Data Mining

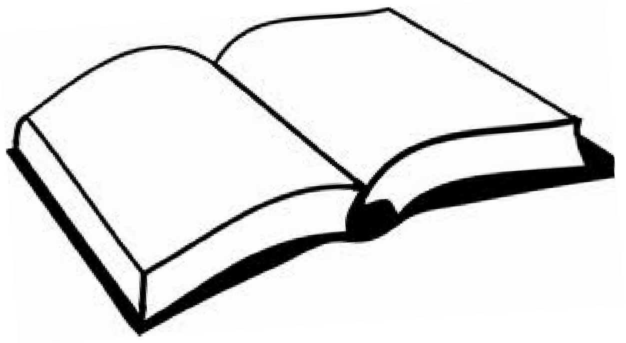
DATA PREPROCESSING

Prof. Dr. Hikmat Ullah Khan
Department of Information Technology
UNIVERSITY OF SARGODAHA

Lesson from Holy Quran

Outline

- Introduction to Data Preprocessing
- Data Quality
- Steps of Data Preprocessing
  - Data cleaning
  - Data integration
  - Data reduction
  - Data transformation



Data Preprocessing

Data Quality

- Measures for data quality:
  - **Accuracy:** correct or wrong
  - **Completeness:** not recorded, unavailable, …
  - **Consistency**: some updated/modified but some not.
  - **Timeliness:** timely update?
  - **Believability:** how trustable the data is?
  - **Interpretability:** how easily 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 or files, diverse sources
- **Data reduction**
  - Dimensionality reduction
  - Data compression
- **Data transformation**
  - Normalization

Data Cleaning

- Data in the Real World Is Dirty: (More thanks to Social Web)
- Lots of potentially incorrect data, e.g., human or computer error, extraction error
  - <u>incomplete</u>: lacking attribute values,
    - *e.g., Occupation=" " (missing data)*
  - <u>noisy</u>: containing noise, errors
    - e.g., *Salary*="−10" (an error)
  - <u>inconsistent</u>: containing discrepancies in codes or names, e.g.,
    - *Age*="42", *Birthday*="03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
  - <u>Intentional</u> (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

How to Handle Missing Data?

- Ignore the tuple:
  - usually done when class label is missing
- 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 median value

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - e.g., Bin ages of the students of undergraduate
  - smooth by bin means, smooth by bin median, 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)

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

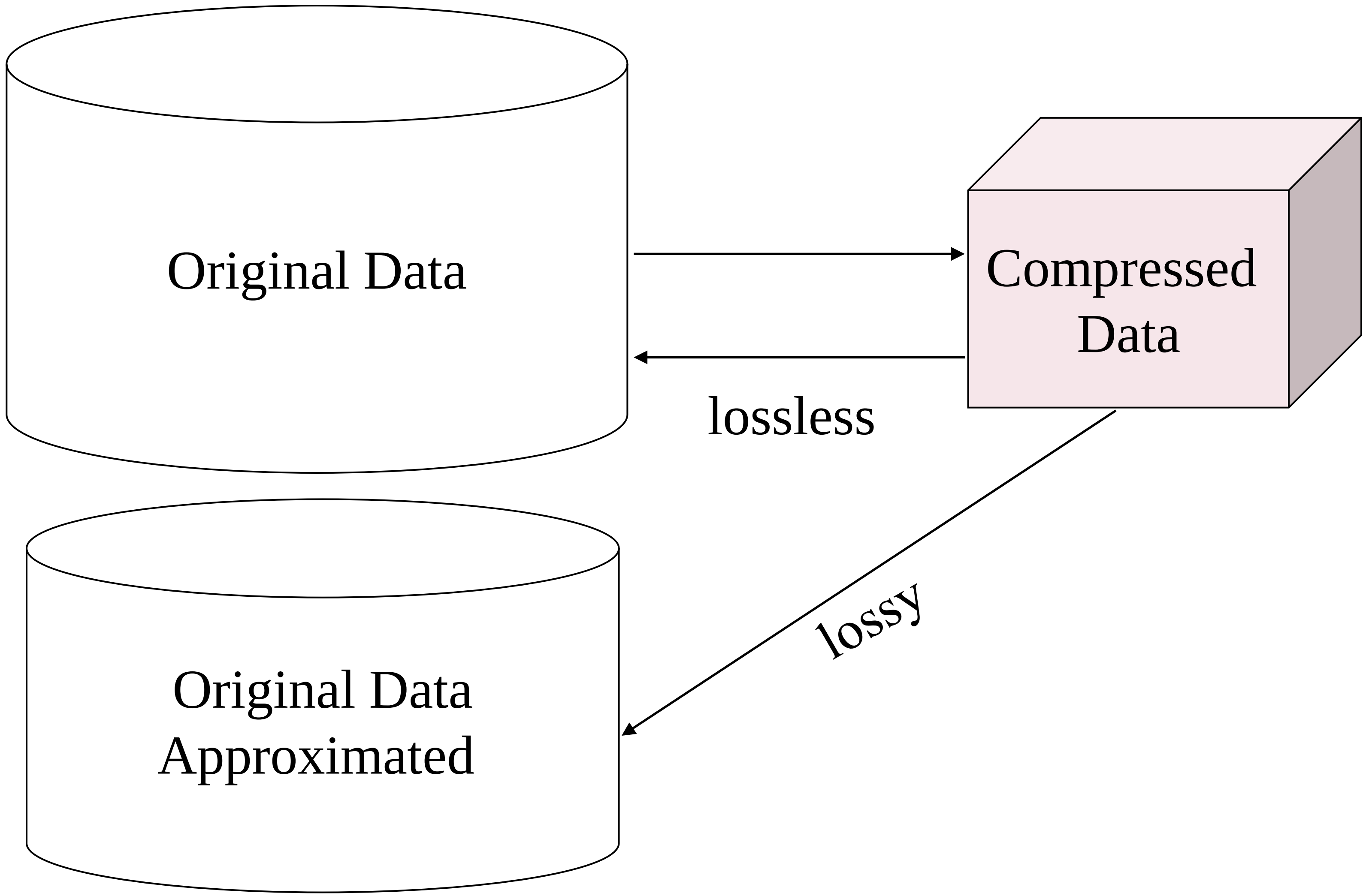
Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id $\equiv$ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem (Name Disambiguation)
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - Possible reasons:
  - different representations: Rs vs. US Dollars
  - different scales, e.g., metric vs. British units

Data Reduction Strategies

- **Data reduction**:
  - Obtain a reduced representation
  - Produces the same (or almost the same)
- Why data reduction?
  - Huge volume (terabytes)
  - Complex data – difficult to analysis
  - Time consuming -
- Data reduction strategies
  - Dimensionality reduction, e.g., remove unimportant attributes
  - Feature subset selection algorithms
    - Info Gain
    - Principal Components Analysis (PCA)

Data Compression



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
  - **Attribute/feature construction**
    - **Derived attributes** constructed from the given ones
    - E.g. Age as new attribute instead of Date of Birth
  - **Normalization:**
  - Scaled to fall within a smaller, specified range
    - min-max normalization

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

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

  - E.g., Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to

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


WHAT YOU SEEK
IS SEEKING YOU
~
RUMI