Dr. Roberto Yus
https://robertoyus.com/

Data Privacy

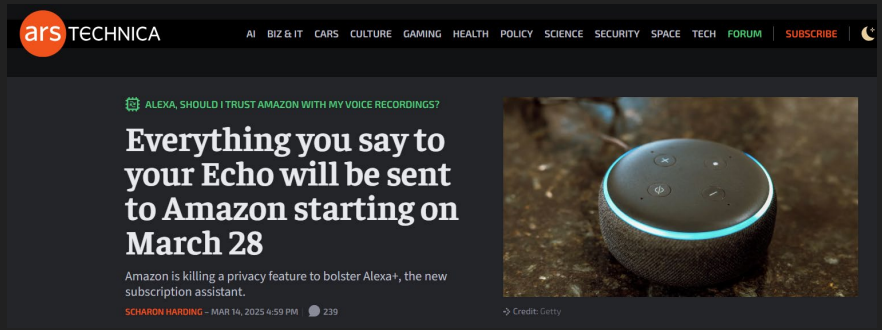
CMSC 463/663

L07 – Differential Privacy

Icons from https://thenounproject.com/

Previously on…

- Need to share data
- Classification of attributes (key, quasi-identifiers, sensitive)
- k-Anonymity and de-anonymization attacks
- l-diversity, t-closeness…
  - Still de-anonymization attacks are possible, and data becomes useless



In the news!

The Problem of Background Knowledge

Race	Zip	HIV status	Condition
Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

Bob is Caucasian and I've heard he was admitted to a hospital with flu…

Imagine a table which is:

- k-anonymous,
- l-diverse,
- and t-close table

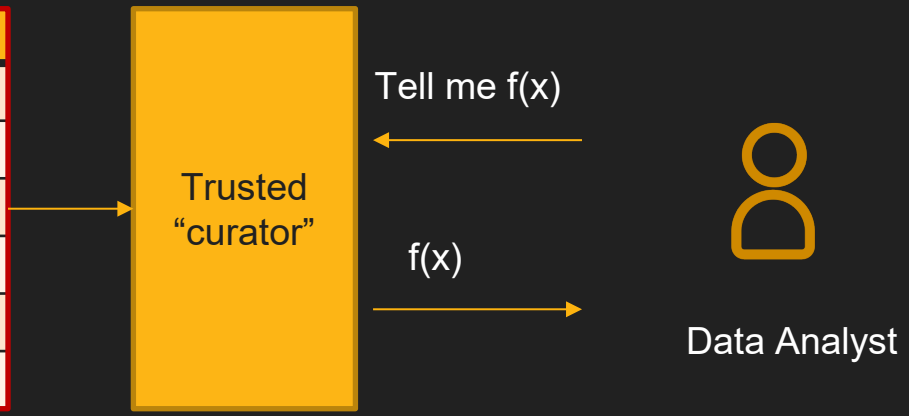
Perfect privacy?

This goes against the rules! "flu" is not a quasi-identifier

Yes… and this is yet another problem with k-anonymity

Mediate Access & Statistical Releases

Name	Race	HIV status	Condition
Bob	Caucas	HIV+	Flu
Mary	Asian/AfrAm	HIV-	Flu
John	Asian/AfrAm	HIV+	Shingles
Tony	Caucas	HIV-	Acne
Anna	Caucas	HIV-	Shingles
Lisa	Caucas	HIV-	Acne



- f(x) → some operation
  - E.g., “What fraction of people are Caucasian and HIV positive?”

Reconstruction Attack

- Reconstruct records using statistical data
- Example: US Census 2010 reconstruction attack

Age	Race

1. There are four people in total
2. Two of these people have age 17
3. Two of these people self-identify as White
4. Two of these people self-identify as Asian
5. The average age of people who self-identify as White is 30
6. The average age of people who self-identify as Asian is 32

Can you reconstruct the table?

https://desfontain.es/privacy/us-census-reconstruction-attack.html

Reconstruction Attack

- Reconstruct records using statistical data
- Example: US Census 2010 reconstruction attack

Age	Race
17	White
17	Asian
43	White
47	Asian

1. There are four people in total
2. Two of these people have age 17
3. Two of these people self-identify as White
4. Two of these people self-identify as Asian
5. The average age of people who self-identify as White is 30
6. The average age of people who self-identify as Asian is 32

Can you reconstruct the table?

