“Identification and Evaluation of Concepts for Privacy-Enhancing Big Data Analytics Using De-Identification Methods on Wrist-Worn Wearable Data”

Kevin Baumer, 08.06.2020 – Final Presentation Master Thesis

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

1) Motivation & Problem
2) Foundations
3) Research Questions & Approach
4) Results
    1) Overview & Classification of De-Identification Methods
    2) Requirements
    3) Concept Development
5) Conclusion & Limitations

Motivation & Problem Statement

Big Data (Analytics)

- Basis for emerging technologies
- Enormous growth in data

→ Analytics + Data-driven decision-making

Privacy

- Increased privacy awareness
- Privacy is valued (67 % → no control)[1]
- Maturing privacy laws (GDPR, HIPAA, BDSG)

Wrist-worn Wearable Data

- Sensitive Health data
- Increasing popularity (36% in Germany)[2]
- Lots of data points
- E.g. heart rates, blood oxygen saturation, location data

[1] (European Commission, 2018)
[2] https://de-statista-com.eaccess.ub.tum.de/statistik/daten/studie/1047586/umfrage/anteil-der-smartwatch-nutzer-in-deutschland/

Motivation & Problem Statement

Risks & Problems:

- Platform-providers (trustful?)
- 3rd party providers (usage / sale)
- no information about storage, processing of data (despite privacy policy)

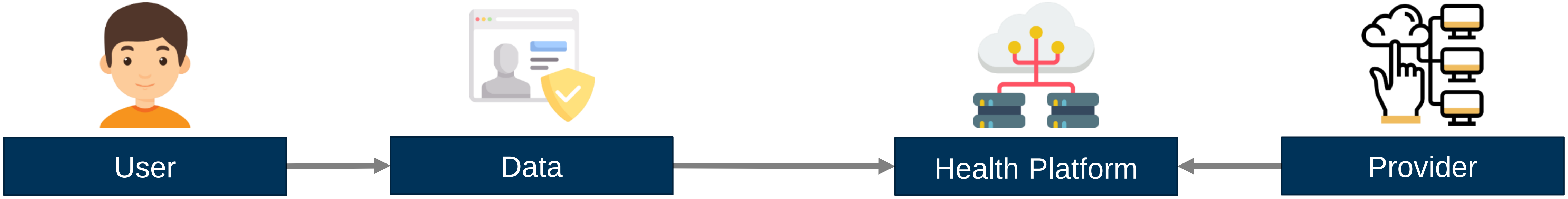
→ Disclosure of personal health information
→ Disclosure of your location data

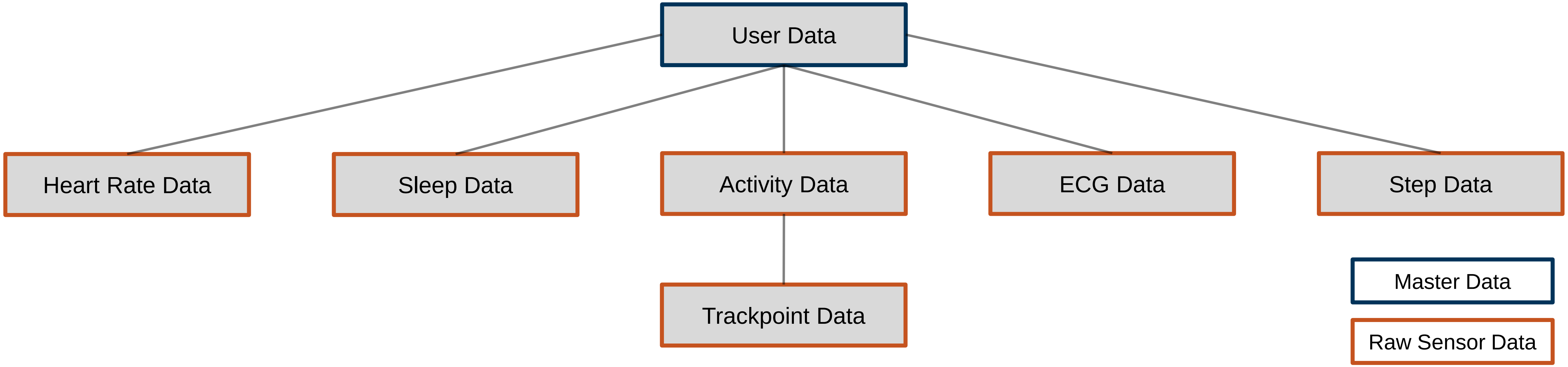
Wrist-worn Wearable Data

- Sensitive Health data
- Increasing popularity (36% in Germany)[2]
- Lots of data points
- E.g. heart rates, blood oxygen saturation, location data

