



Exploratory Data Analysis

Nguyen Ngoc Thao
nnthao@fit.hcmus.edu.vn

Content outline

- Data quality
- Major tasks in Data preprocessing

Data quality

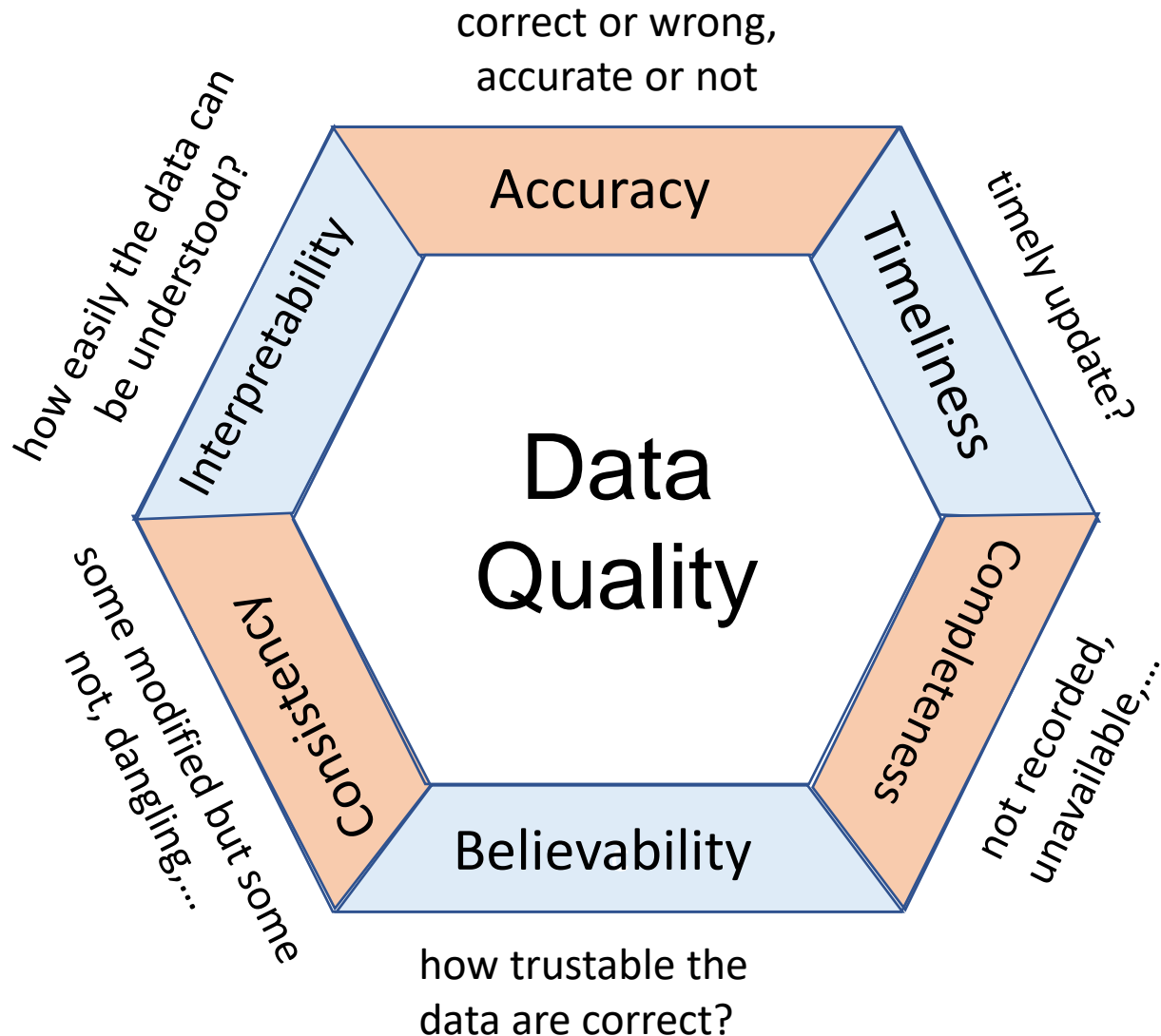
An example of data analytics



- A branch manager analyzes the sales data by inspecting the company's data warehouse to include the necessary attributes.
- HOWEVER, the data being considered has many problems
 - Information needed for the analysis has not been recorded.
 - Many errors and unusual values for some transactions have been reported.

Inaccurate, incomplete, and inconsistent data are commonplace properties of large real-world databases and data warehouses.

Measures of data quality

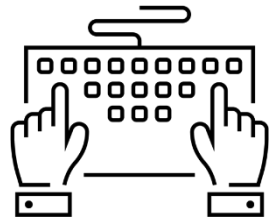


Data accuracy

- **Inaccurate data** means having **incorrect attribute values**.

Incorrect values submitted for mandatory fields

- E.g., negative weight, inappropriate range of ages, etc.
- *Disguised missing data*: many users have the same birthday, e.g., Jan 01

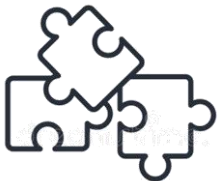


Faulty data collection instruments

Data transmission errors due to technology limitations



- E.g., limited buffer size for coordinating synchronized data transfer



Incorrect data may also result from inconsistencies

Data completeness

- The attributes of interest may not always be available or contain only aggregated data.
 - E.g., study the shopping habits in festive seasons while only the annual sales are available
- Many causes are leading to missing data.
 - Equipment malfunction
 - Some records are deleted due to inconsistency with other records.
 - Data is not entered due to misunderstanding.
 - Certain data may not be considered vital at the time of entry.
 - The recording of data history may have been overlooked.

Data consistency

- **Inconsistencies** in naming conventions or data codes
 - E.g., USA vs. US, alternative name (Bill Clinton vs. William Clinton), author name in reference: Li Fei-Fei vs. Fei-Fei, L.
- **Incompatible formats** for input fields
 - E.g., datetime format (dd/mm/yy vs. mm/dd/yy), rating scale ([1..5] vs. [1..10]), decimal and thousand separators
- **Duplicate tuples** also require data cleaning.

Data timeliness

- Suppose you are overseeing the data of monthly sales
- For a while after each month, the data stored is incomplete.
 - Several sales representatives fail to submit their sales records on time at the end of the month.
 - There are also some corrections and adjustments flowing in after the month's end.
- However, once all the data is received, it is correct.
- The month-end data are **not updated in a timely fashion**, harming the data quality.

Believability and Interpretability

- Suppose that a database, at one point, had several errors, all of which have since been corrected.



Believability: how trustily is the data correct?

- E.g., the past errors had caused many problems for sales department users → they no longer trust the data



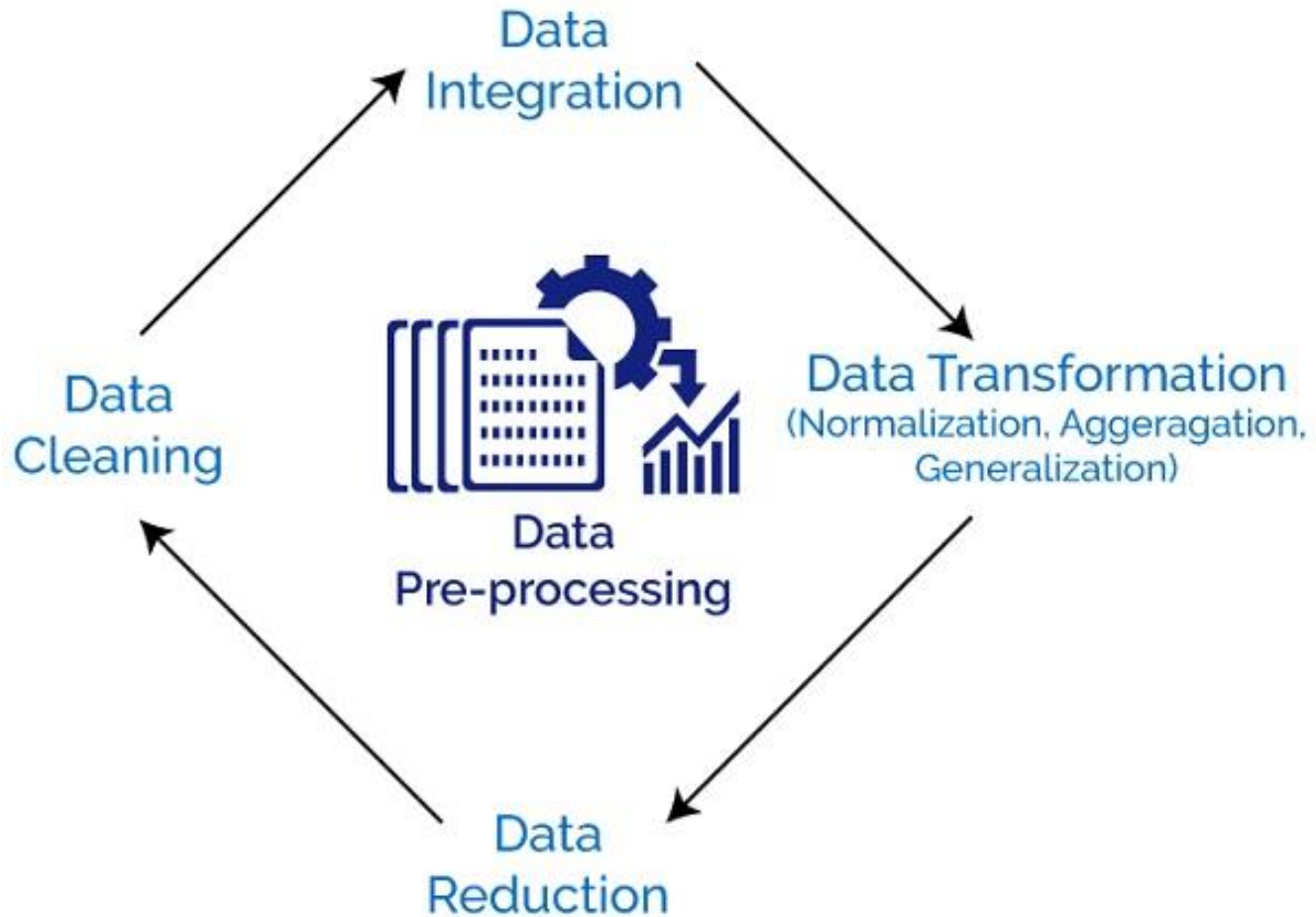
Interpretability: how easily is the data interpreted?

