byDesign

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

DPbD – Purpose limitation, Data minimization & Accuracy

Facilitating GDPR compliance for SMEs and promoting Data Protection by Design in ICT products and services *(www.bydesign-project.eu)*

ΑΡΧΗ ΠΡΟΣΤΑΣΙΑΣ
ΔΕΔΟΜΕΝΩΝ

ΚΕΝΤΡΟ ΕΡΕΥΝΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΕΙΡΑΙΩΣ

ICT abovo
Information & Communication Technologies

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

DPbD – Purpose limitation

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Key design elements for purpose limitation

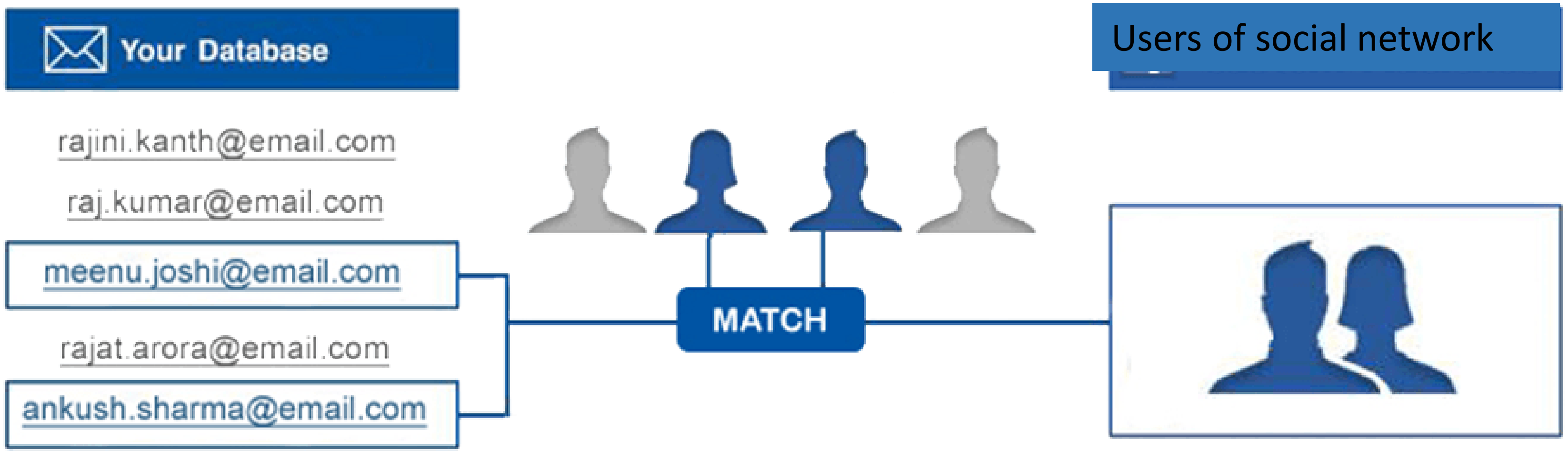
- Predetermination – The legitimate purposes shall be determined before the design of the processing.
- Specificity – The purposes shall be specified and explicit.
- Purpose orientation – The purpose of processing should guide the design of the processing and set processing boundaries.
- Necessity – The purpose determines what personal data is necessary for the processing.
- Compatibility – Any new purpose must be compatible with the original purpose for which the data was collected and guide relevant changes in design.
- Limit further processing – No connecting datasets or perform any further processing for new incompatible purposes.
- Limitations of reuse – Using technical measures, including hashing and encryption, as well as organisational measure, such as polices, to limit the possibility of repurposing personal data.
- Review – Regularly test the design against purpose limitation.

