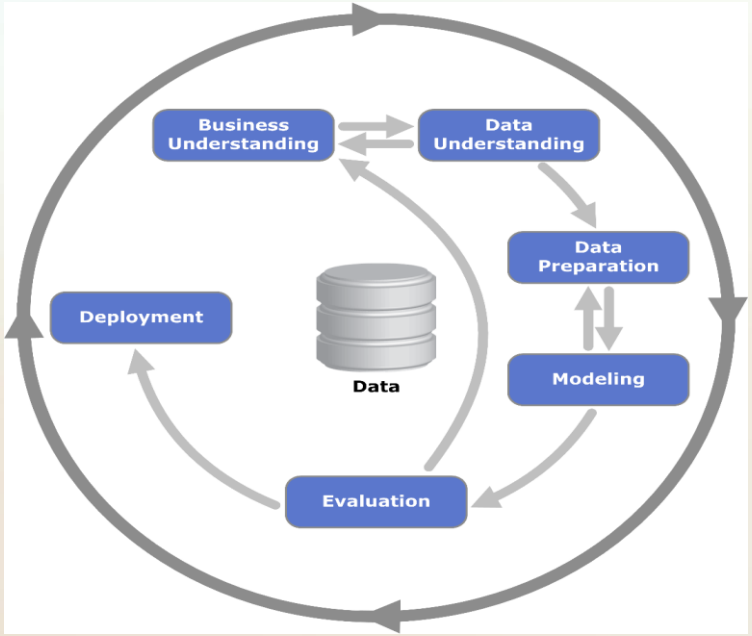
# Artificial Intelligence

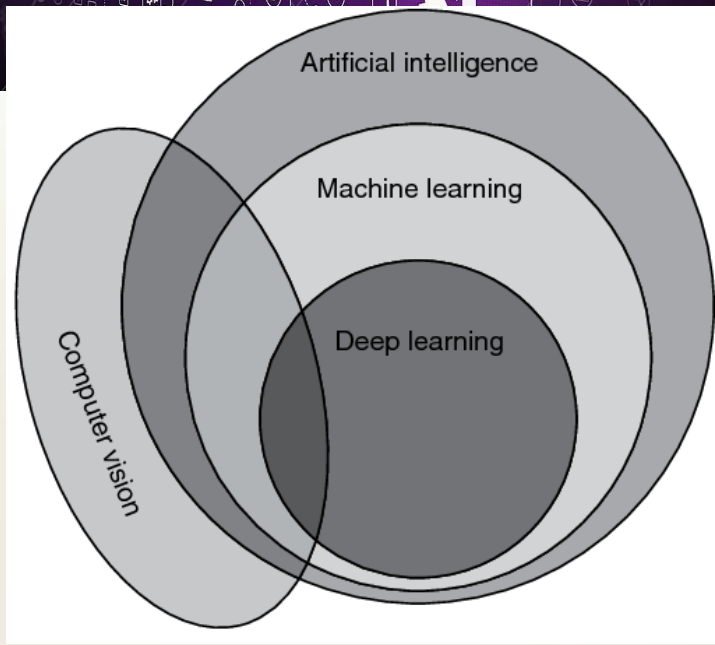
Anita Budhiraja

Scientist ‘D’

# Agenda

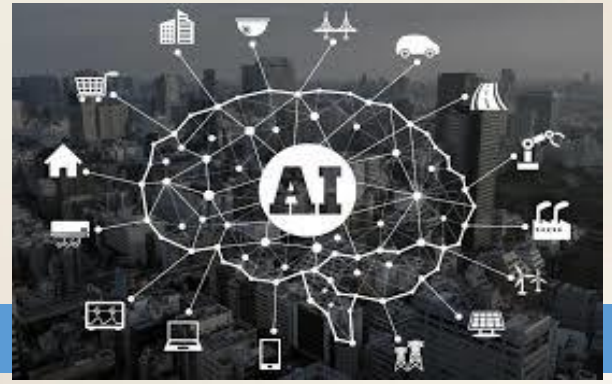
- What is AI ?
- Nomenclature in AI
- Business Understanding of AI
- CRISP - DM Framework

# What is artificial intelligence?

- Artificial intelligence can be loosely interpreted to mean incorporating human intelligence to machines.
- Artificial intelligence (**AI**) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.
- Whenever a machine completes tasks based on a set of stipulated rules that solve problems (algorithms), then such an “intelligent” behavior is called artificial intelligence.



# What is artificial intelligence?

- Artificial intelligence (AI) makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks.
- Most AI examples that you hear about today – from chess-playing computers to self-driving cars – rely heavily on deep learning and natural language processing.
- Using these technologies, computers can be trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data.
- Machine learning is also a core part of AI.
- Learning without any kind of supervision requires an ability to identify patterns in streams of inputs, whereas learning with adequate supervision involves classification and numerical regressions.
- Mathematical analysis of machine learning algorithms and their performance is a well-defined branch of theoretical computer science often referred to as computational learning theory.
- Artificial intelligence enables a computer system to be trained and apply the gained knowledge to new inputs. This ability rests upon math and algorithms and is applicable only to the tasks that the system has been trained to perform.

# Artificial intelligence and its Future

- **Artificial intelligence** is impacting the **future** of virtually every industry and every human being.
- **Artificial intelligence** has acted as the main driver of emerging technologies like big data, robotics and IoT, and it will continue to act as a technological innovator for the foreseeable **future**.
- Machines understand verbal commands, distinguish pictures, drive cars and play games better than we do. How much longer can it be before they walk among us?

# Business Understanding of AI

- AI has become a trend that cannot be disregarded: hardly any report or survey doesn't mention its growing importance.
- Gartner names AI a top priority for Businesses.
- Accenture Technology Vision 2019 Report cites that 89% of businesses either have already adopted or are experimenting with AI.
- And PWC 2019 AI Predictions Survey of 1,000 executives reveals that they expect the following value from AI investments: increased revenue and profits (48%), better customer experience (46%) and improved decision-making (40%).

# Nomenclature of AI

| Conventional Nomenclature | AI Nomenclature |
|---|---|
| Data | Pattern / Knowledge |
| Program / Algorithm | Reasoning / Planning / Problem Solving |
| GIGO ( Garbage In Garbage Out ) | Garbage In Useful Out |
| Deterministic | Deterministic / Probabilistic |
| Hard Wired | Evolution / Mutation / Adaptation |
| Serial | Parallel |

# Nomenclature of Artificial Intelligence

- **Big Data** refers to extremely large data sets that are computationally analyzed to reveal patterns, trends and unique associations related to human interactions and behaviors.
- **Collaborative systems** related to those models and algorithms for development of autonomous systems that can work collaboratively with other systems as well as with human entities.
- **Computer vision** is the most prominent form of machine perception currently available. This sub-area of AI has been significantly transformed by “deep learning.” Many computers are now able to perform some vision tasks better than humans. At present, significant research is underway on the further advancement of “computer vision” in the areas of automatic image and video captioning.

# Nomenclature of Artificial Intelligence

- **Deep learning** is a form of learning that has facilitated object recognition via images and video, along with activity recognition. Research is underway into other areas of perception, including audio, speech and natural language processing.
- The **Internet of Things** (IoT) encompasses concepts related to an array of devices, many of which are used in our everyday lives. Things like your appliances, home, office building, cameras and vehicles may all be (or soon will be) connected via the internet to permit the collection and sharing of information for intelligent purposes.

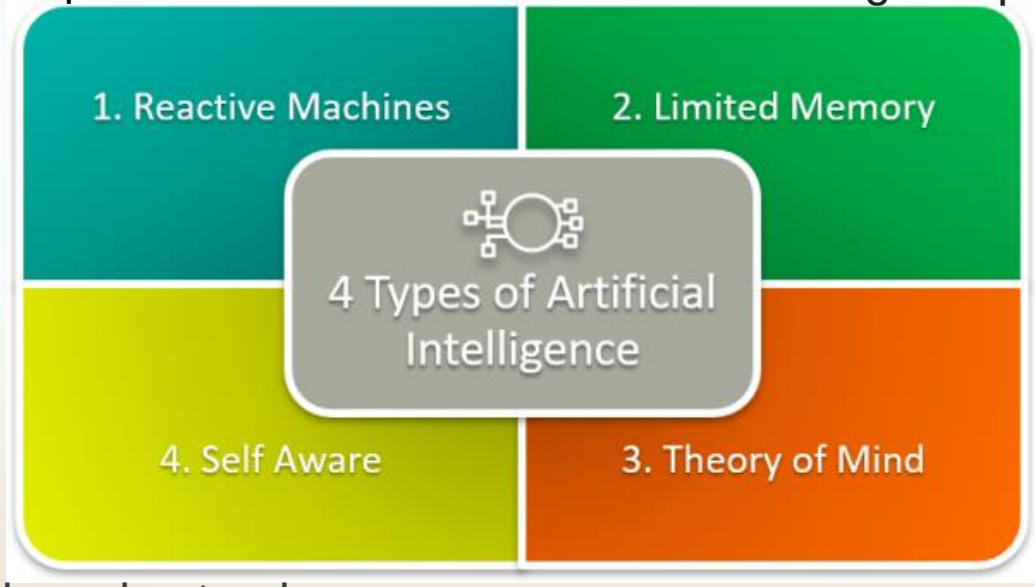
# Nomenclature of Artificial Intelligence

- **Natural Language Processing** is sometimes referred to, or coupled with, speech recognition. This form of AI is quickly developing for widely spoken languages associated with large data sets. At the same time, developers are refining NLP systems so that they can interact directly with people through simple dialog rather than specifically stylized requests. As part of this emerging form of AI, multi-lingual forms of NLP are being designed so that systems can interact with anyone speaking any language on the planet.
- **Reinforcement learning** shifts the focus of machine learning from pattern recognition to experience-driven decision-making. This technology will bring AI to the real world, and in doing so, impact millions of lives. Strides continue to be made in the practical implementation of this form of learning as part of a broadening of AI real-world environments.
- **Robotics** is the process of developing and training robots to interact with the world in predictable ways. This includes the facilitation and manipulation of objects in interactive environments and with people. Substantial advances in robotics have been made in the past few years based upon the successes of other AI related technologies, including computer vision and other forms of machine perception

# Types of Artificial Intelligence

**Reactive machines** are the simplest level of robot. They cannot create **memories** or use information learnt to influence future decisions – they are only able to react to presently existing situations. Eg. Deep Blue, IBM's chess-playing supercomputer and Google AlphaGo

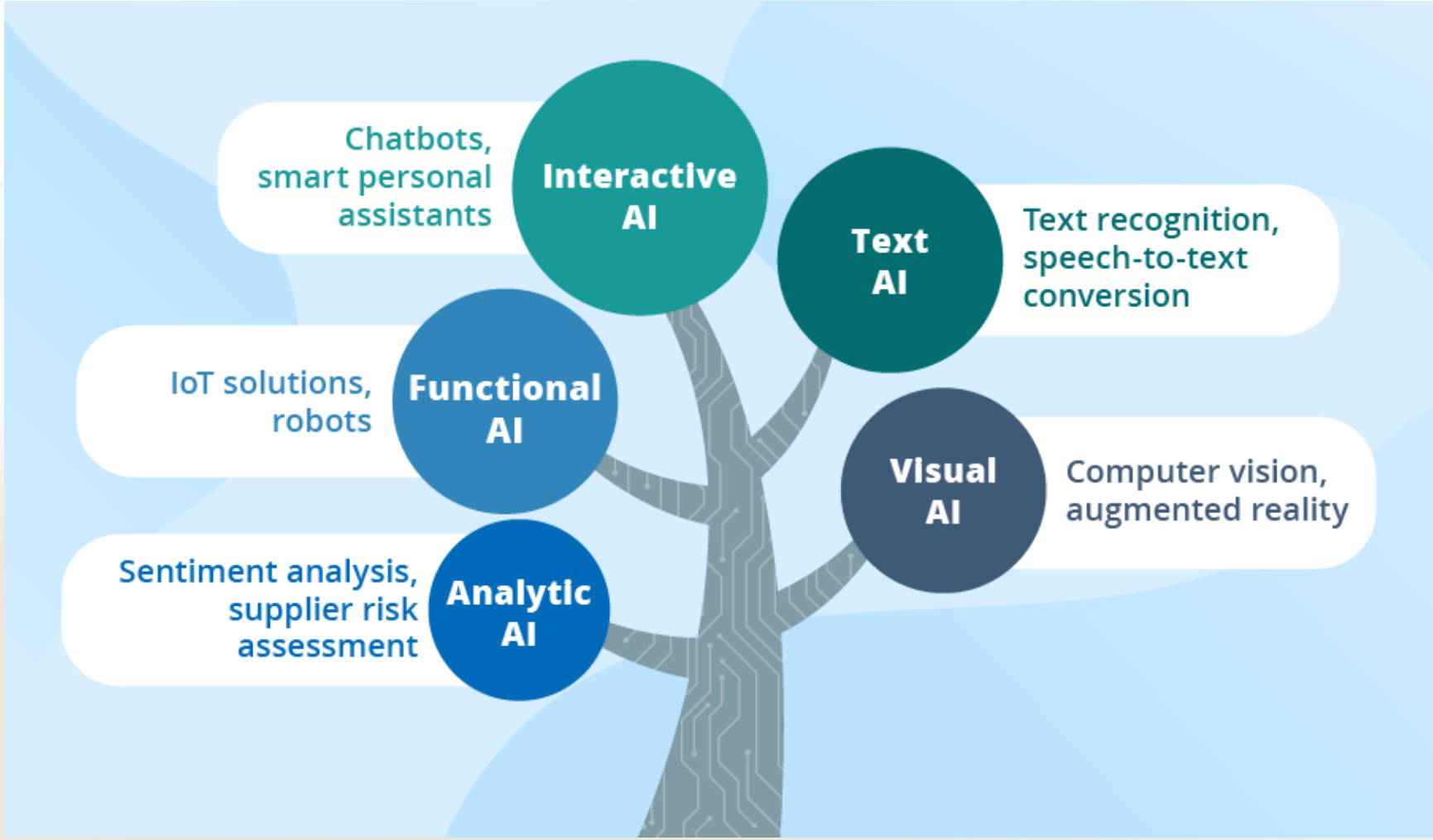
**Limited memory** types refer to an **A.I.**'s ability to store previous data and/or predictions, using that data to make better predictions. ... Every machine learning model requires **limited memory** to be created, but the model can get deployed as a reactive machine type.




