A Dozen Challenges in Causality and Causal Inference

arxiv.org/abs/2508.17099

[Part of the “Grand Challenges in Statistics” Series]

Avi Feller (UC Berkeley)

Guido Imbens
(Stanford)

Sara Magliacane
(Amsterdam)

Jose Zubizarreta
(Harvard)

Carlos Cinelli
(UW)

Edward Kennedy
(CMU)

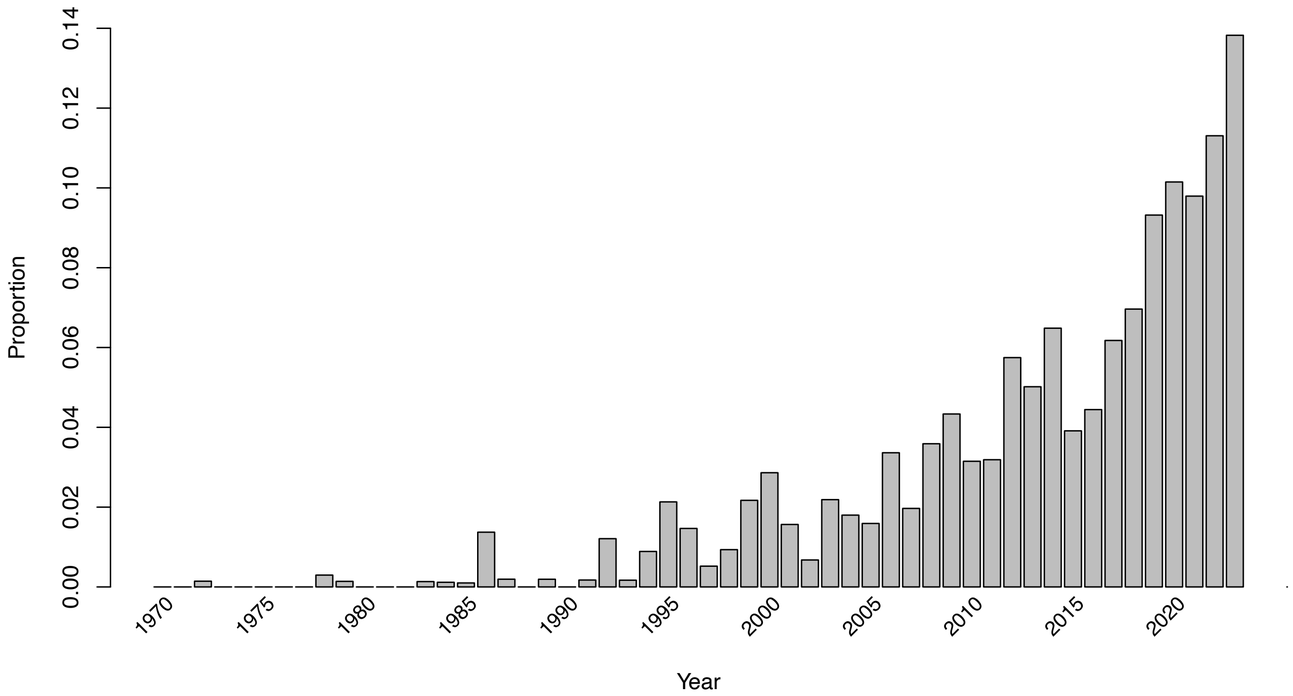
Challenges in Statistics: A Dozen Challenges in Causality and Causal Inference*

Carlos Cinelli, Avi Feller, Guido Imbens, Edward Kennedy, Sara Magliacane, Jose Zubizarreta

Abstract. Causality and causal inference have emerged as core research areas at the interface of modern statistics and domains including biomedical sciences, social sciences, computer science, and beyond. The field's inherently interdisciplinary nature—particularly the central role of incorporating domain knowledge—creates a rich and varied set of statistical challenges. Much progress has been made, especially in the last three decades, but there remain many open questions. Our goal in this discussion is to outline research directions and open problems we view as particularly promising for future work. Throughout we emphasize that advancing causal research requires a wide range of contributions, from novel theory and methodological innovations to improved software tools and closer engagement with domain scientists and practitioners.

