Contents

Data Reduction (데이터 축소)

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction?
  - A database may store tera-, peta-, or exa bytes of data.
  - Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - **Dimensionality reduction**, e.g., remove unimportant attributes
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - **Numerosity reduction** (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
  - **Data compression**

Data Reduction 1: Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Distance between points, which is critical to clustering, outlier analysis, becomes less meaningful

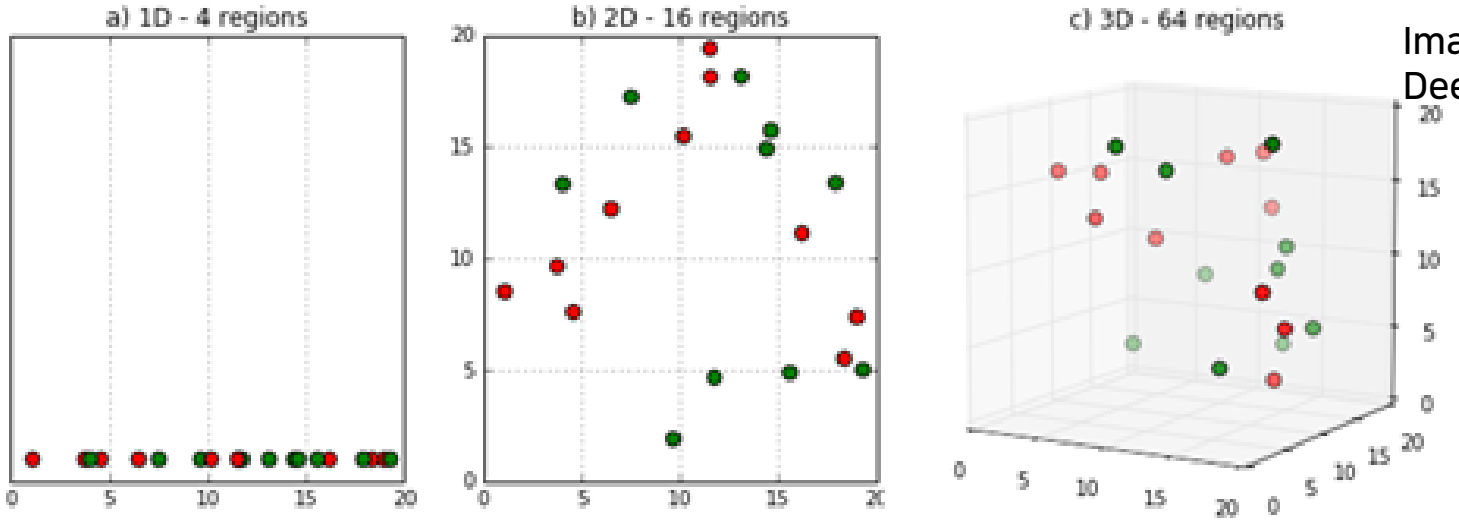


Image source: DeepAI

Num. Dimension smaller larger

Distance more meaningful less meaningful

Data Reduction 1: Dimensionality Reduction

- **Dimensionality reduction (차원 축소)**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

