Andmete kaitsmise hea tava riiklikus statistikas ja teadusuuringutes

Statistikanõukogu koosolekul

Tuulikki Sillajõe

12.06.2019

eesti statistika

Ajend

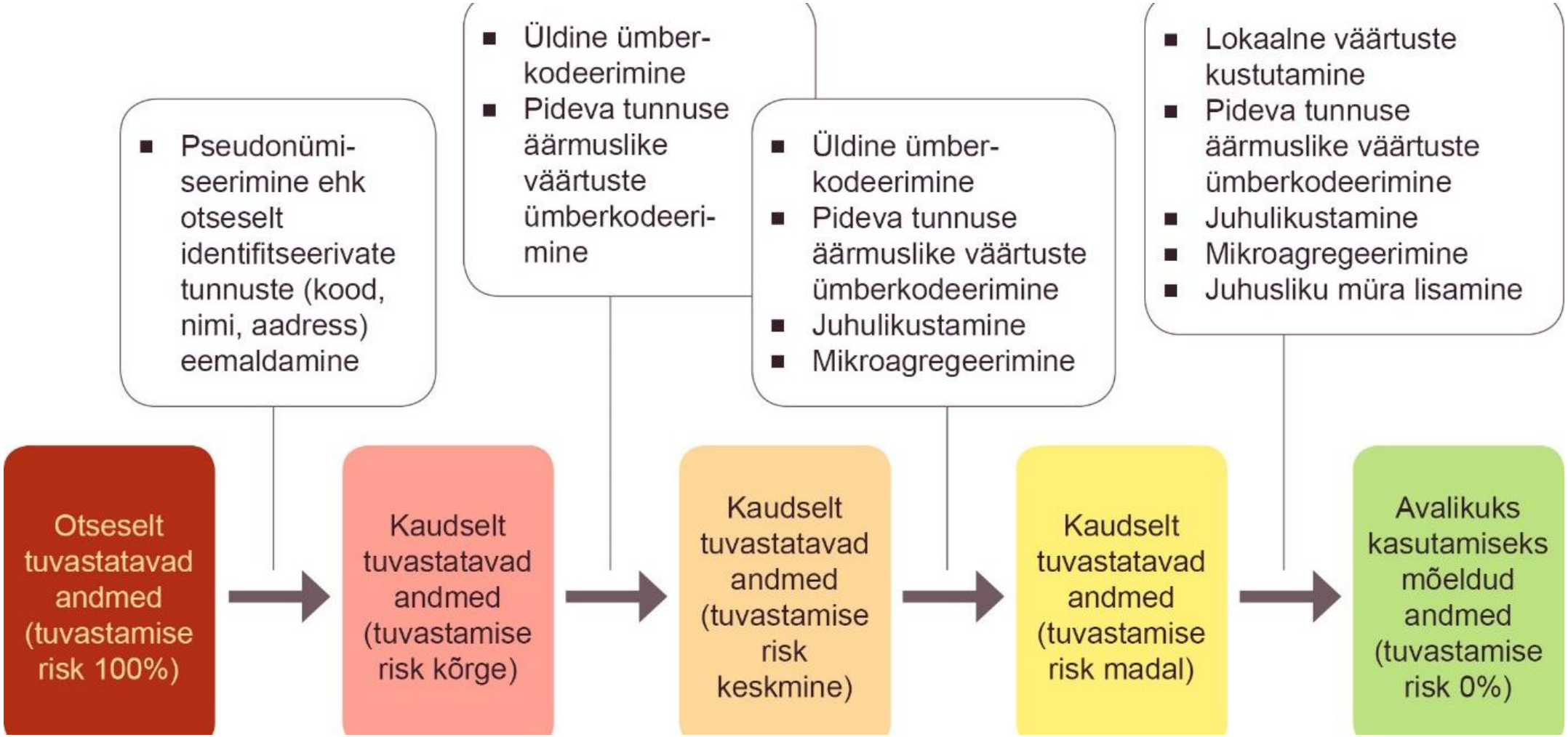
- Nõudlus üksikandmete laiema kasutuse järele, nt mõjuanalüüsides
- Üksteisest aru saamist raskendav mõistete kasutamine eri valdkondades
- GDPR tõi muutuse mõistete kasutuses (enne teadsime mõistet „anonüümimine“ pigem „pseudonüümimise“ tähenduses)
- Cybernetica paber „ Privaatsustehnoloogiate rakendamine andmekaitses“
- JuM-i töögrupp „Vastutustundliku andmetöötluse põhimõtted“

