Faulty of Computer Science and Engineering
Ho Chi Minh City University of Technology

Chapter 2
Data Preprocessing

Assoc. Prof. TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://researchmap.jp/quang

2020/4/7

CONTENT

1. Introduction to Data Preprocessing

- Cluster students

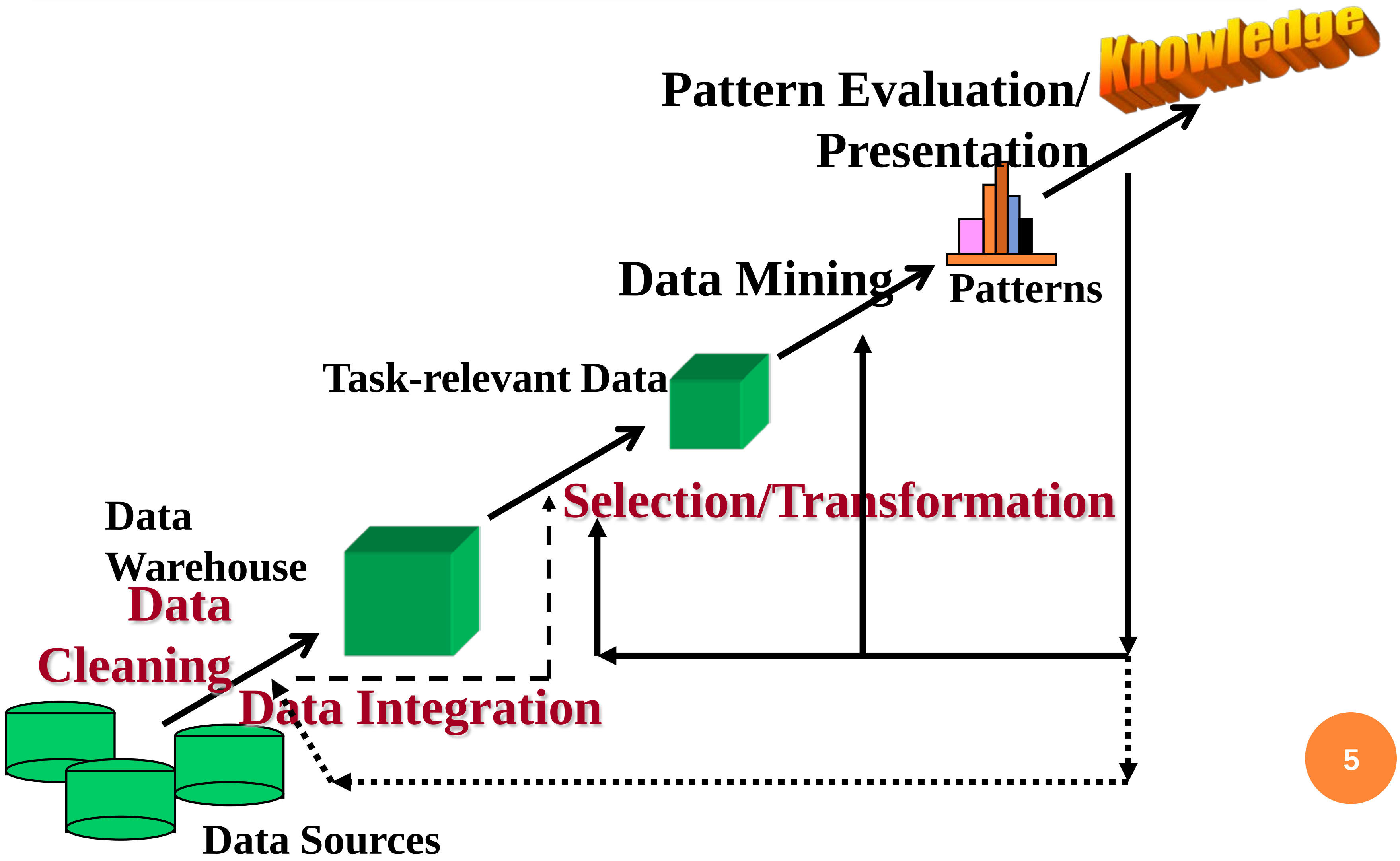
ID	CourseId	Year	Semester	Midterm test	Final exam
50503660	001001	2005	1	6	5.5
50503660	004010	2005	1	NULL	8
50503660	004009	2005	1	NULL	7
50503660	006004	2005	1	3.5	13
50503660	007005	2005	1	NULL	4
50501879	007005	2005	1	5	10
50501879	006001	2005	1	4	13

- Issue:
  - "NULL"
  - Score domain: [0,1]; [0,10], {Good, Fair, not good,..}
  - Every student and every course are taken into account?
  - What other features beside score?

