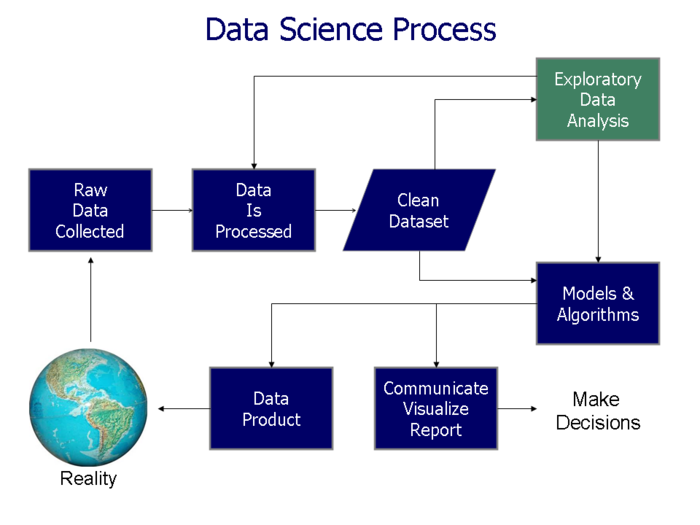
**Data science**, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes and systems to extract [knowledge](https://www.wikiwand.com/en/Knowledge) or insights from [data](https://www.wikiwand.com/en/Data) in various forms, either structured or unstructured,similar to [Knowledge Discovery in Databases](https://www.wikiwand.com/en/Knowledge_Discovery_in_Databases) (KDD).

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.

It employs techniques and theories drawn from many fields within the broad areas of [mathematics](https://www.wikiwand.com/en/Mathematics), [statistics](https://www.wikiwand.com/en/Statistics), [information science](https://www.wikiwand.com/en/Information_science), and [computer science](https://www.wikiwand.com/en/Computer_science), in particular from the subdomains of [machine learning](https://www.wikiwand.com/en/Machine_learning), [classification](https://www.wikiwand.com/en/Classification), [cluster analysis](https://www.wikiwand.com/en/Cluster_analysis), [data mining](https://www.wikiwand.com/en/Data_mining), [databases](https://www.wikiwand.com/en/Database), and [visualization](https://www.wikiwand.com/en/Data_visualization).



**Data science process flowchart**

## Data scientist

Data scientists use their data and analytical ability to find and interpret rich data sources; manage large amounts of data despite hardware, software, and bandwidth constraints; merge data sources; ensure consistency of datasets; create visualizations to aid in understanding data; build mathematical models using the data; and present and communicate the data insights/findings. They are often expected to produce answers in days rather than months, work by exploratory analysis and rapid iteration, and to produce and present results with dashboards (displays of current values) rather than papers/reports, as statisticians normally do.

"Data Scientist" has become a popular occupation with [Harvard Business Review](https://www.wikiwand.com/en/Harvard_Business_Review) dubbing it "The Sexiest Job of the 21st Century"  and [McKinsey & Company](https://www.wikiwand.com/en/McKinsey_%26_Company) projecting a global excess demand of 1.5 million new data scientists.[[24]](https://www.wikiwand.com/en/Data_science#citenote24) Universities are offering masters courses in data science.[[25]](https://www.wikiwand.com/en/Data_science#citenote25) Shorter private bootcamps are also offering data science certificates including student-paid programs like [General Assembly](https://www.wikiwand.com/en/General_Assembly_(school)) to employer-paid programs like [The Data Incubator](https://www.wikiwand.com/en/The_Data_Incubator).

Machine learning

Evolved from the study of [pattern recognition](https://www.wikiwand.com/en/Pattern_recognition) and [computational learning theory](https://www.wikiwand.com/en/Computational_learning_theory) in [artificial intelligence](https://www.wikiwand.com/en/Artificial_intelligence), machine learning explores the study and construction of [algorithms](https://www.wikiwand.com/en/Algorithm) that can learn from and make predictions on [data](https://www.wikiwand.com/en/Data)– such algorithms overcome following strictly static [program instructions](https://www.wikiwand.com/en/Computer_program) by making data-driven predictions or decisions, through building a [model](https://www.wikiwand.com/en/Mathematical_model) from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible; example applications include [email filtering](https://www.wikiwand.com/en/Email_filtering), detection of network intruders or malicious insiders working towards a [data breach](https://www.wikiwand.com/en/Data_breach), [optical character recognition](https://www.wikiwand.com/en/Optical_character_recognition) (OCR), [learning to rank](https://www.wikiwand.com/en/Learning_to_rank) and [computer vision](https://www.wikiwand.com/en/Computer_vision).

