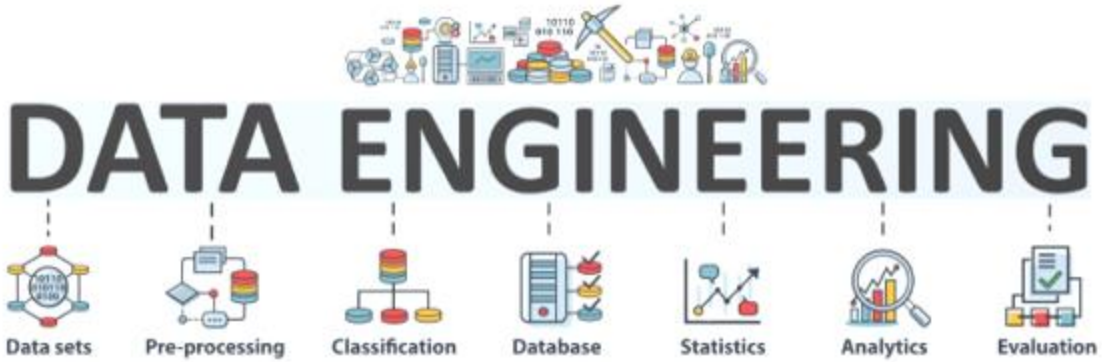


香港城市大學
City University of Hong Kong

Lecture 3: Data Preprocessing

CS5481 Data Engineering

Instructor: Linqi Song

Outline

1. Why data preprocessing?
2. Data Cleaning
3. Data Integration
4. Data Transformation
5. Data Reduction
6. Data Discretization

Why data preprocessing - dirty data

- 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?

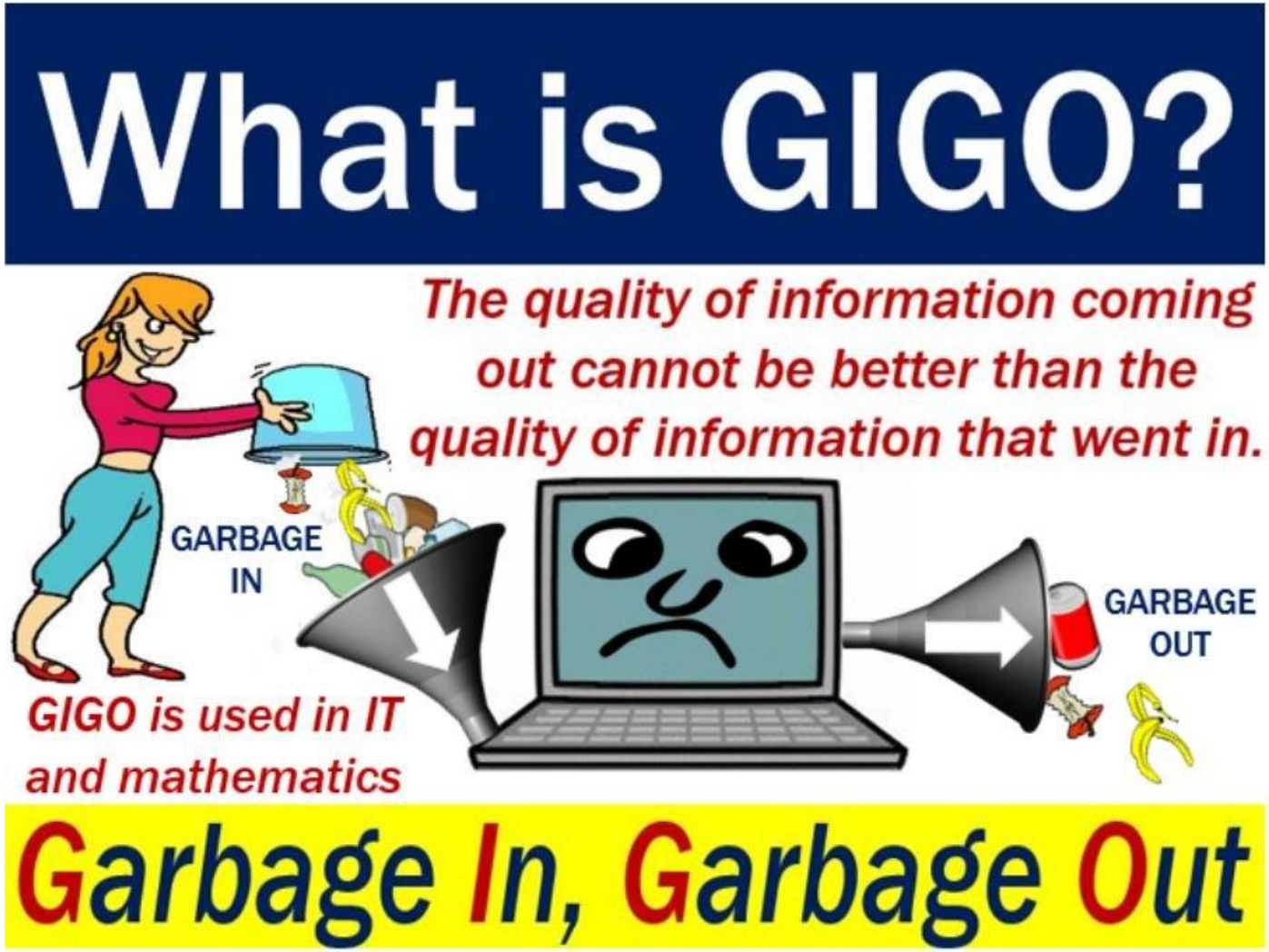
Dirty data example

- **All of these are commonly seen in the real-world (an example of IMDb movie rating data)**
  - Zeros or nulls replace missing values
  - Spelling inconsistency (usa/USA), errors and noises (meaningless symbols)
  - Rows are duplicated
  - Others: inconsistent date formats (e.g. 10/9/15 vs. 9/10/15), units not specified (kg or pound)

color	director_name	duration	gross	movie_title	anguage	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall StreetÂ	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3Â	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful EightÂ	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	MargaretÂ	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of SmaugÂ	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of JusticeÂ	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected JourneyÂ	English	USA	180000000	2012	7.9
Color	Edward Hall	180		RestlessÂ	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud AtlasÂ	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon TattooÂ	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of GodÂ	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Battle of the Five ArmiesÂ	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6

