CompSci 295, Causal Inference

Rina Dechter, UCI

Lecture 1: Introduction

Causal Inference in Statistics, A primer, J. Pearl, M Glymur and N. Jewell Ch1, Why, ch1

Class Information

Course Topics

- Introduction: Causal Hierarchy
- Structural Causal Models
- Identification of Causal Effects
- The Problem of Confounding and the Back-doorCriterion
- Causal Calculus
- Linear Structural Causal Models
- Counterfactuals
- Structural Learning

Textbooks

[P] Judea Pearl, Madelyn Glymour, Nicholas P. Jewell, Causality: Models, Reasoning and Inference, Cambridge Pess, 2016.

[C] Judea Pearl, Causality: Models, Reasoning and Inference, Cambridge Press, 2000.

[W] Judea Pearl, Dana Mackenzie, Causality: Models, Reasoning and Inference, Basic books, 2018.

Class page

Grading

- few homeworks
- Project: Class presentation and a report: Students will present a paper and write a report

http://bayes.cs.ucla.edu/WHY/

Outline

- Simpson Paradox
- The causal Hierarchy
- Structural Causal Models
- Linear Regression

Outline

- Simpson Paradox
- The causal Hierarchy
- Structural Causal Models
- Graphs
- Linear Regression

The Simpson Paradox

- It refers to data in which a statistical association that holds for an entire population is reversed in every subpopulation.
- (Simpson 1951) a group of sick patients are given the option to try a new drug. Among those who took the drug, a lower percentage recover than among those who did not. However, when we partition by gender, we see that more men taking the drug recover than do men not taking the drug, and more women taking the drug recover than do women not taking the drug! In other words, the drug appears to help men and help women, but hurt the general population.
- Example 1.2.1 We record the recovery rates of 700 patients who were given access to the drug. 350 patients chose to take the drug and 350 patients did not. We got:

Table 1.1 Results of a study into a new drug, with gender taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

The Simpson Paradox

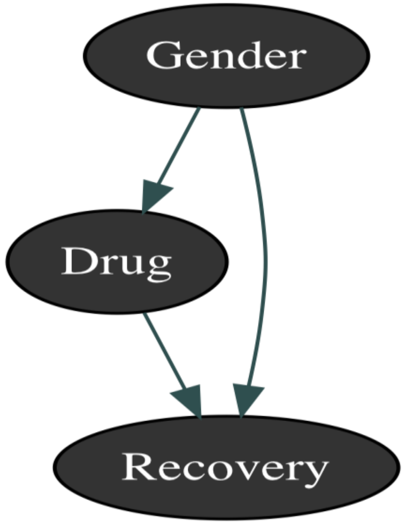
- Example 1.2.1 We record the recovery rates of 700 patients who were given access to the drug. 350 patients chose to take the drug and 350 patients did not. We got:

Table 1.1 Results of a study into a new drug, with gender taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- The data says that if we know the gender of the patient we can prescribe the drug, but if not we should not…. Which is ridiculous.
- So, given the results of the study, should the doctor prescribe the drug for a man? For a woman? Or when gender is unknown?
- **The answer cannot be found in the data!! We need to know the story behind the data- the causal mechanism that lead to, or generated the results we see.**

Simpson’s Paradox

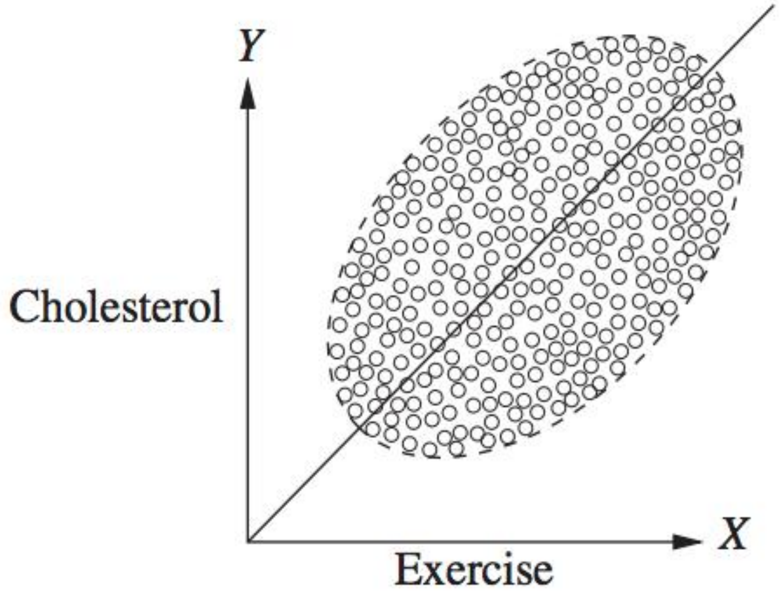


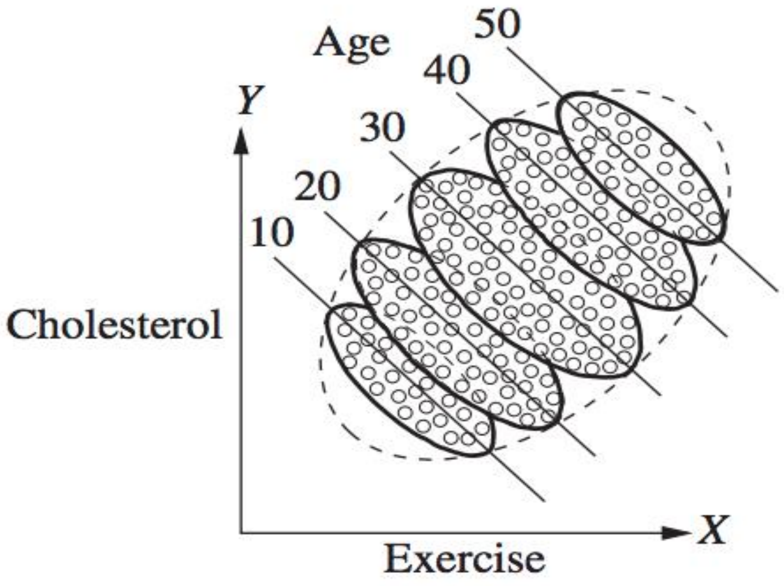
- Also known as Simpson’s Reversal
- No statistical method that can aid our understanding!

	Drug	No drug
Men	81/87 recovered (**93%**)	234/270 recovered (**87%**)
Women	192/263 recovered (**73%**)	55/80 recovered (**69%**)
Combined	273/350 recovered (**78%**)	289/350 recovered (**83%**)

The Simpson Paradox

- The same phenomenon with continuous variables. Example: Impact of exercise on Cholesterol for different age groups:





- Because, Age is a common cause of both treatment (exercise) and outcome (cholesterol). So we should look at the age-segregated data in order to compare same-age people, and thereby eliminate the possibility that the high exercisers in each group we examine are more likely to have high cholesterol due to their age, and not due to exercising.

The Simpson Paradox

- Segregated data is not always the right way. What if we record blood (BP) pressure instead of gender?
- We know that drug lower blood pressure but also has a toxic effect.

- Would you recommend the drug to a patient?

