

# CS6220: DATA MINING TECHNIQUES

## 2: Data Pre-Processing

---

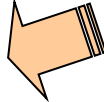
**Instructor: Yizhou Sun**

[yzsun@ccs.neu.edu](mailto:yzsun@ccs.neu.edu)

September 10, 2013

# 2: Data Pre-Processing

---

- Getting to know your data 
  - Basic Statistical Descriptions of Data
  - Data Visualization
- Data Pre-Processing
  - Data Cleaning
  - Data Integration
  - Data Reduction
  - Data Transformation and Data Discretization

# Basic Statistical Descriptions of Data

---

- Central Tendency
- Dispersion of the Data
- Graphic Displays

# Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

Note:  $n$  is sample size and  $N$  is population size.

- Weighted arithmetic mean:
- Trimmed mean: chopping extreme values

- Median:

- Middle value if odd number of values, or average of the middle two values otherwise

- Estimated by interpolation (for *grouped data*):

$$median = L_1 + \left( \frac{n/2 - (\sum freq)l}{freq_{median}} \right) width$$

- Mode

- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula:  $mean - mode = 3 \times (mean - median)$

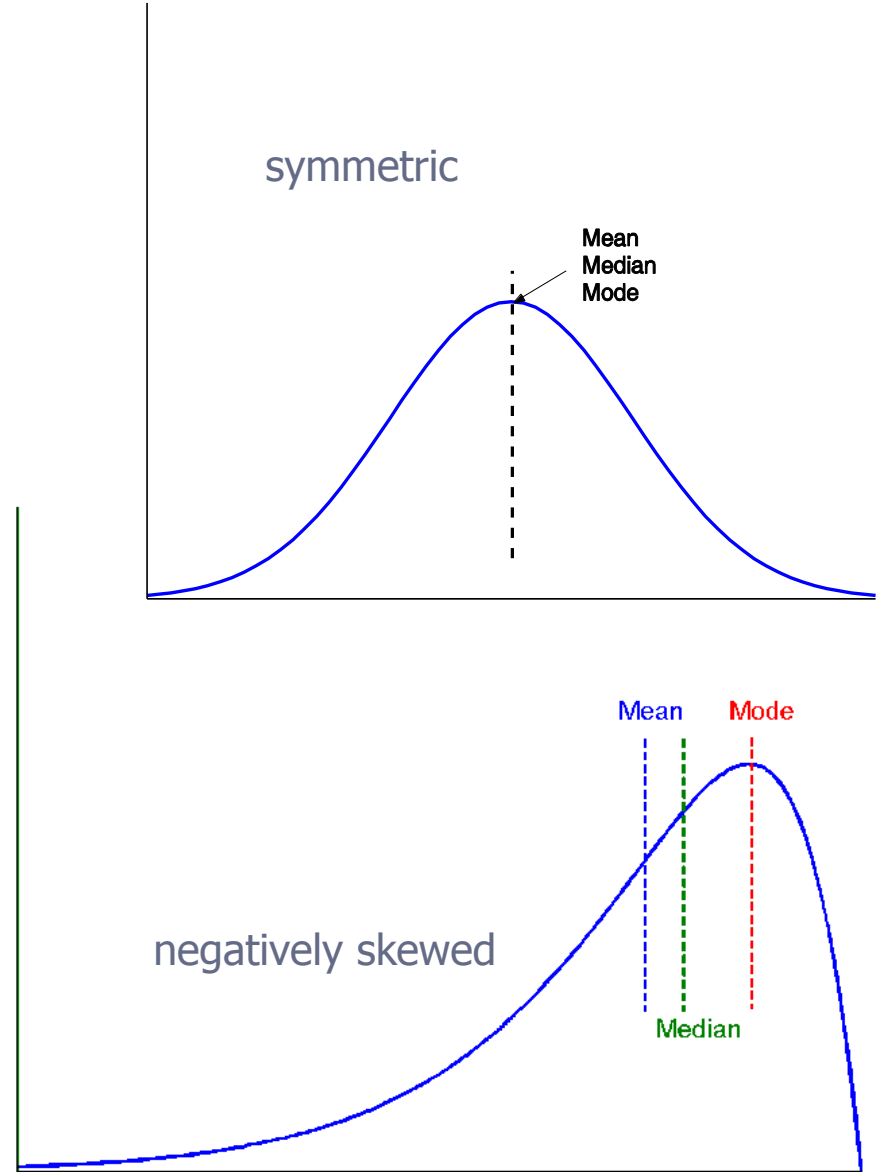
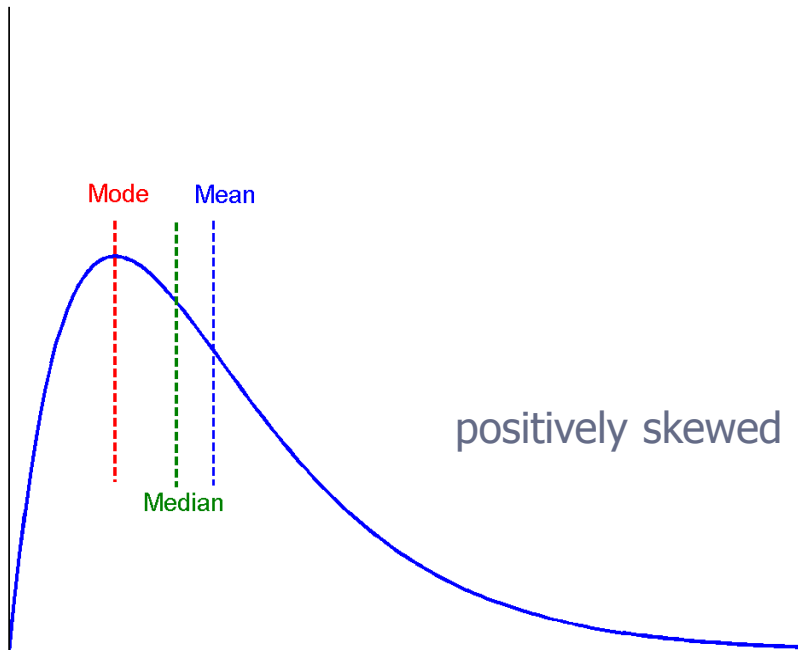
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

<i>age</i>	<i>frequency</i>
1-5	200
6-15	450
16-20	300
21-50	1500
51-80	700
81-110	44

# Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



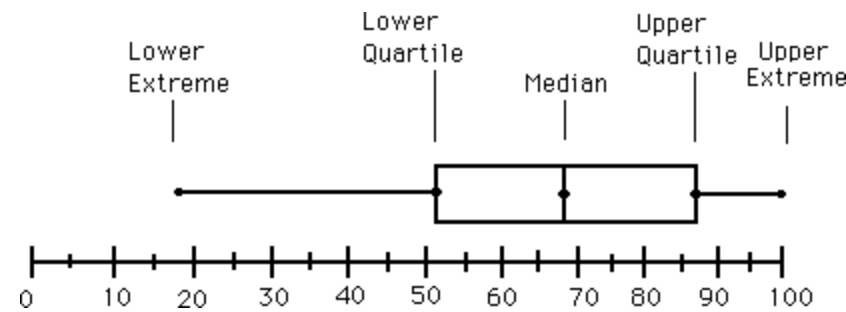
# Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles:**  $Q_1$  (25<sup>th</sup> percentile),  $Q_3$  (75<sup>th</sup> percentile)
  - **Inter-quartile range:**  $IQR = Q_3 - Q_1$
  - **Five number summary:** min,  $Q_1$ , median,  $Q_3$ , max
  - **Boxplot:** ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier:** usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample:  $s$ , population:  $\sigma$* )
  - **Variance:** (algebraic, scalable computation)