https://desfontain.es/privacy/us-census-reconstruction-attack.html

Reconstruction Attack

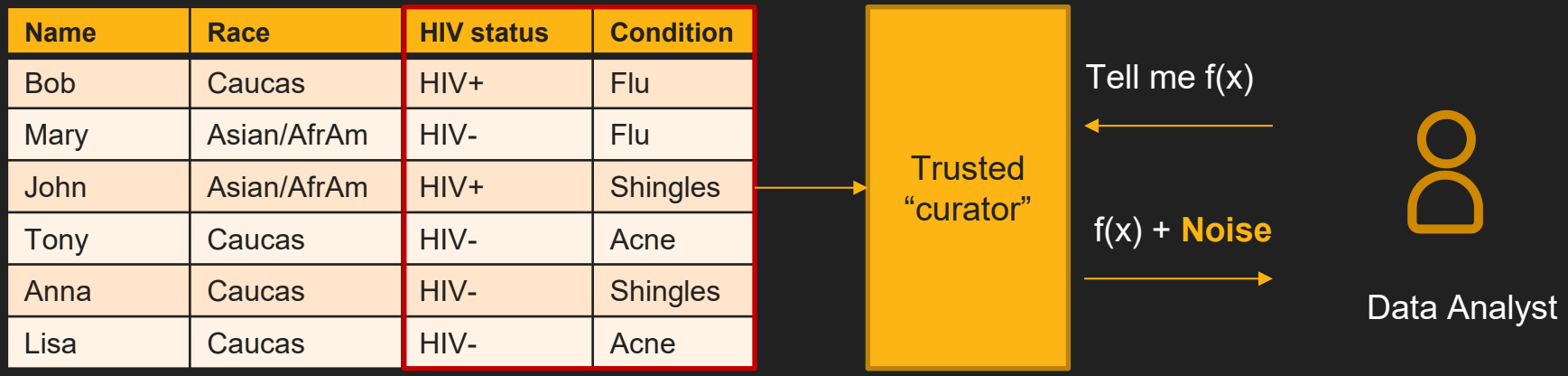
- Example: US Census 2010 reconstruction attack

Age	Race
17	White
17	Asian
43	White
47	Asian

- Team at the Census Bureau reconstructed 46% of all the records using a “small” fraction of statistics
- Re-identification after reconstruction!
  - De-anonymization attack
  - Scary!

https://desfontain.es/privacy/us-census-reconstruction-attack.html

Perturb Output?



- Add noise to the output to prevent reconstruction?

Dinur-Nissim Attack

- Even if we perturb the output of statistical queries, we can still reconstruct the whole table

- Dinur-Nissim Attack (heavily paraphrased):
  - Given a database with *n* rows, if roughly *n* queries are made to the database, then essentially the entire database can be reconstructed *even if* $O(n^{\frac{1}{2}})$ *noise is added to each answer*

Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In 22nd ACM SIGMOD-SIGACT-SIGART (PODS '03).

Formally Defining Privacy

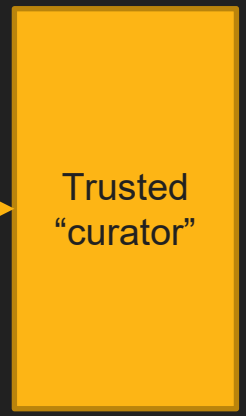
- A problem inherent in all the approaches we have discussed so far (and the source of many of the problems we have seen) is that no definition of “privacy” is offered

- **Differential Privacy is a formal definition of privacy**
  - *“The outcome of any analysis is essentially equally likely, independent of whether any individual joins, or refrains from joining, the dataset”*
  - Based on Dinur-Nissim result that adding *some* (carefully-generated) noise, and limiting the number of queries, *can* be proven to achieve privacy

Differential Privacy

Carefully chosen!
E.g., Magnitude of noise depends on range of plausible sensitive values

Race	HIV status	Condition
Caucas	HIV+	Flu
Asian/AfrAm	HIV-	Flu
Asian/AfrAm	HIV+	Shingles
Caucas	HIV-	Acne
Caucas	HIV-	Shingles
Caucas	HIV-	Acne



Tell me f(x)

f(x) + Noise

Data Analyst

Requirement: Effect of each individual should be “hidden”

- *“A record’s presence or absence from input of an analysis is not revealed by its result”*

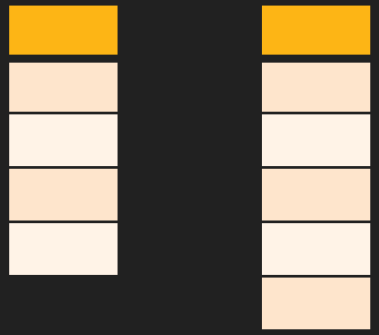
Dwork, Cynthia. "Differential privacy: A survey of results." International conference on theory and applications of models of computation, 2008.

