ORACLE
Certified Professional

Java
Oracle Certified Professional Java Programmer

Microsoft Certified Professional

sohailimran@yahoo.com

BIG DATA

Sohail IMRAN سهيل عمران

ML intro

Ref: Web

Introduction

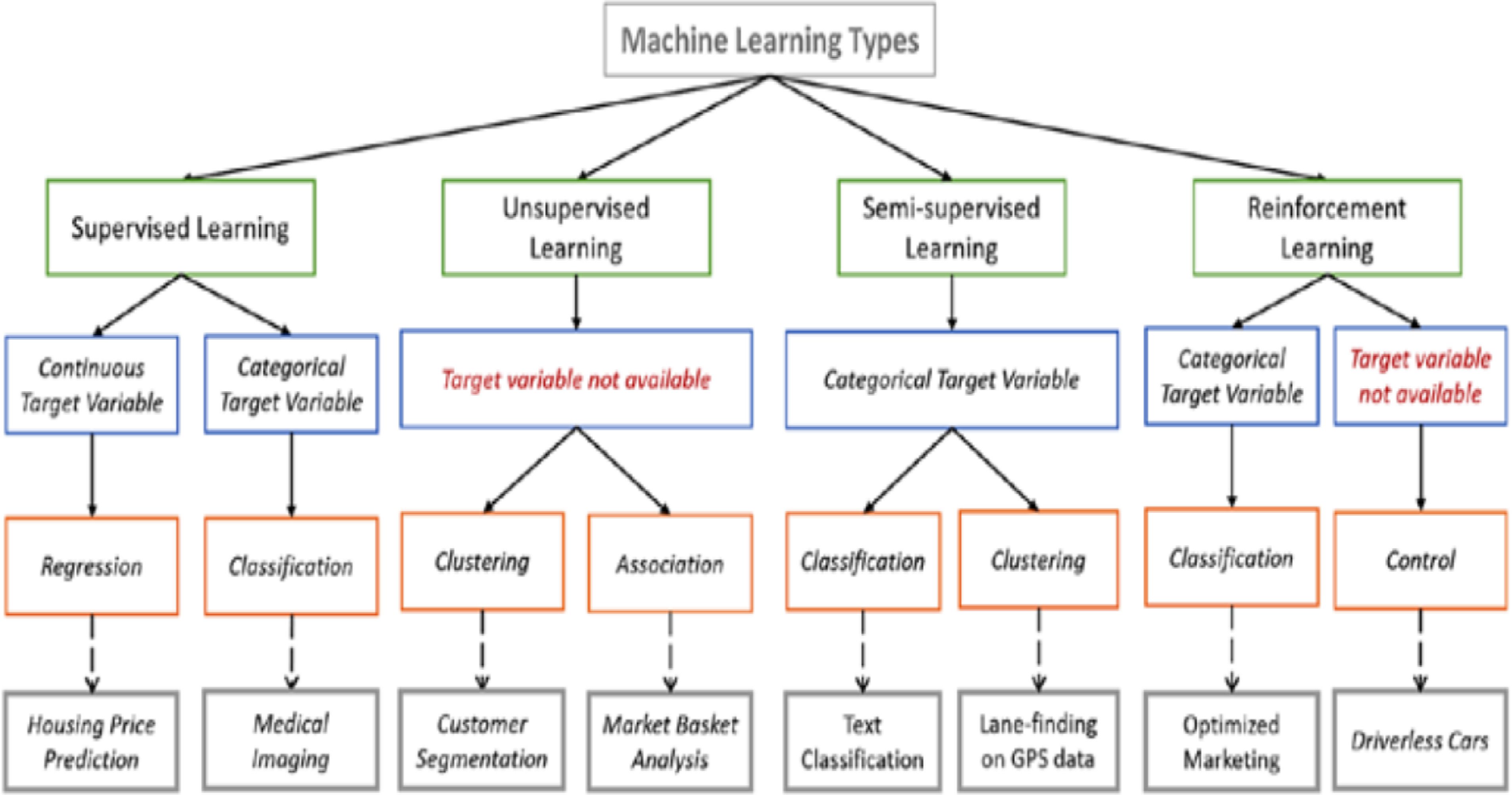
Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.
Machine learning algorithms use historical data as input to predict new output values.

