# Data science using Python

======================================================================================================================================================================

What is Probability?

Probability is a mathematical subject which enables us in determining or predicting how likely it is that an event will happen. The probability of occurrence is assigned a value from 0 to 1. When the value assigned is 1, it implies that the event will happen with all certainty. On the other hand when it is 0, it implies that the event will never take place. Thus, we can be more certain of an event's occurrence when its probability is higher.

What is Statistics?

Statistics is another mathematical subject which deals primarily with data. It helps us draw inferences from data by having procedures in place for collecting, classifying and presenting the data in an organized manner. The analysis and interpretation of the refined data helps in providing further insights.

Role of probability, statistics and computation in data science

When studying and exploring an event, we make use of probability to quantify how likely it is that an event will occur. On the other hand, we use statistics to observe patterns in data samples to draw inferences about a population. We must note that statistics is not completely independent of probability, as statistical analysis involves probability distributions.

Since, both statistics and probability have their roots in mathematics, computation as a tool is needed to perform quantitative analysis. The use of computers is also necessary to perform complex calculations while processing the statistical data.

=============================================================================

============== # linear Algebra

What is Linear Algebra?

Linear Algebra is a mathematical subject the deals with the theory of systems of linear equations, matrices, vector spaces and linear transformations.

Why Linear Algebra?

Linear Algebra is critically used in almost all peripheries of science, practically solving most of the problems using linear models.

Most of the complex science problems are converted into problems of vectors and matrices and then solving it with linear models.

In the world of data (especially, big data), linear algebra can be very handy to process huge chunks of data to accomplish many practical transformations such as graphical transformations, face morphing, object detection and tracking, audio and image compression, edge detection, blurring, and signal processing.

Role of linear algebra in data science

While solving a given business problem, an appropriate statistical computing technique may be used. These algorithms while working on the data, may either use iterative methods or linear algebra techniques for computation.

Linear Algebra works as a computational engine for most of the data science problems because of its performance advantages over iterative methods. Let us discuss a simple example, to understand the difference between the two methods.

============================

Linear Algebra techniques

Say we need to find the Frobenius norm of a matrix(mat1). We can do this using either the iterative method or the linear algebra technique, as shown below.

No\_cols <- 10000

no\_rows <- 1000

frobenius\_norm <- norm(as.numeric(mat1),type = 'F')

# here norm is R function

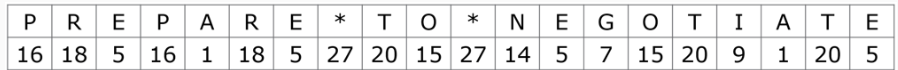
=============================================================================

