Most of the students start learning **Data Science** with Python, Statistics, Machine learning, Deep learning. it’s really a **misconception**.

Technical skills are important, but without a **Solid understanding of data fundamentals is much important than technical skills**, without knowledge on data fundamentals those skills can be much less effective.

**1. First things first What is Data?**

**Data** refers to **raw facts or figures** without context. Its information collected for reference, analysis, or processing.

* Example: "24, 18, 12, 45" — These are data points.
* When organized and analyzed, data becomes **information**.

**2. Where does this data come from? (Data Collection Sources)**

* *Voluntary Data:* Data that people choose to give **freely**, often for surveys, apps, forms, or research.

Example: When you **fill out a survey** about your shopping habits — that's voluntary data.

***Scenario:*** A fitness app wants to learn about users' eating habits.

***How it's collected:*** The app asks users to log their meals and answer optional surveys.

***Result:*** Users willingly provide data about their food intake. The data scientist uses it to improve nutrition tips.

* *Infrared Data:* **Infrared data** refers to information collected using **infrared radiation**, which is **invisible heat energy** detected by sensors.

Example: When a drone uses infrared cameras to scan crop health in a field — that's infrared data.

***Scenario:*** A data scientist is helping farmers monitor crop health.

***How it's collected:*** They use drones with infrared cameras to fly over fields. Healthy plants reflect more infrared light, while stressed or dry plants reflect less.

***Result:*** The infrared data helps the data scientist map plant health, detect issues early, and improve crop yield.

* *Observed Data:* **Observed data** refers to information **collected through direct observation**, often without interference or manipulation by the observer.

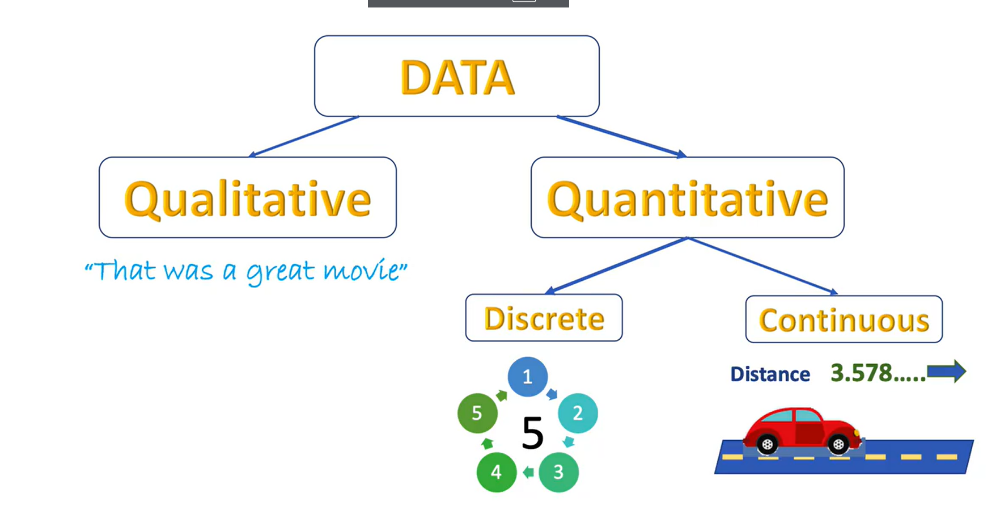
Example: When a store watches how customers move around using security footage — that's observed data.

***Scenario:*** A shopping mall wants to understand customer behavior.

***How it's collected:*** Security cameras record how people move through the mall. A data scientist analyzes video footage to track paths and time spent in different areas.

***Result:*** They find popular zones and help improve store placement.

**3. What Types of Data? (DATA VALUES)**

****

1. **Qualitative Data**: Descriptive, non-numeric information.

* **Example**: Colors, names, labels (e.g., "red," "small," or "happy").
* Purpose: To categorize or describe.

**Types of Qualitative Data**

**a. Nominal Data (Categorical Data):** The categories of data that have no specific order or ranking are called nominal data.

