University of Reading

Institute for Environmental Analytics

BIG DATA AND THE ENVIRONMENT: A-Z GLOSSARY

This glossary explains some of the words and phrases that we use in the course. It's a work in progress, so if there's anything you think is missing, let us know in the comments and we'll add them to the document for next time.

Name	Description
Algorithm	instructions to perform calculations and solve problems, often in computer programming.
Analytics	gathering, processing and exploring data to generate solutions.
API	Application Programme Interface; set of programming commands, functions and protocols to interact with an external system. Useful for software developers, data scientists and others. A website may provide an API that allows straightforward access to specific information from the site.
Big data	a broad term for datasets that cannot reasonably be handled by traditional computers or tools, due to their size (volume) and high rate of update with additional data (velocity). Additionally, two further challenges typify big data, the diversity of different data types and sources (variety) which may need to be dealt with at the same time, and the quality of the data (veracity). Big data is also applied to the technologies and strategies to work with this type of data.
Bioinformatics	collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied in molecular genetics and genomics.

Bit	short for 'binary digit'; the smallest unit of measurement used in computer data, with a value of 0 or 1, abbreviated as 'b'.
Byte	a unit of measurement of data, abbreviated as 'B'. One byte contains eight bits, or a series of eight zeros and 1s. Each byte can be used to represent 256 different values.
BOINC	an open source software for volunteer computing; Berkeley Open Infrastructure for Network Computing https://boinc.berkeley.edu/
Cartography	the science and practice of making maps.
Climate	long-term weather patterns for a location or area, measured in averages, maxima and minima. Typically a minimum of 30 years of weather is considered to be the basis of a climate.
Cumulus	Puffy clouds with flat bases, Formed by convection.
Climate simulation	using computer models and quantitative methods to represent the atmosphere, oceans, land, ice and energy budget of the Earth.
Cloud computing	remote servers used to store, manage and process data rather than a local or desktop computer, often via the Internet.
Cluster computing	using multiple machines linked together and managing their collective capabilities to complete tasks. Computer clusters require a cluster management layer which handles communication between the individual nodes and coordinates work assignment.
Copernicus	Earth Observation programme of the European Space Agency, primarily using the Sentinel series of satellites, to improve the understanding of and management of the environment.
Core	a processor in a computer's central processing unit.
Cryosphere	the part of the Earth-system where water is frozen, including glaciers and sea-ice.
CSV	comma Separated Values; numbers and text stored as plain text. Each record is separated by commas.
Data driven	analysis and decision making led by the numbers, facts and statistical analysis, rather than intuition or experience.
Data mining	the practice of trying to find patterns in large sets of data. It is the process of trying to refine a mass of data into a more understandable and cohesive set of information.
Data point	one measurement, observation or element, a single member of a larger dataset.
Data quality	consistent, robust, validated and reliable information (see 'data challenges' video).

