University of Colorado Boulder

CSCI 4502/5502

Data Mining

Fall 2019
Lecture 04 (Sep 5)

Announcements (1)

- Homework 1
  - Posted at moodle, submit online
  - CSCI 4502 vs. 5502, separate grading
  - due at 9:30am, Th Sep 12
  - access ACM DL on campus or via VPN
  - Jupyter Notebook for Question 3
  - SUBMIT your attempt before deadline

Announcements (2)

- No regular instructor office hours next week
  - please contact me for appointment
- Tuesday, Sep 10
  - TBD
- Thursday, Sep 12
  - Guest lecture by Professor Claire Monteleoni on Machine Learning and Environmental Informatics

Review

- Chapter 2: Getting to know your data
  - data objects and attribute types
  - basic statistical description of data
  - data visualization
  - measuring data similarity and dissimilarity

Chapter 3:
Data Preprocessing

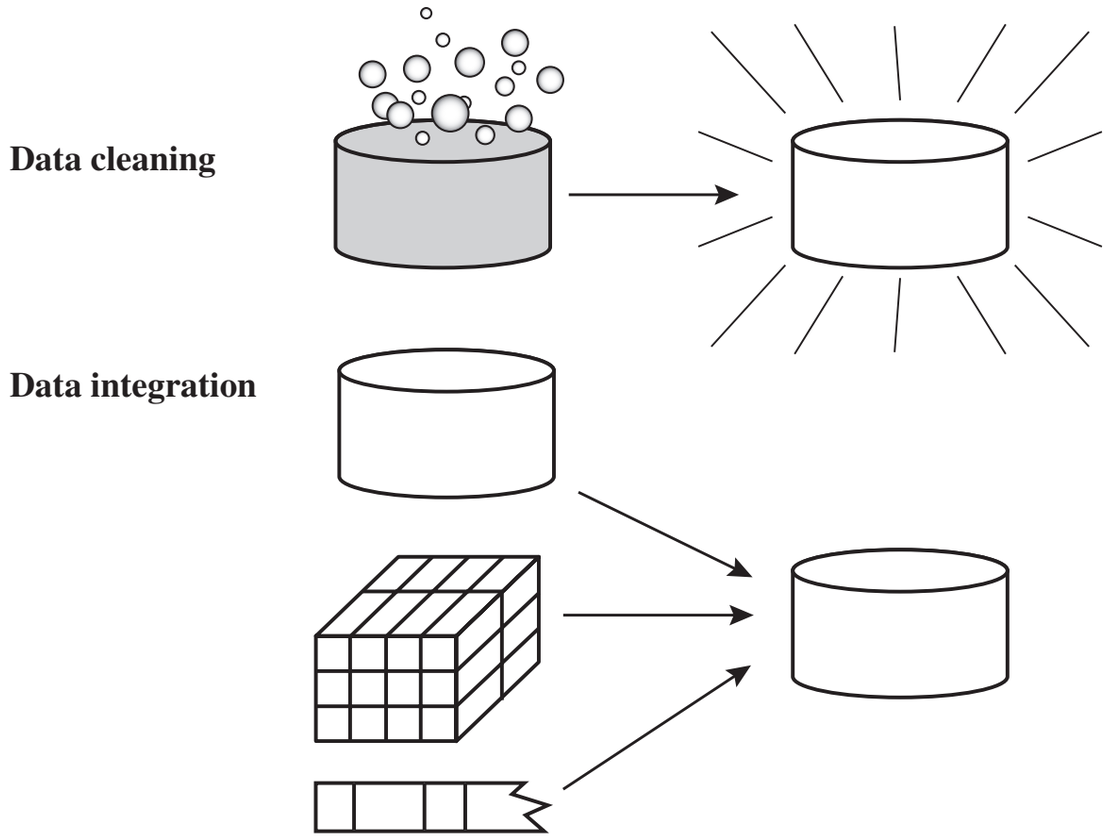
Chap 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization
- Summary

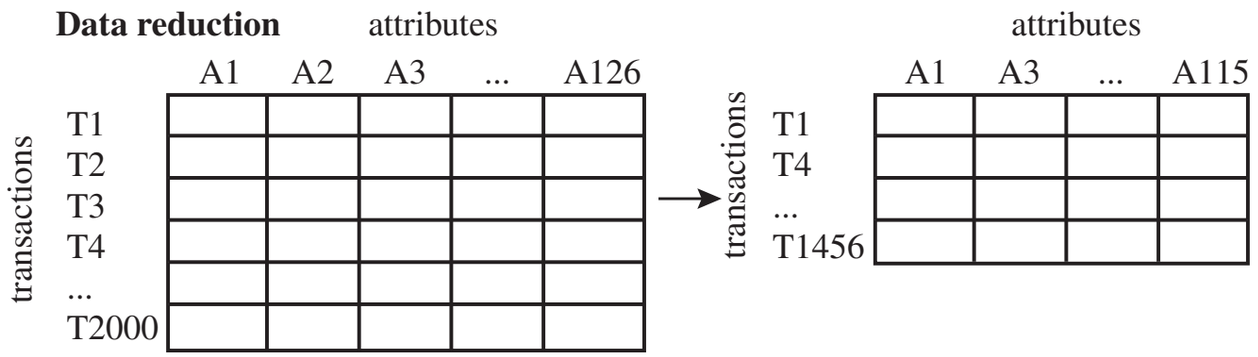
Measures of Data Quality

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability
- Accessibility

Major Tasks in Preprocessing



Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$



Major Tasks in Preprocessing

- Data cleaning
  - fill in missing values, smooth noisy data, identify or remove outliers, resolve inconsistencies
- Data integration
  - integration of multiple data sources
- Data reduction
  - dimensionality, numerosity, compression
- Data transformation and data discretization
  - normalization, concept hierarchy generation

Chap 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization
- Summary

Why Data Cleaning?

