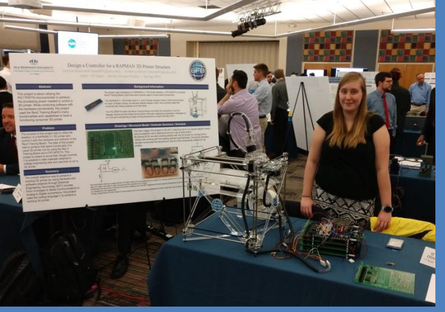
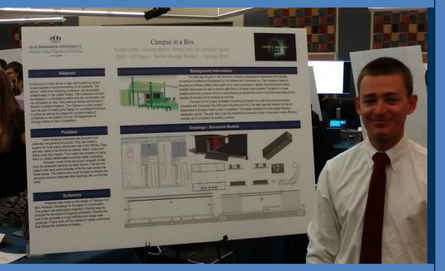




OLD DOMINION UNIVERSITY

ENGT 375:
Applied Machine Learning for Engineering Technology

Lecture 1A: Introduction to Data Science and Machine Learning

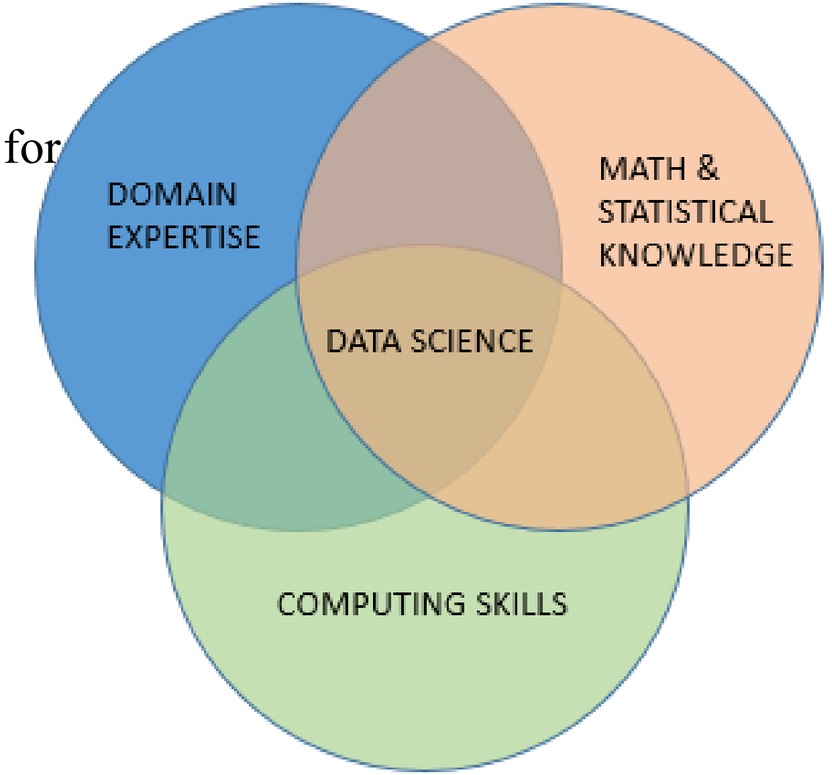
- Dr. Murat Kuzlu
- Department of Engineering Technology

Content

- What is Data Science
- Relationship Data Science and Machine Learning
- Who is a Data Scientist
- What is Big Data
- What is Machine Learning
- Machine Learning Methods
- Common Python Packages for Data Science

What is Data Science

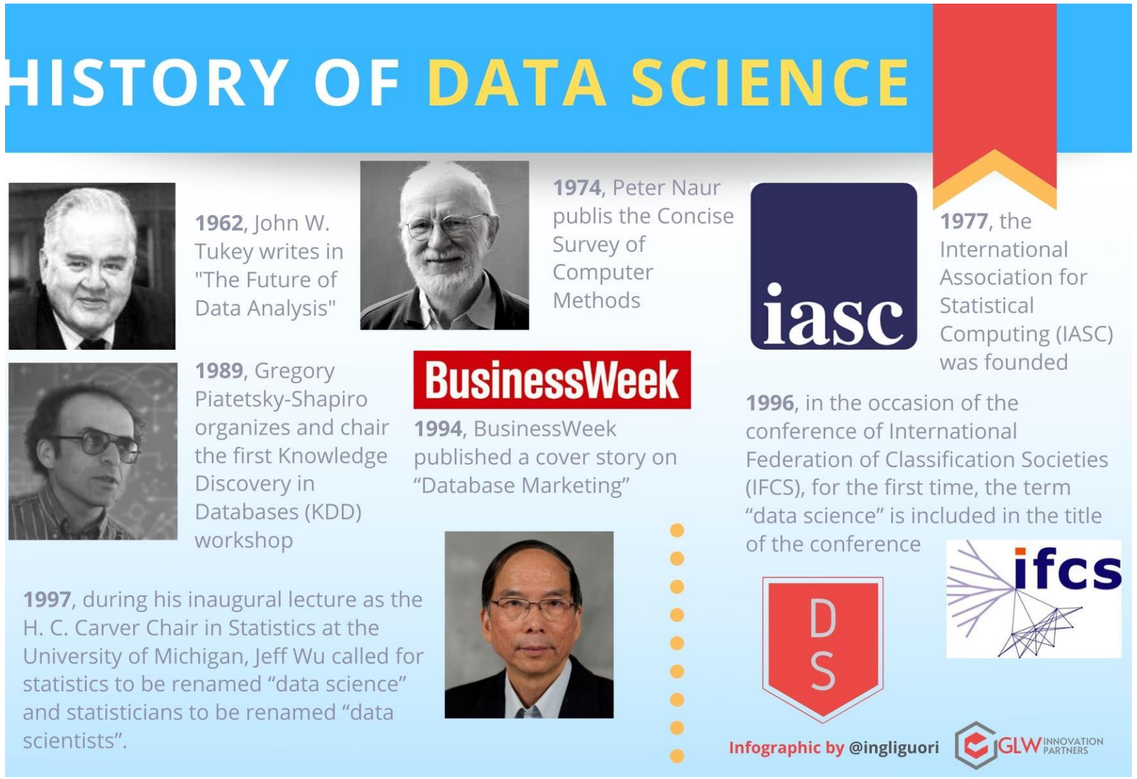
- The **study of data** to extract meaningful insights for business.
- **A multidisciplinary approach** (mathematics, statistics, artificial intelligence, and computer engineering) to analyze **large amounts of data**.
- Helps data scientists to ask and answer questions like **what** happened, **why** it happened, what will happen, and what can be done with the results.



https://medium.com/@anandr42/the-data-science-delusion-7759f4eaac8e

What is Data Science a New Field?

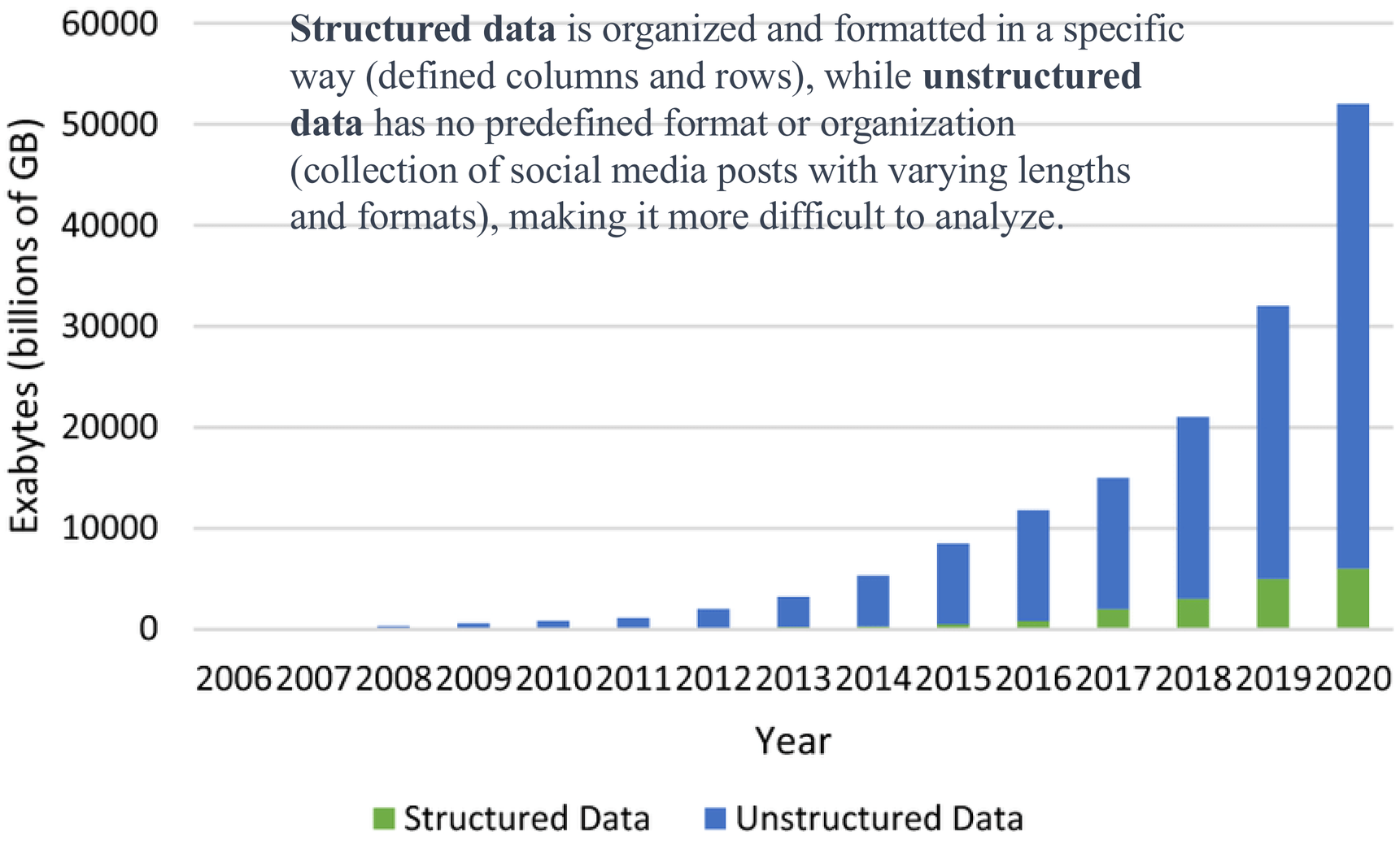
- Data originates from the Latin word, “**datum**,” which means a “**something given**” . The expression “**data**” has been utilized since 1500s.
- The modern practice began during the **1940s and 1950s**. In fact, the moderate development of Data Science is slow.
- The area of data science also expended to biological sciences, medical informatics, health care, social sciences and humanities.
- Data science reduces the workload on data scientists, as they have no need to waste their precious time in complex algorithms.



https://www.glweb.eu/blog/digital-transformation/150-brief-history-of-data-science

The Growth of Structured vs Unstructured Data

Structured data is organized and formatted in a specific way (defined columns and rows), while **unstructured data** has no predefined format or organization (collection of social media posts with varying lengths and formats), making it more difficult to analyze.



