Course Reminders

- Final Project due Wed, June 9th (11:59 PM)
  - Report (GitHub)
  - Video (put on cloud service of choice, MAKE VIEWABLE, link in final report)
  - Team Evaluation Survey: https://forms.gle/XjZfAvFDi34kbuhg6 (link also on Canvas; required)
- Post COGS 108 Survey: https://forms.gle/MymGCmcQdJsNKbxX6 (link also on Canvas; *optional* for EC)
- CAPEs: http://www.cape.ucsd.edu/ (~45%, EC>75%!)

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

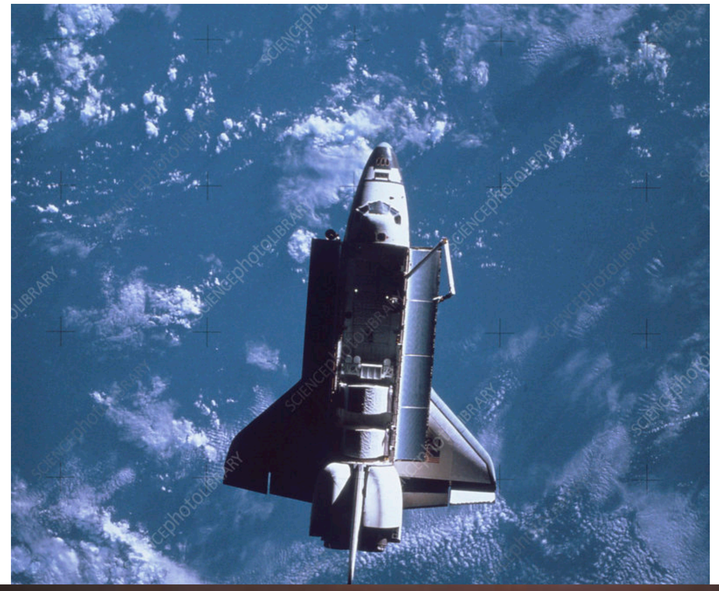
Errors of analysis - did we use the right methods to address the question?

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

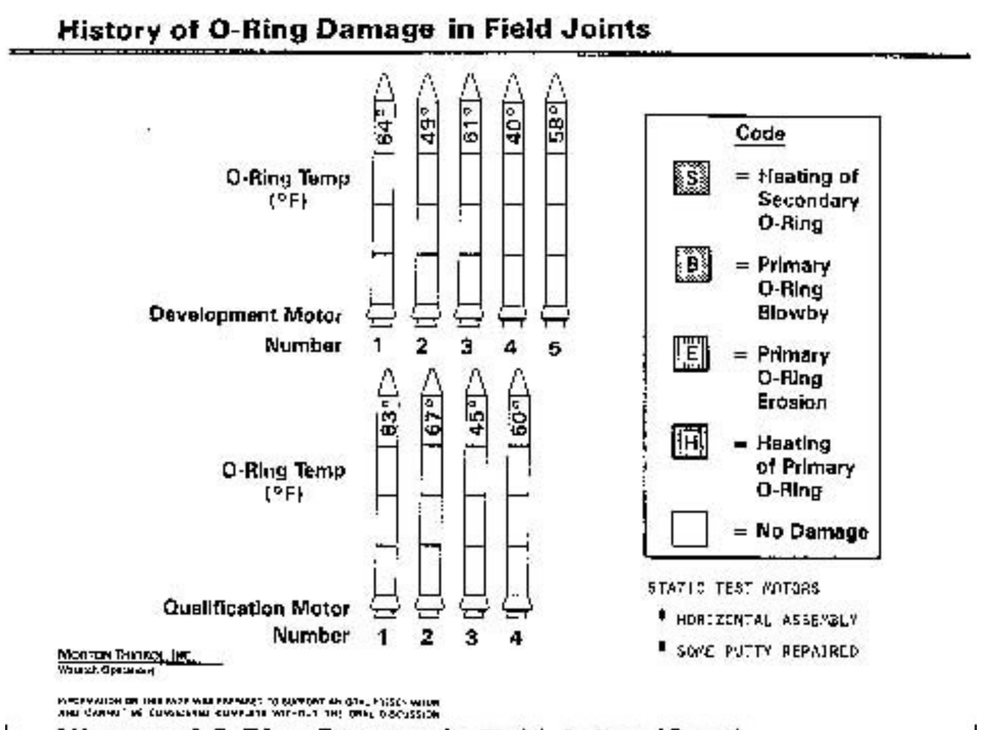
Errors of communication - sometimes you get everything right, but the group and the decision makers never understand properly

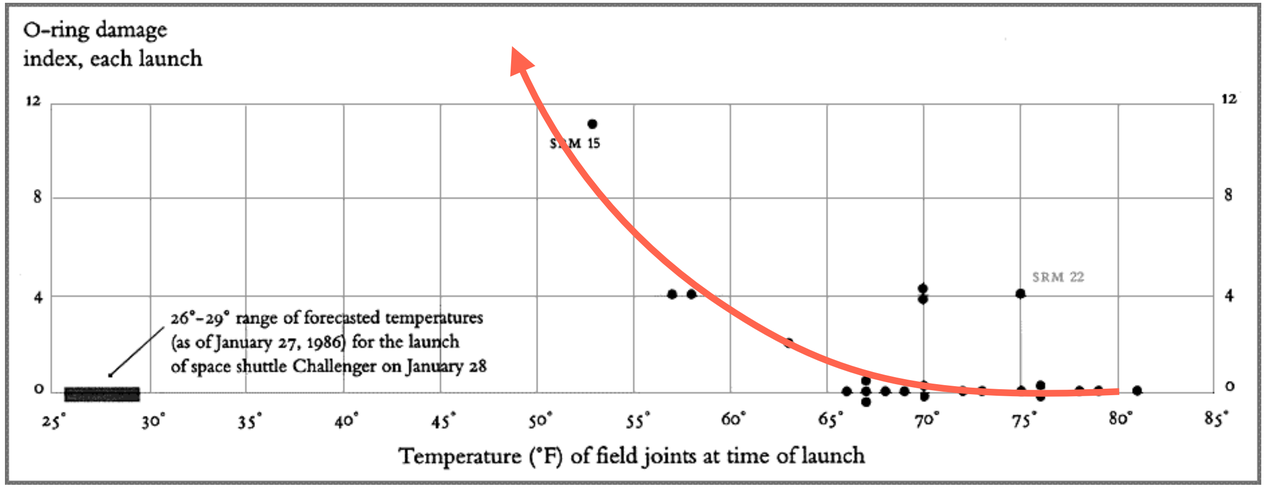
Errors of communication



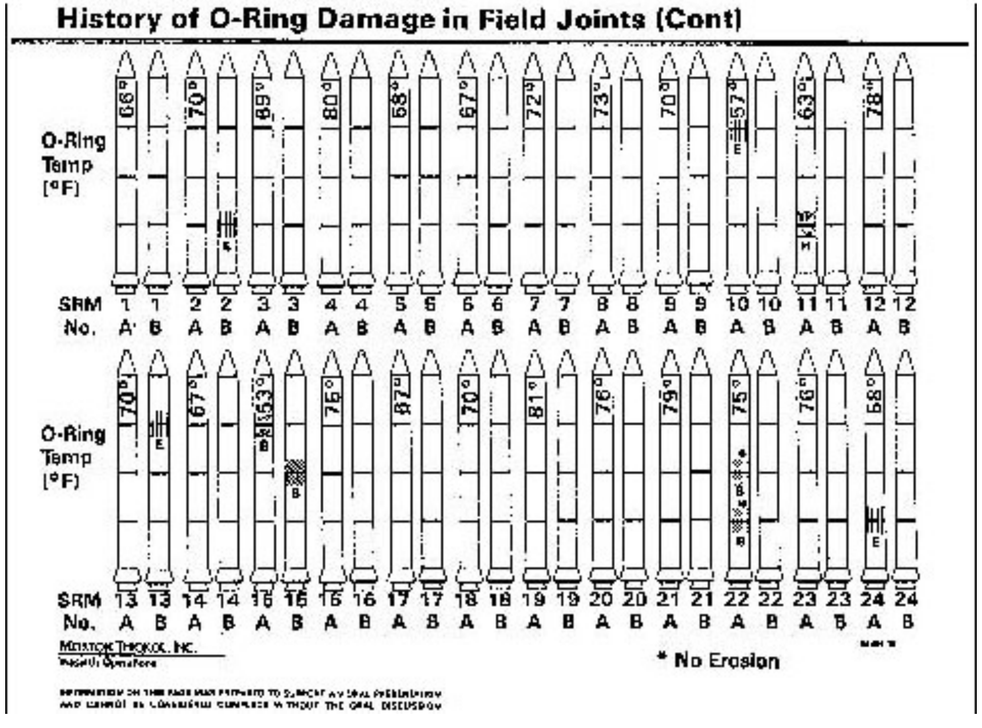
Jan 28, 1986

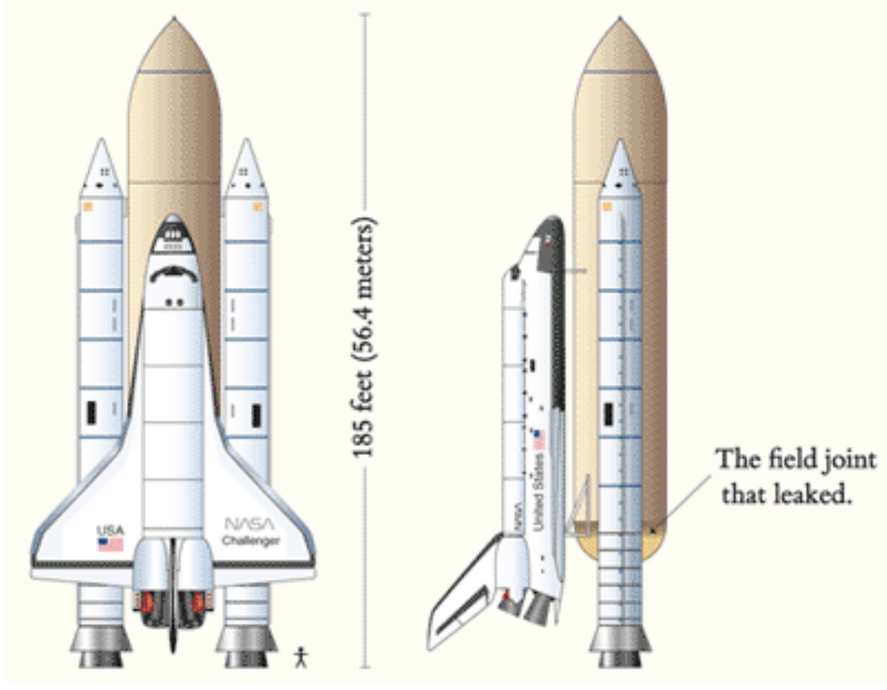
Feb 1, 2003





Images and graphs from Edward Tufte





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.

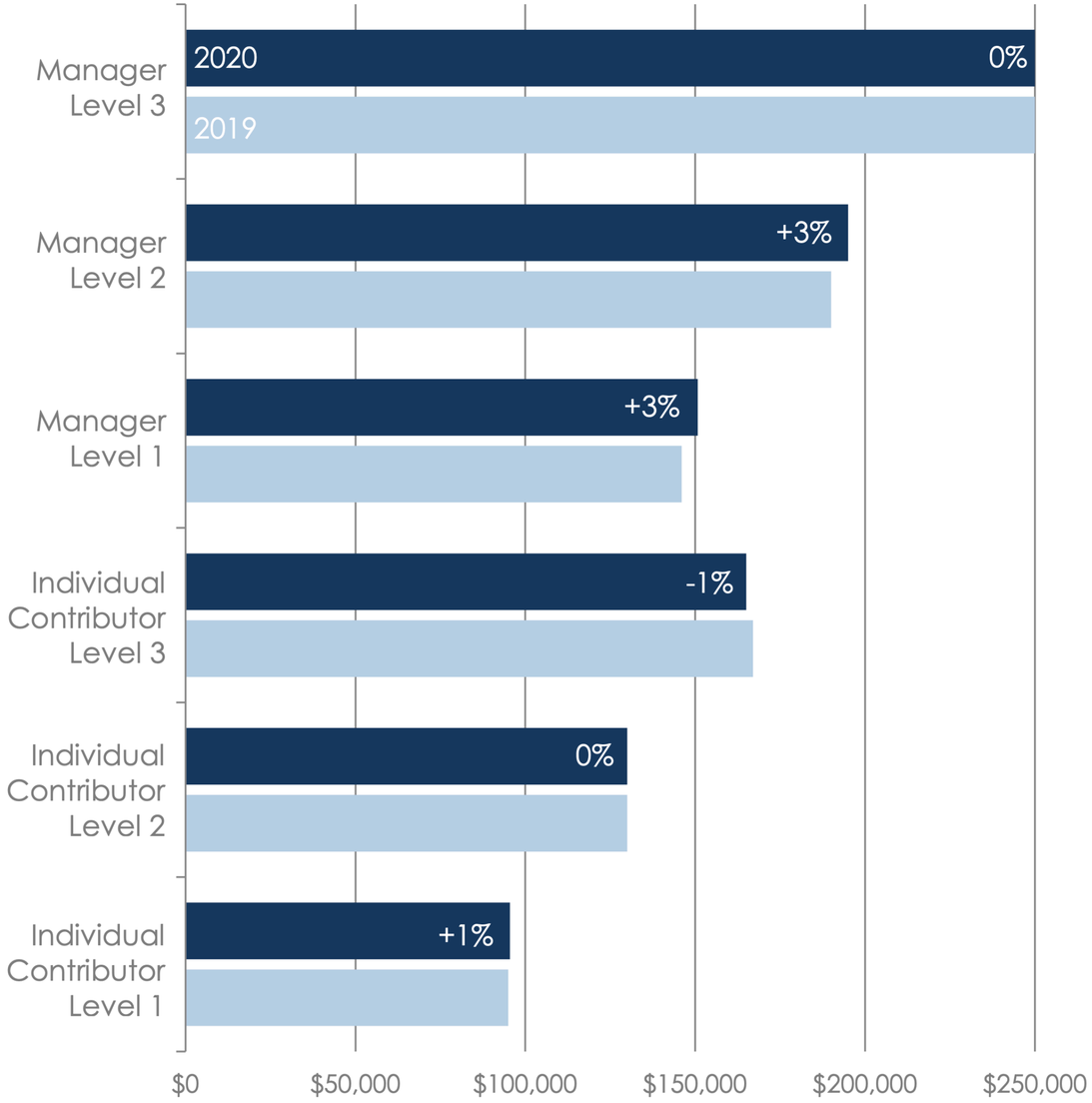
https://veekaybee.github.io/2019/02/13/data-science-is-different/

	Job Title	Median Base Salary	Job Satisfaction	Job Openings
#1	Java Developer	$90,830	4.2/5	10,103
#2	Data Scientist	$113,736	4.1/5	5,971
#3	Product Manager	$121,107	3.9/5	14,515

Source: Glassdoor (2021)

Burtchworks annual predictions and report on DS hiring

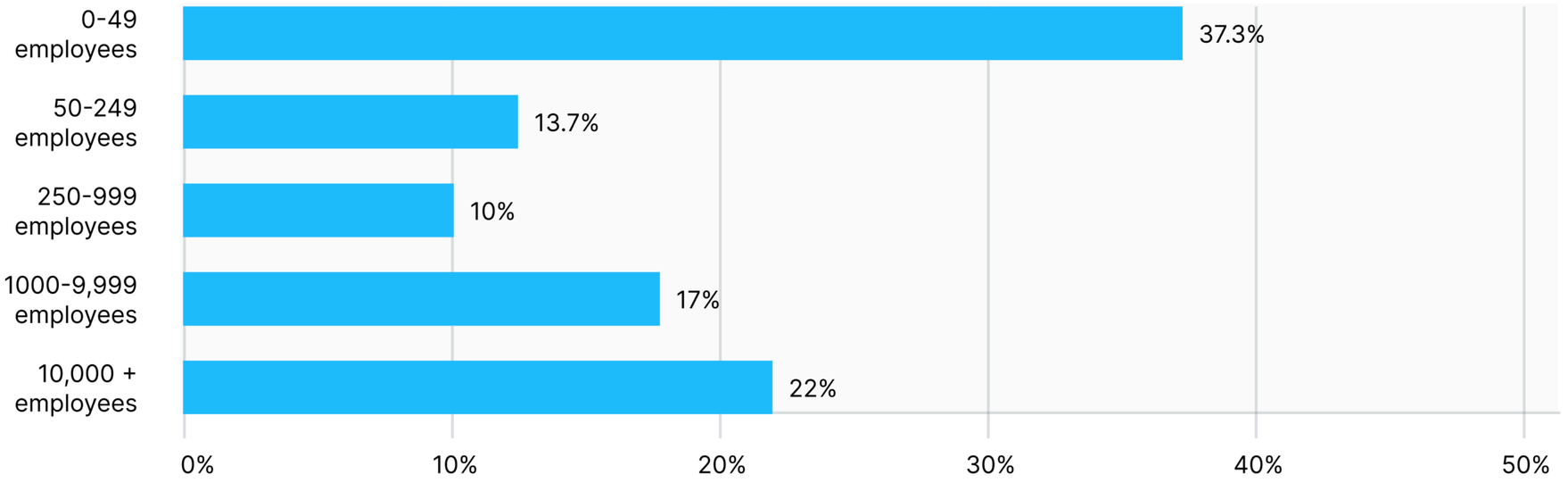
Figure 2 Comparison of **Data Scientists'** Median Base Salaries by Job Category



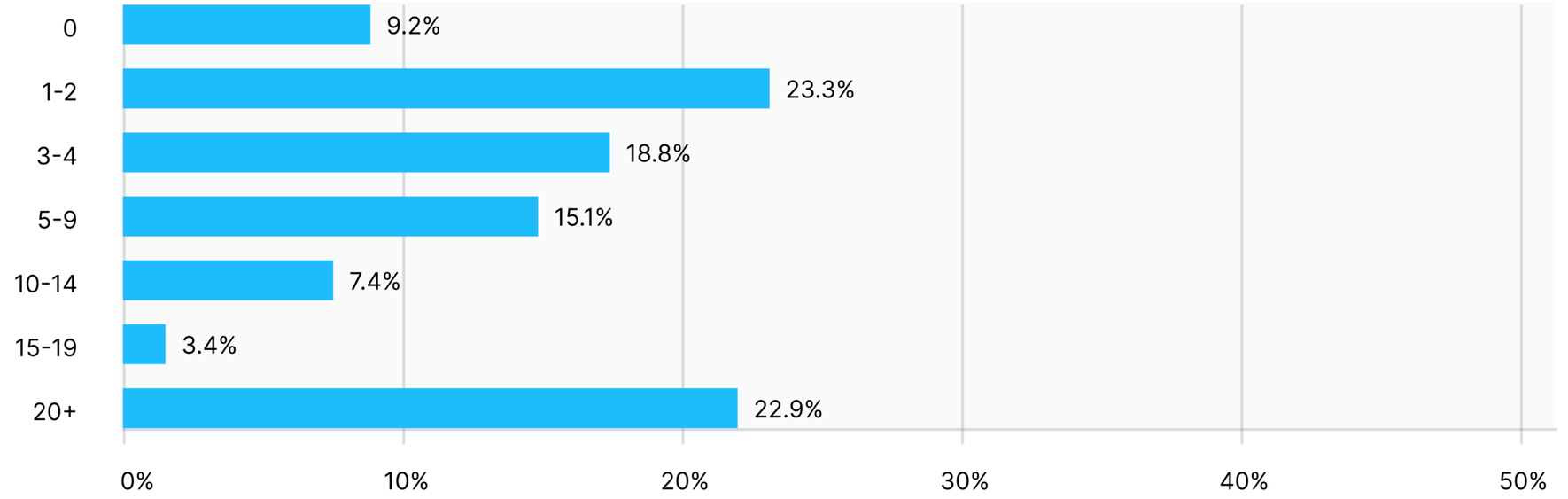
- Salary remained strong, pandemic may change that
- Concerns that supply + demand for DS may be narrowing at entry level
- WFH becomes normal, people move out of hot cities
- Current hot industries: Health, Supply chain

Kaggle 2020 State of ML & DS

COMPANY SIZE (# OF EMPLOYEES)



DATA SCIENCE TEAMS (# OF EMPLOYEES)

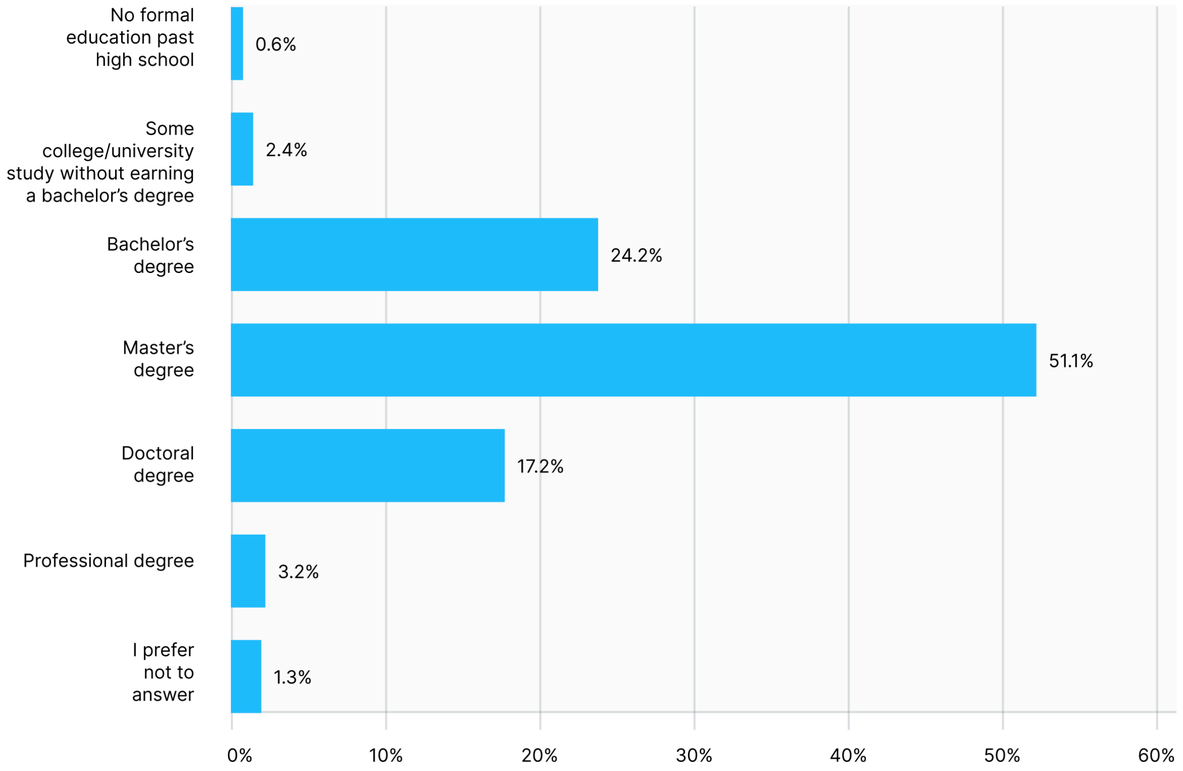


Ongoing Learning

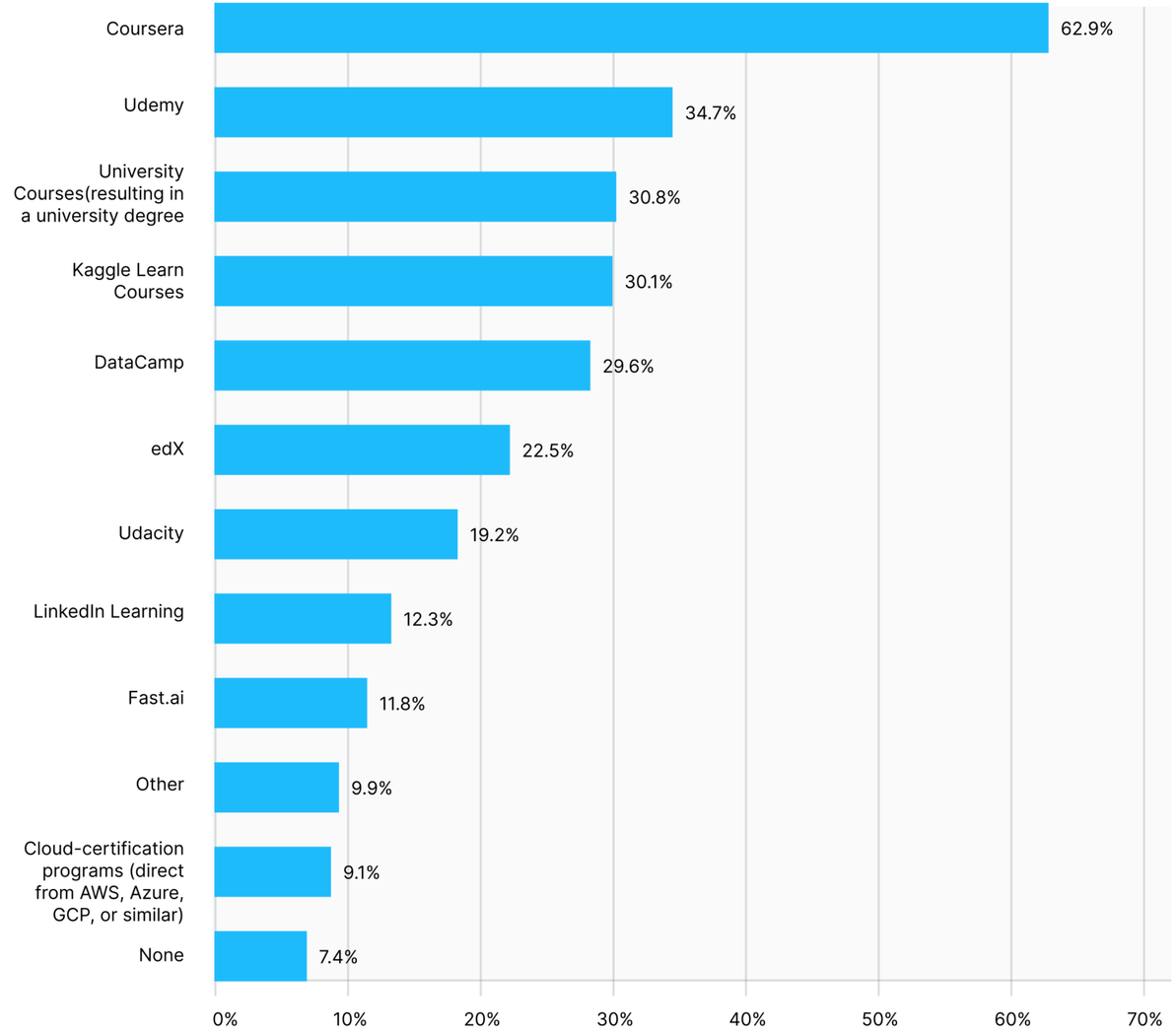
Data science and machine learning are quickly changing, so it's no surprise over 90% of Kaggle data scientists maintain ongoing education. While about 30% take traditional higher education courses, many more learn through online materials.

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

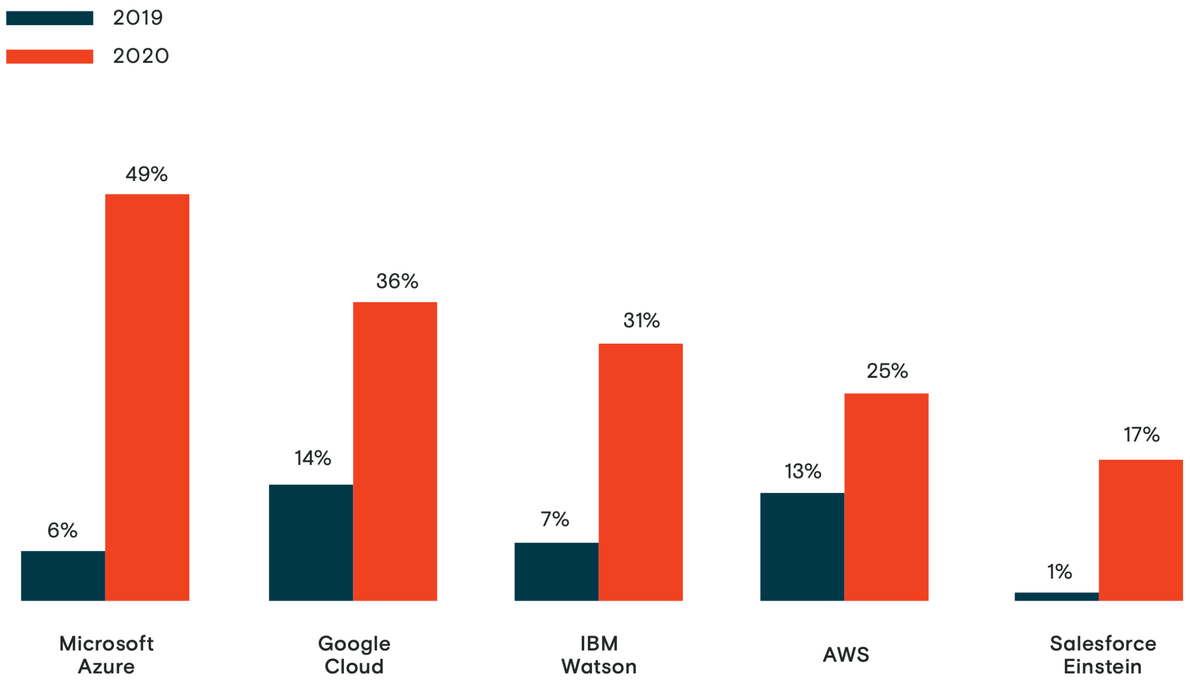


POPULAR ONGOING LEARNING RESOURCES



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

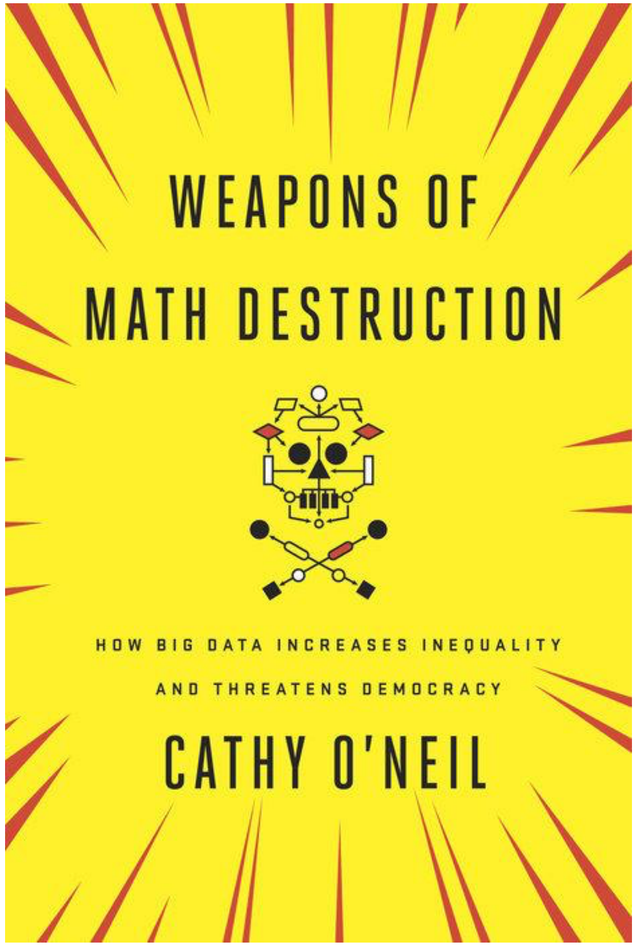
Figure 7: What data science and machine learning tools/frameworks do you use?



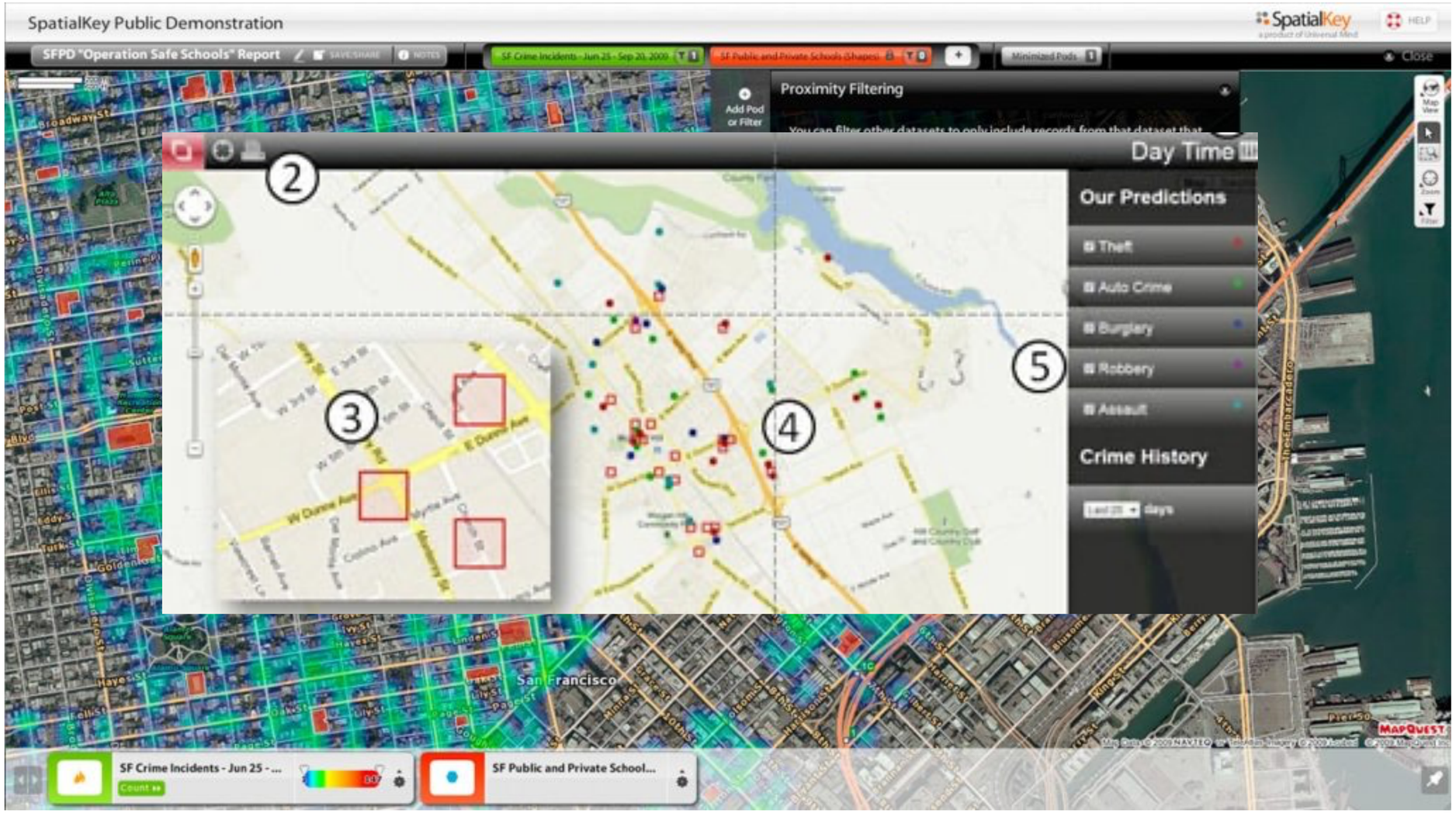
- AI/ML is a big deal at corporate level
- Everything is moving to the cloud (accelerating from WFH?)
- Lots of love for Azure, slowing growth for AWS

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



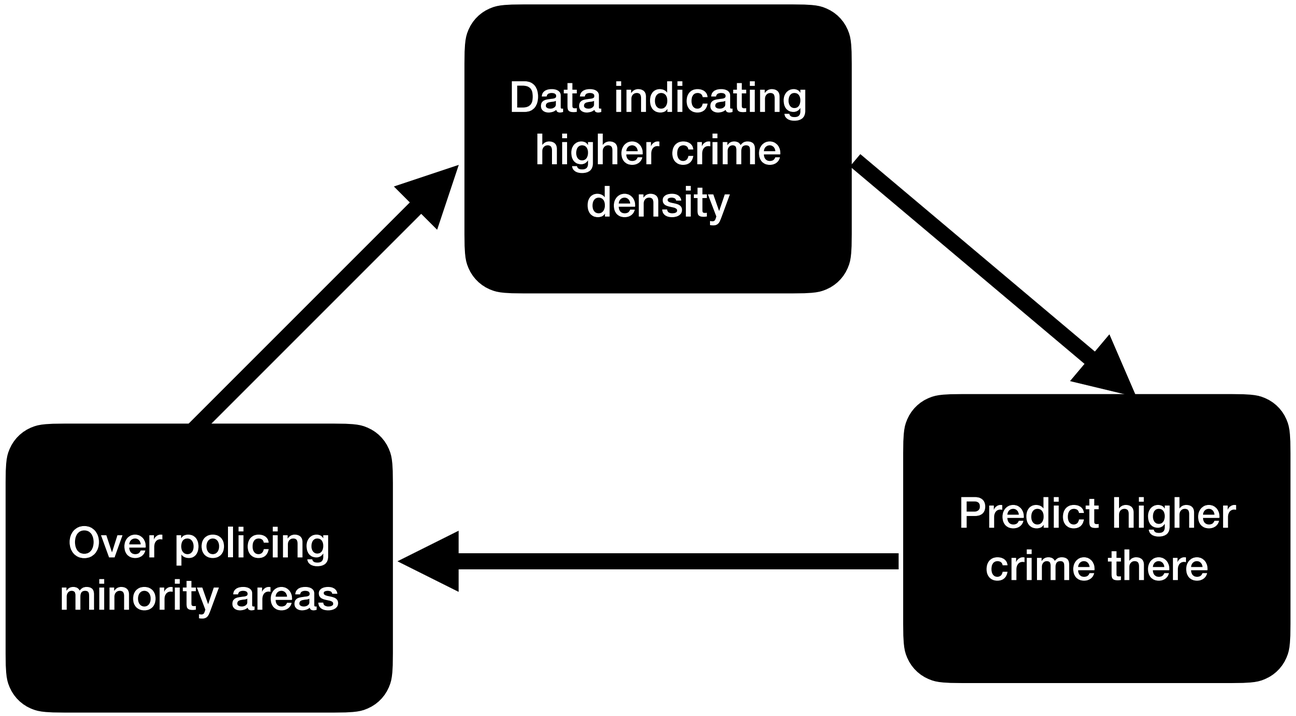
Predictive policing & sentencing



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 go on to (continue to) be good people who are happy & successful

Thank you!

Teaching Assistants:

- **Pooja Pathak**
- **Areeb Syed**
- **Stephen Jarrell**
- **Matthew Fiegelis**

Instructional Assistants:

- **Ruoxuan Li**
- **Scott Yang**
- **Viki Zhao**
- **Sahithi Chimmula**
- **David Bian**
- **Zhigang Lin**

And thanks to YOU!