

Data Analytics Data Mining Comparisons Data Analysis Big Data Data Science
Machine Learning Computer Science

What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data?

Ad by Asana

...

What are great uses for Asana?

I work at Asana so I'm both biased and a power user. Here's a snapshot into a few ways you can use Asana effectively: Remote work Asana is the best way to work from anywhere and sta...(Continue reading)

100+ Answers



Rohit Malshe, Chem Engineer, Programmer, Amazon research scientist

Answered April 17, 2017 · Upvoted by Alan Davis, M.S. Computer Science & Machine Learning, The University of Texas at Dallas (2010) and Tanisha Sharma, B.E. CSE Computer Science (2020) · Author has **352** answers and **9.7M** answer views

I had been wanting to take a stab at this one since a few days, but it always looked like an enormous task, because this question has used too many words. In addition, this is a question on which a lot of people have their eyes, and a lot of others have already written elaborate answers.

Let me first re-order all the important words:

- Big data
- Data mining
- Data analysis
- Analytics
- Machine learning
- Data science

Imagine that you want to become a data scientist, and work in a big organization like Amazon, Intel, Google, FB, Apple and so on.

How would that look like?

- You would have to deal with **big data**, you would have to write computer programs in SQL, Python, R, C++, Java, Scala, Ruby...and so on, to only maintain big-data databases. You would be called a database manager.
- As an engineer working on process control, or someone wanting to streamline operations of the company, you would perform **Data Mining**, and **Data Analysis**; You may use simple software to do this where you would only run a lot of codes written by others, or you may be writing your elaborate codes in SQL, Python, R and you would be doing data mining, data cleaning, data analysis, modeling, predictive modeling and so on.
- All this will be called **Analytics**. Several software exist to do this. One popular one is Tableau. Some others are JMP and SAS. Lot of people do everything online where a SAP based business intelligence setup can be used. Here, simple reporting can be done easily.
- Further, you would then be able to use machine learning to derive conclusions, and come up with predictions, wherever analytical answers are not possible. Think of analytical answers as [If/then] type of computer programs, where all the input conditions are already known, and only a few parameters change.
- Machine learning uses statistical analysis to partition data. An example would be this: Read the comments written by various people on Yelp, and predict from the comments whether the person would have marked a restaurant 4 star



Sign in to Quora with Google



Michael Ryan Villanueva
michaelryanvillanueva@gmail.com



Michael Ryan Villanueva
thisismichael011@gmail.com

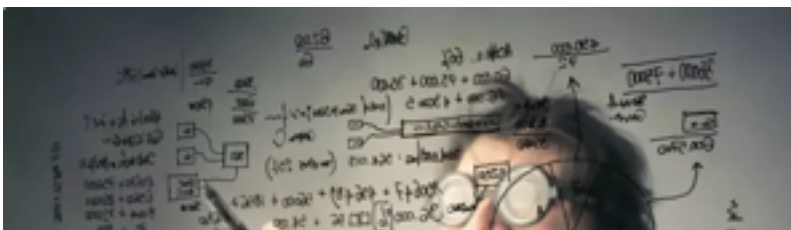
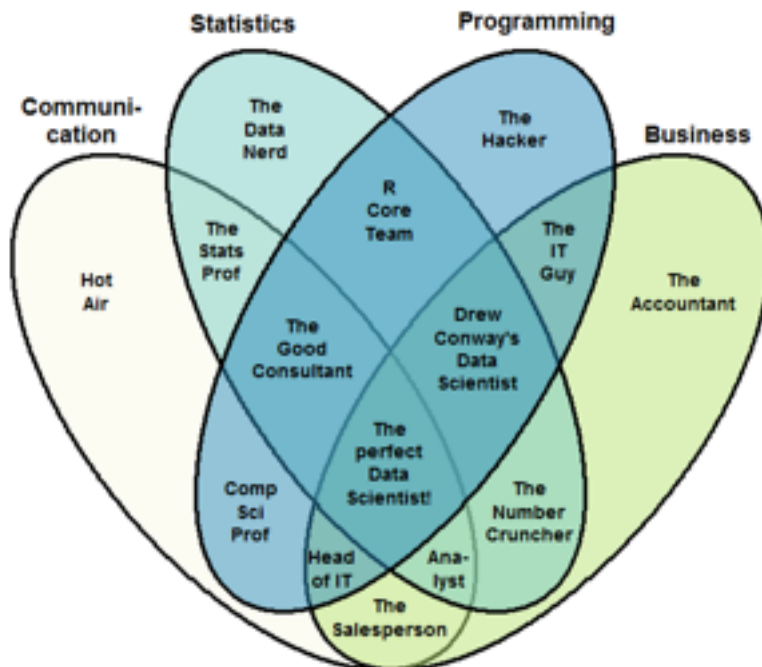
Which one is easy, Data Science or Cloud Computing?

