Knowledge Discovery & Data Mining

– Data Preprocessing –

Data Cleaning & Data Integration

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Based on the original version by Professor Jiliang Tang

Recall

- Basic statistical data description
  - Central Tendency
    - Mean, Median, Mode, Midrange
    - Symmetric vs. Skewed Data
  - Dispersion of Data
    - Range, Quantiles, Quartiles, Interquartile Range(IQR),
    - Variance, Standard Deviation

Recall

- Graphic displays of basic statistics of data
  - Quantile plots
  - Quantile-Quantile plots
  - Histograms

  } Univariate distributions

  - Scatter plots → Bivariate distributions

Outline

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration

Why Preprocess the Data? Data Quality Matters

Source: https://www.researchgate.net/figure/Data-quality-and-standards-garbage-in-data-garbage-out-results_fig4_333491695

Measures for Data Quality

- **Accuracy:** correct or wrong, accurate or not
- **Completeness:** not recorded, unavailable, …
- **Consistency:** some modified but some not, …
- **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
- Resolve inconsistencies

Data Integration

- Databases
- Data cubes
- Files

Data Reduction

- Dimensionality reduction
- Numerosity reduction
- Data Compression

Data Transformation

- Normalization
- Discretization

Outline

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - **Missing values:** Attribute values are missing, some attributes of interest are missing, or only aggregate data is included:
    - e.g., Occupation=“ ” (missing data)
  - **Noisy:** Contains noise, errors, or outliers:
    - e.g., Salary=“−10” (an error)
  - **Inconsistent:** Contains differences in codes or names, e.g.,
    - Age=“42”, Birthday=“03/07/2010”
    - Was rating "1, 2, 3", now rating "A, B, C"
    - Differences between duplicate records
  - **Intentional** (e.g., disguised missing data)
    - Jan. 1 as everyone’s birthday

Data Cleaning: Missing Values

- Attribute values unavailable when collecting data
  - Usually encoded as null values in the database

index	Name	Age	Gender	Salary
0	John Doe	28.0	Male	50000.0
1	Jane Smith	NaN	Female	60000.0
2	Alice Johnson	35.0	NaN	NaN
3	NaN	22.0	Male	45000.0
4	Chris Ray	NaN	Male	70000.0

