Webinar

Machine Learning and Big Data

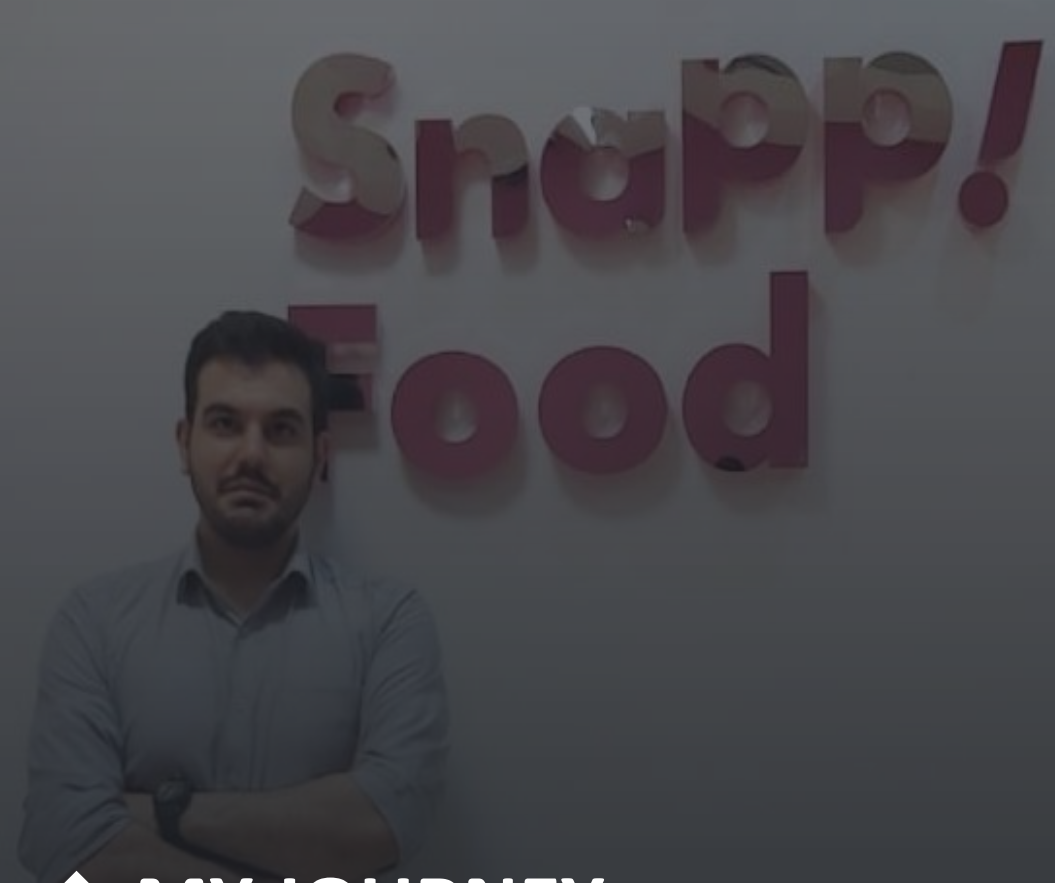
Why and How?

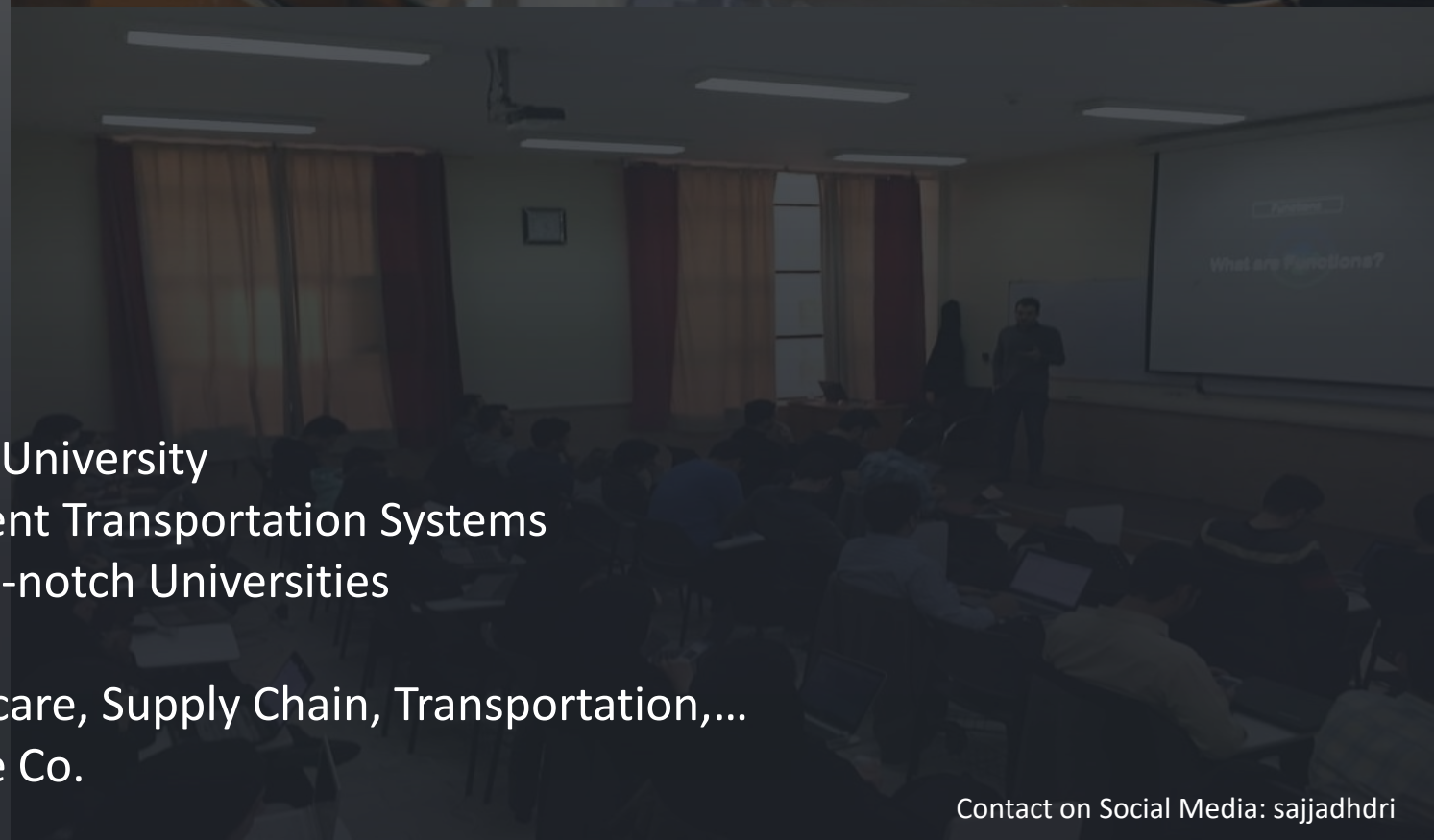
Lecturer:

Sajad Heydari

Aspiring Data Scientist
MSc. Industrial Engineering

Winter 2022





❖ MY JOURNEY

MSc. Industrial Engineering, Tarbiat Modares University
Data Science Researcher especially in Intelligent Transportation Systems
Data Science Courses Instructor in Iranian top-notch Universities
Data Scientist at Snapp! Food
Data Science Consultant in Industries: Healthcare, Supply Chain, Transportation,…
Marketing Data Analyst at Pasargad Insurance Co.

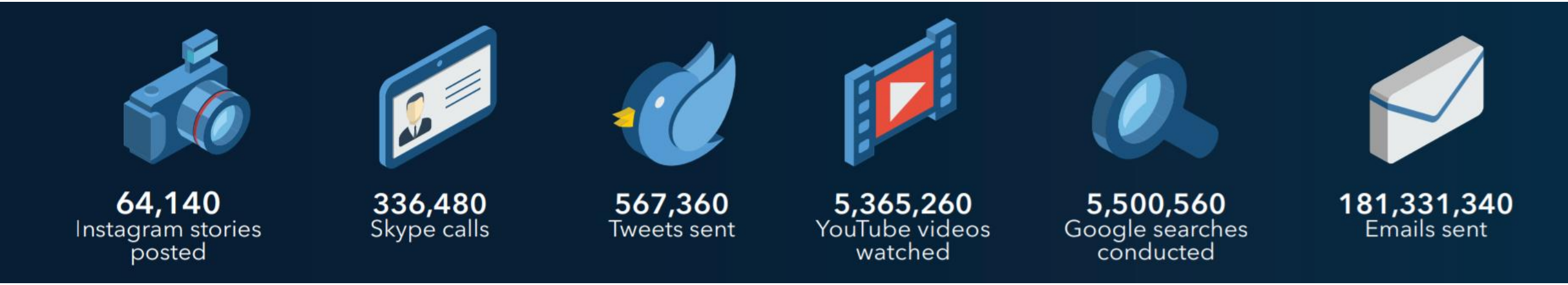
Contact on Social Media: sajjadhdri

Importance of Data

5 Exabyte Data generated every minutes NOW!

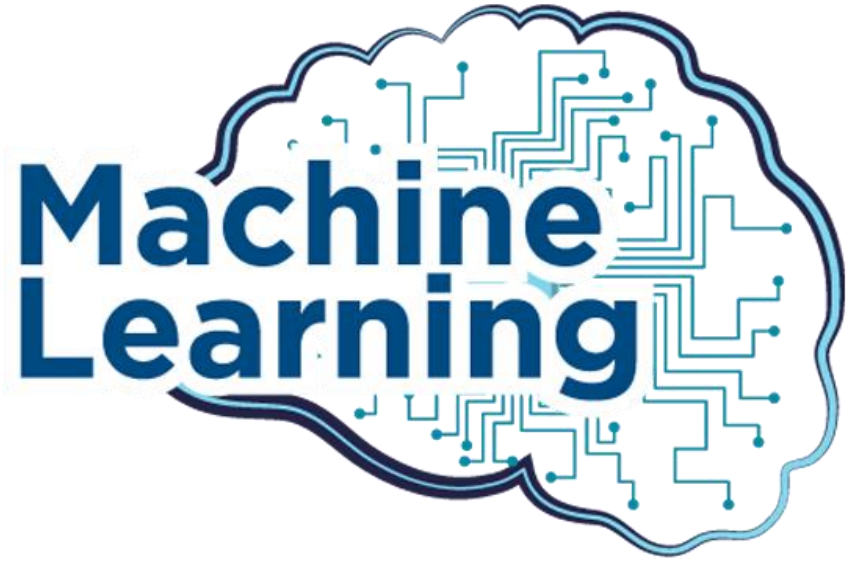
ONLY 0.5% of generated Data are analyzed

