CS 521
Data Mining Techniques
Instructor: Abdullah Mueen

LECTURE 1: OVERVIEWS, DATA TYPES AND SIMILARITY

John Snow and the Broad St. Pump

John Snow (15 March 1813 – 16 June 1858) was an **English physician** and a leader in the adoption of anaesthesia and medical hygiene. He is considered **one of the fathers of modern epidemiology**, in part because of his work in tracing the source of a cholera outbreak in Soho, London, in 1854.

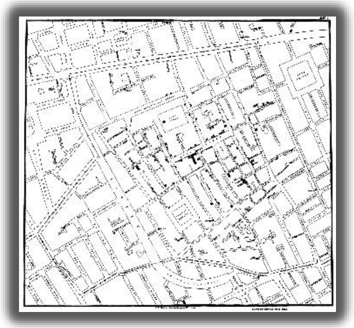
On 31 August 1854, after several other outbreaks had occurred elsewhere in the city, a major outbreak of cholera struck Soho. Over the next three days, 127 people on or near Broad Street died. In the next week, three quarters of the residents had fled the area. By 10 September, 500 people had died and the mortality rate was 12.8 percent in some parts of the city. By the end of the outbreak, 616 people had died.

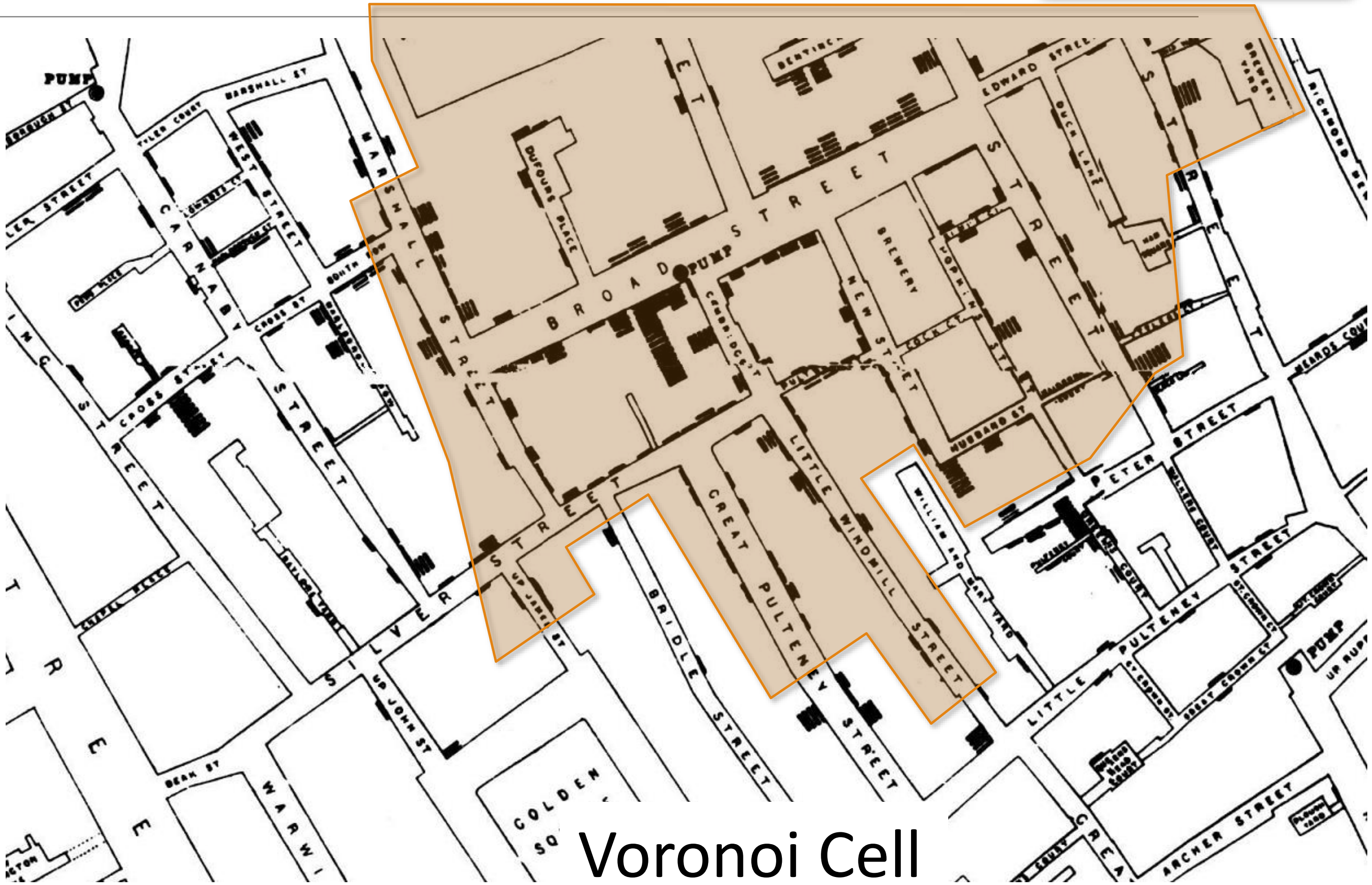
He identified the source of the outbreak. It was the public water pump on Broad Street