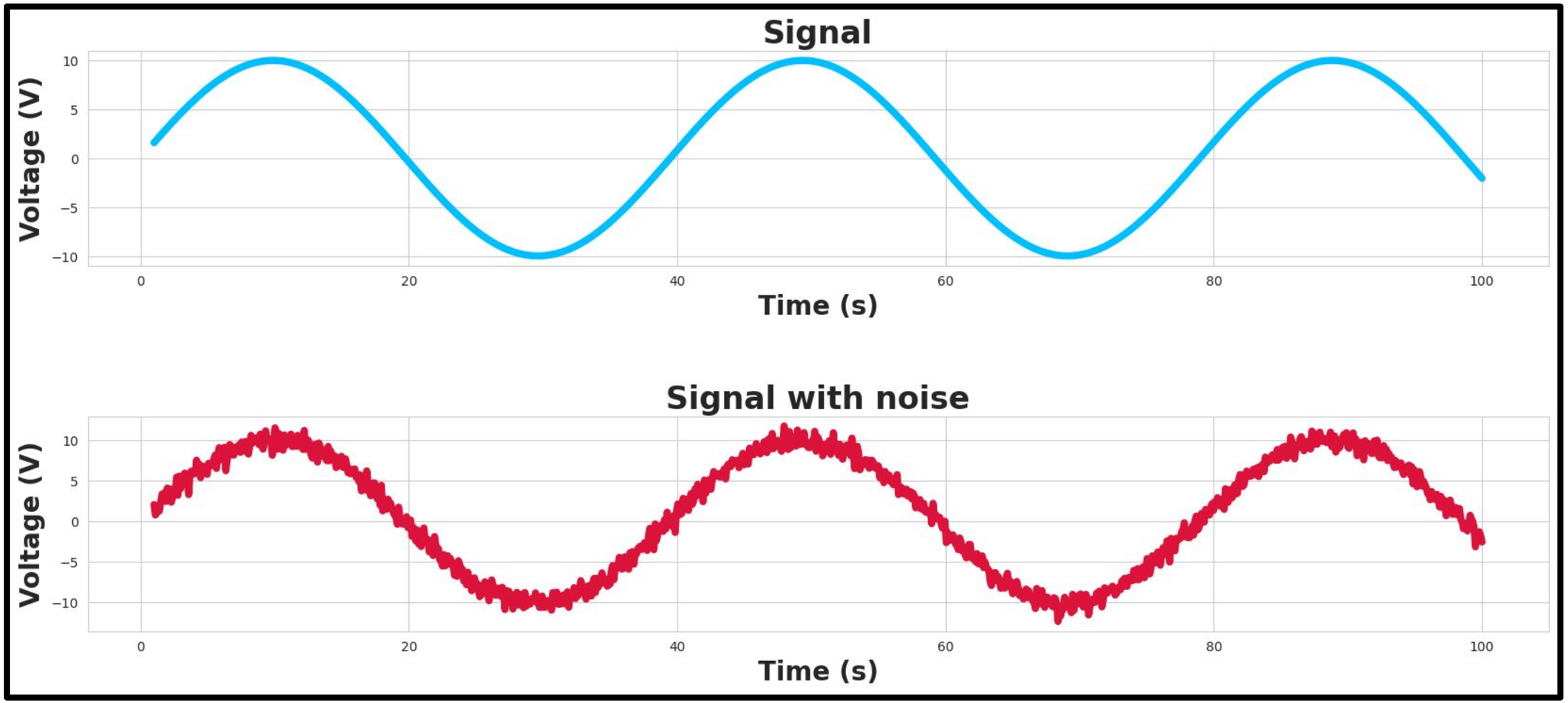
- Examples:
  - The equipment used to gather the data might not work properly.
  - Some data might not match other data, so it's removed.
  - Maybe someone didn't understand how to input the data.
  - At times, people might not think some data is important, so they don't add it.
  - The data's history or any changes might not be recorded.

How to Handle Missing Data?

- **Ignore the tuple:** Discard all data objects with missing values
  - not effective when the % of missing values per attribute varies greatly.
- **Fill in the missing value manually:**
  - is time consuming and may not be feasible given a large data set with many missing values.
- **Fill in it automatically with**
  - a global constant : e.g., Replace all missing attribute values by “unknown”.
    - the mining program may mistakenly think that they form an interesting concept.
  - the attribute mean, median, or mode
- **Model-based approach**
  - regression or inference-based methods such as Bayesian formula or decision tree

Data Cleaning: Noise

- **Noise:** random error or variance in a measured variable, i.e., corrupted data



Data Cleaning: Noise

- How do data get corrupted?
  - Error in measurement due to faulty or low-resolution sensors
  - Error in data recording
  - External (environmental) factors that affect the measurement process
- Are noisy data useful or should they be discarded?
- Are there any reasons to intentionally add noise to data?

Data Cleaning: Denoising

- **Binning (is used as a discretization technique.)**
  - 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
- **Low-pass filter**
  - Allow the low-frequency components of an input signal to pass through while reducing high-frequency components.
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Denoising: Binning

- **Binning:** smooth a sorted data value by consulting its "neighborhood," the values around it.

  Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

  - The data for price are first sorted and then partitioned into equal-frequency bins of size 3
  - **Smooth by bin means**: each value in a bin is replaced by the mean value of the bin
  - **Smooth by bin medians:** each bin value is replaced by the bin median
  - **Smooth by bin boundaries:** the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

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:

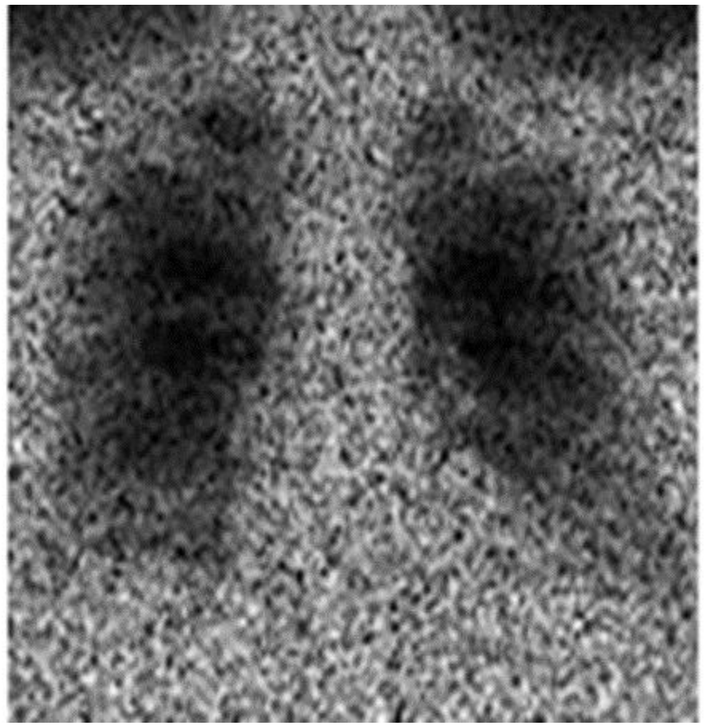
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

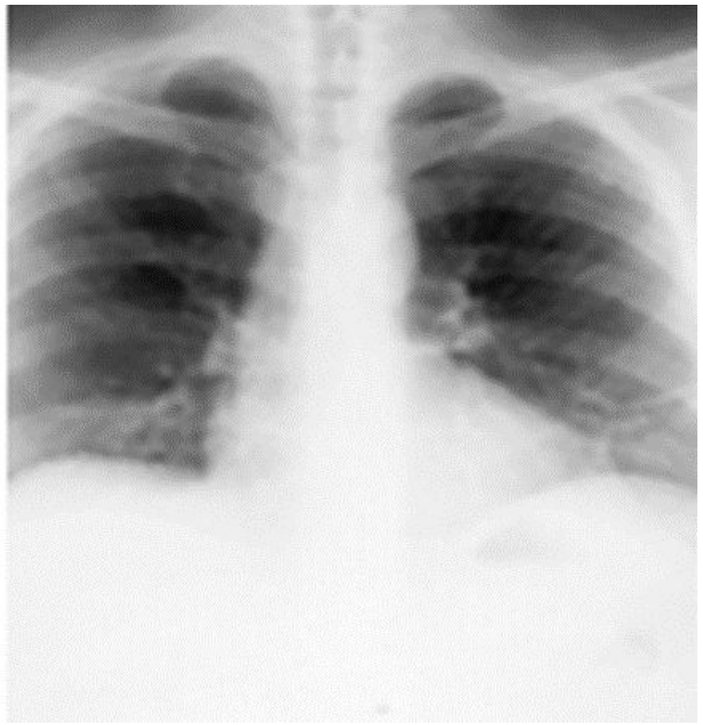
Denoising: Regression

- **Regression:** smooth by fitting the data into regression functions
  - **Linear regression** involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other.
  - **Multiple linear regression:** more than two attributes are involved, and the data are fit to a multidimensional surface

Denoising: Low-pass filter

- **Low-pass filter:** Low-pass filters allow the low-frequency components of an input signal to pass through while attenuating (reducing) high-frequency components. Measurement noise falls into the high-frequency range of the signal spectrum, while the underlying process signal usually lies towards the low-frequency end.

