
Lecture # 1
Overview of the Course
DATA SCIENCE
jupyter
python
QGIS
NumPy
TensorFlow
SciPy
scikit learn
Keras
pandas
matplotlib
BASIC STATISTICS
REGRESSION
CLASSIFICATION
RECOMMENDATION SYSTEM
DIMENSIONALITY REDUCTION
FEATURE EXTRACTION
OPTIMIZATION
CLUSTERING
Dr. Muhammad Arif Butt, Ph.D.

# Today's Agenda

- About Myself
- Course Information and Protocols
- Data Data Everywhere
- Categories of Data
- What is Data Science?
- Factors making Data Science Ubiquitous
- Applications of Data Science
- How to Do Data Science?
- Languages, Tools and Techniques
- Life Cycle of a Data Science Project
- Industry Job Roles in Data Science
- Discussion on Course Matrix

# Myself

Not Secure — arifbutt.me

Home Teaching Video Lectures Research Publications CV Contact

Muhammad Arif Butt

I am working as an Assistant Professor at College of Information Technology, University of the Punjab, since 2005. I am an ex-Pakistan Army officer, who joined the Pak Army in 1988 and left it after serving for nearly thirteen years. During my stay in uniform, I had the honor to serve in the snowy mountains of Kashmir and Siachin. I also have the honor to serve as an instructor at the School of Infantry and Tactics Quetta, which is a prestigious training institute of Pakistan Army. After the service, my thrust for knowledge and passion for teaching and learning moved me to one of the leading IT institutes of Pakistan - Punjab University College of Information Technology (PUCIT). I completed my MSc (CS) and MPhil (CS) both with a Gold Medal, and started teaching as a full time faculty at PUCIT. My teaching interests are Operating Systems, Embedded Systems, and System Programming. The focus of my PhD was on applying Fuzzy inference models in Operating System modules, where decision making is done based on imprecise and vague inputs, with the intent of enhancing performance and making the beast more user friendly. I am running a Kernel Fuzzification and Embedded Systems Lab at PUCIT, where I, along with my students are working on development and fuzzification of Linux Kernel modules and device drivers.


**E-mail:** arif@pucit.edu.pk
**Office:** Principal Office (New and Old Campus)
**Phone:** +92 42 111 923 923 (Extn: 104)

**Office hours:**
Mondays (Old Campus): 12:01 to 13:00
Tuesdays (New Campus): 12:01 to 13:00
Wednesdays (Old Campus): 12:01 to 13:00
Thursdays (New Campus): 12:01 to 13:00

**Mailing address:** Punjab University College of Information Technology, University of the Punjab, Allama Iqbal (Old Campus), Shahrah-e-Quaid-e-Azam (The Mall) Lahore, Pakistan

# Course Information and Protocols

# Course Info

- **Textbook(s):** Python for Data Analysis, by, Wes McKinney, 2nd Edition, Published in 2017, ISBN-13: 9781491957660
- **Lectures Slides Available at:** http://arifbutt.me
- **Video Lectures Available at:** https://youtube.com/learnwitharif
- **Codes Hosted at:** https://github.com/arifpucit/data-science
- **Grades Website:** http://online.pucit.edu.pk
- **Prerequisites: Basic Programming skills**
- **Office:** Building-C, FCIT (NC)
- **Students Counseling hours:**
  - Tues: 11:30 hrs – 1:30 hrs
  - Thu: 11:30 hrs – 1:30 hrs
- **24 hour turnaround for email:** arif@pucit.edu.pk

# Who cares to get an A

**Final exam:** 40

**Mid-exam:** 35

**Sessionals:** 25

- Quizzes: 30%
- Programming Assignments : 30%
- Research Papers : 40%


**MPhil.**

**Minimum GP to pass a course:** GP >=2.3 [C+ or 61 mks]

**Degree Completion Requirement:** CGPA >=2.5

**Probation:** 2.3 <= CGPA < 2.5 [Only one probation allowed]

**Dropped out:** CGPA < 2.3

**Ph.D.**

**Minimum GP to pass a course:** GP >=2.7 [B- or 65 mks]

**Degree Completion Requirement:** CGPA >=2.8

**Probation:** 2.8 <= CGPA < 3.0

**Dropped out:** CGPA < 2.8

# Cheating Policy

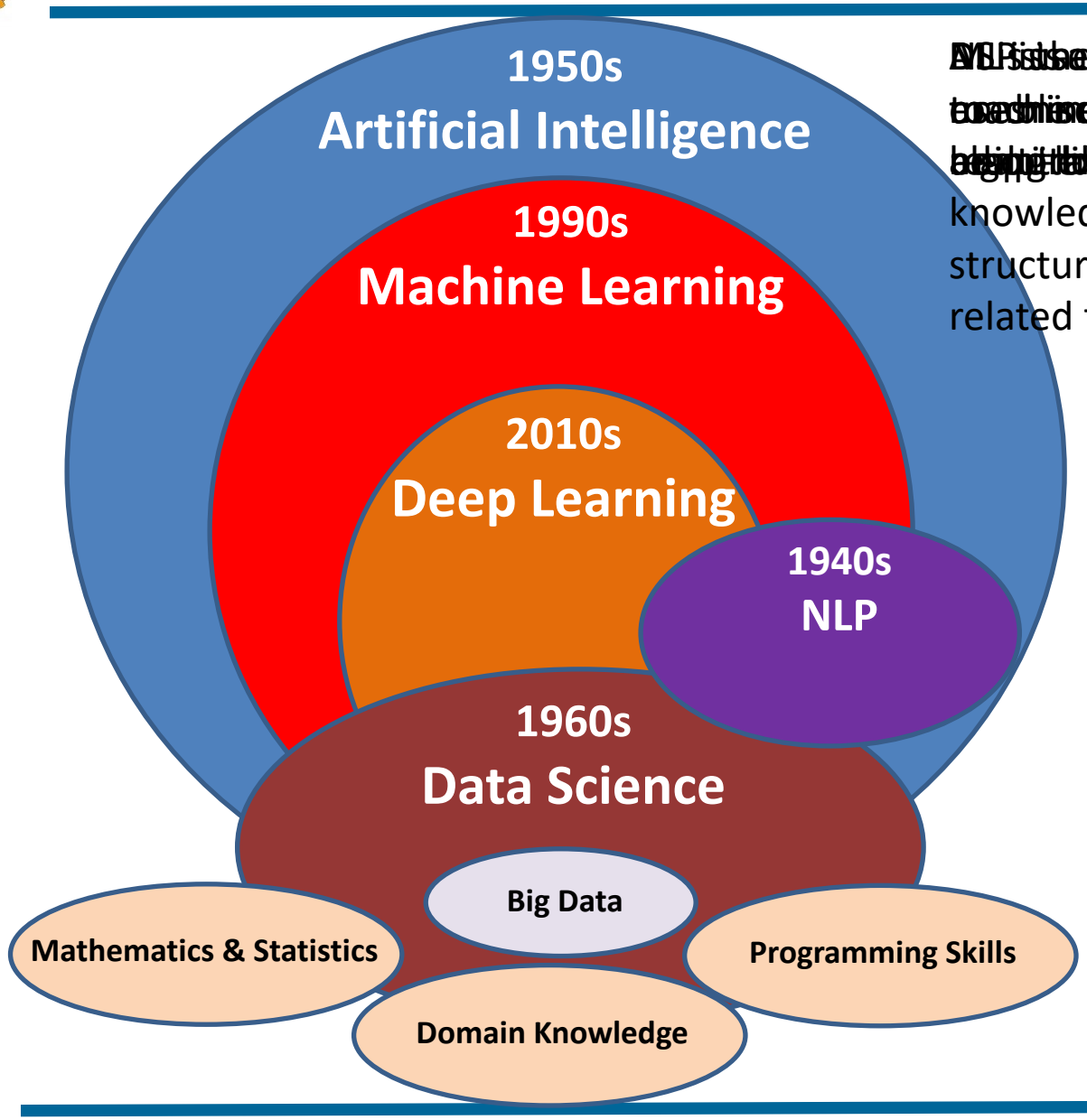
- Academic integrity
- Both the cheater and the student who aided the cheater will be held responsible for the cheating
- The instructor may take actions such as:
  - require repetition of the subject work,
  - assign 'zero' or may be 'negative' marks for the subject work,
  - for serious offenses, assign an F grade for the course

# Late Policy for Home Works and PA

- Late policy for Assignment, Quizzes, and other deliverables
  - No late Assignment submissions!
  - No late quizzes or exams!
- Sticking to dates is your responsibility!
  - Check announcements on lecture notes regularly
- Your best strategy is to play it safe – submit everything on time

# Lecture Format

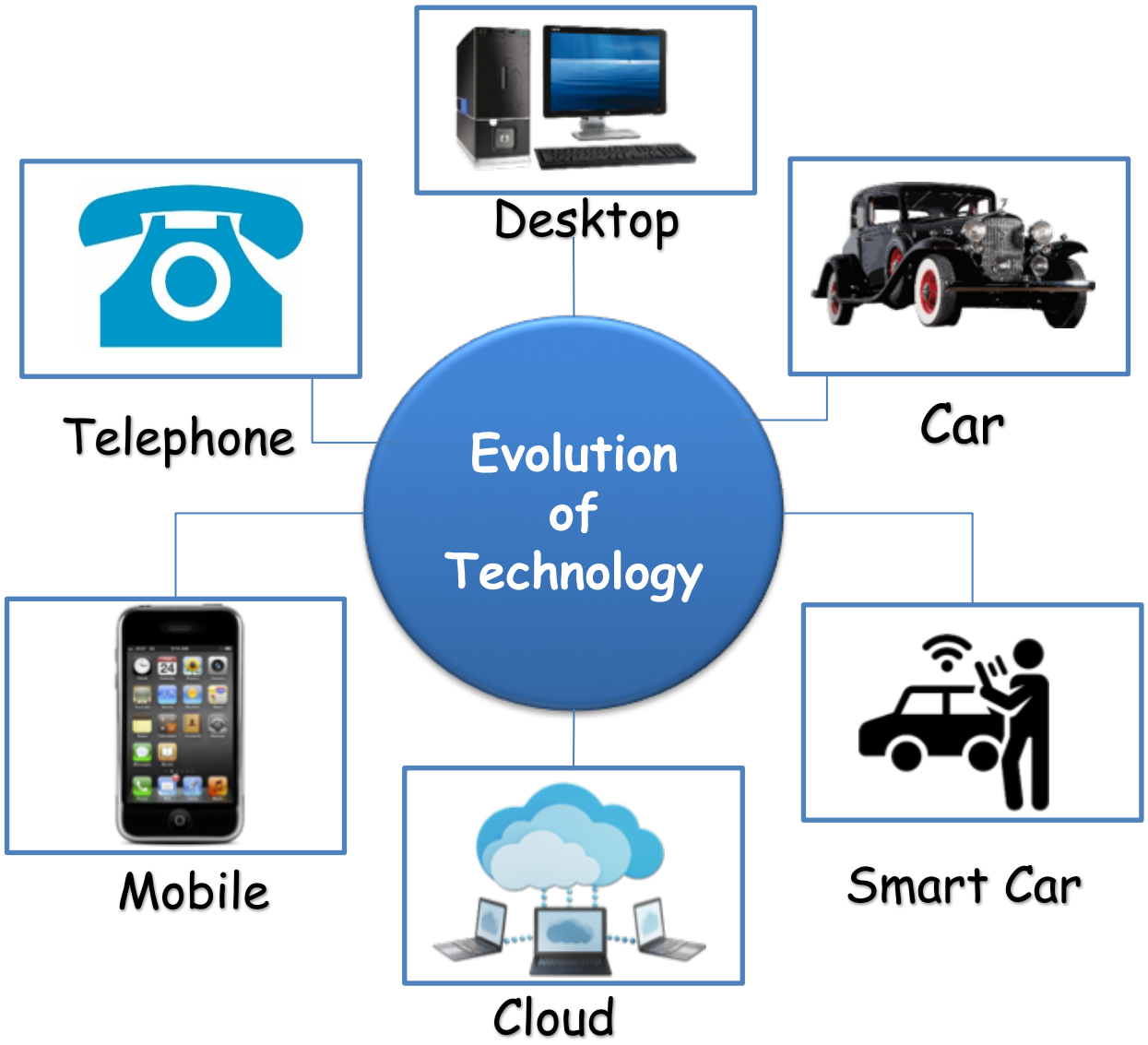
# The Big Picture




knowledge and insights from structured and unstructured data related to business problem.

