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

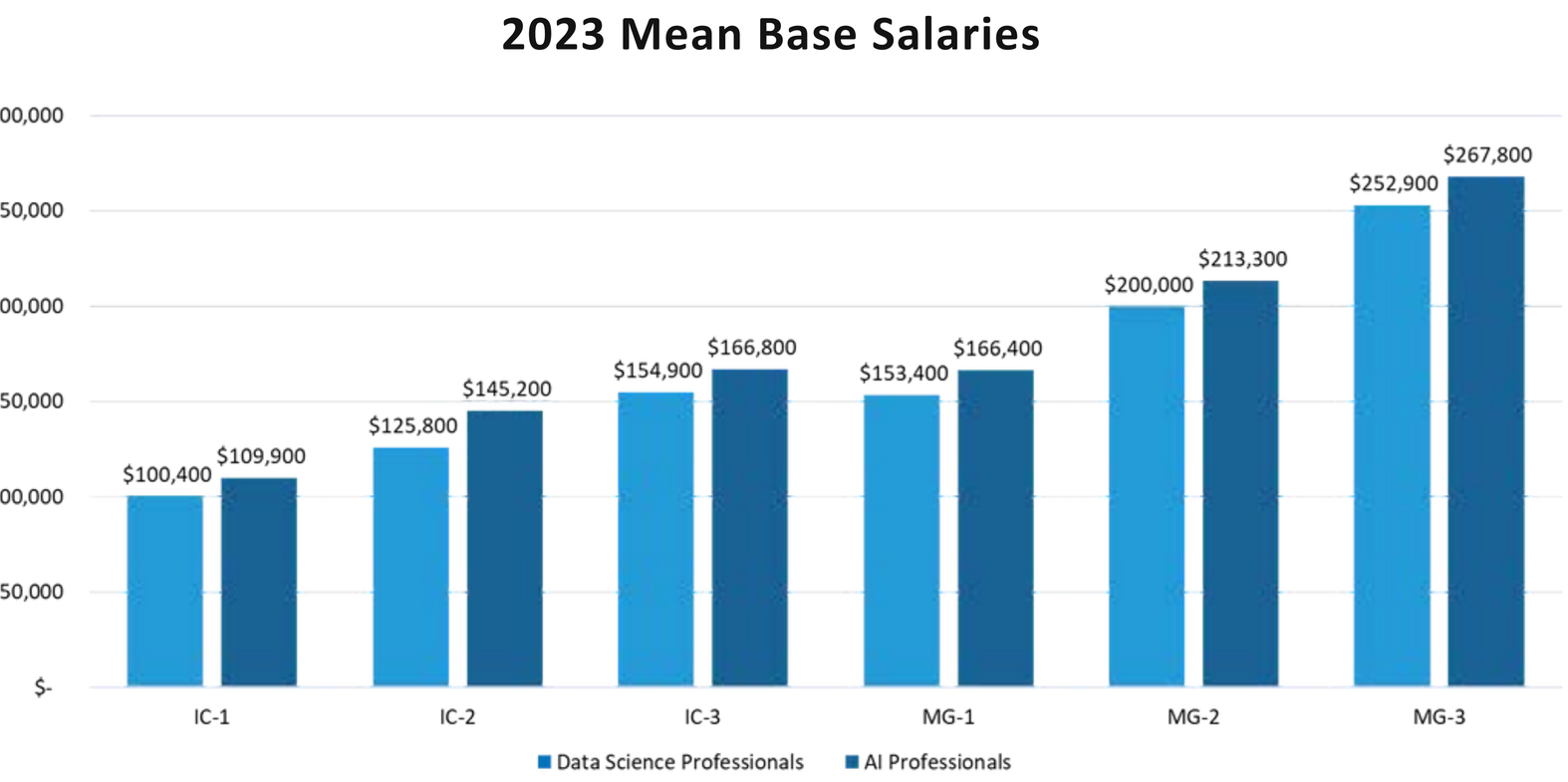
What should you learn next?

- Data science?
  - Statistics + Modeling
  - SQL and/or Data viz packages
  - Cloud
  - Become fantastic at one programming language and learn many packages
- AI?
  - Statistics, Vector calculus, ML theory
  - Cloud
  - Foundation models
  - Become usefully good at several frameworks (PyTorch, Go, Tensorflow, etc)
  - Read arXiv and NeurIPS/ICLR papers!

	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)