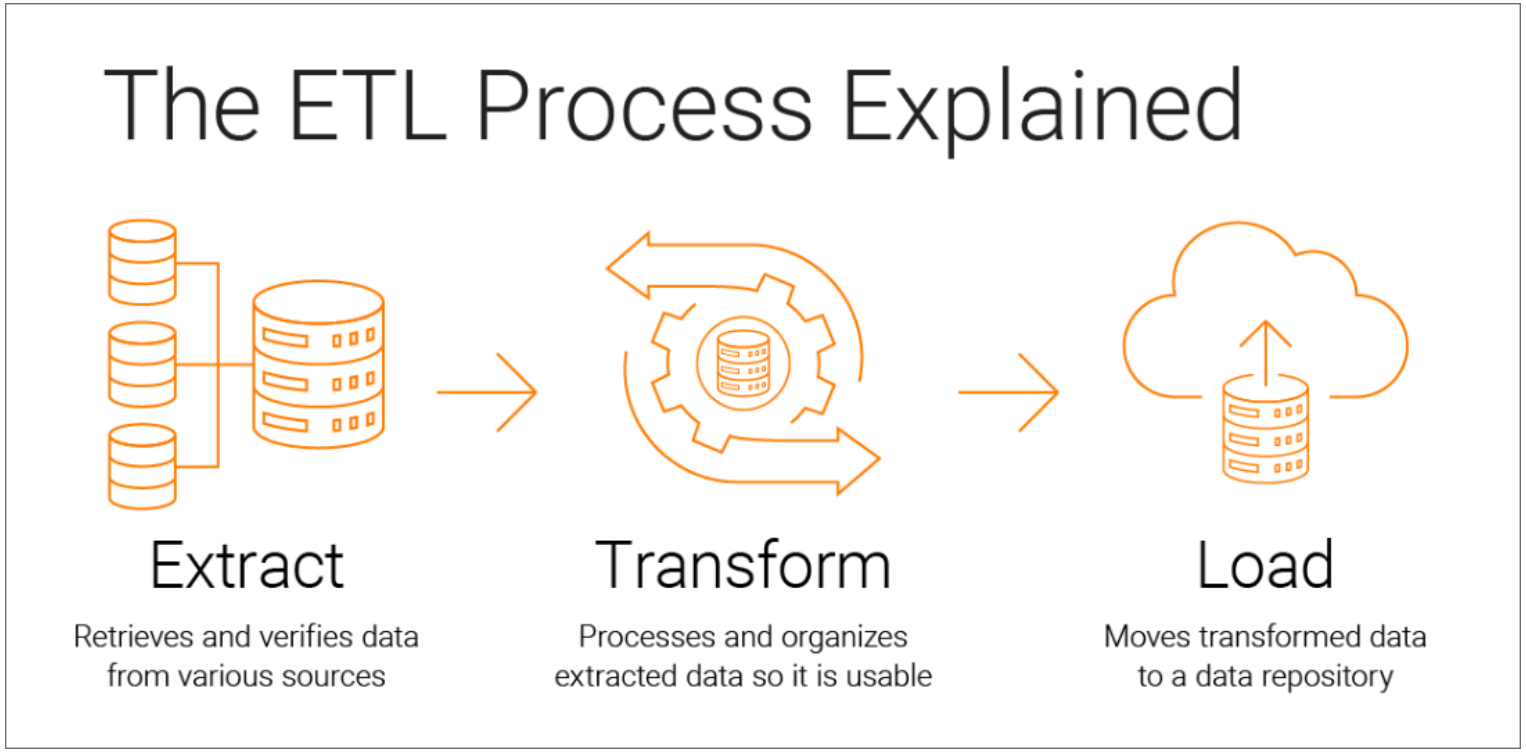
Why data preprocessing - GIGO

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



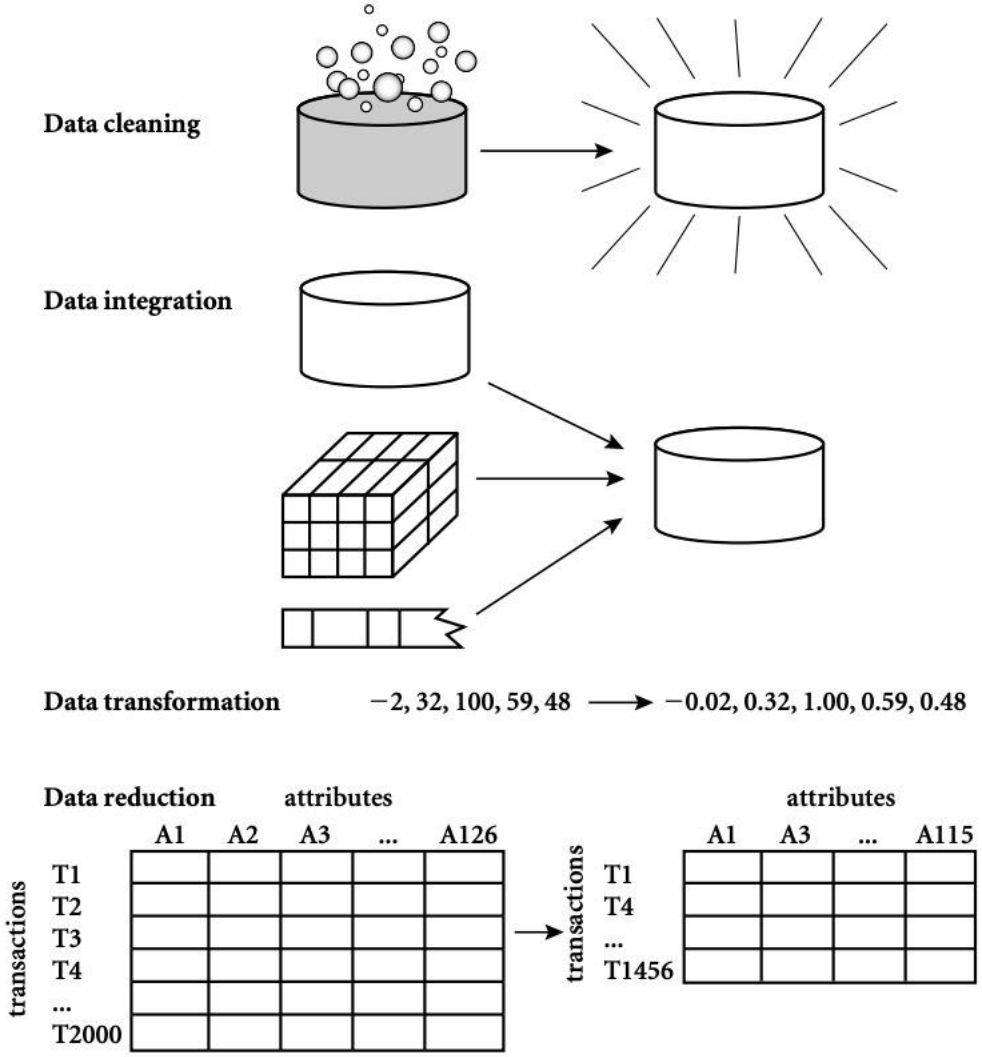
ETL process: Extract, Transform, and Load

- **It's a three-step data integration process used by organizations to combine and synthesize raw data from multiple data sources into a data warehouse, data lake, data store, relational database or any other application.**



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 cleaning

What is data cleaning?

The process of transforming raw data to facilitate subsequent analysis.

Major tasks

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

Example: filling in missing values in blood test.

