Course Introduction

Manoel Horta Ribeiro
manoel@cs.princeton.edu

humans & machines lab

COS 598D / Spring 2026

MENU

—

APPETIZER (5 minutes)

Manoel’s intro

—

1st COURSE (50 minutes)

Why do we need empirical reasoning?

—

2nd COURSE (30 minutes)

Course logistics

—

DESSERT (Rest of class)

Intros

MENU

APPETIZER (5 minutes)
Manoel's intro

1st COURSE (50 minutes)
Why do we need empirical reasoning?

2nd COURSE (30 minutes)
Course logistics

DESSERT (Rest of class)
Intros

BSc and MSc @ UFMG, Brazil 🇧🇷

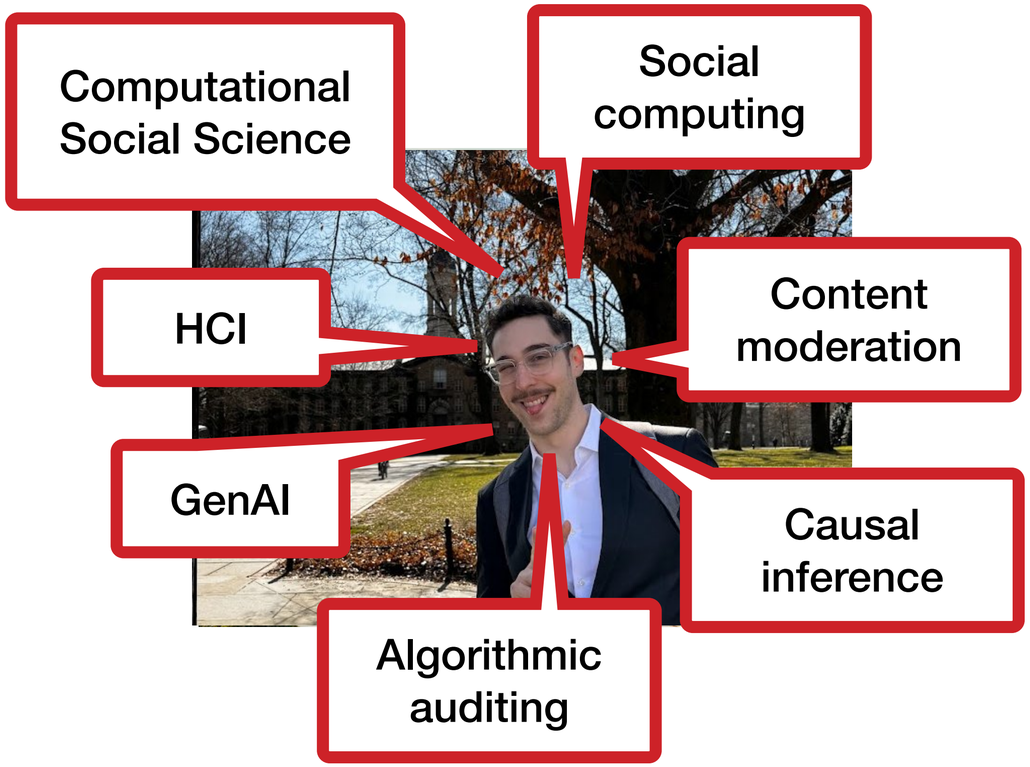
PhD @ EPFL, Switzerland, 🇨🇭

Since 2025 @ Princeton, 🇺🇸

Meta

Microsoft Research

reddit



WHAT?

- How to make online platforms better?
- What is the impact of GenAI on society?
- How can we use AI to understand humans better?

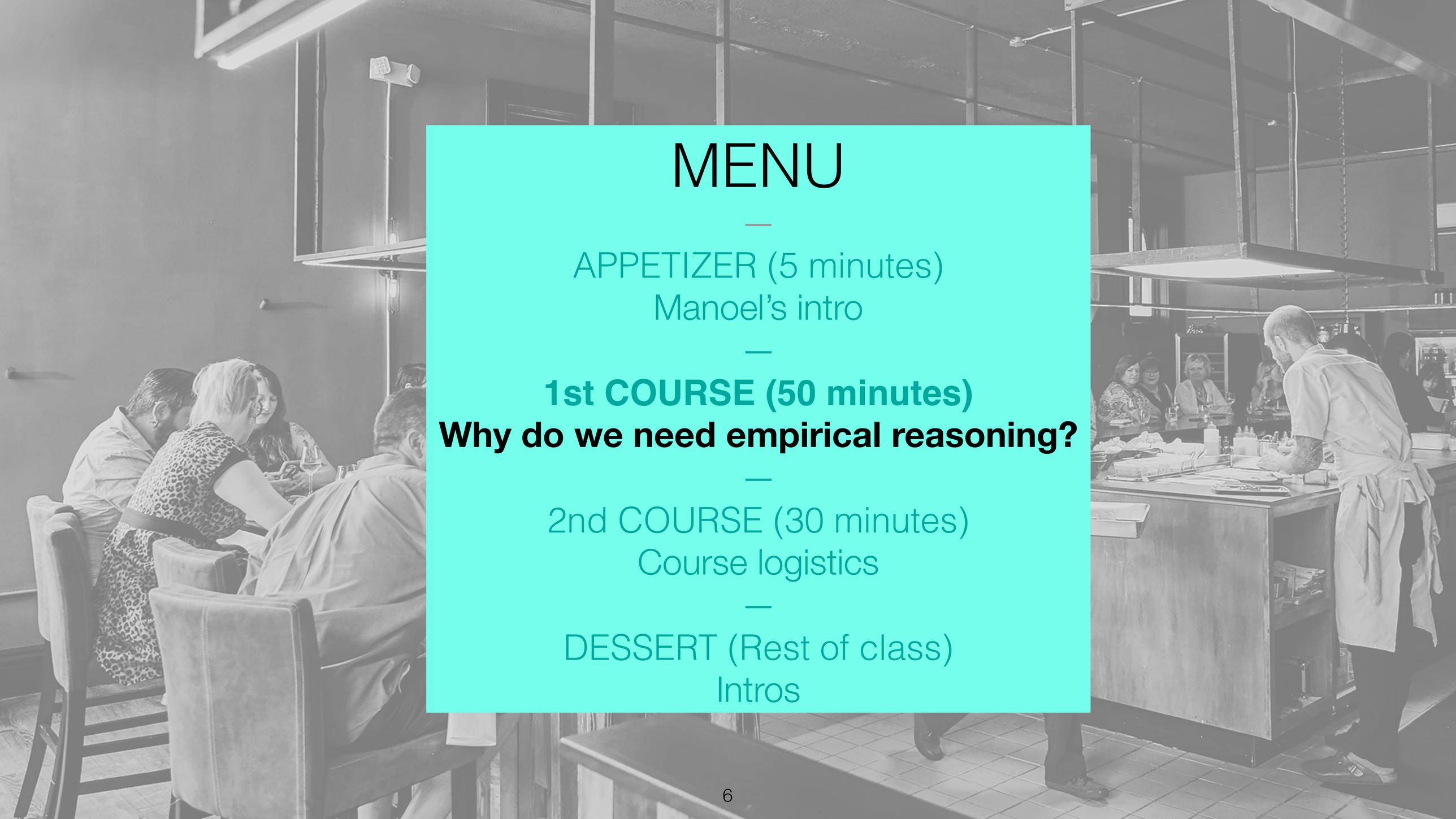
humans & machines lab

WHY?

- To improve our online spaces and our information
- To ensure we can reap the benefits and mitigate harms of new AI
- To further our understanding of how humans & machines

HOW?

- Asking (the right) causal questions!
- Developing neat models and methods!
- Running experiments!
- Studying “found” data capturing human (and machine) behavior!



MENU

—

APPETIZER (5 minutes)
Manoel's intro

—

1st COURSE (50 minutes)
Why do we need empirical reasoning?

