A hierarchy of limitations in machine learning: Data biases and the social sciences

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

Universitat Oberta de Catalunya, Faculty of Psychology and Education
Webinar Series: Data Cultures in Higher Education
29 September 2020

Outline

- Types of inquiry
- Quantification and measurement
- Prediction vs. explanation
- Using correlations
- Model performance
- The future of machine learning in social research

Types of inquiry: My background

DEPARTMENT OF THE HISTORY OF SCIENCE HARVARD UNIVERSITY

Berkman
The Berkman Center for Internet & Society at Harvard University

OXFORD INTERNET INSTITUTE
UNIVERSITY OF OXFORD

Carnegie Mellon University
School of Computer Science

Data Science For Social Good
Summer Fellowship

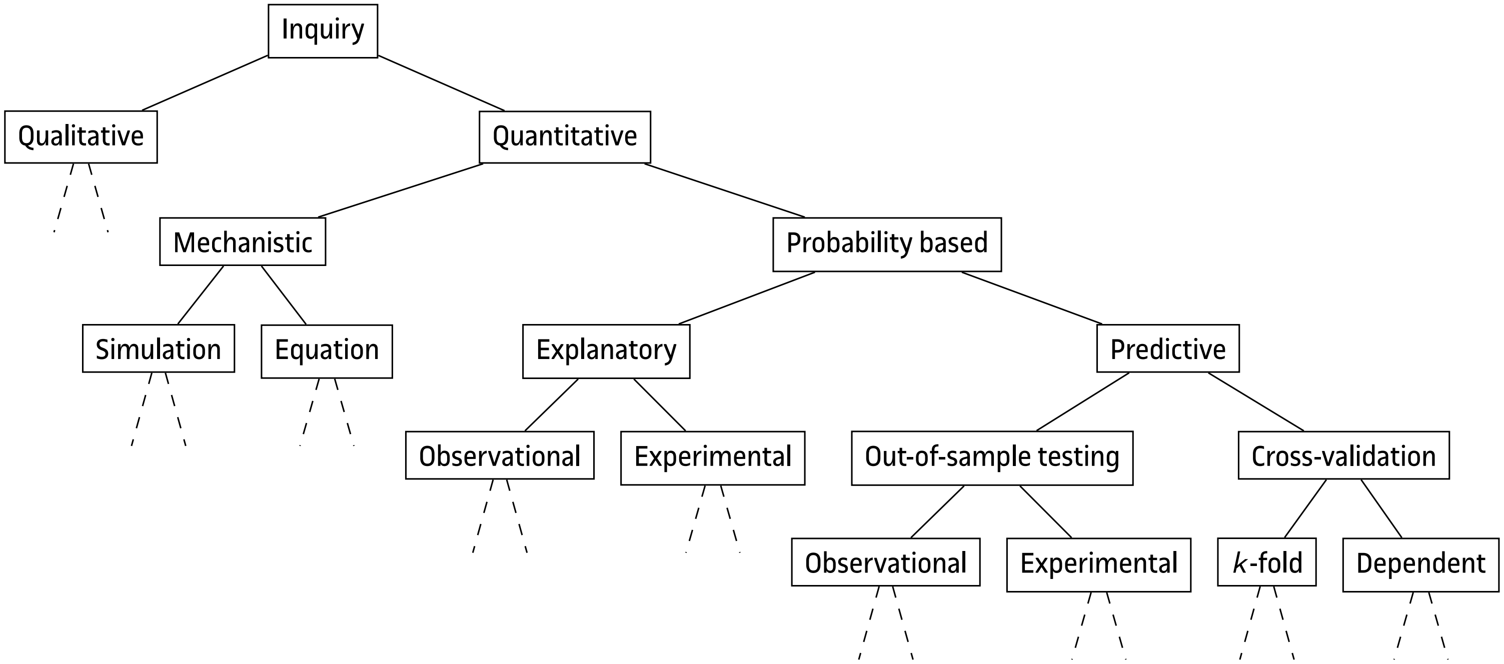
BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

MOVE ON

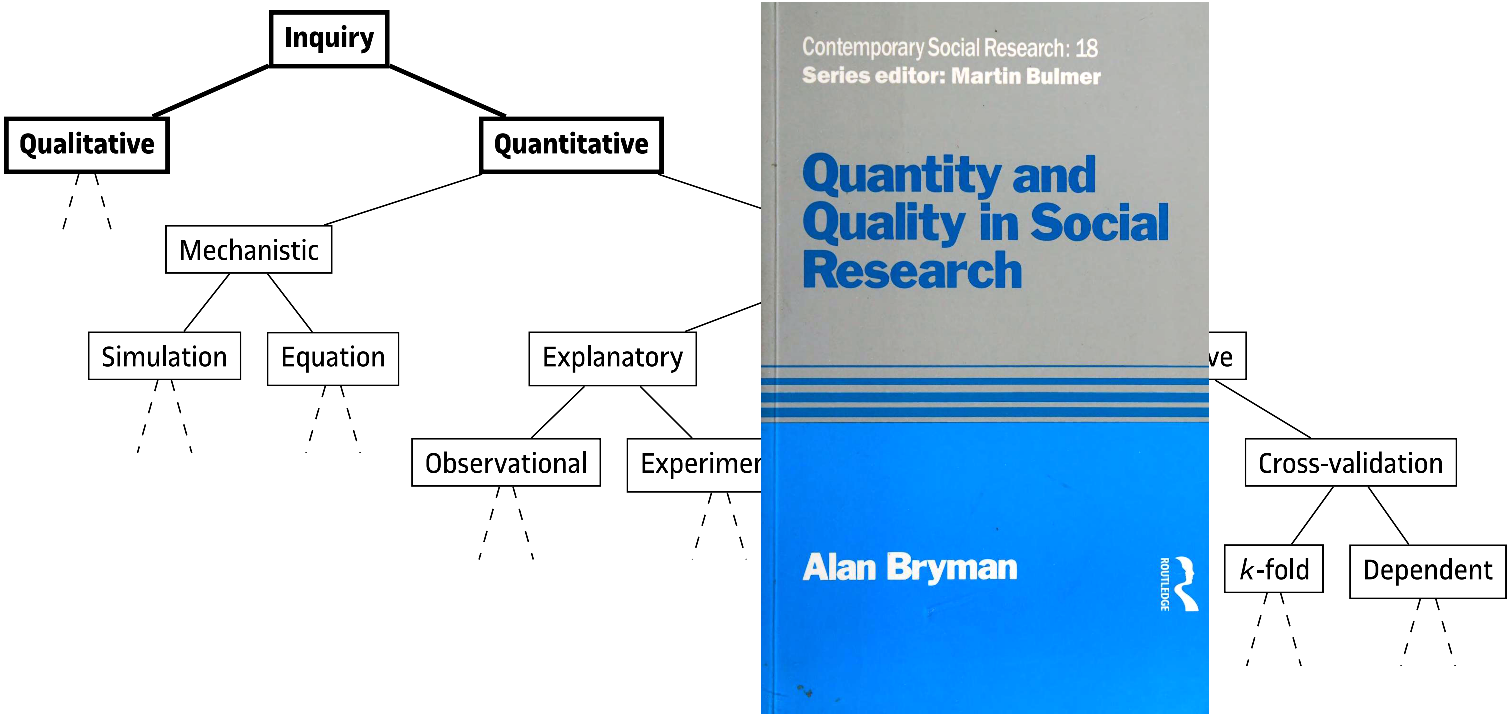
Does modeling really work? Is it better than other approaches?

Sort of. Sometimes. Maybe. If it’s done right and we’re lucky. (And that’s when we can even tell.)

