Aalto University

Privacy and Data Protection

Tuomas Aura
CS-C3130 Information security

Aalto University, autumn 2019

Outline

- Anonymity and pseudonymity
- Anonymization techniques
- Differential privacy
- EU General Data Protection Regulation (GDPR)

Based on a lecture by Andrew Paverd

ANONYMITY AND PSEUDONYMITY

Goal of data anonymization

- Dataset, such as customer database, contains
  - Identifying information of the data subjects
  - Sensitive data about the data subjects

→ Dataset reveals sensitive data about individual subjects

- How can we release the dataset to research, statistical analysis, market research etc. while preserving privacy?

(For simplicity, let’s assume the identifying and sensitive attributes are separate.)

Data anonymization

Customer id	Name	Social security number	Street address	City	ZIP code	Gender	Date of birth	Date registered	Open orders	Purchase history
2	Samu Sirkka	11071998-8976	Jämerän taival 1A	Espoo	02150	M	1998-07-11	2019-09-10	*Beef* *Cigars*	*xxxx*
3	Heli Keiju	104032000A992B	Servin Maijan tie 2B	Espoo	02150	F	2000-03-04	2019-09-10	*Honey* *Sugar* *Cola*	*xxxxx*
…	…	…	…	…	…	…	…	…	…	

Identifying information

Sensitive data

Data anonymization

Open orders	Purchase history
Beef *Cigars*	*xxxx*
Honey *Sugar* *Cola*	*xxxxx*
…	

Identifying information

Sensitive data

If all identifying attributes are removed, the dataset may become worthless

Data anonymization

City	ZIP code	Gender	Date birth	Open orders	Purchase history
Espoo	02150	M	1998-	*Beef* *Cigars*	*xxxx*
Espoo	02150	F	2000-	*Honey* *Sugar* *Cola*	*xxxxx*
…	…	…	…	…	

Identifying information

Sensitive data

Goal of anonymization: remove enough identifying information to prevent association of sensitive data with individual data subjects

Example: de-anonymizing health records

Health records: Massachusetts Group Insurance Commission released “anonymized” health records of state employees

- Removed names, social security numbers, etc.

ZIP code	Gender	Date of birth	Medical record
123456	M	1988.03.17	[hospital visits]
123789	F	1967.07.11	[diagnoses]
123456	F	1984.05.30	[prescriptions]

Example: de-anonymizing health records

Voter records: Purchased legally for $20 (auxiliary information)

ZIP code	Gender	Date of birth	Name
123456	M	1988.03.17	James
123789	F	1967.07.11	Mary
123456	F	1984.05.30	Elisa

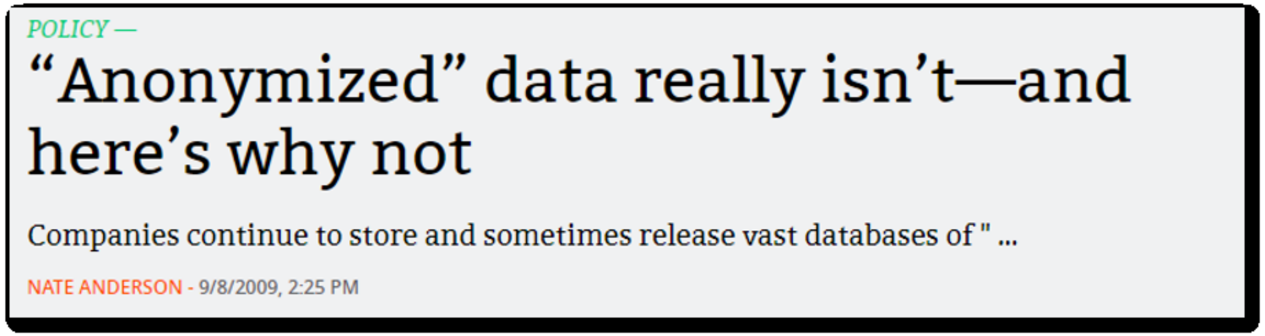
Example: de-anonymizing health records

Combining “anonymized” health records with voter data enabled “de-anonymization” of the Governor of Massachusetts

ZIP code	Gender	Date of birth	Medical record
123456	M	1988.03.17	[hospital visits]
123789	F	1967.07.11	[diagnoses]
123456	F	1984.05.30	[prescriptions]

ZIP code	Gender	Date of birth	Name
123456	M	1988.03.17	James
123789	F	1967.07.11	Mary
123456	F	1984.05.30	Elisa

