Week 13

Validity

INFO 3402: Information Exposition

Brian Keegan, Ph.D.
Assistant Professor, Information Science
brian.keegan@colorado.edu

University of Colorado Boulder

Course Overview

Module	Week	Dates	Type	Skill
Shaping	1	Jan 11, Jan 13	Computation	Loading
	2	Jan 18, Jan 20	Computation	Aggregating
	3	Jan 25, Jan 27	Computation	Joining
	4	Feb 1, Feb 3	Computation	Tidying
Distribution	5	Feb 8, Feb 10	Computation	Histograms
	6	Feb 15, Feb 17	Communication	Audience
Comparison	7	Feb 22, Feb 24	Computation	Cat plots
	8	Mar 1, Mar 3	Communication	Persuasion
Trend	9	Mar 8, Mar 10	Computation	Time series
	10	Mar 15, Mar 17	Communication	Uncertainty
	11	Mar 22, Mar 24	Spring Break	
Relationship	12	Mar 29, Mar 31	Computation	Scatter plots
	13	Apr 5, Apr 7	Communication	~~Fallacies~~ Validity
Spatial	14	Apr 12, Apr 14	Computation	Choropleths
	15	Apr 19, Apr 21	Communication	Conventions
Projects	16	Apr 26, Apr 28	Projects	

Readings

- Questions for Friday's Weekly Quiz 13 will be drawn from these readings
  - Conjointly – Internal Validity.
    - Read all four sub-sections (Establishing Cause & Effect and various threats), "Regression to the Mean" can be skipped
  - Matthay & Glymour. (2020). A Graphical Catalog of Threats to Validity.
    - Focus on introduction and section in internal validity.
- Play GuessTheCorrelation and skim through Spurious Correlations

Weekly Assignment 13

- **Skills**: Reviewing correlations, counterfactual thinking, identifying threats to internal validity
- **Data**: County data

Module Assignment 05

Module Assignment 05

- Use the U.S. county data to identify an *unusual* relationship between variables
  - ~1,000 different county-level variables in “us_counties.csv”, “analytic_data2021.csv”, “Unemployment.csv”
  - Must be unusual: A relationship between poverty and income is **not** unusual—it’s to be expected!
  - Some relationships might be trivial: employment from one dataset is likely similar with employment from another
  - The relationship should be strong-ish: if it’s linear, a correlation above 0.2 or below -0.2
  - Identifying this relationship can be top-down (sorting, correlograms, pairplots, *etc*.) or bottom-up (exploring pairs)
  - Like WA12 Questions 2 & 3, explore whether this relationship shows up or disappears in related variables
  - Make a case for there being a causal rather than a random relationship between these variables

- 700 – 1000 words with at least one visualization

- Module Assignment 05 will be due on **Wednesday, April 13 by 11:59pm**
  - Submit URL of your Medium post to Canvas or save and submit as an HTML file
  - Tag your post on Medium with “INFO3402S22A5” and whatever other tags you’d like

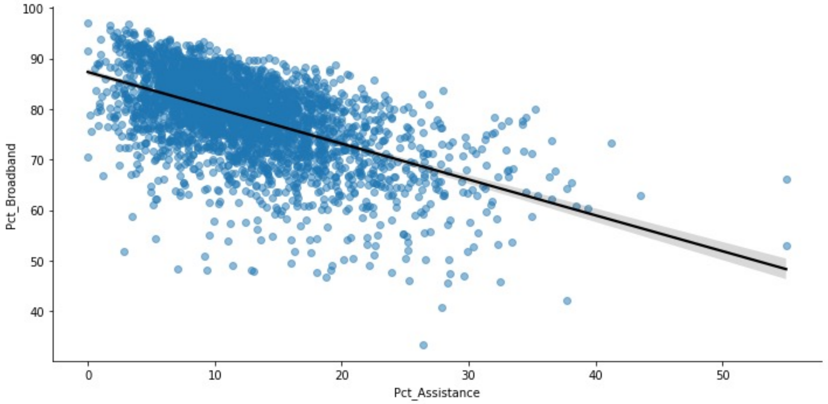
Module Assignment 3 shout-outs

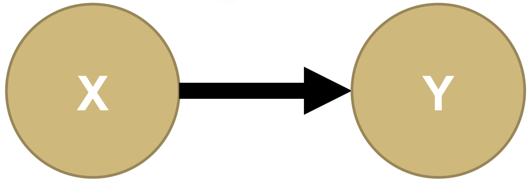
- Jessie Bart – Worldwide Income Inequality
  - Exploring changing in Gini coefficient of income inequality over time and by country
- Max Blanco – World War II's Effects on Population Growth
  - Comparing demographics of US, Germany, and Russia
- Chris Davis – Free and Independent Media Trends
  - Contextualizing Russian media propaganda with Democracy Index and Independent media scores
- Bailey Gimpel – A Humanitarian Crisis: Civil Liberties on the Decline
  - Identifying interesting tension between declining civil liberties scores but increasing political participation
- Jasmine Rivera – Birthrates declining at an Alarming Rate in Japan
  - Nice demographic pyramids, motivation from a 2017 article, comparing different kinds of evidence
- John Ross Greene – The Growing Inequality of Wealth and Power in the United States
  - Visualization stacked with lots of accessibly information, strong voice throughout writeup
- Brian Lee – South Korea retiring into Economic Collapse
  - Nice demographic pyramids and explanations, strong motivation and voice, triangulating evidence
- Max Vali – Income, Life Expectancy, and Access to Sanitation
  - Exploring trends and interactions across three variables over time

Validity

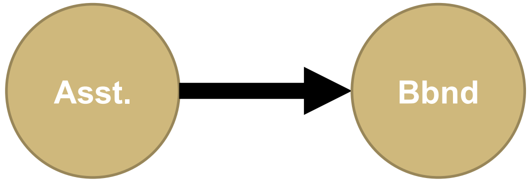
Making a causal case

- MA05 ask you to "make a case for there being a causal rather than a random relationship between these variables"
- Assistance and broadband have a -0.54 correlation
  - Correlation coefficient is a measure of relationship strength
  - Always in range [-1,1]
- A linear regression model gives a slope parameter -0.906
  - Not the same as the correlation coefficient!
  - Can be any real number and has interpretable units
  - A one percent increase in public assistance is correlated with a 0.906 percent drop in broadband adoption
- It's a strong relationship, but is it a *causal* relationship?



