Glossary of terms for AI for Earth Monitoring

Analysis Ready Data are time-series stacks of overhead imagery that are prepared for a user to analyse without having to pre-process the imagery themselves. This may include the geospatial coregistration, resampling, atmospheric and radiometric corrections, cloud masking, etc.

Artificial Intelligence – a branch of computer science that deals with smart machines or robots that can think and perform tasks like a human. In Artificial Intelligence you can teach the computer to learn and make its own decisions.

AI is intelligence exhibited by machines that can observe, perceive and act upon their environment to maximise their performance over a variety of tasks. It refers to the capacity of an algorithm for assimilating information to perform tasks that are characteristic of human intelligence, such as recognising objects and sounds, contextualising language, learning from the environment, and problem solving. [PhiLab]

Artificial Narrow Intelligence – also known as Weak AI, this is artificial intelligence that is good at one or a few specific tasks. We interact with these on a day-to-day basis, for example: map directions, voice assistants or social media recommendations

Artificial General Intelligence – also known as Strong AI or Deep AI, this is artificial intelligence that mimics human intelligence or behaviours, with the ability to learn and apply its intelligence to solve any problem.

Algorithm – Set of rules, usually on a computer. AI, play a central role, they are what makes the computer seem intelligence. An AI algorithm takes data and creates new rules based on what it has learned on that data.

Bayesian Optimisation – AI tool to find the centre of mass. It is a sequential design strategy for global optimization of black-box functions that does not assume any functional forms. It is usually employed to optimize expensive-to-evaluate functions.

Big Data – A large amount of data which is structured or unstructured. Big data is hard to access with traditional methods – for example satellite images fill huge archives of data, and it is hard to access it, let alone work with it.

Big Data Analytics is a suite of analysis techniques aiming to deliver "value" from big datasets, whose Volume, Velocity, Variety, Veracity, and Value is beyond the ability of traditional tools to capture, store, manage and analyse. Within this report, the word big data analytics will mainly address ML. [PhiLab]

Cloud-free Compositing – an AI algorithm to composite multiple video frames segment the clouds and then removes them to create a cloud-free satellite image

Glossary of terms for AI for Earth Monitoring

Compressed sensing (also known as compressive sensing, compressive sampling, or sparse sampling) is a signal processing technique for efficiently acquiring and reconstructing a signal, by finding solutions to underdetermined linear systems. This is based on the principle that, through optimization, the sparsity of a signal can be exploited to recover it from far fewer samples than required by the Nyquist–Shannon sampling theorem. In a nutshell, a small number of measurements obtained by multiplying a sufficiently large signal by a well-defined matrix can be recovered with high probability. As a result, Compressed Sensing gained tremendous interest from research and industry as a powerful signal processing tool for images and audio compression.

Computer Vision – Computer vision creates algorithms that derive information from images. Computer vision algorithms see images as array of number and it then extracts patterns from these numbers: such as - basic shapes, distinct object or 3dD images. Machine Learning helps us to teach computers to do these tasks which help tasks such as creating 3D models, helping smartphones detect faces.

Concept Drift – this is when the relationship between input and output changes. For example, if you ride a bike it becomes harder through time as the pressure in the tyre reduces – so the relationship between the input (pedal power) and the output (speed) changes.

Copernicus In Situ Component:
The Copernicus Services rely on a combination of satellite data and environmental measurements, collected by data providers external to Copernicus, from ground-based, sea-borne or air-borne monitoring systems. This includes, for example, data from sensors placed on the banks of rivers, tall towers, carried on weather balloons or airplanes, pulled through the sea by ships, and drifting in the ocean on floats or buoys. These non-space data are collectively referred to as “in situ” data (named using the Latin for “in position”, “local” or “on site”).
In situ data also include a specific category of information known as geospatial reference data. This refers to background topographic information, such as transportation network maps, administrative boundaries and digital elevation models.

