Course Reminders

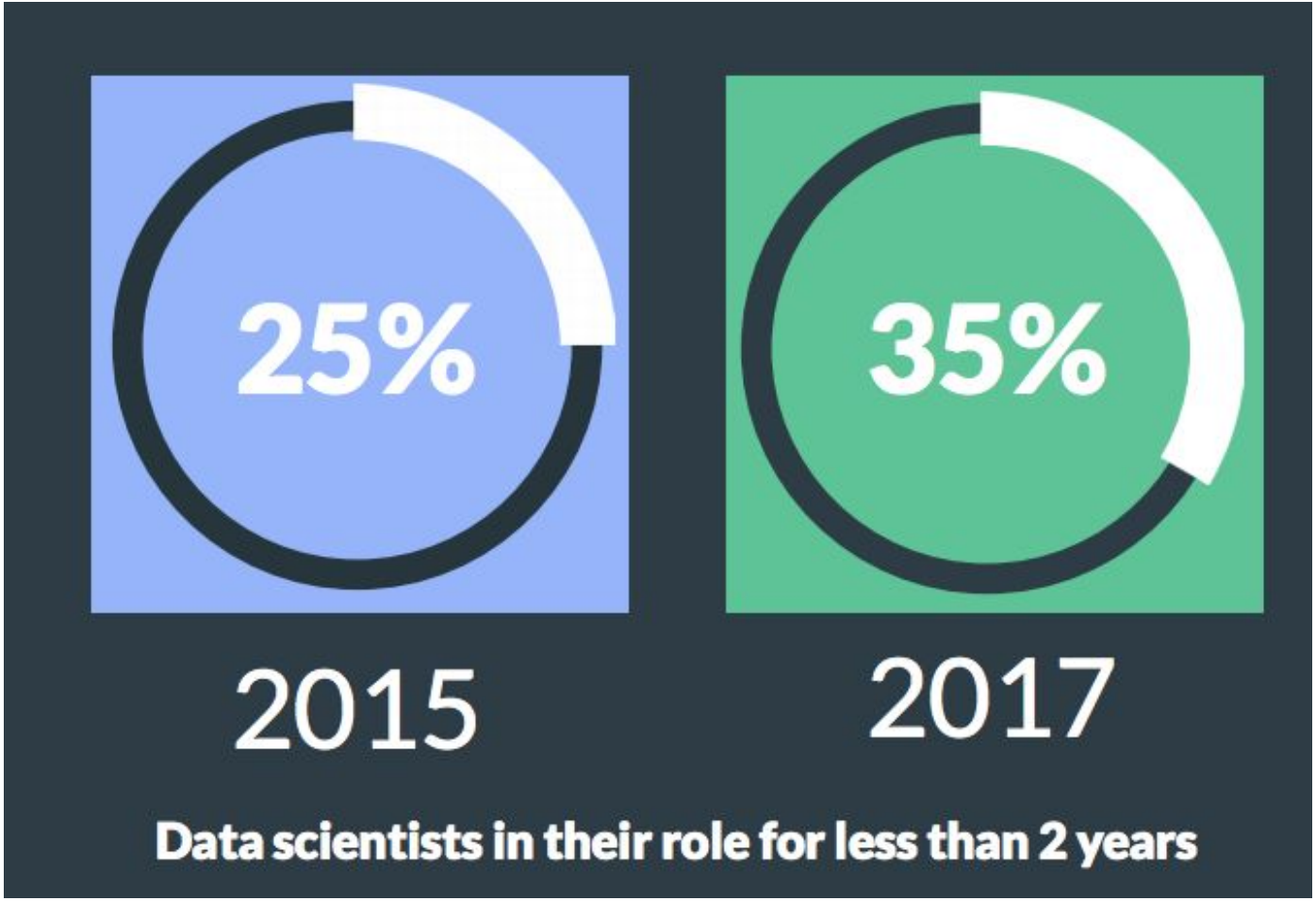
- Final Project due Thurs, March 19th (11:59 PM)
  - Group Work Survey: http://bit.ly/teams_wi20
- Post COGS 108 Survey: http://bit.ly/post_wi20
- CAPEs: http://www.cape.ucsd.edu/ (65%)

The Future of Data Science

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

Where we are now



Source: CrowdFlower, *Data Science Report* (2017)

#1 Data Scientist

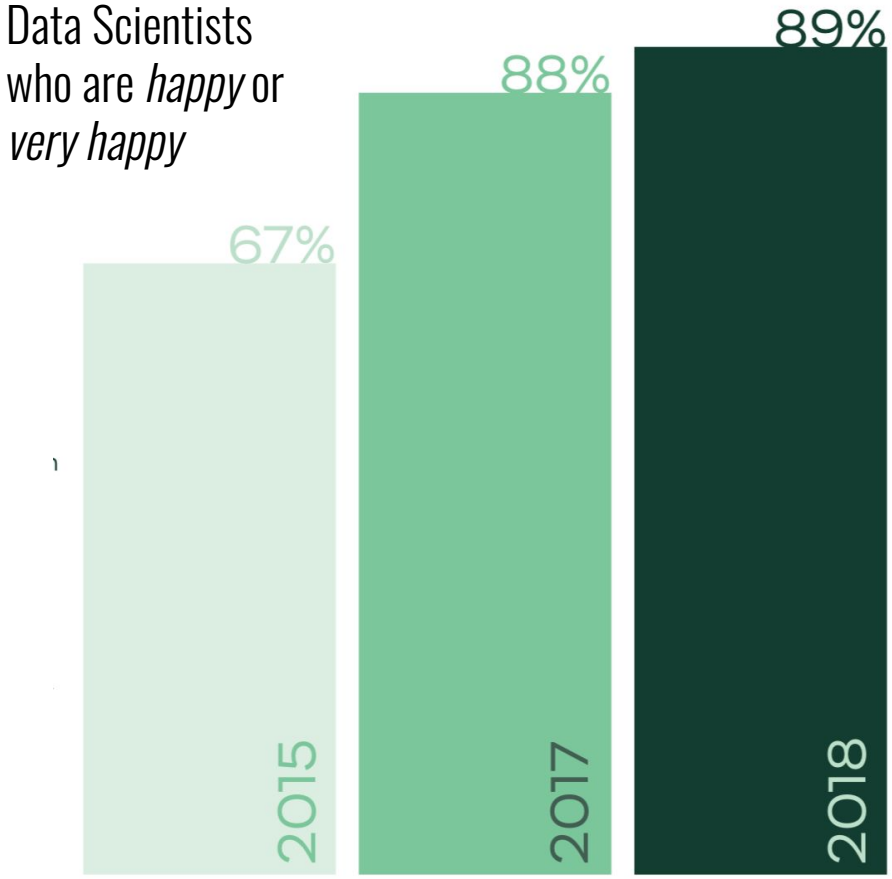
$108,000
Median Base Salary

4.3/5
Job Satisfaction

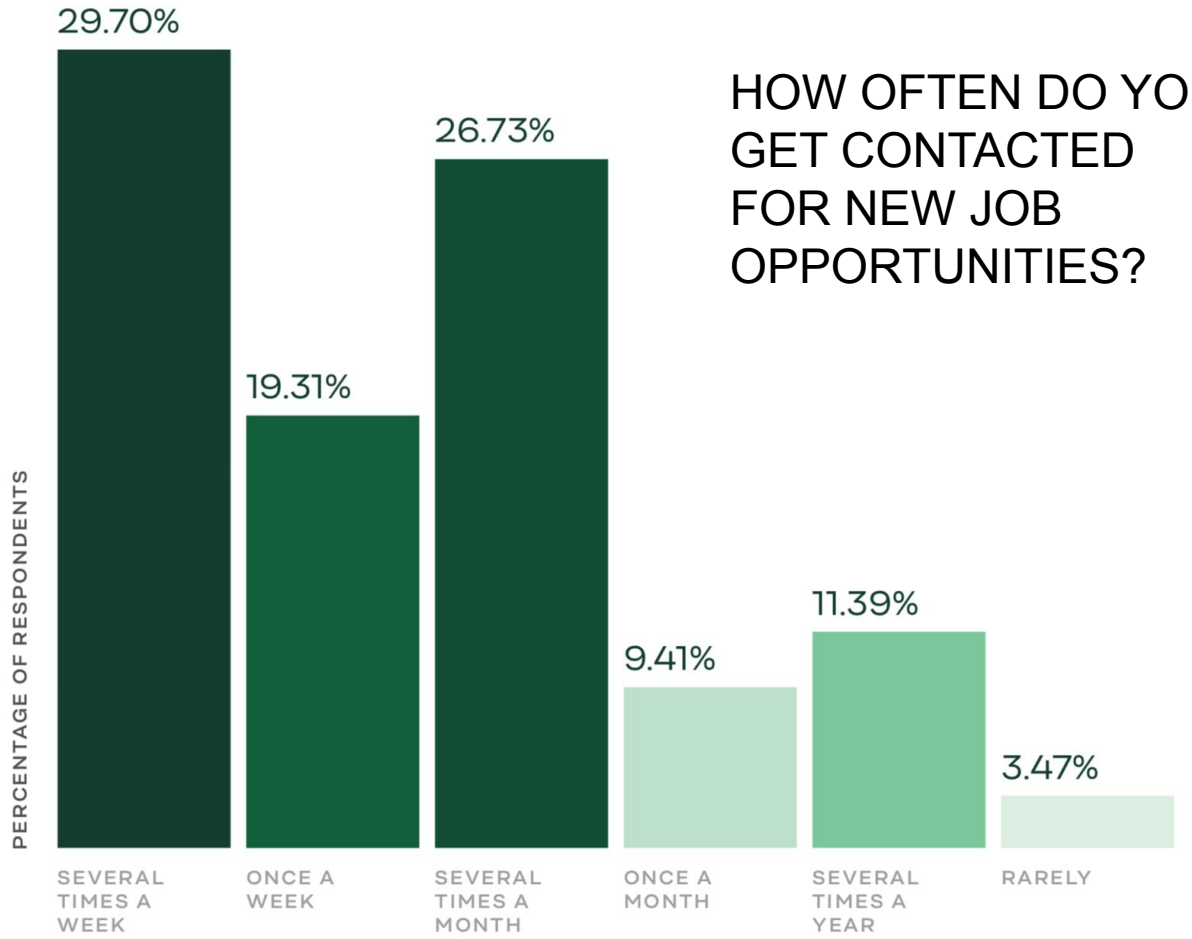
6,510
Job Openings

View Jobs

Source: Glassdoor (2019)



Source: Figure Eight, *Data Science Report* (2018)



Source: Figure Eight, *Data Science Report* (2018)

The Ten Most Common Data Science Skills in Job Postings

Skill	Percentage of Job Listings
Python	**72%**
R	**64%**
SQL	**51%**
Hadoop	**39%**
Java	**33%**
SAS	**30%**
Spark	**27%**
Matlab	**20%**
Hive	**17%**
Tableau	**14%**

Source: Glassdoor Economic Research.