[1] (European Commission, 2018)
[2] https://de-statista-com.eaccess.ub.tum.de/statistik/daten/studie/1047586/umfrage/anteil-der-smartwatch-nutzer-in-deutschland/

Use Case: Wrist-Worn Wearable Data





Foundations

“**Privacy** "[...] is defined not by what it is, but by what it is not - it is the absence of a privacy breach that defines a state of privacy." (Wu, 2012)

De-Identification method: “method for transforming a *dataset* with the objective of reducing the extent to which information is able to be associated with individual *data principals” (ISO, 2018)*

Terminology of De-Identification: (Tomashchuk et. al., 2019)

→ *“a concept of higher level, which covers both anonymization and pseudonymization”*

Research Questions / Research Approach

RQ 1	What is the **state of the art** of approaches using de-identification methods for privacy-enhancing Big Data Analytics and how can they be distinguished from other approaches?	▶	Status Quo
RQ 2	What are **requirements** for privacy-enhancing analytics of wrist-worn wearable data in the cloud?	▶	Requirements
RQ 3	What are **concepts** enabling data privacy for wrist-worn wearable data in the cloud based on **de-identification methods**?	▶	Concepts

RQ 1 approach: Literature Research

RQ 2 approach: Literature Research; Expert interviews; Regulations

RQ 3 approach: Results RQ1 / RQ2; Validation with Experts

Research Question 1: What is the state of the art of approaches using de-identification methods for privacy-enhancing Big Data Analytics and how can they be distinguished from other approaches?

De-Identification – Extensive Literature Review

Databases:

ScienceDirect | IEEE Xplore® Digital Library | WEB OF SCIENCE™ | ACM DL DIGITAL LIBRARY | SpringerLink

