CS440/ECE448 Lecture 12: Privacy

Mark Hasegawa-Johnson, 2/2023

“One nation under CCTV,” Banksy, 2008. CC-SA 2.0, https://commons.wikimedia.org/wiki/File:Bansky_one_nation_under_cctv.jpg

Outline

- Why it matters
  - Privacy laws and lawsuits
  - On the other hand: the right to representation in machine learning
- Data security
  - The failure of k-anonymity
  - Data-centric information security
- Algorithmic methods
  - Federated learning
  - Differential privacy
- How to collect data so that you can share it legally, and why you should do so

Illinois Biometric Information Privacy Act

740 ILCS 14/15: An individual or company can hold biometric data (voice, face) of any person living in Illinois only if:

a) They have a written policy
b) They have obtained your consent
c) They do not profit from it
d) They don't give it away without your consent
e) They protect it from data theft

If any of the above is violated, you can sue them, even if the violation didn't hurt you.

General Data Privacy Regulation (GDPR)

Europeans have the right to:

- Learn where their data is stored, and access to it
- Have their data stored in a manner that prevents unauthorized release
- Correct their data if there are mistakes
- Object to processing of their data, using a binary option that is clearly described and that does not try to hide the “no” option

Data may not be transferred to other countries or international organizations unless the EU has determined that the recipient has adequate data privacy safeguards.

GDPR violations may be fined up to 10 million Euros, or 2% of your global gross revenue, whichever is higher!!!

Outline

- Why it matters
  - Privacy laws and lawsuits
  - On the other hand: the right to representation in machine learning
- Data security
  - The failure of k-anonymity
  - Data-centric information security
- Algorithmic methods
  - Federated learning
  - Differential privacy
- How to collect data so that you can share it legally, and why you should do so

On the other hand: the right to representation in machine learning

- Koenecke et al.(doi:10.1073/pnas.1915768117, 2020) tested automatic speech recognition software published by Amazon, Apple, Google, IBM and Microsoft
- Data: autobiographical monologs by black (73) and white (42) people
- Result: word error rate was 35% for black speakers, 19% for white speakers
- Why:
  - Training data includes more white people than black people.
  - The variability in the speaking styles of different white people is well-represented in training data, but the variability in speaking styles of different black people is not well-represented.

Outline

- Why it matters
    - Privacy laws and lawsuits
    - On the other hand: the right to representation in machine learning
- Data security
    - The failure of k-anonymity
    - Data-centric information security
- Algorithmic methods
    - Federated learning
    - Differential privacy
- How to collect data so that you can share it legally, and why you should do so

The failure of k-anonymity

- K-anonymity is an intuitively obvious idea: if data are binned in buckets (lower table at right), then each person is identical to K-1 other people.
- Unfortunately, no guaranteed K-anonymizing algorithms exist. Many datasets that seem to be K-anonymized have been successfully de-anonymized.

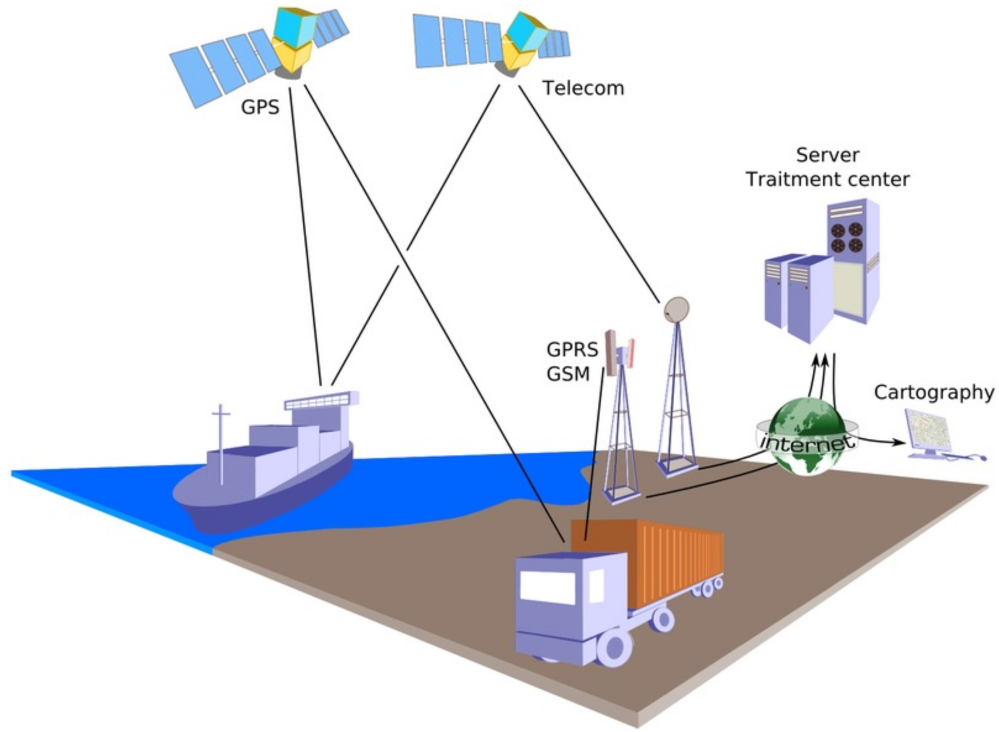
Name	Age	Gender	Height	Weight	State of domicile	Religion	Disease
Ramsha	30	Female	165cm	72kg	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	162cm	70kg	Kerala	Hindu	Viral infection
Salima	28	Female	170cm	68kg	Tamil Nadu	Muslim	Tuberculosis
Sunny	27	Male	170cm	75kg	Karnataka	Parsi	No illness
Joan	24	Female	165cm	71kg	Kerala	Christian	Heart-related
Bahuksana	23	Male	160cm	69kg	Karnataka	Buddhist	Tuberculosis
Rambha	19	Male	167cm	85kg	Kerala	Hindu	Cancer
Kishor	29	Male	180cm	81kg	Karnataka	Hindu	Heart-related
Johnson	17	Male	175cm	79kg	Kerala	Christian	Heart-related

