Data Science

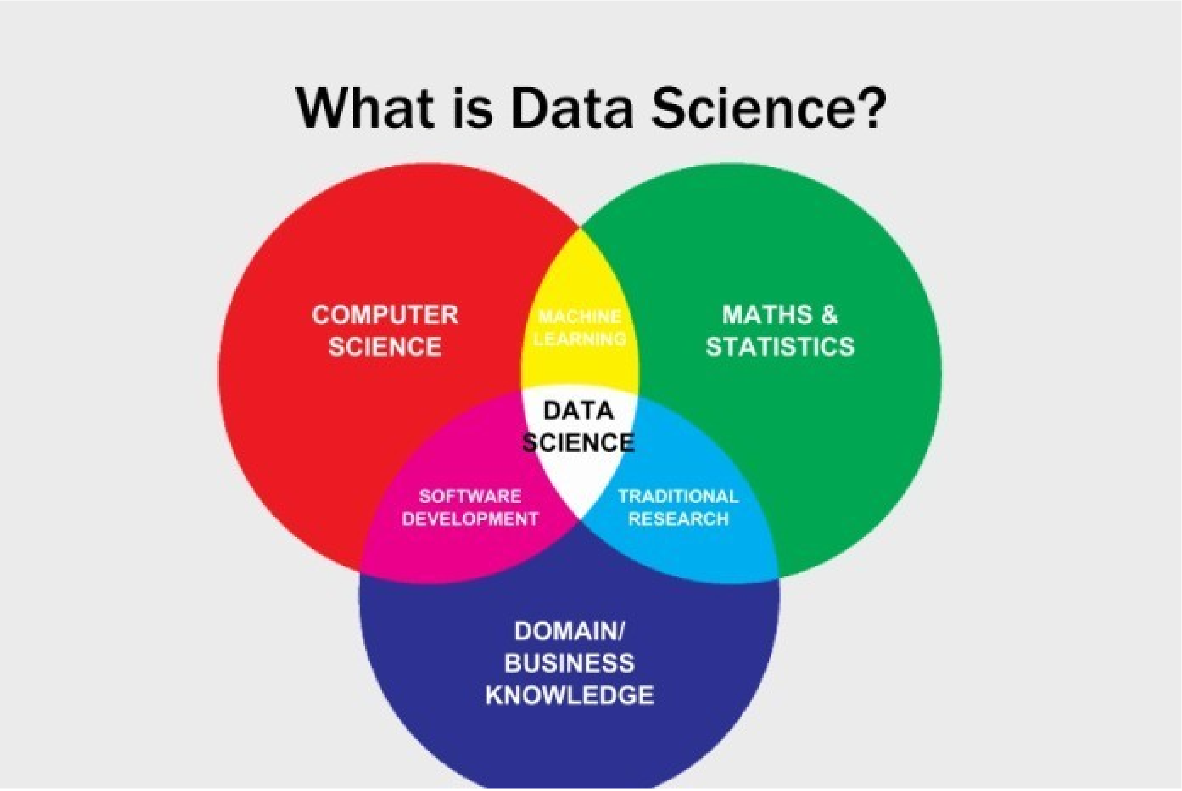
MODCOM

Institute of Technology

Introduction to Data science

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms to extract knowledge and insights from many structural and unstructured data.

Many disciplines……….



Data science……

Sports are always interesting to watch and we all have our favourites clubs.

And we always discuss with our friends and colleagues on which club will win the EPL .

A better way to make an informed guess is to involve the data to predict the EPL winner.

Data science……

Have you ever wondered how Amazon, eBay, Jumia suggest items for you to buy?

How Gmail filters your emails in the spam and non-spam categories?

How Netflix predicts the shows of your liking?

How a Bank mobile app would deny you a loan?

MODCOM Institute of Technology

Data science……

How do they do it?

These are the few questions we ponder from time to time. In reality, doing such tasks are impossible without the availability of data.

Data science is all about using data to solve problems.

Data science……

The problem could be decision making such as identifying which email is spam and which is not.

Or a product recommendation such as which movie to watch?

Or predicting the outcome such as who will be the next President of the USA?

