

# **1.1 – Introduction to Econometrics**

## **ECON 480 • Econometrics • Fall 2022**

Dr. Ryan Safner

Associate Professor of Economics

[safner@hood.edu](mailto:safner@hood.edu)

[ryansafner/metricsF22](https://ryansafner/metricsF22)

[metricsF22.classes.ryansafner.com](https://metricsF22.classes.ryansafner.com)



# About Me



- Ph.D (Economics) – George Mason University, 2015
- B.A. (Economics) – University of Connecticut, 2011
- Specializations:
  - Law and Economics
  - Austrian Economics
- Research interests
  - modeling innovation & economic growth
  - political economy & economic history of intellectual property

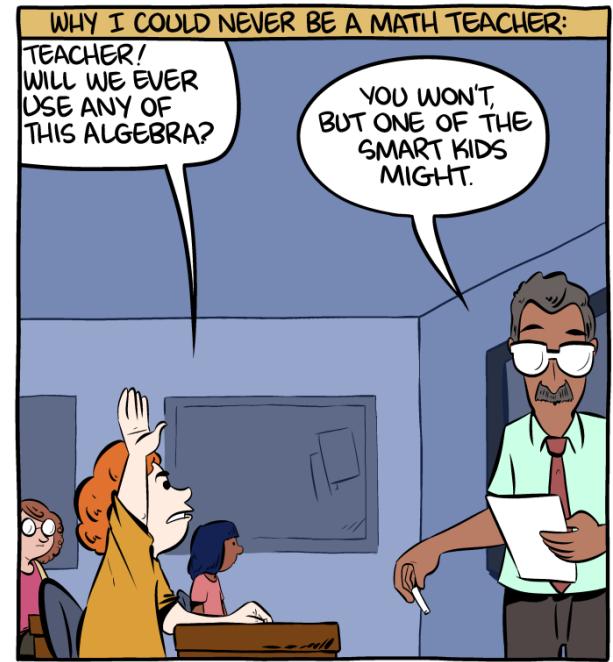


# The Reason I'm Busy Behind the Scenes



# What is Econometrics?

# Why Everyone, Yes *Everyone*, Should Learn Statistics



SMBC

THIS IS WHY PEOPLE SHOULD LEARN STATISTICS:



SMBC



# We're Not So Good at Statistics: Votes I

- Votes in the U.S. House of Representatives in favor of **passing** the *Civil Rights Act of 1964*:

Democrat	Republican
61%	80%

- Simple enough: “on average, Republicans tended to vote for passage more than Democrats”



# We're Not So Good at Statistics: Votes

- Votes in the U.S. House of Representatives in favor of **passing** the *Civil Rights Act of 1964*:

	<b>Democrat</b>	<b>Republican</b>
<b>North</b>	<b>94%</b>	85%
	(145/154)	(138/162)
<b>South</b>	<b>7%</b>	0%
	(7/94)	(0/10)
<b>Overall</b>	<b>61%</b>	<b>80%</b>
	(152/248)	(138/172)

- Larger proportion of Democrats ( $\frac{94}{248}$ , 38%) than Republicans ( $\frac{10}{172}$ , 6%) were from South
- The 7% of southern Democrats voting *for* the Act dragged down the Democrats' *overall* percentage more than the 0% of southern Republicans



# We're Not So Good at Statistics: Kidney Stones

- Suppose you suffer from kidney stones, your doctor offers you **treatment A** or **treatment B**
- In clinical trials, **Treatment A** was effective for a higher percentage of patients with *large* stones and a higher percentage of patients with *small* stones
- **Treatment B** was effective for a larger percentage of patients overall than **treatment A**
- Wait, what?



# We're Not So Good at Statistics: Kidney Stones

From a real [medical study](#):

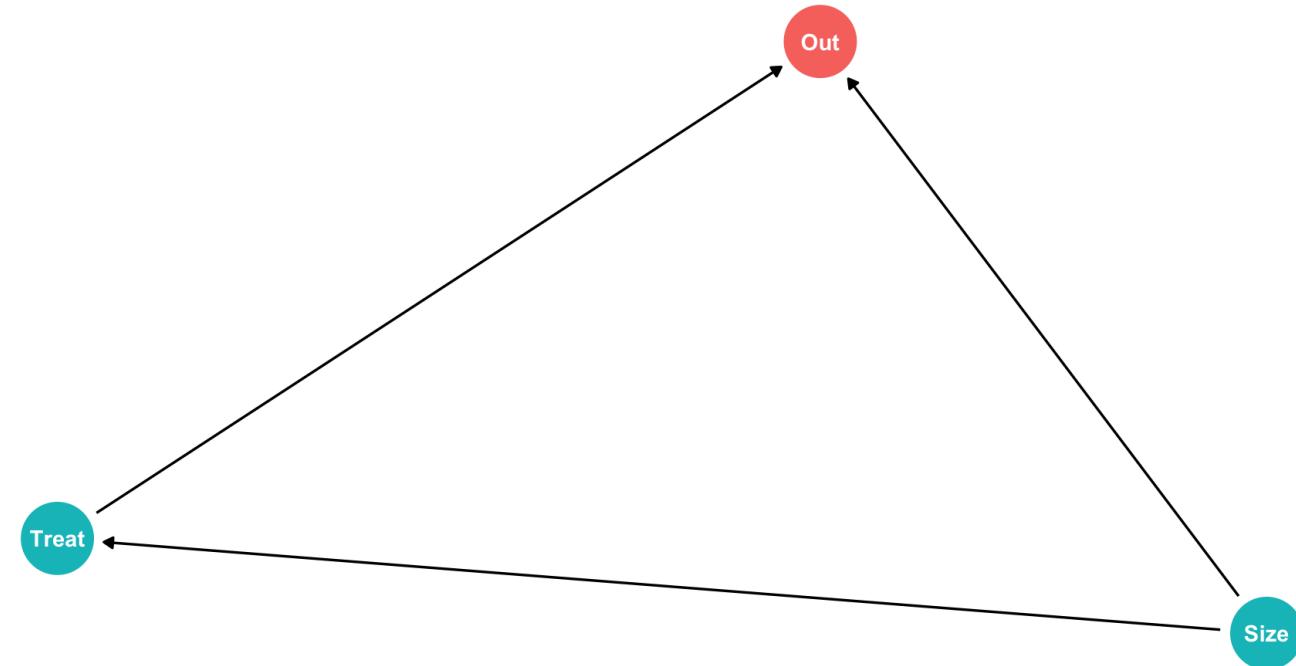
	Treatment A	Treatment B
<b>Small Stones</b>	<b>93%</b> (81/87)	87% (234/270)
<b>Large Stones</b>	<b>73%</b> (192/263)	69% (55/80)
<b>Overall</b>	78% (273/350)	<b>83%</b> (289/350)

C R Charig, D R Webb, S R Payne, and J E Wickham, 1986, "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy," *Br Med J (Clin Res Ed)* 292(6524): 879–882.

- The sizes of the two groups (i.e. who gets A vs B) are *very* different



# We're Not So Good at Statistics: Kidney Stones



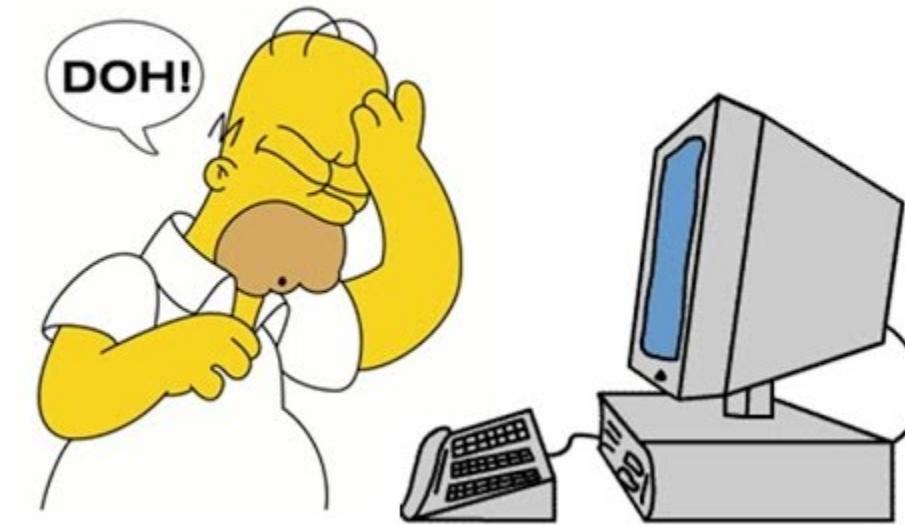
- The sizes of the two groups (i.e. who gets A vs B) are *very* different
- A **lurking variable** in the study is the severity of the case: doctors tended to give treatment B for less severe cases



# Simpson's Paradox

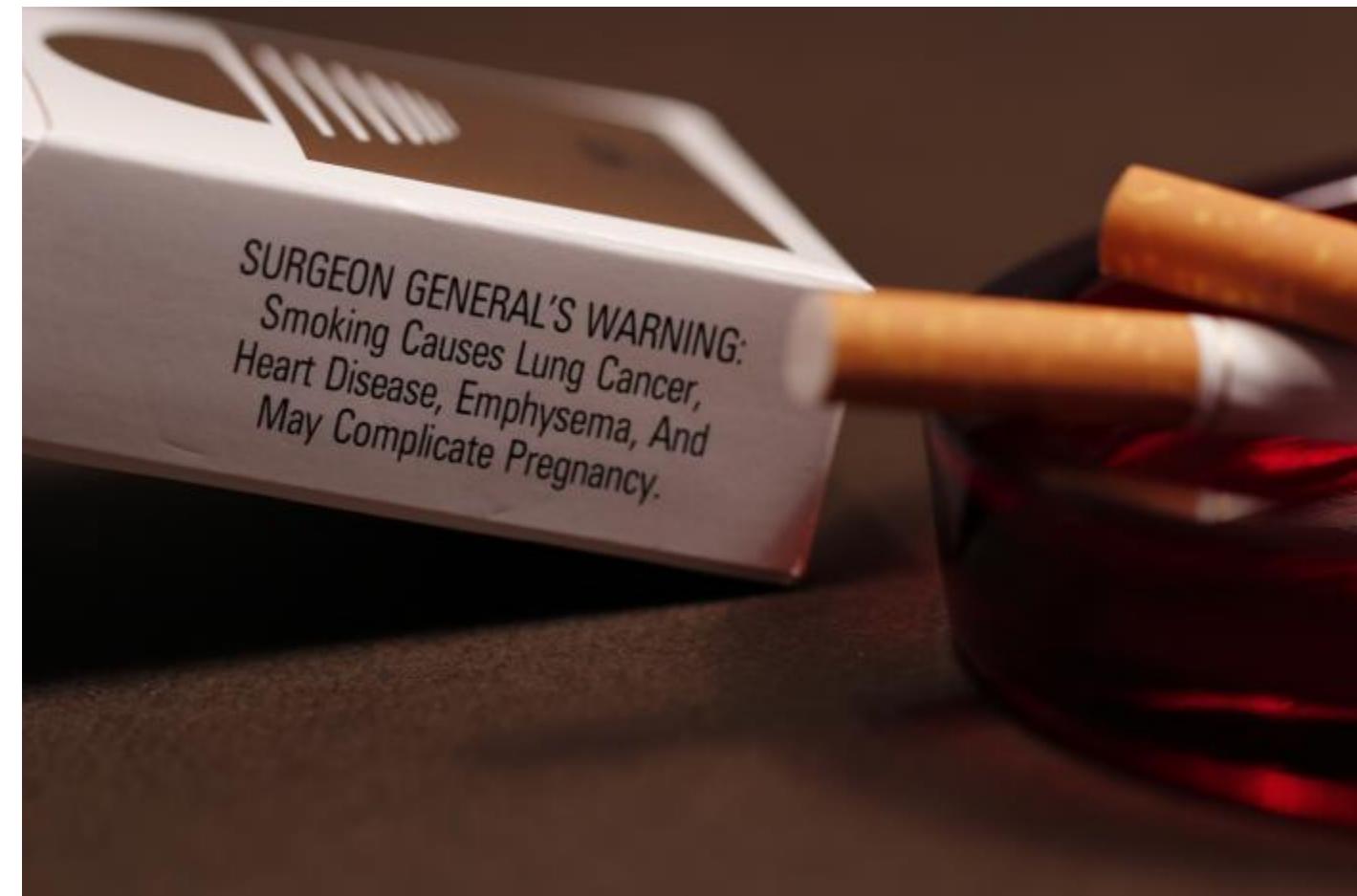


**Simpson's Paradox:** The correlation between two variables can change (even reverse!) when additional variables are considered]



