



Data Science for Researchers and Scholars

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
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
<http://faculty.ist.psu.edu/vhonavar>
<http://ailab.ist.psu.edu>

Three tiers of models

- Descriptive models or descriptive statistics
- Predictive models or supervised machine learning
- Causal models

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 far¹

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

**PennState**
Institute for Computational
and Data Sciences


Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

**PennState**
Clinical and Translational
Science Institute


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!**

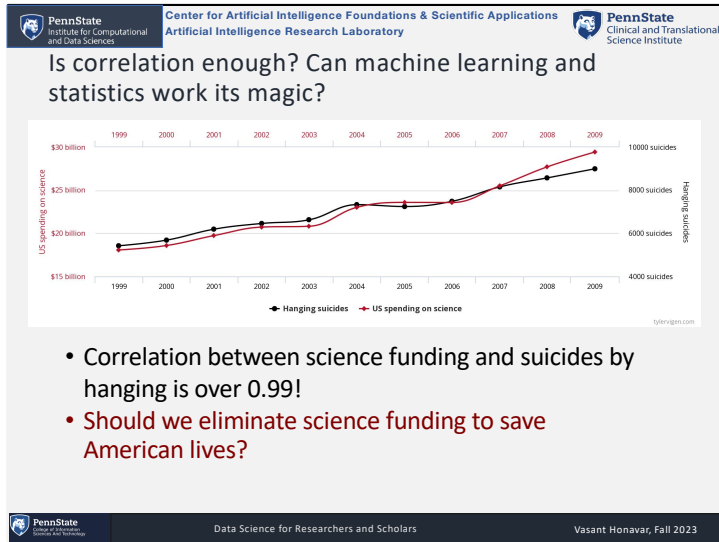
**PennState**
Institute for Computational
and Data Sciences

Data Science: for Researchers and Scholars

PennState
Clinical and Translational
Science Institute
Vasant Honavar, Fall 2023

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



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?

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!





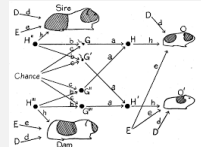

PennState
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory




PennState
Clinical and Translational
Science Institute

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!
- Lesson: **Some** correlations do imply **causation**!



PennState
Institute for Computational
and Data Sciences

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023

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!



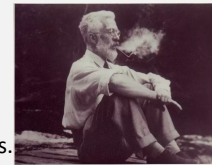
Ronald A. Fisher reduces statistics to data reduction


- “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







**PennState**
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory


**PennState**
Clinical and Translational
Science Institute

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



SimonOtisDuncanGoldberger

- 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

**PennState**
Institute for Computational
and Data Sciences

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023

Bradford Hill Guidelines for Causation in Medical Research (1965)

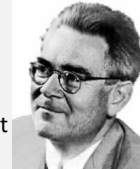
- Strength of association
- Consistency
- Specificity
- Temporality
- Dose-response relationship
- Plausibility
- Coherence
- Experimental evidence
- Analogy
- An interesting checklist of considerations, initially proposed for epidemiological studies
- However, none of the criteria, except “temporality” (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 is 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 outside the scope of statistics, and by implication, scientific analysis of data
- 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

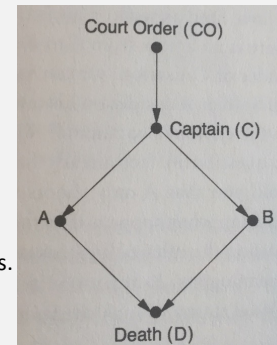
What can a causal reasoner do?

- **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

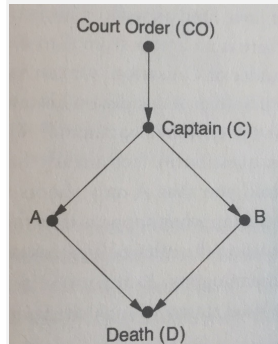
AI Mantra: Representation before anything else

- 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

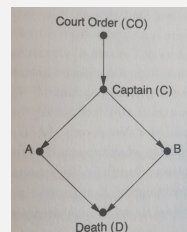
- What would observational data look like in this case?

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

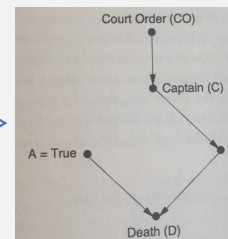
- Prisoner is found dead. Was court order given?
 - Yes
 - Why?
 - CO and D are perfectly correlated
 - $P(CO = 1|D = 1) = 1$

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

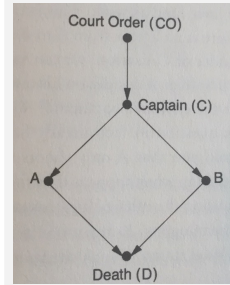


Mutilated
Causal Graph



Source: Book of Why, Pearl & Mackenzie

Seeing \neq 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

PennState

Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications

Artificial Intelligence Research Laboratory

PennState

Clinical and Translational Science Institute

Answering questions of imagination

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

```

graph TD
    CO((Court Order (CO))) --> C((Captain (C)))
    C --> A((A))
    C --> B((B))
    A --> D((Death (D)))
    B --> D

```

Mutilated Causal Graph

```

graph TD
    COTrue["Court Order (CO) = True"] --> CTrue["Captain (C) = True"]
    CTrue --> AFalse["A = False"]
    CTrue --> BTrue["B = True"]
    AFalse --> DQ["Death (D) = ?"]
    BTrue --> DQ

```

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....

