Class 1: Introduction

MGSC 310

Prof. Jonathan Hersh

Welcome to MGSC 310!

Be glad you're not in this intro to data science class @ UC Berkeley!

Class 1: Outline

1. **Syllabus (On Canvas)**
2. About Me & TAs
3. About You!
4. What is Machine Learning?
5. Installing R, Rstudio, Miktek and RTools
6. Predictive vs Causal Inference

Syllabus on Canvas

Statistical Models in Business FALL2020S MGSC-310-01 › Syllabus

Fall 2020

- Home
- Syllabus
- Modules
- Assignments
- Discussions
- Grades
- Panopto Video
- Zoom
- Course Evaluations
- Conferences

MGSC-310-01

Jump to Today

MGSC310: Statistical Models for Business Analytics (Introduction to Machine Learning)

Argyros School of Business and Economics
Chapman University

	Course details
Instructor:	Jonathan Hersh, Ph.D Assistant Professor, Economics and Management Science Argyros School of Business and Economics
Teaching Assistants:	Joshua Anderson (ander428@mail.chapman.edu) Cady Stringer (cstringer@chapman.edu) Sam Webster (swebster@chapman.edu)

Class 1: Outline

1. Syllabus (On Canvas)
2. **About Me & TAs**
3. About You!
4. What is Machine Learning?
5. Installing R, Rstudio, Miktek and RTools
6. Predictive vs Causal Inference

Teaching Assistants

Joshua Anderson
(MSc Data Science Student)
ander428@mail.chapman.edu

Cady Stringer
(4th Year Student)
cstringer@chapman.edu

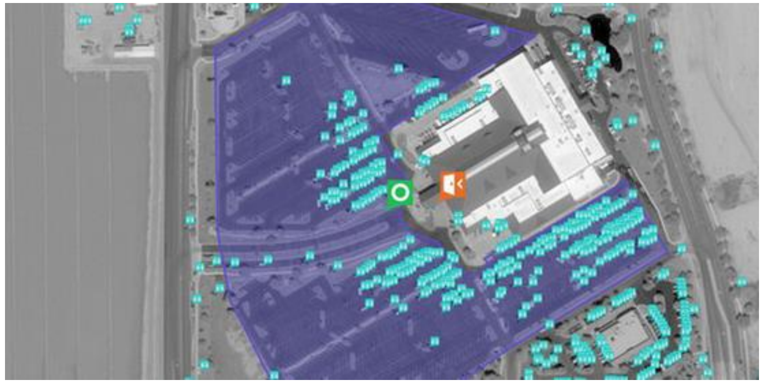
Sam Webster
(4th year CS Major)
swebster@chapman.edu

Who am I?

- Sometimes an economist (PhD in econ) who uses machine learning
- Worked as Data Scientist for the World Bank, economic consultant, coder
- Research in Information Systems and Development Economics

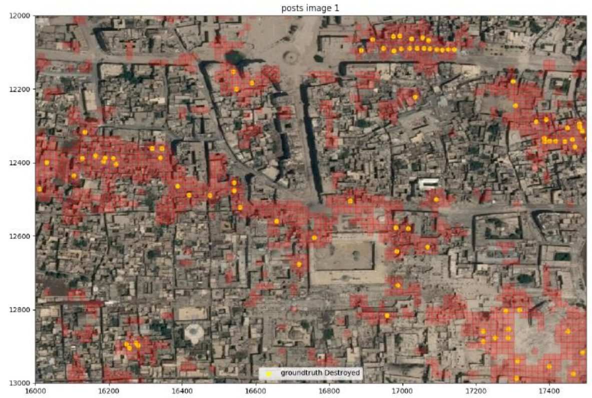
My Research

- Satellite Imagery + Computer Vision + Machine Learning
- Advised World Bank/IDB on COVID poverty transfers (using methods in this course!) in Belize, Togo, Guinea

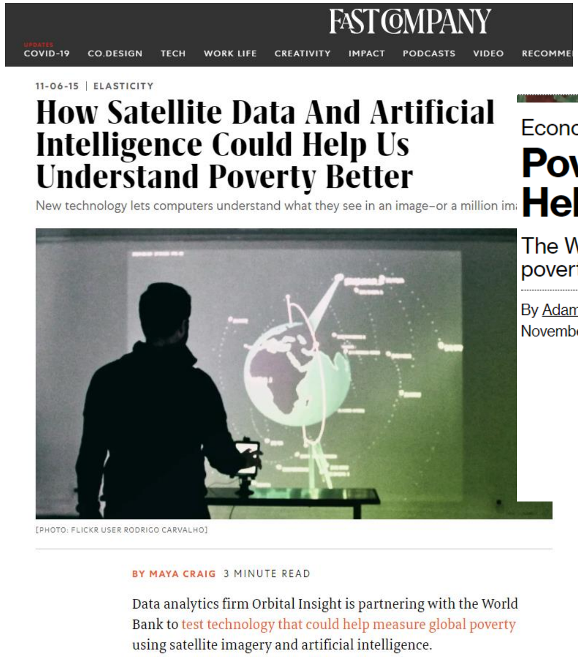


Count cars in parking lots!

Dense Prediction: Scanning Aleppo



Damaged buildings in Syria!



FAST COMPANY

COVID-19 CO.DESIGN TECH WORK LIFE CREATIVITY IMPACT PODCASTS VIDEO RECOMME

11-06-15 | ELASTICITY

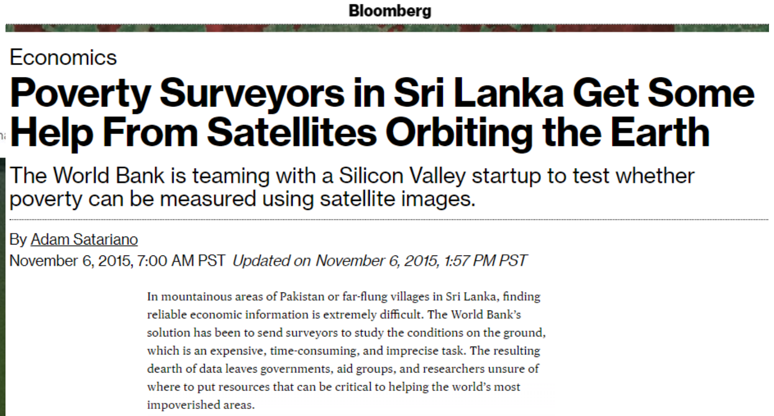
How Satellite Data And Artificial Intelligence Could Help Us Understand Poverty Better

New technology lets computers understand what they see in an image–or a million ima

[PHOTO: FLICKR USER RODRIGO CARVALHO]

BY MAYA CRAIG 3 MINUTE READ

Data analytics firm Orbital Insight is partnering with the World Bank to test technology that could help measure global poverty using satellite imagery and artificial intelligence.



Bloomberg

Economics

Poverty Surveyors in Sri Lanka Get Some Help From Satellites Orbiting the Earth

The World Bank is teaming with a Silicon Valley startup to test whether poverty can be measured using satellite images.

By Adam Satariano

November 6, 2015, 7:00 AM PST *Updated on November 6, 2015, 1:57 PM PST*

In mountainous areas of Pakistan or far-flung villages in Sri Lanka, finding reliable economic information is extremely difficult. The World Bank's solution has been to send surveyors to study the conditions on the ground, which is an expensive, time-consuming, and imprecise task. The resulting dearth of data leaves governments, aid groups, and researchers unsure of where to put resources that can be critical to helping the world's most impoverished areas.

More "Business" Research

- Online Media Piracy

Forbes

There's Hope To Combat Piracy If Hollywood, Industry, and Government Unite

Nelson Granados Contributor
Hollywood & Entertainment
I cover digital trends in travel, media and entertainment.

This article is more than 5 years old.

Several studies have shown that piracy hurts the revenues of content owners, and instead pirate sites are reaping hundreds of millions of dollars in online advertising. Yet theft of movies and TV content seems to be as rampant today as ever. The Motion Picture Association of America (MPAA) reports that in 2014, just in the U.S. alone, 710 million movies and TV shows were shared via BitTorrent sites. Extrapolating to a global scale (the U.S. is less than 5% of the world's population) and adding streaming and other piracy methods, losses were likely in the billions of dollars. The staggering order of magnitude may lead some to wonder if it's even worth fighting the battle, or if it has been lost already. Can the battle against piracy be won? If so, how?

