Unit II : Machine Learning (ML)

Dr. Mukti Padhya
Assistant Professor, NFSU

What is Data?

Data denotes the individual pieces of factual information collected from various sources. It is stored, processed and later used for analysis

Data in various forms

Performing analytics to derive insights

What is Data?

- Collection of data objects and their attributes
- An attribute (in Data Mining and Machine learning often "feature") is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - —Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - —Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Data

- There are different types of attributes
  - —Nominal
    - Examples: ID numbers, eye color, zip codes
  - —Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - —Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - —Ratio
    - Examples: temperature in Kelvin, length, time, counts

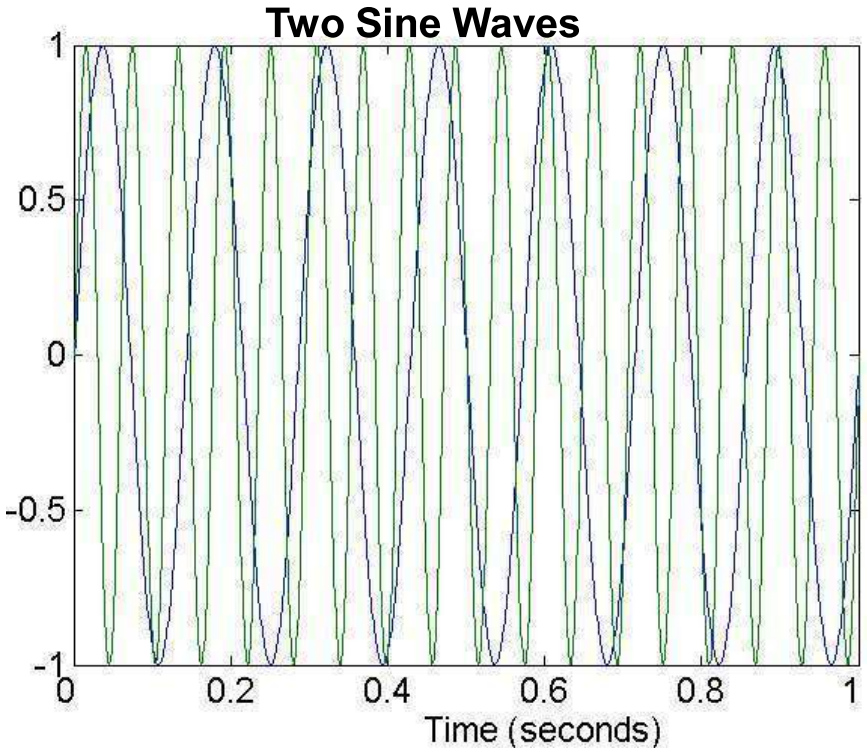
Categorical, Qualitative

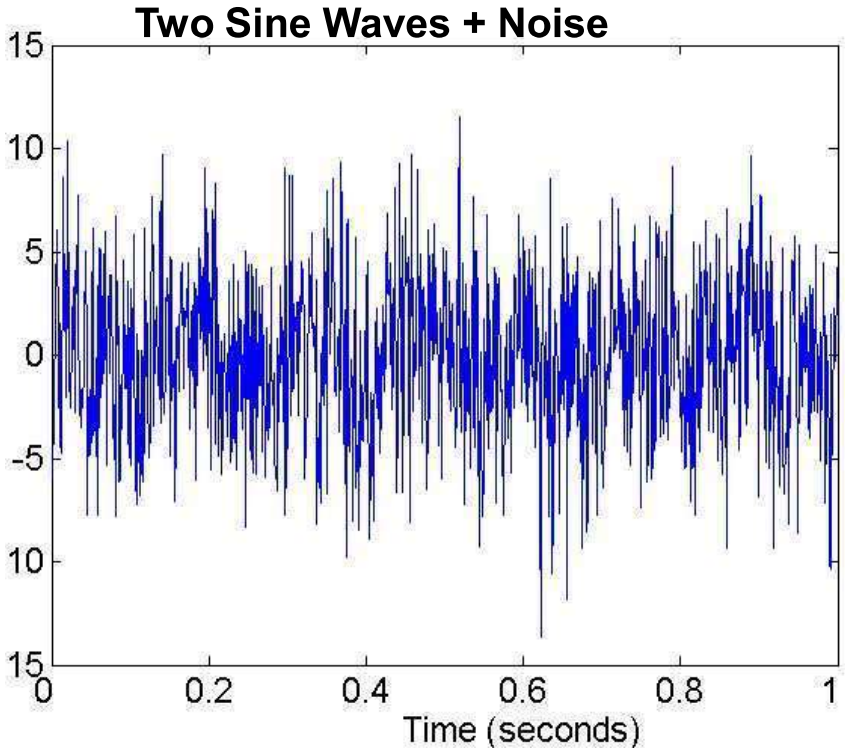
Quantitative

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone, "snow" on television screen, measurement errors.

