Differential Privacy

CS@CU

Data—The New Oil or the New Plutonium?

- Data can be very useful, both to its collectors and to others
  - Marketing
  - “Suggestion” algorithms
  - Researchers
  - Public issues, e.g., election districts, federal aid, etc.
- Data can be very dangerous if compromised
  - Intentional, controlled release
  - Hackers
  - Legal process

Protecting Released Data: First Attempts

- Remove PII
- No problem, right?

Early Privacy Techniques

- *k*-Anonymity
- Statistical queries only

k-Anonymity

- Ensure that any record in a dataset cannot be distinguished from at least *k*-1 others
- Similar to what HIPAA requires
- Often unclear how much privacy is guaranteed

HIPAA Privacy Safe Harbor

- Delete a list of 18 specific identifiers
  - Name, address, SSN, birth date (but not age unless over 90), etc.
- Delete other items known to be identifying
- Can include important demographic information such as gender and race
- Aggregate by state or by the first three digits of zip code if that's more than 20,000 people
- Note: HIPAA applies to health care providers, health care plans, and "business associates"—but not to Google, Microsoft, etc.
- *Note well: this is **very** oversimplified*

CS@CU

Statistical Queries

- Don’t release the dataset; allow only specific queries
- But—if powerful-enough queries are allowed, it is mathematically impossible to achieve good privacy

CS@CU

Raw Data

(Note: Randomly generated)

Last Name	First Name	Phone	Zip Code	Random ID	Gender	Birthdate	Salary	Race	Medical
Bryan	Frank	805-269-4479	60629	6973721572	M	Jun 12, 1958	$52,400	Black	Heart
Dery	Douglas	708-781-4211	79936	3389209159	M	Apr 28, 1985	$118,300	White	
Dube	Bessie	859-817-1388	90011	7574713594	F	May 20, 1938	$91,000	White	Heart
Haines	Fernando	414-614-0455	11385	2741335550	M	Apr 13, 1977	$115,600	-	Cancer
Jones	Naomi	816-202-7762	90650	3717441036	F	Sep 2, 1960	$136,800	White	Cancer
Neary	Hai	706-415-9488	77494	1561829881	NB	Apr 29, 1983	$141,300	White	
Razo	Jesabel	507-454-2166	91331	9037803106	F	Feb 18, 1951	$113,800	Hispanic	
Romano	Carlos	480-391-4486	90201	5132078469	M	Jun 22, 1988	$102,200	Hispanic	
Worley	Elizabeth	617-298-9122	11226	3819315445	F	May 27, 1952	$100,000	White	
Martinez	Mary	775-551-5327	10467	6730204579	F	Jun 13, 1978	$82,200	Hispanic	HIV

Sanitized Data

Zip Code	Random ID	Gender	Birthdate	Salary	Race	Medical
60629	6973721572	M	Jun 12, 1958	$52,400	Black	Heart
79936	3389209159	M	Apr 28, 1985	$118,300	White	
90011	7574713594	F	May 20, 1938	$91,000	White	Heart
11385	2741335550	M	Apr 13, 1977	$115,600	-	Cancer
90650	3717441036	F	Sep 2, 1960	$136,800	White	Cancer
77494	1561829881	NB	Apr 29, 1983	$141,300	White	
91331	9037803106	F	Feb 18, 1951	$113,800	Hispanic	
90201	5132078469	M	Jun 22, 1988	$102,200	Hispanic	
11226	3819315445	F	May 27, 1952	$100,000	White	
10467	6730204579	F	Jun 13, 1978	$82,200	Hispanic	HIV

Is This Data Sufficiently Protected?

