

Causal Inference and the Data-Fusion Problem

Elias Bareinboim

eb@purdue.edu

(Joint work with J. Pearl)

Symposium on Accelerating Science:

A Grand Challenge for AI

November, 2016

Science News

One Drink Of Red Wine Drinks Are Stressful

Feb. 13, 2008 — One drink of alcohol slightly but the positive disappear with Peter Munk Ca Hospital.

from universities, journals, and other research organizations

... To Circulation, But Two

The NEW ENGLAND JOURNAL of MEDICINE



Association of Nut Consumption with Total and Cause-Specific Mortality

Original Article
Jing Bao, M.D., Ph.D., Jiell Han, Ph.D., Frank B. Hu, M.D., Ph.D., Edward L. Giovannucci, M.D., Sc.D., Walter C. Willett, M.D., Dr.P.H., and Charles S. Fuchs, M.D., M.P.H.
N Engl J Med 2013; 369:2001-2011 | November 21, 2013 DOI: 10.1056/NEJMoa1307652

Abstract

BACKGROUND

Increased nut consumption has been associated with a reduced risk of major chronic diseases, including cardiovascular disease and type 2 diabetes mellitus. However, the association between nut consumption and mortality remains unclear.

METHODS

We examined the association between nut consumption and

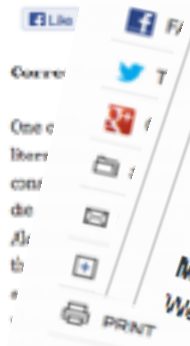


MEDIA IN THIS ARTICLE
Video



Why I...

One of the...
By John C...



August 19
08:00 PM ET

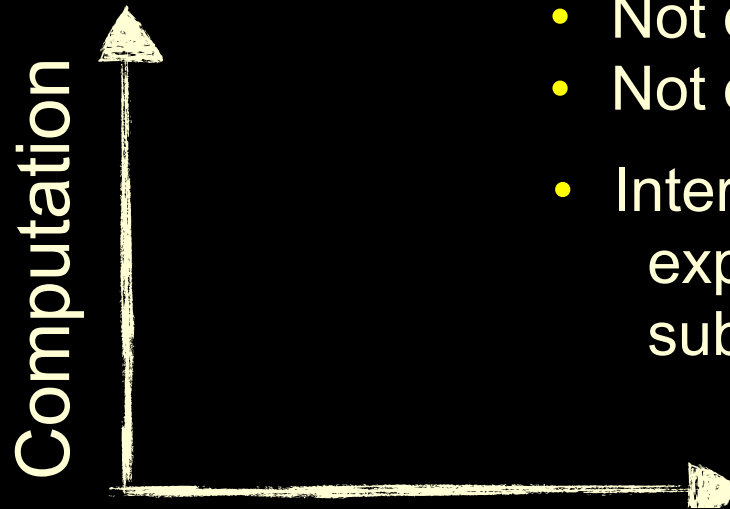
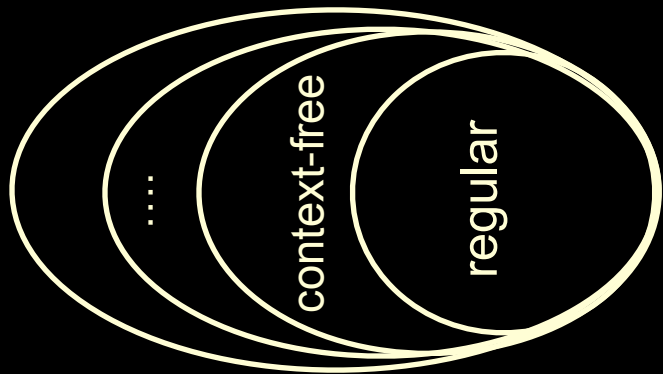
4 Comments / Shares

WHAT'S GOING ON HERE?

Language of Science

- Perhaps surprising to some...

Chomsky Hierarchy

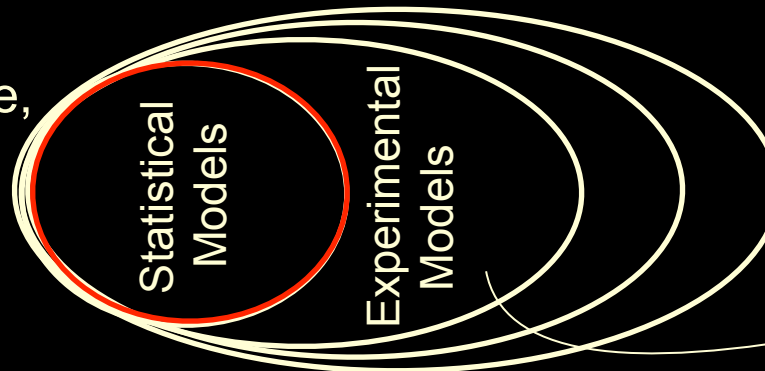


- Not only computational
- Not only sample size
- Interplay of observations, experiments, and substantive knowledge

Data collection (not # samples)

The Grammar of Science,
Pearson, 1892

Noise, uncertainty,
and variability.



tension between
layers..

