Data Science Terms and Jargon: A Glossary

Source link : https://www.dataquest.io/blog/data-science-glossary/

Getting started in data science can be overwhelming, especially when you consider the variety of concepts and techniques a data scienctist needs to master in order to do her job effectively. Even the term “data science” can be somewhat nebulous, and as the field gains popularity it seems to lose definition. To help those new to the field stay on top of industry jargon and terminology, we’ve put together this glossary of data science terms. We hope it will serve as your handy quick reference whenever you’re working on a project, or reading an article and find you can’t quite remember what “ETL” means. *Are we missing a term? Get in touch.*

Fundamentals

These are some baseline concepts that are helpful to grasp when getting started in data science. While you probably won’t have to work with every concept mentioned here, knowing what the terms mean will help when reading articles or discussing topics with fellow data lovers.

Algorithms

An algorithm is a set of instructions we give a computer so it can take values and manipulate them into a usable form. This can be as easy as finding and removing every comma in a paragraph, or as complex as building an equation that predicts how many home runs a baseball player will hit in 2018.

Back End

The back end is all of the code and technology that works behind the scenes to populate the front end with useful information. This includes databases, servers, authentication procedures, and much more. You can think of the back end as the frame, the plumbing, and the wiring of an apartment.

Big Data

Big data is a term that suffers from being too broad to be useful. It's more helpful to read it as, "so much data that you need to take careful steps to avoid week-long script runtimes." Big data is more about strategies and tools that help computers do complex analysis of very large (read: 1+ TB) data sets. The problems we must address with big data are categorized by the 4 V's: volume, variety, veracity, and velocity.

Classification

Classification is a supervised machine learning problem. It deals with categorizing a data point based on its similarity to other data points. You take a set of data where every item already has a category and look at common traits between each item. You then use those common traits as a guide for what category the new item might have.

Database

As simply as possible, this is a storage space for data. We mostly use databases with a Database Management System (DBMS), like PostgreSQL or MySQL. These are computer applications that allow us to interact with a database to collect and analyze the information inside.

Data Warehouse

A data warehouse is a system used to do quick analysis of business trends using data from many sources. They're designed to make it easy for people to answer important statistical questions without a Ph.D. in database architecture.

Front End

The front end is everything a client or user gets to see and interact with directly. This includes data dashboards, web pages, and forms.

Fuzzy Algorithms

Algorithms that use fuzzy logic to decrease the runtime of a script. Fuzzy algorithms tend to be less precise than those that use Boolean logic. They also tend to be faster, and computational speed sometimes outweighs the loss in precision.

Fuzzy Logic

An abstraction of Boolean logic that substitutes the usual True and False and for a range of values between 0 and 1. That is, fuzzy logic allows statements like “a little true” or “mostly false.”

Greedy Algorithms

A greedy algorithm will break a problem down into a series of steps. It will then look for the best possible solution at each step, aiming to find the best overall solution available. A good example is Dijkstra’s algorithm, which looks for the shortest possible path in a graph.

Machine Learning

A process where a computer uses an algorithm to gain understanding about a set of data, then makes predictions based on its understanding. There are many types of machine learning techniques; most are classified as either supervised or unsupervised techniques.

Overfitting

Overfitting happens when a model considers too much information. It's like asking a person to read a sentence while looking at a page through a microscope. The patterns that enable understanding get lost in the noise.

Regression

Regression is another supervised machine learning problem. It focuses on how a target value changes as other values within a data set change. Regression problems generally deal with continuous variables, like how square footage and location affect the price of a house.

Statistic vs. Statistics

Statistics (plural) is the entire set of tools and methods used to analyze a set of data. A statistic (singular) is a value that we calculate or infer from data. We get the median (a statistic) of a set of numbers by using techniques from the field of statistics.

Training and Testing

This is part of the machine learning workflow. When making a predictive model, you first offer it a set of training data so it can build understanding. Then you pass the model a test set, where it applies its understanding and tries to predict a target value.

Underfitting

Underfitting happens when you don't offer a model enough information. An example of underfitting would be asking someone to graph the change in temperature over a day and only giving them the high and low. Instead of the smooth curve one might expect, you only have enough information to draw a straight line.

