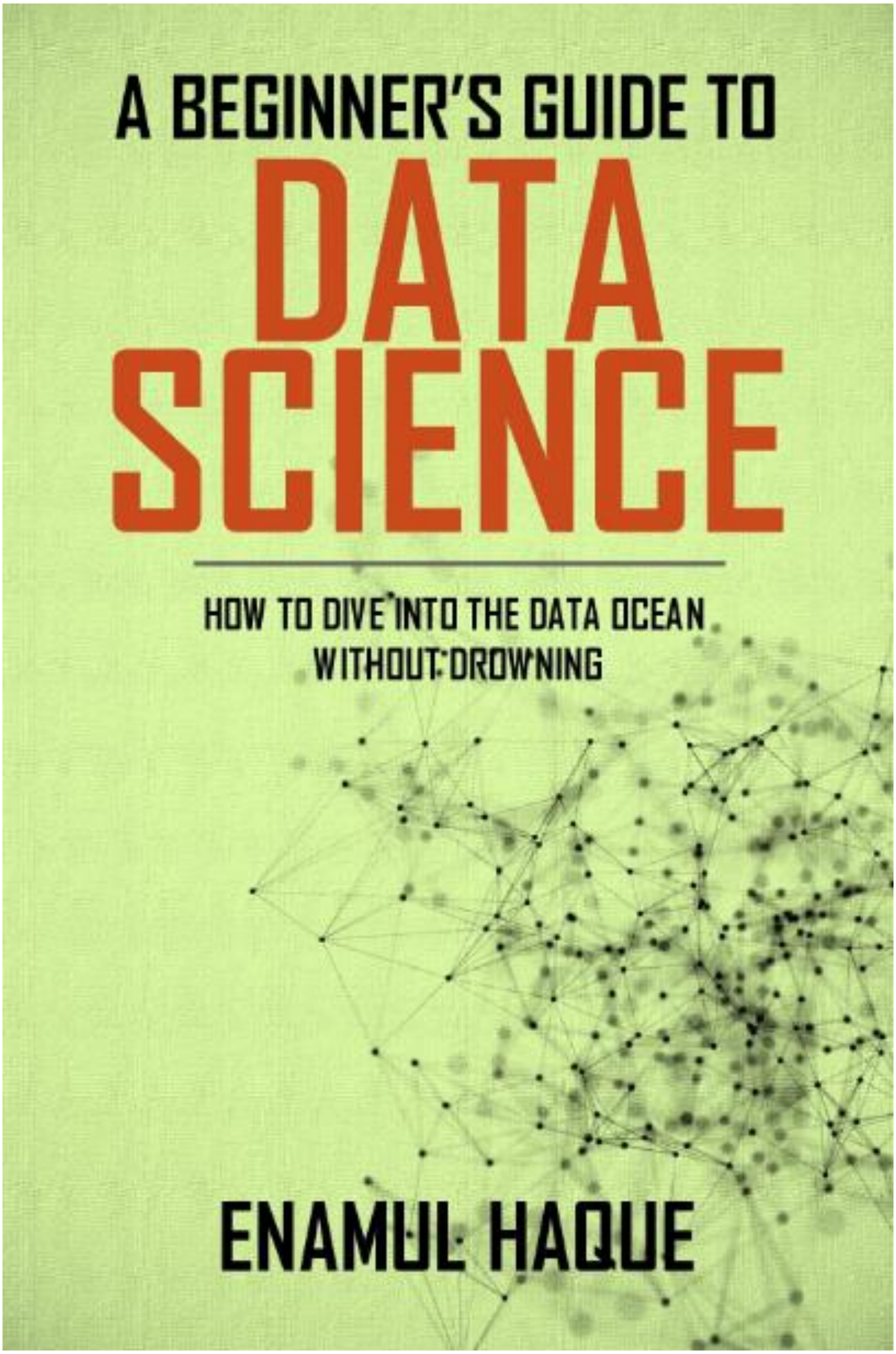


A BEGINNER'S GUIDE TO DATA SCIENCE

HOW TO DIVE INTO THE DATA OCEAN WITHOUT DROWNING

ENAMUL HAQUE

A Beginners Guide To DATA SCIENCE

How to dive into the data ocean without drowning

ENAMUL HAQUE

Enel Publications
London, UK
Amazon Kindle Direct Publishing

First Printing Edition, April 2021 (Revision 2)

ISBN 9798731261074

CHAPTER FOUR: DATA SCIENCE APPLICATIONS

"Data science is the future, and it is better to be on the cutting-edge than left behind." - Arnaud Perigord, Data scientist for the French Ministry of Social Affairs.

Basic Statistical Concepts for Data Science

Statistics is a branch of mathematics that deals with collecting, analysing, interpreting, and presenting masses of numerical data. If you include programming and machine learning in the mix, you have a pretty good description of data science's core competencies. Statistics are used in almost all areas of data science. It is used to analyse, transform, and clean up data. Evaluate and optimise machine learning algorithms. It is also used to present insights and insights.

The statistics field is vast, and it can be challenging to determine precisely what you need to learn and in what order. In addition, a lot of material for learning this topic is very complex. It can be very difficult to digest in some cases, especially if you don't have an advanced math degree and are moving from a field such as software engineering to data science.

Data scientist Rebecca Vickery[1] describes some of the basic statistical concepts you need to understand when learning data science. These are not particularly advanced techniques, but they are selecting the basic requirements that you need to know before you learn more complex methods.

Statistical sample

Statistics refer to the entire set of raw data that you might have available for a test or experiment as a population. For several reasons, you can't necessarily measure patterns and trends across the population. For this reason, we can use statistics to sample, perform some calculations for this data set, and, using the probability and some assumptions, to understand trends for the entire population or to predict future events.

Using statistics, we can take a sample of part of the population, perform some calculations for this data set, and use the probability and assumptions to understand trends with certainty that understand trends for the entire population.

