Machine learning in the hierarchy of methodological limitations

Momin M. Malik

Tuesday, 21 September 2021

TILT Seminar Series, Tilburg University

Tilburg Institute for Law, Technology, and Society

Overview

- Introduction: The hierarchy
- Consequences of quantification
- The problem with “prediction”
- Machine learning logic and the law
- Case study: “e-discovery”

Introduction

Pre-print, "A hierarchy of limitations"

- Massive review paper
- Meant to be a one-stop shop about ML, and indeed quantitative methodologies
- Key message: machine learning *does not, will not, and cannot overcome the limitations of quantification*
  - Indeed, it inherits them all

arXiv:2002.05193v2 [cs.CY] 29 Feb 2020

A Hierarchy of Limitations in Machine Learning

Momin M. Malik
Berkman Klein Center for Internet & Society at Harvard University
momin_malik@cyber.harvard.edu

29 February 2020*

Abstract

"All models are wrong, but some are useful," wrote George E. P. Box (1979). Machine learning has focused on the *usefulness* of probability models for prediction in social systems, but is only now coming to grips with the ways in which these models are *wrong*—and the consequences of those shortcomings. This paper attempts a comprehensive, structured overview of the specific conceptual, procedural, and statistical limitations of models in machine learning when applied to society. Machine learning modelers themselves can use the described hierarchy to identify possible failure points and think through how to address them, and consumers of machine learning models can know what to question when confronted with the decision about if, where, and how to apply machine learning. The limitations go from commitments inherent in quantification itself, through to showing how unmodeled dependencies can lead to cross-validation being overly optimistic as a way of assessing model performance.

Introduction

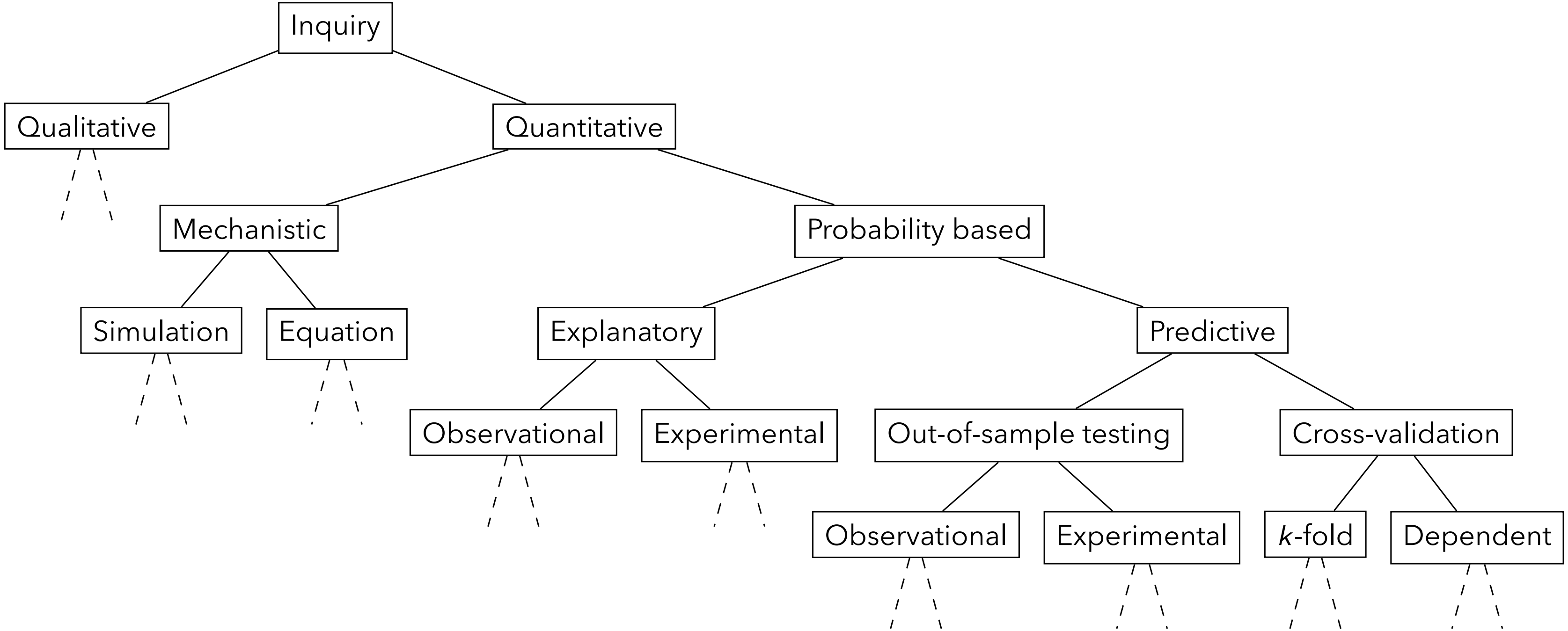
There is little argument about whether or not machine learning models are *useful* for applying to social systems. But if we take seriously George Box's dictum, or indeed the even older one that "the map is not the territory' (Korzybski, 1933), then there has been comparatively less systematic attention paid within the field to how machine learning models are *wrong* (Selbst et al., 2019) and seeing possible harms in that light. By "wrong" I do not mean in terms of making misclassifications, or even fitting over the 'wrong' class of functions, but more fundamental mathematical/statistical assumptions, philosophical (in the sense used by Abbott, 1988) commitments about how we represent the world, and sociological processes of how models interact with target phenomena.

This paper takes a particular model of machine learning research or application: one that its creators and deployers think provides a reliable way of interacting with the social world (whether that is through understanding, or in making predictions) without any intent to cause harm (McQuillan, 2018) and, in fact, a desire to not cause harm and instead improve the world,[1] for example as most explicitly in the various "{Data [Science], Machine Learning, Artificial Intelligence} for [Social] Good" initiatives, and more widely in framings around "fairness" or "ethics." I focus on the almost entirely statistical modern version of machine learning, rather than eclipsed older visions (see section 3). While many of the limitations I discuss apply to the use of machine learning in any domain, I focus on applications to the social world in order to explore the domain where limitations are strongest and stickiest. I consider limitations in machine learning such that, contrary to the expectations

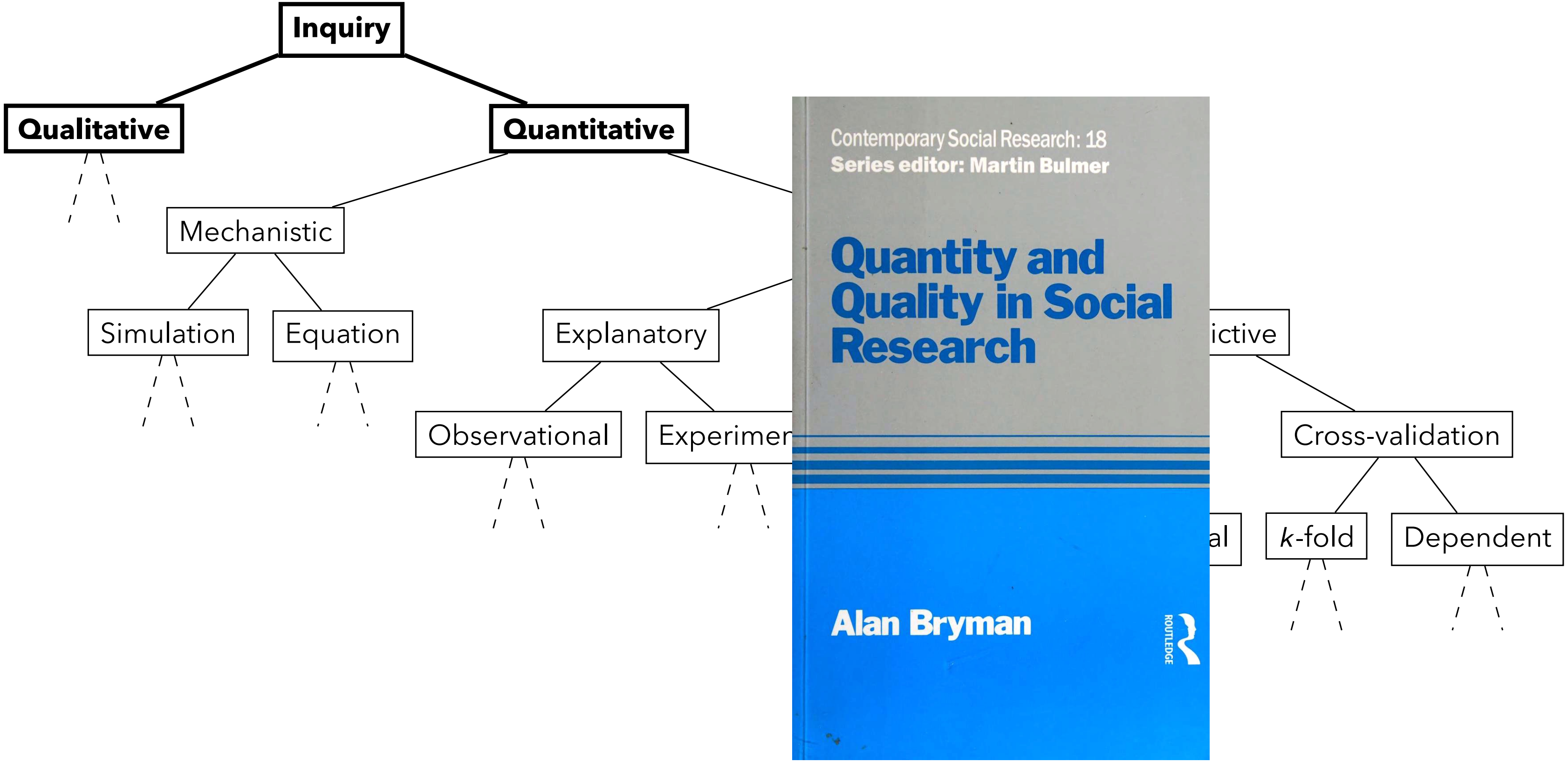
*Draft version 0.3.06. In submission. Please cite with link https://arxiv.org/abs/2002.05193.
[1] I thank John Basl for encouraging me to make clear that I consider both methodological and ethical limitations.