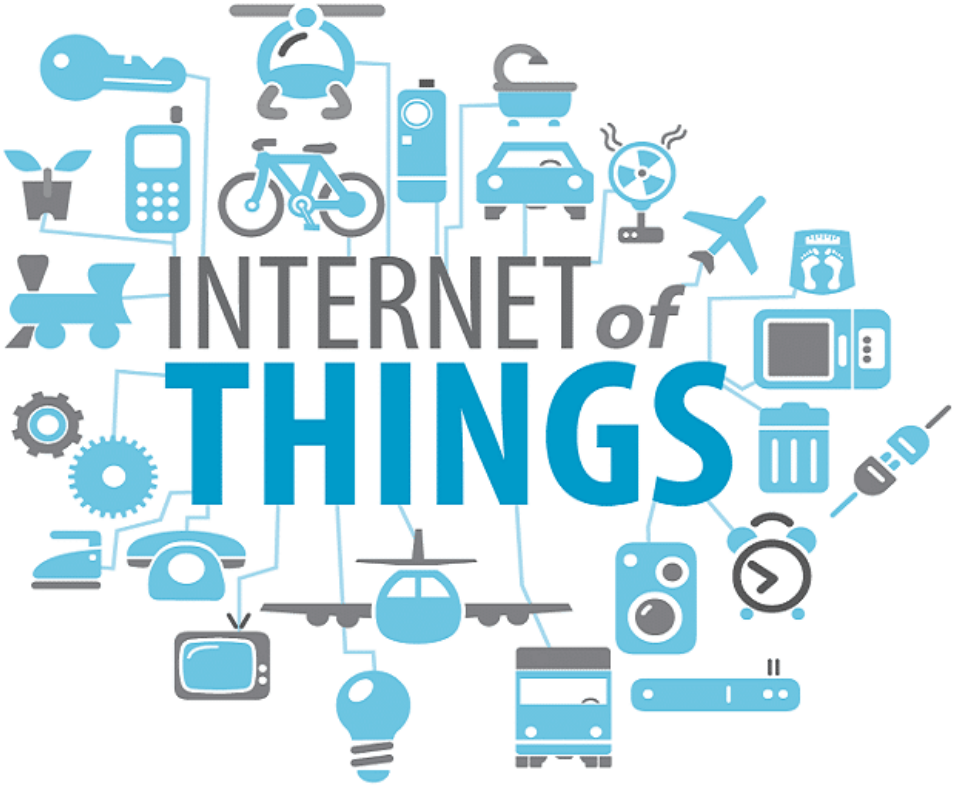
# Data Data Everywhere
# Data Sources

# Data Sources: Evolution of Technology



# Data Sources: IoT

Collection of interconnected devices that communicate and transfer data through the Internet





As per CISCO recent survey, IoT is generating more than 500 ZB of data per year

# Data Sources: Social Media


347,222 tweets

1,736,111 pictures

4,166,667 likes &
200,000 photos

204,000,000 emails

300 hrs of video uploaded
2.78 million video views

342,000 apps downloaded

70,017 hours watched

2.4 million search queries

Imagine processing & analyzing this much data, and then trying to figure out important insights from it

# Data Sources: Other Factors

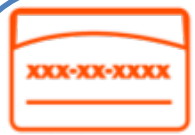

Data Science is all about extracting the useful insights from data and using it to grow your business

