PRAGMATIC
INSTITUTE

DATA SCIENCE GLOSSARY

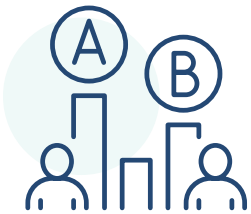
INDUSTRY TERMS

Pragmatic Institute®

DATA SCIENCE **INDUSTRY TERMS**

A

A/B testing



A statistical way of comparing two (or more) techniques, typically an incumbent against a new rival. A/B testing aims to determine not only which technique performs better but also to understand whether the difference is statistically significant. A/B testing usually considers only two techniques using one measurement, but it can be applied to any finite number of techniques and measures.

accuracy

In classification, accuracy is defined as the number of observations that are correctly labeled by the algorithm as a fraction of the total number of observations the algorithm attempted to label. Colloquially, it is the fraction of times the algorithm guessed "right."

anomaly detection

Anomaly detection, also known as outlier detection, is the identification of rare items, events, observations, or patterns which raise suspicions by differing significantly from the majority of the data.

artificial intelligence (AI)

The ability to have machines act with apparent intelligence, although varying definitions of "intelligence" lead to a range of meanings for the artificial variety. In AI's early days in the 1950s, researchers sought general principles of intelligence to implement, often using symbolic logic to automate reasoning. As the cost of computing resources dropped, the focus moved more toward statistical analysis of large amounts of data to drive decision making that gives the appearance of intelligence. See also machine learning, data mining, and expert systems.

B

backtesting

Periodic evaluation of a trained machine learning algorithm to check whether the predictions of the algorithm have degraded over time. Backtesting is a critical component of model maintenance.

baseline

A model or heuristic used as reference point for comparing how well a machine learning model is performing. A baseline helps model developers quantify the minimal, expected performance on a particular problem. Generally, baselines are set to simulate the performance of a model that doesn't actually make use of our data to make predictions. This is called a naive benchmark.

batch

A set of observations that are fed into a machine learning model to train it. Batch training is a counterpart to online learning, in which data are fed sequentially instead of all at once.

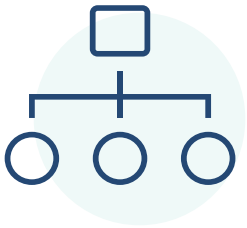
bias

Bias is a source of error that emerges from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the

relevant relations between features and labels. Bias can be mitigated by adding additional features to the data or using a more flexible algorithm. See also variance, cross-validation.

C

classification

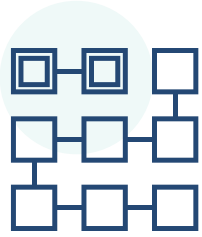


Classification is one of the two major types of supervised learning models in which the labels we train the algorithm to predict are distinct categories. Usually these categories are binary (yes/no, innocent/guilty, 0/1) but classification algorithms can typically be extended to handle multiple classes (peach, plum, pear) or, in a more limited set of cases, multiple labels (an object can belong to more than one category). See also regression, supervised learning

cloud computing

A computing paradigm in which the storage and processing of data or the hosting of computing services such as databases or websites takes place on a remote system comprised of multiple individual computing units acting as one and typically owned by a cloud computing service provider.

clustering



An unsupervised learning technique that identifies group structures in data. Clusters are, loosely speaking, groups of observations that are similar to other observations in the same cluster and different from those belonging to different clusters. The center of each cluster is known by the excellent name "centroid." Importantly, clustering algorithms only consider the relationships between features in the data mathematically and not conceptually; as such, the clusters identified by these algorithms may not reflect any grouping structure that would be sensible to a human being. See also classification, supervised learning, unsupervised learning, k-means clustering.

cross-validation

The name given to a set of techniques that split data into training sets and test sets when using data with an algorithm. The training set is given to the algorithm, along with the correct answers (labels), and becomes the set used to make predictions. The algorithm is then asked to make predictions for each item in the test set. The answers it gives are compared to the correct answers, and an overall score for how well the algorithm did is calculated. Cross-validation repeats this splitting procedure several times and computes an average score based on the scores from each split.