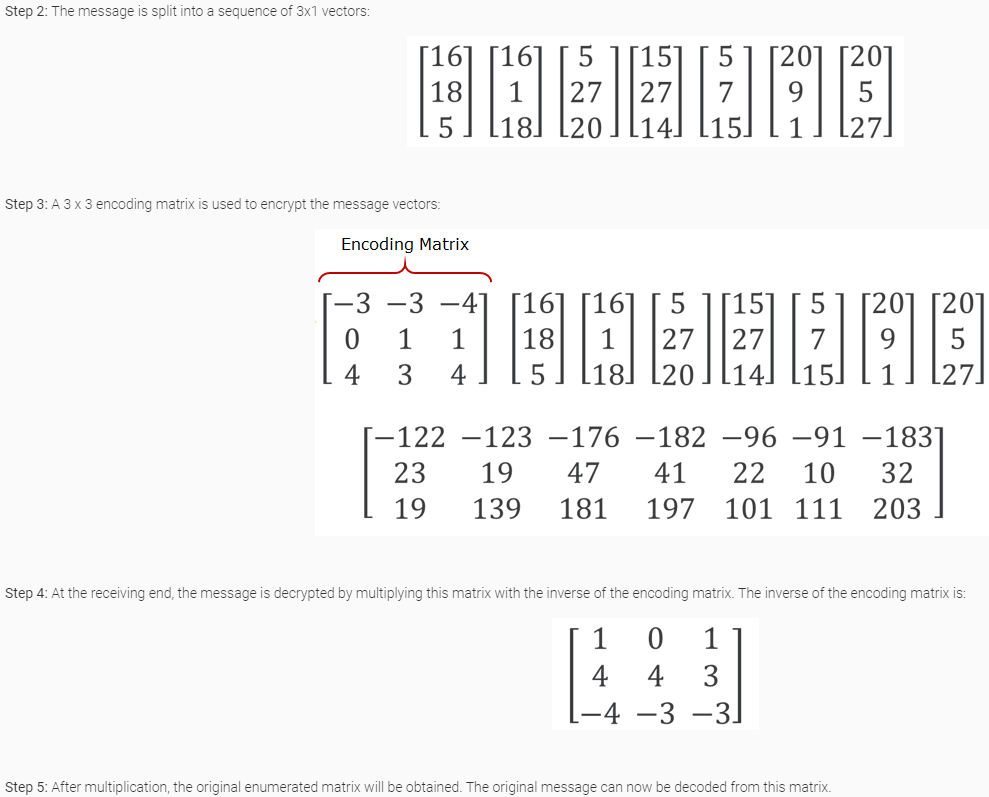
**Example 1: Message Transmission**

**Problem statement**

* We need to transmit a message over the network: “PREPARE to NEGOTIATE”.
* When transmitting we need to encrypt the message and at the receiving end we need to decrypt the message.
* To encrypt and decrypt, we need to use a confidential piece of information, usually referred to as a key.
* The prime objective is to ensure confidentiality and privacy of data during transmission.

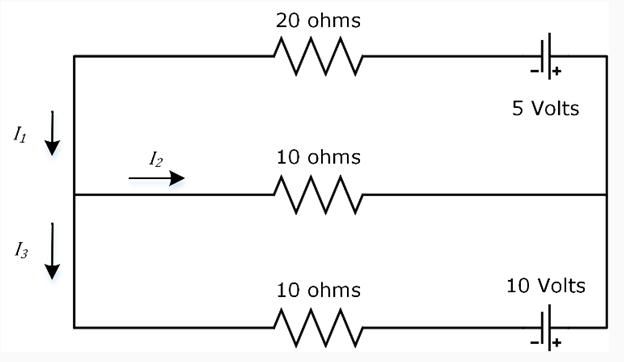
**Solution :Step 1:**The message is encrypted by assigning a number for each letter in the message. Thus, the message becomes:



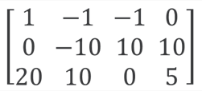
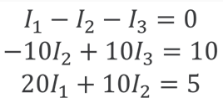


# Example 2: Solving an electrical network

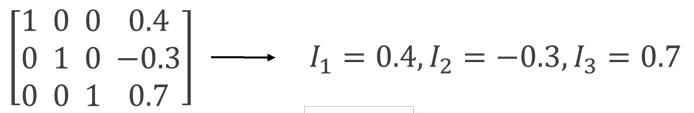
Problem statement: Currents I1, I2 and I3 need to be determined for the following electrical network:



# Solution: **Step 1:** The equations for current are written based on Kirchhoff’s Law.

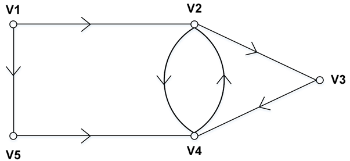
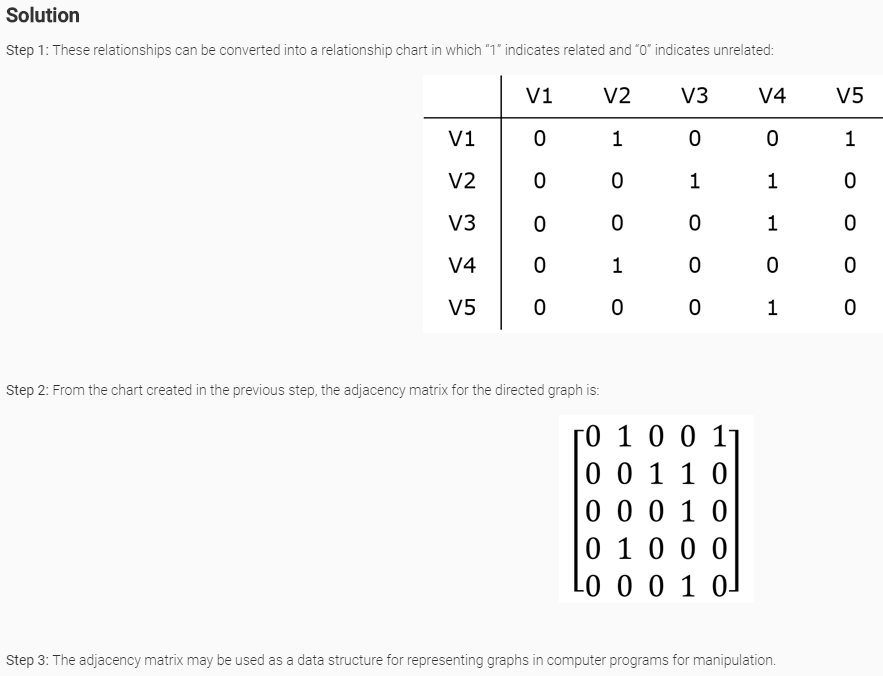
**Step 2:** These equations are converted into a matrix.

