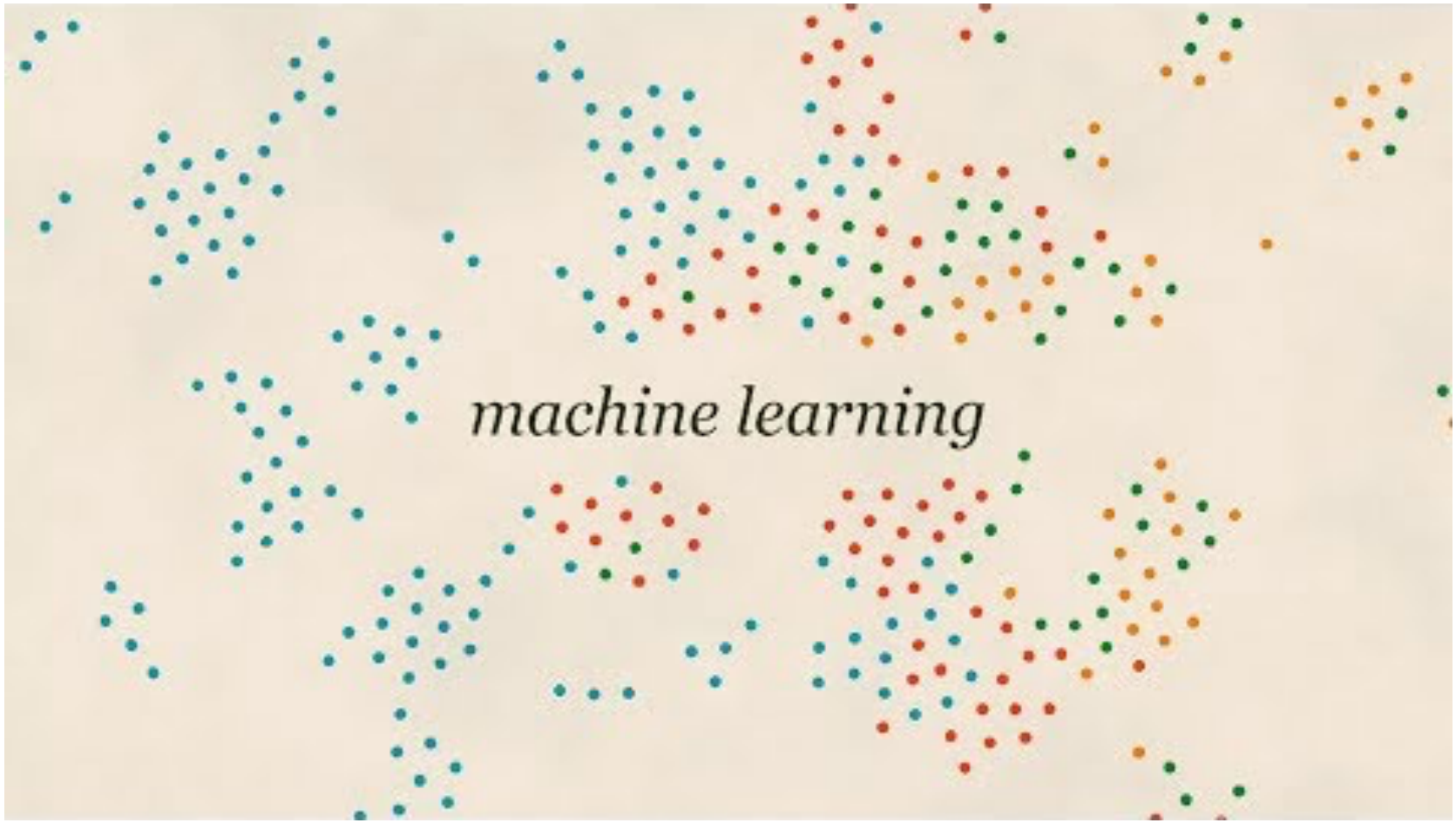

machine learning

Let’s discuss why…

Ethics

João Sedoc

Johns Hopkins University

December 4, 2019

Some content from E. Heizer & A. Roth & Z. Ives and L. Ungar & Yulia Tsvetkov and Alan Black

A problem for AI and Ethics

Consider the following rudimentary ethical questions about AI:

- What should the ultimate good of AI?
- What makes an AI innovation good vs. bad in a moral sense?
- How should AI function such that it promotes its ultimate good?

Problems:

- We're building artificial intelligence that is increasingly taking on the role of thought partner, information broker, medical expert, and social engineer
- There are no robust frameworks for evaluating the ethics of AI
- Industry won't figure this out for us

“Automated” decisions impact every aspect of our lives

- Admission to schools
- Who to hire and who to fire
- Work schedule
- Who to date
- Whether to grant a loan
- What ads are shown, discounts are given
- ….

PREDICTIVE POLICING: USING MACHINE LEARNING TO DETECT PATTERNS OF CRIME

DECISION MAKING

Want Less-Biased Decisions? Use Algorithms.

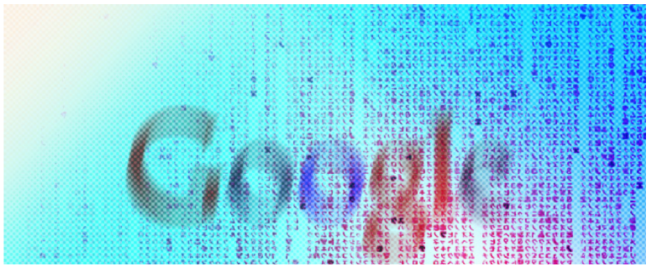
by Alex P. Miller

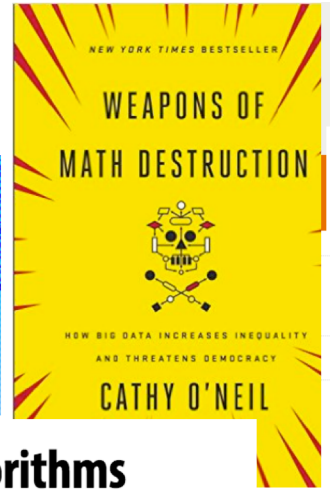
July 26, 2018

Online Ads for High-Paying Jobs Are Targeting Men More Than Women

New study uncovers gender bias

By Garett Sloane | July 7, 2015





Do Google's 'unprofessional hair' results show it is racist?

