

Privacy Module Week 1:

Privacy Attacks and Differential Privacy

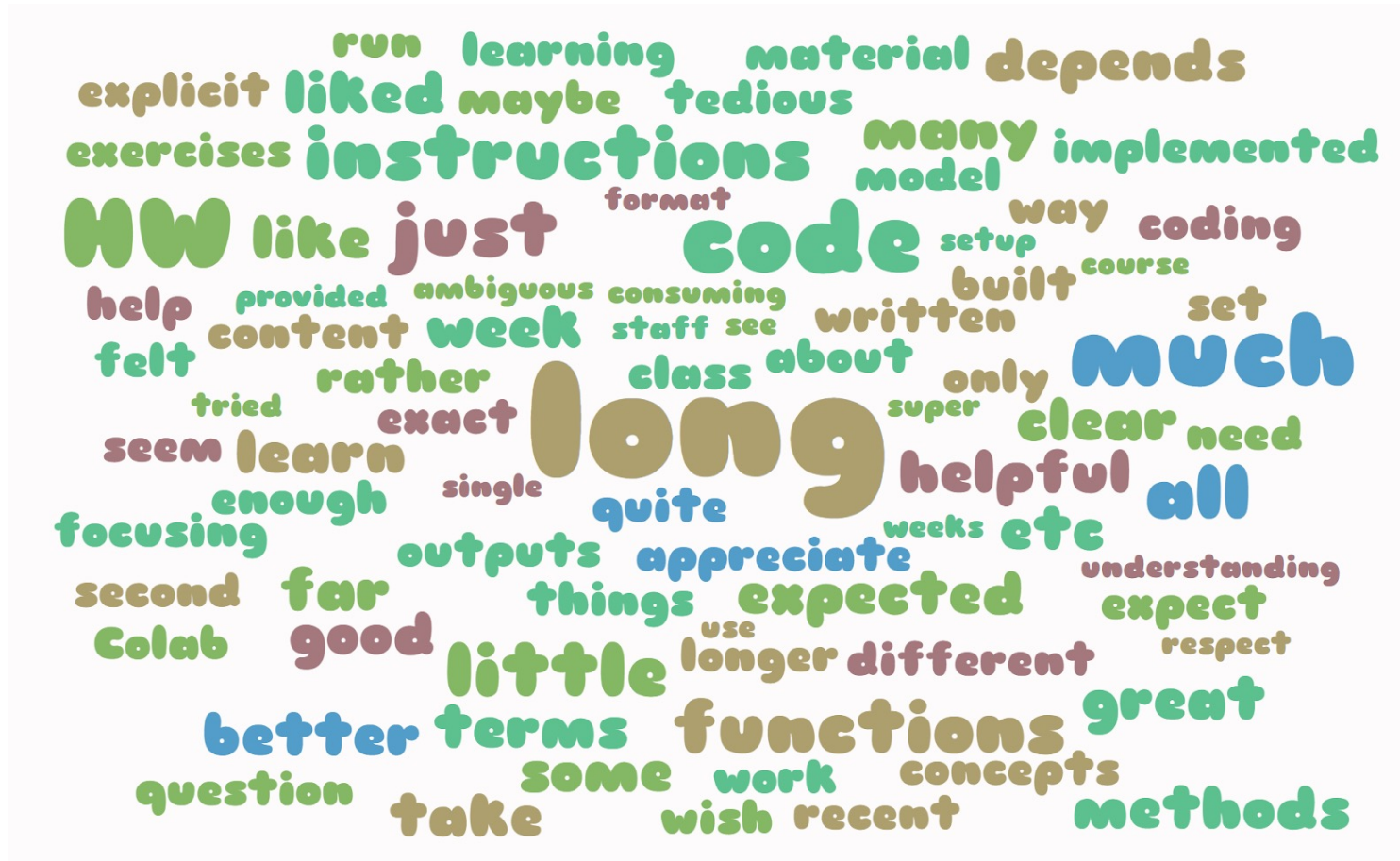
CS 329T

Stanford Spring 2021

Course survey results

- The course material is popular
- There are some suggestions for improvements
 - Homework length
 - Overlaps with video, fireside chat, lab
 - Timing of video, lab, homework
- Homework 3
clarification: <https://piazza.com/class/kmp7hadai9l73y?cid=137>

Word cloud for homework



We will shorten the remaining homework

Word cloud for lab



We will refine labs to help on current homework

Lab feedback

“The lab sections have gotten better; initially I was a little bored, but now that they're most hands-on, I enjoy them.”

“The lab sections are also a bit redundant with the prerecorded lectures and fireside chats. It would be helpful if these were more focused on homework related prep”

“It is a bit difficult to work through a Colab with little direction - it is usually hard to know where to start and we normally don't get too far. But I enjoy working on them and think they are instructive”

Some problems with weekly schedule

- Video release over the weekend; fireside chat on Tuesday



We will work to release video earlier

- Homework due Saturday (this week); lab Wed-Fri
 - Too late to provide help with homework due this week
 - Too early to discuss privacy homework that will be posted this week