# We're Not so Good at Statistics: Smoking

- 1964: U.S. Surgeon General issued a **report** claiming that cigarette smoking causes lung cancer
- Evidence based primarily on *correlations* between cigarette smoking and lung cancer

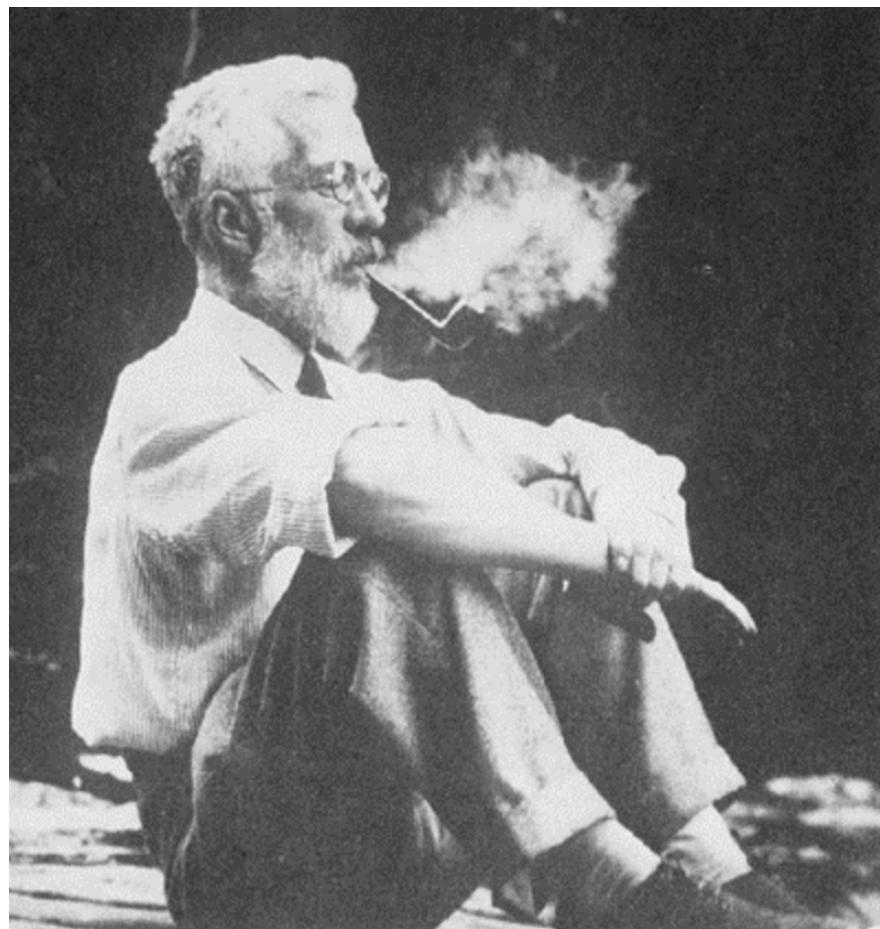


# We're Not so Good at Statistics: Smoking

- Tobacco companies attacked the report, naturally



# We're Not so Good at Statistics: Smoking



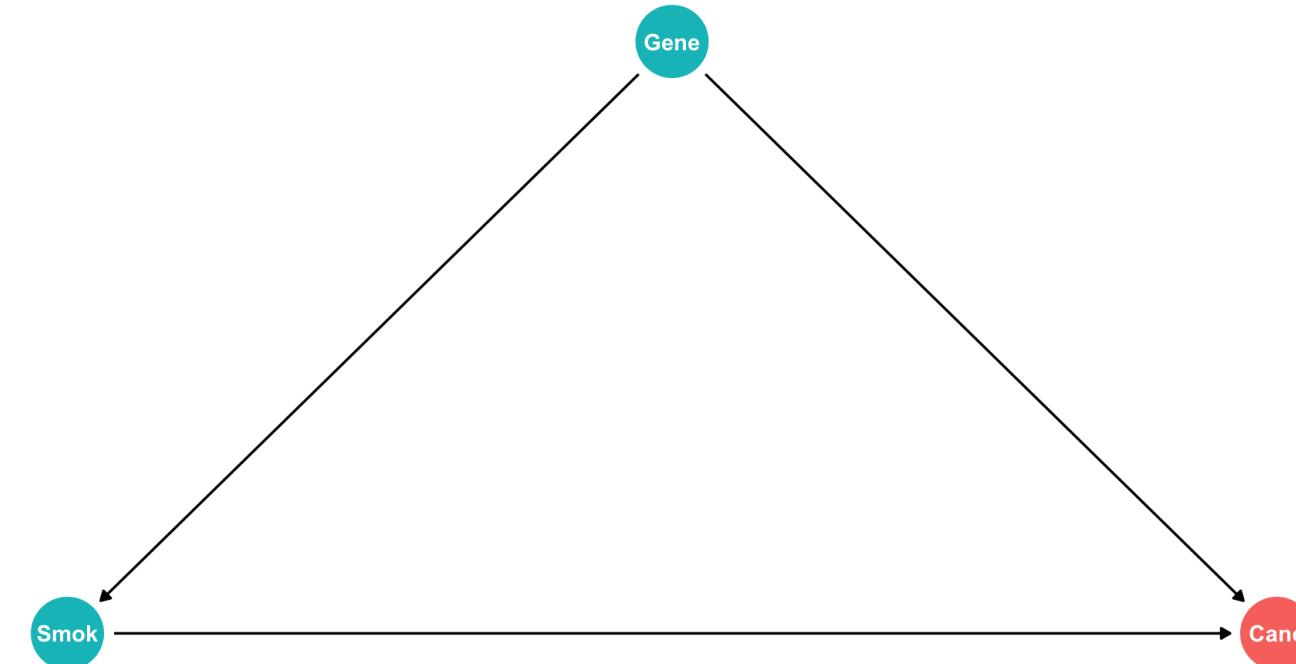
- But **so did R. A. Fisher**, the “father of modern statistics”

