Fourth Industrial Summer School

Module 4:

Introduction to Machine Learning

Session Objectives

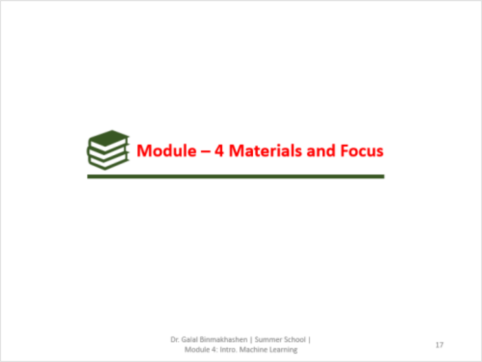
- ✓Introduction
  - ✓ Analytical Models
  - ✓ Predictive Analytics – ML
- ✓ML Unsupervised
  - ✓ Relation between PA/ML
  - ✓ What is Machine learning

Analytics Model

Dr. Galal Binmakhashen | Summer School | Module 4: Intro. Machine Learning 3

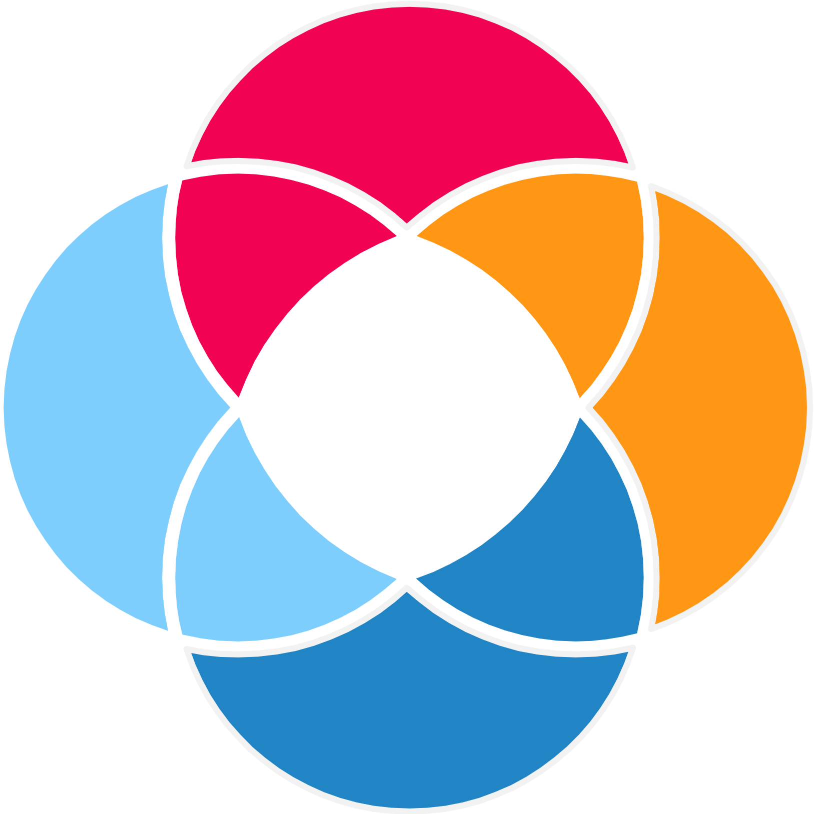
ML: Introduction

Dr. Galal Binmakhashen | Summer School | Module 4: Intro. Machine Learning 7



Module – 4 Materials and Focus

Dr. Galal Binmakhashen | Summer School | Module 4: Intro. Machine Learning 17



Analytics Model

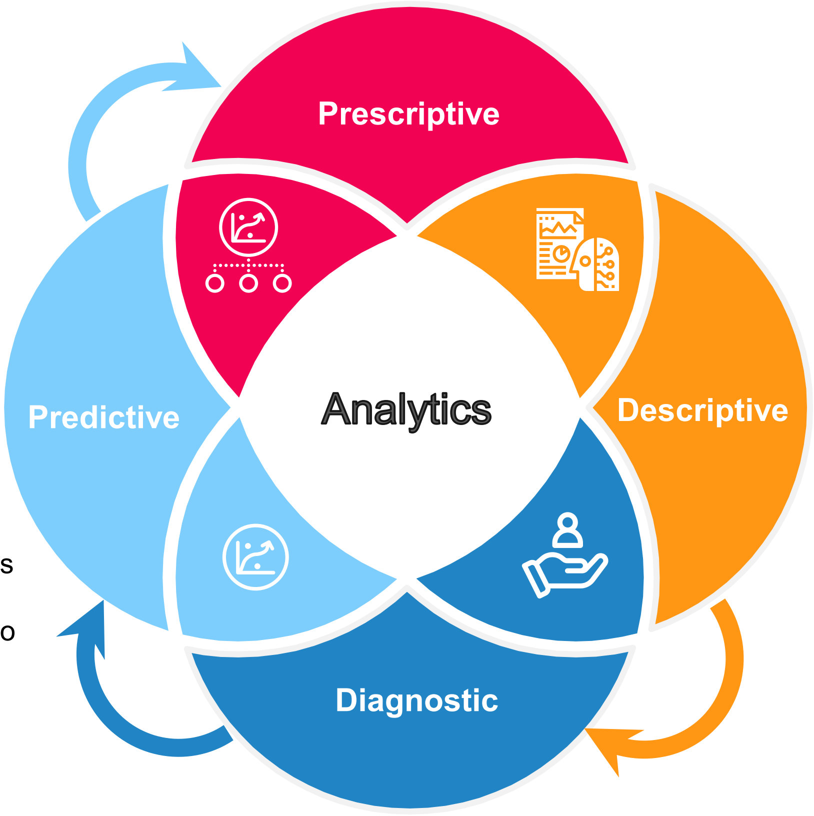
Analytical Models

4 Prescriptive

