

# The Future of Data Science

Shannon E. Ellis, Ph.D  
UC San Diego



Department of Cognitive Science  
[sellis@ucsd.edu](mailto:sellis@ucsd.edu)

**What you all have done**

---

# COGS 108: What we've learned

01: Data Science & Ethics

02: Version Control & Python

03: Data & Data Wrangling

04: Data Visualization & EDA

05: Inference

06: Text Analysis

07: Machine Learning

08: Non-parametric Statistics & Geospatial Analysis

09: Dimensionality Reduction

10: DS Jobs & Communication

Guest Lectures: Vivian Peng & Tyler Richards

# 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: Atman, Ganesh, Sidharth

IAs: Abby, Andrew, Emily, Fei, & Michael

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

---

# 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!**

---

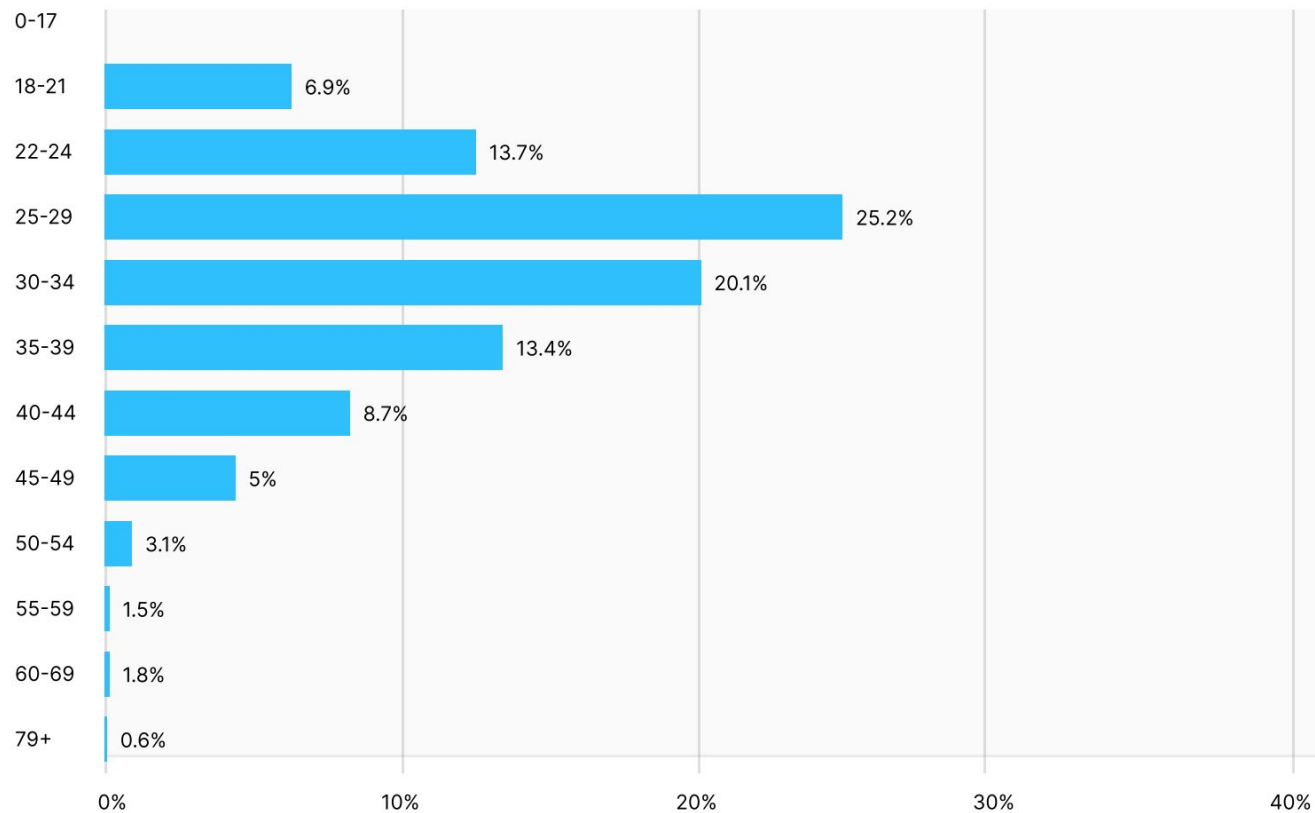


**Where we are now**

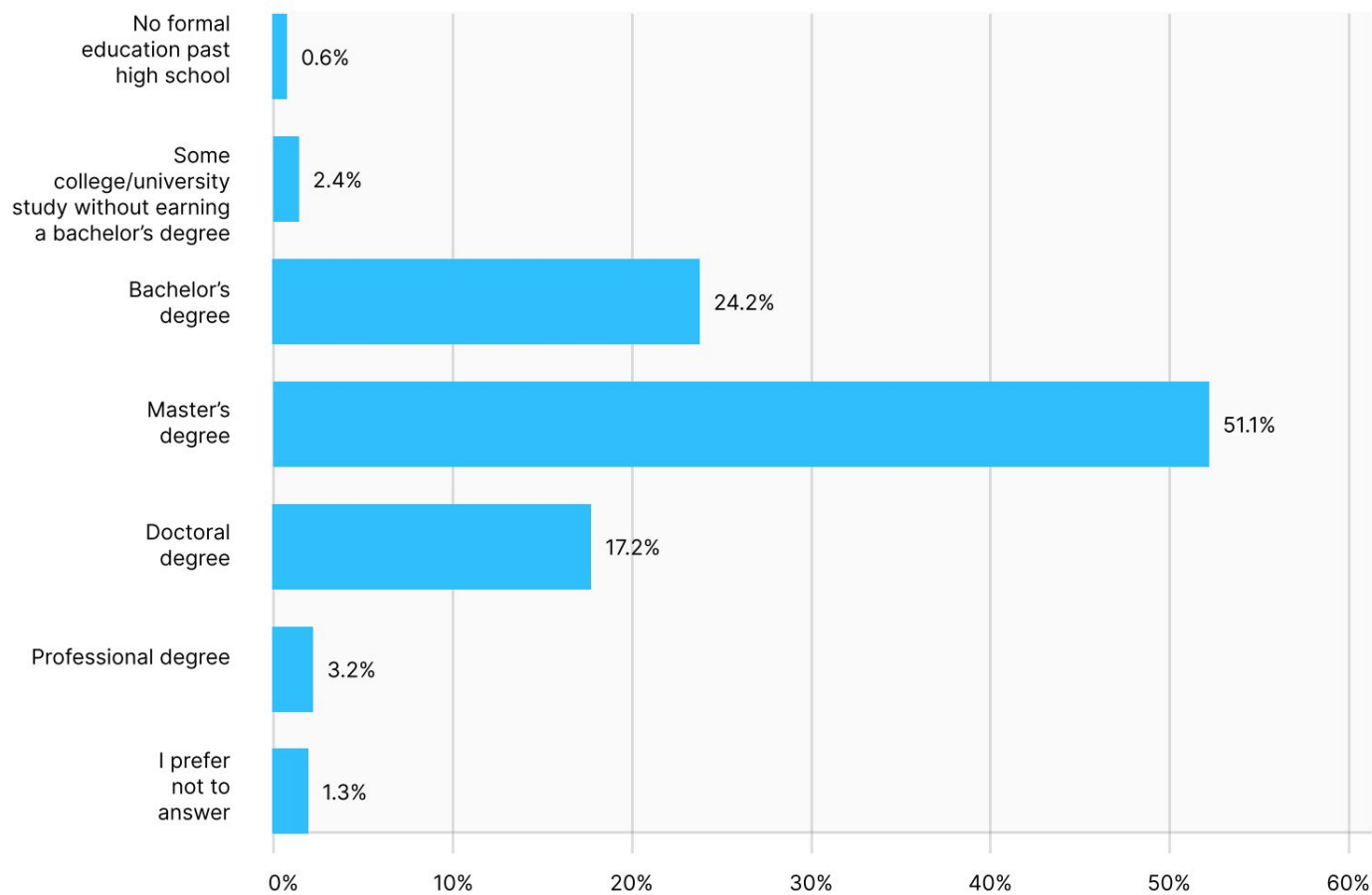
---

Rank	Job Title	Median Base Salary	Job Satisfaction	Job Openings
1	Front End Engineer	\$105,240	3.9	13,122
2	Java Developer	\$83,589	3.9	16,136
3	Data Scientist	\$107,801	4.0	6,542

