Introduction to Data Science
20CS2031

Dr. Esther Daniel

Asso. Prof, Department of CSE

Course Objectives

- understand the key concepts of data science and its applications
- gain in-depth knowledge on statistical and machine learning techniques
- implement simple applications and analyze the results using relevant tools

Course Outcomes

- remember the key concepts of data science, data characteristics, its applications and the toolkit used by data scientists
- recall the mathematical concepts for descriptive and statistical analysis of the given dataset
- discuss on the principle operation of various supervised and unsupervised machine learning techniques
- select appropriate mathematical machine learning techniques for solving real world problems.
- apply the relevant techniques for implementing solutions to solve real world problems
- access the performance of prediction, classification and recommendation of machine learning techniques

Syllabus

Module 1: Basics of Data Science

Introduction to Data science: Applications of data science - Properties of Data: Exploring various dataset in different repositories - Tool Boxes for Data Scientist

Module 2: Understanding Data

Working with Data: Import, Select, Filter, Manipulate, sort, group, rearrange, rank and analyze the data for missing data values. Data visualization: Plot various plots for the given dataset

Module 3: Statistical Inference

Descriptive statistics, Exploratory Data Analysis: Calculate the mean, median, variance and standard deviation for the given small and large dataset, analyze the correlation between the variables in the dataset, Estimation, Hypothesis Testing: Formulate Null and Alternative hypothesis for real world use cases

Syllabus

Module 4: Supervised Learning

Introduction to machine learning, Types of machine learning, Linear, Multiple, Logistic and Polynomial Regression: Applications in transport, gaming and banking. KNN, Decision Trees: Applications in precision farming and smart building, calculate the performance metrics of regression and classification techniques.

Module 5: Unsupervised Learning

Clustering, Similarity and Distance measure, K means clustering: sentiment analysis. Agglomerative Clustering: gene expression data analysis, Graph based clustering techniques: smart city application

Module 6: Recommender System

Content Based Filtering, Collaborative Filtering: Developing a retail recommendation system, Hybrid Recommenders: Hotel recommendation system - Evaluating Recommenders

Text Books

- Laura Igual, Santi Seguí, “Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications”, Springer, 1st ed. 2017 Edition, ISBN 978-3-319-50016-4e-ISBN 978-3-319-50017-1.
- Steven S. Skiena, “The Data Science Design Manual”, Springer, 1st ed. 2017, ISBN 978-3-319-55443-3.

Reference Books/ links:

- Steven Cooper, “Data Science from Scratch: The #1 Data Science Guide for Everything A Data Scientist Needs to Know: Python, Linear Algebra, Statistics, Coding, Applications, Neural Networks, and Decision Tree”, 2018, ISBN-10: 1723141208, ISBN-13: 978-1723141201.
- Cathy O’Neil and Rachel Schutt, “Doing Data Science, Straight Talk from The Frontline”. O’Reilly, 2014. ISBN: 978-1-449-35865-5.
- Sinan Ozdemir, “Principles of Data Science”, Packt Publishing Limited,2016 ISBN10: 9781785887918, ISBN-13: 978-1785887918
- V. K. Jain, “Data Science and Analytics”,Khanna PublishingFirst edition, 2018, ISBN10: 9789386173676, ISBN-13: 978-9386173676.
- Jake VanderPlas, “Python Data Science Handbook: Essential Tools for Working with Data”, Shroff/O'Reilly,2016, ISBN-10: 9352134915, ISBN-13: 978-9352134915
- Peter Morgan, “Data Science from Scratch with Python: Step-By-Step Guide”, Createspace Independent Publishing Platform, 2nd edition, 2018, ISBN-10: 1726020681, ISBN-13: 978-1726020688.

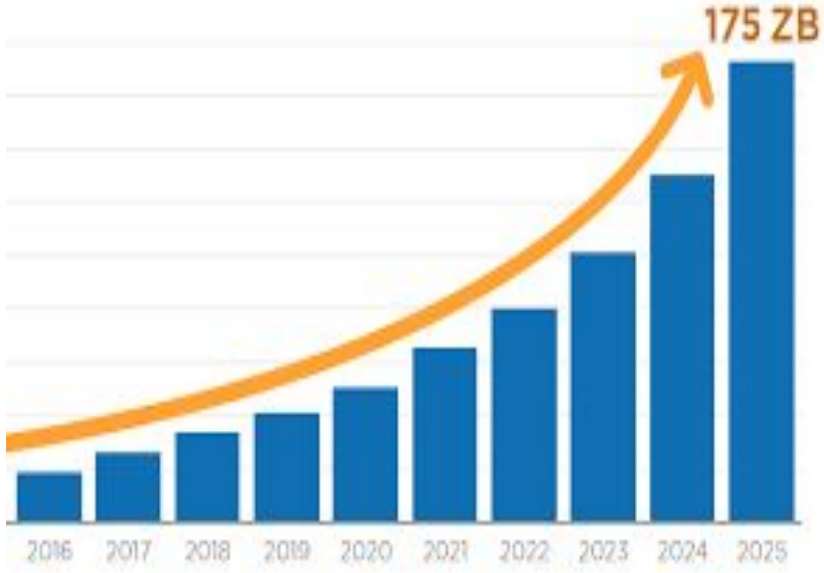
DATA & Types of DATA


quantities
numbers
information
What Is Data?
facts
graphs
observations
measurement

DATA GROWTH

1 exabyte (EB) = 1,000,000,000,000,000,000 bytes

One zettabyte is equal to 1, 000 exabytes or 1, 000, 000, 000, 000, 000, 000, 000 bytes

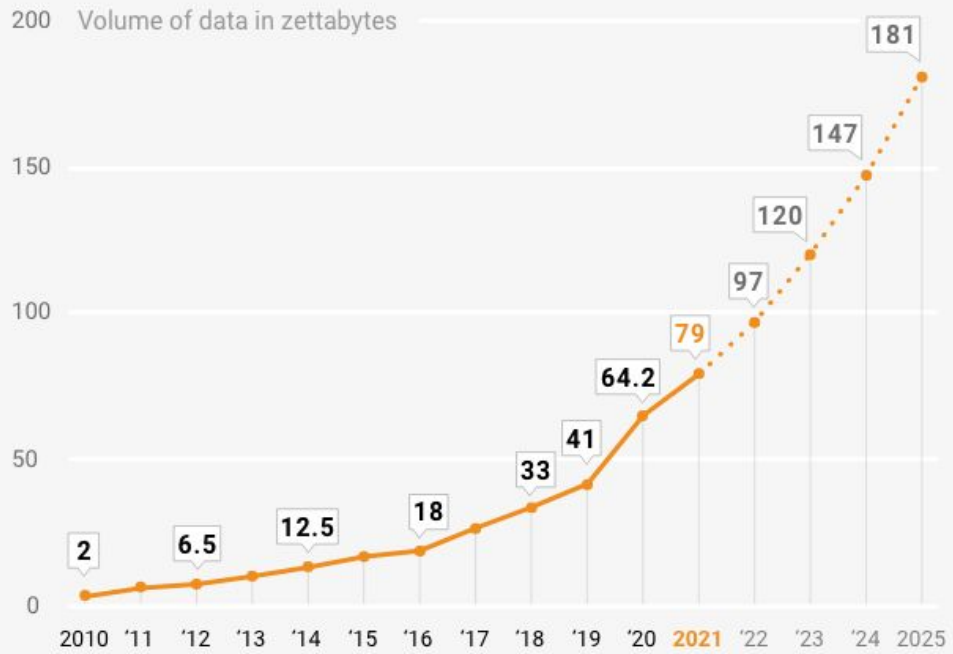


http://www.infosysblogs.com/brandedge/2013/04/20130419infographic.html

https://studentforce.wordpress.com/2013/09/21/umuc-big-data-revolution-is-here/

Volume of data created, captured, copied, and consumed worldwide

The volume of data generated, consumed, copied, and stored is projected to exceed 180 zettabytes by 2025



Source: statista.com

firstsiteguide.com

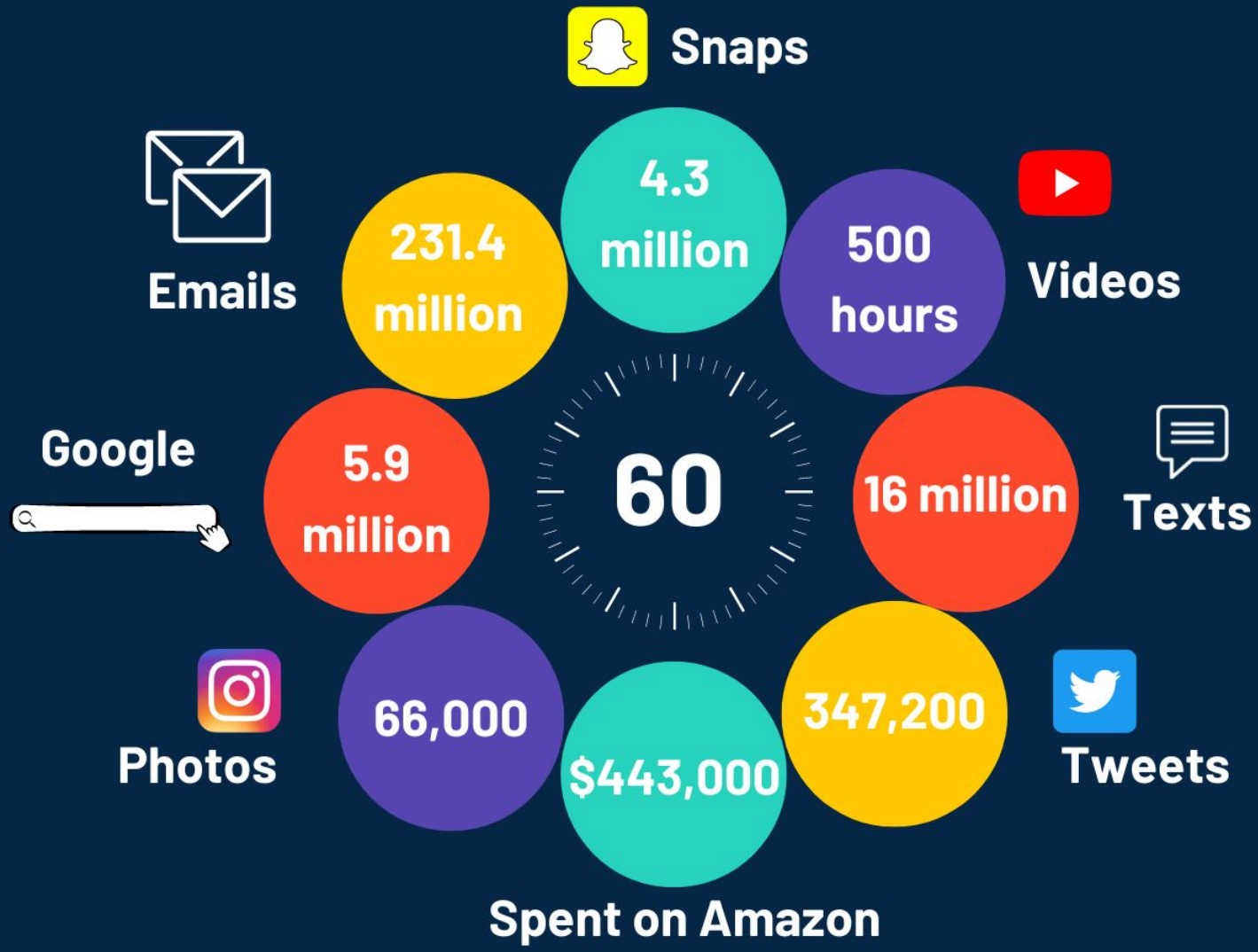
(credit Roy Williams, Center for Advanced Computing Research at the California Institute of Technology).

- Kilo- means 1,000; a Kilobyte is one thousand bytes.
- Mega- means 1,000,000; a Megabyte is a million bytes.
- Giga- means 1,000,000,000; a Gigabyte is a billion bytes.
- Tera- means 1,000,000,000,000; a Terabyte is a trillion bytes.
- Peta- means 1,000,000,000,000,000; a Petabyte is 1,000 Terabytes.
- Exa- means 1,000,000,000,000,000,000; an Exabyte is 1,000 Petabytes.
- Zetta- means 1,000,000,000,000,000,000,000; a Zettabyte is 1,000 Exabytes.
- Yotta- means 1,000,000,000,000,000,000,000,000; a Yottabyte is 1,000 Zettabytes.

Examples of Data Volumes

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	a paragraph of a text document
Megabytes (MB)	1,000 Kilobytes	a small novel
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5th Symphony
Terabytes (TB)	1,000 Gigabytes	all the X-rays in a large hospital
Petabytes (PB)	1,000 Terabytes	half the contents of all US academic research libraries
Exabytes (EB)	1,000 Petabytes	about one fifth of the words people have ever spoken
Zettabytes (ZB)	1,000 Exabytes	as much information as there are grains of sand on all the world's beaches
Yottabytes (YB)	1,000 Zettabytes	as much information as there are atoms in 7,000 human bodies

Data We Create Online in 60 Seconds



G2.com

Source: DOMO

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

320bn

emails to be sent each day by 2021

294bn

billion emails are sent

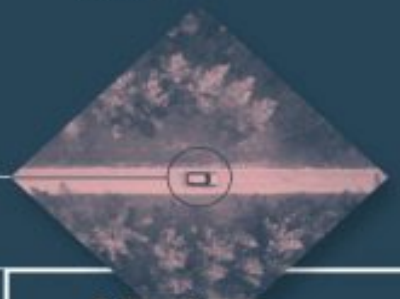
Radicati Group

306bn

emails to be sent each day by 2020

3.9bn

people use emails



4TB

of data produced by a connected car

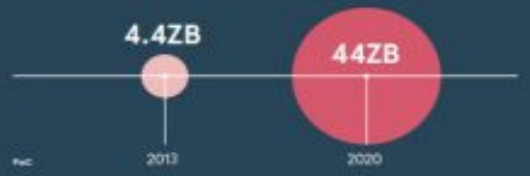
Intel

65bn

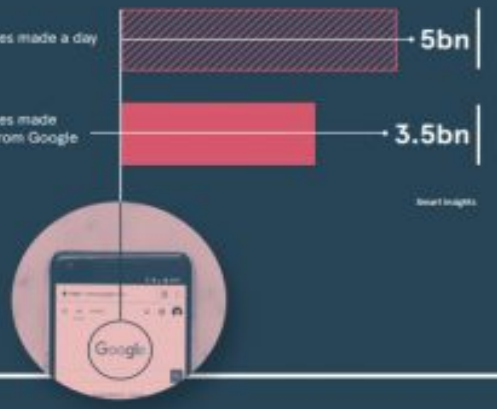
messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