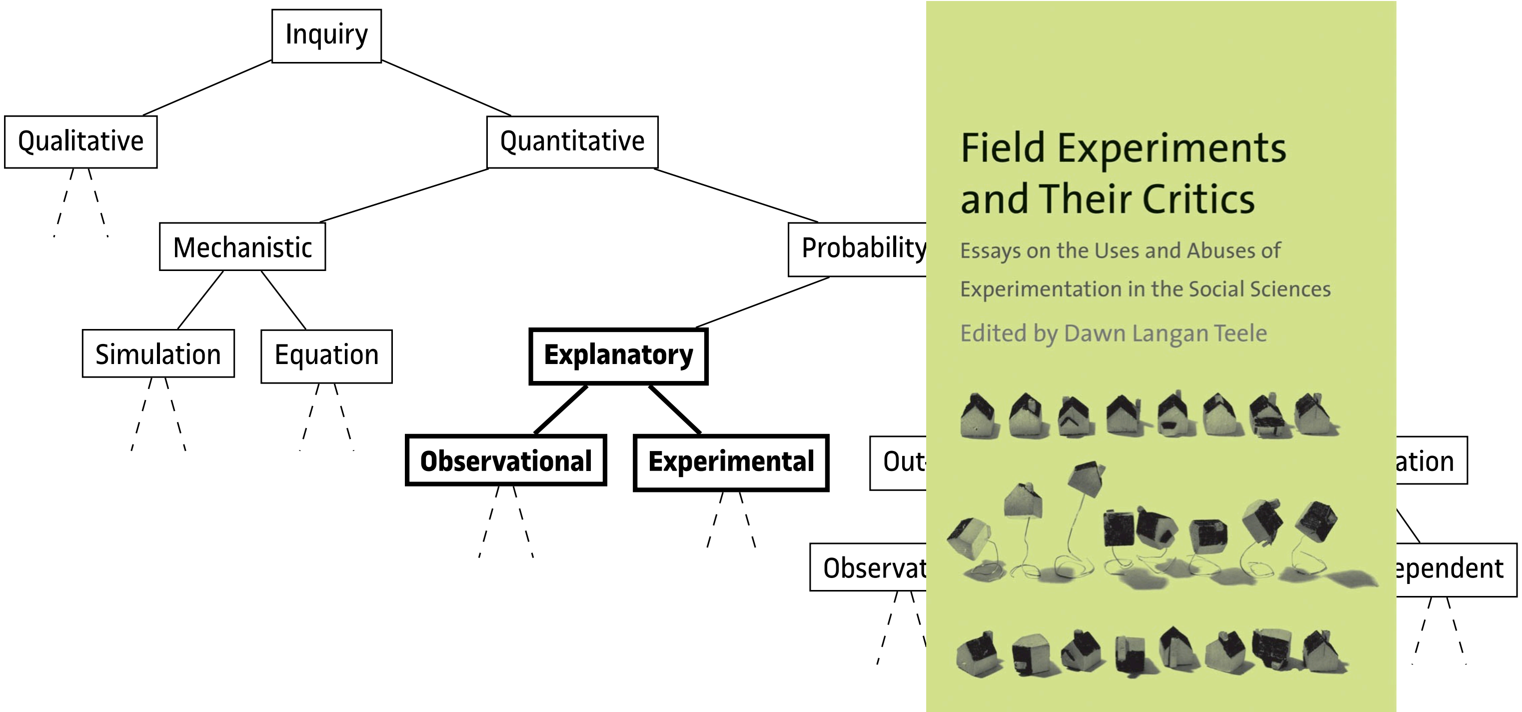
Types of inquiry: Methodological trade-offs



Types of inquiry: Methodological trade-offs



Types of inquiry: Methodological trade-offs



Types of inquiry: Methodological trade-offs

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

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

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

1. INTRODUCTION

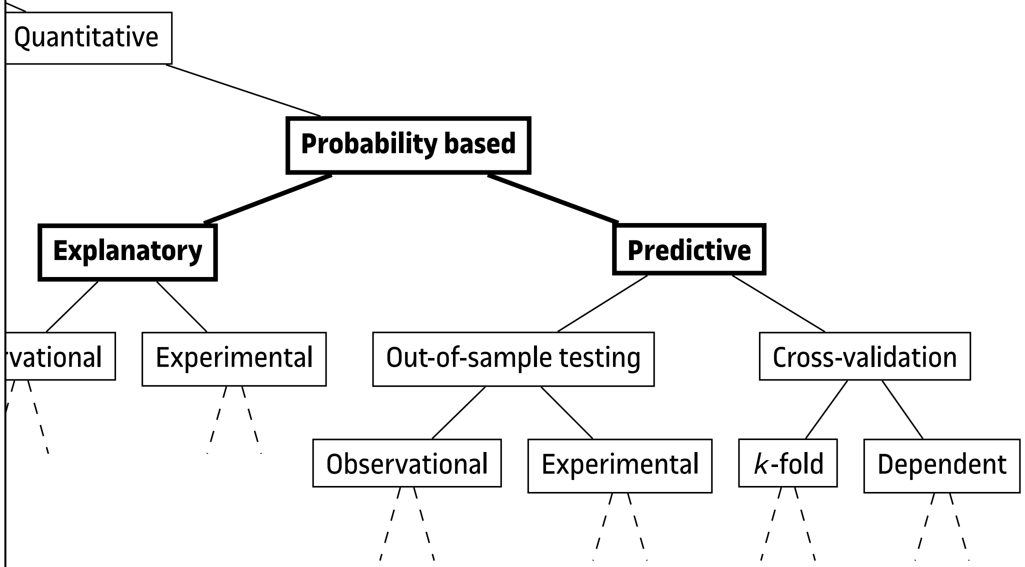
Looking at how statistical models are used in different scientific disciplines for the purpose of theory building and testing, one finds a range of perceptions regarding the relationship between causal explanation and empirical prediction. In many scientific fields such as economics, psychology, education, and environmental science, statistical models are used almost exclusively for causal explanation, and models that possess high explanatory power are often assumed to inherently possess predictive power. In fields such as natural language processing and bioinformatics, the focus is on empirical prediction with only a slight and indirect relation to causal explanation. And yet in other research fields, such as epidemiology, the emphasis on causal explanation versus empirical prediction is more mixed. Statistical modeling for description, where the purpose is to capture the data structure parsimoniously, and which is the most commonly developed within the field of statistics, is not commonly used for theory building and testing in other disciplines. Hence, in this article I focus on the use of statistical modeling for causal explanation and for prediction. My main premise is that the two are often conflated, yet the causal versus predictive distinction has a large impact on each step of the statistical modeling process and on its consequences. Although not explicitly stated in the statistics methodology literature, applied statisticians instinctively sense that predicting and explaining are different. This article aims to fill a critical void: to tackle the distinction between explanatory modeling and predictive modeling.

Clearing the current ambiguity between the two is critical not only for proper statistical modeling, but more importantly, for proper scientific usage. Both explanation and prediction are necessary for generating and testing theories, yet each plays a different role in doing so. The lack of a clear distinction within statistics has created a lack of understanding in many disciplines of the difference between building sound explanatory models versus creating powerful predictive models, as well as confusing explanatory power with predictive power. The implications of this omission and the lack of clear guidelines on how to model for explanatory versus predictive goals are considerable for both scientific research and practice and have also contributed to the gap between academia and practice.

