**CODATA/RDA Data Science Summer Schools Vocabulary List**

Module: Open and Responsible Science Citizenship

| Term | Definition |
| --- | --- |
| Open Science | Umbrella term for open movements, including open software, data, publications, methodologies, hardware, and educational materials  The practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods  The belief that the products of scientific research should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.  Useful ref: <https://www.fosteropenscience.eu/> |
| Responsible Conduct of Research | The practice of scientific investigation with integrity.  Involves the awareness and application of established professional norms and ethical principles in all activities related to scientific research.  Useful ref: <https://www.ncbi.nlm.nih.gov/pubmed/25009901> |
| Beneficence | To ensure that your actions cause good |
| Non-maleficence | To avoid causing harm through one’s actions |
| Integrity | 1. to have strong moral principles 2. to be whole and undivided |
| Bias | The inclination or prejudice for or against one person or group, especially in a way considered to be unfair. |
| Open and Responsible Science Citizenship | Extension of RCR and OS values that highlights how membership to research community comes with freedom, rights and duties. Science citizens should ensure they uphold key rules and legislation, participate in community activities and offer service, and protect the community and its resources from misuse. |

Module: Research Data Management

| Term | Definition |
| --- | --- |
| Data steward | Data stewardship is a shared responsibility between Principal Investigators and data stewards. Principal Investigators are responsible for, and data stewards provide support for: (a) Data collection, data integration, or reuse of existing data; (b) Review of data quality; (c) Description of scientific workflow/process; (d) Provision of standards-compliant metadata; and, (e) Submission of data and data productions. Data stewards are responsible for, and Principle Investigators are consulted and informed on: (a) Preservation of data and data products; and, (b) Provision of formats (e.g., web services, NetCDF, etc.) for data discovery and integration. In addition, Principal Investigators are also responsible for data citation, as appropriate, when preparing documentation, reports, or references. <https://dictionary.casrai.org/Data_steward> |
| Research data | Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data. <https://dictionary.casrai.org/Research_data> |
| Open data | Structured data that are accessible, machine-readable, usable, intelligible, and freely shared. Open data can be freely used, re-used, built on, and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike. <https://dictionary.casrai.org/Open_data> |
| Curation lifecycle model | A curation lifecycle model documents the relationships between all the stages in the existence of digital information, to enable active management of the resource over time thus maintaining accessibility and usability.  <http://www.dcc.ac.uk/digital-curation/glossary> |
|  |  |
| Curation | The activity of managing and promoting the use of data from their point of creation to ensure that they are fit for contemporary purpose and available for discovery and reuse. For dynamic datasets this may mean continuous enrichment or updating to keep them fit for purpose. Higher levels of curation will also involve links with annotation and with other published materials.<https://dictionary.casrai.org/Curation> |
| Digital preservation | The series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation is defined very broadly and refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change. Those materials may be records created during the day-to-day business of an organisation; ""born-digital"" materials created for a specific purpose (e.g. teaching resources); or the products of digitisation projects. This definition specifically excludes the potential use of digital technology to preserve the original artefacts through digitisation. RELATED TERM. Digitisation; Preservation <https://dictionary.casrai.org/Digital_preservation> |
| Research data management (RDM) | Data Management refers to the storage, access and preservation of data produced from a given investigation. Data management practices cover the entire lifecycle of the data, from planning the investigation to conducting it, and from backing up data as it is created and used to long term preservation of data deliverables after the research investigation has concluded. Specific activities and issues that fall within the category of data management include: File naming (the proper way to name computer files); data quality control and quality assurance; data access; data documentation (including levels of uncertainty); metadata creation and controlled vocabularies; data storage; data archiving and preservation; data sharing and reuse; data integrity; data security; data privacy; data rights; notebook protocols (lab or field). RELATED TERM. Data stewardship <https://dictionary.casrai.org/Research_data_management> |
| FAIR | This term has been gaining traction in recent years and refers to research data that are findable, accessible, interoperable and reusable. |
