Aalto University

Personal data (pseudo-)anonymisation

Enrico Glerean, Staff Scientist/Data Agent, @eglerean

27/10/2020

Outline of today's workshop

1. What is (pseudo-)anonymisation and why is it important?
2. Re-identification examples
3. Anonymisation before and after data collection
4. Workflows when anonymisation is not possible
5. Anonymisation, ethics, and open science

Let’s start with the references

Where to read and learn

Great amount of resources at:
https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html

Page about handling personal data at Aalto:
https://www.aalto.fi/en/services/how-to-handle-personal-data-in-research

When in doubt, ask! researchdata@aalto.fi or your dept. legal advisor or data agent

1.1 Concepts and definitions

Context

- We consider anonymisation in the context of **research data with human participants,** i.e. **personal data**.
- Personal data is a broad concept under the **EU's General Data Protection Regulation (GDPR)**
- **"Personal data" is <u>any data about living people from which they can be identified</u>**
  - If you collect <u>information **from** or **of** persons</u>, consider it as **personal data**

Direct and indirect identifiers

a) information which is sufficient on its own to identify an individual ("direct identifiers"):

- e.g. a person's name, email address (containing the person's name), personal identification number, fingerprints, a facial image, a person's voice, video, brain scan images, dental records, DNA

b) information which can be used to identify an individual fairly easily ("strong indirect identifiers"):

- e.g. a postal address, a phone number, a vehicle registration number, bibliographic citation of a publication, an email address not in the form of the personal name, an unusual job title, a very rare disease, a position held by only one person at a time (for example the managing director of a named company), a student ID number, a bank account number, IP address of a computer

- location data, online identifiers, such as internet protocol addresses and cookie identifiers, and other identifiers, such as radio frequency identification tags.

Direct and indirect identifiers

c) information that on its own is not enough to identify someone but, when linked with other available information, could be used to deduce the identity of a person ("indirect identifiers"):

- e.g. age, gender, education, status in employment, economic activity and occupational status, socio-economic status, household composition, income, marital status, mother tongue, ethnic background, place of work or study, postal code, municipality, major region.

Techniques to reduce risk of processing to the data subject: pseudonymisation and anonymisation

Pseudonymisation :

- **removal or replacement of identifiers with pseudonyms or codes, which are kept separately and protected by technical and organisational measures**

- **the data are pseudonymous (and hence personal data) as long as the additional identifying information exists**

Techniques to reduce risk of processing to the data subject: pseudonymisation and anonymisation

Anonymisation:

- to anonymise personal data means to irreversibly remove identifying information from the data so that a person cannot be identified based on the data
- all the means "reasonably likely" to be used for the identification of individuals must be considered when assessing whether the data has been anonymised
- information available from other data sources shall also be taken into account, when considering if persons are reasonably likely to be identified

1.2 Why do we need to anonymise personal data in research?

Why anonymisation?

- **Legal requirements of GDPR**, as we should not store personal data indefinitely
  - https://gdpr.eu/recital-26-not-applicable-to-anonymous-data/
  - https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
- **Ethical considerations**
  - "personal data must be removed from research data when it is no longer necessary in order to carry out the research"
  - Protecting privacy in research publications

https://tenk.fi/sites/tenk.fi/files/Ihmistieteiden_eettisen_ennakkoarvioinnin_ohje_2019.pdf

