Introduction to Machine Learning

Sadia Islam
Assistant Professor
Department of Computer Science and engineering
United International University

Outline

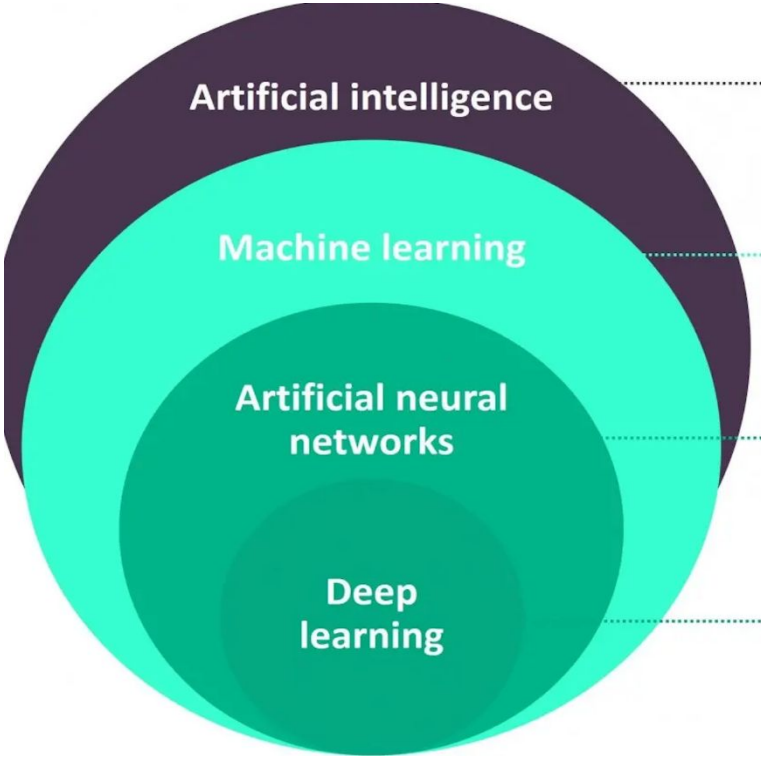
- ❏ Computational Intelligence
- ❏ Artificial Intelligence
- ❏ Machine Learning
- ❏ Neural Network
- ❏ Deep Learning
- ❏ Machine learning processes
- ❏ Data analysis

Computational Intelligence

Computational Intelligence (CI) or **Intelligent Computing** also known as **Soft Computing**, is the process of Knowledge Engineering and advance information processing that deals with creating mining algorithms and systems that can learn from Big Data and make decisions.

- **Knowledge Engineering:** Knowledge engineering is a field of artificial intelligence (AI) that tries to emulate the judgment and behavior of a human expert in a given field.

Artificial Intelligence, Machine Learning, Neural Network and Deep Learning



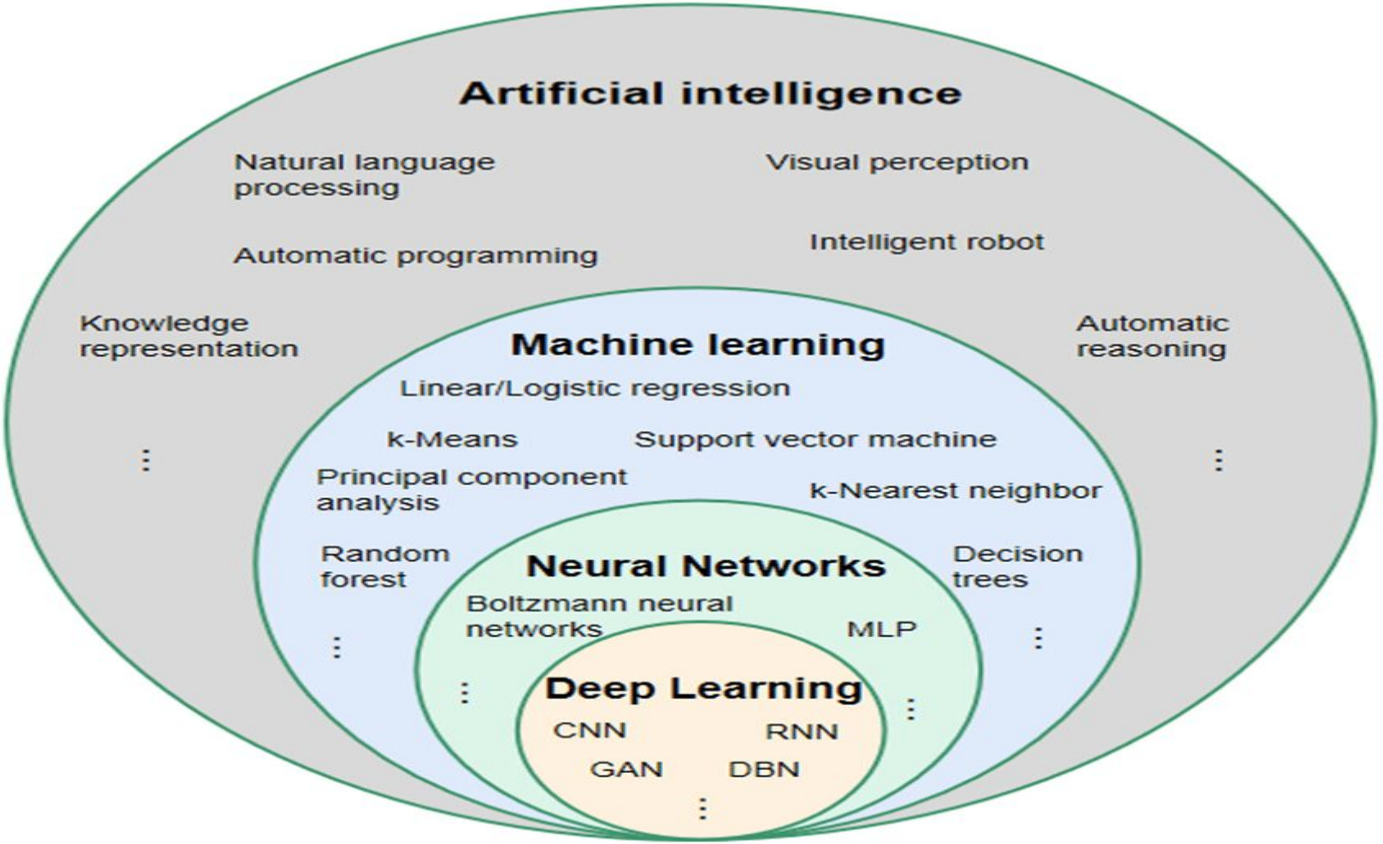
Artificial intelligence (AI)
Any techniques that enable machines to solve a task in a way like humans do