Ronald A. Fisher  
1890–1924



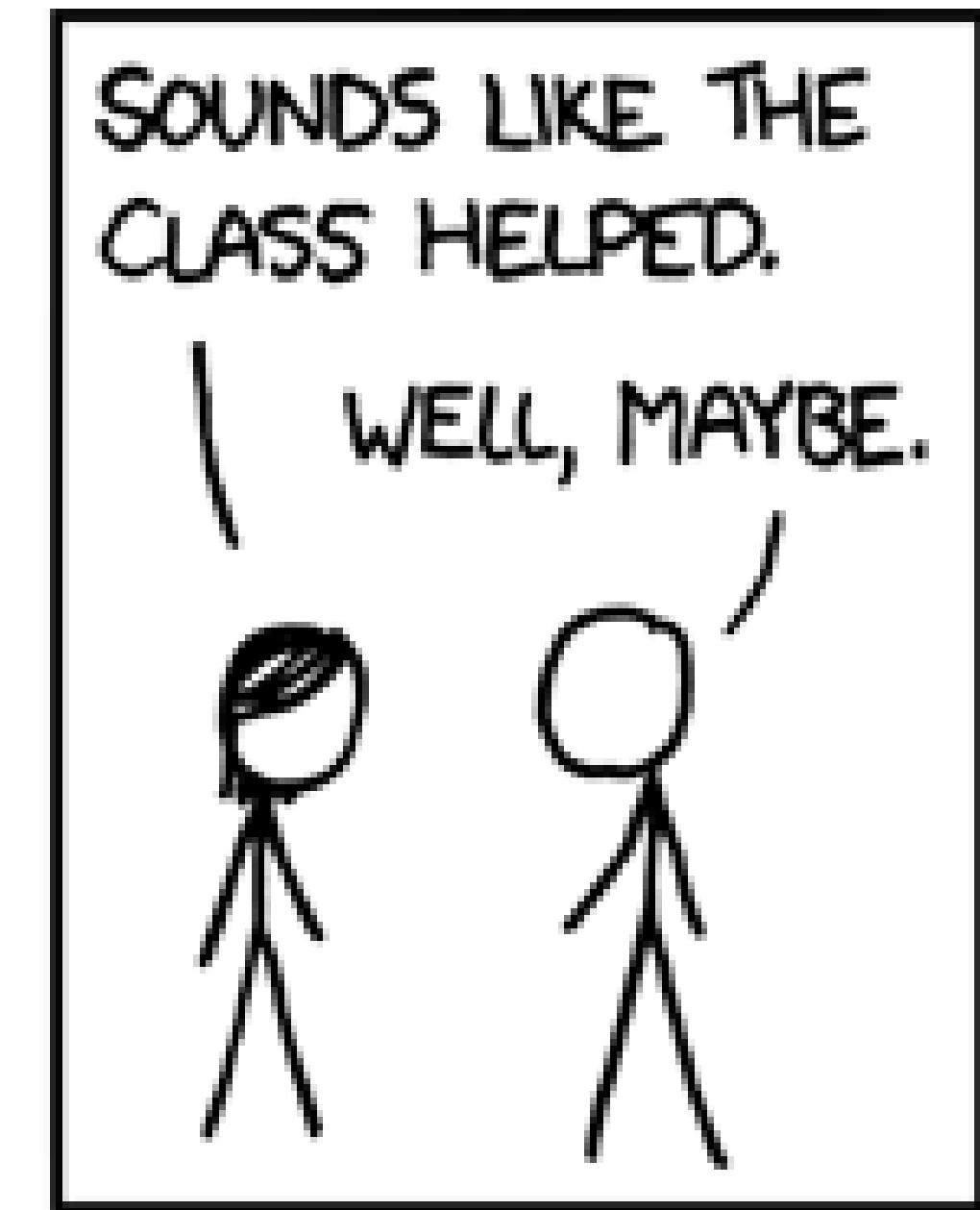
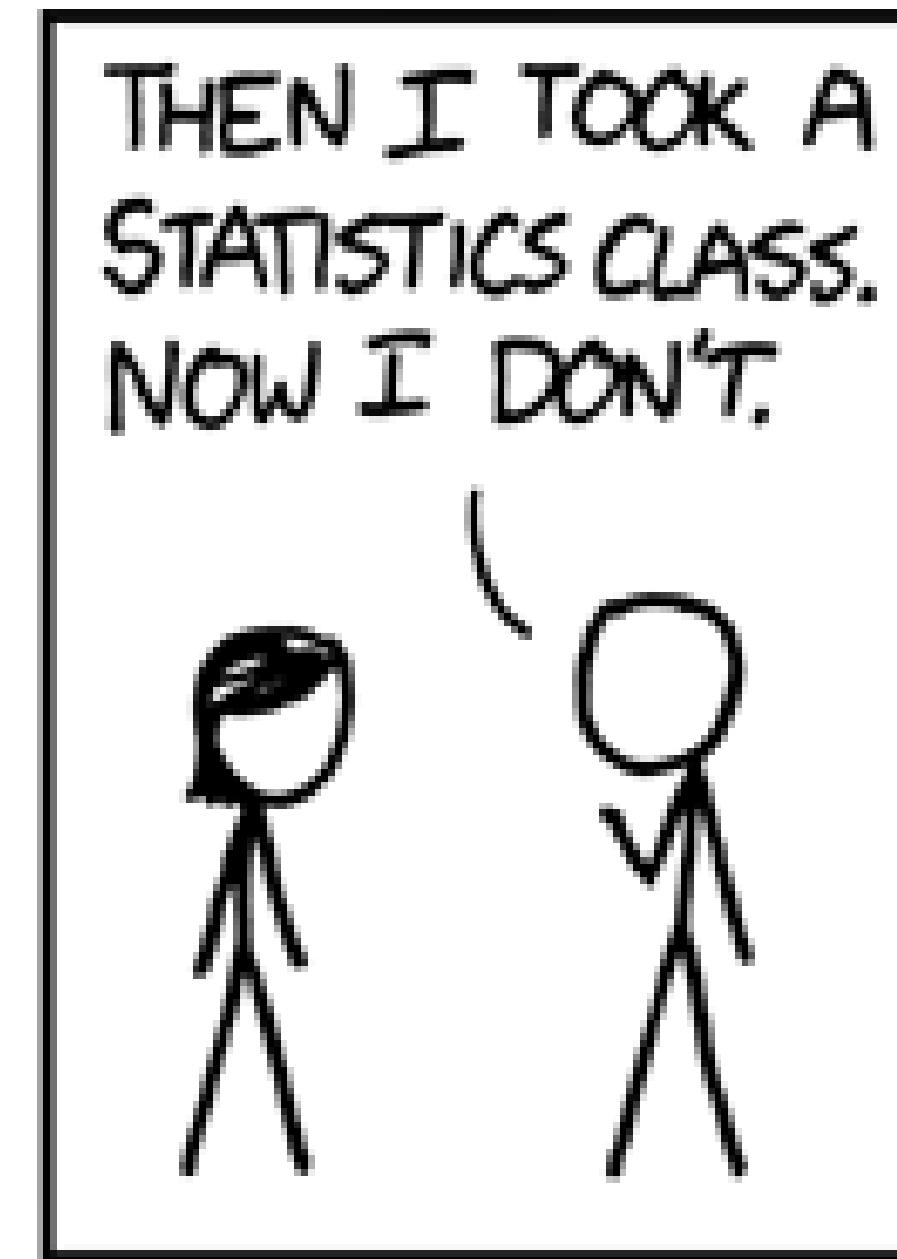
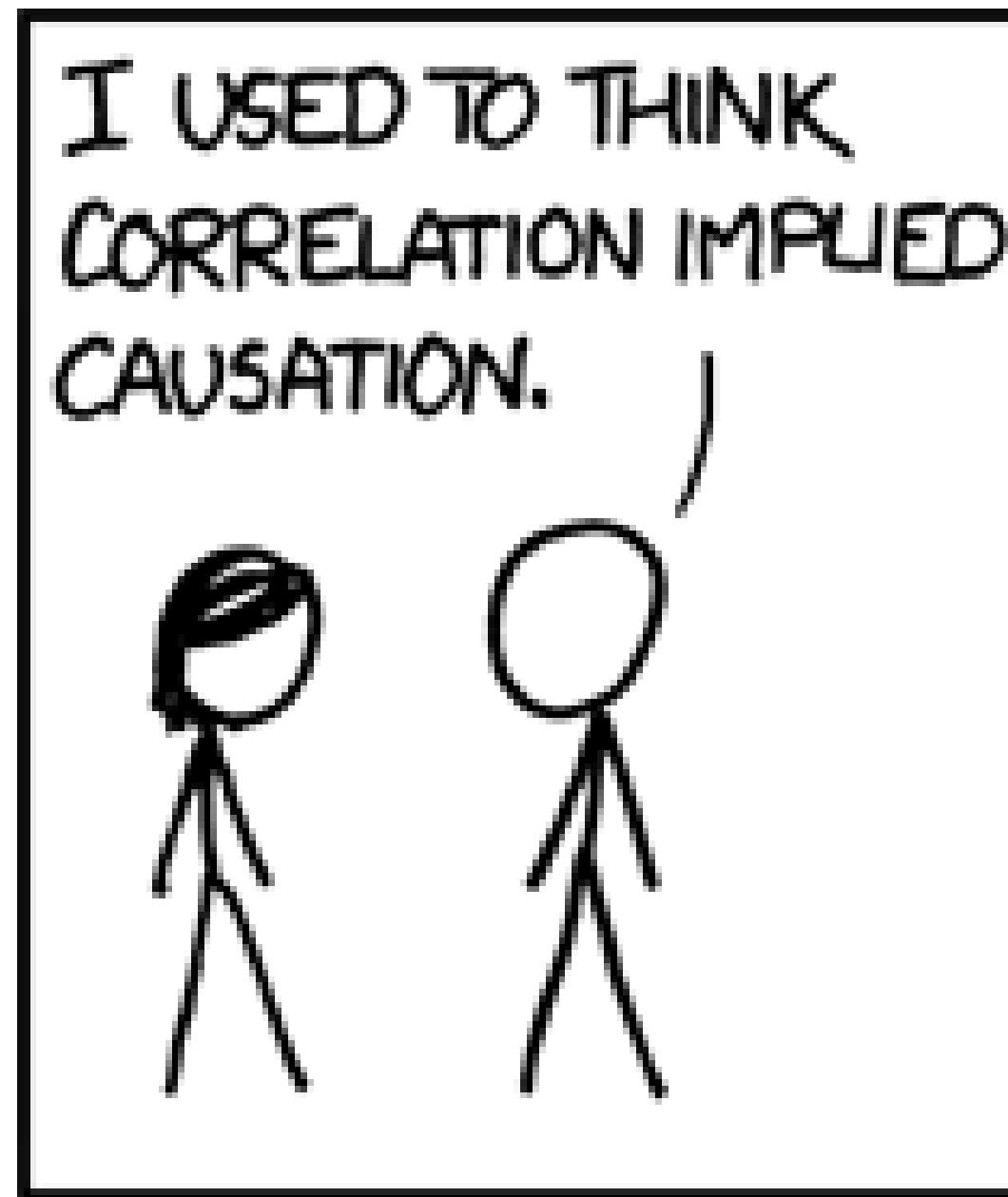
# We're Not so Good at Statistics: Smoking

- There could be a confounding variable (“smoking gene”) that causes *both* lung cancer *and* the urge to smoke
- Would imply: decision to smoke or not would have *no impact* on lung cancer!
- Correlation between smoking and cancer is spurious!



# Correlation Does Not Imply Causation

- The goal of every intro statistics class ever



XKCD: Correlation



# Correlation Does Not Imply Causation

**Number of people who drowned by falling into a pool**  
correlates with  
**Films Nicolas Cage appeared in**



Spurious Correlations



# Correlation Does Not Imply Causation...

- It's always good to be skeptical of causal claims
- But this is actually where **econometrics** shines



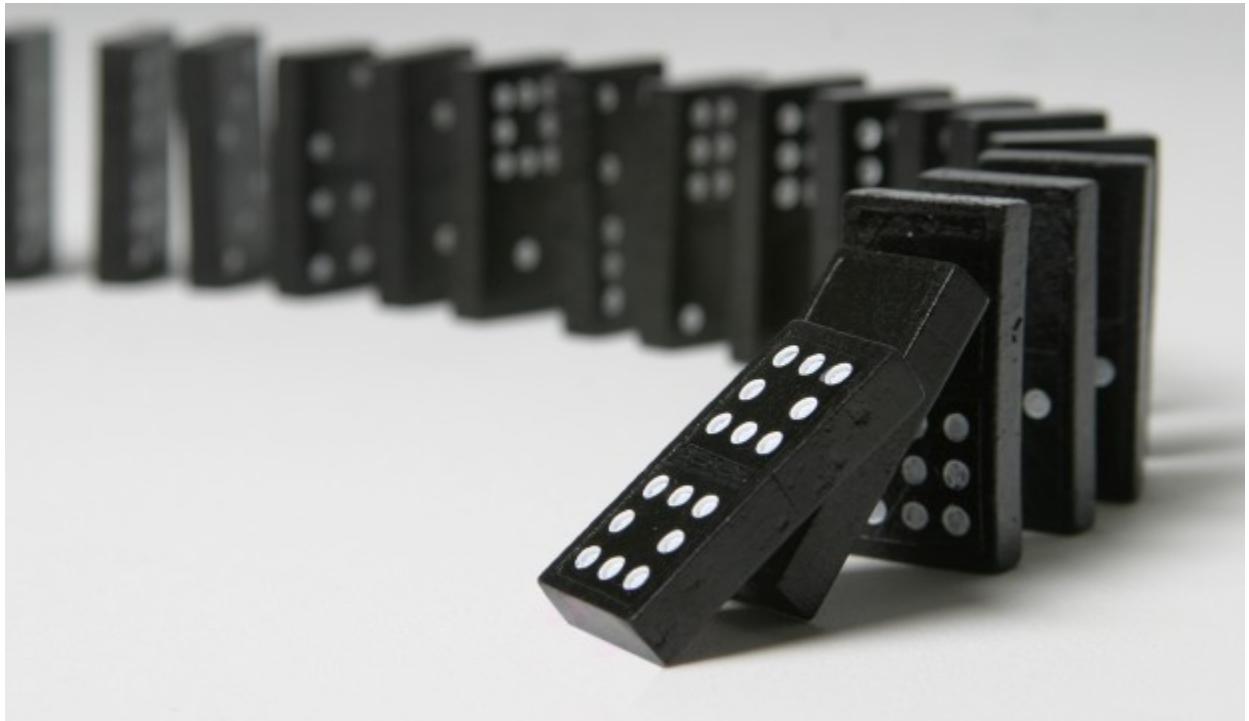
# Econometrics

- **Econometrics** is the application of statistical tools to *quantify* economic relationships in the real world
- Uses real data to
  - test economic hypotheses
  - quantitatively estimate the magnitude of relationships between economic variables
  - forecast future events



# Econometrics and Causal Inference

- What sets econometrics apart from mere statistics (or uses of statistics in other disciplines) is its role in **causal inference**
- We can, with proper tools and interpretations, make *quantitative causal* claims
  - about the effects of individual choices
  - about the effects of policy interventions
  - about the impact of political institutions
  - about economic history and economic development
  - etc...



# Causal Inference: Examples

A 50% increase in police presence in a metropolitan area lowers crime rates by 15%, on average<sup>1</sup>

Being an incumbent in office raises the probability of re-election by 40-45 percentage points<sup>2</sup>

European cities with at least one printing press in 1500 were at least 29% more likely to become Protestant by 1600<sup>3</sup>



1. Klick, Jonathan and Alexander Tabarrok, 2005, "Using Terror Alert Levels to Estimate the Effect of Police on Crime," *Journal of Law and Economics* 48(1): 267-279

2. Lee, David S, 2001, "The Electoral Advantage to Incumbency and Voters' Valuation of Politicians' Experience: A Regression Discontinuity Analysis of Elections to the U.S," *NBER Working Paper* 8441



# Example 1: Education

## Example

- Does reducing class sizes improve student performance?
- ...
- A policy-relevant tradeoff with a budget constraint
- What is the *precise* effect of class size on performance?
- Is it worth hiring new teachers and building more schools over?