D

data cleansing

The act of reviewing and revising data to remove duplicate entries, correct misspellings, add missing data and provide more consistency.

data collection

Any process that captures any type of data.

data custodian

A person responsible for the database structure and the technical environment, including the storage of data.

data dictionary

A set of information describing the contents, format, and structure of a database and the relationship between its elements, used to control access to and manipulation of the database.

data-directed decision making

The use of data to support making crucial decisions.

data engineer

A specialist in data wrangling.

data exhaust

The data that a person creates as a byproduct of a common activity—for example, a cell call log or web search history.

data feed

A means for a person to receive a stream of data. Examples of data feed mechanisms include RSS or Twitter.

data integrity

The measure of trust an organization has in the accuracy, completeness, timeliness and validity of the data.

data mart

The access layer of a data warehouse used to provide data to users.

data migration

The process of moving data between different storage types or formats, or between different computer systems.

data mining

The process of deriving patterns or knowledge from large data sets.

data model, data modeling

An agreed upon data structure. This structure is used to pass data from one individual, group, or organization to another, so that all parties know what the different data components mean. Often meant for both technical and non-technical users.

data profiling

The process of collecting statistics and information about data in an existing source.

data quality

The measure of data to determine its worthiness for decision making, planning or operations.

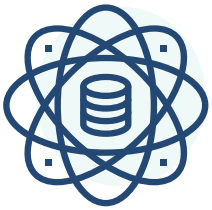
data replication

The process of sharing information to ensure consistency between redundant sources.

data repository

The location of permanently stored data.

data science



The discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning and database engineering to solve complex problems.

data scientist

A practitioner of data science.

data security

The practice of protecting data from destruction or unauthorized access.

data steward

A person responsible for data stored in a data field.

data structure

A specific way of storing and organizing data.

data visualization

A visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively.

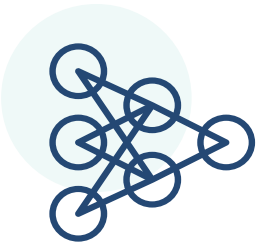
data warehouse

A place to store data for the purpose of reporting and analysis.

data wrangling

The process of transforming and cleaning data from raw formats to appropriate formats for later use. Also called data munging.

deep learning



A multilevel algorithm that gradually identifies things at higher levels of abstraction. For example, the first level identifies certain lines. The second identifies combinations of lines as shapes. Then the third identifies combinations of shapes as specific objects. Deep learning is popular for image classification. See also neural network.

data munging

See data wrangling.

E

errors-at-random

Errors-at-random are data errors such as missing or mismeasured data that are random with respect to the data we observe. Errors are not-at-random if the probability that an observation is missing or erroneous is correlated with the observed data. Errors-not-at-random are especially problematic if errors are correlated with labels.

ETL

ETL is short for extract, transform, load, three database functions that are combined into one tool to pull data from a primary source and place it into a database.

expert system

An expert system is a computer system that emulates the decision-making ability of a human expert. Expert systems are designed to solve complex problems processing data describing the context of the decision being made and applying logic, mainly in the form of if-then rules.

H

Hadoop

Hadoop is a collection of software that facilitate using a network of many computers to solve problems involving large amounts of data and computation. It consists of two main functional components. One, the Hadoop Distributed File System (HDFS), is a utility that allows data to be stored over multiple networked machines in a failure-tolerant manner while still being treated as a single file from the perspective of the user. The other, Hadoop MapReduce, is a programming paradigm that allows the user to process and analyze this data in parallel over large numbers of individual processing units located across multiple machines.

HiPPO

Highest Paid Person's Opinion. A paradigm for decision-making within businesses that is inconsistent with data-driven cultures.

Hive

Hive is a data warehouse software project built on top of Hadoop for providing data query and analysis. Hive gives a SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

I

Internet of Things (IoT)

The Internet of things (IoT) is the extension of internet connectivity into physical devices and everyday objects. Embedded with electronics, internet connectivity, artificial intelligence, and other forms of hardware, these devices can communicate

and interact with others over the Internet, and they can be remotely monitored and controlled

J

Java

Java is a general-purpose, object-oriented, compiled programming language. While it is not among the most common languages used by data scientists, it and its close relative Scala are the native language of many distributed computing frameworks such as Hadoop and Spark.

