LECTURER: TAI LE QUY

MACHINE LEARNING

UNSUPERVISED LEARNING AND FEATURE ENGINEERING

Who am I?

- Name: Tai Le Quy
- PhD at L3S Research Center – Leibniz University Hannover
  - Topic: Fairness-aware machine learning in educational data mining
- MSc in Information Technology at National University of Vietnam
- Profile: tailequy.github.io
- Email: tai.le-quy@iu.org
- Materials: https://github.com/tailequy/IU-ML-Unsupervised

INTRODUCTORY ROUND

Who are you?

- Name
- Employer
- Position/responsibilities
- Fun Fact
- Previous knowledge? Expectations?

UNIT 1

INTRODUCTION TO UNSUPERVISED MACHINE LEARNING AND FEATURE ENGINEERING

- Explain the **general principal** of unsupervised machine learning and its **applications** to real-life problems.
- Define what **features** are, their **types**, their interest for unsupervised machine learning, and their **challenges**.
- Explain the **steps of designing** an unsupervised machine learning **model**.
- **Adapt or transform features** for an unsupervised machine learning model.
- **Evaluate** and **improve** the **performance** of an unsupervised machine learning model.

EXPLAIN SIMPLY

1. Explain the main difference between clustering and dimensionality reduction.

2. Describe the main goals of feature engineering.

3. Briefly describe the main steps to build a successful unsupervised machine learning model?

Unsupervised machine learning

- Clustering algorithms
- Dimensionality reduction
- Real-life applications

Feature engineering

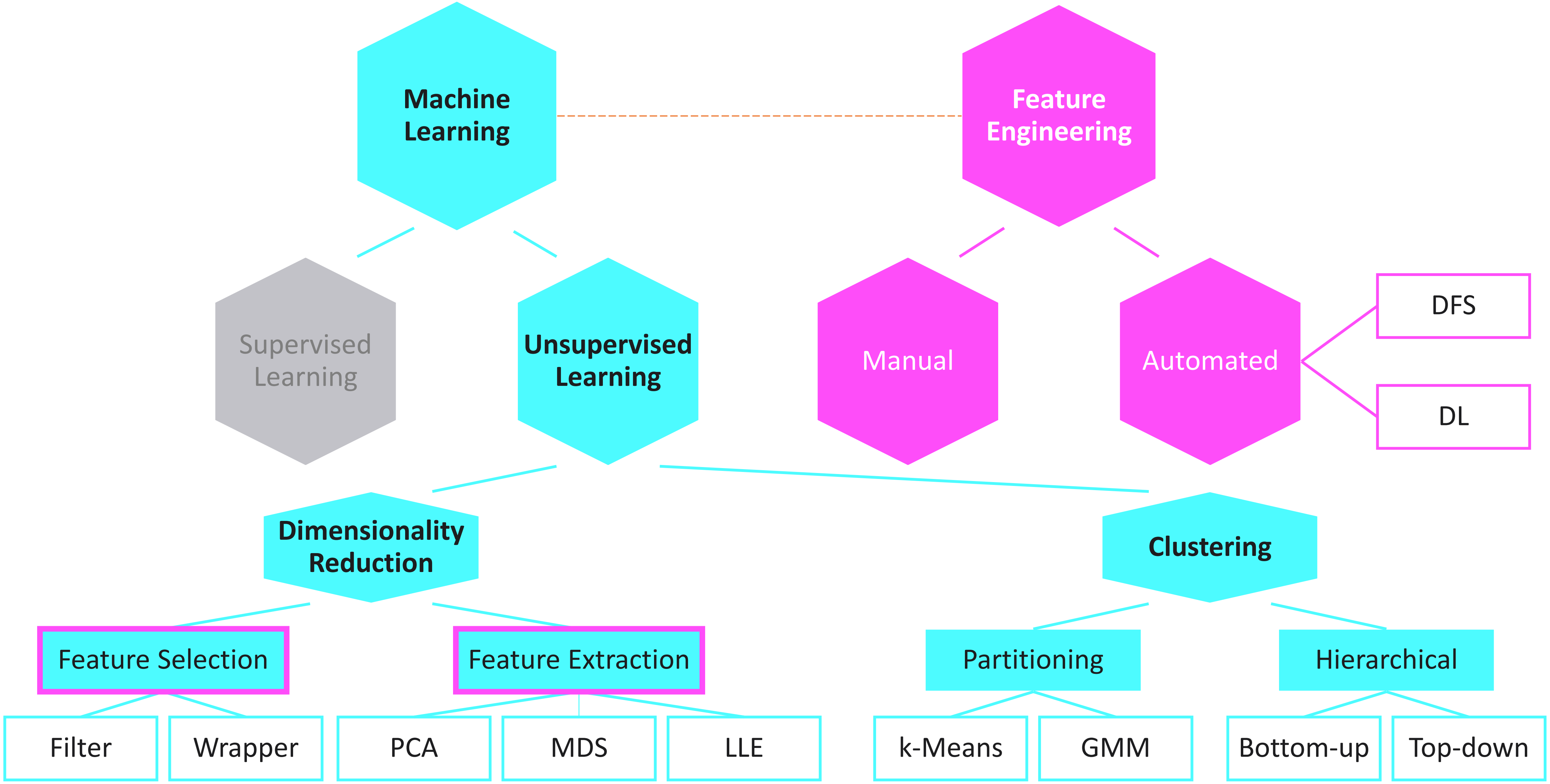
- Definition, types, and motivation
- Types and steps

Steps to build a successful unsupervised learning model

This presenation does **not cover the entire content** of the coursebook unit! It focusses on some aspects.