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

DPbD - Bad practices in purpose limitation

- Common bad practices:
  - The process does not “stick” to what it has been said in the privacy notice
  - Consent is being “translated” as permission to do anything with the data collected
  - Personal data are being transmitted to other parties – e.g. for behavioral advertising, or, more generally, for creating profiling
  - Personal data are being connecting with other available datasets, where some users are common

DPbD – A bad example on purpose limitation



- Sending users e-mail addresses to the social network, for advertisments through the network, is a different purpose
- Matching of common addresses (for the same purpose as above) is also a different purpose
    - Even if a proper legal basis exists for this new process (e.g. consent), emphasis should be put on data minimization (discussed further subsequently)

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

DPbD – Data minimization

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Key design elements for data minimization

- Data avoidance – Avoid processing personal data altogether when this is possible for the relevant purpose.
- Limitation – Limit the amount of personal data collected to what is necessary for the purpose
- Access limitation – Shape the data processing in a way that a minimal number of people need access to personal data to perform their duties, and limit access accordingly.
- Relevance – Personal data should be relevant to the processing in question
- Necessity – Each personal data category shall be absolutely necessary for the specified purposes
- Aggregation – Use aggregated data when possible.
- Pseudonymization – Pseudonymize personal data as soon as it is no longer necessary to have directly identifiable personal data
- Anonymization and deletion – Where personal data is not, or no longer necessary for the purpose, personal data shall be anonymized or deleted.
- Data flow – The data flow should be made efficient enough to not create more copies than necessary.
- “State of the art” – Application of up to date appropriate technologies

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

DPbD - Bad practices in data minimisation

- Common bad practices:
  - Process of more personal data that are needed – e.g. collecting data that are not actually needed
    - E.g. Applications request high-precision location when they only need to know the city or the country.
  - Collecting data for several different purposes which, if combined, allow processing for a new purpose (not transparent to the users)
  - During the data flow, extensive personal data may be available to people/systems that should not have access to such data (violation of the “need-to-know” principle)
  - Bad pseudonymisation – it is believed that identities are protected, but this is not the case
    - I.e. bad pseudonymisation design/technique, insufficient protection of data allowing pseudonymisation reversal etc.
  - Data are being considered as anonymous but this is not the case
    - Anonymisation is a non-trivial task that needs much attention…

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

An example

- Web form of a bookstore:
  - Asking information including the customer’s date of birth, phone number and home address
  - Are all these necessary?
    - If the user pays for the product up front, the user’s date of birth and phone number are not necessary for the purchase of the product
    - The home address may be also unnecessary – e.g. in case of an e-book or in cases that the user chooses to go to the physical store
- **Possible best practice**:
  - Two web forms: one for ordering books, with a field for the customer’s address and one web form for ordering eBooks without a field for the customer’s address.
  - In any case, the mandatory information is explicitly described and justified

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

An example

- A public transportation company wishes to gather statistical information based on travellers' routes
- The passengers have to pass their ticket through a reader every time the enter a transport (e-ticket system)
- The information collected by company may allow identification of the passengers in some circumstances, based on single route identification thanks to the ticket identifier.
  - E.g. if they live or work in scarcely populated areas
- **Solution**: the ticket identifier is not stored
- Challenge: What if the ticket identifier is needed for another legal purpose?
- => the overall process should be carefully designed (e.g. proper pseudonymisation, logical and physical separation of databases for different purpose so as to be unlinkable etc.)