Data cleaning – missing data

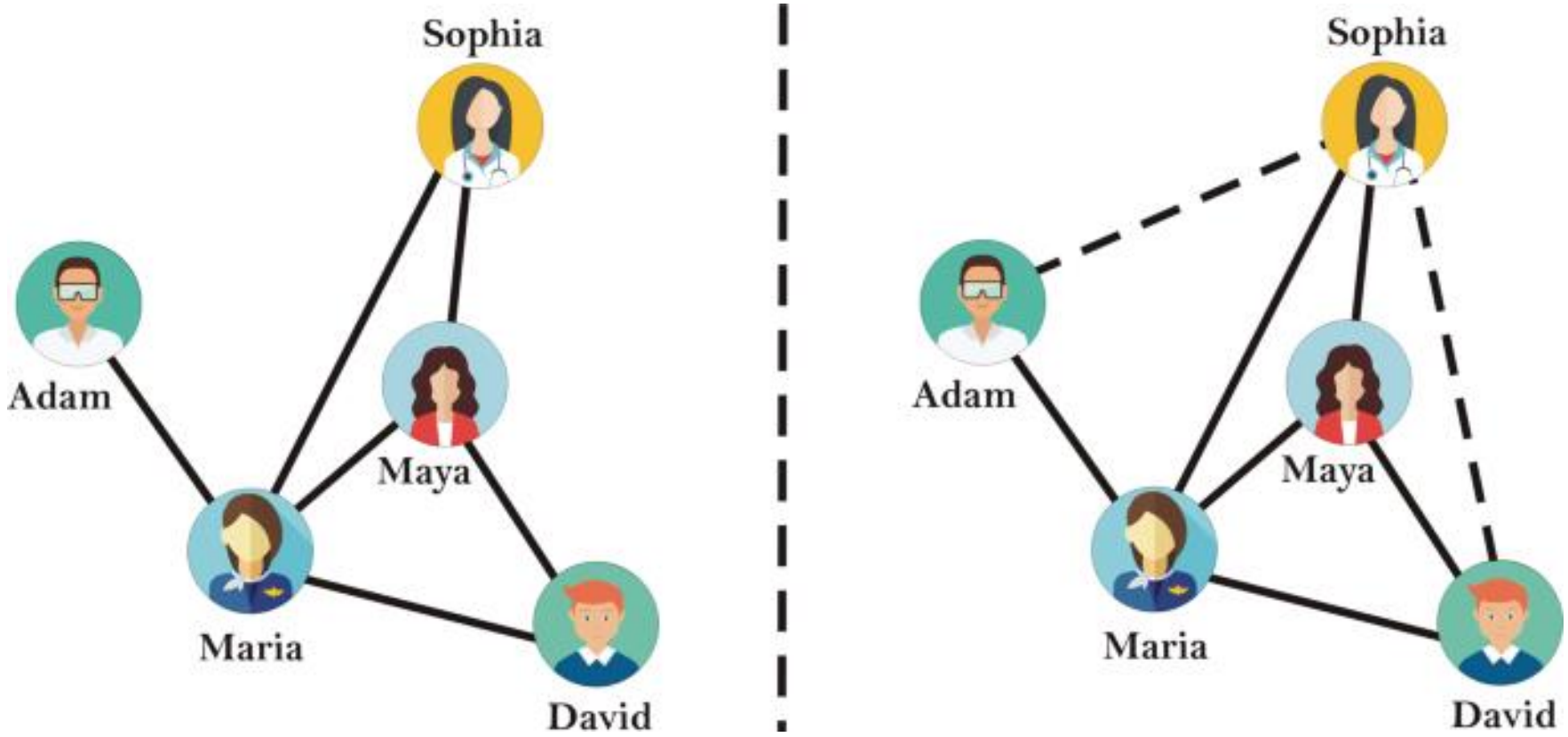
- **Data is not always available**
  - Many tuples have no recorded value for several attributes, such as customer income in sales data
- **Missing data may be due to**
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - 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**

Data cleaning – handling missing data

1. **Ignore** the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
2. Fill in the missing value **manually**: tedious + infeasible
3. Automatically:
    1) Use a **global constant** to fill in the missing value: e.g., “unknown”, a new class
    2) Use the **attribute mean** to fill in the missing value
    3) Use the **attribute mean for all samples of the same class** to fill in the missing value: smarter
    4) Use the **most probable value** to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Data cleaning – example of handling missing data

Example: Link Prediction



Data cleaning – noisy data

- **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

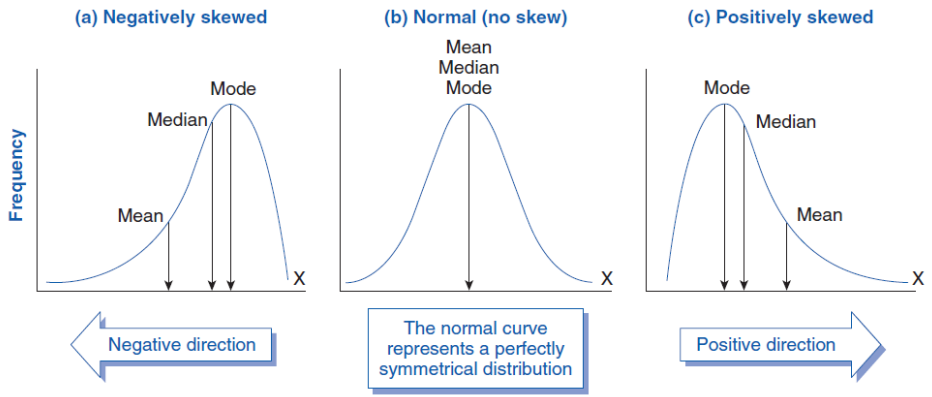
  Duplicate records, incomplete data, inconsistent data

Data cleaning – handling noisy data

- **Binning method**
  - First sort data and partition into bins
  - Then one can **smooth by bin means, median, and boundaries**, etc.
  - Used also for discretization
- **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

Data cleaning – equal-width data binning

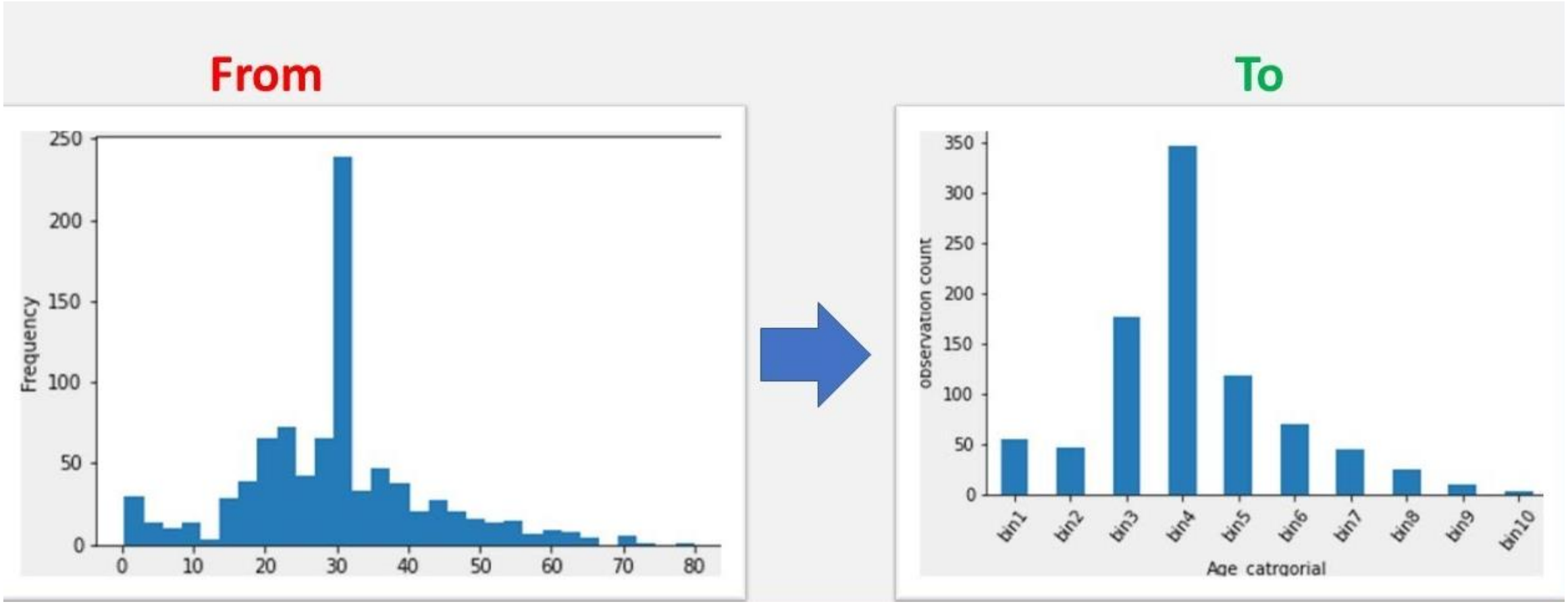
- **Equal-width (distance) partitioning**
  - ➢It divides the range into N intervals of equal size: uniform grid
  - ➢if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B-A)/N.
  - ➢The most straightforward
  - ➢But outliers may dominate presentation
  - ➢Skewed data may not be handled well
    - ➢Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.