Leigh Alexander

When it Comes to Policing, Data Is Not Benign

OPINION BY PRESTON GRALLA

Amazon Prime and the racist algorithms

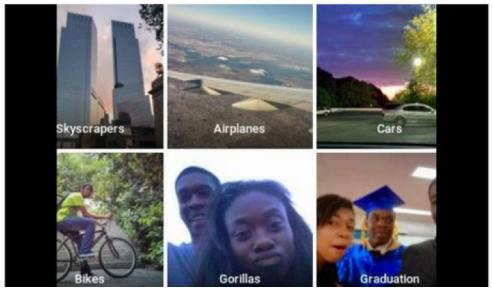
The company's algorithms told it where to offer its Prime Free Same-Day Delivery service, but an algorithm that uses data tainted by racism will be racist in its out

Facebook's Bias Is Built-In, and Bears Watching

Google apologises for Photos app's racist blunder

1 July 2015 | Technology

Share



When Algorithms Discriminate

Who's a CEO? Google image results can shift gender biases

UNIVERSITY OF WASHINGTON

SHARE PRINT E-MAIL

IMAGE: PERCENTAGE OF WOMEN IN TOP 100 GOOGLE IMAGE SEARCH RESULTS FOR CEO IS: 11 PERCENT. PERCENTAGE OF US CEOS WHO ARE WOMEN IS: 27 PERCENT. view more ›

ecm JULY 9, 2015

shaped by forces beyond our control, determining the Facebook, the people we meet on OkCupid and the e on Google. Big data is used to make decisions about ment, housing, education and policing.

rograms be discriminatory?

The Long History of Ethics and AI

Well I'm just an engineer?

You need to understand the ethical issues surrounding the data you obtain/use, the algorithms you employ, and its impact on people.

The Dual Use of A.I. Technologies

- Who should be responsible?
  - The person who uses the technology?
  - The researcher/developer?
  - Paper reviewers?
  - University?
  - Society as a whole?

We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

Deepfakes

Misconceptions about AI that can lead to harm

- Engineers are only responsible for their code, not how the code is used or the data quality
- Humans and computers are interchangeable; replacing humans with computers results in better outcomes
- Regulating the tech industry is too hard and won't be effective
- Our job in tech is just to optimize metrics and respond to customer demand

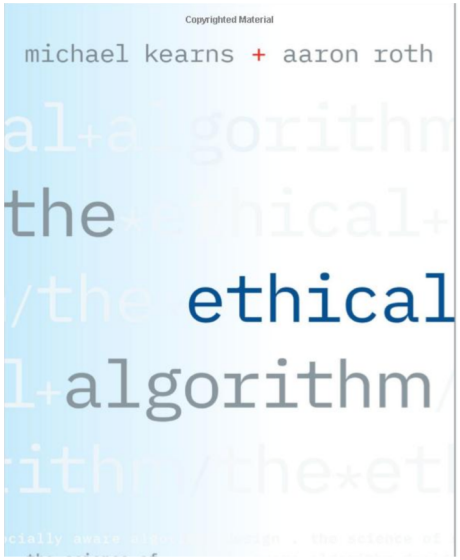
From Rachel Thomas FastAI

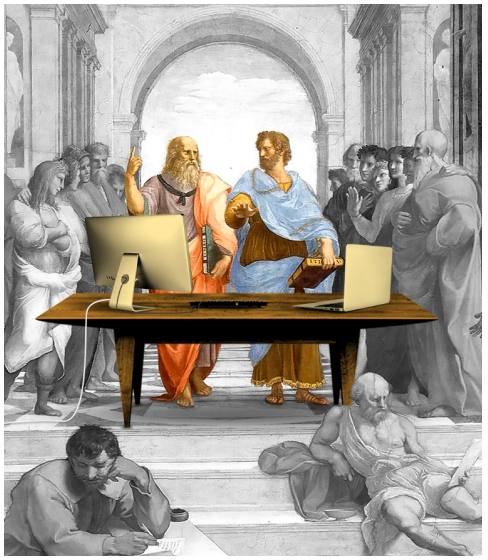
What we'll discuss...

- **What are ethics?**
- **Ethics surrounding data**
  - Privacy and informed consent
  - Ownership and intellectual property
- **Ethics surrounding algorithms**
  - Biased algorithms
  - Bad results from good data
    - reproducibility
  - Fairness

Virtue ethics for humans

- Virtue ethics offers an alternative to rule-based ethical systems (e.g., deontology, utilitarianism)
- Virtues are the qualities of people that promote human flourishing
- Virtue is attained by:
  - performing one's distinctive function well
  - cultivating intellectual and moral excellence
  - achieving proper inner states; i.e., those consistent with virtue
- Virtues may be cultivated based on learning and emulation





Algorithmic Virtue ???

Ethical Algorithms

Identifying algorithmic virtue

We want to make normative claims about what makes an algorithm good (and they kind of sound like Aristotle)

To be good, algorithms must *perform their function well* (intellectual virtue)

- *The distinctive function of a predictive algorithm is to predict correctly; the virtuous algorithm is one that attains excellence in prediction*

It is not enough that predictive algorithms make accurate predictions; to be good, they must arrive at their predictions *for the right – i.e., normative and just – reasons* (moral virtue)

- *An algorithm that makes an accurate prediction based on discriminatory/ prejudicial information acts contrary to virtue*

Need to *embed* social values in algorithms

- Requires being precise about the definitions, developing their consequences.
  - Privacy
  - Fairness
  - Accountability
  - Interpretability

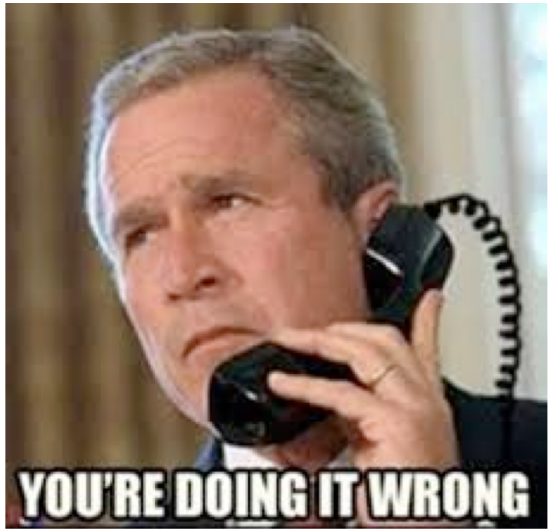
Ethics Surrounding Data

Ethical Principles

- **Autonomy**
  - The right to control your data, possibly via *surrogates*
- **Informed Consent**
  - You should explicitly approve use of your data based on understanding
- **Beneficence**
  - People using your data should do it for your benefit
- **Non-maleficence**
  - Do no harm

Data Collection

- Data is constantly being collected about us
  - Cameras
  - Location reporting
  - Accelerometers
  - Social media
- Do I own data collected about me?
- What if I don't like what the data says about me?
- Can I control how the data is used?