- Imperfect data in the real world
- Incomplete: missing attributes, values
  - e.g., age = “”, occupation = “”
- Noisy: containing errors or outliers
  - e.g., salary = “-10”
- Inconsistent: containing discrepancies
  - e.g., age = “42”, birthday = “03/07/1997”
  - e.g., ratings of “1, 2, 3” and “A, B, C”

Why Are Data Imperfect?

- Incomplete data
  - "not applicable" values
  - time between collection and analysis
  - human/hardware/software problems
- Noisy data
  - faulty data collection instruments
  - human or computer error at data entry
  - errors in data transmission

Why Are Data Imperfect?

- Inconsistent data
  - different data sources
  - naming conventions, data formats
    - e.g., date "03/07/11"
  - functional dependency violation
    - e.g., modify some linked data

- No quality data, no quality data mining results!

How to Handle Missing Data?

- Ignore the tuple
- Fill in the missing value manually
- Fill in it automatically with
  - a global constant
  - the attribute mean
  - the attribute mean of the same class
  - the most probable value: E.g., regression, Bayesian inference, decision tree

How to Handle Noisy Data?

- Binning
  - first sort & partition data into bins
  - then smooth by
    - bin means
    - bin median
    - bin boundaries

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

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

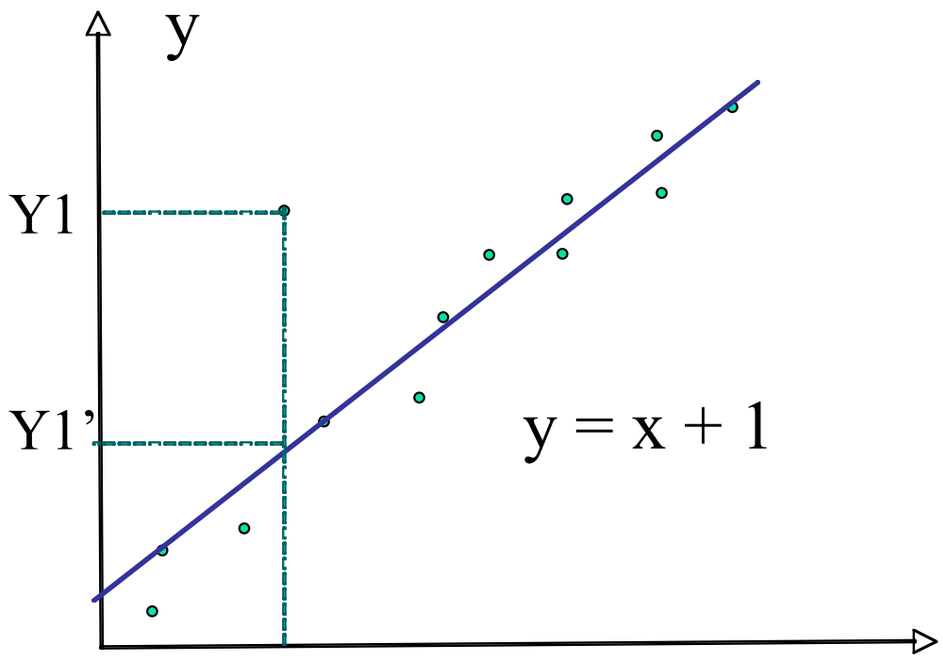
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

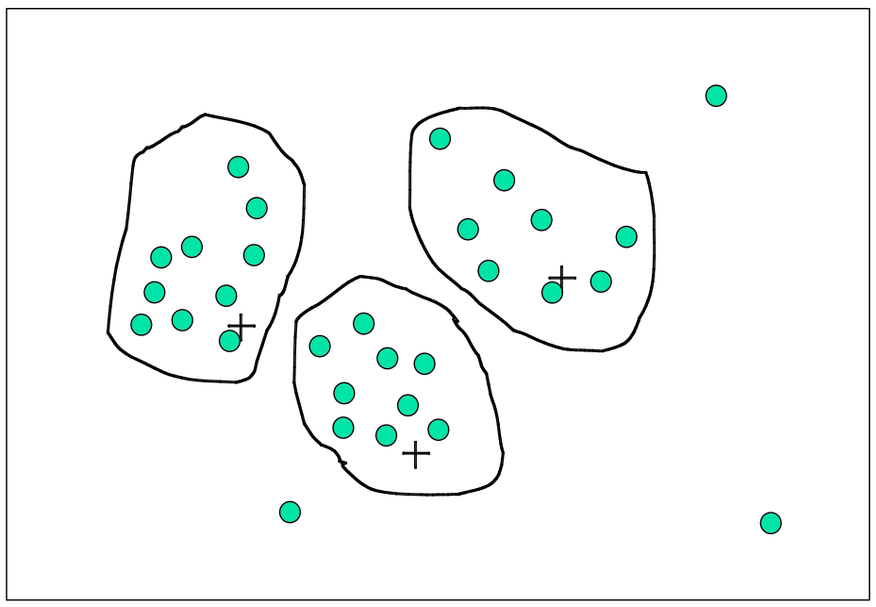
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

How to Handle Noisy Data?