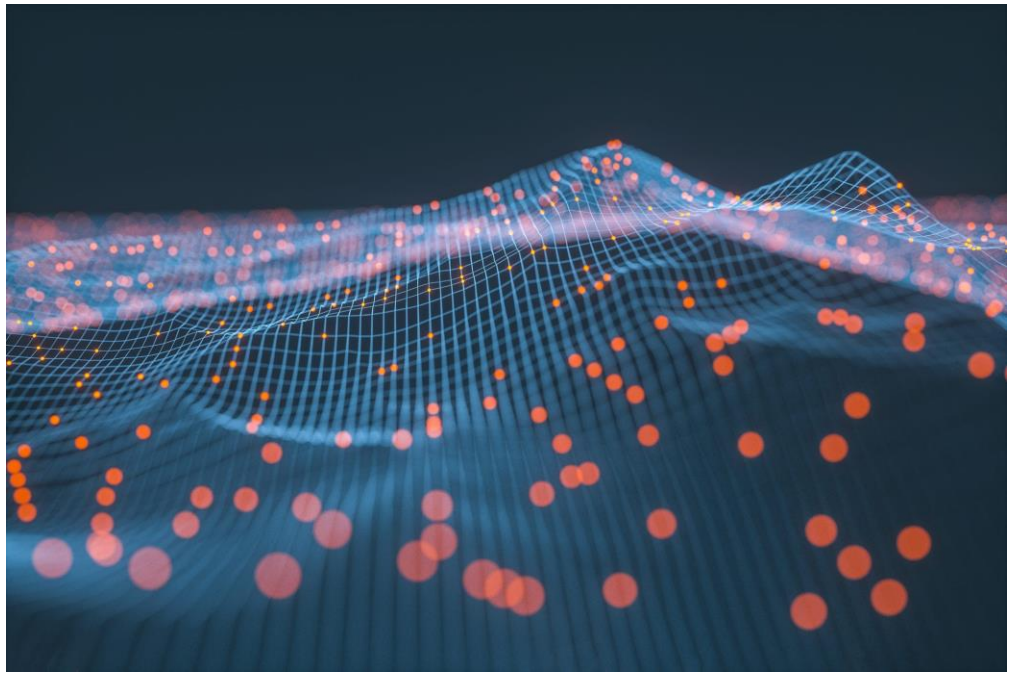
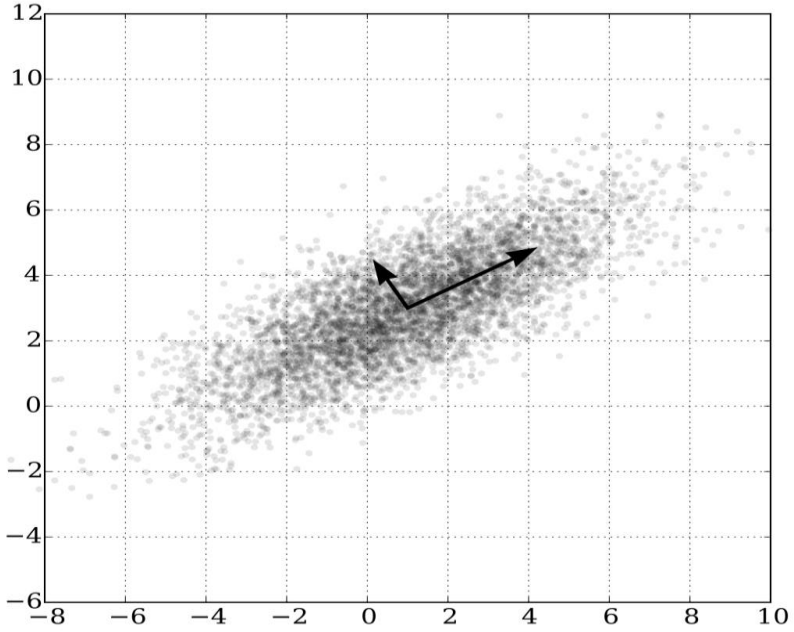