Machine learning (ML)
Algorithms that allow computers to learn from examples without being explicitly programmed

Artificial neural networks (ANN)
Brain-inspired machine learning models

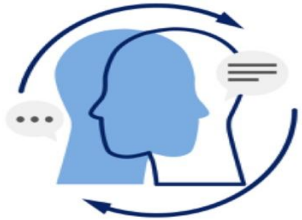
Deep learning (DL)
A subset of ML which uses deep artificial neural networks as models and automatically builds a hierarchy of data representations

Some available algorithms of each type

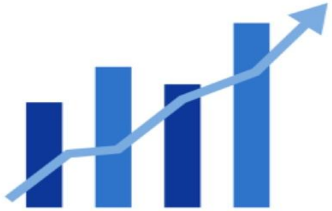


Why Machine Learning

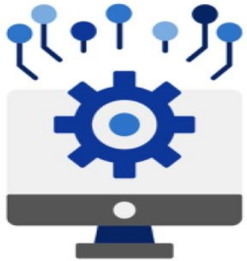
Automation



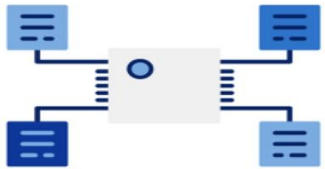
Less reliance on human interaction



Scope of improvement

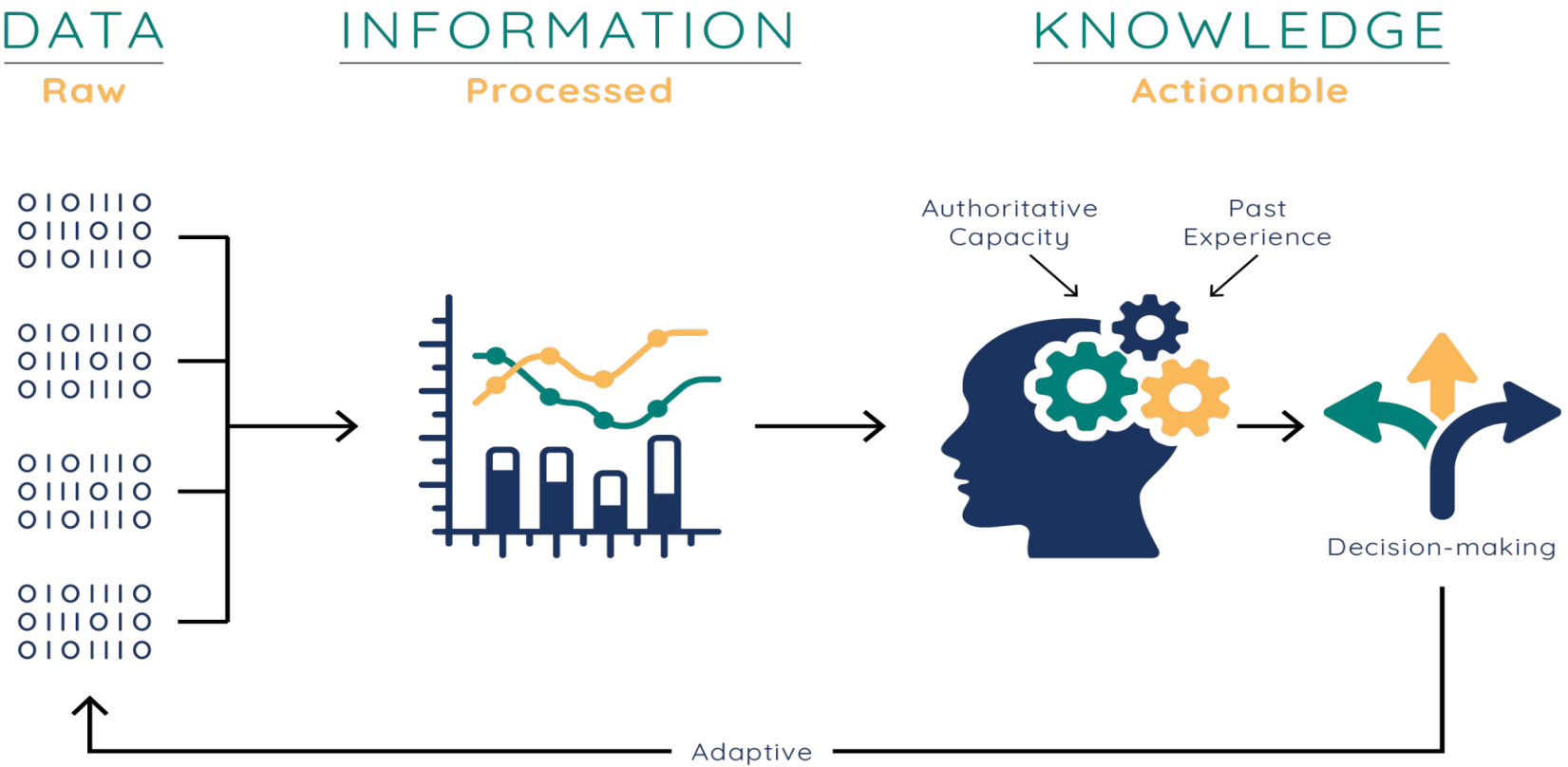


Efficient data handling



Wide range of applications

Data, Information, Knowledge and Wisdom



Data, Information, Knowledge and Wisdom Cont.

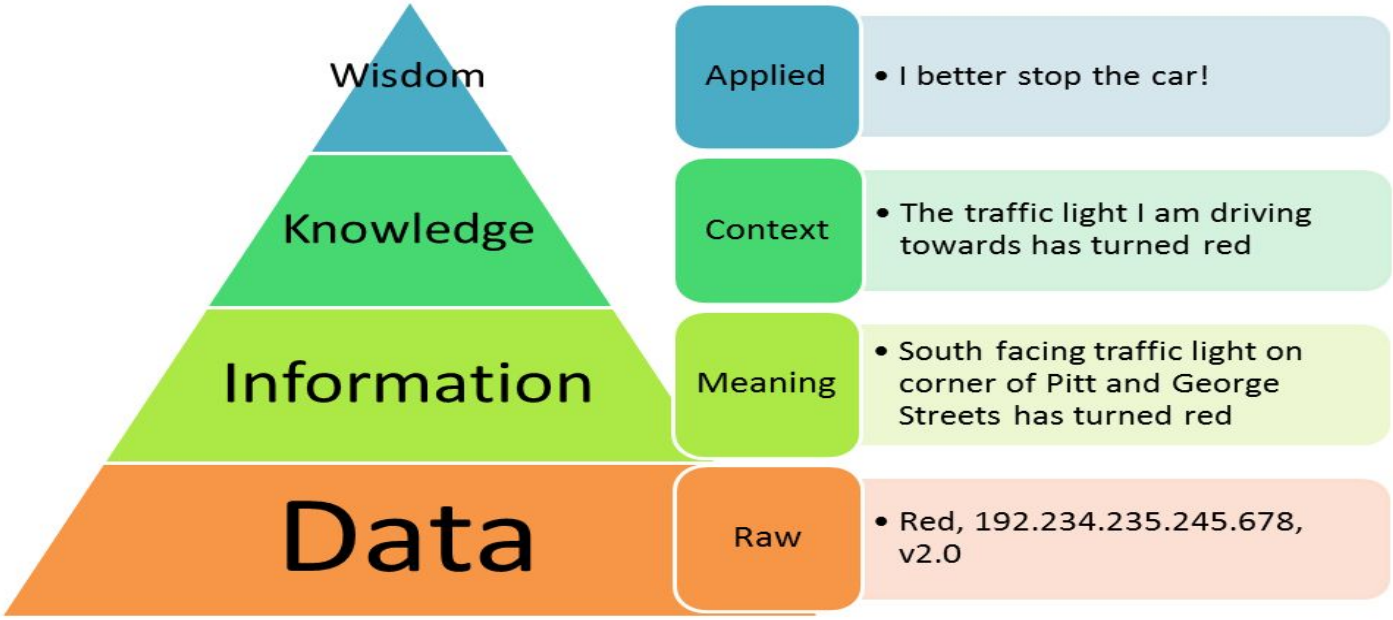
Data refers to raw, unprocessed facts and figures without context. It is the foundation for all subsequent layers but holds limited value in isolation.

Information is organized, structured, and contextualized data. Information is useful for answering basic questions like "who," "what," "where," and "when."