Data and Informed Consent

- In human subjects research, there is a notion of *informed consent*
  - must *understand* what is being done
  - must *voluntarily consent* to the experiment
  - must have the right to withdraw consent at any time
- Not required in "ordinary conduct of business"
  - E.g. A/B testing
  - But this is a very thin line….

Informed Consent

- Informed consent is often buried in the fine print
- Data is often collected first; the experiment comes later.
- How the data, once collected, is going to be used is difficult to control.

OKCupid Experiments with Customers

- **Love Is Blind day**
  - Suppressed photographs in user profiles so that the customer would not be able to see the profile photographs of potential people to meet.
  - Customers could still read what was written in profiles, e.g. interests.

- Impact on success of a date on being told you are compatible? In addition to reporting accurate compatibility scores:
  - Took people with low compatibility score, told them it was high
  - Took people with high compatibility score, told them it was low
  - Three by three test matrix: 30, 60 and 90% compatibility score actual versus 30, 60, and 90% declared

Was this legal/ethical?

- CEO Christian Rudder: "But guess what, everybody: if you use the Internet, you're the subject of hundreds of experiments at any given time, on every site. That's how websites work."
- On the other hand, having a company intentionally lie to you, intentionally give you a wrong score, is something that many people consider socially unacceptable.

Intellectual Property

- Artistic expression can be **copyrighted**: exclusive legal right to print, publish, perform, film or record and authorize others to do the same.
- **Derivative** work can be created with permission.
- There's also a notion of **citation**, in which we give credit to the owner.
- What about data?
  - Wikipedia, Yelp, Rotten Tomatoes, TripAdvisor

Intellectual Property - Copyrights

Neural Networks for Music Generation

Can we reproduce artists' creativity through AI?

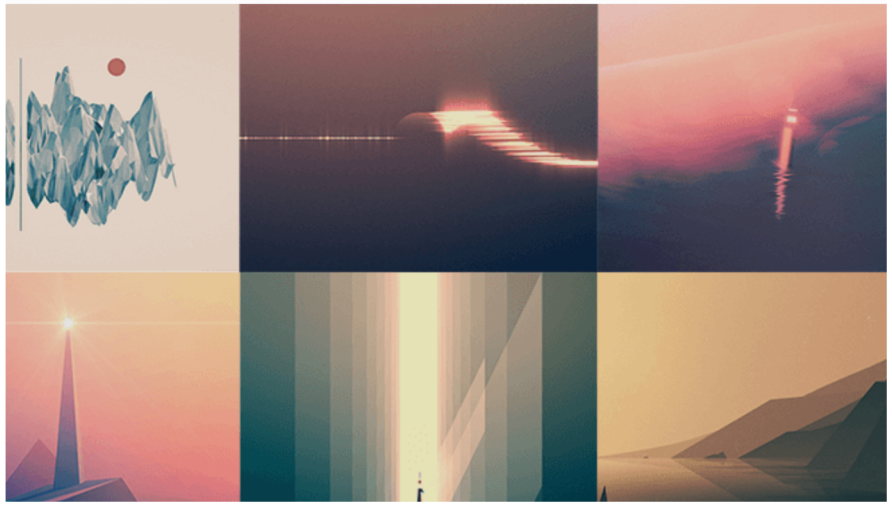
Andy Spezzatti Follow

Jun 24 · 13 min read ★

An exciting application of the recent advance in AI is Artificial Music Generation. Can we reproduce artists' creativity through AI? Can a Deep Learning model be an inspiration or a productivity tool for musicians? Those questions bring us to the definition of creativity and the

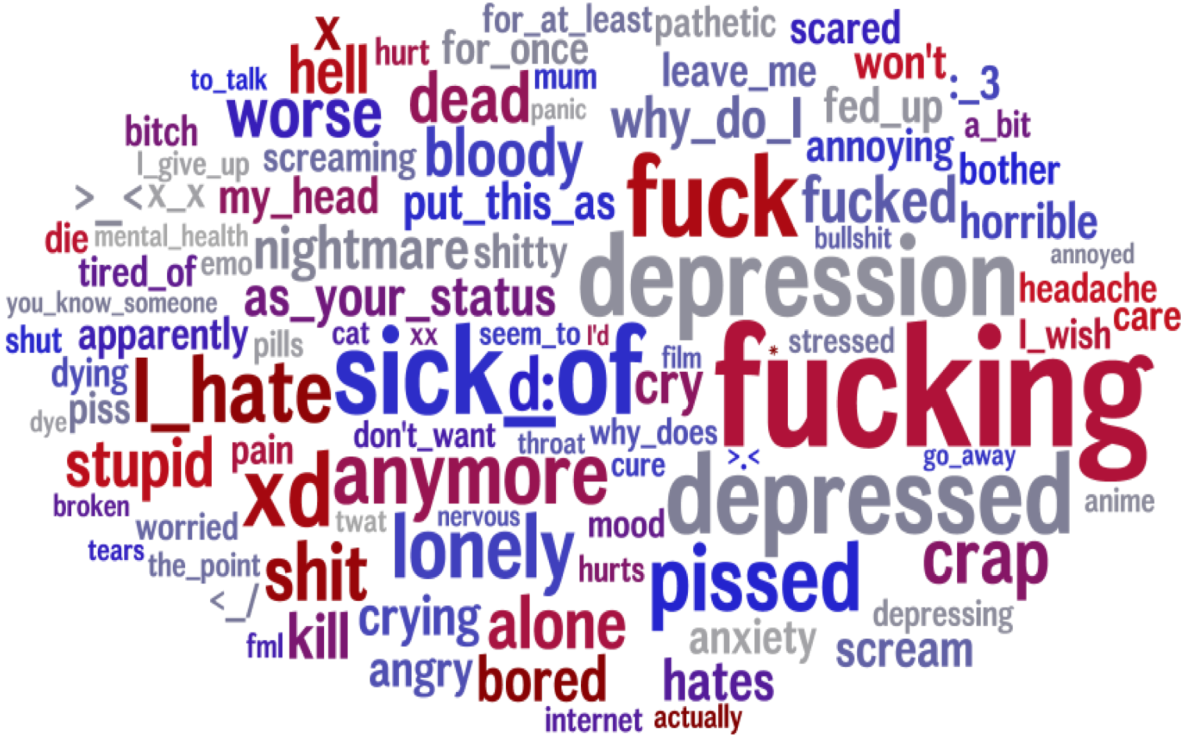
Creative Tools to Generate AI Art

Wondering how to make AI art? Scroll down for our list of tools to generate AI art.



What can we learn from your social media posts?