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

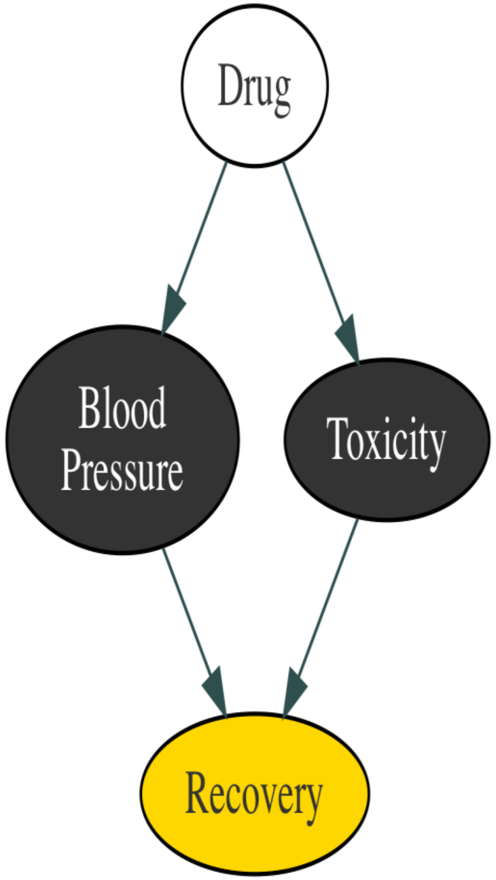
	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- In the general population, the drug might improve recovery rates because of its effect on blood pressure. But in the subpopulations—the group of people whose post-treatment BP is high and the group whose post-treatment BP is low—we of course would not see that effect; we would only see the drug's toxic effect.
- In this case the aggregated data should be consulted.
- **Same data opposite conclusions!!!**

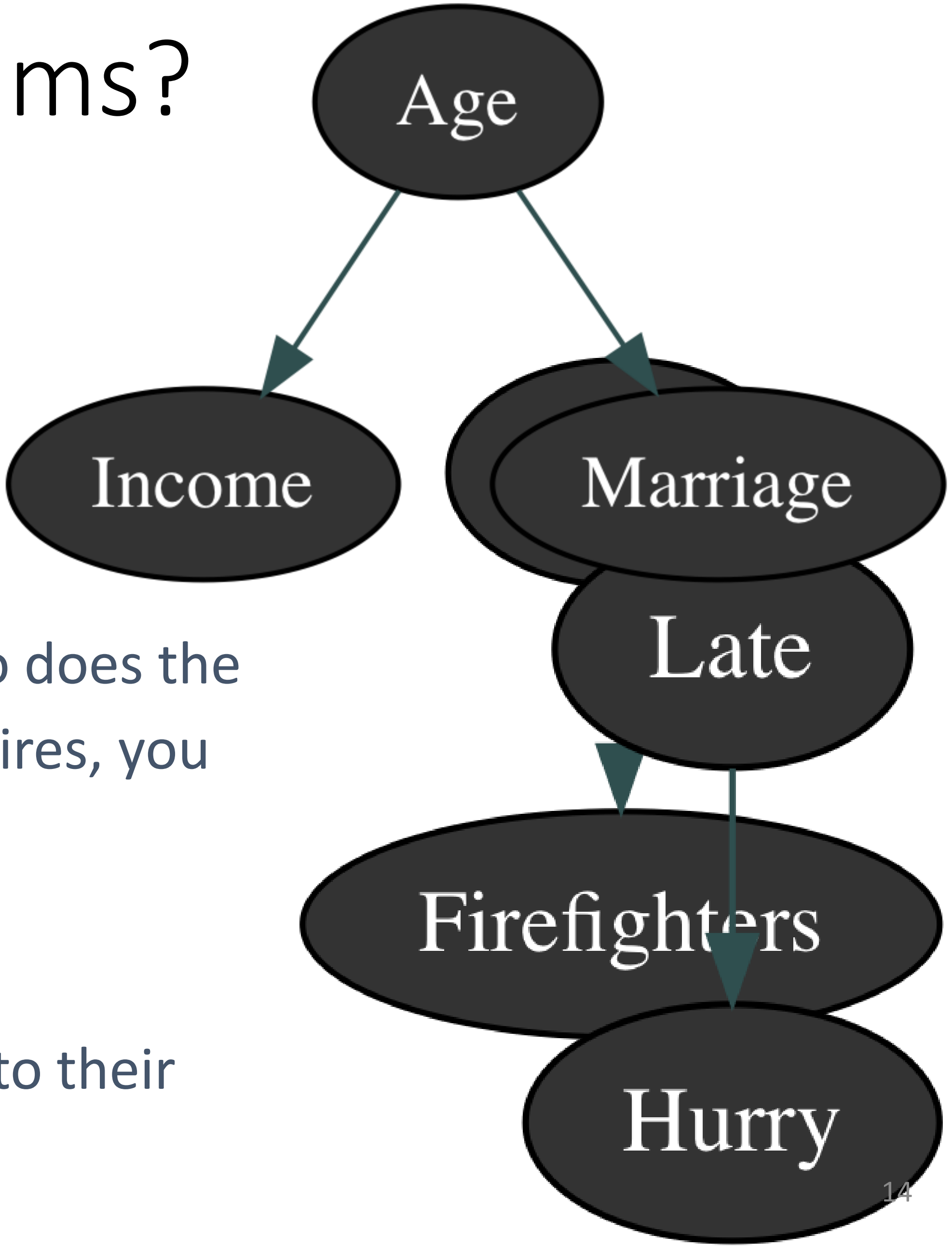
Simpson's Paradox (Aggregated)

- Segregation not good
- Drug lowers blood pressure
- Also has a toxic effect
- 3 groups of people
  - BP high *after* treatment
  - BP low *before and after** treatment
  - Everyone else
- Segregating causes selection bias

	No drug	Drug
Low BP	81/87 recovered (**93%**)	234/270 recovered (**87%**)
High BP	192/263 recovered (**73%**)	55/80 recovered (**69%**)
Combined	273/350 recovered (**78%**)	289/350 recovered (**83%**)



What is Wrong With These Claims?



- "Data show that income and marriage have a high positive correlation. Therefore, your earnings will increase if you get married."

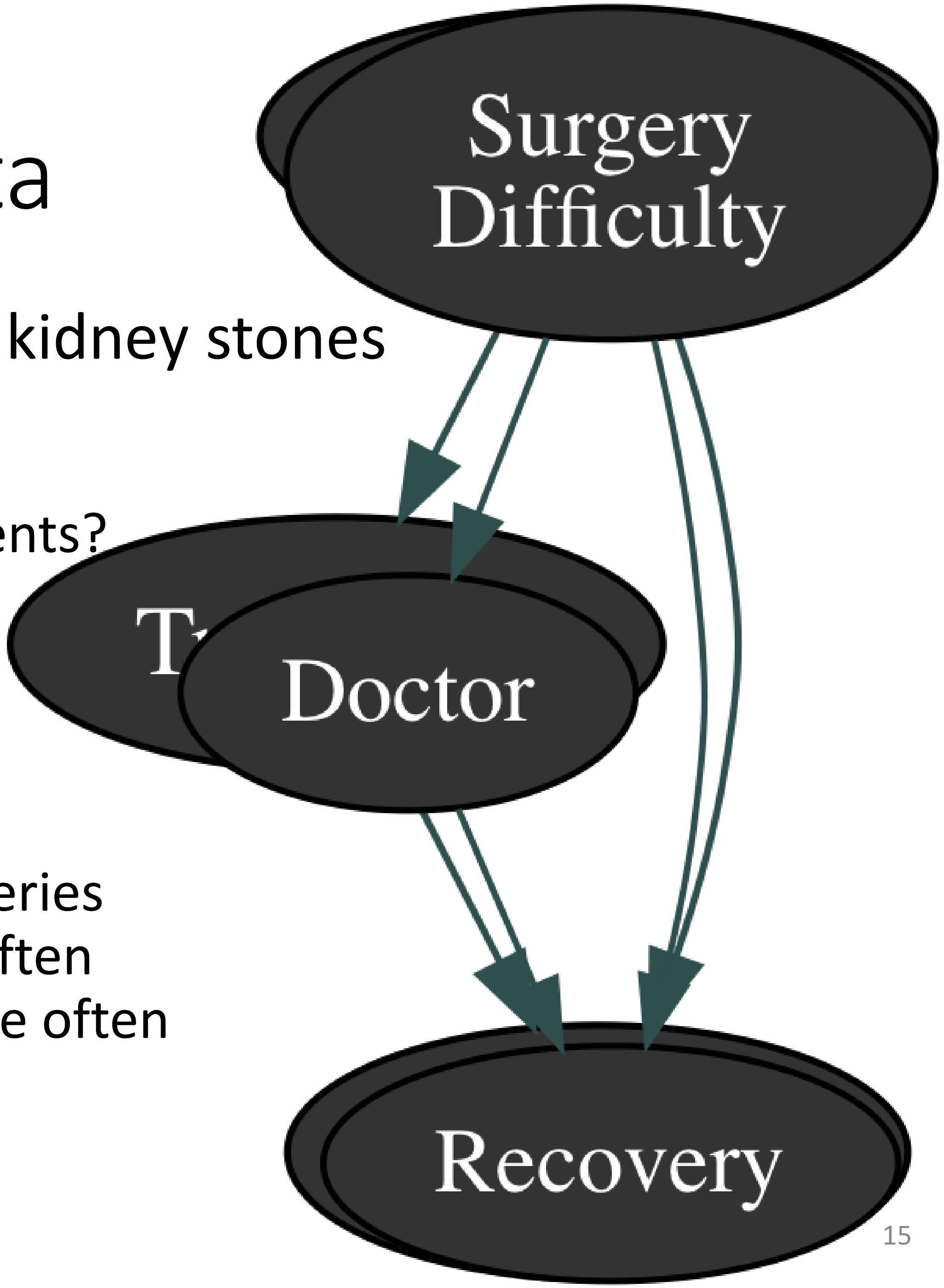
- "Data show that as the number of fires increase, so does the number of firefighters. Therefore, to cut down on fires, you should reduce the number of firefighters."