- IT Strategy

The Paradox of Openness: Exposure vs. Efficiency of APIs

Seth G. Benzell[*]
Guillermo Lagarda[†]
Jonathan Hersh[‡]
Marshall Van Alstyne[§]

August 3, 2019

ABSTRACT

APIs are the building blocks of digital platforms, yet there is little quantitative evidence on their use. Do API adopting firms do better? Do such firms change their operating procedures? Using proprietary data from a major API tools provider, we explore the impact of API use on firm value and operations. We find evidence that API use increases market capitalization and lowers R&D expenditures. We then document an important downside. API adoption increases the risk of data breaches, a risk that rises when APIs are more open or place less emphasis on security. Firms reduce API data flows in the month before a hack announcement, consistent with a conscious attempt to limit breach scope. In the same period, however, the variance of API data flows increases, consistent with heterogeneity in firms' ability to detect and shut down unauthorized data access. Our findings highlight a fundamental paradox of openness: It increases upside value and downside risk at the same time. We document that firms respond to these trade-offs in logical ways and conclude that the benefits of opening APIs exceed the risks for firms situated to adopt a platform strategy.

Keywords: Platforms, APIs, Information Security, Technology Strategy, Market Capitalization

Most Proud of: Cited on the Wikipedia Page for "Waffle"

Not logged in Talk Contributions Create account Log in

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Article Talk Read View source View history Search Wikipedia

Waffle

From Wikipedia, the free encyclopedia

This article is about the batter/dough-based food. For other uses, see Waffle (disambiguation).

A **waffle** is a dish made from leavened batter or dough that is cooked between two plates that are patterned to give a characteristic size, shape, and surface impression. There are many variations based on the type of waffle iron and recipe used. Waffles are eaten throughout the world, particularly in Belgium, which has over a dozen regional varieties.[1] Waffles may be made fresh or simply heated after having been commercially cooked and frozen.

Waffle

Place of origin	France, Belgium
Main ingredients	Batter or dough
Variations	Liège waffle, Brussels Waffle, Flemish Waffle, Bergische waffle, Stroopwafel and others

Cookbook: Waffle
Media: Waffle

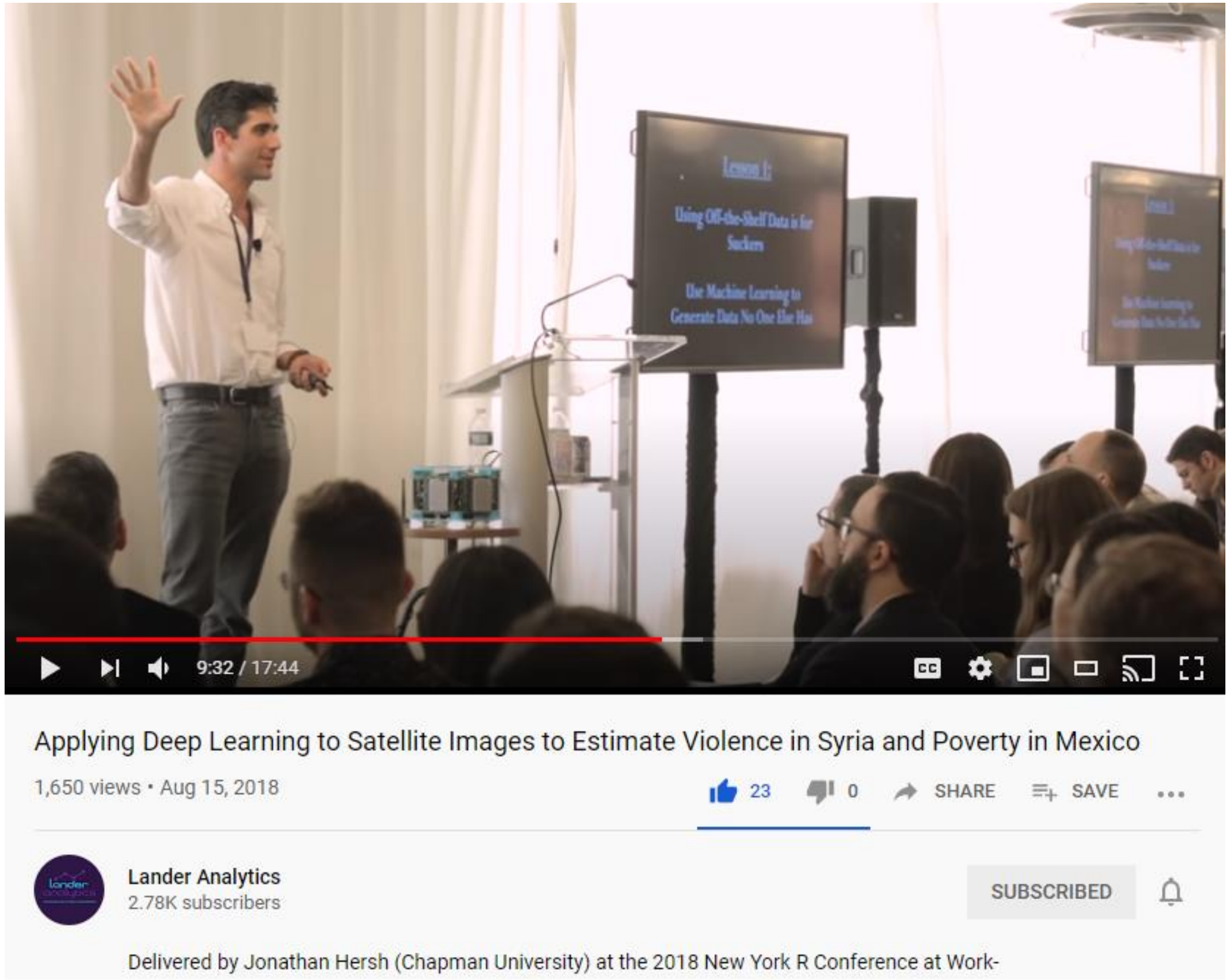
Contents [hide]

52. ^ a b "Sweet Diversity: Overseas Trade and Gains from Variety after 1492" Archived 2013-07-26 at the Wayback Machine, Jonathan Hersh, Hans-Joachim Voth, Real Sugar Prices and Sugar Consumption Per Capita in England, 1600–1850, p.42

References

1. ^ "Les Gaufres Belges" Archived 2012-08-20 at the Wayback Machine. Gaufresbelges.com. Retrieved on 2013-04-07.
2. ^ Robert Smith (1725). *Court Cookery*. p. 176.
3. ^ "Waffle" Archived 2013-04-07 at the Wayback Machine, The Merriam-Webster Unabridged Dictionary

Given Talks for the R Community



Class 1: Outline

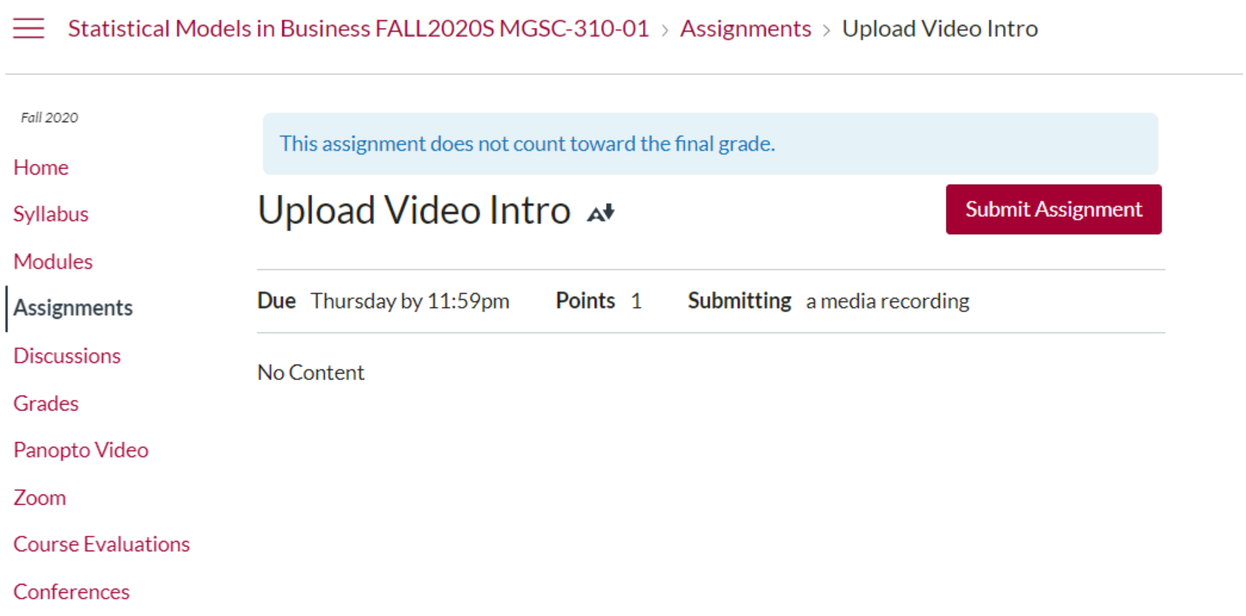
1. Syllabus (On Canvas)
2. About Me & TAs
3. **About You!**
4. What is Machine Learning?
5. Installing R, Rstudio, Miktek and RTools
6. Predictive vs Causal Inference

About You

- **Participation: Upload a Short Video of Yourself Telling Us:**

1. Name and major
2. Work experience
3. Hometown
4. Fun fact about yourself!

- (Let us know if we can play it for the class!)