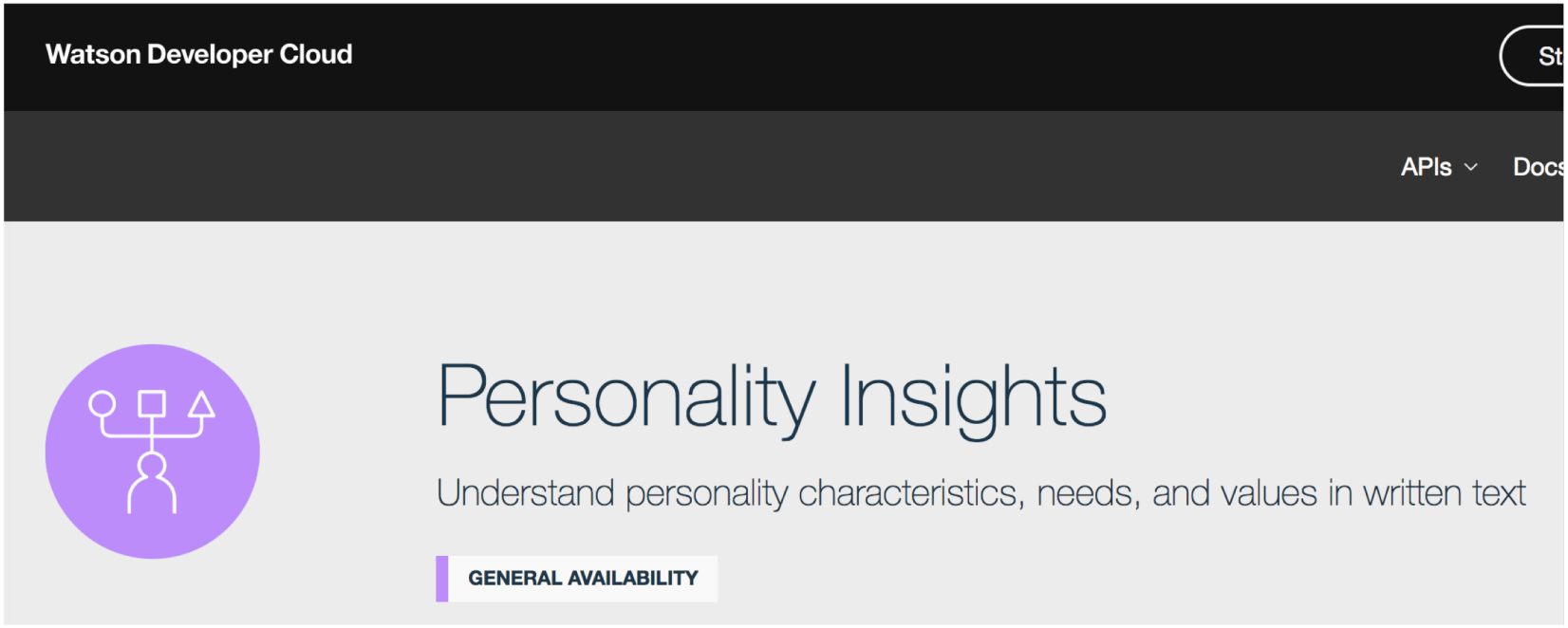
Neurotic



What can we learn from your social media posts?

Well adjusted

Targeted marketing from IBM



Admiral to price car insurance based on Facebook posts

Insurer's algorithm analyses social media usage to identify safe drivers in unprecedented use of customer data

theguardian

Targeted marketing: Cambridge Analytica

Privacy

OKCupid Data Publicly Released

WIRED, Michael Zimmer 5/14/16

- On May 8, a group of Danish researchers publicly released a dataset of nearly 70,000 users of the online dating site OkCupid, including usernames, age, gender, location, what kind of relationship (or sex) they're interested in, personality traits, and answers to thousands of profiling questions used by the site.
- When asked whether the researchers attempted to anonymize the dataset the response was: "… all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form."

Was the OKCupid data public?

- The methodology used to obtain the data was not fully explained, but involved a scraping bot.
  - Likely from an OkCupid profile researchers created.
- Since OkCupid users have the option to restrict the visibility of their profiles to logged-in users only, it is likely the researchers collected—and subsequently released—profiles that were intended to *not* be publicly viewable.

Privacy is not simple

- Many rules governing use of collected information
  - **HIPAA:** Health Insurance Portability and Accountability Act
  - **FERPA:** Family Educational Rights and Privacy Act
  - **GDPR** General Data Protection Regulation (Europe)
- However, “information leakage” can lead to unexpected disclosures
  - e.g. smart water meters
- “Privacy by trust” versus “privacy by design”

Anonymity?

"On the Internet, nobody knows you're a dog."

Correlating data

- **Netflix Prize Competition:** released a de-identified data set with user ID, date, movie name, and the rating given by the user for that movie.
  - Researchers were able to link users with IMDb's system where the users were identified, and talked about (some of) the movies they watched.
- **Problem: "Sparsity" of data**
  - In Netflix data, no two profiles are more than 50% similar.
  - If a Netflix profile is more than 50% similar to a profile in IMDB, then there is a high probability that the two profiles are of the same person

A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets …," Proc. 29th IEEE Symp. Security and Privacy, 2008.

Differential Privacy

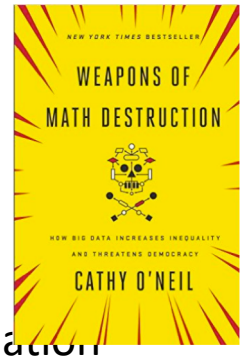
- When do you feel safe releasing personal information, e.g. completing a survey about your tastes in movies?
  - My answers have no impact on the privatized released result?
  - With high probability, an attacker looking at the privatized released result cannot learn any new information about me?
  - **These are not achievable.**
- **Differential privacy** aims to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records.
  - **The privatized released result is nearly the same whether or not I submit my information.**

Dwork and Roth, "Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Computer Science (2014).

