sicherheitskultur.at

Smart City und Smart Living

Anonymisierung und Pseudonymisierung

Philipp Schaumann
philippschaumann@mailbox.org

Disclaimer:
- Alle hier präsentierten Positionen sind rein privater Natur
- Die technischen Details haben keinen Zusammenhang mit Angeboten oder Software meines Arbeitgebers

"Smart" oder "4.0" – Die Basis ist Big Data

- Smart …..
  bedeutet heute: Wir wollen Deine Daten – wenn wir alles über Dich wissen, dann können wir alle Deine Probleme lösen
- Verkehr, Wohnung, Logistic, Dating, . . .

Smart anything = Big Data everywhere

Technology

Tomorrow's cities: Ho
changing the world

By Jane Wakefield
Technology reporter

Guardian sustainable business Technology and Innovation

A truly smart city is more than sensors, big data and an all-seeing internet

Investing billions in big data and smart technology isn't the only answer to building more sustainable urban areas. We need to focus on the big levers

Cities Get Smart with Big Data

by **Ellis Booker** | September 25, 2014 5:30 am | 0 Comments

Forget moving sidewalks and robot police. The biggest technology change in our cities will involve data, and lots of it.

Big Data is incompatible with data protection

<u>Data protection:</u>

- Data can only be collected for a well-defined purpose
- Data can only be collected on well-defined legal basis or with consent

<u>Big Data</u>

- As much Data as possible is collected without knowing what it can be used for
- Consent of the person is assumed or data is considered “anonymous”

Privatsphäre??
Kein Problem – wir anonymisieren

Anonymity = no name (no identifying data field)

. . . that data can no longer be used to
identify a (natural) person by using
“all the means likely reasonably to be used”
by either the controller or a third party –
the processing must be **irreversible**..

Also, „events“ that are relating to 1 person shall not be linkable (no „tracking“)

Source: Definition of
ARTICLE 29 DATA PROTECTION WORKING PARTY

Pseudonymity

Pseudonymity = name and other identifying data fields have been replaced by values that allows for the **original data controller** to match the data with other data – e.g. it allows matching of events concerning the same person, etc.

The **recipient of the data** cannot (by legal means) identify the person.

Anonymity / Pseudonymity Legal Position

Anonymous data are no longer protected by D.P. law

Pseudonymous data:
- still protected by D.P. law
- in AT: privileged as indirect personal data