Data cleaning – equal-depth data binning

- **Equal-depth (frequency) partitioning**
  - ➢ It divides the range into N intervals, each containing approximately same number of samples
  - ➢ Good data scaling
  - ➢ Managing categorical attributes can be tricky.

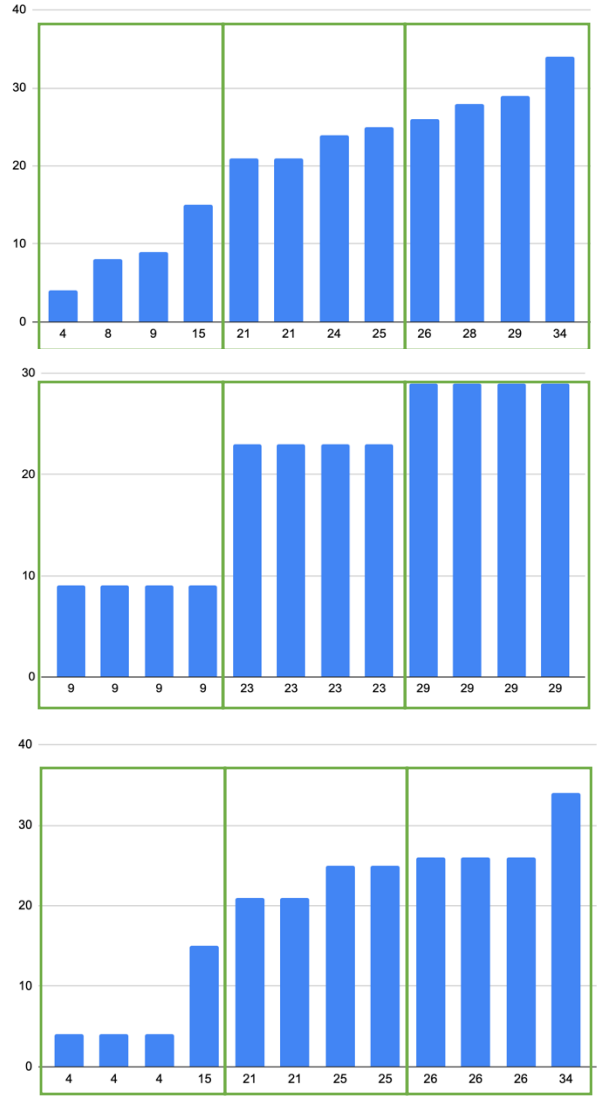
Data cleaning – examples of data binning



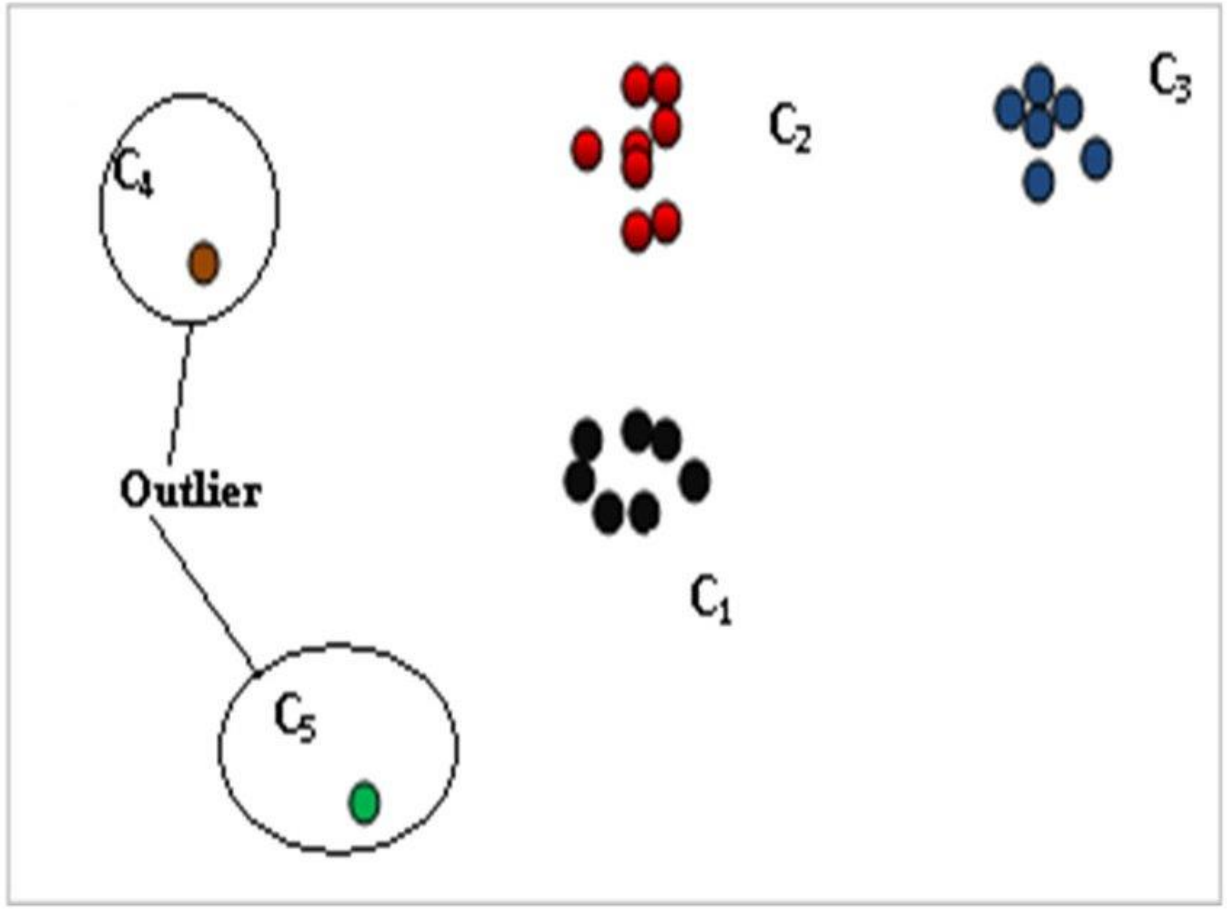
Examples of data binning: number of COVID-19 vaccine doses administered per 100 people