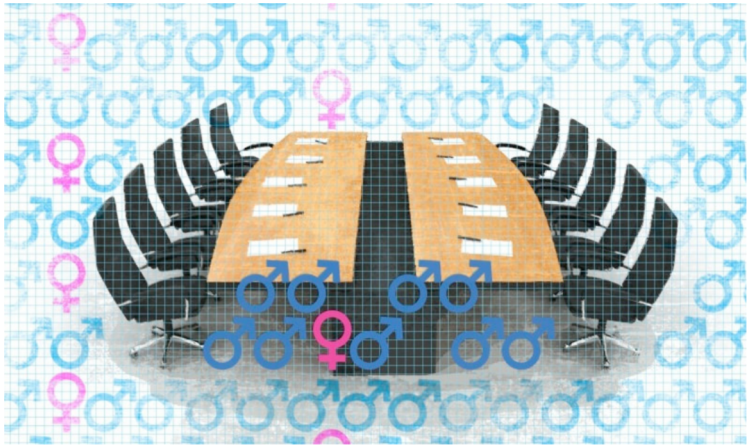
Ethics Surrounding Algorithms

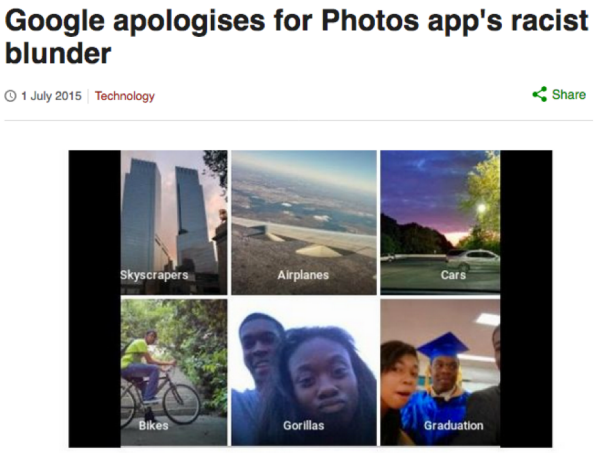
Fairness

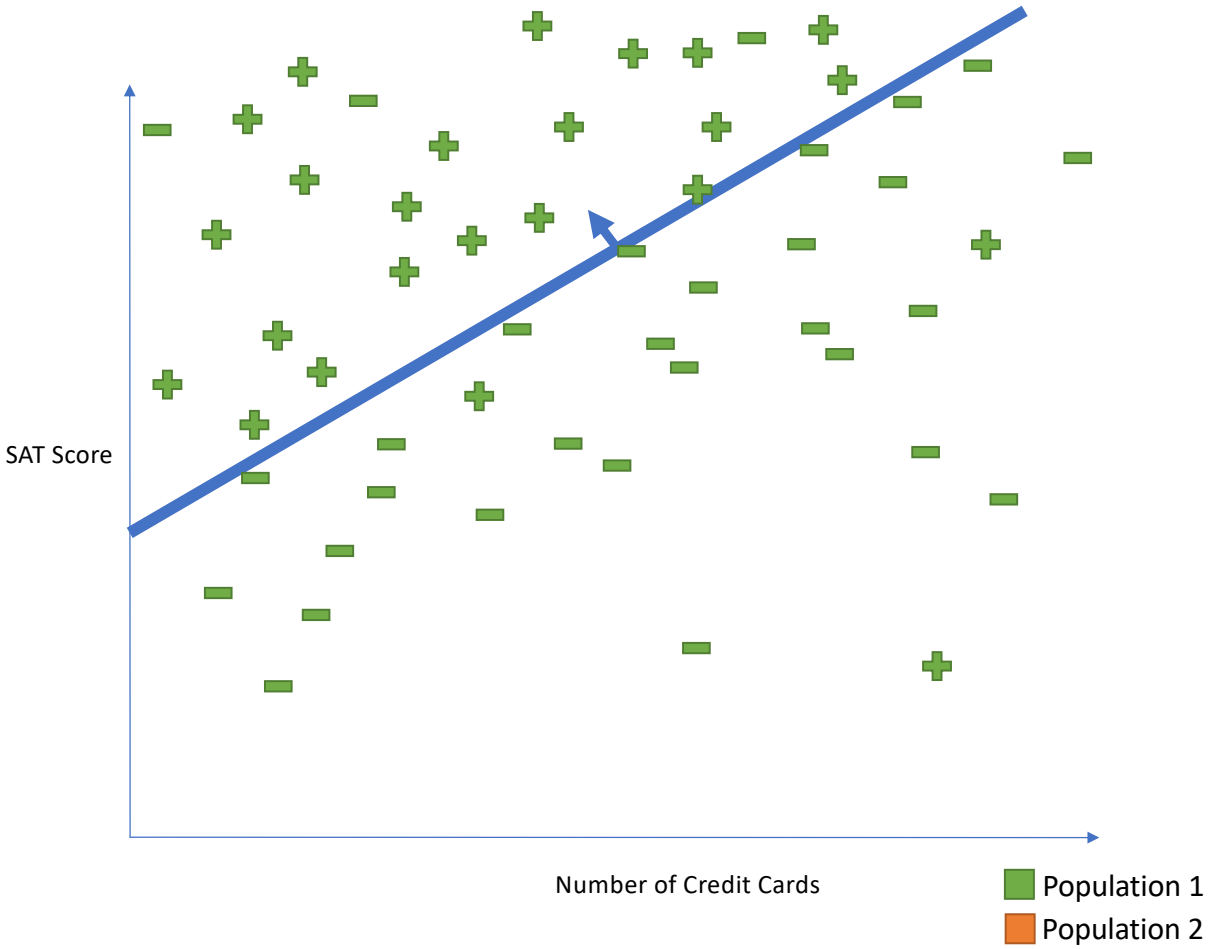
Algorithms are not neutral



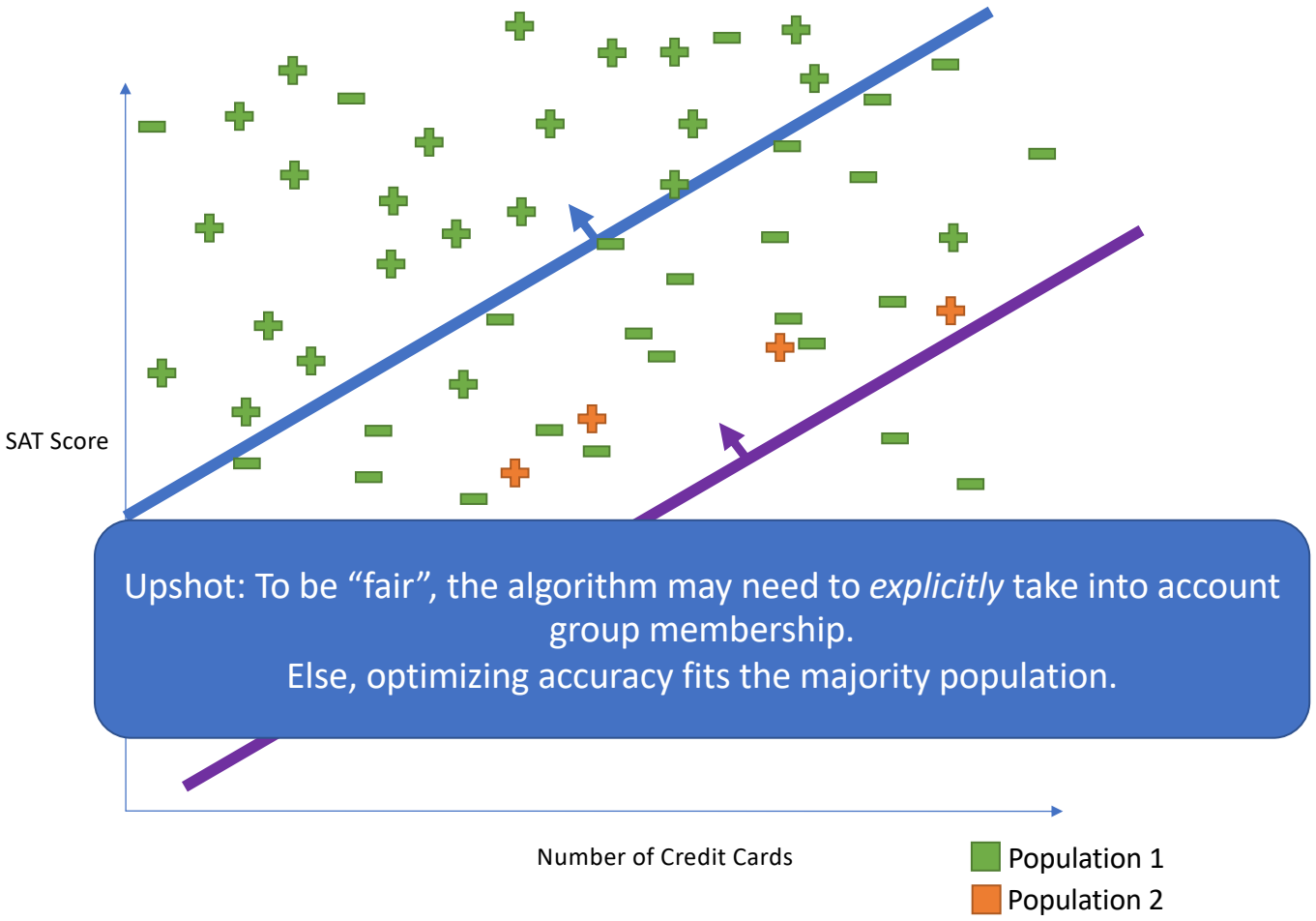
- Algorithms encode our biases.
  - Training data set isn't representative
  - Past population is not representative of the future population






SAT Score
Number of Credit Cards
Population 1
Population 2


SAT Score
Number of Credit Cards
