

RDM


RESEARCH DATA
MANAGEMENT

▶▶ **UHASSELT**

**Anonymisation
Pseudonymisation**

**Access control
Encryption**

SECURE YOUR DATA!

 CC BY-NC-SA RDM stewards, Hasselt University



DATA Security

Index

PART 1 **Hanne Vlietinck**

9.30-10.30

- Basic principles
- Usability versus protecting
- Techniques

10.30-10.45 BREAK

PART 2 **Afshin Amighi**

10.45-11.45

- Fundamental concepts

11.45-12.00 BREAK

12.00-12.45

- Automated solutions

12.45-13.15

- Own research



Part1 and 2 applied exercises



UHASSELT

KNOWLEDGE IN ACTION



DATA Security

Introduction

Meet the RDM team

Afshin Amighi

Margriet Miedema

Hanne Vlietinck

Profession and background



DATA Security

Personal – Sensitive data

Personal data

= Data relating to **identified/identifiable** natural person.

Sensitive data

= Personal data that are **particularly sensitive by their nature**

- *political opinions*
- *religion*
- *philosophical beliefs*
- *racial/ethnic data*
- *trade union membership*
- *sexual orientation*
- *sexual behaviour*
- ...
- *health data*
- *genetic data*
- *biometric data*
(*fingerprint, iris scan, etc.*)





DATA Security

Personal – Sensitive data



National
Coordination Point
Research Data
Management

[DOI 10.5281/ZENODO.3584842](https://doi.org/10.5281/ZENODO.3584842)



UHASSELT

KNOWLEDGE IN ACTION



DATA Security

In numbers data breach

On the (un)informative principle

- 1 in the world: $\approx -\log(1/7,394,000,000) \approx 34$ bits
- 1 in Europe: $\approx -\log(1/742,452,170) \approx 30$ bits
- 1 in NL: $\approx -\log(1/16,802,463) \approx 24$ bits
- 1 in Rotterdam: $\approx -\log(1/625.472) \approx 19$ bits
- 1 in HR: $-\log(1/34.408) \approx 15$ bits
- 1 in this class: $-\log(1/33) \approx 5$ bits
- 1 in the front row: $-\log(1/5) \approx 2$ bits
- 1 is Mr. X: $-\log(1/1) \approx 0$ bits





DATA Security

Data publishing

Why is data publishing important?

- The more you're **cited** it makes you a **more known** among society and researchers
- **Don't invent the wheel again!**
Research and analysis shouldn't be done over and over again. Save time, money! ... for yourself and the community
- Everybody can help each other
Sharing is caring



DATA Security

Data publishing

Why is securing your data necessary?

- Important because of **protecting sensitive data**
BUT:
 - **As open as possible as closed as needed**
 - Under the **correct license**
- **Reasonable reasons no to:**
 - The de-identification gives to much data loss
 - IP & Further research



DATA Security

Steps to secure data

Access control
*Authorization,
Authentication*



De-identification



UHASSELT

KNOWLEDGE IN ACTION



DATA Security

Steps to secure data

De-identification

Minimisation

"only min. personal data necessary"

Storage limitation

"personal data no longer need should be erased"

RISK-Analysis APP*

Techniques

Sort of identifiers

"Define direct, strong indirect, indirect identifiers"

* APP=Anonymisation Pseudonymisation plan



DATA Security

Steps to secure data

Kind of identifiers



Data can be identified by:

1. Direct identifier
2. Strong Indirect identifier
3. Indirect identifier



DATA Security

De-identification

DE-IDENTIFICATION

" Removes association identifying information"