# Categories of Data

# Structured Data

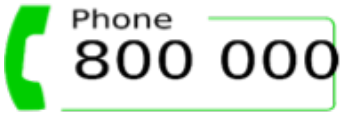
## Structured Data

### Examples




Social Security Number

Date




Phone Numbers

Customer Name

Transaction information

Credit Card number

### Characteristics

Pre-defined Data Model

Easy to Search

Text-based

### Applications


Airline reservation systems

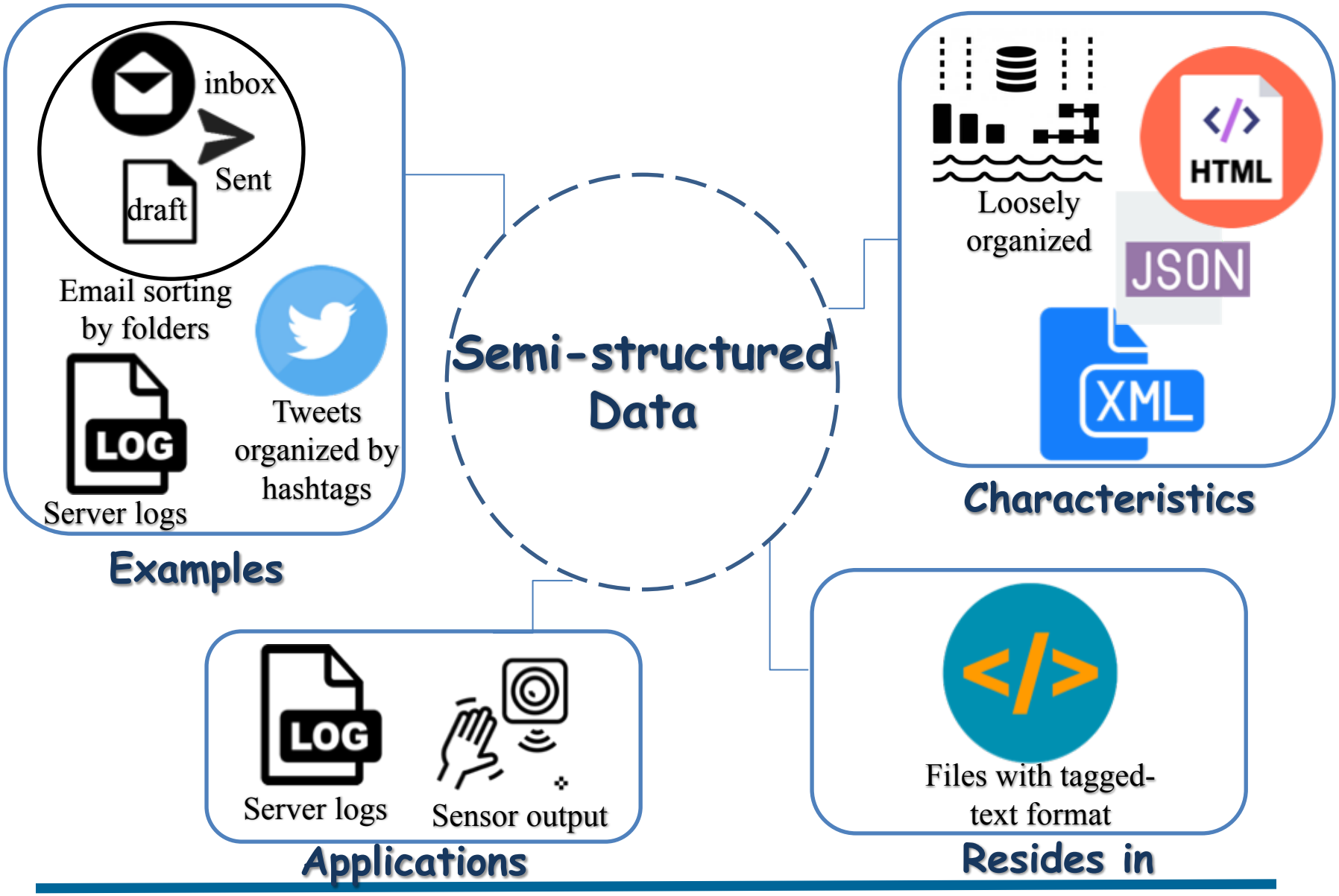
Inventory control

### Resides in

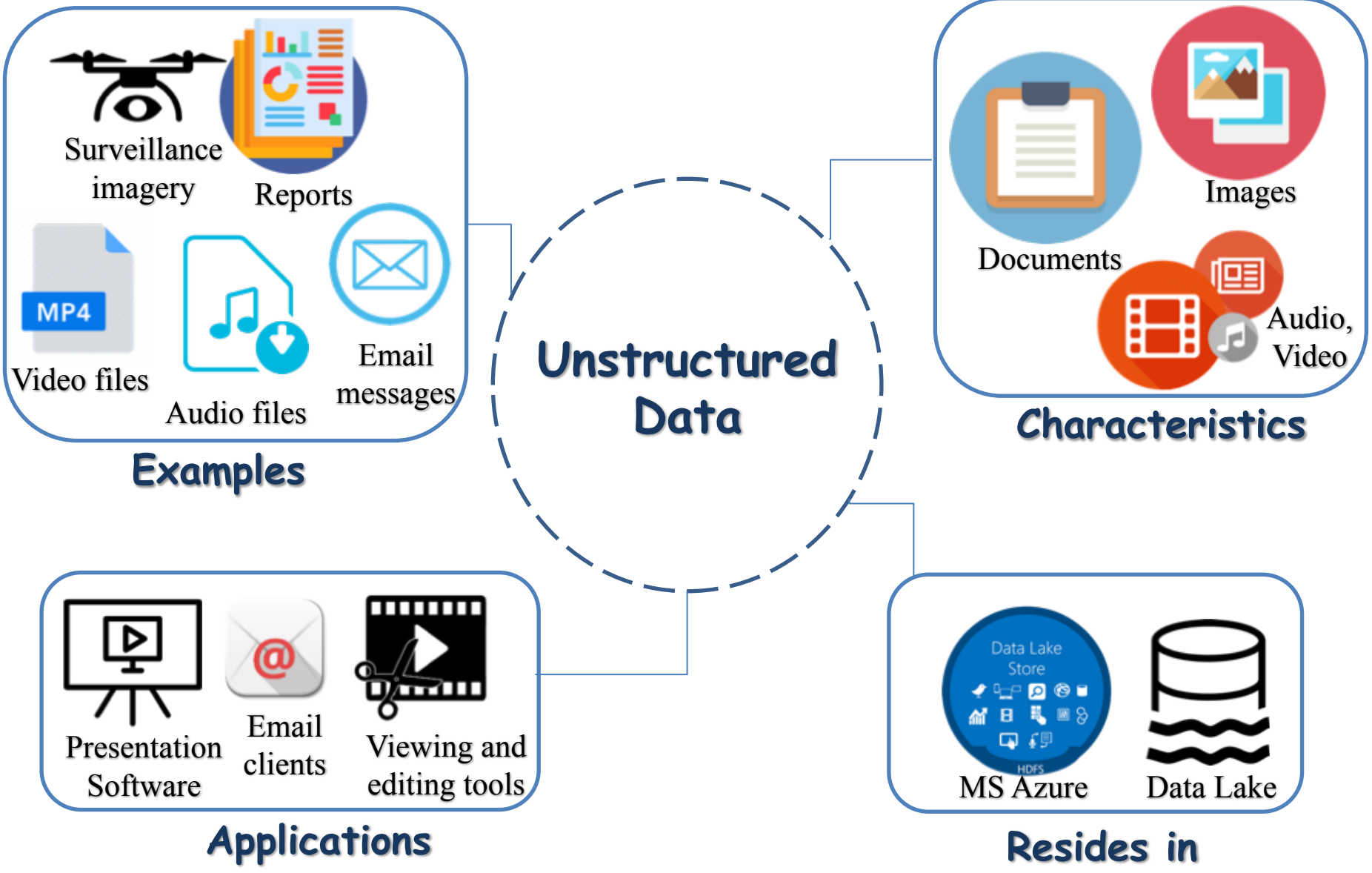

Data Mart

Database

# Semi-structured Data



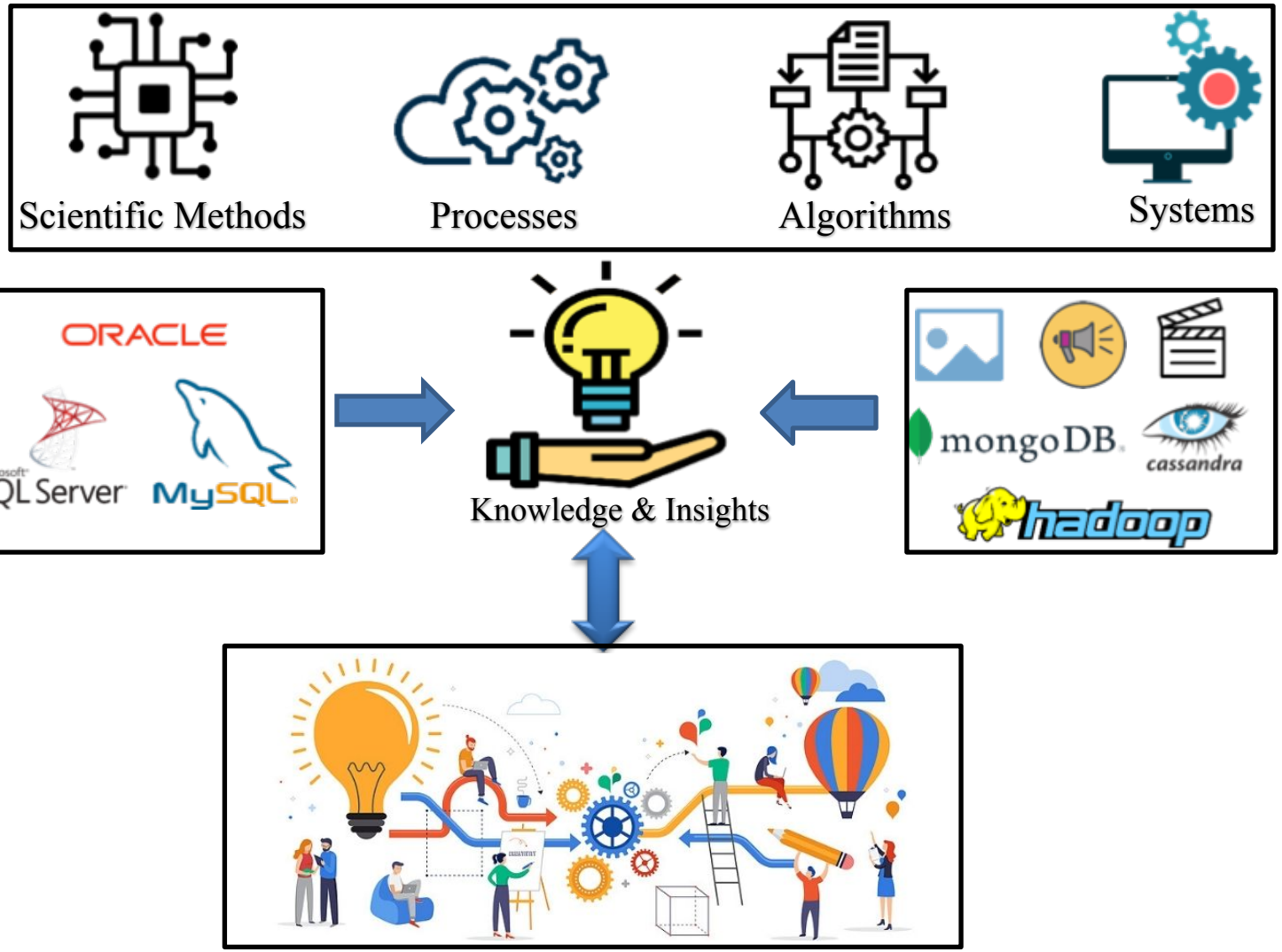
# Unstructured Data



# What is Data Science?

# What is Data Science?

Data Science is an Inter-Disciplinary Field that uses

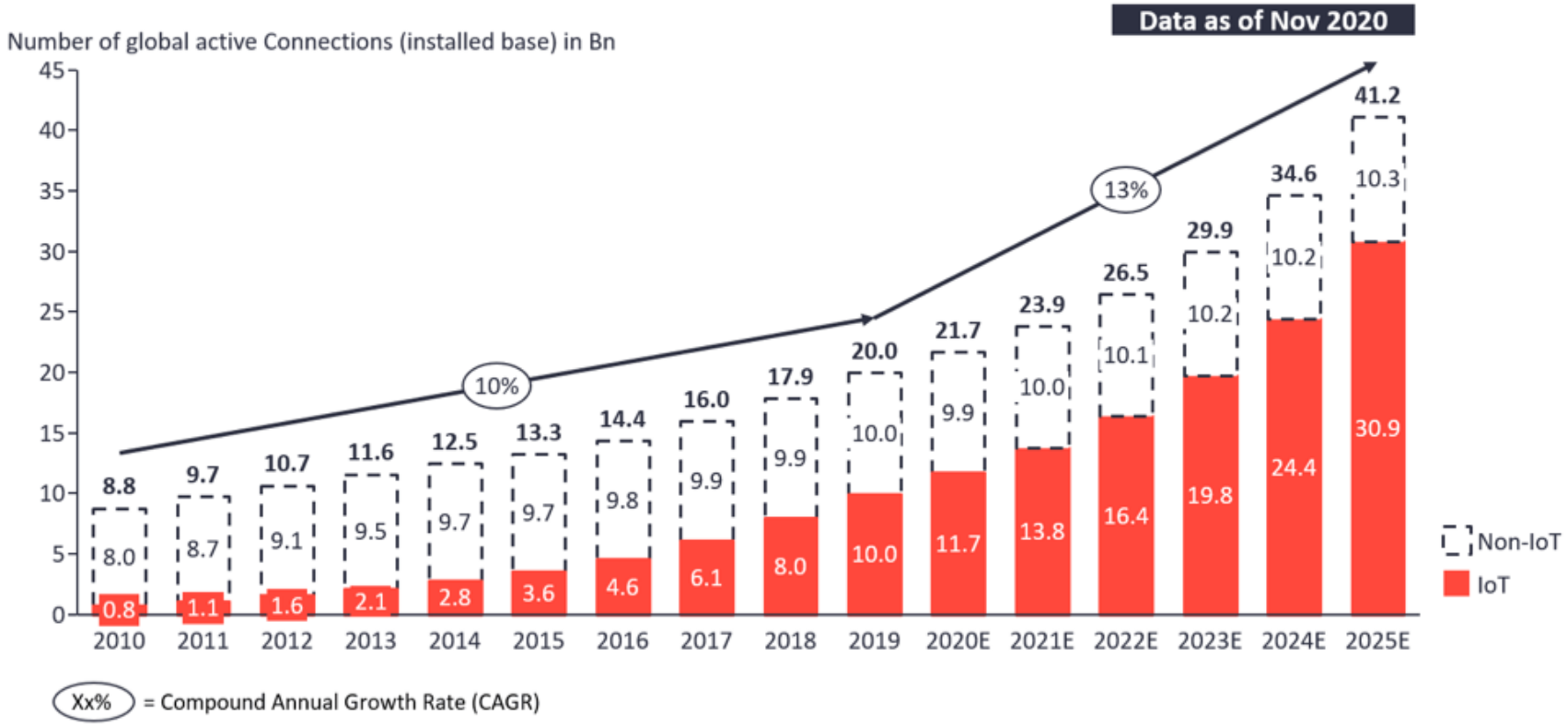


# Factors Making Data Science Ubiquitous

# Increasing Number of Connected Devices

## Total number of device connections (incl. Non-IoT)

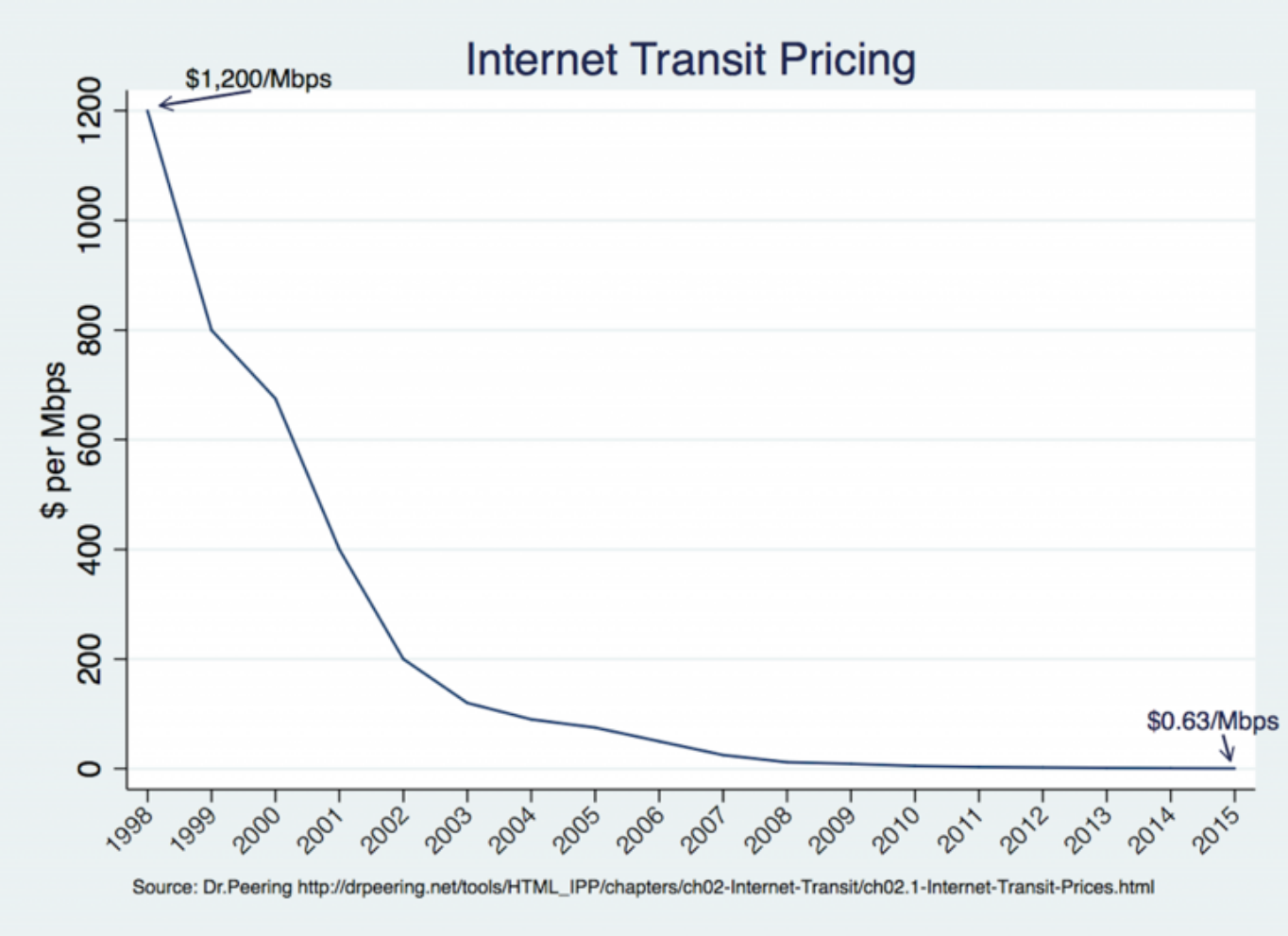
**20.0Bn in 2019– expected to grow 13% to 41.2Bn in 2025**