**Step 3:** The matrix is solved to get the values of the currents.

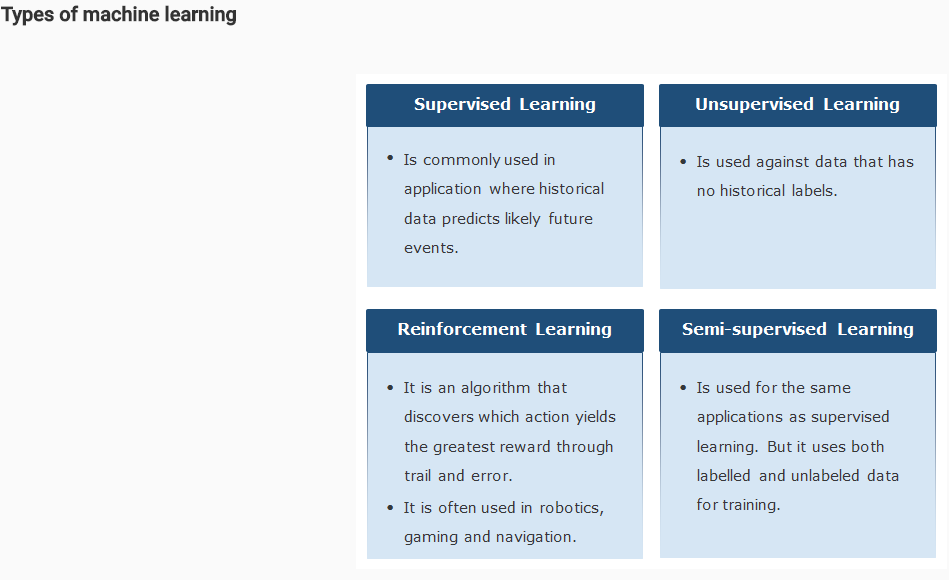


# Example 3: Finding relationships on a social networking site

Problem statement: Five visitors of a social networking site are linked with each other as depicted by the directed graph G below: how can we use these relationships to extract more information about them and predict their proposed activities?



=============================================================================



# Supervised machine learning model: Testing phase

A new set of images is given to this model as  test data so that it can classify different fruits.

**Supervised machine learning model: Types**

There can be two types of supervised machine learning techniques as shown below:

* **Classification:** Used to predict discrete results.

For example, assume a company wants to predict the budget period of a new project that they have acquired as short-term/long-term, based on various input attributes about the project such as number of resources required, software requirement, hardware requirement etc., then we use the classification technique.

* **Regression:** Used to predict continuous numeric results

For example, if we are trying to predict the approximate budget requirement of a new project that the company has acquired in actual quantifiable figures, based on various input attributes about the project such as number of resources required, software requirement, hardware requirement etc., then we use the regression technique.

=============================================================================

# Unsupervised machine learning model

There is a basket filled with some fresh fruits. The machine’s task is to group similar colored fruits together.

# Unsupervised machine learning model – Clustering

Machine identifies four clusters of fruits based on their color, as shown below.

# Semi-supervised machine learning

In real-time, it may so happen that the unlabeled data points exceed the number of labeled data points in a data set. In order to fit a model to such data we use the semi-supervised machine learning technique, wherein we perform the following steps:

* **Step 1:** Train the model with labeled data points only.
* **Step 2:** Use the above model to predict the labels of the unlabeled data points
* **Step 3:** Combine the existing labeled data points with the newly labeled data points and use it to retrain the model
* **Step 4:** Repeat the 2nd and 3rd steps until it converges

Applications of semi-supervised learning are text processing, video-indexing, bioinformatics, web page classification and news classification among others.

