MPSSDEGB
कौशल से रोजगार, समृद्धि अपार

FITT
Foundation For Innovation And Technology Transfer

SANKALP
Ministry of Skill Development & Entrepreneurship

Data Science in IoT

Presenter Name

MPSSDEGB
कौशल से रोजगार, समृद्धि अपार

FITT
Foundation For Innovation And Technology Transfer

SANKALP
Ministry of Skill Development & Entrepreneurship

Recommended Book

https://www.amazon.in/Internet-Things-Surya-Durbha/dp/0190121092/ref=cm_cr_arp_d_bdcrb_top?ie=UTF8

Contents

- Intro to Data Science
- Data Science Processes
- IoT & Big Data relation
- Overview of AI
- Hands on using python

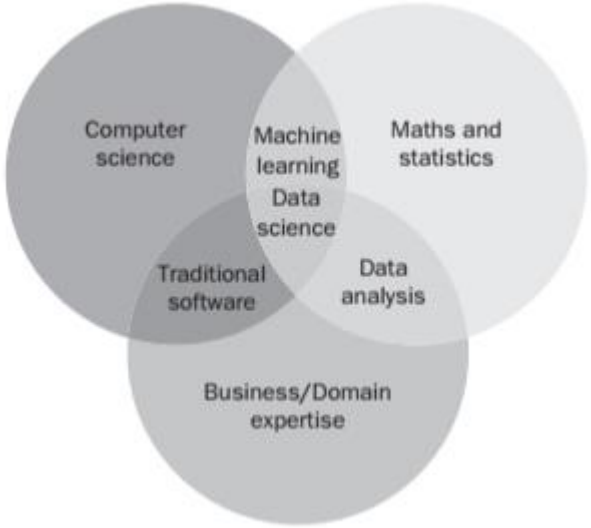
Data Science for IoT

Introduction

- Internet of Things (IoT) has emerged as a game-changing technology, revolutionizing the way we collect, process, and utilize data
- Vast network of interconnected devices and sensors, combined with data science, has opened up endless possibilities for businesses, industries, and everyday life
- Data from IoT is real time, not static helping to develop more accurate evaluations almost instantly

What is Data Science?

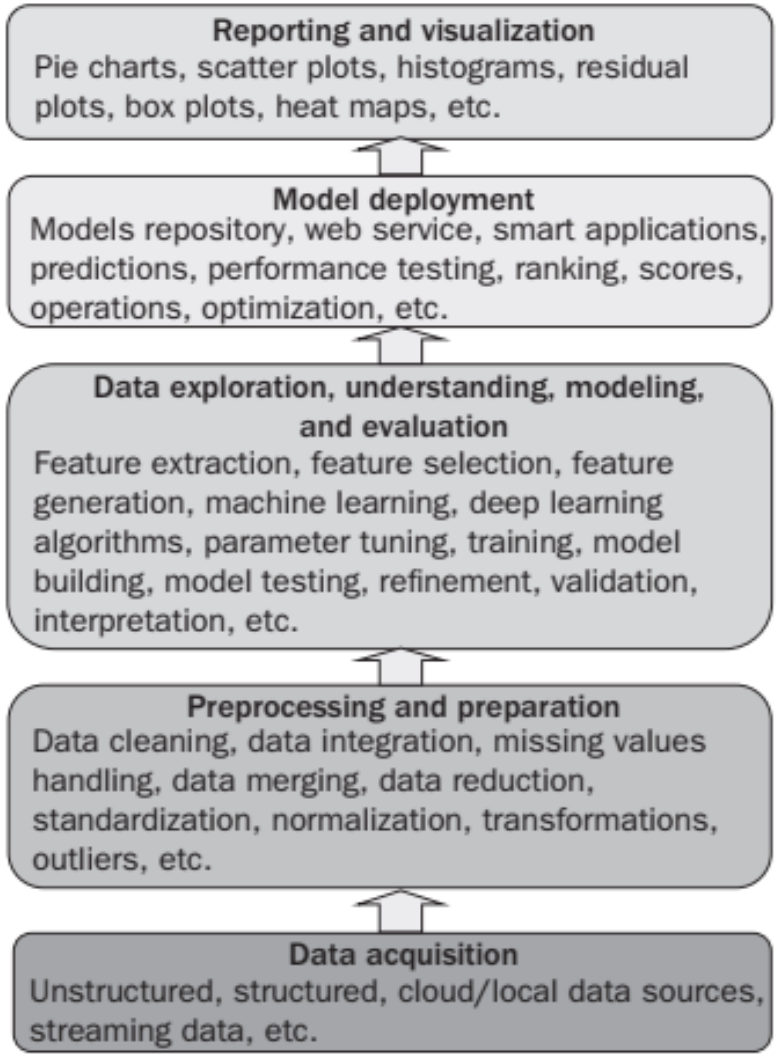
- Data science refers to the study of data in scientific manner involving several disciplines
- “An emerging area of work concerned with collection, preparation, analysis, visualization, management & preservation of collection of information” - Jeffrey Stanton