OVERVIEW OF UNSUPERVISED MACHINE LEARNING AND FEATURE ENGINEERING TECHNIQUES

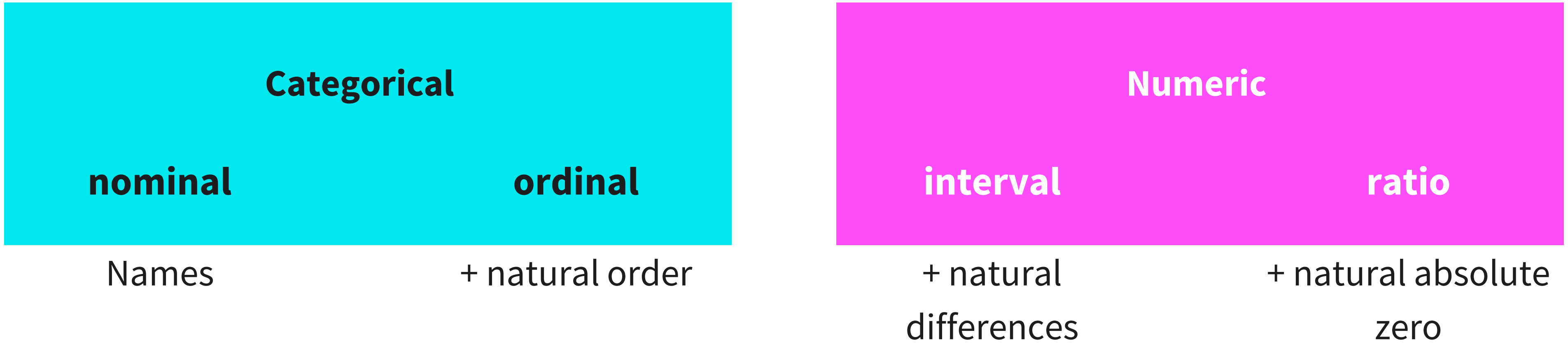
Image 1: Unsupervised ML and feature engineering overview



Source of the image: Müller-Kett, 2021.

LEVEL OF MEASUREMENT

Image 2: Levels of measurement



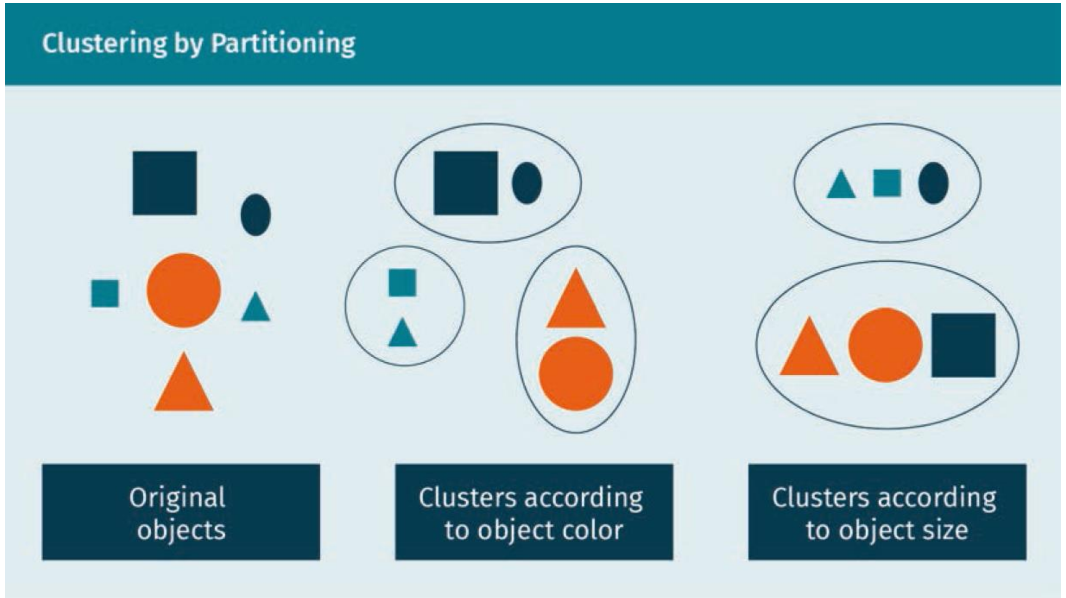
Source of the image: Müller-Kett, 2018.