- Medical information is quite sensitive
  - (N.Y. PBH §2782: No person who obtains confidential HIV related information in the course of providing any health or social service or pursuant to a release of confidential HIV related information may disclose or be compelled to disclose such information, except to the following…)
  - And there's HIPAA
- *Can the individuals be identified without the redacted PII?*
- Sometimes, yes…

CS@CU

Reidentification

- Latanya Sweeney: 87% of Americans are uniquely identified by birthdate, gender, and zip code
  - These fields are not considered PII!
  - Many of the remainder were in places like college towns, with many similar-age people
  - (Another study put the number at 63%—but that's still a lot)
- Now factor in race and likely income bracket

CS@CU

Reidentification Works

- Sweeney located the health records of the governor of Massachusetts
- The NY Times identified some individuals in anonymized AOL query data
- Narayanan and Shmatikov identified individuals in a released Netflix dataset

Several Strategies for Reidentification

- Outside data
  - Narayanan and Shmatikov use IMDb ratings
- Uniqueness of birthdate/gender/zipcode
- Uniqueness of query strings

Now what?

Differential Privacy

- Differential privacy provides a mathematical guarantee of a certain amount of privacy
- But—there is a tradeoff with accuracy
- The parameter ε specifies the tradeoff
- And: differential privacy is a *property*, not an algorithm (but there are such algorithms)

Defining Differential Privacy

- Assume that there are two datasets, D_1 and D_2, differing in at most one element
- Assume a “randomized function” $\mathcal{K}$ and a set $S \subseteq Range(\mathcal{K})$
- Then $\mathcal{K}$ is “ε-differentially private” if

$$\Pr[\mathcal{K}(D_1) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{K}(D_2) \in S]$$

CS@CU

Huh?

$$\Pr[\mathcal{K}(D_1) \in S] \le e^{\varepsilon} \cdot \Pr[\mathcal{K}((D_2) \in S]$$

- D_1 and D_2 are, for example, sets of medical history data
- $\mathcal{K}$ is a function that predicts medical outcomes. The “range” of $\mathcal{K}$ is the set of possible values, i.e., the possible outcomes
- S is a set of particular outcomes, e.g., cancer
- $\Pr[\mathcal{K}(D_i) \in S]$ is the probability of predicting some medical outcome, e.g., cancer
- e is the base of the natural logarithms, 2.71828…
- $\Pr[\mathcal{K}(D_1) \in S] \le e^{\varepsilon} \cdot \Pr[\mathcal{K}(D_2) \in S]$ shows that the change in the probability of predicting an outcome from a small change in the dataset is bounded by a factor of e^{ε}

Intuitive Translation

The privacy risk to an individual should not change significantly whether or not their data is included in the dataset.

The privacy parameter ε should be a small, positive number—the smaller it is, the more privacy you obtain, but the less accurate your results will be

The smaller ε is, the larger your dataset must be to produce useful statistical results

CS@CU

How Do We Achieve Differential Privacy?

- General technique: add “noise”
- That is: generate a random value from a (carefully chosen) distribution parameterized by ε
- Add this value to numeric fields in the dataset
- Note well: if you want privacy for c individuals, the effective privacy bound is

$$c \cdot \varepsilon$$

- In other words, you get good privacy protection for up to $c \approx 1/\varepsilon$ individuals, and very little for $c \approx 10/\varepsilon$ individuals

CS@CU

Why is Differential Privacy Good?

- It's a precise mathematical definition
- It gives an *exact* guarantee of privacy
- It's mathematically tractable, so we can prove theorems about privacy

Real-World Challenges

- Generating differentially private datasets can be expensive
  - (In our census experiments, we needed a fair number of AWS cores for several days, and we were only doing a few counties)
- Differential privacy doesn't work well for "high-dimensionality" datasets—ones with many columns
  - Some things, e.g., text, are hard to fit into the DP model
- There are real-world constraints on some values—you can't have a non-integral number of family members; children can't be older than their (biological) parents, etc.
- Is the accuracy good enough? Recall that lower values of ε imply more privacy but less accuracy.

CS@CU

Who Uses Differential Privacy?