Search: methods for De-Identification, Anonymization, Pseudonymization

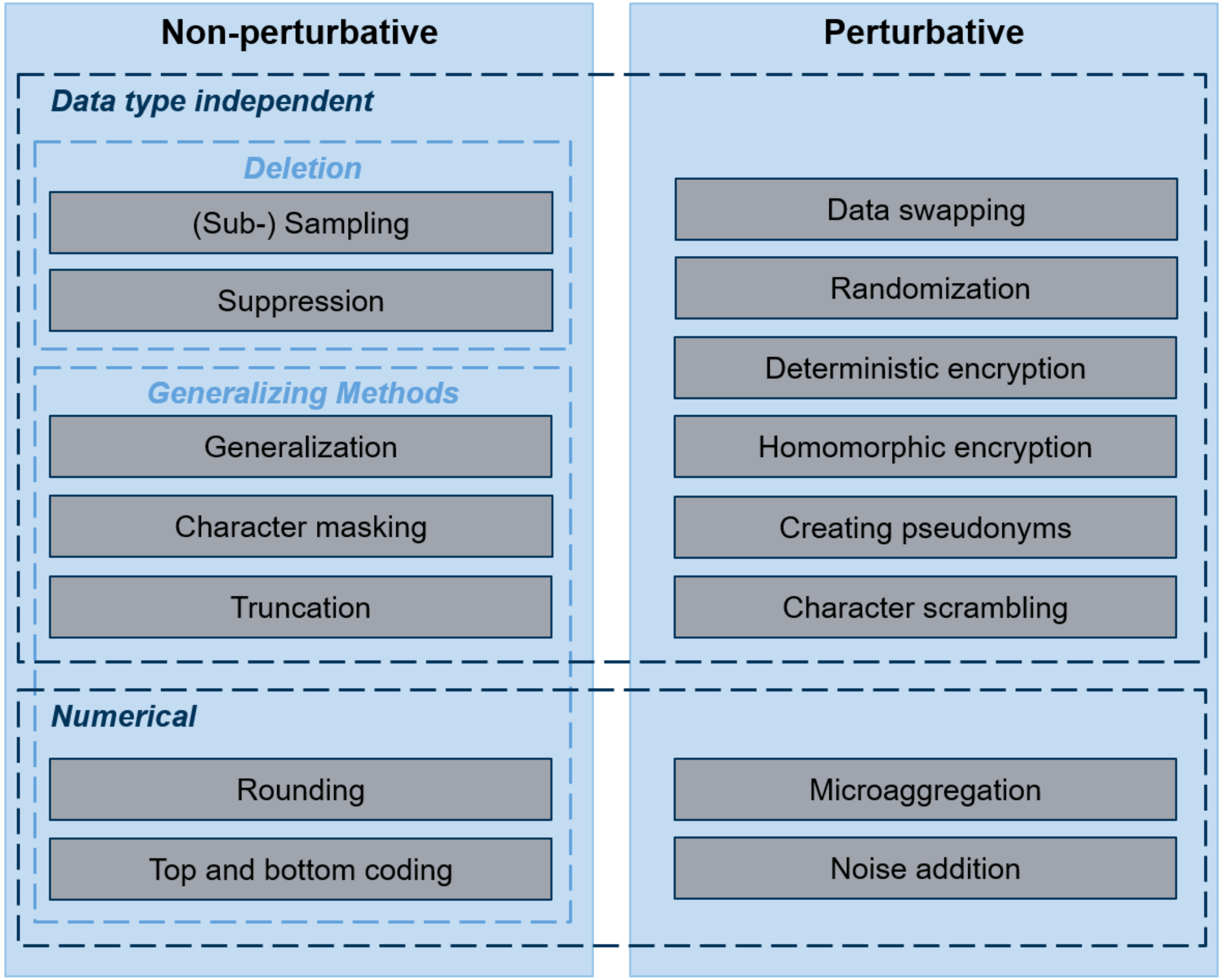
Total
38 sources

→

Comprehensive overviews
7 sources

Source	# of methods
Domingo-Ferrer et. al (2019)	6
Mansfield-Devine (2014)	6
Nelson (2015)	14
Tomashchuk (2019)	7
DP Working Party Art. 29 (2014)	5
Bourka, Drogkaris (2018)	5
ISO/IEC 20889 (2018)	16

Classification of De-Identification Methods



Non-perturbative
→ *Data stays truthful; Accuracy might be reduced*
→ *data truthfulness at the record level*

Perturbative
→ *Transformed values not truthful in general*
→ *statistical properties may be preserved*

Research Question 2: What are requirements for privacy-enhancing analytics of wrist-worn wearable data in the cloud?

Expert Interviews

- 12 semi-structured interviews with industry experts
- Goal: practical insights and implications to derive requirements

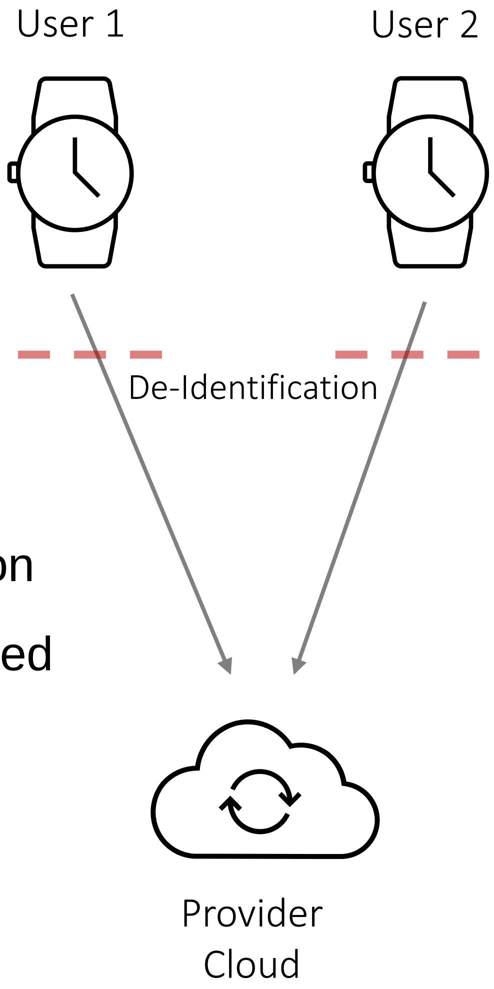
ID	Role	Relevant experience (in years)	No. of employees	Duration (hh:mm:ss)
I1	Consultant Data Privacy & Information Security	>5	1-10	01:04:02
I2	Consultant Data Security & Data Privacy	13	1-10	00:56:49
I3	Head of Data Privacy	20	10,001-50,000	00:28:23
I4	Managing Director & Lawyer	11	1-10	00:42:57
I5	Researcher Digital Health	2	11-50	00:28:13
I6	Data-driven Development & Data Privacy Expert	8	>100,000	0:33:43
I7	Information Security Officer & Data Protection Officer	20	10,001-50,000	00:35:51
I8	Head of Data Privacy	23	>100,000	00:39:57
I9	Head of Data Privacy	19	>100,000	00:28:58
I10	Consultant Data Privacy	7	51-250	00:34:30
I11	Chief Information Security Officer	8	251-1,000	00:32:55
I12	Key Expert Data Privacy	6	>100,000	00:52:21

