Trustworthy AI Autonomy

Human Centricity

Ding Zhao

Assistant Professor

Carnegie Mellon University

Carnegie Mellon University

Safe AI Lab @CMU

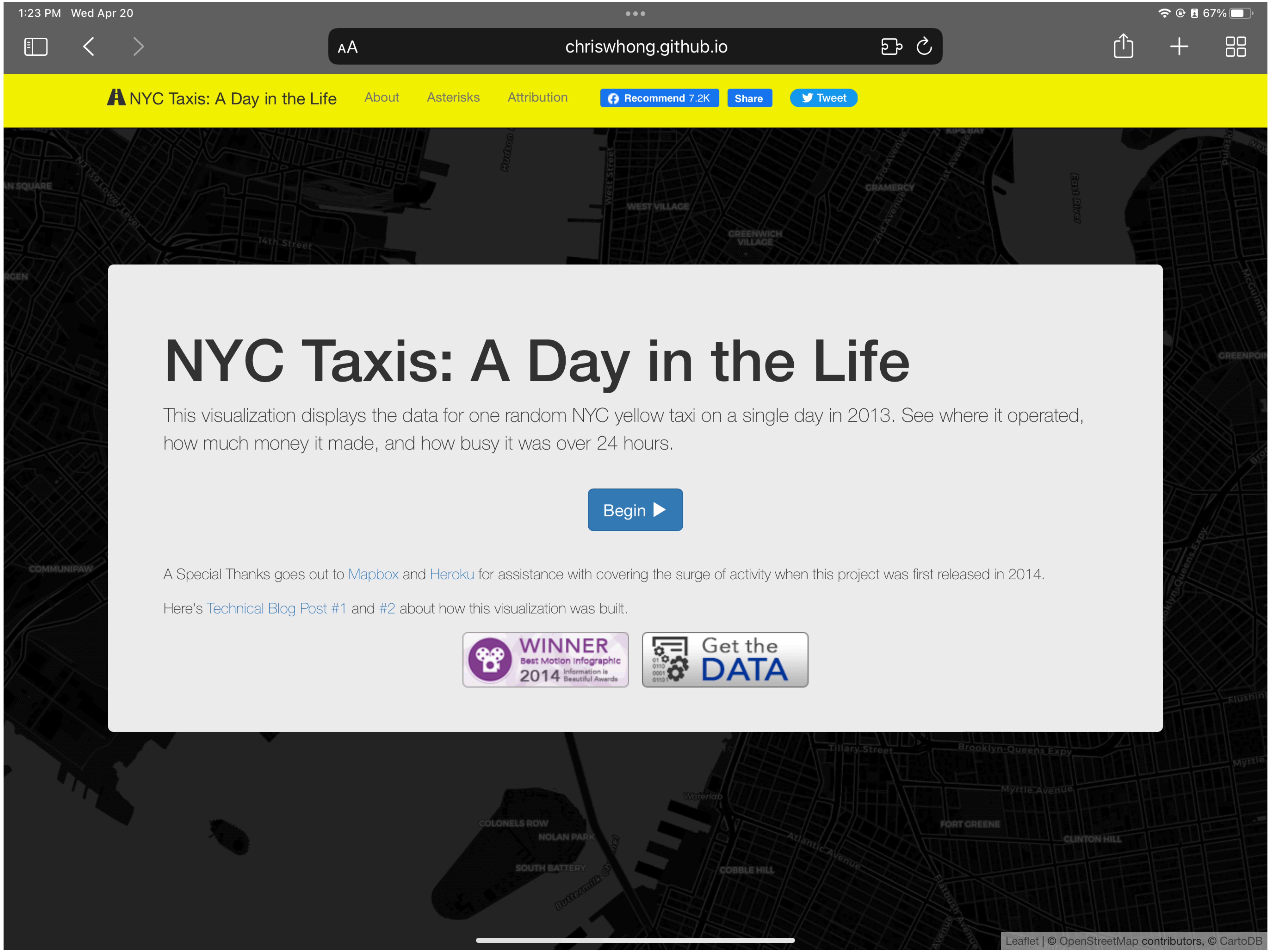
Plan for today

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

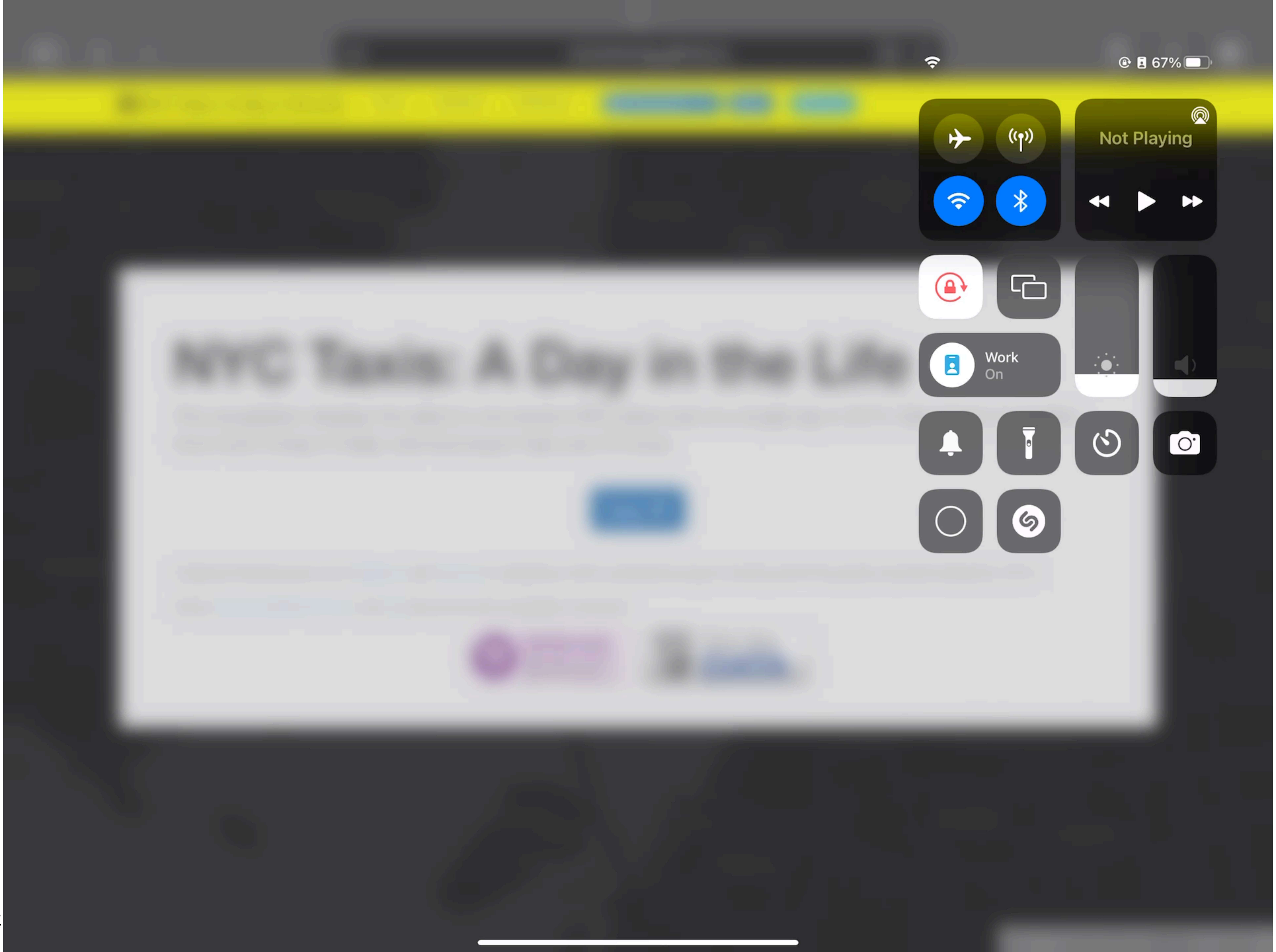
Why should I be mindful of privacy as an engineer?