ACCUMULATED DIGITAL UNIVERSE OF DATA



IDC

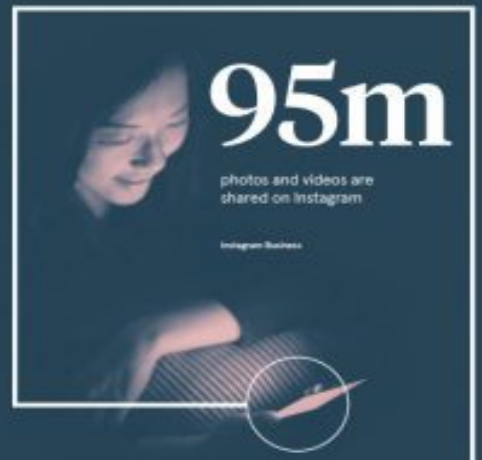


Smart Insights

463EB

of data will be created every day by 2025

IDC



95m

photos and videos are shared on Instagram

Instagram Business

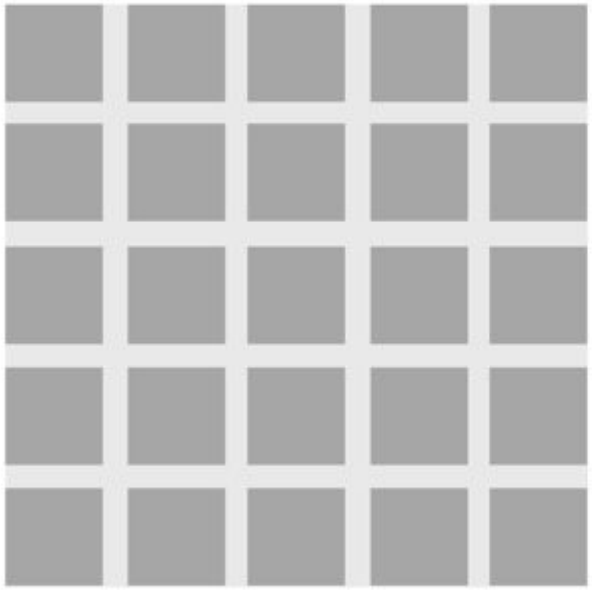
28PB

to be generated from wearable devices by 2020

Statista

RACONTEUR

Structured data



Database, CRM, ERP

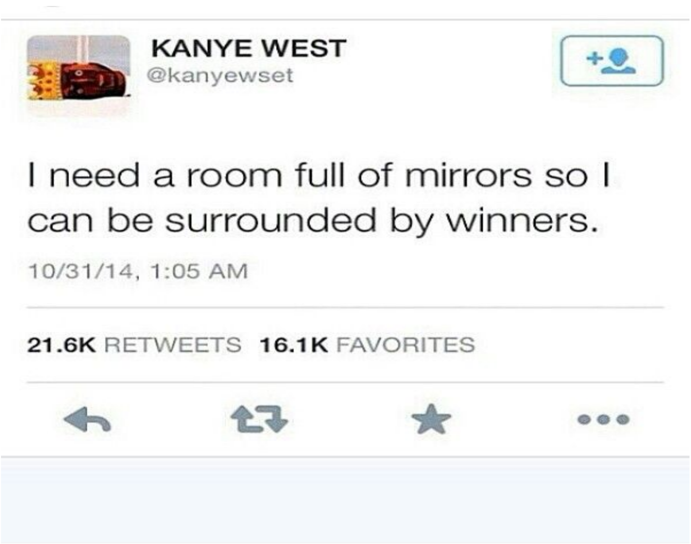
Unstructured data

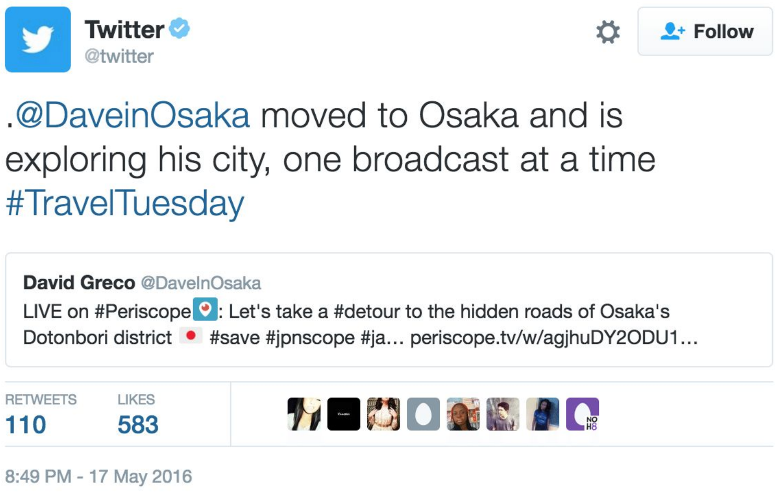
Text, audio, videos

Structured Data

0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

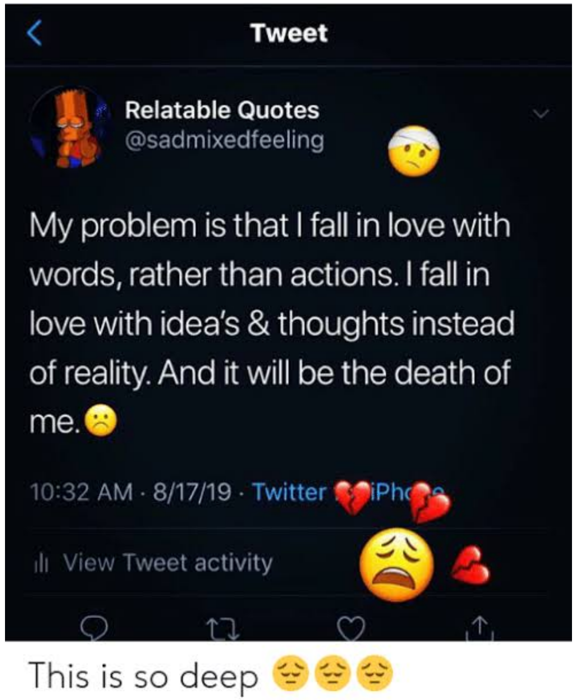
Unstructured Data


KANYE WEST
@kanyewset
I need a room full of mirrors so I can be surrounded by winners.
10/31/14, 1:05 AM
21.6K RETWEETS 16.1K FAVORITES

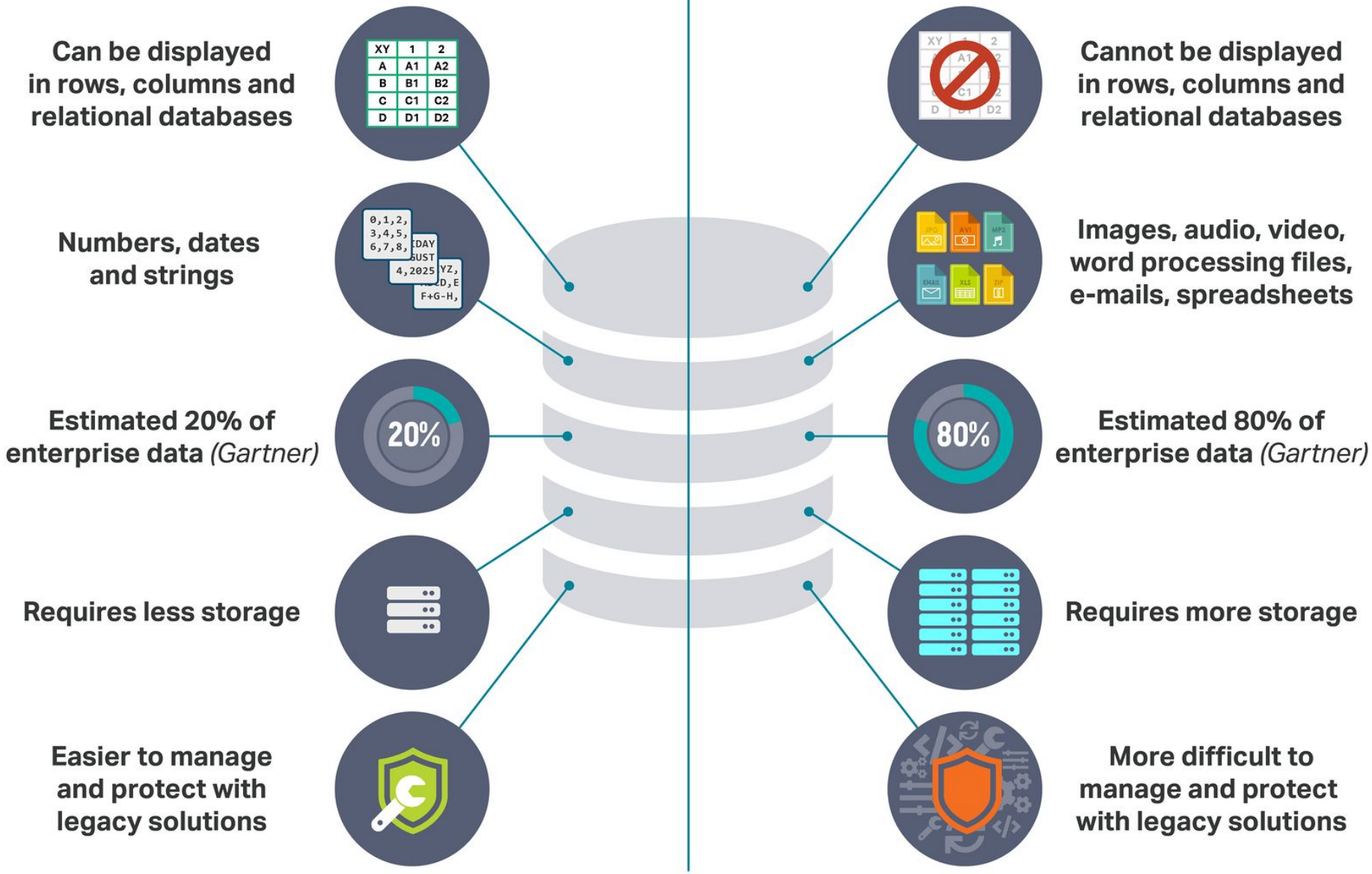

Twitter
@twitter
Follow
.@DaveinOsaka moved to Osaka and is exploring his city, one broadcast at a time #TravelTuesday
David Greco @DaveInOsaka
LIVE on #Periscope: Let's take a #detour to the hidden roads of Osaka's Dotonbori district #save #jpnscope #ja... periscope.tv/w/agjhuDY2ODU1...
RETWEETS 110
LIKES 583
8:49 PM - 17 May 2016


●●●○○ TFW
1:12 PM
69%
Tweet
Sassy
@sassynotez
If you think I'm ignoring you... I am. My phones in my hand 24/7
11/4/15, 3:17 PM
546 RETWEETS 400 LIKES
Destinee Barber @23Destinee 11/5/15
@sassynotez @lostworry @fondie922
Reply to Sassy
Home
Notifications
Moments
Messages
Me


140
Donald J. Trump
@realDonaldTrump
Me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me
Donald J. Trump
@realDonaldTrump
Me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me me
280


Tweet
Relatable Quotes
@sadmixedfeeling
My problem is that I fall in love with words, rather than actions. I fall in love with idea's & thoughts instead of reality. And it will be the death of me.
10:32 AM · 8/17/19 · Twitter
View Tweet activity
This is so deep

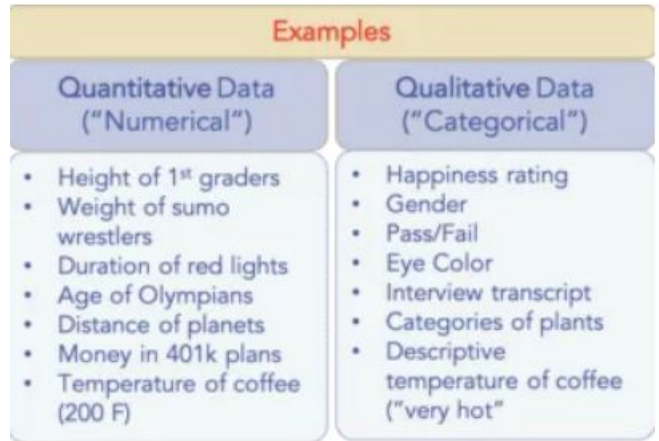
Structured Data vs Unstructured Data



Categorical vs Quantitative Data

Karunya

Categorical Data	Quantitative Data
• Deals with descriptions.	• Deals with numbers.
• Data can be observed but not measured.	•Data which can be measured.
•Colors, textures, smells, tastes, appearance, beauty, etc.	• Length, height, area, volume, weight, speed, time, cost, age, etc.
• Categorical → Description	• Quantitative → Quantity



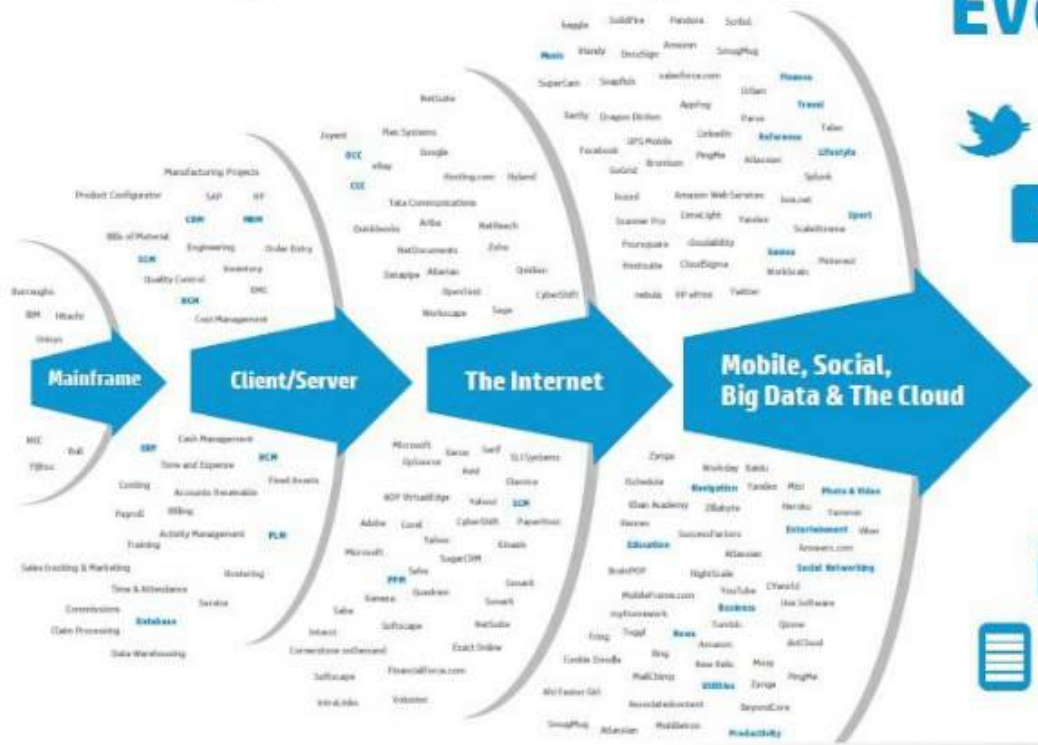
Name	Type	Appearance	Examples
Discrete	Numeric	Integers	pairs of shoes, books owned, children
Continuous	Numeric	Decimals	Time spent, speed, weight
Nominal	Categorical	Words no order	Race, shoe color, car type
Ordinal	Categorical	Words with order	Education, happiness