Causality: So hot right now

% articles* with “causal” in title/abstract/keyword



* In leading statistics journals: *Annals of Statistics, Annals of Applied Statistics, Biometrics, Biometrika, Journal of the American Statistical Association, Journal of the Royal Statistical Society,* and *Statistical Science*

- Key topic in modern statistics, ML
- Multiple major conferences, journals
  - ACIC, CLeaR, EuroCIM
  - Journal of Causal Inference, Obs Studies
  - Society of Causality
- University research centers
  - Berkeley, Stanford, Penn, Harvard, …
- Mainstream stats + core grad curriculum

ACIC @ Berkeley

Largest ever!

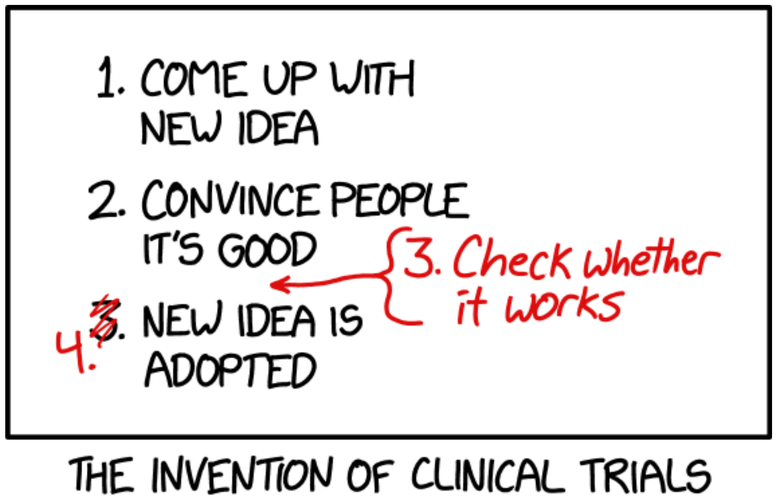
Guido Imbens

From the left: Elizabeth Stuart, Guido Imbens, Luke Keele, Fabrizia Mealli, Jas Sekhon. Pic by Avi Feller.

Mike Jordan

Causality: Some successes! 🚀

- The Randomized Trial [Fisher, Neyman, Bradford-Hill…]
  - Drug trials, policy experiments, A/B testing 🏭, …
  - Both what works and what does *not* work
- "Methodological arbitrage," clarifying assumptions
  - Common language to share ideas, best practices across fields
  - e.g., spread of instrumental variables, regression discontinuity
- Some celebrated analyses
  - Cholera [Snow, 1856], smoking + cancer [Cornfield et al., 1959]



Causality: Some failures 😞

- High-profile failures of clinical review
  - Failure to generalize: Thalidomide → birth defects [see Melchert & List, 2007]
- Excessive faith in strong, implausible assumptions
  - *Credibility crisis:* over-reliance on implausible assumptions [e.g., Leamer; Angrist & Pischke]
  - Real-world complications / implementation issues often dominate "causal evidence"
- Often prioritizes regulation over science
  - Causal inference methods used "for support rather than illumination" [see Recht, 2025]
- Exclude—rather than expand—communities of practice
  - e.g., dismissive of qualitative methods, case studies

Central role of binary, ignorable treatment

Some context: 1980s to 2010s

- Historically, causal inference research focused on binary treatment, ignorable given X
  - Rosenbaum and Rubin [1983] propensity score; Robins et al., [1996] doubly robust estimation; etc
  - Rich literature with novel methods, best practices, complications in this common scenario

- Hugely influence for empirical practice; clearly changed applied social science
  - Coupled with "credibility revolution" in economics, political science; trial emulation in biomedical sciences

- Over-emphasis on this case → slowed uptake of other methodological advances

A Dozen Challenges

A Dozen Challenges 

Experimentation: Better, faster, stronger

- Complex Experiments and Experimental Design
- Interference and Complex Systems
- Heterogeneous Effects and Policy Learning
- Mediation and Causal Mechanisms

Causal inference in the real world

- Sensitivity Analysis and Robustness
- Benchmarks, Evaluation, and Validation
- Aggregation and Synthesis of Causal Knowledge
- New Identification Strategies

Better technical tools, better computation

- Reliable and Scalable Causal Discovery
- Optimality and Minimaxity
- Automating the Causal Inference Pipeline
- Large Language Models and Causality