- Regression
  - fit data into regression functions



- Clustering
  - detect and remove outliers



Chap 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization
- Summary

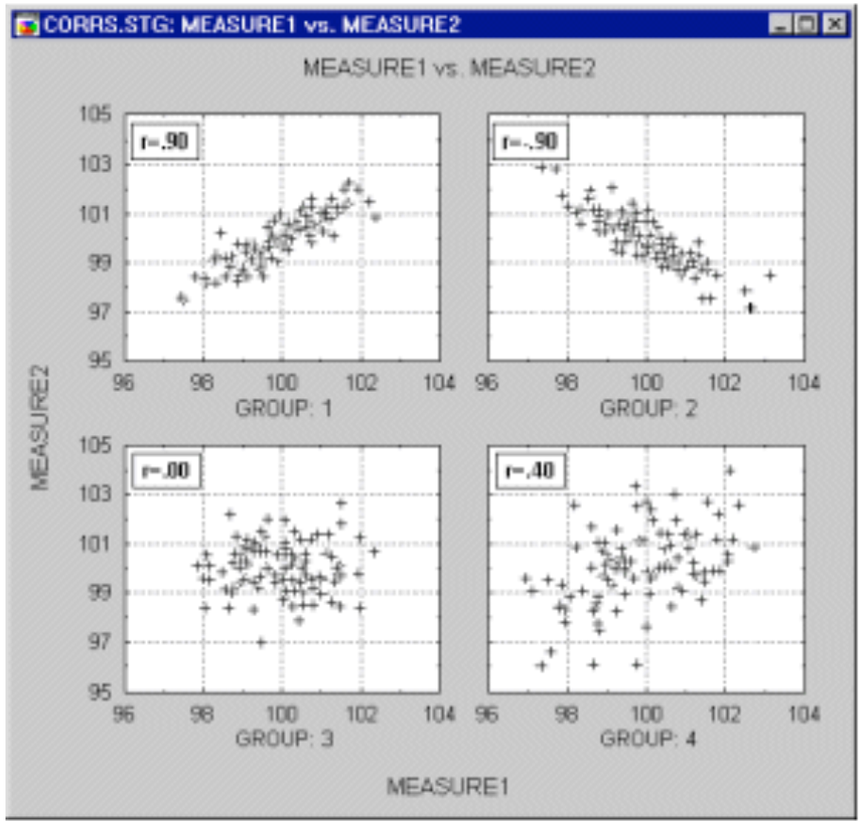
Data Integration

- Combines data from multiple sources
- Entity identification
  - schema integration, object matching
  - e.g., customer_id vs. cust_number
  - e.g., Bill Clinton vs. William Clinton
- Redundant data
  - different naming, derived data
  - may be detected by correlation analysis

Correlation Analysis (1)

- Correlation coefficient (numerical data)

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$



http://www.uta.edu/faculty/sawasthi/Statistics/popups/popup2.gif

Correlation Analysis (2)

- X^2 (chi-square) test (categorical data)

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

Chi-Square Test: An Example

	play chess	not play chess	total
like fiction	250 (90)	200 (360)	450
not like fiction	50 (210)	1000 (840)	1050
total	300	1200	1500

$$e_{11} = \frac{\#(like\ fiction) \times \#(play\ chess)}{N} = \frac{300 \times 450}{1500} = 90$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Correlation Analysis (3)

- Does correlation imply causality?
  - sleeping with one's shoes on is strongly correlated with waking up with a headache
  - the more fireman fighting a damage, the more damage there is going to be
  - as ice cream sales increases, the rate of drowning deaths increases sharply
  - correlation does not imply causality!

http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Chap 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization
- Summary

Data Reduction

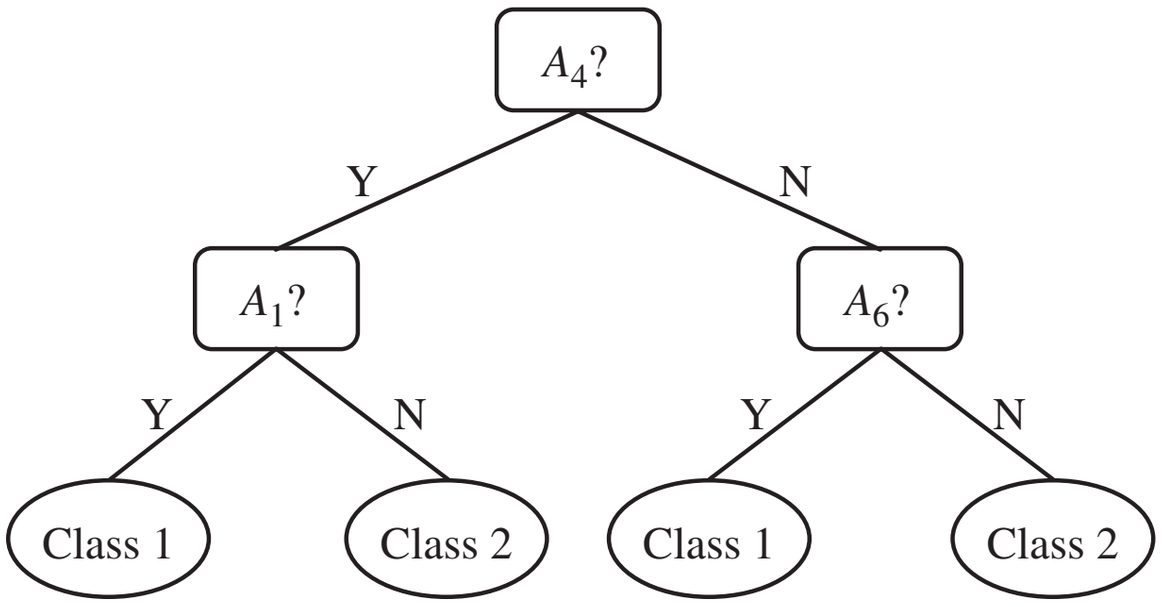
- Why data reduction?
  - massive data sets
  - mining takes a long time
- Goal of data reduction
  - data set is much smaller in volume
  - produces (almost) the same mining results