We will use this week's lab to catch up; cover privacy homework next week

Plan for Today: Three takeaways

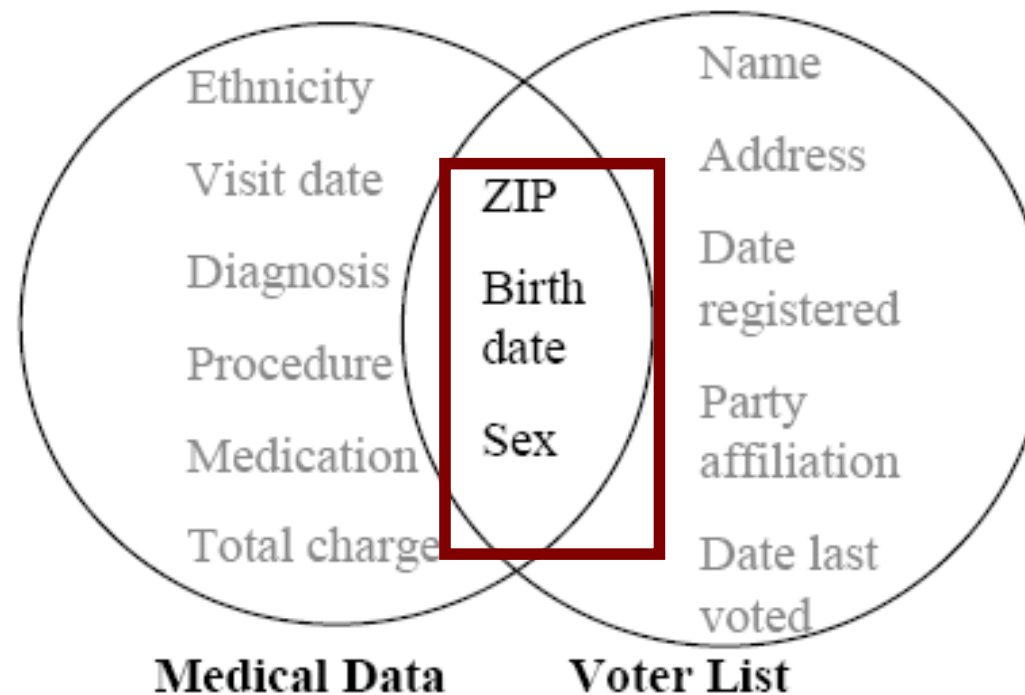
1. Privacy as disclosure limitation (Dalenius) is impractical/impossible to achieve
2. Differential privacy focuses on limiting incremental disclosure from participating in a data collection process
3. Privacy vs. Utility/Accuracy tradeoffs

Takeaway 1

- Privacy as **disclosure limitation** (Dalenius) is impractical/impossible to achieve
- Important attacks link "anonymized" data with side information
 - Latanya Sweeney and Massachusetts medical records
 - Netflix de-anonymization attack
 - Model inversion attack
- There is a general impossibility theorem
 - Theorem assumes adversary has particular background information
 - Practically, it is hard to constrain background knowledge

Latanya Sweeney medical record linking

Massachusetts Group Insurance Commission released anonymized records in 1990s
Latanya Sweeney identified the medical record of Gov Weld of Massachusetts



87 % of US population uniquely identifiable by 5-digit ZIP, gender, DOB

Netflix-IMDb Empirical Attack [Narayanan et al 2008]

Anonymized Netflix DB

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Publicly available IMDb ratings
(noisy)

		Titanic	Heidi
	Bob	2	1

Used as auxiliary information



Weighted Scoring Algorithm

How do you
measure similarity
of this record with
Bob's record?
(Similarity Metric)

What does **auxiliary
information** about a
record mean?