“Privacy-preserving” data release

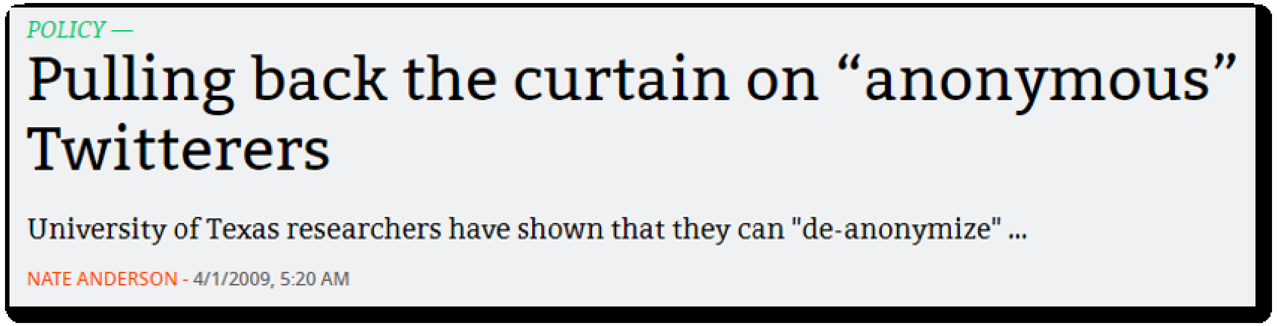


POLICY —

“Anonymized” data really isn’t—and here’s why not

Companies continue to store and sometimes release vast databases of " ...

NATE ANDERSON - 9/8/2009, 2:25 PM

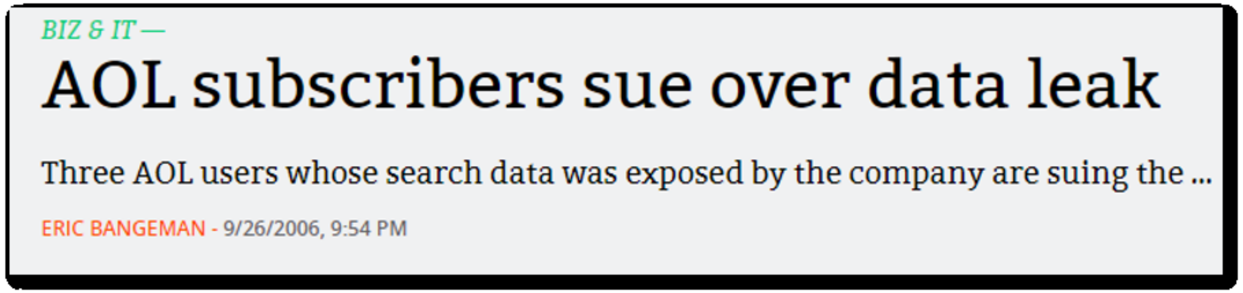


POLICY —

Pulling back the curtain on “anonymous” Twitterers

University of Texas researchers have shown that they can "de-anonymize" ...

NATE ANDERSON - 4/1/2009, 5:20 AM



BIZ & IT —

AOL subscribers sue over data leak

Three AOL users whose search data was exposed by the company are suing the ...

ERIC BANGEMAN - 9/26/2006, 9:54 PM

Pseudonymity

!

- Pseudonym = “a fictitious name”
- Not the real name of a subject, but used wherever the real name would be
- Pseudonym links together different records relating to the same subject
  - E.g. student number

- Pseudonymity is broken if adversary can link pseudonym to real identity (re-identification)

How difficult is re-identification?

Quasi-identifiers

- Data fields that are not identifiers can reveal identify when combined

- ~87% of U.S. population likely to be uniquely identified by: {5-digit ZIP, gender, date of birth}
- ~50% of U.S. population likely to be uniquely identified by only {place, gender, date of birth}
  - where place = city, town, or municipality in which the person resides

L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

Example: Unique in the crowd

Extra material

- Dataset: 15 months of mobility data for 1.5 million users
  - – Time granularity: hourly (~10,000 data points per individual)
  - – Spatial granularity: same as mobile network (e.g. cell tower ID)

Pseudonym	**09:00**	**10:00**	**11:00**	**12:00**	**13:00**	**14:00**	**...**
User 1	13	22	22	22	22	13	...
User 2	44	55	13	13	22	55	...
User 3	27	22	22	22	22	22	...

Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen & Vincent D. Blondel
“Unique in the Crowd: The privacy bounds of human mobility” Scientific Reports, 2013.

Unique in the crowd

Extra material

- Dataset: 15 months of mobility data for 1.5 million users
  - Time granularity: hourly (~10,000 data points per individual)
  - Spatial granularity: same as mobile network (e.g. Cell tower ID)
  - **4 spatio-temporal data points were sufficient to re-identify 95% of users**

Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen & Vincent D. Blondel
"Unique in the Crowd: The privacy bounds of human mobility" Scientific Reports, 2013.

Unique in the shopping mall

Extra material

- Dataset: 3 months of credit card records for 1.1 million individuals

- Uniqueness: How many transactions are needed to re-identify an individual?
  - **4 spatio-temporal points sufficient to re-identify 90% of individuals**

Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, Alex Pentland
"Unique in the shopping mall: On the reidentifiability of credit card metadata" Science, 2015.

ANONYMIZATION TECHNIQUES

k-anonymity

!

A released dataset provides k-anonymity if information about each individual in the dataset is indistinguishable among at least *k* individuals in the dataset, with respect to their identifying information.

Anonymity set

ZIP code	Gender	Date of birth	Medical record
123XXX	X	1967-1988	[hospital visits]
123XXX	X	1967-1988	[diagnoses]
123XXX	X	1967-1988	[prescriptions]

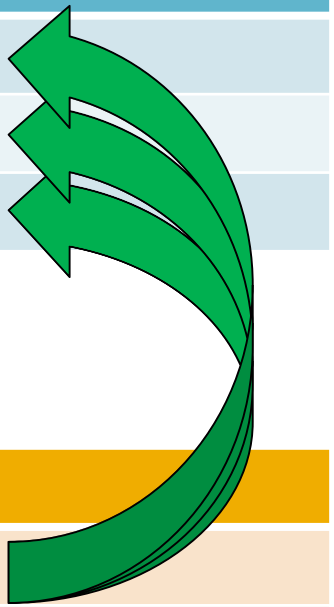
Equivalence class

L. Sweeney. "k-Anonymity: A Model for Protecting Privacy". International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (Oct. 2002), pp. 557–570.

k-anonymity

ZIP code	Gender	Date of birth	Medical record
123XXX	X	1967-1988	[hospital visits]
123XXX	X	1967-1988	[diagnoses]
123XXX	X	1967-1988	[prescriptions]

Equivalence class



ZIP code	Gender	Date of birth	Name
123456	M	1988.03.17	James
123789	F	1967.07.11	Mary
123456	F	1984.05.30	Elisa

L. Sweeney. "k-Anonymity: A Model for Protecting Privacy". International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (Oct. 2002), pp. 557–570.

Breaking *k*-anonymity

- Situation: An equivalence class is k-anonymous, but all *k* entries have the same value of the sensitive attribute.
- Consequence: Sensitive attribute of an individual can be inferred without knowing which record corresponds to the individual.

ZIP code	Gender	Date of birth	Medical record
123XXX	X	1967-1988	Diagnosis XYZ
123XXX	X	1967-1988	Diagnosis XYZ
123XXX	X	1967-1988	Diagnosis XYZ

A. Machanavajjhala et al. “l-Diversity: Privacy Beyond k-Anonymity”. ACM Transactions on Knowledge Discovery from Data 1.1 (Mar. 2007).

l-diversity

An equivalence class is *l*-diverse if it contains at least *l* "well-represented" values for the sensitive attribute. A table is *l*-diverse if every equivalence class is *l*-diverse.

ZIP code	Gender	Date of birth	Medical record
123XXX	X	1967-1988	Diagnosis ABC
123XXX	X	1967-1988	Diagnosis DEF
123XXX	X	1967-1988	Diagnosis GHI

Equivalence class

A. Machanavajjhala et al. "l-Diversity: Privacy Beyond k-Anonymity". ACM Transactions on Knowledge Discovery from Data 1.1 (Mar. 2007).

Breaking *l*-diversity

- Situation: *k*-anonymous and *l*-diverse equivalence class contains only semantically similar values of the sensitive attribute
  - *E.g. ABC, ABD, and BCD are very different from XYZ*
- Consequence: Semantic information about sensitive attribute can be inferred.

