Data Science and PA

Day 1

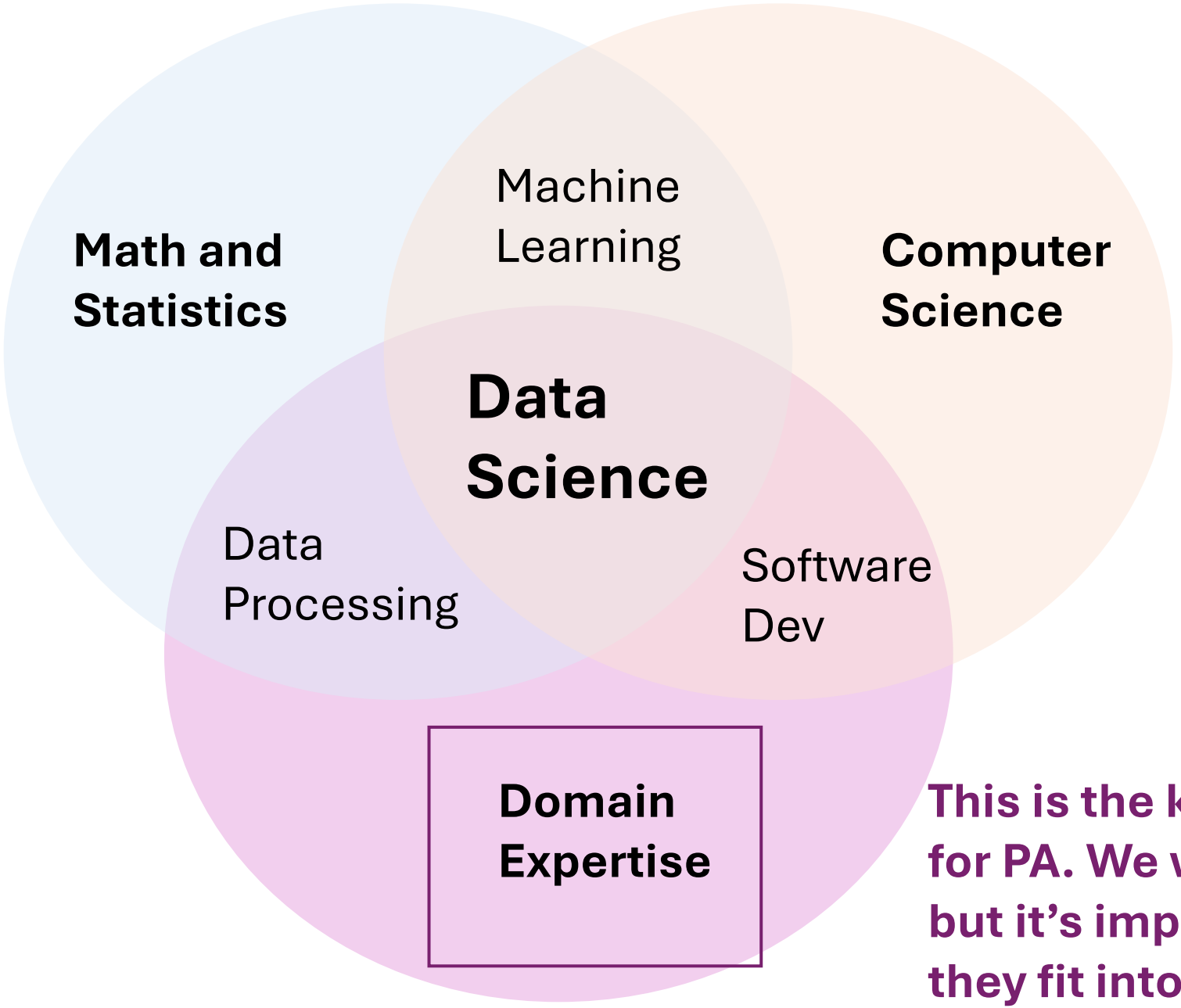
Name

Major

Best activity in US so far

By: Tiana Marrese, PhD

What is data science?

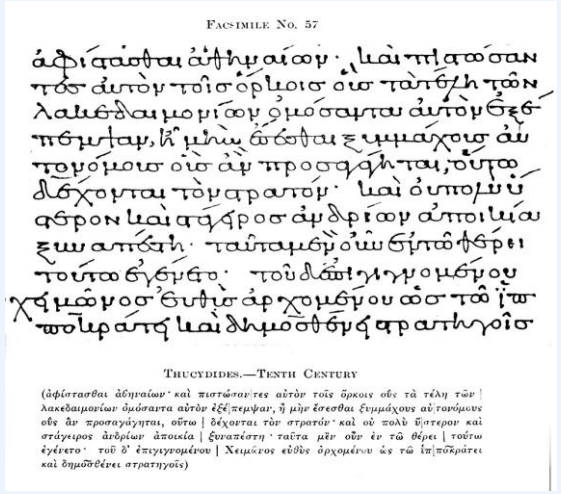


Data Science: extracting insights from data

This is the key feature of data science for PA. We will learn some basic tools but it's important to always ask how they fit into public administration and law enforcement

Source: https://medium.com/analytics-vidhya/introduction-to-data-science-28deb32878e7

Math and Statistics

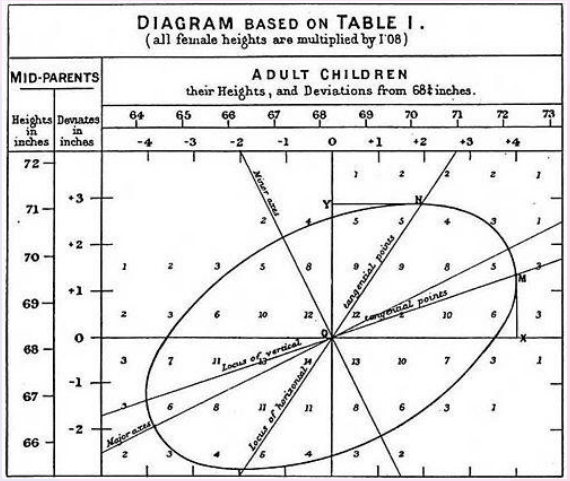


FACSIMILE No. 57

THUCYDIDES.—TENTH CENTURY

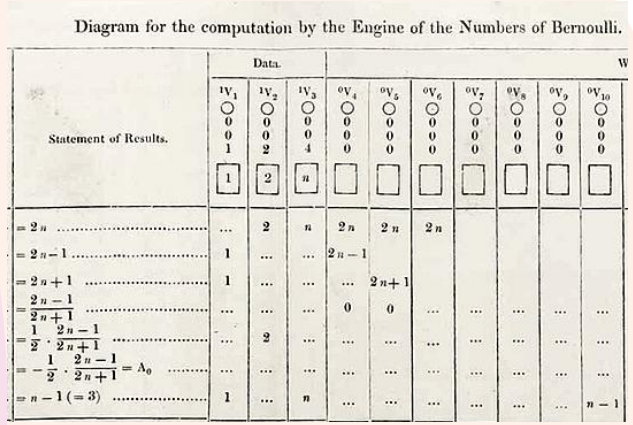
Pelopennesian War, 400BC:
-Multiple soldiers count bricks in wall
-Multiple most common number by height of a brick to get ladder height

Domain Expertise



Sir Francis Galton, 1886:
-Extreme traits are not consistently seen in offspring
-Characteristics regress to some average

Computer Science



Ada Lovelace, 1840:
-First notes on an algorithm to calculate a sequence of numbers

Data Science Today

- **Data collection** like web scraping of tweets
- **Data storage** like cloud storage of all emails from your boss
- **Data visualization** like a pie chart of how you spend your day
- **Data analysis** like calculating average crime rates per year
- **Data predictions*** like finding cities that are most likely to have high crime rates next year

***predictive data science should be handled with caution**

Why is data science important for your job?
(how are you already using it?)

There’s potential for…

- Improving program effectiveness
- Developing new strategies
- Informing funding decisions
- Informing training
- Informing research

Source: https://crimesolutions.ojp.gov/about/how-use-information-you-find

Example: Using Data to Drive Down Traffic Fatalities

- Sergeant James Williams of Nashville Police Department
- **Issue:** High number of vehicle crashes, straining Nashville resources
- **Question:** How can policy reduce crashes and their harms?
- **Existing Evidence:** visibility in crash hot spots reduce crashes
- **Data Collection:** Internal “blind spots” of the existing data, turned to various sources to get a clearer picture of crash problems
- **Analysis:** Found locations, times, severity, and related factors
- **Solution:** Targeted officer deployment around rush hour times, two days a week, one week a month. Officers looked for violation that correlated to factors discovered in data
- **Effect:** Crashes dropped significantly for about three weeks. Cost-benefit analysis was performed to understand best continual implementation
- Source: https://nij.ojp.gov/topics/articles/evidence-based-practices-using-data-drive-down-traffic-fatalities

It all comes down to…

better decision making

SEARCH TO SUCCESS:

Turning CrimeSolutions Insights into Action

NATIONAL INSTITUTE OF JUSTICE | Crime SOLUTIONS

1 IDENTIFY COMMUNITY NEEDS

Determine, define, and prioritize specific challenges faced in the community. Narrow the scope of focus to target critical aspects of the issues.

2 NEEDS ASSESSMENT & RESOURCE REVIEW

Gather current data, outline available resources, and engage with stakeholders to identify gaps and set priorities for action.

3 SEEK POTENTIAL INTERVENTIONS WITH *CRIMESOLUTIONS*

Search crimesolutions.ojp.gov for potential programs. Filters for topic, program type, target population, setting, geography, & rating.

4 ALIGN PROGRAMS WITH GOALS

Evaluate potential programs to determine if they align with available resources, staffing, and infrastructure.

5 EXPLORE RESEARCH PARTNERSHIPS

Outreach to local universities or researchers to discuss program guidelines and protocols to ensure the implementation of the intervention goes as planned.

6 IMPLEMENTATION & MONITORING

Launch the program, continuously monitor performance using data, and make adjustments to enhance the program's success.