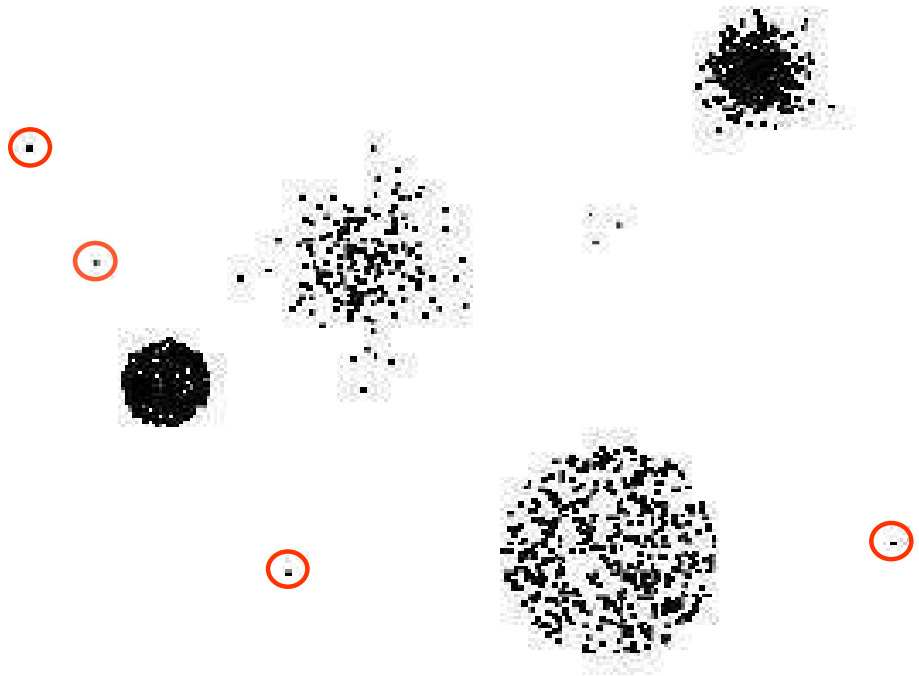




- Find less noisy data
- De-noise (signal processing)

Outliers

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



- Outlier detection + remove outliers

Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate data objects with missing value
  - Eliminate feature with missing values
  - Ignore the missing value during analysis
  - Estimate missing values = Imputation
    (e.g., replace with mean or weighted mean where all possible values are weighted by their probabilities)

Duplicate Data

- Data set may include data objects that are duplicates, or "close duplicates" of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues
  - ETL tools typically support deduplication

Topics

- Why data preprocessing?
- Data cleaning
- Data integration and transformation
- Data reduction

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Data Quality: Why Preprocess the Data?

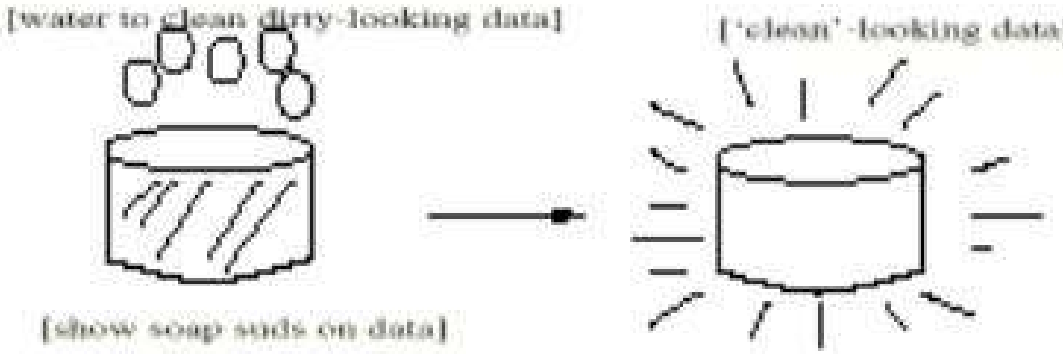
- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

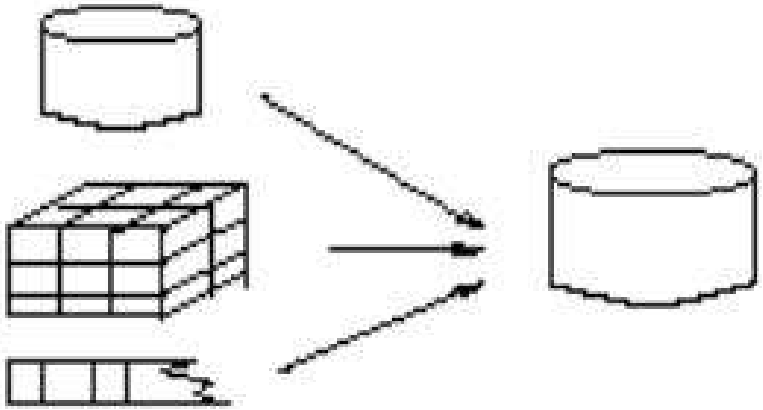
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, files, or notes
- Data transformation
  - Normalization (scaling to a specific range)
  - Aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - Data discretization: with particular importance, especially for numerical data
  - Data aggregation, dimensionality reduction, data compression,generalization

