The data economy
A glossary

The data economy comes with its own terminologies and buzzwords (Table 1), stakeholders (Table 2), and activities (Table 3). This glossary aims to help readers navigate this data-driven environment.

Table 1. Key terms

Term	Definition	Source
Personal data	• Any information relating to an identified or identifiable living individual ("natural person"; "data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data. • Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data.	GDPR Article 4 (1) https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en
Genetic data	• Personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question.	GDPR Article 4 (13)
Biometric data	• Personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic (e.g., fingerprints) data.	GDPR Article 4 (14)
Sensitive personal data	• Personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms and the context of their processing could create significant risks to those fundamental rights and freedoms. • Example of sensitive data are race or ethnic origin, political opinions, religion or beliefs, trade union membership, genetic data, data on health status or sexual orientation.	GDPR Preamble 10, 51, 71
Health data or data concerning health	• Personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status.	GDPR Article 4 (15)
Protected personal data (PPD)	• Any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.	EMA Policy 0070
Individual patient data (IPD)	• Individual data separately recorded for each participant in a clinical study.	EMA Policy 0070
Aggregated data	• In the context of clinical studies, represent statistical data about several individuals that has been combined to show general trends or values without identifying individuals within the data.	EMA Policy 0070

Term	Definition	Source
Anonymous data	• Information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable, also called "de-identified data". • Anonymised data is not considered personal data and is out of scope of GDPR. For data to be truly anonymised, the anonymisation must be irreversible. • Clinical trial data are not fully anonymised.	GDPR Preamble 26 EMA Policy 0070
Pseudonymous data	• Processing of personal data in such a manner that the 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. Personal data that has been de-identified, encrypted or pseudonymised but can be used to re-identify a person remains personal data and falls within the scope of the GDPR. • Clinical trial data are pseudonymised data.	EMA Policy 0070
Identifier	• Information that can directly or indirectly identify a data subject.	EMA Policy 0070
Direct identifiers	• Data elements that permit direct recognition or communication with the corresponding individuals. • Examples: name, email, phone number, address, patient identification number.	EMA Policy 0070
Indirect ("quasi") identifiers	• Data elements representing an individual's background information that can indirectly identify data subjects. • Examples: demographics, characteristics, attributes, socio-economic information.	EMA Policy 0070
Big data	• Term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Big data has one or more of the following characteristics "the three Vs"): high Volume, high Velocity or wide Variety. Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data. For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media – much of it generated in real time and at a very large scale.	https://www.ibm.com/analytics/hadoop/big-data-analytics
Real world data (RWD)	• Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.	https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
Real world evidence (RWE)	• Clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD. RWE can be generated by different study designs or analyses, including but not limited to, randomised trials, including large simple trials, pragmatic trials, and observational studies (prospective and/or retrospective).	https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

Table 2. Key stakeholders

Term	Definition	Source
Data subject	• Any natural ("living") person whose personal data is being processed; see definition of data processing.	GDPR
Identifiable natural person	• One who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.	GDPR Article 4 (1)
Data controller	• Natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes, conditions and means of the processing of personal data.	GDPR Article 4 (7)
Data processor	• A natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller.	GDPR Article 4 (8)
Data recipient	• A natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not. However, public authorities which may receive personal data in the framework of a particular inquiry in accordance with Union or Member State law shall not be regarded as recipients; the processing of those data by those public authorities shall be in compliance with the applicable data protection rules according to the purposes of the processing.	GDPR Article 4 (9)
Data exporter	• A natural or legal person, public authority, agency or another body who transfers personal data from the EU/EEA to a non-EEA country or international organisation. • In the context of international cooperation due to COVID-19, international transfers of health data for the purpose of scientific research outside of the EEA are allowed under certain conditions. The exporter must meet GDPR requirements for data transfers.	European Data Protection Board Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak
Data scientist	• A professional engaged in data science. A data scientist requires an integrated skillset spanning mathematics, machine learning, AI, statistics, databases, and optimisation.	Dhar V. Data Science and Prediction Communications of the ACM, December 2013, Vol. 56 No. 12, pp. 64–73 10.1145/2500499
Data protection authorities (DPA)	• Independent public authorities in each EU member state that supervise, through investigative and corrective powers, the application of the data protection law.	GDPR