Source: Internet of Things, Durbha et.al.

Data Science Processes

Data Science Process



Source: Internet of Things, Durbha et.al.

Data Acquisition

- Data comes from variety of IoT devices -sensors, actuators
- Ancillary data (other supporting data such as device health)
- Acquisition mode can be online or offline
- Data generated by IoT devices is highly heterogeneous

source: https://www.upsolver.com/wp-content/uploads/2019/08/pasted-image-0-8.png

Nature and forms of data

Unstructured data - not fitting into a row/column format (non-relational database), no predefined data model associated

Structured data - data has predefined record length and associated data model, rarely used in IoT data

Data understanding, preprocessing & preparation

Involved processes:

- Importing data
- Data cleaning
- Missing value handling
- Data standardization
- Data normalization

Importing Data

Different ways including reading data from excel sheets, tables, comma separated values

Data Cleaning

Major step in preprocessing, helps to make data in form that is usable for further analysis

Rectify issues like missing values, outliers, malformed records

Hands on in Python

Importing and cleaning in python using colab

Scikit-learn (sklearn)

scikit-learn is an open-source Python library that implements a range of machine learning, pre-processing, cross-validation, and visualization algorithms using a unified interface

to install **pip install scikit-learn**

Important features of scikit-learn

- Simple and efficient tools for data mining and data analysis. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, etc.
- Accessible to everybody and reusable in various contexts.
- Built on the top of NumPy, SciPy, and matplotlib.
- Open source, commercially usable

Data Normalization

- Process of scaling the features between a predefined maximum and minimum.
- Scaling between 0 and 1 is commonly done

Normalization Formula

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

source: https://www.codingninjas.com/studio/library/normalisation-vs-standardisation

Data Standardization

- Process in which the data is restructured in a uniform format
- In statistics, standardization compares the variables by putting all the variables on the same scale
- Common way is to bring data to zero mean and unit variance

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

source: https://www.codingninjas.com/studio/library/normalisation-vs-standardisation

Hands on in Python

Colab Link

Data Exploration

Exploratory Data Analysis - done to gain basic intuition and understanding of the data to further prepare for modeling and analysis