Let’s Practice Some Skills on our Own

Tiana’s Disclaimer:

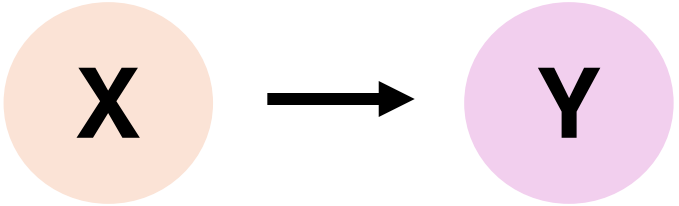
You **will not be** a data scientist/statistician/computer scientist after this lecture

But you will have a few tools in your pocket and ideas for how you can further your skills in the future

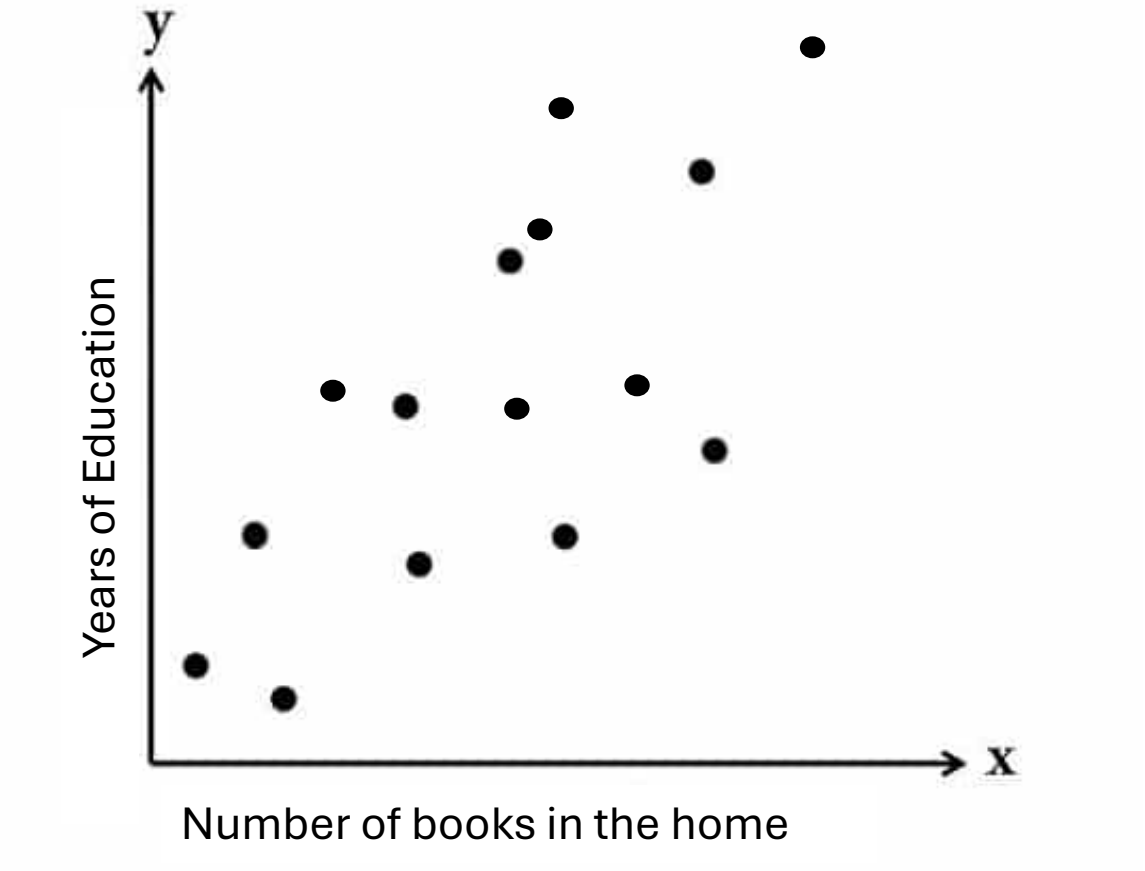
Correlation versus Causation

- Correlation: X is related to Y

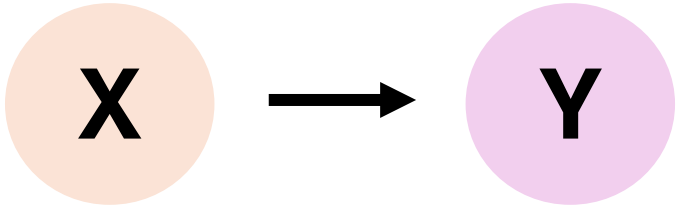
- Causation: X causes Y



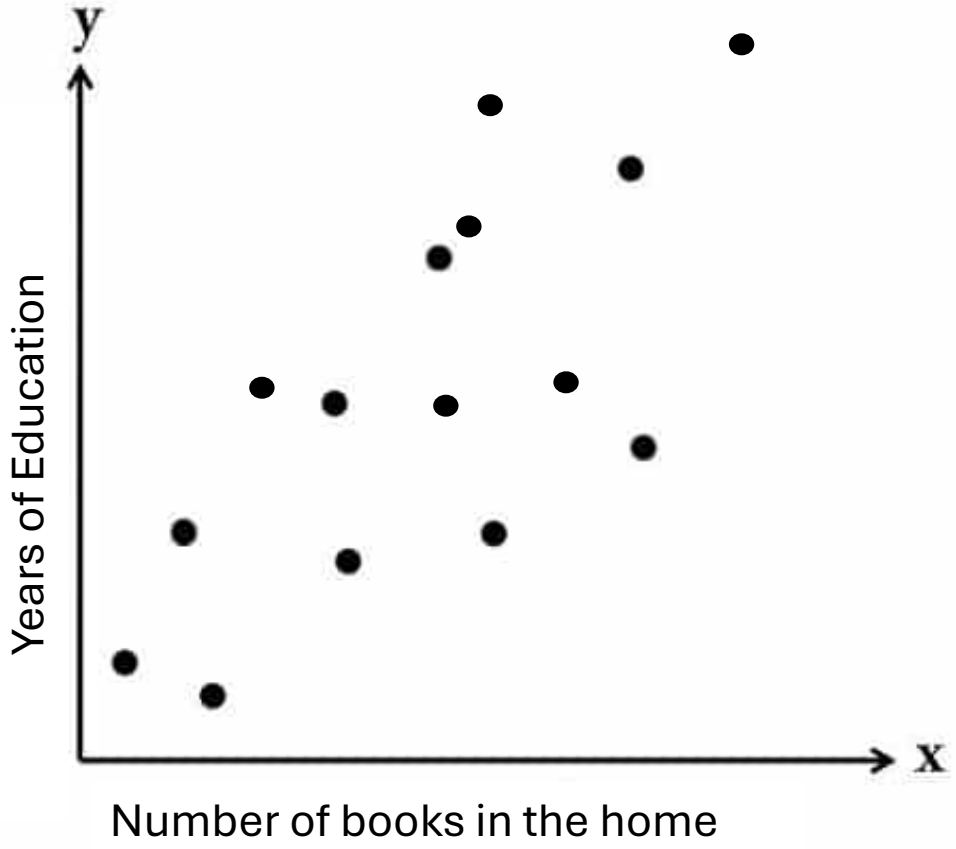
Correlation versus Causation



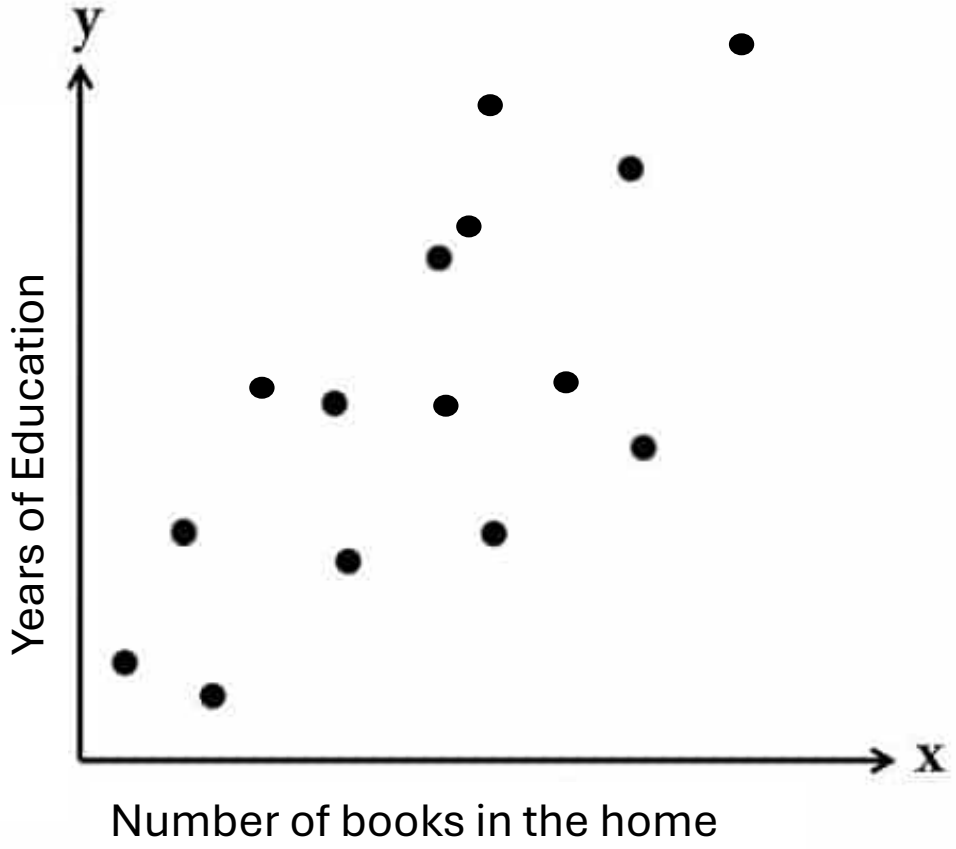
Correlation versus Causation

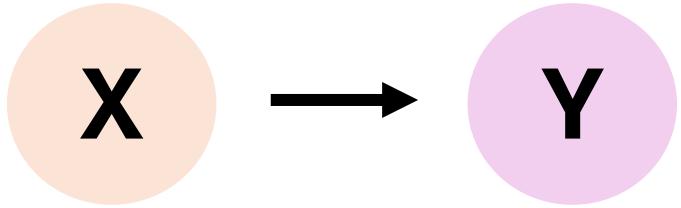


More books at home cause higher educational attainment



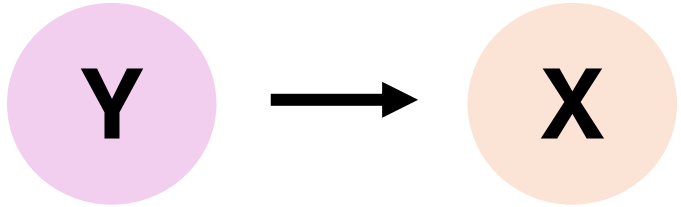
Correlation versus Causation





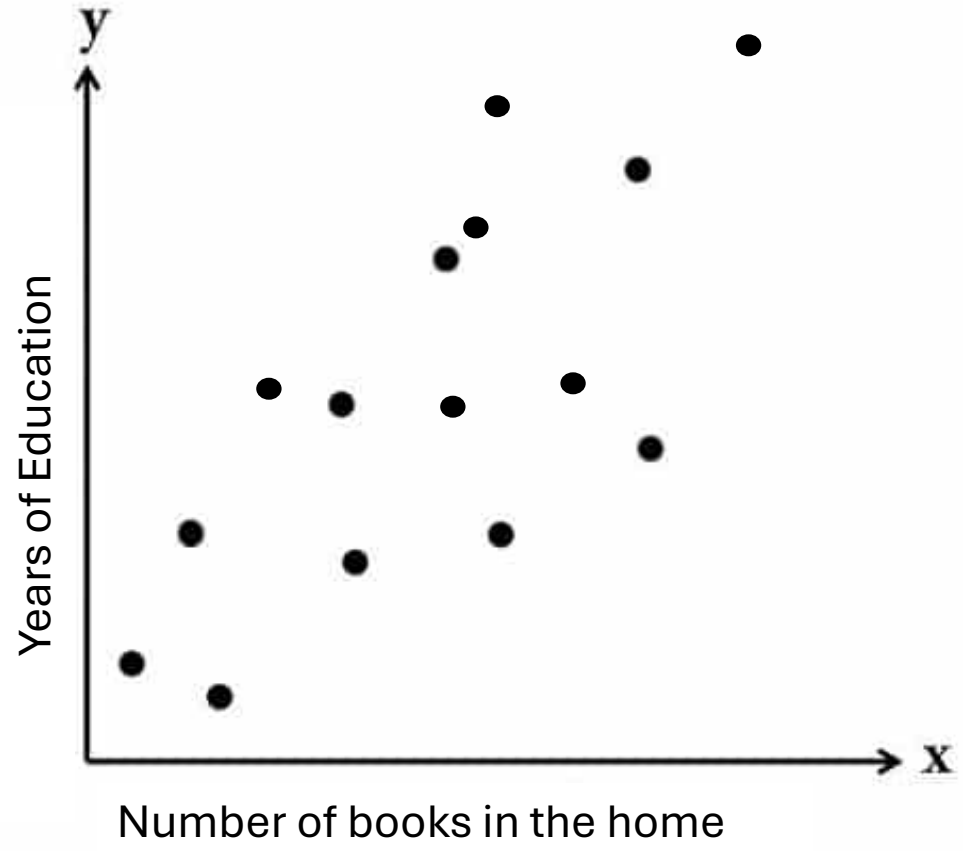
More books at home cause higher educational attainment