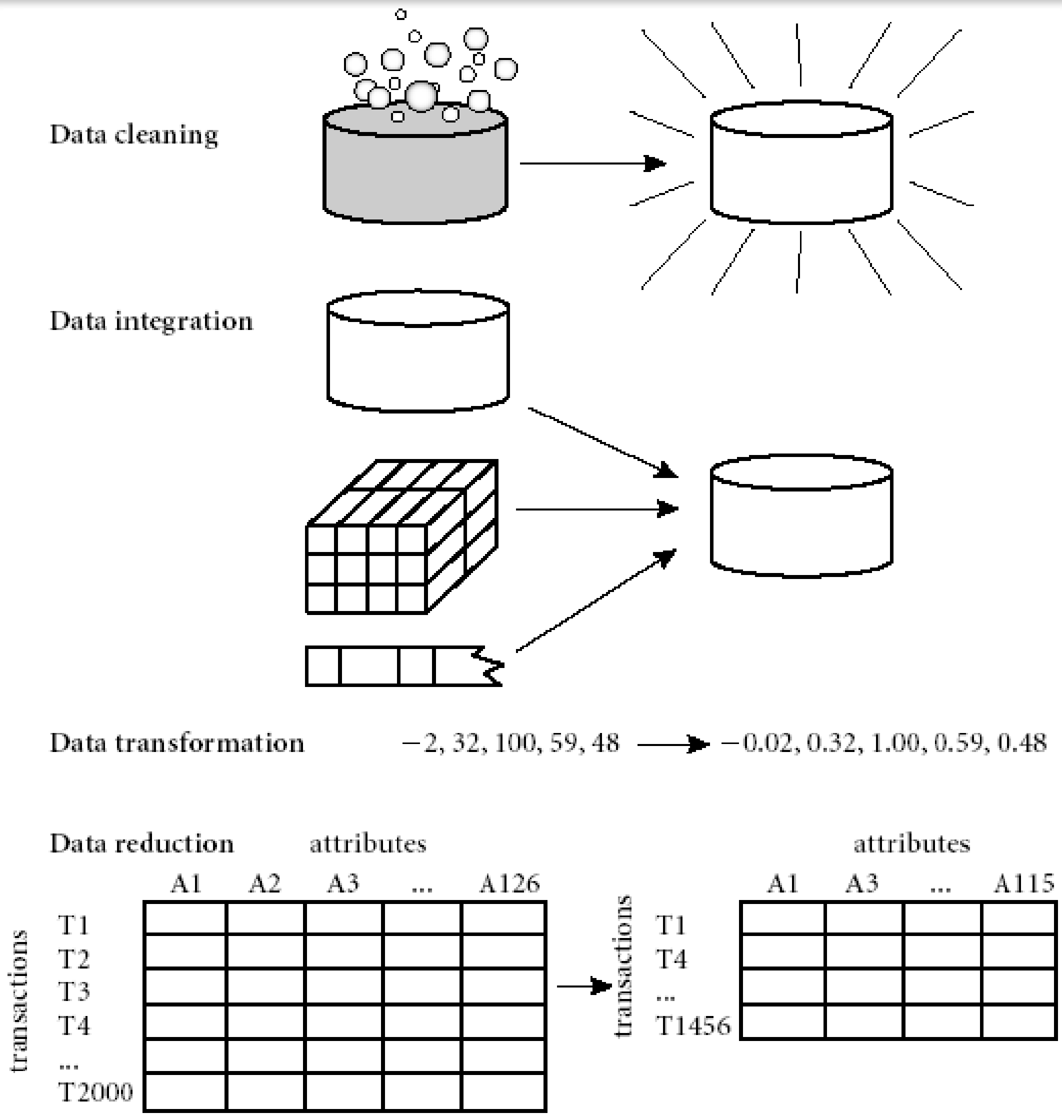
1. Introduction to Data Preprocessing

- The processes on raw/original data to increase the quality of the data, hence the quality of data mining results
- Data quality: Accuracy, currency/timeliness, completeness, consistency,…
  - Accuracy: real/truth values are recorded
  - Currency/timeliness: current availability
  - Completeness: all values for a variable/attribute are recorded
  - Consistency: All data (in the same type) are presented in the same way/format

1. Introduction to Data Preprocessing



1. INTRODUCTION TO DATA PREPROCESSING



1. Introduction to data preprocessing

- Data preprocessing techniques

1. Data cleaning/cleansing: remove noise, correct data inconsistencies,…
2. Data integration: from several sources-> data warehouse
3. Data transformation: data normalization
4. Data reduction: reduce data sizes, dimensions,…

1. Introduction to Data Preprocessing

- Data preprocessing techniques
  1. Data cleaning/cleansing
     - ✓ Data summarization: identify common features of the data and outliers
     - ✓ Resolve the missing and noise data
  2. Data integration:
     - ✓ Schema integration and object matching
     - ✓ Data redundancy issues
     - ✓ Detection and resolution of data value conflicts

1. Introduction to Data Preprocessing

- Data preprocessing techniques

3. Data transformation
   - ✓ Smoothing, aggregation, generalization, normalization
   - ✓ Attribute/feature construction

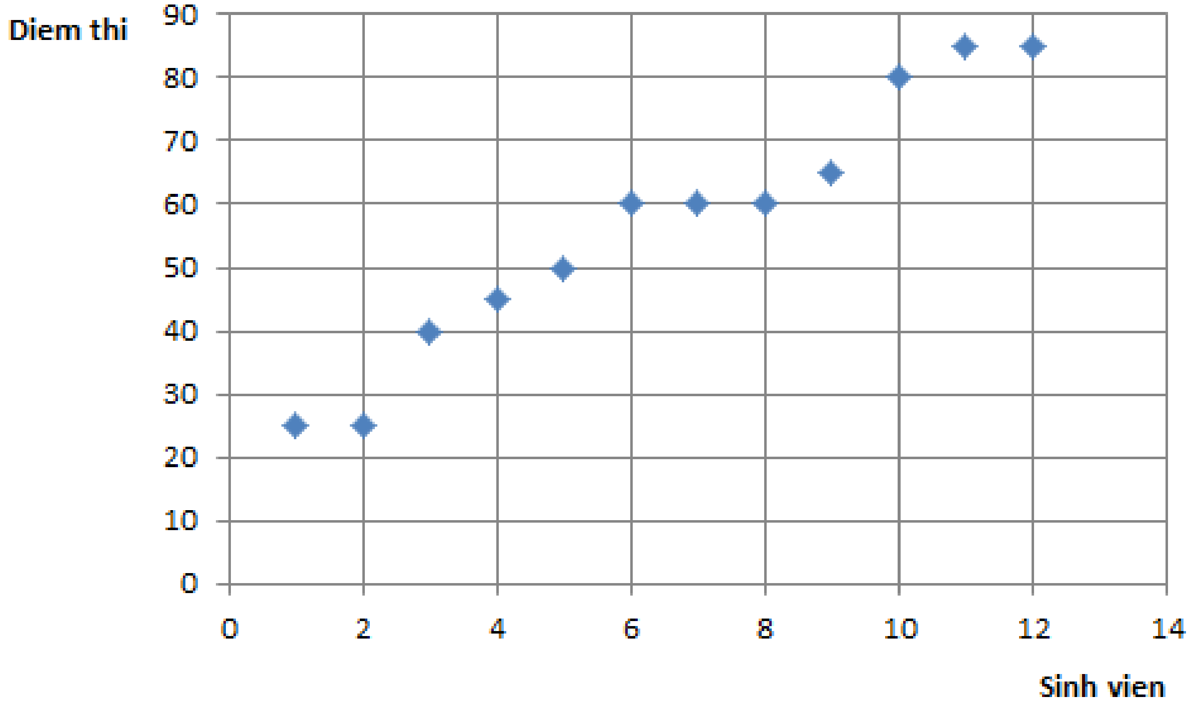
4. Data reduction
   - ✓ Reduce the data size (reduce the number of records/objects): data aggregation, data cube, clustering, concept hierarchy generation, discretization,…
   - ✓ Remove redundant features: reduce the number of dimensions/features/attributes (attribute subset selection)

2. Data summarization

- Identify typical properties of the data such as its central tendency and dispersion
  - Central tendency measures: mean, median, mode, midrange
  - Dispersion measures: quartiles, inter-quartile range (IQR), variance
- Identify noise or outliers, provides a summary about data

2. DATA SUMMARIZATION

Sinh vien	Diem thi
1	25
2	25
3	40
4	45
5	50
6	60
7	60
8	60
9	65
10	80
11	85
12	85



Identify the tendency and dispersion of the data ?

Is there any specific characteristic?