Knowledge is the result of analyzing and interpreting information to uncover patterns, trends, and relationships. It provides an understanding of "how" and "why" certain phenomena occur.

Wisdom is the ability to make well-informed decisions and take effective action based on understanding of the underlying knowledge.

Data, Information, Knowledge and Wisdom Example

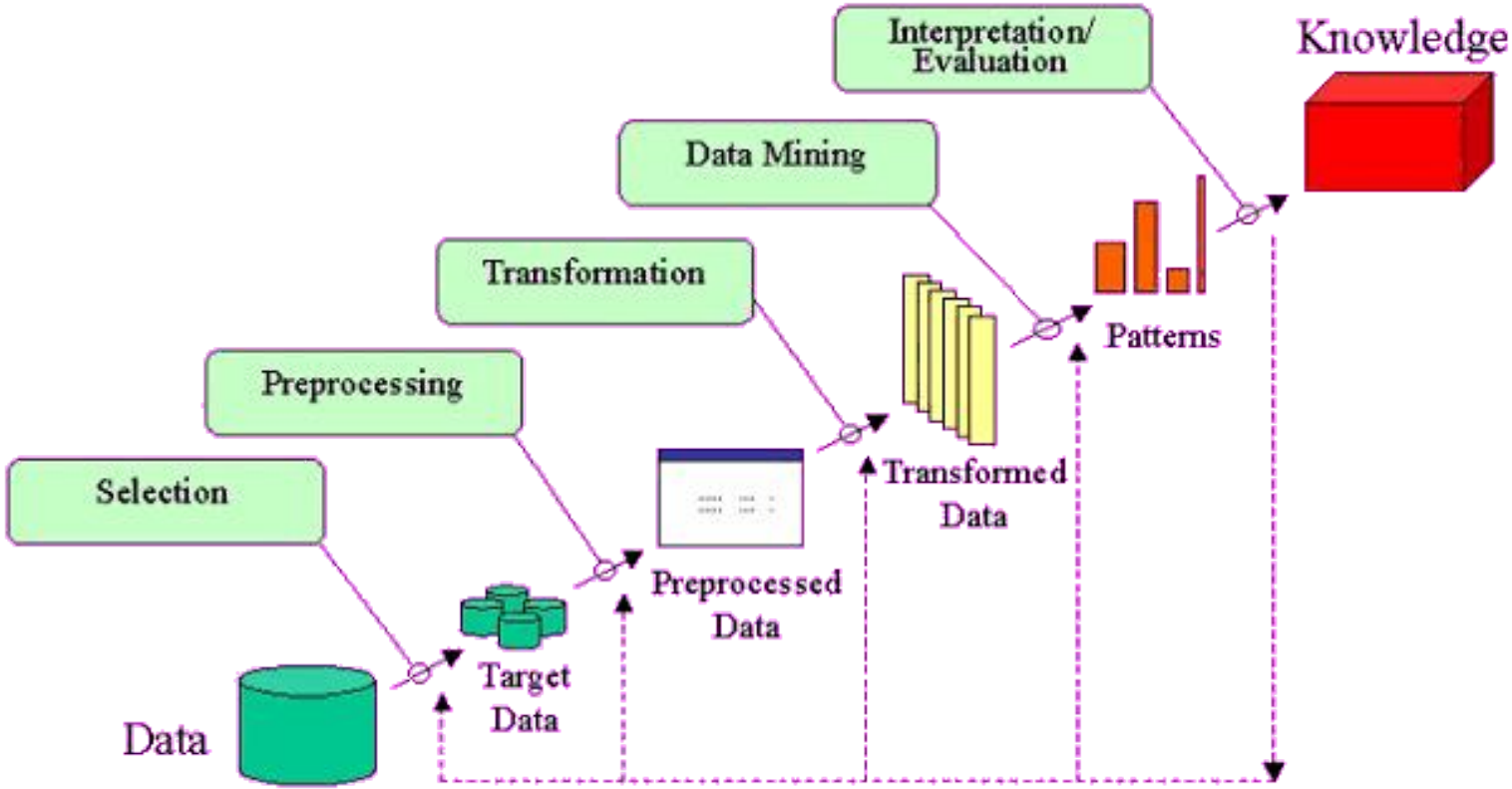


Basic definitions

Data Science: Data Science is the field of study that combines knowledge of Artificial Intelligence, Machine Learning, Data Mining to extract knowledge and insights from structured and unstructured Big Data.

Big Data: Big Data is defined by 3 V's: (1) Volume, (2) Variety, and (3) velocity. We can add few more V's e.g. Variability, Veracity, Value, and Visualisation.