**Example**:

* + Eye color: **Blue, Brown, Green**
  + Types of fruit: **Apple, Banana, Orange**

Key Point: No inherent order. You can't say **one is better than the other**.

**b. Ordinal Data (Ordered Categories)**: The Categories that have a specific order or ranking, but the difference between categories isn't meaningful, are called ordinal Data.

**Example**:

* + Rating scale: **1 star, 2 stars, 3 stars**
  + Education level: **High School, College, Graduate**

Key Point: There's an order, but the **difference** between them isn't precisely measurable.

2. **Quantitative Data**: Numeric data that can be measured and counted.

* **Example**: Age, height, temperature (e.g., 25 years, 5.5 ft).
* Purpose: To quantify or measure.

**Types of Quantitative Data**

**a. Discrete Data (Countable):** Data that can only take specific, countable values. No fractions or decimals.

**Example**:

* + Number of students in a class: **25 students**
  + Number of cars in a parking lot: **50 cars**

Key Point: You can **count** the data, but it’s not continuous.

**b. Continuous Data (Measurable):** Data that can take any value within a given range, including fractions or decimals.

**Example**:

* + Height of a person: **5.7 feet, 5.75 feet**
  + Temperature: **72.3°F, 72.35°F**

Key Point: The data can have **infinite possible values** within a range.

***Commonly Data in:***

***Integer***: It is Whole numbers (like 1, 100, -5, 0). Think of them as numbers without any decimal points.

***Float***: Numbers with decimal points (like 3.14, -0.5, 2.0). They allow for more precise measurements.

***String***: Text (like "hello", "world", "123"). It's a sequence of characters.

***Boolean***: Represents truth or falsehood. It has only two possible values: True or False. Think of it as a simple yes/no or on/off switch.

**4. Data Formats? (Variety of Data/ Data Representation)**

Data is all around us, and it is coming at ever-increasing rates and in many different formats.

🔹 ***Structured Data:*** Structured data makes up about 10%-20% of generated data and has clearly defined data types and patterns that make them easily stored and organized into columns and rows. This organized structure makes easier to search and analyze.

Example:  sales records, airline reservation systems

(Structured data usually stored in Relational databases - SQL and in Excel spreadsheets)

🔹 **Unstructured Data:** Unstructured data makes up most data that is generated, about 80%, and cannot be organized into row and columns. This makes unstructured data difficult to search, manage, and analyze.

Example:  images, PDFs, sensor data, and social media posts

(Unstructured data is usually stored in a non-relational database- NOSQL and also known as NoSQL Database.)

🔹 **Semi-structured Data:** Semi-structured data sits between structured and unstructured data. It does not follow strict rows and columns, but it has some level of organization using tags, keys, or markers — making it easier to analyze than unstructured data.

Example: JSON files, XML, HTML, email metadata, MongoDB document   
(Semi-structured data is usually stored in NoSQL databases like MongoDB, CouchDB, or in file formats like JSON and XML.)

**Before diving into any data analysis, a data scientist or data analyst should definitely verify these crucial steps?**

* **Define Your Question: (Know what you're looking for)**
* **Understand Your Data: (Look at your information and Understand the Data)**
* **Determine Relevance: (Does this information help answer your question?)**

**5. Data Structures?**

**1. Arrays**

* From **NumPy**, used for numerical computations.
* Support multi-dimensional data and vectorized operations.
* Example: np.array([[1, 2], [3, 4]])

**2. Lists**

* Built-in Python collection, ordered and mutable.
* Can hold mixed data types.
* Example: [1, 'apple', 3.14]

**3. Tuples**

* Ordered, **immutable** Python collection.
* Often used for fixed-size records or function returns.
* Example: (42, 'banana')

**4. Dictionaries**

* Unordered **key-value** pair collections in Python.
* Keys must be unique; great for lookups.
* Example: {'name': 'Vicky', 'age': 25}

**5. Data Frames**

* From **Pandas**, used for tabular data (like spreadsheets or SQL tables).
* Each column can have a different type; supports indexing, filtering, grouping.
* Example: pd.DataFrame({'A': [1, 2], 'B': [3, 4]})

**6. Series**

* Also from **Pandas**; a single column of data with an index.
* Essentially a labeled 1D array.
* Example: pd.Series([10, 20, 30], index=['a', 'b', 'c'])

**6. BIG DATA?**

Big data is a term used to describe the massive volumes of digital data generated, collected, and processed. Every

Most of them already studied about Big Data and

**Big Data Characteristics**

1.Volume (Size of Data): Volume describes the amount of data transported and stored.