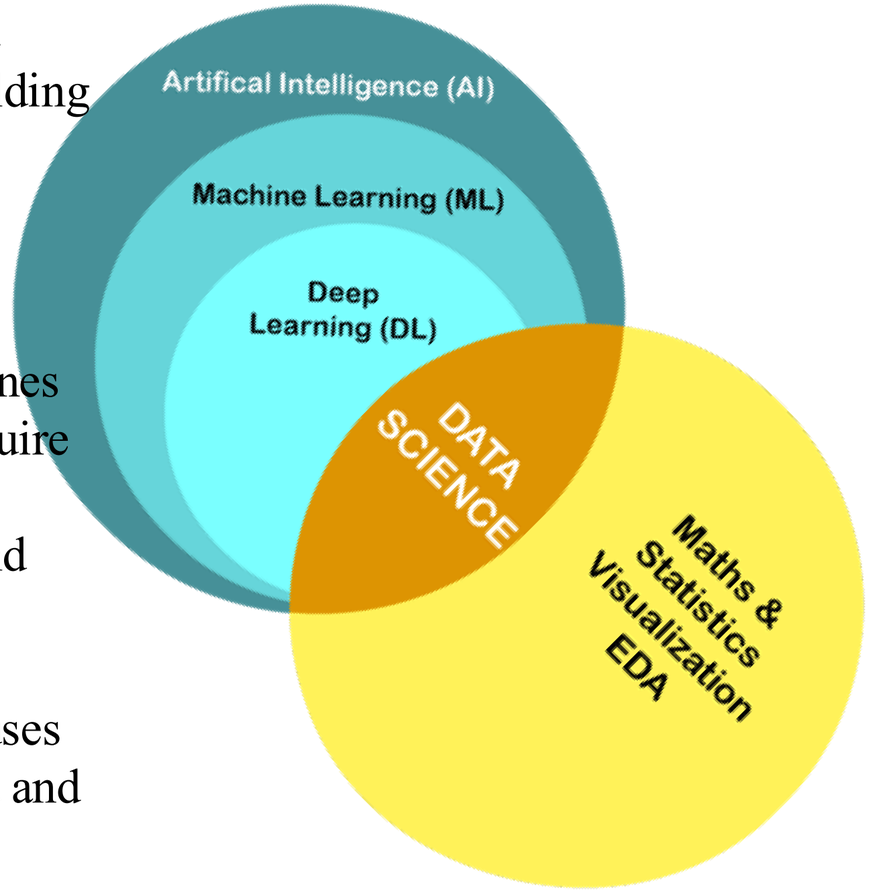
In an Internet Minute

- Instagram users went from posting 55,140 photos per minute two years ago to 65,000 per minute this year.
- Google went from 4.5 million searches per minute in 2019 to 5.7 million per minute currently.
- Netflix users went from streaming 694,444 hrs of video per minute in 2019 to 452,000 hours per minute in 2021

https://www.pcmag.com/news/in-an-internet-minute-way-too-much-is-happening-all-the-time

Relationship Data Science and Machine Learning

- **Data science** is a **broader field** that includes various techniques and methods for extracting insights and knowledge from data, while **machine learning** is a specific **subset of data science** that focuses on building algorithms



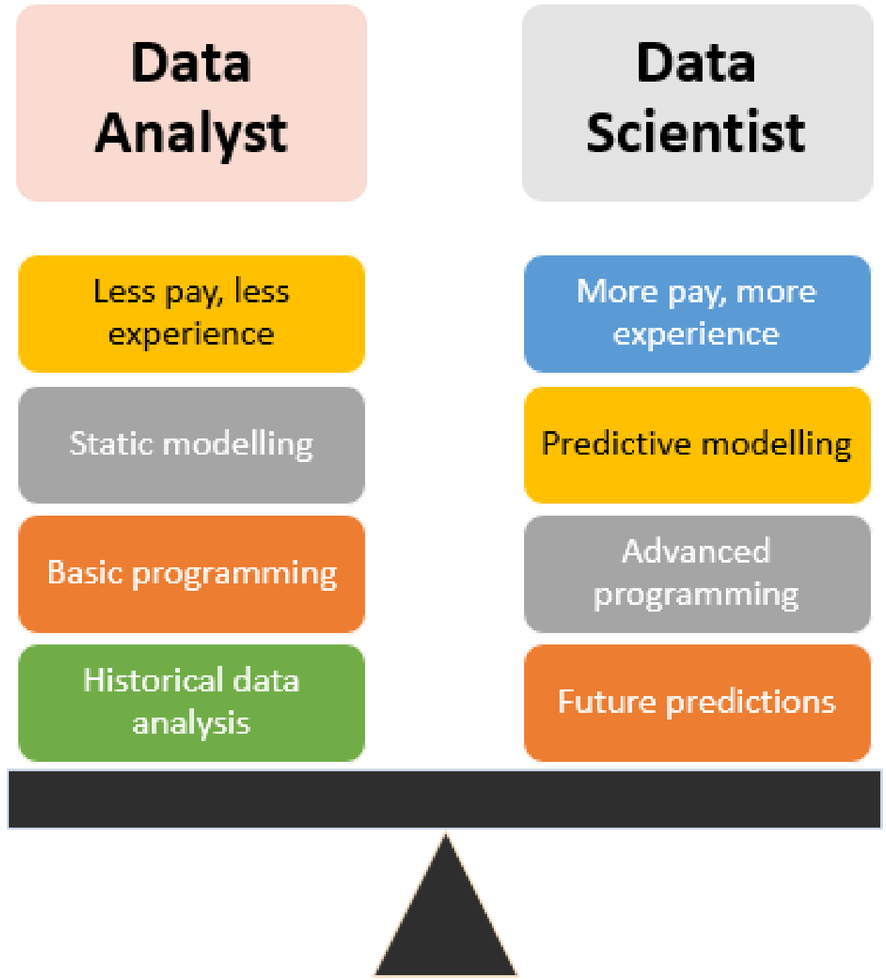
- **AI (Artificial Intelligence)** aims to develop machines or systems that can perform tasks that typically require human intelligence, such as natural language processing, image recognition, decision making, and problem-solving.

- **Machine learning (ML)** is a subset of AI that focuses on **developing algorithms** that can learn from data and make predictions or decisions based on that data.

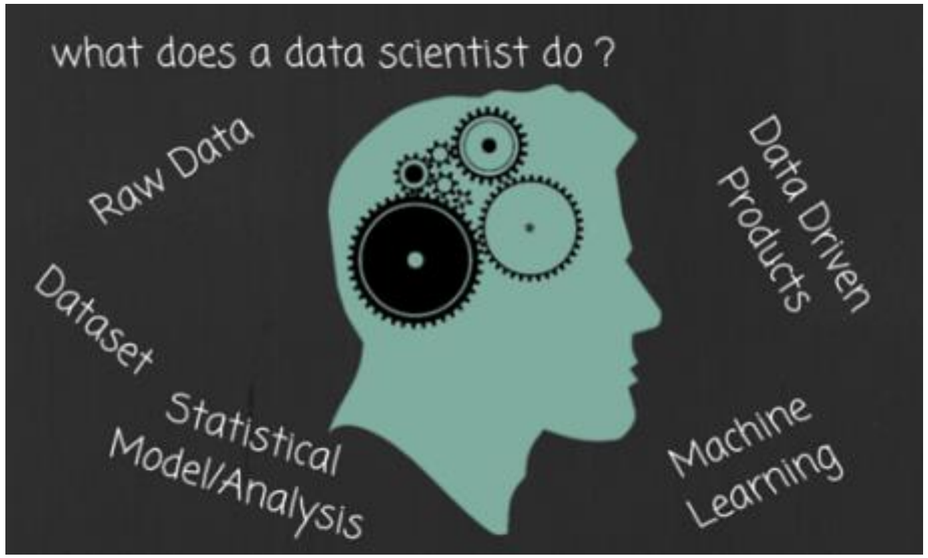
- **Deep learning (DL) is** a subset of ML that uses artificial neural networks to learn from data, which allows it to handle complex tasks such as image and speech recognition.

Data Analyst vs Data Scientist

- **A Data Scientist is** a multidisciplinary profile whose primary mission will be to extract useful information (insights) from raw data.
- Data Scientist will **explore and exploit** the company's data pools to apply machine learning techniques to them.
- **A Data Analyst** is someone who is responsible for collecting and analyzing data to identify patterns and trends
- A Data Analyst has a strong understanding of the **business domain**.

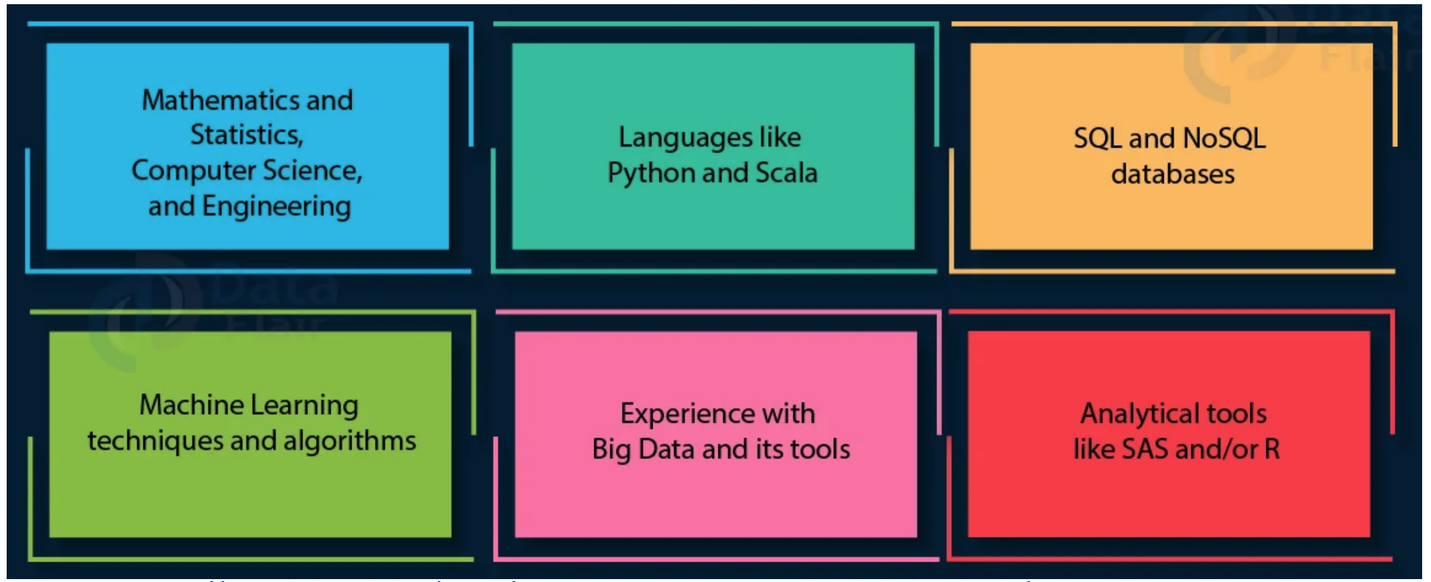


What Does a Data Scientist Do? and How?



https://medium.com/dataseries/what-does-a-data-scientist-do-a6553dc720f

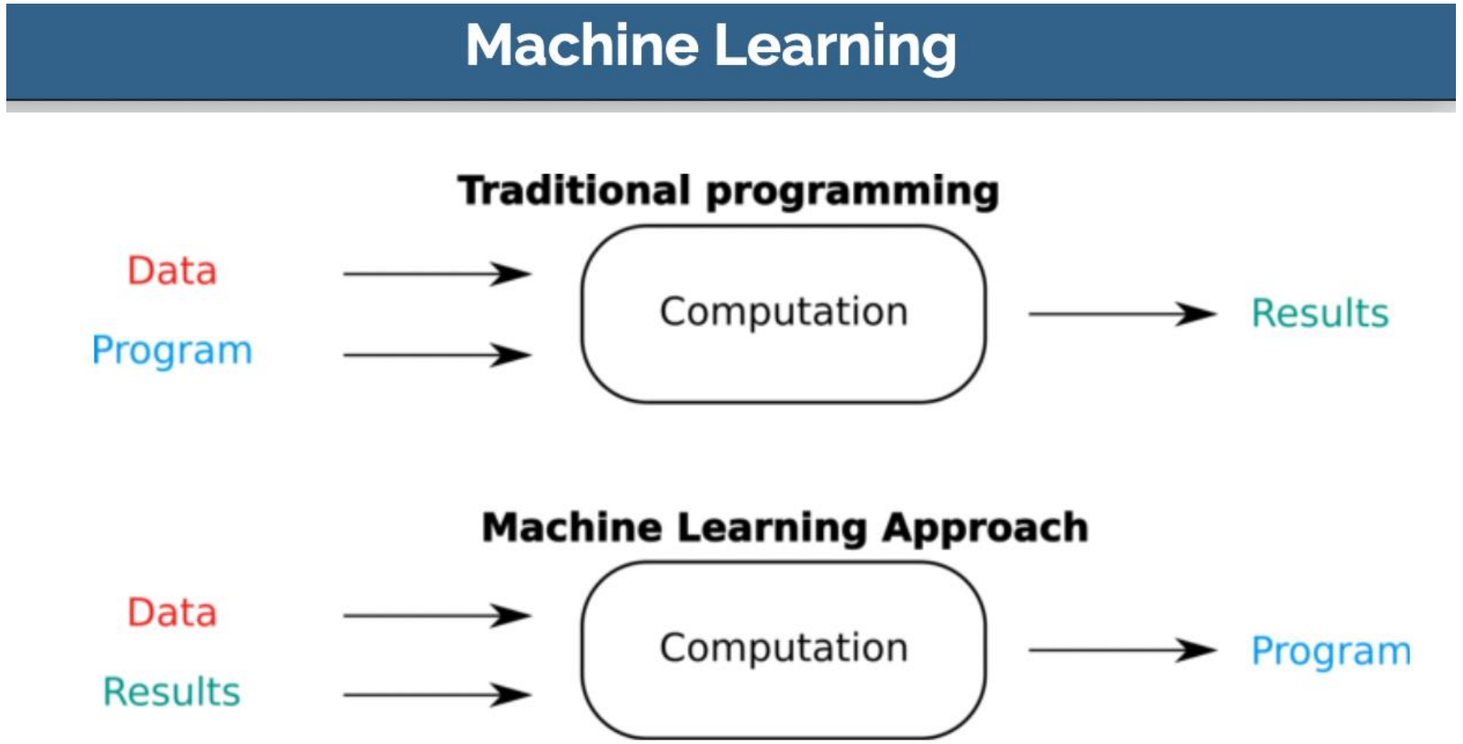
- Ask a question
- Collect data?
- Explore the data?
- Model the data
- Evaluate the results



https://data-flair.training/blogs/how-to-become-data-scientist-infographic/

Machine Learning vs Traditional Programming?