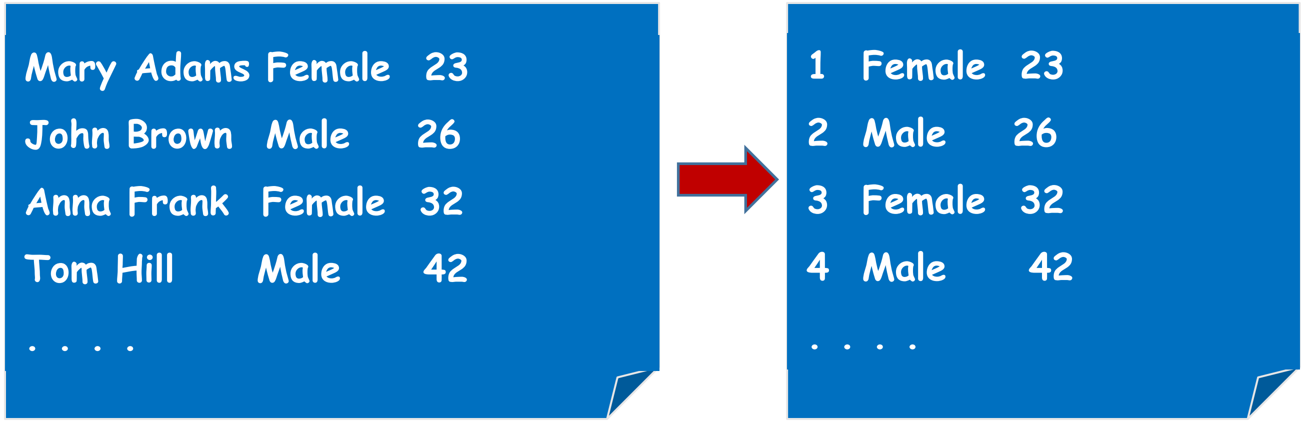
An example

- A hospital utilizes an electronic health record
- By default, access is granted to only those members of the medical staff who are assigned to the treatment of the respective patient in the specific department she/he is assigned to
- The group of people with access to a patient’s file is enlarged if other departments or diagnostic units are involved in the treatment.
- After the patient is discharged, and billing is completed, access is reduced to a small group of employees per speciality department who answer requests for medical information or a consultation made or asked for by other medical service providers upon authorization by the respective patient.
- If access is needed for research purposes, the researcher should not know the exact identity of the patients
    - This may achieved by, e.g, proper pseudonymisation (for example, through a deterministic pseudoynimasation scheme on the Social Security Number)

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Why pseudonymisation?

1. Hiding identities (related to confidentiality)



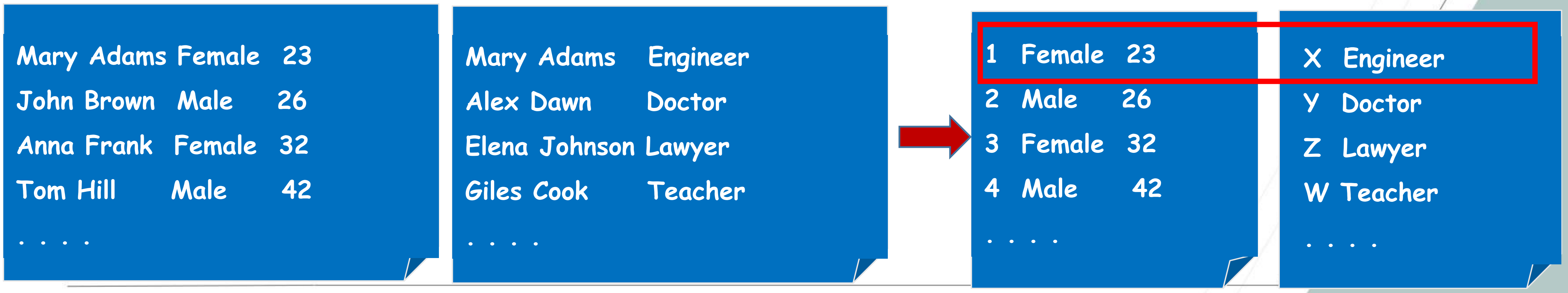
Mary Adams	Female	23
John Brown	Male	26
Anna Frank	Female	32
Tom Hill	Male	42
. . . .		

1	Female	23
2	Male	26
3	Female	32
4	Male	42
. . . .		

- Pseudonymisation is explicitly mentioned in the GDPR as a possible safeguard towards achieving data protection by design (see subsequent seminar)
- Special attention:
    - State-of-the-art
    - Is re-identification possible through the remaining information?

A risk-based approach

2. Unlinkability



Mary Adams	Female	23
John Brown	Male	26
Anna Frank	Female	32
Tom Hill	Male	42
. . . .		

Mary Adams	Engineer
Alex Dawn	Doctor
Elena Johnson	Lawyer
Giles Cook	Teacher
. . . .	

