Principles of Causal Inference

Vasant G. Honavar

Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics,
Public Health Sciences and Neuroscience
Center for Artificial Intelligence Foundations and Scientific Applications
Institute for Computational and Data Sciences
Huck Institutes of the Life Sciences
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Introductions

- Vasant Honavar
    - Professor
        - Graduate programs in
            - Informatics
            - Computer Science and Engineering
            - Bioinformatics and Genomics
            - Operations Research
            - Neuroscience
            - Public Health Sciences
        - Undergraduate program in
            - Data Sciences
        - E335 Westgate Bldg
        - vhonavar@psu.edu
        - http://faculty.ist.psu.edu/vhonavar
- Students?

What is this course about?

- Modern machine learning methods have been incredibly successful in building predictive models from observational data
- Such predictive models are fundamentally incapable of answering causal questions e.g., How would one's risk of heart disease change if one were to quit smoking?
- Drawing valid causal conclusions from data calls for tools for
  - Specifying and reasoning with causal assumptions
  - Identifying and adjusting for confounders
  - Estimating causal effects from causal assumptions and observational and experimental data
  - Generalizing findings across experimental settings
  - Learning causal models from data

What is this course about?

- An introduction to the theoretical underpinnings, and practice of causal inference from observational and experimental data.
- Topics to be covered include:
  - Association versus causation;
  - Causal inference in the absence of randomized control trials; causal effects and counterfactuals;
  - Eliciting causal effects from observations – Potential outcomes framework;
  - Causal Bayesian Network framework for causal inference - do-calculus, identifiability of causal effects from observations and experiments;
  - The relationship between the Potential Outcomes and Causal Bayes Networks; and learning causal models from observations and experiments.
  - Treatment of confounders as well as practical approaches to cope with confounders;
  - Additional topics: mediation analysis; causal transportability, selection bias, meta analysis, applications of causal inference in uncovering algorithmic bias, explaining black box predictive models, etc.
- Course projects will focus on theory, principles and advanced methods for causal inference or applications of causal models and causal inference in the empirical sciences.

About the course

- Who is this course for?
  - Graduate students interested in
    - Causal data science
    - Causal AI
    - Research methods for causal inference from data
- What are the prerequisites for taking the course?
  - Graduate standing in informatics or a related discipline
  - Exposure to basic probability and statistics and discrete math
  - Basic proficiency in programming

Course objectives: What

- Upon successful completion of the course, students should:
  - Demonstrate a broad understanding of the principles of causal inference, including the Potential Outcomes and Causal Bayesian networks frameworks, as well as their applications in the data sciences.
  - Demonstrate an understanding of the implementation, adaptation, and applications of several causal inference algorithms in a high-level programming language (e.g., Python).
  - Identify, formulate, and solve causal inference problems that arise in the empirical sciences.
- Students with the necessary computational and mathematical background will also be prepared to pursue advanced research on the foundations of, and methods for causal inference in Data Sciences and Artificial Intelligence.

Course objectives: How

- To attain the learning goals, you will have to:
  - Actively participate in class
  - Complete assigned readings
  - Familiarize yourself with the relevant software tools
  - Solve assigned problems and lab exercises
  - Complete a class project
  - Take turns being the scribe for lecture notes

Course Policies

- https://faculty.ist.psu.edu/vhonavar/Courses/causality/policies.html
- Grading
  - Class participation (including scribing lecture notes)
  - Assignments
  - Term project
  - Exams
- Academic honesty

Course materials

Study guide, assignments, etc.

- Canvas
- https://faculty.ist.psu.edu/vhonavar/Courses/causality/homepage.html

Recommended books

1. Pearl, J., Glymour, M. and Jewell, N.P., 2016. Causal inference in statistics: A primer. John Wiley & Sons.
2. Hernán, M.A. and Robins, J.M., 2020. Causal inference: what if. Boca Raton: Chapman & Hill/CRC, 2020.
3. Neal, Brady. 2020. Introduction to Causal Inference from a Machine Learning Perspective

Reference books

1. Pearl, J. and Mackenzie, D. (2018). The book of Why. The new science of cause and effect. Basic Books.
2. Cunningham, D. (2021) Causal Inference. The Mixtape. Yale University Press.
3. Huntington-Klein, N. (2021). The Effect: An Introduction to Research Design and Causality. CRC Press.
4. Pearl, J., 2009. Causality. Cambridge university press.
5. Rosenbaum, P.R., 2017. Observation and experiment. Harvard University Press.
6. Imbens, G.W. and Rubin, D.B., 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.
7. Morgan, S.L. and Winship, C., 2015. Counterfactuals and causal inference. Cambridge University Press.
8. Spirtes, P., Glymour, C.N., Scheines, R. and Heckerman, D., 2000. Causation, prediction, and search. MIT press.
9. Berzuini, C., Dawid, P. and Bernardinell, L. eds., 2012. Causality: Statistical perspectives and applications. John Wiley & Sons.
10. Brumback, B. (2022). Fundamentals of Causal Inference, CRC Press.
11. Shipley, B. (2000). Cause and Correlation in Biology. Oxford University Press.
12. Sloman, S. (2009). Causal Models: How People Think About the World and its Alternatives. Oxford Univ. Press.

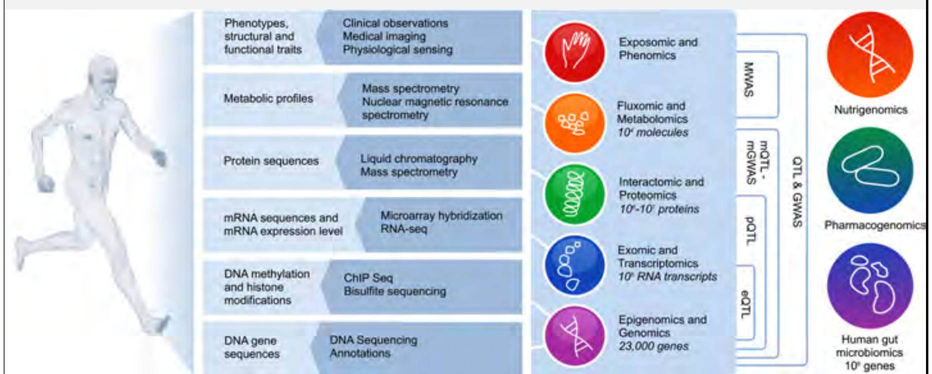
Our perspective on causality

- Philosophers and scientists have thought about causality forever.
- Many opinions and ideas (Hume alone had 3 definitions, Aristotle had 4).
- Our stance on causality is one that is most relevant for scientific inquiry.
- We will focus on notions of causality that we can operationalize, and not just philosophize about.
- Example “causal effect”: difference in outcomes of actual or hypothetical interventions.