Cross-cutting issues 🚧

Cross-cutting issues:

- Bridging the gap between theory and practice
- Gains from incorporating ML and computational tools
- Importance of building an open and inclusive research community

↝ *Asking good questions*

Recurring theme: applications in industry 🏭

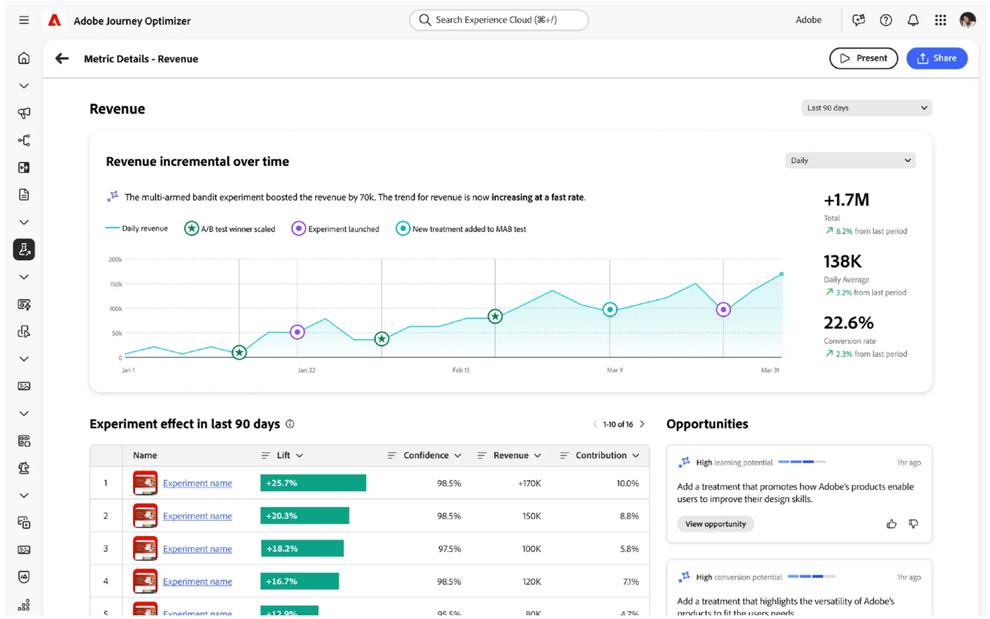
↝ *Key driver of causal inference research*

Experimentation:
Better, faster, stronger 💪

Complex experiments & experimental design

- Traditional experiments: large, simple, offline
  - Anchored in agricultural, drug trials; some industrial design

- Changing landscape, driven by new tech 🏭
  - Renaissance in online, adaptive experiments
  - Interacting, complex treatments (text/images)



- Many barriers to adoption 🚧
  - Despite huge literature, limited use of adaptive designs/bandits in practice (including industry)
  - Sequential treatments, increased adaptivity → better combine (valid) inference and decisions
  - Platform trials in health sciences: master protocol, multiple interventions
  - Automation and reducing frictions: AI-driven experimentation? 🤖

Interference and complex systems

- SUTVA / no interference → implausible in many cases
  - Classically: $Y_i(\mathbf{W}) = Y_i(\mathbf{W}')$ when $W_i = W_i'$ [Cox, 1958; Rubin, 1980]
  - Many important violations: peer effects; marketplaces; infectious diseases
  - Put the "social" in "social sciences"
- Rapid recent progress, esp around defining the problem
  - Key role of *exposure mappings*: low-dim treatment summary
  - *New designs:* bipartite graphs, two-sided marketplaces
- This is hard! 🚧
  - *Anna Karenina* problem: each case violates SUTVA in its own way
  - *Highly fractured literature:* need sufficiently general technical frameworks
  - Better incorporate substantive theory, richer data (incl. time)
  - Practical trade-offs, statistical power

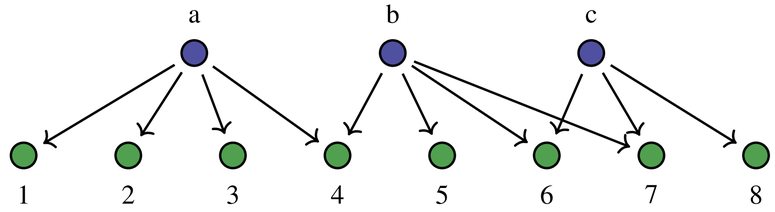


Fig 3: General bipartite interference graph with 3 interventional units (labeled a, b, c) and 8 outcome units (labeled 1, 2, 3, 4, 5, 6, 7, 8).