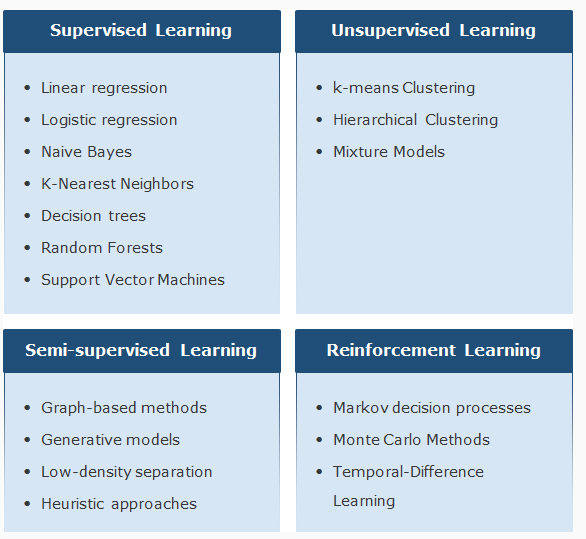
# Reinforcement learning

Reinforcement machine learning algorithm is a reward based and immediate feedback technique. Here the machine's goal is to maximize the numerical reward at each and every step. In the process of learning, the machine is not provided any supervision as opposed to the previous ML algorithms we discussed till now. Instead the machine is expected to figure out the optimum actions which will reap the maximum reward at each step, all on their own, without any interference.

The actions that the machine takes at each step might not only affect the immediate reward but may also affect all the subsequent rewards. The ultimate aim is to reach the max possible reward in the least amount of steps possible.Thus, trial and error search methodology and immediate feedback in the form of a numerical reward are the two main characteristics of reinforcement learning.

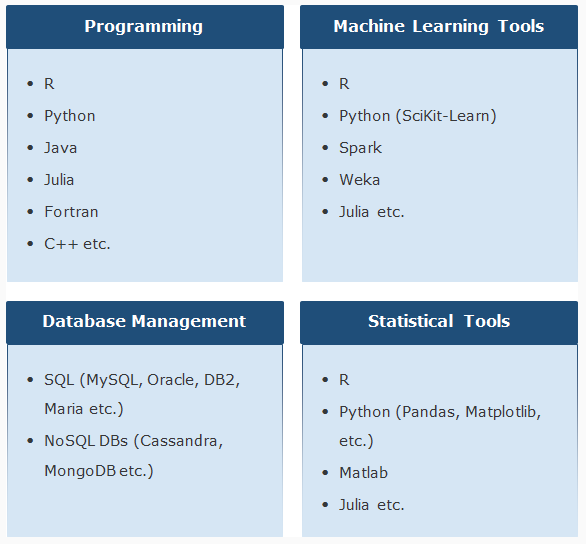
An example of reinforcement learning would be when a machine learning to play chess decides whether a move is right by planning the possible moves, anticipating the corresponding counter moves and finally choosing one based on reward based appeal for a particular position or set of moves. Another example could be when a trash collecting bot's charge is about to reach critical levels and it needs to make a decision, to clean one more room before reaching out for the charging station or to immediately rush to the nearest charging station. The decision taken by the bot depends on the ease with which it can reach the charging station, based on its prior knowledge.

# Some famous machine learning algorithms



# Tools and packages for Data Science

Following a are a few tools and packages available which enable in the application of data science solutions on huge amounts of digital data.