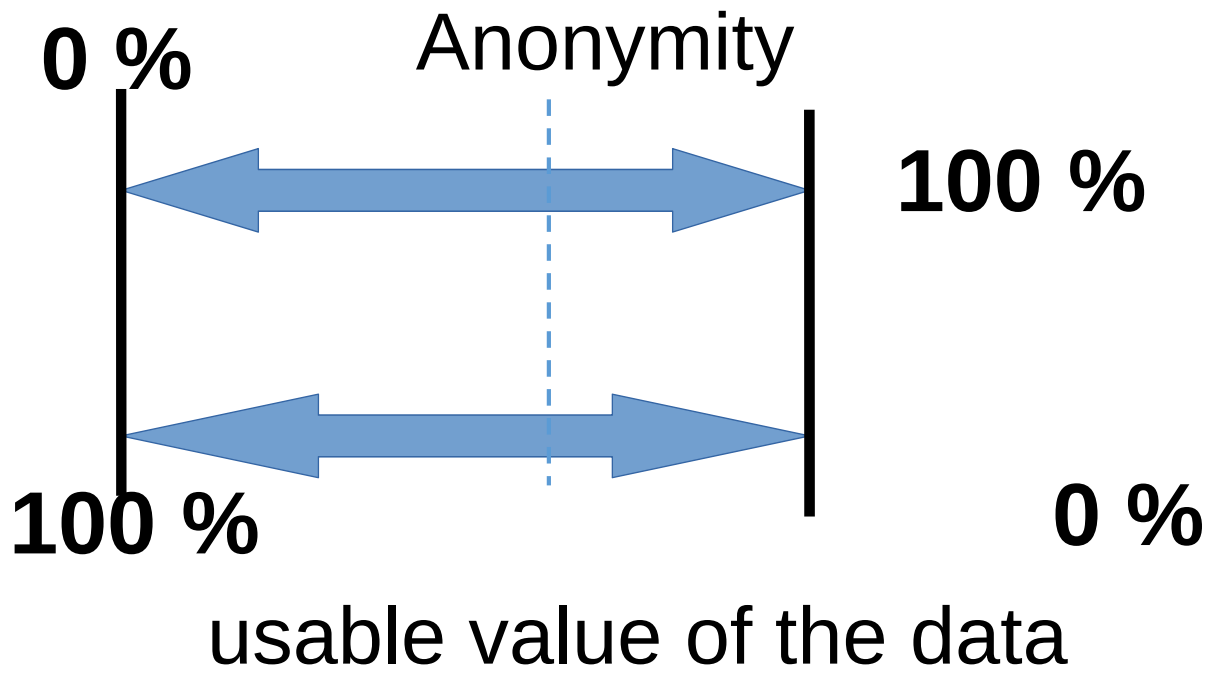
Degree of Anonymization

100 % Anonymity means that
- no entity in the (known) universe can,
- even with lots of compute power and
- unlimited access to all other data sources,
- now and in the forseable future

re-identify the person

Complete anonymity is very hard to do and the data looses most of its value – it is equal to synthetic data

Anonymity and usable value of the data



The Key Questions

1. What should the data be used for after anonymization?

2. Who is the recipient of the data and what other data sources does the recipient have that allows re-identification?

Multiple events per person

- If there are multiple events per person (either in the same record or in multiple records) then pseudonymization and even in-complete anonymization will still allow „tracking“ of the events of 1 person = violation of strict anonymity

What should the data be used for after anonymization? - Debugging

- Using a single table for debugging should allow for complete anonymization – e.g. replacing all sensitive data items with random values
- Care must be taken if there are several tables with referential constraints between them (linkages between data fields that contains the same value) – then referential integrity must be preserved this means total random data will not work

What should the data be used for after anonymization? - Statistical analysis

- If the goal is a statistic over the whole data set (1 record per person) then perfect anonymization will still work
- If the statistics should analyse the influence of parameters like age, education, location of residence, then some identity-preserving values need to be kept
- If multiple events per person should be analyzed, perfect anonymization will not work

What should the data be used for after anonymization? - Checking the Correctness of Financial Calculations

- Referential integrity must usually be preserved – complete anonymization is usually not possible
- Possible solution depends on the testing needs and the other data that the data recipient has to de-anonymize the data

De-Anonymization
AOL search data (2006)

- **AOL apologizes for exposing search data**

- **A spokesman for the ISP-turned-portal says the release of keyword search information from about 658,000 anonymous AOL users was a "screw up" that was based on good intentions.**

The Netflix case (2007)

Projecting two databases onto the same subspace

Each record consists of anonymized personal data plus movie ratings with the date of the rating. (movie name, date, rating value)

Each record can be represented by a point in a multidimensional space, where each attribute is a coordinate.

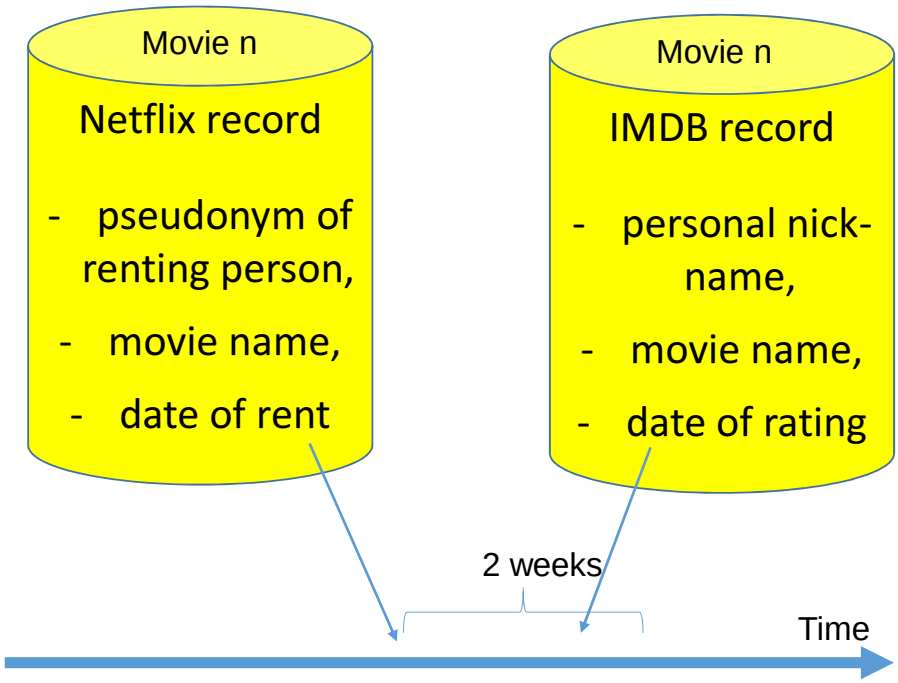
They are so distant that after partitioning the space into regions, each region contains only one record. In the Netflix case, records were sufficiently unique with just 8 movie ratings given within 14 days of distance.