# Example 2: Discrimination in Lending

## Example

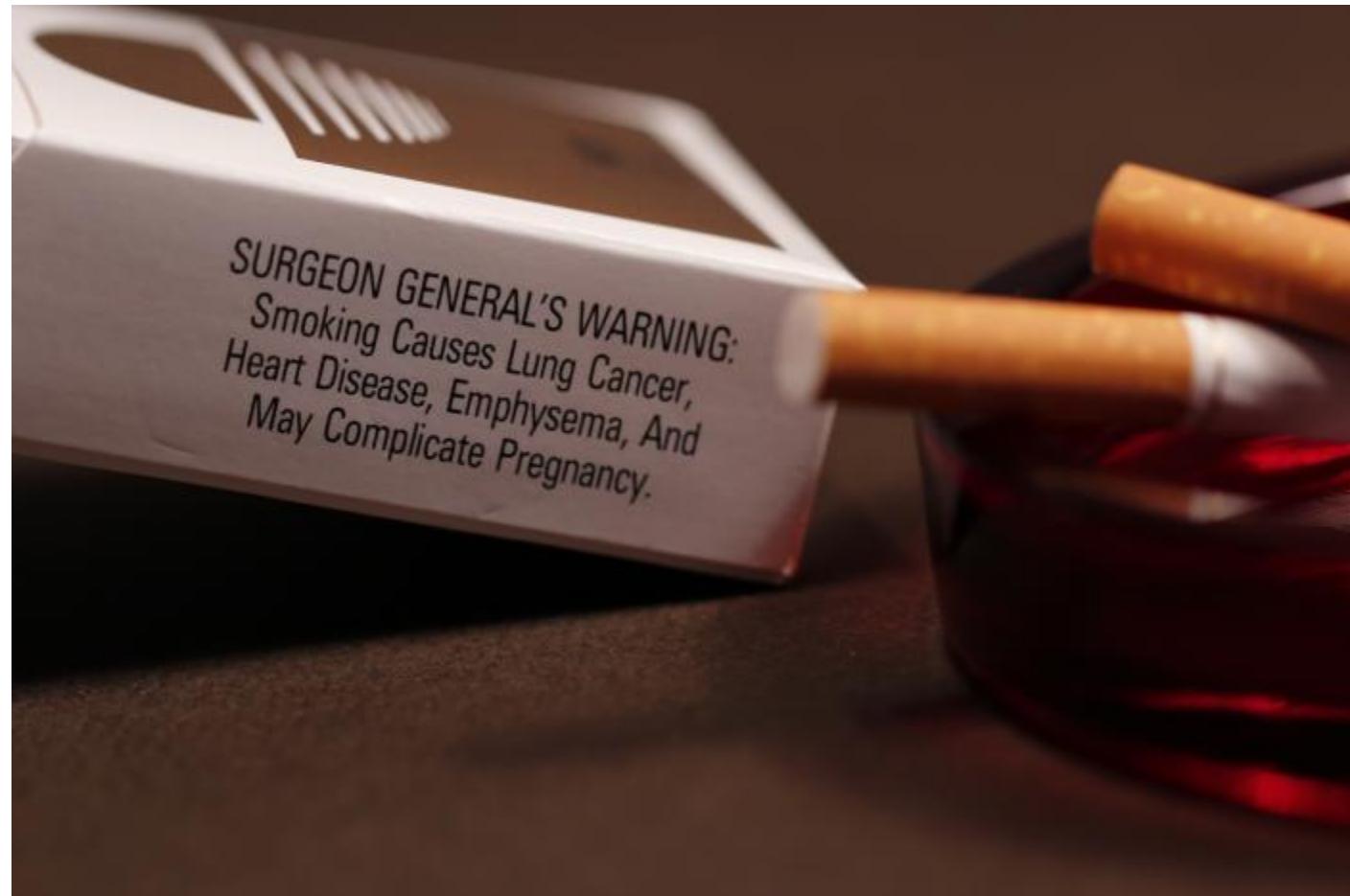
- Is there racial discrimination in home mortgage lending?
- ...
- Boston Fed: 28% of African-Americans are denied mortgages compared to only 9% of White Americans
  - Is this due to factors such as credit history, income, or discrimination *purely* because of race?



# Example 3: Public Health and Public Finance

## Example

- How much do state cigarette taxes reduce smoking rates?  
...
- Econ 101: raise price  $\implies$  lower quantity consumed
- What is the *price elasticity of demand* for smoking?
- How much tax revenue will this generate?
- Probably: *Taxes  $\rightarrow$  Smokers*
- Maybe?: *Taxes  $\leftarrow$  Smokers*



# About This Course



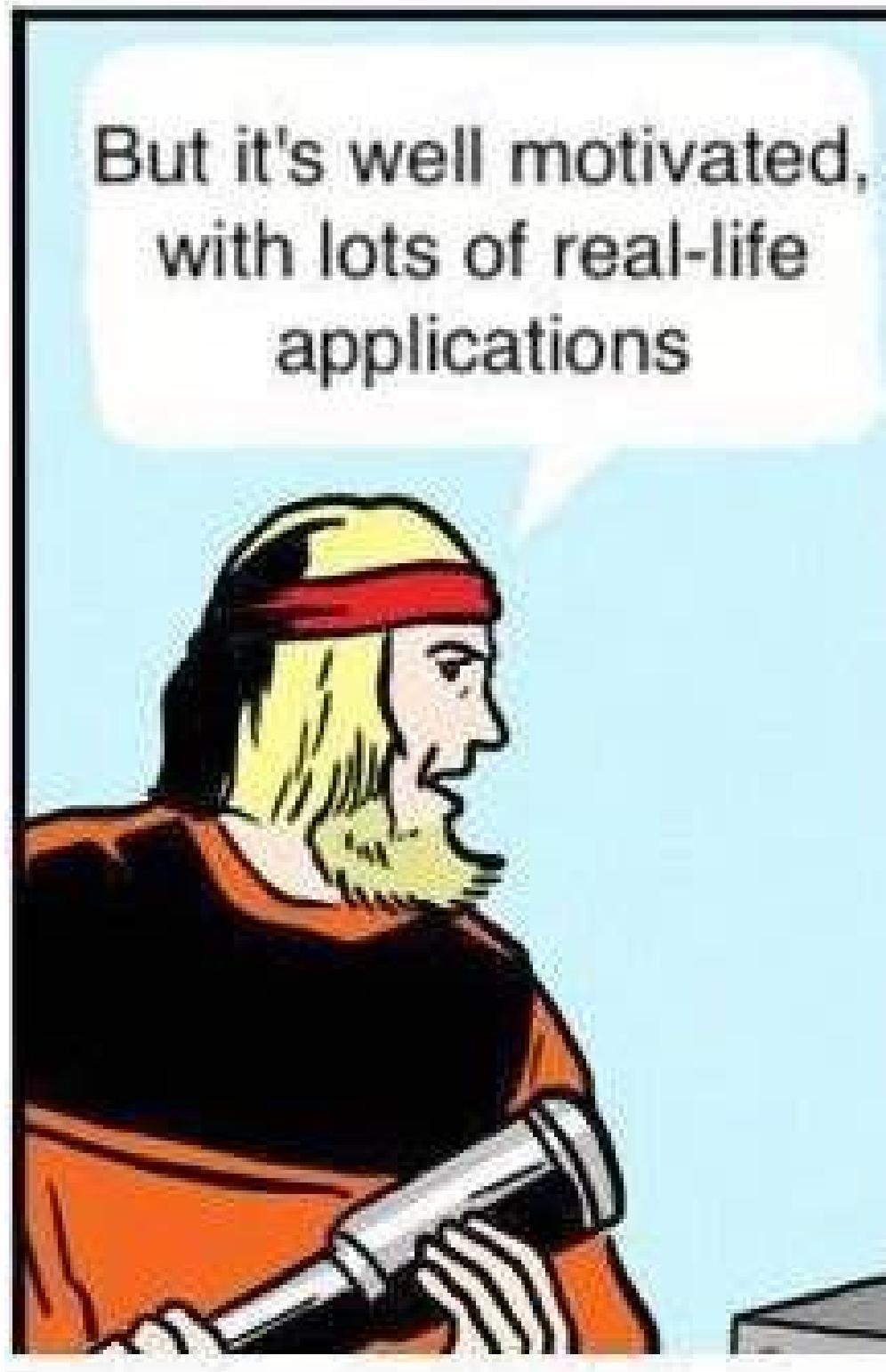
# Real Talk: The Math



# Real Talk: The Math



# Real Talk: The Math



But it's well motivated,  
with lots of real-life  
applications



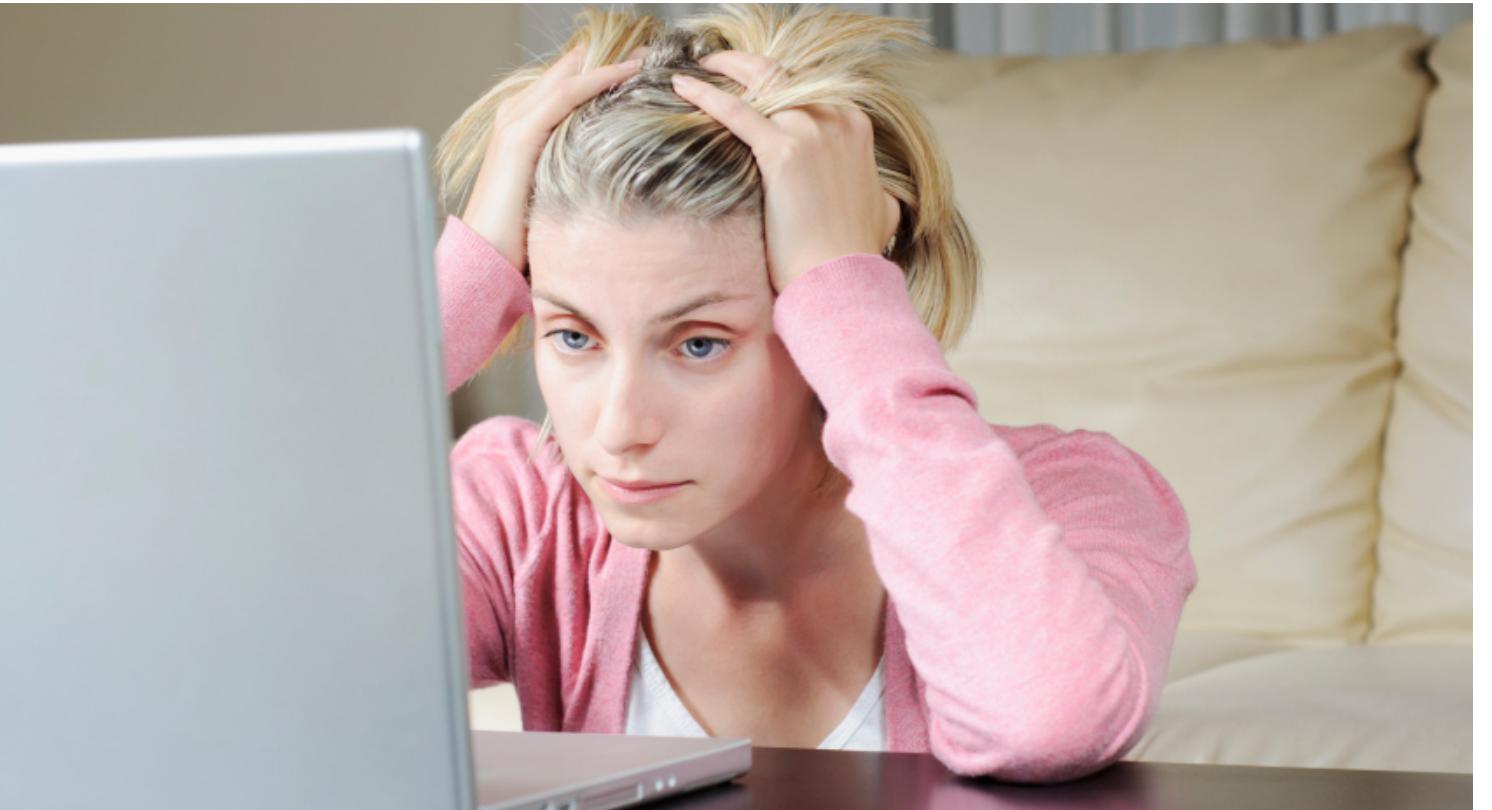
memecenter.com

MemeCenter



# Real Talk: Difficulty

- This will be one of the hardest courses you take at Hood
- There will be moments where you have no idea WTF is going on (*this is normal*)
- Yes, you can still get an **A**



# This Class Is

- **Economics:** take your *preexisting* intuition and models for causal inference
- **Statistics:** add regression and statistical inference
- **Computer Programming:** using [R](#) and [R Studio](#) for analyzing and presenting data ]

Economics

Computer  
Programming

Statistics



# This Class Is

## Old School Statistics Courses

- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- $\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$
- $r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$
- Use pre-cleaned “toy” data, if at all

## Hip New “Data Science” Courses

- 1 `mean(x)`
- 2 `sd(x)`
- 3 `cor(x, y)`

- Import, tidy, and manipulate raw data from scratch (like *real life!*)



# Prerequisites

- **Officially (Courses):**
  - ECON 205
  - ECON 206
  - ECON 305 or ECON 306
  - MATH 112 or ECMG 212
- **Math Skills:**
  - Basic algebra
  - Probability-ish
  - Statistics-ish
- **Computer Science Skills:**
  - None



# What You'll Get Out of This Class

By the end of this semester, you will:

1. understand how to evaluate statistical and empirical claims;
2. use the fundamental models of causal inference and research design;
3. gather, analyze, and communicate with real data in R.