2.Variety (Different Types of Data): Variety describes the many forms data can take, most of which are rarely in a ready state for processing and analysis.

3. Velocity (Speed of Data generation): Velocity describes the rate at which this data is generated.

4.Veracity(accuracy): Veracity is the process of preventing inaccurate data from spoiling your data sets.

***Data Pipelines:***

A data pipeline is a series of interconnected data processing steps that move data from one or more **sources** to a **destination**, with the goal of making that data more useful.  It eliminates the manual task and allows the data to move smoothly. Thus, it also eliminates manual errors. It divides the data into small chunks and processes it parallelly, thus reducing the computing power.

* *Why Do we need?*

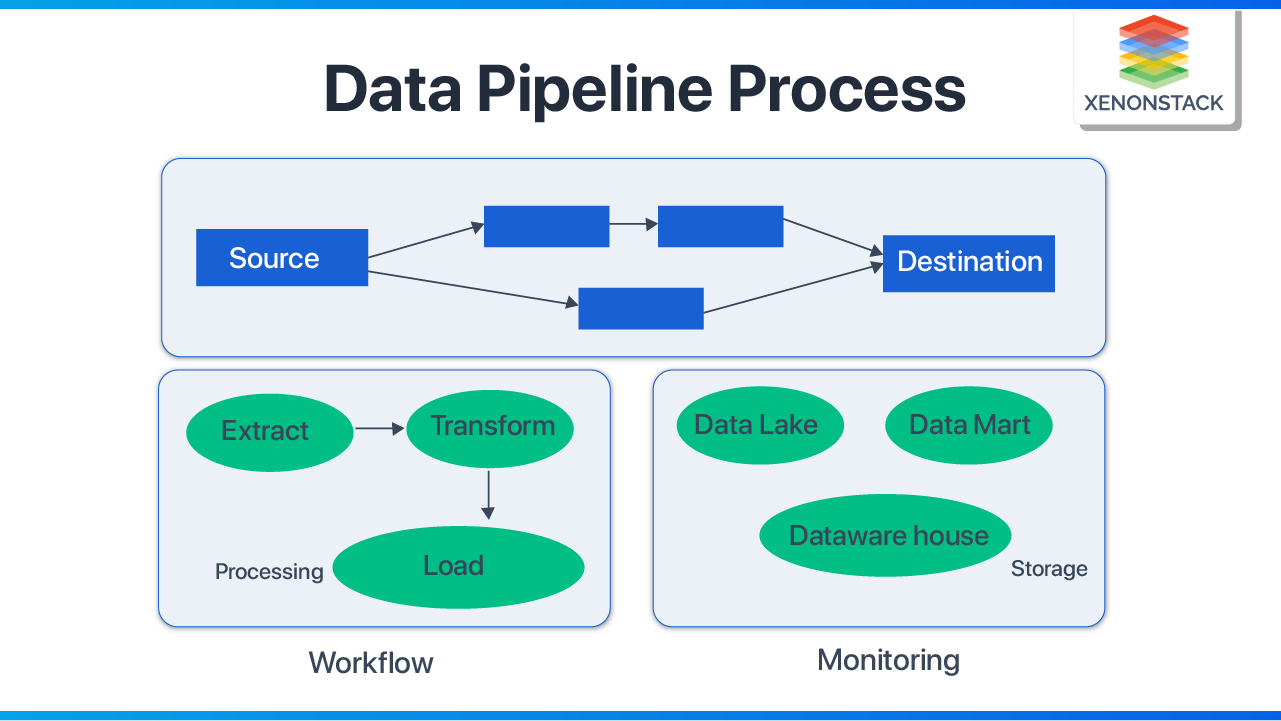
We need data pipelines to make raw, fragmented data usable for analysis and decision-making.