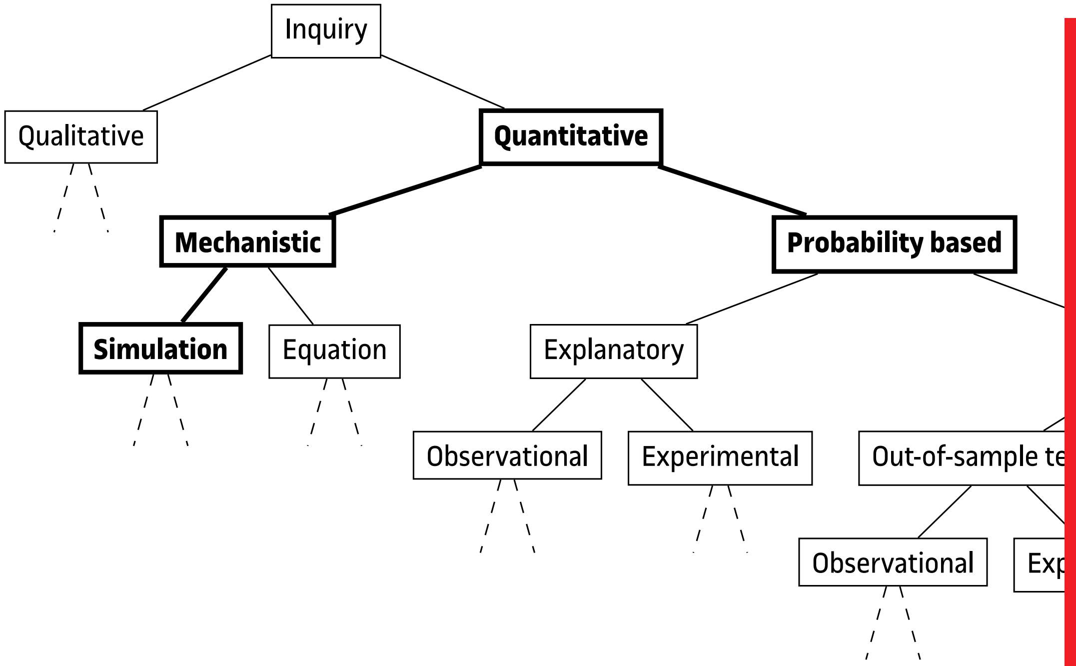
I start by defining what I term *explaining* and *predicting*. These definitions are chosen to reflect the dis-

Galit Shmueli is Associate Professor of Statistics, Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742, USA (e-mail: gshmueli@umd.edu).

