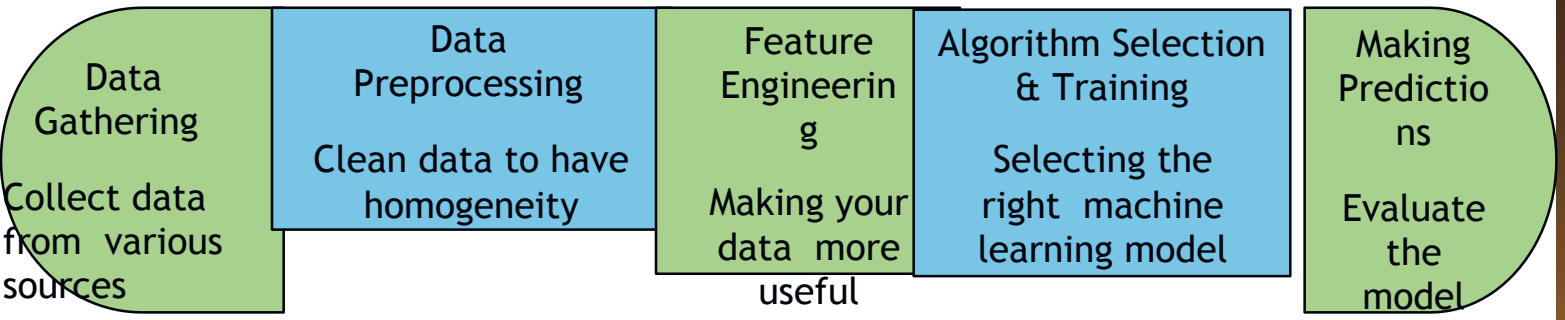
Steps to Solve a Machine Learning Problem



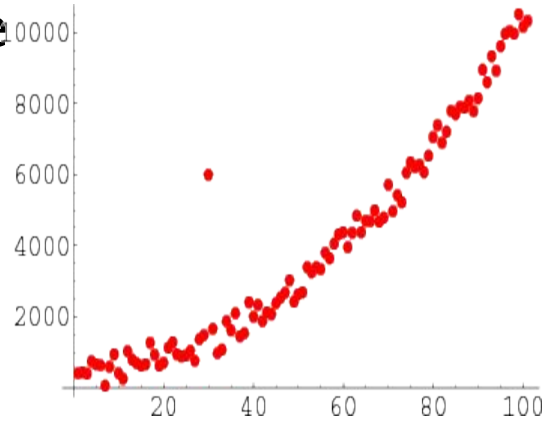
Data Gathering

INTERNSHIPSTUDIO

- Most of the time dependent on human work
- Manual labeling for supervised learning.
- Domain knowledge. Maybe even experts.
- May come for free at times, or “sort of”, E.g., Machine Translation.
- The more the better: Some algorithms need large amounts of data to be useful (e.g., neural networks).
- The quantity and quality of data dictate the model accuracy

Data Preprocessing

- Is there anything wrong with the data?
  - Missing values
  - Outliers
  - Bad encoding (for text)
  - Wrongly-labeled examples
  - Biased data
- Many more samples of one class than the rest?
- Need to fix/remove data?



Feature Engineering

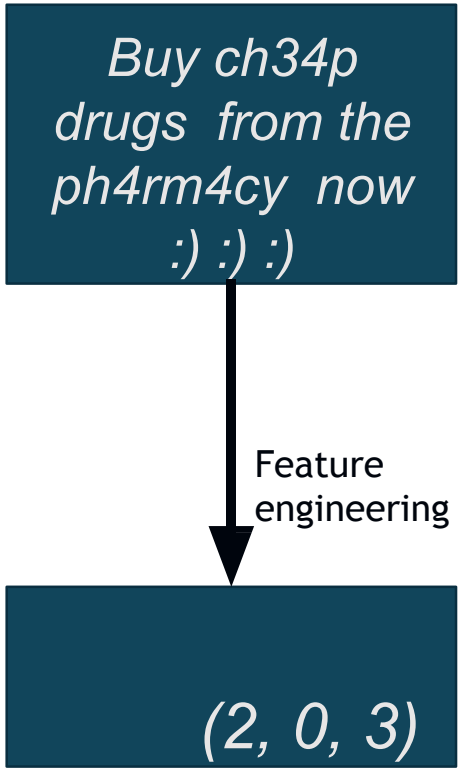
INTERNSHIPSTUDIO

- What is FE?

Given a multidimension dataset, extract information based on interrelations among variables.

- A feature is an individual measurable property of a phenomenon being observed
- Our inputs are represented by a set of features.

- To classify spam email, features could be
  - Number of words that have been ch4ng3d like this.
  - Language of the email (0=English, 1=Spanish)
  - Number of emojis



Feature Engineering (Cont)

INTERNSHIPSTUDIO

- Extract more information from existing data without adding "new" data
- With good features, most algorithms can learn faster
- Requires thought and knowledge of

  the data

Two steps:

- Variable transformation (e.g., dates into weekdays, normalizing)
- Feature creation (e.g., n-grams for texts, if word is capitalized to detect names, etc.)