Identified Requirements

1. Local transformation of data
2. k-anonymity enforcement
3. Privacy levels
4. Generic wrist-worn data model
5. Compliance with regulations
6. Low performance overhead
7. Constraints from a privacy perspective
8. Protection against complete disclosure
9. Transparency
10. Transformation of identifiers

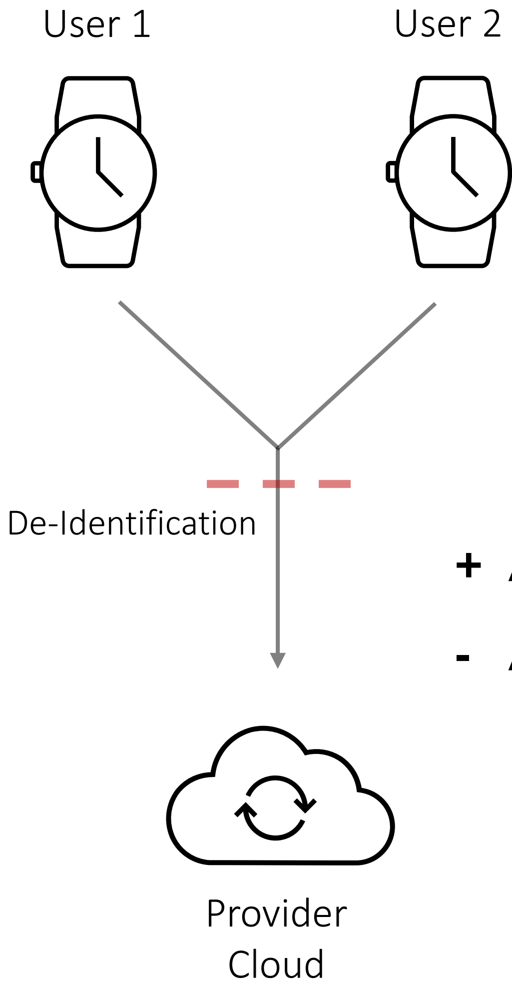
Research Question 3: What are concepts enabling data privacy for wrist-worn wearable data in the cloud based on de-identification methods?

Technical Architecture – Two Options



\+ No additional location

\- Privacy models limited



\+ All models possible

\- Additional point of attack

K-anonymity & AVC

k-anonymity: A table satisfies k-anonymity if every record in the table is indistinguishable from at least k-1 other records with respect to every set of quasi-identifier attributes. (Sweeney, 2002)

Attribute value combinations (AVC): The attribute value combinations represent the total number of possible value combinations for all attributes in a data set.