Data Explosion



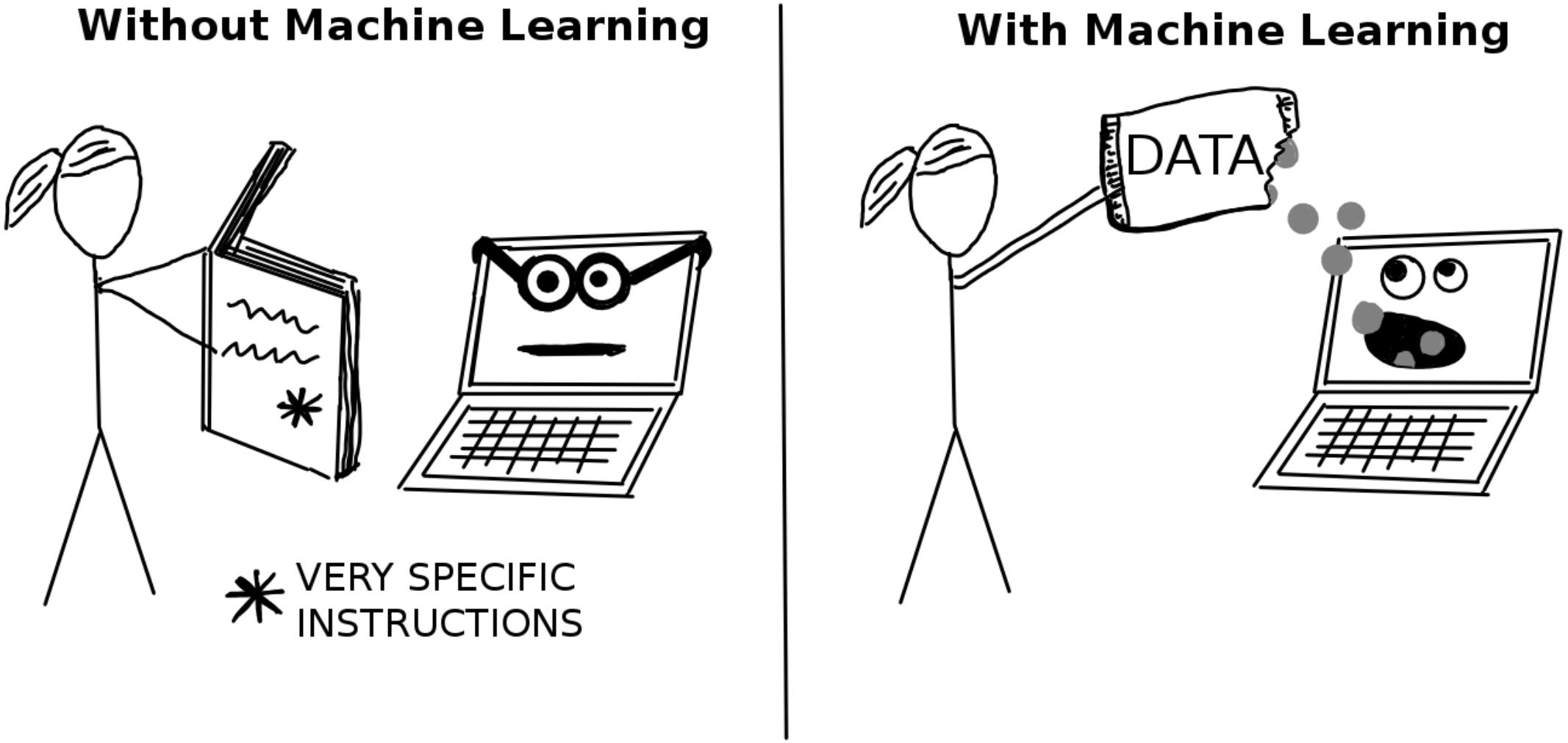
Picture source: SAS

What is Machine Learning?

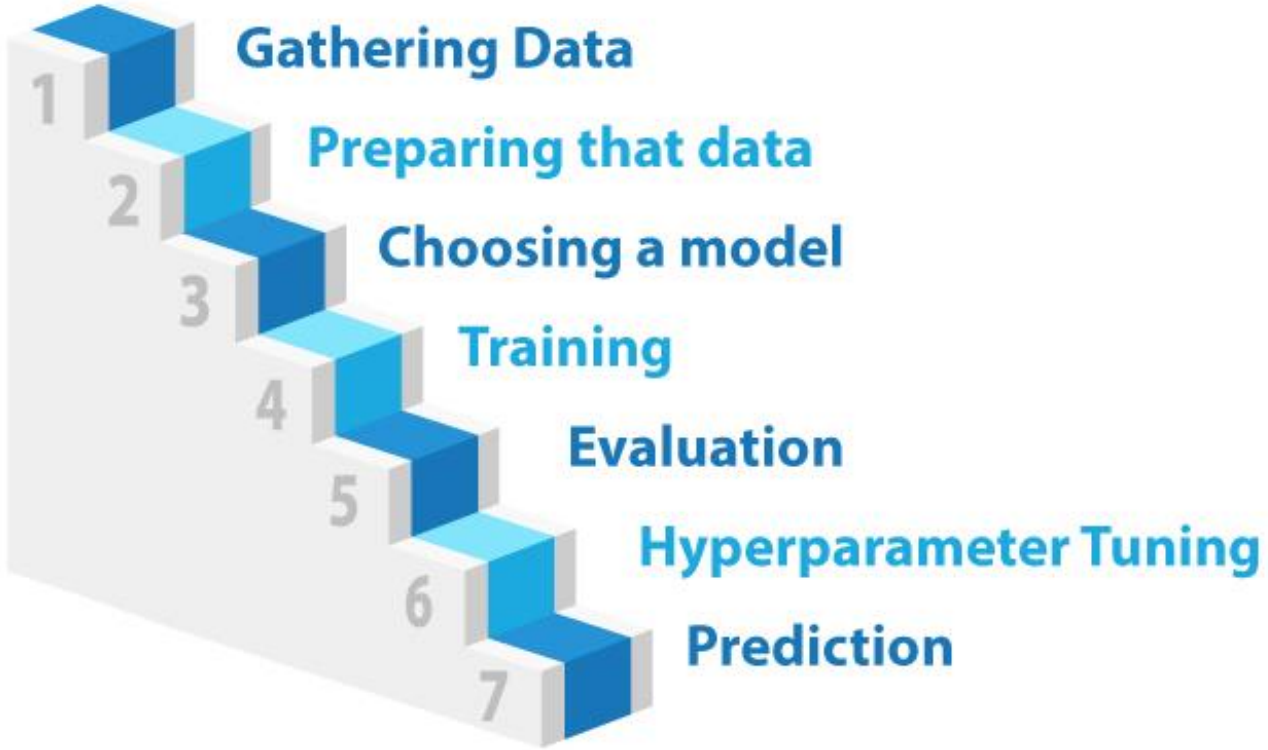


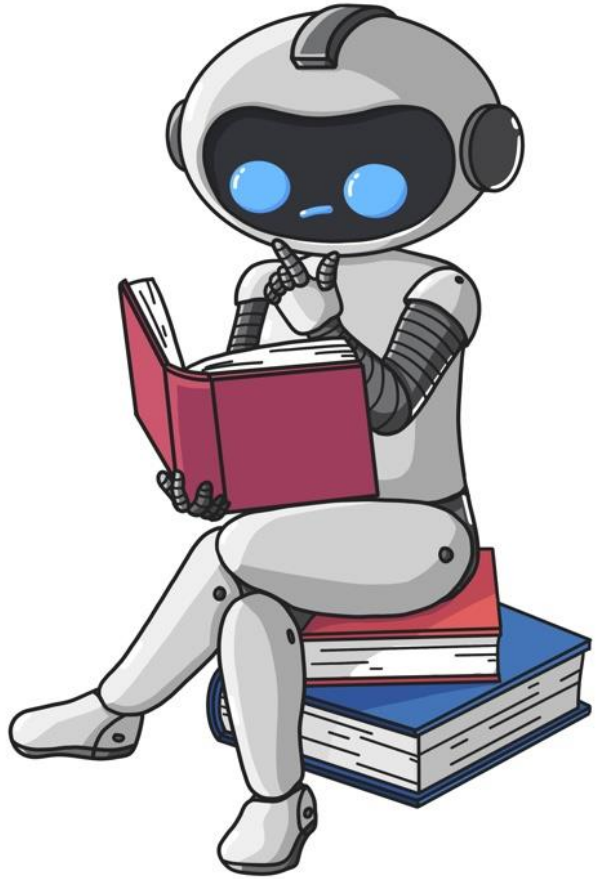
Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

What is Machine Learning?