1

The hierarchy



Qualitative vs. quantitative



Introduction

Consequen-ces of quantificat-ion

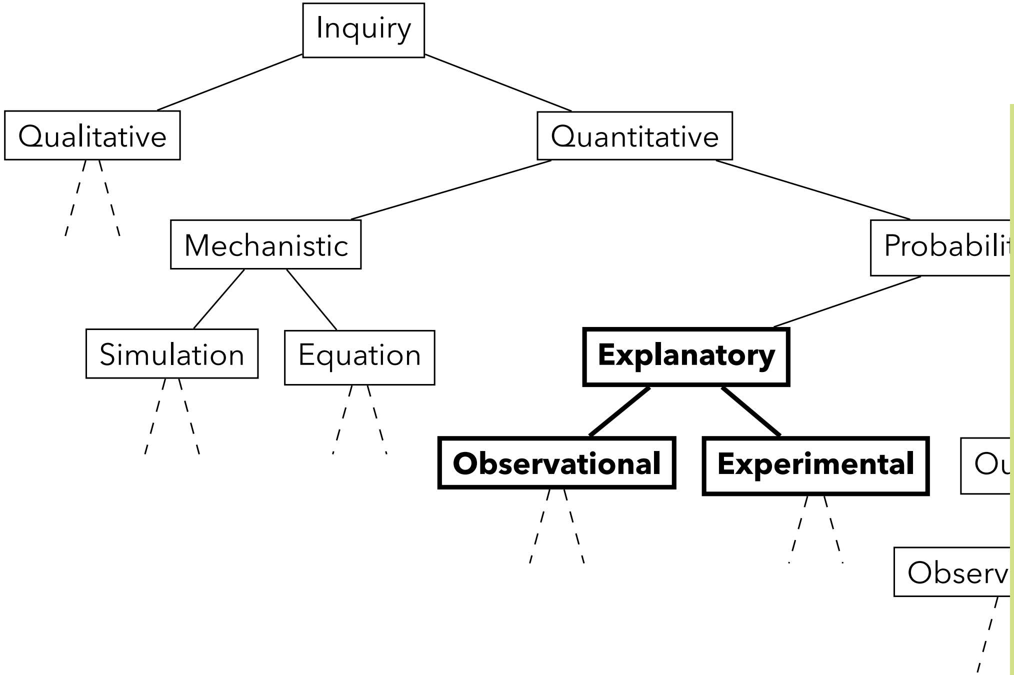
The problem with "prediction"

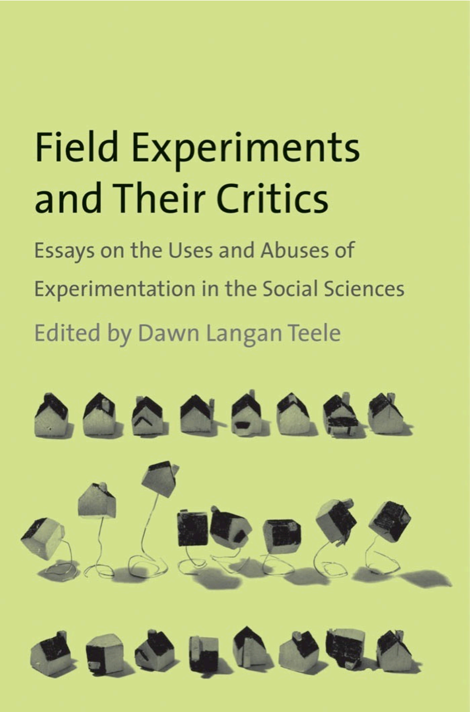
Machine learning logic and the law

Case study: "e-discovery"

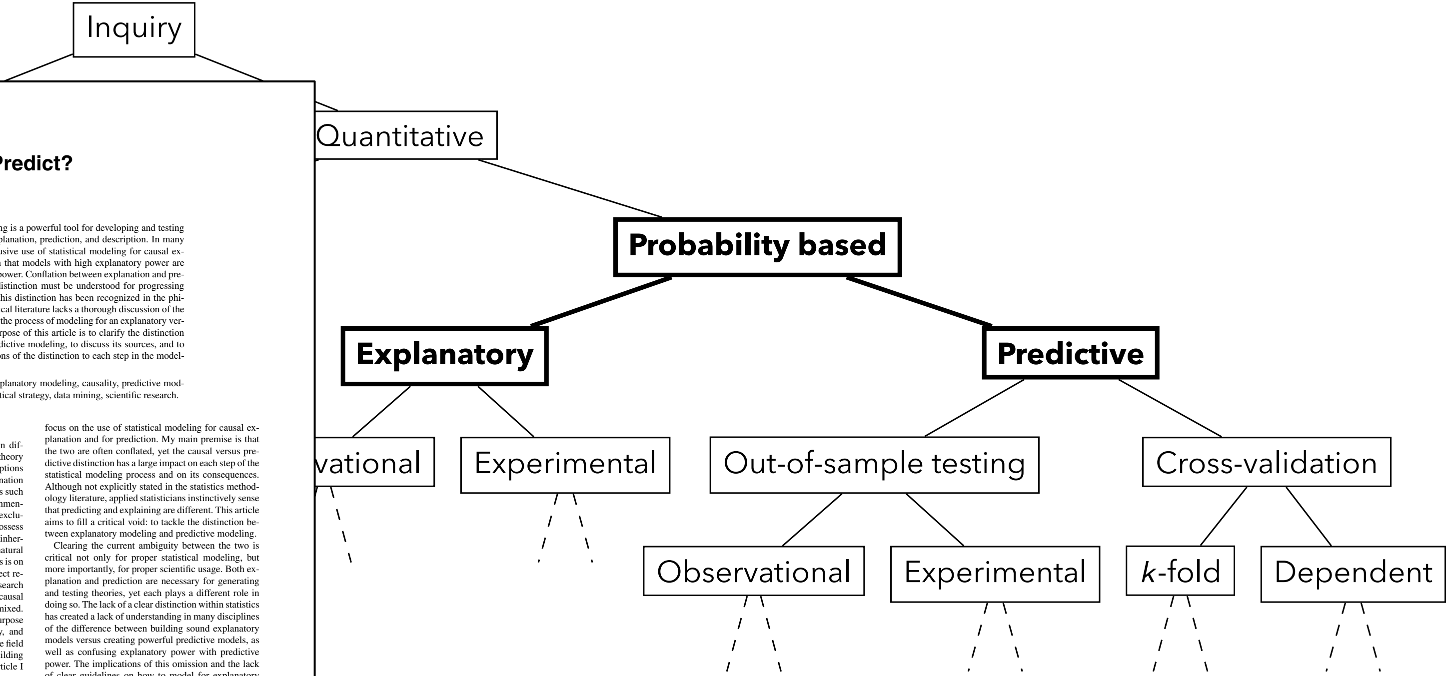
References

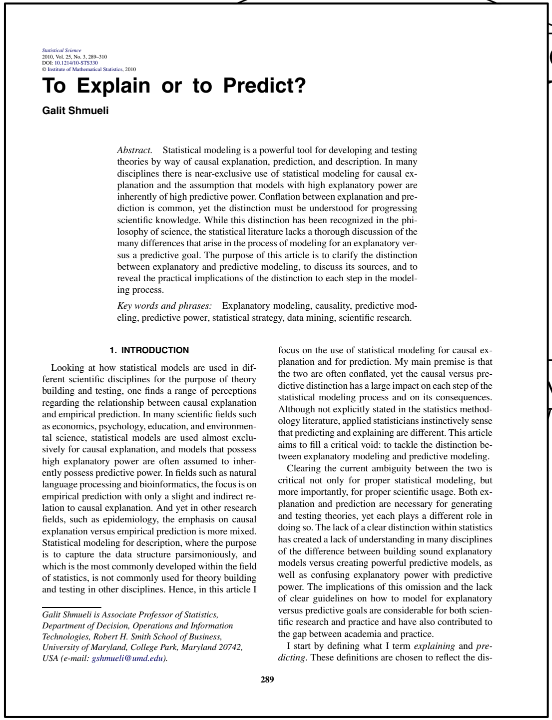
Observational vs. experimental





Statistics vs. machine Learning





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

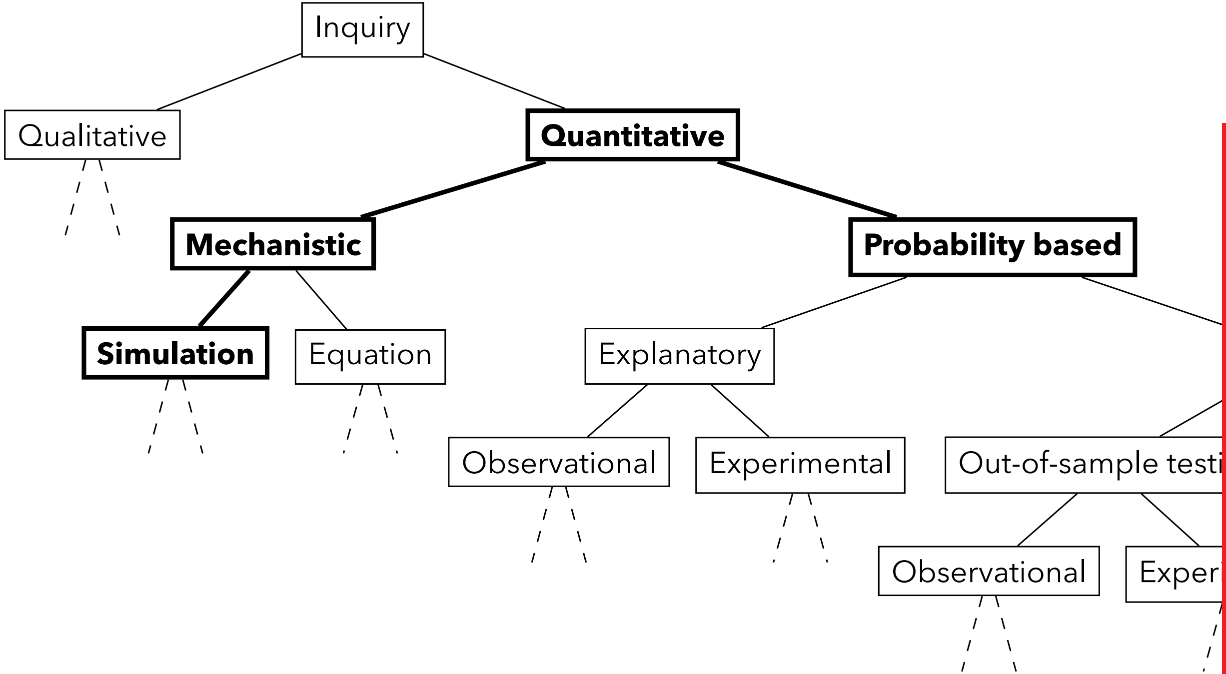
Looking at how statistical models are used in different scientific disciplines for the purpose of theory building and testing, one finds a range of perceptions regarding the relationship between causal explanation and empirical prediction. In many scientific fields such as economics, psychology, education,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

