Human-Centered Computing
Research Group

Freie Universität Berlin

«Human-Centered Data Science»

Privacy - Protecting Individuals' Sensitive Information in Data

Dr. Daniel Franzen & Prof. Dr. Claudia Müller-Birn

Human-Centered Computing (HCC), Institute of Computer Science

Freie Universität Berlin

14 July 2022

- Dr. Daniel Franzen
- PostDoc at HCC
- FreeMove: Data Donation of movement data
  - Communication of privacy guarantees
  - Goal: Informed decision
  - Crowd-sourcing studies

Outline

- What is privacy and why is it important?
- Privacy protection
    - Privacy Guarantees and Properties
    - Algorithms
- Privacy considerations

What is Privacy?

What is Privacy? - Personally identifiable information (PII)

> **any information relating to an identified or identifiable natural person ...**
> in particular by reference to an identifier such as a
> **name, an identification number, location data, an online identifier ...**

Article 4 of the GDPR

- **Phone-book**
- **Credit Card information**
- **Netflix watch history**

- **Traffic information**
- **Webpage visitor data**

- **Position of charging stations**
- **Weather data**

What is Privacy? - Sensitive Information

The following personal data is considered ‘sensitive’ and is subject to specific processing conditions:

- personal data revealing **racial** or **ethnic** origin, **political** opinions, **religious** or **philosophical** beliefs;
- trade-union **membership**;
- genetic data, **biometric data** processed solely to identify a human being;
- **health**-related data;
- data concerning a person’s **sex life or sexual orientation**.

Article 4(13), (14) and (15) and Article 9 and Recitals (51) to (56) of the GDPR

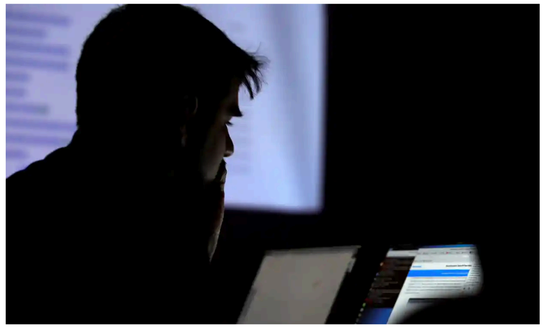
Solution: Anonymization

- Proposal: Just delete names, credit card IDs, … from the data set

  → Data set is not personally identifiable anymore

'Anonymous' browsing data can be easily exposed, researchers reveal

A journalist and a data scientist secured data from three million users easily by creating a fake marketing company, and were able to de-anonymise many users



'We wrote and called nearly a hundred companies, and asked if we could have the raw data, the clickstream from people's lives.' Photograph: Steve Marcus/Reuters

A judge's porn preferences and the medication used by a German MP were among the personal data uncovered by two German researchers who acquired the "anonymous" browsing habits of more than three million German citizens.

"What would you think," asked Svea Eckert, "if somebody showed up at your door saying: 'Hey, I have your complete browsing history - every day, every hour, every minute, every click you did on the web for the last month'? How would you think we got it: some shady hacker? No. It was much easier: you can just buy it."

Conferences > 2008 IEEE Symposium on Securi...

Robust De-anonymization of Large Sparse Datasets

Publisher: IEEE Cite This PDF

Arvind Narayanan ; Vitaly Shmatikov All Authors

905	15	7206
Paper Citations	Patent Citations	Full Text Views

Abstract

Document Sections

1 Introduction

2 Related work

3 Model

4 De-anonymization algorithm

5 Case study: Netflix Prize dataset

Show Full Outline

Authors

Figures

Abstract:
We present a new class of statistical de- anonymization preferences, recommendations, transaction records and tolerate some mistakes in the adversary's background k Prize dataset, which contains anonymous movie ratings rental service. We demonstrate that an adversary who k this subscriber's record in the dataset. Using the Interne successfully identified the Netflix records of known user sensitive information.

Published in: 2008 IEEE Symposium on Security and P

Date of Conference: 18-22 May 2008

Date Added to IEEE *Xplore*: 28 May 2008

Print ISBN:978-0-7695-3168-7

Previous Next

Table of Contents

Related Articles

Home / Proceedings / DSAA / DSAA 2016

2016 IEEE 3rd International Conference on Data Science and Advanced Analytics (DSAA)

Anonymizing NYC Taxi Data: Does It Matter?

Year: 2016, Pages: 140-148
DOI Bookmark: 10.1109/DSAA.2016.21

Authors

Marie Douriez
Harish Doraiswamy
Juliana Freire
Cláudio T. Silva

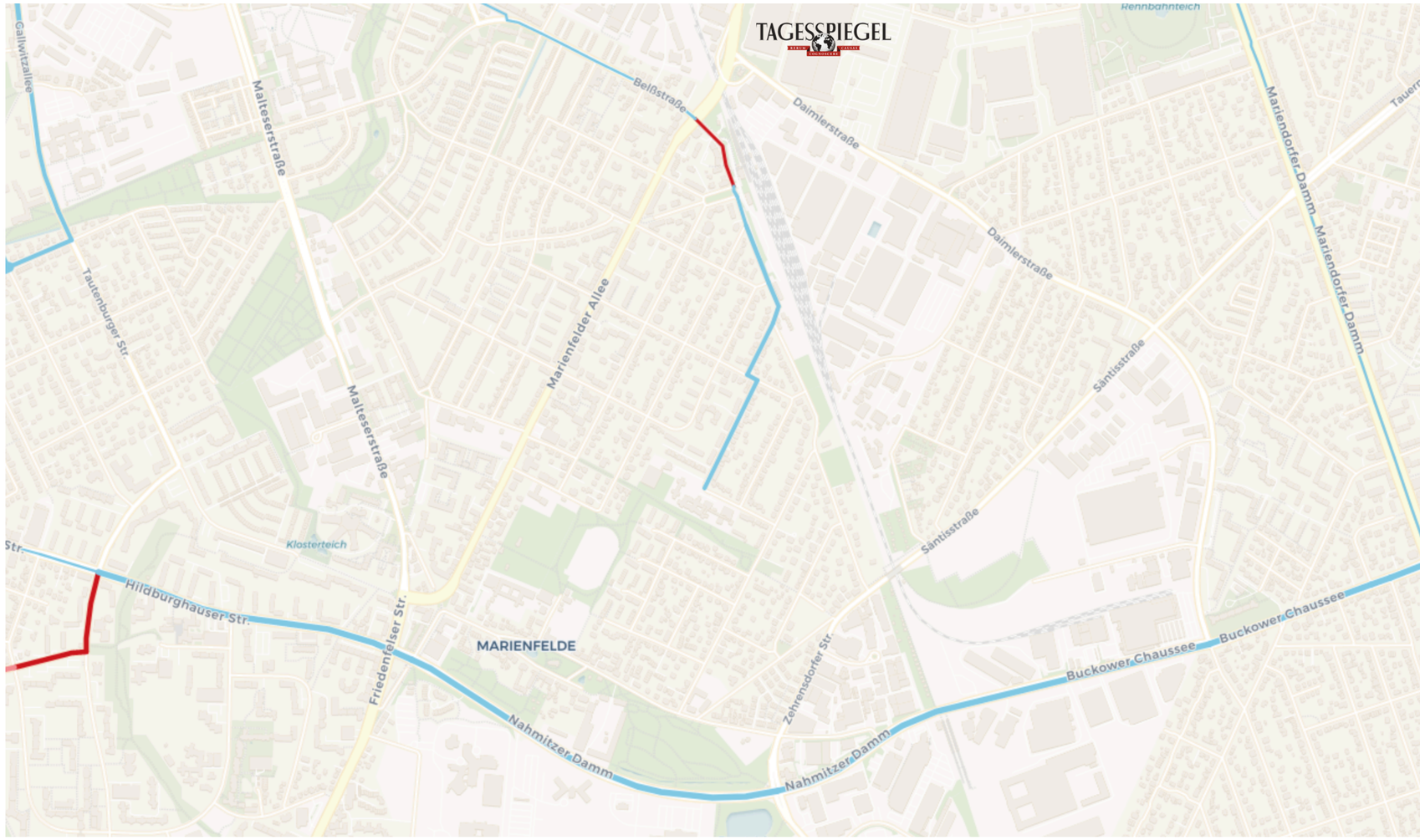
DOWNLOAD PDF SHARE ARTICLE GENERATE CITATION