1.3 Other concepts

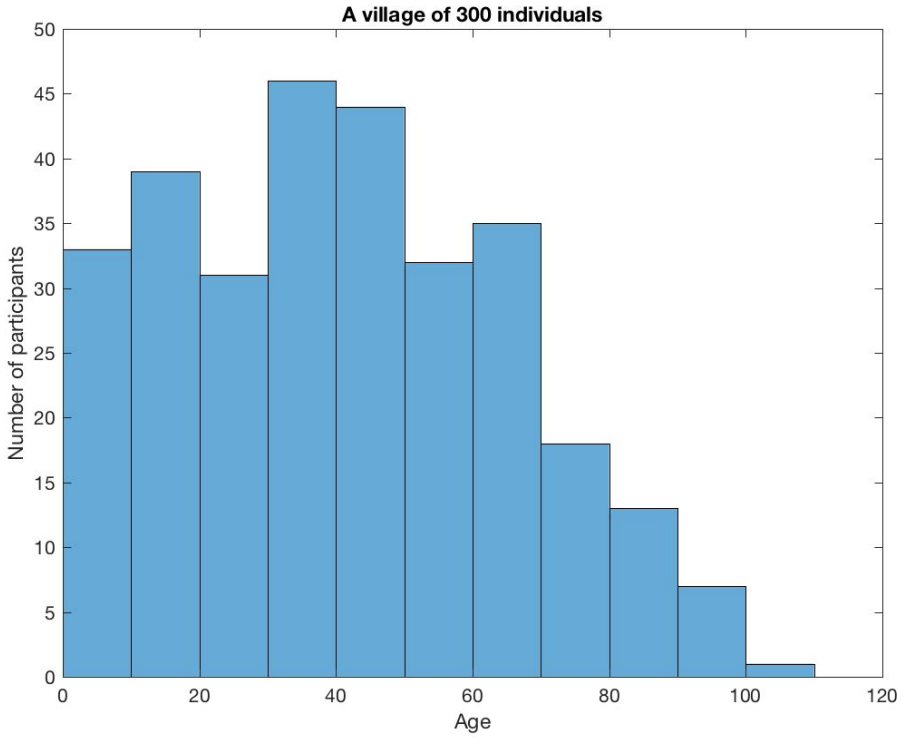
Other concepts

- **De-identification**
  - De-identification often refers to the process of removing or obscuring direct identifiers (Elliot et al. 2016).
- **De-anonymisation or Re-identification**
  - Recovering identities of individuals from an anonymised dataset, because technology has advanced, or more information on the individuals has become available elsewhere.
- **Minimisation:**
  - Only the minimum amount of personal data necessary to accomplish a task (e.g. research) should be collected. Personal data must not be collected just in case they might be useful in the future. There has to be a clear, specified need for collecting the personal data.

https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers /

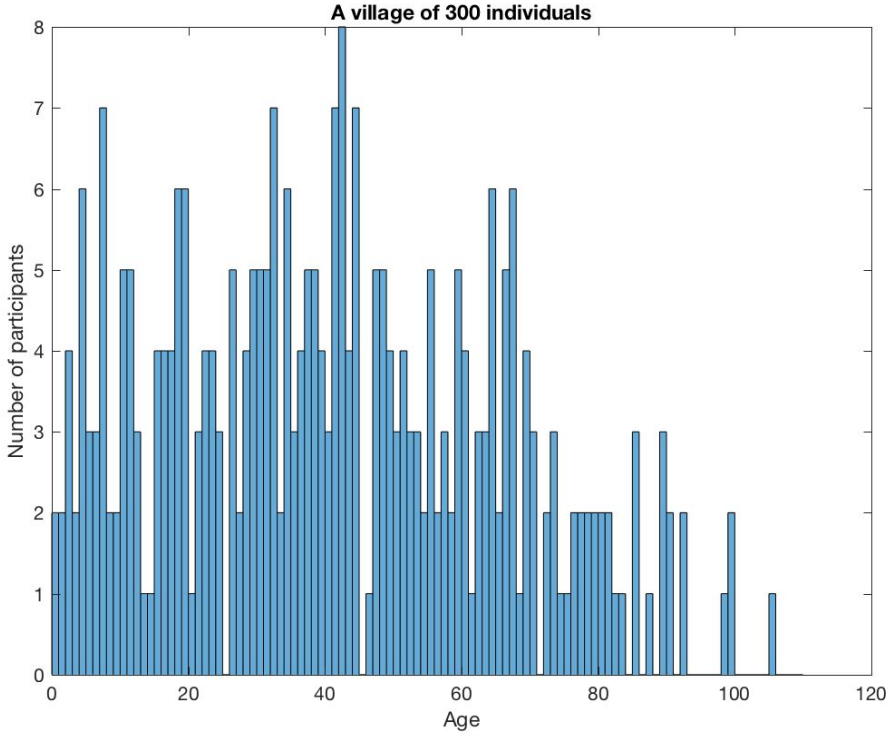
2. Understanding anonymisation through re-identification

Singling out – a small village with 300 individuals



- People counted together based on their “decade” age
- We can single out one very old individual which most likely everybody in the village knows