Zip Code	Age	Nationality
13053	28	Russian
13068	29	American
13068	21	Japanese
13053	23	American
14853	50	Indian
14853	55	Russian
14850	47	American
14850	49	American
13053	31	American
13053	37	Indian
13068	36	Japanese
13068	35	American

(a) k-anonymous table with k=1

Zip Code	Age	Nationality
130**	< 30	*
130**	< 30	*
130**	< 30	*
130**	< 30	*
1485*	≥ 40	*
1485*	≥ 40	*
1485*	≥ 40	*
1485*	≥ 40	*
130**	3*	*
130**	3*	*
130**	3*	*
130**	3*	*

(b) k-anonymous table with k=4

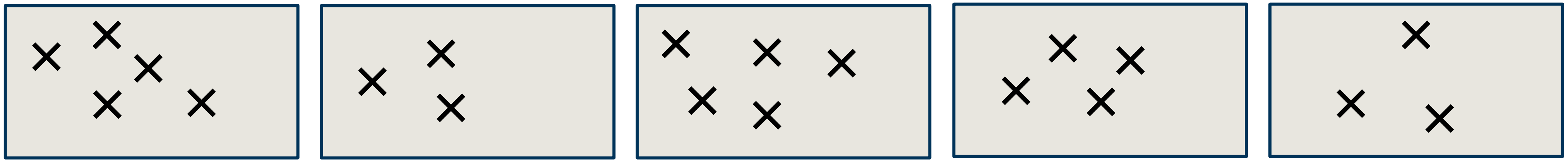
Attribute	# Distinct values
Zip Code	20
Age	80
Nationality	10

→ AVC = 20 * 80 * 10 = 16,000

→ Concept for local probabilistic k-anonymity

Local probabilistic k-anonymity: Monte Carlo simulation

Monte Carlo simulation: repeated random sampling to obtain numerical results



3 parameters:

- g: number of attribute value combinations (AVC) = 5
- r: number of records = 20
- n: number of iterations

→ k: size of the smallest group = 3

Local probabilistic k-anonymity: Monte Carlo simulation (Python)

Input parameters

g = 5 (AVC)

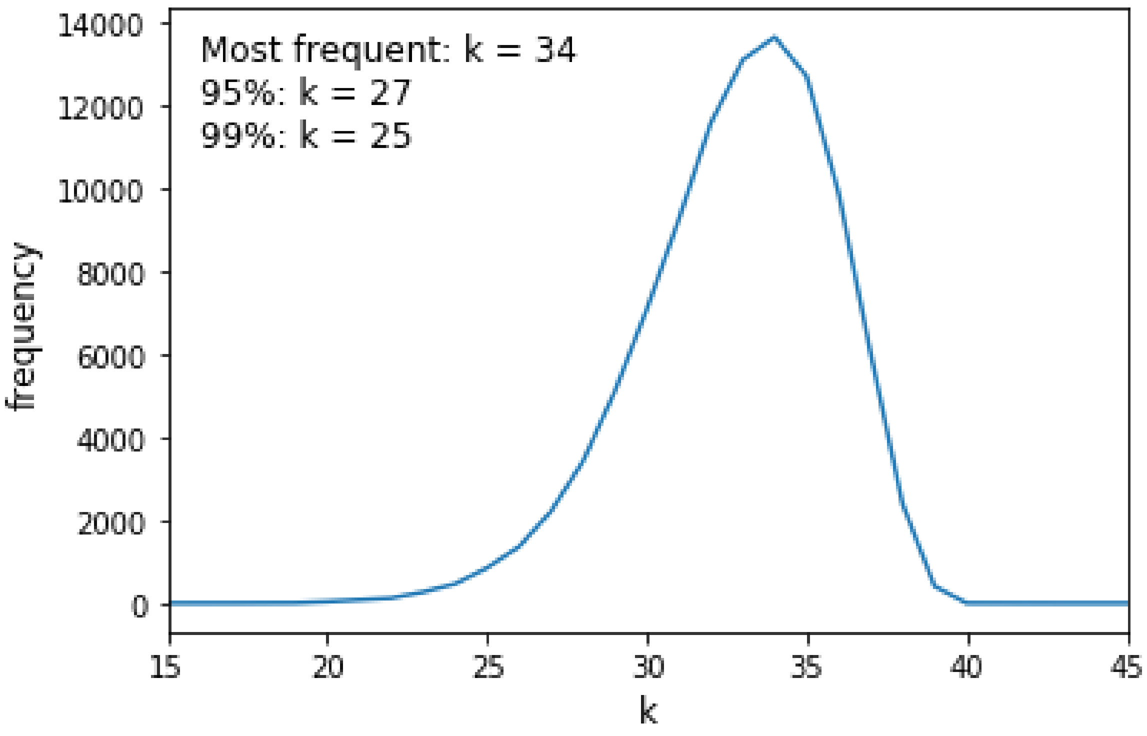
r = 200 (records)