Abstract

The widespread use of location-based services has led to an increasing availability of trajectory data from urban environments. These data carry rich information that are useful for improving cities through traffic management and city planning. Yet, it also contains information about individuals which can jeopardize their privacy. In this study, we work with the New York City (NYC) taxi trips data set publicly released by the Taxi and Limousine Commission (TLC). This data set contains information about every taxi cab ride that happened in NYC. A bad hashing of the medallion numbers (the ID corresponding to a taxi) allowed the

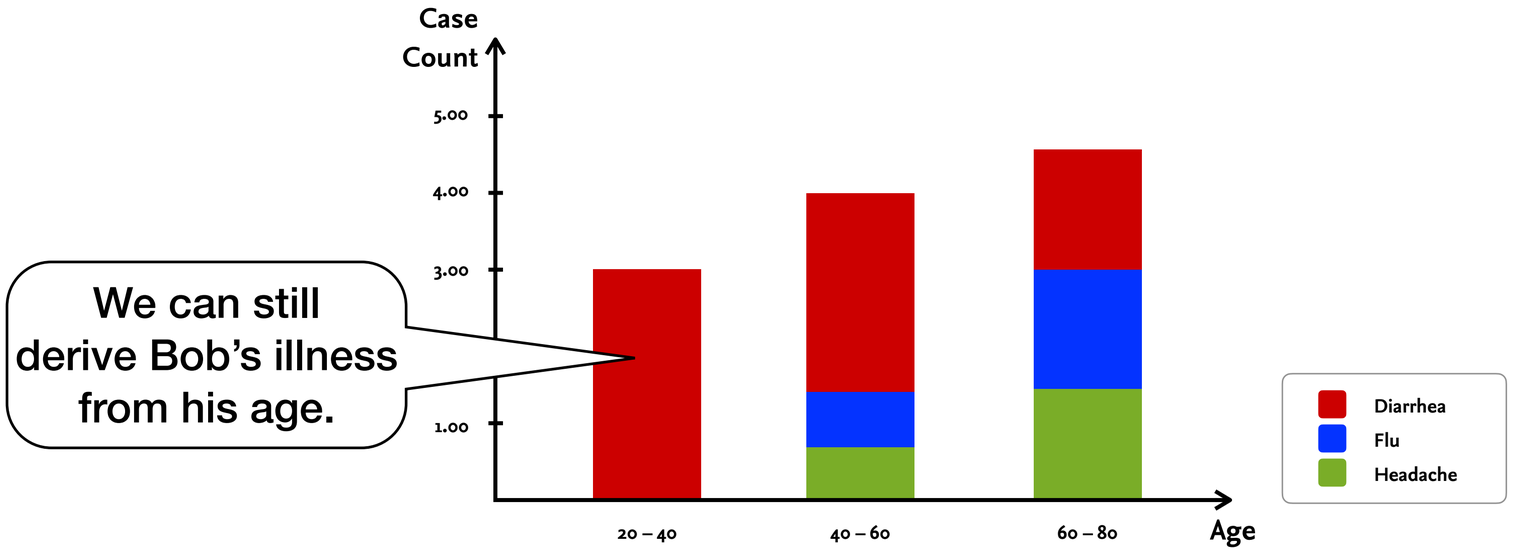
Example: Radmesser - Biking data

- Study:
  Identify dangerous streets for bikes
- Results anonymized and visualized in interactive map
- Newspaper also published interview with picture, name and „living in Marienfelde“



Privacy issues - Aggregation

- Proposal: Publish only aggregated data
  - → no single data entries are released



Why is privacy important?

- Personal
  - Loss of social standing
  - Loss of job / job-opportunities
- Companies
  - personalized advertising
  - increase of insurance fees
- Government
  - increase in traffic fees
  - mass surveillance (presumption of innocence)
- Criminals
  - Blackmailing
  - Identity theft
  - Credit-card fraud
  - Social engineering

Privacy issues and Contexts

Two different kinds of privacy disclosures

- Identity disclosure: Bob left the doctor's appointment **13:30**
- Attribute disclosure: Bob lives in **12345 Berlin**

Time	Billing-Adress	Treatment-reason
11:15	Bahnhofsweg 12, 12345 Berlin	Diarrhea
12:45	Chauseestraße 23, 13456 Berlin	Flu
14:10	Hauptstraße 34, 12345 Berlin	Diarrhea
15:30	Lange Straße 45,14567 Berlin	Headache

Necessity for Privacy

If any sensitive information might be involved

→ Think about privacy, even if anonymized

Privacy Properties & Guarantees
a) deterministic

Definitions

Quasi-Identifier:
Set of attributes whose values when taken together can potentially identify an individual.
Example: {Zip-code, Birthday, Gender}

Sensitive Attribute:
Attribute containing the sensitive information
Example: Disease

Discovery:
Correlation Quasi-Identifier → Sensitive Attribute

Li, Ninghui, et al. 'T-Closeness: Privacy Beyond k-Anonymity and l-Diversity'. *2007.*

Privacy property: k-anonymity, k-anonymous

> For each record there are at least k entries with the same values for the quasi-identifiers

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

- ⊕ protects against Identity disclosure
- ⊕ Easy to understand
- ⊖ Weak protection against attribute disclosure

Li, Ninghui, et al. 'T-Closeness: Privacy Beyond k-Anonymity and l-Diversity'. *2007.*

Privacy property: l-diversity, l-diverse

> For each record there are at least l different values for the sensitive attribute with the same values for the quasi-identifiers