- E.g., the data use many accounting codes → the sales department does not know how to figure out

Data quality is subjective

- Data quality depends on the intended use of the data.
- Two users may assess the quality of a database differently.
- Consider a database in which some customer addresses are outdated or incorrect, yet overall, 80% of them are accurate.
 - A marketing analyst considers the database to be accurate enough for target marketing purposes.
 - However, a sales manager may consider the data inaccurate.

Major tasks in Data preprocessing





Data cleaning

How to handle missing data?



Ignore the tuple

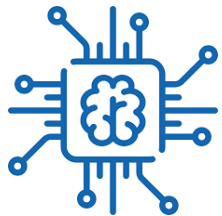
- Usually done when class label is missing
- Not effective when the percentage of missing values per attribute varies considerably

Fill in the missing value manually

- Tedious and infeasible



Fill in it automatically with



- A global constant, e.g., “unknown” or a new class
- The attribute mean (for all samples of the same class)
- The most probable value: Bayesian approach or decision tree

How to handle noisy data?

Binning and smoothing

- First sort data and partition into (equal-frequency) bins
- Then smooth each bin by its mean, median, or boundary, etc.

Regression

- Smooth by fitting the data into regression functions

Clustering

- Detect and remove outliers

Hybrid

- Suspicious values are detected by computers and checked by human

Binning and smoothing: An example

- Consider the following sorted data points

4 8 15 21 21 24 25 28 34

- Partition into equal-frequency bins

Bin 1: 4 8 15

Bin 2: 21 21 24

Bin 3: 25 28 34

- Smooth the bins

Bin 1: 9 9 9

Bin 2: 22 22 22

Bin 3: 29 29 29

By means

Bin 1: 8 8 8

Bin 2: 21 21 21

Bin 3: 28 28 28

By medians

Bin 1: 4 4 15

Bin 2: 21 21 24

Bin 3: 25 25 34

By bin boundaries

Quiz 01: Binning and Smoothing

1. Consider the following 1D data series, which includes 15 data points sorted in ascending order.

21, 25, 27, 29, 32, 36, 36, 48, 67, 80, 84,
85, 89, 92, 97

- Apply equal-frequency and equal-width binnings to the given data series to obtain three bins of data points.
 - Apply smoothing on the bins obtained from equal-width binning.
2. How to perform binning and smoothing, without coding from scratch, in **Python**?



Data integration

Entity identification problem

- Entity identification problem arises during integration.
- Identify real world entities from multiple data sources
 - Differences in representation, scaling, or encoding
 - E.g., metric units in British system and other systems, currencies, grading scheme between schools, time format, etc.
- Matching attributes from one database to another following the ontological structure.
 - An attribute in one system is recorded at, say, a lower abstraction level than the “same” attribute in another.
 - E.g., “Total sales” may refer to one branch or to all stores in a region.
- Careful integration helps improve mining speed and quality.

How to handle redundancy?

- Redundant data often occur when integrating databases.
- **Object identification:** The same attribute or object may have different names in various databases.
 - E.g., the occupation information may be stored in column “job” of the first database and column “career” of the second database.
- **Derivable data:** An attribute is derived from other attributes.
 - E.g., the annual revenue is the sum of monthly revenues.

χ^2 statistics for correlation analysis

- Suppose attribute A has c distinct values and attribute B has r distinct value. There are n data tuples.
- Let (A_i, B_j) denote the joint event that $A = a_i$ and $B = b_j$.
- χ^2 statistic tests the null hypothesis, *A and B are independent*

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- o_{ij} : observed frequency (i.e., actual count) of $(A = a_i, B = b_j)$
- e_{ij} : expected frequency of (A_i, B_j) $e_{ij} = \frac{\text{count}(A=a_i) \times \text{count}(B=b_j)}{n}$
- The larger χ^2 value, the more likely the variables are related.

χ^2 statistics: An example

- Consider the below a contingency table.

	<i>male</i>	<i>female</i>	<i>Total</i>
<i>fiction</i>	250 (90)	200 (360)	450
<i>non_fiction</i>	50 (210)	1000 (840)	1050
Total	300	1200	1500

(Numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

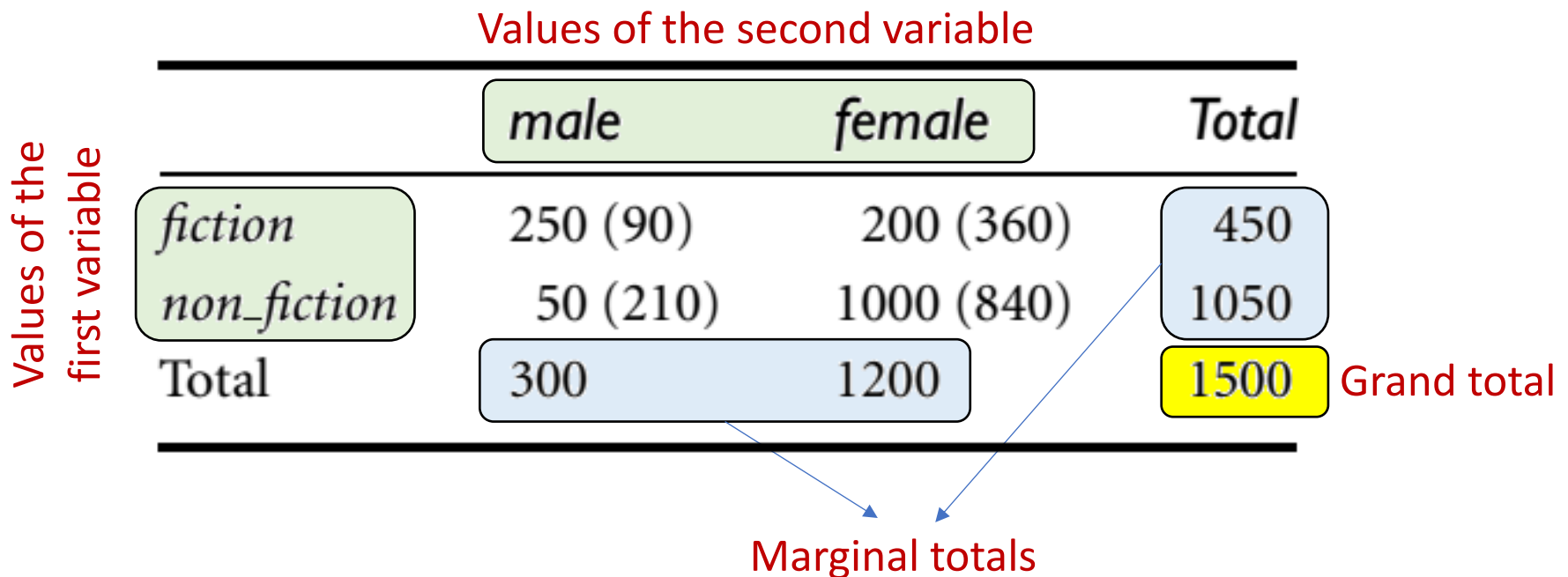
- Are *gender* and *preferred_reading* correlated?

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

- Two attributes are (strongly) correlated for the given group of people
- However, **correlation does not imply causality**.
 - # of hospitals and # of car-theft in a city are correlated
 - However, both are causally linked to the third variable – population.

χ^2 statistics: Contingency table

- A **contingency table** (or **crosstab**) displays the **frequency distribution** of two or more categorical variables.
- It helps to **analyze the relationship between variables**.



The diagram shows a contingency table with the following structure:

		Values of the second variable		
		male	female	Total
Values of the first variable	fiction	250 (90)	200 (360)	450
	non_fiction	50 (210)	1000 (840)	1050
	Total	300	1200	1500

Annotations:

- A vertical label "Values of the first variable" is positioned to the left of the first column.
- A horizontal label "Values of the second variable" is positioned above the first two columns.
- Blue arrows point from the "Total" row and the "Total" column to the label "Marginal totals" at the bottom.
- A yellow box highlights the "Grand total" value of 1500, with a label "Grand total" to its right.

χ^2 statistics: Contingency table

- Is a contingency table able to represent categorical variables that have more than two values?*

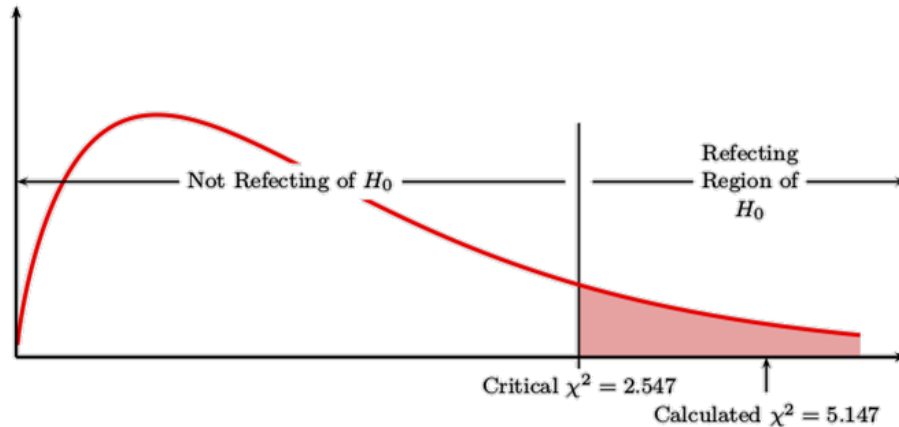
	18-29	30-49	50+	Row Total
Coffee	30	40	20	90
Tea	20	35	25	80
Juice	25	15	10	50
Column Total	75	90	55	220

χ^2 statistics: Contingency table

- Can a contingency table represent the relationship of more than two categorical variables?*

		Response (Y)	
Clinic (Z)	Drug Treatment (X)	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32

χ^2 statistics: An example



- The test is based on a significance level with a DOF of 1.
- If the hypothesis is denied, A and B are statistically correlated.