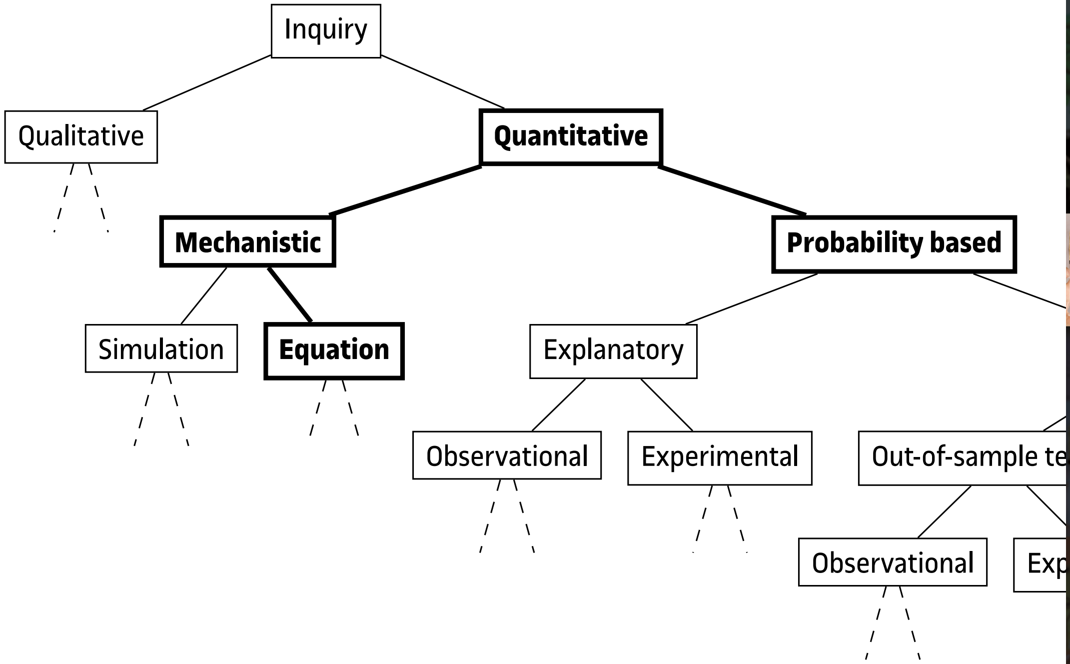
289

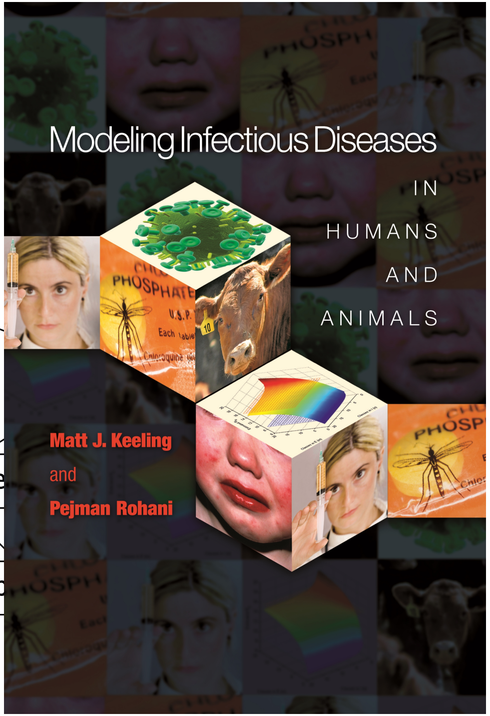


Types of inquiry: Methodological trade-offs

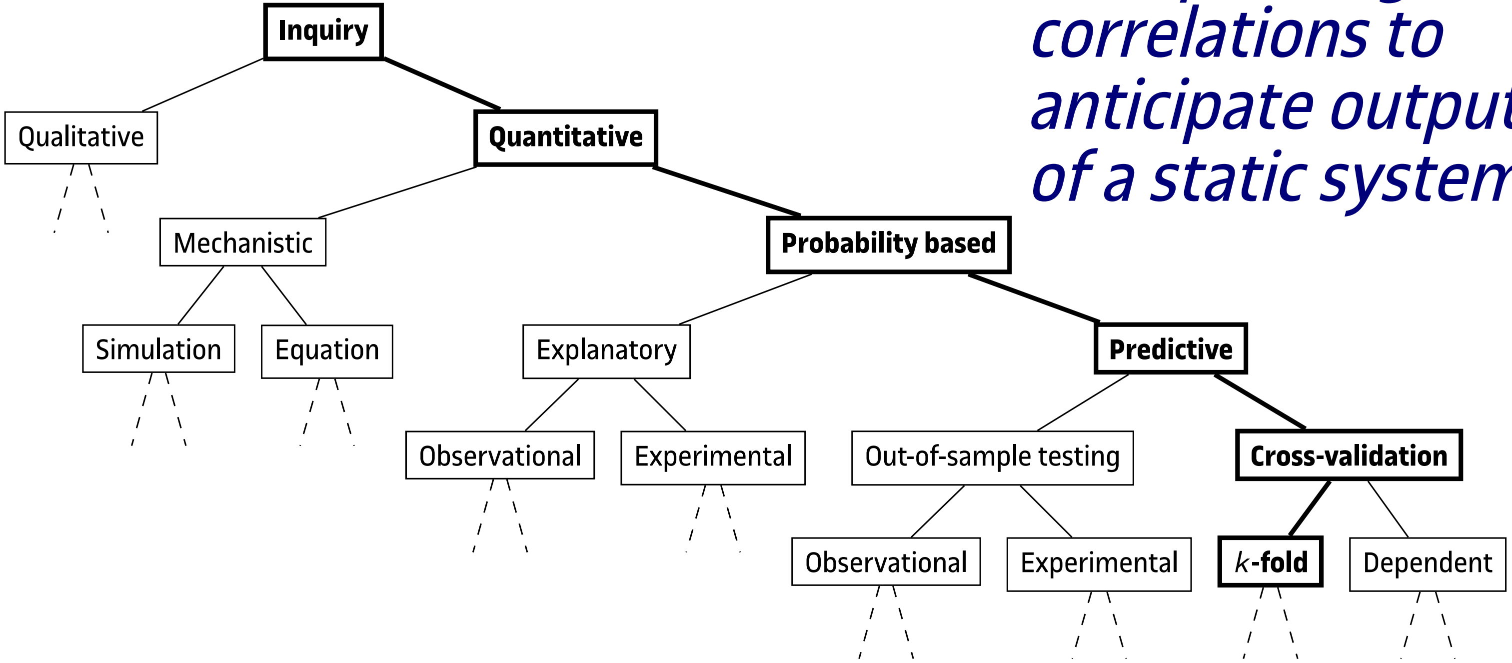


Types of inquiry: Methodological trade-offs





Mainstream machine learning



Extrapolating from correlations to anticipate outputs of a static system

Types of inquiry

Quantification and measurement

Prediction vs. explanation

Using correlations

Model performance

The future

Responsibility for quantification

- Quantification "thins out" meanings (Porter, 2012), solidifying only one set of meanings over all others
- Nothing subsequent can undo this, or transcend it
- Conflating what is *available* with what is *desired* will miss the problems of proxies (e.g., Goodhart's/Campell's Law)
  - Healthcare costs are a poor proxy for 'health' (Obermeyer et al., 2019)
  - Grades are a poor proxy for 'learning'
  - Citations are a poor proxy for 'impact'
  - Both arrests and convictions are poor proxies for 'crime'

Example: Harrisburg study (Withdrawn)

HARRISBURG UNIVERSITY
OF SCIENCE AND TECHNOLOGY

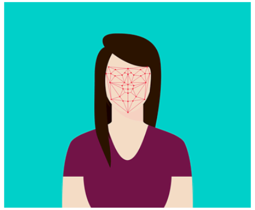
About HU | Admissions | Degrees & Programs | The Campus | Esports | News & Events | Give to HU | Alumni

HU facial recognition software predicts criminality

HOME \ UNCATEGORIZED \

HU facial recognition software predicts criminality

A group of Harrisburg University professors and a Ph.D. student have developed automated computer facial recognition software capable of predicting whether someone is likely going to be a criminal.



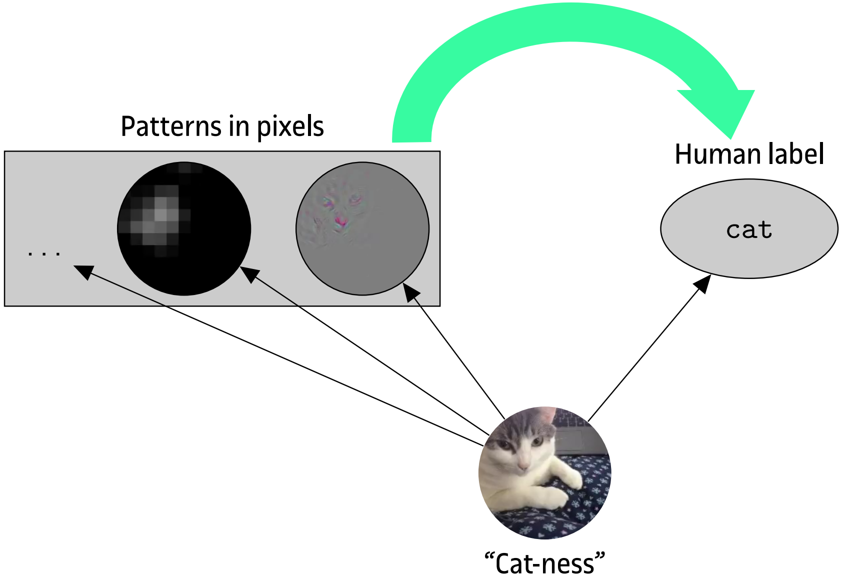
With 80 percent accuracy and with no racial bias, the software can predict if someone is a criminal based solely on a picture of their face. The software is intended to help law enforcement prevent crime.

Ph.D. student and NYPD veteran Jonathan W. Korn, Prof. Nathaniel J.S. Ashby, and Prof. Roozbeh Sadeghian titled their research "A Deep Neural Network Model to Predict Criminality Using Image Processing."

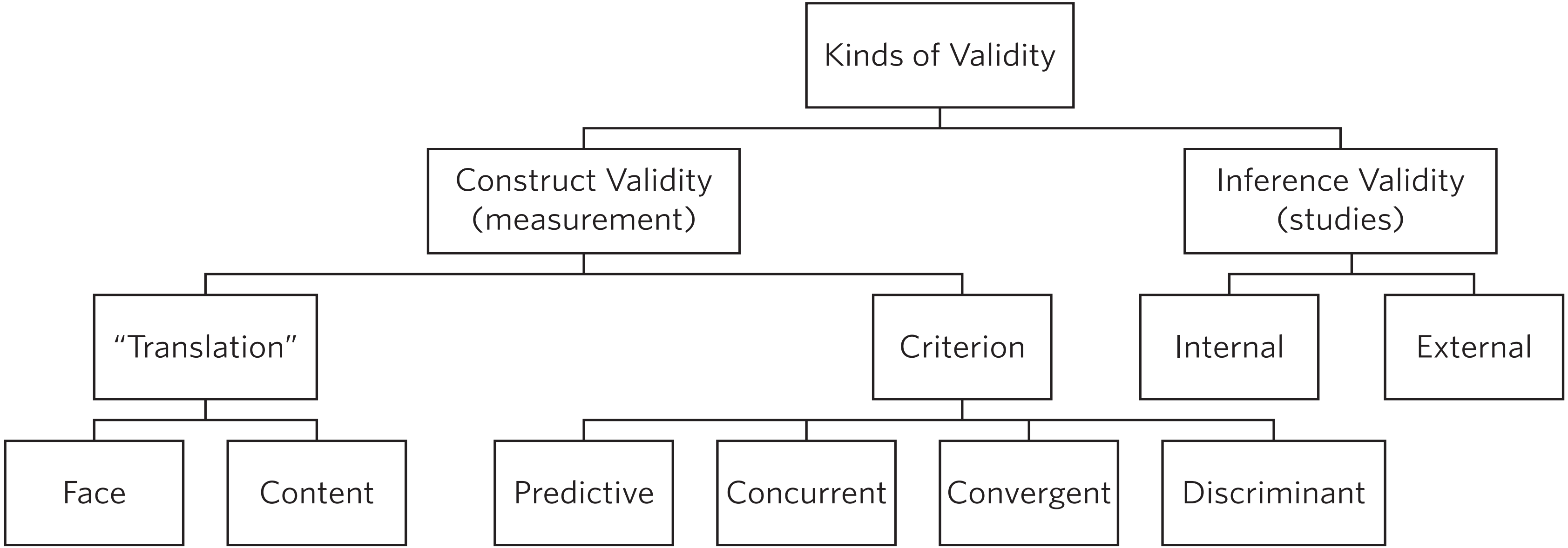
"We already know machine learning techniques can outperform humans on a variety of tasks

- "Criminality" is imposed, not inherent
- Even given a criminal code, we have no crime statistics; we have arrests and convictions
- Their claims were between implausible and categorically impossible (Coalition for Critical Tech, 2020)

