HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

WEB MINING

LECTURE 03: DATA VISUALIZATION

ONE LOVE. ONE FUTURE.

Content

1. Static chart
2. Pixel-based visualization
3. Visualization in vector space
4. Super sphere tree
5. SOM

1. Static chart/ 1.1 Attributes

- Data objects represent entities in the data (e.g. customers, products, transactions)
- Data objects are also known as samples, examples, or data points
- An attribute is a data field that represents a property or features of the data
- Attributes are also known as dimensions, features, or variables

- The values of a given attribute are called observations
- The set of attributes that describe a given object is called an attribute vector (or feature vector).
- The attribute type is determined by the set of attribute values

Nominal attribute

- Valuable as symbols or names
- eg: 'hair color' includes 'blue', 'red', 'black', 'white', 'platinum'
- Description of categories, codes, states
- Common value based on mode function

Binary attribute

- The category attribute has only two categories or two states
  - 0 ~ absent, 1 ~ exists
  - or 0 ~ false, 1 ~ true
- Symmetrical attributes (e.g. 'gender' includes 'male' and 'female')
- Asymmetric attributes (e.g. 'result' includes 'positive' and 'negative'

Ordinal attribute

- The values follow a certain order
- Example: 'size' includes 'small', 'normal', 'large' and 'oversized'
- Typical value based on mode and median function

Interval attribute

- Use to define values measured along a scale, with each point placed at an equal distance from one another.
- Can compare, calculate the distance between values
- For example: Temperature according to the Celsius scale.

Ratio atribute

- Numerical attribute with value 0
- Could multiply values
- E.g: Counts and measures:
  - Quantity
  - Weight
  - Height
  - Money
  - ...

Discrete and Continuous Attributes

- Has only a finite or countably infinite set of values.
- Example:
  - Finite: color, age
  - Countable infinite: Customer ID
- Attribute is continuous if not discrete

1.2 Basic data statistics

- Data Description:
  - Central value
  - Distribution range
  - Visualization based on charts
- Identify outliers

mean

- Values have the same role

$$x = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Values with different weights

$$x = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{n}$$

- The most common measurement, however sensitive to outliers

median

- The median divides the data into larger and smaller parts; These two parts have the same number of elements
- Calculate the median approximation
  - Group data into ranges of values
  - Calculate the frequency of values in each interval
  - Find the interval containing the median frequency

median

- Approximate median by formula:

$$median = L_1 + \left[\frac{N/2 - (\Sigma freq)_l}{freq_{median}} \right] width$$

Where:

- L_1: lower boundary of median
- N : number of value
- $(\Sigma freq)_l$ sum of the frequencies of intervals less than median
- $freq_{median}$: frequency of median range
- *width*: width of the median range

mode

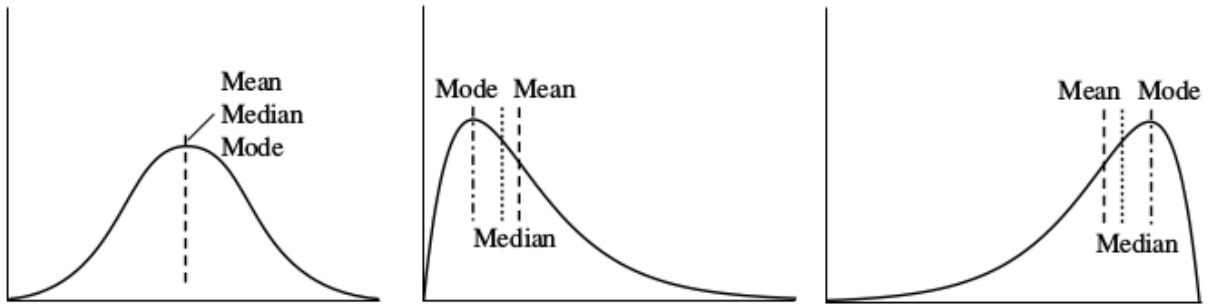
- The most frequency value in the dataset
- Multimodal: more than one mode
- Set containing only unique values without mode
- Với tập unimodal (one mode)