Suppose, for example, that we want to understand the prevalence of a disease such as breast cancer in the whole population of the United Kingdom. For practical reasons, it is not possible to examine the entire population. Instead, we can take a random sample and measure the prevalence among them. Assuming that our sample is sufficiently randomised and representative of the whole population, we can estimate prevalence.

Descriptive statistics

Descriptive statistics help us, as the name suggests, to describe the data. In other words, it allows us to understand the underlying characteristics. It does not predict anything, makes no assumptions or completes nothing. It just describes what the data sample we have looks like.

Descriptive statistics are derived from calculations, often referred to as parameters. These include things like:

Mean - the central value, commonly referred to as average.

Median - the average if we have ordered the data from low to high and divided precisely in half.

Mode - the most common value.

Descriptive statistics are helpful but can often hide important information about the record. For example, suppose a document contains multiple numbers that are much larger than the others. In that case, the mean may be distorted and does not accurately represent the data.

A distribution is a chart, often a histogram, that shows the number of times each value is displayed in a record. This type of chart gives us information about the distribution and skewness of the data.

One of the essential distributions is the normal distribution, usually referred to as a bell curve due to its shape. It has an asymmetric shape, with most values grouped around the central peak and the more distant values evenly distributed on each curve's side.

Probability

The probability, in simple terms, is the likelihood of an event occurring. In statistics, an event results from an experiment that can be something like dice or an AB test result.

The probability for a single event is calculated by dividing the number of events by the total number of possible results. For example, if you throw a six on a dice, there are 6 possible results. So, the chance of dicing a six is 1/6 = 0.167; sometimes, it is expressed as a percentage, i.e., 16.7%.

Events can be either independent or dependent. For dependent events, a previous event affects the subsequent event. Suppose we have a bag of M & Ms and wanted to determine the probability that a red M & M will be selected at random. Each time we remove the selected M & M from the bag, the likelihood of picking red changes due to previous events' effects.

Independent events are not affected by previous events. In the M & M bag case, we put it back in the bag every time we select one. The probability of choosing red remains the same each time.

Whether an event is independent or not is important because we calculate the probability of multiple events changes depending on the type.

The probability of multiple independent events is calculated by simply multiplying the probability of each event. Suppose we wanted to calculate the probability of dicing 6 three times in the example of the dice's roll. This would look like this:

1/6 = 0.167 1/6 = 0.167 1/6 = 0.167

0.167 * 0.167 * 0.167 = 0.005

The calculation is different for dependent events, also known as conditional probability. If we take the example of M & M, let's imagine we have a bag with only two colours, red and yellow, and we know that the pack contains 3 red and 2 yellow, and we want to calculate the probability of selecting two red wines in a row. In the first selection, the probability of making a red selection is 3/5 = 0.6. We removed an M & M that was randomly red in the second selection, so our second probability calculation is 2/4 = 0.5. Therefore, the probability of picking two reds in a row is 0.6 * 0.5 = 0.3.

Bias

As explained in the statistics, we often use data samples to estimate the entire data set. Similarly, we will use some training data for predictive modelling and create a model that can make predictions about new data.

Bias is the tendency of a statistical or predictive model to underestimates a parameter. This is often due to the method of obtaining a sample or the way errors are measured. There are different types of distortions in statistics. Here is a brief description of two of them.

Selection Distortion - This occurs when the sample is not randomly selected. An example of data science can be to stop an AB test prematurely when the test runs or choose data to train a machine learning model from a specific period of time, mask seasonal effects.

Confirmation Distortion - This occurs when the person performing an analysis has a predetermined assumption about the data. In this situation, there may be a tendency to spend more time studying variables that are likely to support this assumption.

As explained earlier, the mean in a data sample is the central value. The variance measures how far each value in the record is from the mean. Essentially, it is a measurement of the variation of numbers in a data set.

The standard deviation is a common measure of the variation of data with the normal distribution. It is a calculation that specifies a value that indicates how far the values are distributed. A low standard deviation indicates that the values tend to be reasonably close to the mean, while a high standard deviation indicates that the values are more distributed.

If the data does not follow a normal distribution, other variance measures are used. The interquartile range is usually used. This measurement is derived by first dividing the values by rank and then dividing the data points into four equal parts, called quartiles. Each quartile describes where 25% of the data points are according to the median. The interquartile range is calculated by subtracting the median for the two central quarters, also known as Q1 and Q3.

Compromise between preload and variance

The concepts of bias and variance are essential for machine learning. When we create a machine learning model, we use a sample of data called a training record. The model learns patterns in this data and generates a mathematical function that can be used to associate the correct target label or target value (y) with a series of inputs (X).

When generating this mapping function, the model uses several assumptions to approximate the target better. For example, the linear regression algorithm assumes a linear relationship (straight line) between the input and the target. These assumptions distort the model.

The variance is the difference between the mean prediction generated by the model and the actual value in the calculation.

If we were to train a model using different training data samples, we would vary the returned predictions. The variance in machine learning is a measure of how big this difference is.

In machine learning, bias and variance are the overall expected flaw for our predictions. In an ideal world, we would have both low distortion and low friction. In practice, however, minimising the preload usually leads to an increase in variance and vice versa. The bias/variance compromise describes the process of compensating these two errors to reduce the overall error for a model.

Correlation

Correlation is a statistical technique used to measure relationships between two variables. The correlation is assumed to be linear (it forms a line when displayed in a chart) and is expressed as a number between +1 and -1. This is called a correlation coefficient.

A correlation coefficient of +1 denotes an entirely positive correlation (if the value for one variable also increases the value of the second variable), a coefficient of 0 does not mean correlation, and a coefficient of -1 denotes an entirely negative correlation.

Statistics is a wide and complex field. This article is intended as a brief introduction to some of the most commonly used statistical techniques in data science. Data science courses often require prior knowledge of these basic concepts or start with descriptions that are too complex and difficult to understand. I hope this article will serve as a refresher for a selection of basic statistical techniques used in data science before going into more advanced topics.