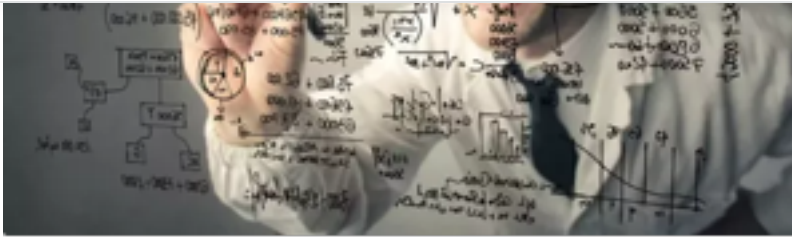
Do I need to know/learn Machine Learning if I want to pursue a career in Data Analytics?

What is the difference between data analytics and data mining?

What is data science?

- ## The Data Scientist Venn Diagram





- Read also:
- [Rohit Malshe's answer to How do I learn machine learning?](#)
- [Rohit Malshe's answer to How should I start learning Python?](#)
- [Rohit Malshe's answer to What is deep learning? Why is this a growing trend in machine learning? Why not use SVMs?](#)
- [Rohit Malshe's answer to Are 'curated paths to a Data Science career' on Coursera worth the money and time?](#)

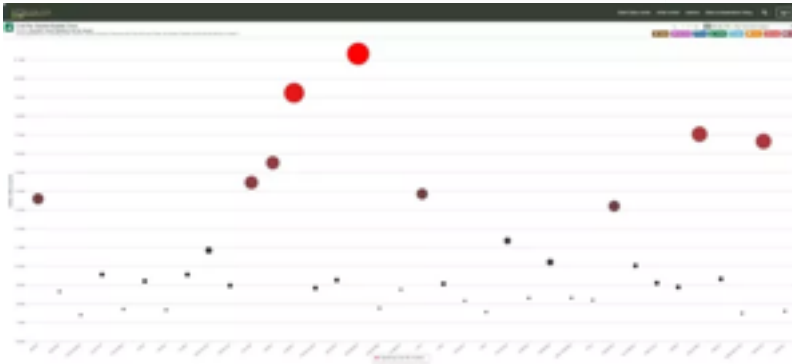
In all the seriousness, if you want a elaborate documentation on all this, I would suggest, go ahead and read this McKinsey report to get a full understanding. I only extracted a few sections out of it conveniently because I only wanted to add on the top of someone else's knowledge, and put together these concepts like a story so as to inspire the people to think about this subject and begin their own journeys.

[Big data: The next frontier for innovation, competition, and productivity](#) 

I will answer a few questions step by step, and wherever possible, I will give a few pictures, or plots to show you how things look like.

McKinsey consultants! You are amazing, so if you read things written in this answer that were typed by you at some point in time, I give full credit to you.

- **What do we mean by "big data"?**
- "Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we need not define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).
- What is a typical size of data I may have to deal with? Sometimes GBs, sometimes just a few MBs, sometimes up to as high as 1TB. Sometimes the complexity is nothing. The data may be representing the same thing. Sometimes the complexity can be very high. I might have a giant file full of a lot of data and logs which can be structured or unstructured.
- Think for example about Macy's. There are thousands of stores, selling thousands of items per day to millions of customers. If Macy's wants to derive a conclusion ~ should they rather diversify in shoes, or should they rather diversify in women's purses? How would they make this decision?
- **Well then, a natural question is: How do we measure the value of big data?**
- **Measuring data** Measuring volumes of data provokes a number of methodological questions. First, how can we distinguish data from information and from insight? Common definitions describe data as being raw indicators, information as the meaningful interpretation of those signals, and insight as an actionable piece of knowledge.