Pseudonymisation

Anonymisation

*Direct identifiers
separately*

Reversible

*Direct identifiers
Deleted*

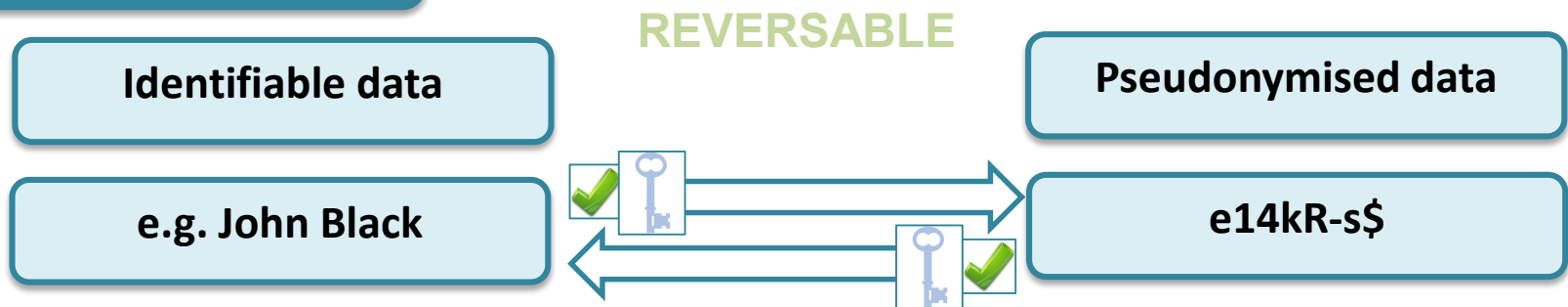
Irreversible



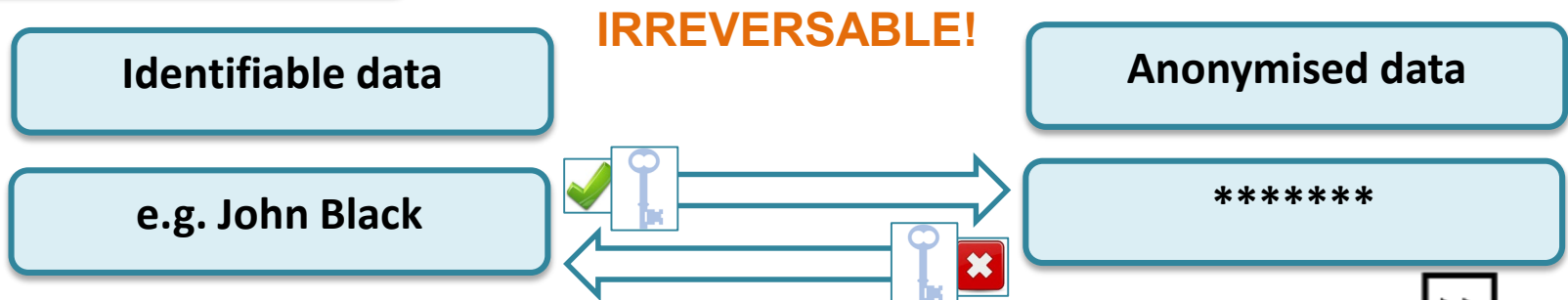
DATA Security

De-identification

Pseudonymisation



Anonymisation





DATA Security

De-identification

PERSONAL DATA		ANONYMOUS DATA
<i>Fully identifiable data</i>	<i>Pseudonymous data</i>	
EXAMPLE		
Patient number 90210	Study subject 47110009	Country Netherlands
City Leeuwarden	Region Friesland	Age 51-60
Date of birth 27-4-1967	Year of birth 1967	Income 5.000 – 15.000
Income 7.861	Income 7.500-10.000	Job Legal
Job Judge	Job Legal	Car sportwagen
Car DeLorean	Car DeLorean	
License Plate SN-09-HN		

**WANT TO KNOW MORE?
TURN CARD OVER!**



National
Coordination Point
Research Data
Management

[DOI:10.5281/zenodo.3584842](https://doi.org/10.5281/zenodo.3584842)