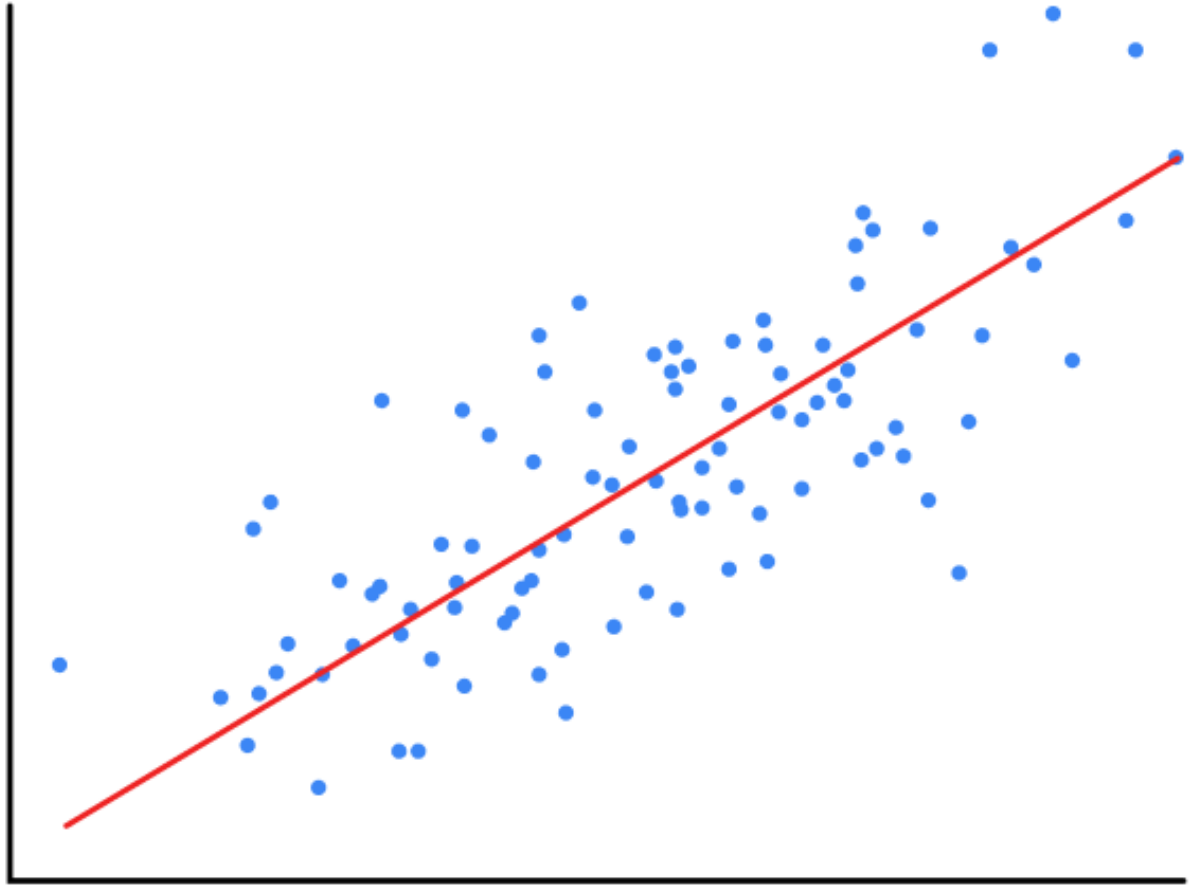
Data cleaning – binning methods for data smoothing

- 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



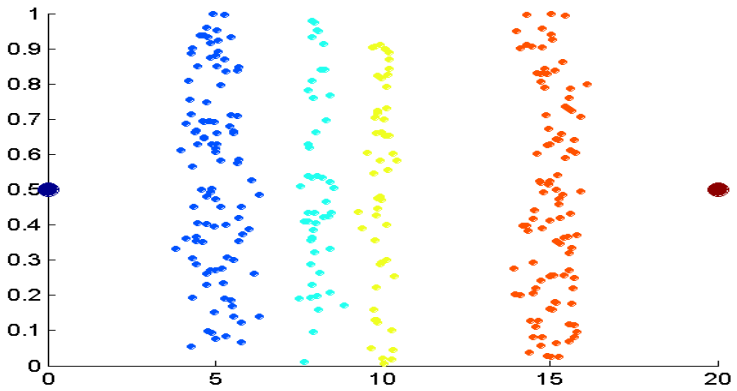
Data cleaning – clustering & regression



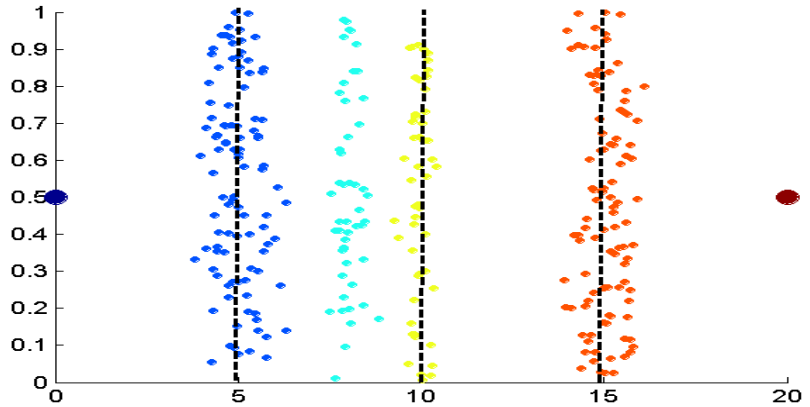


- Linear regression (best line to fit two variables)
- Multiple linear regression (more than two variables, fit to a multidimensional surface