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

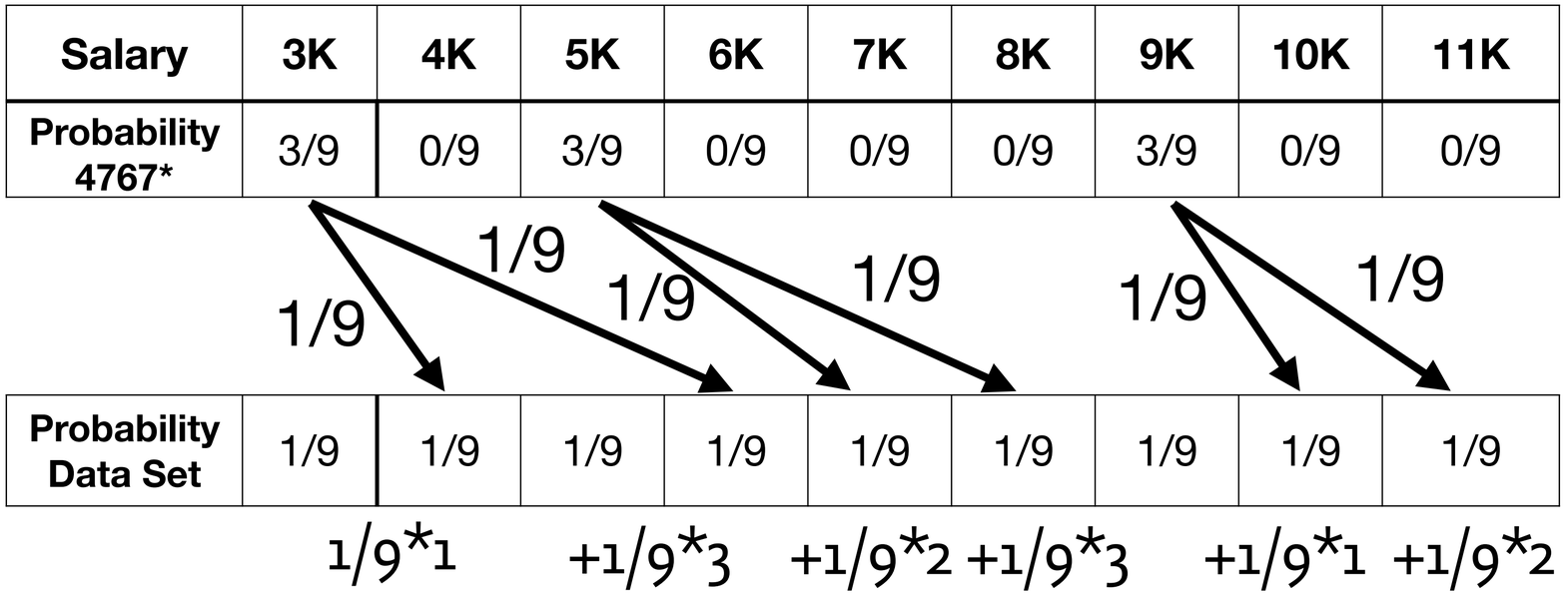
- ➕ Protects against identity and attribute disclosure
- ➖ Harder to achieve
- ➖ Weak protection against **semantic closeness** or **probabilisitic attribute disclosure**

Li, Ninghui, et al. 'T-Closeness: Privacy Beyond k-Anonymity and l-Diversity'. *2007.*

Distance between two distributions: Earthmover

- Find an optimal flow from one distribution to the other
- Distance: weight x distance

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer



Privacy property: t-closeness

> The distance between the distribution of a sensitive attribute with a given combination of quasi-identifiers and the distribution of the attribute in the whole table is no more than t.

- ➕ Protects against **probabilistic attribute disclosure**
- ➖ Even Harder to achieve, domain specific
- ➖ Weak protection against **attribute disclosure**

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Li, Ninghui, et al. 'T-Closeness: Privacy Beyond k-Anonymity and l-Diversity'. *2007.*

Privacy Protection Mechanisms
a) Deterministic

Idea: Make data less sensitive

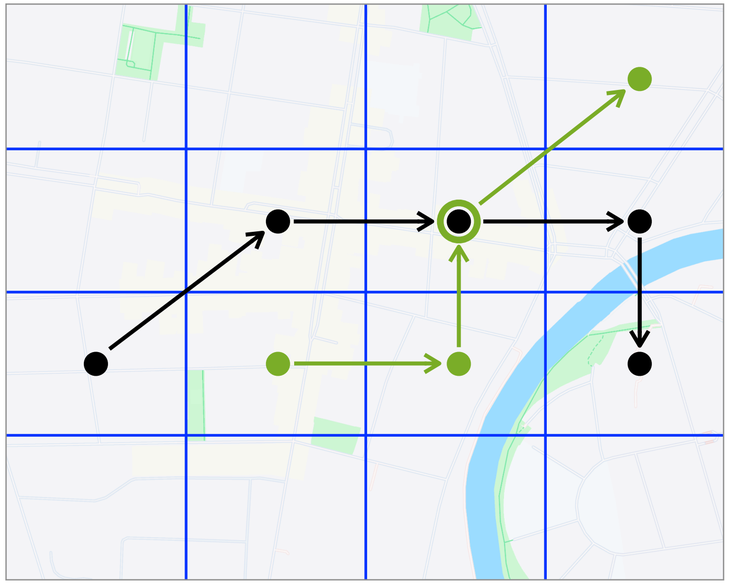
- How to make data less sensitive?