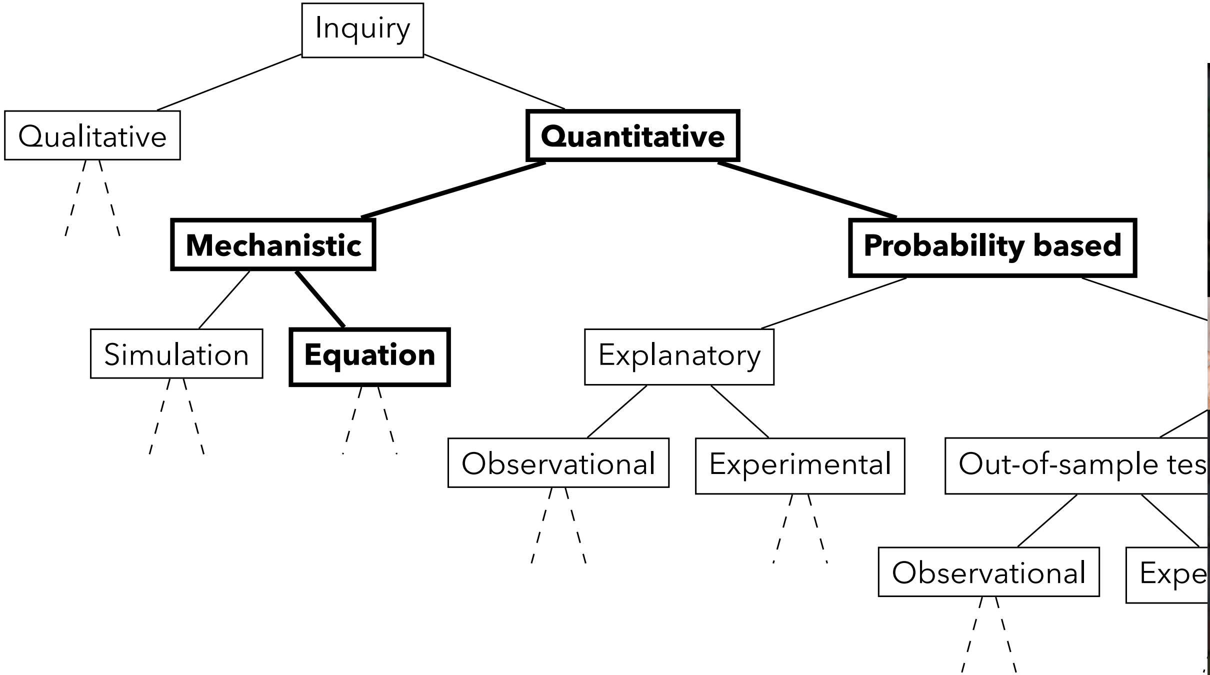
I start by defining what I term *explaining* and *predicting*. These definitions are chosen to reflect the dis-

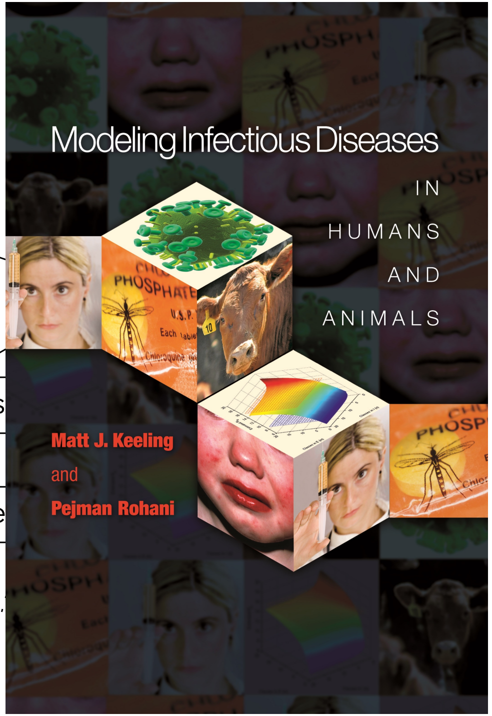
Galit Shmueli is Associate Professor of Statistics, Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742, USA (e-mail: gshmueli@umd.edu).

Simulation modeling



Analytic vs. statistical models





Introduction

Consequen-ces of quantificat-ion

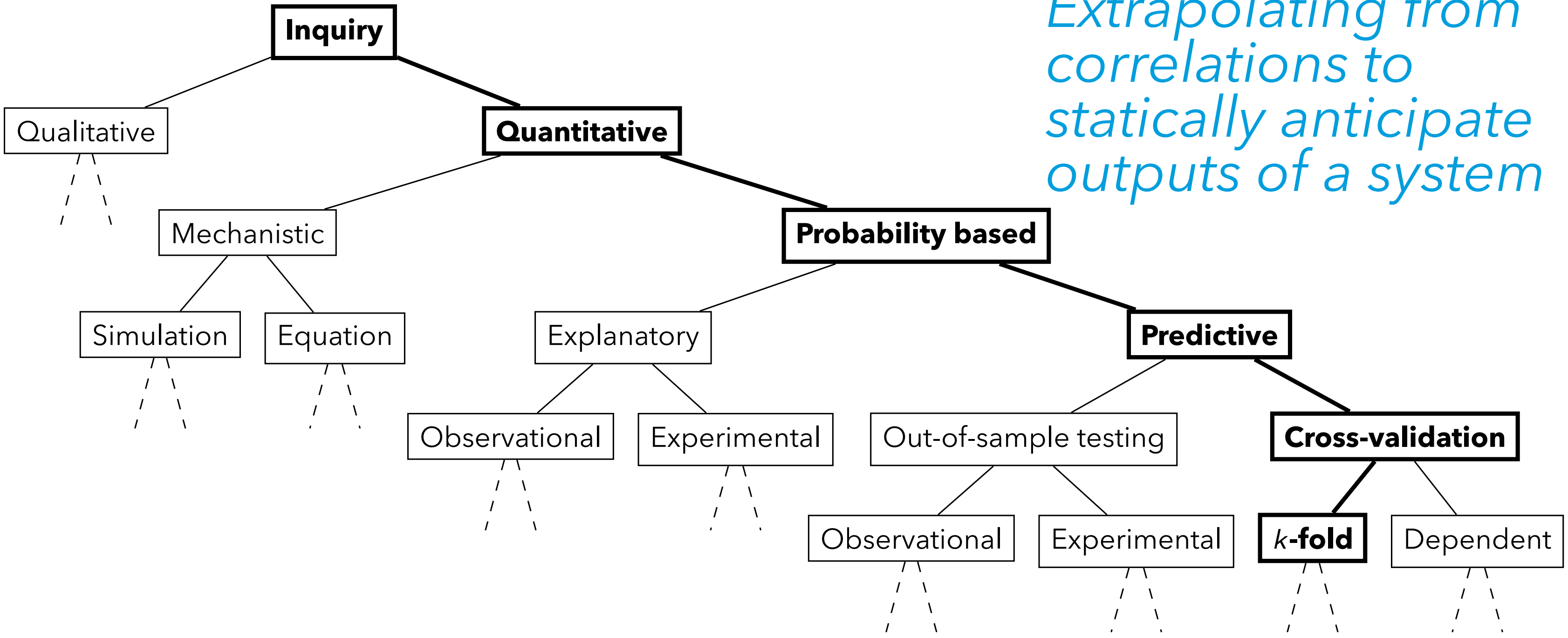
The problem with "prediction"

Machine learning logic and the law

Case study: "e-discovery"