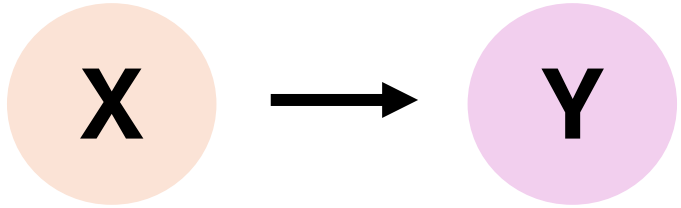
Reverse Causality



Higher educational attainment causes more books at home

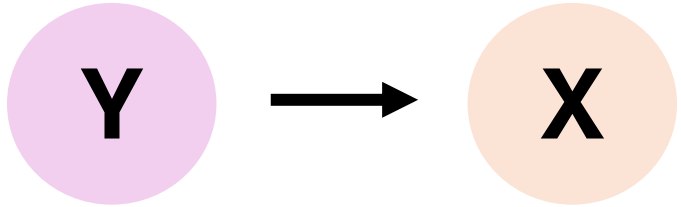
Correlation versus Causation





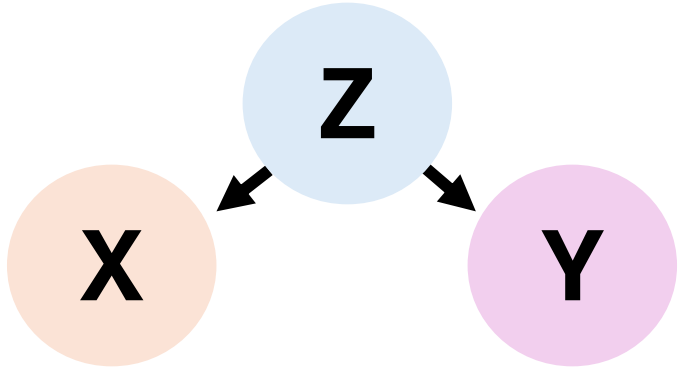
More books at home cause higher educational attainment

Reverse Causality



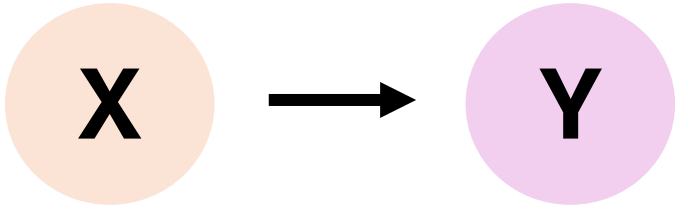
Higher educational attainment causes more books at home

Third Factor



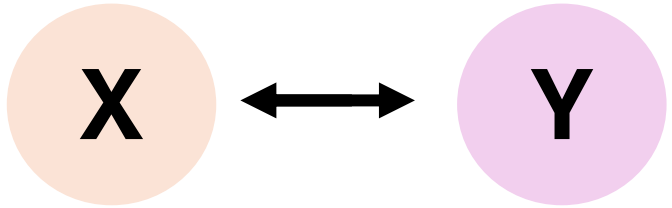
Higher income causes more books at home and higher educational attainment

Correlation versus Causation



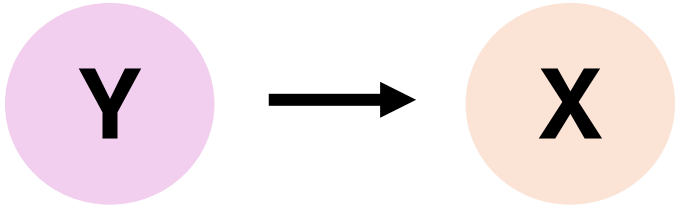
More books at home cause higher educational attainment

Bidirectional Causality



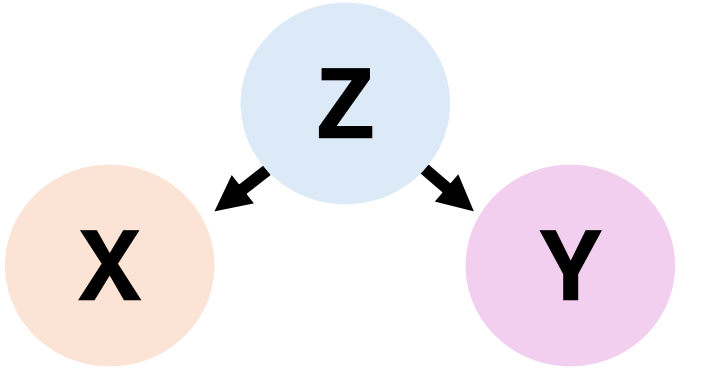
Higher educational attainment causes more books at home and more books at home cause higher educational attainment

Reverse Causality



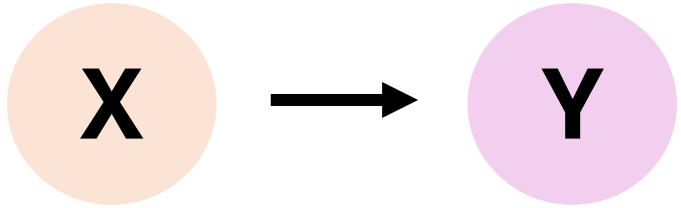
Higher educational attainment causes more books at home

Third Factor



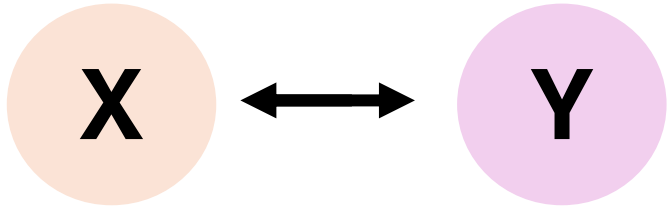
Higher income causes more books at home and higher educational attainment

Correlation versus Causation



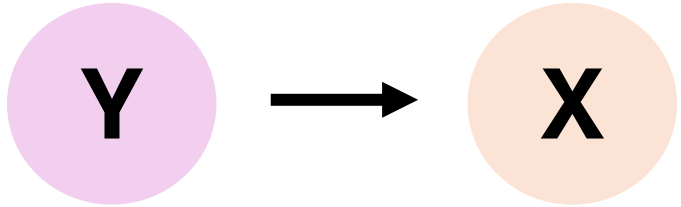
More books at home cause higher educational attainment

Bidirectional Causality



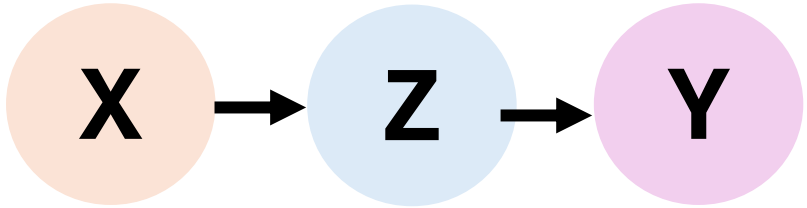
Higher educational attainment causes more books at home and more books at home cause higher educational attainment

Reverse Causality



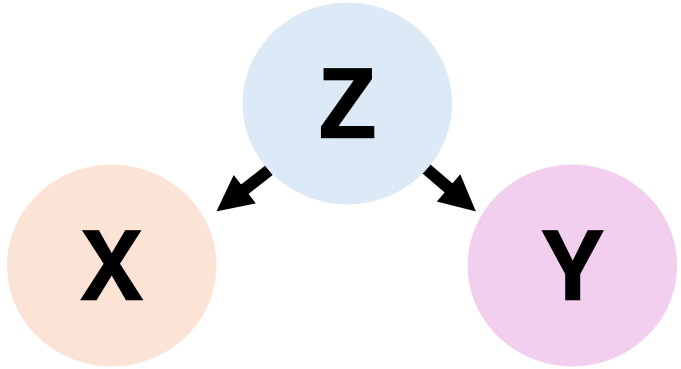
Higher educational attainment causes more books at home

Indirect Causality



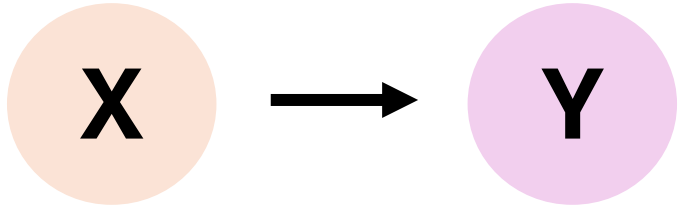
Number of books at home cause higher test scores and higher test score cause higher education

Third Factor



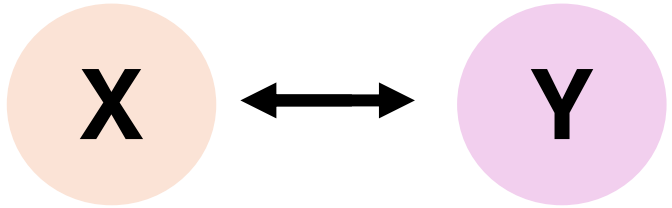
Higher income causes more books at home and higher educational attainment

Correlation versus Causation



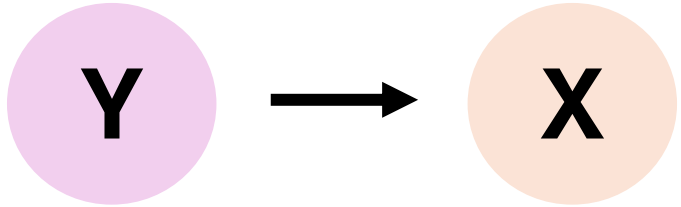
More books at home cause higher educational attainment

Bidirectional Causality



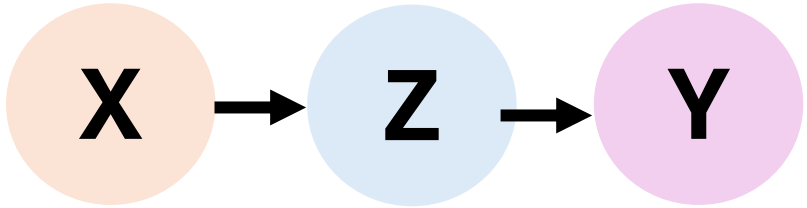
Higher educational attainment causes more books at home and more books at home cause higher educational attainment

Reverse Causality



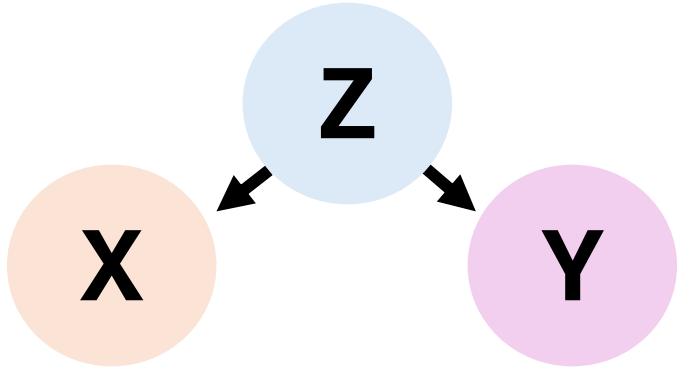
Higher educational attainment causes more books at home

Indirect Causality



Number of books at home cause higher test scores and higher test score cause higher education

Third Factor



Higher income causes more books at home and higher educational attainment

And maybe just spurious correlation!