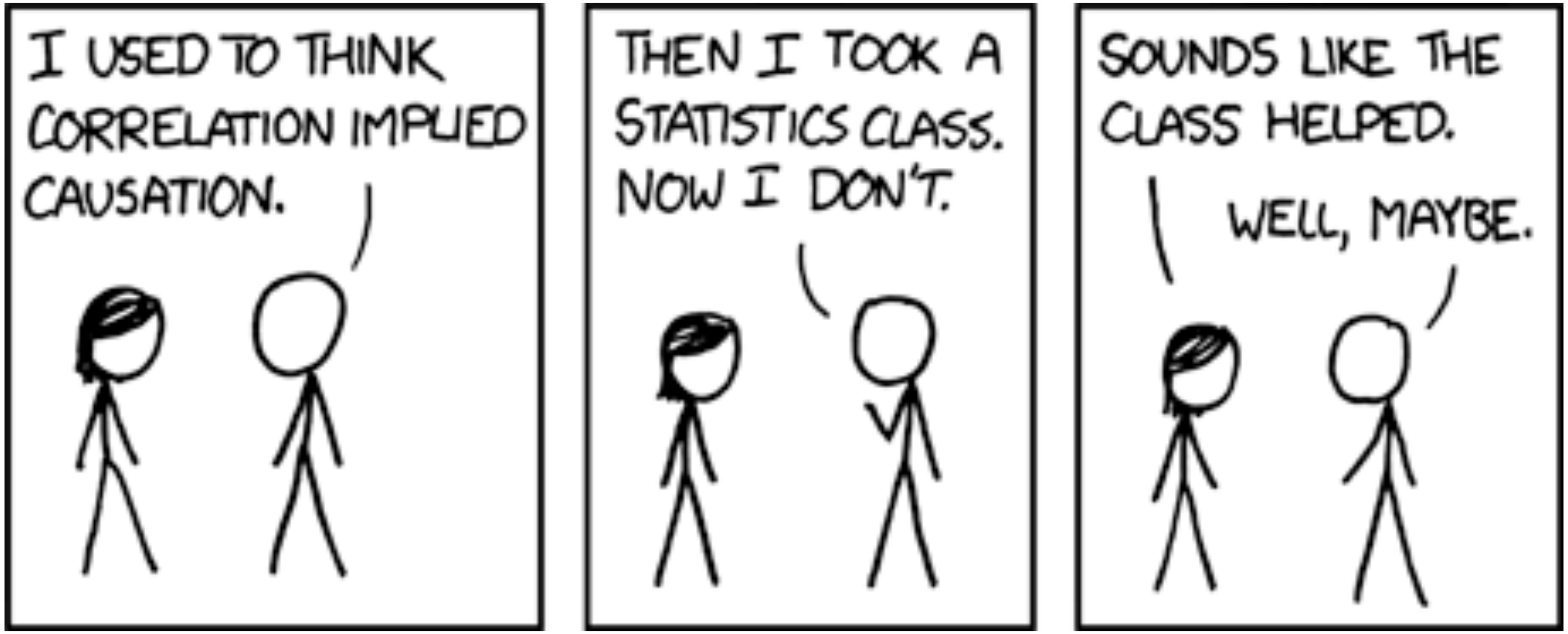


Simple Causation
X causes Y



More Assistance causes Less Broadband

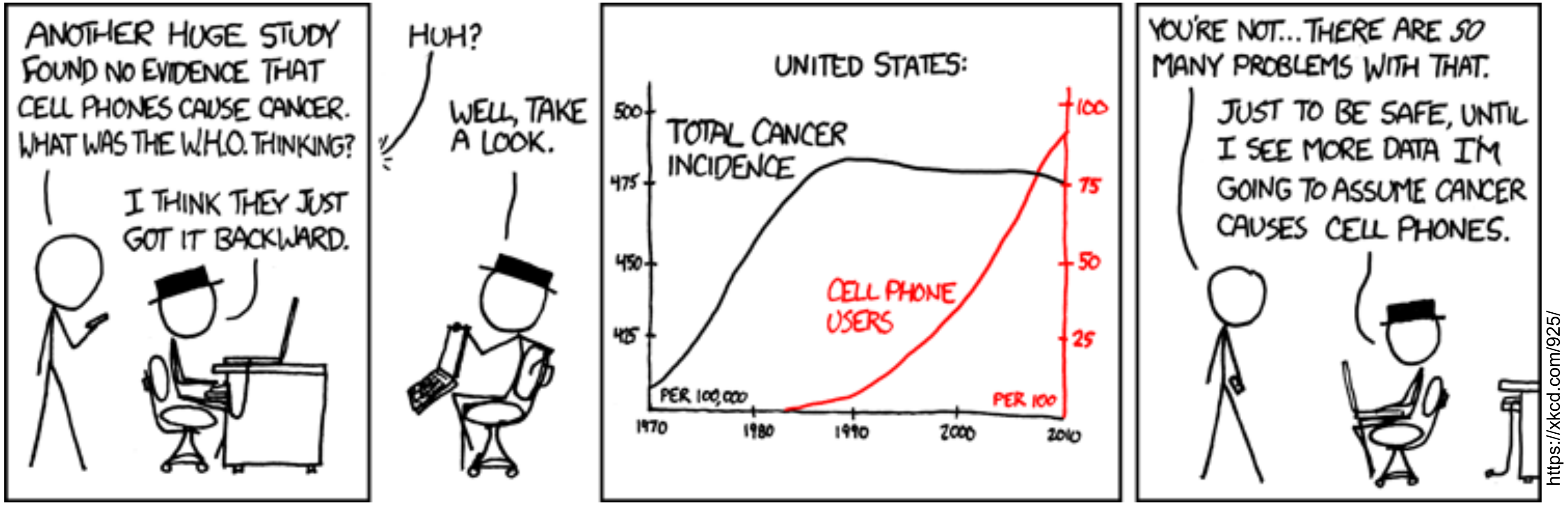
Correlation vs. causation



https://xkcd.com/552/

https://xkcd.com/552/

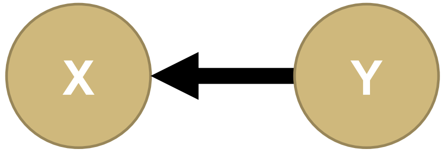
Causal direction



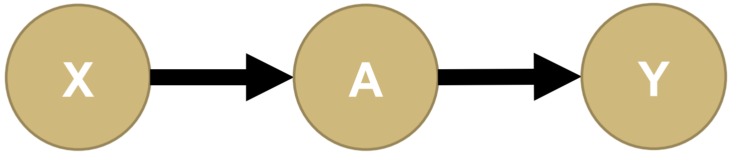
https://xkcd.com/925/

Counterfactual thinking

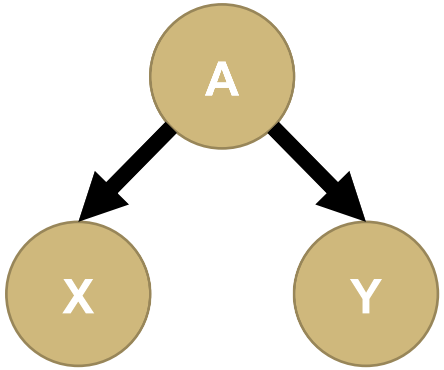
- We must rule out other explanations for causal relationships
- Drawing contrasts to an observation allows us to explore the possibilities of alternative outcomes
- From a simple “X causes Y” relationship, there are many common patterns of other causal relationships that could explain why we observe a correlation between X and Y