Machine learning only matches (central tendency of) labels, not meanings

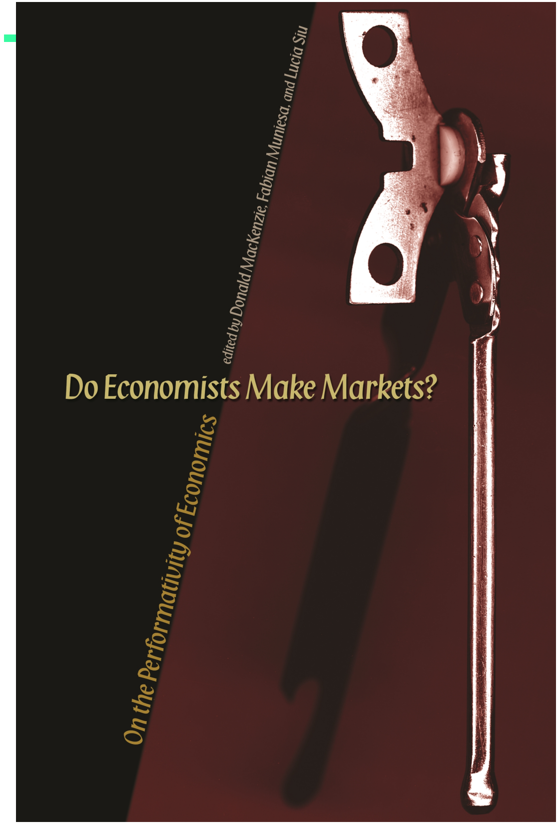


Validating measurements



Adapted from Borgatti, 2012

Performativity: Models making themselves true



"the *performativity thesis* is that economics produces a body of formal models and transportable techniques that, when carried out into the world by its professionals and popularizers, **reformats and reorganizes the phenomena the models purport to describe…**" (Healy, 2015)

“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

Types of inquiry

Quantification and measurement

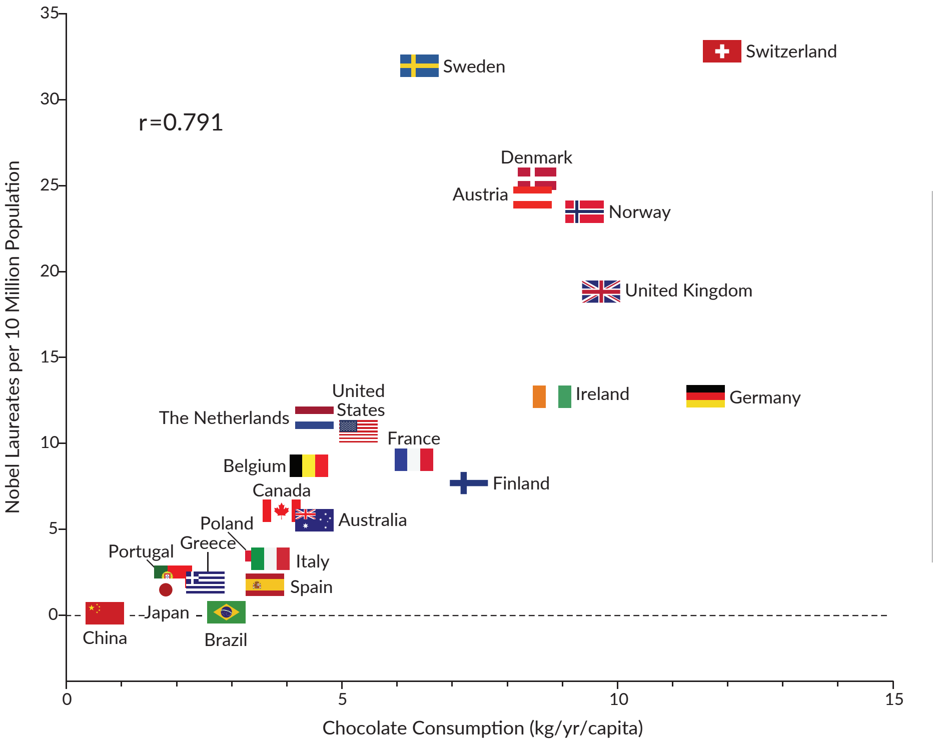
Prediction vs. explanation

Using correlations

Model performance

The future

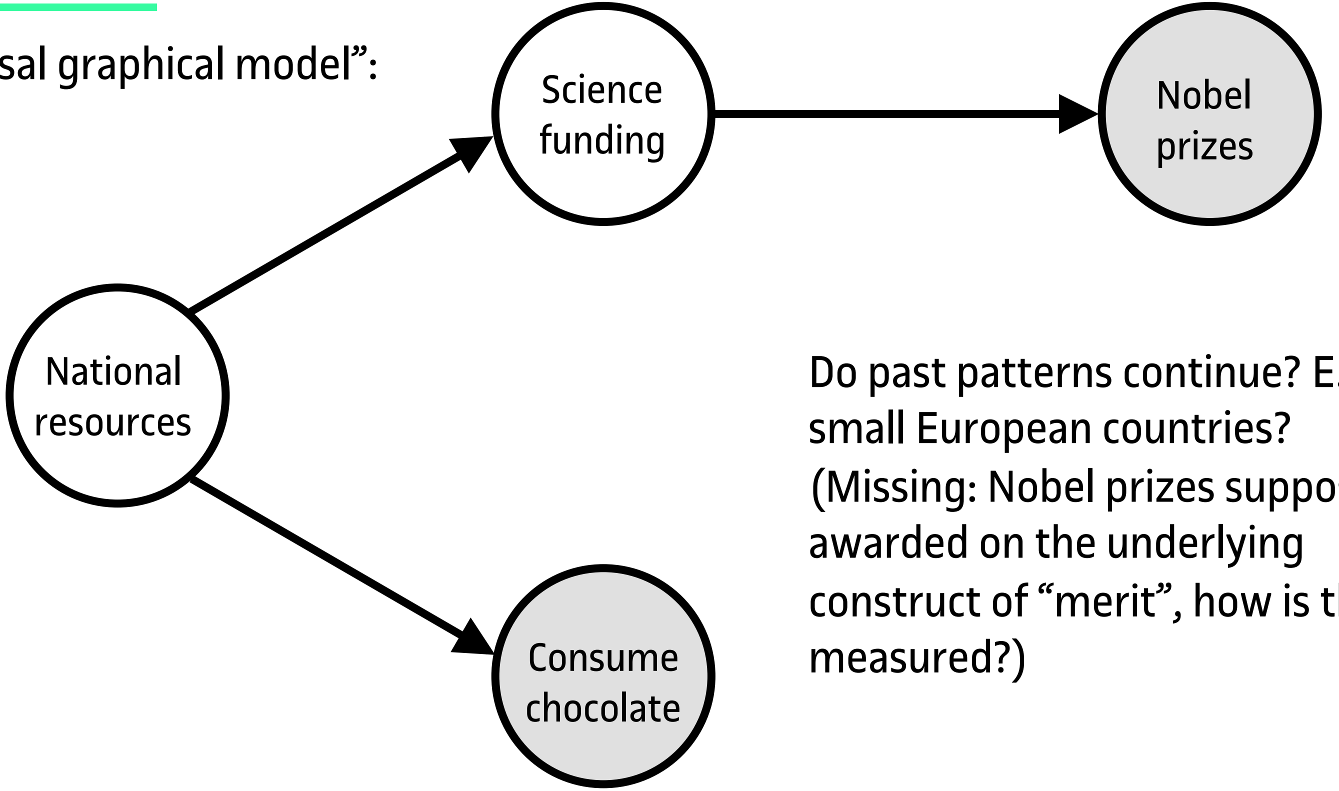
“Prediction” is correlation



Messerli, 2012, *NEJM*

Prediction (correlation) is not explanation (causation)