- "Data show that people who hurry tend to be late to their meetings. Don't hurry, or you'll be late."

Aggregated or Segregated Data

- Doctors choose between 2 treatments for kidney stones
  - Treatment A → large/severe stones
  - Treatment B → small stones
  - What might be reasons for different treatments?
  - Aggregate or segregated data?
  - What does DAG look like?

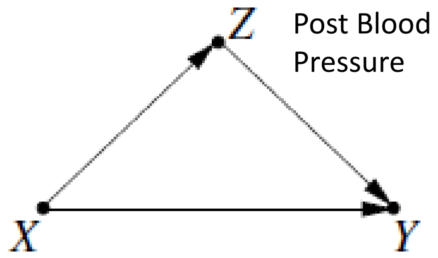
- 2 doctors perform 100 surgeries each
  - Some very difficult and some very easy surgeries
  - Doctor 1 performs easy surgeries far more often
  - Doctor 2 performs difficult surgeries far more often
  - What might be going on?
  - Aggregate or segregated data?
  - What does DAG look like?

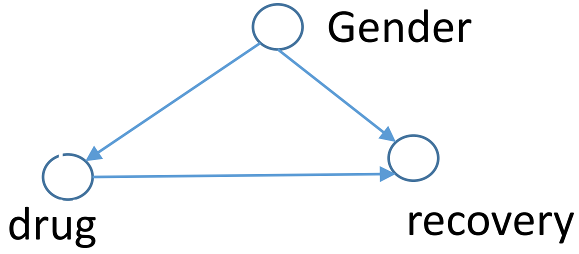


scott

The Simpson Paradox

- The fact that treatment affect BP and not the opposite was not in the data. Indeed in Statistics it is often stressed that “correlation is not causation”, so there is no statistical method that can determine the causal story from the data alone. Therefore, there is no statistical method that can aid in the decision.





- We can make causal assumptions because we know that drug cannot affect gender. “treatment does not cause sex” cannot be expressed in the data.
- So, what do we do? How can we make causal assumptions and make causal inferences?

The Simpson Paradox SCM (Structural Causal Model)

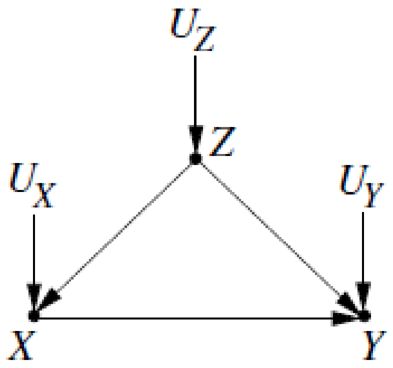


Figure 3.3: A graphical model representing the effects of a new drug, with Z representing gender, X standing for drug usage, and Y standing for recovery

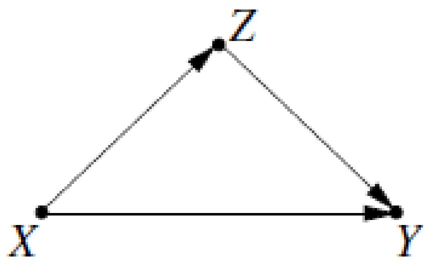


Figure 3.5: A graphical model representing the effects of a new drug, with X representing drug usage, Y representing recovery, and Z representing blood pressure (measured at the end of the study). Exogenous variables are not shown in the graph, implying that they are mutually independent

For Causal Inference We Need:

1. A working definition of “causation”
2. A method by which to formally articulate causal assumptions—that is, to create causal models
3. A method by which to link the structure of a causal model to features of data
4. A method by which to draw conclusions from the combination of causal assumptions embedded in a model and data.

Outline

- Simpson Paradox
- The causal Hierarchy
- Structural Causal Models
- Linear Regression

Motivating Quotes (Book of why)

Adam and Eve:

- When God asks: “Have you eaten from the tree which I forbade you?”
- Adam answers: The woman you gave me for a companion, she gave me fruit from the tree and I ate.
- “What is this you have done?” God asks Eve.
- She replies: “The serpent deceived me, and I ate.”

- God asked for the facts, and they replied with explanations
- Causal explanations, not dry facts, make up the bulk of our knowledge.
- Satisfying our craving for explanation should be the cornerstone of machine intelligence.

- On Machine Learning: no machine can derive explanations from raw data. It needs a push

Hunters Example

- Planning requires imagining the consequences of action (Examples: **hunters of the ICE Age**)

- To imagine and compare the consequences of several hunting strategies. To do this, it must possess, consult, and manipulate a mental model of its reality. Here is how we might draw such a mental model:

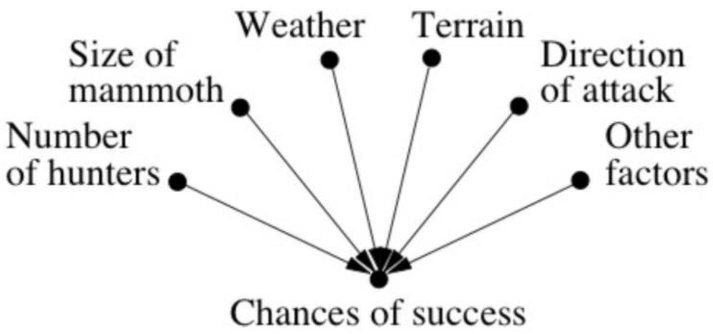
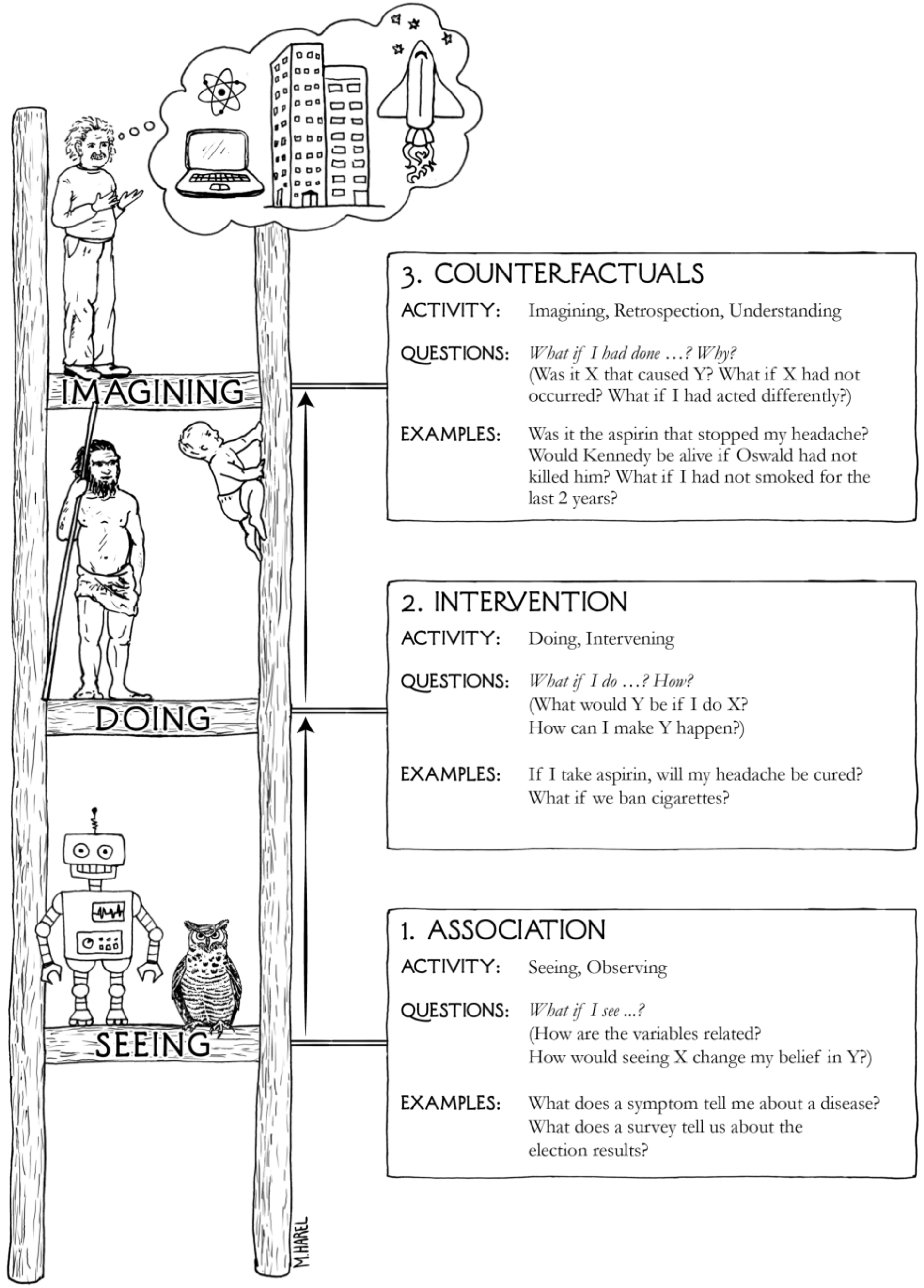


Figure 1. Perceived causes of a successful mammoth hunt.

Ladder of causation: there are at least three distinct levels that need to be conquered by a causal learner: seeing, doing, and imagining.

Ladder of Causation



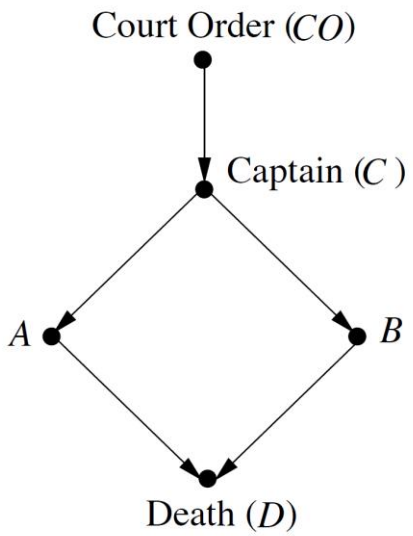
seeing, doing, and imagining.

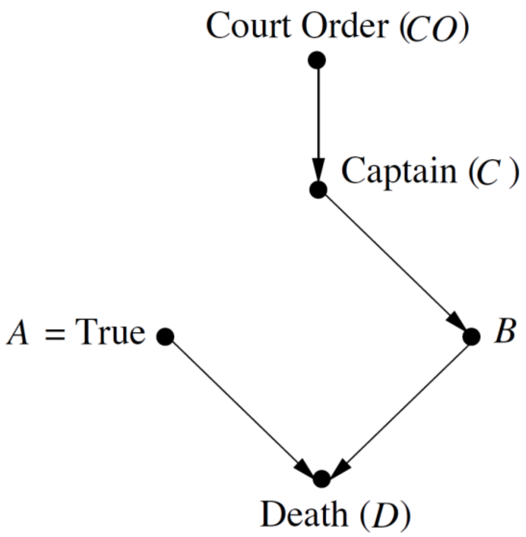
- Most animals, learning machines are on the first rung, learning from association.
- Tool users, such as early humans, are on the second rung, if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge.
- On the top rung, counterfactual learners can imagine worlds that do not exist and infer reasons for observed phenomena.

Darwiche 2017: "Human-Level Intelligence or Animal-Like Abilities?"

The Firing Squad

The story: Suppose that a prisoner is about to be executed. First, the court has to order the execution. The order goes to a captain, who signals the soldiers on the firing squad (A and B) to fire.





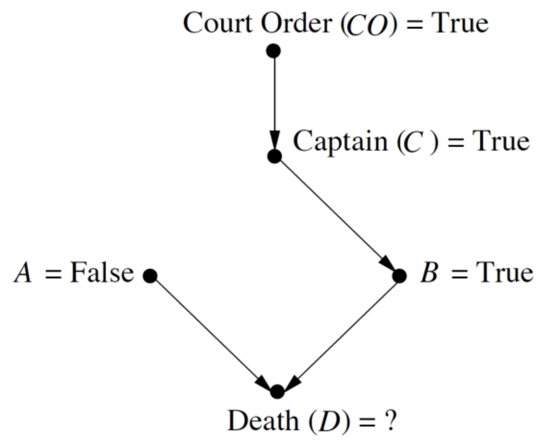
(CO, C, A, B, D) is a true/false variable

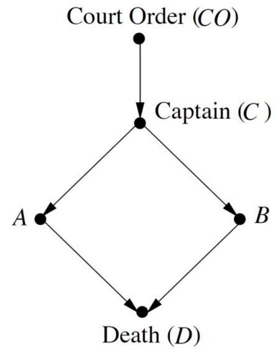
Ladder 1: If the prisoner is dead, does that mean the court order was given?

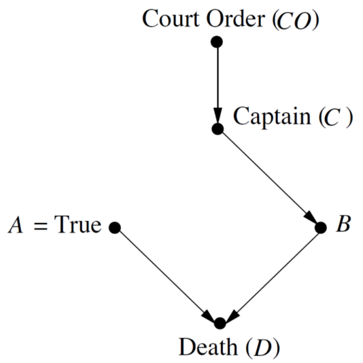
Yes. Logic
Alternatively, suppose we find out that A fired.
What does that tell us about B? Yes.

Ladder 2: we can ask questions of intervention.
What if soldier A decides on his own initiative to fire, without waiting for the captain's command? Will the prisoner be dead or alive?

The Firing Squad, Counterfactuals







- **Ladder 3:** Suppose the prisoner is lying dead on the ground. Using level one implies that A shot, B shot, the captain gave the signal, and the court gave the order.
- If, contrary to fact, A had decided not to shoot, would the prisoner be alive?
- This question requires us to compare the real world with a fictitious and contradictory world where A didn't shoot.
- In the fictitious world, the arrow leading into A is erased and A is set to False, but the past history of A stays the same as it was in the real world.