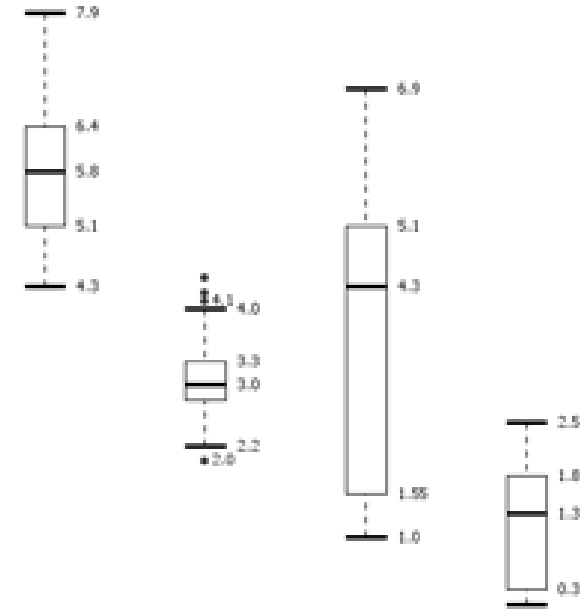
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right] \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

- **Standard deviation  $s$  (or  $\sigma$ )** is the square root of variance  $s^2$  (or  $\sigma^2$ )

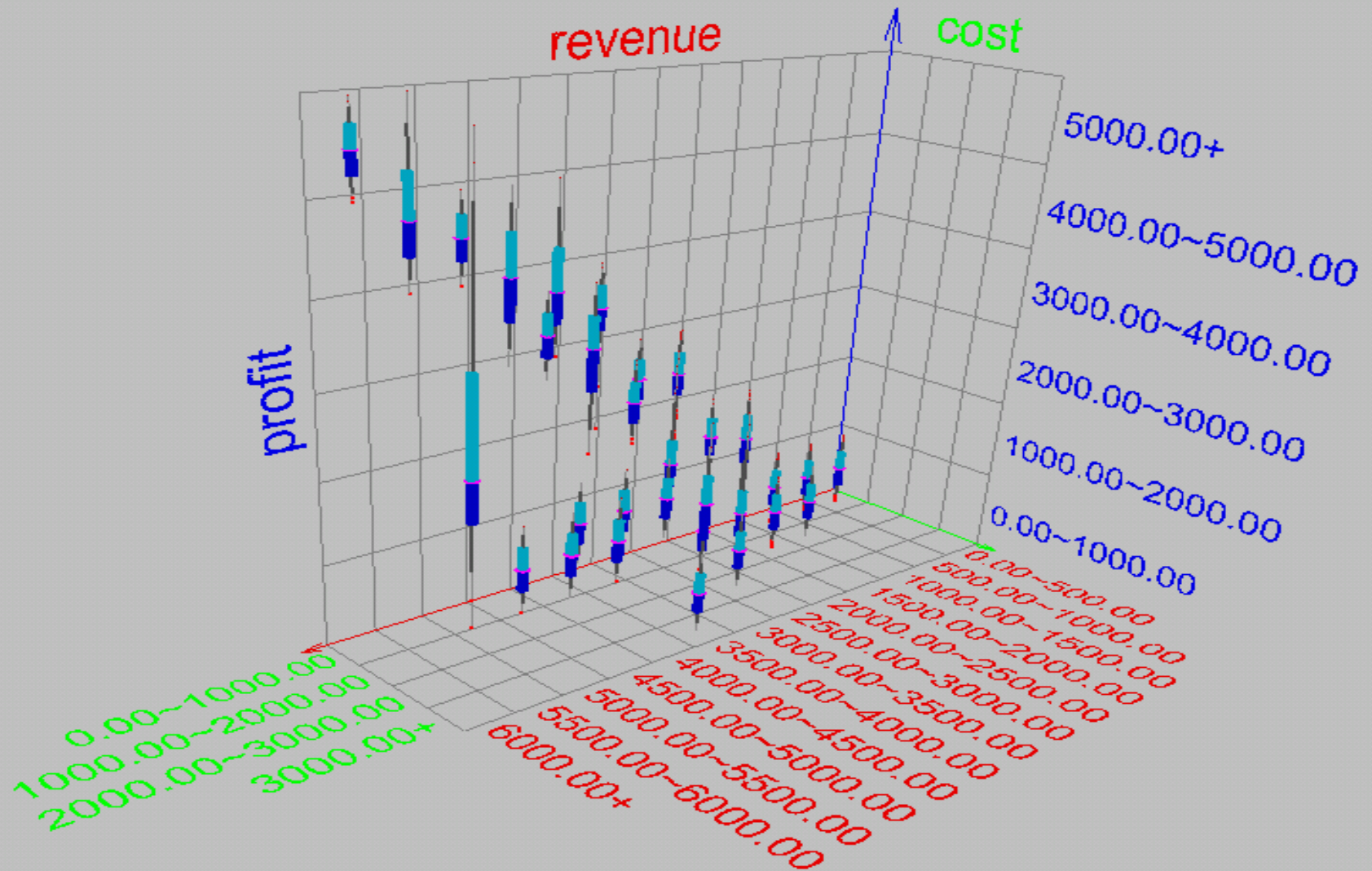
# Boxplot Analysis



- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



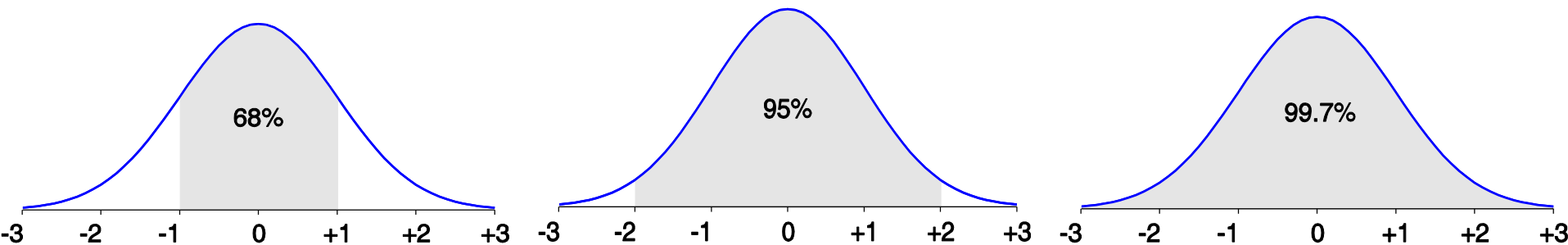
# Visualization of Data Dispersion: 3-D Boxplots





# Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From  $\mu - \sigma$  to  $\mu + \sigma$ : contains about 68% of the measurements ( $\mu$ : mean,  $\sigma$ : standard deviation)
  - From  $\mu - 2\sigma$  to  $\mu + 2\sigma$ : contains about 95% of it
  - From  $\mu - 3\sigma$  to  $\mu + 3\sigma$ : contains about 99.7% of it