Forms of data preprocessing

Data Cleaning



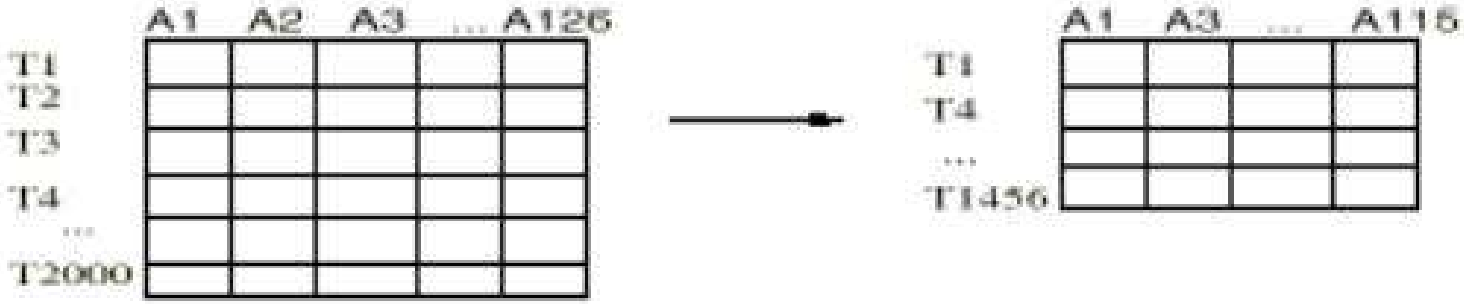
Data Integration



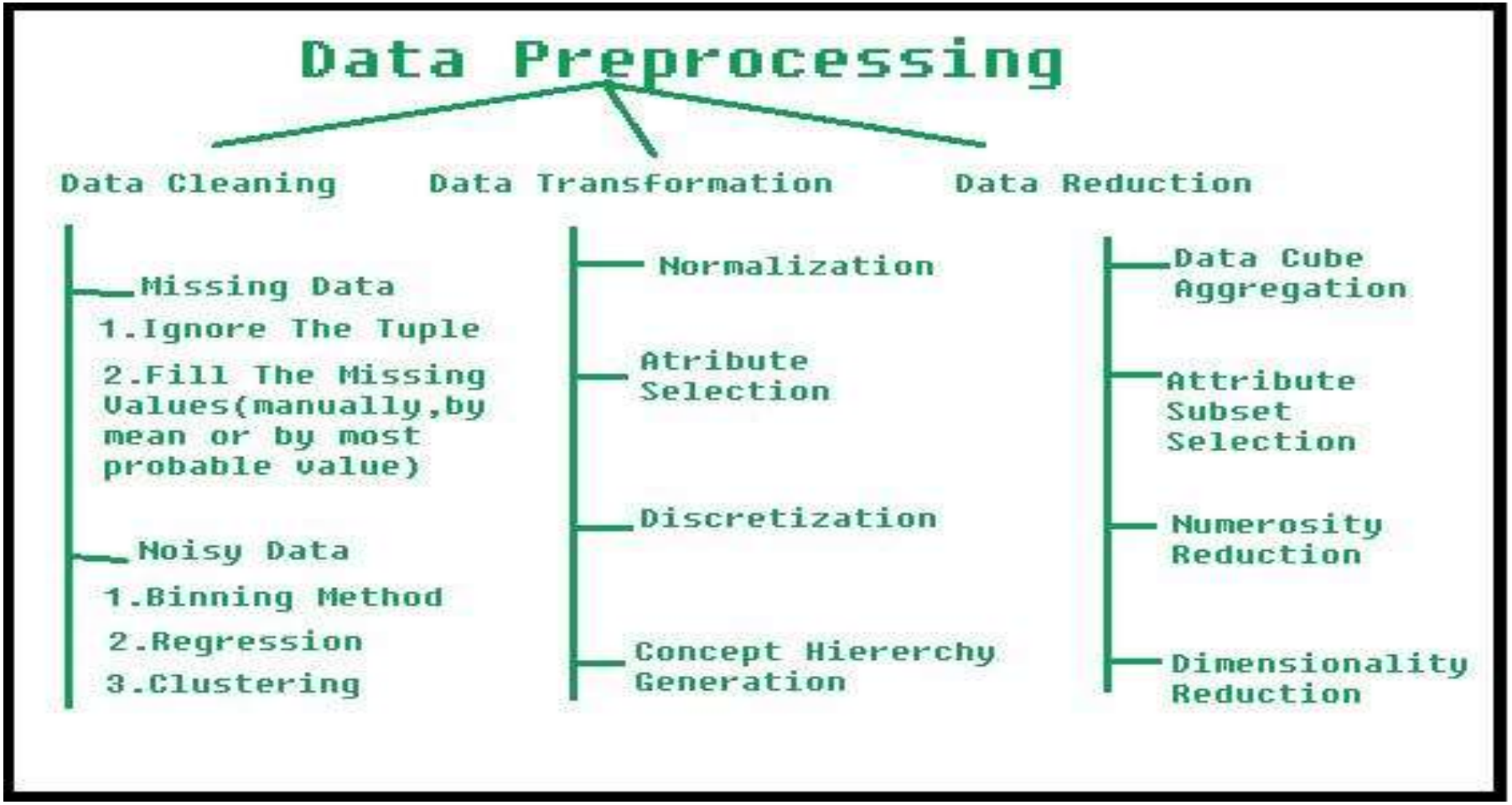
Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Forms of data preprocessing



Topics

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

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?

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 (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
- Use the attribute mean to fill in the missing value
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Noisy Data

- What is noise?
  - Random error 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 requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

