Data Warehousing
資料倉儲

Data Preprocessing: Integration and the ETL process

1001DW03
MI4
Tue. 6,7 (13:10-15:00) B427

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.im.tku.edu.tw/~myday/

2011-09-20

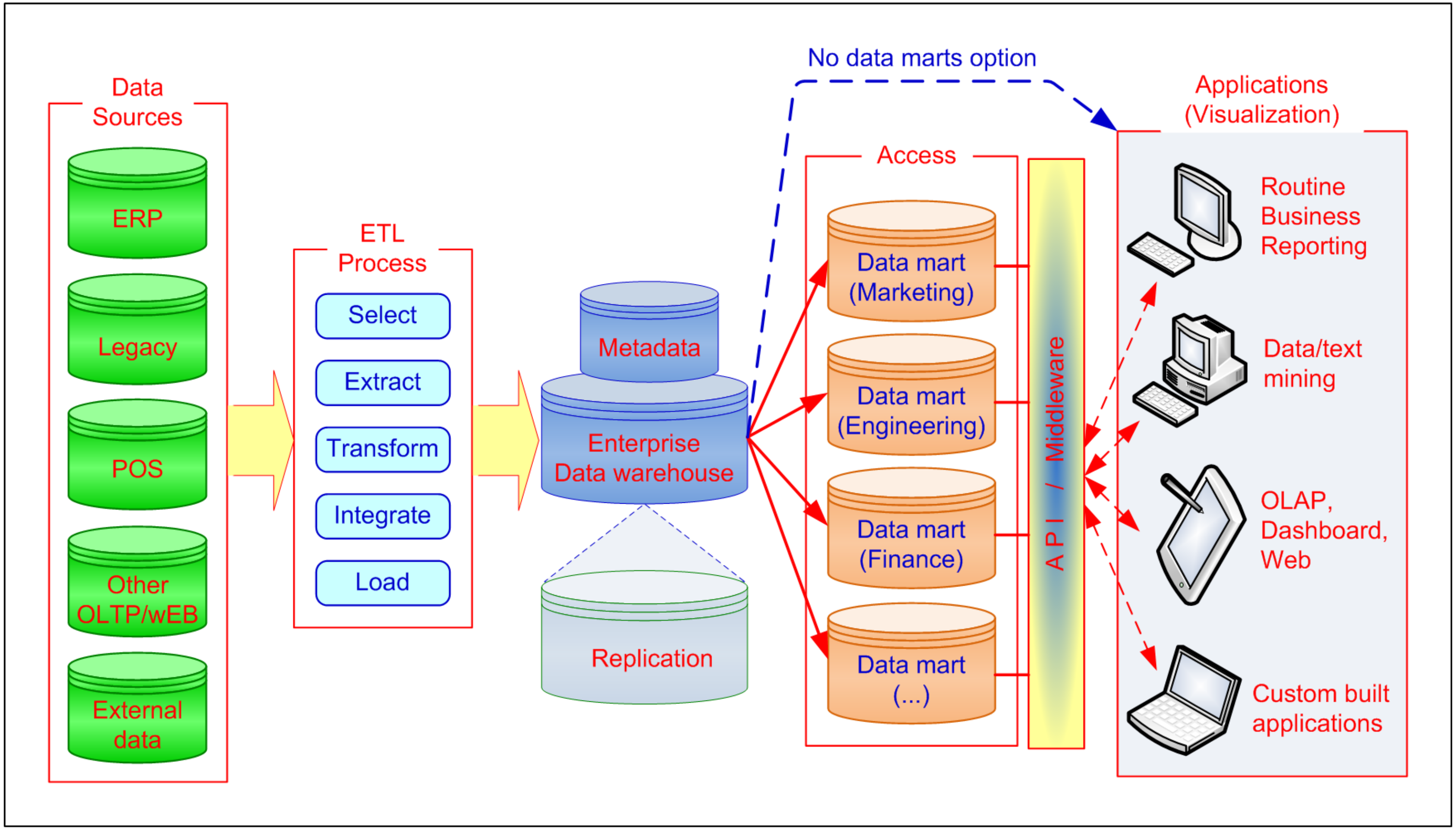
Syllabus

週次	日期	內容（Subject/Topics）
1	100/09/06	Introduction to Data Warehousing
2	100/09/13	Data Warehousing, Data Mining, and Business Intelligence
3	100/09/20	Data Preprocessing: Integration and the ETL process
4	100/09/27	Data Warehouse and OLAP Technology
5	100/10/04	Data Warehouse and OLAP Technology
6	100/10/11	Data Cube Computation and Data Generation
7	100/10/18	Data Cube Computation and Data Generation
8	100/10/25	Project Proposal
9	100/11/01	期中考試週

Syllabus

週次	日期	內容（Subject/Topics）
10	100/11/08	Association Analysis
11	100/11/15	Classification and Prediction
12	100/11/22	Cluster Analysis
13	100/11/29	Sequence Data Mining
14	100/12/06	Social Network Analysis
15	100/12/13	Link Mining
16	100/12/20	Text Mining and Web Mining
17	100/12/27	Project Presentation
18	101/01/03	期末考試週

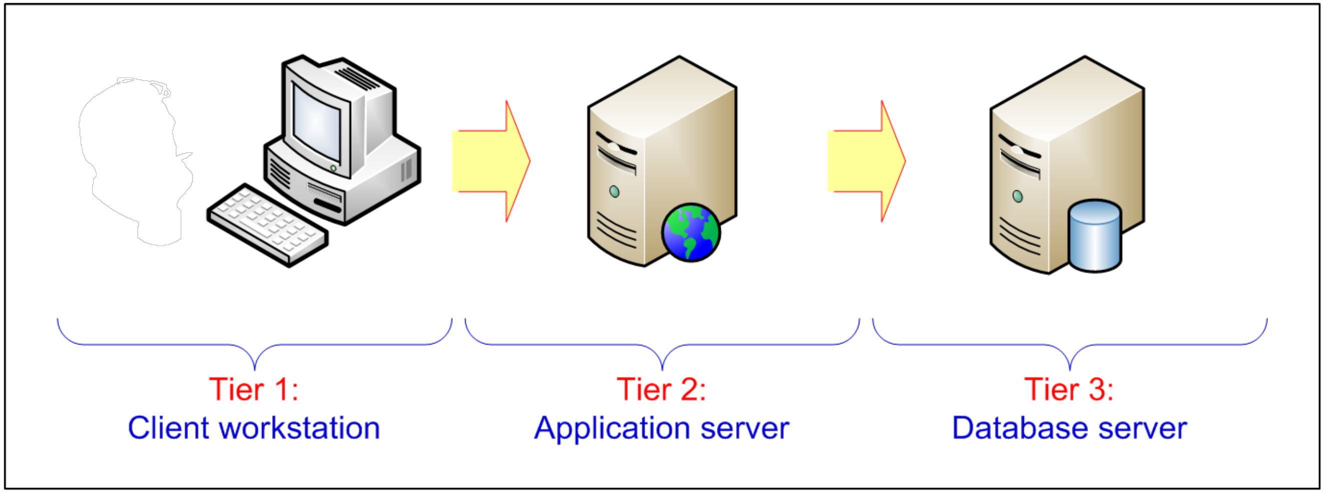
A Conceptual Framework for Data Warehousing