So, the core job of a data scientist is to understand the data, extract useful information out of it and apply this in solving the problems

Data science……

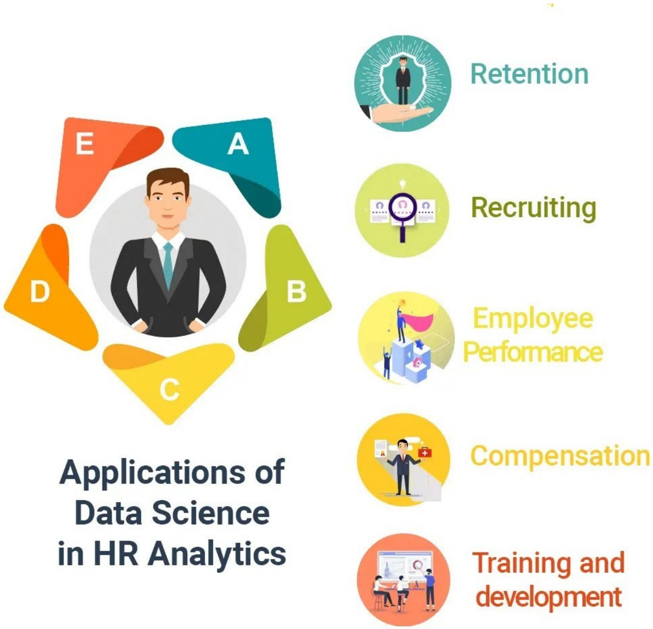
The problem could be decision making such as identifying which email is spam and which is not.

Or a product recommendation such as which movie to watch?

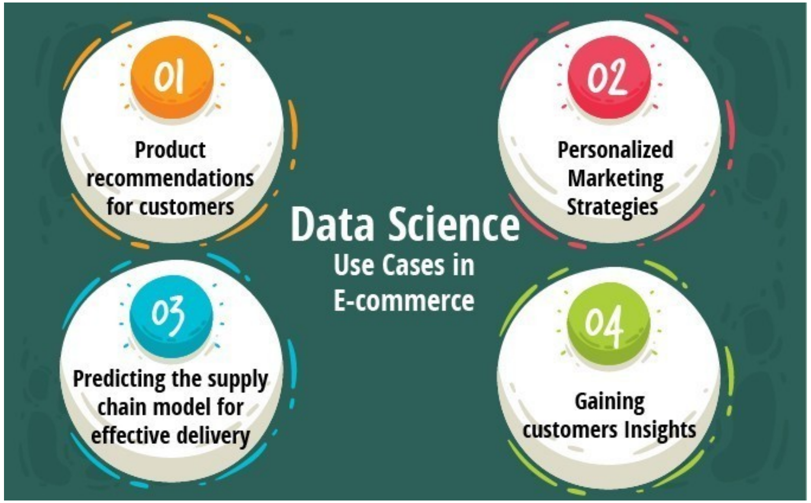
Or predicting the outcome such as who will be the next President of the USA?

So, the core job of a data scientist is to understand the data, extract useful information out of it and apply this in solving the problems

Applications of Data Science

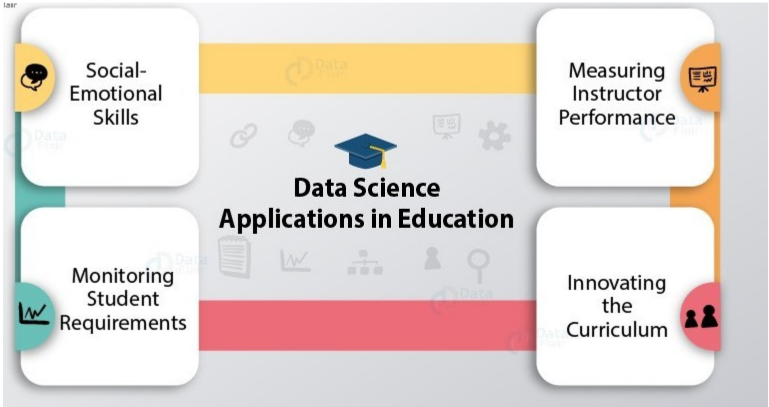


Applications of Data Science in HR Analytics



Data Science Use Cases in E-commerce

01
Product recommendations for customers
02
Personalized Marketing Strategies
03
Predicting the supply chain model for effective delivery
04
Gaining customers Insights

