



## 18. Symposium on Privacy and Security

# Anonymisierung von Daten – die technischen Möglichkeiten



## Big Data to the Extreme: 3× mehr Daten als Sterne



- Volume
- Velocity
- Variety
- Veracity

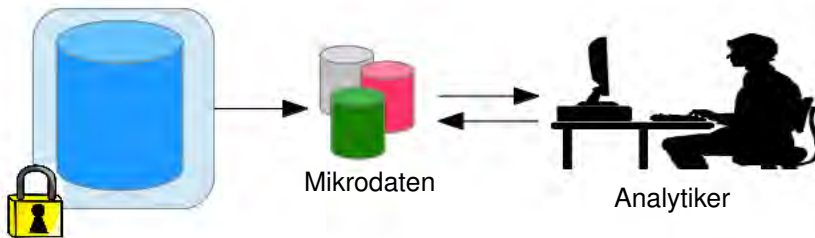
Erwartete Rohdatenrate: 14 Exabytes / Tag

→ HDTV von 20'000 Jahren

## Privacy-Enhancing Cryptography

- Secure Multiparty Computation
    - ☞ Information sharing über private Datensammlungen
  - Private Information Retrieval
    - ☞ schützt die Kriterien der Suchanfrage
  - Suche über verschlüsselte Daten (key words, order-preserving, ...)
    - ☞ Cloud computing
  - Format-preserving Encryption
    - ☞ für Masking-Techniken nützlich
  - Homomorphic Encryption
    - ☞ Laaaaangsam
- ☞ Kryptographie ist gut geeignet, so wenig Information wie möglich über Personen zu sammeln (“Datenminimierung”).

## Offline Data Publishing



👉 Daten sind in der Granularität von Individuen!

## Als anonym veröffentlichte medizinische Daten

SSN	Name	Geb.	Geschl.	PLZ	Ehestand	Krankheit
		09/27/64	W	94139	geschieden	Bluthochdruck
		09/30/64	W	94139	geschieden	Fettsucht
		04/18/64	M	94139	verheiratet	Brustschmerzen
		04/15/64	M	94139	verheiratet	Fettsucht
		03/13/63	M	94138	verheiratet	Bluthochdruck
		03/18/63	M	94138	verheiratet	Kurzatmigkeit
		09/13/64	W	94141	verheiratet	Kurzatmigkeit
		09/07/64	W	94141	verheiratet	Fettsucht
		05/14/61	M	94138	ledig	Brustschmerzen
		05/08/61	M	94138	ledig	Fettsucht
		09/15/61	W	94142	Witwe	Kurzatmigkeit

## Als anonym veröffentlichte medizinische Daten

SSN	Name	Geb.	Geschl.	PLZ	Ehestand	Krankheit
		09/27/64	W	94139	geschieden	Bluthochdruck
		09/30/64	W	94139	geschieden	Fettsucht
		04/18/64	M	94139	verheiratet	Brustschmerzen
		04/15/64	M	94139	verheiratet	Fettsucht
		03/13/63	M	94138	verheiratet	Bluthochdruck
		03/18/63	M	94138	verheiratet	Kurzatmigkeit
		09/13/64	W	94141	verheiratet	Kurzatmigkeit
		09/07/64	W	94141	verheiratet	Fettsucht
		05/14/61	M	94138	ledig	Brustschmerzen
		05/08/61	M	94138	ledig	Fettsucht
		09/15/61	W	94142	Witwe	Kurzatmigkeit

### Wählerliste

Name	Adresse	Stadt	PLZ	Geb.	Geschl.	Partei
Sue. J. Carlson	900 Market St.	San Francisco	94142	9/15/61	W	Demokrat

## Als anonym veröffentlichte medizinische Daten

SSN	Name	Geb.	Geschl.	PLZ	Ehestand	Krankheit
		09/27/64	W	94139	geschieden	Bluthochdruck
		09/30/64	W	94139	geschieden	Fettsucht
		04/18/64	M	94139	verheiratet	Brustschmerzen
		04/15/64	M	94139	verheiratet	Fettsucht
		03/13/63	M	94138	verheiratet	Bluthochdruck
		03/18/63	M	94138	verheiratet	Kurzatmigkeit
		09/13/64	W	94141	verheiratet	Kurzatmigkeit
		09/07/64	W	94141	verheiratet	Fettsucht
		05/14/61	M	94138	ledig	Brustschmerzen
		05/08/61	M	94138	ledig	Fettsucht
		09/15/61	W	94142	Witwe	Kurzatmigkeit