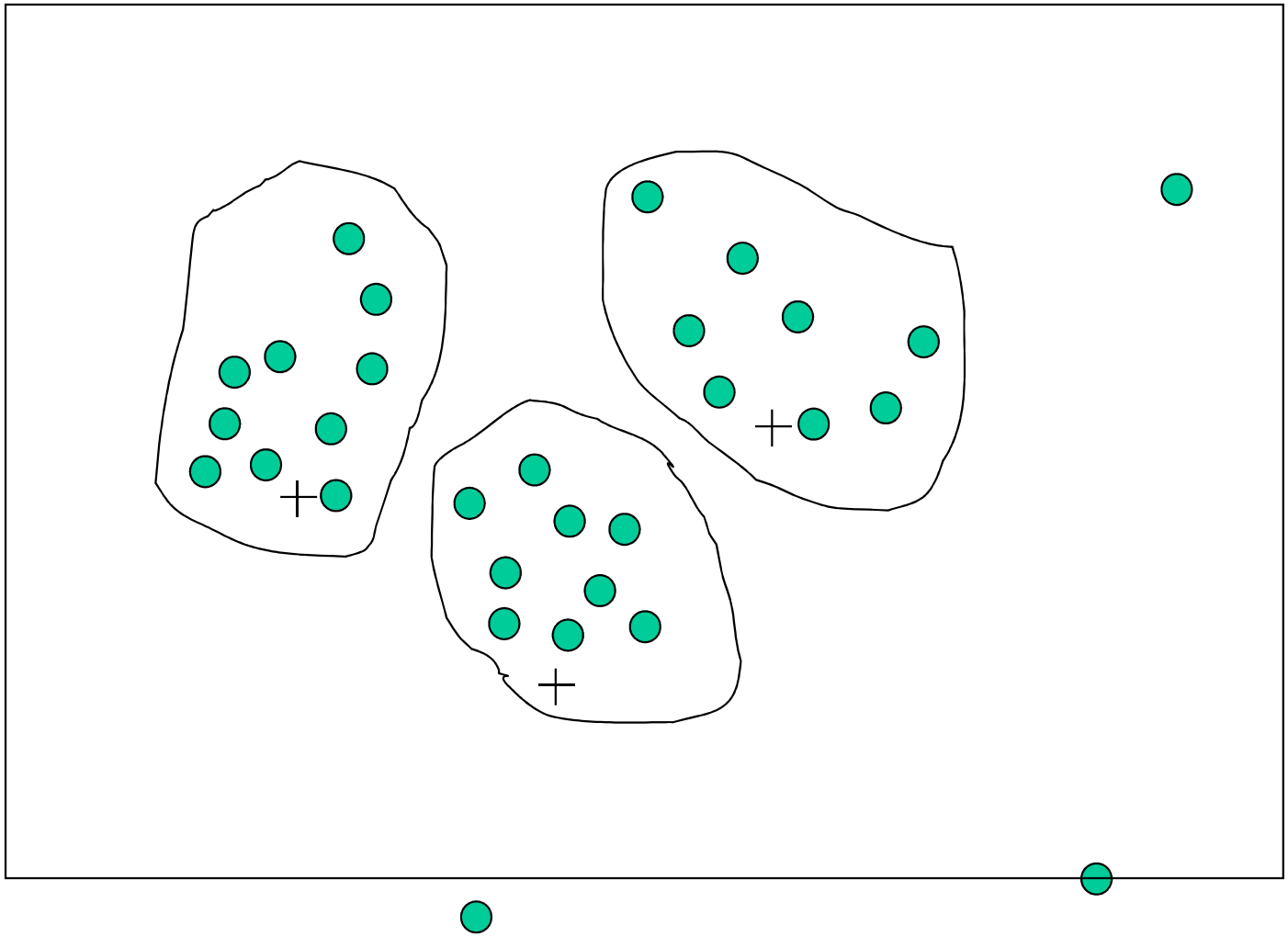
How to Handle Noisy Data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  - used also for discretization (discussed later)
- Clustering
  - detect and remove outliers
- Semi-automated method: combined computer and human inspection
  - detect suspicious values and check manually
- Regression
  - smooth by fitting the data into regression functions

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 (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