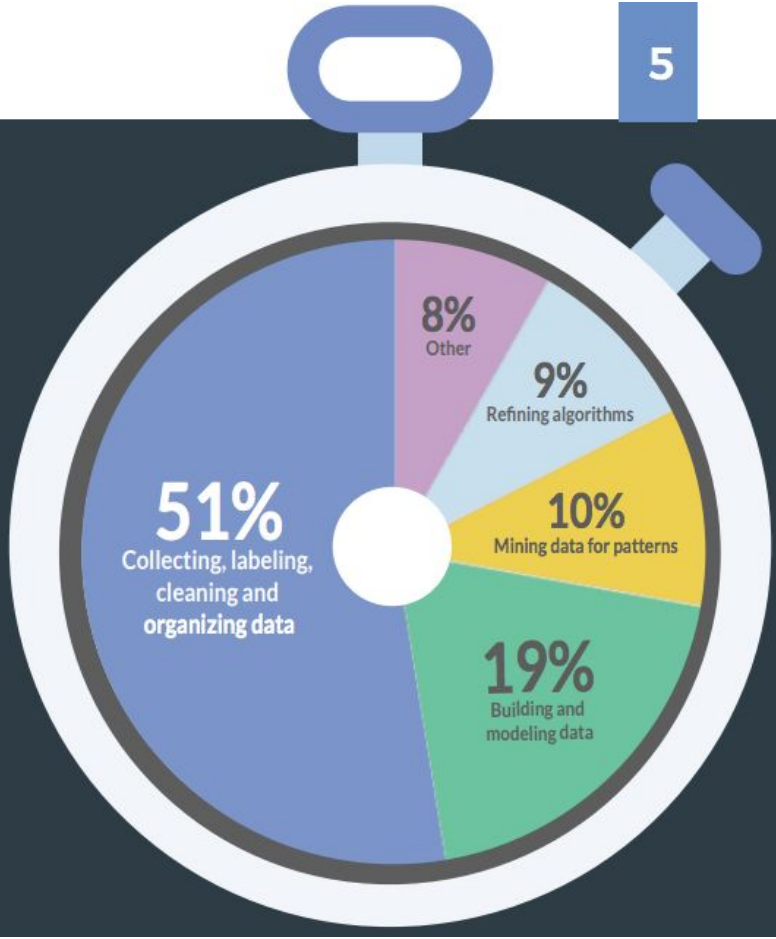
glassdoor

Source: Glassdoor

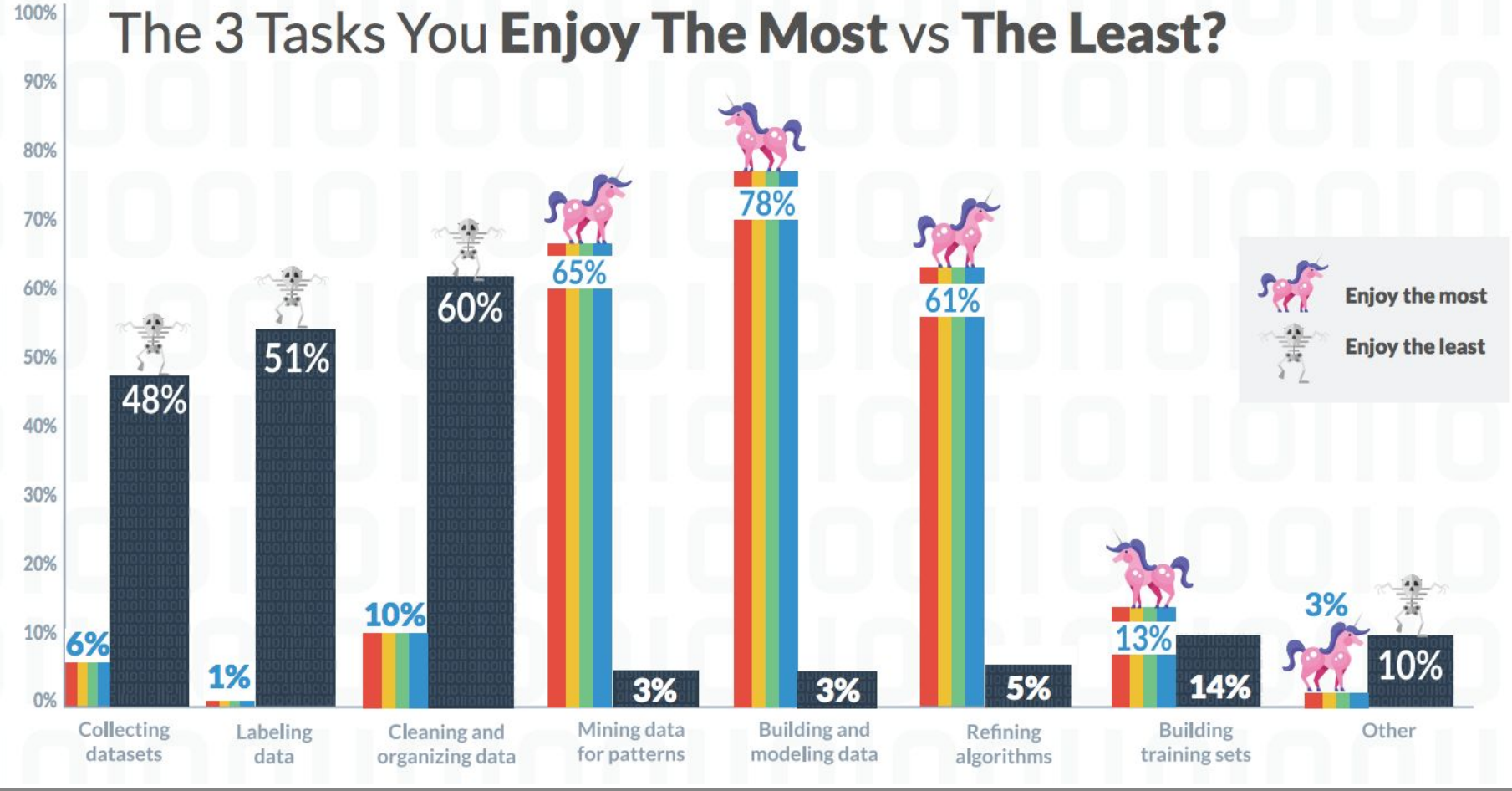
WHAT KEEPS DATA SCIENTISTS HAPPY?

(and why aren't they doing more of it?)

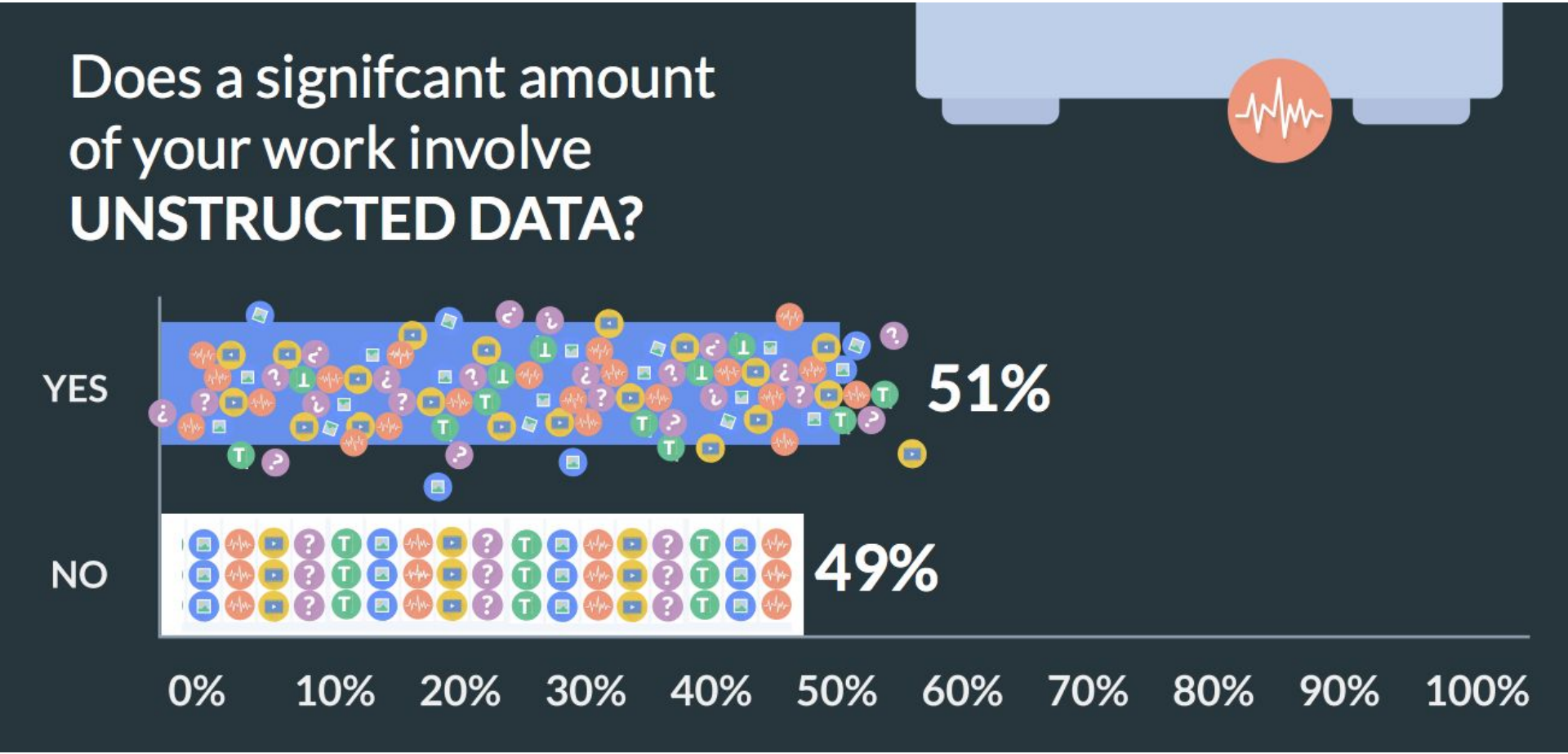
What activity takes up most of your time?