John Snow and the Broad St. Pump

Location of each death in the outbreak and locations of the pumps with the help of Rev. Henry Whitehead

Associate pumps with deaths to support the causal relationship





Components of Data Mining

Data (Images, Files, Tables, Charts)

Tools (Hadoop, Matlab, Algorithms)

Objective (Information integration, organization and scientific discovery)

Data Scientist

Data Domains

Web Sensing

Individual Sensing

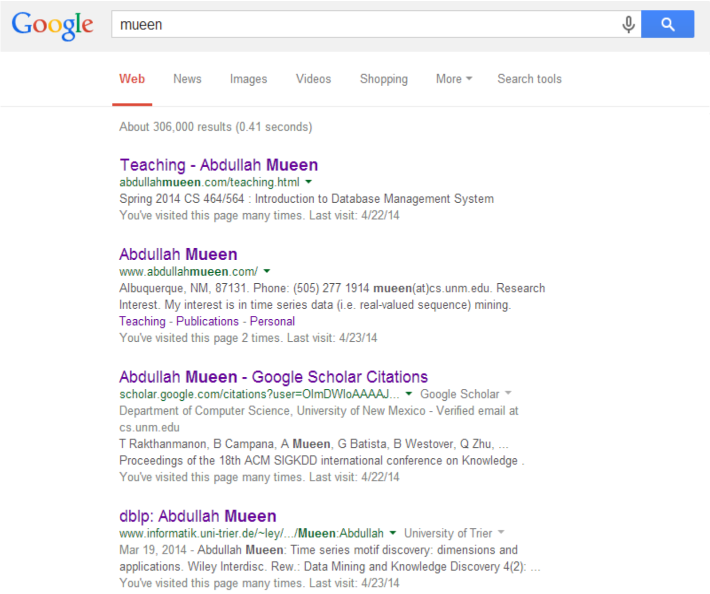
Data:

