CompSci 590.01

Causal Inference in Data Analysis with Applications to Fairness and Explanations

Lecture 1: Overview and Introduction to Causal Inference

Sudeepa Roy

Welcome to

CompSci 590.01:
Causal Inference in Data Analysis
with Applications to
Fairness and Explanations

About me…

- Instructor: Sudeepa Roy
- Associate Professor of Computer Science
- At Duke CS since Fall 2015
- PhD. UPenn, Postdoc: U. of Washington
- Member of two research groups:
- Almost Matching Exactly (AME) lab for causal inference
- "Duke Database Devils" a.k.a. the database research group
- Research interests: Data management and analysis, causality and explanations, data repair, query optimization, database theory, uncertain data, fairness and responsible data science…
- I teach CompSci 316 (undergrad database class) and CompSci 516 (grad database class)

Logistics

Important:
Always feel free to share feedback, stop me and ask questions, ask me to go slower/faster and repeat

No questions are too simple to ask

Logistics

- Course website: https://courses.cs.duke.edu/spring23/compsci590.1/
- Communications through Ed & Sakai
- Prerequisite: No hard prerequisites, but we will use and assume knowledge in CS topics such as probability theory, graphs, algorithms and data structures, some knowledge in machine learning/AI/databases, basic programming
    - It is okay if you do not know some of these but are willing to learn as needed – “willingness to learn” is the key requirement in this course ☺
- Slides will be uploaded after the lectures so that we can brainstorm during the lectures and keep them interactive

Grading

- There are no exams – your active participation, learning, presentation, and project decide your grade.

Class participation (15%):

- This includes both attending lectures and frequent participation in classes including presentations led by other students. If you think you might miss more than 3 classes during the semester, talk to the instructor early.

Grading

Assignments (15%):

- There will be a small number of (2-3) assignments during the semester, and depending on the assignments, we may have peer grading supervised by the instructor.

Grading

Presentation and leading discussion of a research topic (25%):

- We will post a list of potential research papers and topics. You can select a topic and 1-2 important papers on that topic to present and lead the discussion of in a class.
- Depending on the number of students enrolled and their interests, and number of important papers on that topic, it may be done in small groups of 1-2 students. Some topics may require > 1 presentations.
- Feel free to choose a topic related to your class project.
- Students are expected to cover the basics before presenting the research paper -- e.g., if you choose the topic "explainability of GNN", you should first give an overview of GNN.
- Note that all students are expected to read the papers and participate in the discussions, not only the students who are presenting/leading the discussions.

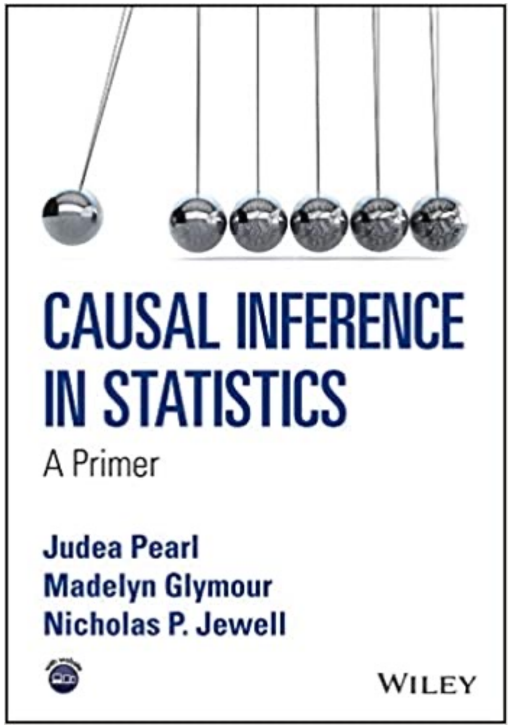
Grading

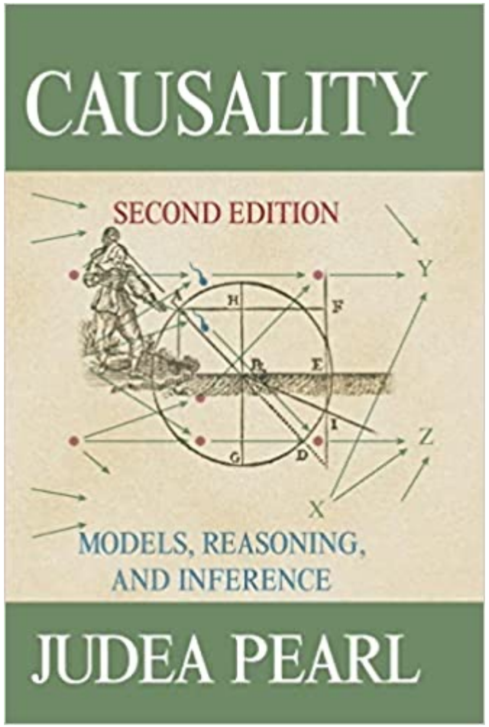
Class project (45%):

- There will be a semester-long class project on a topic of your interest and relevant to the class in small groups of 2-3 students. We will post some possible topics.
- It can be
  - an open-ended research project that can potentially be a paper (you are encouraged to do so, especially if you are a PhD student or an MS/undergraduate student considering doing a PhD later - your effort decides the grade not the end results),
  - implementation and analysis of algorithms,
  - building a tool with GUI for an application related to causal inference, or
  - analyzing real and synthetic datasets for a problem and showing the techniques & results.
- Projects focusing on only reading papers/writing surveys are discouraged.
- There will be three checkpoints - an initial proposal, midterm update, and final report, and you are also encouraged to meet the instructor briefly and every few weeks.
- There will be a short in-class presentation at the end.
- Project grades will take into account your efforts/ results, and quality of related work survey, presentation, and final report

Reading material

- We will read lots of research papers!
- For the first several lectures on Pearl's graphical causal model, we will follow these two books -- there are several research papers, survey articles online