Examples in the News

FOX29 Live News Weather Good Day Sports Co

Top baby names billionaire succe

By Stephanie Weaver | **Published** February 28, 2024 5:32pm EST | Lifestyle | Fox T

Spuric

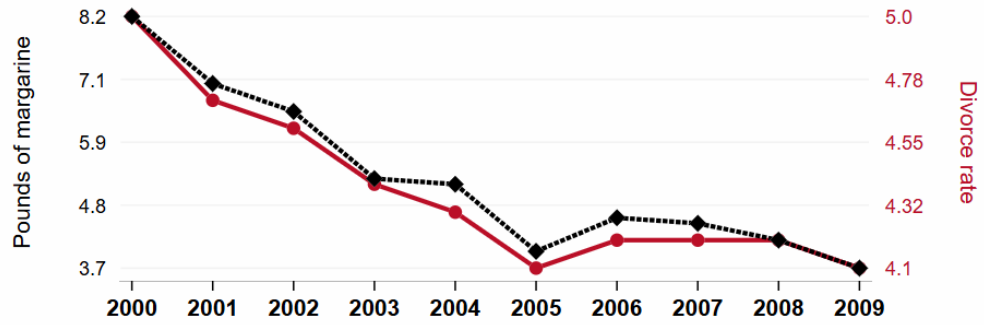
The Washington Post

Over the past 60 years, more spending on police hasn't necessarily meant less crime

June 7, 2020 More than **5 years ago**

correlates with

The divorce rate in Maine



Per capita consumption of margarine in the United States · Source: US Department of Agriculture

The divorce rate in Maine · Source: CDC National Vital Statistics

2000-2009, r=0.993, r^2=0.985, p<0.01 · tylervigen.com/spurious/correlation/5920

The Lesson:

correlation does not mean causation

- Consider the interlocking, complex nature of the problem
- Consider whether the phenomena can react to predictions
- Consider whether randomization (like a medical trial) was implemented

Data Science and PA

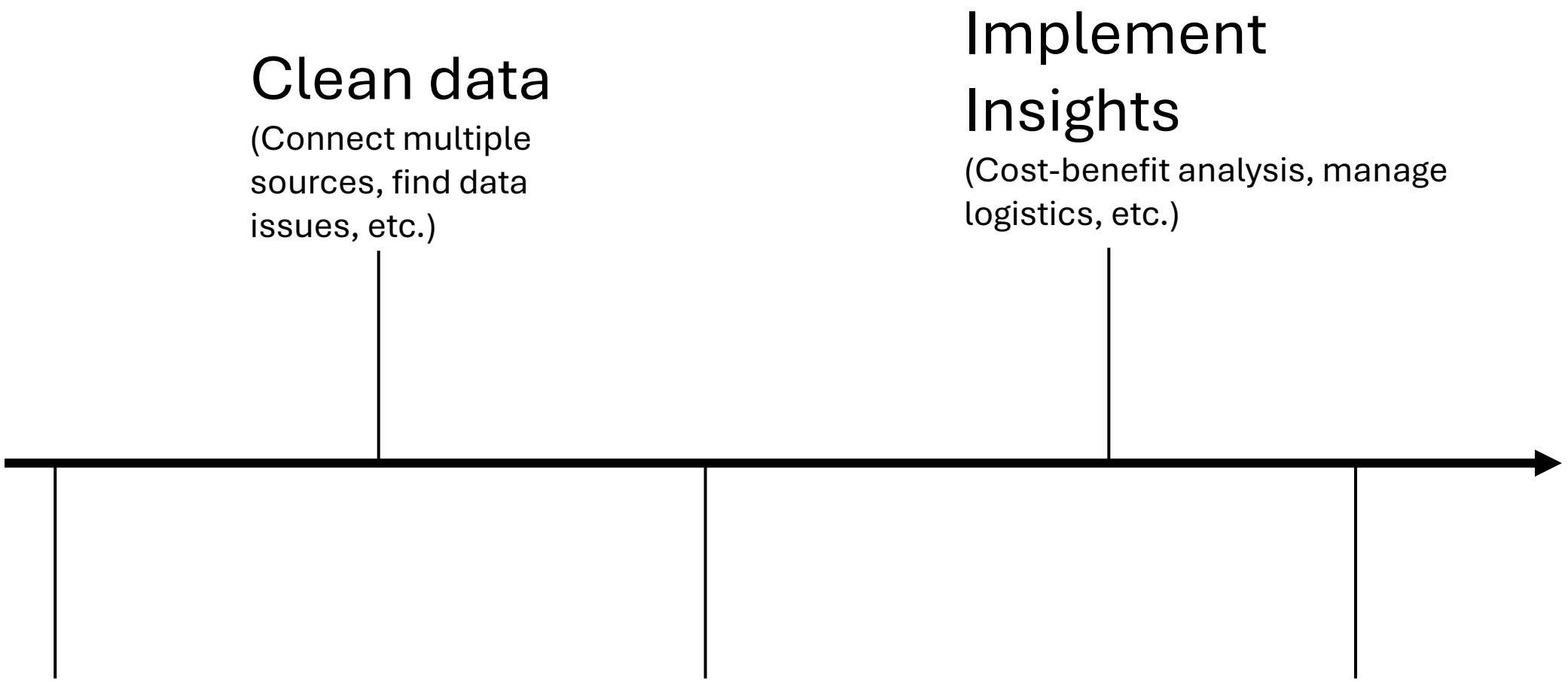
Day 2

By: Tiana Marrese, PhD

Review from Yesterday

- Data science is the practice of extracting insight from data
- Data scientists are trained in math, stats, and computer science
- There is an emphasis on domain expertise
  - Under this component, data science presents a tool that can inform decisions in law enforcement and policing
- There are general steps to implementing data insights
- There are ways that data science can be misused or misunderstood

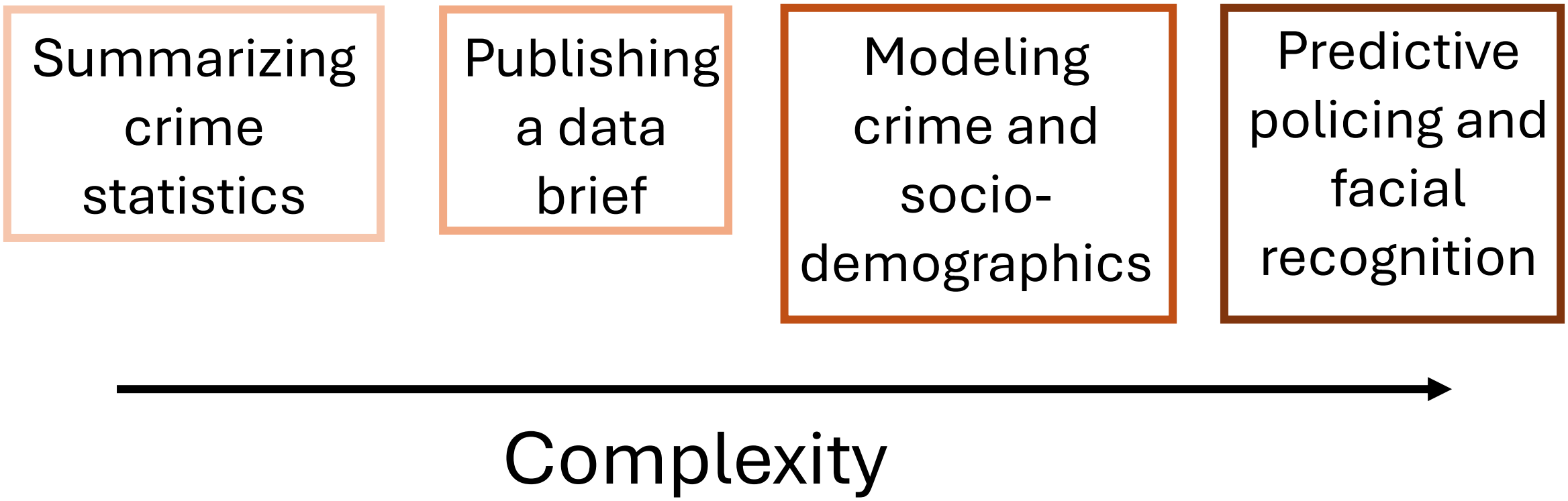
If you wanted to become a data scientist, what are tasks you can specialize in?


Clean data
(Connect multiple sources, find data issues, etc.)
Implement Insights
(Cost-benefit analysis, manage logistics, etc.)
Collect data
(Talk to stakeholders, look at data infrastructures, etc.)
Analyze data
(perform statistical models, check robustness of results, etc.)
Evaluate Implementation
(Measure outcomes, monitor change, maintain policy, etc.)

If you wanted to use data science as a tool, what are good practices?

- Take an introductory statistics course
- Read about other people that have done research on the problem
- Be skeptical of anyone that claims to have the “best” solution
- Practice critical thinking and community involvement
- Be hesitant about causal claims and “easy” solutions
- Encourage transparent governance

Data Science in Law Enforcement



Data Science in Law Enforcement: Example

Patternizr: A pattern-recognition computer software

The Good

- Can digitally compare crimes across entire city
- The computer program picks up key details
- The program is trained on 10 years of really city patterns
- Minimizes the amount of cases necessary for human review

The Bad

- Represents an over-surveillance of citizens
- The calibration of the model is not open to the public
- The 10 years of crime data reflect racial policing
- Places too much trust on automated review

Source: https://www.latimes.com/nation/la-na-new-york-computer-policing-20190310-story.h

https://www.hrw.org/news/2019/12/04/auditing-algorithms-new-york-

In Conclusion

- There’s no quick-and-easy rule book for how to think about data science in law enforcement
- Each initiative should be handled with care
- Data literacy is one of the best defenses against improper data science

Indicators

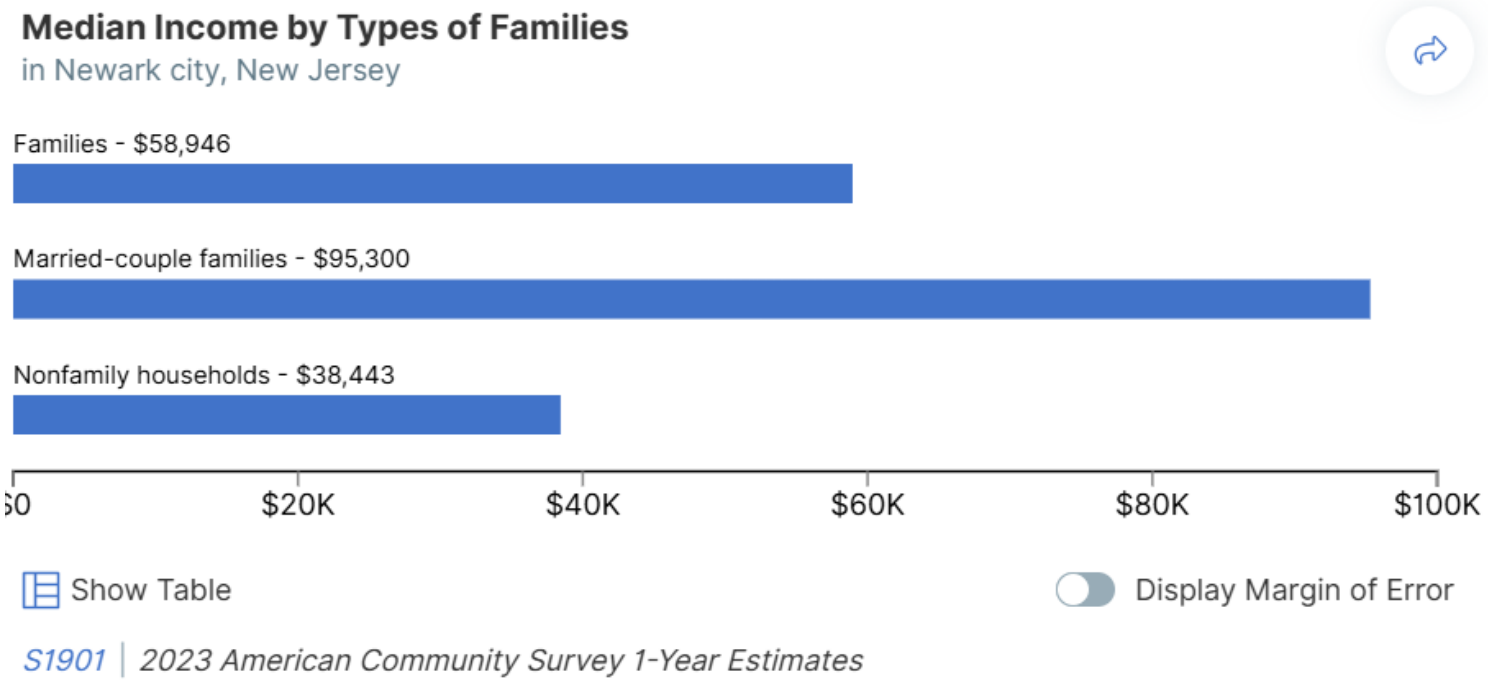
“Indicators are stylized facts that give simple insights into a complicated phenomena”