Name	Age	Gender	Height	Weight	State of domicile	Religion	Disease
*	$20 < \text{Age} \leq 30$	Female	165cm	72kg	Tamil Nadu	*	Cancer
*	$20 < \text{Age} \leq 30$	Female	162cm	70kg	Kerala	*	Viral infection
*	$20 < \text{Age} \leq 30$	Female	170cm	68kg	Tamil Nadu	*	Tuberculosis
*	$20 < \text{Age} \leq 30$	Male	170cm	75kg	Karnataka	*	No illness
*	$20 < \text{Age} \leq 30$	Female	165cm	71kg	Kerala	*	Heart-related
*	$20 < \text{Age} \leq 30$	Male	160cm	69kg	Karnataka	*	Tuberculosis
*	$\text{Age} \leq 20$	Male	167cm	85kg	Kerala	*	Cancer
*	$20 < \text{Age} \leq 30$	Male	180cm	81kg	Karnataka	*	Heart-related
*	$\text{Age} \leq 20$	Male	175cm	79kg	Kerala	*	Heart-related
*	$\text{Age} \leq 20$	Male	169cm	82kg	Kerala	*	Viral infection

Example: Geolocation data

Montjoye et al. (doi:10.1038/srep01376, 2013) showed that,

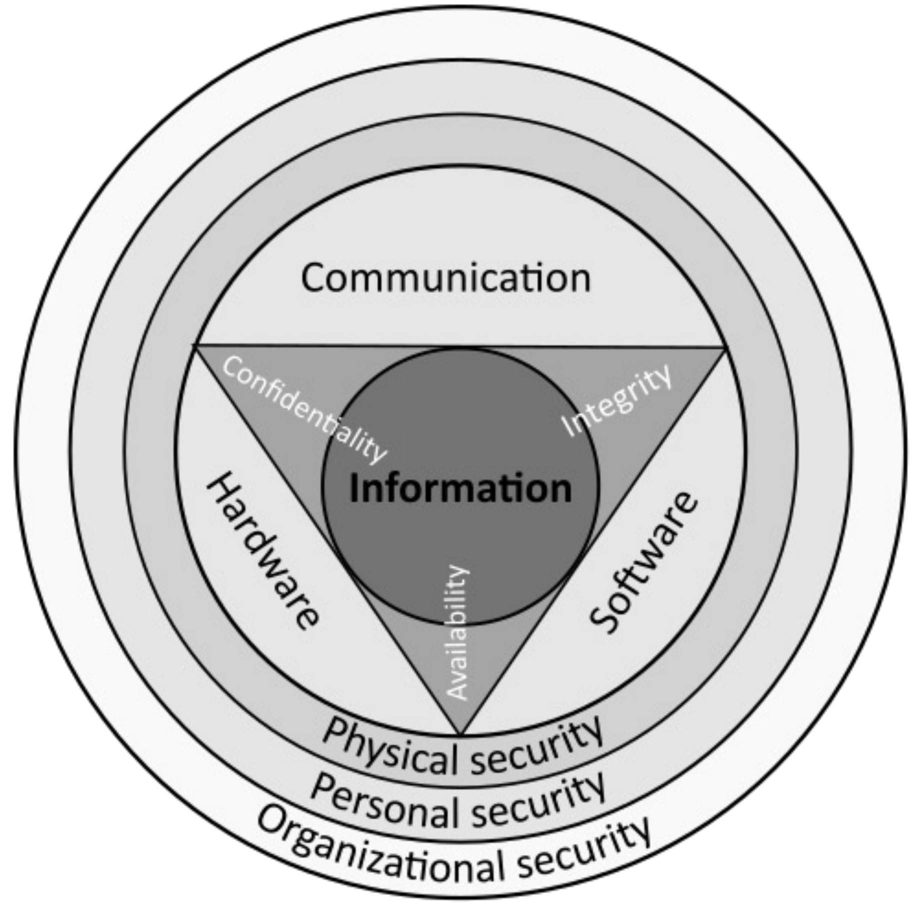
- In a database of 1.5 million people,
- given the ID # of the cell-phone base station closest to a user at 4 different times,
- it is possible to uniquely identify 95% of all users.