Cluster Analysis



Cluster Analysis

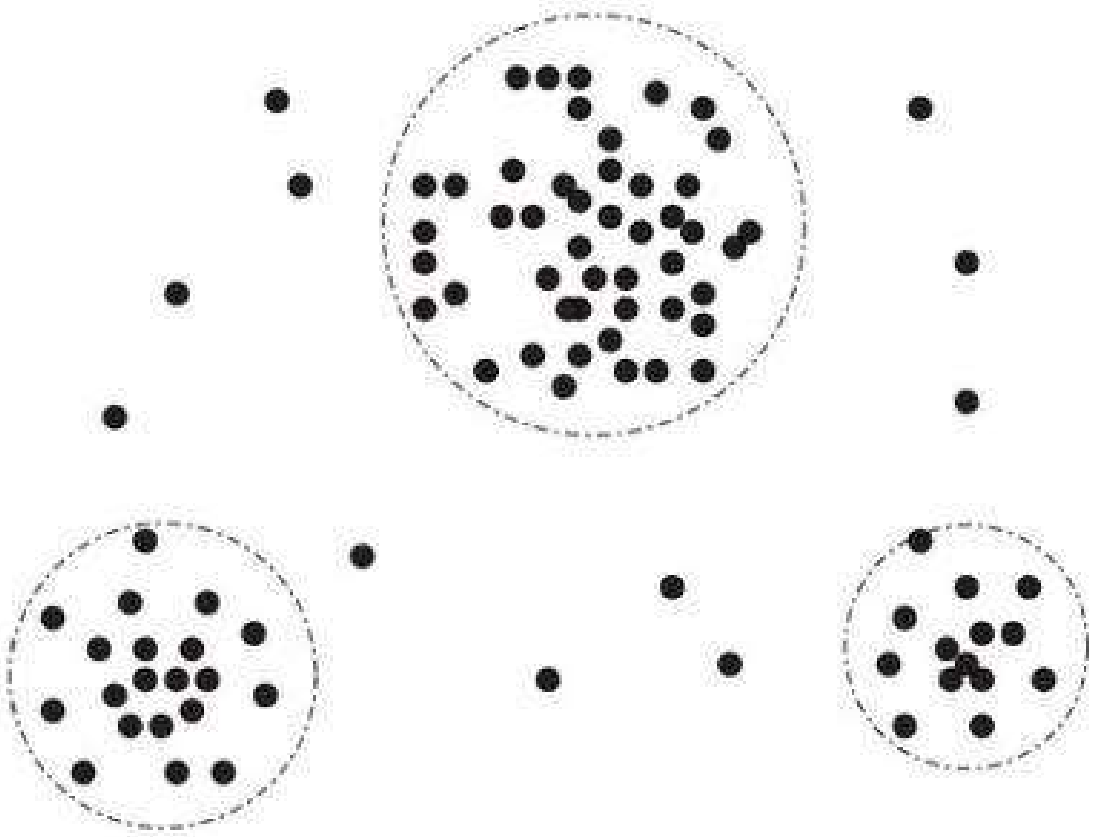
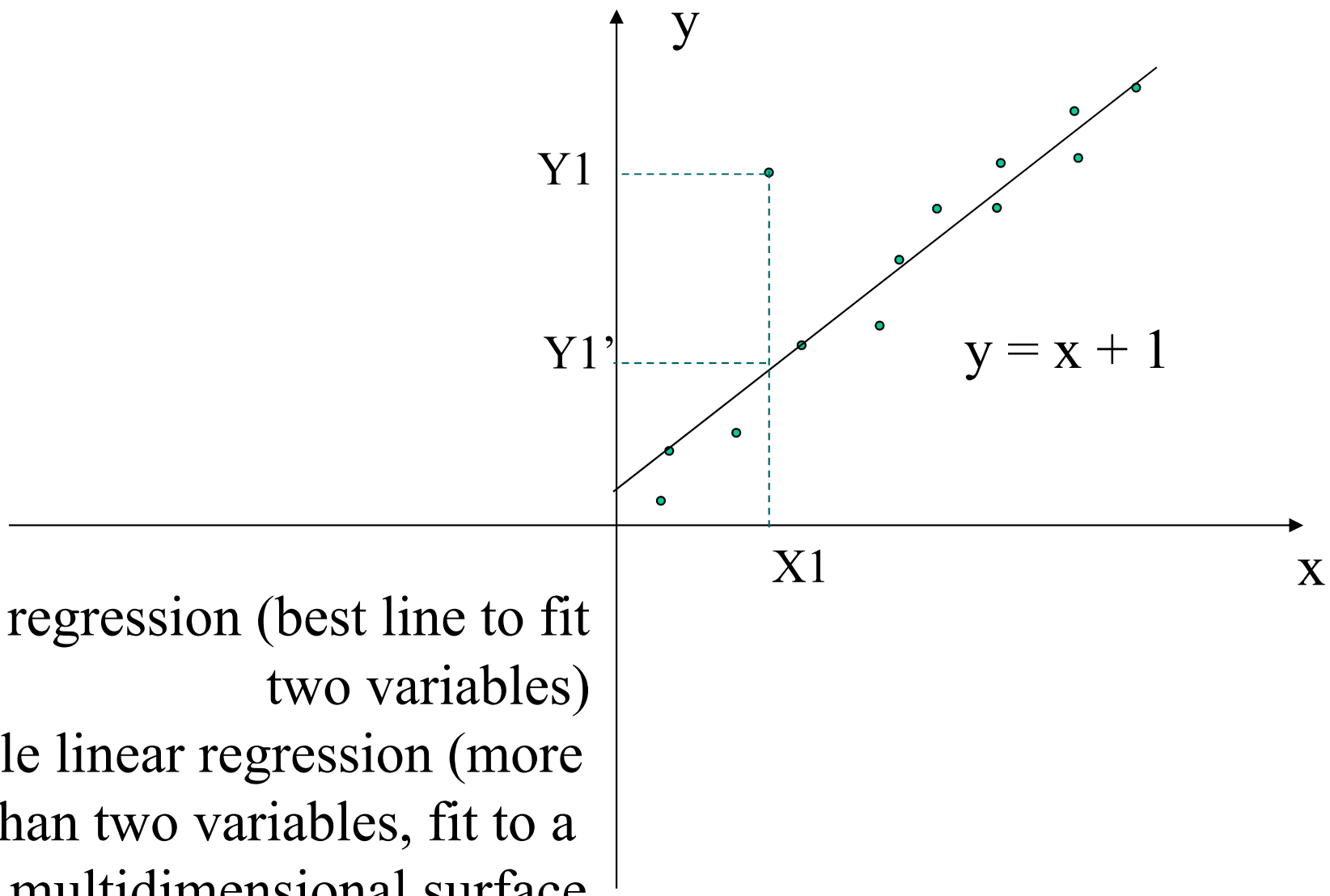


Figure A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

Regression



- Linear regression (best line to fit two variables)
- Multiple linear regression (more than two variables, fit to a multidimensional surface

How to Handle Inconsistent Data?