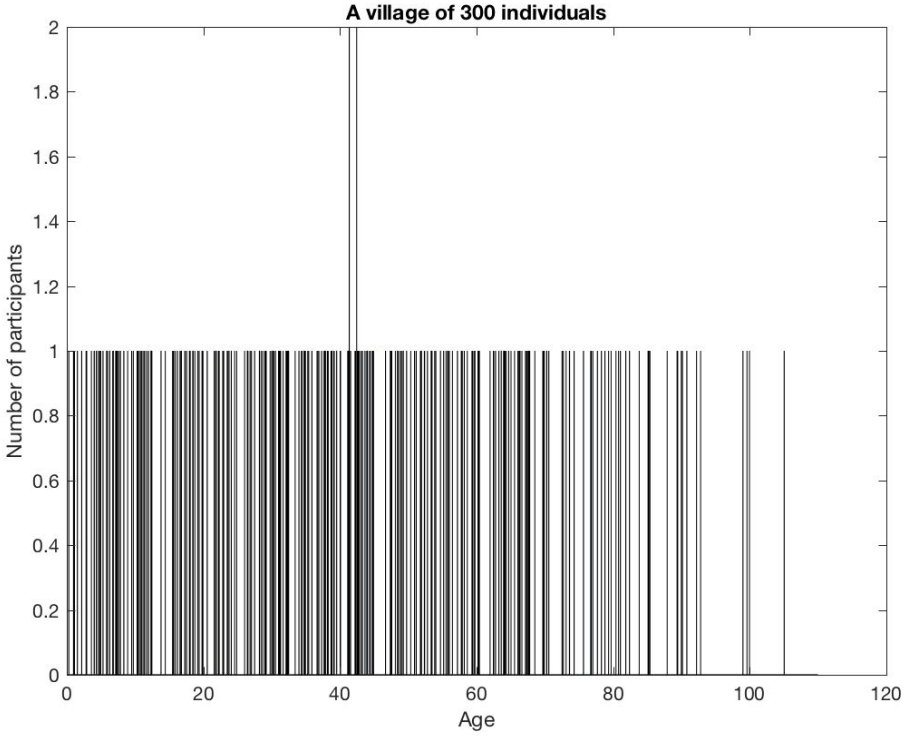
Singling out – a small village with 300 individuals



- People counted together based on their age as integer year
- We can single out more than one individual

Increasing granularity of data, makes the subjects more identifiable.

Singling out – a small village with 300 individuals

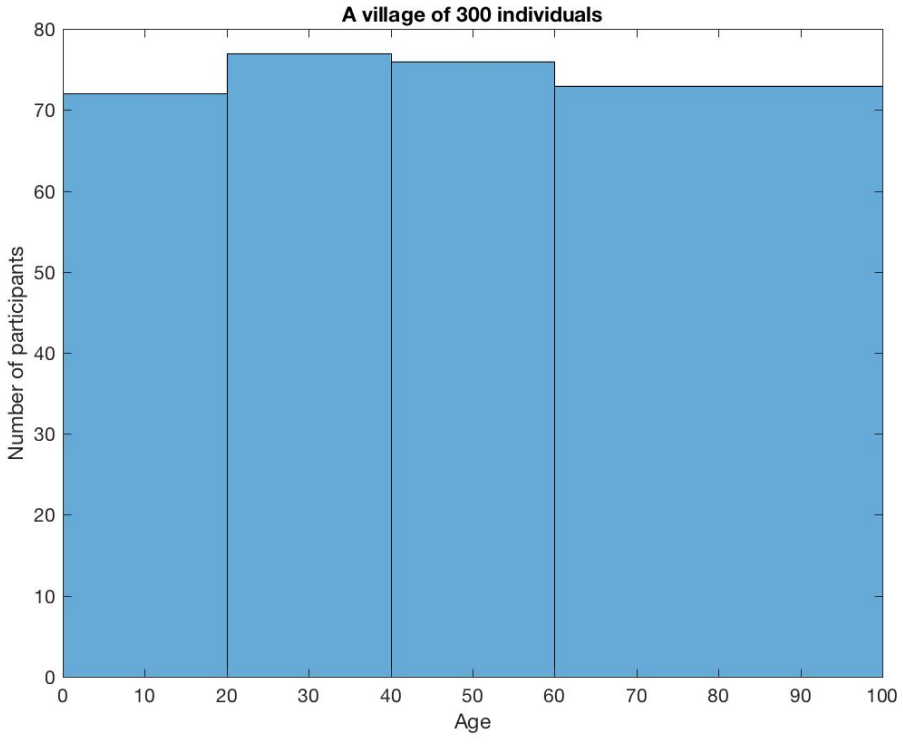


- People counted together based on their age including day and time of birth
- We can single out basically every individual

Increasing granularity of data, makes the subjects more identifiable.

