JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Using Causal Inference To Make Sense of Messy Data

Ilya Shpitser

John C. Malone Assistant Professor of Computer Science

Malone Center for Engineering in Healthcare

The Johns Hopkins University

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

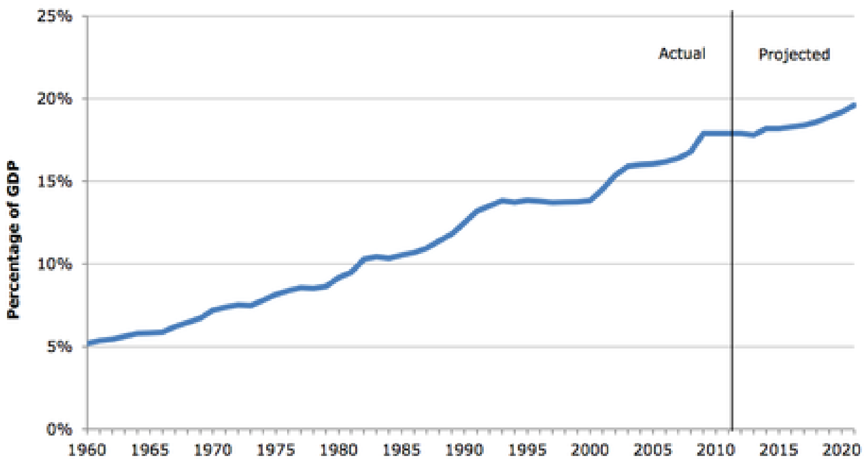
Health Care: Costs

Absolute expenditures – $3.0 trillion 17.5% GDP (2014)

Relative expenditures **–** 50% increase in past 10 years

Potential efficiency gains – $750 billion (2009) – more than 25% of the total

Figure 2: U.S. National Health Expenditures as a Share of GDP, 1960-2021



Source: Centers for Medicare and Medicaid Services.

From "Best Care At Lower Costs: The Path to Continuously Learning Health Care in America" Institute of Medicine, 2012

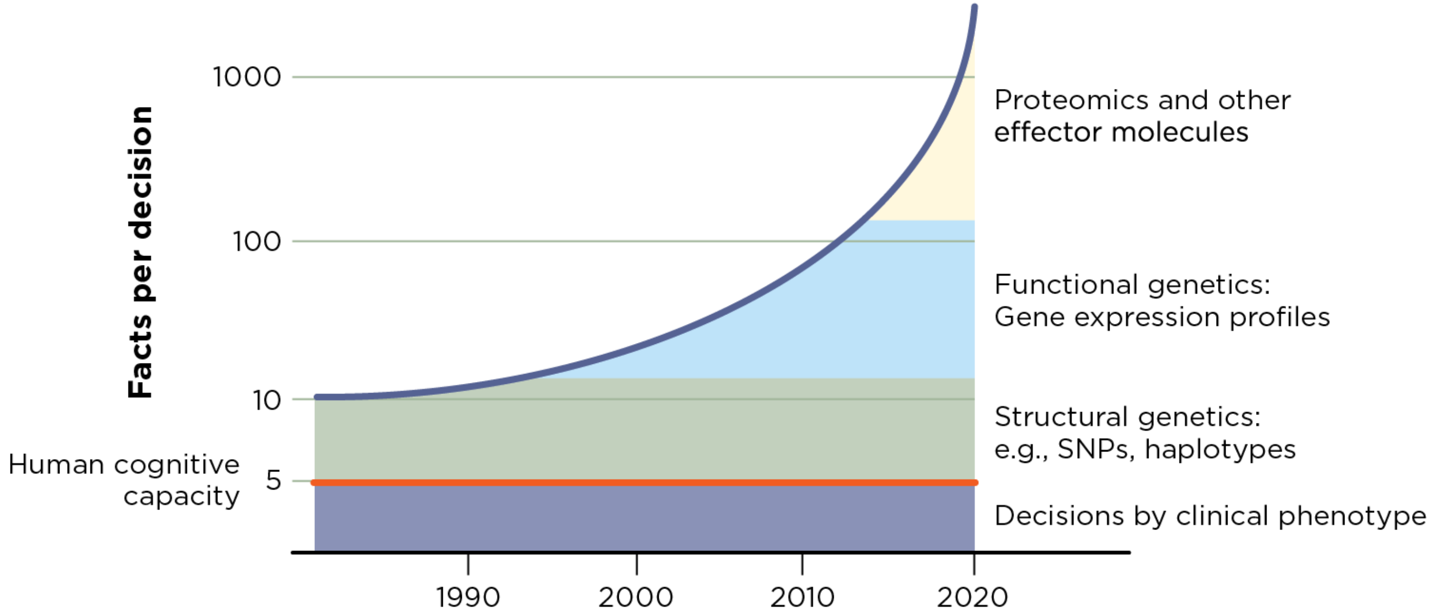
Health Care: Complexity

More conditions – e.g. 79 year old patient with 19 meds per day

More clinicians – e.g. 200 other doctors treating patients of a single primary care doctor

