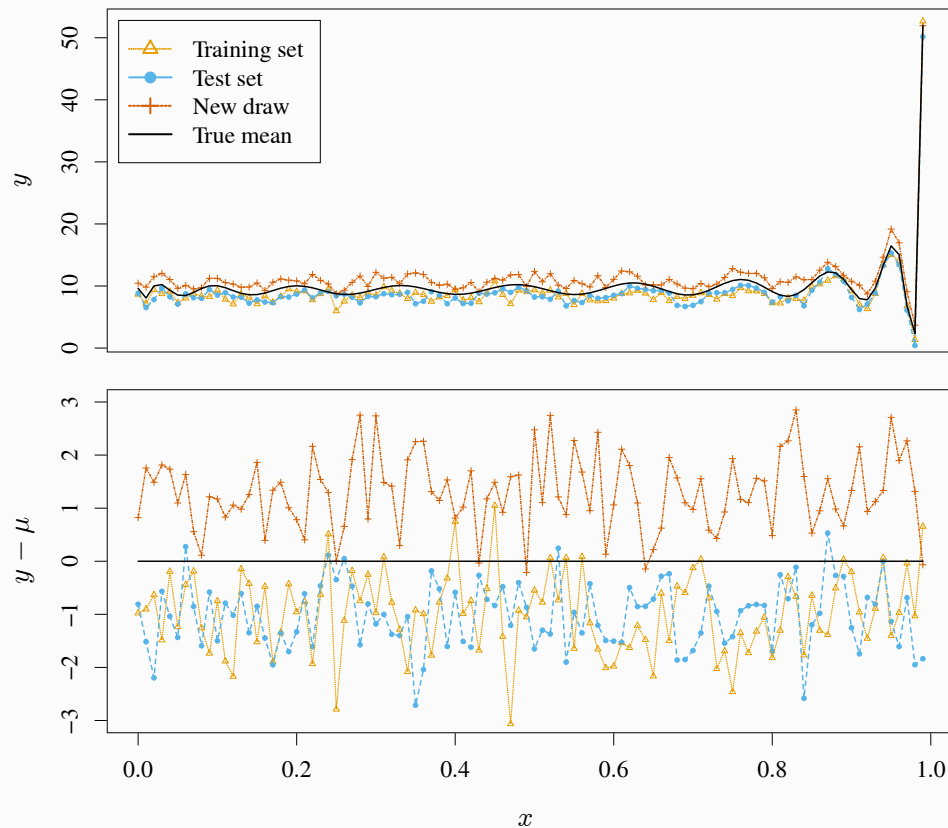
> But test set also has nonzero optimism!

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

# ➤ One draw as an example

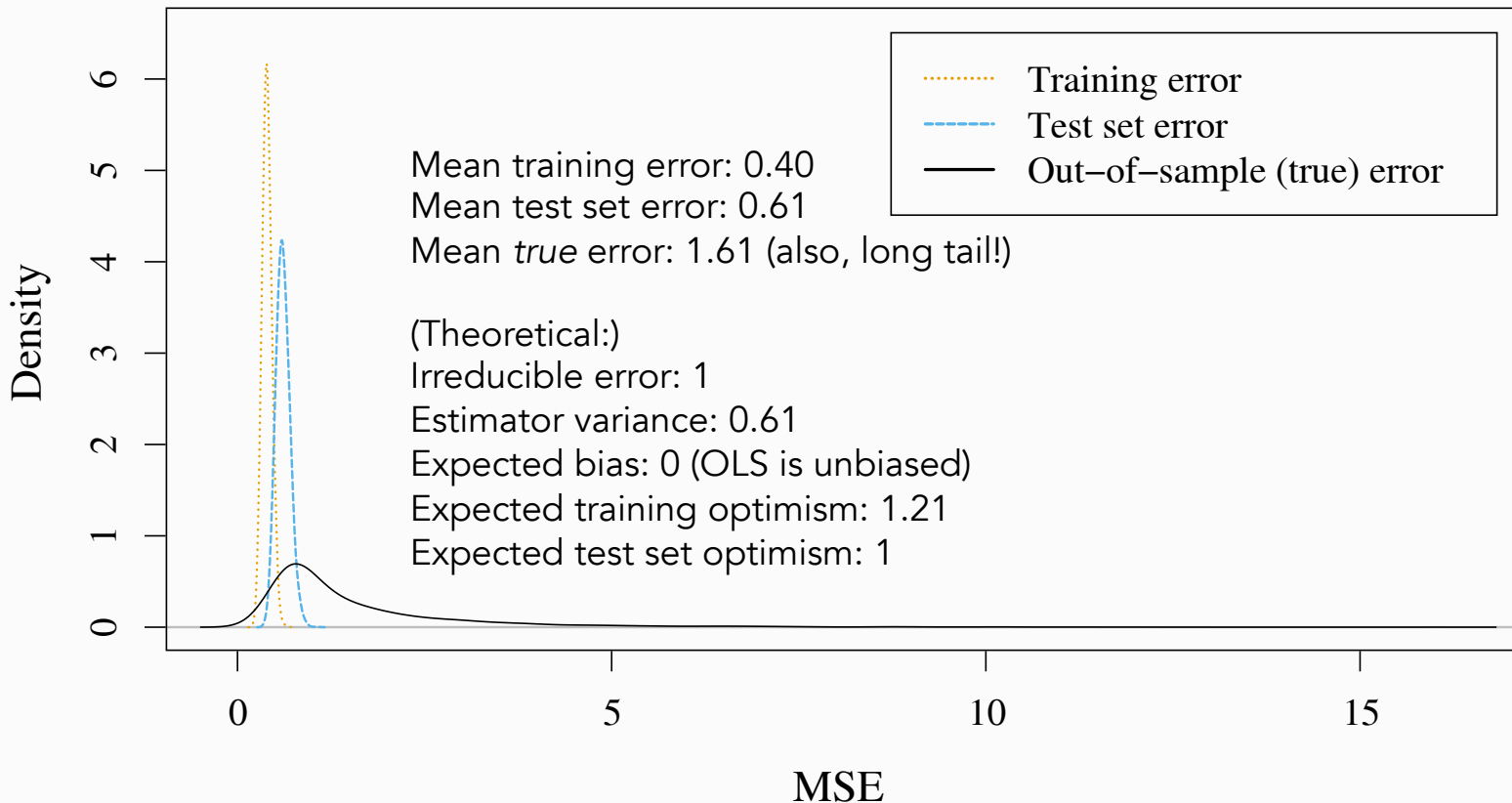
- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References