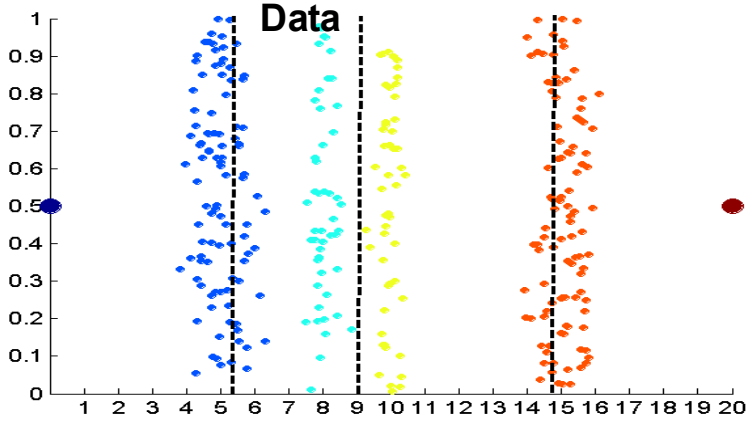
Binning vs. clustering



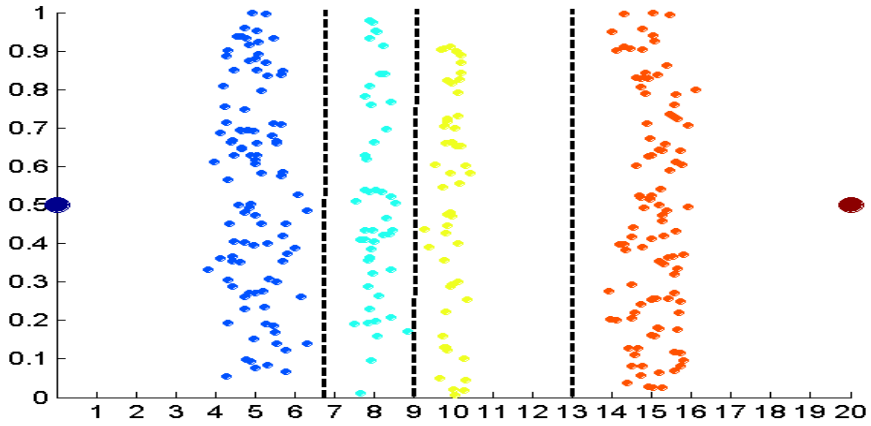
Original data



Equal-width binning



Equal depth (frequency) binning



K-means clustering leads to better results

Data cleaning – handling inconsistent data

- Manual correction using external references
  - Use metadata (metadata is the data about data: e.g., a digital image may include metadata that describes the size of the image, its color depth, resolution, when it was created, the shutter speed, and other data.), such as domain, range, dependency, distribution to check the data inconsistency
  - Check field overloading (e.g., using an unused bit of an attribute whose value range uses only, say, 31 out of 32 bits)
  - Check uniqueness rule, consecutive rule and null rule
- Semi-automatic using various tools
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data
  - 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 integration

- **Data integration:** combines data from multiple sources into a coherent store
- **Schema integration**
  - Integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources
- **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

Data integration – handling redundant data

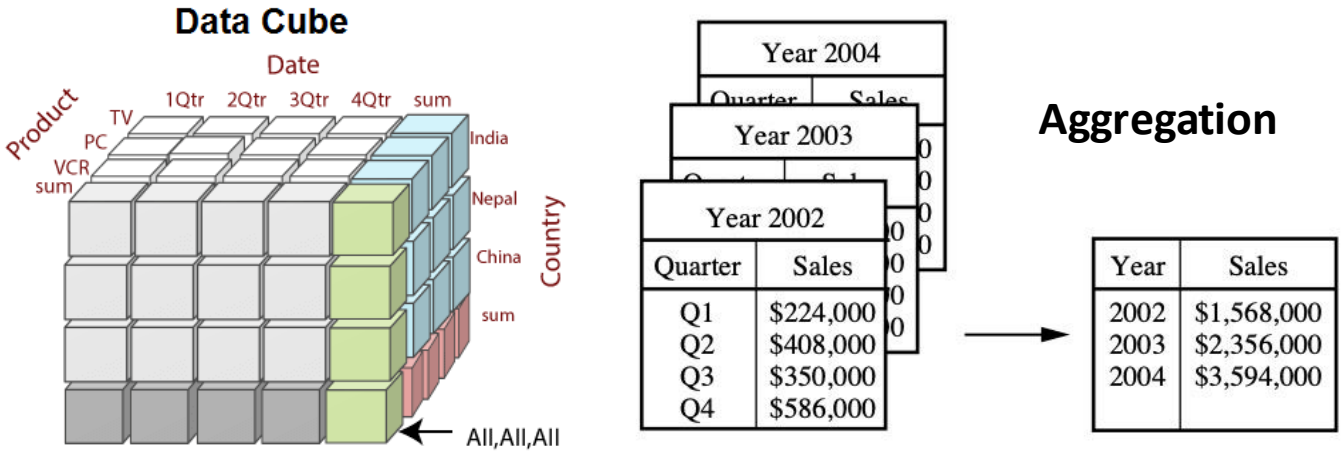
- Redundant data occur often when integrating multiple DBs
  - The same attribute may have different names in different databases
  - Dependency: One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlational analysis

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

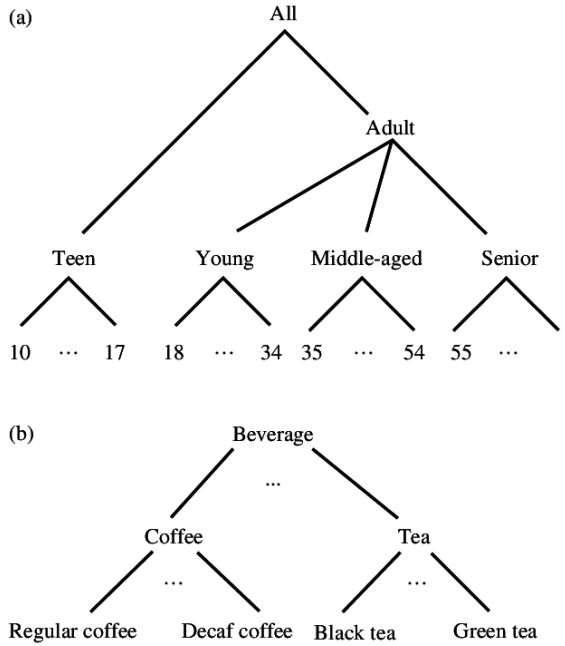
- **Careful integration** can help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data transformation (1)

- **Smoothing:** remove noise from data (binning, clustering, regression)
- **Aggregation:** summarization, data cube construction (the quarterly sales data may be aggregated so as to compute annual total amounts.)
- **Generalization:** concept hierarchy generation



A data cube is **a data structure optimized for fast and efficient analysis**. It enables consolidating or aggregating relevant data into the cube and then drilling down, slicing and dicing, or pivoting data to view it from different angles.



Concept hierarchy

Data transformation (2)

- **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

Data transformation – normalization

- min-max normalization

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

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where *j* is the smallest integer such that Max(| v'|)<1

A useful tool: Regular expressions (Regex)

- **What is Regex**
  - A regular expression is a sequence of characters that specifies a search pattern in text.
  - It can be used when
    - testing the pattern within the string
    - identify specific text in a document, delete the text completely, or replace it with other text
    - extract substrings from strings based on pattern matching
- **Regex examples**

- `.at` matches any three-character string ending with "at", including "hat", "cat", "bat", "4at", "#at" and " at" (starting with a space).
- `[hc]at` matches "hat" and "cat".
- `[^b]at` matches all strings matched by `.at` except "bat".
- `[^hc]at` matches all strings matched by `.at` other than "hat" and "cat".
- `^[hc]at` matches "hat" and "cat", but only at the beginning of the string or line.
- `[hc]at$` matches "hat" and "cat", but only at the end of the string or line.
- `\[.\]` matches any single character surrounded by "[" and "]" since the brackets are escaped, for example: "[a]", "[b]", "[7]", "[@]", "[]]", and "[ ]" (bracket space bracket).
- `s.*` matches s followed by zero or more characters, for example: "s", "saw", "seed", "s3w96.7", and "s6#h%(>>>m n mQ".

