

# Trustworthy AI Autonomy

## Human Centricity

**Ding Zhao**

Assistant Professor  
Carnegie Mellon University

# Plan for today

- Privacy
  - Differential Privacy
  - Federated Learning
- Algorithmic Fairness
- Summary of this course
- Closing Thoughts and Next Steps

# Why should I be mindful of privacy as an engineer?

- AIML requires data to train the model
  - Many take it for granted that one can freely use a dataset to improve our products as long as we are not evil, e.g. illegally download the data and sell it.
  - However, people have concerns about the personal data being collected.

**If we just share use public datasets, then we are safe? — Not really**

1:23 PM Wed Apr 20

AA chriswhong.github.io

NYC Taxis: A Day in the Life About Asterisks Attribution Recommend 7.2K Share Tweet



# NYC Taxis: A Day in the Life

This visualization displays the data for one random NYC yellow taxi on a single day in 2013. See where it operated, how much money it made, and how busy it was over 24 hours.

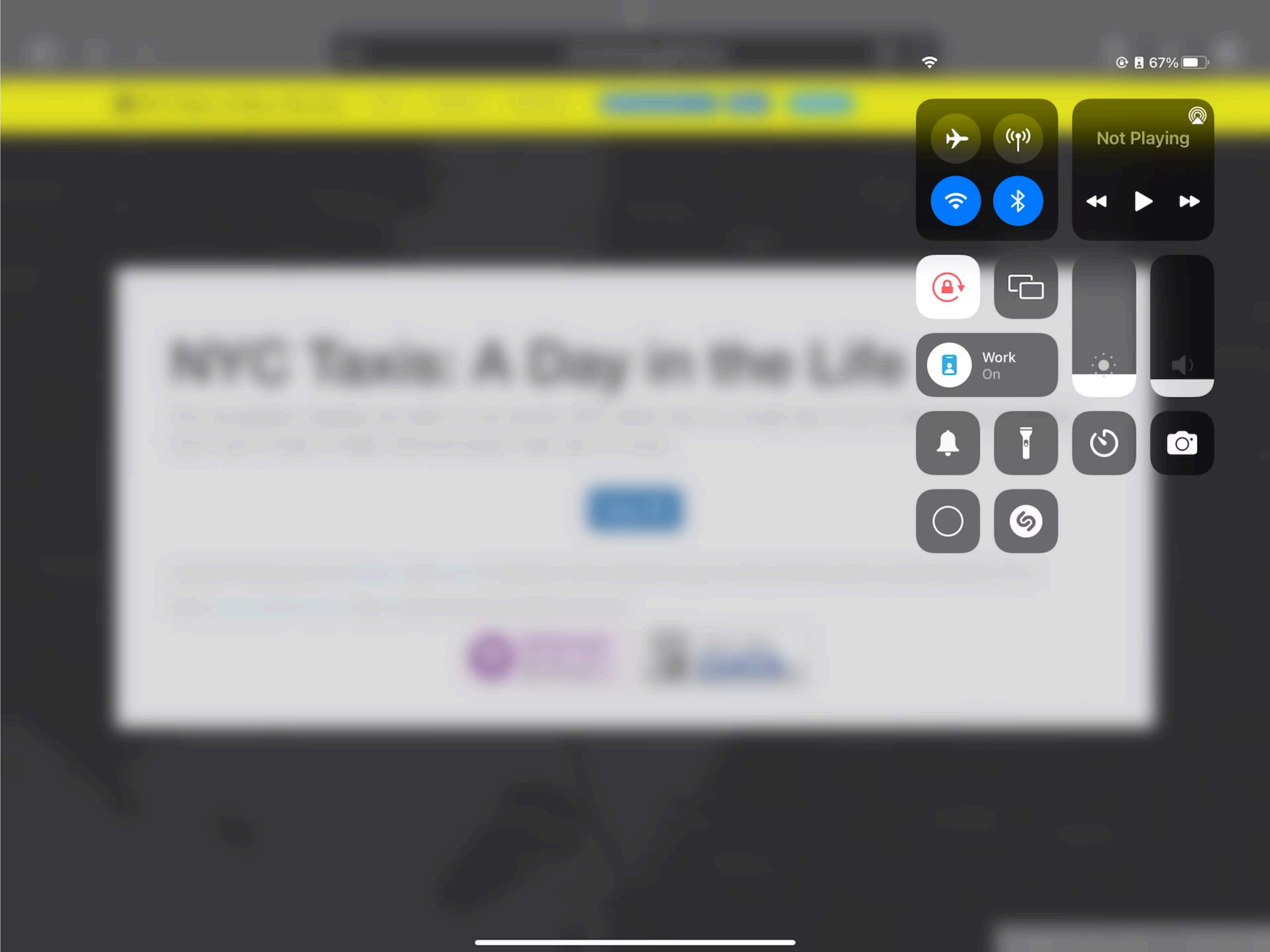
[Begin](#)

A Special Thanks goes out to [Mapbox](#) and [Heroku](#) for assistance with covering the surge of activity when this project was first released in 2014.

Here's [Technical Blog Post #1](#) and [#2](#) about how this visualization was built.