The very same selection of just 8 rated movies constituted a fingerprint of the expressed ratings, not shared between two data subjects within the database.

The researchers matched the supposedly anonymous Netflix data set with another public database with movie ratings (the IMDB), thus finding users who had expressed ratings for the same movies within the same time intervals.

De-Anonymization
The Netflix case (2007)

Projecting two databases onto the same subspace



Identifier vs. Quasi Identifier

- **Identifiers** need to be removed (name, PP#, Soc. Sec. No, email-address, account #, phone #, …..)
- **Quasi-Identifiers** (QI)/ key attributes need to be considered in combination with each other + possible other sources for this data (Zip Code, Gender, Age, Address, education, profession, memberships, participation in groups or events, ….)

De-Anonymization
DNA Study (2013)

•Apr 25, 2013 @ 03:47 PM **18,872** views

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study

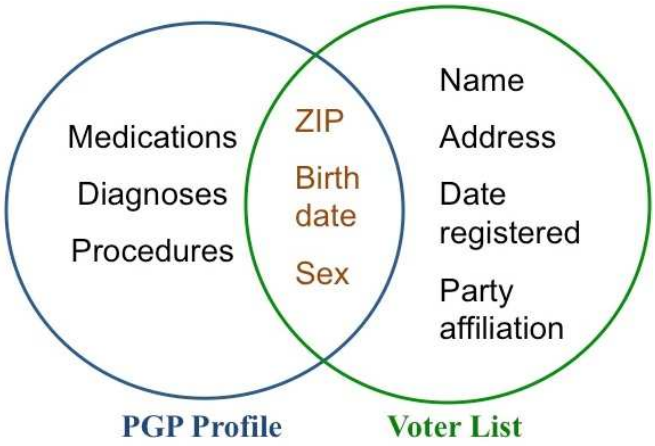
. . . has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.

De-Anonymization
Personal Genome Project

- De-Anonymization is possible, if the combination of several data items generate uniqueness.

E.g.

- sex + birthdate + 5-digit ZIP-code identify 87% of US-population, sex + birthdate + city identify 53 % (all identifiable against voter register)



Source:
Personal Genome Project (PGP) De-Anonymisation

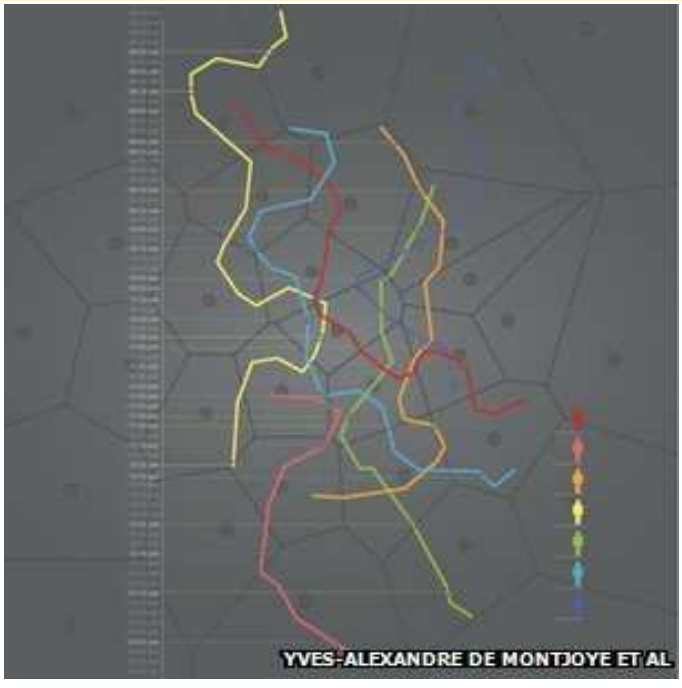
De-Anonymization
Montjoye Mobility Study 2013

- . . . fifteen months . . . for one and a half million individuals:
- . . . in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals

De-Anonymization
Montjoye Mobility Study 2013

- Day-long location tracking (mobility pattern):
  Week-day daylight location + main night location is unique for 90+ % of the people