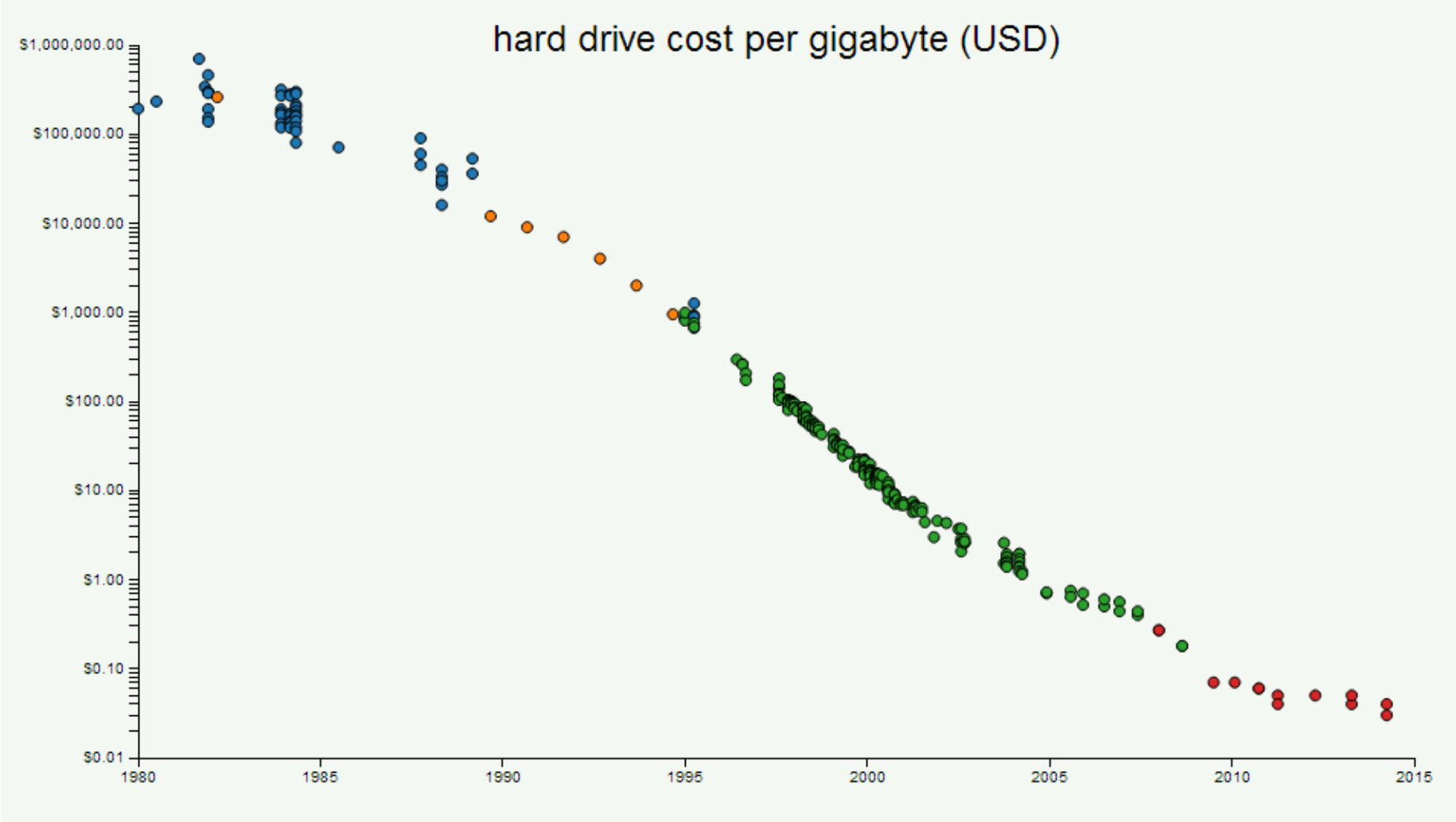

Note: Non-IoT includes all mobile phones, tablets, PCs, laptops, and fixed line phones. IoT includes all consumer and B2B devices connected – see IoT break-down for further details