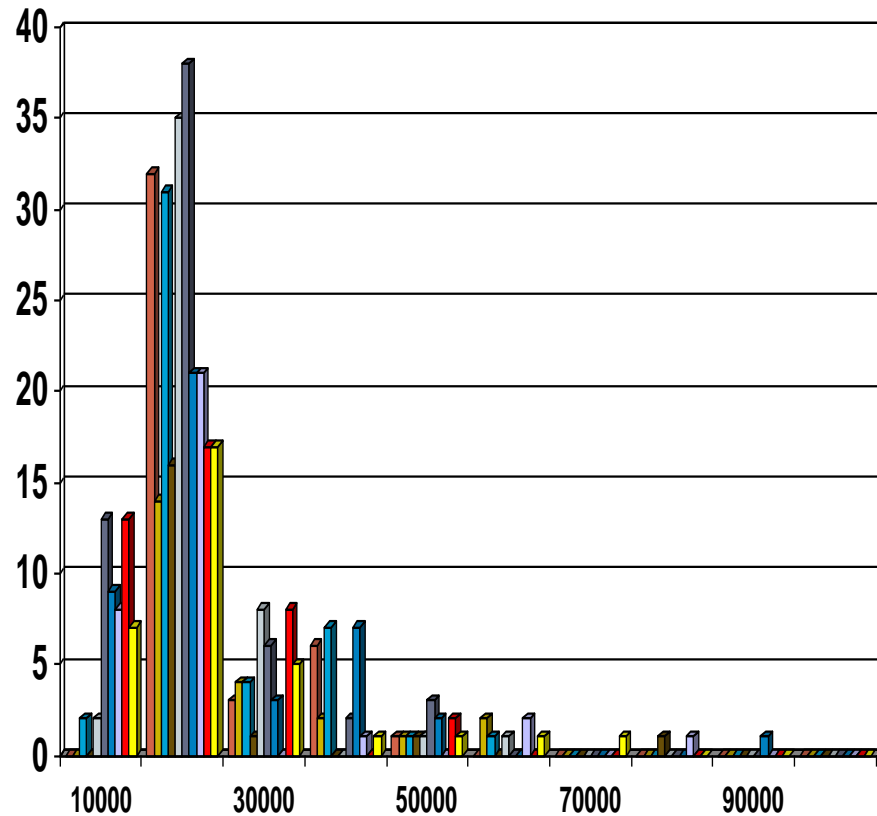
# Graphic Displays of Basic Statistical Descriptions

---

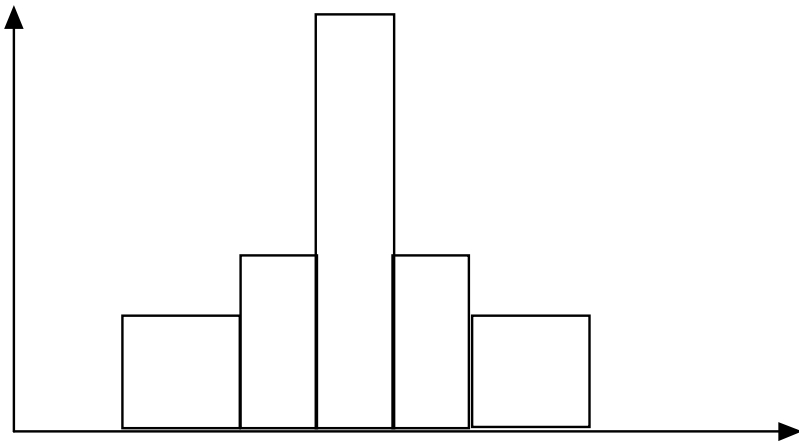
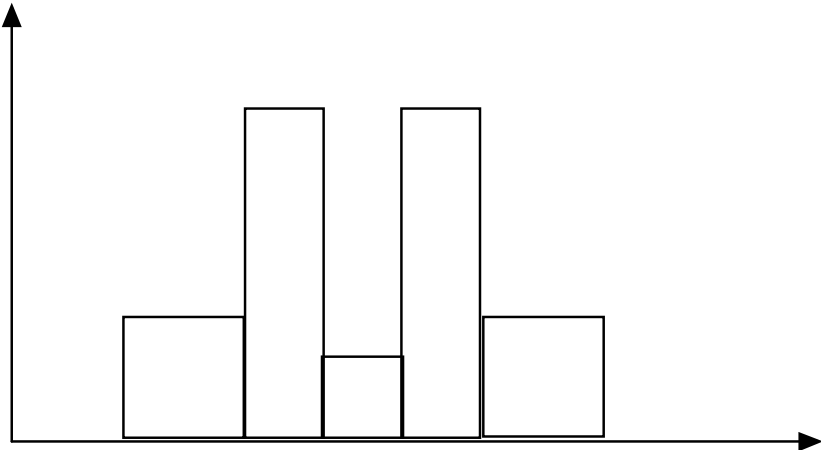
- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis repres. frequencies
- **Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane

# Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



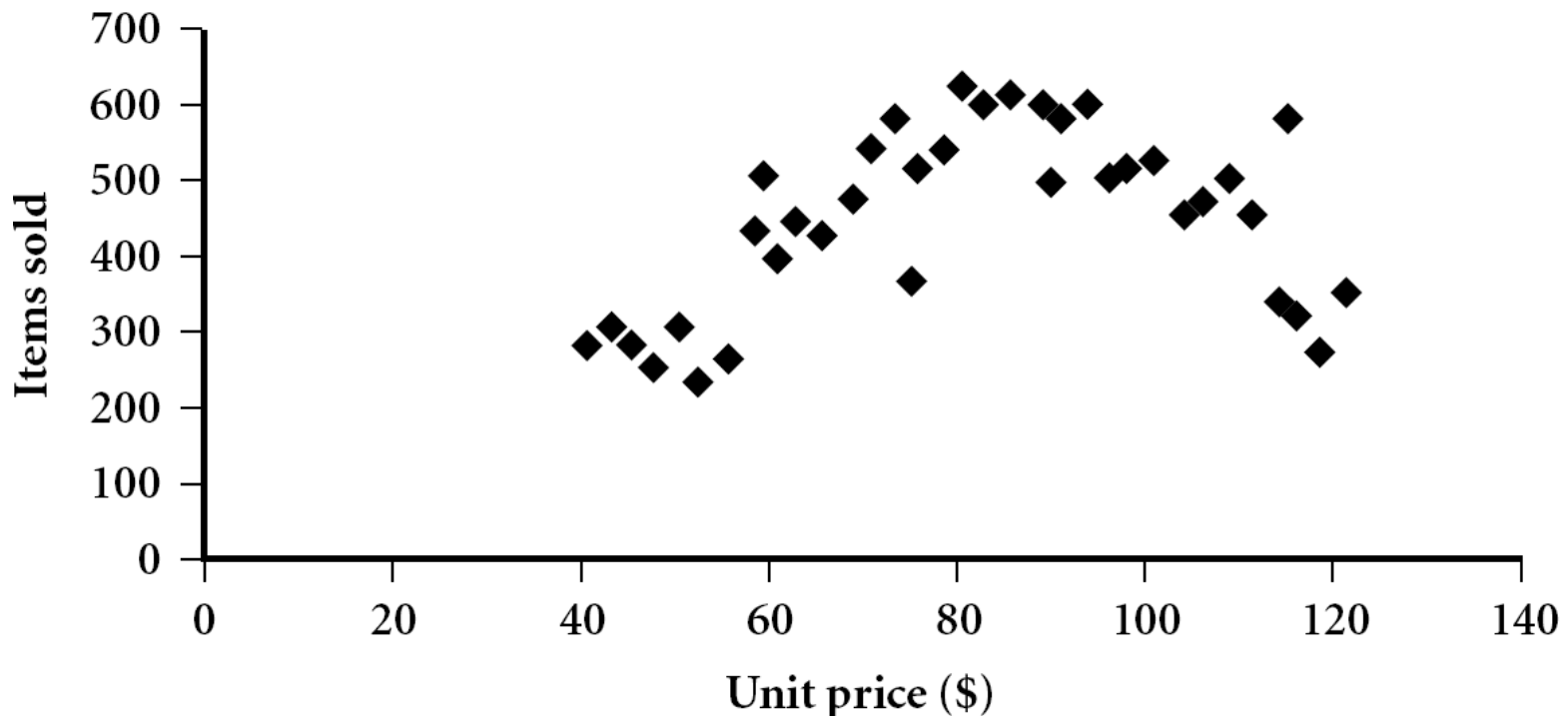
# Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