Good indicators are relatable, simple, relative, and motivate nuance

Example indicators: crime rate, median income, race

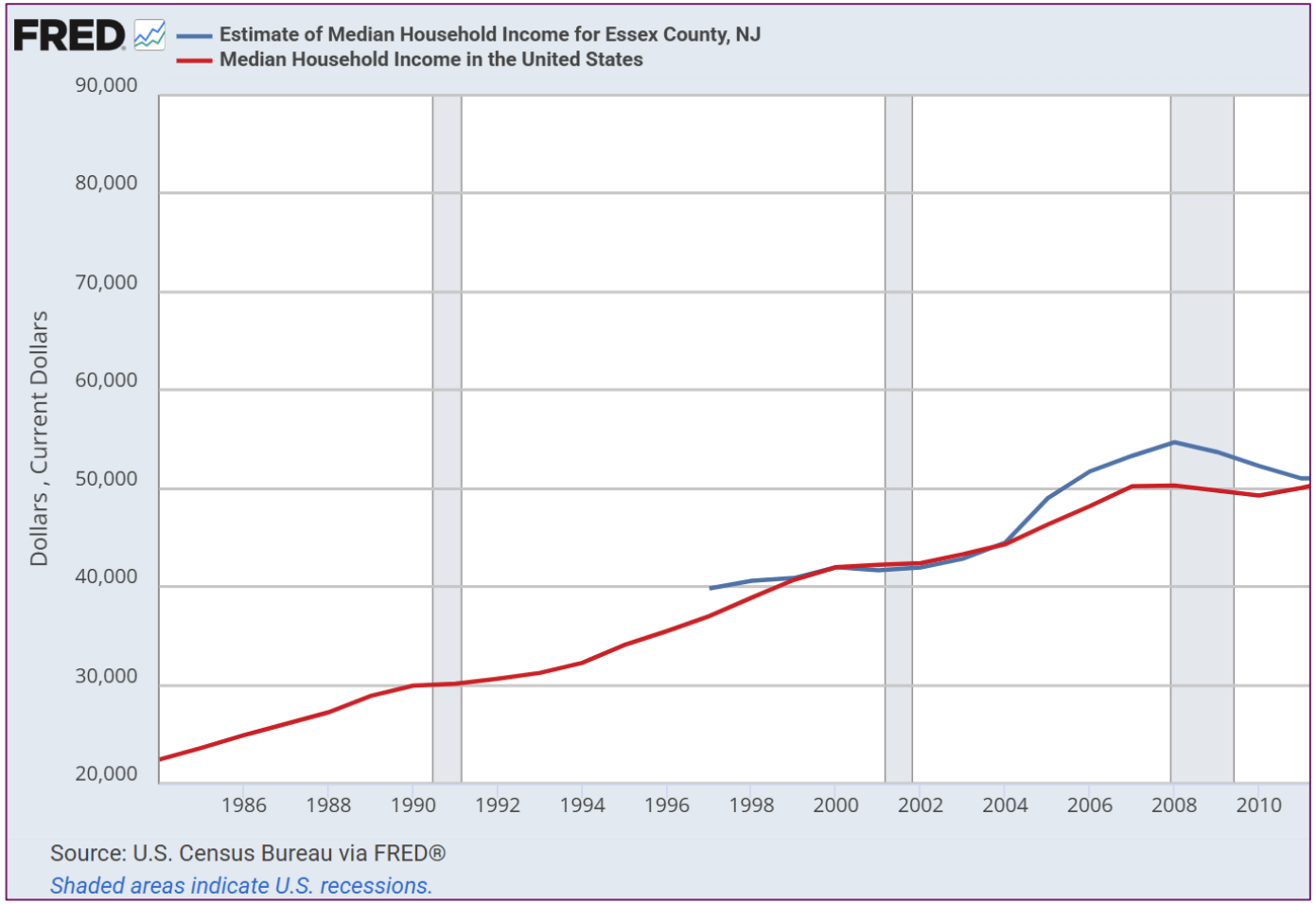
Source: https://urbanspatial.github.io/PublicPolicyAnalytics/TOD.html

A Lesson On Reading Indicator Graphics



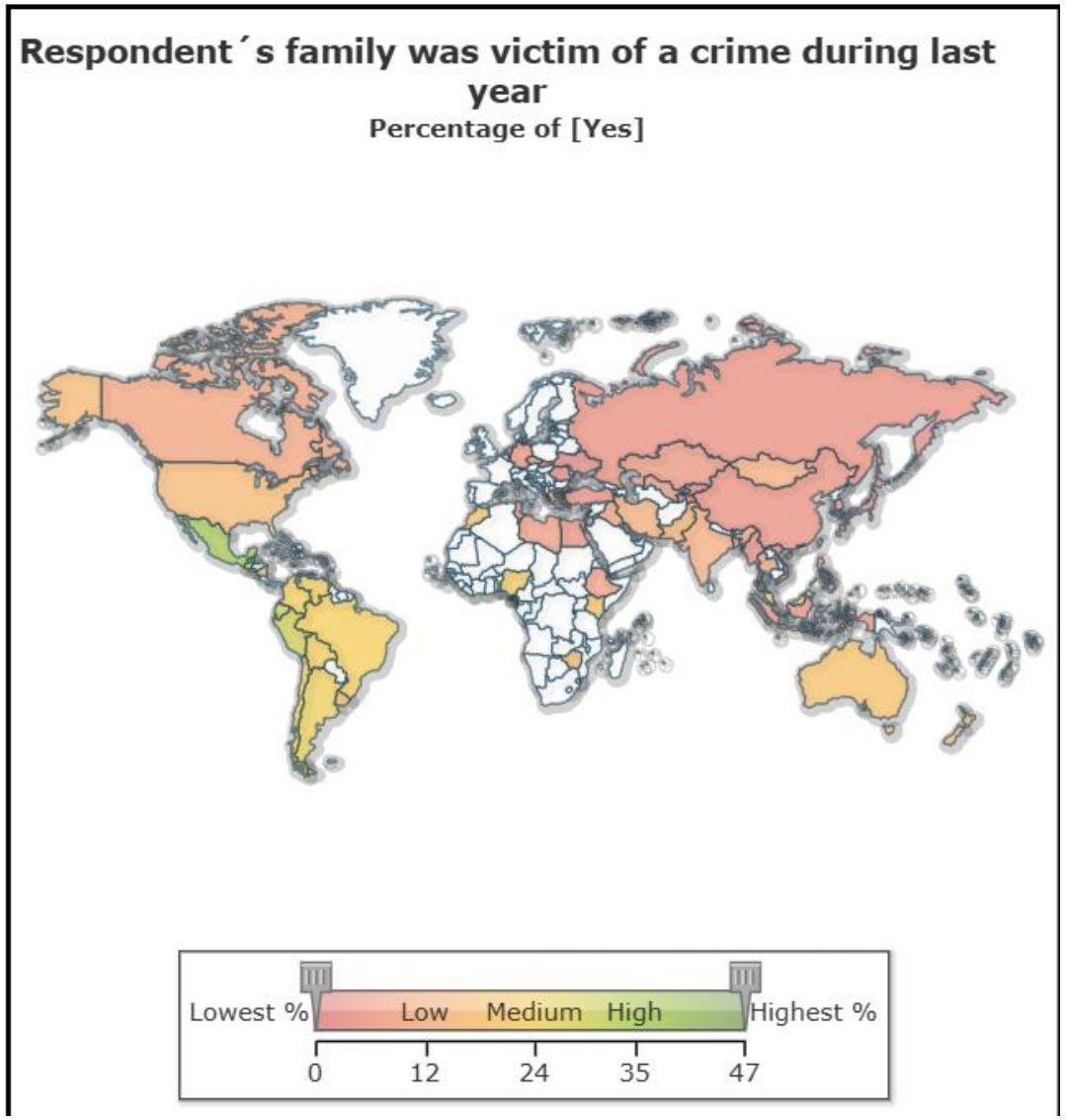
- What is the title and axes of the graph?
- What are the groupings of the data
- What information can I gather within a specific group
- What information can I gather between the groups
- What are the overall take-aways that are very apparent?
- What is unclear or unanswered by the data