* *Where do we need?*

Data pipelines are needed in any system where data from multiple sources requires integration and preparation.

* *How do we use?*

We process and apply data pipelines by defining the flow, building the steps, automating execution, and utilizing the processed data.



* **In an easy way we will understand what is ETL?**

ETL stands for extract transfer load it is a process used in data warehousing to get data from different sources clean it up and put it into a single place for analysis.

**Extracting** is like gathering the raw materials

**Transforming** is like processing them

**Loading** is like putting them into a storage.

If you need a good understanding on ETL or data pipeline the reason video link please watch it you’ll get good understanding [***Data Pipeline Link***](https://www.linkedin.com/posts/vishnukumar-ds_datapipelines-etl-dataengineering-activity-7329367816707461120-q8Pw?utm_source=share&utm_medium=member_desktop&rcm=ACoAADKBHD4BjTBQ73tEzPQbF4xd4fNzgUkIdug)

Data engineers manage data through a data pipeline. The data pipeline has three stages: ingestion, transformation, and storage. These stages preclude any analysis that is to be performed. Data engineers work with two primary sources of data: batch ingestion and streaming ingestion.

Till now we understood about What is data and what type of data structures and What is big data? NEXT IS AI

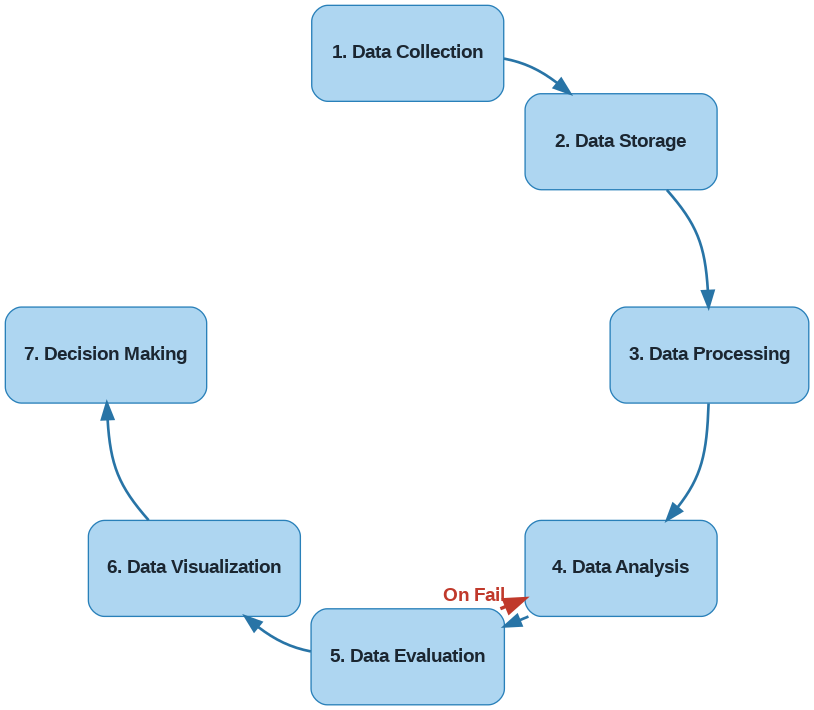
AI – Artificial intelligence

Nowadays AI is more advanced it can Helps us find the cure for cancer and helps in taking financial decisions with Data, Making vehicles self-driving to safer and more.

**Why it is Happening Now?**

Because Storing data and processing it is fast and Cheaper, Algorithms are becoming smart and Ai ecosystem is becoming Larger.

***DATA LIFE CYCLE***

******

**✅ 1. Data Collection & Data Sources**

Understanding:

* Where and how data is collected.
* Structured vs. unstructured sources (e.g., databases vs. logs or social media).
* APIs, web scraping, sensors, surveys, etc.