Reverse Causation
More broadband causes less assistance



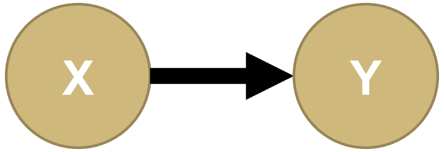
Mediation
Assistance causes no computer causes no broadband



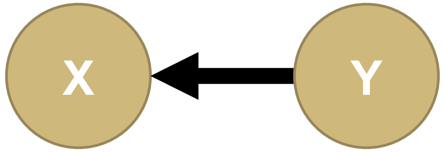
Confounding
Poverty causes assistance and no broadband

Alternative causal relationships

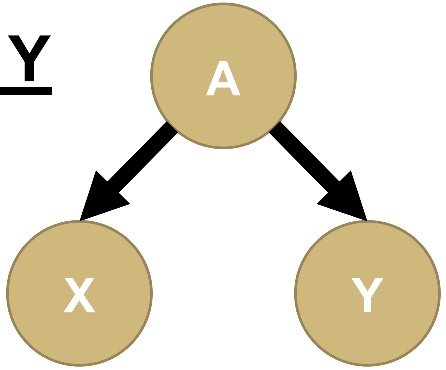
<u>You observe a correlation between X and Y</u>