**Source(s):** IoT Analytics - Cellular IoT & LPWA Connectivity Market Tracker 2010-25

# Decreasing Internet Transit Pricing

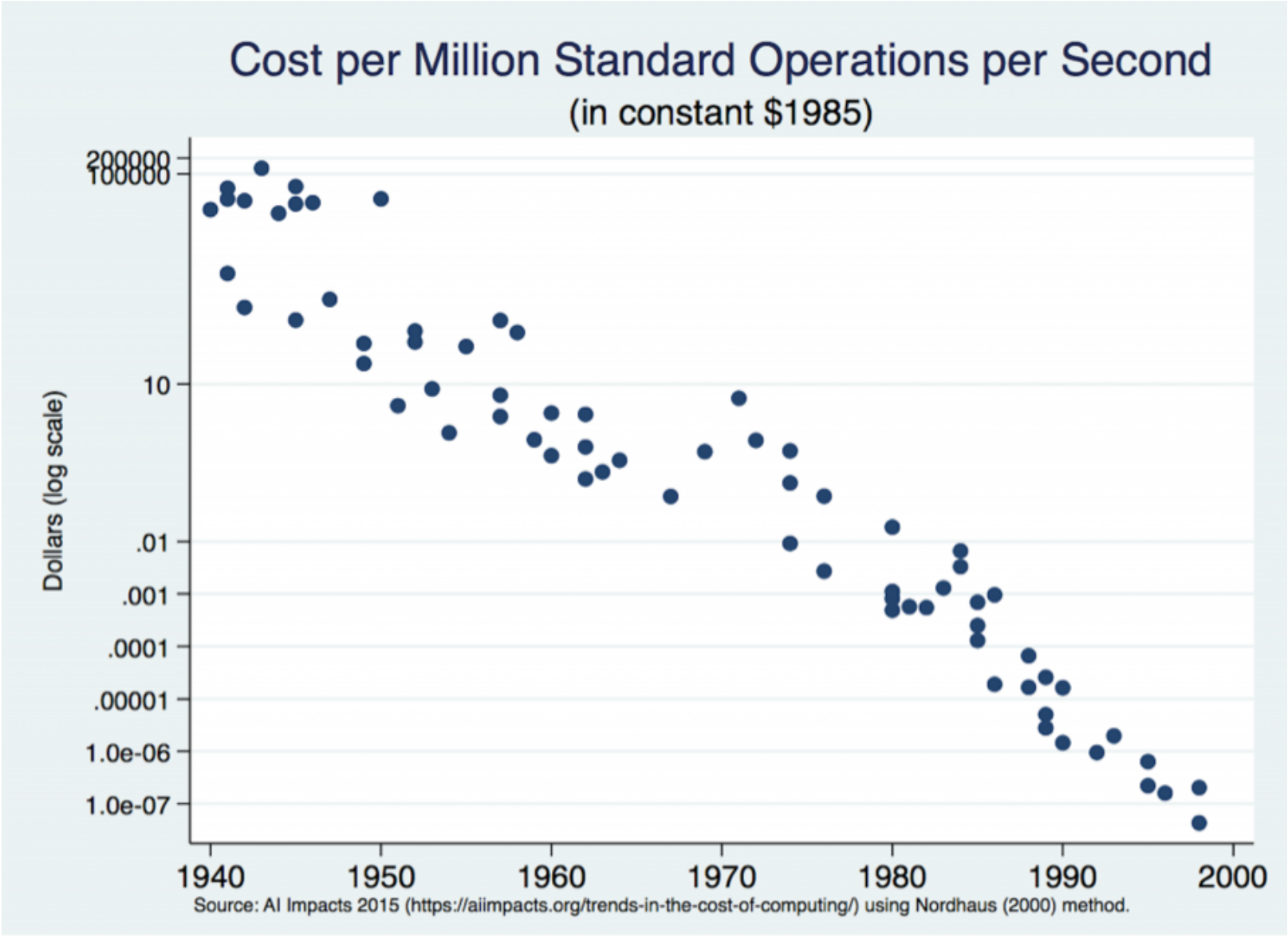


# Decreasing Costs for Data Storage



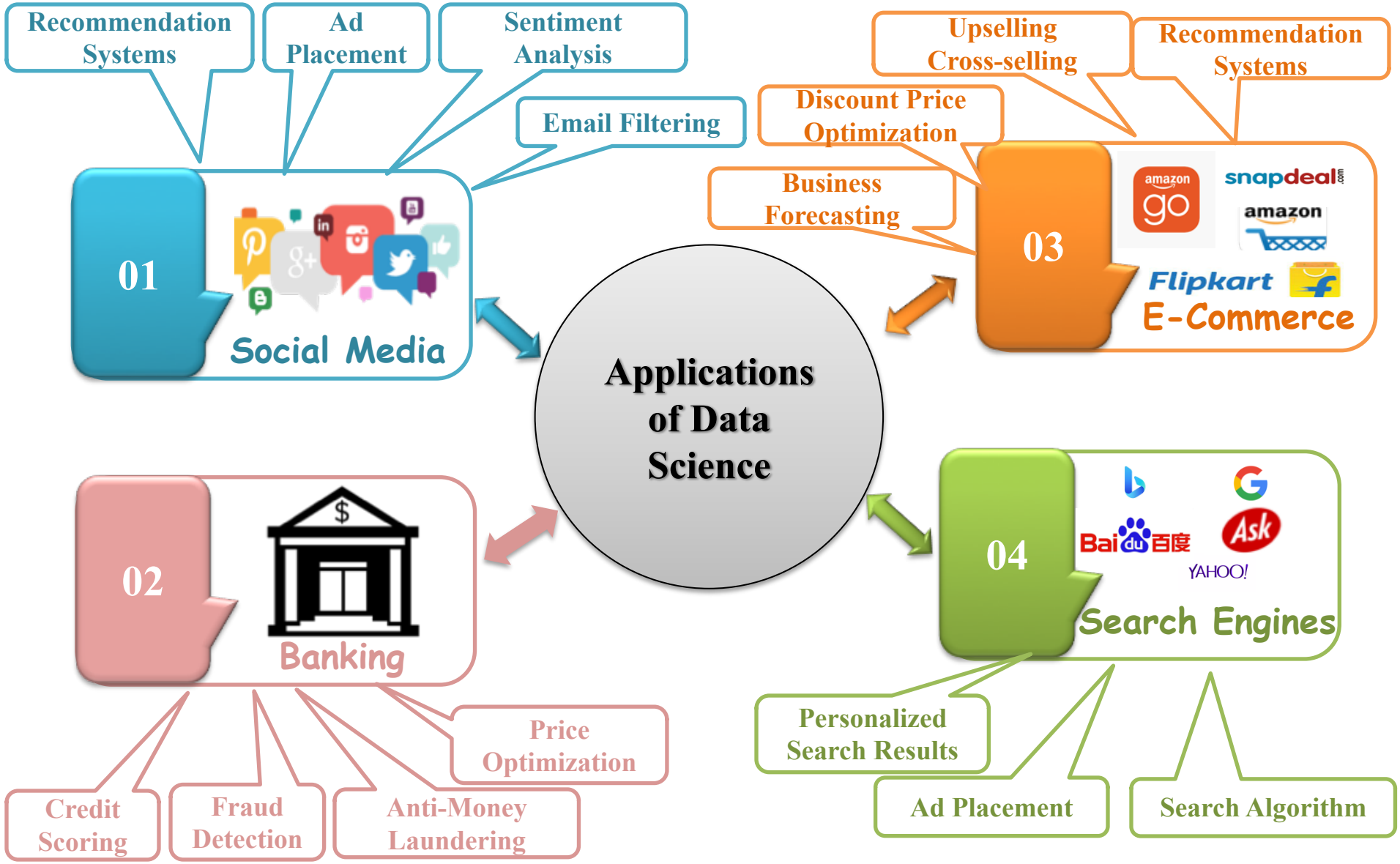

Source: https://community.spiceworks.com

# Decreasing Computational Costs

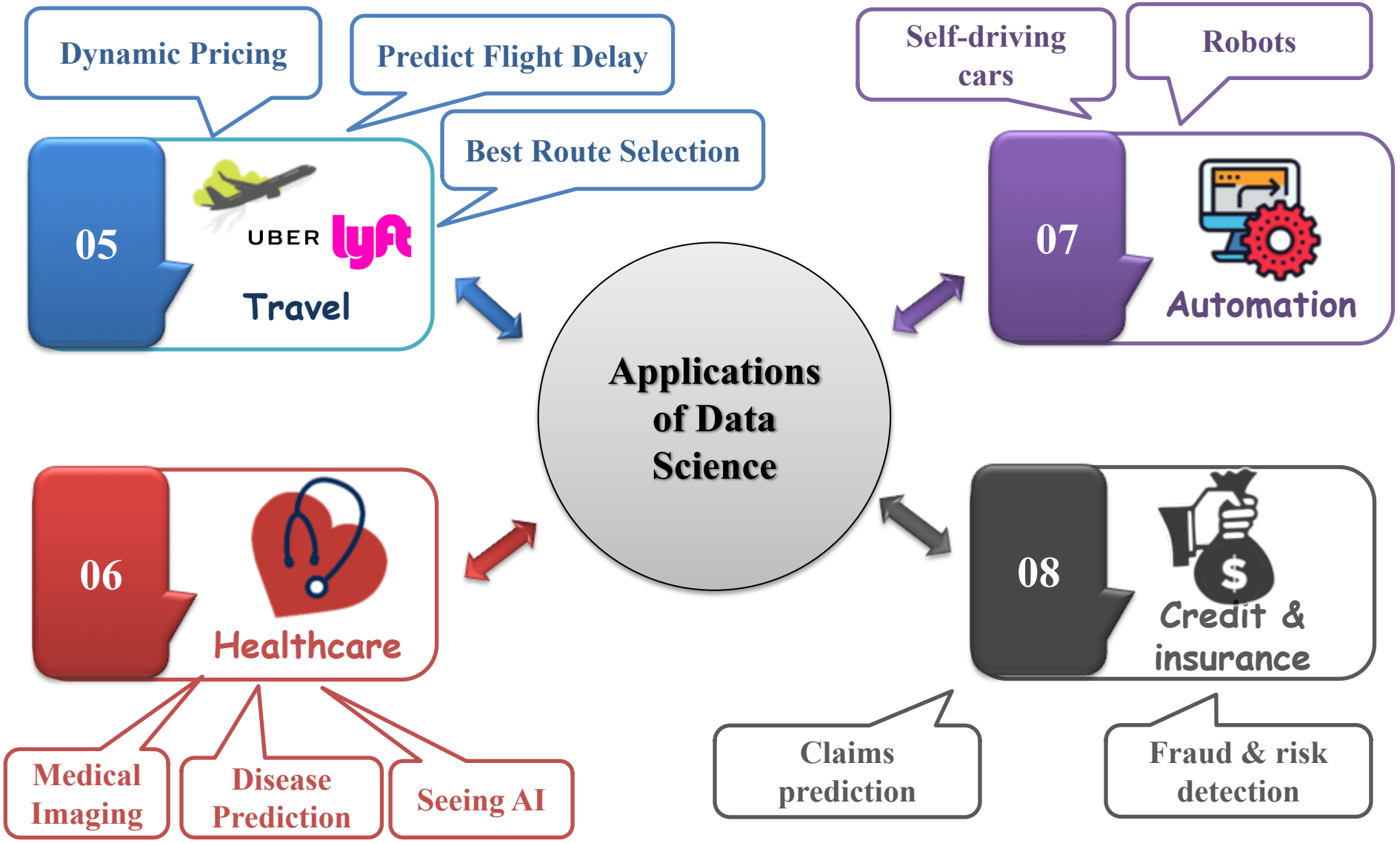


# Applications of Data Science?

# Applications of Data Science



# Applications of Data Science (cont…)

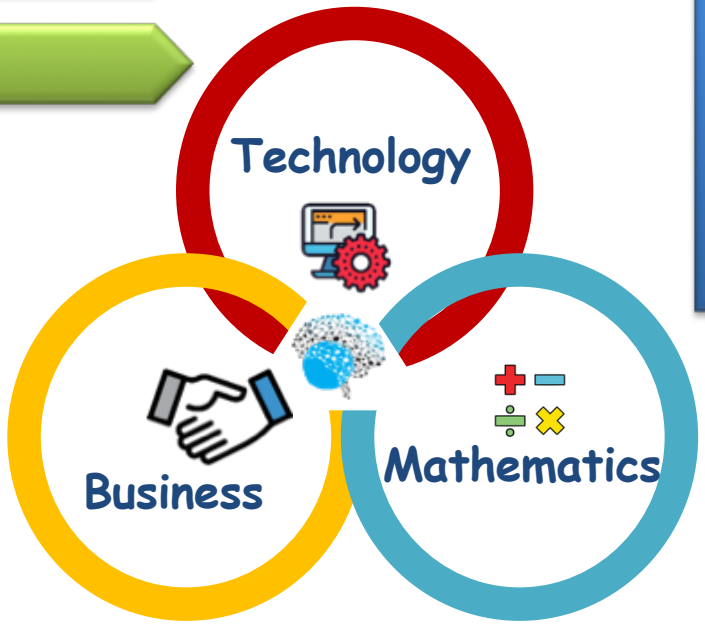


# How to Do Data Science? Languages, Tools and Technologies

# Who is a Data Scientist?

1. Data Scientist
2. Skill Set
3. Programming languages
4. Tools
5. Techniques

A data scientist is a professional responsible for **collecting**, **analyzing** and **interpreting** extremely large amounts of structured and unstructured data in order to gain useful insights to grow the business



# Skill Sets of a Data Scientist

1. Data Scientist
2. **Skill Set**
3. Programming languages
4. Tools
5. Techniques

Statistics


Programming Languages

Data extraction & processing

Big Data processing framework

Machine Learning

Data Visualization

Data wrangling & exploration

# Programming Languages for Data Science

1. Data Scientist
2. Skill Set
3. **Programming language**
4. Tools
5. Techniques

Python

R

Julia

# Tools for Handling this Big Data (3Vs)

**Tools** are softwares that are used to apply DS techniques to perform a task.

1. Data Scientist
2. Skill Set
3. Programming language
4. **Tools**
5. Techniques

**VOLUME**



**VARIETY**

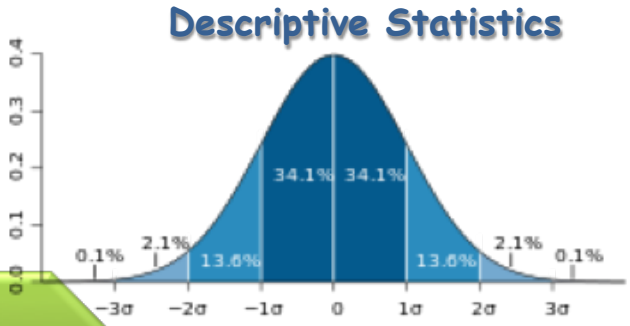

**VELOCITY**

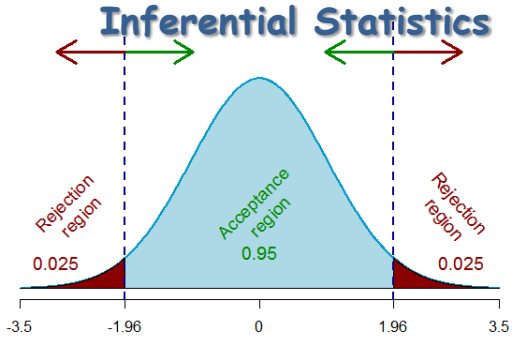

## Python Libraries for Data Science Tasks

# Techniques for Data Science

1. Data Scientist
2. Skill Set
3. Programming language
4. Tools
5. **Techniques**

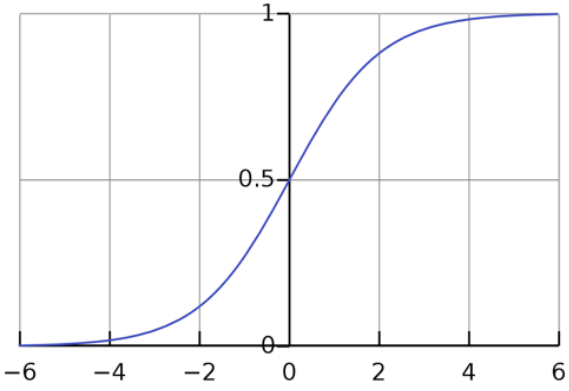
**Techniques** are set of procedures that are followed to perform a task. Tools and techniques together helps in data collection, data storage, data preparation, data analysis, data modeling and data visualization



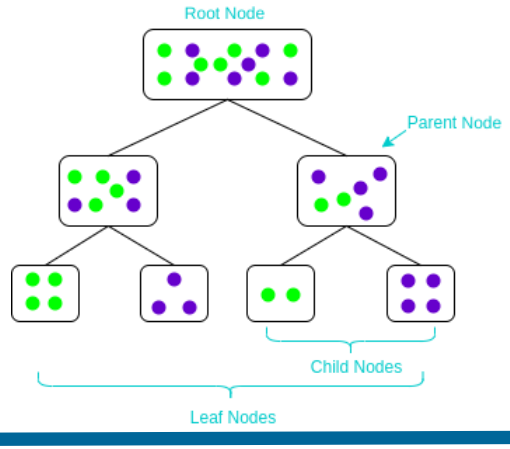




## Classification Techniques

### Logistic Regression




### Decision Tree

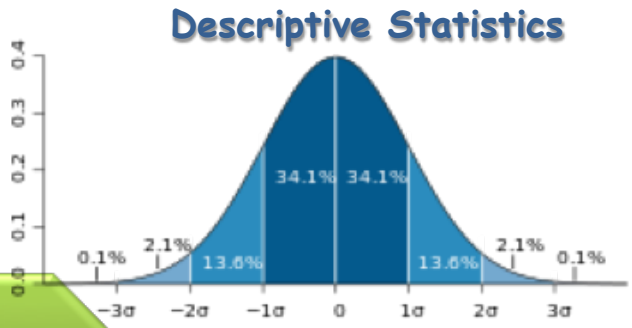


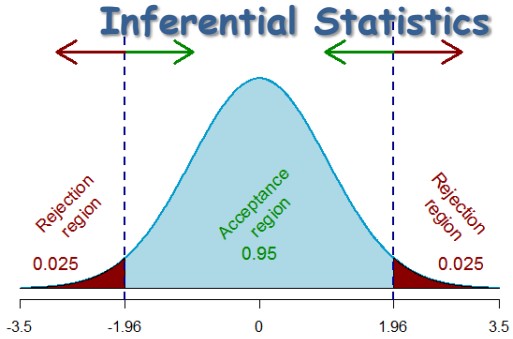

### Random Forest

# Techniques for Data Science

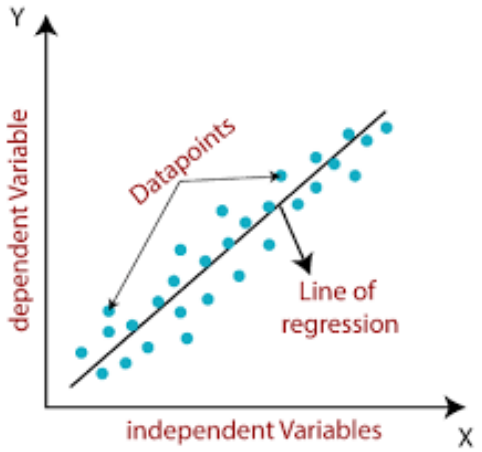
1. Data Scientist
2. Skill Set
3. Programming language
4. Tools
5. **Techniques**