Big Data and the Environment

Data scientist	a person who has the knowledge and skills to conduct sophisticated and systematic analyses of data. A data scientist extracts insights from datasets for research or product development, and evaluates and identifies novel or strategic relationships or opportunities.
Data warehouse	large, ordered repositories of data that can be used for analysis and reporting. In contrast to a *data lake*, a data warehouse is composed of data that has been cleaned, integrated with other sources, and is generally well-ordered. Data warehouses are often spoken about in relation to big data, but typically are components of more conventional systems.
Dataset	a particular group of data, a defined collection of elements collected, stored, manipulated and/or analysed; most often numbers and dealt with as a group by a scientist or computer.
Demonstrator	a one-off system, often software, that shows whether or how data can be used for a specific purpose or task.
Disseminative Visualisation	data visualisation designed as a presentational aid for disseminating information or insight, with no purpose other than communication.
e-Infrastructure	a combination and interworking of digitally-based technology (hardware and software), resources (data, services, digital libraries), communications (protocols, access rights and networks), and the people and organisational structures needed to support modern, internationally leading collaborative research be it in the arts and humanities or the sciences. http://www.rcuk.ac.uk/research/xrcprogrammes/otherprogs/einfrastructure/
Earth Observation	gathering information about the Earth's physical systems via remote sensing technologies, often satellites which look down at the Earth from their orbit.
Electromagnetic Spectrum	the range of wavelengths of electromagnetic radiation, with gamma rays having short wavelengths and high energy, to radio waves with long wavelengths and low energy. Visible light is part of the electromagnetic spectrum. Examples of the use of electromagnetic radiation https://www.bbc.co.uk/education/guides/z66g87h/revision/3
ENIAC	Electronic Numerical Integrator And Computer, the world's first general-purpose computer; designed and built to calculate artillery firing tables in the 1940s and later used for computer weather predictions. https://www.thoughtco.com/history-of-the-eniac-computer-1991601
Ensemble	in weather forecasting an ensemble is a method whereby instead of making a single forecast, a set of forecasts are produced that present a range of future weather possibilities. https://www.ecmwf.int/en/about/media-centre/fact-sheet-ensemble-weather-forecasting
Environmental analytics	analysis of data sourced from the environment, or data with an application relating to the environment.

Environmental consultant	works on a contractual basis for private and public sector clients, addressing environmental issues such as water pollution, air quality and soil contamination. www.sokanu.com
Exa-	prefix denoting a factor of 10^{18} or a billion billion.
FAIR	the principles for scientific data management and stewardship – data should be Findable, Accessible, Interoperable (see below) and Reusable. Ideally, open data adheres to the FAIR principles. https://www.nature.com/articles/sdata201618
Flop	FLOating Point operation, a single calculation on a number with a decimal point i.e. not an integer. In computing, floating point operations per second (FLOPS) is a measure of computer performance, useful in fields of scientific computations that require floating-point calculations.
FTP	File Transfer Protocol, a standardised set of rules to allow upload and download of files between two computers, commonly used for exchanging files over the Internet.
Giga-	prefix denoting a factor of 10^{9} or a billion
Geostationary	a satellite that tracks the Earth's rotation above the equator therefore appearing to remain stationary, viewing the same portion of the Earth's surface, often used for TV or radio broadcasting and some meteorological satellites.
Informatics	the science of collecting, classifying, storing, retrieving and disseminating data and/or knowledge.
Infrared	in the electromagnetic spectrum, the visible light region lies from violet at shorter wavelengths/energies to red at longer wavelengths/energies. Infrared radiation has wavelengths just greater than red light and emitted particularly by heated objects. For example, night vision goggles use infrared radiation.
Integer	a number which is not a fraction; a whole number.
Internet of Things	abbreviated to IoT, a broad term for devices interconnected via the internet enabling sending and receiving of data or instructions. These devices include everyday items such as home appliances or cameras. Sometimes also referred to as 'smart devices'.
Interoperable	as part of the FAIR guidelines (see above) use a formal, accessible, shared, and broadly applicable language and ontology for knowledge and data representation.
JASMIN	petabyte-scale easily accessible storage collocated with data analysis computing facilities run by the Scientific and Technologies Facilities Council for researchers and science community in the UK. http://www.jasmin.ac.uk/
Kilo-	prefix denoting a factor of 10^{3} or a thousand