K-anonymity = 1

Solution: binning the data into uniform sub-groups



- Groups that are less represented should be merged together

This anonymisation method is called **generalisation**

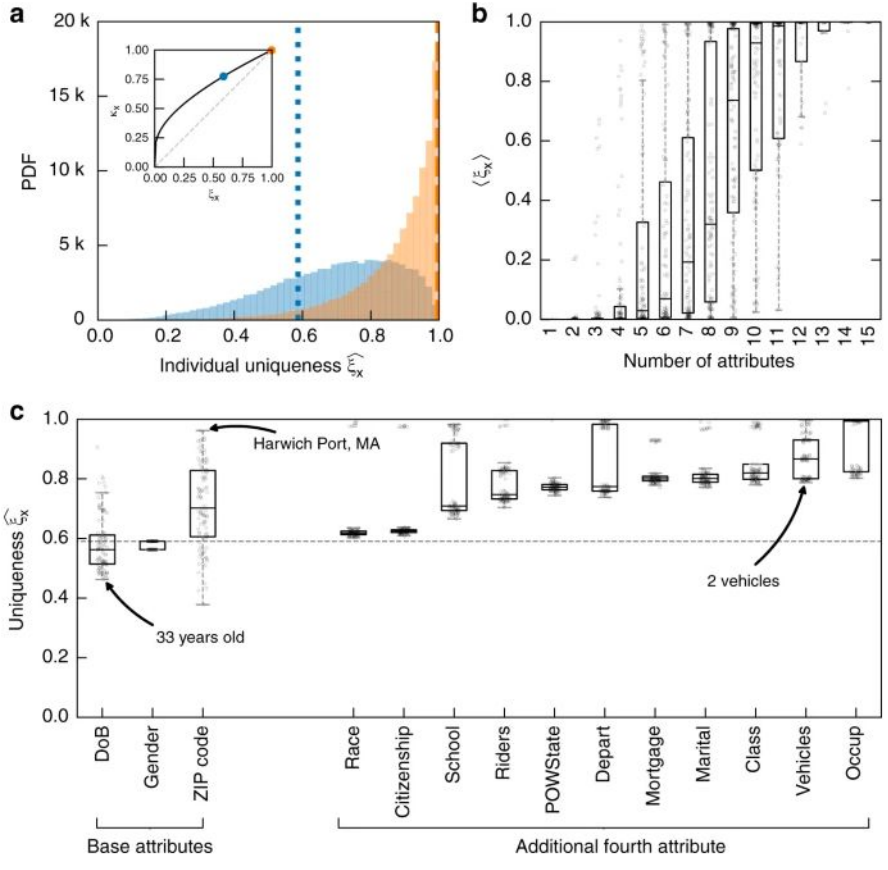
K-anonymity =~ 70

Each person contained in the release cannot be distinguished from at least k - 1 individuals whose information also appear in the release.

Singling out – more data, lower k-anonymity

- With ~4 pieces of information you can uniquely identify a character from “guess who?” board game
- Even though each piece of information could be generalised into a broad category, more information makes the individual more identifiable.