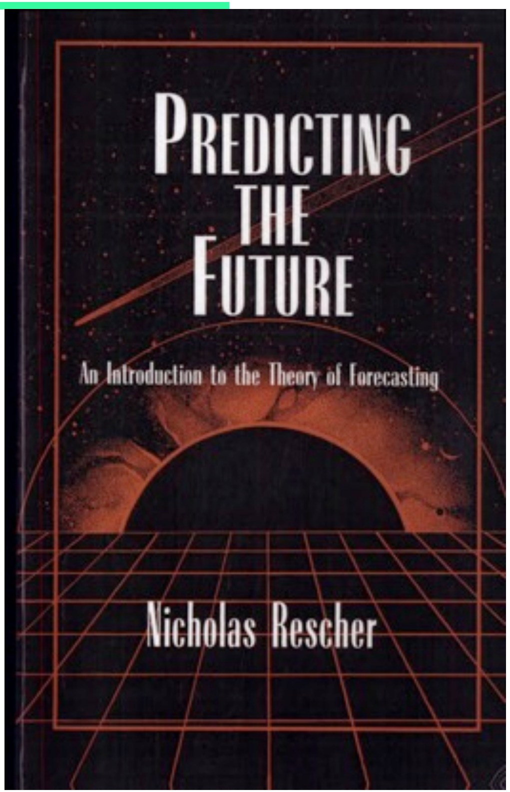
A “causal graphical model”:



Do past patterns continue? E.g., small European countries? (Missing: Nobel prizes supposedly awarded on the underlying construct of “merit”, how is that measured?)

Not obvious usage of “predict”

Not obvious usage of “predict”

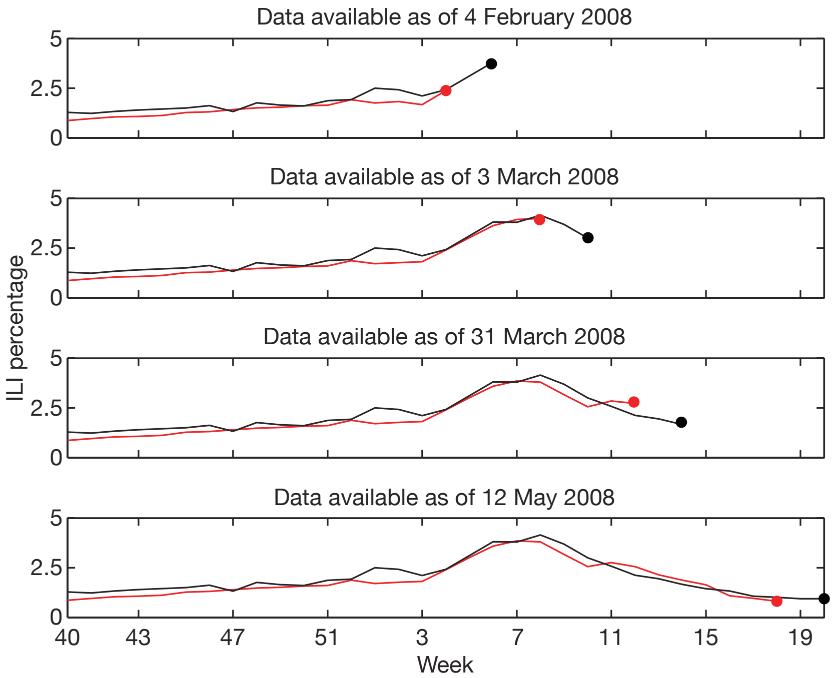


88 ■ PREDICTING THE FUTURE

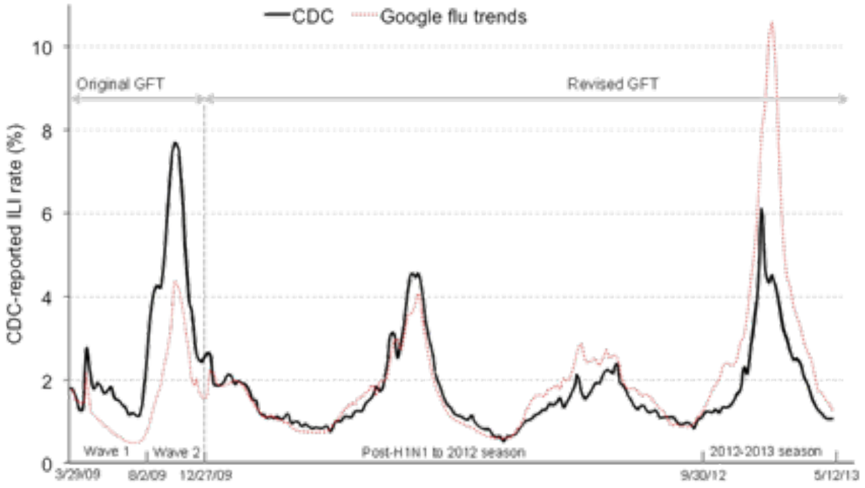
TABLE 6.1: A SURVEY OF PREDICTIVE APPROACHES

Predictive Approaches	Linking Mechanism	Methodology Of Linkage
UNFORMALIZED/JUDGMENTAL		
judgmental estimation	expert informants	informed judgment
FORMALIZED/INFERENTIAL		
RUDIMENTARY (ELEMENTARY)		
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
SCIENTIFIC (SOPHISTICATED)		
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

Extrapolation can fail



Ginsberg et al., 2012, *Nature*



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

Why stick with correlations? Lucrative

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

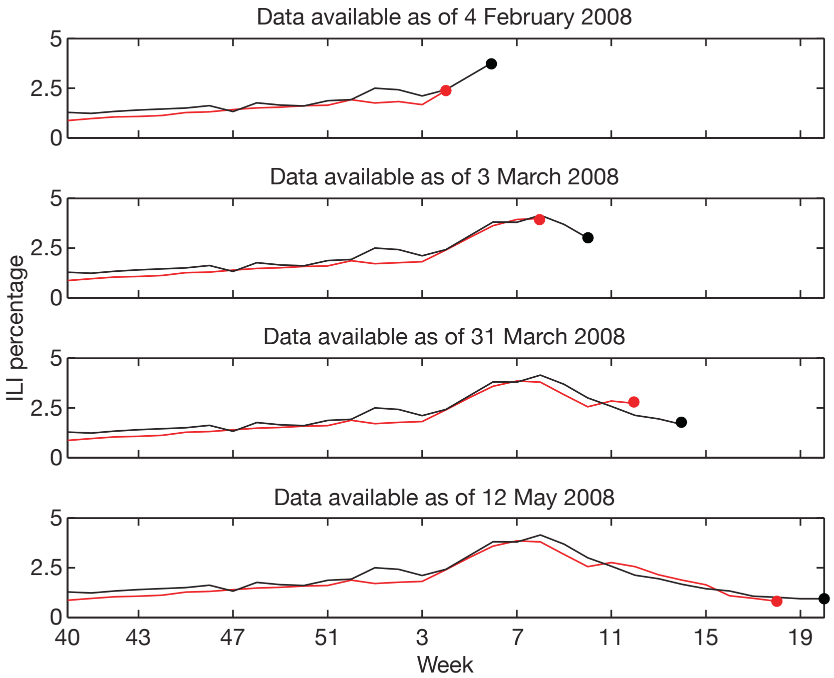
A lawyer opposing the bill "cited statistics from around the nation showing shorter life spans for blacks, including 1870 census figures showing a 17.28 death rate for 'colored people' against 14.74 for whites. These numbers, Williams argued, and not any 'discrimination on the ground of color' motivated 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.'"