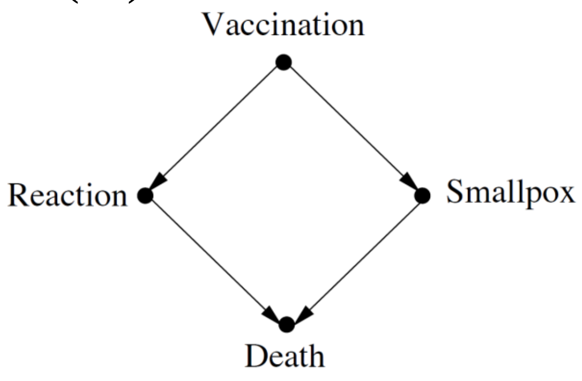
In the firing squad example we ruled out uncertainties: maybe the captain gave his order a split second after rifleman A decided to shoot, maybe rifleman B's gun jammed, etc. To handle uncertainty we need information on how likely the alternatives are to occur.

Smallpox Vaccination: Adding Uncertainties

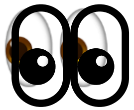
Suppose that out of 1 million children, 99 percent are vaccinated and 1 percent are not. If a child is vaccinated, he or she has 1 chance in 100 of developing a reaction, and the reaction has 1 chance in 100 of being fatal. On the other hand, he or she has no chance of developing smallpox. Meanwhile, if a child is not vaccinated, he or she obviously has zero chance of developing a reaction to the vaccine, but he or she has 1 chance in 50 of developing smallpox. Finally, let's assume that smallpox is fatal in one out of 5 cases.

Data: Out of 1 million children, 990 thousand get vaccinated; 9,900 get the reaction; and 99 die from the reaction. Meanwhile, 10 thousand don't get vaccinated, 200 get smallpox, and 40 die from the disease. In summary, more children die from vaccination (99) than from the disease (40).

- We now ask the counterfactual question: What if we had set the vaccination rate to 0?
- we can conclude that out of 1 million children, 20 thousand would have gotten smallpox and 4,000 would have died.
- Comparing the counterfactual world with the real world, we see that the cost of not vaccinating was the death of 3,861 children (the difference between 4,000 and 139).



Big picture - Pearl's Causal Hierarchy (PCH)

Level (Symbol)	Typical Activity	Typical Question	Examples
1  Association $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?

Big picture - Pearl’s Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Question	Examples
1 👀 Association $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
2 💪 Intervention $P(y \mid do(x), c)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured?

Big picture - Pearl's Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Question	Examples
1 Association $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
2 Intervention $P(y \mid do(x), c)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured?
3 Counterfactual $P(y_x \mid x', y')$	Imagining, Retrospection	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?

Big picture - Pearl's Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Question	Examples
1 Association $P(y \mid x)$	Seeing ML - (Un)Supervised Deep Net, Bayes net, Hierarchical Model, DT	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
2 Intervention $P(y \mid do(x), c)$	Doing ML - Reinforcement Causal Bayes Net, MDP, POMDP	What if? What if I do X?	What if I take aspirin, will my headache be cured?
3 Counterfactual $P(y_x \mid x', y')$	Imagining, Retrospection Structural Causal Model	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?

Big picture - Pearl's Causal Hierarchy

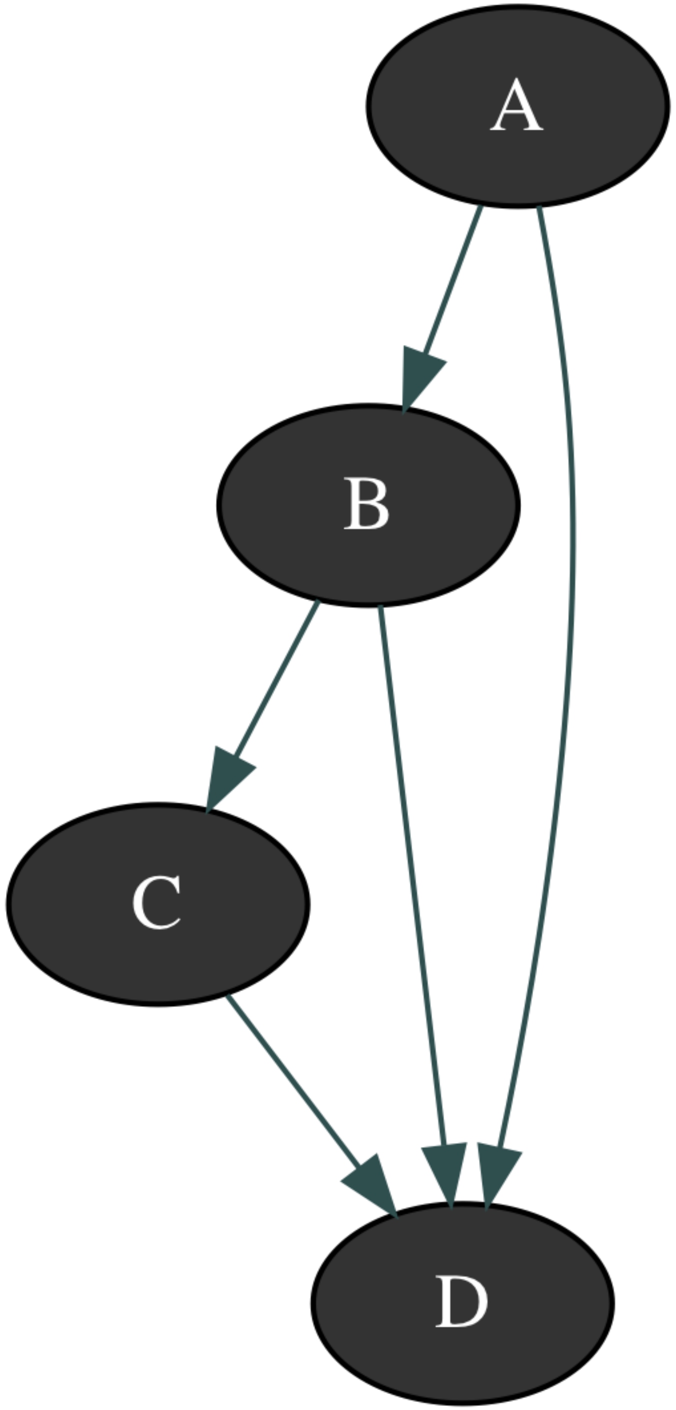
Level (Symbol)	Typical Activity	Typical Question	Examples
1 👀 Association $P(y \mid x)$	Seeing ML - (Un)Supervised Deep Net, Bayes net, Hierarchical Model, DT	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
2 💪 Intervention $P(y \mid do(x), c)$	Doing ML - Reinforcement Causal Bayes Net, MDP, POMDP	What if? What if I do X?	What if I take aspirin, will my headache be cured?
3 🧠 Counterfactual $P(y_x \mid x', y')$	Imagining, Retrospection Structural Causal Model	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?