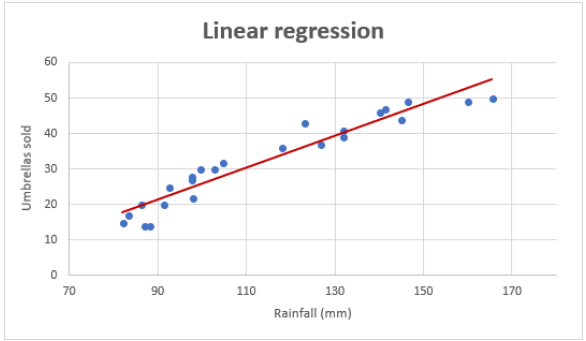
**Techniques** are set of procedures that are followed to perform a task. Tools and techniques together helps in data collection, data storage, data preparation, data analysis, data modeling and data visualization

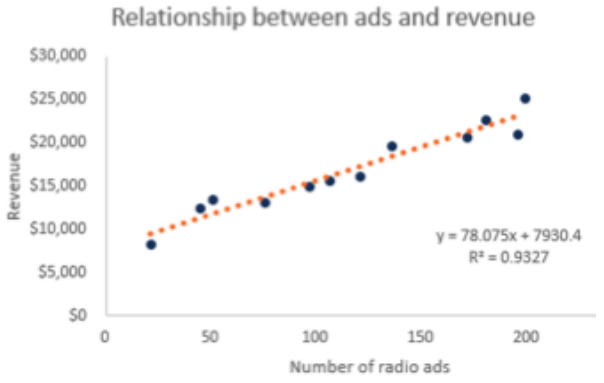



## Regression Techniques



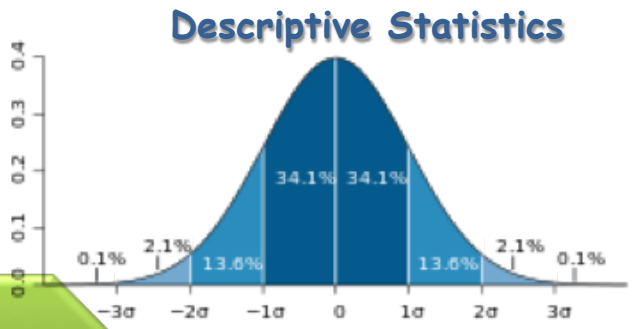



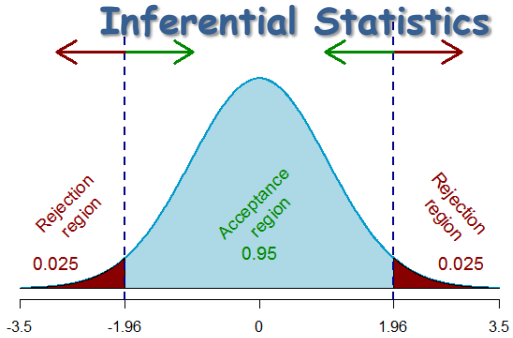



# Techniques for Data Science

1. Data Scientist
2. Skill Set
3. Programming language
4. Tools
5. **Techniques**

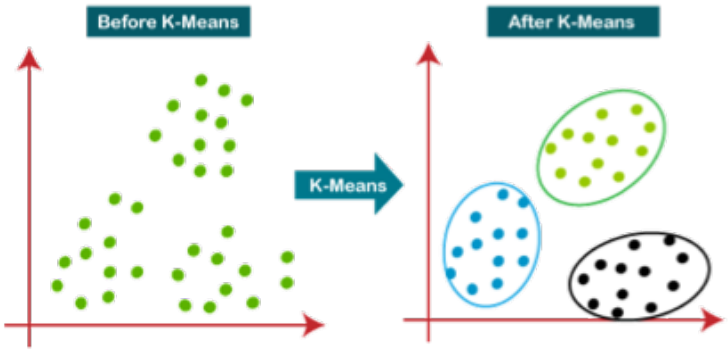
**Techniques** are set of procedures that are followed to perform a task. Tools and techniques together helps in data collection, data storage, data preparation, data analysis, data modeling and data visualization



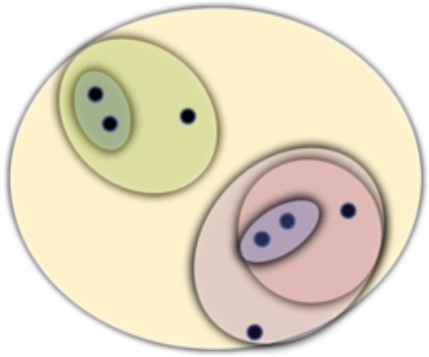




## Clustering Techniques

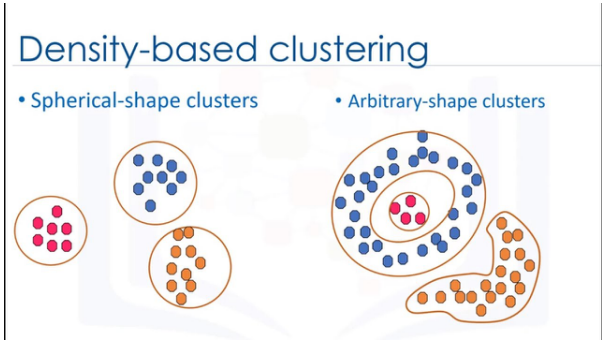
### K-Means Clustering



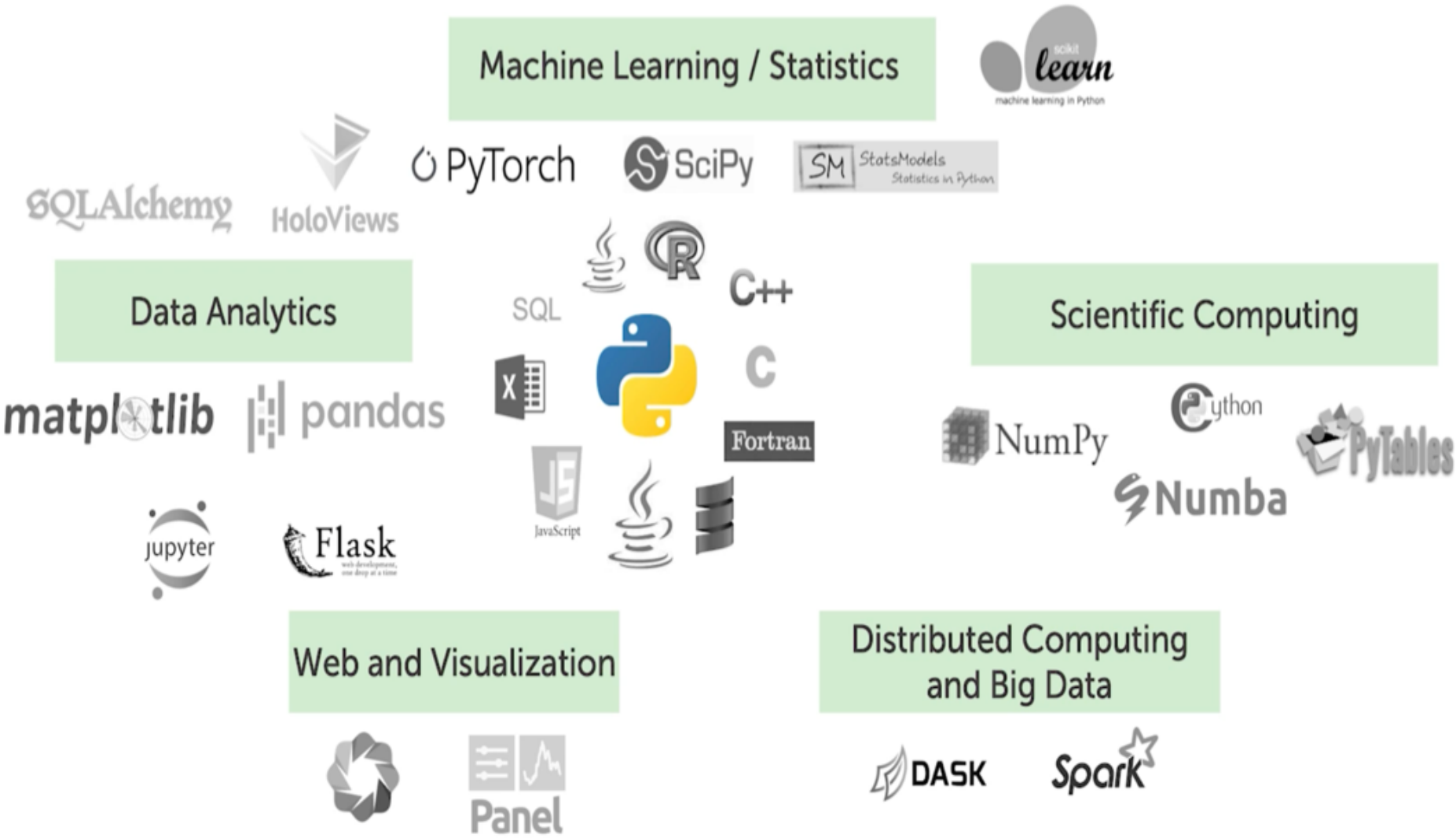

### Hierarchical Clustering



### DB SCAN

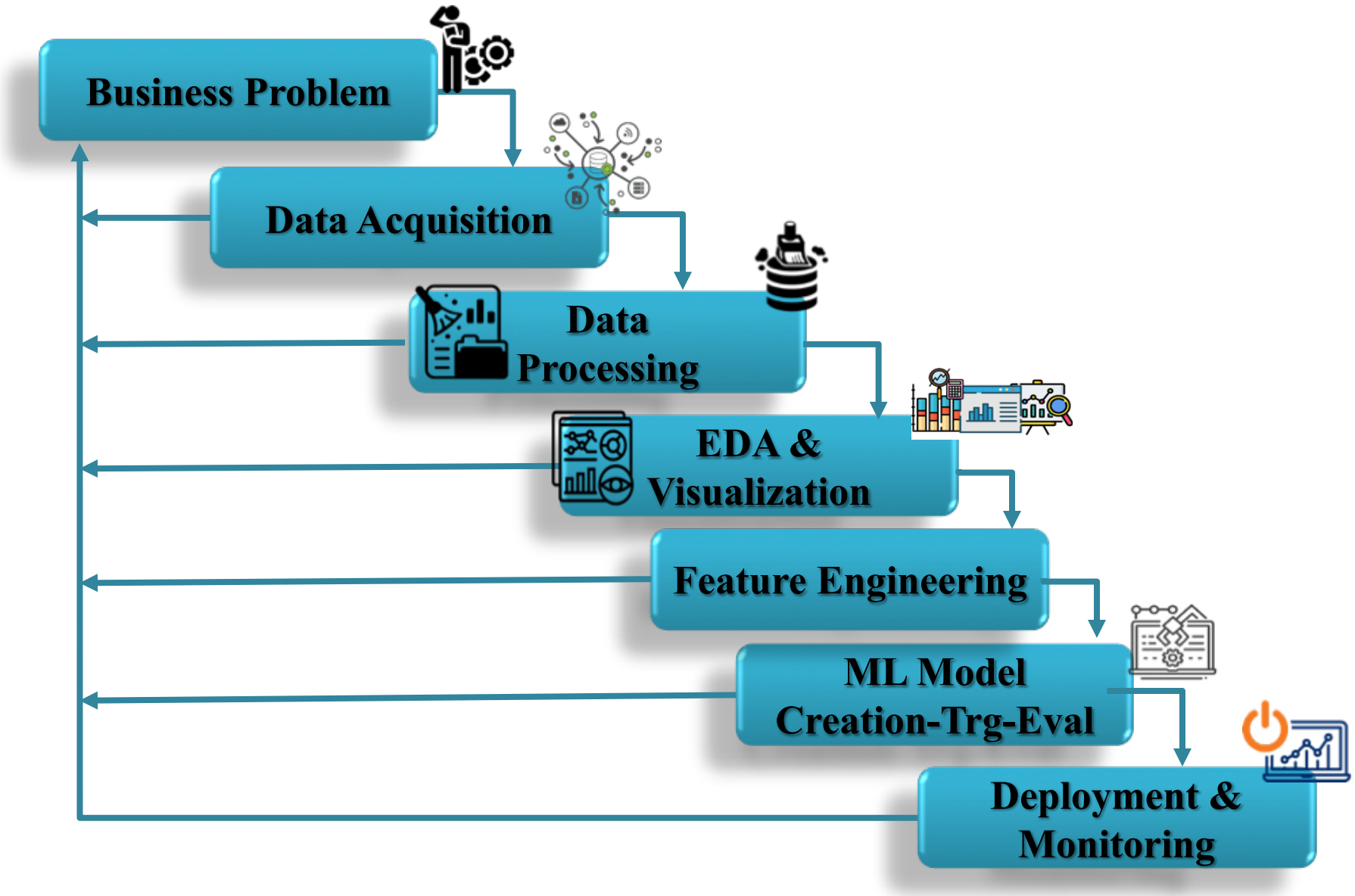


# Why is Data Science so Complicated?



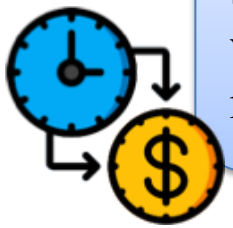
# Data Science Life Cycle

# ML Project Life Cycle



# Understanding Business Problem

1. **Business Problem**
2. Data Acquisition
3. Data Processing
4. EDA & Visualization
5. Feature Engineering
6. Model Creation-Trg-Eval
7. Deployment & Monitoring



Most critical phase of a Data Science Life Cycle, if conducted will save lot of time, money and resources.