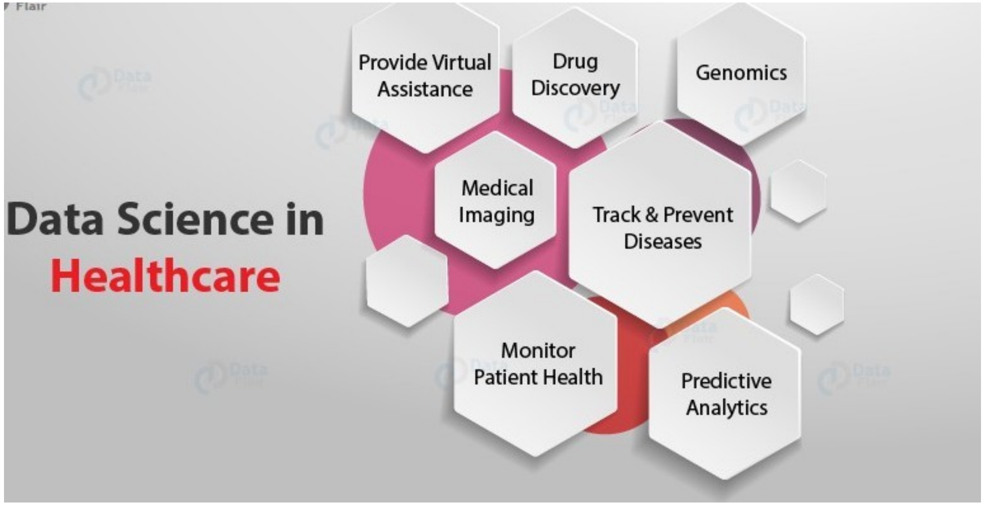


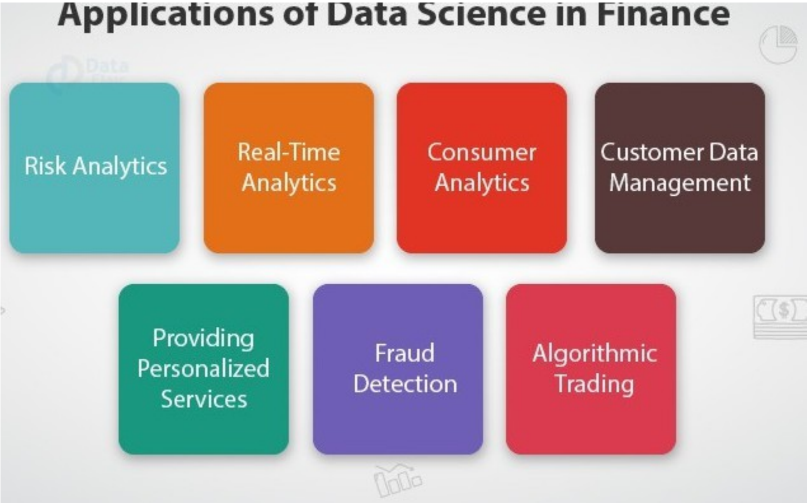
Data Science Applications in Education

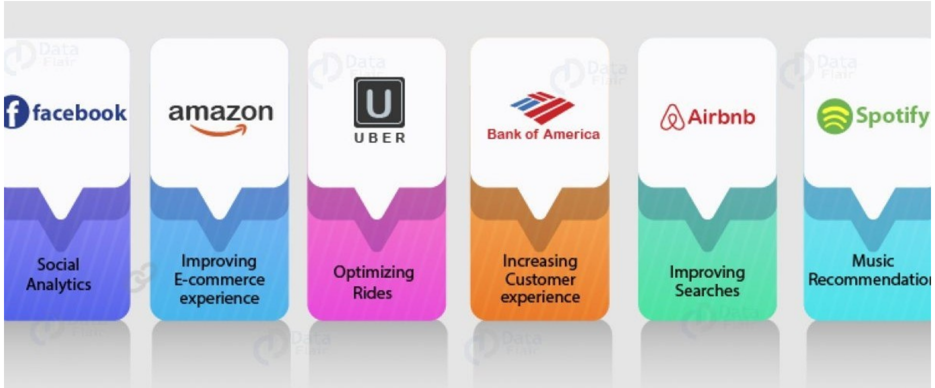
Social-Emotional Skills
Measuring Instructor Performance
Monitoring Student Requirements
Innovating the Curriculum