Outline

- Simpson Paradox
- The causal Hierarchy
- Structural Causal Models
- Linear Regression

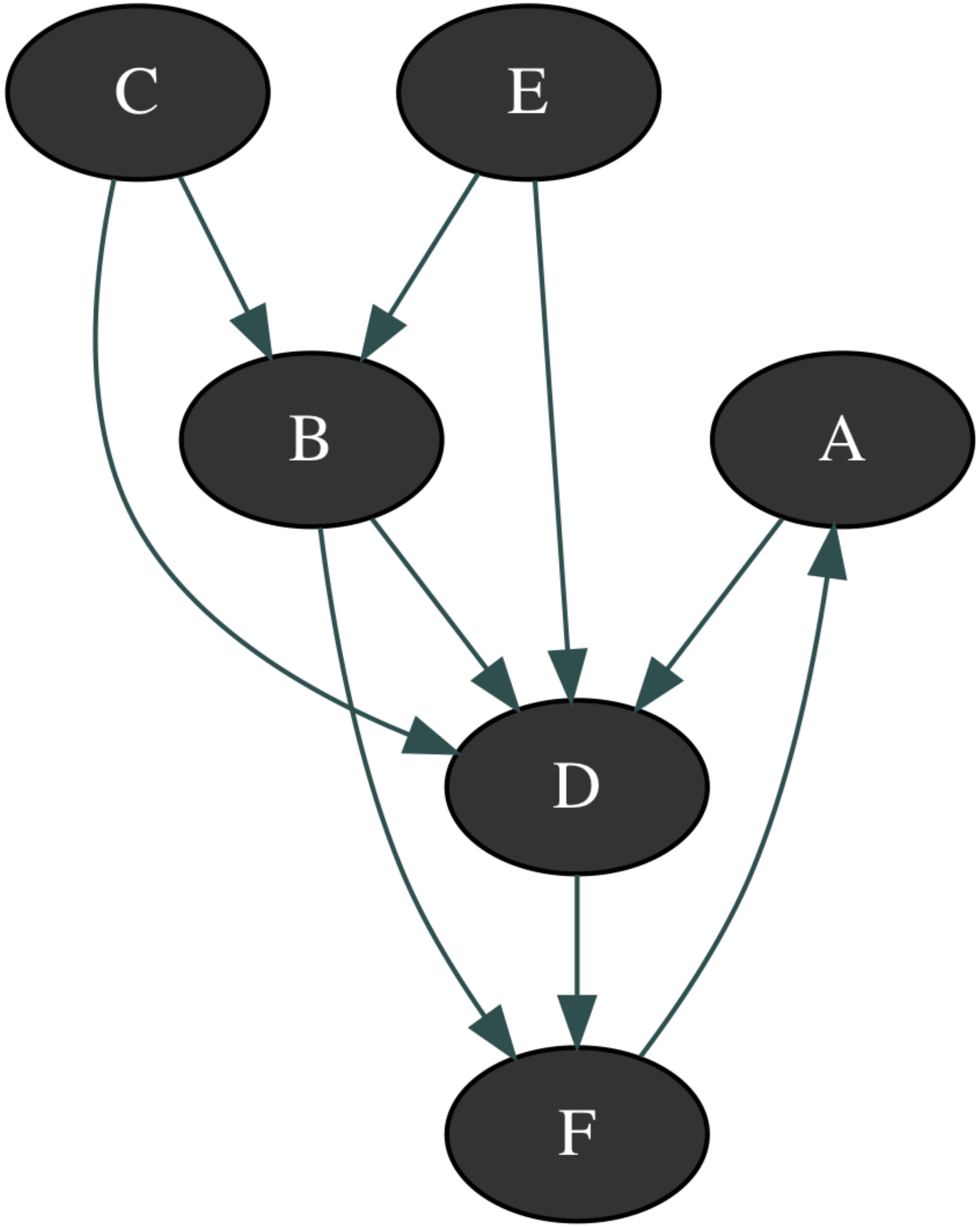
Graphs

- Parents
- Ancestors
- Children
- Descendants
- Paths
- Directed paths



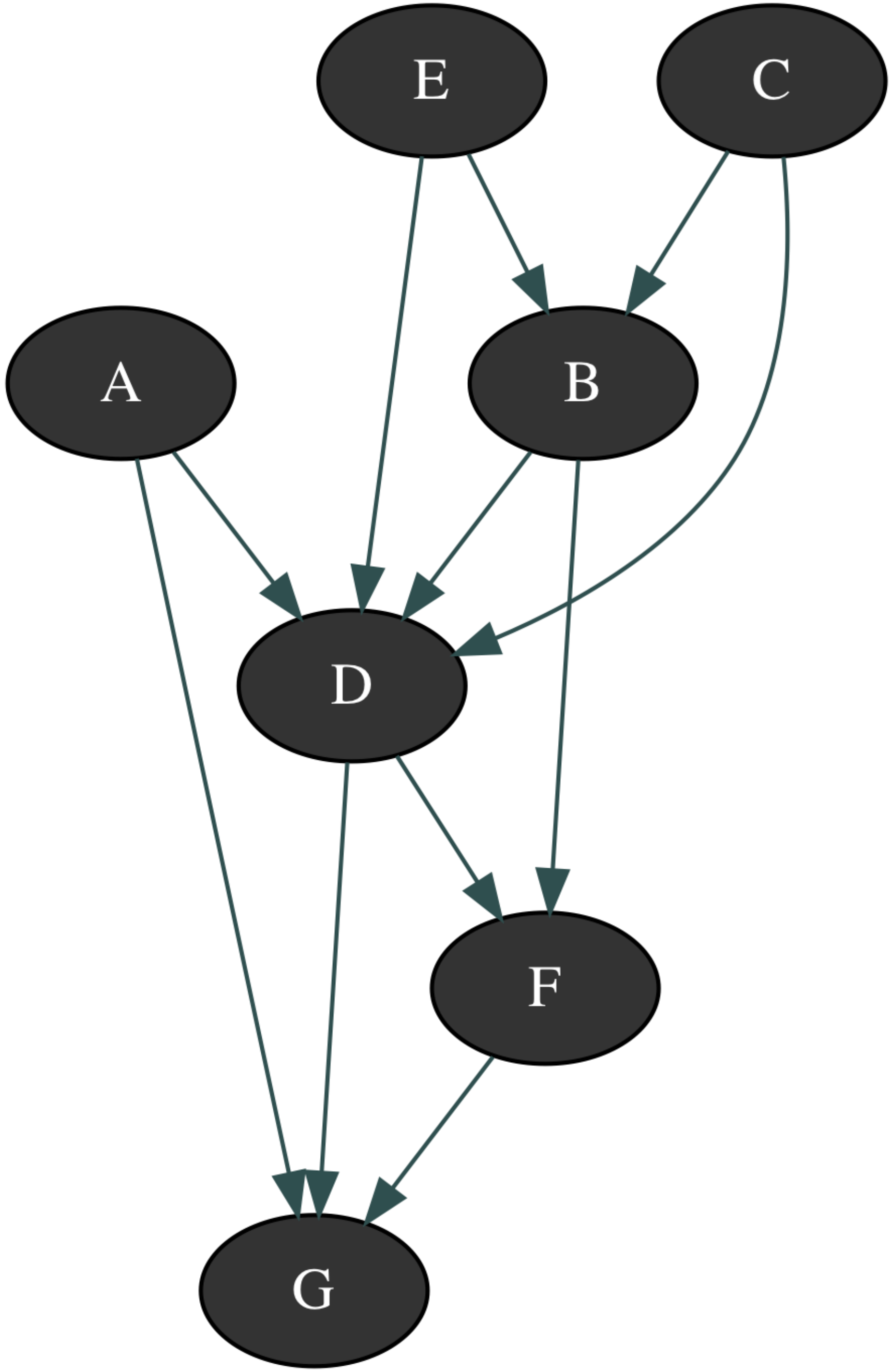
Cyclic?

A → D → F → A → …



Cyclic?

No directed cycles



Outline

- Simpson Paradox
- The causal Hierarchy
- Structural Causal Models
- Product form of Graphical models.
- Linear Regression

Structural Causal Models (SCM), M

A structural causal model describes how nature assigns values to variables of interest.

- Two sets of variables, U and V and a set of functions **F: (U,V,F)**
- Each function assigns value to a variable in V based on the values of the other variables.
- We say that Variable X is a direct cause of Y if it appears in the function of Y.
- U are exogenous variables (external to the model. We do not explain how they are caused)..
- Variables in U have no parents.