$$mean - mode \approx 3 \times (mean - median)$$

midrange

- Average of the maximum and minimum values

$$\text{midrange} = \frac{\max + \min}{2}$$



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

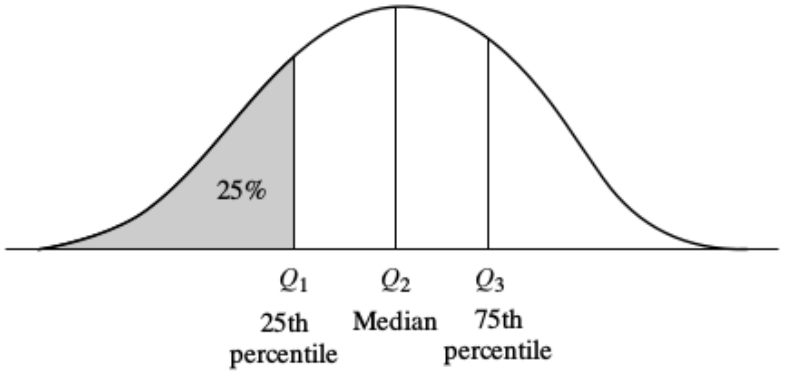
range

- distance between the largest and the smallest value in the set

$$range = max - min$$

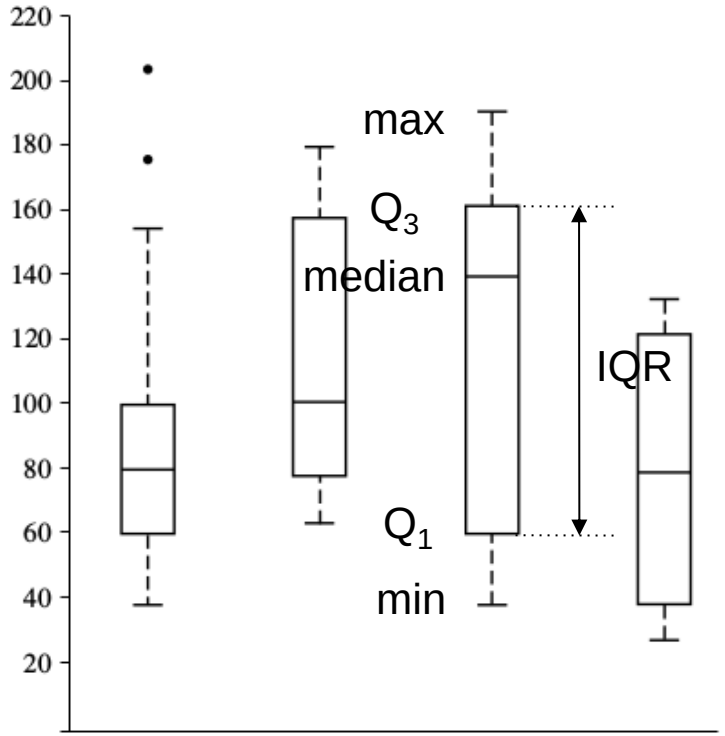
quantile

- Quantiles are points that divide data into (nearly) equal parts (with equal number of elements).
  - 2-quantile: a point that divides data into two equal parts ~ median
  - 4-quantile (quartile)
  - 100-quantile (percentile)
  - Interquartile range IQR = Q3 - Q1
- Interquartile range IQR = $Q_3 - Q_1$



boxplot

- The box chart includes:
  - Q_1, Q_3: the beginning and the end of the box
  - IQR: length of the box
  - Median
  - Min and max values



variance, standard deviation

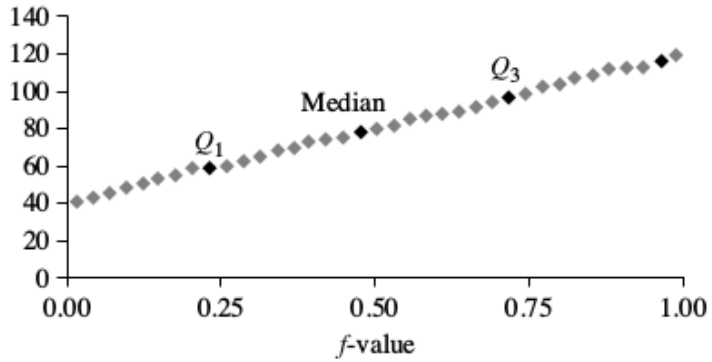
- Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2,$$

- σ: standard deviation represents the dispersion of the data relative to mean

Quantile chart

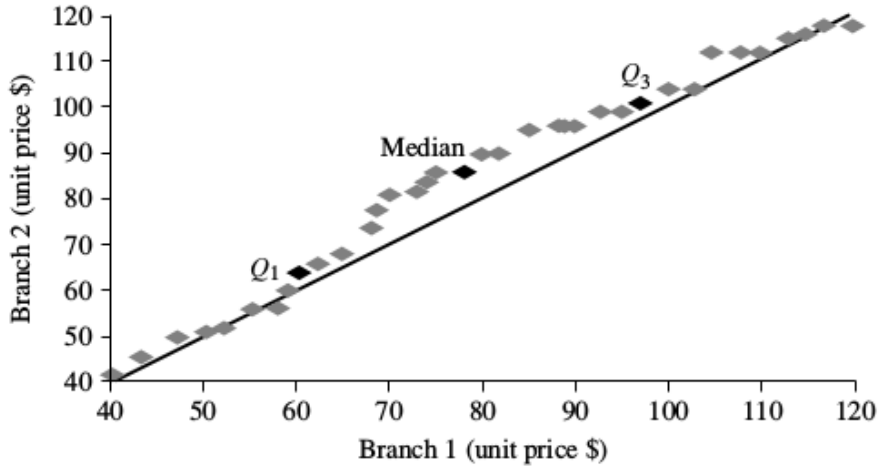
- Sort the values in ascending order $x_1 < x_2 < \ldots x_n$
- Frequency f_i corresponding to x_i là is the percentage of data with values below x_i