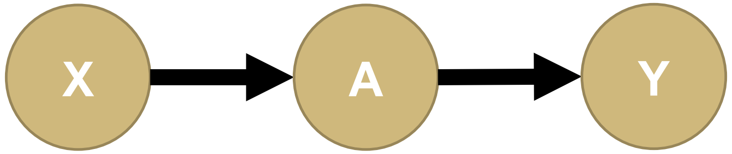
Simple Causation
X causes Y



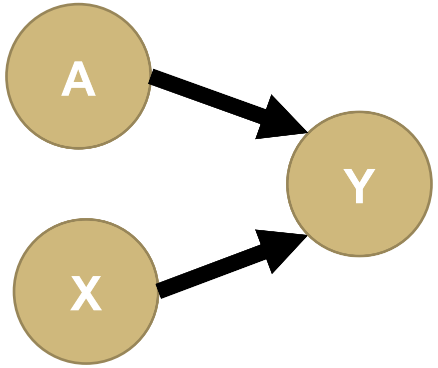
Reverse Causation
Y causes X



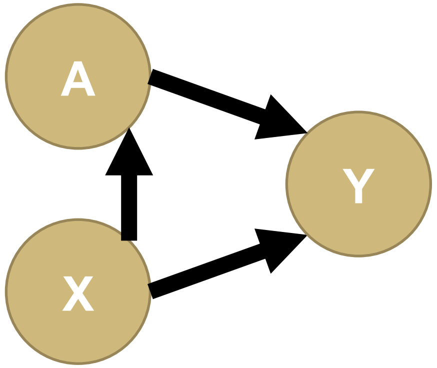
Confounding
A causes X and Y



Mediation
X causes A which causes Y

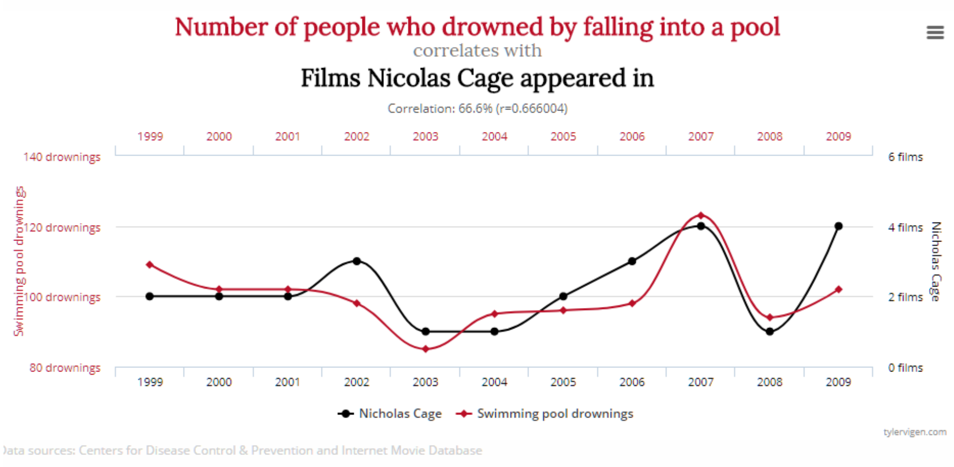


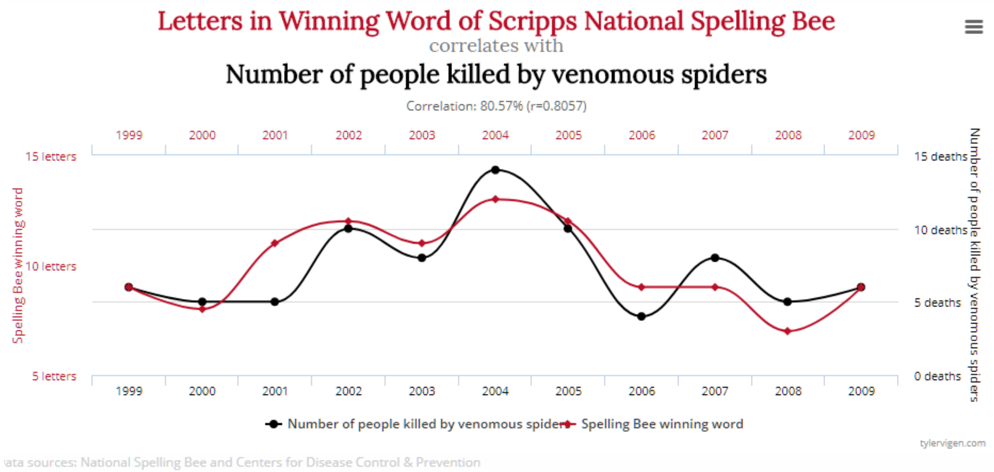
Multiple causation
X and A cause Y



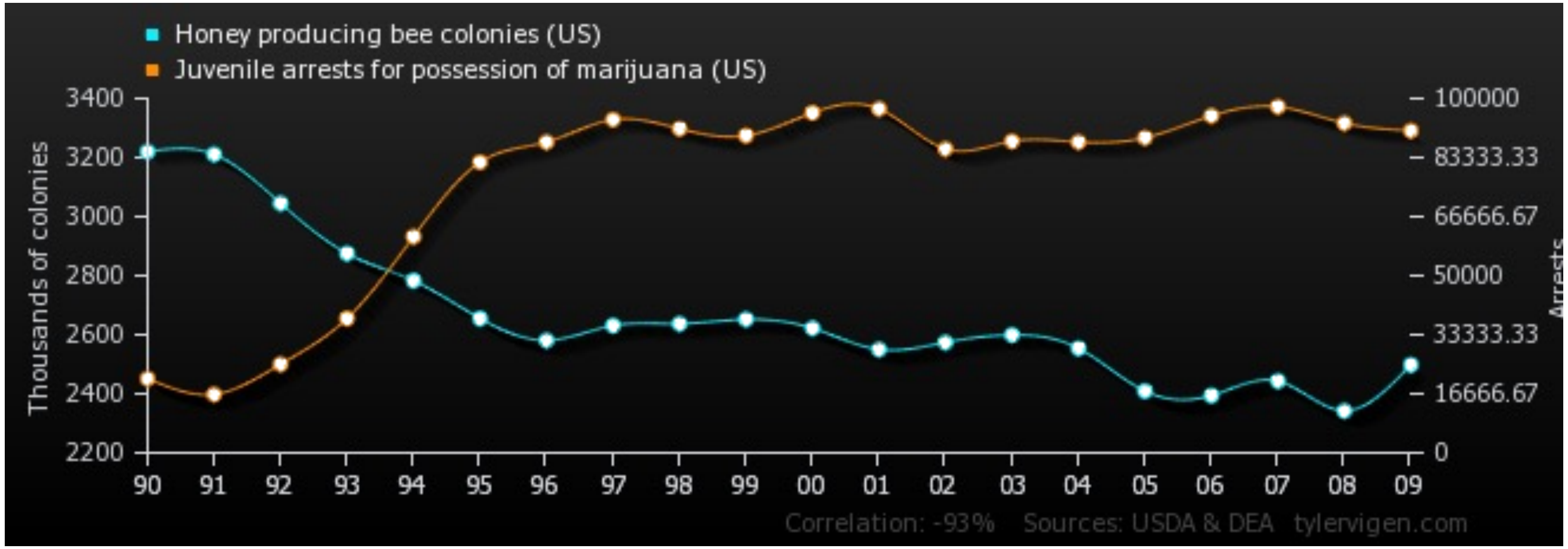
Interaction
X causes A and Y,
A also causes Y

Spurious correlations





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



http://tylervigen.com/spurious-correlations

Validity

- ***Conclusion validity***: is there a relationship between the variables of interest?
- ***Internal validity***: is the observed relationship a causal relationship?
- ***Construct validity***: did you measure what you thought you measured?
- ***External validity***: does the relationship observed generalize to other people, places, and times?

Internal validity

- There are many factors outside of a researcher's control that can change behavior
  - How to rule out other factors when trying to make a causal argument?
- Causal requirements
  - *Temporal precedence*: cause must happen before effect
  - *Covariation*: if more treatment, then more outcome; if less treatment, then less outcome
  - *Alternative explanations:* threats form single/multiple groups, other factors
- Experimental design and causal inference methods → advanced statistics and $$$ in industry

https://conjointly.com/kb/internal-validity/

Hill (1965): Association or Causation?

1. **Strength**: How big is effect you are measuring? (Large effects imply causality, but small effects can)
2. **Consistency**: Can the effect be replicated? (Causal effects should be reproducible)
3. **Specificity**: Can association be pinpointed? (No other mechanisms should plausibly explain)
4. **Temporality**: Do the causes come before effects? (No time traveling)
5. **Gradient**: Do stronger/weaker treatments cause greater/lesser effects? (Same as covariation)
6. **Plausibility**: Does the causal mechanism itself make sense? (How could Cage films cause drowning?)
7. **Coherence**: Is the causal mechanism compatible with other evidence?
8. **Experiment**: Can experiments reproduce the effect?
9. **Analogy**: Is the causal mechanism similar to other established mechanisms?

Hill, A.B. (1965). The Environment and Disease: Association or Causation? *Proc. Royal Society of Medicine*.

Threats to internal validity

1. **Attrition**. Subjects more likely to leave one study condition than another.
2. **Confounding**. A third variable related to your cause is influencing the observed effects.
3. **Diffusion**. An intervention spreads from a treatment group into a control group.
4. **Researcher bias**. Researchers treat one group differently than another.
5. **Events**. Major events happening during a study could similarly influence groups.
6. **Instrumentation**. Subjects act differently because of methods rather than treatments.
7. **Maturation**. Everyone is always getting older, causes based on time need careful design.
8. **Regression**. Subjects with extreme scores at one point in time tend to have normal scores later.
9. **Repetition**. Repeated testing can cause learning, exhaustion, *etc*. separate from intervention.
10. **Selection**. Strategy for assigning subjects to groups can influence results.

https://www.formpl.us/blog/internal-validity; https://www.verywellmind.com/internal-and-external-validity-4584479

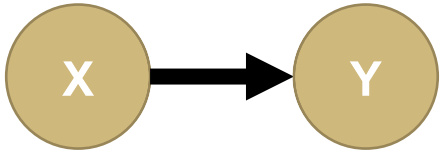
Research question & Hypothesis

- **Research Question**: "Does cannabis legalization reduce opioid overdoses?"