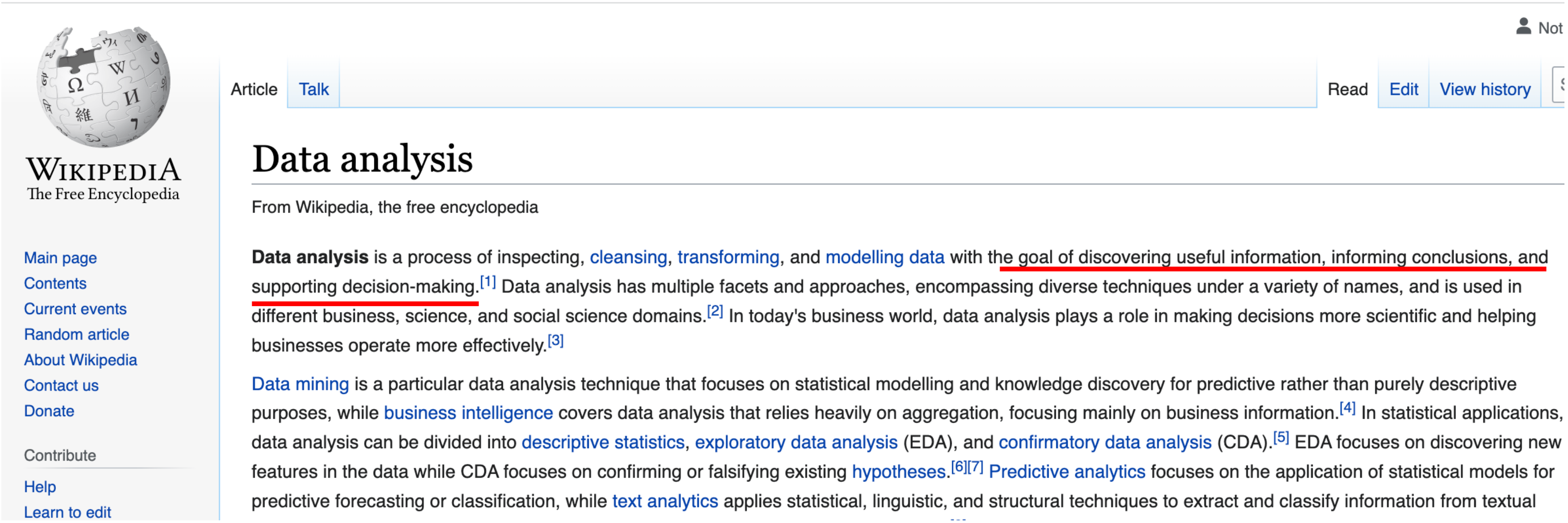




Causal Inference in Data Analysis with Applications to Fairness and Explanations

- What is “Data Analysis”?
- What is “Causal Inference”?
- What is “Fairness”?
- What is “Explanations”?

What is Data Analysis?



Data analysis

From Wikipedia, the free encyclopedia

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.[1] Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains.[2] In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.[3]

Data mining is a particular data analysis technique that focuses on statistical modelling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information.[4] In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA).[5] EDA focuses on discovering new features in the data while CDA focuses on confirming or falsifying existing hypotheses.[6][7] Predictive analytics focuses on the application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual

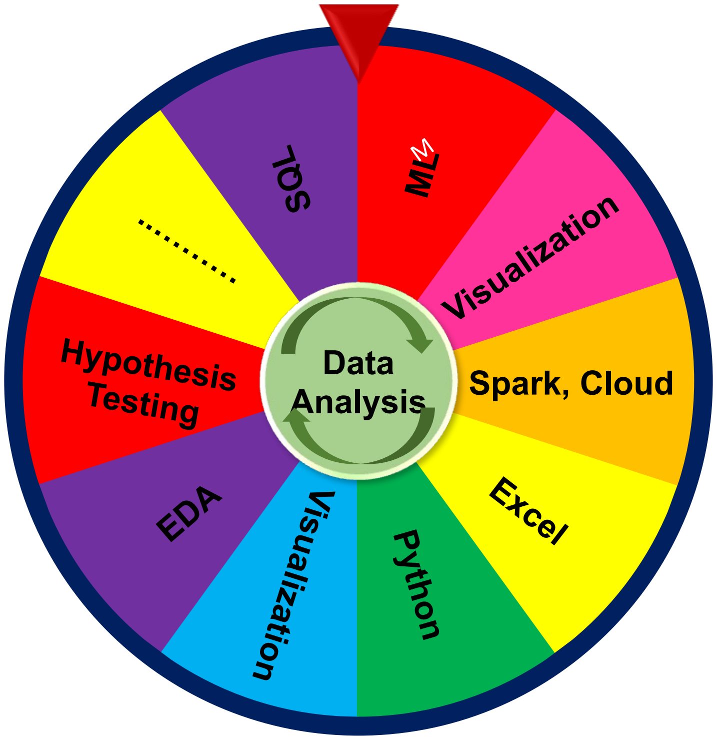
Discover useful information

✓ Inform conclusions

Support decision-making

The high-level goals of data analysis…

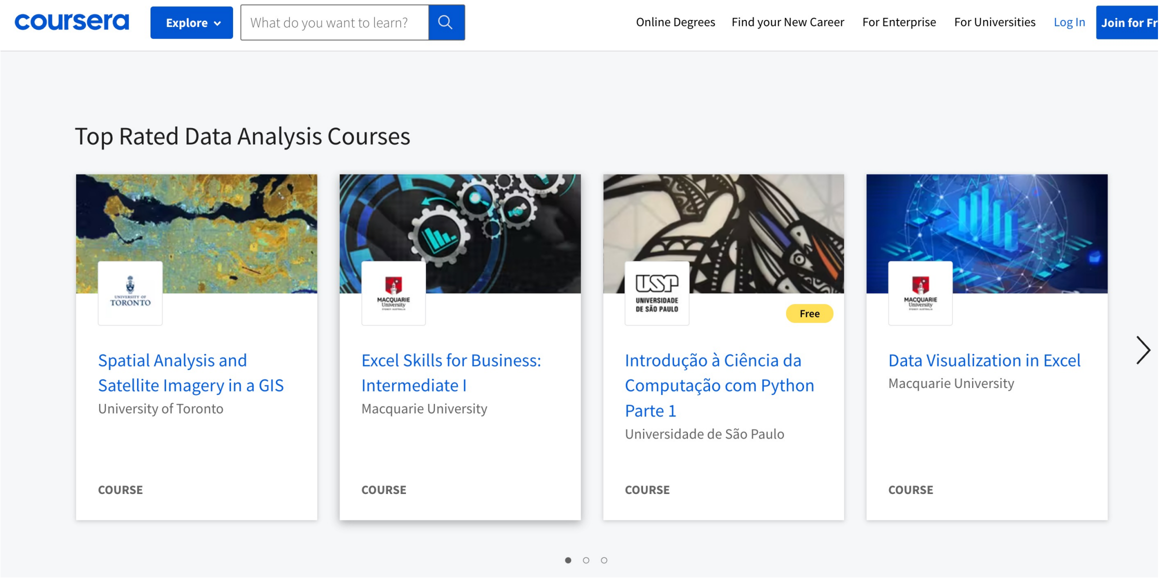
Several tools and techniques for Data Analysis...