AI researchers will have to not only understand consciousness, but build machines that have it.While we are probably far from creating machines that are self-aware, we should focus our efforts toward understanding memory, learning and the ability to base decisions on past experiences.

Theory of mind is the important divide between th machines we have and the machines we will build in th future. AI systems will have to adjust according to ou thoughts, feelings and expectations. Autonomous cars

# AI That brings Value to Business



# AI types and Business Use Cases

- **Analytic AI**

Powered with machine learning (including its most advanced deep learning techniques), analytic AI scans tons of data for dependencies and patterns to ultimately produce recommendations or provide a business with insights, thus contributing to data-driven decision-making.

***Sentiment analysis , inventory optimization and demand forecasting*** are just a few examples of analytic AI in action.

# AI types and Business Use Cases

- **Functional AI**

allows businesses to automate communication without compromising on interactivity. To envisage this type of AI, think of ***chatbots*** and ***smart personal assistants*** whose abilities can vary from answering pre-built questions to understanding the conversation context.

# AI types and Business Use Cases

- **Text AI**

Businesses that use text AI can enjoy ***text recognition***, ***speech-to-text conversion, machine translation***, and ***content generation*** capabilities. Even if a company is not Google or Amazon, or any other giant company that provides text AI as a service, it can still take advantage of this AI type. For example, the company can use text AI to power an internal corporate knowledge base.

AI-powered one can find the document containing the most relevant answer even if the document doesn't have full keywords. This is possible thanks to semantic search and natural language processing, which allow AI to build semantic maps and recognize synonyms to understand the context of the user's question.

# AI types and Business Use Cases

- **Visual AI**

With visual AI, businesses can identify, recognize, classify and sort objects or convert images and videos into insights. A computer system that helps an insurer to estimate damage based on damaged car photos or a machine that grades apples based on their color and size are the examples of visual AI. This type of AI covers ***computer vision*** or ***augmented reality*** fields.

face recognition solution that we developed to help a retailer enhance and personalize their customer service; or about an application for automated inspections that allowed a manufacturer to immediately control the quality of the produced details.

# Sub Fields of Artificial Intelligence

AI now consists of many sub-fields, using a variety of techniques, such as:

- Neural Networks – e.g. brain modelling, time series prediction, classification
- Evolutionary Computation – e.g. genetic algorithms, genetic programming
- Vision – e.g. object recognition, image understanding
- Robotics – e.g. intelligent control, autonomous exploration
- Expert Systems – e.g. decision support systems, teaching systems
- Speech Processing– e.g. speech recognition and production
- Natural Language Processing – e.g. machine translation
- Planning – e.g. scheduling, game playing
- Machine Learning – e.g. decision tree learning, version space learning

# What is Speech Recognition?

- Speech recognition is a technology that enables a computer to identify and interpret words and phrases in spoken language and convert them into texts by computers.
- It is a subfield of computational linguistics that deals with technologies to allow spoken input into systems. It allows user to control their computers with their voice.
- Speech recognition is quickly becoming a mainstay in human-computer interaction. Today, speech recognition tools are used in different types of dictation tasks, such as composing a text message, playing music through a home-connected device, or text-to-speech applications with virtual assistants.
- It has become ubiquitous today as a great way to interact with technology, significantly bridging the gap between human and computer interaction and making it more natural. NLP is one such technology that simplifies speech recognition processes using natural language.

# What is Natural Language Processing?

- NLP is a branch of artificial intelligence that investigates the use of computers to process or to understand human languages for the purpose of performing useful tasks.
- It encompasses approaches to help machines understand, interpret, and generate human language.
- It simply deals with the interaction between humans and computers using a natural language such as English.
- It combines computational linguistics, computing science, cognitive science and artificial intelligence to perform tasks such as translation, automatic summarization, topic segmentation, relationship extraction, information retrieval, machine translation, and speech recognition.

# NLP and Speech Recognition

NLP and Speech Recognition are sometimes used in conjuction in applications such as voice assistants, ASR engines and speech analytical tools.

**Siri** is Apple's personal assistant for iOS, macOS, tvOS and watchOS devices that uses voice recognition and is powered by **artificial intelligence** (**AI**).

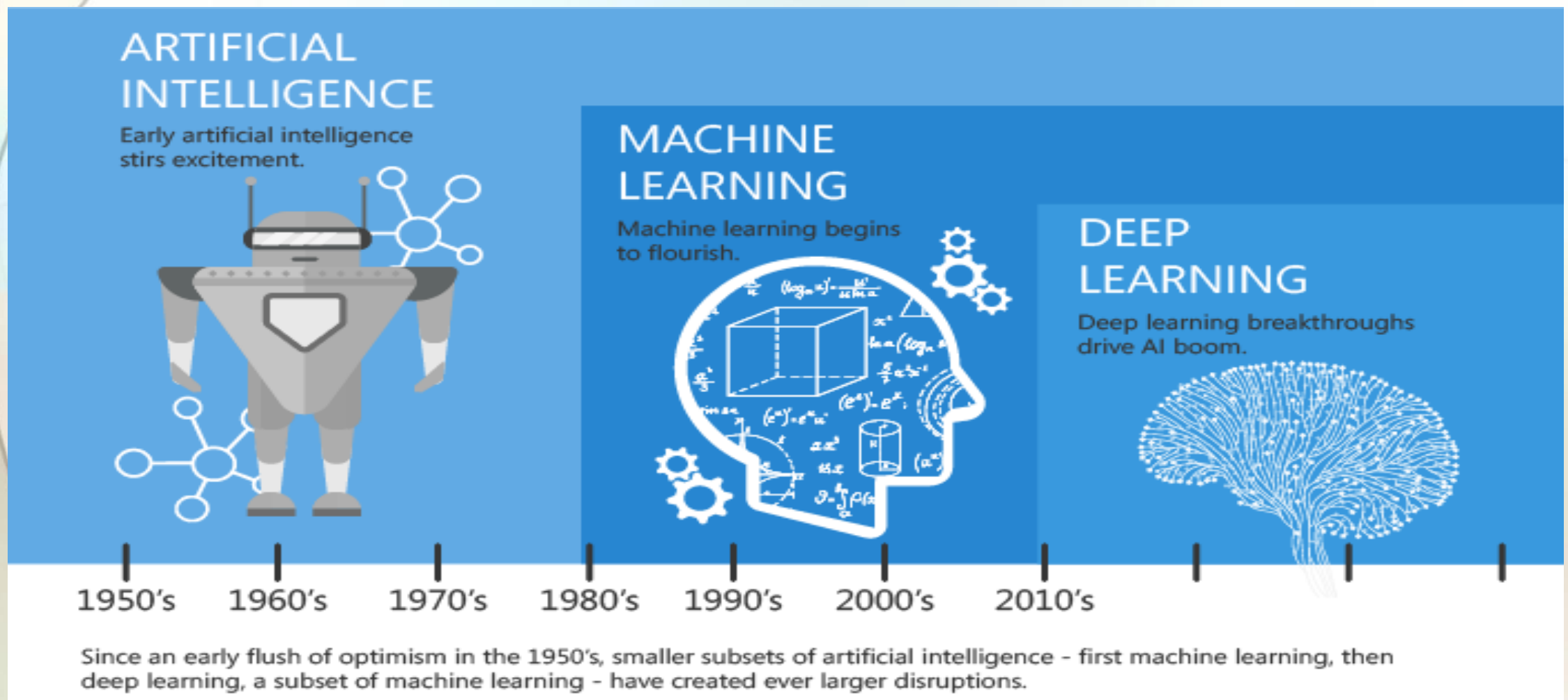
Amazon **Alexa**, also known simply as **Alexa**, is a virtual assistant **AI** technology developed by Amazon,
Google **Voice** Search

As per the Gartner, 30% of interactions with the technology are performed through conversations.

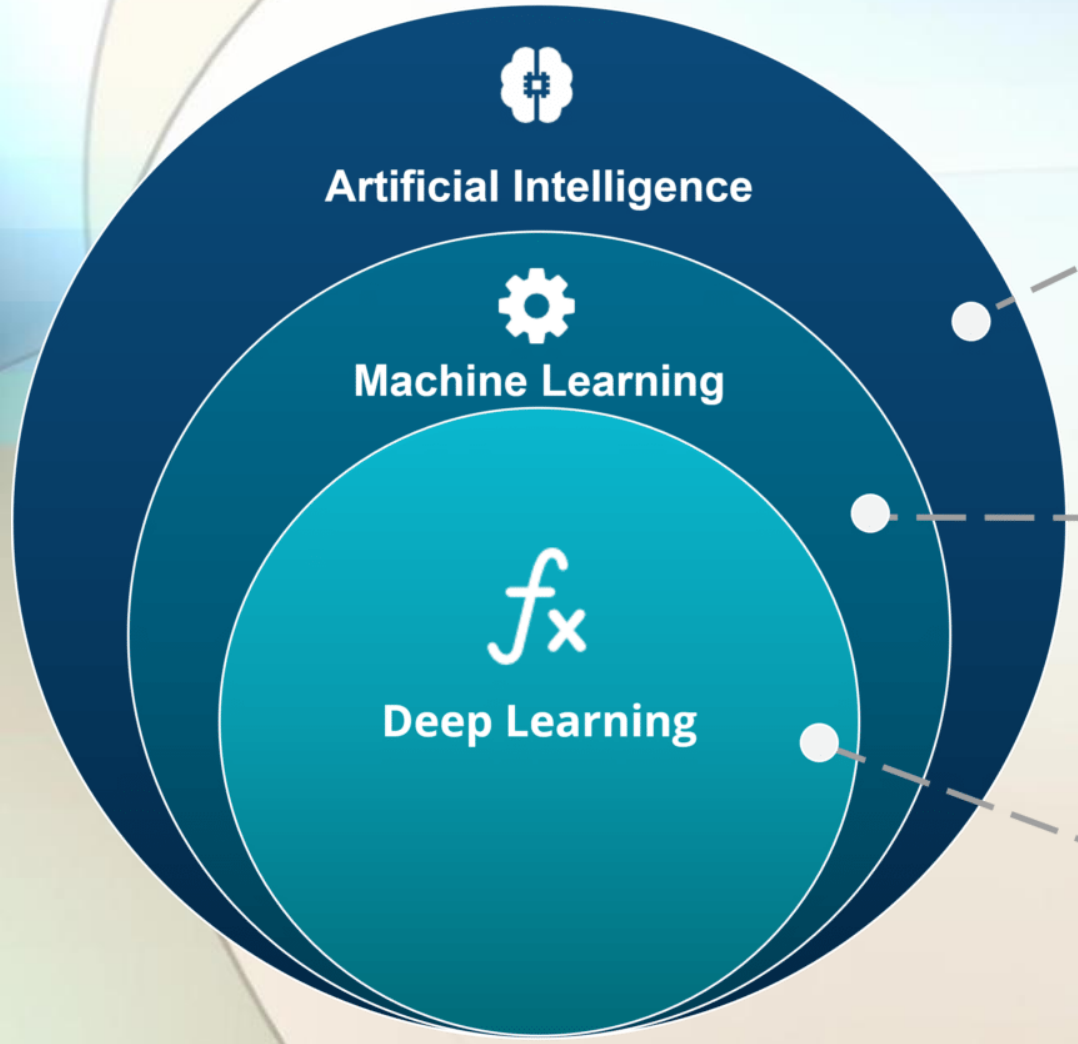
# NLP and Speech Recognition

- Dragon Professional is best as an overall speech recognition software. Dragon Anywhere and Siri are best for iOS users. Cortana is best for Windows users.
- Google Now is best for Android Mobile devices. For the dictation on Google Docs, Google Docs Voice Typing is the best option. For creating Chatbot, Amazon Lex is the best option.

# Evolution of AI, ML and DL



# Relationship between AI,ML and DL

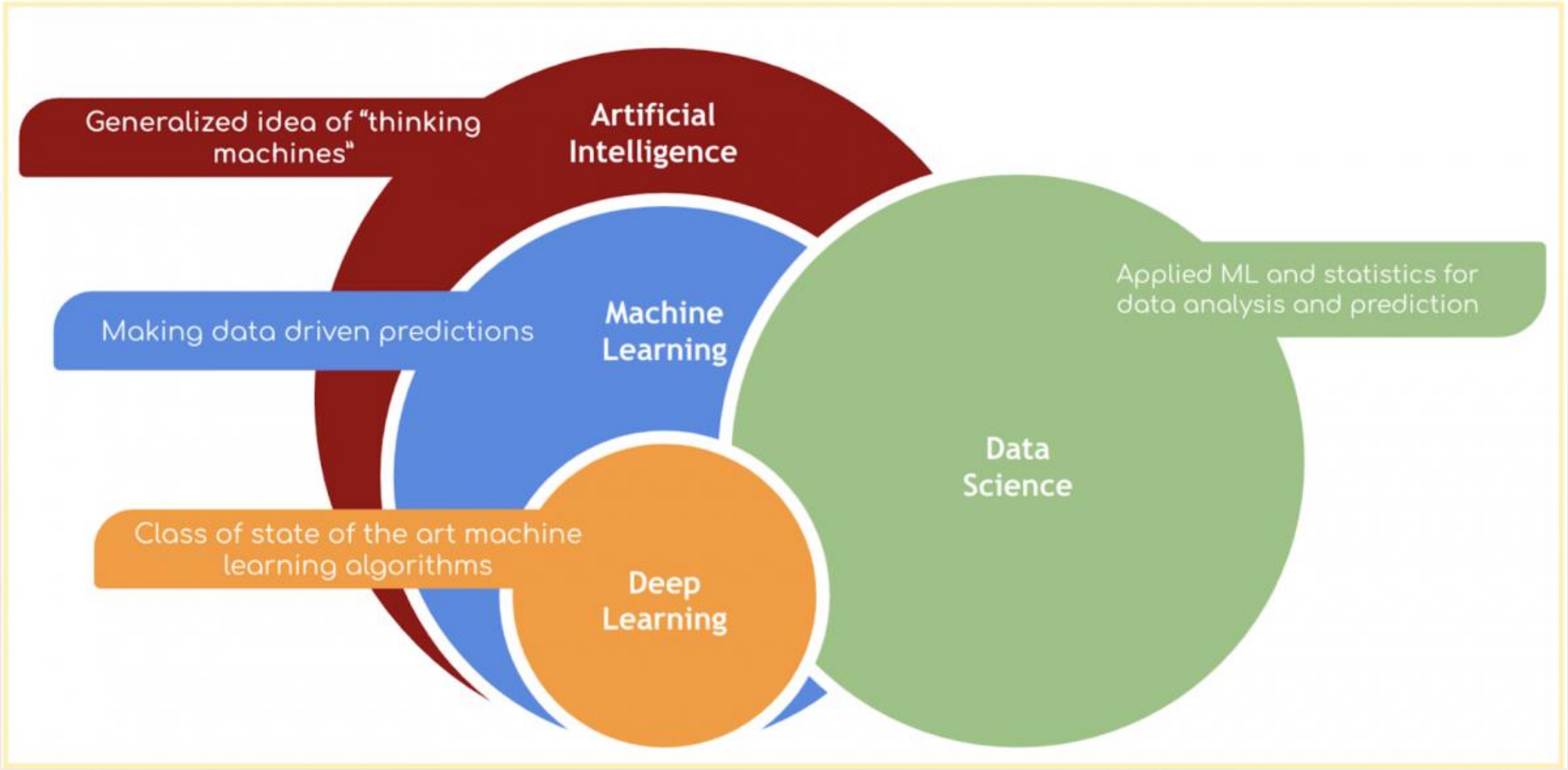



**ARTIFICIAL INTELLIGENCE**
A technique which enables machines to mimic human behaviour

**MACHINE LEARNING**
Subset of AI technique which use statistical methods to enable machines to improve with experience

**DEEP LEARNING**
Subset of ML which make the computation of multi-layer neural network feasible

# AI, ML, DL and Data Science



# Machine Learning

Its not a magic pill

"~~Can machines think~~?"