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions.
There are four basic approaches:

- supervised learning,
- unsupervised learning,
- semi-supervised learning and
- reinforcement learning.

Supervised algorithms use labeled data in which both the input and output are provided to the algorithm.
Unsupervised algorithms do not have the outputs in advance. These algorithms are left to make sense of the data without labels.

Overview of ML Algorithms



Machine Learning Types

Supervised Learning

Unsupervised Learning

Semi-supervised Learning

Reinforcement Learning

Continuous Target Variable

Categorical Target Variable

Target variable not available

Categorical Target Variable

Categorical Target Variable

Target variable not available

Regression

Classification

Clustering

Association

Classification

Clustering

Classification

Control

Housing Price Prediction

Medical Imaging

Customer Segmentation

Market Basket Analysis

Text Classification

Lane-finding on GPS data

Optimized Marketing

Driverless Cars

Classification

Classification is a family of supervised machine learning algorithms that designate input as belonging to one of several pre-defined classes.

Classification data is labeled, for example, as spam/non-spam or fraud/non-fraud. Machine learning assigns a label or class to new data.

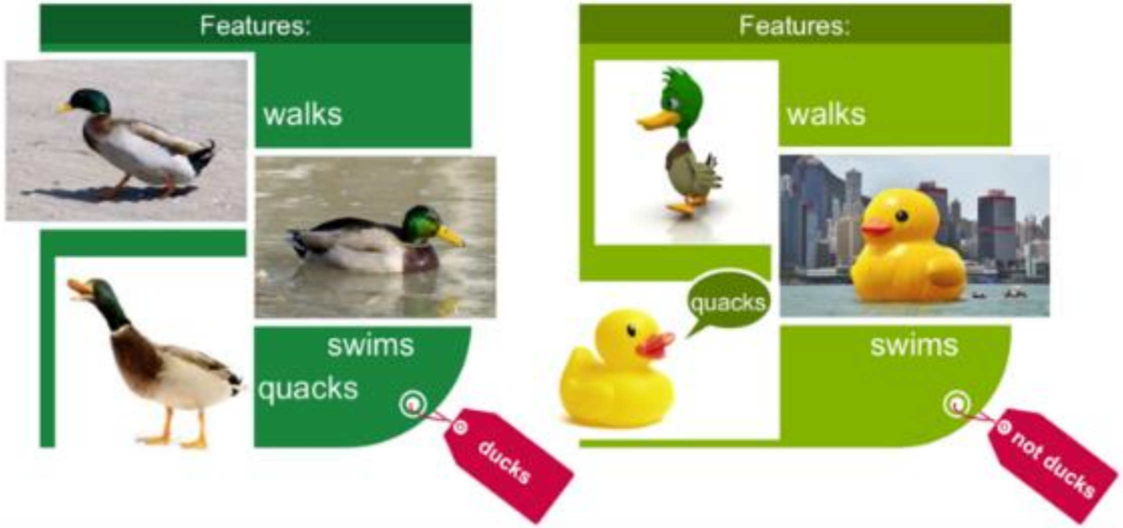
You classify something based on pre-determined features.

Features are the “if questions” that you ask.

The label is the answer to those questions.

In this example, if it walks, swims, and quacks like a duck, then the label is "duck ".