Let us now talk about analysis: This is big part of being a data scientist.

• TECHNIQUES FOR ANALYZING BIG DATA

- There are many techniques that draw on disciplines such as statistics and computer science (particularly machine learning) that can be used to analyze datasets. This list is by no means exhaustive. Indeed, researchers continue to develop new techniques and improve on existing ones, particularly in response to the need to analyze new combinations of data.
- Also, note that not all of these techniques strictly require the use of big data—some of them can be applied effectively to smaller datasets (e.g., A/B testing, regression analysis). However, all of the techniques listed here can be applied to big data and, in general, larger and more diverse datasets can be used to generate more numerous and insightful results than smaller, less diverse ones.
- **A/B testing.** A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate. This technique is also known as split testing or bucket testing. An example application is determining what copy text, layouts, images, or colors will improve conversion rates on an e-commerce Web site. Big data enables huge numbers of tests to be executed and analyzed, ensuring that groups are of sufficient size to detect meaningful (i.e., statistically significant) differences between the control and treatment groups (see statistics). When more than one variable is simultaneously manipulated in the treatment, the multivariate generalization of this technique, which applies statistical modeling, is often called “A/B/N” testing. What would an example look like?
 - *Imagine that Coke signs up with Facebook to work on marketing and sales. Facebook would put advertisements according to the customers. It can create versions of advertisements. Not all versions will suit to every geography. Some will suit to USA, some will suit to India. Some can suit to Indians living in USA. What Facebook can do is to choose a subset of people from a massive pool, and pass advertisements to them in their feed according to whether those people love food or not. For each advertisement, Facebook will collect the responses and accordingly determine which advertisement does better, and on a larger pool of people it will use a better one. Does data science let someone determine better what the answer should be? Absolutely!*
- **Association rule learning.** A set of techniques for discovering interesting relationships, i.e., “association rules,” among variables in large databases. These techniques consist of a variety of algorithms to generate and test possible rules. One application is market basket analysis, in which a retailer can determine which products are frequently bought together and use this information for marketing (a commonly cited example is the discovery that many supermarket shoppers who buy diapers also tend to buy beer).
- **Classification.** A set of techniques to identify the categories in which new data points belong, based on a training set containing data points that have already been categorized. One application is the prediction of segment-specific customer behavior (e.g., buying decisions, churn rate, consumption rate) where there is a clear hypothesis or objective outcome. These techniques are often described as supervised learning because of the

- **Cluster analysis.** A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance. An example of cluster analysis is segmenting consumers into self-similar groups for targeted marketing. This is a type of unsupervised learning because training data are not used. This technique is in contrast to classification, a type of supervised learning.
- **Crowdsourcing.** A technique for collecting data submitted by a large group of people or community (i.e., the “crowd”) through an open call, usually through networked media such as the Web.²⁸ This is a type of mass collaboration and an instance of using Web 2.0.²⁹ Data fusion and data integration.
- A set of techniques that integrate and analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data.
- **Data mining.** A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression. Applications include mining customer data to determine segments most likely to respond to an offer, mining human resources data to identify characteristics of most successful employees, or market basket analysis to model the purchase behavior of customers.
- **Ensemble learning.** Using multiple predictive models (each developed using statistics and/or machine learning) to obtain better predictive performance than could be obtained from any of the constituent models. This is a type of supervised learning.
- **Genetic algorithms.** A technique used for optimization that is inspired by the process of natural evolution or “survival of the fittest.” In this technique, potential solutions are encoded as “chromosomes” that can combine and mutate. These individual chromosomes are selected for survival within a modeled “environment” that determines the fitness or performance of each individual in the population. Often described as a type of “evolutionary algorithm,” these algorithms are well-suited for solving nonlinear problems. Examples of applications include improving job scheduling in manufacturing and optimizing the performance of an investment portfolio.
- **Machine learning.** A subspecialty of computer science (within a field historically called “artificial intelligence”) concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Natural language processing is an example of machine learning.
- **Natural language processing (NLP).** A set of techniques from a subspecialty of computer science (within a field historically called “artificial intelligence”) and linguistics that uses computer algorithms to analyze human (natural) language. Many NLP techniques are types of machine learning. One application of NLP is using sentiment analysis on social media to determine how prospective customers are reacting to a branding campaign. Data from social media, analyzed by natural language processing, can be combined with real-time sales data, in order to determine what effect a marketing campaign is having on customer sentiment and purchasing behavior.
- **Neural networks.** Computational models, inspired by the structure and workings of biological neural networks (i.e., the cells and connections within a brain), that find patterns in data. Neural networks are well-suited for finding nonlinear patterns. They can be used for pattern recognition and optimization. Some neural network applications involve supervised learning and others involve unsupervised learning. Examples of applications include identifying high-value customers that are at risk of leaving a particular company and identifying fraudulent insurance claims.
- **Network analysis.** A set of techniques used to characterize relationships among discrete nodes in a graph or a network. In social network analysis, connections between individuals in a community or organization are analyzed, e.g., how information travels, or who has the most influence over whom.