References

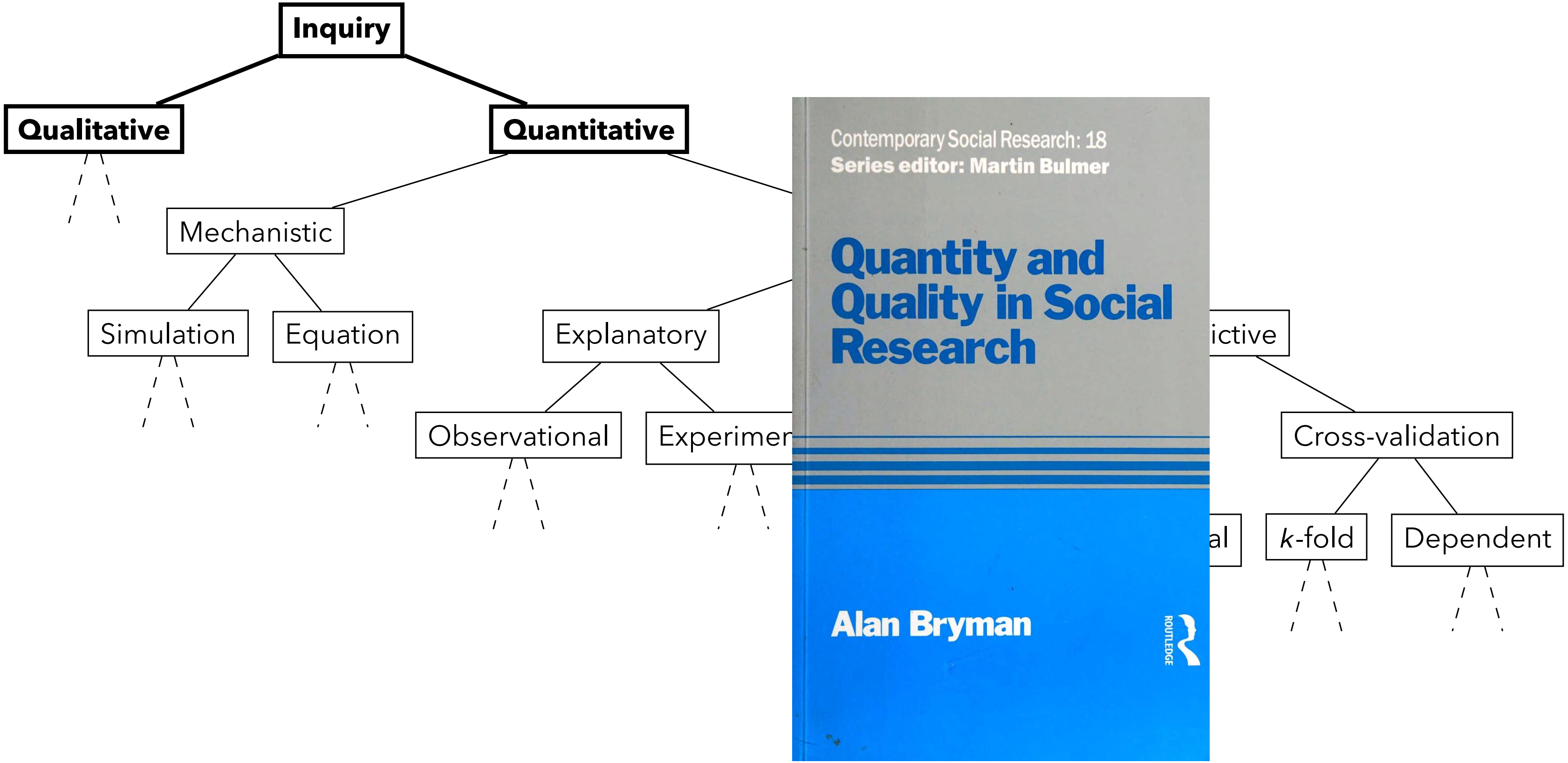
Mainstream machine learning



Extrapolating from correlations to statically anticipate outputs of a system

Consequences of quantification

Qualitative vs. quantitative



Qualitative vs. quantitative

"During the writing of this book, my first grandchild was born, and this book is dedicated to her. The hospital records document her weight, height, health, and *Apgar* score – activity (muscle tone), pulse, grimace (reflex response), appearance, and respiration. The mother's condition, length of labor, time of birth, and hospital stay are all documented... But nowhere in the hospital records will you find anything about what the birth of Calla Quinn means. Her name is recorded but not why it was chosen by her parents and what it means to them. 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 that is essential to understanding what her birth means to family and friends in this time and place." (Patton 2015)

"Understanding a person…" (slide from Barbara Kiviat)

"

	As a case [in data]	In narrative
Context/circumstance	Stripped away	Key
Mental states	Absent (for the most part)	Crucial; constitutive
Relevant features	Determined in advance	Emergent
Orientation to time	Atemporal	Chronological
Ordering of features	Unimportant	Meaningful
Other actors	Invisible	Often present
Causal logic	Mathematical	Theoretical
To boost predictive validity	Add cases	Know person better

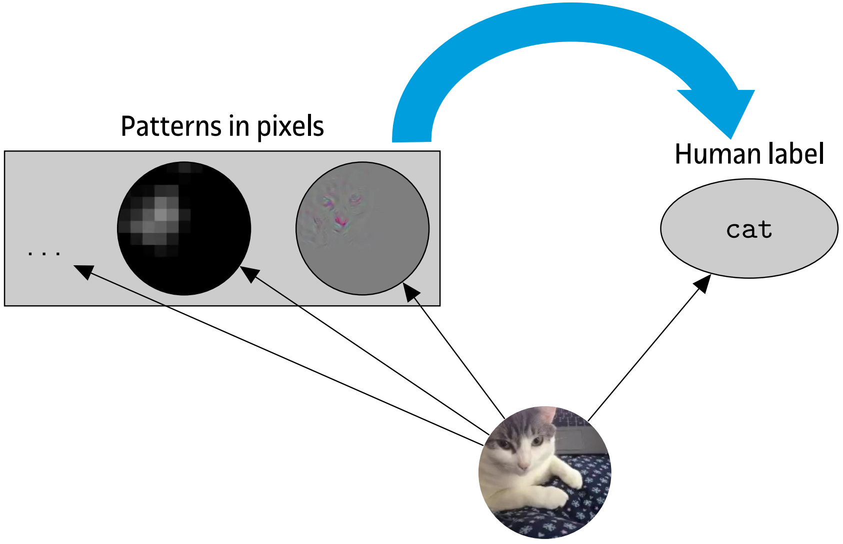
"

"Bowker and Star 2000; Bruner 1986; Desrosières 1998; Espeland 1998; Espeland and Stevens 1998, 2008; Fourcade and Healy 2017; Hacking 1990; Porter 1994, 1995; Ricouer 1998; White 1980, 1984". I would add: Patton 2015; Abbott 1988

Core incompatibility

"*…it is striking how absolutely these assumptions* [of linear models] *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… 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… all approach social causality in terms of stories, rather than in terms of variable attributes." (Abbott 1988)

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



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'

Problem: "Thinning" flattens meanings

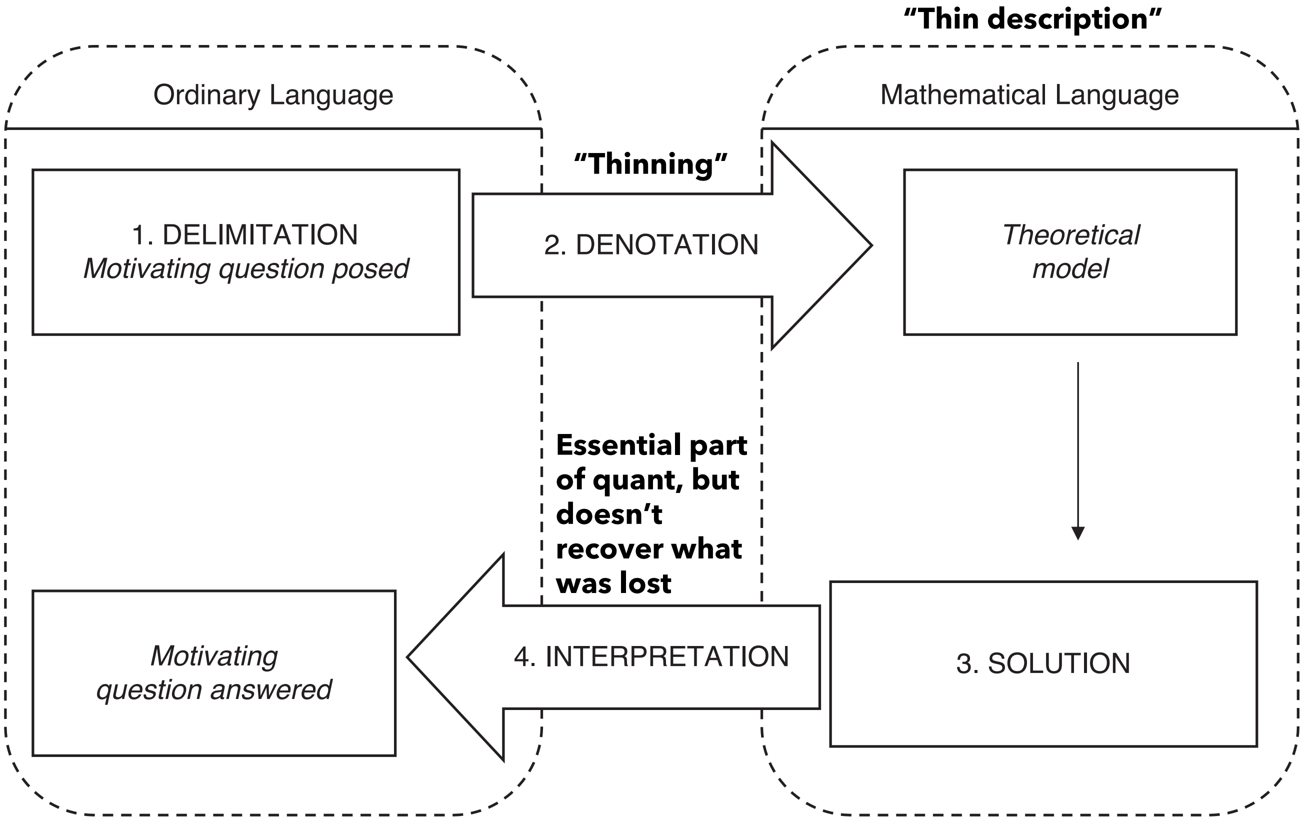
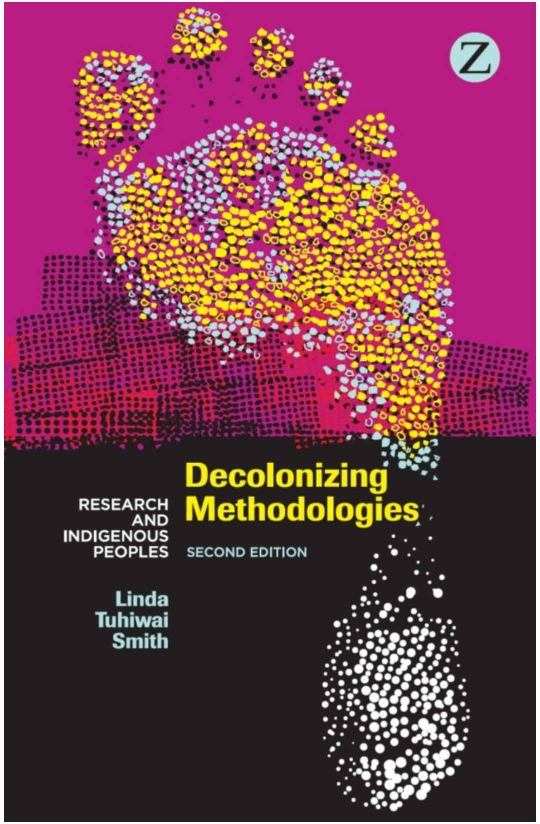