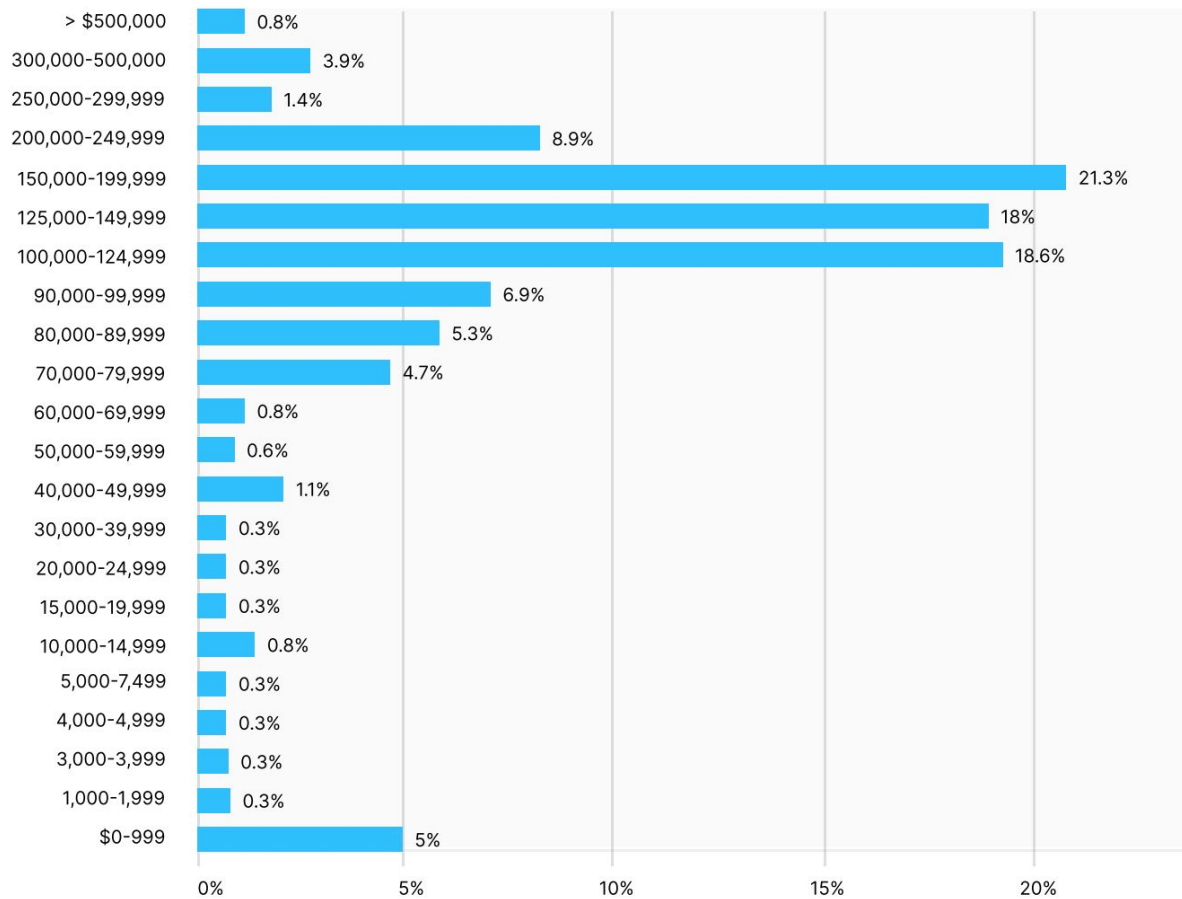
## AGE RANGES OF DATA SCIENTISTS



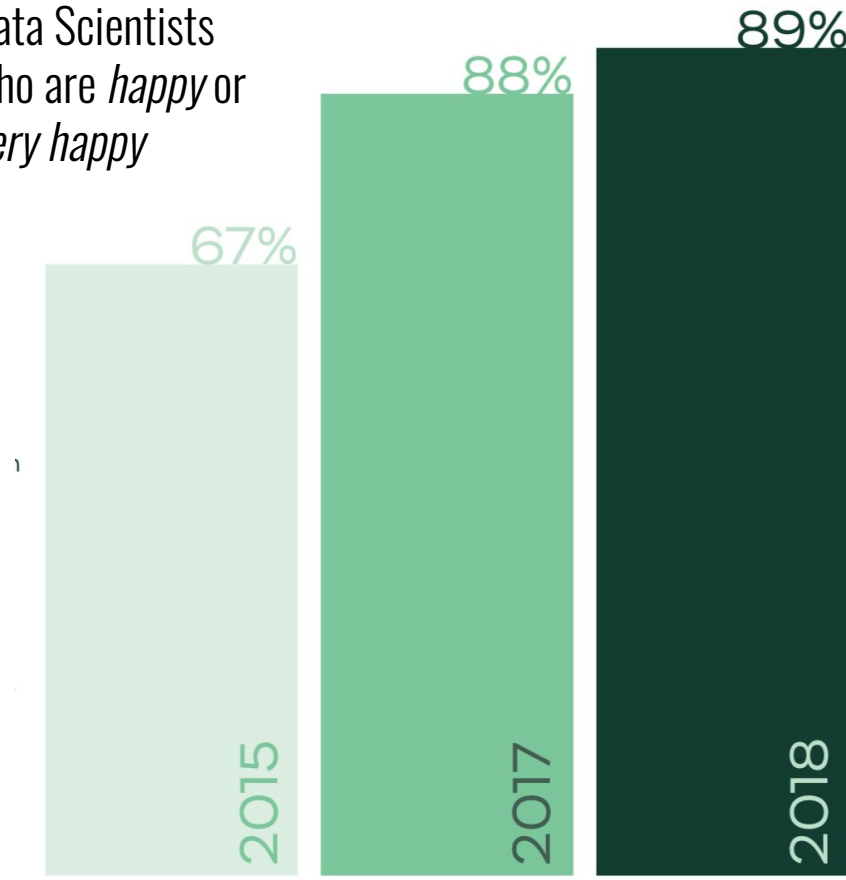
## EDUCATION LEVEL OF KAGGLE DATA SCIENTISTS