2. Data summarization

- Central tendency measures
  - Mean $$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$
  - Weighted arithmetic mean $$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$
  - Median (ordered data) $$Median = \begin{cases} x_{\lceil N/2 \rceil} & if \quad N \quad odd \\ (x_{N/2} + x_{N/2+1})/2 & if \quad N \quad even \end{cases}$$
  - Mode: most frequent data
  - Midrange: average value of maximum and minimum values in the dataset

2. DATA SUMMARIZATION

Student	Score
1	25
2	25
3	40
4	45
5	50
6	60
7	60
8	60
9	65
10	80
11	85
12	85

Score	Number of students
25	2
30	0
35	0
40	1
45	1
50	1
55	0
60	3

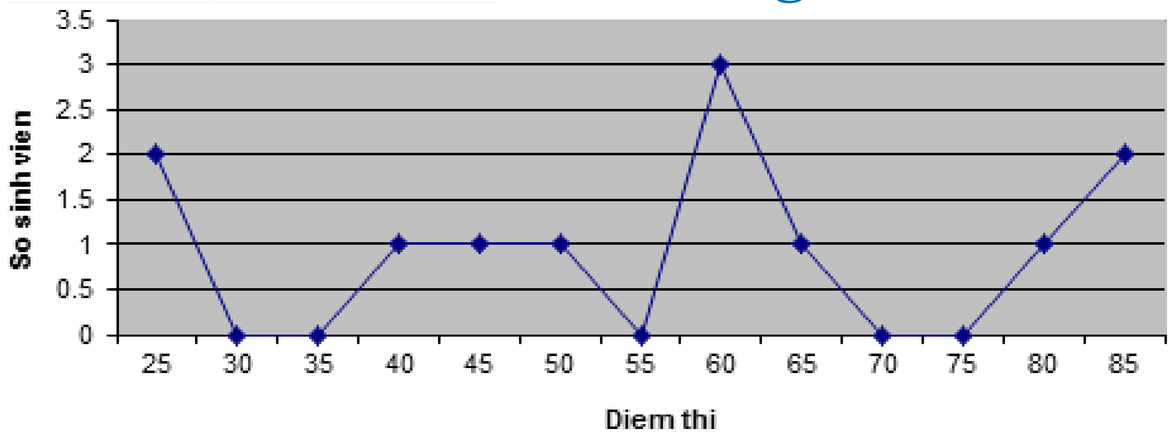
Mean = 56.67

Median = 60

Mode = 60

Midrange = 55

Histogram



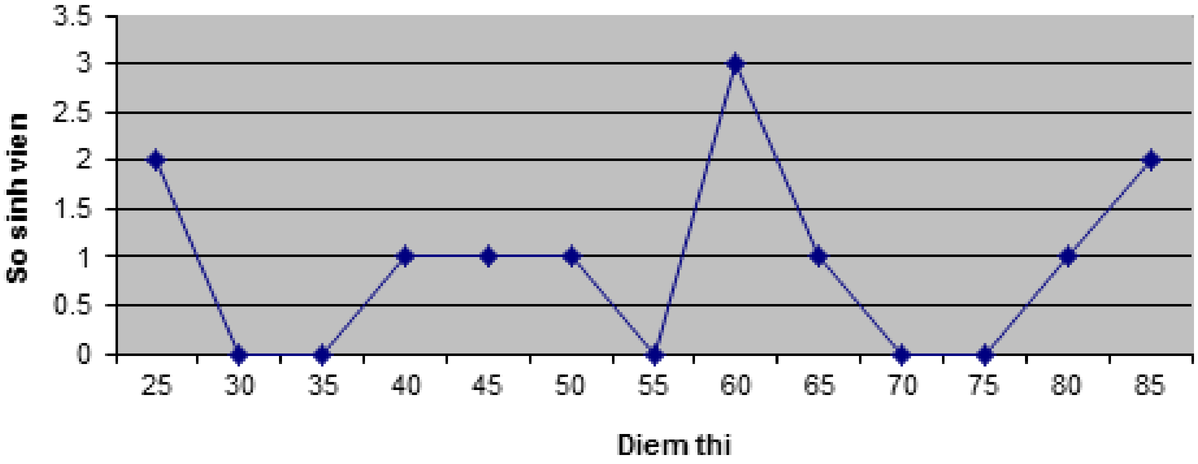
2. Data summarization

- Dispersion measures
  - Quartiles
    - ✓ 1st quartile (Q1): the 25th percentile
    - ✓ 2nd quartile (Q2): the 50th percentile (median)
    - ✓ 3rd quartile (Q3): the 75th percentile
  - Inter-quartile Range (IQR) = Q3 – Q1
    - ✓ Outliers: >=Q3 + 1.5xIQR hay <=Q1 - 1.5xIQR
    - ✓ Extreme: >=Q3 + 3xIQR hay <=Q1 - 3xIQR
  - Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}\left(\sum x_i\right)^2\right]$$

2. DATA SUMMARIZATION

Student	Score	Score - mean
1	25	-31.6667
2	25	-31.6667
3	40	-16.6667
4	45	-11.6667
5	50	-6.66667
6	60	3.333333
7	60	3.333333
8	60	3.333333
9	65	8.333333
10	80	23.33333
11	85	28.33333
12	85	28.33333



Q1 = 42.5

Q2 = median = 60

Q3 = 72.5

IQR = Q3 – Q1 = 30

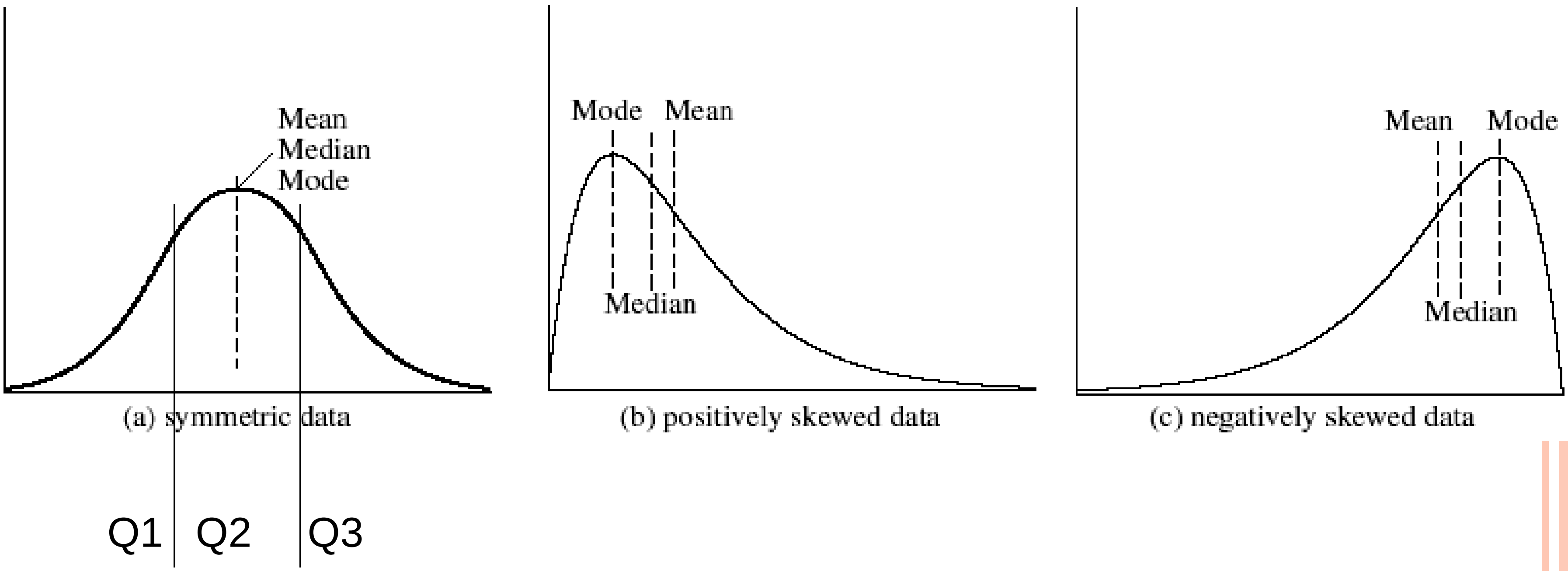
→Outliers = ???