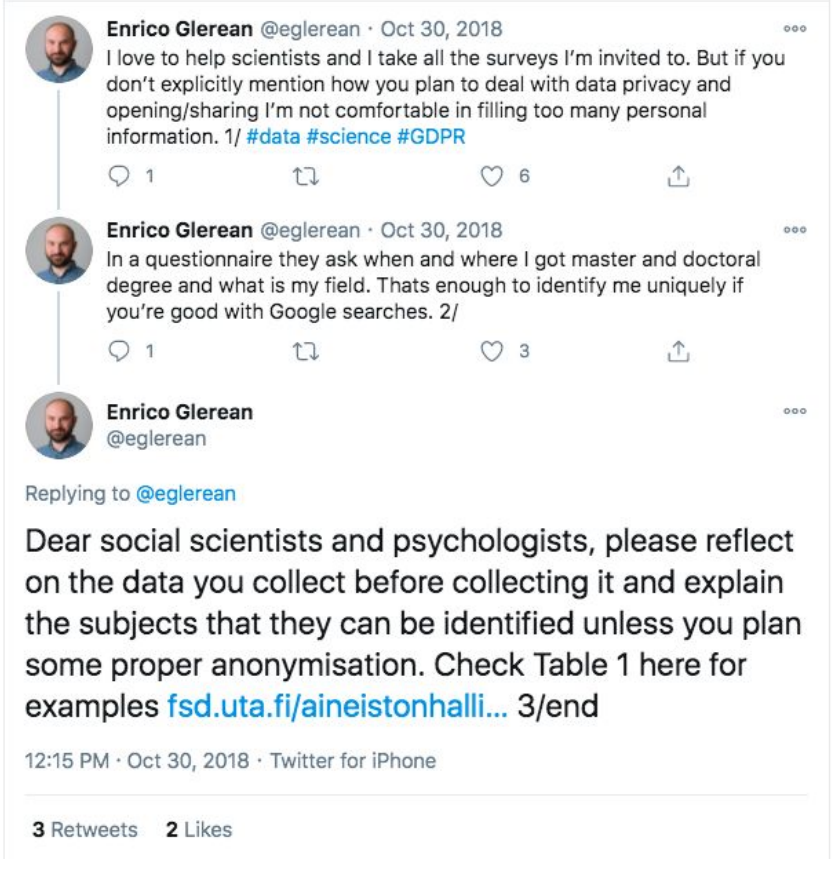
Singling out – more data, lower k-anonymity



- With ~15 pieces of information you can uniquely identify 99.98% of the American population.

"Estimating the success of re-identifications in incomplete datasets using generative models" Rocher et al (2019)
https://www.nature.com/articles/s41467-019-10933-3

Play the “guess who” game with your data



Enrico Glerean @eglerean · Oct 30, 2018

I love to help scientists and I take all the surveys I’m invited to. But if you don’t explicitly mention how you plan to deal with data privacy and opening/sharing I’m not comfortable in filling too many personal information. 1/ #data #science #GDPR

1 6

Enrico Glerean @eglerean · Oct 30, 2018

In a questionnaire they ask when and where I got master and doctoral degree and what is my field. Thats enough to identify me uniquely if you’re good with Google searches. 2/

1 3

Enrico Glerean
@eglerean

Replying to @eglerean

Dear social scientists and psychologists, please reflect on the data you collect before collecting it and explain the subjects that they can be identified unless you plan some proper anonymisation. Check Table 1 here for examples fsd.uta.fi/aineistonhalli... 3/end

12:15 PM · Oct 30, 2018 · Twitter for iPhone

3 Retweets **2** Likes

- **Do not promise anonymity if you are not mathematically sure about it**
- **Do not promise that you will open the data, if you cannot guarantee the privacy of your participants**

Identifier type	Direct identifier	Strong indirect identifier	Indirect identifier	Anonymisation method
Personal identification number	x			Remove
Full name	x			Remove/Change
Email address	x	x		Remove
Phone number		x		Remove
Postal code			x	Remove/Categorise
District/part of town			x	Categorise
Municipality of residence			x	Categorise
Region			x	(Categorise)
Major region			x	
Municipality type			x	
Audio file	x			Remove
Video file displaying person(s)	x			Remove
Photograph of person(s)	x			Remove
Year of birth		x		Categorise
Age			x	Categorise
Gender			x	
Marital status				