But lucrative at the cost of equity

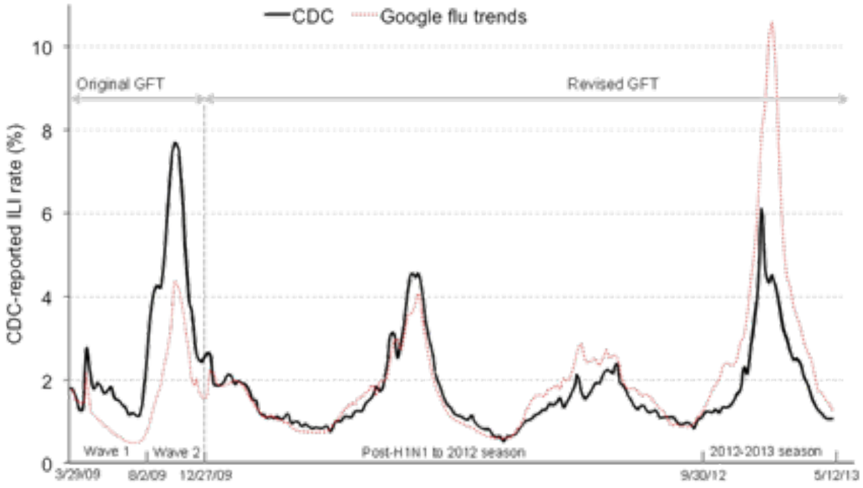
"Chappelle's allies noted that Williams's statistics, while bleak enough, answered the wrong question. The question was not whether blacks in slavery or adjusting to freedom were poor insurance risks, or even whether southern blacks were poor risks. The question was African Americans' potential for equality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

UOC

Model performance (Google Flu Trends)

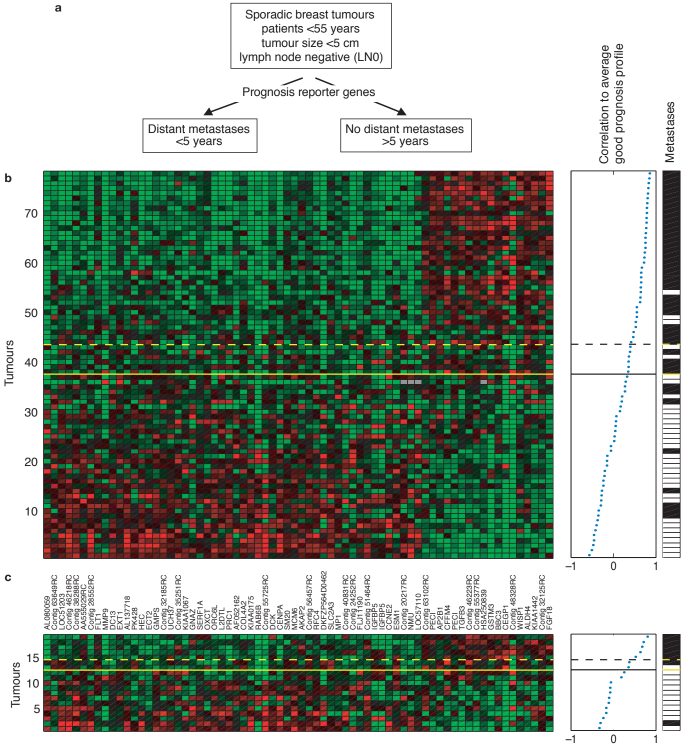


Ginsberg et al., 2012, *Nature*



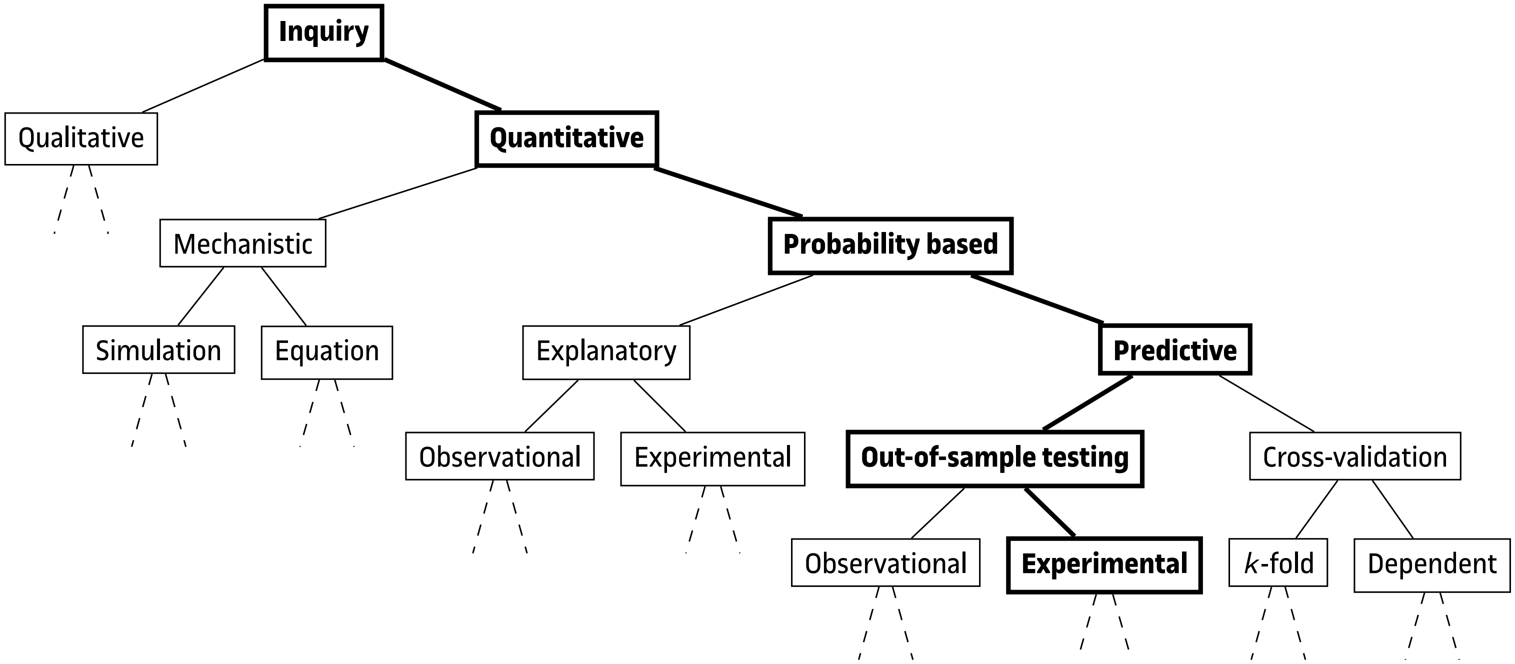
Santillana et al., 2014, *Am. J. Prev. Med.*

Real-world testing of "predictions"