	Customer Experiment							Property Experiment					
Properties →	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Customers ↓														
1	C	C	C	C	C	C	C	C	T	C	T	C	C	T
2	C	C	C	C	C	C	C	C	T	C	T	C	C	T
3	T	T	T	T	T	T	T	T	T	C	T	C	C	T
4	C	C	C	C	C	C	C	C	T	C	T	C	C	T
5	T	T	T	T	T	T	T	T	T	C	T	C	C	T
6	T	T	T	T	T	T	T	T	T	C	T	C	C	T
7	C	C	C	C	C	C	C	C	T	C	T	C	C	T
8	T	T	T	T	T	T	T	T	T	C	T	C	C	T
9	T	T	T	T	T	T	T	T	T	C	T	C	C	T
10	C	C	C	C	C	C	C	C	T	C	T	C	C	T

Fig 4: Possible randomization for customer and properties in a marketplace, following Bajari et al. [2023].

Heterogeneous effects and policy learning

- Effect heterogeneity key for science, engineering
  - Low-dim CATE, $\mathbb{E}[Y(1) - Y(0) \mid X = x]$, or policy learning, $\max_\pi \mathbb{I}\{\text{CATE} > 0\}$
  - Estimating regression function ≠ estimating mean

- CausalML → explosion in new methods, esp for binary treatment
  - Incorporate flexible e.g., R/DR-Learner, (Bayesian) Causal Forests

- Hard to extend beyond base case 🚧
  - Beyond expectations (e.g., quantiles); beyond smoothness; beyond binary treatment
  - More challenging inference, optimization goals
  - Major gaps in implementation, translation → still limited impact on practice

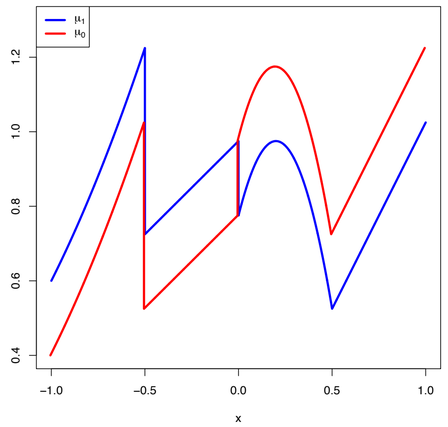
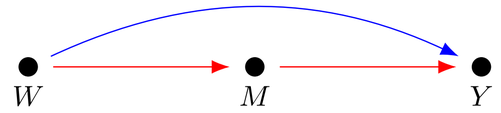


Fig 5: Example regression functions $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$ for which (i) the average causal effect $\mathbb{E}\{\mu_1(X) - \mu_0(X)\}$ is zero (e.g., if $X \sim \text{Unif}[-1, 1]$), (ii) the individual regression functions μ_a are non-smooth and difficult to estimate accurately, and (iii) the CATE $\mu_1(x) - \mu_0(x) = -0.2 \times \text{sign}(x)$ is piecewise constant and very simple.

Mediation and causal mechanisms

- Not just "what works" ⟶ *why* does it work?
  - Central question of "opening the black box" of causal relationships
  - Long history in many fields [Wright, 1921; Baron & Kenny, 1986; Pearl, 2001]
  - Long-standing debates, differences across causal subfields



- Recent unifications, extensions
  - Substantial work clarifying estimands, assumptions
  - Incorporate "modern" doubly robust / causal ML estimation

- Statistics: working to make mediation safe for economists 🚧
  - "Informal" mediation very common ⇝ many barriers to formal analyses
  - Develop practical tools, designs
  - Increasingly complex causal pathways, esp in GenAI era