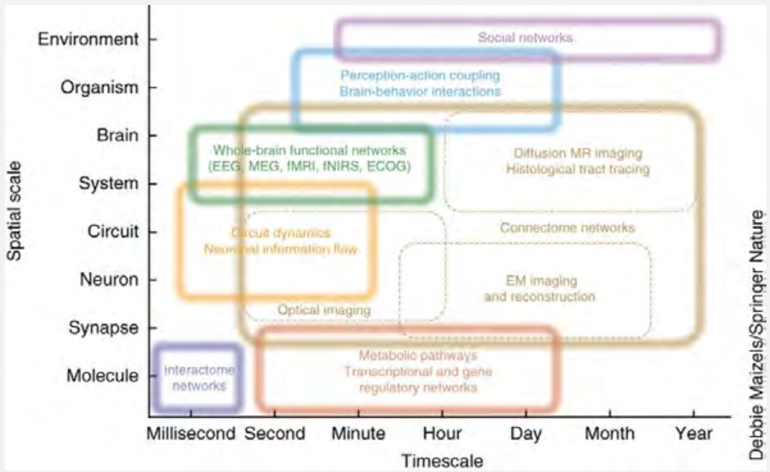
Multiple Views of Causal Inference

- Extending statistical analysis of experimental data to draw causal conclusions in settings where the controlled experiments are not feasible.
- Exploring how to do statistics or machine learning when your data is biased.
- Formalizing and operationalizing philosophical, e.g., David Hume's, notions of causality.
- Exploring why in general ``association is not causation" and the conditions under which it is.

Big Data in Health Sciences



Big Data in Brain Sciences



Big Data = End of the scientific method?

Big Data Doctrine

CHRIS ANDERSON, The end of theory: The data deluge makes the scientific method obsolete, *Wired Magazine* 16.07 (June 23, 2008). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

- "Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."
- "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."
- Most machine learning and data mining algorithms are essentially sophisticated ways of finding correlations from data

Is the Big Data Doctrine True?

CHRIS ANDERSON, The end of theory: The data deluge makes the scientific method obsolete, *Wired Magazine* 16.07 (June 23, 2008). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

- Are big data and powerful computers all we need for understanding complex systems?
  - How a complex gene network orchestrates development, aging and disease?
  - How changes in brain structure impact brain function and behavior?
- Almost certainly not, based on our experience so far[1]

[1] Jonas, E. & Kording, K. (2017) "Could a neuroscientist understand a microprocessor?" PLoS Comput Biol 13(1): e1005268.

Big Data = the end of the scientific method?
A lesson from Physics

Transformation of physics from a descriptive science (pre Newton) into a predictive science (post Newton)

- Brahe gathered 20 years of extremely accurate astronomical measurements: positions of the stars and planets: big data

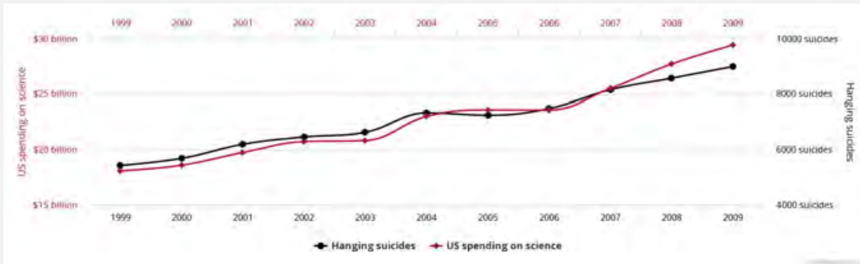
- Kepler, working for Brahe, fit the data in every way imaginable to discover laws of planetary motion: big data analytics and machine learning

- But it is only after Newton and Leibnitz invented calculus that there was language to express the laws of physics: knowledge representation for physics
- Big data did not make obsolete the scientific method then, and it does not do so now!

Big Data Doctrine Echoes Karl Pearson, Godfather of Statistics

- "The ultimate scientific statement of the description of relationship between two things can always be thrown back upon a contingency table" – Karl Pearson, "The Grammar of Science", 1892.
- "One can adopt an essentially model free approach, seeking to understand the data interactively by using a battery of displays, indices, and contrasts" – Samuel Karlin, Stanford, 1989
- Echo of Karl Pearson, the godfather of Statistics
  - Data already contain all scientific wisdom; all we need to do is to cajole the data using our tools to reveal that wisdom
  - There is no need for our analysis to take into account the process that generated the data

Is correlation enough? Can machine learning and statistics work its magic?



- Correlation between science funding and suicides by hanging is over 0.99!
- Should we eliminate science funding to save American lives?

Do we have a problem?

Why Most Published Research Findings Are False

The New York Times
Do We Really Know What Makes Us Healthy?

BIG DATA
The Parable of Google Flu: Traps in Big Data Analysis

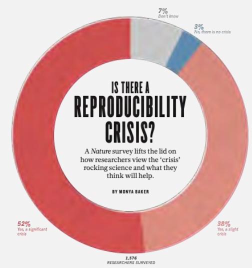
The replication crisis has spread through science – can it be fixed? | New Scientist 1/10/23, 12:25 PM

The replication crisis has spread through science – can it be fixed?

nature human behaviour LETTERS https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[6,16], Brian A. Nosek[7,8,16*], Thomas Pfeiffer[9,16], Adam Altmejd[2], Nick Buttrick[7,8], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,8], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]



- Indeed, we do!
- Why?
  - At least in part because of sloppy data analysis
  - Analyses that ignores the "story behind the data"
- Hence, the importance of causal models in making sense of data!

Cause and Effect

- Questions of cause and effect form the basis of almost all scientific inquiry
  - Medicine: drug trials, effect of a drug
  - Social sciences: effect of a certain policy
  - Genetics: effect of gene mutations on disease
- Causal inference is a central problem of AI

Statistical Association

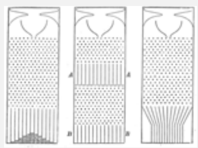
- Any attempt to discover a causal effect often starts by observing a statistical association
- A 'statistical association' between two factors means that they 'tend to appear together'
  - lung cancer is more common among smokers than among non-smokers
  - sickness is more common in hospitals than outside hospitals

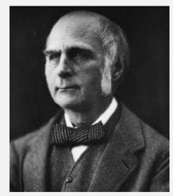
Association versus causation

- We are taught, "Causation is not association"
- What is meant is "Association does not imply causation"
- This begs the question: What is causation?
- Apart from a true causal effect, what could possibly explain the statistical association between
  - Smoking and lung cancer?
  - Hospitals and sickness?
  - Science funding and death by suicide by hanging?
  - Chocolate consumption and Nobel prizes?
  - Divorce rates in Alabama and per capita whole milk consumption?
  - Per capita cheese consumption and the number of lawyers in Iowa?