More choices – e.g. hundreds of diagnostic factors; dozens of treatments

More activities – e.g. ICU clinicians with 180 activities per day



From "Best Care At Lower Costs: The Path to Continuously Learning Health Care in America" Institute of Medicine, 2012

Malone Center Mission:

To catalyze and accelerate the development, translation, and deployment of research-based innovations that advance the effectiveness and efficiency of health care.

Smart Devices and Systems for Healthcare
creating devices and information analytics that enhance care in the clinical environment

Modeling and Optimization for Healthcare Delivery
exploiting traditional and new sources of data to enhance the efficiency and quality of healthcare

Mobile Health and Healthy Living
developing innovations that support individuals outside traditional care environments, that enhance health in everyday life, and that augment traditional health care approaches

My Work at the Malone Center

- **Science from biased data**
  - Poor treatment outcomes: bad treatment, poor adherence, confounding?
- **Decision support**
  - Treatment decisions are a complex combination of medical training and institutional knowledge.
  - Can we use learning algorithms to help?
- **Dealing with missing data**
  - Most datasets in practice have systematically missing entries. This creates **bias** if not properly handled.
  - How do we handle complex missing data?

Science from biased data

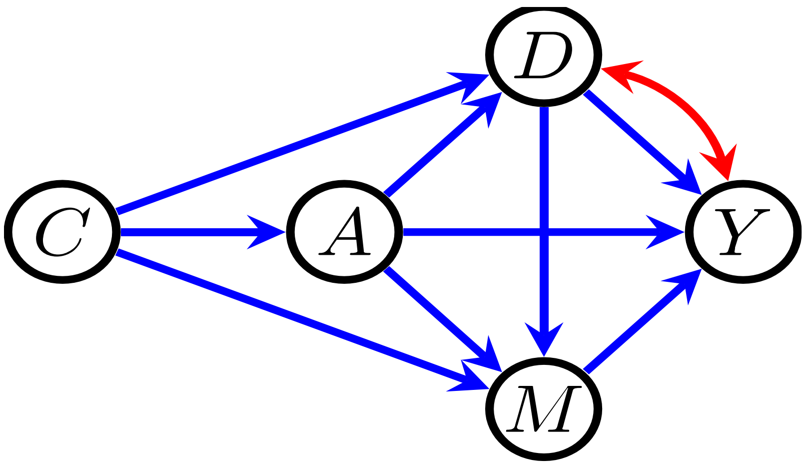
- Better healthcare means making better decisions
- Decisions are about causal efficacy
- Randomized controlled trial data often not available
- Practical data: confounding bias, selection bias, missing data, measurement error
- The field of causal inference aims to provide answers in this challenging setting

Adherence in HIV Patients

- Setting: longitudinal observational studies of HIV patients (PEPFAR program).
- Outcome: viral failure, treatments are anti-retroviral therapies
- Question: how are outcomes affected by:
  - Poor drug choice, or
  - Poor adherence
- Formally, adherence **mediates** (all?, some?) of the effect of the drug.

Adherence as a causal problem

- What causes virological failure in patients?
- A single slice of a longitudinal study:



- C (age, gender, etc.) A (HIV drug), D (white blood cell #), M (% pills taken), Y (outcome)
- Lots of reasons Y might be poor!

Predicting the hypothetical

- Every patient was on some drug, had some toxicity, and some adherence level.
- What would have happened to their outcome
  - If toxicity were low?
  - If adherence were high?
- RCTs possible for this, but expensive, lengthy.
- Alternative approach for existing, messy data:
  - Fit observed data models
  - Combine in a particular way to **mimic** the right RCT.
- Hard in general due to confounding, selection bias.

Predictions under counterfactual adherence

- How would less effective treatment do if adherence was of more effective treatment?