ZIP code	Gender	Date of birth	Medical record
123XXX	X	1967-1988	Diagnosis ABC
123XXX	X	1967-1988	Diagnosis ABD
123XXX	X	1967-1988	Diagnosis BCD

N. Li, T. Li, and S. Venkatasubramanian. "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity". 23rd IEEE International Conference on Data Engineering, 2007.

t-closeness

Extra material

- An equivalence class has *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*.
- A table is said to have *t*-closeness if all equivalence classes have *t*-closeness.

N. Li, T. Li, and S. Venkatasubramanian. "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity". 23rd IEEE International Conference on Data Engineering, 2007.

DIFFERENTIAL PRIVACY

Statistical database privacy

- Situation: Database held by a trusted party, who answers statistical queries about the data

Name	Gender	Age	Likes ice cream
Andrew	M	0-10	Yes
Bob	M	0-10	No
Charles	M	10-20	Yes
David	M	0-10	Yes
Elisa	F	10-20	No

Query 1: How many males like ice cream?

Result: 3

Query 2: How many kids (age 0-10) like ice cream?

Result: 2

Adversary knows Charles is excluded

Inference: Charles likes ice cream

Differential privacy

- Informal Definition: Given two databases that differ by at most one individual, it should not be possible to distinguish between these based on the results of statistical queries.
- Approach: Add random noise to each statistical query result.

!

- *C. Dwork. "Differential Privacy" 33rd International Colloquium on Automata, Languages and Programming, 2006.*
- *C. Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis" Theory of Cryptography Conference, 2006.*

Statistical database privacy

- Situation: Database held by a trusted party, who answers statistical queries about the data

Name	Gender	Age	Likes ice cream
Andrew	M	0-10	Yes
Bob	M	0-10	No
Charles	M	10-20	Yes
David	M	0-10	Yes
Elisa	F	10-20	No

Works for much large datasets. In this example, the noise makes the data unusable.

Query 1: How many males like ice cream?

Result: 3+noise = 3.1

Query 2: How many kids (age 0-10) like ice cream?

Result: 2+noise =2.7

Requirement: random noise must increase for each query ("privacy budget")

Example: US census

- Population counted every 10 years
- Census data used for research and business, but privacy?
  - 17% of household re-identifiable from 2010 data
- 2020 census data release will be based on differential privacy
  - Accurate totals on state level because needed for legal reasons
  - Noise added to data from smaller blocks
  - All data will be releases once; no privacy budget left for additional releases

S.L. Garfinkel, Deploying Differential Privacy for the 2020 Census of Population and Housing, US Census Bureau, July 2019 (link)

Questions about differential privacy

- Currently favored approach to big-data releases

Potential issues:

- Will research and business development stop when the privacy budget is used?
- Medical researchers may want the most accurate data
- Big business uses customer data internally; internal firewalls that limit data releases inside the company are rare

EU GENERAL DATA PROTECTION REGULATION (GDPR)

I’m not a lawyer, though

General Data Protection Regulation

- General Data Protection Regulation (2016/679)
  - – Adopted: 27 April 2016
  - – Effective: 25 May 2018
  - – Available in all 24 EU languages (e.g. yleinen tietosuoja-asetus)

- Regulation takes effect automatically without national laws (cf. directive)

General Data Protection Regulation

Regulation 2016/679

Article 1: Subject-matter and objectives

1. This Regulation lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data.
2. This Regulation **protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data**.
3. The **free movement of personal data within the Union** shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data.

Two aspects of privacy

- European definition of privacy emphasizes control over personal data
  - World-wide impact on global businesses

- Compare with the US: emphasis on right to be left alone
  - Interference, control, discrimination, censorship, home, spam