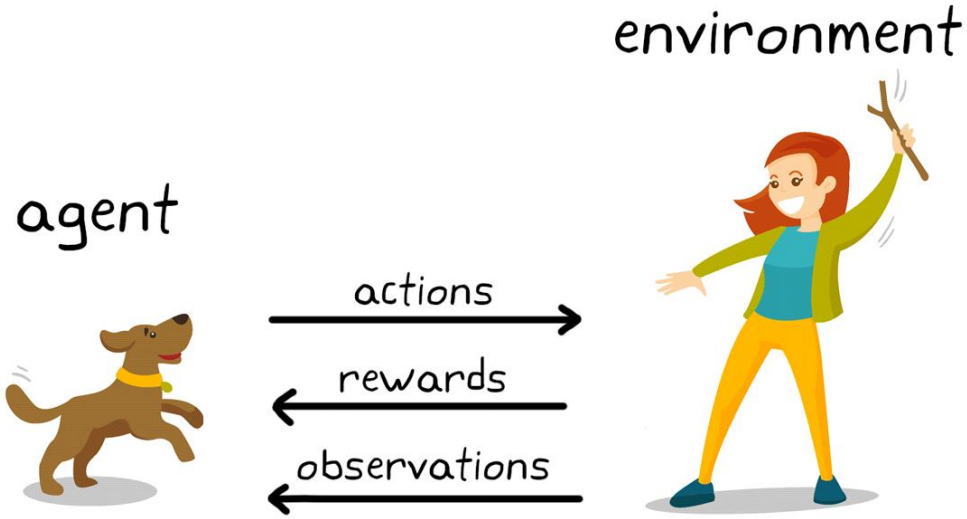
Steps of Machine Learning



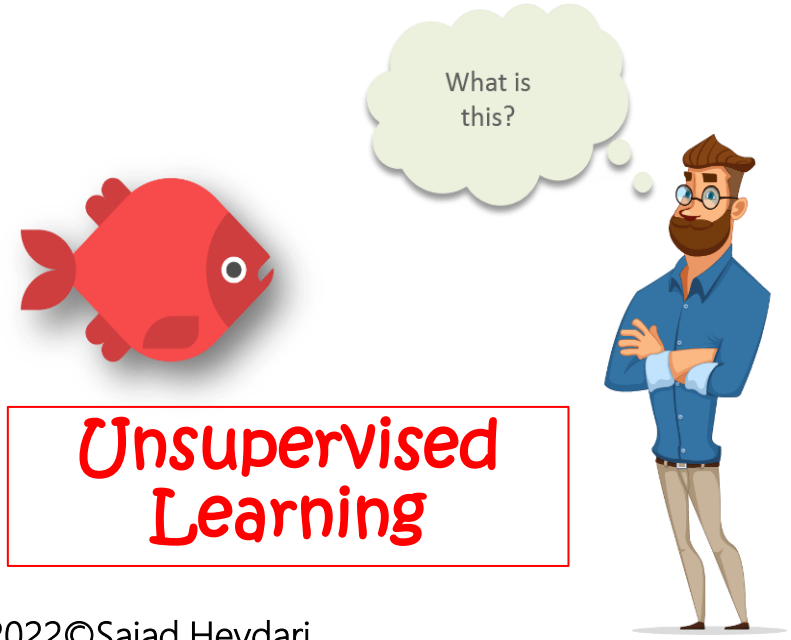


ML Sections

Supervised Learning

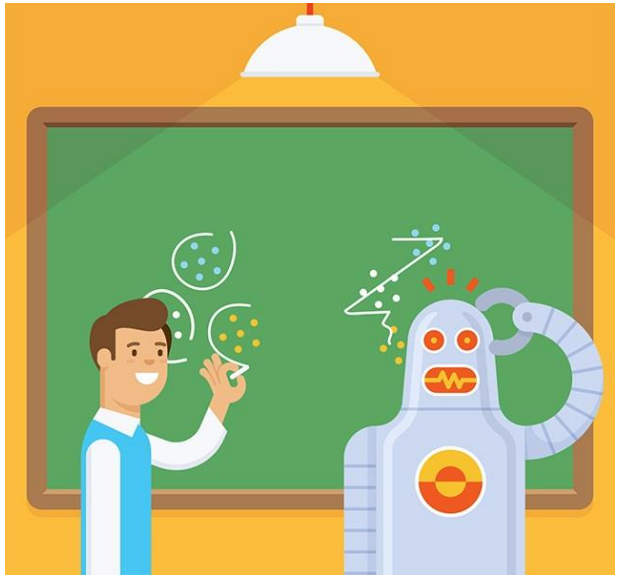


Reinforcement Learning

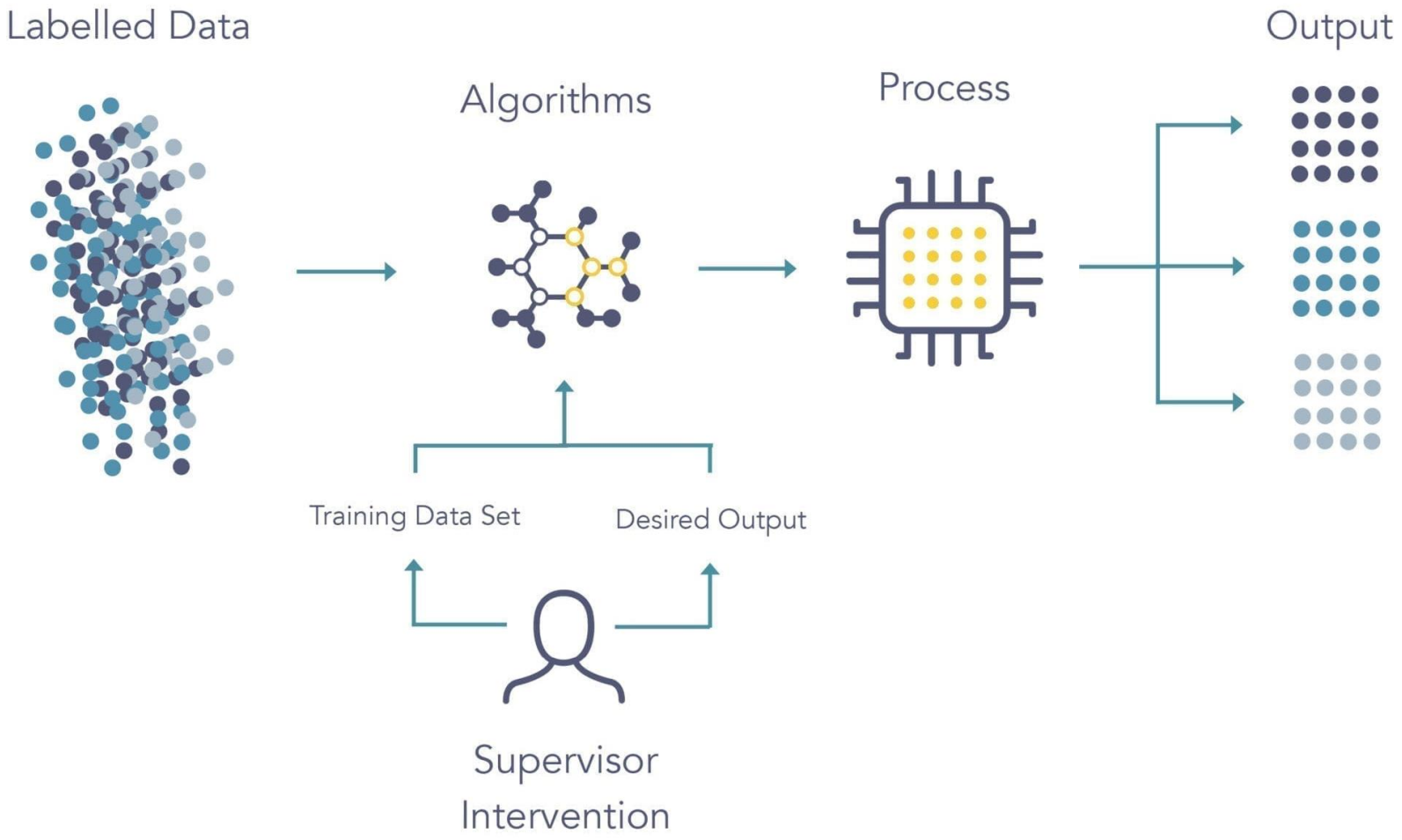


Unsupervised Learning

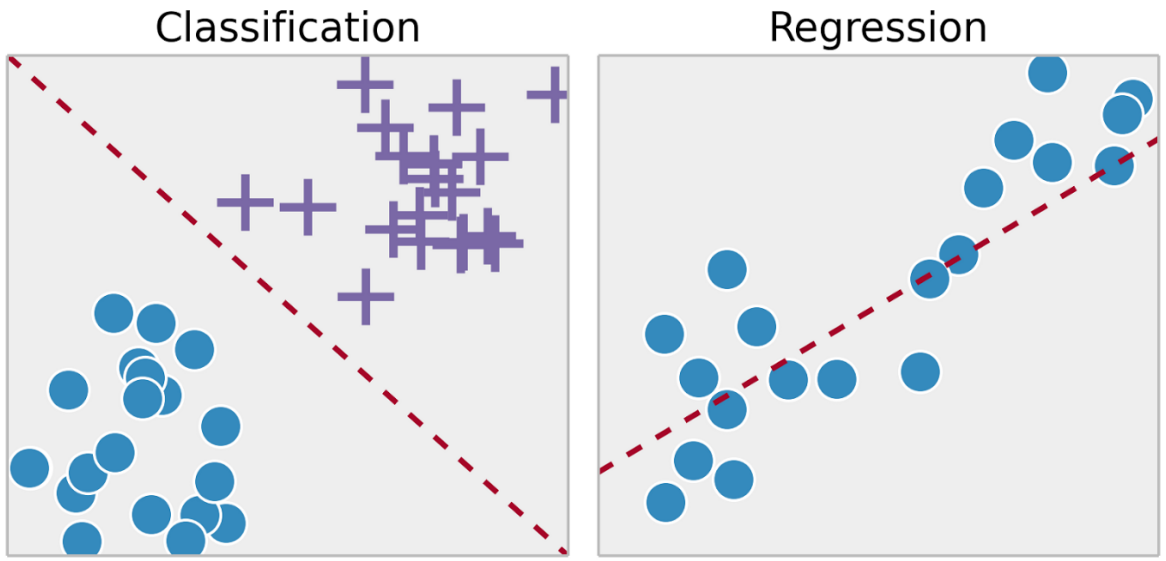
Semi-Supervised Learning



What is Supervised Learning?



Supervised Learning Sections

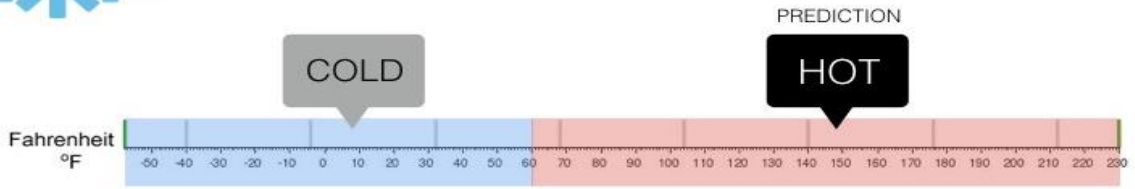


Regression

What is the temperature going to be tomorrow?

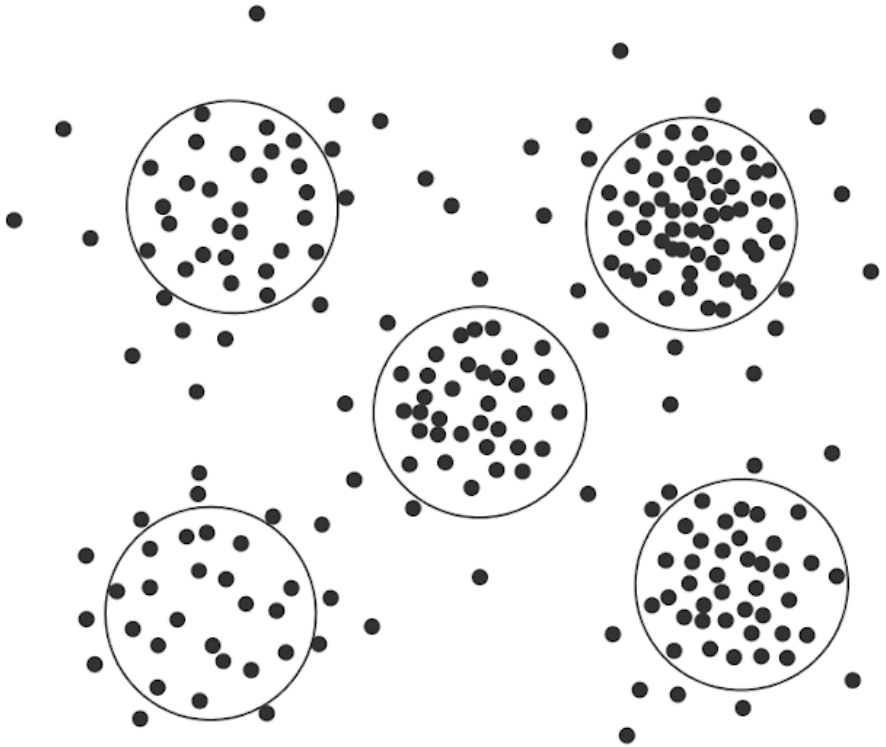
Classification

Will it be Cold or Hot tomorrow?

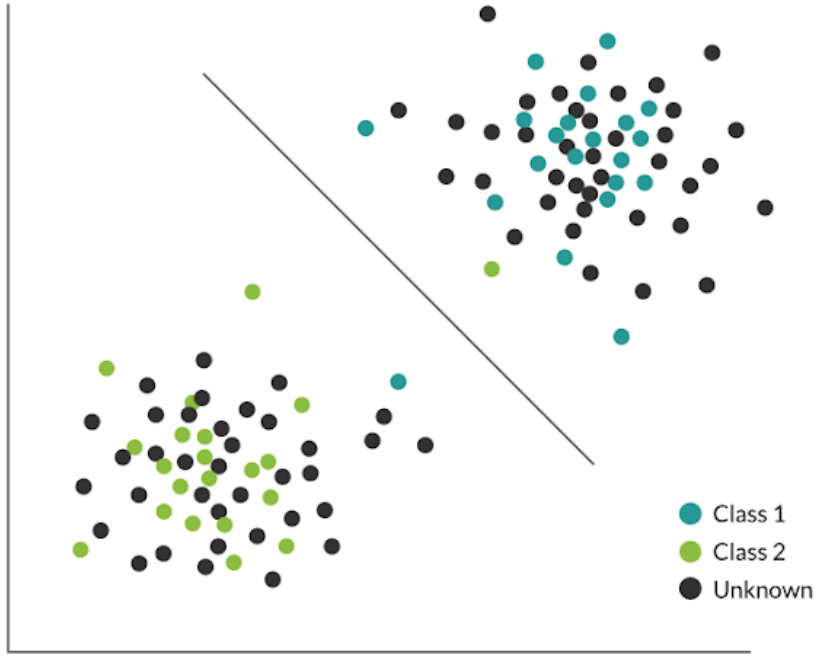


What is Unsupervised Learning?