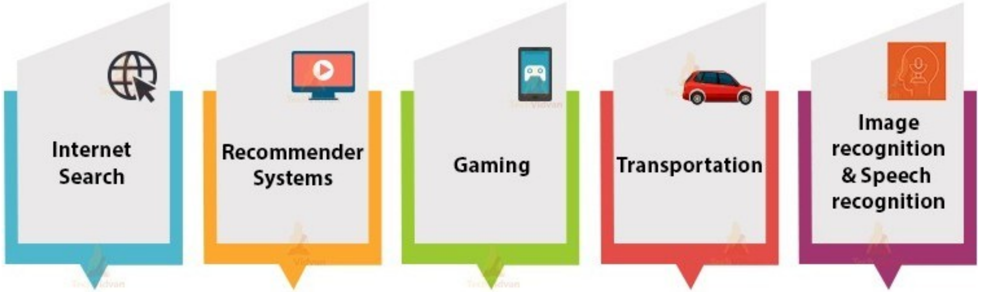
MODCOM Institute of Technology

Applications of Data Science









And many more…...

Data Science Methodology

Business understanding – What does the business need?

Data understanding – What data do we have / need? Is it clean?

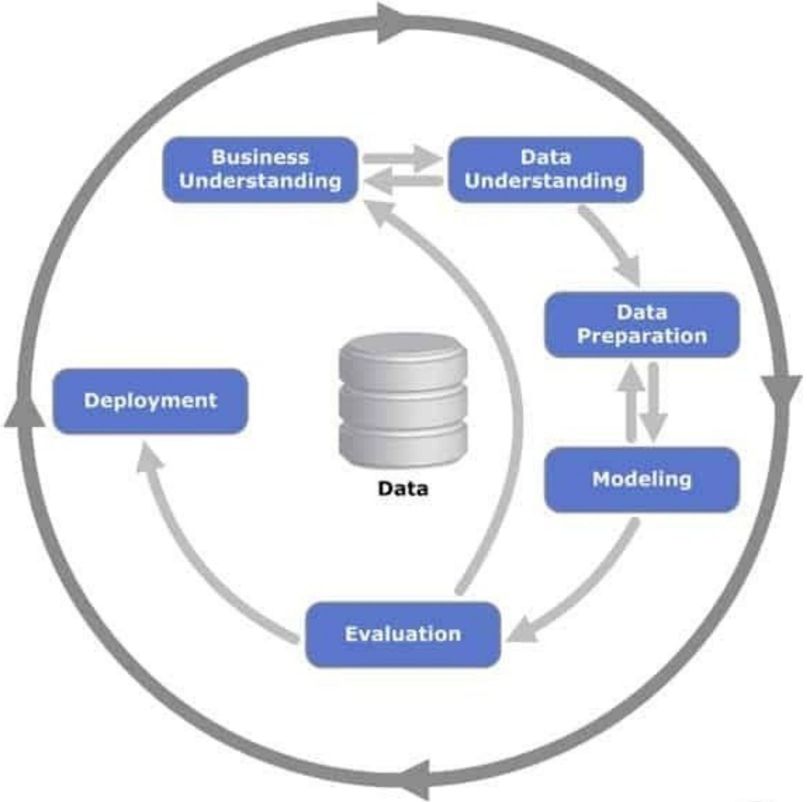
Data preparation – How do we organize the data for modeling?

Modeling – What modeling techniques should we apply?

Evaluation – Which model best meets the business objectives?

Deployment – How do stakeholders access the results?

Feedback - How is the model performing?