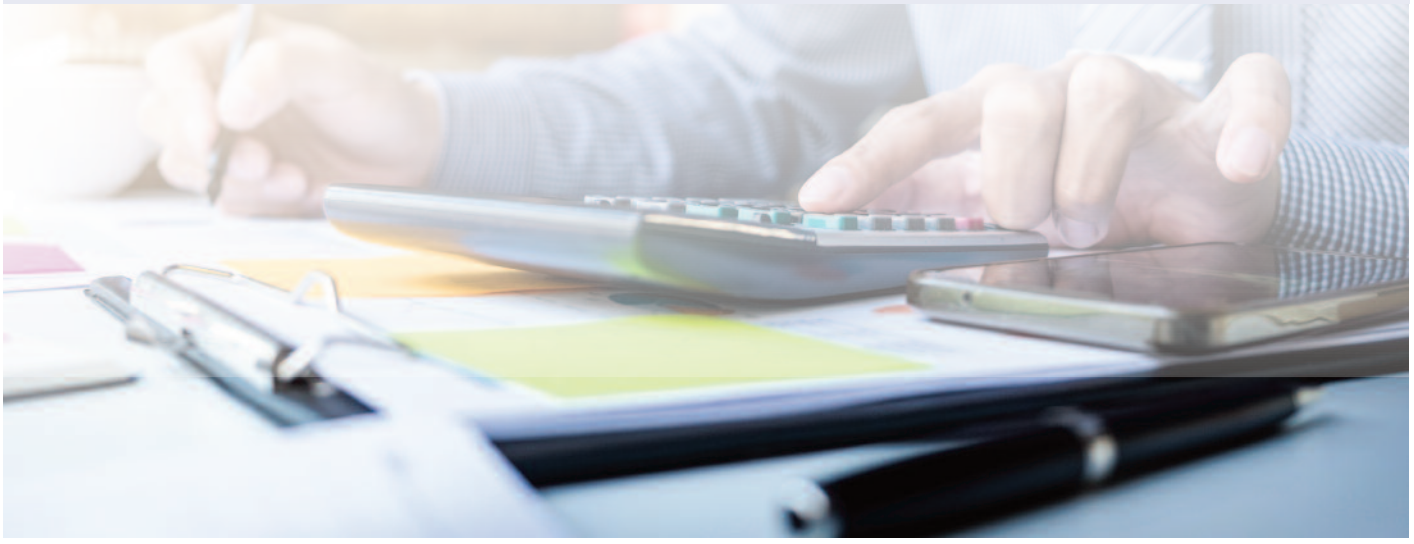

Table 3. Key activities

Term	Definition	Source
Data science	• The study of the generalisable extraction of knowledge from data.	Dhar V. Data Science and Prediction Communications of the ACM, December 2013, Vol. 56 No. 12, pp. 64–73 10.1145/2500499
Data processing	• Processing covers a wide range of operations performed on personal data, including by manual or automated means. It includes the collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction of personal data. • Data collection, analysis, reporting, and sharing during a clinical trial constitute data processing.	https://ec.europa.eu/info/law/ law-topic/data-protection/ reform/what-constitutes-data-processing_en GDPR Article 4 (2)
Data sharing	• Making data available to others.	
Data transparency	• In the context of GDPR, personal data must be processed "lawfully, fairly and in **a transparent manner** in relation to the data subject." This covers the data subject's right to information (at the minimum) about data use, storage, access, and dissemination. • In the context of clinical research, data transparency is sharing of clinical research data to meet regulatory requirements and advance the generation of critical scientific knowledge.	GDPR Article 5 (1a) https://www.phusewiki.org/docs/ WorkingGroups/New%20 Template% 20Deliverables/Data%20Transparency/ Clinical%20Trial%20Transparency% 20and%20Disclosure-%20A%20 Global% 20View.pdf
Personal data protection	• The act of protecting the rights of data subjects.	GDPR
Anonymisation	• The process of rendering data into a form which does not identify individuals and where identification is not likely to take place. Anonymisation is irreversible.	EMA Policy 0070

Table 3 continued opposite

Term	Definition	Source
Pseudonymisation	• The process of replacing one attribute (typically a unique attribute) in a record by another. The natural person is may still be identified indirectly but pseudonymisation reduces the linkability of a dataset with the original identity of a data subject.	EMA Policy 0070
Re-identification or de-anonymisation	• The process of analysing data or combining it with other data with the result that individuals become identifiable.	EMA Policy 0070
Anonymisation techniques	• Techniques to mitigate risks of re-identification of the individual data subjects; Effective ness of these techniques are based on three criteria: singling out; linkability, and inference.	EMA Policy 0070
Proactive anonymisation (in the context of the CSR)	• Use of anonymisation techniques for generating anonymised datasets and generating another copy of the CSR using anonymised datasets. • If it is necessary to discuss any individual subject level information in text, consider using **proactively anonymised clinical document text and data presentations** that maintain data meaning, remain in context AND conform to **current** minimum standards for de-identifying data.	https://www.phusewiki.org/docs/Deliverables/Narratives%20Phuse%20Subgroup%20Writeup_Final_11.21%20(1).pdf https://www.core-reference. org/core-reference/
Retrospective anonymisation	• Generally, the use of redaction or masking to anonymise text and data; also known as reactive data anonymisation.	EMA Policy 0070 https://www.phusewiki.org/docs/Deliverables/Narratives%20Phuse%20Subgroup%20Writeup_Final_11.21%20(1).pdf
Big data analytics	• Use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.	https://www.ibm.com/analytics/hadoop/big-data-analytics

Raquel Billiones
medical.writing@billiones.biz

The data economy