Ethics in Natural Language Processing – SS 2022

TECHNISCHE UNIVERSITÄT DARMSTADT

Lecture 9
Privacy & Security II

Dr. Thomas Arnold
Aniket Pramanik

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Slides and material from Yulia Tsvetkov

Carnegie Mellon University
Language Technologies Institute

Syllabus (tentative)

Nr.	Lecture
01	Introduction, Foundations I
02	Foundations II
03	Bias I
04	Bias II
05	Incivility and Hate Speech I
06	NO LECTURE – Christi Himmelfahrt
07	Incivility and Hate Speech II
08	Low-Resource NLP
09	NO LECTURE - Fronleichnam
10	Privacy and Security I
11	Privacy and Security II
12	Language of Manipulation I
13	Language of Manipulation II

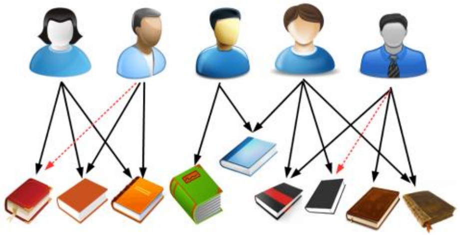
Outline

Recap

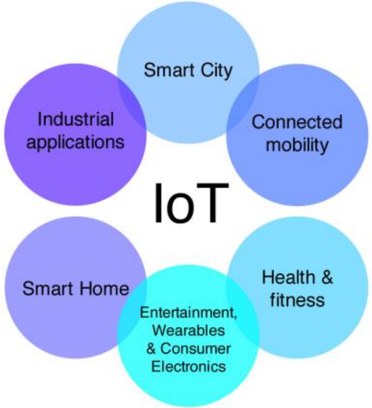
Anonymizing Data

Differential Privacy

Profiling



TSA Pre✓



What is Privacy

https://en.wikipedia.org/wiki/Privacy is the ability of an individual or group to seclude themselves or information about themselves, and thereby express themselves selectively