### Wählerliste

Name	Adresse	Stadt	PLZ	Geb.	Geschl.	Partei
Sue. J. Carlson	900 Market St.	San Francisco	94142	9/15/61	W	Demokrat

## Anonymisierung: Datentransformationsmethoden

### ■ **Perturbative Ansätze**

- Bewahren Aggregatstatistik (Mittelwert, Korrelationskoeffizient, ...), z. B. durch
  - Hinzufügen von Rauschen,
  - Daten vertauschen,
  - Micro-aggregation,
  - Runden
- Verfälschen der Daten

### ■ **Nicht-perturbative Ansätze**

- Verändern die Granularität der veröffentlichten Daten, z. B. durch
  - **Generalisierung**  
PLZ (24103 → 241\*\*), Geschlecht (M → \*), Alter (24 → [20–29])
  - **Unterdrückung** (“Ausreisser”)
- **Keine Verfälschung der Daten!**



## $k$ -Anonymity – Ein Mass für den Schutz von Personendaten

### Quasi-Identifikator

- Eine Untermenge der Attribute, deren Wertekombination für eine Person charakteristisch sein könnte.

## $k$ -Anonymity – Ein Mass für den Schutz von Personendaten

### Quasi-Identifikator

- Eine Untermenge der Attribute, deren Wertekombination für eine Person charakteristisch sein könnte.

### Sensitive Attribute

- Attribute, welches nicht mit einer Person verknüpfbar sein sollen.

## $k$ -Anonymity – Ein Mass für den Schutz von Personendaten

### Quasi-Identifikator

- Eine Untermenge der Attribute, deren Wertekombination für eine Person charakteristisch sein könnte.

### Sensitive Attribute

- Attribute, welches nicht mit einer Person verknüpfbar sein sollen.

### $k$ -Anonymity

- $k$  Datensätze bilden eine Äquivalenzklasse
- schützt mit einer Konfidenz von  $1/k$  vor einer 'korrekten' Verknüpfung einer Person mit ihren sensitiven Attributen

Ein Tabelle ist  $k$ -anonym, wenn jedes Tupel von mindestens  $k - 1$  anderen Tupeln (bis auf die sensitiven Attribute) nicht unterscheidbar ist.

## Eine Tabelle

<b>Geburtstag</b>	<b>Geschl.</b>	<b>PLZ</b>	<b>Ehestand</b>	<b>Krankheit</b>
09/27/64	W	94139	geschieden	Bluthochdruck
09/30/64	W	94139	geschieden	Fettsucht
04/18/64	M	94139	verheiratet	Brustschmerzen
04/15/64	M	94139	verheiratet	Fettsucht
03/13/63	M	94138	verheiratet	Bluthochdruck
03/18/63	M	94138	verheiratet	Kurzatmigkeit
09/13/64	W	94141	verheiratet	Kurzatmigkeit
09/07/64	W	94141	verheiratet	Fettsucht
05/14/61	M	94138	ledig	Brustschmerzen
05/08/61	M	94138	ledig	Fettsucht
09/15/61	W	94142	Witwe	Kurzatmigkeit

## ... und ihre minimale Generalisierung

Generalisierung (spaltenweise): Tag  $\rightarrow$  Jahr  $\rightarrow$  5 Jahre

{M, W}  $\rightarrow$  \*

{verheiratet, geschieden, Witwe}  $\rightarrow$  nicht-ledig  $\rightarrow$  \*

Geburtsjahr	Geschl.	PLZ	Ehestand
64	*	941**	*
64	*	941**	*
64	*	941**	*
64	*	941**	*
63	*	941**	*
63	*	941**	*
64	*	941**	*
64	*	941**	*
61	*	941**	*
61	*	941**	*
61	*	941**	*

## ... und ihre minimale Generalisierung

Generalisierung (spaltenweise): Tag  $\rightarrow$  Jahr  $\rightarrow$  5 Jahre

{M, W}  $\rightarrow$  \*

{verheiratet, geschieden, Witwe}  $\rightarrow$  nicht-ledig  $\rightarrow$  \*

Geburtsjahr	Geschl.	PLZ	Ehestand
64	*	941**	*
64	*	941**	*
64	*	941**	*
64	*	941**	*
63	*	941**	*
63	*	941**	*
64	*	941**	*
64	*	941**	*
61	*	941**	*
61	*	941**	*
61	*	941**	*