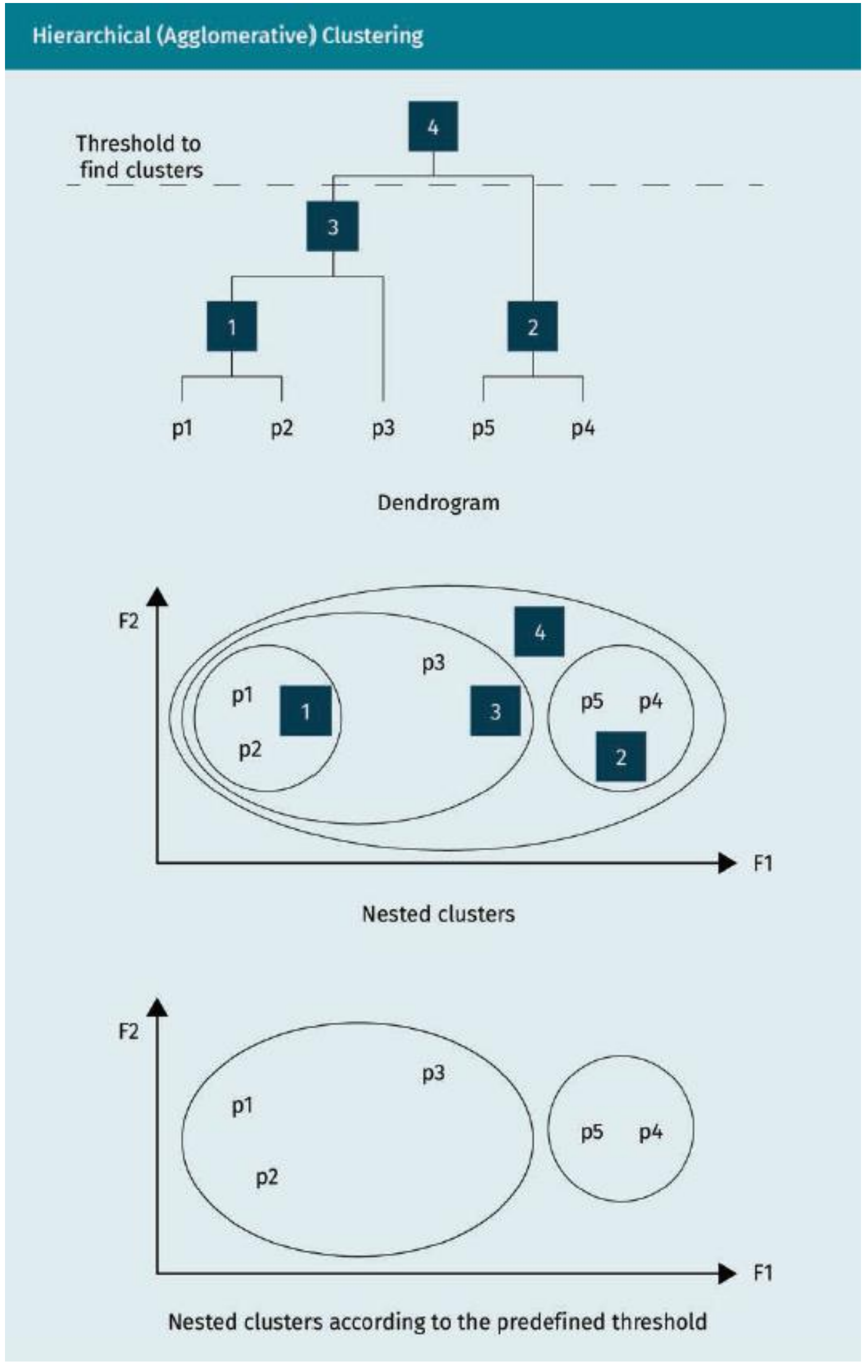
- Subdivide a data set of ***n* samples** into ***k groups***, i.e., clusters.
- **Samples in one cluster** should be **similar.**
- **Sample from different clusters** should be **different** from each other.
- **Different approaches**
  - Partitioning (k-Means, GMM, DBSCAN)
  - Hierarchical clustering

Image. 3: Clustering

Source of the image: Microsoft Archive.

