Intro to Causality

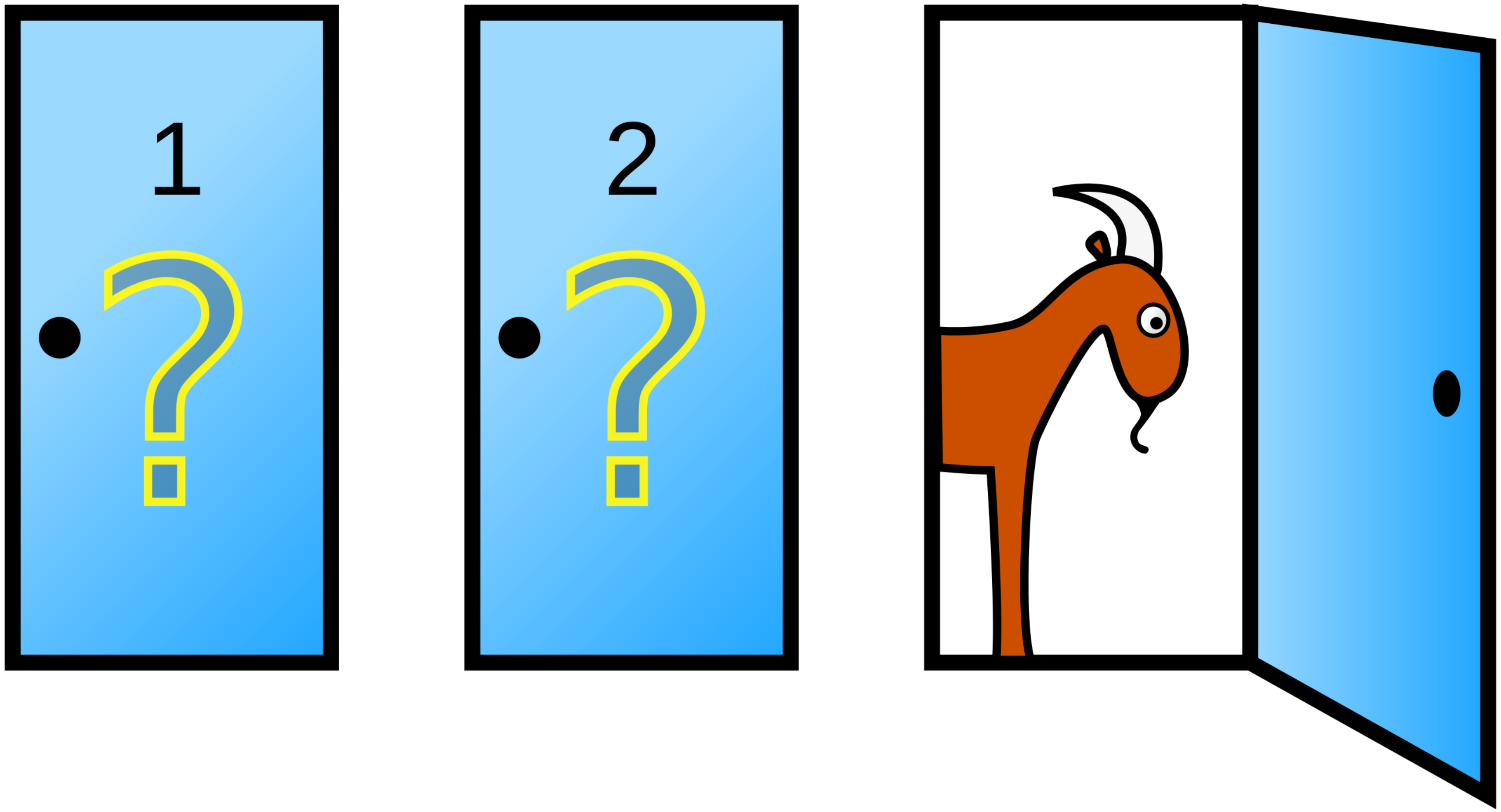
David Madras

October 22, 2019

Simpson's Paradox

Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

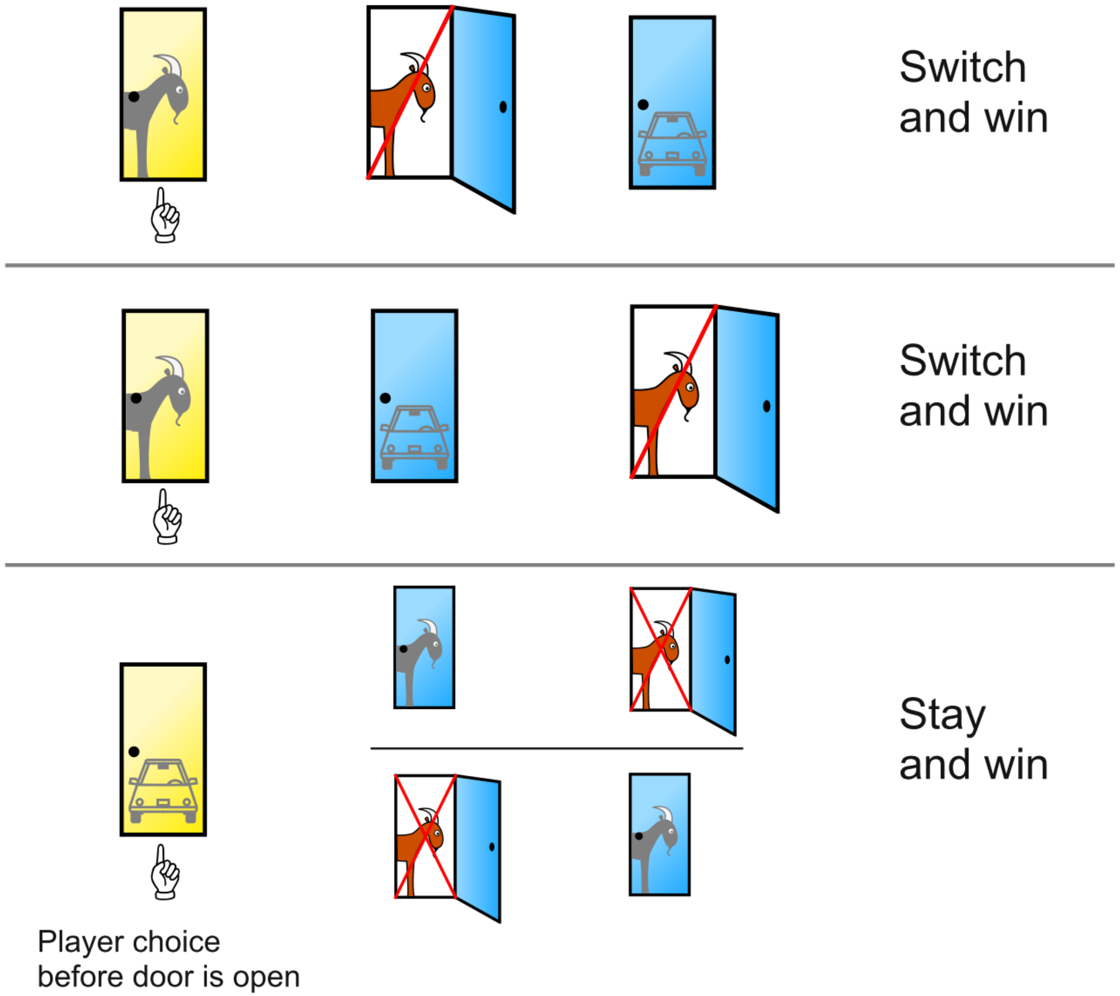
The Monty Hall Problem



The Monty Hall Problem

1. Three doors – 2 have goats behind them, 1 has a car (you want to win the car)
2. You choose a door, but don’t open it
3. The host, Monty, opens *another* door (not the one you chose), and shows you that there is a goat behind that door
4. You now have the option to switch your door from the one you chose to the other unopened door
5. What should you do? Should you switch?