Francis Galton questions if we need causal explanations (1877)

- Data: The distribution of heights of individuals is roughly normal
- Galton's board illustrates why this is so
  - ➢ https://www.youtube.com/watch?v=EvHiee7gs9Y
- But if we make more passes through the Galton's board, the width of the distribution increases

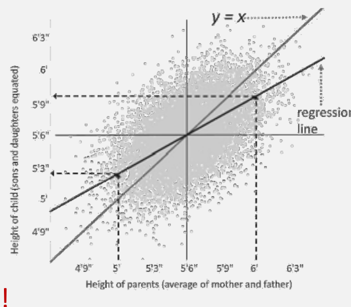




- But in nature, the height distribution is stable across generations
  - Tall parents have children who are shorter than they are
- Is there a causal explanation for this regression towards the mean?

Francis Galton questions if we need causal explanations (1877)

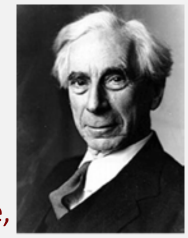
- Is there a causal explanation for this regression towards the mean?
- Galton observed that
  - Tall parents have children who are on average shorter than they are!
  - Tall children have parents who are on average shorter than they are!
  - How can a son be both taller and shorter than his or her parents?
- Not really: We are talking about populations, not individuals
- Galton concluded that there can be no causal explanation for regression to the mean!
- Galton started out trying to answer a causal question but ended up discovering correlation!



Karl Pearson banishes causality from science (1892)

- "The ultimate scientific statement of the description of relationship between two things can always be thrown back upon a contingency table" – Karl Pearson, "The Grammar of Science", 1892.
- "Beyond such discarded fundamentals as `matter' and `force' lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect."
- Pearson founded Biometrika, an influential statistics journal
- Pearson banished dissenters from "church biometric".
- Yet there were cracks in Pearson's edifice of causality free science
    - Spurious correlations!

Bertrand Russell suggests abandoning causality (1912)

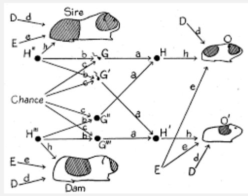


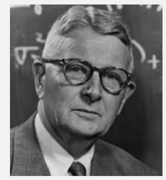
- "All philosophers [...] imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word **`cause'** never occurs."
- "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm."

But he seems to have changed his mind about causality later in his career

- "The power of science is its discovery of causal laws."

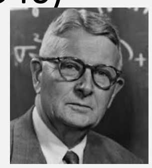
Sewall Wright proves Pearson wrong (1920-1940)





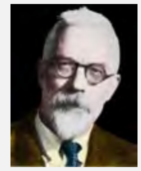
- Sewall Wright joined USDA to assume a job as a caretaker of Guinea pigs (1915) after receiving his PhD in genetics
- Wright found it impossible to breed all white guinea pigs
  - Even the most inbred families had considerable variation in color
  - Contradicts the prediction of Mendelian genetics that the coat color should become fixed after multiple generations of inbreeding
- Wright hypothesized that developmental factors (d) in the womb played a role – a hypothesis that was proven right in hindsight –after the discovery of DNA etc.
- Wright set up a ”path diagram” and solved a set of simultaneous equations to predict the coat color – thus inventing the first causal model!
- Implication: Some correlations do imply causation!

Sewall Wright proves Pearson wrong (1920-1940)



- Wright was attacked by Pearson and his followers
- Wright, by combining qualitative causal assumptions with 20 years of guinea pig breeding data, was able to establish that 42% of the variation in coat color is due to heredity
- Wright laid the foundations of structural causal models which were further developed by Pearl and others nearly 50 years later!
- Wright, a self-taught mathematician, faced the hegemony of the statistical establishment alone!

Ronald A. Fisher reduces statistics to data reduction



- "The object of statistical methods is the reduction of data".
- From 1920s through 1950s the scientific world turned to Fisher as the fountain of all statistical knowledge
- Fisher invented randomized trials
- Fisher believed that smoking did not cause cancer

Notes

- Statistical concepts are those expressible in terms of joint distribution of observed variables.
- The language of statistics cannot express, let alone, answer causal questions

Social scientists discover path analysis (1960s-1980s)


Simon

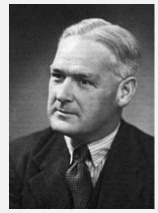

Otis


Duncan


Goldberger

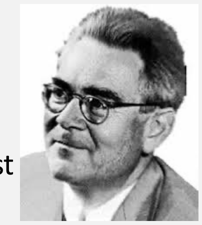
- Path analysis was relabeled as structural equation modeling (SEM)
- Over time, many social scientists used SEM software as a black-box and forgot about their causal underpinnings or causal interpretation

Bradford Hill Guidelines for Causation in Medical Research (1965)



- Strength of association
- Consistency
- Specificity
- Temporality
- Dose-response relationship
- Plausibility
- Coherence
- Experimental evidence
- Analogy
- An interesting checklist of considerations, initially proposed for epidemiological studies
- However, none of the criteria, except "temporality" (cause must precede effect), are either necessary or sufficient!