Generalization

Outlier Removal

Noice

Generalization

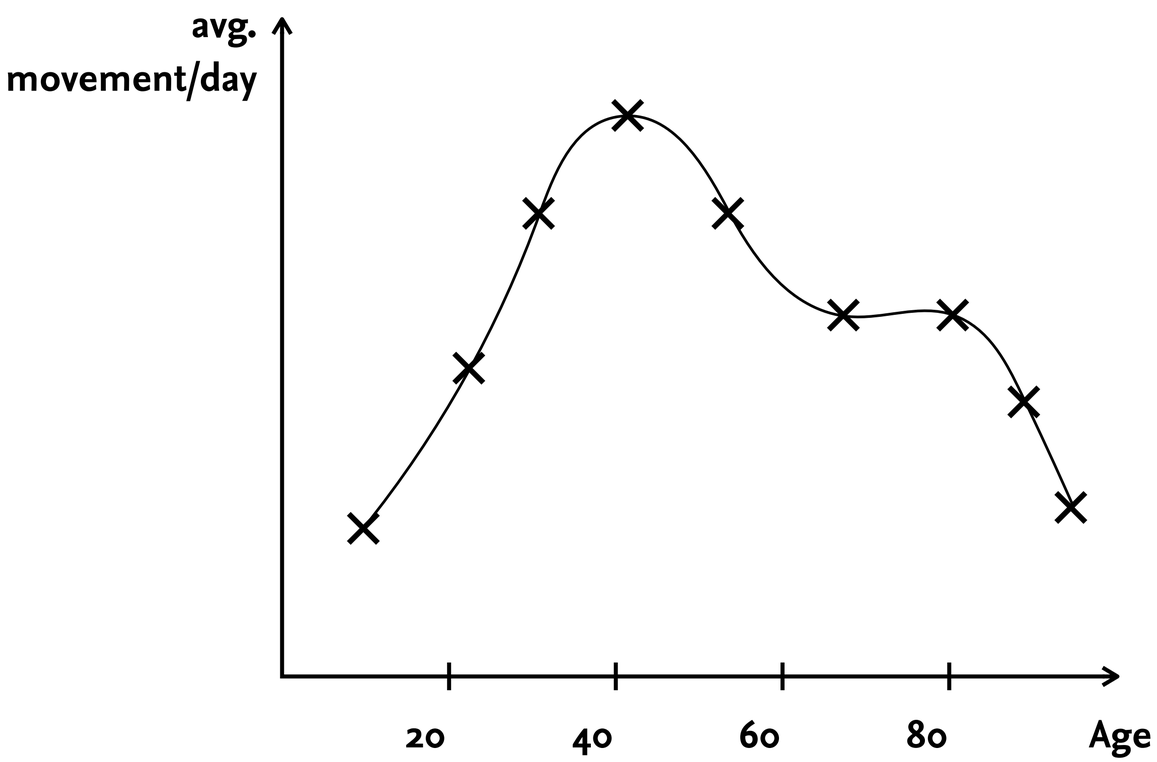


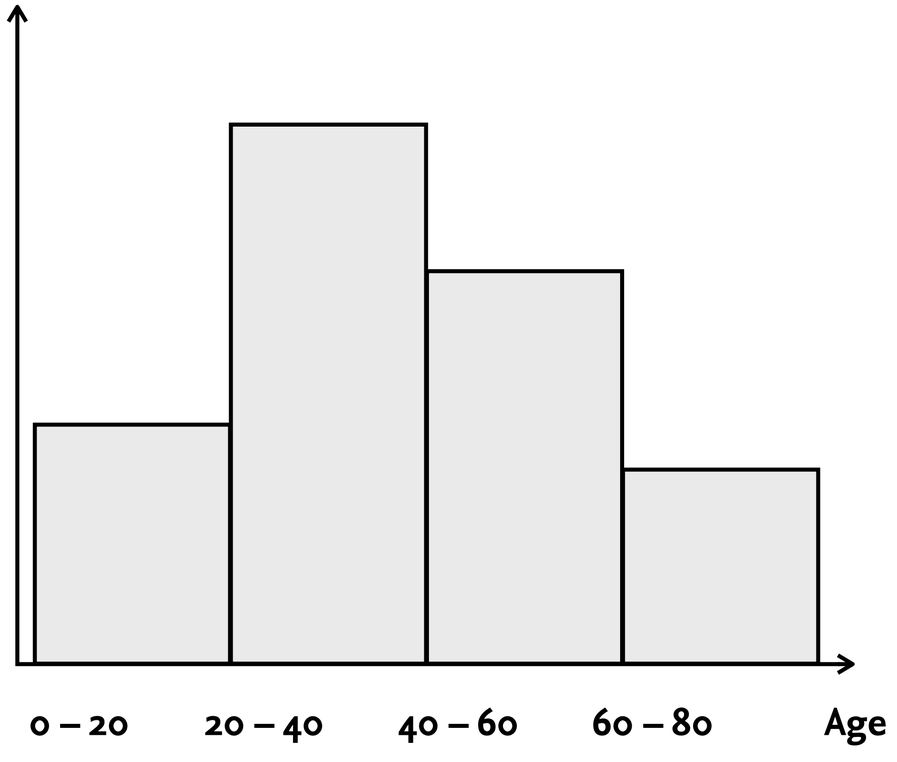
Generalization - decrease granularity

- Reduce location data from precise points to areas → tessellation
- Only show movement from one area to another
- Area could be defined by
  - Grid of predefined size
  - Natural areas
    - Zip-code, district, country
    - Closest public transport station

Generalization - change Visualization

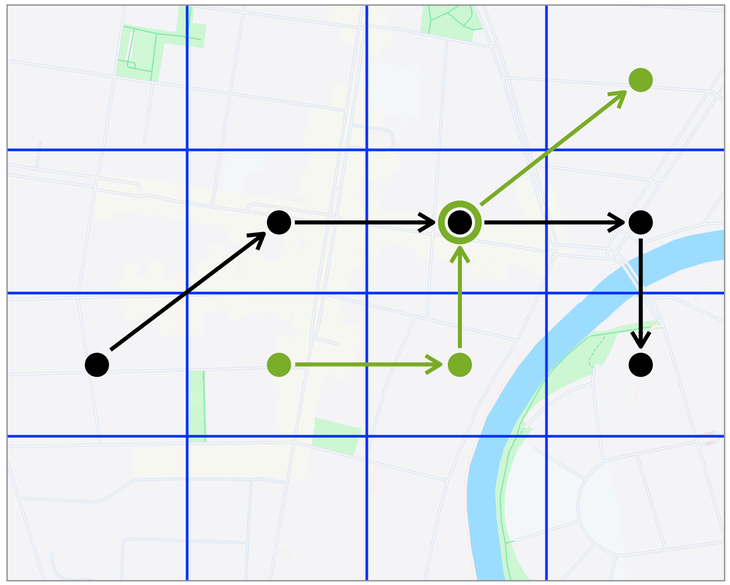
- Generalizing might benefit different visualization types:

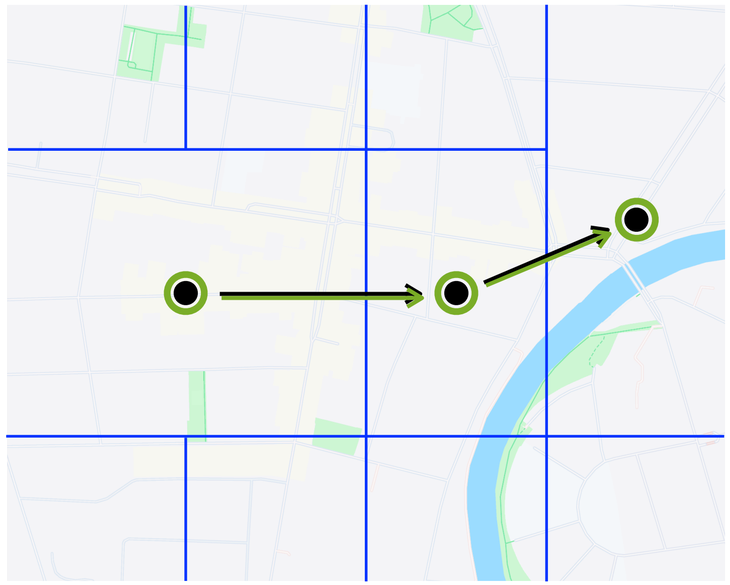




Generalization - dynamic

- Problem: If only one person in one area → potential identity disclosure
  - Coarser generalization → decreases overall accuracy unnecessarily
- Solution: Combine only areas with too few participants