Discover useful information

Inform conclusions

Support decision-making



Top Rated Data Analysis Courses

Spatial Analysis and Satellite Imagery in a GIS
University of Toronto
COURSE

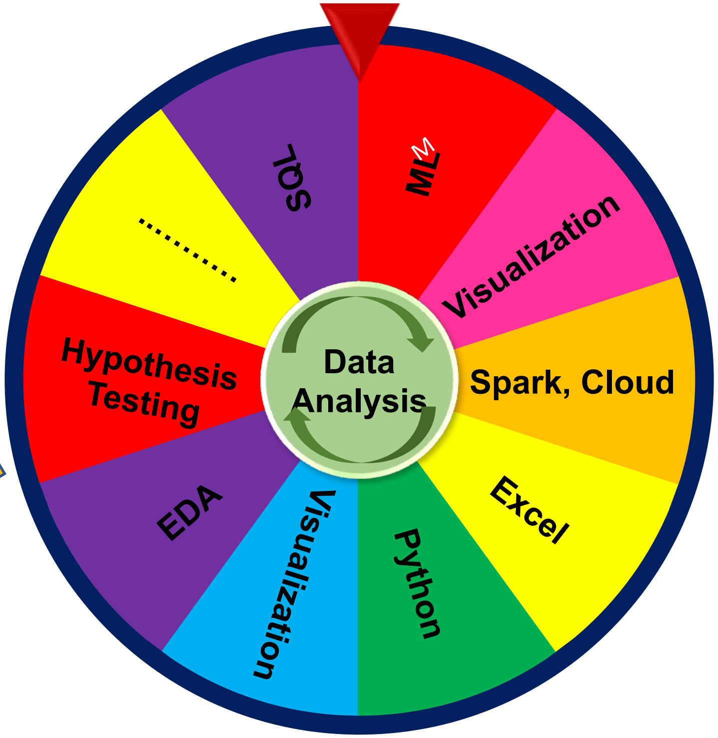
Excel Skills for Business: Intermediate I
Macquarie University
COURSE

Introdução à Ciência da Computação com Python Parte 1
Universidade de São Paulo
COURSE

Data Visualization in Excel
Macquarie University
COURSE

Several tools and techniques for Data Analysis...

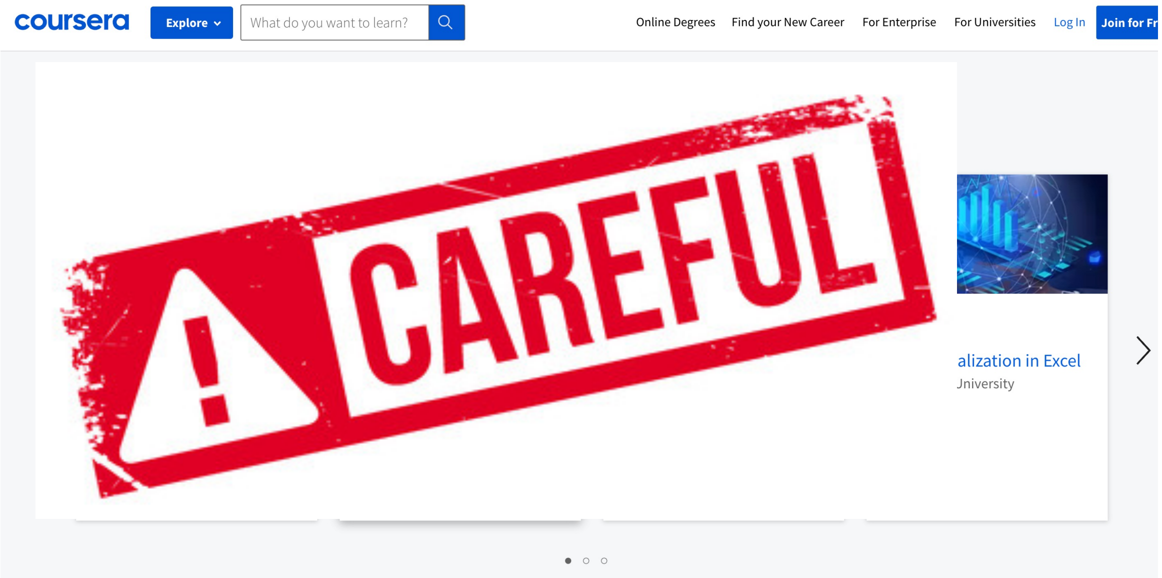
Data mining & knowledge discovery "Data Science" course by Prof. Jian Pei