Methods and Metrics

Model evaluation metrics are used to assess the goodness of fit between model and data, compare different models in the context of model selection, and predict how predictions (associated with a specific model and data set) are expected to be accurate.[2]

Confidence interval

Modern definitions of variance have several desirable properties. Confidence intervals are used to assess how reliable and statistical estimate is. Wide confidence intervals mean that your model is flawed (and it is worth investigating other models) or that your data is very noisy if confidence intervals don't improve by changing the model (that is, testing a different theoretical statistical distribution for your observations). Modern confidence intervals are model-free, data-driven. A more general framework to assess and reduce sources of variance is called the analysis of variance.

Confusion matrix

Used in the context of clustering. These N x N matrices (where N is the number of clusters) are designed as followed: the element in the cell (i, j) represents the number of observations in the test training set (as opposed to the control training set, in a cross-validation setting) that belong to cluster i and are assigned (by the clustering algorithm) to cluster j. When these numbers are transformed into proportions, these matrices are sometimes called contingency tables. A wrongly assigned observation is called false positive (non-fraudulent transaction erroneously labelled as fraudulent) or false negative (fraudulent transaction erroneously labelled as non-fraudulent). The higher the concentration of observations in the confusion matrix's diagonal, the higher the accuracy/predictive power of your clustering algorithm.

Gain and Lift Chart

Lift is a measure of the predictive model's effectiveness calculated as the ratio between the results obtained with and without the predictive model. Cumulative gains and lift charts are visual aids for measuring model performance. Both charts consist of a lift curve and a baseline.

Kolmogorov-Smirnov Chart.

This non-parametric statistical test compares two distributions to assess how close they are to each other. In this context, one of the distributions is the theoretical distribution that the observations are supposed to follow (usually a continuous distribution with one or two parameters, such as Gaussian law), while the other distribution is the actual, empirical, parameter-free, discrete distribution computed on the observations.

Chi Square

It is another statistical test similar to Kolmogorov-Smirnov, but in this case, it is a parametric test. It requires you to aggregate observations in a number of buckets or bins, each with at least 10 observations.

ROC curve

Unlike the lift chart, the ROC curve is almost independent of the response rate. The receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates a binary classifier system's performance as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity or the sensitivity index d', known as "d-prime" in signal detection and biomedical informatics, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as (1 — specificity). The ROC curve is thus the sensitivity as a function of fall-out.

Gini Coefficient

The Gini coefficient is sometimes used in classification problems. Gini = 2*AUC-1, where AUC is the area under the curve (see the ROC curve entry above). A Gini ratio above 60% corresponds to a good model. Not to be confused with the Gini index or Gini impurity, used when building decision trees.

Root Mean Square Error (RMSE)

RMSE is the most used and abused metric to compute goodness of fit. It is defined as the square root of the absolute value of the correlation coefficient between true values and predicted values and is widely used by Excel users.

L^1 version of RMSE

The RMSE metric (see above entry) is an L^2 metric sensitive to outliers. Modern metrics are L^1 and sometimes based on rank statistics rather than raw data. One of these new metrics, developed by our data scientist, is described here.

Cross-Validation

This is a general framework to assess how a model will perform in the future; it is also used for model selection. It consists of splitting your training set into test and control data sets, training your algorithm (classifier or predictive algorithm) on the control data set, and testing it on the test data set. Since the actual values are known on the test data set, you can compare them with your predicted values using one of the other comparison tools mentioned in this article. Usually, the test data set itself is split into multiple subsets or data bins to compute confidence intervals for predicted values. The test data set must be carefully selected and must include different time frames and various types of observations (compared with the control data set), each with enough data points, in order to get sound, reliable conclusions as to how the model will perform on future data, or on data that has slightly involved. Another idea is to introduce noise in the test data set and see how it impacts prediction: this is referred to as model sensitivity analysis.

Predictive Power

It is related to the concept of entropy or the Gini index mentioned above. It was designed as a synthetic metric satisfying, interesting properties and used to select a good subset of features in any machine learning project or as a criterion to decide which node to split at each iteration when building decision trees.

Understanding Data Analysis

Data analysis is often used in enterprises and in government to make decisions. When you sent the last email, you created the data. When you entered the shopping site to make a purchase, you created data. This data is likely to be stored somewhere, usually either on your computer or on the company's servers.

Have you ever thought about asking yourself: what do people do with this data? That's a great question. There is a whole area called data analysis. Which is about finding out what a particular data set is using. Data analysis involves processing, cleaning, and understanding data to find a solution to a problem.

Let's look at the following scenario. The shopping website decides which product they should sell on their next sale. They want to sell a popular product to increase sales. A shopping website can use data analysis to determine which products are most popular. So, they can make a more informed decision about which product to put up for sale.

While people rely on their own intuition when making decisions, data analysis is based on the belief in numbers. As the size of the dataset increases, the reliability of someone's analysis increases. That's why companies collect so much data.

What is a data analyst?

A data analyst is the person charged with addressing what business, government, or other organisations want to answer. The data analyst will have a problem, for example, determining which product should go on sale in an online store. They will then use their knowledge of professional data analysis techniques to solve the problem.

The tasks that a data analyst solves depend on the industry in which they work. Governments use data analysis for applications such as protecting public health and predicting changes in the economy. On the other hand, companies use data analysis for everything, from analysing your app experience to figuring out which features users like best on the website.

What do data analysts do?

Every day, data analysts use technologies such as structured query language and mathematical libraries. Data analysts typically have a specific set of data to work with that contains a set of values. This is the work of a data engineer who works with data that needs to be analysed. It can be data on house sales, employee salaries, earthquakes, or something else, depending on the business's problem.

Data analysts first analyse the data set. To determine what data, it contains and what conclusions can be drawn from this data. They then use their understanding of this data to use different methods of data analysis. Such as statistical analysis in their research.