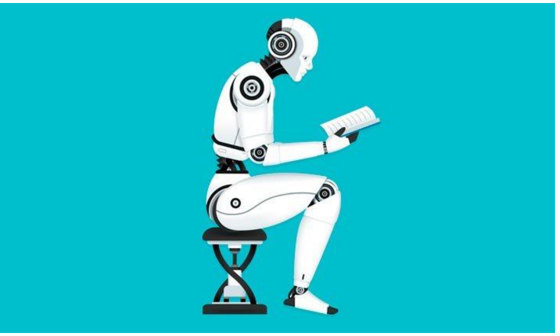


Class 1: Outline

1. Syllabus (On Canvas)
2. About Me & TAs
3. About You!
4. **What is Machine Learning?**
5. Installing R, Rstudio, Miktek and RTools
6. Predictive vs Causal Inference

Free Association with the Phrase Machine Learning...

Public Conception of Machine Learning

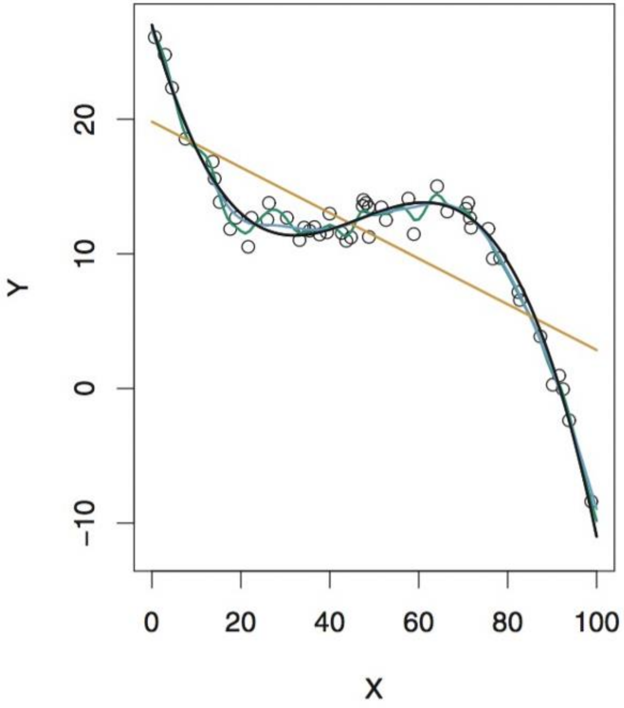


Reality (90% of the time)

Target or Output

Input data

$$\hat{y} = \hat{f}(x)$$

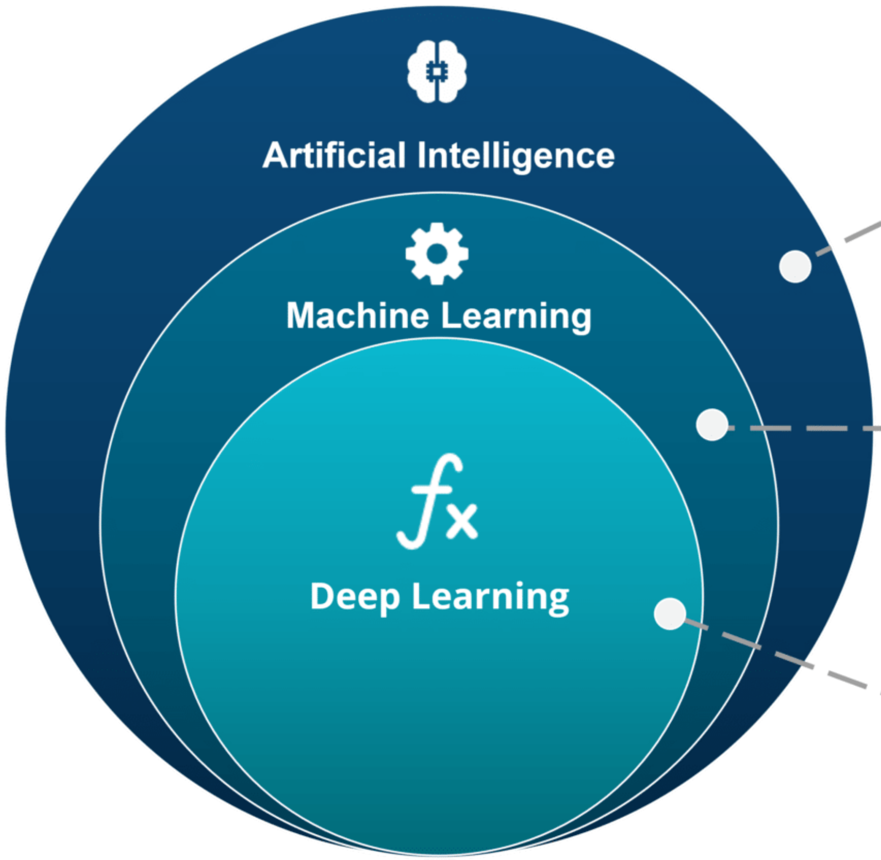


Machine Learning Versus Econometrics

- **Machine Learning**
  - Developed to solve problems in computer science
  - Prediction/classification
  - Desire: goodness of fit
  - Huge Datasets! (Terabytes) Thousands of variables!
  - Whatever works

- **Econometrics**
  - Developed to solve problems in economics
  - Explicitly testing a theory
  - "Statistical significance" more important than model fit
  - Small datasets Few dozen variables
  - "It works in practice, but what about theory?"

What is Machine Learning? What is Artificial Intelligence?



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

Bill Gates Says This Type of AI Will Be Worth “10 Microsofts”

The Motley Fool **Rex Moore, The Motley Fool**

Motley Fool August 24, 2019

Microsoft (NASDAQ: MSFT) founder Bill Gates was speaking to a group of college students in 2004.

According to *The New York Times*, Gates was a bit concerned about the decline in the number of computer science majors, as well as the notion that the field had matured and there weren't many breakthroughs left to achieve in the area.

One student expressed doubt that there would ever be another tech company as successful as Microsoft. Gates' reply is eye-opening:

''If you invent a breakthrough in artificial intelligence, so machines can learn, **that is worth 10 Microsofts**.''

He wasn't kidding...

DATA

Data Scientist: The Sexiest Job of the 21st Century

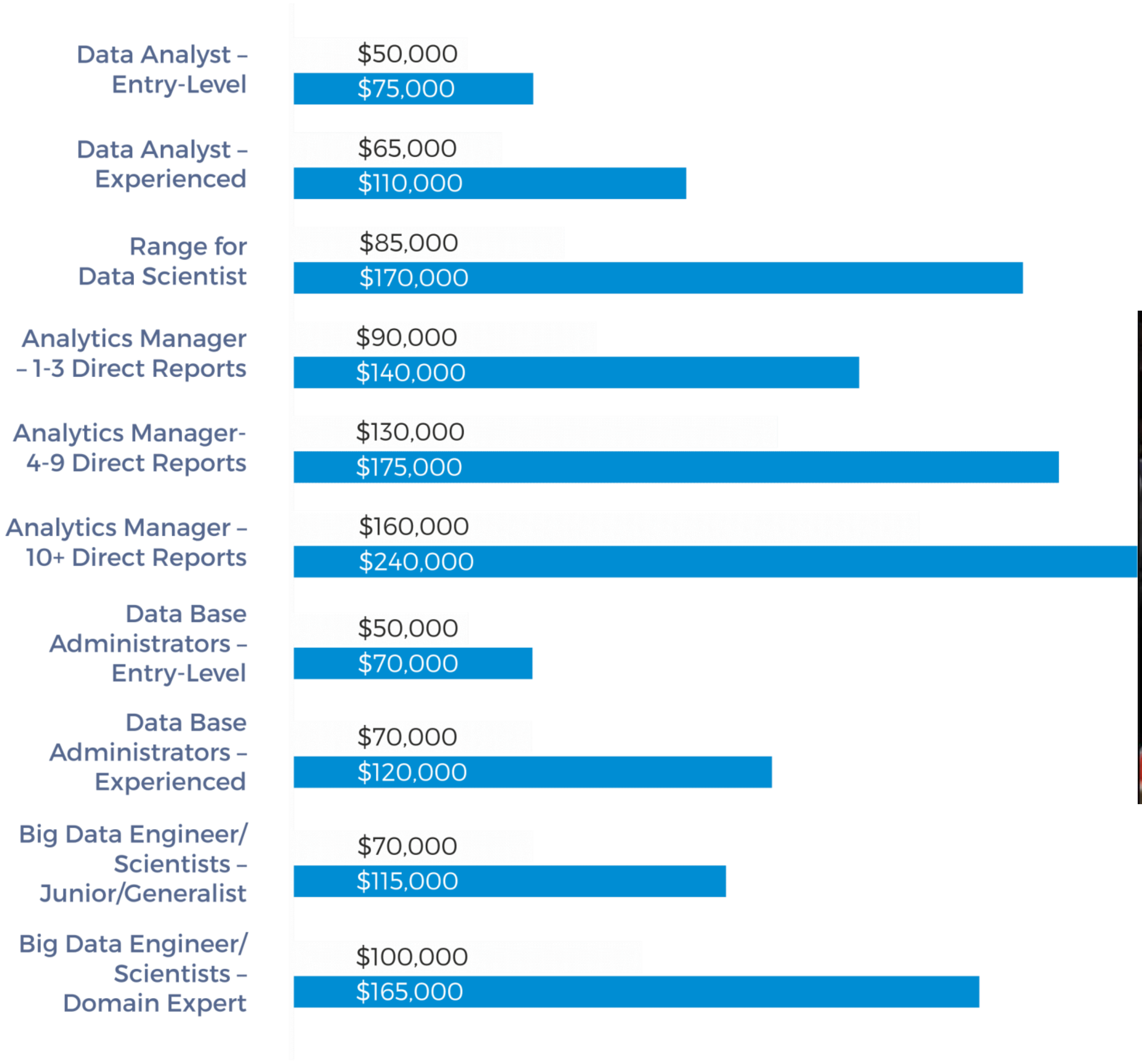
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Summary Save Share 16 Comment Text Size Print PDF $8.95 Buy Copies