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
Non-significant									Significant		

χ^2 statistics: Degree of freedom (DOF)

- **DOF** is the number of independent values that can vary in a statistical calculation without breaking constraints.
- It can be calculated from the contingency table as follows.

$$DOF = (r - 1) \cdot (c - 1)$$

- r is the number of rows and c is the number of columns.
- E.g., in a table with 3 rows and 2 columns, $DOF = (3 - 1)(2 - 1) = 2$.

χ^2 statistics: Significant levels

- The **significance level α** is the probability threshold to decide **whether to reject the null hypothesis**.
- It represents the risk of rejecting a true null hypothesis.
- **Common significance levels**
 - **0.05 (5%)**: The most common choice, indicating a 5% risk of wrongly rejecting the null hypothesis.
 - **0.01 (1%)**: Used for stricter criteria, implying a 1% risk.
 - **0.10 (10%)**: Sometimes used in exploratory studies where a higher risk is acceptable.

Quiz 02: χ^2 statistics

1. Consider the data that relate the sex of children in families who have two children. Apply χ^2 statistics at the 0.001 significance level.

		First child		Total
		Male	Female	
Second child	Male	114 (120.54)	131 (124.46)	245
	Female	132 (125.46)	123 (129.54)	255
Total		246	254	500

Numbers out of parenthesis are actual counts from observations and numbers in parenthesis are expected counts calculated based on the data distribution in the two categories

2. How to perform χ^2 statistics in **Python**?

Pearson correlation coefficient

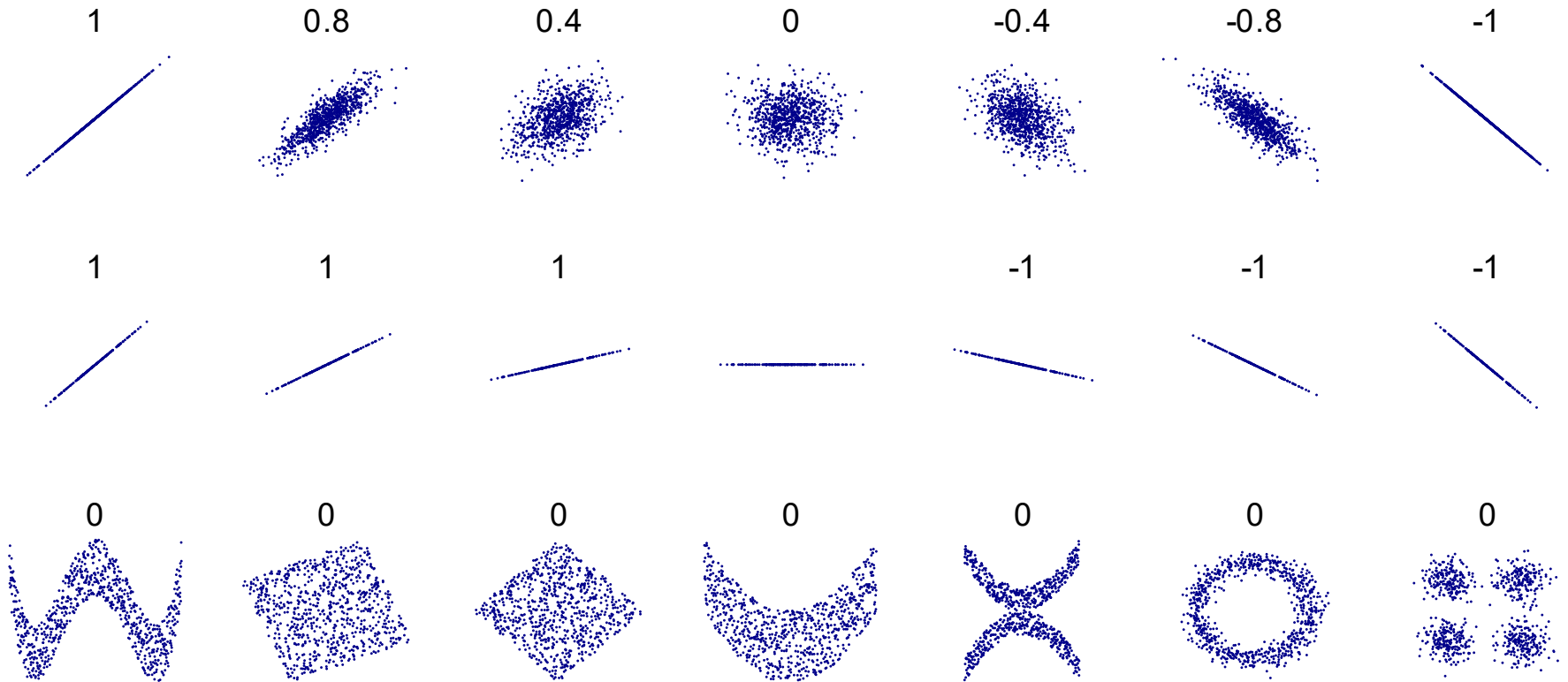
- Consider two numeric attributes A and B , and a set of n observations $\{(a_1, b_1), \dots, (a_n, b_n)\}$.
- **Pearson's product moment coefficient**

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{(\sum_{i=1}^n a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

- $\bar{A}, \bar{B}, \sigma_A, \sigma_B$: means and standard deviations of A and B , respectively
- $\sum a_i b_i$: sum of the AB cross-product

$-1 \leftarrow r_{A,B}$	$r_{A,B} = 0$	$r_{A,B} \rightarrow 1$
Negative correlation	A and B are independent	Positive correlation

Pearson correlation coefficient



Several sets of (x, y) points, with the Pearson correlation coefficient of x and y for each set. The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero. ([Wikipedia](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient))

Covariance analysis

- The **covariance between A and B** is defined as

$$Cov(A, B) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n} = E(A \cdot B) - \bar{A}\bar{B}$$

- where $E(A) = \bar{A} = \frac{\sum_{i=1}^n a_i}{n}$ and $E(B) = \bar{B} = \frac{\sum_{i=1}^n b_i}{n}$ are the expected values of A and B

$Cov(A, B) > 0$	$Cov(A, B) < 0$	$Cov(A, B) = 0$
Positive covariance	Negative covariance	A and B are independent

- Covariance vs. correlation: $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

Covariance analysis: An example

- If the stocks are affected by the same industry trends, will their prices rise or fall together?

Stock Prices for *AllElectronics* and *HighTech*

<i>Time point</i>	<i>AllElectronics</i>	<i>HighTech</i>
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

- $E(\text{AllElectronics}) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \4
- $E(\text{HighTech}) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \10.80
- $\text{Cov}(\text{AllElectronics}, \text{HighTech}) = \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80 = 7$
- Therefore, a positive covariance indicates that stock prices for both companies rise together

Quiz 03: Correlation metrics

1. Consider the following data table, in which there are five tuples of two attributes, A and B.

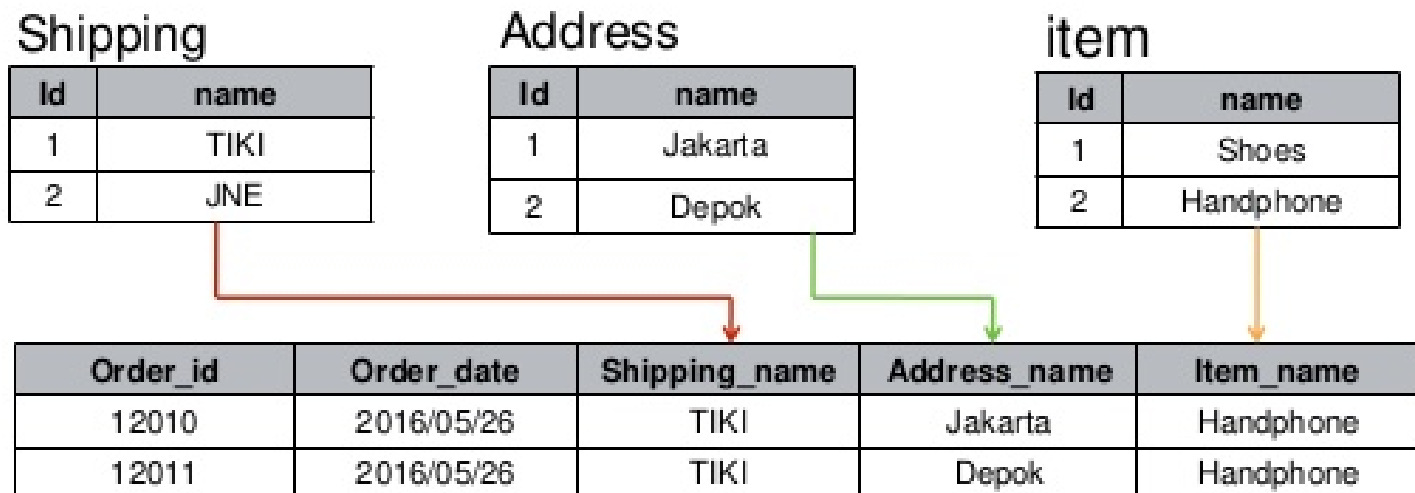
Calculate the Pearson correlation coefficient and Covariance between A and B.

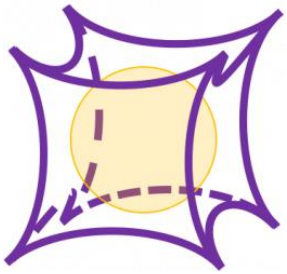
No.	Attributes	
	A	B
1	19	16
2	25	10
3	13	26
4	12	29
5	16	20

2. How to compute the above measurements in **Python**?

Tuple duplication

- Duplication should also be detected at the tuple level.
 - E.g., two or more identical tuples for a given unique data entry case.
- The use of denormalized tables (often done to improve performance by avoiding joins operation) is also a reason.
 - E.g., a purchase order database contains a purchaser's name and address instead of a key to this information in a purchaser database





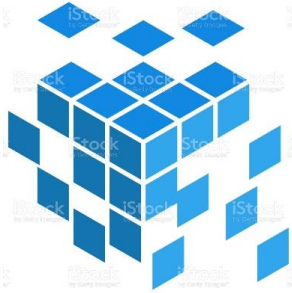
Data reduction

Why data reduction?