Image source: KX systems

Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



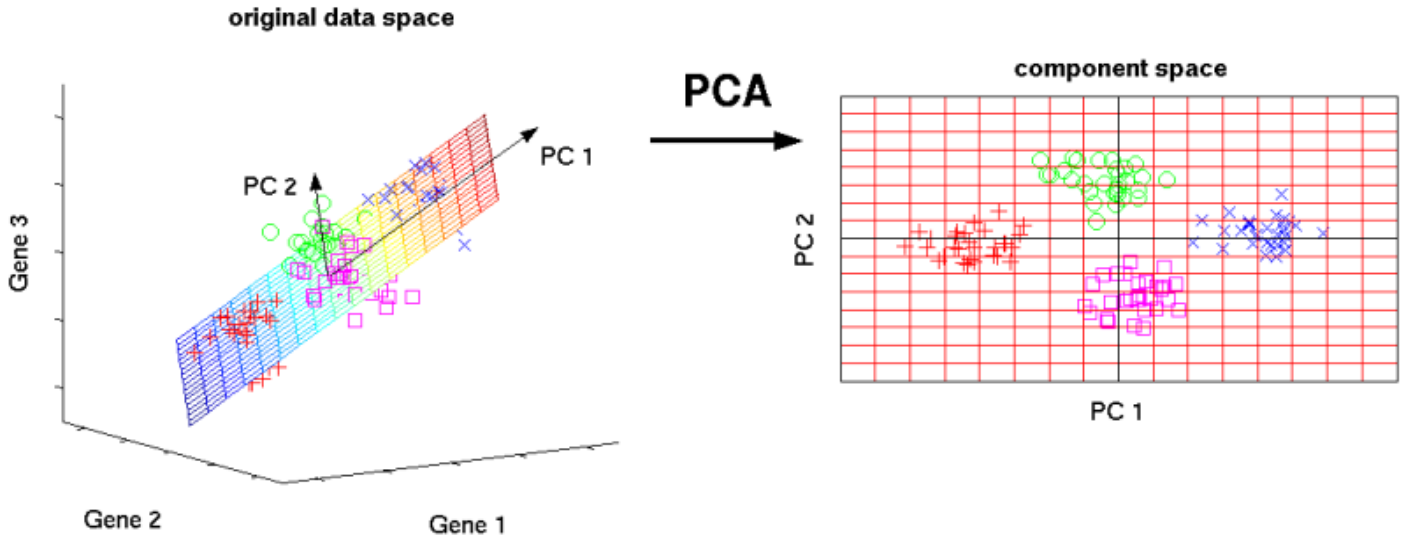


Image source: Wikipedia

PCA

Large Table

X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

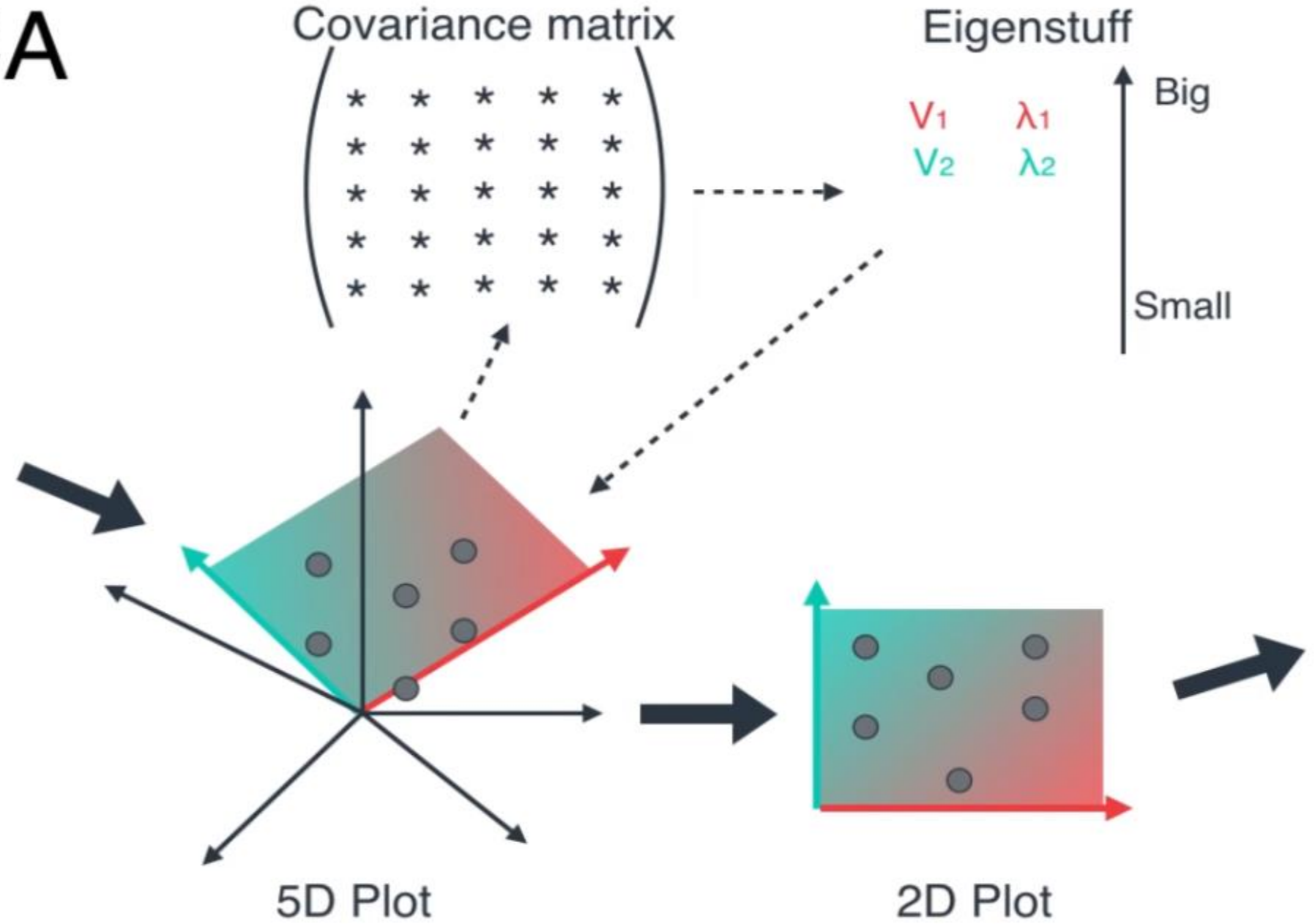


Image source: LinkedIn

Attribute Subset Selection

- Another way to reduce dimensionality of data