- Has the capability of a machine to imitate intelligent human behavior.
- Helps you discover relationships, recognize patterns, predict trends and find associations from your data.

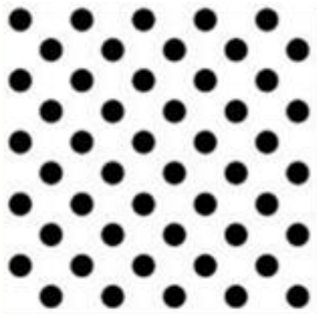
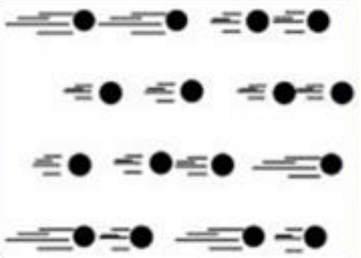
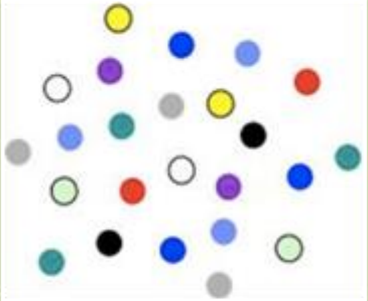
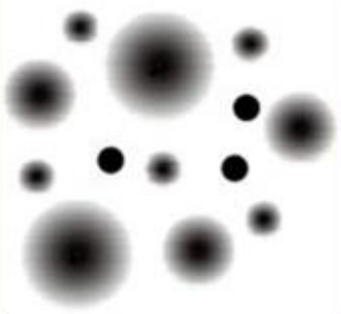


https://datalya.com/blog/machine-learning/machine-learning-vs-traditional-programming-paradigm

What is Big Data?

- Big Data is one of the many concepts that, in recent years, have gained momentum in the technological world.
- **Big data** refers to **data** that is so large, fast or complex that it's difficult or **impossible to process using traditional methods.**
- The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity.

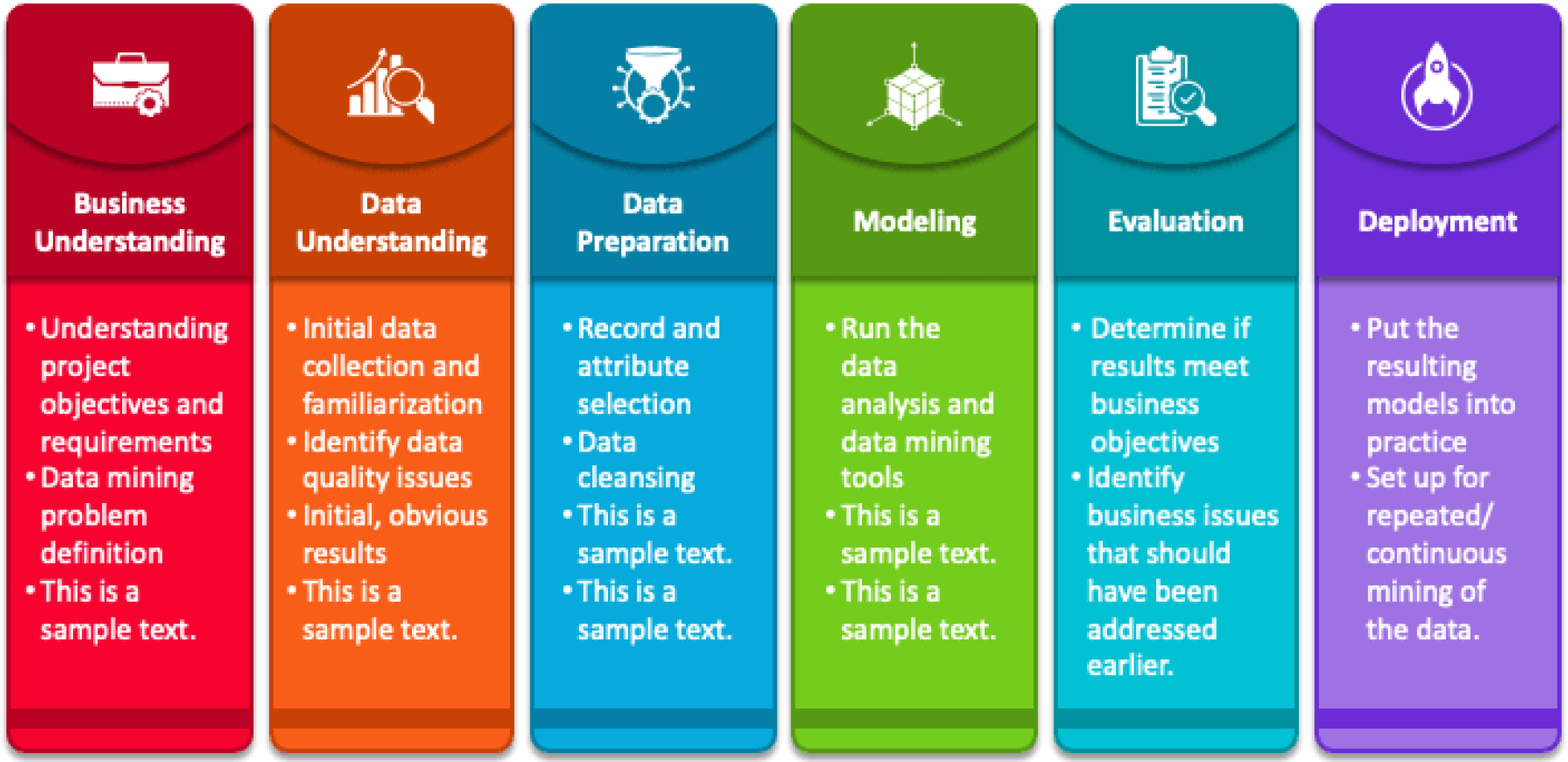
5V's of Big Data

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,…	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

What is CRoss Industry Standard Process for Data Mining (CRISP-DM) ?

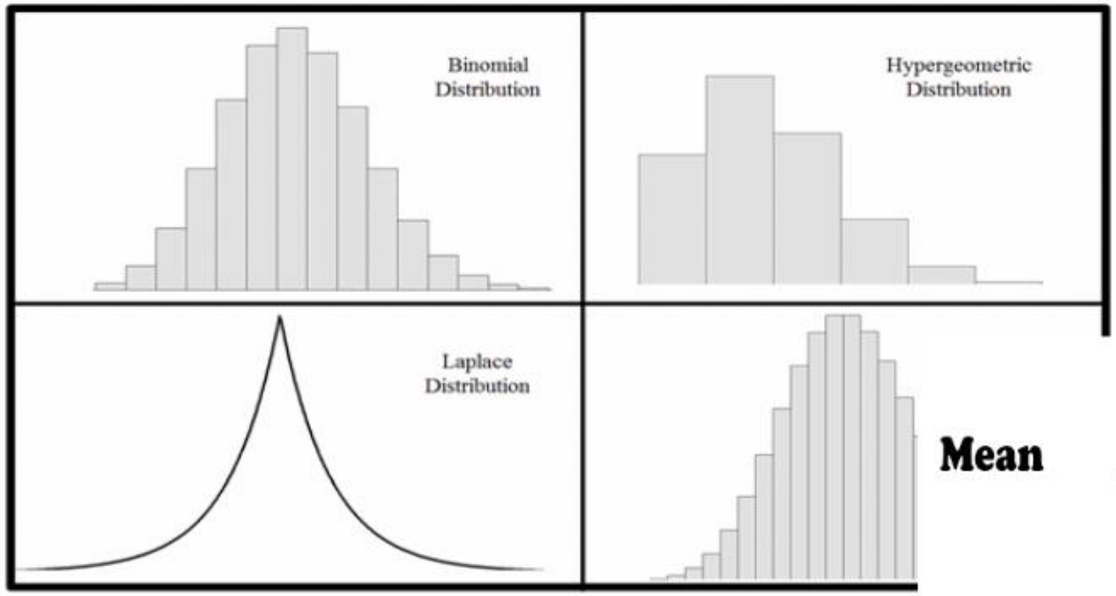
The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process.

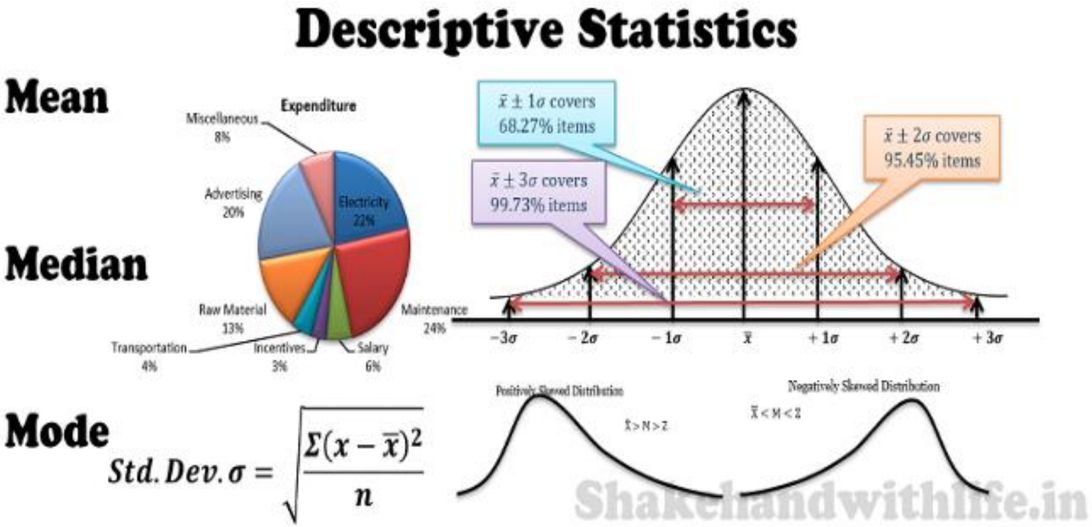


https://www.sketchbubble.com/en/presentation-crisp-dm.html

Statistics

Develop a statistics-based model that analyzes data distributions, relationships, and trends to support data-driven decision-making and model validation.