L

label

In supervised learning applications, labels are the components of the data that indicate the desired predictions or decisions we would like the machine learning algorithm to make for each observation we pass into the algorithm. Supervised learning algorithms learn to use other features in the data to predict labels so that these algorithms can learn to predict labels in other instances when the labels are not known or determined. In certain fields, labels are called targets. See also supervised learning, classification, regression.

leakage

Leakage is the introduction of information during training that will not be germane or available to the deployed algorithm.

length

Length measures the number of observations in our dataset.

linear regression

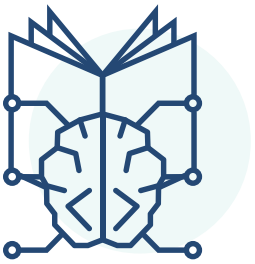
A technique to look for a linear relationship by starting with a set of data points that don't necessarily line up nicely. This is accomplished by computing the "least squares" line: on an x-y graph, the line that has the smallest possible sum of squared distances to the actual data point y values. Statistical software packages and typical spreadsheet packages offer automated ways to calculate this.

linear relationship

The relationship between two varying amounts, such as price and sales, that can be expressed with an equation that can be represented as a straight line on a graph.

M

machine learning

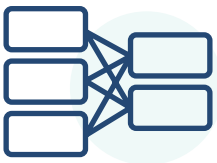


The use of data-driven algorithms that perform better as they have more data to work with, redefining their models or "learning" from this additional data. This involves cross-validation with training and test data sets. Studying the practical application of machine learning usually means researching which machine learning algorithms are best for which situations.

machine learning model

The model artifact that is created in the process of providing a machine learning algorithm with training data from which to learn.

MapReduce



MapReduce is a programming model and implementation designed to work with big data sets in parallel on a distributed cluster system. MapReduce programs consist of two steps. First, a map step takes chunks of data and processes it in some way (e.g. parsing text into words). Second, a reduce step takes the data that are generated by the map step and performs some kind of summary calculation

(e.g. counting word occurrences). In between the map and reduce step, data move between machines using a key-value pair system that guarantees that each reducer has the information it needs to complete its calculation (e.g. all of the occurrences of the word "Python" get routed to a single processor so they can be counted in aggregate).

minimum viable product (MVP)

The minimum viable product is the smallest complete unit of work that would be valuable in its own right, even if the rest of the project fizzled out.

model

The specification of mathematical or probabilistic relationships existing between different variables. Because "modeling" can mean many things, the term "statistical modeling" is often used to more accurately describe the kind of modeling that data scientists do.

N

natural language processing (NLP)

Natural Language Processing (NLP) is a branch of data science that applies machine learning techniques to help machines learn to interpret and process textual data consisting of human language. Applications of NLP include text classification (predicting what type of content a document contains), sentiment analysis (determining whether a statement is positive, negative, or neutral), and translation. NLP also comprises techniques to encode textual content numerically to use in machine learning applications.

Naive Bayes

A classification algorithm that predicts labels from data by assuming that the features of the data are statistically independent from each other. Due to this assumption, Naive Bayes models can be easily fit on distributed systems.

neural network

A machine learning method modeled after the brain. This method is extremely powerful and flexible, as it is created from an arbitrary number of artificial neurons that can be connected in various patterns appropriate to the problem at hand, and the strength of those connections are adjusted during the training process. They are able to learn extremely complex relationships between data and output, at the cost of large computational needs. They have been used to great success in processing image, movie, and text data, and any situation with very large numbers of features.

non-stationarity

Non-stationarity occurs when the mapping between the features of our data and the label we're trying to predict changes from the time our model was trained. Housing prices, for example, are non-stationary: a model fit in the 1930s would make exceptionally poor predictions today, as houses cost a lot less back then. Models fit on non-stationary data must be backtested and adjusted frequently to keep them relevant.

NoSQL

A database management system that uses any of several alternatives to the relational, table-oriented model used by SQL databases. Originally meant as "not SQL," it has come to mean something closer to "not only SQL" due to the specialized nature of NoSQL database management systems. These systems often are tasked with playing specific roles in a larger system that may also include SQL and additional NoSQL systems.

O

online learning

Online learning is a learning paradigm by which machine learning models may be trained by passing

them training data sequentially or in small groups (mini-batches). This is important in instances where the amount of data on hand exceeds the capacity of the RAM of the system on which a model is being developed. Online learning also allows models to be continually updated as new data are produced.

overfitting

See Variance

P

Perl

An older scripting language with roots in pre-Linux UNIX systems. Perl has always been popular for text processing, especially data cleanup and enhancement tasks.