Source: CrowdFlower, *Data Science Report* (2017)

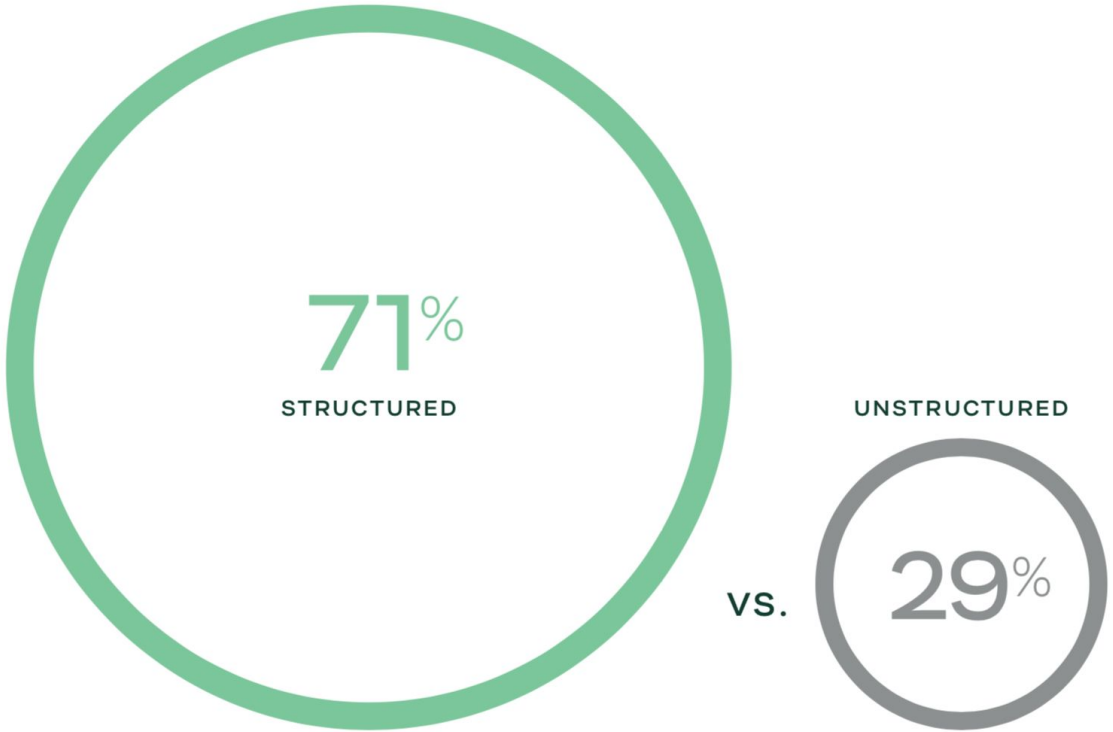


Source: CrowdFlower, *Data Science Report* (2017)

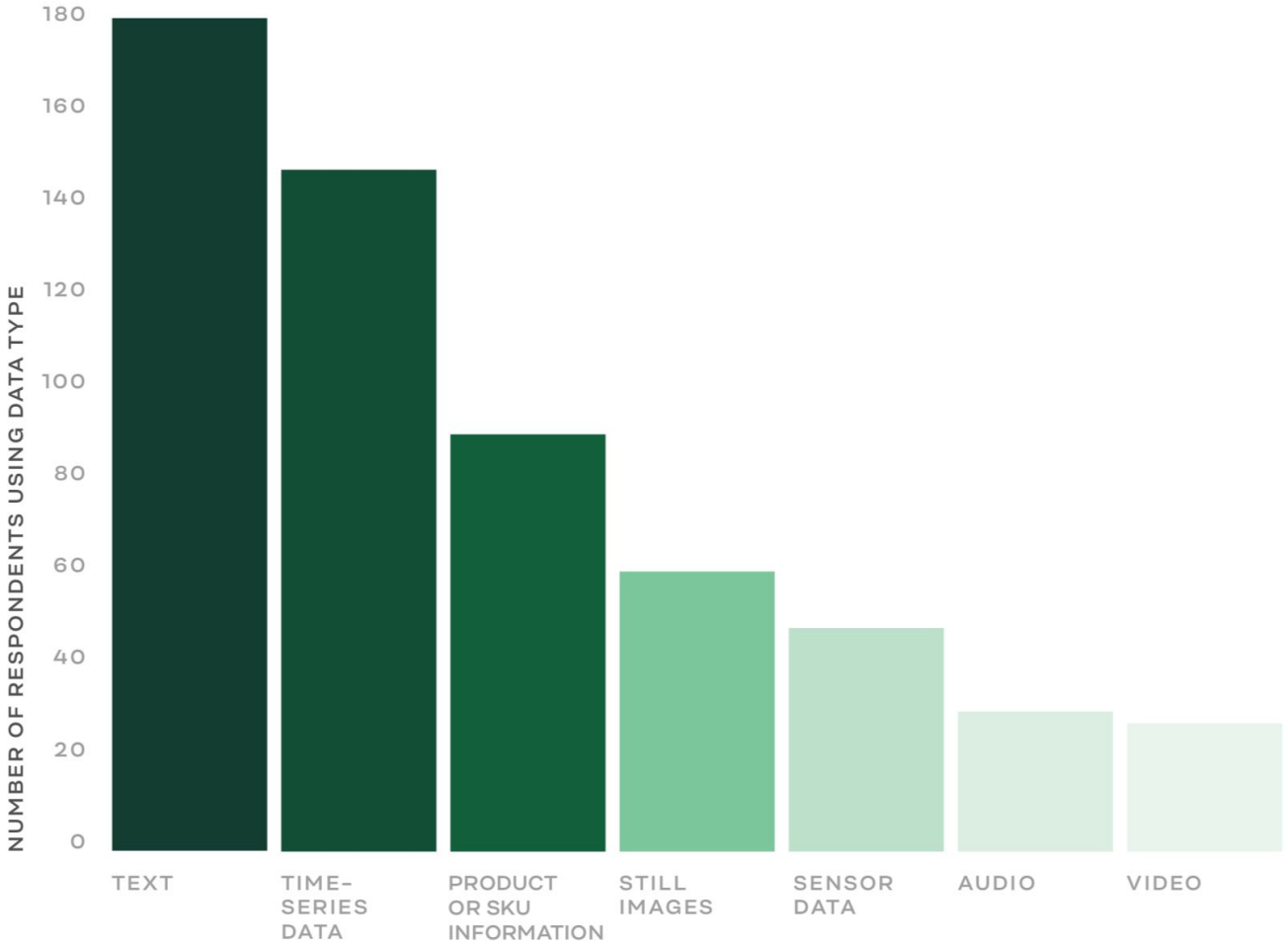


Source: CrowdFlower, *Data Science Report* (2017)