Once the data analyst analyses the dataset, he will combine his findings into a report. Their report should contain a recommendation or a series of recommendations based on data that suggest an answer to a question.

What skills are needed to analyse the data?

Data analysis requires a combination of mathematical skills, programming skills and business information analysis.

Encoding

To become a successful data analyst, you need to be able to program. This is because data analysis is very individual work. Each data set will be different. To be able to work effectively with a dataset, you need to know how to clean, process, and analyse data in a variety of ways. This is usually due to the use of a programming language like Python or R.

Data requirements analysis

It's easy to collect data. It's harder to get the correct data. Before the data analyst gets to work, he has to ask what kind of problem he needs to solve and what data is required in order to solve the problem.

Based on their answers to these questions, the data analyst will tell data engineers and other engineers what data points they need to find the answer to the question successfully.

Statistical analysis

Data analysts use solutions for the statistical analysis of datasets. This includes setting the limits of the data set. Use statistical principles, such as probabilities, to understand the data set and calculate the final results using the same principles.

Data visualisation

Data analysts are responsible for creating visual effects that represent what they discovered after the analysis. This is an important part of the job because data analysts usually answer questions from people with no data analysis experience.

Data analysts should be able to report their findings to other people without technical education. A great way to do this is to use graphs that are easier to interpret than lists of numbers. Data analytics tools, such as matplotlib[3] and Tableau, allow data analysts to create graphics and visuals for their work.

Data storytelling

Data storytelling takes data visualisations to the next level — data storytelling refers to "how" you communicate your insights. Think of it as a picture book. A good picture book has good visuals, but it also has an engaging and powerful narrative that connects the visuals.

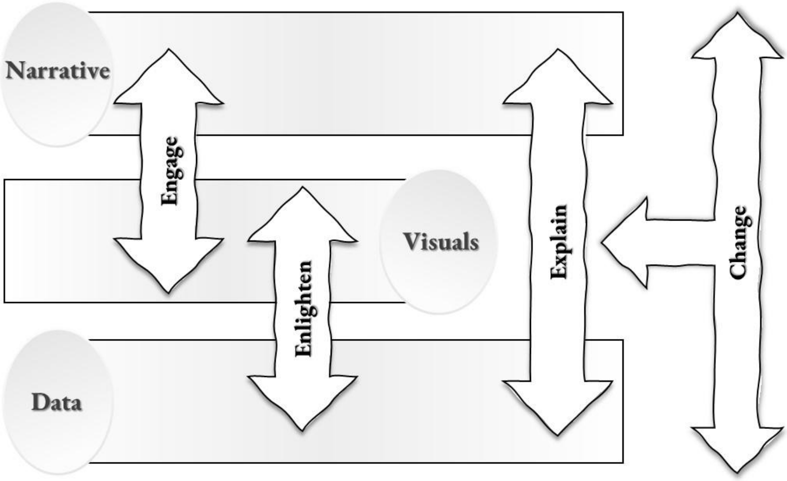


Figure 1 - Data storytelling framework

We can then point out the benefits of storytelling in a business intelligence project as follows:

- ***Make sense of the data:*** This allows you to interpret information to make it more meaningful and interesting.
- ***This is the key to understanding:*** An effective way to understand data and its importance is to tell a story. If there is no prior understanding, there will be no discernible change in decision making or significant action, no matter how important the analysis is.
- ***Build trust:*** Stories that use data and analysis are always more compelling. Involving real people for information and analysis gives you more confidence than a simple graph or obscure numbers.

- ***Provides simplicity:*** By fitting stories and gaining a better understanding, a more condensed view of data analysis and information is simplified, and exposure time is reduced.

Business analytics

While it's easy to think that data analysts are just sitting around analysing data, they need to do so in the context of a much more serious problem. Data analysts need to be well aware of business goals and how data can help them achieve them.

Data analysts work with people across the organisation every day to solve problems. This means that they need to know how to speak the language that engineers, directors, sellers, and other employees understand.

As a result, the problems are as follows: "How can this help achieve our organisation's goals?" These people are called business analysts. They use diagnostic analysis to solve business problems.

Cleaning up the data

The data is not included in a neatly packaged file with instructions. It's raw data. Data analysts need to figure out what to do with this data. Using a cleaning method, the data analyst looks through the dataset and makes sure it's structured the way they want it to.

This includes removing any inappropriate values, changing the value formats. Which are displayed incorrectly, and check the correctness of the values. The analysis can only begin once the dataset has been cleaned up.

Interpretation of data

Data analysts should be able to interpret the data. Not only do you need to know what the dataset can tell you, but it's also important that you can understand what it's telling you. It's useless to just know what data exists. You need to know what this data is talking about.

After the analysis, the data analyst will read the data he works with to determine trends. These trends will be included in the final report along with any visualisations and graphs prepared by the analyst.

Data analysis is critical to our modern economy. Today, data analysis is used by the insurance industry to predict insurance cases, the financial industry to predict the direction of the stock market, technology companies to analyse interactions with users.

Moreover, even the government relies on data to solve some problems. This is because data can help an organisation make a more informed and data-based problem. When you have data to back up a solution, it's easier to be sure you're on the right track.

Typically, data analysts use their knowledge of mathematics, dataset statistical analysis, and programming to solve business problems.

Data Cleanup

When data scientists talk about "cleaning up" data, it's hard to interpret it literally. This is reasonable because data scientists do not clean up the data. Cleaning up the data is to make a valuable dataset by removing and changing erroneous or irrelevant values.

What is data cleanup?

Cleaning up data is when a programmer removes incorrect and repetitive values from a dataset and ensures that all values are formatted the way they want them to. Cleaning up the data is called because it involves cleaning up "dirty data."

Rarely raw data comes in the form of a neatly packaged file that considers everything you need to do with the dataset. That's where the cleaning comes in.

When a data scientist receives a dataset, the first task he has to do is clean up the data. They need to take the time to read the data set. To make sure they can use it in their program.

