KDD Phases

- Terminology – Misnomer!
- Data Mining Phases

Cleaning

Integration

Selection

Transformation

Data Mining

Pattern Evaluation

Knowledge Presentation

Data Preprocessing
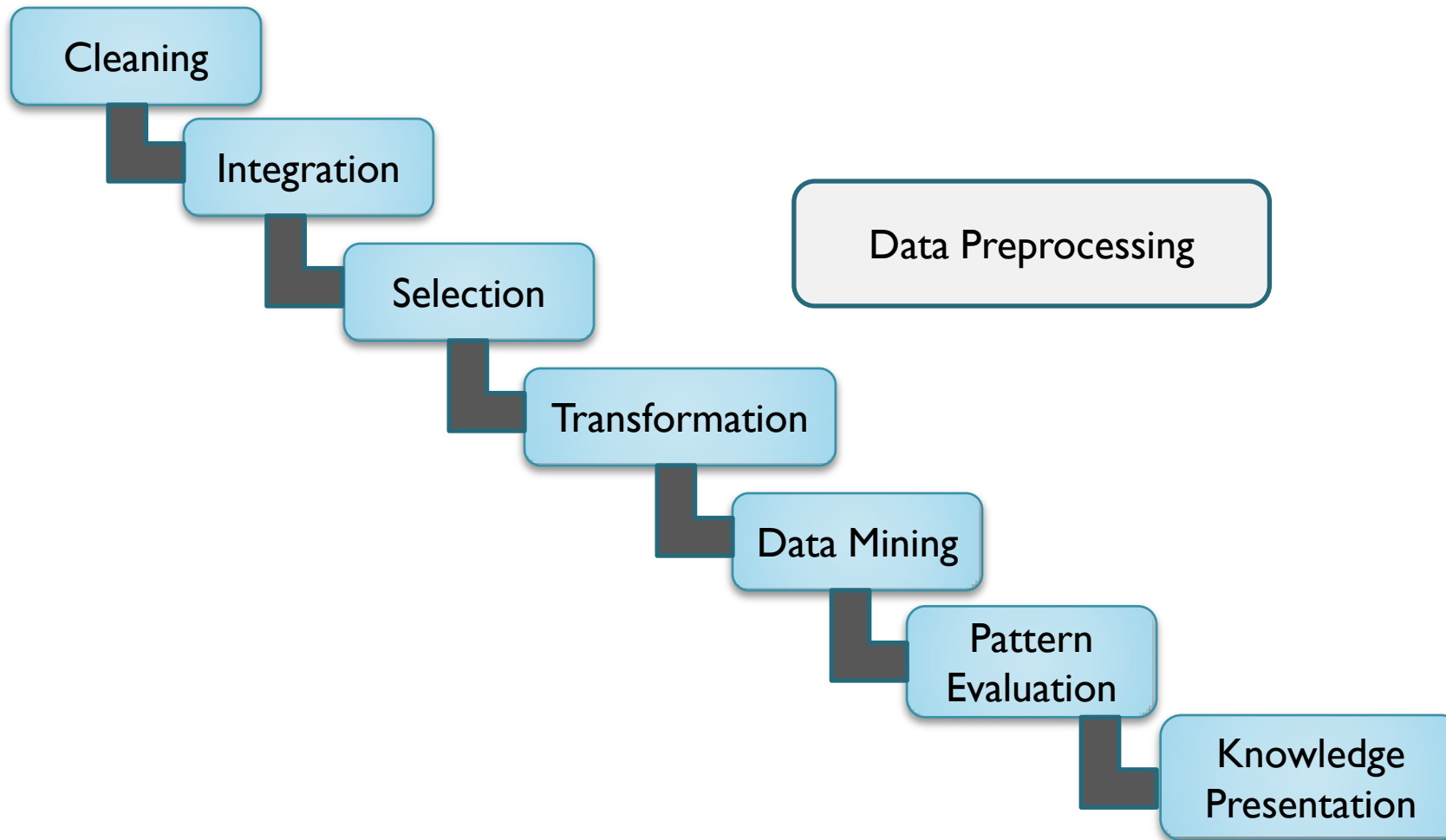
Data Mining Techniques

- Data Classification
- Association Rule Mining
- Clustering
- Prediction
- Outlier Mining
- Multimedia Data Mining – Text / Audio / Video

Issues in Data Mining

- Pattern Interestingness ?
- Knowledge Presentation – Decision Trees, Rules
- Major Issues in Data Mining:
  - Efficiency / Scalability / Parallelization of / Incremental Data Mining Techniques
  - Multiple Levels of Abstraction of Data
    - Drill Down
    - Roll Up

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view

  ◦ Accuracy: correct or wrong, accurate or not

  ◦ Completeness: not recorded, unavailable, …

  ◦ Consistency: some modified but some not, dangling, …

  ◦ Timeliness: timely update?