Population 1
Population 2


SAT Score
Upshot: To be “fair”, the algorithm may need to *explicitly* take into account group membership.
Else, optimizing accuracy fits the majority population.
Number of Credit Cards
Population 1
Population 2

Online data is riddled with **SOCIAL STEREOTYPES**

Consequence: models are biased

Image Search

- June 2017: image search query “Doctor”

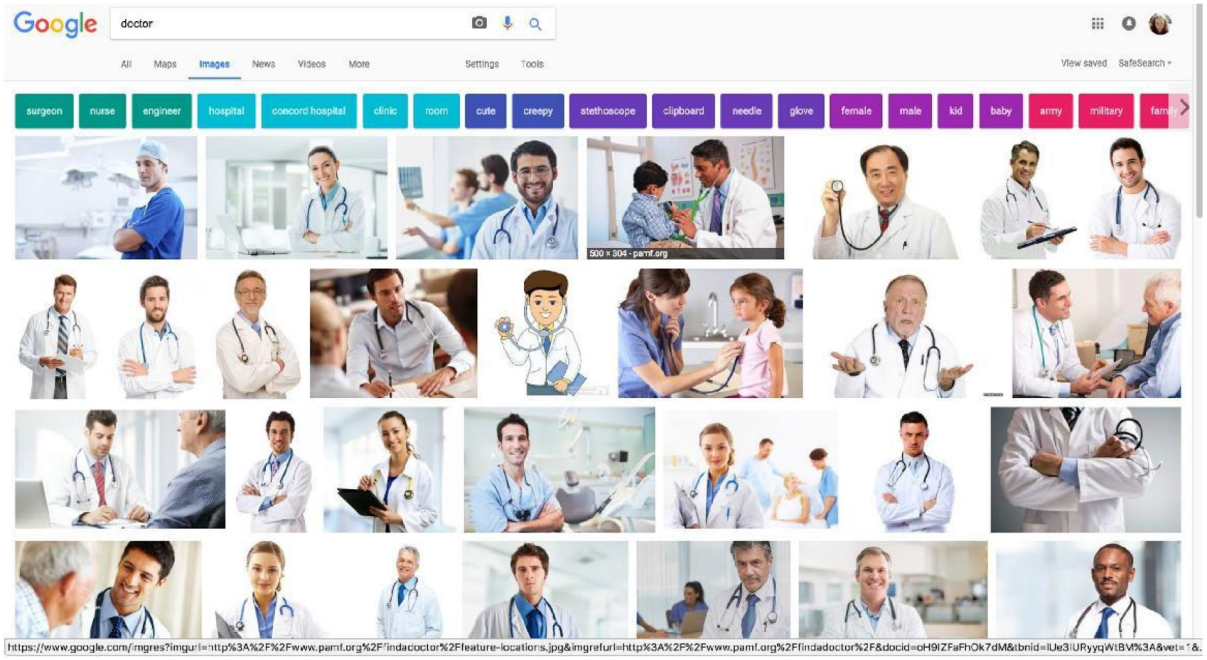


Image Search

- June 2017: image search query “Nurse”