- Manual correction using external references
- Semi-automatic using various tools
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data

Topics

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction

Data Integration

- Data integration:
  - combines data from multiple sources into a coherent store
  - 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, different currency

Handling Redundant Data in Data Integration

- Redundant data occur often when integrating multiple DBs
  - The same attribute may have different names in different databases
  - One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlational analysis

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

- Careful integration can help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - —Simple functions: $x^k, \log(x), e^x, |x|$
  - —Standardization and Normalization
    The z-score normalizes data roughly to an interval of $[-3,3]$.

$$x' = \frac{x - \bar{x}}{s_x}$$

$\bar{x}$ ... column (attribute) mean

s_x ... column (attribute) standard deviation

Data Transformation

- Aggregation: summarization, data cube construction
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

Data Transformation: Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More "stable" data
    - Aggregated data tends to have less variability (e.g., reduce seasonality by aggregation to yearly data)

Data Transformation: Aggregation

Quarter	Sales	Quarter	Sales	Quarter	Sales	Year	Sales
Year 2010		Year 2011		Year 2012		Year Sales	
Q1	Rs.10000	Q1	Rs.8000	Q1	Rs.15000	2010	Rs.1,30,000
Q2	Rs.50000	Q2	Rs.15000	Q2	Rs.20000	2011	Rs.53000
Q3	Rs.40000	Q3	Rs.10000	Q3	Rs.40000	2012	Rs.1,05,000
Q4	Rs.30000	Q4	Rs.20000	Q4	Rs.30000		

Sales per quarter from year 2010 to 2012 get aggregated into a single annual sales record.

Data Transformation: Conversion

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Topics

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction

Data Reduction

- Problem:

  Data Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

- Solution?
  - Data reduction…

Data Reduction

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

- Data reduction strategies
  - Dimensionality reduction
  - Sampling
  - Data compression
  - Feature Subset Selection
  - Feature Creation

Sampling

Sampling

- Sampling is the main technique employed for data selection.
  - —It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because **obtaining** the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because **processing** the entire set of data of interest is too expensive (e.g., does not fit into memory or is too slow).

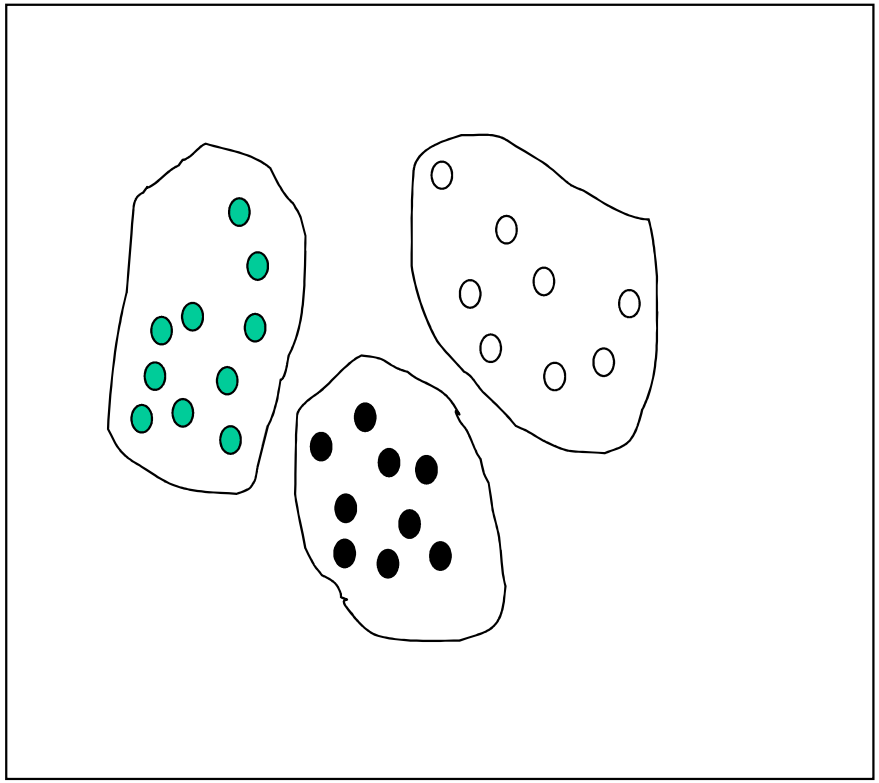
Sampling

Sampling ...

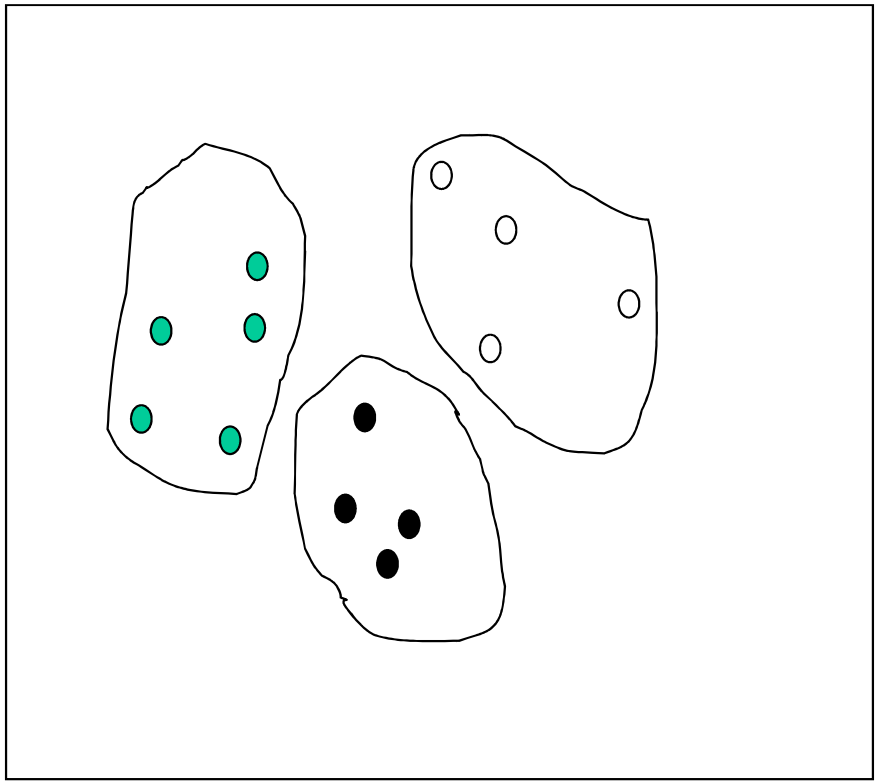
- The key principle for effective sampling is the following:
  - —using a sample will work almost as well as using the entire data sets, if the sample is **representative**.
  - —A sample is representative if it has approximately the same property (of interest) as the original set of data.