Data Reduction Strategies

- Dimensionality reduction
  - attribute subset selection
  - Wavelet transform
  - principle component analysis (PCA)
- Numerosity reduction
  - regression, log-linear models
  - data cube aggregation
  - histograms, clustering, sampling
- Data compression

Attribute Subset Selection

- Remove irrelevant or redundant attributes

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Dimensionality Reduction

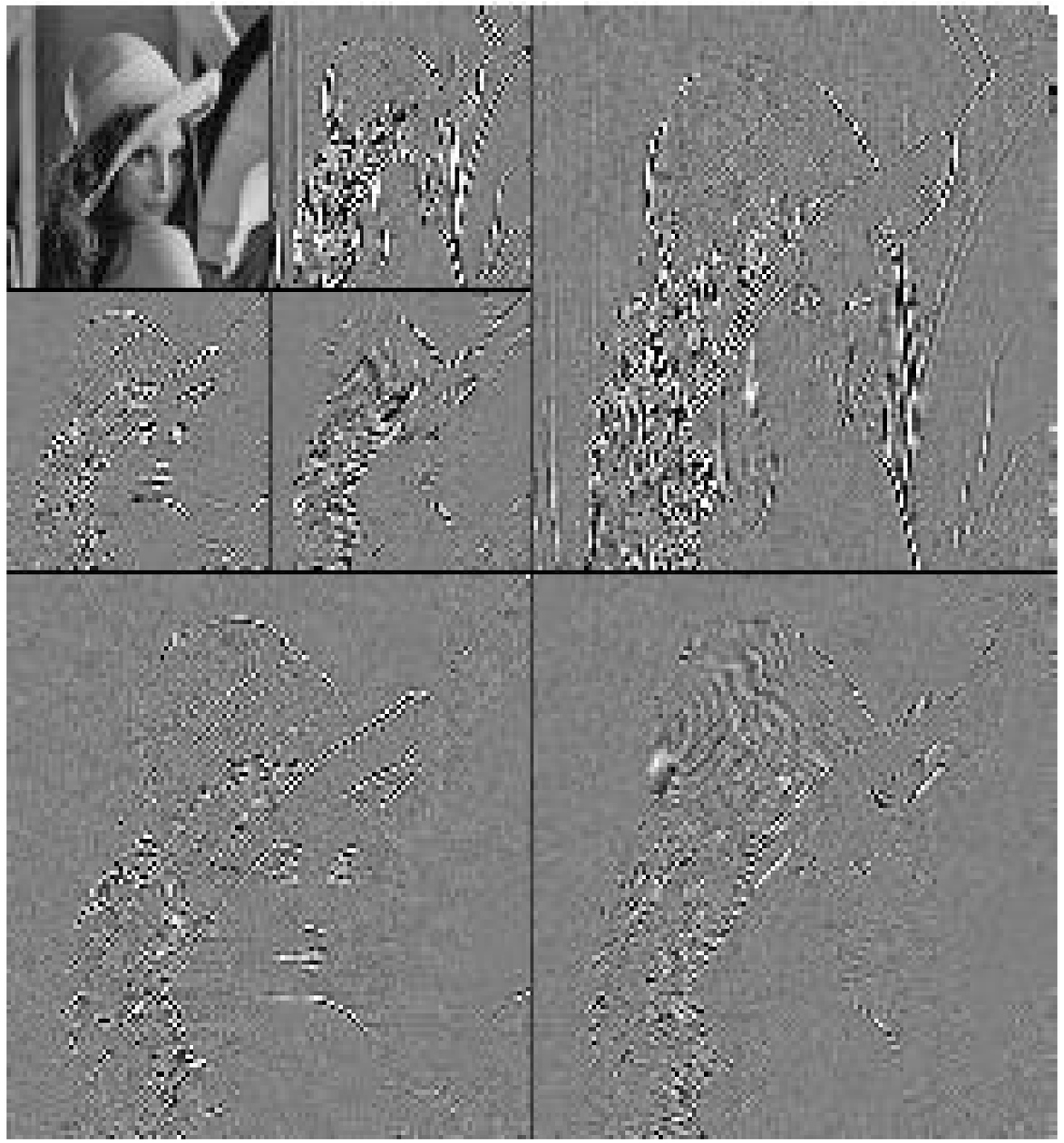
- Discrete wavelet transform (DWT)
  - linear signal processing, multi-resolution
  - store only a small fraction of the strongest wavelet coefficients