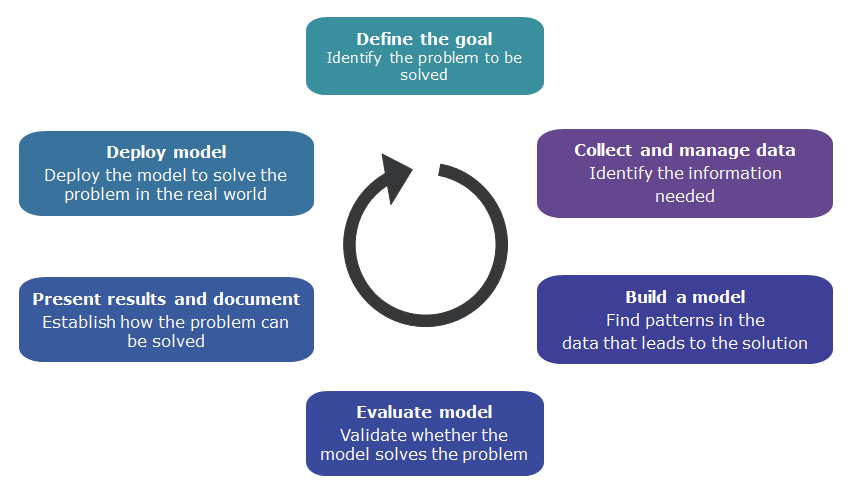


# Putting the concepts together: Data Science process

# 

Historical data/ Past data from various sources are cleaned and subjected to Feature Engineering, which is the process of using domain knowledge to select/create significant features from the historical data relevant to the problem statement. This engineered data is divided into two sets: Train data and Test data. Data science models are built using train data which are validated using test data. This validated model is used for taking various decisions on new/unseen data points.

**The complete life cycle of a data science project is shown below.**



# Data Science implementation - Business use case

Country Bank of India wants to cut down on their losses due to bad loans. It approaches a data analytics firm to help them reduce these losses by X%.

# Step 1: Define the goal

The first and foremost step in any project is to define a clear goal. Hence, at this point, it is important to learn every minute detail about the project such as:

* Why is the project being started? What is missing currently and what exactly is required?
* What are they currently doing to fix the problem, and why isn’t it working?
* What all resources are needed? What kind of data is available? Is domain expertise available within the team? What are the computational resources available/required?
* How does the business organization plan to deploy the derived results? What kind of problems need to be addressed for successful deployment?

# Bad loan use-case: Define the goal

The goal is to lessen the bank's losses caused by bad loans. To do this the firm intends to create a tool to help the bank's loan officers to improve their accuracy in identifying bad loan applicants, thereby lowering the number of bad loans being authorized. For this purpose, the goal defined should be to the point and unambiguous. For example, a goal which states "We want to reduce the rate of loan charge-offs by at least 10%, using a model which predicts whether loan applicants are likely to default" is preferred over "We want to get better at finding bad loans".

# Step 2: Collect and manage data

Now that the goal has been set, the next step is to find, explore and clean the data necessary for analysis. This stage takes up a lot of time but helps in finding answers to many important questions, such as:

* What all data is available?
* Will it help in solving the problem?
* Is the data enough to carry out analysis?
* Is the quality of data up to the mark?

# Bad loan use-case: Collect and manage data

* Collect the data about each loan application with relevant attributes such as status of loan, duration, credit history of the applicant, present employment status, residing at an address since, number of dependents, and the number of active loans under the applicant’s name.
* Collect the data across a reasonable span of time such as one year or one decade.
* Conduct initial exploration (using data visualization and summary statistics) and clean the data.
* While refining the data, it may so happen that the data identified earlier turns out to be not adequate to perform the analysis.
* There might also be a situation wherein we encounter various new problematic areas within the data, which we disregarded as not being a problem at all, previously! For example, if the data set we took contained most of the defaulters or just a few defaulters, our analysis may result in a biased conclusion.

# Step 3: Build a model

Once the data is ready, the next step is to find meaningful insights from the data. Depending on the nature of the business problem we are dealing with we can make use of any of the following data modelling techniques to gather such insights.

* **Classification:** Determining which among the given categories a data point falls under
* **Scoring:** Predicting or estimating a quantifiable value
* **Ranking:** Ordering the data points depending on the priorities involved
* **Clustering:** Grouping similar items based on certain parameters
* **Finding relations:** Finding associations between various features of the data
* **Characterization:** Creating plots, graphs and various reports for understanding the data better

# Bad loan use-case: Build a model

* In the bank scenario, the problem we are dealing with is classification. We wish to classify bank customers who apply for loans as  probable defaulters or non-defaulters. Hence, we need to train our model in such a way that it covers the entire range of the available data, thus enabling it to learn about most of the probable loan defaulter cases.
* Given the preceding requirement, we decide on a suitable approach to build the model. We can choose from either logistic regression, naive Bayes, k-nearest neighbours or decision trees, among other available classification techniques.
* We also need to be aware of why a model is taking a particular decision and how confident it is in its prediction. Ultimately, our model should be able to answer the question "how likely is an applicant to be a defaulter"?