Generalization - general Categories

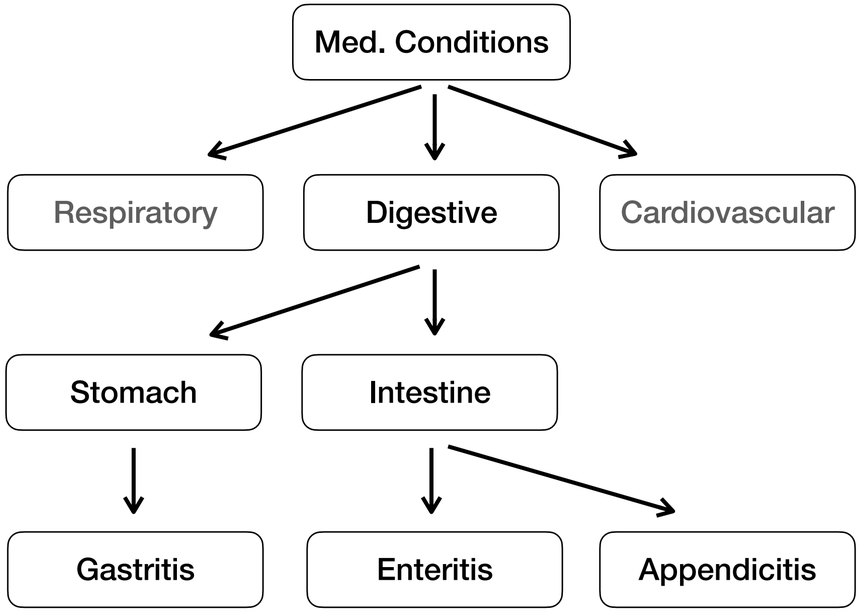
Ordered attributes

According to order

Attribute	Static	Dynamic
Age	[20-39], [40-59], [60-79]	[<40], [4*], [>=50]
Names	A., B., C.,	Al., B., Chr.

Unordered attributes

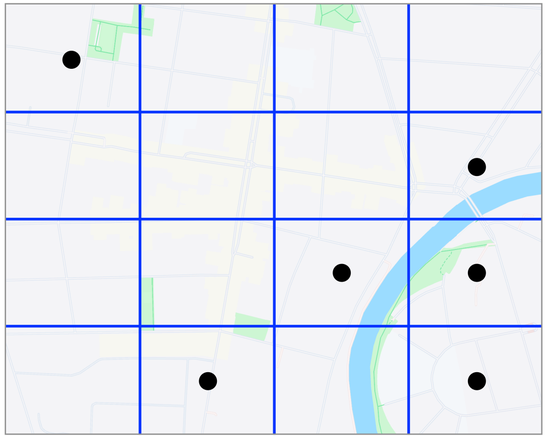
Taxonomy-Tree

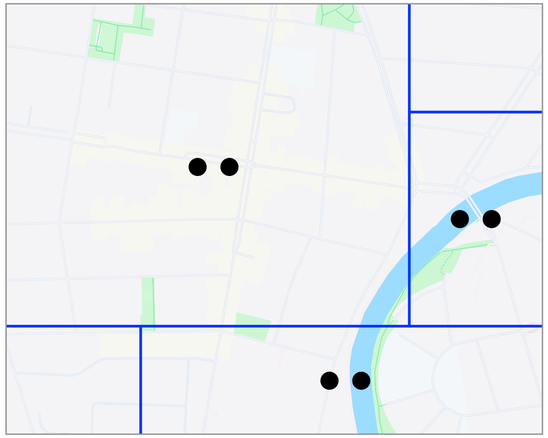


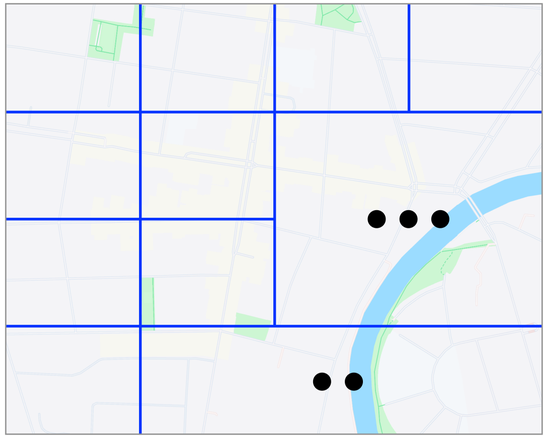
Outlier removal

- Problem: Generalization not feasible for extreme outliers
- Solution: Remove extreme outliers, for better utility

  → Preserves properties of the dataset







Generalization + Outlier removal = k-Anonymity

- Idea:
  - Choose k, t or l before collection
  - Privacy guarantee
  - Use Generalization + Outlier removal
- Goal: Find combination with Privacy guarantee with highest utility
- BUT: Ideal combination of generalization + outlier removal not known and dependent on data

  → Some intuition needed.

Privacy Properties & Guarantees
b) Probabilistic

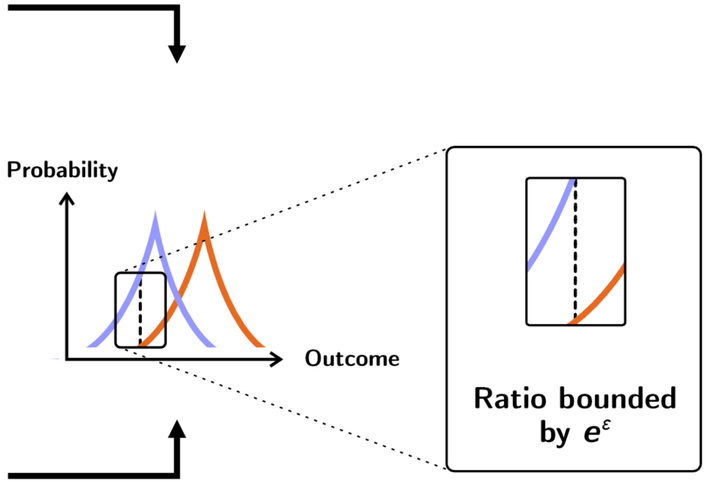
Differential Privacy: Idea

- Worst case: Mallory knows everything about everybody but Bob
  - Computes expected result of analysis
  - Compare to actual result -> Difference is caused by Bob

- Gold standard: Analysis with Bob = Analysis without Bob
  - But: anyone is protected like Bob
  - -> same results can also be obtained **without data**

ε-Differential Privacy

- Instead: **Plausible deniability** with a certain probability
  - M: „I learned something about you by looking at the data“
  - B: „That cannot be, I **was not involved** in the data“
  - M: „Sure, **how probable** is that?“
- Method
  - Compute result randomly (close to actual result)
  - Compare distribution with/without Bob (for all Bobs, for all results)
    - In the worst case difference it should be „less than ε“



Differential Privacy vs. t-closeness

- t-closeness: compares distribution of group to whole data set
  - Data-subjects are as protected as anyone **in the data set**
  - Highly dependent on data set