| Data management plan (DMP) | A formal statement describing how research data will be managed and documented throughout a research project and the terms regarding the subsequent deposit of the data with a data repository for long-term management and preservation.<https://dictionary.casrai.org/Data_management_plan>  Data management plans outline how data will be created, managed, shared and preserved, justifying any restrictions that need to be applied.<http://www.dcc.ac.uk/resources/how-guides/develop-data-plan> |
| Metadata | Literally, "data about data"; data that defines and describes the characteristics of other data, used to improve both business and technical understanding of data and data-related processes. Business metadata includes the names and business definitions of subject areas, entities and attributes, attribute data types and other attribute properties, range descriptions, valid domain values and their definitions. Technical metadata includes physical database table and column names, column properties, and the properties of other database objects, including how data is stored. Process metadata is data that defines and describes the characteristics of other system elements (processes, business rules, programs, jobs, tools, etc.). Data stewardship metadata is data about data stewards, stewardship processes and responsibility assignments. <https://dictionary.casrai.org/Metadata> |
| Persistent identifier (PID) | A persistent identifier is a long-lasting reference to a digital object that gives information about that object regardless of what happens to it. Developed to address "link rot," a persistent identifier can be resolved to provide an appropriate representation of an object whether that objects changes its online location or goes offline. SYNONYM. PID <https://dictionary.casrai.org/Persistent_identifier> |
| Digital Object Identifier (DOI) | A name (not a location) for an entity on digital networks. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks. A DOI is a type of Persistent Identifier (PID) issued by the International DOI Foundation. This permanent identifier is associated with a digital object that permits it to be referenced reliably even if its location and metadata undergo change over time. <https://dictionary.casrai.org/Digital_Object_Identifier> |
| Data repository | Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis.<https://dictionary.casrai.org/Repository> |
| Trusted data repository | Trusted data repositories undergo regular assessments according to a set of rules such as defined by Data Seal of Approval (DSA) or TRAC (ISO 16363). <https://dictionary.casrai.org/Trusted_Digital_Repository> SYNONYM: TDR  Please note that the term “trustworthy” is sometimes preferred to “trusted”, because it’s awkward to claim yourself that your repository is “trusted”.  Please note also that the CoreTrustSeal has succeeded the Data Seal of Approval: <https://www.coretrustseal.org/> |
| OAIS | An Open Archival Information System (or OAIS) is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. The term OAIS also refers, by extension, to the ISO 14721:2012 OAIS Reference Model for an OAIS. <https://en.wikipedia.org/wiki/Open_Archival_Information_System> |
| Ontology | A formal, explicit specification of a shared conceptualisation. <https://doi.org/10.1016/S0169-023X(97)00056-6> |
| License | Signed agreements to exploit a piece of IP such as a process, product, data, or software. <https://dictionary.casrai.org/License> |
| Linked open data | Data where relationships/connections between them are available to allow easy data access. A typical case of a large Linked dataset is DBPedia (<http://dbpedia.org/>), which essentially makes the content of Wikipedia available in RDF. This related collection of interrelated datasets is stored on the Web and available via a common format -RDF.<http://www.w3.org/standards/semanticweb/data#summary>  <https://dictionary.casrai.org/Linked_open_data> |
| Uniform Resource Identifier (URI) | A string of characters used to identify or name a resource on the Internet. Such identification enables interaction with representations of the resource over a network, typically the World Wide Web, using specific protocols. SYNONYM. URI <https://dictionary.casrai.org/Uniform_resource_identifier> |

Module: Git

| Term | Definition |
| --- | --- |
| Git | an open source, distributed version-control system |
| GitHub | a platform for hosting and collaborating on Git repositories |
| Commit | a Git object, a snapshot of your entire repository compressed into a SHA |
| Branch | a lightweight movable pointer to a commit |
| Clone | a local version of a repository, including all commits and branches |
| Remote | a common repository on GitHub that all team member use to exchange their changes |
| Fork | a copy of a repository on GitHub owned by a different user |
| Pull request | a place to compare and discuss the differences introduced on a branch with reviews, comments, integrated tests, and more |
| HEAD | representing your current working directory, the HEAD pointer can be moved to different branches, tags, or commits when using git checkout |