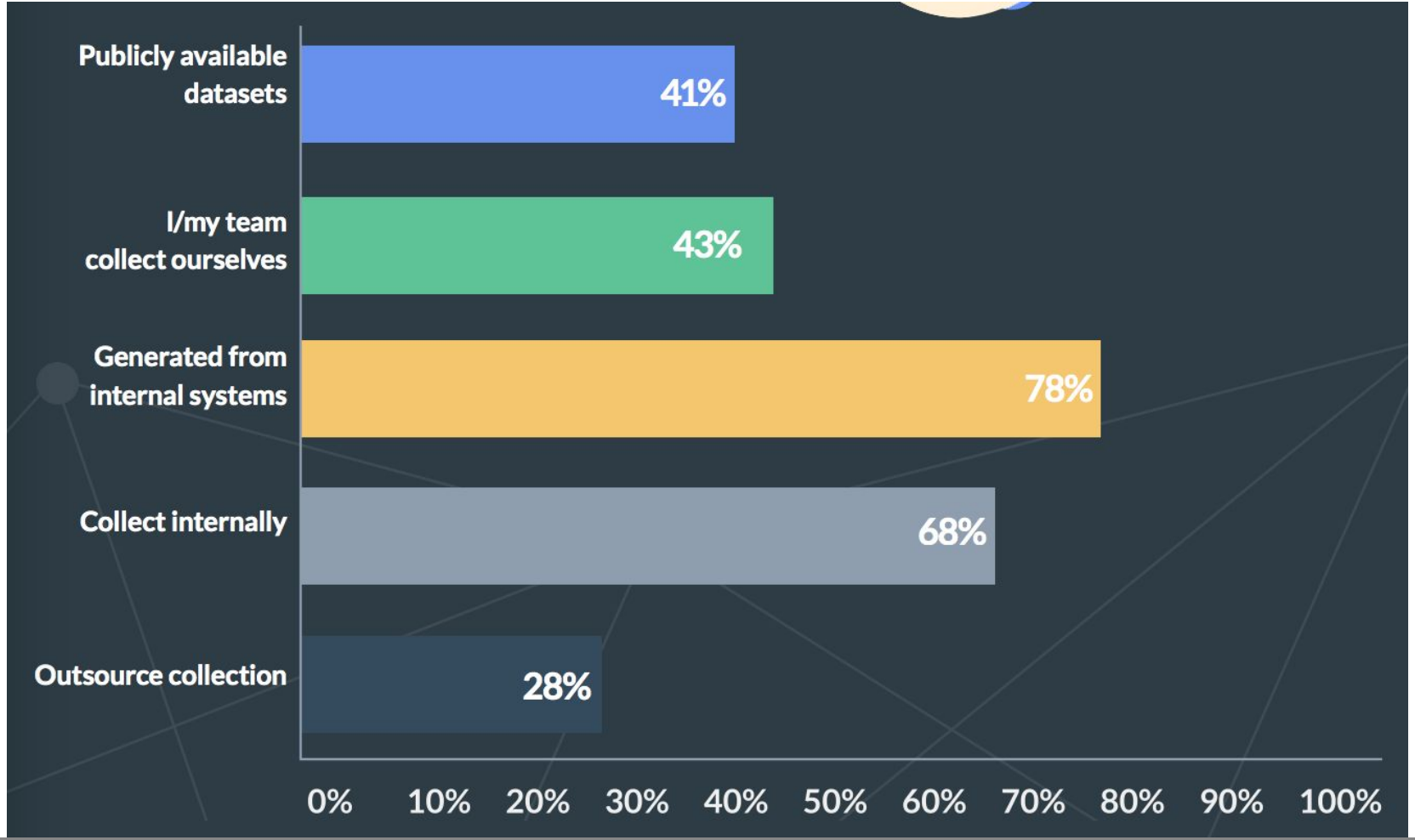
DO YOU WORK PRIMARILY WITH STRUCTURED OR UNSTRUCTURED DATA?



Source: Figure Eight, *Data Science Report* (2018)



Source: Figure Eight, *Data Science Report* (2018)



Source: CrowdFlower, *Data Science Report* (2017)

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/

https://gph.is/17VPNIG

What you all have done

COGS 108: What we've learned

01: Data Science, Ethics, & Version Control
02: Python Programming & pandas
03: Data & Data Wrangling
04: Data Intuition & Visualization
05: Descriptive & Exploratory Data Analysis
06: Inference
07: Machine Learning
08: Text Analysis
09: Geospatial Analysis & Non-parametric Statistics
10: Dimensionality Reduction & DS Jobs

Guest Lectures: Agustin Lebron & Andy White

COGS 108: Final Project Lessons

1. Asking the right question up front really helps
2. Finding the data you need is a skill
    a. ...so is knowing if the data are reliable
    b. ...and if they can answer your question
    c. ....and recognizing what information you don't have
3. Data Visualization and storytelling are important skills.
4. Determining which analytical approach is best is HARD.
5. Programming is merely a piece of the puzzle for data scientists.

COGS 108 Thank yous!

TAs: Matt, Sam, & Shreenivas

IAs: Amir, Emily, John, Josh

All of you for your patience, feedback, and time!

COGS108 thoughts

Which of the following best describes you?

A
I understand what data science is *less* than before taking COGS108.

B
I knew what data science was before taking COGS108 and my understanding hasn't changed.

C
I knew what data science was before taking COGS108 but have a better understanding now.

D
I didn't really know what data science was before and have a better understanding having taken COGS108.

COGS108 thoughts

Which of the following best describes you?

A
I was excited before about being a data scientist and am just as excited after COGS108.

B
I never wanted to be a data scientist, but think what I learned in COGS108 could help me in the future

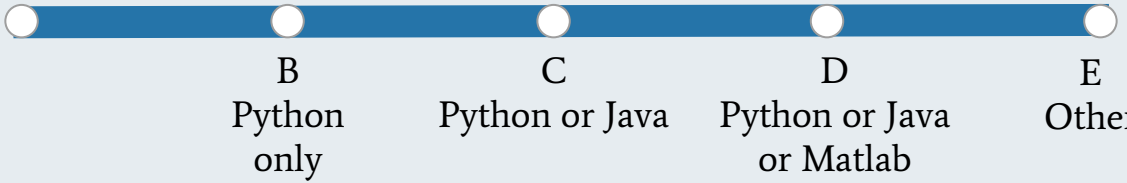
C
I never wanted to be a data scientist & don't think COGS108 topics will help me in the future.

D
I'm more interested in becoming a data scientist after COGS108.

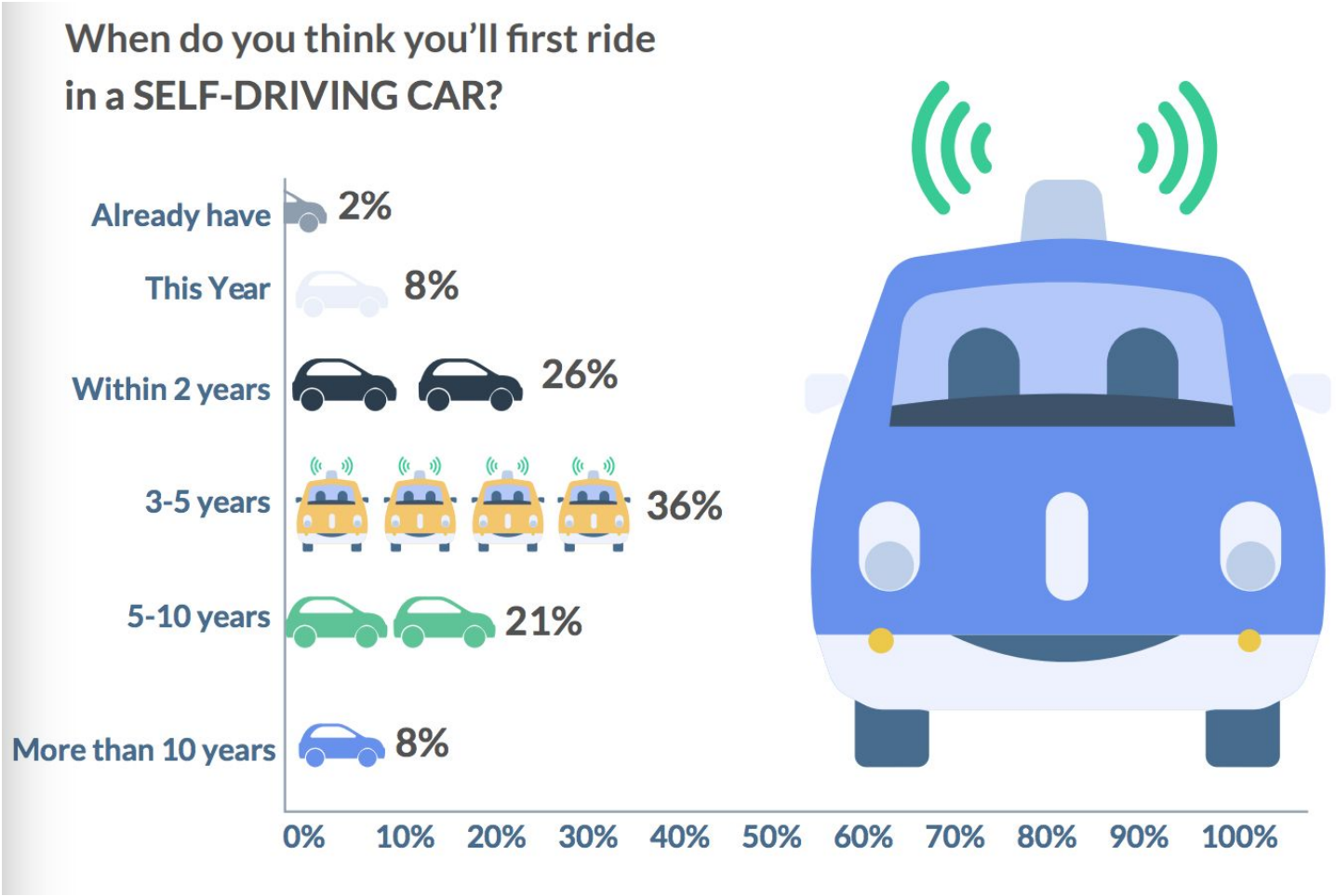
E
I'm less interested in becoming a data scientist than I was before taking COGS108.

COGS108 thoughts

What should the prerequisite for this course be?