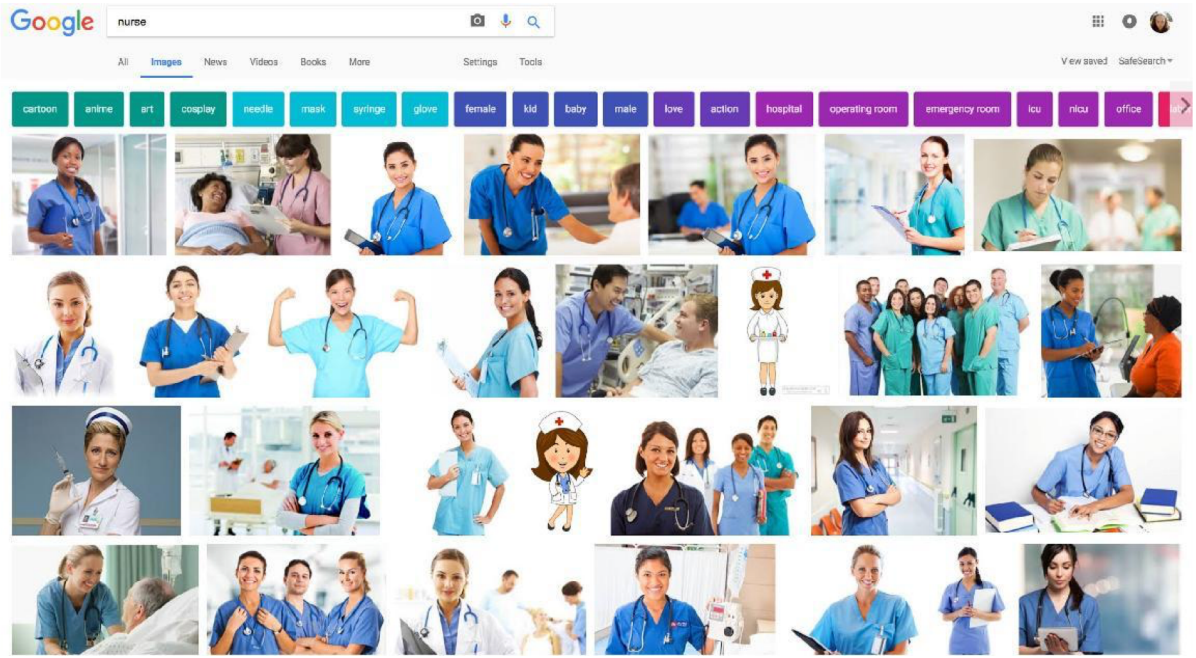


Image Search

- June 2017: image search query “Homemaker”

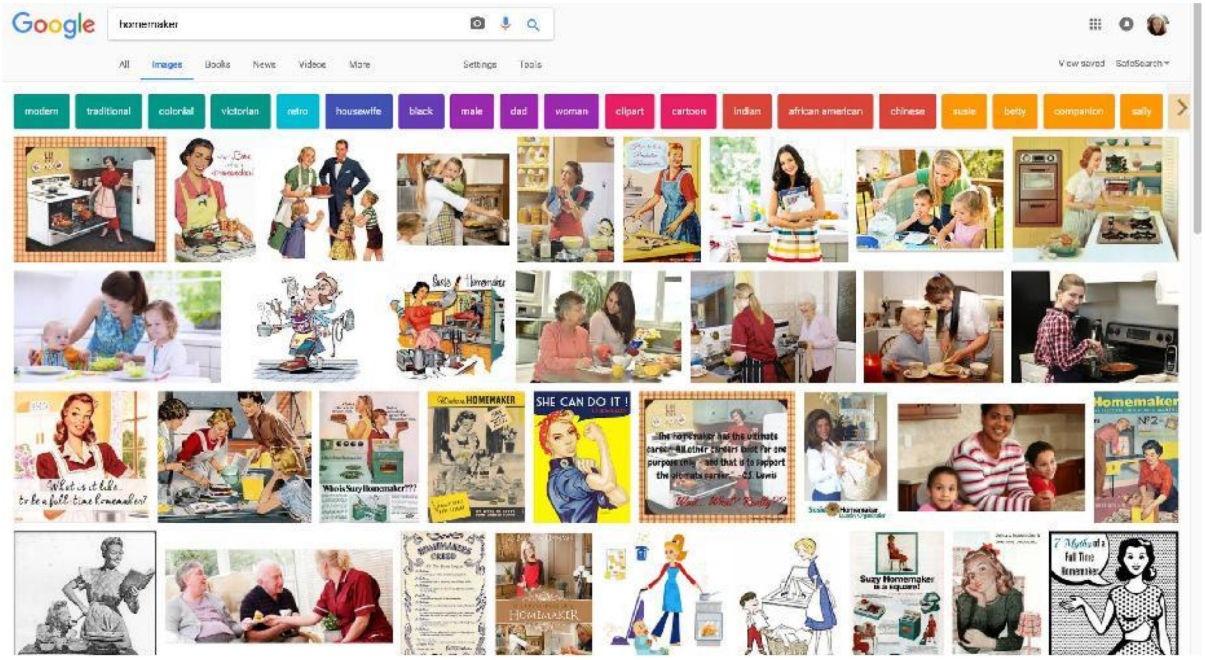


Image Search

- June 2017: image search query “CEO”

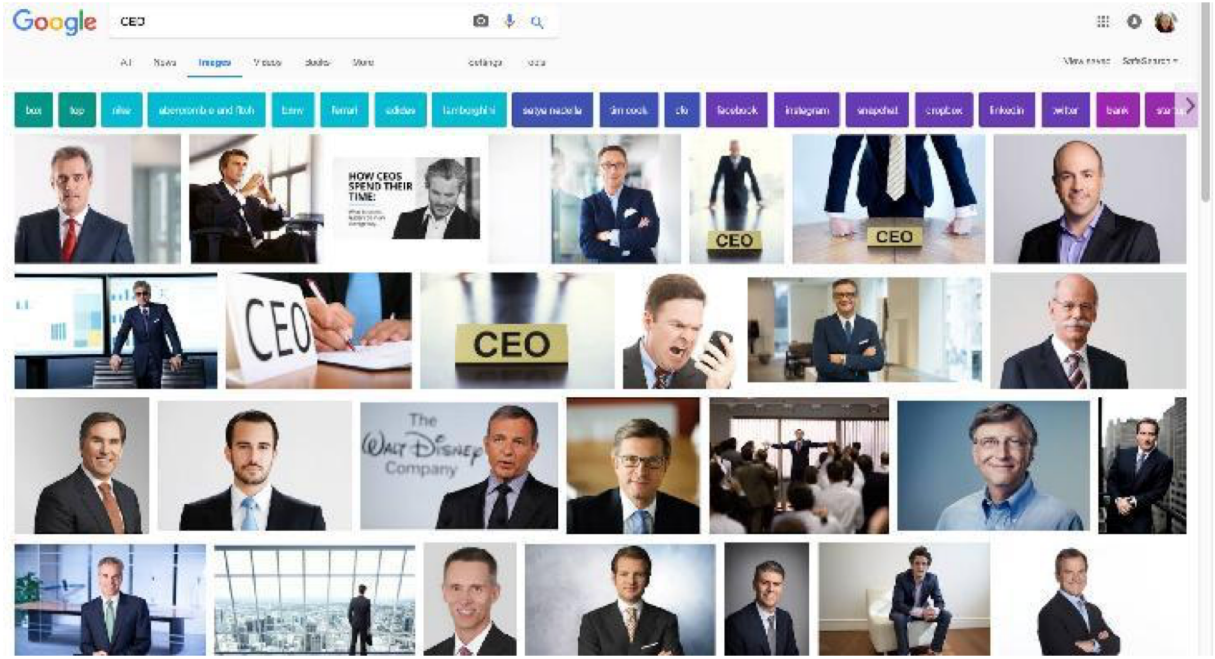
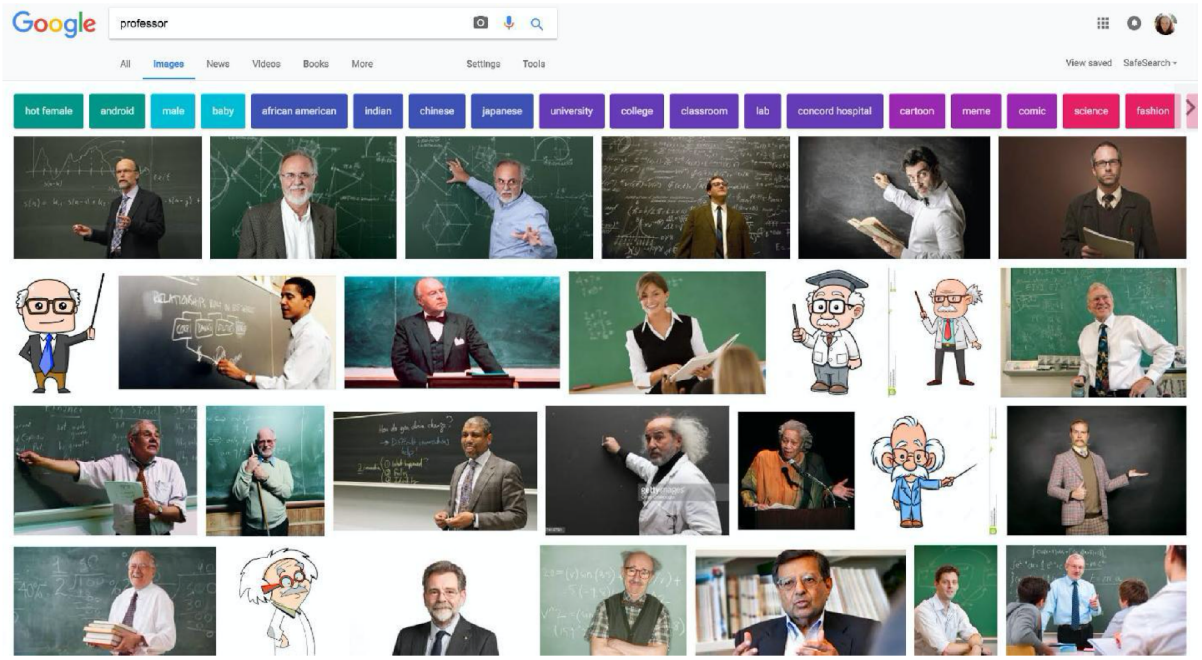