Based on our data (past knowledge) and possible future (predictive analytics), what decisions/actions can be made?
Tools: Python, R etc.

3 Predictive
Machine Learning

It involves beyond diagnostic analytics, automatically answering what will happen in the future. This includes building statistical and AI models that generalize and scale to our data!
Tools: Python, R etc.



1 Descriptive

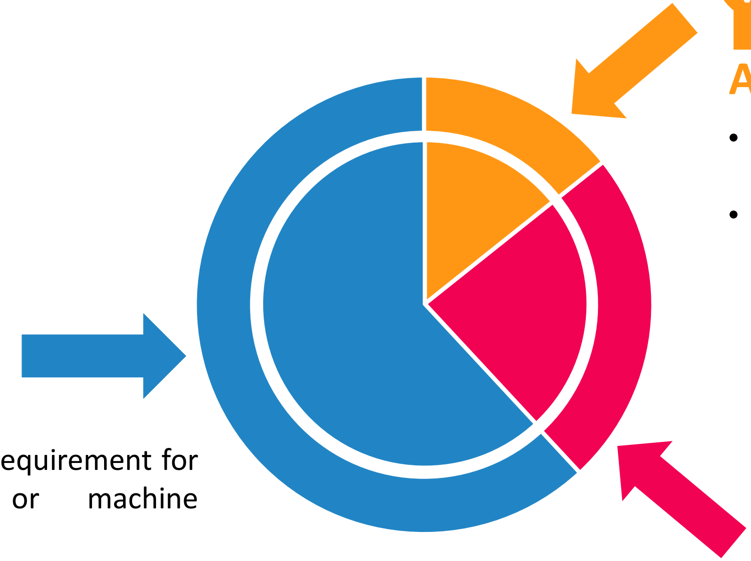
It involves gathering, cleaning, preparing summarizing, and visualizing the data.
Tools: SQL, R, Python, Excel etc.

2 Diagnostic

It involves relating, discovering causation, quantifying correlations
Tools: R, Python, Excel etc.

Predictive analytics is the practice of extracting insights from some existing data using statistical and machine-learning modeling.