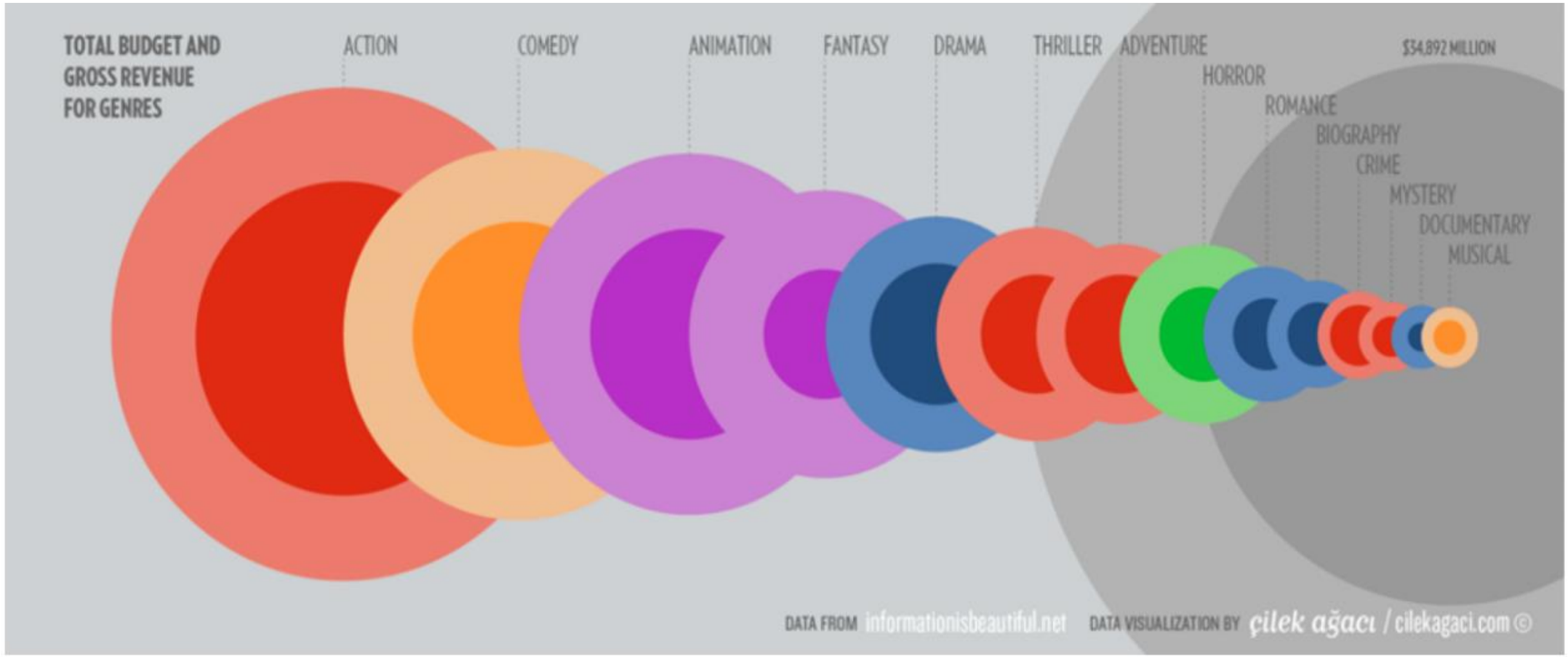




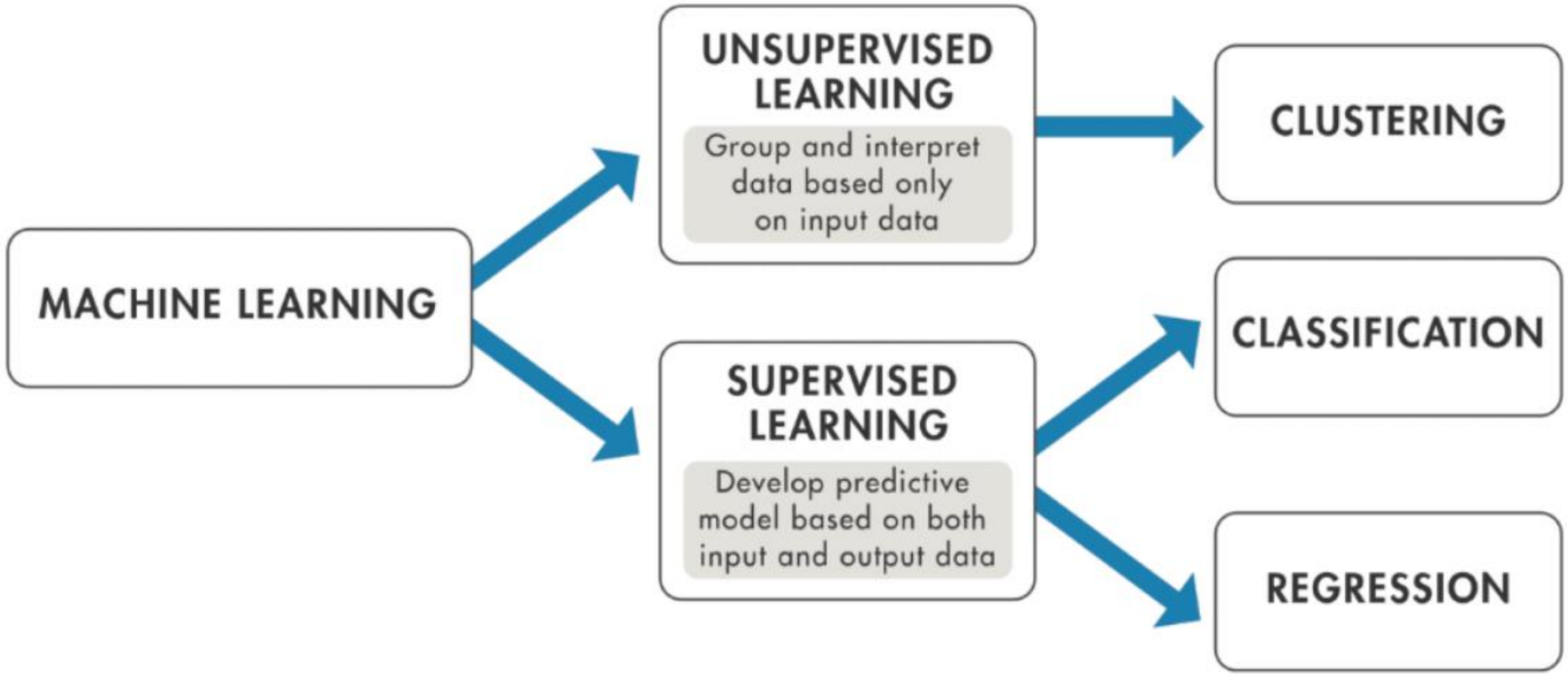
Data Visualization

Develop a data visualization project that presents complex datasets through clear, interactive, and insightful visual representations to reveal patterns and trends.

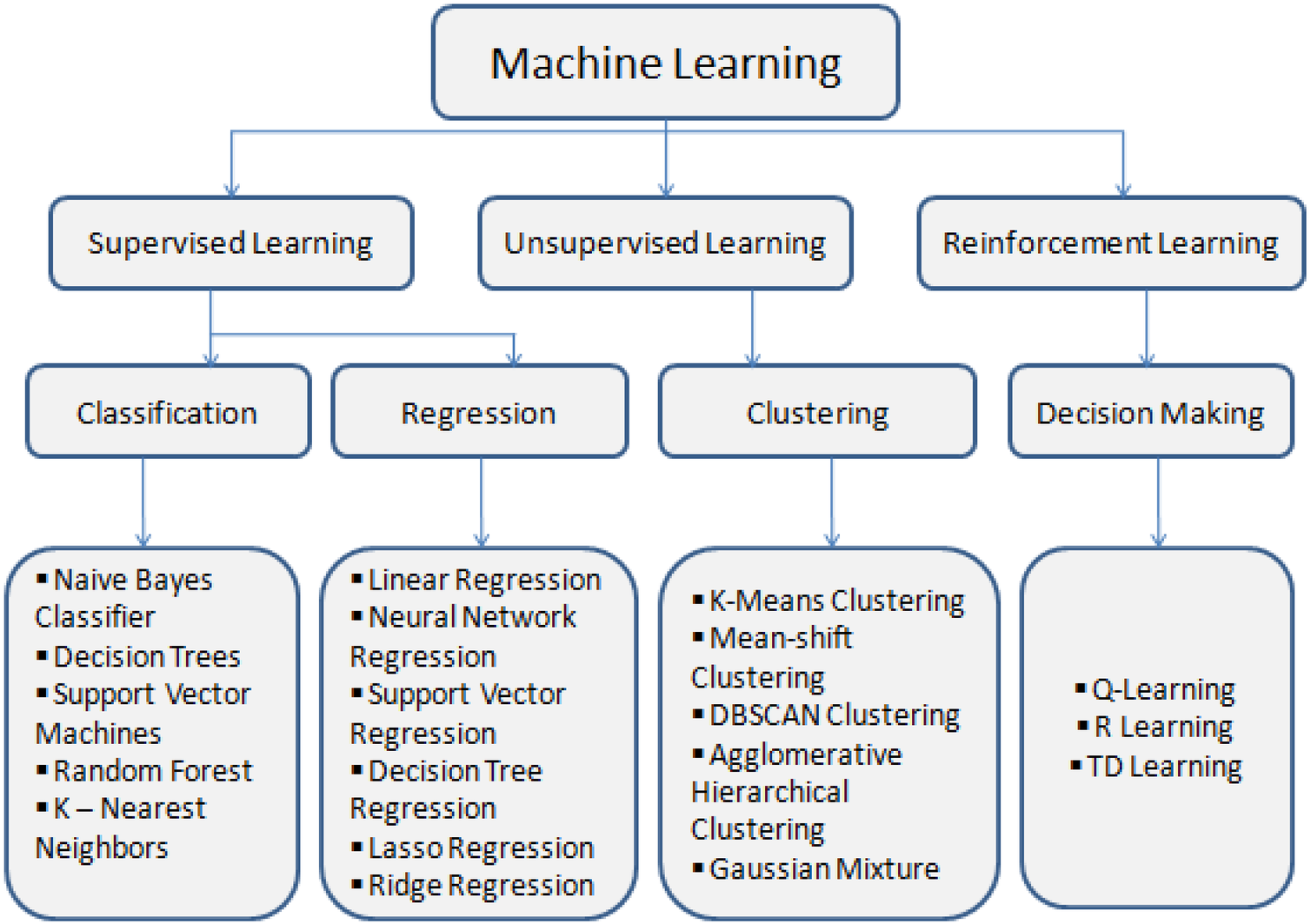
Hollywood Economics



Categorical Types of Machine-Learning Algorithms



Top Machine-Learning Algorithms



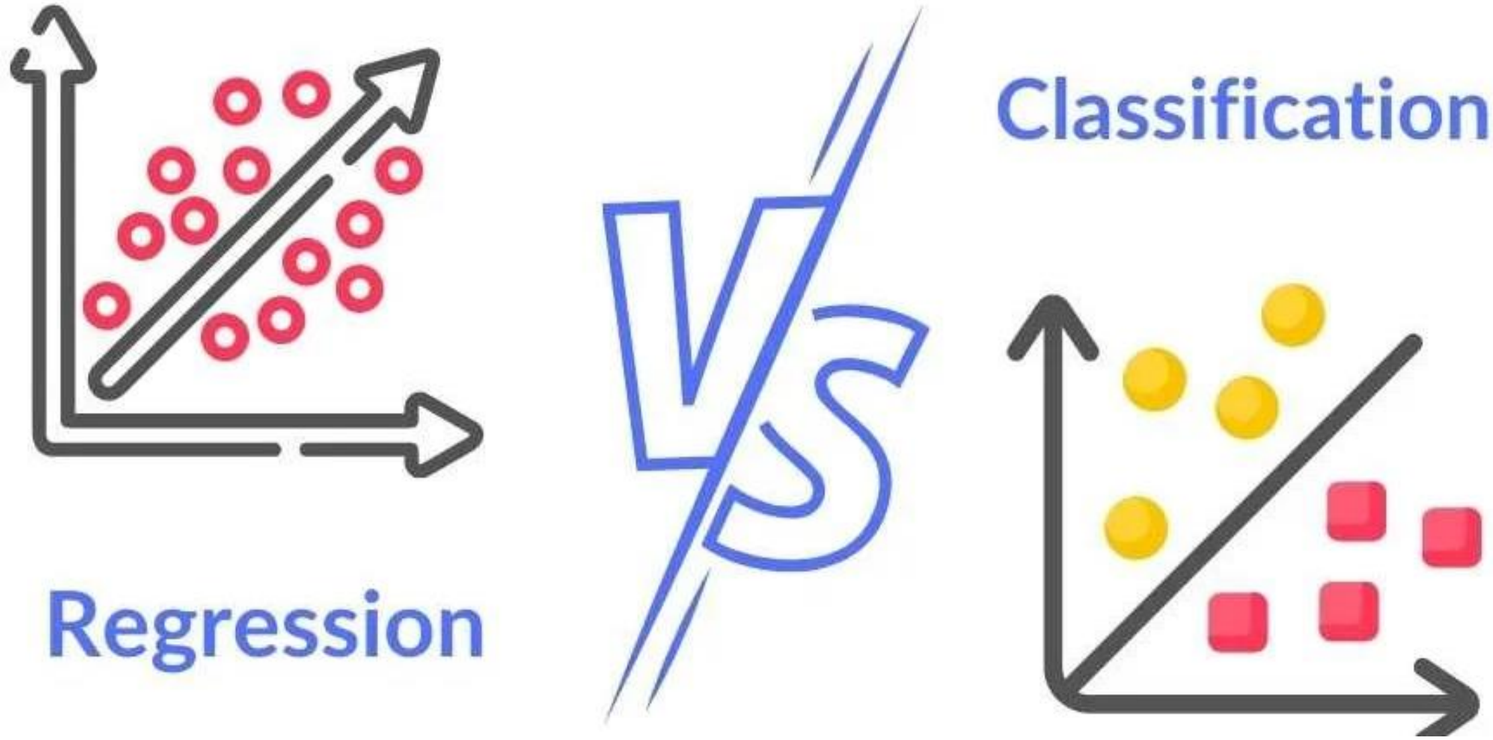
https://medium.com/@sarthakkumargupta/semi-supervised-and-supervised-machine-learning-750e2dade6d

