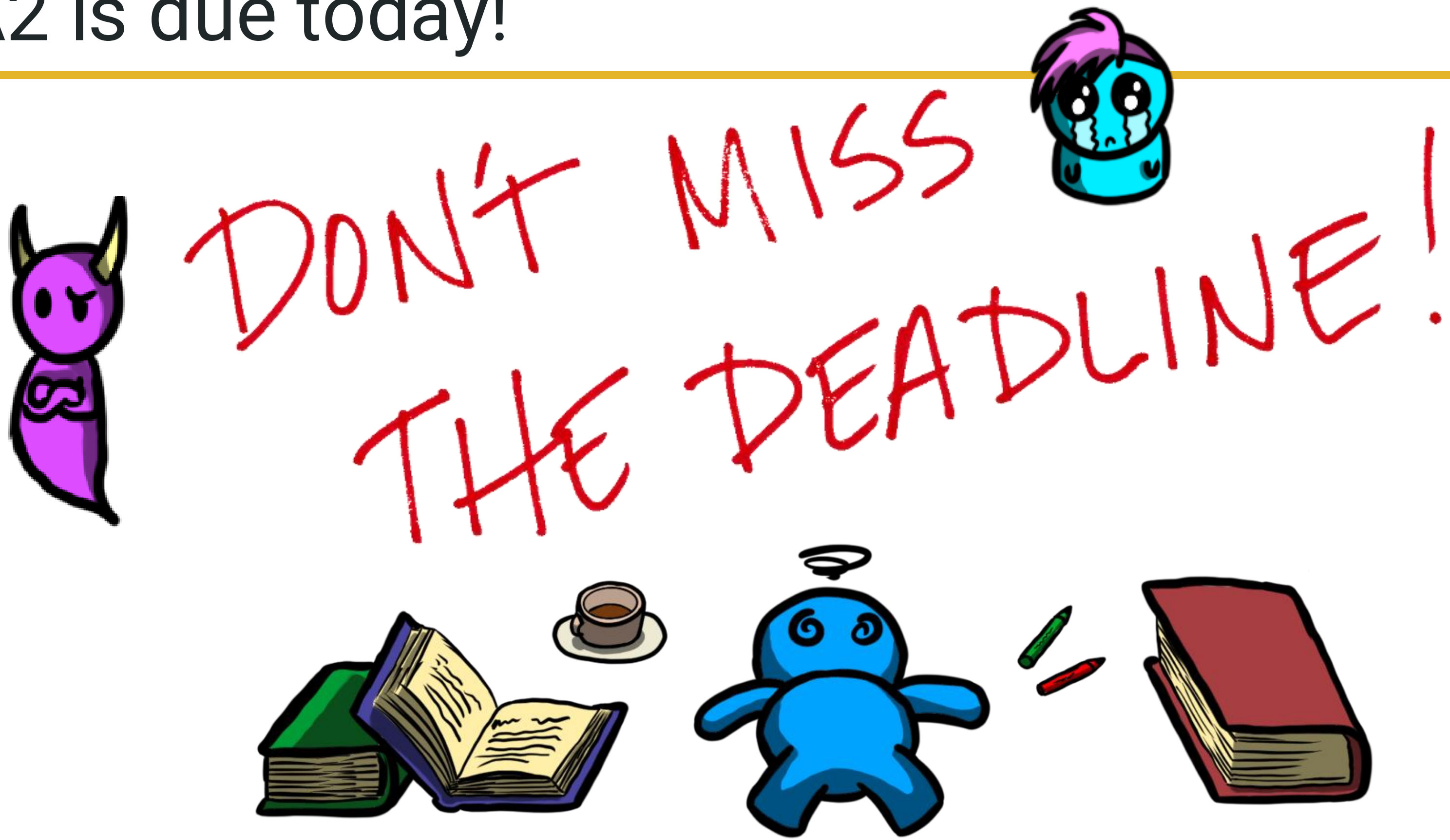
CS459/698
Privacy, Cryptography, Network and Data Security

Inference Attacks

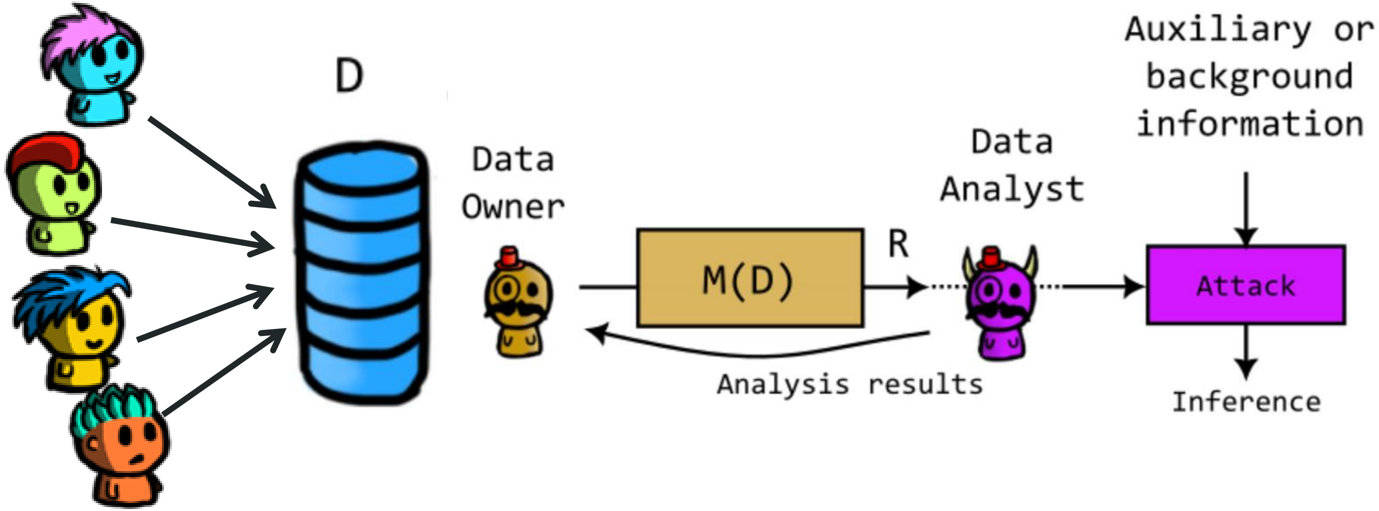
Fall 2024, Tuesday/Thursday 02:30pm-03:50pm

A2 is due today!



What are inference attacks?

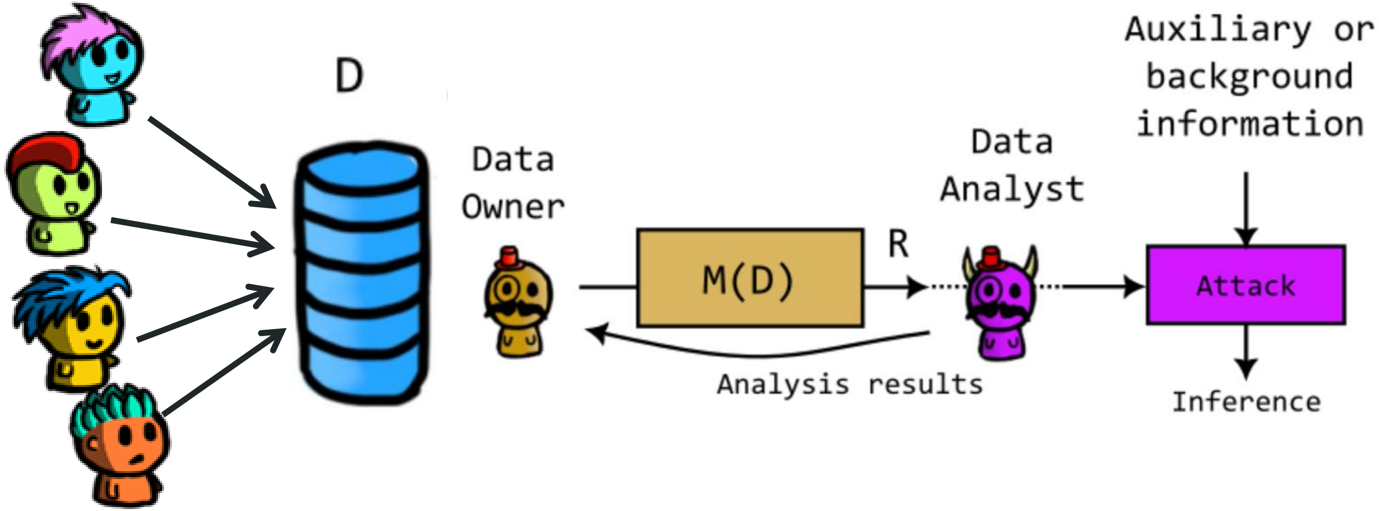
The adversary (e.g., the system provider, a data analyst, an eavesdropper, etc.) gains access to some (sanitized) data.



What are inference attacks?

The adversary (e.g., the system provider, a data analyst, an eavesdropper, etc.) gains access to some (sanitized) data.

The adversary could have auxiliary/background information.

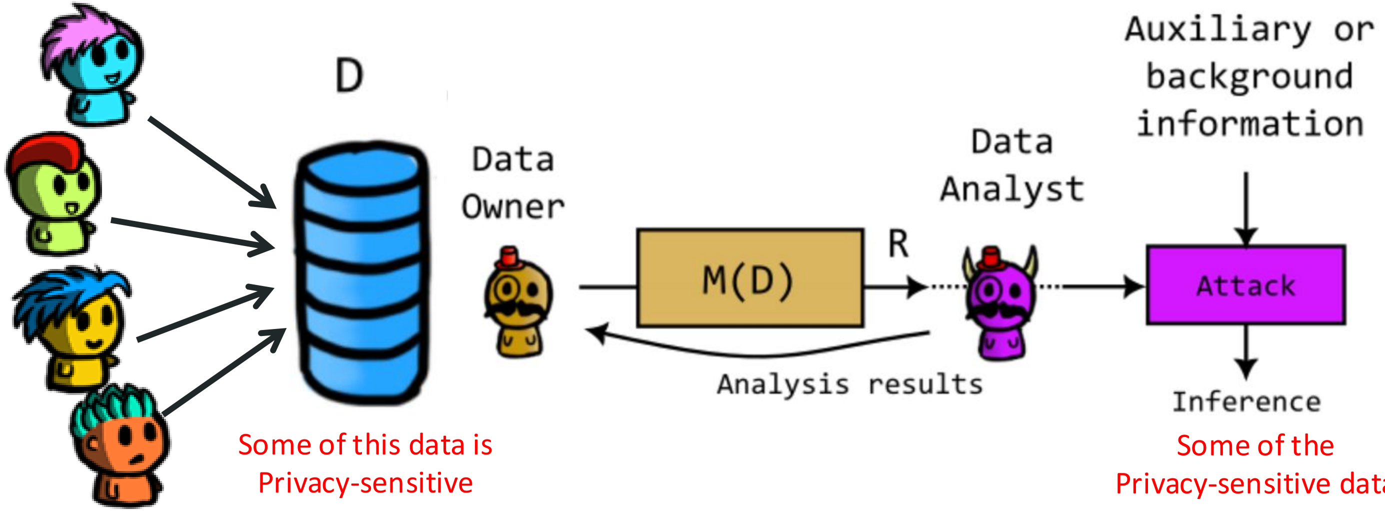


What are inference attacks?

