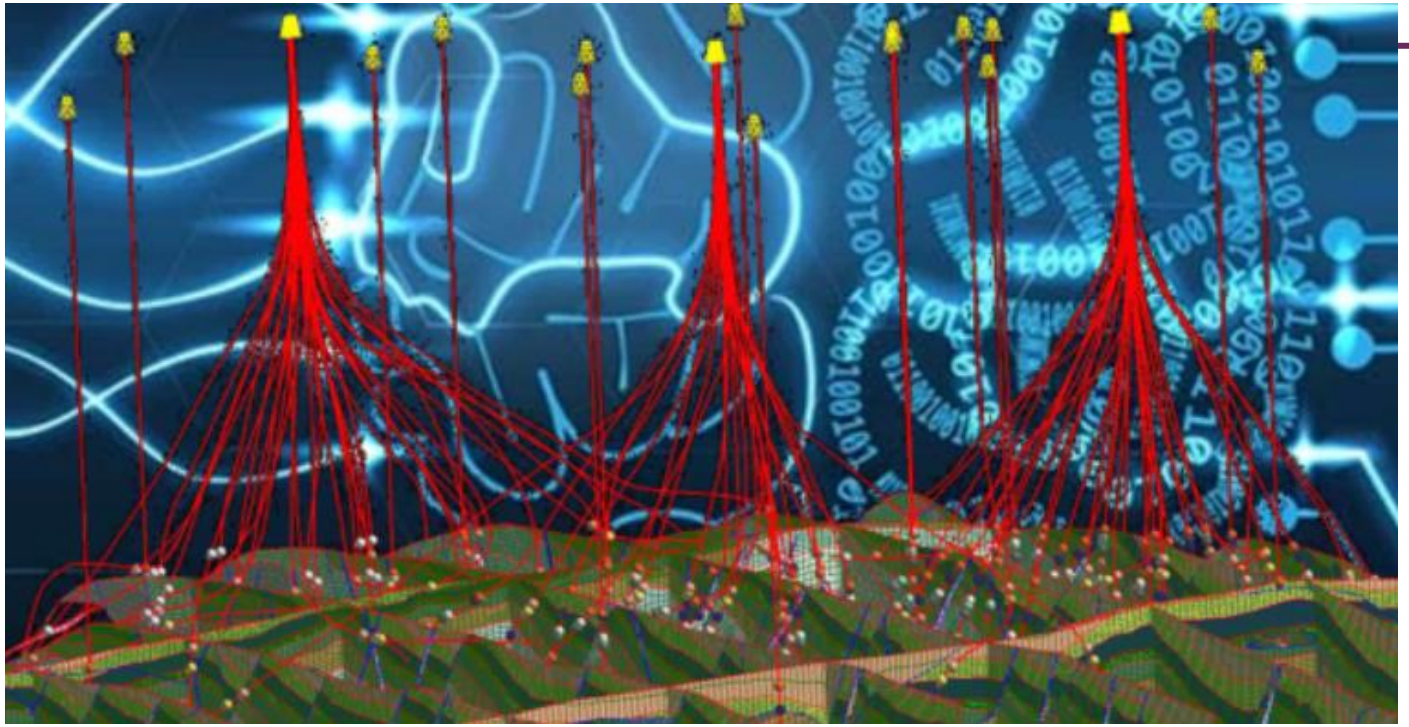
Covenant University

Raising a new Generation of Leaders

PDA_SIG Learn

A PEEP INTO DATA ANALYTICS AND MACHINE LEARNING

A PEEP INTO DATA ANALYTICS AND MACHINE LEARNING



Olatunde O. Mosobalaje (PhD)

Department of Petroleum Engineering,
Covenant University, Ota
Nigeria

OUTLINE

- **Data Mining: Data Analytics and Machine Learning**
  - Data Age
  - Data versus Information
  - Data Analytics
  - Machine Learning
  - Knowledge Domains
  - Oil and Gas Use Cases
- **Machine Learning: Taxonomy and Terminology**
  - Supervised Machine Learning
  - Unsupervised Machine Learning
  - Data Terminology
- **Machine Learning Workflow**
  - Feature Extraction
  - Data Splitting
  - Model Training
  - Model Evaluation

Data Mining: Data Analytics and Machine Learning

- It is the age of data. Data is everywhere!!!

Stone age

Iron age

Industrial age

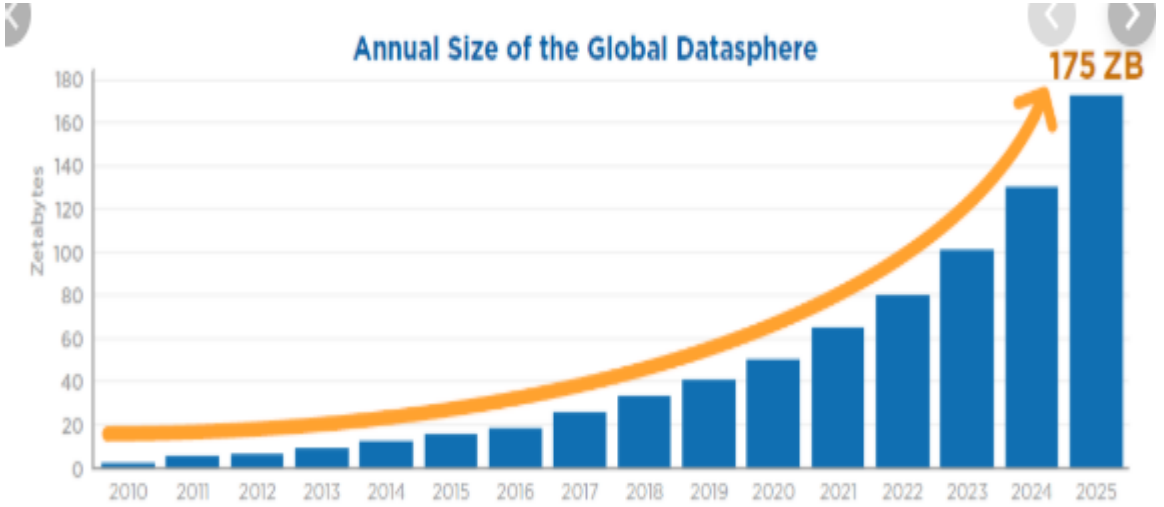
Electronic age

Data age

- Technology makes it possible to capture and store vast quantities of data

- Amount of data is growing. 2.5 billion gigabytes of data generated everyday in 2012.*

* https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/



Source: IDC Data Age 2025 Whitepaper

Data Mining: Data Analytics and Machine Learning

Data versus Information

Data: recorded facts.

Information: patterns, trends, insights and relationships that underlie data

- Information is what is needed for decision-making and problem-solving processes.
- The process of extracting information from data is generally known as Data Mining (DM). The required body of knowledge is Data Science (DS).
- Data Analytics (DA), Machine Learning (ML) and Artificial Intelligence (AI) are all parts of Data Science

Data Mining: Data Analytics and Machine Learning

Data Analytics

- Data Analytics: a set of tasks performed on data with the aid of specialized systems and softwares in order to describe or infer the information contained in such data.