# Step 4: Model evaluation

Now that we have built our model, we need to determine whether it meets our goals by asking the following questions:

* Is the model accurate enough for our needs?
* Does the model meet the expectations? Is it better than the methodology being currently used?

If for either of the above questions, the answer is NO, we need to revisit the previous steps.

# Bad loan use-case: Model evaluation

* Check whether the evaluation parameters from the suggested model apply in our scenario.
* Proceed to calculate the model evaluation parameters (such as accuracy and precision) based on the predefined rules and observe how many predicted values match the actual values.

# Step 5: Present results and document

At this stage we have achieved a desirable model. A model that meets all the requirements and goals we set for ourselves at the beginning of the project. The next step is to showcase the project to various audience as follows:

* Present the details of the model to all the collaborators, clients and sponsors.
* Provide everyone in charge of usage and maintenance of the model, once deployed, with documentation that covers all aspects of the working of the model.

We must keep in mind that each group of people involved in the project require different kind of treatment, when it comes to presentations and providing documentation. Hence, specific data visualization techniques must be used for each of them. What might work for one audience, might not work for the other.

# Bad loan use-case: Present results and document

In the bad loans scenario, we need to showcase the efficacy of our model to the bank officials. For example, if our model is able to identify a set of bad loans that amount to 36% of the total money lost to defaulters, we need to emphasize the same in the executive summary. In addition to this, any interesting insights uncovered such as, new auto loans are more riskier than used auto loans; or less losses are tied to home loans, must also be presented.

# Step 6: Deploy model

The last and final step is to deploy the model. Usually from this point ahead the data scientist is no longer associated with the operations of the model. But before they are off the job they must make sure that the following are in place:

* The model has been tested thoroughly and generalizes well.
* The model should be able to adjust well to unforeseen environmental changes.
* The model has been deployed in a pilot program and any problems that cropped up in the last moment were taken care of by updating the model accordingly.

# Bad loan use-case: Deploy model

There may arise a situation wherein experienced loan officers might veto the decision taken by the model that we created as it opposes their instincts. Hence, we need to be always on the look out for which is correct, our model or their intuition?

Data Science Top 10 Real World Use Cases

# Churn Prediction

Churn implies loss of customers to competition. For any company, it costs more to acquire new customers than to retain the old ones. As churn prediction aids in customer retention, it is extremely important especially for businesses with a repeat customer base. The application of this model cuts across domains such as Banking, E-Retail, Telecom, Energy and Utilities.

# Sentiment Analysis

Also referred to as opinion mining, it is the process of computationally identifying what customers like and dislike about a product or a brand. A domain which relentlessly makes use of sentiment analysis is the Retail industry. Companies like Amazon, Flipkart, Reliance, Paytm use customer feedback from social networking sites like Facebook, Twitter, etc. or their own company websites to find out what their customers are talking about and how they feel i.e. positive, negative, or neutral. They leverage this information to reposition their products and provide better/new services.

# Online Advertisement

The incremental growth in the complexity of ads industry is due to the ease of access to the internet via a wide variety of devices around the world. This gives the advertisers an opportunity to study user preferences and online trends. The insights offered to them through these analysis, translates to actionable items on issues and opportunities such as reducing ad-blindness or optimizing cost-per-action (CPA) and click-through-rates (CTR).

# Recommendations

Many e-retail companies like Amazon, Netflix, Spotify, Best Buy, You Tube among many others use recommender systems to improve a customer's shopping experience. This offers the companies a chance to gather information on customer's preferences, purchases and other browsing patterns which lend insights that can amplify their return on investment.

# Truth and Veracity

In today's digital world, the quantity of fake news is on the rise. Not only does vast majorities of population fall prey to misinformation but it effects businesses negatively. Data Science is being used to ensure data veracity or in other words, verify the truthfulness of data based on both accuracy and context. Companies such Facebook, Twitter, Starbucks, Costco and many others are combating fake news currently with the help of various data science techniques.