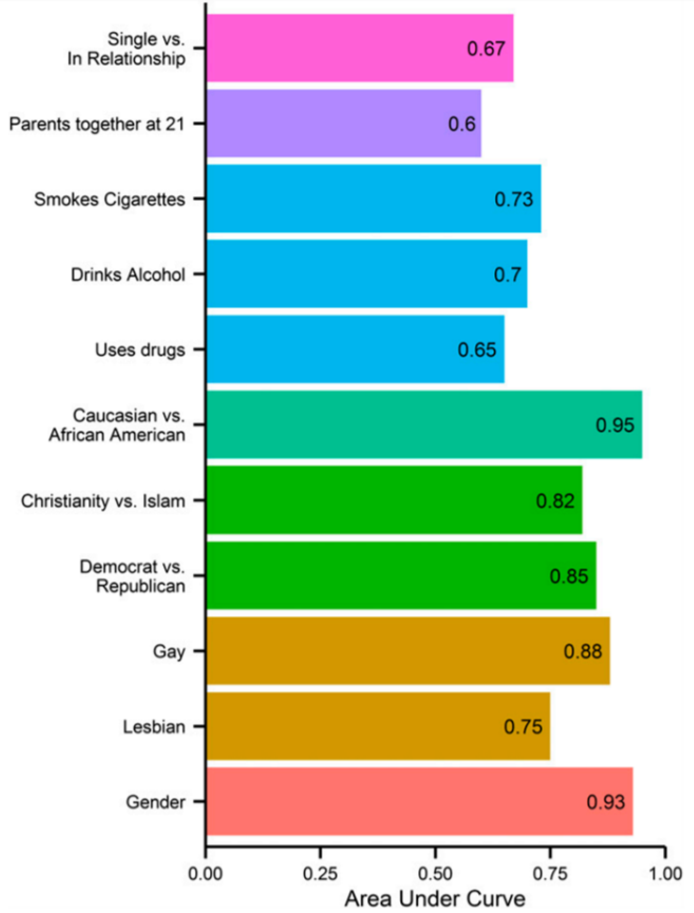


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

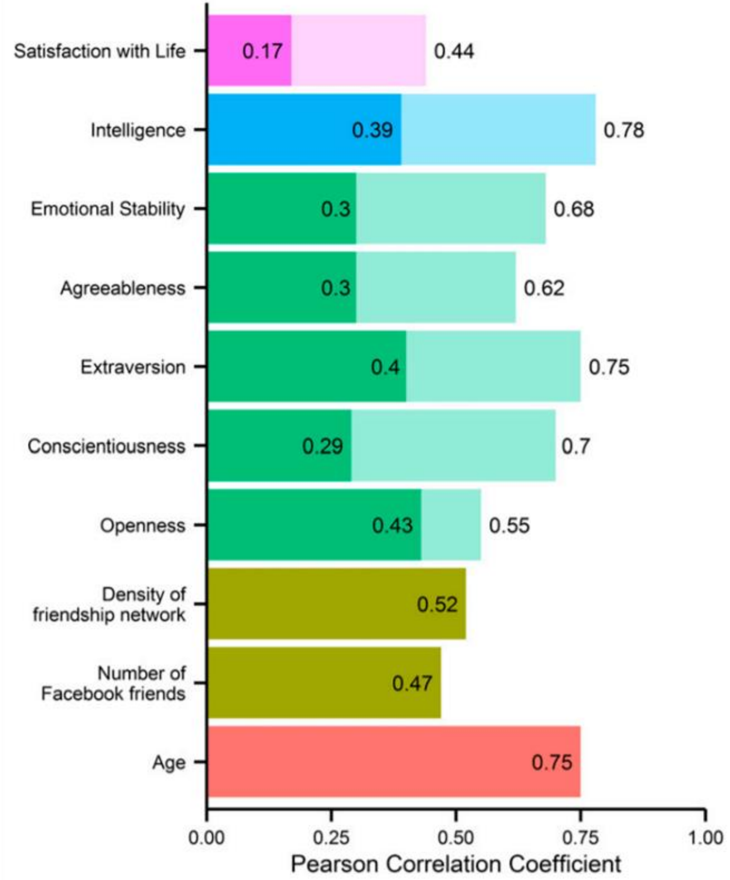


Fig. 3. Prediction accuracy of regression for numeric attributes and traits expressed by the Pearson correlation coefficient between predicted and actual attribute values; all correlations are significant at the $P < 0.001$ level. The

*58K people

Privacy vs. Utility

Image: Mostly AI

Learning Goals

TECHNISCHE UNIVERSITÄT DARMSTADT

After hearing this lecture, you should be able to…

- determine if a dataset has a certain level of k-anonymity, l-diversity or t-closeness
- describe possible attacks on tables with k-anonymity
- explain a simple form of differential privacy, and calculate true distributions from distributions with random noise

Outline

Recap

Anonymizing Data

Differential Privacy

Privacy and Anonymity

TECHNISCHE UNIVERSITÄT DARMSTADT

- Being on-line without giving up everything about you
- Ensuring collected data doesn’t reveal its users data
- Privacy in
  - Structured Data: k-anonymity, differential privacy
  - Text: obfuscating authorship
  - Speech: speaker id and de-identification

Companies Getting Your Data

- They actually don’t want your data, they want to upsell
  - They want to be able to do tasks (recommendations)
  - They actually don’t care about the individual you
- Can they process data to never have identifiable content?
  - Cumulated statistics
  - Averages, counts, for classes
- How many examples before it is anonymous?

k-anonymity

- Latanya Sweeney and Pierangela Samarati 1998
- Given some table for data with features and values
- Release data that guarantees individuals can’t be identifyied
  - **Suppresion:** Delete entries that are too “unique”
  - **Generalization:** relax specificness of fields,
    e.g. age to age-range or city to region

k-anonymity

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	29	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

- From wikipedia: K-anonymity

k-anonymity

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

- From wikipedia: K-anonymity

k-anonymity

A dataset has k-anonymity if the information for each person cannot be distinguished from at least k - 1 other individuals