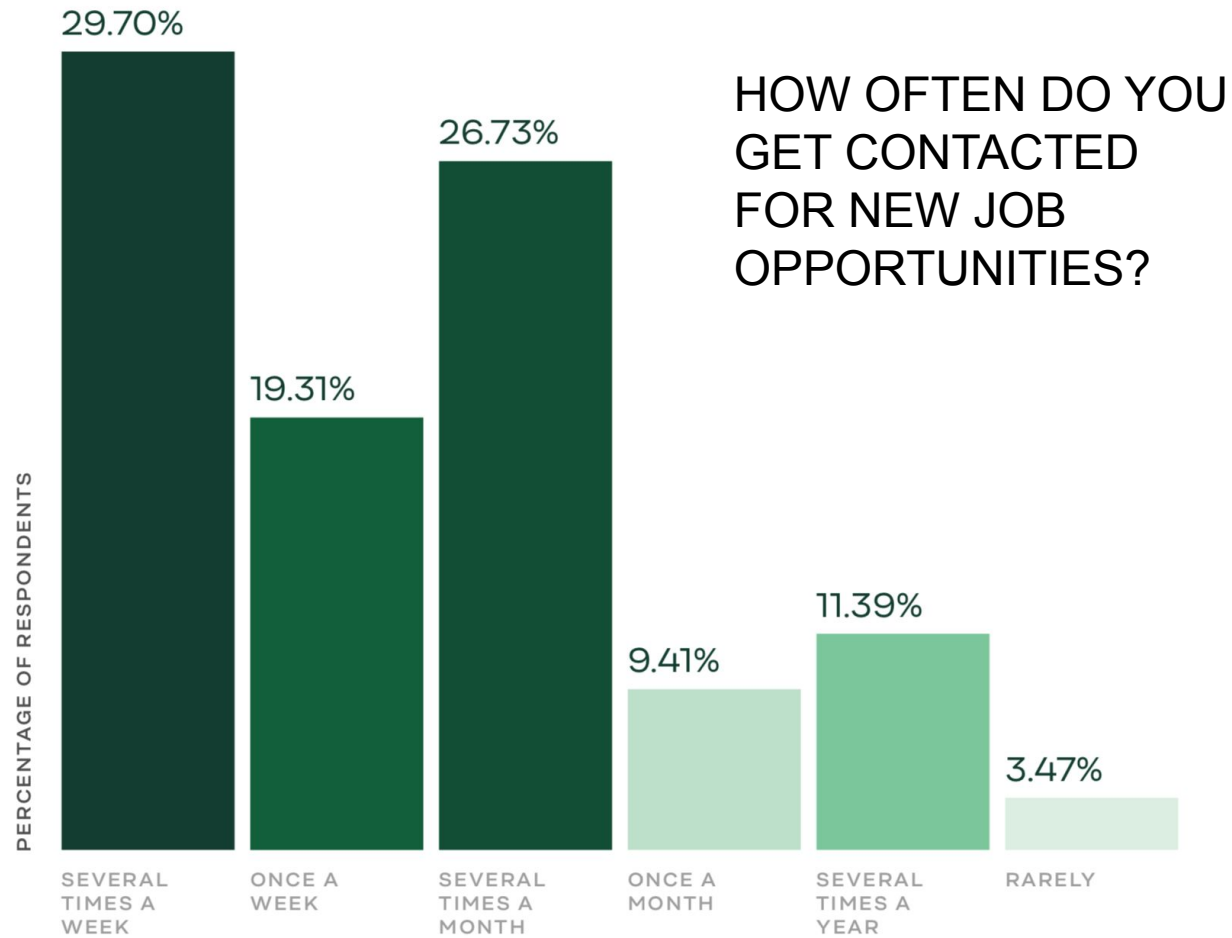


# SALARY DISTRIBUTION FOR US-BASED DATA SCIENTISTS



Data Scientists  
who are *happy* or  
*very happy*





## The Ten Most Common Data Science Skills in Job Postings

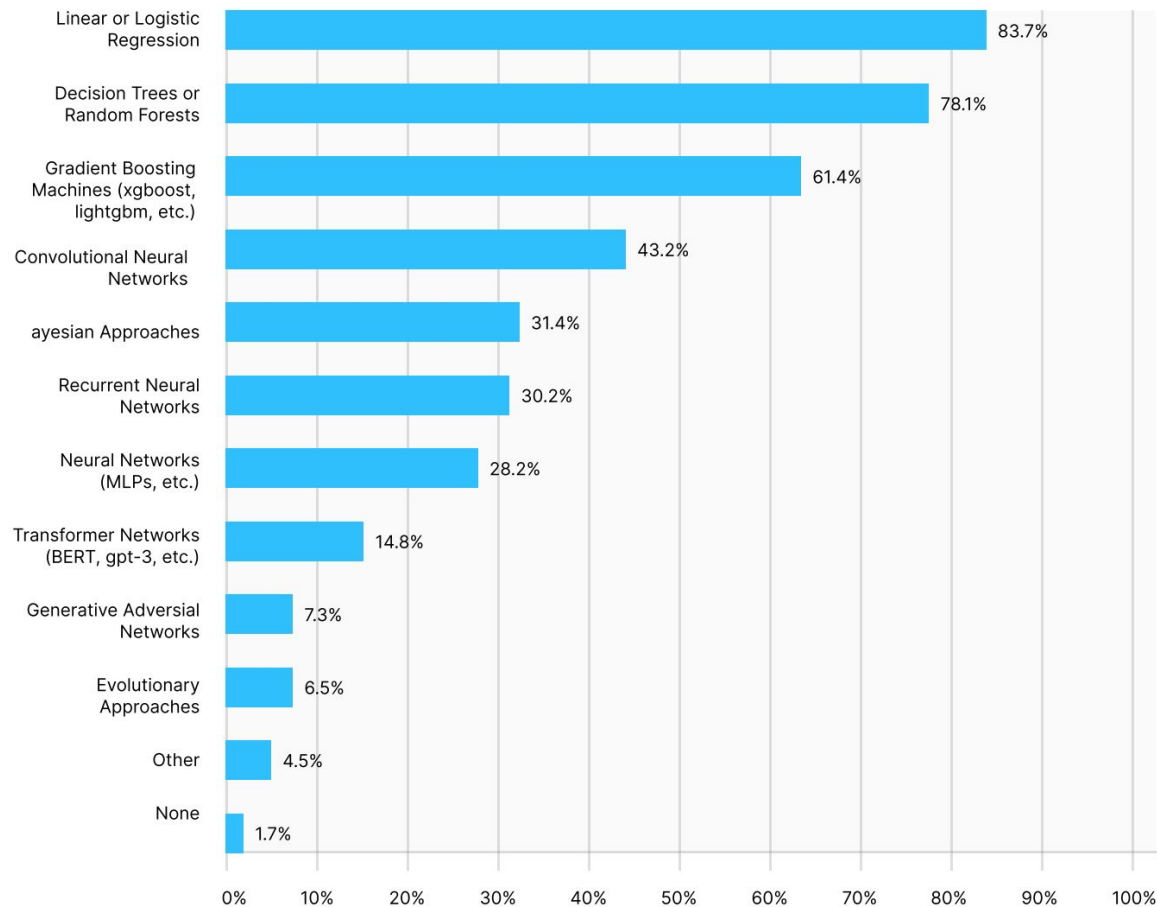
Skill	Percentage of Job Listings
Python	72%
R	64%
SQL	51%
Hadoop	39%
Java	33%
SAS	30%
Spark	27%
Matlab	20%
Hive	17%
Tableau	14%

Source: Glassdoor Economic Research.

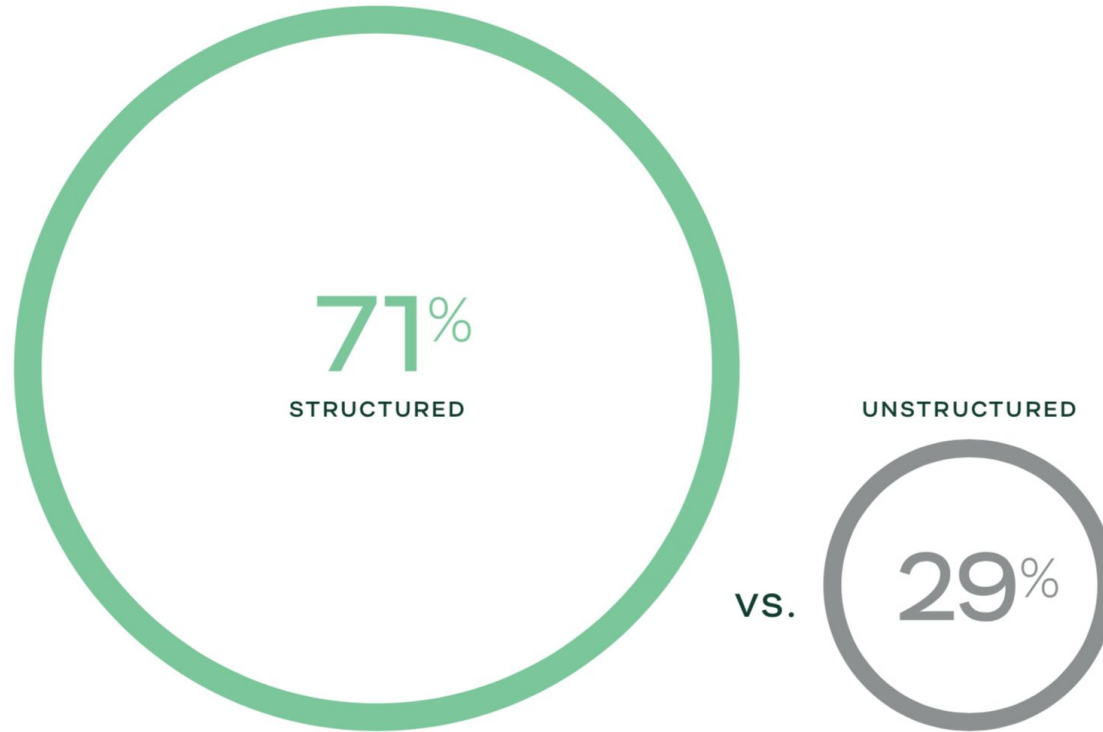
**glassdoor**

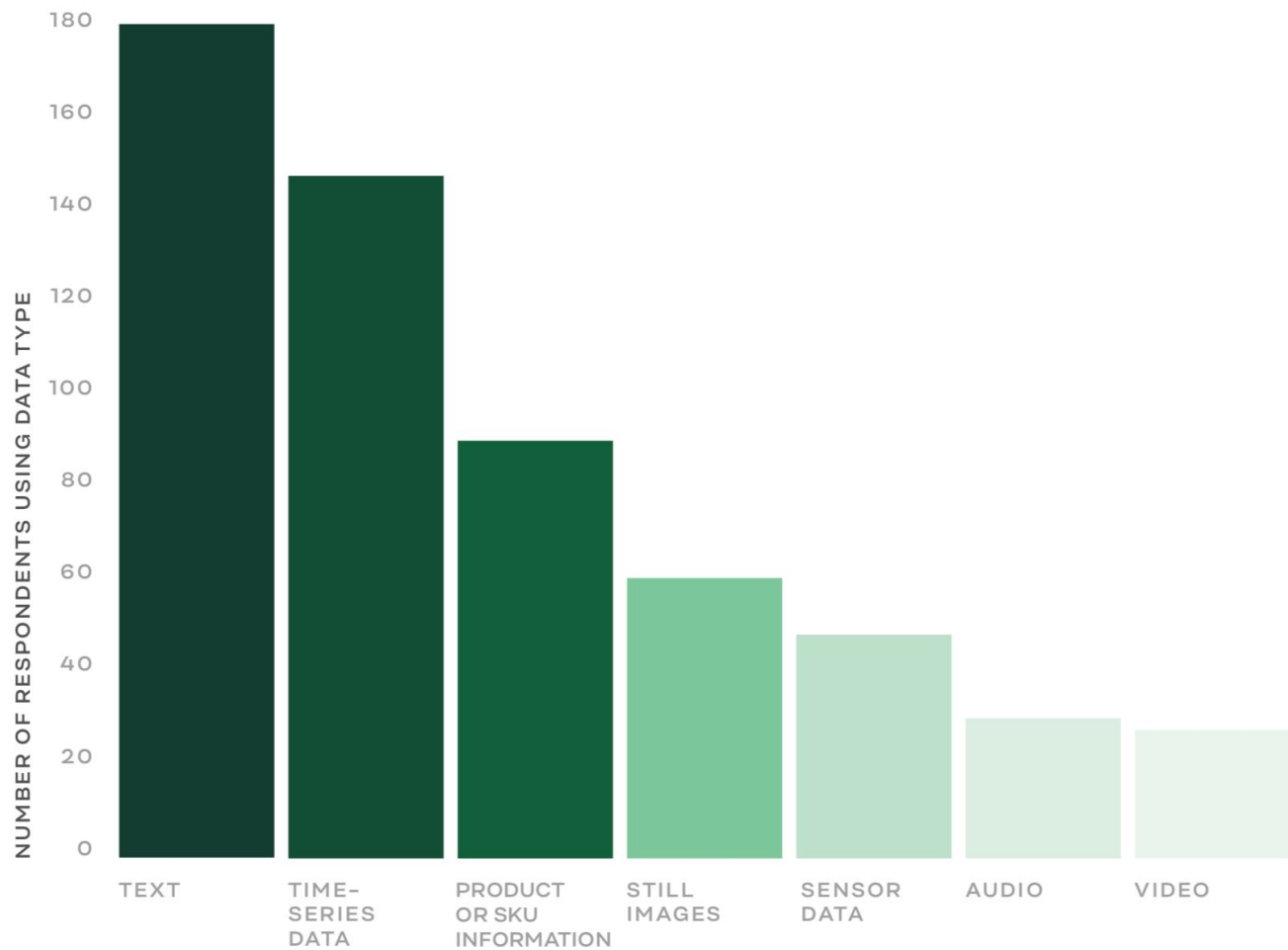


## METHODS AND ALGORITHMS USAGE



DO YOU WORK PRIMARILY  
WITH STRUCTURED OR  
UNSTRUCTURED DATA?