Machine learning is closely related to (and often overlaps with) [computational statistics](https://www.wikiwand.com/en/Computational_statistics), which also focuses on prediction-making through the use of computers. It has strong ties to [mathematical optimization](https://www.wikiwand.com/en/Mathematical_optimization), which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with [data mining](https://www.wikiwand.com/en/Data_mining), where the latter subfield focuses more on exploratory data analysis and is known as [unsupervised learning](https://www.wikiwand.com/en/Unsupervised_learning). Machine learning can also be unsupervised and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies.

Within the field of [data analytics](https://www.wikiwand.com/en/Data_analytics), machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as [predictive analytics](https://www.wikiwand.com/en/Predictive_analytics). These analytical models allow researchers, [data scientists](https://www.wikiwand.com/en/Data_science), engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

### Types of problems and tasks

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are

* [Supervised learning](https://www.wikiwand.com/en/Supervised_learning): The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that [maps](https://www.wikiwand.com/en/Map_(mathematics)) inputs to outputs.
* [Unsupervised learning](https://www.wikiwand.com/en/Unsupervised_learning): No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end ([feature learning](https://www.wikiwand.com/en/Feature_learning)).
* [Reinforcement learning](https://www.wikiwand.com/en/Reinforcement_learning): A computer program interacts with a dynamic environment in which it must perform a certain goal (such as [driving a vehicle](https://www.wikiwand.com/en/Autonomous_car) or playing a game against an opponent). The program is provided feedback in terms of rewards and punishments as it navigates its problem space.

Between supervised and unsupervised learning is [semi-supervised learning](https://www.wikiwand.com/en/Semi-supervised_learning), where the teacher gives an incomplete training signal: a training set with some (often many) of the target outputs missing. [Transduction](https://www.wikiwand.com/en/Transduction_(machine_learning)) is a special case of this principle where the entire set of problem instances is known at learning time, except that part of the targets are missing.

Among other categories of machine learning problems, [learning to learn](https://www.wikiwand.com/en/Meta_learning_(computer_science)) learns its own [inductive bias](https://www.wikiwand.com/en/Inductive_bias) based on previous experience. [Developmental learning](https://www.wikiwand.com/en/Developmental_robotics), elaborated for [robot learning](https://www.wikiwand.com/en/Robot_learning), generates its own sequences (also called curriculum) of learning situations to cumulatively acquire repertoires of novel skills through autonomous self-exploration and social interaction with human teachers and using guidance mechanisms such as active learning, maturation, motor synergies, and imitation.

Another categorization of machine learning tasks arises when one considers the desired *output* of a machine-learned system:

* In [classification](https://www.wikiwand.com/en/Statistical_classification), inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more ([multi-label classification](https://www.wikiwand.com/en/Multi-label_classification)) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam".
* In [regression](https://www.wikiwand.com/en/Regression_analysis), also a supervised problem, the outputs are continuous rather than discrete.
* In [clustering](https://www.wikiwand.com/en/Cluster_analysis), a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
* [Density estimation](https://www.wikiwand.com/en/Density_estimation) finds the [distribution](https://www.wikiwand.com/en/Probability_distribution) of inputs in some space.
* [Dimensionality reduction](https://www.wikiwand.com/en/Dimensionality_reduction) simplifies inputs by mapping them into a lower-dimensional space. [Topic modeling](https://www.wikiwand.com/en/Topic_modeling) is a related problem, where a program is given a list of [human language](https://www.wikiwand.com/en/Natural_language) documents and is tasked to find out which documents cover similar topics.

## Approaches

### Decision tree learning

|  |
| --- |
|  |

Decision tree learning uses a [decision tree](https://www.wikiwand.com/en/Decision_tree) as a [predictive model](https://www.wikiwand.com/en/Predictive_modelling), which maps observations about an item to conclusions about the item's target value.