*is replaced with the question*

"Can machines do what we (as thinking entities) can do?"

Alan Turing

- "A computer program is said to learn

  from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

  *Tom M. Mitchell*

# What is Machine Learning?

- As the name suggests, machine learning can be loosely interpreted to mean empowering computer systems with the ability to “learn”.
- The intention of ML is to enable machines to learn by themselves using the provided data and make accurate predictions.
- ML is a subset of **artificial intelligence**; in fact, it's simply a technique for realizing AI.
- It is a method of training algorithms such that they can **learn** how to make decisions.

# What is Machine Learning?

- For example, here is a table that identifies the type of fruit based on its characteristics:

| Weight (grams) | Texture | Type of Fruit |
|---|---|---|
| 155 | Rough | Orange |
| 180 | Rough | Orange |
| 135 | Smooth | Apple |
| 110 | Smooth | Apple |
| 120 | Smooth | ? |

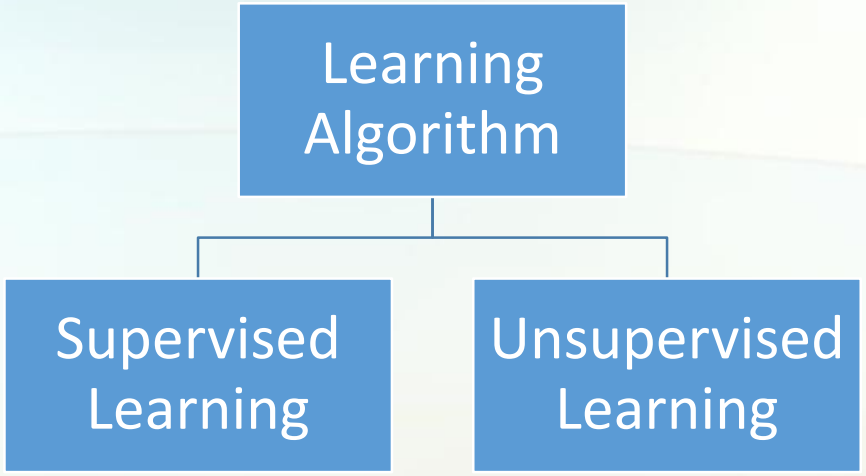
Problem –

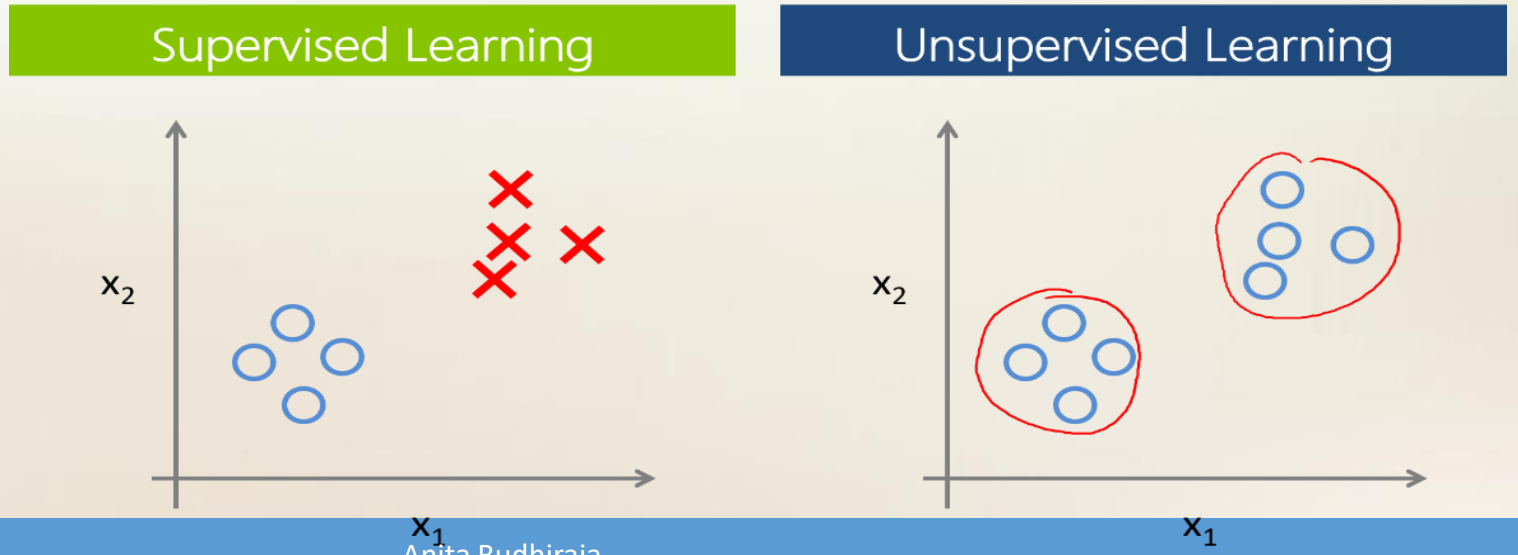
Distinguish Apple from Orange

- The last row gives only the weight and texture, without the type of fruit.
- A machine learning algorithm can be developed to try to identify whether the fruit is an orange or an apple.
- After the algorithm is fed with the training data, it will learn the differing characteristics between an orange and an apple.
- Therefore, if provided with data of weight and texture, it can predict accurately the type of fruit with those characteristics.

# Major Classes of Learning Algorithms

Better to think of machine learning as a *means of building models of Data*



# Why Machine Learning is needed?

- Rule Based Systems become complex and are static as we make use cases / rules for complex system.
- Machine Learning is needed in cases where humans cannot directly write a program to handle each and every case.
- So it's better to have a machine (~~rather than human~~) that learns from a large training set.

*according to the definition earlier:*

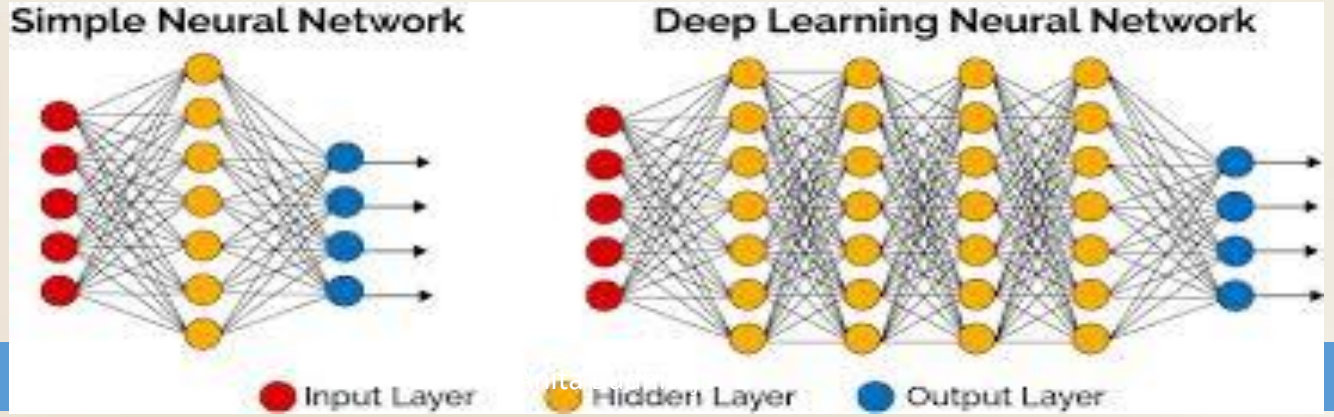
- Task (T): recognizing and classifying handwritten words within images
- Performance measure (P): percent of words correctly classified
- Training experience (E): a database of handwritten words with given classifications

# What is deep learning?

- Deep learning is a subset of ML.
- In fact, it's simply a technique for realizing machine learning. In other words, DL is the next evolution of machine learning.
- DL algorithms are roughly inspired by the information processing patterns found in the human brain. Just like we use our brains to identify patterns and classify various types of information, deep learning algorithms can be taught to accomplish the same tasks for machines.

# What is deep learning?

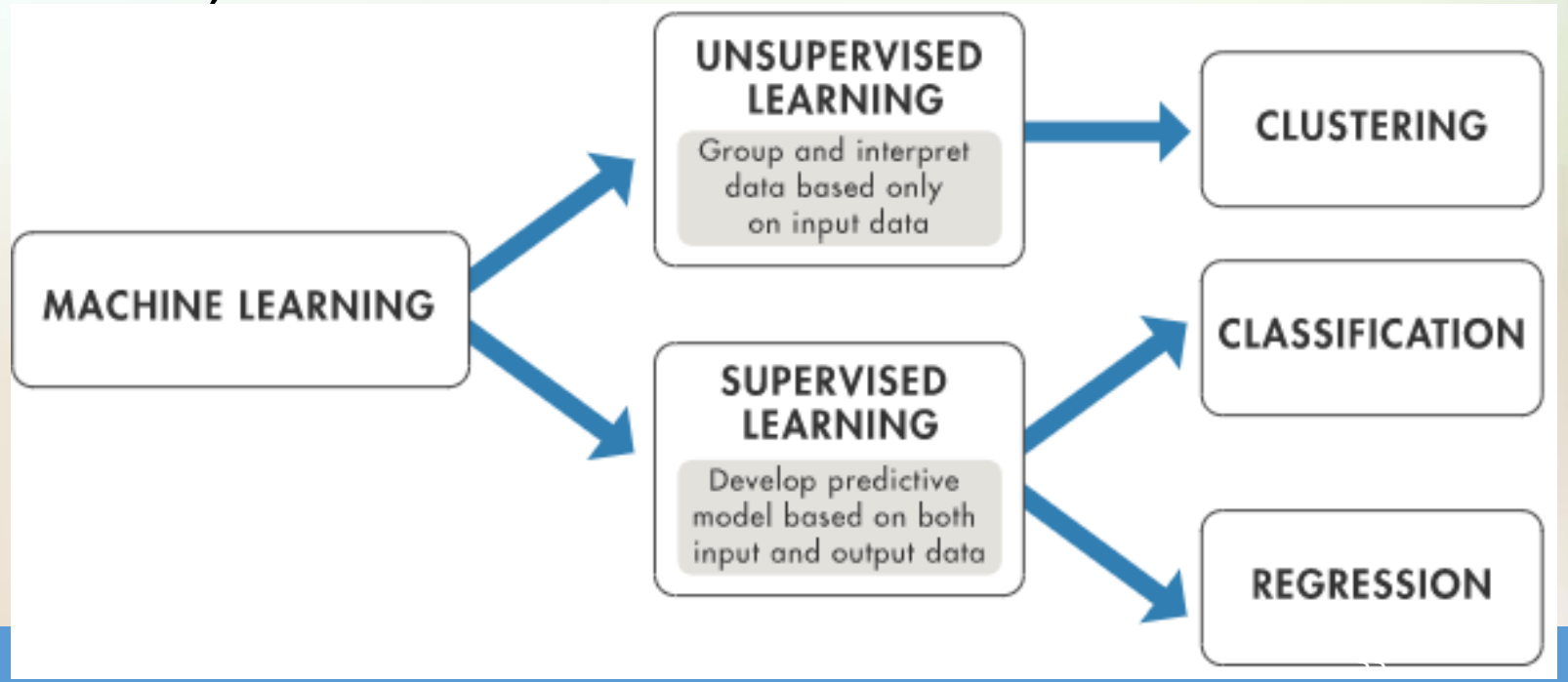
- For example, while DL can automatically discover the features to be used for classification, ML requires these features to be provided manually.
- Furthermore, in contrast to ML, DL needs high-end machines and considerably big amounts of training data to deliver accurate results.
- For example, artificial neural networks (ANNs) are a type of algorithms that aim to imitate the way our brains make decisions.