20 coeffs 100 coeffs 400 coeffs 16,000 coeffs

http://grail.cs.washington.edu/projects/query/

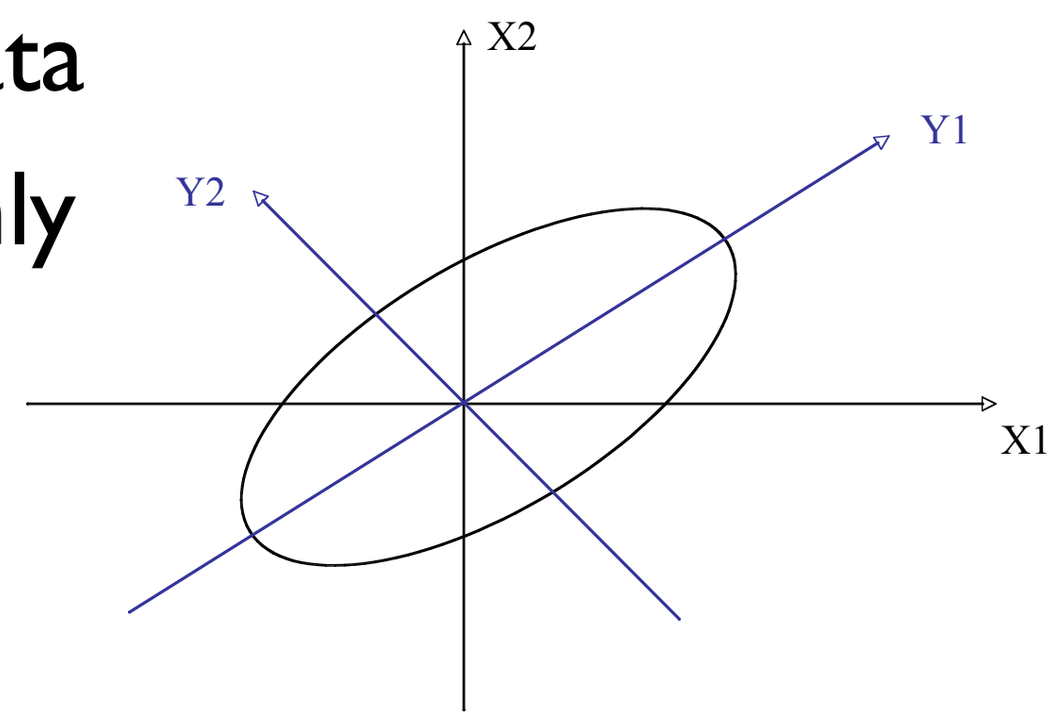
DWT for Image Compression

- Image
  - high pass
  - low pass
    - high pass
    - low pass
      - high pass
      - low pass



Dimensionality Reduction

- Principal component analysis (PCA)
  - given N data vectors of n dimensions
  - find k <= n orthogonal vectors (principal components) that can
  - best represent the data
  - for numerical data only
  - used when n is large

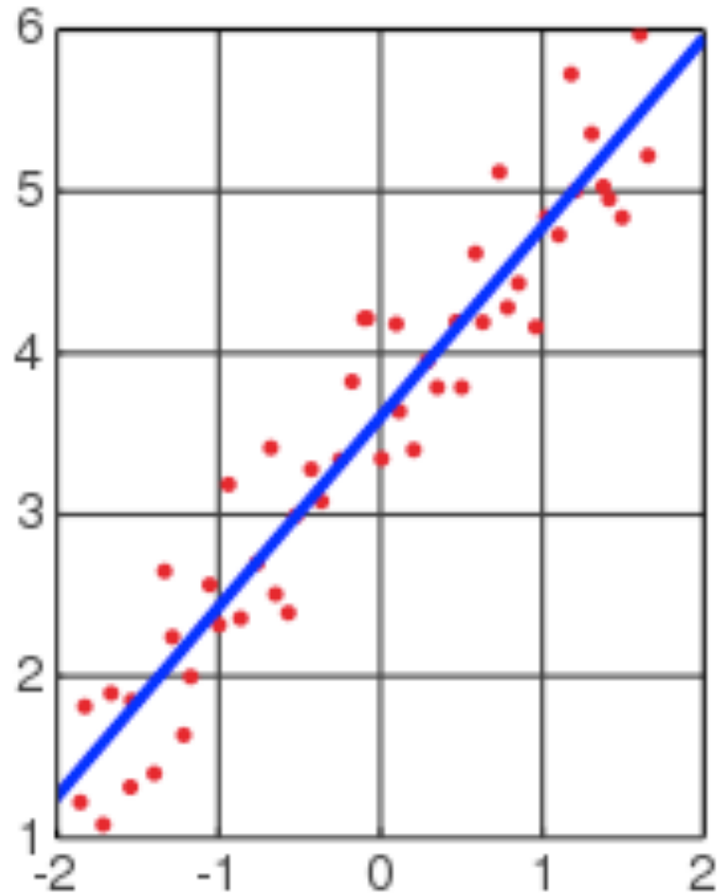


Numerosity Reduction

- Use alternative, smaller data representations
- Parametric methods
  - assume the data fits some model
  - estimate model parameters
  - store the parameters, discard the data
- Non-parametric methods
  - do not assume models
  - e.g., histograms, clustering, sampling

Regression & Log-Linear Models

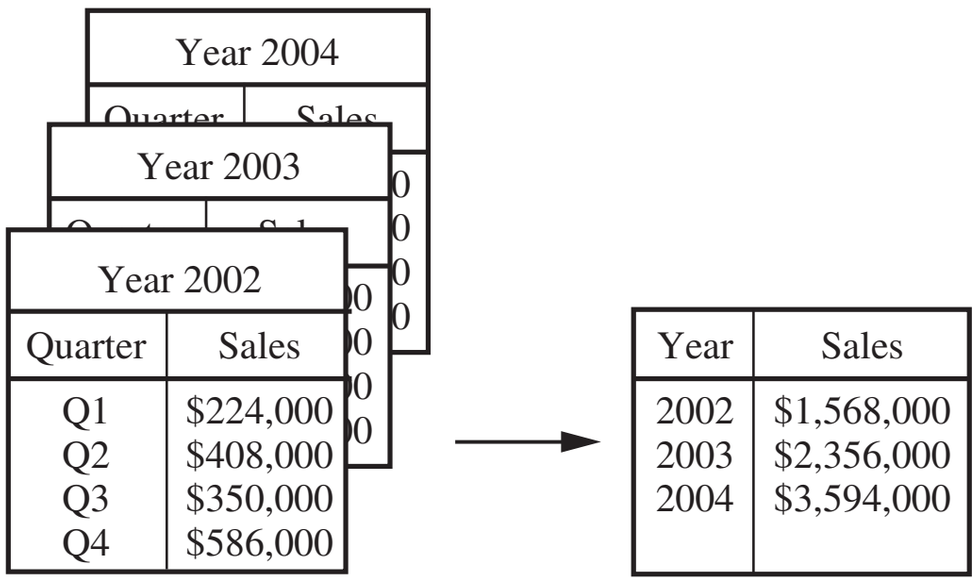
- Linear regression
  - Y = w X + b
- Multiple regression
  - Y = b0 + b1 X1 + b2 X2
- Log-linear models
  - approximate multi-dimensional probability distributions with lower-dimensional distributions