Big Data


VOLUME
Huge amount of data
VERACITY
Inconsistencies and uncertainty in data
VARIETY
Different formats of data from various sources
BIG DATA
VELOCITY
High speed of accumulation of data
VALUE
Extract useful data

Changes in IT Systems

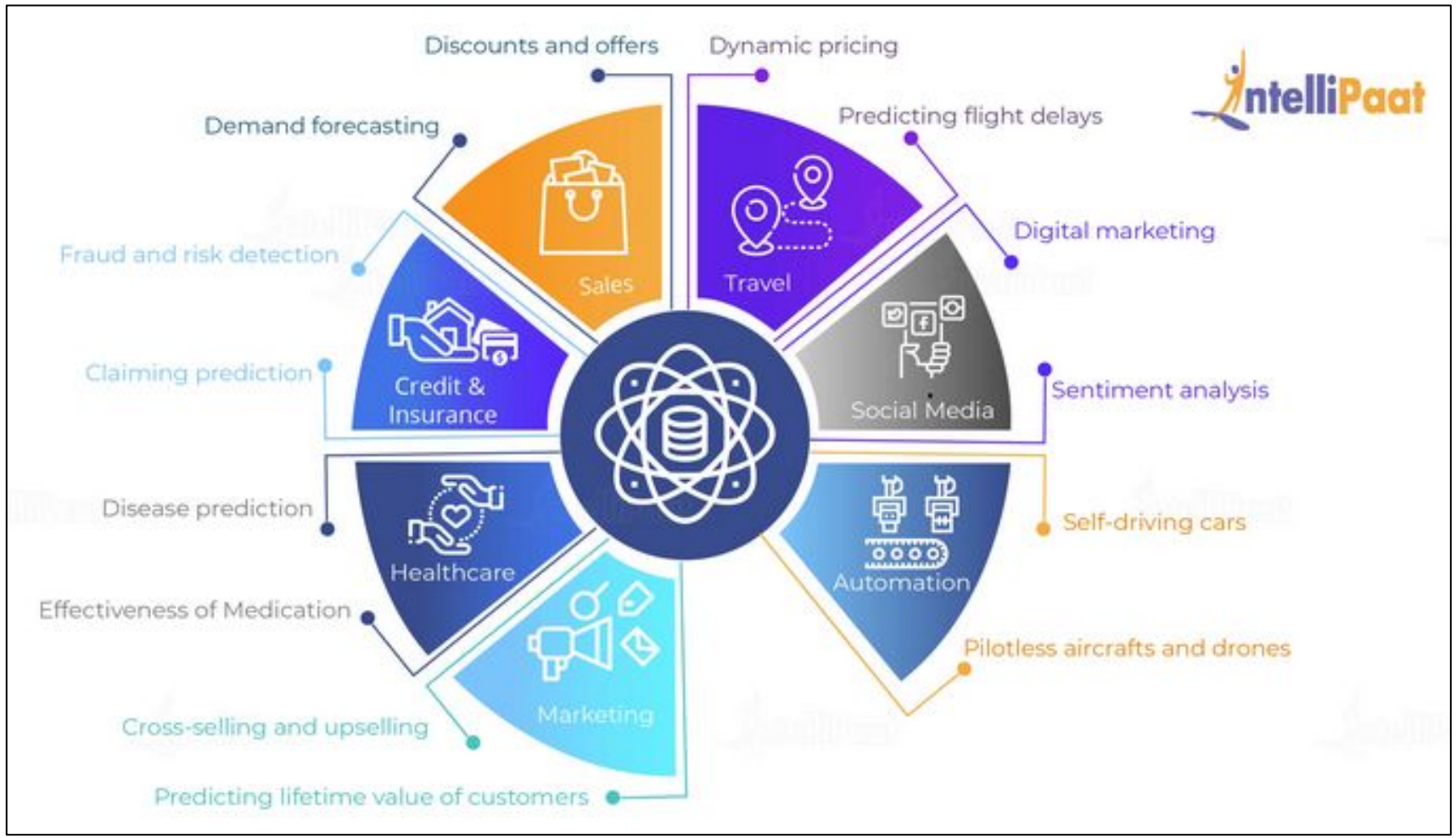
A new style of IT emerging



Every 60 seconds

- **98,000+** tweets
- **695,000** status updates
- **11million** instant messages
- **698,445** Google searches
- **168 million+** emails sent
- **1,820TB** of data created
- **217** new mobile web users

Big Data challenges


Discounts and offers
Dynamic pricing
Demand forecasting
Predicting flight delays
Fraud and risk detection
Digital marketing
Claiming prediction
Sentiment analysis
Disease prediction
Self-driving cars
Effectiveness of Medication
Pilotless aircrafts and drones
Cross-selling and upselling
Predicting lifetime value of customers
Sales
Travel
Credit & Insurance
Social Media
Healthcare
Automation
Marketing
IntelliPaat

DATA SCIENCE

Why Data Science?

- We hear a lot about how artificial intelligence and machine learning are going to change the world and how the Internet of Things will make everyone's life easier.

- But in reality, the one thing that underpins all of these **revolutionary** technologies is "data".

- In a world that is approaching a digital space, organizations deal with an immeasurable amount of structured and unstructured data every day. This data can be collected from various possible sources, out of which the most common sources are the self-directed interviews, surveys, observations, and experiments.

- The data can also be collected from other sources such as research done by various researchers, online surveys, various government organizations, social media accounts, etc.

- Well, this data is known as **Big Data**.

Data Science-Definition

- **Data Science** is a “detailed study of the flow of information from the colossal amounts of data present in an organization. Data Science can be simply defined as the process of analyzing data for making a decision/marketing decision”.
  - Involves obtaining meaningful insights from new and unstructured data which can be processed through analytical, programming and business skills.
  - Deals with nature of data, types of data, visualizing, analyzing, modeling, Machine Learning and Neural Network etc.
  - It brings together a lot of skills like Machine learning, Statistics, Mathematics and business domain knowledge.

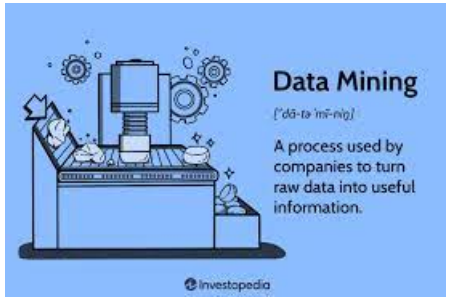
Way of Data Science…

Statistics

Data Mining

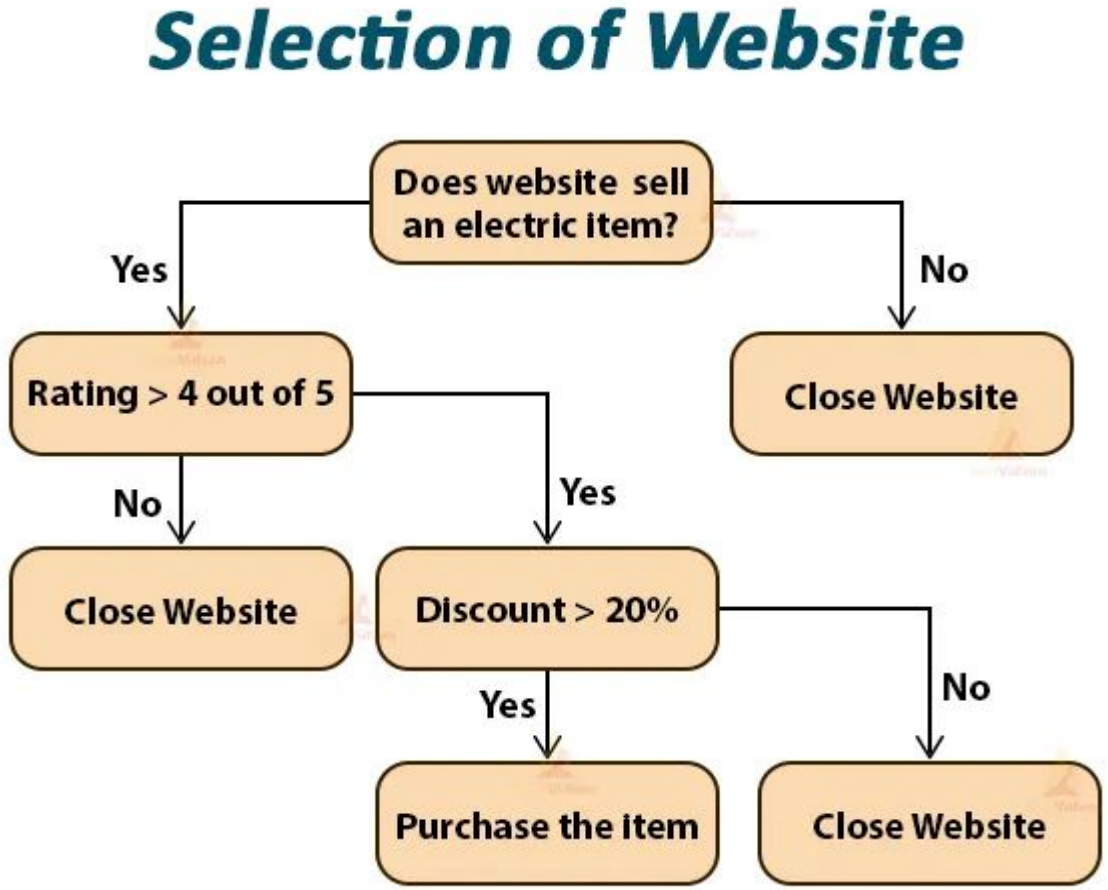
Predictive Analysis

Data Science



Example

- Let's consider an **example**, suppose you have decided to buy electronic items for your home online. So you will have to take a sequence of decisions for buying the items.

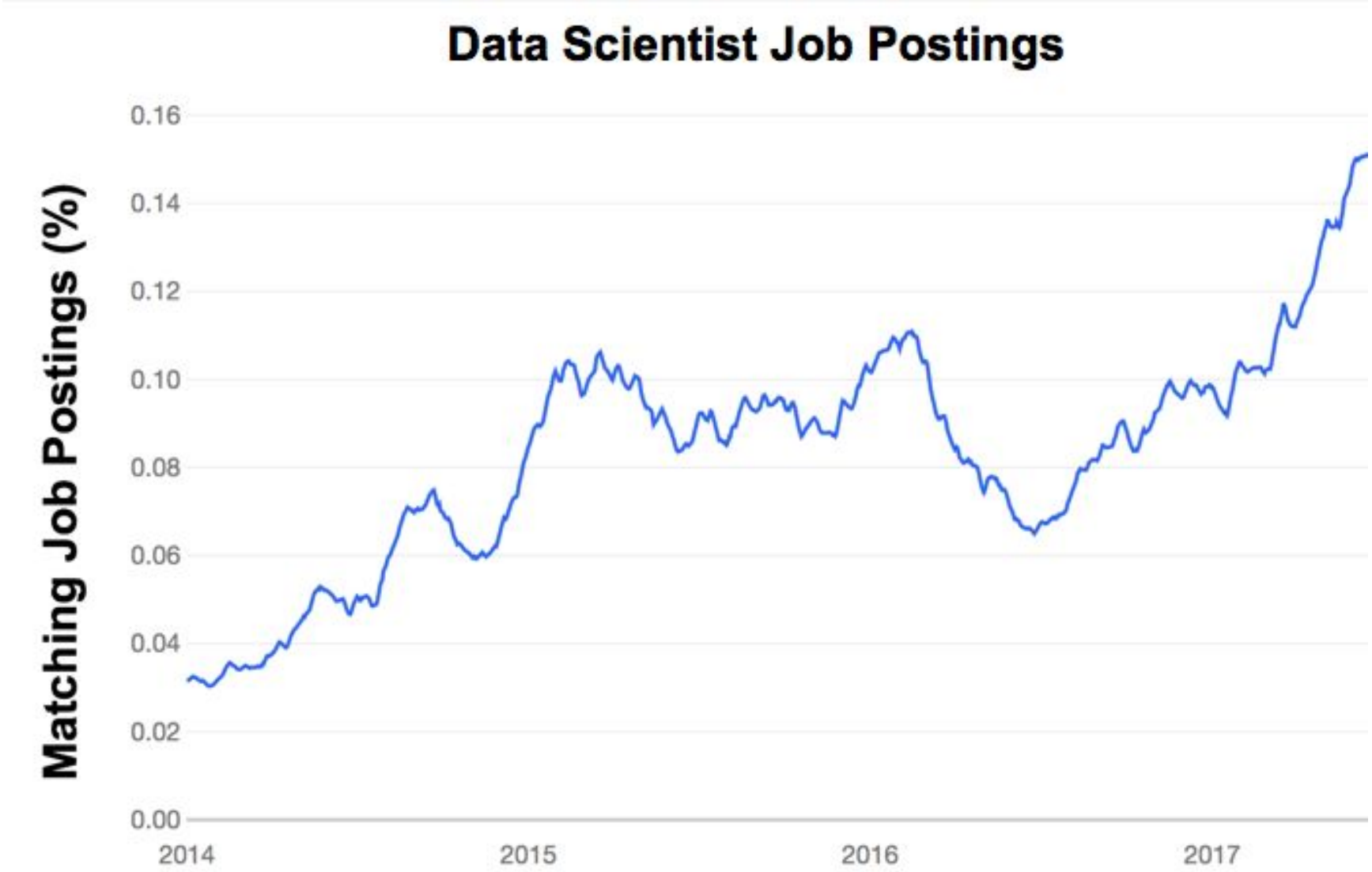


Data Scientist

- A Data Scientist is a professional who collects and organizes the data and then analyzes it for meaningful and actionable insights by using various statistical approaches.

- The job of a Data Scientist requires knowledge of **math**, **science**, **statistics**, and **computer**.

SCOPE OF DATA SCIENCE


Data Scientist Job Postings
Matching Job Postings (%)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2014
2015
2016
2017

1.5 MILLION

Career Opportunities for managers capable of reaping actionable insights from big data

Top Industries:

Companies that hire Big Data Hadoop Architect -

DELL, LinkedIn, accenture, IBM, CISCO, facebook, ORACLE, amazon

Gita Gopinath | CHIEF ECONOMIST, IMF

TCS doubles pay for fresh hires with new-age skills

Offers 1,000 New Recruits ₹6.5L Instead Of Usual ₹3.5L

Avik Das & Shilpa Phadnis | TNN

Bengaluru: Tata Consultancy Services (TCS) has offered about 1,000 freshers with new-age digital skills almost double the salary it normally offers those coming out of campuses.

While the entry level salary of Indian engineers in the IT industry has been stuck at about Rs 3.5 lakh per annum for the past decade, TCS is offering those with digital skills a starting salary of about Rs 6.5 lakh.