The adversary (e.g., the system provider, a data analyst, an eavesdropper, etc.) gains access to some (sanitized) data.

The adversary could have auxiliary/background information.

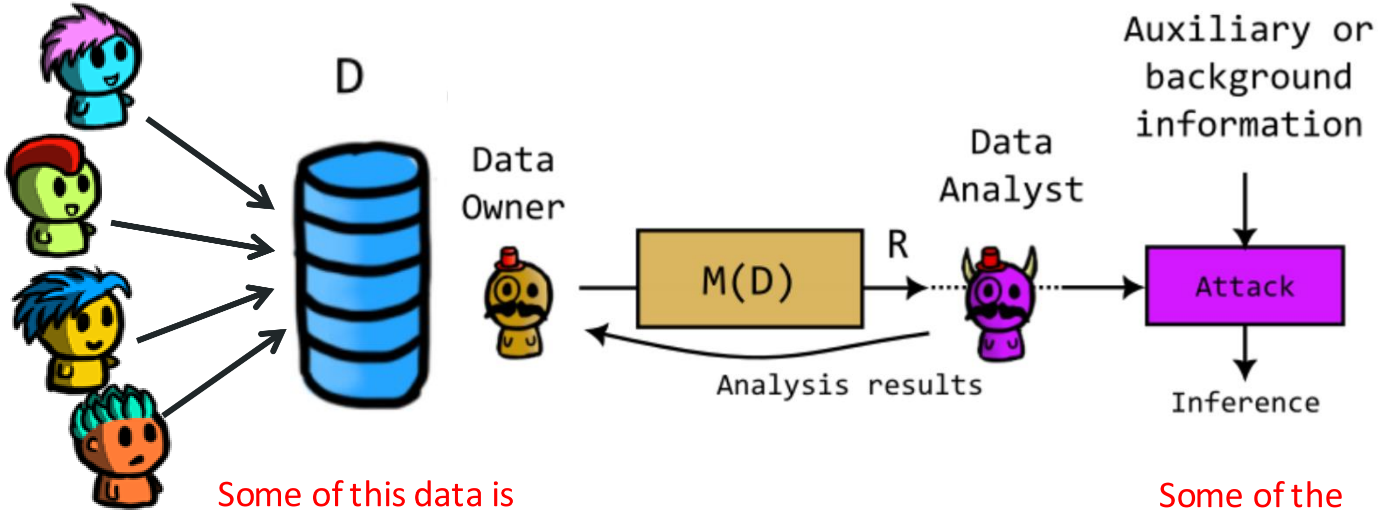
An inference attack infers privacy-sensitive information from this information.



Some of this data is Privacy-sensitive

Some of the Privacy-sensitive data

What are inference attacks?

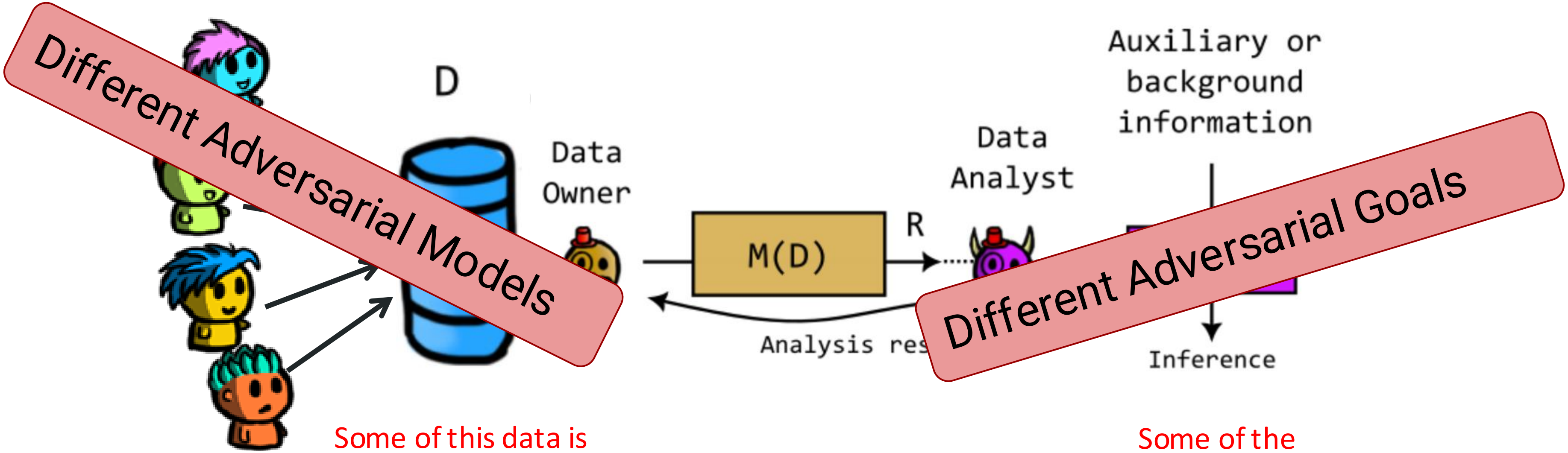


Some of this data is
Privacy-sensitive

Some of the
Privacy-sensitive data

Goal: Learn something (non-trivial) and privacy sensitive from the system

What are inference attacks?



Some of this data is
Privacy-sensitive

Some of the
Privacy-sensitive data

Goal: Learn something (non-trivial) and privacy sensitive from the system

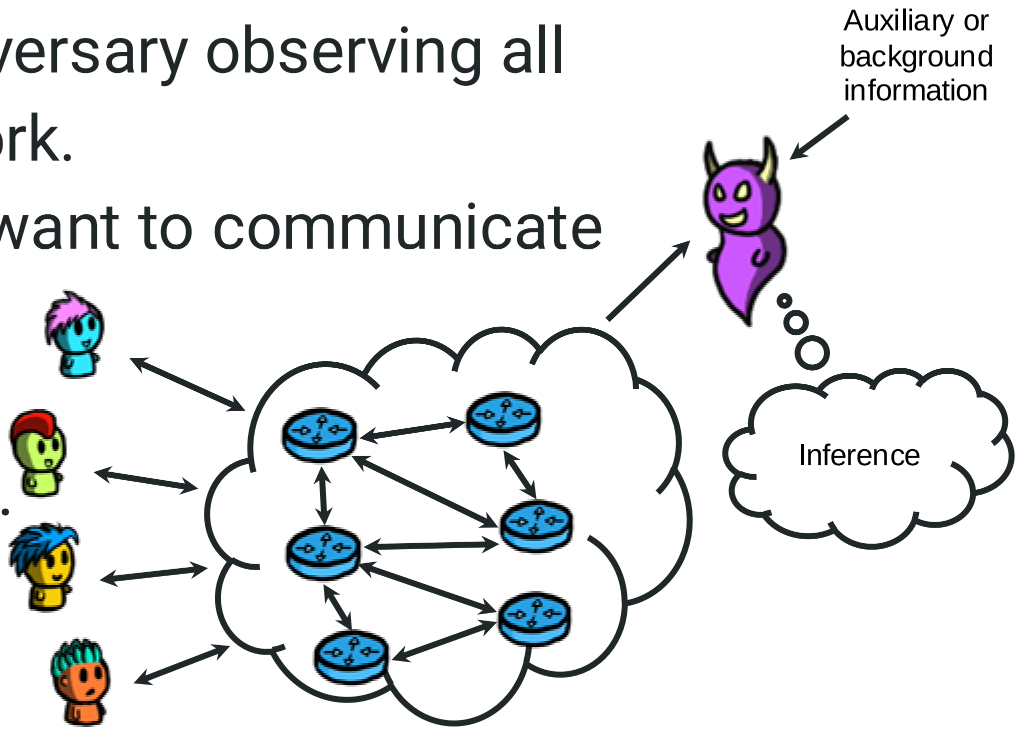
Context for Inference Attacks: The Model

- Attacks generally rely on information “leakage”
- The leakage can be intentional:
  - Sending usage statistics to a service provider (Microsoft, Apple, …)
  - Reporting our location to Google Maps
  - Publishing census data
- Some leakage is unintentional:
  - E.g., side-channels: you saw these earlier!

Attacks can combine all leaked information with auxiliary information to infer non-trivial sensitive data!

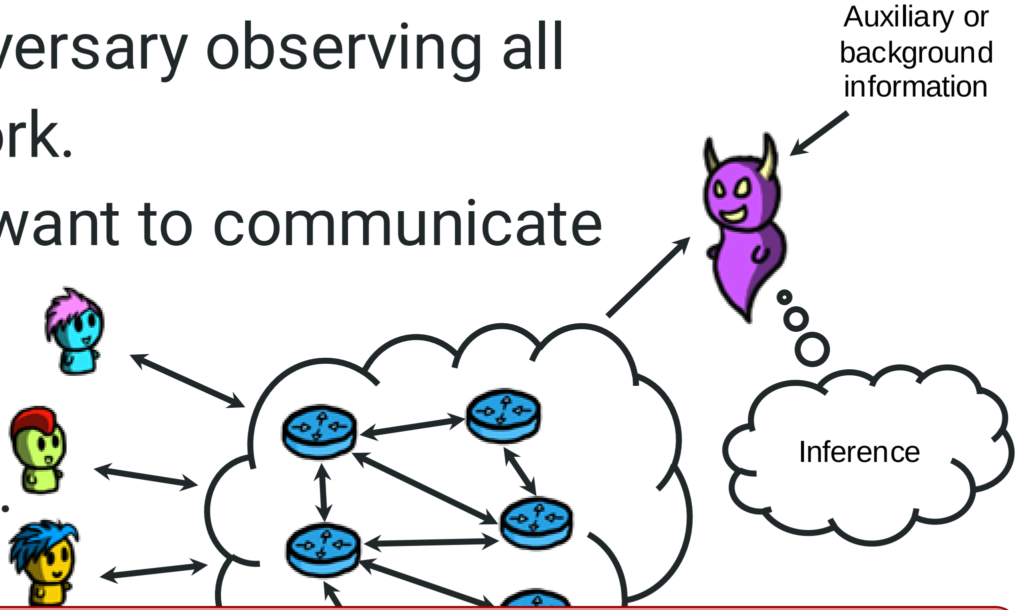
Example 1: Communication Systems

- **Adversary:** A passive adversary observing all flows of traffic in a network.
- **Functionality:** The users want to communicate with each other (they don't intend to leak anything to an adversary).



Example 1: Communication Systems

- **Adversary:** A passive adversary observing all flows of traffic in a network.
- **Functionality:** The users want to communicate with each other (they don't intend to leak anything to an adversary).