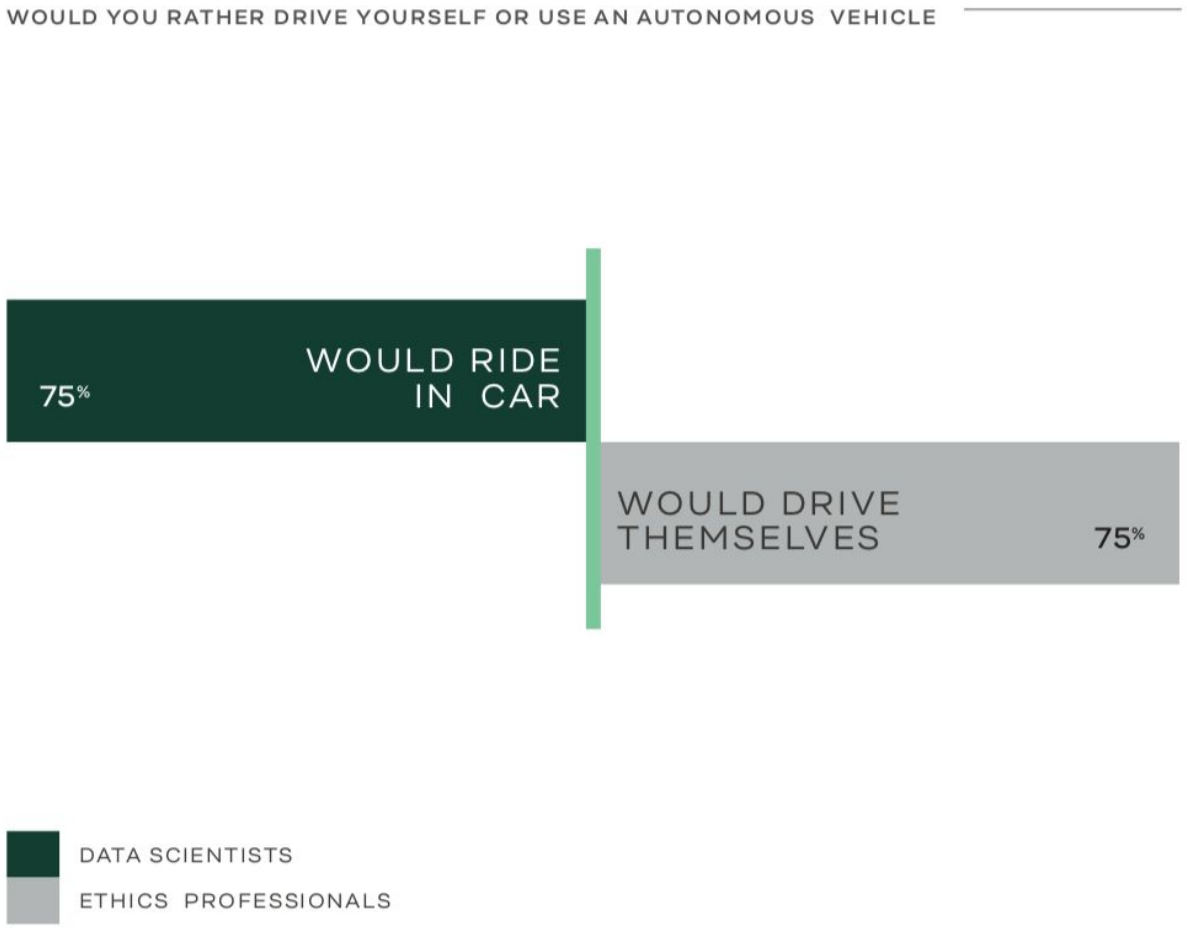


...so where are we going?



Source: CrowdFlower, *Data Science Report* (2017)

Source: What a Driverless World Could Look Like (Wanis Kabbaj, TED)



Source: Figure Eight, *Data Science Report* (2018)

Algorithms are fragile

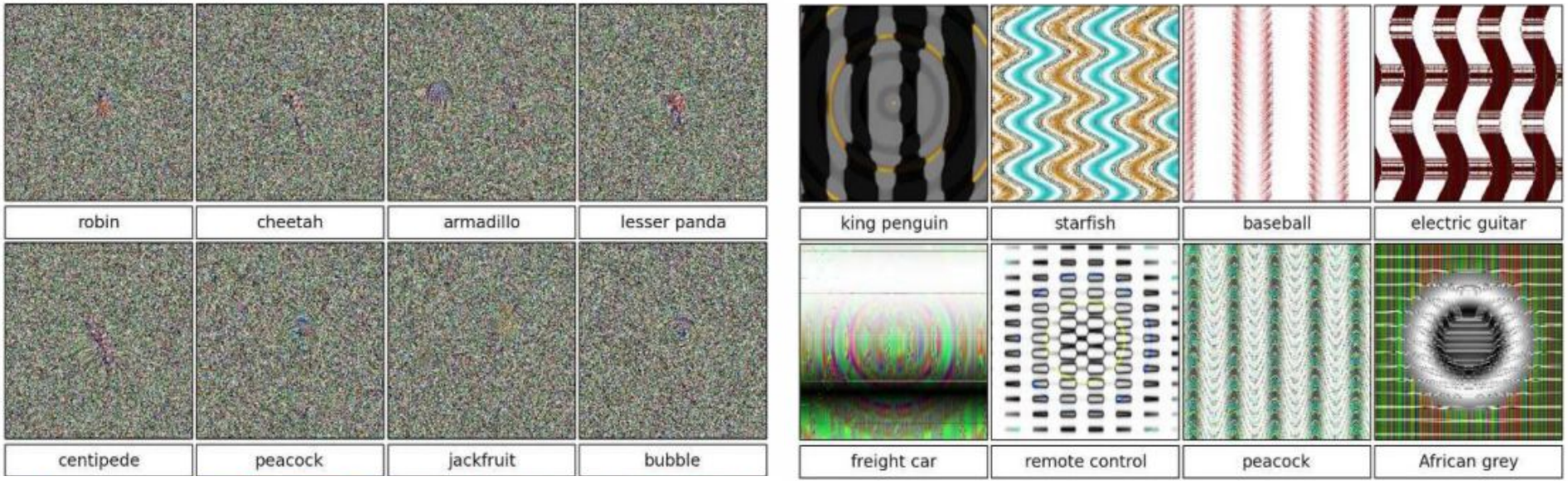
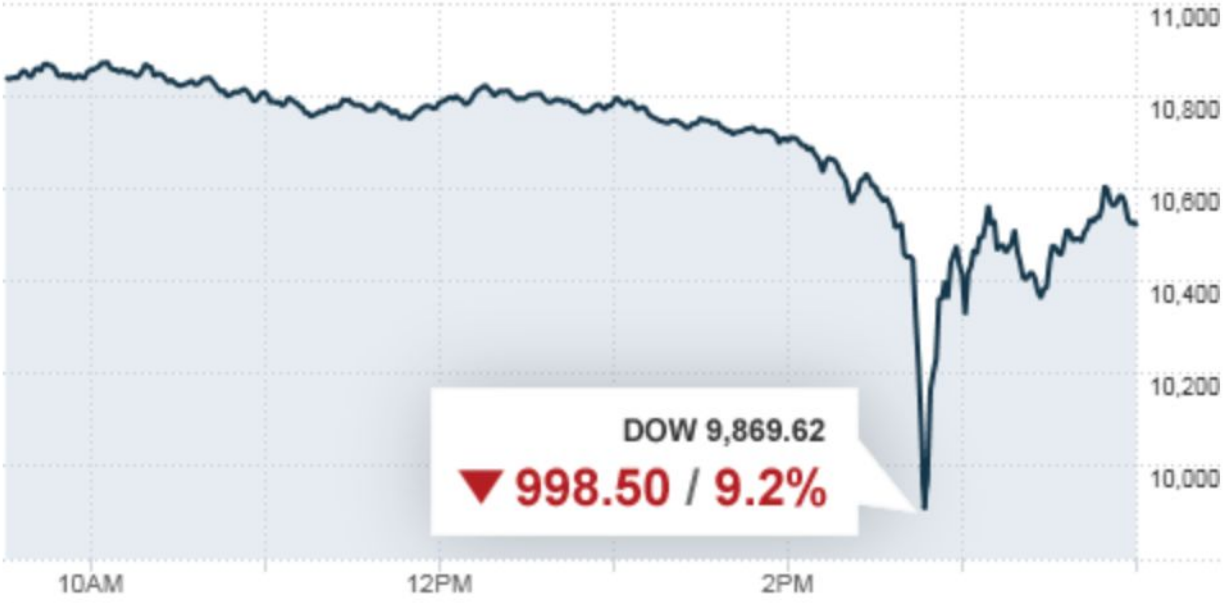


Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with $\geq$ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly (*bottom*) encoded.

Deep Neural Networks are Easily Fooled (IEEE, 2015)

Trading program sparked May 'flash crash'



Government regulators say a trading program was behind the massive stock slide on May 6.

Automatic computerized traders on the stock market shut down as they detected the sharp rise in buying and selling. (*NYT*)

Source: CNN Money

Algorithms are fragile & powerful

Human-based computation

Source: Massive-scale online collaboration (Luis von Ahn, TEDx)

Source: Massive-scale online collaboration (Luis von Ahn, TEDx)

From the amusing...