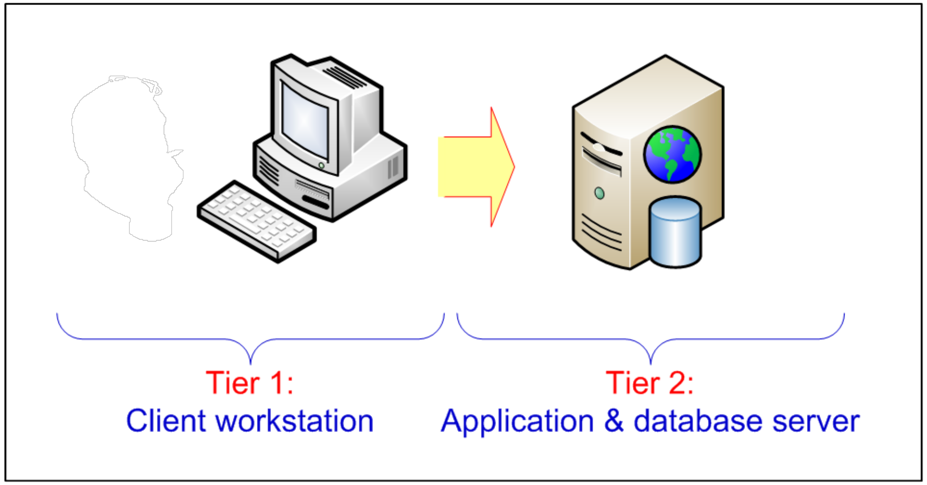
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Generic DW Architectures

3-tier architecture



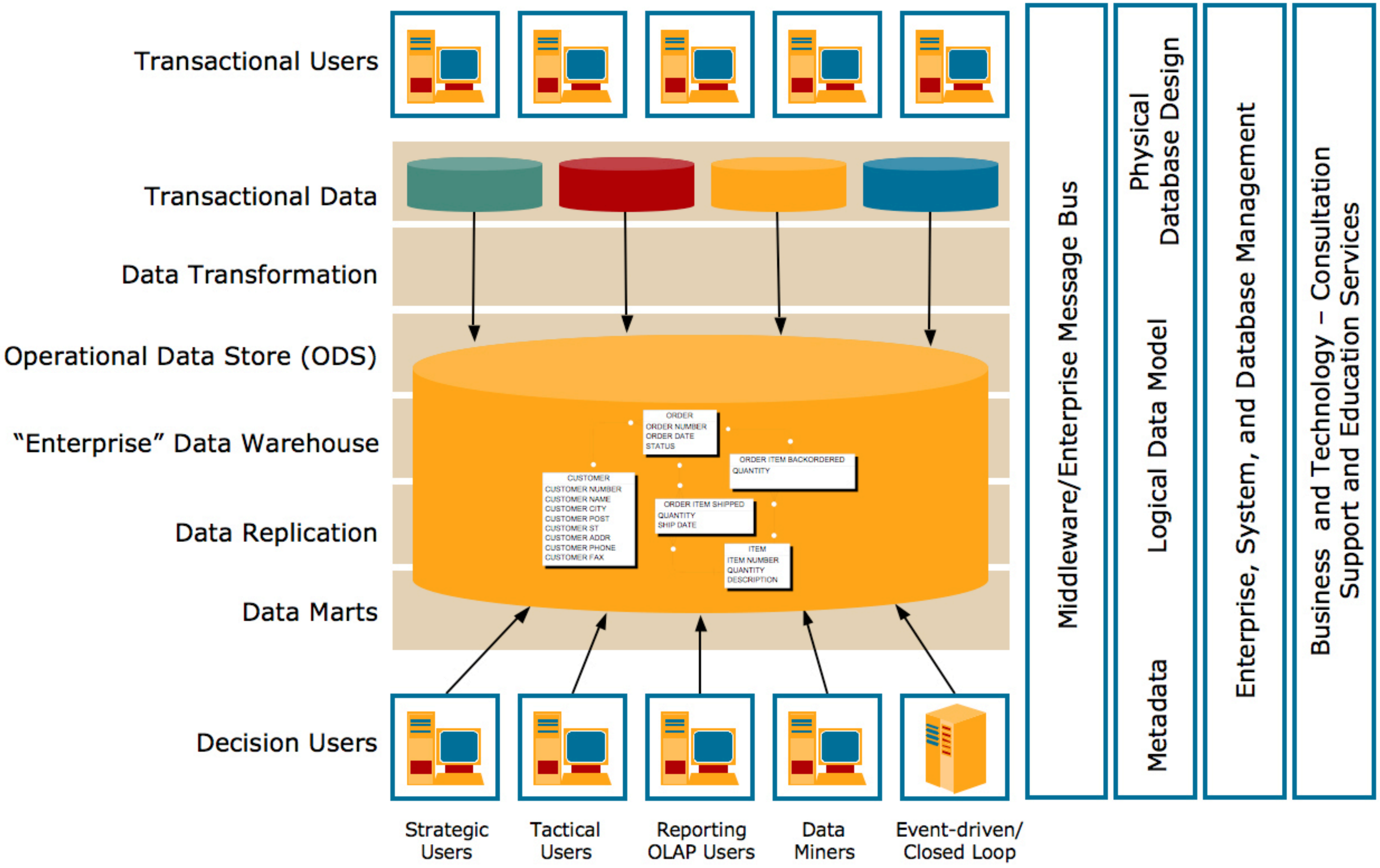
2-tier architecture



1-tier Architecture ?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Enterprise Data Warehouse (by Teradata Corporation)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Integration and the Extraction, Transformation, and Load (ETL) Process