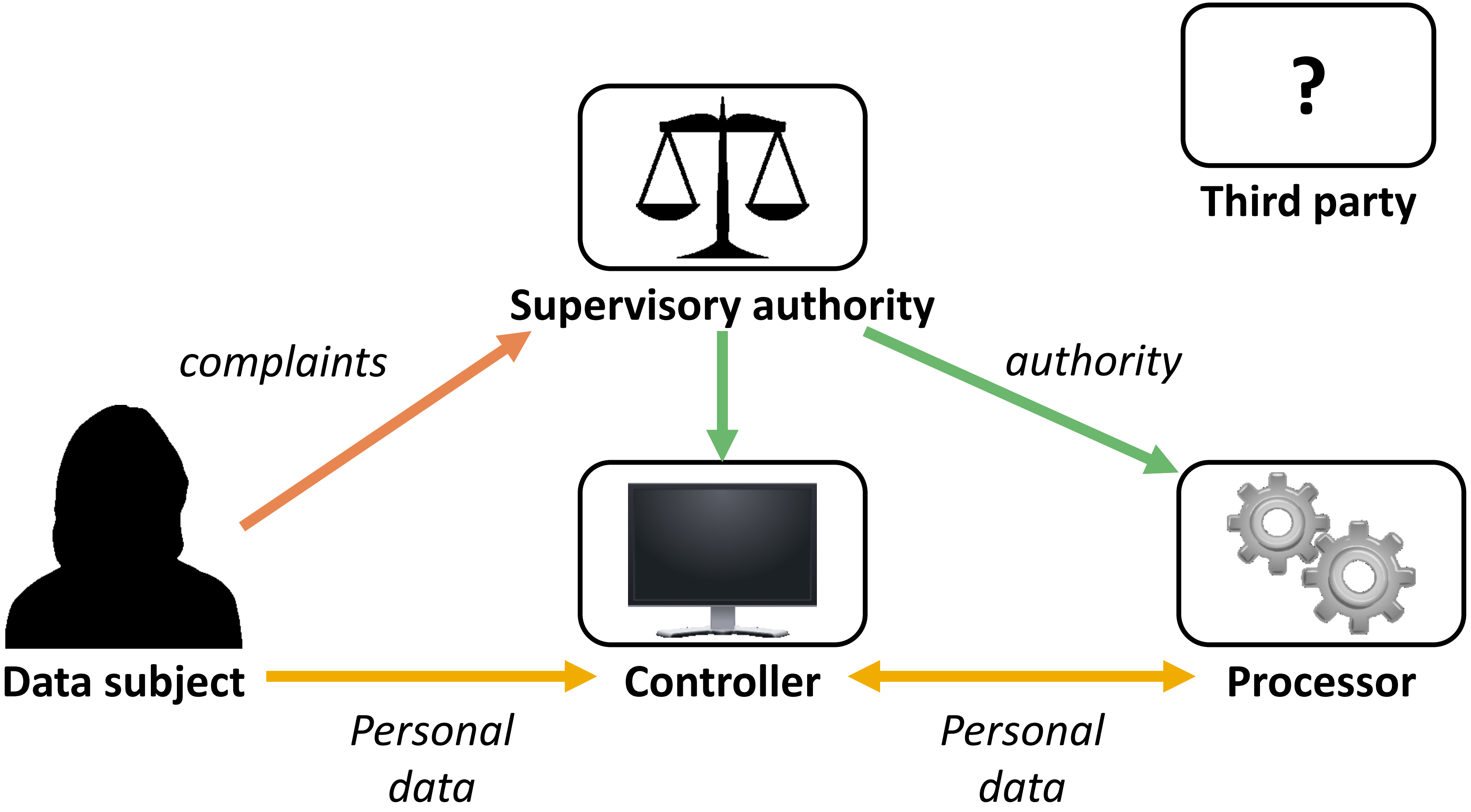
S. Warren and L. Brandeis, The Right to Privacy, 4 Harvard Law Review 193 (Dec. 15, 1890)

Chapters of the GDPR

Extra material

1. General provisions (Articles 1-4)
2. Principles (5-11)
3. Rights of the data subject (12-23)
4. Controller and processor (24-43)
5. Transfers of personal data to third countries or international organisations (44-50)
6. Independent supervisory authorities (51-59)
7. Cooperation and consistency (60-76)
8. Remedies, liability and penalties (77-84)
9. Provisions relating to specific processing situations (85-91)
10. Delegated acts and implementing acts (92-93)
11. Final provisions (94-99)

Entities in GDPR



Prohibition with exceptions

Regulation 2016/679

Article 6: Lawfulness of processing

1. Processing shall be lawful only if and to the extent that at least one of the following applies:

   a) the data subject has given **consent** to the processing of his or her personal data for one or more specific purposes;

   b) processing is **This presentation does not constitute legal** to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;

   c) processing is necessary for **compliance with a legal obligation** to which the controller is subject;

   d) processing is necessary in order to protect the This presentation does not constitute legal of the data subject or of another natural person;

   e) processing is necessary for the performance of a task carried out in the **public interest** or in the exercise of official authority vested in the controller;

MAIN CONTENT OF GDPR

Broad definition of personal data

Regulation 2016/679

30)Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags. This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them.