The selection of these candidates was based on their clearing a test focused on new digital areas. From this year, candidates who perform exceedingly well in its online National Qualifier Test (NQT), about which TOI reported recently, will also get an opportunity to take a shot at that examination.

TCS, one of the biggest recruiters from Indian engineering colleges, usually visits its accredited colleges to conduct a test followed by an interview. This process is going to be largely replaced by the NQT. "People who have done well in the NQT will get a chance to write another test for the digital talent pool, and if they clear and go through the interview, then they will get into the digital pool and their compensation will be differentiated," Ajoy Mukherjee, executive VP and head of global human resources, told TOI. The test involves programming with a higher degree of difficulty compared to the NQT. The test is longer and requires good coding skills.

TCS's move shows the lengths to which companies are ready to go to hire good talent. Employees armed with skills in the fields of machine learning, AI, data analysis are getting better appraisals across levels. Such specialists are few and in much demand. A recent report by LinkedIn said machine learning engineer, application development analyst, back-end developer, full stack engineer and data scientist constitute the top five jobs in India, with demand for them growing by 43, 32, 23, 18 and 14 times respectively in the last five years. Companies often have to spend significantly on training employees to acquire such skills.

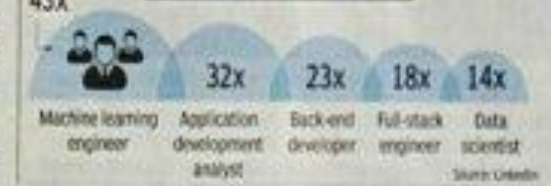
TCS's NQT, launched this year, has enabled it to reach out to a far larger student talent base, as also complete the recruitment process in 3-4 weeks, compared to the 3-4 months it took under the traditional process.

Mukherjee would not comment on the number of people who would be hired this year as part of the digital pool, saying that the process is still on, but added that the numbers would be nearly the same or higher compared to the 1,000 it did in the last academic year.

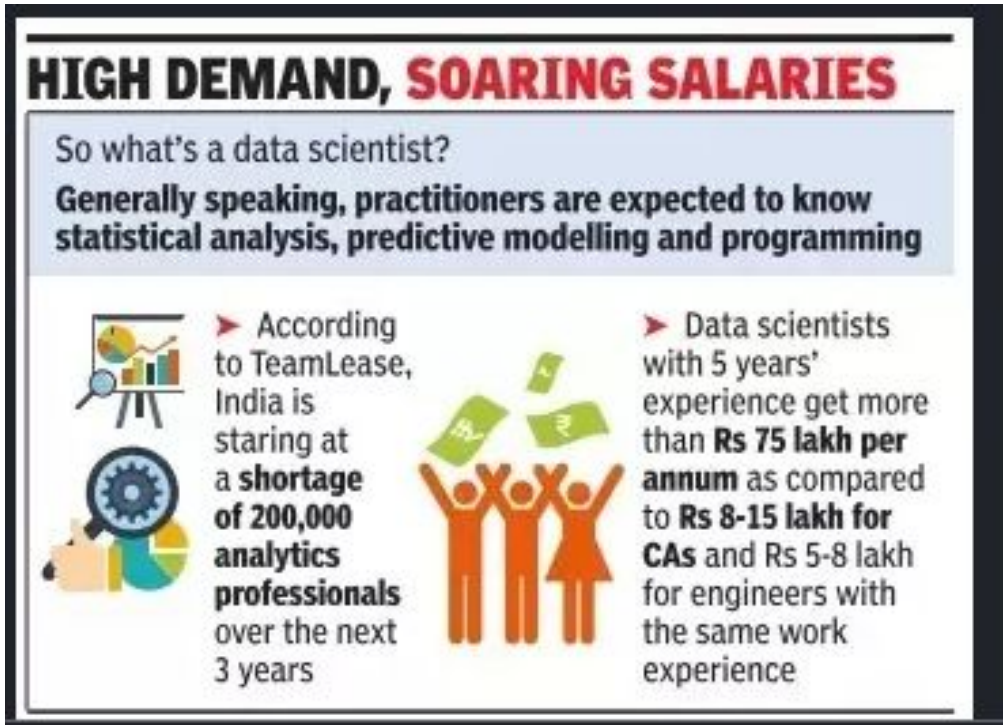
Apart from this process, TCS also does select hiring from the IITs and NITs, where it offers compensation packages that are even higher.

INDIA IT JOBS IN GREAT DEMAND

Demand growth over 5 years



Source: LinkedIn



Data Flair

Why Learn Data Science?

- Fuel of 21st Century
- Problem of Demand & Supply
- A Lucrative Career
- Data Science is Changing the World
- Data Science is Future

Karunya

Important Data Scientist Job Roles

- Data Scientist
- Data Engineer
- Data Analyst
- Statistician
- Data Architect
- Data Admin
- Business Analyst
- Data/Analytics Manager

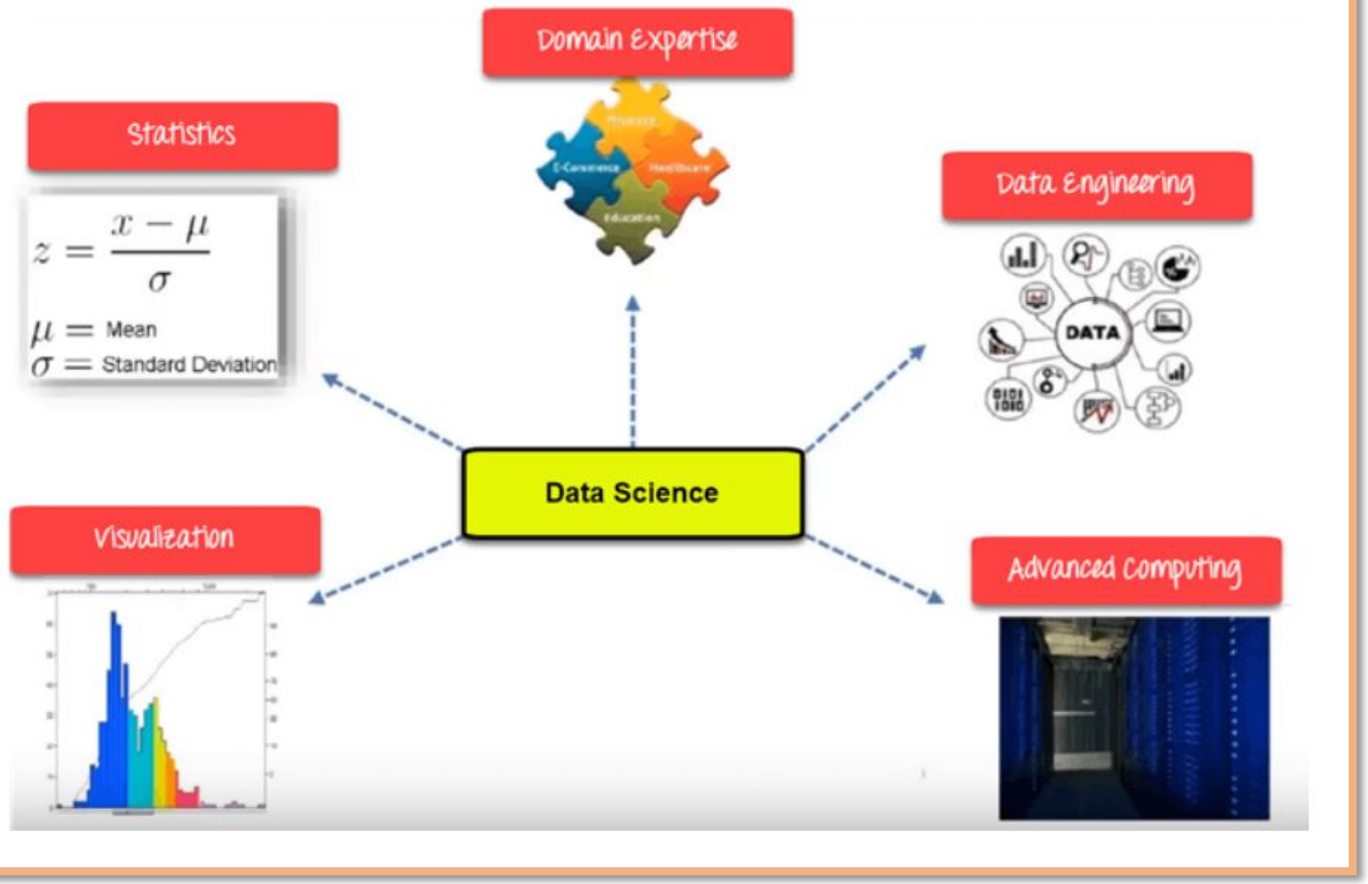
Data Scientist

- Manages enormous amounts of data to come up with compelling business visions by using various tools, techniques, methodologies, algorithms, etc.

Data Engineer

- working with large amounts of data. He develops, constructs, tests, and maintains architectures like large scale processing system and databases.

Data Science Components



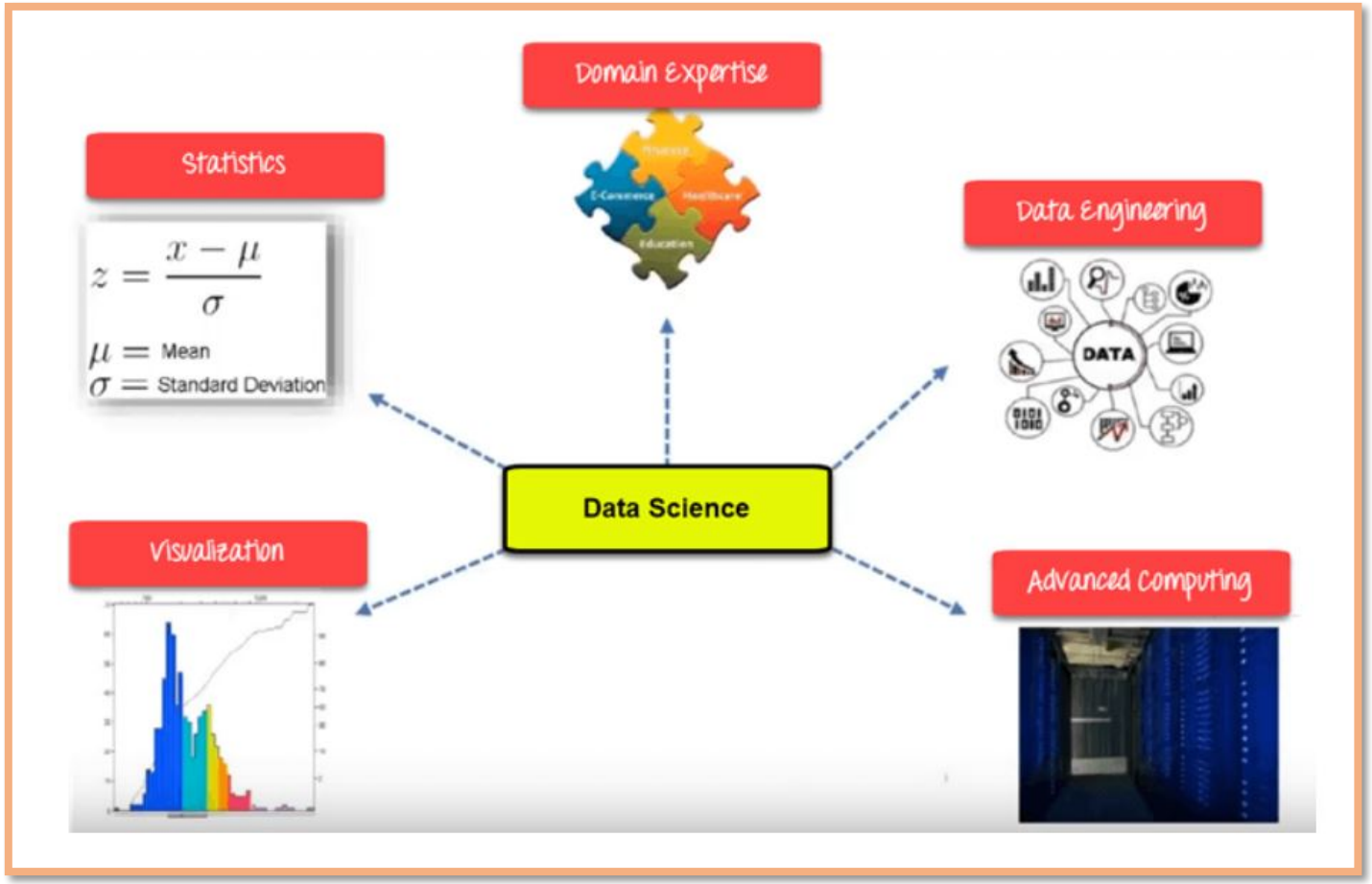
Statistics:

Statistics is the most critical unit in Data science. It is the method or science of collecting and analyzing numerical data in large quantities to get useful insights.

Visualization:

Visualization technique helps you to access huge amounts of data in easy to understand and digestible visuals.