Diagram from Spiegler (2015); science (and, I would say, modeling) as "thin" description from Porter (2012)

Caution: Qual not intrinsically better

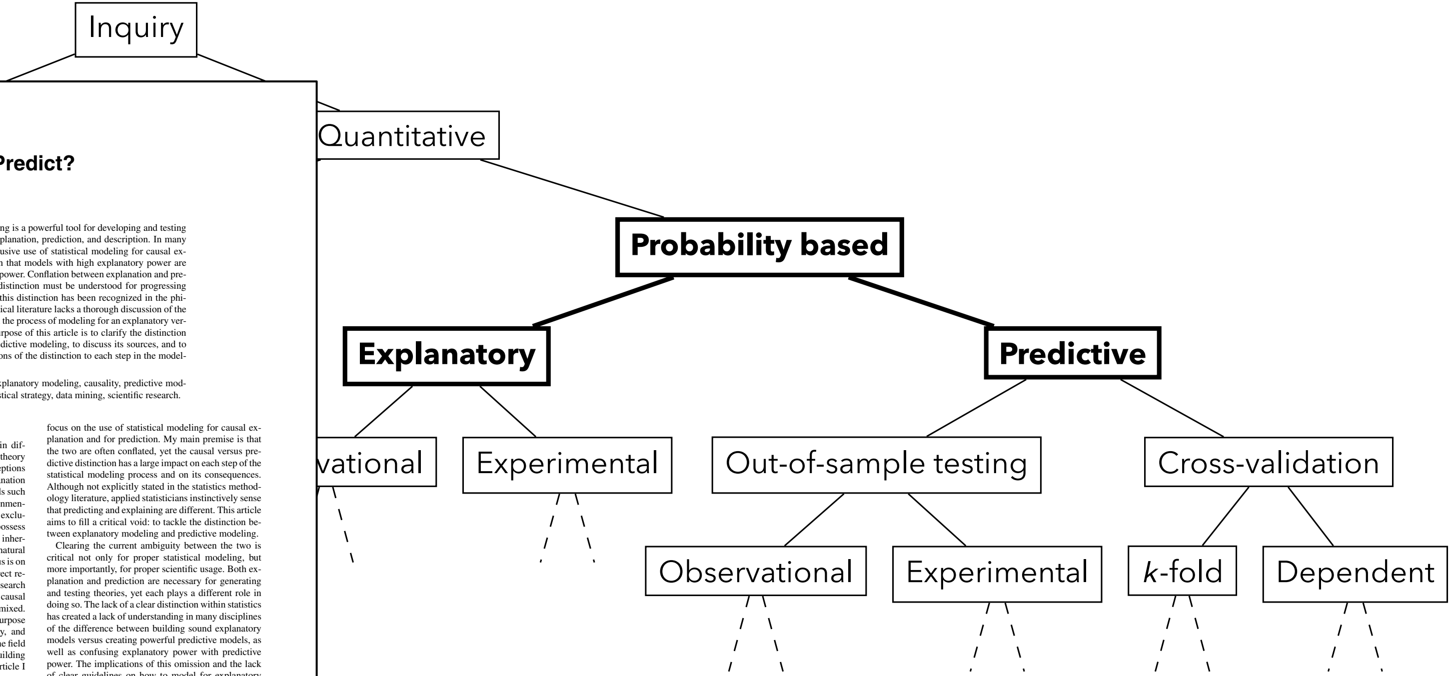
“we are suggesting that anthropological analyses (of pain and passion and power), when they are experience-distant, are at risk of delegitimating their subject matter's human conditions. The anthropologist thereby constitutes a false subject; she can engage in a professional discourse every bit as dehumanizing as that of colleagues who unreflectively draw upon the tropes of biomedicine or behaviorism to create their subject matter.” (Kleinman and Kleinman 1991; also, Tuhiwai Smith 2012 →)



The problem with "prediction"

- Introduction
- Consequences of quantification
- **The problem with "prediction"**
- Machine learning logic and the law
- Case study: "e-discovery"
- References

Statistics vs. machine Learning



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.

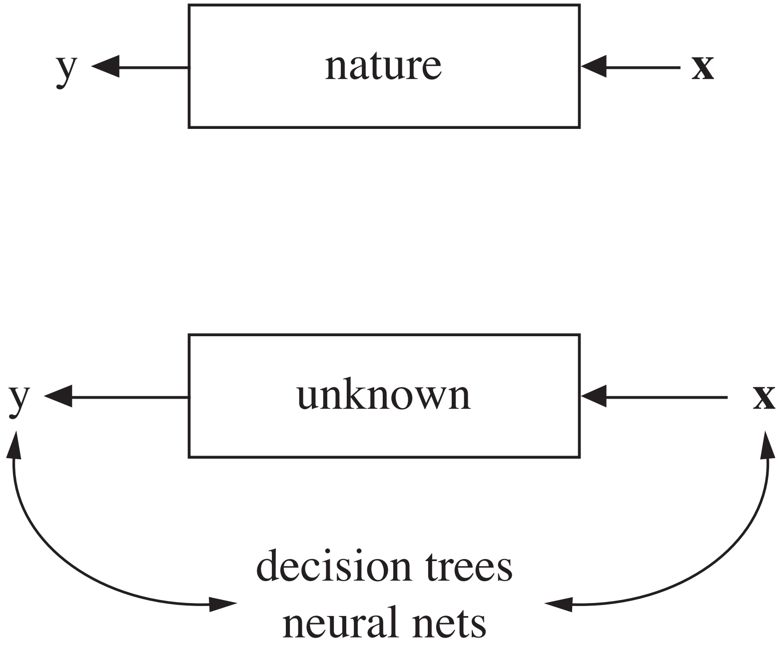
Clearing the current ambiguity between the two is critical not only for proper statistical modeling, but more importantly, for proper scientific usage. Both explanation and prediction are necessary for generating and testing theories, yet each plays a different role in doing so. The lack of a clear distinction 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

Defining machine learning



My definition: An instrumental use of statistical correlations to *mimic* the output of a target process, rather than understand the *relationship* between inputs and outputs. Involves finding expressions that maximize correlation.

Breiman 2001. See also Jones 2018.

Introduction

Consequences of quantification

The problem with "prediction"

Machine learning logic and the law

Case study: "e-discovery"

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 2012, "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper")

Tilt
Tilburg Institute for Law, Technology, and Society

Introduction

Consequences of quantification

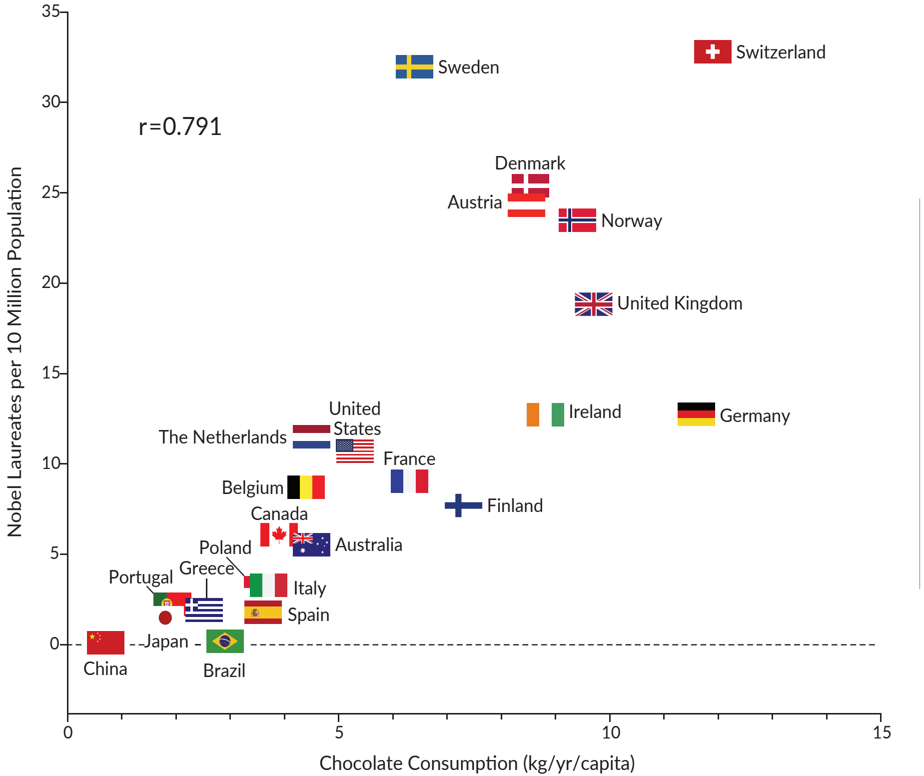
The problem with "prediction"

Machine learning logic and the law

Case study: "e-discovery"