Cleaning up the data is a good opportunity for a data specialist to get to know the dataset. By cleaning up the dataset, the data scientist learns more about what data is included in the dataset, how it is formatted, and what data they don't like.

Why is data cleanup so important?

Cleaning up data helps people working in data science improve the accuracy of their findings. The data specialist's job is to find answers to questions using data. If a data scientist works with incorrect data, their conclusion is unlikely to be accurate.

What's more, cleaning up the data helps save time in the future. Cleaning up the data precedes the analysis. This means that the data scientist analyses the data, and long before he draws any conclusions by the time, he draws any conclusions. Their dataset will be prepared exactly as they want.

Having a clean data set means that the data scientist can move forward to the analysis. Knowing that he doesn't have to go back and fix incorrectly formatted or remove inaccurate values. Ultimately, the data scientist wants their dataset to make sense and include all the data. Necessary to draw a reasonable conclusion on the issue.

How do you clean up your data?

Each data scientist follows their own data-cleaning procedure. Many organisations have their own standard rules. Make sure the dataset has been thoroughly cleaned before it can be used in any data analysis.

View missing data

Data scientists want all the data needed to be analysed ready before it can work. That's why the data analyst checks any missing data during the cleanup process. If the data isn't in the dataset, the data analyst can change their plan to not rely on that data. This needs to be carefully considered because this can change the final conclusions that a data scientist can draw.

A data scientist can decide to calculate missing values based on existing data. For example, if a data scientist needs an average number, he can calculate it with a program. They don't need to delete any analysis that depends on the average of their analysis. The data scientist can also add values such as 0 to make sure the program can easily process the dataset. These values will replace empty gaps in the dataset, which can cause structural errors.

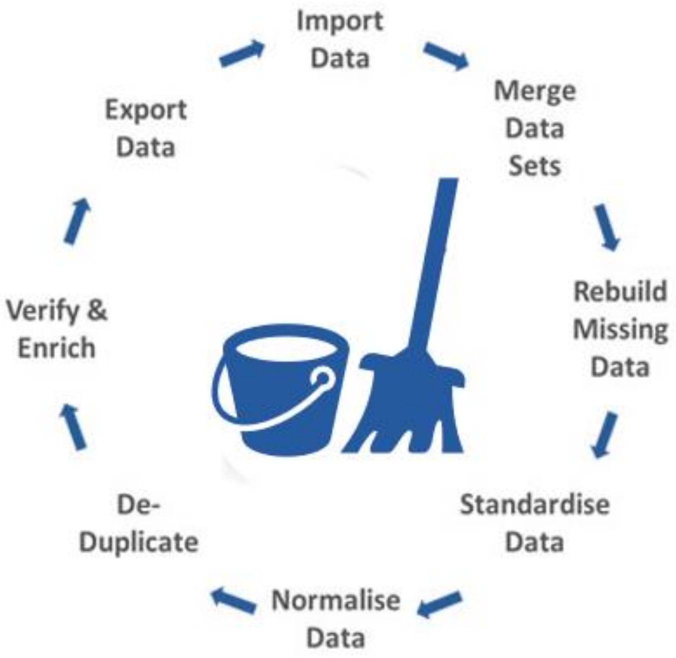


Figure 2 - Effective data cleaning

Remove useless data

Some of the data in the dataset doesn't add value to the data set. While it may be useful to have more data, some data points can distract the data specialist during the analysis. Before the analysis begins using data analysis tools, the data scientist will delete all data unrelated to its research. This will reduce the size of their dataset, thus making it easier to work with it.

Delete repetitive data

When the dataset is collected, it will likely get repetitive records. This can happen if the dataset has not been verified when collected or multiple datasets with combined data points.

Removing redundant data ensures that the findings are based on the right values. If there is to be repetitive data in the dataset, the data may deviate from one output over another. This will have a significant impact on the accuracy of the final conclusions.

Processing emissions data

The dataset may contain emissions values. For example, there could be one blank value or a damaged record. The data analyst will examine the dataset and make sure there are no emissions. If there is value ejecting, there are two options. A data analyst can completely remove emissions from the dataset. This is probably if the ejection value has a low chance of being accurate. The data scientist may also decide to double-check the value. This allows the data scientist to check errors when entering or collecting data before deleting the value.

Cleaning up data is a fundamental part of the data analysis process. The cleanup takes place after the data is collected and before the analysis. During the cleanup process, the data scientist will work to ensure that the dataset is valid, accurate and includes all the necessary values.

Without cleaning up the data, data scientists would have to switch between analysing the dataset and fixing basic data issues. This can confuse the data analysis process to such an extent that the conclusion loses its accuracy.

Data Mining vs Data Science

Choosing between two very similar fields makes it difficult to determine which one is best for you. In data mining and data science, it can be difficult even to find differences between them. Therefore, in such areas, it becomes even more challenging to decide before.

What is data mining?

Data mining is a term coined to transform raw data into useful and more understandable information. Many companies use data mining to find and analyse patterns in their marketing, revenue, expense, and sales data. This type of information is then used to make important decisions about marketing strategies and financial management.[4]

The data mining process begins with companies collecting data and uploading it to the data warehouses where it is stored and managed. Often it also loads into the cloud for storage. Business analysts are then involved in studying the data and deciding how best to organise and display it. This information is transferred to software designed to sort the data and then sorted and displayed as a diagram, graph, or table.

Data analysts do quite a lot of work to find these useful ideas. Typically, they are instructed to search for suitable data sets and variables to study, collect both structured and unstructured data, analyse and interpret data, and explain their findings to stakeholders in an understandable way.

Key differences between data mining and data science

Data mining is a process, and data science is an area of research.

What they are

The biggest difference between data mining and data science is simply what it is. Although data science is an extensive field of science, data mining is just a method used in this area. This means that data science encompasses a wider range of research and methods, while data mining focuses solely on data collection and transformation within a single process.