Mean: 56.67

Variance = σ^2 = 428.78

σ = 20.7

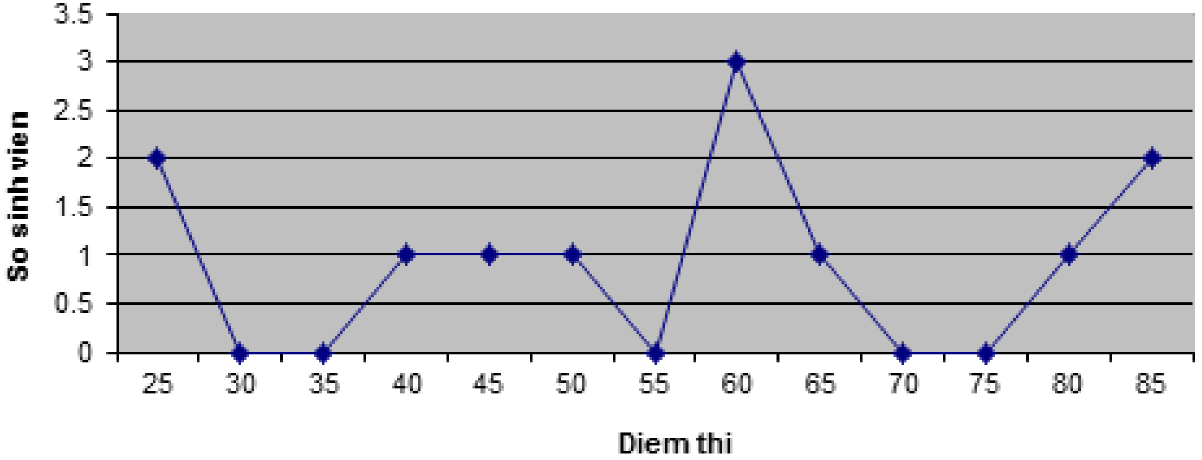
2. DATA SUMMARIZATION



Data distribution can be described by 5 main measures: median, Q1, Q3, max, và min (order by: Minimum, Q1, Median, Q3, Maximum)

2. DATA SUMMARIZATION

Sinh vien	Diem thi
1	25
2	25
3	40
4	45
5	50
6	60
7	60
8	60
9	65
10	80
11	85
12	85



Mean = 56.67 < Mode = Median = 60

→ Negatively skewed data

Minimum, Q1, Median, Q3, Maximum

25, 42.5, 60, 72.5, 85

3. Data cleansing

1. Resolve data missing issues
2. Identify outliers and remove noise from data
3. Resolve data inconsistency issues

3.1 Resolve data missing

- Missing data: not available when it is needed
- Cause:
  - ✓ Objective: Data itself does not exist, system error,…
  - ✓ Subjective: mistake from human
- Solution:
  - ✓ Don't use it
  - ✓ Update manually
  - ✓ Update automatically by replacing values: a global constant, frequent values, average (local, global), predicted value,…
  - ✓ Preventing issues from design: DB design and integrity constraints,…

3.2 Outlier detection & Noise removal

- *Outliers*: objects that do not follow common characteristics (behaviors) of the dataset.
- *Noisy data*: rejected/discarded outliers, exceptions
- Causes
  - ✓ Objective: system errors, communication issues, technical issues,…
  - ✓ Subjective: Human errors

3.2 Outlier detection & Noise removal

- Outlier detection
  - ✓ Statistical distribution-based
  - ✓ Distance-based
  - ✓ Density-based
  - ✓ Deviation-based
- Noise removal
  - ✓ Binning
  - ✓ Regression
  - ✓ Cluster analysis

3.2 Outlier detection & Noise removal

- Noise removal
  - Binning (by bin means, bin median, bin boundaries)
    - ✓ Ordered data
    - ✓ Distribute data into bins (buckets)
    - ✓ Bin boundaries: min and max values

Sorted data for *price* (in dollars): 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

Smoothing by bin means:

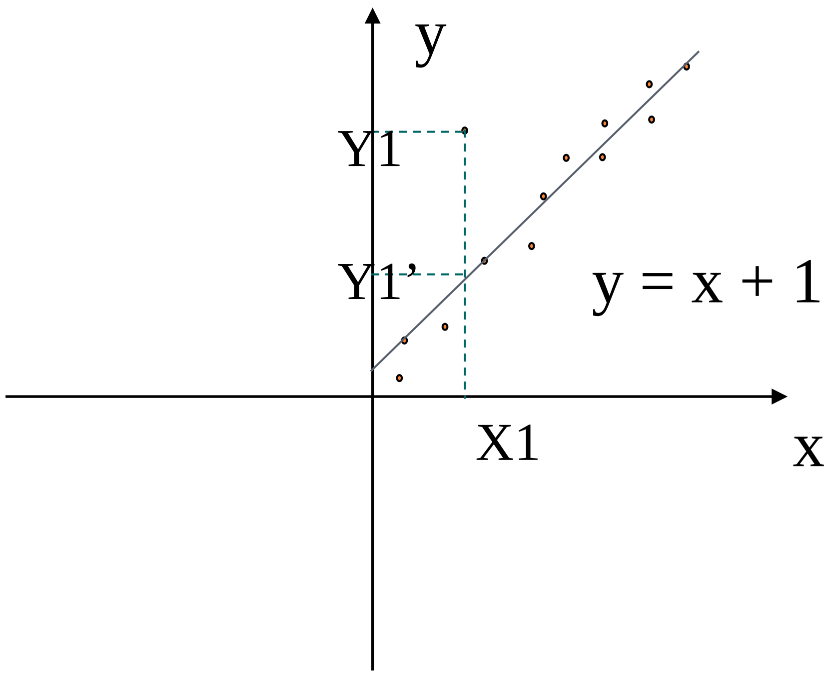
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