Refer to https://en.wikibooks.org/wiki/Regular_Expressions

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 – 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

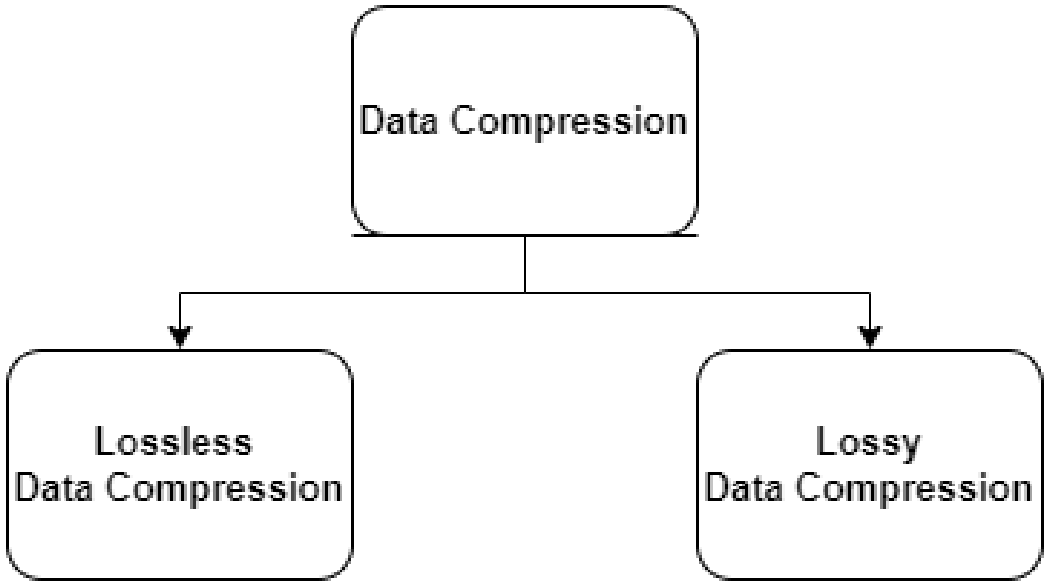
- Data cube aggregation
- Dimensionality reduction
- Data compression
- Numerosity reduction
- Discretization and concept hierarchy generation

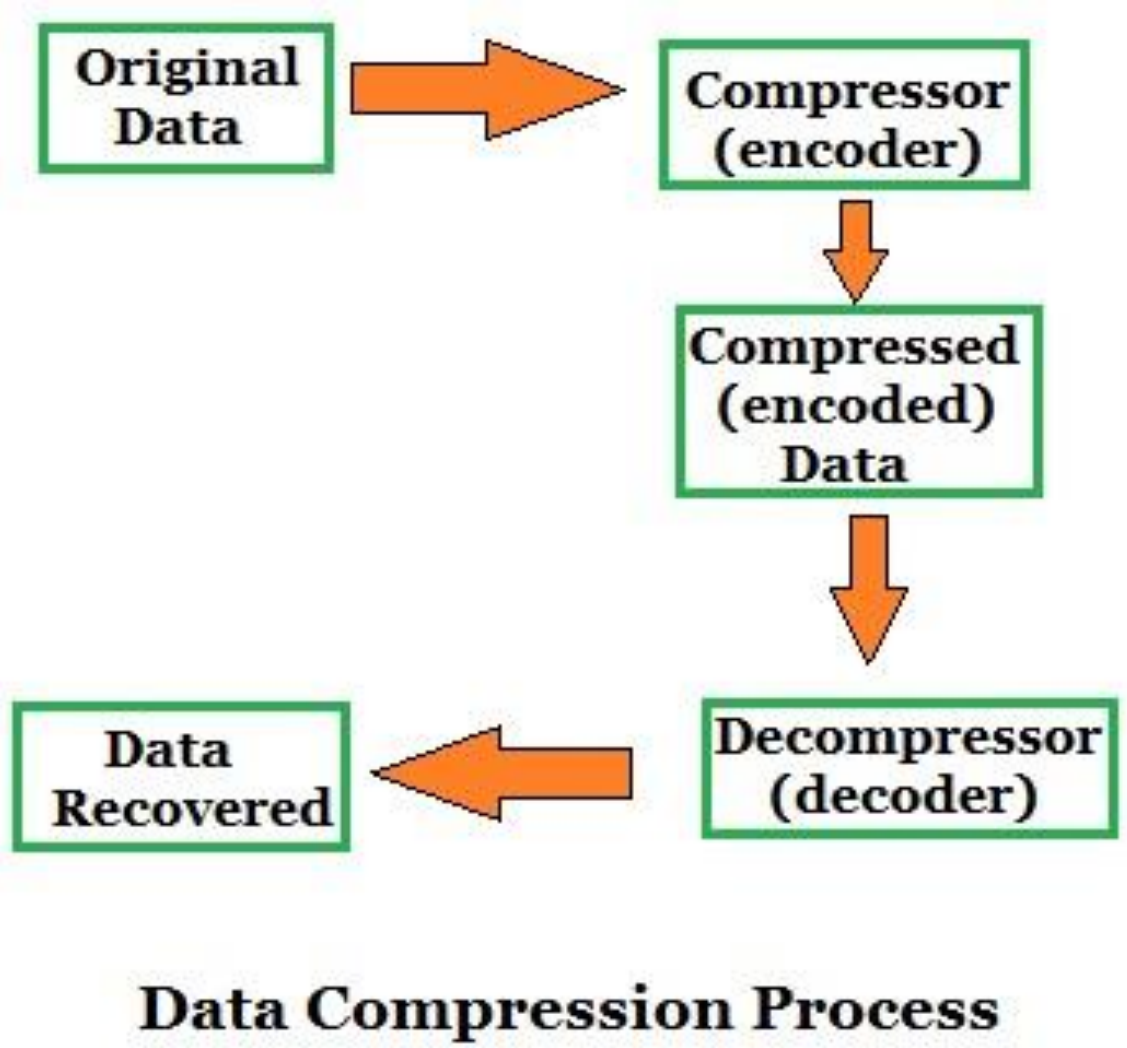
Data reduction – dimensionality reduction

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.
- Nice side-effect: reduces # of attributes in the discovered patterns, easier to understand.

Data reduction – data compression





Data Compression Process

Data reduction – data compression

- **String compression**
  - There are extensive theories and well-tuned algorithms (Shannon source coding theory)
  - 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 sequences (e.g., stock prices over time)**
  - Typically short and vary slowly with time