Leaflet | © OpenStreetMap contributors, © CartoDB



# Public NYC Taxicab Database Lets You See How Celebrities Tip



J.K. Trotter

10/23/14 12:00PM Filed to: DATA

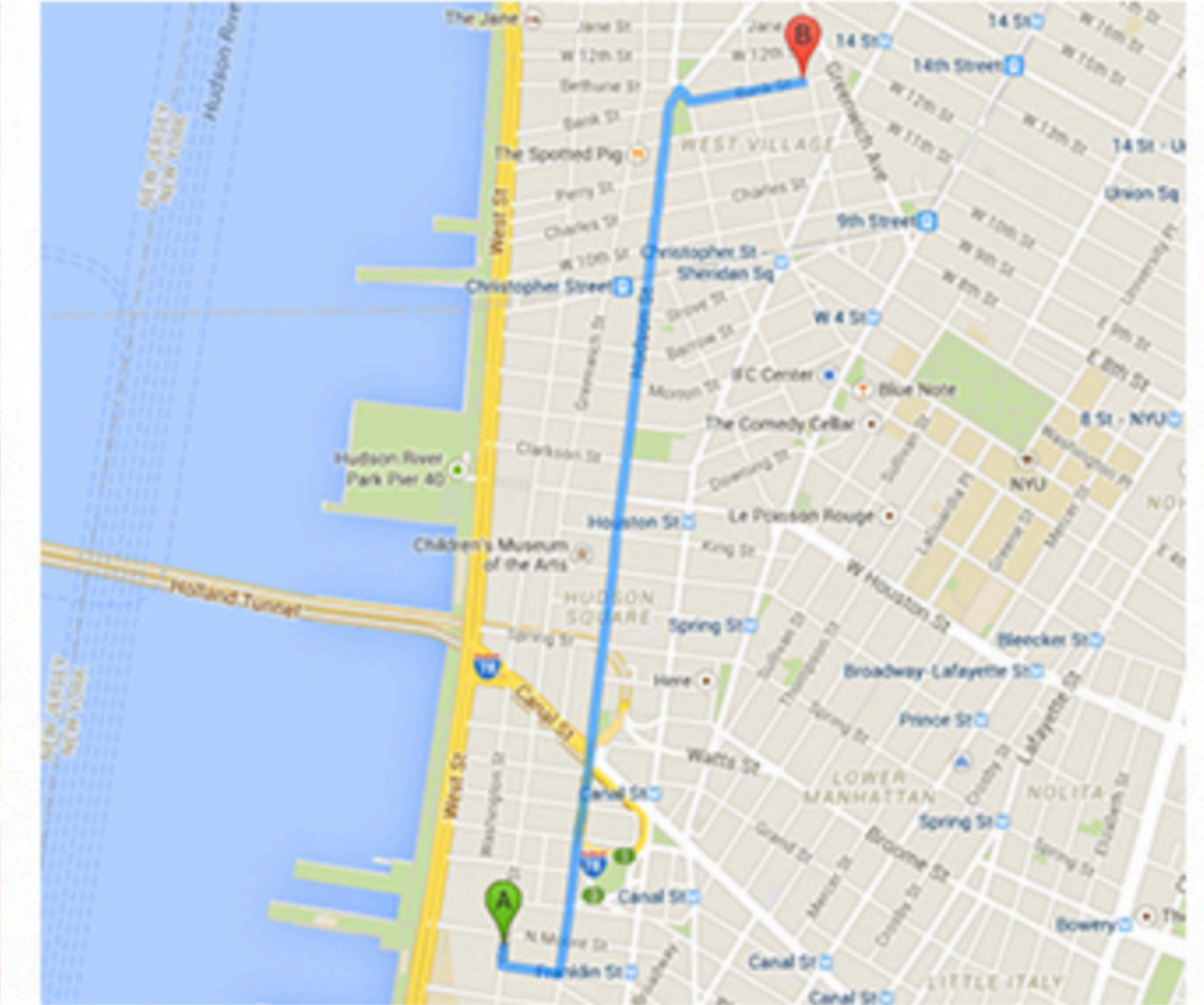


142.43K

Would it be safe to use data published voluntarily? - No



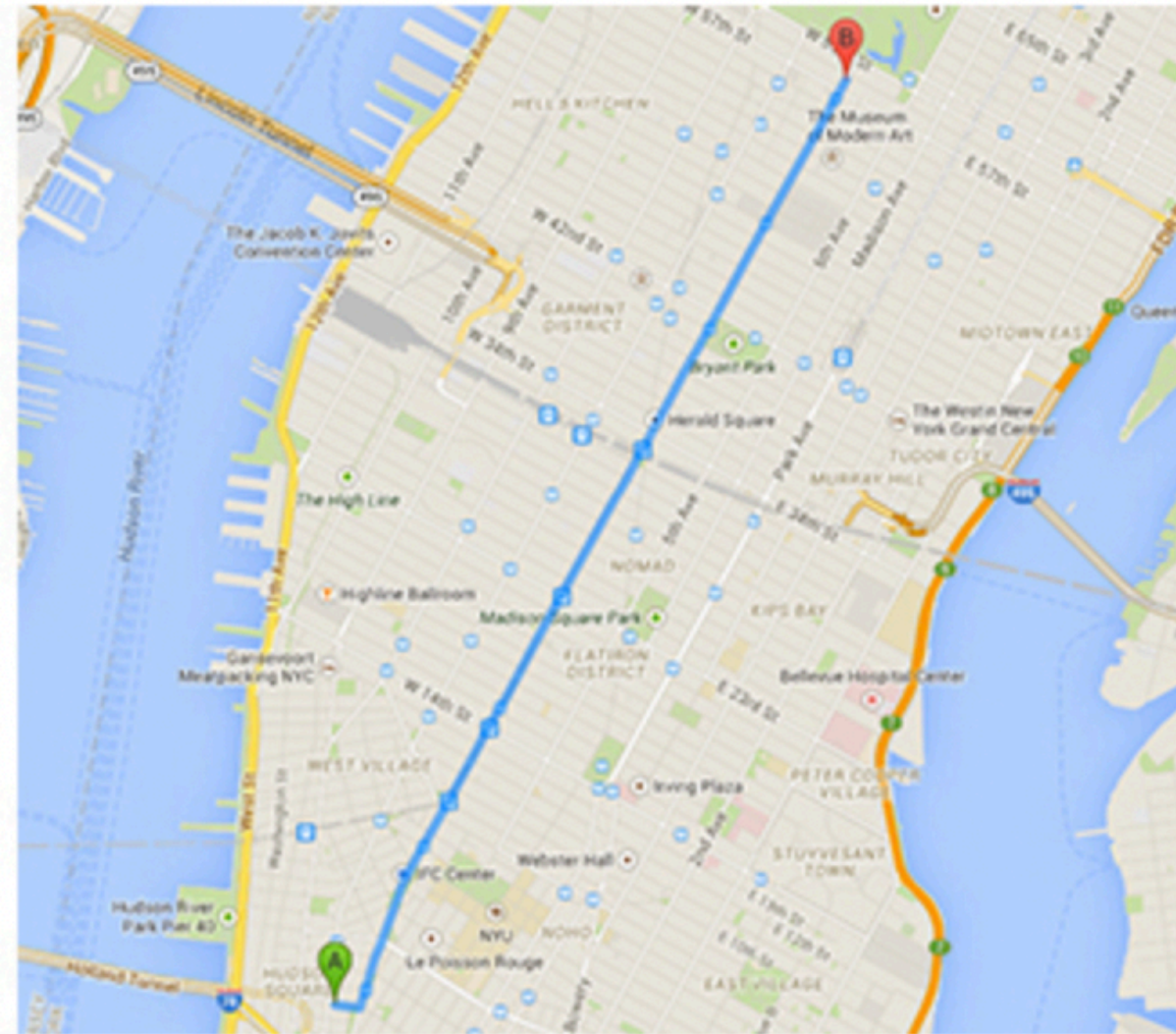
BRADLEY COOPER



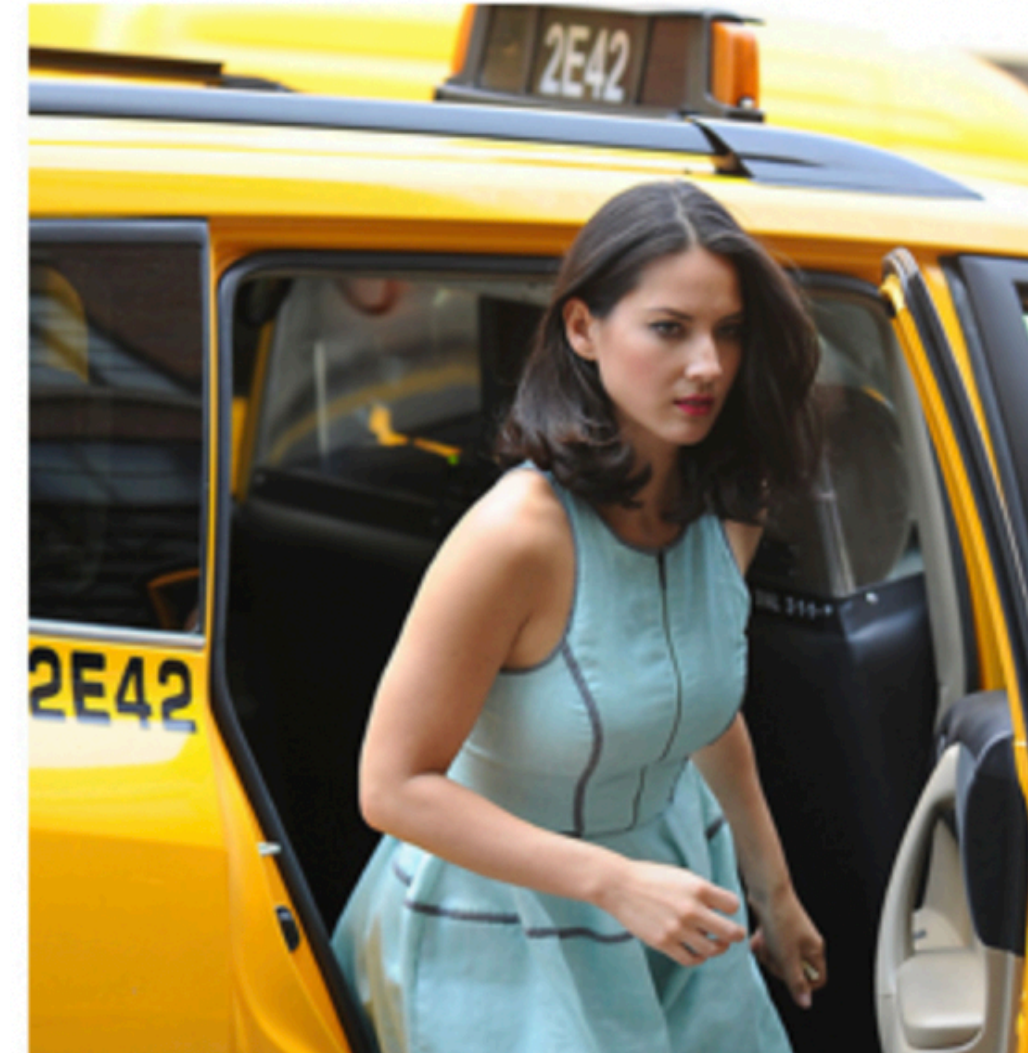
JULY 8, 2013 • 7:34 PM - 7:44 PM  
376 GREENWICH ST. TO 13 BANK ST.