- Optimal k-anonymity is an NP-hard problem

k-anonymity

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	Parking Violation
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Speeding
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

k-anonymity

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	Parking Violation
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Speeding
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

Personal attributes

Sensitive attribute(s)

k-anonymity

	Name	Age	Gender	City	Religion	Crime
	…	…	…	…	…	…
	*	31 – 40	Male	Griesheim	*	Parking Violation
3	*	31 – 40	Male	Griesheim	*	Murder
	*	31 – 40	Male	Griesheim	*	Speeding
	*	31 – 40	Male	Darmstadt	*	Speeding
2	*	31 – 40	Male	Darmstadt	*	Robbery
	…	…	…	…	…	…

Equivalence class: Entries that have the same personal attributes

k-anonymity

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	Parking Violation
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Speeding
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

3

2

If all **equivanlence classes** are at least size 2, this table has 2-anonymity

k-anonymity

A dataset has k-anonymity if the information for each person cannot be distinguished from at least k - 1 other individuals

- Optimal k-anonymity is an NP-hard problem
- Homogeneity Attack: All sensitive values within a set can be identical
- Background Knowledge Attack: Association between one or more quasi-identifier attributes with the sensitive attribute

Homogeneity Attack

TECHNISCHE UNIVERSITÄT DARMSTADT

I am male, 39 and live in Griesheim. What was my crime?

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

Homogeneity Attack

I am male, 39 and live in Griesheim. What was my crime?

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	**Murder**
*	31 – 40	Male	Griesheim	*	**Murder**
*	31 – 40	Male	Griesheim	*	**Murder**
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

Background Knowledge Attack

I am male, 39 and live in Griesheim. I always go by bike. What was my crime?

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	Parking Violation
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Speeding
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

Background Knowledge Attack

I am male, 39 and live in Griesheim. I always go by bike. What was my crime?

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	Parking Violation
*	31 – 40	Male	Griesheim	*	**Murder**
*	31 – 40	Male	Griesheim	*	Speeding
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

very unlikely

l-diversity

An equivalence class has l-diversity if there are at least l "well-represented" values for the sensitive attribute. A dataset has l-diversity if every equivalence class of the dataset has l-diversity.

l-diversity

An equivalence class has l-diversity if there are at least l "well-represented" values for the sensitive attribute. A dataset has l-diversity if every equivalence class of the dataset has l-diversity.

What means "well-represented" values?

- **Distinct l-diversity: At least l distinct values (simplest definition)**
- Entropy l-diversity: Calculates entropy of sensitive values (most complex)
- Recursive l-diversity: Compromise definition

(Distinct) l-diversity

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

(Distinct) l-diversity

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	**Murder**
*	31 – 40	Male	Griesheim	*	**Murder**
*	31 – 40	Male	Griesheim	*	**Murder**
*	31 – 40	Male	Darmstadt	*	**Speeding**
*	31 – 40	Male	Darmstadt	*	**Robbery**
…	…	…	…	…	…

?

?

How many distinct **sensible values** are in each eq. Class?

(Distinct) l-diversity

	Name	Age	Gender	City	Religion	Crime
	…	…	…	…	…	…
	*	31 – 40	Male	Griesheim	*	**Murder**
1	*	31 – 40	Male	Griesheim	*	**Murder**
	*	31 – 40	Male	Griesheim	*	**Murder**
2	*	31 – 40	Male	Darmstadt	*	**Speeding**
	*	31 – 40	Male	Darmstadt	*	**Robbery**
	…	…	…	…	…	…

This dataset has 1-diversity (lowest value)

(Distinct) l-diversity

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	Parking Violation
*	31 – 40	Male	Griesheim	*	Murder
*	31 – 40	Male	Griesheim	*	Speeding
*	31 – 40	Male	Darmstadt	*	Speeding
*	31 – 40	Male	Darmstadt	*	Robbery
…	…	…	…	…	…