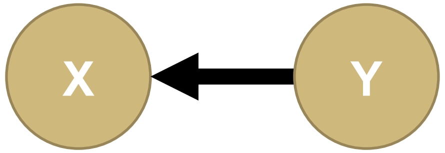
- Federal government regulates cannabis as having a high risk of abuse and no medical uses
  - https://www.dea.gov/sites/default/files/2020-06/Marijuana-Cannabis-2020_0.pdf
- The risks of abuse and harm from cannabis are actually lower than alcohol or tobacco
  - Nutt, King, Saulsbury, Blakemore (2007); Nutt, King, & Phillips (2010); Lachenmeier & Rehm (2015)
- Cannabis use may substitute for more dangerous drugs for treating pain, anxiety, and sleep
  - Lucas & Walsh (2017); Corroon, Mischley, & Sexton (2017); Piper, DeKeuster, et al. (2017)

- **Hypothesis**: If greater availability of legal cannabis reduces use of higher-risk substances, use of higher-risk substances should decrease after cannabis legalization

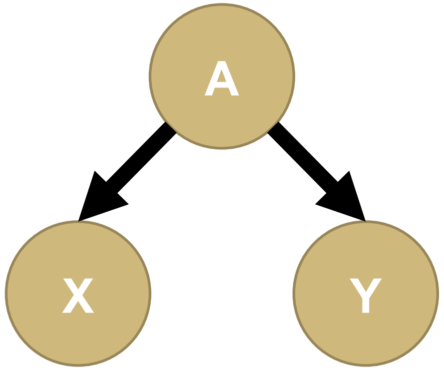
Alternative causal relationships



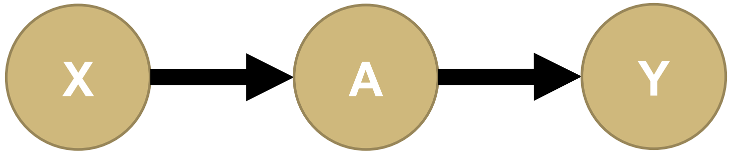
Simple Causation
Legalization (X) causes ↓ opioids (Y)



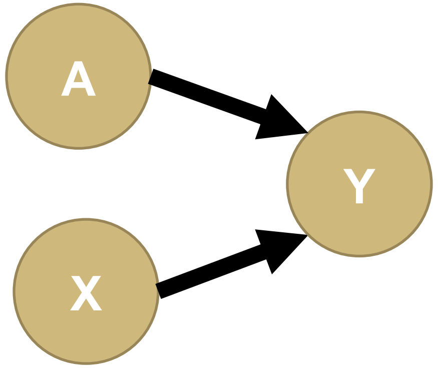
Reverse Causation
↓ opioids causes legalization



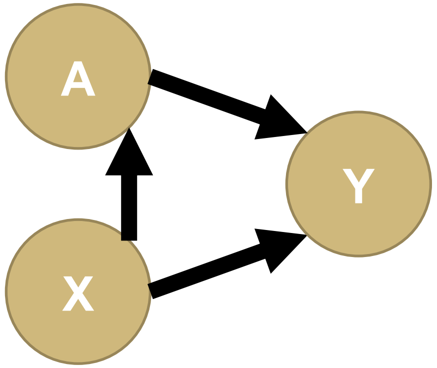
Confounding
Health campaign (A) causes
↓ opioids and legalization



Mediation
Legalization causes enforcement dollars (A)
causes ↓ opioids



Multiple causation
Legalization and health campaign
causes ↓ opioids



Interaction
Legalization causes ↓ opioids and
health campaign causing ↓ opioids

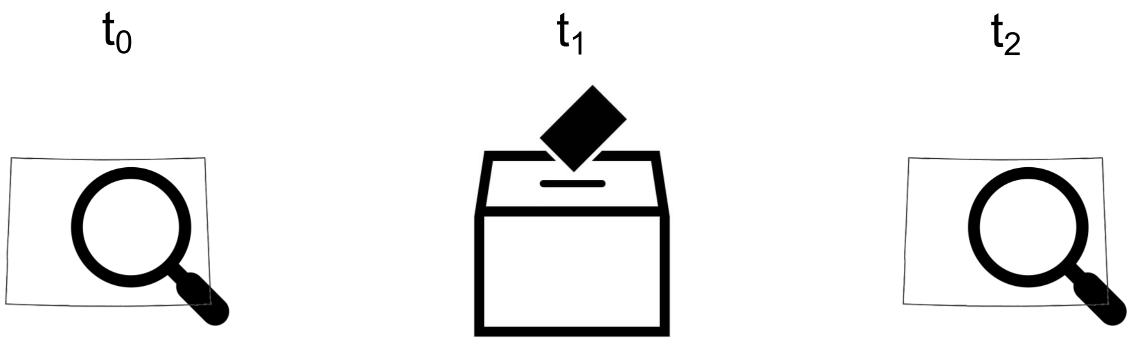
Threats to internal validity

1. **Attrition**. Population pre and post legalization should be similar, admittedly not identical.
2. **Confounding**. Lots of threats here because other causes could influence opioid and cannabis use.
3. **Diffusion**. Depends on design, no time travelers is more convincing than no spillovers.
4. **Researcher bias**. Depends on design.
5. **Events**. Major events or innovations could shift opioid use independently of cannabis legalization.
6. **Instrumentation**. Design of opioid surveys could have changed.
7. **Maturation**. Aging population may have more injuries and need more pain treatment.
8. **Regression**. Colorado's opioid use may have been unusually high or low pre-2014.
9. **Repetition**. Asking about opioid use may make people less truthful, more interested, *etc*.
10. **Selection**. Sampling strategy for opioid surveys could have changed.

https://www.formpl.us/blog/internal-validity; https://www.verywellmind.com/internal-and-external-validity-4584479

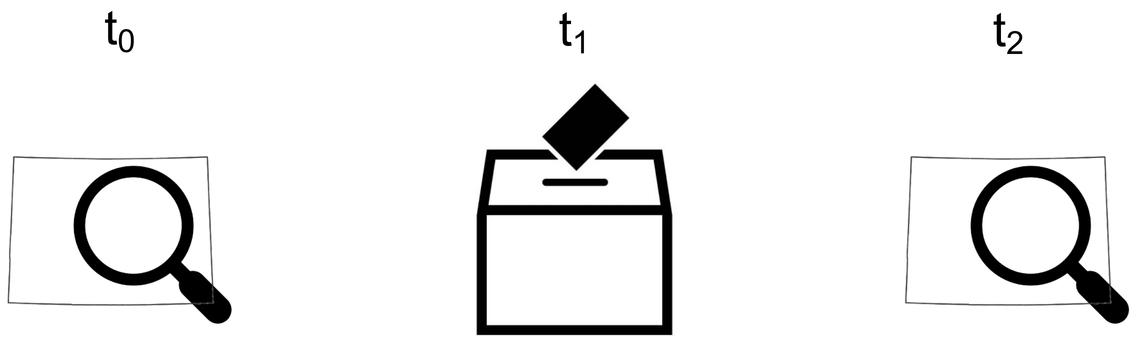
Single group design

- **<u>Single group design</u>**: measure a baseline, give the group a treatment, measure the outcome
  - Measure opioid use before legalization in Colorado (past – 2014)
  - Identify when cannabis legalization started in Colorado (2014)
  - Measure opioid use after legalization in Colorado (2015 – present)
  - Compare values after legalization to values before legalization to evaluate effect of treatment