UHASSELT

KNOWLEDGE IN ACTION



DATA Security

De-identification

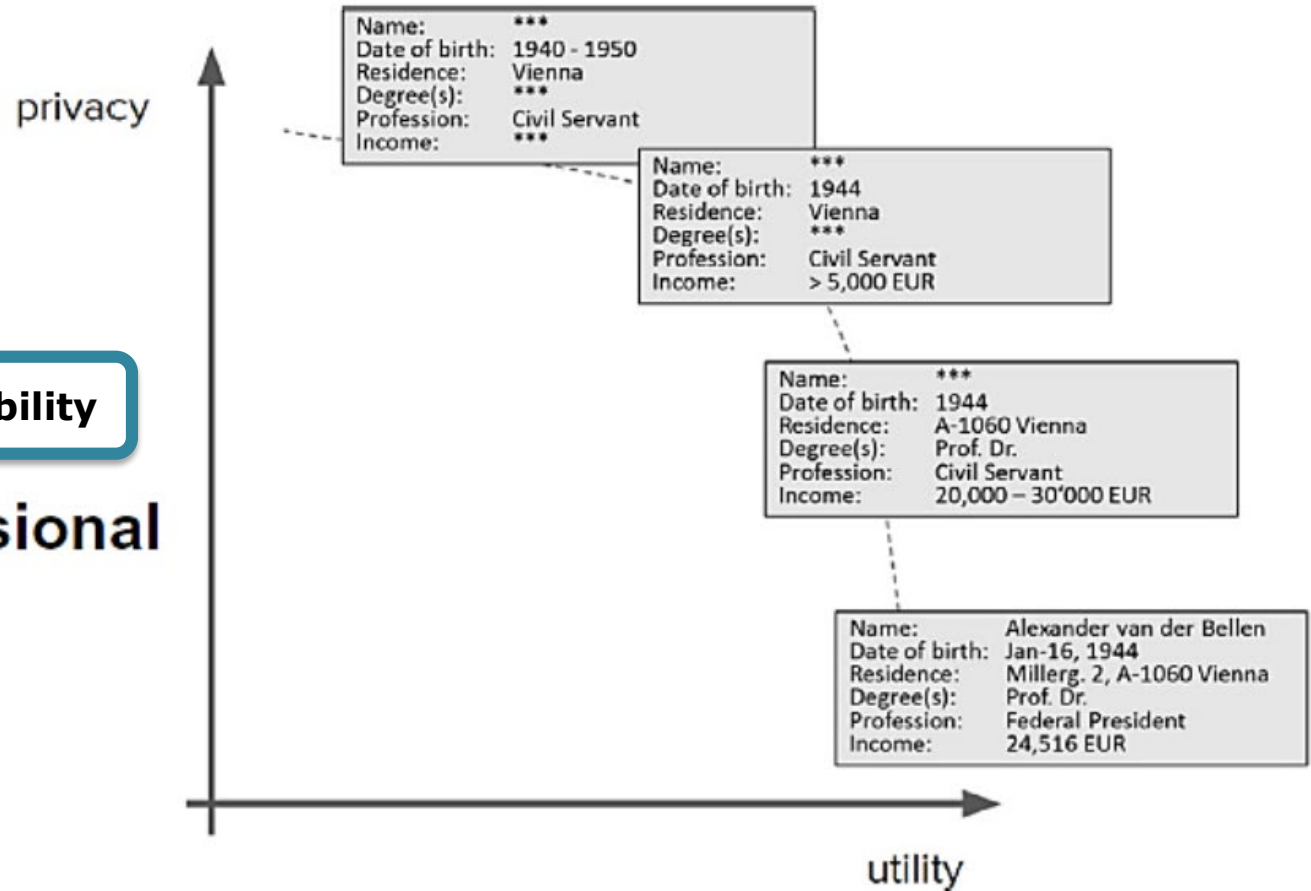


*Find a balance
between
protection and usability*



DATA Security

De-identification



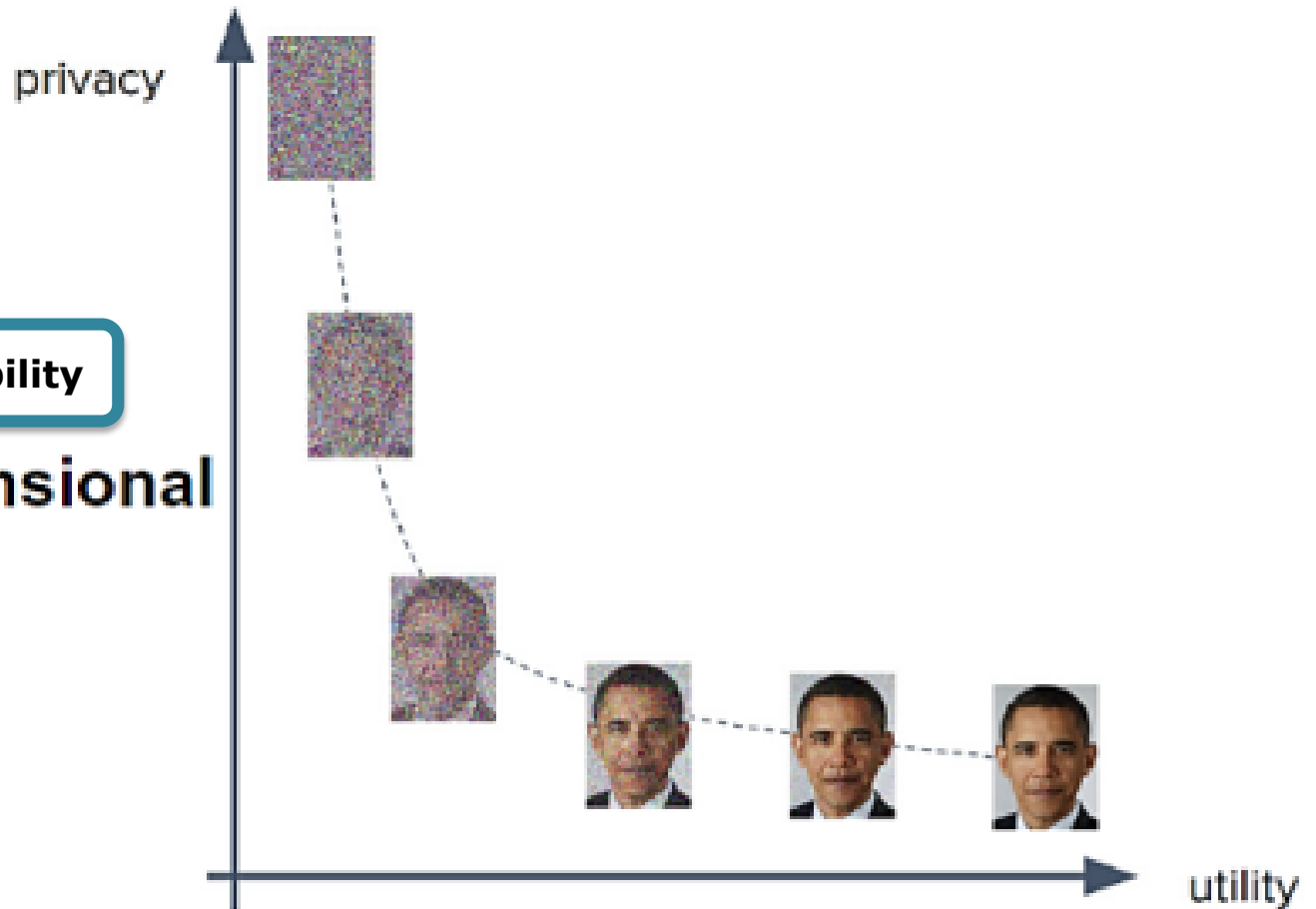
Degrees of identifiability

(a) Low-dimensional case



DATA Security

De-identification



Degrees of identifiability

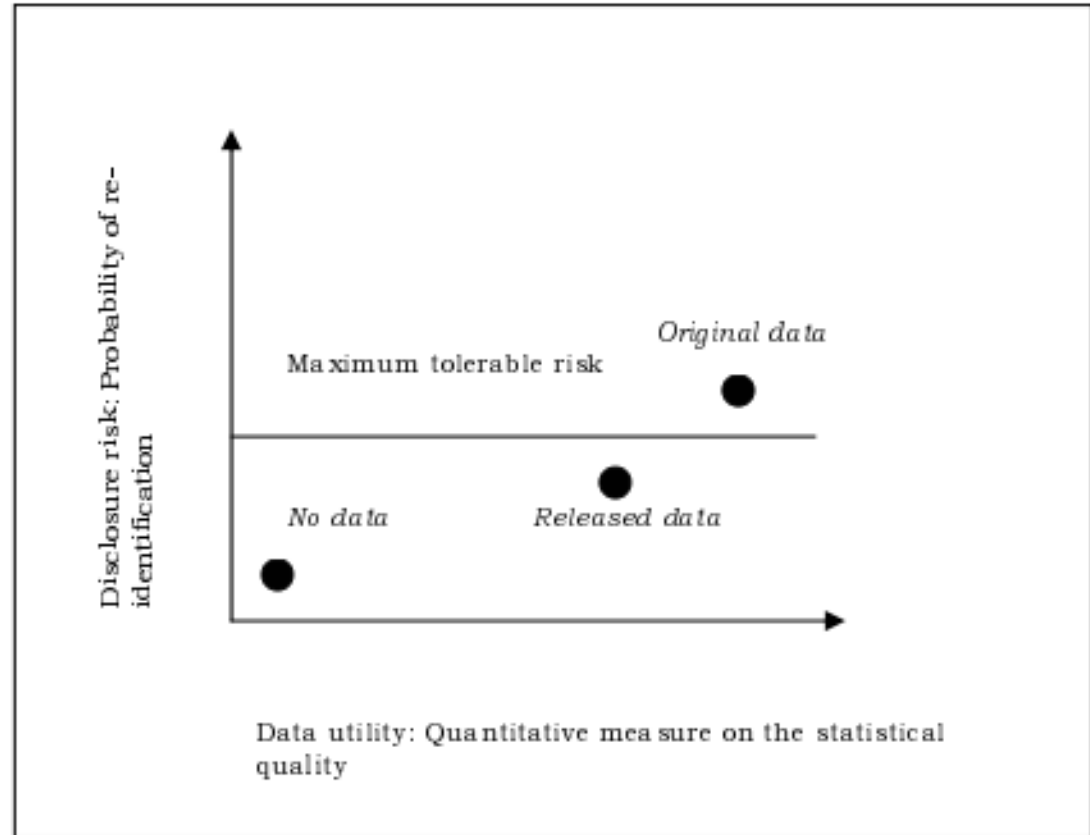
(b) High-dimensional case



DATA Security

De-identification

Degrees of identifiability





DATA Security

De-identification

Methodology

Anonymisation

Pseudonymisation

Perturbative

Non-perturbative

Protection model

K-anonymity, L-Diversity, T-Closeness

Techniques

Quantitative

Qualitative



UHASSELT

KNOWLEDGE IN ACTION



DATA Security

Steps to secure data

Anonymisation-Pseudonymisation plan

Make a anonymisation-pseudonymisation plan.

