Introduction

Machine Learning

Data Science (M.A.), RUPP
21-AUG-2023

周南公立大学
SHUNAN UNIVERSITY

Kiyota HASHIMOTO, D.Eng.
hash@g.shunan-u.ac.jp

Who we are

- Kiyota Hashimoto
  - Professor, Shunan University, Japan and Prince of Songkla University, Thailand
- Hidekazu Yanagimoto
  - Assoc. Prof., Osaka Metropolitan University, Japan
- Expertise:
  - NLP, ML, Data Science, IoT

Our courses here

- Machine Learning (Kiyota & Hide)
- Big Data & Deep Learning (Kiyota & Hide)
- Research Methodology (Kiyota)

- As our stay is limited, so all courses are mixture of face-to-face and online classes.

Machine Learning

Course Description	Machine learning is a key technology of data-oriented tasks nowadays, and its wide and deep understanding is essential for proper applications of them. Based on this recognition, this course provides not only how to use various machine learning techniques but also when, where, why, and how to apply machine learning approaches with appropariate evaluations. Most of the classes consist of both lecturing and practice with an emphasis on good recognition of the validity of the chosen approach based on task understanding.
Learning Outcome	By studying this course, students are expected to be able - to understand theoretical keys of machine learning approaches, inlucluding necessary mathematics and optimization - to use various machine learning techniques -to evaluate the results and how to improve - to design and conduct machine learning projects

Machine Learning

	Date	Contents		Form	Assignment
1	21Aug	Introduction	On site	Lecture / Exercise	Python environment building
2	22Aug	Linear classification / Regression	On site	Lecture / Exercise	Optimization
3	23Aug	Evaluation / Metrics	On site	Lecture / Exercise	Classification(Iris)
4	24Aug	Neural Network / DL	On site	Lecture / Exercise	Performance evaluation
5		**Mathematical Backgrounds**	Online	Lecture / Exercise	Classification(Iris)
6		Pre-processing	Online	Lecture / Exercise	Classification(MNIST)
7		Nonlinear classification (SVM)	Online	Lecture / Exercise	Visualization
8		Logistic regression / Naïve Bayes	Online	Lecture / Exercise	Classification
9		Ensemble Learning	Online	Lecture / Exercise	Classification
10		Clustering (unsupervised)	Online	Lecture / Exercise	Classification
11		Time series data analysis	Online	Lecture / Exercise	Classification
12		Anomally detection	Online	Lecture / Exercise	Clustering
13		How to proceed projects with machine learning	Online	Lecture	
14		Mini Project	Online	Exercises	
15		Mini Project	Online	Exercises	

Big Data & Deep Learning

Course Description	It is all the more important to make decision making and to automize various tasks with data. In particular, such data are collected from various sources, periodically updated, but also are often imperfect and/or lack key insights by humans. Such data processing can be partly done by traditional machine learning approaches, but often needs more advanced approaches, one of which is to use Deep Learning. Deep Learning is simply applications of many layered neural networks, which is expected to automize feature engineering together. This course gives the basic understanding skills for Big Data analysis and Deep Learning. Each class consists of lecture and practice with more practices by assignment.
Learning Outcome	By studying this course, students are expected to be able -to know theoretical and practical knowledge to cope with Big Data -to acquire basic skills to conduct deep learning projects - to design projects with Big Data and Deep Learning by using various platforms

Schedule	Date	Contents	On site/online	Form	Assignment
1		Introduction	Onsite	Lecture/Exercises	
2		BIgData	Onsite	Lecture/Exercises	
3		Deep Learning	Onsite	Lecture/Exercises	
4		Reinforcement Learning	Onsite	Lecture/Exercises	
5		BigData & Spaarsity	Online	Lecture/Exercises	
6		Visualization	Online	Lecture/Exercises	
7		Deep Learning implementation	Online	Lecture/Exercises	
8		Image processing	Online	Lecture/Exercises	
9		Transformer	Online	Lecture/Exercises	
10		Generative AI	Online	Lecture/Exercises	
11		Reinforcement Learning Aplication	Online	Lecture/Exercises	
12		GIS	Online	Lecture/Exercises	
13		Mini project	Online	Exercises	
14		Mini project	Online	Exercises	
15		Final Presentation	Online	Exercises	

#1-4 will be done in the middle of Sept.