PA: Components



Data

- Data quality is a crucial requirement for predictive analytics or machine learning, in general.
- Predictive modeling is strongly depending on the quality of the historical data (training).

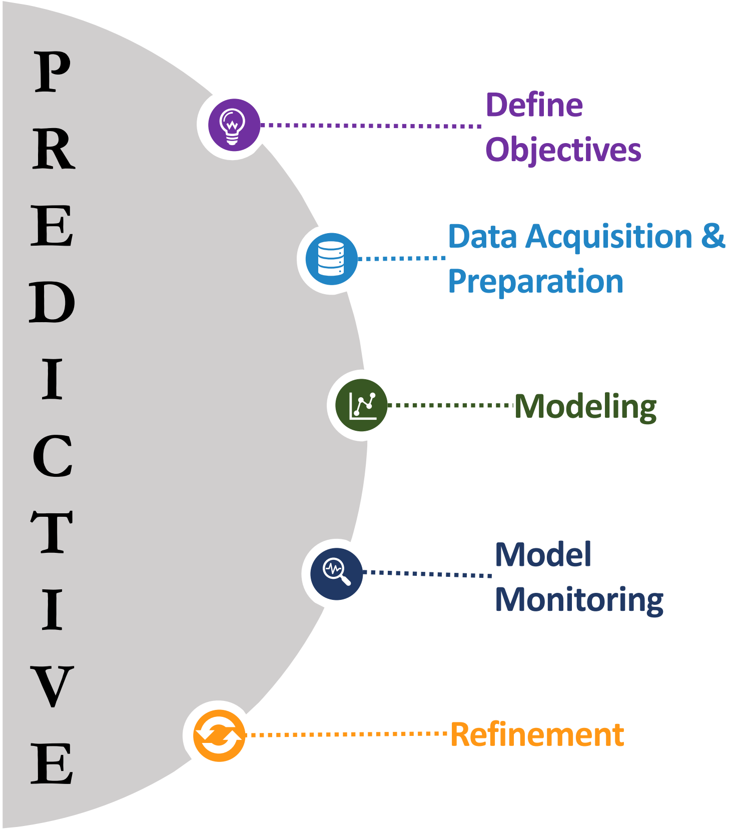
Assumptions

- The conclusions are **usually assuming the future will follow past patterns**!
- Assumptions may change by time, so in cases, we need to refine and adapt, e.g., customer behavior may change!!

Modeling

- It requires proper knowledge of each ML algorithm.
- Such knowledge qualifies users to select the right ML algorithm and possible ways to optimize them

PA: Life Cycle



Define Objectives

It is very important to have a predefined goal(s) at this stage. such as detecting fraud transaction in banking, etc.

Data Acquisition & Preparation

It includes 1) collecting the right data, 2) preprocess them, 3) clean them, and most importantly set the depended and independent variables (label/tag records).

Modeling

Assuming data is in good shape, building a model is a crucial step. The model and process pipeline should be generalized so that new (future) data can be predicted.

Model Monitoring

Beta AI models require monitoring to excel. So, if there is any unexpected outcomes should be traced back to data. In other words, redo the previous steps. Repeat until pass!

Refinement

The above two step usually repeat many times and re-evaluation is done at this stage.

ML: Introduction

Machine Learning: Definition

A science that develops algorithms to perform a specific task without explicit instructions but relying on ***patterns and inference instead****!*

Source Oxford Dictionary : https://www.lexico.com/en/definition/machine_learning

Example

- Suppose a fish packing plant wants **to automate** the process of sorting fish on a conveyor belt according to the species