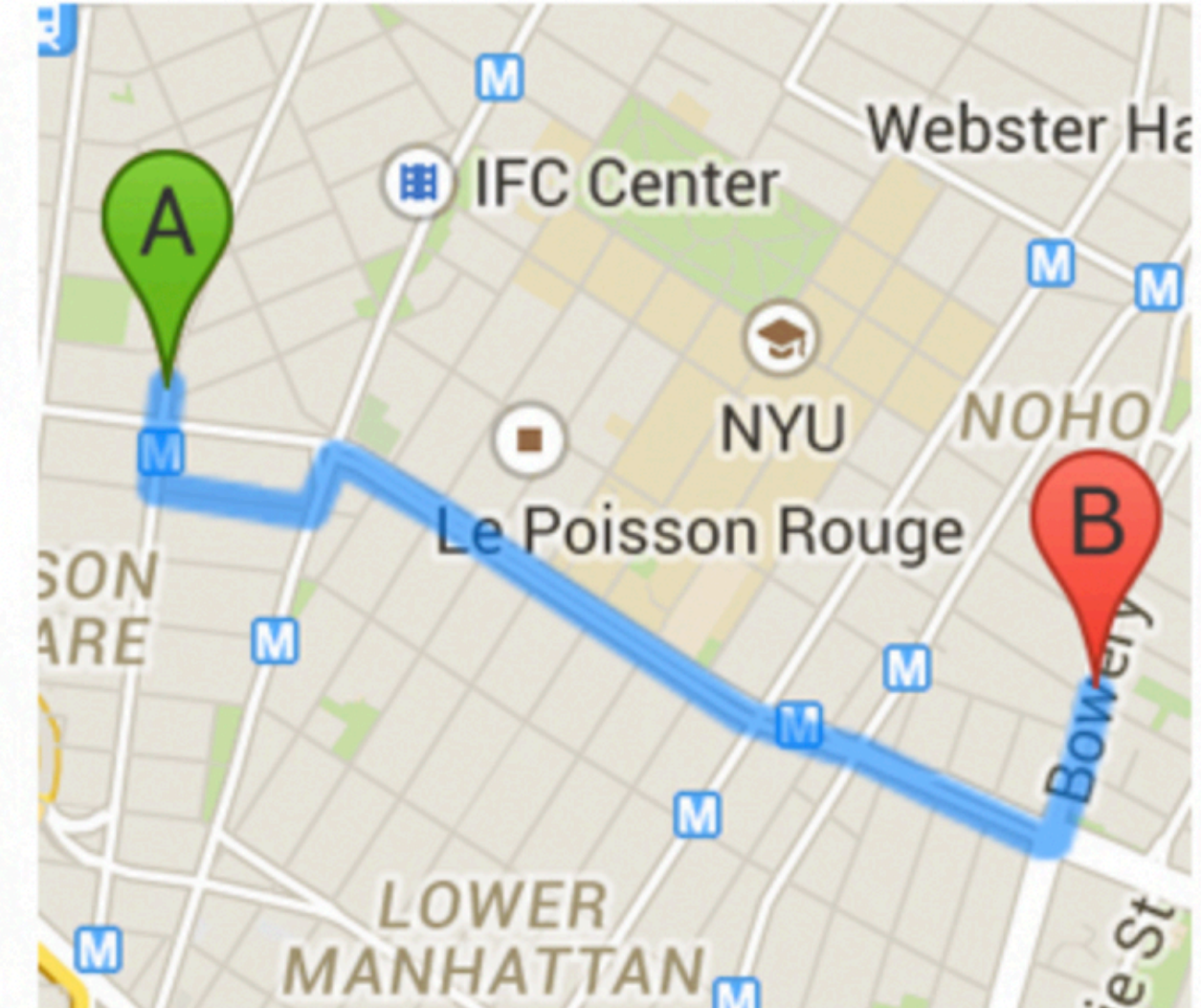
KOURTNEY KARDASHIAN  
SCOTT DISICK



NOVEMBER 4, 2013 • 12:11 PM - 12:36 PM  
246 SPRING ST. TO 1412 6TH AVE  
\$16.50 FARE • \$3.40 TIP • ©SPLASH



OLIVIA MUNN



JULY 8, 2013 • 11:20 AM - 11:26 AM  
225 VARICK ST. TO 325 BOWERY  
\$6.00 FARE • CASH; UNKNOWN TIP • ©SPLASH

# Recover location from data volunteered published data

**Support the Guardian**  
Available for everyone, funded by readers  
[Contribute](#) → [Subscribe](#) →

Search jobs | Sign in | Search | US edition ▾

**The Guardian**  
For 200 years

[News](#) | [Opinion](#) | [Sport](#) | [Culture](#) | [Lifestyle](#) | [More](#) ▾

**GPS**

This article is more than 4 years old

Advertisement


## Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

- [Latest: Strava suggests military users 'opt out' of heatmap as row deepens](#)

**Alex Hern**  
@alexhern  
Sun 28 Jan 2018 16.51 EST

[f](#) [t](#) [e](#)



Would it be safe to use anonymous data? – No

# Antonymy is not enough



Bits

Business, Innovation, Technology, Society

The New York Times

## PRIVACY

### With a Few Bits of Data, Researchers Identify 'Anonymous' People

BY NATASHA SINGER JANUARY 29, 2015 2:01 PM 12

Email

Share

Tweet

Save

More



Yves-Alexandre de Montjoye, a graduate student at the

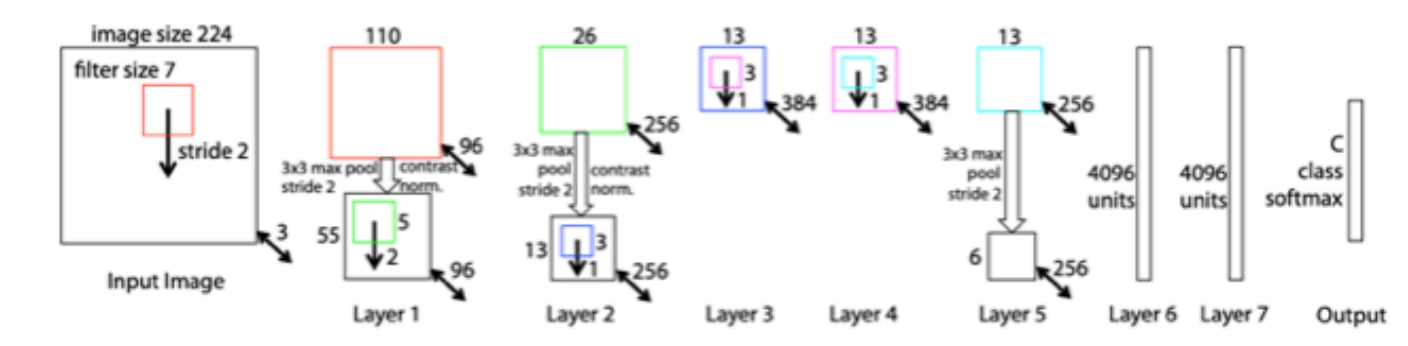
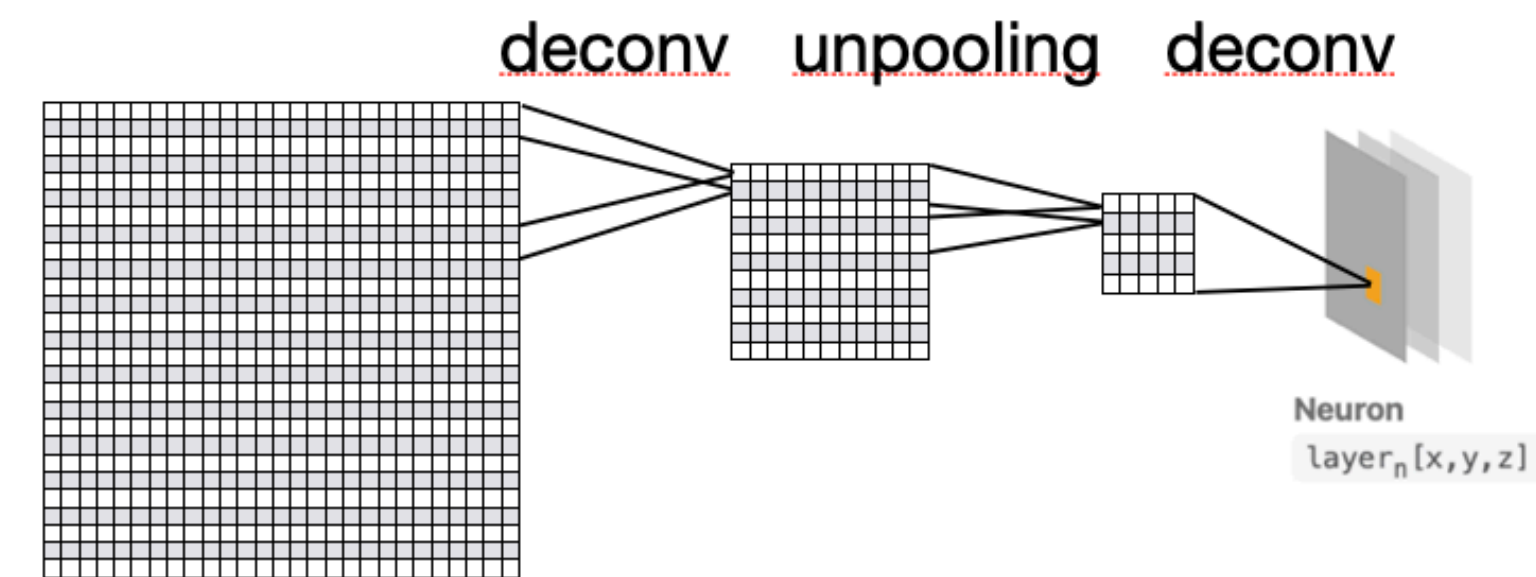
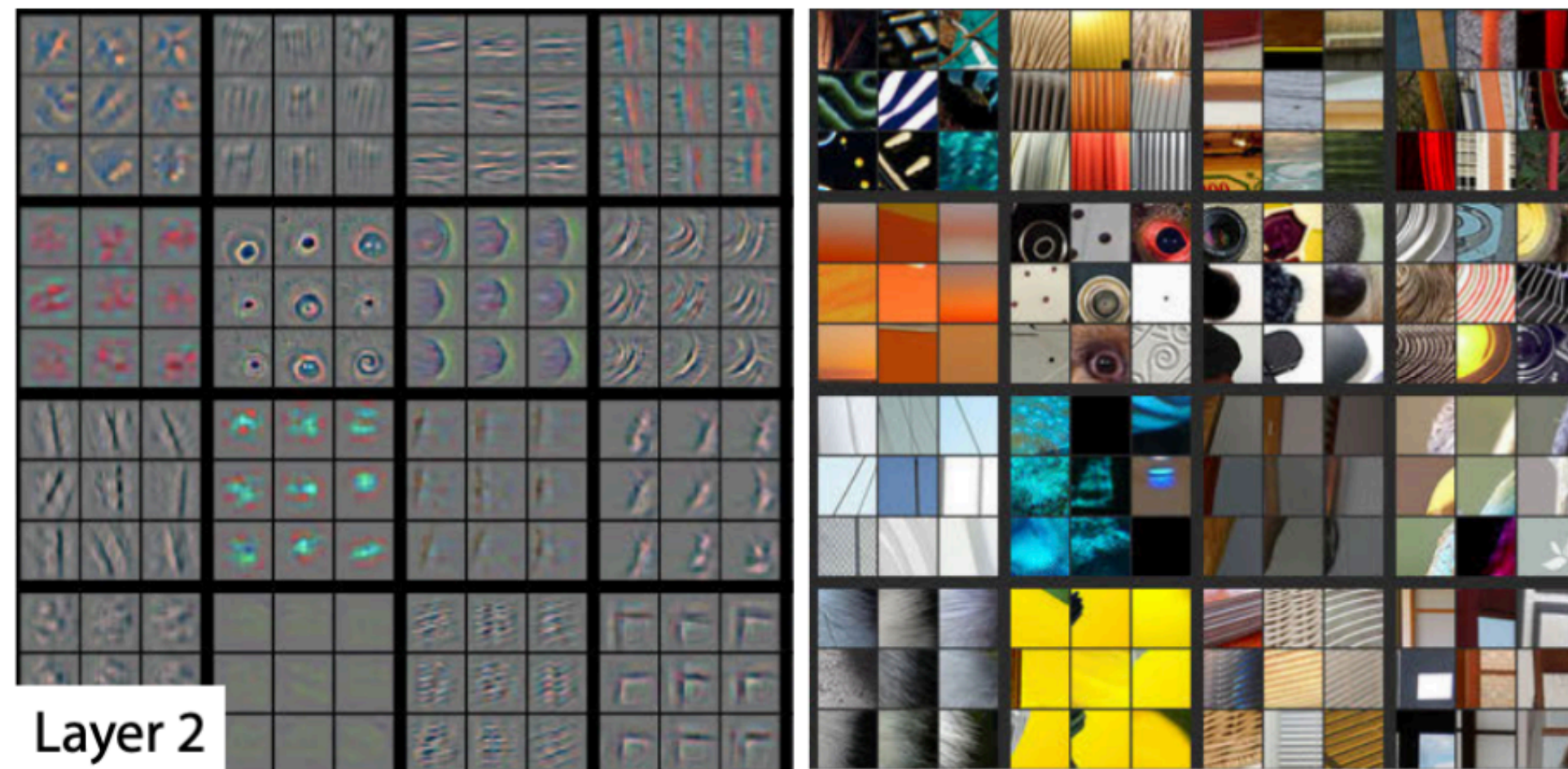
So, it seems we should not share any data, then we are safe? — still not true.