—

2nd COURSE (30 minutes)
Course logistics

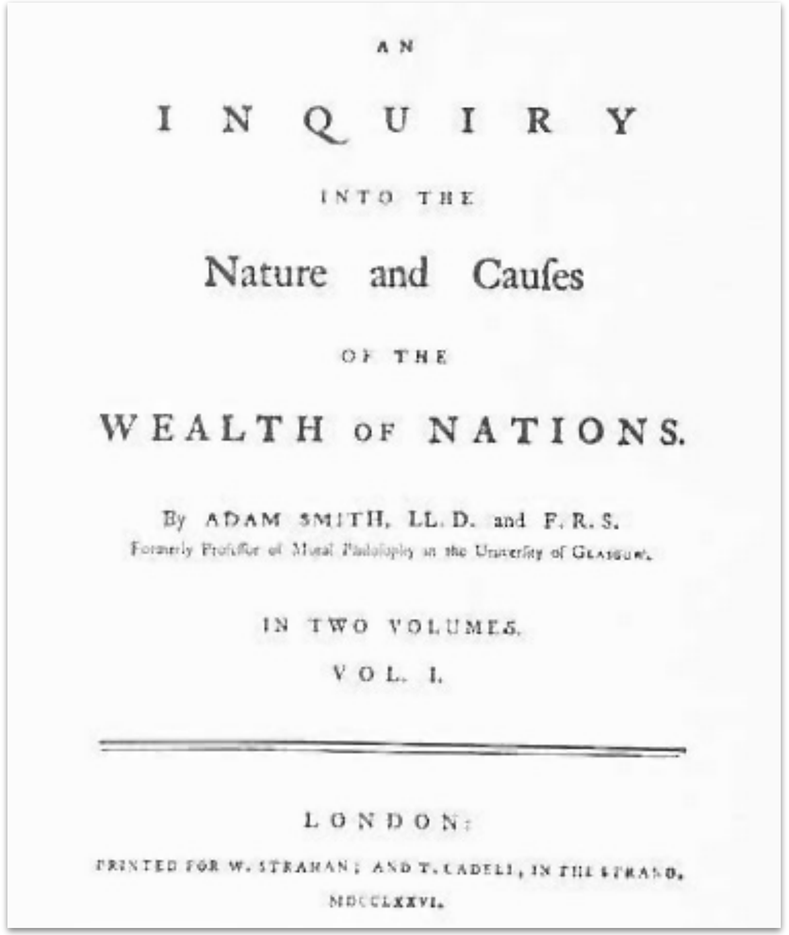
—

DESSERT (Rest of class)
Intros

Founding Myths

Scientific disciplines have "founding myths" that shape:

- How they see themselves;
- What they value;
- What they believe their role ought to be in society;



AN

INQUIRY

INTO THE

Nature and Cauſes

OF THE

WEALTH OF NATIONS.

By ADAM SMITH, LL.D. and F.R.S.

Formerly Profeſſor of Moral Philoſophy in the Univerſity of GLASGOW.

IN TWO VOLUMES.

VOL. I.

LONDON:

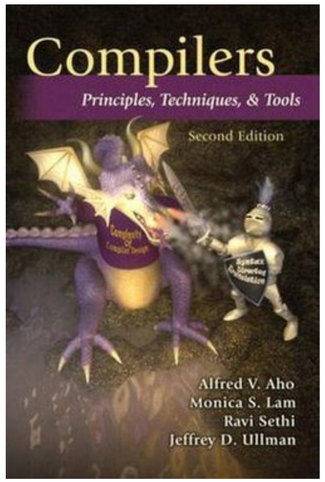
PRINTED FOR W. STRAHAN; AND T. CADELL, IN THE STRAND.

MDCCLXXVI.

Computer Science's Myths

CS's founding myths:

- theory;
- system building;



COS 583: Great Moments in Computing

Foundations of Digital Logic [Boole, 1854] [Shannon, 1938]	Operating Systems [Ritchie & Thompson, 1974] [Engler et al. 1995]
Crypto and Encryption [Diffie & Hellman, 1976] [Rivest et al., 1978]	Compilers [Hopper, 1952] [Backus et al. 1957]

Yet, Modern CS is Empirical

Modern CS is driven by observation and experiment!

Characterizing and Detecting Propaganda-Spreading Accounts on Telegram

Klim Kireev[§†], Yevhen Mykhno, Carmela Troncoso[§†], Rebekah Overdorf[‡⋆]
§ *EPFL,* † *MPI-SP Max Plank Institute for Security and Privacy*
‡ *Ruhr University Bochum (RUB), Research Center Trustworthy Data Science and Security in University Alliance Ruhr*
⋆ *University of Lausanne*

USENIX 2025 (Security & Privacy)

"I've talked to ChatGPT about my issues last night.": Examining Mental Health Conversations with Large Language Models through Reddit Analysis

KYUHA JUNG, University of California, Irvine, USA
GYUHO LEE, Seoul National University, Republic of Korea
YUANHUI HUANG, University of California, Irvine, USA
YUNAN CHEN, University of California, Irvine, USA

CSCW 2025 (Human Comp. Inter.)

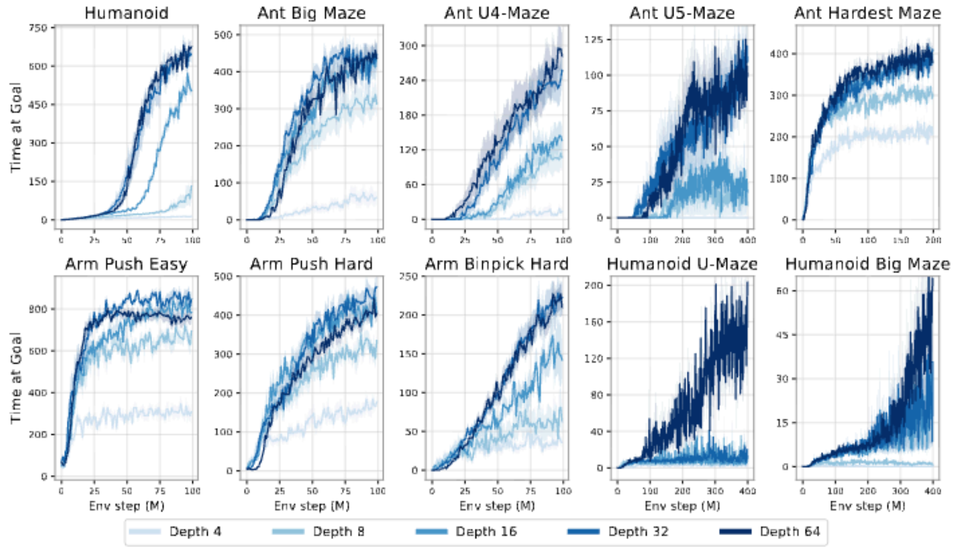


Figure 1: **Scaling network depth yields performance gains** across a suite of locomotion, navigation, and manipulation tasks, ranging from doubling performance to 50× improvements on Humanoid-based tasks. Notably, rather than scaling smoothly, performance often jumps at specific critical depths (e.g., 8 layers on Ant Big Maze, 64 on Humanoid U-Maze), which correspond to the emergence of qualitatively distinct policies (see Section 4).

Wang et al.; NeurIPS 2025 (Machine Learning)

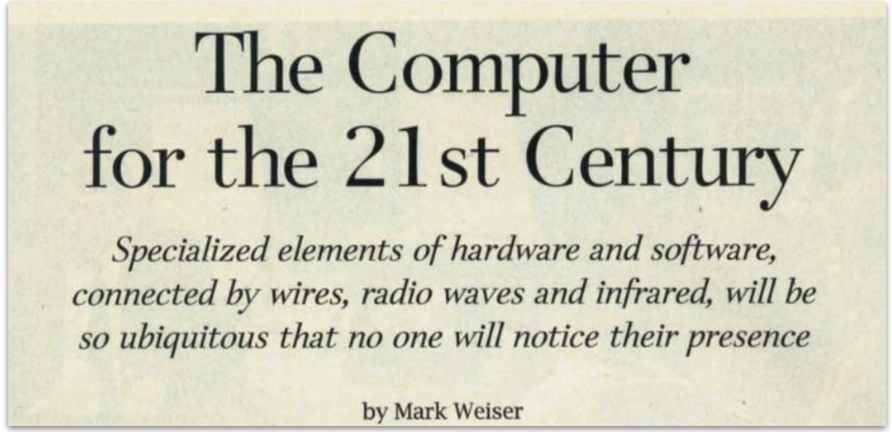
Why did this happen?

Reason #1: Socio-technical integration of computing

“Scope Creep”

Computer Science Broadened its scope. Why?

- Computing became ubiquitous.
- Problems started involving humans using computers.



The Computer for the 21st Century

Specialized elements of hardware and software, connected by wires, radio waves and infrared, will be so ubiquitous that no one will notice their presence

by Mark Weiser

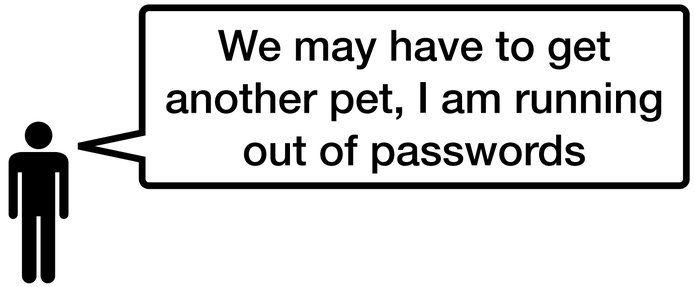
(Link)

> *“Only when things disappear in this way are we freed to use them without thinking and so to focus on beyond them on new goals”*

Security and Privacy

Original framing:

- Cryptography;
- Privacy = technical property.