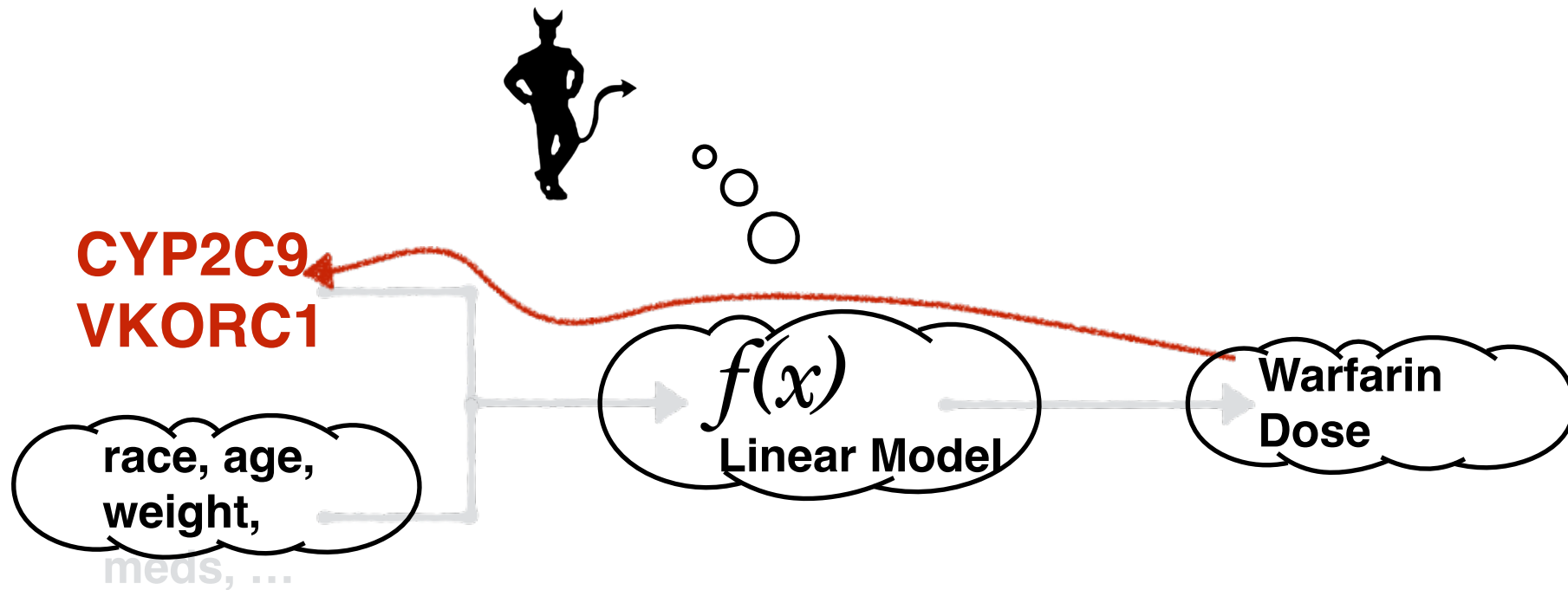


	r_1	4	1	0
---	-------	---	---	---

Model Inversion

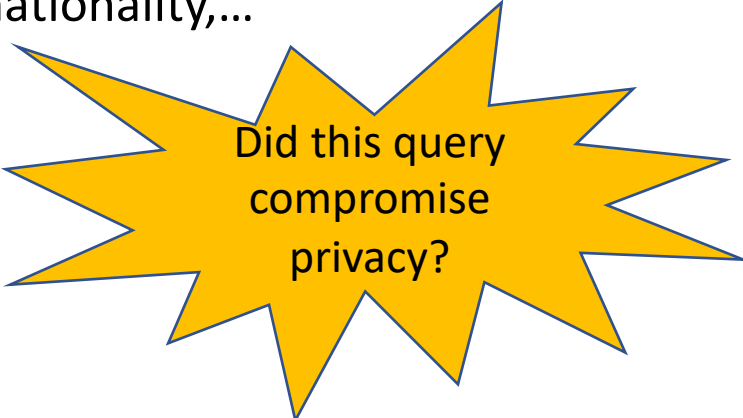
Disclosure in Pharmacogenetics

age	height	weight	race	history	vkorc1	cyp2c9	dose
50-60	176.2	185.7	asian	cancer	A/G	*1/*3	42.0



Impossibility Result [Dwork, Naor 2006]

- Suppose
 - San is a sanitization answering some class of queries Q over a class \mathbf{D} of databases, and
 - P is a class of private queries that are prohibited on sanitized databases
- Then under very general conditions, for any private query p
 - There is a query q and side information $Side(D,p,q)$ for database D such that
 - p cannot be inferred from either $San(D,q)$ or $Side(D,p,q)$ but can be from their combination
- Example
 - Consider databases \mathbf{D} with individual height, weight, eye color, nationality,...
 - Sanitized queries: average characteristic by nationality
 - Private query: find the height of an individual, e.g., Terry Gross
 - Side information that breaks privacy:
 - Terry Gross is two inches shorter than the average Lithuanian woman

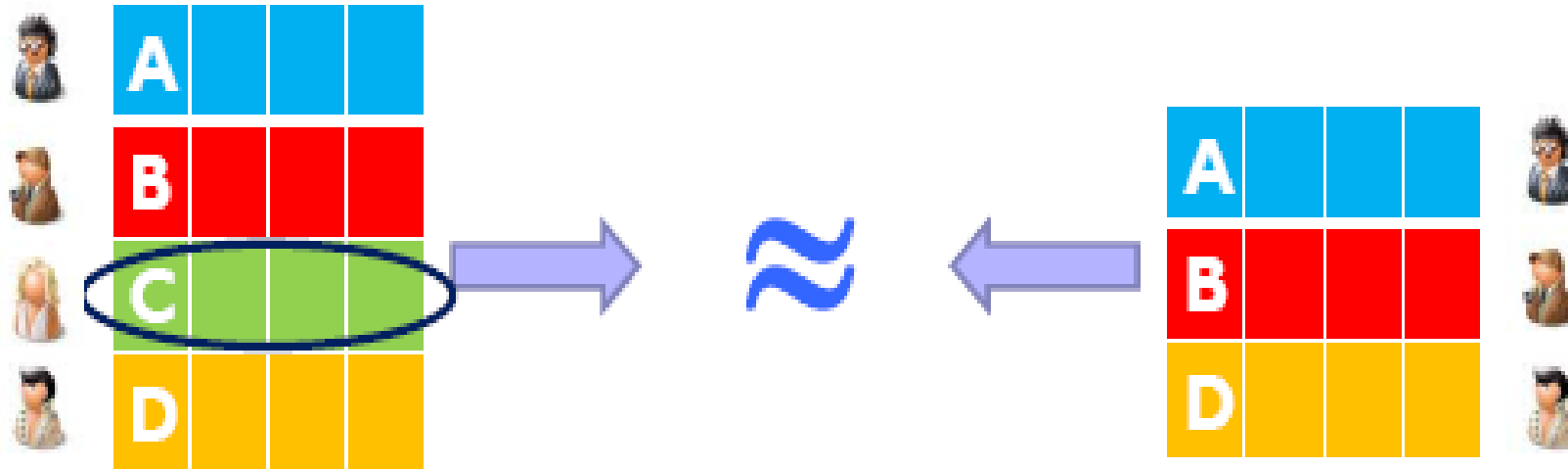


Did this query
compromise
privacy?

Takeaway 2

- Differential privacy focuses on limiting incremental **disclosure** from participating in a data collection process
 - Terry Gross is no worse off from a privacy standpoint whether her height is in the dataset or not (i.e., whether she participates in the survey)
 - Key Concept: Sensitivity of a function to individual inputs
 - Key Concept: Randomized mechanism that adds noise calibrated to sensitivity

Differential Privacy: Idea



Released statistic is about the same
if any individual's record is
removed from the database

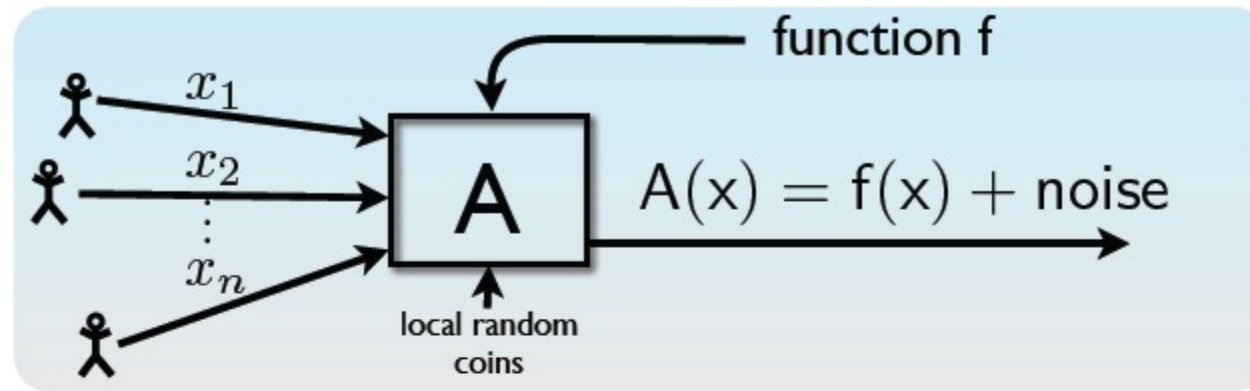
Differential Privacy: Definition

Randomized sanitization function κ has ϵ -differential privacy if for all data sets D_1 and D_2 differing by at most one element and all subsets S of the range of κ ,