### Association rule learning

|  |
| --- |
|  |

Association rule learning is a method for discovering interesting relations between variables in large databases.

### Artificial neural networks

|  |
| --- |
|  |

An [artificial neural network](https://www.wikiwand.com/en/Artificial_neural_network) (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is inspired by the structure and functional aspects of [biological neural networks](https://www.wikiwand.com/en/Biological_neural_networks). Computations are structured in terms of an interconnected group of [artificial neurons](https://www.wikiwand.com/en/Artificial_neuron), processing information using a [connectionist](https://www.wikiwand.com/en/Connectionism) approach to [computation](https://www.wikiwand.com/en/Computation). Modern neural networks are [non-linear](https://www.wikiwand.com/en/Non-linear) [statistical](https://www.wikiwand.com/en/Statistical) [data modeling](https://www.wikiwand.com/en/Data_modeling) tools. They are usually used to model complex relationships between inputs and outputs, to [find patterns](https://www.wikiwand.com/en/Pattern_recognition) in data, or to capture the statistical structure in an unknown [joint probability distribution](https://www.wikiwand.com/en/Joint_probability_distribution) between observed variables.

### Deep learning

|  |
| --- |
|  |

Falling hardware prices and the development of [GPUs](https://www.wikiwand.com/en/GPU) for personal use in the last few years have contributed to the development of the concept of [Deep learning](https://www.wikiwand.com/en/Deep_learning) which consists of multiple hidden layers in an artificial neural network. This approach tries to model the way the human brain processes light and sound into vision and hearing. Some successful applications of deep learning are [computer vision](https://www.wikiwand.com/en/Computer_vision) and [speech recognition](https://www.wikiwand.com/en/Speech_recognition).[[23]](https://www.wikiwand.com/en/Machine_learning#citenote23)

### Inductive logic programming

|  |
| --- |
|  |

Inductive logic programming (ILP) is an approach to rule learning using [logic programming](https://www.wikiwand.com/en/Logic_programming) as a uniform representation for input examples, background knowledge, and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system will derive a hypothesized logic program that [entails](https://www.wikiwand.com/en/Entailment) all positive and no negative examples. [Inductive programming](https://www.wikiwand.com/en/Inductive_programming) is a related field that considers any kind of programming languages for representing hypotheses (and not only logic programming), such as [functional programs](https://www.wikiwand.com/en/Functional_programming).

### Support vector machines

|  |
| --- |
|  |

Support vector machines (SVMs) are a set of related [supervised learning](https://www.wikiwand.com/en/Supervised_learning) methods used for [classification](https://www.wikiwand.com/en/Statistical_classification) and [regression](https://www.wikiwand.com/en/Regression_analysis). Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.

### Clustering

|  |
| --- |
|  |

Cluster analysis is the assignment of a set of observations into subsets (called *clusters*) so that observations within the same cluster are similar according to some predesignated criterion or criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some *similarity metric* and evaluated for example by *internal compactness* (similarity between members of the same cluster) and *separation* between different clusters. Other methods are based on *estimated density* and *graph connectivity*. Clustering is a method of [unsupervised learning](https://www.wikiwand.com/en/Unsupervised_learning), and a common technique for [statistical](https://www.wikiwand.com/en/Statistics) [data analysis](https://www.wikiwand.com/en/Data_analysis).

### Bayesian networks

|  |
| --- |
|  |

A Bayesian network, belief network or directed acyclic graphical model is a [probabilistic graphical model](https://www.wikiwand.com/en/Graphical_model) that represents a set of [random variables](https://www.wikiwand.com/en/Random_variables) and their [conditional independencies](https://www.wikiwand.com/en/Conditional_independence) via a [directed acyclic graph](https://www.wikiwand.com/en/Directed_acyclic_graph) (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Efficient algorithms exist that perform [inference](https://www.wikiwand.com/en/Inference) and learning.

### Reinforcement learning

|  |
| --- |
|  |

Reinforcement learning is concerned with how an *agent* ought to take *actions* in an *environment* so as to maximize some notion of long-term *reward*. Reinforcement learning algorithms attempt to find a *policy* that maps *states* of the world to the actions the agent ought to take in those states. Reinforcement learning differs from the [supervised learning](https://www.wikiwand.com/en/Supervised_learning) problem in that correct input/output pairs are never presented, nor sub-optimal actions explicitly corrected.