- A data collection stores terabytes of data → complex data analysis on the entire dataset may take a long time.
- Data reduction **reduces the dataset in volume** to achieve **(almost) the same analytical results**.



Why data reduction?



Avoid the curse of dimensionality



Eliminate irrelevant features and reduce noise



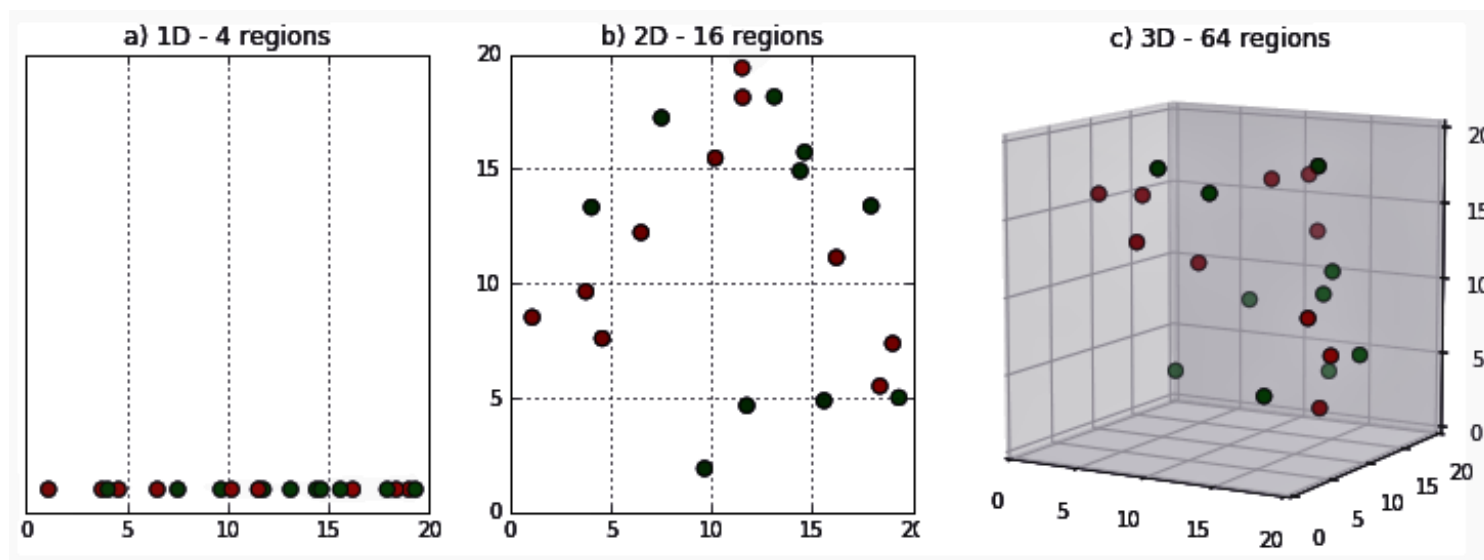
Reduce time and space required in data mining



Allow easier visualization

Curse of dimensionality

- As the dimensionality increases, the volume of the space increases so fast that the available data become sparse.



- The data must grow exponentially with the dimensionality to obtain a reliable result.

Data reduction techniques

Dimensionality reduction

- Data encoding schemes are applied for a compressed representation of the original data.

Numerosity reduction

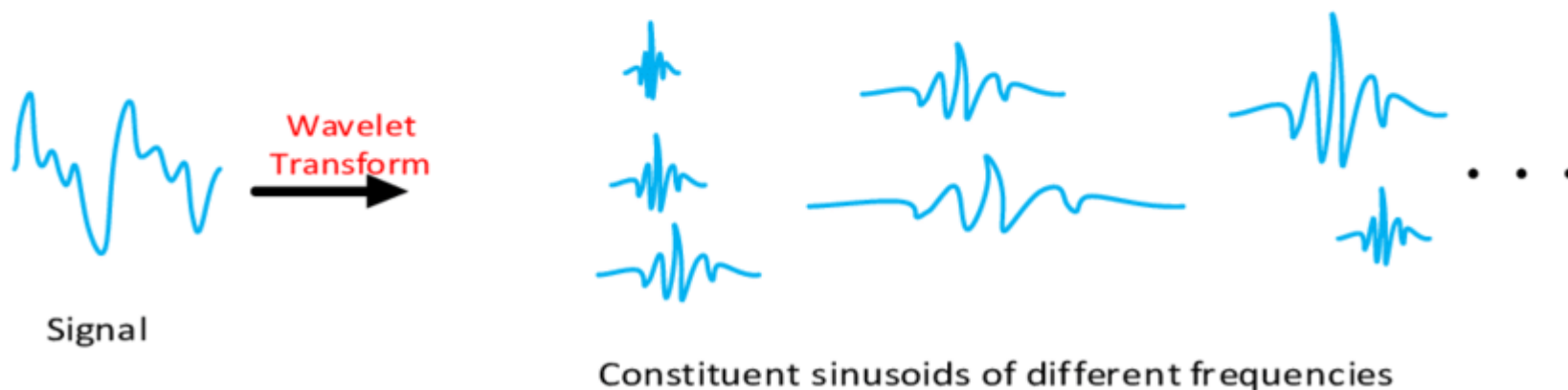
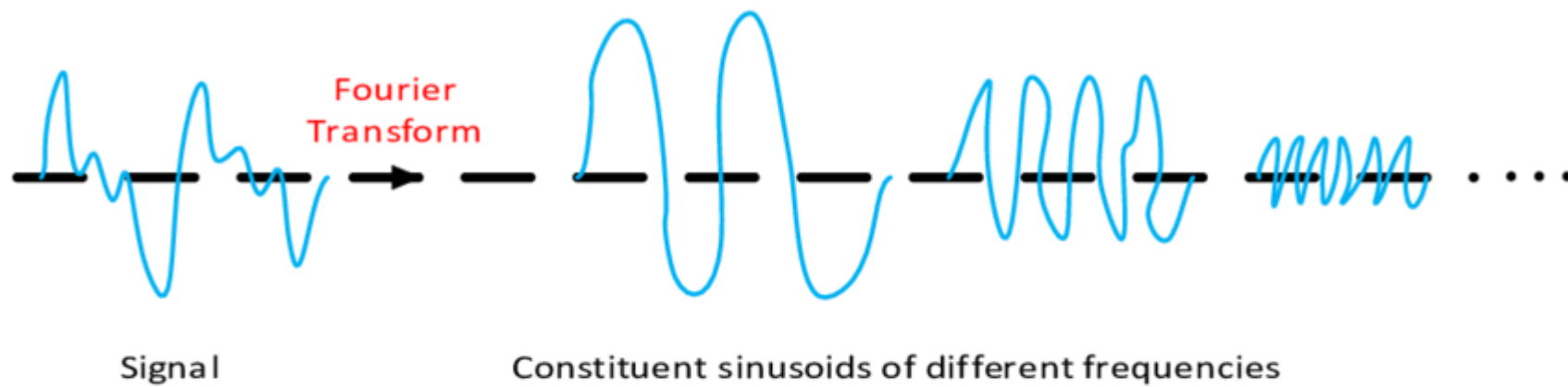
- Data volume is reduced by choosing alternative, smaller forms of data representation

Data compression

- The data is encoded using fewer bits than the original representation.

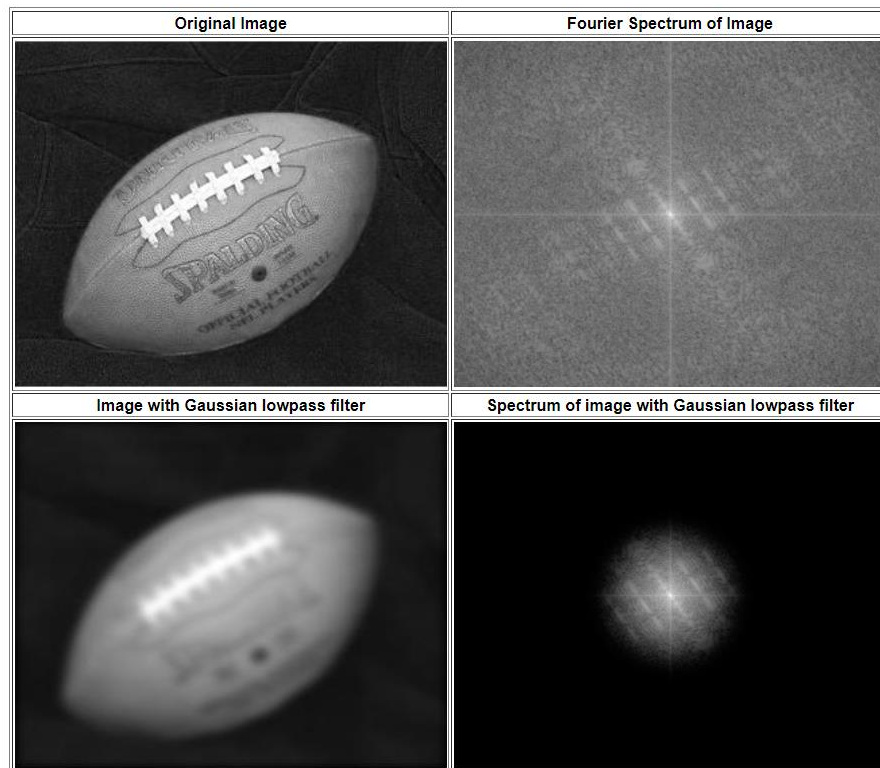
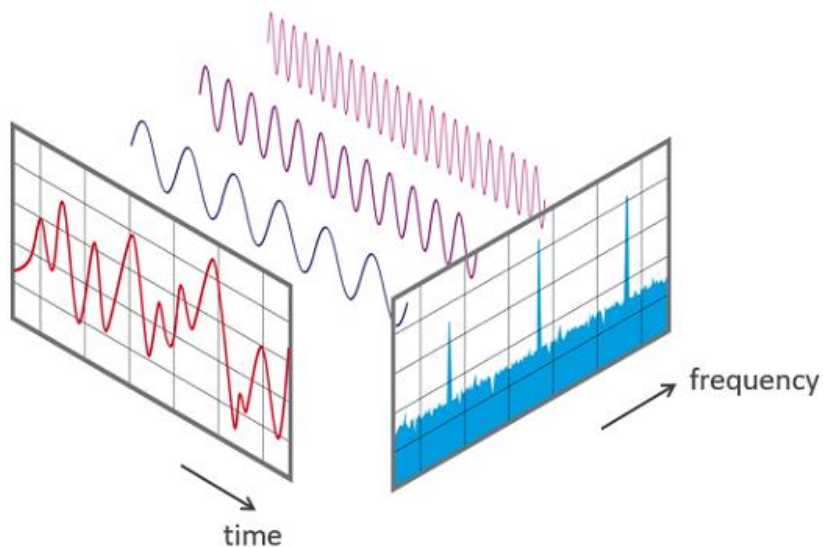
Mathematical transform