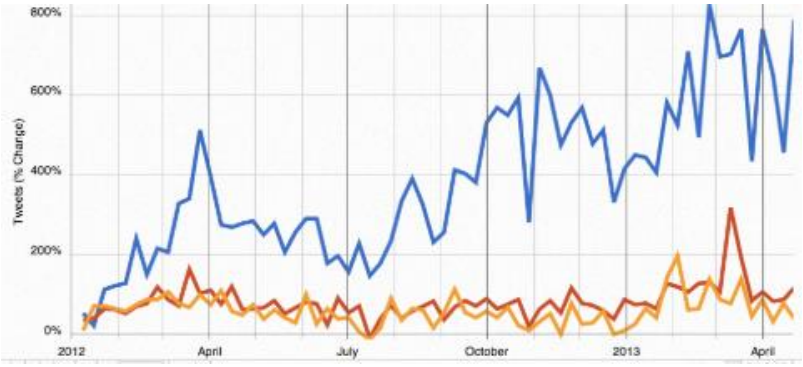
- DA tasks include:
  - Data collection – identifying sources, subsetting, assembling
  - Data integration – combining data from different sources into a common format
  - Data preparation – manipulating and organizing data to conform it to analytics requirements
  - Data cleaning – fixing quality problems: errors, suspicious, omitted and duplicated data
  - Data modeling – fitting data into conventional models: linear regression etc.
  - Data visualization – presenting data with charts, graphs to aid information mining
  - Data interpretation

Data Mining: Data Analytics and Machine Learning

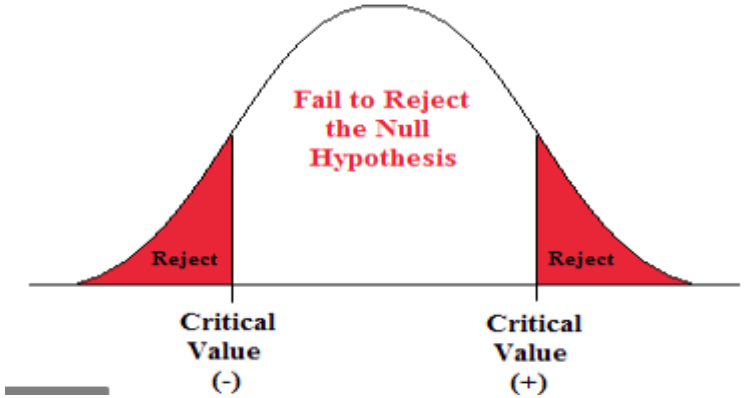
Data analytics

Exploratory or confirmatory

- Exploratory Data Analysis (EDA): finds patterns and relationship – more like the work of a crime investigator*



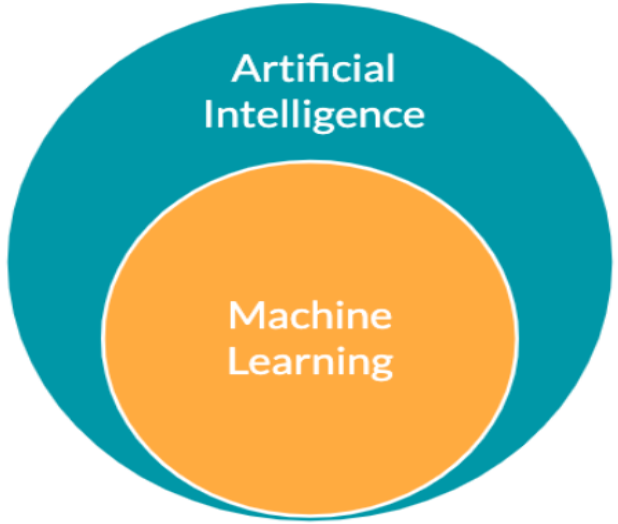
- Confirmatory Data Analysis (CDA): determines if hypothesis are true or false – more like the work of a judge*



* Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Data Mining: Data Analytics and Machine Learning

Machine Learning



- Machine learning refers to techniques by which machines (computers) are made to analyze data and recognize (learn) patterns and trends in data without relying on standard rule-based programming practices.

- Core of ML: making machines perform tasks based on past experiences passed to the machines as training data

- ML could handle larger and more complicated data and could find patterns more quickly than conventional data analytics tools.

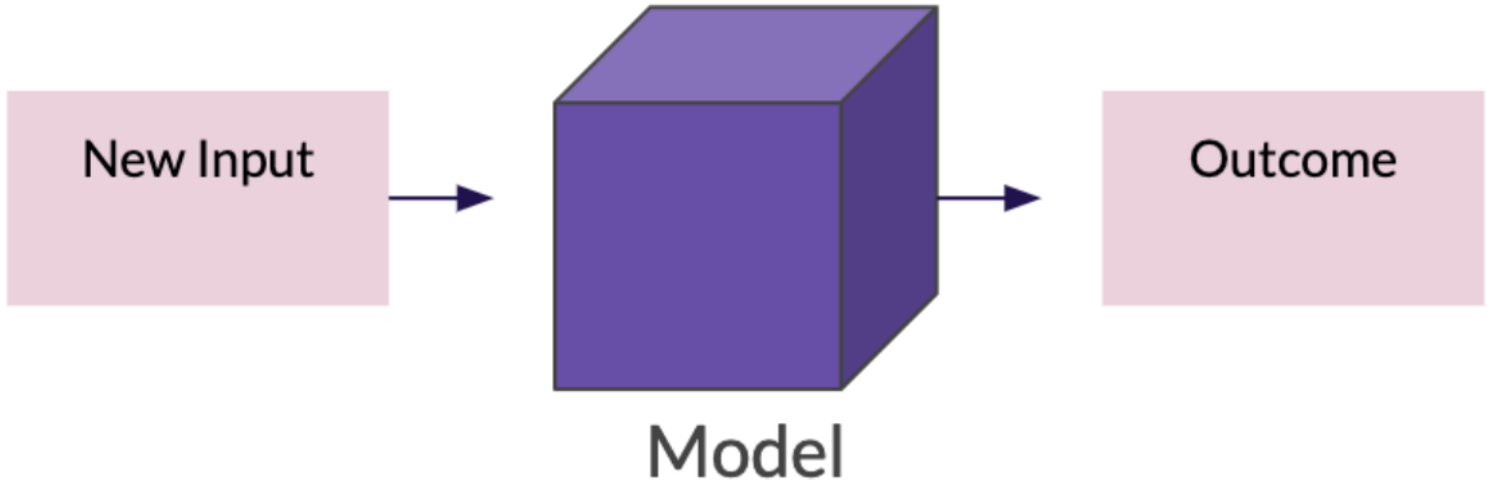
- ML requires less human effort and less assumptions .

- A subfield of computer science and AI

Data Mining: Data Analytics and Machine Learning

Machine Learning

- Machine learning: a data-driven model representation of a physical process



Data Mining: Data Analytics and Machine Learning

Machine Learning

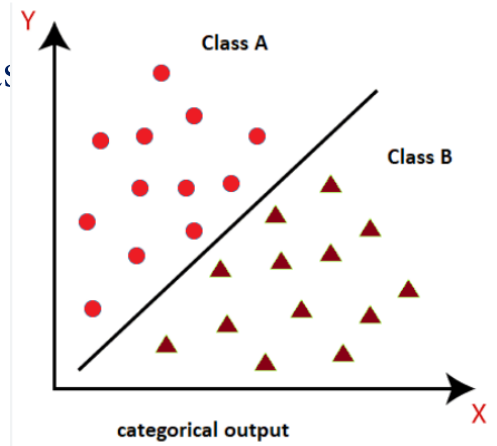
- Machine learning: learning information and insights from data without explicit equations – i.e. learning from experience
  - The 'learning' gets better as the number of data points increases – like humans

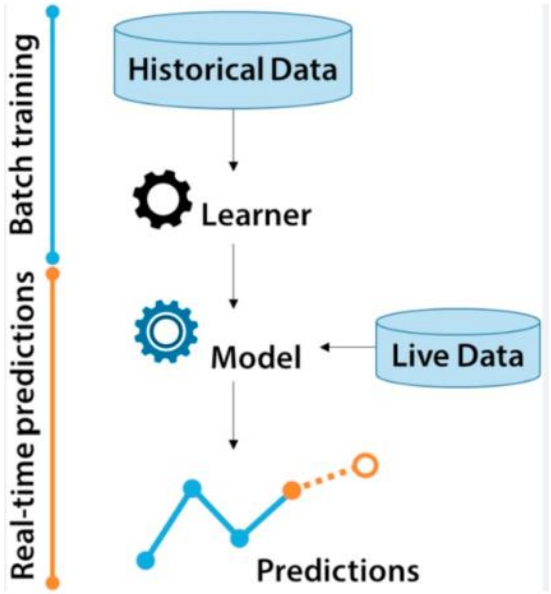
Data Mining: Data Analytics and Machine Learning