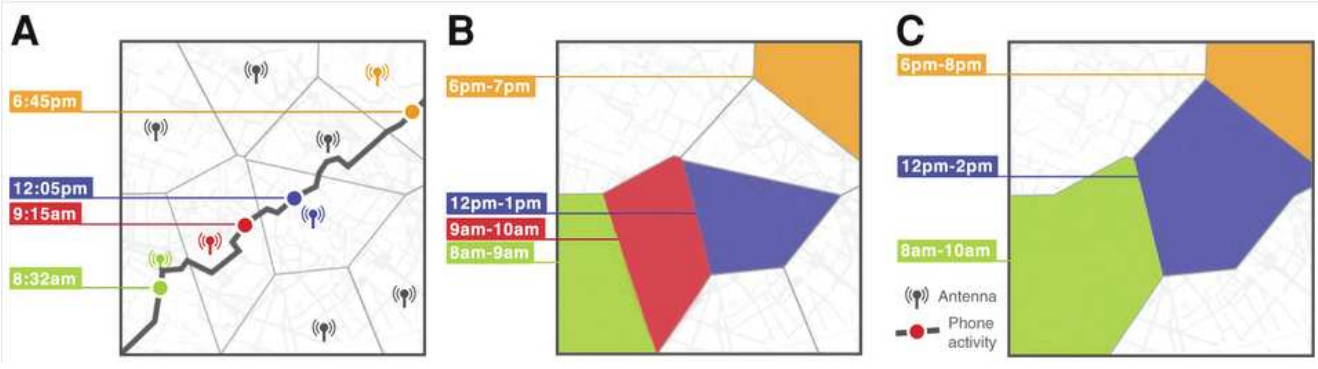
- In a village, we might have only a single male/female customer in a certain age range – how many of our customers are living in a mid-size town in Norway?
  Probably just 1



Your smartphone trail: Location data is very Talkative

- Where do you spend some of your nights? (full night, only short visits)
- Which pubs, clubs do you visit?
- Which sport clubs do you visit?
- Which churches do you visit?
- Which political meeting do you visit?
- Where do you shop?
- What movies do you watch?

Decreasing the resolution of Location Data



- A shows the full tracking details
- B and C show increasing coarseness being applied (location and time)

NY Taxi Rides (2014)

- Dataset released by the New York City Taxi and a Limousine Commission - every taxi ride in New York in 2013,
- including the pickup and drop off times, locations, fare and tip amounts, as well as anonymized (hashed) versions of the taxi's license.

NY Taxi Rides (2014) - "Chasing" Celebrities

- searching through images of "celebrities in taxis in Manhattan in 2013" to find enough information to identify the correct record in the database.

Jessica Alba (Click to Explore)

And the solution for Smart Cities?

1. Collect single data points wherever possible, never full travel data (e.g. for road congestion data)

2. If single data points are not possible, make data very very coarse grained (time + location)

And the solution for Smart Cities?

3. Watch out for uniqueness in the data set! (“quasi identifier”)

4. Hashing without salt allows easy brute force attacks

5. Remove as many data items as possible

And for Consumer

Be very suspicious if somebody promises you to “anonymize your data”

Leave as little traces as possible

Sources for De-anonymization Examples

2000 L. Sweeney study: http://dataprivacylab.org/projects/identifiability/paper1.pdf
2006 AOL search data: http://searchsecurity.techtarget.com/news/1208972/AOL-apologizes-for-exposing-search-data
2006 AOL search data: http://usatoday30.usatoday.com/tech/columnist/andrewkantor/2006-08-17-aol-data_x.htm
2007 Netflix: http://archive.wired.com/politics/security/commentary/securitymatters/2007/12/securitymatters_1213
2007 Netflix: https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
2013 DNA study: http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/
2013 Montjoye mobility study: http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html
2013 Sweeney Patient records: http://dataprivacylab.org/projects/wa/1089-1.pdf
2014 NY taxi rides: http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/ (with tutorial about differential privacy)
2014 Taxi rides NY: http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546
2015 Uber rides: https://web.archive.org/web/20140828024924/http://blog.uber.com/ridesofglory
2015 credit card data: http://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/
2015 credit card data: http://www.sciencemag.org/content/347/6221/468.summary?sid=de616c3b-4b69-416c-b497-1df7cec3d033

Overview (German language)
http://sicherheitskultur.at/Glaeserner_Mensch.htm#deanom

Where to learn more about Pseudonymization?

Much more techncial tutorial from the

ARTICLE 29 DATA PROTECTION WORKING PARTY of the EU

http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Thanks

Philipp Schaumann

philippschaumann@mailbox.org