- 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



https://chriswhong.github.io/nyctaxi/

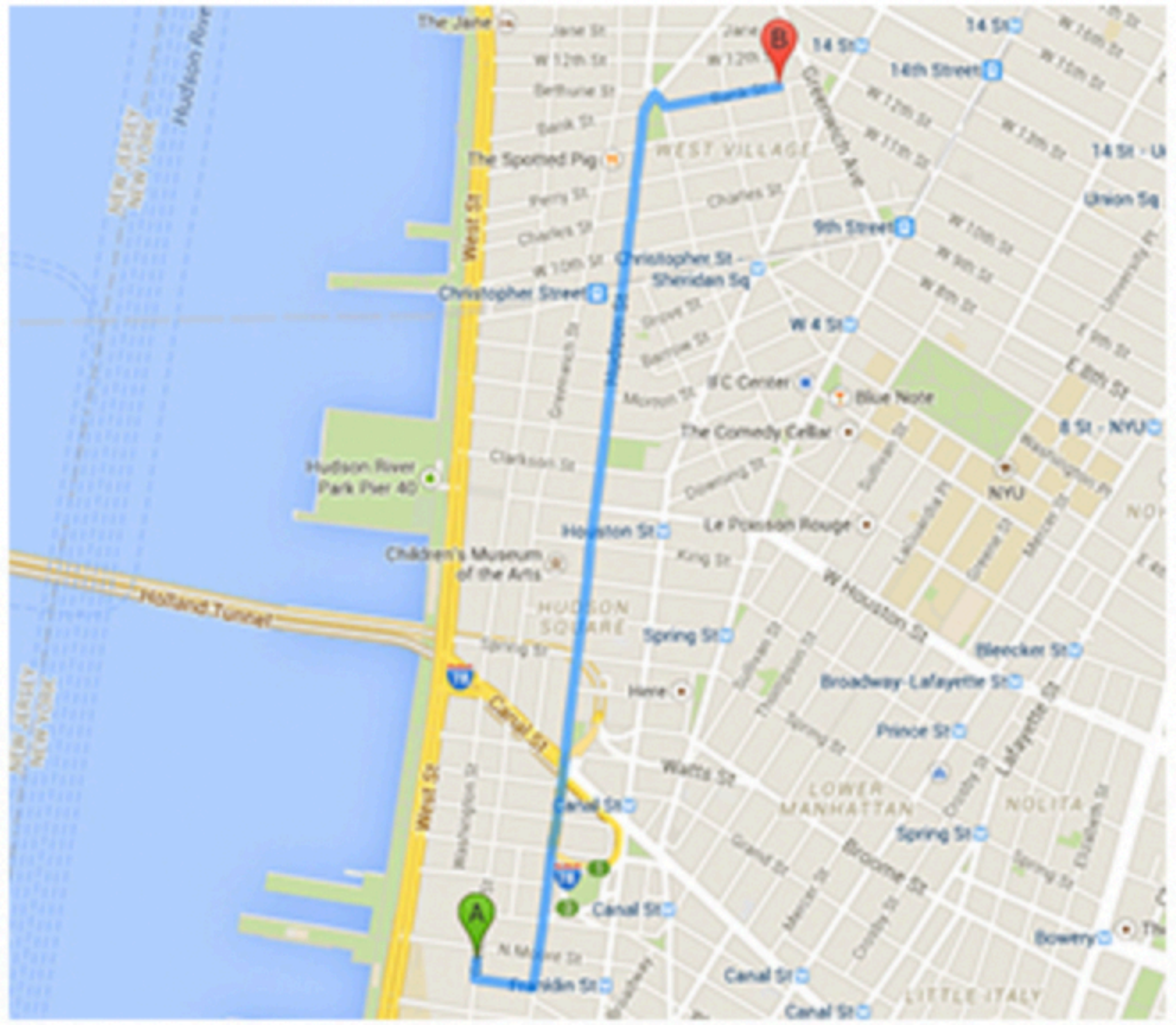

67%
Not Playing
Work
On

Public NYC Taxicab Database Lets You See How Celebrities Tip

J.K. Trotter 142.43K
10/23/14 12:00PM Filed to: DATA

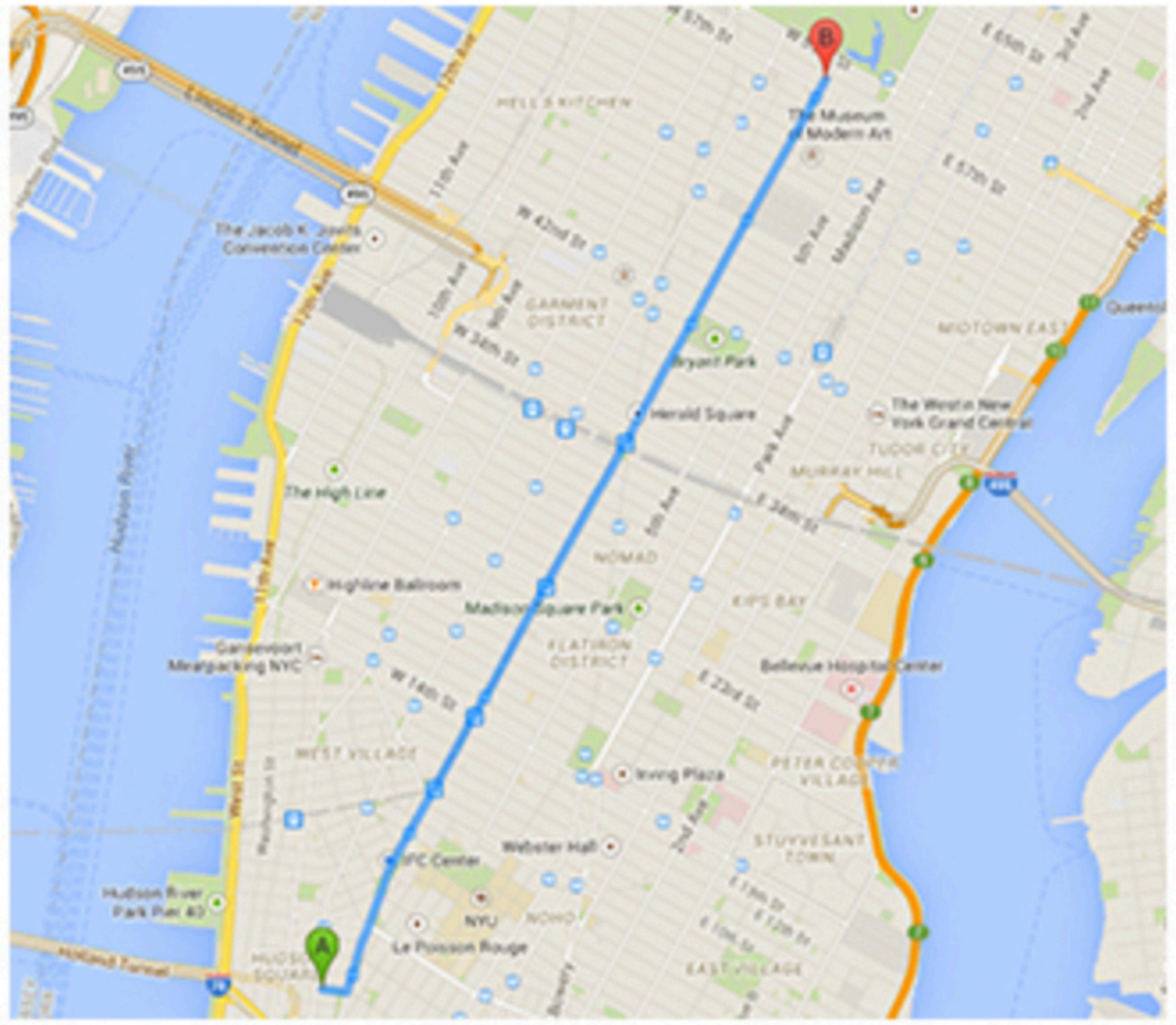
Would it be safe to use data published voluntarily? - No