# Categories of Learning

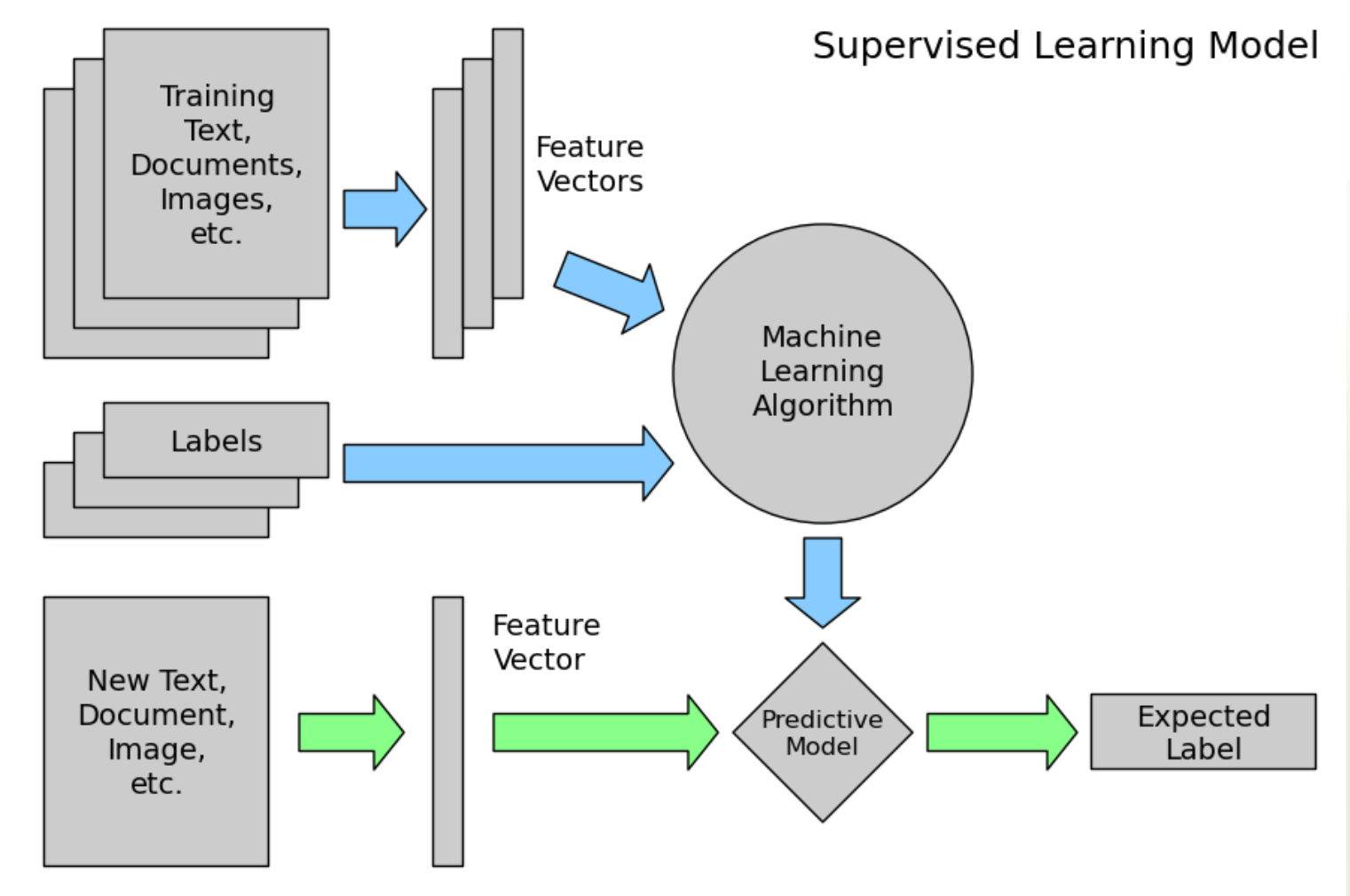
learning from exemplars

- The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- The training process continues until the model achieves a desired level of accuracy on the training data. once this model is determined, it can be used to apply labels to new, unknown data.



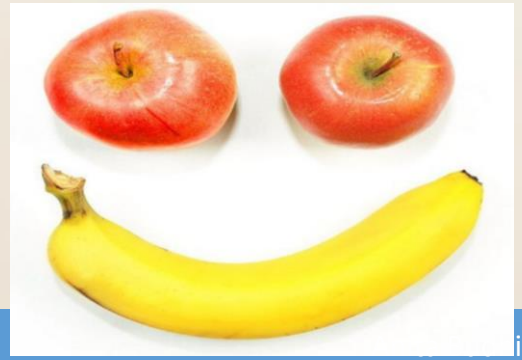

Source : wikipedia

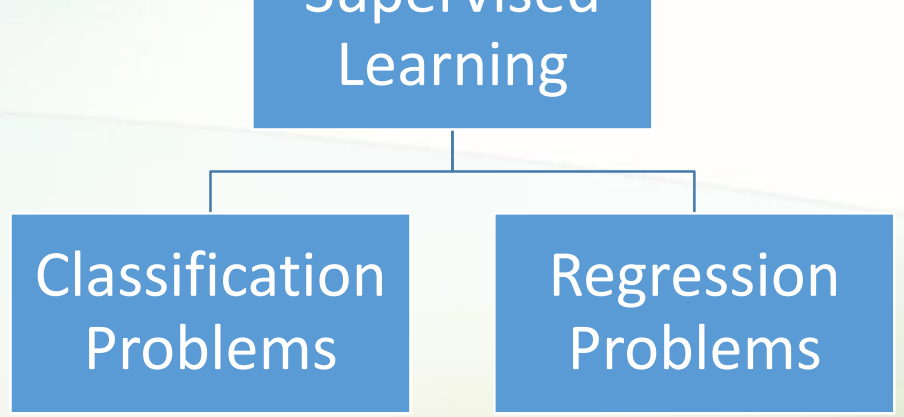
# Supervised Learning

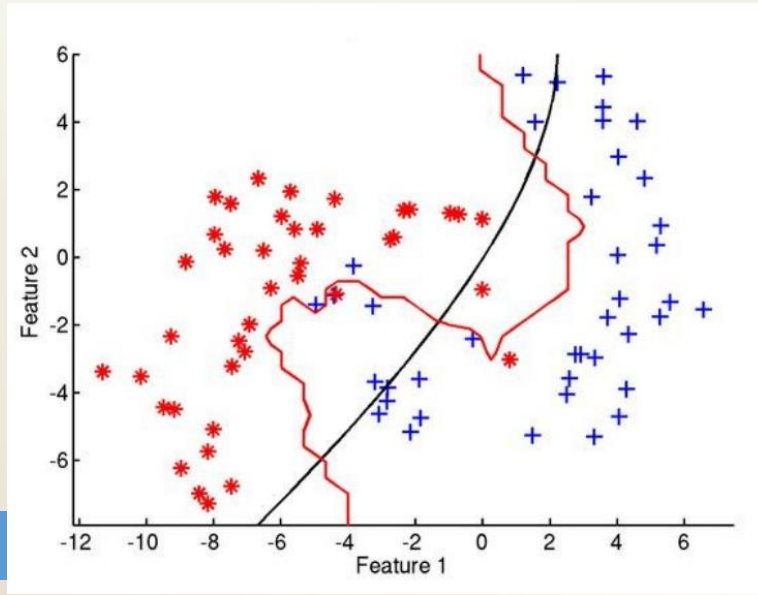


# Supervised Learning: Classification Problems

- Consists of taking input vectors and deciding which of the N classes they belong to, based on training from exemplars of each class
- Find 'decision boundaries' that can be used to separate out the different classes.
- It is to decide which class the current input belongs to.



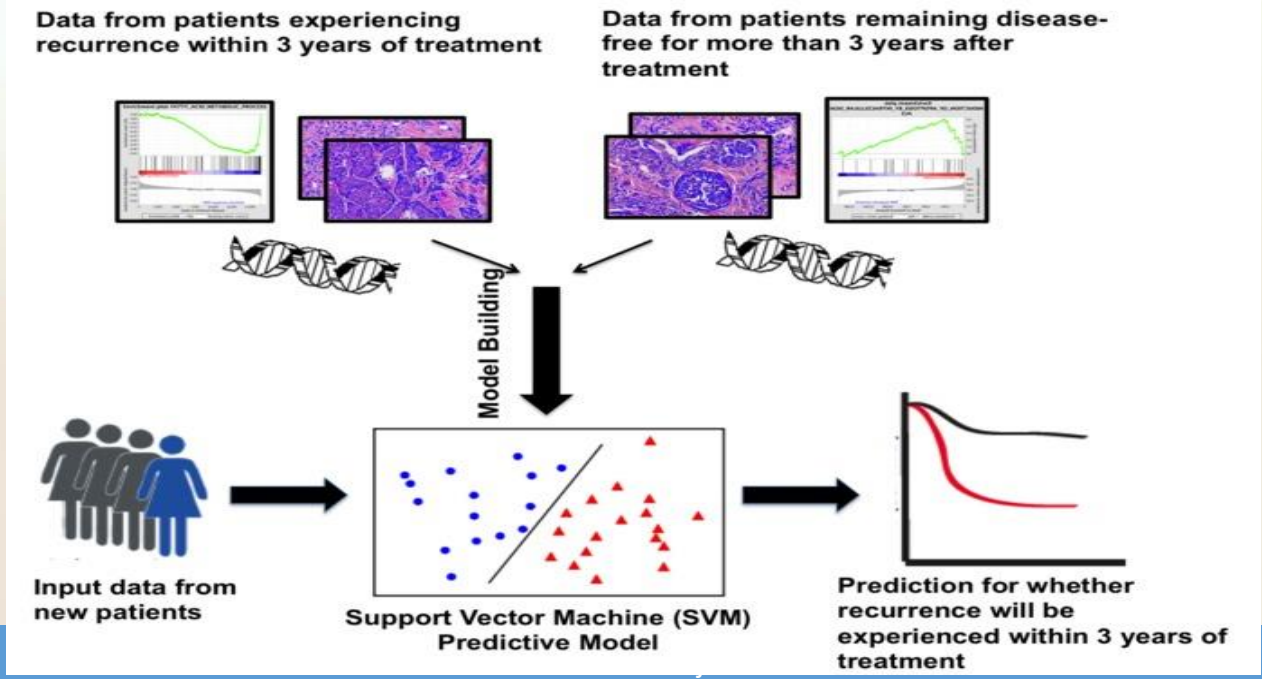


# Supervised Learning: Classification

- Classification is finding the category of the input variable, ie mapping input variables into discrete categories.
- Answer to classification problem would be, *whether this or that*, like, yes or no, 0 or 1, true or false.

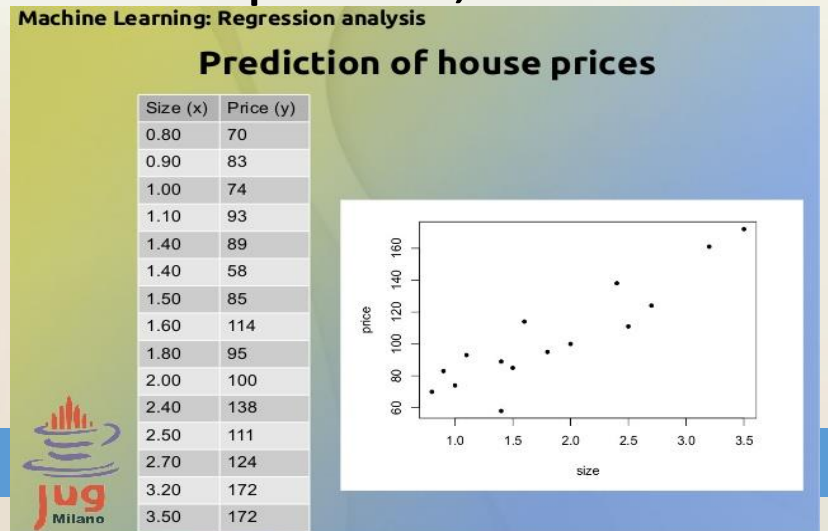
Another example for breast cancer prognosis



# Supervised Learning: Regression

***Regression***

- In regression problem, we are trying to ***predict*** results within a continuous output, meaning that we are trying to map input variables to some continuous function.
- Therefore regression is for predicting a quantity.
- Eg : Given data about the size of houses on the real estate market, try to predict their price.
- Eg: Given a picture of a person, we have to *predict* their age or gender.



# Unsupervised learning

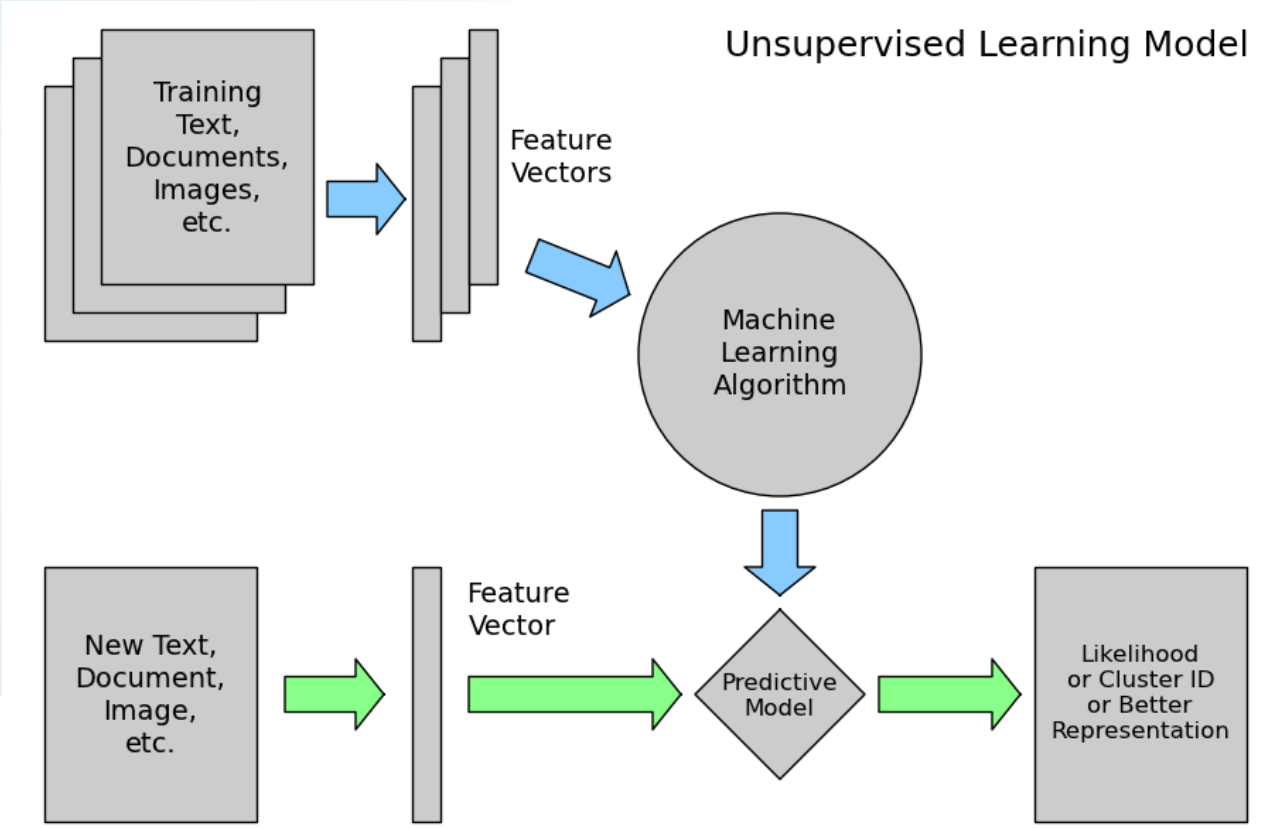
" letting the dataset speak for itself "