# 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](#).

# 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.

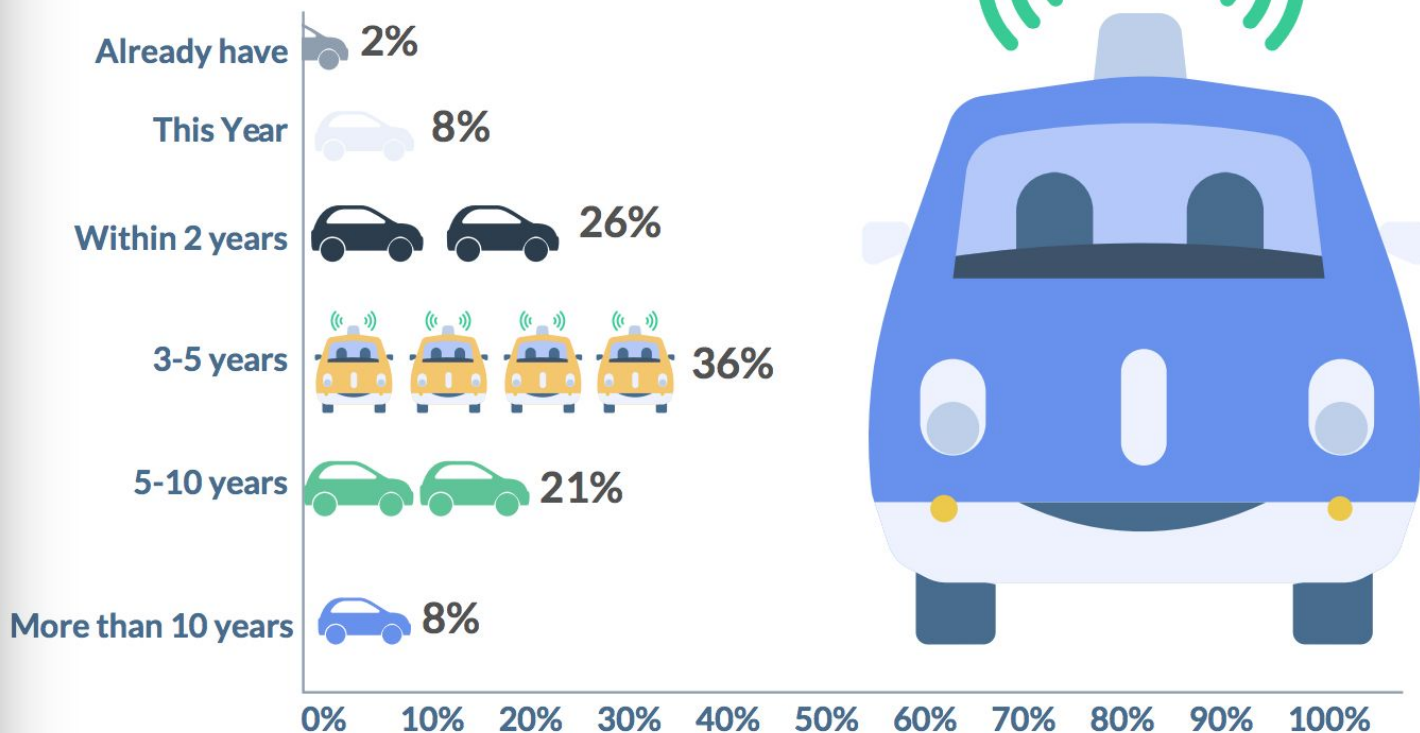


Hard things are hard.

