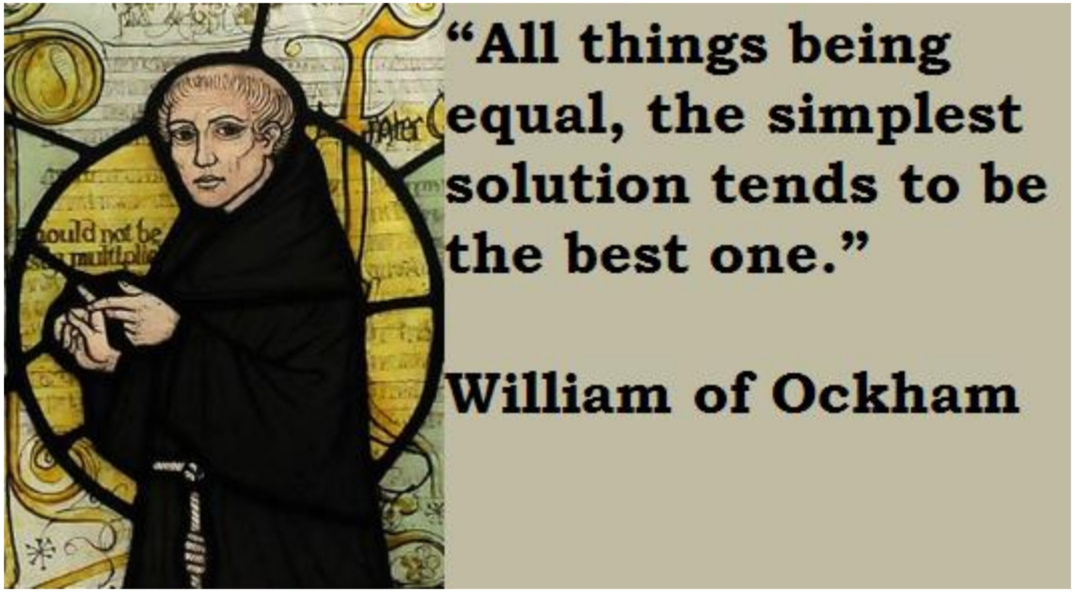


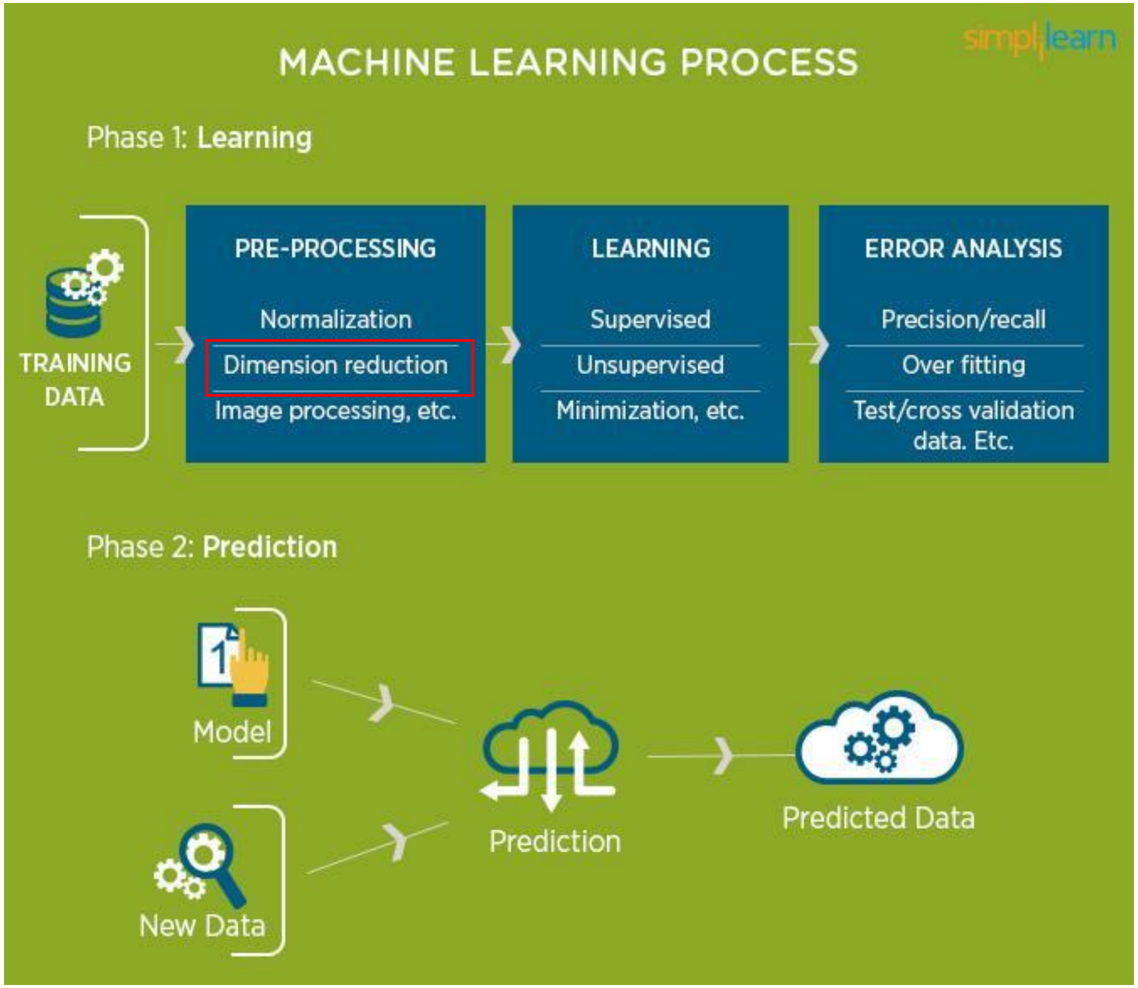
Dimensionality Reduction

Pilsung Kang
School of Industrial Management Engineering
Korea University

AGENDA

Data Analytics Process

- Process of Business Analytics with Machine Learning



https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article