- No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or

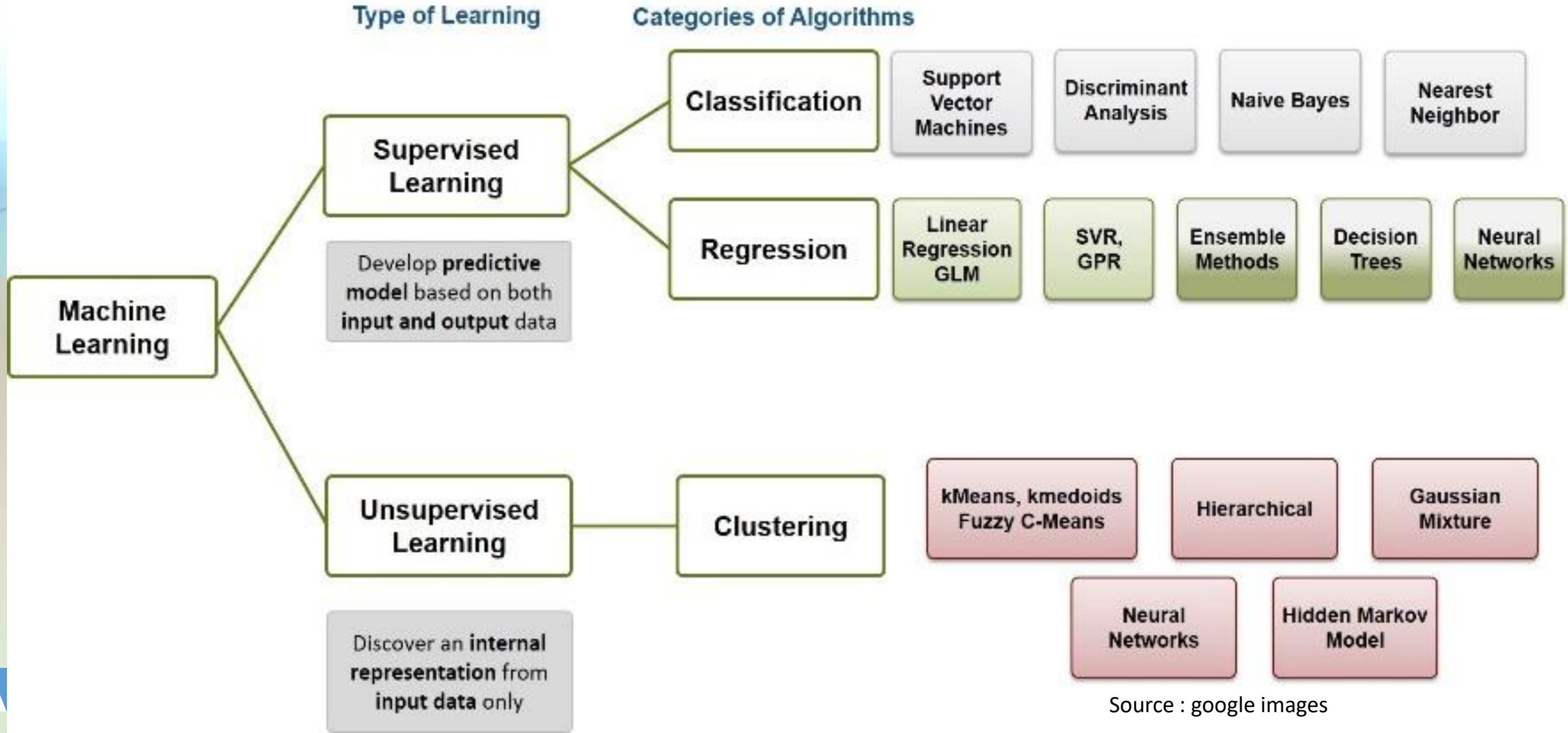
  a means towards an end (feature learning).
- It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention.
- distinct groups of data, while
- Dimensionality reduction algorithms is the.

# Unsupervised learning - Clustering

- The aim of unsupervised learning is to find clusters of similar inputs in the data without being explicitly told that some datapoints belong to one class and the other in other classes.
- The algorithm has to discover this similarity by itself



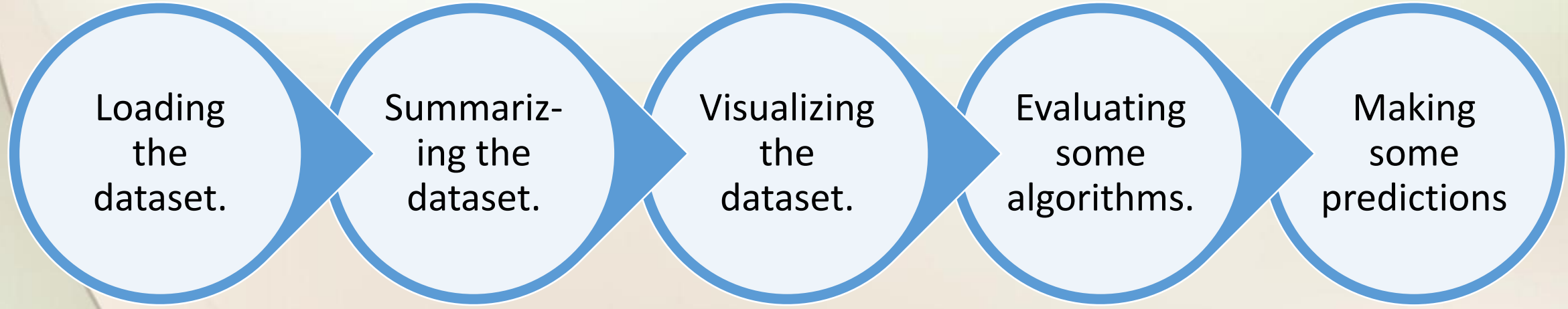

# Category of Algorithms




Source : google images

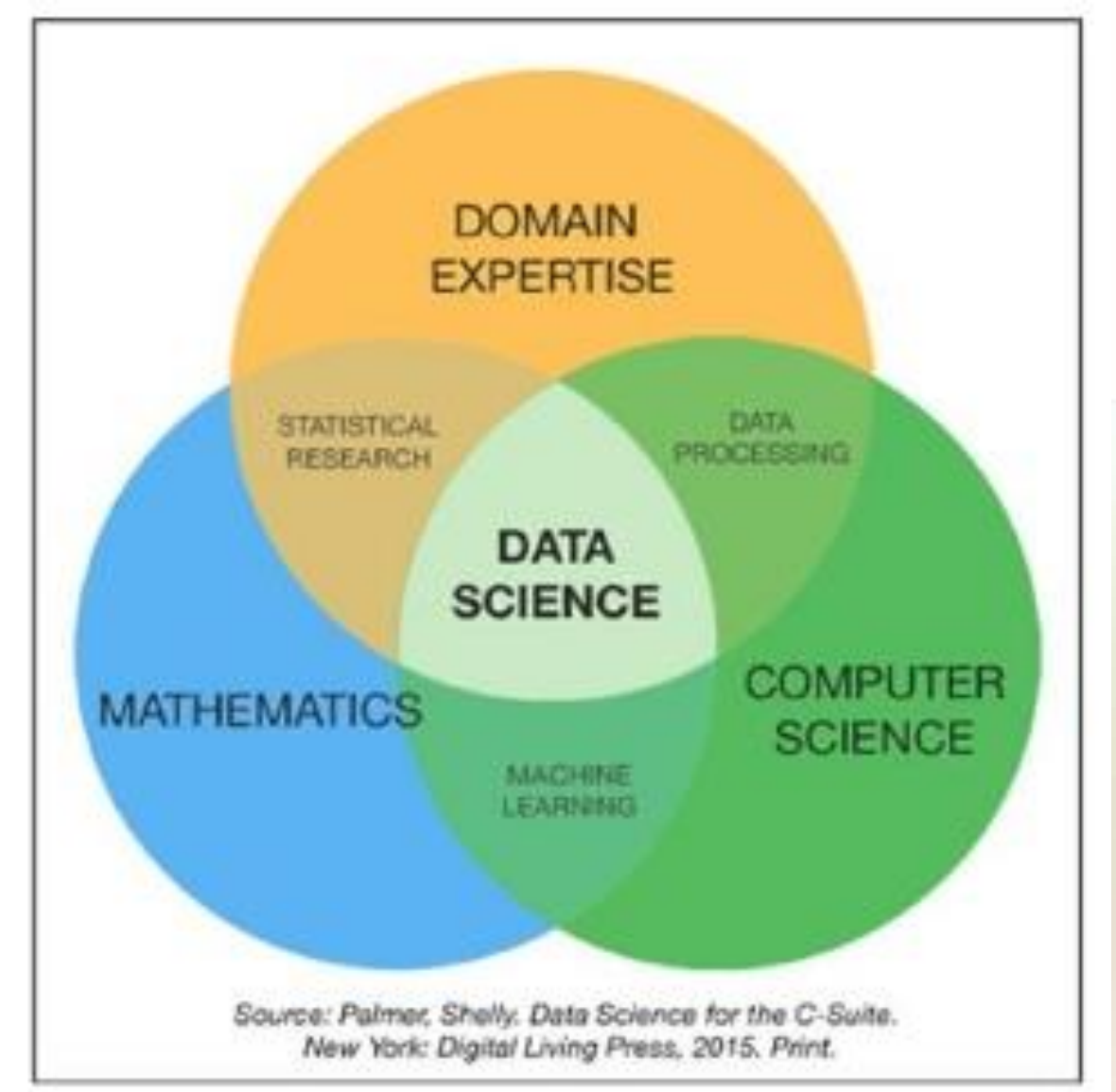
# Layman's view of Machine Learning

- Loading the dataset.
- Summarizing the dataset.
- Visualizing the dataset.
- Evaluating some algorithms.
- Making some predictions.



# What is Data Science

- Data science is a broad field of study pertaining to data systems and processes, aimed at maintaining data sets and deriving meaning out of them.
- The information extracted through data science applications are used to guide business processes and reach organisational goals.
- One of the domains that data science influences directly is business intelligence.


DOMAIN EXPERTISE
STATISTICAL RESEARCH
DATA PROCESSING
DATA SCIENCE
MATHEMATICS
COMPUTER SCIENCE
MACHINE LEARNING
Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

# Difference between AI, ML and Data Science

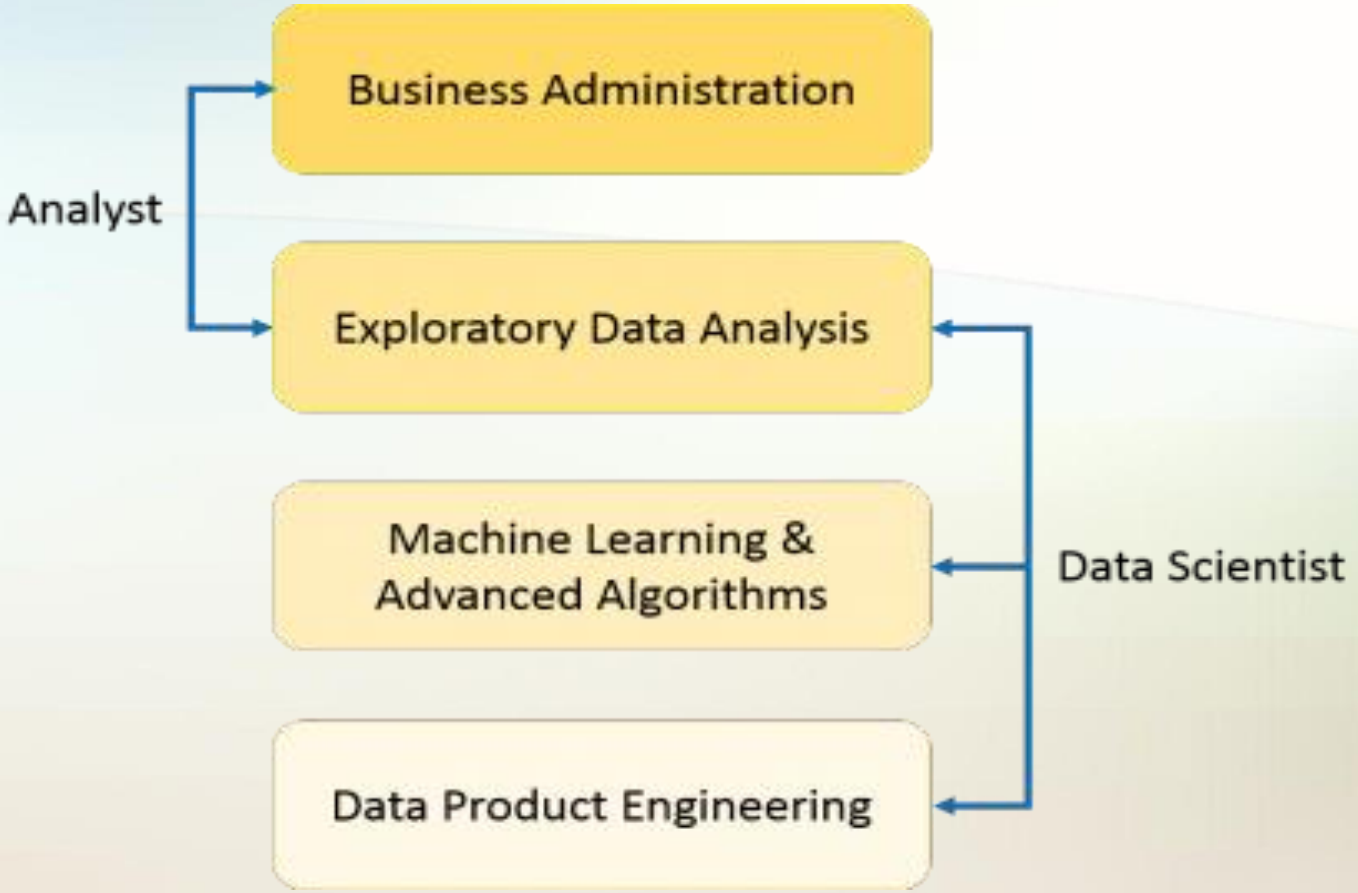
| Artificial Intelligence | Machine Learning | Data Science |
|---|---|---|
| • Includes Machine Learning. | • Subset of Artificial Intelligence. | • Includes various Data Operations. |
| • Artificial Intelligence combines large amounts of data through iterative processing and intelligent algorithms to help computers learn automatically. | • Machine Learning uses efficient programs that can use data without being explicitly told to do so. | • Data Science works by sourcing, cleaning, and processing data to extract meaning out of it for analytical purposes. |
| • Artificial Intelligence uses logic and decision trees. | • Machine Learning uses statistical models. | • Data Science deals with structured and unstructured data. |
| • Chatbots, and Voice assistants are popular applications of AI. | • Recommendation Systems such as Spotify, and Facial Recognition are popular examples. | • Fraud Detection and Healthcare analysis are popular examples of Data Science. |