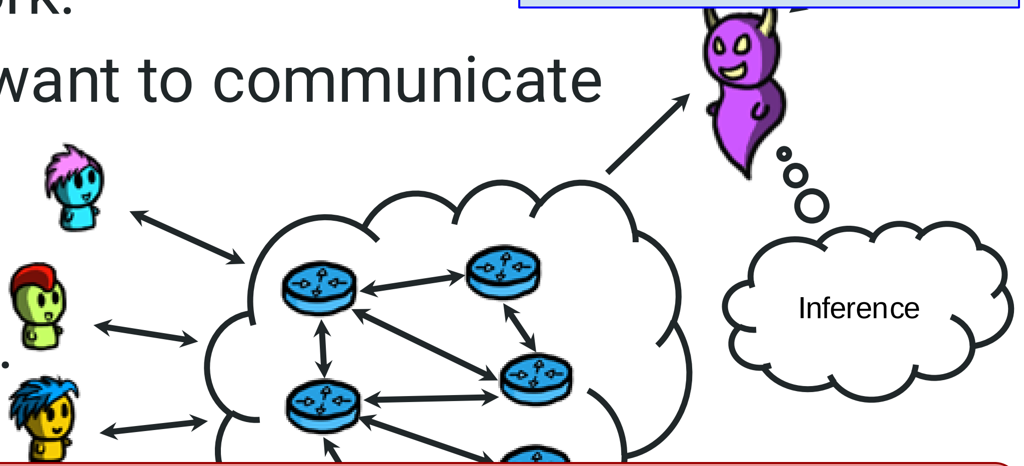
Q: What non-trivial privacy-sensitive information could the adversary infer?

Example 1: Communication Systems

Leakage:

- Packet payload
- Packet headers
- Timing information

- **Adversary:** A passive adversary observing a flows of traffic in a network.
- **Functionality:** The users want to communicate with each other (they don't intend to leak anything to an adversary).



Q: What non-trivial privacy-sensitive information could the adversary infer?

Example 1: Communication Systems

Leakage:
- Packet payload
- Packet headers
- Timing information

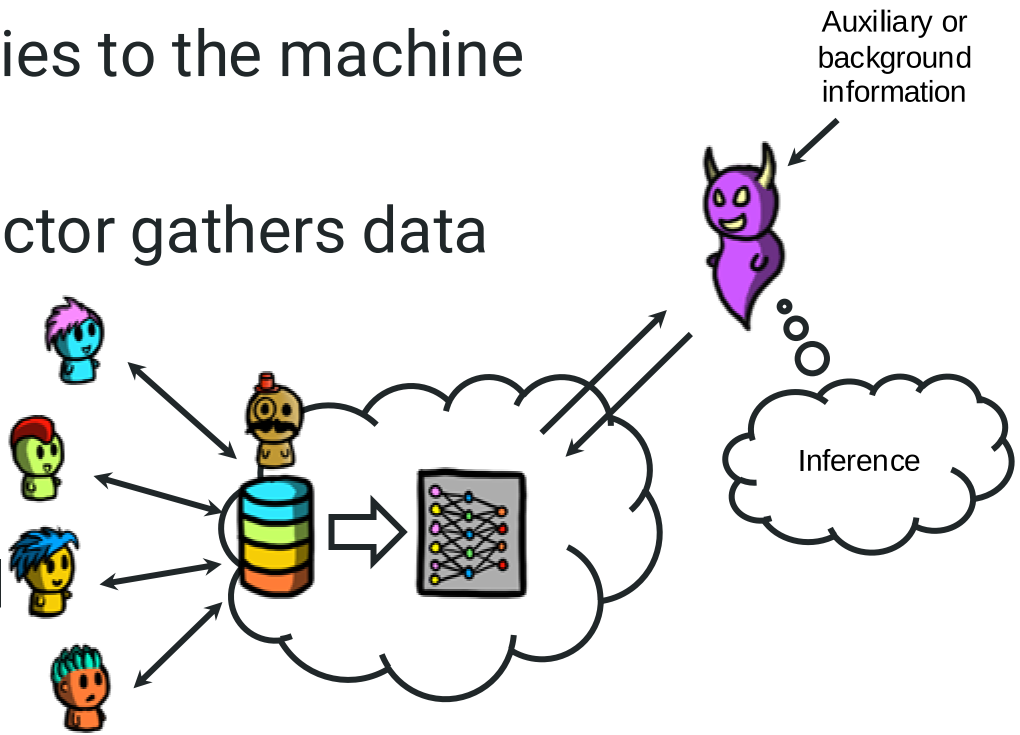
- **Adver**
  flows
- **Func**
  with
  (they
  anyth

A:
- What the users are talking about
- Who is talking with whom
- The social graph of the users
- How often two users communicate
- How often a user participates in a system
- Whether or not a user communicates at all
- …

Q: What non-trivial privacy-sensitive information could the adversary infer?

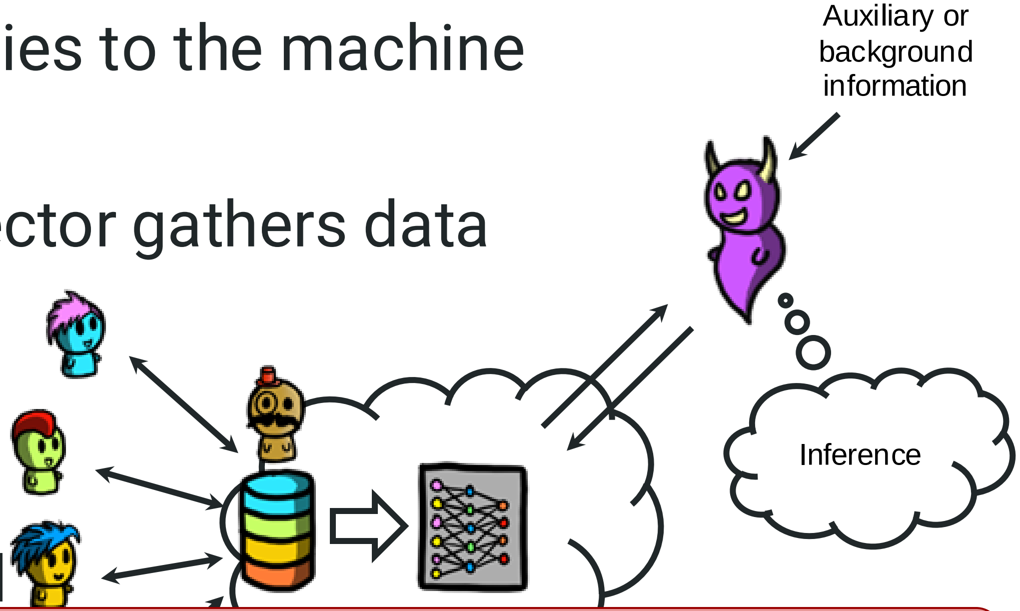
Example 2: Machine Learning

- **Adversary:** can issue queries to the machine learning model.
- **Functionality:** A data collector gathers data from users and trains a machine learning model with it (they don't intend to leak anything non-trivial to the adversary).



Example 2: Machine Learning

- **Adversary:** can issue queries to the machine learning model.
- **Functionality:** A data collector gathers data from users and trains a machine learning model with it (they don't intend to leak anything non-trivial

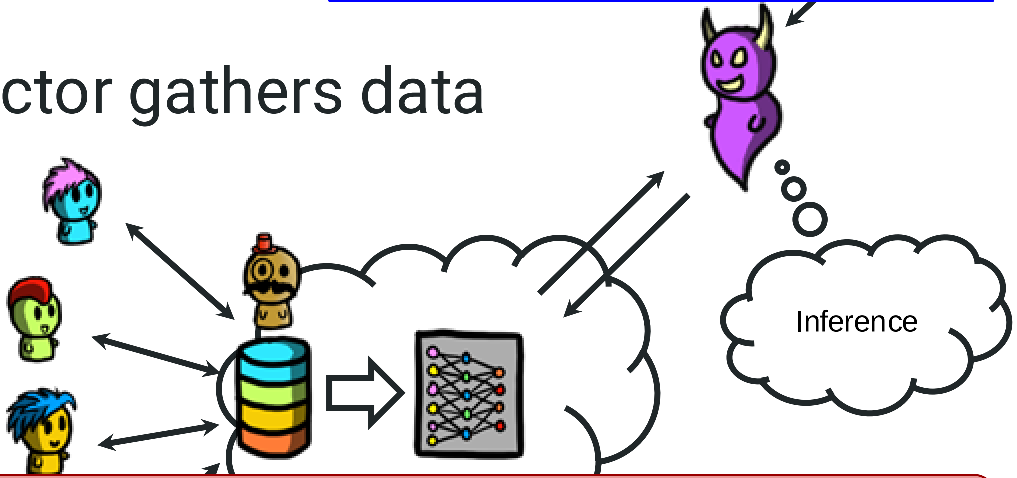


Q: What non-trivial privacy-sensitive information could the adversary infer?

Example 2: Machine Learning

- **Adversary:** can issue queries to the m
  learning model.
- **Functionality:** A data collector gathers data from users and trains a machine learning model with it (they don't intend to leak anything non-trivial

Leakage:
- Inferences from the ML model



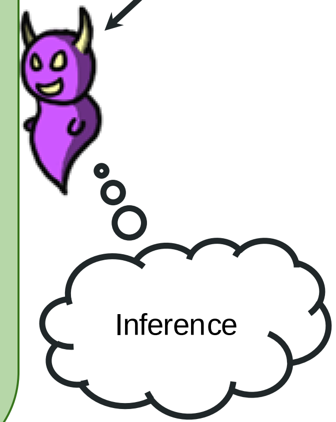
Q: What non-trivial privacy-sensitive information could the adversary infer?

Example 2: Machine Learning

- **Advers**
  learn
- **Funct**
  from
  mach
  with i
  to leak

A:
- Each user's data (the whole training dataset)
- Whether or not a particular data sample was in the training set
- A general property of the training population
- Given partial data about a user, learn other attributes about the user
- …

Leakage:
- Inferences from the ML model



Q: What non-trivial privacy-sensitive information could the adversary infer?

Why study inference attacks?

Adversarial Thinking

- Think like an adversary to understand the ***vulnerabilities*** of a system and develop ***protection techniques***.

- With inference attacks, we also apply **Kerckhoff's principle** (or Shannon's maxim), adapted to privacy

Adversarial Thinking

- Think like an adversary to understand the ***vulnerabilities*** of a system and develop ***protection techniques***.

- With inference attacks, we also apply **Kerckhoff's principle** (or Shannon's maxim), adapted to privacy

Assume the adversary knows how the system works
- there are no hidden parameters other than the users' data
- the adversary can even know some rough distribution that the users' data follows.

Designing a System Aware of Inference Attacks

For any system that relies on users' data, there are two goals:

- **Utility:** Design a system that provides benefits to its users and the service provider
- **Privacy:** Design a system that provides protection against inference attacks

Q: What are "utility" and "privacy"? How do we "measure" them?

Designing a System Aware of Inference Attacks

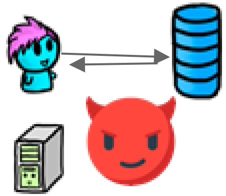
For any system that relies on users' data, there are two goals:

- **Utility:** Design a system that provides benefits to its users and the service provider
- **Privacy:** Design a system that provides protection against inference attacks

Q: What are “utility” and “privacy”? How do we “meas

It's complicated…

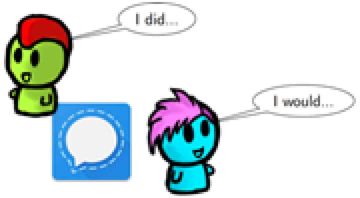
Recall, What is privacy?



