



Digitala verktyg för riskbedömning och skyddsåtgärder i tabulära data med personuppgifter

Dimitar Popovski & Björn Sjögren

Workshoppresentation, 2024-04-24

Citering:

Swedish National Data Service. (2024). *Digitala verktyg för riskbedömning och skyddsåtgärder i tabulära data med personuppgifter*. Zenodo.
<https://doi.org/10.5281/zenodo.10777054>

Detta verk är licensierat under en
[Creative Commons Erkännande 4.0 Internationell Licens](https://creativecommons.org/licenses/by/4.0/)





Syftet med workshopen

- Förstå vad bakvägsidentifiering är och hur det går till
- Förstå och tillämpa vanliga metoder för att minska risken för bakvägsidentifiering i kvantitativa tabulära data
- Lära sig baskunskaper i ett digitalt verktyg för statistisk röjandekontroll
- Ha förståelse för begränsningar inom statistisk röjandekontroll

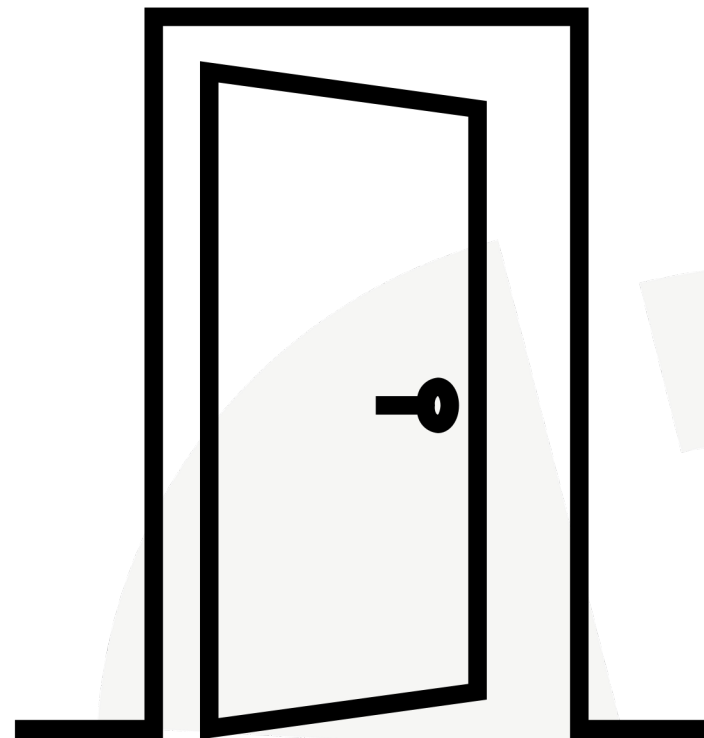


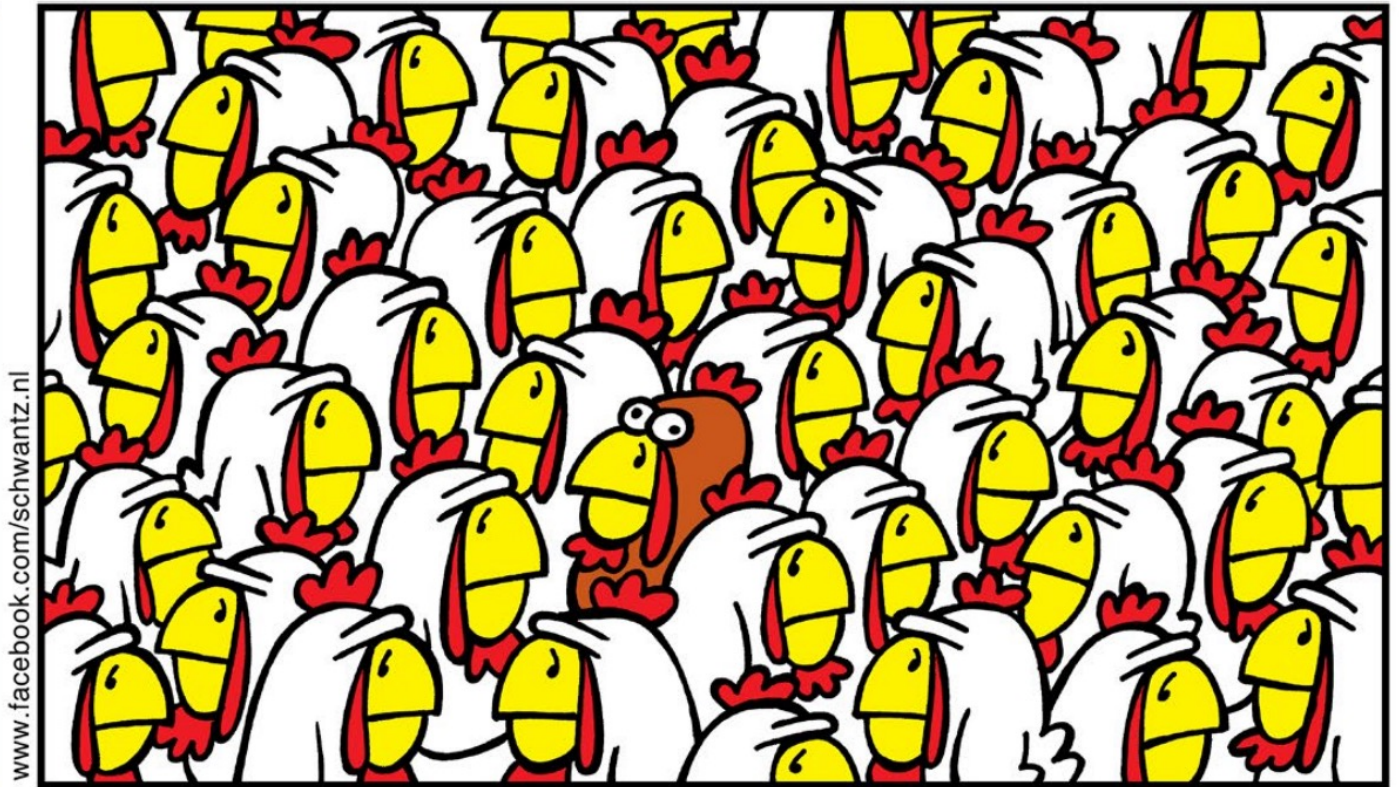
Hålltider

- **20 minuter** – Introduktion till grundläggande begrepp.
- **30 minuter** – Workshopledaren visar funktioner i sdcApp
- **10 minuter** – Gruppindelning, start av sdcApp
- **10 minuter** – Paus
- **50 minuter** – Deltagarna läser instruktioner och pseudonymiserar ett syntetiskt dataset med hjälp av sdcApp
- **30 minuter** – Avslutande genomgång, frågor och diskussion

Bakvägsidentifiering – vad är det?

- Man tittar på unika kombinationer av indirekta identifierare och jämför dem med andra tillgängliga data
- Kompletterande uppgifter i andra datakällor används för att identifiera en enskild person i ett dataset – ”pusselläggande”
- Personer som har unika attribut/egenskaper som sticker ut från mängden löper större risk att bli identifierade