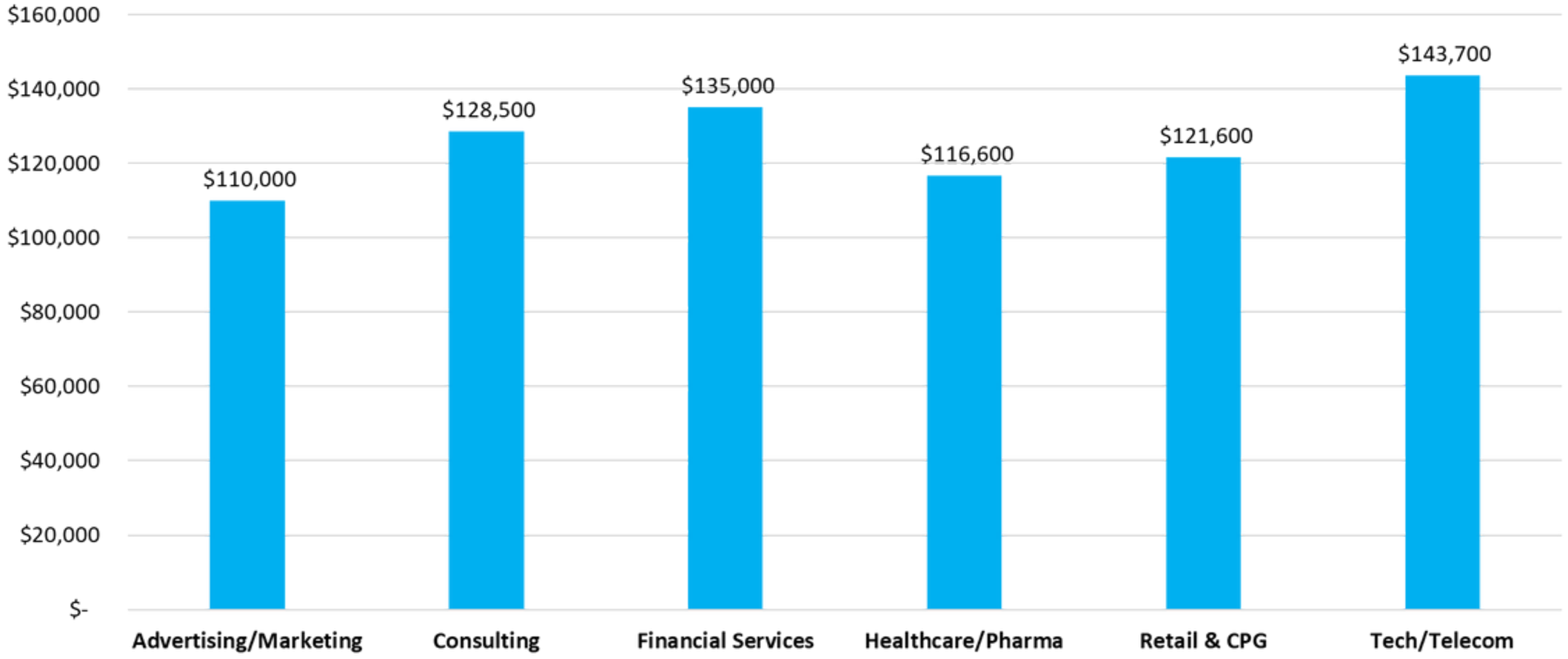
Burtchworks annual predictions and report on DS hiring



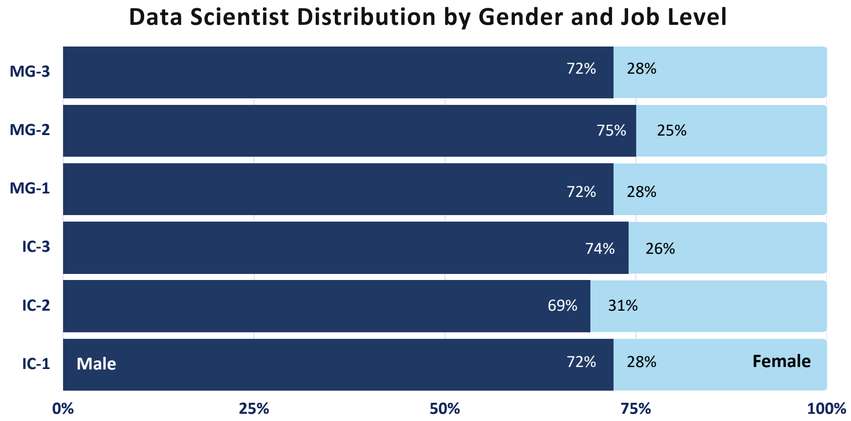
Consistent with our findings from previous reports, it appears that AI professionals may earn more base salary when compared to data science professionals across all job levels.

- Great resignation -> ’22 had huge salary increases which continue but slower now
- Market contraction in ‘23, hiring low but picking back up now
- 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

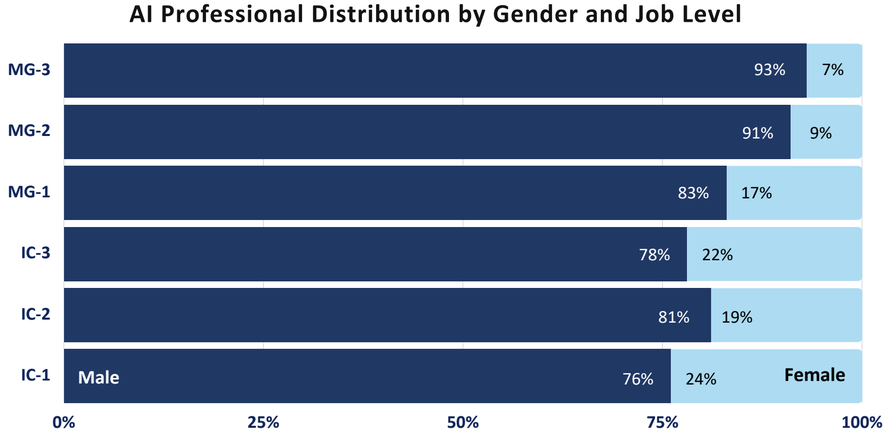
Comparison of Data Science IC-2 Mean Base Salaries Across Various Industries



DISTRIBUTION OF DATA SCIENTISTS & AI PROFESSIONALS BY GENDER AND JOB LEVEL



Similarly to previous years, the proportion of females among Data Scientists and AI Professionals is highest at the junior levels. Women become less prevalent among high-level individual contributors and senior managers.