Technical Privacy

Conceptual Privacy

Legal Privacy



Usable Privacy

What is privacy?

- Useful definition: informational self-determination

"The right of the individual to decide what information about himself should be communicated to others and under what circumstances" (Westin, 1970)

- Privacy is having control over:
  - Who we share our data with
  - Who they can share it with
  - For what purpose they use it
  - Etc.

Quantifying Privacy?

- Protecting the sensitive information e.g., not just data, also meta-data, relationships, timing, whether a user participated in a system, etc.
- Quantifying privacy is very hard

There is **no "one solution fits all" metric** for privacy, measuring privacy can be computationally intractable, etc.

Quantifying Privacy: Theoretical Notions

- **Syntactic** notions of privacy: these are computed on the leaked or released data. They are data dependent
  - K-anonymity, l-diversity, t-closeness, etc

Quantifying Privacy: Theoretical Notions

- **Syntactic** notions of privacy: these are computed on the leaked or released data. They are data dependent
  - K-anonymity, l-diversity, t-closeness, etc
- **Semantic** notions of privacy: these are computed on the data release mechanism itself, and they hold regardless of the data (data independent)
  - Differential Privacy

Quantifying Privacy: Empirical Notions

- The performance of an **inference attack** e.g., the attacker error, accuracy, true positive rate, false positive rate, etc

→ Can provide an **upper bound** on privacy

Quantifying Privacy: Empirical Notions

- The performance of an **inference attack** e.g., the attacker error, accuracy, true positive rate, false positive rate, etc

→ Can provide an **upper bound** on privacy

Q: Why an upper bound?

Quantifying Privacy: Empirical Notions

- The performance of an **inference attack** e.g., the attacker error, accuracy, true positive rate, false positive rate, etc
- Can provide an **upper bound** on privacy

Q: Why an upper bound?

A: Can't get more privacy if this attack succeeds

Utility and Privacy

Utility

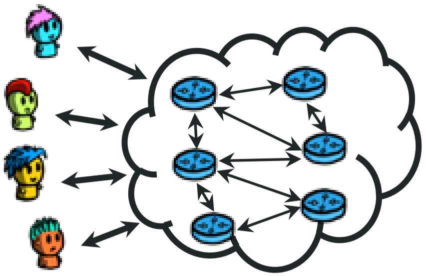
Definition: the benefit that users (and the provider) get from using the system.

Utility

Definition: the benefit that users (and the provider) get from using the system.

Communications system:

- For users: being able to communicate

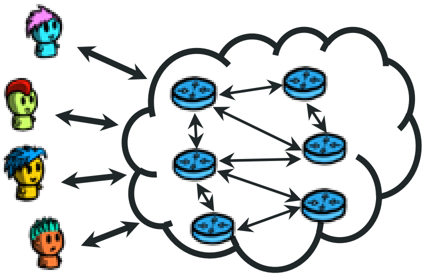


Utility

Definition: the benefit that users (and the provider) get from using the system.

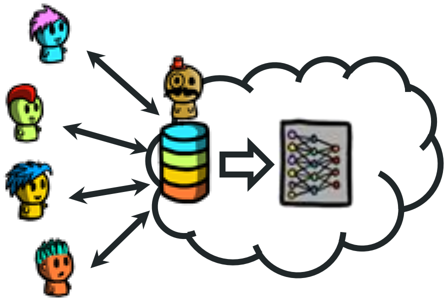
Communications system:

- For users: being able to communicate



Machine learning:

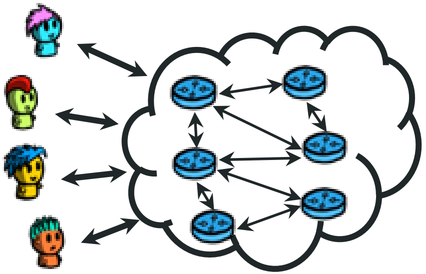
- For participants: maybe they get compensation?
- For data owner: it can sell access to the model for revenue
- Analysts: they pay to get benefits from the model's outputs
- General public: maybe the model outputs are good for society?



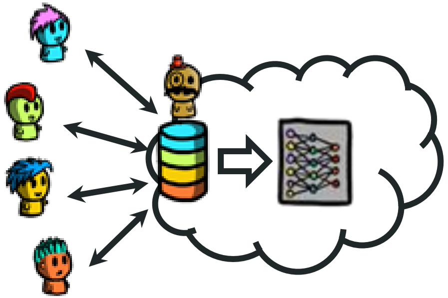
Quantifying Utility

Q: How do we *quantify* utility?

Communications system:



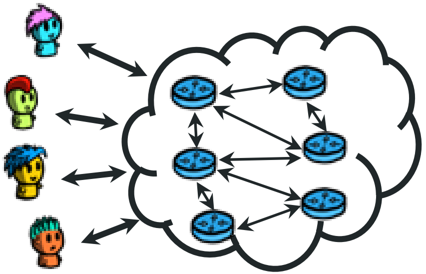
Machine learning:



Quantifying Utility

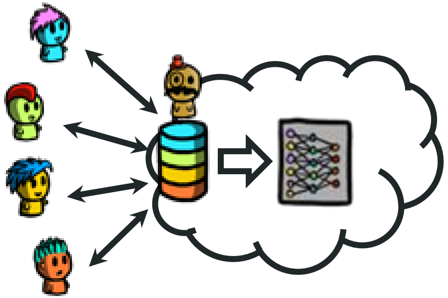
Q: How do we *quantify* utility?

Communications system:



- Few packets dropped
- High bandwidth/throughput
- Low latency/delay…

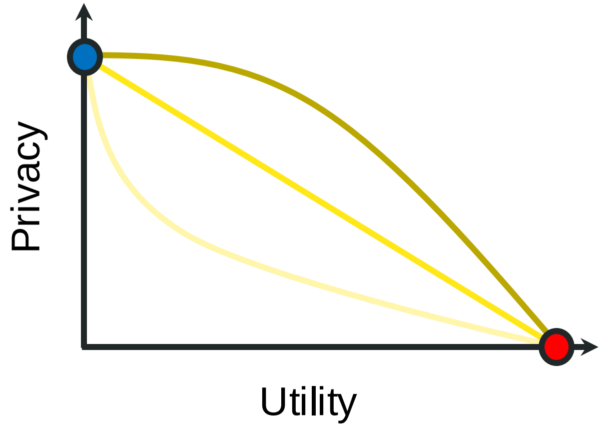
Machine learning:



- Useful model (high test accuracy)
- Unbiased model (low disparity among subpopulations)
- Low computational requirements to build the model
- Fast training algorithm…

The Privacy-Utility trade-off

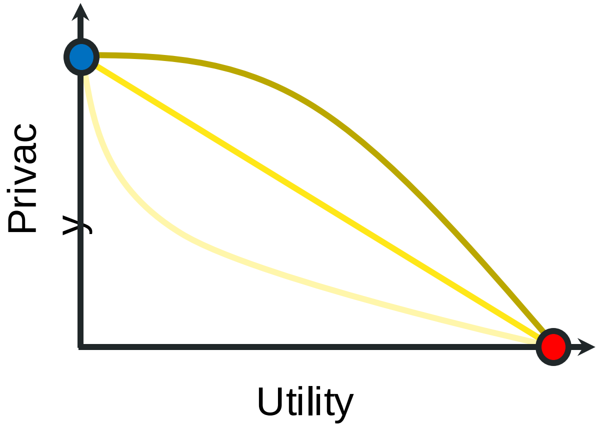
- Given any metric for privacy and for utility, they are usually at odds:



- **Q:** How do you design a system that provides maximum utility?
- **Q:** How do you design a system that provides maximum privacy?
- Designing a system that provides a good privacy-utility trade-off is hard!

The Privacy-Utility trade-off

- Given any metric for privacy and for utility, they are usually at odds:



- How do you design a system that provides maximum utility?
  - You design it without privacy in mind
- How do you design a system that provides maximum privacy?
  - You don't design it
- Designing a system that provides a good privacy-utility trade-off is hard!

Inference Attacks: Goals and Techniques

- As we saw before, the attacker can have different **goals**:
    - Infer data
    - Infer a property of the data
    - Infer the presence (membership) of some data
    - Infer the behavior of a user
    - Infer some attributes of a data sample
    - Infer dependencies among the data
    - …

Inference Attacks: Goals and Techniques

- As we saw before, the attacker can have different **goals**:
  - Infer data
  - Infer a property of the data
  - Infer the presence (membership) of some data
  - Infer the behavior of a user
  - Infer some attributes of a data sample
  - Infer dependencies among the data
  - …
- There are different **techniques** to perform an inference attack:
  - Statistics (estimation theory, maximum likelihood, Bayesian inference…)
  - Combinatorics
  - Heuristics
  - Machine learning
  - …

Inference Attack Examples

Inference attacks: examples

- For the rest of the lecture, we will see examples of inference attacks with different **goals** and **techniques**.

- You need to understand these attacks, their goal, the leakage they exploit and the techniques they use.
  - Given a new system, with some leakage specification and an attack goal, you should be able to come up with reasonable privacy/utility metrics and an inference attack.

Inference attacks: examples

We will see:

1. Census reconstruction attacks
2. SQL inference attacks (tracker attacks)
3. Database reconstruction attacks
4. Statistical inference attacks
   - Maximum Likelihood
   - Maximum A-Posteriori
5. ML Inference attacks
6. Linking attacks

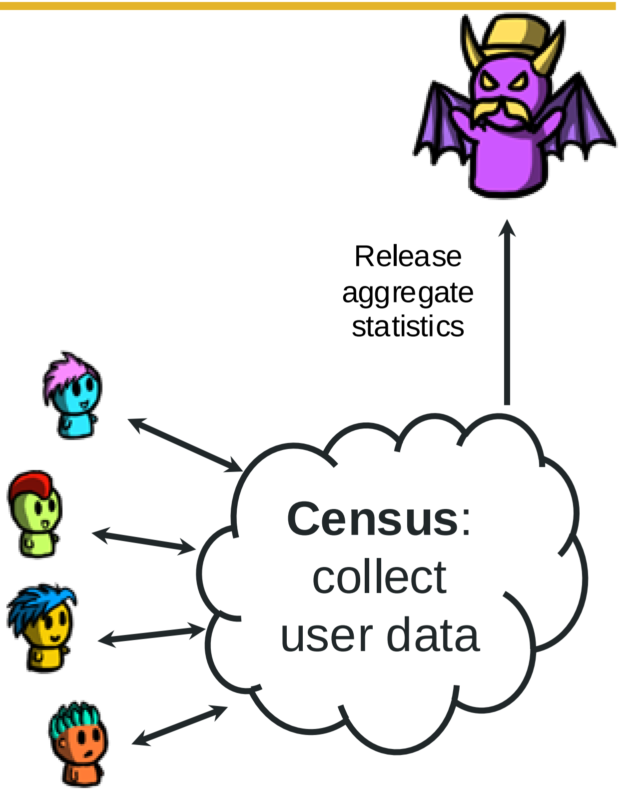
1. Census Reconstruction Attacks

U.S. census data vulnerable to attack without enhanced privacy measures

A new PNAS study shows that statistics released by the U.S. Census Bureau can be reverse engineered to reveal protected information about individual respondents.