The Monty Hall Problem

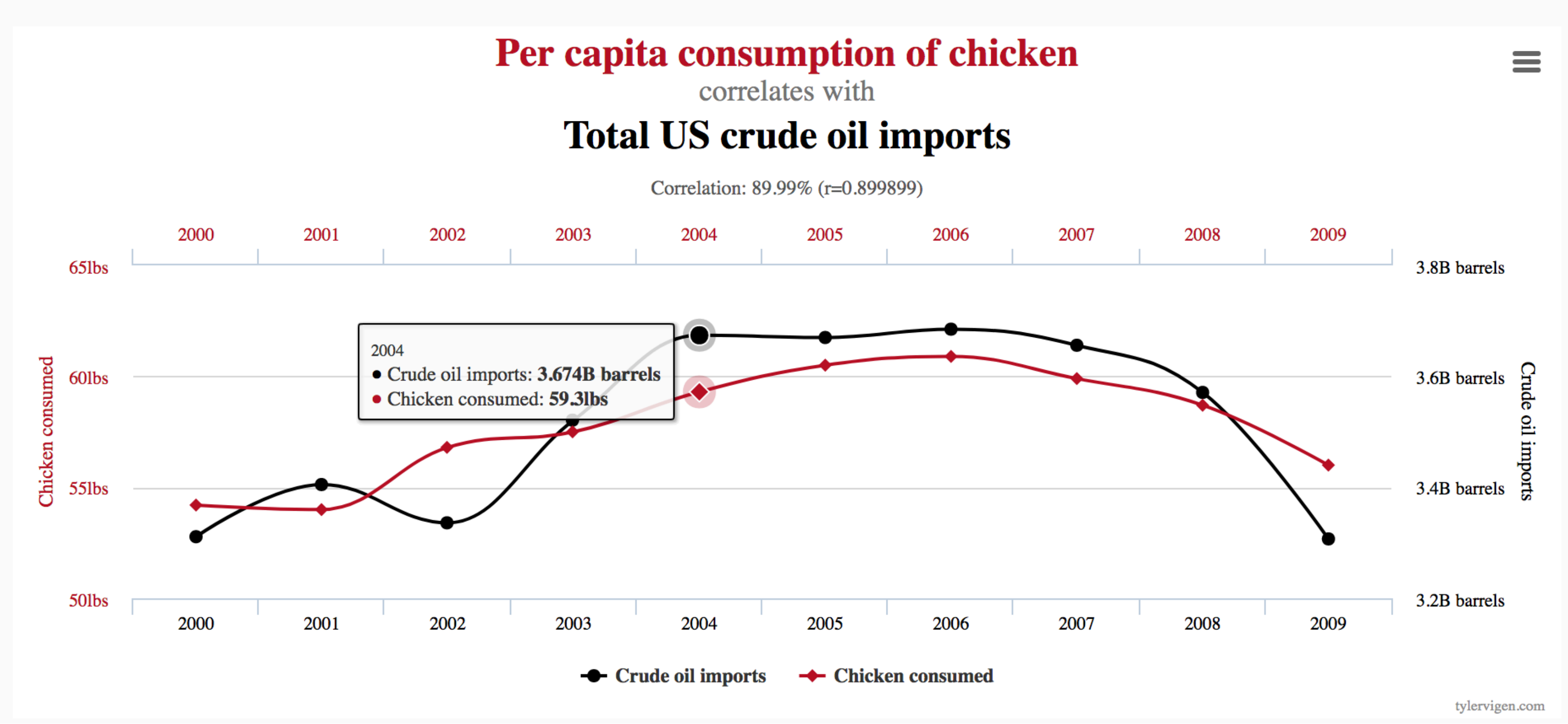


What's Going On?

Causation != Correlation

- In machine learning, we try to learn correlations from data
  - "When can we predict X from Y?"
- In causal inference, we try to model **causation**
  - "When does X **cause** Y?"
- These are not the same!
  - Ice cream consumption **correlates** with murder rates
  - Ice cream does not **cause** murder (usually)

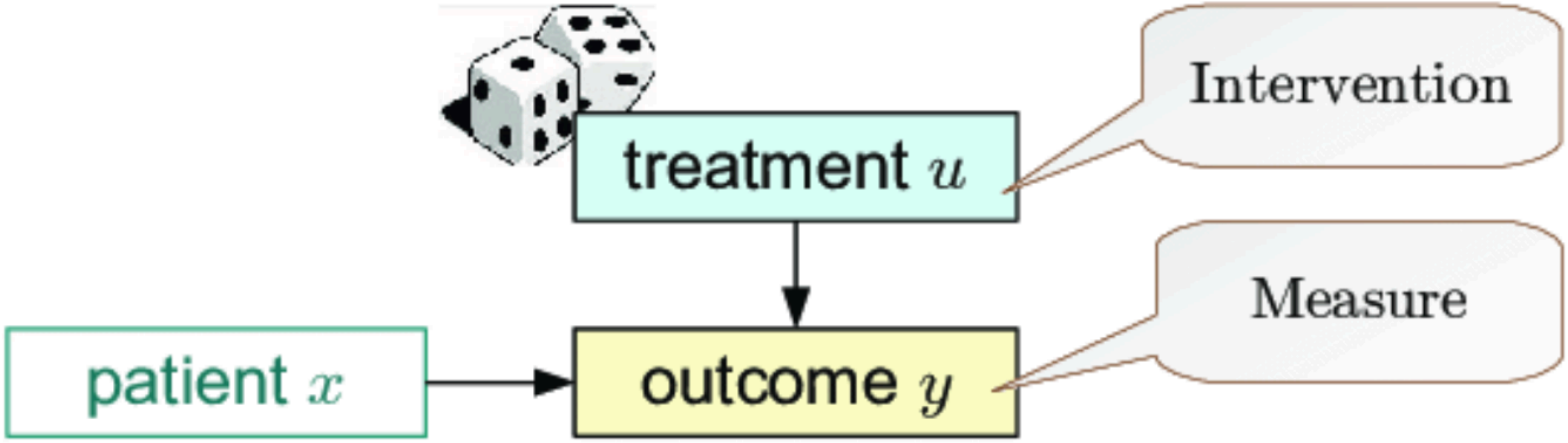
Correlations Can Be Misleading



https://www.tylervigen.com/spurious-correlations

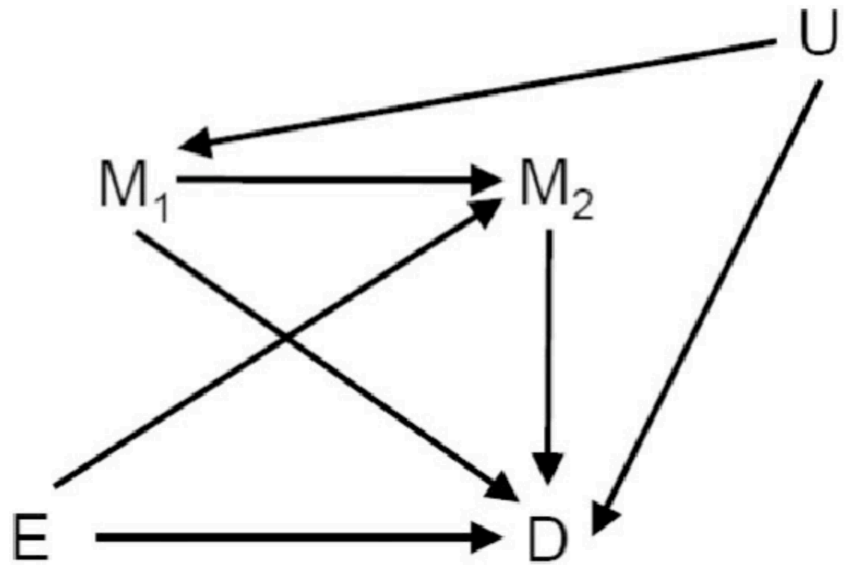
Causal Modelling

- Two options:
    1. Run a **randomized experiment**



Causal Modelling

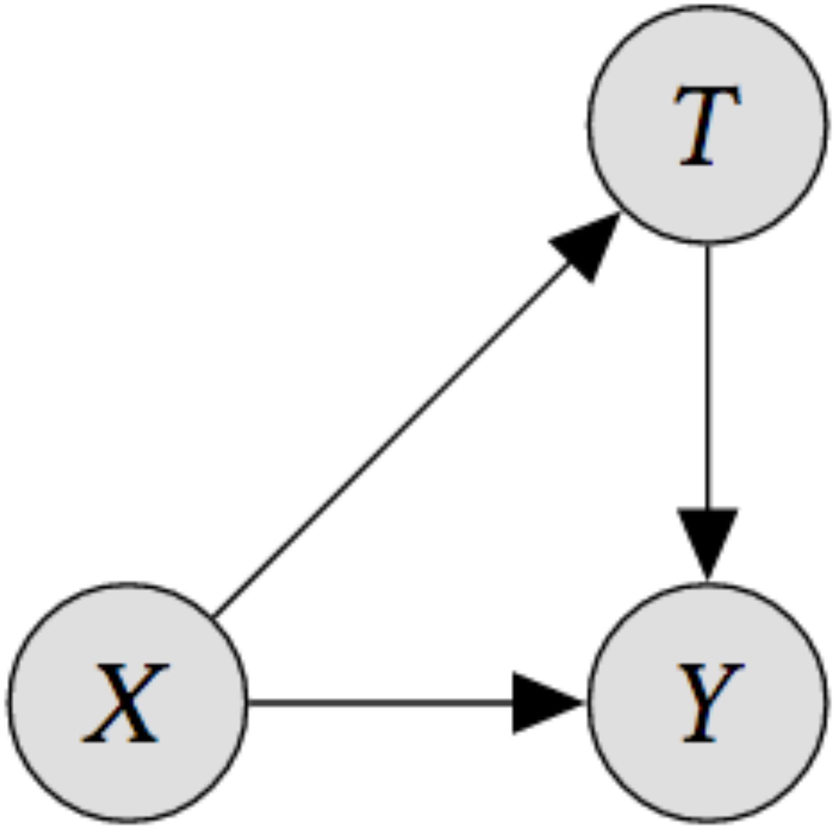
- Two options:
  1. Run a randomized experiment
  2. Make assumptions about **how our data is generated**