Fields of Focus

As businesses become more data-focused, new opportunities open up for people of various skill sets to become part of the data community. These are some of the areas of specialization that exist within the data science realm.

Artificial Intelligence (AI)

A discipline involving research and development of machines that are aware of their surroundings. Most work in A.I. centers on using machine awareness to solve problems or accomplish some task. In case you didn't know, A.I. is already here: think self-driving cars, robot surgeons, and the bad guys in your favorite video game.

Business Intelligence (BI)

Similar to data analysis, but more narrowly focused on business metrics. The technical side of BI involves learning how to effectively use software to generate reports and find important trends. It's descriptive, rather than predictive.

Data Analysis

This discipline is the little brother of data science. Data analysis is focused more on answering questions about the present and the past. It uses less complex statistics and generally tries to identify patterns that can improve an organization.

Data Engineering

Data engineering is all about the back end. These are the people that build systems to make it easy for data scientists to do their analysis. In smaller teams, a data scientist may also be a data engineer. In larger groups, engineers are able to focus solely on speeding up analysis and keeping a data well organized and easy to access.

Data Journalism

This discipline is all about telling interesting and important stories with a data focused approach. It has come about naturally with more information becoming available as data. A story may be about the data or informed by data. There's a full handbook if you'd like to learn more.

Data Science

Given the rapid expansion of the field, the definition of data science can be hard to nail down. Basically, it's the discipline of using data and advanced statistics to make predictions. Data science is also focused on creating understanding among messy and disparate data. The "what" a scientist is tackling will differ greatly by employer.

Data Visualization

The art of communicating meaningful data visually. This can involve infographics, traditional plots, or even full data dashboards.

Nicholas Felton is a pioneer in this field, and Edward Tufte literally wrote the book.

Quantitiative Analysis:

This field is highly focused on using alogrithms for to gain an edge in the financial sector. These algorithms either recommend or make trading decisions based on a huge amount of data, often on the order of picoseconds. Quantitative analysts are often called "quants."

Statistical Tools

There are a number of statistics data professionals use to reason and communicate information about their data. These are some of the most basic and vital statistical tools to help you get started.

Correlation

Correlation is the measure of how much one set of values depends on another. If values increase together, they are positively correlated. If one values from one set increase as the other decreases, they are negatively correlated. There is no correlation when a change in one set has nothing to do with a change in the other.

Mean (Average, Expected Value)

A calculation that gives us a sense of a “typical” value for a group of numbers. The mean is the sum of a list of values divided by the number of values in that list. It can be deceiving used on its own, and in practice we use the mean with other statistical values to gain intuition about our data.

Median

In a set of values listed in order, the median is whatever value is in the middle. We often use the median along with the mean to judge if there are values that are unusually high or low in the set. This is an early hint to explore outliers.

Normalize

A set of data is said to be normalized when all of the values have been adjusted to fall within a common range. We normalize data sets to make comparisons easier and more meaningful. For instance, taking movie ratings from a bunch of different websites and adjusting them so they all fall on a scale of 0 to 100.

Outlier

An outlier is a data point that is considered extremely far from other points. They are generally the result of exceptional cases or errors in measurement, and should always be investigated early in a data analysis workflow.

Sample

The sample is the collection of data points we have access to. We use the sample to make inferences about a larger population. For instance, a political poll takes a sample of 1,000 Greek citizens to infer the opinions of all of Greece.

Standard Deviation

The standard deviation of a set of values helps us understand how spread out those values are. This statistic is more useful than the variance because it's expressed in the same units as the values themselves. Mathematically, the standard deviation is the square root of the variance of a set. It's often represented by the greek symbol sigma, σ.

Statistical Significance

A result is stasticially significant when we judge that it probably didn't happen due to chance. It is highly used in surveys and statistical studies, though not always an indication of pratical value. The mathematical details of statistical significance are beyond the scope of this post, but a fuller explanation can be found here.

Summary Statistics

Summary statistics are the measures we use to communicate insights about our data in a simple way. Examples of summary statistics are the mean, median and standard deviation.