# Recall this slide in M1-2 Explanation

## Deconvnet of a single neuron: Layer 2

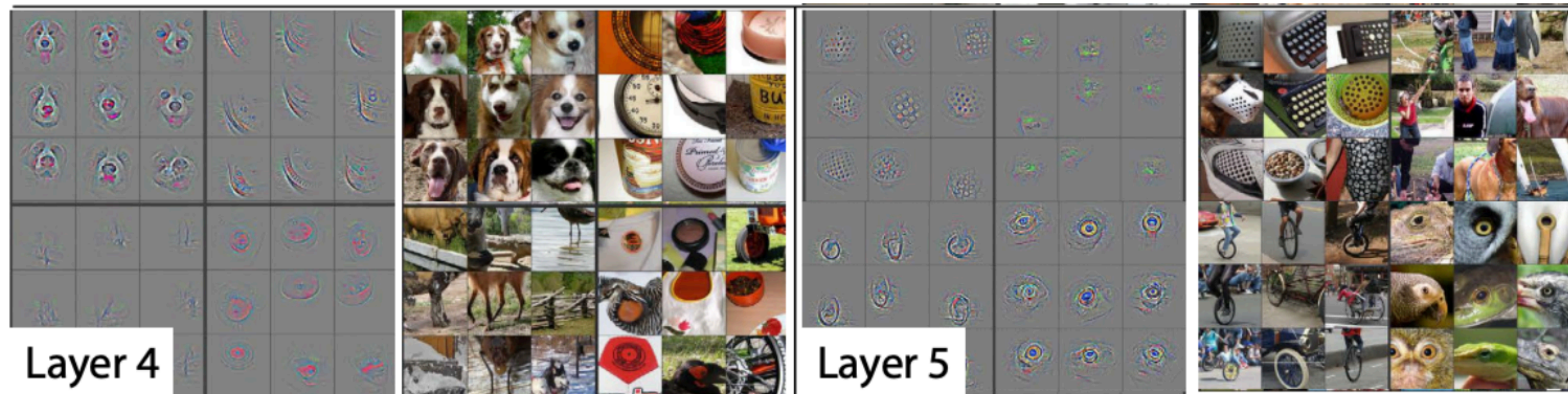
Projection to the pixel space    Corresponding image patches



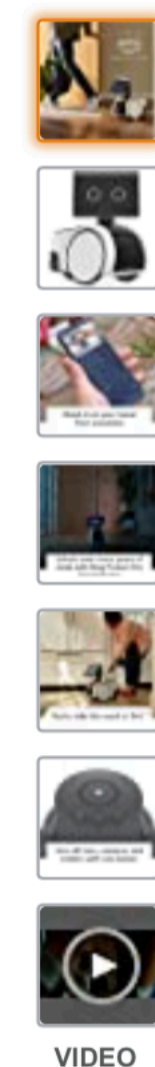
# Recover images from algorithms

## Deconvnet

- Final layers identify informative complex features for final prediction



Day1 Editions Be among the first to access new innovations from Amazon



Introducing Amazon Astro, Household Robot for Home Monitoring, with Alexa, Includes 6-month Free Trial of Ring Protect Pro