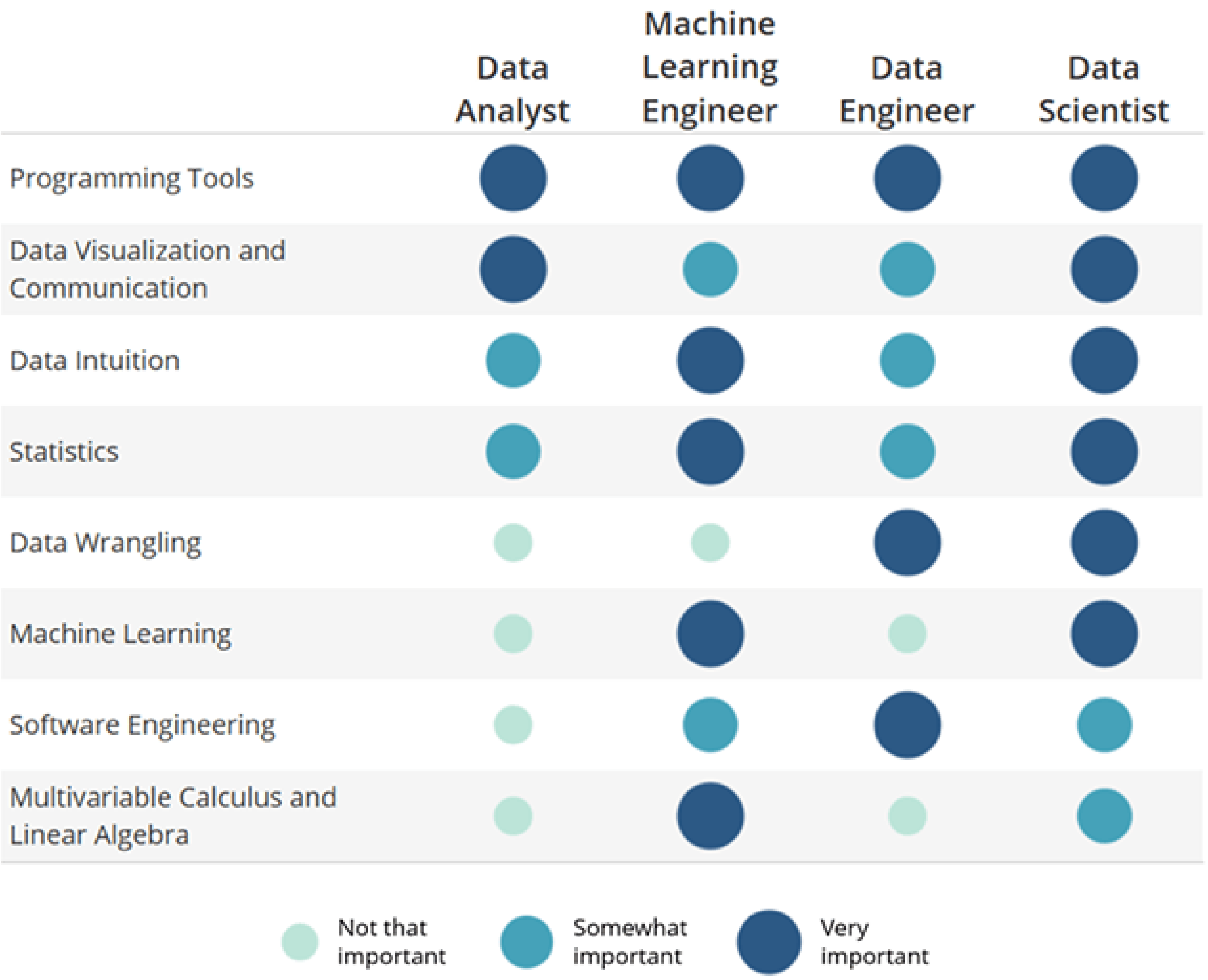
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early." Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see possibilities. He began forming theories, testing hunches, and finding patterns that allowed him to predict whose networks a given profile would land in. He could imagine that new features capitalizing on the heuristics he was developing might


Big Data, Big Paycheck
Median salary for analytics professionals and those specifically within data science, by level of experience.
Up to 3 years
Analytics professionals
$65,000
Data scientists
$80,000
4 to 8 years
$85,000
$120,000
9+ years
$115,000
$150,000
Note: Data do not include managers
Source: Burtch Works
The Wall Street Journal


Data Analyst – Entry-Level
$50,000
$75,000
Data Analyst – Experienced
$65,000
$110,000
Range for Data Scientist
$85,000
$170,000
Analytics Manager – 1-3 Direct Reports
$90,000
$140,000
Analytics Manager- 4-9 Direct Reports
$130,000
$175,000
Analytics Manager – 10+ Direct Reports
$160,000
$240,000
Data Base Administrators – Entry-Level
$50,000
$70,000
Data Base Administrators – Experienced
$70,000
$120,000
Big Data Engineer/ Scientists – Junior/Generalist
$70,000
$115,000
Big Data Engineer/ Scientists – Domain Expert
$100,000
$165,000

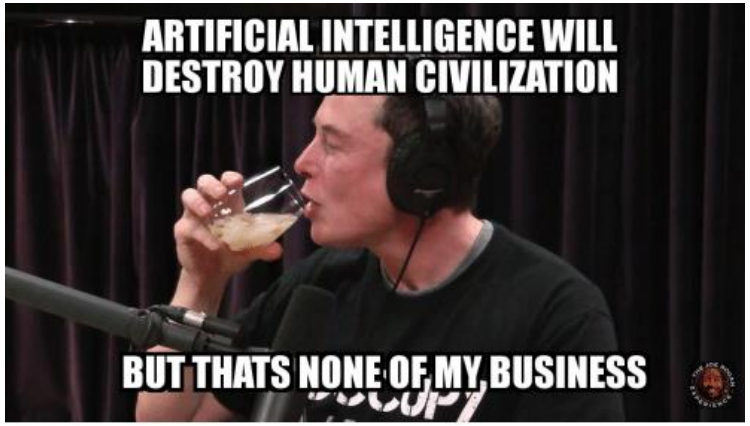

MACHINE LEARNING
SO HOT RIGHT NOW

Varied Skills in the Data Science Landscape

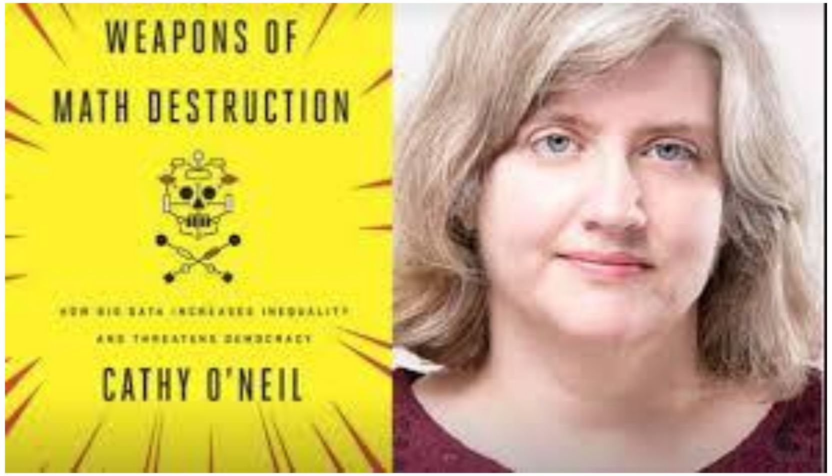


You can Help Answer: Policy Problems of AI/Big Data

We need a blend of humanistic and scientific understanding

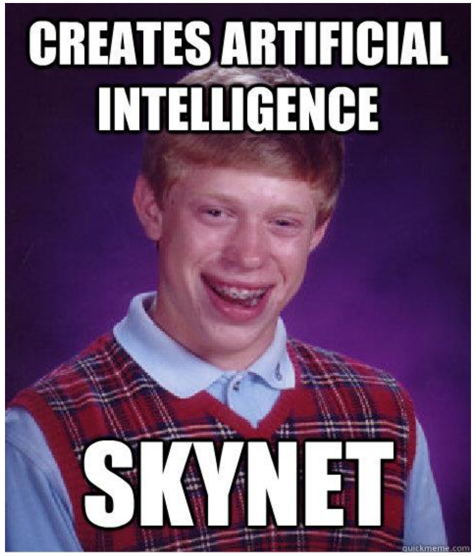


Unintended Consequences of AI



Algorithmic Bias

Will AI grow too powerful?