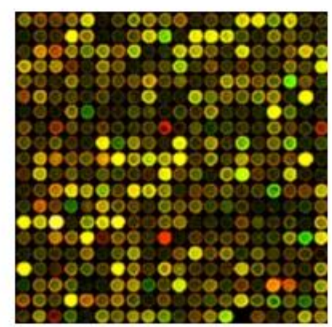
High-dimensional Data

- Examples of high dimensional data

Document classification:
Billions of documents x Thousands/ Millions of words/bigrams matrix

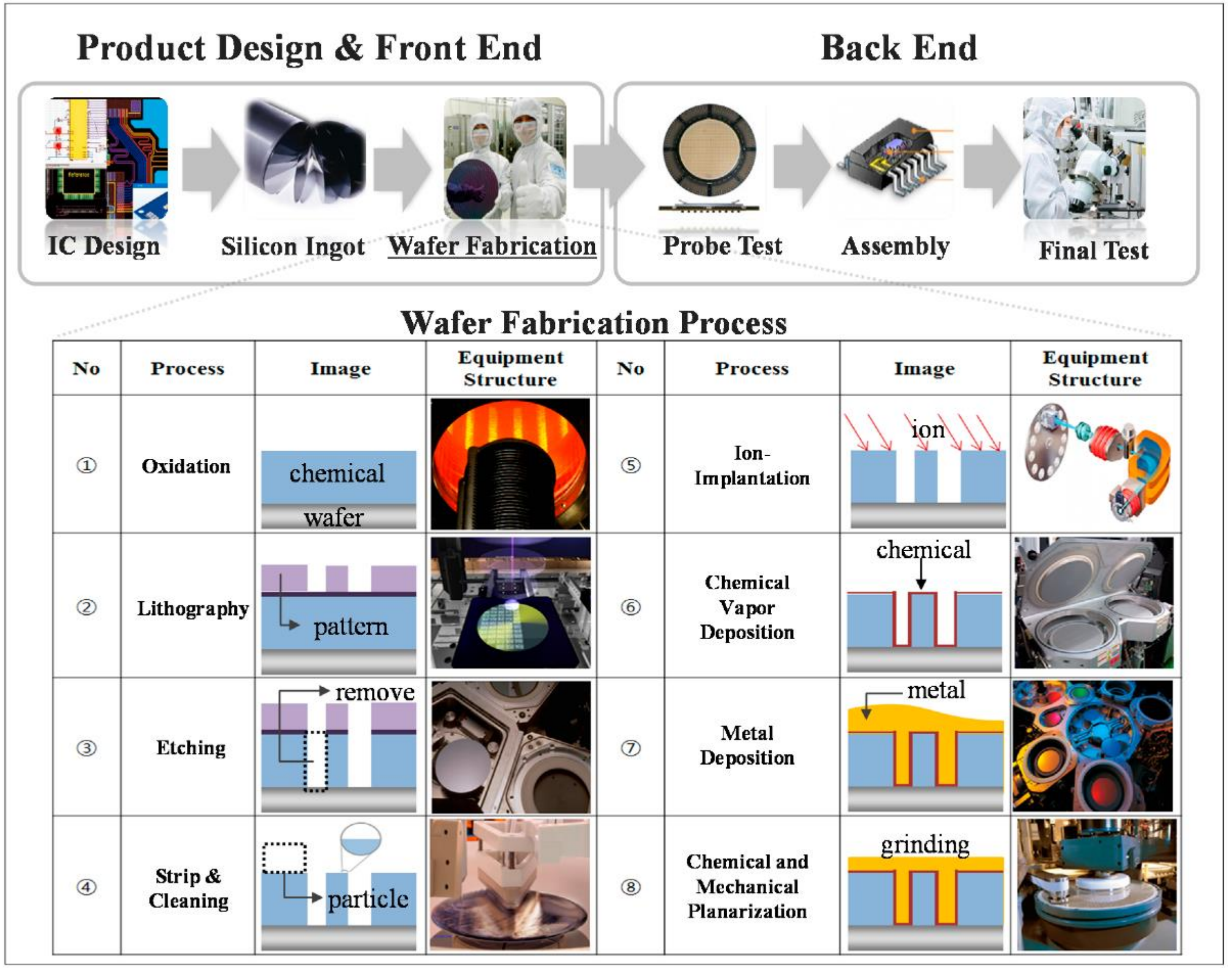
Recommendation systems:
480,189 users x 17,770 movies matrix