Both ML and BDDL

- We assume that all of you have some familiarity with Python
- We use Anaconda 3 (Version 2023/7) and Google Colab
- All data & script can be downloaded via GitHub

Grading	Assignments: 50% (must submit at least more than half of assignments) Mini project accomplishment including final presentation: 50%

What is ML?

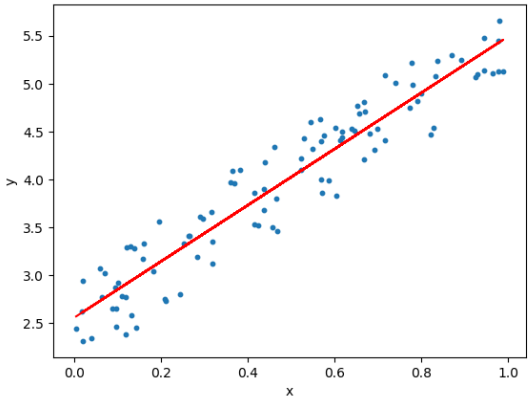
- Usual definition
  - **Machine learning is an application of artificial intelligence that uses statistical techniques to enable computers to learn and make decisions without being explicitly programmed.** It is predicated on the notion that computers can learn from data, spot patterns, and make judgments with little assistance from humans.

What is ML?

- Usual definition
  - **Machine learning is an application of artificial intelligence that uses statistical techniques to enable computers to learn and make decisions without being explicitly programmed.** It is predicated on the notion that computers can learn from data, spot patterns, and make judgments with little assistance from humans.
- More practical definition
  - Machine Learning is to build a numerical model for something using optimization with data.

A Simplest Example

- Numerical model: a function such as $y = ax + b$

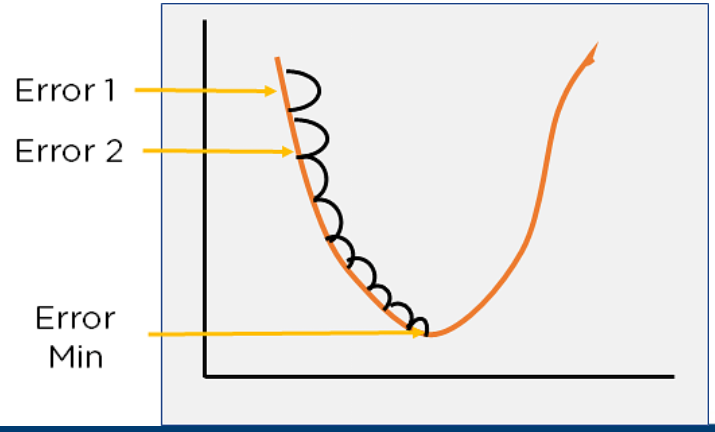


- We want to know **a** and **b**
- For each data (in this example), we know the actual data (x_i, y_i) and $(x_i, \hat{y}_i)$ (on the line), and we want to minimize the difference between these y_i and $\hat{y}_i$
- If the data is small, we can solve analytically, but if data is bigger, we need trial-error optimization.

- Optimization should minimize the total differences among all data.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

- To minimize this by choosing the best **a** and **b**, Machine Learning takes an trial-error approach.



So, ML is

- Numerical modeling + Optimization with Data
- There are a variety of machine learning algorithm, and each adopts a particular function or probability model
  - In other words, ML can learning only based on the assumed function/probability model.
- Each also adopts a specific optimization technique
- The problem is that we usually don't know which ML algorithm is most appropriate for a task.
  - Thus we usually apply several ML methods to find the best one.

Machine learning is a magic wand?