Data Science Components

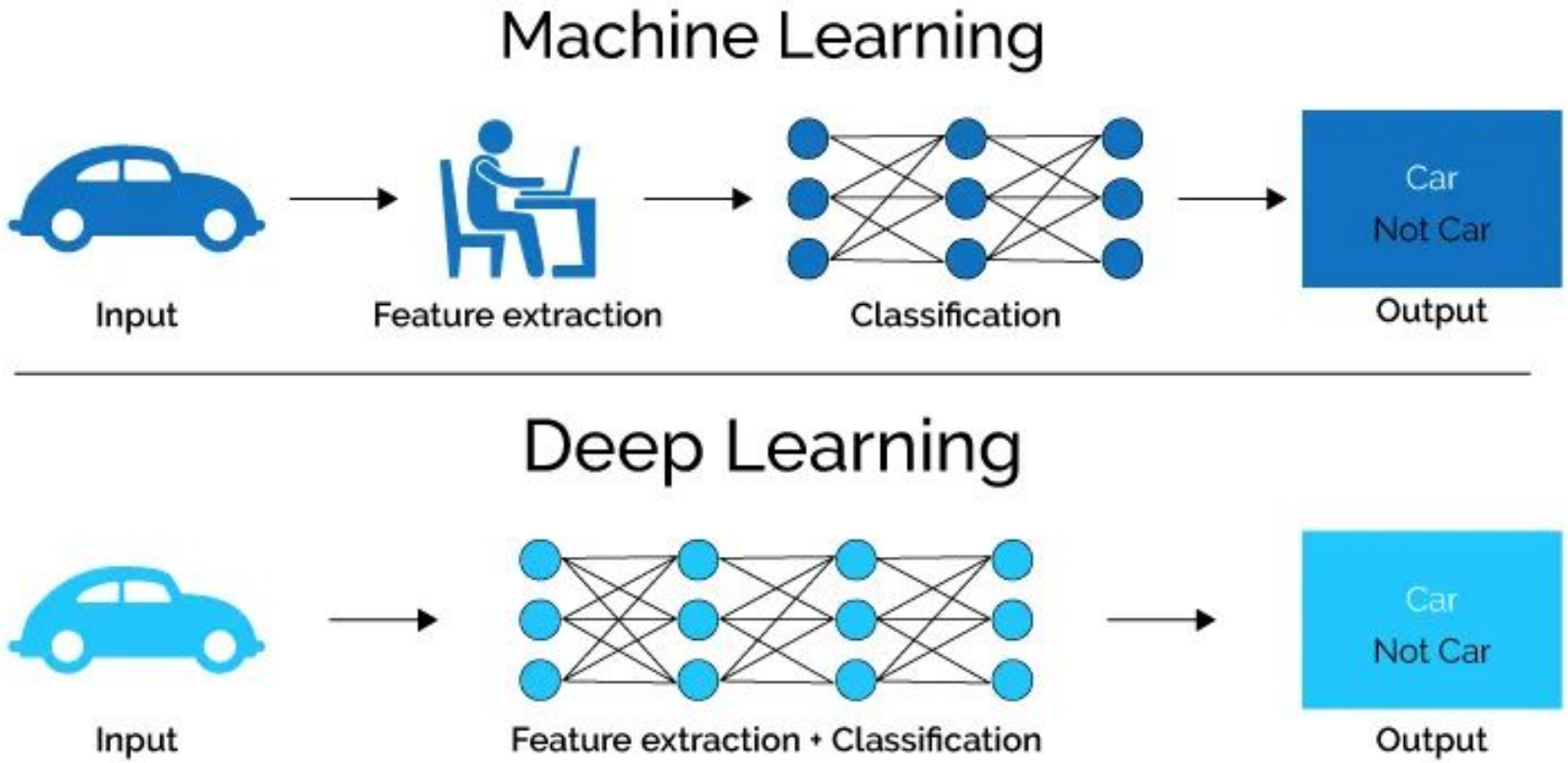


Machine Learning:

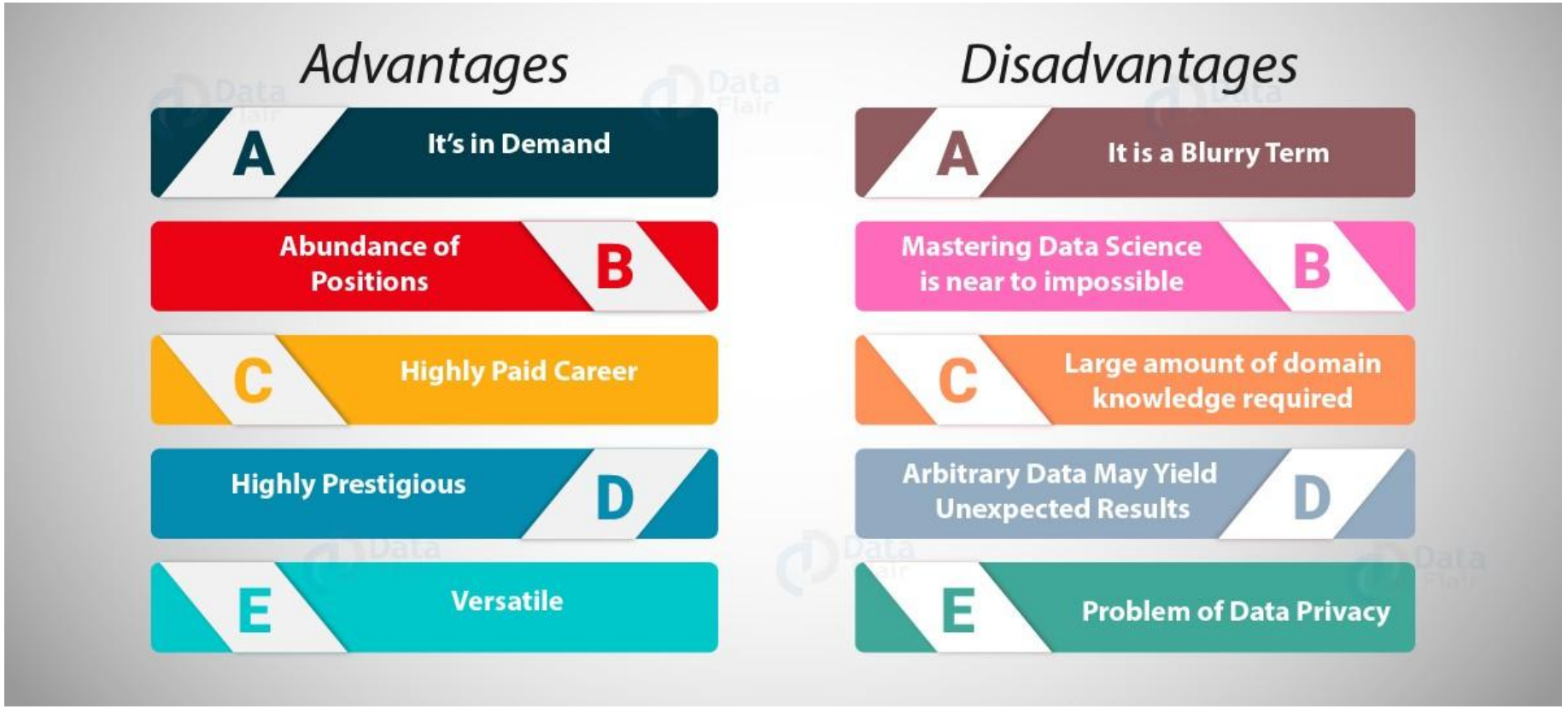
Machine Learning explores the building and study of algorithms which learn to make predictions about unforeseen/future data.

Deep Learning:

Deep Learning method is new machine learning research where the algorithm selects the analysis model to follow.


Machine Learning
Input
Feature extraction
Classification
Car
Not Car
Output
Deep Learning
Input
Feature extraction + Classification
Car
Not Car
Output

Advantages and Disadvantages



Applications of Data science

Internet Search:

- Google search use Data science technology to search a specific result within a fraction of a second

Recommendation Systems:

- To create a recommendation system. Example, "suggested friends" on Facebook or suggested videos" on YouTube, everything is done with the help of Data Science.

Image & Speech Recognition:

- Speech recognizes system like Siri, Google assistant, Alexa runs on the technique of Data science. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.

Applications of Data science

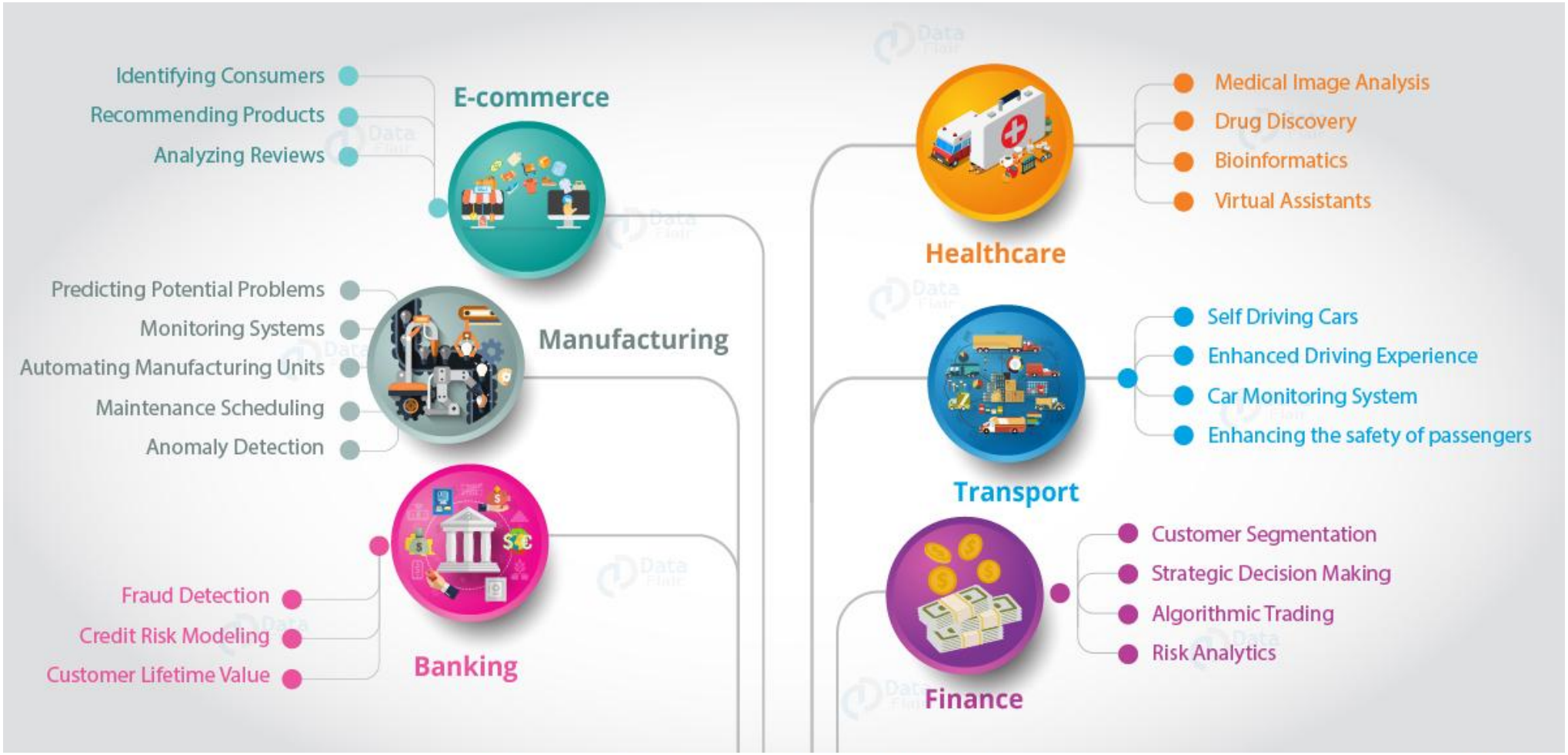
Gaming world:

- EA Sports, Sony, Nintendo, are using Data science technology. This enhances your gaming experience. Games are now developed using Machine Learning technique. It can update itself when you move to higher levels.

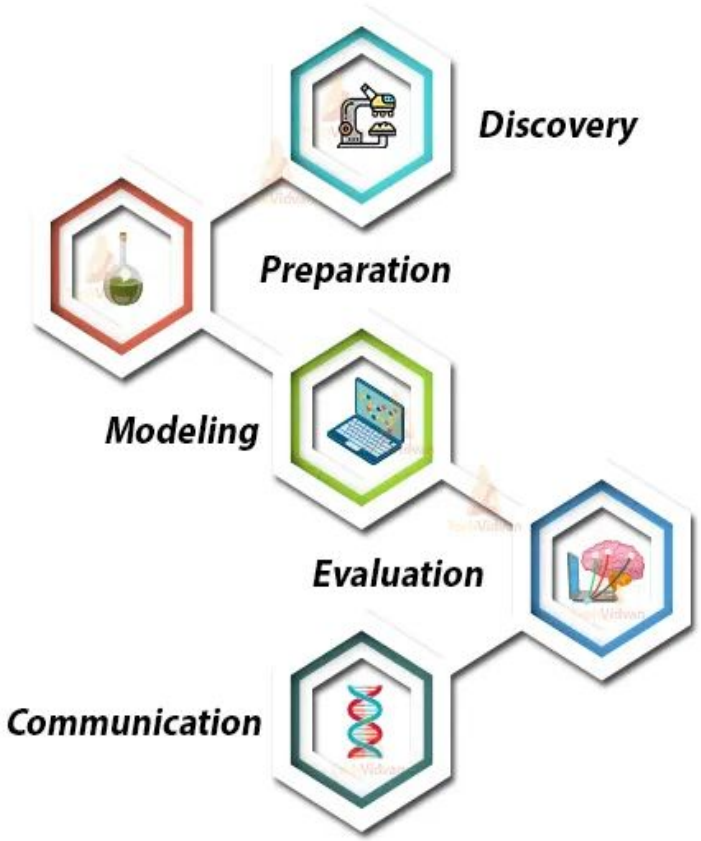
Online Price Comparison:

- PriceRunner, Junglee, Shopzilla work on the Data science mechanism. Here, data is fetched from the relevant websites using APIs.