Clustering gene expression profiles:
10,000 genes x 1,000 conditions



High-dimensional Data

- Examples of high dimensional data

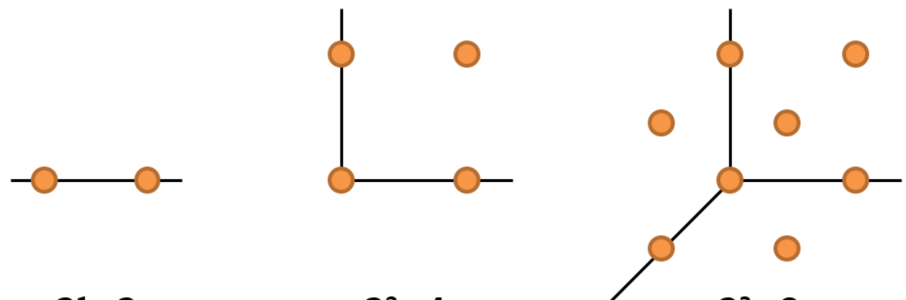


Park, S. H., Kim, S., & Baek, J. G. (2018). Kernel-Density-Based Particle Defect Management for Semiconductor Manufacturing Facilities. Applied Sciences, 8(2), 224.

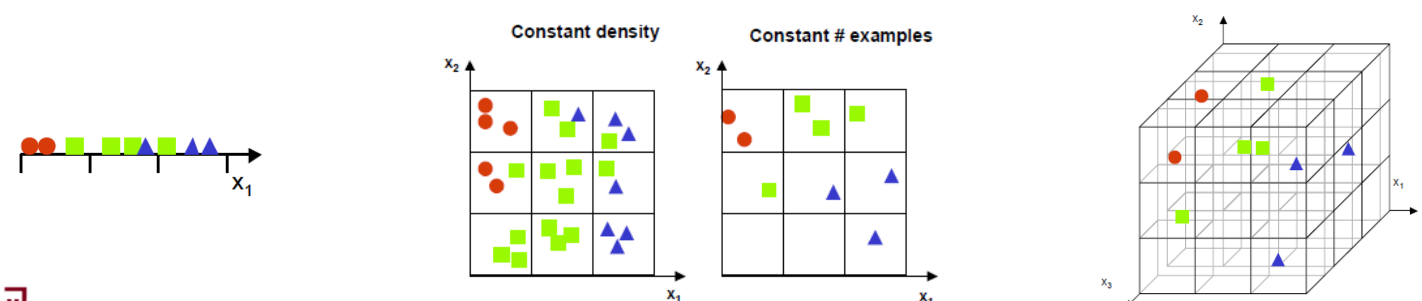
Dimensionality Reduction: Overview

- Curse of dimensionality
  - ✓ The number of instances increases exponentially to achieve the same explanation ability when the number of variables increases