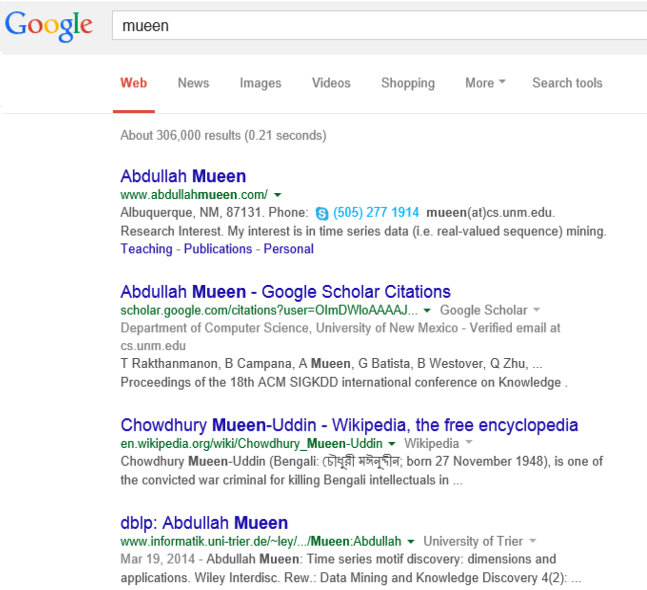
1. Search Query Logs: Mostly Tabular. Query, IP address/Account, Time, Link Clicked
2. Action Sequence: Every Click you make is being recorded across devices
3. Key Sequence: Text, Reviews, Comments, Survey, Instant messaging
4. Voice/Video Data: Video Conferencing
5. Spatio-temporal Data: Check-in Services

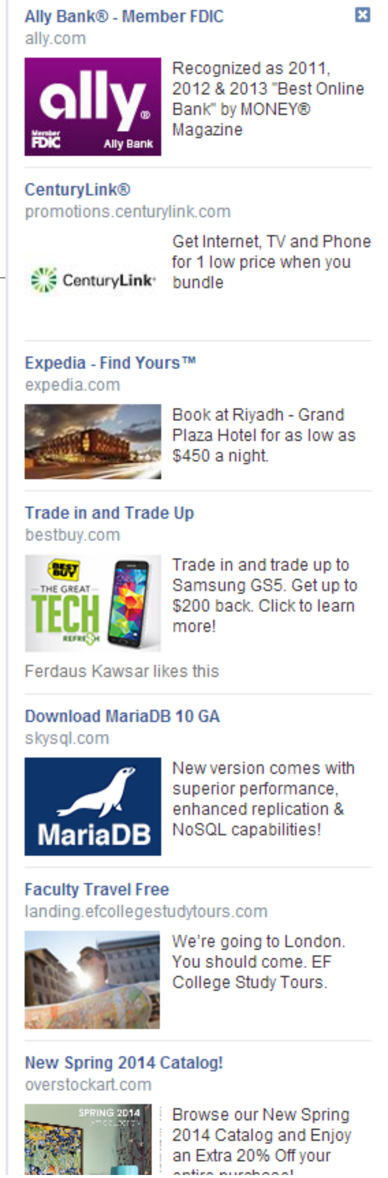
Web Sensing

Applications Targeted to Individuals

1. [Targeted advertisement]
2. Personalized Search Results







Web Sensing

Social/Community Sensing

Data:

Networks: Friend Net, Call Net, Follower Net,

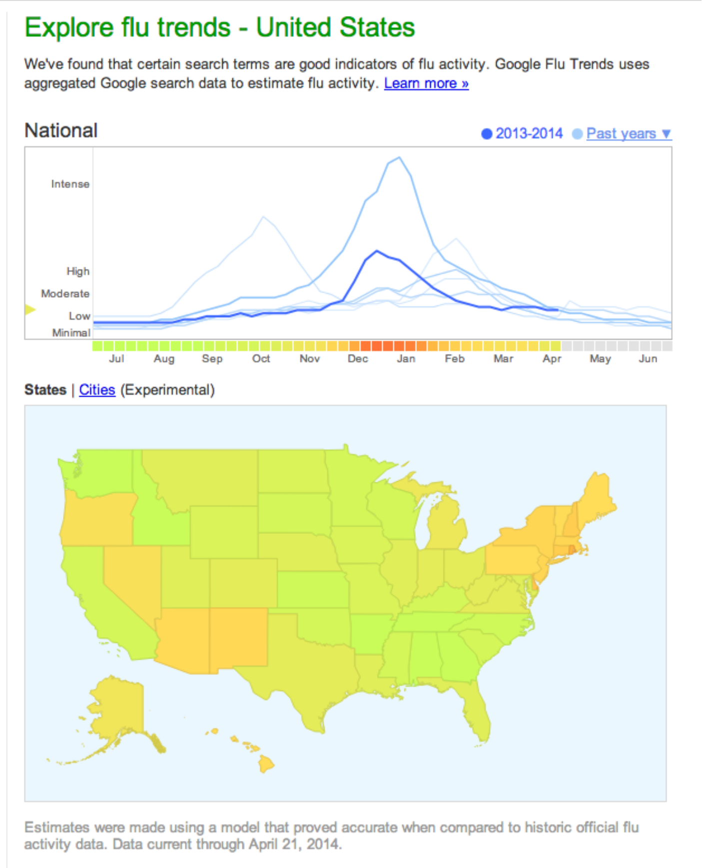
Text: News, Reviews, Comments, Tweets

Census Data

Applications:

[Flue Trends]

[BoxOffice Prediction]



Business

Stock market

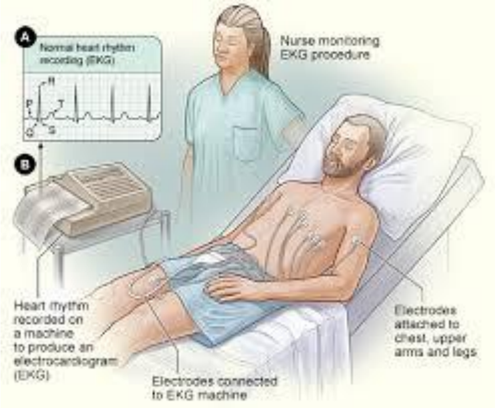
Banks

Insurance…

Health and Medicine

Patient Records (Clinical, Pathological etc.)

Sequencing Data…



[Success Stories in Data/Text Mining] by Christophe Giraud-Carrier

Remote Sensing

From Earth to the Outer Space

From Space to the Earth

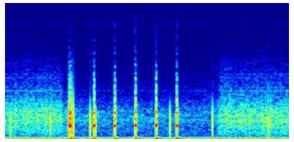
Data:

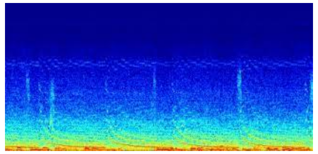
Images and spectrograms

Derived Data:

Vegetation Index

Sea-surface Height





Remote Sensing

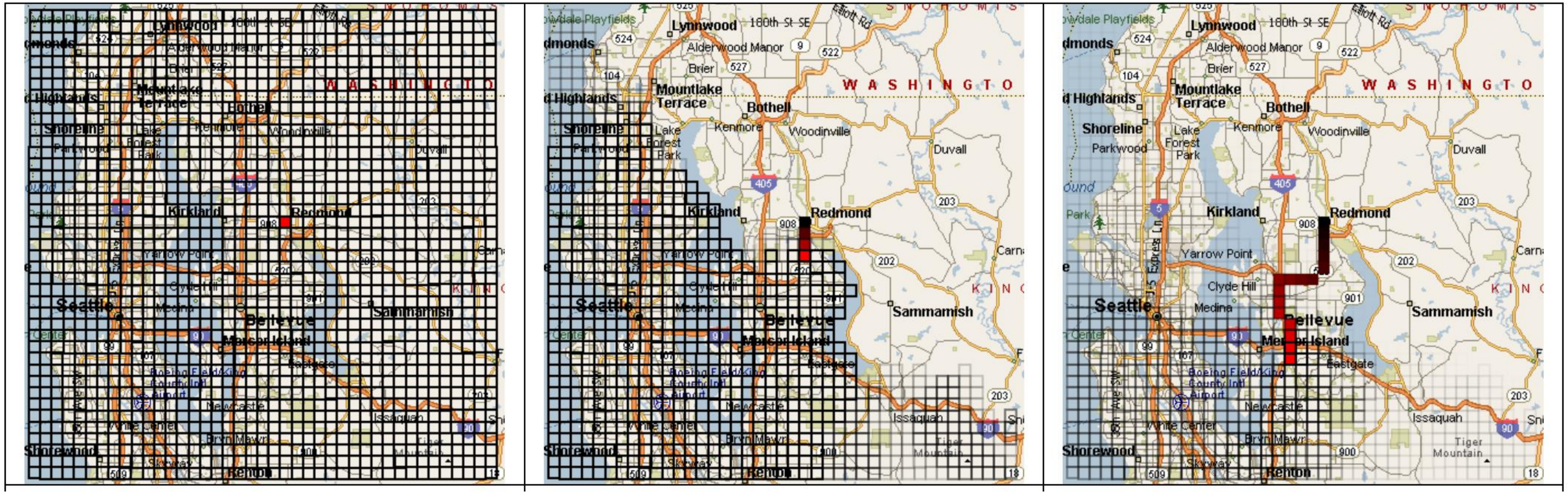
Applications in Space Exploration

1. Detecting, Tracking, categorizing asteroids
   - TopCoder Contest
2. Categorizing stars based on types and their remaining life using light curves

Applications in Observing Earth

1. Modeling and Validating Climate Changes
2. Predicting storm formation
3. Detecting forest fire, deep ocean eddies, air pollution, etc. [Expedition]

Movement Sensing



Data: GPS Traces of Human and Animals, Maps

Applications

1. Traffic based route planning
2. Destination Prediction
3. Opportunistic Crowdsourcing

Government Data

Data:

Transportation Data

Environmental Data

Utility Data

Police Data

http://www.cabq.gov/abq-data

Applications:

Smart City Applications

Energy Efficient Building, Transportation etc.

Anthropology

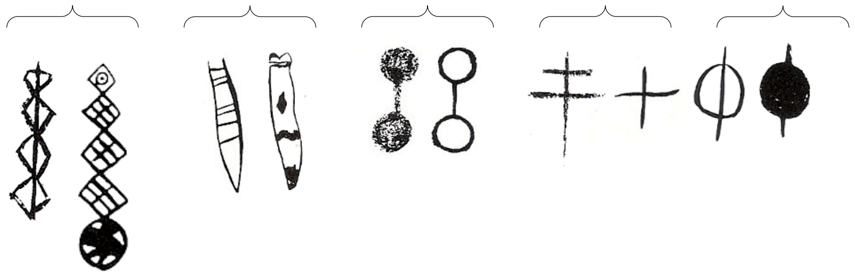
Data:

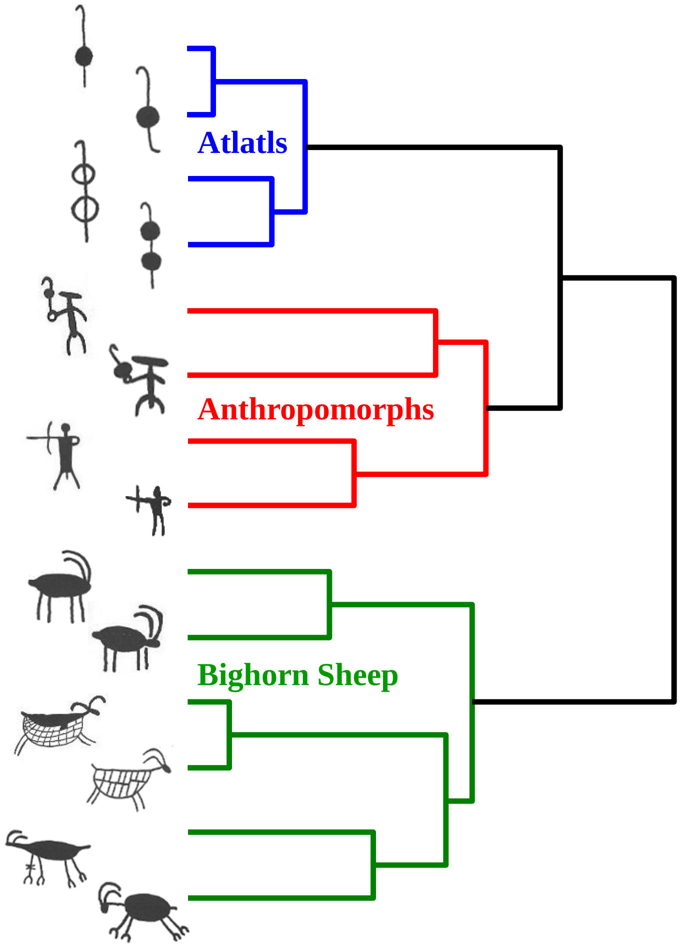
Images and Shapes of the Petroglyphs and Petrographs