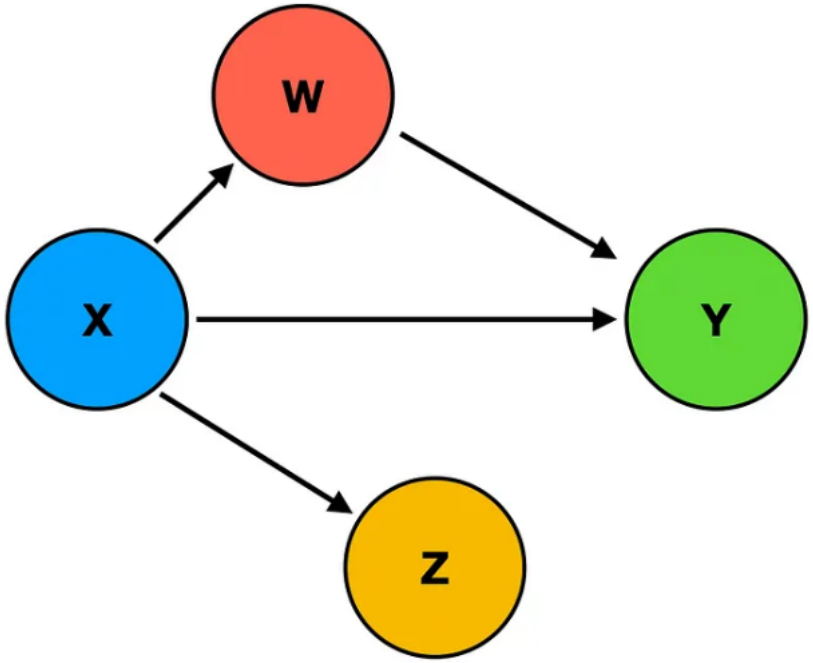
Machine Learning

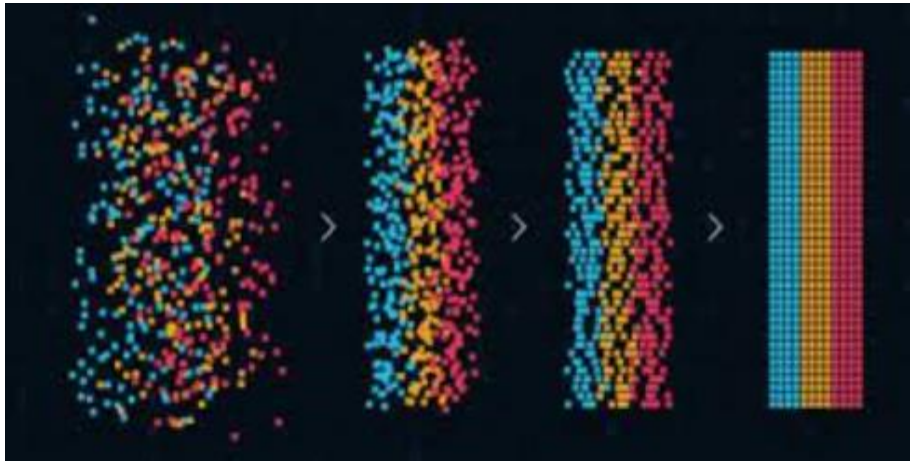
What can machine learning algorithm do?

- Predict the future
- Infer causes
- Patterns recognition
- Clas





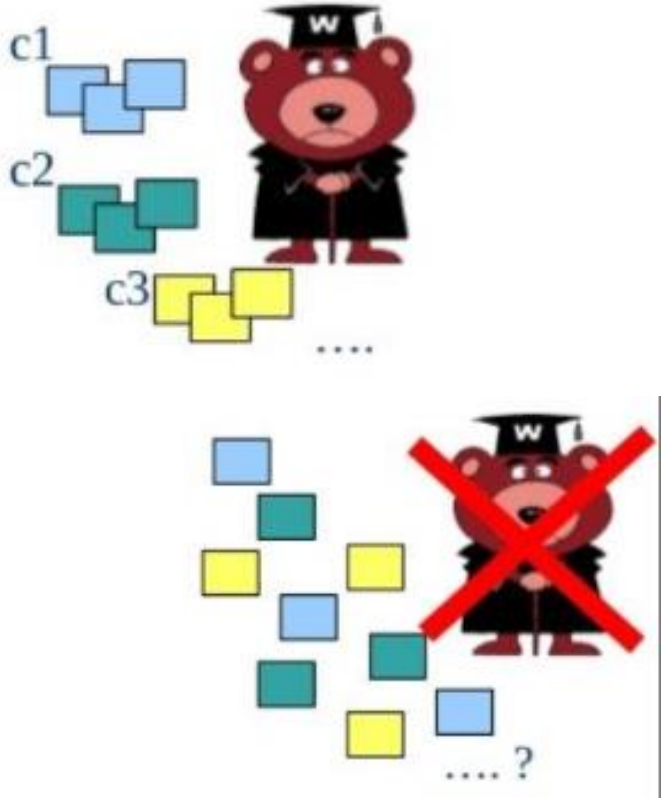




Data Mining: Data Analytics and Machine Learning

Machine Learning

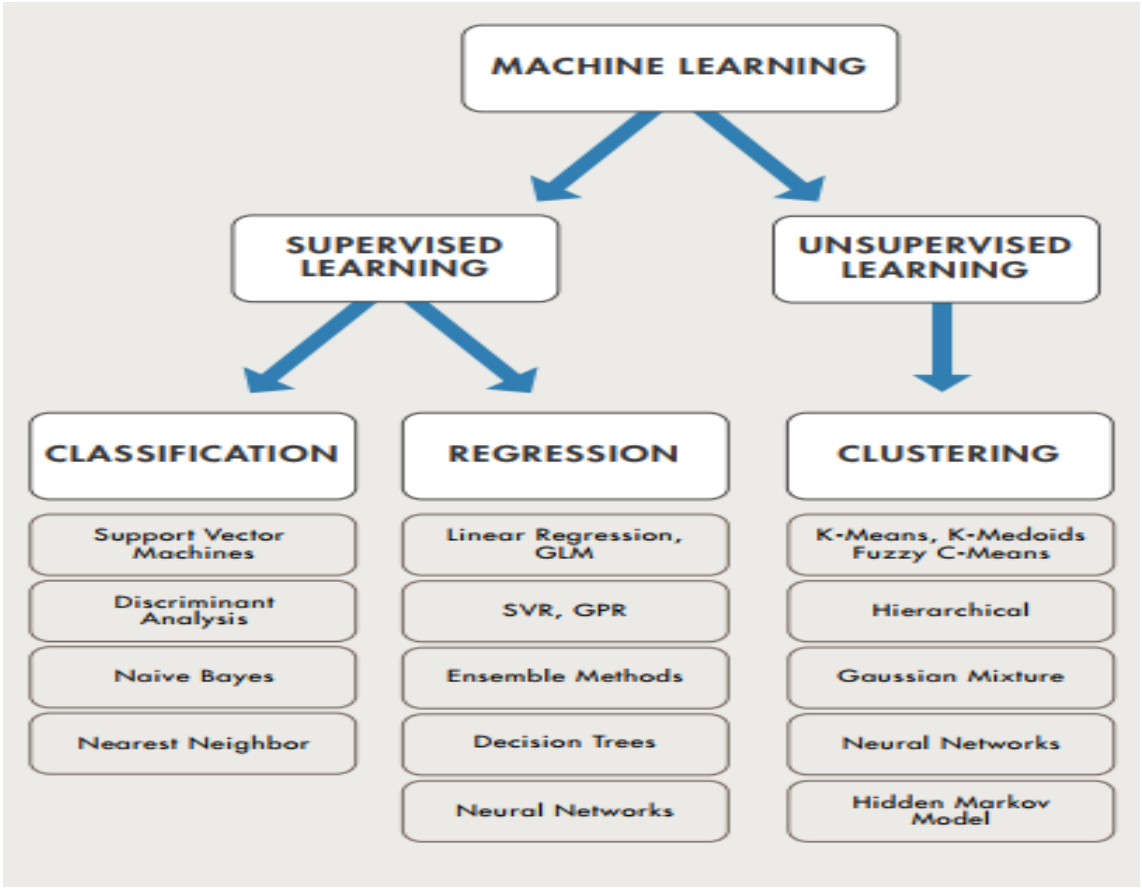
- Two major approaches to machine learning:
  - Supervised learning – train with labeled (input/output) data – to predict output for new inputs. Examples – regression, classification.
  - Unsupervised learning – train with unlabeled data – to detect patterns. Examples – clustering analysis.