### Representation learning

|  |
| --- |
|  |

Several learning algorithms, mostly [unsupervised learning](https://www.wikiwand.com/en/Unsupervised_learning) algorithms, aim at discovering better representations of the inputs provided during training. Classical examples include [principal components analysis](https://www.wikiwand.com/en/Principal_components_analysis) and [cluster analysis](https://www.wikiwand.com/en/Cluster_analysis). Representation learning algorithms often attempt to preserve the information in their input but transform it in a way that makes it useful, often as a pre-processing step before performing classification or predictions, allowing reconstruction of the inputs coming from the unknown data generating distribution, while not being necessarily faithful for configurations that are implausible under that distribution.

[Manifold learning](https://www.wikiwand.com/en/Manifold_learning) algorithms attempt to do so under the constraint that the learned representation is low-dimensional. [Sparse coding](https://www.wikiwand.com/en/Sparse_coding) algorithms attempt to do so under the constraint that the learned representation is sparse (has many zeros). [Multilinear subspace learning](https://www.wikiwand.com/en/Multilinear_subspace_learning) algorithms aim to learn low-dimensional representations directly from [tensor](https://www.wikiwand.com/en/Tensor) representations for multidimensional data, without reshaping them into (high-dimensional) vectors. [Deep learning](https://www.wikiwand.com/en/Deep_learning) algorithms discover multiple levels of representation, or a hierarchy of features, with higher-level, more abstract features defined in terms of (or generating) lower-level features. It has been argued that an intelligent machine is one that learns a representation that disentangles the underlying factors of variation that explain the observed data.

### Similarity and metric learning

|  |
| --- |
|  |

In this problem, the learning machine is given pairs of examples that are considered similar and pairs of less similar objects. It then needs to learn a similarity function (or a distance metric function) that can predict if new objects are similar. It is sometimes used in [Recommendation systems](https://www.wikiwand.com/en/Recommendation_systems).

### Sparse dictionary learning

In this method, a datum is represented as a linear combination of [basis functions](https://www.wikiwand.com/en/Basis_function), and the coefficients are assumed to be sparse. Let *x* be a *d*-dimensional datum, *D* be a *d* by *n* matrix, where each column of *D* represents a basis function. *r* is the coefficient to represent *x* using *D*. Mathematically, sparse dictionary learning means solving  where *r* is sparse. Generally speaking, *n* is assumed to be larger than *d* to allow the freedom for a sparse representation.

Learning a dictionary along with sparse representations is [strongly NP-hard](https://www.wikiwand.com/en/Strongly_NP-hard) and also difficult to solve approximately.A popular heuristic method for sparse dictionary learning is [K-SVD](https://www.wikiwand.com/en/K-SVD).

Sparse dictionary learning has been applied in several contexts. In classification, the problem is to determine which classes a previously unseen datum belongs to. Suppose a dictionary for each class has already been built. Then a new datum is associated with the class such that it's best sparsely represented by the corresponding dictionary. Sparse dictionary learning has also been applied in [image de-noising](https://www.wikiwand.com/en/Image_de-noising). The key idea is that a clean image patch can be sparsely represented by an image dictionary, but the noise cannot.

### Genetic algorithms

|  |
| --- |
|  |

A genetic algorithm (GA) is a [search](https://www.wikiwand.com/en/Search_algorithm) [heuristic](https://www.wikiwand.com/en/Heuristic_(computer_science)) that mimics the process of [natural selection](https://www.wikiwand.com/en/Natural_selection), and uses methods such as [mutation](https://www.wikiwand.com/en/Mutation_(genetic_algorithm)) and [crossover](https://www.wikiwand.com/en/Crossover_(genetic_algorithm)) to generate new [genotype](https://www.wikiwand.com/en/Chromosome_(genetic_algorithm)) in the hope of finding good solutions to a given problem. In machine learning, genetic algorithms found some uses in the 1980s and 1990s.Vice versa, machine learning techniques have been used to improve the performance of genetic and [evolutionary algorithms](https://www.wikiwand.com/en/Evolutionary_algorithm).