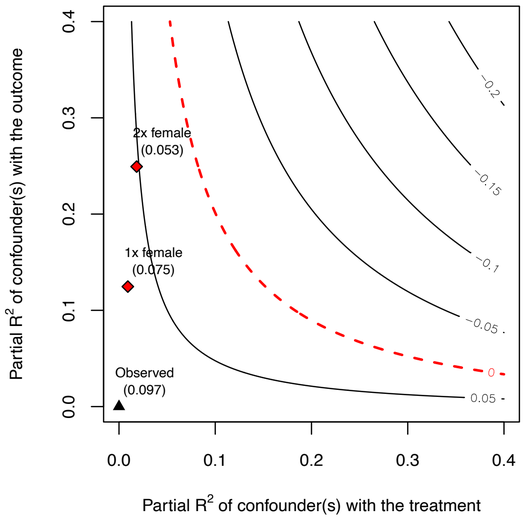
BY NBER SI METHODS LECTURE!

Causal inference in the real world

Sensitivity analysis and robustness

- Strong identifying assumptions in obs causal inference
  - Cornfield [1959]: Sensitivity analysis to argue smoking → cancer
  - Many modern frameworks; typically assume ignorability given X, U [Rosenbaum and Rubin, 1983; Rosenbaum, 1987; Robins, 1999]
  - Adapted to Bayes, CausalML; simplified regression frameworks [e.g., depend on partial R^2 for U]

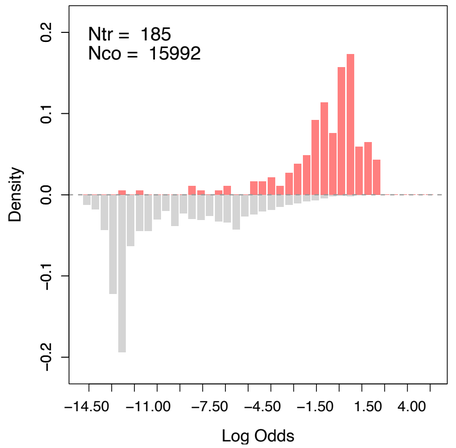
- Barriers to adoption: both technical and non-technical 🚧
  - Non-statistical reasons: sociological/norms; publishing incentives
  - But also: little consensus on statistical sensitivity/robustness frameworks
  - Remains challenging to calibrate sensitivity, incorporate subject knowledge



Benchmarks, evaluation, and validation

- Community benchmarks critical for rapid progress in ML
  - "Common task framework" [Donoho, 2023]; e.g., MNIST, CIFAR
- Limited benchmarks, coordination in causality research
  - Never observe ground truth: fundamental problem of causal inference
  - Experimental [LaLonde, 1986]; synthetic / semi-synthetic [ACIC challenge]
  - Causal discovery challenges: e.g., estimate protein signaling network in cells
- Need to **coordinate and incentivize progress** 🚧
  - Better benchmarks for high-dimensional, complex data
  - Better tailor evaluations/simulations to deployment, realism
  - Need novel validation, assessment in biological and social science applications
  - Improved incentives for this type of work and for subject matter experts

Fig 6: Assessing the Overlap in Lalonde-Dehejia-Wahba (LDW) Data



(a) LDW-CPS

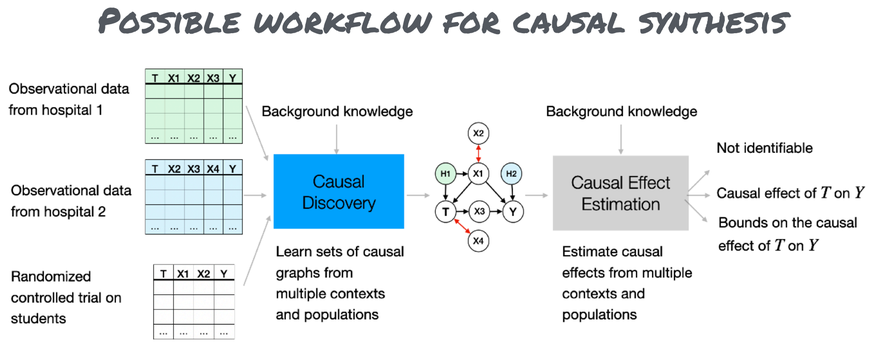
Note: Histograms depict the log odds ratios, i.e., $\log \frac{\hat{e}}{1-\hat{e}}$, using propensity score estimated through generalized random forest. The data are the observational CPS sample from the Lalonde-Dehejia-Wahba data.