Data Mining: Data Analytics and Machine Learning

Machine Learning

- Some machine learning algorithms:



Data Mining: Data Analytics and Machine Learning

Knowledge Domains

Certain domains of knowledge are necessary to attain competence in DA & ML

Mathematics and Statistics

- Matrix (linear) algebra: forms the basis of many ML techniques

- Multivariable calculus
- Data distributions, statistical estimators, hypothesis tests

Computer Programming

- Often, there will be the need to write codes to implement workflows for specific datasets and objectives.
- For automation and flexibility
- Statistical programming language like R and data querying language (SQL)

Data Visualization

- Turning numeric data to visual objects aids communication and decision making process.
- Understanding the principles of visual encoding of data.

Creativity, critical and intuitive thinking, and problem solving skills also needed

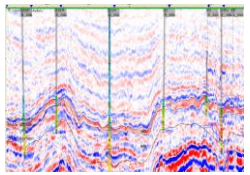
Data Mining: Data Analytics and Machine Learning

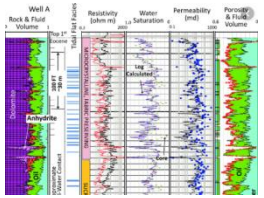
Oil and Gas Use Cases

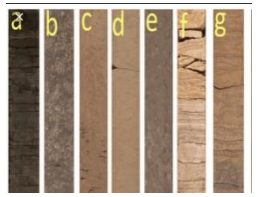
The oil and gas sector is prolific in data generation and measurements.

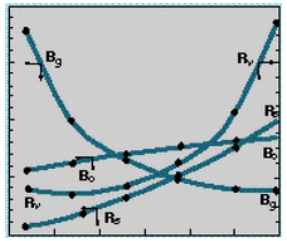
Data items include:

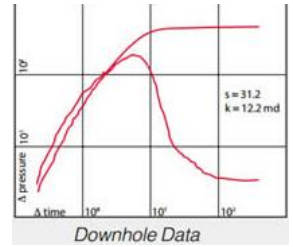
- Seismic surveys
- Well logs
- Core data
- Fluid data
- Pressure and temperature data
- Production test data
- Production and injection data – volumes, rates
- Drilling performance data
- Pipeline inspection data
- Equipment maintenance and failure data
- Crude trading data

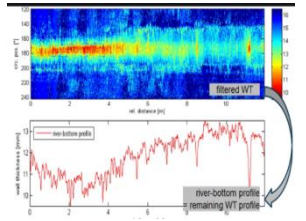


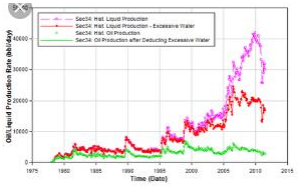








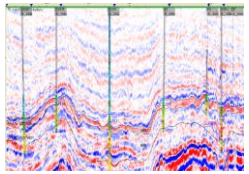


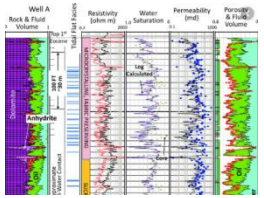


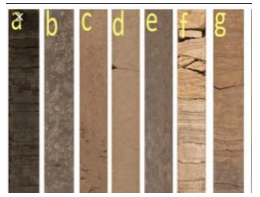
Data Mining: Data Analytics and Machine Learning

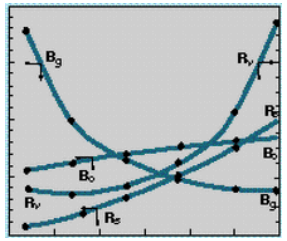
Oil and gas use cases

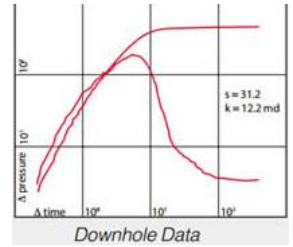
- Resource estimation – using ML to predict UR
- Subsurface modeling and analysis – using ML to find correlations among geological variables
- Wells analysis – historical and real-time well data analysis
- In-fill drilling – using AI to optimize number and locations of wells
- Well logging – data visualization
- Lithofacies classification using ML

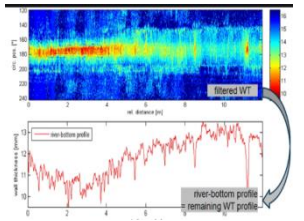


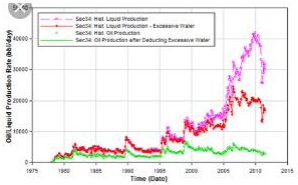








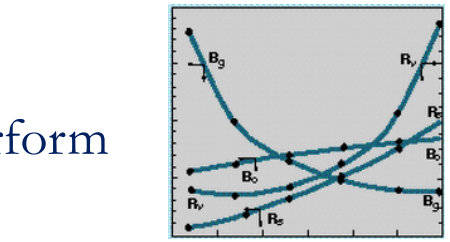


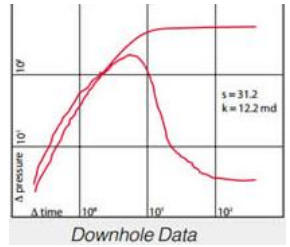


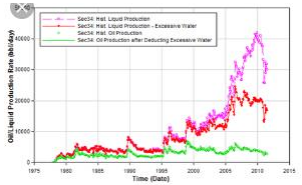
Data Mining: Data Analytics and Machine Learning

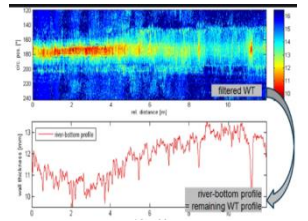
Oil and gas use cases

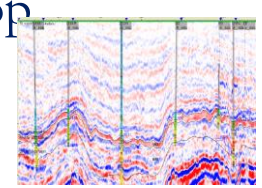
- PVT Fluid properties – using ML algorithms to develop correlations
- Investment portfolio management – using AI/ML to evaluate opportunities
- Health, Safety and Environment – using AI/ML to perform root cause analysis
- Inventory management – using AI/ML to predict demands, stock levels, warehouse utilization etc
- Finance – using AL/ML for cost allocations