Knowledge Discovery in database (KDD)



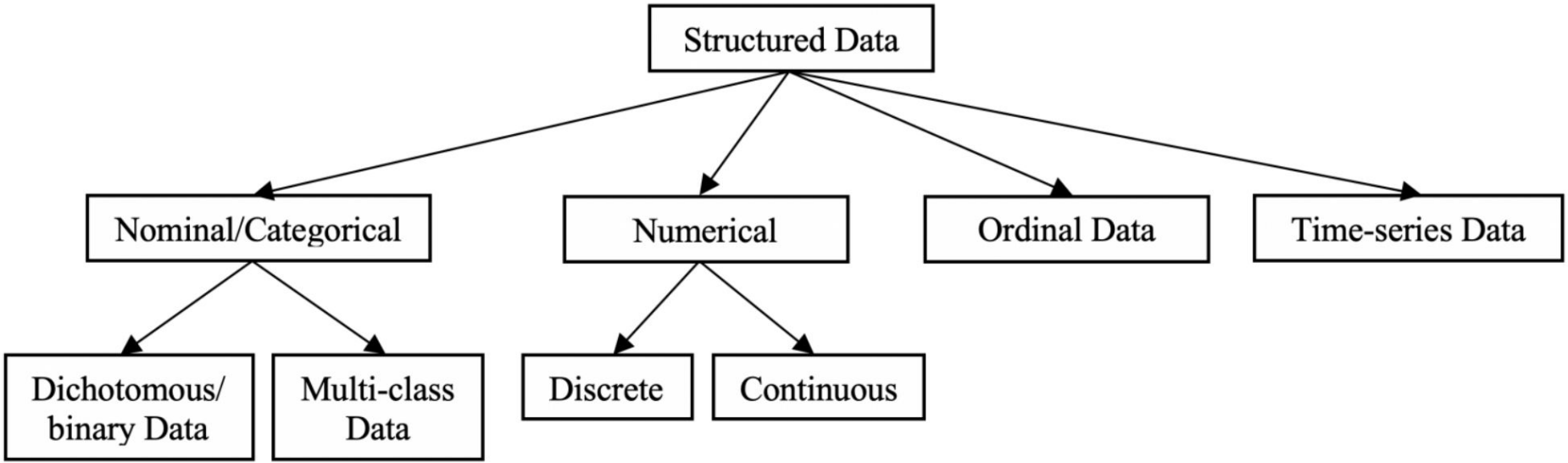
Data Analysis vs Data Mining

- **Data Analysis :** Data Analysis involves extraction, cleaning, transformation, modeling and visualization of data with an objective to extract important and helpful information which can be additional helpful in deriving conclusions and make choices.
  The main purpose of data analysis is to search out some important information in raw data so the derived knowledge is often used to create vital choices.

- **Data Mining :** Data mining could be called as a subset of Data Analysis. It is the exploration and analysis of huge knowledge to find important patterns and rules.

Types of data

- Unstructured data: Unstructured data have no rigid structure, e.g. images, video, natural language text, and speech etc.
- Structured data: Structured data is like a table.



Machine Learning

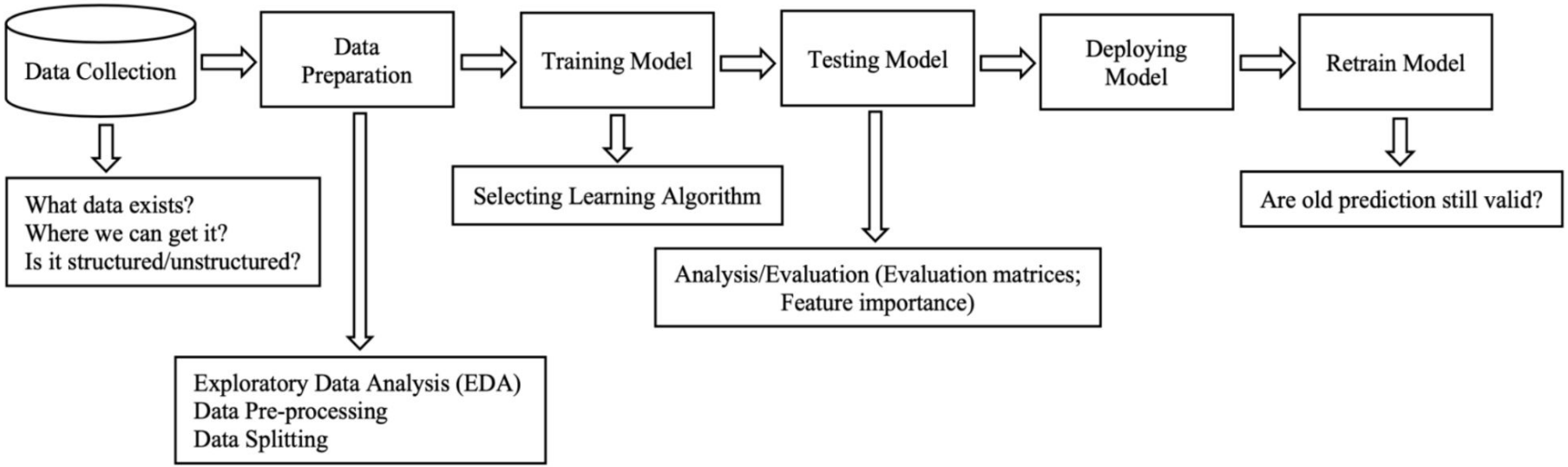
Machine learning is made up of three parts:

- The computational algorithm at the core of making determinations.
- Variables and features that make up the decision.
- Base knowledge for which the answer is known that enables (trains) the system to learn.