### Rule-based machine learning

[Rule-based machine learning](https://www.wikiwand.com/en/Rule-based_machine_learning) is a general term for any machine learning method that identifies, learns, or evolves `rules’ to store, manipulate or apply, knowledge. The defining characteristic of a rule-based machine learner is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. This is in contrast to other machine learners that commonly identify a singular model that can be universally applied to any instance in order to make a prediction. Rule-based machine learning approaches include [learning classifier systems](https://www.wikiwand.com/en/Learning_classifier_system), [association rule learning](https://www.wikiwand.com/en/Association_rule_learning), and [artificial immune systems](https://www.wikiwand.com/en/Artificial_immune_system).

### Learning classifier systems

|  |
| --- |
|  |

Learning classifier systems (LCS) are a family of [rule-based machine learning](https://www.wikiwand.com/en/Rule-based_machine_learning) algorithms that combine a discovery component (e.g. typically a [genetic algorithm](https://www.wikiwand.com/en/Genetic_algorithm)) with a learning component (performing either [supervised learning](https://www.wikiwand.com/en/Supervised_learning), [reinforcement learning](https://www.wikiwand.com/en/Reinforcement_learning), or [unsupervised learning](https://www.wikiwand.com/en/Unsupervised_learning)). They seek to identify a set of context-dependent rules that collectively store and apply knowledge in a [piecewise](https://www.wikiwand.com/en/Piecewise) manner in order to make predictions.

## Applications

Applications for machine learning include:

* [Adaptive websites](https://www.wikiwand.com/en/Adaptive_website)
* [Affective computing](https://www.wikiwand.com/en/Affective_computing)
* [Bioinformatics](https://www.wikiwand.com/en/Bioinformatics)
* [Brain-machine interfaces](https://www.wikiwand.com/en/Brain-machine_interfaces)
* [Cheminformatics](https://www.wikiwand.com/en/Cheminformatics)
* Classifying [DNA sequences](https://www.wikiwand.com/en/DNA_sequence)
* [Computational anatomy](https://www.wikiwand.com/en/Computational_anatomy)
* [Computer vision](https://www.wikiwand.com/en/Computer_vision), including [object recognition](https://www.wikiwand.com/en/Object_recognition)
* Detecting [credit card fraud](https://www.wikiwand.com/en/Credit_card_fraud)
* [Game playing](https://www.wikiwand.com/en/Strategy_game)
* [Information retrieval](https://www.wikiwand.com/en/Information_retrieval)
* [Internet fraud](https://www.wikiwand.com/en/Internet_fraud) detection
* [Marketing](https://www.wikiwand.com/en/Marketing)
* [Machine perception](https://www.wikiwand.com/en/Machine_perception)
* [Medical diagnosis](https://www.wikiwand.com/en/Diagnosis_(artificial_intelligence))
* [Economics](https://www.wikiwand.com/en/Economics)
* [Natural language processing](https://www.wikiwand.com/en/Natural_language_processing)
* [Natural language understanding](https://www.wikiwand.com/en/Natural_language_understanding)
* [Optimization](https://www.wikiwand.com/en/Mathematical_optimization) and [metaheuristic](https://www.wikiwand.com/en/Metaheuristic)
* [Online advertising](https://www.wikiwand.com/en/Online_advertising)
* [Recommender systems](https://www.wikiwand.com/en/Recommender_system)
* [Robot locomotion](https://www.wikiwand.com/en/Robot_locomotion)
* [Search engines](https://www.wikiwand.com/en/Search_engines)
* [Sentiment analysis](https://www.wikiwand.com/en/Sentiment_analysis) (or opinion mining)
* [Sequence mining](https://www.wikiwand.com/en/Sequence_mining)
* [Software engineering](https://www.wikiwand.com/en/Software_engineering)
* [Speech](https://www.wikiwand.com/en/Speech_recognition) and [handwriting recognition](https://www.wikiwand.com/en/Handwriting_recognition)
* [Financial market](https://www.wikiwand.com/en/Financial_market) analysis
* [Structural health monitoring](https://www.wikiwand.com/en/Structural_health_monitoring)
* [Syntactic pattern recognition](https://www.wikiwand.com/en/Syntactic_pattern_recognition)
* [User behavior analytics](https://www.wikiwand.com/en/User_behavior_analytics)
* [Translation](https://www.wikiwand.com/en/Translation)