- **Data integration**

  Integration that comprises three major processes: data access, data federation, and change capture.
- **Enterprise application integration (EAI)**

  A technology that provides a vehicle for pushing data from source systems into a data warehouse
- **Enterprise information integration (EII)**

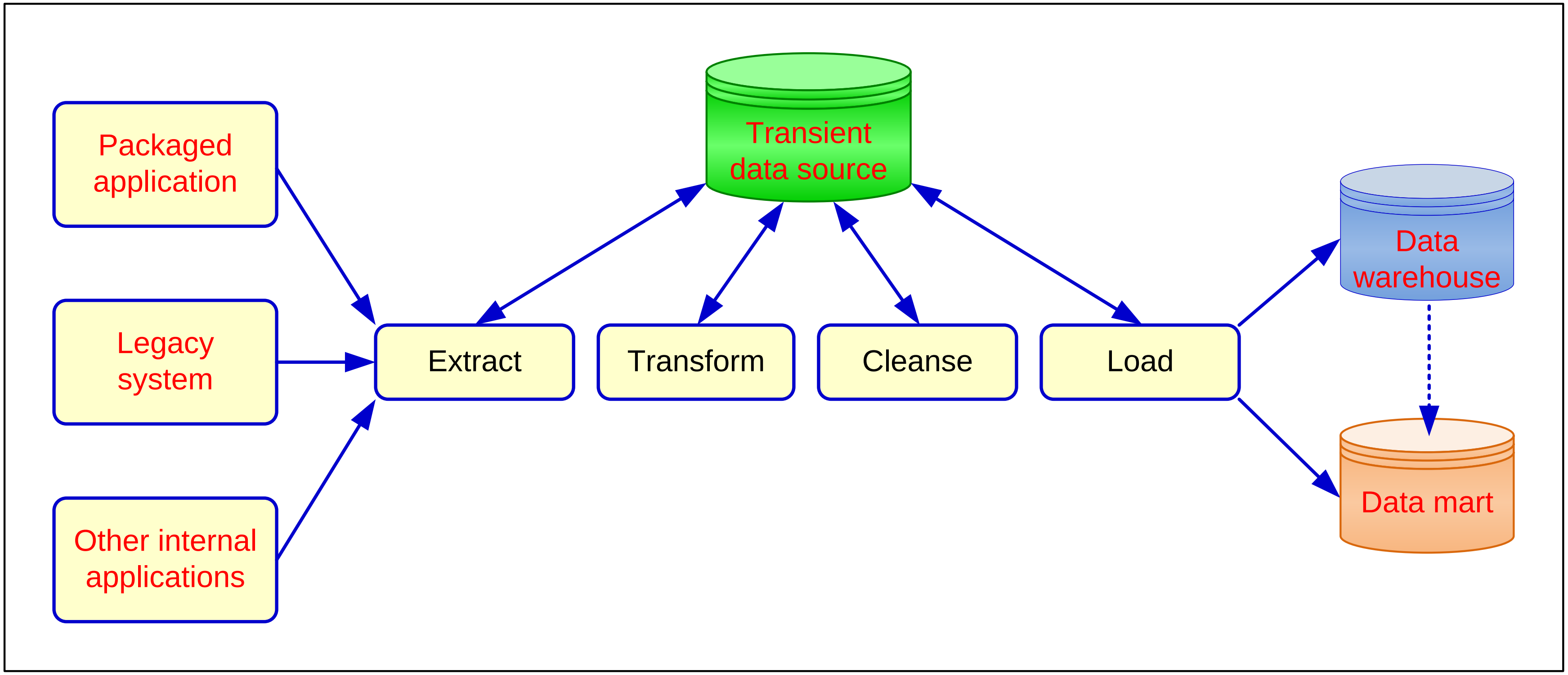
  An evolving tool space that promises real-time data integration from a variety of sources
- **Service-oriented architecture (SOA)**

  A new way of integrating information systems

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Integration and the Extraction, Transformation, and Load (ETL) Process

Extraction, Transformation, and Load (ETL) process



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

ETL

- Issues affecting the purchase of and ETL tool
  - Data transformation tools are expensive
  - Data transformation tools may have a long learning curve
- Important criteria in selecting an ETL tool
  - Ability to read from and write to an unlimited number of data sources/architectures
  - Automatic capturing and delivery of metadata
  - A history of conforming to open standards
  - An easy-to-use interface for the developer and the functional user

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

Source: Han & Kamber (2006)

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Source: Han & Kamber (2006)

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Source: Han & Kamber (2006)

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility
- Broad categories:
  - Intrinsic, contextual, representational, and accessibility