Categorical Types of Machine-Learning Algorithms

https://www.researchgate.net/publication/330217507_Application_of_machine_learning_in_rheumatic_disease_research/figures?lo=1

Regression vs Classification

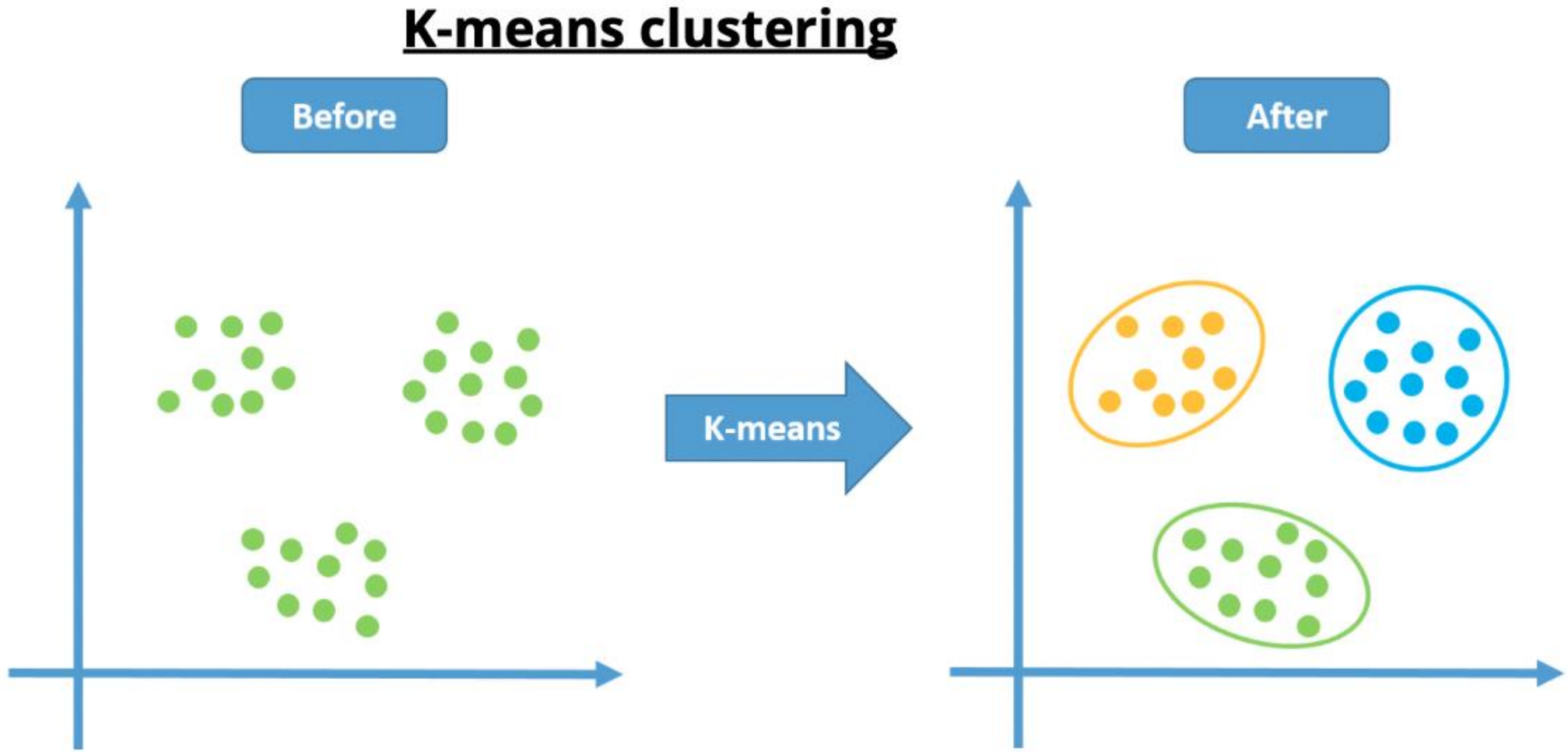
Develop a regression or classification model that uses supervised learning to predict continuous values (regression) or assign labels to categories (classification) based on input data.



https://spotintelligence.com/2023/05/02/regression-vs-classification/

Clustering

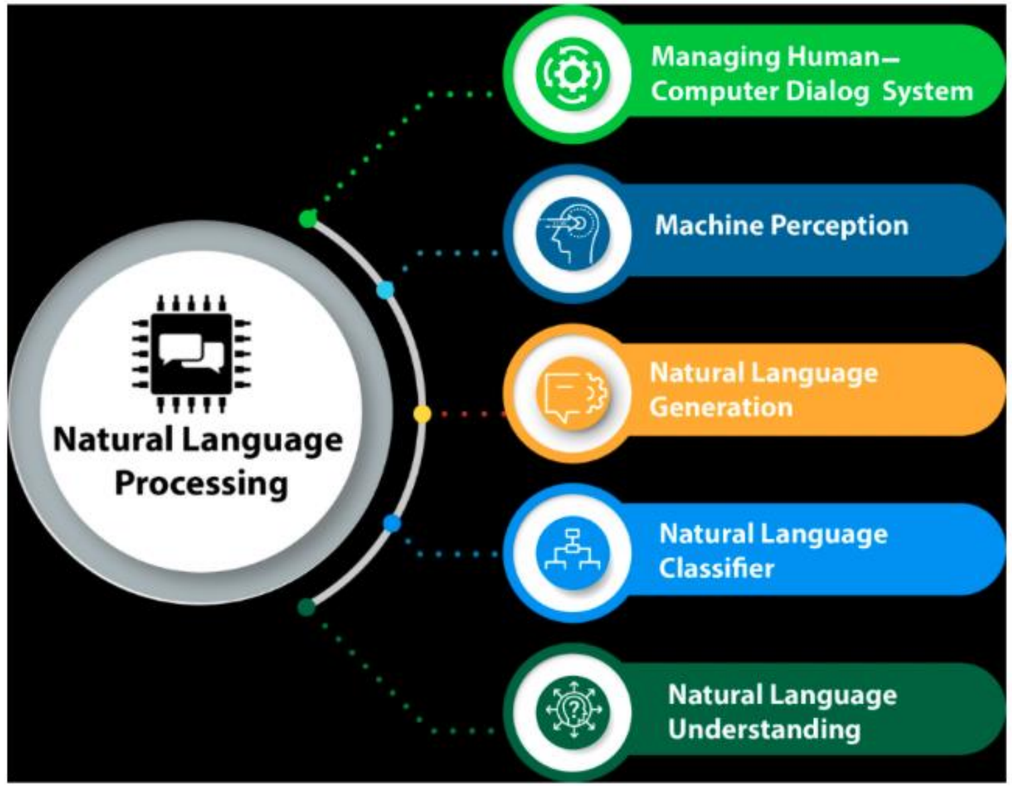
Develop a clustering model that groups unlabeled data into meaningful segments (e.g., using k-means, hierarchical, or DBSCAN) and evaluates cohesion/separation (e.g., silhouette score).



Natural Language Processing

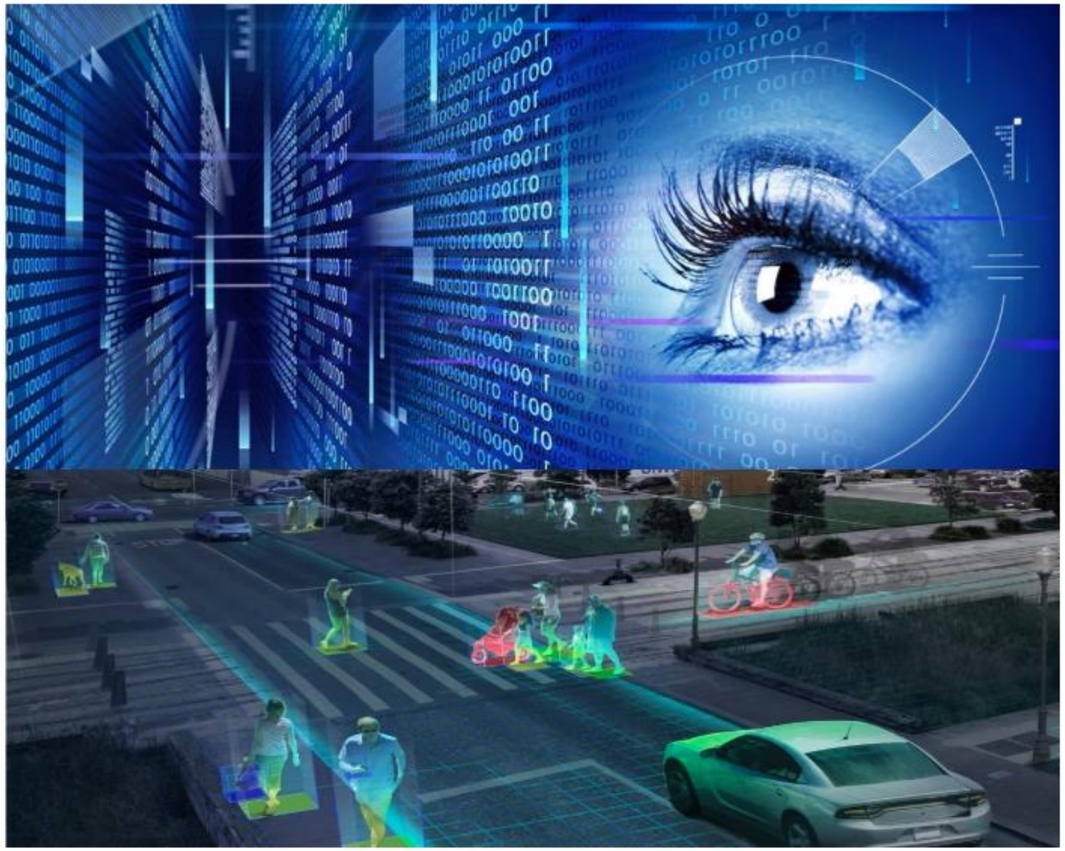
Text Mining/Text Analysis is the process of deriving meaningful information from natural language text.

NLP: Natural Language Processing is a part of computer science and artificial intelligence which deals with natural language. NLP is the ability of computers to recognize and respond to natural spoken words and written text.