Causal DAGs

- Pioneered by Judea Pearl
- Describes generative process of data

$$X = f_X(\epsilon_X)$$
$$T = f_T(X, \epsilon_T)$$
$$Y = f_Y(T, X, \epsilon_Y)$$



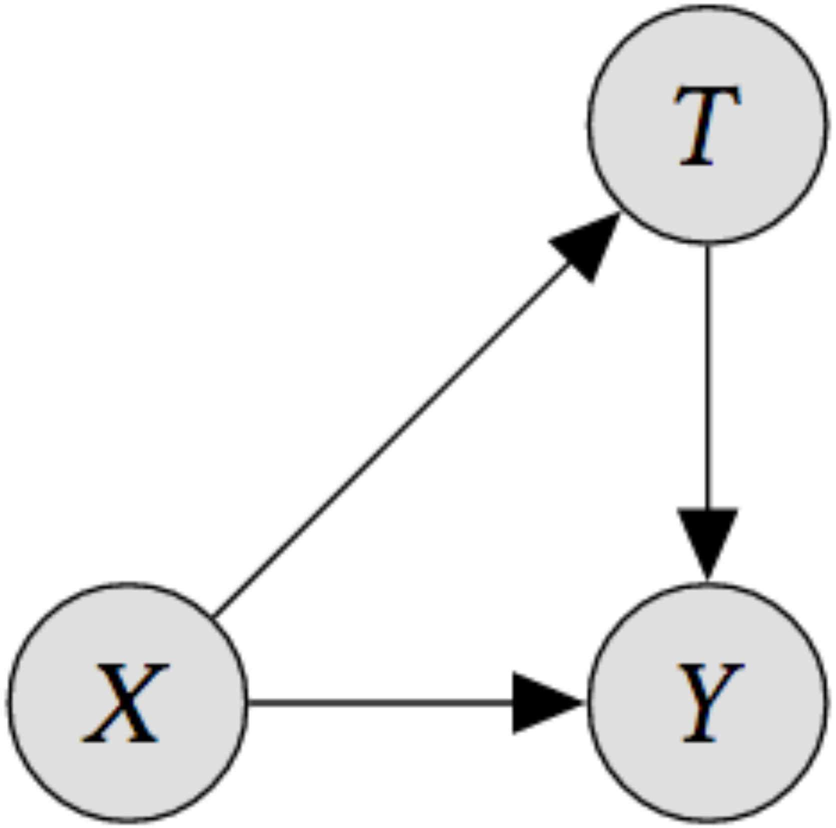
Causal DAGs

- Pioneered by Judea Pearl
- Describes (stochastic) generative process of data

$$X \sim P_X$$

$$T \sim P_T | X$$

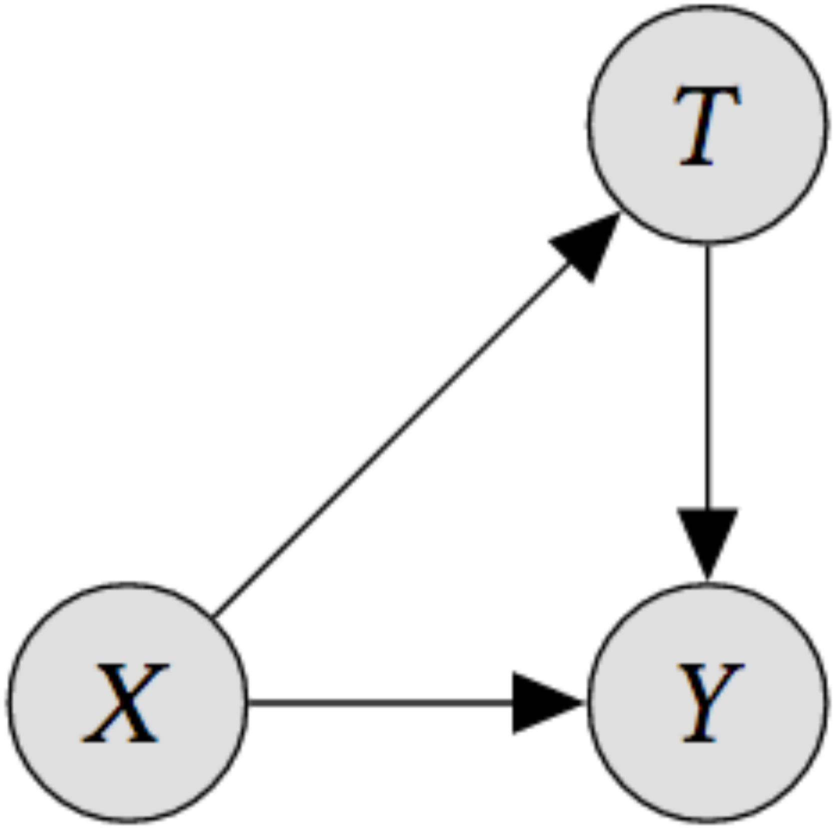
$$Y \sim P_Y | X, T$$



Causal DAGs

- T is a medical treatment
- Y is a disease
- X are other features about patients (say, age)

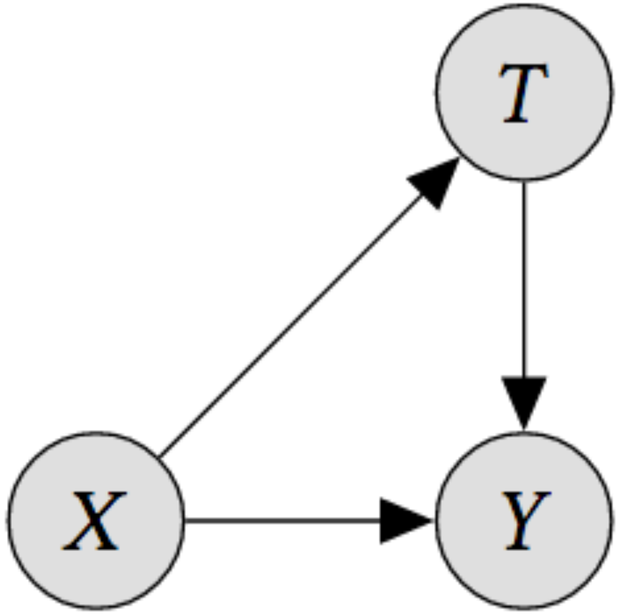
- **We want to know the *causal effect* of our treatment on the disease.**



Causal DAGs

- Experimental data: randomized experiment
  - We decide which people should take *T*
- Observational data: no experiment
  - People chose whether or not to take *T*

- Experiments are expensive and rare
- Observations can be **biased**
  - E.g. What if mostly young people choose *T*?



Asking Causal Questions

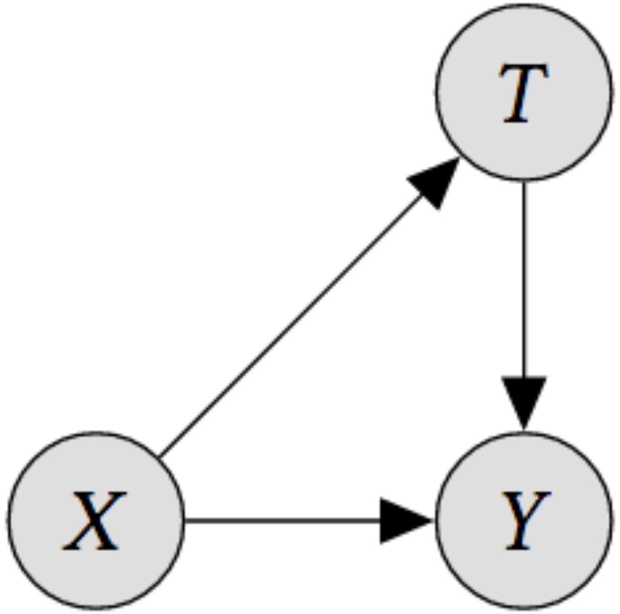
- Suppose *T* is binary (1: received treatment, 0: did not)
- Suppose *Y* is binary (1: disease cured, 0: disease not cured)
- We want to know "If we give someone the treatment *(T = 1),* what is the probability they are cured *(Y = 1)?"*

- This is **not** equal to $P(Y = 1 \mid T = 1)$
- Suppose mostly young people take the treatment, and most were cured, i.e. $P(Y = 1 \mid T = 1)$ is high
  - Is this because the treatment is good? Or because they are young?