$$f_i = \frac{i - 0.5}{N}$$

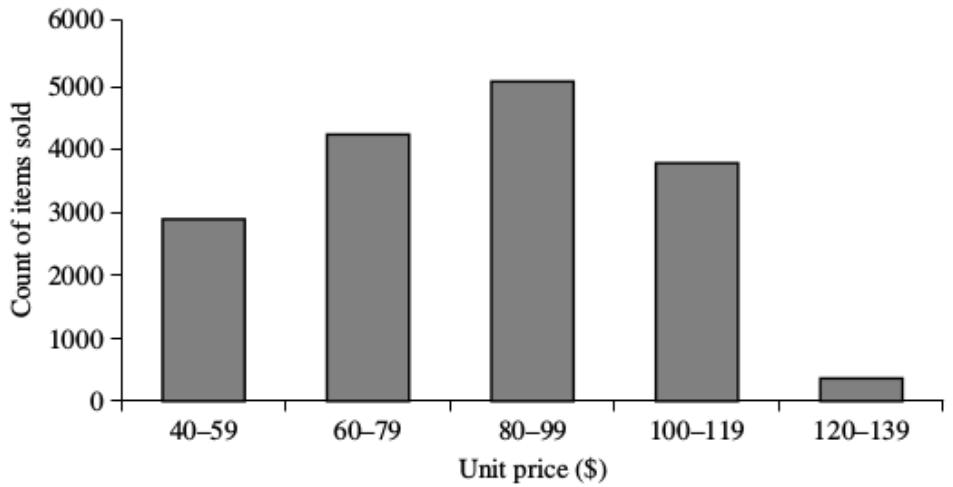
Quantile - quantile chart

- Show the relationship between quantile values of two univariate distributions



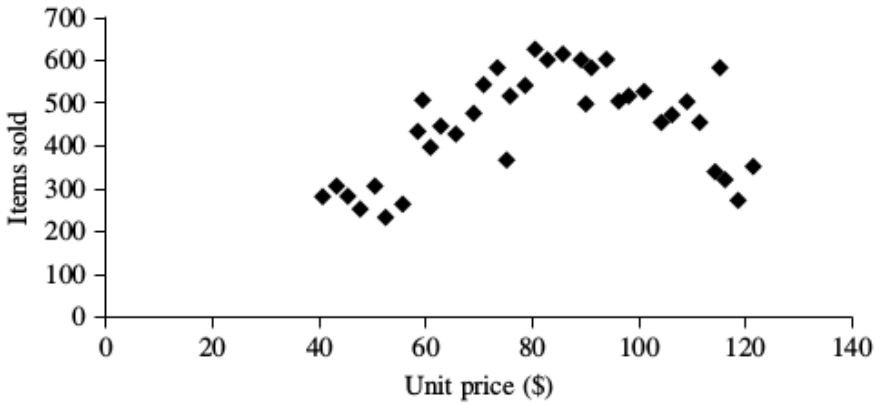
Histogram

- Values are grouped into equal intervals called buckets.

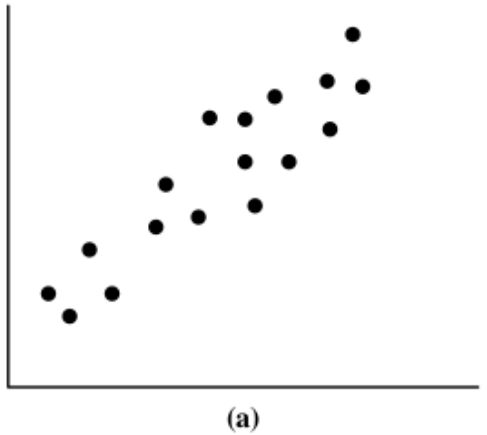


Scatter chart

- Determine the reciprocity between two numeric attributes

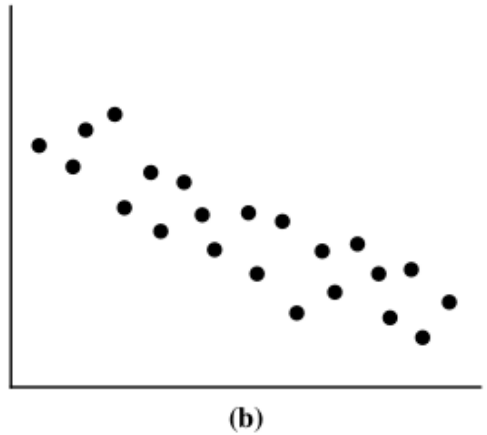


Scatter chart (cont.)



