Data Analytics Glossary

|  |  |
| --- | --- |
| General Data Terms | |
| Word | Meaning |
| Aggregation | The process of combining individual data to summarise, group or categorise. |
| Analysis | The process of examining data to spot patterns, trends or insights. |
| Anonymisation | The process of removing or encrypting personal identifiable information from data to protect privacy. |
| Archive | An archive is a physical or virtual repository which holds information which no longer has a primary business value but is maintained for a further period for research or historical value. |
| Big Data | Large and complex datasets that require advanced processing techniques. |
| Caldicott Guardian | This is the person appointed to ensure that patient personal health information is adequately safe-guarded (in accordance with the NHS Caldicott Principles). |
| Cloud | This describes where data is held, and data services (such as storing, processing and software) can be conducted off-premises (usually by a ‘cloud provider’ that may be in any part of the world). This occurs when the cloud provider is deemed to have the technology and capacity to do this more effectively than ‘in-house’ for a particular purpose. |
| Correlation | A relationship or connection between two or more values. |
| Dashboard | The display of multiple visuals made to represent data. |
| Data | Information, facts or statistics collected together for reference or analysis. |
| Data Breach | This is when something has occurred which may contravene data protection or other law. This could for example be when data is lost, shared inappropriately or where any of the data protection principles have been breached (e.g. lack of transparency). |
| Data Ethics | Over and above the legal frameworks which exist such as Data Protection, there are still questions around whether it is right and proper for certain data processing to occur. See the SCC Ethical Data Stewardship Charter for more detail. |
| Data Mashing | Data Mashing is the process of integrating and transforming data from multiple sources to derive insight and inform decision making. Providing a unified “big picture” view of data, it provides analysts with the ability to utilise collective datasets to gain a more comprehensive understanding of the underlying information. |
| Data Matching (for identification) | This is where data from one dataset is read alongside data from other datasets in order to identify a person with an agreed level of accuracy. |
| Data Measure | Applying a measure to data involves using a mathematical equation to analyse, summarise or transform data to answer specific questions or gain new insight. |
| Data Model | An organisation of various datasets which standardises how they relate to each other. |
| Data Opt-in | Under data protection law where the lawful basis requires ‘consent’, the individual will need to consciously opt-in (rather than merely opt-out) before processing can commence. Note:  most of the data processing undertaken by SODA is based on other lawful bases and so an opt-in or out does not arise. |
| Data Opt-out | This is where an individual is given the opportunity to decline having their data captured and processed. This can be due to data protection law (e.g. an existing customer declining to be on a mailing list) or because of national or local policy (e.g. NHS allowing persons to decline having data used for non-direct care purposes). |
| Data Owner | This is the person or organisation which is responsible for data on a day-to-day basis. Note: this is not the same as ‘data controller’ which is a legal term to describe the person or entity which is responsible for the purposes and manner to which personal data is processed in accordance with GDPR/DPA. |
| Data Owner/Controller | An organisation, team or individual responsible for the possession and storage of logically grouped data. The Data Owner/Controller also determines the means and reasons for which the data is used. |
| Data Processors | An organisation, team or individual who changes and reorganises data on behalf of a Data Controller. |
| Data Protection Act (DPA) | The Data Protection Act 2018 is the UK’s implementation of the General Data Protection Regulation (GDPR). Controls established to dictate how your personal information is used by organisations, businesses or the government. The use of personal data requires data processors and owners to follow a strict set of governance rules known as the ‘data protection principles’. |
| Data Protection Impact Assessment (DPIA) | The process used to identify and minimise the risk that occurs from the processing of data. Designed to minimise risk at the earliest opportunity, an DPIA is also required to comply with GDPR. |
| Data Protection Officer | This is the individual appointed under the terms of GDPR/DPA to advise on data protection compliance. |
| Data Protection Principles | To comply with the Data Protection Act data processors must follow the ‘data protection principles’. To ensure the information is:   * used fairly, lawfully and transparently * used for specified, explicit purposes * used in a way that is adequate, relevant and limited to only what is necessary * accurate and, where necessary, kept up to date * kept for no longer than is necessary * handled in a way that ensures appropriate security, including protection against unlawful or unauthorised processing, access, loss, destruction or damage   Sensitive information requires stricter governance compromised of the following information:   * race * ethnic background * political opinions * religious beliefs * trade union membership * genetics * biometrics (where used for identification) * health * sex life or orientation   There are separate safeguards for personal data relating to criminal convictions and offences. |