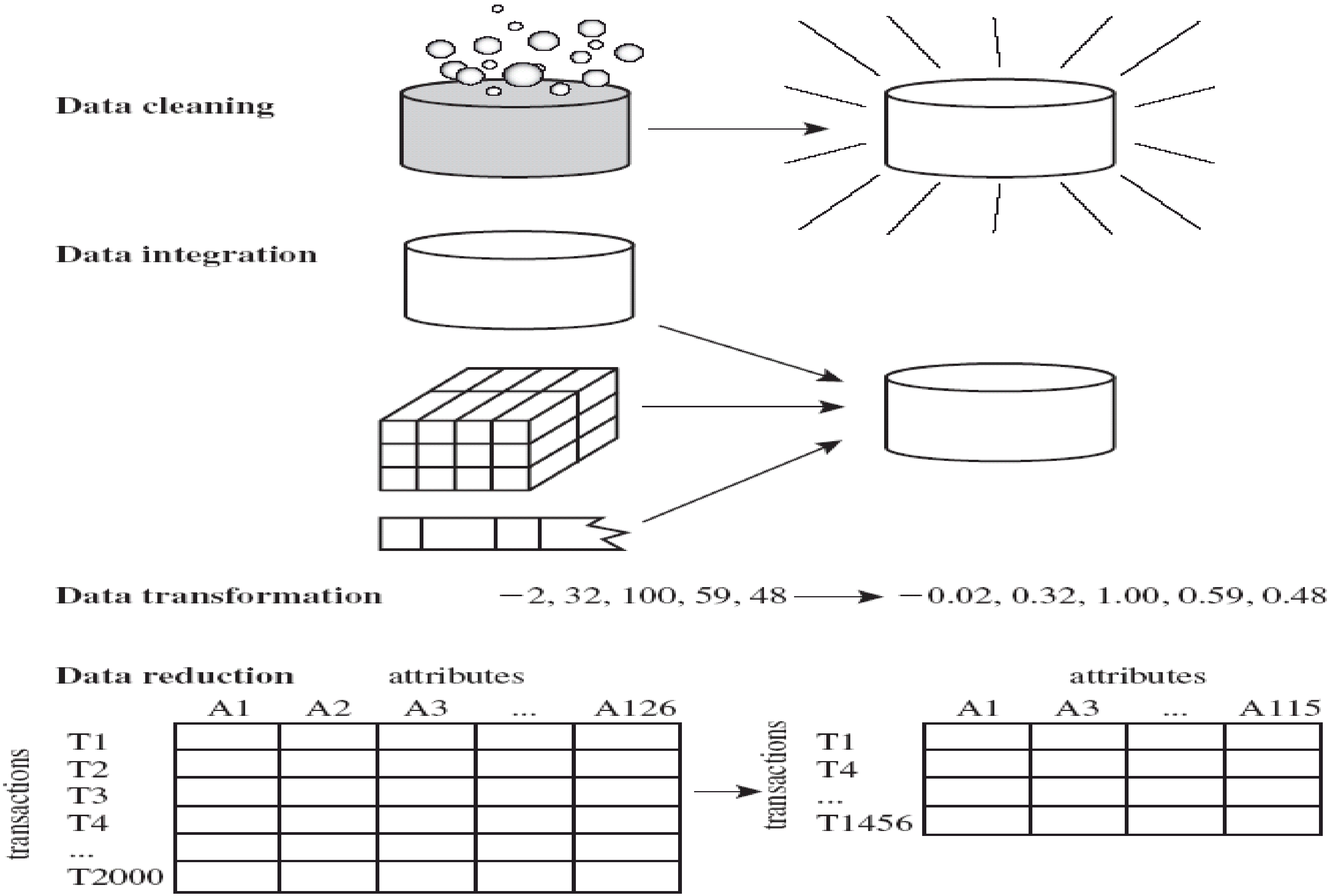
Source: Han & Kamber (2006)

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Source: Han & Kamber (2006)

Forms of Data Preprocessing



Source: Han & Kamber (2006)

Mining Data Descriptive Characteristics

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Source: Han & Kamber (2006)

Data Cleaning

- Importance
  - “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
  - “Data cleaning is the number one problem in data warehousing”—DCI survey

Source: Han & Kamber (2006)

Data cleaning tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy caused by data integration

Source: Han & Kamber (2006)

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

Source: Han & Kamber (2006)

How to Handle Missing Data?

- Ignore the tuple
- Fill in the missing value manually
- Fill in it automatically with
  - a global constant
    - e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Source: Han & Kamber (2006)

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

Source: Han & Kamber (2006)

How to Handle Noisy Data?

1. Binning
   - first sort data and partition into (equal-frequency) bins
   - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
2. Regression
   - smooth by fitting the data into regression functions
3. Clustering
   - detect and remove outliers

- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Source: Han & Kamber (2006)

Simple Discretization Methods: 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
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Source: Han & Kamber (2006)

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 (equi-depth) bins:

  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

* Smoothing by bin means:

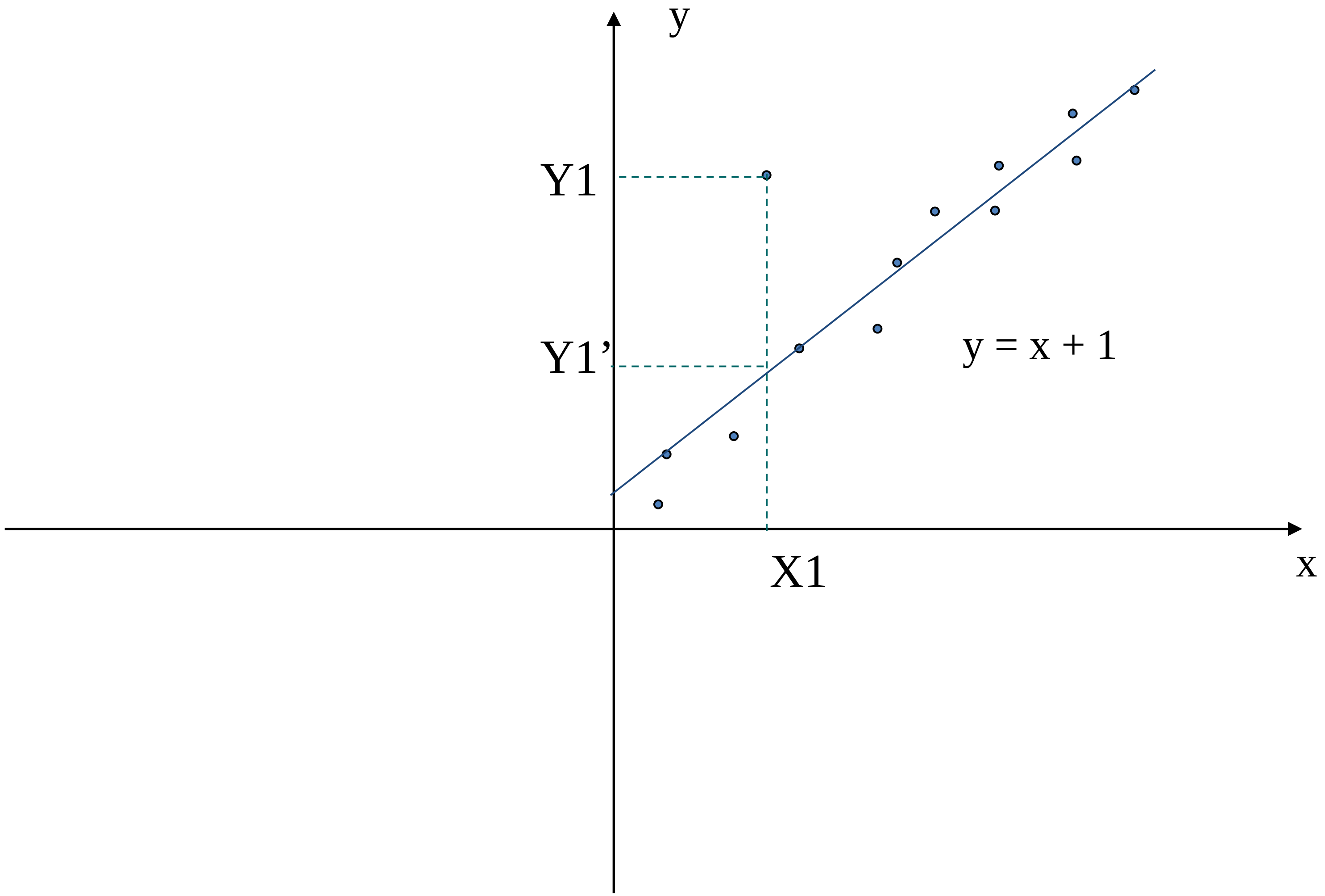
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:

  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

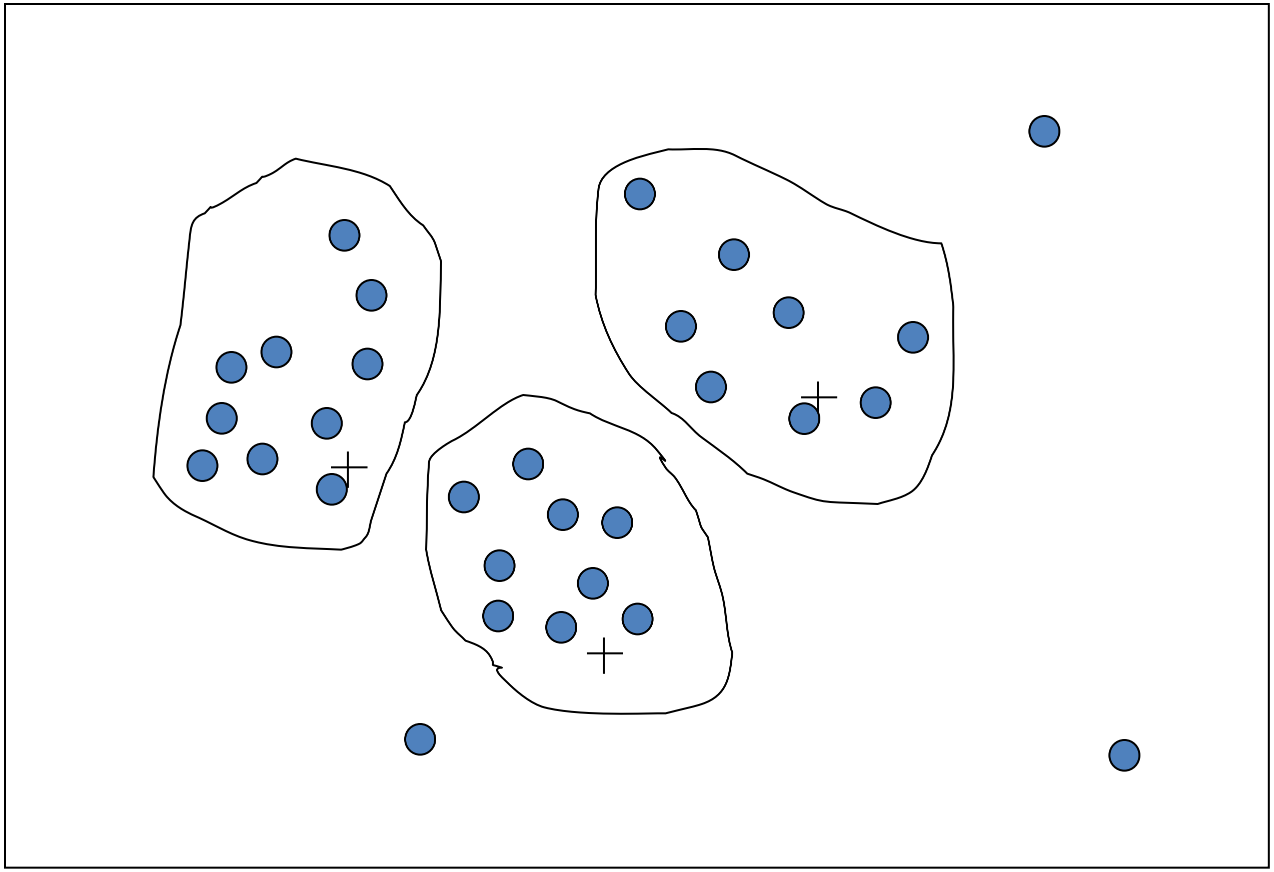
Source: Han & Kamber (2006)

Regression



Source: Han & Kamber (2006)

Cluster Analysis



Source: Han & Kamber (2006)

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata
    - (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing
      - use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing
      - by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

Source: Han & Kamber (2006)

Data Cleaning as a Process

- Data migration and integration
  - Data migration tools
    - allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools
    - allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

Source: Han & Kamber (2006)

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Source: Han & Kamber (2006)

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Source: Han & Kamber (2006)

Correlation Analysis (Numerical Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum (A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\overline{A}\overline{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, $\overline{A}$ and $\overline{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and Σ(AB) is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated

Source: Han & Kamber (2006)

Correlation Analysis (Categorical Data)

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Source: Han & Kamber (2006)

Chi-Square Calculation: An Example

$$e_{11} = \frac{300 \times 450}{1500} = 90$$

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

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

- It shows that like_science_fiction and play_chess are correlated in the group

Source: Han & Kamber (2006)

Data Transformation

- Smoothing
  - remove noise from data
- Aggregation
  - summarization, data cube construction
- Generalization
  - concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Source: Han & Kamber (2006)

Data Transformation: 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 $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- 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(|ν'|) < 1

Source: Han & Kamber (2006)

Why data reduction?