If it Walks/Swims/Quacks Like a Duck ...... Then It Must Be a Duck



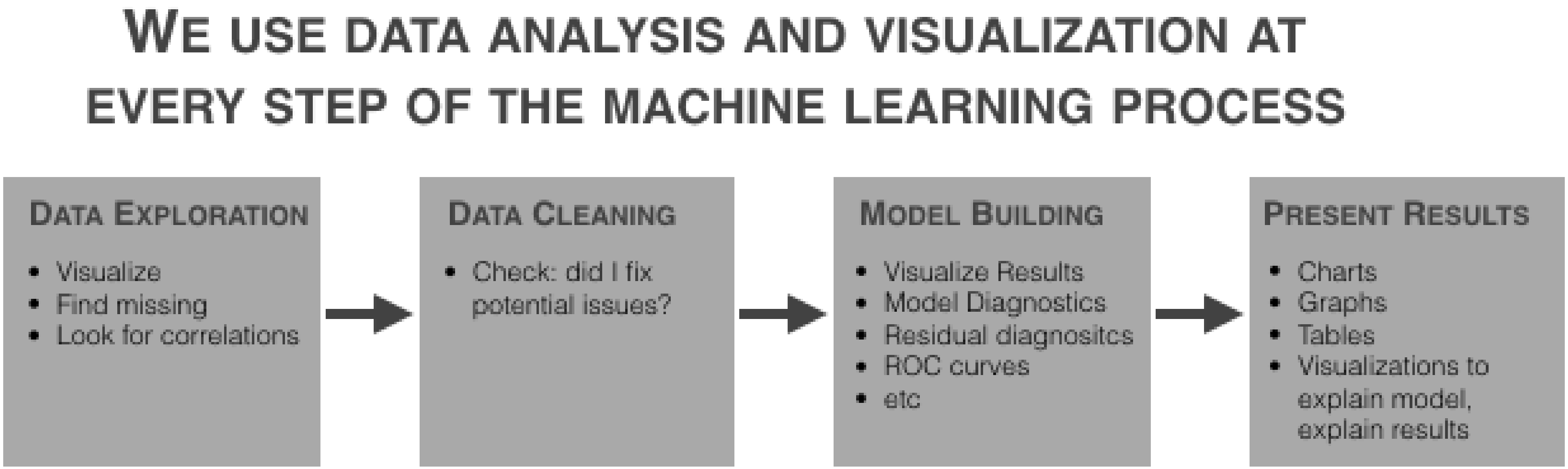
Clustering

In clustering, an algorithm groups objects into categories by analyzing similarities between input examples.

Clustering uses unsupervised algorithms, which do not have the outputs in advance.

Clustering applications include: Grouping of customers, Anomaly detection

Exploratory Data Analysis (EDA)



We use data analysis and visualization at every step of the machine learning process

Data Exploration

- Visualize
- Find missing
- Look for correlations

Data Cleaning

- Check: did I fix potential issues?