Understand the problem by talking to the stake holders & domain experts to get the clear understanding of the problem and document all the requirements.



Identify the key business variables that need to be predicted
Define the success criteria and success measuring metrics (KPIs & SLAs)

# Data Acquisition

1. Business Problem
2. **Data Acquisition**
3. Data Processing
4. EDA & Visualization
5. Feature Engineering
6. Model Creation-Trg-Eval
7. Deployment & Monitoring

What data do we need for our project?

What are the data sources and data format?
Where is the data located?

How can we obtain the data?


What is the most efficient way to store and access all of it for later processing?

# Data Processing

1. Business Problem
2. Data Acquisition
3. **Data Processing**
4. EDA & Visualization
5. Feature Engineering
6. Model Creation-Trg-Eval
7. Deployment & Monitoring

**Extract:** Acquire data from single or multiple sources

**Transform**

Data Wrangling/Munging: Transform collected data into desired format for later analysis

Data Cleansing: Handling missing data, duplicate values, null values, mis-spelled attributes, inconsistent data types and outliers

**Load:** The transformed data is loaded into the target data source or data warehouse

rapidminer Informatica NumPy pandas

# EDA and Visualization

1. Business Problem
2. Data Acquisition
3. Data Processing
4. **EDA & Visualization**
5. Feature Engineering
6. Model Creation-Trg-Eval
7. Deployment & Monitoring

EDA involves understanding your data and identifying patterns. It involves identifying relationships and correlations between variables using visual as well as statistical techniques

NumPy pandas sas

matplotlib seaborn Qlik plotly

These patterns are not evident when you are looking at data in tables. A correct visualization tool can help you quickly gain a deeper understanding of your data

# Feature Engineering

1. Business Problem
2. Data Acquisition
3. Data Processing
4. EDA & Visualization
5. **Feature Engineering**
6. Model Creation-Trg-Eval
7. Deployment & Monitoring

## Housing Data Set

| City | Size | Covered Area | No of bedrooms | Trees near by | No of bathrooms | Schools near by | Construction Date | Price |
|---|---|---|---|---|---|---|---|---|
| Lahore | 2000 | 3500 | 3 | 1 | 3 | 1 | 25/10/2001 | 20.5 M |
| Karachi | 2600 | 3000 | 2 | 0 | 4 | 1 | 16/05/1990 | 18 M |
| Islamabad | 1800 | 2000 | 3 | 1 | 3 | 2 | 25/11/1995 | 20 M |
| Shaikhupura | 1600 | 2600 | 1 | 2 | 0 | 0 | 08/06/2020 | 5 M |
| Lahore | 2600 | 2000 | 3 | 3 | 1 | 1 | 03/09/2016 | 4 M |
| Karachi | 3000 | 1000 | 2 | 2 | 1 | 0 | 19/01/1980 | 6 M |
| Islamabad | 2000 | 3600 | 4 | 4 | 3 | 3 | 21/07/1999 | 30 M |
| Lahore | 1000 | 2000 | 3 | 0 | 1 | 2 | 12/04/2015 | 10 M |

## Merge the Features

Size of house
Number of bedrooms
Number of bathrooms
Schools around
Park around

→

Size of house
Number of bedrooms
Number of bathrooms
→ Size feature

Schools around
Park around
→ Location feature

**Feature Engineering is the process of using domain knowledge to extract features from raw data via data mining techniques**

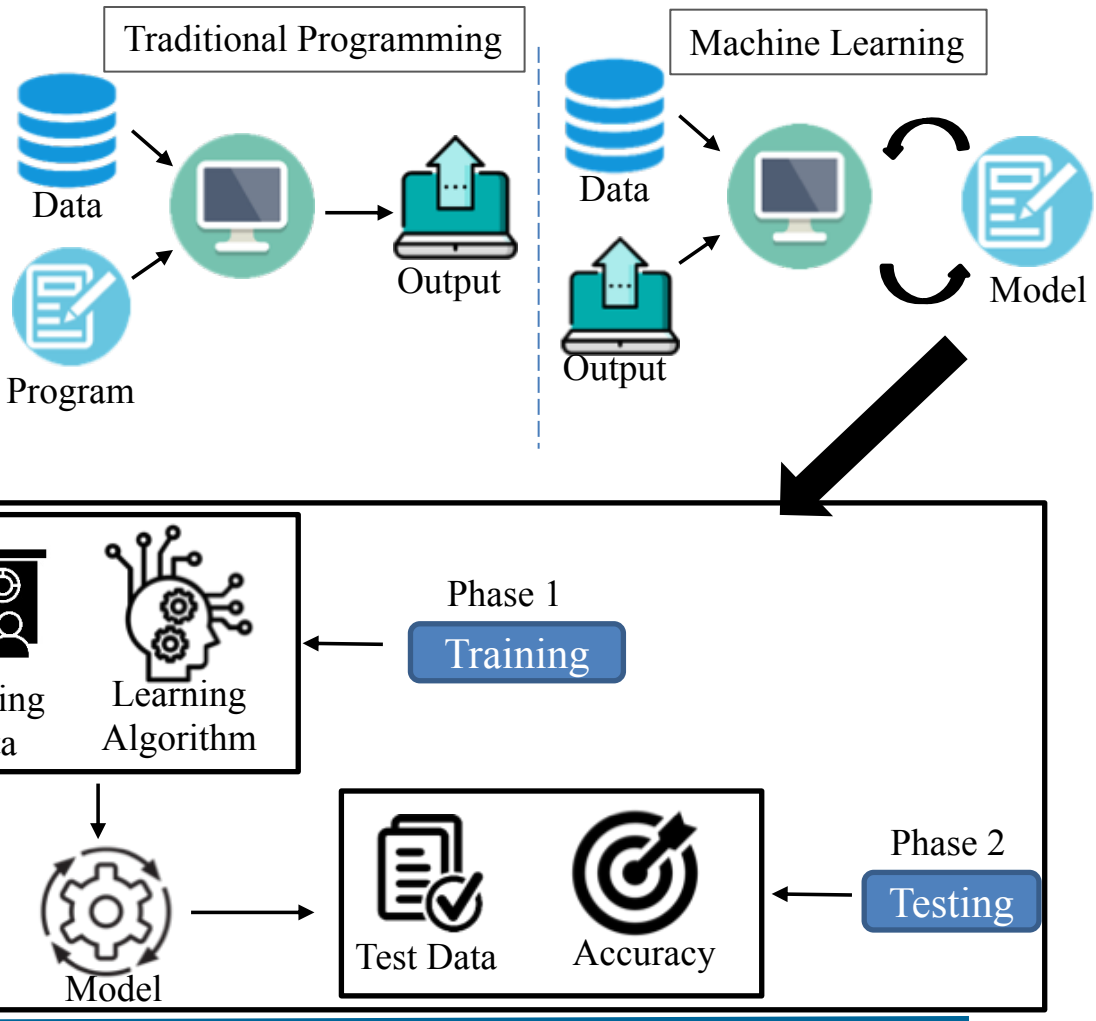
- Extracting Information
- Combining Information
- Transforming Information

# ML Model: Creation-Training-Evaluation

1. Business Problem
2. Data Acquisition
3. Data Processing
4. EDA & Visualization
5. Feature Engineering
6. **ML Model Creation-Trg-Eval**
7. Deployment & Monitoring

Use different but appropriate machine learning algorithms like Decision Tree, Linear Regression, K-Nearest Neighbour to the data to identify the model that best fits the business requirements

ML is an application of AI that gives computers the ability to learn without being explicitly programmed. [Arthur Samuel]



# Model Deployment and Monitoring

1. Business Problem
2. Data Acquisition
3. Data Processing
4. EDA & Visualization
5. Feature Engineering
6. Model Creation-Trg-Eval
7. **Deployment & Monitoring**

After a model is trained, tuned and tested, you can deploy the model into production and make inferences (predictions)

Check the deployment environment for dependency issues
Deploy the model first in the test and then in the production environment

Cloud Deployment


Most of the times the live real world data differ from the data that was used to train the model, thus making the model less accurate. To handle this, build a model monitor that detects deviations such as data drift and alerts you to take remedial actions

# Industry Job Roles in Data Science

# Industry Job Roles: Data Scientist

1. **Data Scientist**
2. Data Engineer
3. Data Analyst
4. Database Administrator
5. ML Engineer

- Senior most in the team and take inputs from the rest to formulate actionable insight for the business
- Makes use of the latest tools and technologies in finding solutions and reaching conclusions that are crucial for an organization's growth and development


# Industry Job Roles: Data Engineer/Architect

1. Data Scientist
2. **Data Engineer**
3. Data Analyst
4. Database Administrator
5. ML Engineer

- Scrape data and store it in warehouses using ETL
- Handle databases and create data warehouses
- Design, build, and manage the big data infrastructure
- Build data pipelines for easy access of data
- Big Data Tools (Apache Spark, Apache Hive, Hadoop)
- Cloud Platforms (AWS, Google Cloud Platform)


# Industry Job Roles: Data Analyst

1. Data Scientist
2. Data Engineer
3. **Data Analyst**
4. Database Administrator
5. ML Engineer

- Data Analyst is an entry level member into the data analytics team
- Needs to have good technical skills and know the basics of statistics, data munging, data utilization, and exploratory data analysis
- Generate reports after analyzing the data
- Can move to the role of Data engineer and Data scientist with more experience