Machine learning input/output

1. Input: Concept, Instance and Features
2. Output: Label

Process of Machine learning

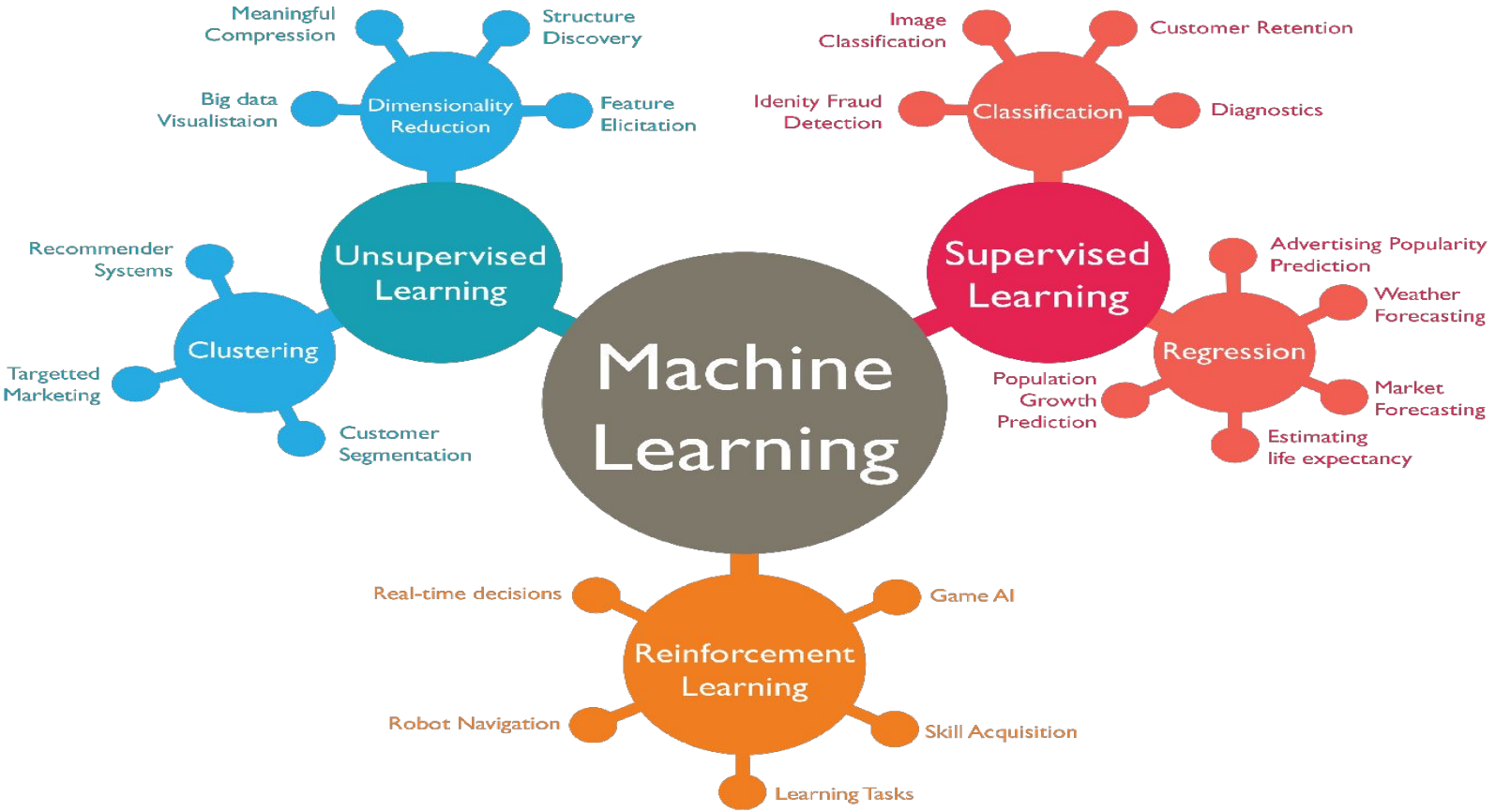


Exploratory Data Analysis (EDA) involves analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.

Learning about the data/understanding and knowing the data:

- What are the feature variables (input) and the target variable (output)?
- What's the kind of data? e.g. structured/unstructured
- Are there missing values?
- Where are the outliers? how many of them are there? why are they there?

Categories of Machine Learning processes



A. Supervised Learning

1. Classification

Predicting if a student will pass or fail based on features like study hours, sleep hours, attendance, and previous scores.