References

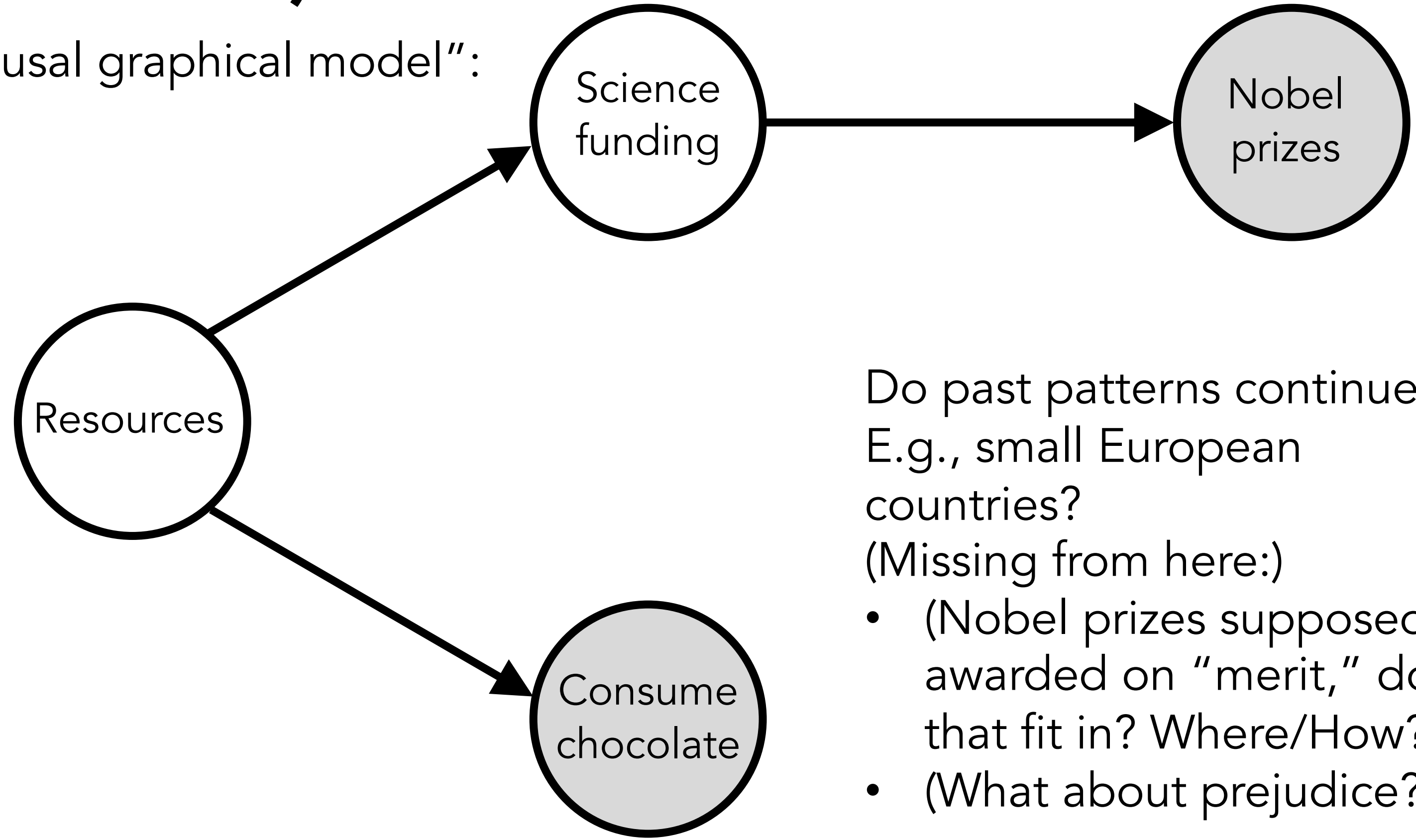
"Prediction" is correlation



Messerli 2012

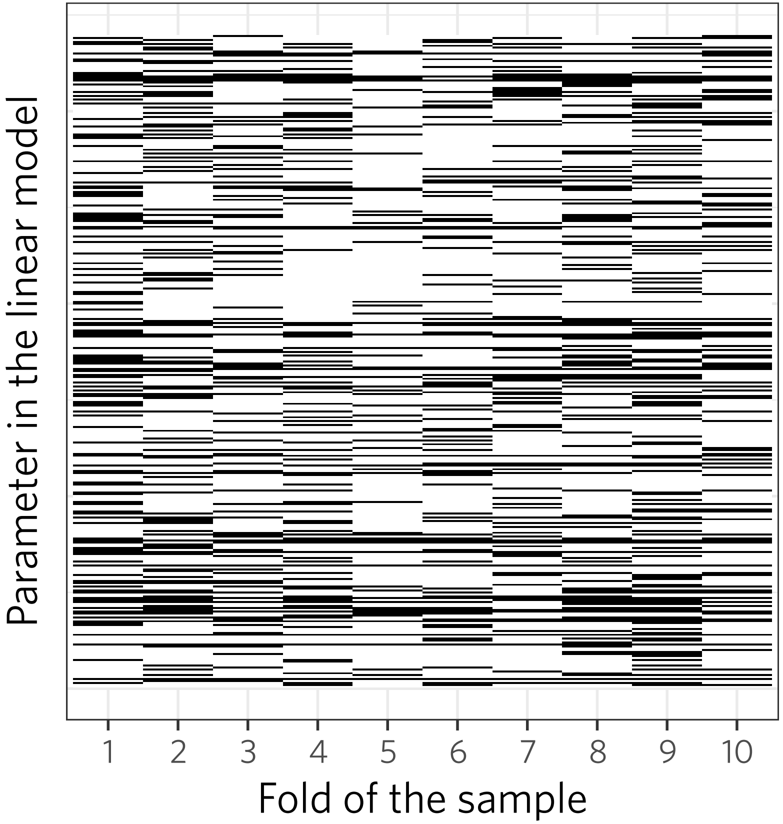
Prediction (correlation) is not explanation (causation)

A "causal graphical model":



Do past patterns continue? E.g., small European countries?
(Missing from here:)
- (Nobel prizes supposedly awarded on "merit," does that fit in? Where/How?)
- (What about prejudice?)

Can't *intervene* based on correlations



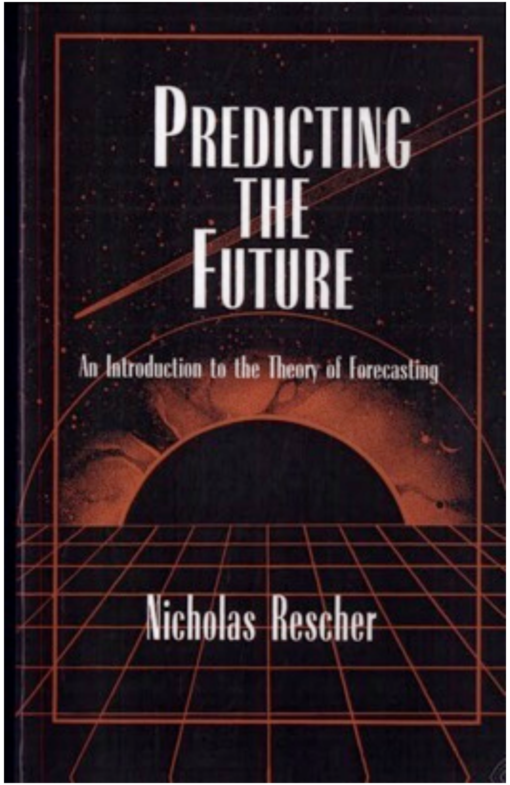
- Probably won't win more Nobel prizes by feeding population more chocolate
- Very different sets of correlations can "predict" equally well (Mullainathan and Spiess 2017)

The surprising part

- *The best-fitting* (most accurate*) *model does not necessarily reflect how the world works*
- This has been shocking in statistics for decades (Stein's paradox, Leo Breiman's "two cultures"), but little known outside
- We can "predict" without "explaining"!

* Or other relevant metric of success

Not obvious usage of "predict"