A Lesson On Reading Indicator Graphics



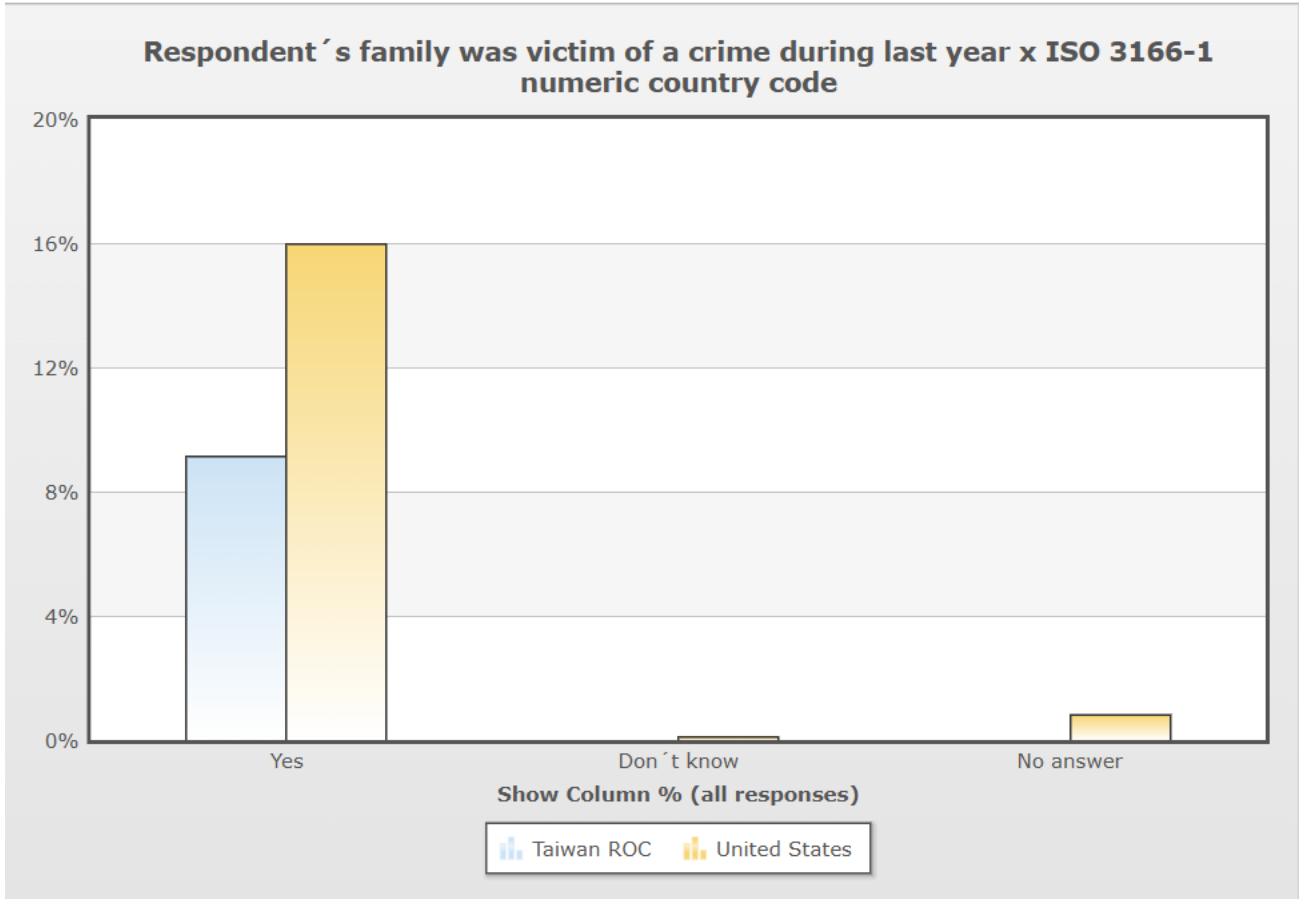
- What is the title and axes of the graph?
- What are the groupings of the data
- What information can I gather within a specific group
- What information can I gather between the groups
- What are the overall take-aways that are very apparent?
- What is unclear or unanswered by the data

A Lesson On Reading Indicator Graphics



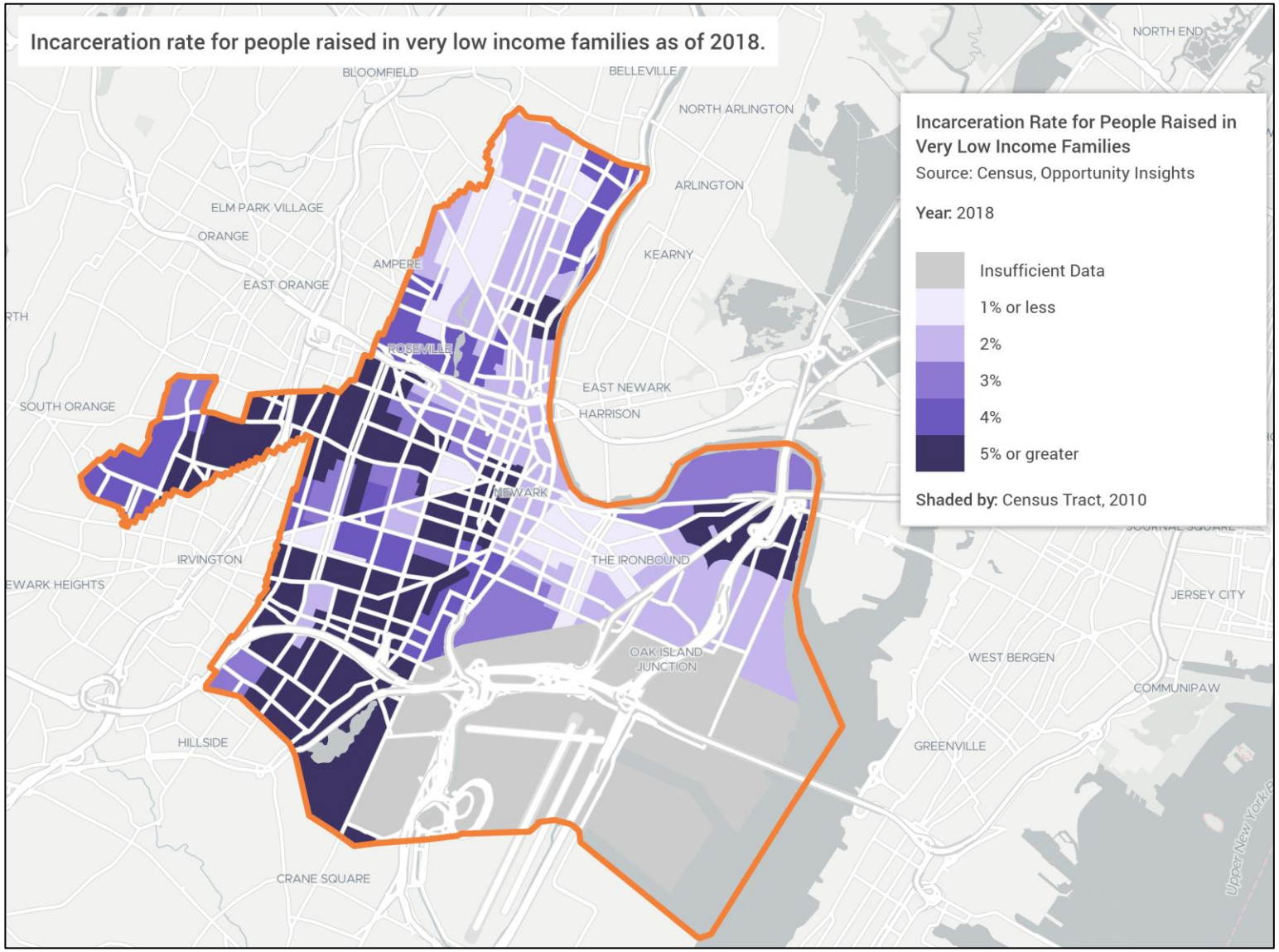
- What is the title and axes of the graph?
- What are the groupings of the data
- What information can I gather within a specific group
- What information can I gather between the groups
- What are the overall take-aways that are very apparent?
- What is unclear or unanswered by the data

A Lesson On Reading Indicator Graphics



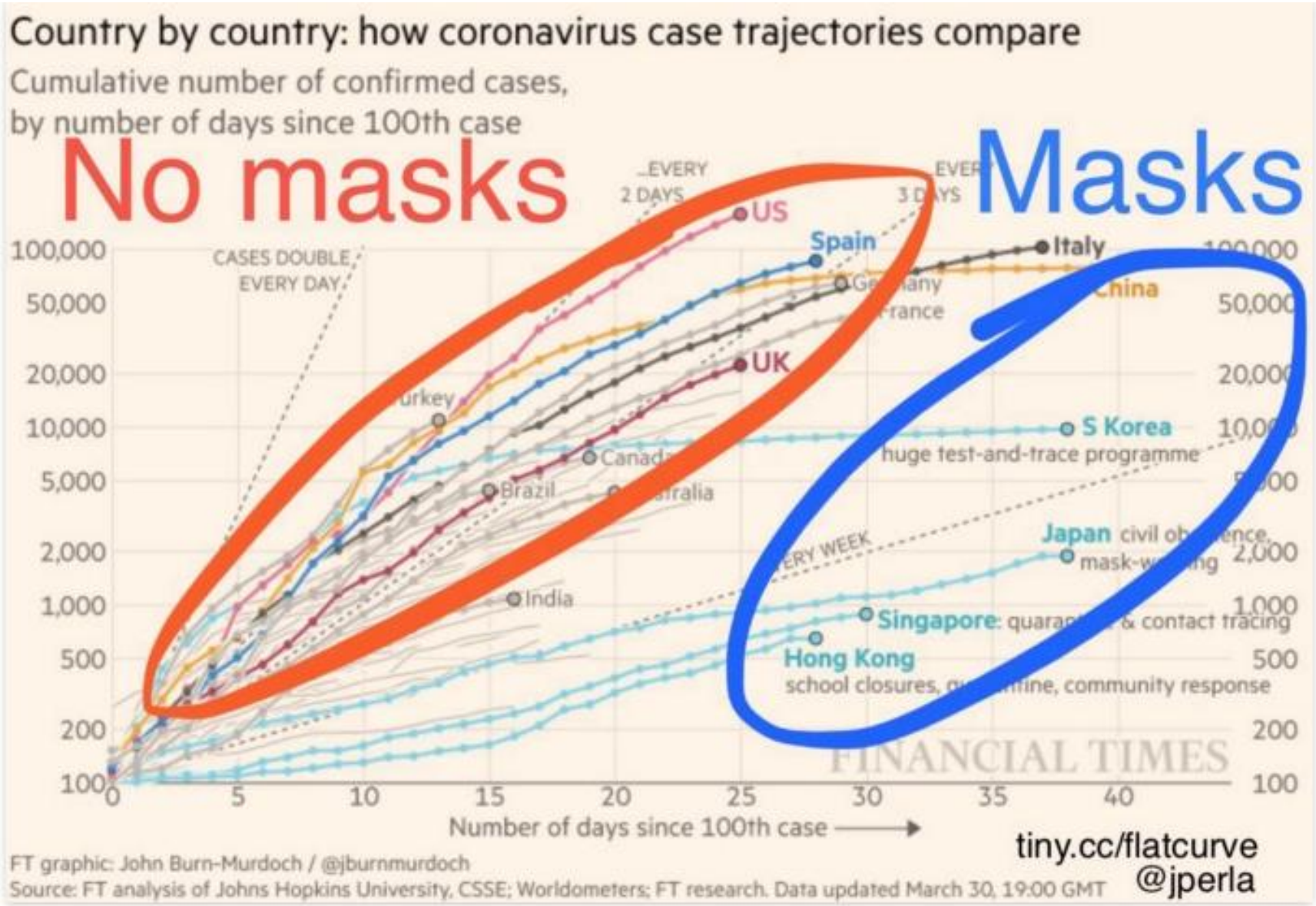
- What is the title and axes of the graph?
- What are the groupings of the data
- What information can I gather within a specific group
- What information can I gather between the groups
- What are the overall take-aways that are very apparent?
- What is unclear or unanswered by the data