(Distinct) l-diversity

	Name	Age	Gender	City	Religion	Crime
	…	…	…	…	…	…
	*	31 – 40	Male	Griesheim	*	**Parking Violation**
3	*	31 – 40	Male	Griesheim	*	**Murder**
	*	31 – 40	Male	Griesheim	*	**Speeding**
2	*	31 – 40	Male	Darmstadt	*	**Speeding**
	*	31 – 40	Male	Darmstadt	*	**Robbery**
	…	…	…	…	…	…

If all **equivanlence classes** have at least 2 distinct sensible values, this table has 2-diversity

t-closeness

An equivalence class has t-closeness if the distance between the distribution of the sensitive attribute in the class and the distribution of the attribute in the whole data set is no more than threshold t.

A dataset has t-closeness if every eq. class of the dataset has t-closeness.

t-closeness

An equivalence class has t-closeness if the distance between the distribution of the sensitive attribute in the class and the distribution of the attribute in the whole data set is no more than threshold t.

A dataset has t-closeness if every eq. class of the dataset has t-closeness.

t is a tradeoff between security and utility!

0.0-closeness: most secure, no utility

1.0-closeness: lowest security, highest utility

t-closeness

There are several ways to measure the distance between distributions. The easiest is the variational distance:

For two distributions $P = (p_1, p_2, \dots, p_m), Q = (q_1, q_2, \dots, q_m)$

$$D(P, Q) = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|$$

t-closeness

Example: The whole data set contains 4 sensitive attribute classes:

500 Parking Violation, 100 Murder, 200 Speeding, 200 Robbery

Distribution P = (0.5, 0.1, 0.2, 0.2)

t-closeness

Example: The whole data set contains 4 sensitive attribute classes:

500 Parking Violation, 100 Murder, 200 Speeding, 200 Robbery

Distribution P = (0.5, 0.1, 0.2, 0.2)

t-closeness

Name	Age	Gender	City	Religion	Crime
…	…	…	…	…	…
*	31 – 40	Male	Griesheim	*	**Parking Violation**
*	31 – 40	Male	Griesheim	*	**Murder**
*	31 – 40	Male	Griesheim	*	**Speeding**
*	31 – 40	Male	Darmstadt	*	**Speeding**
*	31 – 40	Male	Darmstadt	*	**Robbery**
…	…	…	…	…	…

Distribution Q = (0, 0, 0.5, 0.5)

t-closeness

Example: The whole data set contains 4 sensitive attribute classes:

500 Parking Violation, 100 Murder, 200 Speeding, 200 Robbery

Distribution P = (0.5, 0.1, 0.2, 0.2) Distribution Q = (0, 0, 0.5, 0.5)

$$D(P,Q) = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|$$

t-closeness

Example: The whole data set contains 4 sensitive attribute classes:

500 Parking Violation, 100 Murder, 200 Speeding, 200 Robbery

Distribution P = (0.5, 0.1, 0.2, 0.2) Distribution Q = (0, 0, 0.5, 0.5)

$$D(P,Q) = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|$$

= 0.5 * (|0.5 – 0| + |0.1 – 0| + |0.2 – 0.5| + |0.2 - 0.5|)

t-closeness

Example: The whole data set contains 4 sensitive attribute classes:

500 Parking Violation, 100 Murder, 200 Speeding, 200 Robbery

Distribution P = (0.5, 0.1, 0.2, 0.2) Distribution Q = (0, 0, 0.5, 0.5)

$$D(P,Q) = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|$$

= 0.5 * (|0.5 – 0| + |0.1 – 0| + |0.2 – 0.5| + |0.2 - 0.5|)

= 0.5 * (0.5 + 0.1 + 0.3 + 0.3)

t-closeness

Example: The whole data set contains 4 sensitive attribute classes:

500 Parking Violation, 100 Murder, 200 Speeding, 200 Robbery

Distribution P = (0.5, 0.1, 0.2, 0.2) Distribution Q = (0, 0, 0.5, 0.5)

$$D(P,Q) = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|$$

= 0.5 * (|0.5 – 0| + |0.1 – 0| + |0.2 – 0.5| + |0.2 - 0.5|)