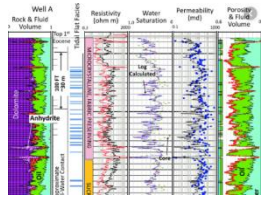


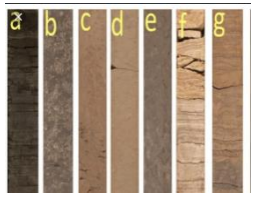








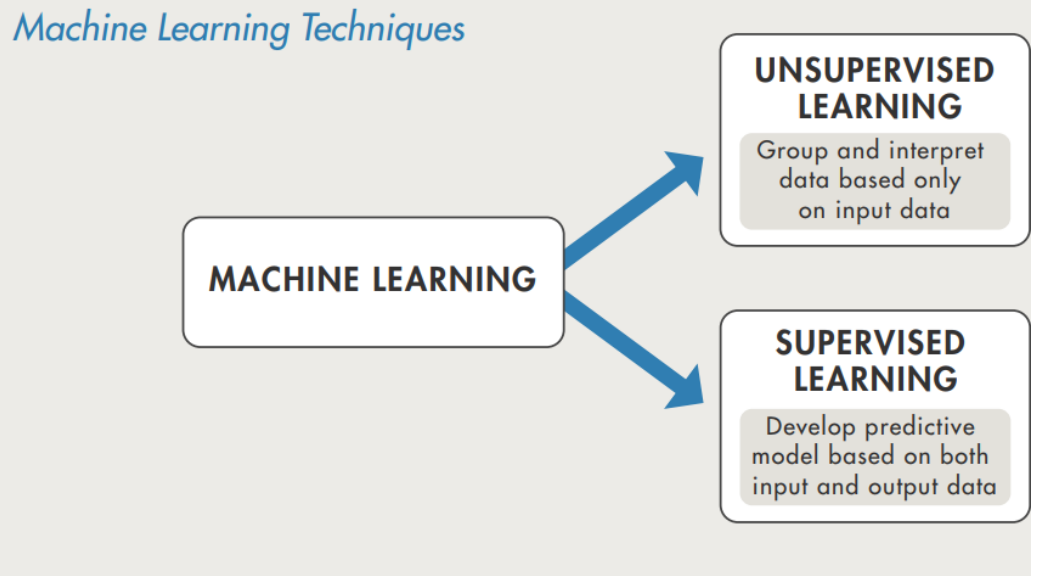


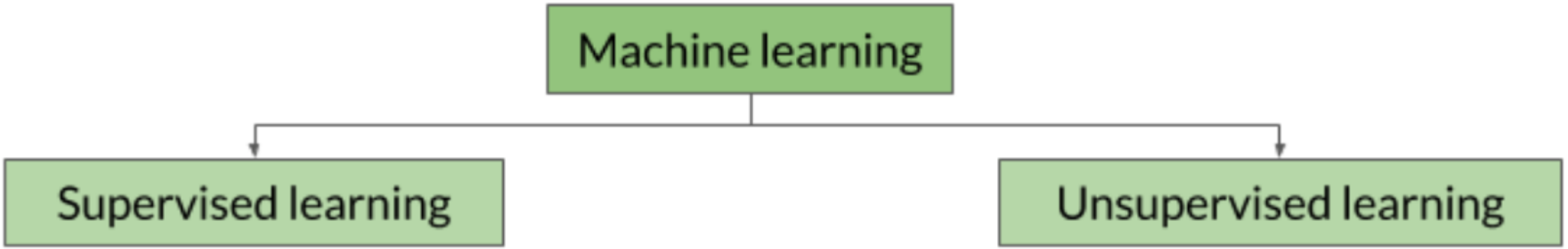


Machine Learning: Taxonomy & Terminology

"Taxonomy" – a scheme of classification

- Machine Learning is broadly classified into:
  - Supervised Learning
  - Unsupervised Learning





Machine Learning: Taxonomy & Terminology

Supervised Machine Learning

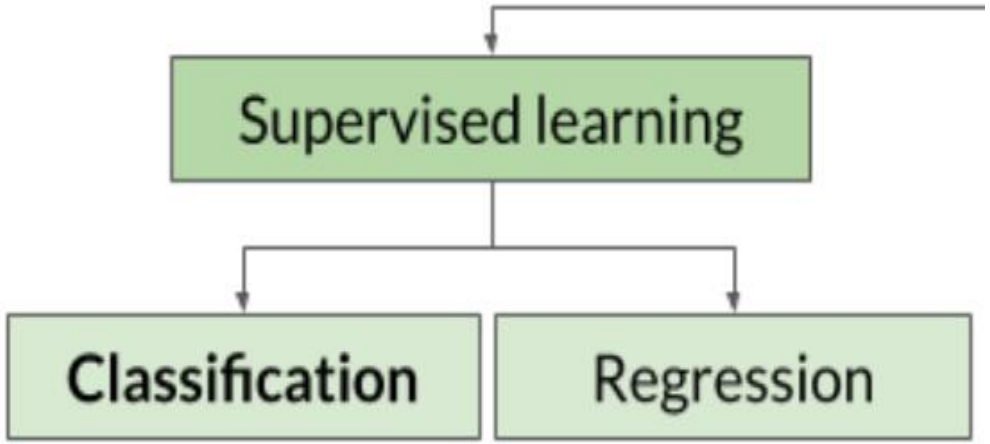
- Supervised Machine Learning:
  - Trains (make to learn) a model with a set of data containing several values of input(s) and output.
  - Predict out output value for a new input value
  - Examples:
    - Predicting Ultimate Recovery (output) from Sand-Water Ratio, GIP, Well Spacing, Proppant Loading, BTU content etc (input variables)
    - Predicting occurrence or non-occurrence of equipment failure (output) based on equipment performance data, weather and environmental conditions (input)

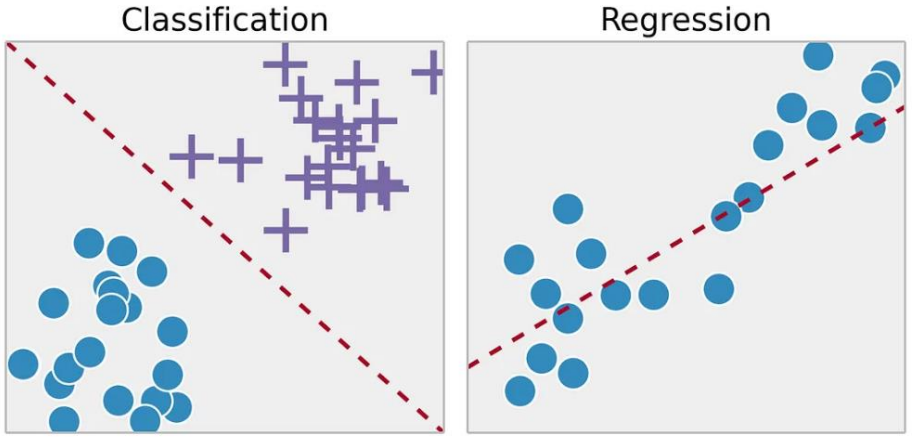
  - The learning is termed 'supervised' because known output values are available to compare with model's prediction

Machine Learning: Taxonomy & Terminology

Supervised Machine Learning

- Supervised machine learning is categorized into two types
  - Regression ML: when output to be predicted is a continuous variable, any value within an interval (e.g. porosity, oil viscosity)
  - Example:
    - Predicting permeability (output) from GR, neutron, density, resistivity and sonic logs (input variables)
  - Classification ML: when output to be created is categorical, one of a few discrete values (e.g. lithofacies)
    - Example:
      - Predicting occurrence or non-occurrence of equipment failure (output) based on equipment performance data, weather and environmental conditions (input)

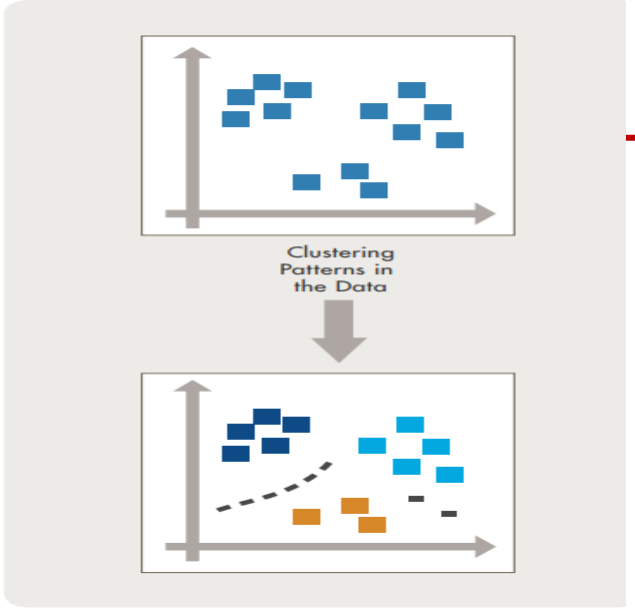


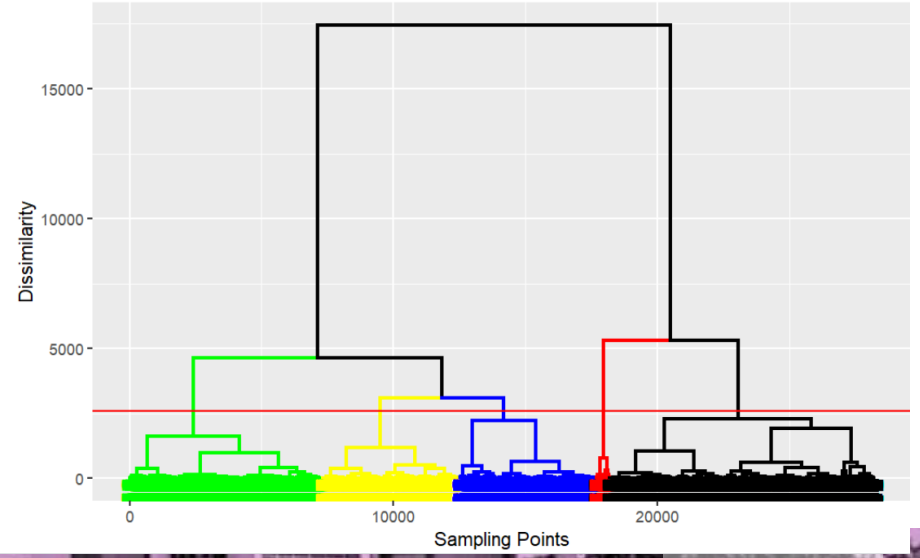


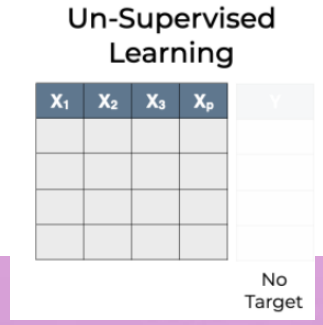
Machine Learning: Taxonomy & Terminology