$$\Pr[\kappa(D_1) \in S] \leq e^\epsilon \Pr[\kappa(D_2) \in S]$$

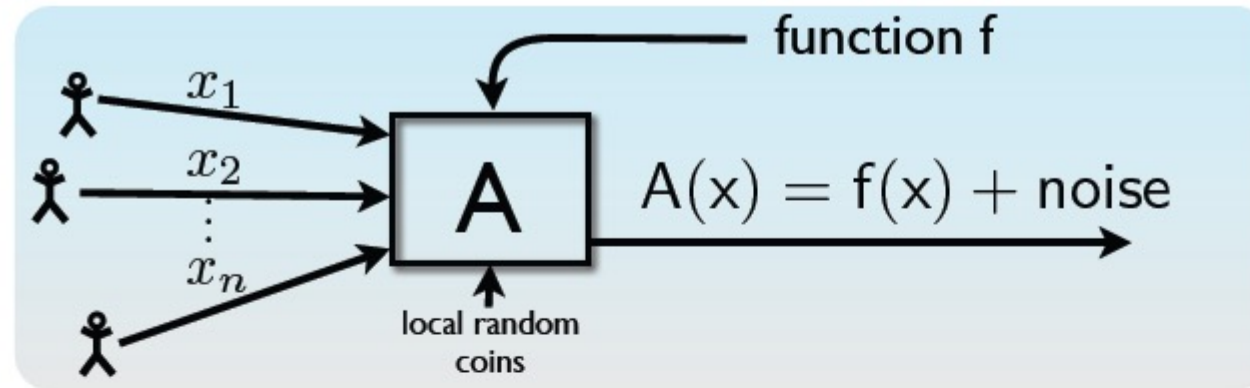
Answer to query # individuals with salary > \$30K is in range [100, 110] with approximately the same probability in D_1 and D_2

Example: Noise Addition



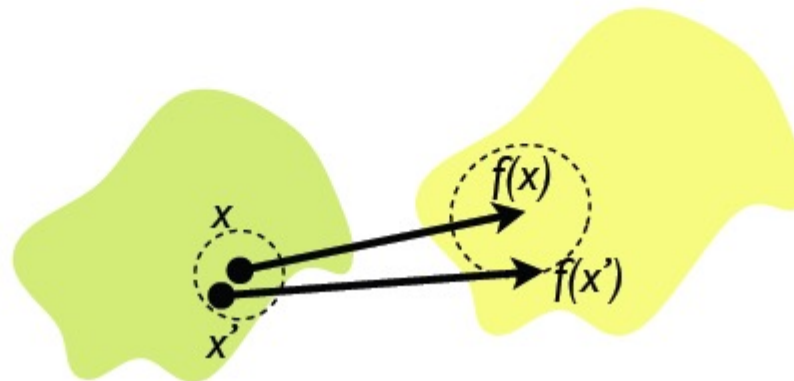
- Say we want to release a summary $f(x) \in \mathbb{R}^p$
 - e.g., proportion of diabetics: $x_i \in \{0, 1\}$, $f(x) = \frac{1}{n} \sum x_i$
- Simple approach: add noise to $f(x)$
 - How much noise is needed?
- **Intuition:** $f(x)$ can be released accurately when f is insensitive to individual entries x_1, x_2, \dots, x_n

Global Sensitivity



- **Global Sensitivity:** $GS_f = \max_{\text{neighbors } x, x'} \|f(x) - f(x')\|_1$

➤ Example: $GS_{\text{proportion}} = \frac{1}{n}$



Laplacian Mechanism

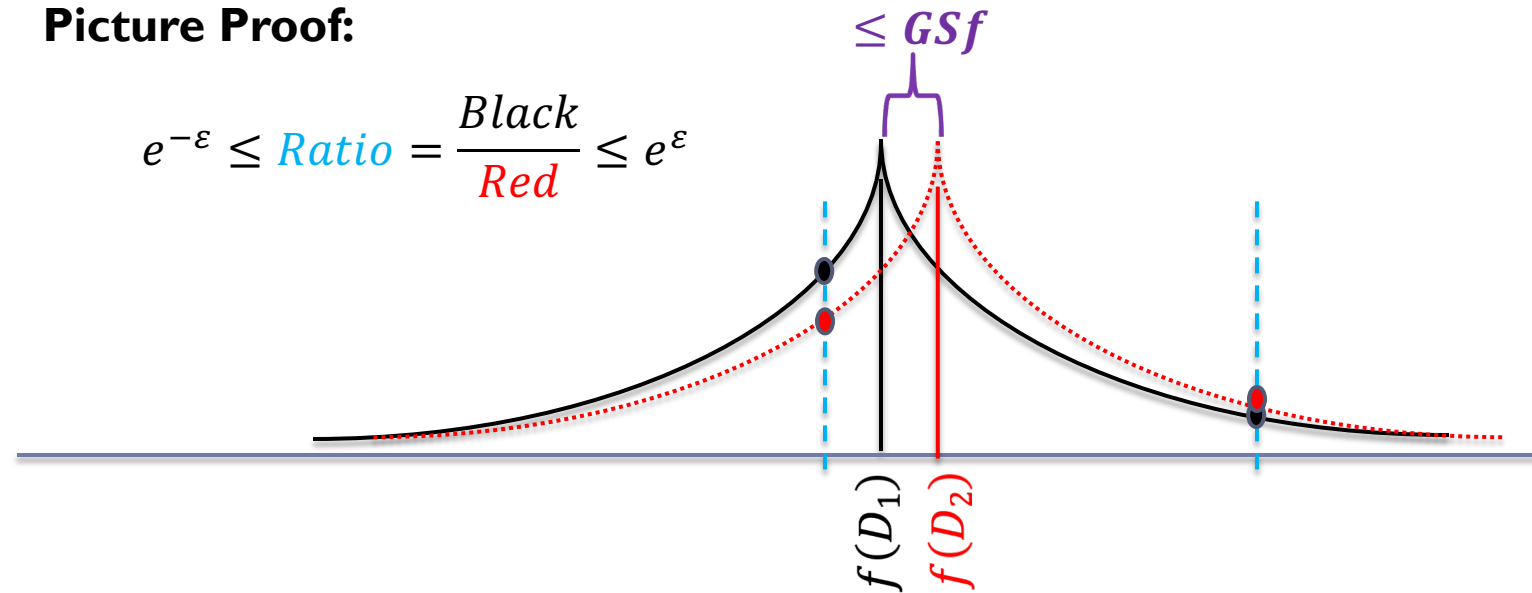
$$K(D) = f(D) + \text{Lap}\left(\frac{GS_f}{\epsilon}\right)$$