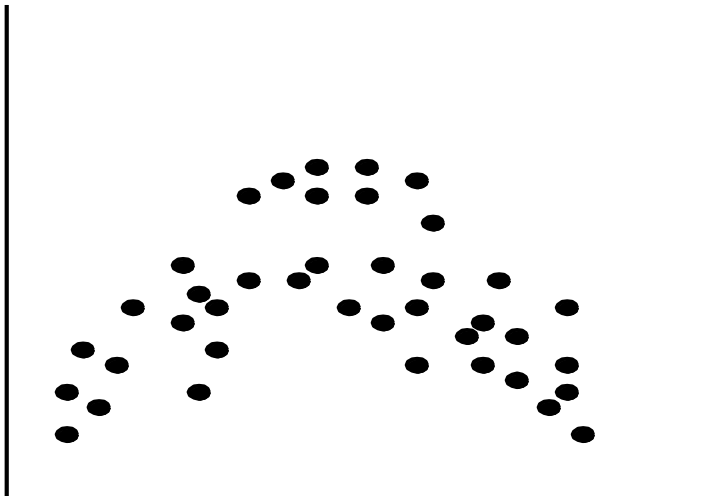
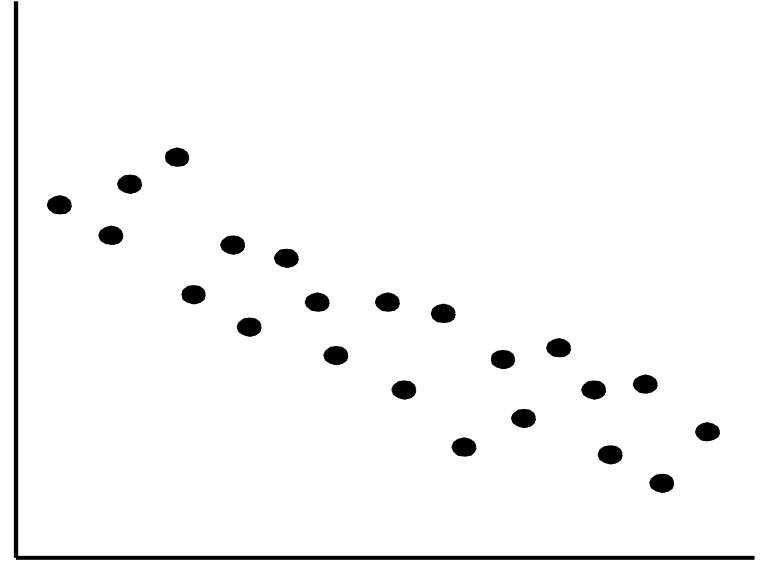
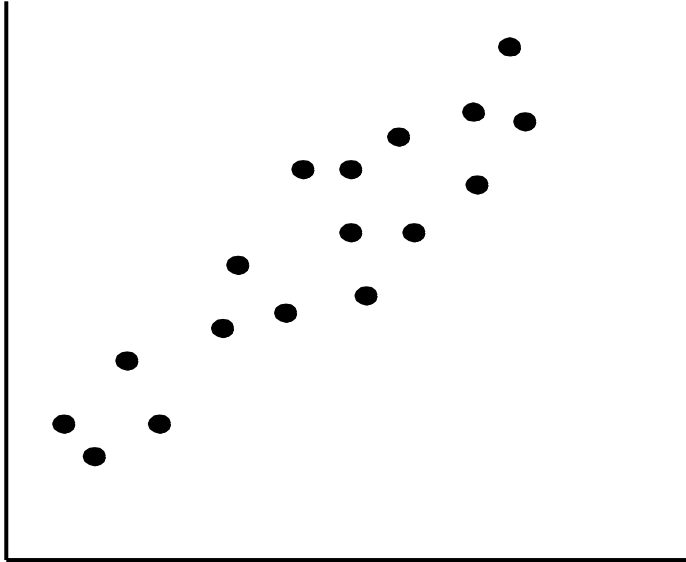
# Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



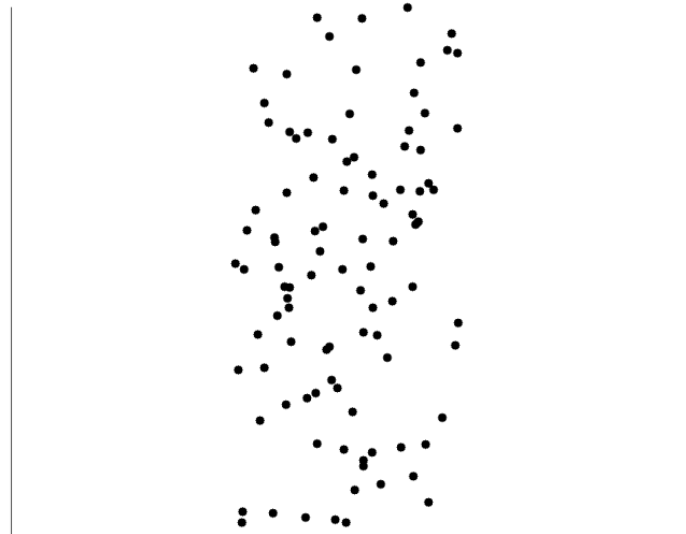
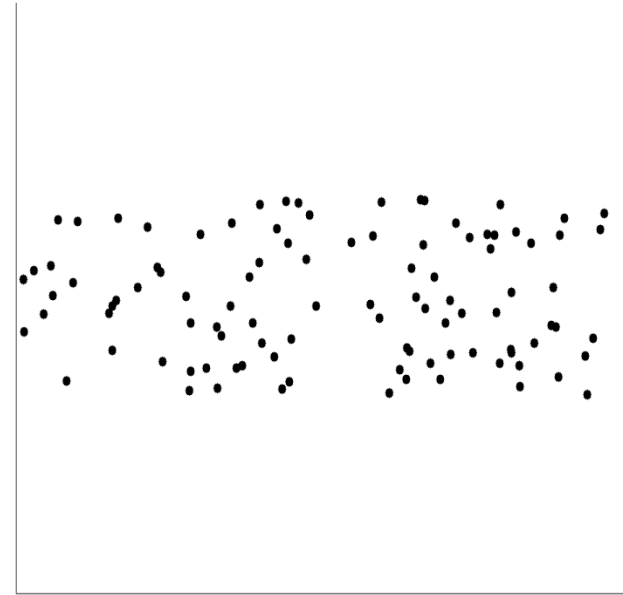
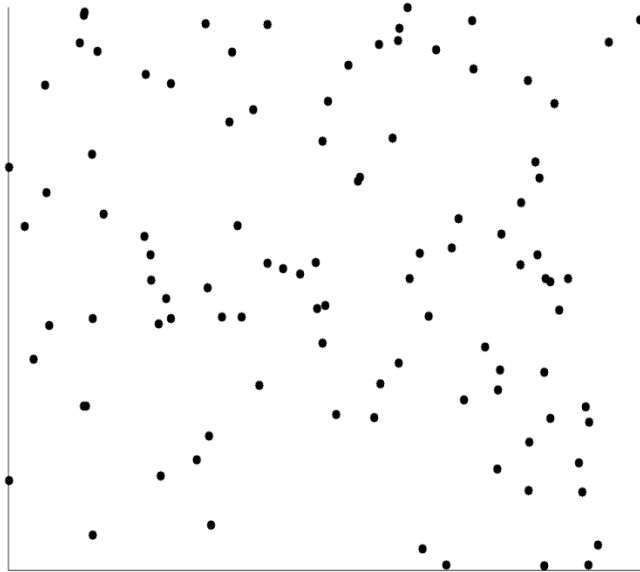
# Positively and Negatively Correlated Data