Differential Privacy can’t “make” privacy

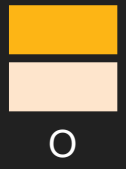
- Imagine that a DP analysis teaches us that smokers are at risk for cancer, and also you smoke in public
- DP has not violated your privacy. All conclusions about you could be reached without your secrets
- DP masks the nature of one’s participation in surveys and prevents the mishandling of individuals’ records
- It does not manufacture privacy where none exists

Differential Privacy

For every pair of inputs that differ in one row



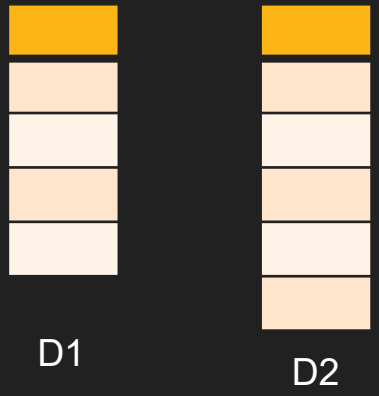
For every output…



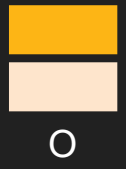
Adversary should not be able to distinguish between any D1 and D2 based on any O

Why pairs of datasets *that differ in one row?*

For every pair of inputs that differ in one row



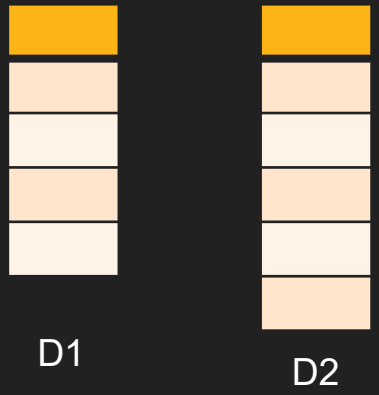
For every output…



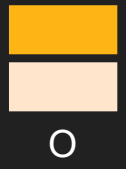
Simulate the presence or absence of a single record

Why *all* pairs of datasets?

For every pair of inputs that differ in one row



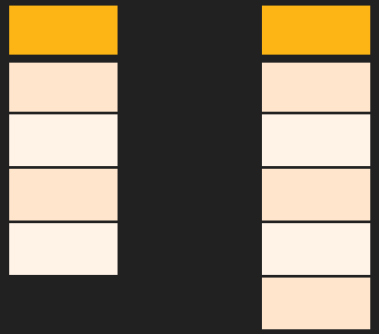
For every output…



Guarantee holds no matter what the other records are

What does it mean *not to be able to distinguish*…?

For every pair of inputs that differ in one row



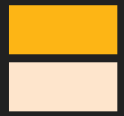
D1

D2

$$\ln\left(\frac{\Pr[A(D1)=O]}{\Pr[A(D2)=O]}\right) \leq \varepsilon\,,\ \varepsilon > 0$$

A → ε-differentially private algorithm

For every output…



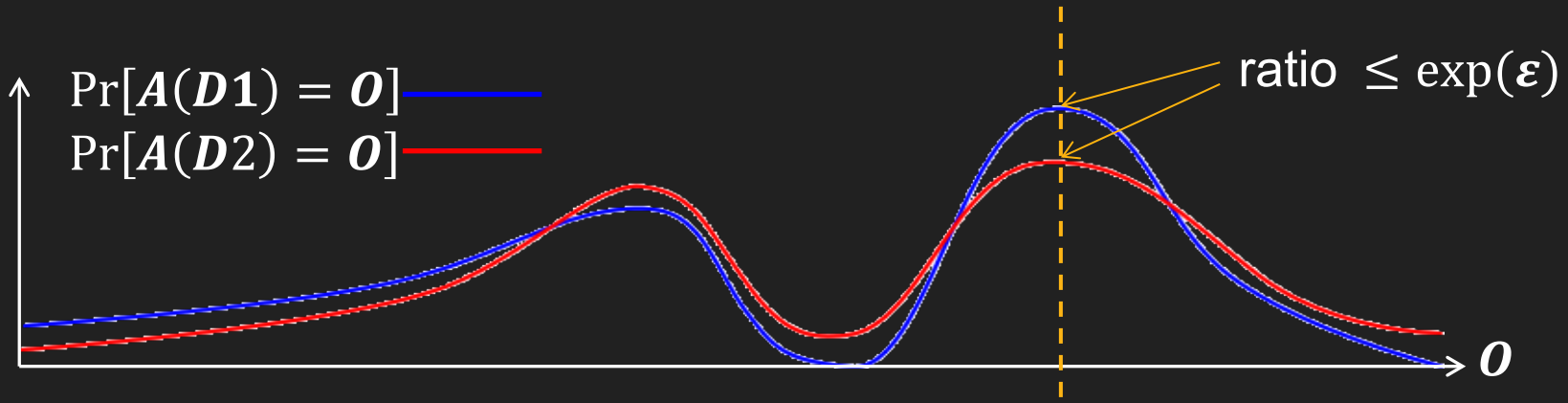
O

Privacy parameter ε controls the degree to which D1 and D2 can be distinguished

Smaller the ε more privacy
…and worse utility

$$\ln\left(\frac{\Pr[A(D1)=O]}{\Pr[A(D2)=O]}\right) \leq \varepsilon\,,\ \varepsilon > 0$$

$$\Pr[A(D1) = O] \leq \exp(\varepsilon) * \Pr[A(D2) = O]\,,\ \varepsilon > 0$$



Useful Properties of Differential Privacy

- Postprocessing
- Composability
- Group privacy
- …

Post-processing

- If A(D) is ε-private, and f is any (randomized) function,

  then f(A(D)) is ε-private.