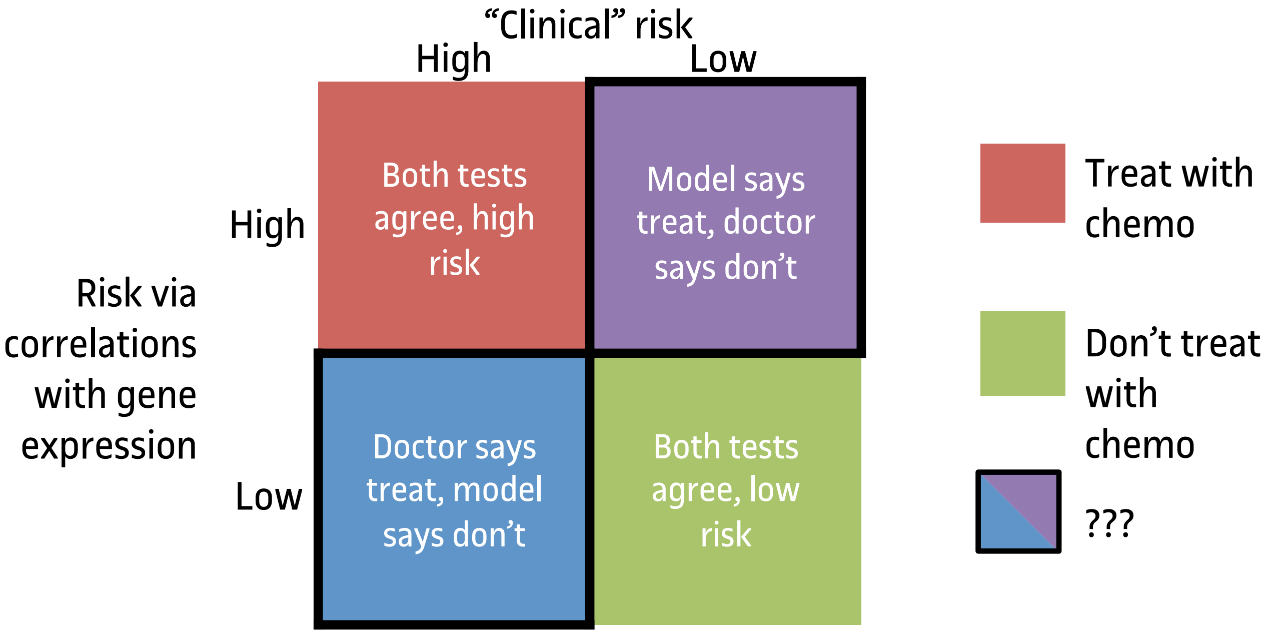


- van't Veer et al. (2002) found 70 genes correlated with developing breast cancer
- Of course the correlations were optimal, post-hoc. But did it generalize?

Real-world testing of “predictions”

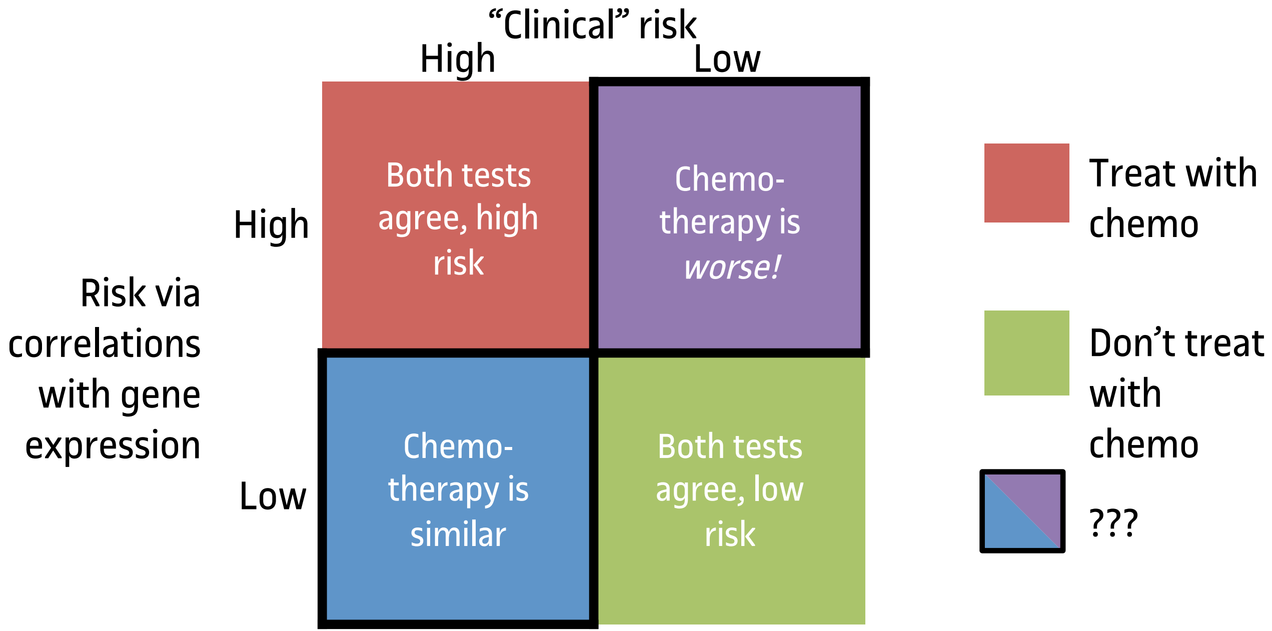


Real-world testing of “predictions”



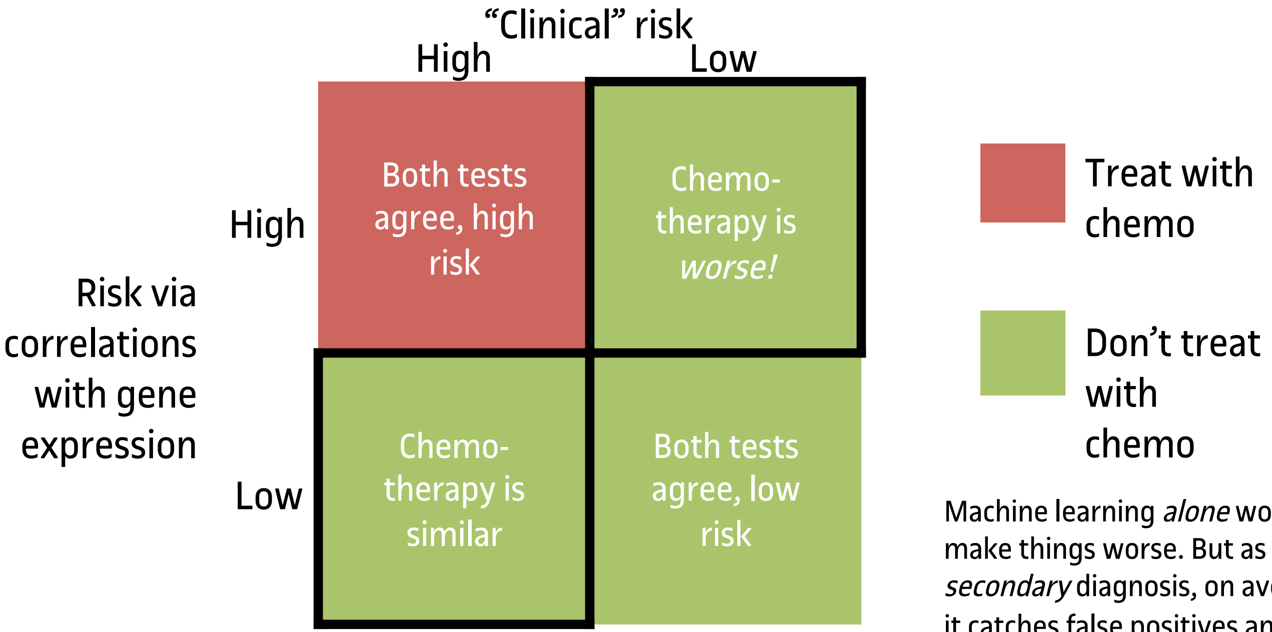
Cardoso et al., 2016, *NEJM*

Real-world testing of "predictions"



Cardoso et al., 2016, *NEJM*

Real-world testing of "predictions"



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*

The future

- Real-world, holistic testing before accepting claims
  - How much does it cost to build and maintain a “predictive” system? What if that was spent elsewhere?
- *Qualitative* assessments of “predictive” systems
- For labeling, use qualitative best practices (develop a codebook, recognize which set of meanings we are committing to)
- Rejecting new (and existing) governance via correlations