Modern framing:

- All of the above;
- Also: social, norms, and context.
- Privacy may fail at the human interface.


WWW 2007 / Track: Security, Privacy, Reliability, and Ethics Session: Passwords and Phishing

A Large-Scale Study of Web Password Habits

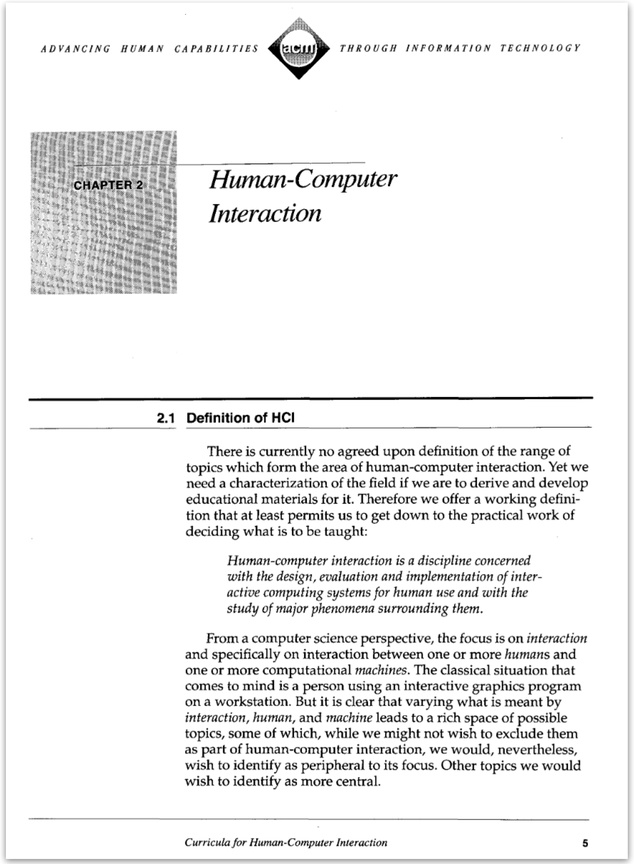
Dinei Florêncio and Cormac Herley
Microsoft Research
One Microsoft Way
Redmond, WA, USA
dinei@microsoft.com, c.herley@ieee.org

(Link)

Human Computer Interaction

The design, evaluation, and implementation of interactive systems "for human use," and with "the study of major phenomena

- Empiricism is no longer optional! You've got to do user studies!
- Over time, HCI has broadened the scope of what it considers "interaction" to be.



ADVANCING HUMAN CAPABILITIES THROUGH INFORMATION TECHNOLOGY

CHAPTER 2

Human-Computer Interaction

2.1 Definition of HCI

There is currently no agreed upon definition of the range of topics which form the area of human-computer interaction. Yet we need a characterization of the field if we are to derive and develop educational materials for it. Therefore we offer a working definition that at least permits us to get down to the practical work of deciding what is to be taught:

> *Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.*

From a computer science perspective, the focus is on *interaction* and specifically on interaction between one or more *humans* and one or more computational *machines*. The classical situation that comes to mind is a person using an interactive graphics program on a workstation. But it is clear that varying what is meant by *interaction*, *human*, and *machine* leads to a rich space of possible topics, some of which, while we might not wish to exclude them as part of human-computer interaction, we would, nevertheless, wish to identify as peripheral to its focus. Other topics we would wish to identify as more central.

Curricula for Human-Computer Interaction 5

ACM SIGCHI Curricula for HCI (1992)

SIGMETRICS

- Systems and networking communities institutionalized *measurement* itself.
- After the advent of the Internet, no one understood how it worked “in the wild.”

> *“Submissions should contribute to the current understanding of how to collect or analyze Internet measurements, or give insight into how the Internet behaves.”*

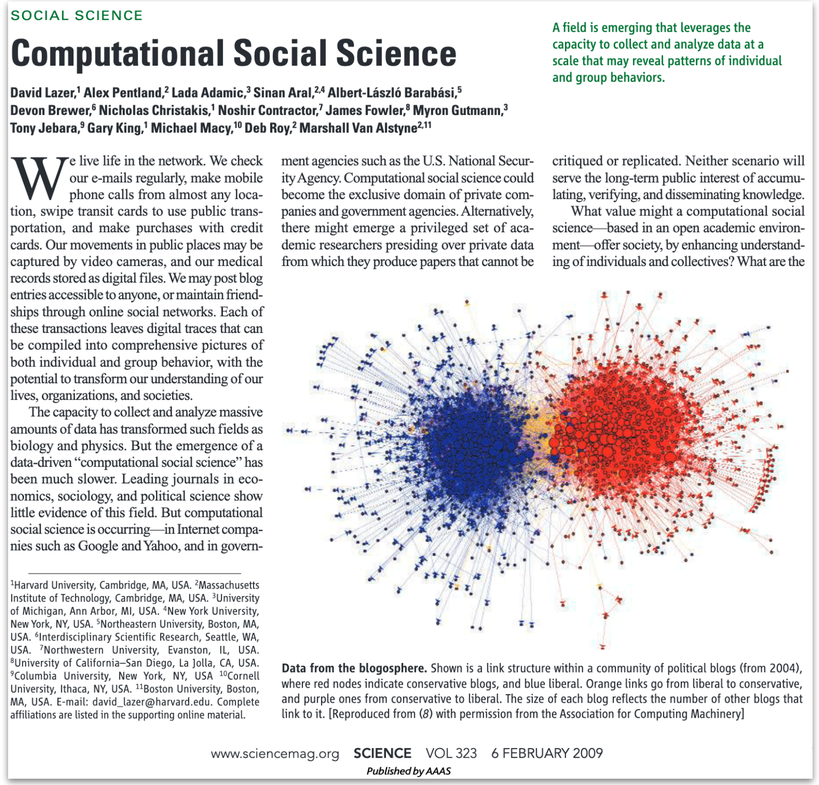
- SIGCOMM Workshop 2001

> *In a recent comparison test, six computer manufacturers were asked to code a particular program loop to run as quickly as possible on their machine. (…) We have reduced the number of Instructions for the loop by an average of one instruction per machine, a 15% decrease. It appears that conclusions might more appropriately be drawn about manufacturers' software.*

– Kernighan et al. 2002

Computational Social Science

- The study of society through digital traces.
- CSS blurs the boundary between “CS papers” and “social science papers.”
- IC2S2: hard to tell if CS or PoliSci.



SOCIAL SCIENCE

Computational Social Science

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven “computational social science” has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.

Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (*8*) with permission from the Association for Computing Machinery]

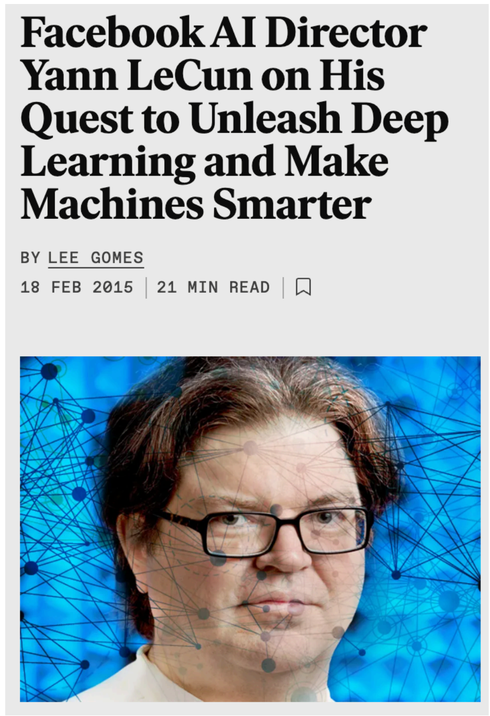
www.sciencemag.org **SCIENCE** VOL 323 6 FEBRUARY 2009
Published by AAAS

(Link)

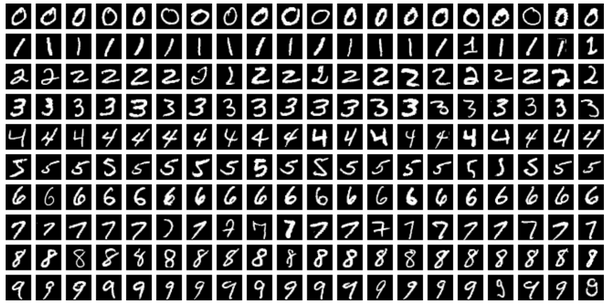
Reason #2: The Triumph of Empiricism

The Rise of “Connectionism”

- Case study: *Deep Learning*.
- 90s: Unfashionable
  - *“People used other methods because doing theory on them was easier.”*
- But… *benchmarks*!