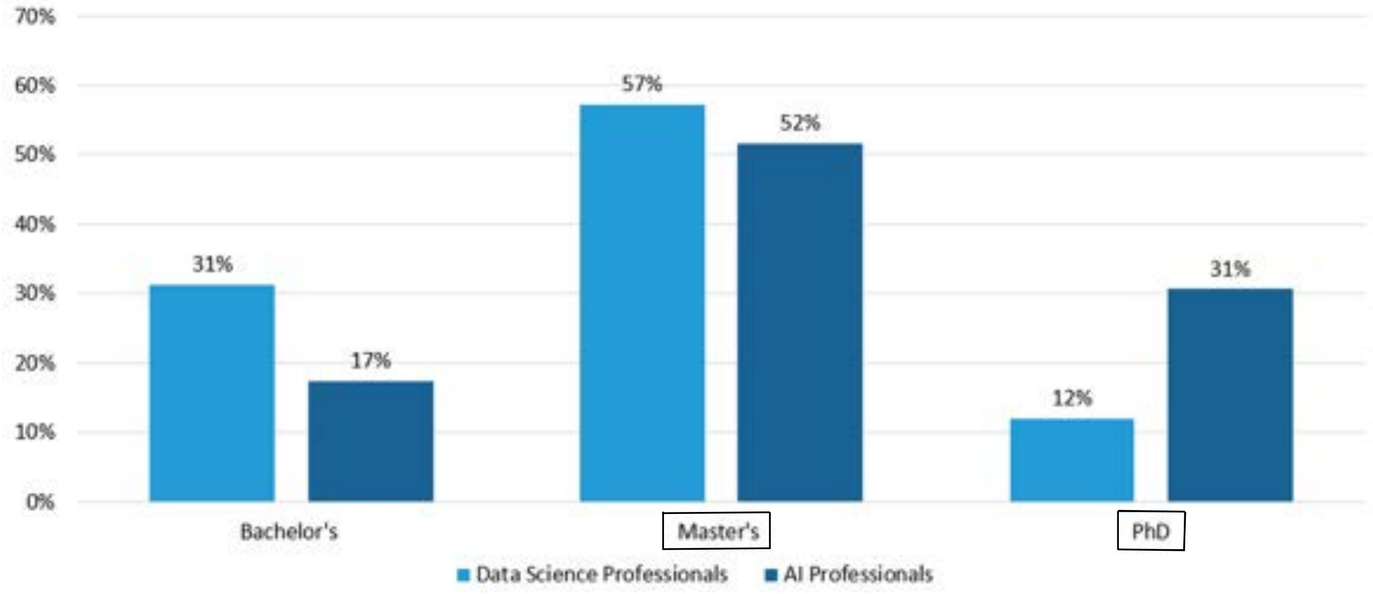


EDUCATION: COMPARISON OF DEGREE LEVEL

- ***72% of all data scientists and AI professionals surveyed held an advanced degree.***
- *Education level has historically had a marked effect on salary.*
- *The proportion of AI Professionals with a Master's and/or PhD as their highest degree earned is higher than data scientists and is a statistically significant difference.*

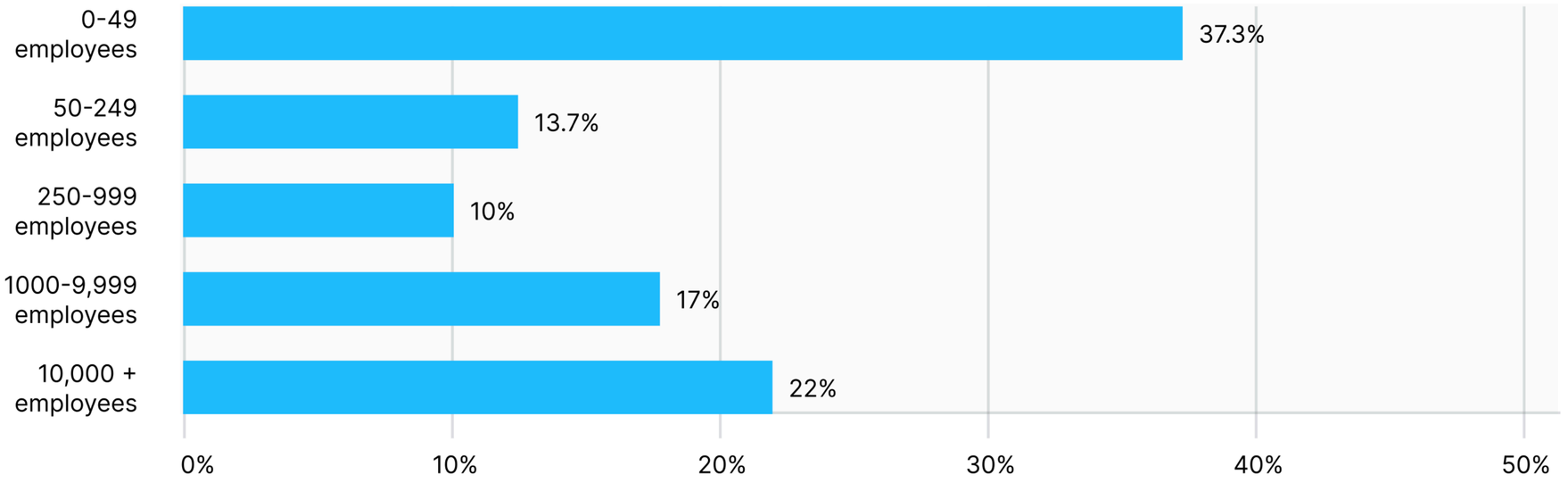
COMPARISON OF DEGREE LEVEL *(for highest degree earned)*