Crowdsourcing (CS) is the practice of public participation and collaboration in a common goal. Within this report, the term will also be used as a synonym for Citizen Science when the goal is to do research. Citizen scientists can help in processing / analysing EO data (e.g. visual interpretation of land cover and identification of other features visible from VHR images) but also in generating new observations (e.g. air quality measurements using a mobile phone (ispex.nl) or a variety of new mobile apps for recording observations on the ground) for a myriad of applications, ranging from land cover validation to animal tracking to humanitarian response. There are also many emerging synergies between AI and CS, in both directions, where AI can help analyse the data of citizens, and where citizens can train the AI through generation of data sets (e.g. labelled observations). [PhiLab]

Glossary of terms for AI for Earth Monitoring

Computer Vision (CV) is a field concerned with the automatic extraction, analysis, and understanding of useful information from a single image or sequence of images (e.g. videos). It involves the development of a theoretical and algorithmic basis to achieve automatic visual understanding. Over the last decade, CV has dramatically improved its capability enabling challenging tasks, such as identifying cars, buildings or changes, to be done by a machine in an automatic manner. Such potential of CV has not yet been fully harnessed in EO where specific problems are still to be explored (e.g., smaller objects, more spectral bands), thereby holding great promises for EO.

Data Cube The interpretation of the term data cube in the EO domain usually depends on the current context. It may refer to a data service such as Sentinel Hub, to some abstract API, or to a concrete set of spatial images that form a time-series.

Decision Tree - A decision tree is a useful machine learning algorithm used for both regression and classification tasks. The name “decision tree” comes from the fact that the algorithm keeps dividing the dataset down into smaller and smaller portions until the data has been divided into single instances, which are then classified. [https://www.unite.ai/what-is-a-decision-tree/]

Deep Learning is a type of ML algorithm that aims to solve the same kind of problems by mimicking the biological structure of the brain and construct hierarchical architectures of increasing sophistication. There is a wide variety of network architectures including Convolutional Neural Networks (CNN) (e.g., GoogleNet, Res-Net, YOLO), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Generative Adversarial Networks (GAN), Deep Belief Networks, and stacked auto-encoders. Today, DL is reaching high-level accuracy going beyond human performance, with the potential to substitute handcrafted feature extraction, thereby enabling totally automatic image recognition of big data (including EO) and opening huge opportunities for new science and business. [PhiLab]

Digital Twin is a virtual representation of a physical asset enabled through data and simulations for real-time prediction, monitoring, control and optimisation of the asset for improved decision-making throughout the life-cycle of the asset and be- yond [Rasdeed et al., 2019].

Digital Twin Earth (DTE) is an interactive “digital replica” of the entire planet that can facilitate a shared understanding of the multiple relationships between the physical and natural environments and society. DTE enables scientists and users to quantify past, present and future changes on our planet, by integrating data from models and observations and technologies such as AI to advance our understanding of the impact of human activities on our global environment and society. [PhiLab]

Glossary of terms for AI for Earth Monitoring

Dimensionality Reduction – the transformation of data from a high-dimensional space into a low-dimensional space, so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension [Wikipedia via Fabio]

Earth Observation (EO) is the gathering of information about our planet's physical, chemical and biological systems. Within this document, EO refers to the measurements specifically made from space via satellites. EO remote sensing data are used to monitor and assess the state and changes in our environment. EO satellites are providing unprecedented volumes of data and Artificial Intelligence (AI) techniques offer the potential to fully exploit this Big Data resource to extract relevant information for science and society. [PhiLab]

Explainability The wide-ranging concept of explainability is about making explanations on an algorithmic decision-making system available. The requirement for explainable AI addresses the fact that complex machines and algorithms often cannot provide insights into their behaviour and processes. This sometimes results in a black box effect, i.e. a situation where AI systems are capable of producing results, but the process by which the results are produced and the reasons why the algorithm makes specific decisions are not fully understandable by humans.
Explainability is therefore particularly important to ensure fairness in the use of algorithms and to identify potential bias in the training data. This far-reaching requirement means that an explanation should be available on how AI systems influence and shape the decision-making process, on how they are designed, and on what is the rationale for deploying them. Explainability must address both the technical processes of an AI system and the related human decisions taken in accordance with the EU guidelines. (EU guidelines on ethics in artificial intelligence: Context and implementation)