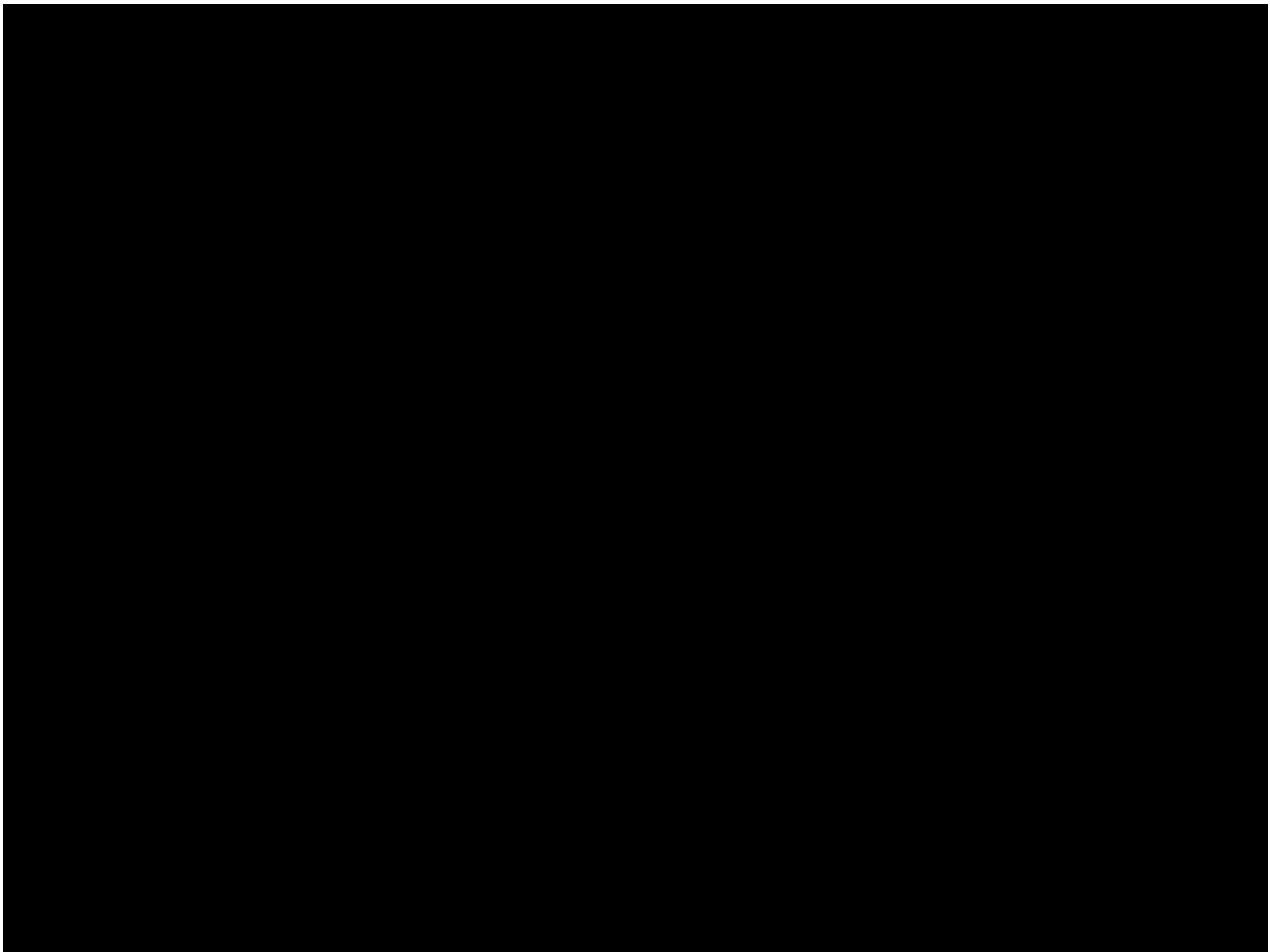
**...so where are we going?**

---

## When do you think you'll first ride in a SELF-DRIVING CAR?









**Algorithms are fragile**

---

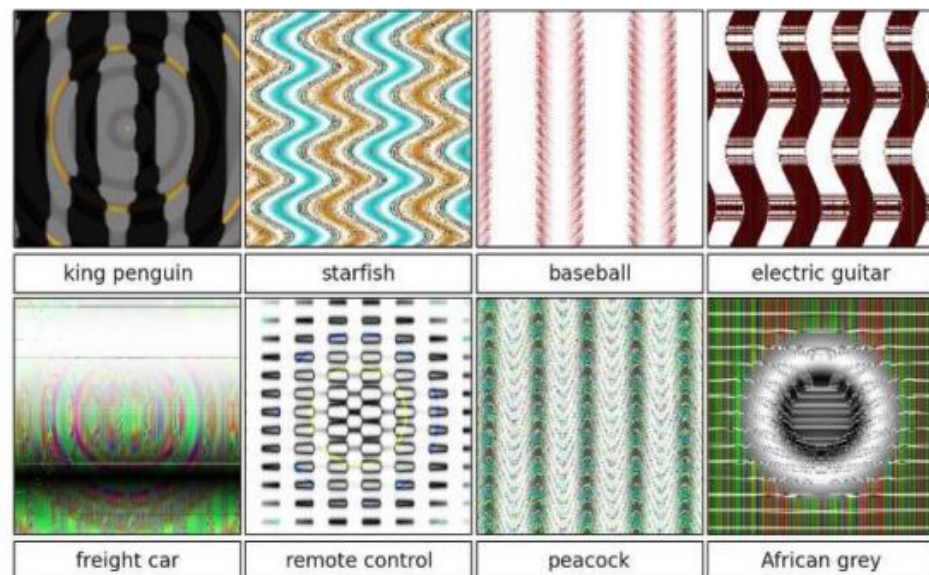
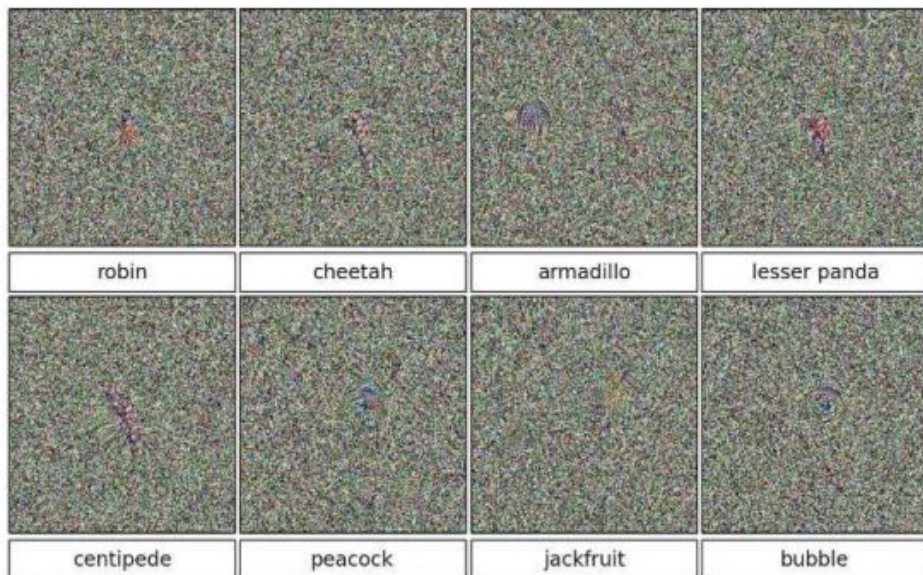


Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with  $\geq 99.6\%$  certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly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

# Trading program sparked May 'flash crash'



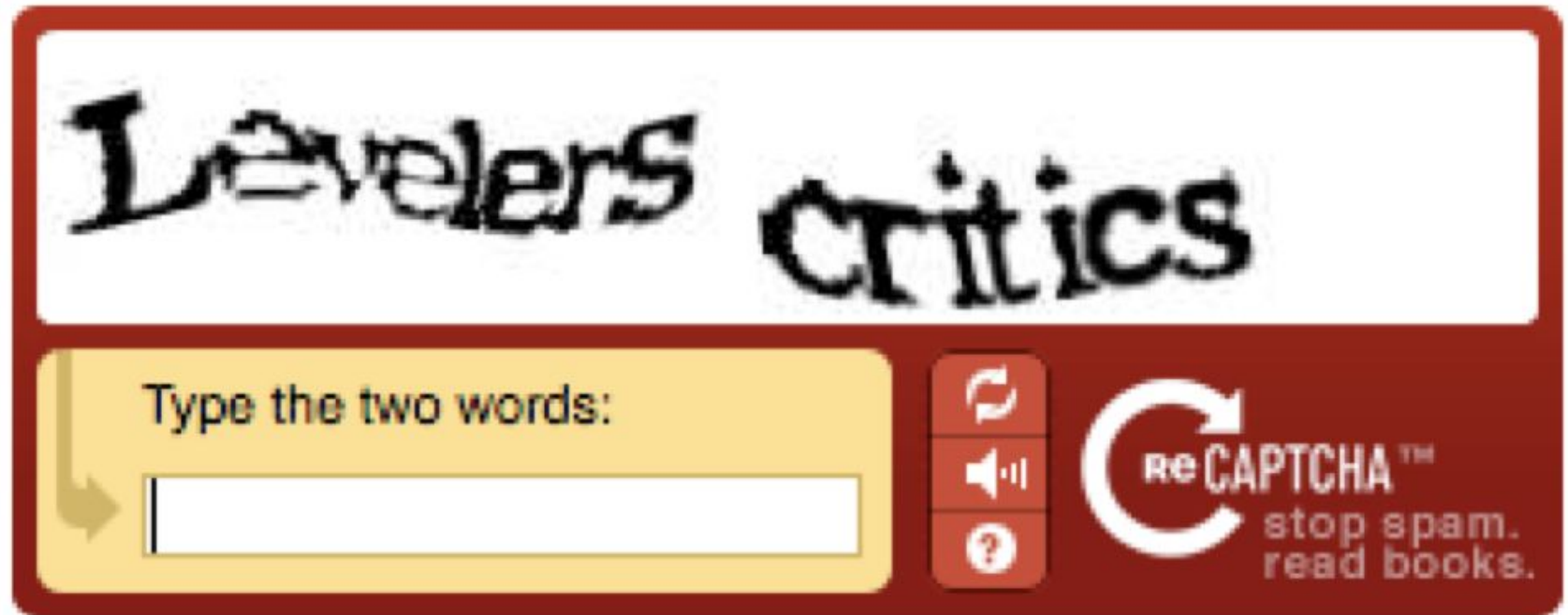
Automatic computerized traders on the stock market shut down as they detected the sharp rise in buying and selling. (NYT)

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

**Algorithms are fragile & powerful**

---

# Human-based computation



Levelers critics

Type the two words:

reCAPTCHA™  
stop spam.  
read books.

The image shows a reCAPTCHA interface. At the top, a white rectangular area contains a distorted, low-resolution image of the words "Levelers" and "critics" in a black, serif font. Below this, a yellow rectangular box contains the text "Type the two words:" followed by a white text input field. To the right of the input field is a vertical stack of three red buttons: the top button has a circular arrow icon, the middle button has a speaker icon, and the bottom button has a question mark icon. Further to the right is the reCAPTCHA logo, which consists of a large white 'C' with an arrow, followed by the text "reCAPTCHA™" and "stop spam. read books." below it.



Unfortunately, it would be extremely expensive.





Unfortunately, it would be extremely expensive.

**From the amusing...**

---

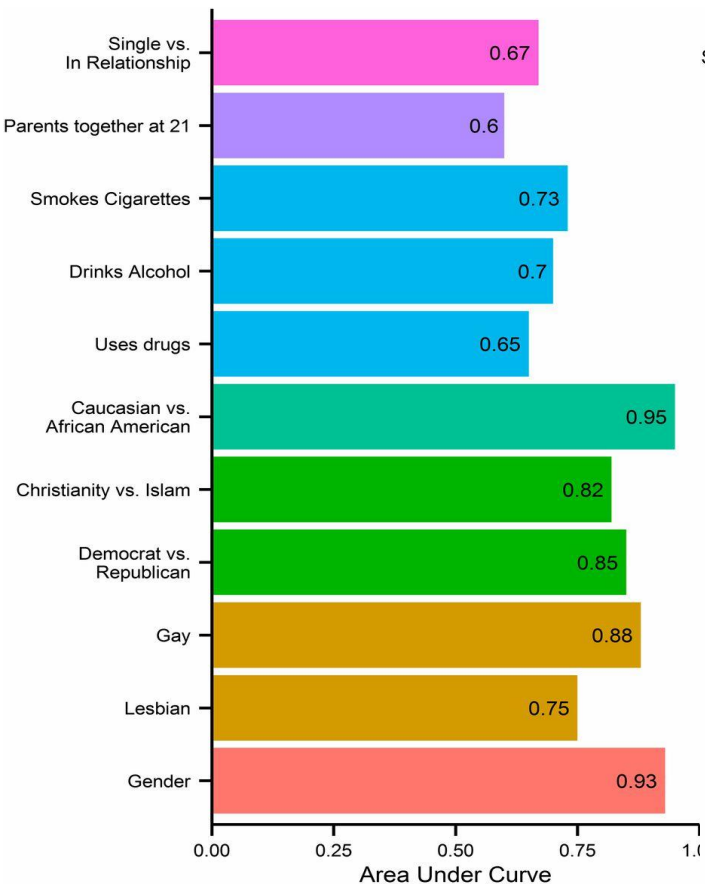


Fig 2: Prediction accuracy of classification of dichotomous/dichotomized attributes expressed by the AUC