https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html

Re-identification from fingerprinting

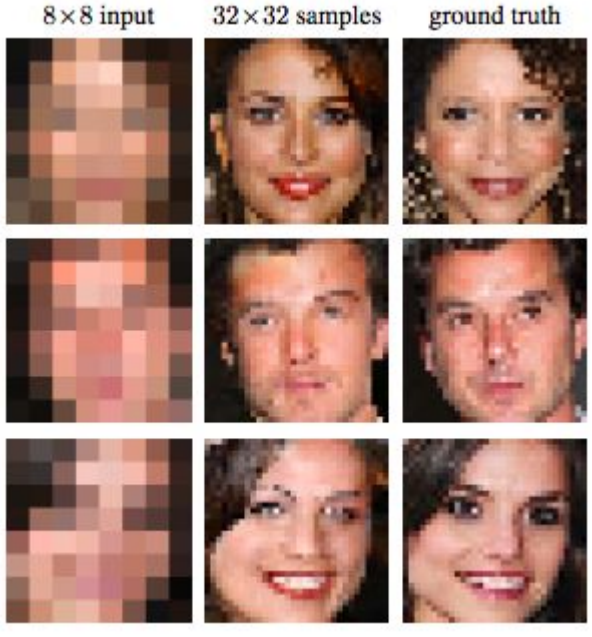
- **When you start collecting more rich data from individuals** (basically anything related to their body: fingerprints, DNA, eye movements, brain activity, brain morphology, electrocardiogram, gait, voice, face traits) **anonymisation becomes impossible, unless you want to make the data unusable.**
- **Pseudo-anonymisation is still necessary** (e.g. removing direct identifiers), and data need to be handle as personal data

Re-identification through synthetic data

Advances in machine learning technology can potentially allow re-identification of de-identified data.

- **Reconstructing faces from low quality images:**

  https://arxiv.org/pdf/1702.00783.pdf



Re-identification through synthetic data

Advances in machine learning technology can potentially allow re-identification of de-identified data.

- **Reconstructing faces from defaced MRIs of the head.**

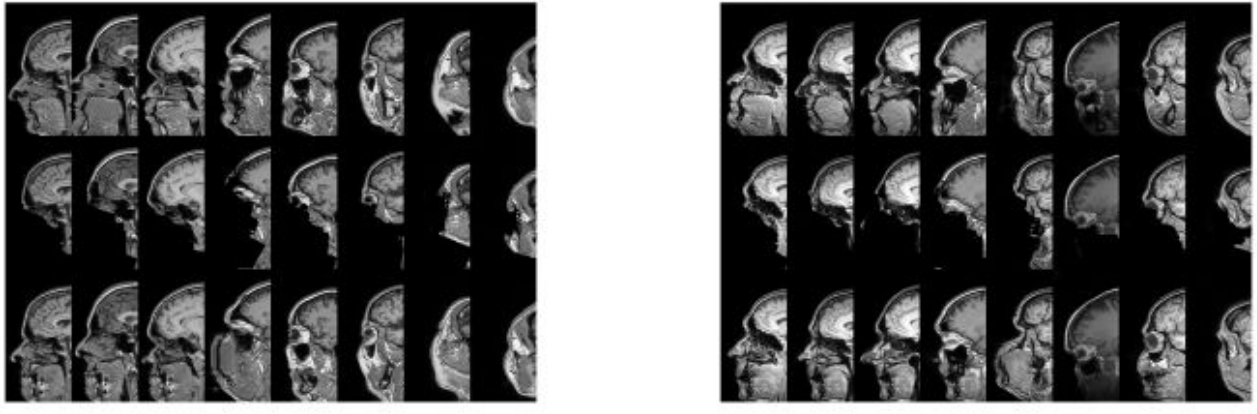


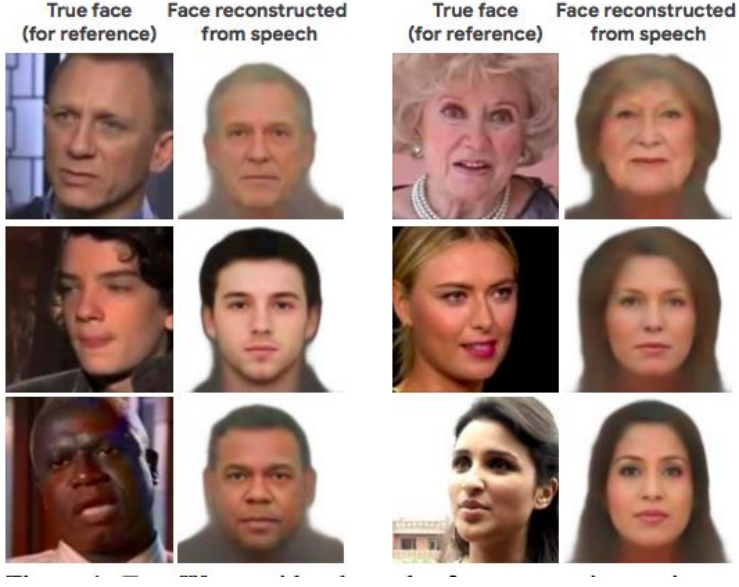
Fig. 2. Typical results of refacing face-removed images. Left: results for training using only subjects from Guy's hospital, Right: results for training using data from all 3 sites. Top row: original image, middle row: face-removed image, bottom row: reconstructed image. CycleGAN learns to add a face, but in many cases it is not the correct face.