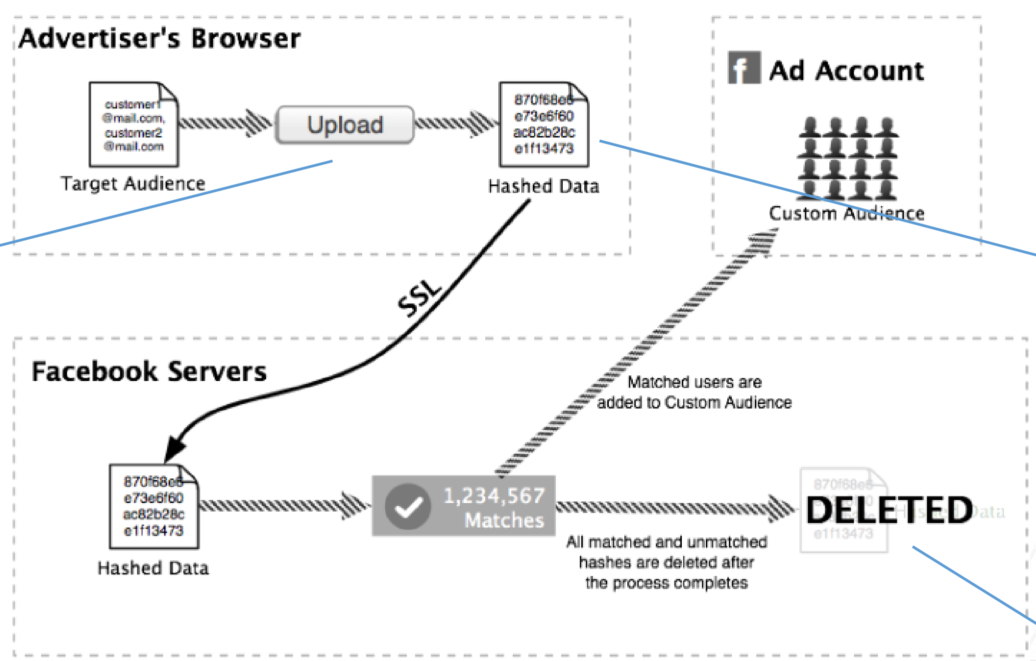
1	Female	23
2	Male	26
3	Female	32
4	Male	42
. . . .		

X	Engineer
Y	Doctor
Z	Lawyer
W	Teacher
. . . .	

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

An example of (bad?) pseudonymisation

- Custom audience match process in Facebook (Under the assumption that valid users consents exist)
- (source: https://3qdigital.com/wp-content/uploads/2016/06/facebook_audiences_data_security_overview.pdf)


Advertiser's Browser
customer1 @mail.com, customer2 @mail.com
Target Audience
Upload
870f68e6 e73e6f60 ac82b28c e1f13473
Hashed Data
Ad Account
Custom Audience
SSL
Facebook Servers
Matched users are added to Custom Audience
870f68e6 e73e6f60 ac82b28c e1f13473
Hashed Data
1,234,567 Matches
DELETED
All matched and unmatched hashes are deleted after the process completes

E.g. email addresses from customers

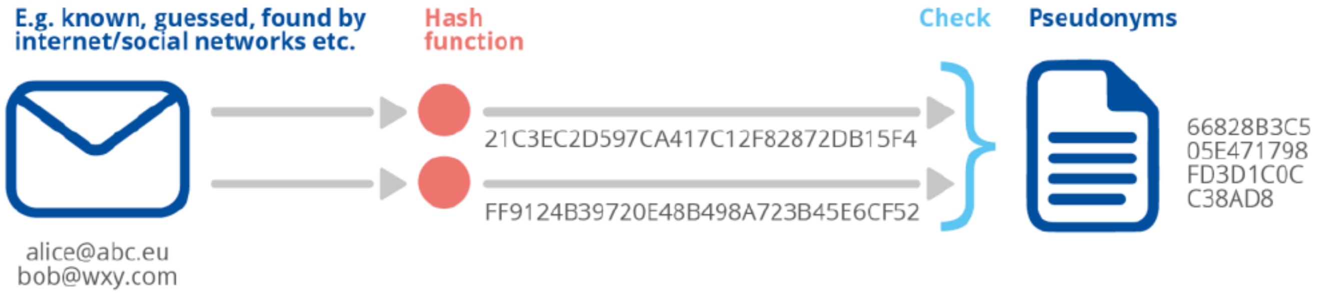
These are actually pseudonyms (hashed email addresses)

Necessary, due to data minimization)

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

An example of (bad?) pseudonymisation (Cont.)