Results

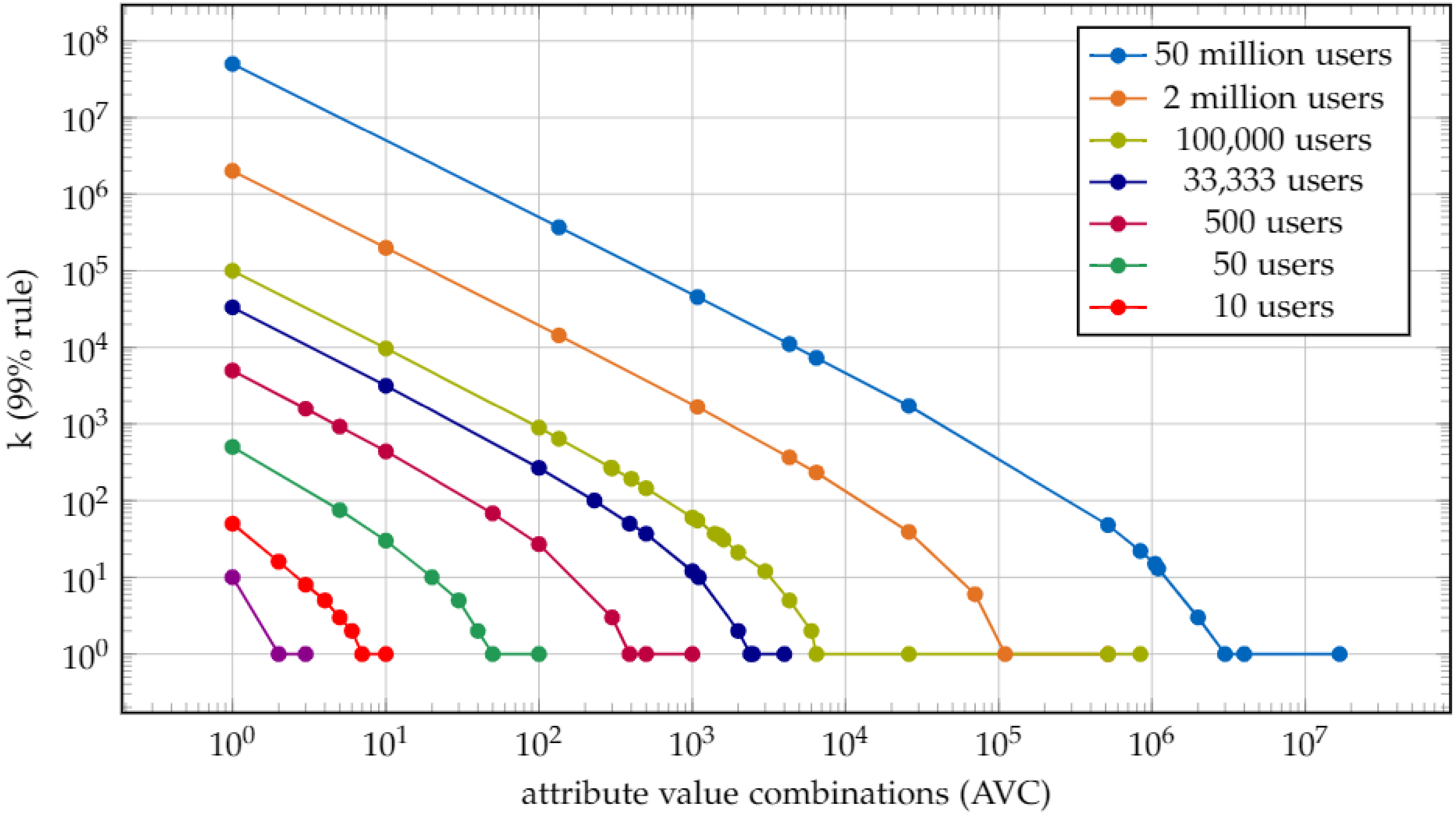
99%: k = 25

→ 1% risk remains