Unsupervised Machine Learning

- Unsupervised machine learning:
  - Explores a dataset (input only) to uncover patterns such as clustering, correlation or association
  - No variable in the dataset is designated as output; hence, the term 'unsupervised'; model finds its own pattern.
  - Examples:
    - Clustering well log data (GR, NPHI, RDEEP, RHOB & RSHAL) to determine possible number of lithofacies
    - Clustering petrophysical properties to determine number of geological horizons or flow units in a formation



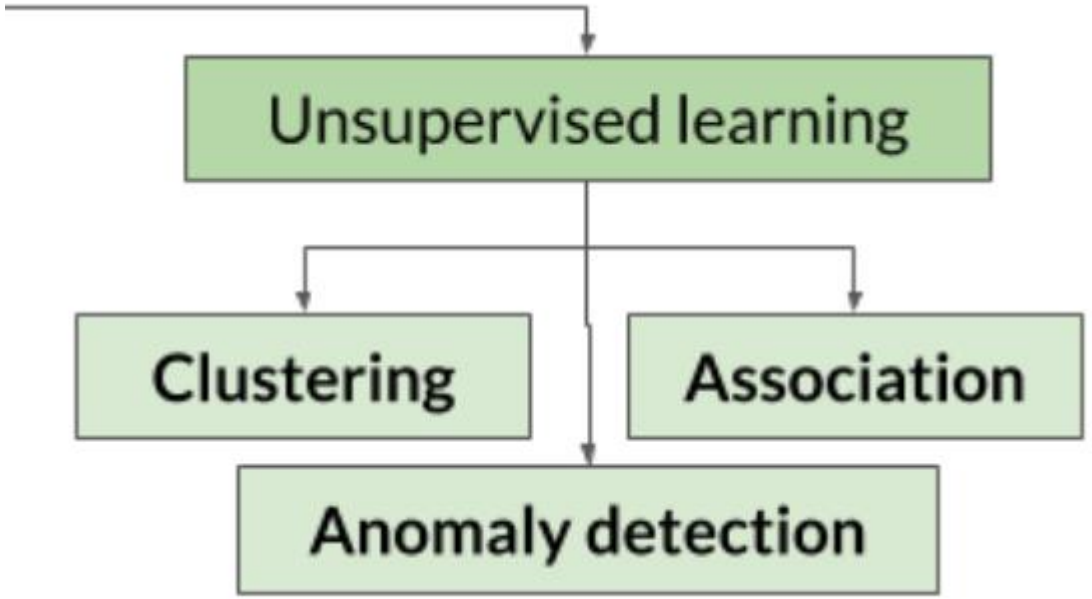




Machine Learning: Taxonomy & Terminology

Unsupervised Machine Learning

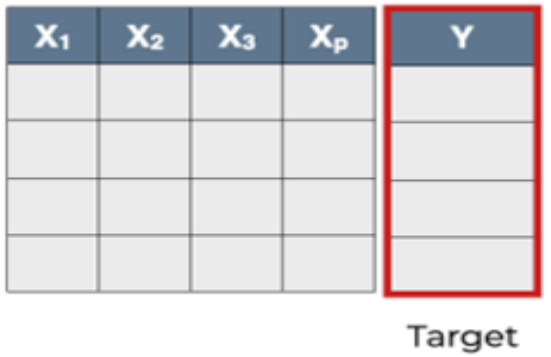
- Unsupervised machine learning may be categorized into:
  - Clustering
  - Association
  - Anomaly detection



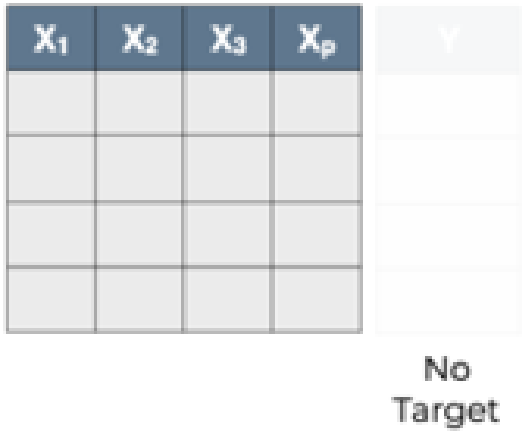
Machine Learning: Taxonomy & Terminology

Data Terminology

- Labelled versus Unlabelled
  - Labelled Data: a dataset containing values of input variable(s) and their corresponding output value (label, response, target)



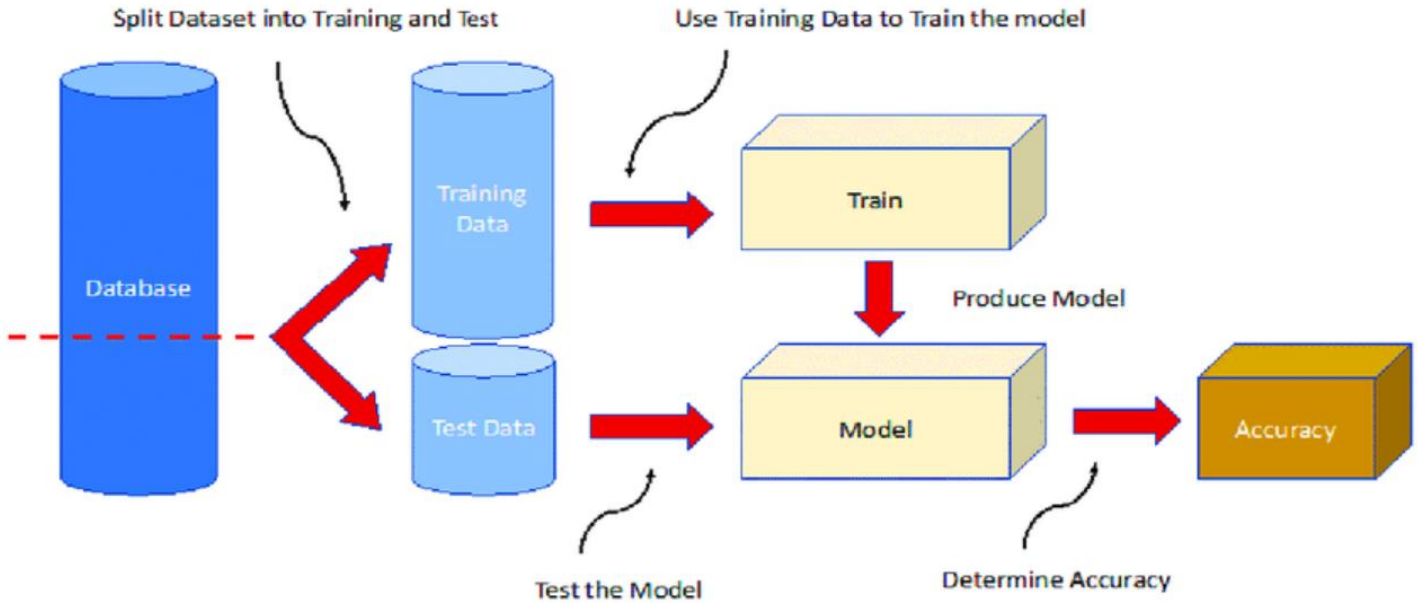
  - Un-labelled Data: a dataset with no designated output variablevariable in the dataset is designated as output



Machine Learning: Taxonomy & Terminology

Data Terminology

- Training Data versus Testing Data
  - Training Data: a portion of the dataset made available for training the ML algorithm to build a model
  - Testing Data: a portion of the dataset reserved (unused in training) for testing the accuracy of the model



- The process of allocating datapoints to either training or testing data is known as Data Splitting (Vertical)