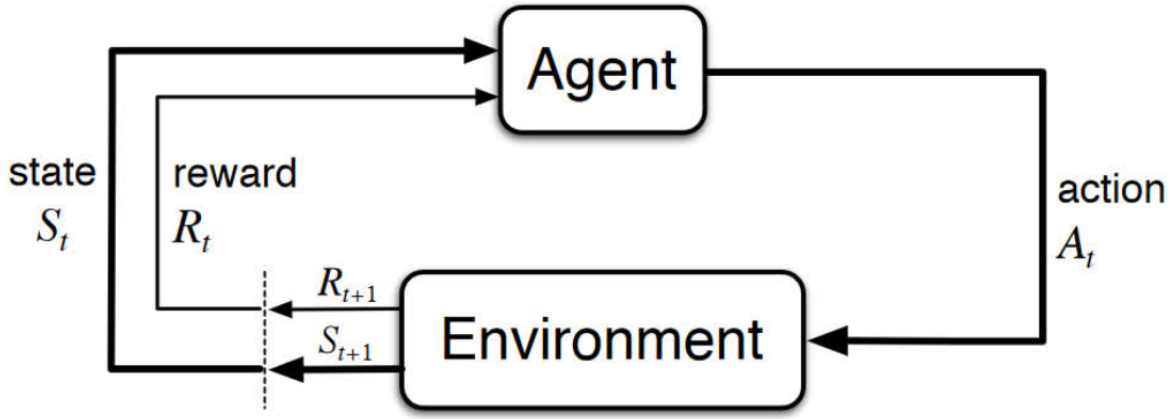
Unsupervised



Supervised



What is Reinforcement Learning?

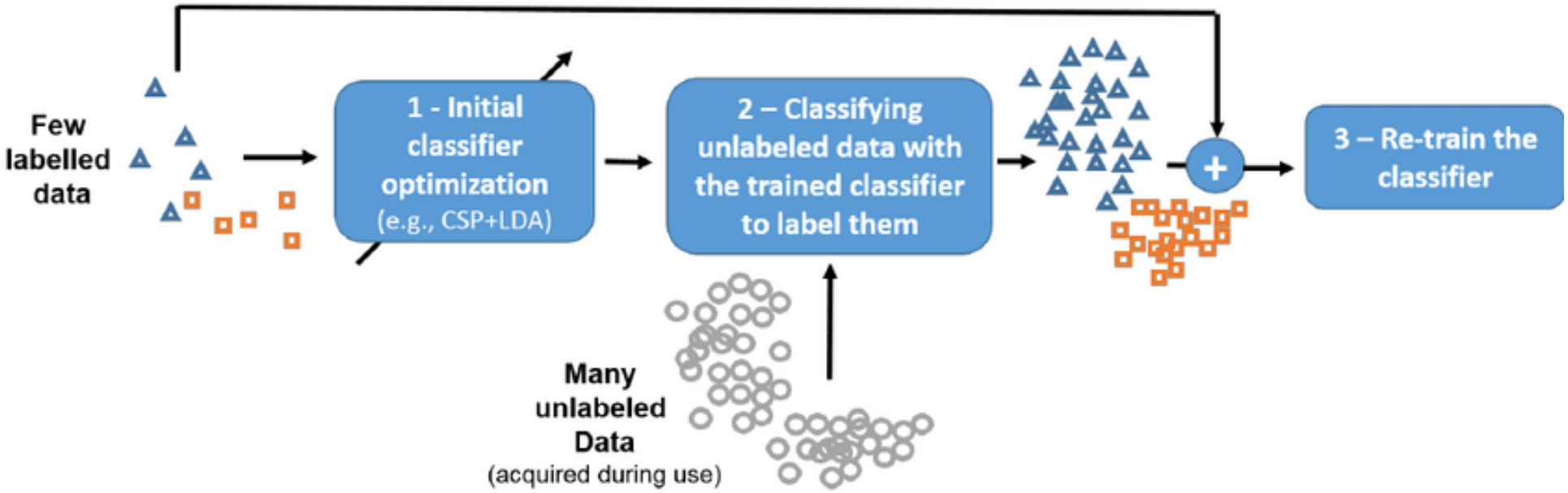


Let's break down this diagram into steps.

1. At time t, the environment is in state S_t.
2. The agent observes the current state and selects action A_t.
3. The environment transitions to state S_{t+1} and grants the agent reward R_{t+1}.
4. This process then starts over for the next time step, $t+1$.
    - Note, $t+1$ is no longer in the future, but is now the present. When we cross the dotted on the bottom left, the diagram shows $t+1$ transforming into the current time step t sc that S_{t+1} and R_{t+1} are now S_t and R_t.

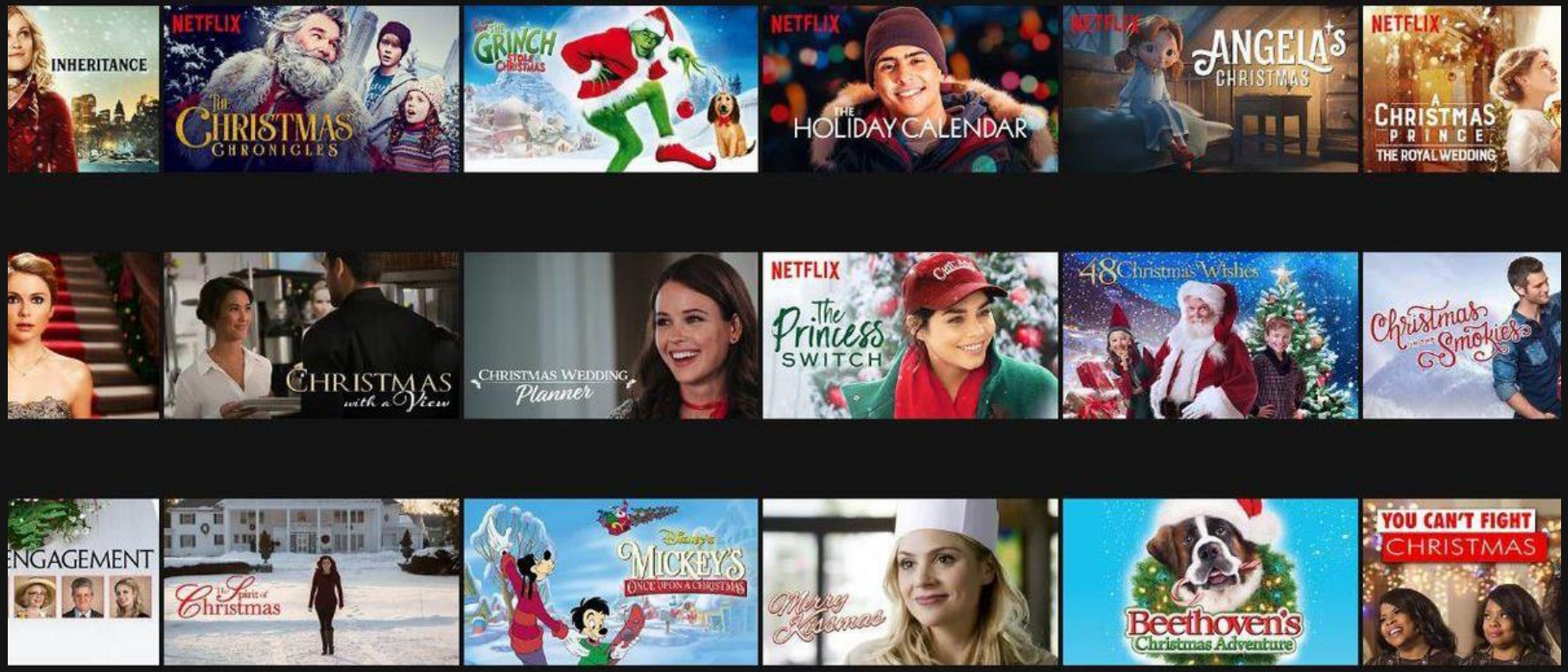
Reinforcement learning is the science of decision-making.

What is Semi-Supervised Learning?



Source: Lotte, F., (2015). Signal Processing Approaches to Minimize or Suppress Calibration Time in Oscillatory Activity-Based Brain–Computer Interfaces, Proceedings of the IEEE, 103(6), pp. 871-890.

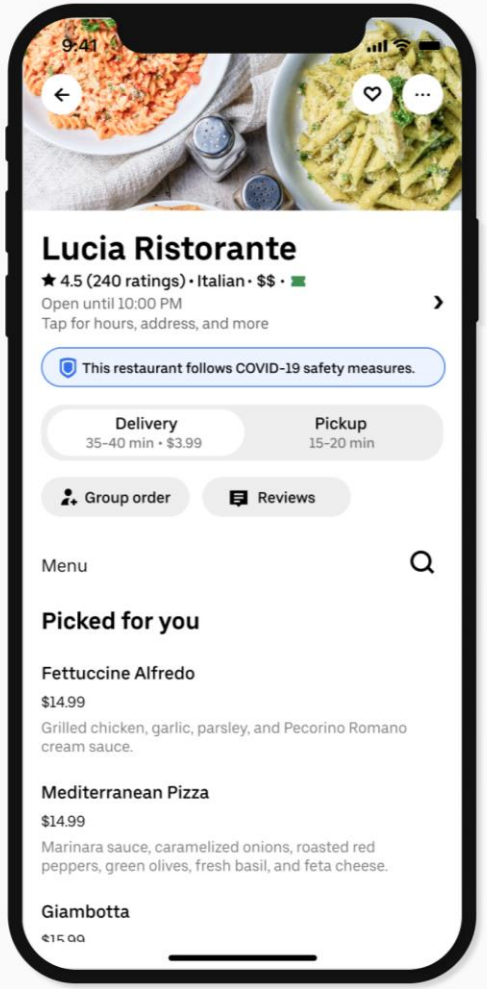
ML Cases

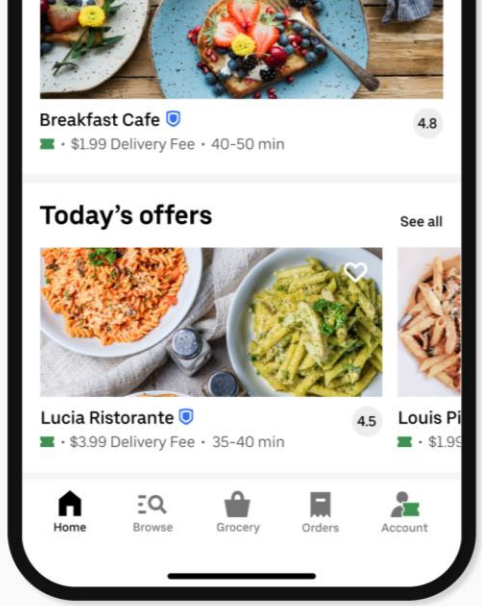


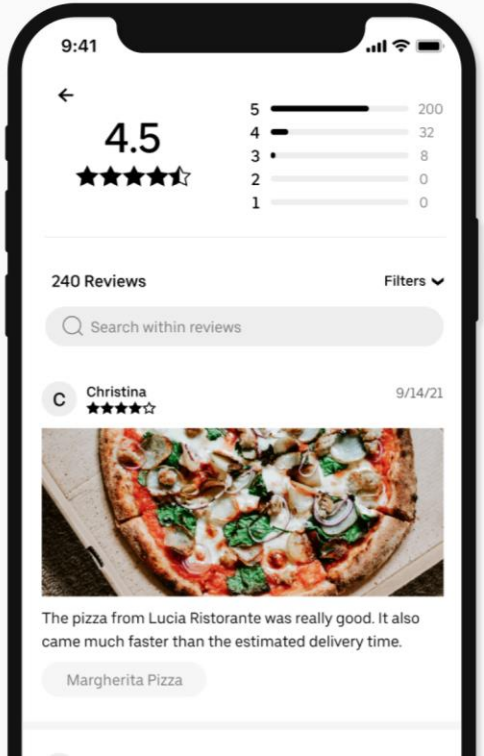
Machine Learning Project on Netflix Recommendation System