Geburtsjahr	Geschl.	PLZ	Ehestand
[60 – 64]	W	9413*	nicht-ledig
[60 – 64]	W	9413*	nicht-ledig
[60 – 64]	M	9413*	nicht-ledig
[60 – 64]	M	9413*	nicht-ledig
[60 – 64]	M	9413*	nicht-ledig
[60 – 64]	W	9414*	nicht-ledig
[60 – 64]	W	9414*	nicht-ledig
[60 – 64]	M	9413*	ledig
[60 – 64]	M	9413*	ledig
[60 – 64]	W	9414*	nicht-ledig

## Beispiele von Angriffen

### Homogenität

#### Urs

PLZ	Alter
74678	26

### Hintergrundwissen

#### Satoshi (Japaner)

PLZ	Alter
74673	36

### Eine 3-diverse Patiententabelle

PLZ	Alter	Gehalt	Krankheit
746**	2*	20K	Herzerkrankung
746**	2*	30K	Herzerkrankung
746**	2*	40K	Herzerkrankung
7490*	≥ 40	50K	Gastritis
7490*	≥ 40	100K	Grippe
7490*	≥ 40	70K	Bronchitis
746**	3*	60K	Herzerkrankung
746**	3*	80K	Krebs
746**	3*	90K	Krebs

$k$ -Anonymity kann versagen, falls

- es den sensitiven Werten in einer Äquivalenzklasse an **Vielfalt** mangelt oder
- der Angreifer **Hintergrundwissen** besitzt.

## /Diversity

Jeder  $q^*$ -Block enthält mindestens  $l$  “wohl-vertretene” Werte des sensitiven Attributes  $s$ .

Alter	Geschlecht	Krankheit
[26 – 27]	M	Grippe
[26 – 27]	M	Grippe
[23 – 25]	*	Erkältung
[23 – 25]	*	Diabetes
22	M	Grippe
22	M	Krebs

$$k = 2$$



## I-Diversity

Jeder  $q^*$ -Block enthält mindestens  $l$  “wohl-vertretene” Werte des sensitiven Attributes  $s$ .

Alter	Geschlecht	Krankheit
[26 – 27]	M	Grippe
[26 – 27]	M	Grippe
[23 – 25]	*	Erkältung
[23 – 25]	*	Diabetes
22	M	Grippe
22	M	Krebs

$$k = 2$$

Alter	Geschlecht	Krankheit
[25 – 27]	M	Grippe
[25 – 27]	M	Grippe
[25 – 27]	M	Erkältung
[22 – 23]	*	Diabetes
[22 – 23]	*	Grippe
[22 – 23]	*	Krebs

$$k = 3, E \geq \log(1.9)$$

- Datenveröffentlicher benötigt weniger Information als der Angreifer
- berücksichtigt *instance-level knowledge* (“mein Nachbar hat keine Diabetes”)

## Utility Measure und Risk Assessment

Alter	Geschl.	Krankheit
[26 – 27]	M	Grippe
[26 – 27]	M	Grippe
[23 – 25]	*	Erkältung
[23 – 25]	*	Diabetes
22	M	Grippe
22	M	Krebs

$k = 2$

Alter	Geschl.	Krankheit
[25 – 27]	M	Grippe
[25 – 27]	M	Grippe
[25 – 27]	M	Erkältung
[22 – 23]	*	Diabetes
[22 – 23]	*	Grippe
[22 – 23]	*	Krebs

$k = 3$

Alter	Geschl.	Krankheit
[22 – 27]	*	Grippe
[22 – 27]	*	Grippe
[22 – 27]	*	Erkältung
[22 – 27]	*	Diabetes
[22 – 27]	*	Grippe
[22 – 27]	*	Krebs

$k = 6$

Wird die Anonymitätsgarantie verstärkt, verringert sich die Datenqualität: Es benötigt eine Güterabwägung zwischen Nutzwert und Datenschutz.

## Daten-Representation

- Relationale Daten
  - Registrierungs- und demographische Daten
- Transactional (set-valued) Daten
  - Abrechnungen
- Sequentielle Daten
  - DNA
- Trajektorien (Bahnkurven)
  - Ortsdaten von Mobiltelefonen
- Graphen
  - Soziale Netzwerke
- Text
  - Klinische Aufzeichnungen
  - Tweets

Electronic Medical Records			
Name	Geburtsjahr	ICD	DNA
Beat	1955	493.00, 185	C ... T
Vreni	1943	185, 157.3	A ... G
Vreni	1943	493.01	C ... G
Ursula	1965	493.02	C ... G
Urs	1973	157.9, 493.03	G ... C
Urs	1973	157.3	A ... T