This should include the following information:

- ✓ creator(s) of the plan
- ✓ person(s) carrying out the anonymisation
- ✓ features in the data that have an impact on anonymisation
- ✓ assessment of the disclosure risk of respondents' personal data
- ✓ anonymisation techniques used along with the rationale for using them.

[Anonymisation plan template \(PDF\)](#)





DATA Security

De-identification

⌚ 15 m. 👤 6 p.

BREAK-OUT
RISK-analyse



Discuss your scope of sensitive data,
(future) actions taken

- *First give extra info about you and your data.*
 - *Function*
 - *Dataset*
 - *Which techniques manually/ automatically (software) already used?*
- *Create a Anonymisation-Pseudonymisation plan*
[Link to template](#)

Only share what you may share



DATA Security

De-identification

Quantitative data

*By Encryption**

1. Coding by lookuptable
2. Hashing and Salted hash

Non-perturbative

1. Deletion of variables
2. Sampling
3. Generalisation
4. Top/Bottom coding
5. Local suppression

Perturbative

1. Noise addition
2. Post-randomisation
3. Overimputation
4. Microaggregation
5. Data swapping

* Conversion to code or symbols contents not understandable if intercepted



DATA Security

De-identification

Quantitative data

By Encryption

Coding by lookup table

Pseudonymization Data

Name	Age	Disabilities
18w8fy1uitxg	42	Vision impairment
sjjinsx53ccm	21	None
ta6n4md6cosk	74	deaf or hard of hearing
dhkg1ufzkkp6	44	None
xo2f42372wfc	32	Mental health

Personal Data

Data Subject Name (personal data identifier)
Alice
Bob
Dave
Eve
Grace



Pseudonym (Token)	Data Subject Name (personal data identifier)
18w8fy1uitxg	Alice
sjjinsx53ccm	Bob
ta6n4md6cosk	Dave
dhkg1ufzkkp6	Eve
xo2f42372wfc	Grace



Pseudo lookup table



UHASSELT

KNOWLEDGE IN ACTION



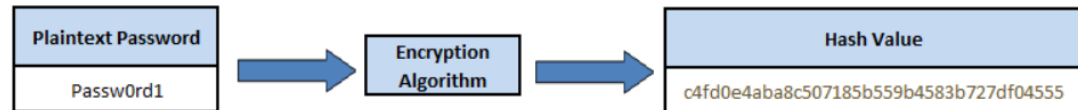
DATA Security

De-identification

Quantitative data

By Encryption

Hashing



Username	Password (hashed)
Alice	c4fd0e4aba8c507185b559b4583b727df04555
Bob	1809c34c89e13c9f056d461f5e252d1d24f188eb
Dave	317db6ad17e98c88c064debd5ae8d274b6b2433
Eve	c051dfe501477eadeae8d14340123ac9f25b9b6a
Grace	41882e3b02d38b11bf6fc739b2a07bf321c7



DATA Security

De-identification

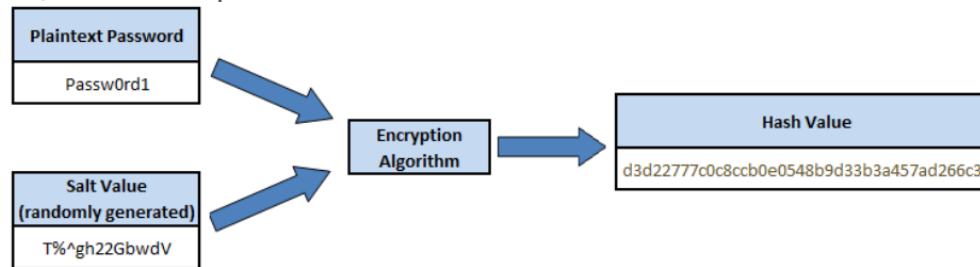
Quantitative data

Salted hashing

By Encryption

Why do we call it salted hashing?

The use of the word "salt" is probably a reference to warfare in ancient times, when people would **salt the wells or farmland to make it less hospitable**. The **Romans** are sometimes supposed to have done this to Carthage in 146 BC. In the context of passwords, a "salted" password is harder to crack.