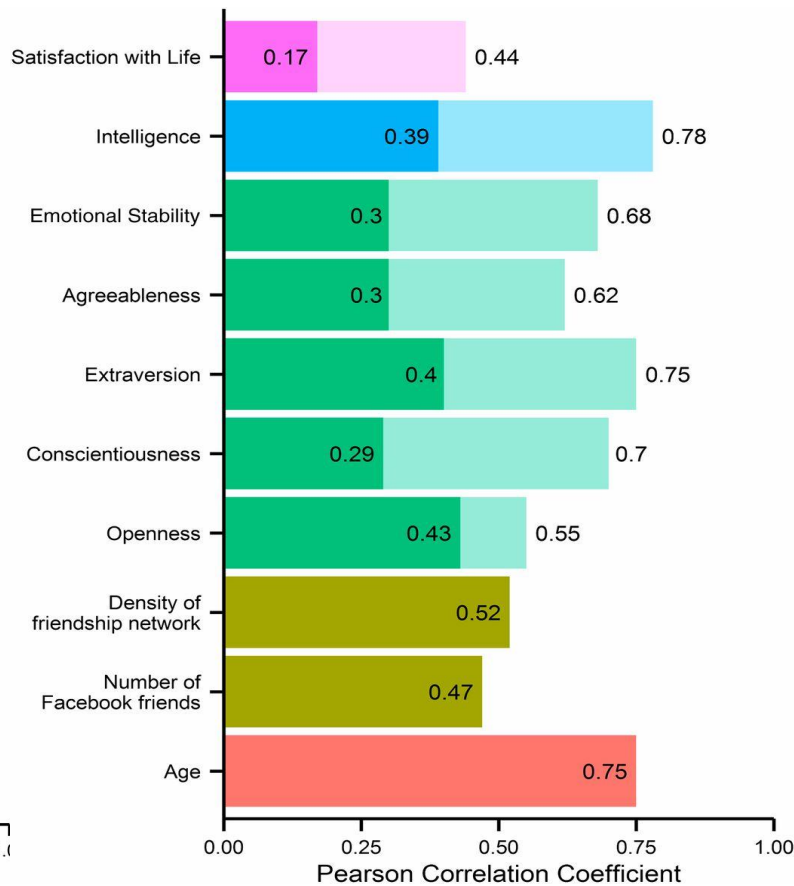


Fig 3: Prediction accuracy of regression for numeric attributes and traits expressed by the Pearson correlation coefficient between predicted and actual attribute values; all correlations are significant at the  $P < 0.001$  level. The transparent bars indicate the questionnaire's baseline accuracy,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

### Predictive Power of Likes.

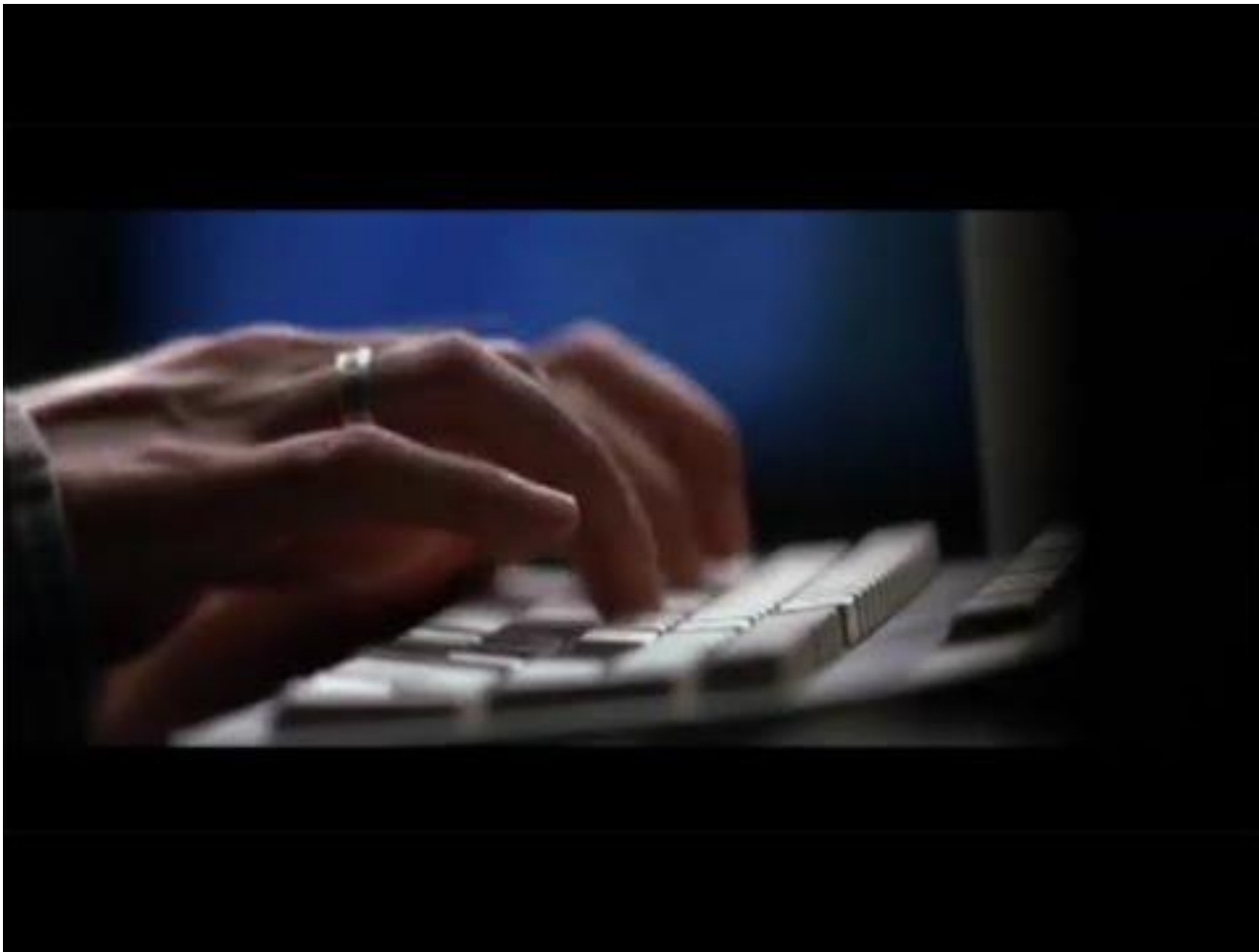
Individual traits and attributes can be predicted to a high degree of accuracy based on records of users' Likes. [Table S1](#) presents a sample of highly predictive Likes related to each of the attributes. For example, the best predictors of high intelligence include "Thunderstorms," "The Colbert Report," "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.

**From the amusing...**

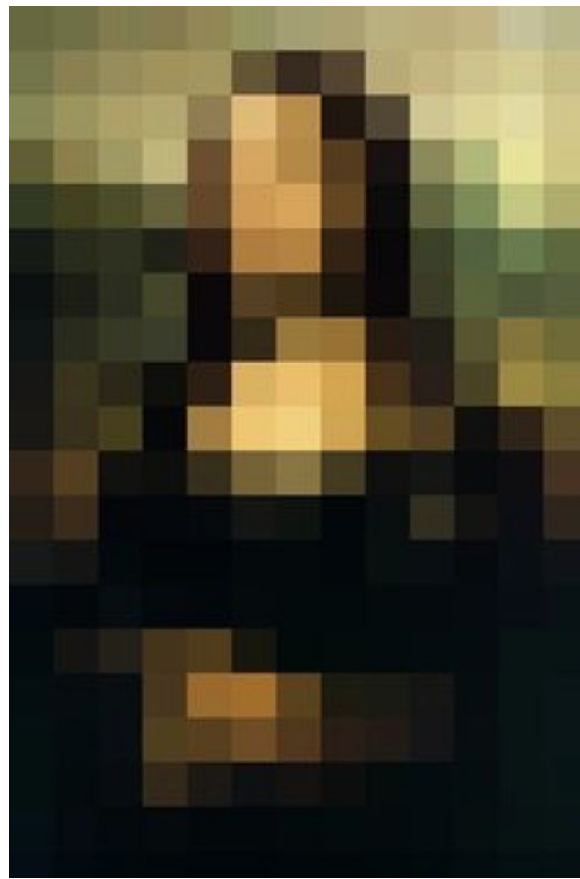
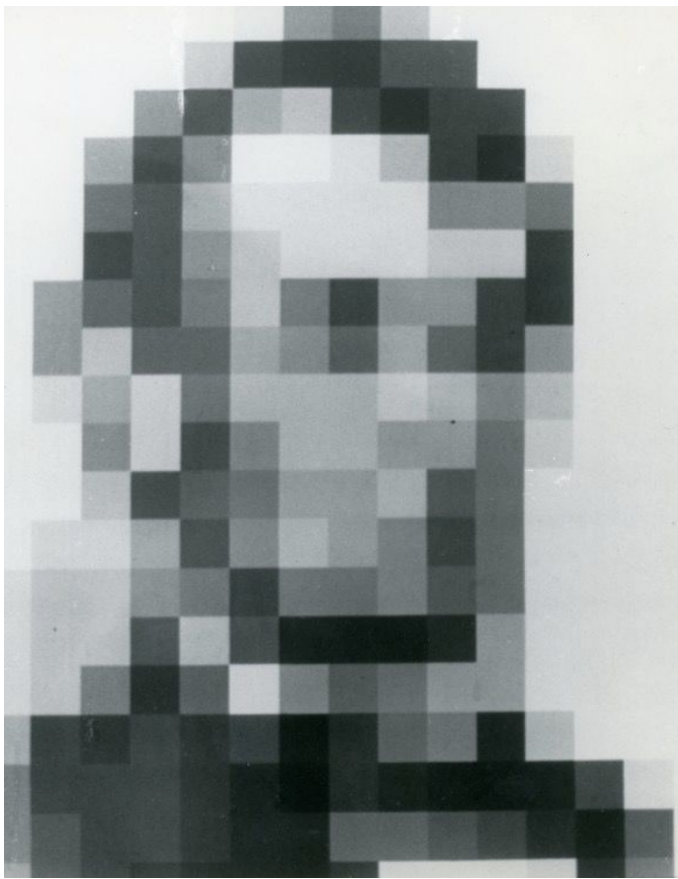
**...to the amazing**