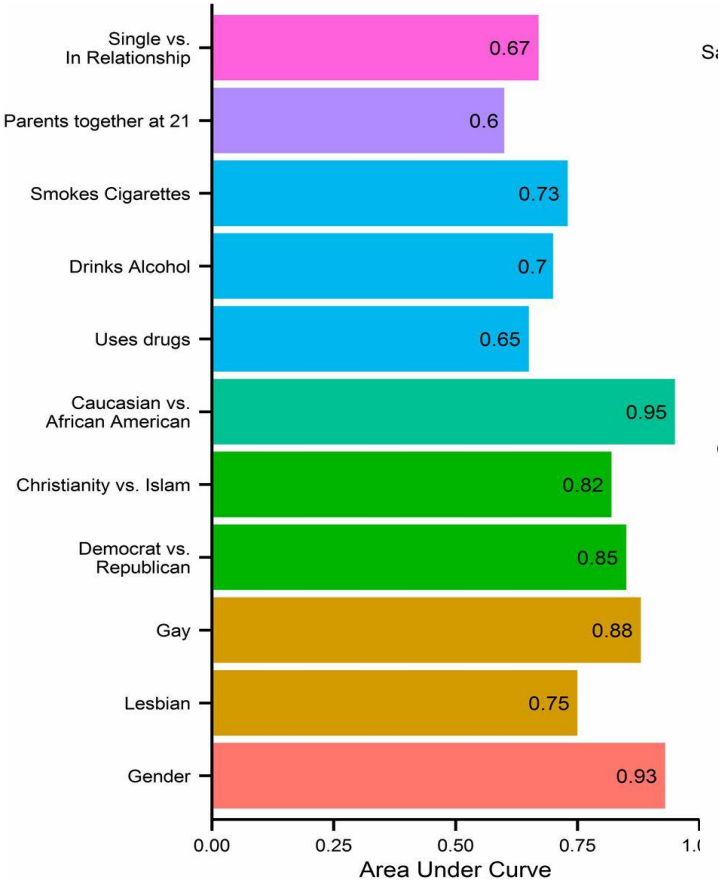


Fig 2: Prediction accuracy of classification of dichotomous/dichotomized attributes expressed by the AUC

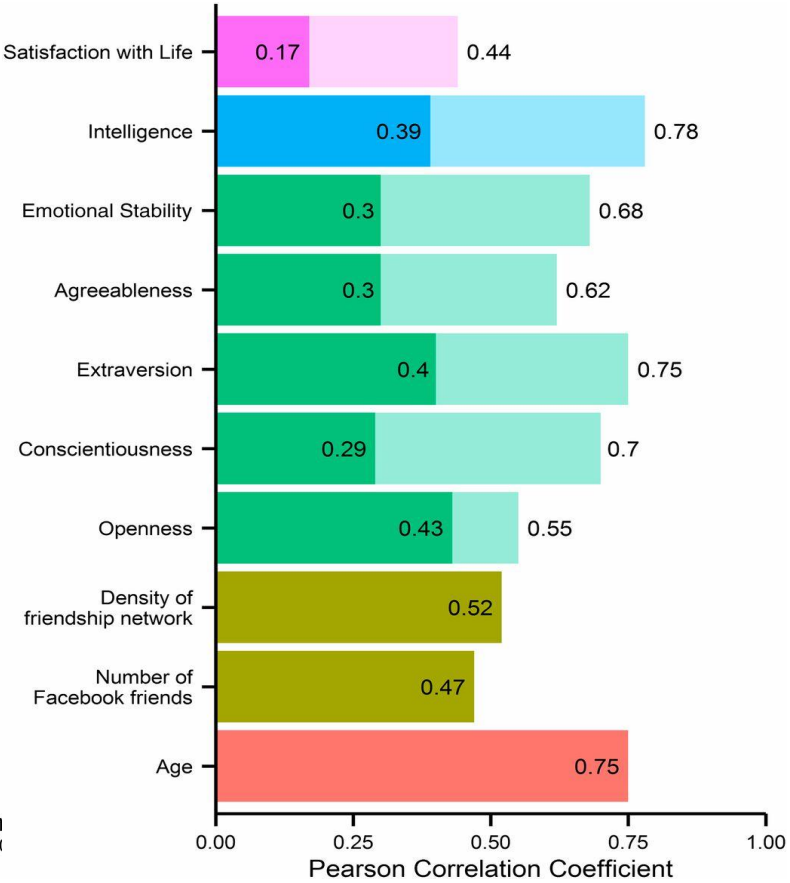


Fig 3: Prediction accuracy of regression for numeric attributes and traits expressed by the Pearson correlation coefficient between predicted and actual attribute values; all correlations are significant at the $P < 0.001$ level. The transparent bars indicate the questionnaire's baseline accuracy, expressed in terms of test–retest reliability.

Predictive Power of Likes. Individual traits and attributes can be predicted to a high degree of accuracy based on records of users' Likes. Table S1 presents a sample of highly predictive Likes related to each of the attributes. For example, the best predictors of high intelligence include "Thunderstorms," "The Colbert Report," "Science," and "Curly Fries," whereas low intelligence was indicated by "Sephora," "I Love Being A Mom," "Harley Davidson," and "Lady Antebellum." Good predictors of male homosexuality included "No H8 Campaign," "Mac Cosmetics," and "Wicked The Musical," whereas strong predictors of male heterosexuality included "Wu-Tang Clan," "Shaq," and "Being Confused After Waking Up From Naps." Although some of the Likes clearly relate to their predicted attribute, as in the case of No H8 Campaign and homosexuality, other pairs are more elusive; there is no obvious connection between Curly Fries and high intelligence.

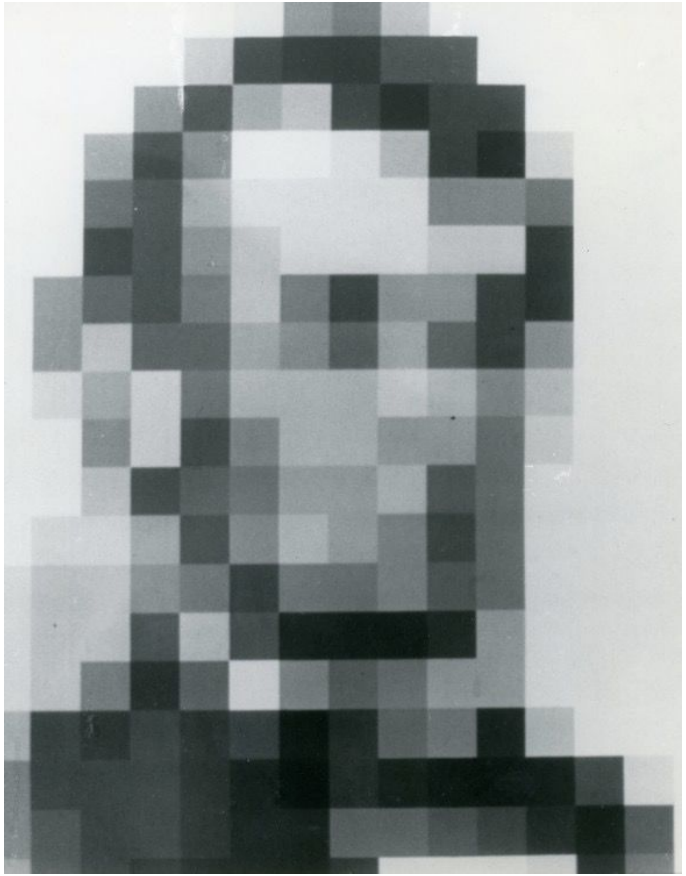
Private traits and attributes are predictable from digital records of human behavior (*PNAS*, 2013)

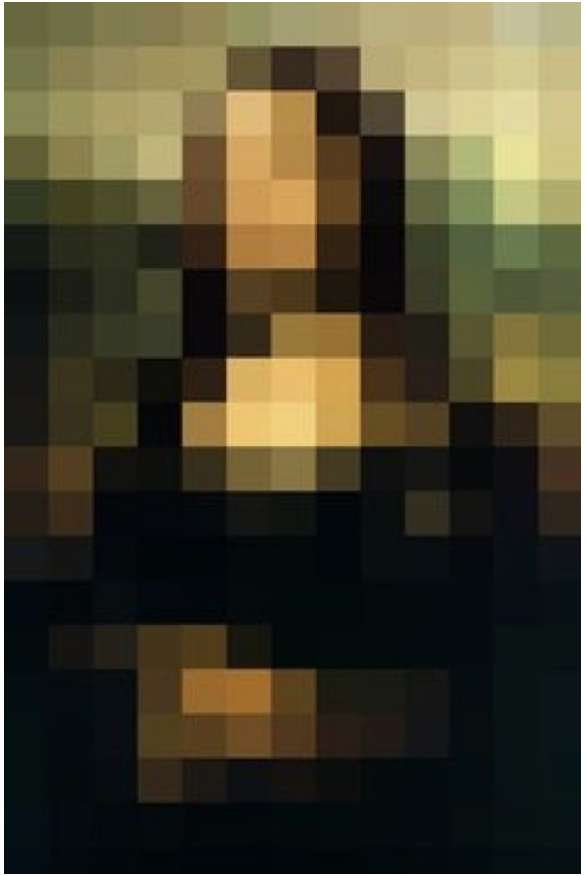
From the amusing...