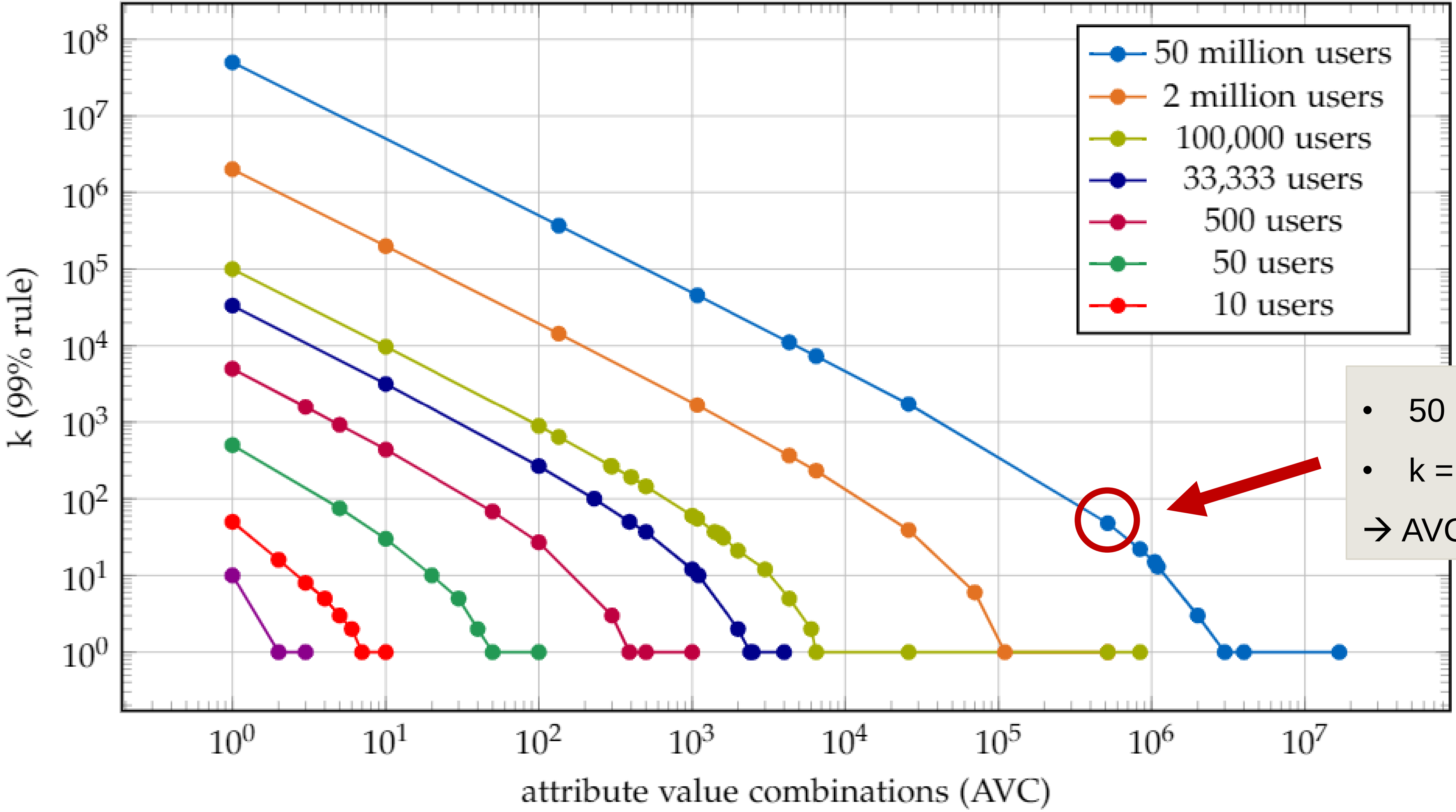
(anonymity threshold)