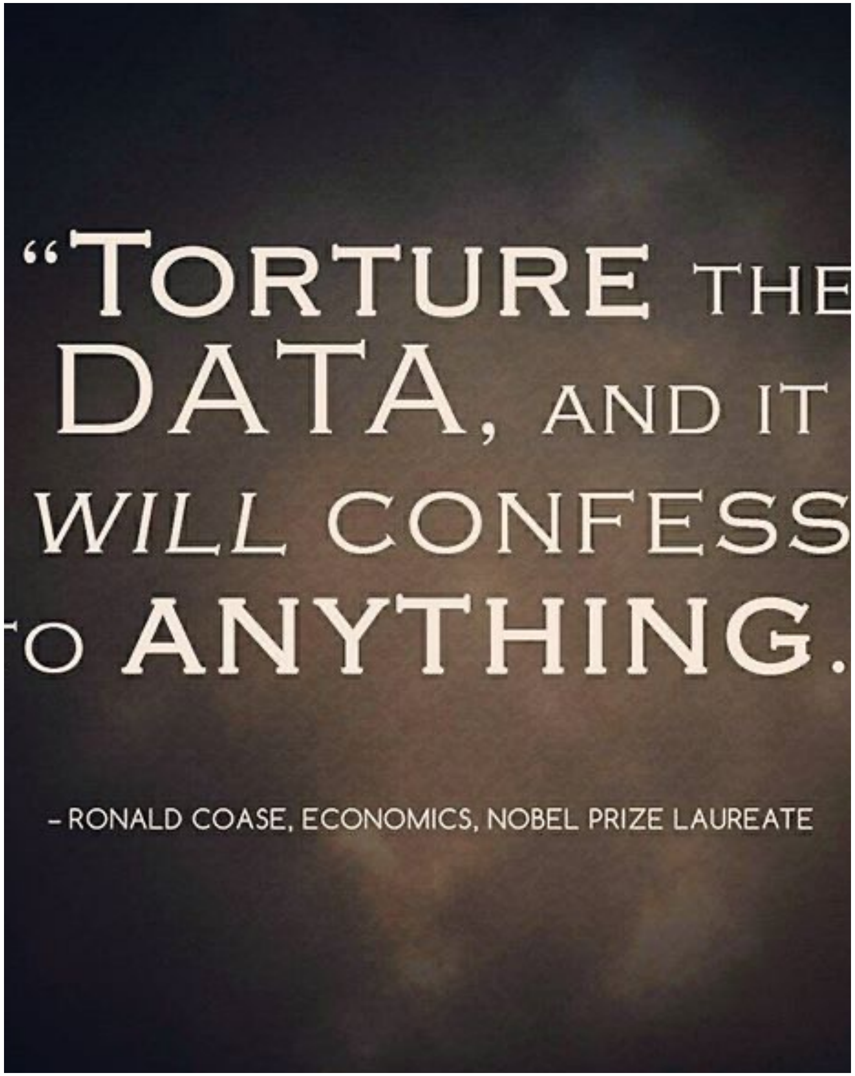


Discover useful information

Inform conclusions

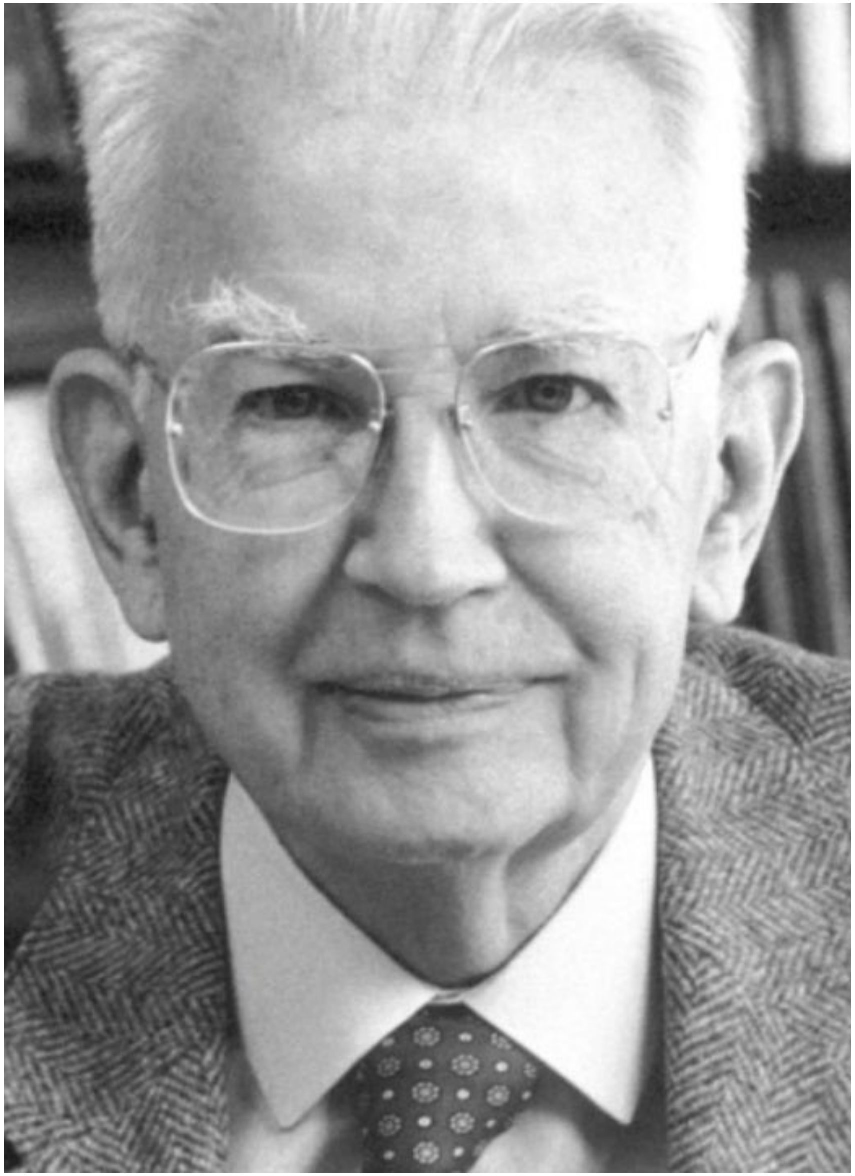
Support decision-making





“TORTURE THE DATA, AND IT *WILL* CONFESS TO ANYTHING.

– RONALD COASE, ECONOMICS, NOBEL PRIZE LAUREATE



Q: Is having an MS better for PhD admission?

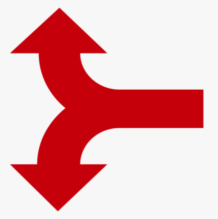
(Hypothetical data)

	Admitted	Total	%
MS	27	100	27%
No MS	60	200	30% 🏅

Department A

	Admitted	Total	%
MS	150	200	75%
No MS	78	100	78% 🏅

Department B



Total

	Admitted	Total	%
MS	177	300	59% 🏅
No MS	138	300	46%

Which version do we report?
What variables do we condition on?

Simpson Paradox

[Sex bias in graduate admissions: Data from Berkeley Bickel et al., Science, 1975]

	Admitted	Total	%
Male	27	100	27%
Female	60	200	30% 🏅

Department A

	Admitted	Total	%
Male	150	200	75%
Female	78	100	78% 🏅

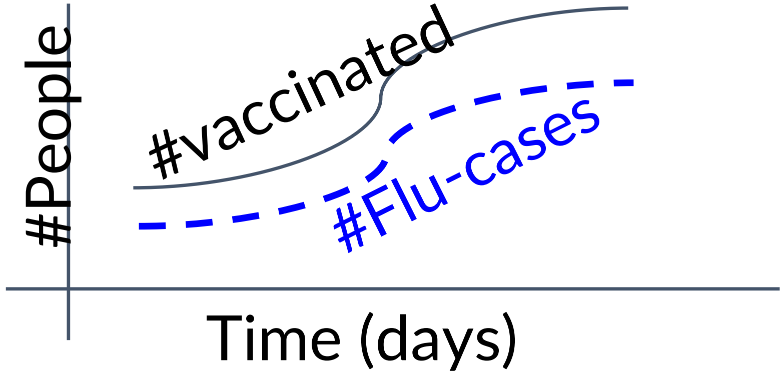
Department B

Total

	Admitted	Total	%
Male	177	300	59% 🏅
Female	138	300	46%

Change MS/No MS with Male/Female
(apologies for illustrating with binary gender)
"Does "Gender" affect admission decision?"

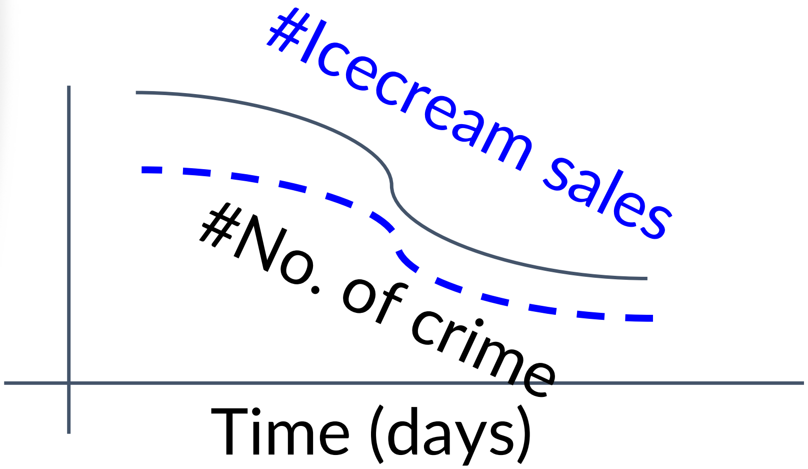
Q: Does taking flu vaccine help prevent flu infections?
Q: Does eating icecreams cause more crime?