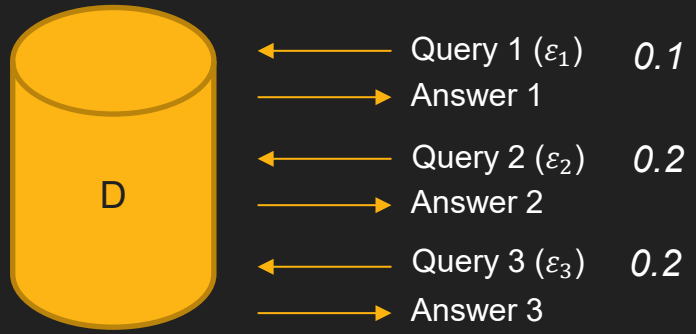
- In other words, differential privacy is robust against further process of a previous database output
  - Future-proof → Current and future side information

Composability

- Composability is the ability to join the output of two (or more) differentially privacy mechanisms
- Why?
  - Reasoning about privacy of a complex algorithm is hard
  - Helps software design process
  - If building blocks are proven to be private, it would be easy to reason about privacy of a complex algorithm built entirely using these building blocks

Composability

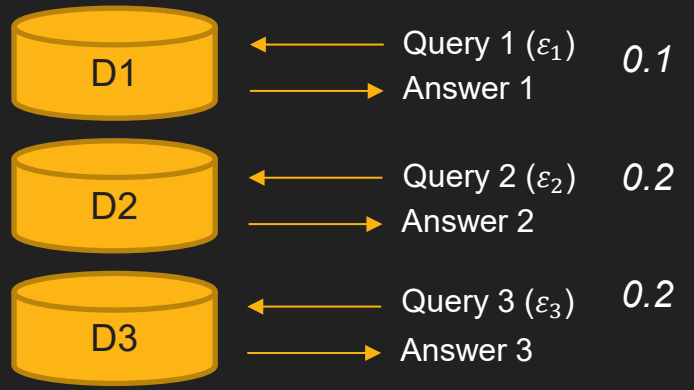
- Arbitrary composition (sequential and/or parallel) of *k* differentially private algorithms is still differentially private



Sequential composition

$\sum_i \varepsilon_i$– differential privacy

0.5



Parallel composition

$\max(\varepsilon_i)$– differential privacy

0.2

How to Achieve Differential Privacy?

- Basic algorithms:
  - Randomized response
  - **Laplace mechanism**
  - Exponential mechanism
- Advanced algorithms:
  - histograms [DMNS06]
  - contingency tables [BCDKMT07, GHRU11, TUV12, DNT14],
  - machine learning [BDMN05,KLNRS08],
  - regression & statistical estimation [CMS11,S11,KST11,ST12,JT13]
  - clustering [BDMN05,NRS07]
  - social network analysis [HLMJ09,GRU11,KRSY11,KNRS13,BBDS13]
  - approximation algorithms [GLMRT10]
  - singular value decomposition [HR12, HR13, KT13, DTTZ14]
  - streaming algorithms [DNRY10,DNPR10,MMNW11]
  - mechanism design [MT07,NST10,X11,NOS12,CCKMV12,HK12,KPRU12]
  - …

Sensitivity

- Measure how much the answer of a function can change when we change one of the input rows

$$f: D \rightarrow R^d$$

(Sensitivity)

$$\Delta f = max \| f(D1) - f(D2) \|_1$$

Name	HIV+
John	1
Mary	0
Anna	1
Tom	0

D1

Name	HIV+
John	1
Mary	0
Anna	1
Tom	**1**

D2

(Average)

$$f(x) = \frac{1}{n} \sum_{i=0}^{n} x_i \qquad \Delta = \frac{1}{n}$$

The average can change at most by 1/n if we change one single record!