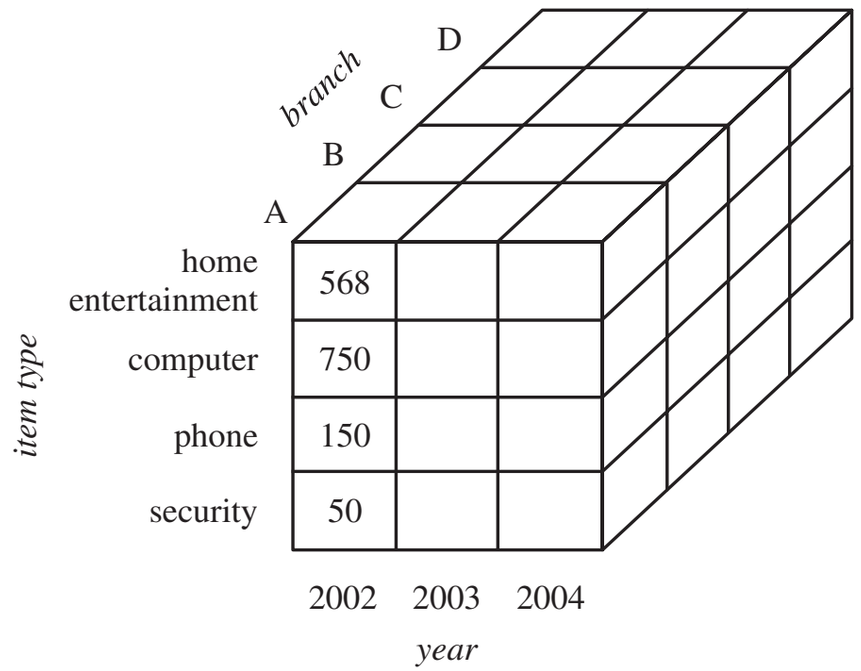


http://en.wikipedia.org/wiki/Linear_regression

Data Cube Aggregation

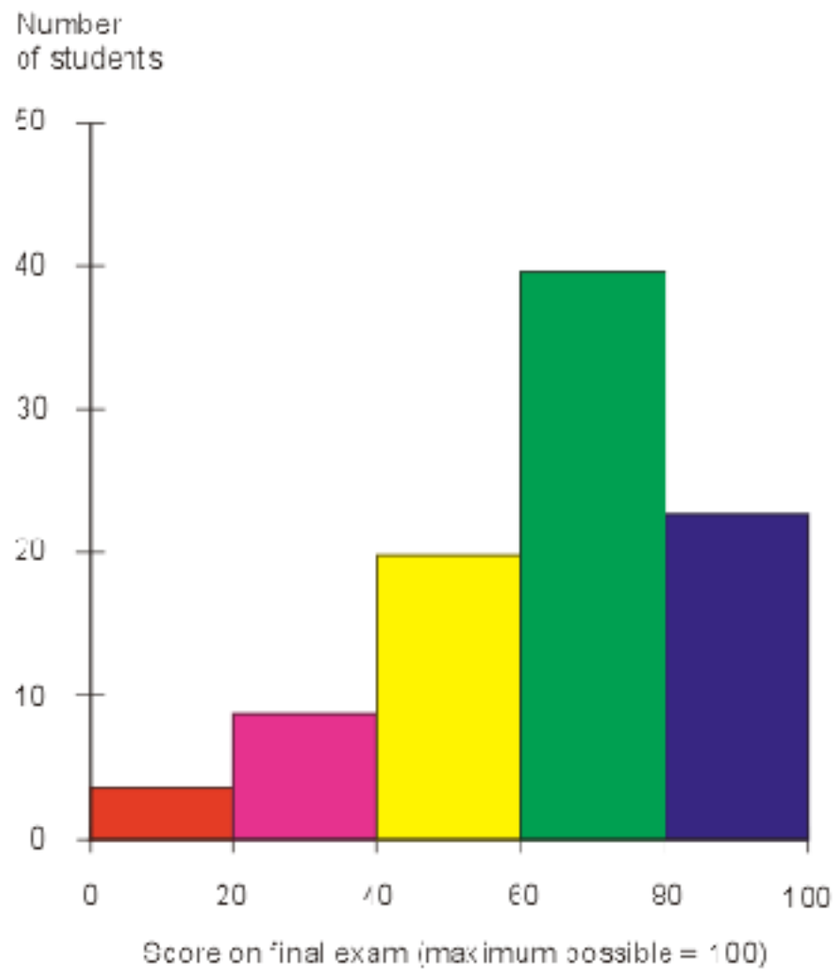
- E.g., quarterly sales => annual sales
- Multiple levels of aggregation may be possible
- Use the smallest representation which is enough for the task