- Map the data to a new space and store only a small fraction of the strongest of the signal coefficients



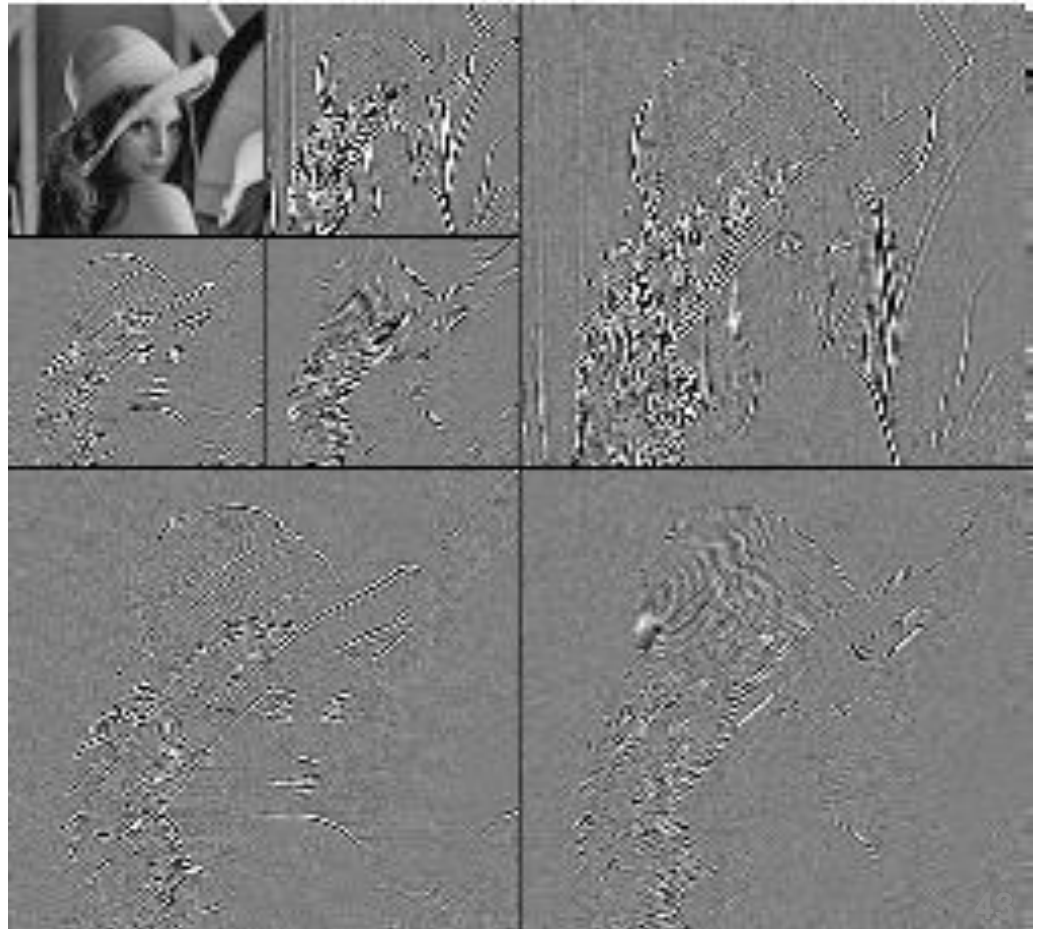
Mathematical transform: DFT

- **Discrete Fourier transform** decomposes a function in time domain into the frequency one
 - E.g., decompose an audio wave in the time domain into its constituent frequencies and volume (amplitude)



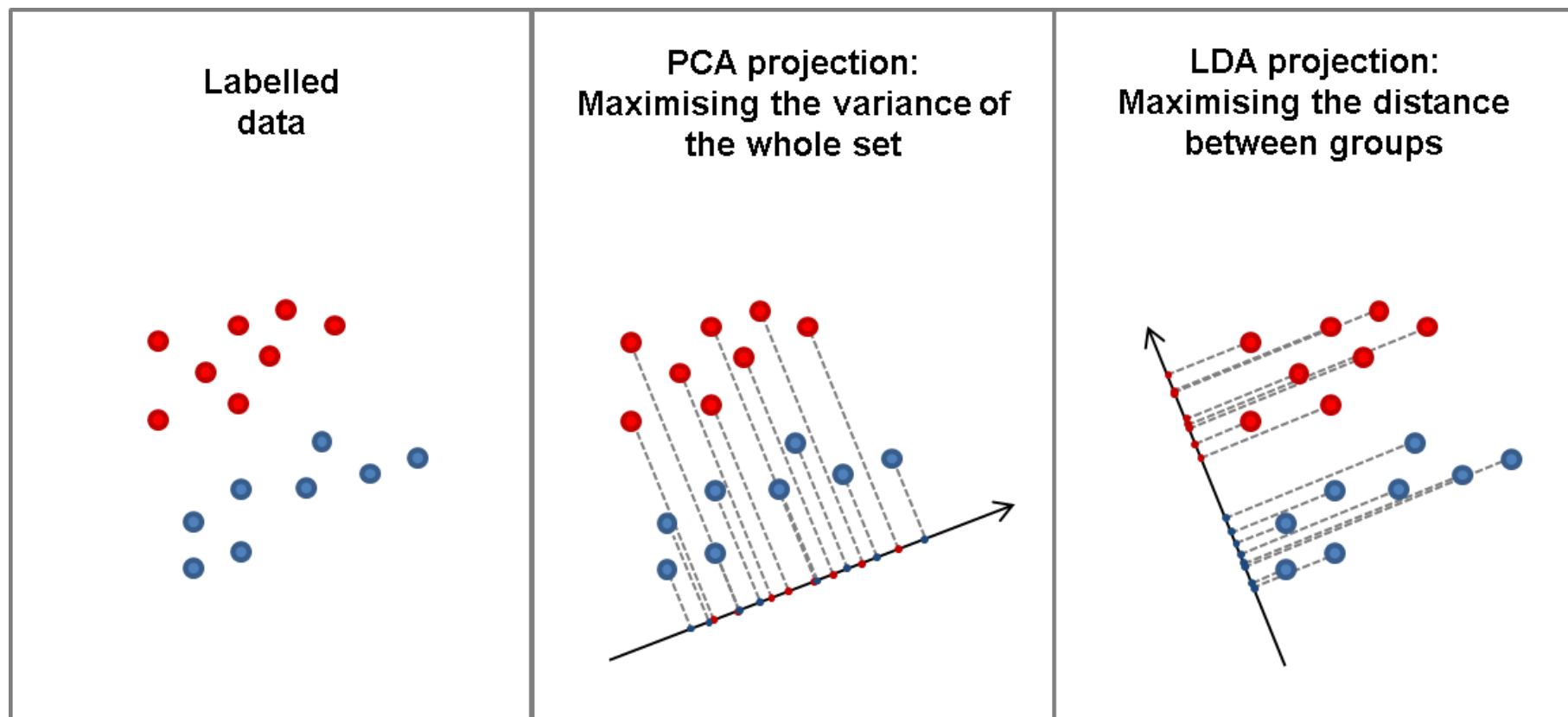
Mathematical transform: Wavelet

- **Wavelet transform**: decompose a (n-dimensional) signal into different frequency sub-bands
- Preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression



Subspace transform

- PCA and LDA both look for linear combinations of variables which best explain the data.

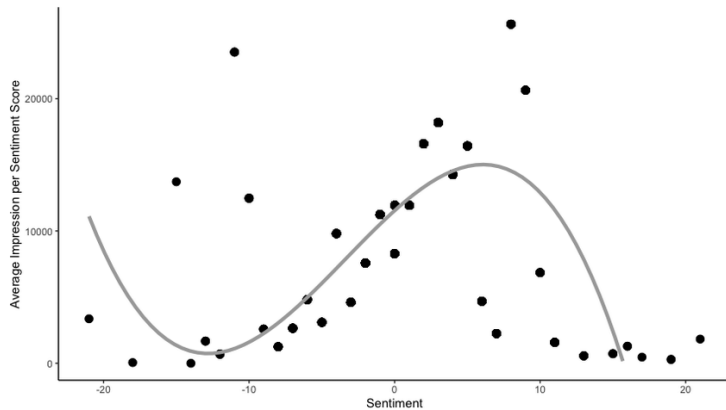


Attribute subset selection

- A way to remove redundant and/or irrelevant attributes
 - The purchase price of a product already includes the amount of sales tax paid → redundant
 - Predicting a student's GPA does not require his full name → irrelevant
- There are 2^d possible attribute combinations of d attributes.