(Link)

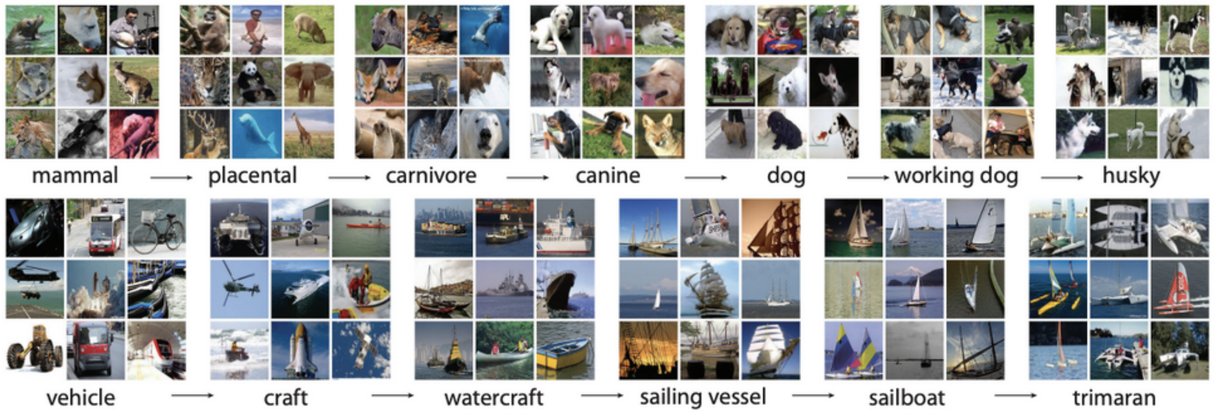


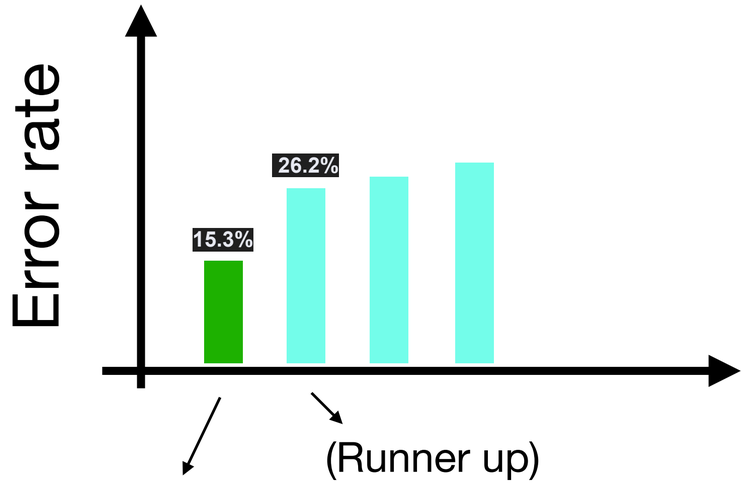
“In 1992, NIST and the Census Bureau sponsored a competition and a conference to determine the state of the art in this industry.”

(Link)

A Tipping Point: ImageNet

- Enter ImageNet (10M+ images; 1k+ labels)
- Neural networks outperformed traditional methods by wide margins!
- Performance started driving scientific progress.





AlexNet sparked a revolution because it *worked*!

Check the papers: Deng et al. 2009 (CVPR); Russakovsky et al. 2015 (IJCV)

How is Knowledge Accumulated?

Before: through theory

Vapnik (1999): Use SVMs because they are based on the developed theory.

Intuition → Theory → Model → Experiment

After: through experimentation

Srivastava (2014): "A motivation for dropout comes from a theory of the role of sex in evolution (…). A closely related motivation for dropout comes from thinking about successful conspiracies."

Intuition → Model → Experiment → Theory

Theory Often Comes *After*

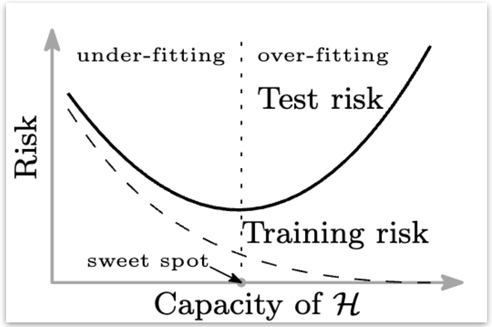
Reconciling modern machine-learning practice and the classical bias–variance trade-off

Mikhail Belkin[a,b,1], **Daniel Hsu**[c], **Siyuan Ma**[a], **and Soumik Mandal**[a]

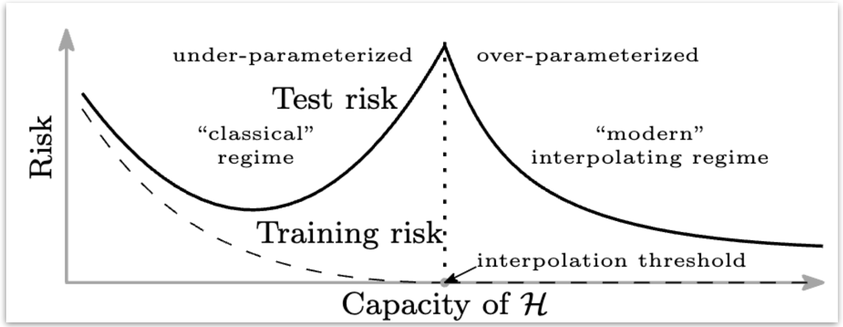
[a]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210; [b]Department of Statistics, The Ohio State University, Columbus, OH 43210; and [c]Computer Science Department and Data Science Institute, Columbia University, New York, NY 10027

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved July 2, 2019 (received for review February 21, 2019)

Theory



Practice



Understanding Dropout

Pierre Baldi
Department of Computer Science
University of California, Irvine
Irvine, CA 92697
pfbaldi@uci.edu

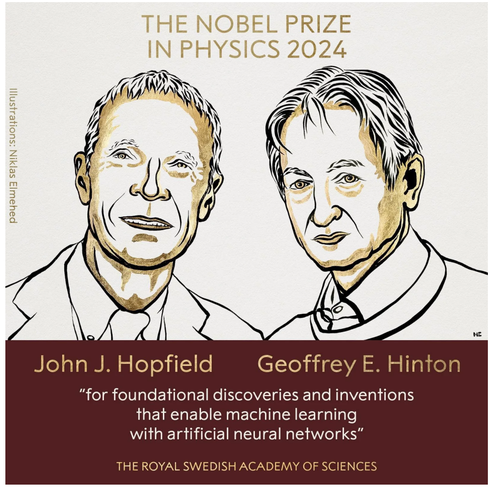
Peter Sadowski
Department of Computer Science
University of California, Irvine
Irvine, CA 92697
pjsadows@ics.uci.edu

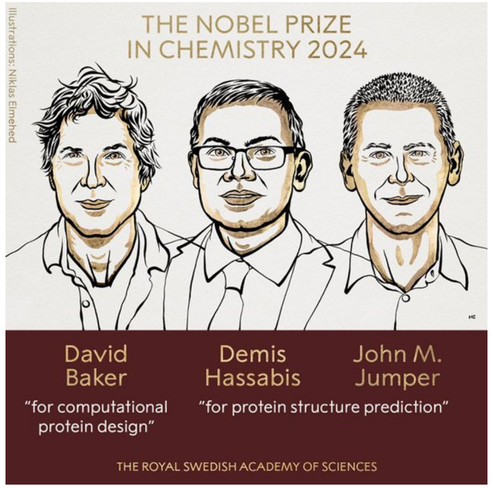
Abstract

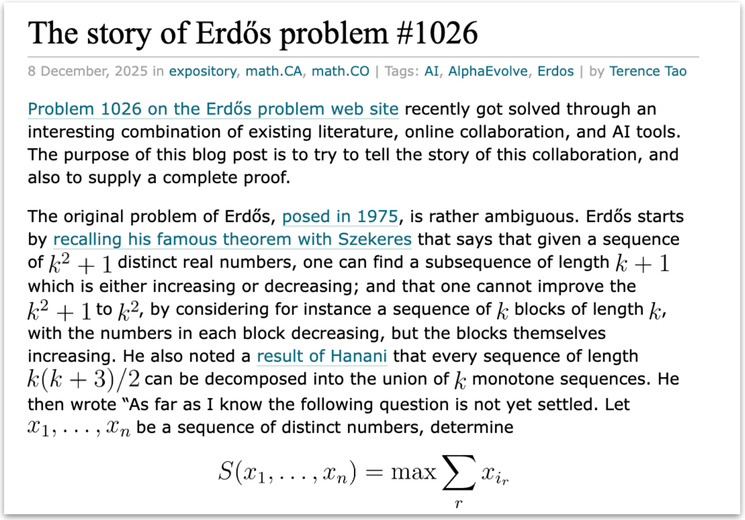
Dropout is a relatively new algorithm for training neural networks which relies on stochastically "dropping out" neurons during training in order to avoid the co-adaptation of feature detectors. We introduce a general formalism for studying dropout on either units or connections, with arbitrary probability values, and use it to analyze the averaging and regularizing properties of dropout in both linear and non-linear networks. For deep neural networks, the averaging properties of dropout are characterized by three recursive equations, including the approximation of expectations by normalized weighted geometric means. We provide estimates and bounds for these approximations and corroborate the results with simulations. Among other results, we also show how dropout performs stochastic gradient descent on a regularized error function.