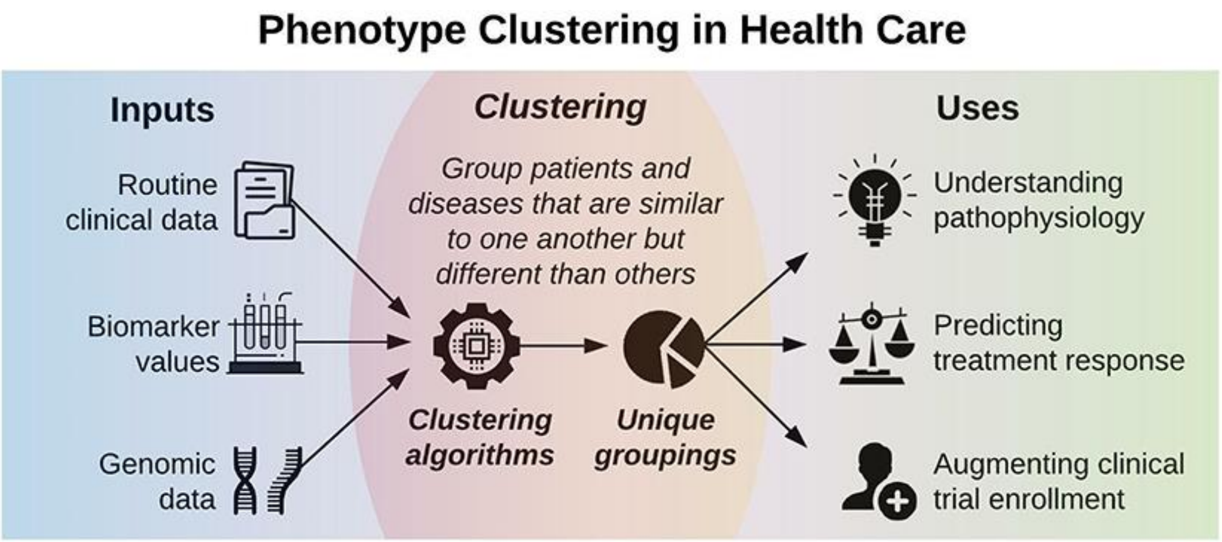
CLUSTERING





Source of the image: Course book.

- Medical diagnosis
- Fault diagnosis of industrial systems
- Customer segmentation or client profiling
- Crime and fraud detection

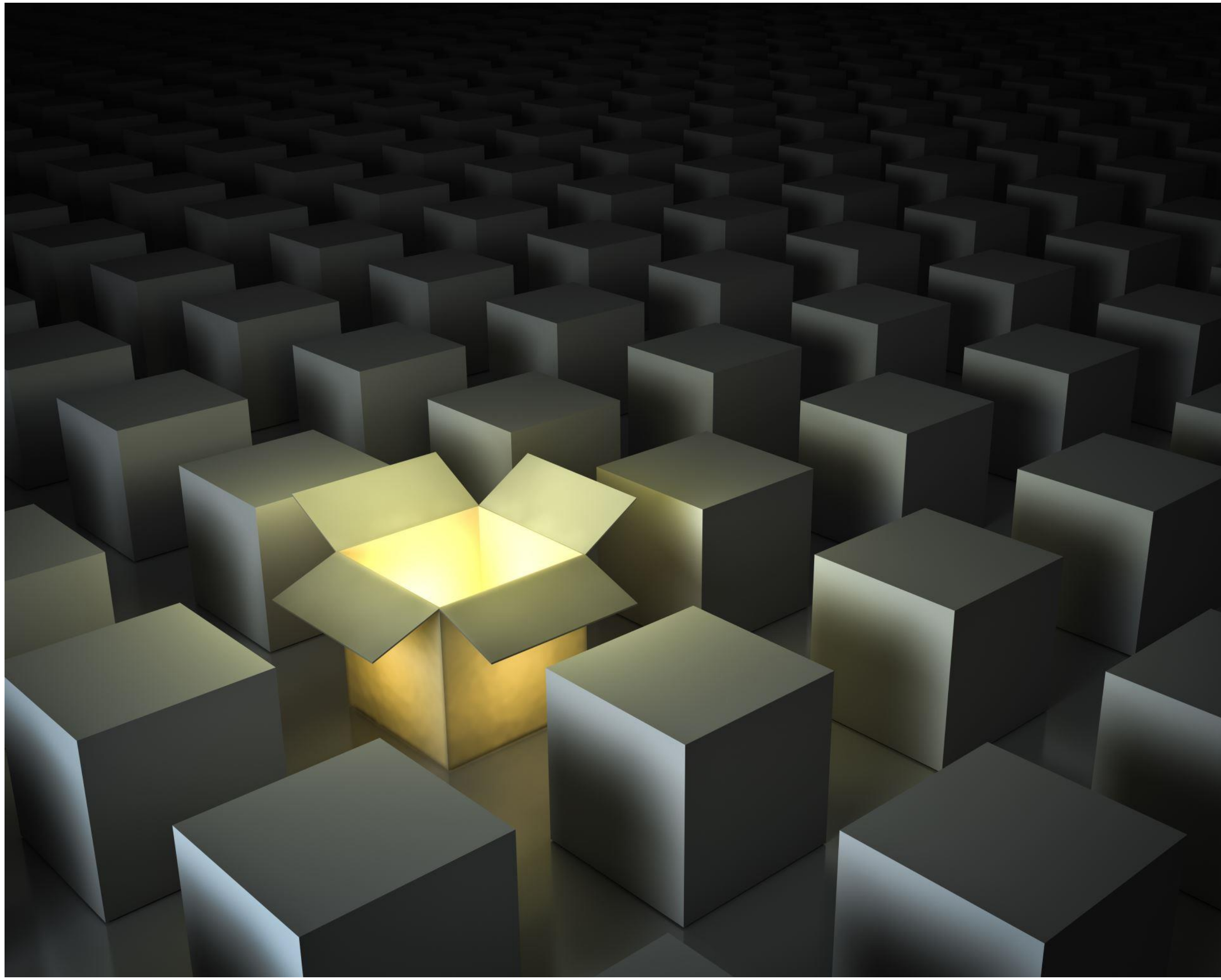


Source of the image: https://powerslides.com/powerpoint-marketing/analytics-templates/customer-profile-template/
https://doi.org/10.3389/frai.2022.842306