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





# Simulated MSE



- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References

# › Consequences

- › Non-experimental results are always preliminary
- › Can try to split data around covariance (Bergmeir et al., 2018; Hammerla & Plötz, 2015)
  - 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!*

## › Summary

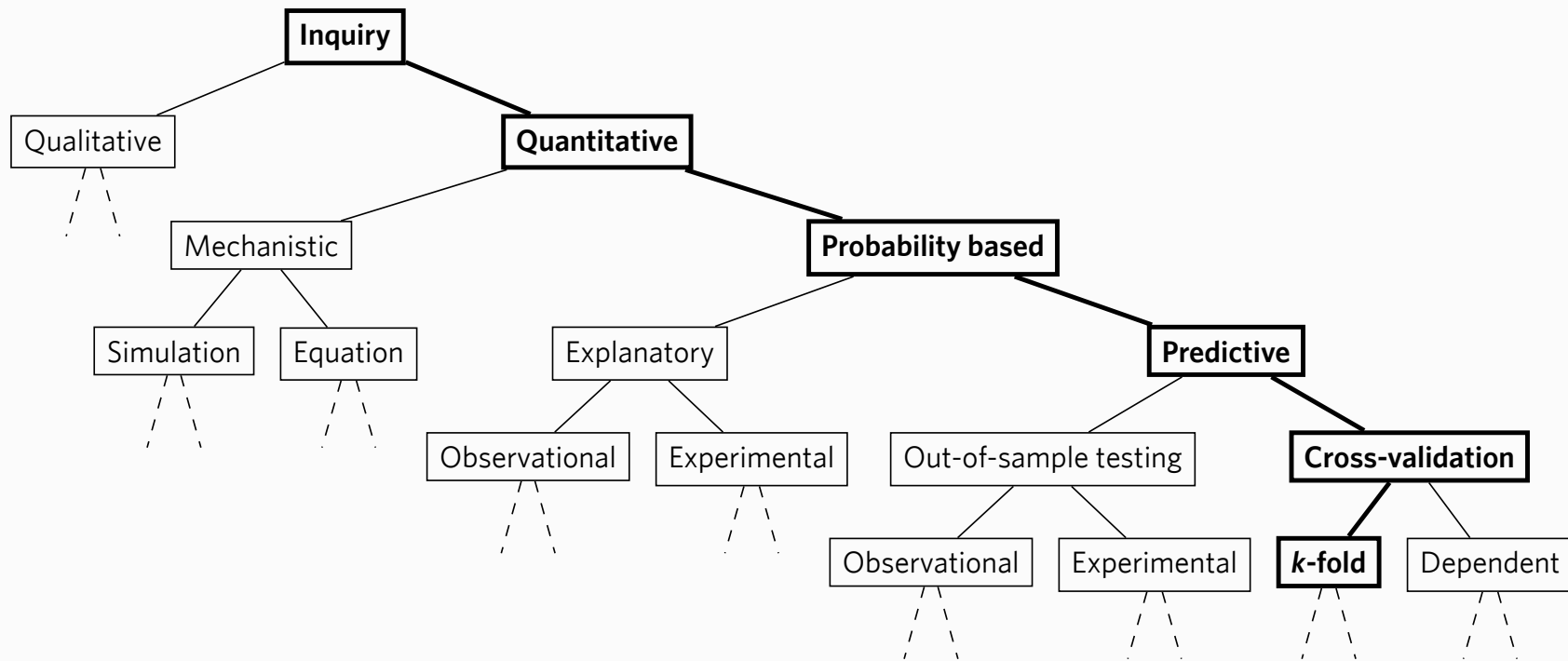
- › Introduction
  - › Quantitative: meanings, measurement, and constructs
  - › Probability-based: Central tendency, variability
  - › Predictive: Correlation vs. causation
  - › Cross-validation: Dependencies and optimism
  - › Summary
  - › References
- › Quantification sacrifices multiplicity and depth of meaning, and is at the mercy of measurement processes that only imperfectly capture constructs
  - › Probability-based modeling requires multiple observations, and uses central tendencies which exclude outliers
  - › “Prediction” is based on correlation, which can sidestep responsibility, and are fragile to causation
  - › Cross-validation can fail if there are dependencies, or other problems



# ➤ Thank you!

# ➤ <momin.malik@gmail.com>

- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- References



# ➤ 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.
- Bouk, Dan. *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual*. University of Chicago Press, 2015.
- 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>.