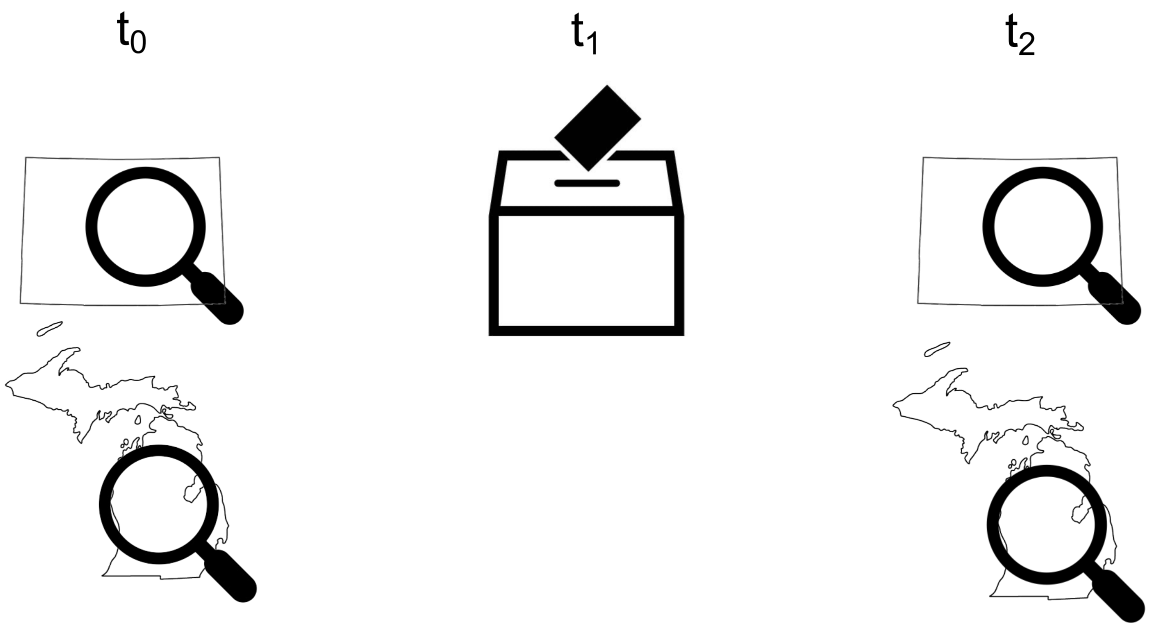
Alternative explanations – single group

- **History**: Legalization didn’t reduce overdoses, an unrelated 2014 mental health program did
- **Maturation**: Risky behavior goes down with age, ODs went down because population got older
- **Testing**: Asking about opioid consumption before 2014 caused people to OD
- **Instrumentation**: Method for measuring opioid ODs changed between 2010 and 2018
- **Attrition**: Lots of Coloradans moved in/out between after 2014, different groups now
- **Regression**: Opioid ODs were unusually high around 2010, they could only come down



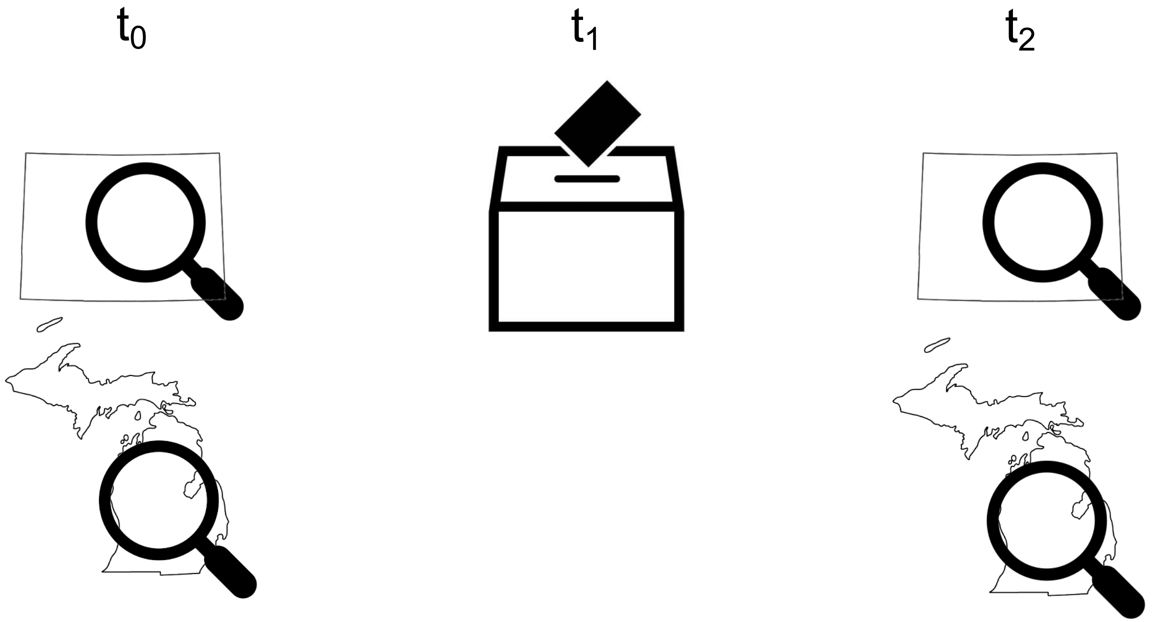
Multiple group design

- **<u>Multiple group designs</u>**: Compare a treatment group to a control group without treatment
  - Identify a state similar to Colorado but did not legalize cannabis: Michigan
  - Measure opioid use before legalization in Colorado and Michigan (past – 2014)
  - Identify when cannabis legalization started in Colorado (2014)
  - Measure opioid use after legalization in Colorado and Michigan (2015 – present)
  - Compare CO & MI values after legalization to CO & MI values from before to evaluate effect of treatment