Module: Information Security

| Term | Definition |
| --- | --- |
| Malicious | Something with an intent to do harm. |
| Vulnerability | A weakness that can be exploited by a malicious attacker. <https://en.wikipedia.org/wiki/Vulnerability_(computing)> |
| Encryption | The process of converting information or data into a code, especially to prevent unauthorized access. |
| Cipher | An algorithm for performing encryption. <https://en.wikipedia.org/wiki/Cipher> |
| Integrity | Internal consistency or lack of corruption in data. |
| Hash | A cryptographic hash function is a hash function which takes an input (or 'message') and returns a fixed-size string of bytes. <https://simple.wikipedia.org/wiki/Cryptographic_hash_function> |
| Non-repudiation | A situation where someone cannot deny having done something, often linked to being authenticated. |
| Symmetric encryption | Encryption that uses a single key to encrypt and decrypt. |
| Asymmetric encryption | Encryption that uses one key to encrypt and a different key to decrypt. |
| Public and Private keys | A pair of linked keys used for asymmetric encryption, a message encrypted with one of the keys can be decrypted with the other. Typically one key is made public (the Public Key) and the other is kept private (the Private Key). By using different methods for asymmetric encryption it is possible to transmit confidential, authenticated messages. |

Module: Data visualization

| Term | Definition |
| --- | --- |
| Color blindness | Color blindness, also known as color vision deficiency, is the decreased ability to see color or differences in color. Colorblindness is not a total loss of color vision. Colorblind people can recognize a wide range of colors. But certain ranges of colors are hard to distinguish. Red-green color blindness is the most common form, followed by blue-yellow color blindness and total color blindness.Red-green color blindness affects up to 8% of males and 0.5% of females of Northern European descent. The ability to see color also decreases in old age. |
| Color blind friendly color palettes | <http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/#a-colorblind-friendly-palette>  Good graphic design avoids using color coding or using color contrasts alone to express information; this not only helps color blind people, but also aids understanding by normally sighted people by providing them with multiple reinforcing cues. |
| Misleading graph | Misleading graphs are sometimes deliberately misleading and sometimes it’s just a case of people not understanding the data behind the graph they create. The “classic” types of misleading graphs include cases where:   * The Vertical scale is too big or too small, or skips numbers, or doesn’t start at zero. * The graph isn’t labeled properly. * Data is left out. |
| Data (variable) type | There are two types of variables—Quantitative(numericals) and Qualitative(categoricals). |
| Qualitative (categorical) variable | Descriptions of groups or things, like “breeds of dog” or “voting preference”. The categorical variable can be divided into ordinals and nominals (non-ordinals).  The ordinal variable can be ordered. For example, spiciness can be mild, medium, or hot. Even if they are not numbers per se, they can still be ordered. While nominals can’t be ordered. Two simple examples are sex, male or female, or regions of the states Northeast, South, Central, West. |
| Quantitative(numerical) variable | Quantitative variables are numerical variables: counts, percents, or numbers.  We can further divide numerical data into continuous and discrete. Continuous variables are those that can take any value such as heights if measured with enough precision. For example, a pair of twins maybe 68.12 inches and 68.11 inches respectively.  Counts such as population sizes are discrete because they have to be round numbers. |

Module: Computational Infrastructures

| Term | Definition |
| --- | --- |
| Clustered Computing | A computer cluster is a set of loosely or tightly connected computers that work together so that, in many respects, they can be viewed as a single system. In most cases computers that are clustered are all within the same organization domain. |
| High Performance Computing (HPC) | High Performance Computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering, or business. HPC is synonymous with parallel computing methods. |
| High Throughput Computing (HTC) | High-throughput computing describes the use of many computing resources over long periods of time to accomplish a computational task. |
| Cloud Computing | Cloud computing is the on-demand availability ofcomputer system resources, especiallydata storage andcomputing power, without direct active management by the user. Most cloud resources are provisioned in the form of Virtual Machines (or hardware-level virtualization). |
| Containers | Containers or OS-level virtualization refers to an operating system paradigm in which the kernel allows the existence of multiple isolated user-space instances. |
| HTCondor, Slurm, PBS, SGE, LSF | Examples of batch computing software. These allow headless (non-interactive) workflows to be managed. |
| Virtual Machines | Provide hardware-level virtualization which emulate a computer system through the use of software. |