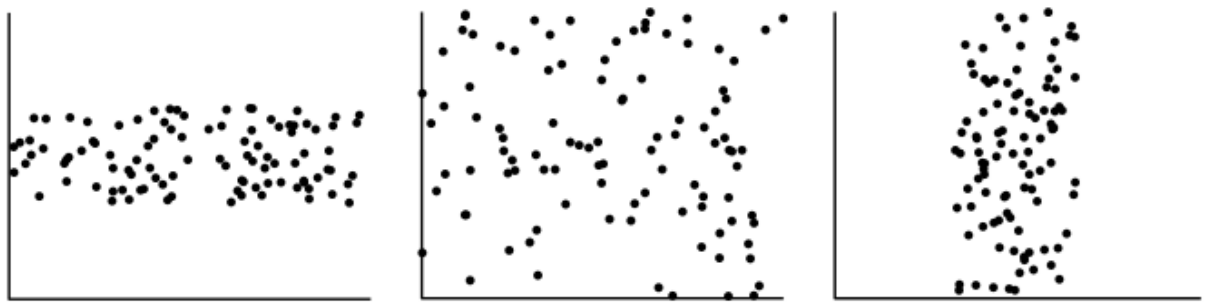
(a)

positive reciprocity



(b)

negative reciprocity

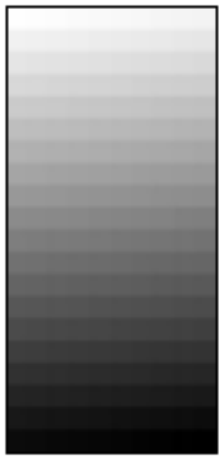


Null reciprocity

2. Pixel based visualization

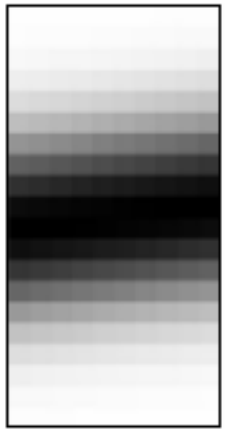
- The value of a data dimension represented by a pixel with the color corresponding to the value
- Example: Small value corresponds to light color, large value corresponds to dark color
- *m* dimension corresponds to *m* window. A data point with *m* dimensions is represented by *m* pixels at corresponding positions in each window

- Records are usually sorted in a dimension of interest
- Correlation, if any, between the data dimensions is expressed through the color distribution (data values) over the windows

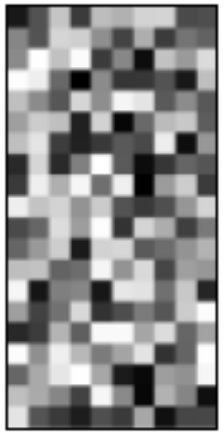


(a) *income*

(b) *credit_limit*



(c) *transaction_volume*



(d) *age*

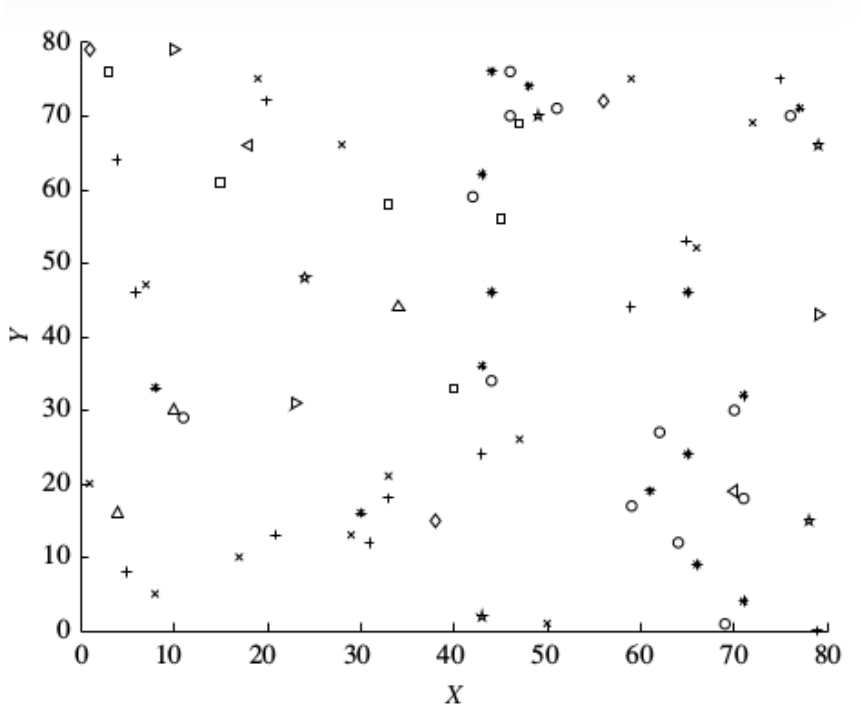
3. Visualization in vector space

- Pixel-based visualization does not represent the density of data points
- Visualization on vector space based on projection technique to represent multidimensional data in 2D space

Scatter Charts

- How to represent:
  - Two axes X and Y used to represent two-dimensional numbers in Cartesian coordinates
  - The third dimension is represented by different shapes
  - The fourth dimension can be represented by color

Scatter Chart (cont)