Hans Reichenbach posits a connection between causation and correlation (1956)

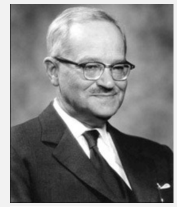


Reichenbach's Common Cause Principle (RCCP)

- No correlation without causation
- More explicitly RCCP claims that if two events *A* and *B* are correlated, then one of the following must be true:
  - *A* causes *B,* or
  - *B* causes *A* or
  - *A* and *B* share a common cause *C.*

Neyman & Rubin’s Potential Outcomes model (1970s)

- Jerzy Neyman and Donald Rubin’s potential outcomes model
  - Offers a formal definition of causal effects
  - Practical methods for estimating causal effects from observational data
  - Specifies the assumptions under which such estimates can be accurately obtained



Robins' marginal and nested structural models (1980s)

James Robins addresses causal inference from longitudinal data

- Marginal structural models
- Structural nested models

Prominent statisticians remain dismissive of causality

- "Considerations of causality should be treated as they have always been in statistics: preferably not at all." (Terry Speed, 1990)

Pearl's structural models and do-calculus (1990s-2010's)

Judea Pearl

- (Re)introduces and generalizes path diagrams as structural causal models
- Introduces do-calculus for reasoning with causal effects
- Establishes identifiability of causal effects from causal assumptions
- Establishes conditions for generalizability of causal effects
- Provides the language for expressing and answering causal questions

How we got here

- Godfathers of Statistics claimed Statistics to be the language of science
- The primary concern of statistics was to summarize the data
- Science became an exercise in correlation analysis and hypothesis testing
- All other questions, especially those having to do with causality were dismissed as ad hocery
- Language of statistics is inadequate for expressing, let alone answering causal questions
- We now have the language and tools to ask and answer causal questions

Where are we now

- "More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all of prior recorded history." (Gary King, 2014)
- Emergence of causality from exile
  - Many workshops, including one at NAS
  - Papers in AAAI, NIPS, ICML, UAI, PNAS, JSSM…
  - Applications – algorithmic fairness, explaining deep neural network predictions …
- The emergence of causality from exile makes it fun to solve important problems that Pearson, Fisher, and most of their successors. . . were not able to articulate, let alone solve!
- This is just the beginning!

What is causality?

- What does it mean to *cause* something?
- Cause and effect have been topics of deep philosophical debates since Aristotle!
- Our view: Meaning of causal claims can be understood in terms of, roughly speaking, conditionals of the form "If *A* had not occurred, *C* would not have occurred"
- We will make this much more precise

Cause and effect

Interventionist definition of causality

- *T* causes *Y* iff changing *T* leads to a change in *Y*, all else being held constant
- The causal effect of *T* on *Y* is the magnitude by which *Y* is changed by a unit change in *T*.

Causality and counterfactuals

  - How would the economy have responded had the interest rate not been raised?
  - Would the patient have been alive had he not suffered a stroke?

Why just statistics and machine learning are not enough

- There are tasks of **prediction**, **control** and **explanation**.
- Prediction is the focus of most of machine learning, statistics, predictive analytics etc.
- Control is about **taking actions** to achieve a particular outcome.
  - How should I change my diet to reduce the risk of heart disease?
- Explanation concerns what the outcome would be if you had done something differently.
  - Would Jane have recovered had she taken the drug?

Causality and scientific enquiry are inseparable!

The central concern of all sciences, has to do with discovering, representing, and reasoning about causal relationships

- How does a gene mutation impact cancer?
- What would happen to economic growth if taxes were lowered?
- Would you have been hired had you been female?
- How should I change my diet to reduce my risk of heart disease?

A causal model allows us to

- Understand mechanisms
- Predict the results of interventions
- Control events

What is Causal Inference really about?

- Causal inference is NOT about definitively establishing that A causes B
- Causal inference is about reasoning about cause-effect relationships from causal assumptions and data
- Causal assumptions are subjective, not verifiable from available data
  - You and I can disagree about the assumptions
  - But once you accept the assumptions and the data, we cannot disagree about the conclusions
- Causal assumptions are not the same as "priors" in Bayesian statistics
  - Why? Priors can be expressed in the language of statistics
  - Causal assumptions cannot be expressed in the language of statistics

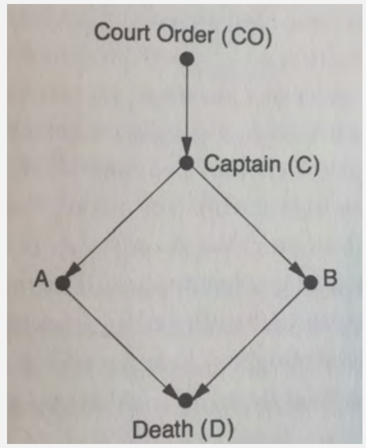
What can a causal reasoner do?

- Association
  - Activity: Seeing (Observation)
  - Question: How would seeing X change my belief about Y?
  - Methods: Statistics, Traditional machine learning
  - Powerful methods for summarizing data!
- Intervention
  - Activity: Doing (Intervention)
  - Question: What would Y be if I do X?
  - Statistics and traditional machine learning don't offer the means to even pose the question, let alone answer it!
- Counterfactuals
  - Activity: Imagining (Retrospection)
  - Question: What would Y be if I had not done X?
  - Statistics and traditional machine learning don't offer the means to even pose the question, let alone answer it

Representing causal assumptions

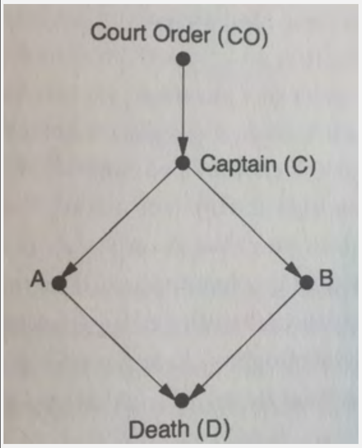
AI Mantra: Representation before anything else

- How can we represent causal knowledge?
- Causal diagrams
  - Nodes denote variables
  - Links denote direct causes
- Boolean Causal model
  - If a court order is given captain orders soldiers A and B to fire.
  - If at least one fires, prisoner dies.



Source: Book of Why, Pearl & Mackenzie

Answering questions of association



Source: Book of Why, Pearl & Mackenzie

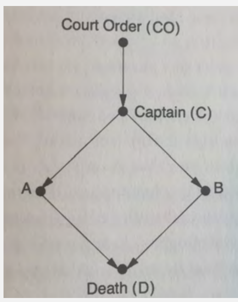
- What would observational data look like in this case?