...to the amazing

Andrew Ochoa, Waverly Labs

https://www.youtube.com/watch?v=Vxq9yj2pVWk&list=RDVxq9yj2pVWk&start_radio=1

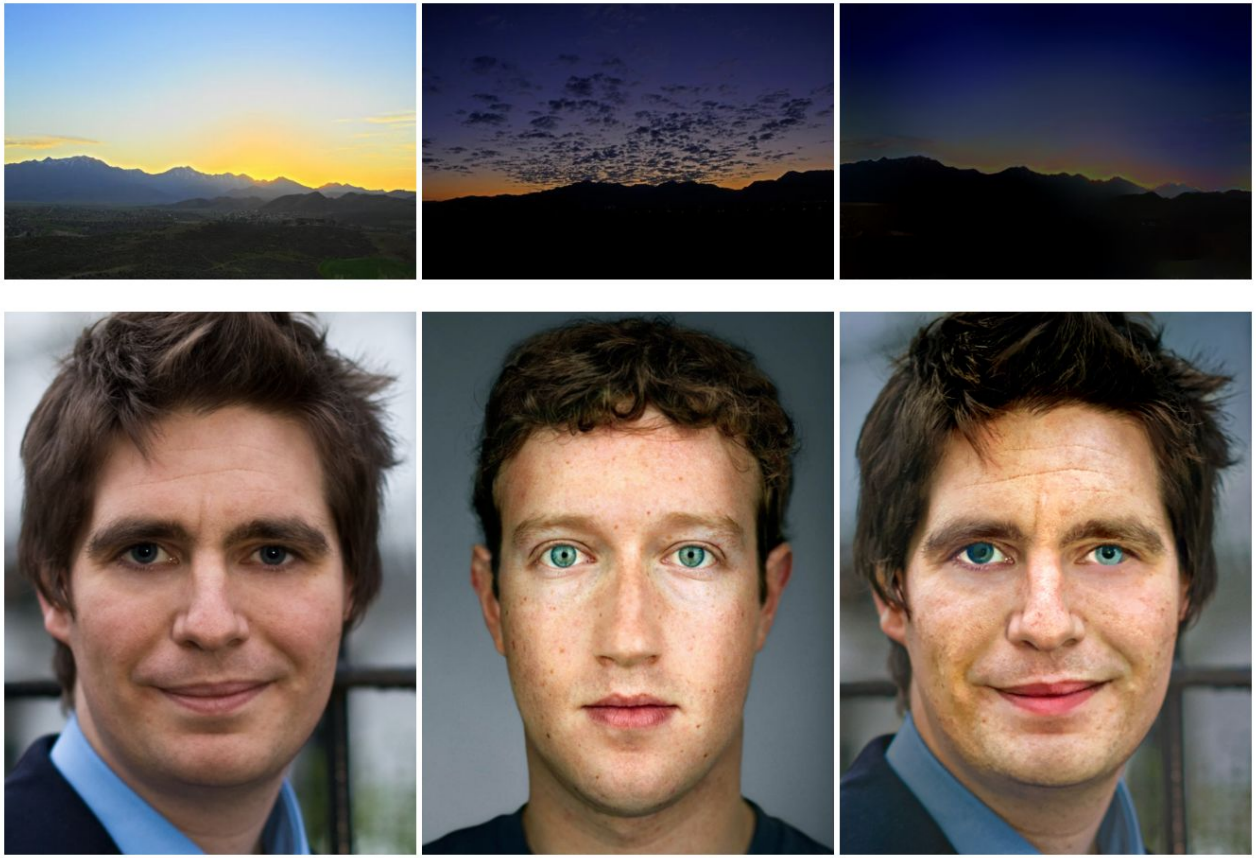




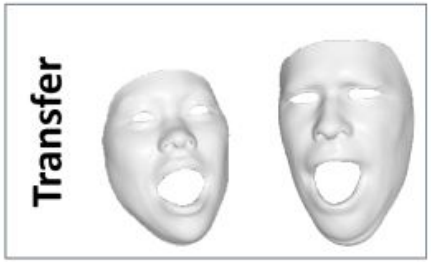
c.f. Harmon & Julesz, *Science* 1973

“...the first column is the 16x16 input image, the second one is what you would get from a standard bicubic interpolation, the third is the output generated by the neural net, and on the right is the ground truth.”

https://github.com/luanfujun/deep-photo-styletransfer



Reality manipulation



Proposed online reenactment setup: a monocular target video sequence (e.g., from Youtube) is reenacted based on the expressions of a source actor who is recorded live with a commodity webcam.

Personal Privacy

Source: Deb Roy, TED (2011)

60%
match

Strolling along the Seashore

Joaquín Sorolla y Bastida

346,054 views | Jan 17, 2019, 12:35pm

Was The Facebook '10 Year Challenge' A Way To Mine Data For Facial Recognition AI?

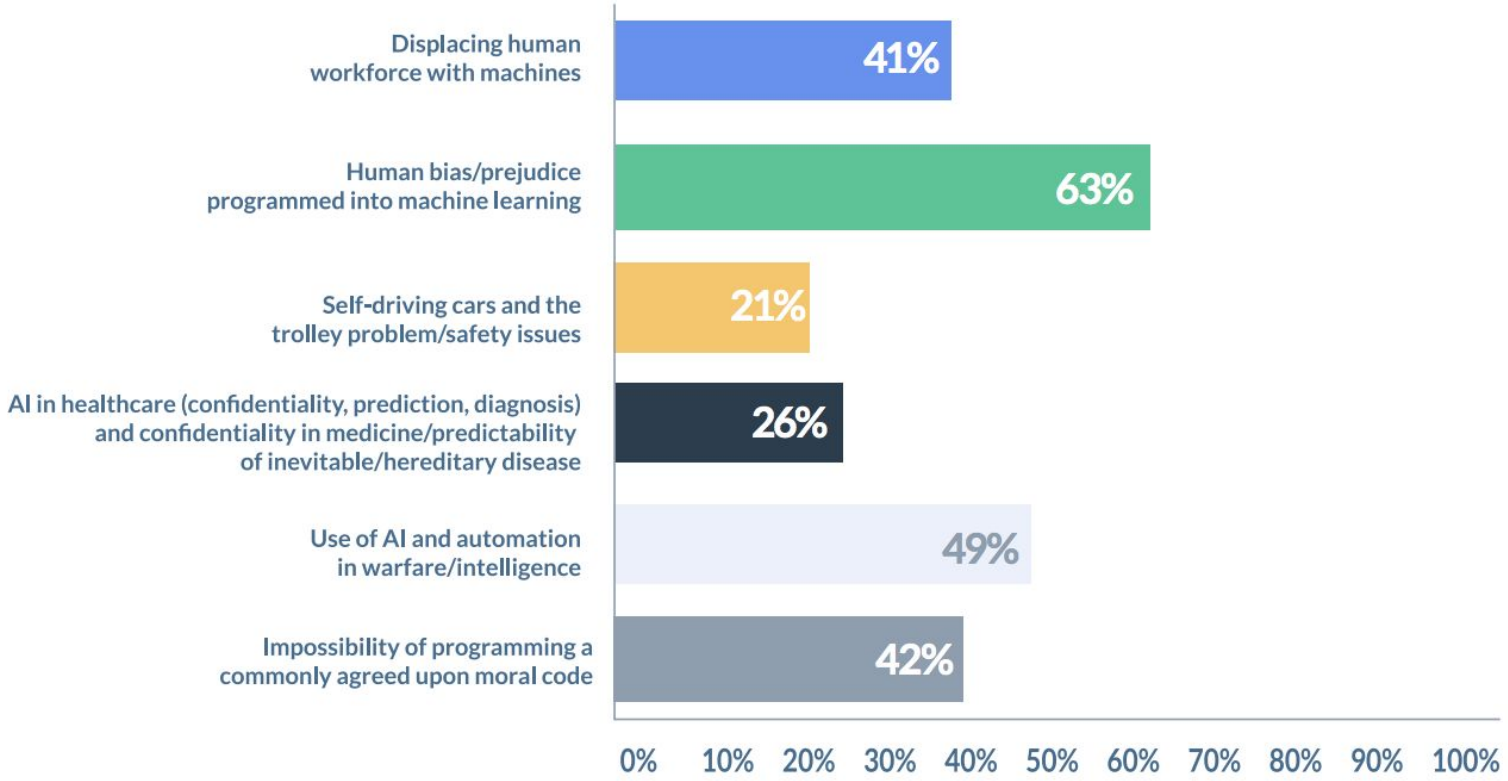
Nicole Martin Contributor ⓘ

AI & Big Data

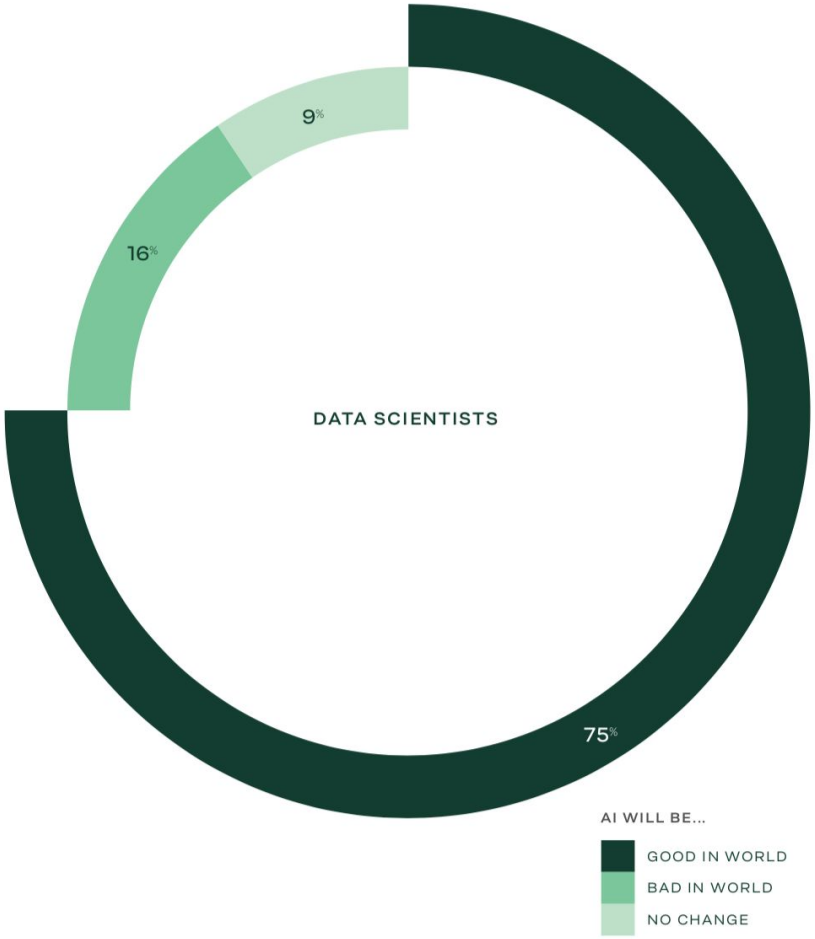
I write about technology, data and privacy.