Machine Learning: Taxonomy & Terminology

Data Terminology

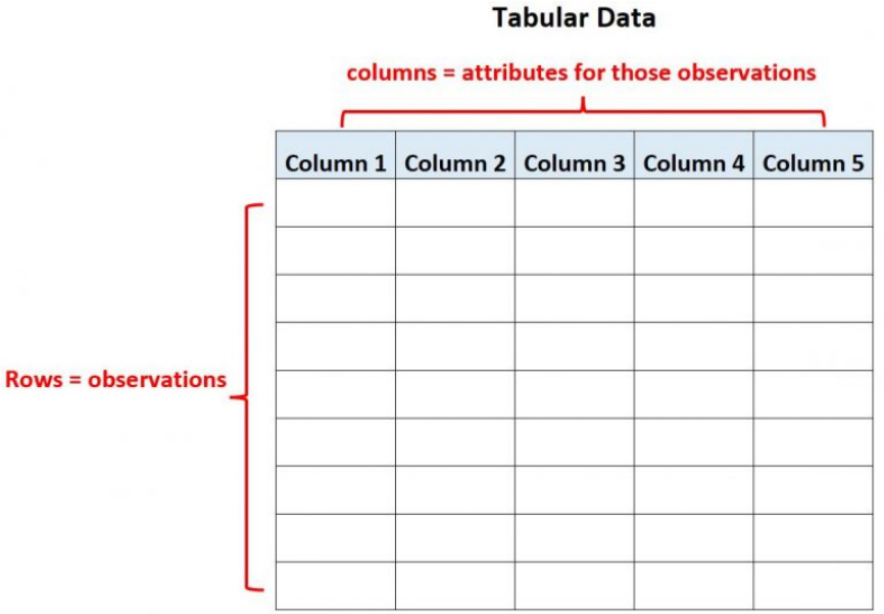
- Features versus Target
  - Feature Variables: variables that are to be used as inputs (a.k.a.: predictors, regressors, attributes, independent variables) in ML
  - Target Variable: the variable to be predicted (a.k.a.: response, output)
  - Of course, an un-labelled data (for unsupervised learning) only contains feature variables

	Features		Target

Machine Learning: Taxonomy & Terminology

Data Terminology

- **Features versus Target**
  - In tabular data:
    - Each column represents either a feature or the target variable
    - Each row represents an observation or sample
    - Collectively, the columns of features are known Feature Matrix while the target column is known as Response Vector.
- The process of selecting relevant feature variable(s) from a raw dataset containing several variables is known as Feature Extraction



Machine Learning: Taxonomy & Terminology

Data Terminology

- Training (Supervised Learning): both feature(s) and the target variables of the training data are fed into the ML algorithm, to train the model (i.e. get the algorithm to 'learn' the relationship between the features and the target variables)

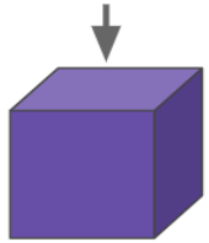
Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
50	F	196	0	False	non-anginal pain	98	False
53	F	215	0	True	asymptomatic	110	True

Machine Learning: Taxonomy & Terminology

Data Terminology

- Testing (Supervised Learning): only the feature(s) of the testing data are fed into the ML model to predict the target values which are compared with original target values

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
65	F	208	2	False	typical angina	105	???



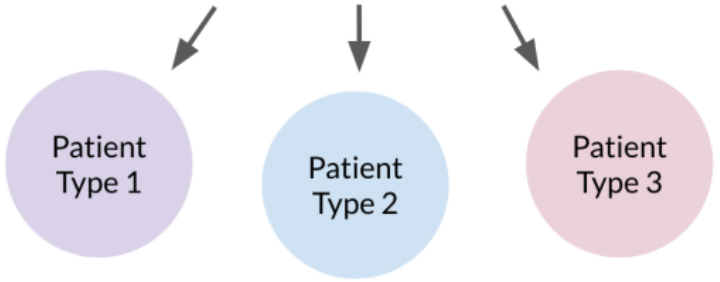
Heart disease
False

Machine Learning: Taxonomy & Terminology

Data Terminology

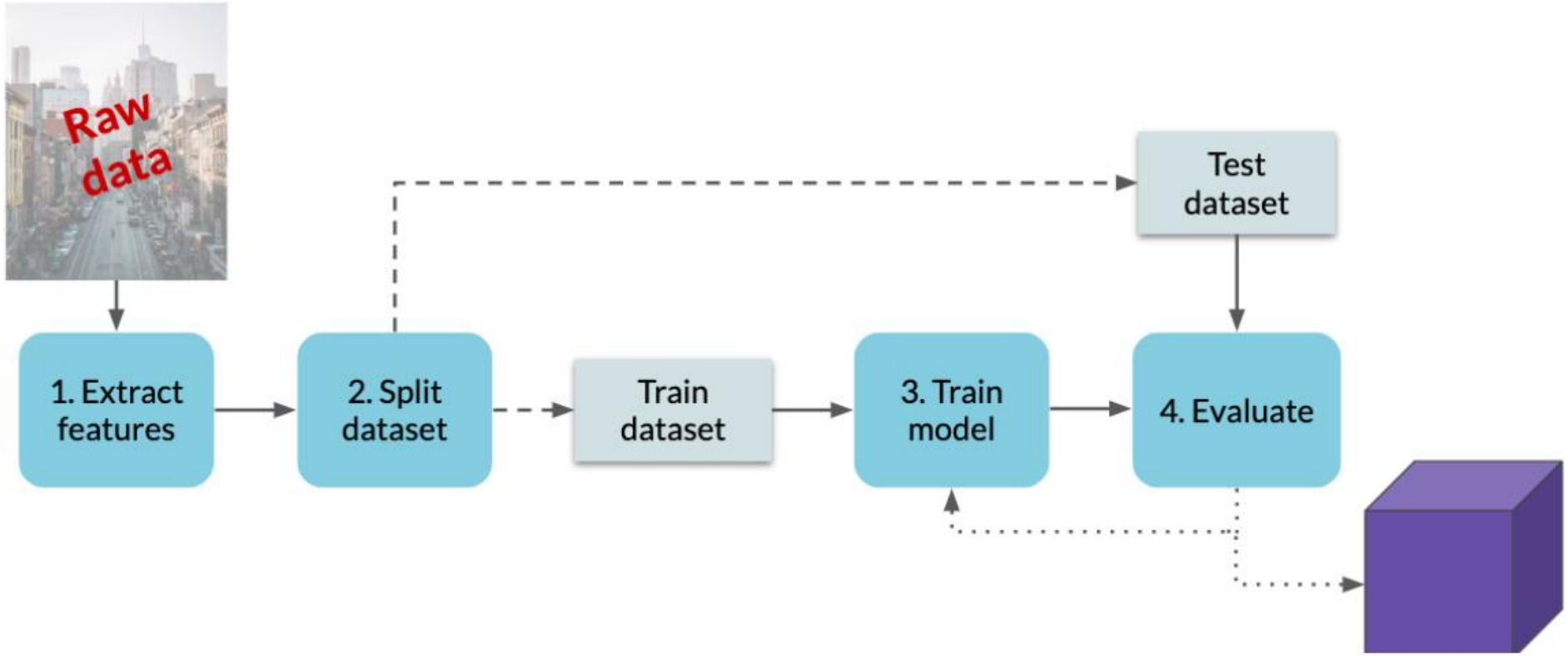
- Training (Unsupervised Learning): feature(s) of the training data are fed into the ML algorithm, to train the model

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
53	F	199	0	True	non-angin al pain	98	True
53	F	215	0	True	asymptoma tic	110	True
62	M	245	3	False	typical angina	126	True
...	...	...	...	...	...	...	...