- Oops!

- What do you think might be the reason?

(hypothetical graphs)



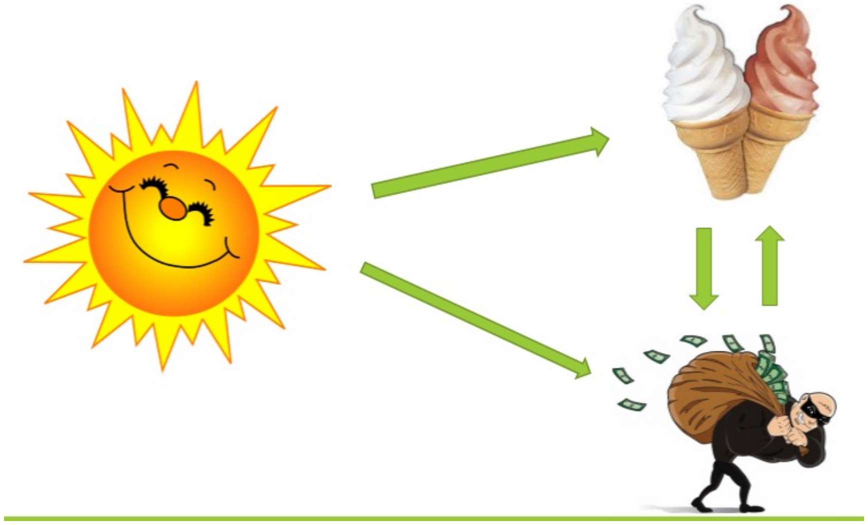
We should learn ML predictions, correlations, association etc. …

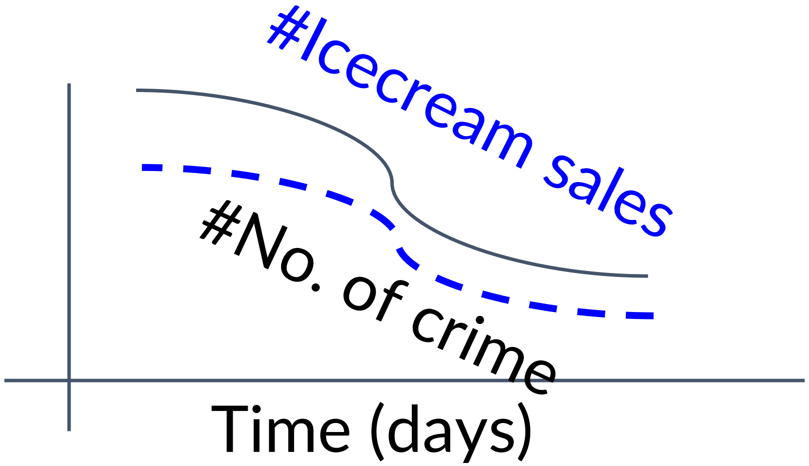
But, Before forming any conclusions and taking actions, we need to do a sound “causal” analysis

What is Causal Inference?

Image from the “Causal Inference in Statistics: A Primer” book by Pearl et al.

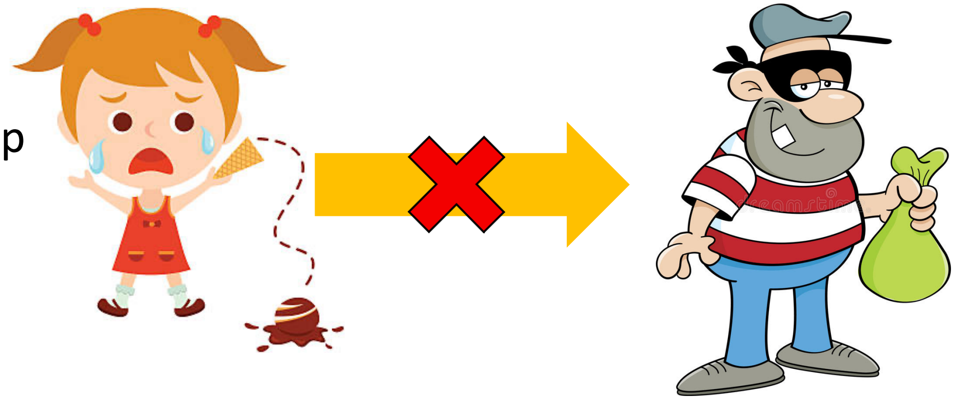
Correlation ≠ Causation





Common cause (confounder)

We need to understand causal relationship
before making a decision or policy



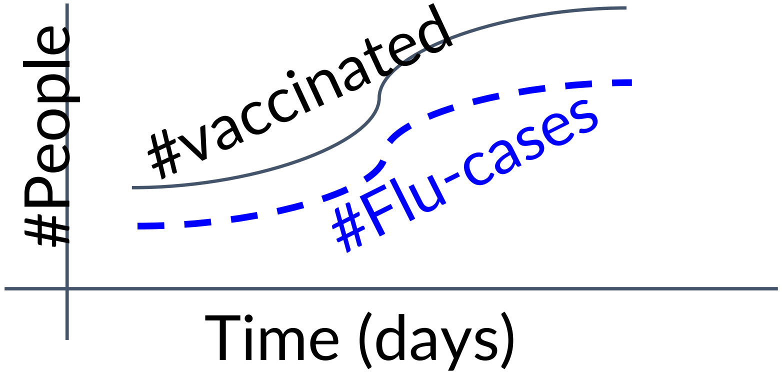
Correlation ≠ Causation

Positive Correlation

During a Flu Season -

- More Flu Infection
- More Flu Vaccination

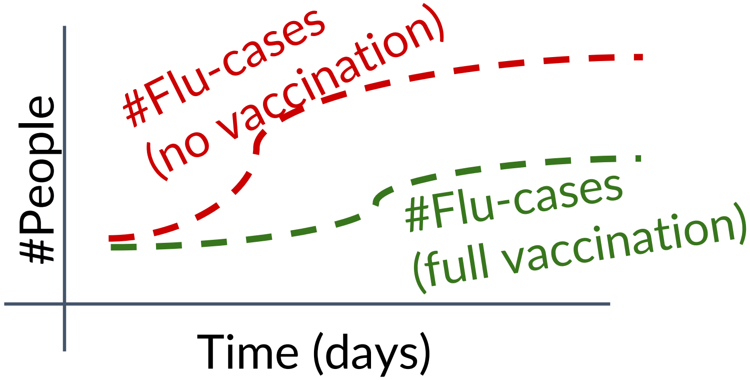
→ Doesn't Imply Vaccines causes Flu!