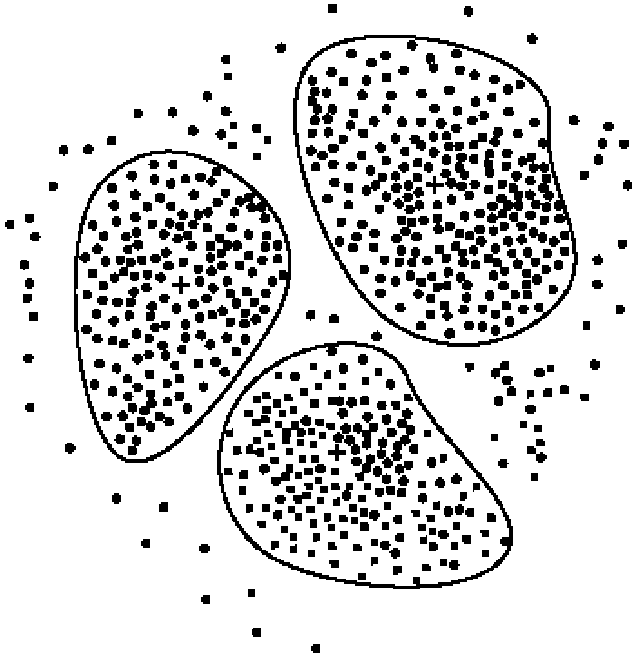
Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

3.2 Outlier detection & Noise removal

- Noise removal
  - Regression and cluster analysis





3.3 Resolve data inconsistency issues

- Discrepancies from inconsistent data representations

  -> Ex. 2004/12/25 and 25/12/2004
- Data violates integrity constraints: Ex., referential key violation
- Causes
  - ✓ Inconsistent in naming or coding methods
  - ✓ Inconsistence in data format
  - ✓ System errors, human mistake,…
- Solution
  - ✓ Use metadata for correction, apply data constraints
  - ✓ Manual and/or automatic correction

4. Data Integration

- Is to integrate data from multiple sources to a data warehouse which is ready for data mining
  - Entity identification issues
    - Schema integration
    - Object matching
  - Redundancy issues: the same data available at different sources
  - Data value conflicts: which should be the correct one?

→ The issues relates to different data structures, heterogeneity, and data semantics

→ Need to reduce/avoid data redundancy and inconsistency (DB technologies can help) → improve the accuracy and efficiency of the data mining process

4. DATA INTEGRATION

- Entity identification issues
  - Object/entity/attribute: come from multiple sources
  - Two different names but have the same meaning
  - E.x., schema level: *cust_id* in S1 and *cust_No* in S2
  - E.x., instance level: *"R & D"* in S1 and *"Research & Development"* in S2. *"Male"* and *"Nam"* in two sources

→ Metadata plays an important role for resolving these issues

4. Data Integration

- Data redundancy issues
  - Fact: value of attribute A can be inferred from B and both of them are available in the dataset (-> data duplication)
  - Cause: bad data management, inconsistency in dimension/attribute naming
  - Redundancy detection: correlation analysis
    - ✓ Detect the capability to infer A from B
    - ✓ Numerical attributes: correlation coefficient, aka Pearson's product moment coefficient
    - ✓ Categorical attributes: analyze the correlation between two attributes using chi-square (χ^2) analysis

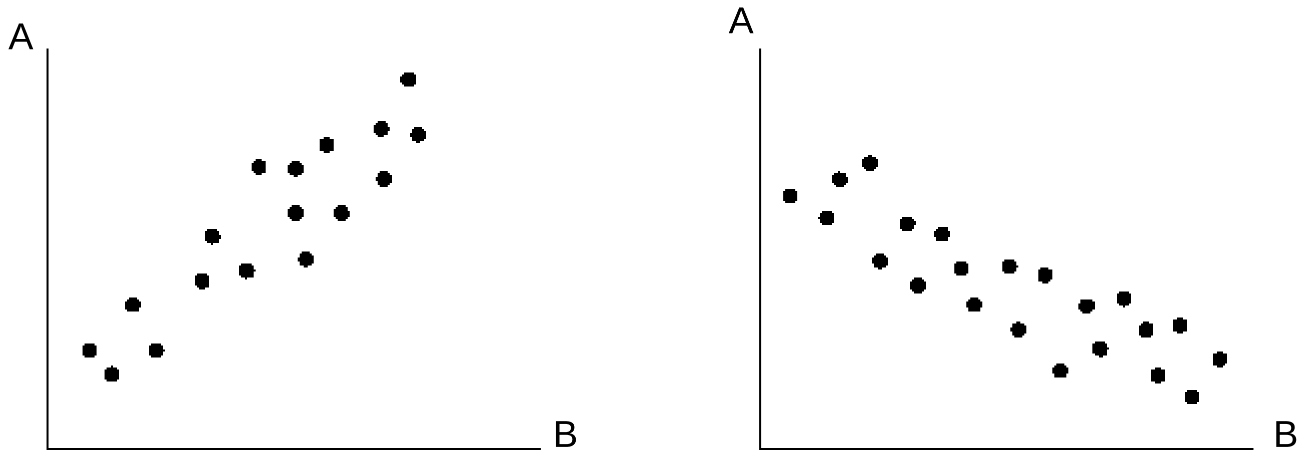
4. Data Integration

- Numerical attributes: correlation analysis between two attributes A and B
  - $r_{A,B} \in [-1, 1]$
  - $r_{A,B} > 0$: A & B correlate with each other, the more $r_{A,B}$ the higher correlation between them -> A or B can be removed (for data mining)
  - $r_{A,B} = 0$: A and B are independent
  - $r_{A,B} < 0$: A and B are inversely correlated with each other (A increases then B decreases and vice versa)

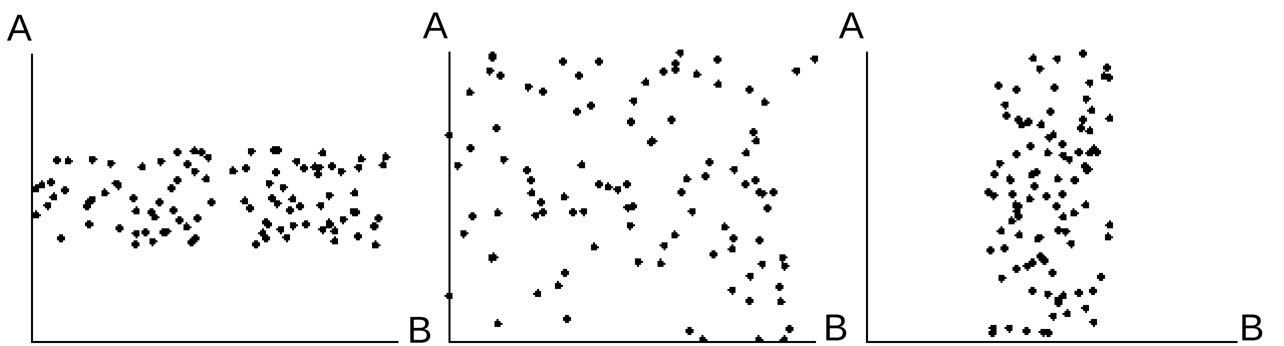
  => can we remove A or B?