---



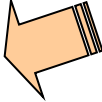
- The left half fragment is positively correlated
- The right half is negative correlated

# Uncorrelated Data



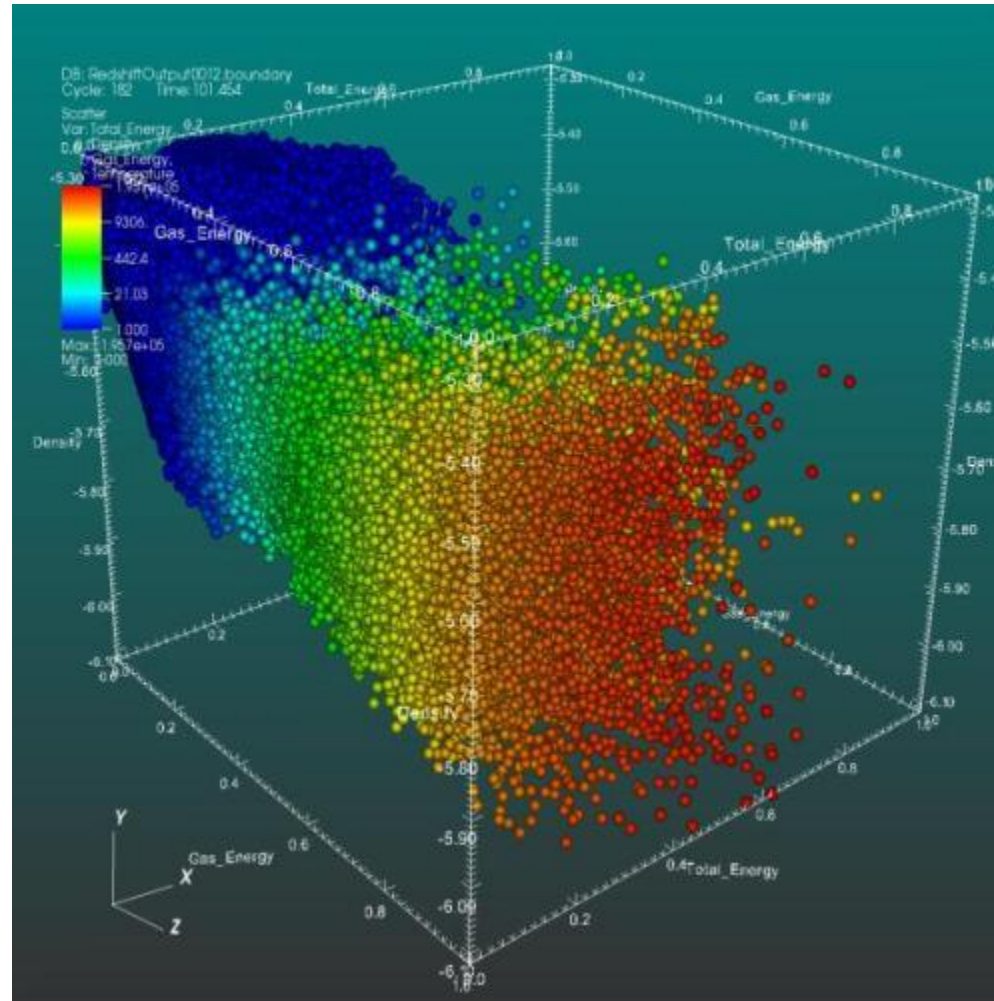
# 2: Data Pre-Processing

---

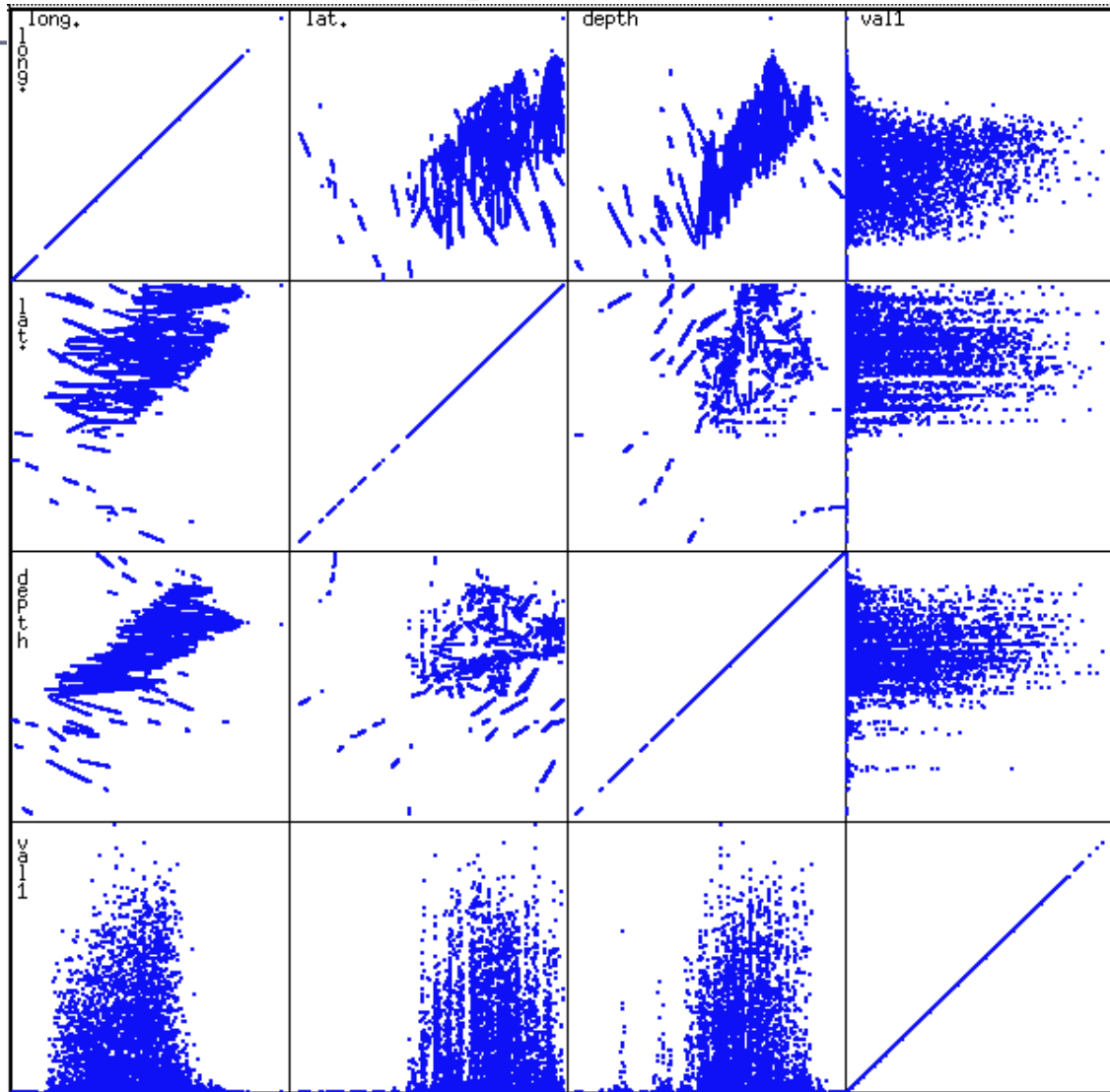
- Getting to know your data
  - Basic Statistical Descriptions of Data
  - Data Visualization 
- Data Pre-Processing
  - Data Cleaning
  - Data Integration
  - Data Reduction
  - Data Transformation and Data Discretization