Does AI Create an Unfair Advantage for Incumbents / Big firms?

The New York Times

Good for Google, Bad for America

At its core, artificial intelligence is a military technology. Why is the company sharing it with a rival?

By Peter Thiel
Mr. Thiel is an entrepreneur and investor.

Aug. 1, 2019

A "Manhattan Project" for artificial intelligence is how Demis Hassabis, the founder of DeepMind, described his company in 2010, when I was one of its first investors. I took it as figurative grandiosity. I should have taken it as a literal warning sign, because that is how it was taken in foreign capitals that were paying close attention.

Now almost a decade later, DeepMind is the crown jewel of Google's A.I. effort. It has been the object of intense fascination in East Asia especially since March 2016 when its AlphaGo software project beat Lee Sedol, a champion of the ancient strategic board game of Go.

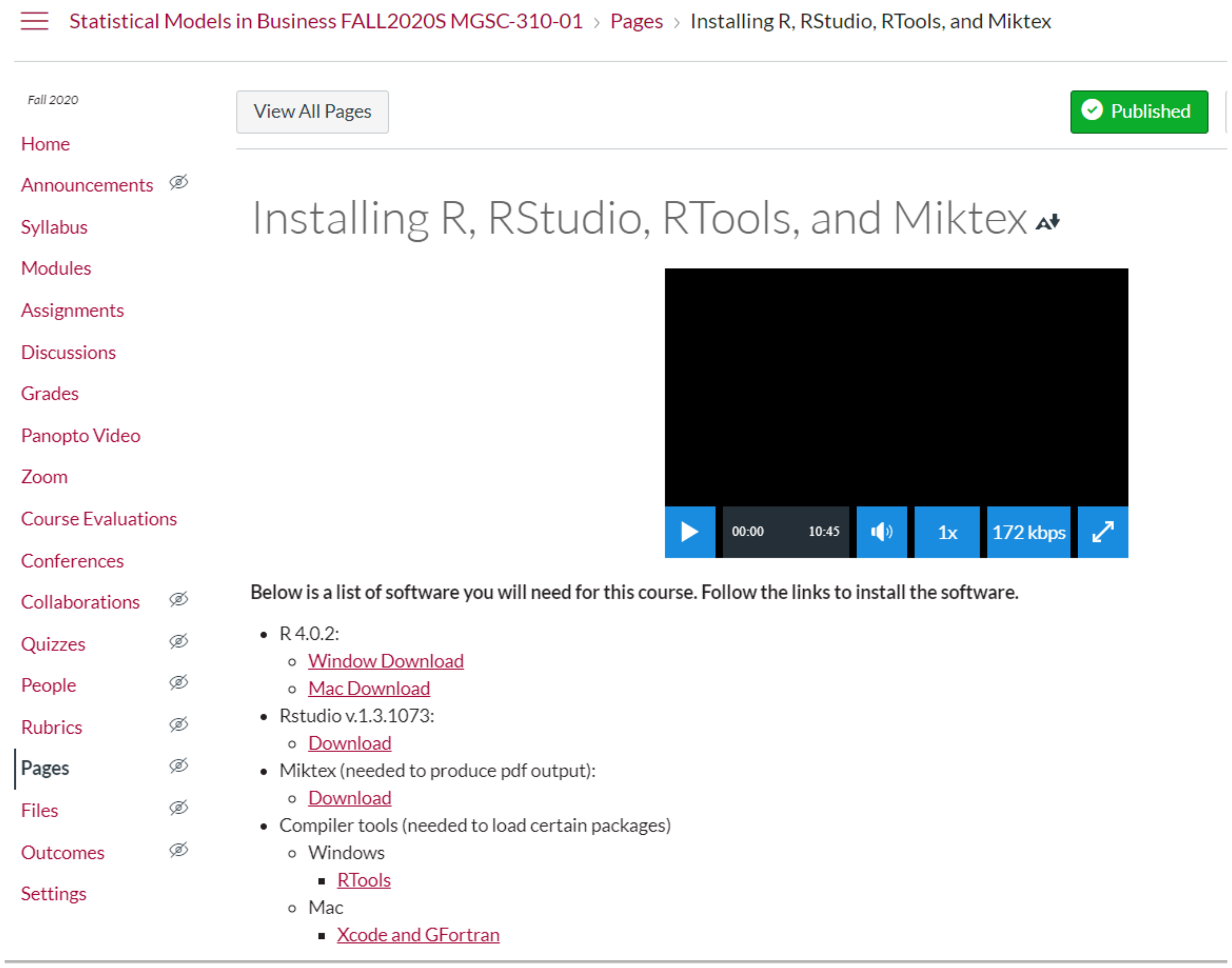
Class 1: Outline

1. Syllabus (On Canvas)
2. About Me & TAs
3. About You!
4. What is Machine Learning?
5. **Installing R, Rstudio, Miktek and RTools**
6. Predictive vs Causal Inference

Class 1: Outline

1. Syllabus (On Canvas)
2. About Me & TAs
3. About You!
4. What is Machine Learning?
5. **Installing R, Rstudio, Miktek and RTools**
6. Predictive vs Causal Inference

Please Follow Instructions to Install Computer Tools



Statistical Models in Business FALL2020S MGSC-310-01 › Pages › Installing R, RStudio, RTools, and Miktex

Fall 2020

- Home
- Announcements
- Syllabus
- Modules
- Assignments
- Discussions
- Grades
- Panopto Video
- Zoom
- Course Evaluations
- Conferences
- Collaborations
- Quizzes
- People
- Rubrics
- Pages
- Files
- Outcomes
- Settings

View All Pages

Published

Installing R, RStudio, RTools, and Miktex

00:00 10:45 1x 172 kbps

Below is a list of software you will need for this course. Follow the links to install the software.

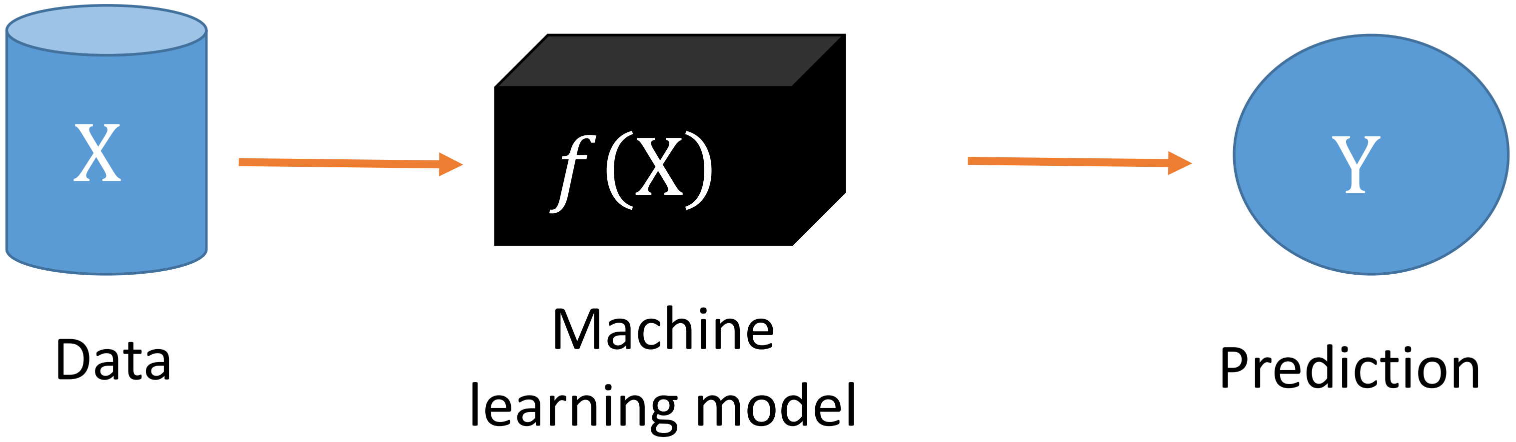
- R 4.0.2:
  - Window Download
  - Mac Download
- Rstudio v.1.3.1073:
  - Download
- Miktex (needed to produce pdf output):
  - Download
- Compiler tools (needed to load certain packages)
  - Windows
    - RTools
  - Mac
    - Xcode and GFortran