- A database/data warehouse may store terabytes of data
- Complex data analysis/mining may take a very long time to run on the complete data set

Source: Han & Kamber (2006)

Data reduction

- Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

Source: Han & Kamber (2006)

Data Reduction Strategies

- Data cube aggregation:
- Dimensionality reduction
  - e.g., remove unimportant attributes
- Data Compression
- Numerosity reduction
  - e.g., fit data into models
- Discretization and concept hierarchy generation

Source: Han & Kamber (2006)

Data Cube Aggregation

- The lowest level of a data cube (base cuboid)
  - The aggregated data for an individual entity of interest
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

Source: Han & Kamber (2006)

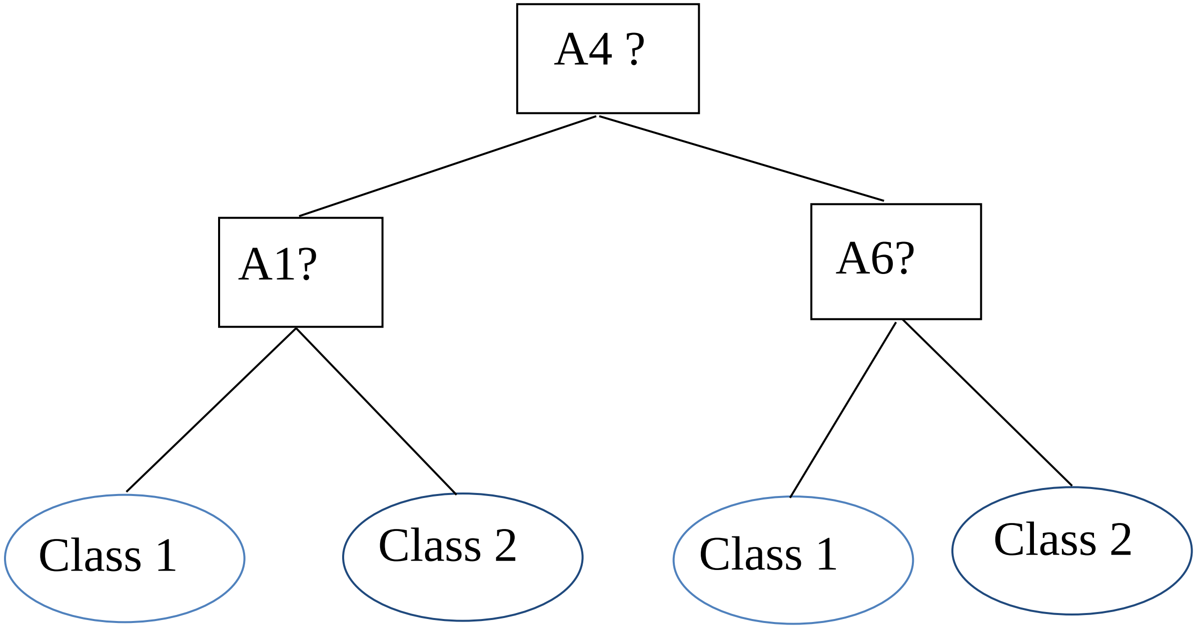
Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination
  - Decision-tree induction

Source: Han & Kamber (2006)

Example of Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



-------> Reduced attribute set: {A1, A4, A6}

Source: Han & Kamber (2006)

Heuristic Feature Selection Methods

- There are *2^d* possible sub-features of *d* features
- Several heuristic feature selection methods:
  - Best single features under the feature independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination:
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination
  - Optimal branch and bound:
    - Use feature elimination and backtracking

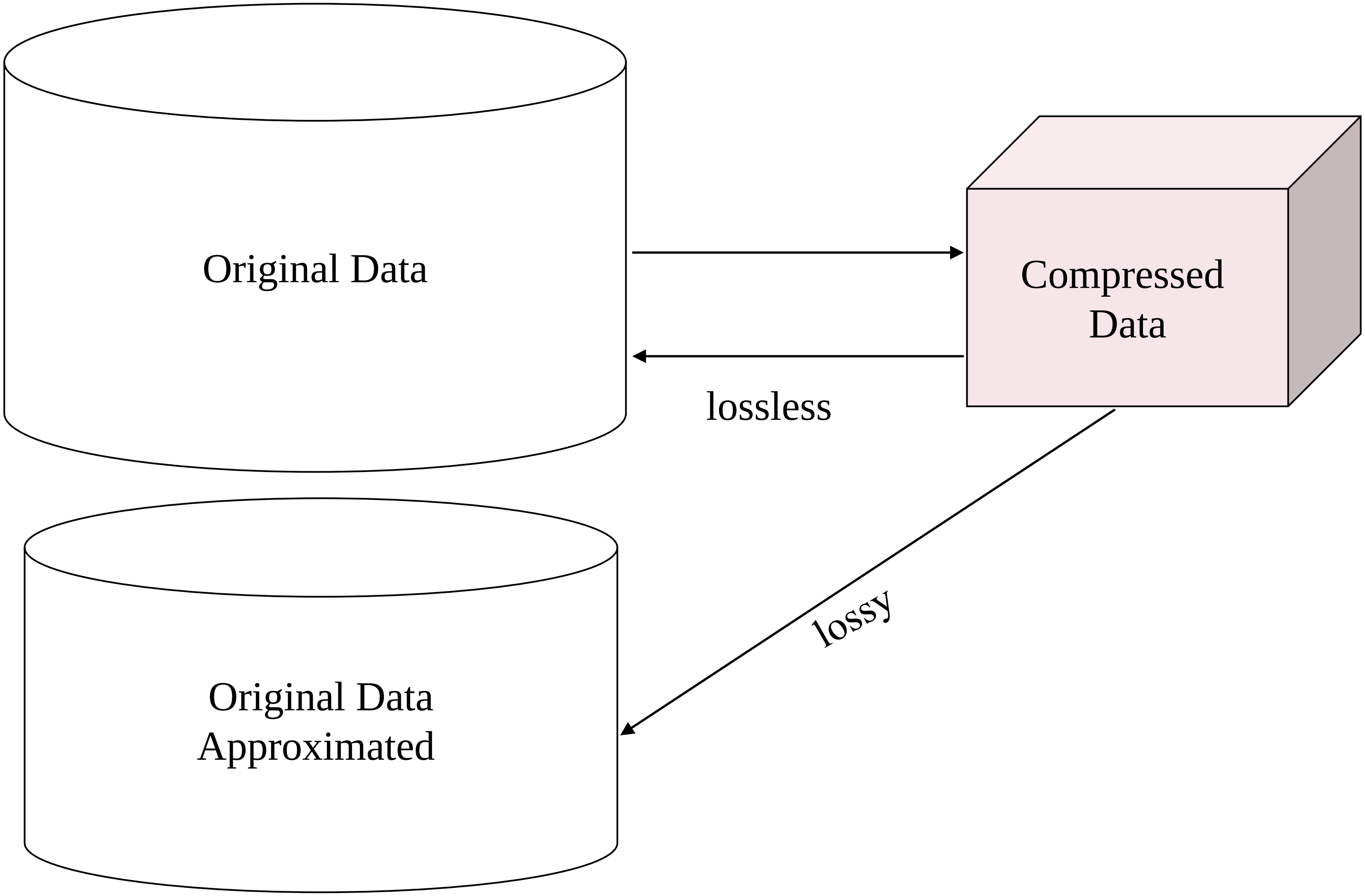
Source: Han & Kamber (2006)