Module: Applications -Classification networks

| ANN | Artificial Neural Network |
| --- | --- |
| Neural Network | Network of neurons, as in the brain |
| Classification | Assigning a data object to one of two (or more,, but usually just two) classes |
| Signal and Background | Example of two such classes. Objects which are signal are interesting and analysed further: background objects are ignored |
| Neuron | The brain is made of billions of these |
| Node | Equivalent in ANN of a brain neuron |
| weight | In this context: numerical values belonging to a node that dictate what it does |
| sigmoid | The function 1/(1+exp(-x)). Rises from 0 to 1 |
| tanh | Hyperbolic tangent. The function (exp(x)-exp(-x))/(exp(x)+exp(-x)). Rises from -1 to +1. |
| Perceptron | A type of neural network, with nodes arranged in layers |
| HIdden layers | Layers in a perceptron that are not the first (input) or last (output) layer |
| Synchronised | Happens at the same time |
| Training | Adjusting the weights in a network to improve its performance |
| back-propagation | A method of training |
| Supervised learning | Training that involves running the system on samples for which the classification is known already |
| Unsupervised learning | Training that involves running the system on samples for which the classification is not known |
| Event | Equivalent to ‘data object’ |
| Type I error | Rejecting a wanted signal event |
| Type II error | Accepting an unwanted (background) event |
| Prior probabilities | The fractions of the sample in the various classes, before any classification |
| Over training | Training a network on a sample so intensely that it recognises features of specific individual events, giving a misleading indication of its performance on general samples. |
| Training sample | Sample used for training |
| Testing sample | Sample not used for training, but to measure progress. `Avoids being misled by over training |
| Validation | Establishing how well the network performs |
| ROC plot | “Receiver Operating Characteristic”. A meaningless acronym. Actually a 2D plot of the probability of accepting signal and backgRound events, for the range of cut values. |
| Regression | Using one (or more) variables to predict the value of another |
| PCA | Principal Component Analysis. Transforming a set of variables by a generalised rotation so that they are uncorrelated. |
| BDT | Boosted Decision Trees. An alternative to Neural Networks. See Lecture 2 slide 5. |
| SVM | Support Vector Machines. An alternative to Neural Networks. See Lecture 2 slide 6. |
| BLT | Bacon, Lettuce and Tomato. In this table by mistake |
| GAN | Global Adversarial Networks. Used to generate apparently real data. See lecture 2 slide 9 |
| Deep learning | Means whatever you want it to mean. I take it as neutral networks with many (>5) layers. |
| Kohonen Network | An unsupervised learning network. See lecture 2 slides 10-17 |
| SOM | Self-organising Map. What Kohonen (and some others) calls a Kohonen network. |

Module: Machine Learning - Recommender Systems

| AI | Artificial Intelligence - A branch of computer science |
| --- | --- |
| ML | Machine Learning - one of several AI techniques |
| Supervised Learning | A method in ML where we already know the answers we want |
| Unsupervised Learning | We want to find unknown structures, patterns or trends within a data set |
| Item | The entities a system recommends |
| Attribute | The characteristics of an item |
| Query | The information a system uses to make recommendations |
| Embedding | A mapping from a discrete set (in this case, the set of queries, or the set of items to recommend) to a vector space called the embedding space. Many recommendation systems rely on learning an appropriate embedding representation of the queries and items. |
| Recommender/Recommendation Systems | A subclass of information filtering system that seek to predict the ‘rating’ or ‘preference’ that user would give to an item |
| Information Filtering | Delivery of information that the use us likely to find interesting or useful |
| Content-based Filtering | A content-based filtering system selects items based on the correlation between the content of the items and the user’s preferences |
| Collaborative Filtering | A collaborative filtering system chooses items based on the correlation between people with similar preferences. |
| Measure of Effectiveness | The effectiveness of a recommender system is usually determined by the ratios precision and recall. |
| Precision | Precision indicates how well the retrieved information match a user’s interest |
| Recall | Recall is the ratio of number of relevant documents retrieved to the total number of relevant documents in the collection |
| Vectors | An efficient way to represent data as an array stored in memory as the models used for recommender systems are fundamentally mathematical in nature. |
| TF-IDF | TF- IDF stands for Term Frequency and Inverse Document Frequency .TF-IDF helps in evaluating importance of a word in a document. |
| Cosine Similarity | Cosine similarity is a measure of similarity between two non-zero vectors. |