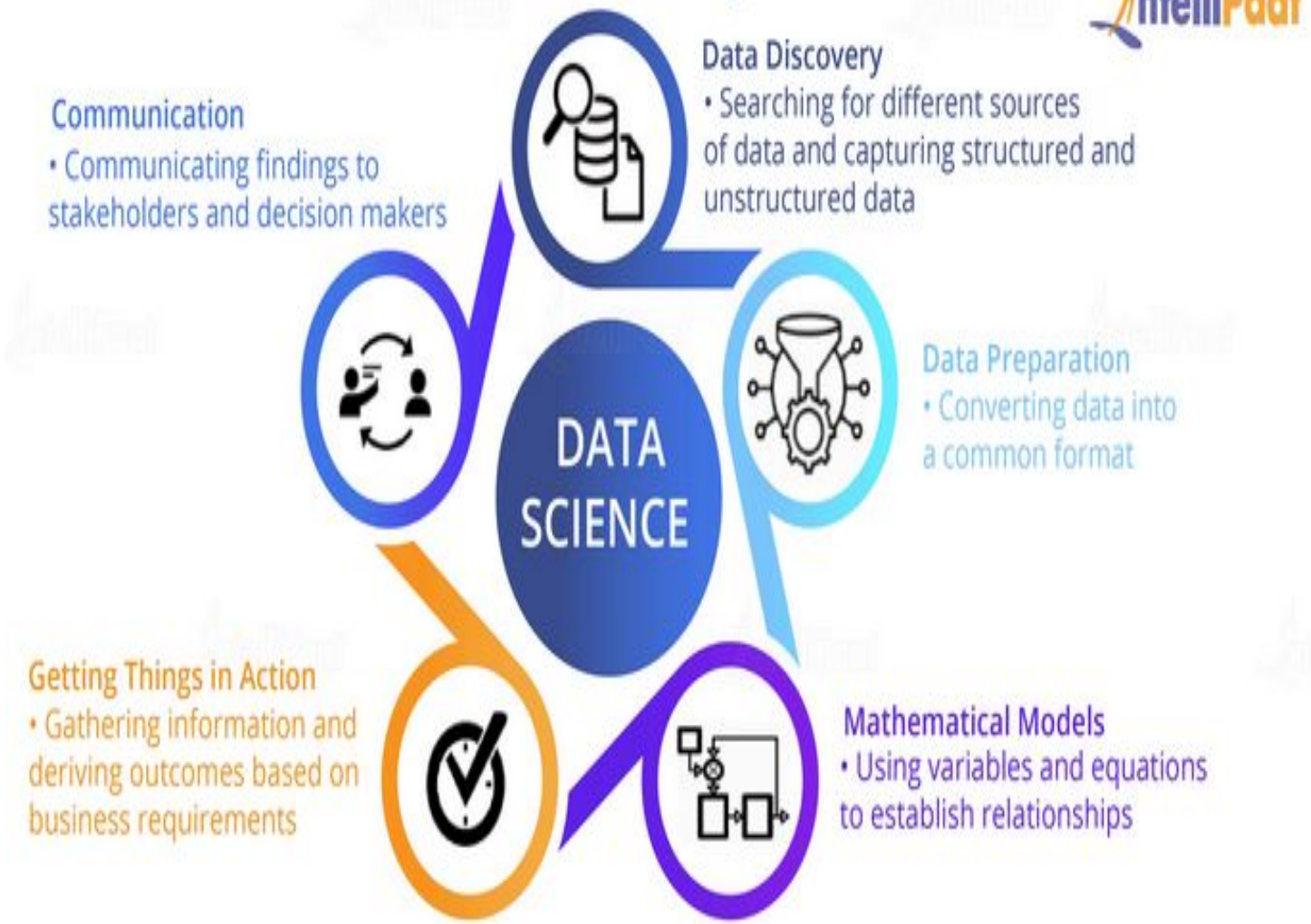
Applications of Data Science



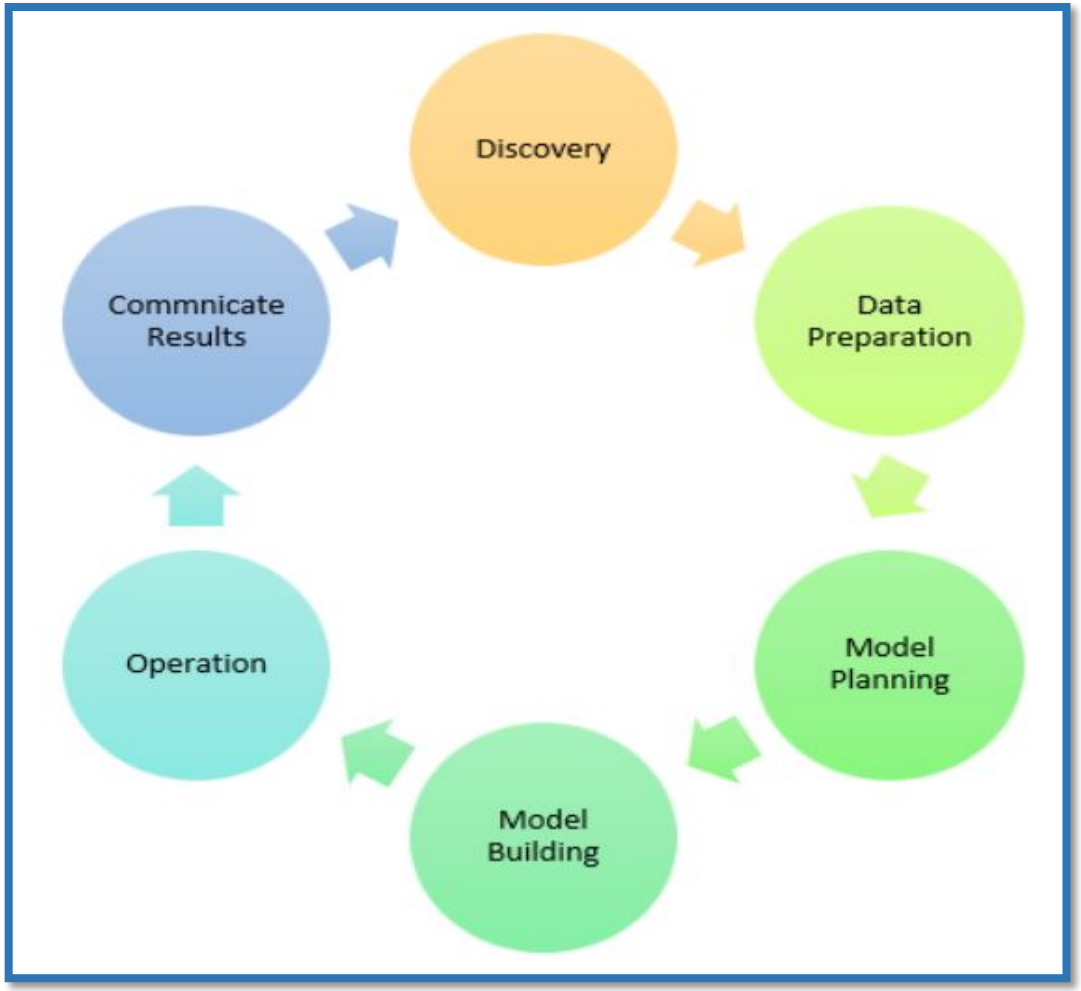
Data Science Life Cycle



Data Science Life Cycle



Data Science Process

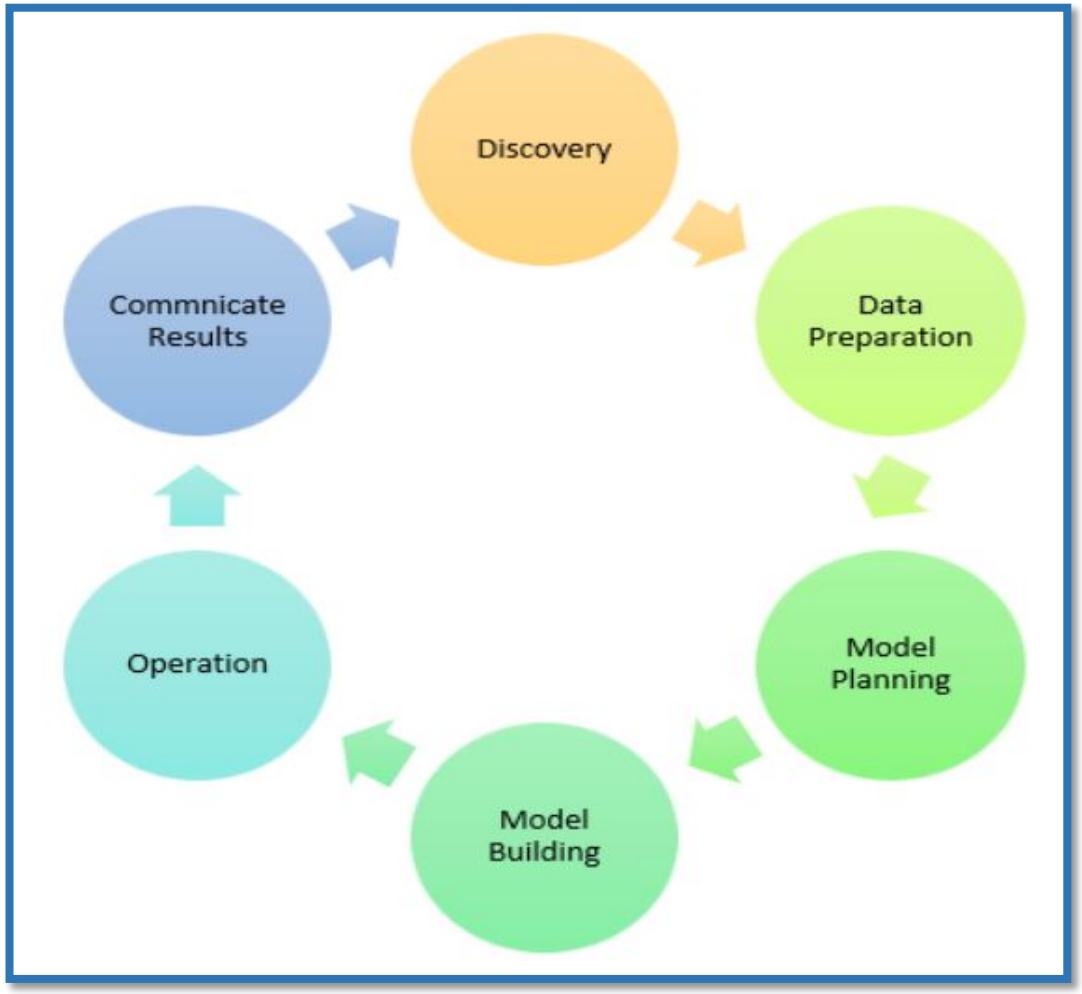


Discovery:

- Discovery step involves acquiring data from all the identified internal & external sources which helps you to answer the business question.

- The data can be:
  - Logs from webservers
  - Data gathered from social media
  - Census datasets
  - Data streamed from online sources using APIs

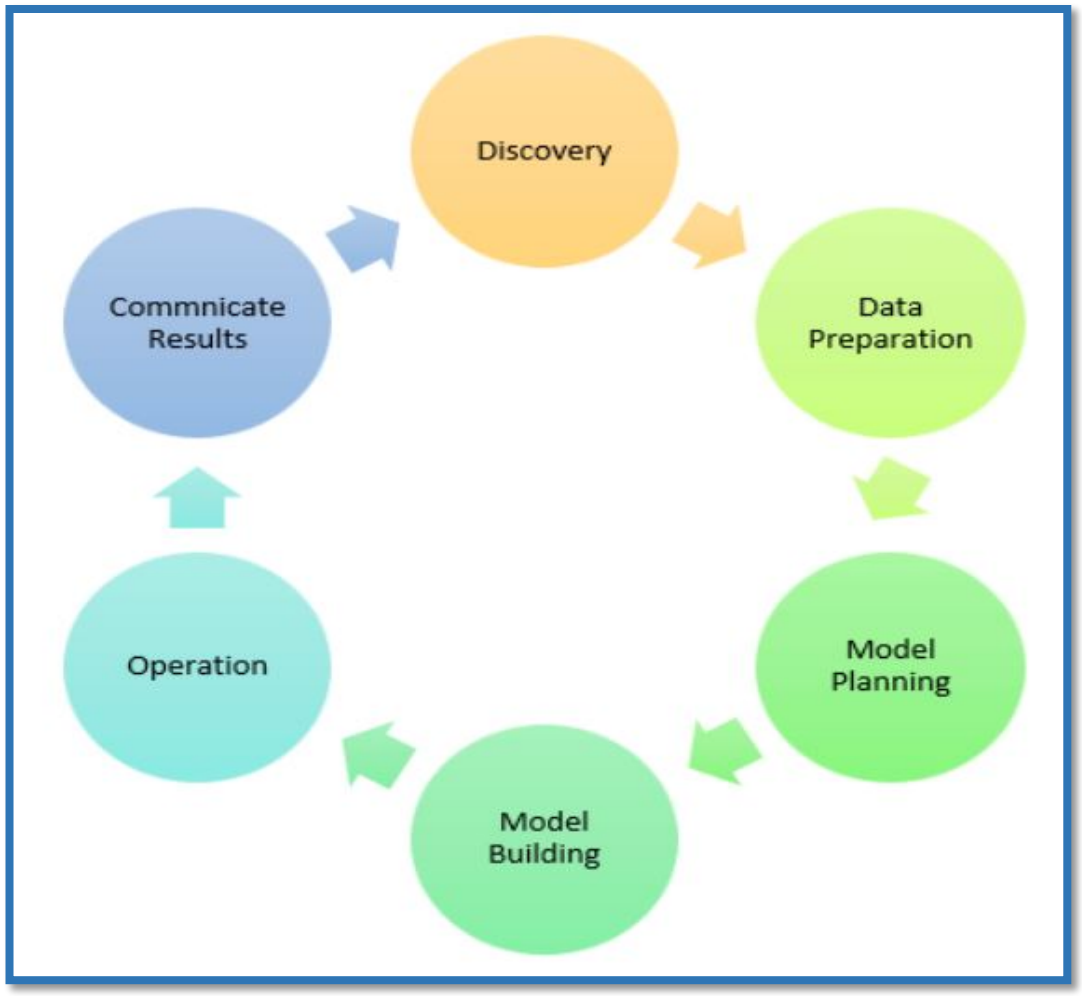
Data Science Process



Data Preparation:

- Data can have lots of inconsistencies like missing value, blank columns, incorrect data format which needs to be cleaned.
- Need to process, explore, and condition data before modeling.
- The cleaner the data, the better are the predictions.

Data Science Process

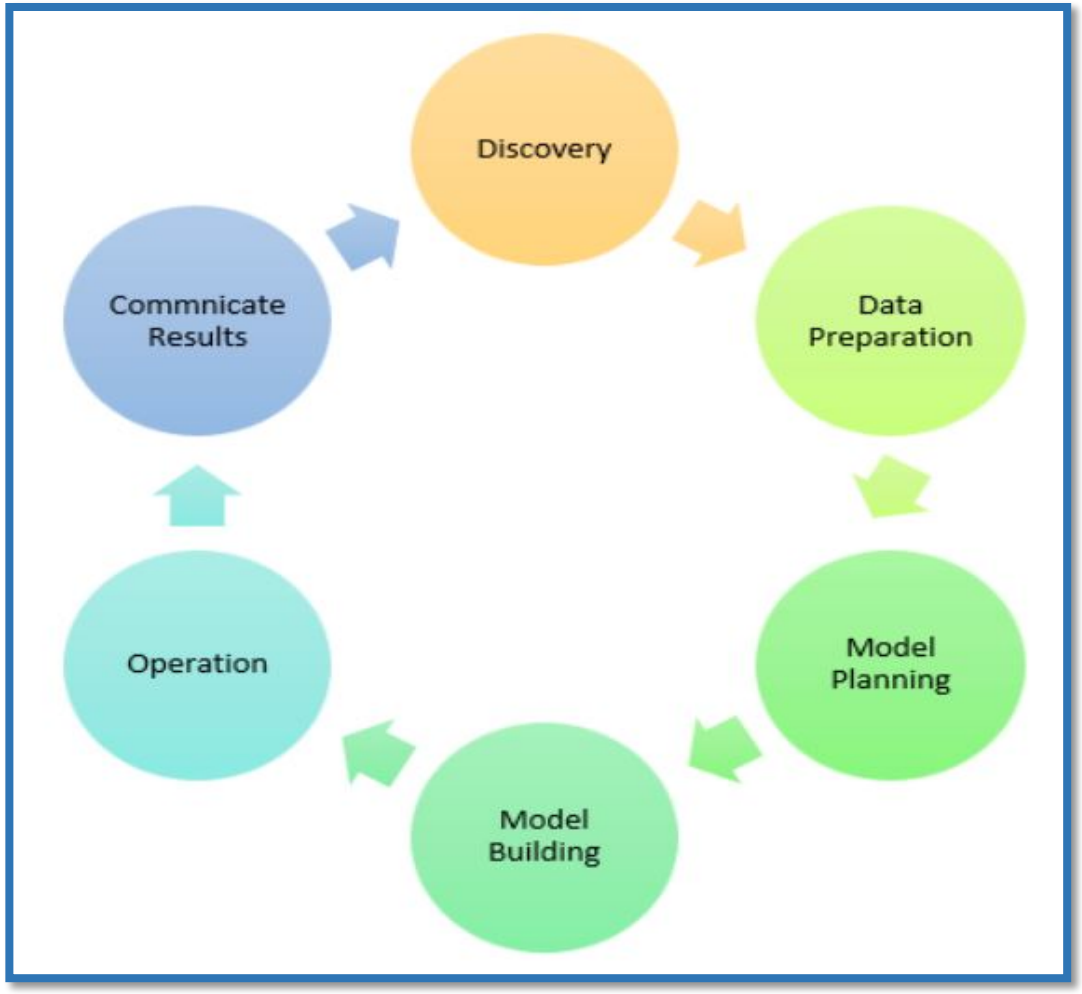


Model Planning:

- This stage will help to determine the method and technique to draw the relation between input variables.