NETFLIX

ML Cases







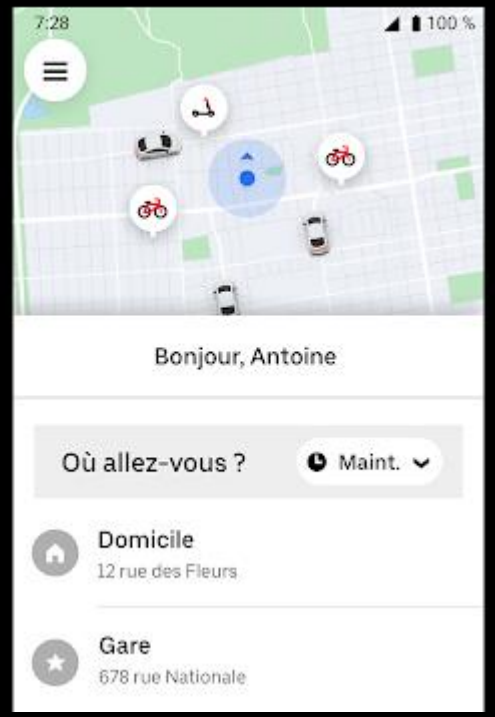
ML Cases

amazon

ML Cases

Find Best Price

Optimize Waiting Time and Idle Vehicle



UBER

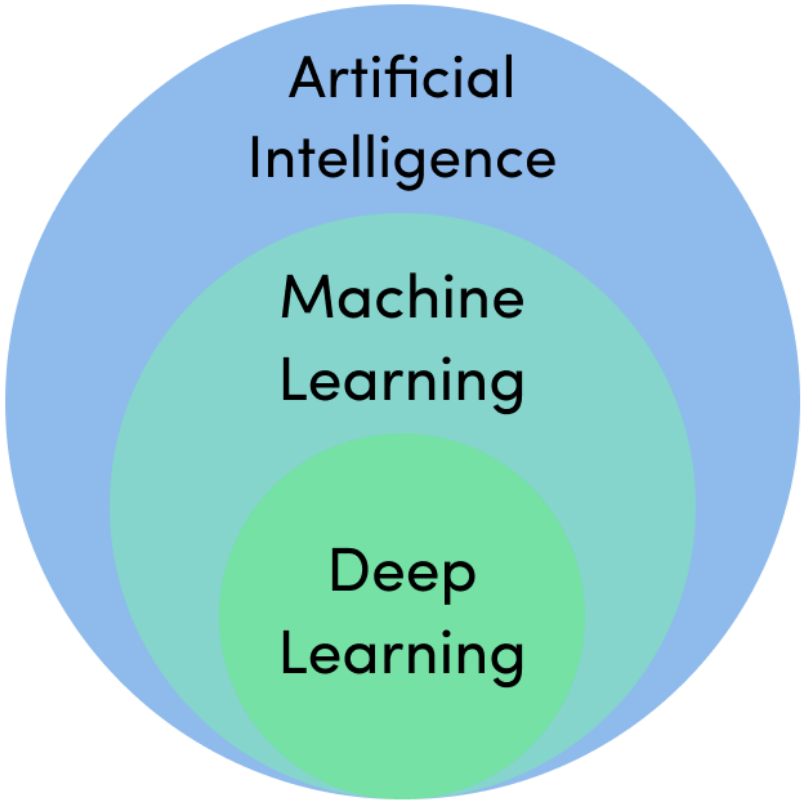
ML Cases

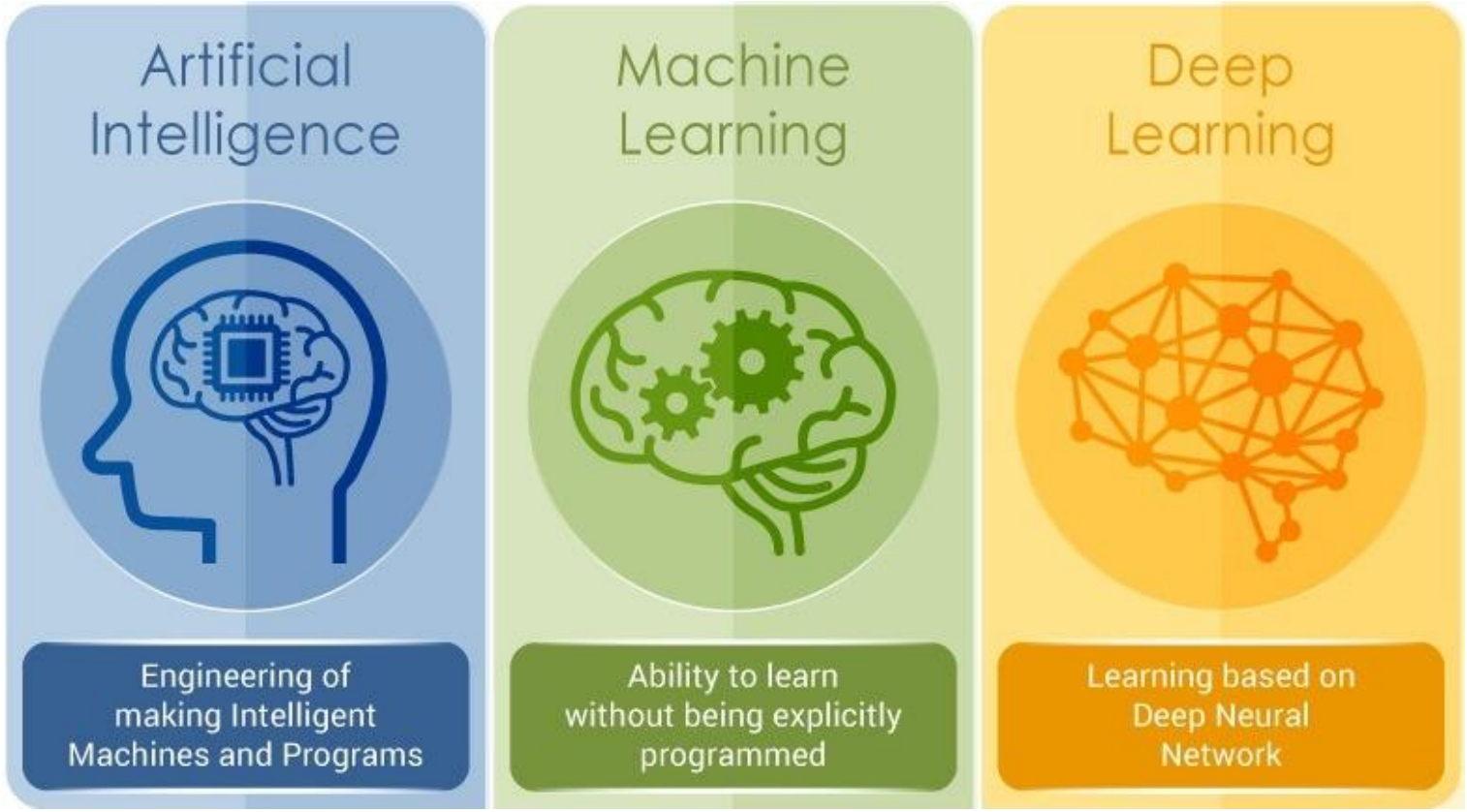
EA SPORTS FUT22

ML Cases

QATAR AIRWAYS القطرية

Deep Learning





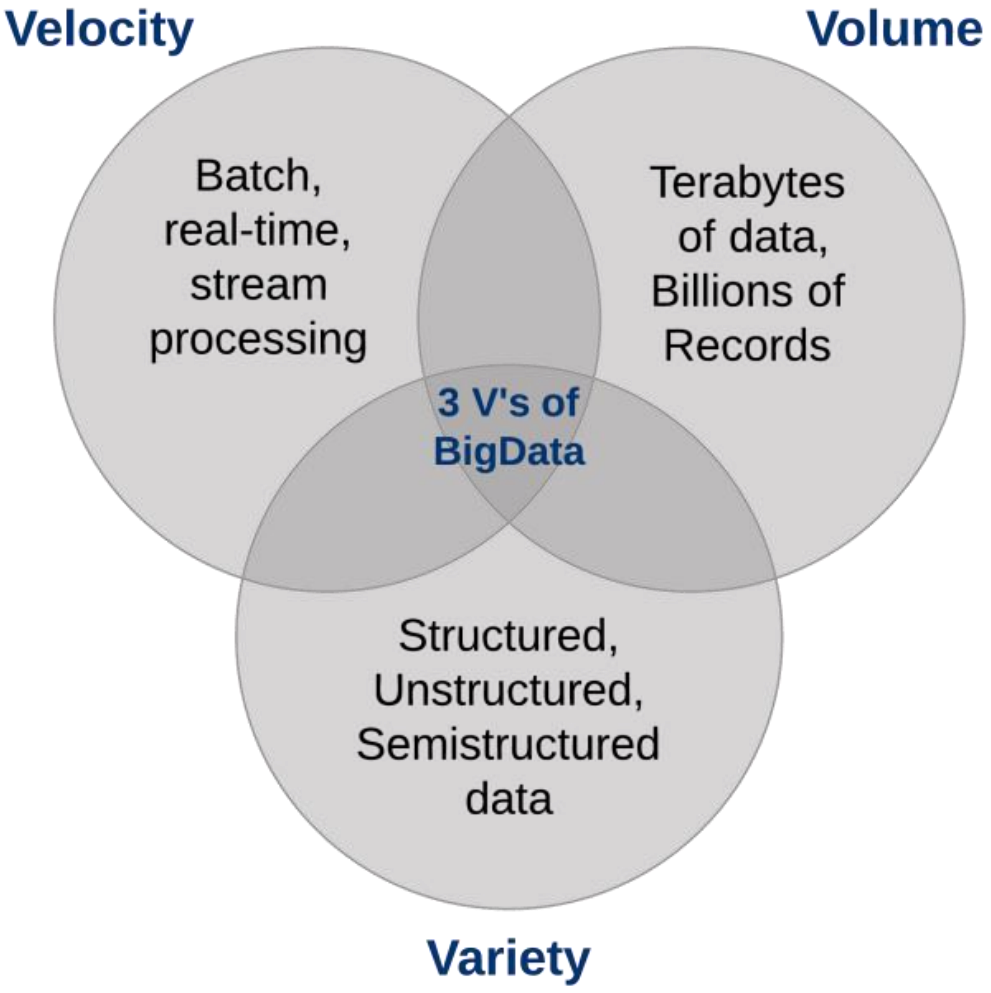
What is Big Data?

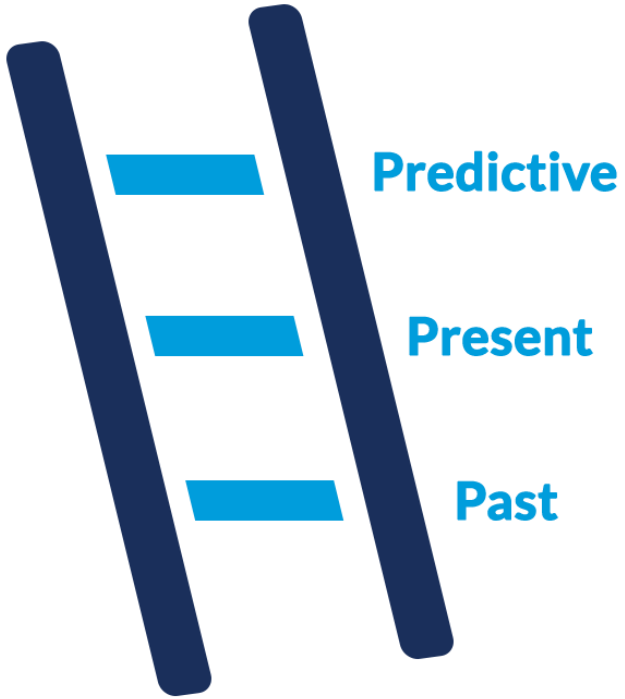
Big Data is larger, more complex Datasets, especially from new data sources

Traditional Data Processing software just can't manage them

This is also known as the three Vs

What is Big Data?