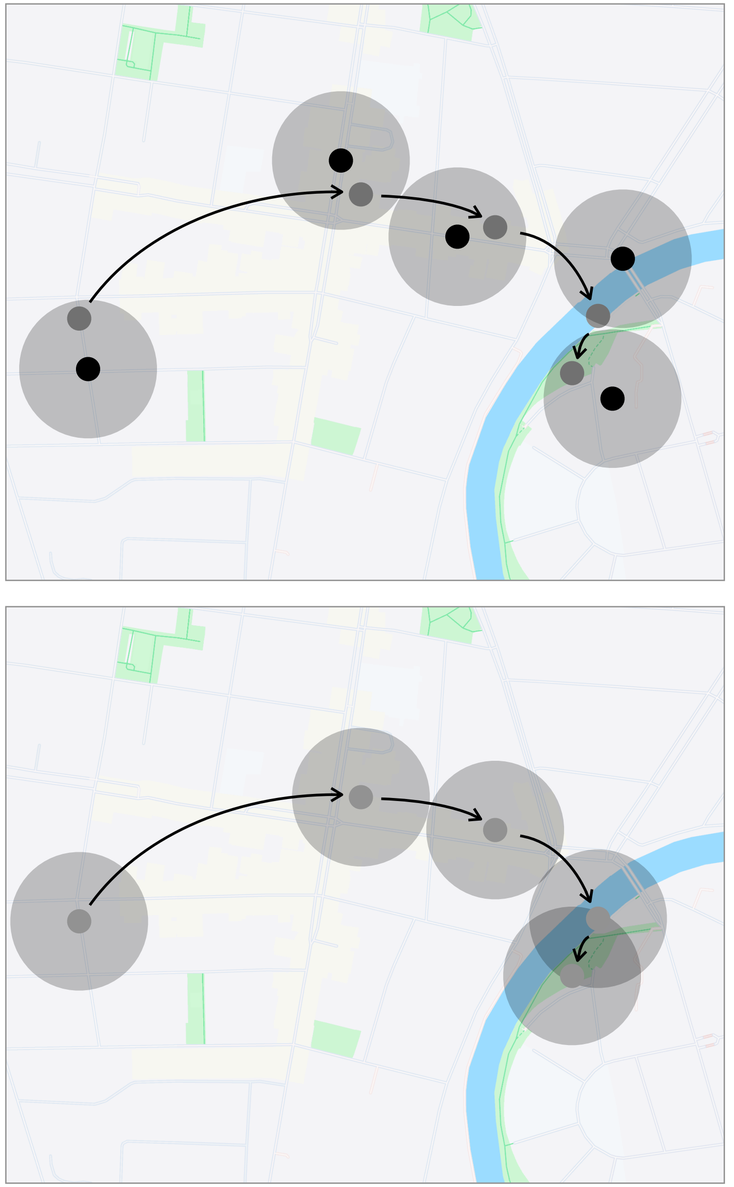
- Differential Privacy: compares distribution of results with Bob and without Bob
  - Protects Bob, even if he does not participate
  - Independent of the data set, only dependent on analysis
  - But: Hard to explain

Privacy Protection Mechanisms
b) Probabilistic

Achieving Differential Privacy

- Different mechanisms to achieve DP
- Noice
  - Modify each value by random amount
  - Randomness will cancel out in aggregated data

    → properties of dataset are preserved (ideally)
  - Calculate max amount of noice according to ε



Dwork, Cynthia. 'Differential Privacy'. *Automata, Languages and Programming*, edited by Michele Bugliesi et al., Springer, 2006, pp. 1–12. *Springer Link*, https://doi.org/10.1007/11787006_1.

Privacy Considerations

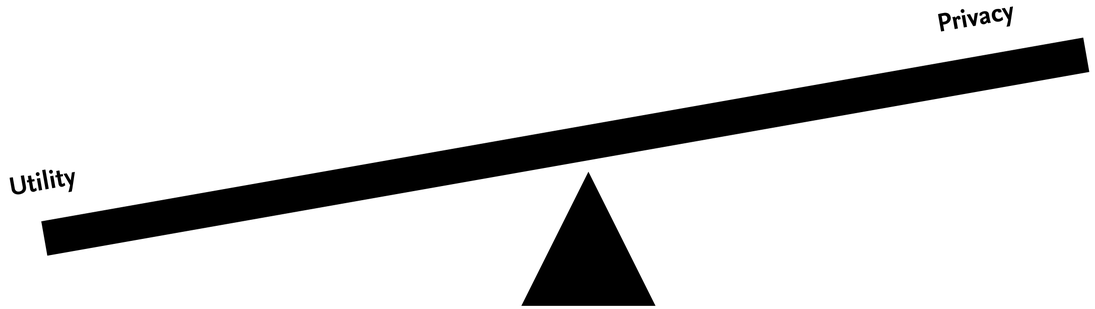
Necessity for Privacy (Ceterum censeo …)

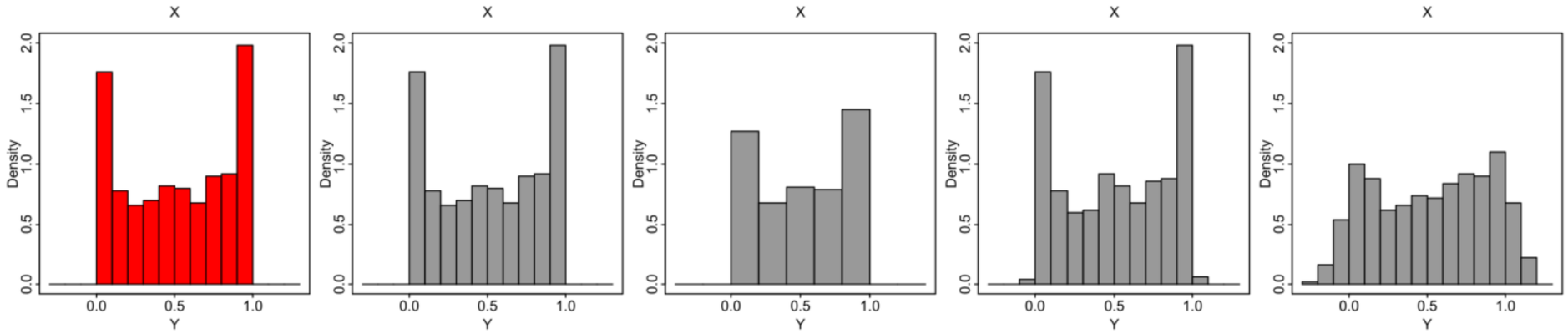
If any sensitive information might be involved