Causation ("intervention")

During a Flu Season -

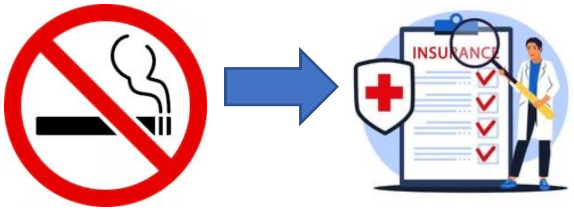
- What-if no-one was vaccinated?
    - Will the number of cases be more?
- What-if everyone was vaccinate?
    - Will the number of cases be small?



“Causal Analysis” is Important

Clinical Trial (COVID vaccine!)

Drug and vaccine discovery, and healthcare



Does quitting smoking reduce insurance premium?

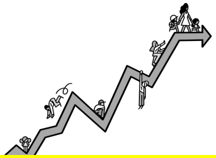
https://www.nytimes.com/2015/01/01/opinion/social-programs-that-work.html

The New York Times

Opinion

OP-ED CONTRIBUTOR

Social Programs That Work



By Ron Haskins

Dec. 31, 2014

Does a teen outreach program help reduce school dropouts?

At 24 mostly rural locations in Florida, Wyman's Teen Outreach Program works with 6,000 ninth graders a year to promote healthy behaviors, life skills and a sense of purpose. Evaluat program, which is based on a nine-month curriculur helped reduce teen pregnancies and lowered the risl suspension and dropout.

At 160 elementary schools in low-income communities in California, Colorado, Maryland, New York, Oklahoma, South Carolina, Texas, Washington and the District of Columbia, a program called Reading Partners pairs volunteer tutors with children for twice-weekly 45-minute sessions. An evaluation of the schools across three states by the research firm d substantial improvements in reading skills.

Do reading sessions by volunteers help improve reading skills of children?

performance level are grouped together and receive daily, 90-minute reading classes, as well as one-on-one tutoring and cooperative learning activities. We know it works because a study that randomly assigned 41 schools across 11 states to an experimental or control group found impr including comprehension, in students in tl Most of the students were black or Hispar families. Success for All was awarded $50 double its network of schools over five yea to improve effectiveness in new sites.

Social Studies

In Lancaster County, Pa., the Nurse-Family Partnership serves 175 low-income, first-time moms. Nurses start visiting the mothers before birth and continue, with diminishing frequency, until the child is 2. The nurses are trained to form a close relationship with the mother and advise her on prenatal health and child-rearing issues — including smoking and drinking during pregnancy and

Do home-visits of expecting low-income mothers by nurses help children’s well being later?

is 15. The mothers who participated were less likely to abuse or neglect their kids, and more likely to be working, and their kids were more likely to be healthy and ready for school.

Causality: A (really) long history

Aristotle
(384-322 BC)
Metaphysics / Four Causes

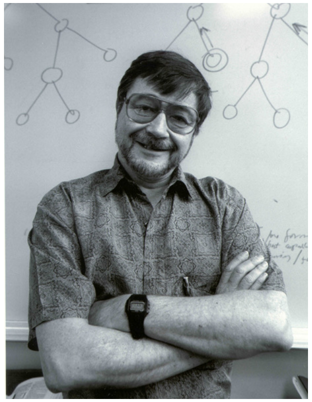
David Hume
(1738)
A Treatise of Human Nature

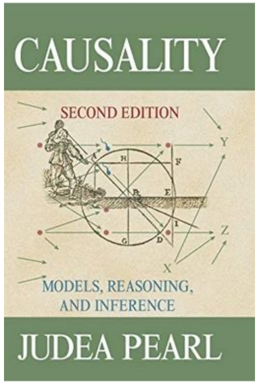
Karl Pearson
(1911)
The Grammar of Science, 3rd ed.

Jerzy Neyman
(1923)
Master's thesis: On the Application of Probability Theory to Agricultural Experiments. Essay on Principles

- *"We do not have knowledge of a thing until we have grasped its why, that is to say, its cause."* — Aristotle
- *"…Thus we remember to have seen that species of object we call Flame, and to have felt that species of sensation we call Heat. .. Without any farther ceremony, we call the one Cause and the other Effect, and infer the existence of the one from that of the other."* -- Hume
- "..before we can accept [any cause of a progressive change] as a factor we must have not only shown its plausibility but if possible have demonstrated its quantitative ability" - Pearson

Two Popular *Formal* Causal Models



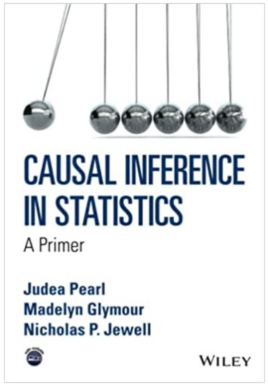


CAUSALITY

SECOND EDITION

MODELS, REASONING, AND INFERENCE

JUDEA PEARL



CAUSAL INFERENCE IN STATISTICS

A Primer

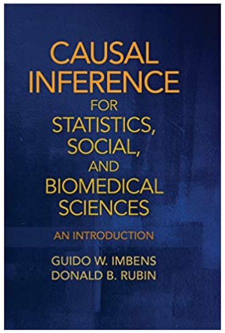
Judea Pearl
Madelyn Glymour
Nicholas P. Jewell

WILEY

Both are used in research and practice in recent times

We will study both

Pearl’s ”Graphical Causal Model” (1985, 1999 - ….)
(AI)



CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES

AN INTRODUCTION

GUIDO W. IMBENS
DONALD B. RUBIN

Rubin’s (Neyman-Rubin’s) “Potential Outcome Model” (1923, 1974 - …..)
(Statistics)