Thm: K is ϵ -differentially private

Probability Density Function

$$\text{Lap}(x, 0, \sigma) \propto \frac{1}{2\sigma} \exp\left(\frac{-|x|}{\sigma}\right)$$

Picture Proof:



Takeaway 3: Privacy vs. Utility

- Privacy vs. Utility/Accuracy tradeoffs
 - K-anonymity
 - Recommender systems
 - Laplacian vs. Gaussian mechanism for DP/approxDP
 - DP in model inversion
 - Adversarial training

Takeaway 3: Privacy vs. Utility

- K-anonymity

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

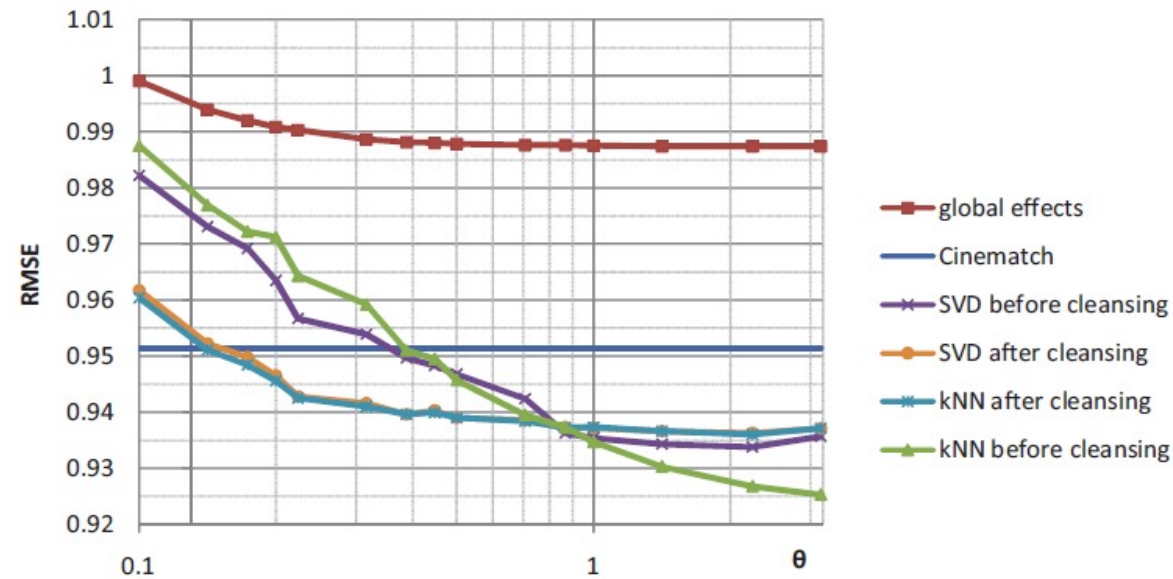
	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Takeaway 3: Privacy vs. Utility