Rakendatud tehnoloogia	Midagi rakendatud ei ole	Isikukood asendatud pseudonüümiga	Reeglipõhine pseudonüümimine	Reeglipõhine anonüümimine	Müraga anonüümimine	Otse isikustavad tunnused eemaldatud	Andmete krüpteerimine	Andmete töötlemata jätmine
	49403136525	6574d303dbf63c66e	6574d303dbf63c66e	6574d303dbf63c66e	6574d303dbf63c66e	6574d303dbf63c66e	6574d303dbf63c66e	-
	Mari Maasikas	Mari Maasikas	Kaire Kartul	Kaire Kartul	Kaire Kartul	-	yJsplhVFt7m3k2Yv	-
	Kraavi 27	Kraavi 27	Kuuse 3	Kuuse 3	Kuuse 5	-	XHDMuUaldYGOEQ3	-
	Põlva 63304	Põlva 63304	Põlva 63304	Põlva 63xxx	Põlva 63303	-	NFDtEnjlxa20yq4w	-
	13. märts 1994	13. märts 1994	13. märts 1994	1990-1995	28. aprill 1993	-	rpxMCgvFPAIMViLQ	-
	55 123 9876	55 123 9876	55 555 0001	55 555 0001	55 555 0001	-	VEgTgTSecCVUvTV0	-
	mari@maasik.as	mari@maasik.as	kairekartul@e-po.st	kairekartul@e-po.st	kairekartul@e-po.st	-	UU95zrEoWppl9llJ	-
	9:00 - Põlva	9:00 - Põlva	9:00 - Põlva	9:00 - Põlva	9:00 - Põlva	9:00 - Põlva	fb7gLGLtpyhj1uaR	-
	9:15 - Valgemetsa	9:15 - Valgemetsa	9:15 - Valgemetsa	9:15 - Valgemetsa	9:15 - Valgemetsa	9:15 - Valgemetsa	QlYTt5wMYnGXK3P	-
	9:30 - Ülenurme	9:30 - Ülenurme	9:30 - Ülenurme	9:30 - Ülenurme	9:30 - Ülenurme	9:30 - Ülenurme	SJWnMDWb72atfavh	-
	9:45 - Tartu	9:45 - Tartu	9:45 - Tartu	9:45 - Tartu	9:45 - Tartu	9:45 - Tartu	ebxbmXsiJOluFX5O	-
Isikustamise võimalus	Otseselt, isikut määravate kirjete kaudu	Otseselt, isikut määravate kirjete kaudu	Kaudselt, väliste andmestike toel	Kaudselt, liikumismustrite järgi, väliste andmestike toel	Kaudselt, liikumismustrite järgi, väliste andmestike toel	Kaudselt, liikumismustrite järgi, väliste andmestike toel	Võimatu, kui pole ligipääsu võtmele	Võimatu
Töötlemise vahendid	Tavapärased vahendid	Tavapärased vahendid	Tavapärased vahendid	Tavapärased vahendid	Tavapärased vahendid	Tavapärased vahendid	Privaatsustehnoloogiad võimaldavad töödelda ilma võtmele ligipääsuta (turvaline ühisarvutus või turvalised käivituskeskkonnad)	Tuleb töödelda mõnes teises süsteemis

Cybernetica paberist „Privaatsustehnoloogiate rakendamine andmekaitses“

- Tööjõu-uuringu näide

Üksikandmestikes andmesubjekti aimatavuse vähendamise meetodid ja aimatavuse riski tasemed



A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by FUTURE OF PRIVACY FORUM FPF.ORG