Study Hours	Sleep Hours	Attendance (%)	Previous Scores (%)	Pass/Fail
4	7	80	75	Pass
2	6	60	55	Fail
3	8	90	65	Pass
5	5	70	70	Pass
1	4	50	45	Fail

Types of classification problems

- Binary-class
- Multi-class
- Multi-label

Age	Inc ($)	Visits	Purchase (Yes/No)
28	30000	10	No
34	70000	5	Yes
22	25000	8	No
45	90000	2	Yes
37	50000	6	Yes

Wt (kg)	Eng (L)	HP	Vehicle Type
1500	2.0	150	Sedan
1800	2.5	200	SUV
1200	1.6	100	Hatchback
2000	3.0	250	Truck
1400	1.8	130	Coupe

Bud (M)	Dur (min)	Dir Score	Genres
100	120	8.0	Action, Sci-Fi
50	90	6.5	Romance, Drama
30	110	7.0	Comedy, Family
150	140	9.0	Action, Thriller
70	130	8.5	Drama, Biography

Supervised Learning Cont.

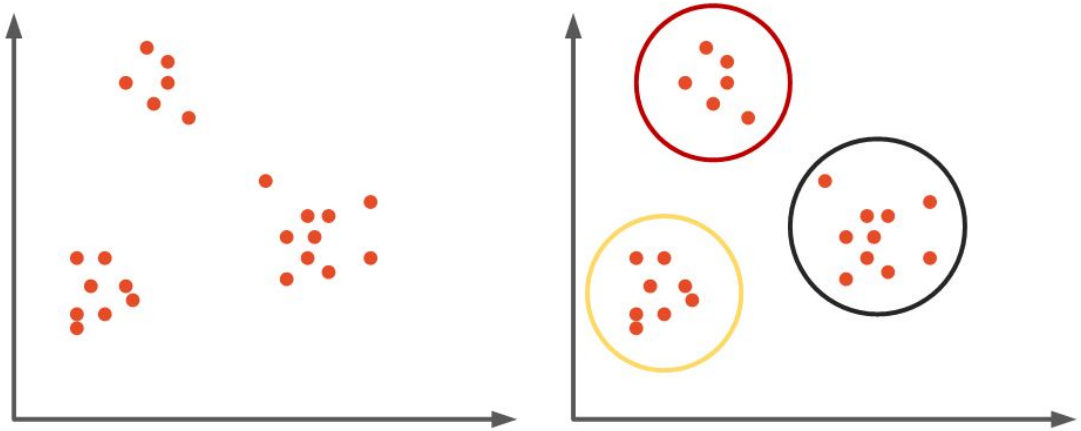
2. Regression

- Maps input to output where output is
  - Numerical
  - Real
  - Continuous
- Example of predicting house prices based on square footage, number of bedrooms, number of bathrooms, and age of the house.

Square Footage	Bedrooms	Bathrooms	Age (years)	House Price ($)
1500	3	2	10	350000
2000	4	3	5	500000
1200	2	1	30	200000
1800	3	2	15	450000
2500	4	3	8	600000

B. Unsupervised

- Data having no label
- ML tries to find inherent structures in the data based on the features
- Clustering(K Means, Hierarchical etc.)



Semi Supervised

- In between supervised and unsupervised learning
- Both labeled and unlabelled data

Semi Supervised Example

- Training a model to classify emails as spam or not spam.
- Labeled data

Words	Length (words)	Label
Buy now	20	Spam
Meeting at 10	50	Not Spam
Limited offer	15	Spam

- Unlabeled data

Words	Length (words)
Free gift	30
Lunch at noon	40
Congratulations, you won	25

- Train the model on the labeled data, then use that model to predict labels for the unlabeled data.

C. Reinforcement

- Training an agent to make a sequence of decisions by rewarding or punishing its actions.
- Rewarding or punishing based on an objective function.

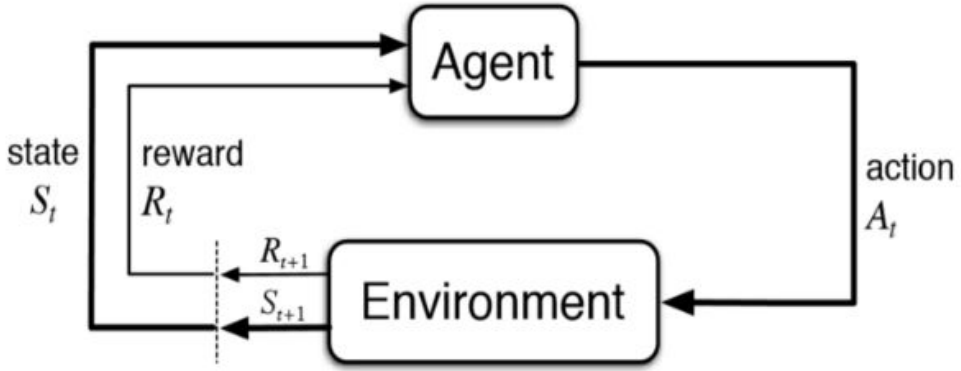


Figure 1.4: Reinforcement Learning.

Reinforcement - Example

- AI agent is playing Tic-Tac-Toe.
- The agent (player) interacts with the environment (game board) by taking actions (placing X or O). Each action leads to a new state (updated board) and the agent gets a reward (win, lose, or draw). The goal is to maximize cumulative rewards (winning more games).

