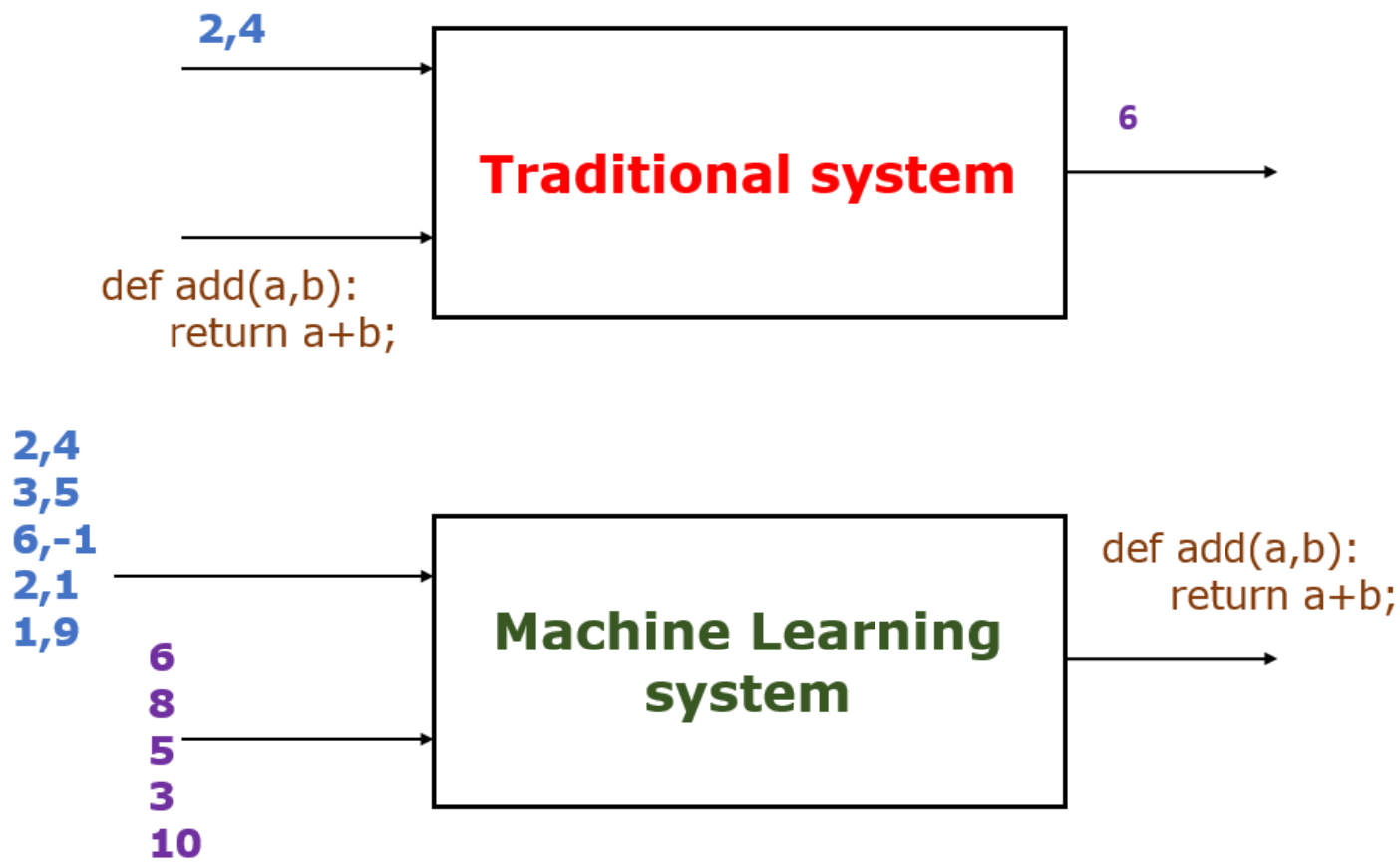
What is machine learning

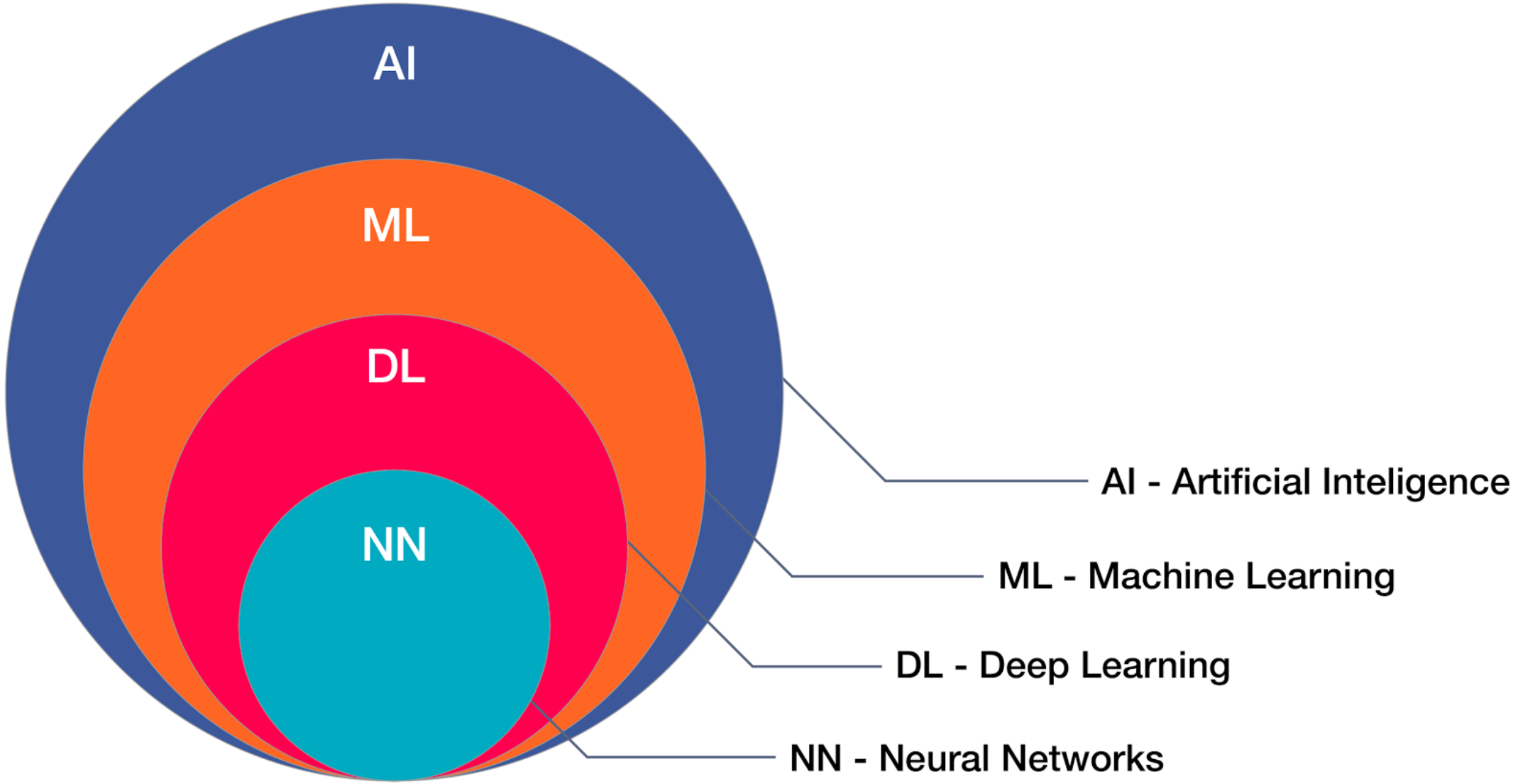
Machine Learning definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