Correlation vs. Causation

- Correlation

$$P(Y=1|T=1)=\sum_x P(Y=1,X=x|T=1)$$
$$=\sum_x P(Y=1|T=1,X=x)P(X=x|T=1)$$



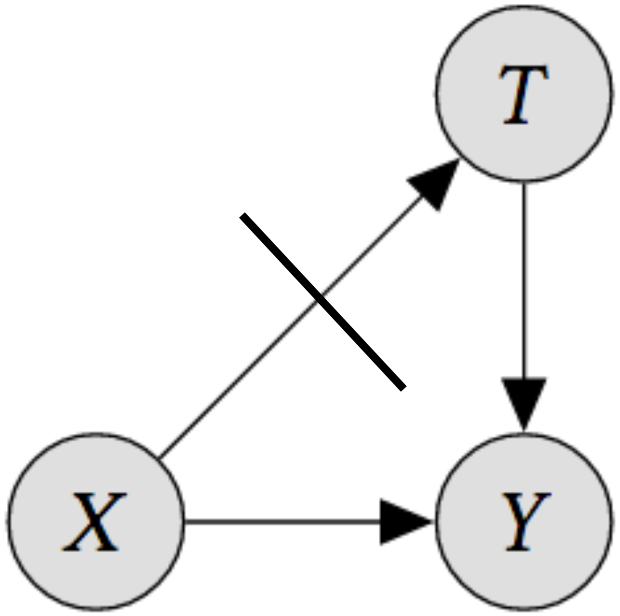
- **In the observed data**, how often do people who take the treatment become cured?
- <u>**The observed data may be biased!!**</u>

Correlation vs. **Causation**

- Let's **simulate** a randomized experiment
  - i.e. $T \perp X$
  - Cut the arrow from X to T
  - This is called a ***do***-operation

- Then, we can estimate causation:

$$P(Y = 1|do(T = 1)) = \sum_x P(Y = 1, X = x|do(T = 1))$$
$$= \sum_x P(Y = 1|T = 1, X = x)P(X = x)$$



Correlation vs. Causation

- Correlation

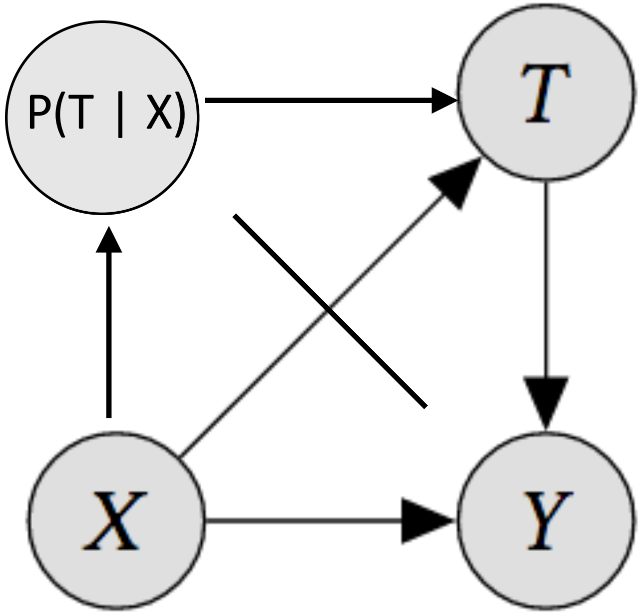
$$P(Y=1|T=1) = \sum_x P(Y=1, X=x|T=1)$$
$$= \sum_x P(Y=1|T=1, X=x)\boxed{P(X=x|T=1)}$$

- Causation – treatment is **independent of X**

$$P(Y=1|do(T=1)) = \sum_x P(Y=1, X=x|do(T=1))$$
$$= \sum_x P(Y=1|T=1, X=x)\boxed{P(X=x)}$$

Inverse Propensity Weighting

- Can calculate this using *inverse propensity scores*
- Rather than adjusting for X, sufficient to adjust for P(T | X)



Inverse Propensity Weighting

- Can calculate this using *inverse propensity scores*
- These are called *stabilized weights*

$$P(Y=1|do(T=1)) = \sum_x P(Y=1, X=x|do(T=1))$$

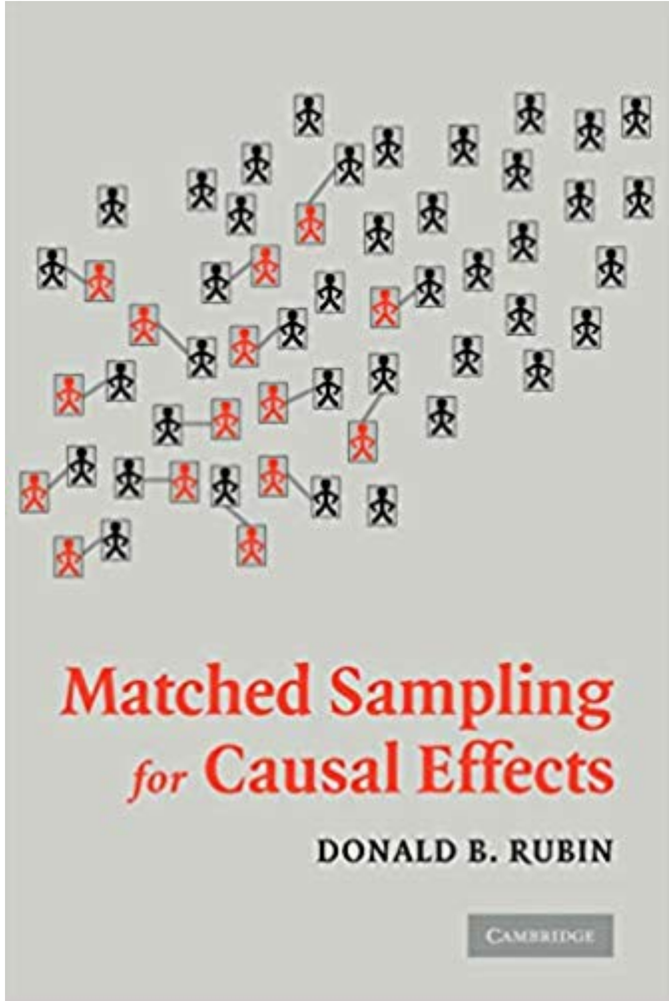
$$= \sum_x P(Y=1|T=1, X=x)P(X=x)$$

$$= \sum_x P(Y=1|T=1, X=x)P(X=x|T=1)\frac{P(T=1)}{P(T=1|X=x)}$$

$$= \sum_x P(Y=1, X=x|T=1)\boxed{\frac{P(T=1)}{P(T=1|X=x)}}$$

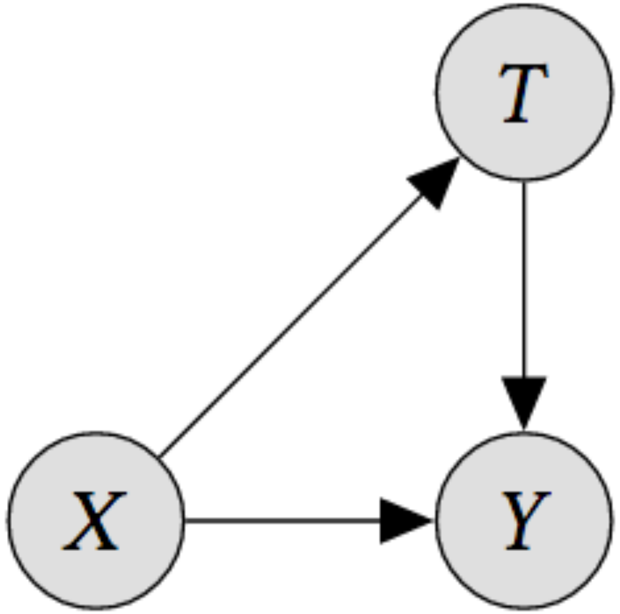
Matching Estimators

- Match up samples with different treatments that are near to each other
- Similar to reweighting



Review: What to **do** with a causal DAG

$$P(Y=1|do(T=1)) = \sum_x P(Y=1, X=x|do(T=1))$$
$$= \sum_x P(Y=1|T=1, X=x)P(X=x)$$