Parametric numerosity reduction

- **Parametric methods** stores only the model's parameters, while discarding the original data (except possible outliers)

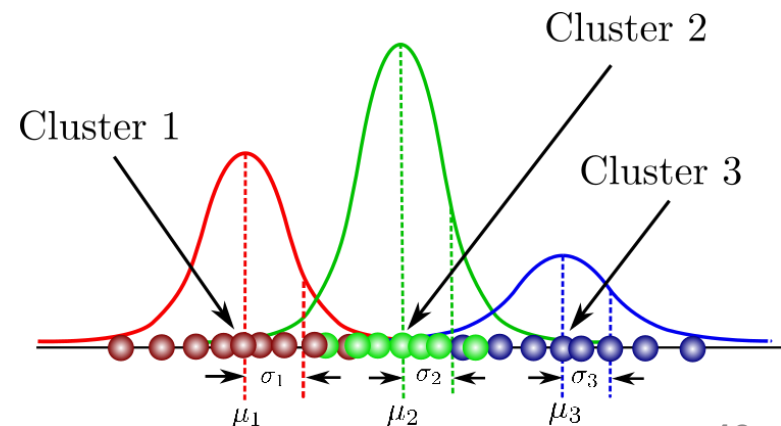


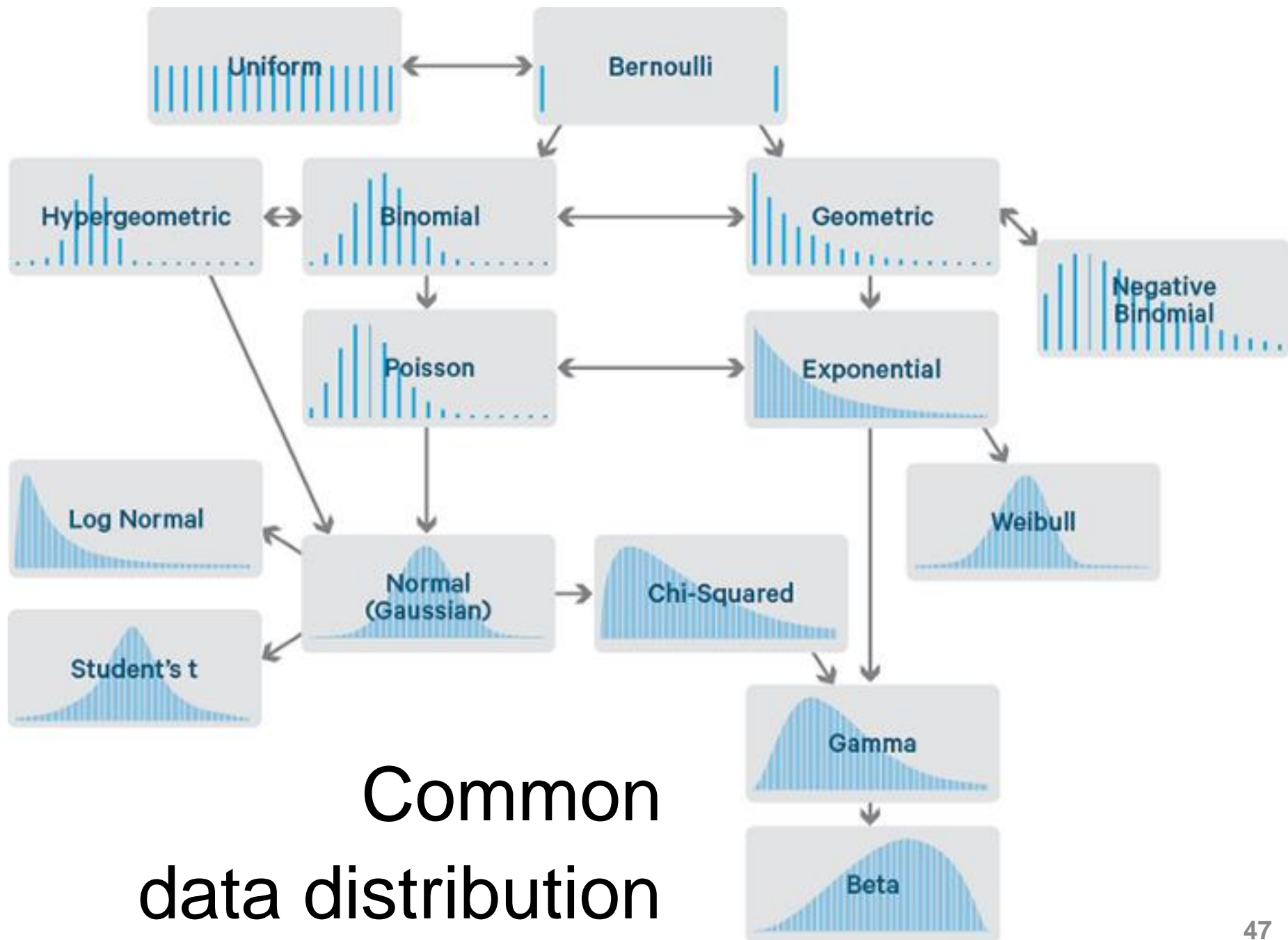
Regression analysis

- The parameters are estimated to give a "best fit" of the data
- The best fit is evaluated by using the **least squares method** or other criteria

Gaussian mixture model

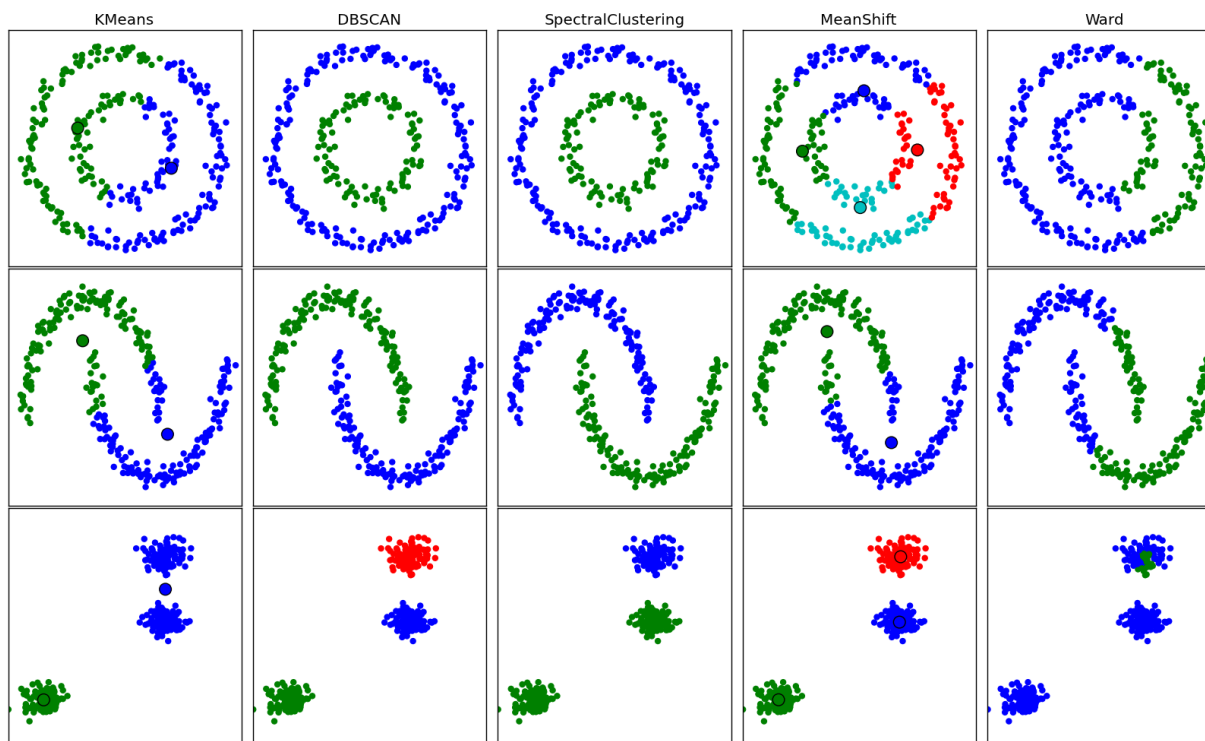
- All data points are assumed to come from a mixture of Gaussian distributions
- The best fit is estimated by using the **Expectation-Maximization** algorithm.





Why data distribution is important?

- Correctly identifying the data distribution helps apply suitable data analysis or machine learning methods.



Clustering methods work differently on
three exemplar data distributions.

Non-parametric numerosity reduction

- Non-parametric methods do not assume models.
- A histogram divides the data into buckets and stores the average sum for each bucket
- Equal-width (distance) binning
 - Divide the range into N intervals of equal width, $W = (B - A)/N$
 - where A and B are the lowest and highest values of the attribute
 - Outliers may dominate presentation, skewed data is not handled well
- Equal-depth (frequency) binning
 - Divide the range into N intervals, each containing approximately same number of samples → good data scaling

Histogram analysis: An example

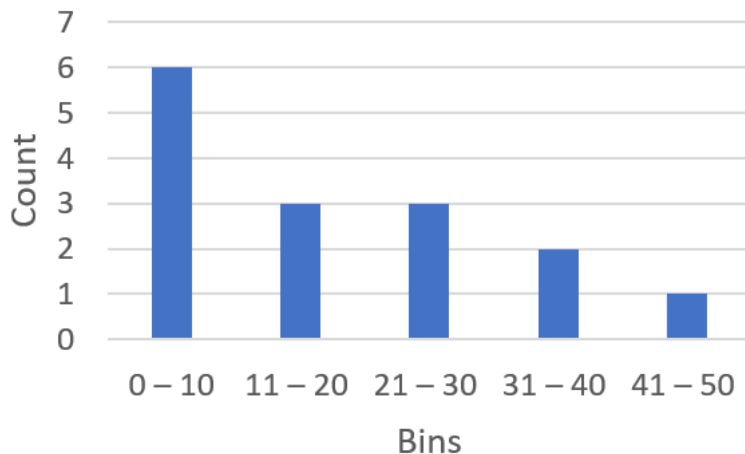
- Consider the values aside

0, 2, 5, 8, 8, 10, 15, 15, 20, 25, 25, 30, 35, 40, 49

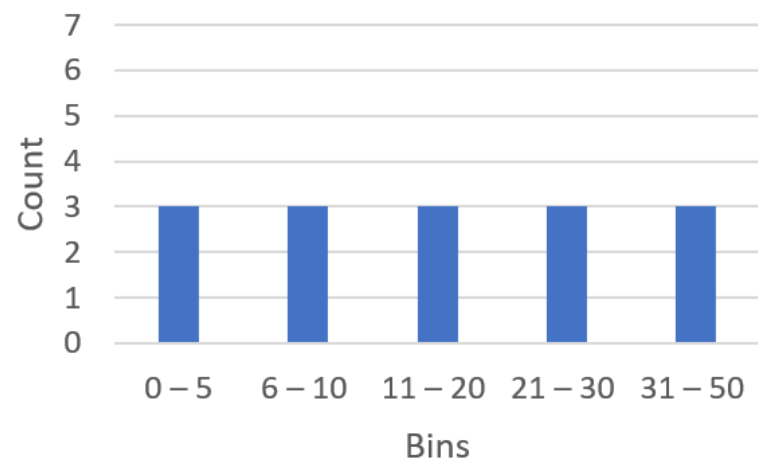
- Partition the above data into 5 bins

Equal-width binning

	Bin range	Values
Bin 1	0 – 10	0, 2, 5, 8, 8, 10
Bin 2	11 – 20	15, 15, 20
Bin 3	21 – 30	25, 25, 30
Bin 4	31 – 40	35, 40
Bin 5	41 – 50	49