- Recommender systems

$$\text{Cov} = \sum_u w_u \hat{r}_u \hat{r}_u^T + \text{Noise}^{d \times d}$$



Privacy decreases →

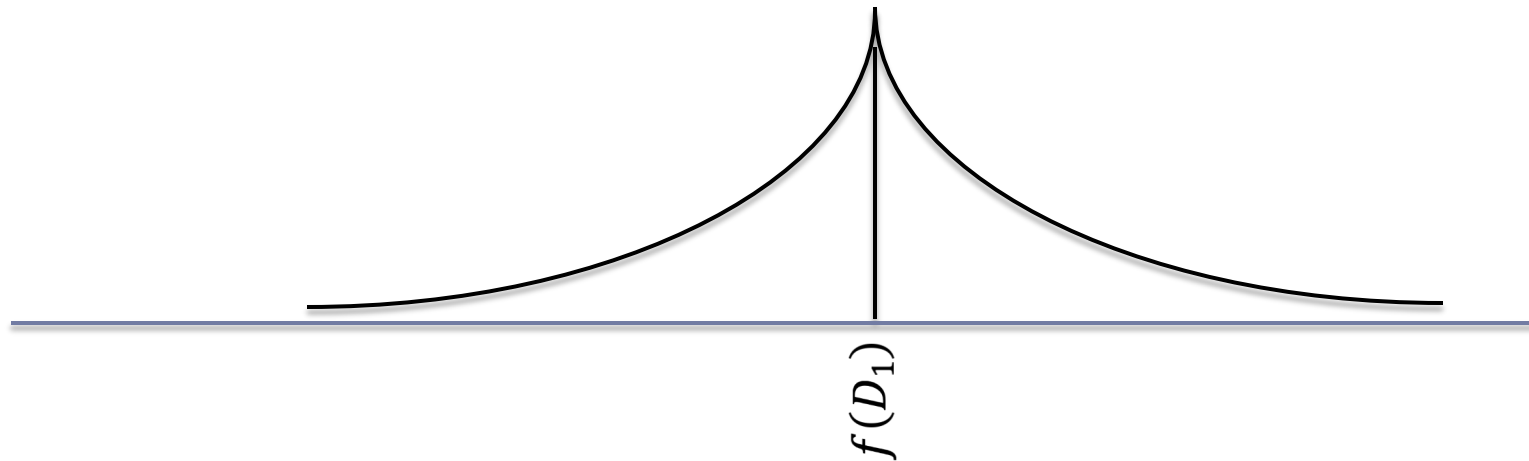
Takeaway 3: Privacy vs. Utility

- Laplacian vs. Gaussian mechanism for DP/approxDP

Review: Laplacian Mechanism

$$K(D) = f(D) + \text{Lap}\left(\frac{GS_f}{\epsilon}\right)$$

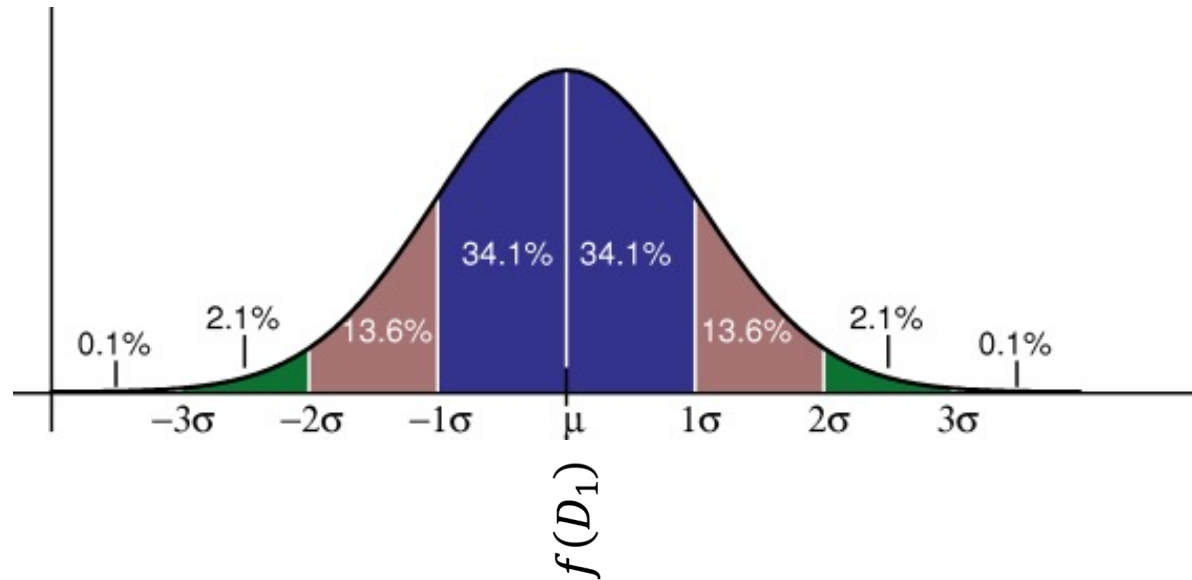
Thm: K is ϵ -differentially private



Review: Gaussian Mechanism

$$K(D) = f(D) + N(\sigma^2)$$

Thm K is (ϵ, δ) -differentially private as long as $\sigma \geq \frac{\sqrt{2 \ln(2/\delta)}}{\epsilon} \times GS_f$