GFDL, Éric Chassaing,
https://commons.wikimedia.org/wiki/File:Geolocation.png

Data security

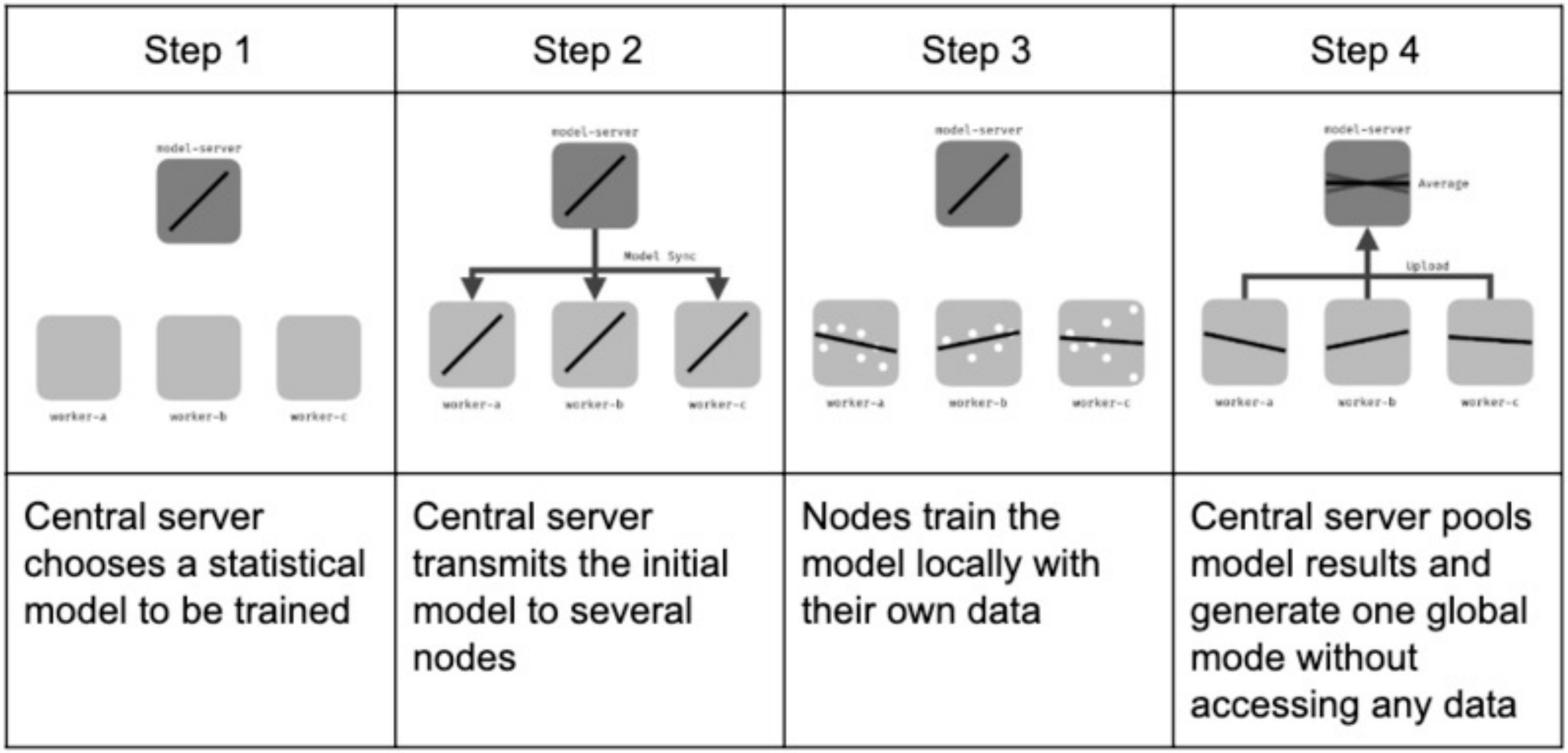
- Discover: know what data you have
- Manage: create a policy specifying who has access to each byte of data
- Protect:
  - Software: ensure that data can only be communicated via tools that guarantee the management policy
  - Hardware: ensure that when you throw hardware away, the data is wiped first
- Monitor: monitor data usage to detect deviation from normal behavior