The causal effect of T on Y is

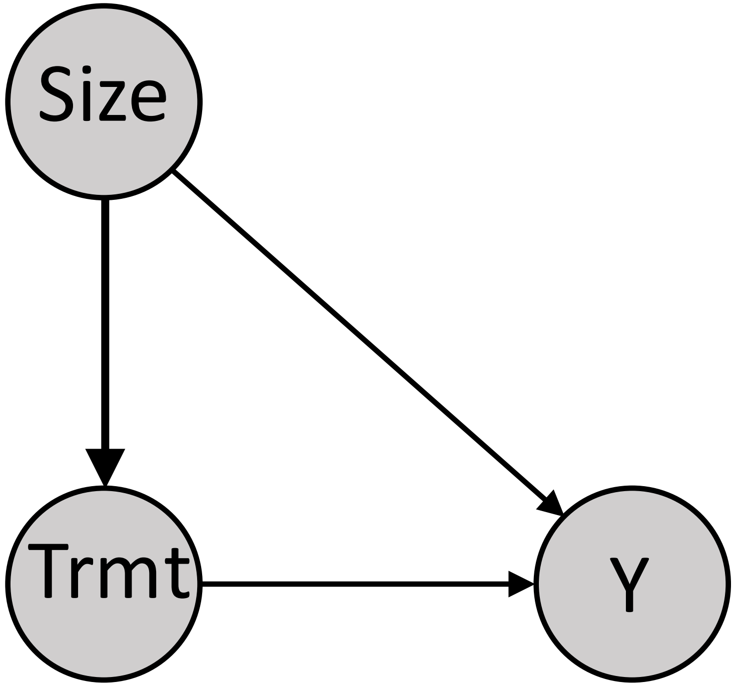
$$CE_{T \to Y} = E[Y|do(T=1)] - E[Y|do(T=0)]$$

This is great! But we've made some assumptions.

Simpson's Paradox, Explained

Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Simpson's Paradox, Explained

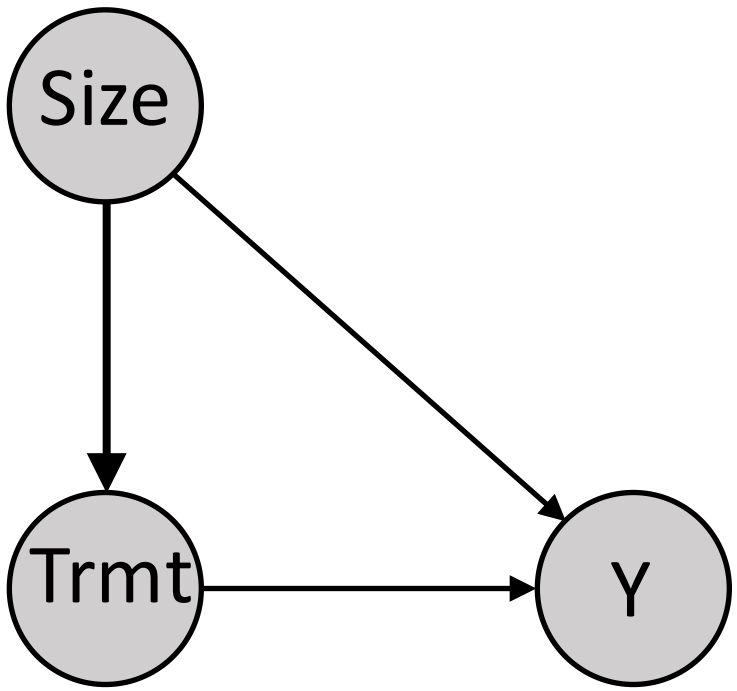


Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

$$P(Y=1|T=A) = \sum_s P(Y=1, Size=s|T=A)$$
$$= \sum_s P(Y=1|T=A, Size=s)P(Size=s|T=A)$$
$$= 0.93 * 0.25 + 0.73 * 0.75 = 0.78$$

$$P(Y=1|T=B) = \sum_s P(Y=1, Size=s|T=B)$$
$$= \sum_s P(Y=1|T=B, Size=s)P(Size=s|T=B)$$
$$= 0.87 * 0.77 + 0.69 * 0.23 = 0.83$$

Simpson's Paradox, Explained



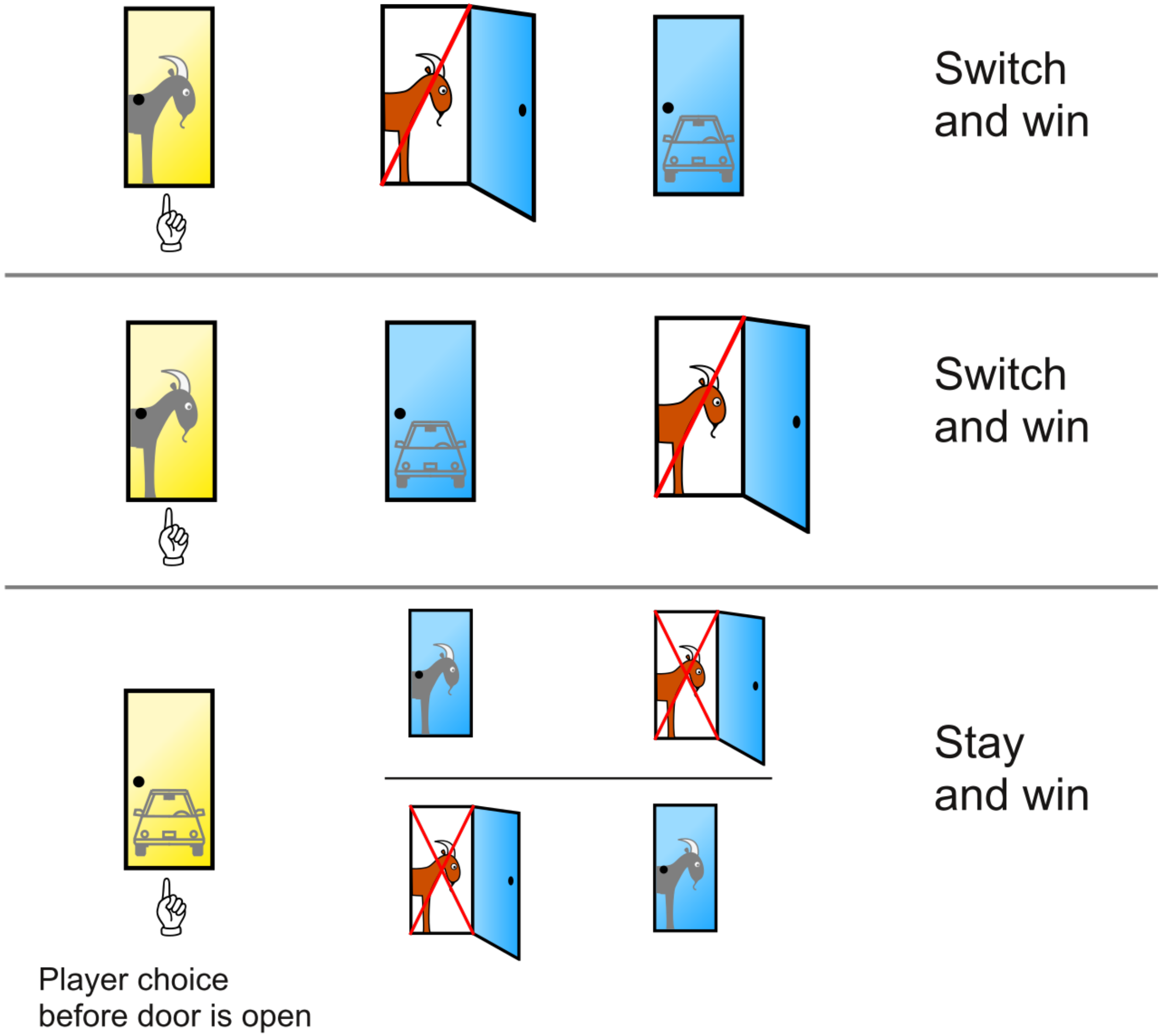
Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

$$P(Y=1|do(T=A)) = \sum_s P(Y=1, Size=s|do(T=A))$$
$$= \sum_s P(Y=1|T=A, Size=s)P(Size=s)$$
$$= 0.93 * 0.51 + 0.73 * 0.49 = 0.83$$

$$P(Y=1|do(T=B)) = \sum_s P(Y=1, Size=s|do(T=B))$$
$$= \sum_s P(Y=1|T=B, Size=s)P(Size=s)$$
$$= 0.87 * 0.51 + 0.69 * 0.49 = 0.78$$