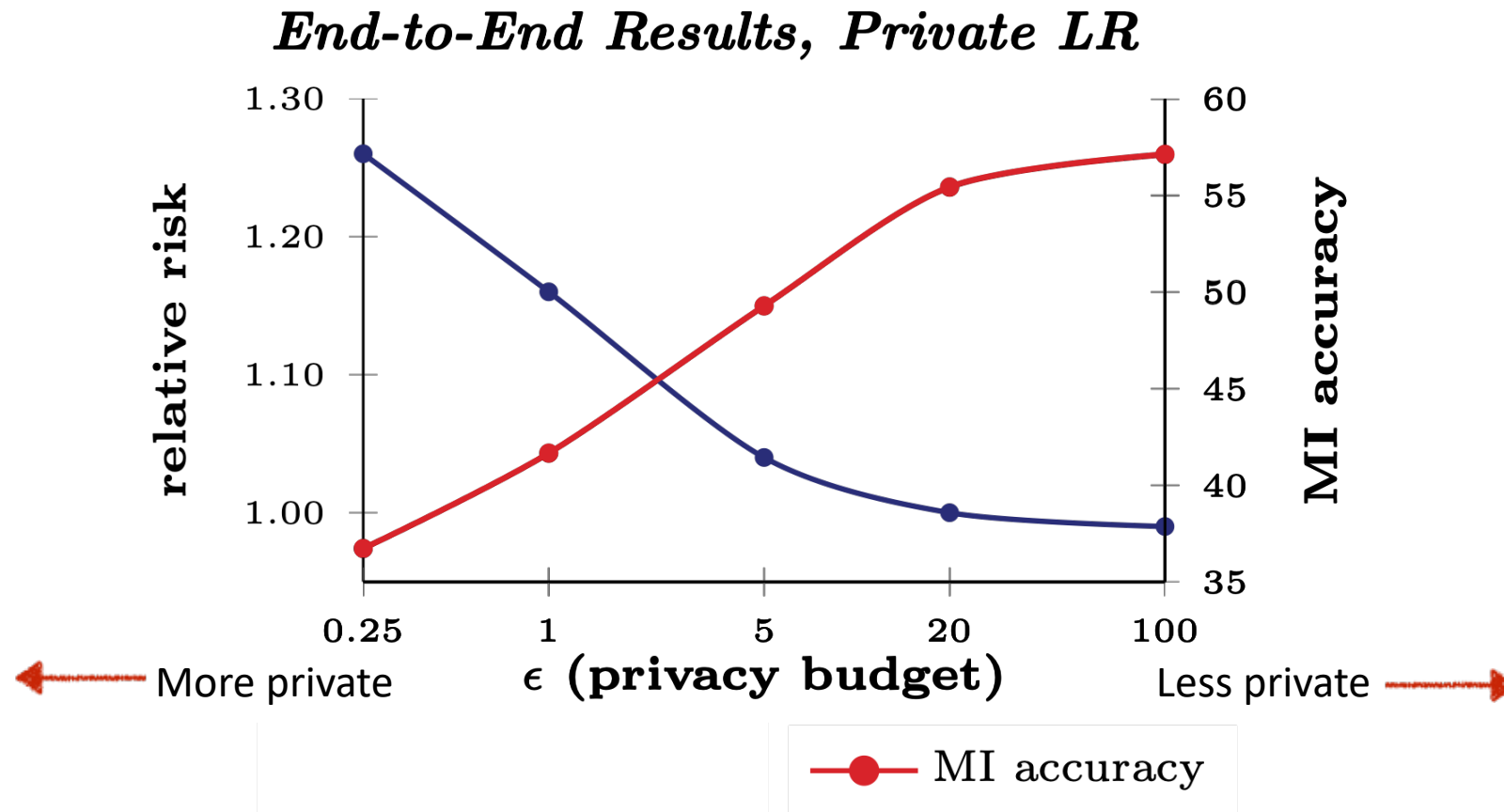


Approximate Differential Privacy

The privacy guarantees made by (ϵ, δ) -*differential privacy* are not as strong as ϵ -differential privacy, but less noise is required to achieve (ϵ, δ) -*differential privacy*. So, it provides better utility, i.e. more accurate answers

Takeaway 3: Privacy vs. Utility

- DP in model inversion



Takeaway 3: Privacy vs. Utility

- Adversarial training

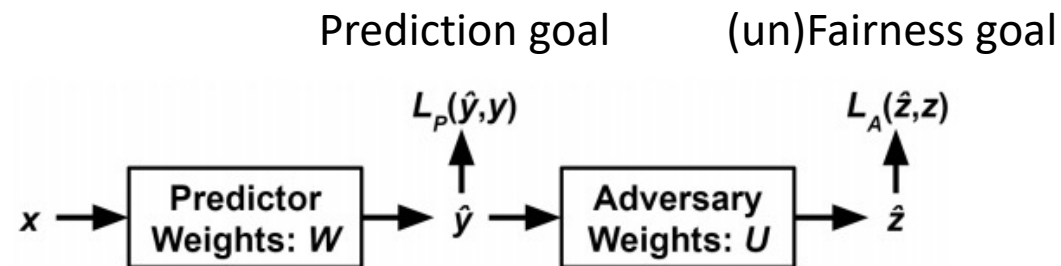


Figure 1: The architecture of the adversarial network.

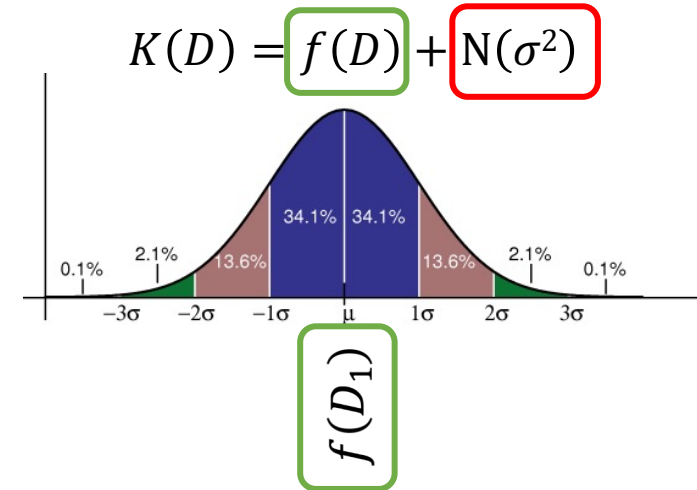
Takeaway 3: Privacy vs. Utility

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	40	*	Cancer
6	1485*	40	*	Heart Disease
7	1485*	40	*	Viral Infection
8	1485*	40	*	Viral Infection
9	130**	*	*	Cancer
10	130**	*	*	Cancer
11	130**	*	*	Cancer
12	130**	*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata



$$\text{Cov} = \sum_u w_u \hat{r}_u \hat{r}_u^T + \text{Noise}^{d \times d}$$

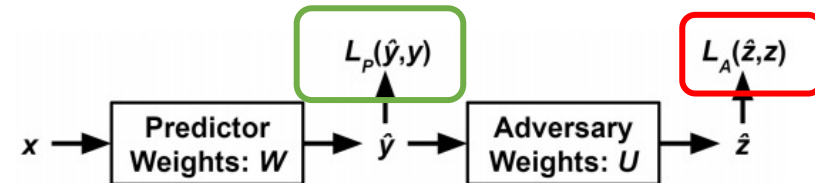
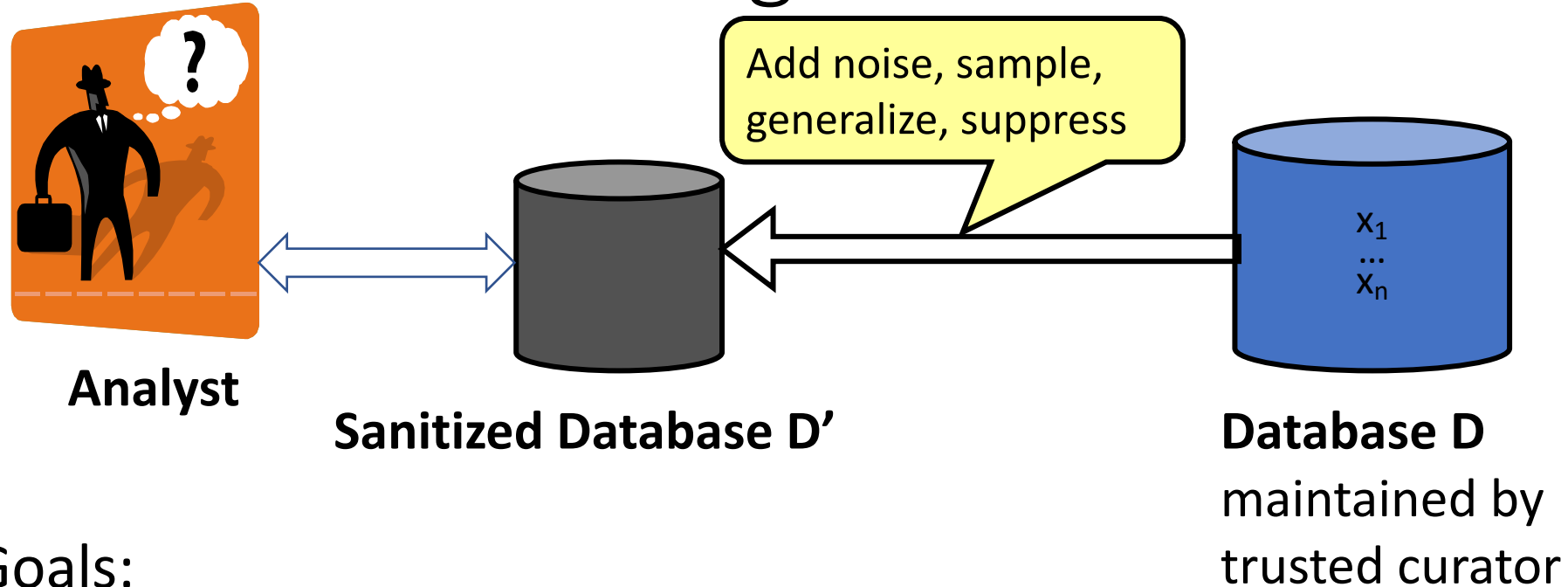


Figure 1: The architecture of the adversarial network.

Privacy-Preserving Statistics: Non-Interactive Setting

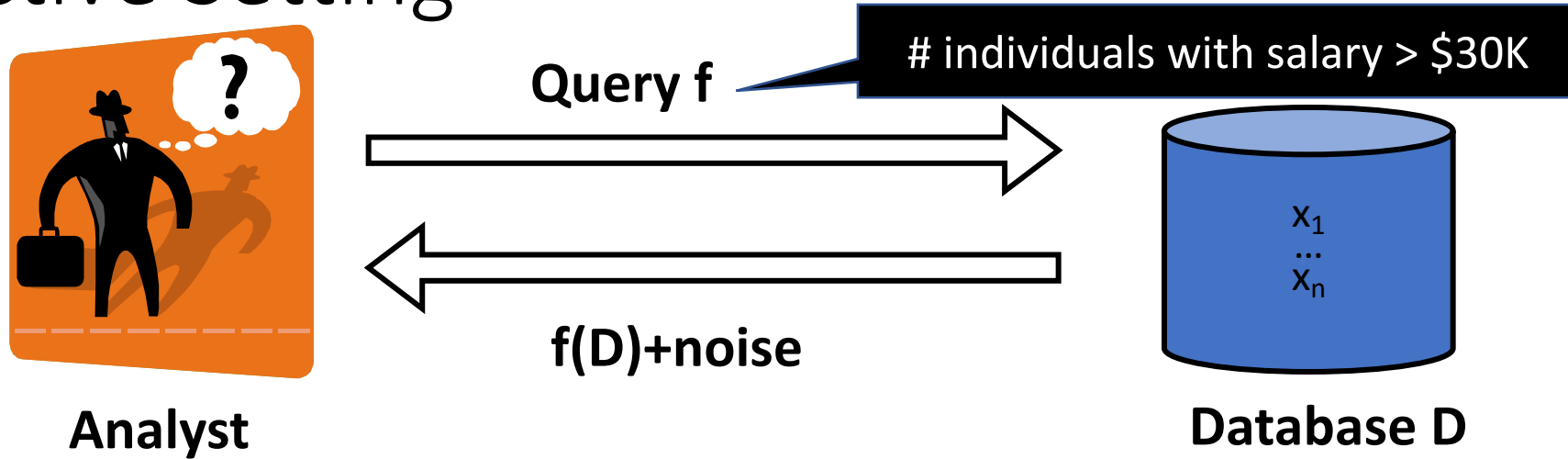


Goals:

- Accurate statistics (low noise)
- Preserve individual privacy
(what does that mean?)

- Census data
- Health data
- Network data
- ...

Privacy-Preserving Statistics: Interactive Setting



Goals:

- Accurate statistics (low noise)
- Preserve individual privacy
(what does that mean?)

- Census data
- Health data
- Network data
- ...

Possible things to discuss

- Privacy attacks
 - R. Shokri, M. Stronati, C. Song and V. Shmatikov, Membership inference attacks against machine learning models, in 2017 IEEE Symposium on Security and Privacy (SP), 2017. <https://arxiv.org/abs/1610.05820>
- Differentially private machine learning
 - Differential Privacy Has Disparate Impact on Model Accuracy <https://arxiv.org/abs/1905.12101>
- Soham: include homework 3 clarifications reminder, pointing students to here: <https://piazza.com/class/kmp7hadai9l73y?cid=137>
- Questions posed in model inversion video:
 - How does sensitive information end up in a model (or dataset)?
 - Is the model or the background knowledge more responsible for model inversion attack success?
- What does differential privacy protect against?
- Interactive vs. offline setting differences
- Discussion of “Terry Gross” example.
 - Who/what is responsible for the privacy violation?
- Role of background knowledge.
 - Differential Privacy: A Survey of Results
 - No Free Lunch in Data Privacy
- Privacy vs. explanations