= 0.5 * (0.5 + 0.1 + 0.3 + 0.3)

= 0.5 * 1.2 = **0.6**

t-closeness

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: The whole data set contains 4 sensitive attribute classes:

500 Parking Violation, 100 Murder, 200 Speeding, 200 Robbery

Distribution P = (0.5, 0.1, 0.2, 0.2)

Distribution Q = (0, 0, 0.5, 0.5)

$$D(P,Q) = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|$$

= 0.5 * (|0.5 – 0| + |0.1 – 0| + |0.2 – 0.5| + |0.2 - 0.5|)

= 0.5 * (0.5 + 0.1 + 0.3 + 0.3)

= 0.5 * 1.2 = **0.6**

This eq. class has 0.6 closeness (and higher)

Take-Home Message

- k-anonymity provides some anonymity, but can be vulnerable to certain weaknesses (Homogeneity, Background Knowledge)

- l-diversity improves k-anonymity by adding contraints to the diversity of the sensitive values

- t-closeness compares the distribution of sensitive values to the overall distribution (no explicit statements about eq. class size)

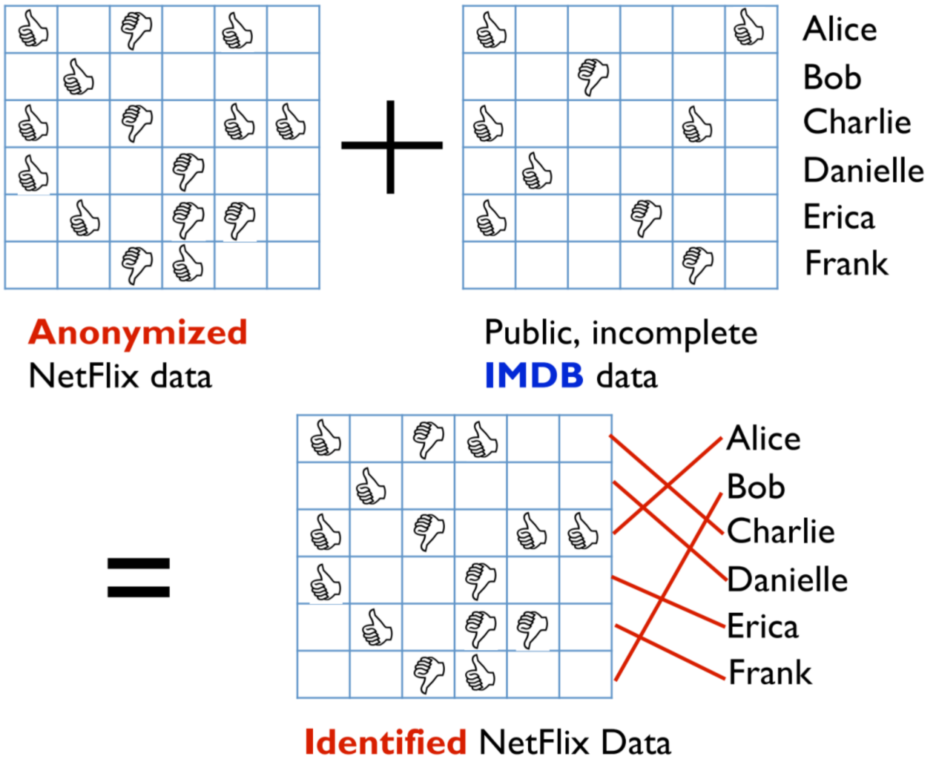
Outline

Recap

Anonymizing Data

Differential Privacy

Linkage Attacks Are Still Possible

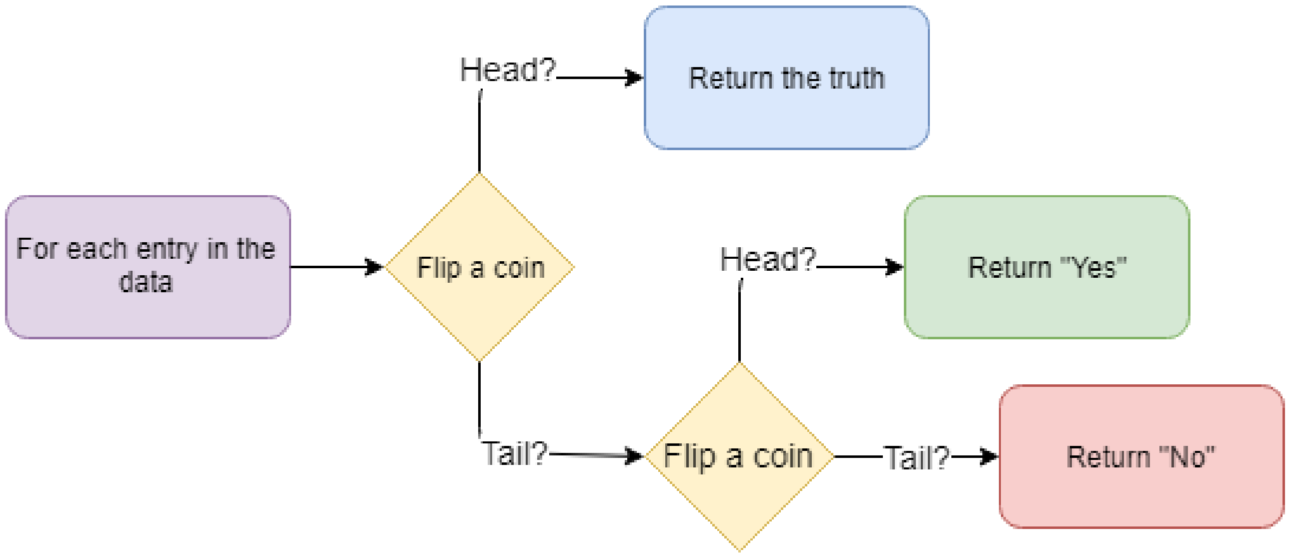


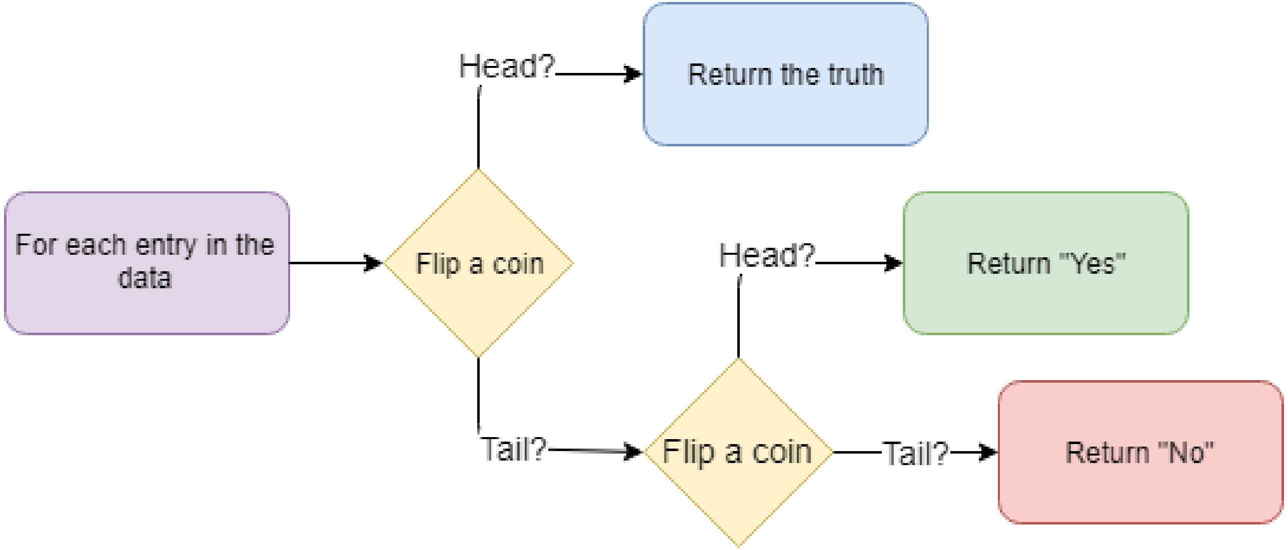
Is there a better way to hide identification?