# This Class Opens Doors

Regressions!



[REDACTED]@hood.edu>

to Ryan ▾

Hi Dr. Safner,

I hope all is well and you are enjoying the start to summer. I changed jobs in March to work as a researcher at an investment fund right outside of DC in Virginia.

I often find myself running regressions (unfortunately my boss doesn't understand R so I have to use Python!) and using time series data; however I am in need of some notes from our courses together. Would you be willing to pass along past lecture presentations from Econometrics? I have hard copy notes to go along, but they are less useful than with the slides.

That is probably the single most influential class I have taken, especially given my new job function and I would certainty benefit from a refresher.

Best Regards,

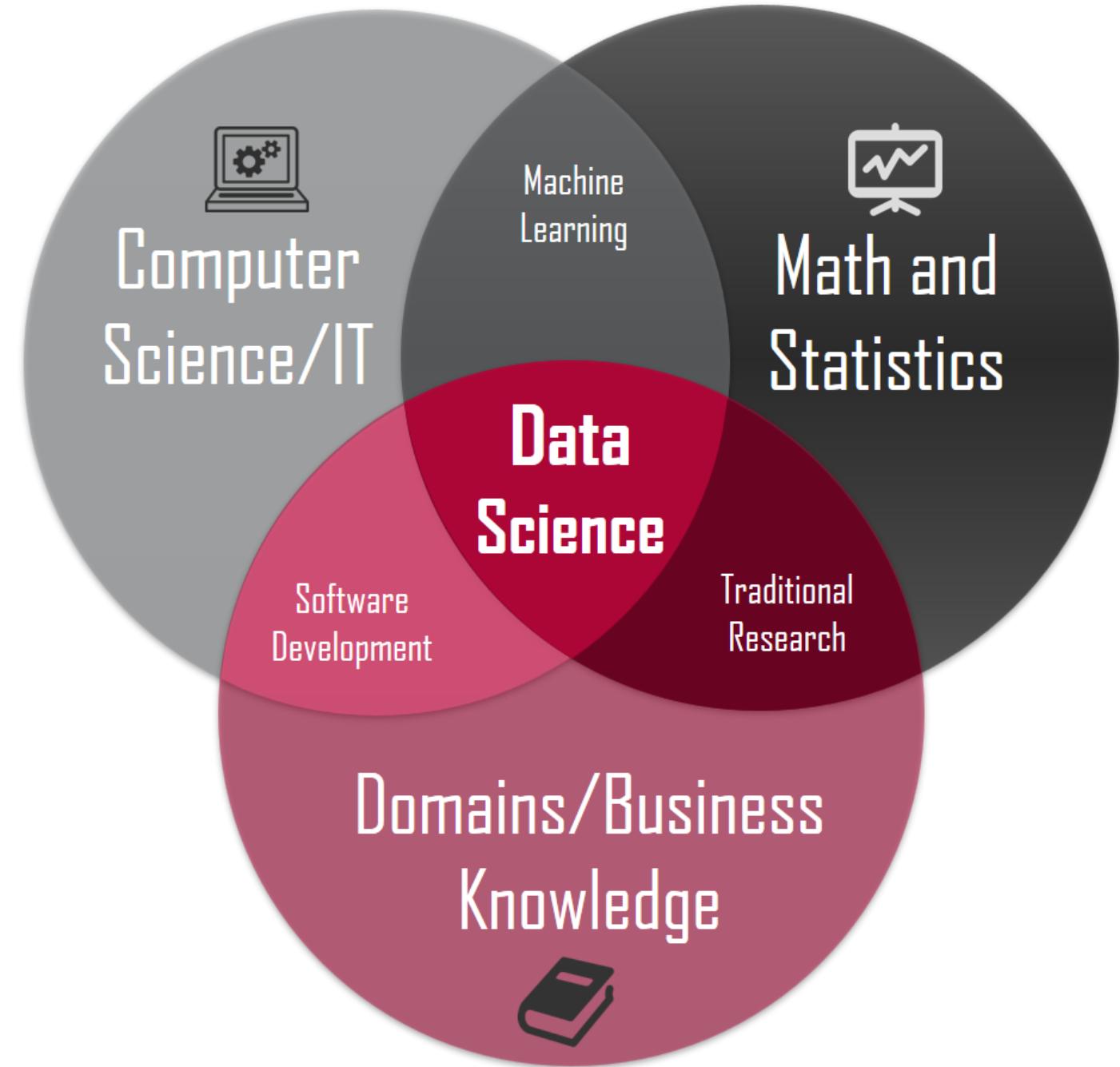
Hood College, B.A. Economics  
Class of 2018



# Building Industry-Demanded Data Science Skills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

— Josh Wills (@josh\_wills) May 3, 2012



# Building Industry-Demanded Data Science Skills

hbr.org

Harvard Business Review

Subscribe

Latest Magazine Popular Topics Podcasts Video Store The Big Idea Visual Library Reading Lists Case Selections



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

**DATA**

# Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Summary Save Share Comment 16 Text Size Print \$8.95 Buy

4/6 FREE ARTICLES LEFT > SUBSCRIBE TO ACCESS THE ARCHIVE

LinkedIn Economic Graph

[Research](#) [Resources](#) [Blog](#) [About](#)

---

6. **Relationship Consultant** (5.5X growth)

- **Top Skills:** Banking, Retail Banking, Loans, Consumer Lending, Credit
- **Where They Work:** [Regions Bank](#), [Merrill Edge](#), [Vanguard](#)
- **Top Industries:** Banking, Financial Services, Insurance
- **Cities Where Demand is High:** Jacksonville, New York City, St. Louis

7. **Data Science Specialist** (5X growth)

- **Top Skills:** Machine Learning, Data Science, Python, R, Apache Spark
- **Where They Work:** [IBM](#), [Facebook](#), [McKinsey & Company](#)
- **Top Industries:** Higher Education, Information Technology & Services, Computer Software
- **Cities Where Demand is High:** New York City, San Francisco, Chicago

8. **Assurance Staff** (5X growth)

- **Top Skills:** Auditing, Accounting, Financial Reporting, Internal Controls
- **Where They Work:** [EY](#), [Plante Moran](#), [Moss Adams](#)
- **Top Industries:** Accounting, Higher Education, Financial Services
- **Cities Where Demand is High:** Detroit, Philadelphia, Boston

9. **Sales Development Representative** (4X growth)

- **Top Skills:** Salesforce, Cold Calling, Software-as-a-Service, Lead Generation, Sales Prospecting