FEATURE ENGINEERING

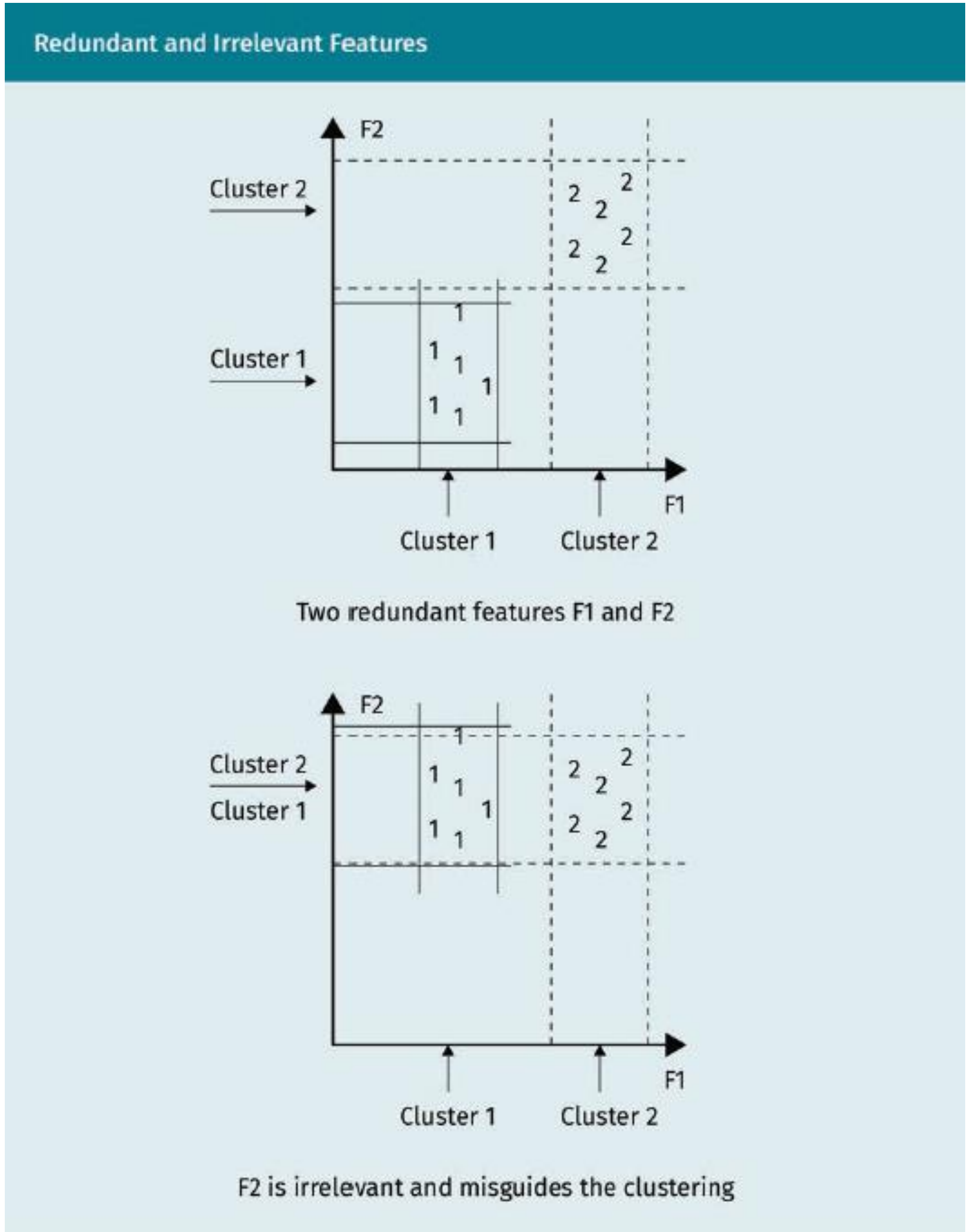
- Avoid **overfitting** and keep models **simple** (*curse of dimensionality*).
- Only use **relevant features.**
- Feature which contain **unique information.**
- **Feature Selection**
  - Wrapper methods
  - Filter methods
  - Embedded methods

