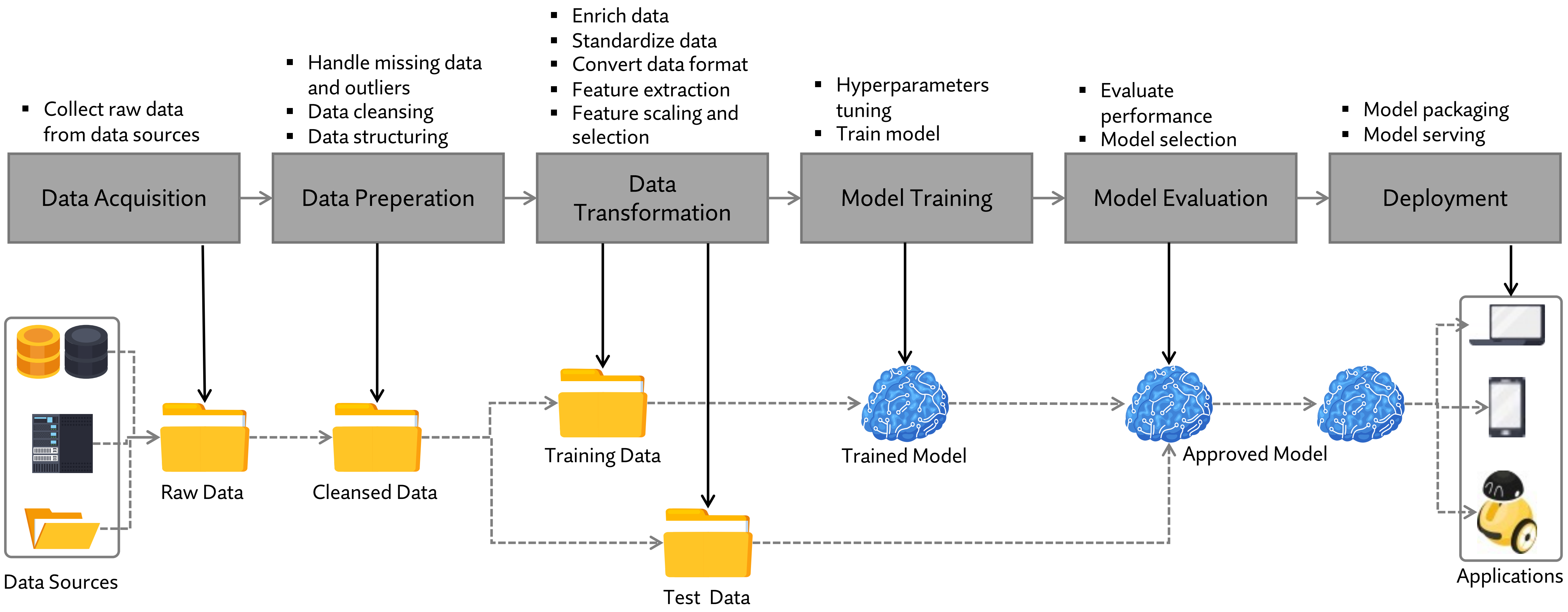
Feature Engineering

Papangkorn Inkeaw, Ph.D.

Overview of Feature Engineering

Chapter 1

Machine Learning Workflow



Machine Learning Workflow

Example 1

- AI for identifying plant species from plant leaf photo.
- Machine learning pipeline:
  - Collection plant leaf photos with their species from many data sources.
  - Combine data from several sources, standardize file format and organize file structure.
  - Convert RGB images to binary images, reduce noise and resize images to have the same size.
  - Extract region of interest (ROI).
  - Extract features from images.
  - Separate training and test datasets.
  - Built a classifier (using the training dataset)
  - Evaluate classification performance
  - Develop an application for identifing plant species from plant leaf photo using the model (classifier).