Machine Learning is Eating ~~CS~~ Science

- Already true in the Deep Learning era — two Nobel prizes in 2024!
- Generative AI may become a “method layer” that percolates science, and even mathematics (see Terrence Tao’s account of Erdős #1026)
- *In CS*: New conferences, like MLSys, so much of HCI is about LLMs







The story of Erdős problem #1026

8 December, 2025 in expository, math.CA, math.CO | Tags: AI, AlphaEvolve, Erdos | by Terence Tao

Problem 1026 on the Erdős problem web site recently got solved through an interesting combination of existing literature, online collaboration, and AI tools. The purpose of this blog post is to try to tell the story of this collaboration, and also to supply a complete proof.

The original problem of Erdős, posed in 1975, is rather ambiguous. Erdős starts by recalling his famous theorem with Szekeres that says that given a sequence of k^2+1 distinct real numbers, one can find a subsequence of length $k+1$ which is either increasing or decreasing; and that one cannot improve the k^2+1 to k^2, by considering for instance a sequence of k blocks of length k, with the numbers in each block decreasing, but the blocks themselves increasing. He also noted a result of Hanani that every sequence of length $k(k+3)/2$ can be decomposed into the union of k monotone sequences. He then wrote “As far as I know the following question is not yet settled. Let $x_1,\dots,x_n$ be a sequence of distinct numbers, determine

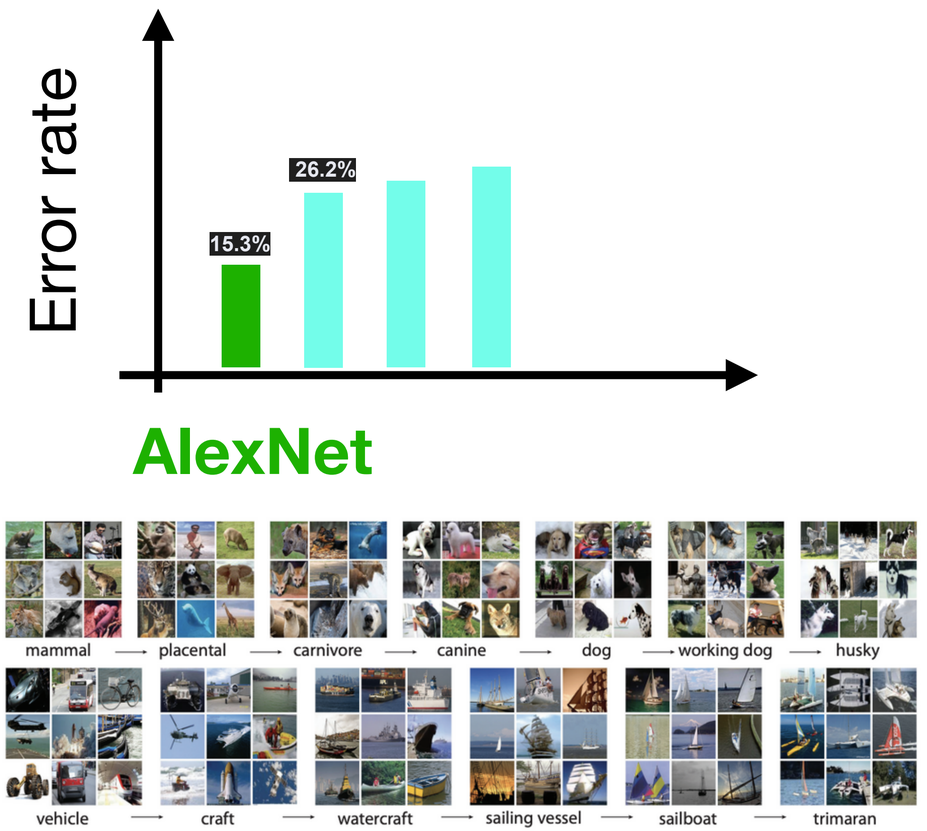
$$S(x_1,\dots,x_n) = \max \sum_r x_{i_r}$$

Wait… But…

- This whole rant started as a “we don’t have empiricism as a founding myth.”
- But then I spent so much ink trying to convince you that CS is already empirical.
- Am I just reminiscing? What’s my point?

Build-and-Test Empiricism

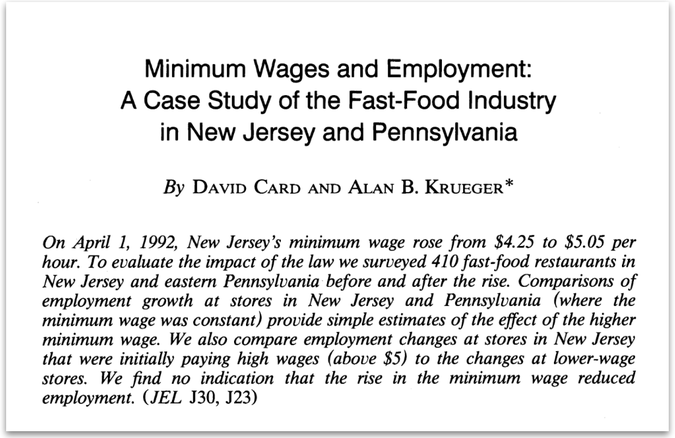
- *Engine*: Construction
  - Create a system or a model
  - Evaluate your system
  - Compare your system to others
- *Works best when* there's a shared yardstick.
- *Knowledge accumulates through* artifacts + evidence, e.g., AlexNet.



Check the papers: Deng et al. 2009 (CVPR); Russakovsky et al. 2015 (IJCV)

Describe-and-Defend Empiricism

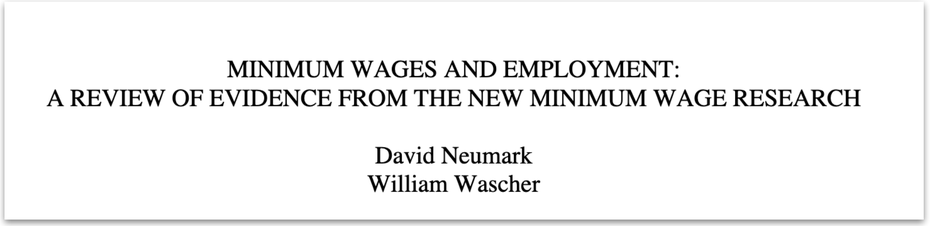
- *Engine*: Inference
  - Formulate a claim about the world
  - Assemble quantitative evidence
  - Argue why said data supports claim
- *Works best when* there is a shared language for what counts as evidence
- *Knowledge accumulates by* refining and defending generalizable statements about reality



Minimum Wages and Employment:
A Case Study of the Fast-Food Industry
in New Jersey and Pennsylvania

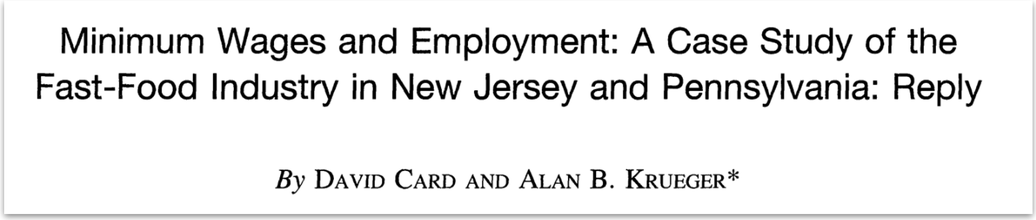
By David Card and Alan B. Krueger*

On April 1, 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above $5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment. (*JEL* J30, J23)



MINIMUM WAGES AND EMPLOYMENT:
A REVIEW OF EVIDENCE FROM THE NEW MINIMUM WAGE RESEARCH

David Neumark
William Wascher



Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply

By David Card and Alan B. Krueger*

Maintenance Layer

Engine

How results get produced.

The maintenance layer

How to ensure the engine doesn't produce nonsense.

Build-and-Test Empiricism

Maintenance layer: rituals

NeurIPS Paper Checklist Guidelines

The NeurIPS Paper Checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. The checklist is included in the LateX style file. Do not remove the checklist: **The papers not including the checklist will be desk rejected**.