Class 1: Outline

1. Syllabus (On Canvas)
2. About Me & TAs
3. About You!
4. What is Machine Learning?
5. Installing R, Rstudio, Miktek and RTools
6. **Predictive vs Causal Inference**

Predictive Analytics

- This course will primary cover **predictive analytics**



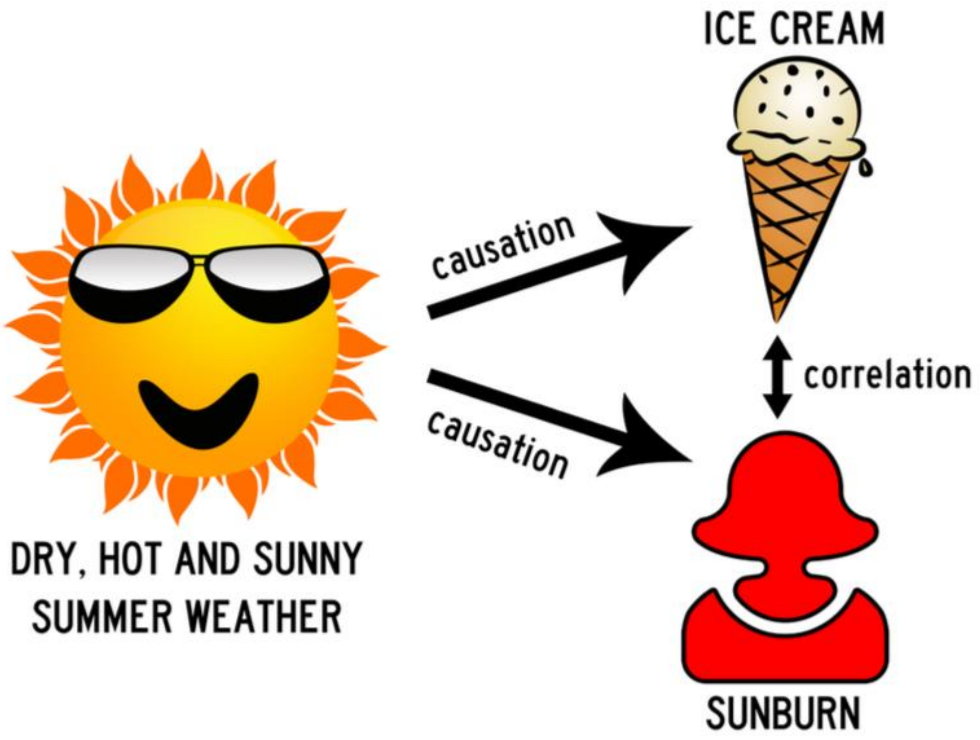
- This type of analysis **assumes the world stays the same**
- It cannot tell us **what would have happened** if the world was different

Distinction Between Causal and Predictive Analysis is Subtle!

Causal model asks: If I were to make X happen, what would happen to Y?

Predictive model ask: If I observe X, what do I know about Y?

Predictive vs Causal Analysis

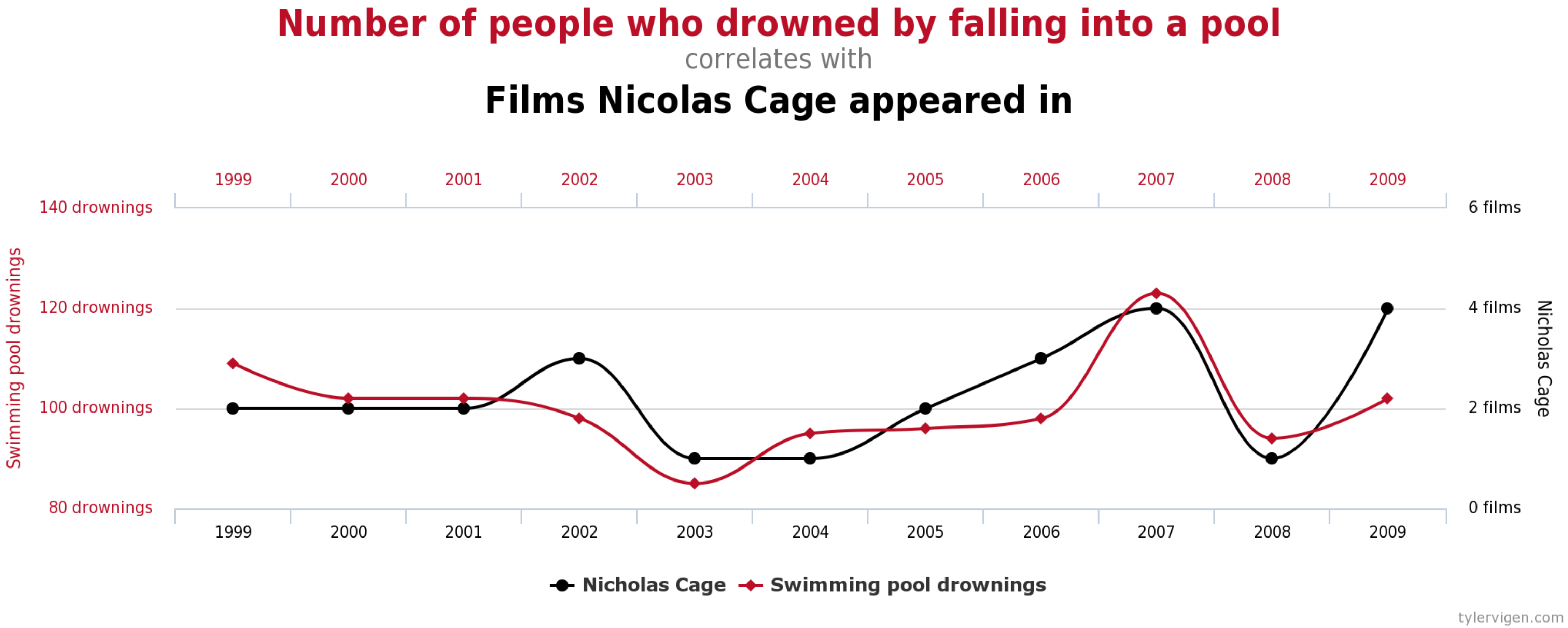


- Ice cream sales are predictive of sunburns but do not cause sunburns
- We all know that correlation =! Causation
  - Correlation = two series move together!
- But sometimes knowing two things are correlated – even if the causal link is unknown -- is useful!

Example of Useful Correlation: Tesla Stock Price Correlated With Bitcoin Price

1. Is there a causal link?
2. How can you benefit even if there isn't a causal link?

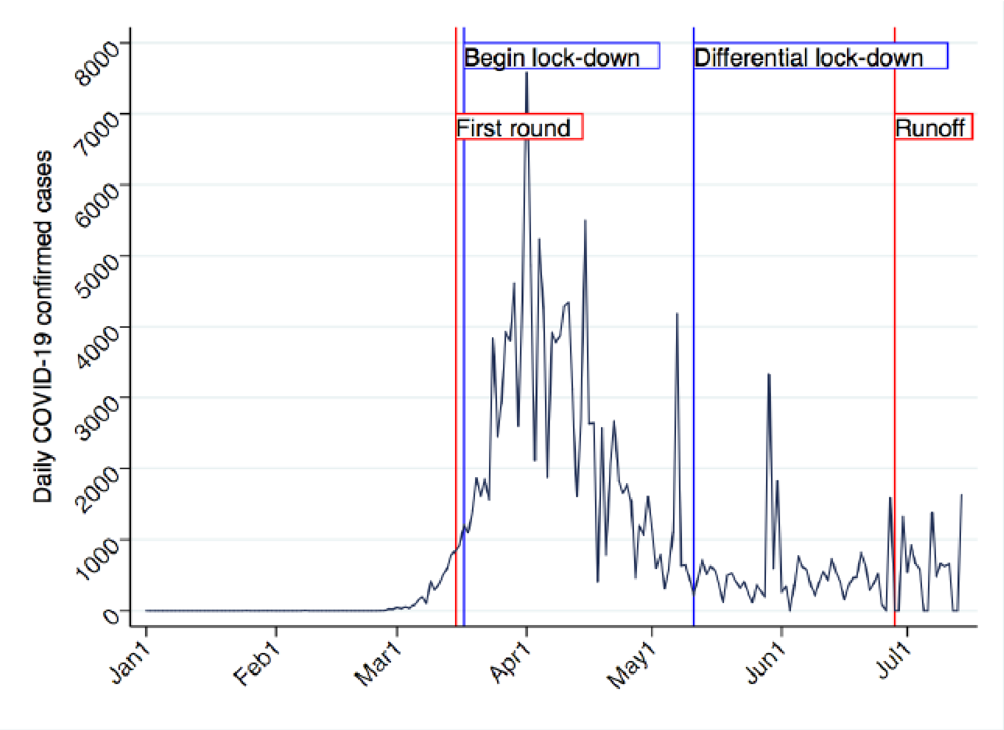
Silly Correlations Are Fun But Deadly



Predictive or Causal: "Do Lockdowns or Quarantines Impact the Spread of COVID?"