The Design of Experiments
Fisher, 1935

GOAL

- Develop machinery (language, conditions, and algorithms) for performing two tasks:

1. Learning about **population-level causal effects** by cohesively combining multiple heterogeneous datasets.

Bareinboim, Pearl. Causal Inference and the Data-Fusion Problem. PNAS'16.

2. Deciding **individual-level treatments** by leveraging population-level fused data.

Bareinboim, Forney, Pearl. Bandits with Unobserved Confounders. NIPS'15.

BIG PICTURE

(“All data is not created equal”)

- **Heterogenous datasets** are pervasive in the empirical sciences since the data is collected:
 - (1) under different **experimental conditions**,
 - (2) the underlying **populations** are different,
 - (3) the **sampling procedure** is not random,
 - (4) the **treatment assignment** is not random,
 - (5) many variables are **not measured**.
- All these dimensions are now formalized.
- And there are conditions and algorithms to decide what is “entailed” from a certain data collection.

MOTIVATION FOR DATA-FUSION

Target population Π^*

Query of interest $Q = P^*(y \mid do(x))$

(a) **US**

Census data
available

(b) **New York**

Survey data
resembling target

(c) **Los Angeles**

Survey data
younger population

(d) **Boston**

Age not recorded
Mostly successful
lawyers

(e) **San Francisco**

High post-treatment
blood pressure

(f) **Texas**

Mostly Spanish
subjects
High attrition

(g) **Arkansas**

Randomized trial
College students

(h) **Utah**

RCT, paid
volunteers, mainly
unemployed

(i) **Wyoming**

Natural experiment
young athletes

HETEROGENEOUS DATASETS



d ₁	Population	Los Angeles	New York	Texas
d ₂	Obs. / Exp.	Experimental	Observational	Experimental
	Treat. Assign.	Randomized Z ₁	-	Randomized Z ₂
d ₃	Sampling	Selection on Age	Selection on SES	-
d ₄	Measured	X ₁ , Z ₁ , W, M, Y ₁	X ₁ , X ₂ , Z ₁ , N, Y ₂	X ₂ , Z ₁ , W, L, M, Y ₁

DATA-FUSION TASKS

Description of each dataset: tuple (d_1, d_2, d_3, d_4)
(population, obs./exp., sampling, measure.)

Dimensions

1. *Causal Inference* (observational studies)
 $(d_1, \text{Observ.}, d_3, d_4) \rightarrow (d_1, \text{Experiment}(X), d_3, d_4)$
2. *Sampling Selection Bias*
 $(d_1, d_2, \text{Select}(\text{Salary}), d_4) \rightarrow (d_1, d_2, \{\}, d_4)$
3. *Transportability* (External Validity)
 $(\text{Bonobos}, d_2, d_3, d_4) \rightarrow (\text{Humans}, d_2, d_3, d_4)$

Data-fusion:

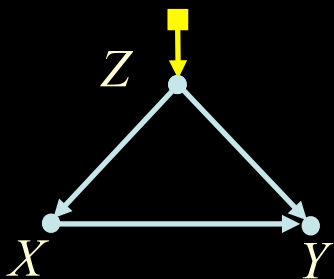
$$\{(d_1, d_2, d_3, d_4)\} \rightarrow (d'_1, d'_2, d'_3, d'_4)$$

BIG PICTURE

1 query

$$Q = P^*(y \mid \text{do}(x))$$

2 model



3 data

$$P^*(x, y, z) + P(y \mid \text{do}(x), z)$$

With the current scientific knowledge about the problem (2) and the available data (3), is it possible to answer the research question (1)?

inference engine

solution

(yes / no)

$$P^*(y \mid \text{do}(x)) =$$

$$\sum_z P(y \mid \text{do}(x), z) P^*(z)$$

DEMO

CONCLUSIONS

- Data-fusion from big data requires encoding of structural features of the data-generating model.
 - Even when the ‘gold standard’ (RCTs) is available, it is still not direct to compute effects of interventions.
- There are necessary and sufficient conditions (and algorithms) that fully characterize sampling selection bias and transportability (non-parametrically).
- Principled framework for data-fusion — pooling and aggregating observational and experimental information spread throughout heterogeneous domains for population- and individual-level causal inference.