Machine Learning Workflow

A typical machine learning project goes through the following procedures



Machine Learning Workflow

Feature Extraction

- Not every variable in the raw data may be necessary to be included in the ML project.
  - Reasons some variables may be dropped:
    - Irrelevance: no correlation with the target; no useful information added
    - Collinearity: strong correlation between two or more variables: i.e. they add same information to the model: e.g. amplitude and porosity
    - Low rank in Feature Ranking.
    - Dimensionality reduction (curse of dimensionality), to increase accuracy, explainability, and avoid over-fitting

Machine Learning Workflow

Feature Extraction

- Sometimes, some variables need to transformed before being used: e.g. pressure transformed to squared pressure or pseudo pressures, for gas reservoirs
- Some others need to be combined into a more informative variable: e.g. inlet pressure (P_i) and outlet pressure (P_o) combined as pressure difference ($\Delta P = P_o - P_o$)

Machine Learning Workflow

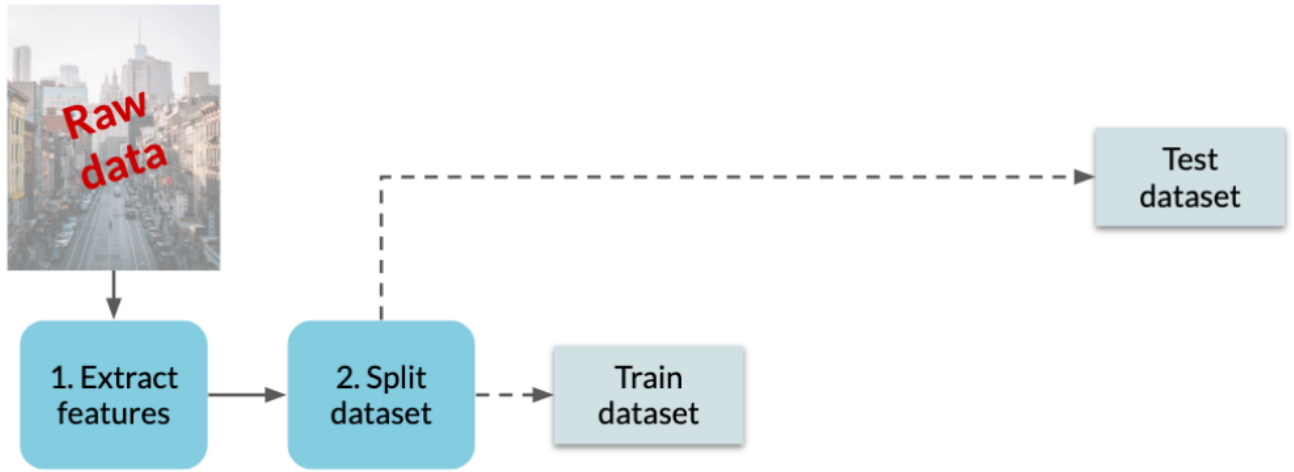
Feature Extraction

- Feature Extraction (or Feature Engineering) refers to all steps taken to determine and separate relevant features into a separate dataset in preparation for modelling.
- The popular Python DataFrame object is used to store such dataset
- Domain expertise is needed here.

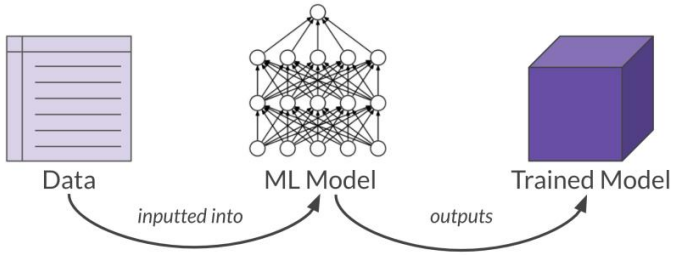
Machine Learning Workflow

Data Splitting

- This is the stage where the dataset is split into training dataset and testing dataset
- Typically, the train:test ratio is 80:20
- Sometimes, there is a third portion known as validation dataset.
- Python DataFrame's *.drop* method can be used to implement this splitting.

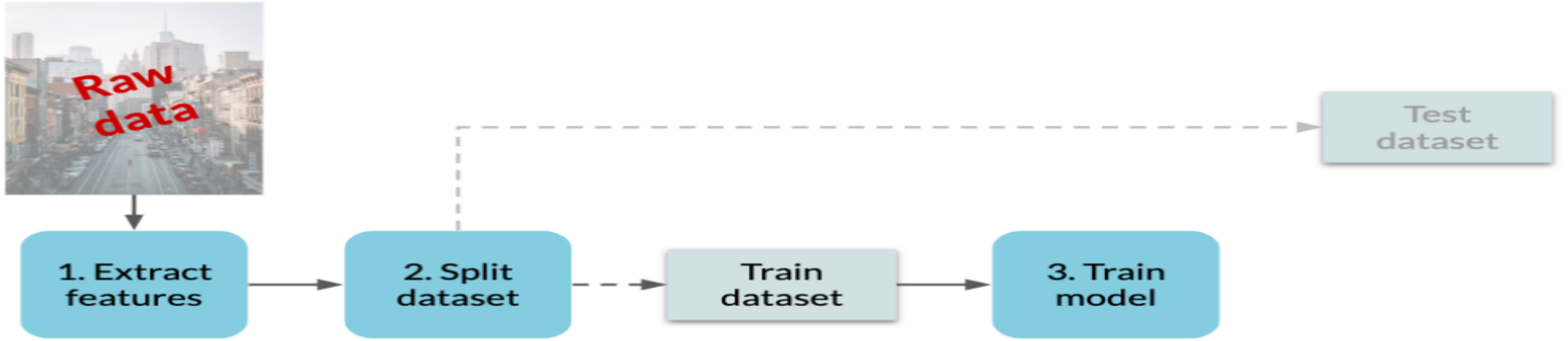


Machine Learning Workflow



Model Training

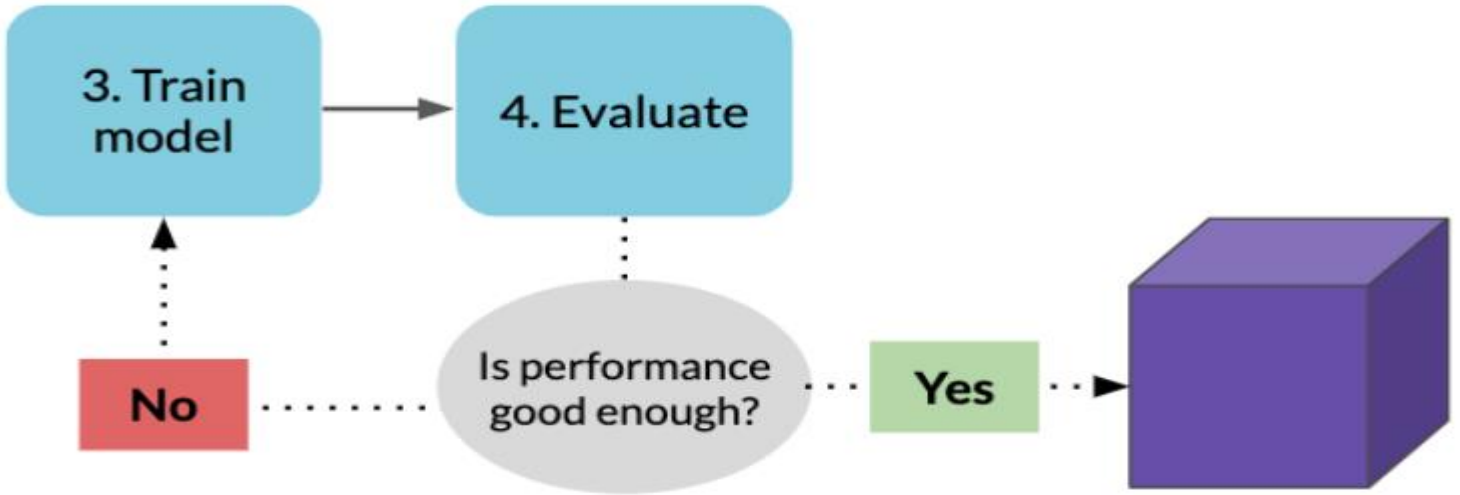
- At this stage, a choice is made of what ML algorithm is to be used for the model.
- Various algorithms should be evaluated for performance to pick the best
- It s recommended to start with simpler algorithms
- Feed the training data into the algorithm, and the algorithm 'learns' the relationship between the features and the target variables.



Machine Learning Workflow

Model Evaluation

At this stage, the model is evaluated (tested) for accuracy, by passing the feature matrix of the testing data to it, and comparing its predicted response values to the original response values (target) in the test data.



Machine Learning Workflow

Putting it all together