Image. 4: Feature engineering



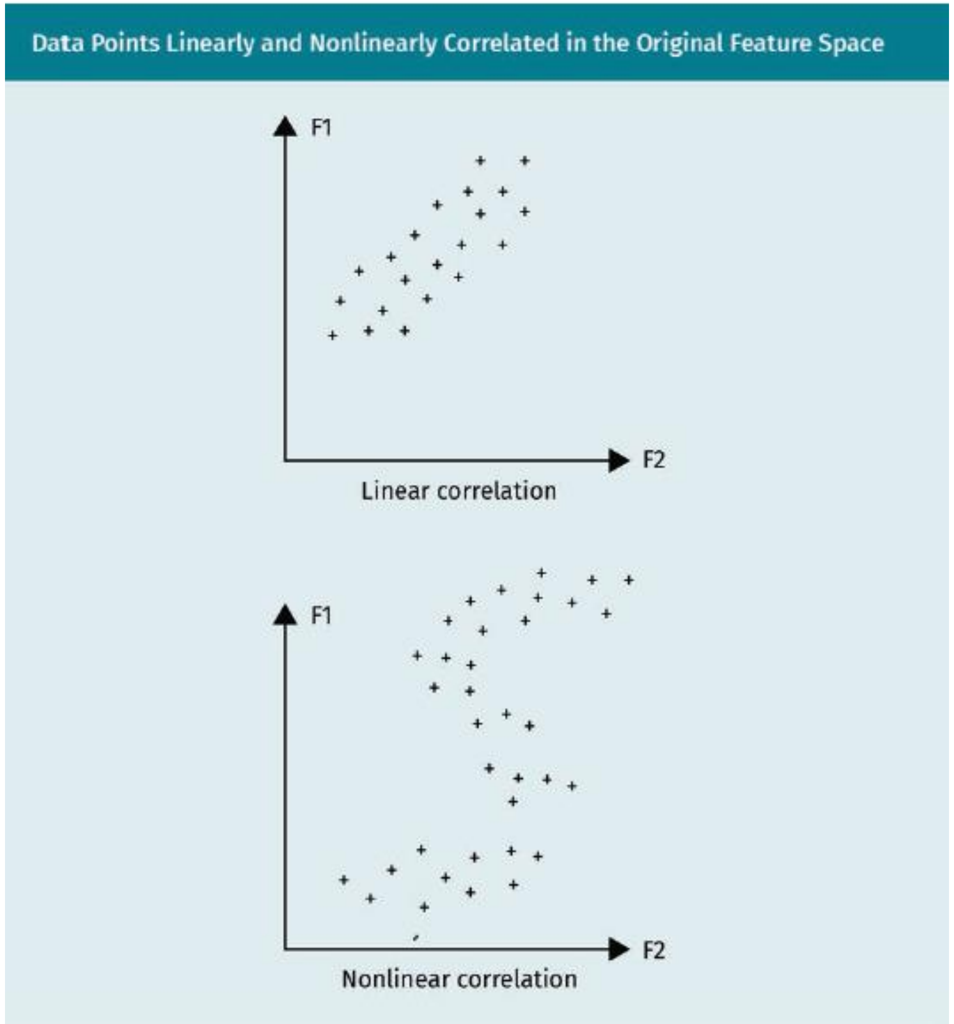
Source of the image: Microsoft Archive.

FEATURE SELECTION



Source of the image: Course book.

- Linear dimensionality reduction methods
  - Principal component analysis (PCA), Factor Analysis, and Linear Discriminant Analysis
  - The original data points are linearly correlated and thus can be linearly transformed and projected into a reduced new feature space.
- Nonlinear dimensionality reduction methods
  - Multi-Dimensional Scaling (MDS), Locally Linear Embedding (LLE), and Kernel PCA
  - The original data points are correlated in the feature space in nonlinear way

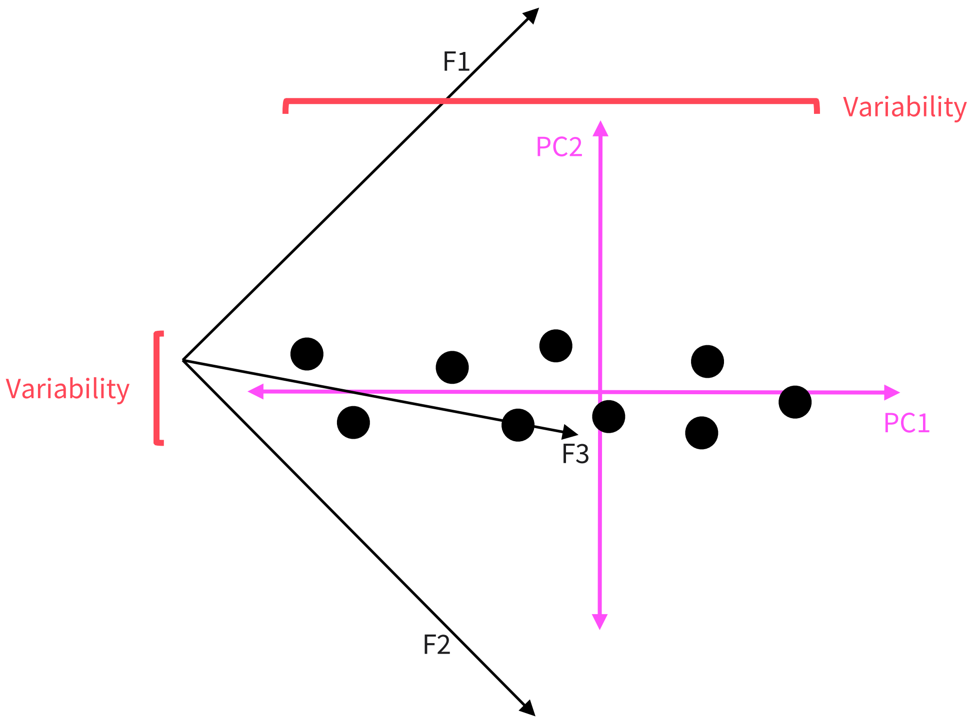


Source of the image: Course book.