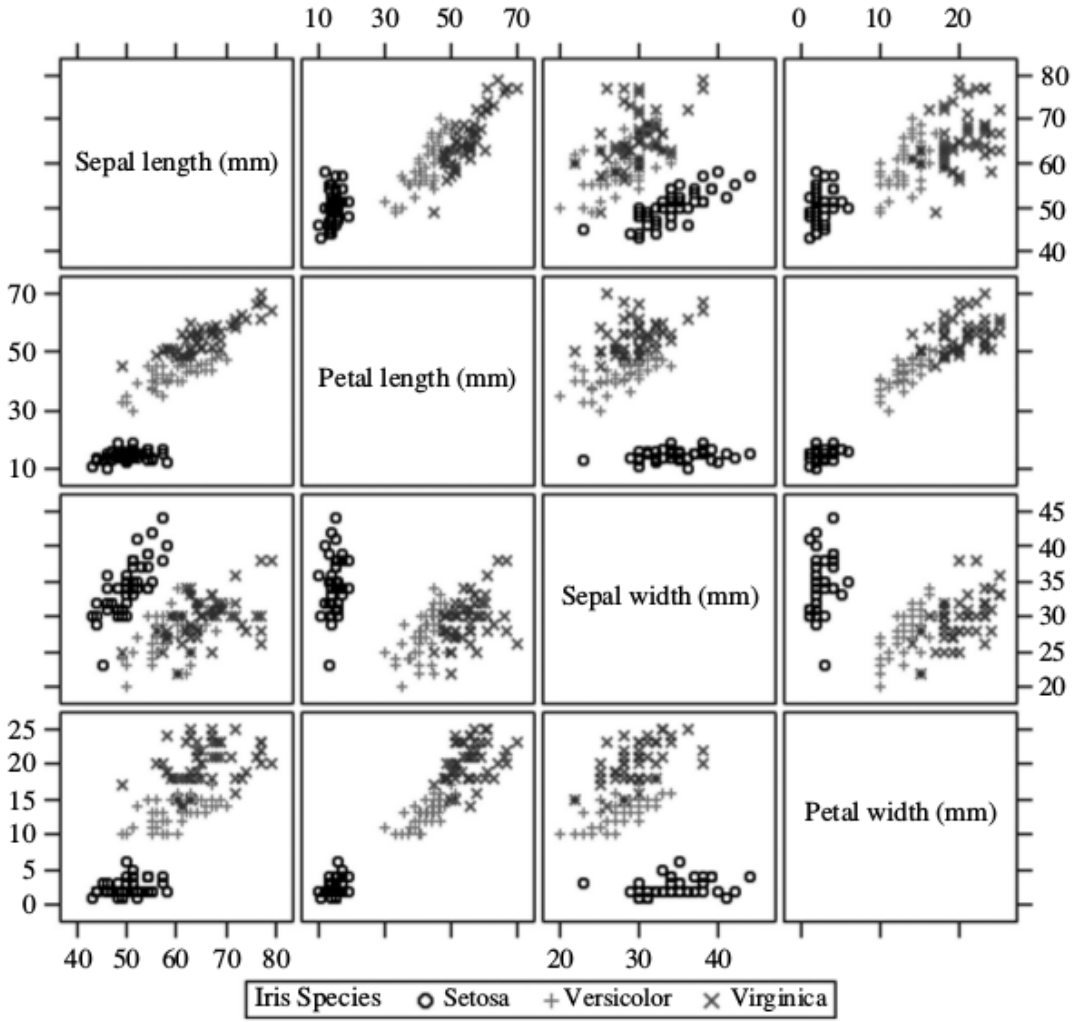


www.cs.sfu.ca/jpei/publications/rareevent-geoinformatica06.pd

Scatter chart matrix

- Scatter charts can only represent up to 4 dimensions
- For data with more than 4 dimensions, use the scatter plot method
  - Data has m dimensions
  - Use a 2D scatter plot m x m matrix to represent each data dimension with the remaining data dimensions
  - Example: The Iris dataset has 5 dimensions visualized by a 4 x 4 matrix consisting of 24 3D scatter plots