BRADLEY COOPER



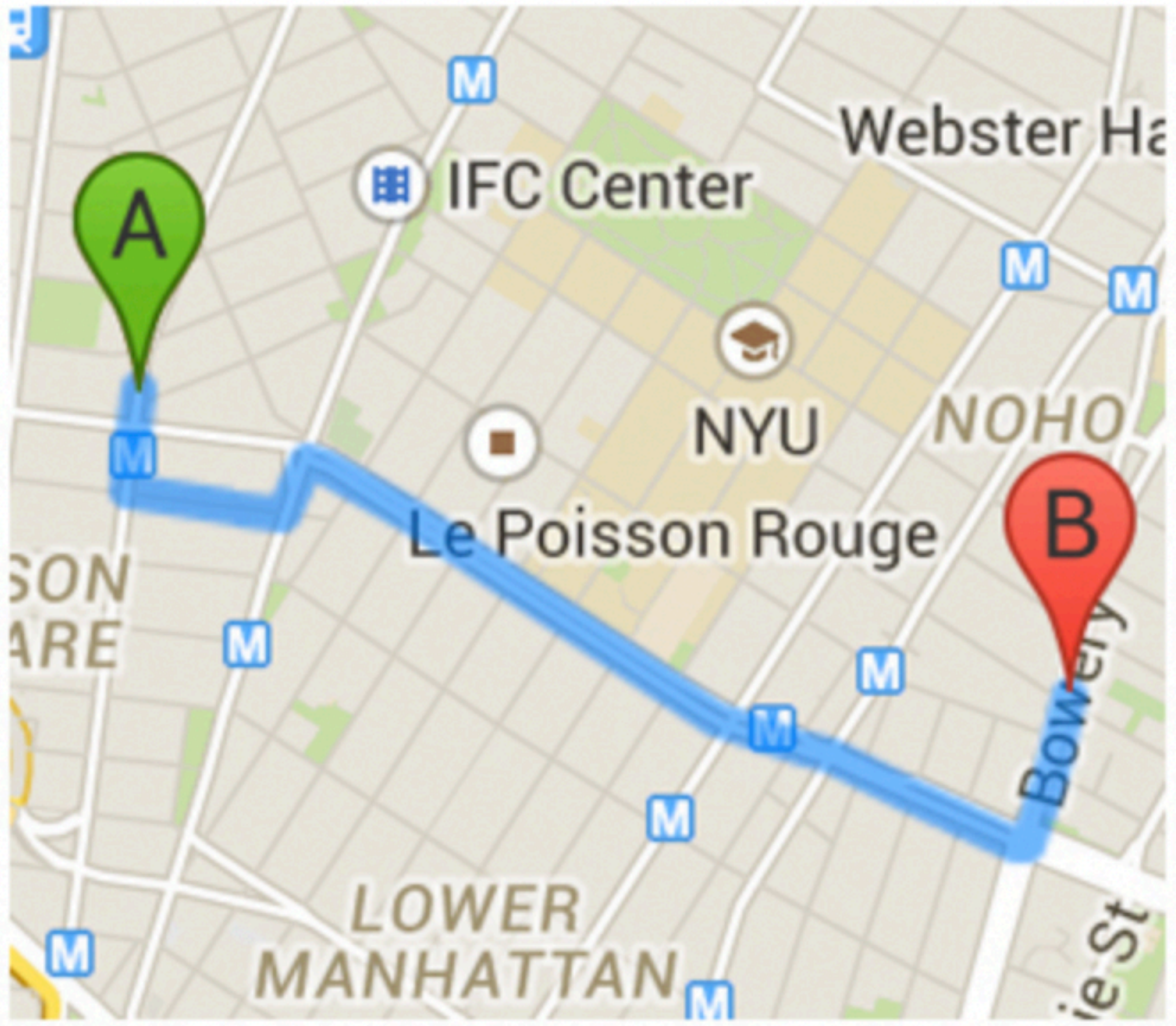
JULY 8, 2013 • 7:34 PM - 7:44 PM
376 GREENWICH ST. TO 13 BANK ST.

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

https://www.gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546?utm_campaign=socialflow_gawker_twitter&utm_source=gawker_twitter&utm_medium=socialflow