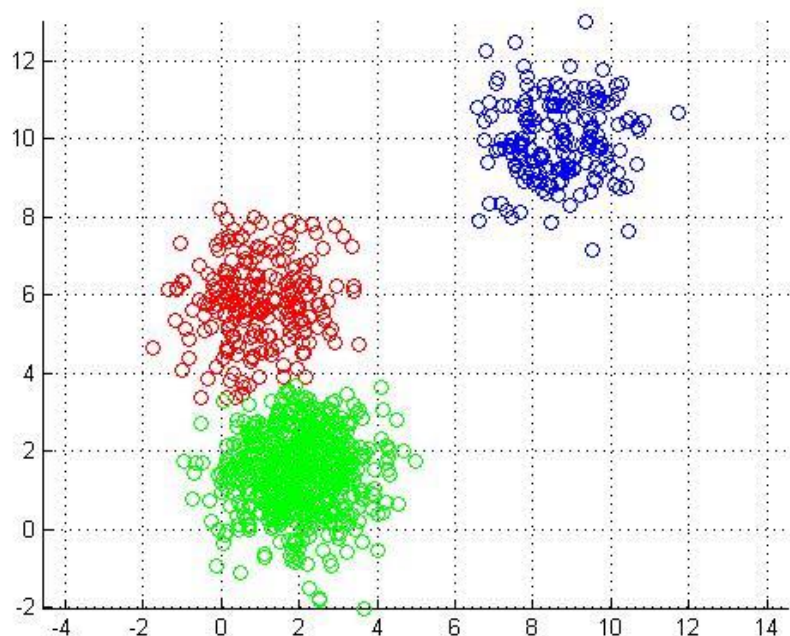
	Bin range	Values
Bin 1	0 – 5	0, 2, 5
Bin 2	6 – 10	8, 8, 10
Bin 3	11 – 20	15, 15, 20
Bin 4	21 – 30	25, 25, 30
Bin 5	31 – 50	35, 40, 49



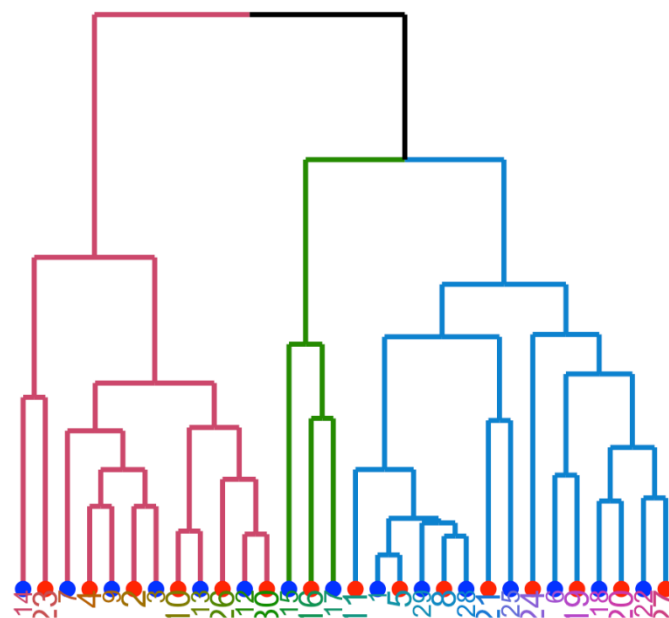
Equal-frequency binning

Clustering

- Partition the data into clusters based on similarity, and **store cluster representation** (e.g., centroid and diameter) only
- Very effective if data is clustered but not if data is “smeared”



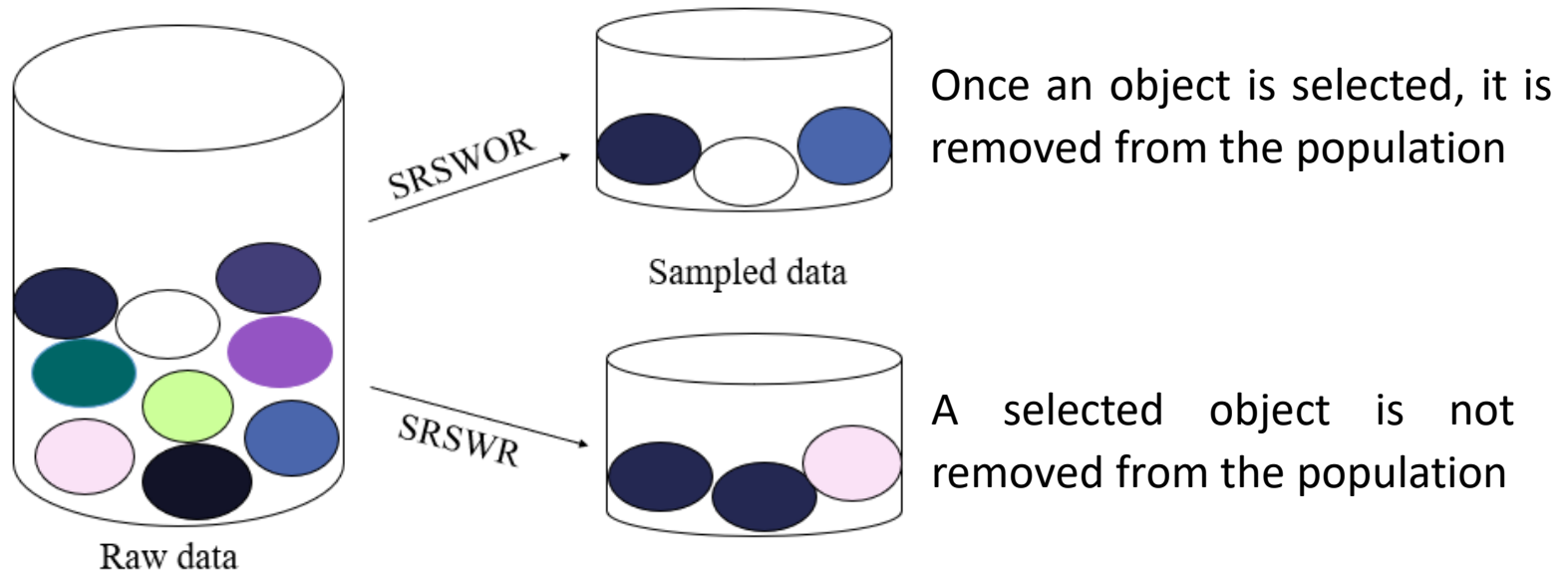
Distance-based clustering



Hierarchical clustering

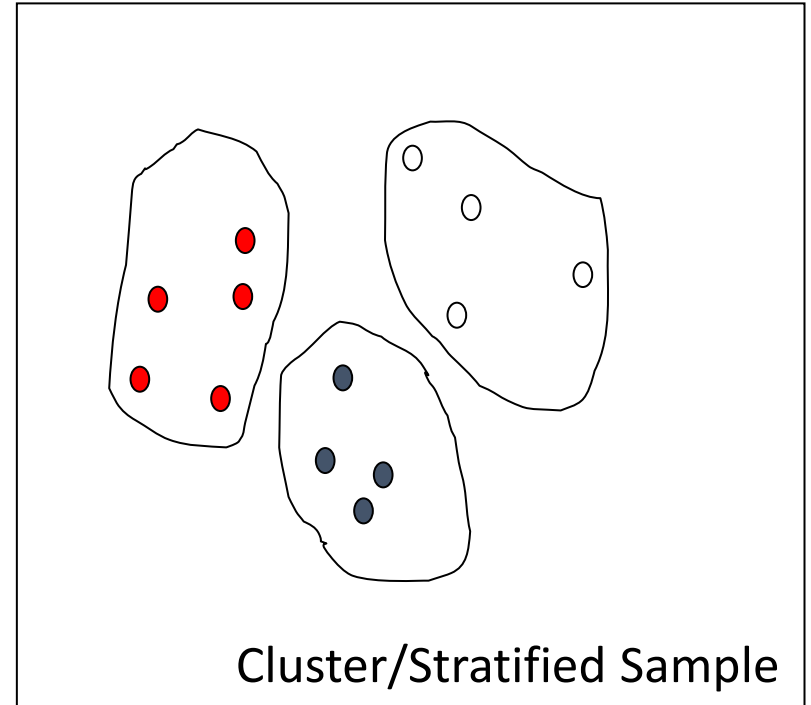
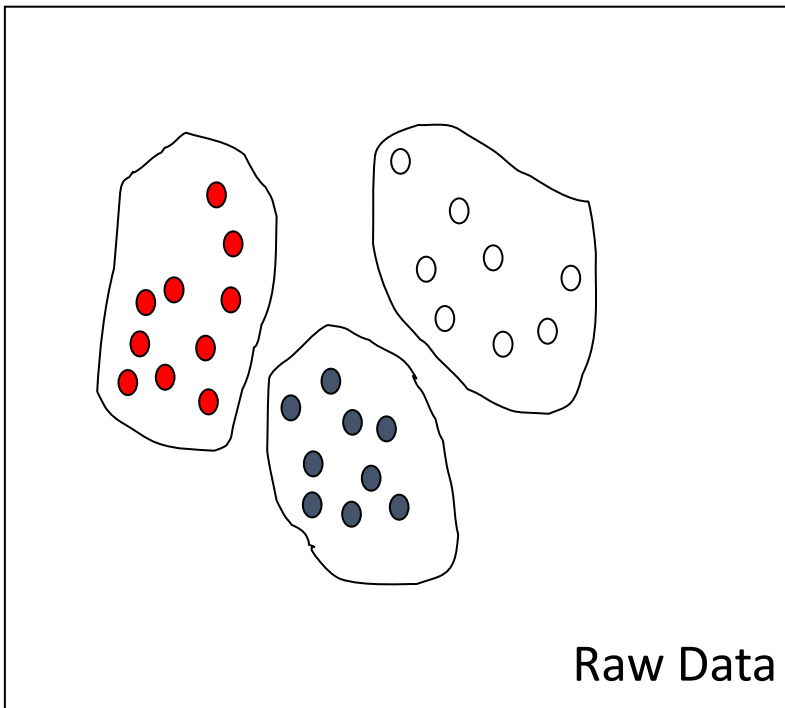
Random sampling