# 3D Scatter Plot



# Scatterplot Matrices

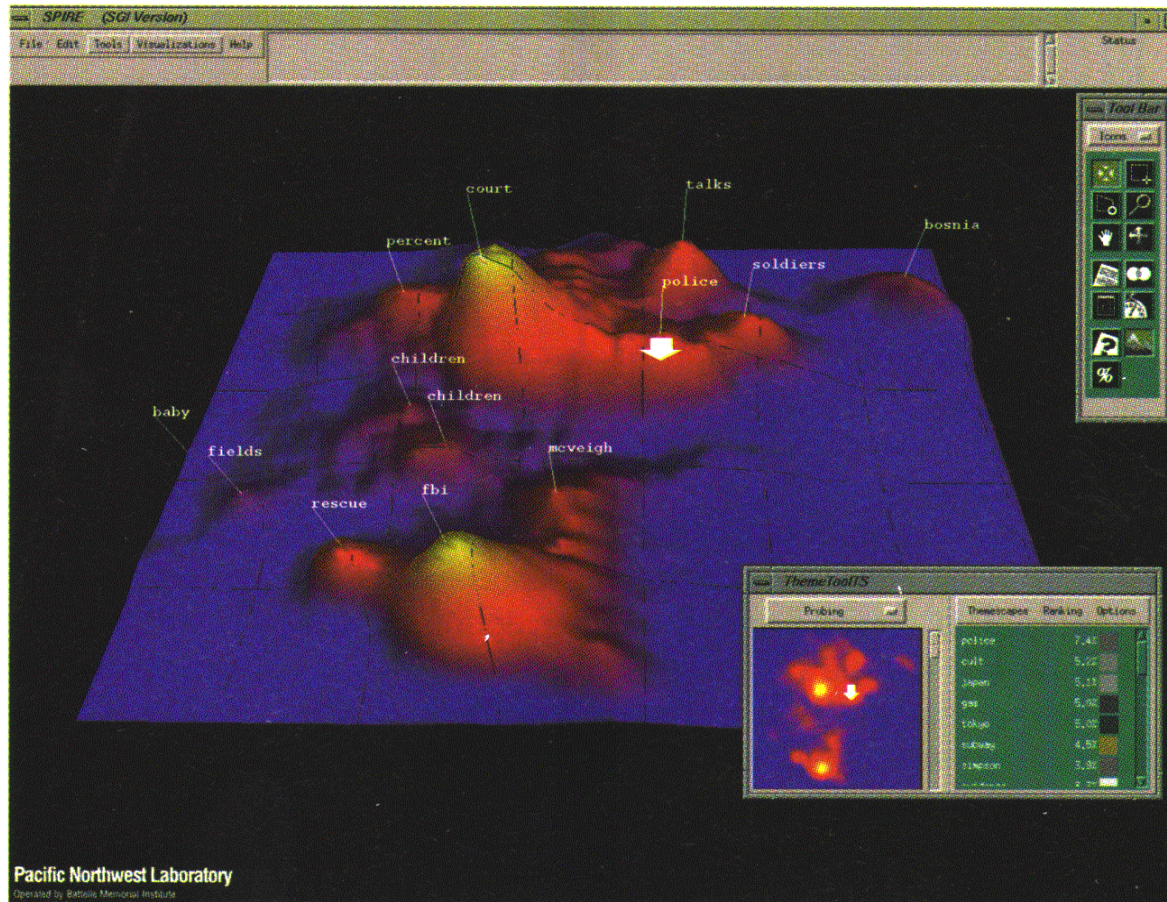


Used by permission of M. Ward, Worcester Polytechnic Institute

Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of  $(k^2/2-k)$  scatterplots]

# Landscapes

Used by permission of B. Wright, Visible Decisions Inc.