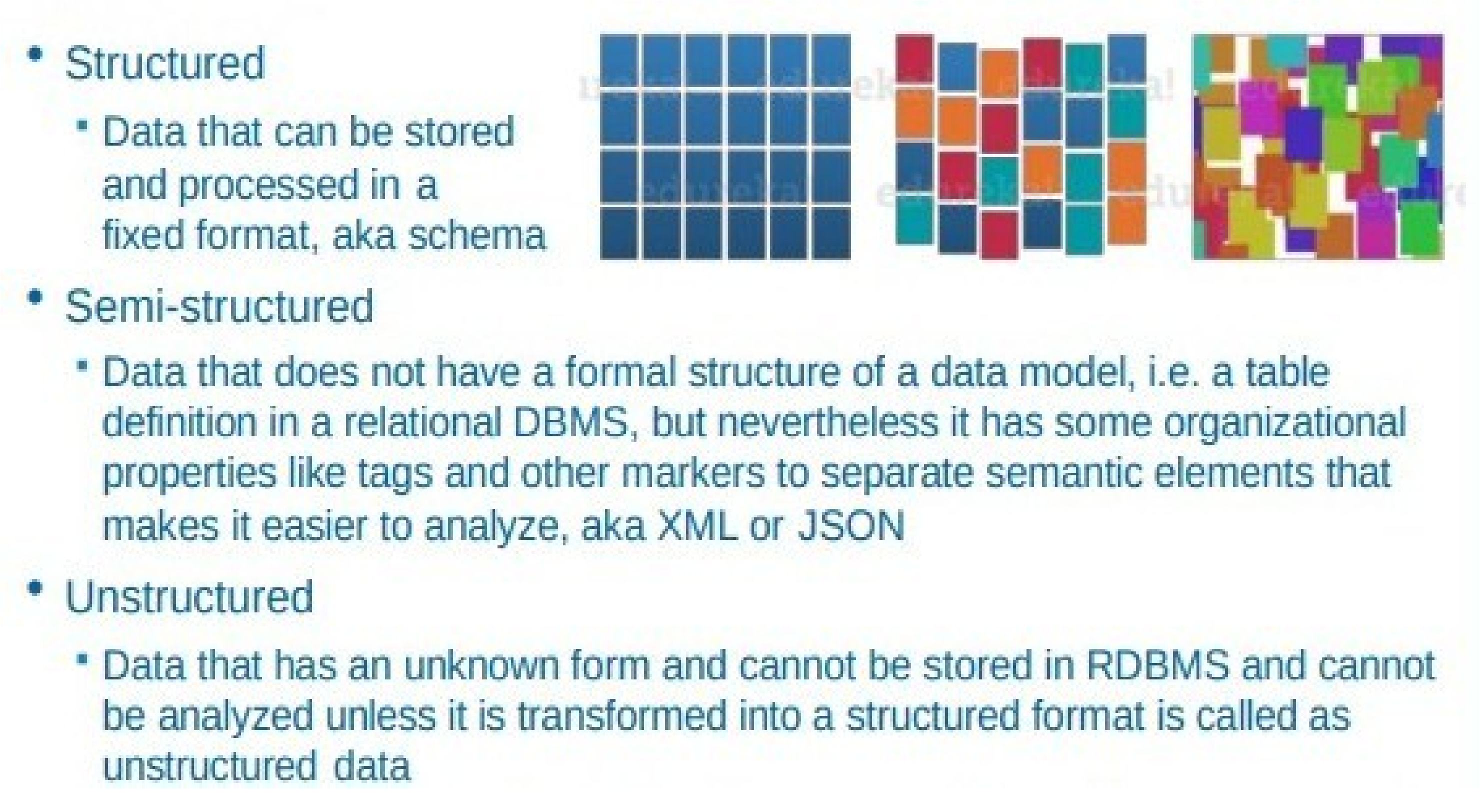


Read more..