Sampling

Raw Data

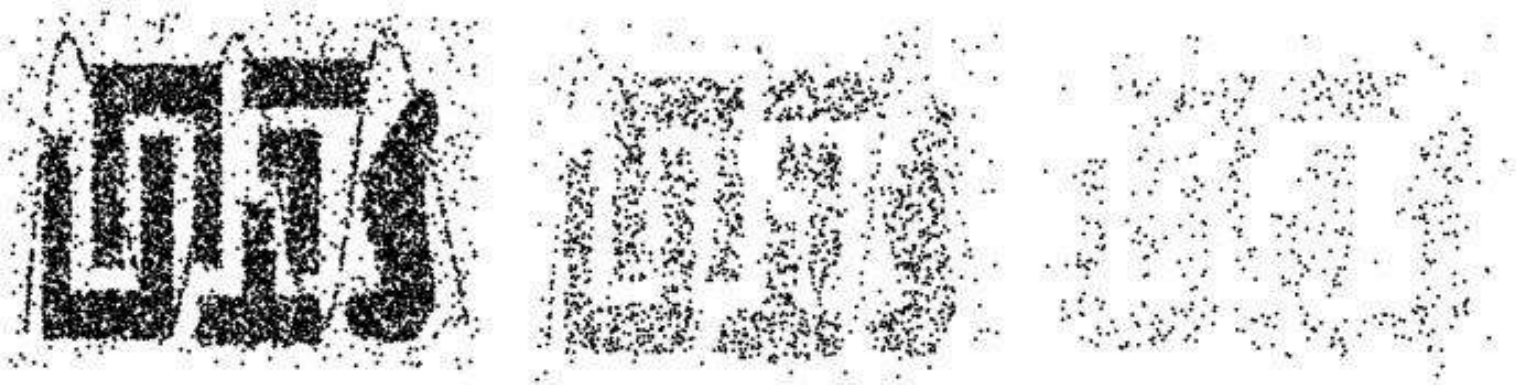


Cluster/Stratified Sample



Sampling

Sample Size



8000 points

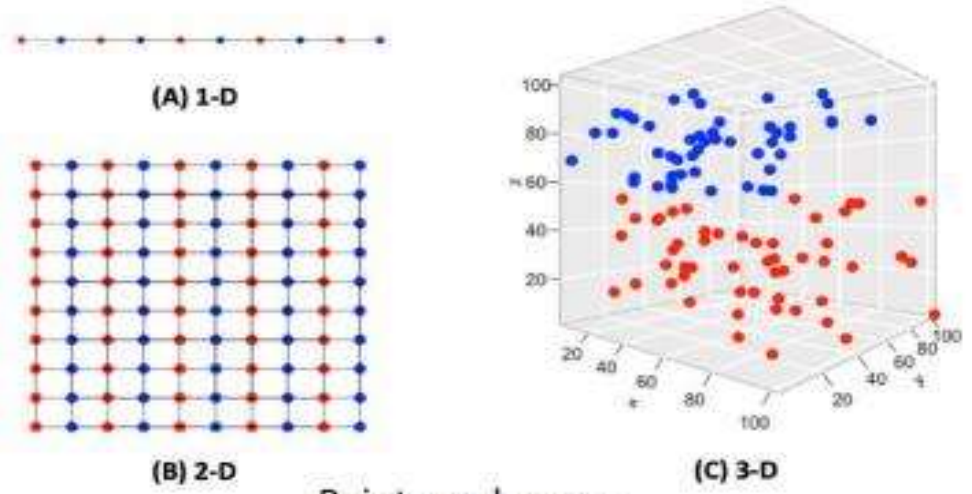
2000 Points

500 Points

Dimensionality Reduction

Curse of Dimensionality

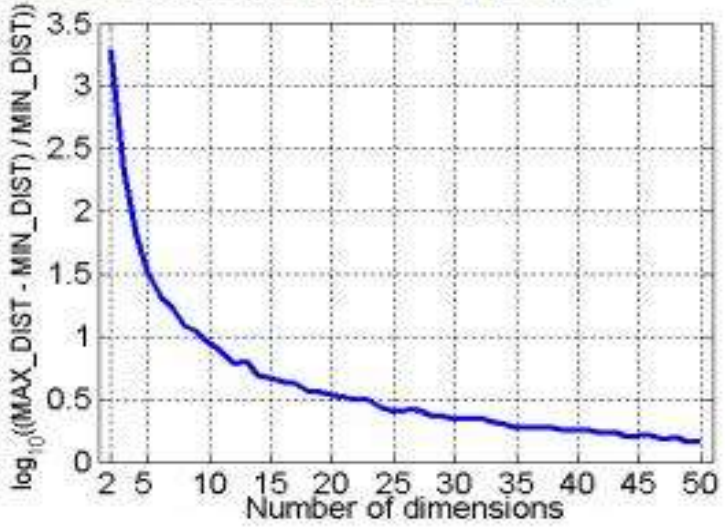
- When dimensionality increases, the size of the data space grows exponentially.