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Table 5: Ablation over the pre-training tasks using the $\text{BERT}_{\text{BASE}}$ architecture. "No NSP" is trained without the next sentence prediction task. "LTR & No NSP" is trained as a left-to-right LM without the next sentence prediction, like OpenAI GPT. "+ BiLSTM" adds a randomly initialized BiLSTM on top of the "LTR + No NSP" model during fine-tuning.

Describe and Defend Empiricism

Maintenance layer: robustness checks

Appendix Table of Contents

Causal Panel Analysis under Parallel Trends: Lessons from a Large Reanalysis Study

ALBERT CHIU *Stanford University, United States*
XINGCHEN LAN *New York University, United States*
ZIYI LIU *University of California, Berkeley, United States*
YIQING XU *Stanford University, United States*

Two-way fixed effects (TWFE) models are widely used in political science to establish causality, but recent methodological discussions highlight their limitations under heterogeneous treatment effects (HTE) and violations of the parallel trends (PT) assumption. This growing literature has introduced numerous new estimators and procedures, causing confusion among researchers about the reliability of existing results and best practices. To address these concerns, we replicated and reanalyzed 49 studies from leading journals that employ TWFE models for causal inference using observational panel data with binary treatments. Using six HTE-robust estimators, diagnostic tests, and sensitivity analyses, we find: (i) HTE-robust estimators yield qualitatively similar but highly variable results; (ii) while a few studies show clear signs of PT violations, many lack evidence to support this assumption; and (iii) many studies are underpowered when accounting for HTE and potential PT violations. We emphasize the importance of strong research designs and rigorous validation of key identifying assumptions.

Two Stories

On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach

STEVEN L. SALZBERG salzberg@cs.jhu.edu
Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

Editor: Usama Fayyad

Abstract. An important component of many data mining projects is finding a good classification algorithm, a process that requires very careful thought about experimental design. If not done very carefully, comparative studies of classification and other types of algorithms can easily result in statistically invalid conclusions. This is especially true when one is using data mining techniques to analyze very large databases, which inevitably contain some statistically unlikely data. This paper describes several phenomena that can, if ignored, invalidate an experimental comparison. These phenomena and the conclusions that follow apply not only to classification, but to computational experiments in almost any aspect of data mining. The paper also discusses why comparative analysis is more important in evaluating some types of algorithms than for others, and provides some suggestions about how to avoid the pitfalls suffered by many experimental studies.

Keywords: classification, comparative studies, statistical methods

1. Introduction

Data mining researchers often use classifiers to identify important classes of objects within a data repository. Classification is particularly useful when a database contains examples that can be used as the basis for future decision making; e.g., for assessing credit risks, for medical diagnosis, or for scientific data analysis. Researchers have a range of different types of classification algorithms at their disposal, including nearest neighbor methods, decision tree induction, error back propagation, reinforcement learning, and rule learning. Over the years, many variations of these algorithms have been developed and many studies have been produced comparing their effectiveness on different data sets, both real and artificial. The productiveness of classification research in the past means that researchers today confront a problem in using those algorithms, namely: how does one choose which algorithm to use for a new problem? This paper addresses the methodology that one can use to answer this question, and discusses how it has been addressed in the classification community. It also discusses some of the pitfalls that confront anyone trying to answer this question, and demonstrates how misleading results can easily follow from a lack of attention to methodology. Below, I will use examples from the machine learning community which illustrate how careful one must be when using fast computational methods to mine a large database. These examples show that when one repeatedly searches a large database with powerful algorithms, it is all too easy to "find" a phenomenon or pattern that looks impressive, even when there is nothing to discover.

It is natural for experimental researchers to want to use real data to validate their systems. A culture has evolved in the machine learning community that now insists on a convincing evaluation of new ideas, which very often takes the form of experimental testing. This is a

Journal of Econometrics 225 (2021) 254–277

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Difference-in-differences with variation in treatment timing☆

Andrew Goodman-Bacon *
Opportunity and Inclusive Growth Institute, Federal Reserve Bank of Minneapolis, 90 Hennepin Ave, Minneapolis, MN 55401, USA
National Bureau of Economic Research, USA

ARTICLE INFO

Article history:
Received 19 January 2021
Received in revised form 19 January 2021
Accepted 17 March 2021
Available online 12 June 2021

Keywords:
Difference-in-differences
Variation in treatment timing
Two-way fixed effects
Treatment effect heterogeneity

ABSTRACT

The canonical difference-in-differences (DD) estimator contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper shows that the two-way fixed effects estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. A causal interpretation of two-way fixed effects DD estimates requires both a parallel trends assumption and treatment effects that are constant over time. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls.

Published by Elsevier B.V.

1. Introduction

Difference-in-differences (DD) is both the most common and the oldest quasi-experimental research design, dating back to Snow's (1855) analysis of a London cholera outbreak.[1] A DD estimate is the difference between the change in outcomes before and after a treatment (difference one) in a treatment versus control group (difference two): $\left(\bar{y}^{POST}_{TREAT} - \bar{y}^{PRE}_{TREAT}\right) - \left(\bar{y}^{POST}_{CONTROL} - \bar{y}^{PRE}_{CONTROL}\right)$. That simple quantity also equals the estimated coefficient on the interaction of a treatment group dummy and a post-treatment period dummy in the following regression:

$$y_{it} = \gamma + \gamma_i TREAT_i + \gamma_t POST_t + \beta^{2x2} TREAT_i \times POST_t + u_{it}. \quad (1)$$

The elegance of DD makes it clear which comparisons generate the estimate, what leads to bias, and how to test the design. The expression in terms of sample means connects the regression to potential outcomes and shows that, under a common trends assumption, a two-group/two-period (2x2) DD identifies the average treatment effect on the treated. Almost all econometrics textbooks and survey articles describe this structure,[2] and recent methodological extensions build on it.[3]

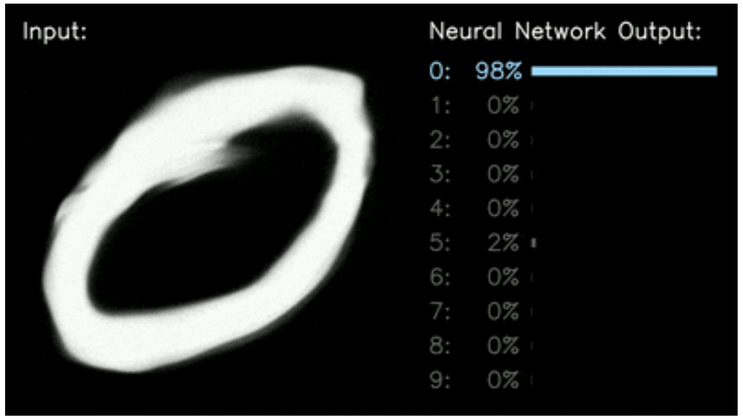
But Why Isn’t Build-and-Test Empiricism Enough?

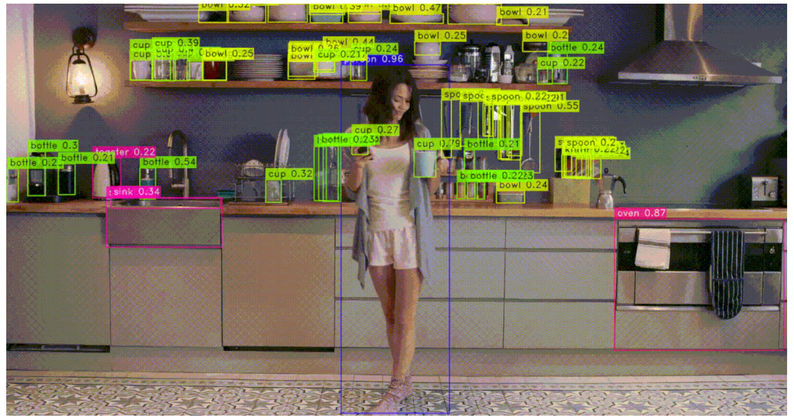
B&T is actually remarkable

The Emerging Science of Machine Learning Benchmarks

Moritz Hardt

"Researchers might've indeed dismissed the idea of benchmarking as unreasonable, if it hadn't been so successful."





We are studying non-benchmarkable things

- Mismatch between “build-and-test” habits and “claim-centered” research.
  - This design choice changes behavior, or
  - This intervention improves real outcomes.
- High-stakes claims, unserious methods.