Sensitivity

- Measure how much the answer of a function can change when we change one of the input rows

$$f: D \rightarrow R^d$$

(Sensitivity)

$$\Delta f = max \| f(D1) - f(D2) \|_1$$

Name	HIV+
John	1
Mary	0
Anna	1
Tom	0

D1

Name	HIV+
John	1
Mary	0
Anna	1
Tom	**1**

D2

(Average)

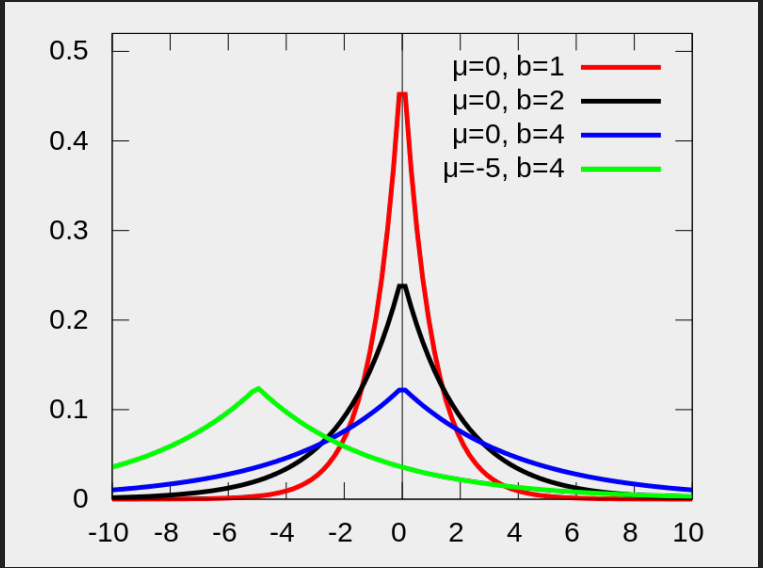
$$f(x) = \frac{1}{n} \sum_{i=0}^{n} x_i \qquad \Delta = \frac{1}{n}$$

How much is the sensitivity of the count?

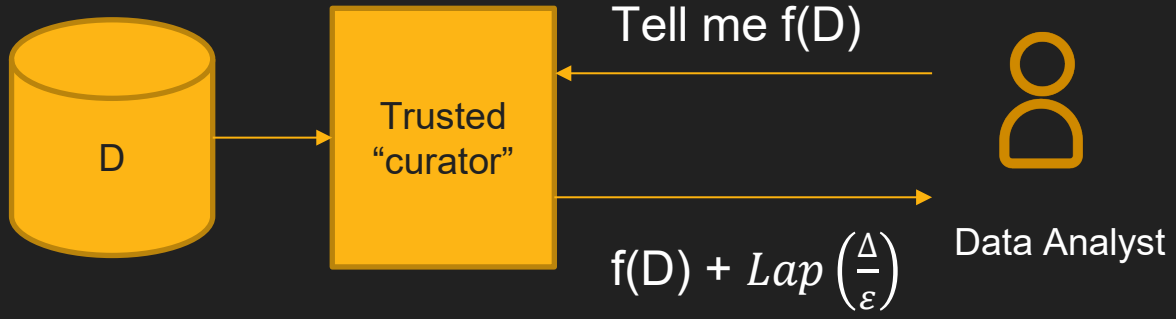
Laplace Distribution

$$pdf(x) = \frac{1}{2b}\exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{b}}\right)$$

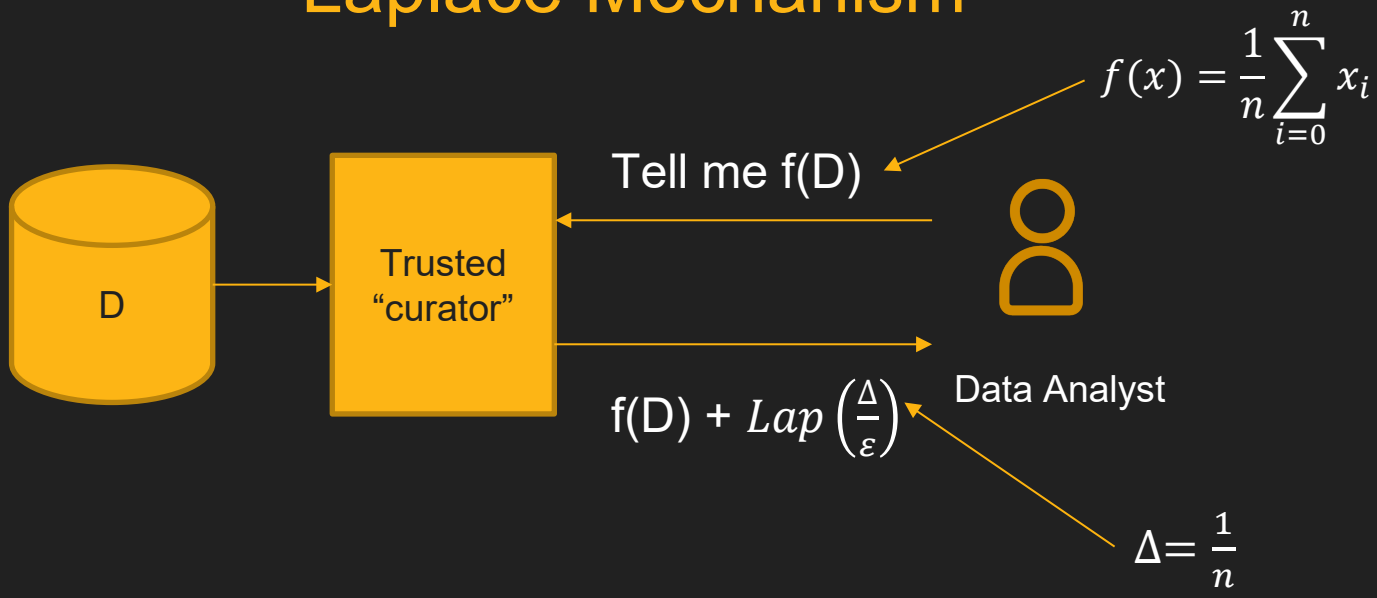
variance: $2b^2$, b is referred as the scale