1. Census Reconstruction Attacks

- A census involves collecting lots of privacy-sensitive data.
- Some useful aggregate statistics are released.
- The adversary tries to infer (reconstruct) some individuals' data.



1. Census Reconstruction Attacks

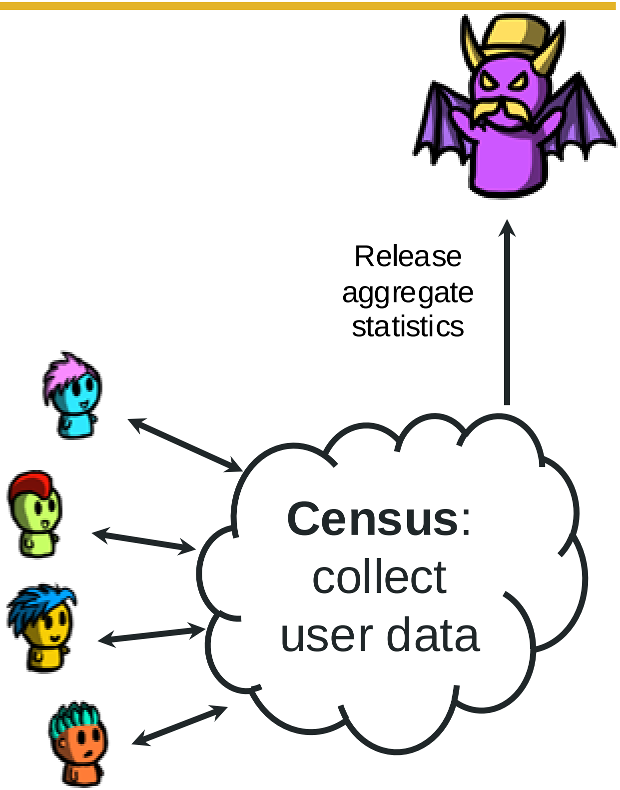
- A census involves collecting lots of privacy-sensitive data.
- Some useful aggregate statistics are released.
- The adversary tries to infer (reconstruct) some individuals' data.

➢ Example:

Background data: adversary knows a participant that self-identifies as white is 35 years old.

Released aggregates:

	COUNT	AGE MEAN
Total population	4	24
White	2	26
Asian	2	22



1. Census Reconstruction Attacks

- A census involves collecting lots of privacy-sensitive data.
- Some useful aggregate statistics are released.
- The adversary tries to infer (reconstruct) some individuals' data.

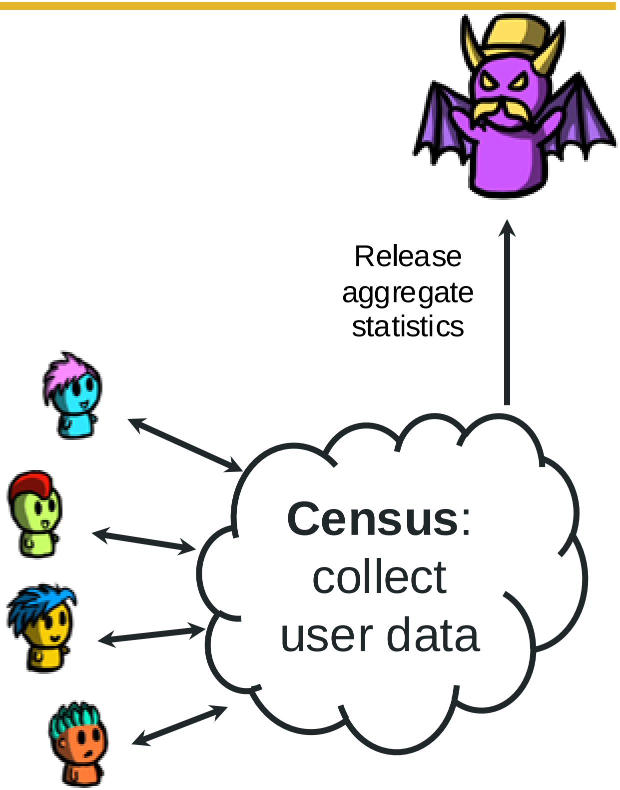
➢ Example:

Background data: adversary knows a participant that self-identifies as white is 35 years old.

Released aggregates:

	COUNT	AGE MEAN
Total population	4	24
White	2	26
Asian	2	22

Q: Can you guess the age and self-identified race of every participant?



1. Census Reconstruction Attacks

- A census involves collecting lots of privacy-sensitive data.
- Some useful aggregate statistics are released.
- The adversary tries to infer (reconstruct) some individuals' data.

➢ Example:

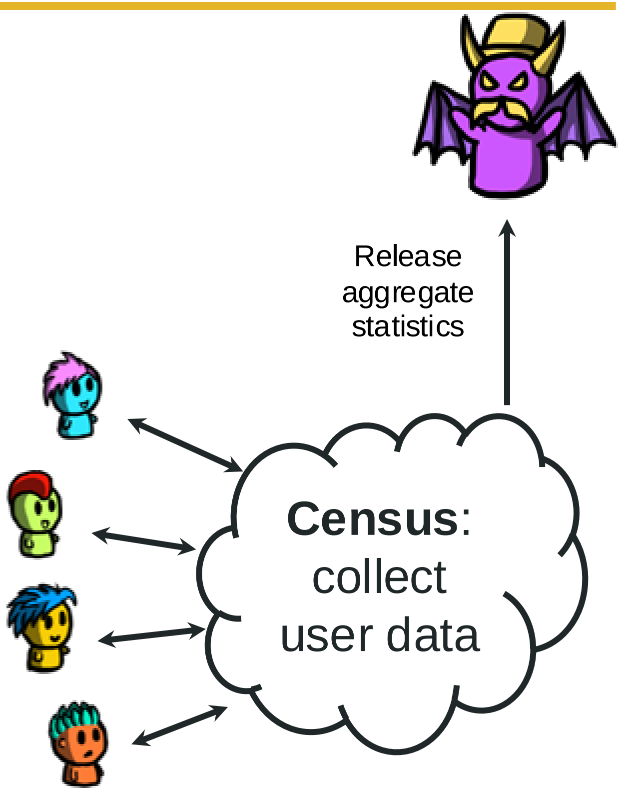
Background data: adversary knows a participant that self-identifies as white is 35 years old.

Released aggregates:

	COUNT	AGE MEAN
Total population	4	24
White	2	26
Asian	2	22

Q: Can you guess the age and self-identified race of every participant?

A: W1=17, W2=35, A1=21, A2=23



1. Census reconstruction attacks

- Another example, no background information:

Q: Can you guess the self-identified race, age, and marital status?

	COUNT	AGE MEAN	AGE MEDIAN
Total population	4	37.5	35.5
White	2	42.5	42.5
Asian	2	32.5	32.5
Single	1	25	25
Married	3	41.66	31

1. Census reconstruction attacks

- Another example, no background information:

Q: Can you guess the self-identified race, age, and marital status?

	COUNT	AGE MEAN	AGE MEDIAN
Total population	4	37.5	35.5
White	2	42.5	42.5
Asian	2	32.5	32.5
Single	1	25	25
Married	3	41.66	31

A: If you **assume the single person is Asian**, A1=25, then A2=40.
One white has to be W=31 (because that's the median of married), and the other white is W=54.
These values meet the total population age median.

1. Census reconstruction attacks

- Another example, no background information:

Q: Can you guess the self-identified race, age, and marital status?

	COUNT	AGE MEAN	AGE MEDIAN
Total population	4	37.5	35.5
White	2	42.5	42.5
Asian	2	32.5	32.5
Single	1	25	25
Married	3	41.66	31

A: If you **do the same assuming the single is White**, you get W1=25, W2=54, A1=31, A2=34, which does not meet the age median result, **so it can't be true.**

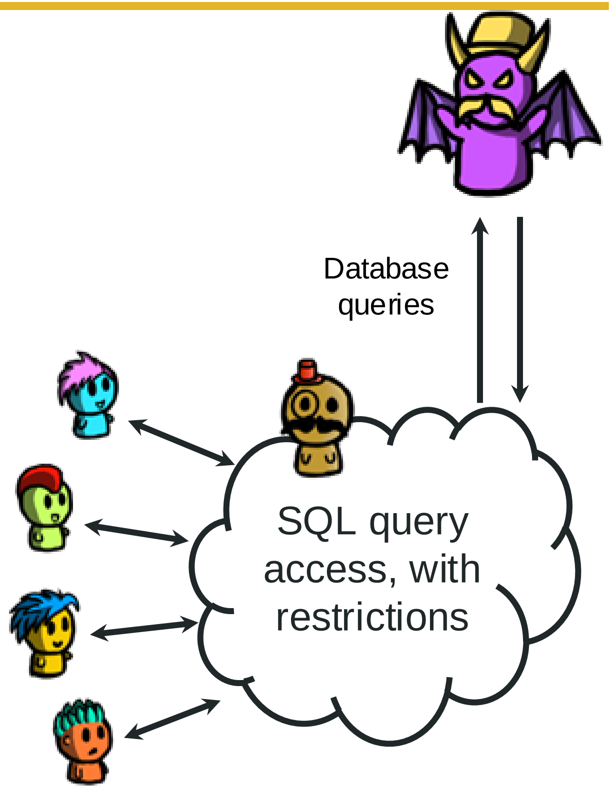
2. SQL Query Attacks

2. SQL query attacks

- A data collector creates a relational database (table) with data from different clients.
- An adversary can issue SQL queries to gather data from the table.
- The database management system allows queries with the following syntax:

  SELECT SUM(ATTRIBUTE) FROM (TABLE) WHERE (CONDITION)

- However, any queries that match less than X entries or more than N-X entries are discarded.



2. SQL query attacks: example

Name	Age	Position	Salary
Alice	40	full time	120k
…	…	…	…
Carol	…	…	…
…	…	…	…

- The table Employees has four attributes:
  - Names are unique
  - Ages are between 18 and 65
  - Position is either 'full time' or 'part time'
  - Salaries are between 50k and 500k
- You know Carol is in the dataset, and that around 50% of the people are 'full time'.
- There are **N** records in the dataset; any query that matches less than $\frac{N}{10}$ or more than $\frac{9N}{10}$ entries *is discarded.*
- Can you recover Carol's salary**?** How many queries do you need**?**

SELECT SUM(ATTRIBUTE) FROM (TABLE) WHERE (CONDITION)

2. SQL query attacks: solution

Name	Age	Position	Salary
Alice	40	full time	120k
…	…	…	…
Carol	…	…	…
…	…	…	…

- There are N records in the dataset; any query that matches less than $\frac{N}{10}$ or more than $\frac{9N}{10}$ entries *is discarded.*

Solution:

Q1=SELECT SUM(Salary) FROM Employees WHERE (Position='full time' OR Name=Carol)
Q2=SELECT SUM(Salary) FROM Employees WHERE (Position='full time' AND Name!=Carol)
Salary=Q1-Q2

If Carol is part time:

		Q1	Q2	Q1-Q2
Full time				
Part time				
	Carol			

If Carol is full time:

		Q1	Q2	Q1-Q2
Full time				
	Carol			
Part time				

Q1-Q2 always gets Carol's salary!

2. SQL query attacks:

- The lesson is: even if the result of a query is harmless (too general), the combination of two or more queries can be very dangerous (very specific).