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

Environmental Models - Environmental modelling deals with representation of processes that occur in the real world in space and time. The processes that transform the environment through time are mostly described by dynamic models based on differential equations. The spatial interactions and topological rules are mostly managed by geographic information systems (GIS). [https://www.sciencedirect.com/topics/earth-and-planetary-sciences/environmental-modeling]

Favourability Functions - The favourability function assesses the variation in the probability of occurrence of an event in certain conditions with respect to the overall prevalence of the event. Consequently, it has potential to be applied in the cases where the probability of occurrence of an event is analysed, such as species distribution modelling or among others, habitat selection and epidemiological studies [https://link.springer.com/article/10.1007/s00114-012-0926-0]

Frame extraction – an algorithm to break the video down into individual images for 2D image processing

Generalised Linear Models - GLMs are a class of models that are appleid in cases where linear regression isn’t applicable or fail to make appropriate predictions. A GLM consists of three components: Random component (an exponential family of probability distributions), a systematic component (a linear predictor) and a link function (that generalises linear regression)
[https://towardsdatascience.com/linear-regression-or-generalized-linear-model-1636e29803d0]

Generative Adversarial Networks (GANS) - AI tool to modify VAE output.
GANs are a clever way of training a generative model by framing the problem as a supervised learning problem with two sub-models: the generator model that we train to generate new examples, and the discriminator model that tries to classify examples as either real (from the domain) or fake (generated). The two models are trained together in a zero-sum game, adversarial, until the discriminator model is fooled about half the time, meaning the generator model is generating plausible examples.

Genetic Algorithms for Rule Set Production - a genetic algorithm that creates models describing environmental conditions under which the species should be able to maintain populations. For input, GARP uses a set of point localities where the species is known to occur and a set of geographic layers representing the environmental parameters that might limit the species' capabilities to survive [https://old.dataone.org/software-tools/desktopgarp]

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

Labelled Data – Data used to train a ML application during supervised learning. It is an expensive process that involves human assign text to data, so the ML can learn to distinguish between them. This is like when you have to label features on a Captcha on the internet.

Heuristic – A rule drawn from experience, used to give a quick approximated solution for a hard to solve problem.

Hyperspectral data consists of many bands -- up to hundreds of bands -- that cover the electromagnetic spectrum. The NEON imaging spectrometer collects data within the 380nm to 2510nm portions of the electromagnetic spectrum within bands that are approximately 5nm in width.

Image Recognition Challenges - have played a key role in the AI renaissance, development of ML algorithms and breakthrough of DL schemes. They provide the benchmarking framework necessary to test / rank a suite of algorithms but train them through the availability of labelled data sets (being a pre-requisite to train ML algorithms). Here are a few examples.

Interference is the application of pre-trained Machine Learning algorithms on newly sensed and real-world data. The inference is the result of a trained neural network making predictions based on new data input.

A **land-parcel identification system (LPIS)** is a system to identify land use for a given country. It utilises orthophotos – basically aerial photographs and high precision satellite images that are digitally rendered to extract as much meaningful spatial information as possible. A unique number is given to each land parcel to provide a unique identification in space and time. This information is then updated regularly to monitor the evolution of the land cover and the management of the crops.

Machine Learning (ML) is a branch of AI relying on algorithms that are capable of learning from both data and through human interactions (e.g. supervision) to enable prediction, and are also used for data mining (i.e. discovery of unknown properties and patterns). ML is a field of statistical research for training computational algorithms that split, sort, transform a set of data in order to maximise the ability to classify, predict, cluster or discover new patterns in target datasets. ML is all about using computers to learn how to deal with problems without programming. In fact, ML generates models by taking some data for training a model, and then makes predictions. ML relies on a wide variety of algorithms (supervised and un-supervised), ranging from simple Symbolic Regression, Neural Network, decision tree, Support Vector Machine, up to genetic programming and ensemble methods such as random forest. [PhiLab]

Machine Learning Algorithms are programs that can learn from data and improve from experience, without human intervention.