Monty Hall Problem, Explained

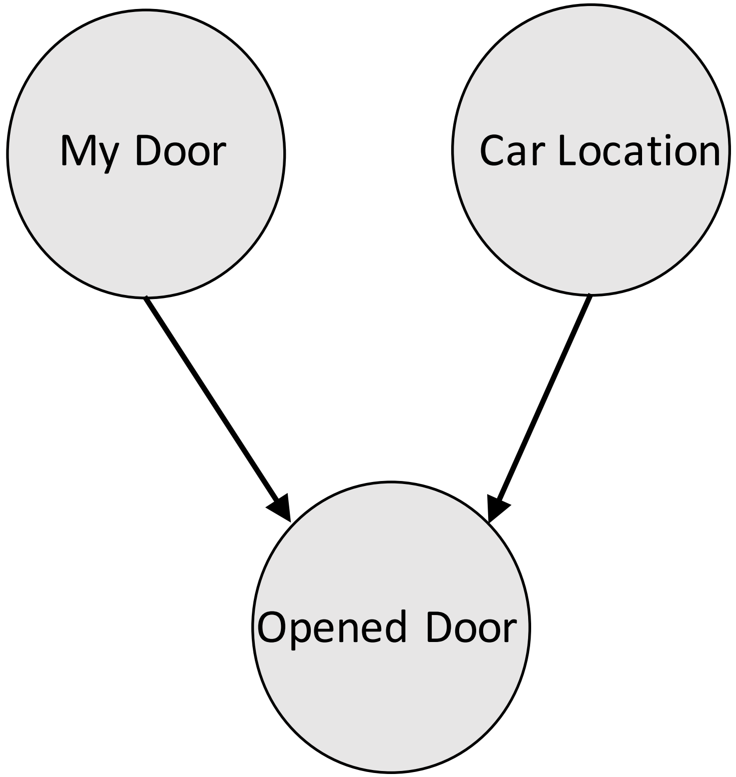
Boring explanation:



Monty Hall Problem, Explained

Causal explanation:

- My door location is correlated with the car location, **conditioned** on which door Monty opens!

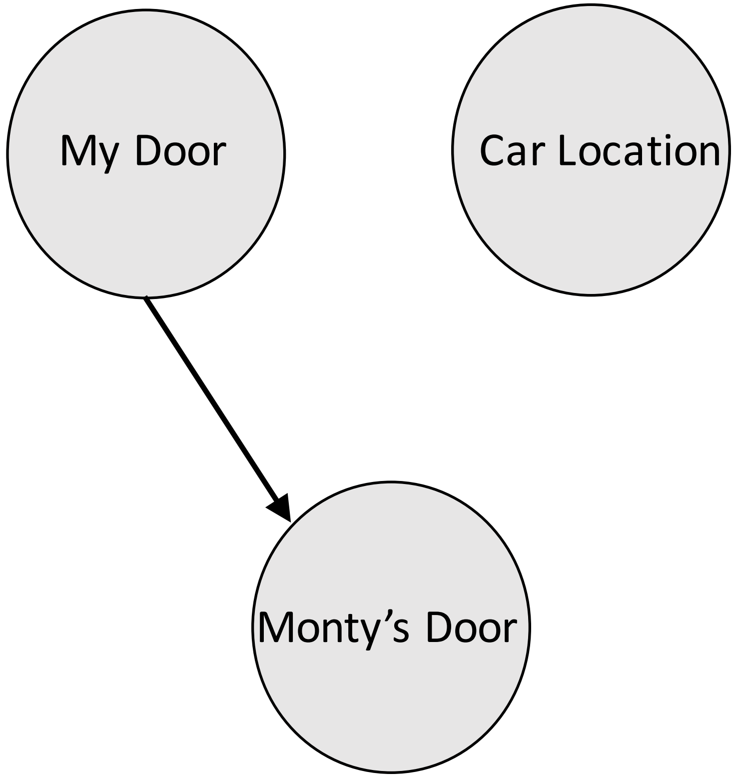


https://twitter.com/EpiEllie/status/1020772459128197121

Monty Hall Problem, Explained

Causal explanation:

- My door location is correlated with the car location, **conditioned** on which door Monty opens!
- This is because Monty won't show me the car
- If he's guessing also, then correlation disappears



Structural Assumptions

- All of this assumes that our assumptions about the DAG that generated our data are correct

- Specifically, we assume that there are *no hidden confounders*
  - Confounder: a variable which causally effects both the treatment (T) and the outcome (Y)
  - **No hidden confounders** means that we have observed all confounders

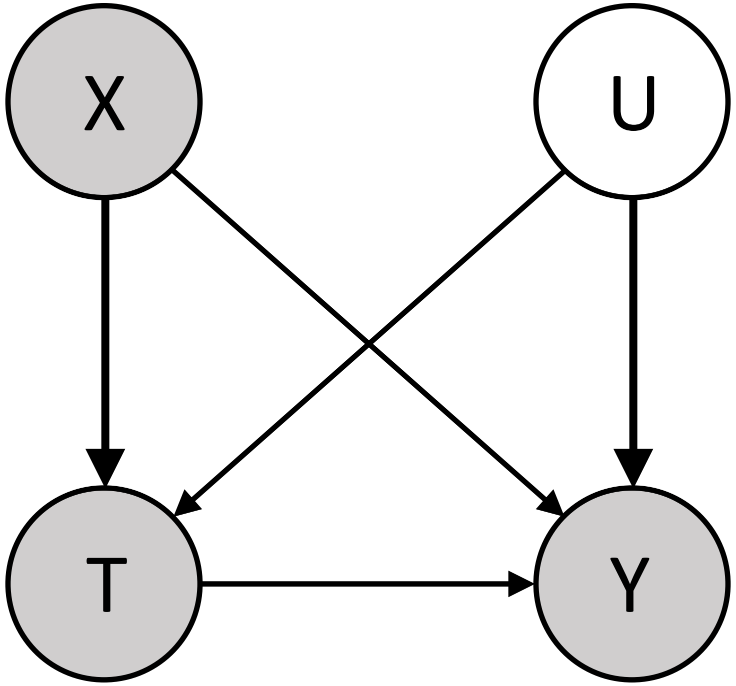
- This is a strong assumption!

Hidden Confounders

- Cannot calculate P(Y | do(T)) here, since U is unobserved

$$P(Y=1|do(T=1)) = \sum_{x,u} P(Y=1, X=x, U=u|do(T=1))$$
$$= \sum_{x,u} P(Y=1|T=1, X=x, U=u)P(X=x, U=u)$$

- We say in this case that the causal effect is **unidentifiable**
  - Even in the case of infinite data and computation, we can never calculate this quantity

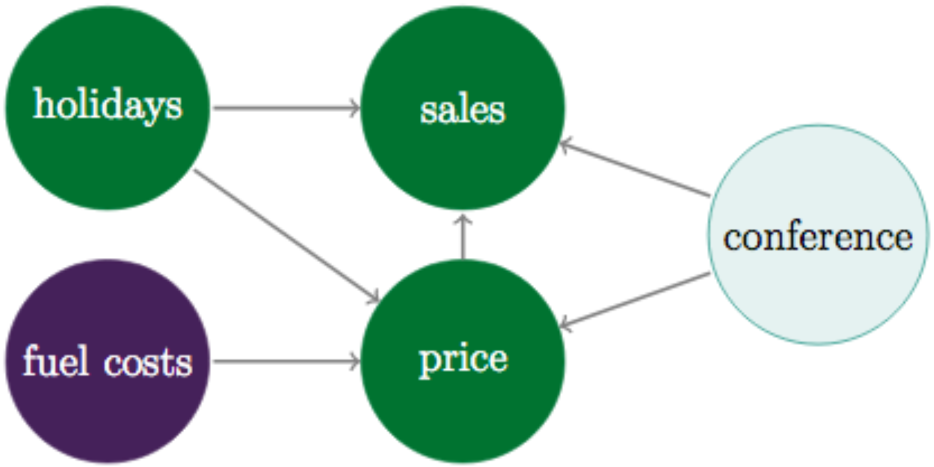


What Can We Do with Hidden Confounders?