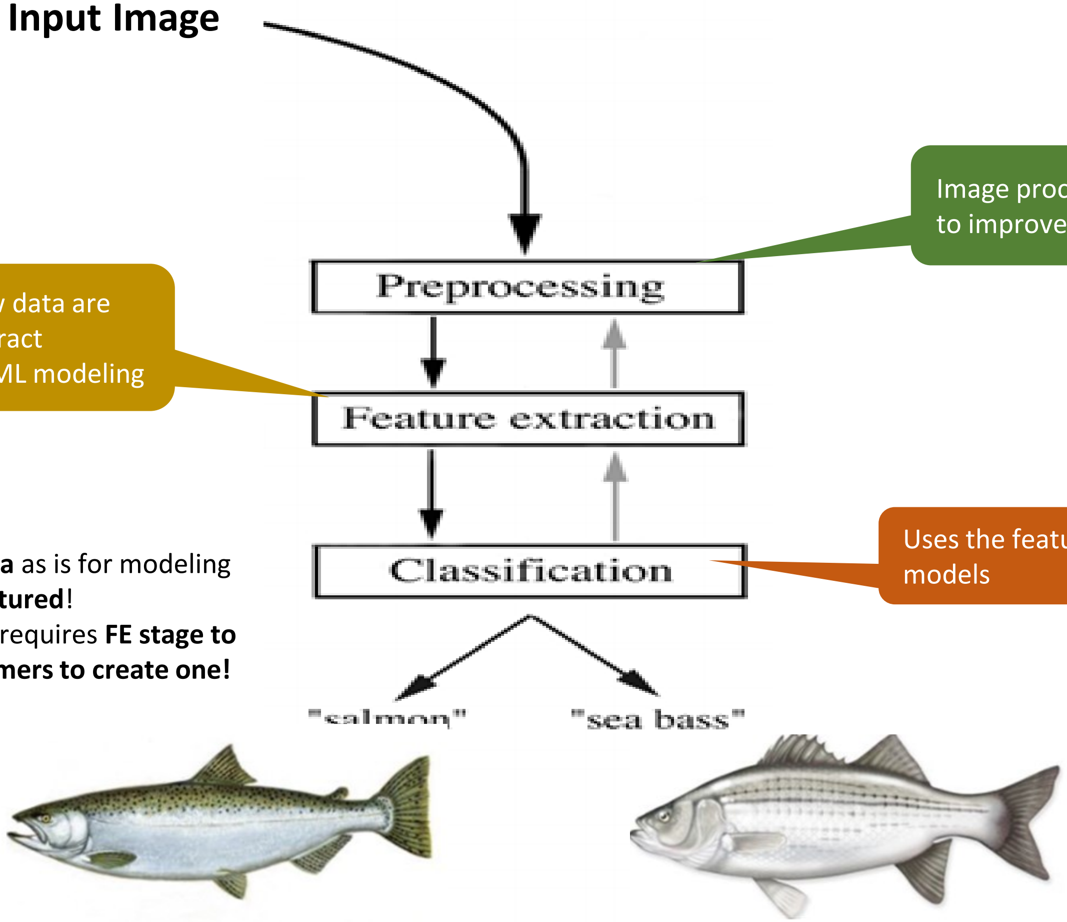


Image processing could be used to improve images before FE.

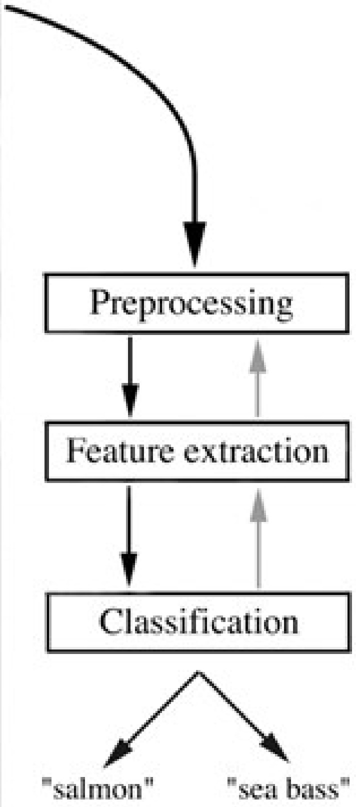
At this stage, raw data are processed to extract information for ML modeling