Points and space

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful
  - Density → 0
  - All points tend to have the same Euclidean distance to each other.

Experiment: Randomly generate 500 points. Compute difference between max and min distance between any pair of points



Dimensionality Reduction

- Purpose:
  - —Avoid curse of dimensionality
  - —Reduce amount of time and memory required by data mining algorithms
  - —Allow data to be more easily visualized
  - —May help to eliminate irrelevant features or reduce noise

- Techniques
  - —Principle Component Analysis
  - —Singular Value Decomposition
  - —Others: supervised and non-linear techniques

Dimensionality Reduction

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features
  - Nice side-effect: reduces # of attributes in the discovered patterns (which are now easier to understand)

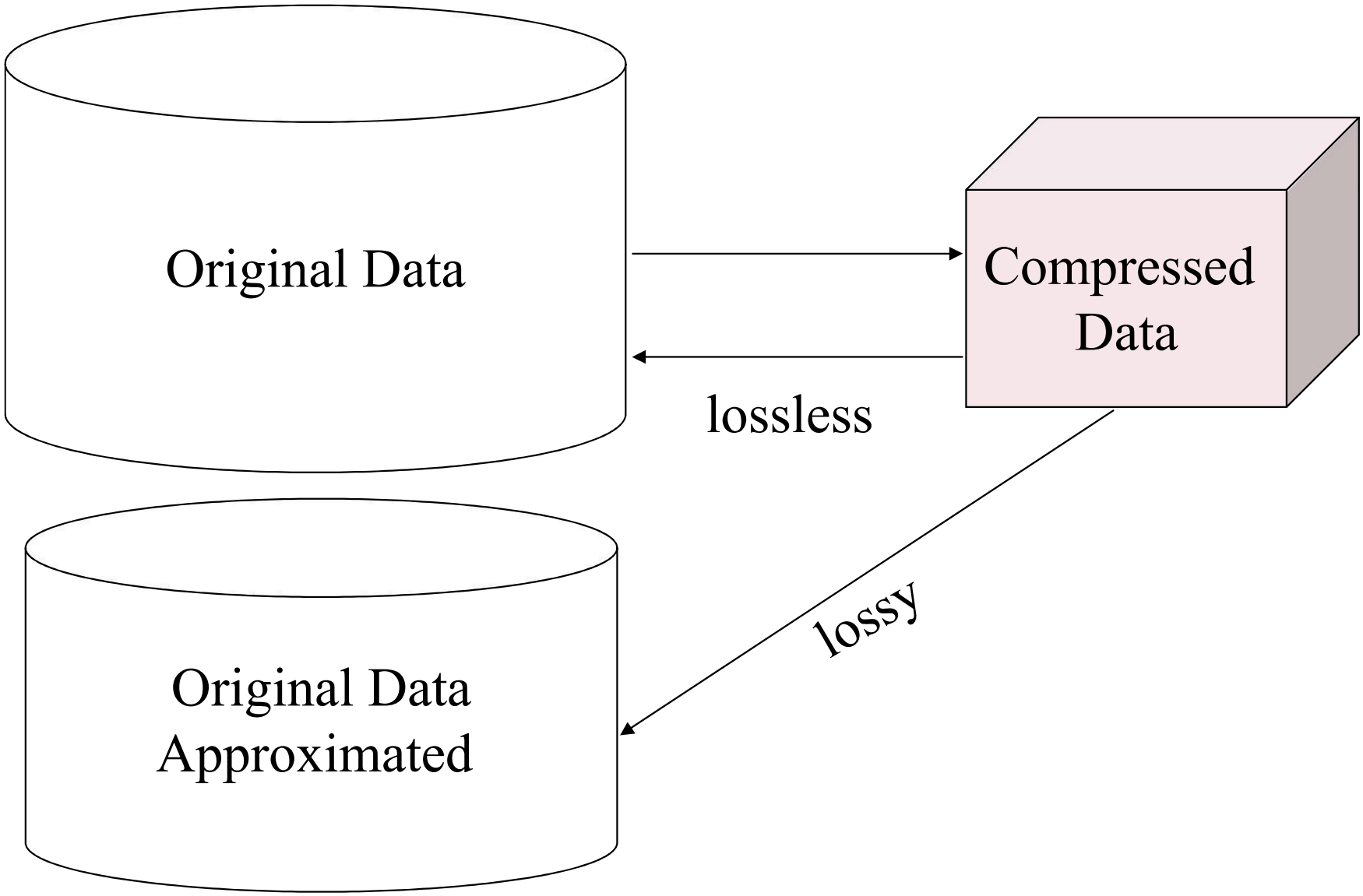
- Redundant features
  - —duplicate much or all of the information contained in one or more other attributes (are correlated)
  - —Example: purchase price of a product and the amount of sales tax paid

- Irrelevant features
  - —contain no information that is useful for the data mining task at hand
  - —Example: students' ID is often irrelevant to the task of predicting students' GPA

Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video, image 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

Data Compression



Wavelet Transforms

- Discrete wavelet transform (DWT): linear signal processing
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients

- Fourier transform
- Wavelet transform