Histograms

- Divide data into buckets and store average (or sum) for each bucket
- Partitioning rules
  - equal-width
  - equal-frequency
  - v-optimal
  - max-diff



http://media.techtarget.com/digitalguide/images/Misc/iw_histogram.gif

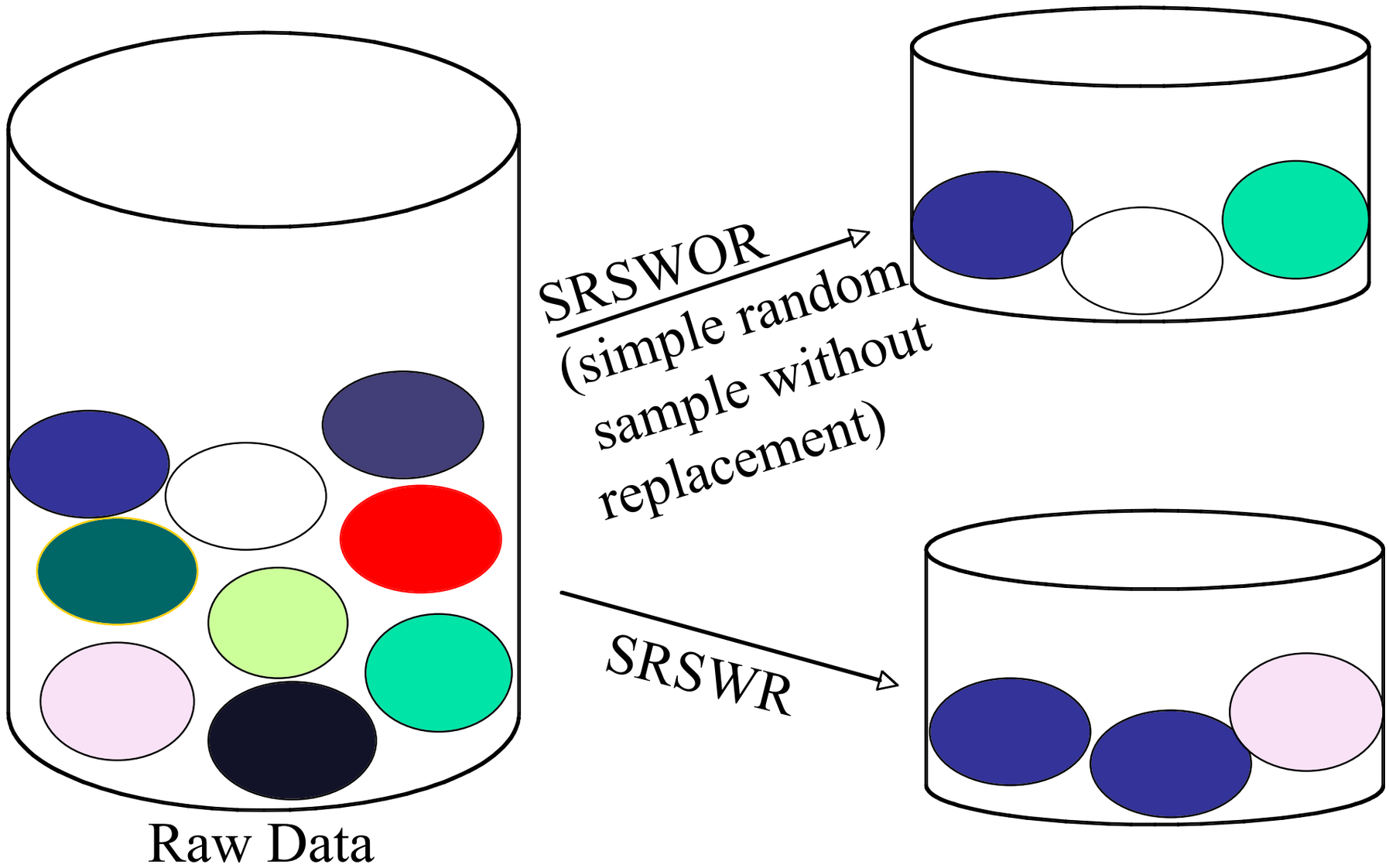
Clustering

- Partition data into clusters based on similarity
- Store cluster representations only
  - e.g., centroid and diameter
- Can have hierarchical clustering
- Many choices of clustering definitions and clustering algorithms