Uses the features to build ML models

- We may not use **raw data** as is for modeling since the data are **unstructured**!
- Usually, the ML pipeline requires **FE stage to make it easy for programmers to create one!**

Rule-based classification

Source: https://g.co/kgs/8jTDgF



Preprocessing

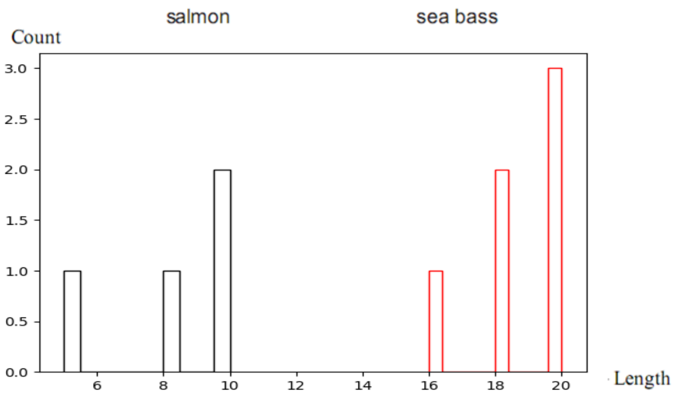
Feature extraction

Classification

"salmon"

"sea bass"

- We may develop some ***rules*** to classify these species as Salmon or Sea-bass.
- Let us start with one feature (variable); the **length** of the species.
- We sample some and draw the below figure,
- Then, any fish longer than l_1^*, classify it as **Sea-bass** otherwise **Salmon**



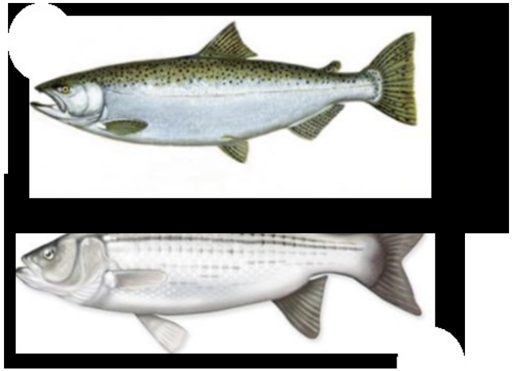
Will that work for future cases?
How about sampling more data!

Rule-based classification

For many reasons, the fish-length alone may not work well for all examples. It is not easy to determine the length threshold.

We may need to extract and include other factors (features) from the images such as:

- Width
- Number and shape of fins
- Position of fish mouth
- Etc.



The main issue of the hand-crafted rules, they depend on the programmer wrote them for classification, which may suffer a generalization issue to new data.