Lockdown Correlated With Subsequent COVID Growth But Is It Causal?

Figure 2: Evolution of Covid-19 confirmed cases in France



Notes: The plot shows the total number of confirmed COVID-19 cases in France starting from January 1st 2020. The red lines indicate the dates of 2020 local elections (first round -March 15th-, runoff -June 28th-), the blue lines indicate the dates of the modification in the lockdown policy (introduction of the lockdown -March 17th-, first relaxation of the lockdown -May 11th-). The source is the French Government data portal (`https://www.data.gouv.fr/fr/`).

Did The CA Lockdown Lower COVID Cases? Synthetic Control for Causal Estimation

ABSTRACT

Did California's Shelter-In-Place Order Work? Early Coronavirus-Related Public Health Effects*

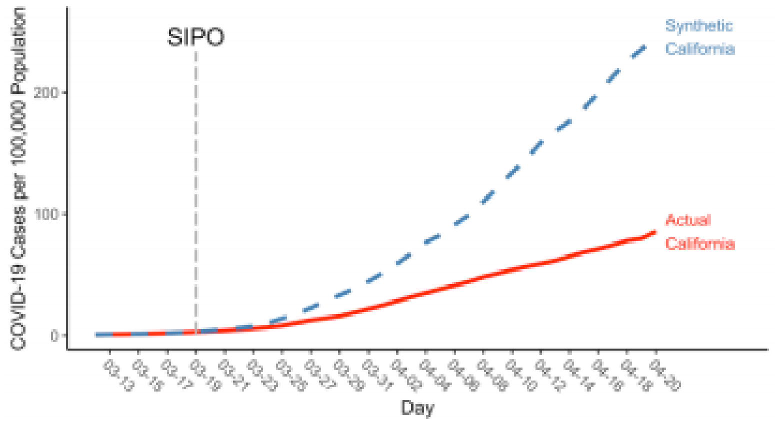
On March 19, 2020, California Governor Gavin Newsom issued Executive Order N-33-20 2020, which required all residents of the state of California to shelter in place for all but essential activities such as grocery shopping, retrieving prescriptions from a pharmacy, or caring for relatives. This shelter-in-place order (SIPO), the first such statewide order issued in the United States, was designed to reduce COVID-19 cases and mortality. While the White House Task Force on the Coronavirus has credited the State of California for taking early action to prevent a statewide COVID-19 outbreak, no study has examined the impact of California's SIPO. Using daily state-level coronavirus data and a synthetic control research design, we find that California's statewide SIPO reduced COVID-19 cases by 125.5 to 219.7 per 100,000 population by April 20, one month following the order. We further find that California's SIPO led to as many as 1,661 fewer COVID-19 deaths during the first four weeks following its enactment. Back-of-the-envelope calculations suggest that there were about 400 job losses per life saved during this short-run post-treatment period.

JEL Classification: H75, I18

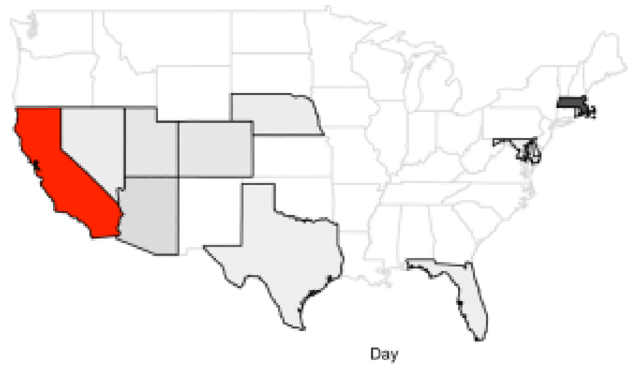
Keywords: coronavirus, COVID-19, shelter in place order, synthetic control

Figure 6: Synthetic Control Estimates for Cases Per 100,000
Iatching Variables: COVID-19 Cases on 3 Pre-Treatment Days & Urbanicity]

(a) Synthetic California v. Actual California Cases Per 100,000



(b) Donor States for Synthetic California Cases



Notes: Estimate is generated using synthetic control methods. The matching was based on three days of pre-SIPO COVID-19 cases per 100,000 and urbanicity measure. The donor states shaded in Figure 6b are those that received a weight of at least .015 in the estimation of the synthetic control counterfactual for California. Darker shaded states received more weight. Synthetic California is comprised of MA (.265), HI (.153), AZ (.051), DC (.036), UT (.033), CO (.031), RI (.028), NE (.022), NV (.021), FL (.019), DE (.017), MD (.017), and TX (.016). In addition, 17 states each contributed a weight between .010 and .015.

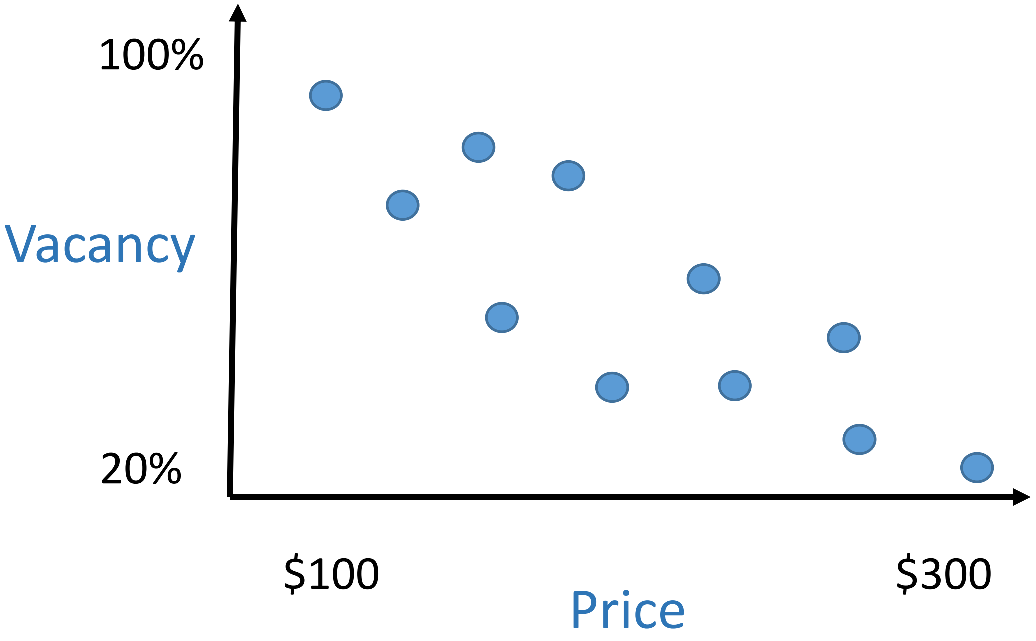
Predictive or Causal? "Do Netflix Users Who Watch 'Stranger Things' Also Watch 'Tiger King'"

Predictive or Causal: "What Advertisements Should We Show To Maximize Purchase/Click-Through?"