# Why We Need Data Science?

- How about if you could understand the precise requirements of your customers from the existing data like the customer's past browsing history, purchase history, age and income.
- Data Science can be used in predictive analytics. Let's take weather forecasting as an example. Data from ships, aircrafts, radars, satellites can be collected and analyzed to build models.
- These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities

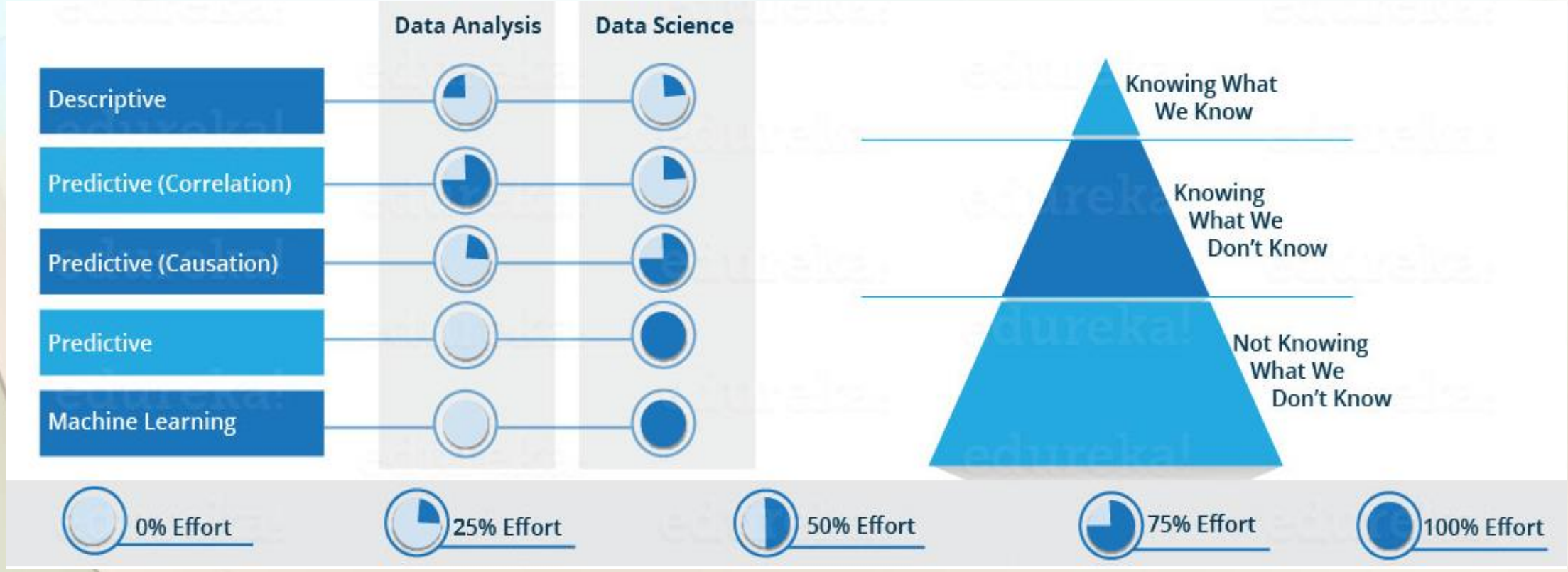
# Why We Need Data Science?

- Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.
- Data Scientist not only does the exploratory analysis to discover insights from it, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future.



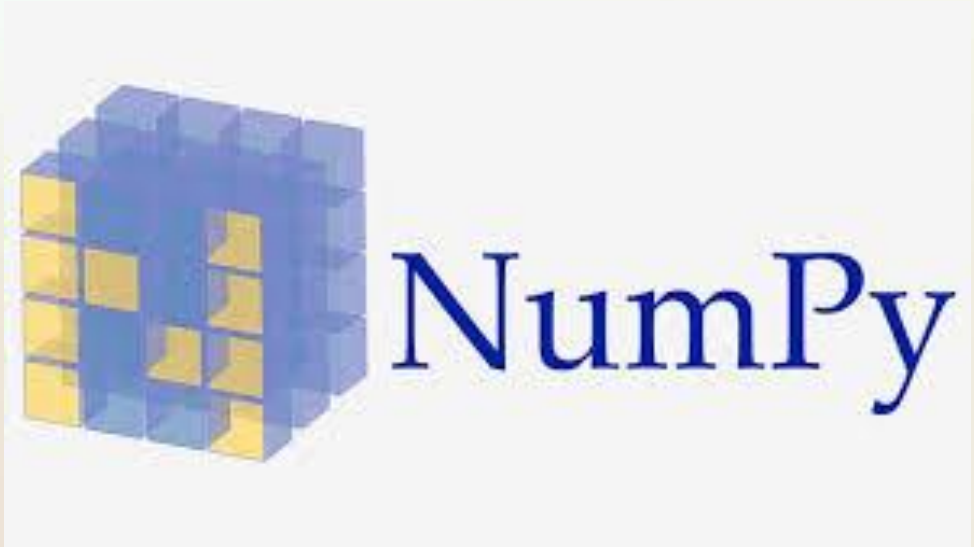
# Data Science For Predictions

- Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.



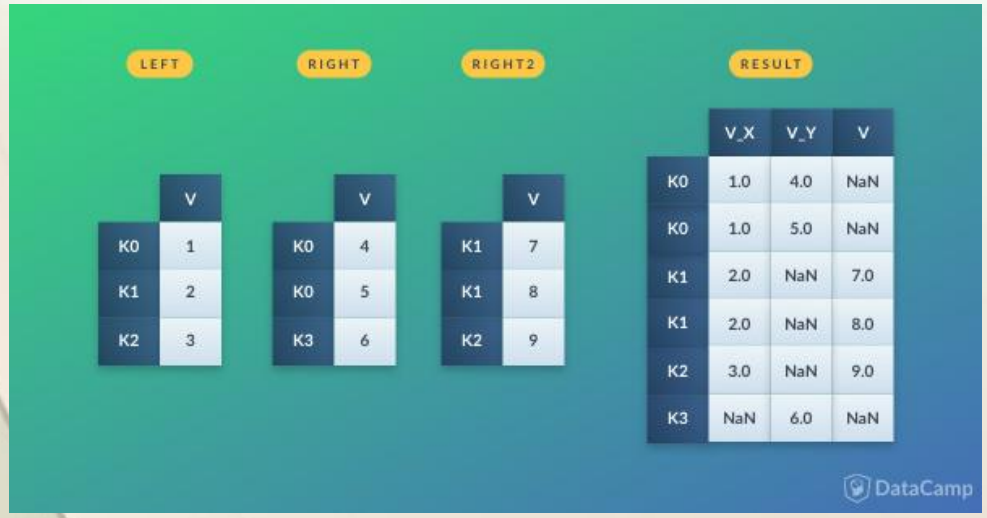
# Numpy

- NumPy is the fundamental package for scientific computing with Python. It is used for performing numeric operations on arrays
- NumPy is better than python list in terms of size, speed and functionality.



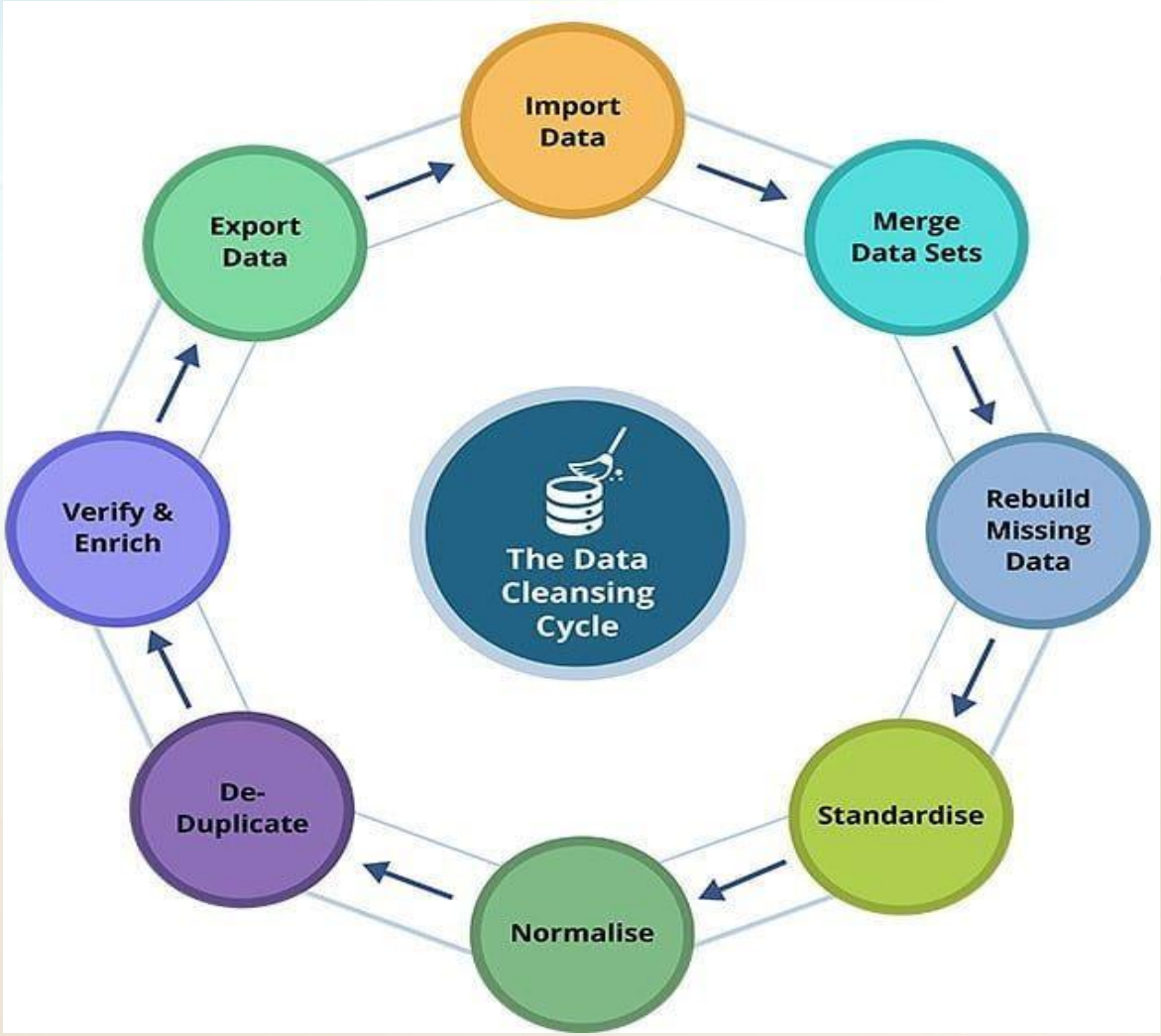
# Pandas

- **pandas** is a software library written for the **Python** programming language for data manipulation and analysis.
- Pandas DataFrames make manipulating your data easy, from selecting or replacing columns and indices to reshaping your data.




Pandas

# The Data Cleansing Cycle



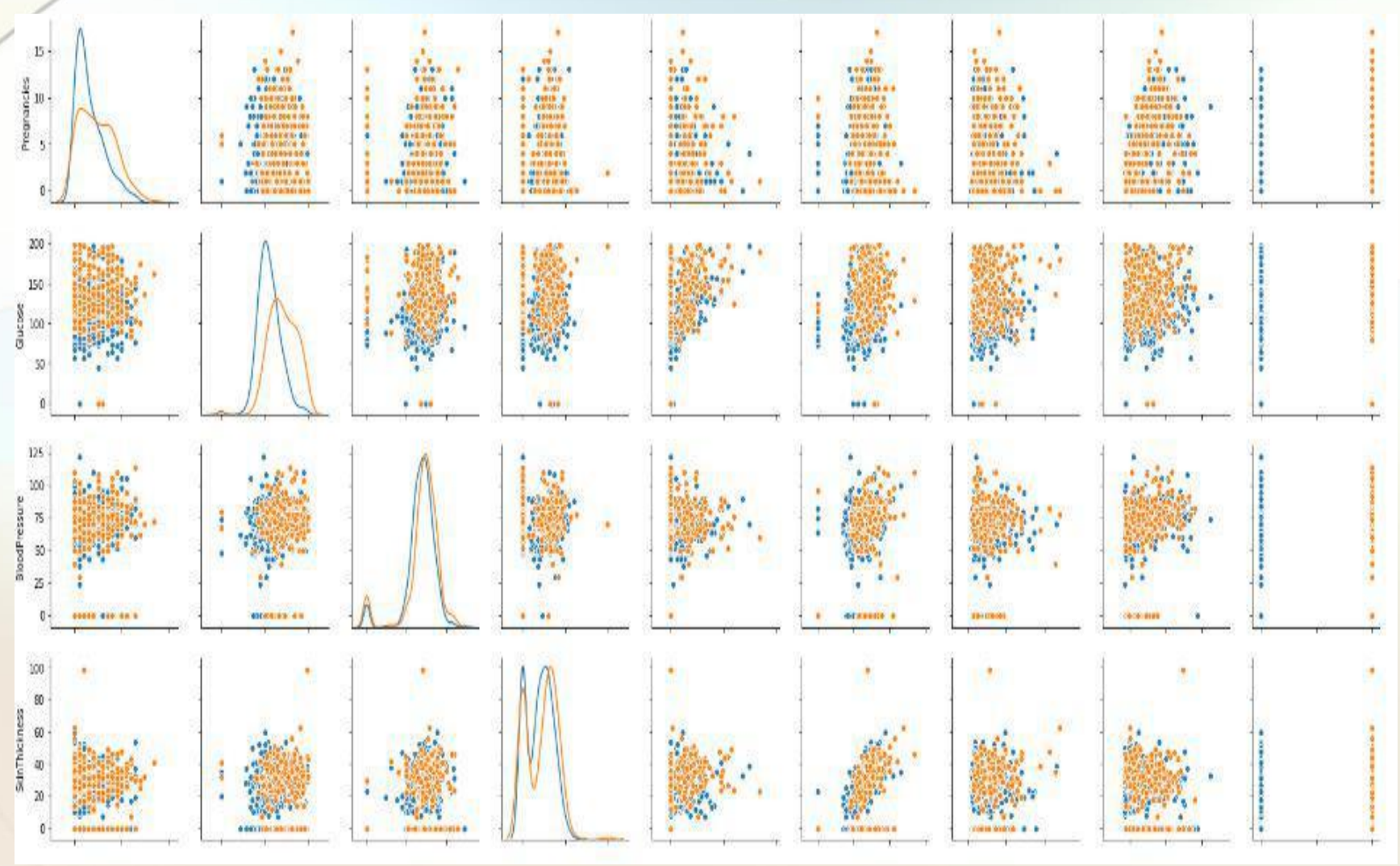
# Classification : DIABETES ANALYSIS AND PREDICTION

