Course Reminders

- Final Project due Tue Jun 8th (10AM)
  - Report (GitHub)
  - Video (put on YouTube [unlisted if desired], link in final report)
  - Team Evaluation Survey: https://forms.gle/92WekBxi4BnvTb2o9 (link also on Canvas; required)
- Post COGS 108 Survey: https://forms.gle/MfXnUHdHMP4HJeQy5 (link also on Canvas; *optional* for EC)
- CAPEs: http://cape.ucsd.edu/ (~45%, EC>75%!)

Errors of measurement - are we measuring what we think we are?

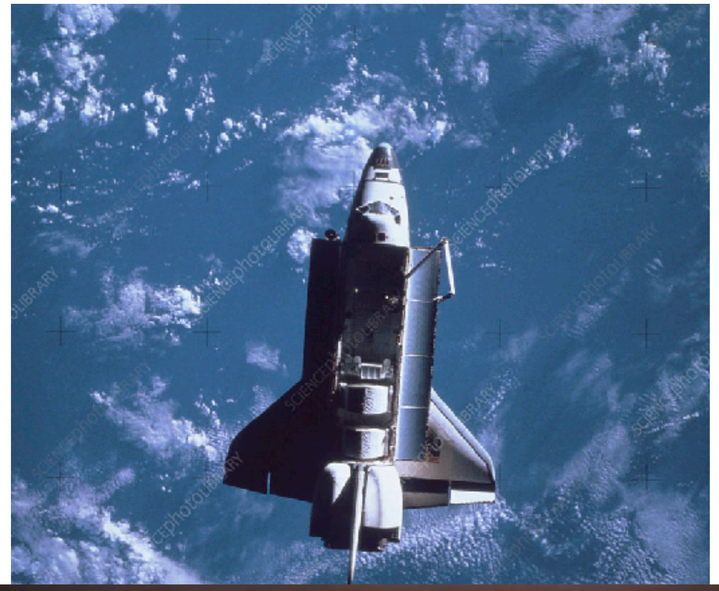
Errors of analysis - did we use the right methods to address the question?

Errors of borked tools - choosing the wrong tools or using them poorly leads to bad results

Errors of human cognition - data science is a human endeavor with all the usual frailties and foibles

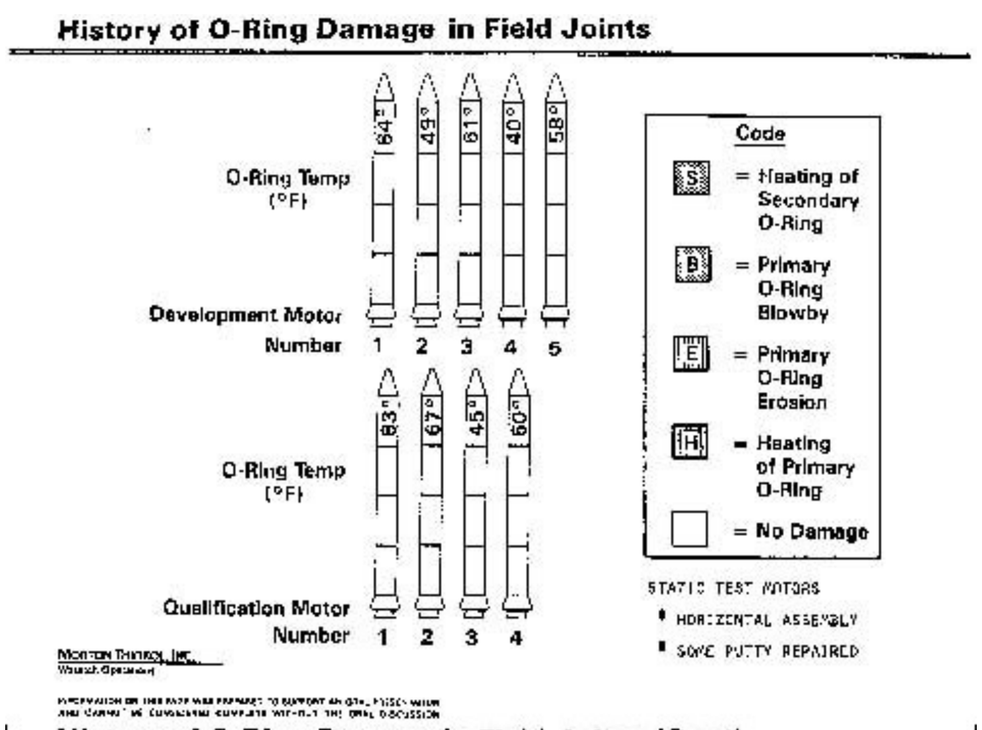
Errors of communication - sometimes you get everything right, but the group and the decision makers never understand properly