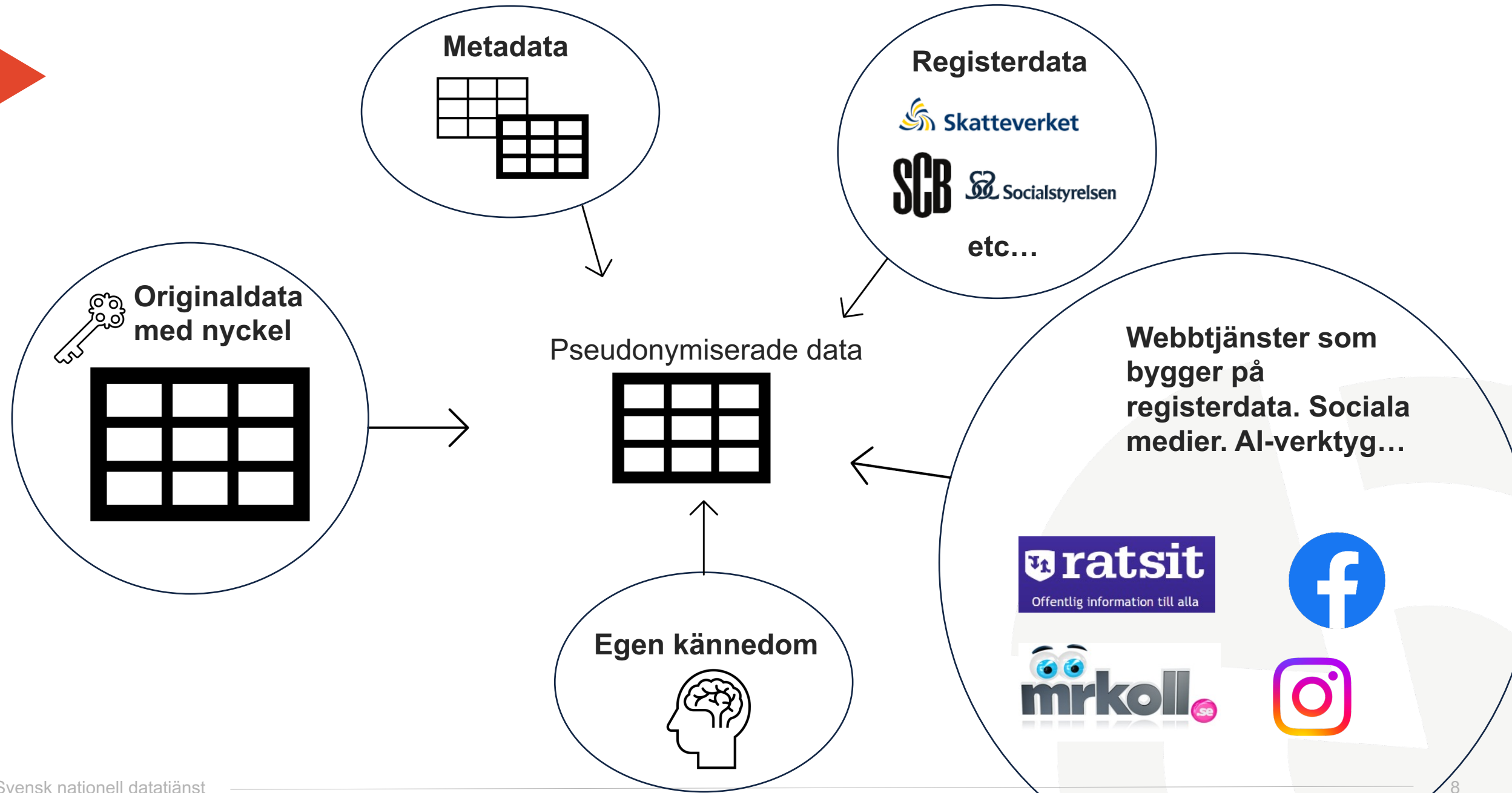


Bakvägsidentifiering – hur går det till?



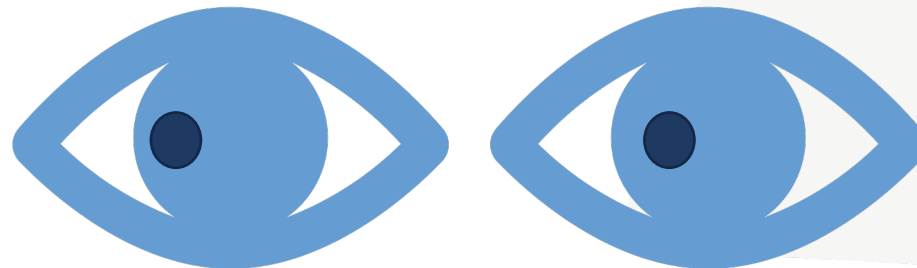
**VEM
DÄR**
?

Emma	Benjamin	Olivia	Henry	Evelyn
Michael	Sophia	Thomas	Sarah	David
Emily	Alex	GUESS WHO	Jack	Mary
Matthew	Karen	Samuel	Ashley	Anthony
Carol	Joseph	Kimberly	Daniel	Dorothy



Några saker att ha koll på

- Kodnycklar och direkta identifierare
- Bakgrundsvariabler som potentiella indirekta identifierare
- Urval
- Typ av studie
- Studiedesign



Riskbedömning (k-anonymity)

- **Statistisk metod** för pseudonymisering av kvantitativa data
- **Mått** på hur svårt det är att bakvägsidentifiera individer i ett enskilt dataset – hur många unika variabelkombinationer finns det?
- **k:et** i k-anonymity står för hur många observationer i ett dataset som har samma unika kombination av egenskaper/attribut (k2, k3 osv)
- Ett av de vanligaste sätten att sänka detaljrikedomen i kvantitativa tabulära data är att minska antalet unika kombinationer (koda om variabler)
- När man kodar om skapar man lite förenklat ”**datatvillingar**”



Exempel

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

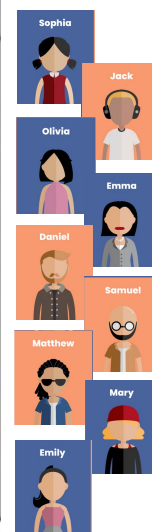


Table 1. Original Patients Table

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



Table 2. A 3-Anonymous Version of Table 1

Räkna på risk – egenskaper inom en grupp (l-diversitet)

- En förlängning av k-anonymitet
- Tar även hänsyn till om individer inom specifika grupper (t.ex. postnr 476** och ålder 2*) har samma värden på andra variabler i datasetet.

	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



Table 2. A 3-Anonymous Version of Table 1

Data vs verkligheten

- Även om ett dataset inte uppfyller en viss nivå av k-anonymitet behöver det inte betyda att det går att identifiera en individ i verkligheten – **och tvärtom.**
- Beror på hur **verkligheten** förhåller sig gällande:
 - Hur många med samma attribut/egenskaper finns i verkligheten?
 - Hur ser tillgången på kompletterande data ut?
 - Hur ska data hanteras?
 - Vem ska hantera data och i vilket syfte?
 - etc...

Verktyg för pseudonymisering och riskbedömning



ARX – öppen källkod-programvara för känsliga personuppgifter.



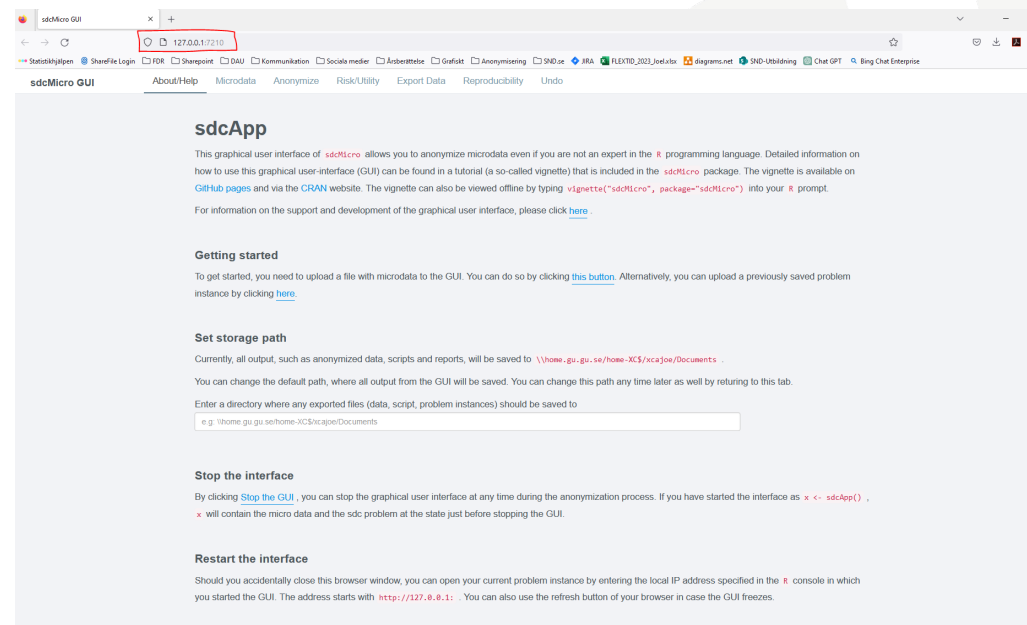
Amnesia Tool från OpenAIRE



R-paketet sdcMicro och gränssnittet sdcApp

sdcMicro och gränssnittet sdcApp

- Ett R-paket
- Väletablerat verktyg
- Används av forskare och internationella organisationer som exempelvis Världsbanken och FN-organet OCHA





Frågor?





Genomgång sdcApp



Kom igång med sdcApp via RStudio

1. Du bör redan ha installerat R och RStudio. Starta RStudio.
2. Ladda in paketet sdcMicro genom att skriva kommandot: `library(sdcMicro)`

Bekräfta med enterslag.

3. Starta sdcApp med kommandot: `sdcApp()`

... sdcMicro GUI bör öppnas i din standardwebbläsare!

Om det inte öppnas automatiskt anger du i din webbläsare webbadressen som du får i R och som har det här formatet:

`http://127.0.0.1: följt av ett fyrsiffrigt nummer`

4. Välj fliken "Microdata" och ladda in datasetet med .csv-importfunktionen genom att bläddra fram filen på din dator i dialogrutan som kommer fram.



Paus





Instruktion för labb

Huvuduppgiften är att skapa ett dataset med så liten risk för bakvägsidentifiering som möjligt, samtidigt som datasetets användbarhet bibehålls i så stor utsträckning som möjligt.

1. Läs avsnittet **Huvuduppgift labb**
2. Läs avsnittet **Studie, data och kontext**
3. Använd avsnittet **Instruktion steg-för-steg**



Frågor att svara på

1. Hur väl lyckades ni med era skyddsåtgärder utifrån den grad av k-anonymitet som ni satt som mål?
2. Hur användbara är de modifierade data ni skapat jämfört med originaldata? Går det fortfarande att återanvända data för den sambandsanalys som specificeras i Studie, data och kontext?
3. Finns det andra lämpliga skyddsåtgärder som skulle kunna komplettera eller ersätta de åtgärder som ni genomfört?



Diskussion och frågor



Slutsatser

Möjligheter:

- Ger relativt snabbt en överblick av potentiella risker i en datamängd.
- Gör det möjligt att relativt enkelt tillämpa metoder för att minska/hantera risk.
- Mycket användbart vid hantering av stora kvantitativa datamängder.

Begränsningar:

- Tar ingen som helst hänsyn till utomliggande kontextuella faktorer. Kan endast förhålla sig till de data du matar in.
- Det mått på risk som du får ut av sdcApp säger inte så mycket om du inte också stämmer av med verkligheten.

Slutsatser:

- Forskningsdata existerar inte i ett vakuum. Det finns alltid en verklighet att ta hänsyn till vad gäller skyddsåtgärder för data med personuppgifter.
- Digitala hjälpmedel är ett av flera verktyg som kan användas för att förstå och förminska risker i data.
- Avvägning mellan risk och nytta.
- Forskningsdata ska inte bli oanvändbara.



Veta mer?

- [Webbinarium från CESSDA](#)
- [University of Utrecht: Data privacy handbook](#)
- [Finnish Social Science Data Archive](#)
- [5 Safes framework](#)

A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.



DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.



PEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.



DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.



ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

This is a primer on how to distinguish different categories of data.

EXPLICITLY PERSONAL POTENTIALLY IDENTIFIABLE NOT READILY IDENTIFIABLE KEY CODED PEUDONYMOUS PROTECTED PEUDONYMOUS DE-IDENTIFIED PROTECTED DE-IDENTIFIED ANONYMOUS AGGREGATED ANONYMOUS



DIRECT IDENTIFIERS
Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)



INDIRECT IDENTIFIERS
Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)



SAFEGUARDS and CONTROLS
Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals

INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation

SELECTED EXAMPLES

Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)

Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)

Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)

Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)

Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)

Same as Pseudonymous, except data are also protected by safeguards and controls

Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)

Same as De-Identified, except data are also protected by safeguards and controls

For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)

Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)