Focus

Data mining is usually used as part of the business analysis process. Typically, data mining is not used outside the business environment because it is explicitly designed to help companies collect and understand their data. On the other hand, data science is a scientific study. Data scientists use this study, among other things, to create predictive models, conduct experiments and social analysis.

Professionals in this field

In data mining, professionals are only expected to understand how to collect, organise, understand, and accurately display data. On the other hand, data scientists should have at least some qualifications in many fields, such as AI research, data engineering, data analysis, programming, and domain knowledge. To use data mining, you need to have some of the knowledge and skills that data scientists have, but not as great.

Data type

As a rule, data mining focuses only on structured data, although unstructured data can be used. For data scientists, using structured, unstructured and semi-structured data is common. Data mining is a little easier in this aspect because professionals cannot know how to work with all types of data, while data experts will probably need to know all types.

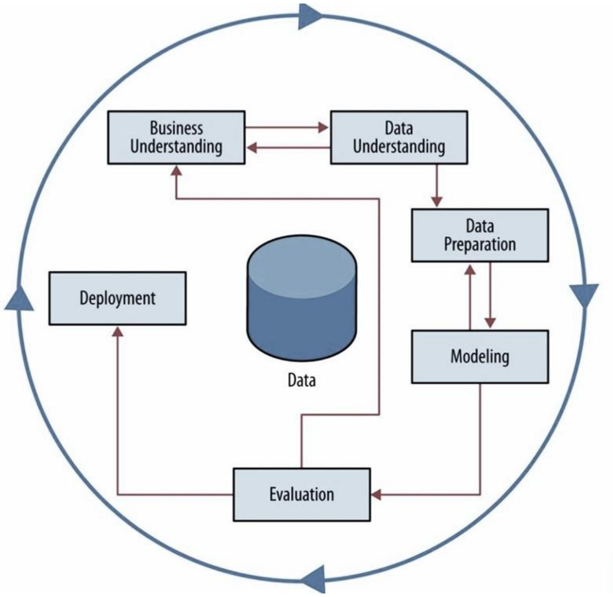


Figure 3 - Basics of data mining

Goal

Data mining's primary goal is to make business data easy to understand and therefore available for use. Data science aims to achieve scientific advances and create data-driven products for use by different organisations. In general, data mining has a much more specific purpose than data science. The whole purpose of data mining is to study and organise company data and identify previously unknown trends.

Data Science in Self-Driving Cars

According to Gartner, in 2018, just over 137,000 driverless cars were produced, and in 2019 - more than 330,000. Let us explore the basic concepts to help you navigate the topic and understand how data science makes a whole new chapter with this technology. This simulates the human brain and its cognitive networks, which works as a basis for a self-driving car.

Data scientists are the pioneers behind perfecting the brain of the beast (driverless cars). We must somehow figure out how to develop algorithms that master Perception, Localisation, Prediction, Planning, and Control.[5]

"Perception merges several different sensors to know where the road is and what is the state (type, position, speed) of each obstacle. Localisation uses precise maps and sensors to understand where the car is in its environment at the centimetre level. Prediction allows the car to anticipate the behaviour of objects in its surrounding. Planning uses the knowledge of the car's position and obstacles to planning routes to a destination. The application of the law is coded here, and the algorithms define waypoints. Control is to develop algorithms to follow the waypoints efficiently."[6]

Radars

Radars use radio waves to determine the distance to objects and the trajectory of their movement. As a rule, the radars on the drone are four pieces. The pulses they emit are reflected from objects, even if they are far away, and are transmitted to the receiving antenna. Thanks to radars, the system can instantly respond to changes in space.

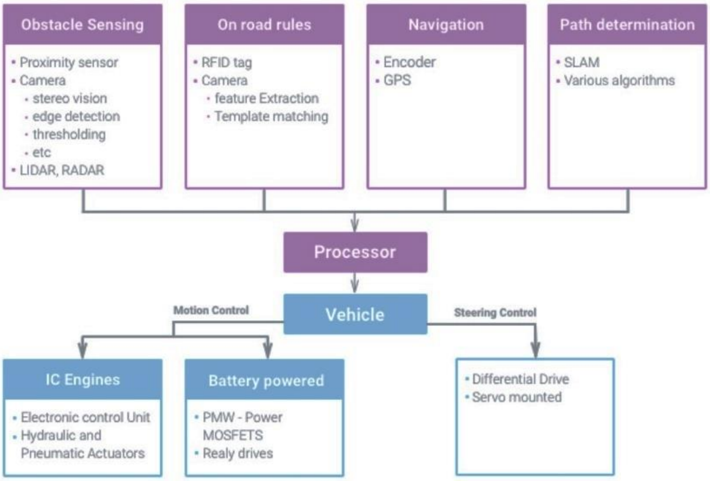


Figure 4 - Self-driving car architecture

Lidar

Lidar on the principle of action is similar to radar, but instead of radio waves uses laser beams. Today it is the most accurate tool for measuring the distance to objects - from a couple of centimetres to hundreds of meters - and their recognition. Lidar is installed on the car's roof or, if there are several lidars, along its perimeter. The device scans the space and creates a 3D map of the area.

Sensors

Sensors are the common name of lidars and radars. They scan the traffic scene around the car and avoid accidents. The driverless car can have as many as you want. For example, Roborace's first Robocar drone race car is equipped with 18 sensors.

Position sensor

The position sensor is a device that determines the drone's position on the map up to its coordinates.

Video Camera

The video camera is needed to distinguish traffic lights' colours to recognise road signs, markings, and people.

Computer

The computer is in the trunk of a driverless car and, in real-time, analyses all the data that comes from sensors and sensors. The power of the computer allows it to process a vast array of information.

Maps

High-precision maps allow drones to drive even on roads that do not have markings. They are needed in order for sensors and sensors to react only to changes in the situation on the ground. Otherwise, constant scanning of the surrounding space requires huge computing power. In addition, thanks to the maps, the car "understands" what is behind the turn - cameras and sensors cannot.