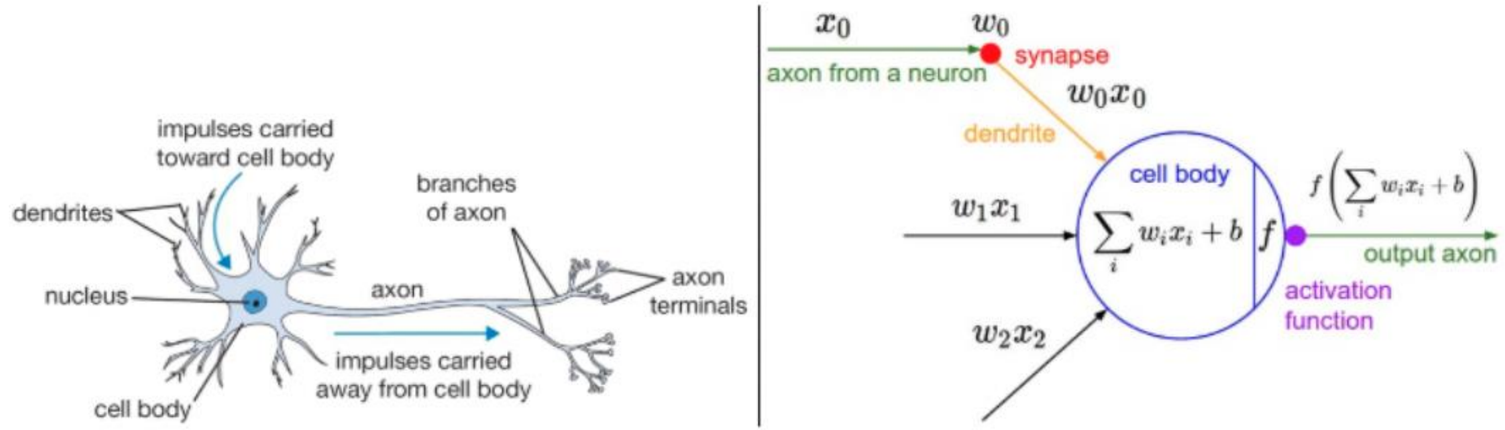
Computer Vision

Develop a computer vision–based system using machine learning techniques to analyze, detect, or classify visual data from images or videos.



Deep Learning

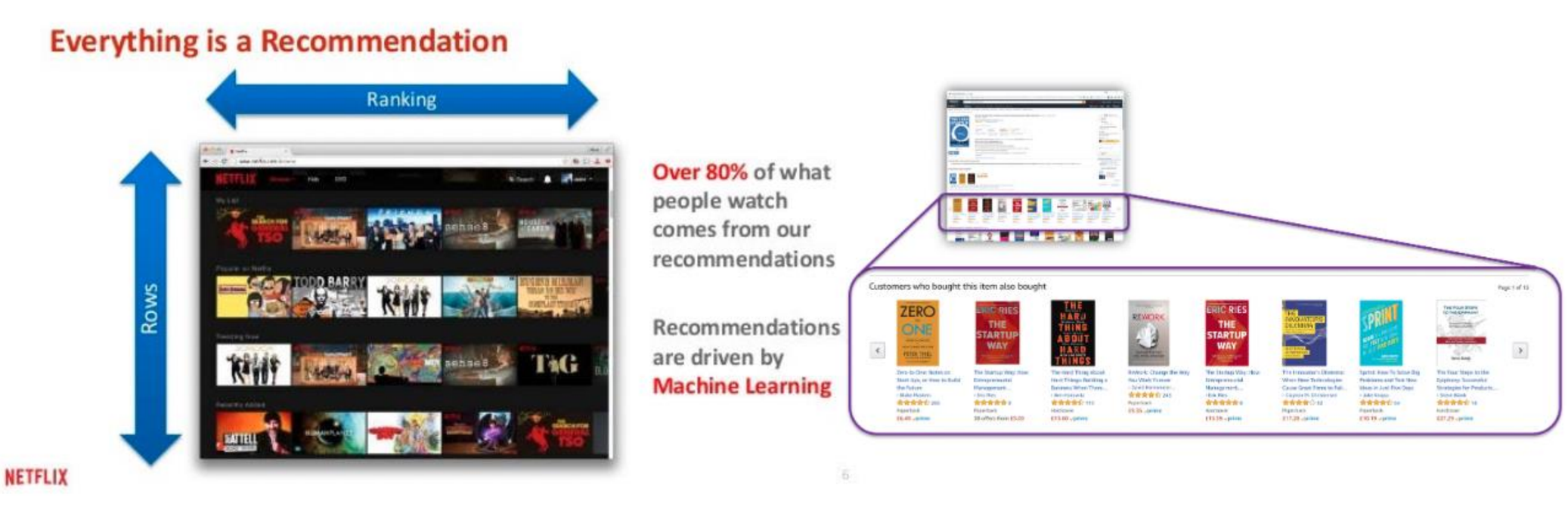
Develop a deep learning–based system that uses neural networks to model, analyze, or predict complex patterns from data.



A cartoon drawing of a biological neuron (left) and its mathematical model (right).

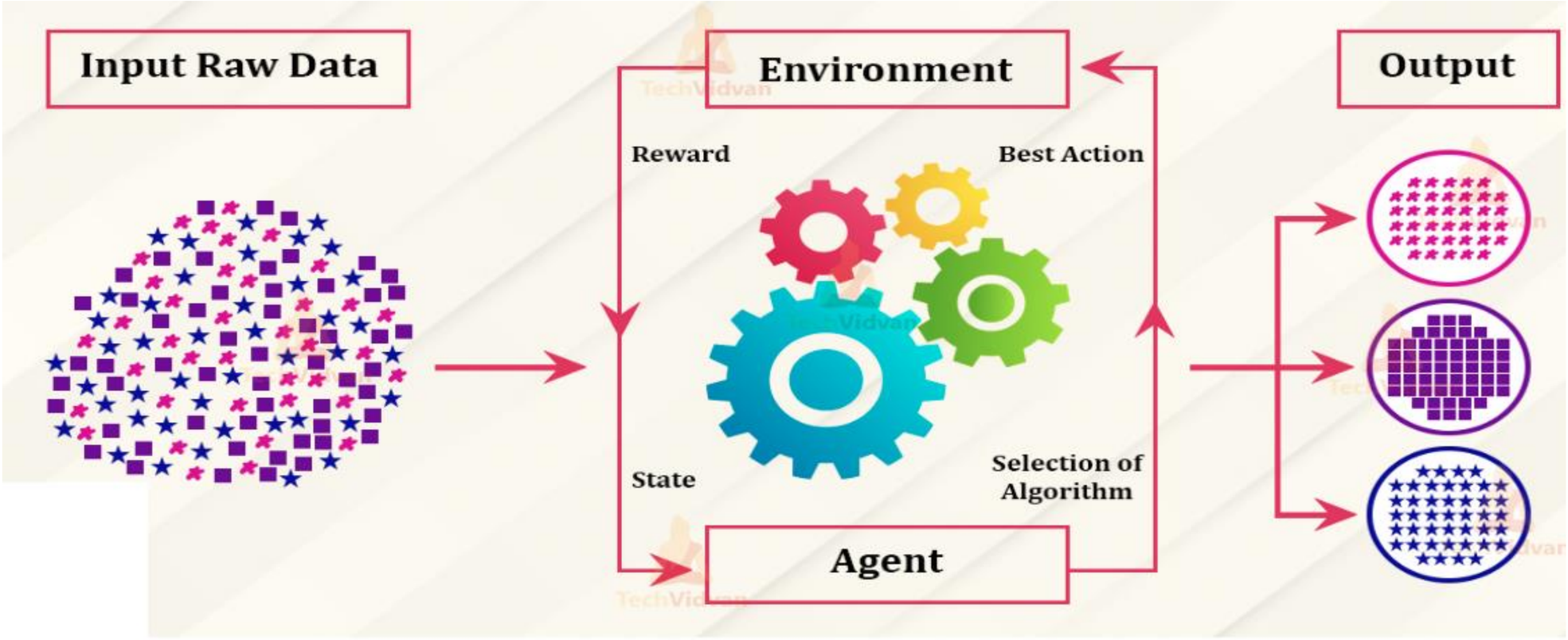
Recommender Systems

Develop a recommender system that uses machine learning algorithms to suggest relevant items, products, or content based on user preferences or behavior.



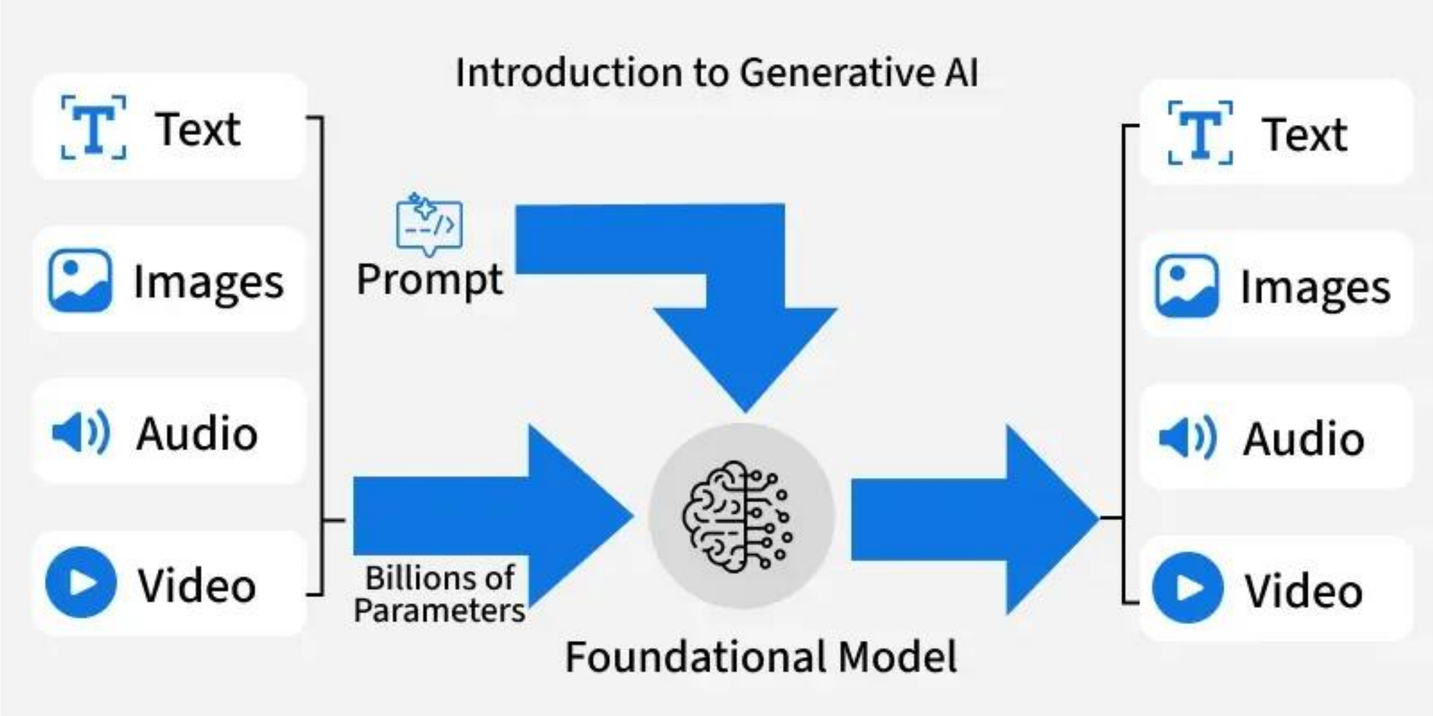
Reinforcement Learning

Develop a reinforcement learning–based system that learns optimal decisions or actions through interaction with its environment to maximize performance or rewards.