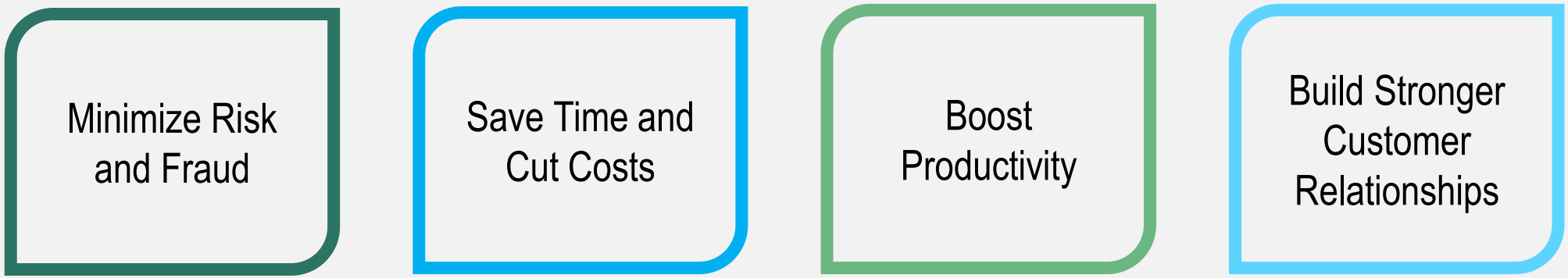




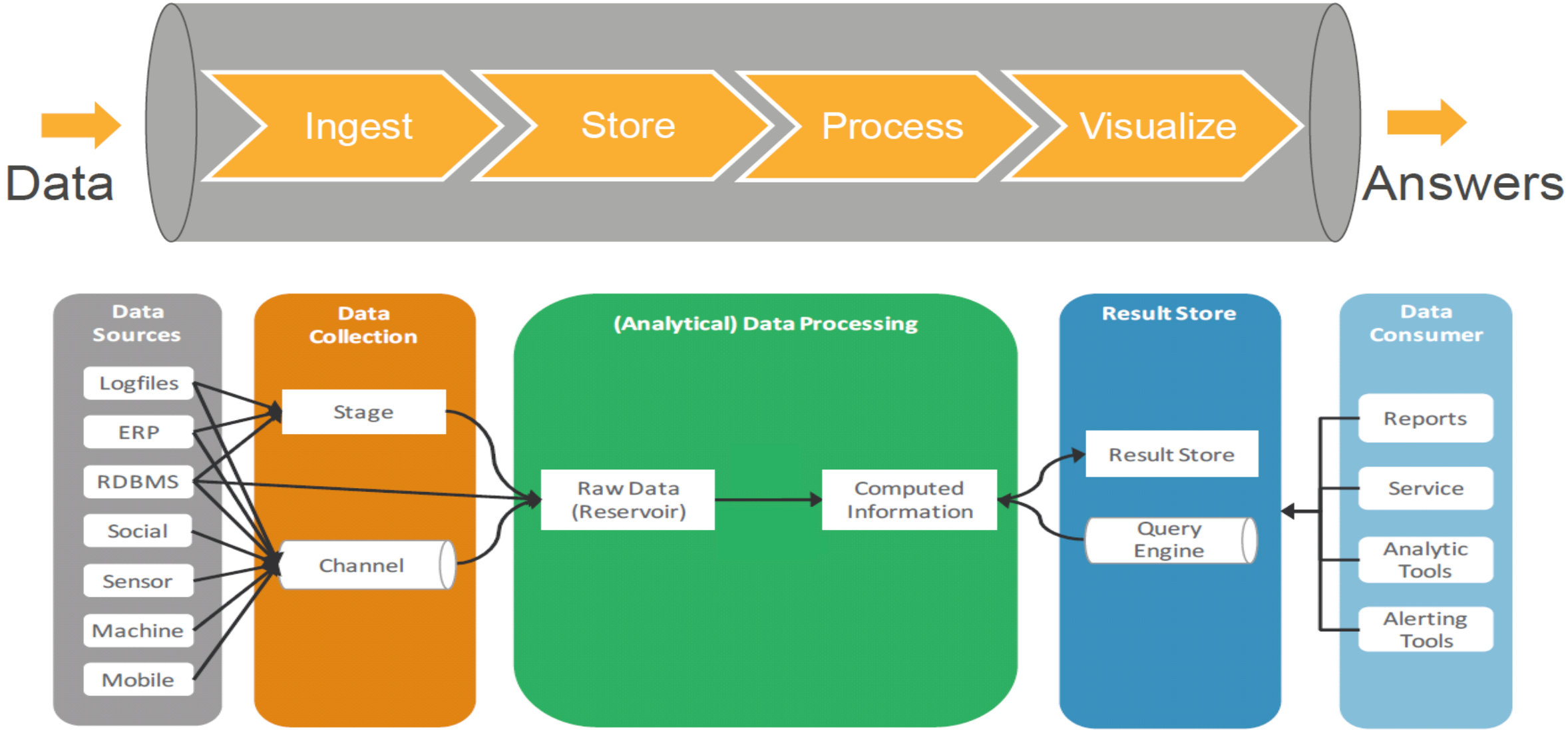
Why Is Big Data Important?

You don’t have to be big to use big data

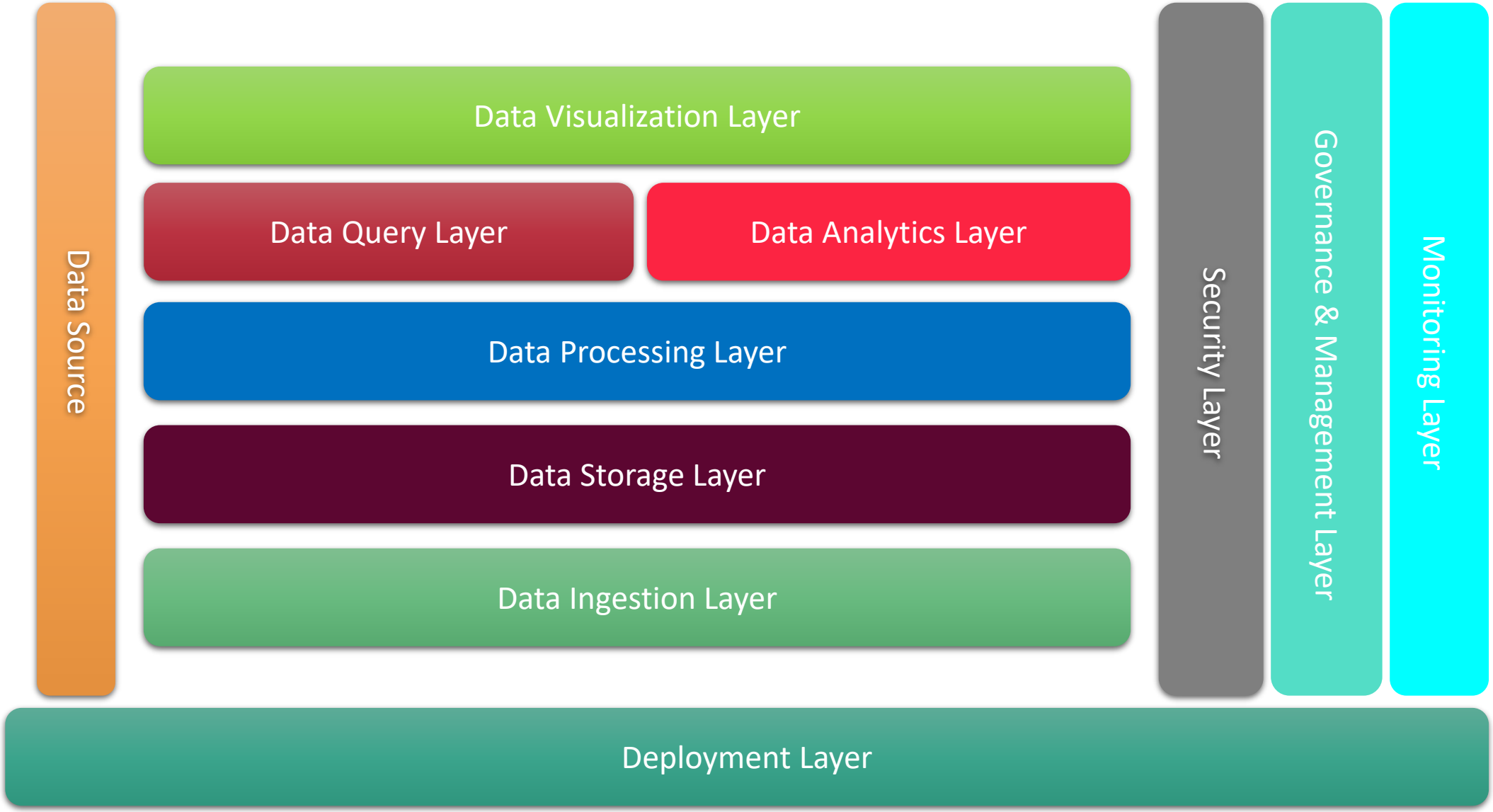
- Businesses of all sizes use big data with analytics and the cloud to be more competitive, achieve digital transformation or dominate in a market. In turn, they can:



How Big Data Works?



Big Data Layers



Big Data Layers: Data Ingestion

- Scalable, Extensible to capture streaming and batch data
- Provide capability to business logic, filters, validation, data quality, routing, etc. business requirements

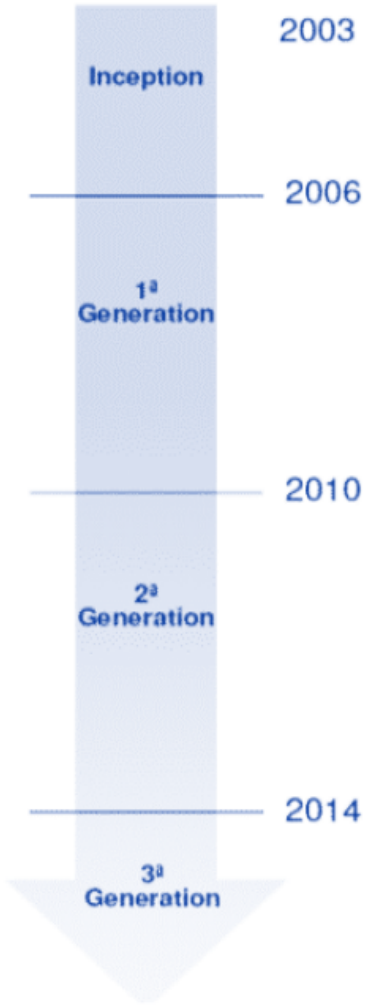
- Technology Stack:
  - Apache Kafka
  - Apache Sqoop
  - Apache NiFi
  - Logstash
  - Fluentd
  - Facebook Scribe
  - Amazon Kinesis
  - …

Big Data Layers: Data Storage

- Depending on the requirements data can placed into Distributed File System, Object Storage, Nosql Databases, etc.