> **"If there are various logical ways to explain a certain phenomenon, the simplest is the best" - Occam's Razor**

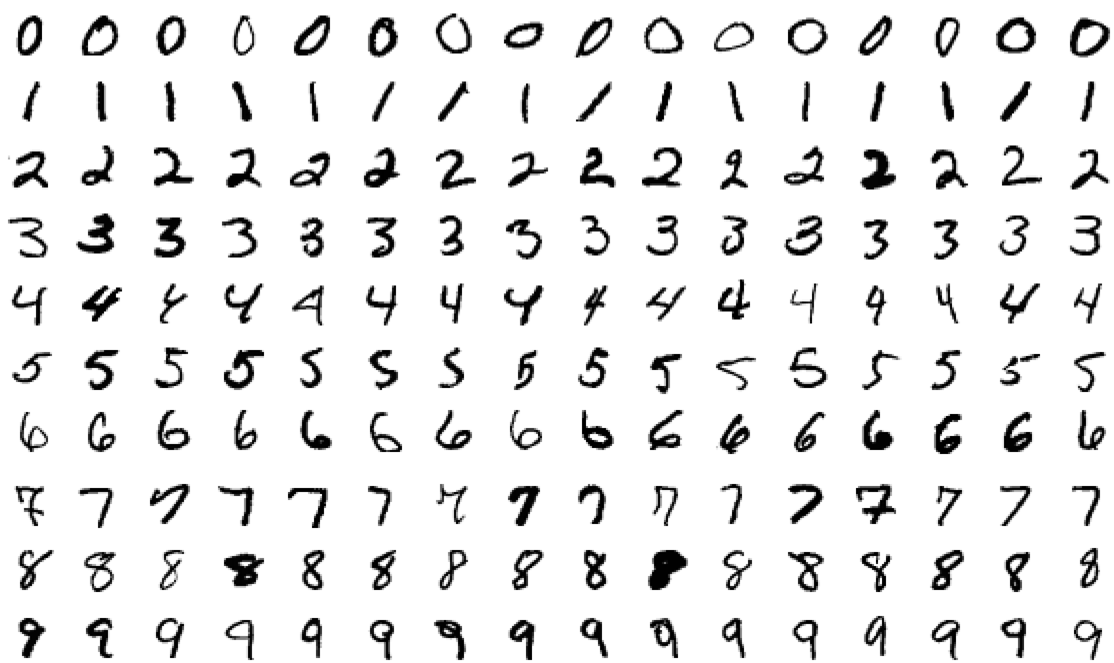


$2^1=2$ $2^2=4$ $2^3=8$



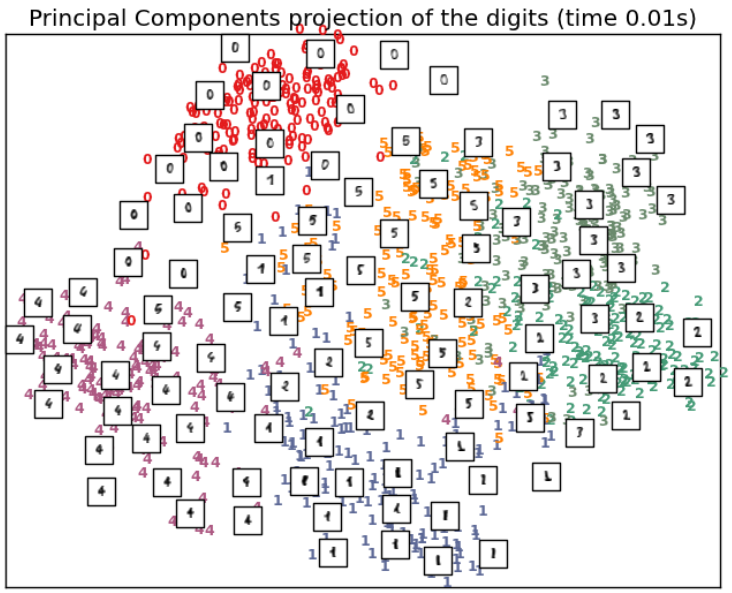
Dimensionality Reduction: Overview

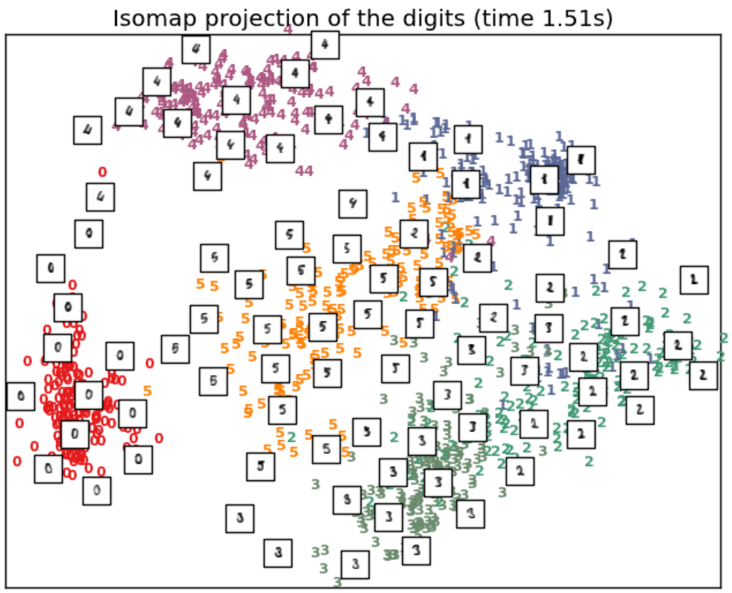
- Curse of dimensionality
  - ✓ Sometimes, an intrinsic dimension is relatively low compared to the original dimension.
    - Ex: handwritten digits in a 16 by 16 pixel (256 dimensions)



Dimensionality Reduction: Overview

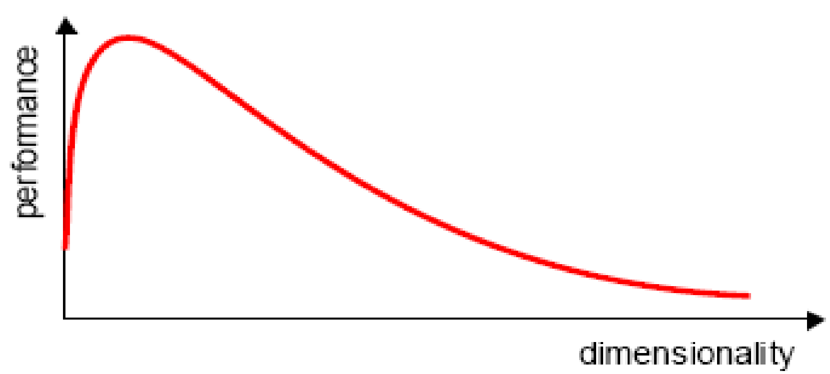
- Curse of dimensionality
  - ✓ Sometimes, an intrinsic dimension is relatively low compared to the original dimension.
    - Ex: handwritten digits in a 16 by 16 pixel (256 dimensions)
    - Reduced to two dimensions by PCA and ISOMAP





Dimensionality Reduction: Overview

- Curse of dimensionality
  - ✓ Problems caused by high-dimensionality
    - Increase the probability of having noise in data → degenerate the prediction performance
    - Increase computational burden for training/applying prediction models
    - Require more number of examples to secure generalization ability of prediction model
  - ✓ To resolve the curse of dimensionality
    - Utilize domain knowledge
    - Use a regularization term in objective function
    - Employ a quantitative reduction technique