Brand: Amazon

Day1 Editions

Price: \$999.99 & FREE Returns

**Introductory price.** After introductory period, price will be \$1,449.99. Terms and conditions apply.

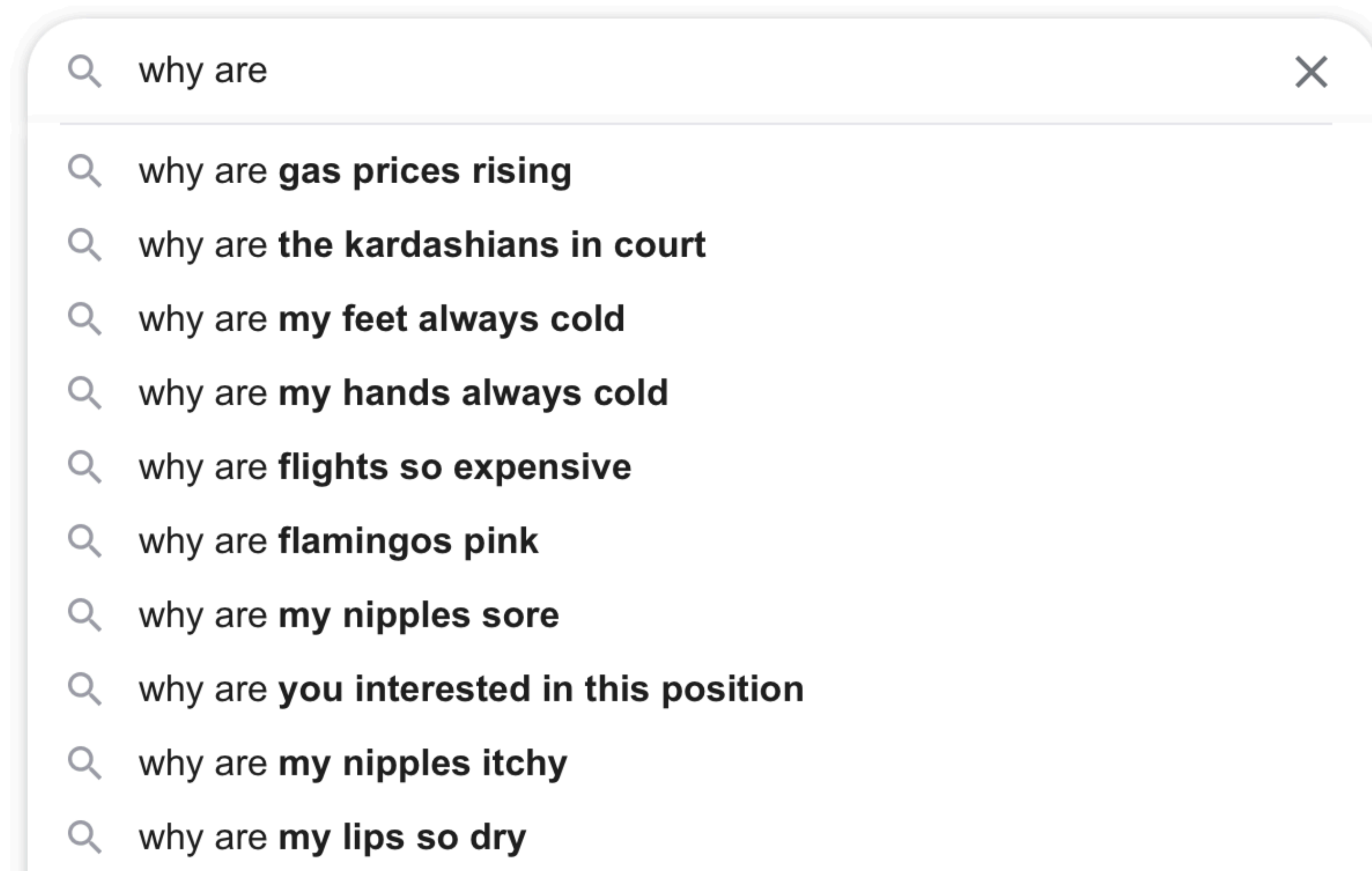
Available exclusively by invitation

Ships from and sold by Amazon.com.

Please note this product can only ship to addresses in the 50 US states.



# Infer historical data from the output of algorithms



# What is a privacy-preserving algorithm?

- What is Prof Zhao's salary?
- What is the average salary of CMU's professors?

The screenshot shows the Glassdoor website interface. At the top, the Glassdoor logo is on the left, and search filters for 'Carnegie Mellon University' and 'Pittsburgh, PA' are in the center. A 'Sign In' button is on the right. Below the search bar, there are navigation tabs for 'Jobs', 'Companies', 'Salaries', and 'Careers'. The 'Salaries' tab is selected. The main content area displays the 'Carnegie Mellon University' profile, including a logo, the name 'Carnegie Mellon University', and a status of 'Engaged Employer'. Below this, there are statistics for 'Overview', 'Reviews', 'Jobs', 'Salaries', 'Interviews', 'Benefits', and 'Photos'. The 'Salaries' section is highlighted, showing 'Carnegie Mellon University Professor Salaries' with a 'Low Confidence' warning. A horizontal bar chart shows the salary distribution with an average of \$234,379/yr. Below the chart, a text box explains that the typical salary is \$234,379 per year, ranging from \$218,071 to \$317,455. To the right of the main content, there are two sidebars: one for 'Work in HR or Marketing?' and another for 'Top Companies for Compensation and Benefits Near You' listing the University of Pittsburgh, Penn State, and the University of Pennsylvania.

**Carnegie Mellon University Professor Salaries**  
Updated Mar 11, 2022

United States | Any Experience | Search

**\$234,379/yr**  
Base Pay Average  
5 Salaries

Average

\$218K | \$223K (25%) | \$288K (75%) | \$317K

Cash Bonus, Stock Bonus, Profit Sharing, Commission Sharing, Tips have not been reported for this role

**How much does a Professor at Carnegie Mellon University make?**

The typical Carnegie Mellon University Professor salary is \$234,379 per year. Professor salaries at Carnegie Mellon University can range from \$218,071 - \$317,455 per year. This estimate is based upon 5 Carnegie Mellon University Professor salary report(s) provided by employees or estimated based upon statistical methods. When factoring in bonuses and additional compensation, a Professor at Carnegie Mellon University can expect to make an average total pay of \$234,379 per year.

**\$234,379**  
Total Pay Average ⓘ

Work in HR or Marketing?  
Get a free employer account

Top Companies for "Compensation and Benefits" Near You

- University of Pittsburgh  
Compensation & Benefits  
3.7 ★
- Penn State  
Compensation & Benefits  
3.7 ★
- University of Pennsylvania  
Compensation & Benefits  
3.8 ★

# Definition of a privacy-preserving algorithm

- Version-1
  - Analysis of dataset D is private if:
    - Analyst knows no more about me after analysis than before.
    - - It is strict but not very realistic.
  - Was my salary privacy violated if someone gets the average salary information of CMU or even UW?
    - Yes, under such a definition.

Overview 2.0k 99 3.5k 171 368 52  
Reviews Jobs Salaries Interviews Benefits Photos

Follow + Add a Salary

Cash Bonus, Stock Bonus, Profit Sharing, Commission Sharing, Tips have not been reported for this role

How much does a Professor at Carnegie Mellon University make?

The typical Carnegie Mellon University Professor salary is \$234,379 per year. Professor salaries at Carnegie Mellon University can range from \$218,071 - \$317,455 per year. This estimate is based upon 5 Carnegie Mellon University Professor salary report(s) provided by employees or estimated based upon statistical methods. When factoring in bonuses and additional compensation, a Professor at Carnegie Mellon University can expect to make an average total pay of \$234,379 per year.

**\$234,379**  
Total Pay Average ⓘ

Related Searches: All Professor Salaries | All Carnegie Mellon University Salaries

How does this pay data of \$234,379 look to you? Your input helps Glassdoor refine our pay estimates over time.

Right High Low

Salaries > Professor > Carnegie Mellon University

**Related Salaries**

Company	Average Base Salaries in (USD)	Low	High
SCAD Savannah College of Art and Design 79 salaries	\$79,991/yr	\$50K	\$118K
University of Washington 50 salaries	\$165,000/yr	\$77K	\$287K
Penn State 49 salaries	\$174,423/yr	\$100K	\$312K

Job Seekers Also Viewed < >

University of Pittsburgh  
4.2 ★  
2,310 Reviews

How satisfied or dissatisfied are you with the information you found on this page?

Extremely dissatisfied

Somewhat dissatisfied

Neutral

Somewhat satisfied

Extremely satisfied

SUBMIT

**It seems privacy intrusion is almost unavoidable if we ever collect any data. Well, yes, but we could constrain the privacy budget to a certain degree by defining such a soft privacy constraint**

# A more useful definition

- Version 2: Analysis of dataset  $D$  is private if:
  - analyst knows **almost** no more about me after analysis than he would have,
  - had he conducted the same analysis on
  - an identical dataset **with my data removed**
- **Mathematically, this leads to a famous privacy definition**
  - **Differential Privacy**

# Neighboring

- Two data sets  $D_1$  and  $D_2$  if differ on  $\leq 1$  entry

Name	Salary
Farnam Jahanian	\$xxxxxxx
• • •	• • •
Ding Zhao	\$xxxxxxx
• • •	• • •
Jon Cagan	\$xxxxxxx

$D_1$

Name	Salary
Farnam Jahanian	\$xxxxxxx
• • •	• • •
Joe Biden	\$xxxxxxx
• • •	• • •
Jon Cagan	\$xxxxxxx

$D_2$



# $\epsilon$ -differential privacy

- Algorithm  $\mathcal{A}$  is  $\epsilon$ -differentially private if:
- For all pairs of neighboring sets  $D_1, D_2$  and any set  $R$  of possible output (response)