Principal Component Analysis (PCA)

- **New axis, maximizing the variance** in the data along this axis (**PC1**).
- **PC2**: Orthogonal to PC1
- …
- **Rotate and center** to PC feature space.
- **PC1** contains **most of the variability.**
- PC2 less than PC1
- …
- **Feature selection**
  - Use PCs for modeling.
  - Use **loading scores** to identify informative original features.

Image. 5: PCA



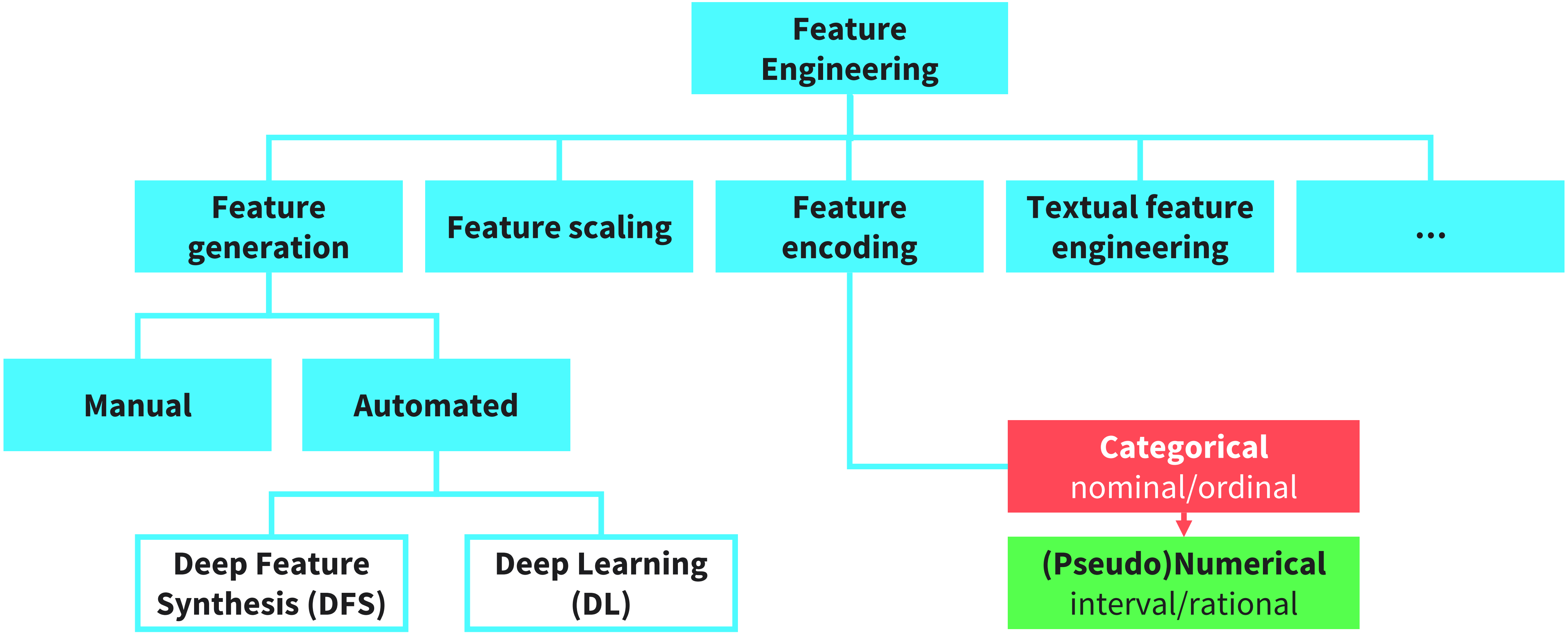
Source of the image: Müller-Kett, 2021.

- The **performance** of machine learning models largely **depends on the input features.**
- **Relevant information** might not be directly **accessible** by the ML algorithms.
- **Expose relevant information** for the modeling step by…
  - Extraction
  - Aggregation
  - Filtering
  - …

Image. 6: Hidden information

Source of the image: Microsoft Archive.