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

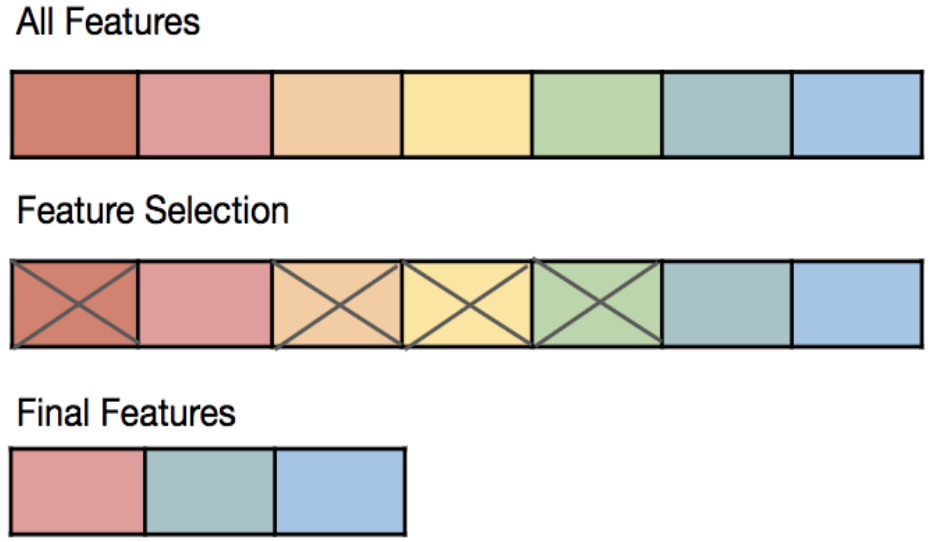


Image source: KDnuggets

Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of d attributes
- Typical methods (heuristic)
  - Step-wise feature selection
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination
    - Starting from all the feature sets, repeatedly eliminate the worst attribute