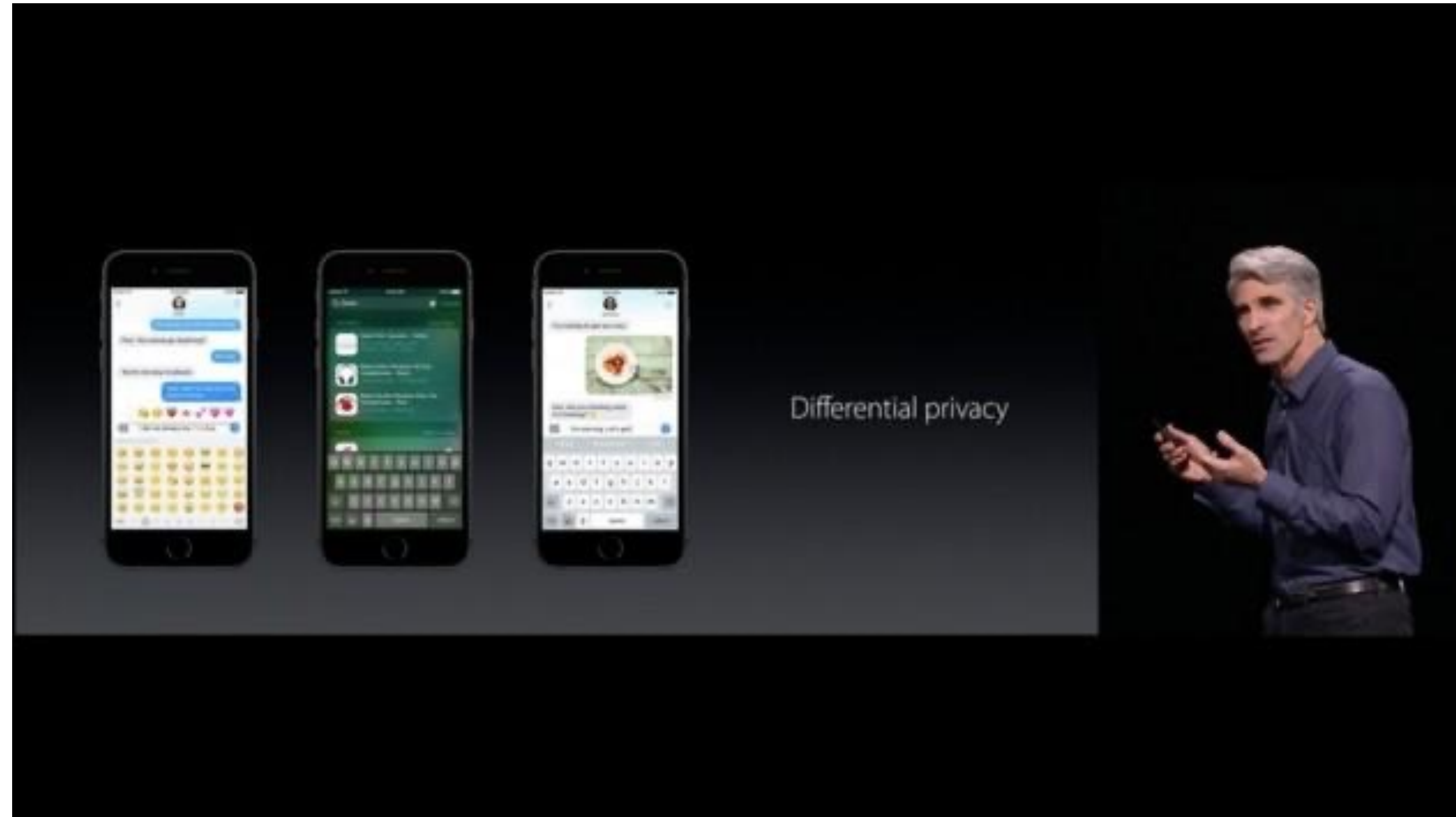
$$\Pr[\mathcal{A}(D_1) \in R] \leq e^\epsilon \Pr[\mathcal{A}(D_2) \in R]$$

- Note: for small  $\epsilon$ ,  $e^\epsilon \approx 1 + \epsilon$
- A consequence: for any possible response  $y$

$$\exp(-\epsilon) \leq \frac{\Pr(\mathcal{A}(\mathcal{D}_1) = y)}{\Pr(\mathcal{A}(\mathcal{D}_2) = y)} \leq \exp(\epsilon)$$

# DP has been used in the industry

- Apple has adopted and further developed a technique known in the academic world as local differential privacy to do something really exciting: gain insight into what many Apple users are doing, while helping to preserve the privacy of individual users. It is a technique that enables Apple to learn about the user community without learning about individuals in the community.



# Tools for designing privacy-preserving algorithms

- Key idea: add noise to the output of the analysis  $\mathcal{A}$ , such that the output of the analysis  $\mathcal{A}(D)$  is insensitive to the addition of my salary to  $D$ .
- For example, if  $\mathcal{A}$  is to take the average. For different  $D$ s we may need to add different level of noise to be  $\epsilon$ -differential private given a fixed  $\epsilon$ .
- The more sensitive, the bigger noise we need to add to the output.