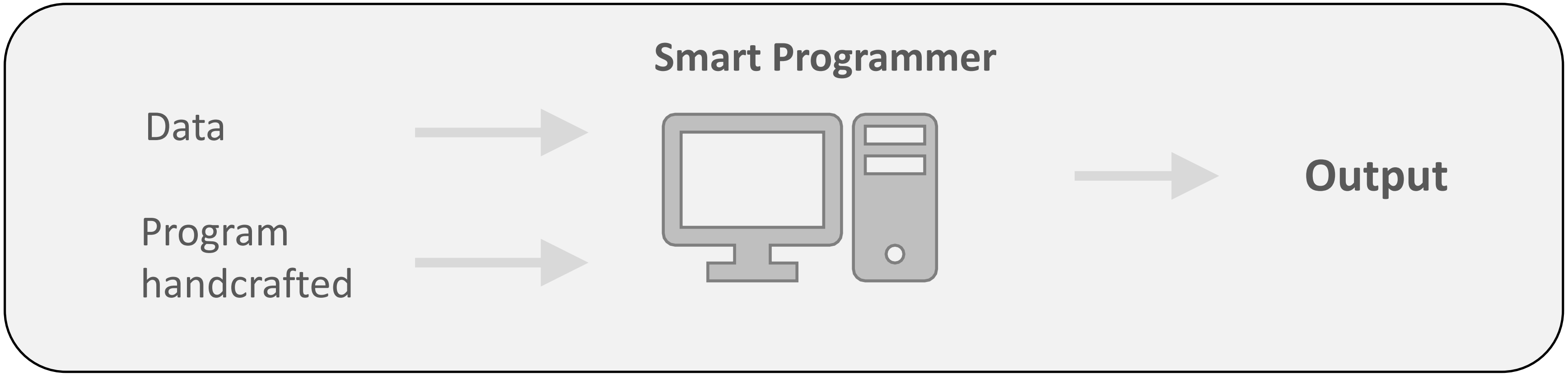
ML: Automated algorithms

In Machine Learning:

- ML algorithms – not programmers – create the rules.
- Requires training data to generate new rules.
- The more training-data the better ML algorithms learn.

Internet Society: https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/?gclid=CjwKCAjw8pH3BRAXEiwA1pvMsQNVjdVzTYdG303BSjNVJ3xDil9ru5K1Z28RCQ-VDBep1qI0WmQjBBoCDpsQAvD_BwE

Conceptual Picture of ML



Smart Programmer

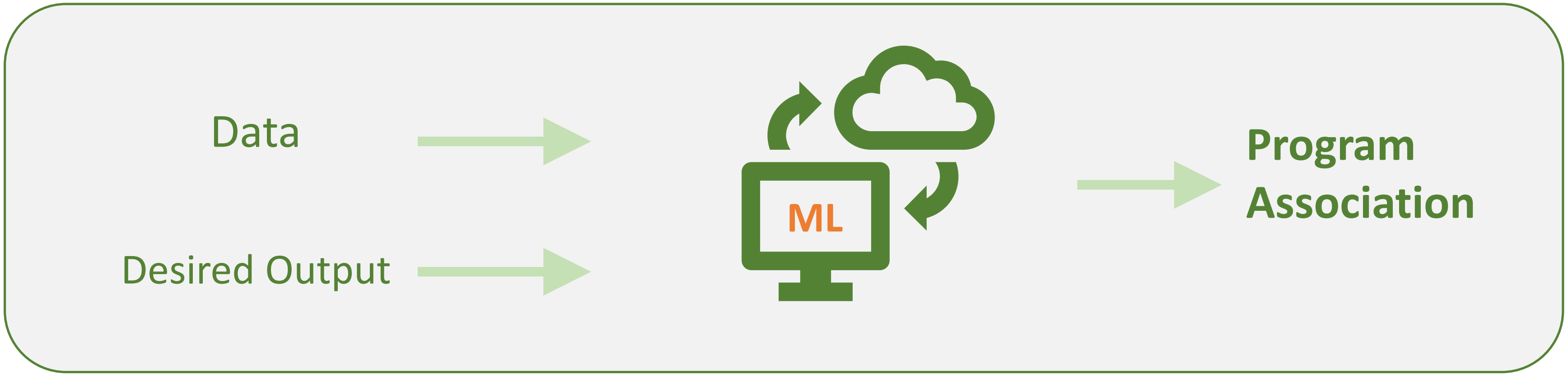
Data

Program handcrafted

Output

Dead Technology

Current Technology



Data

Desired Output

ML

Program Association

Source: https://courses.cs.washington.edu/courses/csep573/11wi/lectures/19-mlintro.pdf Dr. Pedro Domingos

ML: Definition

Tom Mitchell

*A computer program (ML model) is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E***