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

Would it be safe to use anomymous data? — No

Antonymy is not enough

Bits 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

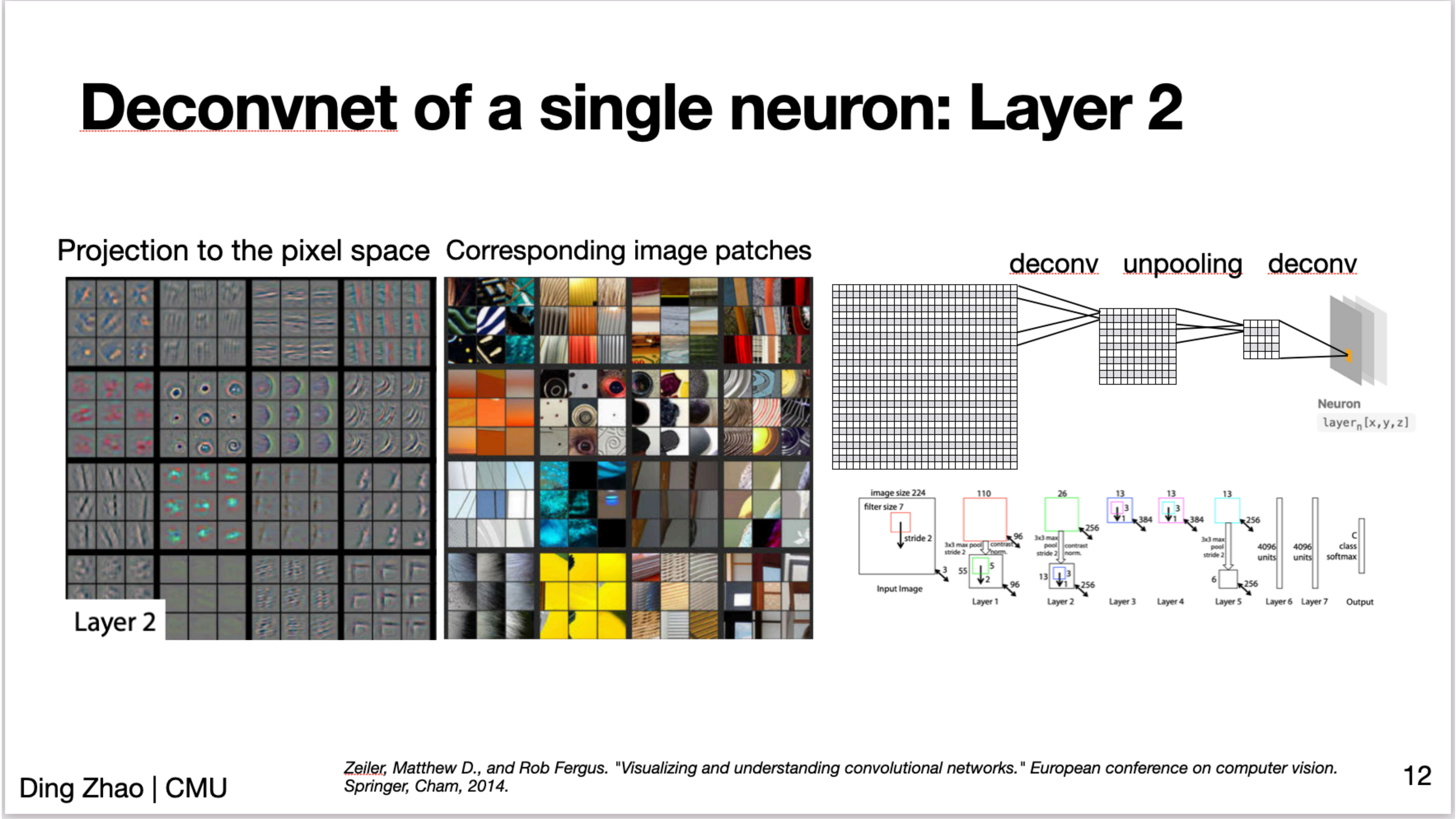
Save

More

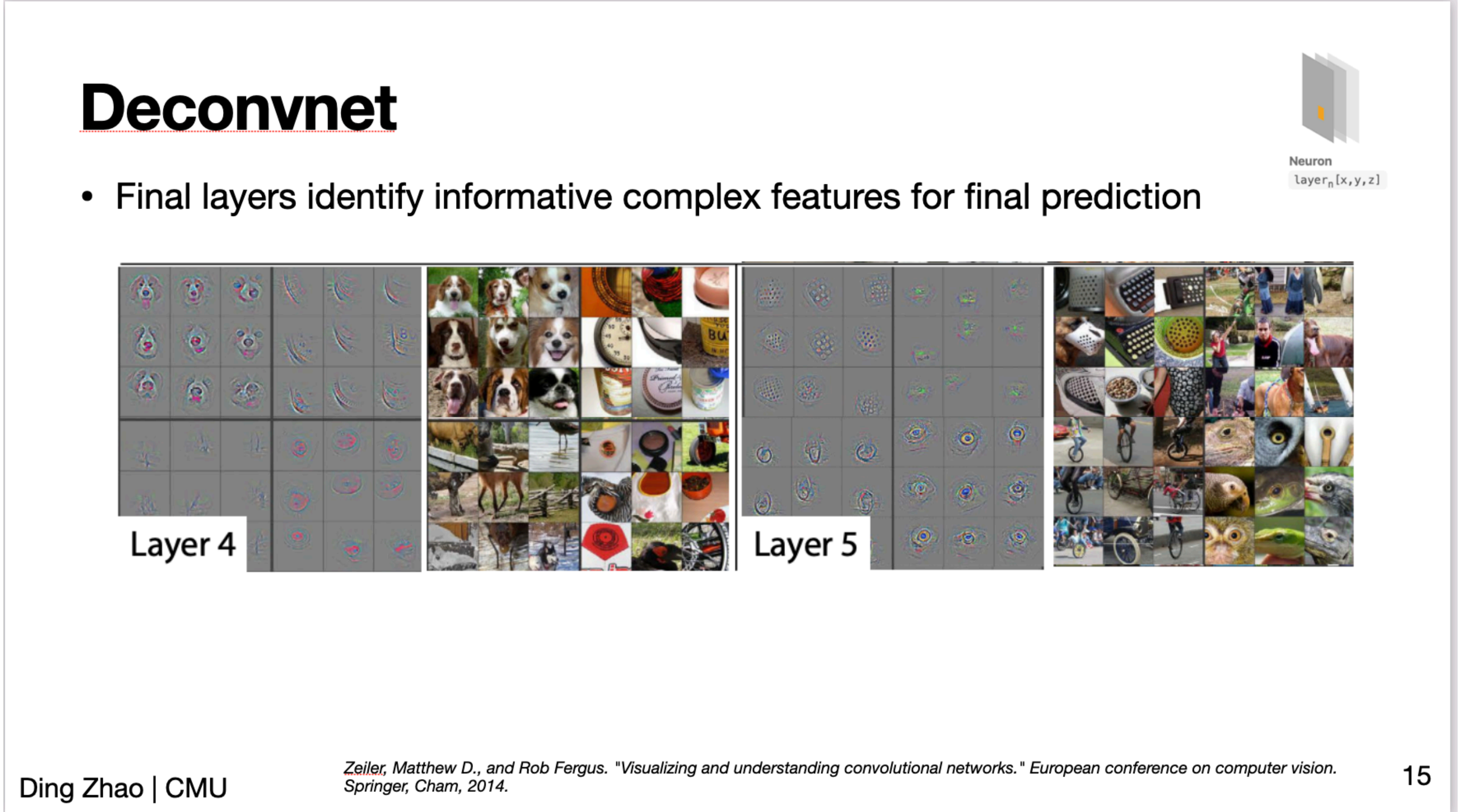
Yves-Alexandre de Montjoye, a graduate student at the

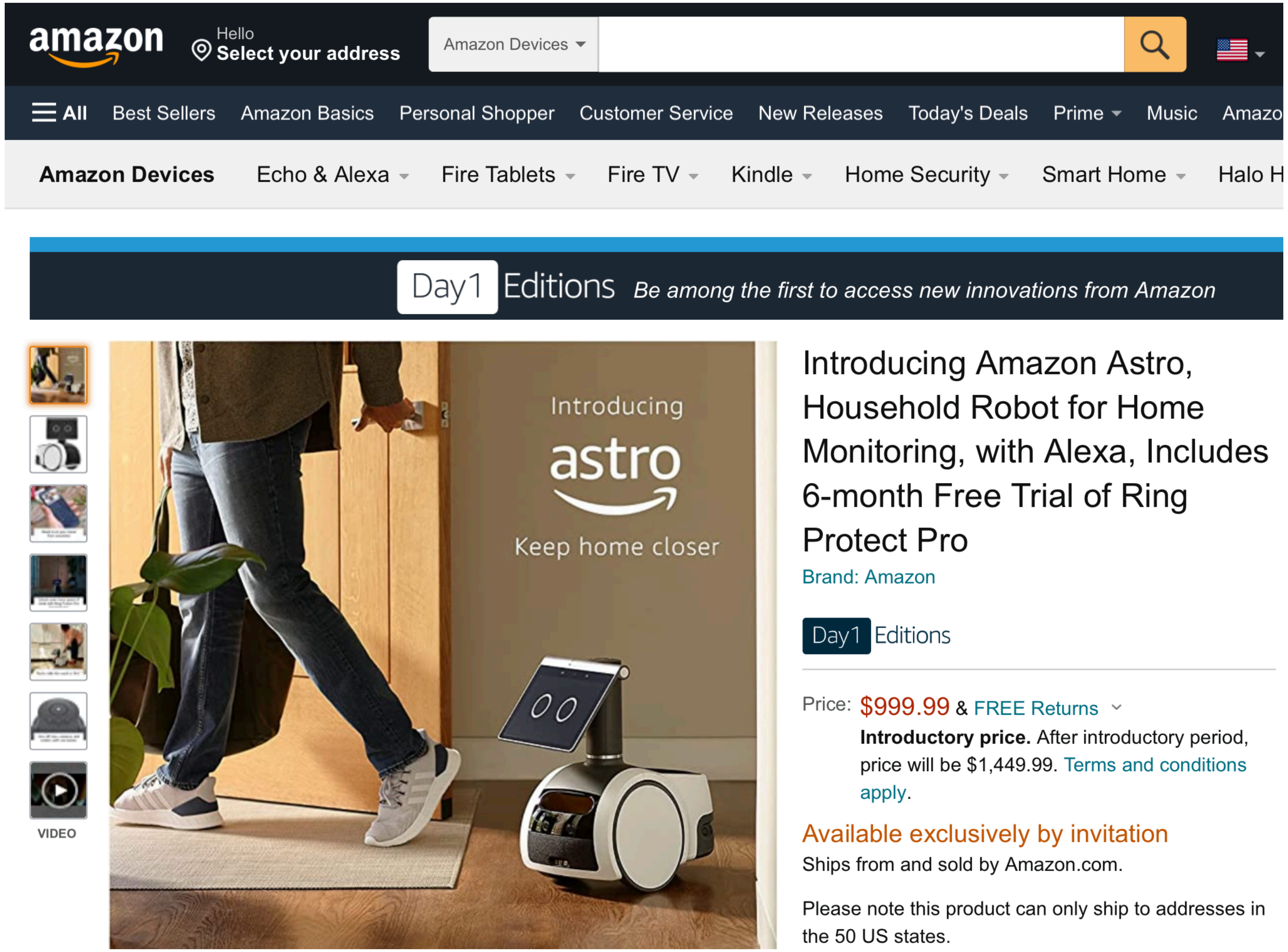
So, it seems we should not share any data, then we are safe? — still not true.