- Placing restrictions on individual queries, while still reporting exactly values, does not work.

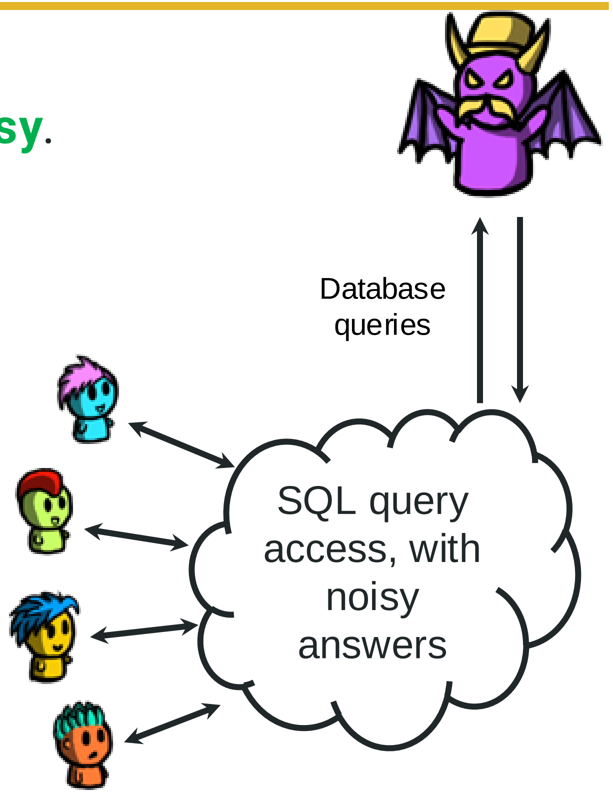
- When coming up with SQL query attacks in this setting:
    - Look for an attribute that you can use to make sure you always bypass the restriction so that the query goes through.
    - After you design the queries, check that they get the desired value regardless of the values of other attributes in the dataset (e.g,. whether Carol was full or part time in the example)

3. Database Reconstruction Attacks: Dinur-Nissim

3. Database Reconstruction Attacks: Dinur-Nissim

- Now we are going to see an example where the adversary can issue queries but the answers are **noisy**.
- We consider the case where the adversary knows everything in the database except for one binary attribute, e.g.,

Name	Age	Position	Salary
Alice	40	?	120k
Bob	40	?	80k
Carol	32	?	150k



3. Database Reconstruction Attacks: Dinur-Nissim

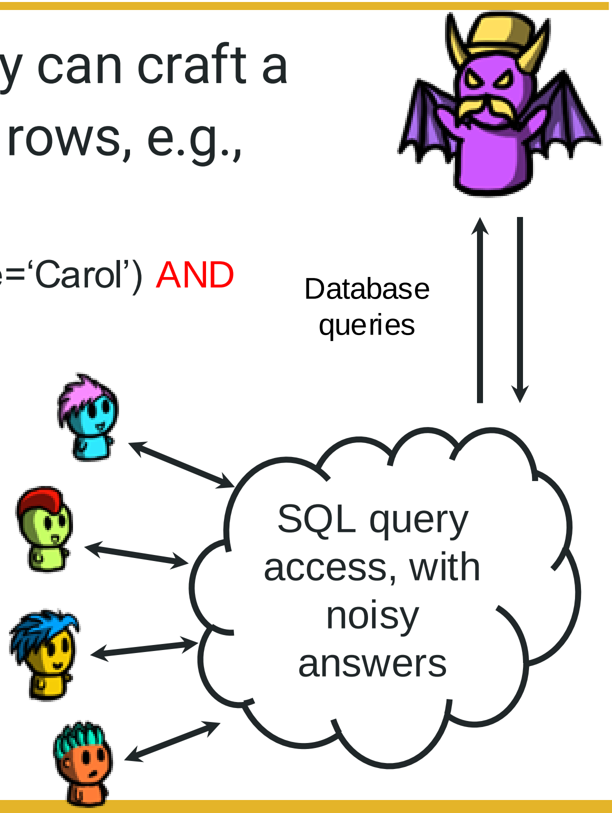
- Since the adversary knows the primary key, they can craft a condition that matches any specific number of rows, e.g.,

SELECT COUNT(*) FROM Employees WHERE (Name=‘Alice’ OR Name=‘Carol’) AND Position=‘full time’

Name	Age	Position	Salary
Alice	40	?	120k
Bob	40	?	80k
Carol	32	?	150k

True output:
- 0 if none are full time
- 1 if one is full time
- 2 if both are full time

(But the system will only report noisy outputs)



3. Dinur-Nissim attack: intuition

- Example: the DBMS adds noise uniformly chosen between -1, 0, 1.
- The server queries for the sum of rows that have ‘full time’ AND a certain combination of names, and does this for every possible combination of names. These are the outputs:

Name	Position
Alice	?
Bob	?
Carol	?

Rows	Output
Alice	2
Bob	1
Bob, Alice	0
Carol	1
Carol, Alice	2
Carol, Bob	2
Carol, Bob, Alice	1

Q: Can you tell who is full time and who is part time?

3. Dinur-Nissim attack: intuition

- Example: the DBMS adds noise uniformly chosen between -1, 0, 1.
- The server queries for the sum of rows that have ‘full time’ AND a certain combination of names, and does this for every possible combination of names. These are the outputs:

Name	Position
Alice	?
Bob	?
Carol	?

Rows	Output
Alice	2
Bob	1
Bob, Alice	0
Carol	1
Carol, Alice	2
Carol, Bob	2
Carol, Bob, Alice	1

Alice is full time

Q: Can you tell who is full time and who is part time?

3. Dinur-Nissim attack: intuition

- Example: the DBMS adds noise uniformly chosen between -1, 0, 1.
- The server queries for the sum of rows that have 'full time' AND a certain combination of names, and does this for every possible combination of names. These are the outputs:

Name	Position
Alice	?
Bob	?
Carol	?

Rows	Output
Alice	2
Bob	1
Bob, Alice	0
Carol	1
Carol, Alice	2
Carol, Bob	2
Carol, Bob, Alice	1

Alice is full time

??

Q: Can you tell who is full time and who is part time?

3. Dinur-Nissim attack: intuition

- Example: the DBMS adds noise uniformly chosen between -1, 0, 1.
- The server queries for the sum of rows that have ‘full time’ AND a certain combination of names, and does this for every possible combination of names. These are the outputs:

Name	Position
Alice	?
Bob	?
Carol	?

Rows	Output
Alice	2
Bob	1
Bob, Alice	0
Carol	1
Carol, Alice	2
Carol, Bob	2
Carol, Bob, Alice	1

← Alice is full time

← ??

← Bob is part time

Q: Can you tell who is full time and who is part time?

3. Dinur-Nissim attack: intuition

- Example: the DBMS adds noise uniformly chosen between -1, 0, 1.
- The server queries for the sum of rows that have ‘full time’ AND a certain combination of names, and does this for every possible combination of names. These are the outputs:

Name	Position
Alice	?
Bob	?
Carol	?

Rows	Output	
Alice	2	← Alice is full time
Bob	1	← ??
Bob, Alice	0	← Bob is part time
Carol	1	← ??
Carol, Alice	2	
Carol, Bob	2	
Carol, Bob, Alice	1	

Q: Can you tell who is full time and who is part time?

3. Dinur-Nissim attack: intuition

- Example: the DBMS adds noise uniformly chosen between -1, 0, 1.
- The server queries for the sum of rows that have ‘full time’ AND a certain combination of names, and does this for every possible combination of names. These are the outputs:

Name	Position
Alice	?
Bob	?
Carol	?

Rows	Output	
Alice	2	← Alice is full time
Bob	1	← ??
Bob, Alice	0	← Bob is part time
Carol	1	← ??
Carol, Alice	2	← ??
Carol, Bob	2	← Carol is full time
Carol, Bob, Alice	1	

Q: Can you tell who is full time and who is part time?

3. Dinur-Nissim attack: intuition

- Example: the DBMS adds noise uniformly chosen between -1, 0, 1.
- The server queries for the sum of rows that have ‘full time’ AND a certain combination of names, and does this for every possible combination of names. These are the outputs:

Name	Position
Alice	?
Bob	?
Carol	?

Rows	Output	
Alice	2	Alice is full time
Bob	1	??
Bob, Alice	0	Bob is part time
Carol	1	??
Carol, Alice	2	??
Carol, Bob	2	Carol is full time
Carol, Bob, Alice	1	Possible if both Carol and Alice are full time

Q: Can you tell who is full time and who is part time?

4. Statistical Inference Attacks

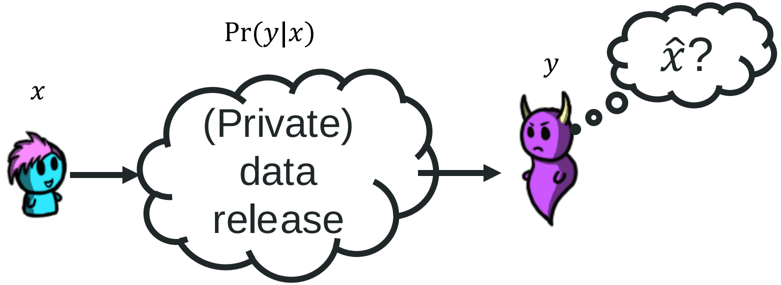
4. Statistical Inference: Probability recap

The following attacks require some basic knowledge of probability and statistics.
Let's do a recap.

- For simplicity, we assume *discrete* random variables here.
- $\boldsymbol{x}$ is Alice's private information, $\boldsymbol{y}$ is the leakage; usually $\hat{\boldsymbol{x}}$ is the adversary's estimate of x.
- $\boldsymbol{Pr(x)}$: the *prior* probability distribution of Alice's secret value
- $\boldsymbol{Pr(y|x)}$: the *mechanism* that models the leakage given Alice's secret information
    - In Bayesian inference, $\Pr(y|x)$ is also called the *likelihood* (of x having generated y)
- $\mathbf{Pr}(\boldsymbol{x}|\boldsymbol{y})$: the *posterior* probability distribution (the probability that x took a certain value given the observed leakage y)
- **Bayes' theorem** connects these concepts:

$$\Pr(x|y) = \frac{\Pr(y|x) \cdot \Pr(x)}{\Pr(y)}$$

- **Law of total** probability: $\Pr(y) = \sum_x \Pr(x) \Pr(y|x)$



4. Statistical Inference: Probability recap

- Recall the expected value of a random variable:

$$E\{x\} = \sum_x x \cdot \Pr(x)$$

- When the adversary sees y, they can compute the conditional expectation of x (leveraging the leakage y):

$$E\{x|Y = y\} = \sum_x x \cdot \Pr(x|y)$$

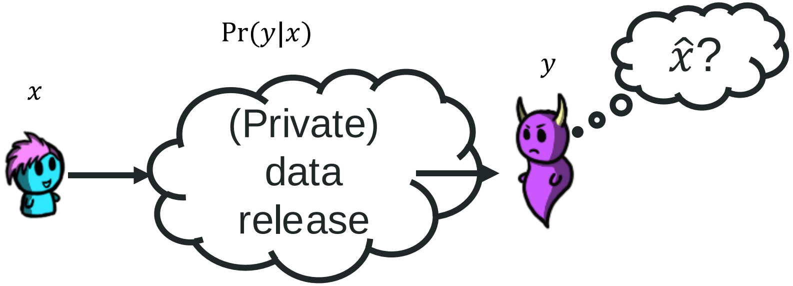
- Given y, $\Pr(x)$, and $\Pr(y|x)$, how do we run an attack (i.e., find x)?
  - There are many options!