CO	C	A	B	D
1	1	1	1	1
0	0	0	0	0
1	1	1	1	1
1	1	1	1	1
0	0	0	0	0

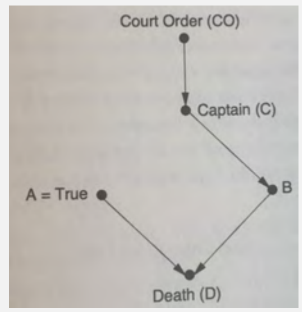
- Prisoner is found dead. Was court order given?
  - Yes
  - Why?
  - CO and D are perfectly correlated
  - $P(CO = 1|D = 1) = 1$

Answering questions of intervention

- Intervention: If soldier A goes rogue and shoots (without captain's order), would the prisoner die?
- We have no way to answer this question from observational data
- Why? This scenario is not observed!
- But if we have the causal graph, we can answer the question

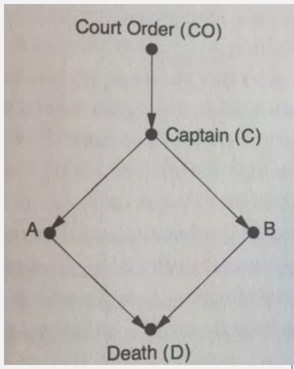


Mutilated Causal Graph



Source: Book of Why, Pearl & Mackenzie

Seeing ≠ Doing



Source: Book of Why, Pearl & Mackenzie

- Observational data

CO	C	A	B	D
1	1	1	1	1
0	0	0	0	0
1	1	1	1	1
1	1	1	1	1
0	0	0	0	0

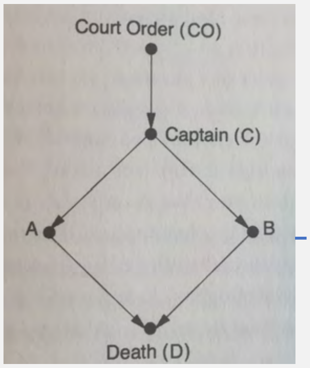
There is perfect correlation between the observed variables!!

- Given only such data, without a causal model, there is no way to know what happens when A goes rogue and fires in the absence of captain's order!

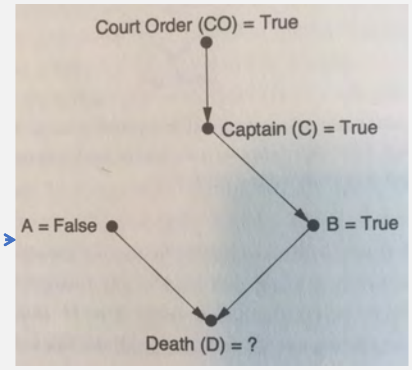
Seeing: If we see that A shoots, we can conclude that B shoots as well (correlation);
Doing: If A is forced to shoot, we can't say what B does, but we can say prisoner dies

Answering questions of imagination

- Counterfactual: Suppose the prisoner is found dead. Would he have died had A's gun failed to shoot?



Mutilated Causal Graph



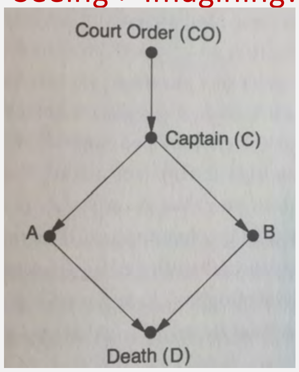
Source: Book of Why, Pearl & Mackenzie

Seeing ≠ Imagining!

Seeing: If D is dead, A and B must have shot (correlation)

Imagining: If A failed to shoot, and D is dead, B must have shot….

Seeing ≠ Imagining!



Source: Book of Why, Pearl & Mackenzie

- Observational data

CO	C	A	B	D
1	1	1	1	1
0	0	0	0	0
1	1	1	1	1
1	1	1	1	1
0	0	0	0	0

There is perfect correlation between the observed variables!!

- Given only such data, without a causal model, there is no way to explain the prisoner's death if A had not shot!

Seeing: If we see that A does not shoot, we can conclude that neither does B (correlation);
Imagining: If A is failed to shoot, we can attribute the prisoner's death to B having shot upon receiving a court order

Example: Should we ban vaccination?

Data:

- Out of 1 million people, 990,000 are vaccinated for COVID of whom 9900 have a reaction, and among those, 99 die
- 10,000 are not vaccinated, 200 get COVID of whom 40 die

Fact:

- More people die from COVID vaccine than those that die from COVID

Question:

- Should we ban vaccination?

Example: Should we ban vaccination?

Data:

- Out of 1 million people, 990,000 are vaccinated for COVID of whom 9900 have a reaction, and among those, 99 die
- 10,000 are not vaccinated, 200 get COVID, of whom 40 die

Fact:

- More people die from COVID vaccine than those that die from COVID

Question:

- Should we ban vaccination?
- Can you answer this question from the given data alone?
  - Why or why not?

Example: Should we ban vaccination?

Data:

- Out of 1 million people, 990,000 are vaccinated for COVID, of whom 9900 have a reaction, and among those, 99 die
- 10,000 are not vaccinated for COVID, of whom 200 get COVID of whom 40 die

From the data we can infer

- 99% of people are vaccinated, 1% are not
- A vaccinated person has a 1 in 100 chance of a reaction; and a reaction has a 1 in 100 chance of being fatal
- A person who is not vaccinated has 0 chance of reaction, but 1 in 50 chance of COVID which is fatal in 1 in 5 cases

Question:

- Should we ban vaccination?
- Can you answer this question from given data alone?
  - Why or why not?

Example: Should we ban vaccination?

Question:

- Should we ban vaccination?

Answer:

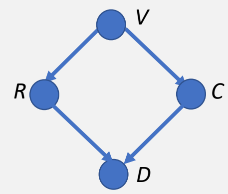
- Depends.
- On what?
- On how many would have died had no one been vaccinated!

Question:

- Can the data alone tell us how many would have died had no one been vaccinated? No!

Example: Should we ban vaccination?

- Suppose we know the story behind the data
- The story is expressed by the causal diagram shown



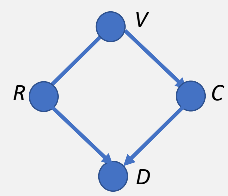
Data:

- 99% of of the people are vaccinated, 1% are not
- A vaccinated person has a 1 in 100 chance of a reaction; and a reaction has a 1 in 100 chance of being fatal
- A person who is not vaccinated has 0 chance of reaction, but 1 in 50 chance of COVID which is fatal in 1 in 5 cases

Question:

- Should we ban vaccination?
- We know how many died when 99% were vaccinated. We need to know how many would have died had no one been vaccinated.

Example: Should we ban vaccination?



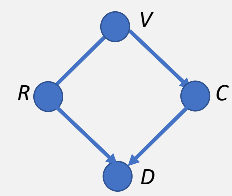
From the data we can infer

- 99% of people are vaccinated, 1% are not
- A vaccinated person has a 1 in 100 chance of a reaction; and a reaction has a 1 in 100 chance of being fatal
- A person who is not vaccinated has 0 chance of reaction, but 1 in 50 chance of COVID which is fatal in 1 in 5 cases

From data informed by causal diagram we can infer

- Out of 1 million people
  - If none were vaccinated, (1/50)(1/5)(1000000) = 4000 would have died
  - If 99% are vaccinated, 99+(10000)(1/50)(1/5) = 99+40= 139 would die

Example: Should we ban vaccination?