SCM 1.5.1 (Salary Based on Education and Experience)

$$U = \{X, Y\}, \quad V = \{Z\}, \quad F = \{f_Z\}$$
$$f_Z : Z = 2X + 3Y$$

Z- salary, X – years in school, Y – years in the profession

Structural Causal Models (SCM), M

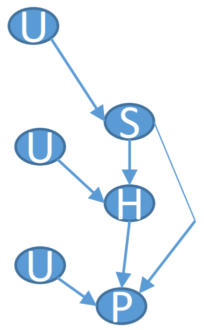
- Every SCM is associated with a graphical causal model.
- The graphical model G for an SCM M contains one node for each variable in M. If, in M, the function f_X for a variable X contains variable Y (i.e., if X depends on Y for its value), then, in G, there will be a directed edge from Y to X.
- We will deal primarily with SCMs that are acyclic graphs (DAGs).
- A graphical definition of causation: If, in a graphical model, a variable X is the child of another variable Y then Y is a direct cause of X; if X is a descendant of Y , then Y is a potential cause of X .

SCM 1.5.1 (Salary Based on Education and Experience)

$$U = \{X, Y\}, \quad V = \{Z\}, \quad F = \{f_Z\}$$
$$f_Z : Z = 2X + 3Y$$

X and Y are direct causes for Z

Structural Causal Models (SCM)



SCM 1.5.2 (Basketball Performance Based on Height and Sex)

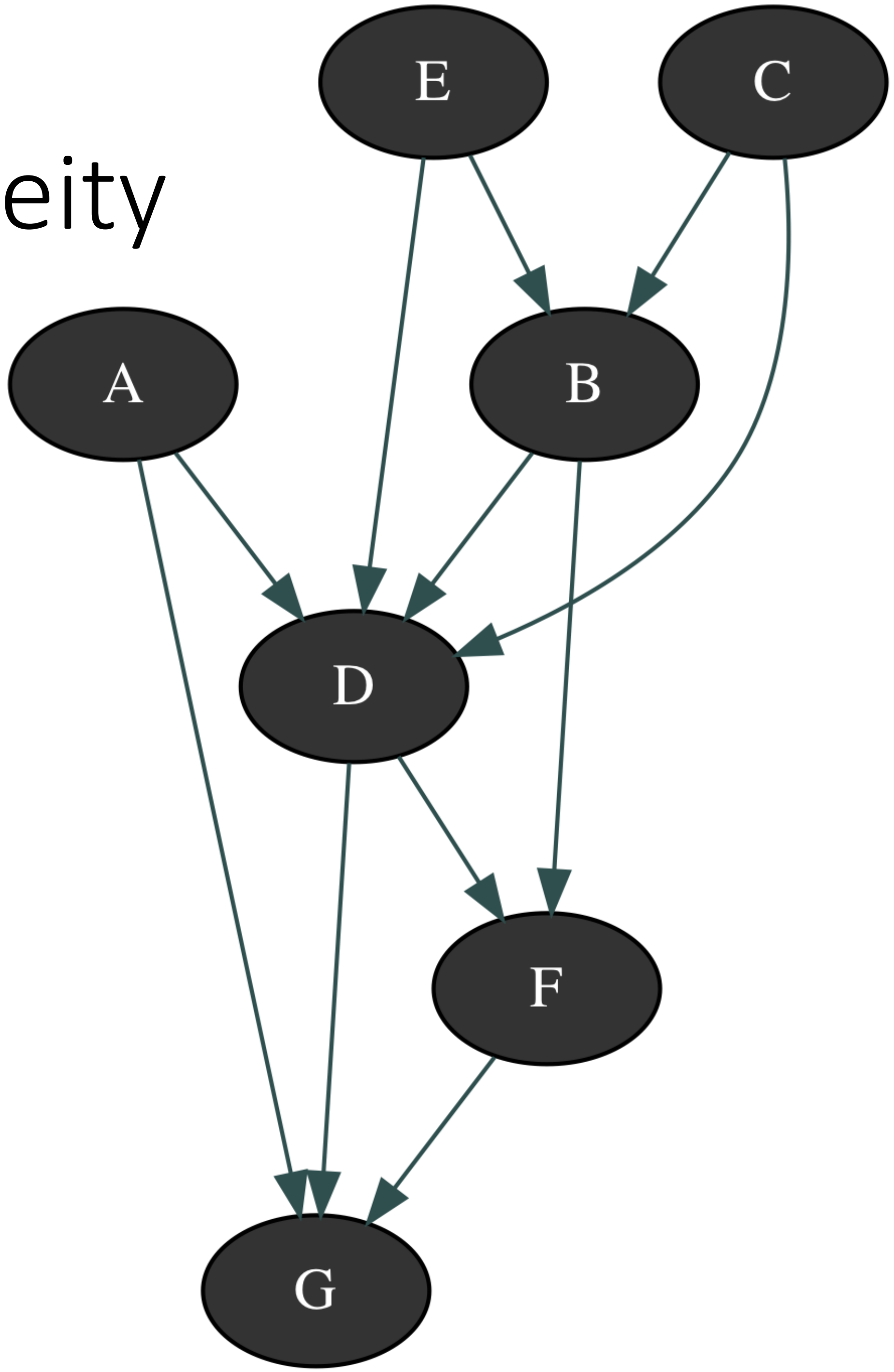
$$
\begin{aligned}
V = \ & \{\text{Height}, \text{Sex}, \text{Performance}\}, \quad U = \{U_1, U_2, U_3\}, \quad F = \{f1, f2\} \\
& \text{Sex} = U_1 \\
& \text{Height} = f_1(\text{Sex}, U_2) \\
& \text{Performance} = f_2(\text{Height}, \text{Sex}, U_3)
\end{aligned}
$$

U are unmeasured terms that we do not care to name. Random causes we do not care about.
U are sometime called error terms.

The graphical causal model provides lots of information about what is going on: X causes Y and Y causes Z

More on Exogeneity/Endogeneity

- Exogenous variables?
  - {A, E, C}
- Endogenous variables?
  - {B, D, F, G}
- Functions and their inputs?
  - f_A(?)
    - No functions for exogenous variables
  - f_B(C, E)
  - f_D(A, B, C, E)
  - f_F(B, D)
  - f_G(A, D, F)



Outline

- Simpson Paradox
- The causal Hierarchy
- Structural Causal Models
- Product form of Graphical models.
- Linear Regression

Data: Joint Probability Distributions

- X = eye color (amber, blue, brown, gray, green, hazel, red)
- Y = hair color (blonde, brown, black, red, white, purple)
-

X	Y	P(x, y)
amber	blonde	0.03
amber	brown	0.09
…	…	…

- How many rows in our joint distribution table?
- What if there were 10 variables?
- How do SCMs help?