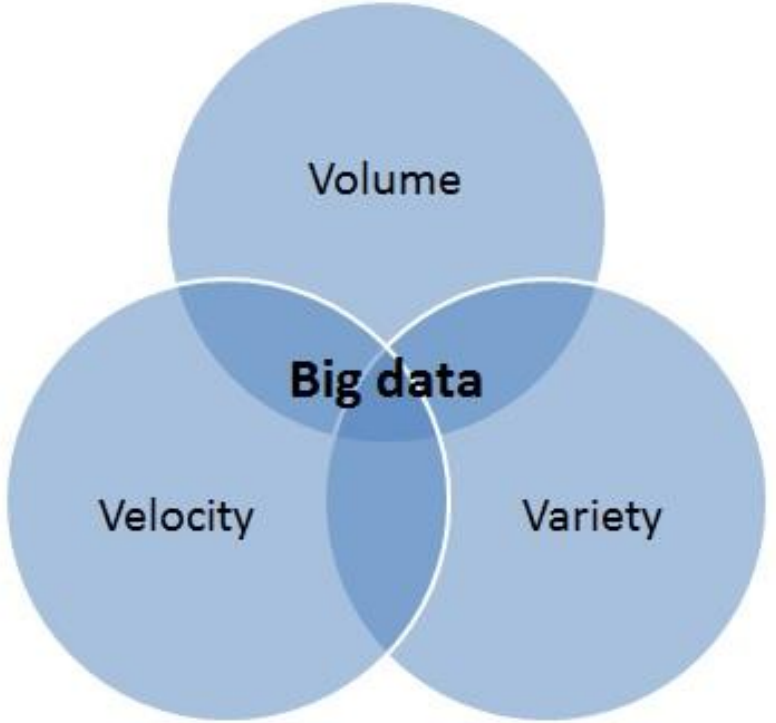
Tasks involved in EDA:

- Data description
- Sampling the data
- Data querying

Hands on in Python

Colab Link

Relation Between IoT & Big Data



source:https://bigdataldn.com/news/big-data-the-3-vs-explained/

Relation Between IoT & Big Data

IoT devices are generating tremendous amounts of data at an unprecedented rate. Characteristics of data in terms of big data:

Volume/Scale : millions of devices connected to internet, connecting people, devices and applications in a massive scale

Velocity : extremely high rate of data generation

Relation Between IoT & Big Data

Variety : data from diverse types of devices is in a variety of data models: structured, unstructured, semi-structured etc.

Heterogeneity : data for IoT based applications is usually gathered from heterogeneous data sources having multiple characteristics and structures