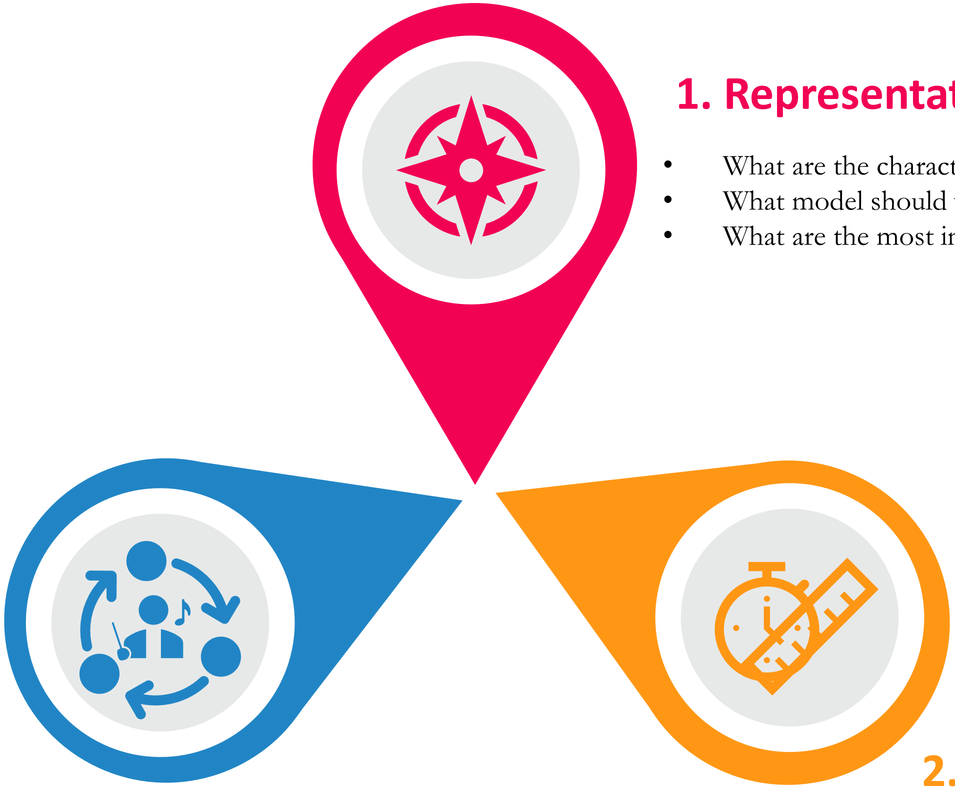
http://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf

Handwriting Recognition
Task (T): Recognizing handwritten words within images.
Experience (E): A labeled dataset of handwritten words
Performance (P): A rate of words correctly classified.

Autonomous Driving
Task (T): Driving on public highways using vision cameras.
Experience (E): A sequence of images and steering commands recorded while observing a human driver.
Performance (P): Average distance travelled before an error had happened (as judged by a human observer).

ML Key Factors

Learning = Representation + Evaluation + Optimization



1. Representation

- What are the characteristics of the problem?
- What model should we use?
- What are the most important hyperparameters to tune?

2. Evaluation

- How to evaluate the performance of the model ?
- What captures the model's errors?

3. Optimization

- Allow a selection among variety of ML models.
- Improving the leaner performance by tuning its parameters with more data