Maximum Entropy Reinforcement Learning - The principle of maximum entropy is a model creation rule that requires selecting the most unpredictable (maximum entropy) prior assumption if only a single parameter is known about a probability distribution. The goal is to maximize “uniformitiveness,” or uncertainty when making a prior probability assumption so that subjective bias is minimized in the model’s results. [https://deepai.org/machine-learning-glossary-and-terms/principle-of-maximum-entropy]

Motion Tracking – an AI algorithm to track velocity and direction of vessels at sea

Glossary of terms for AI for Earth Monitoring

Multivariate Adaptive Regression Splines - s an algorithm for complex non-linear regression problems.
The algorithm involves finding a set of simple linear functions that in aggregate result in the best predictive performance. In this way, MARS is a type of ensemble of simple linear functions and can achieve good performance on challenging regression problems with many input variables and complex non-linear relationships.
[https://machinelearningmastery.com/multivariate-adaptive-regression-splines-mars-in-python/]

Neural networks are a set of algorithms modelling loosely the human brain connections, that are designed to recognise patterns. They interpret sensory data through a kind of machine perception, labelling or clustering raw input. There are a lot of variety of NN architectures including for example: [PhiLab]

Normalized difference vegetation index (NDVI) is a simple graphical indicator that can be used to analyze remote sensing measurements, often from a space platform, assessing whether or not the target being observed contains live green vegetation. NDVI quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs).

A **Convolutional Neural Network (ConvNet/CNN)** is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

A CNN is a class of Deep Learning networks (and one of the most influential innovations in the field of computer vision), most commonly applied to analysing visual imagery. A CNN takes the input (a tensor with a shape depending on number of images, width and height and number of bands), and passes it through a convolutional layer (instead of the more traditional matrix multiplications) abstracting into a feature map. A CNN is able to successfully capture the spatial and temporal dependencies in an image through the application of relevant filters - and are an ideal tool for processing regularly sampled data (such as is the case of satellite data). [PhiLab]

Logistic Regression - Logistic Regression is a type of classification algorithm, used when the value of the target variable is catrogrical in nature. It is most commonly used when the data in question has a binary output, so when it belongs to one class or another. For example: is an email spam or not, or is a mushroom poisonous or edible.
[https://kambria.io/blog/logistic-regression-for-machine-learning/]

Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems [https://christophm.github.io/interpretable-ml-book/logistic.html]

On-board processing – On-board payload data processing encompasses the data acquisition, analysis, transfer, storage, compression or reduction and transmission to ground of instrument and sensor data. Quite often the amount of raw data generated by modern instruments is in excess of what can be transmitted to ground. This makes it is necessary to use new technologies, for example based on AI, to reduce the amount of data. [ESA via Fabio]

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

Recurrent neural networks (RNN) is a class of neural networks where the connections between nodes from a direct- ed graph run along a temporal sequence, allowing to exhibit temporal dynamic behaviour. As such, RNNs can use their internal state (memory) to process variable length sequences of inputs, and is a powerful technique to enhance temporal information. [PhiLab]

Regression analysis consists of a set of **machine learning** methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x). Briefly, the goal of **regression model** is to build a mathematical equation that defines y as a function of the x variables.

Reinforcement learning is the training of machine learning models to make a sequence of decisions. It is an area of ML concerned with how software agents ought to take actions in an environment in order to maximise some notion of cumulative reward. e.g. learning from "mistakes" [PhiLab] The agent learns to achieve a goal in an uncertain, potentially complex environment. In reinforcement learning, an artificial intelligence faces a game-like situation. The computer employs trial and error to come up with a solution to the problem. To get the machine to do what the programmer wants, the artificial intelligence gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward.