From data informed by causal diagram we can infer

- Out of 1 million people
  - If none were vaccinated, 4000 would die
  - If 99% are vaccinated, 99+40 = 139 would die
  - Fewer people (139) die with the vaccination policy in place than not (4000)
- Should we ban vaccination?
- Obviously not!

What did we just do?

- Causal inference from observational data!
- Using the causal diagram, and data, we were able to answer the question as to whether vaccination should be banned based on an imagined intervention!
- The imagined intervention is unethical to do in the real world if we believed that vaccination is beneficial because it would have meant withholding vaccination from people and watching them succumb to COVID!
- A causal diagram encodes causal assumptions and permits thinking about interventions
- Semantics
  - nodes represent observable variables
  - represent direct causal dependencies

Pearl's *do* notation

- We distinguish random X from X fixed by intervention by the notation "do(X)"
- X observed (seeing) is not the same as X fixed by intervention (doing)
- Average causal effect of X on Y

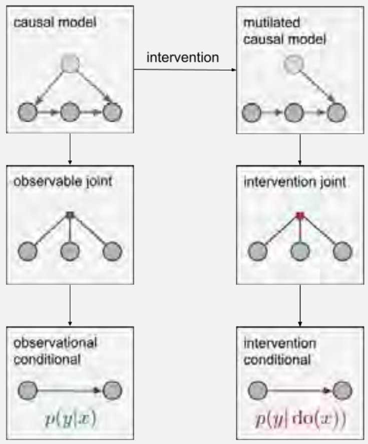
$$P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0))$$

- In the terminology of statistics, this is the estimand, the quantity to be estimated.
- We may derive the estimand from a causal graph and estimate it from data using an estimator

Do not conflate the estimand, model, and estimator

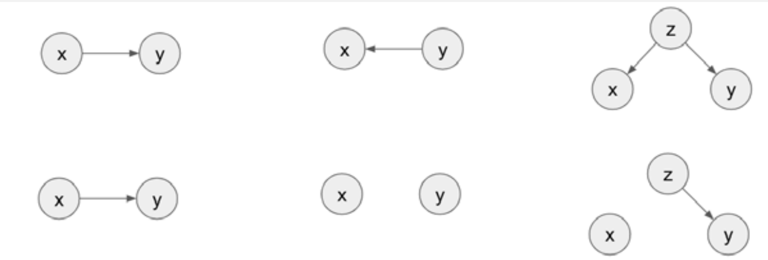
- Conflation between estimand, model, and estimator is a major source of confusion in debates about causal inference between warring camps
- The primary focus of causal graphs is on establishing the estimands
- We can leverage existing techniques (including statistical methods, machine learning , even deep learning) for estimating the estimand from data
- Postulating a causal graph allows us to
  - See how the causal estimand of interest can be written as a function of it, and
  - Check whether it can be estimated from observations
  - That is, determine if the causal effect of interest is identifiable

Seeing ≠ Doing!



Causal model allows reasoning about interventions

Seeing ≠ Doing!



$P(y|do(X)) = p(y|x)$ $\quad$ $P(y|do(X)) = p(y)$ $\quad$ $P(y|do(X)) = p(y)$

Causal model allows reasoning about interventions

Observations versus experiments

- Suppose you want to see whether a drug helps cure COVID
- Experiment: If you have the budget, and willing COVID-positive study participants, and can pass IRB review, you can run a randomized experiment (RCT)
  - Randomly assign study participants to treatment and control groups
  - Those in treatment group receive drug; those in control group receive placebo
  - Compare the two groups in terms their recovery from COVID
- RCT not always feasible due to cost, ethical considerations, etc.
- Observation: Prospectively choose two groups of COVID patients
  - The first group took the drug
  - The second did not
- Compare the two groups in terms of recovery from COVID

Observations versus experiments

- Suppose you want to see whether a drug helps reduce the risk of heart disease
- You could not run an RCT for whatever reason and had to make do with an observational study
- Prospectively choose two groups of individuals
  - The first group took the drug
  - The other did not
- Compare the two groups on incidence of heart disease
- Do you see a problem with this setup?
  - May be there are factors that impact both
    - adherence to drug prescription and
    - predisposition to heart disease
  - Confounding bias!

The curious case of a drug that is bad for men, bad for women, and good for people

	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart attack	*No heart attack*	*Heart attack*	*No heart attack*
Female	1	19	3	37
Male	12	28	8	12
Total	13	47	11	49

Source: Book of Why, Pearl & Mackenzie

- For women, the rate of heart attack was 1 in 20 (5%) without the drug and 3 in 40 (7.5%) with the drug – The drug is bad for women
- For men, the rate of heart attack was 12 in 40 (30%) without the drug and 8 in 20 (40%) with the drug – The drug is bad for men
- But paradoxically, the rate of heart attack was 13 in 60 without the drug and 11 out of 60 with the drug – The drug is good for people!
- Hmm!!!! How can a drug that is bad for men and for women be good for people?

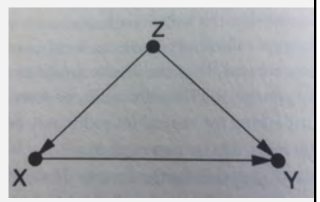
The curious case of a drug that is bad for men, bad for women, and good for people

	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart attack	*No heart attack*	*Heart attack*	*No heart attack*
Female	1	19	3	37
Male	12	28	8	12
Total	13	47	11	49

Source: Book of Why, Pearl & Mackenzie

- The data present an instance of Simpson's paradox which has puzzled statisticians since 1956
- There are dozens of papers and PhD theses in Statistics attempting to "explain" the Simpson's paradox
- Simpson's paradox underscores the pitfalls of analyzing observational data without causal assumptions
- Causal models provide a way to resolve the paradox

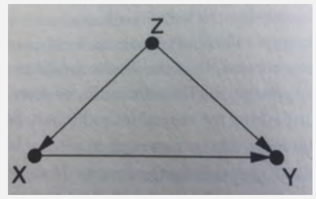
Confounding bias



- Suppose the treated group is healthier than the control group to start with
- Confounding bias arises whenever a variable influences both who is selected for treatment and the outcome of the experiment
  - Sometimes the confounders are known
  - Sometimes the confounders are suspected
- The most basic version of confounding
  - The true causal effect X → Y is mixed with the spurious correlation induced by the effect of X on Y that is attributable to Z
  - Example: We are testing a drug but give it to patients who are younger, but not to those who are older – age becomes a confounder