To predict whether the patients have diabetes or not, by passing the features. So first of all we will read the csv file of patients from which we get the data of the patients which contains the following columns:

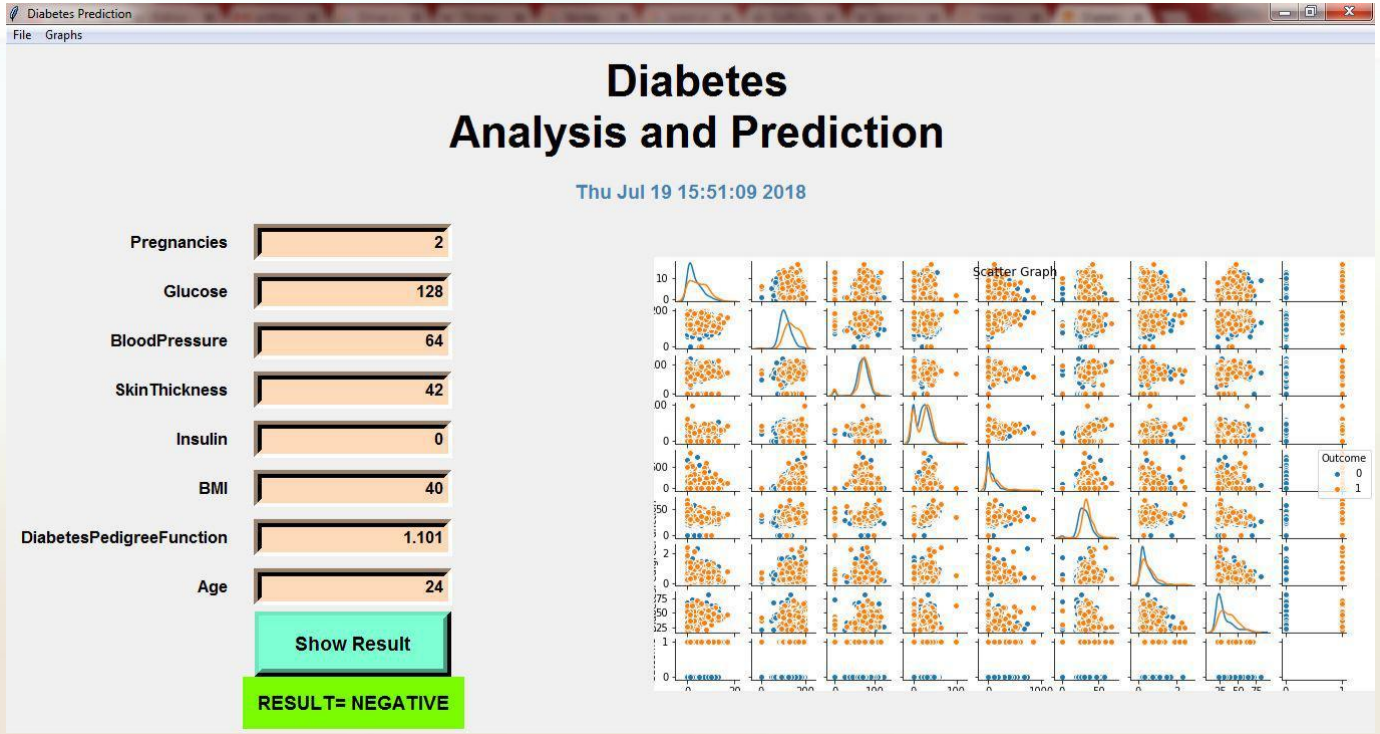
- Pregnancies ,
- Blood pressure
- Skin thickness
- Insulin , BMI
- Diabetes Pedigree Function
- Age , Outcome

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 5 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 6 | 3 | 78 | 50 | 32 | 88 | 31.0 | 0.248 | 26 | 1 |
| 7 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |

# Data Science : Diabetes Analysis: Scatter plot



# Data Science : Diabetes Analysis
# Patients with Diabetes Prediction Negative



# Diving into Machine Learning

**Python :** which is in the first place in the list of all AI development languages due to the simplicity.

There are plenty of libraries in python, which make our tasks easier.

The major deep learning frameworks like Tensorflow, Theano, Keras, neon, Caffe have Python interfaces.

# Diving into Machine Learning

**R:** which is one of the most effective language and environment for analyzing and manipulating the data for statistical purposes.

Using R, we can easily produce well-designed publication-quality plot, including mathematical symbols and formulae where needed.

Apart from being a general purpose language, R has numerous packages which are used in the field of machine learning.

# Diving into Machine Learning

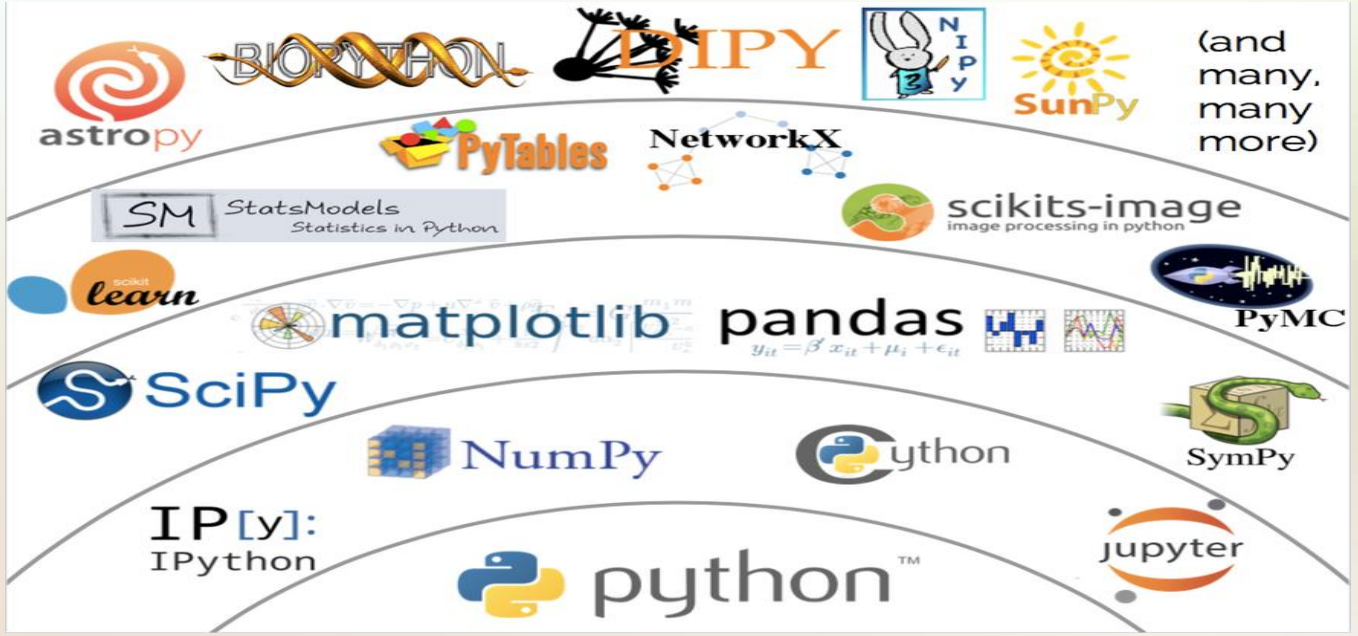
- **Java:** Java can also be considered as a good choice for AI development. Artificial intelligence has lot to do with search algorithms, artificial neural networks and genetic programming.
- **JSAT** which stands for Java Statistical Analysis Tool, is a machine learning library developed in Java for solving machine learning problems.
- **Scala** is a rival of Java and Python in the world of Data Science and becoming more and more popular due to extensive use of Apache Spark in Big data Hadoop.

# Python for Data Science and ML

In a nutshell, **Python** is **better** for data manipulation and repeated tasks, while **R** is good for ad hoc analysis and exploring datasets. **R** has a steep learning curve, and people without programming experience may find it overwhelming.

- **Python** is generally considered easier to pick up.

Python EcoSystem



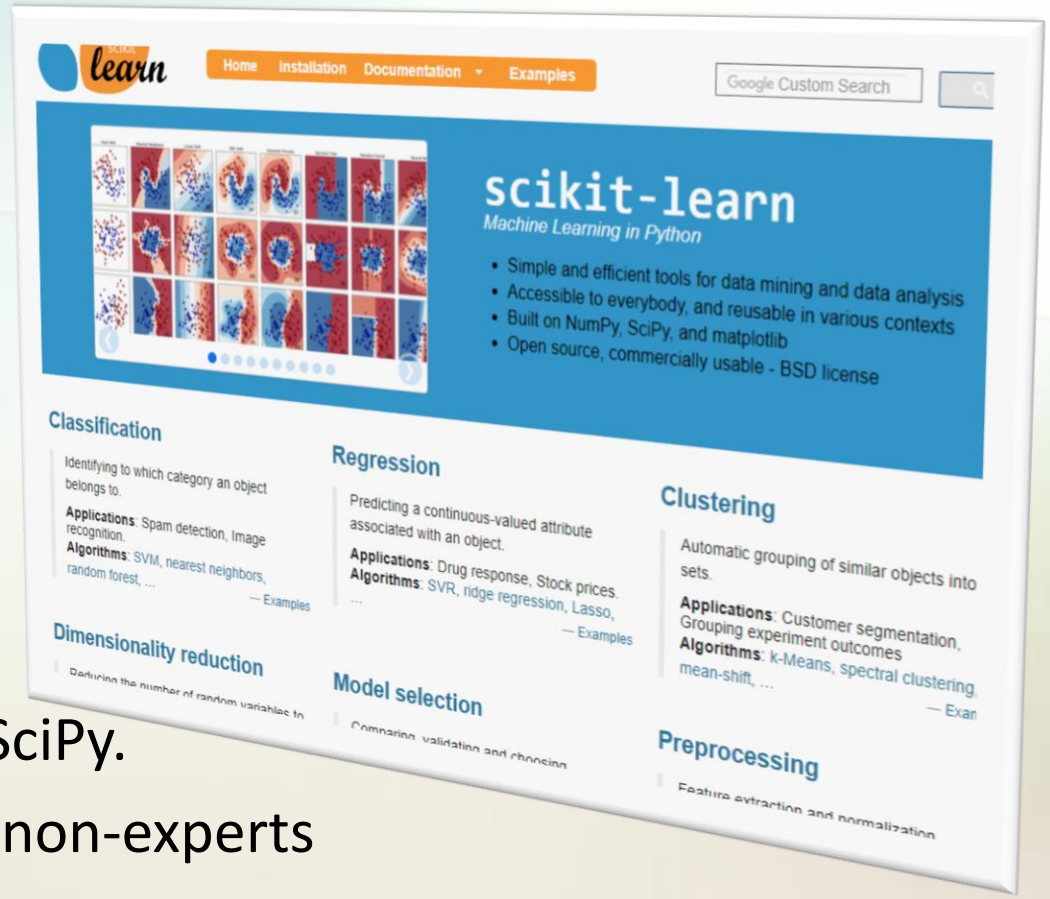
# Ecosystem of Python

Python has a rich ecosystem for scientific inquiry in the form of many proven, and popular open-source packages and machine learning frameworks including:

- **Numpy**: a Python package for scientific computing providing support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- **Pandas:** is an open source library package for data science providing high-performance, easy-to-use data structures like dataframes.
- **Matplotlib:** 2D plotting library which produces publication quality figures
- **Scikit** : provides a range of supervised and unsupervised machine learning algorithms via a consistent interface in **Python**.

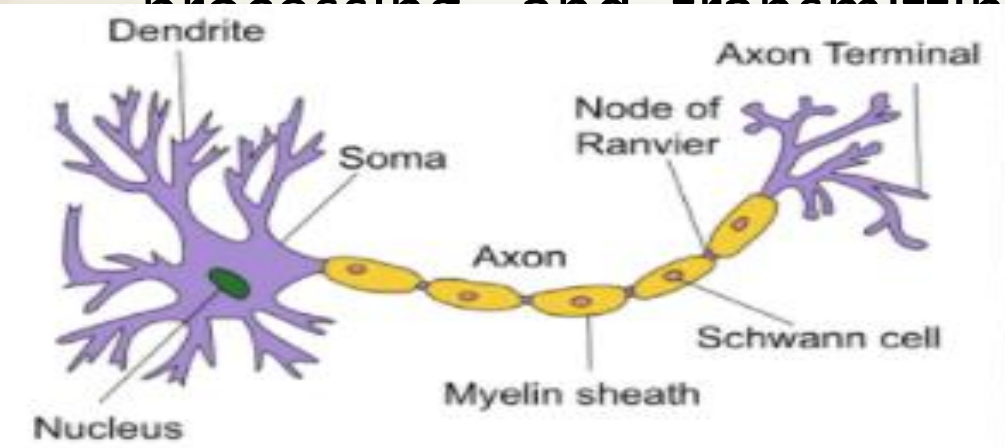
# Machine Learning in Python : Scikit Learn

- Scikit-learn is a free machine learning library for the **Python** programming language.
- It features various classification, regression and clustering algorithms including support vector machines, random forests, k-means, Naïve Bayes.
- It is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- It is Simple and efficient, for both experts and non-experts
- It contains classical and well-established machine learning algorithms