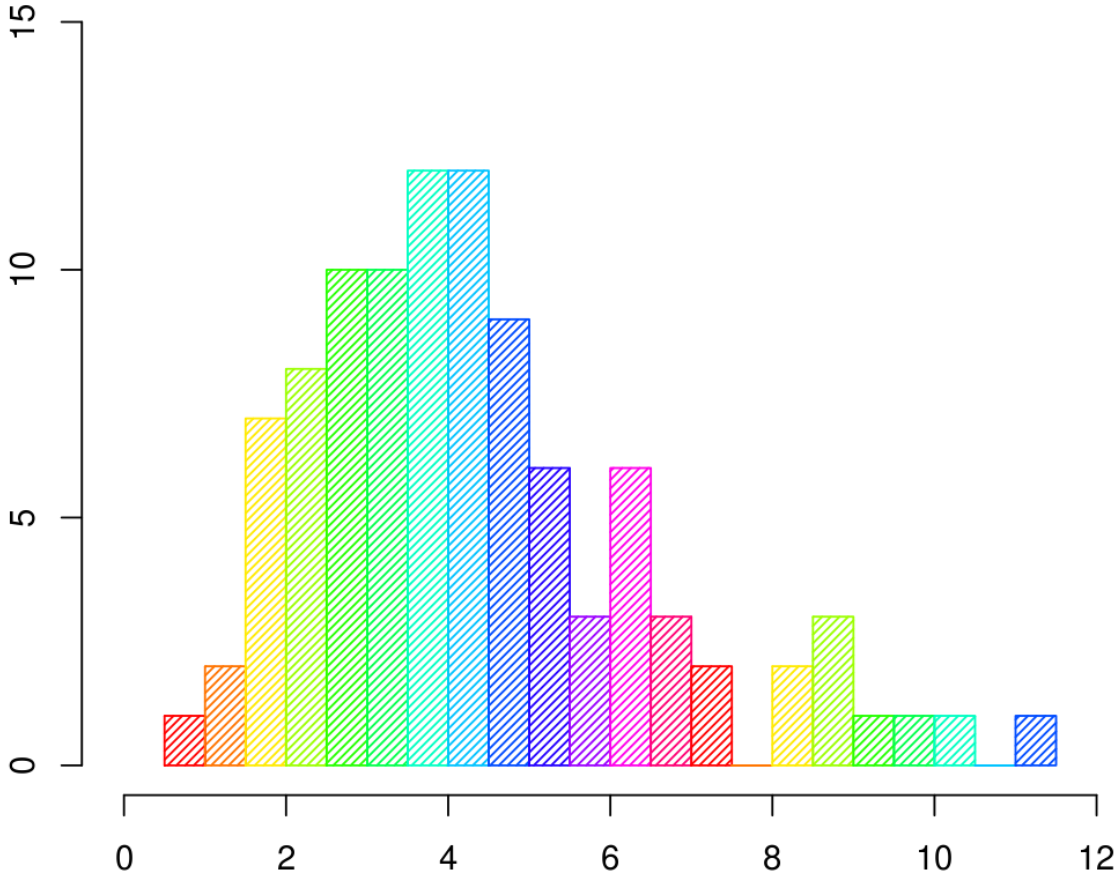
Data reduction – numerosity reduction (1)

- **Parametric methods**
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - E.g.: Log-linear models: obtain value at a point in m-D space as the product on appropriate marginal subspaces
- **Non-parametric methods**
  - Do not assume models
  - Major families: **Histograms, Clustering, Sampling**

Data reduction – numerosity reduction (2)

Numerosity reduction – histogram

- Approximate data distributions
- Divide data into buckets and store average (sum) for each bucket
- A bucket represents an attribute-value/frequency pair
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.



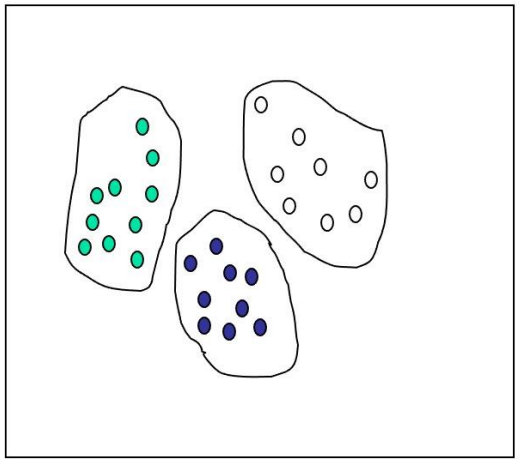
Data reduction – numerosity reduction (3)

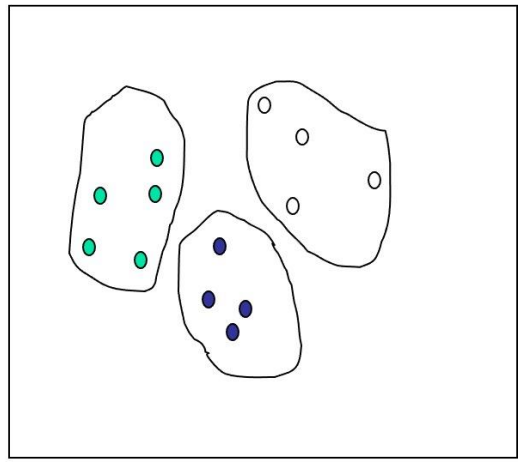
Numerosity reduction – clustering

- Partition data set into clusters, and **store cluster representation only**
- Quality of clusters measured by their diameter (max distance between any two objects in the cluster) or centroid distance (avg. distance of each cluster object from its centroid)
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering (possibly stored in multi-dimensional index tree structures (B+-tree, R-tree, quad-tree, etc))

Data reduction – numerosity reduction (4)

Numerosity reduction – sampling





- Choose a **representative** subset of the data
- 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

Data discretization (1)

- **Three types of attributes:**
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- **Discretization/Quantization:**
  - 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

Data discretization (2)

- **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.

- **Concept Hierarchies**

  Reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

Data discretization (3)

Discretization and concept hierarchy generation for **numeric data**

1. **Hierarchical and recursive decomposition** using:
   - Binning (data smoothing)
   - Histogram analysis (numerosity reduction)
   - Clustering analysis (numerosity reduction)
2. **Entropy-based discretization**
3. **Segmentation by natural partitioning**

Data discretization (4)

Entropy-Based Discretization

- Given a set of samples S, if S is partitioned into two intervals S1 and S2 using threshold T on the value of attribute A, the information gain resulting from the partitioning is:

$$I(S,T) = \frac{|S_1|}{|S|} E(S_1) + \frac{|S_2|}{|S|} E(S_2)$$

  where the entropy function E for a given set is calculated based on the class distribution of the samples in the set. Given m classes the entropy of S1 is:

$$E(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

  where pi is the probability of class i in S1.

- The threshold that maximizes the information gain over all possible thresholds is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g., $E(S) - I(S,T) < \delta$
- Experiments show that it may reduce data size and improve classification accuracy

Data discretization

Segmentation by natural partitioning

- 3-4-5 rule can be used to segment numeric data into relatively uniform, “natural” intervals.
- It **partitions** a given range **into 3,4, or 5 equal-width intervals recursively** level-by-level based on the value range of the most significant digit.
  - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals.
  - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals.
  - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals.

Data discretization

Discretization and concept hierarchy generation for **Categorical Data**

- Categorical data: no ordering among values
- Specification of a partial ordering of attributes explicitly at the schema level by users or experts
- Specification of a portion of a hierarchy by explicit data grouping
- Specification of a set of attributes, but not of their partial ordering
- Specification of only a partial set of attributes

Summary

- Data preparation is a big issue for both warehousing and mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but still an active area of research

香港城市大學
City University of Hong Kong

Thanks for your attention!

Appendix

1. https://en.wikipedia.org/wiki/Data_cleansing

2.https://docs.google.com/presentation/d/1cLSOsGYrzOZo1awGJoaQFQRkYq4shzkGbjXIHBsLS0E/htmlpresent

3. http://hanj.cs.illinois.edu/cs412/bk3/03.pdf