---





[https://www.youtube.com/watch?v=Vxq9vi2pVWk&list=RDVxq9vi2pVWk&start\\_radio=1](https://www.youtube.com/watch?v=Vxq9vi2pVWk&list=RDVxq9vi2pVWk&start_radio=1)

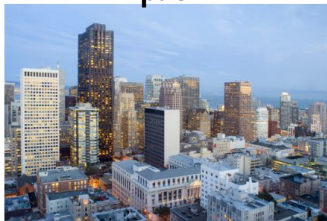




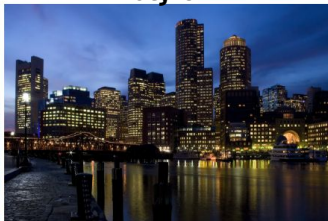
“...the first column is the 16x16 input image, the second one is what you would get from a standard bicubic interpolation, the third is the output generated by the neural net, and on the right is the ground truth.”



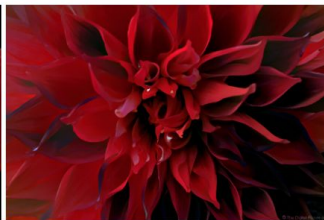
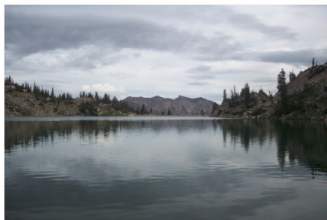
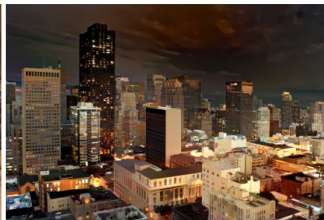
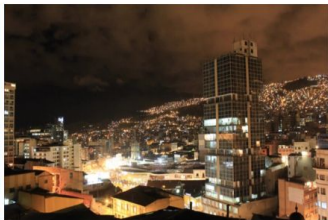
input

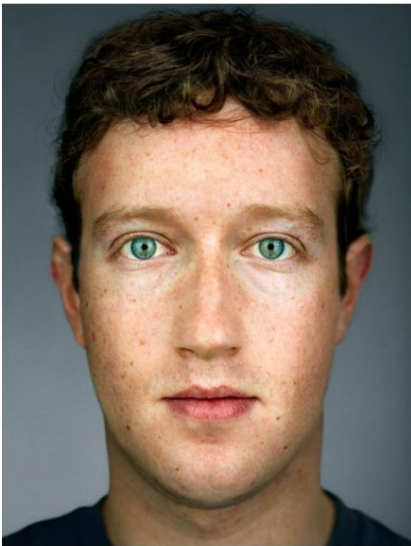
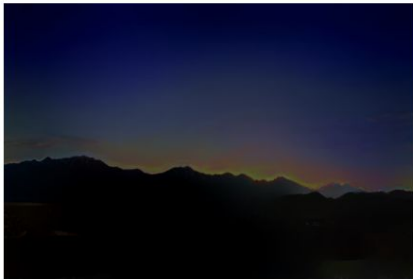
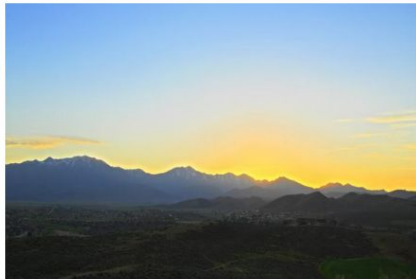


style

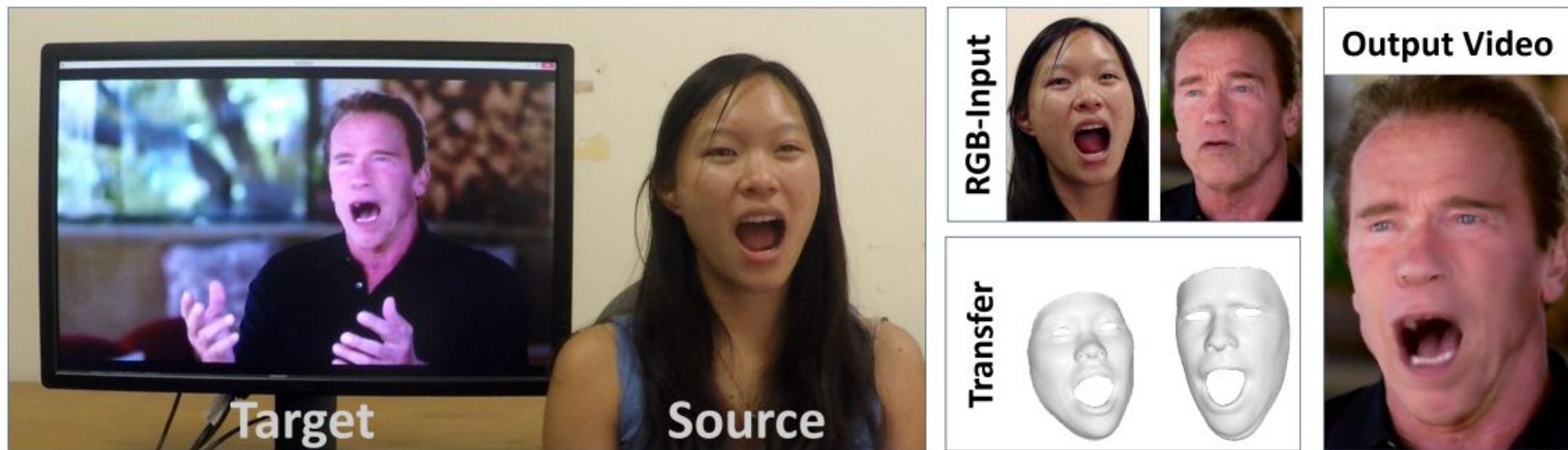


output



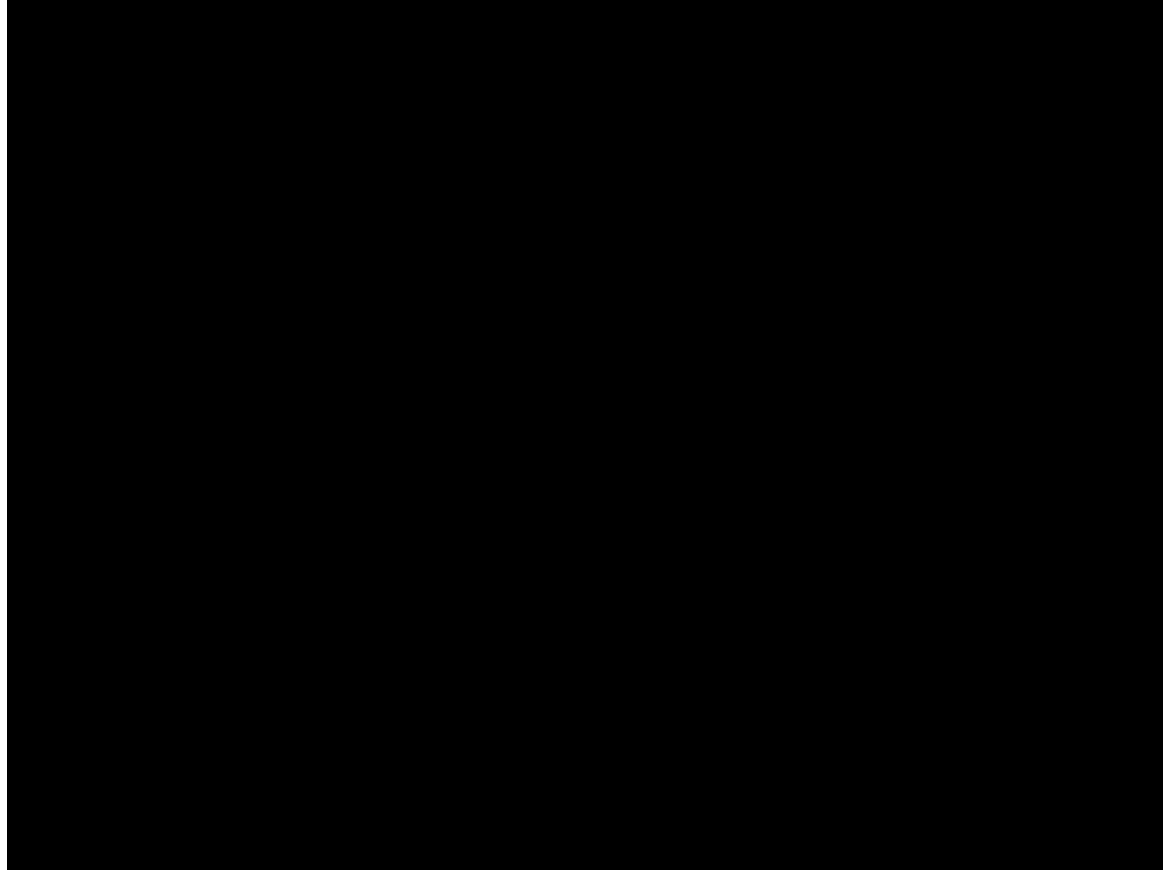


# Reality manipulation



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





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?