Outline

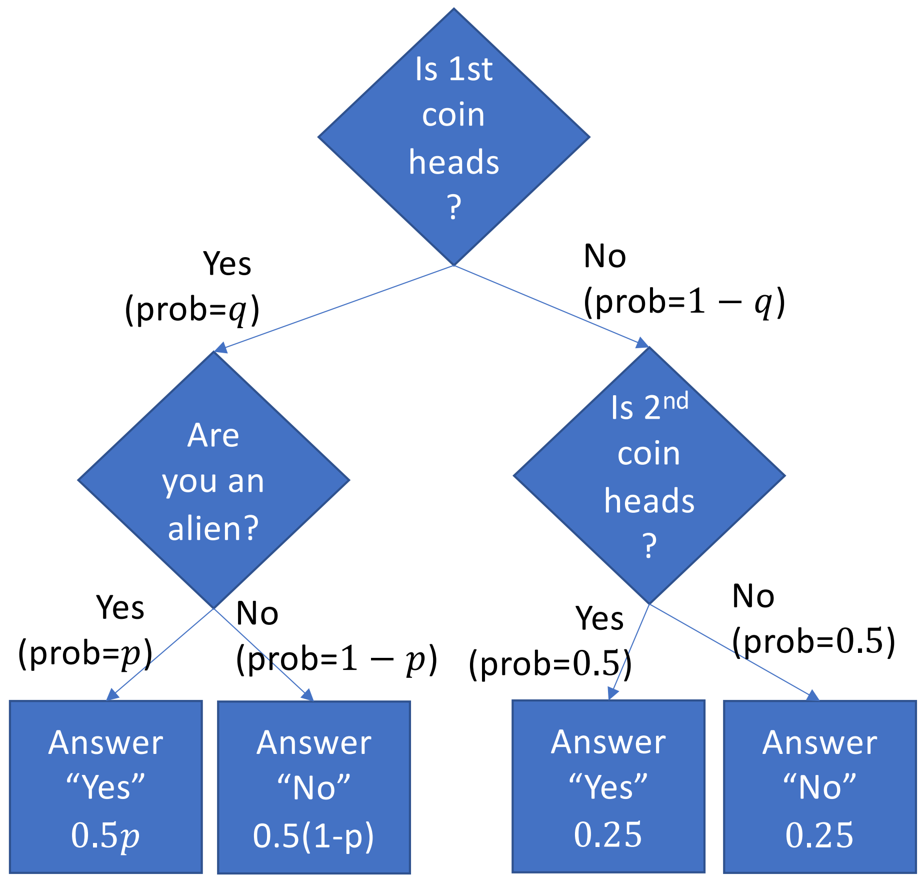
- Why it matters
  - Privacy laws and lawsuits
  - On the other hand: the right to representation in machine learning
- Data security
  - The failure of k-anonymity
  - Data-centric information security
- Algorithmic methods
  - Federated learning
  - Differential privacy
- How to collect data so that you can share it legally, and why you should do so