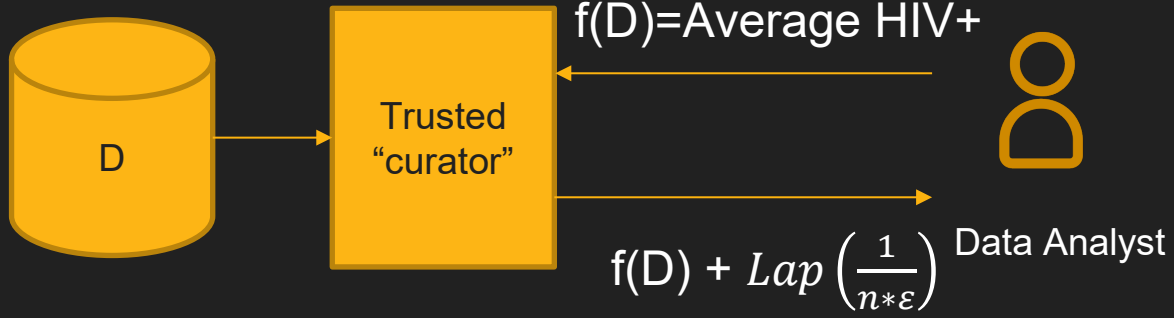
Laplace Mechanism



Laplace Mechanism



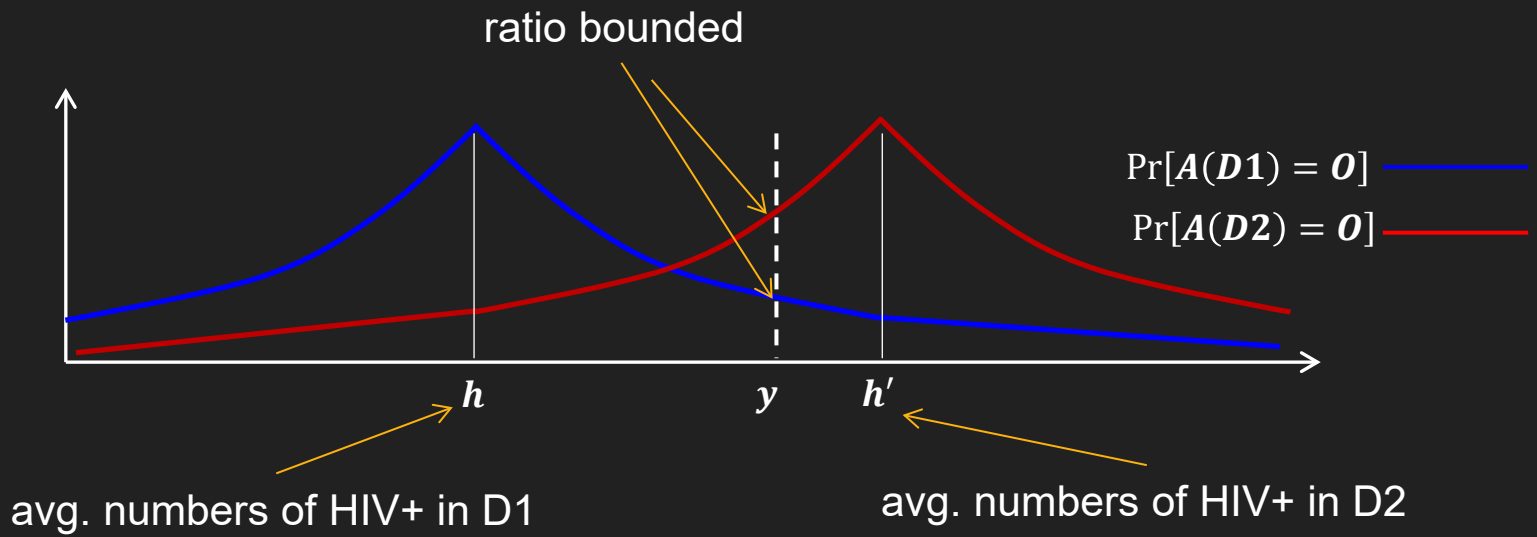
Laplace Mechanism



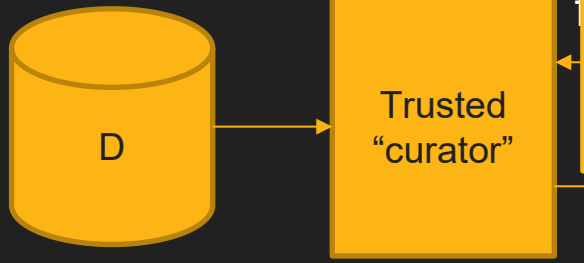
$$\mathrm{M}(D) = f(D) + Lap\left(\frac{1}{n * \varepsilon}\right)$$

Is ε-differentially private!

Laplace Mechanism

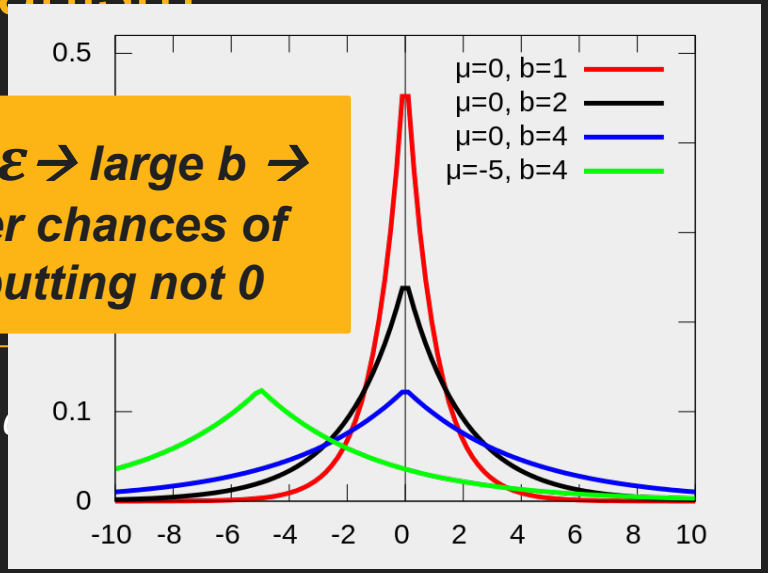


Laplace Mechanism



Small ε → large b → higher chances of outputting not 0

f(D) + L



$$\mathrm{M}(D) = f(D) + Lap\left(\frac{1}{n * \varepsilon}\right)$$

Is ε-differentially private!

Small ε → more noise → more privacy → less utility!

How to choose ε?

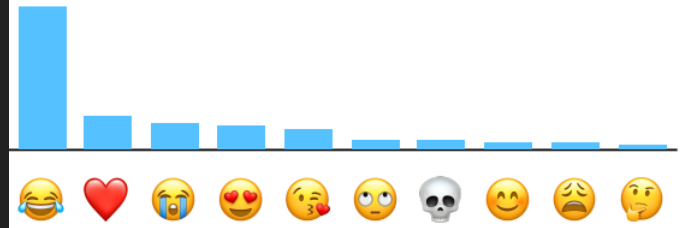
Hard problem!

- Perspective taken by theory: Pick ε, prove (or evaluate) accuracy
- Realities of practice: Hard accuracy requirements
  - Find the smallest level of ε consistent with accuracy targets.