- Instrumental variables
  - Find some variable which effects **only** the treatment
- Sensitivity analysis
  - Essentially, assume some maximum amount of confounding
  - Yields confidence interval
- Proxies
  - Other observed features give us information about the hidden confounder

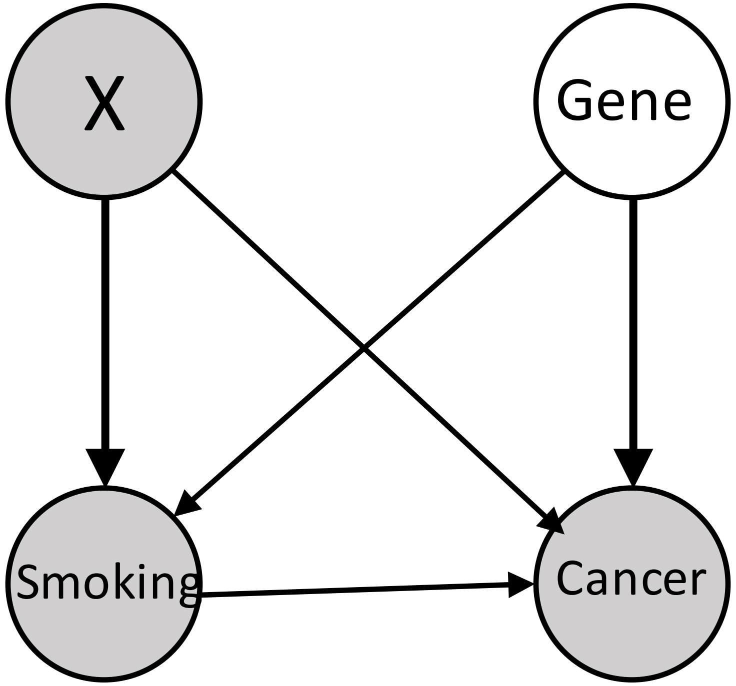
Instrumental Variables

- Find an *instrument* – variable which only affects treatment
  - Decouples treatment and outcome variation
- With linear functions, solve analytically
- But can also use any function approximators



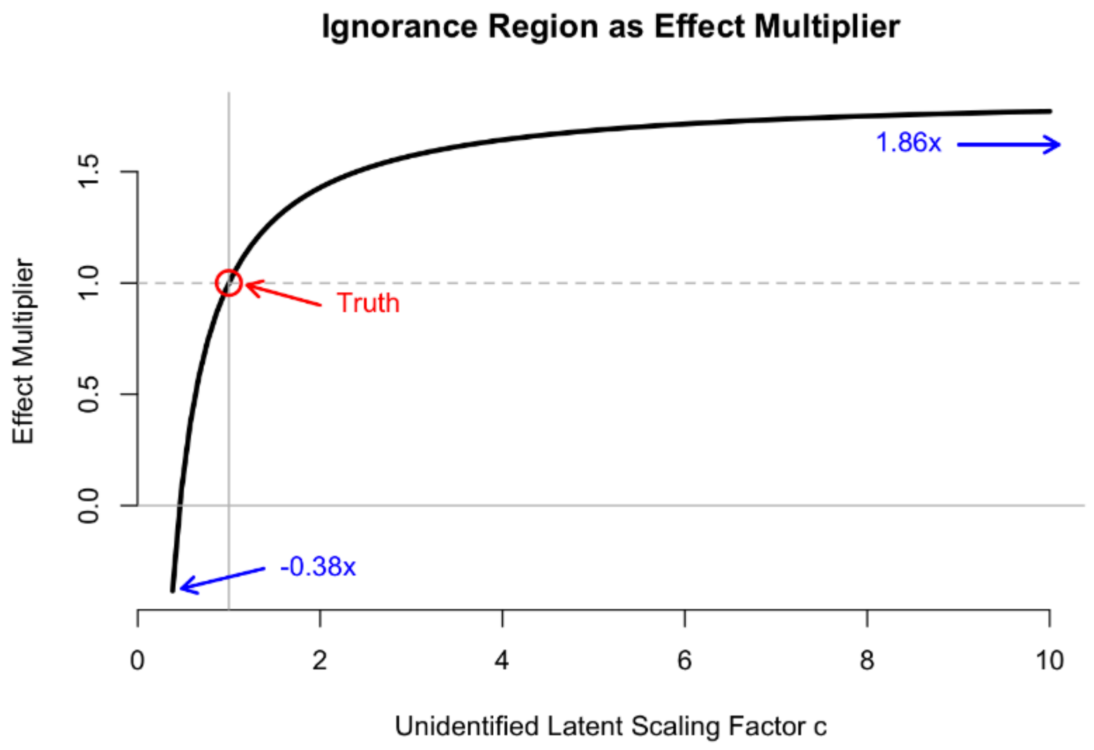
Sensitivity Analysis

- Determine the relationship between **strength of confounding** and **causal effect**
- Example: Does smoking cause lung cancer? (we now know, yes)
  - There *may* be a gene that causes lung cancer and smoking
  - We can't know for sure!
  - However, we can figure out **how strong this gene would need to be** to result in the observed effect
  - Turns out – very strong



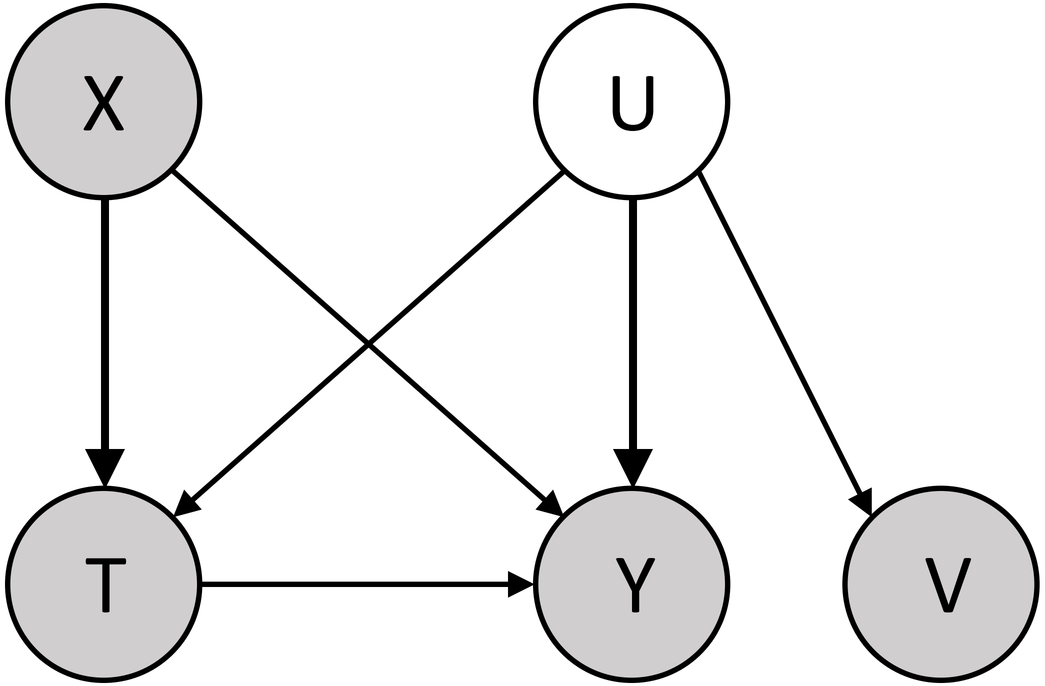
Sensitivity Analysis

- The idea is: parametrize your uncertainty, and then decide which values of that parameter are reasonable



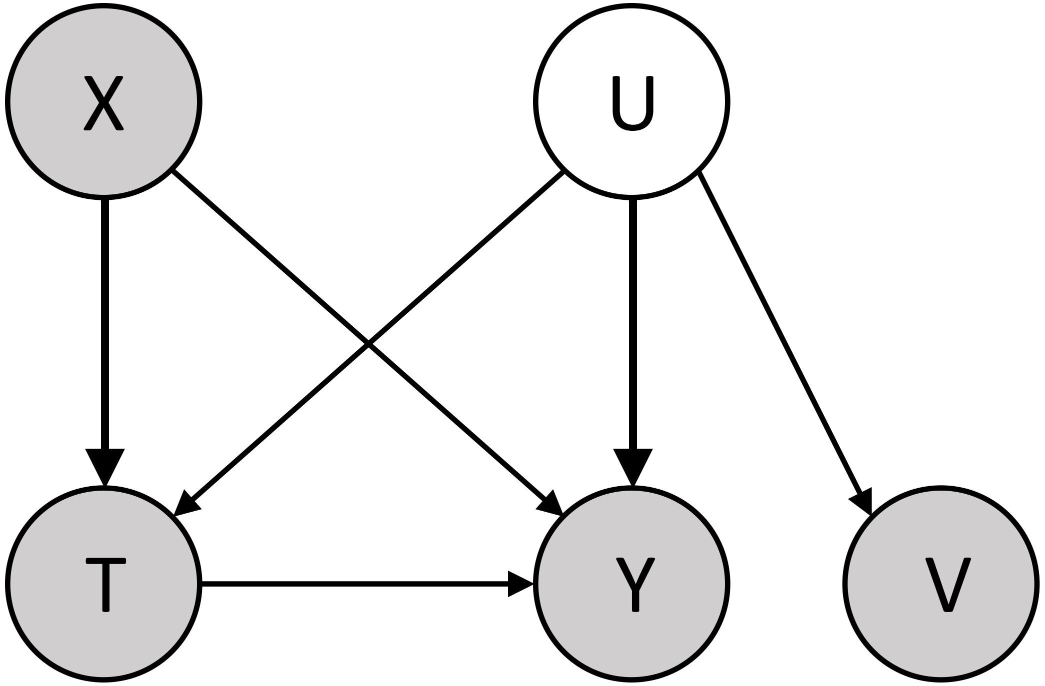
Using Proxies

- Instead of measuring the hidden confounder, measure some **proxies** ($V = f_{prox}(U)$)
  - <u>Proxies</u>: variables that are caused by the confounder
  - If U is a child's age, V might be height
- If f_{prox} is known or linear, we can estimate this effect



Using Proxies

- If f_{prox} is non-linear, we might try the Causal Effect VAE
- Learn a posterior distribution P(U | V) with variational methods
- However, this method does not provide theoretical guarantees
- Results may be unverifiable: proceed with caution!