This example was adapted from Kaur, Surleen & Kaur, Prabhpreet. (2019). Plant Species Identification based on Plant Leaf Using Computer Vision and Machine Learning Techniques. Journal of Multimedia Information System. 6. 49-60. 10.33851/JMIS.2019.6.2.49.

Data & Features

Data are:

- Recording of fact.
- Observations of real-world phenomena.

Feature is:

- An attribute of data that is meaningful to the machine learning process.
- A numeric representation of raw data.

Feature Vector is:

- A representation of a datum.
- An <u>ordered list</u> of <u>numerical properties</u> of <u>observed phenomena</u>.

Data & Features

Mathematical aspect

- Machine learning model is a function that transform input (in form of a feature vector) to desired output.

$$f: U \rightarrow Y$$

- Where the domain U (here, called feature space) has the form $U \subseteq X_1 \times X_2 \times \cdots \times X_d$ and the dimension of U is d.
- Y is the range or image of a function.

Data & Features

Attribute, Property, Feature, Dimension, Variable, Field

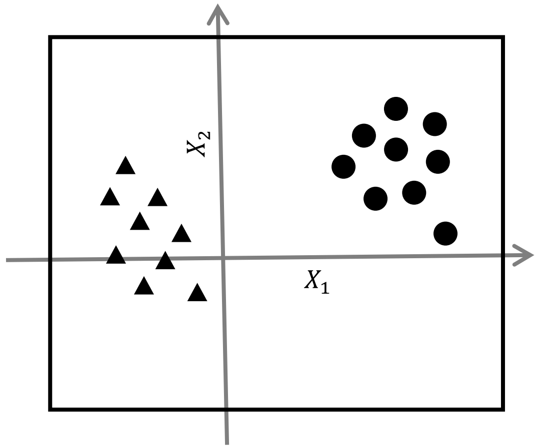
$$
\begin{array}{cc} & X_1 \quad X_2 \quad \cdots \quad X_d \\
D = \begin{array}{c} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \vdots \\ \mathrm{x}_n \end{array} & \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \end{array}
$$

Entity, Instance, Example, Record, Transction, Data point, Tuple, Feature vector

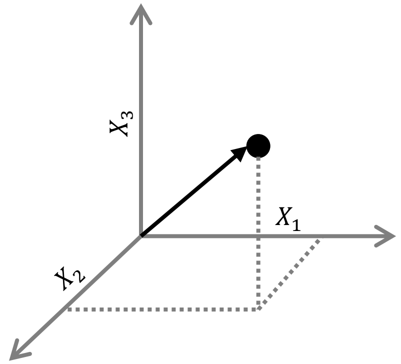
Samples

$$\mathrm{x_i} = \left(x_{i,1}, x_{i,2}, \dots, x_{i,d}\right)$$

Feature Vector



Feature Space (2D)



A feature vector in feature space (3D)

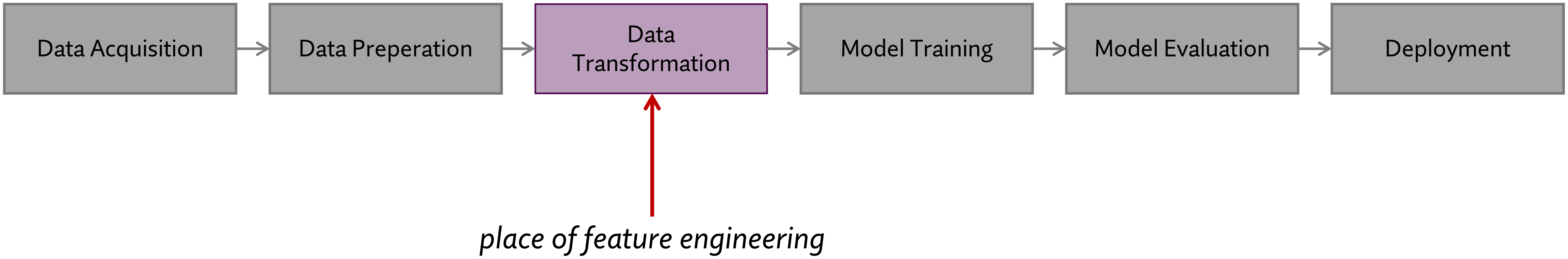
Feature Engineering in Machine Learning

Feature Engineering

task *objective*

"The process of transforming data into features that better represent the underlying problem, resulting in improved machine learning performance."