Recall this slide in M1-2 Explanation

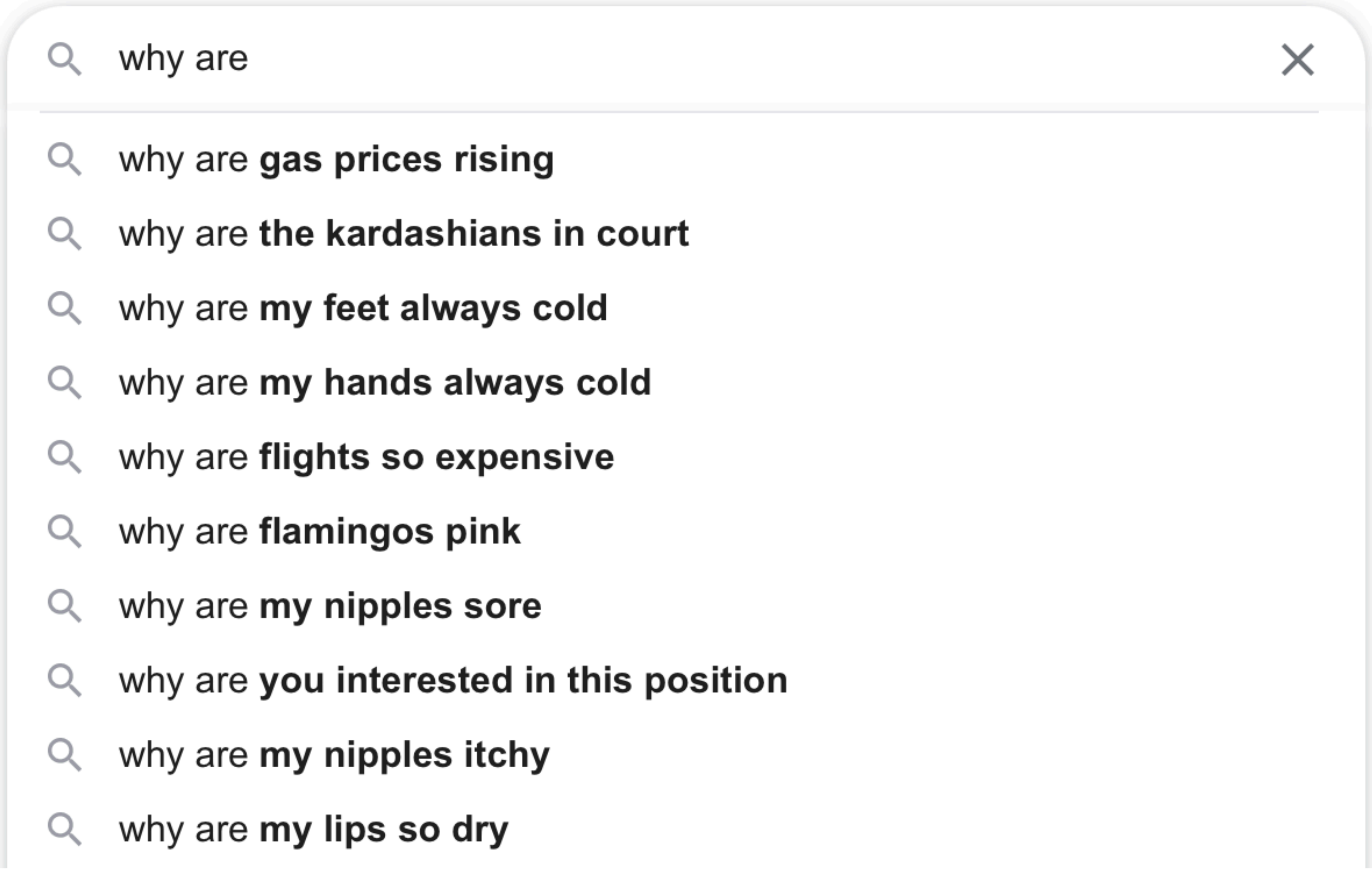


Recover images from algorithms



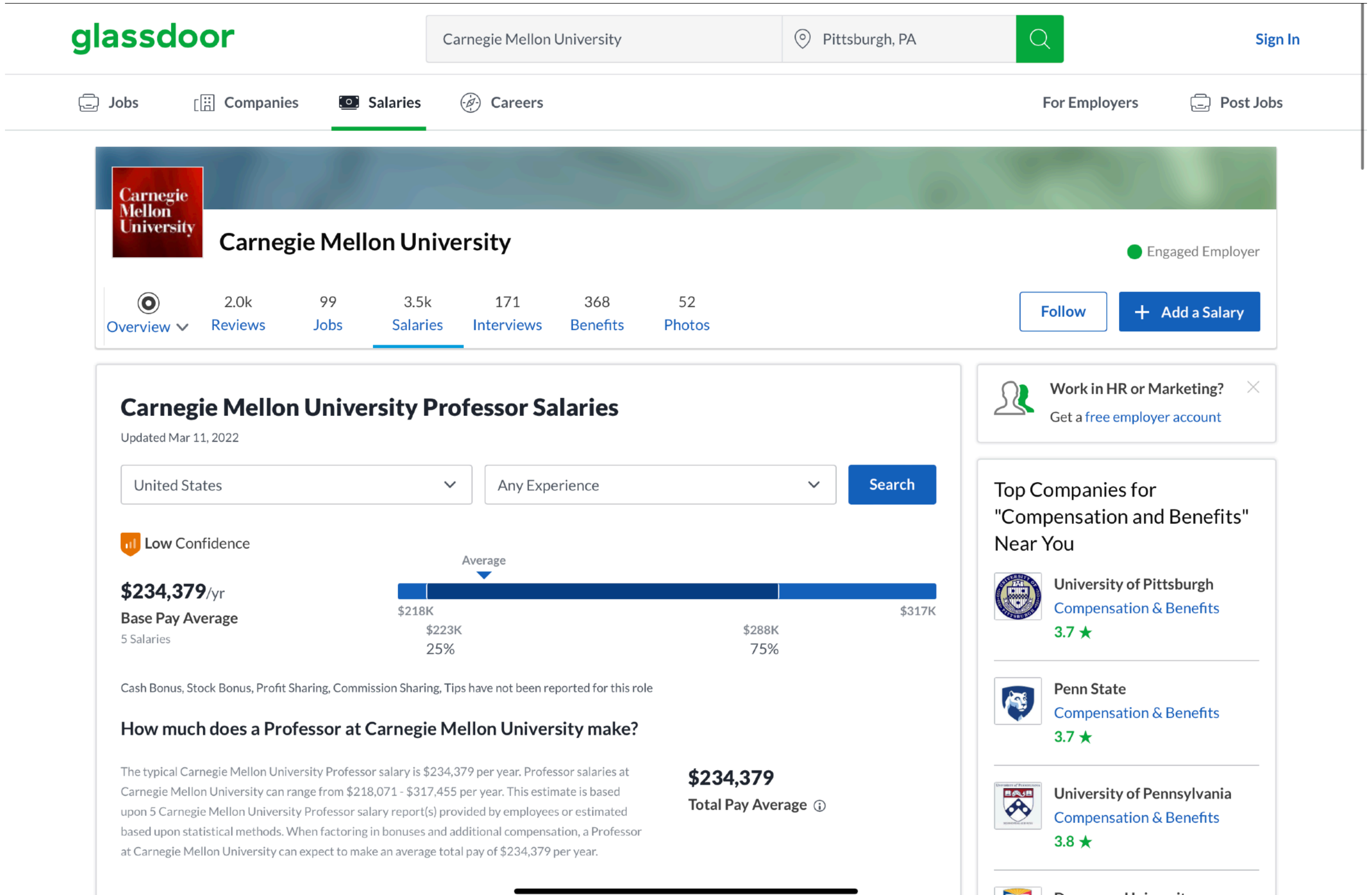

amazon
Hello
Select your address
Amazon Devices
All
Best Sellers
Amazon Basics
Personal Shopper
Customer Service
New Releases
Today's Deals
Prime
Music
Amazon Devices
Echo & Alexa
Fire Tablets
Fire TV
Kindle
Home Security
Smart Home
Day1 Editions
Be among the first to access new innovations from Amazon
Introducing
astro
Keep home closer
VIDEO
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



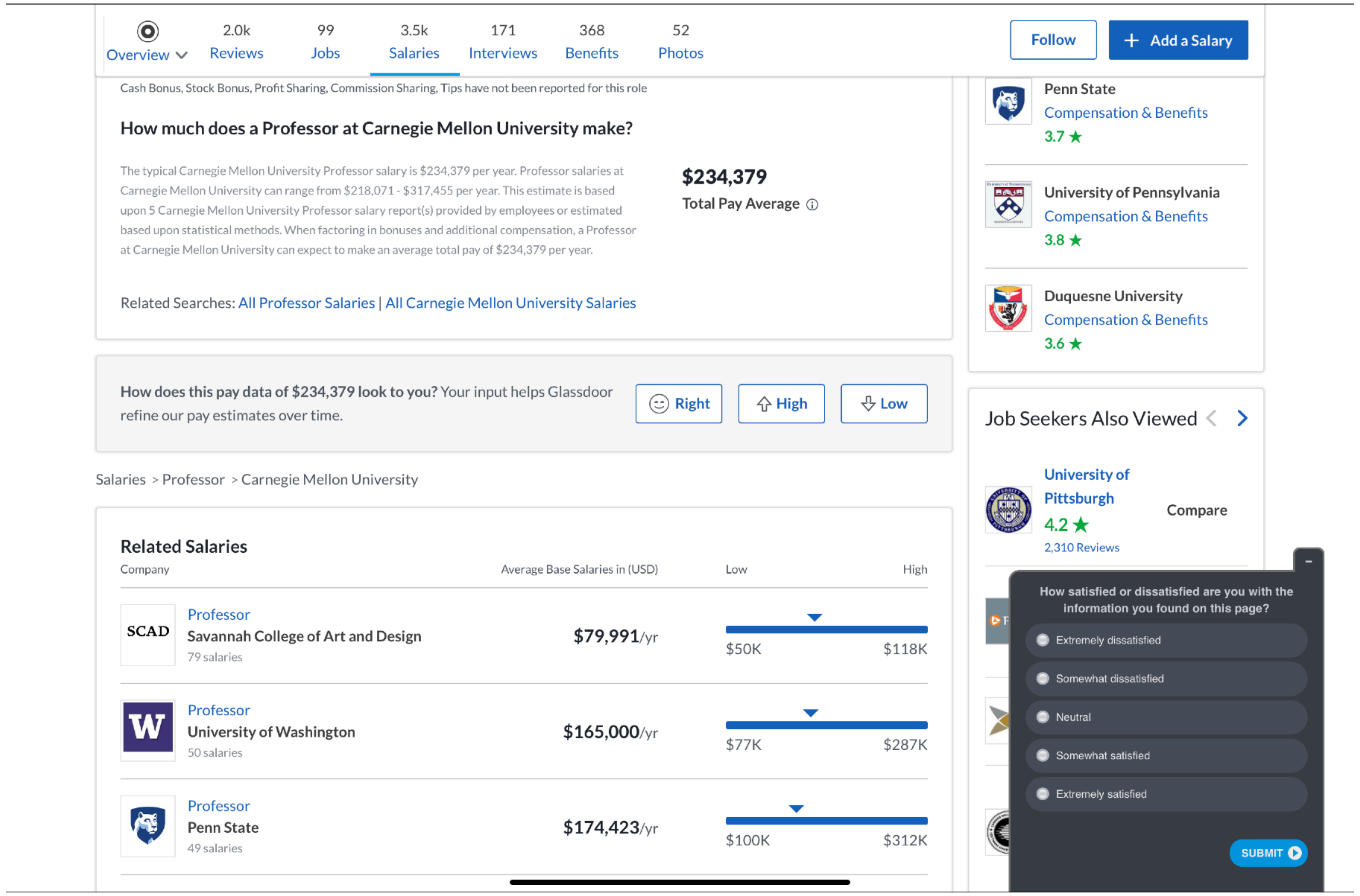
What is a privacy-preserving algorithm?

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



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.



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**

“The Algorithmic Foundations of Differential Privacy”. Dwork and Roth. Foundations and Trends in Theoretical Computer Science, NOW Publishers. 201

Neighboring

- Two data sets D_1 and D_2 if differ on ≤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

ε-differential privacy

- Algorithm $\mathscr{A}$ is ε-differentially private if:

- For all pairs of neighboring sets D_1, D_2 and any set R of possible output (response)