In collaboration with EY

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, record numbe cookie, IP add (e.g., MAC add 68:A8:6D:35:6	Same as Potentially	Clinical or research	Unique, artificial	Same as Pseudonymous,	Data are suppressed, GPA:	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)

Kommenteerimiseks saadetud

- Tallinna Ülikool
- Tartu Ülikooli mobiilsusuuringute labor
- Tartu Ülikooli inimgeograafia ja regionaalplaneerimise õppetool
- TalTechi IT-teaduskond
- Cybernetica
- Positium
- Eesti Pank
- Majandus- ja Kommunikatsiooni-minsiteerium
- Sotsiaalministeerium
- Tervise Arengu Instituut

Kokkuvõte

- pseudonüümimine ≠ anonüümimine & umbisikustamine
- anonüümimine = umbisikustamine = aimatavuse vähendamise meetodite rakendamine
- Anonüümsete üksikandmete kasutusvõimalused teadusuuringuteks ja mõju analüüsideks on piiratud
- Eesmärk on üksikandmete kasutust suurendada, sh nii SA uuringutega kogutud, registritest saadavad jms nii teadusuuringuteks kui ka mõjuanalüüsideks

Järgmised sammud „Hea tavaga“

- Täiendame mõisteid (koos definitsioonidega)
- Lisame juurde näiteid
- Lisame eri tasemele vähendatud aimatavusega andmestike kasutamisel rakendatavate kaitse meetmete hea tava
- Arutame asjassepuutuvatega

Andmejagamisteenuse laiendamine

- Teadusuuringuteks üksikandmete kasutada andmise olemasoleva teenuse re-disain ja riikliku statistika seaduse muudatusest tuleneva andmejagamisteenuse disain
- Disaimimeeskonnas lisaks SA esindajatele (5) ka RM ja JuM esindaja
- Disainimeeskond osaleb Avalike Teenuste Arendamise Programmis (Praxis ja Velvet)
- Intervjueeritakse ca 30 teenuse kasutajat
- Õppekäik Taani ja Norra statistikaametitesse

The 'five safes' framework for the IDI and LBD

1. **Safe people** – researchers can be trusted to use data appropriately and follow procedures.
   Researchers must pass referee checks before we allow them to work with data. We require them to sign a declaration of secrecy under the Statistics Act 1975 and follow our rules and protocols. Researchers who break our protocols can be banned, blacklisted, or prosecuted.
2. **Safe projects** – the project has a statistical purpose and is in the public interest.
   Research is restricted to the analysis of groups, not individuals, and must be in the public interest. This means that the research is focused on finding solutions to issues that are likely to have a wide public benefit. The Government Statistician or delegated authorised person signs off all research proposals.
3. **Safe settings** – security arrangements prevent unauthorised access to the data.
   Data can only be accessed through a secure Data Lab environment. Computers are not connected network and only Stats NZ staff can release data to researchers.
4. **Safe data** – the data inherently limits the risk of disclosure.
   We de-identify data, which means we remove personal identifying information such as names a addresses, and encrypt (ie replace with another number) identifiers such as IRD and NHI n our de-identified data fact sheet for more information about the benefits, risks, and possibl identified data. Researchers in the IDI get access only to the data relating to their research; rese the LBD get access to all LBD data.
5. **Safe output** – the statistical results produced do not contain any identifying results.
   Researchers must confidentialise output before it can be released from the Data Lab, and Stats NZ st double-check results to ensure individuals cannot be identified. See Microdata output guide for the methods and rules researchers must use for confidentialising output produced from Stats NZ's microdata.

1. Konfidentsiaalsus-kohustus

2. Leping asutusega, avaliku huvi olemasolu

3.Teadlaste infosüsteemsüsteem

4. Pseudonüümitud ja ainult uurimiseesmärgi täitmiseks vajalikud andmed

5. Tulemite aimatavuse kontroll

Uus-Meremaa põhimõtted üksikandmete kasutamisel

Tänan tähelepanu eest!

tuulikki.sillajoe@stat.ee

eesti statistika