indicator



Feature Engineering Types

Categories of feature engineering:

1. **Feature Improvement**: Making existing features more usable through mathematical transformations.
   - Imputing (filling in) missing data
   - Removing harmful data
   - Normalizing/standardizing data
   - Feature encoding and transformation
2. **Feature Construction**: Augmenting the dataset by creating net new interpretable features from existing interpretable features.
   - Combining features
   - Expanding features
   - Deriving new features

Feature Engineering Types

Categories of feature engineering:

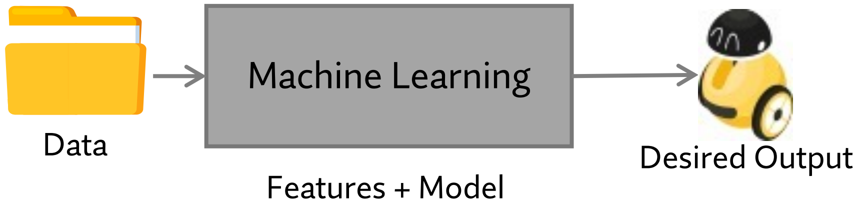
3. **Feature Extraction**: Relying on algorithms to automatically create new sometimes uninterpretable features usually based on making parametric assumptions about the data.
4. **Feature Selection**: Choosing the best subset of features from an existing set of features.
5. **Feature Learning**: Automatically generating a brand new set of features usually by extracting structure and learning representations from raw unstructured data such as text, images, and videos, often using deep learning.

Feature Engineering Types

Curse of Dimensionality

- In higher-dimensional spaces, everything starts to become very close.
- That is the case as meaningful differences on a few key coordinates are drowned over similar values for the remaining coordinates.
- Peaking phenomenon: with a fixed number of training samples, the average (expected) predictive power of a classifier or regressor first increases as the number of dimensions or features used is increased but beyond a certain dimensionality it starts deteriorating instead of improving steadily.
- More dimensions present in the data, the more training data needed (a rule of thumb says an absolute minimum of five training instances per dimension.)