Local probabilistic k-anonymity - Results



Local probabilistic k-anonymity - Results



- 50 million users
- k = 48

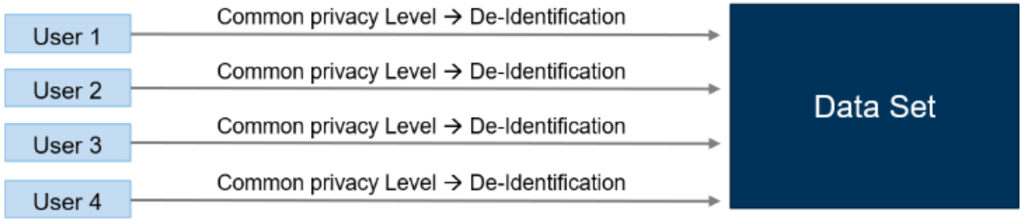
→ AVC = 518,400

De-identification methods to reduce AVC

Quasi Identifiers	Distinct values (initially)	Methods	Distinct values
handedness	2	Suppression	-
gender	3	-	3
height	51	Generalization (5 cm)	10
weight	81	Generalization (5 kg)	16
country	195	Generalization (continent)	6
currentGear	2,000	Generalization (brand)	20
createdDate	3,650	Suppression	-
birthdate	15,695	Generalization (5 years)	9
eMail	~ ∞	Creating pseudonym	-
firstName	~ ∞	Suppression	-
lastName	~ ∞	Suppression	-
profileImage	~ ∞	Suppression	-
AVC	**~ ∞**		**518,400**

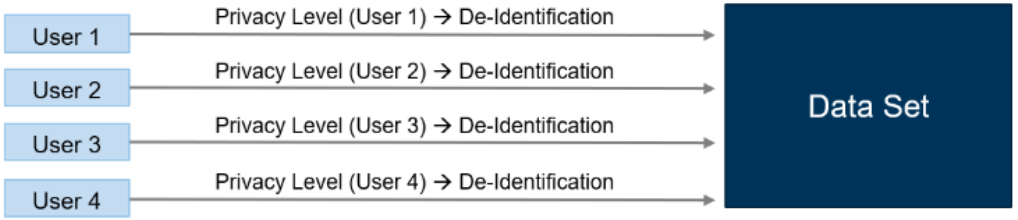
Local probabilistic k-anonymity – 4 Scenarios

Scenario A: Common privacy level



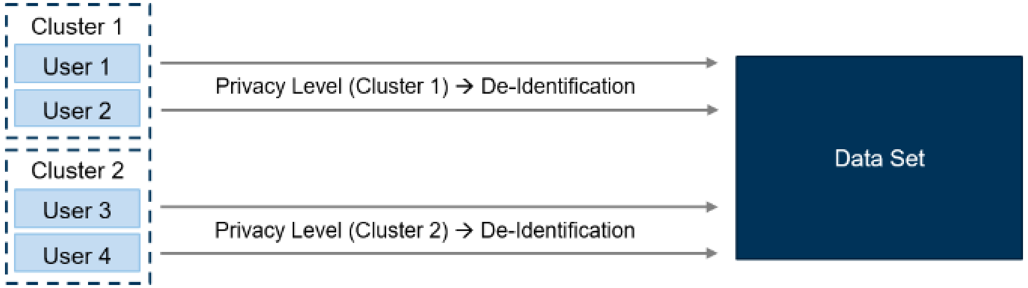
→ No individual privacy levels

Scenario B: Independent & individual privacy levels



→ Mutual weakening

Scenario C / D: Privacy clusters with equal / unequal distribution



→ Best approach to achieve individual privacy levels

Implications and Results for wrist-worn wearable data

- Proposed approach: privacy clusters
- Limiting the number of different clusters (low / medium / high)
- Optimization problem to investigate combination of clusters (unequal distribution)
- Consider fluctuations between clusters (e.g. risk calculation with 80% of users)

Summary & Results

RQ 1

Complete & comprehensive overview of de-identification methods

→ Clear picture and impression of available methods

→ Relation and differentiation between methods

→ Supporting decision process for method choice

RQ 2

10 identified concept requirements

→ Generic wrist-worn data model

RQ 3

Local probabilistic k-anonymity concept for wrist-worn wearable data

→ Privacy estimates based on Monte Carlo simulation

→ Concept involving privacy clusters

Limitations & Future Work

Limitations

- Lacking availability of a data set
- Validation in real application scenario
- Suitable domain experts

Future work

- Extension of the local probabilistic k-anonymity concept
- Evaluation and testing on data set
- Benchmark against local differential privacy

Kevin Baumer

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

matthes@in.tum.de
wwwmatthes.in.tum.de