Step	Description
1.	Initialize: Agent starts knowing nothing, randomly placing moves.
2.	Play Game: Agent plays against an opponent (could be itself).
3.	Receive Feedback: Agent gets feedback (reward) based on game outcome. - Win: +1 - Draw: 0 - Lose: -1
4.	Update Strategy: Agent uses feedback to update its strategy, learning to play better.

Transfer Learning

- Lack of data.
- Use a pretrained model trained with similar types of data
- No need to start from scratch

Active Learning

- Label and unlabeled data
- Need to label the unlabeled data
- Choosing some(not all) most informative data points to label manually
- Trained on labeled + newly labeled data points

Data Collection

To collect the data we need to ask the following questions:

- What kind of problems are we trying to solve?
- What data sources already exist?
- What privacy concerns are there?
- Is the data public?
- Where should we store the data?

Data Preparation

- Transforming raw data into a meaningful one.
- **EDA**: Preparing raw data for further processing.
- **Data preprocessing**: Transforming to machine learnable format to be modeled.
- **Data Splitting**

The **data pre-processing** process involves the following stages,

1. Data cleaning
2. Data integration
3. Data transformation
4. Data reduction
5. Data discretisation

1. Data Cleaning

- Imputing missing values
- Smoothing noisy data
- Resolving inconsistencies
    - Imputing(Mean) example

Original Data:

Age	Income
25	50000
30	NA
45	75000
NA	62000
50	80000

Using mean for Age and Income:

Age	Income
25	50000
30	67000
45	75000
37.5	62000
50	80000

2. Data Integration

- Different representations(features) are merged together
- From different databases, files, tables etc.
- Example of integrating two tables.

Dataset 1: Customer Info

Customer ID	Name	Age
101	John	30
102	Alice	25
103	Bob	35

Dataset 2: Purchase Records

Customer ID	Product	Amount (\$)
101	Laptop	1200
102	Smartphone	800
103	Tablet	300

Integrated Dataset:

Customer ID	Name	Age	Product	Amount (\$)
101	John	30	Laptop	1200
102	Alice	25	Smartphone	800
103	Bob	35	Tablet	300

3. Data Transformation

- Encoding (turning values into numbers)
- Normalization (Scaling)
- Aggregation

Data Transformation Cont.

- **Encoding**

  Convert categorical data into numerical form. Consider a "Color" column:

Color
Red
Blue
Green

  Encoded Color using One-Hot Encoding:

Red	Blue	Green
1	0	0
0	1	0
0	0	1

Data Transformation Cont.

- **Normalization**

  Transforming features to a common scale, like converting values between 0 and 1.
  Suppose we have a column for age:

Age
20
40
60

  Normalizing Age using Min-Max scaling:

$$\text{Normalized Age} = \frac{\text{Age} - \min(\text{Age})}{\max(\text{Age}) - \min(\text{Age})}$$

Age	Normalized Age
20	0.0
40	0.5
60	1.0

Data Transformation Cont.

- **Aggregation**

  Let's take student scores in different subjects and aggregate them to get their total and average score.

Original Data

Student ID	Math	Science	English
101	85	90	75
102	78	88	82
103	92	85	80

Aggregated Data

Student ID	Total Score	Average Score
101	250	83.33
102	248	82.67
103	257	85.67

4. Data Reduction

- Feature Selection (selecting most valuable features)
- Dimensionality reduction [Principal Component Analysis (PCA)]
- Clustering (Dividing the data into multiple clusters)

Initial Features

Age	BMI	Blood Pressure	Glucose Level	Insulin Level	Diabetes (Yes/No)
50	30	80	150	85	Yes
40	25	75	140	78	No
60	35	90	160	95	Yes
45	28	85	145	80	No
55	32	88	155	87	Yes

Selected Features

Using feature selection, we determine that the most important features for predicting diabetes are Age, BMI, and Glucose Level.

Age	BMI	Glucose Level	Diabetes (Yes/No)
50	30	150	Yes
40	25	140	No
60	35	160	Yes
45	28	145	No
55	32	155	Yes

4. Data Discretisation

- Converting data from Continuous to Discrete

- Imagine having a list of ages: 23, 45, 31, 50, 27
  - Using data discretization, we can convert these continuous values into categories (or bins). For example:
    - Ages 20-29 -> Young Adult
    - Ages 30-39 -> Adult
    - Ages 40-49 -> Middle-aged
    - Ages 50+ -> Senior
  - So, the list would become
    - 23 -> Young Adult
    - 45 -> Middle-aged
    - 31 -> Adult
    - 50 -> Senior
    - 27 -> Young Adult

Data Splitting

Splitted to:

- **Training set:** Used to train the machine learning model, this is the core dataset where the model learns to understand patterns and relationships in the data.
- **Validation set:** Assists in fine-tuning the model. It evaluates the model's performance during the training phase, helping adjust hyper-parameters and prevent over-fitting.
- **Test set:**Provides a fair evaluation of the model's performance on unseen data. This is crucial for assessing the model's ability to generalize to unknown data.

Splitting types:

- Random Splitting
- Stratified Splitting
- Time Series Splitting
- K-Fold Cross-Validation
- Leave-One-Out Cross-Validation

Thank You