**✅ 2. Data Cleaning (Data Wrangling / Preprocessing)**

This is one of the most critical steps:

* Handling missing data
* Removing duplicates
* Standardizing formats (dates, categories)
* Outlier detection and treatment
* Encoding categorical variables

**✅ 3. Exploratory Data Analysis (EDA)**

This includes:

* Descriptive statistics (mean, median, mode, standard deviation, etc.)
* Data visualization (histograms, box plots, scatter plots)
* Understanding distributions, trends, patterns, and correlations

**✅ 4. Data Transformation**

Transforming data for modeling:

* Normalization and standardization
* Feature scaling
* Log transformations
* Binning, polynomial features
* Feature engineering

**✅ 5. Data Storage & Management**

Understanding how to store and retrieve data efficiently:

* File formats: CSV, JSON, Parquet, etc.
* Databases: SQL vs. NoSQL
* Data warehouses and data lakes
* Indexing, partitioning, schema design

**✅ 6. Data Ethics & Privacy**

As data becomes powerful, responsibility grows:

* Bias and fairness
* Data privacy (e.g., GDPR, HIPAA)
* Ethical handling and informed consent

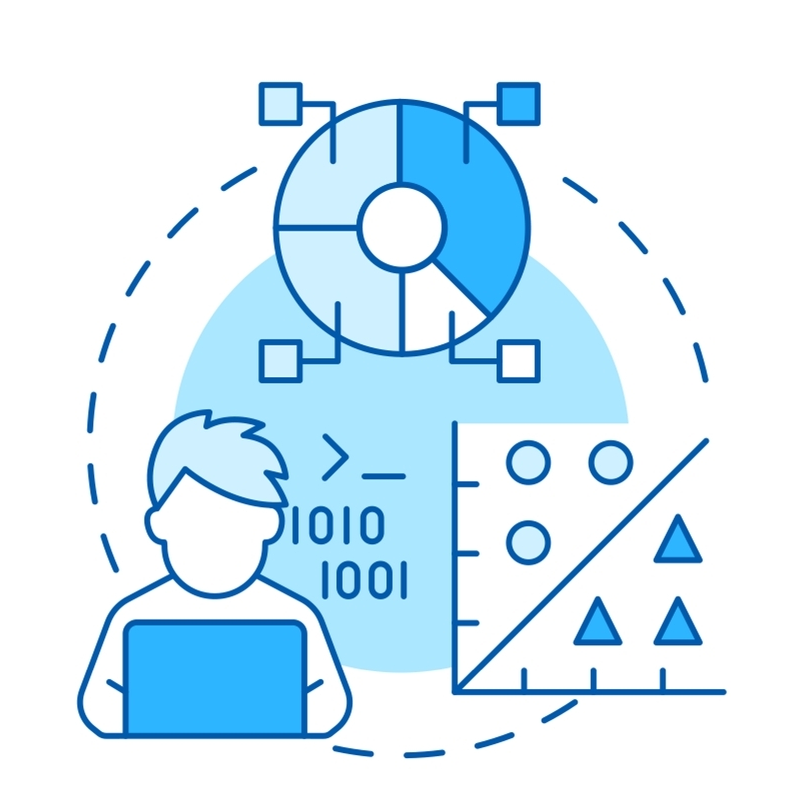
***Video Link (Above Concept)***

NEXT IS MACHINE LEARNING

Machine Learning: The ability of computer solved problems through programs or instructions given to them by human. Machine learning means the computer can figure out a solution without being specifically programmed. These machines have the ability to continuous learning by looking at the data they can find the patterns and they do much as faster than humans.