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

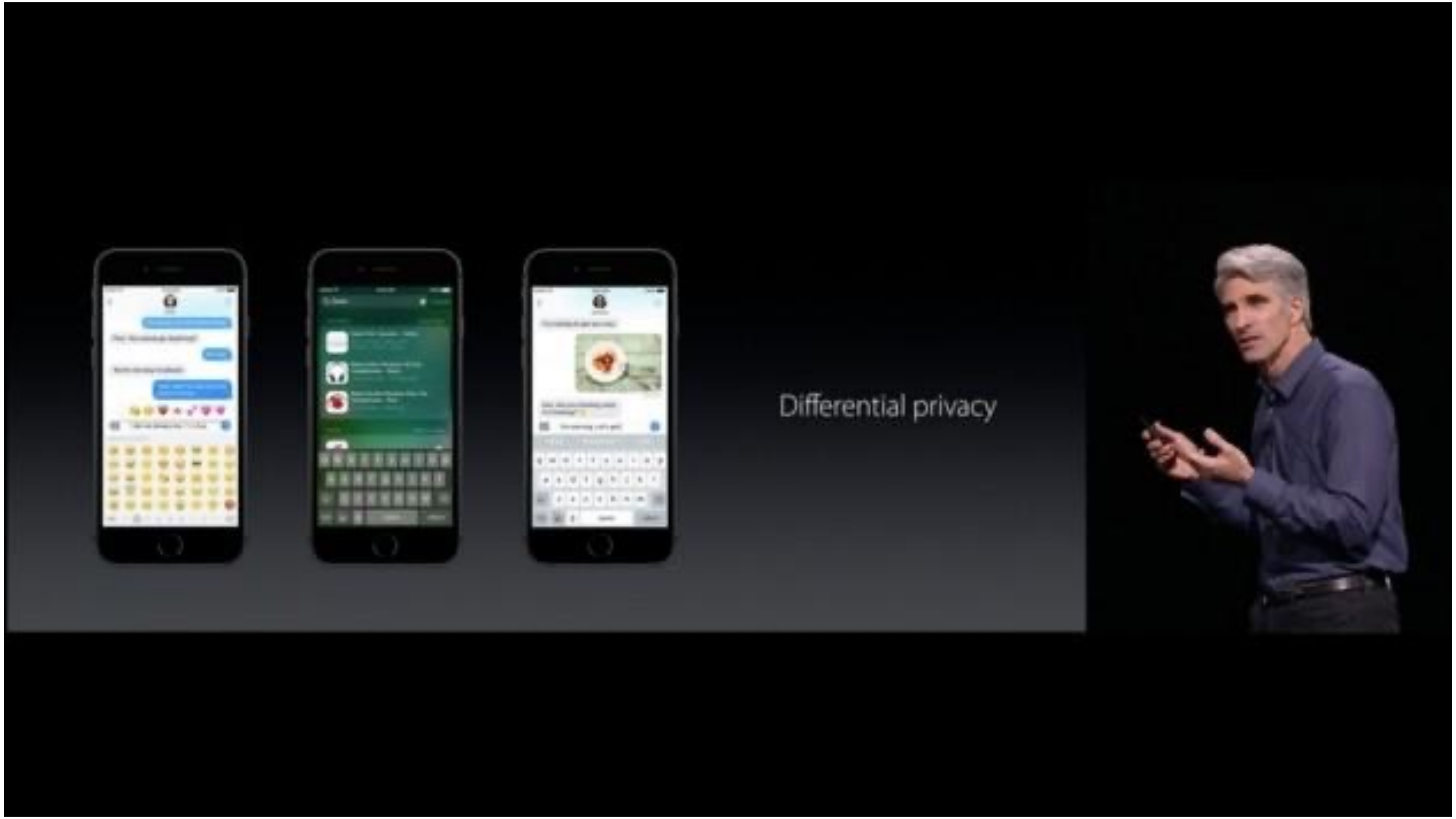
- Note: for small ε, $e^{\varepsilon} \approx 1 + \varepsilon$

- A consequence: for any possible response y

$$\exp(-\varepsilon) \leq \frac{\Pr(\mathscr{A}(\mathscr{D}_1) = y)}{\Pr\left(\mathscr{A}(\mathscr{D}_2) = y\right)} \leq \exp(\varepsilon)$$

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.

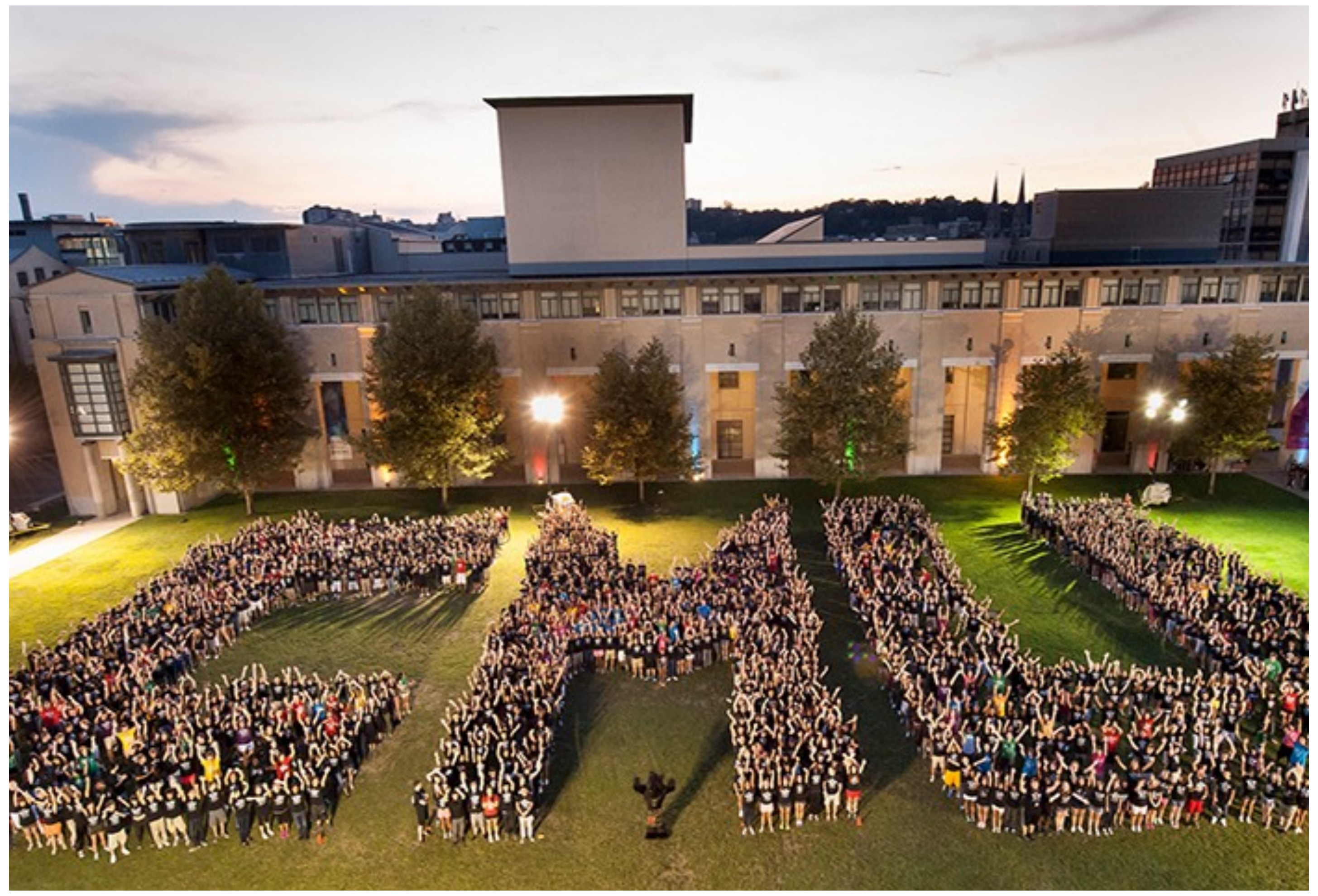


https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Tools for designing privacy-preserving algorithms