Deplatforming Norm-Violating Influencers on Social Media Reduces Overall Online Attention Toward Them

MANOEL HORTA RIBEIRO, EPFL, Switzerland
SHAGUN JHAVER, Rutgers University, USA
JORDI CLUET-I-MARTINELL, EPFL, Switzerland
MARIE REIGNIER-TAYAR, EPFL, Switzerland
ROBERT WEST, EPFL, Switzerland

From politicians to podcast hosts, online platforms have systematically banned (“deplatformed”) influential users for breaking platform guidelines. Previous inquiries on the effectiveness of this intervention are inconclusive because 1) they consider only a few deplatforming events; 2) they consider only overt engagement traces (e.g., likes and posts) but not passive engagement (e.g., views); 3) they do not consider all the potential places influencers impacted by the deplatforming event might migrate to. We address these limitations in a longitudinal, quasi-experimental study of 165 deplatforming events targeting 101 influencers. We identify deplatforming events through Reddit posts and then manually curate the data, ensuring the correctness of a large dataset of deplatforming events. Then, we link these events to Google Trends and Wikipedia page views, platform-agnostic measures of online attention that capture the general public's interest in specific influencers. Through a difference-in-differences approach, we find that deplatforming reduces online attention toward influencers. After 12 months, we estimate that online attention toward deplatformed influencers is reduced by −63% (95% CI [−75%,−46%]) on Google and by −43% (95% CI [−57%,−24%]) on Wikipedia. Further, as we study over a hundred deplatforming events, we can analyze in which cases deplatforming is more or less impactful, revealing nuances about the intervention. Notably, we find that both permanent and temporary deplatforming reduces online attention toward influencers and that deplatforming influencers from multiple platforms further reduces the online attention they receive. Overall, this work contributes to the ongoing effort to map the effectiveness of content moderation interventions, driving platform governance away from speculation.

CCS Concepts: • **Information systems → Social networking sites**; **Social networks**; • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: online communities; fringe online communities; content moderation; online radicalization; deplatforming; social networks

ACM Reference Format:
Manoel Horta Ribeiro, Shagun Jhaver, Jordi Cluet-i-Martinell, Marie Reignier-Tayar, and Robert West. 2025. Deplatforming Norm-Violating Influencers on Social Media Reduces Overall Online Attention Toward Them. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW062 (April 2025), 25 pages. https://doi.org/10.1145/3710960

Authors' Contact Information: Manoel Horta Ribeiro, EPFL, Switzerland, manoel.hortaribeiro@epfl.ch; Shagun Jhaver, Rutgers University, USA, sj917@comminfo.rutgers.edu; Jordi Cluet-i-Martinell, EPFL, Switzerland, jordi.cluetimartinell@epfl.ch; Marie Reignier-Tayar, EPFL, Switzerland, marie.reignier@epfl.ch; Robert West, robert.west@epfl.ch, EPFL, Switzerland, robert.west@epfl.ch.

ACM 2573-0142/2025/4-ARTCSCW062
https://doi.org/10.1145/3710960

Going from benchmarks to the real world is hard

Examples:

- Self-driving cars
- COVID side projects

GenAI widens this crisis!

- No crisp yardstick
- Test set?

This article is more than **10 years old**

Self-driving cars: from 2020 you will become a permanent backseat driver

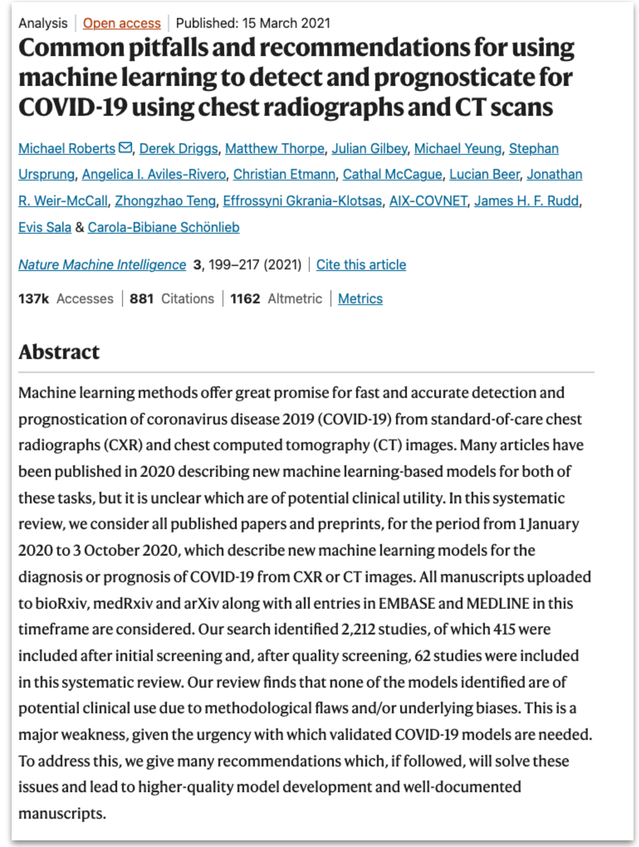
Driverless cars will revolutionise motoring, claim the manufacturers. But is the greatest danger that they will be too safe?

A BMW 'highly automated' prototype on the German autobahn. Photograph: PR

Despite High Hopes, Self-Driving Cars Are 'Way in the Future'

Ford and other companies say the industry overestimated the arrival of autonomous vehicles, which still struggle to anticipate what other drivers and pedestrians will do.

Share full article 523



Analysis | Open access | Published: 15 March 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence **3**, 199–217 (2021) | Cite this article

137k Accesses | **881** Citations | **1162** Altmetric | Metrics

Abstract

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

We can learn from D&D

Evals: “Apples-to-oranges”

People ought to think about

- What construct are you measuring?
- How is it operationalized?
- What are the sources of bias?

Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge

Hanna Wallach [1] Meera Desai [2] A. Feder Cooper [1] Angelina Wang [3] Chad Atalla [1] Solon Barocas [1] Su Lin Blodgett [1] Alexandra Chouldechova [1] Emily Corvi [1] P. Alex Dow [1] Jean Garcia-Gathright [1] Alexandra Olteanu [1] Nicholas Pangakis [1] Stefanie Reed [1] Emily Sheng [1] Dan Vann [1] Jennifer Wortman Vaughan [1] Matthew Vogel [1] Hannah Washington [1] Abigail Z. Jacobs [2]

Abstract

The measurement tasks involved in evaluating generative AI (GenAI) systems lack sufficient scientific rigor, leading to what has been described as “a tangle of sloppy tests [and] apples-to-oranges comparisons” (Roose, 2024). In this position paper, we argue that the ML community would benefit from learning from and drawing on the social sciences when developing and using measurement instruments for evaluating GenAI systems. Specifically, our position is that evaluating GenAI systems is a social science measurement challenge. We present a four-level framework, grounded in measurement theory from the social sciences, for measuring concepts related to the capabilities, behaviors, and impacts of GenAI systems. This framework has two important implications: First, it can broaden the expertise involved in evaluating GenAI systems by enabling stakeholders with different perspectives to participate in conceptual debates. Second, it brings rigor to both conceptual and operational debates by offering a set of lenses for interrogating validity.

1. Evaluating GenAI Systems

Evaluating a generative AI (GenAI) system[1]—i.e., making and justifying evaluative claims about that system—is critical for making decisions about whether it should be used for a particular purpose, whether it should be deployed in a particular context, or even whether it should be redesigned. The *process of evaluation*[2] necessarily requires information about the system’s capabilities (like its mathematical reasoning skills), behaviors (like regurgitating pieces of its training data), and impacts (like causing its users to feel harmed). Often, this information takes the form of *measurements* on nominal, ordinal, interval, and ratio scales (Hand, 2004), where each measurement reflects the amount of some *concept of interest* exhibited by that system (related to its capabilities, behaviors, or impacts) in some *context of interest*. Such measurements are obtained via the *process of measurement*, which uses *measurement instruments*[3] (e.g., datasets, classifiers, annotation guidelines, scoring rubrics, and aggregation functions) that instantiate a particular *measurement approach* (e.g., benchmarking, automated red teaming, real-world evaluations, and user studies).

Across academia, industry, and government (e.g., National Institute for Standards and Technology, 2024; Cooper et al., 2023; Perez et al., 2022; Weidinger et al., 2023), there is an increasing awareness that the measurement tasks involved in evaluating GenAI systems are more difficult than those involved in evaluating traditional ML systems. This is because GenAI systems accept a variety of inputs, produce diverse outputs, support a wide range of use cases, and have potential impacts on people and society that range from mundane to catastrophic. As a result, concepts related to the capabilities, behaviors, and impacts of GenAI systems—the concepts to be measured when evaluating GenAI systems—are often abstract and deeply intertwined with people and society. Abstract concepts cannot be directly measured and must therefore be indirectly measured from other observable phenomena. In addition, their meanings and understandings are often contested (e.g., Mulligan et al., 2016; 2019) across—and within—use cases, cultures, and languages.