- Hashed data are mathematically irreversible, but..
  - If a controller knows hashed email addresses, recovering some of them is possible…



E.g. known, guessed, found by internet/social networks etc.

Hash function

Check

Pseudonyms

21C3EC2D597CA417C12F82872DB15F4

FF9124B39720E48B498A723B45E6CF52

66828B3C5
05E471798
FD3D1C0C
C38AD8

alice@abc.eu
bob@wxy.com

Source: ENISA Report, Pseudonymisation techniques and best practices, 2019.

- Not a state-of-the-art solution => data minimization is put at risk
- More advanced state-of-the-art techniques should be considered in the design process – e.g. private set intersection techniques (see ENISA Report, Data Pseudonymisation: Advanced Techniques and Use Cases, 2021)

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Example of bad anonymisation – The famous AOL example

August 2006: research.aol.com

AOL is embarking on a new direction for its business making its content and products freely available to all consumers. To support those goals, AOL is also embracing the vision of an open research community. To get started, we invite you to visit us at http://research.aol.com, where you will find:

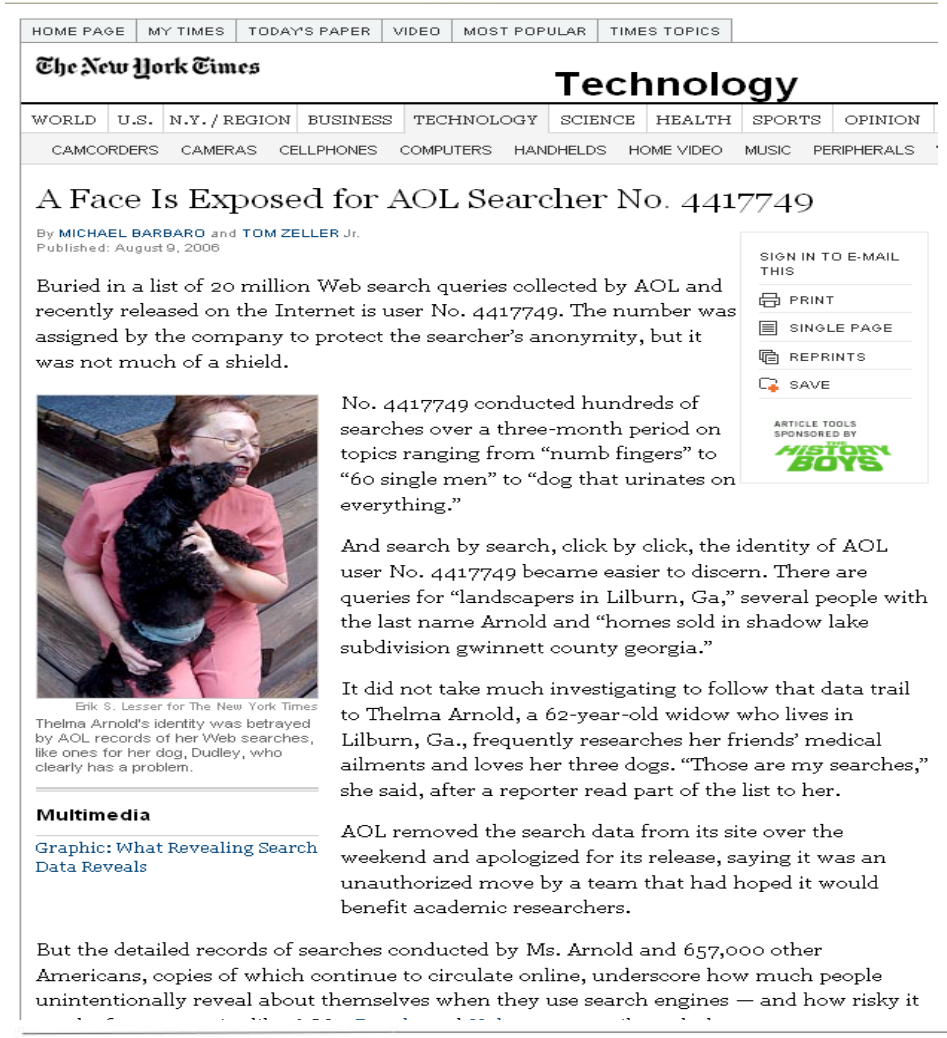
- *…*
- ***Query streams for 500,000 users over 3 months (20 million queries)***
- *….*