4. Statistical Inference: Maximum Likelihood

- The **Maximum Likelihood** (ML) approach simply looks for the x that is *most likely* to have generated y, i.e.,

$$\hat{x} = \operatorname*{argmax}_{x} \Pr(y|x)$$

Q: what is the downside of this?



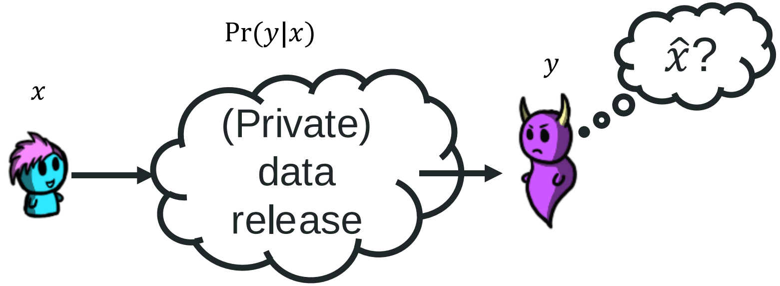
4. Statistical Inference: Maximum Likelihood

- The **Maximum Likelihood** (ML) approach simply looks for the x that is *most likely* to have generated y, i.e.,

$$\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(y|x)$$

Q: what is the downside of this?

A: Maybe that x had a very low prior probability.
- However, if the adversary does not know the prior, this is reasonable
  → They can still perform an inference attack by maximizing the **likelihood** P(D | x) alone.
  → Without the prior, the adversary doesn't need to compute the full **posterior** making the inference attack simpler (though potentially less accurate or informed).



4. Statistical Inference: Maximum A-Posteriori

- The **Maximum A-Posteriori** (MAP) approach chooses the x that maximizes the posterior probability:

$$\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(x|y)$$

- **Q:** Expand the posterior and simplify the expression:

$$\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(x|y) = \underset{x}{\operatorname{argmax}} \Pr(x) \cdot \Pr(y|x)$$

This is like ML, but taking into account the posterior.
Note that we do not need to compute $\Pr(y)$!

- **Q:** when are MAP and ML equivalent?
- When the prior is uniform! (every secret value x is just as likely)

4. Statistical Inference: other attacks

- MAP and ML choose an x that maximizes a probability.
  Sometimes the attacker just wants to get an x that is “as close as possible” to the real x.
- Let $d(x, \hat{x})$ be a distance measuring how different x and $\hat{x}$ are.

Q: What is the estimation of x (i.e., $\hat{x}$), that *minimizes* the *average distance* to x?

4. Statistical Inference: other attacks

- MAP and ML choose an x that maximizes a probability. Sometimes the attacker just wants to get an x that is "as close as possible" to the real x.
- Let $d(x, \hat{x})$ be a distance measuring how different x and $\hat{x}$ are.

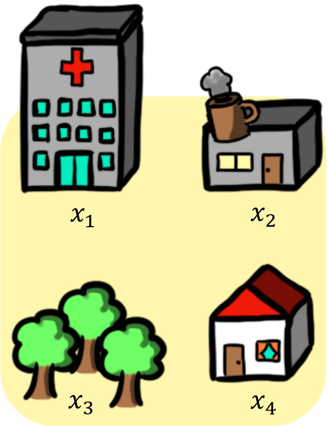
Q: What is the estimation of x (i.e., $\hat{x}$), that *minimizes* the *average (expected) distance* to x?

A:

$$\hat{x} = \underset{x'}{\operatorname{argmin}}\, E\{d(x, x')\} = \underset{x'}{\operatorname{argmin}} \sum_x \sum_y \Pr(x) \cdot \Pr(y|x) \cdot \mathrm{d}(\mathrm{x}, x')$$

Statistical Inference example: location privacy

- Alice wants to query for a location-based service, without revealing her real location x to the service provider. She runs a randomized mechanism $\Pr(y|x)$ and reports an obfuscated location y.
- Consider all locations are in a discrete set of only 4 possible locations: a hospital (x_1), a café (x_2), a forest (x_3), and a house (x_4).



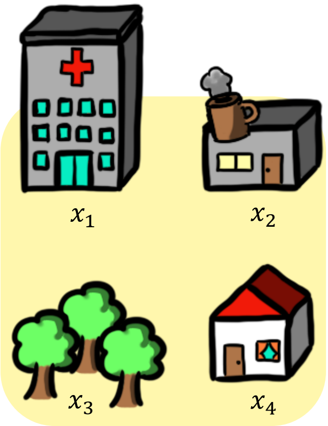
Pos.	Coordinates	$\Pr(x)$
x_1	(0,1)	0.2
x_2	(1,1)	0.4
x_3	(0,0)	0.1
x_4	(1,0)	0.3

$\Pr(y\|x)$	$y = x_1$	$y = x_2$	$y = x_3$	$y = x_4$
$x = x_1$	0.5	0.2	0.2	0.1
$x = x_2$	0.2	0.5	0.1	0.2
$x = x_3$	0.2	0.1	0.5	0.2
$x = x_4$	0.1	0.2	0.2	0.5

The mechanism

Statistical Inference example: location privacy

- Alice wants to query for a location-based service, without revealing her real location x to the service provider. She runs a randomized mechanism $\Pr(y|x)$ and reports an obfuscated location y.
- Consider all locations are in a discrete set of only 4 possible locations: a hospital (x_1), a café (x_2), a forest (x_3), and a house (x_4).



Pos.	Coordinates	$\Pr(x)$
x_1	(0,1)	0.2
x_2	(1,1)	0.4
x_3	(0,0)	0.1
x_4	(1,0)	0.3

$\Pr(y\|x)$	$y = x_1$	$y = x_2$	$y = x_3$	$y = x_4$
$x = x_1$	0.5	0.2	0.2	0.1
$x = x_2$	0.2	0.5	0.1	0.2
$x = x_3$	0.2	0.1	0.5	0.2
$x = x_4$	0.1	0.2	0.2	0.5

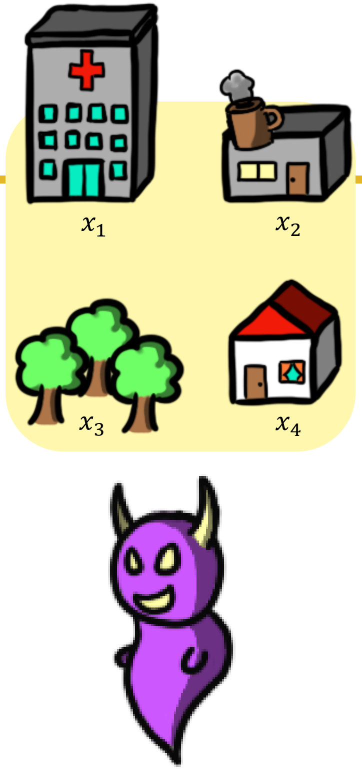
The mechanism

Alice reports that she is in the forest ($y = x_3$).
Q: What are the ML and MAP estimates of x?

Location privacy: solutions

Pos.	Coordinates	Pr(x)
x_1	(0,1)	0.2
x_2	(1,1)	0.4
x_3	(0,0)	0.1
x_4	(1,0)	0.3

$\Pr(y\|x)$	$y = x_1$	$y = x_2$	$y = x_3$	$y = x_4$
$x = x_1$	0.5	0.2	0.2	0.1
$x = x_2$	0.2	0.5	0.1	0.2
$x = x_3$	0.2	0.1	0.5	0.2
$x = x_4$	0.1	0.2	0.2	0.5

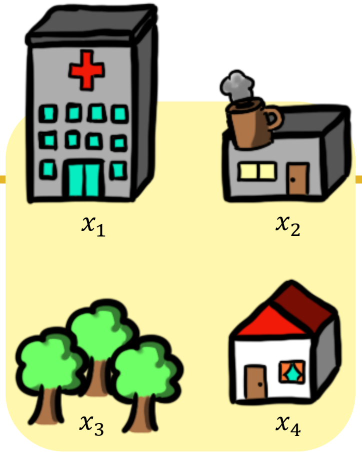


- ML: $\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(y|x) = x_3$

Location privacy: solutions

Pos.	Coordinates	Pr(x)
x_1	(0,1)	0.2
x_2	(1,1)	0.4
x_3	(0,0)	0.1
x_4	(1,0)	0.3

$\Pr(y\|x)$	$y = x_1$	$y = x_2$	$y = x_3$	$y = x_4$
$x = x_1$	0.5	0.2	0.2	0.1
$x = x_2$	0.2	0.5	0.1	0.2
$x = x_3$	0.2	0.1	$\boxed{0.5}$	0.2
$x = x_4$	0.1	0.2	0.2	0.5

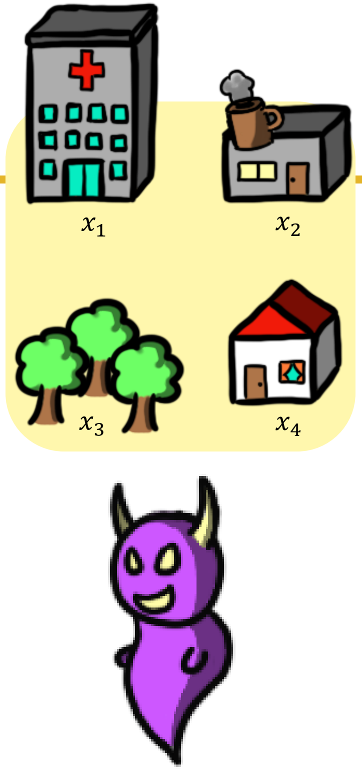


- ML: $\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(y|x) = x_3$

Location privacy: solutions

Pos.	Coordinates	Pr(x)
x_1	(0,1)	0.2
x_2	(1,1)	0.4
x_3	(0,0)	0.1
x_4	(1,0)	0.3

$\Pr(y\|x)$	$y = x_1$	$y = x_2$	$y = x_3$	$y = x_4$
$x = x_1$	0.5	0.2	0.2	0.1
$x = x_2$	0.2	0.5	0.1	0.2
$x = x_3$	0.2	0.1	0.5	0.2
$x = x_4$	0.1	0.2	0.2	0.5

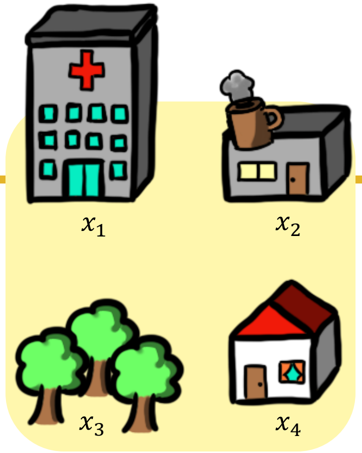


- ML: $\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(y|x) = x_3$
- MAP: $\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(x) \cdot \Pr(y|x) = x_4$