Generative AI

A type of artificial intelligence designed to create new content such as text, images, music or even code by learning patterns from existing data.



https://www.geeksforgeeks.org/artificial-intelligence/what-is-generative-ai/

Data Center – Cloud Computing

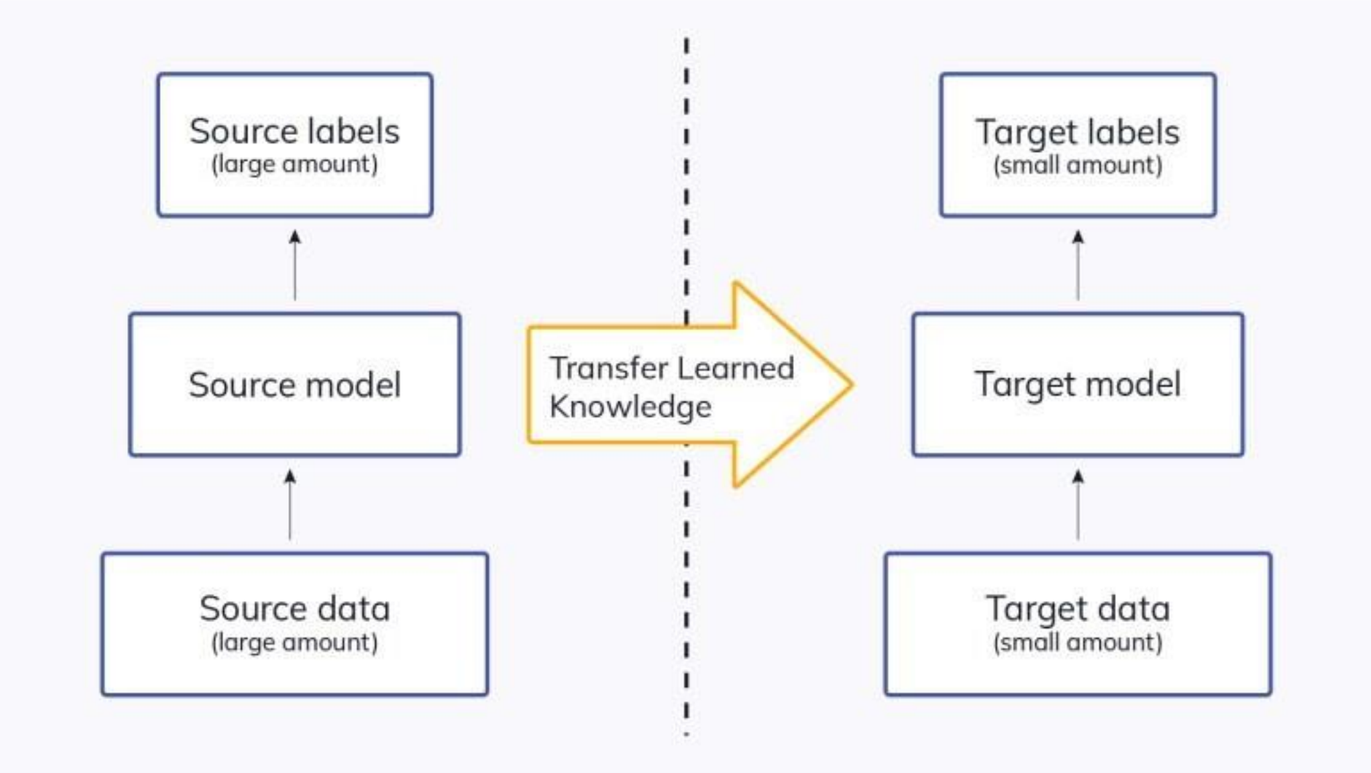
Develop a machine learning system that optimizes data center or cloud computing operations, improving performance, energy efficiency, and resource allocation.

Quantum Computers

Develop a machine learning model that leverages quantum computing principles to enhance computation speed, optimization, or data processing efficiency.

Transfer Learning

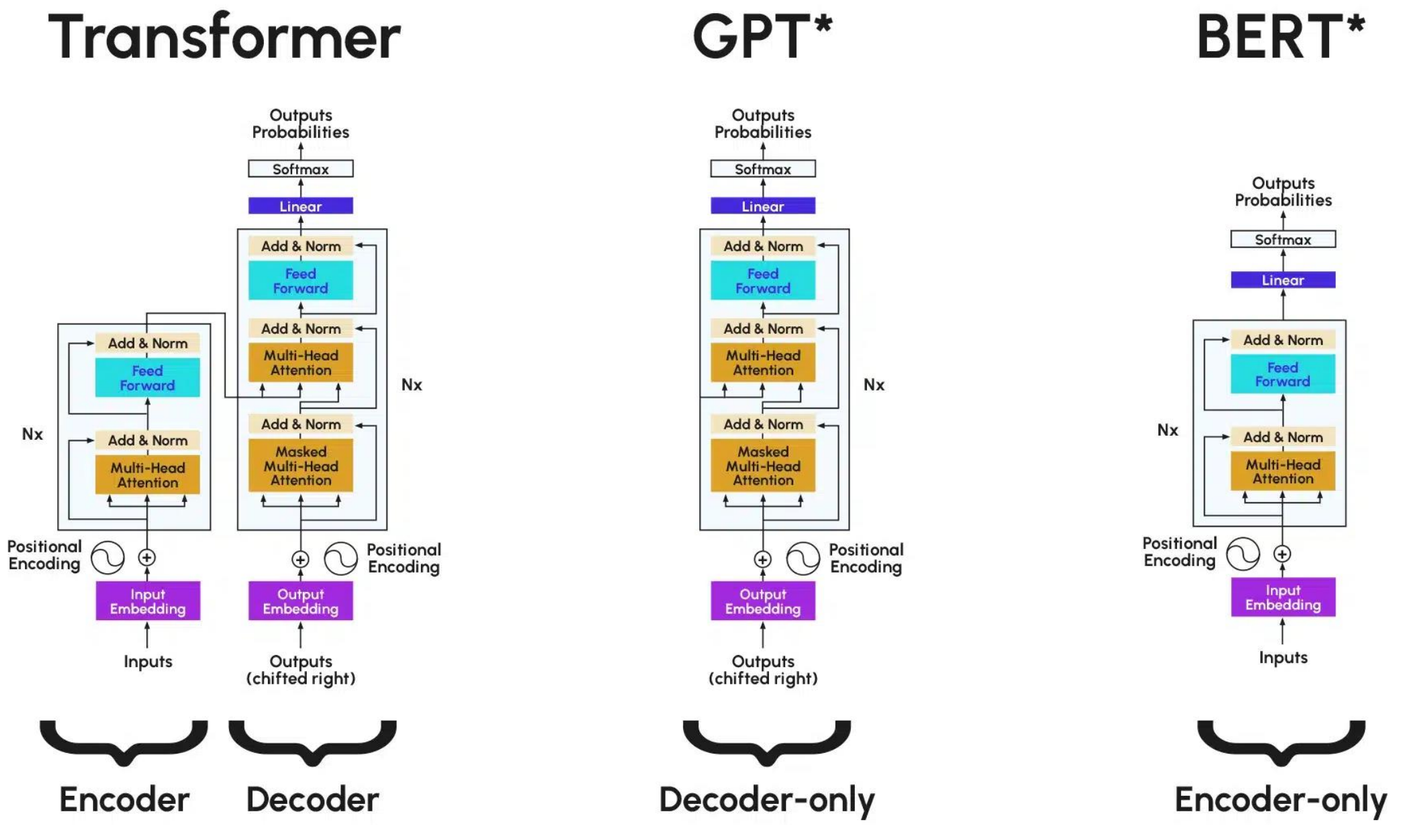
Develop a transfer learning–based model that adapts knowledge from a pre-trained network to solve a new but related problem efficiently with limited data.



https://www.v7labs.com/blog/transfer-learning-guide

Transformers

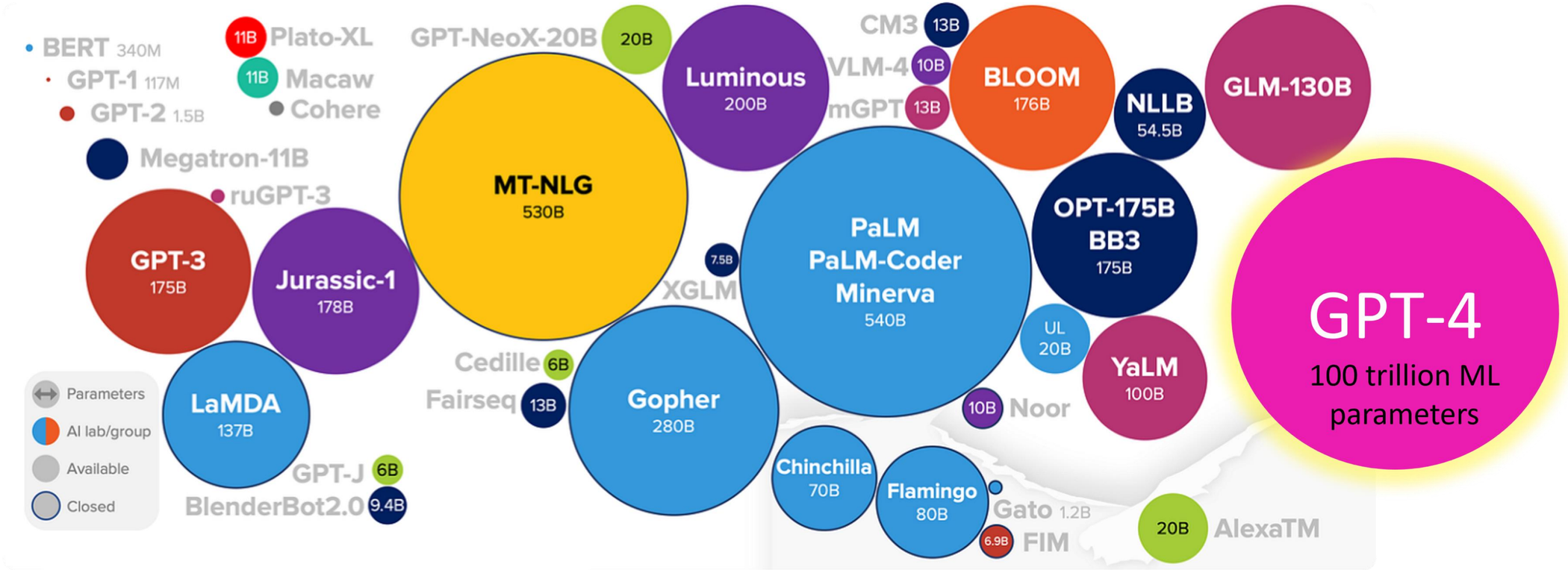
Develop a transformer-based AI system that leverages attention mechanisms to efficiently model relationships within data for tasks like text generation, translation, or image analysis.



https://datascientest.com/en/all-about-transformers-models

Large Language Models (LLMs)

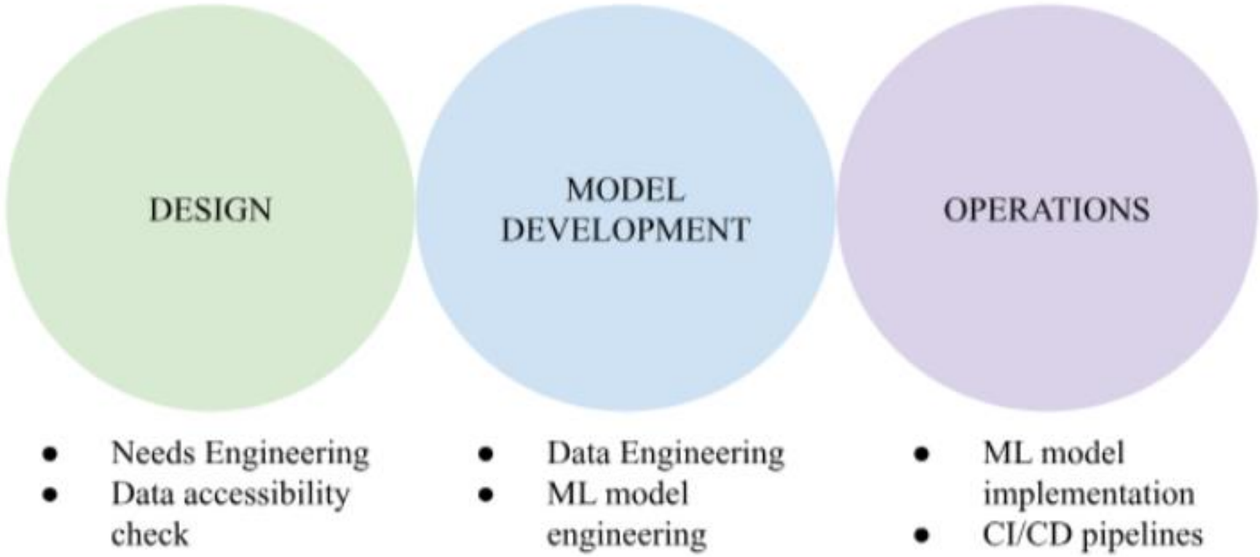
Develop a large language model–based system that understands and generates human-like text for applications such as summarization, question answering, or dialogue generation.



https://utcn-robotica.ro/expanding-creative-possibilities-exploring-the-synergy-between-large-language-models-llm-and-theory-of-inventive-problem-solving-triz/