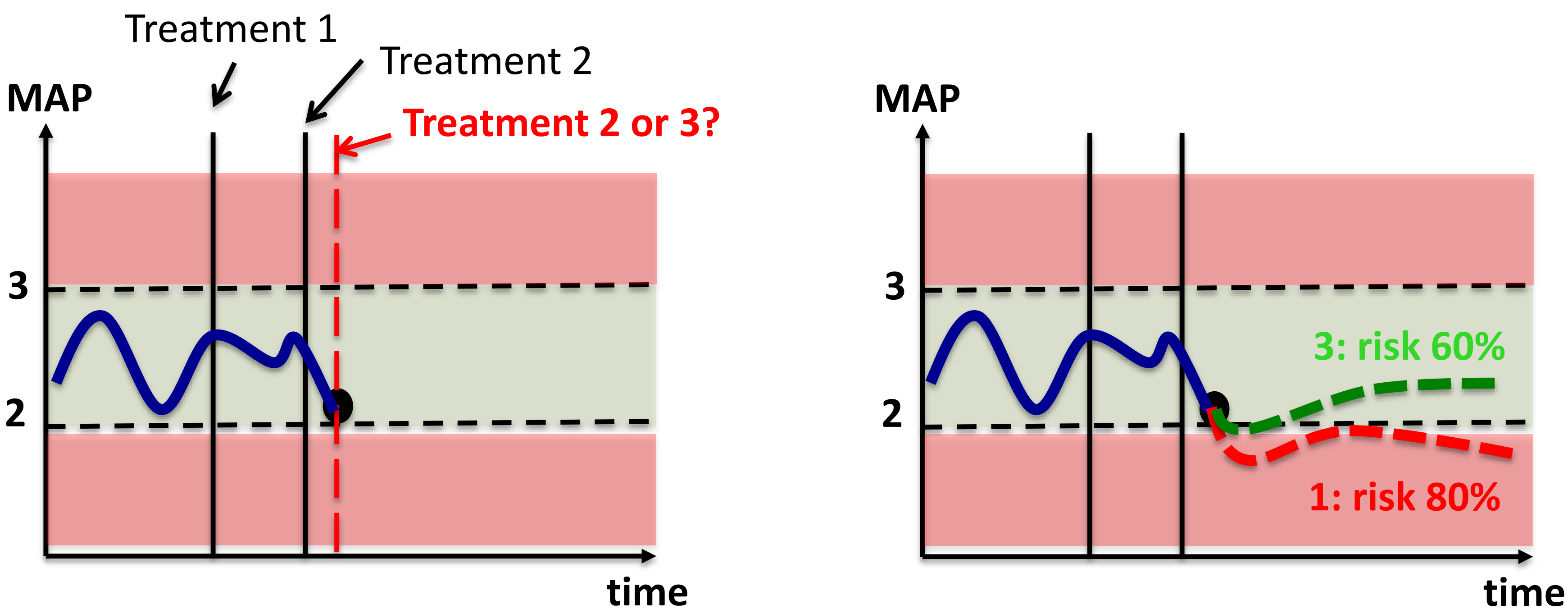
		Baseline treatment			
		2	3	4	5
Comparison treatments	1	0.412*	0.210	-0.059	-0.068*
	2	-	0.132	-0.495*	-0.198*
	3	-	-	-0.566*	-0.135*
	4	-	-	-	-0.027*

- Most significant effects negative. Meaning:
- More effective treatments are "harder to take."
- Effectiveness driven by biochemistry, not adherence.

Clinical decision support

- Exploiting patterns in complex data is difficult for (unaided) humans, even very experienced clinically.
- Naïve analysis can be misleading
  - Example: in crashing sepsis patients, treatment is associated with worse outcomes.
- Wanted: a tool that can output counterfactual outcomes at a complex decision point