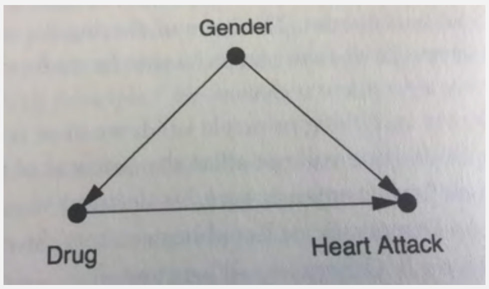
Confounding bias

- Suppose Z is a confounder



- If we have measurements on the confounder Z, we will see that it is easy to de-confound the true and spurious causal effects – by adjusting for Z
  - Compare treatment and control groups for each value of Z
  - Take a weighted average where the weights correspond to the fraction of the population represented by each value of Z

Back to the drug that is bad for men, bad for women, but good for people



- Suppose gender is unaffected by the drug
- Suppose gender affects both heart attack risk and whether the patient takes the drug
- Gender is a confounder that needs to be controlled for in assessing the effect of the drug on heart attack

Adjusting for gender resolves the paradox!

	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart attack	*No heart attack*	*Heart attack*	*No heart attack*
Female	1	19	3	37
Male	12	28	8	12
Total	13	47	11	49

Source: Book of Why, Pearl & Mackenzie

- For women, the rate of heart attack was 1 in 20 (5%) without the drug and 3 in 40 (7.5%) with the drug: The drug is bad for women
- For men, the rate of heart attach was 12 in 40 (30%) without the drug and 8 in 20 (40%) with the drug: The drug is bad for men
- Adjusting for the confounder, with the proportion of men and women being the same, we simply average the gender-specific heart attack rates to get the population heart attack rates
  - $(5 + 30)/2 \ = \ 17.5\%$ without the drug
  - $(40 + 7.5)/2 \ = \ 23.75\%$ with the drug
- The drug is bad for people. Paradox resolved!

Adjusting for confounders

- You can correctly determine causal effects by controlling for confounders
- Standard statistical methodology provides little guidance for what variables to control for
  - You may end up controlling for variables that you did not need to control for
  - You may fail to control for confounder(s) that you should have controlled for
  - In both scenarios, you can end up with incorrect causal conclusions
- Even if you get lucky and control for the exact set of variables that should have been controlled for,
  - you have no way of knowing that you did so, and therefore
  - avoid making causal claims even if they are justified

Adjusting for confounders

- We shall see that the determining the exact set of confounders to control for requires a causal graph
- Given a causal graph, we can determine the confounders we need to control for
- If the confounders are measured in the data, we can control for them and determine the causal effect of interest

Simpson's paradox and supervised machine learning

- Suppose you are asked to train an ML model to predict the benefit of a drug for heart patients
- A hospital supplies you some training data
- Because they thought gender did not matter or because they thought aggregating data across genders gave a larger sample, or because they did not want the predictive model to discriminate based on gender, they gave you data without gender information
- You train the model and deploy it.
- On some new patients, suppose the model predicts that the drug is beneficial.
- Should you trust the predictions?

Simpson's paradox and supervised machine learning

- Suppose you are asked to train an ML model to predict the benefit of a drug for heart patients
- A hospital supplies you some training data
- Suppose you have gender information along with other variables, and have reason to believe that gender is a confounder
  - What should you do?
- Suppose you don't know that gender is a confounder, but you suspect that gender, and perhaps some other variables could be confounders
  - What should you do?
  - Can you think of ways to identify the confounders?

Story behind the data: Data generating process

- Science presupposes that nature is governed by laws
- The laws work behind the scenes, and generate the data that we observe
- Observation: If we let go of a ball, it drops to the ground
- Data generating process (DGP): $F = G\,\frac{m_1 m_2}{r^2}$
- DGP can be far more complex in life sciences, behavioral sciences, social sciences
- Regardless of how complex DGP is, science presupposes the existence of DGP
- In practice, DGP consists of parts we know, and parts we don't

Story behind the data: Data generating process

- In practice, DGP consists of parts we know, and parts we don't
- If he was starting with nothing, Newton would have no way to figure out the law of gravitation
- If we know nothing, we can't rule out the possibility that planets move the way they do because of magic
- But if we know about mass, forces, momentum, velocity, can we learn about gravity?
- History tells us that we can

Shadows: Shadow Puppetry :: Data : Data Generating Process

Image source: Annie Katsura Rollins, Ballard Institute and Museum of Puppetry, photo by Kenneth Best

Data generating process

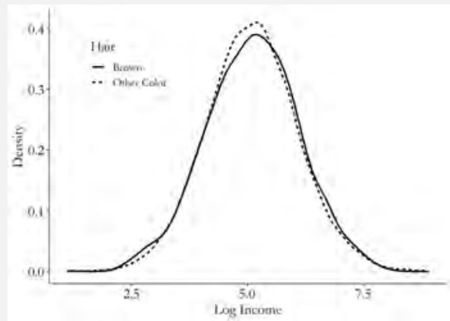
- Consider the following DGP
  - Income is log normally distributed
  - Being brown haired gives you a 10% log income boost
  - Having a college degree gives you a 20% log income boost
  - 30% of the population have college degrees
  - 20% of people are naturally brown haired
  - 40% of people who don't have brown hair or a college degree will dye their hair brown

$$P(College) = 0.3$$
$$P(BrownHair) = 0.2 + (0.8)(0.4)(\neg College)$$
$$\log(Income) = (0.1)BrownHair + (0.2)College + \epsilon$$

where ϵ is normally distributed

DGP and Causal Effect Identification

- Suppose we have some data generated by the DGP shown
- But we have no idea what the DGP is
- Suppose we are interested in the effect of $BrownHair$ on $Income$
- Suppose we start by looking at the observed distribution of log income by hair color
- We find that $BrownHair$ gives only 1% NOT 10% $\log(Income)$ boost
- Wrong answer



Hair	Mean Log Income
Brown	5.111
Other Color	5.095

DGP and Causal Effect Identification

- What can we do to get the right answer?
- Nothing, unless we have additional knowledge
- Suppose we know everything about the DGP except the effect of $BrownHair$ on $\log(Income)$
- Now can you figure out the right answer?
- How?