# Industry Job Roles: Database Administrator

1. Data Scientist
2. Data Engineer
3. Data Analyst
4. **Database Administrator**
5. ML Engineer

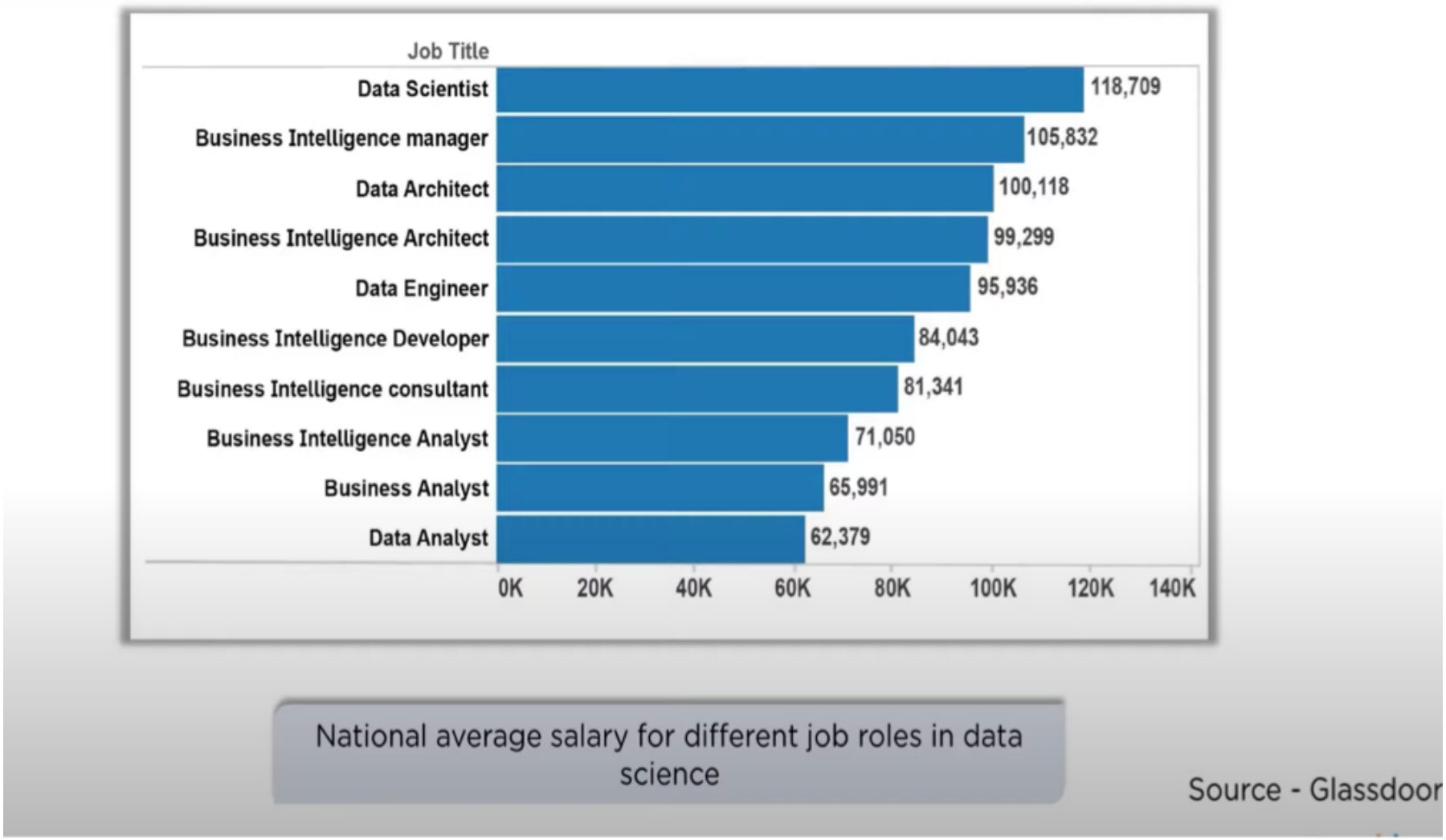
- Responsible for administering the collected data by installing, configuring, monitoring, operating, and maintaining database
- Ensure that all databases are available to all relevant users, and is protected securely from any malicious activity


# Industry Job Roles: Machine Learning Engineer

1. Data Scientist
2. Data Engineer
3. Data Analyst
4. Database Administrator
5. **ML Engineer**

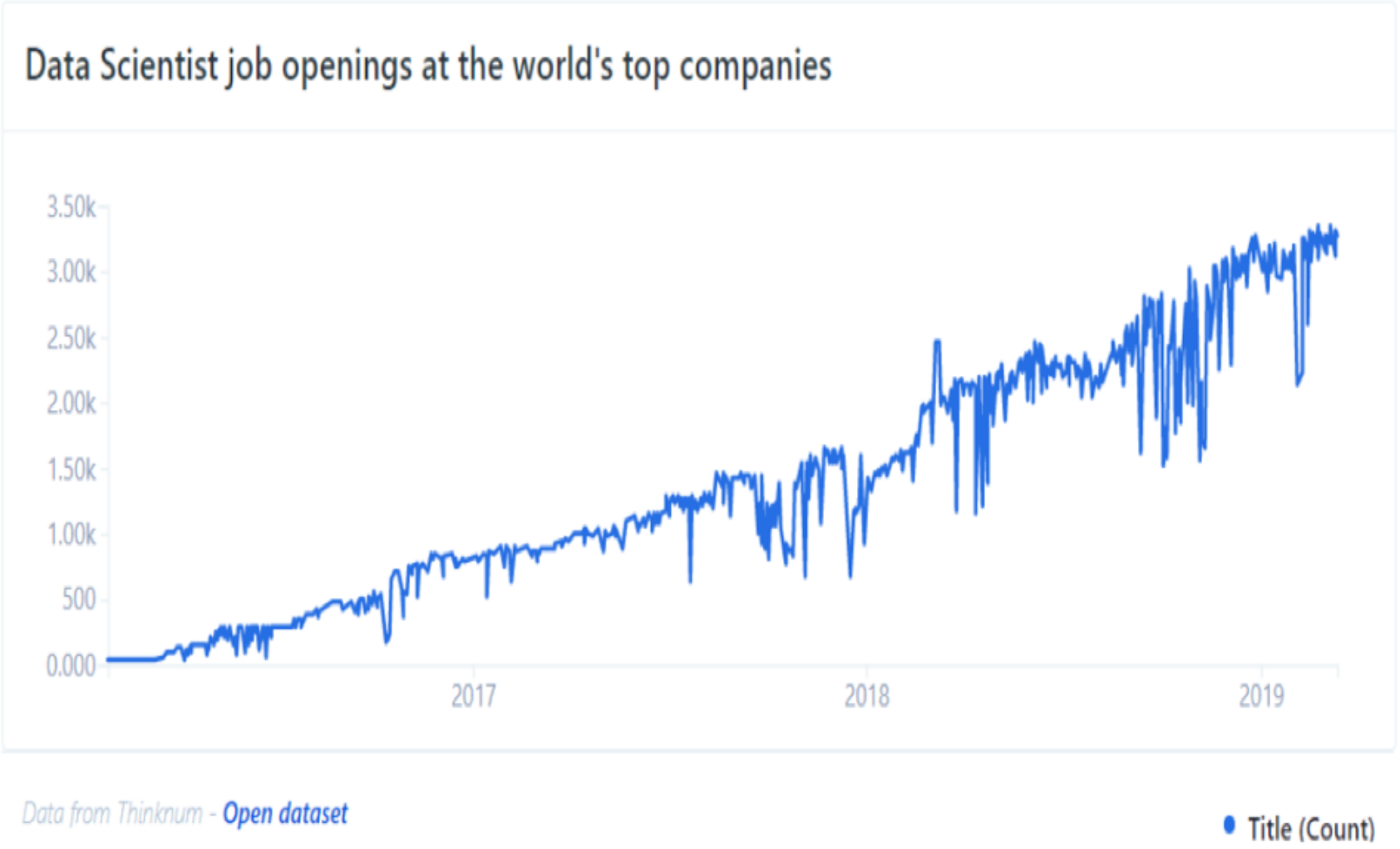
- Machine learning engineer works as a part of large data science team
- Responsible to design and create all algorithms capable of learning and making predictions
- They are expected to perform A/B testing, build data pipelines, and implement algorithms for classification, clustering, regression, anomaly detection etc.


# History: Data Science Salary Trends




Source: https://towardsdatascience.com/why-learn-data-science-in-2020-d3f54123b2e4

# History: Job trends




Source: https://www.tecla.io/blog/the-high-demand-for-data-scientists-and-how-to-hire-for-them/

# Discussion on Course Matrix

Visit Course Website: http://www.arifbutt.me/

# What we will do in this course?

## Module 1: (Overview of the course)

- What is Data Science?
- Why/How to do Data Science?
- Structured vs Unstructured data
- Applications of Data Science
- Tools and Technologies for Data Science
- Life Cycle of a Data Science Project
- Job Roles in the Industry
- Data Science Use Cases from real life
- Git and Github for Data Scientists

## Reading Tasks:

- …

# What we will do in this course?

## Module 2: (Basics of Python Programming)

- Overview of Python programming language
- Python programming environments
- Python intrinsic data types and operators
- Python data structures
- Selection and Repetition structures
- Functions in Python
- Exception handling
- Modules, packages and libraries
- Basic file handling in Python
- Regular Expressions in Python

## Reading Tasks:

- …

# What we will do in this course?

## Module 3: (Python for Data Scientists)

- Overview of Python libraries for Data Science
- Reading data in Python (csv, xlsx, json)
- Data manipulation with NumPy
- Scientific computation with SciPy
- Data manipulation with Pandas
- Visualization with Matplotlib and Seaborn

## Reading Tasks:

- …

# What we will do in this course?

**Module 4: (Mathematics for Data Scientists)**

- Applied Linear Algebra for Data Scientists
- Applied Descriptive Statistics for Data Scientists
- Applied Inferential Statistics for Data Scientists
- Applied Calculus for Data Scientists

**Reading Tasks:**

- …

# What we will do in this course?

## Module 5: (Data Acquisition)

- Overview of Data Acquisition
- Data Acquisition from Websites (Web Scraping)
- Data Acquisition from SQL Databases
- Data Acquisition from NoSQL Databases

## Reading Tasks:

- …

# What we will do in this course?

## Module 6: (Machine Learning : A Bird's-eye View)

- Overview of Machine Learning
- Categories of Machine Learning Types and Algorithms
- Python for Machine Learning (Scikit-learn)
- Will do hands on practice for
  - ✓ Model creation
  - ✓ Model training
  - ✓ Model evaluation
  - ✓ Feature engineering
  - ✓ Dimensionality reduction

## Reading Tasks:

- …

# What we will do in this course?

## Module 7: (NLP : A Bird's-eye View)

- Overview of NLP
- Text Pre-Processing Techniques
- Text Vectorization Techniques
- Applying Machine Learning Models on Textual Data

## Reading Tasks:

- …

# What we will do in this course?

## Module 8: (Deep Learning: A Bird's-eye View)

- Machine Learning vs Deep Learning
- Overview of Deep Learning Models (CNN vs RNN)
- Deep Learning Applications
  - ✓ Image recognition
  - ✓ Self-driving cars
  - ✓ Language translation services
- A Hello World on Deep Learning Project using
  - ✓ TensorFlow/Keras/Theano/Pytorch/Caffe

## Reading Tasks:

- …

# What we will do in this course?

## Module 9: (Big Data: A Bird's-eye View)

- What is Big Data?
- Big Data Storage and Processing Frameworks
  - ✓ Apache Hadoop with MapReduce (used by Alibaba, AOL)
  - ✓ Apache Storm (used by Twitter, Spotify)
  - ✓ Apache Spark (used by Netflix, Yahoo, eBay)
  - ✓ Apache Hive (used by Facebook, Walmart)
- An Overview of Hadoop Ecosystem
  - ✓ Data Storage (HDFS, HBASE)
  - ✓ Data Processing (YARN, Map Reduce)
  - ✓ Data Access (Hive, Pig, Mahout, Avro, Sqoop)
  - ✓ Data Management (Oozie, Chukwa, Flume, ZooKeeper)

## Reading Tasks:

- …

# Things To Do

- Should have a very clear understanding of different data sources, its types and storage
- Must know the applications of data science in different domains.
- While going through todays lecture slides click all the tools and technologies, which have been hyperlinked to respective web sites.


- Have a very clear understanding of Data Science Life Cycle, the tools & the technologies used in each phase.
- Think of few use cases where you can apply Data Science, Machine Learning and Deep Learning technologies and make a list of the skill set you need to develop/learn to implement and deploy such projects.

**Coming to office hours does NOT mean you are academically weak!**