Differential Privacy

Basic idea: Introduce randomness

- Coin Toss Example: When asked about feature x for record y
  - Toss a coin: if heads give right answer
  - If tails: throw coin again, answer yes if heads, no if tails
- Still has accuracy at some level of confidence
- Still has privacy at some level of confidence (plausible deniability)


For each entry in the data
Flip a coin
Head?
Return the truth
Tail?
Flip a coin
Head?
Return "Yes"
Tail?
Return "No"



First coin toss: Privacy parameter

Always heads: No privacy, perfect accuracy

Always tails: Perfect privacy, no accuracy

Differential Privacy

The distribution of attributes can still be estimated

If person X has attribute A, then

$P(A|X) = 0.75$

$P(\sim A|X) = 0.25$

If p is the true proportion of people with attribute A, then we expect

$(1 / 4) + p/2$ positive responses

Differential Privacy

Example: 700 people say they like ice-cream, 300 say they do not like ice-cream

What is the **estimated true distribution p** of people that like ice-cream?

P(Likes ice-cream) =

P(Coin = tails) * P("I like ice-cream" | Coin = tails)

+

P(Coin = heads) * P("I like ice-cream | Coin = heads)

Differential Privacy

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: 700 people say they like ice-cream, 300 say they do not like ice-cream

What is the **estimated true distribution p** of people that like ice-cream?

P(Likes ice-cream) =

P(Coin = tails) * P("I like ice-cream" | Coin = tails)

+

P(Coin = heads) * P("I like ice-cream | Coin = heads)

p = true ice-cream lovers

Differential Privacy

Example: 700 people say they like ice-cream, 300 say they do not like ice-cream

What is the **estimated true distribution p** of people that like ice-cream?

P(Likes ice-cream) =

P(Coin = tails) * P("I like ice-cream" | Coin = tails)

+

P(Coin = heads) * P("I like ice-cream | Coin = heads)

0.5 (fair coin toss)

p = true ice-cream lovers

Differential Privacy

Example: 700 people say they like ice-cream, 300 say they do not like ice-cream

What is the **estimated true distribution p** of people that like ice-cream?

P(Likes ice-cream) =

P(Coin = tails) * P("I like ice-cream" | Coin = tails)

+

P(Coin = heads) * P("I like ice-cream | Coin = heads)

0.5 (decided by another coin toss)

0.5 (fair coin toss)

p = true ice-cream lovers

Differential Privacy

Example: 700 people say they like ice-cream, 300 say they do not like ice-cream

What is the **estimated true distribution p** of people that like ice-cream?

0.7 (700 out of 1000)

P(Likes ice-cream) =

P(Coin = tails) * P("I like ice-cream" | Coin = tails)

+

P(Coin = heads) * P("I like ice-cream | Coin = heads)

0.5 (decided by another coin toss)

0.5 (fair coin toss)

p = true ice-cream lovers

Differential Privacy

Example: 700 people say they like ice-cream, 300 say they do not like ice-cream

What is the **estimated true distribution p** of people that like ice-cream?

$$0.7 = 0.5 * 0.5 + 0.5 * p$$

Differential Privacy

Example: 700 people say they like ice-cream, 300 say they do not like ice-cream

What is the **estimated true distribution p** of people that like ice-cream?

0.7 = 0.5 * 0.5 + 0.5 * p

0.7 = 0.25 + 0.5 * p

0.45 = 0.5 * p

P = 0.9

Differential Privacy

Example: 700 people say they like ice-cream, 300 say they do not like ice-cream

What is the **estimated true distribution p** of people that like ice-cream?