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).

Porter, Theodore M. "Thin Description: Surface and Depth in Science and Science Studies." *Osiris* 27, no. 1 (2012). <https://dx.doi.org/10.1086/667828>.

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>.

# ➤ References (7/8)

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>.

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
- › Quantitative: meanings, measurement, and constructs
- › Probability-based: Central tendency, variability
- › Predictive: Correlation vs. causation
- › Cross-validation: Dependencies and optimism
- › Summary
- › References
- › Backup slides

# › 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

➤ Introduction

➤ Quantitative:  
meanings,  
measurement,  
and constructs

➤ Probability-  
based: Central  
tendency,  
variability

➤ Predictive:  
Correlation vs.  
causation

➤ Cross-  
validation:  
Dependencies  
and optimism

➤ Summary

➤ References

➤ Backup slides

# ➤ 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

➤ Introduction

➤ Quantitative:  
meanings,  
measurement,  
and constructs

➤ Probability-  
based: Central  
tendency,  
variability

➤ Predictive:  
Correlation vs.  
causation

➤ Cross-  
validation:  
Dependencies  
and optimism

➤ Summary

➤ References

➤ Backup slides



# ➤ But language not enough

- Introduction
- Quantitative: meanings, measurement, and constructs
- Probability-based: Central tendency, variability
- Predictive: Correlation vs. causation
- Cross-validation: Dependencies and optimism
- Summary
- 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*

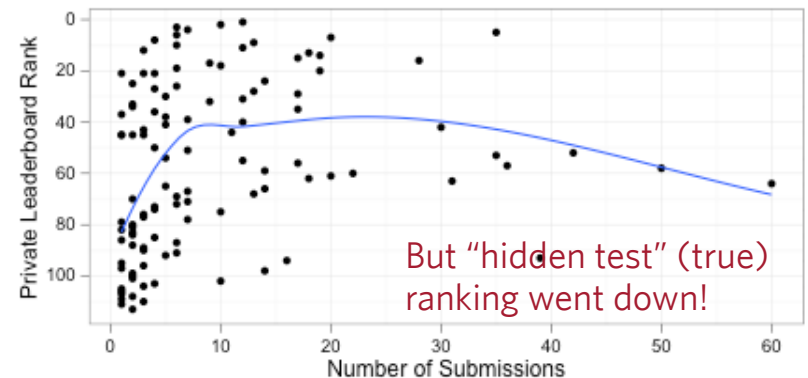
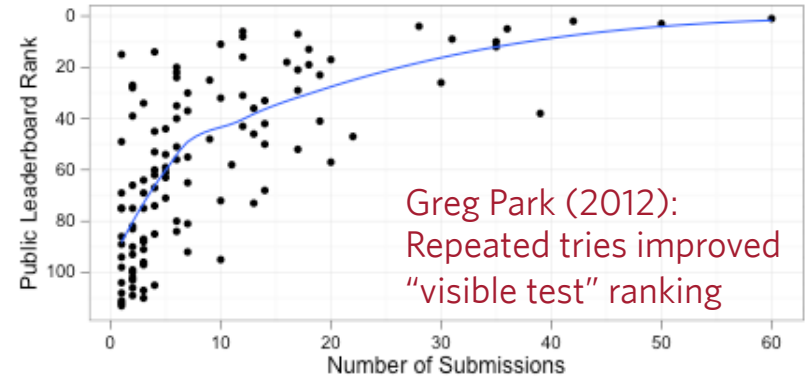
(I.e., using “backtest” in place of “predict” has not prevented financial analysts from unwitting overfitting)

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

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

- > Introduction
- > Quantitative: meanings, measurement, and constructs
- > Probability-based: Central tendency, variability
- > Predictive: Correlation vs. causation
- > Cross-validation: Dependencies and optimism
- > Summary
- > References
- > Backup slides

# › Critical technical practice (1)

- › Introduction
  - › Quantitative: meanings, measurement, and constructs
  - › Probability-based: Central tendency, variability
  - › Predictive: Correlation vs. causation
  - › Cross-validation: Dependencies and optimism
  - › Summary
  - › References
  - › Backup slides
- › 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
  - › Quantitative: meanings, measurement, and constructs
  - › Probability-based: Central tendency, variability
  - › Predictive: Correlation vs. causation
  - › Cross-validation: Dependencies and optimism
  - › Summary
  - › References
  - › Backup slides
- › “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)

- > “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.**”