Example: *Iris* Dataset

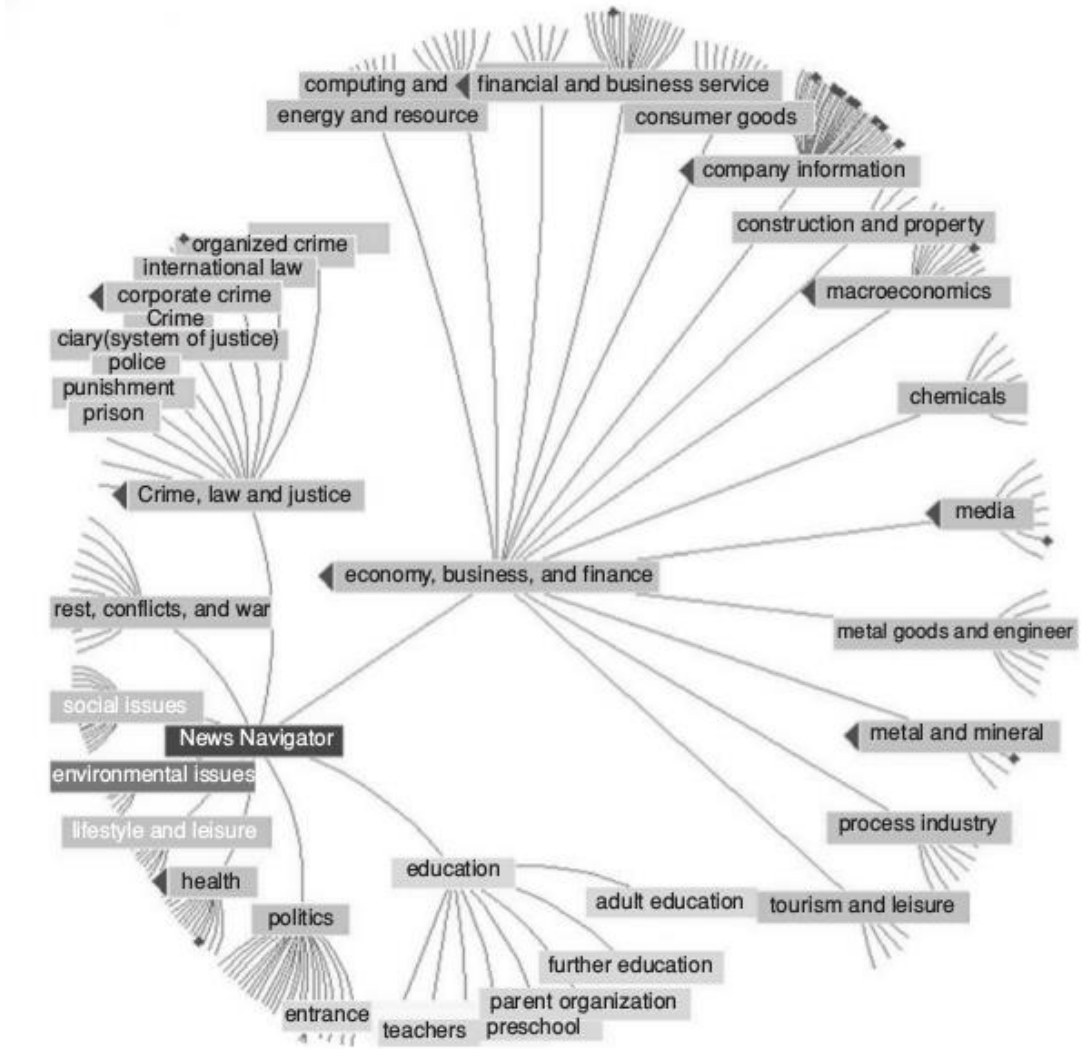


http://support.sas.com/documentation/cdl/en/grstatproc/61948/HTML/default/images/gsgscmat.gif

4. Super sphere tree

- Visualize large amounts of data
- Data has a tree structure
- focus on a part of the data, on the other hand still represent the general context of the data
- Fish eye properties:
  - The size of unfocused buttons rapidly decreases when
  - The size of the buttons are focused is rapidly increasing