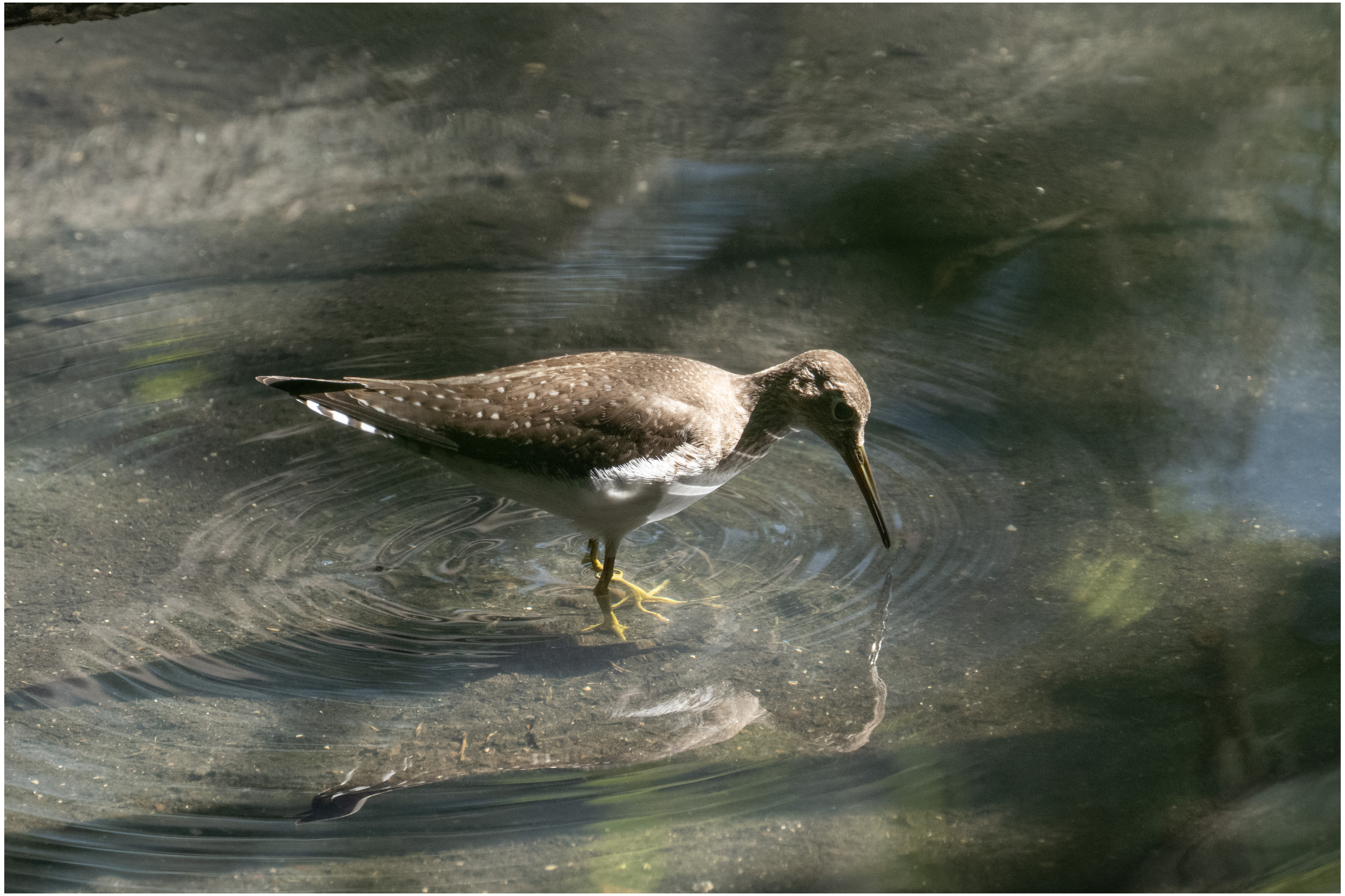
- Apple, to protect collected user data
- Google
- Microsoft, for telemetry in Windows
- Facebook
- Wikimedia
- The Census Bureau, for the 2020 census
- (Not used for HIPAA—legal liability issue…)

The Census Bureau

- By law, individuals' data must be kept confidential
- 13 U.S.C. §9(a)(2): "Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof… may… make any publication whereby the data furnished by any particular establishment or individual under this title can be identified"
- Researchers there found that their older privacy mechanism, swapping, was subject to attack
- But—census data *must* be accurate (enough)
- The *American Community Survey* is high-dimensional—and the Census Bureau has given up on using differential privacy for it

CS@CU

Daily Bird



Solitary sandpiper, Central Park, September 20, 2023