  ◦ Believability: how trustable the data are correct?

  ◦ Interpretability: how easily the data can be understood?

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 reduction**
  - ◦ Dimensionality reduction
  - ◦ Numerosity reduction
  - ◦ Data compression
- **Data transformation and data discretization**
  - ◦ Normalization
  - ◦ Concept hierarchy generation

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=" " (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*="−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*="42", *Birthday*="03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Incomplete (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

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably

- Fill in the missing value manually: tedious + infeasible?

- 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

Noisy Data

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

How to Handle Noisy Data?

- 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.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

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

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* and *covariance analysis*

- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Reduction Strategies

- **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/data warehouse may store terabytes 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
    - Wavelet transforms
    - 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 cube aggregation
  - Data compression

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

Data Transformation

- 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, data cube construction

  ◦ Normalization: Scaled to fall within a smaller, specified range

    • min-max normalization

    • z-score normalization

    • normalization by decimal scaling

  ◦ Discretization: Concept hierarchy climbing

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 $73,600 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($|v'|$) < 1

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

Concept Hierarchy Generation for Nominal 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*}

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
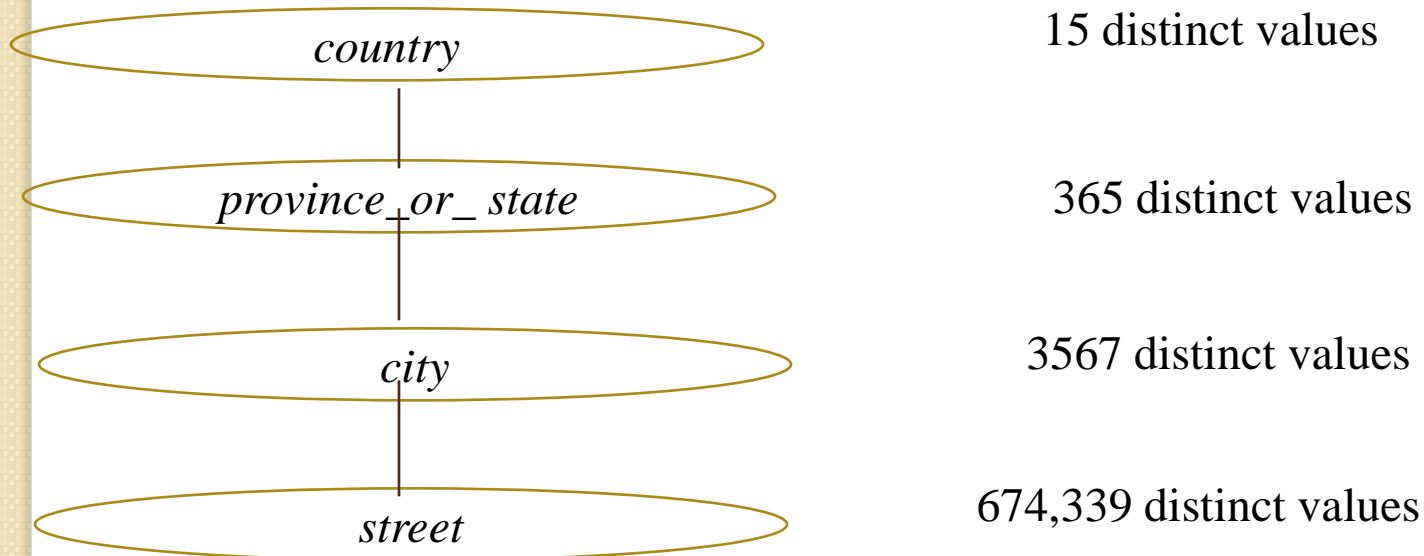
Perspectives of Data Mining

- Induction
- Data Compression
- Approximation
- Search and
- Querying