Username	Password (hashed)	Unique Salt
Alice	d3d22777c0c8ccb0e0548b9d33b3a457ad266c39	T%^gh22GbwdV
Bob	5ef000cb4f7f7d62dad6924e9a55012f9fa578	caca71a46aeb
Dave	91d12a6eca1784fef394ad7f265bd122f8e07daa	e31f31d9761f
Eve	017f545cfcba28d9121b18d1656fd41f30190ac8	7524acf2458c
Grace	19b5e89fa2e871fa6f0e3ac194bd4687471r56hf6	8r540hd1100R

Once you **add salt to food the real taste is no longer visible**.

So basically this is a figurative saying; **add** a little salt and it **changes the original dish**



DATA Security

De-identification

Quantitative data

Non-perturbative – Does not distort the data

1. Deletion of variables
- 2. Sampling**
Taking less data than analysed. All the people of Utrecht. Only 80% percentage of Utrecht
3. Generalisation
Generalise to less specific categories. E.g. Delorean → Sportcar
4. Top/Bottom coding
Interval upper and lower limit. E.g. Salary 3000-5000
5. Local suppression
Suppression so a quasi-identifier can't be identified. E.g. code 3600 => 36**



DATA Security

De-identification

Quantitative data

Perturbative – Distorts data

1. Noise addition

Only for numerical data. Ex.: Adding noise to data e.g. birthday + 3 months

2. Post-randomisation - –Synthesing data

Replacing data by simulated data

3. Microaggregation

Ordering in groups as homogeneous as possible where only the average is published. Ex.: average of age 46 (the real age can be 45/46/47)

4. Data swapping

2 respondents form a "swapping pair" by same age

Ex.: Income categories highly identifiable swapped. Reduce chance of data disclosure





DATA Security

De-identification

Anonymisation tools

- ✓ Arx
- ✓ Amnesia
- ✓ sdcMicro
- ✓ Scrubber





DATA Security

De-identification

Qualitative data

Transcriptions
audio- or video

1. Search replace function
2. Distorting
3. Blurring

Neuroimaging
MRI data

1. Scrubbing
2. Defacing



UHASSELT

KNOWLEDGE IN ACTION



DATA Security

De-identification

Qualitative data

Transcriptions
audio- or video

Search replace function

Transcriptions

Annotation replacement between [bracket] or <angle brackets>

- “Advanced Find > Wildcards > [A-Z] > Reading Highlights”
- “Advanced Find > Use Wildcards > [0-9] > Reading Highlights”

Distorting sounds

Audio

Beeping, distortion voices

Tools: Opensource programme :Audacity

Blurring

Video

Pixel enlargement image, blurring



DATA Security

De-identification

Qualitative data

Neuroimaging
MRI data

Defacing

Facial features

Algorithms to disguise facial features

Tools: pydeface, mri_deface

Scrubbing

Identifiable info image files

To delete or generalize identifiable info in image files, DICOM-files or filepath names. E.g. Age instead of birth date

Tools: DeID, MITRE Identification Scrubber Toolkit



UHASSELT

KNOWLEDGE IN ACTION



DATA Security

De-identification

Depseudonimyzje



Description of
Techniques separate place



Key in different place



4-eyes principle



DATA Security

De-identification

Best practices

- ✓ Audit by external company
- ✓ Castor EDC
- ✓ Find an expert in using tools/ coding techniques
- ✓ Share experiences → Workgroup (Start May)
[Register](#)



DATA Security

De-identification

⌚ 7 m. 👤 6 p.

BREAK-OUT
Future actions



Discuss future actions

- *Which of these techniques would you apply in the future?*

Only share what you may share



DATA Security

De-identification

Protection model

K-Anonymity L-Diversity, T-Closeness

K-Anonymity

X individual matches

L-Diversity

Every generalized block at least
l different sensitive values

T-Closeness

Reducing the granularity

A good technique is to go for a high K-anonymity, L-Diversity and T-Closeness



UHASSELT

KNOWLEDGE IN ACTION



DATA Security

De-identification

Protection model

K-anonymity L-Diversity, T-Closeness

Concepts

Key Attribute

Quasi-identifier

Sensitive attribute

Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

Key Attribute: Direct identifiers: name, address, phone number removed before releasing the data

Quasi-identifier: Not unique identifiers but combined with other quasi-identifiers creates a unique identifier

Sensitive attribute: Medical records, salaries, attributes that researchers need, so always released directly

✓ 'k-Anonymity and cluster based methods for privacy | ~elf11.github.io'. <https://elf11.github.io/2017/04/22/kanonymity.html> (accessed Mar. 14, 2022) available under a [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/)



DATA Security

De-identification

Protection model

K-anonymity

X individual matches

K-ANONYMITY

(lowest) Individual matches of quasi-identifiers=2

→ **2-anonymity**

TECHNIQUES

✓ **Generalisation**

(Birthyear instead of exact data)

✓ **Suppression**

Generalisation causes too much information loss
quasi-identifier not released at all.

Common for outliers

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

✓ 'k-Anonymity and cluster based methods for privacy | ~elf11.github.io'. <https://elf11.github.io/2017/04/22/kanonymity.html> (accessed Mar. 14, 2022)
available under a [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/)



DATA Security

De-identification

Protection model

K-anonymity

Sort of attacks

- 1. Unsorted matching attack :**
Same order records released table, original table.
→ Randomize order before releasing
- 2. Homogeneity Attack:** the background knowledge attack



DATA Security

De-identification

⌚ 10 m. 👤 3 p.

BREAK-OUT
Protection
model



Apply k-anonymity

- *Which techniques did you apply?*
- *How much is the k-anonymity?*



DATA Security

De-identification

BREAK-OUT
*Protection
model*



Apply k-anonymity Solution

- *Which techniques did you apply?
Suppression, Top Bottom coding*
- *How much is the k-anonymity? **4***



UHASSELT

KNOWLEDGE IN ACTION



DATA Security Techniques

15 m.



UHASSELT

KNOWLEDGE IN ACTION

RDM
RESEARCH DATA
MANAGEMENT

▶▶ **UHASSELT**



Authentication

Authorization

**Security
Background info**



Access control

ACCESS CONTROL

Dictates who's allowed to access

AUTHENTICATION

AUTHORISATION

- ✓ WHO gets access
- ✓ WHAT KIND of access

RECHECK! ...





Access control

AUTHENTICATION

Ways to authenticate

✓ Is it you??

- ✓ Passwords
- ✓ IP-access
- ✓ VPN
- ✓ SSO (SAML) or Google sign on

AUTHORISATION

Your rights

✓ What are you allowed to do?

- ✓ Files, Folders, Drives
- ✓ Platforms, Software



Access control

⌚ 10 m. 👤 3 p.

BREAK-OUT RISK-analyse



Discuss your levels of security

- *What are the general risks breaches digitally, physically in your project?*
- *What actions are taken based on access control?*
- *Which storage locations are you using?*

Only share what you may share

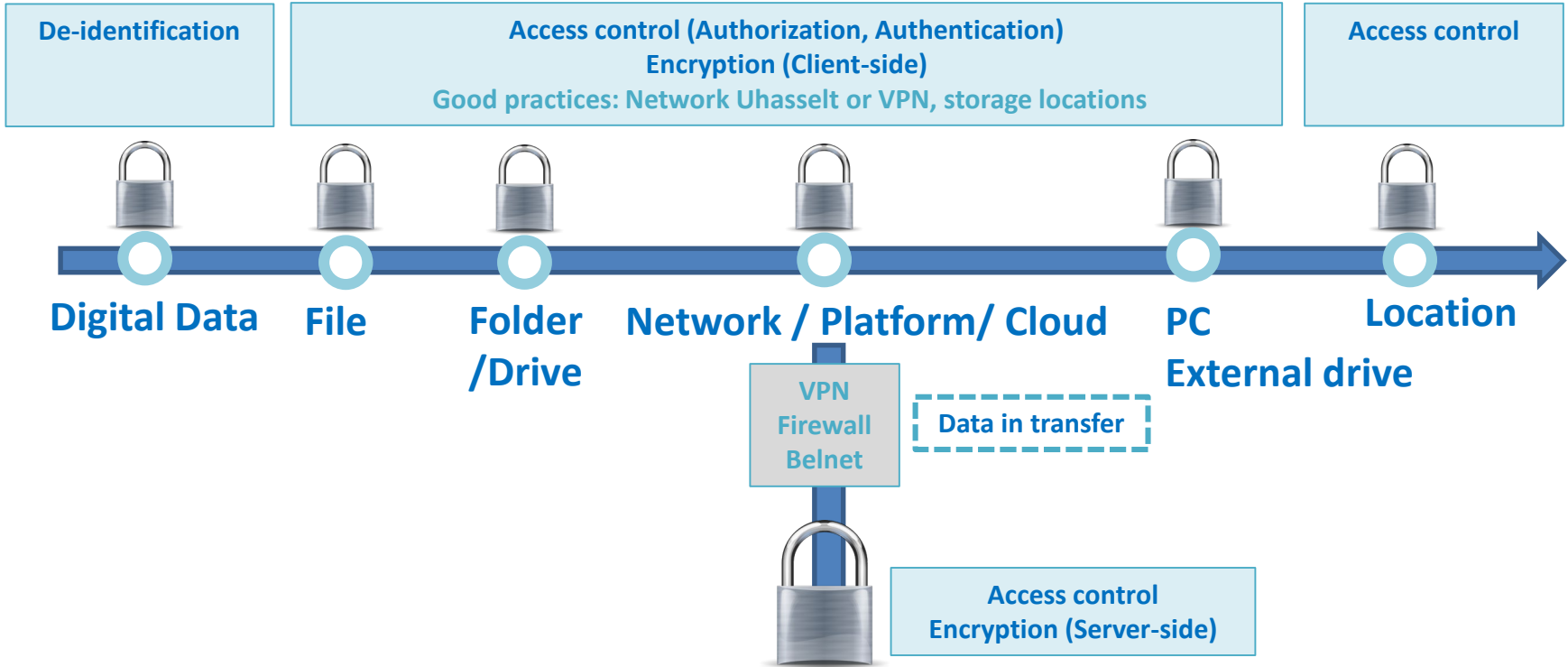


DATA Security

Access control

Several layers of security

DIGITAL



Externally



DATA Security

Access control

Several layers of security

PHYSICAL

Access control (Authorization, Authentication)



Physical Data
(Samples, ...)

Container

Lab

Location

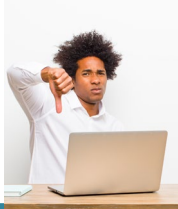




DATA Security

Access control

Password use



BAD PRACTICES

Don't use

- ✓ Tricky, common character substitutions
- ✓ Apparent Sequences
- ✓ Neighbouring keystrokes
- ✓ Repeated characters
- ✓ Complete dates



GOOD PRACTICES

Make it

- ✓ Long
 - ✓ Diverse
 - ✓ (Semi-)impersonal
 - ✓ Different
 - ✓ Random set of words
- OR use a password manager



Access control

Password managers



KeePass

LastPass...
by LogMeIn

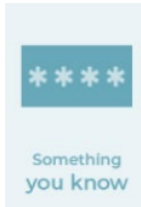




DATA Security

Access control

Combination of authentication



SFA

Single factor authentication



2FA

2 Factor-authentication



MFA

Multi-factor authentication





Access control

ENCRYPTION

Tools:

- ✓ Bitlocker
- ✓ Veracrypt
- ✓ Boxcrypter

Be carefull! Losing the key means losing the data



DATA Security

Secure your sensitive data

⌚ 5m. 👤 3 p.

BREAK-OUT *Improvements*



Discuss improvements

- *What improvements can you take based on access control?*

Only share what you may share



Secure your sensitive data

References

- ✓ N. Vollmer, 'Article 4 EU General Data Protection Regulation (EU-GDPR)', Jul. 02, 2021. <https://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm> (accessed Mar. 14, 2022)
- ✓ E. J. Hrudey et al., 'Anonymization - reference card for researchers', Dec. 2019, doi: 10.5281/zenodo.3584842
- ✓ F. S. S. D. Archive (FSD), 'Data Management Guidelines', Finnish Social Science Data Archive (FSD). <https://www.fsd.tuni.fi/en/services/data-management-guidelines/> (accessed Mar. 14, 2022),
- ✓ ANDS, 'De-identifying your data', ANDS. <https://www.ands.org.au/working-with-data/sensitive-data/de-identifying-data> (accessed Mar. 11, 2022)
- ✓ A. Hundepool et al., 'Statistical Disclosure Control', 2012.
- ✓ 'Minimizing application privacy risk', IBM Developer, May 25, 2018. <https://developer.ibm.com/articles/s-gdpr3/> (accessed Mar. 14, 2022)
- ✓ N. Van Wettere, 'Anonymization tools and techniques', May 25, 2020. doi: 10.5281/zenodo.3843319
- ✓ 'k-Anonymity and cluster based methods for privacy | ~elf11.github.io'. <https://elf11.github.io/2017/04/22/kanonymity.html> (accessed Mar. 14, 2022)
- ✓ J. Wieringa, P. K. Kannan, X. Ma, T. Reutterer, H. Risselada, and B. Skiera, 'Data analytics in a privacy-concerned world', *J. Bus. Res.*, vol. 122, May 2019, doi: 10.1016/j.jbusres.2019.05.005.
- ✓ A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramaniam, 'L-diversity: Privacy beyond k-anonymity', *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 3-es, maart 2007, doi: 10.1145/1217299.1217302.



Secure your sensitive data

Meet the RDM team

Swing by@Data Stewards

Tuesdays (9.00-16.00)

Campus Diepenbeek

Location see contact page



Info & Contact

rdm@uhasselt.be

bibliotheek.uhasselt.be/rdm



DATA Security

Secure your sensitive data

Any Questions?



[Improvement form](#)