**Types of Machine Learning Analysis:**

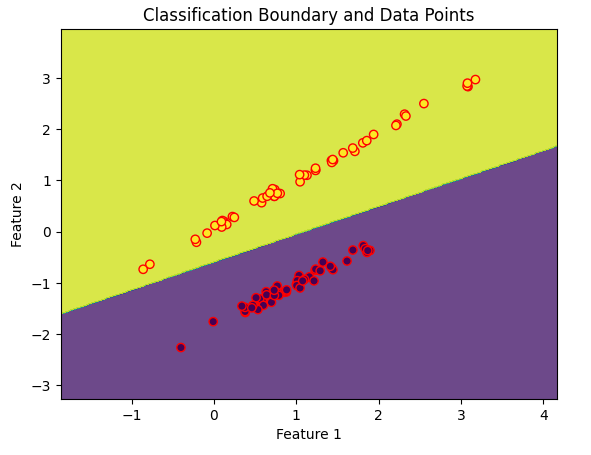
1. Supervised
2. Unsupervised
3. Reinforcement



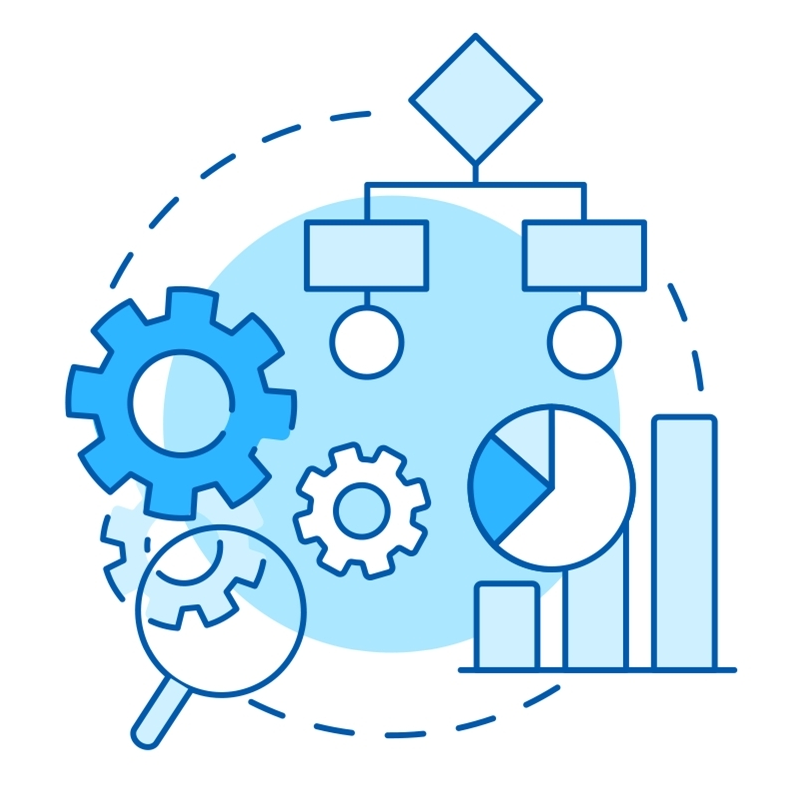
1.Supervised Machine Learning: **Supervised machine learning** algorithms are the most commonly used for predictive analytics. Supervised machine learning requires human interaction to label data read for accurate supervised learning.

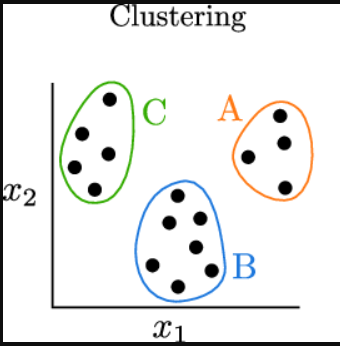


* Regression – Regression involves estimating the mathematical relationship between a continuous variable and one or more other variable (dependent variable and one or more independent variable)
* Classification – Classification consist of a discrete unknown variables in this we need to estimate the new sample category belongs to the set of predefined categories.

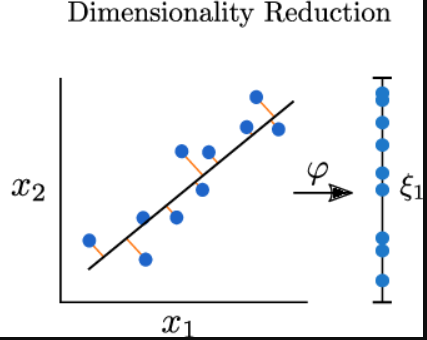


2.Unsupervised Learning: **Unsupervised machine learning** algorithms do not require human experts but autonomously discover patterns in data. Unsupervised learning mainly deals with unlabeled data.