Location privacy: solutions

Pos.	Coordinates	Pr(x)
x_1	(0,1)	0.2
x_2	(1,1)	0.4
x_3	(0,0)	0.1
x_4	(1,0)	0.3

$\Pr(y\|x)$	$y = x_1$	$y = x_2$	$y = x_3$	$y = x_4$
$x = x_1$	0.5	0.2	0.2	0.1
$x = x_2$	0.2	0.5	0.1	0.2
$x = x_3$	0.2	0.1	0.5	0.2
$x = x_4$	0.1	0.2	0.2	0.5



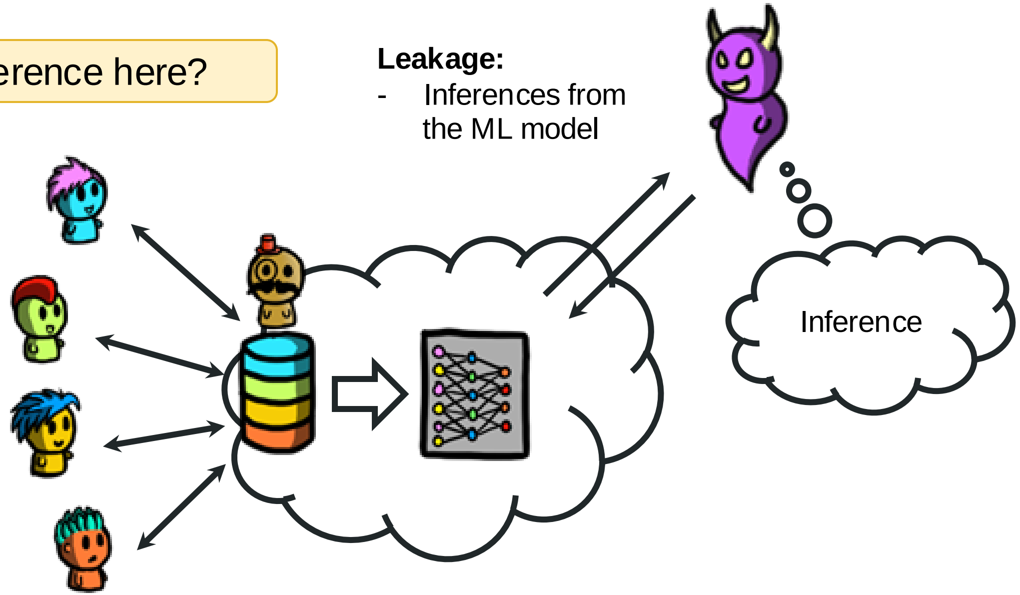
- ML: $\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(y|x) = x_3$
- MAP: $\hat{x} = \underset{x}{\operatorname{argmax}} \Pr(x) \cdot \Pr(y|x) = x_4$

5. Inference Attacks in Machine Learning

5. Inference attacks in Machine Learning

- There are many possible inference attacks in ML.
- For now, we will just think about the adversary **goals** and possible **techniques**

Q: We saw this before: what could be an inference here?

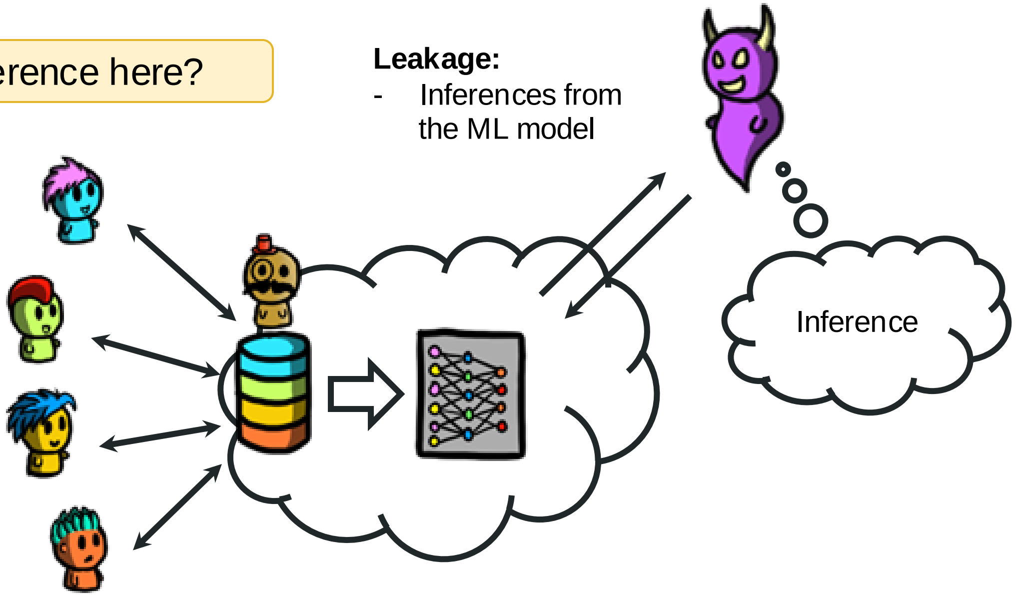


5. Inference attacks in Machine Learning

- There are many possible inference attacks in ML.
- For now, we will just think about the adversary **goals** and possible **techniques**

Q: We saw this before: what could be an inference here?

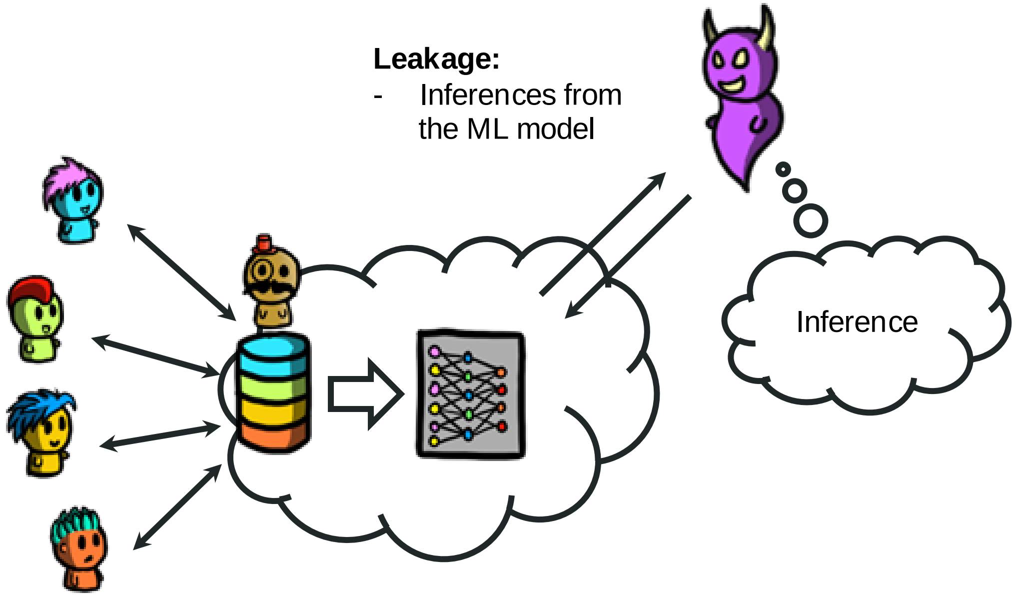
A:
- Membership inference
- Attribute inference (parts of a data sample)
- Property inference (property of the whole training set)
- Reconstruction attack (infer a whole training set)
- …



5. Inference attacks in Machine Learning

- There are many possible inference attacks in ML.
- For now, we will just think about the adversary **goals** and possible **techniques**

Q: If you were the adversary, which *techniques* would you use to run an attack in this scenario?

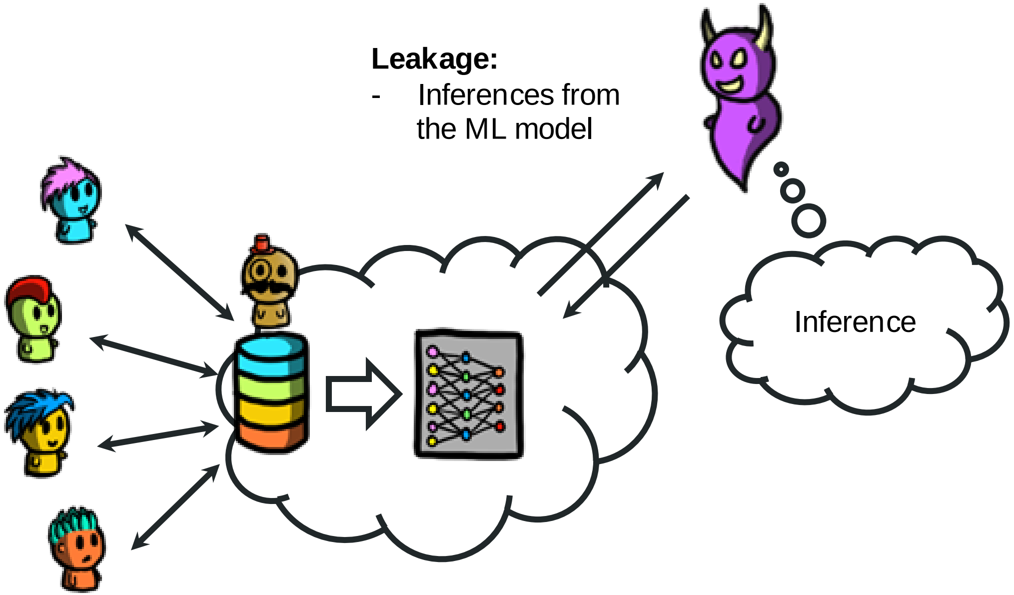


5. Inference attacks in Machine Learning

- There are many possible inference attacks in ML.
- For now, we will just think about the adversary **goals** and possible **techniques**

Q: If you were the adversary, which *techniques* would you use to run an attack in this scenario?

A: the idea is to use the fact that the model is more "confident" on samples it has trained on. We can use the confidence score, we can use thresholding techniques or train an ML model as an attack, etc.



6. Linking Attacks

6. Linking attacks

- As the name suggests, linking attacks find connections between two different sources of leakage that, alone, seem harmless.
- Famous example, from [1]:

For $20, you can purchase the voter registration list for Cambridge, Massachusetts

The Group Insurance Comission (GIC) in Massachusetts, sold data from 135,000 state employees to industry and researchers. They believed it was anonymous, so it was fine.

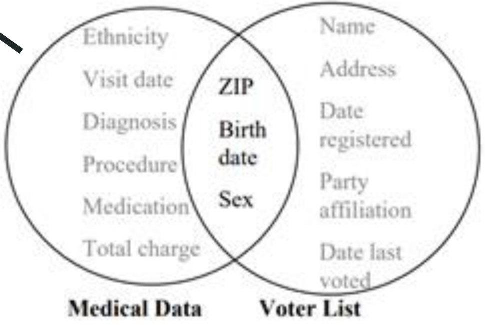


Figure 1 Linking to re-identify data

Fun fact: 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them **unique** based only on {5-digit ZIP, gender, date of birth}

[1] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002): 557-570.

Conclusion

- Inference attacks are one way of quantifying the leakage of a mechanism empirically
    - Need to be cautious as:
        - What if a better attack is developed later
        - What if the assumptions of the attacks do not represent real world threats

- Next we'll look at defences
    - More theoretical way to measure privacy
    - Usually a lower bound on privacy