  - How to do this?
    - Search incurs privacy overhead…
    - Privacy parameter is now a data-dependent quantity. What is the semantics?
    - Constant factors can be meaningful…

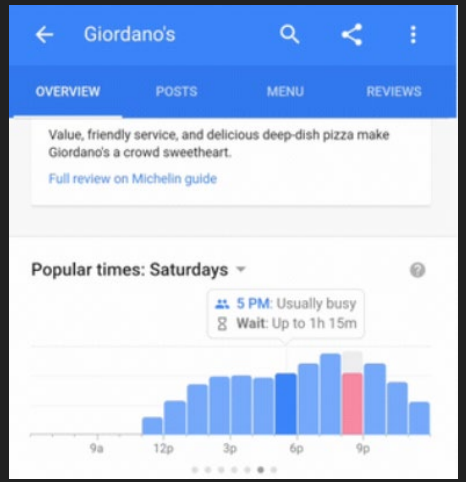
Lee, Jaewoo, and Chris Clifton. "How much is enough? choosing ε for differential privacy." International Conference on Information Security, 2011.

Use Cases

There are situations where Apple can improve the user experience by getting insight from what many of our users are doing, for example: What new words are trending and might make the most relevant suggestions? What websites have problems that could affect battery life? Which emoji are chosen most often? The challenge is that the data which could drive the answers to those questions—such as what the users type on their keyboards—is personal.