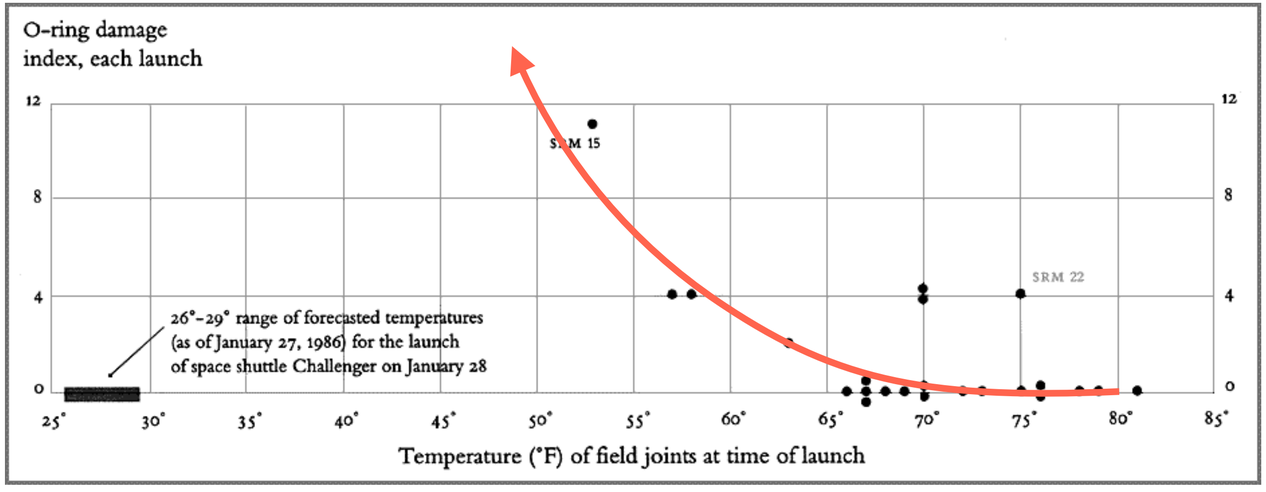
Errors of communication



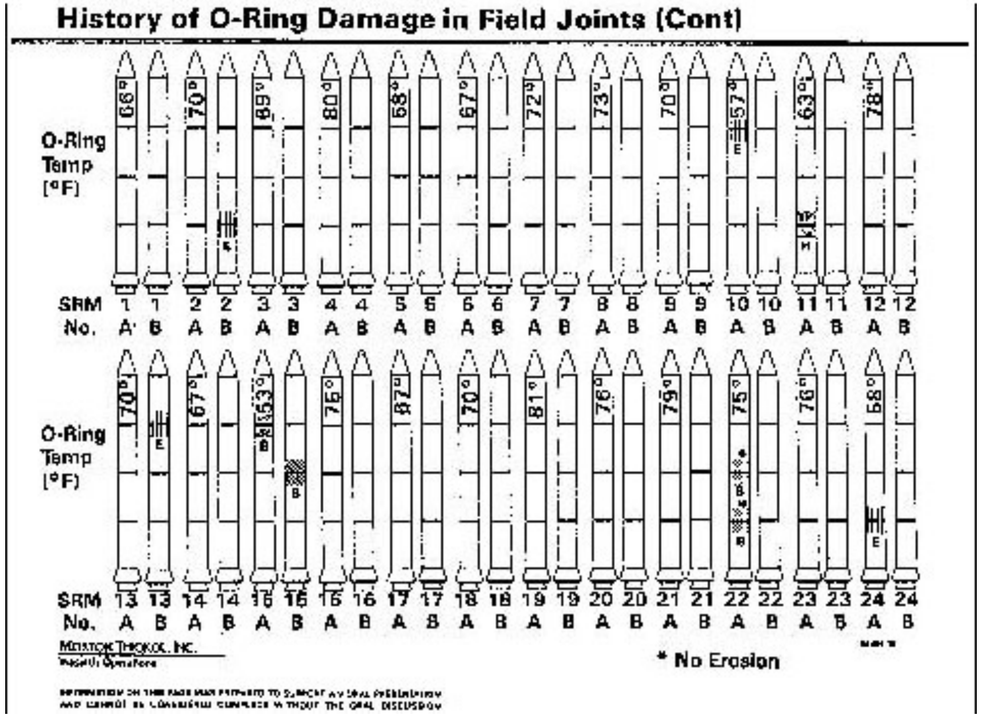
Jan 28, 1986

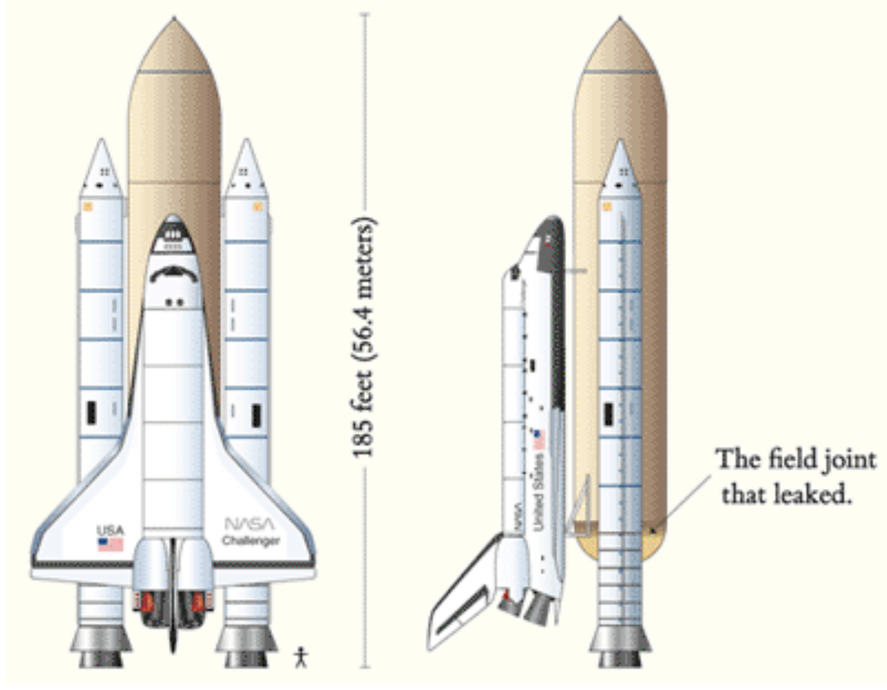
Feb 1, 2003



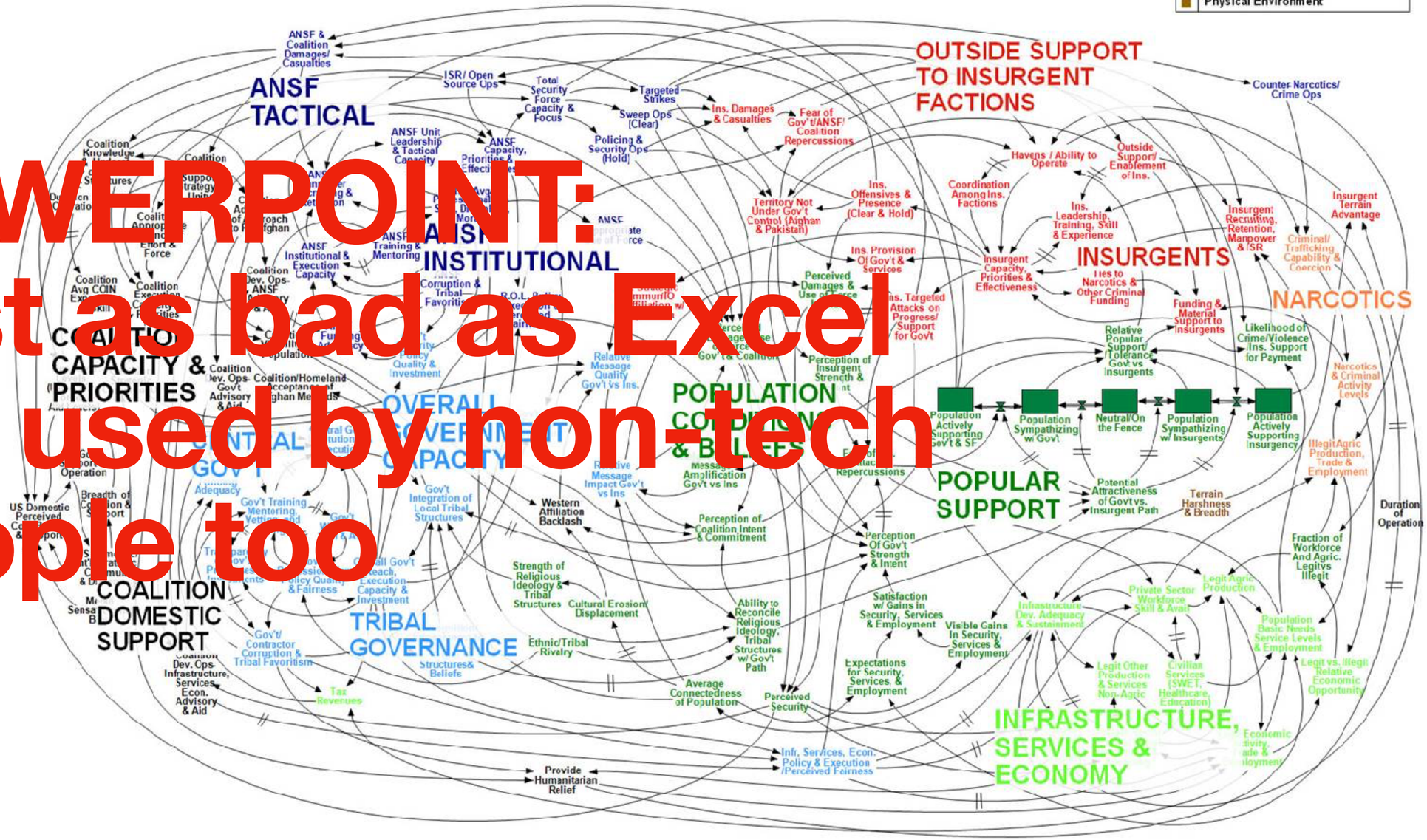


Images and graphs from Edward Tufte





Afghanistan Stability / COIN Dynamics



POWERPOINT: Just as bad as Excel but used by non-tech people too

On this one Columbia slide, a PowerPoint festival of bureaucratic hyper-rationalism, 6 different levels of hierarchy are used to display, classify, and arrange 11 phrases:

Level 1	Title of Slide
Level 2	● Very Big Bullet
Level 3	– big dash
Level 4	◆ medium-small diamond
Level 5	• tiny square bullet
Level 6	( ) parentheses ending level 5

The analysis begins with the dreaded Executive Summary, with a conclusion presented as a headline: "Test Data Indicates Conservatism for Tile Penetration." This turns out to be unmerited reassurance. Executives, at least those who don't want to get fooled, had better read far beyond the title.

The "conservatism" concerns the *choice of models* used to predict damage. But why, after 112 flights, are foam-debris models being calibrated during a crisis? How can "conservatism" be inferred from a loose comparison of a spreadsheet model and some thin data? Divergent evidence means divergent evidence, not inferential security. Claims of analytic "conservatism" should be viewed with skepticism by presentation consumers. Such claims are often a rhetorical tactic that substitutes verbal fudge factors for quantitative assessments.

As the bullet points march on, the seemingly reassuring headline fades away. Lower-level bullets at the end of the slide undermine the executive summary. This third-level point notes that "Flight condition [that is, the debris hit on the Columbia] is significantly outside of test database." How far outside? The final bullet will tell us.

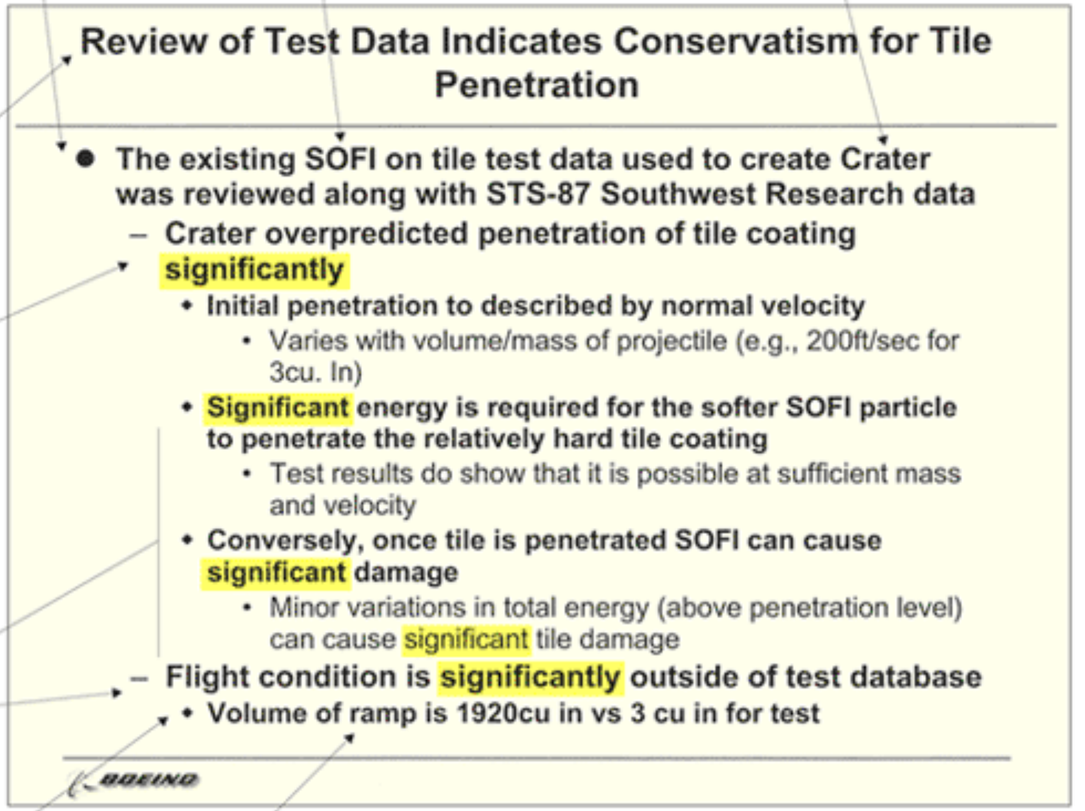
This fourth-level bullet concluding the slide reports that the debris hitting the Columbia is estimated to be 1920/3 = 640 times larger than data used in the tests of the model! The correct headline should be "Review of Test Data Indicates Irrelevance of Two Models." This is a powerful conclusion, indicating that pre-launch safety standards no longer hold. The original optimistic headline has been eviscerated by the lower-level bullets.

Note how close readings can help consumers of presentations evaluate the presenter's reasoning and credibility.

The Very-Big-Bullet phrase fragment does not seem to make sense. No other VBB's appear in the rest of the slide, so this VBB is not necessary.

Spray On Foam Insulation, a fragment of which caused the hole in the wing

A model to estimate damage to the tiles protecting flat surfaces of the wing



Here "ramp" refers to foam debris (from the bipod ramp) that hit Columbia. Instead of the cryptic "Volume of ramp," say "estimated volume of foam debris that hit the wing." Such clarifying phrases, which may help upper level executives understand what is going on, are too long to fit on low-resolution bullet outline formats. PP demands the shorthand of acronyms, phrase fragments, and clipped jargon in order to get at least some information into the tight format.

Edward Tufte

Our models are irrelevant

Debris hitting the wing was 640x larger than the experimental data used to build these models

We have no clue what will happen on re-entry

Communication is key

Identify audience & setting

Identify key insight, main points of evidence, and assumptions

Organize into a story focussed on ⬆️

Create supporting visualizations

Revise to be as precise and concise as possible

Your future in DS

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com

Slides in this presentation are from material kindly provided by
Shannon Ellis and Brad Voytek

Courses in DS and ML at UCSD

- DS
- CSE
- CS
- ECE
- COGS
- But also many other departments like ECON, MATH, LING, BENG, etc

My list of '20-21 ML (and ML adjacent) courses

Some job titles and what they do

- Analytics or statistician: data handling, analysis
- Data scientist: programming, data handling, analysis
- Data engineer: programming, databases, management
- Data architect: programming, databases, design
- Data manager: databases, design, management
- *OPs (eg, devOPs, dataOPs, full stack): programming, tool development, mangagement concentrating on end to end process
- ML Engineer: programming, tool development, management of infrastructure
- ML researcher: programming, algorithm design and testing

Glut of new data scientists

First, let's talk about the oversupply of junior data scientists. The continuing media hype cycle around data science has enormously exploded the amount of junior talent available on the market over the past five years.

This is purely anecdotal evidence, so take it with a large grain of salt. But, based on my own participation as a resume screener, mentor to data scientists leaving boot camps, interviewer, interviewee, and from conversations with friends and colleagues in similar positions, I've developed an intuition that the number of candidates per any given data science position, particularly at the entry level, has grown from 20 or so per slot, to 100 or more. I was talking to a friend recently who had to go through 500 resumes for a single opening.

This is not abnormal. More anecdotal evidence comes from job openings like this one, from machine learning's godfather, Andrew Ng, whose AI startup demanded 70-80 hours a week. He was flooded with applications, after blithely noting that previously many people had tried to volunteer for free. As of this latest writing, they ran out of space in their current office.

It's very, very hard to estimate the true gap between market demand and supply, but here's a starting point.

https://veekaybee.github.io/2019/02/13/data-science-is-different/

Advice from Vicki Boykis

Sr. Manager, Data Science + Engineering at CapTech Ventures, Inc

1. Learn SQL
2. Learn a programming language extremely well and learn programming concepts.
3. Learn how to work in the cloud.
4. This stuff is really hard for everyone, and there are a million things it seems like you have to know. Don’t get discouraged.

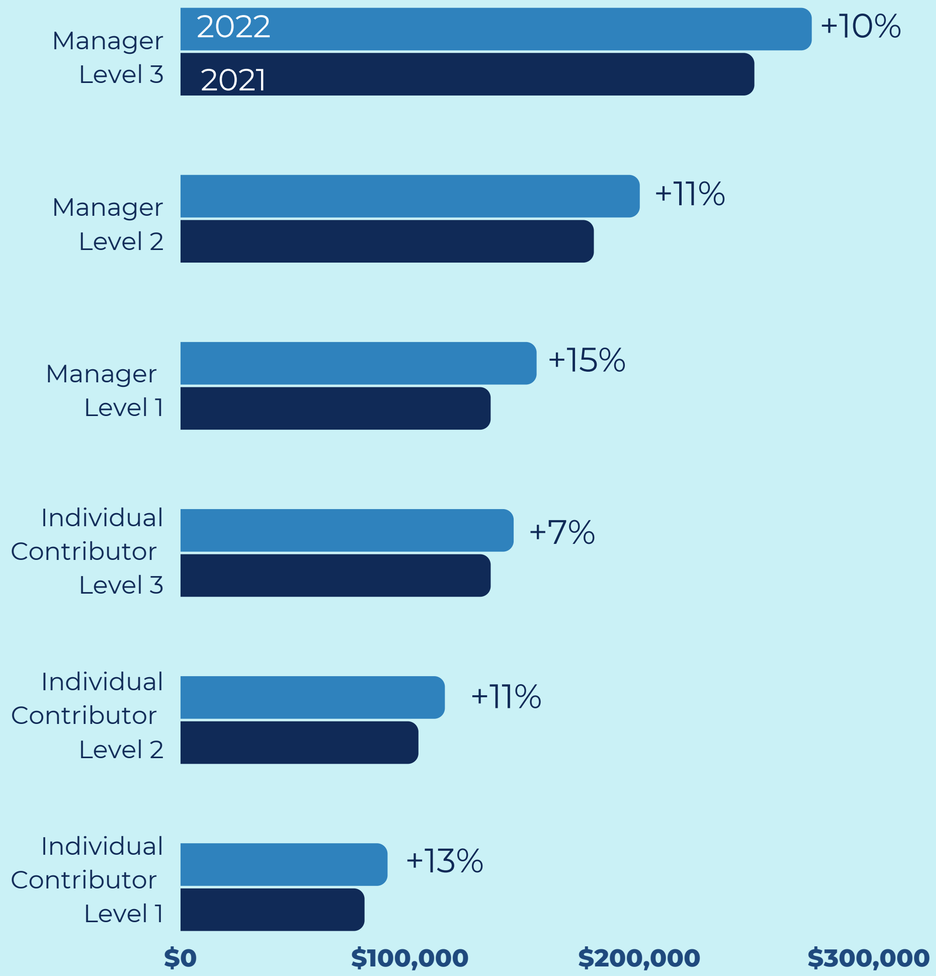
https://veekaybee.github.io/2019/02/13/data-science-is-different/

	Job Title	Median Base Salary	Job Satisfaction	Job Openings
#1	Enterprise Architect	$144,997	4.1/5	14,021
#2	Full Stack Engineer	$101,794	4.3/5	11,252
#3	Data Scientist	$120,000	4.1/5	10,071
#4	Devops Engineer	$120,095	4.2/5	8,548
#5	Strategy Manager	$140,000	4.2/5	6,977
#6	Machine Learning Engineer	$130,489	4.3/5	6,801

Source: Glassdoor (2023)

Burtchworks annual predictions and report on DS hiring

Comparison of **Median Base Salaries** by Job Level for **Data Science Professionals**

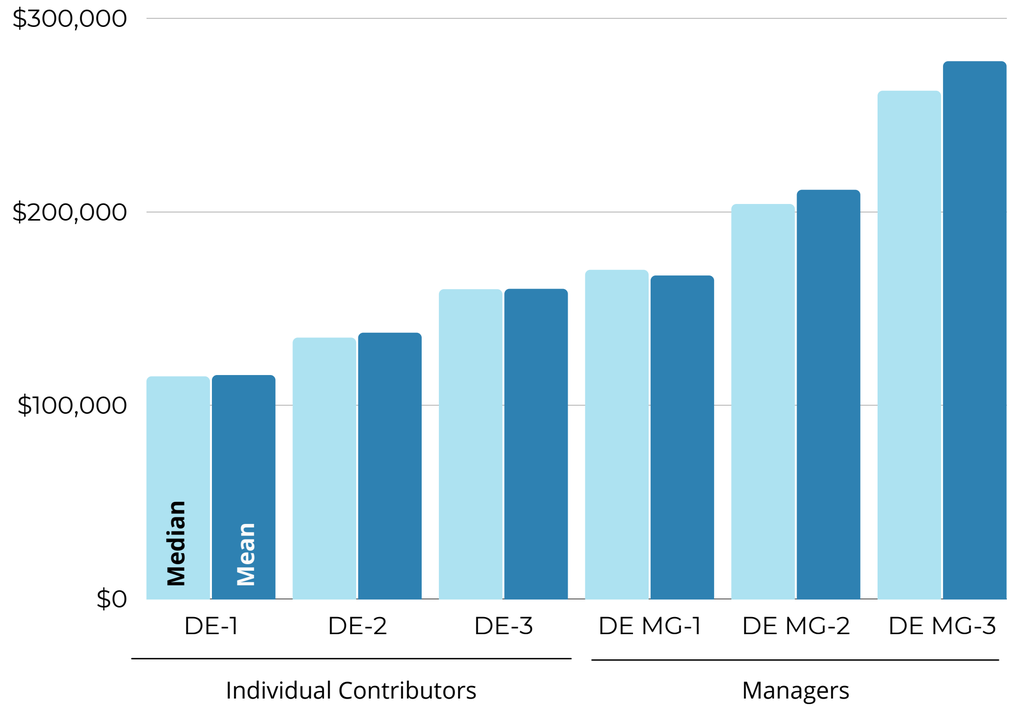


*See page 45 for job category definitions

- Salary remained strong, 2022 largest increase ever seen
- Concerns with tech sector shedding jobs, especially tough on entry level
- Most DS candidates have a MS
- Gender imbalance: only 25% female
- Job seekers desire WFH, companies are pushing back
- Current hot industries: Financial sector, Health

Compensation Changes Over Time

Salary Median and Means for Data Engineers - 2022



Burtch Works separated Data Engineers into six job levels based on their function.

DE IC 1: Early Career Professionals, generally under 3 years of professional experience

DE IC 2: Data Engineers with 4 to 8 years of experience, typically titles include Sr. Data Engineer or Lead Data Engineer

DE IC 3: Experienced Data Engineer, typically above 9+ years of experience, titles include Principal Data Engineer

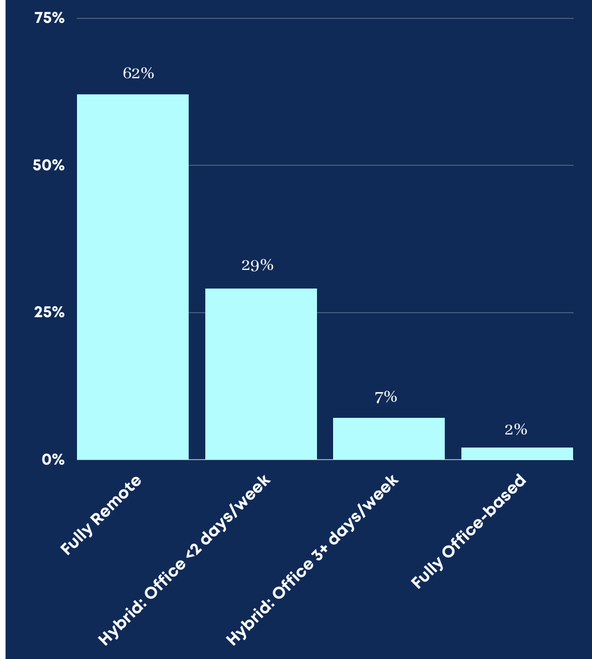
DE MG 1: Typically 0- 10 years of management experience, supervises a functional team under 3 direct reports, titles include Manager or Sr. Manager

DE MG 2: Leads function and executes on strategy, titles include Associate Director, Director, Sr. Director

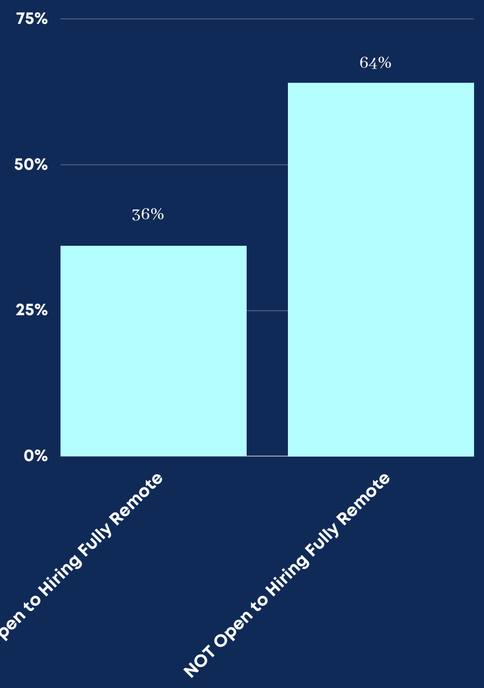
DE MG 3: Strategic leaders that are responsible for determining data strategy, titles include Vice Presidents, Heads of Data Engineering/Architecture

- Salary remained strong, 2022 largest increase ever seen
- Concerns with tech sector shedding jobs, especially tough on entry level
- Most DS candidates have a MS
- Gender imbalance: only 25% female
- Job seekers desire WFH, companies are pushing back
- Current hot industries: Financial sector, Health

As Teams Balance Employee Preference with Company Policy, **WFH Continues to Evolve**



CANDIDATES



CLIENTS

With many companies evaluating their WFH and remote work policies going forward, we sent out a survey earlier this year to gauge candidate and client WFH preferences.

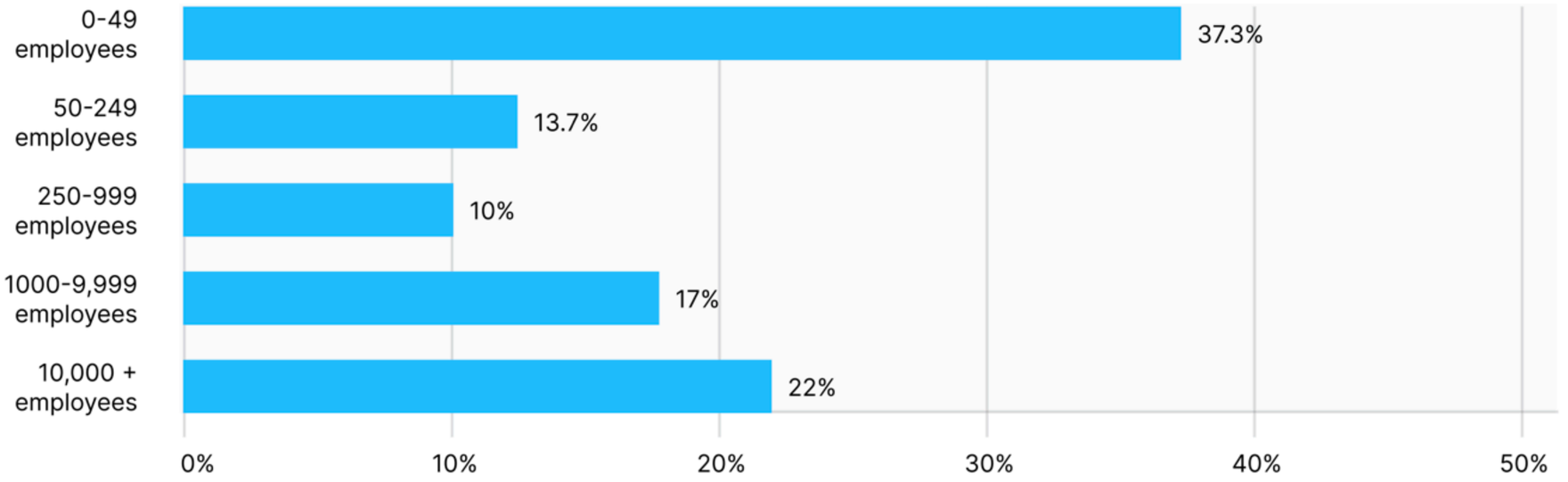
It is evident that fully remote positions are still heavily favored amongst candidates due to a increased flexibility. Contrary to that, company leaders like Jamie Dimon (JPMorgan Chase), Elon Musk (Tesla Motors) and even Howard Schultz (currently leading Starbucks) are pushing for corporate workers to come back to the office.

The idea that remote work has opened doors for individuals to work in cities across the country is shifting quickly towards more and more requests to relocate. **This evolution in policies is generally translating to a hybrid model where individuals are expected to come into the office on a partial or an as-needed basis.** With that said, there are countless roles and opportunities open to those that are seeking a fully remote position, but they are not the majority of roles available as often assumed or reported by the media.

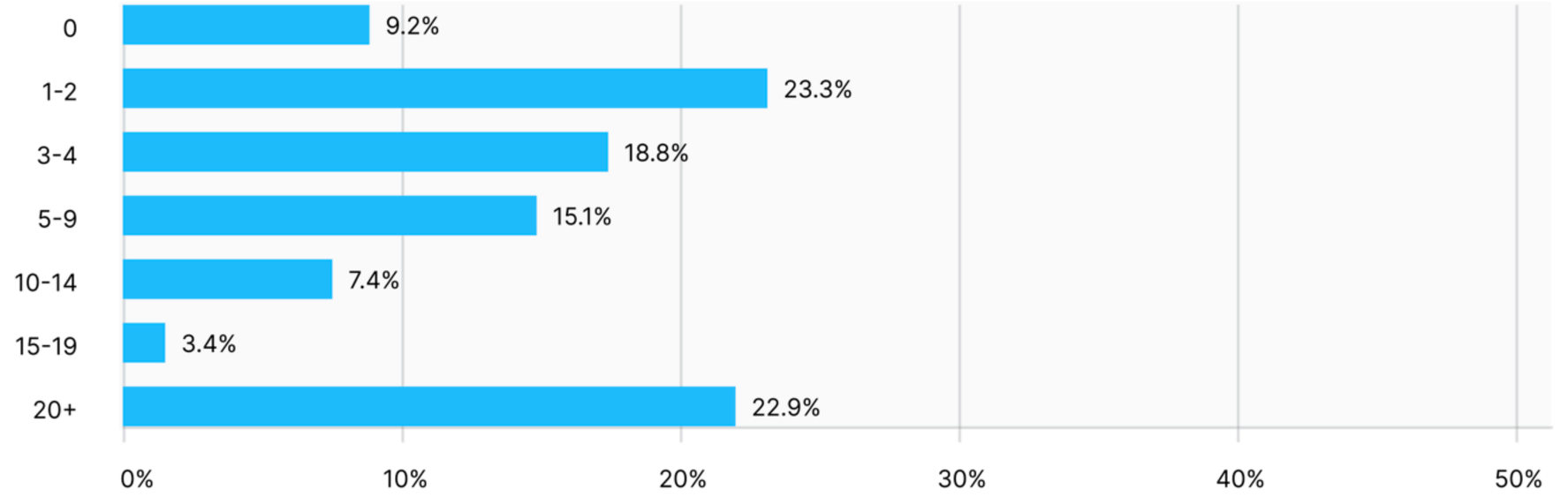
- Salary remained strong, 2022 largest increase ever seen
- Concerns with tech sector shedding jobs, especially tough on entry level
- Most DS candidates have a MS
- Gender imbalance: only 25% female
- Job seekers desire WFH, companies are pushing back
- Current hot industries: Financial sector, Health

Kaggle 2020 State of ML & DS

COMPANY SIZE (# OF EMPLOYEES)



DATA SCIENCE TEAMS (# OF EMPLOYEES)

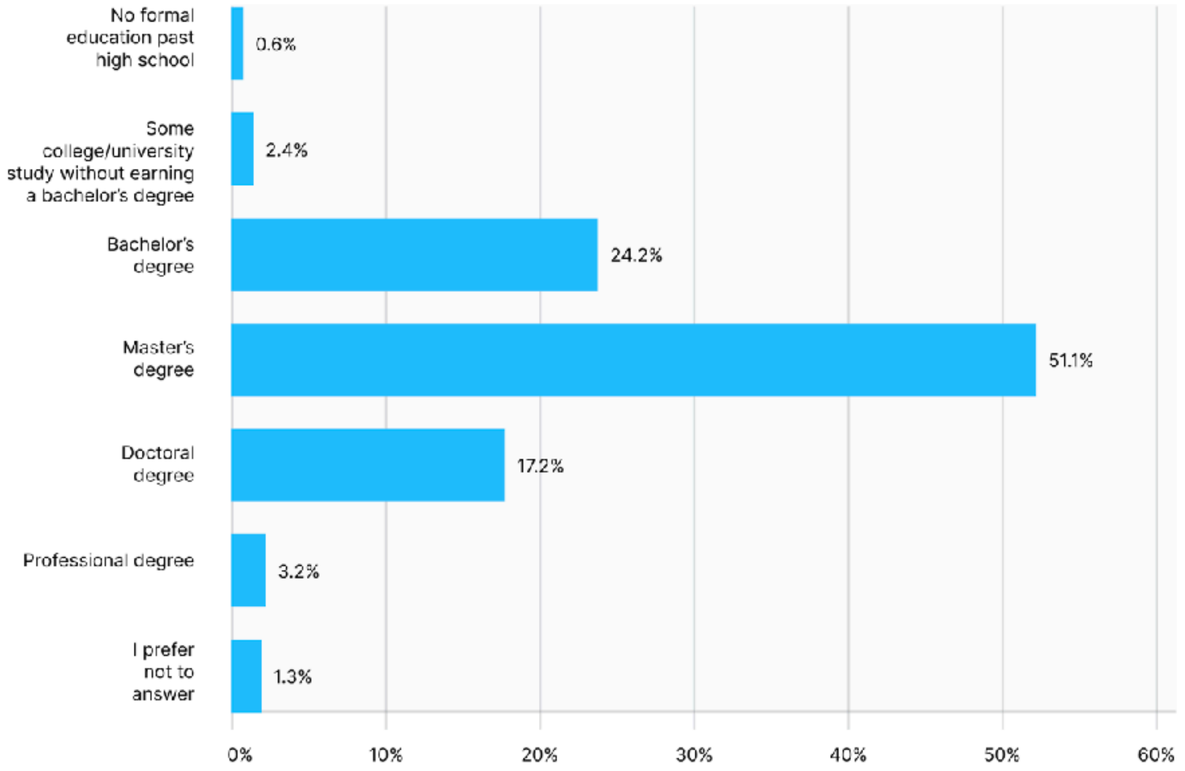


Ongoing Learning

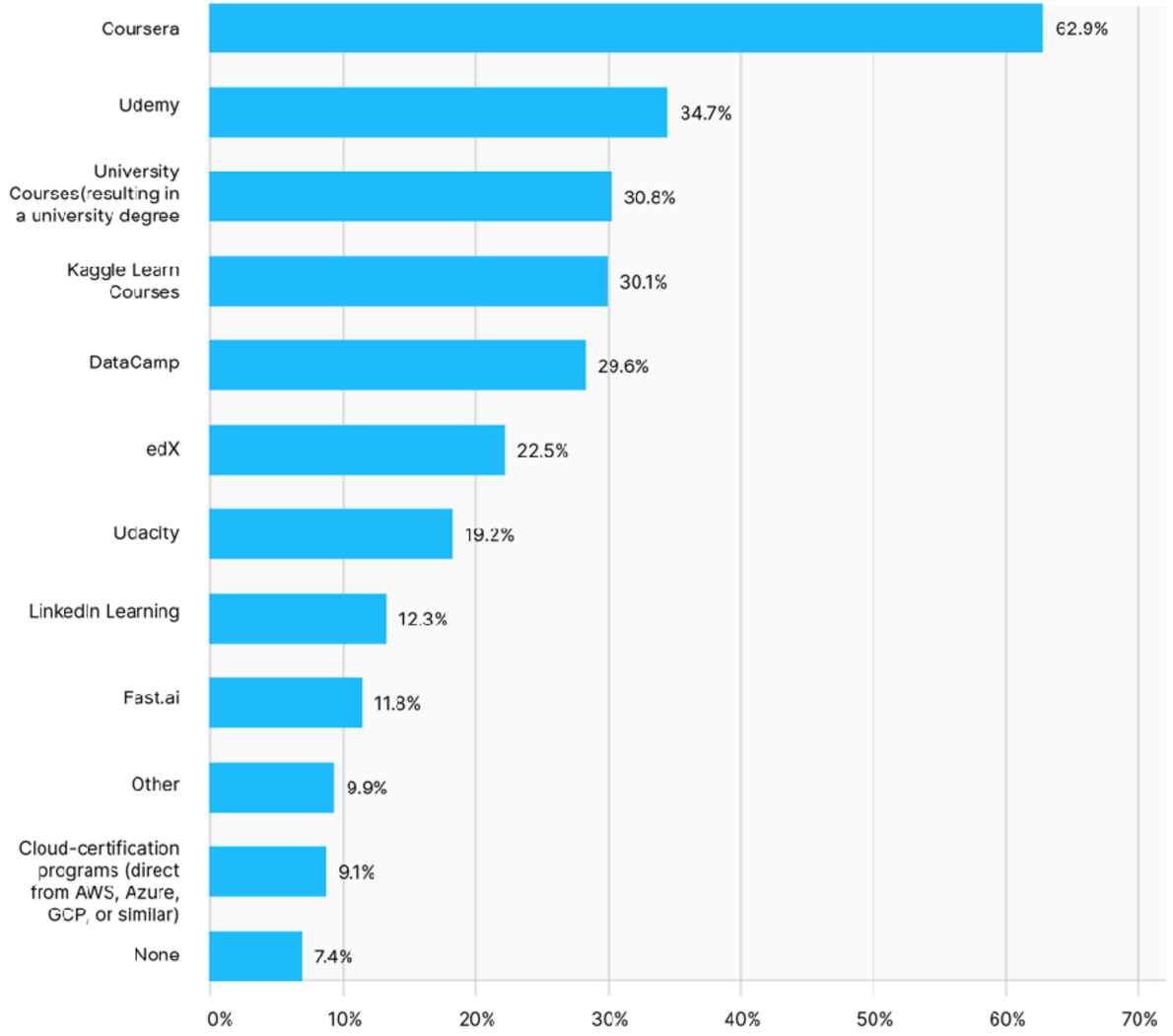
Data science and machine learning are quickly changing, so it's no surprise over 90% of Kaggle data scientists maintain ongoing education. While about 30% take traditional higher education courses, many more learn through online materials.

Coursera, Udemy, and Kaggle Learn top the most common mediums in our survey. Unsurprisingly, many Kaggle data scientists chose multiple resources in the survey, with an average of 2.8 mediums selected.

EDUCATION LEVEL OF KAGGLE DATA SCIENTISTS

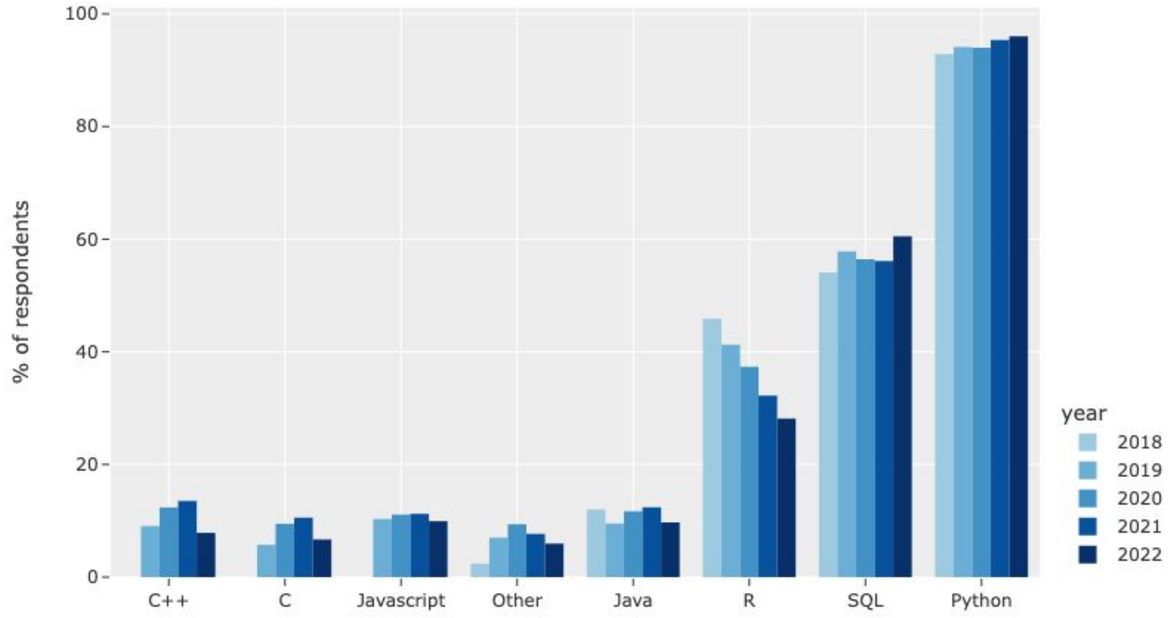


POPULAR ONGOING LEARNING RESOURCES



Kaggle DS & ML Survey 2022

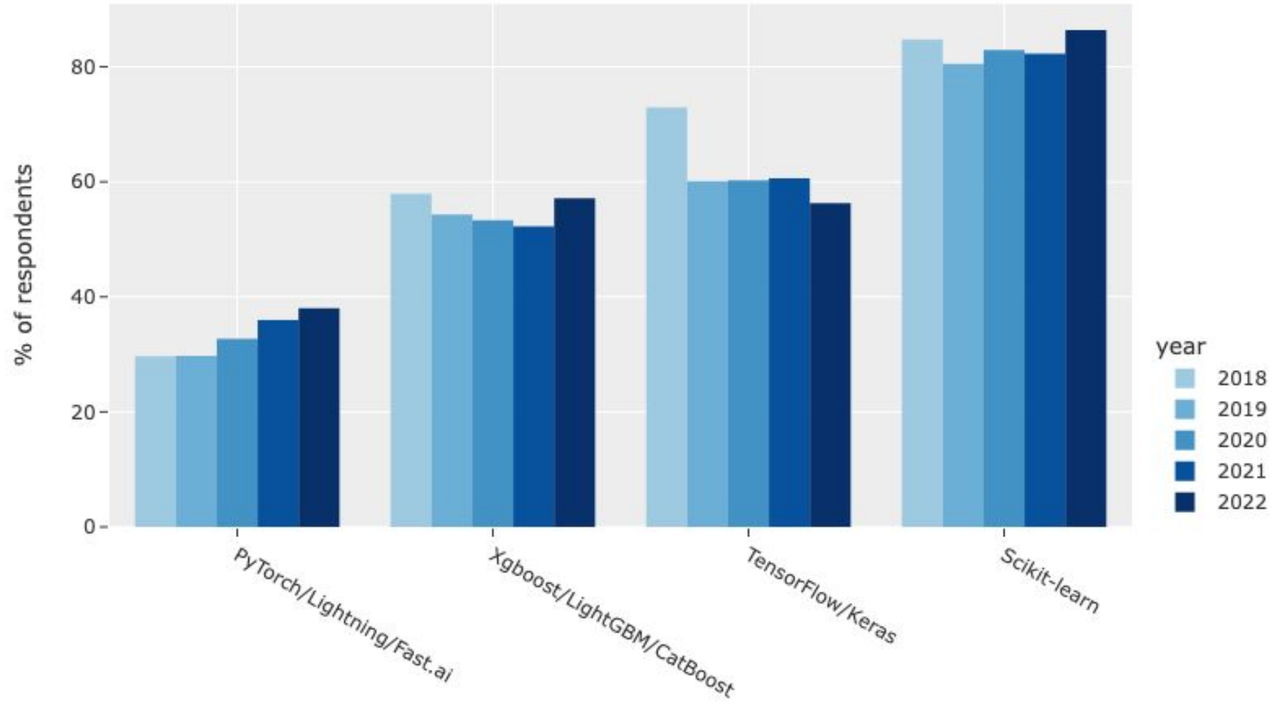
Python and SQL remain the two most common programming skills for data scientists



VSCode usage > 50%
Jupyter usage > 80%

Kaggle DS & ML Survey 2022

Scikit-learn is the most popular ML framework while PyTorch has been growing steadily year-over-year



Appen (aka Figure-Eight aka Crowdflower) State of AI report

Key Takeaways

1

SOURCING

Considered a challenging step of the AI lifecycle, data sourcing remains an obstacle.

42% of technologists say the data sourcing stage of the AI lifecycle is very challenging. However, business leaders were less likely to report data sourcing as very challenging (24%).

2

QUALITY

Business leaders and technologists report a gap in the ideal vs. reality of data accuracy.

More than half of respondents say data accuracy is critical to the success of AI, but only 6% reported achieving data accuracy higher than 90%.

3

EVALUATION

AI will not be replacing humans any time soon.

There's a strong consensus around the importance of human-in-the-loop machine learning with 81% stating it's very or extremely important and 97% reporting human-in-the-loop evaluation is important for accurate model performance.

4

ADOPTION

Perceptions regarding the prominence of AI in business may be shifting.

Technologists are split on whether their organization is ahead or even with others in their industry. Respondents in the US are more likely than their European counterparts to say their organizations are ahead of others in their industry at adopting AI.

5

ETHICS

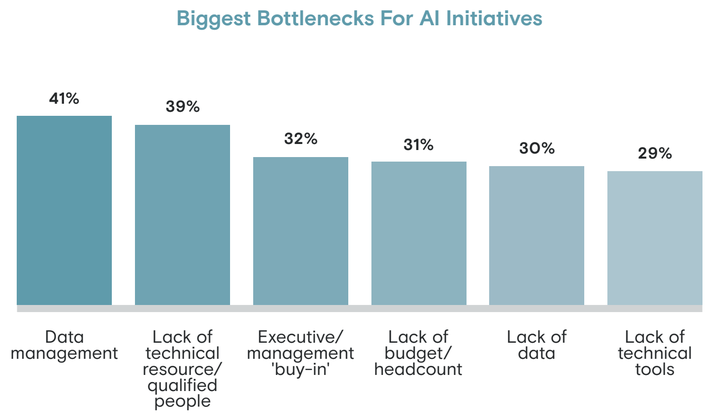
Responsible AI is the foundation of all AI projects.

93% of respondents agree that responsible AI is a foundation for all AI projects within their organization.

The greatest hurdle for AI initiatives is data management.

The greatest hurdle for AI initiatives is data management, with 41% indicating it as the biggest bottleneck. Right behind, 39% of respondents reported a lack of qualified talent--data scientists and technologists, data architects and engineers are scarce. 31% indicated a lack of budget for adequate headcount, adding to the challenge of properly staffing data management teams. This shortage of qualified data scientists and technologists emphasizes the importance of ensuring critical talent is focused on activities that require their valuable skills. To remedy this, companies look to external data providers to reduce their workload in areas such as data sourcing, freeing up scientists' time for other AI initiatives.

What do you consider the biggest bottleneck to any of your AI initiatives or projects?



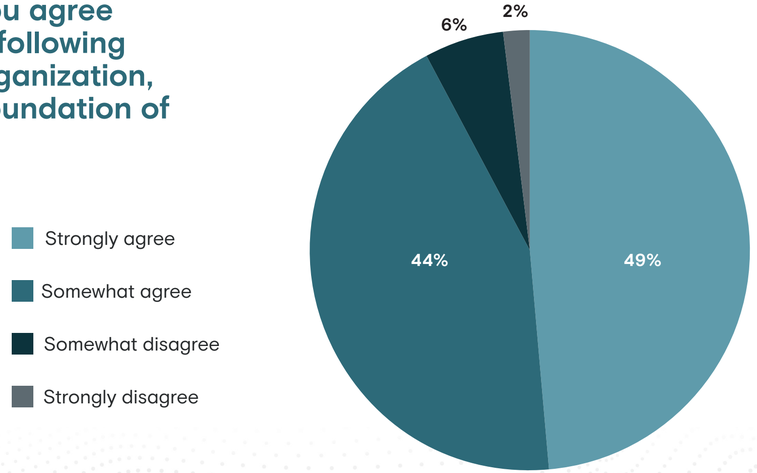
Ethics

One of the challenges in our industry is the perception that artificial intelligence poses ethical risks.

We are pleased to report that the large majority (93%) agree that responsible AI is a foundation for all AI projects within their organization. As diversity and inclusion become more prominent parts of mainstream AI and ML conversation, ethics at all stages of the AI lifecycle is more important than ever—especially regarding reducing bias and ethical data sourcing.

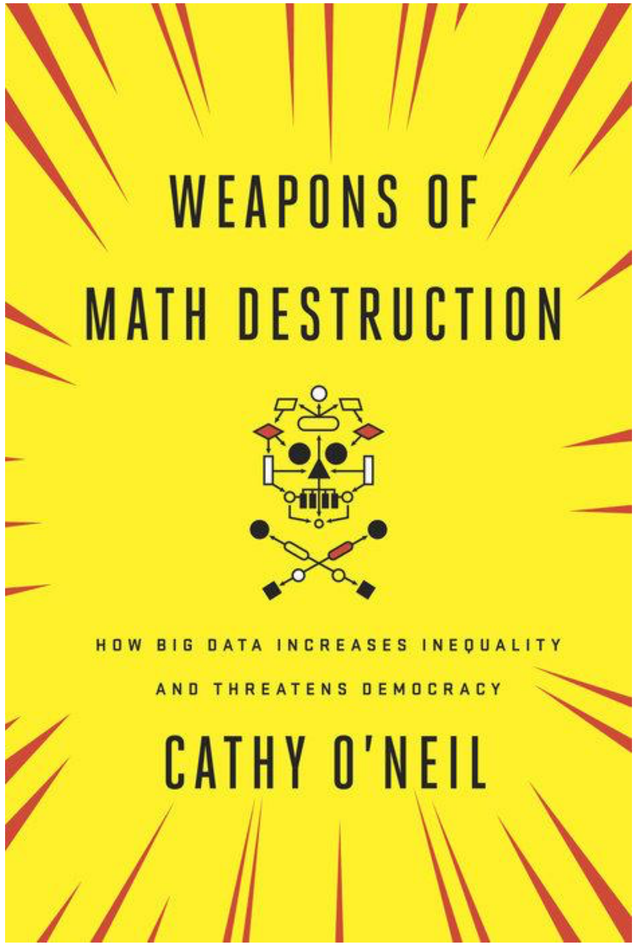
"Data ethics isn't just about doing the right thing, it's about maintaining the trust and safety of everyone along the value chain from contributor to consumer."
Erik Vogt
VP of Enterprise Solutions - Appen

To what extent do you agree or disagree with the following statement? At my organization, responsible AI is a foundation of all AI projects.

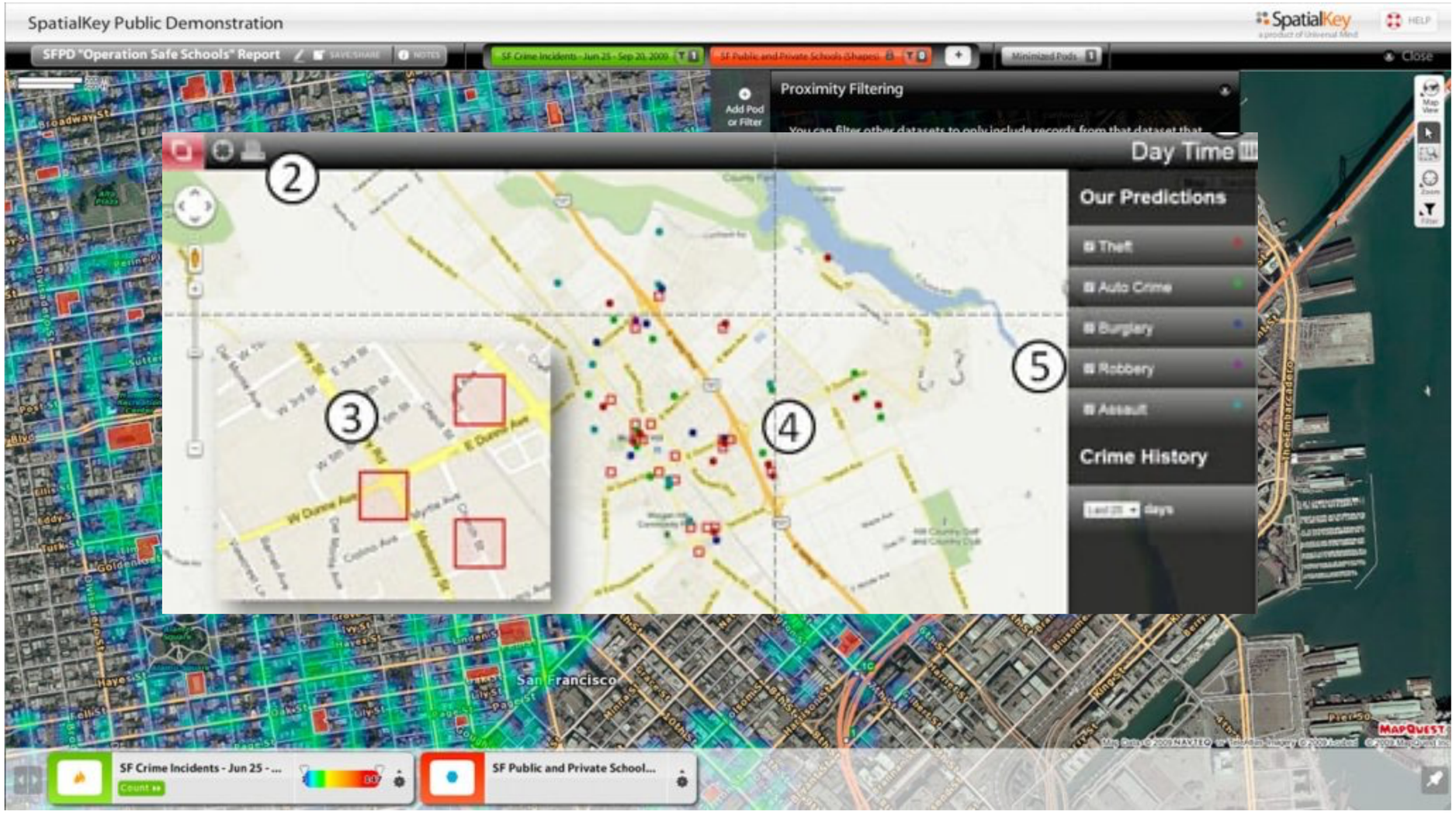


Don’t be a tool for creating WMDs

- Algorithms (and DS!) implement our biases, yet look objective
- Can implement our biases at scale
- Can have huge impacts on people’s lives
- Are not transparent or accountable to the people being impacted



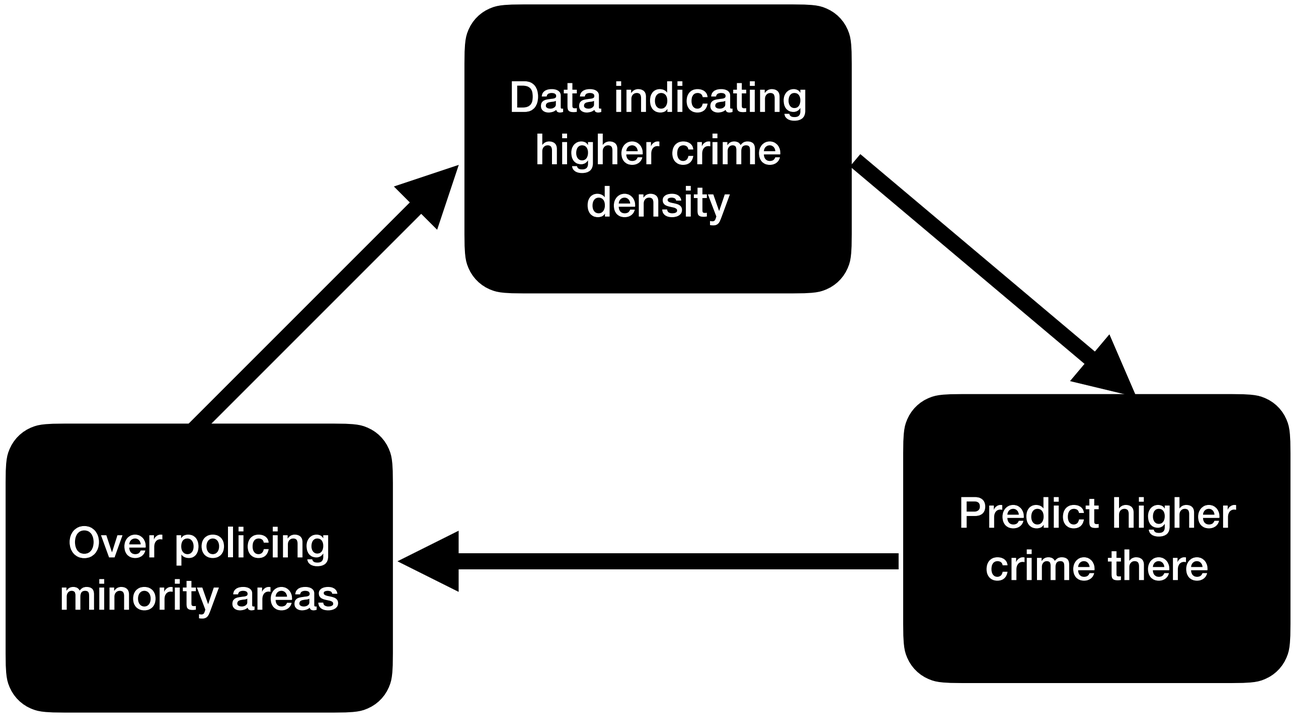
Predictive policing & sentencing



https://americanpoliceofficersalliance.com/police-tech-predictive-technology/#prettyPhoto

Predictive policing & sentencing

Blacks arrested for possession at 4x the rate of whites
Usage rates the same



https://norml.org/marijuana/fact-sheets/racial-disparity-in-marijuana-arrests/

https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x

https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them

“A lot of times, people are talking about bias in the sense of equalizing performance across groups. They’re not thinking about the underlying foundation, whether a task should exist in the first place, who creates it, who will deploy it on which population, who owns the data, and how is it used?”

-Timnit Gebru

You all are the future of data science!

So, if you remember anything from this course...

Ethics should always be a priority in your work.

Data wrangling is a puzzle and a big part of the job. When done well, it's not boring!

Data science is a competitive, but rewarding field. You have a chance to make a big difference!

Your grade in this course is probably *not* predictive of future success.

My hope is that all of you stay:

- happy & balanced in your life
- good people who think about how to make the world better, especially about how the systems around us constrain us
- interested, curious, and engaged with the world

and that you go on to find success and fulfillment!

Thank you!

Teaching Assistants:

- **Ruby Ying**
- **Fuling Sun**
- **Heeket Mehta**
- **Shanay Shah**

Instructional Assistants:

- **Cindy Wang**
- **Nathaniel Mackler**
- **Jack Zhao**

And thanks to YOU!