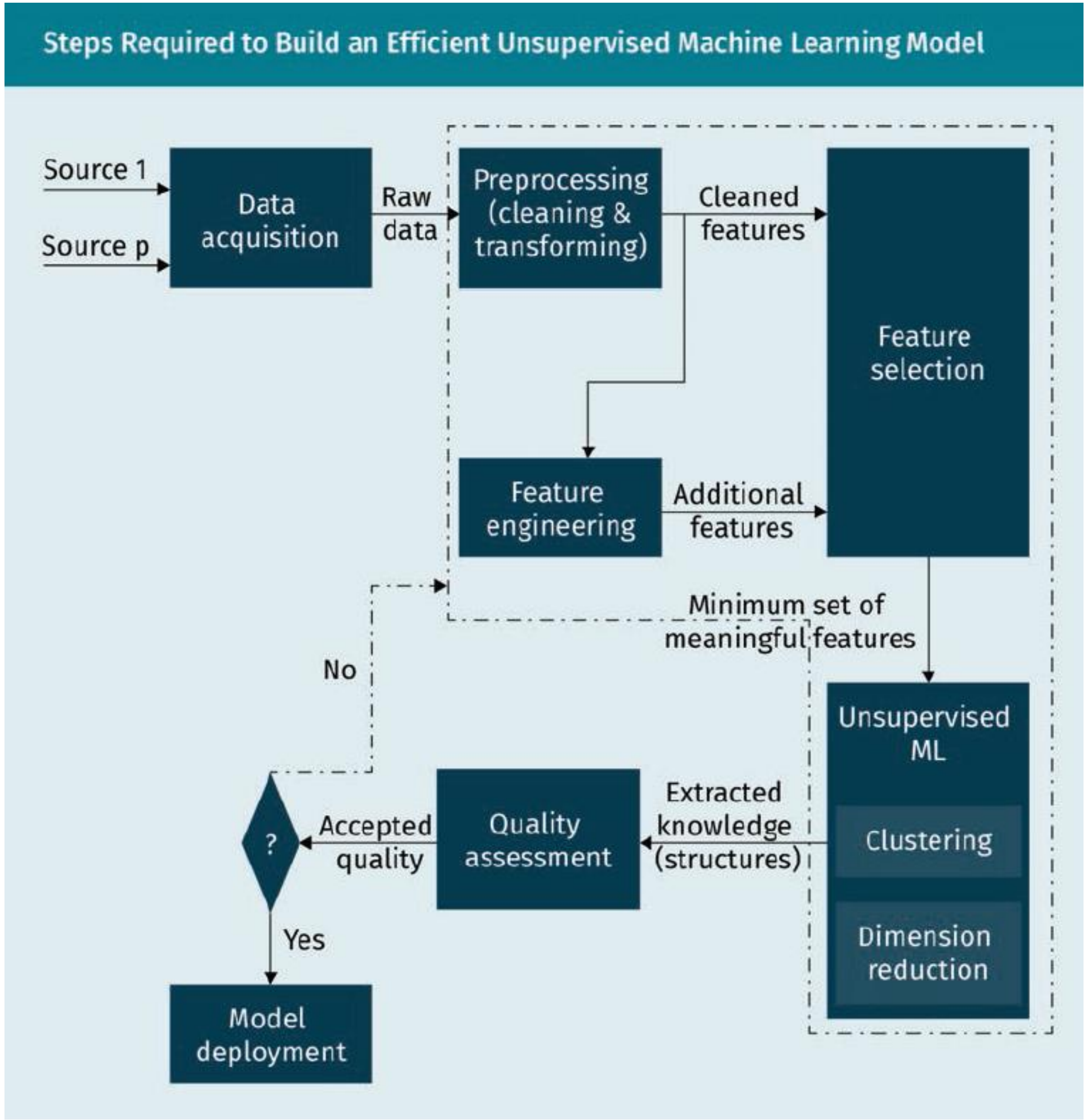
Image. 7: Feature engineering techniques



Source of the image: Müller-Kett, 2021.

STEPS TO BUILD A SUCCESSFUL UNSUPERVISED LEARNING MODEL

- Data acquisition
- Preprocessing
- Feature selection/engineering
- Unsupervised ML
- Quality assessment
- Model deployment



Source of the image: Microsoft archive.

REVIEW STUDY GOALS

- Explain the **general principal** of unsupervised machine learning and its **applications** to real-life problems.
- Define what **features** are, their **types**, their interest for unsupervised machine learning, and their **challenges**.
- Explain the **steps of designing** an unsupervised machine learning **model**.
- **Adapt or transform features** for an unsupervised machine learning model.
- **Evaluate** and **improve** the **performance** of an unsupervised machine learning model.

SESSION 1

TRANSFER TASK

A start-up that sells **sustainable products in smaller stores** has been very successful in recent years. As a result, more stores are to be opened worldwide.

To keep an **overview of the offered products**, you and your team of Data Scientists are tasked to **define homogeneous groups of products** to facilitate ordering, marketing, and distribution.

Create a rough project plan to achieve this goal. **For each phase** of this plan, explain which **unsupervised machine learning and feature engineering techniques** might be applied.

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.
The results will be discussed in plenary.

1. Which of the following techniques is used to transform an original feature space into a new, smaller feature space?
    a) dimensionality reduction
    b) feature selection
    c) hierarchical decomposition
    d) partitioning techniques

2. Which of the following methods is a recommended technique for non-linear dimensionality reduction?
   a) k-means
   b) Gaussian mixture model
   c) Principal components analysis
   d) Multi-dimensional scaling

3. At what point does a learned model become susceptible to overfitting?
   a) When the number of data points is high.
   b) When the number of features grows for a given number of data points.
   c) When the data points are non-linearly correlated.
   d) When the data points are linearly correlated.

LIST OF SOURCES

Images

Müller-Kett, 2018.

Müller-Kett, 2021.

Microsoft Archive.