Pig

Apache Pig is a high-level platform for creating programs that run on Hadoop. Pig is designed to make it easier to create data processing and analysis workflows that can be executed in MapReduce, Spark, or other distributed frameworks.

precision

A performance measure for classification models. Precision measures the fraction of all of the observations that a classification algorithm flagged positively that were flagged correctly. For example, if our algorithm were judging suspects, precision would measure the percentage of all the suspects declared guilty by the algorithm who actually were guilty. See also recall.

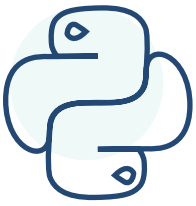
predictive analytics

The analysis of data to predict future events, typically to aid in business planning. This incorporates predictive modeling and other techniques. Machine learning might be considered a set of algorithms to help implement predictive analytics.

predictive modeling

The development of statistical models to predict future events.

Python



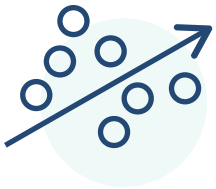
A programming language available since 1994 that is popular with people doing data science. Python is noted for ease of use among beginners and great power when used by advanced users, especially when taking advantage of specialized libraries such as those designed for machine learning and graph generation.

R

R

An open-source programming language and environment for statistical computing and graph generation available for Linux, Windows and Mac. Along with Python, R is among the most popular software packages used by data scientists.

regression



Regression is one of the two major types of supervised learning models in which the labels we train the algorithm to predict are ordered quantities like prices or numerical amounts. One might use a regression, for instance, to predict temperatures over time or housing prices within a city. See also classification, supervised learning

recall

A performance measure for classification models. Recall measures the fraction of all of the observations that a classification algorithm should have flagged positively that were actually flagged by the algorithm. For example, if our algorithm were judging suspects, recall would measure the

percentage of all guilty suspects that the algorithm correctly identified as such. See also precision.

Ruby

A scripting language that first appeared in 1996. Ruby is popular in the data science community, but not as popular as Python, which has more specialized libraries available for data science tasks.

S

SAS

A commercial statistical software suite that includes a programming language also known as SAS.

Scala

Scala is a Java-like programming language commonly used by data scientists. It is the native language of Spark.

Scikit-Learn

The most common Python package for machine learning

shell

A computer's operating system when used from the command line. Along with scripting languages such as Perl and Python, Linux-based shell tools (included and available for Mac and Windows computers) such as grep, diff, split, comm, head and tail are popular for data wrangling. A series of shell commands stored in a file that lets you execute the series by entering the file's name is known as a shell script.

Simpson's paradox

Simpson's paradox is a phenomenon in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

Spark

Apache Spark is a high-level open-source distributed cloud computing framework. Spark is particularly valuable because it contains libraries that support the querying of distributed databases, distributed processing and wrangling, and distributed machine learning. As such, it provides end-to-end solutions that allow data scientists to take full advantage of cloud computing resources.

SPSS

A commercial statistical software package, or predictive analytics software, popular in the social sciences. It has been available since 1968 and was acquired by IBM in 2009.

SQL

Stands for "Structured Query Language." The ISO standard query language for relational databases. This language is used to ask structured databases for information out of one or more data tables stored in the database. Variations of this extremely popular language are often available for data storage systems that aren't strictly relational. Watch for the phrase "SQL-like."

Stata

A commercial statistical software package commonly used by academics, particularly in the social sciences.

supervised learning

A type of machine learning algorithm in which a system learns to predicts labels after being shown a set of training data and identifying statistical associations between features in the data and the labels it is given. The classic example is sorting email into spam versus ham. See also unsupervised learning, machine learning.

T

Tableau

A commercial data visualization package often used in data science projects.

U

unsupervised learning

A class of machine learning algorithms designed to identify (potentially) useful patterns or structures in data without being directed to perform a specific prediction or decision task.

V

variance

Variance is the amount that the estimate of the target function will change if different training data was used. Another way of saying this is that variance measures the degree to which a model picks up noise as opposed to signal. High-variance is synonymous with overfitting.

W

width

Width measures the number of features in a dataset.

PRAGMATIC
INSTITUTE

MARKET-DRIVEN. DATA-LED.

Pragmatic Institute is a trusted learning partner to professionals across data, product, and design—providing training, support and resources. With a focus on dynamic instruction, continued learning, and what works for today's businesses, Pragmatic delivers engaging and impactful education to thousands of companies worldwide.

PragmaticInstitute.com