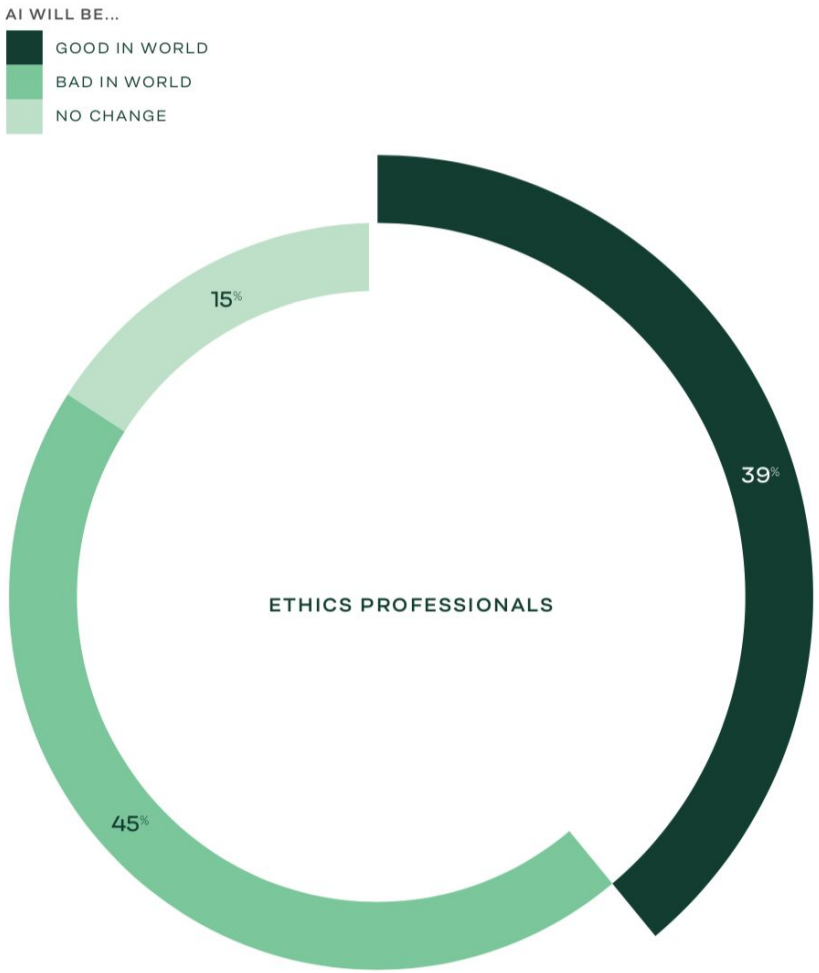
Source: Forbes

Which of the following do you personally think might be issues regarding ethics and AI?

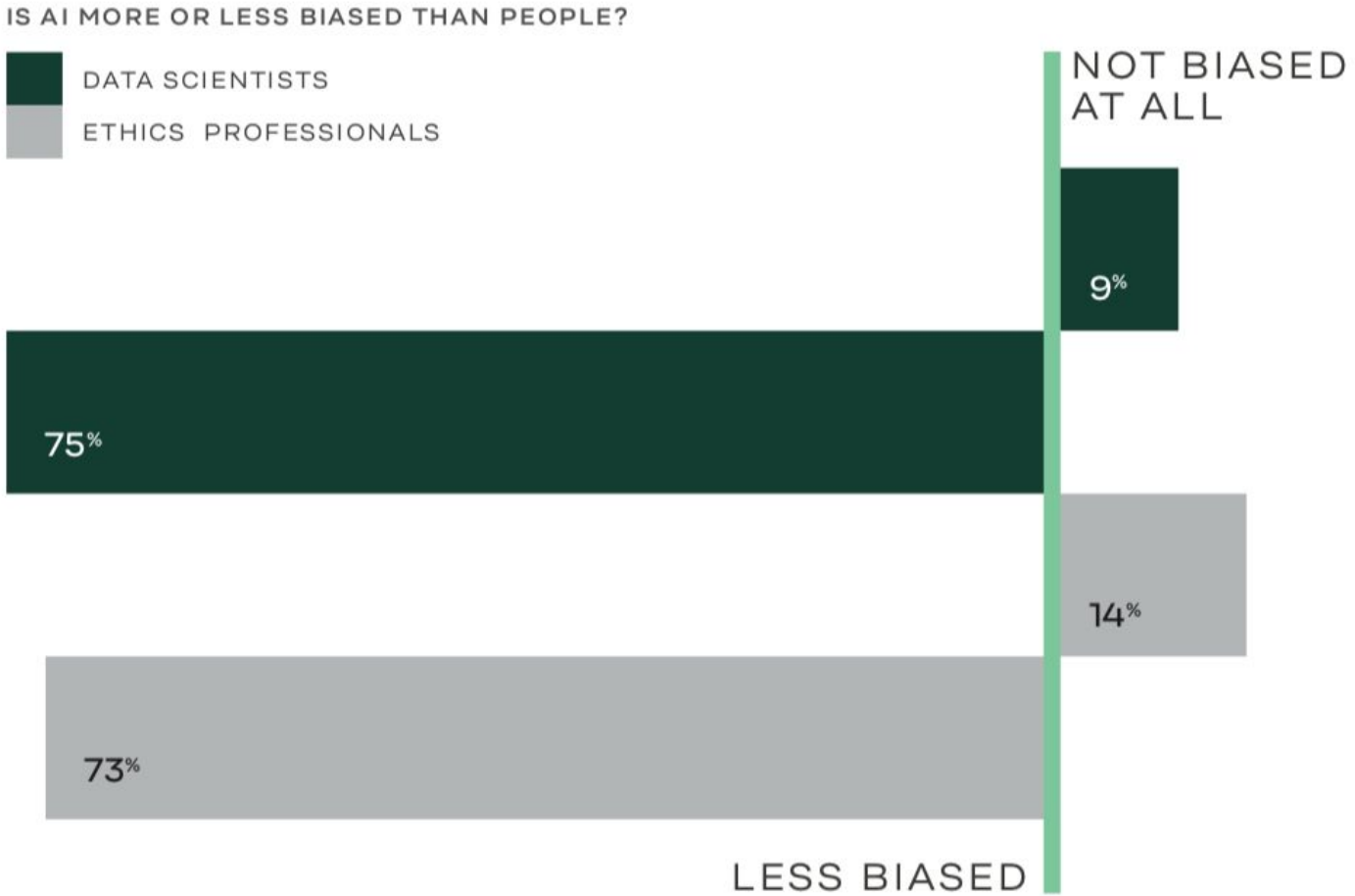


Source: CrowdFlower, *Data Science Report* (2017)





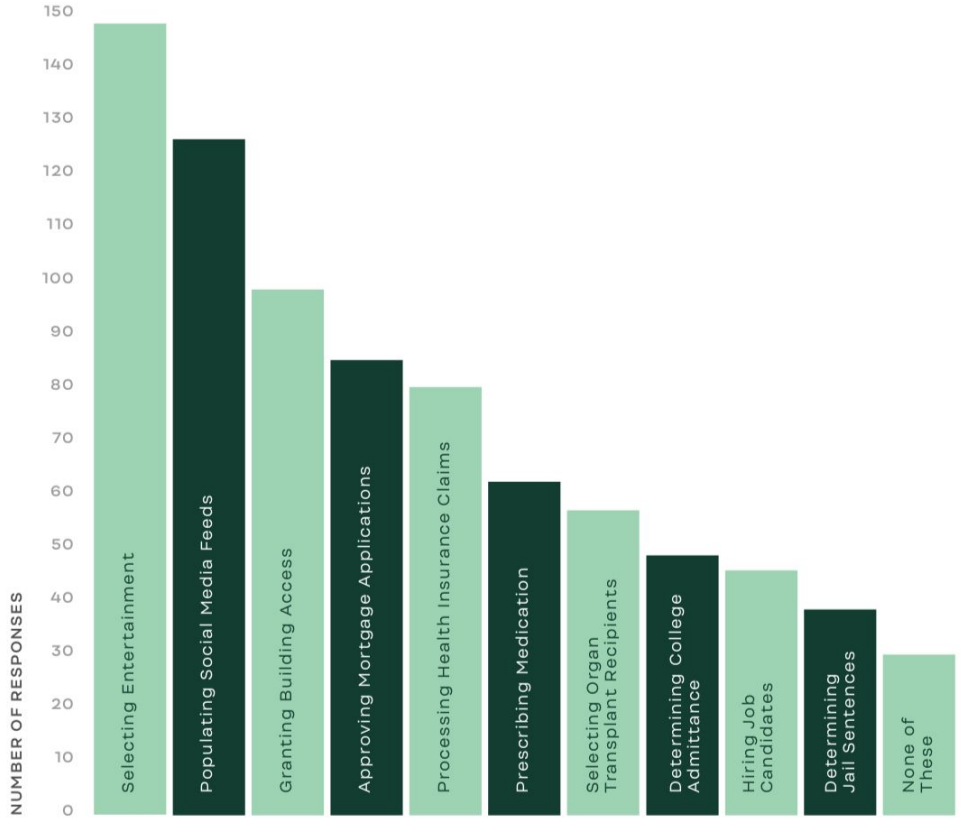
Source: Figure Eight, *Data Science Report* (2018)



Source: Figure Eight, *Data Science Report* (2018)

ETHICS: AI DECISION MAKING

IN WHICH OF THE FOLLOWING SCENARIOS WOULD IT BE APPROPRIATE FOR AI TO MAKE DECISIONS WITHOUT HUMAN INTERACTIONS?



Source: Figure Eight, *Data Science Report* (2018)

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

Thanks for taking COGS 108!