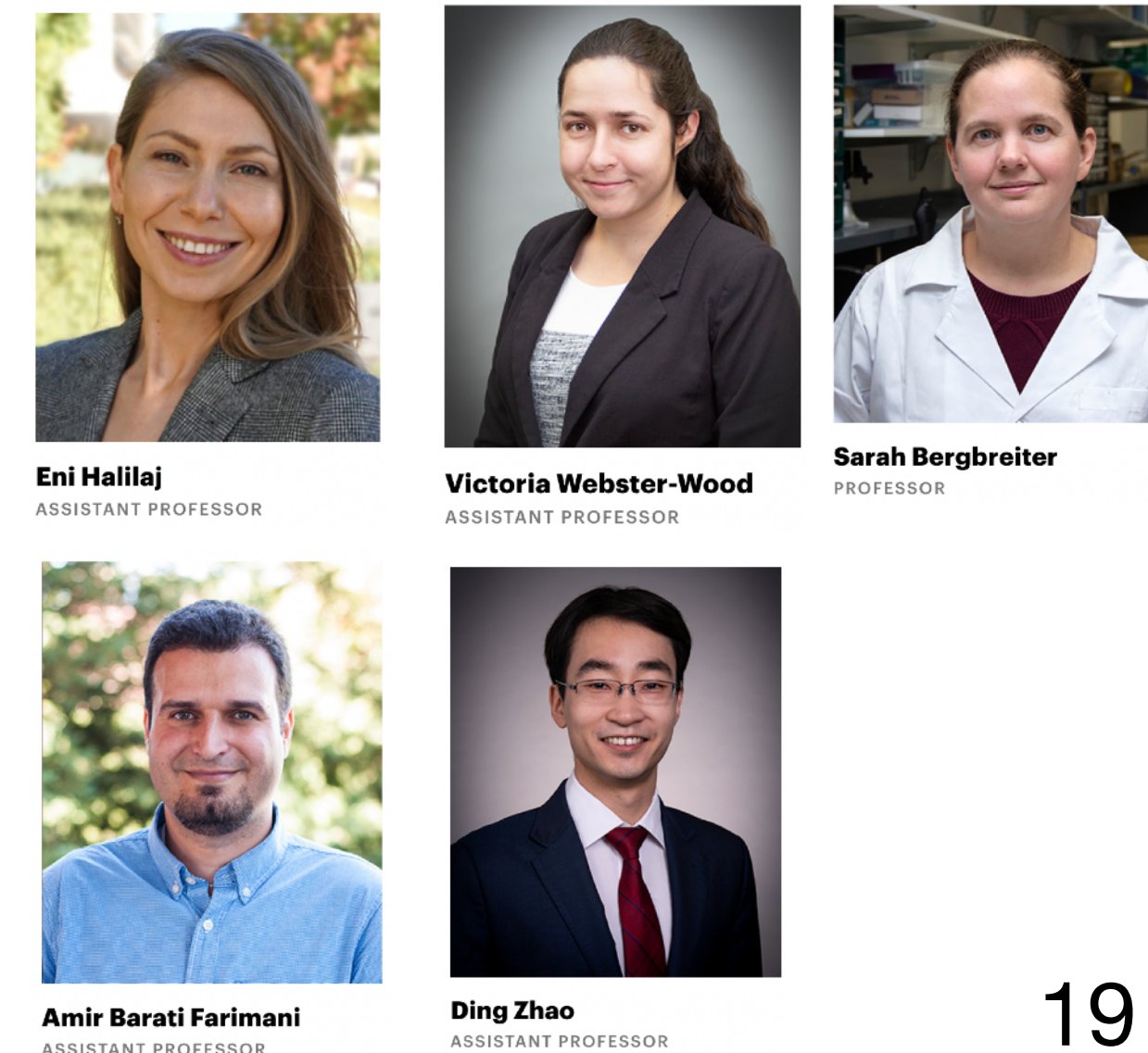
CMU faculty



ME faculty

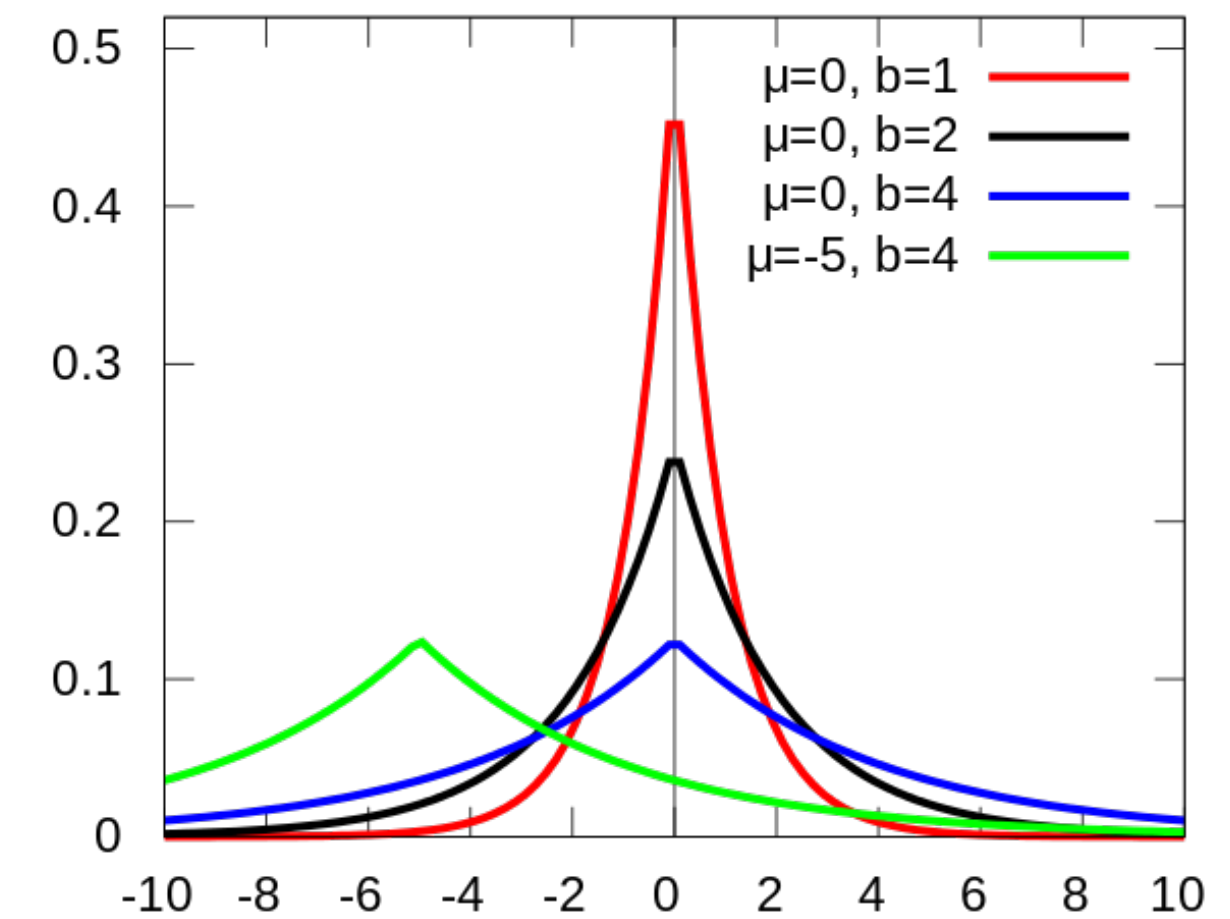


ME faculty joined in 2018



# Laplace mechanism

- Goal: Evaluate  $f:D \rightarrow \mathbb{R}$  mapping datasets to  $\mathbb{R}$ ; preserve  $\epsilon$ -DP
  - For example,  $f$  is the mean salary of people in  $D$
- Idea: add noise to  $f$  to hide any individual info
- Sensitivity of  $f$  over  $D$ :  $\Delta_f = \max_{D_1, D_2 \text{ neighboring}} |f(D_1) - f(D_2)|$
- Laplace Mechanism outputs:  $Z_D \sim \text{Lap}(f(D), \frac{\Delta_f}{\epsilon})$
- Note: adding Gaussian will violate the DP requirement.



$$\text{Lap}(x | \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

# Federated Learning

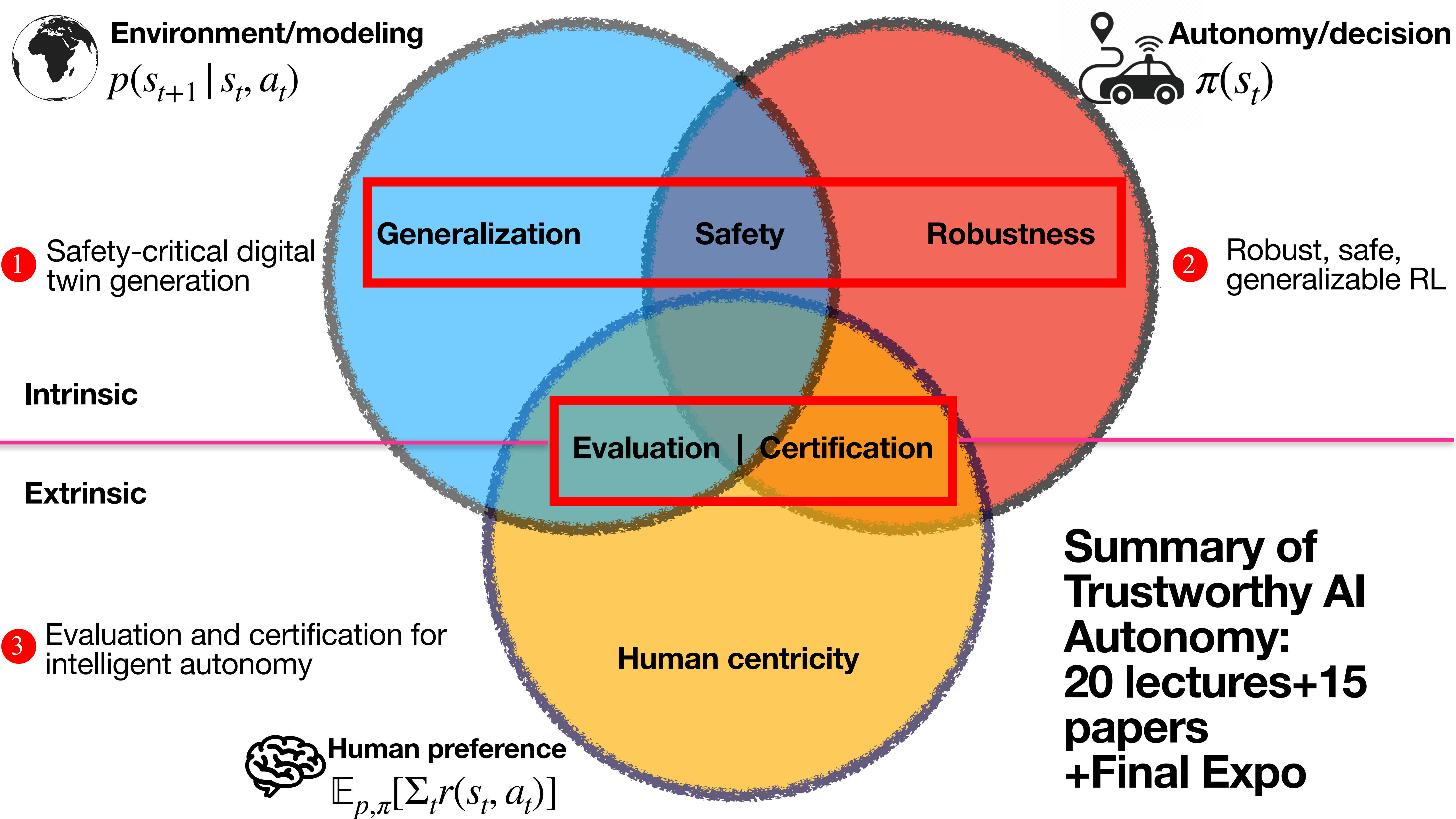
- So far, we've assumed there's a curator who we trust with access to all the raw data.
- What if a company (say Google) wants to learn a classifier from the images stored on everyone's phones, but without having to send the images to Google?
- Federated learning: learning a model without any centralized entity having access to all the data
  - Google sends the phone the current weights of the network
  - The phone does a small number of steps of gradient descent, and communicates the local update back to Google
  - Google updates their network by adding the local update
- Does this satisfy differential privacy?
  - Not automatically, but the local updates could be randomized in a way that makes them differentially private.