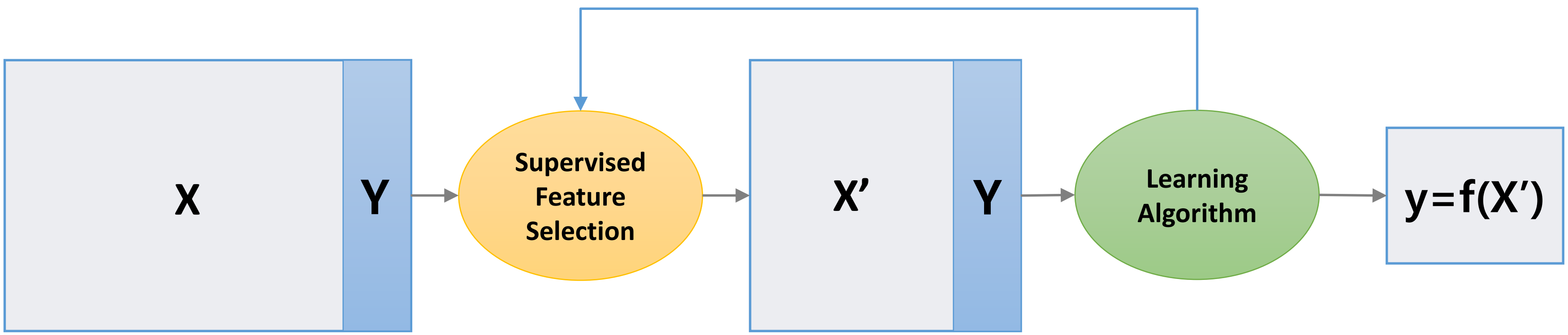


Dimensionality Reduction: Overview

- Backgrounds
  - ✓ Theoretically, model performance improves when the number of variables increases (Under variable independence condition)
  - ✓ In reality, model performance degenerates due to variable dependence, existence of noise, etc.
- Purpose
  - ✓ Identify a subset of variables that best fit the model
- Effect
  - ✓ Remove correlations between variables
  - ✓ Simplified post-processing
  - ✓ Remove redundant or unnecessary variables while keeping relevant information
  - ✓ Visualization can be possible

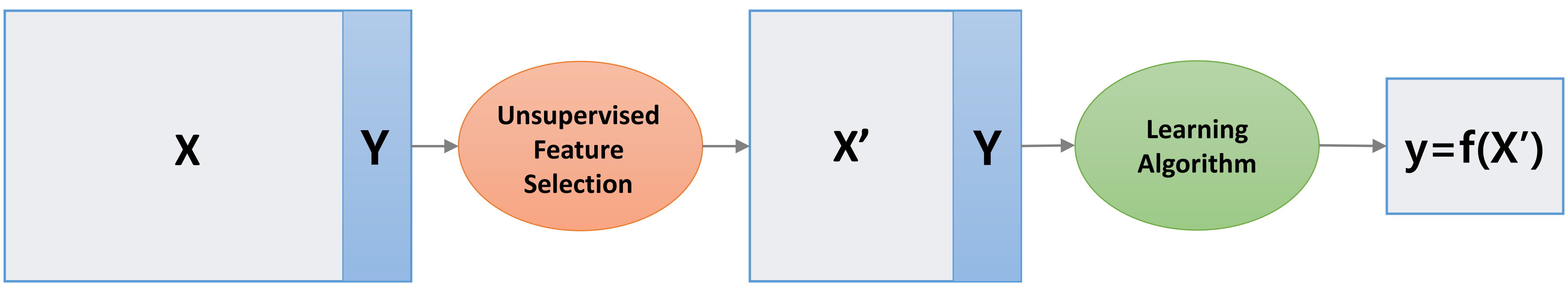
Dimensionality Reduction: Overview

- Supervised vs. Unsupervised Dimensionality Reduction
  - ✓ Supervised dimensionality reduction
    - Use data mining models to verify the reduced dimensions
    - Dimensionality reduction results can be different according to the data mining algorithms employed



Dimensionality Reduction: Overview

- Supervised vs. Unsupervised Dimensionality Reduction
  - ✓ Unsupervised dimensionality reduction
    - Find a set of coordinate systems in a lower dimension that preserve the information (e.g., variance, distance, etc.) in the original input space as much as possible
    - Do not use data mining models during the process
    - Dimensionality reduction results are identical if the data and method is same



Dimensionality Reduction: Overview

- Dimensionality reduction techniques
  - ✓ Variable/feature selection
    - Select a subset of variables from the original variable set
    - Filter – Variable selection and model training are independent
    - Wrapper – Variable selection is done to optimizes the result of the considered data mining model
  - ✓ Variable/feature extraction
    - Extract a new smaller set of variables that preserve the characteristics of the original data
    - Performance metric that is independent from data mining models is used

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \\ x_N \end{bmatrix} \xrightarrow{\text{feature selection}} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \\ x_{i_M} \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \\ x_N \end{bmatrix} \xrightarrow{\text{feature extraction}} \begin{bmatrix} y_1 \\ y_2 \\ \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \\ x_N \end{bmatrix} \right)$$

Dimensionality Reduction: Overview

- Selection vs. Extraction
  - ✓ Conceptual difference between variable selection and variable extraction

X_1	X_2	X_3	…	X_n
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

Variable selection

X_1	X_5	X_8
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…

Variable extraction

Z_1	Z_2	Z_3
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…

$Z_1 = X_1 + 0.2*X_2$

$Z_2 = X_3 - 2*X_5$

$Z_3 = X_4 + X_6 - X_9$


ANY
questions?