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
- Time sequence is not audio
  - Typically short and vary slowly with time

Source: Han & Kamber (2006)

Data Compression



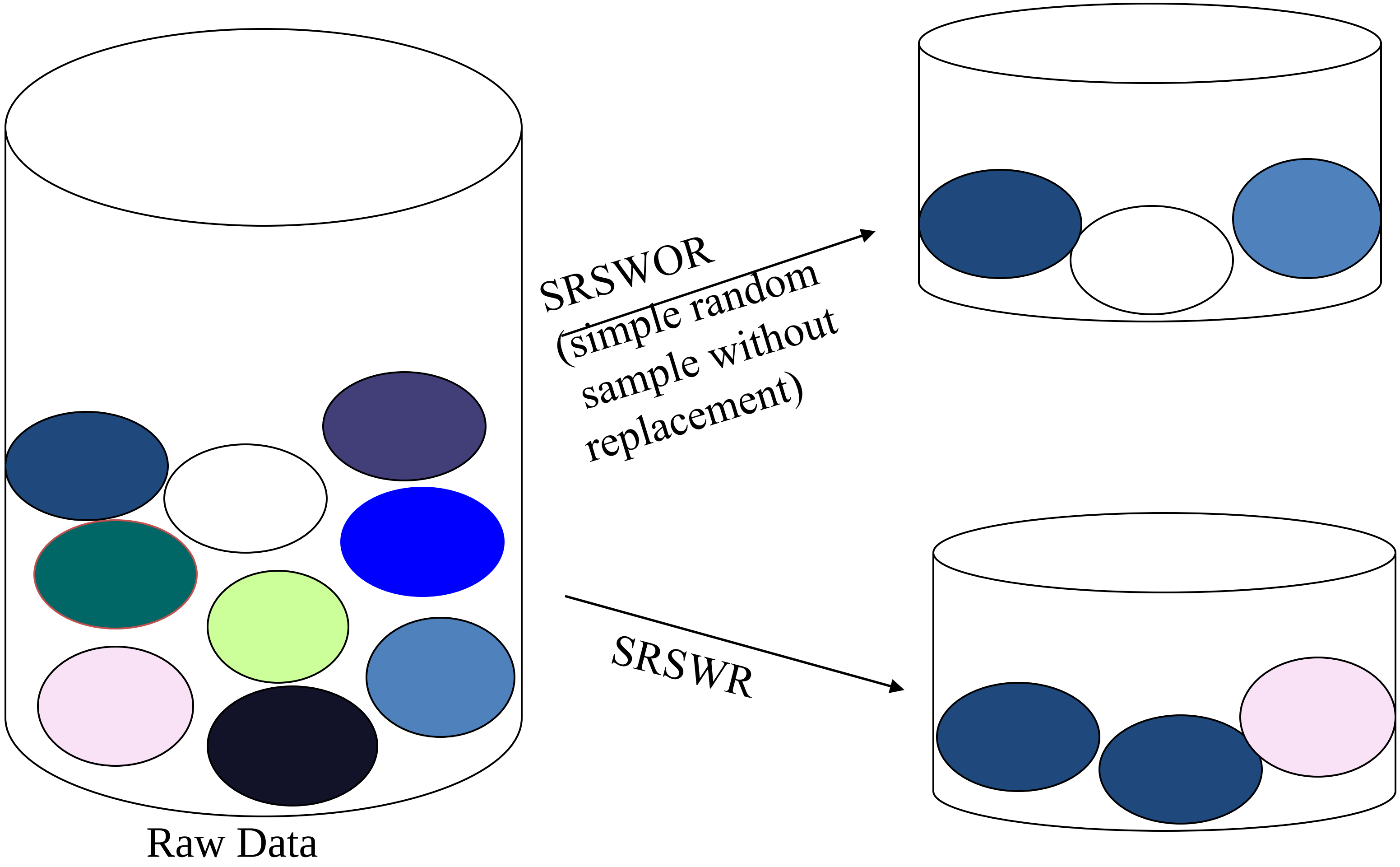
Source: Han & Kamber (2006)

Data Reduction Method: Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Note: Sampling may not reduce database I/Os (page at a time)

Source: Han & Kamber (2006)