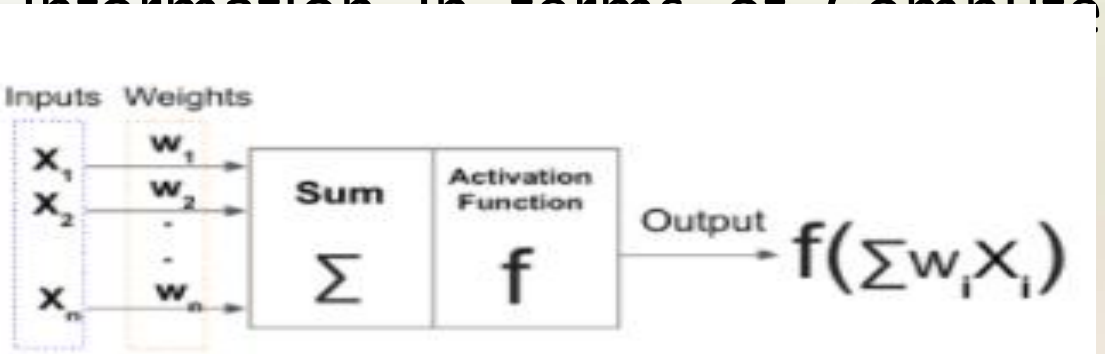


# Artificial Neural Networks

- An Artificial **Neuron** Network (**ANN**), popularly known as Neural Network comprises of Set of Techniques which were inspired by human nervous system.
- It is s a computational model based on the structure and functions of biological neural networks.
- It is like an artificial human nervous system for receiving,



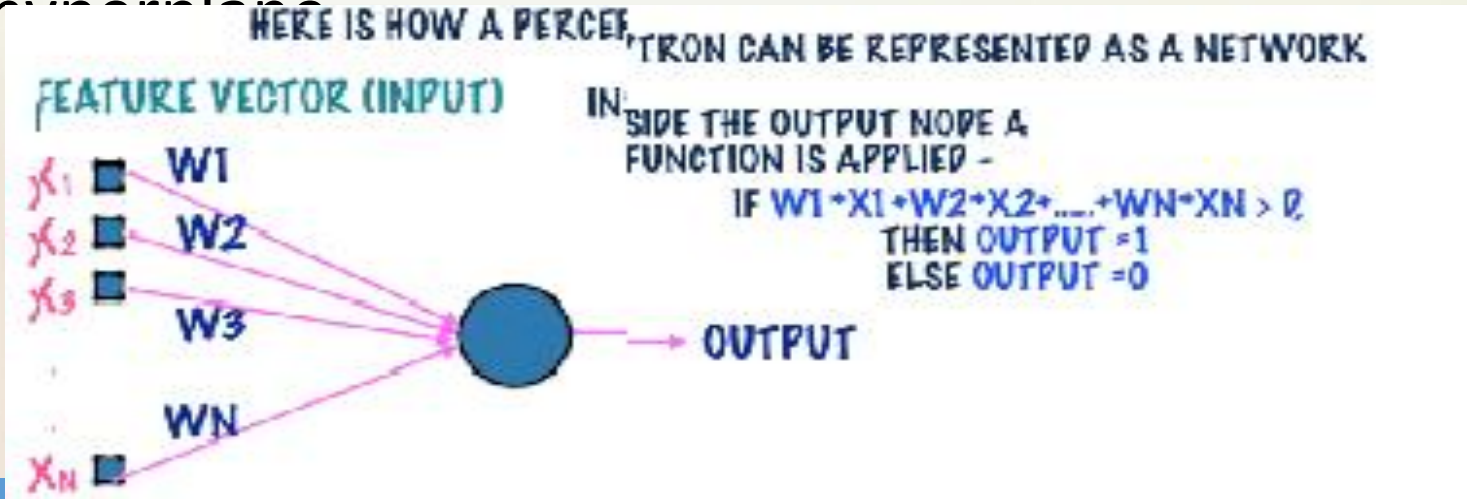
Structure of a typical neuron
(source: Wikipedia)



Structure of artificial neuron

# Artificial Neural Networks

- A Perceptron is a type of artificial neuron which takes in several binary inputs x1, x2, … , xn and produces a single binary output. In the example shown above, there are three inputs x1, x2 and x3.
- A perceptron is a specific algorithm for determining some hyperplane that separates data of two categories.
- The SVM finds the best such hyperplane namely the maximum margin h



# Examples of AI

## Visual Recognition

- An unsupervised learner processes 10 million videos together with related textual data such as descriptions and comments.
- The learner models images in the videos using statistical analysis that allows it to identify visual patterns.
- These patterns can then be correlated with text to develop theories about the visual traits of various things.
- For example, such a learner might be able to build a solid model that can identify skateboards in videos. The learner is never given the right answer but can gain confidence based on a large number of samples.
- Likewise, the learner will discard a large number of models that don't appear to be correct.

# Examples of AI

## Visual Recognition

- An AI that is learning to identify pedestrians on a street is trained with 2 million short videos of street scenes from self-driving cars.
- Some of the videos contain no pedestrians at all while others have up to 25.
- A variety of learning algorithms are trained on the data with each having access to the correct answers
- Each algorithm develops a variety of models to identify pedestrians in fast moving scenes. The algorithms are then tested against another set of data to evaluate accuracy and precision.

# Examples of AI

## Decision Support

- An AI is learning to estimate investing risk.
- It is fed a large number of trades that real investors made and asked to estimate a risk/reward ratio for each trade based on company fundamentals, price and other factors such as volume.
- The estimated risk/reward ratio is then compared to the historical results of the trade at a variety of time intervals such as a day or a year.

# Examples of Unsupervised Learning

**Human Behavior**

- A learner that possesses visual highly developed visual and speech recognition capabilities could watch a large number of television shows to learn about human behavior.
- For example, a learner might be able to build a model that detects when people are smiling based on correlation of facial patterns and words such as "what are you smiling about?“
- Personalised Recommendation Systems?

# Example of Deep Learning

Speech Recognition

Self Driving Car

Robotics

Face Detection

Signature Detection

Customer Discovery

Character Recognition (Multiple cat.)

# Natural Language Processing

- It's a form of artificial intelligence that focuses on analyzing the human **language** to draw insights, create advertisements, help you text (yes, really) and more.
- A field of computer science related to understanding  and generating human language.
  - Auto Summarization of Text
  - Spelling Correction
  - Text Classification
  - Sentiment Analysis
  - Topic Analysis

# Natural Language Processing

- **Natural Language Toolkit**, or **NLTK**, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English.

- Natural language processing (NLP) is about developing applications and services that are able to understand human languages.

- Practical examples of natural language processing (NLP) are :
  - understanding complete sentences,
  - understanding synonyms of matching words,
  - writing complete grammatically correct sentences and paragraphs,
  - speech recognition, speech translation.

# CRISP-DM Framework

- **CRISP-DM** framework – **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- It is an open standard process model that describes common approaches used by data mining experts and is the most widely-used analytics model as it solved the existing problems in the data-mining industries.
- Implemented in any Data Science project irrespective of its domain.
- CRISP-DM was conceived in 1996 and became a European Union project under the ESPRIT funding initiative in 1997. The project was led by five companies: Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation and OHRA, an insurance company.

# CRISP-DM Framework

- **CRISP-DM** encourages best practices and allows projects to replicate. This methodology provides a uniform framework for planning and managing a project.
- Typical analytics projects involve multiple steps like data cleaning, preparation, modelling, model evaluation etc. It may take several months, and thus it is important to have a structure for it.

# CRISP-DM Framework

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

# 1. Business Understanding

- Understanding the business requirement is of paramount importance. You must understand the problem clearly to convert it into a well-defined analytics problem.
- **Determine the business objectives clearly**:

- **Assess the situation**:
- **Determine the goals of data analysis**:
- **Produce a project plan**:

# 2. Data Understanding

- The data understanding phase goes hand in hand with the business understanding phase and encourages the focus to ascertain, assemble, and scrutinize the data sets that can help you achieve the project goals.
- **Collect initial data**
- **Describe data**
- **Explore data**
- **Verify data quality**
- **Data Preparation**
  - Select, clean, construct, integrate, format

# 4. Modelling

- This task refers to the specific modelling technique, e.g., decision-tree or random forest building, or neural network generation with backpropagation.
- build and assess various models based on several different modelling techniques.
- **Select modelling techniques**
- **Generate test design**
- **Build model**
- **Assess model**

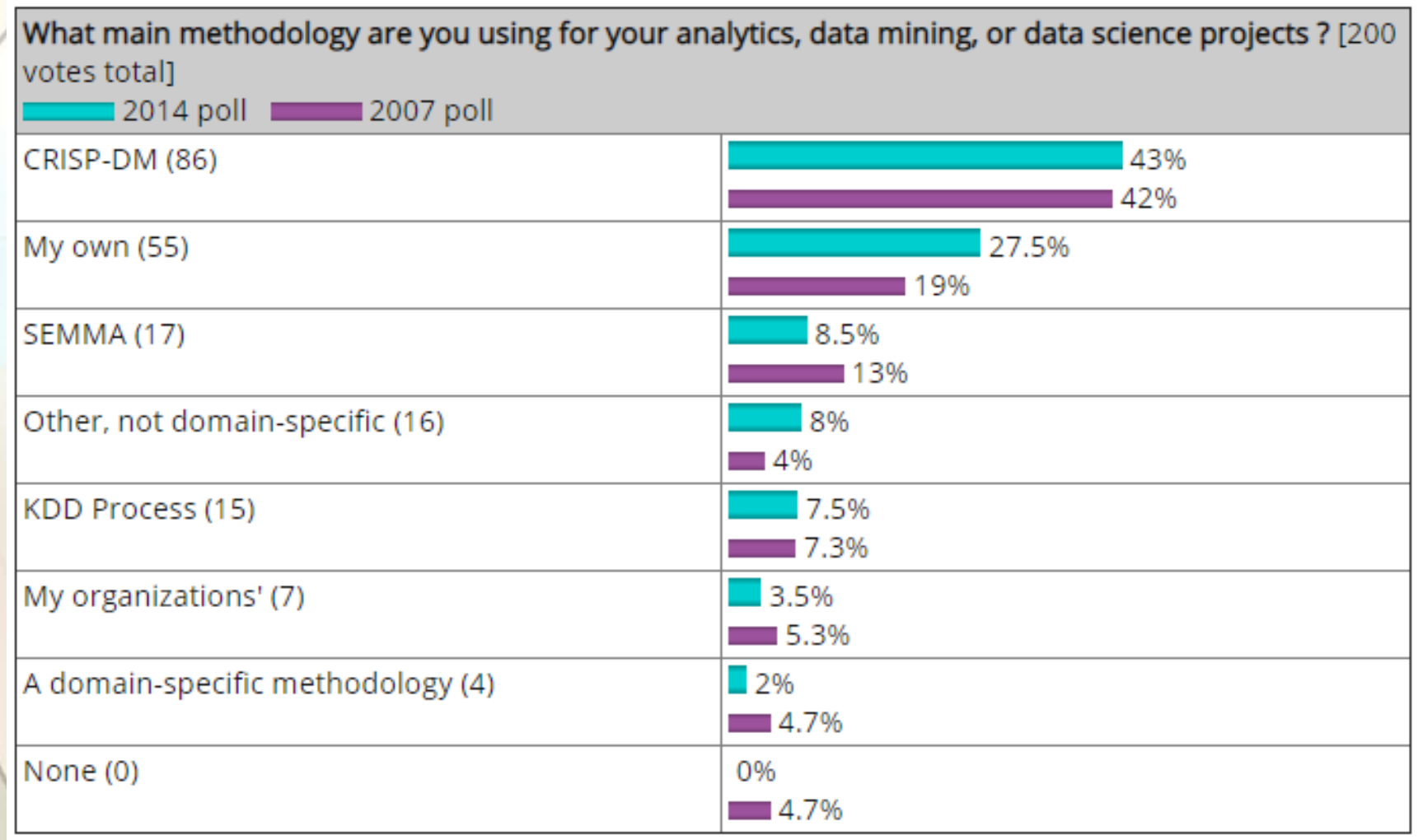
# 5. Evaluation

- thoroughly evaluate the model and review the steps executed to construct the model to ascertain that it properly achieves the business objectives.
- The evaluation phase looks more broadly at which model best meets the business and what to do next
- **Evaluate results**
- **Review process**
- **Determine next steps**

# CRISP-DM in Detail

| Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment |
|---|---|---|---|---|---|
| **Determine Business Objectives**<br>*Background*<br>*Business Objectives*<br>*Business Success Criteria* | **Collect Initial Data**<br>*Initial Data Collection Report* | **Select Data**<br>*Rationale for Inclusion/ Exclusion* | **Select Modeling Techniques**<br>*Modeling Technique*<br>*Modeling Assumptions* | **Evaluate Results**<br>*Assessment of Data Mining Results w.r.t. Business Success Criteria*<br>*Approved Models* | **Plan Deployment**<br>*Deployment Plan* |
| **Assess Situation**<br>*Inventory of Resources*<br>*Requirements, Assumptions, and Constraints*<br>*Risks and Contingencies*<br>*Terminology*<br>*Costs and Benefits* | **Describe Data**<br>*Data Description Report* | **Clean Data**<br>*Data Cleaning Report* | **Generate Test Design**<br>*Test Design* | **Review Process**<br>*Review of Process* | **Plan Monitoring and Maintenance**<br>*Monitoring and Maintenance Plan* |
| **Determine Data Mining Goals**<br>*Data Mining Goals*<br>*Data Mining Success Criteria* | **Explore Data**<br>*Data Exploration Report* | **Construct Data**<br>*Derived Attributes*<br>*Generated Records* | **Build Model**<br>*Parameter Settings*<br>*Models*<br>*Model Descriptions* | **Determine Next Steps**<br>*List of Possible Actions*<br>*Decision* | **Produce Final Report**<br>*Final Report*<br>*Final Presentation* |
| **Produce Project Plan**<br>*Project Plan*<br>*Initial Assessment of Tools and Techniques* | **Verify Data Quality**<br>*Data Quality Report* | **Integrate Data**<br>*Merged Data* | **Assess Model**<br>*Model Assessment*<br>*Revised Parameter Settings* | | **Review Project**<br>*Experience Documentation* |
| | | **Format Data**<br>*Reformatted Data*<br><br>*Dataset*<br>*Dataset Description* | | | |

# Global Acceptance of CRISP-DM



# Why does CRISP-DM make you a better Data Scientist?

- CRISP-DM provides a uniform framework for
  - Guidelines
  - experience documentation
- This methodology is cost-effective as it includes a number of processes to take out simple data mining tasks and the processes are well established across industry.
- CRISP-DM encourages best practices and allows projects to replicate.
- This methodology provides a uniform framework for planning and managing a project.
- Being cross-industry standard, CRISP-DM can be implemented in any Data Science project irrespective of its domain.
- CRISP-DM has been the de-facto industry standard process model for data mining, with an expanding number of applications across a wide array of industries.
- It is extremely important that every data scientist and data miner must understand the different steps of this model.

# References


- https://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist/
- Cross-industry standard process for data mining, https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining
- What is the CRISP-DM methodology, https://www.sv-europe.com/crisp-dm-methodology/
- Kdnuggets poll on CRISP-DM: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html