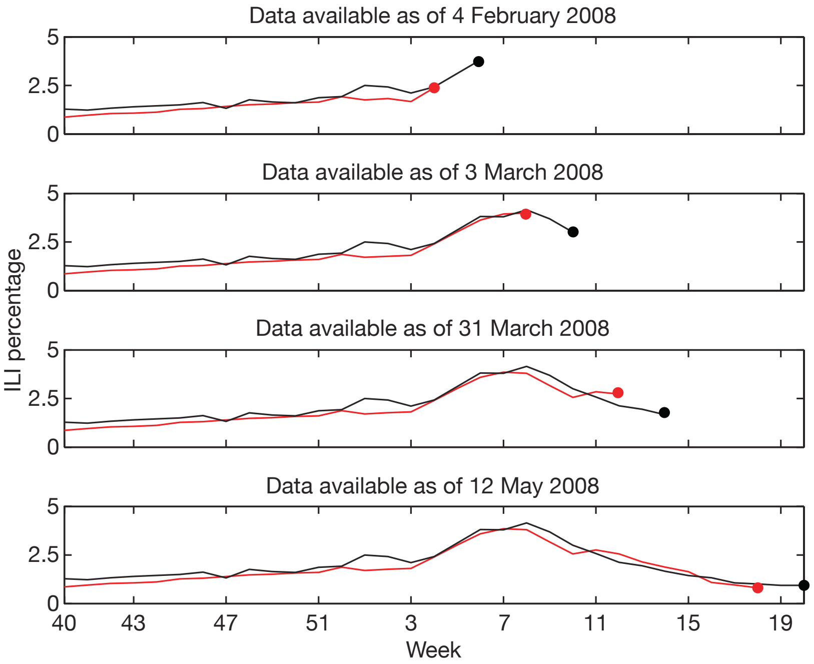


88 ■ PREDICTING THE FUTURE

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

Santillana et al. 2014

Why stick with correlations? Lucrative

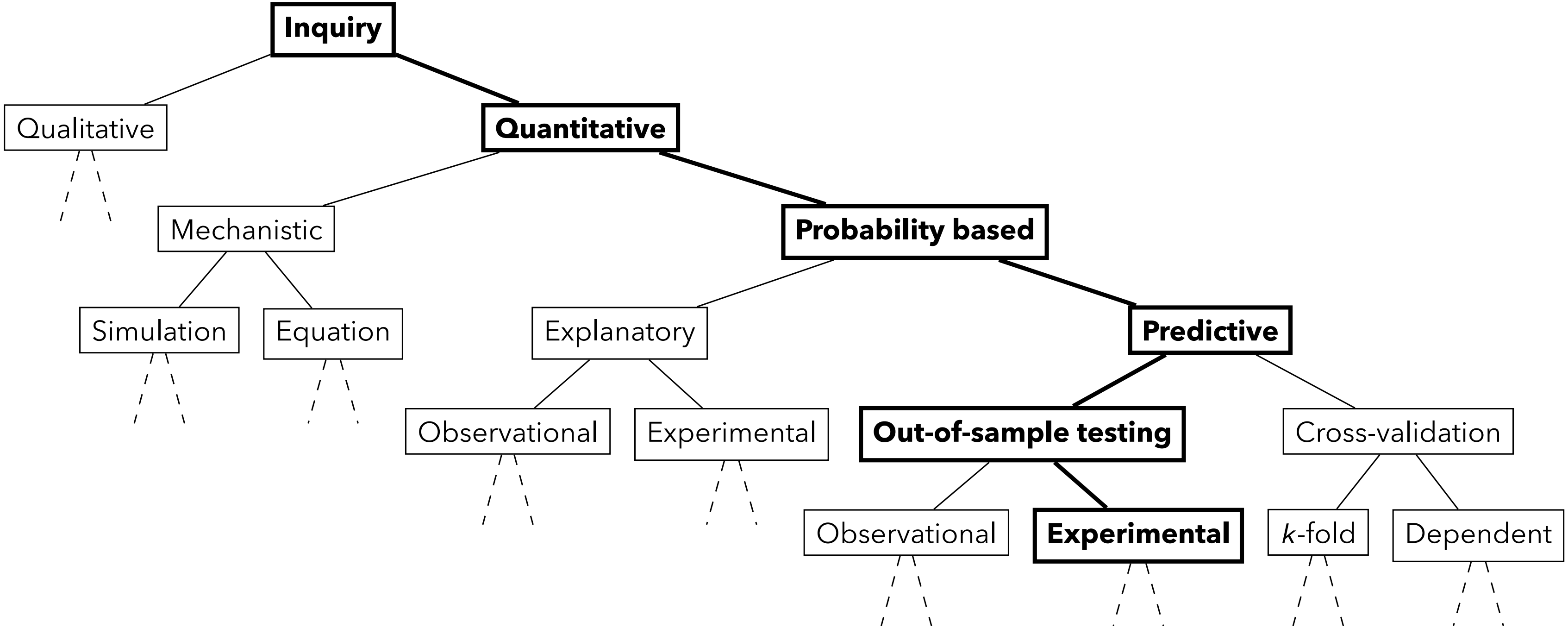
Julius C. Chappelle proposed a bill in Massachusetts to ban charging Black people more for life insurance

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

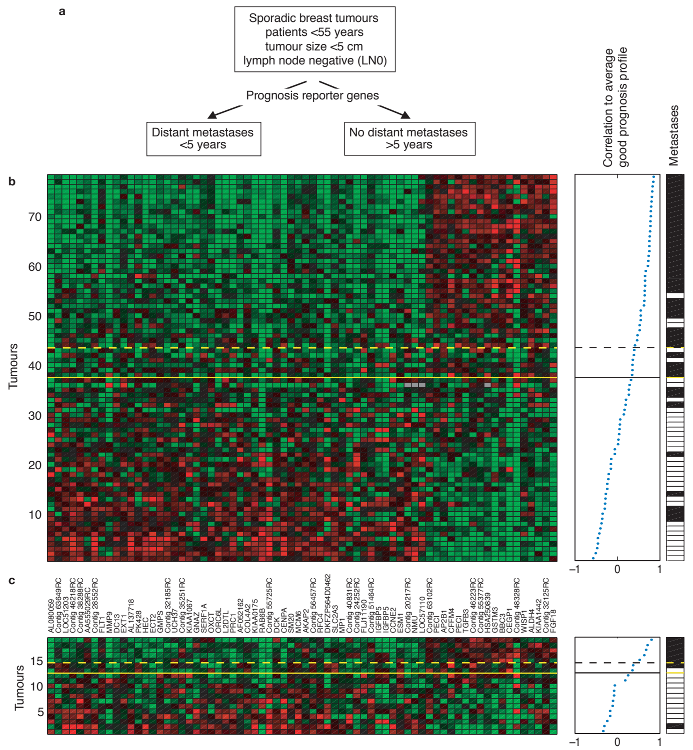
"Chappelle's allies noted that Williams's statistics, while bleak enough, answered the wrong question. The question was not whether blacks in slavery or adjusting to freedom were poor insurance risks, or even whether southern blacks were poor risks. The question was African Americans' potential for equality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

An alternative branch to the mainstream

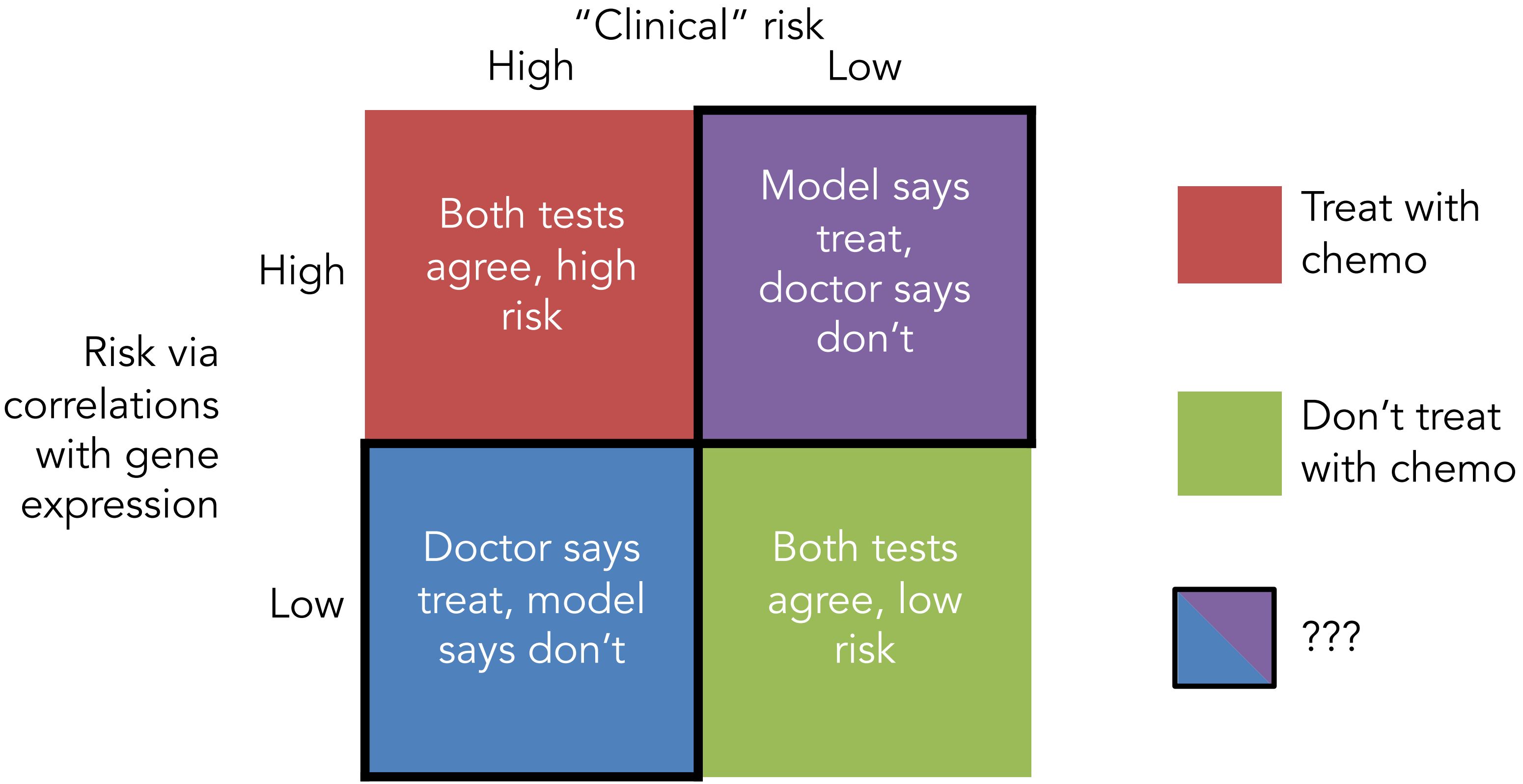


Real-world testing of "predictions"

- 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



Real-world testing

Risk via correlations with gene expression	"Clinical" risk: High	"Clinical" risk: Low
High	Both tests agree, high risk	Chemo-therapy is *worse!*
Low	Chemo-therapy is similar	Both tests agree, low risk

- Treat with chemo
- Don't treat with chemo
- ???

Real-world testing

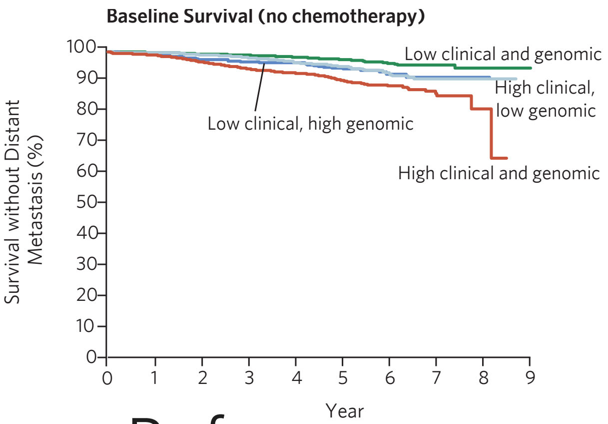
		"Clinical" risk: High	"Clinical" risk: Low
Risk via correlations with gene expression	High	Both tests agree, high risk	Chemo-therapy is *worse!*
	Low	Chemo-therapy is similar	Both tests agree, low risk

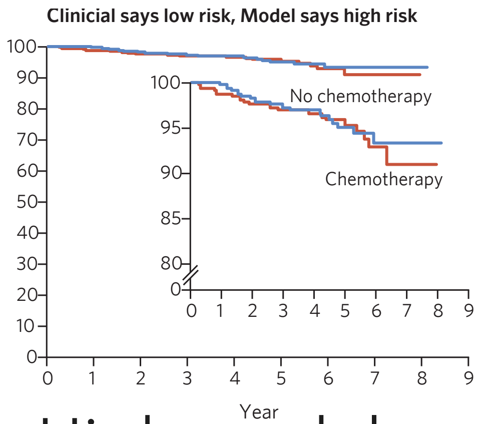
Treat with chemo

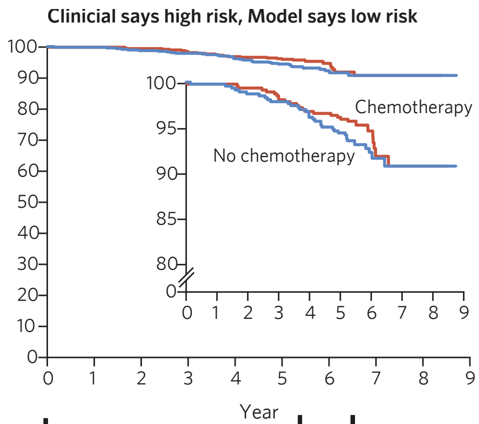
Don't treat with chemo

(Still: whose data went into the model? Who were the subjects in the experiment?)