Although ML researchers and practitioners have proposed myriad measurement instruments for evaluating GenAI

[1]Microsoft Research [2]University of Michigan [3]Stanford University. Correspondence to: Hanna Wallach <wallach@microsoft.com>.

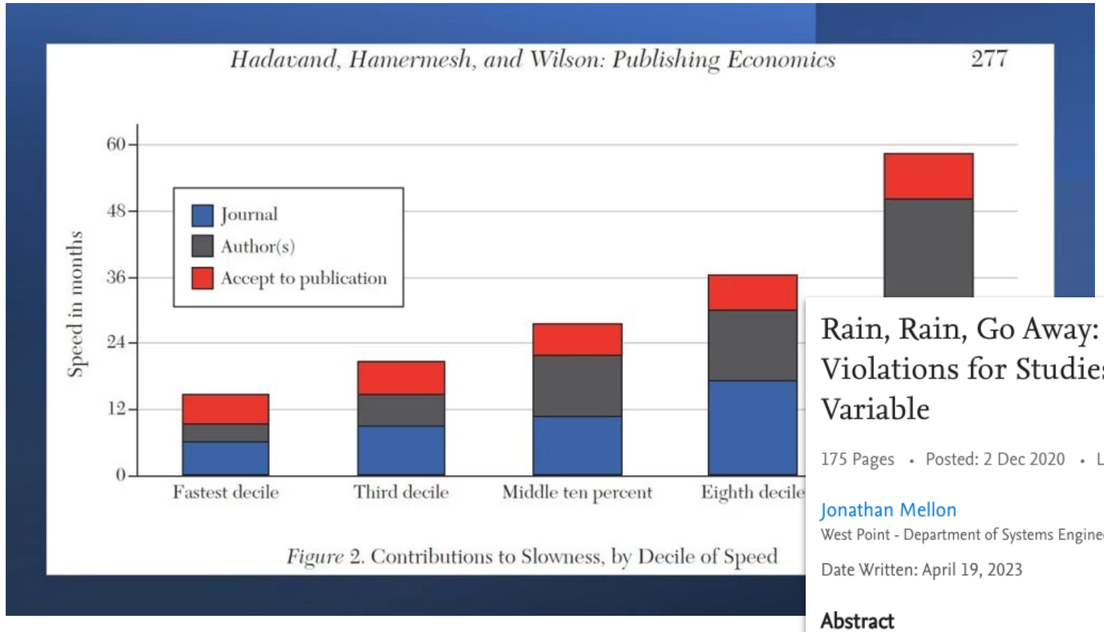
Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

[1]We use the term “GenAI system” to refer to either 1) a single GenAI model or 2) one or more integrated software components, where at least one component is an GenAI model. When we wish to refer to a single GenAI model, we use the term “GenAI model.”

[2]We provide definitions for italicized terms in Appendix A.

[3]We note that measurement instruments can be qualitative or quantitative, but must collectively result in measurements.

But we should also be wary of D&D's limitations



Hadavand, Hamermesh, and Wilson: Publishing Economics 277

Figure 2. Contributions to Slowness, by Decile of Speed

Rain, Rain, Go Away: 195 Potential Exclusion-Restrictio Violations for Studies Using Weather as an Instrument Variable

175 Pages • Posted: 2 Dec 2020 • Last revised: 10 Sep 2023

Jonathan Mellon
West Point - Department of Systems Engineering

Date Written: April 19, 2023

Abstract

Instrumental variable (IV) analysis assumes the instrument only affects the dependent variable via its relationship with the independent variable. Other possible causal routes from the IV to the dependent variable are exclusion-restriction violations and invalidate the instrument. Weather has been widely used as an instrumental variable in social science to predict many different variables. The use of weather to instrument different independent variables represents strong prima facie evidence of exclusion violations for all studies using weather IVs. A review of 289 studies reveals 195 variables previously linked to weather: all representing potential exclusion violations. Using sensitivity analysis, I show that the magnitude of many of these violations is sufficient to overturn numerous existing IV results. I conclude with practical steps to systematically review existing literature to identify possible exclusion violations when using IV designs.

Keywords: instrumental variable, IV regression, weather, rain, exclusion restriction

JQD:DM
Journal of Quantitative Description: Digital Media

We introduce the rationale for a new peer-reviewed scholarly journal, the *Journal of Quantitative Description: Digital Media*. The journal is intended to create a new venue for research on digital media and address several deficiencies in the current social science publishing landscape. First, descriptive research is undersupplied and undervalued. Second, research questions too often only reflect dominant theories and received wisdom. Third, journals are constrained by unnecessary boundaries defined by discipline, geography, and length. Fourth, peer review is inefficient and unnecessarily burdensome for both referees and authors. We outline the journal's scope and structure, which is open access, fee-free and relies on a Letter of Inquiry (LOI) model. Quantitative description can appeal to social scientists of all stripes and is a crucial methodology for understanding the continuing evolution of digital media and its relationship to important questions of interest to social scientists.

Why not do both?

MENU

APPETIZER (5 minutes)
Manoel's intro

1st COURSE (50 minutes)
Why do we need empirical reasoning?

2nd COURSE (30 minutes)
Course logistics

DESSERT (Rest of class)
Intros

What are we doing here?

A deep dive into causal inference and quantitative methods.

(Focused on things I deem particularly helpful)

Module #1: “Basics”

- **29 January (Today)**: Class intro
- **5 February**: Potential Outcomes
- **12 February**: Experiments
- **19 February**: Thinking with DAGs
- **26 February**: Regression (Part 1)
- **5 March**: Regression (Part 2)

Homeworks #1-2 (Individual)

Jupyter notebooks

- Experiments (Due Feb. 26th)
- Regression (Due Mar. 19th)

Readings

- Come prepared to have a discussion at the end of class!

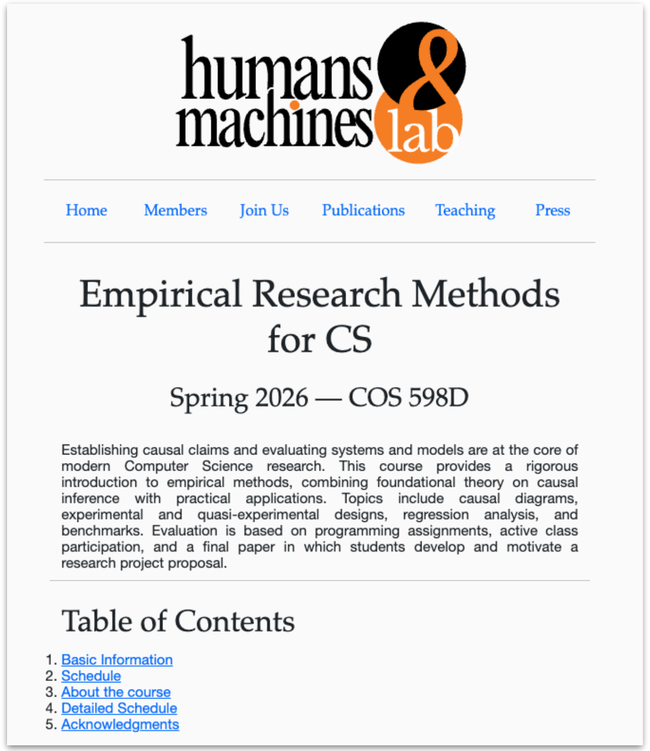
Module #2: Advanced Topics

- **19 March**: Benchmarks
  - Olawale Salaudeen (MIT)
- **26 March**: Quasi-experiments
  - Pietro Lesci (Cambridge)
- **02 April:** Causal ML
  - Drew Dimmery (Hertie School)
- **09 April:** Labeling with LLMs
  - Kristina Gligorić (Johns Hopkins)
- **16 April**: No classes (Manoel goes to CHI)
- **23 April: Final project presentations**

Homework #3:

- Project proposal
- “A paper up to the methods section”
- Experimental study / Benchmark / Observational study
- Deliverables (Due 23rd of April):
  - Presentation
  - Report (~4-5 pages)

Where to find details!



humans & machines lab

Home Members Join Us Publications Teaching Press

Empirical Research Methods for CS

Spring 2026 — COS 598D

Establishing causal claims and evaluating systems and models are at the core of modern Computer Science research. This course provides a rigorous introduction to empirical methods, combining foundational theory on causal inference with practical applications. Topics include causal diagrams, experimental and quasi-experimental designs, regression analysis, and benchmarks. Evaluation is based on programming assignments, active class participation, and a final paper in which students develop and motivate a research project proposal.

Table of Contents

https://humans.cs.princeton.edu/teaching/spring2026_empirical_methods.html

MENU

APPETIZER (5 minutes)
Manoel's intro

1st COURSE (50 minutes)
Why do we need empirical reasoning?

2nd COURSE (30 minutes)
Course logistics

DESSERT (Rest of class)
Intros