4. Data Integration

- Correlation between two numerical attributes A and B



Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.



Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

4. Data Integration

- Correlation of categorical attributes: A & B
  - A consists of *c* separate values, $a_1, a_2, \ldots, a_c$.
  - B consists of *r* separate values, $b_1, b_2, \ldots, b_r$.
  - o_{ij}: number of objects (tuples) to which value of A is a_i and value of B is b_j.
  - count($A=a_i$): number of objects who A's value is a_i.
  - count($B=b_j$): number of objects who B's value is b_j.

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

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

4. Data Integration

- Correlation of categorical attributes: A & B
  - The chi-square (χ^2) statistics will evaluate the hypothesis "A and B are independent with a *significance level (Sig)* and a *degree of freedom (DoF)*"
  - If the above hypothesis is not acceptable the A and B are correlated with each other (based on statistics)
  - Degree of freedom: (r-1)*(c-1)
    - ✓ Map (*Sig* and *DoF*) to the chi-square table to identify the value χ^2
    - ✓ If the calculated χ^2 (from the previous slide) is greater than or equal to χ^2 extracted from the table then **A and B are correlated** (the hypothesis is wrong)

4. Data Integration

- Correlation of categorical attributes: A & B
  - Investigate 1500 persons with 2 attributes *gender* and *preferred_reading* -> **whether they are correlated**?

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

→ Use χ^2 to evaluate the hypothesis that *gender* and *preferred_reading* are independent

4. Data Integration

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

o_{11} = 250; o_{12} = 200; o_{21} = 50; o_{22} = 1000

e_{11} = (count(male)*count(fiction))/N = (300*450)/1500 = 90

e_{12} = (count(female)*count(fiction))/N = (1200*450)/1500 = 360

e_{21} = (count(male)*count(non_fiction))/N = (300*1050)/1500 = 210

e_{22} = (count(female)*count(non_fiction))/N = (1200*1050)/1500 = 840

$$\begin{aligned}\chi^2 &= \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} \\ &= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.\end{aligned}$$

Degree of freedom = (2-1)*(2-1) = 1; Significance level = 0.001

From: χ^2 = 10.828 <<< χ^2 calculated from dataset (507.93)

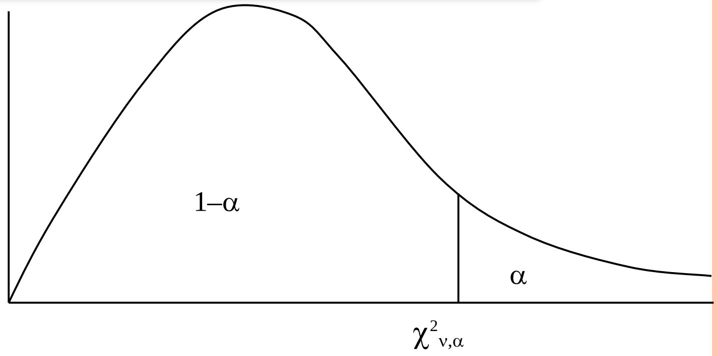
→ The hypothesis is rejected -> *gender* and *preferred_reading* are correlated

4. DATA INTEGRATION

CHI-SQUARE Distribution

(Given alpha=0.1, degree of freedom = 6, Chi-square $_{alpha}$ is10.64.
Meaning: P(Chi-quare > Chi-square $_{alpha}$) = alpha)

Statistical distribution



=CHIINV(v,α)

Freedom v	Chi-Square Alpha									
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005	**0.001**
1	3.93E-05	1.57E-04	9.82E-04	3.93E-03	0.0158	2.71	3.84	5.02	6.63	7.88	**10.8276**
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.61	5.99	7.38	9.21	10.60	13.8155
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84	16.2662
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86	18.4668
5	0.412	0.554	0.831	1.145	1.610	9.24	11.07	12.83	15.09	16.75	20.515
6	0.676	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55	22.4577
7	0.989	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28	24.3219
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95	26.1245
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59	27.8772
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19	29.5883
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76	31.2641
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30	32.9095
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82	34.5282
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32	36.1233
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80	37.6973
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27	39.2524
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72	40.7902
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16	42.3124
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58	43.8202
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00	45.3147
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40	46.797

4. DATA INTEGRATION

- Data value confliction issues
  - Given a real object, its values come from different sources might be different in terms of representation, scaling, encoding, …
    - ✓ Representation: “2004/12/25” với “25/12/2004”.
    - ✓ Scaling: *GPA* : [0, 4] hay [0, 10]; *Price* in different currency systems
    - ✓ Encoding: “yes” vs. “no” or “1” vs. “0”

5. Data Transformation

- Is a process that transforms or aggregates data into appropriate forms/formats for the KDD
  - Data smoothing
  - Aggregation
  - Generalization
  - Normalization
  - Attribute/feature construction

5. DATA TRANSFORMATION

- Smoothing
  - Binning (bin means, bin medians, bin boundaries)
  - Regression: to predict a new value
  - Clustering (Outlier analysis)
  - Data discretization (Conceptual hierarchy)

  → Remove noises from data
- Aggregation
  - Data summarization: Detailed data -> aggregated data (min, max, average, sum,…)
  - Multi-dimensional data cubes with different levels of granularity (e.g., sum by week/month/quarter…)

  → Data reduction

5. DATA TRANSFORMATION

- Generalization
  - Atomic data or from lower levels -> higher level based on conceptual hierarchy
  - Ex. Detailed score -> GPA -> Student classification (excellent, good, fair,…)

  → Data reduction
- Normalization
  - min-max normalization
  - z-score normalization
  - Normalization by decimal scaling

  → Data values are transformed to values in a pre-defined domains

5. DATA TRANSFORMATION

- Normalization
  - min-max normalization
    - ✓ Current value: v ∈[minA, maxA]
    - ✓ New value: v’ ∈ [new_minA, new_maxA]
    - ✓ Ex: normalize the score from [0, 4] to [0,10].