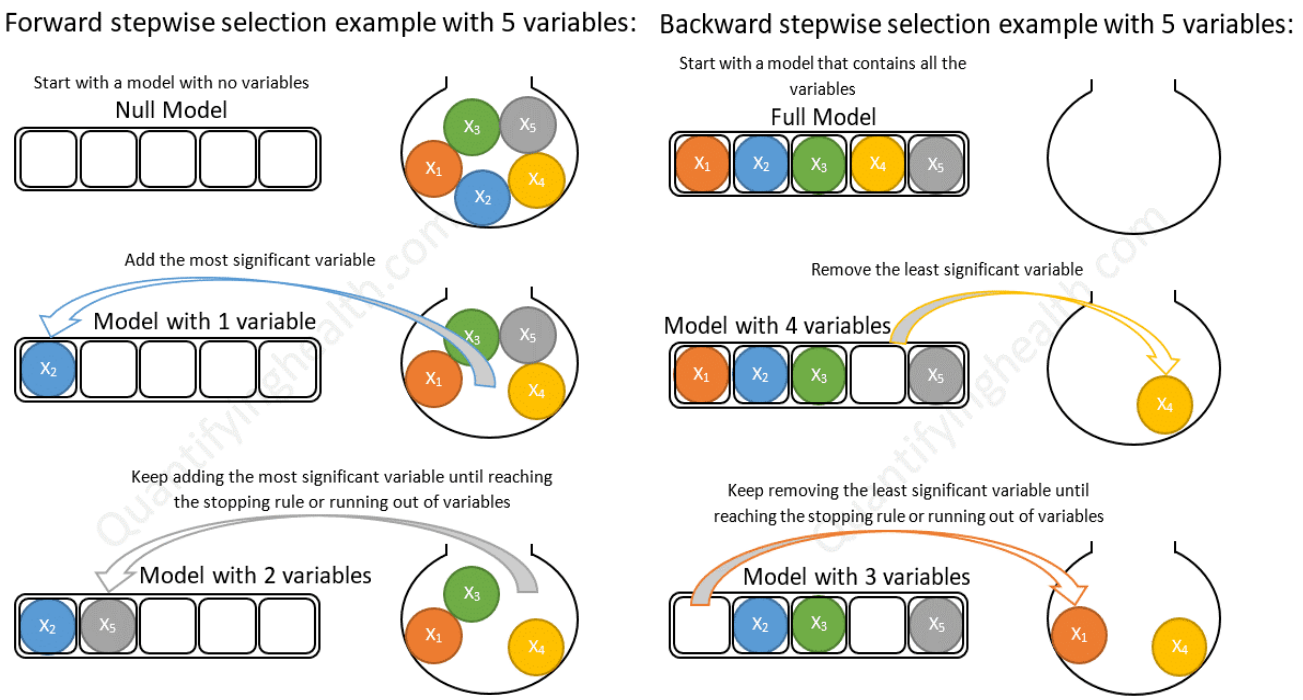


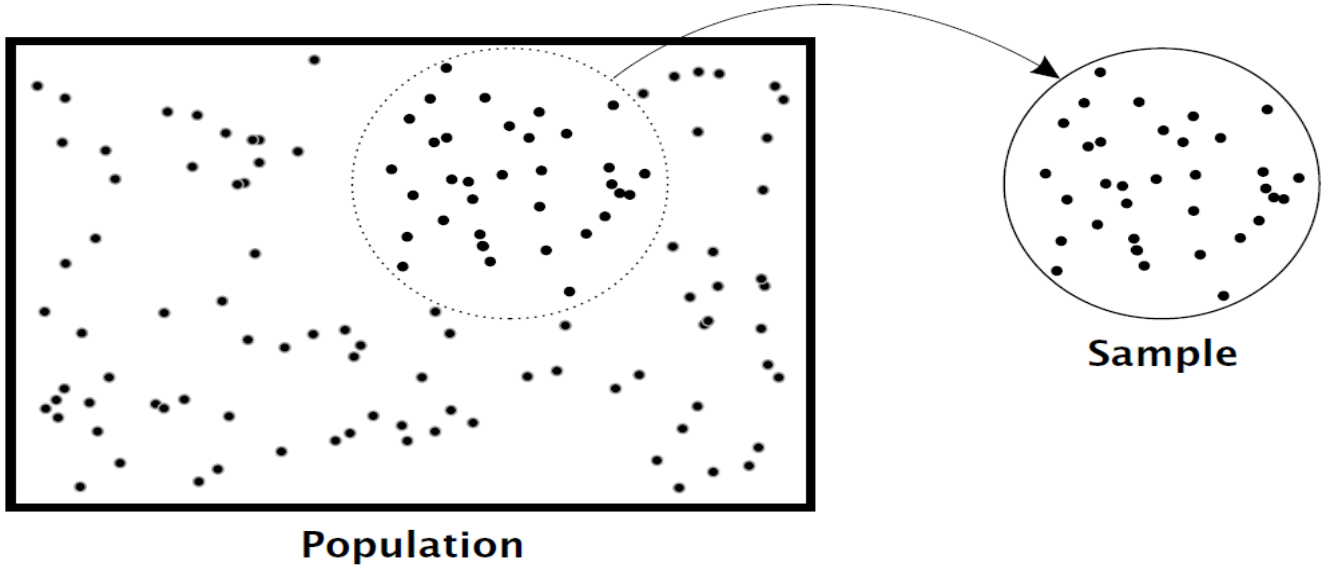
Image source: QUANTIFYING HEALTH

Data Reduction 2: Numerosity Reduction

- Reduce data volume by choosing alternative, *smaller* forms of data representation
- **Parametric methods** (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- **Non-parametric methods**
  - Do not assume models
  - Major families: histograms, clustering, sampling, …