Threats to internal validity – multiple group

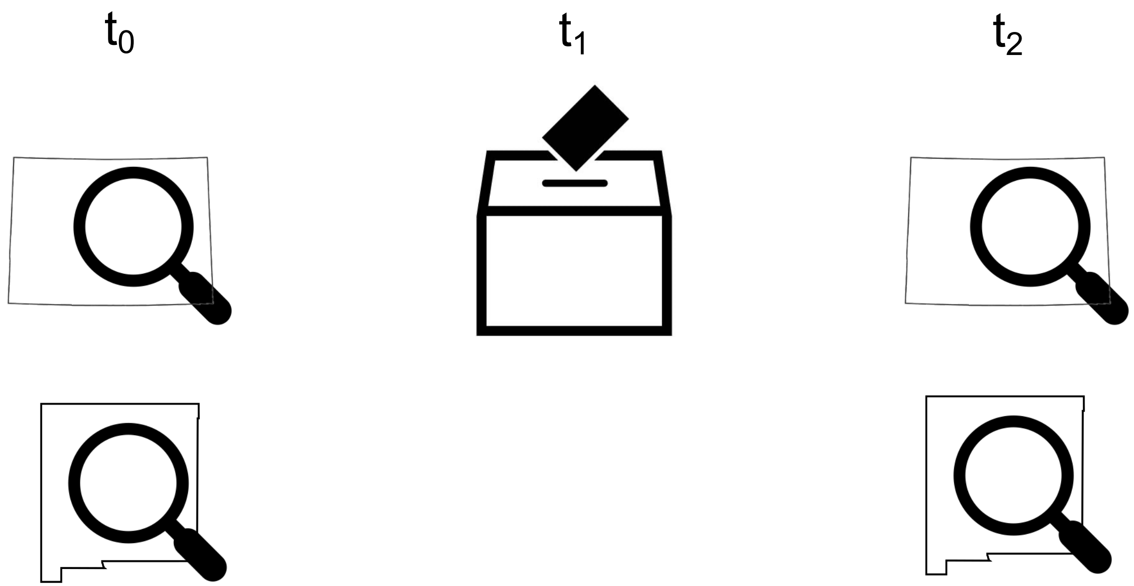
- **History**: Population in CO reacts to Obama's 2012 re-election differently than MI, CO has less abuse
- **Maturation**: Population in CO matures faster than MI, MI has greater risk-taking behavior around opioids
- **Testing**: Pre-2014 surveys caused MI people to be more likely to start abusing opioids
- **Instrumentation**: Method for measuring overdoses differs between CO and MI
- **Mortality**: At-risk people in CO are more likely to move than MI, more of them to drop out of statistics
- **Regression**: MI had unusually high rates that had to come down, regardless of treatment/control



https://conjointly.com/kb/multiple-group-threats/

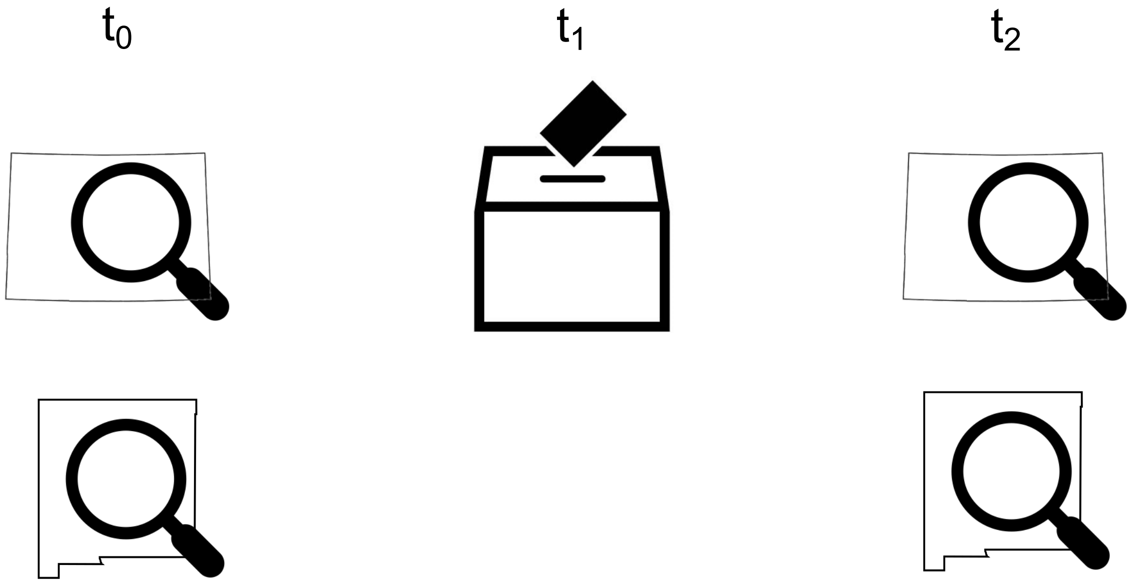
Alternative multiple group design

- **Multiple group designs**: Compare a treatment group to a control group
  - Identify a state similar to Colorado but did not legalize cannabis: New Mexico
  - Measure opioid use before legalization in Colorado and New Mexico (past – 2014)
  - Identify when cannabis legalization started in Colorado (2014)
  - Measure opioid use after legalization in Colorado and New Mexico (2015 – present)
  - Compare CO & NM values after legalization to CO & NM values from before to evaluate effect of treatment



Threats to internal validity – social interaction

- **Diffusion/Imitation of Treatment**: NM sees CO, enforcement becomes lax
- **Compensatory Rivalry**: NM sees CO, starts similar anti-OD program
- **Resentful Demoralization**: NM sees CO, increases enforcement
- **Compensatory Equalization**: Federal agency sees CO, increases NM’s enforcement budget



https://conjointly.com/kb/social-interaction-threats/

Notebook

Notebook

- Download ”us_counties.csv” and “Week 13 – Lecture.ipynb”