# News Aggregation

A news aggregator gathers and clusters stories of the same topic from several leading news websites and also traces the genuine source of a news item and what is the course of the story. It has a special interactive timeline that allows the reader to flip swiftly between headlines, refining their search by country or by specific news sites. Notable examples include Google News, Reddit, Flipboard, Pulse etc.

# Scalability

Scalability refers to an enterprise's ability to handle increased demands. In the corporate environment, a scalable company is one that can maintain or improve profit margins while sales volume increases. Many a times the process is slowed down by human intervention for decisions. For example, the credit operations in banks invest substantial time in assessing the credit worthiness of a client. Client management teams take long time to suggest the right product to the customer/suggest alternatives. The client help desk takes long time to provide the desired info to the client. If these processes can be automated, the business can scale up. Data science helps build systems like recommender systems, Chabot’s etc. to achieve scalability.

# Content Discovery/ Search

Content discovery involves using predictive algorithms to help make content recommendations to users based on how they search. Search engines such as Google, Bing and Yahoo and various other platforms are now using intelligent learning mechanisms to understand user preferences to be able to suggest content that’s most suitable for them.

Few more platforms that use content discovery algorithms are Facebook and YouTube. The content that appears in an individual's Facebook news feed and the videos that appear in the "Recommended for You" section of YouTube user's account, are both altered according to each user's past behavior and personal preferences.

# Intelligent Learning

Intelligent learning has become a part of our day to day lives in various forms. For example, Google Maps uses undesignated location data from various smart devices to predict the flow of traffic in real time. It also utilizes user based reports on incidents that might affect the traffic, like road construction and accidents, to help suggest fastest routes for travel, to users.

Another example would be ride sharing apps like Uber and Ola. They optimize the ride experience by not only minimizing the ride time but also by matching users with other passengers for least amount of detours in shared rides

Other examples of intelligent learning include self-driving cars, smart-email categorization, credit-card fraud detection, etc.

# Personalized Medicine

In many cases, the success of a particular treatment for a patients' condition cannot be predicted beforehand. Thus, many medical practitioners follow a non-optimal trial-and-error approach.

In personalized medicine, a doctor needs to study an individual's genes, environment and lifestyle. This would help tailor treatments for specific medical conditions as opposed to a trial and error approach. This would also enable pharmaceutical researchers to create combination drugs targeting a specific genomic profile which in turn increases safety and efficiency.

Companies that are active in the field of personalized medicine are Roche, Novartis, Johnson & Johnson among others.

=============================================================================

Probability and Statistics using Python

# What is Probability?

By Probability, we measure the likelihood of an event in which we are interested.

# What is the Likelihood of an event?

The Likelihood of an event is the frequency with which the event may occur.

Probabilistic Model

The probabilistic model is a generic structure which describes the random outcomes of an activity.

This model helps assign the likelihood of occurrence to a collection (set) of the random outcomes.

The process of observation of an activity is termed as an **experiment**.

The results of an observation are termed as **outcomes**of the Experiment.

The events for which we cannot calculate the outcomes, those experiments are called **random experiments**.

**Sample Space** is the set of **all the possible outcomes of a random experiment**. It can be of two types which are as follows:

# ****Continuous sample space****

Example: Value of stock for a company can be between $15 and $25 = {15.01, 15.02, …,24.98, 24.99, 25}

# ****Discrete sample space****

Example: Number of people attending the meeting can be between 10 and 20 = {10, 11, 12,…,20}

A Set is a collection of distinct objects/elements. It is a collection of all possible outcomes.

The Sets help translate the problems to mathematical representations using set notations.

* If a set contains finite no. of elements (x1,x2,x3) it can be represented as S={x1,x2,x3}
  + Example: Set of all possible outcomes of a die roll S={1,2,3,4,5,6 }
* If a set contains infinite no. of elements (x1,x2,x3,…..) it can be represented as S={x1,x2,x3,…..}
  + Example: Set of all possible positive real numbers S={0,0.01,1,1.5,2,……}

**Intersection**: The intersection of two sets A and B is represented as **A ∩ B**

**Union**: The union of two sets A and B is represented as **A ∪ B**

**Complement**: The complement of Set A is represented as **A**

**Difference**: The difference of Set A and B is represented as  **A — B**

**Symmetric Difference**: The symmetric difference of Set A and B is represented as **A Δ B**

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================

=============================================================================