Machine learning	the study and practice of designing systems that can learn, adjust, and improve automatically, based on the data fed to them. This typically involves implementation of predictive and statistical algorithms that focus on "correct" behaviour and insights as data flows through the system.
MapReduce	a big data algorithm for scheduling work on a computing cluster. The process involves splitting the problem set up, mapping it to different nodes (map), and computing over them to produce intermediate results, shuffling the results to align like sets, and then reducing the results by outputting a single value for each set (reduce).
Mega-	prefix denoting a factor of 10^6 or a million
METAR	METeorological Aviation Report, a weather observation taken at a certain location, most likely an airfield, for use by pilots and weather forecasters. The METAR coding standard is agreed between civil aviation and weather authorities.
Metadata	a set of data that describes and gives information about other data. The purpose of metadata is to make finding, tracking and working with datasets easier for example by tagging with keywords.
Model	representation of a real world situation.
Natural Capital	can be defined as the world's stocks of natural assets which includes soil, water, air, flora and fauna.
Near-infrared	in the electromagnetic spectrum, near-infrared lies between red visible light and infrared. See also Infrared.
Nimbus	a programme of seven NASA missions of Earth Observation satellites, starting in 1964. Nimbus is Latin for rain cloud.
Noise	noise in data is meaningless data or unexplained variation in data which might be due to instrument errors, corruption or other issues. Noise disguises and/or distorts the underlying data which make it harder to analyse, just as noisy environments make it more difficult to hear the sound on which you wish to focus.
Open data	data that can anyone can access, use or share, often free of cost and subject only to attribution requirements (a longer basic definition by the Canadian government http://open.canada.ca/en/open-data-principles#toc94).
Open source	software for which the original source code is made freely available and may be redistributed and modified.
Outlier	a data point showing an unexpected relationship or large difference to the remainder of the dataset.
Peta-	prefix denoting a factor of 10^{15} or a million billion

Polar orbiting	a satellite orbit passing above or nearly above both poles on each orbit. Polar orbiting satellites have a lower altitude above the Earth's surface than geostationary satellites and therefore an increased resolution.
Proxy	in the case of data, other data that you may use and/or transform when you do not have a direct measurement of the data you require.
Provenance	in the case of data, the process of tracing and recording the origins of data and its movements between databases. Data's full history including how and why it got to its present palace.
RGB	in satellite imagery, the satellite's sensors operate in three channels, red, green and blue separately, and can be combined to give a colour image.
Repository	store datasets and provide access to users.
Satellite imagery	an image of part of the Earth taken using artificial satellites in orbit around the Earth. These images have a variety of uses including: mapping, military intelligence and meteorology. Satellites can carry a range of instruments each with a specific purpose, for example, visible light images are photographs of the Earth and only useful during daylight hours.
Sentinel satellites	a family of Earth Observation satellite missions by the European Space Agency http://m.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview4
Signal to noise ratio	a measure of how much useful information there is in a system, a phrase applied generally but originating in electrical systems to indicate the strength of the information (signal) compared to unwanted interference (noise), a low signal to noise ratio means that it is difficult to determine the useful information.
Smart Meter	a new kind of energy meter that can digitally send meter readings to your energy supplier and come with in home display units, to see in real-time how much energy is being used in a household.
Software developer	a person who researches, designs, programs and tests computer code.
Stream processing	the practice of computing over individual data items as they move through a system. This allows for real-time analysis of the data being fed to the system and is useful for time-sensitive operations using high velocity metrics.
Tera-	prefix denoting a factor of 10^{12} or a thousand billion (also a million million).
Urban heat island	a built-up area that is warmer than the surrounding rural areas due to human activities.
Version control	formally and methodically tracking changes to data, documents or programming code. Also known as revision control.

Visualisation	representing data visually to enable understanding of its significance; to highlight patterns and trends that might otherwise be missed; to communicate data quickly and in a meaningful way.
Weather	specific atmospheric conditions around us which can change minute-by-minute, day-to-day.
Weather forecast	a prediction of specific future weather conditions, such as daily maximum temperature at a location, up to several days ahead, with the estimate frequently becoming more uncertain with increasing lead-time. Weather forecasts are often based on computer simulations of the atmosphere known as NWP, Numerical Weather Prediction.
Weather observation	also known as a weather report, is a snapshot of the weather conditions at a certain location and at a certain time. An observation may be as basic as an air temperature reading but can include wind speed and direction, visibility, humidity, precipitation, cloud cover or soil surface temperature. A long-term average of a location's weather observations e.g. 30 years, determines the location's climate.
Yotta-	prefix denoting a factor of 10^{24} or a million billion billion and the largest unit prefix in the metric system.
Zetta-	prefix denoting a factor of 10^{21} or a thousand billion billion.