*19 Jahre alter Mann mit Vorgeschichte Ekzem im Kleinkindalter, jetzt sporadische lokale Beschwerden im Mund nach Erdnussverzehr und Rhinokonjunktivitis während der Pollensaison.*

## Einzigartigkeit von Daten

Wieviele persönliche Daten sind notwendig, um jemand re-identifizieren zu können:

- (Geburtsjahr, Geschlecht, 3-stellige PLZ)  
→ 0.04% der amerikanischen Bevölkerung
- (Geburtsdatum, Geschlecht, 5-stellige PLZ)  
→ 63–87 % der amerikanischen Bevölkerung

## Einzigartigkeit von Daten

Wieviele persönliche Daten sind notwendig, um jemand re-identifizieren zu können:

- (Geburtsjahr, Geschlecht, 3-stellige PLZ)  
→ 0.04% der amerikanischen Bevölkerung
- (Geburtsdatum, Geschlecht, 5-stellige PLZ)  
→ 63–87 % der amerikanischen Bevölkerung
- 2 spatio-temporale Punkte → 50%
- 4 spatio-temporale Punkte → 95%

## Einzigartigkeit von Daten

Wieviele persönliche Daten sind notwendig, um jemand re-identifizieren zu können:

- (Geburtsjahr, Geschlecht, 3-stellige PLZ)  
→ 0.04% der amerikanischen Bevölkerung
- (Geburtsdatum, Geschlecht, 5-stellige PLZ)  
→ 63–87 % der amerikanischen Bevölkerung
- 2 spatio-temporale Punkte → 50%
- 4 spatio-temporale Punkte → 95%
- 2 ICD Nummern → > 90%

## Text De-Identifikation

*Ein Einwohner von **Aarau** kaufte **Marihuana** gegen **lumbale Schmerzen**,  
verursacht durch **Leberkrebs**.*

## Text De-Identifikation

*Ein Einwohner von **Aarau** kaufte **Marihuana** gegen **lumbale Schmerzen**,  
verursacht durch **Leberkrebs**.*

*Ein Einwohner von ~~Aarau~~ kaufte ~~Marihuana~~ gegen ~~lumbale Schmerzen~~,  
verursacht durch ~~Leberkrebs~~.*



## Text De-Identifikation

*Ein Einwohner von Aarau kaufte Marihuana gegen lumbale Schmerzen, verursacht durch Leberkrebs.*

*Ein Einwohner von Aarau kaufte Marihuana gegen ~~lumbale~~ Schmerzen, verursacht durch ~~Leberkrebs~~.*

$t$ -Plausibility verallgemeinert sensitive Terme zu semantisch ähnlichen Termen, z. B. “Tuberkulose” → “Infektion”.

Ist eine Wortontologie und ein Grenzwert  $t$  gegeben, kann der gesäuberte Text mindestens  $t - 1$  anderen Texten zugeordnet werden.

## Text De-Identifikation

Ein Einwohner von **Aarau** kaufte **Marihuana** gegen **lumbale Schmerzen**,  
verursacht durch **Leberkrebs**.

Ein Einwohner von ~~Aarau~~ kaufte ~~Marihuana~~ gegen ~~lumbale Schmerzen~~,  
verursacht durch ~~Leberkrebs~~.

$t$ -Plausibility verallgemeinert sensitive Terme zu semantisch ähnlichen Termen,  
z. B. “Tuberkulose” → “Infektion”.

Ist eine Wortontologie und ein Grenzwert  $t$  gegeben, kann der gesäuberte Text  
mindestens  $t - 1$  anderen Texten zugeordnet werden.

Ein Einwohner von **Kantonshauptstadt** kaufte **Droge** gegen **Schmerzen**,  
verursacht durch **Karzinom**.

## Resümee

Der Wert von Daten mit Personenbezug erschöpft sich nicht schon in ihrer ersten Verwendung. Aber wie können sie Dritten sicher zugänglich gemacht werden?

- Verschiedene Anonymisierungsmethoden und -masse
- Eine De-Identifikation von Daten gibt keine (strikte) Garantie der Anonymität!
  - *k*-Anonymity – Schutz gegen Verknüpfung von Identitäten
  - *l*-Diversity – Schutz gegen die Offenlegung von Attributen

## Big Data

- Masking (engl. Redaction) erweitert mit Generalisierung
- Erste Anonymitätsmasse, z. B. *t*-Plausibility

Vielen Dank für ihre Aufmerksamkeit!



Quelle: Jorge Stolfi

## Acknowledgements

Aris Gkoulalas-Divanis (IBM Research – Dublin)