The Count Mean Sketch technique allows Apple to determine the most popular emoji to help design better ways to find and use our favorite emoji. The top emoji for US English speakers contained some surprising favorites.



Impact of Differential Privacy on Congressional Districts (2010)

	Total population				White non-Hispanic				Black/African-American non-Hispanic			
Congressional District	Summary File (2010)	DP (2010)	Numeric Difference	Percent Difference	Summary File (2010)	DP (2010)	Numeric Difference	Percent Difference	Summary File (2010)	DP (2010)	Numeric Difference	Percent Difference
1	644,787	644,782	-5	0.0%	573,596	573,468	-128	0.0%	13,642	13,607	-35	-0.3%
2	732,515	732,687	172	0.0%	626,655	626,757	102	0.0%	23,650	23,704	54	0.2%
3	650,185	650,212	27	0.0%	512,639	512,584	-55	0.0%	50,236	50,308	72	0.1%
4	614,624	614,539	-85	0.0%	424,833	424,717	-116	0.0%	59,514	59,563	49	0.1%
5	616,482	616,431	-51	0.0%	402,523	402,449	-74	0.0%	93,434	93,482	48	0.1%
6	759,478	759,432	-46	0.0%	685,794	685,845	51	0.0%	18,221	18,229	8	0.0%
7	625,512	625,486	-26	0.0%	565,870	565,682	-188	0.0%	4,701	4,600	-101	-2.1%
8	660,342	660,356	14	0.0%	613,232	613,636	404	0.1%	5,743	5,647	-96	-1.7%
Minnesota (all)	**5,303,925**	**5,303,925**	**0**	**0.0%**	**4,405,142**	**4,405,138**	**-4**	**0.0%**	**269,141**	**269,140**	**-1**	**0.0%**

	Hispanic/Latino				Asian				American Indian			
Congressional District	Summary File (2010)	DP (2010)	Numeric Difference	Percent Difference	Summary File (2010)	DP (2010)	Numeric Difference	Percent Difference	Summary File (2010)	DP (2010)	Numeric Difference	Percent Difference
1	33,517	33,756	239	0.7%	14,325	14,297	-28	-0.2%	1,438	1,387	-51	-3.5%
2	34,803	34,862	59	0.2%	29,412	29,411	-1	0.0%	3,155	3,180	25	0.8%
3	25,801	25,915	114	0.4%	43,855	43,862	7	0.0%	2,043	2,034	-9	-0.4%
4	46,505	46,454	-51	-0.1%	62,836	62,911	75	0.1%	3,594	3,559	-35	-1.0%
5	58,639	58,583	-56	-0.1%	32,477	32,538	61	0.2%	7,766	7,731	-35	-0.5%
6	18,361	18,297	-64	-0.3%	21,542	21,532	-10	0.0%	2,988	3,049	61	2.0%
7	24,063	24,130	67	0.3%	4,761	4,725	-36	-0.8%	17,064	17,086	22	0.1%
8	8,569	8,278	-291	-3.4%	3,788	3,748	-40	-1.1%	17,373	17,394	21	0.1%
Minnesota (all)	**250,258**	**250,275**	**17**	**0.0%**	**212,996**	**213,024**	**28**	**0.0%**	**55,421**	**55,420**	**-1**	**0.0%**

Source: NHGIS Privacy Protected Microdata File, University of Minnesota, from U.S. Census Bureau data

Impact of Differential Privacy on House Legislative Districts (2010)

	Total population		White non-Hispanic		Black/African-American non-Hispanic		Hispanic/Latino		Asian non-Hispanic		American Indian non-Hispanic	
	# Dif.	% Dif.	# Dif.	% Dif.	# Dif.	% Dif.	# Dif.	% Dif.	# Dif.	% Dif.	# Dif.	% Dif.
Largest Positive Difference	120	0.3%	96	0.3%	52	12.2%	94	6.1%	92	15.3%	34	20.3%
Largest Negative Difference	-93	-0.3%	-124	-0.5%	-55	-23.3%	-106	-15.7%	-53	-30.6%	-27	-27.6%
MAPE		0.0%		0.0%		-1.2%		-0.4%		-1.3%		-0.4%

Source: NHGIS Privacy Protected Microdata File, University of Minnesota, from U.S. Census Bureau data

Conclusions

- Current state of the art for privacy protection
- DP mechanisms use parameters like ε to adjust the tradeoff between the level of privacy loss and data quality.
- Works well when you have a lot of data
- Works well to learn about the average population but not about outliers
- Offers strong mathematical guarantees about privacy, not so much about utility
- Tends to be less effective when there exist correlations among the tuples

Group Activity

- Think about your group project
- What statistical data would you want to release?
  - How much data?
  - What operations?
  - Would differential privacy help?