Aggregating/synthesizing causal knowledge

- Beyond single study ⇝ aggregate causal knowledge
  - Fundamental to science—but field focused on single studies

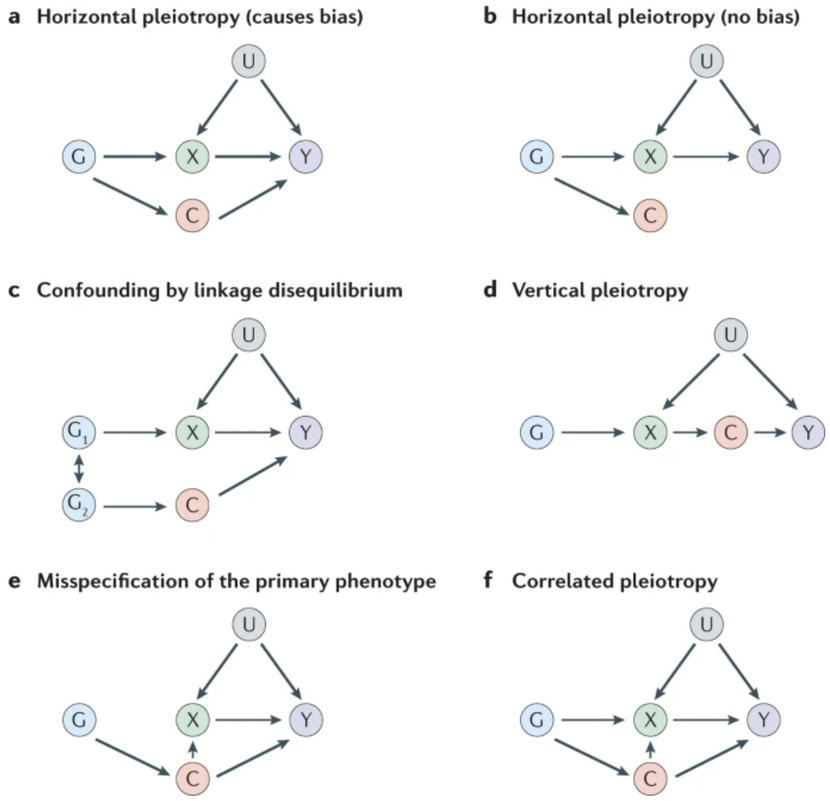
- Learning across studies typically *ad hoc*
  - Traditionally synthesize via "expert knowledge," meta-analysis
  - Difficult to weigh studies by strength of causal evidence
  - Recent work: surrogates and long-term effects; historical controls

- Hard to extend to many disparate studies 🚧
  - Analogy: completing a crossword puzzle [Rosenbaum, 2015]
  - Need for *causal* meta-analysis, causal data fusion, systems analysis
  - Key role for mechanisms and causal pathways



New identification strategies

- *Social science:* now standard **research designs**
  - e.g., RCTs, obs studies, diff-in-diff, IV, RDD [aka., my course!]
- Newer methods gaining popularity
  - Synthetic controls, Mendelian Randomization now more common
  - Graphical model-based strategies, e.g., front-door criterion
- Promising new directions
  - Active research on new panel data methods [see Arkhangelesky and Imbens, 2024]
  - Proximal methods
  - Shift-share IVs, bunching



Better computation,
Better technical tools

Reliable, scalable causal discovery

- Huge but distinct lit on causal discovery [Glymour et al., 2019]
  - e.g., *Protein signaling network:* want to predict and control how certain signals affect a cell. Can we learn causal structure from data?
  - *Promise:* Active field with many computational, theoretical advances
  - *Reality:* limited practical success (so far)

- Major barriers to adoption, deploying for new use cases 🚧
  - Current algos scale poorly, can be unreliable in complex systems
  - Goal of *local, targeted* causal discovery for only a part of a system
  - Better link causal discovery with "downstream" causal inference

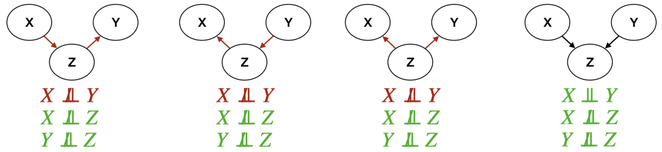


Fig 7: V-structure example: only the rightmost causal structure satisfies all the dependences and independences between the variables X, Y, Z. The statements in red are not correct in the original distribution.