Time Series

A time series is a set of data that's ordered by when each data point ocurred. Think of stock market prices over the course of a month, or the temperature throughout a day.

Residual (Error)

The residual is a measure of how much a real value differs from some stastical value we calculated based on the set of data. So given a

prediction that it will be 20 degrees fahrenheit at noon tomorrow, when noon hits and its only 18 degrees, we have an error of 2 degrees. This is often used interchangably with the term “error,” even though, technically, error is a purely theoretical value.

Variance

The variance of a set of values measures how spread out those values are. Mathematically, it is the average difference between individual values and the mean for the set of values. The square root of the variance for a set gives us the standard deviation, which is more intuitively useful.

Parts of a Workflow

While every workflow is different, these are some of the general processes that data professionals use to derive insights from data.

Data Exploration

The part of the data science process where a scientist will ask basic questions that helps her understand the context of a data set. What you learn during the exploration phase will guide more in-depth analysis later. Further, it helps you recognize when a result might be surprising and warrant further investigation.

Data Mining

The process of pulling actionable insight out of a set of data and putting it to good use. This includes everything from cleaning and organizing the data; to analyzing it to find meaningful patterns and connections; to communicating those connections in a way that helps decision-makers improve their product or organization.

Data Pipelines

A collection of scripts or functions that pass data along in a series. The output of the first method becomes the input of the second. This continues until the data is appropriately cleaned and transformed for whatever task a team is working on.

Data Wrangling (Munging)

The process of taking data in its original form and “taming” it until it works better in a broader workflow or project. Taming means making values consistent with a larger data set, replacing or removing values that might affect analysis or performance later, etc. Wrangling and munging are used interchangeably.

ETL (Extract, Transform, Load)

This process is key to data warehouses. It describes the three stages of bringing data from numerous places in a raw form to a screen, ready for analysis. ETL systems are generally gifted to us by data engineers and run behind the scenes.

Web Scraping

Web scraping is the process of pulling data from a website’s source code. It generally involves writing a script that will identify the information a user wants and pull it into a new file for later analysis.

Machine Learning Techniques

The field of machine learning has grown so large that there are now positions for Machine Learning Engineers. The terms below offer a broad overview of some common techniques used in machine learning.

Clustering

Clustering techniques attempt to collect and categorize sets of points into groups that are “sufficiently similar,” or “close” to one another. “Close”

varies depending on how you choose to measure distance. Complexity increases as the more features are added to a problem space.

Decision Trees

This machine learning method uses a line of branching questions or observations about a given data set to predict a target value. They tend to over-fit models as data sets grow large.Random forests are a type of decision tree algorithm designed to reduce over-fitting.

Deep Learning

Deep learning models use very large neural networks — called deep nets — to solve complex problems, such as facial recognition. The layers in a model start with identifying very simple patterns and then build in complexity. By the end the net (hopefully) has a nuanced understanding that can accurately classify or predict values.

Feature Engineering

The process of taking knowledge we have as humans and translating it into a quantitative value that a computer can understand. For example, we can translate our visual understanding of the image of a mug into a representation of pixel intensities.

Feature Selection

The process of identifying what traits of a data set are going to be the most valuable when building a model. It's especially helpful with large data sets, as using fewer features will decrease the amount of time and complexity involved in training and testing a model. The process begins with measuring how relevant each feature in a data set is for predicting your target variable. You then choose a subset of features that will lead to a high-performance model.

Neural Networks

A machine learning method that's *very* loosely based on neural connections in the brain. Neural networks are a system of connected nodes that are segmented into layers — input, output, and hidden layers. The hidden layers (there can be many) are the heavy lifters used to make predictions. Values from one layer are filtered by the connections to the next layer, until the final set of outputs is given and a prediction is made. A nice video explanation can be found here.

Supervised Machine Learning

With supervised learning techniques, the data scientist gives the computer a well-defined set of data. All of the columns are labelled and the computer knows exactly what it's looking for. It's similar to a professor handing you a syllabus and telling you what to expect on the final.

Unsupervised Machine Learning

In unsupervised learning techniques, the computer builds its own understanding of a set of unlabeled data. Unsupervised ML techniques look for patterns within data, and often deal with classifying items based on shared traits.