Sampling

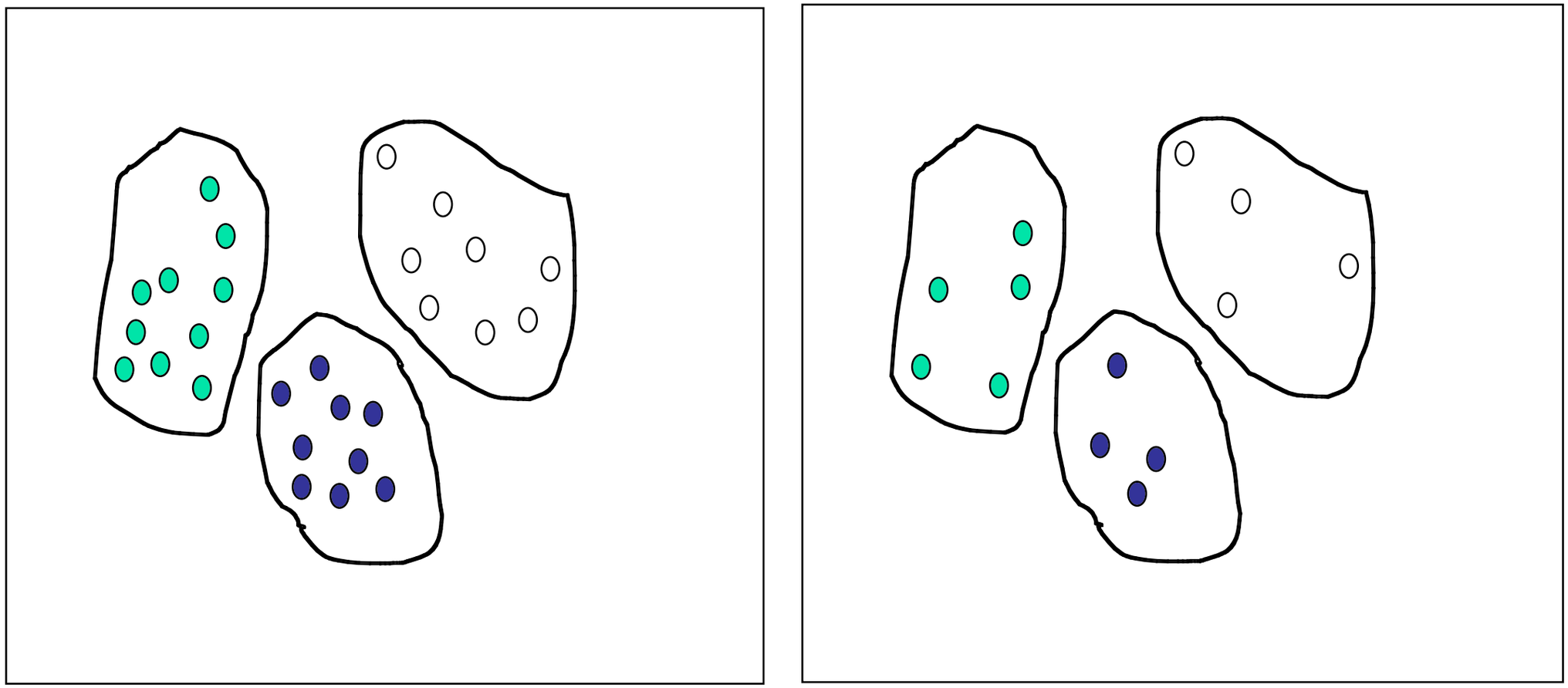
- Use a small sample to represent whole data
- Choose a representative subset of the data
  - simple random sampling may have very poor performance in the presence of skew
- Simple random sample without replacement
- Simple random sample with replacement
- Cluster sample
- Stratified sample

Sample w/ or w/o Replacement



Cluster or Stratified Sampling

- Approximate the percentage of each class



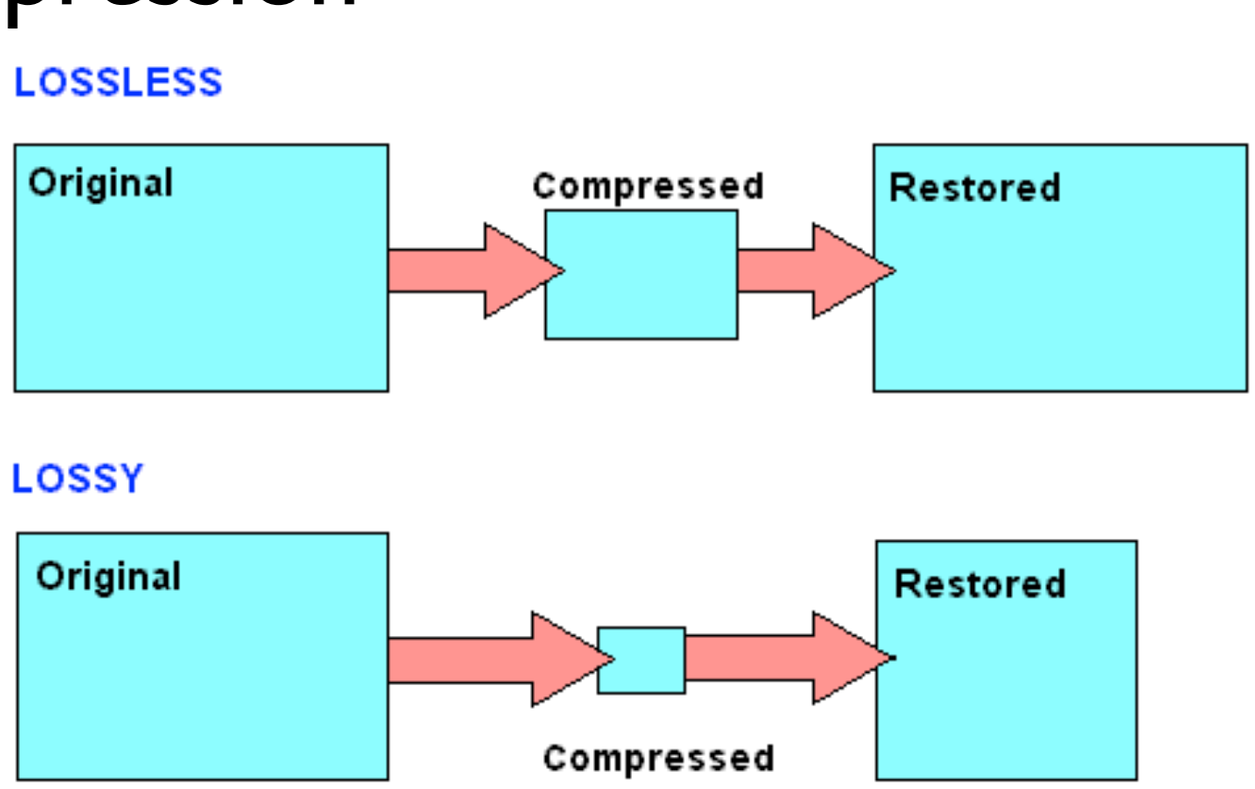
raw data cluster/stratified sample

Data Compression

- Lossy vs. lossless compression
- Example
  - string compression
  - audio/video compression

From Computer Desktop Encyclopedia
© 1998 The Computer Language Co. Inc.

http://img.tfd.com/cde/LOSSY.GIF



Chap 3: Data Preprocessing

- Data preprocessing overview
  - data quality
  - major tasks in data preprocessing
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization
- Summary