- **Optimization.** A portfolio of numerical techniques used to redesign complex systems and processes to improve their performance according to one or more objective measures (e.g., cost, speed, or reliability). Examples of applications include improving operational processes such as scheduling, routing, and floor layout, and making strategic decisions such as product range strategy, linked investment analysis, and R&D portfolio strategy. Genetic algorithms are an example of an optimization technique. Same way, mixed integer programming is another way.
- **Pattern recognition.** A set of machine learning techniques that assign some sort of output value (or label) to a given input value (or instance) according to a specific algorithm. Classification techniques are an example.
- **Predictive modeling.** A set of techniques in which a mathematical model is created or chosen to best predict the probability of an outcome. An example of an application in customer relationship management is the use of predictive models to estimate the likelihood that a customer will “churn” (i.e., change providers) or the likelihood that a customer can be cross-sold another product. Regression is one example of the many predictive modeling techniques.
- **Regression.** A set of statistical techniques to determine how the value of the dependent variable changes when one or more independent variables is modified. Often used for forecasting or prediction. Examples of applications include forecasting sales volumes based on various market and economic variables or determining what measurable manufacturing parameters most influence customer satisfaction. Used for data mining.
- **Sentiment analysis.** Application of natural language processing and other analytic techniques to identify and extract subjective information from source text material. Key aspects of these analyses include identifying the feature, aspect, or product about which a sentiment is being expressed, and determining the type, “polarity” (i.e., positive, negative, or neutral) and the degree and strength of the sentiment. Examples of applications include companies applying sentiment analysis to analyze social media (e.g., blogs, microblogs, and social networks) to determine how different customer segments and stakeholders are reacting to their products and actions.
- **Signal processing.** A set of techniques from electrical engineering and applied mathematics originally developed to analyze discrete and continuous signals, i.e., representations of analog physical quantities (even if represented digitally) such as radio signals, sounds, and images. This category includes techniques from signal detection theory, which quantifies the ability to discern between signal and noise. Sample applications include modeling for time series analysis or implementing data fusion to determine a more precise reading by combining data from a set of less precise data sources (i.e., extracting the signal from the noise). Signal processing techniques can be used to implement some types of data fusion. One example of an application is sensor data from the Internet of Things being combined to develop an integrated perspective on the performance of a complex distributed system such as an oil refinery.
- **Spatial analysis.** A set of techniques, some applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data set. Often the data for spatial analysis come from geographic information systems (GIS) that capture data including location information, e.g., addresses or latitude/longitude coordinates. Examples of applications include the incorporation of spatial data into spatial regressions (e.g., how is consumer willingness to purchase a product correlated with location?) or simulations (e.g., how would a manufacturing supply chain network perform with sites in different locations?).
- **Statistics.** The science of the collection, organization, and interpretation of data, including the design of surveys and experiments. Statistical techniques are often used to make judgments about what relationships between variables could have occurred by chance (the “null hypothesis”), and what relationships between variables likely result from some kind of underlying causal relationship (i.e., that are “statistically significant”). Statistical techniques are

determine what types of marketing material will most increase revenue.