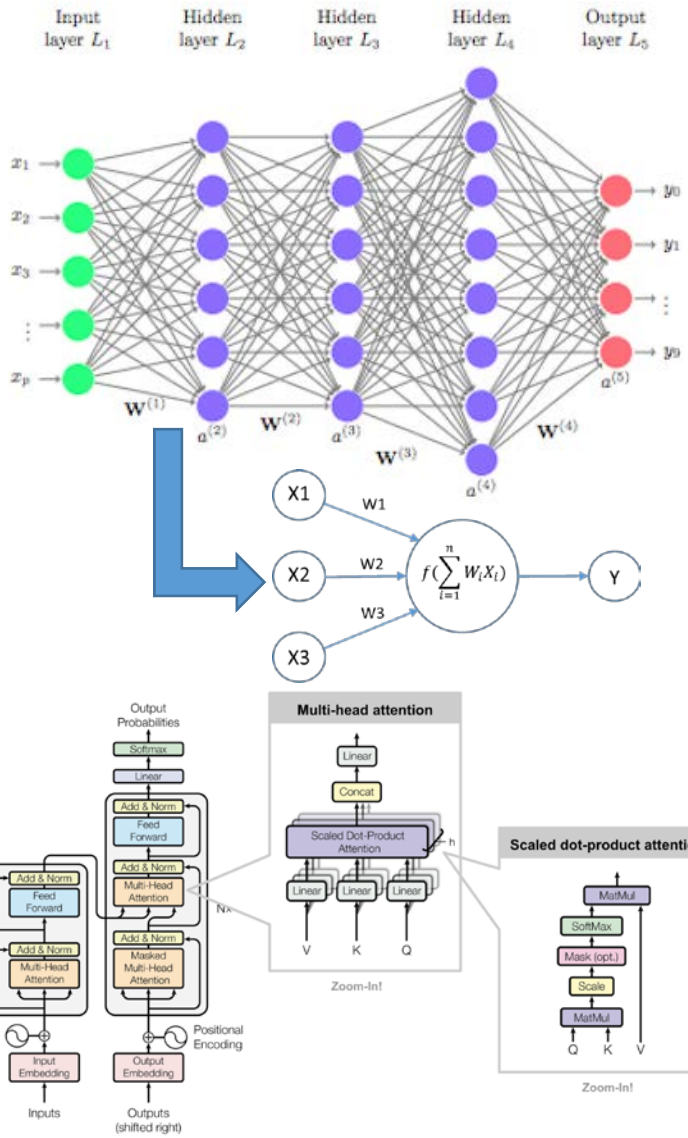
Simply, machine learning seeks to approximate co-efficients in a mathematical model to describe the population with the available sample

- Each machine learning method has its own mathematical model (e.g., $y = a\,x + b$ in the case of linear regression). This means that the application of a particular machine learning method presumes that we expect the characteristics and distribution of the data are well approximated with its model.
- Usually, machine learning attempts are more used for data with many parameters / features. ← Which humans cannot exercise their reasoning and deterministic calculations are not available.
- Unfortunately, we often have little clue to expect what characteristics and distribution our data will have, and that's why we often apply, by trial-and-error, many methods to find which is better than others
- Machine learning is definitely *NOT* a magic wand