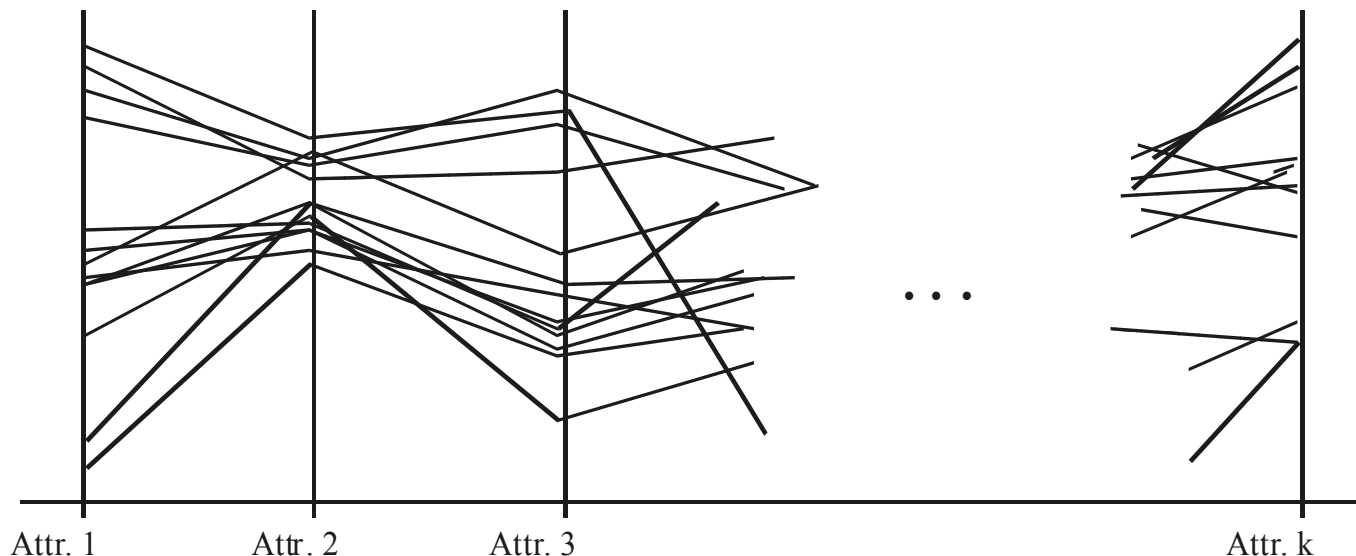


news articles  
visualized as  
a landscape

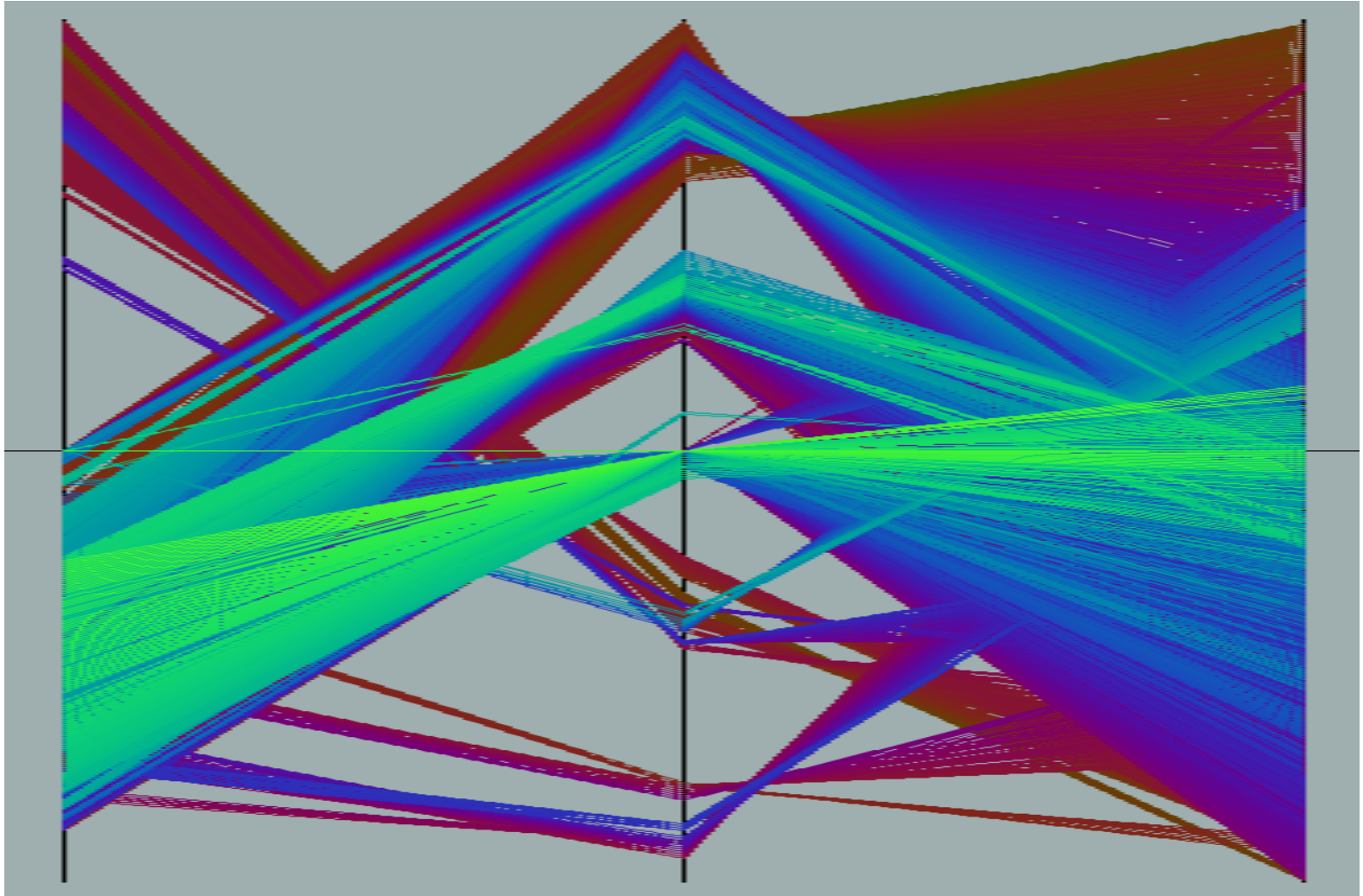
- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

# Parallel Coordinates

- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute



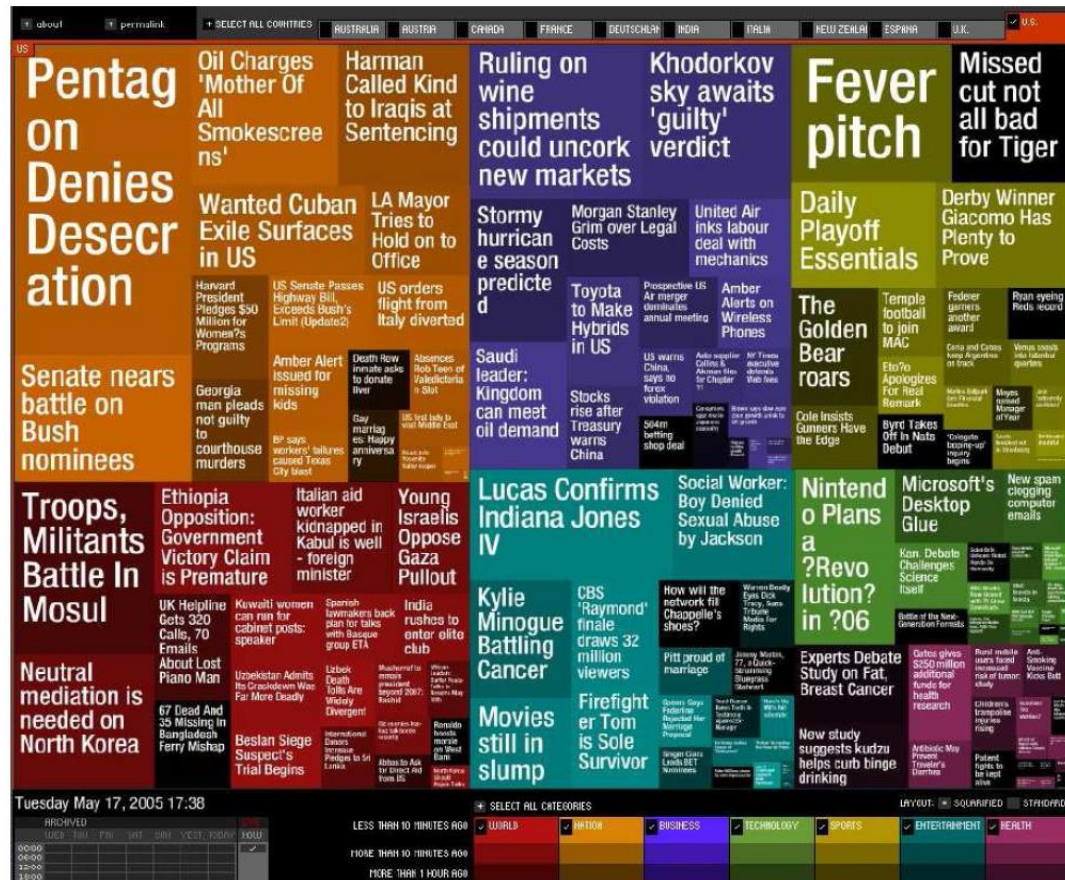
# Parallel Coordinates of a Data Set



# Visualizing Text Data

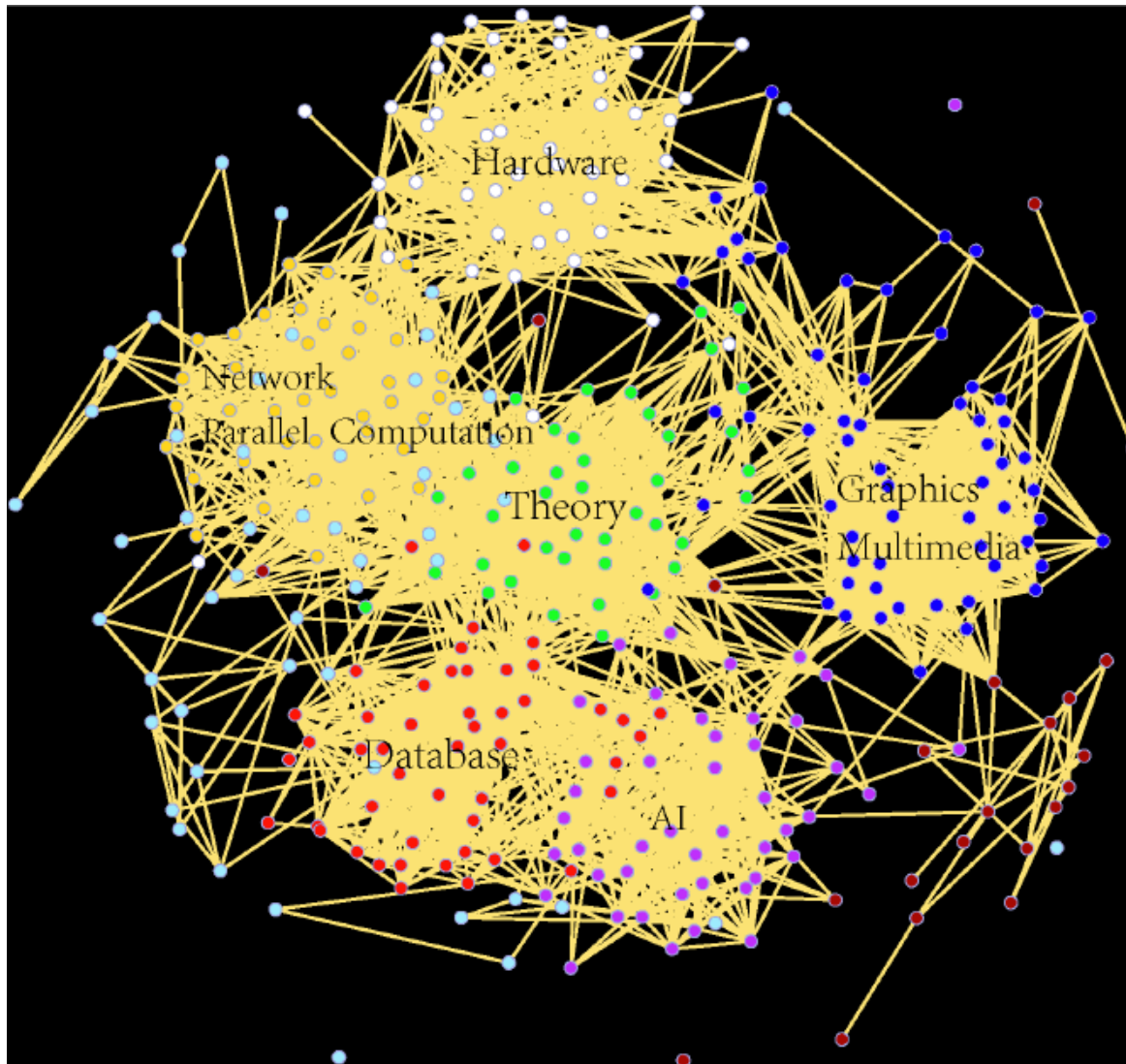
- Tag cloud: visualizing user-generated tags