- **Supervised learning.** The set of machine learning techniques that infer a function or relationship from a set of training data. Examples include classification and support vector machines.³⁰ This is different from unsupervised learning.
- **Simulation.** Modeling the behavior of complex systems, often used for forecasting, predicting and scenario planning. Monte Carlo simulations, for example, are a class of algorithms that rely on repeated random sampling, i.e., running thousands of simulations, each based on different assumptions. The result is a histogram that gives a probability distribution of outcomes. One application is assessing the likelihood of meeting financial targets given uncertainties about the success of various initiatives.
- **Time series analysis.** Set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data. Examples of time series analysis include the hourly value of a stock market index or the number of patients diagnosed with a given condition every day.
- **Time series forecasting.** Time series forecasting is the use of a model to predict future values of a time series based on known past values of the same or other series. Some of these techniques, e.g., structural modeling, decompose a series into trend, seasonal, and residual components, which can be useful for identifying cyclical patterns in the data. Examples of applications include forecasting sales figures, or predicting the number of people who will be diagnosed with an infectious disease.
- **Unsupervised learning.** A set of machine learning techniques that finds hidden structure in unlabeled data. Cluster analysis is an example of unsupervised learning (in contrast to supervised learning).
- **Visualization.** Techniques used for creating images, diagrams, or animations to communicate, understand, and improve the results of big data analyses. This expands into creating dashboards, on web or desktop platforms.



Hope this somewhat elaborate write up gives you some inspiration to hold on to. *Stay blessed and stay inspired!*

151.1K views · View 1.2K Upvoters · View Sharers

Related Questions

More Answers Below

[What is the exact difference between Big Data, Data Science & Data Analytics?](#)

[What is BIG DATA, DATA MINING and DATA ANALYTICS?](#)

[Which one is easy, Data Science or Cloud Computing?](#)

What is the difference between data analytics and data mining?



Naitik Chandak, Spent 5+ years in Machine learning

Answered June 9, 2017 · Author has **367** answers and **559.1K** answer views

I will try to give some brief Introduction about every single term that you have mentioned in your question.! Let's begin..

1. **Data Analytics** : Data Analytics often refer as the techniques of Data Analysis. It includes Algorithms, process of Data Mining methods, etc. Based on this techniques Data Scientist can figure it out which method gives more efficient / quick results with less calculations.

Continue Reading ▾

Promoted by DataCamp

ooo

How can I become a data scientist?



DataCamp, Teaching data science to more than 5.7MM learners worldwide.

Answered April 7, 2020

After being dubbed "sexiest job of the 21st Century" by Harvard Business Review, data science has stirred the interest of the general public. Many people are intrigued by the job and wonder how they themselves can become data scientists. [T? \(Continue reading\)](#)



Gam Dias, Defining Ecosystems for Personal Data

Updated June 19, 2017 · Upvoted by Nate Gadgibalaev, studied Computer Science and Karthikeyan Arasarethinam (கார்த்திகேயன் அரசரெத்தினம்), Microsoft .NET Architect, Learner of data science & analysis · Author has **395** answers and **831.7K** answer views

Lots of good answers already - however the question is such that I think perhaps a business rather than technical description might be warranted.

First things first, doing stuff with data, whatever you want to call it is going to require

Continue Reading ▾



Pallavi SK, Team Lead at Flipkart

Answered May 8, 2020

I will give a short brief on all these, wherein you can easily understand and will get to know the difference and similarities between these terms.

Data Analytics: Data analytics is the method of analyzing data sets to conclude the knowledge stored inside, particularly with the assistance of advanced systems and applications. In commercial enterprises, data analytics tools and techniques are

Continue Reading ▾

Related Questions

More Answers Below

[What is data science?](#)

[What is difference between Data Science and Big Data?](#)

[What are the differences between "Big Data" and Data Analytics?](#)

[What is the difference between the concepts of Data Mining and Big Data?](#)



Debidatta Dwibedi, works at Carnegie Mellon University

Updated January 17, 2017 · Upvoted by Ricardo Vladimiro, Data Science Lead @ Miniclip and Rushabh Shah, M.S. Computer Science, New Jersey Institute of Technology (2018) · Author has **53** answers and **2.1M** answer views

Originally Answered: What are the differences between machine learning and data science?