0.7 = 0.5 * 0.5 + 0.5 * p

0.7 = 0.25 + 0.5 * p

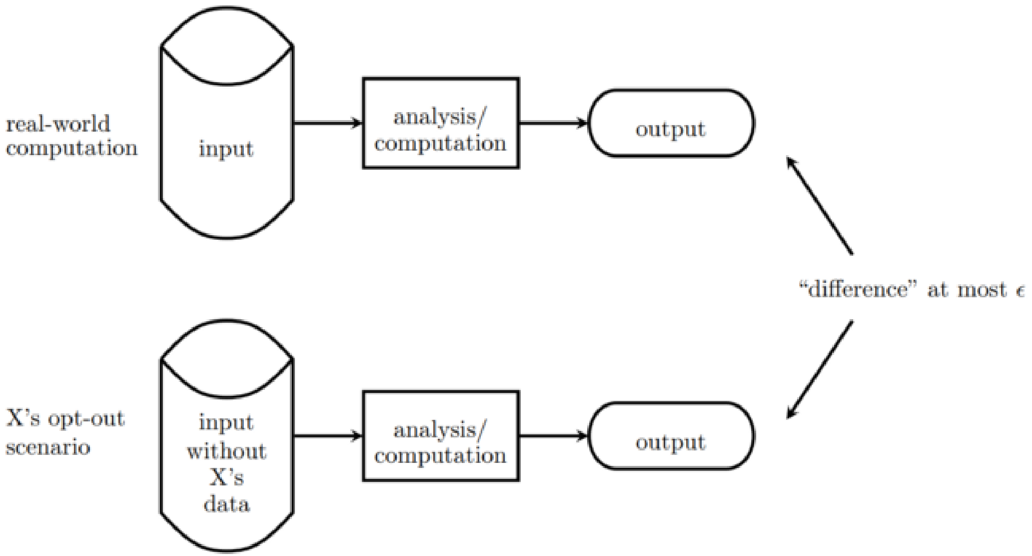
0.45 = 0.5 * p

P = 0.9 <u>Approximately 90 % of the people liked ice-cream</u>

Differential Privacy

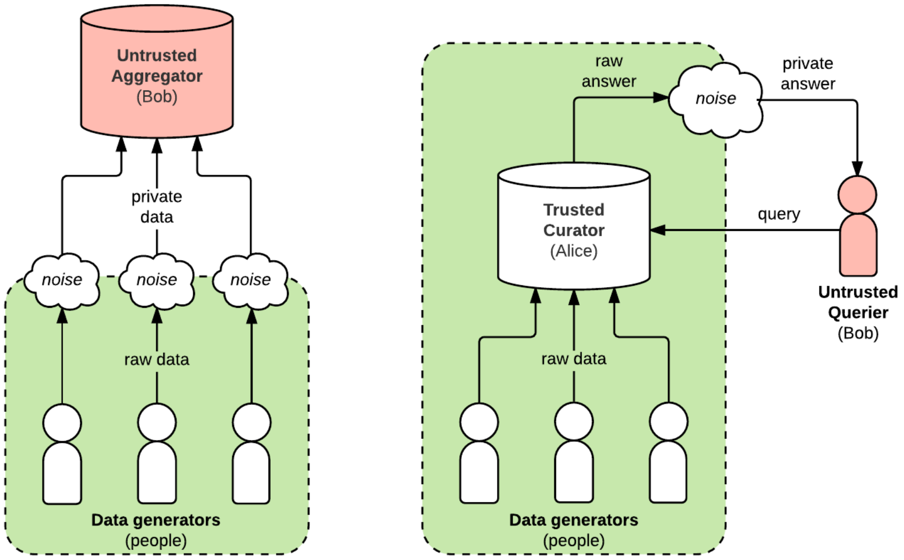
Coin toss is a very simplified version of Differential Privacy

Main idea:



Differential Privacy

Either submit data with noise (like the coin toss) or add global noise



Local privacy   Global privacy

Next Lecture

Language of Manipulation I

Now

TECHNISCHE UNIVERSITÄT DARMSTADT

Lecture Evaluation

The link is in Moodle

Thank you for your feedback!