- Technology Stack:
  - HDFS, Hive
  - Redis, MongoDB, Hbase, Cassandra, ElasticSearch,…
  - RDBMS

Big Data Layers: Data Processing



Batch processing

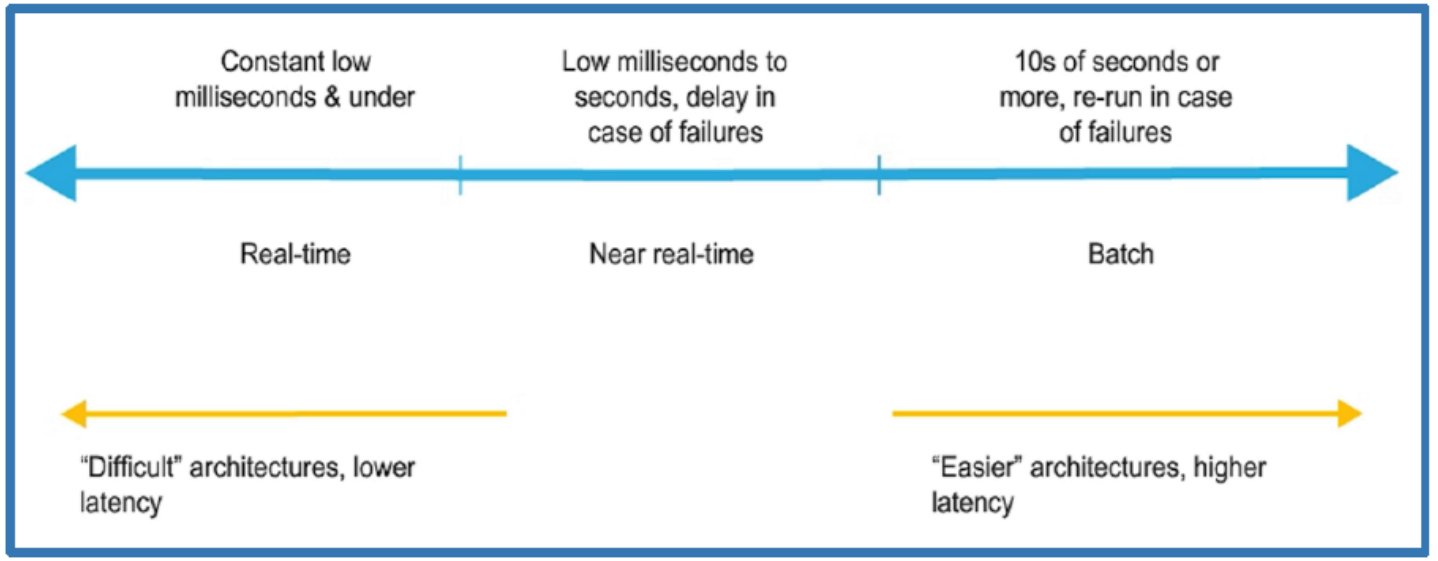
- Large amount of statics data
- Generally incurs a high-latency / Volume

Real-time processing

- Compute streaming data
- Low latency
- Velocity

Hybrid computation

- Lambda Architecture
- Volume + Velocity



Big Data Layers: Data Processing

- Processing is provided for batch, streaming and near real time cases
- Scale-Out Instead of Scale-Up
- Fault-Tolerant based methods

- Technology Stack:
  - MapReduce
  - Spark

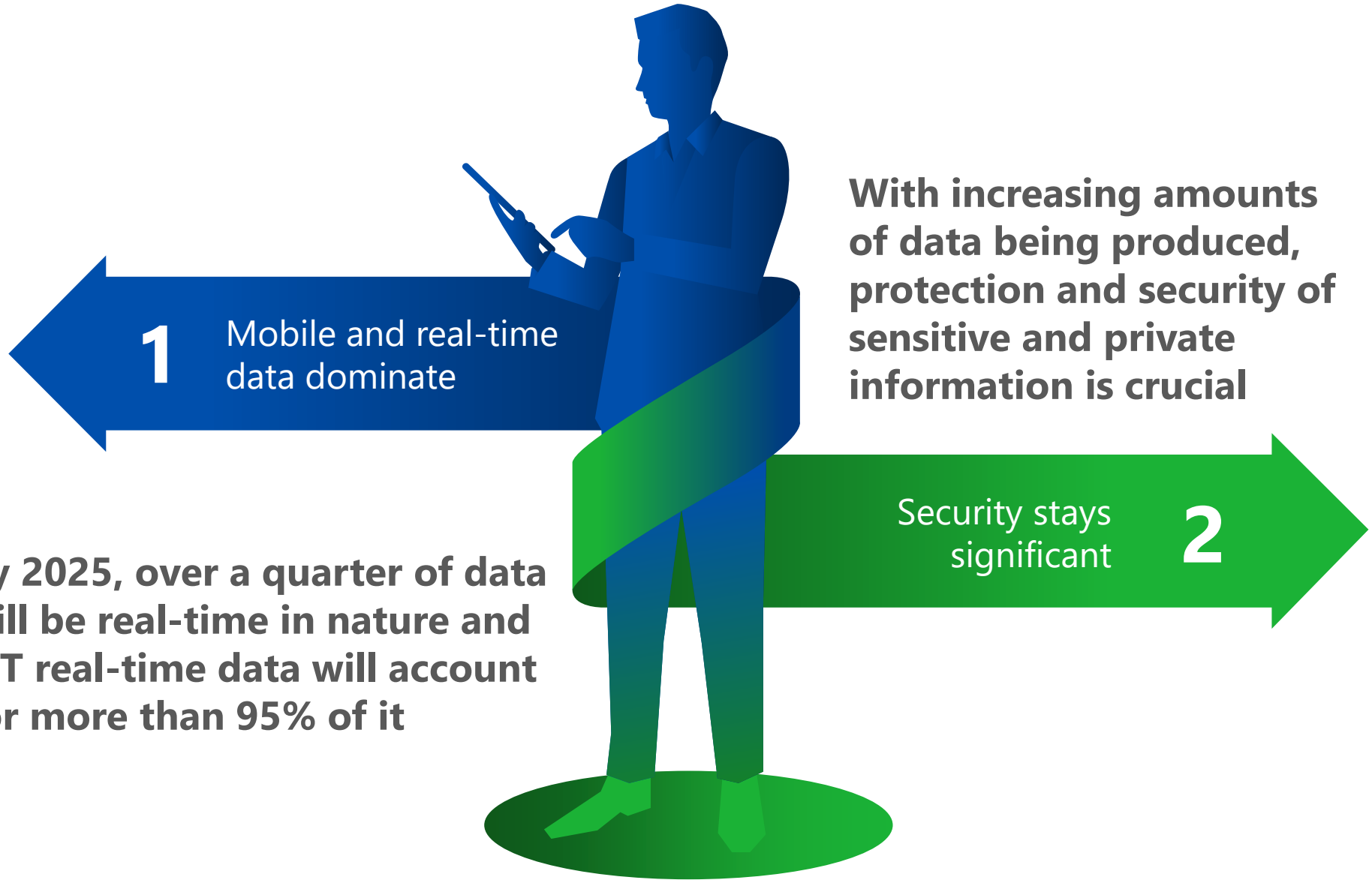
Big Data Layers: Data Visualization

- Dashboard and applications that provides valuable business insights

- Technology Stack:
  - Qlik
  - Tableau
  - Google Data Studio

Trends in Big Data

Source: IDC. Data Age 2025. The Digitization of the World From Edge to Core. David Reinsel, John Gantz and John Rydning. Sponsored by Seagate(SAS). US44413318. November 2018.

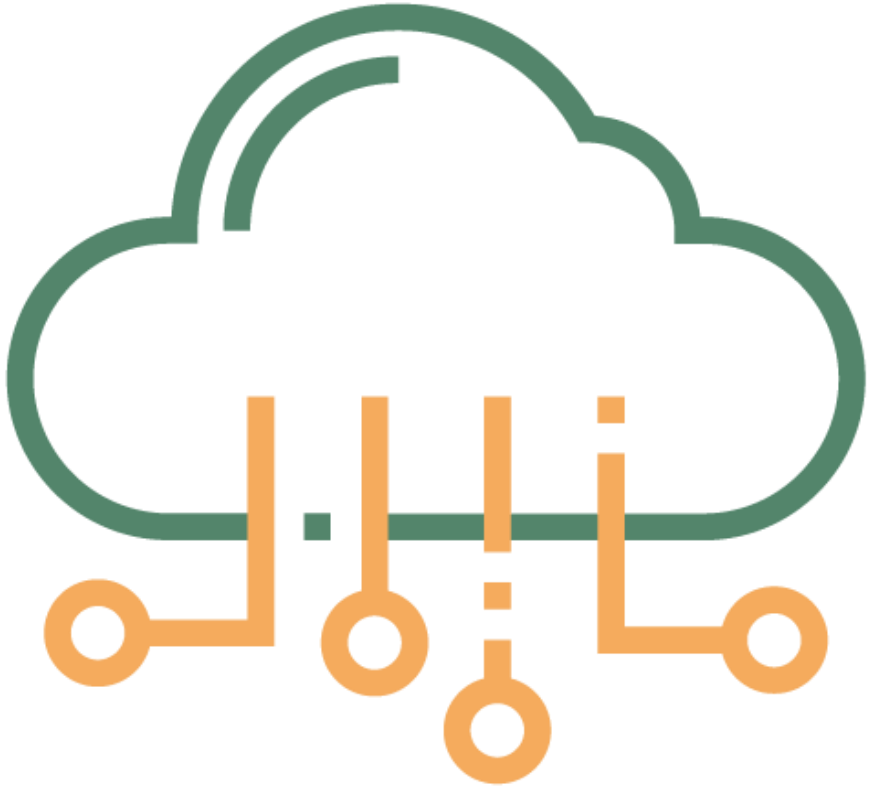


With increasing amounts of data being produced, protection and security of sensitive and private information is crucial

By 2025, over a quarter of data will be real-time in nature and IoT real-time data will account for more than 95% of it

Big Data and Cloud

Cloud computing is a subscription-based delivery model that provides scalability, fast delivery and IT efficiencies