Optimality, causal inference theory

- Causal inference needs strong theoretical foundations
  - *Minimax optimality:* what statistical structure are we willing to assume? Can we build estimators that work well under these assumptions? What is best possible performance?

    ⇝ natural framework to apply to causal inference, esp for semiparametrics [see Kennedy review]
  - Many other theoretical frameworks for understanding different estimators

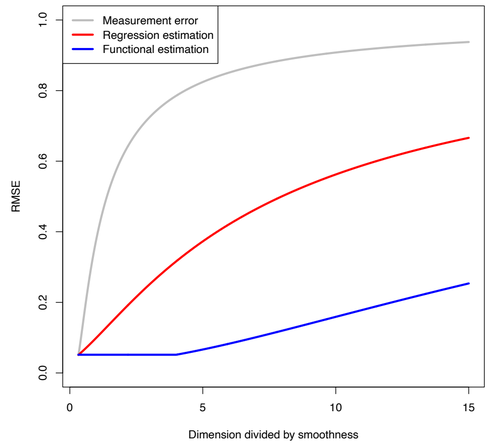
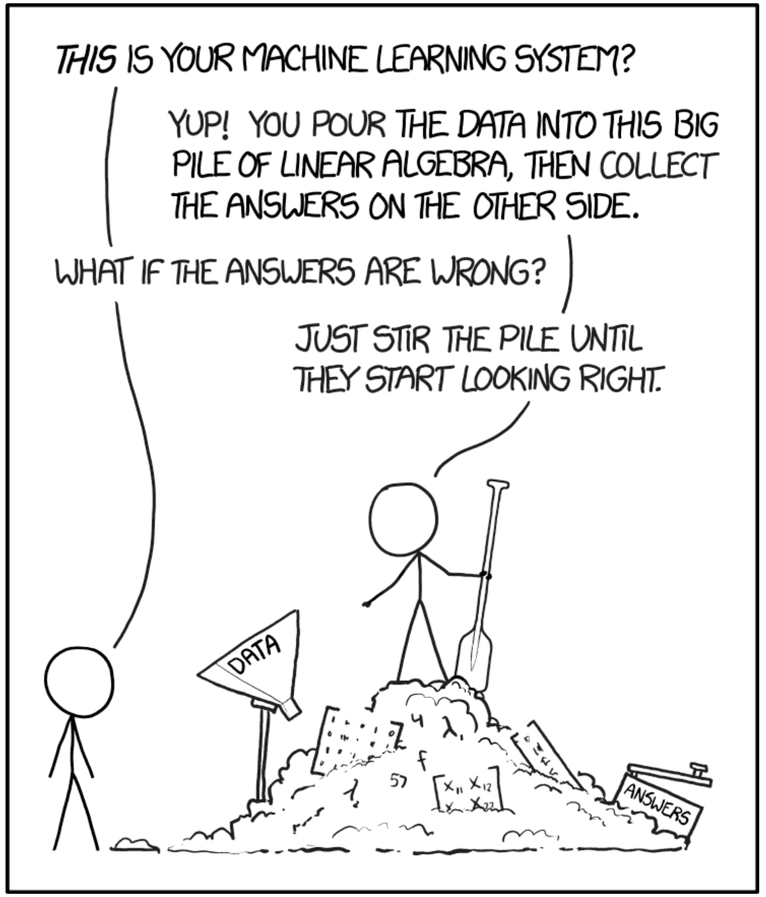


Fig 7: Minimax rates for three classical estimation tasks, as a function of covariate dimension d scaled by smoothness s: functional estimation (*e.g.,* estimating the ATE in causal inference); regression estimation (*e.g.,* the CATE); and density estimation with measurement error (*e.g.,* the counterfactual density when outcomes are measured with error). In causal analogues of these problems the rates further depend on the complexity of nuisance functions (*e.g.,* propensity scores, outcome regressions).