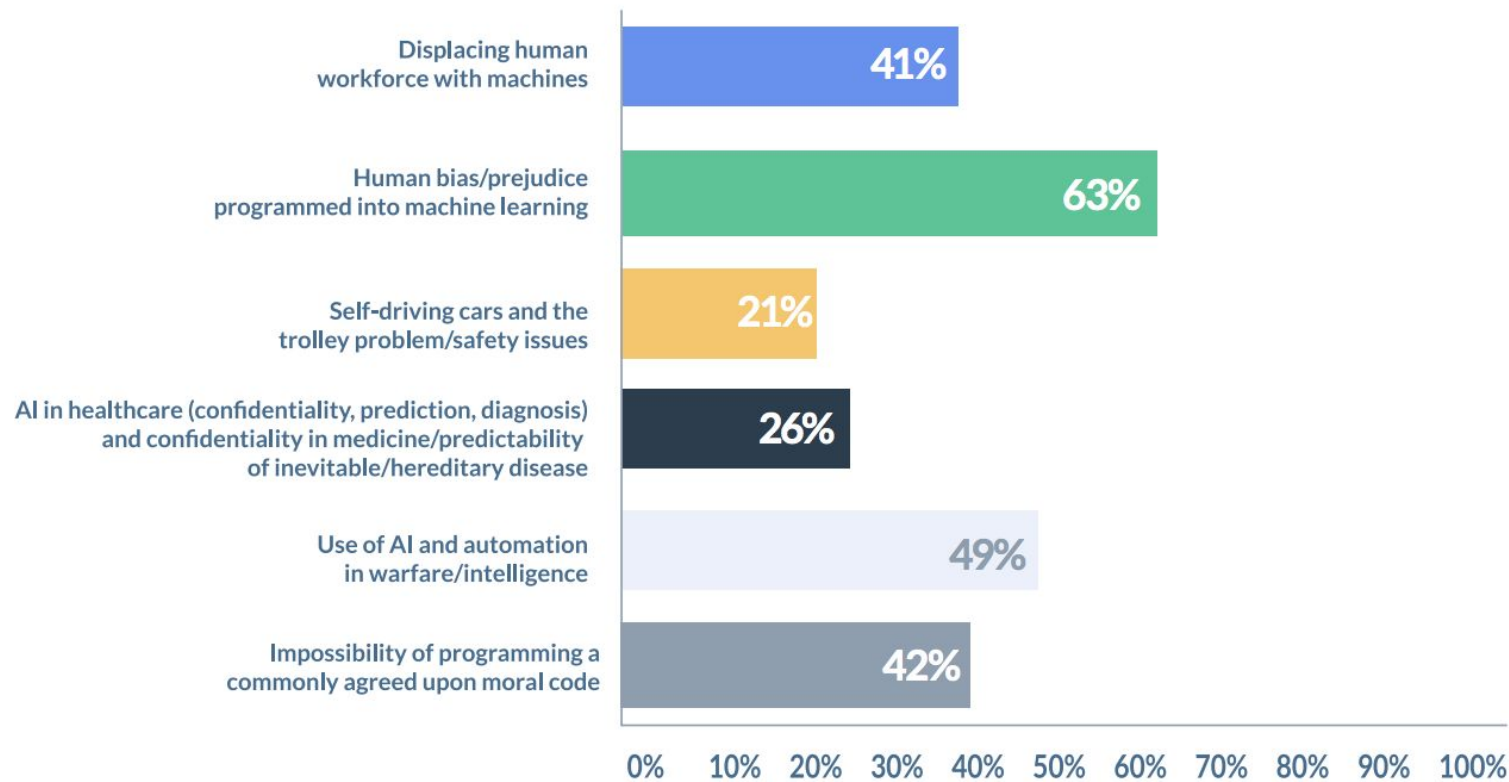


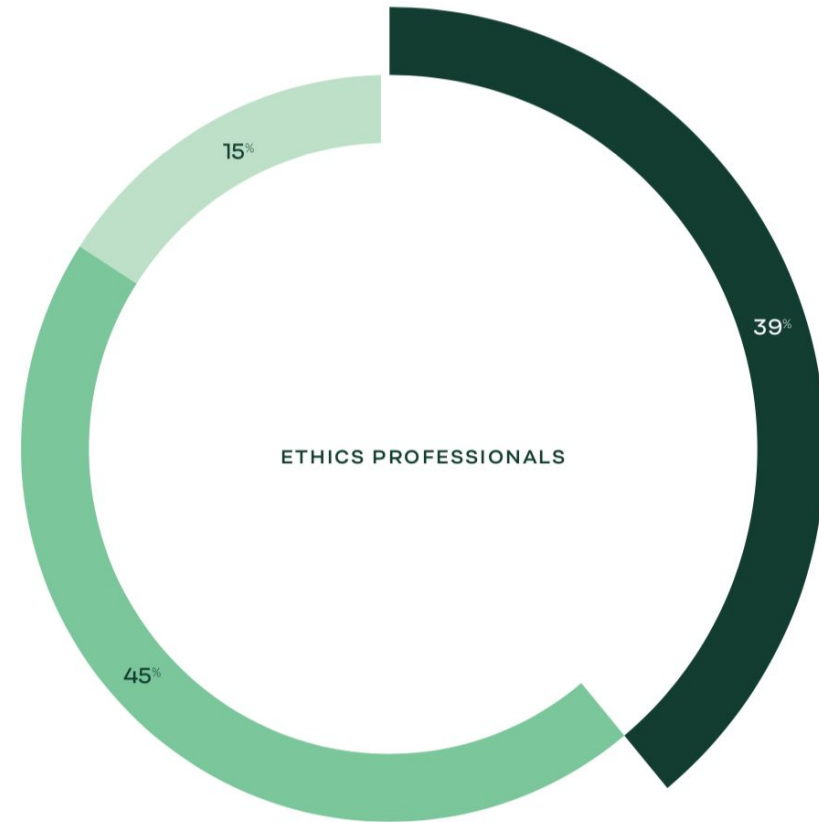
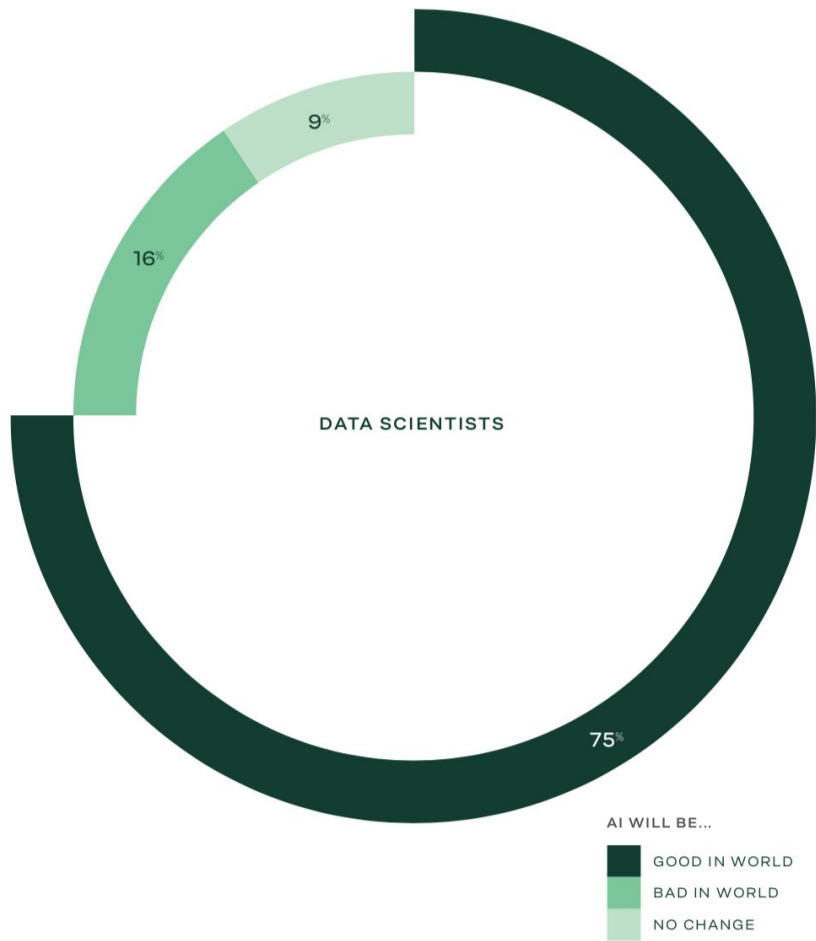
**Nicole Martin** Contributor ⓘ

[AI & Big Data](#)

*I write about technology, data and privacy.*

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







IS AI MORE OR LESS BIASED THAN PEOPLE?



DATA SCIENTISTS

ETHICS PROFESSIONALS

NOT BIASED  
AT ALL

9%

75%

14%

73%

LESS BIASED

## ETHICS: AI DECISION MAKING

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