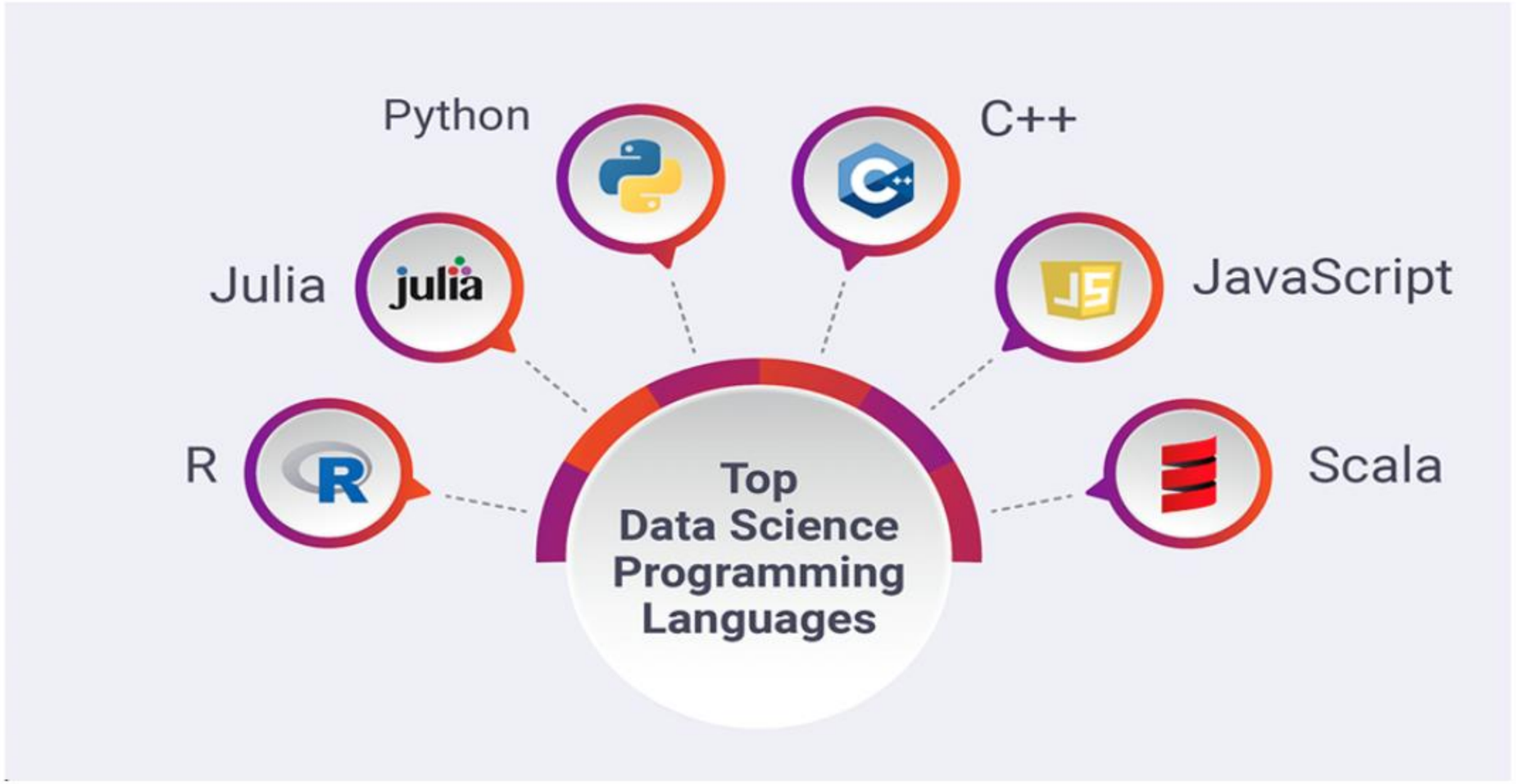
MLOps, or Machine Learning Operations

Develop an MLOps-based system that automates the deployment, monitoring, and maintenance of machine learning models in production environments.



https://www.linkedin.com/pulse/sharing-my-experiential-insights-related-implementing-balaji-t-0v9hc/

Most Popular Data Science Programming Languages



Most Popular Data Science Programming Languages

Python

```
print('Hello World')
```

Java

```
class HelloWorld {
   public static void main(String[] args) {
      System.out.println("Hello World!");
   }
}
```

C++

```
#include <iostream>

int main() {
   std::cout << "Hello World!";
   return 0;
}
```

C

```
#include <stdio.h>
int main() {
   printf("Hello, World!");
  return 0;
}
```

C#

```
namespace HelloWorld
{
   class Hello {
      static void Main(string[] args)
      {
         System.Console.WriteLine("Hello World!");
      }
   }
}
```

Most Popular Data Science Python Libraries for Machine Learning

https://deeptechbytes.com/most-popular-data-science-python-libraries-for-machine-and-deep-learning/

Wrong is Wrong!
Right is Right!

I need a better computer
I need to know English
I need to be an Engineer or Programmer
I need to be good at Math, Calculus, Linear Algebra etc.
Imposter Syndrome
The ship has sailed.

We only need is hard work, dedication, commitment, eager to learn and self confidence.

Future of AI Technology

https://www.fieldengineer.com/blogs/future-of-ai-technology

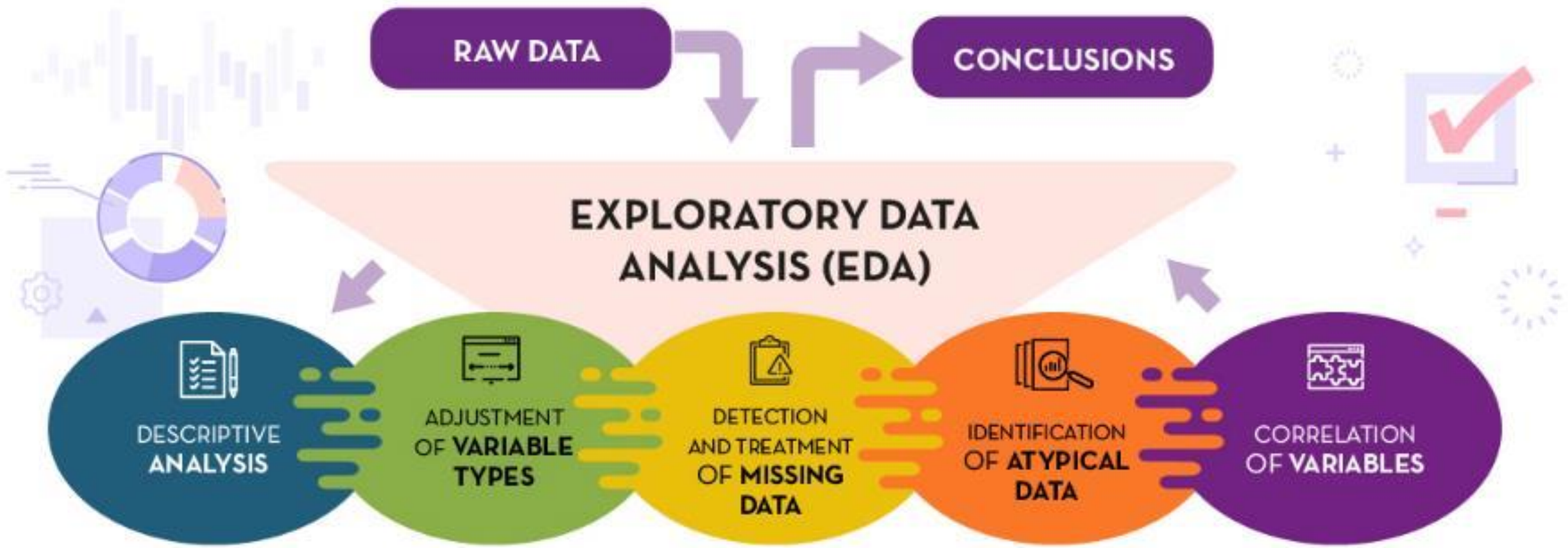
Accounts

www.github.com
www.kaggle.com
https://quantum-computing.ibm.com/
https://appinventor.mit.edu/
https://colab.research.google.com/
https://aws.amazon.com/machine-learning/
https://www.pythonanywhere.com/
https://huggingface.co/


THANK YOU

Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is **used by data scientists to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods.**



https://datos.gob.es/en/documentacion/practical-introductory-guide-exploratory-data-analysis/

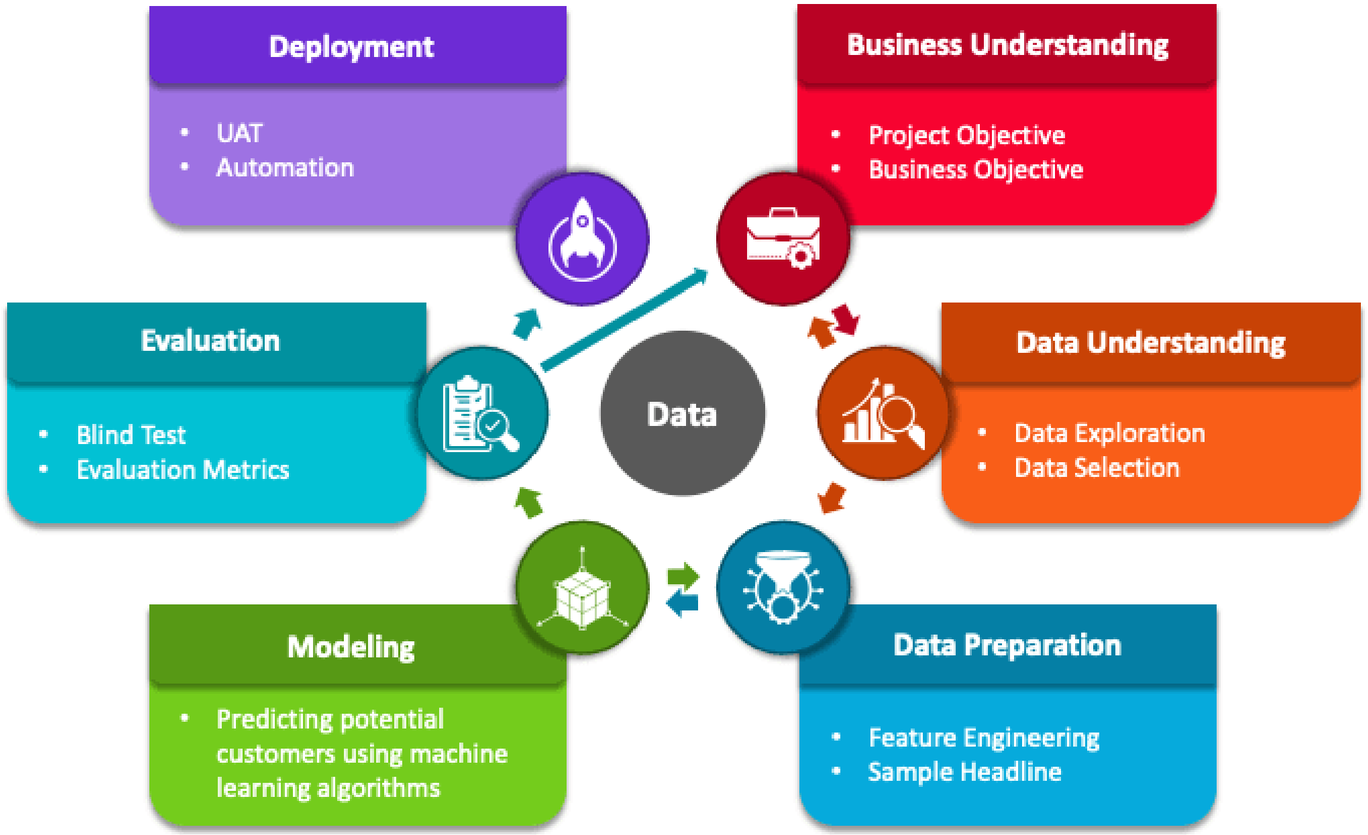
EDA: Data Wrangling Common Tasks (also known as data munging)

Restructuring, cleaning, and enriching the raw data available into a more processed format

https://favtutor.com/blogs/data-wrangling

What is CRoss Industry Standard Process for Data Mining (CRISP-DM) ?

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process.



https://www.sketchbubble.com/en/presentation-crisp-dm.html

Data Wrangling in Python

- Numpy (aka Numerical Python): It's the most basic python package for data science. One can perform operations on n-arrays and matrices in Python using Numpy. It provides vectorization of mathematical operations on the NumPy array type, which helps improve performance and accordingly speeds up the execution of the python code.

- Pandas: It makes data analysis operations faster and easier. Useful for data structures with labeled axes. Some data alignment prevents common errors that can be extracted from misaligned data during data scraping.

- Matplotlib: It's the most common python visualization module. One can create line graphs, pie charts, histograms, and other professional-grade figures.

- Plotly: for interactive, publication-quality graphs. Great for creating line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axis, polar graphs, and bubble charts.

- Theano: A python library for numerical computation similar to Numpy. This library is created to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently.

Application of Machine-Learning