# Algorithmic fairness

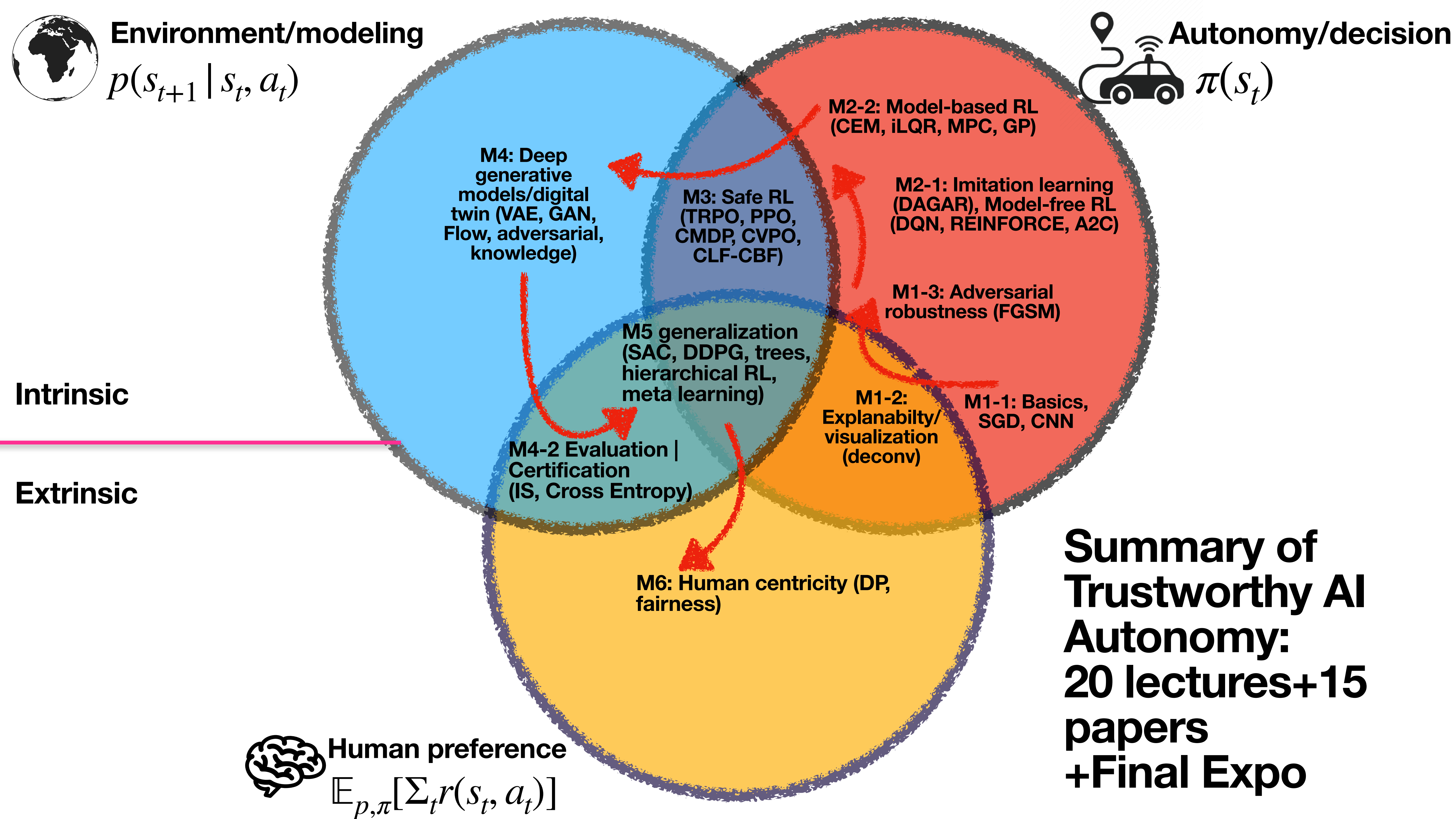
- Goal: identify and mitigate bias in ML-based decision making
- Sources of bias/discrimination
  - Data
    - Imbalanced/impoverished data
    - Labeled data imbalance (more data on white recidivism outcomes)
    - Labeled data incorrect / noisy (historical bias)
  - Model
    - ML prediction error imbalanced
    - Compound injustices
      - One “highly predictive” indicator of recidivism, Hellman posits, is a history of suffering child abuse.<sup>56</sup> Nonetheless, Hellman suggests, the state has “a strong reason” not to include this variable in its predictive model: If the state denies someone early release because he suffered child abuse, it will be adding to the harms caused by that earlier wrong.

# Definition of Fairness

- Notation:
  - $X$ : input to classifier
  - $S$ : sensitive feature (age, gender, race, etc.)
  - $Y$ : prediction
  - $T$ : true label
  - We use capital letters to emphasize that these are random variables
- Most common way to define fair classification is to require some invariance with respect to the sensitive attribute
  - Demographic parity:  $Y \perp S$
  - Equal opportunity:  
 $Y \perp S \mid T = t$ , for some  $t$
  - $\perp$  denotes stochastic independence.







# Closing Thoughts and Next Steps

Where to find good papers to read:  
Google scholar metrics

The image shows a screenshot of the Google Scholar interface. At the top left, the Google Scholar logo is visible. Below it, there is a navigation menu with a blue graduation cap icon and the text "Top publications". The main content area shows a breadcrumb trail: "Categories > Engineering & Computer Science > Subcategories". A dropdown menu is open, displaying a list of subcategories organized into three columns. The subcategories include:

- Architecture
- Artificial Intelligence
- Automation & Control Theory
- Aviation & Aerospace Engineering
- Bioinformatics & Computational Biology
- Biomedical Technology
- Biotechnology
- Ceramic Engineering
- Civil Engineering
- Combustion & Propulsion
- Computational Linguistics
- Computer Graphics
- Computer Hardware Design
- Computer Networks & Wireless Communication
- Computer Security & Cryptography
- Computer Vision & Pattern Recognition
- Computing Systems
- Data Mining & Analysis
- Databases & Information Systems
- Educational Technology
- Engineering & Computer Science (general)
- Environmental & Geological Engineering
- Evolutionary Computation
- Food Science & Technology
- Fuzzy Systems
- Game Theory and Decision Science
- Human Computer Interaction
- Library & Information Science
- Manufacturing & Machinery
- Materials Engineering
- Mechanical Engineering
- Medical Informatics
- Metallurgy
- Microelectronics & Electronic Packaging
- Mining & Mineral Resources
- Multimedia
- Nanotechnology
- Ocean & Marine Engineering
- Oil, Petroleum & Natural Gas
- Operations Research
- Plasma & Fusion
- Power Engineering
- Quality & Reliability
- Radar, Positioning & Navigation
- Remote Sensing
- Robotics
- Signal Processing
- Software Systems
- Structural Engineering
- Sustainable Energy
- Technology Law
- Textile Engineering
- Theoretical Computer Science
- Transportation
- Water Supply & Treatment
- Wood Science & Technology

# Top venues for AI

Premium confs

Top confs

Top journals  
Less theoretical


Premium venues  
emphasizing on  
theories

A lot of free tutorials/  
workshops

Categories > Engineering & Computer Science > Artificial Intelligence ▾

	Publication	h5-index	h5- median
1.	International Conference on Learning Representations	<u>253</u>	470
2.	Neural Information Processing Systems	<u>245</u>	422
3.	International Conference on Machine Learning	<u>204</u>	370
4.	AAAI Conference on Artificial Intelligence	<u>157</u>	240
5.	IEEE Transactions On Systems, Man And Cybernetics Part B, Cybernetics	<u>127</u>	172
6.	IEEE Transactions on Neural Networks and Learning Systems	<u>119</u>	171
7.	Neurocomputing	<u>119</u>	164
8.	Expert Systems with Applications	<u>118</u>	164
9.	International Joint Conference on Artificial Intelligence (IJCAI)	<u>105</u>	174
10.	Applied Soft Computing	<u>103</u>	133
11.	Journal of Machine Learning Research	<u>96</u>	165
12.	IEEE Transactions on Fuzzy Systems	<u>96</u>	128
13.	Knowledge-Based Systems	<u>96</u>	127
14.	Neural Computing and Applications	<u>83</u>	115
15.	Neural Networks	<u>72</u>	105
16.	International Conference on Artificial Intelligence and Statistics	<u>68</u>	101

# Top venues for robotics



Conference on  
Robot Learning

L4DC - Learning for  
Dynamics & Control  
Conference

Categories > Engineering & Computer Science > Robotics ▾

	Publication	h5-index	h5- median
1.	IEEE International Conference on Robotics and Automation	105	178
2.	IEEE Robotics and Automation Letters	74	104
3.	IEEE/RSJ International Conference on Intelligent Robots and Systems	73	108
4.	IEEE/ASME Transactions on Mechatronics	71	95
5.	Science Robotics	67	125
6.	The International Journal of Robotics Research	65	108
7.	IEEE Transactions on Robotics	65	91
8.	Robotics and Autonomous Systems	58	91
9.	Robotics and Computer-Integrated Manufacturing	58	82
10.	Robotics: Science and Systems	50	100
11.	ACM/IEEE International Conference on Human Robot Interaction	50	71
12.	Journal of Field Robotics	48	67
13.	Autonomous Robots	48	63
14.	Journal of Intelligent & Robotic Systems	43	65
15.	Soft Robotics	43	65

Categories > Engineering & Computer Science > Transportation ▾

Publication

1. IEEE Transactions on Intelligent Transportation Systems
2. Transportation Research Part C: Emerging Technologies
3. Transportation Research Part A: Policy and Practice
4. Transportation Research Part B: Methodological