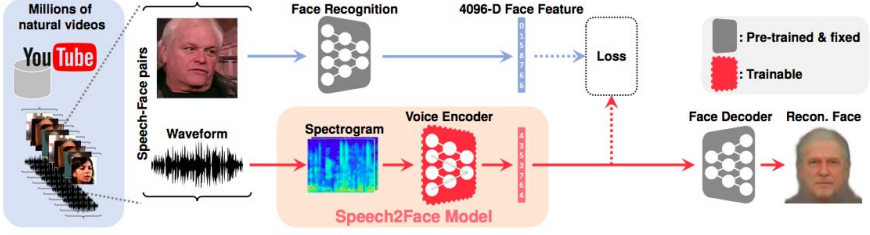
http://www.diva-portal.org/smash/get/diva2:1359959/FULLTEXT01.pdf

Re-identification through synthetic data

Advances in machine learning technology can potentially allow re-identification of de-identified data.

- **Reconstructing faces from voice recordings**





https://openaccess.thecvf.com/content_CVPR_2019/papers/Oh_Speech2Face_Learning_the_Face_Behind_a_Voice_CVPR_2019_paper.pdf

Other re-identification situations

- **When studying a rare population, there are higher chances for the participants to be re-identified**

  Individuals with rare diseases, individuals that are famous (musicians, celebrities, politicians)

3. (Pseudo-) anonymisation in practice

Anonymisation before data collection: Privacy by design

- Plan your research with **minimisation** in mind
  - What is the minimum amount of personal data that I need to answer my research question?
  - Justify your data/protocol whether you have an hypothesis or explicitly mention that it is exploratory (preregistration and DMP)
  - Structured data in favour of unstructured data (i.e. avoid open ended questions)
- **Data Management Plan** should mention about your anonymisation strategy.
- **Consider also the ethical aspects:** e.g. you might collect health data and come across an *incidental finding*, you still need to be able to contact the participant before pseudo-anonymising the data

Anonymisation after data collection

- **Participants' background information**: Destroy, obfuscate, generalise background variables according to the table at https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/
- **Health data**: de-identify file headers, remove direct identifiers (e.g. faces from MRI)
- **Geospatial data:** https://www.sciencedirect.com/science/article/pii/S0198971520302465#f0010
- **…**
- **What is your data?**

Anonymisation after data collection: some tools

- https://amnesia.openaire.eu/ (tabular data)
- https://arx.deidentifier.org/ (tabular data)
- https://sourceforge.net/projects/anony-toolkit/ (tabular data)
- https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface (MRI)
- http://mist-deid.sourceforge.net/ (unstructured medical records)
- https://nlp.stanford.edu/software/CRF-NER.html (NLP for unstructured text)

4. Workflows when anonymisation is not possible

When anonymisation is not possible

- **Adopt secure worflows**
- **Keep the data in the safest place**
- **Bring the code to your data**
  a. Code and data on same system
  b. Federated analysis approach
- **Bring (part of) the data to your code**
  a. Subpart of the data e.g. Beacons in genomics (https://www.nature.com/articles/s41587-019-0046-x)
  b. Synthetic data with same statistical properties of the data you work with (https://arxiv.org/abs/1912.04439)

Data and software on the same organisation system

- **After minimisation** you want to analyse your data in a secure way -> **personal data should not be stored in a location accessible to others**
- **Aalto workflows for data analysis:** https://scicomp.aalto.fi/triton/usage/workflows.html
  a. Use remote computing
  b. VDI https://vdi.aalto.fi
  c. Triton high performance computing cluster via Jupyterhub or slurm batch scripting
- **Higher risks -> higher security. CSC ePouta** https://research.csc.fi/epouta
- **In doubt? Talk to us (SciComp Garage and other tools listed at** https://scicomp.aalto.fi/triton/help.html**)**

Federated analysis approaches

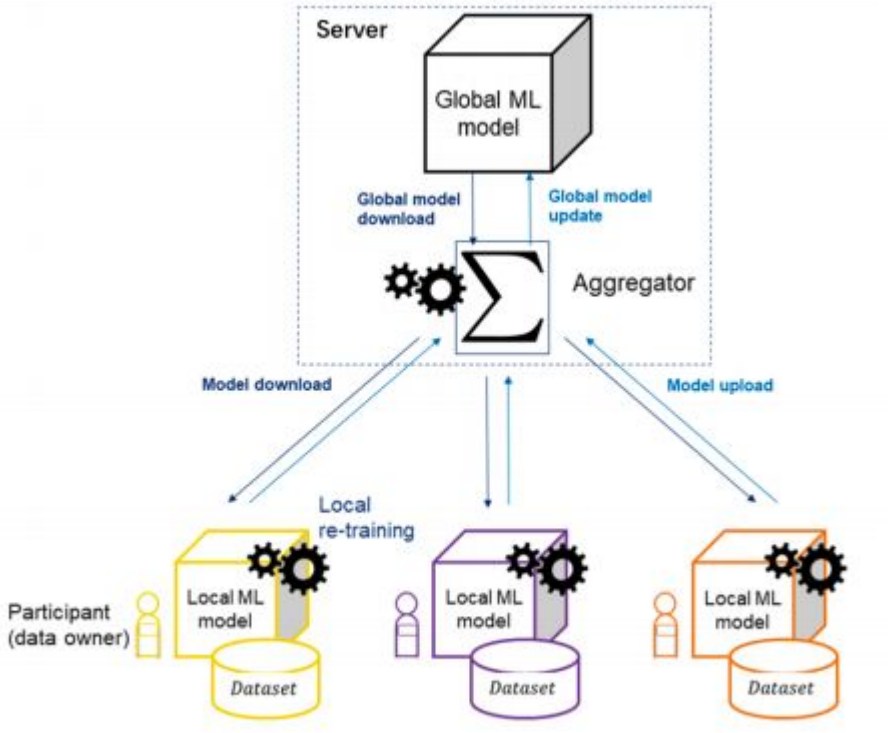


Figure 4: Federated Learning Architecture (client-server FL)

- Data stays with owners who can run the same code
- Aggregator can join models from multiple data owners

Figure from https://aaltodoc.aalto.fi/bitstream/handle/123456789/43561/master_Xia_Yuxi_2020.pdf?sequence=1&isAllowed=y

Other solutions and caveats

- Confidential computing (encrypted containers on secure clusters)
- Blockchain solutions

- Personal comment: keeping data and analysis separated is always the best approach

5. Anonymisation, ethics, and open science

Opening and sharing data must be part of the transparent process of doing research

- **Aalto open science policy** **https://www.aalto.fi/en/open-science-and-research/aalto-university-open-science-and-research-policy**
- **Open science practices** bring **high benefits to researchers** (higher citations, higher impact) and **to the future of science itself** (reproducibility, generalisability, sustainability)
- **“As open as possible, as closed as necessary”**

Can pseudo-anonymised data be opened/shared?

- **Sharing with EU partner institutions can be done**

  It is good to contact the lawyers because you might have joint controllership and you need a data processing agreement
- **Sharing with non-EU partner institutions** requires a legal agreement between parties
- **Data opening:** process is not fully defined yet.
  a. It is not recommended to open personal research data, we can make data available on request.
  b. We need to ask for permissions for "secondary use of data"
  c. Share data under a "Data Use Agreement". See for example: https://data.donders.ru.nl/doc/dua/?0
  d. There are still ethical implications (data leaks, data abuse)

Can anonymised data be opened/shared?

- **Anonymised data are not personal data** so they do not fall under the restriction we have seen before
- However **always consider ethical implications**
  a. What we promise to be anonymous today, might not be anonymous anymore in the future
  b. Although it can be impossible to re-identify an individual, these are still data given by persons who might not be approving re-usage of data for purposes they did not agree with.
  c. See section "2.6 Why ethics is an important issue in Anonymisation" here: http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

Open questions and cases from the participants

Questions from the participants

-

More questions? Contact us: researchdata@aalto.fi

THE END

Thank you!