- Open questions 🚧
  - More complex models, relax smoothness assumptions; more complex targets
  - Adaptivity and performance for different function classes
  - New frameworks, including new losses, locality
  - Role of distributional results, regularity
  - Gap between theory and implementation

Automation 🤖

- New identification, theory typically require manual derivation
  - Deriving results for new estimator = PhD student dissertation chapter
  - Some automatic solutions for graphical models, causal discovery
  - Increasing role of numeric methods, esp constrained optimization
  - End-to-end causal inference still too challenging
- Many new challenges, opportunities! 🚧
  - Automatic construction of efficient estimators [e.g., Leudtke; Zhou et al.]; automatic debiasing [Chernozhukov et al]
  - Automatic identification beyond simple restrictions
  - Need for scalable approaches
  - Better knowledge representation and elicitation
  - Better software and GenAI tools



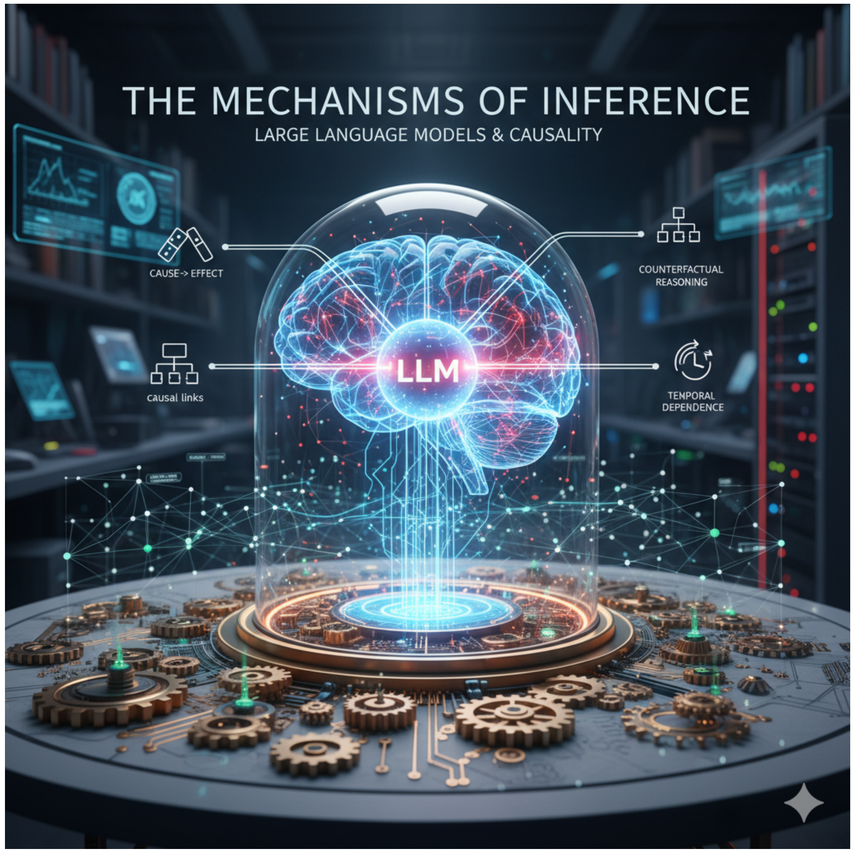
Large Language Models + Causality

- LLMs are apparently this hot new thing 🤷

- LLMs → Causal
  - **Incorporate complex data** into causal inference pipeline [building on long literature on text + causal]
  - **Synthetic experiments and agentic models:** combine real-world experiments with simulated experiments run with AI agents
  - **LLMs and causal reasoning:** how do we know?
  - **AI co-pilots** for causal inference

- Causal → LLMs
  - Frameworks for developing **trustworthy and reliable** AI
  - **Evaluating and monitoring deployed AI systems**

Generated with Gemini



Cross-cutting issues

Cross-cutting themes

- Bridge gap between theory and practice
  - Many opportunities to use theory to improve practice *and vice-versa*
  - Critical need to engage with substantive experts; better software, diagnostics, guidance/outreach
  - Exciting new approaches that marry structural and “statistical” techniques

- Gains from incorporating ML/better computation
  - Starting in ~2010s: important rise in CausalML, flexible prediction tools 🏭
  - Today: similar opportunity for AI/LLMs? 🤖

- Make causality research an inclusive, “big tent” community
  - Historically fractured research communities
  - Today: Much more ecumenical, open → better research!

Cross-cutting themes

The central challenge in causality research is to continue to **ask good questions**.