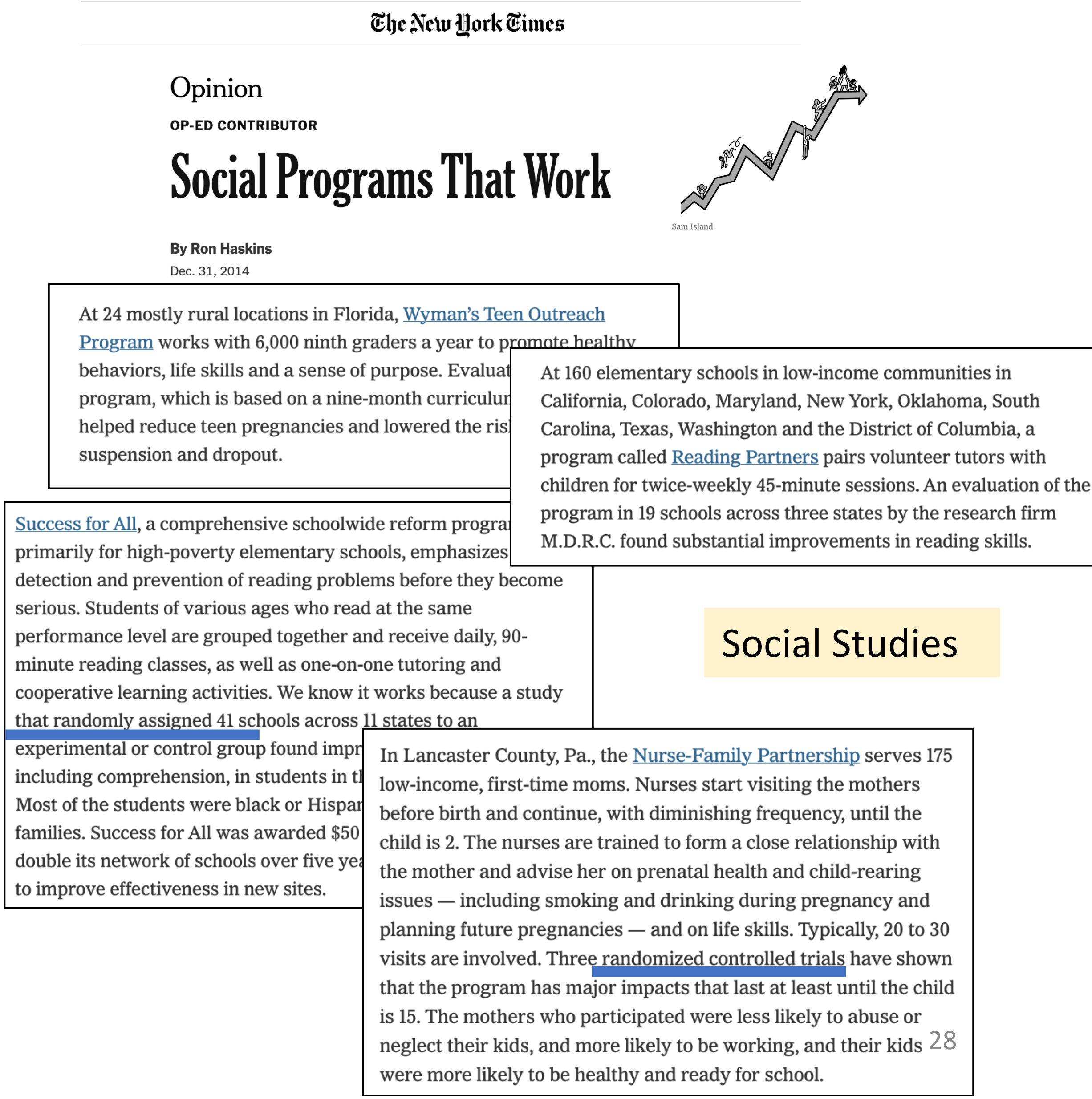
Review on board

- Probabilities, conditional probabilities, independence, Bayes’ rule, expectation

- Graphs – directed, undirected, edges, nodes, paths, reachability

Gold standard of causal inference: Controlled Trial

Clinical Trial



The New York Times

Opinion

OP-ED CONTRIBUTOR

Social Programs That Work

Sam Island

By Ron Haskins
Dec. 31, 2014

At 24 mostly rural locations in Florida, Wyman's Teen Outreach
Program works with 6,000 ninth graders a year to promote healthy
behaviors, life skills and a sense of purpose. Evaluat
program, which is based on a nine-month curriculu
helped reduce teen pregnancies and lowered the risl
suspension and dropout.

At 160 elementary schools in low-income communities in California, Colorado, Maryland, New York, Oklahoma, South Carolina, Texas, Washington and the District of Columbia, a program called Reading Partners pairs volunteer tutors with children for twice-weekly 45-minute sessions. An evaluation of the program in 19 schools across three states by the research firm M.D.R.C. found substantial improvements in reading skills.

Success for All, a comprehensive schoolwide reform progra
primarily for high-poverty elementary schools, emphasizes
detection and prevention of reading problems before they become
serious. Students of various ages who read at the same
performance level are grouped together and receive daily, 90-
minute reading classes, as well as one-on-one tutoring and
cooperative learning activities. We know it works because a study
that randomly assigned 41 schools across 11 states to an
experimental or control group found impr
including comprehension, in students in t
Most of the students were black or Hispa
families. Success for All was awarded $50
double its network of schools over five ye
to improve effectiveness in new sites.

In Lancaster County, Pa., the Nurse-Family Partnership serves 175 low-income, first-time moms. Nurses start visiting the mothers before birth and continue, with diminishing frequency, until the child is 2. The nurses are trained to form a close relationship with the mother and advise her on prenatal health and child-rearing issues — including smoking and drinking during pregnancy and planning future pregnancies — and on life skills. Typically, 20 to 30 visits are involved. Three randomized controlled trials have shown that the program has major impacts that last at least until the child is 15. The mothers who participated were less likely to abuse or neglect their kids, and more likely to be working, and their kids were more likely to be healthy and ready for school.