A Lesson On Reading Indicator Graphics

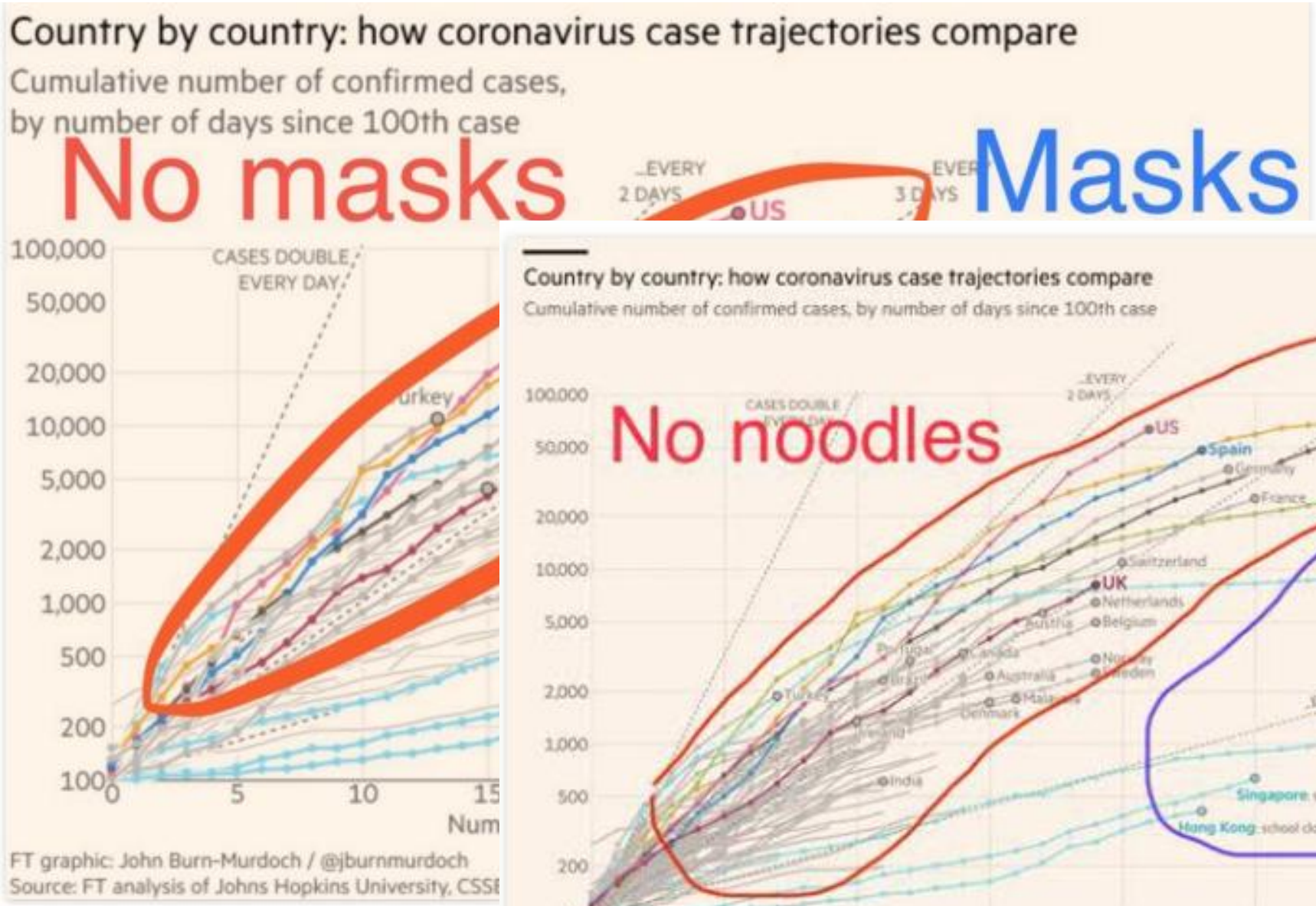


- What is the title and axes of the graph?
- What are the groupings of the data
- What information can I gather within a specific group
- What information can I gather between the groups
- What are the overall take-aways that are very apparent?
- What is unclear or unanswered by the data

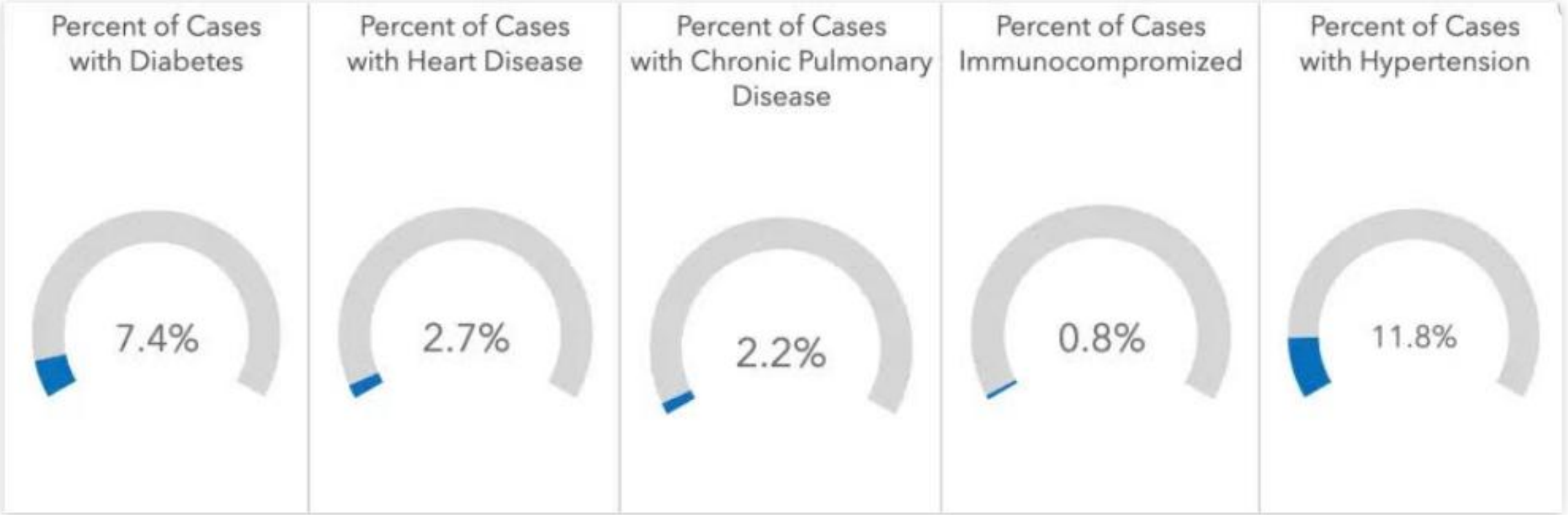
Now time for some bad graphics



Correlation is not causation!


Country by country: how coronavirus case trajectories compare
Cumulative number of confirmed cases,
by number of days since 100th case
No masks
Masks
CASES DOUBLE EVERY DAY
EVERY 2 DAYS
US
100,000
50,000
20,000
10,000
5,000
2,000
1,000
500
200
100
0
5
10
15
Num
FT graphic: John Burn-Murdoch / @jburnmurdoch
Source: FT analysis of Johns Hopkins University, CSSE

Country by country: how coronavirus case trajectories compare
Cumulative number of confirmed cases, by number of days since 100th case
No noodles
Noodles
CASES DOUBLE EVERY DAY
EVERY 2 DAYS
EVERY WEEK
US
Spain
Italy
China had 74,556
France
Iran
Switzerland
UK
Netherlands
Belgium
Canada
Australia
Denmark
India
Japan
strong social norms around obedience and mask-wearing
Singapore strict quarantine rules & contact tracing
Hong Kong school closures, quarantine, community response
100,000
50,000
20,000
10,000
5,000
2,000
1,000
500
200
100
0
5
10
15
20
25
Number of days since 100th case



This data shows the % of COVID cases that also have other severe illnesses. The baseline of 100% is misleading because it makes it seem like this is a small percentage. But it's not really answering any important question

An extremely necessary question would be: what percentage of these cases survived?

Considered timeframe is important to the context too. This graphic makes it seem like U.S. is on the road to economic recovery...

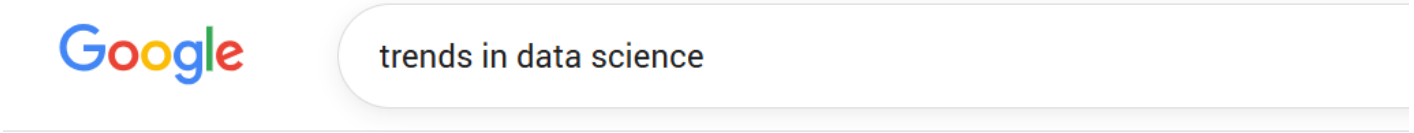
The timeframe is only January 2020 to June 2020, not a very helpful visual when considering the economy leading up to 2020

Practice

- Go to the World Value Survey (https://www.worldvaluessurvey.org/wvs.jsp)
- Select: Data and documentation > online analysis

- Select year, countries, variable of interest
- Try to cross with another variable then graph it

The Current State of Data Science

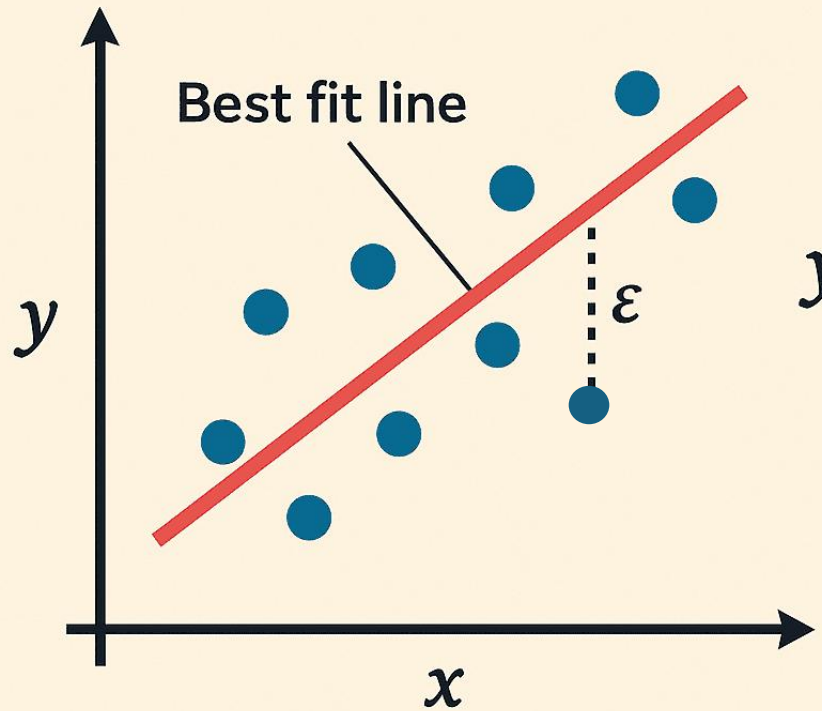


Here are the key trends shaping data science:

Core Technologies & Methodologies

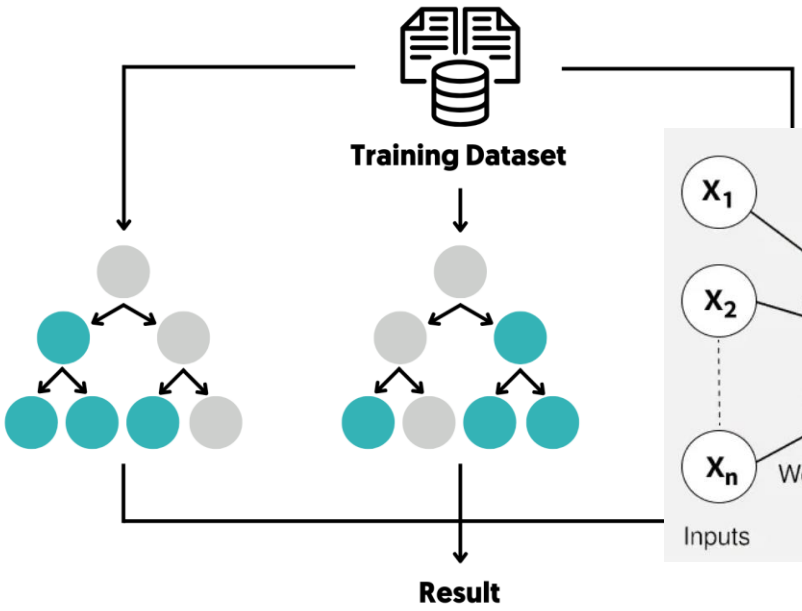
- **AI & Machine Learning (ML) Integration**: Deeper embedding of AI for predictive modeling, enhanced personalization, and automating complex tasks.
- **Generative AI**: Its rapid rise is impacting how data scientists work, with high urgency to adopt for creative and analytical tasks, notes the MIT Sloan Management Review.
- **Automated Machine Learning (AutoML)**: Automates model building (tuning, feature engineering) for faster, more efficient data science, notes Digital Regenesys.
- **Explainable AI (XAI)**: Focus on transparency and understanding complex model decisions (like SHAP, LIME), crucial for trust, notes Digital Regenesys.
- **Augmented Analytics**: AI-driven automation of data prep, insight generation, and discovery, notes Digital Regenesys.
- **Natural Language Processing (NLP)**: Advancements leading to more context-aware language understanding.