Strict definition of ‘consent’

Regulation 2016/679

Article 4: Definitions

11)‘consent’ of the data subject means any **freely given**, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;

- Some conditions for consent *(Article 7)*:
  - Clearly distinguishable from the other matters
  - Subject has the **right to withdraw the consent** at any time
  - Consent not freely given (i.e. not valid) if “the performance of a contract, including the provision of a service, is dependent on the consent, despite such consent not being necessary for such performance.”

Data protection by design and default

Regulation 2016/679

Article 25: Data protection by design and by default

1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, **implement appropriate technical and organisational measures**, such as **pseudonymisation**, which are designed to implement data-protection principles, such as data **minimisation**, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

Pseudonymisation

Regulation 2016/679

Article 4: Definitions

5)‘pseudonymisation’ means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

How does this compare to the type of pseudonymization we saw earlier?

Territorial scope

- Applies to businesses/processing activities outside EU, if related to EU citizens’ data.

Regulation 2016/679

Article 3: Territorial scope

1. This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not.

2. This Regulation applies to the processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union, where the processing activities are related to:

   (a)the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or

   (b)the monitoring of their behaviour as far as their behaviour takes place within the Union.

Right to erasure *(Right to be forgotten)*

Regulation 2016/679

Article 17: Right to erasure ('right to be forgotten')

1. The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies:

 a) the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed

 b) the data subject withdraws consent on which the processing is based … and where there is no other legal ground for the processing

 c) the data subject objects to the processing … and there are no overriding legitimate grounds for the processing, …

 d) the personal data have been unlawfully processed

Data portability

Regulation 2016/679

Article 20: Right to data portability

1. The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, **in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller** without hindrance from the controller to which the personal data have been provided, where:

   a) the processing is based on consent pursuant to point (a) of Article 6(1) or point (a) of Article 9(2) or on a contract pursuant to point (b) of Article 6(1); and

   b) the processing is carried out by automated means.

2. In exercising his or her right to data portability pursuant to paragraph 1, the data subject shall have the right to have the personal data transmitted directly from one controller to another, where technically feasible.

Right to object

Extra material

Regulation 2016/679

Article 21: Right to object

2. Where personal data are processed for direct marketing purposes, the data subject shall have the right to object at any time to processing of personal data concerning him or her for such marketing, which includes profiling to the extent that it is related to such direct marketing.
3. Where the data subject objects to processing for direct marketing purposes, the personal data shall no longer be processed for such purposes.

Automated decision making

Regulation 2016/679

Article 22: Automated individual decision-making, including profiling

1. The data subject shall have the **right not to be subject to a decision based solely on automated processing, including profiling,** which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:

   a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

   b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

   c) is based on the data subject's explicit consent.

Data Breach Notifications

Regulation 2016/679

> **Article 34: Communication of a personal data breach to the data subject**
>
> 1. When the personal data breach is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall communicate the personal data breach to the data subject without undue delay.

- Exceptions:
  - Controller implemented appropriate technical/organisational protection measures that render the data unintelligible to anyone not authorised to access it, such as encryption;
  - The controller has taken subsequent measures which ensure that the high risk is no longer likely to materialise;
  - It would involve disproportionate effort (use public announcement instead)

Data Protection Officer

Extra material

- Required when controller/processor performs regular, systematic monitoring on a large scale
- Selected based on professional qualities and expert knowledge (which their employer is obliged to help them maintain)
- Involved in all issues which relate to the protection of personal data
- Must have adequate resources and report directly to highest level of management
- Tasks and duties
  - Advise and inform
  - Monitor compliance
  - Contact point for supervisory authority

Administrative fines

Regulation 2016/679

Article 83: General conditions for imposing administrative fines

4. Infringements of the following provisions shall, in accordance with paragraph 2, be subject to administrative fines up to 10 000 000 EUR, or in the case of an undertaking, up to 2 % of the total worldwide annual turnover of the preceding financial year, whichever is higher…
5. Infringements of the following provisions shall, in accordance with paragraph 2, be subject to administrative fines up to 20 000 000 EUR, or in the case of an undertaking, up to 4 % of the total worldwide annual turnover of the preceding financial year, whichever is higher…

SUMMARY

Open questions

- Is data protection regulation hurting or helping business?

- Have we given up our expectations of privacy?

List of key concepts

- De-anonymization: pseudonymity, re-identification
- Anonymization techniques: k-anonymity, anonymity set, l-diversity, t-closeness
- Differential privacy: privacy budget
- GDPR: directive vs. regulation, personal data, consent, pseudonymization, right to erasure, data portability