- Choose a representative subset s of the whole dataset D
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- **Simple random sampling:** any item can be selected with an equal probability → poor performance in skewed data



Stratified random sampling

- Partition the dataset and draw samples from each partition proportionally → good for skewed data

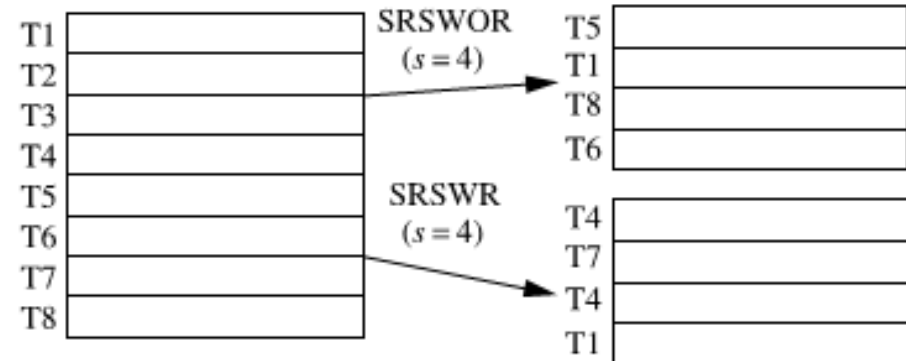


Sampling: An example

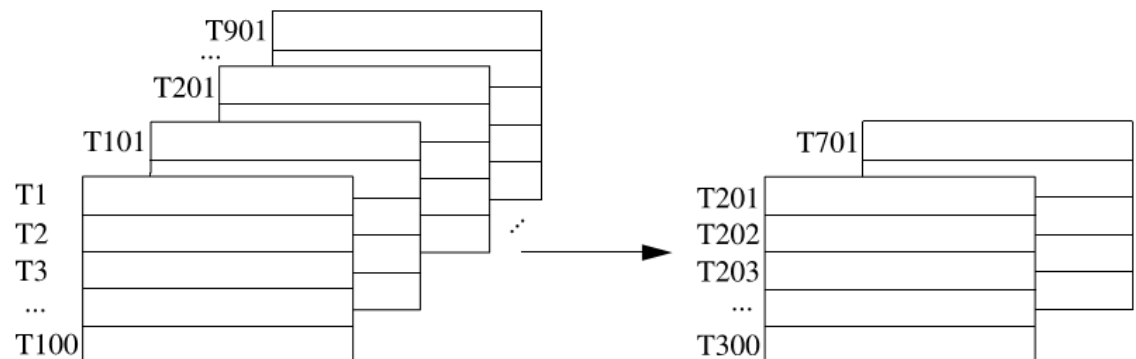
Stratified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

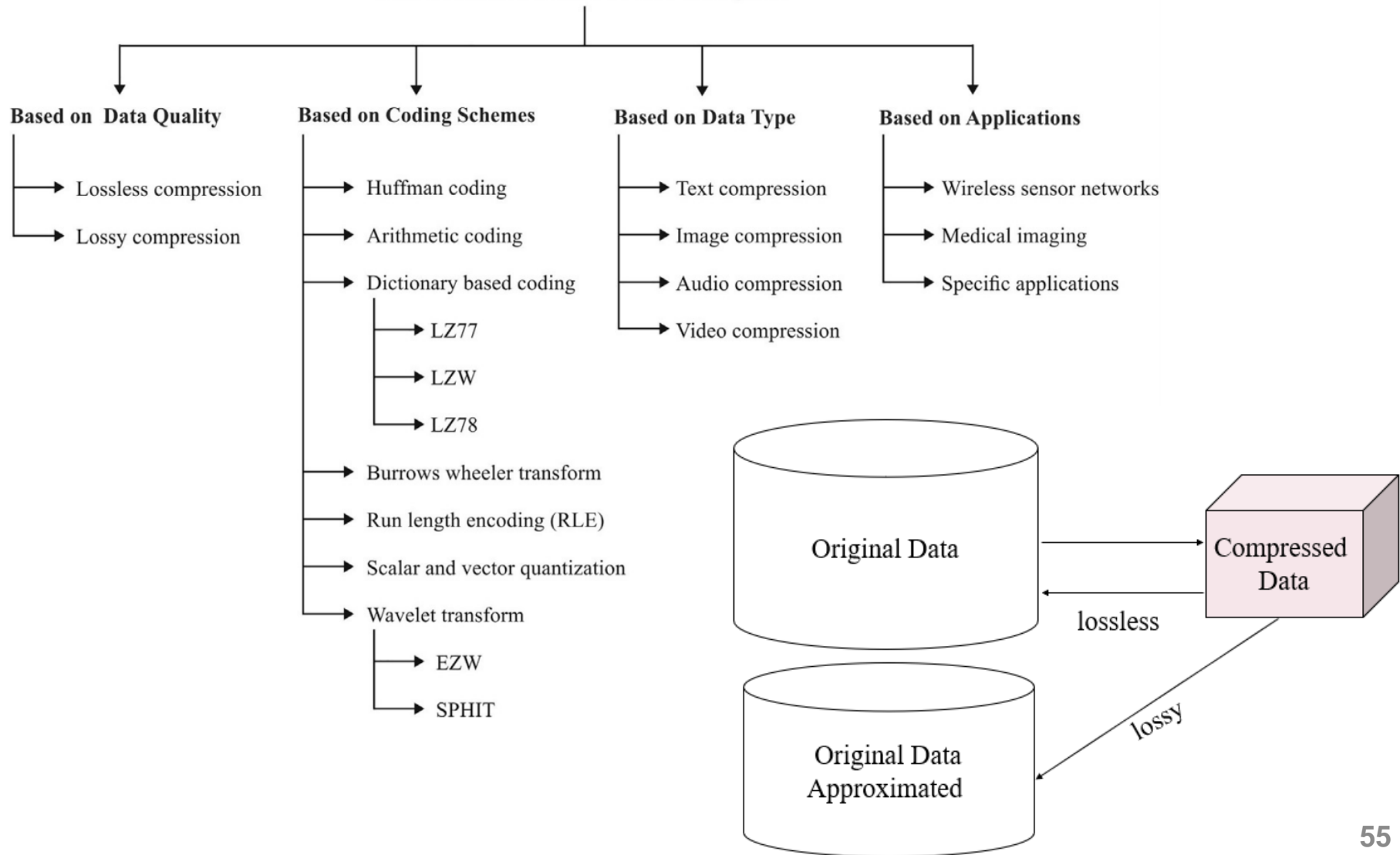


Cluster sample
($s = 2$)



Data compression

DATA COMPRESSION TECHNIQUES





Data transformation

Data normalization

- Let A be a numeric attribute with n observed values, v_1, \dots, v_n

- **Min-max normalization**

$$v'_i = \frac{v_i - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- where \max_A , new_max_A , \min_A and new_min_A are the original and modified maximum and minimum values of attribute A , respectively.

- **Decimal scaling:** $v'_i = \frac{v_i}{10^j}$

- where j is the smallest integer such that $\max(|v'_i|) < 1$
- Move the decimal point of values of A , in which the number of decimal points moved depends on the maximum absolute value of A

Data normalization

- **Z-score normalization:** $v'_i = \frac{v_i - \mu_A}{\sigma_A}$
 - Where μ_A and σ_A are the mean and standard deviation, respectively, of attribute A
- A variation that is more robust to outliers: replace σ_A by mean absolute deviation of A

$$s_A = \frac{1}{n} (|v_1 - \mu_A| + |v_2 - \mu_A| + \cdots + |v_n - \mu_A|)$$

Data normalization: An example

- Consider the following sorted points

4 8 15 21 21 24 25 28 34

- Perform min-max normalization to the new range $[-1, 1]$

- $min = 4, max = 34, new_min = -1, new_max = 1$

-1 -0.733 -0.267 0.133 0.133 0.333 0.4 0.6 1

- Perform decimal scaling normalization

- $max = 34 \rightarrow j = 2$

0.04 0.08 0.15 0.21 0.21 0.24 0.25 0.28 0.34

- Perform Z-score normalization

- $mean = 20, std = 8.994$

-1.779 -1.334 -0.556 0.111 0.111 0.445 0.556 0.889 1.557

Quiz 04: Data normalization

1. Consider the following 1D data series, which includes 10 data points sorted in ascending order.

5, 12, 18, 23, 35, 42, 50, 55, 63, 70

Transform the value 35 to a new value using

- Min-max normalization with the range $[-1.0, 1.0]$
- z-score normalization
- Decimal scaling

2. How to perform normalization, without coding from scratch, in **Python**?

Data discretization

- The range of a continuous attribute is divided into intervals, whose **labels are used to replace actual data values**.
- It aims to reduce data size or prepare for further analysis.
- Typical methods can be applied recursively, e.g., clustering and histogram-binning.

References

- Jiawei Han, Micheline Kamber, and Jian Pei, 2011. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc. Chapter 2.

...the end.