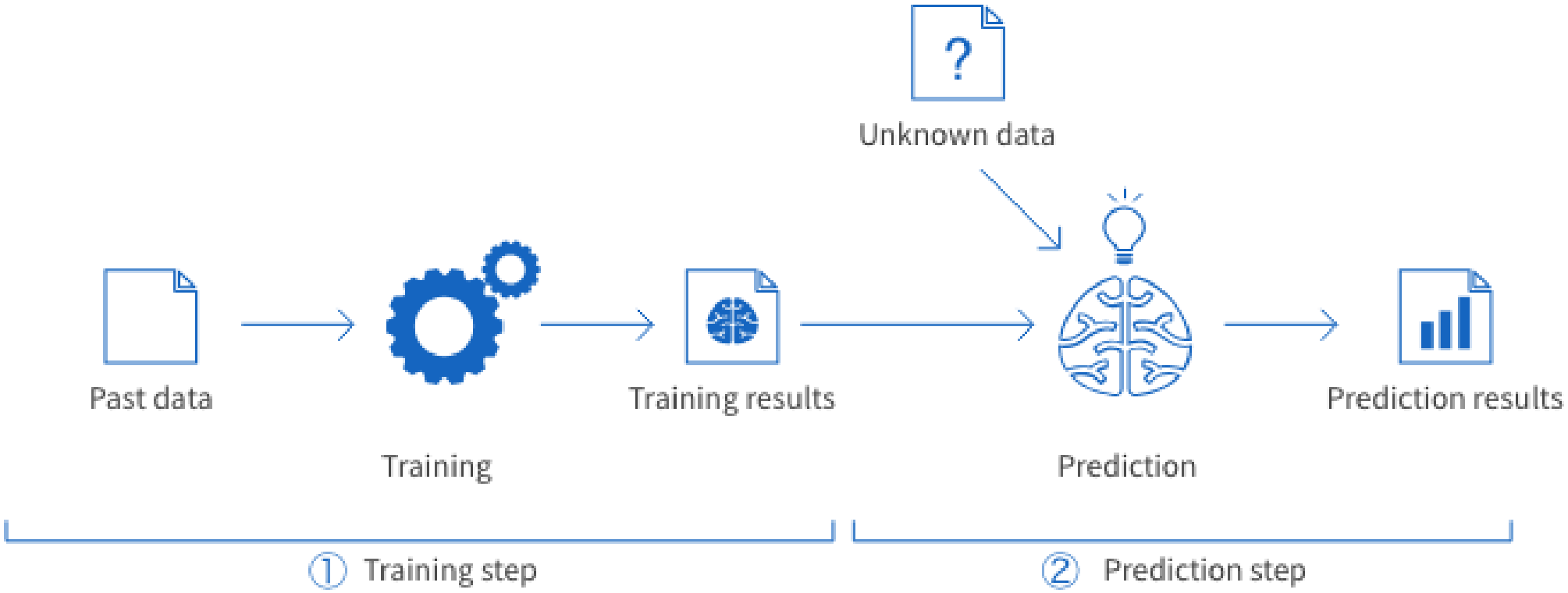
What is machine learning?



What is machine learning?



What ML does?


?
Unknown data
Past data
Training
Training results
Prediction
Prediction results
① Training step
② Prediction step

Why Data Pre-processing

- Before sending data to ML models, we need to prepare data which is acceptable by ML models.
- Its an essential part of machine learning.
- There are pre-defined steps for data pre-processing:
  - Get the dataset.
  - Import the dataset
  - Missing data treatment.
  - Categorical Data and feature transformation
  - Splitting data in train and test set
  - Feature scaling

Variable Selection Methods

What Variable Selection Is

- “Variable selection” means selecting which variables to include in our model.
- Working in machine learning field is not only about building different classification or clustering models.
- It’s more about feeding the right set of features into the training models.
- This process of feeding the right set of features into the model mainly take place after the data collection process.