- A random ID was associated to each user
  - The same (meaningless) ID, for the same user
  - This is actually a pseudonymisation procedure..

- However, a combination of the published information with other available data could allow identification!

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

Re-identification from “anonymous” data



A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher’s anonymity, but it was not much of a shield.

Erik S. Lesser for The New York Times
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.

Multimedia

Graphic: What Revealing Search Data Reveals

AOL removed the search data from its site over the weekend and apologized for its release, saying it was an unauthorized move by a team that had hoped it would benefit academic researchers.

But the detailed records of searches conducted by Ms. Arnold and 657,000 other Americans, copies of which continue to circulate online, underscore how much people unintentionally reveal about themselves when they use search engines — and how risky it

- The characterization of anonymous data is not an easy task
- Simply removing “obvious identifiers” is not adequate
- In other words, the notions of identifiers or “identifying data” is wide
  - Identifier in which context?
- A proper design on the anonymisation/pseudonymisation procedure is needed to be performed from the beginning

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

DPbD – Accuracy

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

Key design elements for accuracy

- Data source – Sources of personal data should be reliable in terms of data accuracy.
- Degree of accuracy – Each personal data element should be as accurate as necessary for the purposes.
- Measurably accurate - Reduce the number of false positives/negatives, for example biases in automated decisions and artificial intelligence.
- Verification – Verify the correctness of personal data with the data subject before and at different stages of the processing
- Erasure/rectification – Without delay.
- Error propagation avoidance – Mitigating the effect of an accumulated error in the processing chain.
- Access – Users should be allowed having effective access to personal data
- Continued accuracy – Tests of accuracy should be carried out at several stages
- Up to date – Personal data shall be updated if necessary for the purpose.
- Data design - Use of appropriate design features to decrease inaccuracy – e.g. present concise predetermined choices instead of free text fields.

Το έργο χρηματοδοτήθηκε από το Πρόγραμμα Δικαιώματα, Ισότητα και Ιθαγένεια 2014-2020 της Ευρωπαϊκής Ένωσης