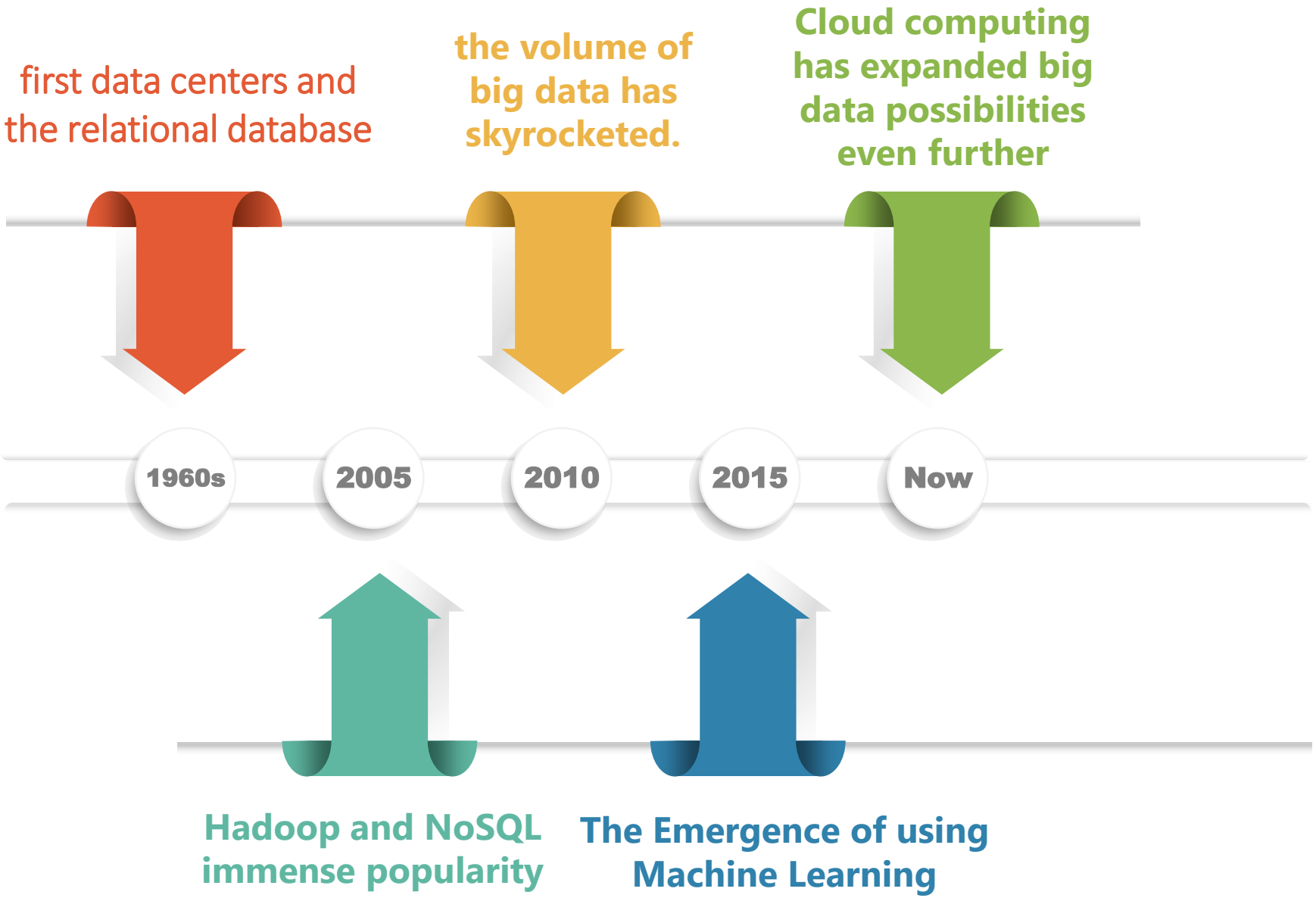


- Important benefits for today's increasingly mobile workforce. With the cloud, you have the flexibility to use data storage, network and compute resources when you need them. And you can easily scale up and down as business needs change.

Public Cloud Private Cloud

A major force for business innovation across all industries

History of Big Data



Big Data Cases

Securities Exchange Commission (SEC)

- ❑ anti-money laundering
- ❑ demand enterprise risk management
- ❑ fraud mitigation

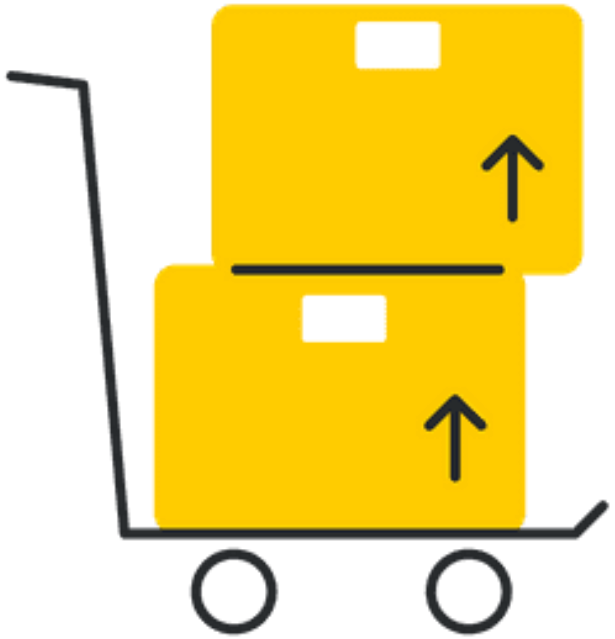
Big Data Cases

Source: edureka!

Big Data Cases

Retail Industry

- Optimized staffing through data from shopping patterns, local events, etc.
- Reducing fraud
- Timely analysis of inventory

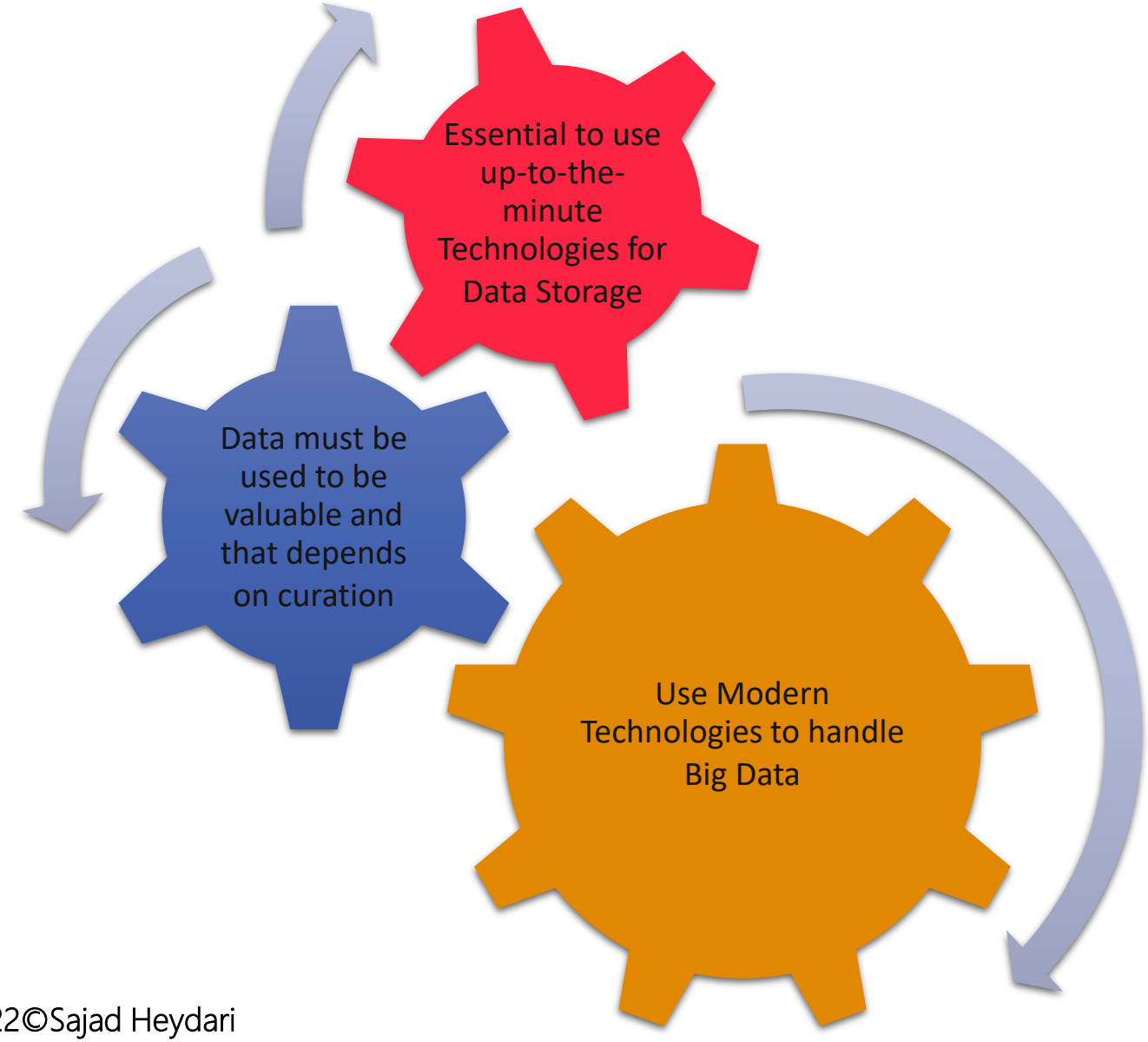


Big Data Cases

Big Data and Formula1

- ❑ Analyzing the relative speed of the car
- ❑ Machine/Engine failure or damage
- ❑ Predicting the outcome of a race
- ❑ Strengths and weaknesses of a car
- ❑ Communicating between the departments

Big data challenges



Related Job Positions

	Data Scientist	Data Engineer	ML Engineer
			
Key Role	Creating Data-Driven Solutions	Develops, Construct, test, maintain	Write and Develop Algorithms
Data Preparation Skill	✓	✓	✓
Data Visualization	✓	✓	✓
Programming Tools	✓	✓	✓
Statistics Linear Algebra	✓	✓	✓
ML, BD	✓	✓	✓

✓ High level

✓ Low level

Study References

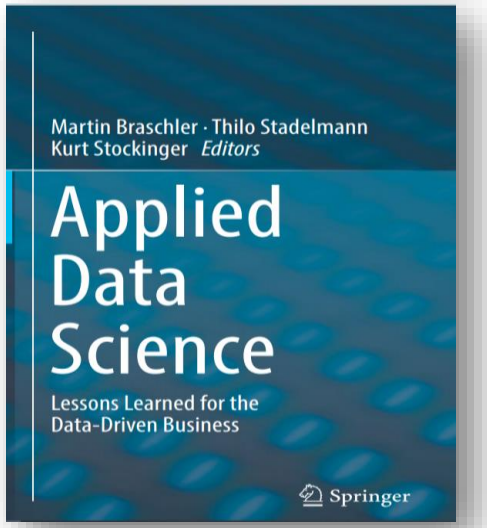
Course:

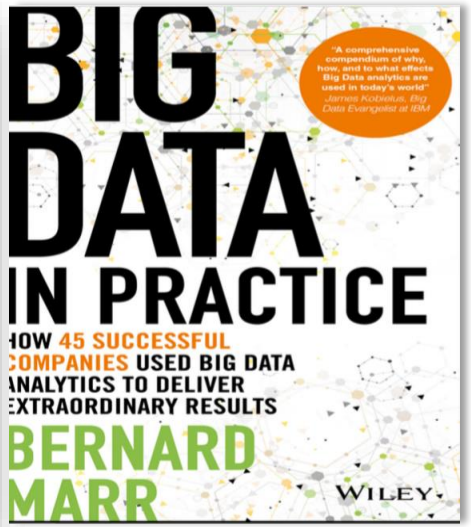
coursera

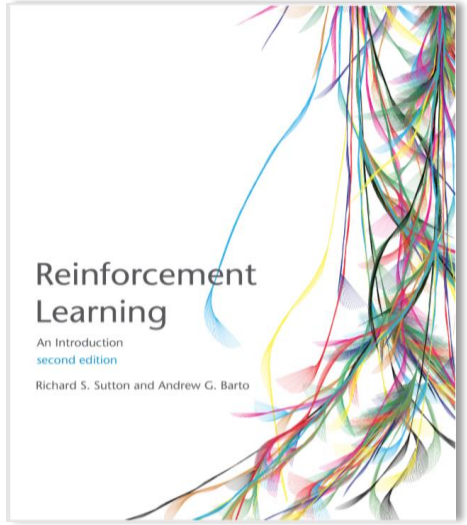
YouTube

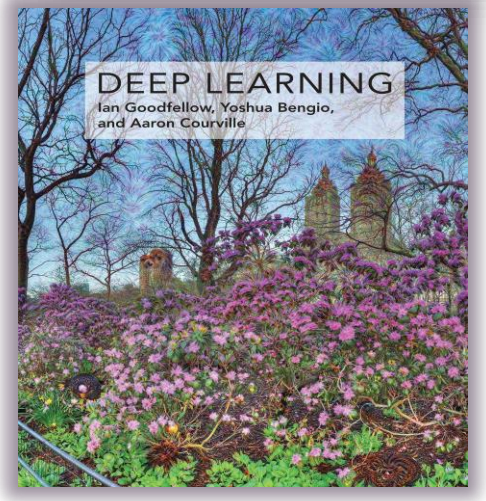
Udemy

Books:









Study References

Competitions:

kaggle

DataCamp

Articles:

KDnuggets

Analytics Vidhya

Conclusion and Questions?

THANK YOU!

@ScienceWave

Sciencewave Academy

ScienceWave

SCIENCEWAVE