- The importance of tag is represented by font size/color



Newsmap: Google News Stories in 2005


# Visualizing Social/Information Networks



Computer Science Conference Network

# 2: Data Pre-Processing

---

- Getting to know your data
  - Basic Statistical Descriptions of Data
  - Data Visualization
- Data Pre-Processing 
  - Data Cleaning
  - Data Integration
  - Data Reduction
  - Data Transformation and Data Discretization



# 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
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization

# 2: Data Pre-Processing

---

- Getting to know your data
  - Basic Statistical Descriptions of Data
  - Data Visualization
- Data Pre-Processing
  - Data Cleaning
  - Data Integration
  - Data Reduction
  - Data Transformation and Data Discretization



# 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?

# How to Handle Missing Data?

---

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- 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 mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

# How to Handle Noisy Data?

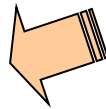
---

- **Binning**
  - 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
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)

# 2: Data Pre-Processing

---

- Getting to know your data
  - Basic Statistical Descriptions of Data
  - Data Visualization
- Data Pre-Processing
  - Data Cleaning
  - Data Integration
  - Data Reduction
  - Data Transformation and Data Discretization




# 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:**
  - 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

# 2: Data Pre-Processing

---

- Getting to know your data
  - Basic Statistical Descriptions of Data
  - Data Visualization
- Data Pre-Processing
  - Data Cleaning
  - Data Integration
  - Data Reduction 
  - Data Transformation and Data Discretization



# Data Reduction Strategies

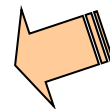
---

- **Data reduction:** Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - **Dimensionality reduction**, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - **Numerosity reduction** (some simply call it: **Data Reduction**)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - **Data compression**

# 2: Data Pre-Processing

---

- Getting to know your data
  - Basic Statistical Descriptions of Data
  - Data Visualization
- Data Pre-Processing
  - Data Cleaning
  - Data Integration
  - Data Reduction
  - Data Transformation and Data Discretization



# 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
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization

# Normalization

- **Min-max normalization:** to  $[\text{new\_min}_A, \text{new\_max}_A]$

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

- Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to  $\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$
- **Z-score normalization** ( $\mu$ : mean,  $\sigma$ : standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Ex. Let  $\mu = 54,000$ ,  $\sigma = 16,000$ . Then  $\frac{73,600 - 54,000}{16,000} = 1.225$
- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|) < 1$$

# Discretization

---

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

# Simple Discretization: Binning

---

- **Equal-width** (distance) partitioning
  - 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 is not handled well
- **Equal-depth** (frequency) partitioning
  - Divides the range into  $N$  intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

# 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

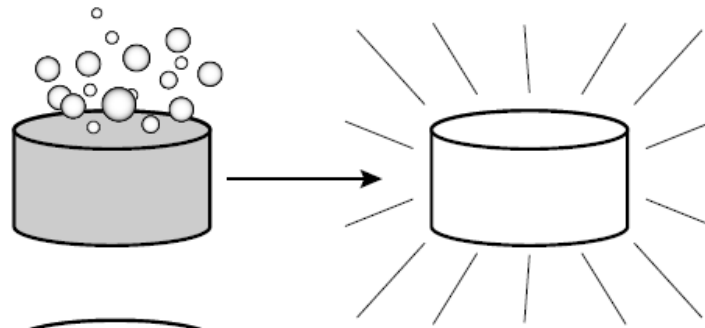
\* Smoothing by **bin means**:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

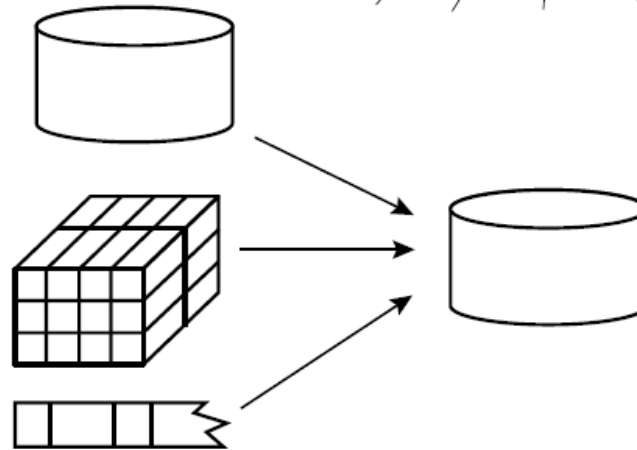
\* Smoothing by **bin boundaries**:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

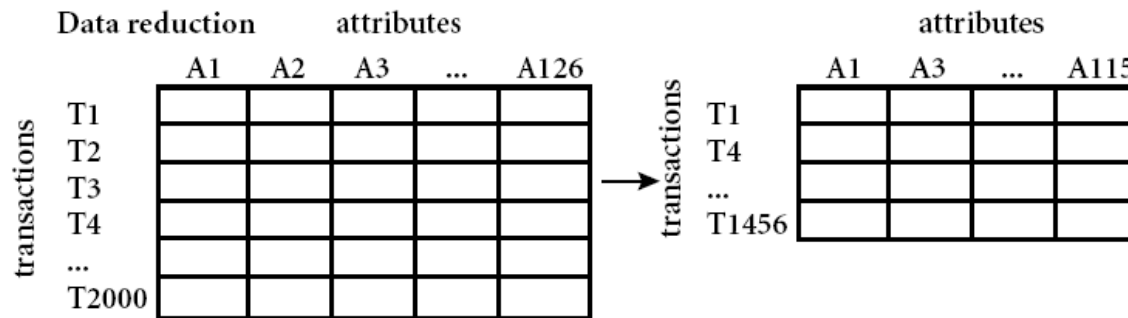
Data cleaning



Data integration



Data reduction



Data transformation

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



# References

---

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Comm. of ACM*, 42:73-78, 1999
- T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. [Mining Database Structure; Or, How to Build a Data Quality Browser](#). SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. *Bulletin of the Technical Committee on Data Engineering*, 20(4), Dec. 1997
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*. Vol.23, No.4
- V. Raman and J. Hellerstein. *Potters Wheel: An Interactive Framework for Data Cleaning and Transformation*, VLDB'2001
- T. Redman. *Data Quality: Management and Technology*. Bantam Books, 1992
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995