Sampling

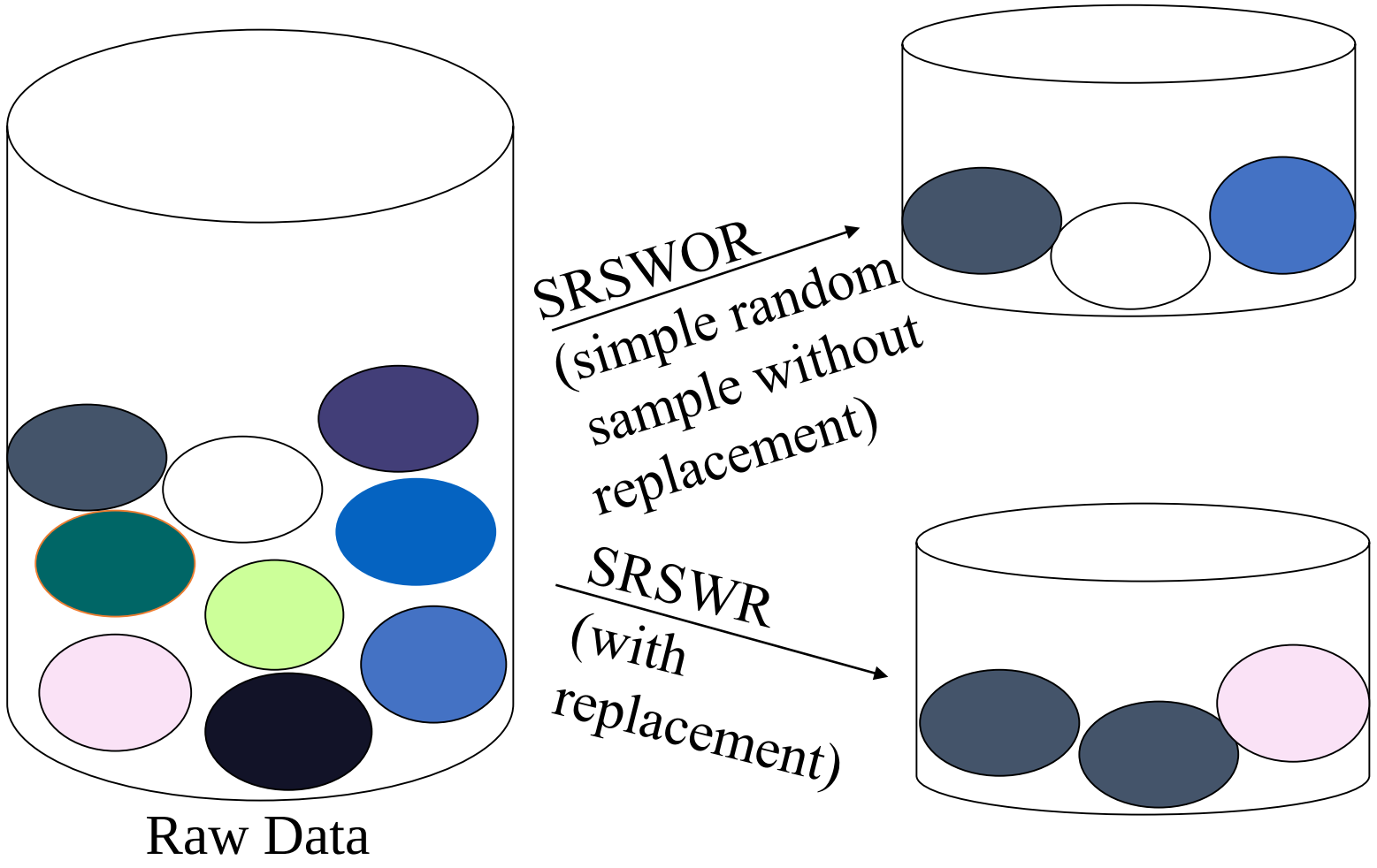
- Sampling: obtaining a small sample s to represent the whole data set N
- **Key principle**: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skewness
  - Develop adaptive sampling methods, e.g., stratified sampling