The level of driverlessness

The level of driverlessness is the degree of automation of the car. There are six of them - from 0 to 5. The zero levels of unmannedness - the car is fully controlled by the driver; level 5 - the car is able to reach its destination entirely on autopilot. To date, the most technically advanced machines are on the fourth level of unmanned.

Data Science Acronyms You Need to Know

ACID: *Atomicity, Consistency, Isolation and Durability*

ANOVA*: Analysis of Variance*

AOSD *: Aspect: Oriented Software Development*

AQL*: Annotation Query Language*

AUC*: Area Under the Curve (ROC curve)*

AUROC*: Area Under Receiver Operating Characteristic*

BDA*: Big Data Analytics*

CART*: Classification and Regression Trees*

CCA*: Canonical Correlational Analysis*

CEP*: Complex Event Processing*

CNN: *Convolutional Neural Network;*

COTS*: Commodity off:the: shelf*

CQL*: Cassandra Query Language*

CQL*: Contextual/Common Query Language*

CV*: Cross-Validation*

DAD*: Discover, Access, Distill*

DAG*: Directed Acyclic Graph*

DHSL *: Distributed Hadoop Storage Layer*

DNN: *Deep Neural Network or Deconvolutional Neural Network*

ECL*: Enterprise Control Language*

EDA*: Exploratory Data Analysis, Event-Driven Architecture*

FUSE*: Filesystem in Userspace*

GBM*: Gradient Boosting Machine*

GEOFF*: Graph Serialization Format*

GLM*: Generalized Linear Model*

GRU: *Gated Recurrent Unit*

HAR*: Hadoop Archive*

HMM*: Hidden Markov Model*

HPCC*: High-Performance Computing Cluster*

HPIL*: Hadoop Physical Infrastructure Layer*

ICA*: Independent Component Analysis*

IDA*: Initial Data Analysis*

JAQL*: JSON Query Language*

JSON: *JavaScriptObjectNotation Query Language*

KFS*: Kosmos File System*

kNN*: k: Nearest Neighbors*

LB*: LeaderBoard*

LDA*: Latent Dirichlet Allocation or Linear Discriminant Analysis*

LKOV*: Leave:k:Outcross:validation*

LLE*: Locally Linear Embedding*

LOOCV: *Leave:One:Outcross: validation*

LpO CV*: Leave:p:outcross: validation*

LSA/LSI*: Latent Semantic Allocation/Indexing*

LSTM*: Long Short Term Memory*

LZO*: Lempel–Ziv–Oberhumer*

MAPE*: Mean Absolute Percentage Error*

EDA*: Exploratory Data Analysis*

EPN*: Event Processing Nodes*

MDS*: Multidimensional Scaling*

MSE*: Mean Squared Error*

NLDR*: Non: Linear Dimensionality Reduction*

NLP *: Natural Language Processing*

NMF*: Non: Negative Matrix Factorization*

OLAP*: Online Analytical Processing*

OLTP*: Online Transactional Processing*

OOF*: Out Of Fold*

PCA*: Principal Component Analysis*

pLDA*: Probabilistic Latent Semantic Allocation*

PMML*: Predictive Model Markup Language*

R2 : *R: squared (regression metrics)*

RDD*: Resilient Distributed Database*

RF: *Random Forest*

RFE*: Recursive Feature Elimination*

RMS*: Root Mean Squared Logarithmic Error*

RNN: *Recurrent Neural Network*

ROC *: Receiver Operating Characteristic*

S4 *: Simple Scalable Streaming System*

SMOTE*: Synthetic Minority Oversampling Technique*

SOA*: Service-Oriented Architecture*

SVM*: Support Vector Machine*

TDA*: Topological Data Analysis*

tf:idf: *term frequency, inverse document frequency*

t:SNE*: t: Distributed Stochastic Neighbor Embedding*

UDTF*: User-Defined Tablegenerating Function*

UIMA*: Unstructured Information Management Architecture*

MCMC*: Markov chain Monte Carlo*

MDM*: Master Data Management*

VC*: Vapnik Chervonekis Dimension*

W3C*: World Wide Web Consortium*

XML*: Extensible Markup Language*

YARN*: Yet Another Resource Manager*

ZFS*: Zettabyte File System by Sun Microsystem*

About the Author

Enamul Haque is an author, researcher, and managing consultant best known for working with global companies such as Microsoft, Capgemini, Nokia, HCL Technologies, and the United Nations High Commissioner for Refugees (UNHCR) and International Telecommunication Union (ITU). He has over 26 years of rich experience in IT transformation and leading people for their professional growth and increase contribution to the organisation. Out of which, he treasured 13 years of experience in remote working and leading virtual teams.

As a consultant, Enamul worked with many of the world's best-known companies on their digital transformation and service integration strategies for improving business performance and value creation, including Alstom, Bayer AG, Bombardier, Britvic, Cadent, Carphone, Chanel, Direct Line Group, Estee Lauder Companies, Heathrow Airport, Neste, Rockwell Automation, Rogers, Sandvik, Shell, SJ Johnson, Terex, True-Value, Unilever, Warner Brothers, among many others. He assists in reskilling technical workforces to stay modern and ensure business continuity and compliance.

Enamul shares his industry knowledge among the MBA students as a guest lecturer at the University of Coventry, London campus. He worked very extensively as contributing writer for various newspapers, magazines, and other publications. Enamul is multilingual and lived and worked in many countries, including the USA, Switzerland, Finland, UAE, UK, India, and Germany.

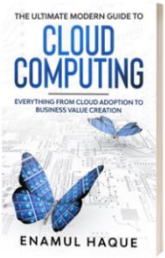
Enamul Haque studied mathematics and analytics (*Cours de mathématiques spéciales*) at the Swiss Federal Institute of Technology (EPFL), Lausanne, and architecture and Technology of computer science (license en science Informatique) at the University of Geneva. He also has a diploma in Artificial Intelligence and Machine Learning from the University of Helsinki. He is currently pursuing a Harvard and Capgemini co-branded program on foundational behaviours of managerial success (proximity, performance, and perspective). The program is based on three key areas, such as understanding the importance of Managerial behaviours and the impact they have on teams (including virtual teams), the ability to demonstrate and apply new managerial practices in a changing environment and to be equipped to enable a cultural shift towards a more substantial employee experience and engagement.