Causality and Other Areas of ML

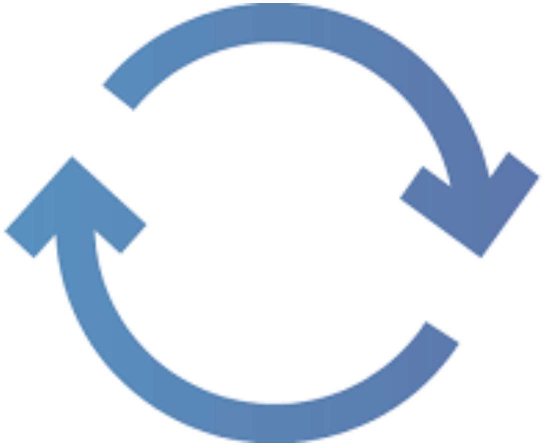
- Reinforcement Learning
  - Natural combination – RL is all about taking actions in the world
  - Off-policy learning already has elements of causal inference
- Robust classification
  - Causality can be natural language for specifying distributional robustness
- Fairness
  - If dataset is biased, ML outputs might be unfair
  - Causality helps us think about dataset bias, and mitigate unfair effects

Quick Note on Fairness and Causality

- Many fairness problems (e.g. loans, medical diagnosis) are actually causal inference problems!
- We talk about the label Y – however, this is not always observable
  - For instance, we can’t know if someone *would* return a loan if we don’t give one to them!
  - This means if we just train a classifier on historical data, our estimate will be biased
  - Biased in the fairness sense <u>and</u> the technical sense
- General takeaway: if your data is generated by past decisions, think very hard about the output of your ML model!

Feedback Loops

- Takes us to part 2… feedback loops
- When ML systems are deployed, they make many decisions over time
- So our past predictions can impact our future predictions!
  - Not good

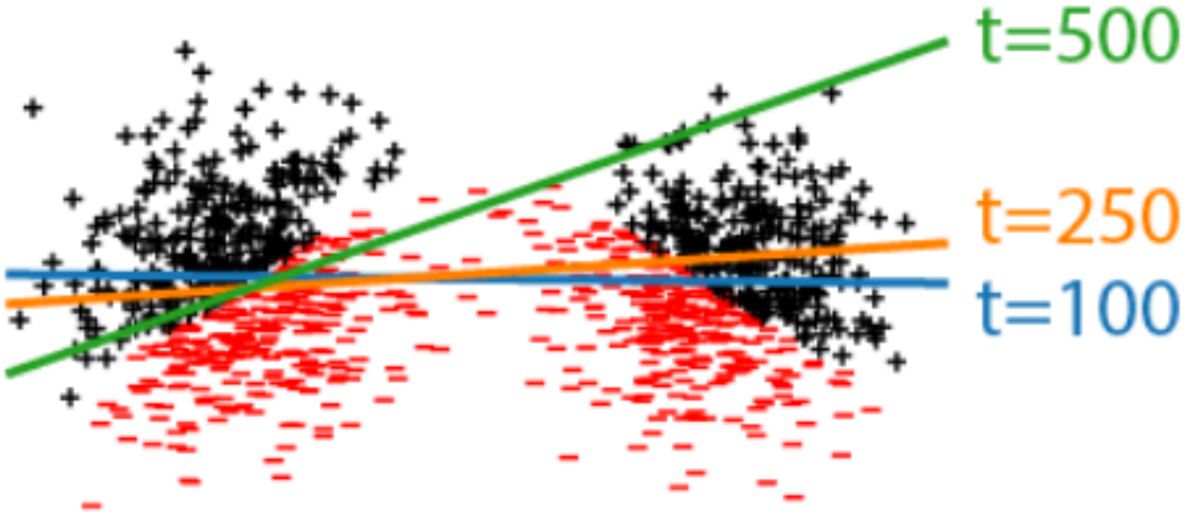


Unfair Feedback Loops

- We'll look at "Fairness Without Demographics in Repeated Loss Minimization" (Hashimoto et al, ICML 2018)
- Domain: recommender systems
- Suppose we have a majority group (A = 1) and minority group (A = 0)
- Our recommender system may have high overall accuracy but low accuracy on the minority group
  - This can happen due to empirical risk minimization (ERM)
- Can also be due to repeated decision-making

Repeated Loss Minimization

- When we give bad recommendations, people leave our system
- Over time, the low-accuracy group will shrink



Distributionally Robust Optimization

- Upweight examples with high loss in order to improve the worst case
- In the long run, this will prevent clusters from being underserved

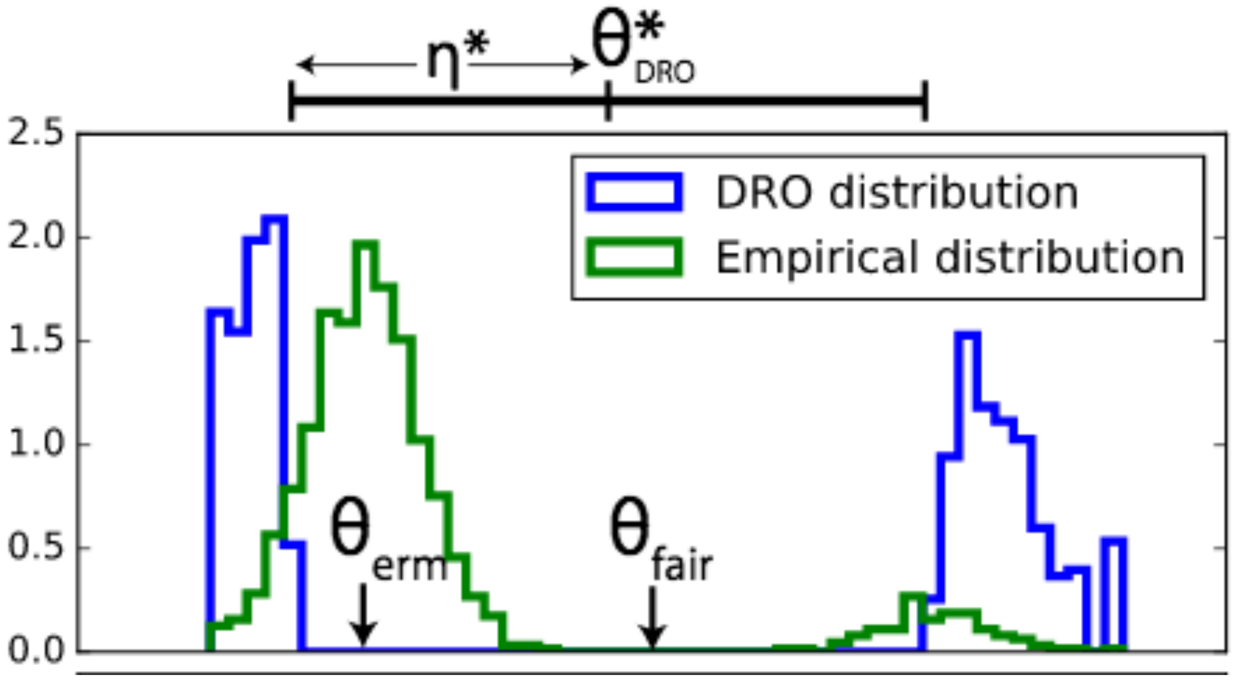
$$\mathcal{R}_{\mathrm{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)].$$

- This ends up being equal to

$$\inf_{\eta \in \mathbb{R}} \left\{ F(\theta; \eta) := C \left(\mathbb{E}_P \left[[\ell(\theta, Z) - \eta]_+^2 \right] \right)^{\frac{1}{2}} + \eta \right\}$$

Distributionally Robust Optimization

- Upweight examples with high loss in order to improve the worst case
- In the long run, this will prevent clusters from being underserved



Conclusion

- Your data is not what it seems
- ML models only work if your training/test set **actually** look like the environment you deploy them in
- This can make your results unfair
  - Or just incorrect
- So examine your model assumptions and data collection carefully!