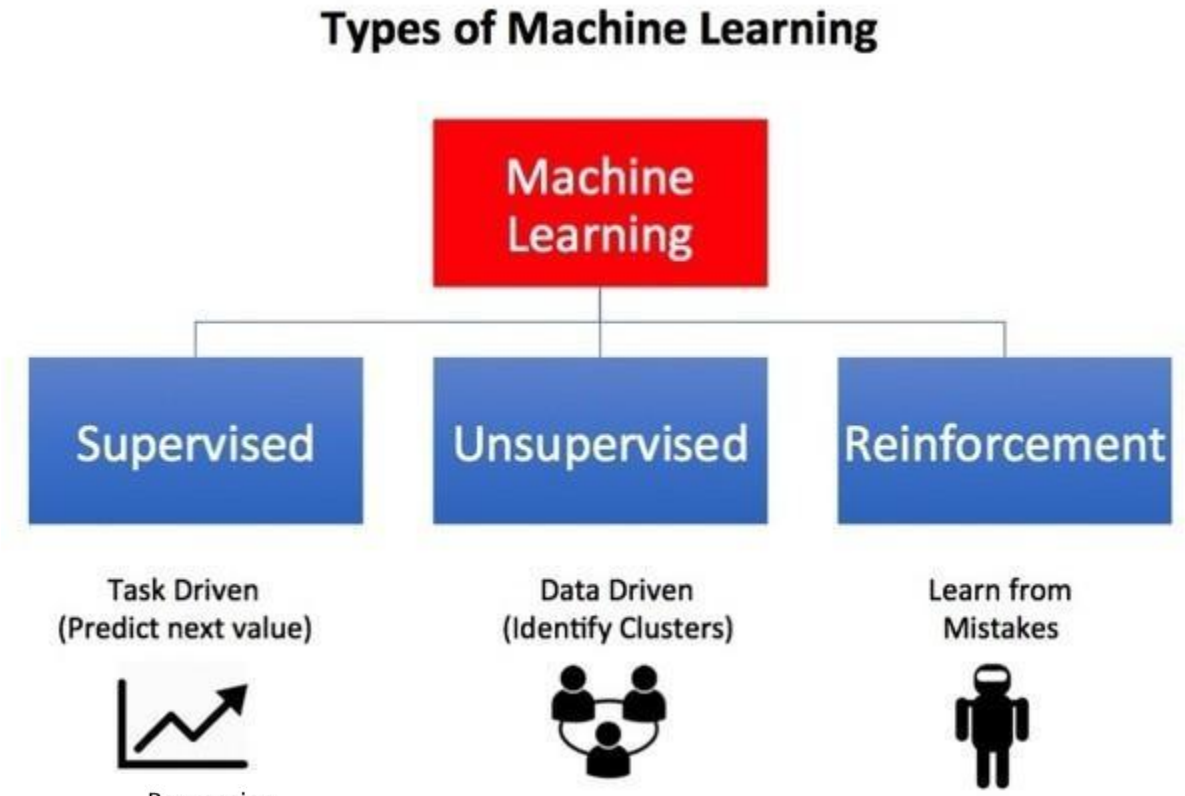
Deep learning, so-called

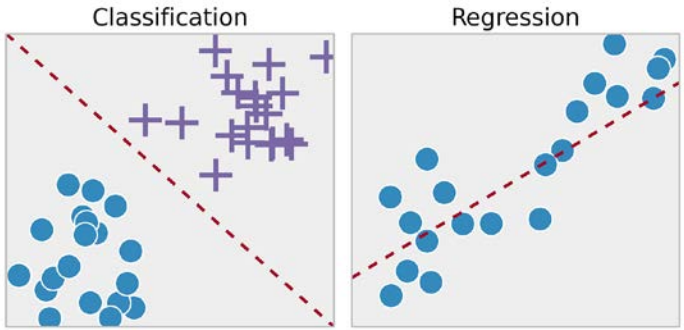
What *deep learning* is new is that deep neural networks, based on multi-stories of non-linearity-flavored polynomial function, can approximate any mathematical model

- Deep learning is a machine learning method to employ (often very) many layers of neural networks
- However, as it contains a huge number of co-efficients, their estimation needs a huge amount of data, and computer resources & electricity to compute
- Particularly useful when our data (in a computable format) are so superficial that we are not sure how such data are linked to the true characteristics of the data → image, language, sound, etc.



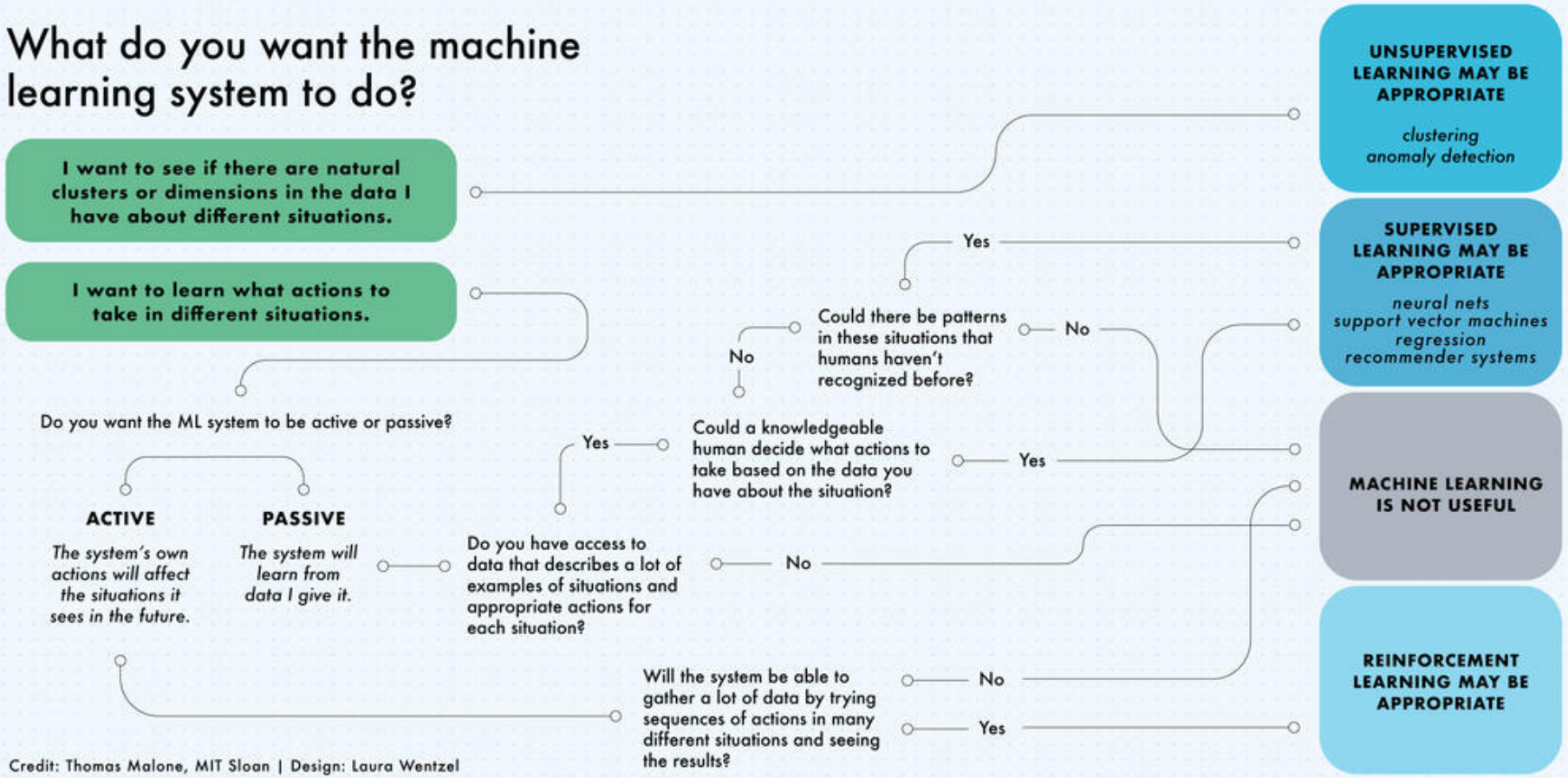
Variety of learning



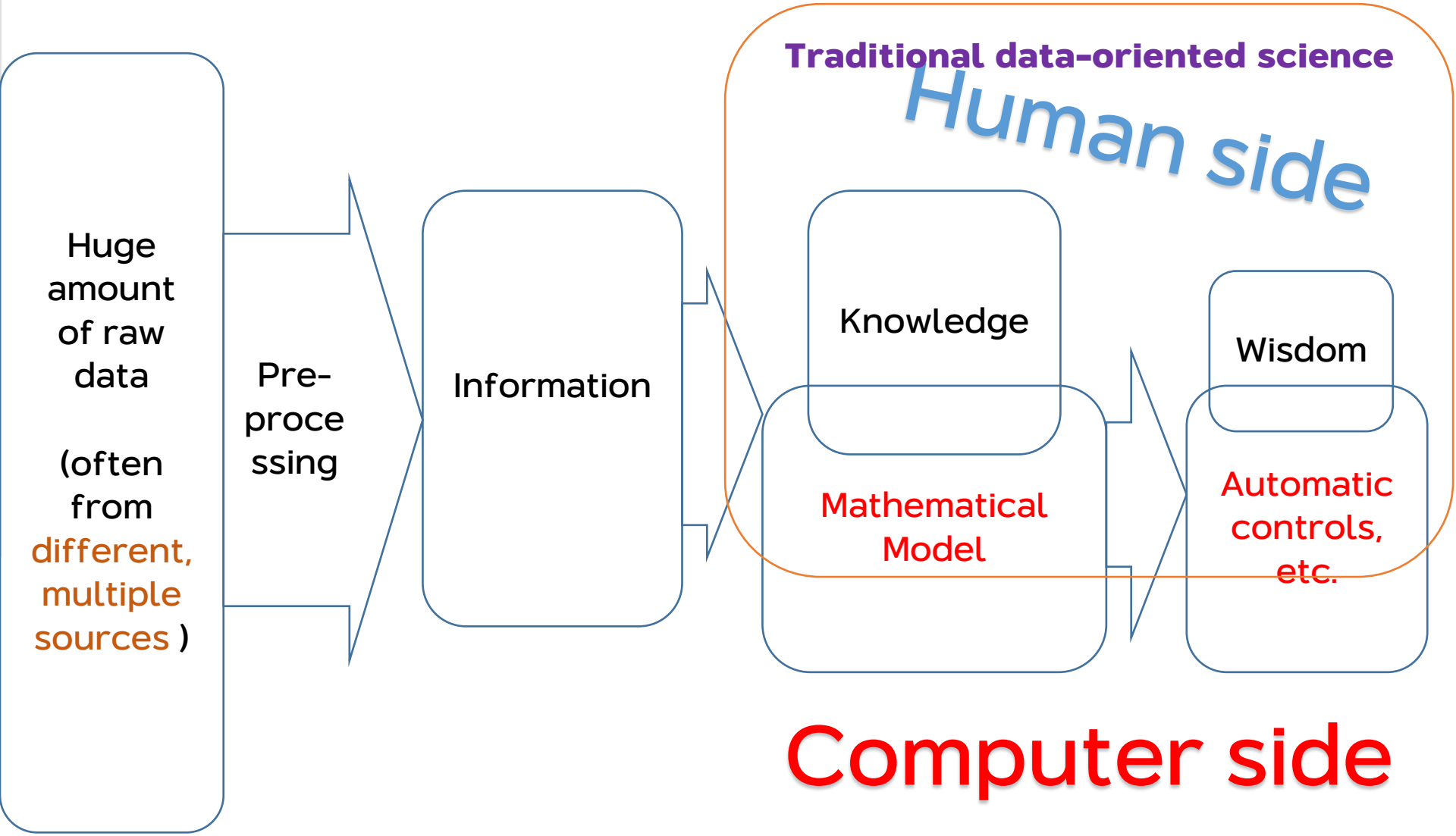


Clustering
Association

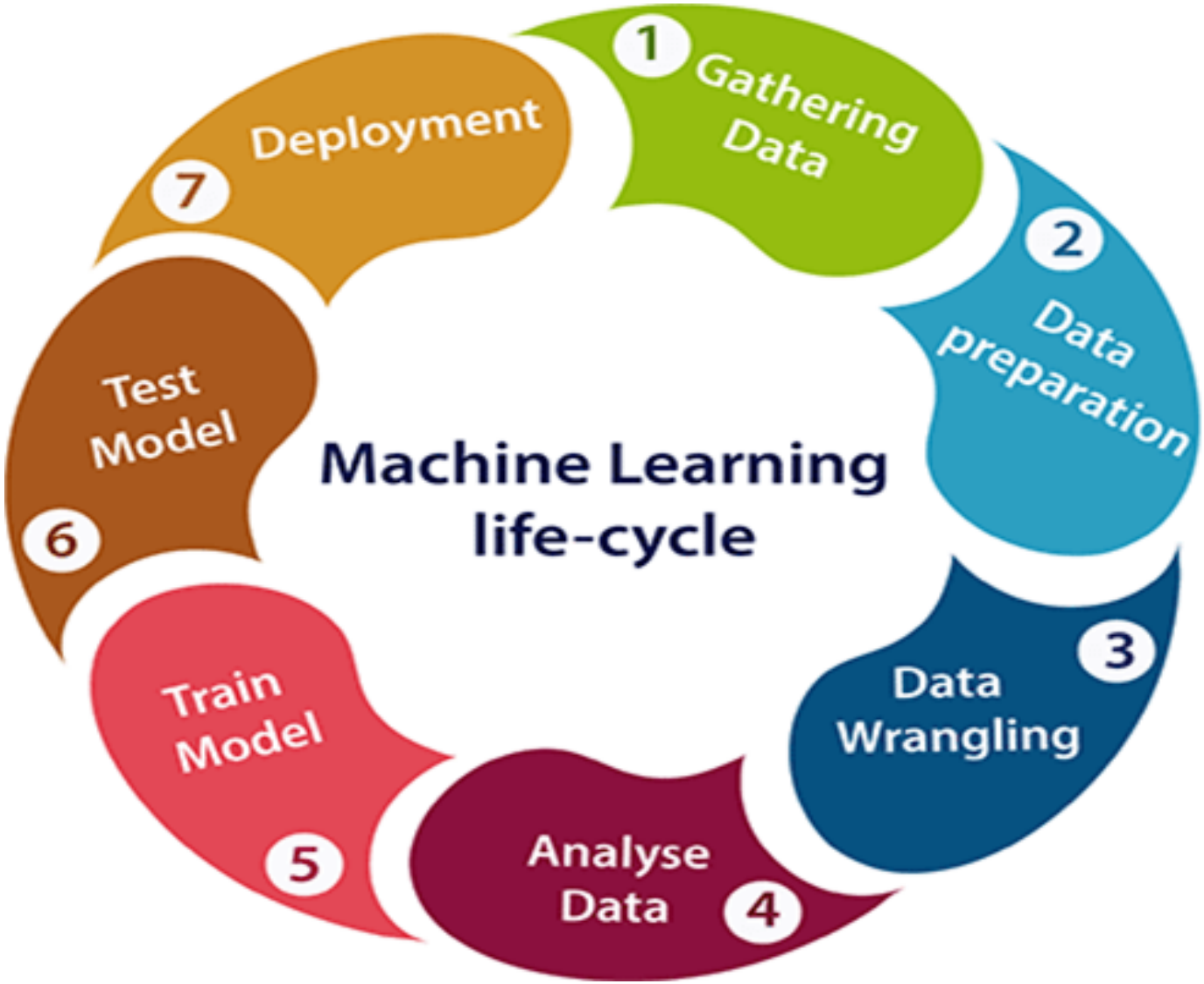
Choice Chart



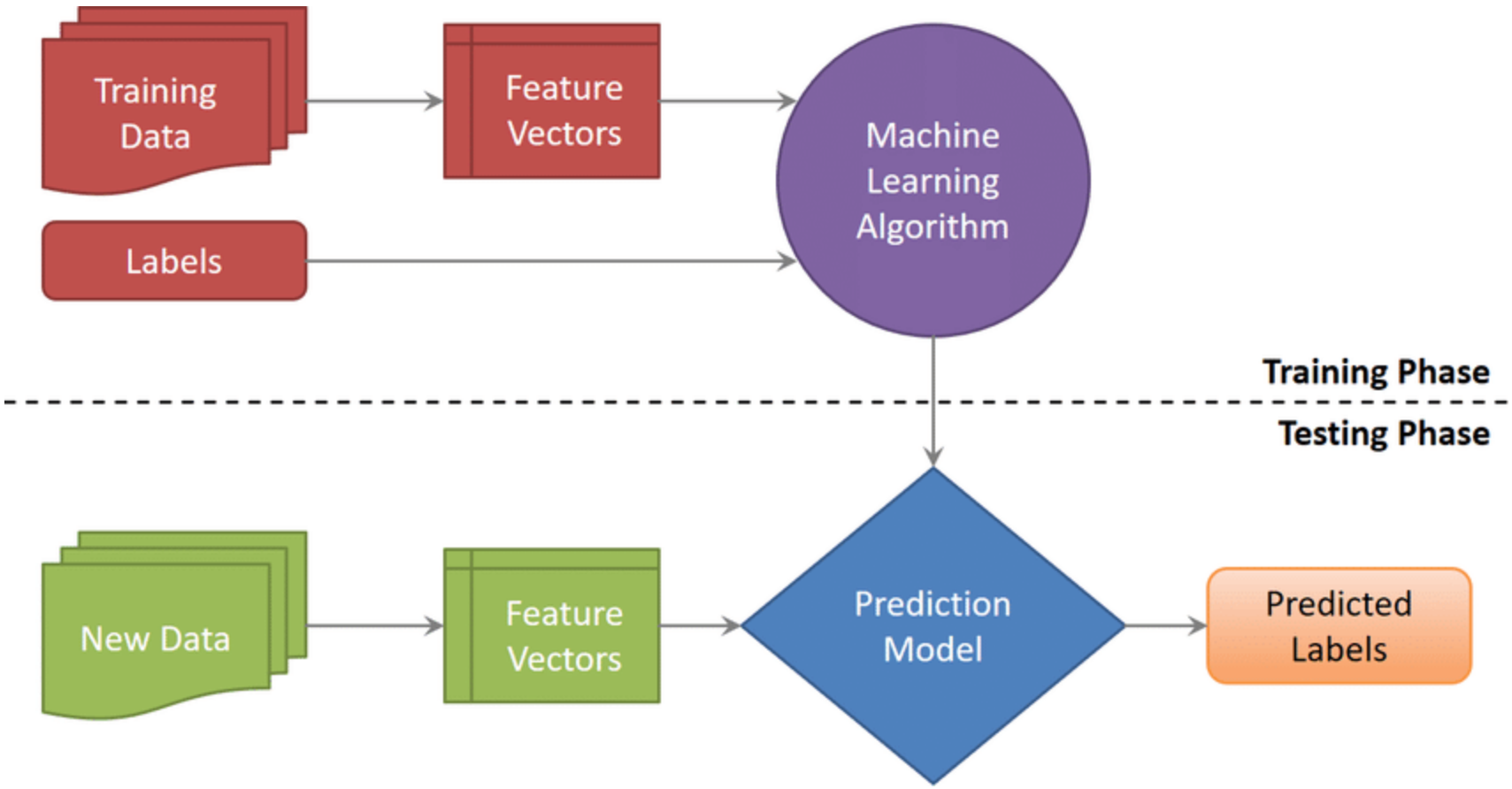
Simple scheme of machine learning



How ML procedure works



More specifically



Caution!

- Mostly, ML tasks focus more on **predictability**, and thus too much fitting with the training data is **un**desirable. → How to **evaluate**?
- Most of ML methods are short of **explainability** (why the obtained model works well?) → If a task needs good explainability but wants to use an ML, we often choose low-performance but easy-to-explain ML models (such as *Decision Tree*)
- Usual ML methods are wrongly affected by too many unrelated features, and thus **feature engineering & selection** is a key but is done by humans
  - Deep Learning covers most of feature engineering (that's why DL is more promising, but much more data is needed)

So in this course you'll learn

- Theoretical understanding of well-known ML methods
  - In order to take advantage of your ML skills, you need to understand theory
- Hands-on applications of well-known ML methods
  - Remember that, even now, many tasks are done with *traditional* ML methods
    - Partly because they are enough
    - Partly because DL approaches are too costly or impossible to apply (due to shortage of data)
- The last mini-project will let you go through the whole process of ML tasks