https://www.datascience-pm.com/crisp-dm-2/

MODCOM
Institute of Technology

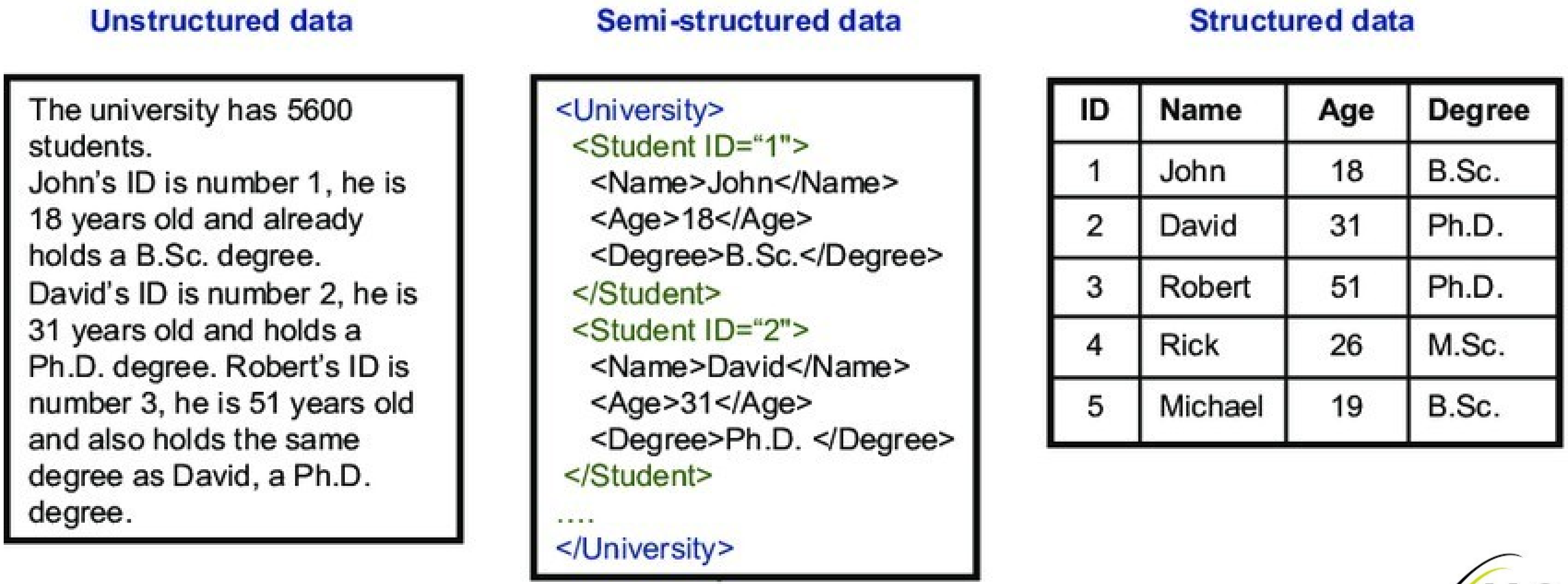
Forms of Data



- Structured
  - Data that can be stored and processed in a fixed format, aka schema
- Semi-structured
  - Data that does not have a formal structure of a data model, i.e. a table definition in a relational DBMS, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that makes it easier to analyze, aka XML or JSON
- Unstructured
  - Data that has an unknown form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format is called as unstructured data

MODCOM Institute of Technology

Forms of Data



Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
 <Student ID=“1">
   <Name>John</Name>
   <Age>18</Age>
   <Degree>B.Sc.</Degree>
 </Student>
 <Student ID=“2">
   <Name>David</Name>
   <Age>31</Age>
   <Degree>Ph.D. </Degree>
</Student>
….
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Data Science in Python

- Python libraries for Data Science
- Standard Python libraries useful for Data Analysis with Python are:
- Pandas - for data munging and preparation
- NumPy - abundance of useful features for operations on n-arrays and matrices in Python
- SciPy - library of software for engineering and science
- Matplotlib - tailored for the generation of simple and powerful visualizations
- Statsmodels - data exploration by using methods of estimation of statistical models
- scikit-learn - concise & consistent interface to the common ML algorithms
- Seaborn - visualization of statistical models; based & highly dependent on Matplotlib
- Bokeh - is aimed at interactive visualizations; independent of Matplotlib - main focus is interactivity, with presentation via modern browsers in the style of Data-Driven Documents
- Plotly - web-based toolbox for building visualizations, exposing APIs to Python, etc.
- Theano / TensorFlow / Keras
- NLTK - Natural Language Toolkit - tasks of symbolic & statistical NL processing
- Gensim / Scrapy

Data Science Tools/Notebooks

- There are several notebooks and tools used by data Scientists
- Apache Zeppelin
- Jupyter Notebook – in VS Code
- Watson Studio
- BeakerX
- Anaconda Distributions.

We will be using the popular Jupyter notebook;

- Can be found in anaconda distribution

- Also Jupyter Notebook in Google Colab
  Can be Used