Causal Versus Predictive Is Subtle

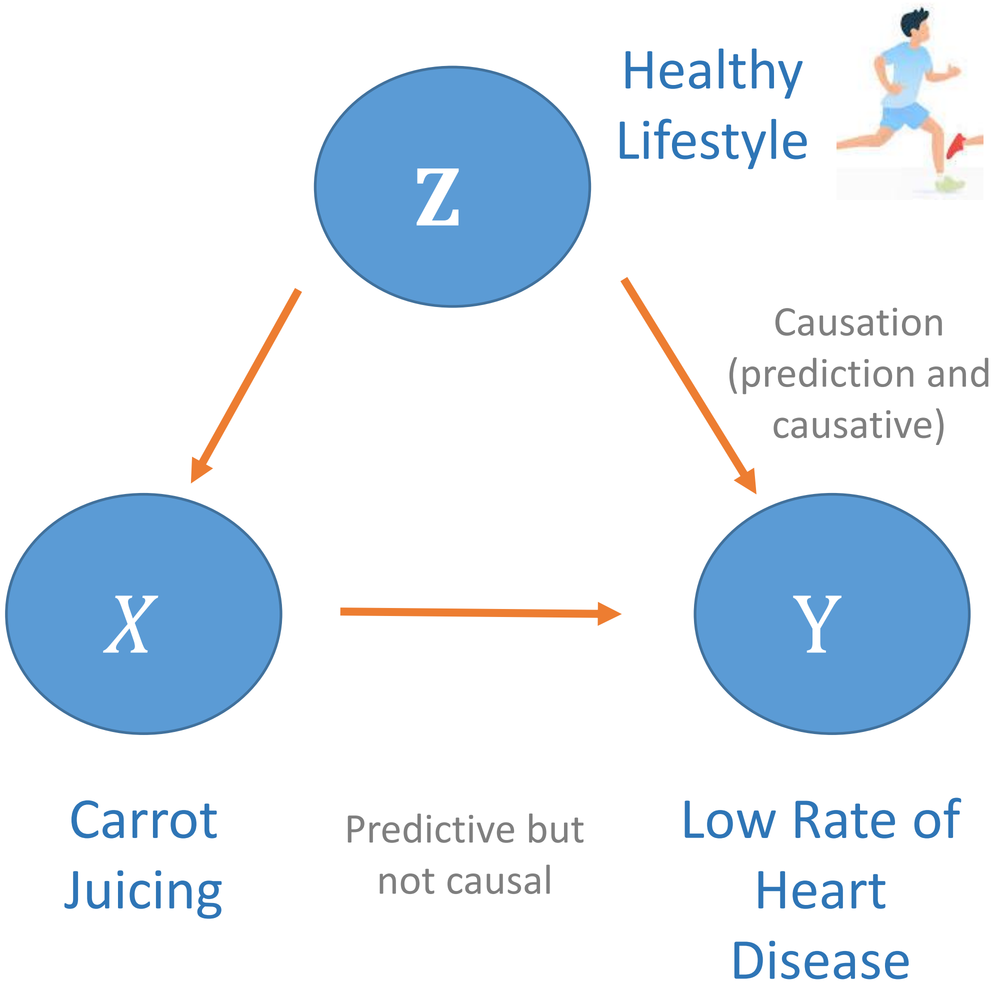
Identify the Error in Logic Here

1. Suppose you are a pricing analyst for a hotel chain
2. You examine one hotel's data and find that when the hotel offers a high room price, their vacancy rate (number of unsold hotel rooms) is low.
3. Do you conclude that the hotel should raise prices to lower vacancies?



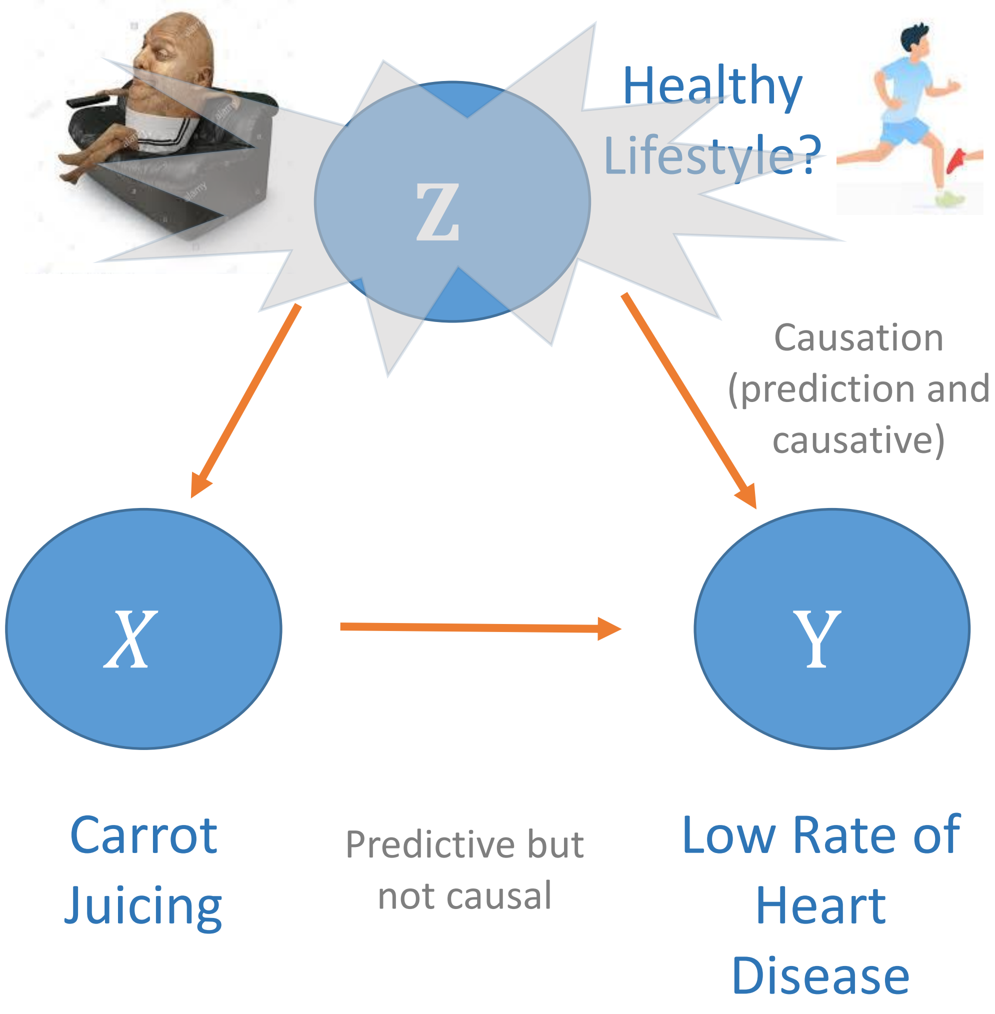
Confounders: Influences Both X and Y

- You are a doctor and want to understand how carrot juicing (X) impacts heart disease (Y).
- You record patient food diaries and find a high prediction between juicing and low heart disease
- Why is that not causal?
- Something else (patients' healthy lifestyle) is a confounder to the causal relationship between X and Y



But Prediction is Very Useful!

- Suppose every patient lies about how active they are. (We cannot observe Z)
- How is this information useful?
- Just knowing whether a patient drinks carrot juice is predictive of whether they have heart disease
- I.e. if we know X, we know to whom we should give anti-cholesterol medication



5-10 Min Breakout Session: Useful Causal and Predictive Analyses

1. Suppose you are a data scientist for the MPAA hired to advise movie theaters

2. Think of **2-4 analyses** – one **causal** and one **predictive** – that would help theater owners as they navigate a post-COVID world

- Examples:
  - After the recession, what snacks do movie-goers want to buy? (predictive)
  - Do masks in theaters reduce COVID transmission? (causal)

Causal Analyses for Theater Owners

Predictive Analyses for Theater Owners

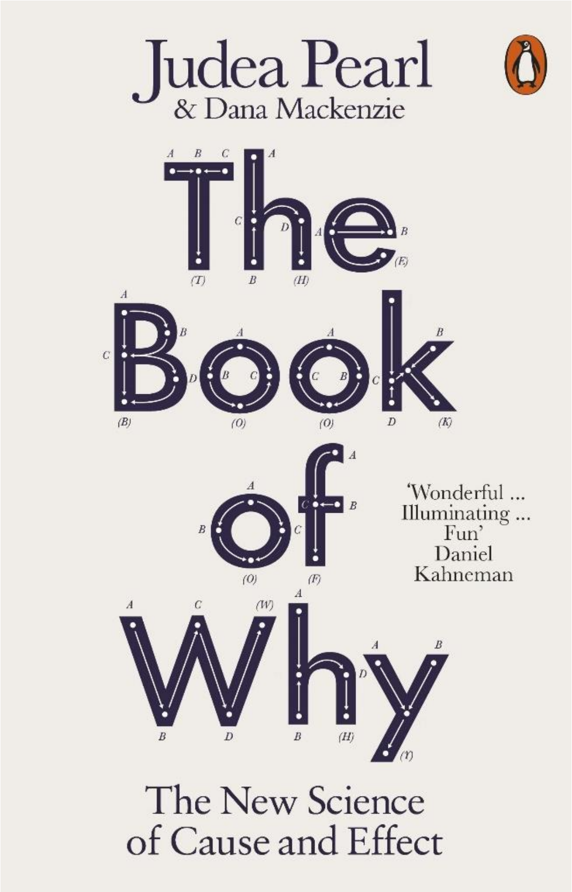
Rest of this Class: Prediction, but for More on Causal Analysis

- ML *can* estimate causal analysis
- Econometrics generally more concerned with causation
- We will focus on predictive methods in this course, but for more on causation take econometrics, biostatistics or read "The Book of Why"

ECON 452 - Econometrics

ECON 452 - Econometrics

Prerequisites, ECON 200, ECON 201, and MGSC 207 or MGSC 220 and MATH 109, or MATH 110, and business administration, or economics major, or computational science, or economics, or mathematics minor. Mathematical and statistical tools to measure economic phenomena. This will involve mathematical formulation of economic theories and statistical inference relating economic theory to empirical analysis. (Offered spring semester.) **3 credits**



Class 1: Summary

- Machine learning is a set of statistical methods used by CS people to learn from data.
- **Predictive analytics**: if I know X, what does this tell me about Y?
- **Causal analysis**: if I changed X, how does that cause Y to change?
- A **confounder** (z) influences both X and Y and results in a spurious correlation (non-causal) between X and Y.
- Both predictive and causal analysis are powerful and useful!
- Many problems can be solved by either (or both!).