Probability distibutions of SCM; Product Decomposition in Bayesian Networks

- How to calculate joint probability from SCM
- $P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$
  - pa_i are the parents of x_i
- X→Y←Z, how to calculate P(x, y, z)?
  - $P(x, y, z) = P(X = x) \cdot P(Y = y | X = x, Z = z) \cdot P(Z = z)$
- Why is this important?
- Y→X→Z, X = eye color, Y = age (between 7 and 14), Z = hair color
  - How many rows of data do we need for joint probability?
    - $8 \cdot 7 \cdot 6$
  - How many rows of data do we need to be able to *calculate* joint probability?
    - $8 + 8 \cdot 7 + 6 \cdot 7$

Curse of Dimensionality

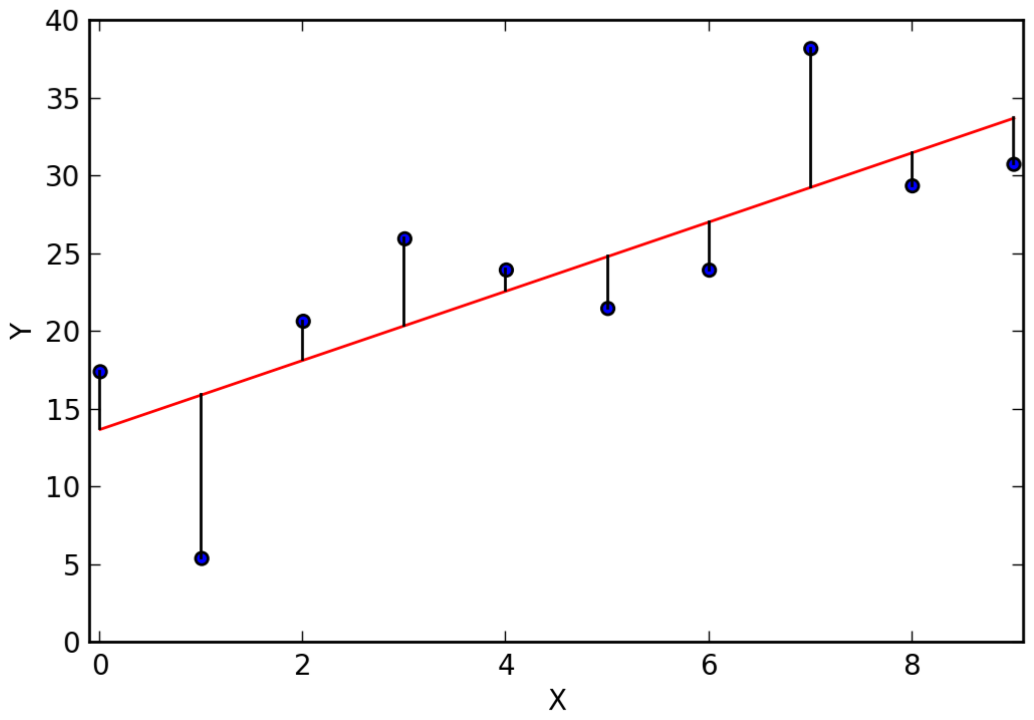
- 5 binary variables: X, Y, Z, W, V
- What is the size of the probability table?
  - $2^5 - 1$
- Assume we recorded 100 samples
  - How many samples per X, Y, Z, W, V bucket?
    - 100/32 ≈ 3
  - Not very accurate
- X → Y → Z → W → V
  - We just need P(x), P(y|x), P(z|y), P(w|z), and P(v|w)
  - P(x) uses entire 100 samples
  - P(y|x) gets split how? How many buckets?
    - 4 buckets ⇒ 100/4 = 25 samples per bucket on average

Outline

- Simpson Paradox
- The causal Hierarchy
- Structural Causal Models
- Product form of Graphical models.
- Linear Regression

Regression

- Predict the value of Y based on X
- Used in Machine Learning too
- Can't we just use E[Y | X = x]?
  - Yes, if know it or can calculate it
- How to create a regression line?
  - Plot data values of X, Y
  - "Fit" them to y = mx + b
  - The least square regression is the line tht minimize the sum of the squared error average $\sum (y - b - mx)^2$
  - Need to find b and m
    - What do they represent on the graph?



Dice

- Can't get all of data for some things, like a roll of dice
- We have to let sample size be infinity, how?
- X = Dice 1, Z = Dice 2, $Y = X + Z$
- $E[Y|X = x]$?
  - $= E[\text{Dice 2} + X|X = x] = E[\text{Dice 2}] + x = 3.5 + 1.0x$
- Intuitive, right?
- It is easy to see that $E[X|Y = y] = 0.5y$, why?
  - $E[X|Y = y] = E[Y - Z|Y = y] = y - E[Z|Y = y]$
  - $E[X|Y = y] = E[Z|Y = y]$
  - $2 \cdot E[X|Y = y] = y$
  - $E[X|Y = y] = \frac{1}{2}y = 0.5y$

Regression Coefficient

- R_{YX} is slope of regression line of Y on X
- $m = R_{YX} = \sigma_{XY}/\sigma_X^2$
  - $R_{YX} = R_{XY}$?
  - When is it?
- Slope gives correlation
  - Positive number → positive correlation
  - Negative number → negative correlation
  - Zero → independent or non-linear

Multiple Regression

- $y = r_0 + r_1 \cdot x + r_2 \cdot z$
- How do we visualize?
  - 3d plane
- What happens if we hold x at a value?
  - $r_1 \cdot x$ becomes a constant
  - r_2 is now the 2d slope of slice along X-axis
- What happens if we hold z at a value?
  - $r_2 \cdot z$ becomes a constant
  - r_1 is now the 2d slope of slice along Z-axis

Partial Regression Coefficient

- Symbol for regression coefficient of Y on X?
  - R_{YX}
- Symbol for regression coefficient of Y on X when holding Z constant?
  - $R_{YX \cdot Z}$
  - Called **partial regression coefficient**
- What happens when R_{YX} is positive and $R_{YX \cdot Z}$ is negative?
- What are partial regression coefficients in $y = r_0 + r_1 \cdot x + r_2 \cdot z$?
  - r_1 and r_2

Orthogonality Principle

- How do we find partial regression coefficients in $y = r_0 + r_1 \cdot x + r_2 \cdot z$?
- Use statistical variables and add noise term ϵ
  - $Y = r_0 + r_1 X_1 + r_2 X_2 + \epsilon$
- Assume ϵ is not correlated with X_1 or X_2
  - $\sigma_{\epsilon X_1} = 0$ $\sigma_{\epsilon X_2} = 0$
- Called **orthogonality principle**

- Remember covariance: $E[(X - E[X]) \cdot (Y - E[Y])] = E(X \cdot Y) - E(X) \cdot E(Y)$
  - What does orthogonality principle say about that?
  - $E(\epsilon X) = E(\epsilon) \cdot E(X)$
  - What if $E(\epsilon) = 0$?

Orthogonality Principle Example

- X = Dice 1, Y = Dice 1 + Dice 2
- $X = \alpha + \beta \cdot Y + \epsilon$
  - What do we want to find to make predictions of X on Y?
  - α and β
- Take expectation of both sides, assuming $E[\epsilon] = 0$
  - $E[X] = E[\alpha] + \beta \cdot E[Y] + E[\epsilon]$
  - $E[X] = \alpha + \beta \cdot E[Y]$
- Also, multiply by Y and take expectation of both sides
  - $E[X \cdot Y] = E[\alpha Y] + \beta \cdot E[Y^2] + E[\epsilon Y]$
  - But $E[\epsilon Y] = 0$
- Now we have 2 equations with 2 unknowns

Regression Equation

- Solve for α and β in orthogonality principle equations
  - $E[X] = \alpha + \beta \cdot E[Y]$
  - $E[X \cdot Y] = E[\alpha Y] + \beta \cdot E[Y^2] + E[\epsilon Y]$
    - $= \alpha E[Y] + \beta \cdot E[Y^2]$
- Remember $\sigma_Y^2 = E[Y^2] - E[Y]^2$
- $\alpha = E[X] - \beta \cdot E[Y]$
  - $E[X \cdot Y] = (E[X] - \beta \cdot E[Y]) \cdot E[Y] + \beta \cdot E[Y^2]$
  - $E[X \cdot Y] = E[X] \cdot E[Y] - \beta \cdot E[Y]^2 + \beta \cdot E[Y^2]$
  - $E[X \cdot Y] - E[X] \cdot E[Y] = -\beta \cdot E[Y]^2 + \beta \cdot E[Y^2]$
  - $E[X \cdot Y] - E[X] \cdot E[Y] = \beta \cdot (E[Y^2] - E[Y]^2)$
  - $\beta = \sigma_{XY} / \sigma_Y^2$
  - Regression coefficient!
  - Could've simply swapped X and Y!