DGP and Causal Effect Identification

- What can we do to get the right answer?
- Nothing, unless we have additional knowledge
- Suppose we know everything about the DGP except the effect of $BrownHair$ on $\log(Income)$
- We know that only people without college degrees dye their hair brown
- But they don't get the wage bump due to college degree, so brown hair does not do much for them
- Knowing the parts of the DGP we do, tells us what we need to do with the data to get the right answer

DGP and Causal Effect Identification

- Limiting yourself to the distribution of log(Income) among those with college degrees, we find from the data

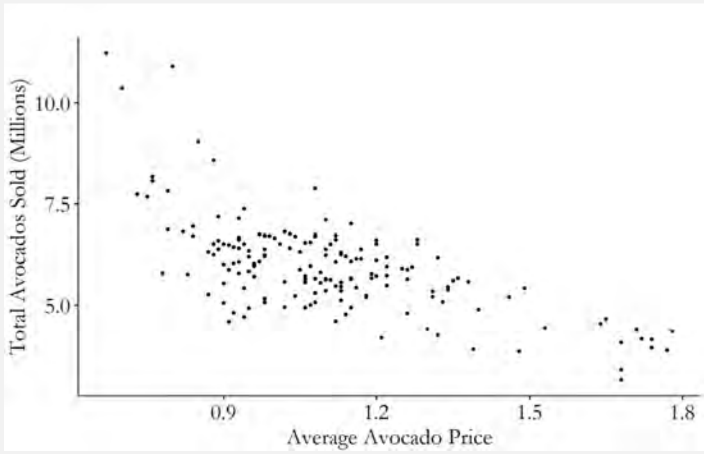
Hair	Mean Log Income for those with College degrees
Brown	5.340
Other Color	5.208

- Now we see that $BrownHair$ gives you approximately 13% boost in $\log(Income)$
- Why not 10%?
- Remember the data we have is a sample from a distribution
- If we repeated our sampling thousands of times, the mean boost in $\log(income)$ from $BrownHair$ will approach 10%
- Exercise – check this

How did we get the right answer (or close to it)?

- Looking for variation
  - DGP tells us about all the different processes working behind the scenes to give us the data we see
  - But to answer a question we have, e.g., regarding the effect of $BrownHair$ on $\log(Income)$, given what we knew about parts of the DGP, it turned out to be the variation in income by hair color just among those with college degrees
  - How can we find the variation we need to focus on?
- Identification
  - How can we use what we know about the DGP to be sure that the variation we are digging out is the right one to focus on?
  - Figuring out what problems in data we need to get rid of, like that resulting from people without college degrees dying their hair is the process of identification

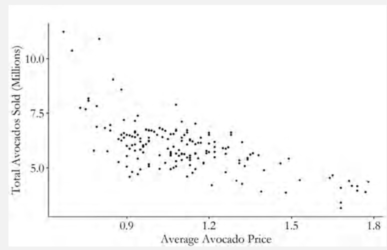
From variation to identification



What does this plot show?

From variation to identification

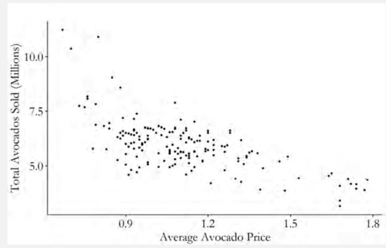
- What does this plot show?
- Negative correlation between price of avocados and avocado sales
- Does increase in avocado prices drive down avocado sales?
- Does increase in avocado prices drive down demand for avocados?



- Can you tell from the data? No!
- All that the graph tells us is the covariation between avocado prices and avocado sales
- But these variables can move around for a host of reasons

From variation to identification

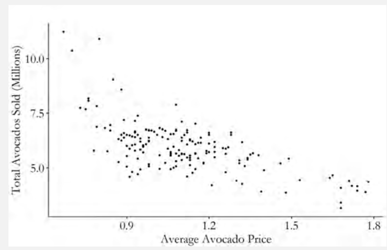
- Suppose we want to find out the effect of avocado prices on avocado sales
- The data is too messy
- How can we find the variation in the data that answers our question?
- We want the variation in people buying avocados (rather than people selling them) that is driven by changes in avocado price (rather than, say, avocados becoming less popular, or Elon Musk tweet discouraging people from buying avocados)



- This is a hopeless task unless we know something about the DGP!

From variation to identification

- Suppose we learn that
  - the sellers set weekly avocado prices at the beginning of each month and
  - don't change them until next month.

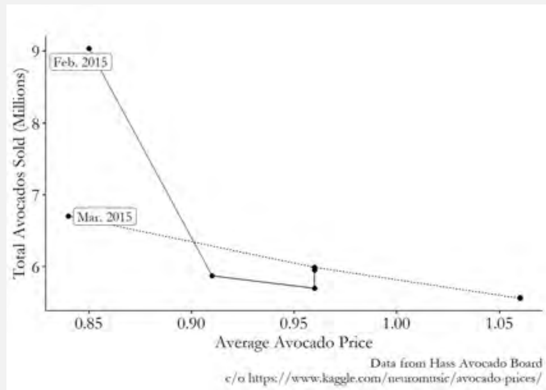


From variation to identification

- Now "suppliers set prices" and "suppliers determine supply" explanations matter only between months, not within each month!
- The variation in price and sales from week to week in the same month will isolate variation in people buying avocados and get rid of variation from people selling avocados
- Further, because the price is set by sellers, we can conclude that any observed variation in sales can only be driven by changes in price
- By tossing out the variation between months, we are eliminating explanations that rely on that variation, e.g., those that appeal to seller behavior, leaving us with only buyer behavior

From variation to identification

- Suppose we look at the data for each month
- We find that there is a negative relationship between avocado prices and sales
- Given the data and what we know about the DGP, we conclude that an increase in avocado prices does reduce avocado sales
  - We used our knowledge of the DGP to answer our question



From variation to identification

To figure out the part the data that answers our question, we must

- Using theory (what we know), specify as much of the DGP as we can
- Use the DGP to figure out the reasons for our data looking the way it does that do NOT answer our question
- Find ways to block out those alternative reasons to unearth the variation that we need to focus on to answer our question
- This process is a lot more difficult than “just looking at the data and seeing what it says”
- Causal graphs allow us to specify the causal structure of DGP, determine whether a causal effect is identifiable from the causal graph and data, and if it is, give us an identification formula that specifies the causal estimand that we need to estimate.