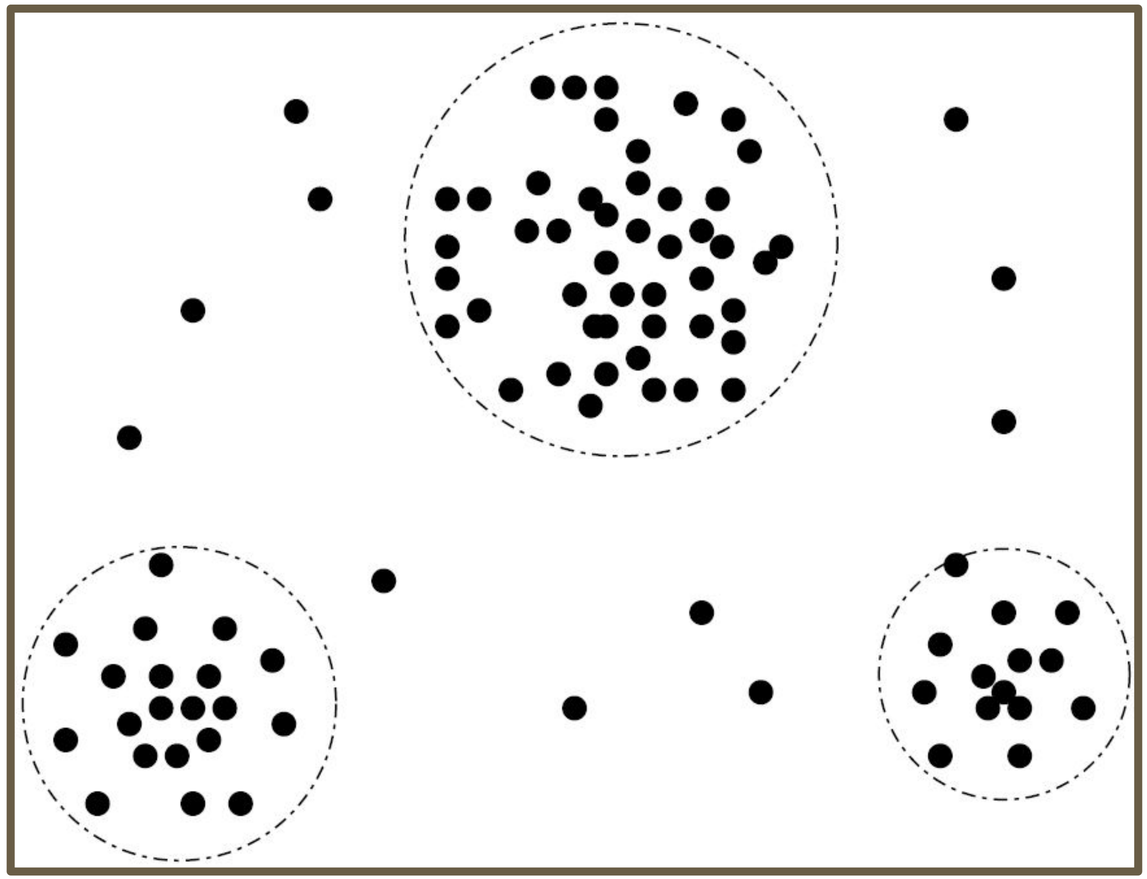




Data Cleaning: Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

- **Outlier analysis:** Outliers may be detected by clustering: similar values are organized into groups or “clusters.” Values that fall outside of the set of clusters may be considered as outliers



Outliers vs. Noise

- Difference between noise and outliers?
- Are outliers useful?

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - 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)
- 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

Outline

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration

Data integration

Data integration: Combines data from multiple data stores(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
- **Detect and resolve 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
  - ***At the tuple level:*** The same object may occur in different databases
    - Duplication detection: Careful integration can help reduce/avoid redundancies and inconsistencies in the resulting data set. This can help improve the accuracy and speed of the subsequent data mining process.
  - ***Between attributes:*** One attribute may be a "derived" attribute in another table, e.g., annual income and monthly income
    - Symbolic values: χ^2 test
    - Numeric values: correlation analysis