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

5. BIẾN ĐỔI DL (DATA TRANSFORMATION)

- Chuẩn hóa (normalization)
  - z-score normalization
    - ✓ Giá trị cũ: v tương ứng với mean Ā và standard deviation $б_A$
    - ✓ Giá trị mới: v'

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

5. Data Transformation

- Normalization
  - Normalization by decimal scaling
    - ✓ Current value: v
    - ✓ New value: v’ as in the equation, where j is the minimum integer that satisfies Max(|v’|) < 1

$$v' = \frac{v}{10^j}$$

5. Data Transformation

- Attribute/feature construction
  - Create new attributes and add to the dataset
  - Support for accuracy evaluation and help to understand the structure of multi-dimensional dataset
  - Support to identify missing data

→ Derived attributes

6. Data reduction

- Transform the original dataset to a smaller one (while keeping the data/information completeness)
- Reduction strategies
  - Data cube aggregation
  - Attribute subset selection
  - Dimensionality reduction
  - Numerosity reduction (reduce the number of objects)
  - Discretization
  - Concept hierarchy generation

  → Data reduction: lossless and lossy

6. Data reduction

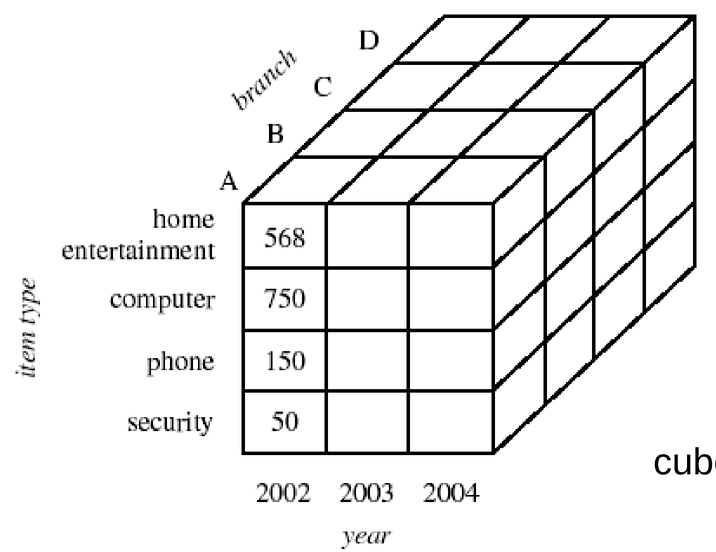
- Data cube aggregation
  - Data type: additive, semi-additive (numerical)
  - Data aggregation: average, min, max, sum, count, …
  - Abstraction/granularity level: the higher level the more data reduction

Year 2004

Year 2003

Year 2002	
Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000



cube: Sale

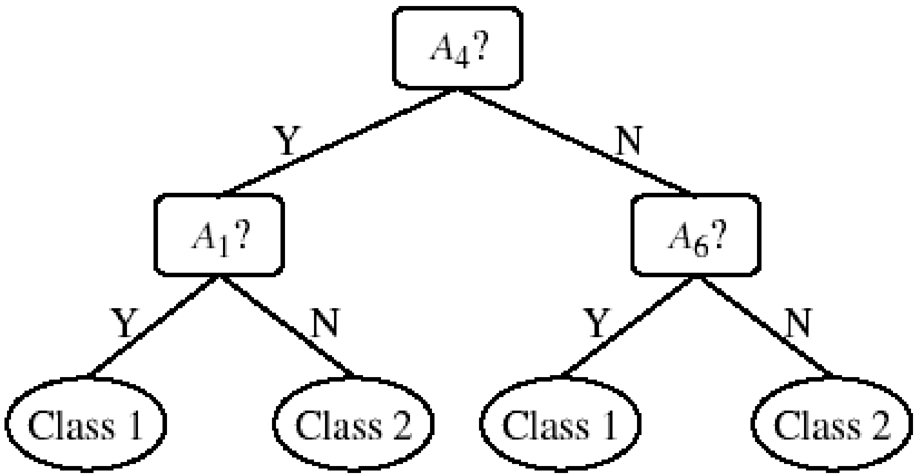
6. DATA REDUCTION

- Attribute subset selection
  - Remove attribute/dimension/feature that are redundant or irrelevant
  - Objective: to get a dataset with the smallest set of attributes while keeping the probability distribution of different object classes in the original dataset

  → This is an optimal problem: Applyheuristics

6. Data reduction

- Attribute subset selection

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: {} => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

6. DATA REDUCTION

- Dimensionality reduction
  - Correlation analysis
  - Wavelet transforms
  - Principal component analysis (PCA)

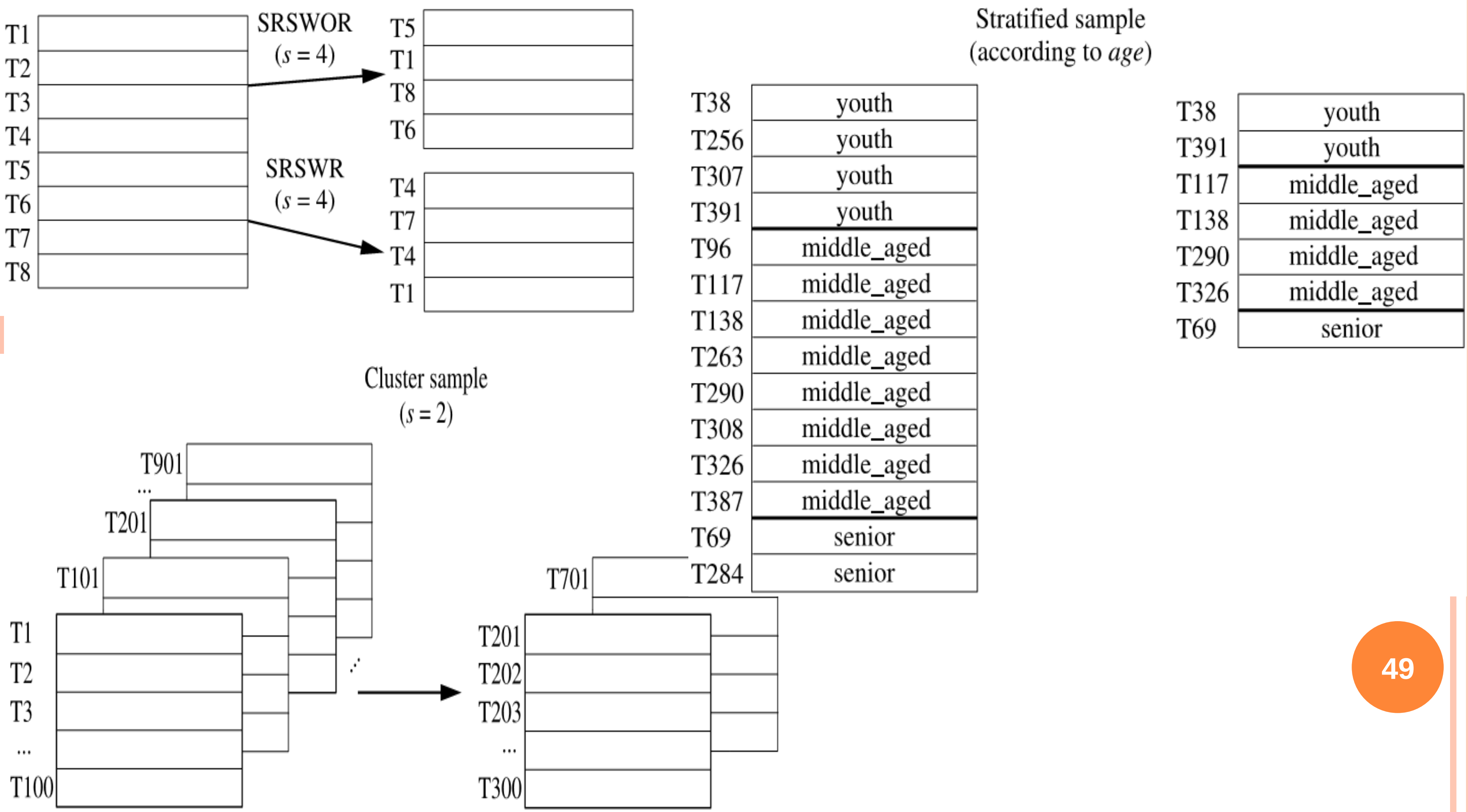
→ Depending data/application characteristics