AUTHOR OFFICIAL WEBSITE: https://www.enamulhaque.co.uk/
ALL BOOKS BY ENAMUL HAQUE: https://enamulhaque.co.uk/my-books
ENAMUL HAQUE BLOG: https://enamulhaque.co.uk/my-articles
GOODREADS AUTHOR PROFILE: https://www.goodreads.com/haquenam
AMAZON AUTHOR PROFILE: https://www.amazon.com/ author/enamulhaque
TWITTER HANDLE @HAQUENAM: https://twitter.com/haquenam
LINKEDIN PROFILE: https://www.linkedin.com/in/haquenam
YOUTUBE TUTORIAL: https://www.youtube.com/c/digitaldeepdive
FACEBOOK AUTHOR PAGE: https://www.facebook.com/authorenam

Other Books by the Author

THE ULTIMATE MODERN GUIDE TO CLOUD COMPUTING

ISBN- 979-8666050637

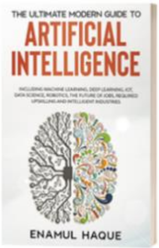

THE ULTIMATE MODERN GUIDE TO
CLOUD COMPUTING
EVERYTHING FROM CLOUD ADOPTION TO BUSINESS VALUE CREATION
ENAMUL HAQUE

This book has the most simplified explanation of Cloud Computing, starting from understanding digital transformation, enabling technologies to define essential characteristics, service models, deployment models, etc., with a pragmatic approach. It provides the path to digital transformation through the adoption of Cloud Computing to help construct Intelligent Enterprises.

THE ULTIMATE MODERN GUIDE TO ARTIFICIAL INTELLIGENCE

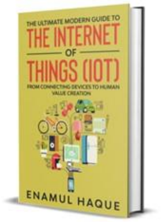
ISBN: 979-8691930768

This book has the most simplified explanation of Cloud Computing, starting from understanding digital transformation, enabling technologies to define essential characteristics, service models, deployment models, etc., with a pragmatic approach. It provides the path to digital transformation through the adoption of Cloud Computing to help construct Intelligent Enterprises.



THE ULTIMATE MODERN GUIDE TO THE INTERNET OF THINGS (IOT)

ISBN- 979-8691930768


THE ULTIMATE MODERN GUIDE TO
THE INTERNET OF THINGS (IOT)
FROM CONNECTING DEVICES TO HUMAN VALUE CREATION
ENAMUL HAQUE

The Internet of Things explained: Simply and Non-Technically. IoT is a computing paradigm in which several technologies that connect various devices based on wireless Internet acquire environmental information through sensors and control. This book provides a rigorous understanding of the IoT framework, characteristics, architecture, applications, technologies etc., in plain English to improve your awareness A key objective of this book is to provide a systematic source of reference for all aspects of IoT.

The Ultimate Modern Guide to Digital Transformation

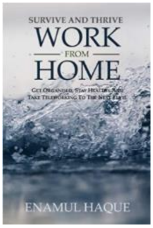
ISBN- 979-8702899572

In this book, you'll learn how new technologies disrupt businesses and how to transform to survive with the convergence of cloud computing, big data, artificial intelligence, the internet of things, and many other emerging technologies and how they are changing how we operate the 21st century. This book will give you the digital practices needed to catapults your organisation into next-level success.

THE ULTIMATE MODERN GUIDE TO
DIGITAL
TRANSFORMATION
THE "THRIVE OR DIE" THING CLARIFIED IN A SIMPLER WAY
ENAMUL HAQUE

Survive and Thrive Work from Home

ISBN- 979-8580562872



The impact of Pandemic and the new shifting trends for work-live balance. How we get there and it a success both for employees and employers. The fundamentals of remote working. Understanding the norms, teleworking history, benefits, challenges, and a very high-level overview of technology and culture's essential aspects. A collection of the best practices to do your work from home work for you. This includes the very best tips and tricks for working remotely by personality, job types etc. This has a selection of tops tools for remote use.

Cloud Service Management and Governance

ISBN- 978-1716788352

Once an organisation adopts cloud computing, it quickly becomes apparent that the traditional IT Service Management processes' traditional approaches will need to undergo drastic changes to integrate and run Bi-Modal IT Service Operations. This book is an alleyway to manage enterprise could services with a framework consisting of progressive Service Management practices to ensure practical, strategic, and modular methodology for the positive transformation of ITSM for cloud delivery models.

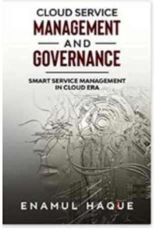


Table of Figures

List of Tables

No table of figures entries found.

Notes and References

[1] **Rebecca Vickery** - 8 Fundamental Statistical Concepts for Data Science - https://towardsdatascience.com/8-fundamental-statistical-concepts-for-data-science-9b4e8a0c6f1c

[2] **L.V** - 11 Important Model Evaluation Techniques Everyone Should Know - https://www.datasciencecentral.com/profiles/blogs/7-important-model-evaluation-error-metrics-everyone-should-know

[3] **Matplotlib** is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. [Wikipedia]

[4] **Saira Tabassum** - Data Mining vs Data Science: The Key Differences for Data Analysts - https://careerkarma.com/blog/data-mining-vs-data-science/

[5] **Fei Qi** - The Data Science Behind Self-Driving Cars - https://medium.com/@feiqi9047/the-data-science-behind-self-driving-cars-eb7d0579c80b

[6] **Jeremy Cohen** - AI & Self-Driving Car Engineer —I teach people how to join the Autonomous Tech world! https://www.thinkautonomous.ai