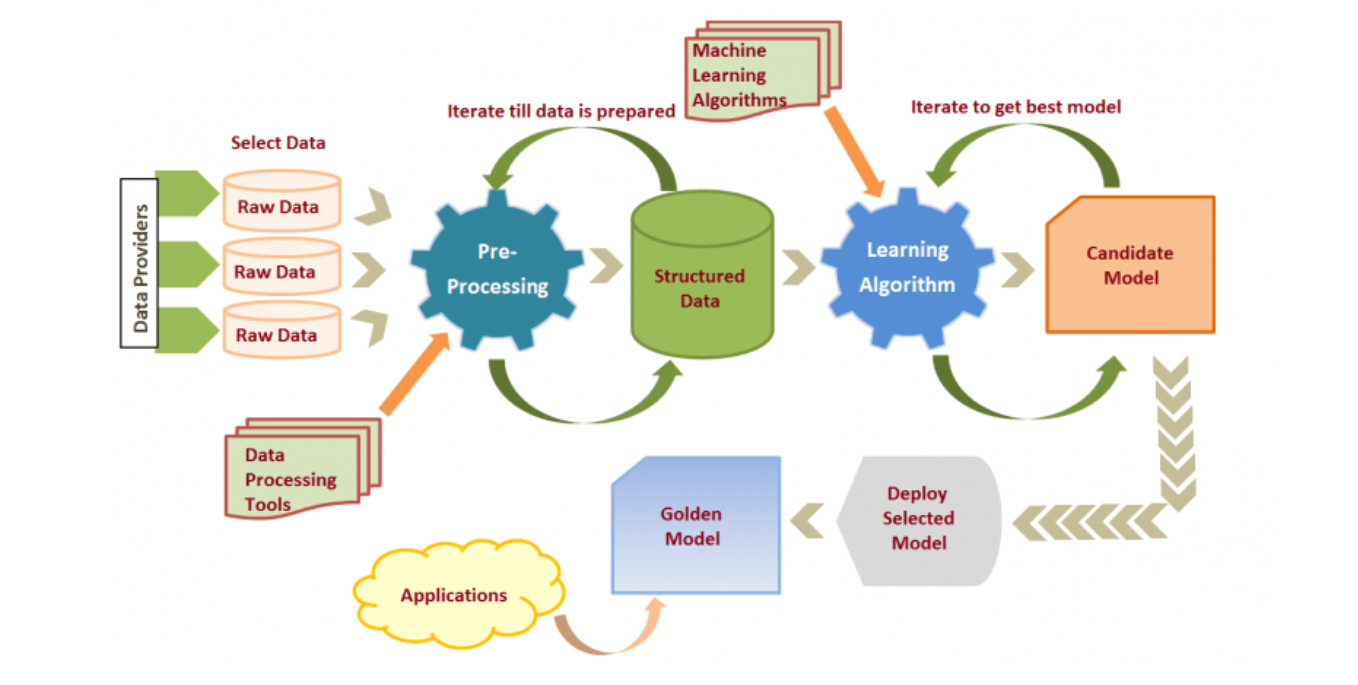




* Clustering -- Clustering is the grouping of data that have similar characteristics. It helps segment data into groups and analyzes each to find patterns.
* Dimensionality Reduction -- Dimensionality reduction is the process of reducing the number of features or variables in a dataset while retaining as much important information as possible.



* Association -- Association learning discovers interesting relationships or co-occurrences among variables in large datasets.

***MACHINE LEARNING PROCESS:***  


* **Raw Data Collection**: Data providers supply raw data which is selected for use.
* **Pre-processing**: Raw data is cleaned and structured using data processing tools.
* **Model Training**: Machine learning algorithms are applied to the structured data to train candidate models.
* **Model Selection**: The process iterates to find the best model, which becomes the **Golden Model**.
* **Deployment**: The selected model is deployed and used in real-world applications.

**This Is the Fundamentals of Data**

**NEXT Go to Stats and Math’s**

**Thank You**