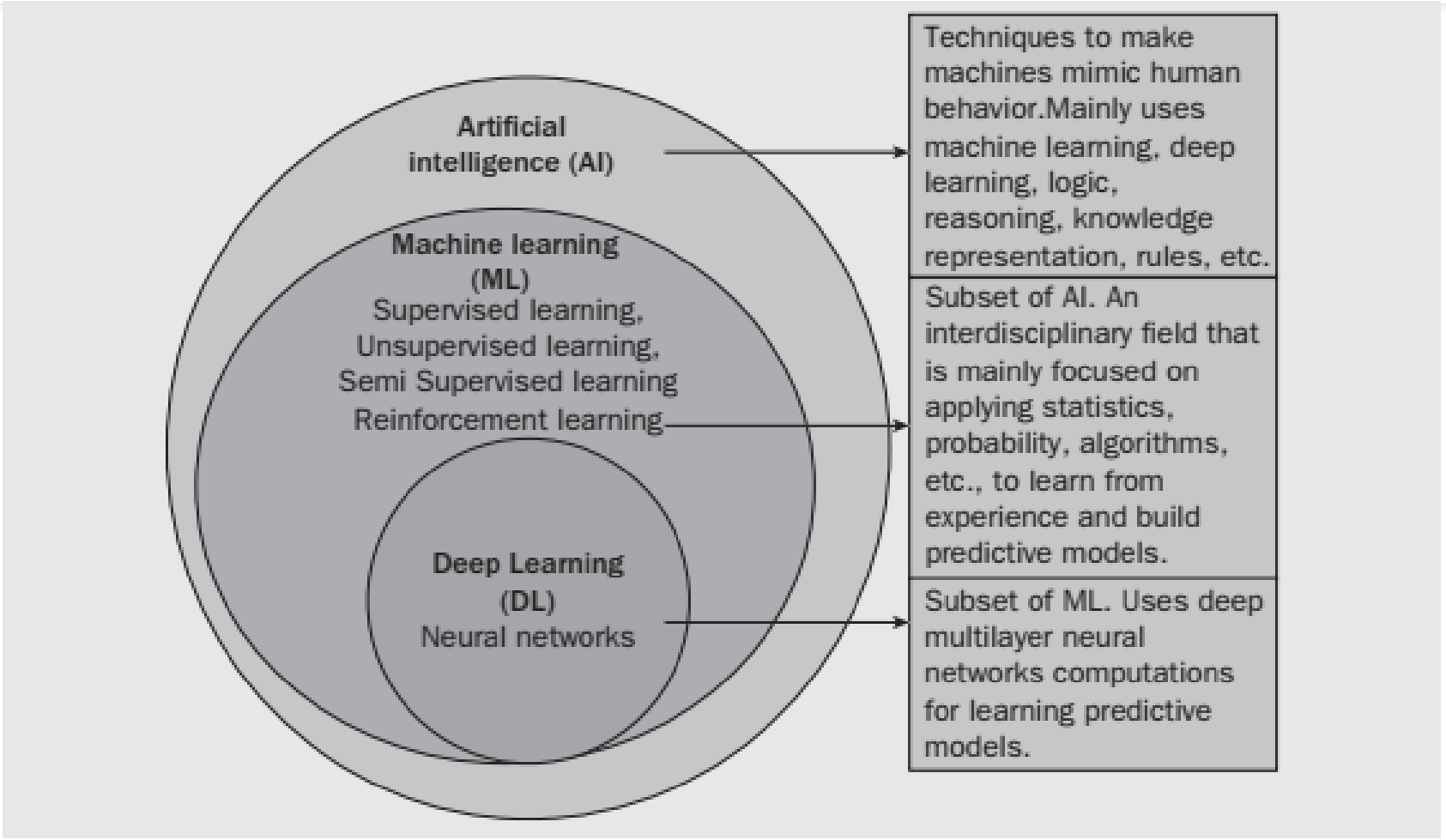
Big Data Analytics in IoT

- It provides a means for analyzing and visualizing data from IoT sensors, actuators, devices and other connected components of the IoT system
- Useful to understand, summarize and obtain useful insights from the large volumes of data

IoT Data Analytics usefulness

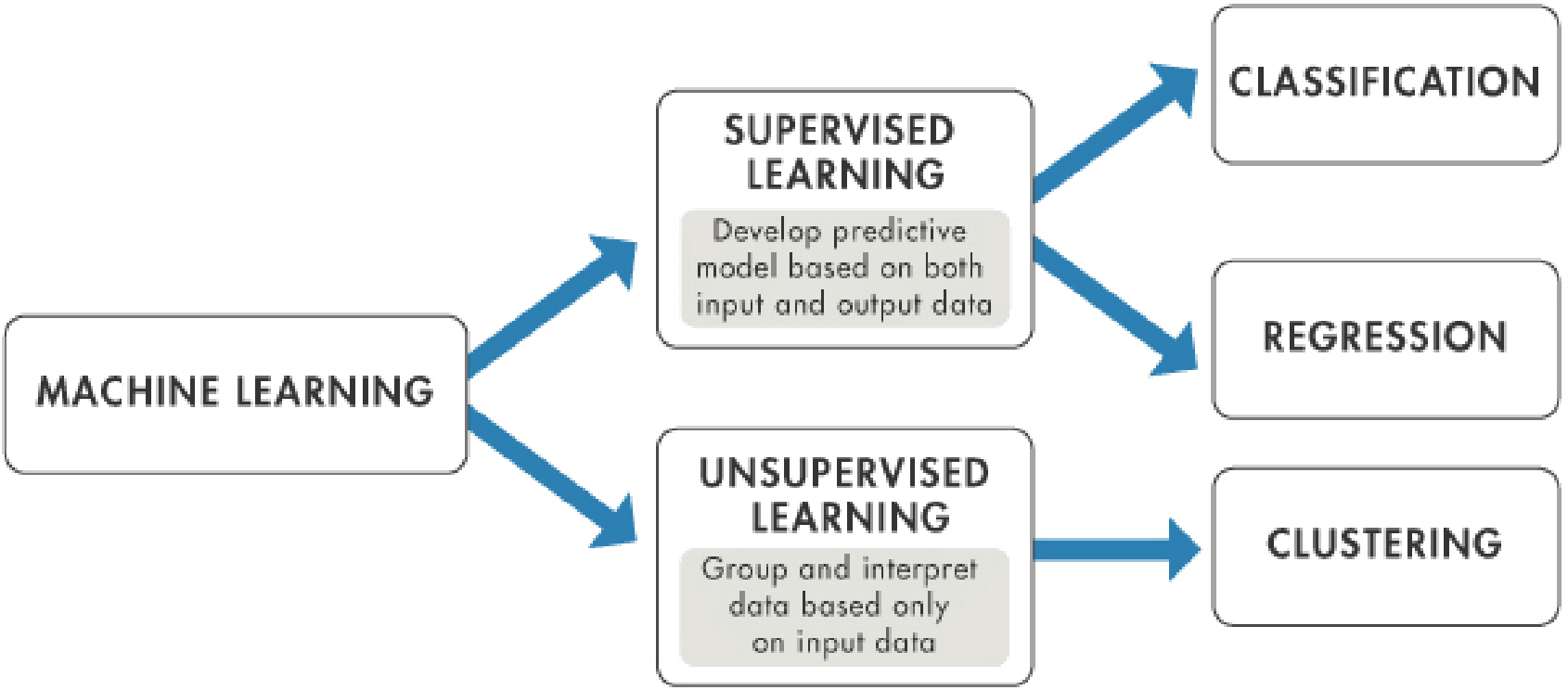
- Automating decision-making processes minimizing human intervention, IoT devices & applications can autonomously perform actions
- Increasing the efficiency with which processes can be executed
- Condition-based monitoring and predictive maintenance of equipment, which is critical in many areas such as industries, manufacturing, healthcare, and transportation
- Service efficiency that encompasses remote management, service chain, material management, etc.
- Analysis of the product usage by customers and accordingly customize the product thus enabling competitive advantage in the market
- Reducing overall operational expenditure and increasing revenue