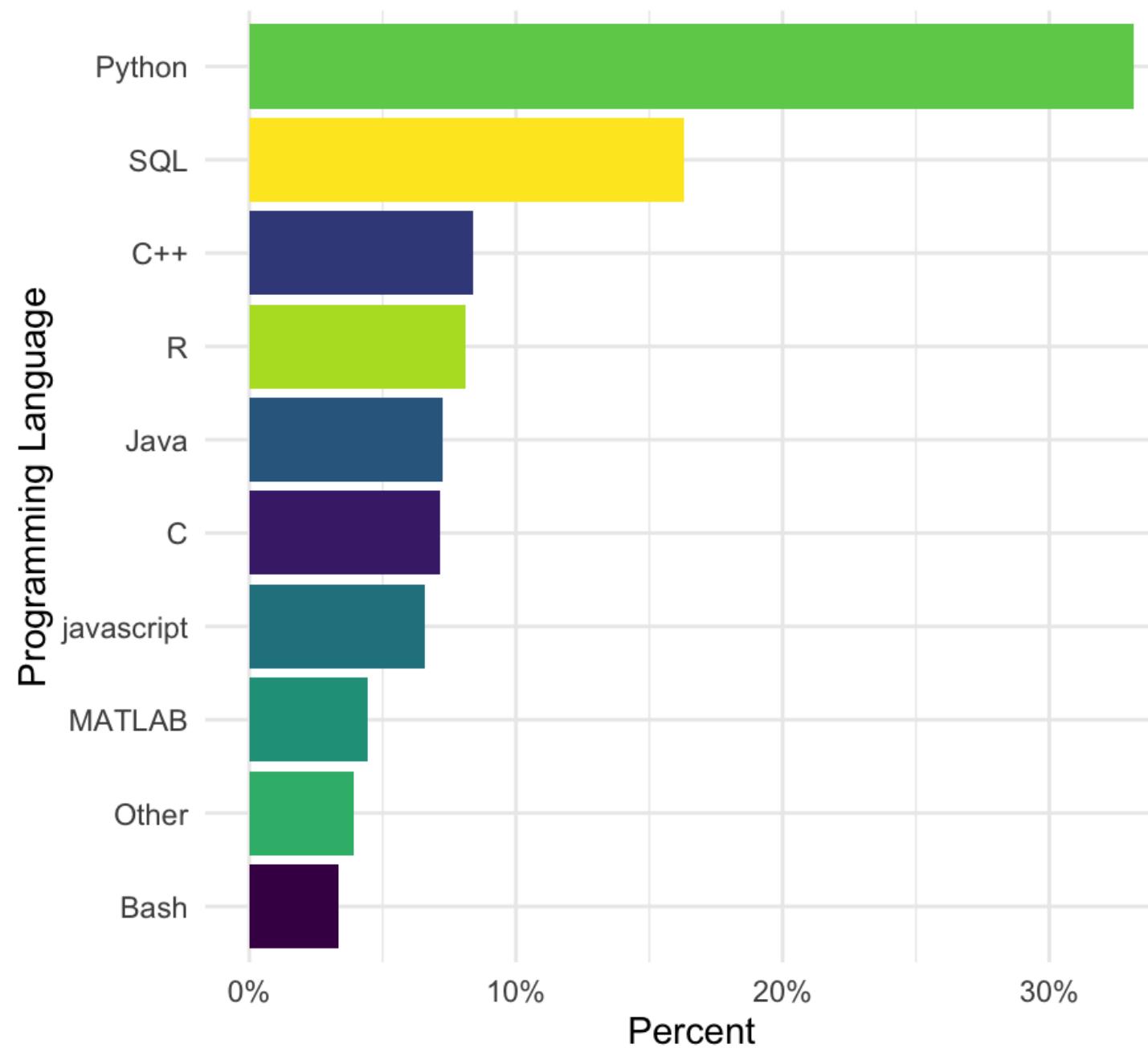
# Harvard Business Review

# LinkedIn 2018 Emerging Jobs Report



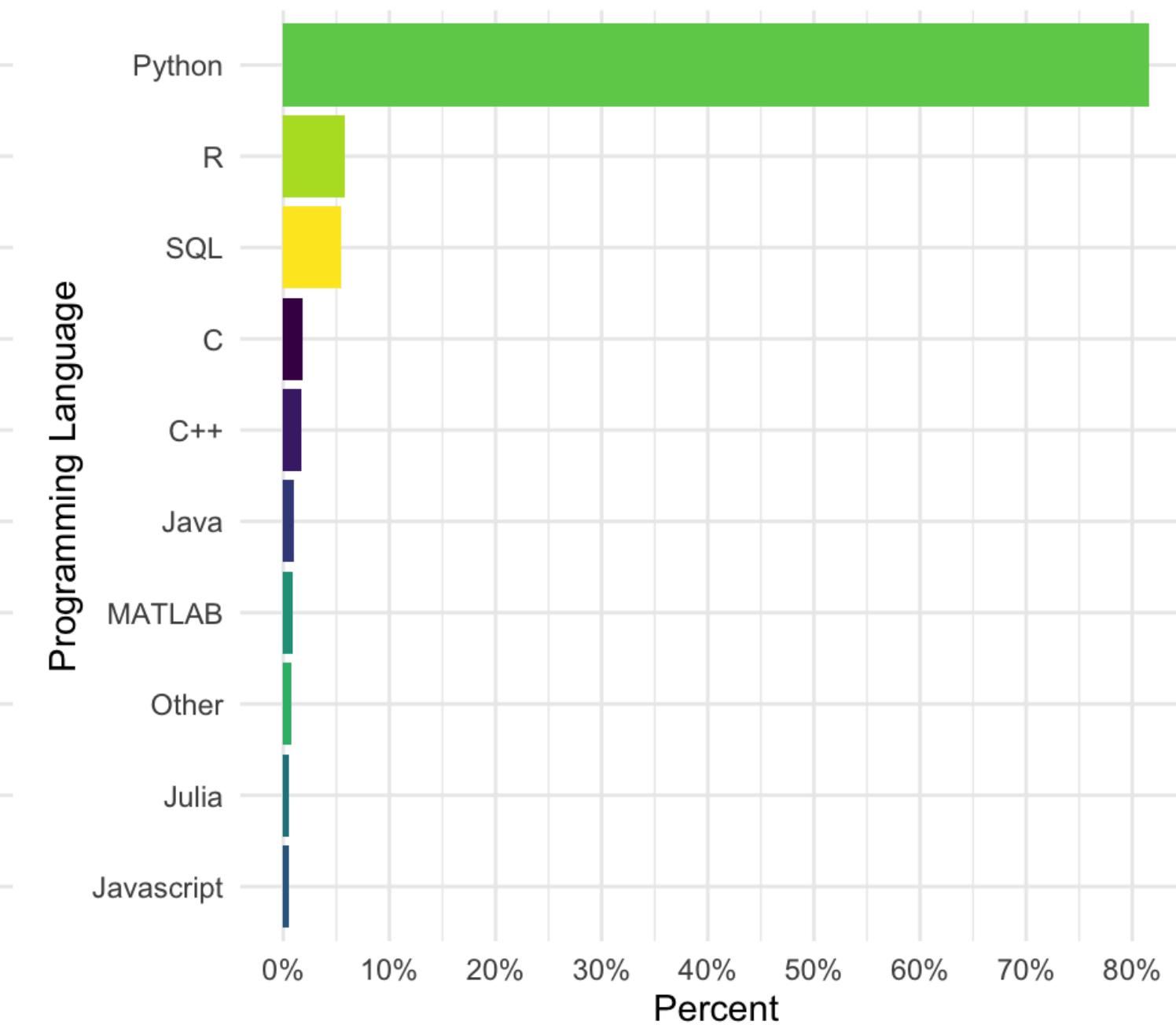
# R Can Be Used for Data Science

Programming Languages Used Most by Data Scientists



Source: 2021 Kaggle Machine Learning and Data Science Survey

Programming Languages Recommended to Learn



Source: 2021 Kaggle Machine Learning and Data Science Survey



# But Remember

- This is an economics course, not a data science course
- Other software economists use (STATA, SASS, SPSS, Excel) is not on here!



# Two Uses For Econometrics

$$Y = f(X)$$

1. **Causal inference**: how changes in  $X$  cause changes in  $Y$

- Care more about accurately estimating  $f$  than getting an accurate  $\hat{Y}$
- Measure the **causal effect** of  $X \mapsto Y$  (e.g.,  $\hat{\beta}_1$ )

2. **Prediction**: predict  $\hat{Y}$  using an estimated  $f$

- Care more about getting  $\hat{Y}$  as accurate as possible,  $f$  is an unknown “black-box”
- **Forecasting**: predict future values of  $Y$



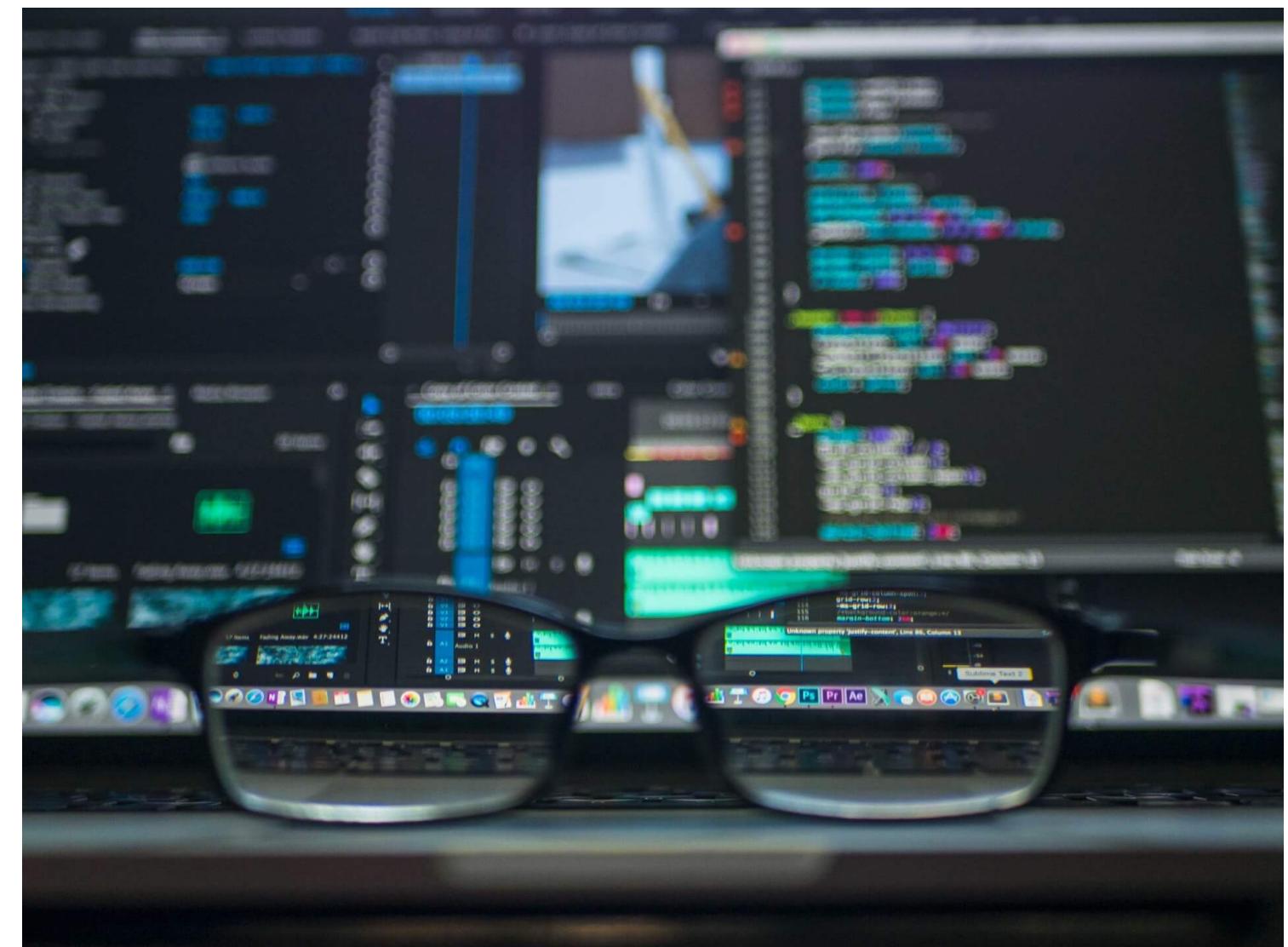
(inflation, sales, GDP)

- **Classification**: predict the *category* of an outcome (success or failure, cat picture or not cat picture)
- We care (in this class at least) only about the first

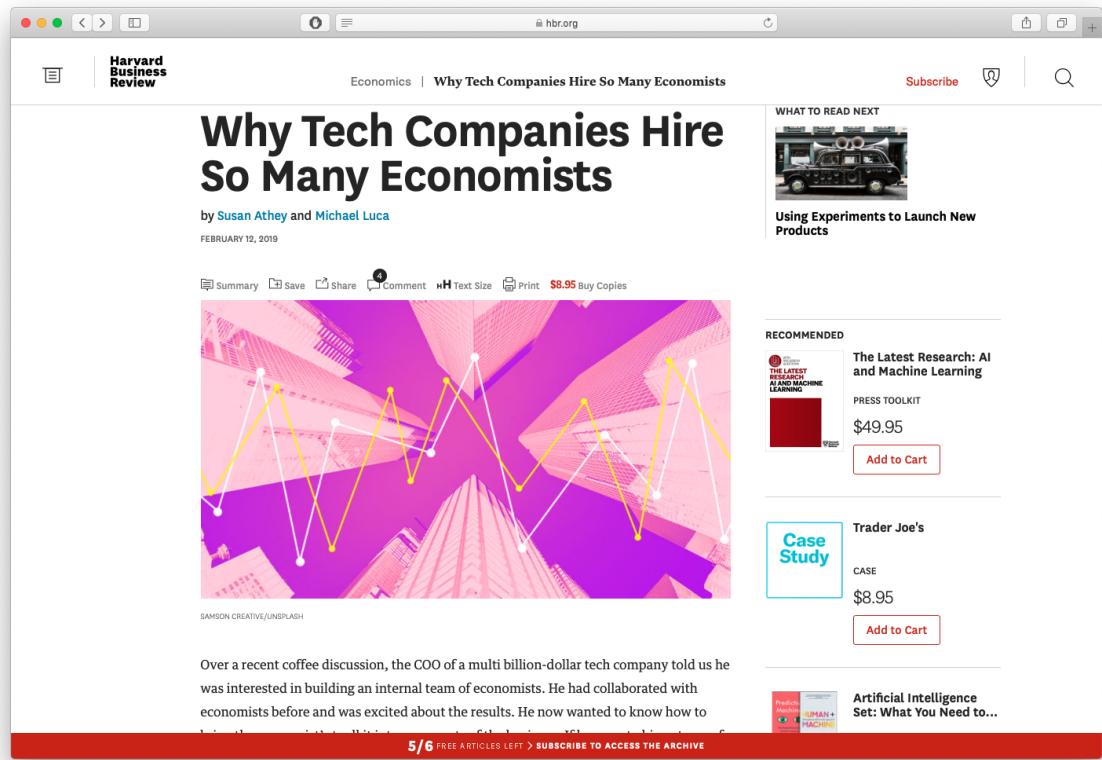


# Causal Inference – Economists' Comparative Advantage

- Machine learning and artificial intelligence are “dumb”<sup>1</sup>
- With the right models and research designs, we *can* say “X causes Y” and quantify it!
- Economists are in a unique position to make *causal* claims that mere statistics cannot