Types of Sampling

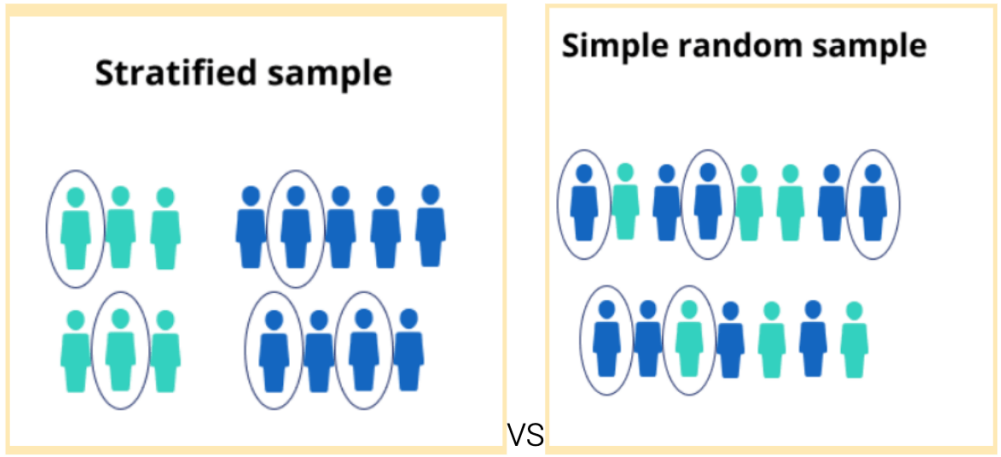
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Sampling: With or without Replacement



Sampling: Stratified Sampling

Motivation



Bias (편향) in the training data:
Training on a skewed data could lead to a poor performance later in the KD steps

Stratified sampling can reduce the bias of selected samples

Process

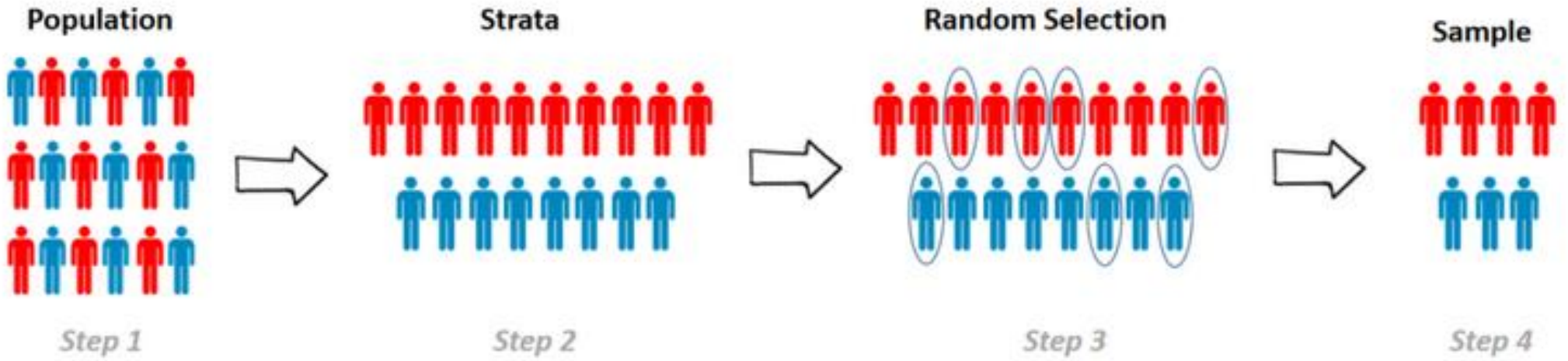
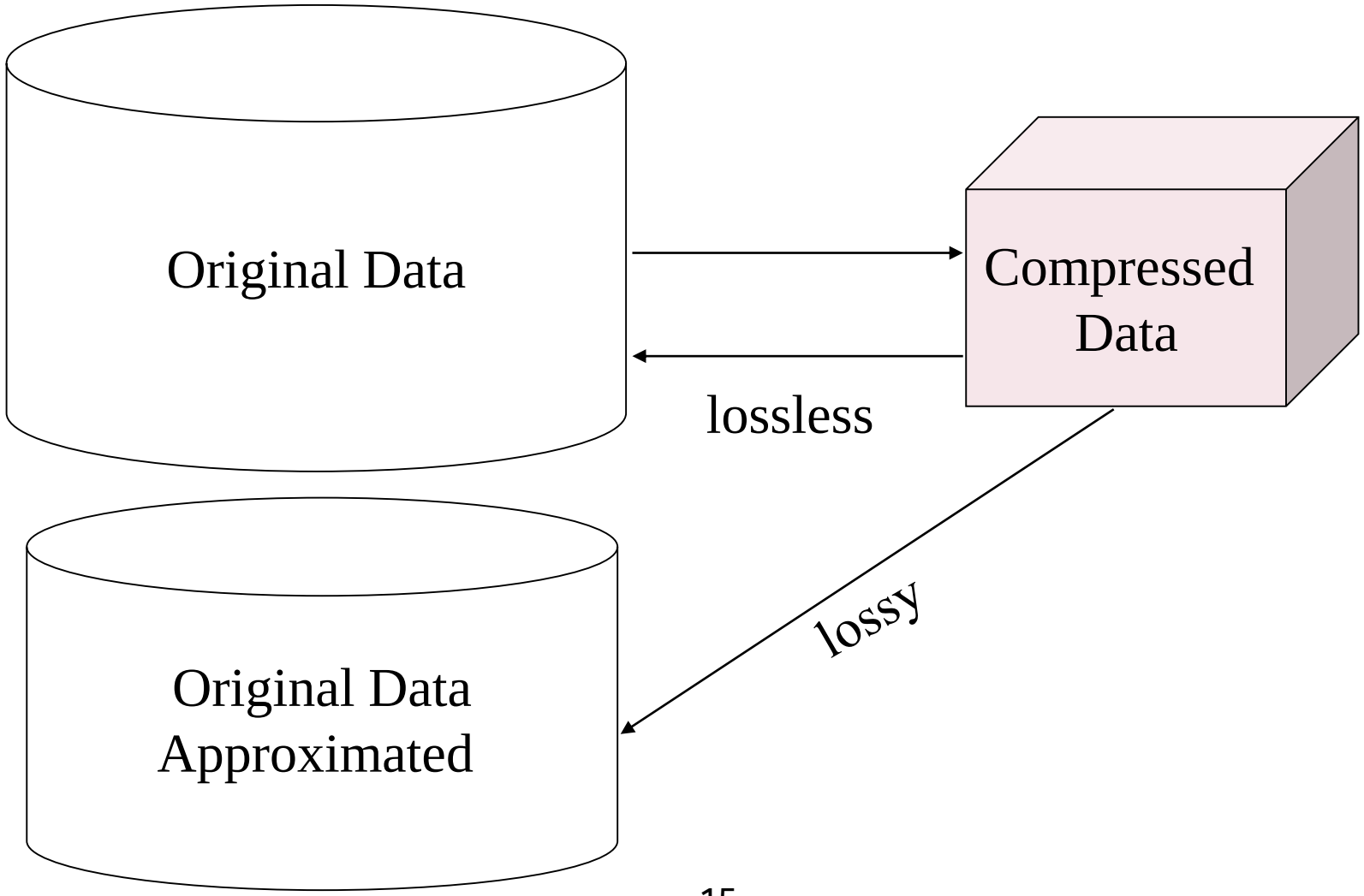