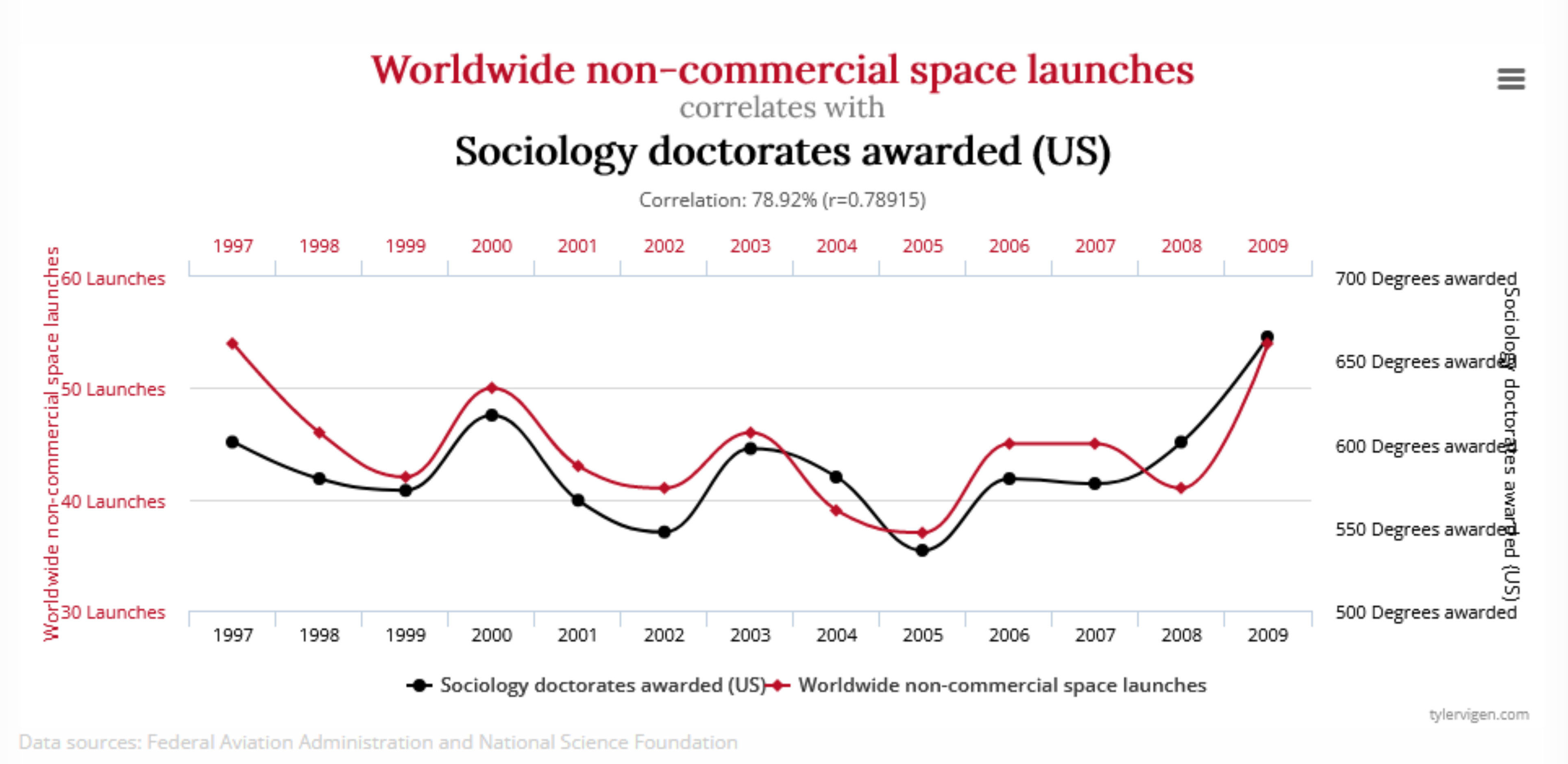
- Analyzing annual county-level opioid deaths from 1999 to 2020
  - Data from the CDC WONDER database
- Reshaping data and computing state-level rates
- Visualizing differences between states legalizing cannabis in 2014 and other states
- Single-group design for Colorado
- Multiple group design comparing Colorado to Tennessee and Maryland

Activity

Activity

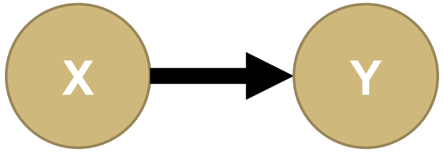
- How spurious are these correlations?
  - Why are they not believable? What would it mean if there was a causal relationship?
  - http://tylervigen.com/spurious-correlations
- Also, guess the correlation
  - http://guessthecorrelation.com/

Space launches and sociology doctorates

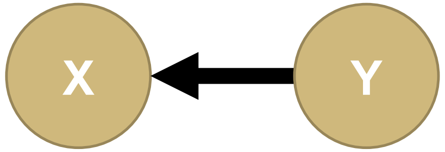


Alternative causal relationships

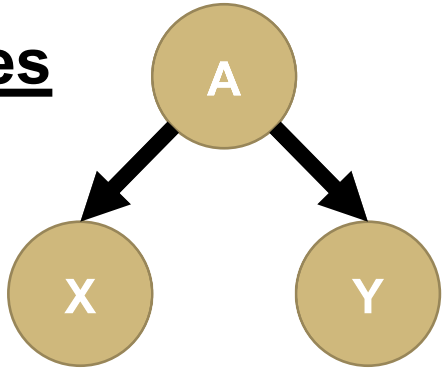
<u>Space launches correlated with sociology doctorates</u>



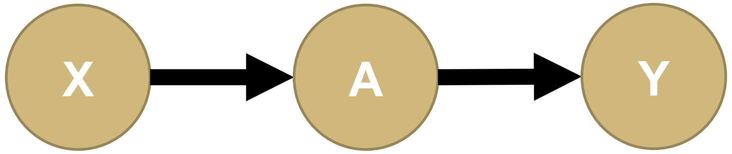
Simple Causation
PhDs (X) cause launches (Y)



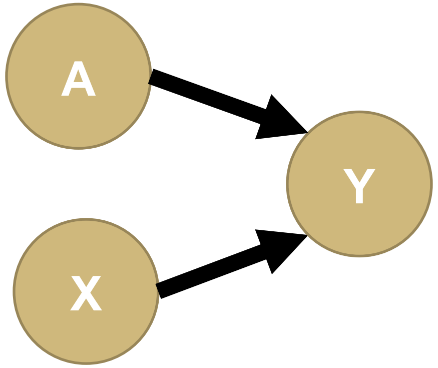
Reverse Causation
Launches cause PhDs



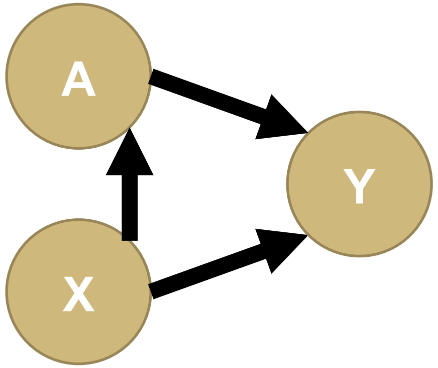
Confounding
Funding (A) causes PhDs and launches



Mediation
PhDs causes researchers (A)
which causes launches



Multiple causation
PhDs and crew dynamics (A)
causes launches



Interaction
PhDs causes crew dynamics (A)
and launches,
Crew dynamics (A) causes launches

Next class

Next Class

- Review concepts and exercises from last class
  - Complete “Thursday Questions” form! https://forms.gle/kQRJdn9XXBMRKw8t6 (ungraded/optional)
- Time to brainstorm and work on Weekly Assignment 13 and Module Assignment 05
- Weekly quiz at the end of class (12:00–12:30)
- Upcoming deadlines
  - Weekly Assignment 13 due Sunday, April 10 before midnight
  - Module Assignment 05 due Wednesday, April 13 before midnight