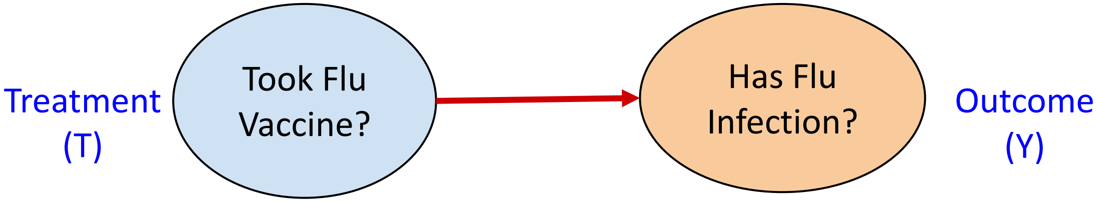
Social Studies

https://www.nytimes.com/2015/01/01/
opinion/social-programs-that-work.html

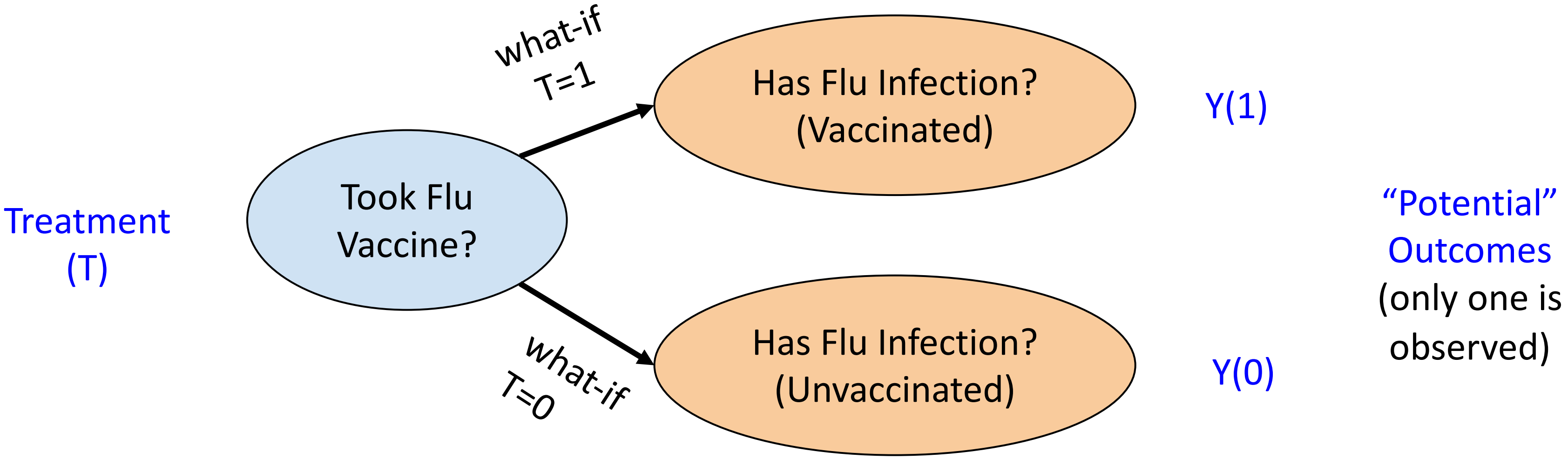
How does a controlled trial help?

… we will see using
Rubin’s Potential Outcome Framework

Treatment (T) & Outcome (Y)



Goal: Average Treatment Effect



Treatment (T)

Took Flu Vaccine?

what-if T=1

what-if T=0

Has Flu Infection? (Vaccinated)

Has Flu Infection? (Unvaccinated)

Y(1)

Y(0)

"Potential" Outcomes (only one is observed)

Average Treatment Effect (ATE) = **E**[Y(1) - Y(0)]

Randomized Controlled Experiments

[Rubin, '74 '05]

Treatment
(vaccine)

Population

Control
(placebo)

At random

$T \perp Y(1), Y(0)$

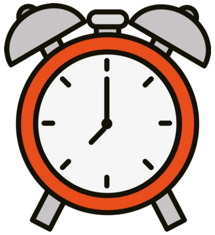
(Only one of Y(1), Y(0) is observed)

Average Treatment Effect (ATE) = $\mathbf{E}[Y(1) - Y(0)]$

$= E[Y(1) \mid T = 1] - E[Y(0) \mid T = 0]$

Can be estimated from experimental observed data!

What if we cannot do randomized controlled experiments?



Due to ethical, time, or cost constraints

- *“Does smoking cause lung cancer?”*
- *“Does growing up in a poor neighborhood make a child earn less as an adult?”*

Fortunately, we can do

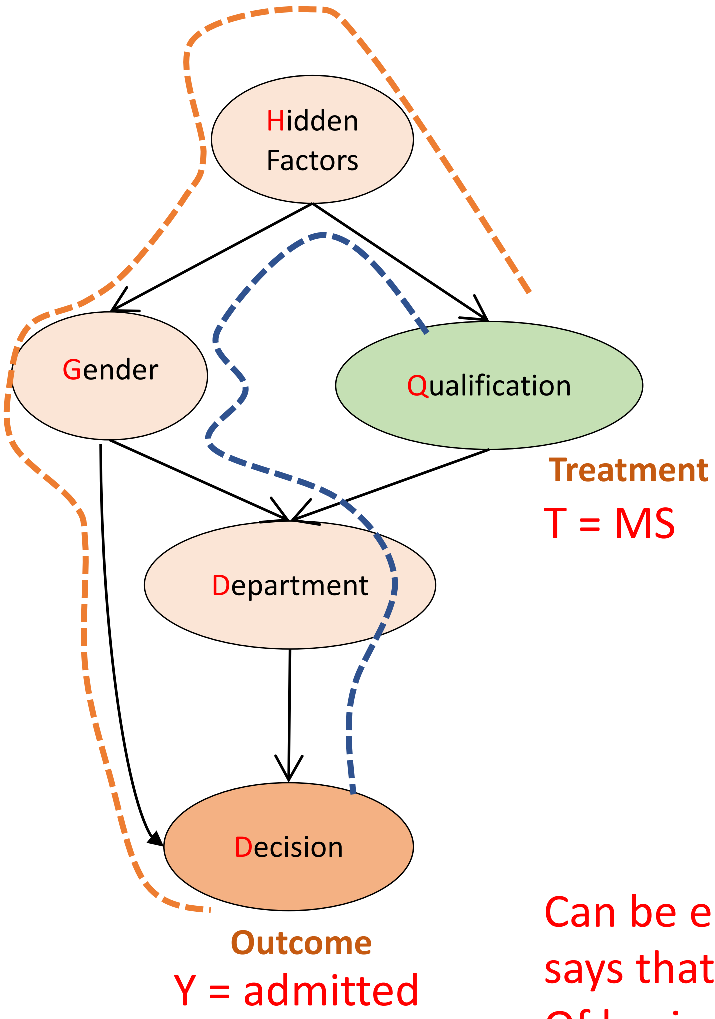
“Observational Causal Studies”

Under certain assumptions

By Rubin’s or Pearl’s model

Glimpse: Observational Studies with Pearl's Graphical Model

We will learn these concepts and techniques in this class



Goal:

Reduce causal relationship as "do-operators" to observed conditional probabilities

$$\Pr(D = \text{yes} \mid do(Q) = MS)$$

- Find the right variables to condition on
  - "d-separation" (from **graphical models** in AI)
  - "Back-door condition"

$$= \sum_g \Pr(D = \text{yes} \mid Q = MS, G = g)\Pr(G = g)$$

Can be estimated from data:
says that to understand the causal effect
Of having an MS on PhD admission decision, condition on gender

What is Fairness?

What are explanations?

How causal inference helps?

Share your thoughts – more later in the class

What’s next?

- Possible topics for presentations and projects will be posted by the next class on Thursday
- Talk to fellow students about forming project teams with 2-3 students
    - aim big!
    - In the last seminar course I taught, students wrote 2 SIGMOD/VLDB research papers and 2 demo papers starting with class projects, although it took more time (> 1 year) after the class