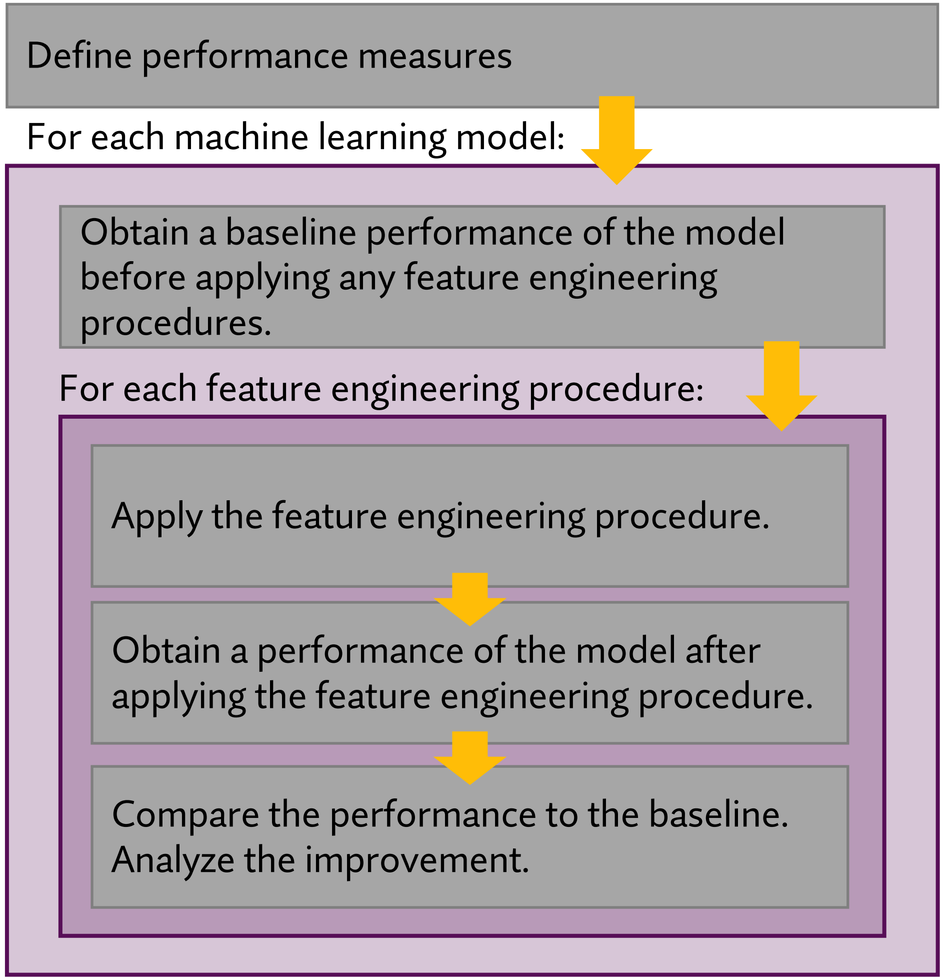
Evaluation of Feature Engineering Procedure



- In a machine learning workflow, we pick both the model and the features.
- A double-jointed lever, the choice of one affects the other.
  - Good features make the subsequent modeling step easy and the resulting model more capable of completing the desired task.
  - Bad features may require a much more complicated model to achieve the same level of performance.
- We cannot evaluate the performance of feature engineering procedure without model
  - The performance of the features is measured after applying them to model.

Evaluation of Feature Engineering Procedure

Steps to evaluate feature engineering procedure:



Model	**Feature Engineering Precedure**				
	Baseline	Precedure 1	Precedure 2	...	Precedure m
Model 1					
Model 2					
...					
Model n					

An example of experimental result record

Evaluation of Feature Engineering Procedure

Performance Matrices

- Supervised Learning
  - Classification
    - Accuray
    - Sensitivity
    - Specificity
    - Precision
    - F1-Score
    - AUC
  - Regression
    - *R Square*
    - *Root Mean Square Error*
    - *Mean Absolute Error*
- *Unsupervised Learning*
  - V-measure
  - Silhouette index
  - Calinski-Harabasz index
  - Davies-Bouldin Index

Evaluation of Feature Engineering Procedure

Overfitting

- Model follows too closely to the original training sample, and it fails to generalize.
- So, we always use a separate test set when training supervised learning models.
- Sometimes, through a series of training and evaluation steps, we might gain insights about the full sample that will lead to overfitting of the ML process, not just on the testing-on-train data sense but on trickier ways.
- So, the test data should be independent from the training dataset.

References & Study Resources

- Alice Zheng and Amanda Casari. (2018). *Feature Engineering for Machine Learning*. O'Reilly Media, Inc.
- Pablo Duboue. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge University Press.
- Sinan Ozdemir. (2022). *Feature Engineering Bookcamp*. Manning Publications Co.
- Sinan Ozdemir and Divya Susarla. (2018). *Feature Engineering Made Easy*. Packt Publishing.
- Kaur, Surleen & Kaur, Prabhpreet. (2019). *Plant Species Identification based on Plant Leaf Using Computer Vision and Machine Learning Techniques*. Journal of Multimedia Information System. 6. 49-60. 10.33851/JMIS.2019.6.2.49.