The following graphic nicely summarizes what all is involved in data science.



Continue Reading ▾

Sponsored by CloudFactory

Outsource confidently with these free resources.

Data drives every aspect of the AI model development life cycle. Which data hurdle is slowing you down?

[🔗 Learn More](#)

Data Mining & Modeling

Answered May 18, 2014 · Author has **65** answers and **174.3K** answer views

Originally Answered: Is there any concept map or description of the difference between the various concepts in data science, such as data analytics, business intelligence, big data, data mining, data warehousing, etc.?

It seems that your comment would be the equivalent of asking if there's a chart to show the overlap between all sports that exist today.

Given the list you have, it'd be hard to build a map with a sufficient level of detail to illustrate how the various methods, tools, and technologies are all groups together as some are as difference as ice hockey is to chess.

Continue Reading ▾



James Lee, Co-Founder of Level Up Academy and Ex-Googler

Answered May 16, 2018 · Upvoted by Frederick T. Williams, M.S. Data Science, Regis University (2020) · Author has **98** answers and **626.1K** answer views

This question has many phrases. Let's explain them one by one to have a deep understanding of the question:

- **Data Analytics**

Big Data analytics is a process in which large sets of data (Big Data) are collected, organized and analysed to discover useful patterns/findings, uncover hidden patterns, market trends and customers preferences. These patterns provide useful information that can help a company to produce future decisions. Data analytics are techniques of data analysis (discussed below). These techniques include algorithms and data mining

Continue Reading ▾

University of Oregon (2017)
Answered December 13, 2019

What is data science?

Data Science deals with both structured and unstructured data.

It is a field that includes everything that is associated with the cleansing, preparation and final analysis of data.

Data science combines the programming, logical reasoning, mathematics and statistics.


Data scientists are responsible for creating the data products and several other data

Continue Reading ▾



Animesh Niraj Ekka, Creative Head + Part time DevOps at Amazing Workz Studios (2011-present)

Answered November 6, 2017

[Data Science](#)  deals with structured and unstructured data. In principle, everything that relates to data cleansing, preparation and analysis lies within the scope of Data Science. There are different terms associated with Data Science.

Let's look how these terms differ.



Continue Reading ▾



Michael O. Church, studied at Carleton College

Answered September 8, 2014 · Upvoted by Ricardo Vladimiro, Data Science Lead @ Miniclip and Arjun Narayan, Ph.D student in computer science · Author has **1.5K** answers and **15.4M** answer views

Originally Answered: What is the connection between data science and machine learning?

massive amounts of data that humans would be unlikely to find. Like much of AI, it's an attempt to replace explicit programming (which becomes inflexible, costly, and illegible

[Continue Reading](#) ▾

Chinmayi Kashyap, B.E from Ramaiah Institute of Technology (2019)

Answered July 8, 2020

The funny thing is that the different roles that are mentioned in this question, which barely sound different from one other, belongs to one field that is Data Science. The term 'data' is similar in most of these roles because the main fuel for all these roles is the data, they all use data but their approach towards it will differ based on the

[Continue Reading](#) ▾

Related Questions

[What is the exact difference between Big Data, Data Science & Data Analytics?](#)

[What is BIG DATA, DATA MINING and DATA ANALYTICS?](#)

[Which one is easy, Data Science or Cloud Computing?](#)

[Do I need to know/learn Machine Learning if I want to pursue a career in Data Analytics?](#)

[What is the difference between data analytics and data mining?](#)

[What is data science?](#)

[What is difference between Data Science and Big Data?](#)

[What are the differences between "Big Data" and Data Analytics?](#)

What are the differences between Data Science and Data Mining, are they same?

What is the difference between Data analysis and Data analytics?

What is the difference between Data Management and Data Analytics?

What is difference between Big Data and Machine Learning?

Is data analytics and machine learning the same? Why?

Which is better, business analytics or data analytics?