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

CMU faculty

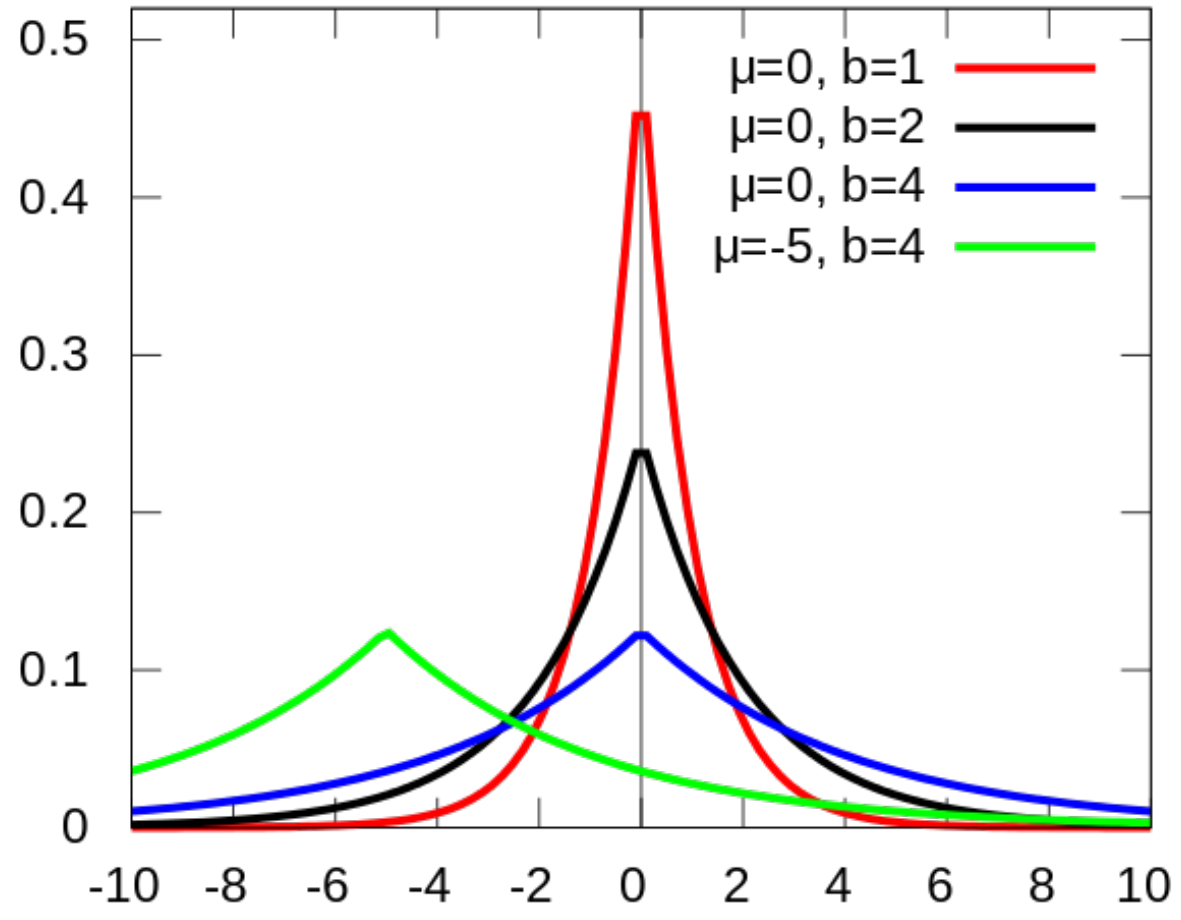


ME faculty

ME faculty joined in 2018

Laplace mechanism

- Goal: Evaluate $f{:}D \to \mathbb{R}$ mapping datasets to $\mathbb{R}$; preserve ϵ-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: $Z_D \sim \mathrm{Lap}(f(D), \frac{\Delta_f}{\varepsilon})$
- Note: adding Gaussian will violate the DP requirement.



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

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.

Slides adopted from Roger Grosse
https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

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.56 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

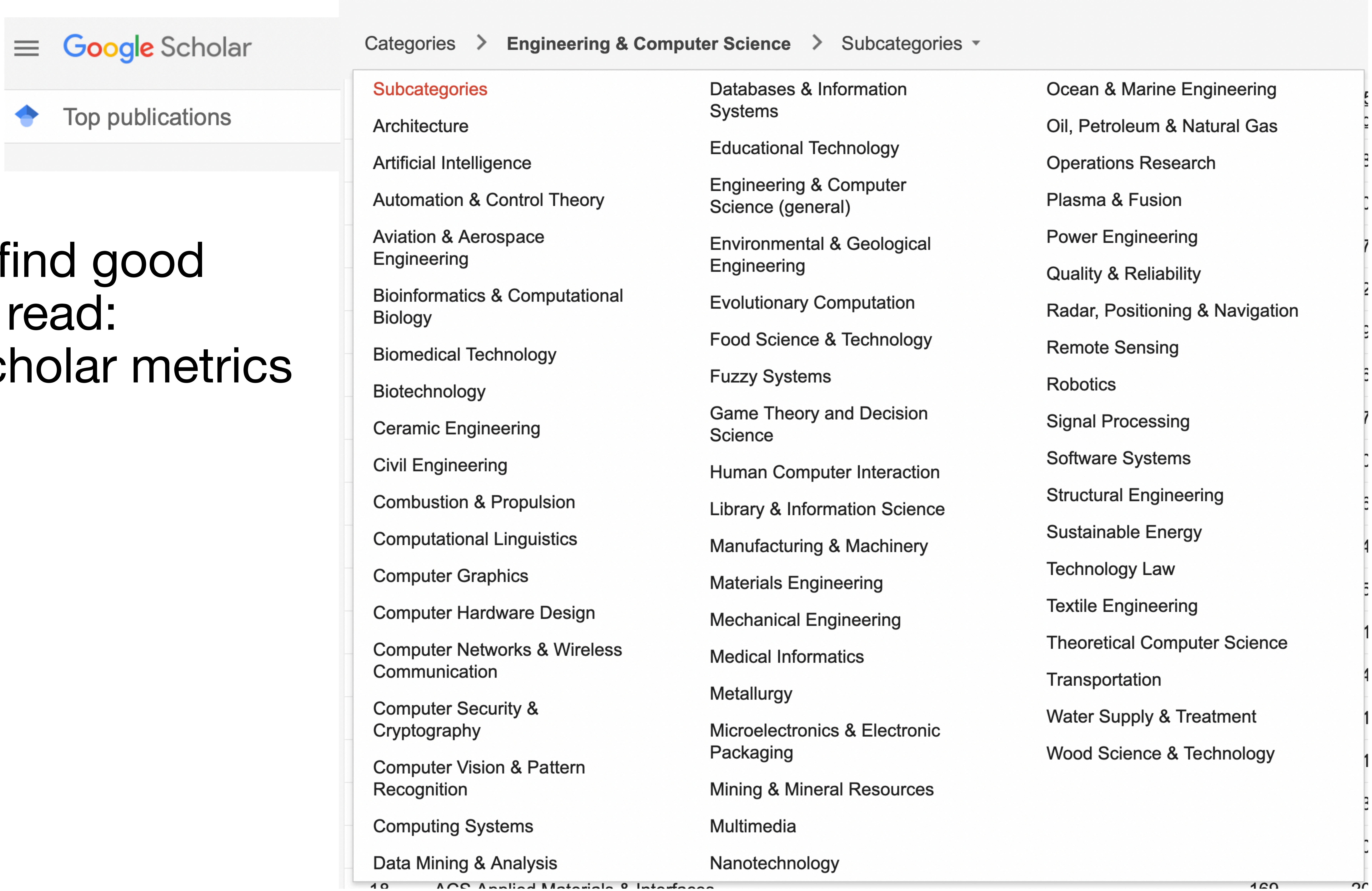
- Notation:
    - X: input to classifier
    - S: sensitive feature (age, gender, race, etc.)
    - Y: prediction
    - T: true label
    - We use capital letters to emphasize that these are random variables

- 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.


Environment/modeling
$p(s_{t+1} \mid s_t, a_t)$
Autonomy/decision
$\pi(s_t)$
Generalization
Safety
Robustness
1 Safety-critical digital twin generation
2 Robust, safe, generalizable RL
Intrinsic
Evaluation | Certification
Extrinsic
Human centricity
3 Evaluation and certification for intelligent autonomy
Human preference
$\mathbb{E}_{p,\pi}[\Sigma_t r(s_t, a_t)]$
Summary of Trustworthy AI Autonomy: 20 lectures+15 papers +Final Expo

Summary of Trustworthy AI Autonomy: 20 lectures+15 papers +Final Expo

Closing Thoughts and Next Steps



Where to find good papers to read:
Google scholar metrics

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

Top venues for robotics

Conference on Robot Learning

L4DC - Learning for Dynamics & Control Conference

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

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