- Planning for a model is performed by using different statistical formulas and visualization tools. SQL analysis services, R, and SAS/ACCESS are some of the tools used for this purpose

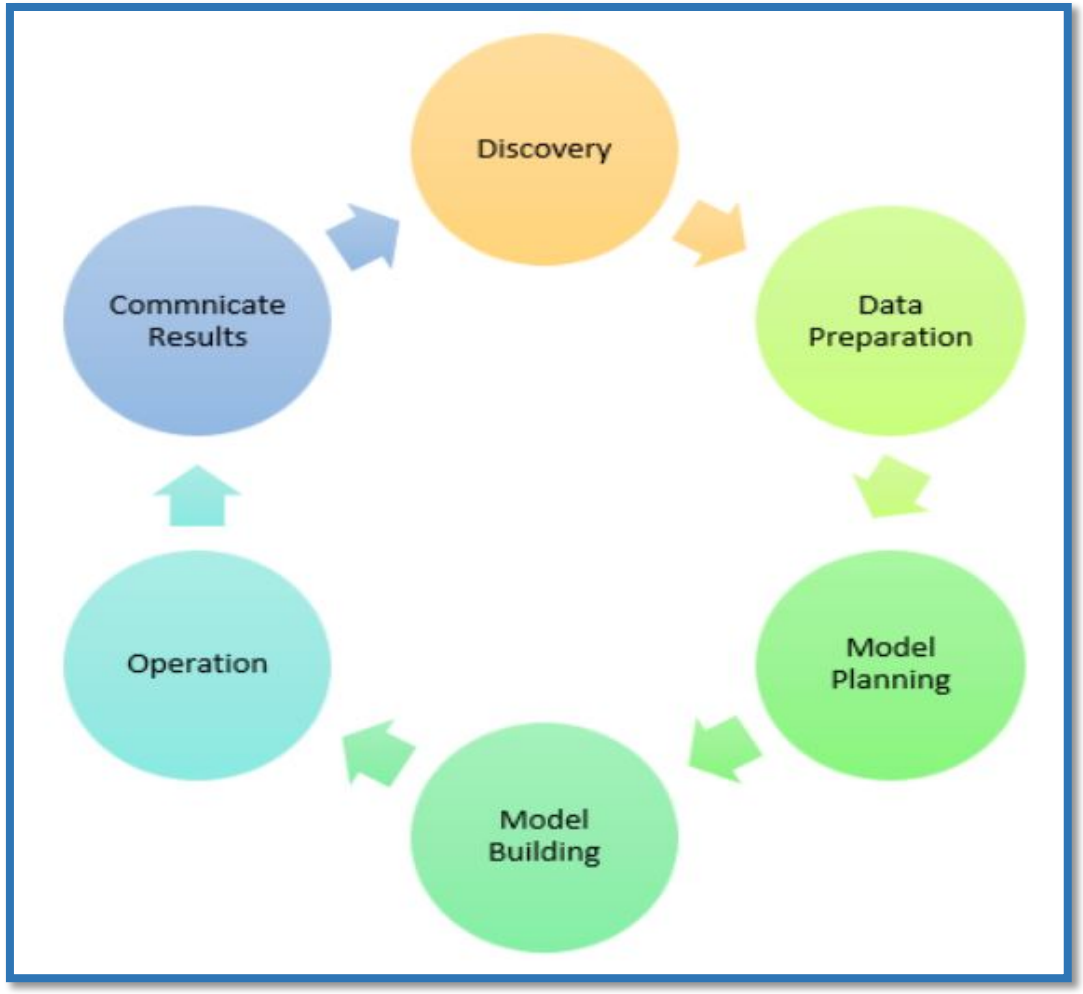
Data Science Process



Model Building:

- In this step, the actual model building process starts.
- Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model once prepared is tested against the "testing" dataset.

Data Science Process



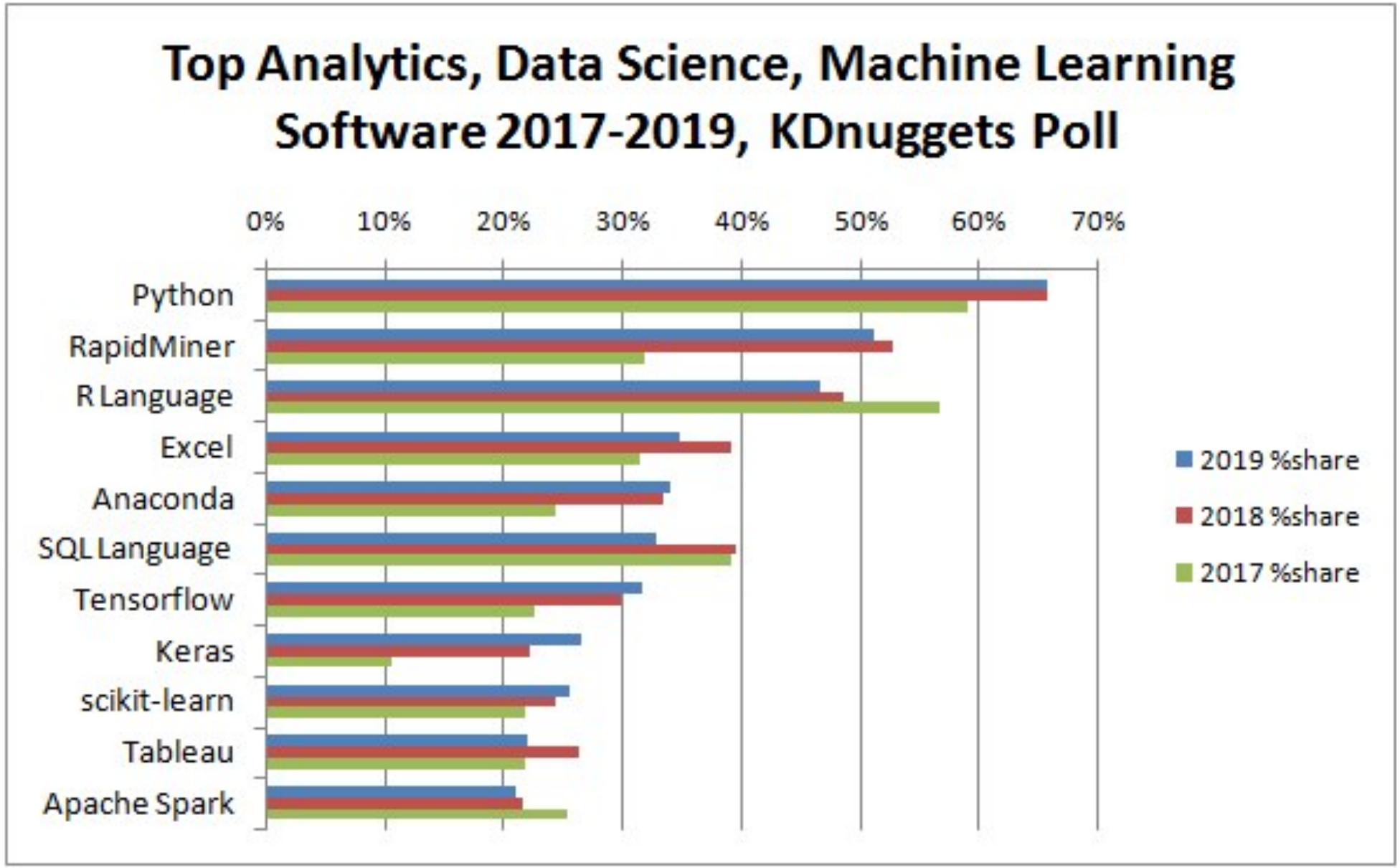
Operationalize:

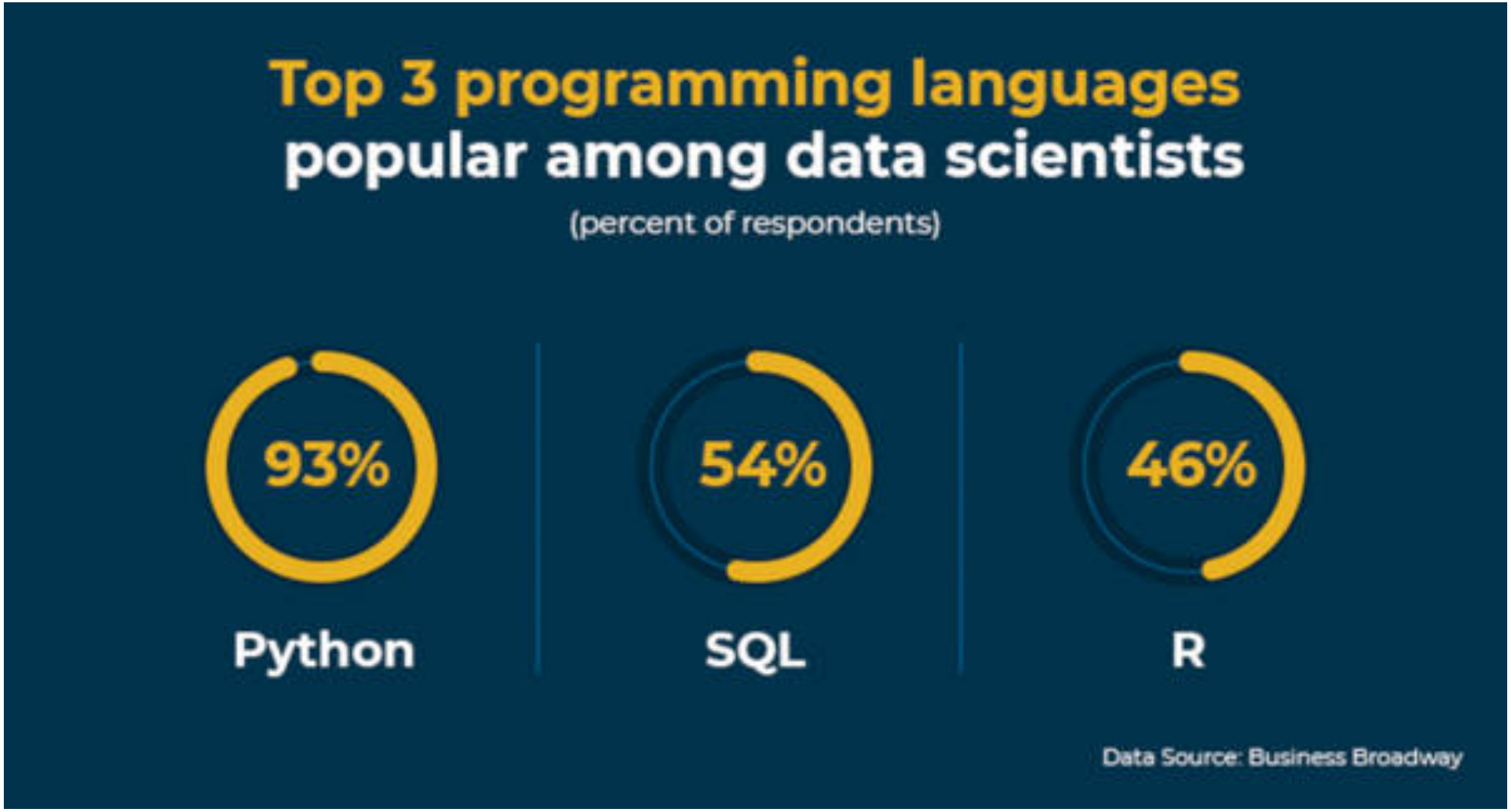
In this stage, you deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing.

Communicate Results:

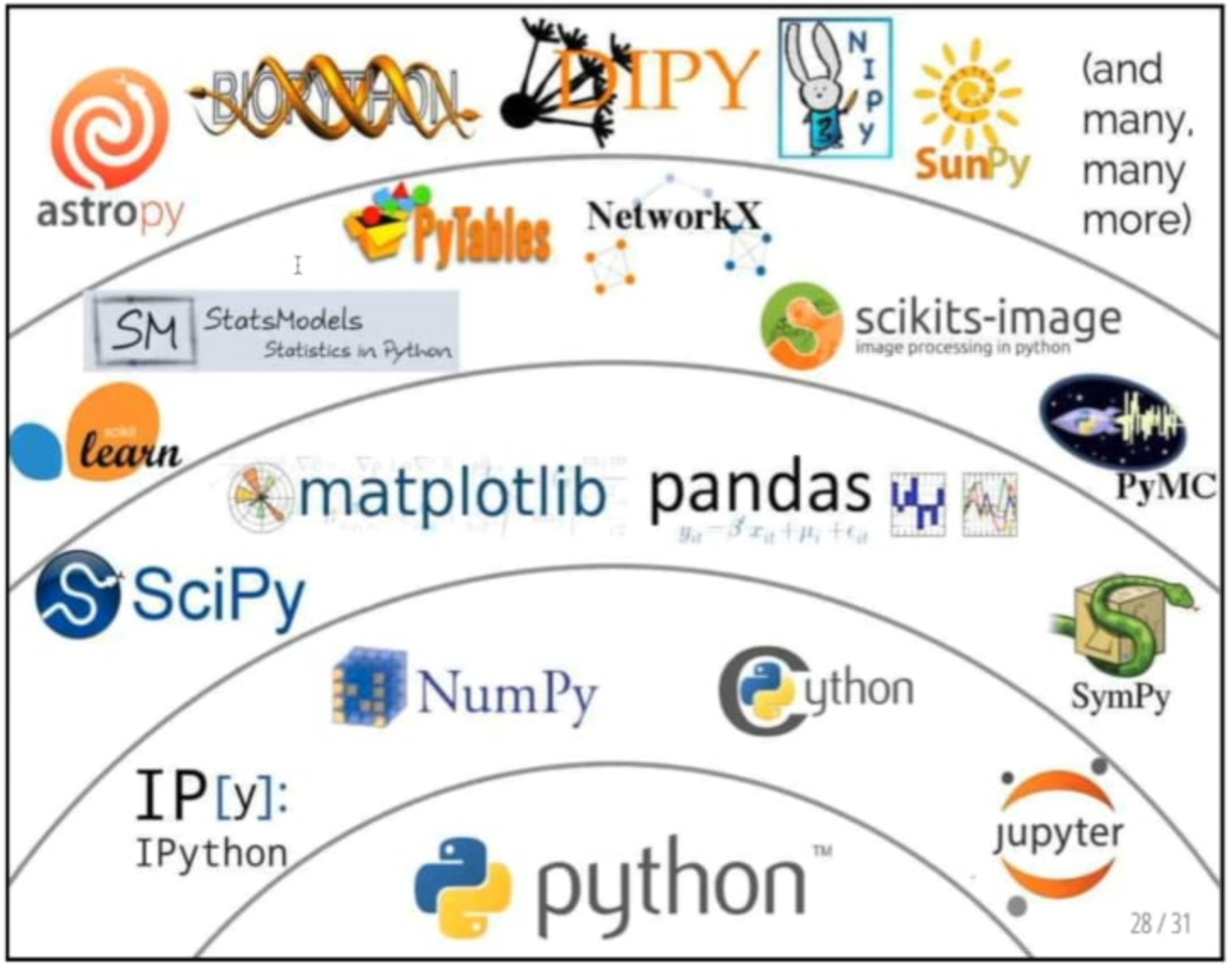
In this stage, the key findings are communicated to all stakeholders. This helps you to decide if the results of the project are a success or a failure based on the inputs from the model.