Image Search

- June 2017: image search query “Professor”



Fairness

- Fairness has been studied in social choice theory, game theory, economics and law.
- Currently trendy in theoretical computer science
  - **Discrimination of an individual:** An individual from the target group gets treated differently from an otherwise identical individual not from the target group.
  - **Discrimination in aggregate outcome**: the percentage success of the target group compared to that of the general population.

Dwork, Hardt, Pitassi, Reingold and Zemel,
"Fairness through Awareness"
Proc. 3rd Innovations in Theoretical Computer Science, 2012.

The Inquirer
DAILY NEWS philly.com

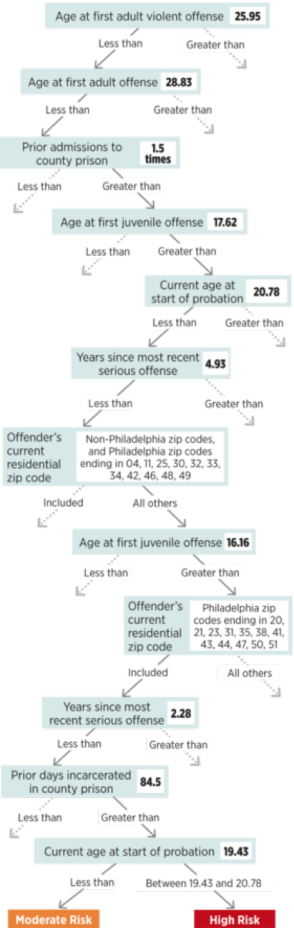
News — Crime

How computers are predicting crime - and potentially impacting your future

Updated: SEPTEMBER 21, 2017 — 12:53 PM EDT

How the Model Works

The risk-assessment tool used by the city's probation and parole department has 500 trees. This chart shows a potential path through one of them.

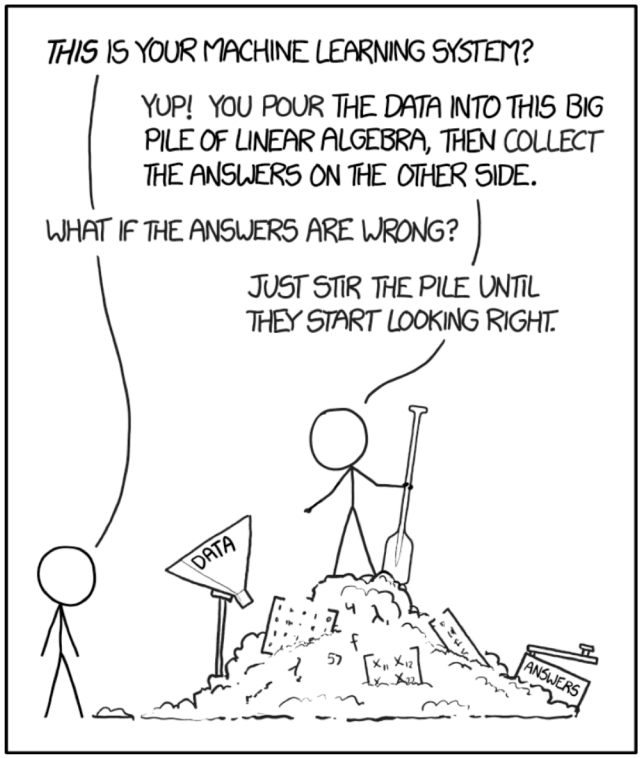


Why is computer advice on parole controversial?

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Machine Learning

|< < Prev Random Next > >|



|< < Prev Random Next > >|

Permanent link to this comic: https://xkcd.com/1838/

The pile gets soaked with data and starts to get mushy over time, so it's technically recurrent.

Reproducibility

- We need to be able to reproduce results but …
- The algorithms used in data science are complicated
  - When things “go wrong”, we need to understand why
  - Research code is often shared; commercial algorithms less so.
- The data often cannot be shared

Ethics summary

- Codes of conduct for research are fairly well understood
  - Get IRB approval
  - obtain informed consent
  - protect the privacy of subjects
  - maintain the confidentiality of data collected
  - minimize harm
- Fairness is more subtle
  - What is fair treatment of a group: equal accuracy? FP rate?