Source: https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

ML Three Key Factor

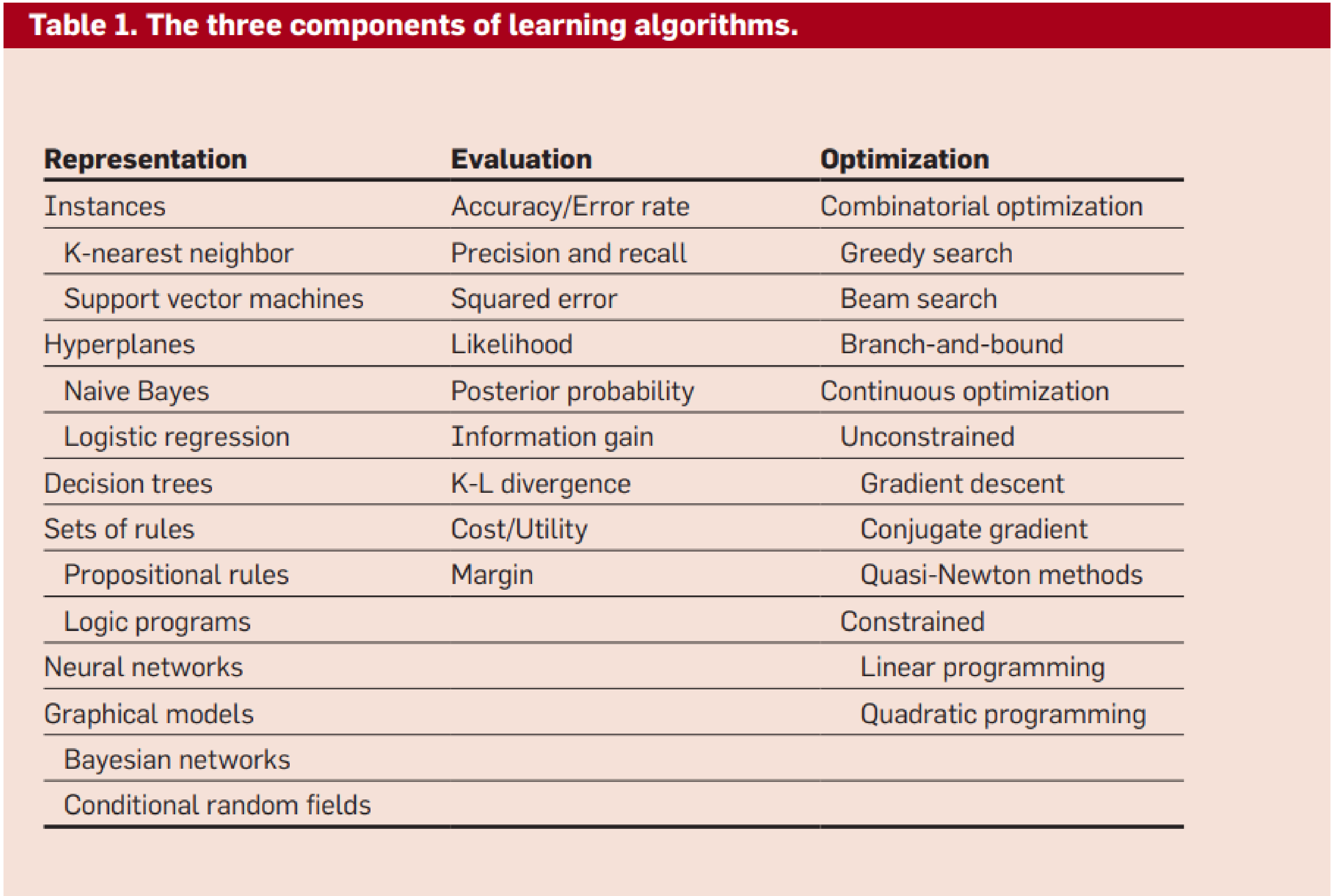


Table 1. The three components of learning algorithms.

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

Source: https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Module – 4 Materials and Focus

Module - 4

- In this module, we're covering an introductory material on ML that includes the best practices to develop ML models. In particular, the following types will be covered
  - **Unsupervised learning**
    - **Clustering**, algorithms, and evaluation
  - **Supervised learning**
    - **Regression** from Least Squares to stepwise, evaluation, pipelines, and regularization methods
    - **Classification** using ML algorithms, model selection strategies, evaluation metrics, etc.
- **Scikit-learn library** will be used throughout this module to develop ML models, evaluate them and compare/select ML models.

References

- Textbook:
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). **An introduction to statistical learning** (Vol. 112, pp. 3-7). New York: springer. (R language)
  - Belyadi, H., & Haghighat, A. (2021). **Machine Learning Guide for Oil and Gas Using Python: A Step-by-Step Breakdown with Data, Algorithms, Codes, and Applications.** Gulf Professional Publishing.
  - [CODE available] Müller, A. C., & Guido, S. (2016). **Introduction to machine learning with Python: a guide for data scientists**. " O'Reilly Media, Inc.".
- Libraries:
  - Scikit Learn Library Documentation: https://scikit-learn.org/stable/