AI, ML & DL



Source: Internet of Things, Durbha et.al.

Machine Learning



source: https://www.mathworks.com/help/stats/machine-learning-in-matlab.html

Machine Learning

Supervised Learning

Data: (x, y)
x is an input data, y is a label (e.g. photo with label "cat")

Goal: Learn to map input to output
i.e. $x \rightarrow y$

An example: to classify

This is a cat

Unsupervised Learning

Data: x
x is data, there's no labels!

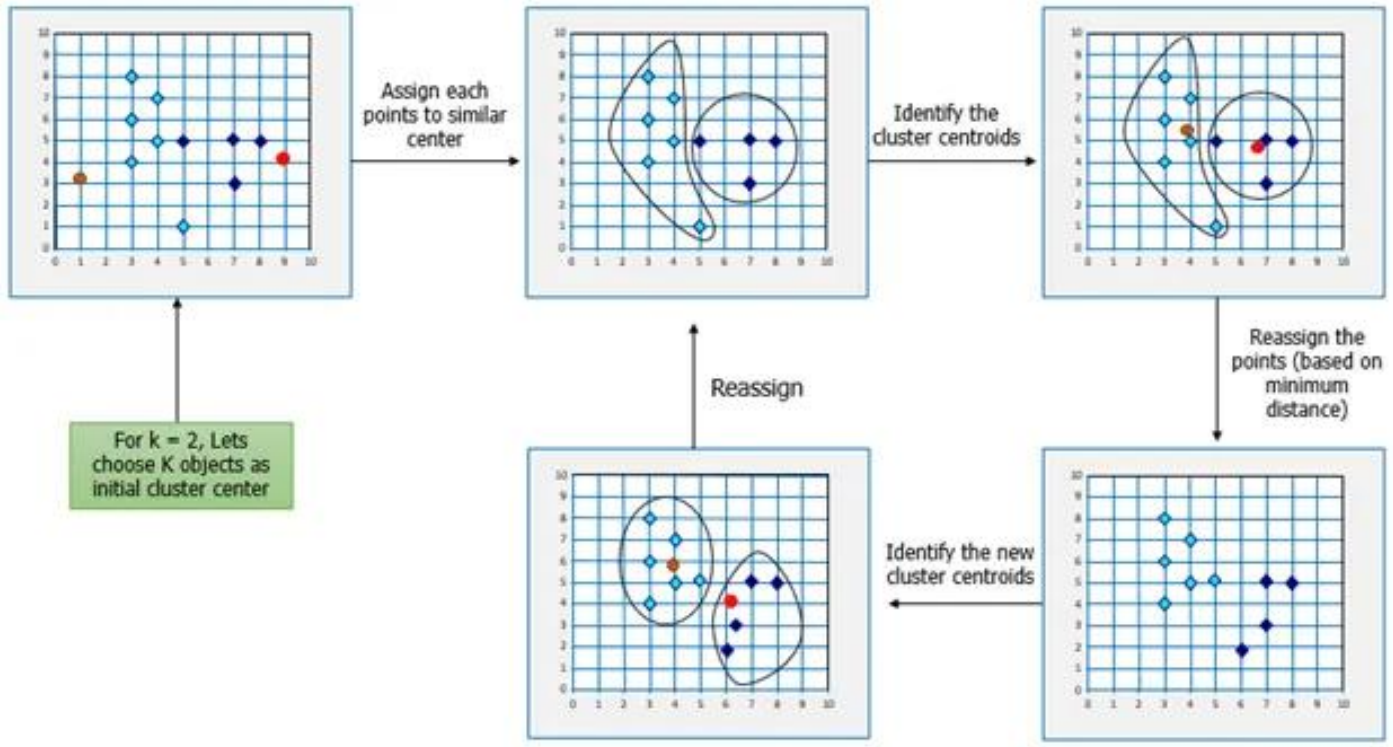
Goal: Learn an underlying structure of the data.

An example: Comparison

The two things are alike

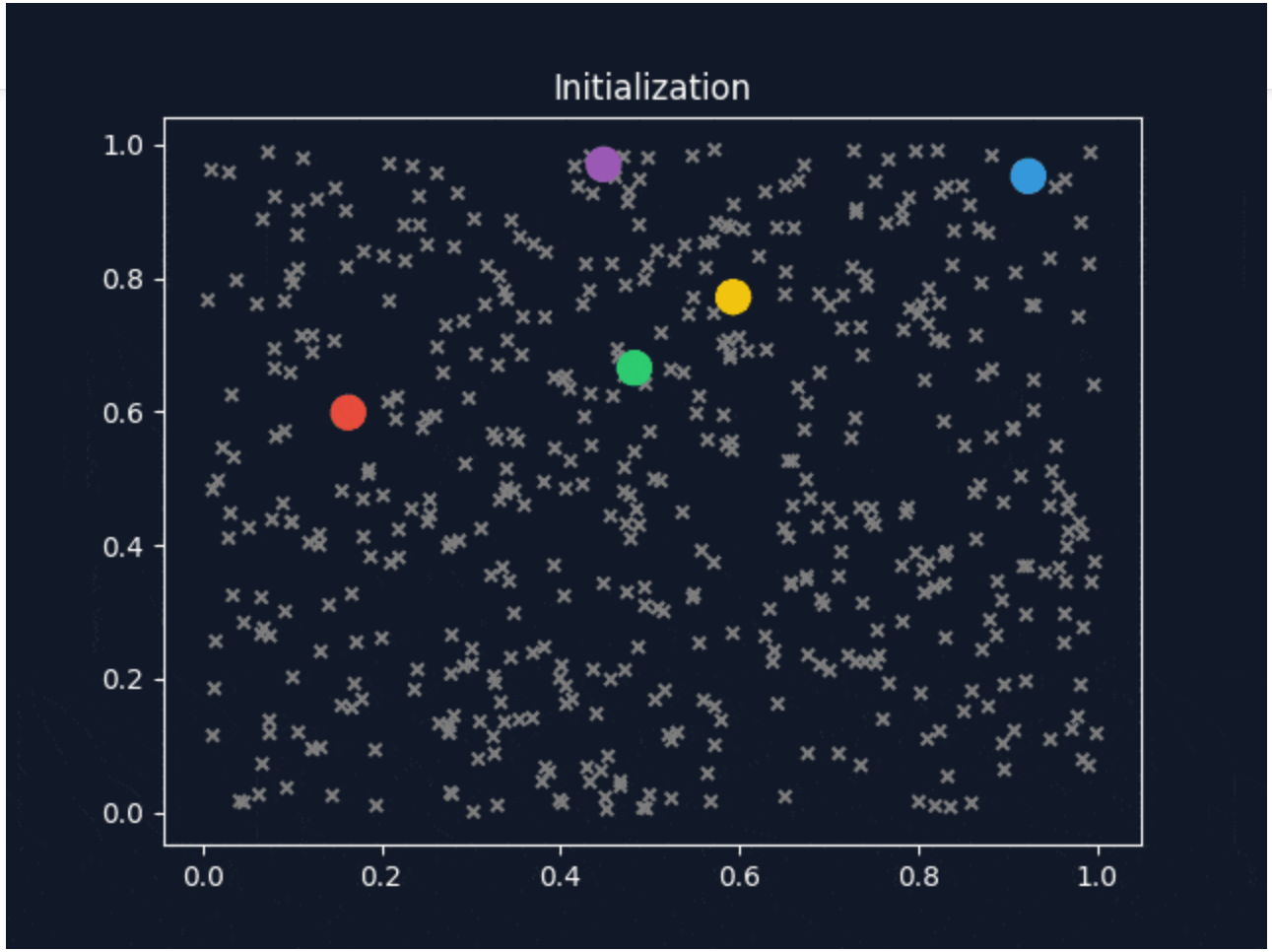
source:https://www.linkedin.com/pulse/decoding-machine-learning-supervised-unsupervised-daniel-wiczew/