Decision Support

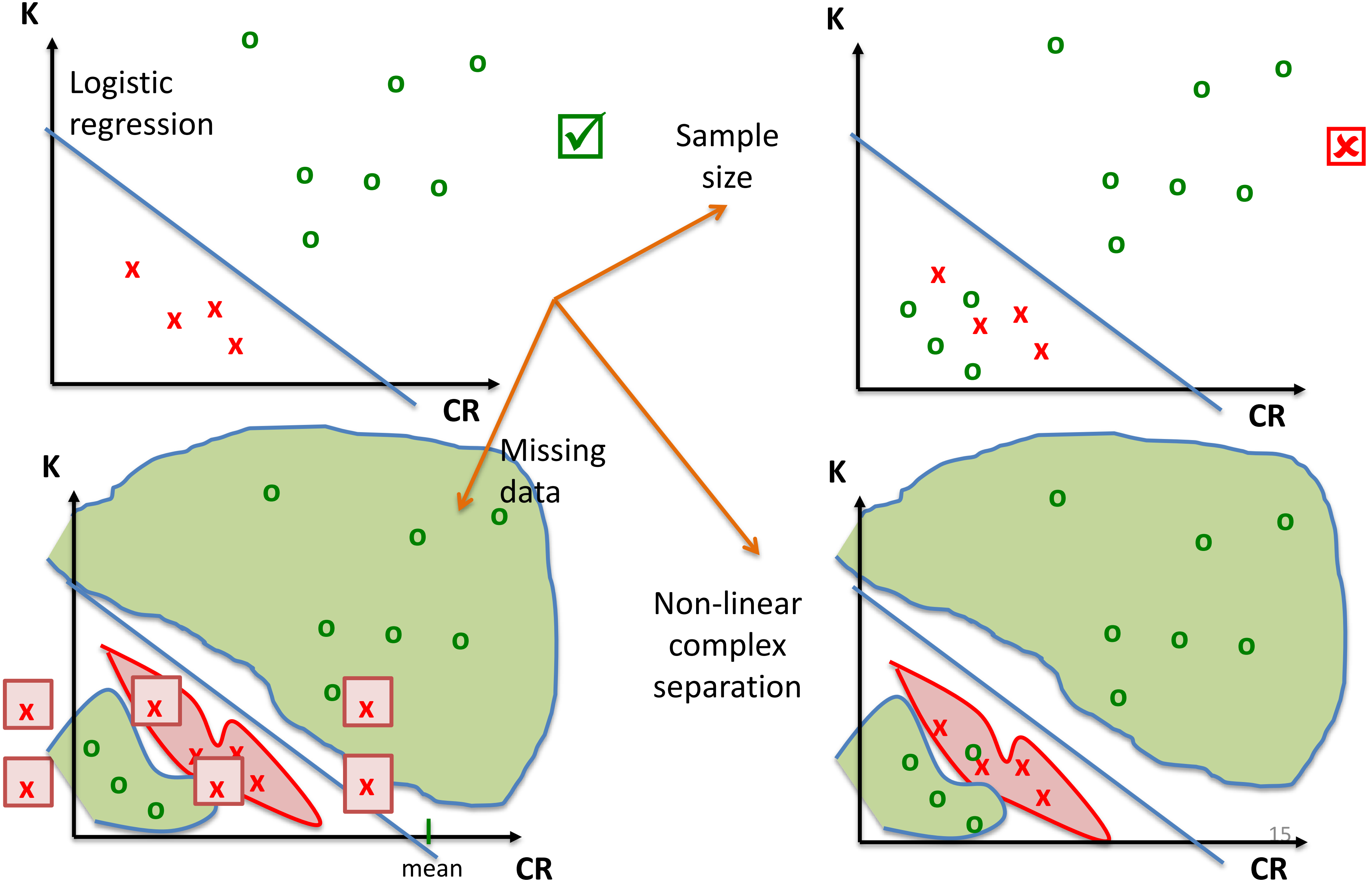


Causal inference methods exist for predicting counterfactual outcomes based on factual data. Work in progress (with Suchi Saria's group) on learning treatment policies.

Missing Data

- Ubiquitous problem.
- Often handled poorly.
- Possibility of severe bias (example):
  - HIV prevalence in Zambia Demographic and Health Survey
  - Sick people (severely) underreport
  - Complete case analysis underestimates prevalence by as much as 10%.

Dealing With Missing Data



New methods for missing data

- Most complex setting is data missing not at random (MNAR)
  - People don't report sexual history **due to** that history.
  - Voting intent, intermittent dropout, etc.
- Easier settings: reweigh observed cases based on typicality (recent NYT article on polls about this)
- Developed new extension of this to MNAR data.
- More generally: work on a complete theory of when missing data is a solvable problem.

Selected projects

- **Decision support in the ICU (with Suchi Saria and Katie Henry)**
- **Next generation methods for data missing not at random (with Eric Tchetgen Tchetgen and James Robins)**
- **Mediation analysis for understanding adherence in HIV studies (with Eric Tchetgen Tchetgen and Phyllis Kanki)**
- **Mediation analysis for study of radiation side effects (with Todd McNutt and the Oncospace Consortium)**

THANK YOU!

Ilya Shpitser (ilyas@cs.jhu.edu)