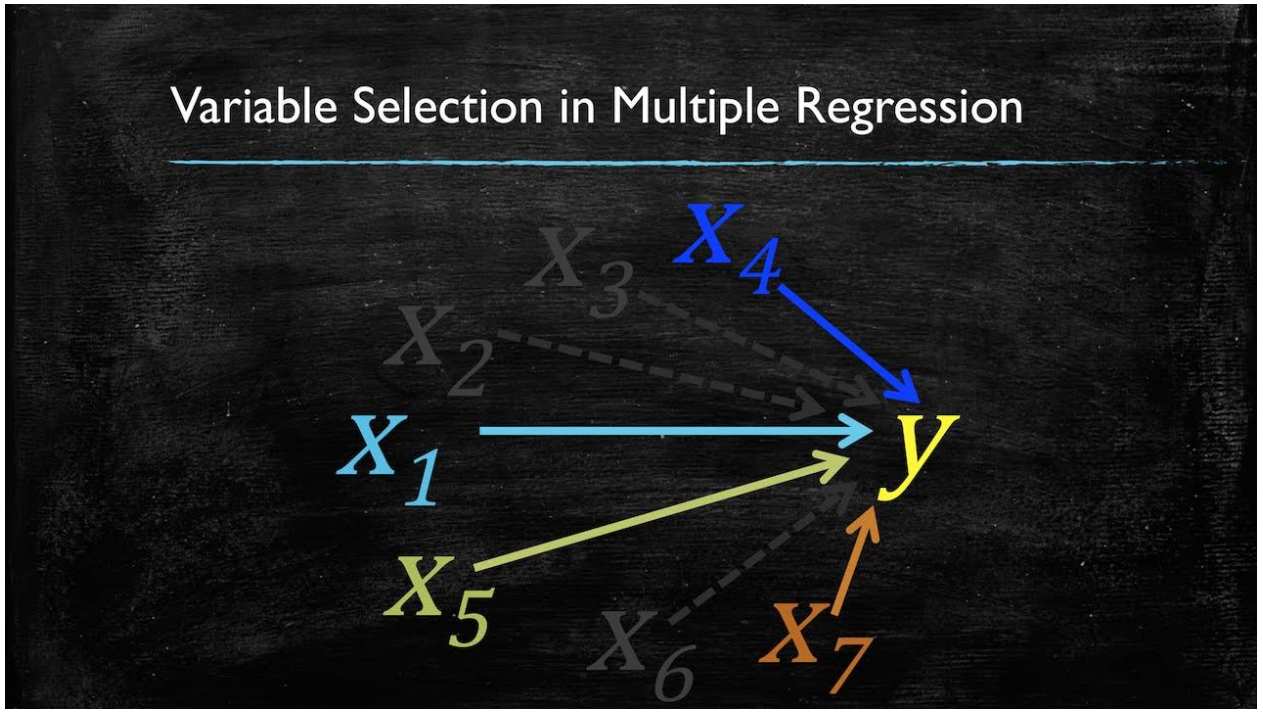
- Example: For prediction of the scores in exams, below variables are given:
  - Student Name
  - Student Roll Number
  - Age
  - Class (1 – 10)
  - Average score in class
  - Student average performance till now
  - Subject
  - Board
  - Exam Centre
- In the example above the Student Name, Roll Number are clearly not the variable needed to predict.

Variable Selection

- Once we have enough data, We won’t feed entire data into the model and expect great results. We need to pre-process the data.
- In fact, the challenging and the key part of machine learning processes is data pre-processing.
- Below are the key things we indented to do in data pre-processing stage.
  - Feature transformation
  - Feature selection
- Feature transformation is to transform the already existed features into other forms.

Variable Selection

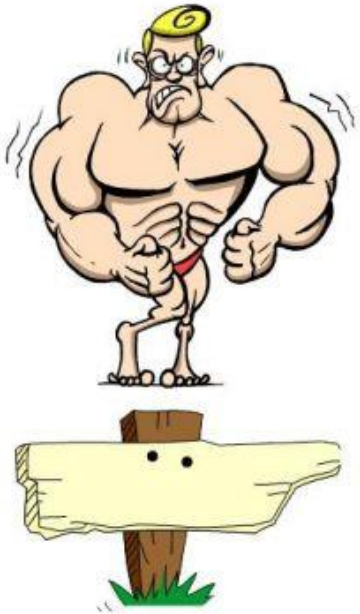
- Feature selection is to select the best features out of already existed features.
- We are going to learn the basic techniques to pick the best features for modelling.



Correlation

Correlation Coefficient

- **Correlation Coefficient**: number between +1 and -1 that represents the strength and direction of the relationship between two variables
- Correlations that are closer to +1 and –1 are stronger and are better able to accurately predict



Variable Selection techniques

- Forward selection:
  - Starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
- Backward selection (or backward elimination):
  - Starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
- Stepwise selection (or sequential replacement):
  - A combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

Thanks