K-Means Clustering



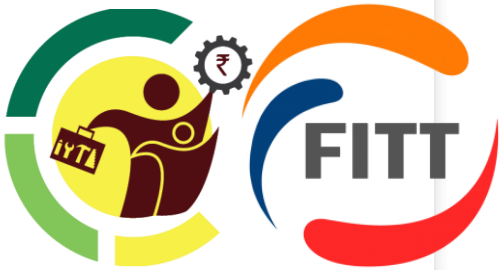
source: https://medium.com/@imparth/k-means-clustering-algorithm-34807a7cec71

K-Means Clustering



source: https://code-specialist.com/python/k-means-algorithm

Python Hands on



[Link for Colab]

Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that is commonly used in a wide range of regression and classification problems to classify labelled data determined by the optimized separating hyperplane.

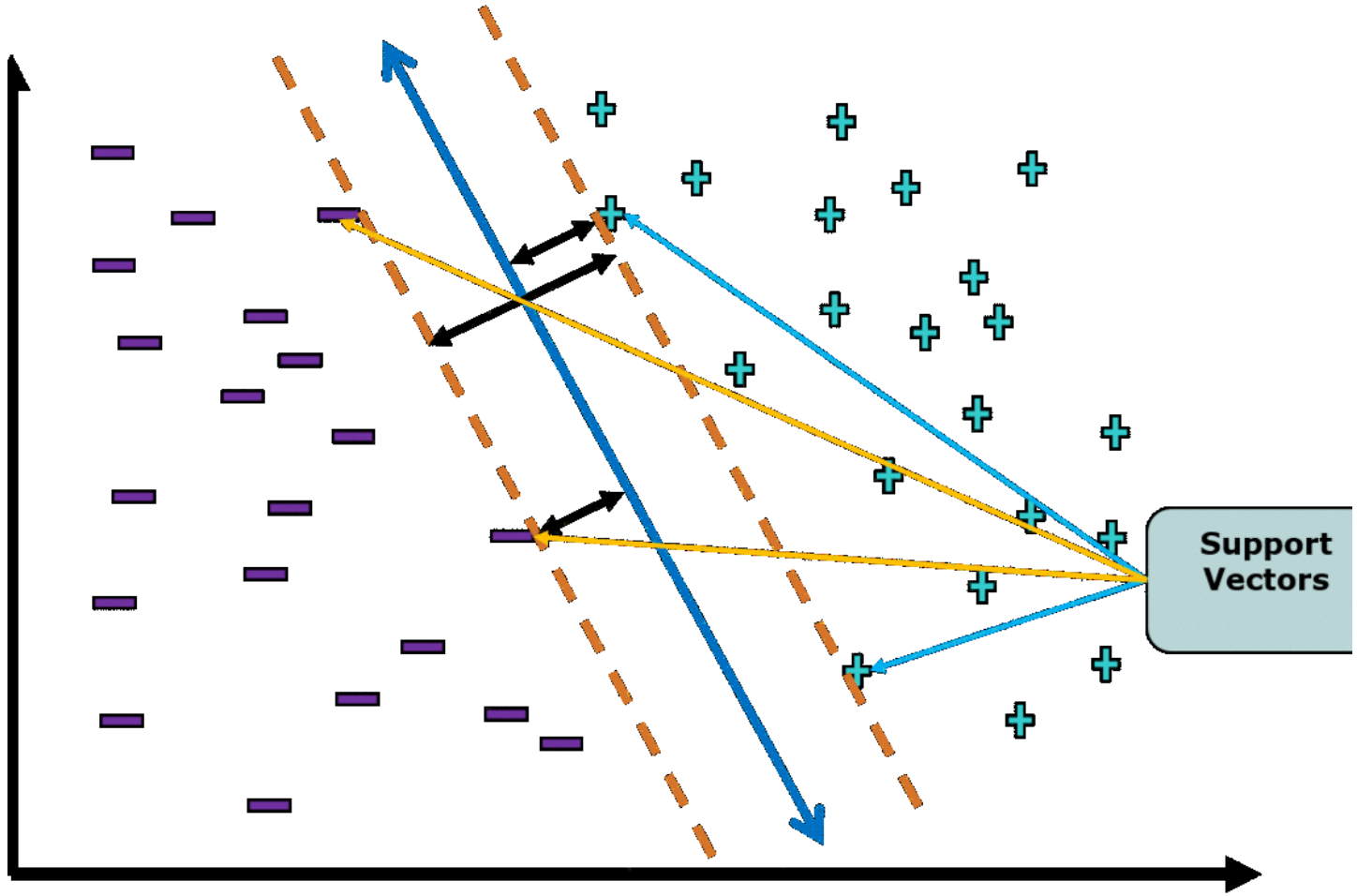
Support Vector Machine (SVM)

Hyperplane – The objective is to maximize the margin of separation between the hypothesis to those of distinct classes in an n-dimensional space. Can be perceived as a decision line that separates/splits the space into two parts of linearly separable data points(samples/observations).

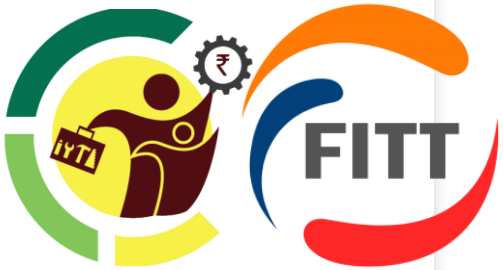
Support Vectors – are the classes (observations/data points) that are relatively close to the decision boundaries.

Soft Margins – are the perpendicular lines in both sides lies close to the support vectors. What soft margin does is to tolerates a few samples to be misclassified and it also performs a trade-off to figure out the line which minimize the misclassification and at the same time maximize the margin.

SVM



Python Hands on



[Link for Colab]

Homework

- What is Data Science and how it is useful for IoT?
- What are the characteristics of big data?
- Why data cleaning, standardization or normalization is important?
- Perform the exercise yourself utilizing the dataset.