PennState


Clinical and Translational Science Institute

Data Science for Researchers and Scholars

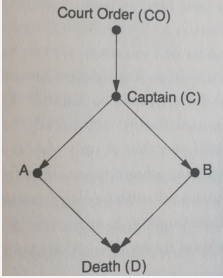
Vasant Honavar, Fall 2023


PennState
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory


PennState
Clinical and Translational
Science Institute

Seeing ≠ Imagining!



```

graph TD
    CO[Court Order (CO)] --> C[Captain (C)]
    C --> A[A]
    C --> B[B]
    A --> D[Death (D)]
    B --> D[Death (D)]

```

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


PennState
Institute for Computational
and Data Sciences

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023

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?

**PennState**
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

**PennState**
Clinical and Translational
Science Institute

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?

**PennState**
Clinical and Translational
Science Institute

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023

Example: Should we ban vaccination?

Question:

- Should we ban vaccination?

Answer:

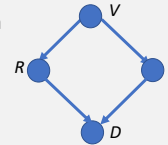
- Depends.
- On what?
- On how many would have died had no one been vaccinated!

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 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

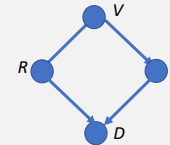
Question:

- 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



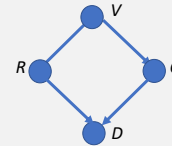
From data informed by causal diagram we can infer

- 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 " $\text{do}(X)$ "
- X observed (seeing) is not the same as X fixed by intervention (doing)
- Average causal effect of X on Y

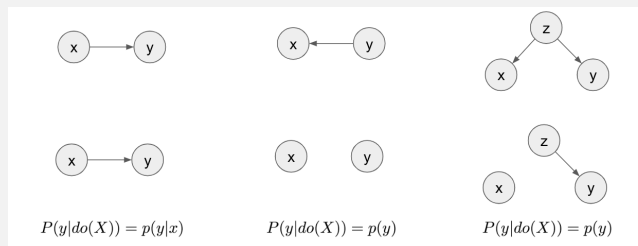
$$P(Y = 1 \mid \text{do}(X = 1)) - P(Y = 1 \mid \text{do}(X = 0))$$

- 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 \neq Doing!



Causal model allows reasoning about interventions

 PennState
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

 PennState
Clinical and Translational
Science Institute


 PennState
Institute for Computational
and Data Sciences

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023


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!



PennState
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory




PennState
Clinical and Translational
Science Institute

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?**



PennState
Institute for Computational
and Data Sciences

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023




PennState

Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications

Artificial Intelligence Research Laboratory



PennState


Clinical and Translational Science Institute

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)	
	<i>Heart attack</i>	<i>No heart attack</i>	<i>Heart attack</i>	<i>No heart attack</i>
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



PennState

Clinical and Translational Science Institute

Data Science for Researchers and Scholars

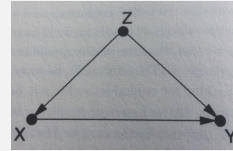
Vasant Honavar, Fall 2023

Observations versus experiments

- Suppose you want to see whether a drug helps reduce the risk of heart disease
- 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!

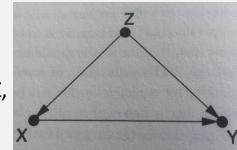
Confounding bias

- 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 $X \rightarrow Y$ is mixed with the spurious correlation induced by the fork $X \leftarrow Z \rightarrow Y$
 - Example: We are testing a drug but give it to patients who are younger, but not to those who are older – age becomes a

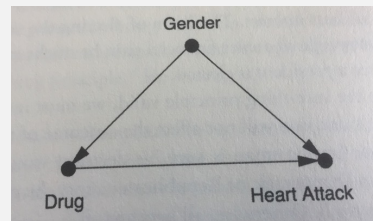


Confounding bias


- 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 chooses to take the drug
- Gender is a confounder that needs to be controlled for in assessing the effect of the drug on heart attack




PennState

Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications

Artificial Intelligence Research Laboratory



PennState


Clinical and Translational Science Institute

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 attack 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!**



PennState

Clinical and Translational Science Institute

Data Science: for Researchers and Scholars

Vasant Honavar, Fall 2023

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


Inferring causal effects from data

- Requires assumptions about the data generating process (DGP)
- Causal graphs allow us to specify the causal structure of DGP
- This allows us to determine causal effects from observations under certain conditions
- Is there a way around making causal assumptions?
- Yes, randomized control trials!




PennState
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory




PennState
Clinical and Translational
Science Institute

A possible solution - randomized control trial (RCT)



The diagram illustrates a randomized control trial (RCT) process. At the top, a group of seven diverse people (represented by colorful icons) are shown. Below them, a hand is shown flipping a coin to randomly assign the participants into two groups. The left group, consisting of four people, is assigned to receive a real drug (represented by a pill bottle with an 'Rx' symbol). The right group, consisting of three people, is assigned to receive a placebo (represented by a pill bottle labeled 'PLACEBO').



PennState
Institute for Computational
and Data Sciences

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023