6. DATA REDUCTION

- Numerosity reduction
  - Numerosity reduction by applying another ways of data representation
  - Parametric methods: Data estimation models → Storing models, parameters rather than storing real data
    - ✓ Ex. A regression model
  - Nonparametric methods: store reduced representation of the data
    - Histogram, Clustering, **Sampling**
      - ✓ Simple random sample without replacement (SRSWOR)
      - ✓ Simple random sample with replacement (SRSWR)
      - ✓ Cluster sample
      - ✓ Stratified sample

6. DATA REDUCTION

- Sampling



7. DATA DISCRETIZATION

- To reduce the number of values of a continuous attribute by dividing the attribute domain into intervals (discrete)
- These intervals are labeled and used instead of original continuous values
- Attribute values can be partitioned following a hierarchy or in multiresolution manner

7. Data discretization

- Discretizing numeric attributes
  - Using conceptual hierarchy: lower concepts (many) are replaced by higher concept
  - The conceptual hierarchy can be built automatically based on analyzing data distribution
  - The data details will be lost
  - The resulted data still remain the meaning for analysis but easier to be presented and required less storage

7. Data Discretization

- Discretizing numeric attributes
  - Binning
  - Histogram analysis
  - Interval merging by χ^2 analysis
  - Cluster analysis
  - Entropy-based discretization
  - Discretization by “natural/intuitive partitioning”

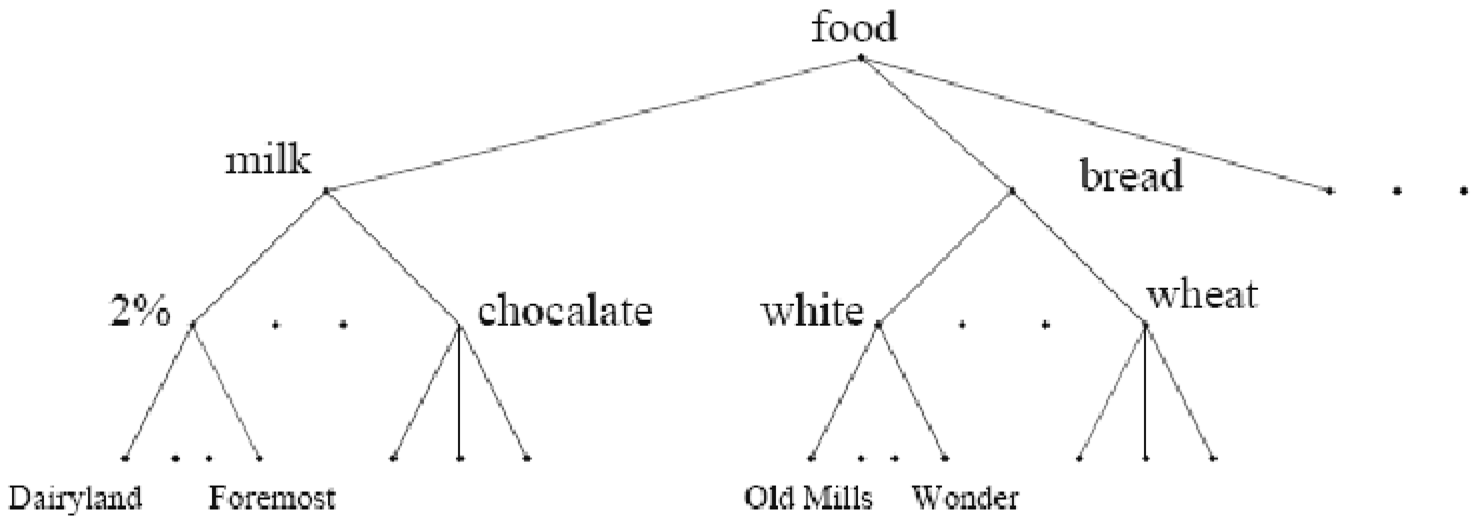
8. Create conceptual hierarchy

- Categorical data
  - Discrete data
  - Categorical attribute domain
    - ✓ Limited number of separate values
    - ✓ Not ordered

→ We can create a conceptual hierarchy for categorical data

8. CREATE CONCEPTUAL HIERARCHY

- Create conceptual hierarchy for categorical/discrete data
  - Describe a hierarchy by explicitly grouping data
  - Create hierarchies by predefined semantic connections



5. SUMMARY

- Real data: incomplete/missing, noisy, inconsistent,…
- Data preprocessing is required
  - Data cleansing: resolve missing data issues, smoothing, outlier detection, correct inconsistent data
  - Data integration: issues in entity identification, redundancy, data value conflicts
  - Data transformation: smoothing, aggregation, generalization, normalization, building new attributes/features
  - Data reduction: aggregated cube, attribute subset selection, dimensional reduction, discretization, conceptual hierarchy

5. SUMMARY

- Data discretization
  - Continuous values -> intervals -> label those intervals
  - Partioned hierarchy/multiresolution: on attribute values → phân cấp ý conceptual hierarchy for numerical attributes
- Conceptual hierarchy
  - Support data mining in multi levels of abstraction
  - Numerical attributes: binning, histogram analysis, entropy-based discretization, χ^2-merging, cluster analysis, discretization by intuitive partitioning
  - Categorical/discrete attributes: explicitly identify by users or experts, explicitly group data, based on number of separate data values of each attribute.

REFERENCE

[1] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2012.

[2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.

[3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.

[4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.

[5] ZhaoHui Tang, Jamie MacLennan, "Data Mining with SQL Server 2005", Wiley Publishing, 2005.

[6] Oracle, "Data Mining Concepts", B28129-01, 2008.

[7] Oracle, "Data Mining Application Developer's Guide", B28131-01, 2008.

[8] Ian H.Witten, Eibe Frank, "Data mining : practical machine learning tools and techniques", 2nd Edition, Elsevier Inc, 2005.

[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.

[10] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2nd Edition, Springer Science + Business Media, LLC 2005, 2010.

Q&A

quangtran@hcmut.edu.vn