PROGRAMMING


Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll
0%
10%
20%
30%
40%
50%
60%
70%
Python
RapidMiner
R Language
Excel
Anaconda
SQL Language
Tensorflow
Keras
scikit-learn
Tableau
Apache Spark
2019 %share
2018 %share
2017 %share


Top 3 programming languages
popular among data scientists
(percent of respondents)
93%
Python
54%
SQL
46%
R
Data Source: Business Broadway

Python Toolbox



1

NYC DATA SCIENCE ACADEMY

Python Libraries for Data Science

NumPy:

- introduces objects for multidimensional arrays and matrices, as well as functions that allow to easily perform advanced mathematical and statistical operations on those objects
- provides vectorization of mathematical operations on arrays and matrices which significantly improves the performance
- many other python libraries are built on NumPy

Link: http://www.numpy.org/

Python Libraries for Data Science

SciPy:

- collection of algorithms for linear algebra, differential equations, numerical integration, optimization, statistics ,Signal processing and more
- part of SciPy Stack
- built on NumPy

Link: https://www.scipy.org/scipylib/

Python Libraries for Data Science

Pandas:

- adds data structures and tools designed to work with table-like data (similar to Series and Data Frames in R)
- provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.
- allows handling missing data

Link: http://pandas.pydata.org/

Python Libraries for Data Science

SciKit-Learn:

- provides machine learning algorithms: classification, regression, clustering, model validation etc.

- built on NumPy, SciPy and matplotlib

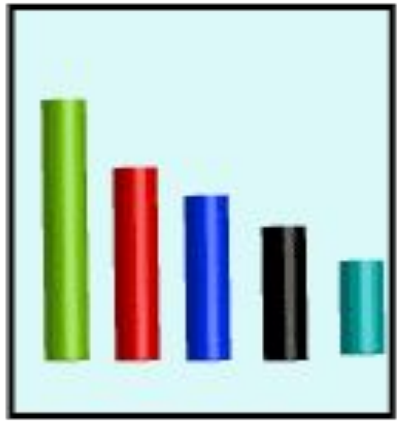
Link: http://scikit-learn.org/

Python Libraries for Data Science

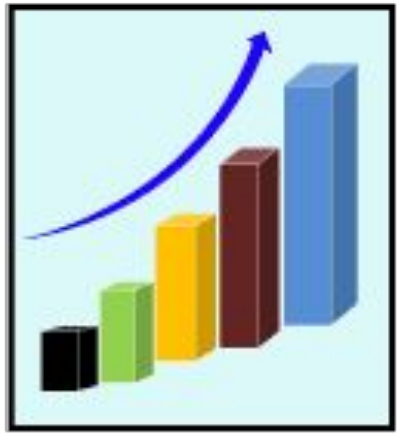
matplotlib:

- python 2D plotting library which produces publication quality figures in a variety of hardcopy formats
- a set of functionalities similar to those of MATLAB
- line plots, scatter plots, barcharts, histograms, pie charts etc.
- relatively low-level; some effort needed to create advanced visualization

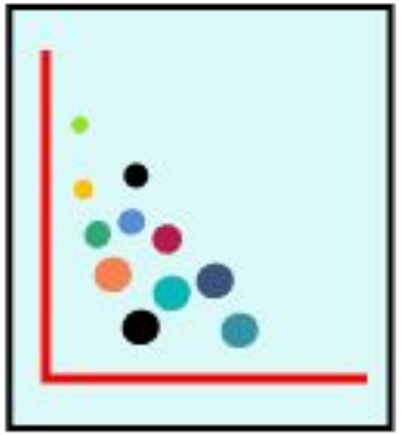
Link: https://matplotlib.org/



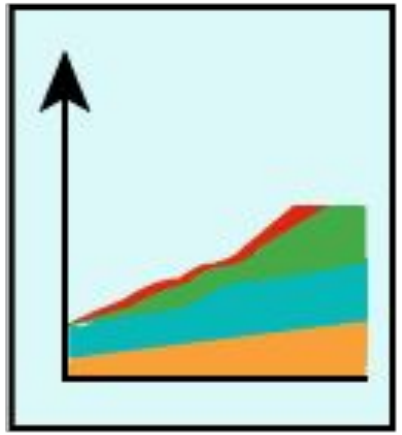
Bar Graph



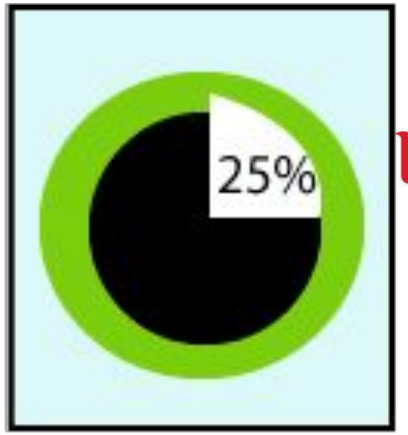
Histogram



Scatter Plot

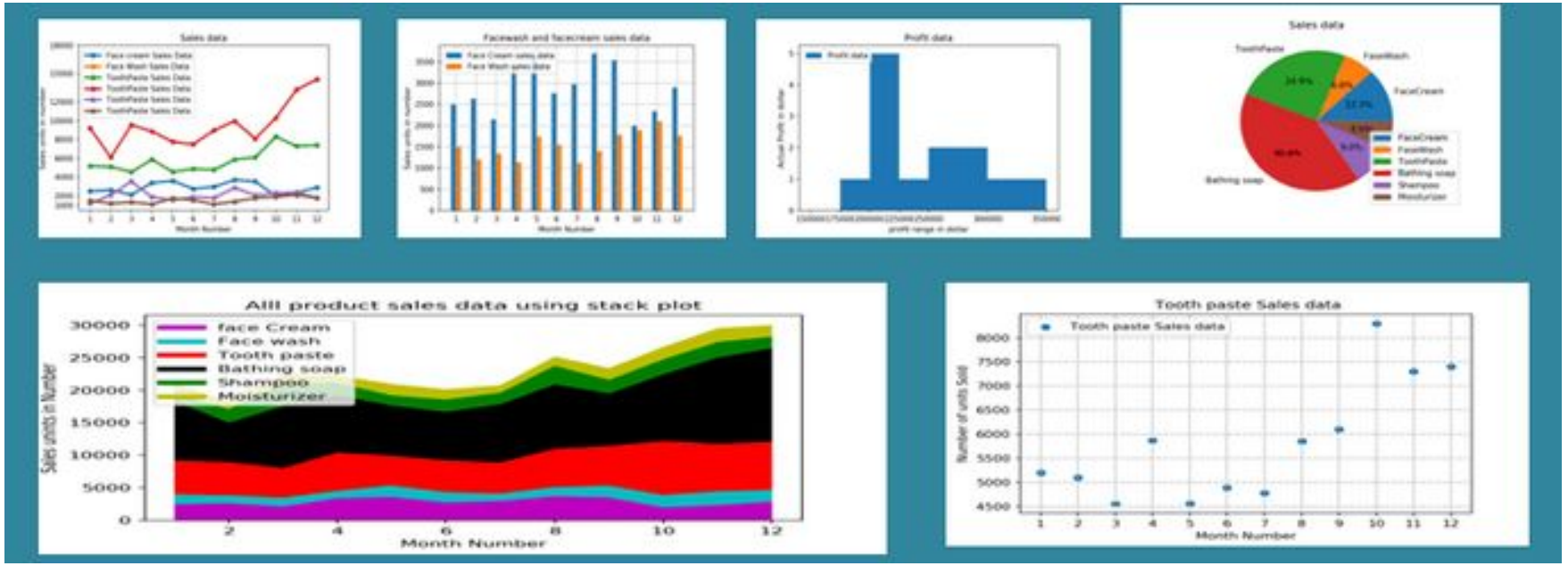


Area Plot



Pie Plot

Karunya



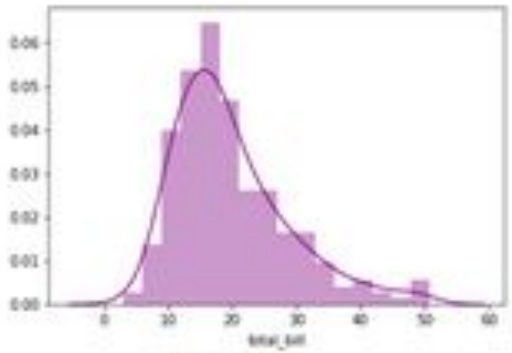
Python Libraries for Data Science

Seaborn:

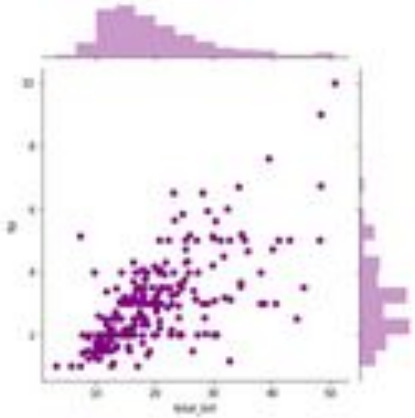
- based on matplotlib
- provides high level interface for drawing attractive statistical graphics
- Similar (in style) to the popular ggplot2 library in R

Link: https://seaborn.pydata.org/

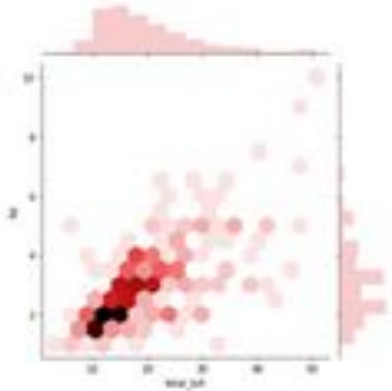
Seaborn Plots



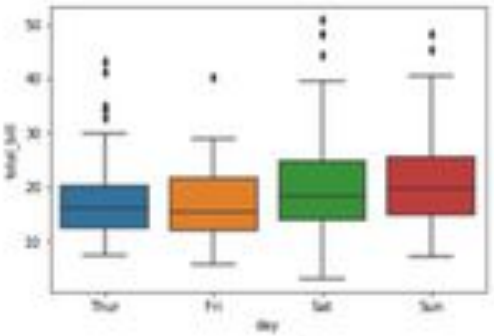
distplot



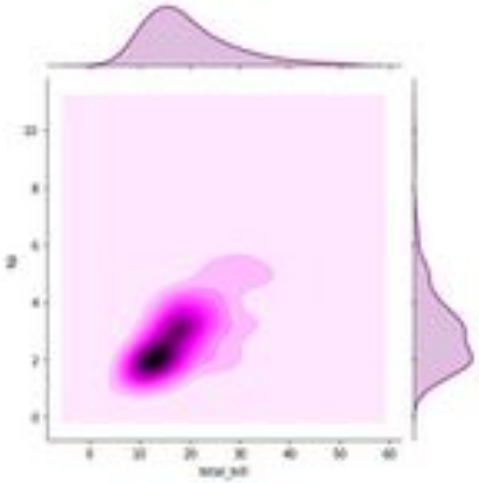
Jointplot



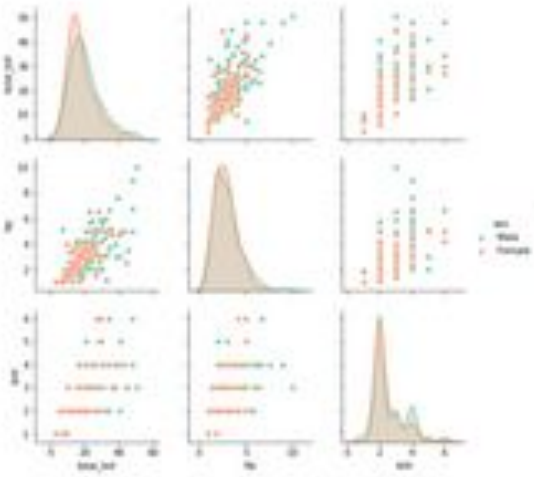
Hexplots



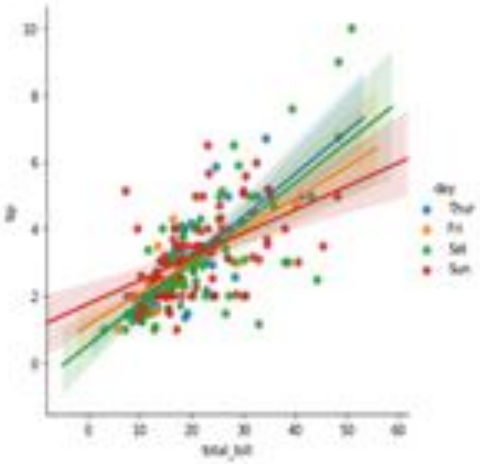
Boxplots



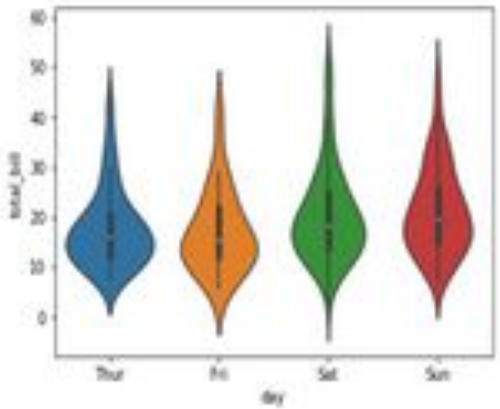
KDE Plot



Pair Plots



LM Plots



Violin Plots

Advantages of Python

- Provides good ecosystem of libraries that are robust and varied
- Tight knit integration with big data frameworks like Hadoop, Spark etc
- Supports both object oriented and functional programming paradigms
- Python is reasonably fast to prototype
- Provides support for reading files from local, databases and cloud

Who invented Python?

Guido Van Rossum

Where was it invented?

National Research Institute for Mathematics and computer science

In Which Country?

Netherland

What is the current version of Python?

3.11

Integrated development environment (IDE)

- Software application consisting of a cohesive unit of tools required for development
- Designed to simplify software development
- Utilities provided by IDEs include tools for managing, compiling, deploying and debugging software

Features of IDE

- IDE should centralize three key tools that form the crux of software development
  - Source code editor
  - Compiler
  - Debugger
- Additional features
  - Syntax and error highlighting
  - Code completion
  - Version control

Commonly used IDEs

- Spyder
- PyCharm
- Jupyter Notebook
- Atom

Top Algorithms For Data Scientists