→ Think about privacy, even if anonymized

Balance: Utility vs. Privacy

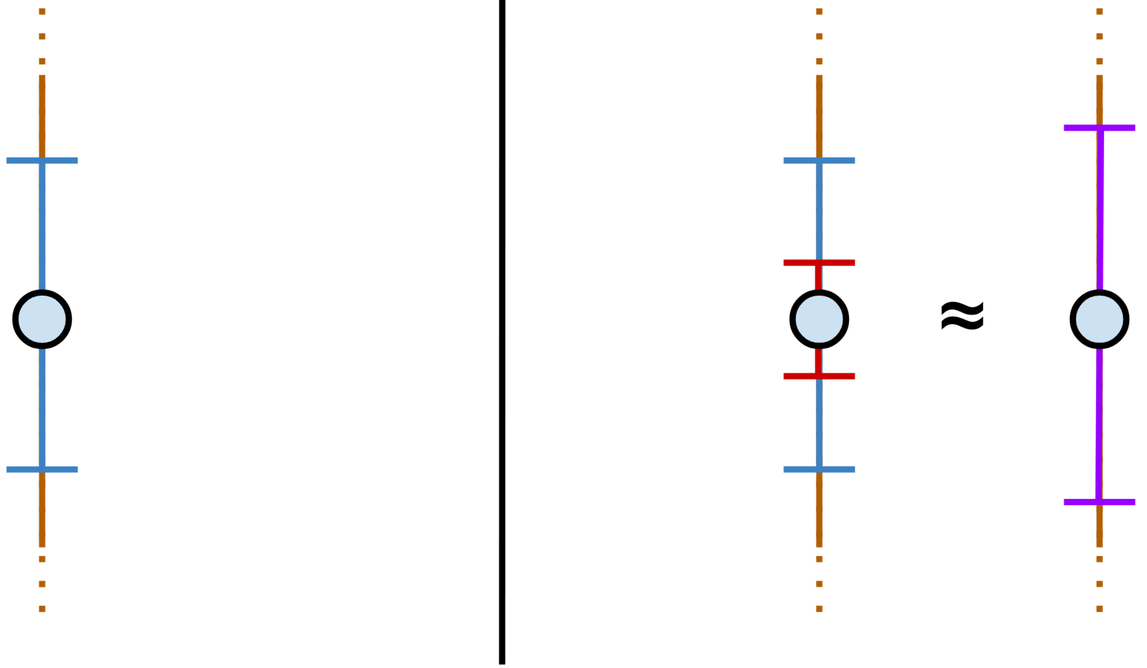
- Generalization, Outlier-removal and Noise reduce accuracy of data
- Data-Consumer: Accuracy = Utility
- More privacy → less utility





(c) Avraam, Demetris, u. a. „Privacy Preserving Data Visualizations“. *EPJ Data Science*, Bd. 10, Nr. 1, 1, Dezember 2021, S. 1–34. *epjdatascience.springeropen.com*, https://doi.org/10.1140/epjds/s13688-020-00257-4.

Accuracy Loss due to Privacy protection



(c) Desfontaines, Damien. *Don't worry, your data's noisy - Ted is writing things*. 27. Juli 2021, https://desfontain.es/privacy/noisy-data.html.

- Decreased accuracy is known → add to uncertainty

Privacy Configurations

- Anonymization is highly configurable
  - Choice of Method (Generalization, Outlier-removal, DP-mechanism)
  - Choice of Parameter (k, l, t, ε)
  - No exact algorithm → intuition necessary
  - Dependents on
    - values collected,
    - context,
    - data-subjects

Sources

- Li, Ninghui, et al. ‘T-Closeness: Privacy Beyond k-Anonymity and l-Diversity’. *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106–15. *IEEE Xplore*, https://doi.org/10.1109/ICDE.2007.367856.
- Avraam, Demetris, et al. ‘Privacy Preserving Data Visualizations’. *EPJ Data Science*, vol. 10, no. 1, 1, Dec. 2021, pp. 1–34. *epjdatascience.springeropen.com*, https://doi.org/10.1140/epjds/s13688-020-00257-4.
- Dwork, Cynthia. ‘Differential Privacy’. *Automata, Languages and Programming*, edited by Michele Bugliesi et al., Springer, 2006, pp. 1–12. *Springer Link*, https://doi.org/10.1007/11787006_1.
- Narayanan, Arvind, and Vitaly Shmatikov. ‘Robust De-Anonymization of Large Sparse Datasets’. *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, 2008, pp. 111–25. *IEEE Xplore*, https://doi.org/10.1109/SP.2008.33.
- Douriez, Marie, et al. ‘Anonymizing NYC Taxi Data: Does It Matter?’ *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 140–48. *IEEE Xplore*, https://doi.org/10.1109/DSAA.2016.21.

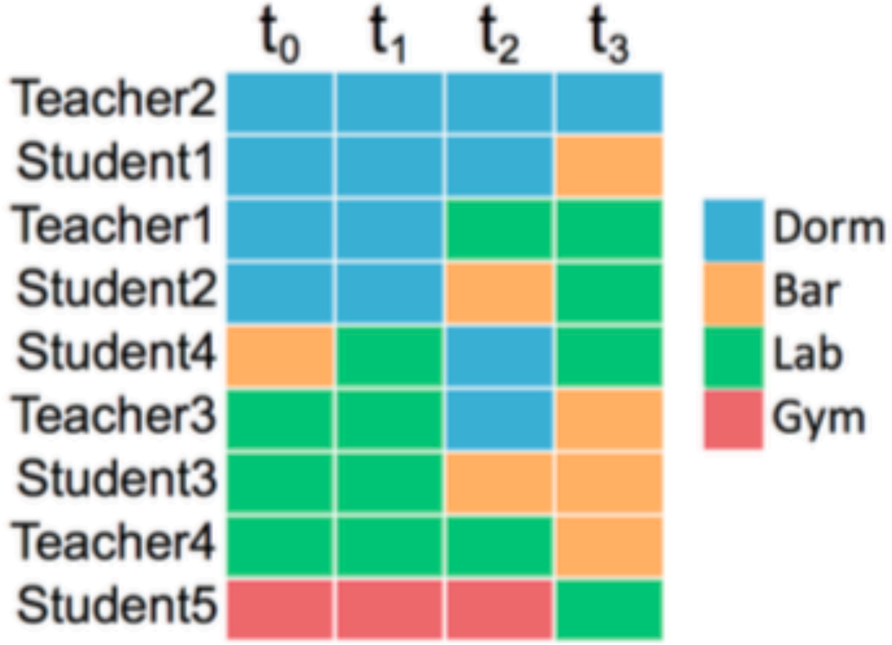
Decoupling

More Privacy → Less Utility ?