Semi-Supervised Learning – This is a combination of supervised and unsupervised machine learning methods. When you don't have enough labelled data to produce an accurate model and you don't have the ability or resources to get more data, you can use semi-supervised techniques to increase the size of your training data. [https://www.datarobot.com/wiki/semi-supervised-machine-learning/] Unlabelled data, when used in conjunction with a small amount of labelled data, can produce considerable improvement in learning accuracy. [Wikipedia]

Glossary of terms for AI for Earth Monitoring

Sentinel is a set of Earth Observation missions from the Copernicus Programme that systematically acquires satellite imagery at high spatial resolution (10m to 60m) over land and coastal waters, developed specifically for the operational needs of the Copernicus programme. Each Sentinel mission is based on a constellation of two satellites to fulfil revisit and coverage requirements, providing robust datasets for Copernicus Services. These missions carry a range of technologies, such as radar and multi-spectral imaging instruments for land, ocean and atmospheric monitoring. The mission supports a broad range of services and applications such as agricultural monitoring, emergencies management, land cover classification or water quality.

Sentinel Hub is an engine for processing of petabytes of satellite data. It is opening the doors for machine learning and helping hundreds of application developers worldwide. It makes Sentinel, Landsat, and other Earth observation imagery easily accessible for browsing, visualization and analysis. Scale your system globally with an intuitive and user-friendly interface, without any hassle. It is a set of RESTful APIs, which make the Sentinel data seamlessly available to the user in analysis ready format. They are designed to be used in a synchronous fashion, in most cases retrieving results in less than one second - fast enough for real-time consumption and for intense compute processes such as machine learning. [**Sentinel Hub]**

Signal processing is an electrical engineering subfield that focuses on analysing, modifying, and synthesizing signals such as sound, images, and scientific measurements. Signal processing is key to embedded Machine Learning. Signal processing allows for the expansion of computing power and data storage capabilities.

Supervised Learning – A method of training an algorithm where a data set consists of two parts: The data itself (eg. an image of an apple) and a label (eg. the label apple). The algorithm is rewarded when it predicts the correct label, and can re-arrange itself if it predicts the wrong label. This produces a system that can highly accurately classify data.

Support Vector Machines - SVMs are a set of supervised learning methods used for classification, regression and outliers detection. [https://scikit-learn.org/stable/modules/svm.html] SVMs produce significant accuracy with less computation power. They can be used for both regression and classification tasks. But, it is widely used in classification objectives. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points. [https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47]

Synthetic-aperture radar (**SAR**) is a form of radar that is used to create two-dimensional images or three-dimensional reconstructions of objects, such as landscapes. SAR uses the motion of the radar antenna over a target region to provide finer spatial resolution than conventional beam-scanning radars.

Glossary of terms for AI for Earth Monitoring

SAR is a type of active data collection where a sensor produces its own energy and then records the amount of that energy reflected back after interacting with the Earth.

Training – This is the process of providing data to the computer so it can learn what we expect from it. For example, providing an computer with lots of images of different kind of dogs, and telling the computer that they are dogs – so in the future when it sees a picture of a new dog, it will be able to tell that it is a dog.

Transfer Learning is a ML method focused on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognise a type of feature that could apply when trying to recognise other similar or related features. Practically in this method a model developed for a task is reused as the starting point for a model on a second task. [PhiLab]

Transformative EO Technologies is also called radical, deep or disruptive technologies refer in this document to technologies that can make a big impact on the EO sector, by helping to address big societal and environmental challenges in a new way with EO data, by shaping the future of new EO “smart” and “connected” satellites, but also by having the power to create their own markets or disrupt existing industries. [PhiLab]

Unsupervised Learning – this is when no labels are given to a learning algorithm and it is left on its own devides to find structure in its input. Unsupervised learning can be a goal in itself (for discovering hidden patterns in data) or a means towards an end (such as feature learning) [https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a]

Variational Auto Encoders (VAEs) – AI tool to train the neural net. In a nutshell, a VAE is an autoencoder whose encodings distribution is regularised during the training in order to ensure that its latent space has good properties allowing us to generate some new data.

Vessel Detection – an AI algorithm to detect and count vessels in satellite video

Video Generation – an algorithm to take multiple 2D images and build it back up into a video, as well as transcoding videos into different formats (e.g. H.264, H.265, MPEG and Motion JPEG)

Video Stabilisation – an AI algorithm to stabilise video by aligning each video frame