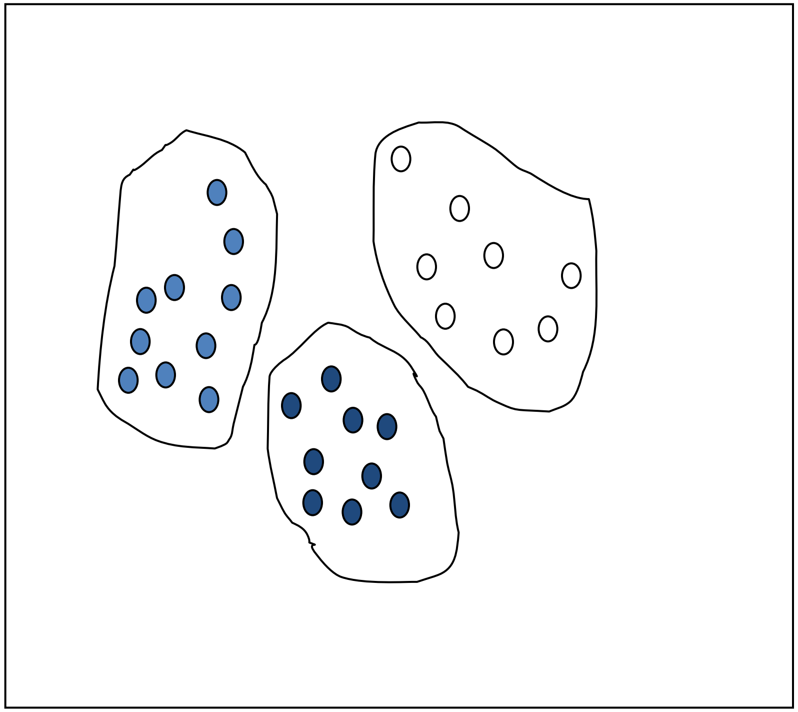
Sampling: with or without Replacement



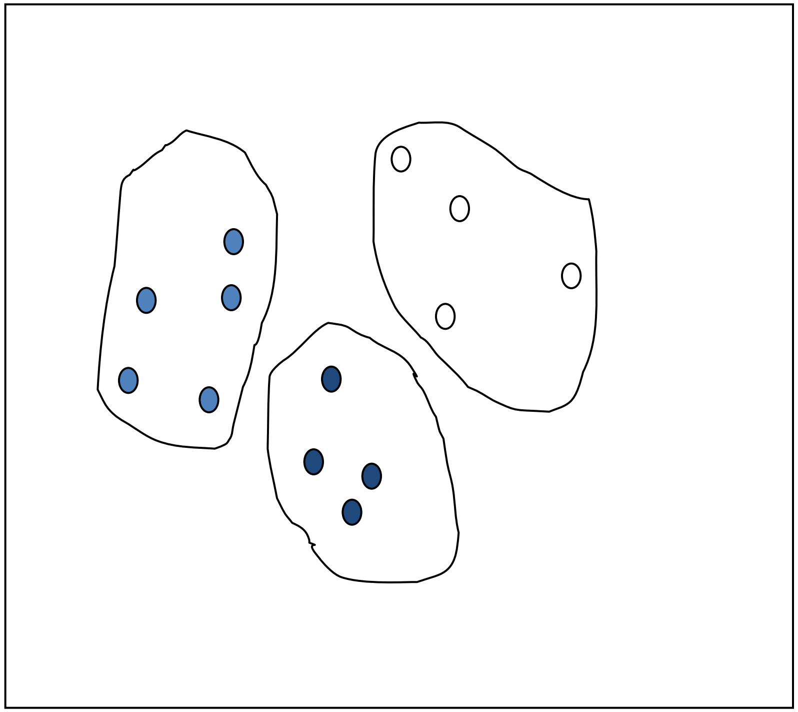
Source: Han & Kamber (2006)

Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Source: Han & Kamber (2006)

Discretization

- Three types of attributes:
  - Nominal — values from an unordered set, e.g., color, profession
  - Ordinal — values from an ordered set, e.g., military or academic rank
  - Continuous — real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Source: Han & Kamber (2006)

Discretization and Concept Hierarchy

- Discretization
  - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
- Concept hierarchy formation
  - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)

Source: Han & Kamber (2006)

Discretization and Concept Hierarchy Generation for Numeric Data

- Typical methods: All the methods can be applied recursively
  - Binning (covered above)
    - Top-down split, unsupervised,
  - Histogram analysis (covered above)
    - Top-down split, unsupervised
  - Clustering analysis (covered above)
    - Either top-down split or bottom-up merge, unsupervised
  - Entropy-based discretization: supervised, top-down split
  - Interval merging by χ^2 Analysis: unsupervised, bottom-up merge
  - Segmentation by natural partitioning: top-down split, unsupervised

Source: Han & Kamber (2006)

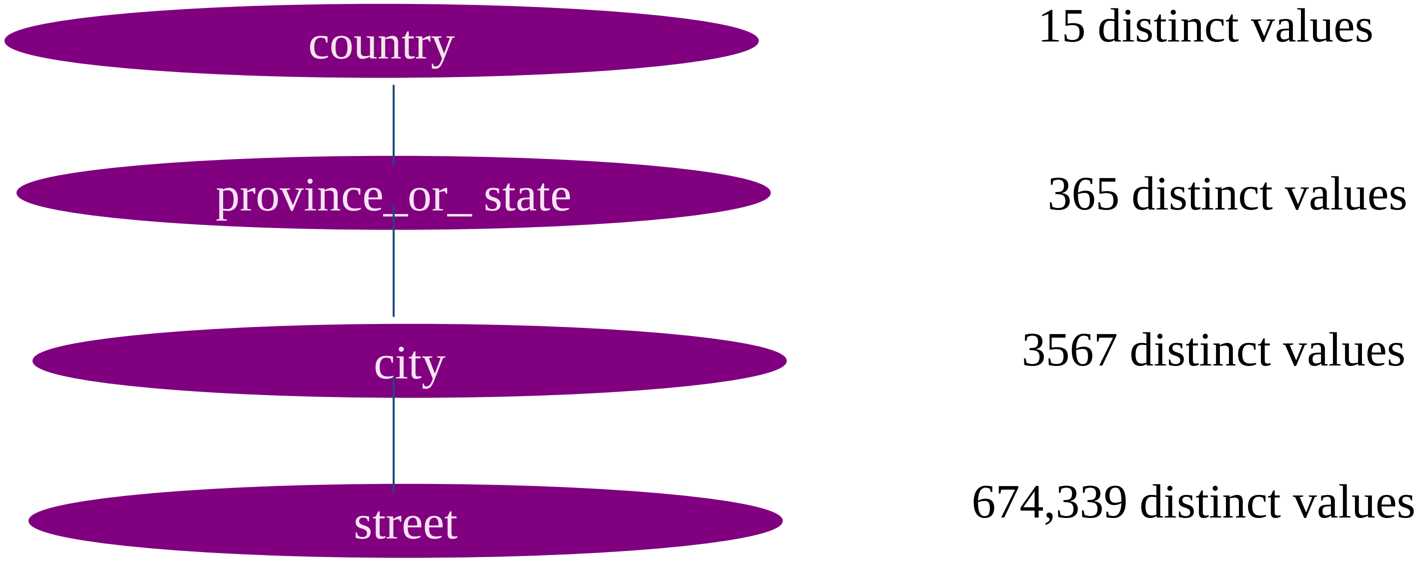
Concept Hierarchy Generation for Categorical Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only street < city, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {street, city, state, country}

Source: Han & Kamber (2006)

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



country — 15 distinct values

province_or_ state — 365 distinct values

city — 3567 distinct values

street — 674,339 distinct values

Source: Han & Kamber (2006)

Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Discriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization

Source: Han & Kamber (2006)

References

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier
- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, 2011, Pearson.