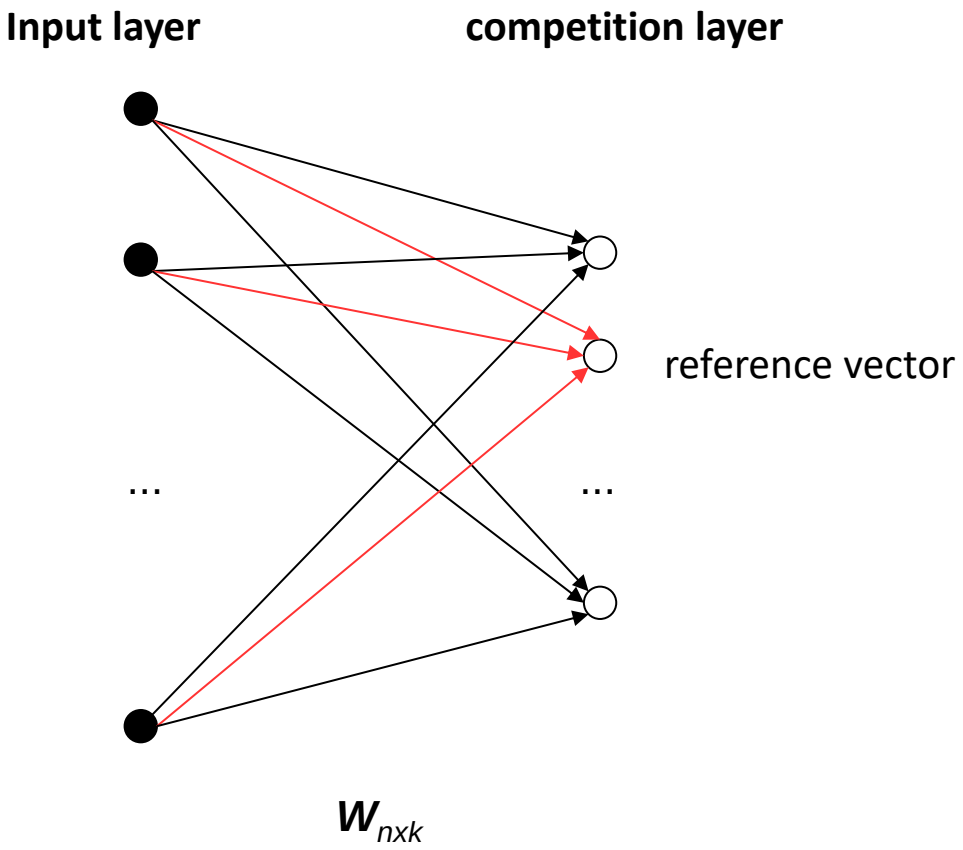
Example


computing and
financial and business service
energy and resource
consumer goods
company information
construction and property
organized crime
international law
corporate crime
Crime
macroeconomics
ciary(system of justice)
police
punishment
prison
chemicals
Crime, law and justice
media
economy, business, and finance
rest, conflicts, and war
metal goods and engineer
social issues
News Navigator
metal and mineral
environmental issues
process industry
lifestyle and leisure
education
health
adult education
tourism and leisure
politics
further education
parent organization
entrance
teachers
preschool

5. SOM (Self Organizing Map)

- SOM learns 2D representation of multidimensional data
- SOM is a feed-forward neural network (2 layers)
  - Input layer receives a signal from the input data, whose dimension is equal to the dimension of the data
  - Competition layer is organized in a certain shape (rectangle, hexagon, etc.) showing the spatial relationship between neurons.
  - Each neuron in competition layer has association weights from the input layer called a reference vector

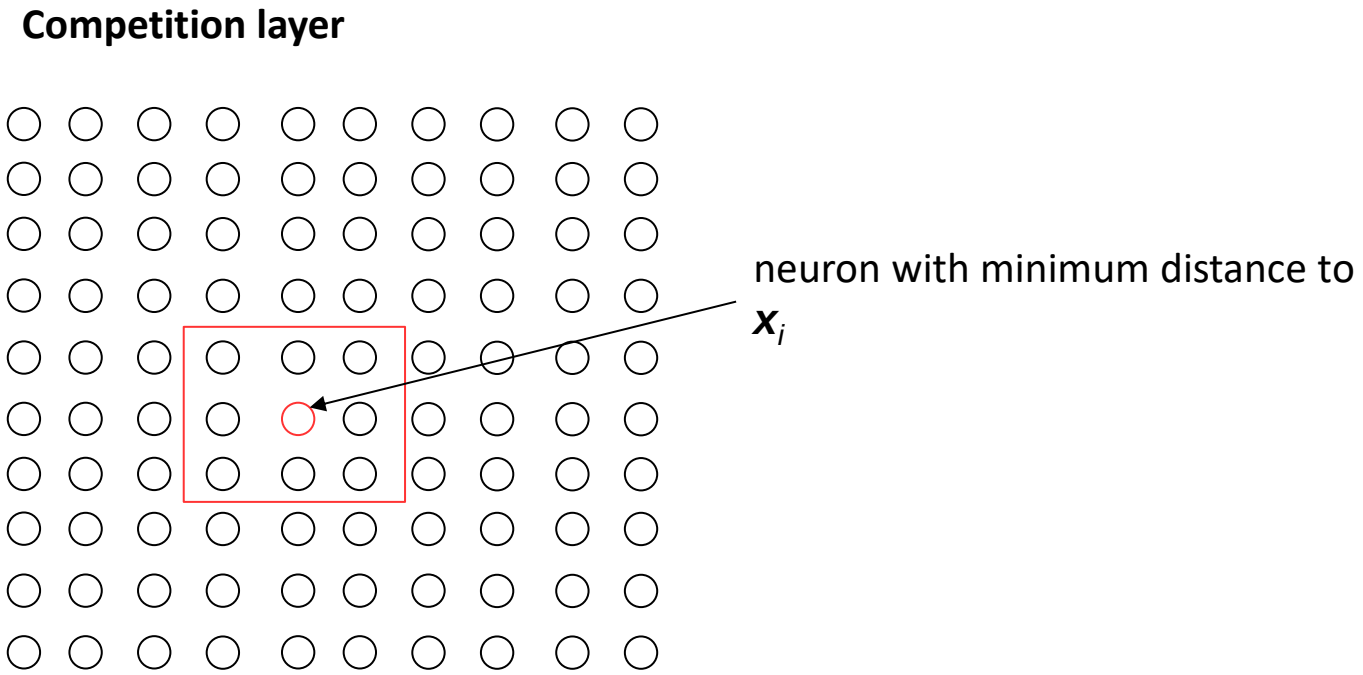
SOM architecture



Competitive learning

- For each input $\boldsymbol{x}_i$, select neuron m_k whose distance to $\boldsymbol{x}_i$ is the smallest
- Distance between $\boldsymbol{x}_i$ and m_k: euclide distance between x_i and reference vector of m_k
- Objective function: Minimize total distance between input and corresponding nearest neuron
- Weight Update: Only update weights of m_k and neighboring neurons