Data Scientists vs. AI Professionals

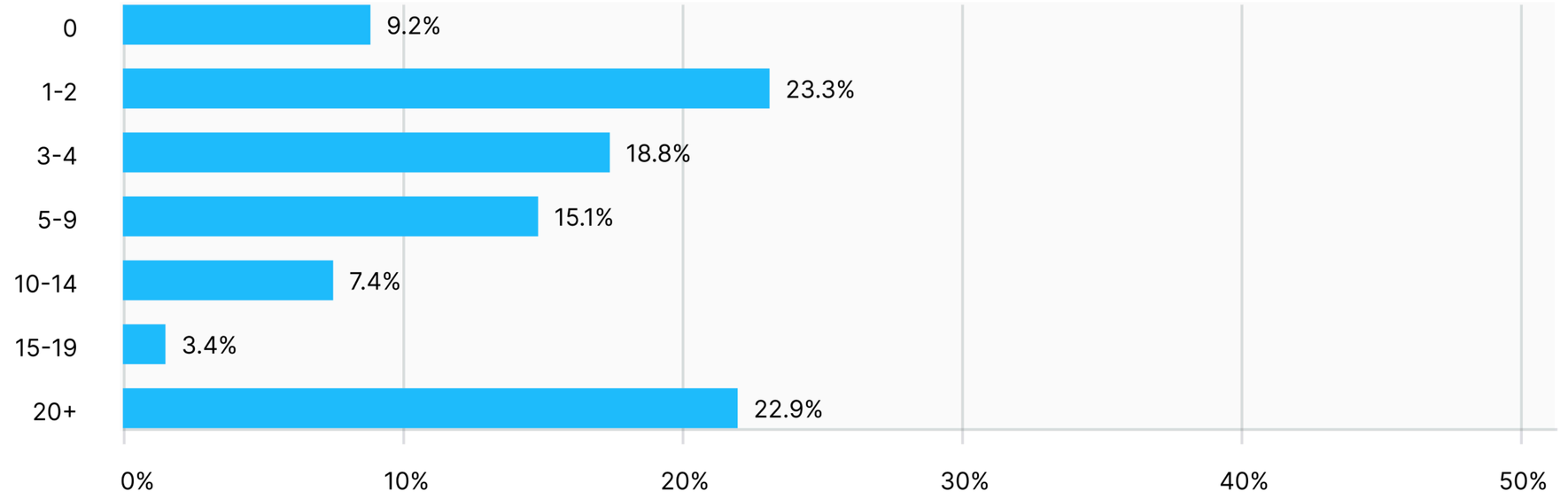


Kaggle 2020 State of ML & DS

COMPANY SIZE (# OF EMPLOYEES)



DATA SCIENCE TEAMS (# OF EMPLOYEES)

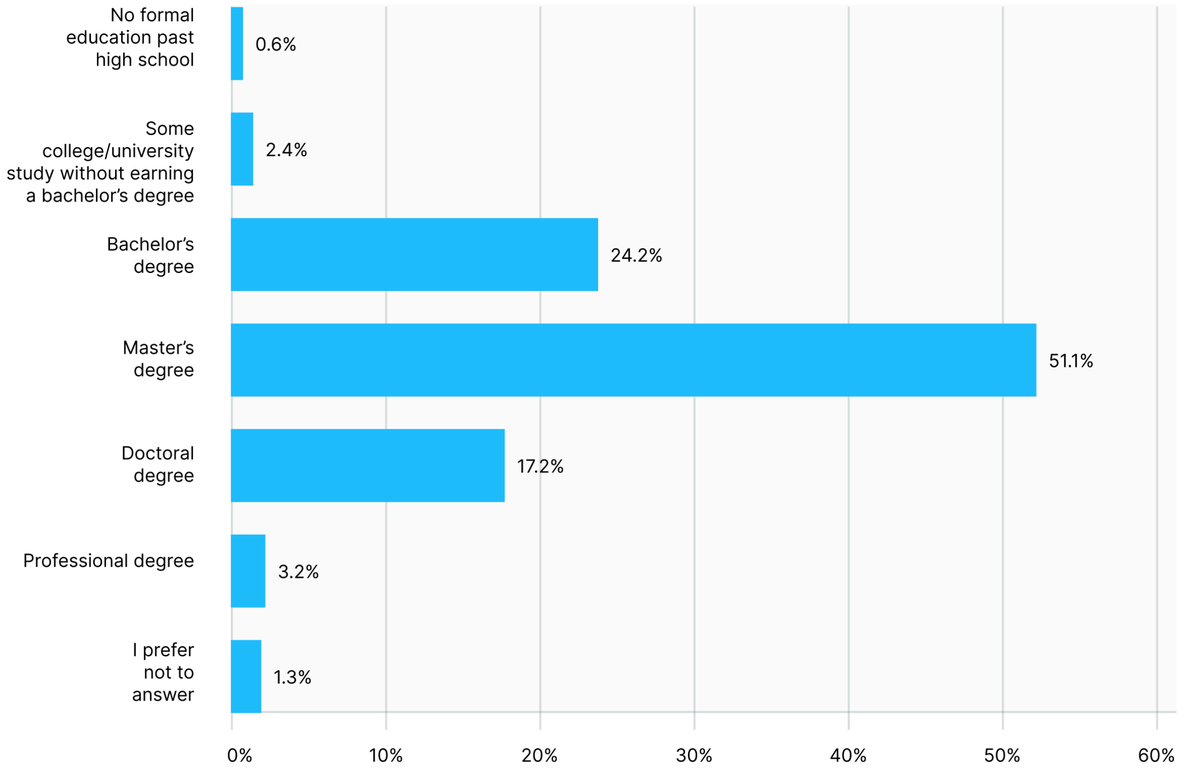


Ongoing Learning

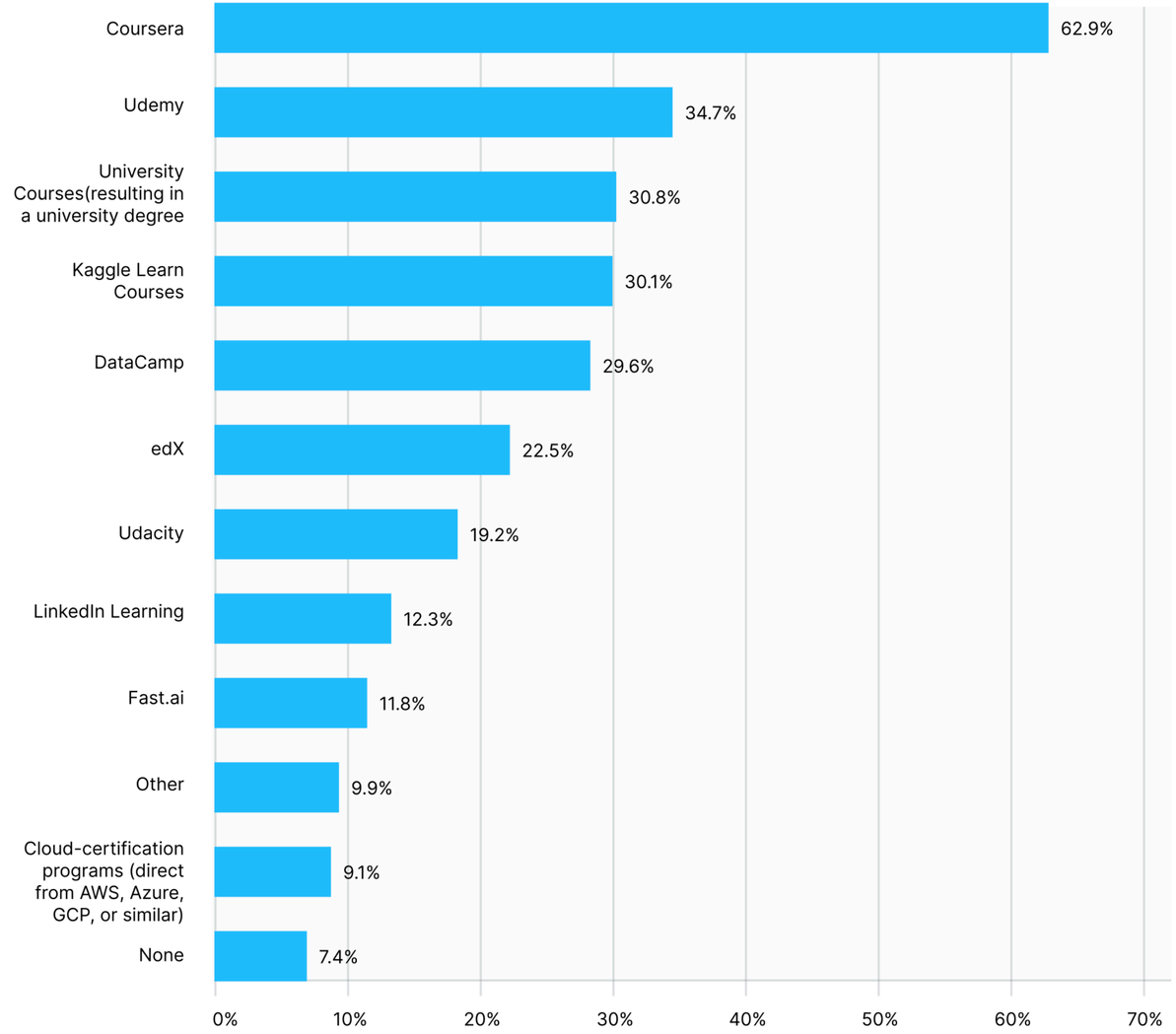
Data science and machine learning are quickly changing, so it's no surprise over 90% of Kaggle data scientists maintain ongoing education. While about 30% take traditional higher education courses, many more learn through online materials.

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

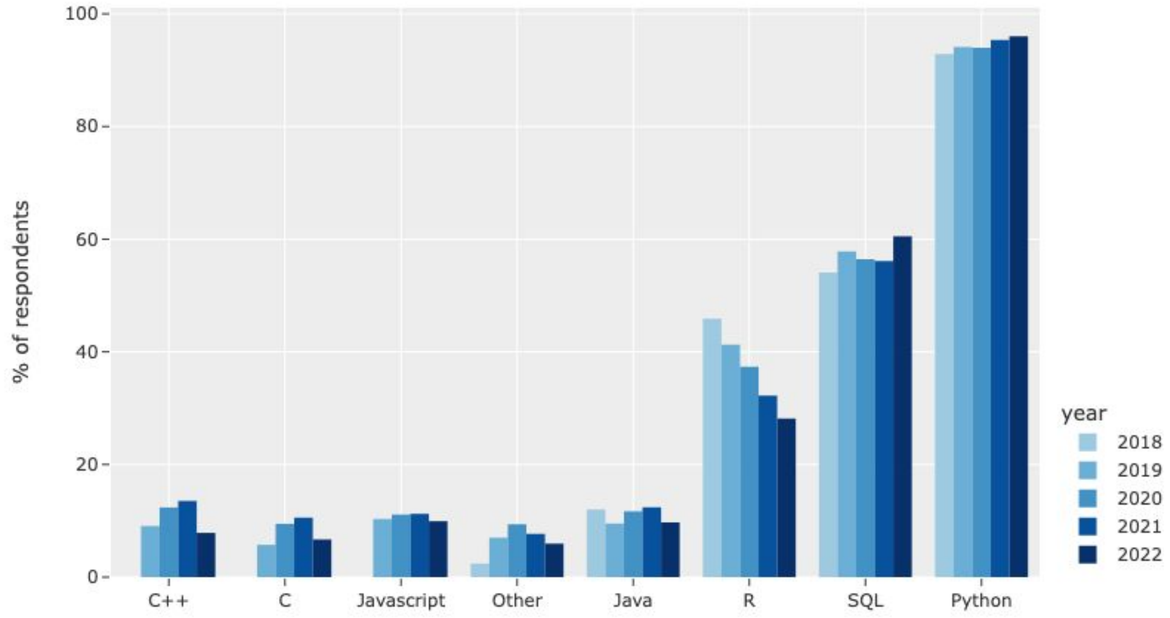


POPULAR ONGOING LEARNING RESOURCES



Kaggle DS & ML Survey 2022

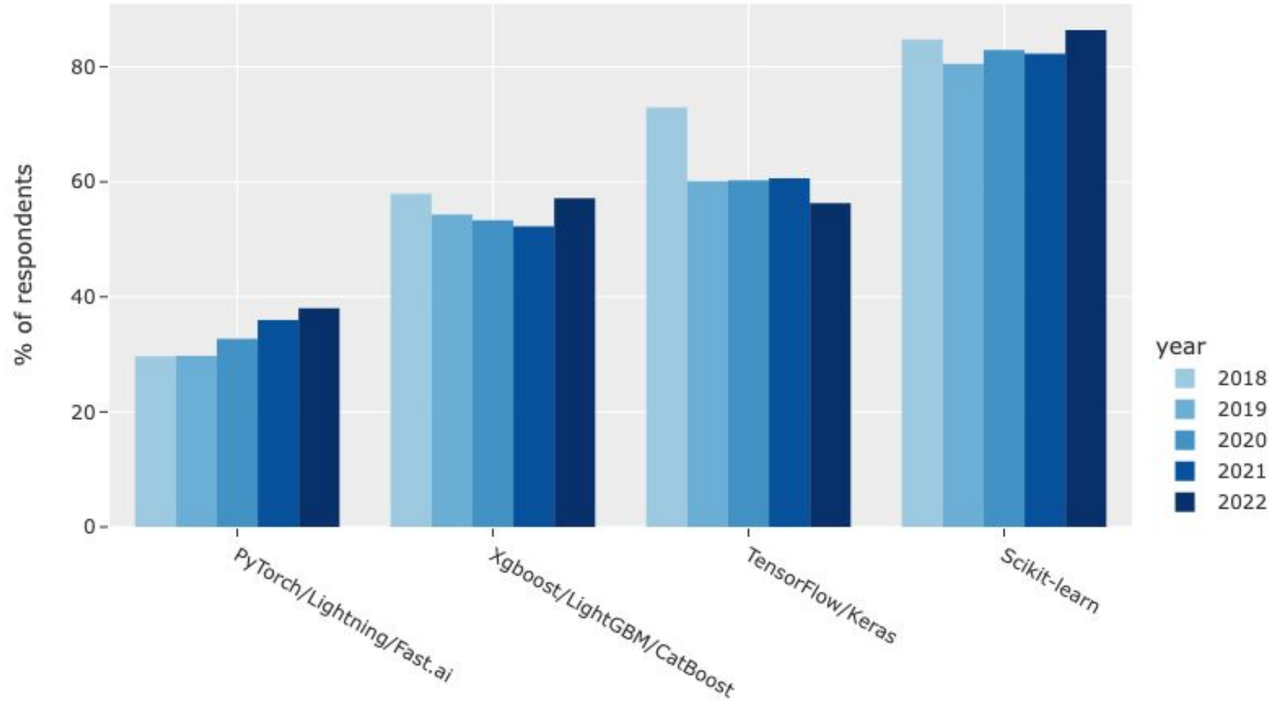
Python and SQL remain the two most common programming skills for data scientists



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

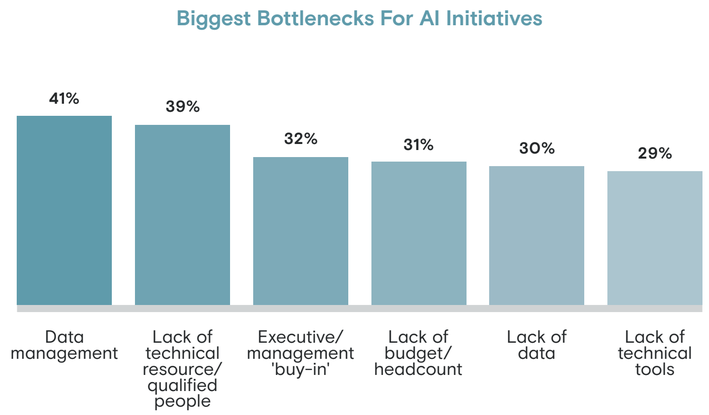
Responsible AI is the foundation of all AI projects.

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?



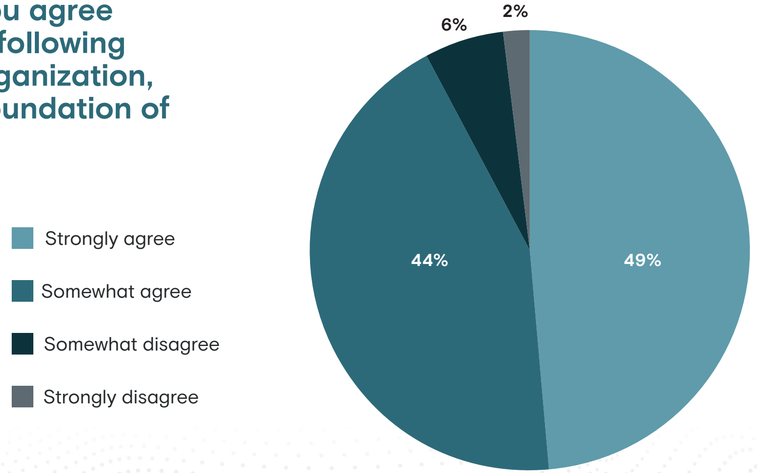
Ethics

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.

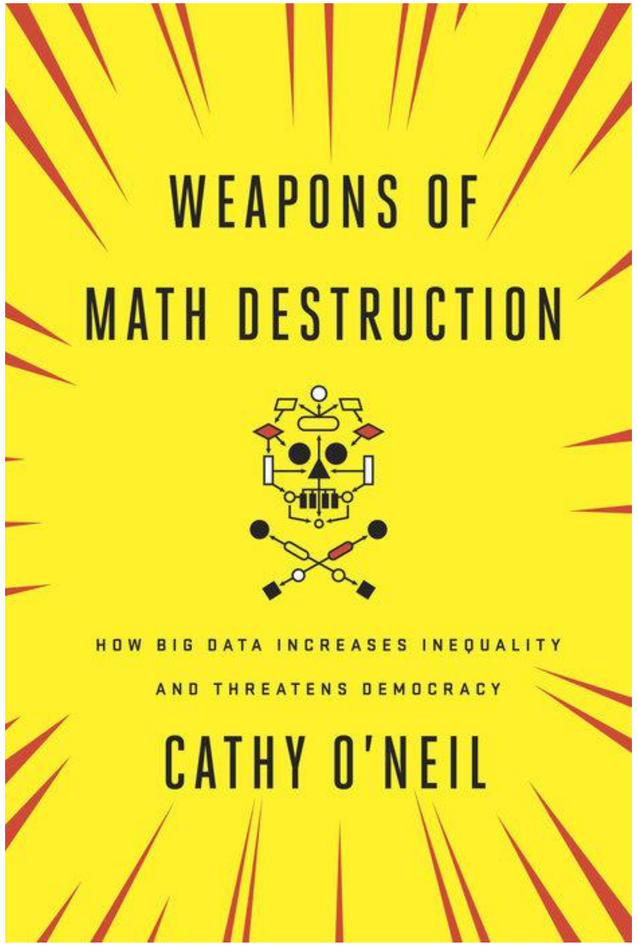
"Data ethics isn't just about doing the right thing, it's about maintaining the trust and safety of everyone along the value chain from contributor to consumer."
Erik Vogt
VP of Enterprise Solutions - Appen

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

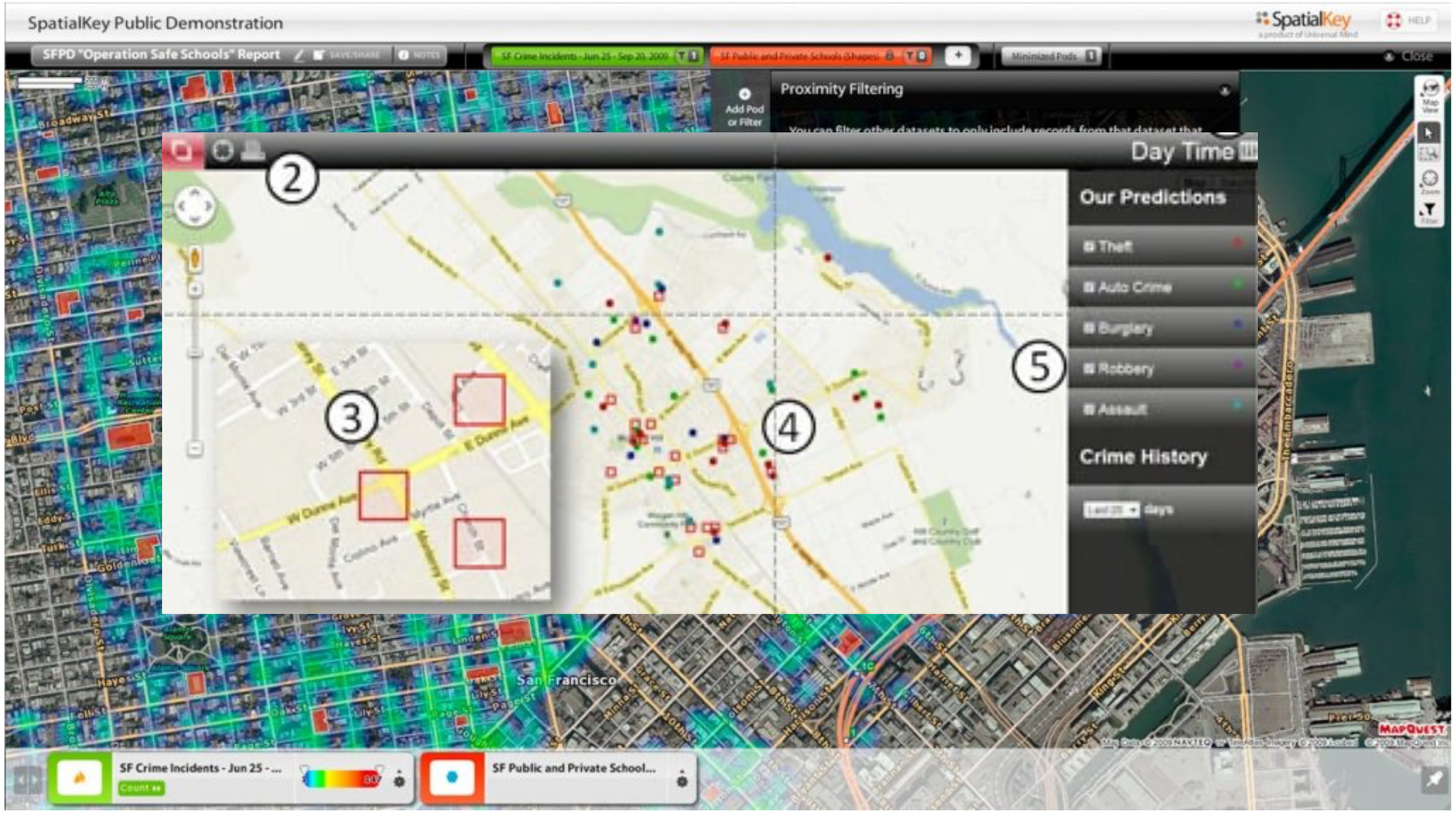


Don’t be a tool for creating WMDs

- 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



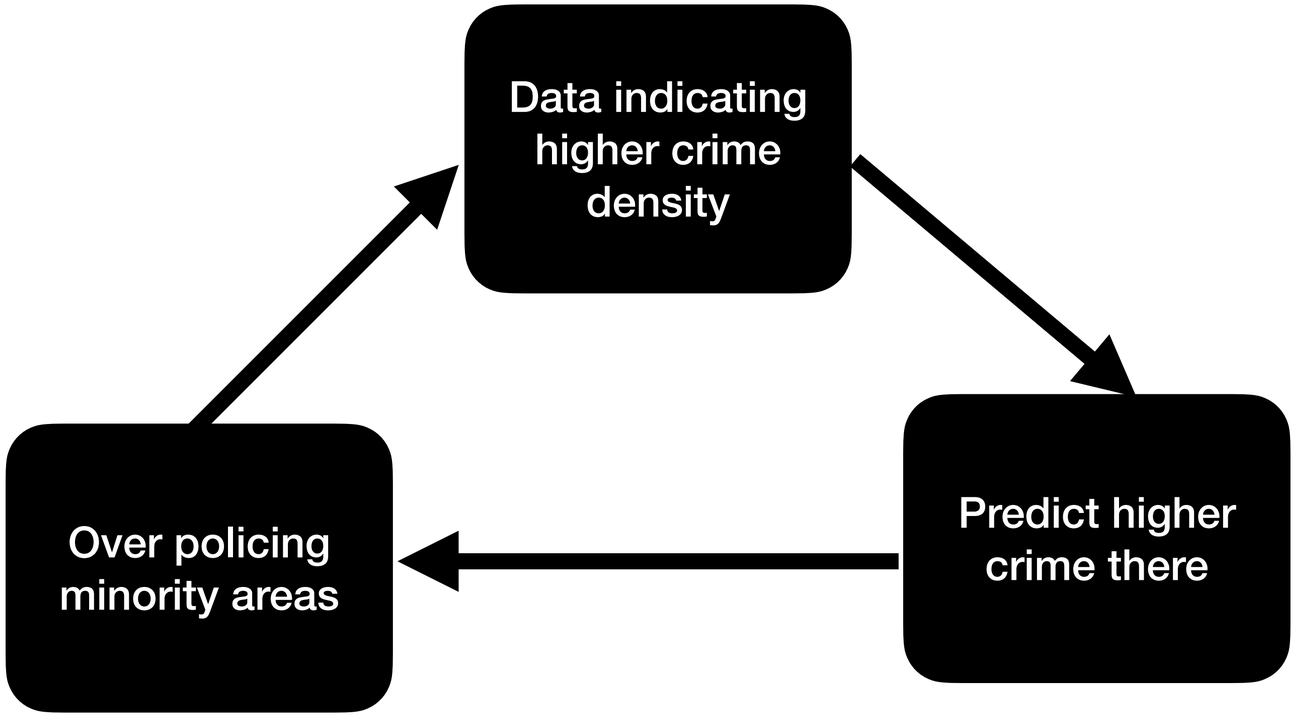
Predictive policing & sentencing



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!

Devamardeep Hayatpur
Joshua Chen
Fuling Sun
Zoe He

Ben Bao
Dhathry Doppalapudi
Fatima Dong
Ricky Zhu
Shreya Musini

And thanks to YOU!