Federated learning



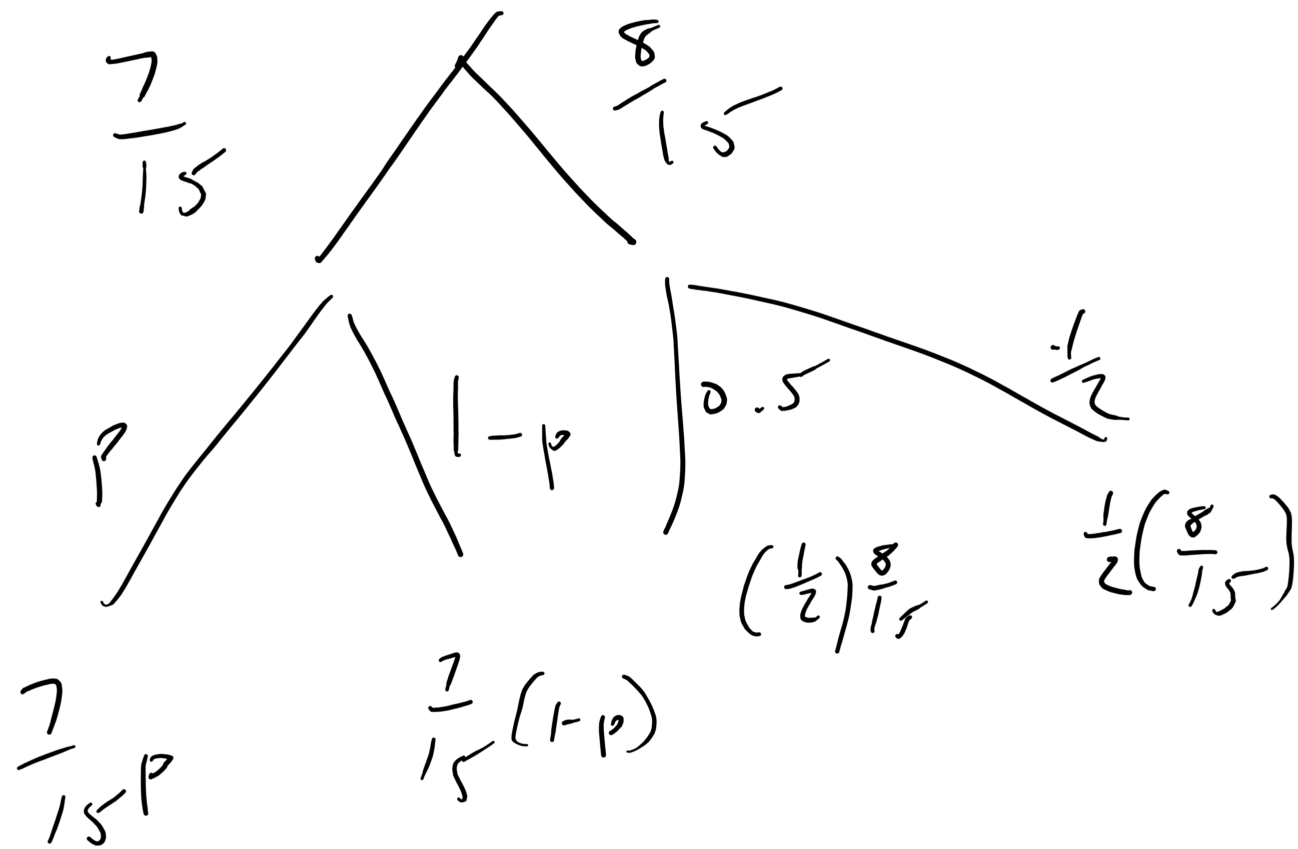
Example Solution: Differential Privacy

- A social scientist wants to know p, the fraction of Americans who are extraterrestrials. She tells people:
  - Toss a coin.
  - If it's heads, answer truthfully. If tails, use a second coin toss to decide what you'll tell me.
- Outcomes:
  - If an individual says "yes," chance is $0.5(1-q)/(0.5(1-q)+pq)$ that they're lying.
  - But we can estimate p with high accuracy: the fraction of people who say "yes" is exactly $0.5(1-q)+qp$.



Quiz

- Try the quiz: https://us.prairielearn.com/pl/course_instance/129874/assessment/2332096



$$\frac{9}{30} = \frac{7}{15}p + \left(\frac{1}{2}\right)\left(\frac{8}{15}\right)$$

Outline

- Why it matters
  - Privacy laws and lawsuits
  - On the other hand: the right to representation in machine learning
- Data security
  - The failure of k-anonymity
  - Data-centric information security
- Algorithmic methods
  - Federated learning
  - Differential privacy
- How to collect data so that you can share it legally, and why you should do so

Why you should collect as much shareable data as you can

1. It makes the world a better place
2. If you publish the first algorithm using a dataset, then:
    1. Everybody else who publishes will cite your paper (in order to say that their performance is better than yours)
    2. By the time they do, you will have even better results, because you started earlier

How to share data

- Method 1: ask your friends to let you record them
  - Pro: easy. Verbal consent is consent.
  - Con: no documentation. Nobody else can use your data, because they can't be sure that your friends gave consent.
- Method 2: ask your contributors to release their data under CC0
  - Pro: easy (https://creativecommons.org/choose/zero/). The data becomes free for anybody to use in any way they wish.
  - Con: not everybody is willing to do this
- Method 3: contributors sign a "consent form," data users sign a "data use agreement," and the terms of the two agreements match
  - Con: hard. You have to design the agreements; contributors have to read & sign.
  - Pro: allows very precise specification of what's allowed and not allowed