Real-world testing: Details







- Before experiment (training data)
- High model risk, low clinical risk: randomize. Chemo *worse*!
- Low model risk, high clinical risk: chemo makes no difference

Cardoso et al. 2016

Key components of a good use case

1. We have "ground truth" (e.g., human labels, previous failures/fraud), and
2. Ground truth is hard to collect, and
3. We have some readily available proxy measure, and
4. *We don't care how or what in the proxy recovers the ground truth, only that it does*

Machine learning and the law

Group to individual inference

- When can statistical evidence be brought to bear in a courtroom?
- E.g., there is statistical evidence of a link between a chemical and a disease
- Or, a forensic test with a particular false positive rate
- There is a fundamental problem of "group to individual" (G2I) inference (Faigman et al. 2014)
- All evidence is based on reasoning based on similarity of *cases*
- What is the justification of applying the pattern to the individual?

Applying probability to individuals

- Dawid (2017) says that the foundational philosophical question of of "individual*ized* risk" is a notion of "individual risk."
- Frequentist notion of individual risk requires the assumption that "the chosen attributes capture 'all relevant characteristics' of the individuals."
- Personalist (Bayesian) notion requires the assumption of "no relevant additional information about [an individual] (or any of the other individuals in the data), and can properly assume exchangeability–conditional on the limited information that is being taken into account."
- "Neither of these requirements is fully realistic."
- See also, "What is the chance of an earthquake?" (Stark and Freedman 2003) where statisticians conclude that it is really hard to make meaning of probability statements about earthquakes

Applying machine learning

- Machine learning is even worse, if applying to law
- At least statistical evidence tries to establish the existence of a causal link that acts in a majority of cases, even if the mechanism is unknown
- A "prediction" from machine learning, i.e. a post-hoc correlation, is even weaker

Counterfactual causal thinking

Introduction

Consequences of quantification

The problem with "prediction"

Machine learning logic and the law

Case study: "e-discovery"

References

Counterfactual causal thinking

- In law: "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination" (Kohler-Hausmann 2019)
- Would somebody have been discriminated against but for the color of their skin (or for their gender) is akin to asking, would they have been discriminated against if they were a completely different person?
  - Can't just "toggle" people's attributes: they are tied up with life history, opportunity, and so much more
- In statistics, the problem is similar! (Hu 2019a, 2019b)
  - Applies to "fairness audits" for ML systems

Counterfactual causal thinking

- "the constructivist view of race claims we can only understand it within a broader system of racial subordination and domination, in which being raced Black, for example, is inextricably (probabilistically) bound up with historic disadvantage, community under-resourcing, forms of state violence, and so on and so forth… ideas about fairness and discrimination do not come to us *ex nihilo* as precise judgments about permissible causal effects, troubling mediators, and ideal adjustment criteria that need only be plugged into technical machinery to generate results that are certifiably fair." (Hu 2019b)
- Or: can't just change "Greg" to "Jamal" on a resume, or in a model. What would it take for for Jamal to be otherwise the same?

Case study: “E-discovery”

TREC Legal Track

- Text Retrieval Conference (TREC), by the United States' National Institute for Standards and Technology (NIST)
- TREC Legal Track, 2006–2011, looked at e-discovery
- How they judged the effectiveness of e-discovery is instructive to think about

E-Discovery

- Discovery phase in legal proceedings can now cover tens of millions of electronic documents
- Far too much to search through manually
- But how do we know how good ML is?

How do we know when something is relevant?

- "In order to measure the efficacy of TREC participant efforts in the two tasks, it is necessary to compare their results to a *gold standard* indicating whether or not each document in the collection is relevant to a particular discovery request." (Cormack et al. 2010)
- "The potential magnitude of the search problem is highlighted by past research indicating that lawyers greatly overestimate their true rate of recall in civil discovery (i.e., how well their searches for responsive documents have uncovered all relevant evidence, or at least all potential 'smoking guns')." (Oard et al. 2010)

Their procedure

- Get pro bono "Topic Authorities" to play the role of senior attorneys in charge of the process
- These authorities guided mostly third-year law students in reviewing and labeling tens of thousands of documents to produce a "gold standard" (label was the majority opinion of labelers)
- To build models for doing the discovery, teams were given a set of exemplars and asked to classify all documents in a corpus
  - To simulate an "interactive" e-discovery, competing teams of modelers could also ask topic authorities questions, e.g. how they would judge a given document

TREC Legal Track: Lessons/Questions

- At best, automated methods can aim to mimic what people do
- But when people are inconsistent, and/or the task is inherently ambiguous, it's hard to tell how well the automated methods have done
- Is their model of "Topic Authorities", and overlapping judgements from law students, a good way to get a "gold standard"?

References (1/2)

Abbott, Andrew. 1988. "Transcending General Linear Reality." *Sociological Theory* 6 (2): 169–186. https://doi.org/10.2307/202114

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726

Bouk, Dan. 2015. *How our Days Became Numbered: Risk and the Rise of the Statistical Individual*. The University of Chicago Press.

Cardoso, Fatima, Laura J. van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, Etienne Brain, et al. 2016. "70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer." *The New England Journal of Medicine* 375 (8): 717–729. https://doi.org/10.1056/NEJMoa1602253

Cormack, Gordon V., Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. 2010. *Overview of the TREC 2010 Legal Track*. https://trec.nist.gov/pubs/trec19/papers/LEGAL10.OVERVIEW.pdf

Dawid, Philip. 2017. "On Individual Risk." *Synthese* 194 (9): 3445–3474. https://dx.doi.org/10.1007/s11229-015-0953-4

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." https://arxiv.org/abs/1702.08608

Efron, Bradley, and Carl Morris. 1977. "Stein's Paradox in Statistics." *Scientific American* 236 (5): 119–127. https://doi.org/10.1038/scientificamerican0577-119.

Faigman, David L., John Monahan, and Christopher Slobogin. 2014. "Group to Individual (G2I) Inference in Scientific Expert Testimony." *University of Chicago Law Review* 81 (1): 417–480. https://chicagounbound.uchicago.edu/uclrev/vol81/iss2/1

Garip, Filiz. 2020. "What Failure to Predict Life Outcomes can Teach Us." *PNAS* 117 (15): 8234–8235. https://doi.org/10.1073/pnas.2003390117

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

Gayo-Avello, Daniel. 2012. "No, You Cannot Predict Elections with Twitter." *IEEE Internet Computing* 16 (6): 91–94. https://doi.org/10. 1109/MIC.2012.137

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data" *Nature* 457: 1012–1014. https://doi.org/10.1038/nature07634

Heimer, Carol A. 2001. "Cases and Biographies: An Essay on Routinization and the Nature of Comparison." *Annual Review of Sociology* 27 (1): 47–76. https://doi.org/10.1146/annurev.soc.27.1.47

Hu, Lily. 2019a. "Disparate Causes, pt. I". *Phenomenal World*, October 11. https://phenomenalworld.org/analysis/disparate-causes-i

Hu, Lily. 2019b. "Disparate Causes, pt. II". *Phenomenal World*, October 17. https://phenomenalworld.org/analysis/disparate-causes-pt-ii

Hu, Lily. 2020. "Direct Effects." *Phenomenal World*, September 25. https://phenomenalworld.org/analysis/direct-effects

Jones, Matthew L. 2018. "How We Became Instrumentalists (Again): Data Positivism since World War II." *Historical Studies in the Natural Sciences* 48 (5): 673–684. https://doi.org/10.1525/hsns.2018.48.5.673.

References (2/2)

Kleinman, Arthur, and Joan Kleinman. 1991. "Suffering and its Professional Transformation: Towards an Ethnography of Interpersonal Experience." *Culture, Medicine and Psychiatry* 15 (3): 275-301. https://doi.org/10.1007/BF00046540

Kohler-Hausmann, Issa. 2019. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination." *Northwestern University Law Review* 113 (5): 1163–1228. https://scholarlycommons.law.northwestern.edu/nulr/vol113/iss5/6/.

Lipton, Zachary C. 2015. "The Myth of Model Interpretability." *KDnuggets* 15 (13) https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability." *ACM Queue* 16 (3): 31–57. https://doi.org/10.1145/32363 86.3241340

Malik, Momin M. 2020. "A Hierarchy of Limitations in Machine Learning." https://www.arxiv.org/abs/2002.05193

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

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87.

Oard, Douglas W., Jason R. Baron, Bruce Hedin, David D. Lewis, and Stephen Tomlinson. 2010. "Evaluation of Information Retrieval for E-discovery." *Artificial Intelligence and Law* 18: 347–386. https://doi.org/10.1007/s10506-010-9093-9

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–453. https://dx.doi.org/10.1126/science.aax2342

Patton, Michael Quinn. 2015. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. 4th ed. SAGE.

Porter, Theodore M. 2012. "Thin Description: Surface and Depth in Science and Science Studies." *Osiris* 27 (1): 209–226. https://doi.org/10.1086/667828

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

Santillana, Mauricio, D. Wendong Zhang, Benjamin M. Althouse, and John W. Ayers. 2014. "What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?" *American Journal of Preventive Medicine* 47 (3): 341–347. https://doi.org/10.1016/j.amepre.2014.05.020

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310. https://doi.org/10.1214/10-STS330

Smith, Linda Tuhiwai. 2012. *Decolonizing Methodologies: Research and Indigenous Peoples*. 2nd ed. London & Zed Books.

Spiegler, Peter. 2015. *Behind the Model: A Constructive Critique of Economic Modeling*. Cambridge University Press.

Stark, Philip B., and David A. Freedman. 2003. "What is the Chance of an Earthquake?" In *Earthquake Science and Seismic Risk Reduction*, edited by Francesco Mulargia and Robert J. Geller, 201–213. NATO Science Series IV: Earth and Environmental Sciences, v. 32. Dordrecht, The Netherlands: Kluwer.

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. 2002. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer." *Nature* 415 (6871): 530–536. https://doi.org/10.1038/415530a