| Data Schema | The blueprint of a database, showing what datasets are included and what variables make up each database. |
| Data Subject | In the context of data protection law, the ‘data subject’ is the living person that the data relates to. He/she has certain ‘data subject rights’ in law. |
| Data Visualisation | A visual made to display data. |
| Data Warehouse | Large store of integrated data, accumulated from a range of sources within an organisation. |
| Database | An organised collection of structured information or data. |
| Dataset | A collection of related data or information. |
| Deviation | The amount by which a single measurement differs from a fixed value. |
| Encryption | The process of encoding to translate data into a different form or code. Protecting the data so that only an authorised person can access via decryption or password access. |
| GDPR | General Data Protection Act (2018). Even with Brexit, the GDPR is still relevant and needs to be read in conjunction with the UK Data Protection Act (2018). |
| General Data Protection Regulation (GDPR) | The GDPR form part of European Union (EU) law that came into effect on 25th May 2018. GDPR governs the way in which we can use, process, and store personal data (information about an identifiable, living person). |
| Governance | This is an established system to control how an organisation operates and the mechanisms by which its employees are held to account. Encompassing risk management, compliance, administration, and ethics. |
| Hypothesis | A proposed explanation based on limited evidence. |
| Inference | A conclusion reached based on evidence or reasoning. |
| Information | Information is something in written form – digital or paper - which is understandable. This can be made up from different building blocks of data. |
| Information Governance (IG) | Information Governance (IG) is the framework for handling information in a secure and confidential manner. Allowing organisations to manage data securely.  Determining legal and regulatory compliance for information storage, use, sharing and disposal. |
| Information Sharing Agreement (ISA) / Data Sharing Agreement (DSA) | An Information Sharing Agreement (ISA)/Data Sharing Agreement (DSA) identifies the lawful basis for the use and sharing of personal information, and the extent and nature of the personal information to be shared. |
| Mean | The average value of a set of numbers. |
| Median | The middle value of a dataset organised from highest to lowest. |
| Metadata | This simply means ‘data about data’. When data is captured and stored by an IT system there is often other data which contains things like date, when modified, when deleted, name of creator etc. |
| Mode | The most frequently occurring value in a dataset. |
| Outlier | Information or value that differs significantly from the dataset it is part of. |
| Positive Patient Identification (PPI) | This describes the process where there is a very high level of assurance that the data/information does relate to the person who is being cared for. |
| Privacy by Design | This is a term from GDPR which describes how all aspects of data protection compliance, privacy and security should be considered, risk assessed and ‘baked into’ the design (rather than trying to retrofit at a later stage). |
| Privacy Notice | One of the key tenets of data protection law is transparency. One of the ways in which an organisation explains to its customers (data subjects) what personal data is collected, for what purposes and whether it is shared etc is via a privacy notice. This can be on a website, leaflet or other medium. Note: a privacy notice is only one tool, and transparency needs to be woven into every aspect of data processing. |
| Pseudonymisation | The process of replacing identifiable information with artificial identifiers or pseudonyms to protect privacy whilst still allowing for data analysis. |
| Qualitative Data | Data that represents information and concepts that are not represented by numbers. |
| Quantitative Data | Data in a number format, any data that can be counted, measured or shown as a number. |
| Record | A record is information which has content, context and structure. This forms the ‘official’ record (e.g. for a patient, taxpayer etc.) as opposed to unstructured information. |
| Reidentification | The process of taking anonymous data and turning it back to its original identifiable form. |
| Retention/disposal schedule | This is the formal policy of an organisation which sets out how long a type or class of record should be retained for legal and business reasons (and when deleted or moved etc.) |
| Sampling | Selecting a subset of information from a larger set of information. |
| Senior Information Responsible Owner (SIRO) | This is used in the context of UK public sector to describe the senior person (e.g. director, Chief Executive etc.) who has been appointed to manage top-level information risks relating to data assets within the organisation. |
| Statistic | A value or data obtained from a study of a large quantity of numerical data. |
| Third Party | In the context of data protection law, a third party is anyone other than the data controller (and their agreed processors) and the data subject. |
| Variable | A characteristic of data that can be measured or observed. |