- Consider: Sensible information = Analysed dimension?
- If NO → decrease accuracy in the sensible information by decoupling
  - Either: Decouple data entries from each other
  - Or: Decouple data dimensions from each other
- No stronger guarantees in the worst-case, but decreases severity of identity disclosure

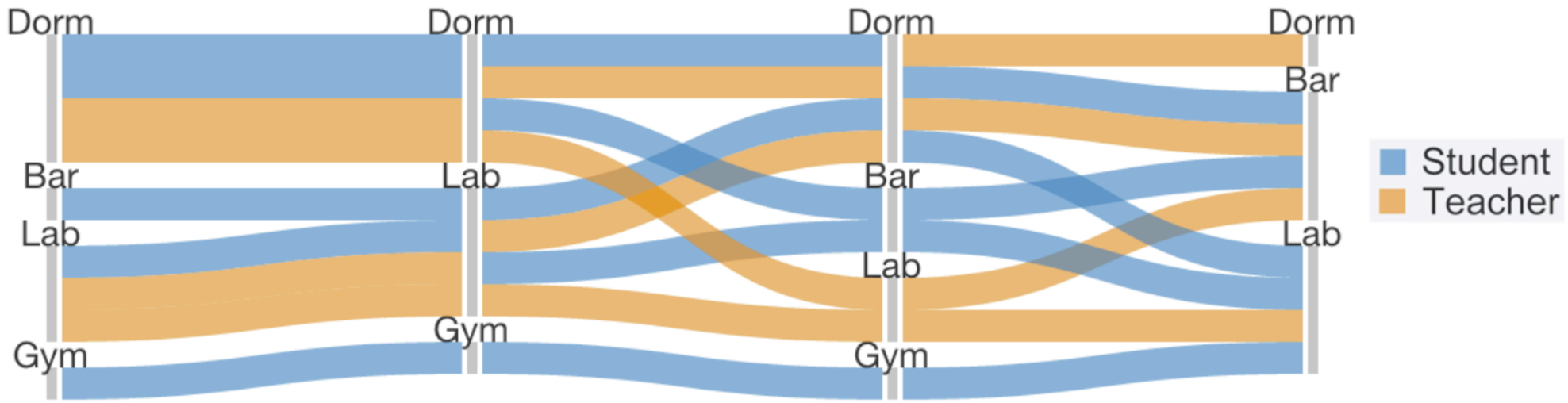
Decouple Data entries

- Scenario: Measure same Attribute over multiple events
  - Example: Location of participants every hour



(c) Chou, Jia-Kai, u. a. „Privacy Preserving Visualization: A Study on Event Sequence Data“. *Computer Graphics Forum*, Bd. 38, Nr. 1, 2019, S. 340–55. *Wiley Online Library*, https://doi.org/10.1111/cgf.13535.

Sankey Diagram

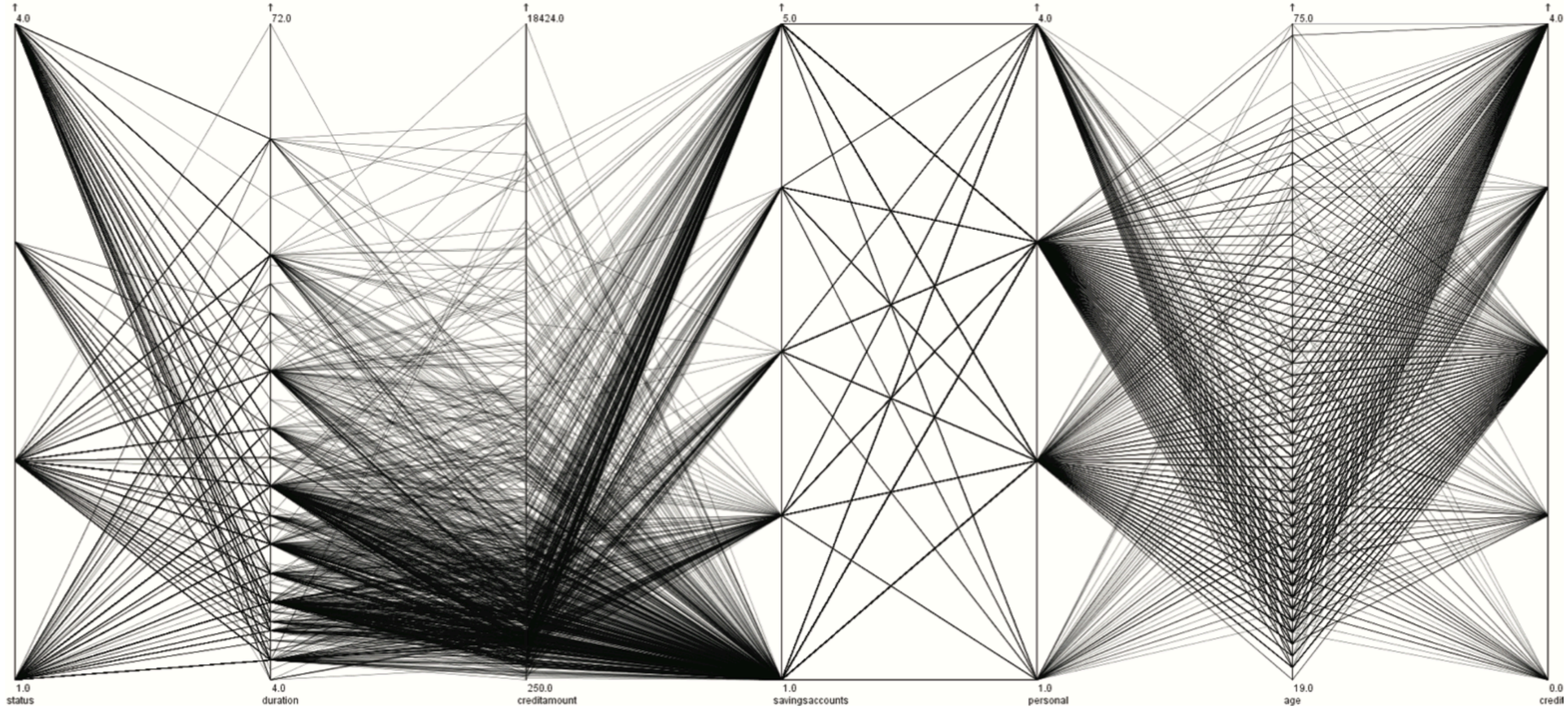


(c) Chou, Jia-Kai, u. a. „Privacy preserving event sequence data visualization using a Sankey diagram-like representation“. *SIGGRAPH ASIA 2016 Symposium on Visualization*, Association for Computing Machinery, 2016, S. 1–8. *ACM Digital Library*, https://doi.org/10.1145/3002151.3002153.

Sankey Diagram

- For data from repeated measurements (event-based)
- Preserves changes between measurements
- Removes connections between non-neighbored measurements in most cases
  - → Attribute Discovery Limited to neighboring measurements
  - → Limited Identity Discovery
- But: still perceptible to outlier identification
  - Is also used together with k-anonymity

Decouple Dimensions: Parallel Coordinates



(a) Default parallel coordinates view of the German credit dataset

(c) Dasgupta, Aritra, und Robert Kosara. „Adaptive Privacy-Preserving Visualization Using Parallel Coordinates“. *IEEE Transactions on Visualization and Computer Graphics*, Bd. 17, Nr. 12, Dezember 2011, S. 2241–48. *IEEE Xplore*, https://doi.org/10.1109/TVCG.2011.163.

Parallel Coordinates

- Decouples Dimensions of Data
- Preserves relations between neighboring dimensions
  - → Attribute Discovery limited to neighboring Attributes
  - → Limited Identity Discovery
- Leaked relations can be tuned
- Still perceptible to outlier identification
  - Modification: Blurring