DPbD - Bad practices in accuracy

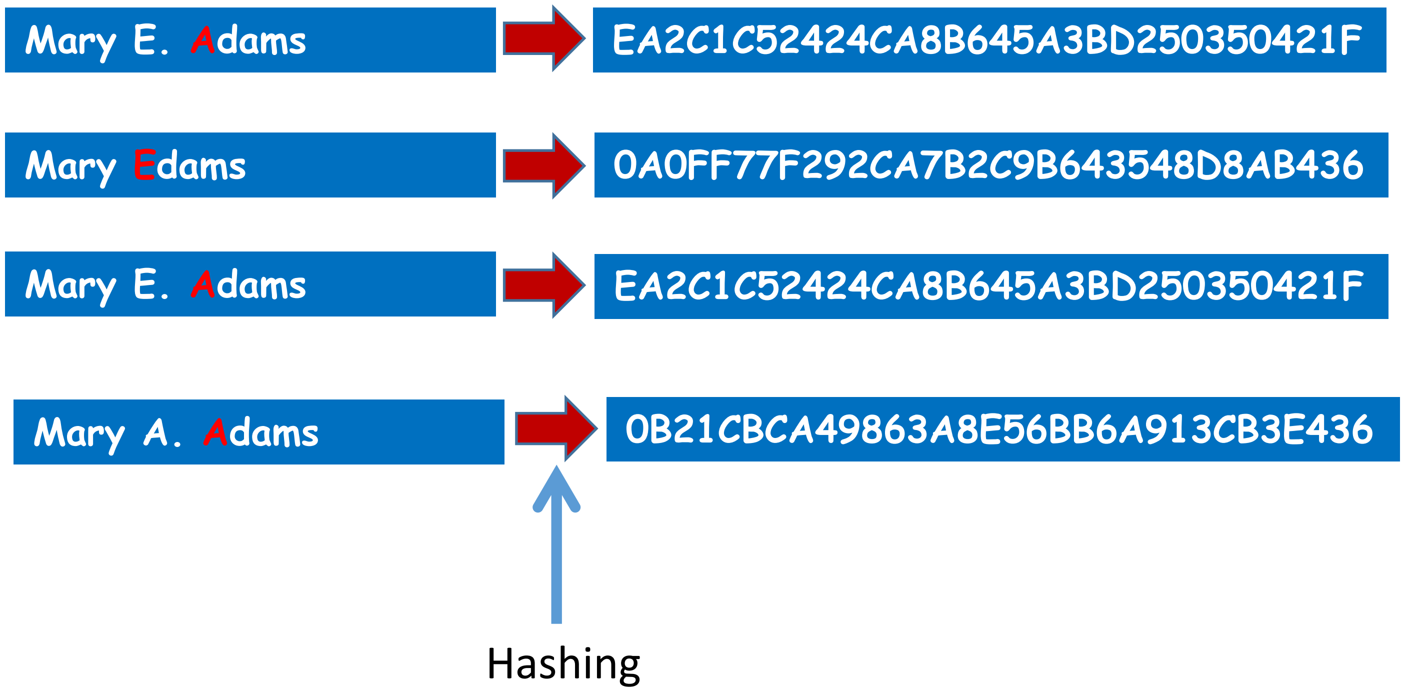
- Common bad practices:
  - Non-verification of accuracy during the collection of data
    - E.g. The validity of the email address is not being checked
  - No appropriate ways to discriminate individuals with “similar” identifiers
    - E.g. Two records for two different individuals with the same name, such as “Mary Adams”
  - Collection of data from untrusted sources
    - E.g. from social networks or public web sites (note that such a collection may not even have a legal basis!)
  - No easy means for the users to update their information that they have provided

An example

- A health institution is looking to find methods to ensure the accuracy of personal data in their client registers.
- Where two persons arrive at the institution at the same time and receive the same treatment, there is a risk of mistaking them if the only parameter to distinguish them is by name.
- Need for a unique indistiguishable identifier for each person.
- Possible solution: Unique identifier per user (pseudonymisation)
  - Via, e.g., cryptographic solutions such as hashing

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

An example (Cont.)



Mary E. Adams → EA2C1C52424CA8B645A3BD250350421F

Mary Edams → 0A0FF77F292CA7B2C9B643548D8AB436

Mary E. Adams → EA2C1C52424CA8B645A3BD250350421F

Mary A. Adams → 0B21CBCA49863A8E56BB6A913CB3E436

Hashing

Το έργο χρηματοδοτήθηκε από το
Πρόγραμμα Δικαιώματα,
Ισότητα και Ιθαγένεια 2014-2020
της Ευρωπαϊκής Ένωσης

An example

- An insurance company wishes to use artificial intelligence (AI) to profile customers buying insurance as a basis for their decision making when calculating the insurance risk
  - Assuming that a valid legal basis is in place
- The model is being trained based on large pool of existing customers
  - Proper pseudonymisation of data to feed the training module
- The accuracy of input data is essential (only trusted sources of data)
- The company should check whether the AI is reliable and provides non-discriminatory results both during its development and finally before the product is released