# Causal Inference – Economists' Comparative Advantage



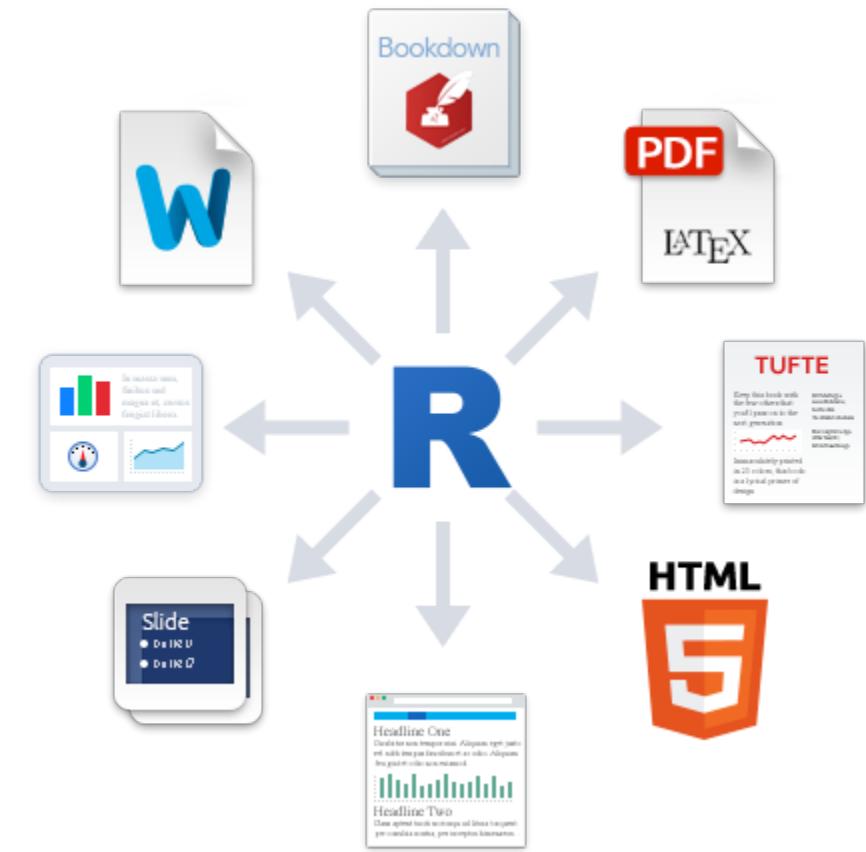
Harvard Business Review

"[T]he field of economics has spent decades developing a toolkit aimed at investigating empirical relationships, focusing on techniques to help understand which correlations speak to a causal relationship and which do not. This comes up all the time – does Uber Express Pool grow the full Uber user base, or simply draw in users from other Uber products? Should eBay advertise on Google, or does this simply siphon off people who would have come through organic search anyway? Are African-American Airbnb users rejected on the basis of their race? These are just a few of the countless questions that tech companies are grappling with, investing heavily in understanding the extent of a causal relationship."



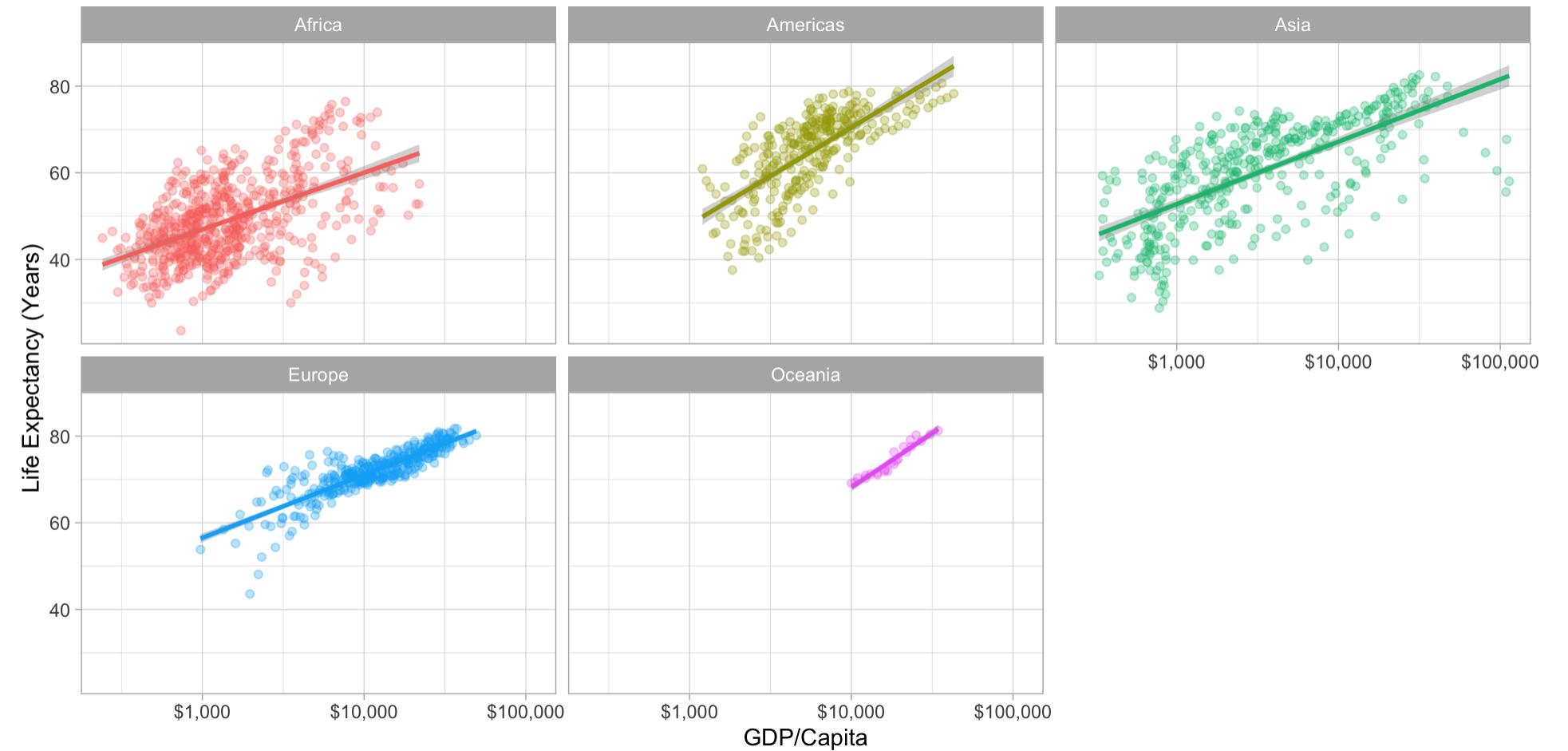
# Building Good Workflow Habits

- I will show you the tools to make your workflow:
  - Reproducible
  - Computer- and Human-Readable (!)
  - Automated
  - All in one program



# For Example

Output    Code



# Assignments

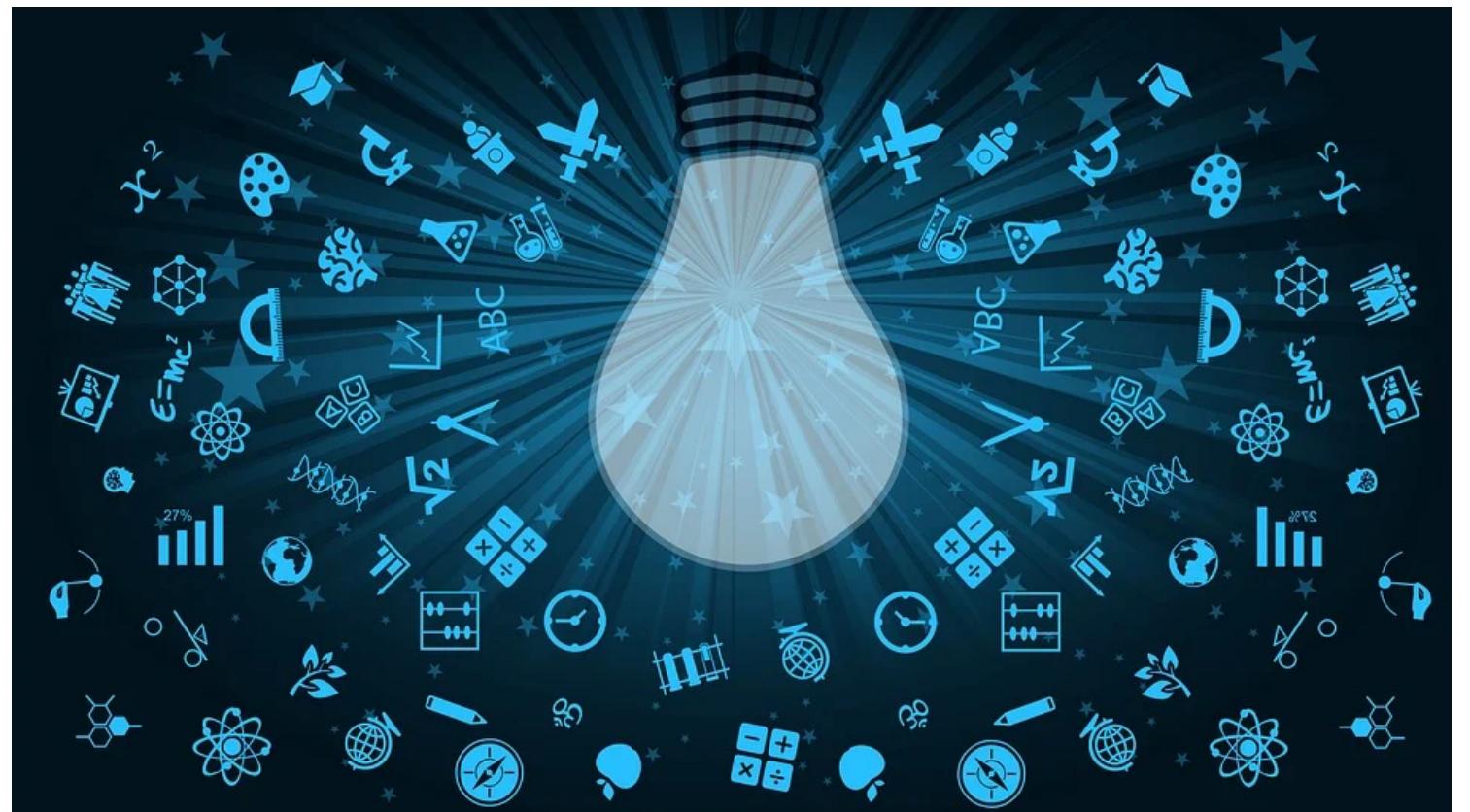
- Research project:
  - Come up with a testable research question
  - Find data
  - Analyze data
  - Present your results (in writing and verbally)
- HWs
- Midterm, Final exam

Assignment	Percent
1 Research Project	30%
n Homeworks (Average)	25%
1 Midterm	20%
1 Final	25%