Image source: Qualtrics, Uedufy

Data Reduction 3: Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically **lossless**, but only limited manipulation is possible without expansion
- Audio/video compression
  - Typically **lossy** compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Dimensionality and numerosity reduction may also be considered as forms of data compression

Data Compression



Recommended reading

- Large language models (LLMs) such as GPT are lossy compression of WEB
- If you are interested in the basic mechanism of ChatGPT, check the inspiring article in the New Yorker by Ted Chiang (SF writer)
- https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

Contents

Data Transformation (데이터 변환)

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - **Smoothing**: Remove noise from data
  - **Attribute/feature construction**: New attributes constructed from the given ones
  - **Aggregation**: Summarization
  - **Normalization**: Scaled to fall within a smaller, specified range
  - **Discretization**
- Data transformation may make the resulting mining process more efficient

Normalization

- **Min-max normalization**: to $[new\_min_A, new\_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to 0.716. $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ where j is the smallest integer such that $\max(|v'|) < 1$

Discretization (이산화)

- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)

Simple Discretization: Binning

- **Equal-width** (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
- **Equal-depth** (frequency) partitioning
  - Divides the range into N intervals, each containing approx. same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
  * Partition into equal-frequency (**equal-depth**) bins:
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
  * Smoothing by **bin means**:
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29
  * Smoothing by **bin boundaries**:
    - Bin 1: 4, 4, 4, 15
    - Bin 2: 21, 21, 25, 25
    - Bin 3: 26, 26, 26, 34

Summary

- **In the preprocessing steps, we want to improve data quality**:
  accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values
- **Data integration** from multiple sources: e.g., remove redundancies
- **Data reduction**:
  Dimensionality reduction, Numerosity reduction, Data compression
- **Data transformation and data discretization**