LINEAR REGRESSION

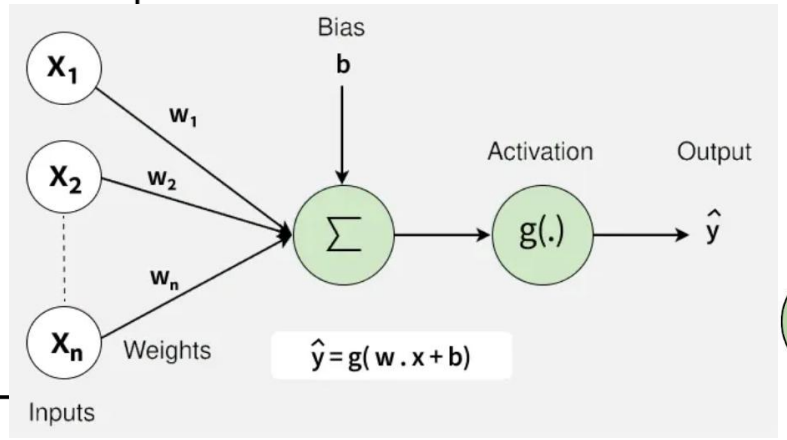


$$y = \beta_0 + \beta_1 x + \varepsilon$$

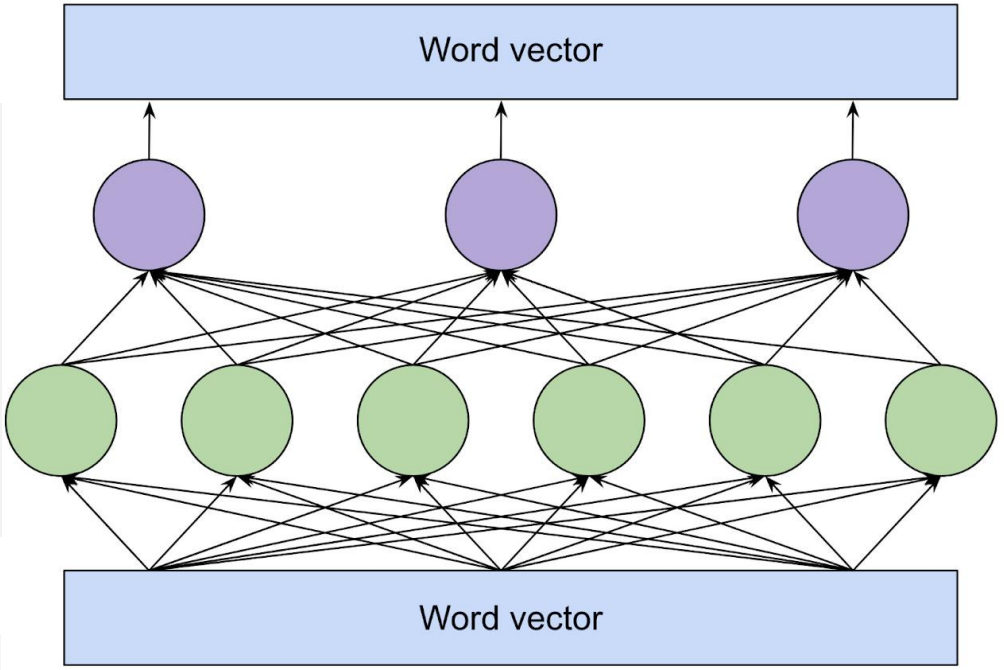
Random Forest Model



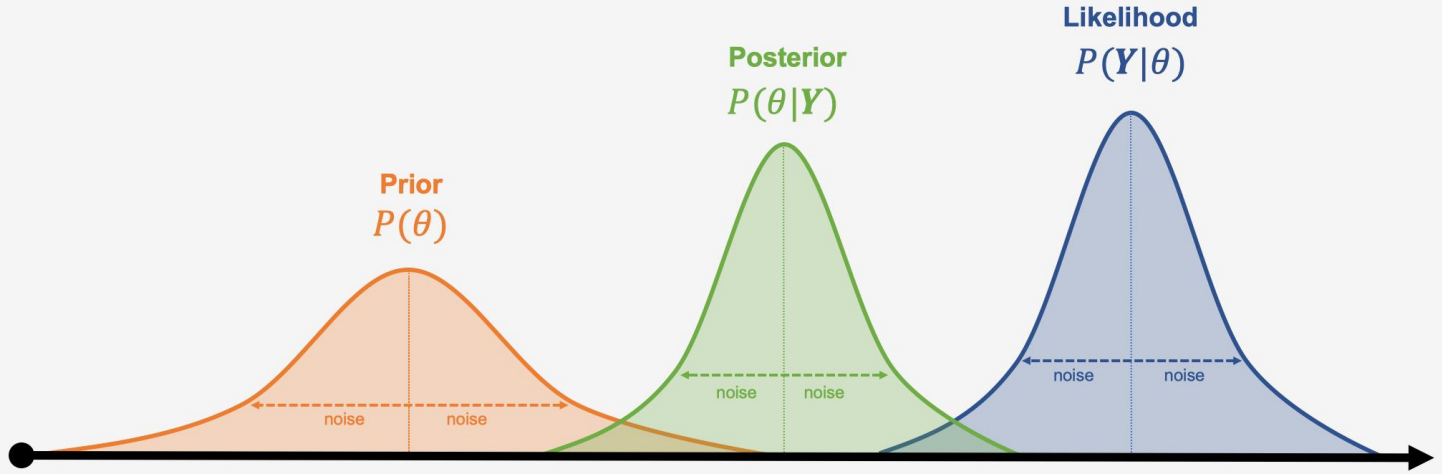
Neural Network



Large Language Model



Bayesian Model



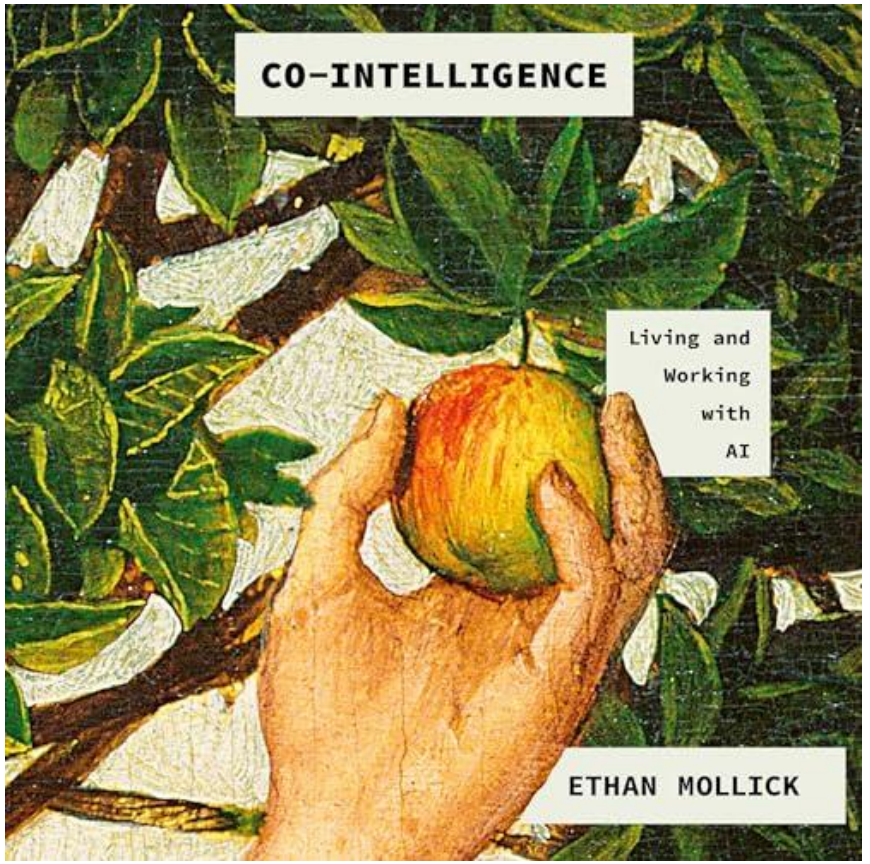
Given the fast-changing world of data science, technology competency is more important than ever

AI – What It Is and How It Relates to Your Work

- **Artificial Intelligence:** machines perform human-like, problem-solving tasks
- **Generative AI:** technology that creates original content
- **Machine learning**: models underneath these applications. They “train” algorithms based on input data (reCapcha) and then predict inferences based on this learning data.

Using AI Effectively

- You need about 10 hours of relationship building. Assign a personality
- Consider it “co-creation”
  - Ask it to generate 10 different opening sentences
  - Ask it to recite the main ideas of your paper
  - Ask it to take notes on your citations
- AI does best with redirection rather than direction. It’s a *learning* model
- Ask how it can be a creative partner rather than how it can give you an answer



Source: https://www.nytimes.com/2024/04/02/opinion/ezra-klein-podcast-ethan-mollick.html

AI USE	ROLE	PEDAGOGICAL BENEFIT	PEDAGOGICAL RISK
MENTOR	Providing feedback	Frequent feedback improves learning outcomes, even if all advice is not taken.	Not critically examining feedback, which may contain errors.
TUTOR	Direct instruction	Personalized direct instruction is very effective.	Uneven knowledge base of AI. Serious confabulation risks.
COACH	Prompt metacognition	Opportunities for reflection and regulation, which improve learning outcomes.	Tone or style of coaching may not match student. Risks of incorrect advice.
TEAMMATE	Increase team performance	Provide alternate viewpoints, help learning teams function better.	Confabulation and errors. "Personality" conflicts with other team members.
STUDENT	Receive explanations	Teaching others is a powerful learning technique.	Confabulation and argumentation may derail the benefits of teaching.
SIMULATOR	Deliberate practice	Practicing and applying knowledge aids transfer.	Inappropriate fidelity.
TOOL	Accomplish tasks	Helps students accomplish more within the same time frame.	Outsourcing thinking, rather than work.

https://www.oneusefulthing.org/p/assigning-ai-seven-ways-of-using

Keep in mind…

- It might not work the first time you try it
- It’s not a real human, even though it might feel that way
- It can make things up
- You are in charge
- Only share things you’re comfortable with
- If one model doesn’t seem to be working, try another
- Give it context, ask questions, seek clarification
- Direct it to the ways that

Keys to Spotting AI Photos

- Weird or misplaced objects including body parts, shadows, and text
- Unrealistic textures for skin or fabric
- Watermarks or icons

Source: https://www.getcybersafe.gc.ca/en/resources/recognize-artificial-intelligence-ai-9-ways-spot-ai-content-online

AI Videos

- https://www.npr.org/2025/11/30/nx-s1-5610951/fake-ai-videos-slop-quiz

Keys to Spotting AI Videos

- **Video length:** AI videos are 8-10 seconds (as of Jan. 2026)
- **Video angle and framing:** AI videos usually occur at angles that appear too good to be true. They also have trouble mimicking security footage features like date and time stamps
- **Context matters:** Consider the place it's being shared, who posted it, reverse image search to find original post

- Importantly: don't assume everything is fake. Liar's dividend is is the phenomenon where bad actors claim real events are fake to evade responsibility.