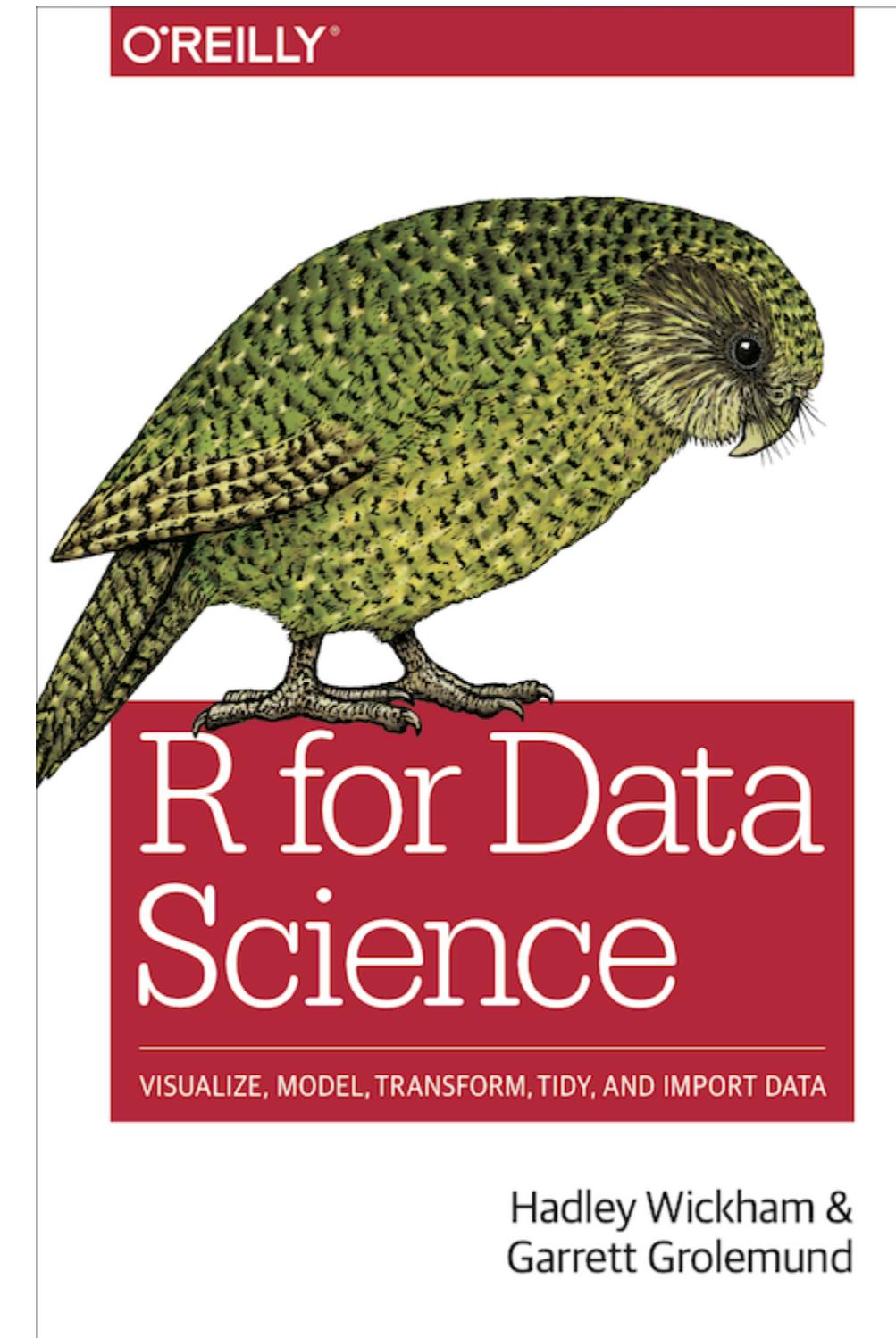
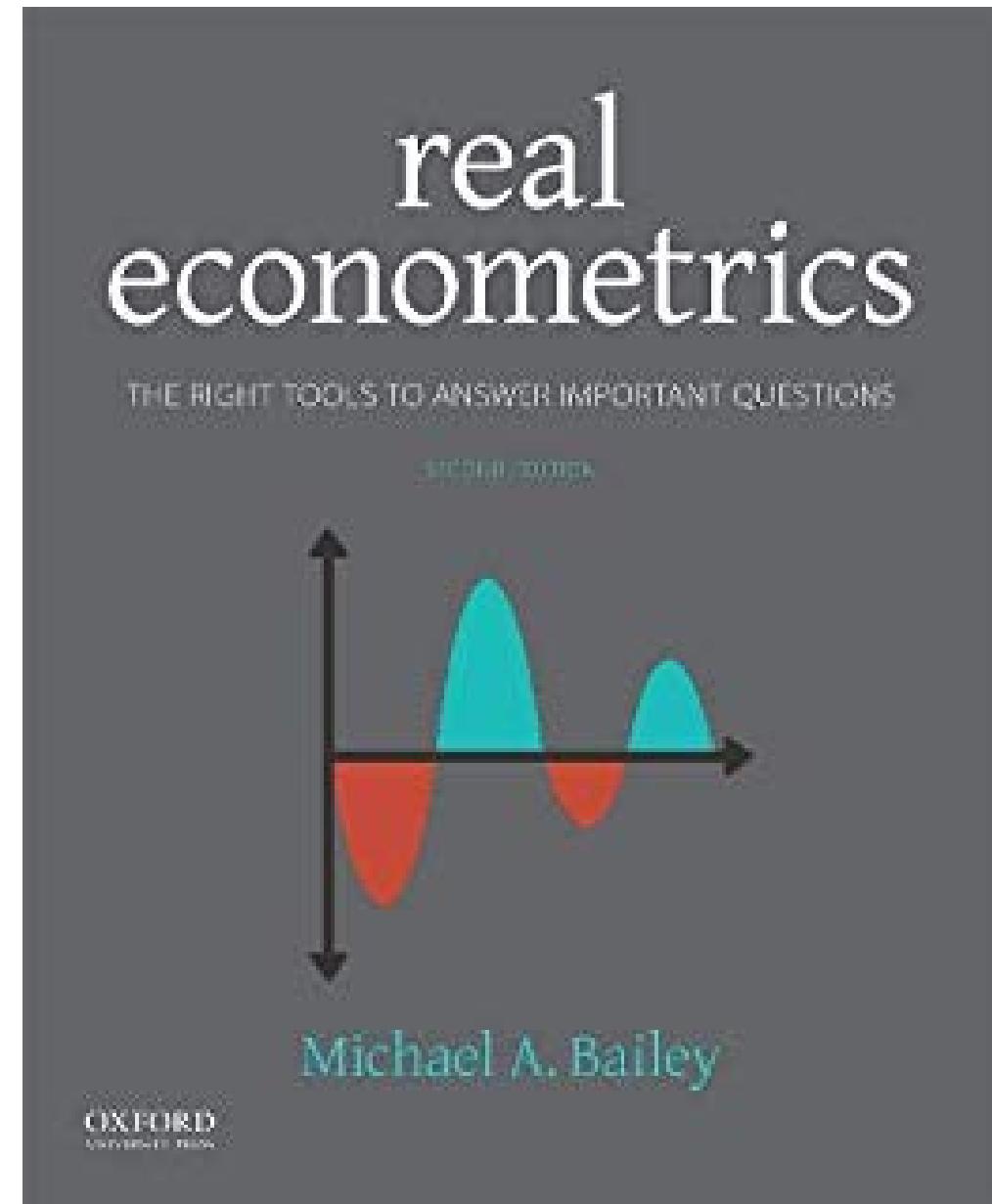


# Logistics

- Office hours: M/W 10:00-11:00 AM & by appt
  - Office: 110 Rosenstock
-  Slack channel [#c-306-metrics](#)
- Recorded videos in Blackboard Panopto
- Attendance — **Kahoot!**s
- Teaching Assistant(s): TBD
  - grade HWs & hold office hours
- See the [resources page](#) for tips for success and more helpful resources



# Your Textbooks



# You Can Do This.



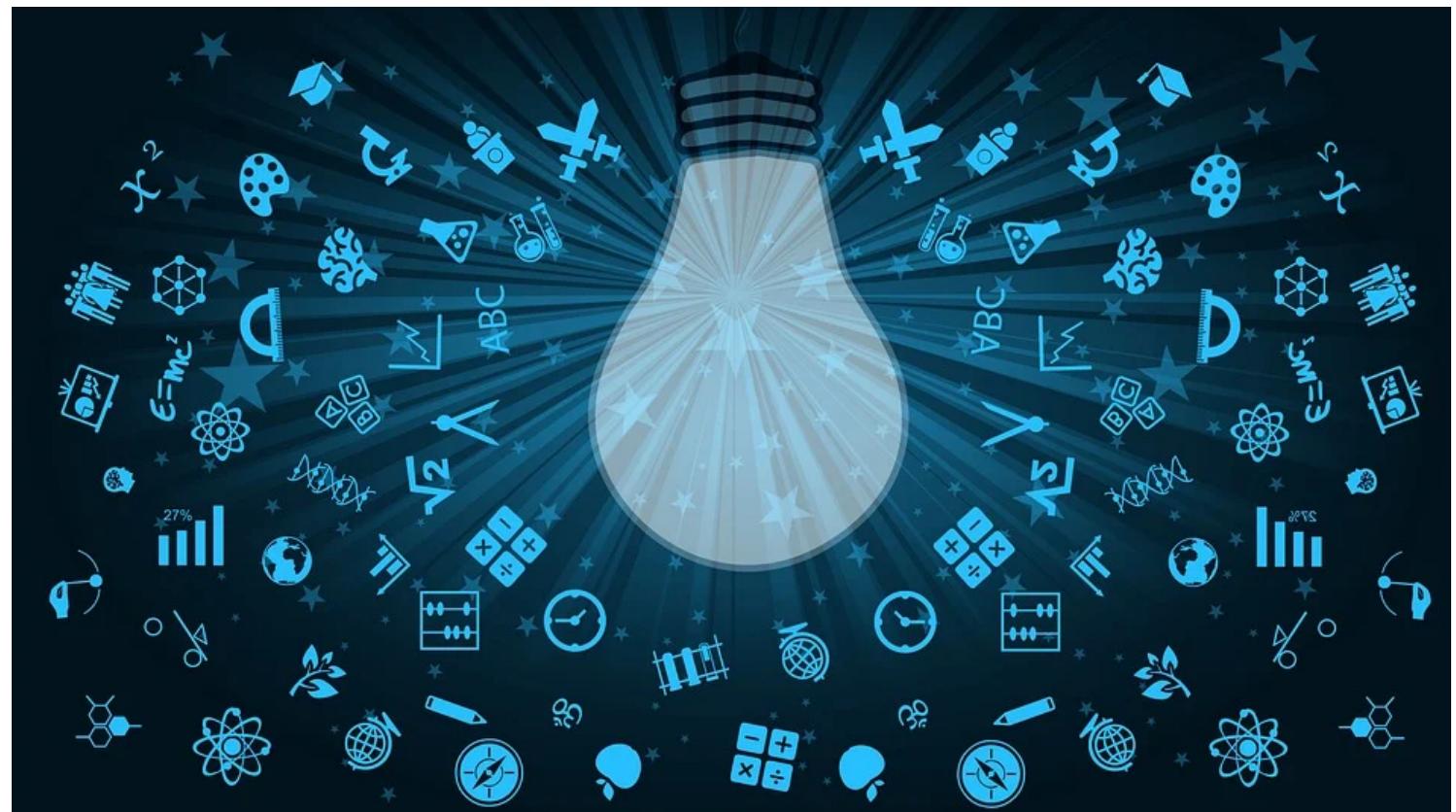
# And I am Here To Help

- Don't forget to prioritize your mental health
- Come talk to me! It's not scary!



# Tips for Success in This Course

- *Take notes. On paper. Really.*
- **Work together** on assignments and study together.
- Ask questions, come to office hours. Don't struggle in silence, you are not alone!
- **The biggest skill you are developing is learning how to learn<sup>1</sup>**
- See the [reference page](#) for more



<sup>1</sup> A properly worded Google search will become your secret weapon. Believe me. It's still mine.



# Course Website



# Econometrics

**ECON 480 • Fall 2021 • Hood College**

**Learn how to analyze data using the canonical models of econometrics with R.**

By the end of this course, you will:

1. understand how to evaluate statistical and empirical claims;
2. use the fundamental models of causal inference and research design;
3. gather, analyze, and communicate with real data in R.

[Schedule](#)[Syllabus](#)

[metricsF21.classes.ryansafner.com](https://metricsF21.classes.ryansafner.com)



# Roadmap for the Semester