Algorithm Selection & Training

Supervised

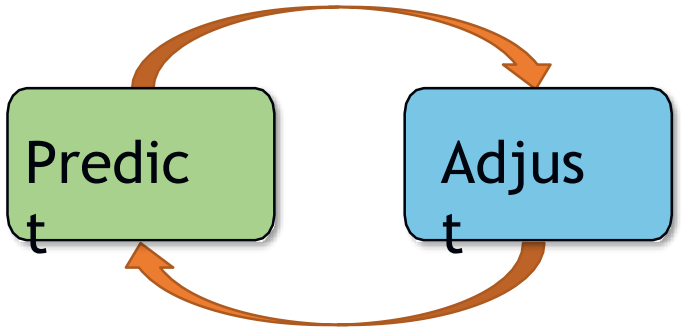
- Linear classifier
- Naive Bayes
- Support Vector Machines (SVM)
- Decision Tree
- Random Forests
- k-Nearest Neighbors
- Neural Networks (Deep learning)

Unsupervised

- PCA
- t-SNE
- k-means
- DBSCAN

Algorithm Selection & Training

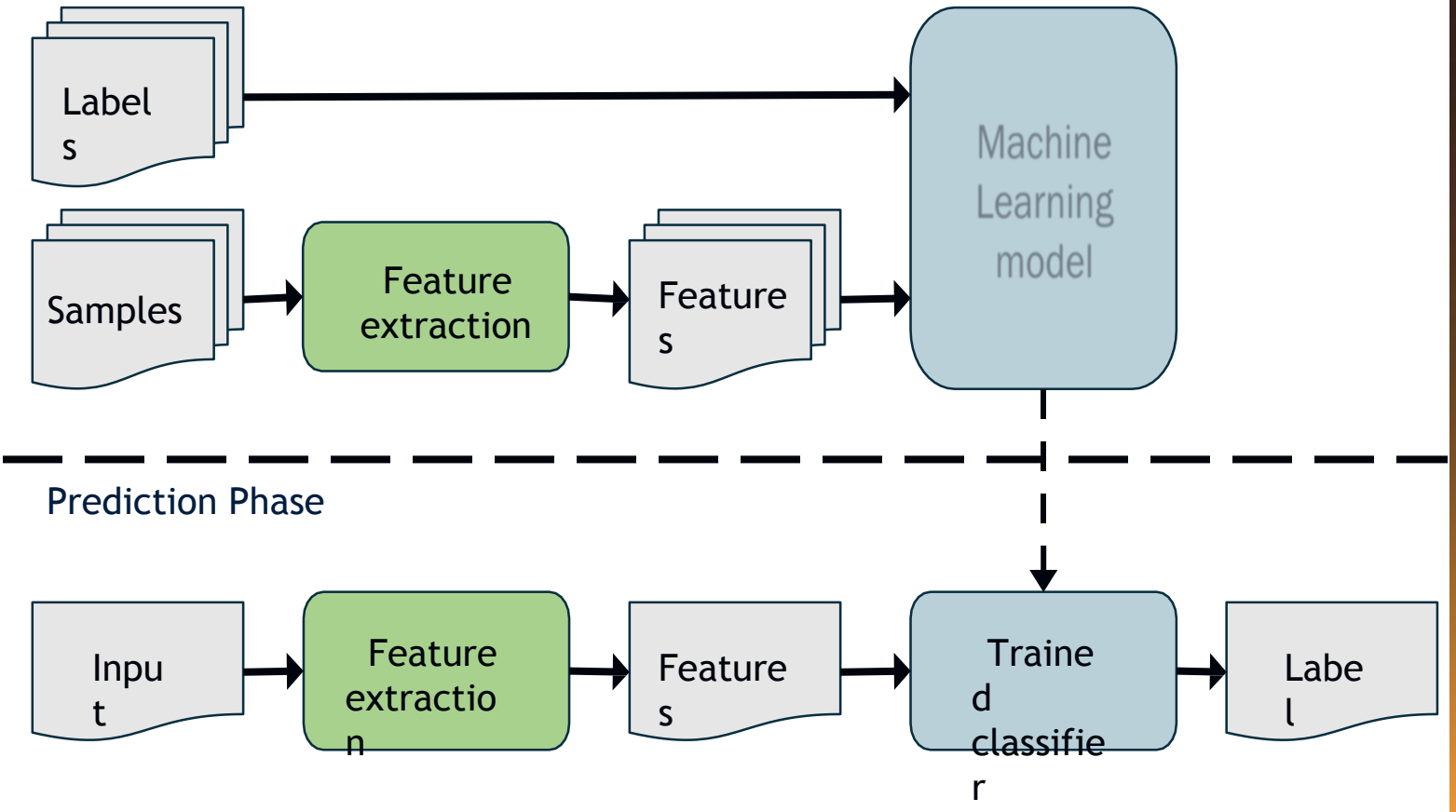
- Goal of training: making the correct prediction as often as possible
- Incremental improvement:



- Use of metrics for evaluating performance and comparing solutions
- Hyperparameter tuning: more an art than a science

Making Predictions

- Training Phase



Machine learning

INTERNSHIPSTUDIO

Q.1 What are the steps to solve ML problems?

Q.2 Explain Data Gathering?

Q.3 Explain Data Pre-Processing?

Q.4 Explain Feature Engineering?

Q.5 What are the various Algorithm Selection & Training?

Q.6 Explain Making predictions in ML?

THANK YOU