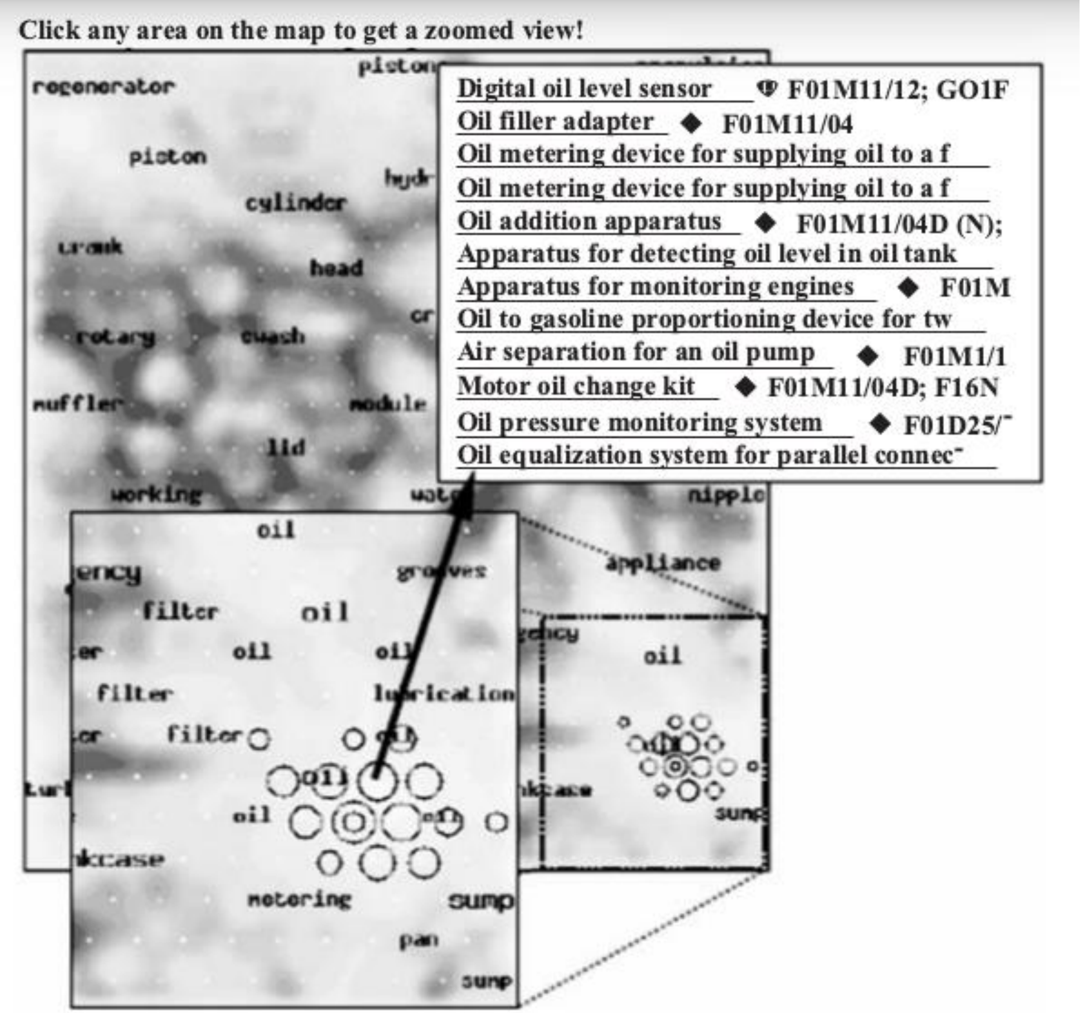
Example: neighboring neurons



WEBSOM

- Representation of a set of documents on a 2D map
- Texts are represented as bags of words
- After learning, each text group can be represented by specific keywords
- Areas with high density are places where a lot of text is concentrated

WEBSOM


Click any area on the map to get a zoomed view!
regenerator
piston
cylinder
crank
head
rotary
muffler
module
lid
working
oil
filter
lubrication
metering
sump
pan
appliance
nipple
Digital oil level sensor F01M11/12; GO1F
Oil filler adapter F01M11/04
Oil metering device for supplying oil to a f
Oil metering device for supplying oil to a f
Oil addition apparatus F01M11/04D (N);
Apparatus for detecting oil level in oil tank
Apparatus for monitoring engines F01M
Oil to gasoline proportioning device for tw
Air separation for an oil pump F01M1/1
Motor oil change kit F01M11/04D; F16N
Oil pressure monitoring system F01D25/
Oil equalization system for parallel connec

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !