Workshop: Skyddsåtgärder och riskbedömning för data med personuppgifter

2023-11-08

Joel Carlsten Rosberg, André Jernung, Marcus Lundberg, Lars Eklund

Syftet med workshopen

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

Hålltider

- **13.00-13.20** – Introduktion till grundläggande begrepp
- **13.20-13.40** – Genomgång av funktioner i sdcApp
- **13.40-13.50** – Gruppindelning, inlogg och start av sdcApp
- **13.50-14.00** – Paus
- **14.00-15.00** – Deltagarna pseudonymiserar ett syntetiskt dataset med hjälp av sdcApp
- **15.00-15.30** – Avslutande genomgång, frågor och diskussion

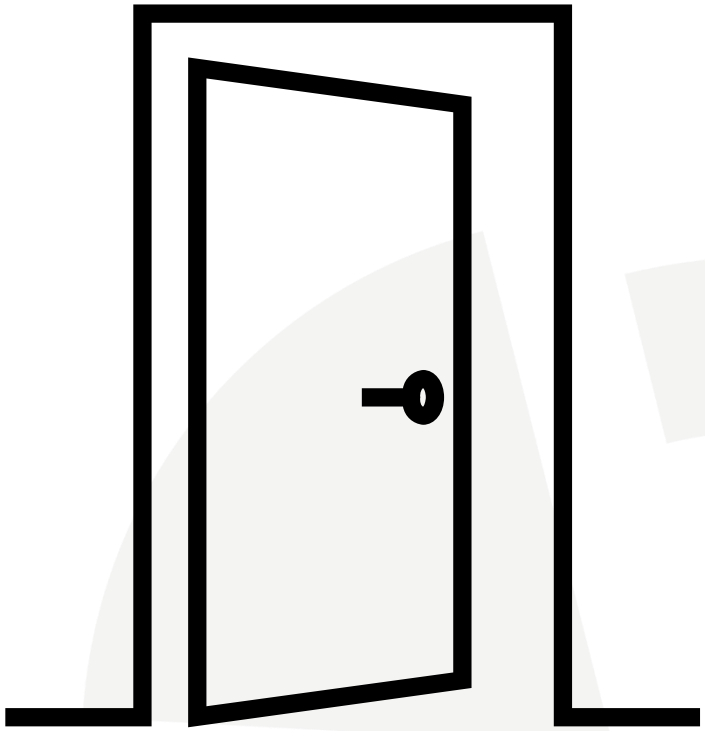
Frågor besvaras efter presentationen.

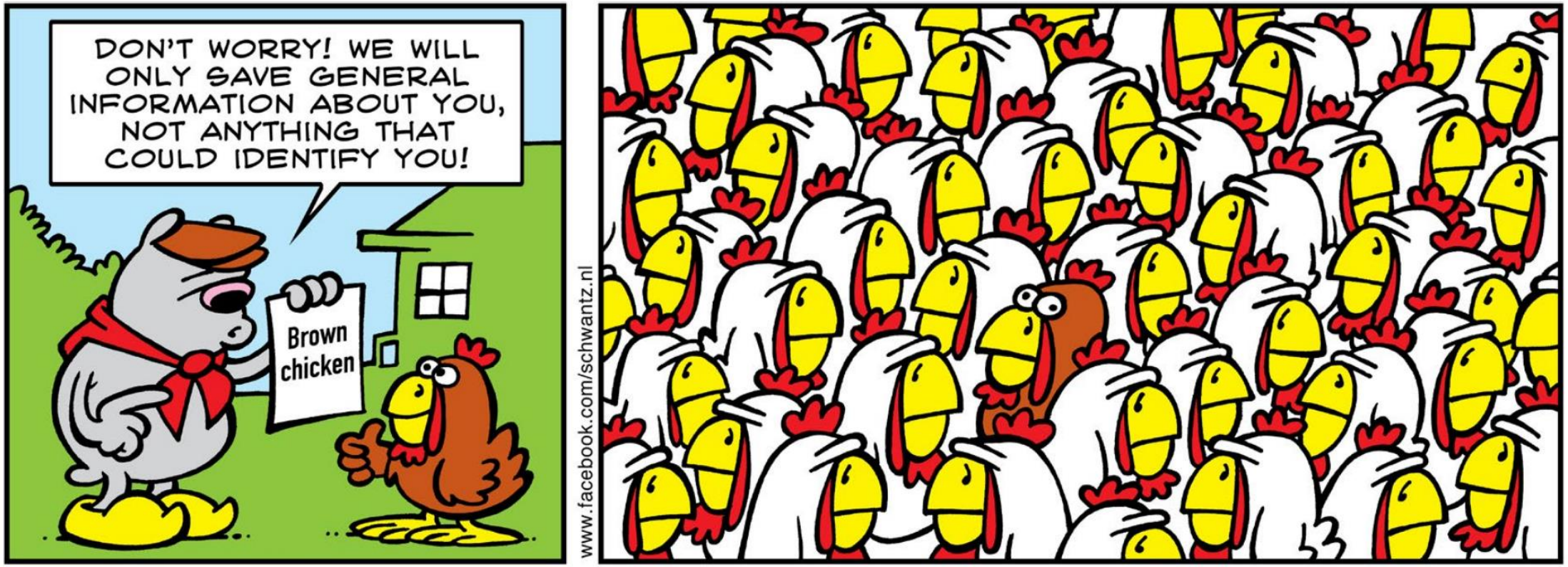
Bakvägsidentifiering – vad är det?

- Kompletterande uppgifter i andra datakällor används för att identifiera en enskild person i ett dataset

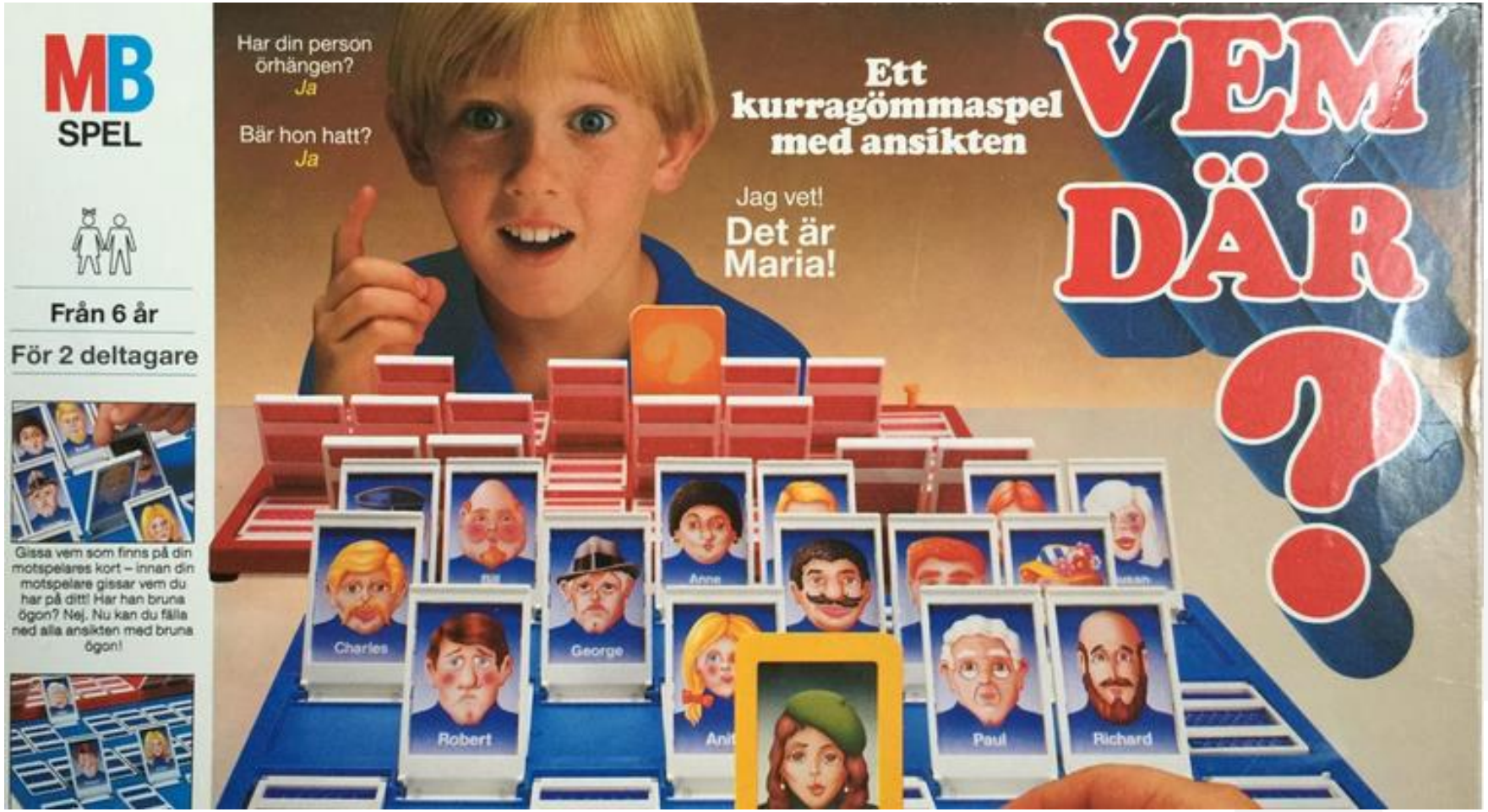
- Man tittar på unika kombinationer av indirekta identifierare och jämför dem med andra tillgängliga data

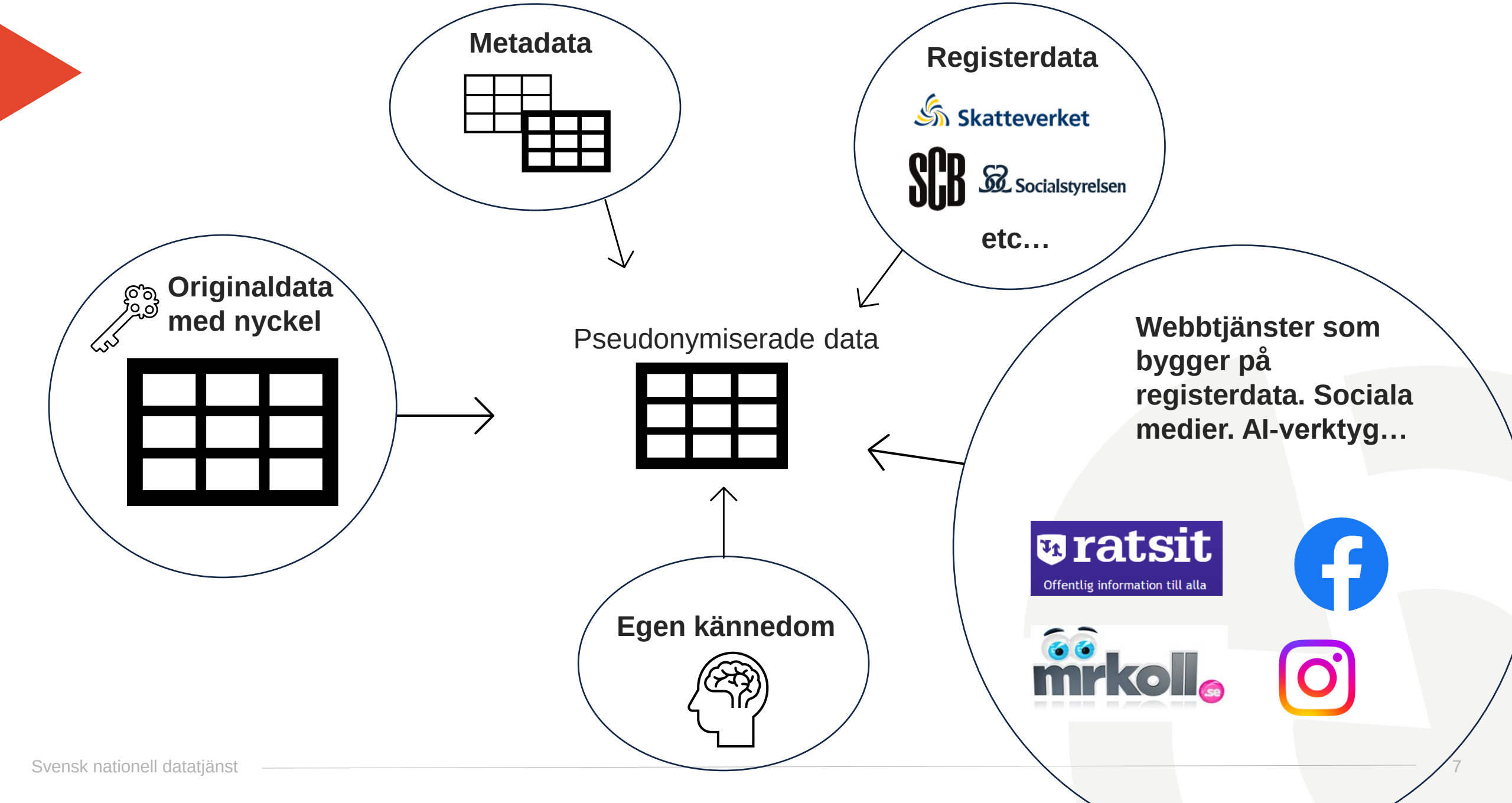
- Personer som har unika attribut/egenskaper sticker ut från mängden löper större risk att bli identifierade




DON'T WORRY! WE WILL ONLY SAVE GENERAL INFORMATION ABOUT YOU, NOT ANYTHING THAT COULD IDENTIFY YOU!
Brown chicken
www.facebook.com/schwantz.nl

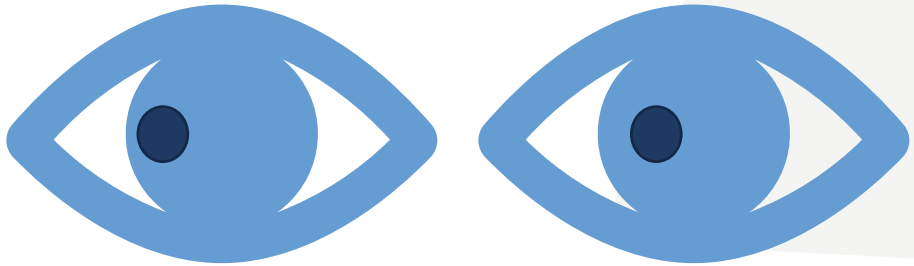
Bakvägsidentifiering – hur går det till?




Metadata
Registerdata
Skatteverket
SCB
Socialstyrelsen
etc…
Originaldata med nyckel
Pseudonymiserade data
Webbtjänster som bygger på registerdata. Sociala medier. AI-verktyg…
ratsit
Offentlig information till alla
mrkoll.se
Egen kännedom

Några saker att ha koll på

- Kodnycklar och direkta identifierare
- Indirekta identifierare och bakgrundsvariabler
- Urval
- Typ av studie
- Studiedesign



Urval

- **Slumpmässigt urval/sannolikhetsurval** → urval som representerar en totalpopulation, exempelvis Sveriges befolkning

➢ Ger oss ganska ospecifik förhandsinformation om personer i data

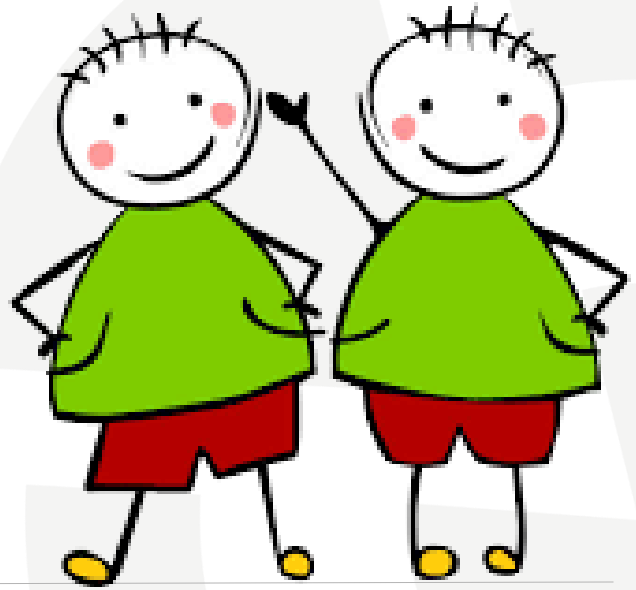
- **Icke-slumpmässigt urval** → används för att undersöka en specifik grupp, exempelvis anställda på Göteborgs universitet, pensionärer, personer med en viss sjukdom etc.

➢ Ger oss förhandsinformation om den grupp som undersöks → kan öka risken för bakvägsidentifiering

Veta mer?

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 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-diversity)

- En förlängning av k-anonymity
- Tar även hänsyn till hur fördelningen av attribut inom specifika grupper ser ut

	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

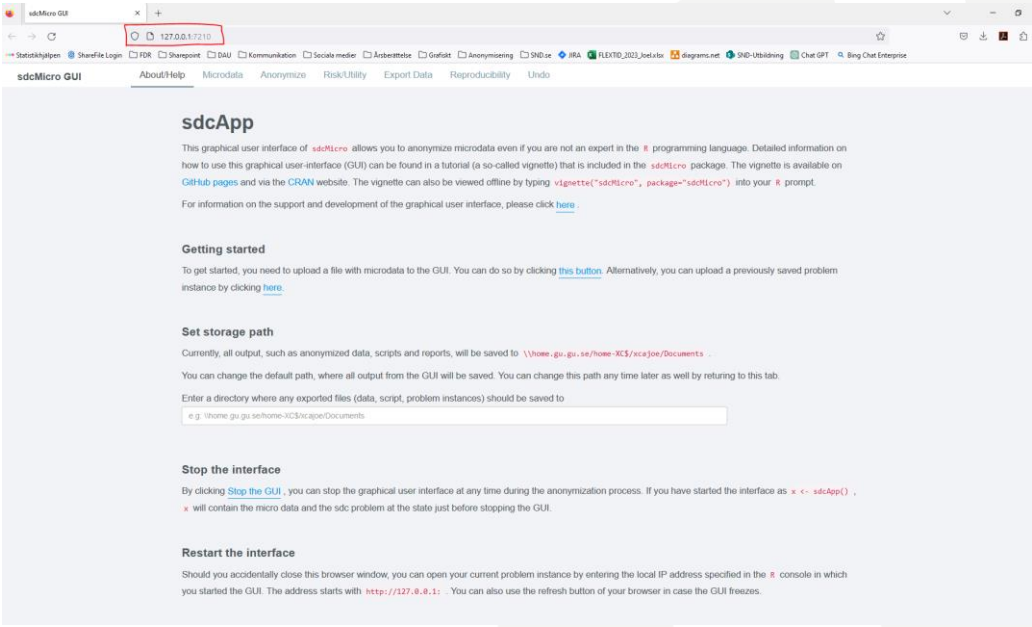
- Även om ett dataset inte uppfyller en viss nivå av k-anonymitet behöver det inte betyda att det går att identifiera någon i verkligheten – **och tvärtom**.

- 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

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

1. Peka en webbläsare till http://130.238.29.100:8787/ och mata in kontonamnet och lösenordet. Logga in.
2. Nu borde ni få upp RStudio Server i ett webbläsarfönster.
3. Ladda in paketet sdcMicro genom att skriva kommandot: `library(sdcMicro)` Kör det med enterslag.
4. Starta sdcApp med kommandot: `sdcApp()`
5. Välj fliken Microdata och ladda in datasetet med .csv-importfunktionen genom att bläddra fram det på din egna dator i dialogrutan som kommer fram.

Paus till 14.00

Instruktion 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 dokumentet **Huvuduppgift labb**

2. Läs dokumentet **Studie, data och kontext**

3. Använd dokumentet **Instruktion steg-för-steg**

Frågor att svara på

1. Hur väl lyckades ni med era skyddsåtgärder utifrån den grad av k-anonymitet 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.

Begränsingar:

- 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 sager inte så mycket utan att stämma av med verkligheten.

Slutsatser:

- Forskningsdata existerar inte i ett vakuum. 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 mitigera 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

Produced by FUTURE OF PRIVACY FORUM
FPF.org

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.

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

DEGREES OF IDENTIFIABILITY
Information containing direct and indirect identifiers.

PSEUDONYMOUS 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.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	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)	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
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	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)