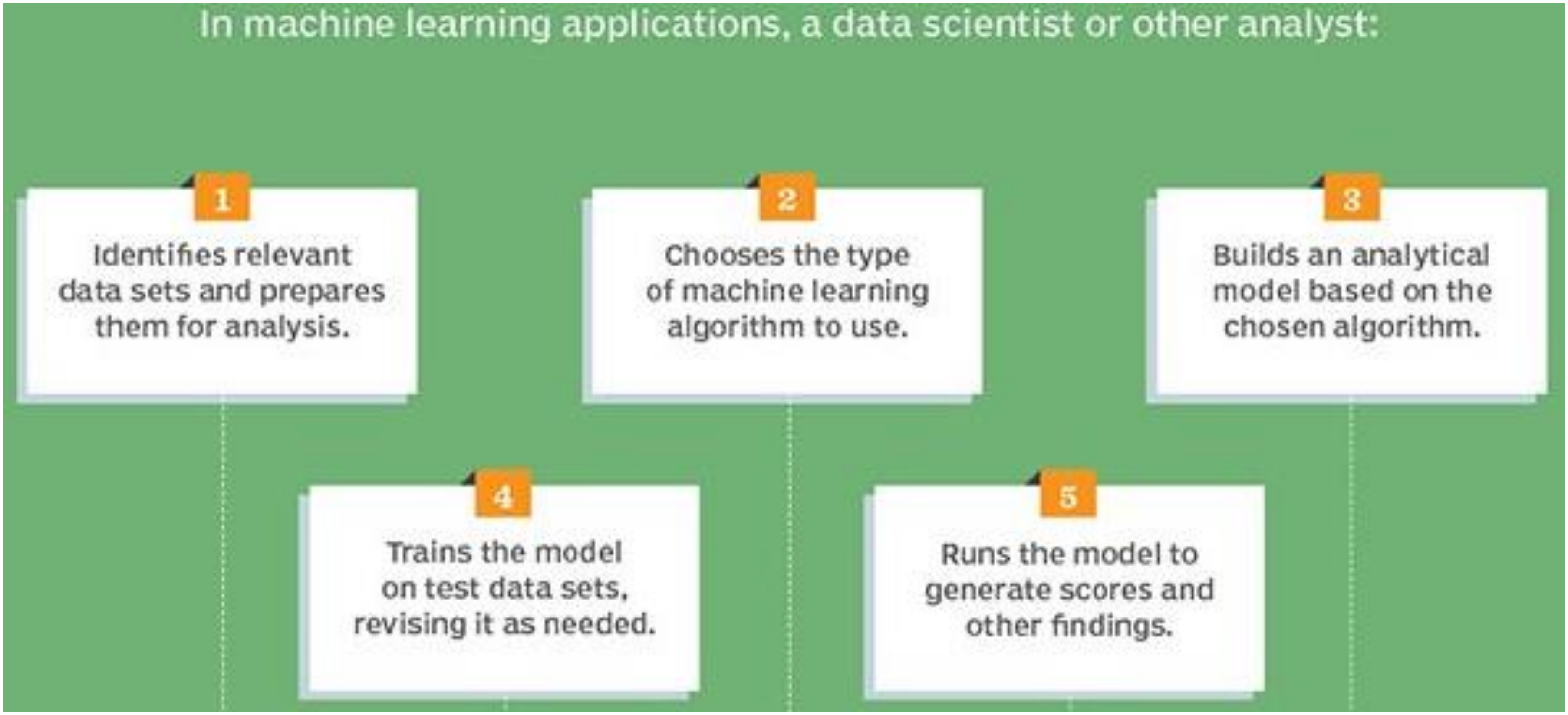
Model Building

- Visualize Results
- Model Diagnostics
- Residual diagnositcs
- ROC curves
- etc

Present Results

- Charts
- Graphs
- Tables
- Visualizations to explain model, explain results

Machine Learning Tasks



In machine learning applications, a data scientist or other analyst:

1. Identifies relevant data sets and prepares them for analysis.
2. Chooses the type of machine learning algorithm to use.
3. Builds an analytical model based on the chosen algorithm.
4. Trains the model on test data sets, revising it as needed.
5. Runs the model to generate scores and other findings.

Machine Learning Process (MLP)

1 - Data Collection

- The quantity & quality of your data dictate how accurate our model is
- The outcome of this step is generally a representation of data (Guo simplifies to specifying a table) which we will use for training
- Using pre-collected data.

2 - Data Preparation

- Wrangle data and prepare it for training
- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets

MLP (contd..)

3 - Choose a Model

- Different algorithms are for different tasks; choose the right one

4 - Train the Model

- The goal of training is to answer a question or make a prediction correctly as often as possible
- Linear regression example: algorithm would need to learn values for *m* (or *W*) and *b* (*x* is input, *y* is output)
- Each iteration of process is a training step

5 - Evaluate the Model

- Uses some metric or combination of metrics to "measure" objective performance of model
- Test the model against previously unseen data
- This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)
- Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.

MLP (contd..)

6 - Parameter Tuning

- This step refers to *hyperparameter* tuning.
- Tune model parameters for improved performance
- Simple model hyperparameters may include: number of training steps, learning rate, initialization values and distribution, etc.

7 - Make Predictions

- Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world