Applications:

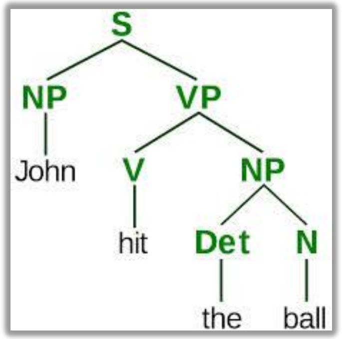
Clustering Petroglyphs

Finding repeated Petroglyphs across states or countries





Linguistics



Data:

Text Data: Books and News

Audio: Audio Corpus

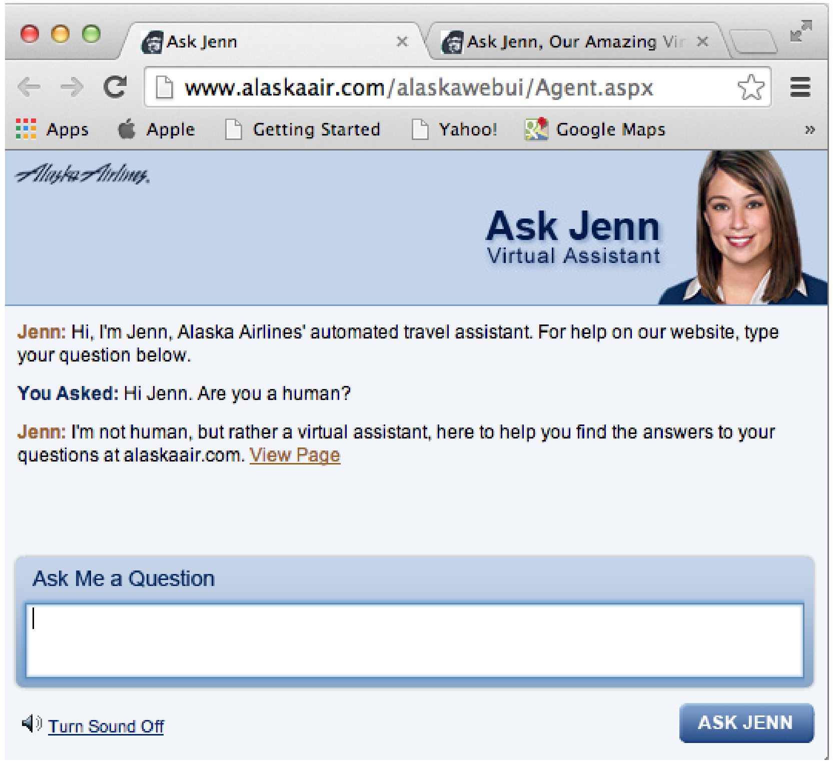
Applications

Machine Translation

Dialogue Processing

NLP for assistive technologies

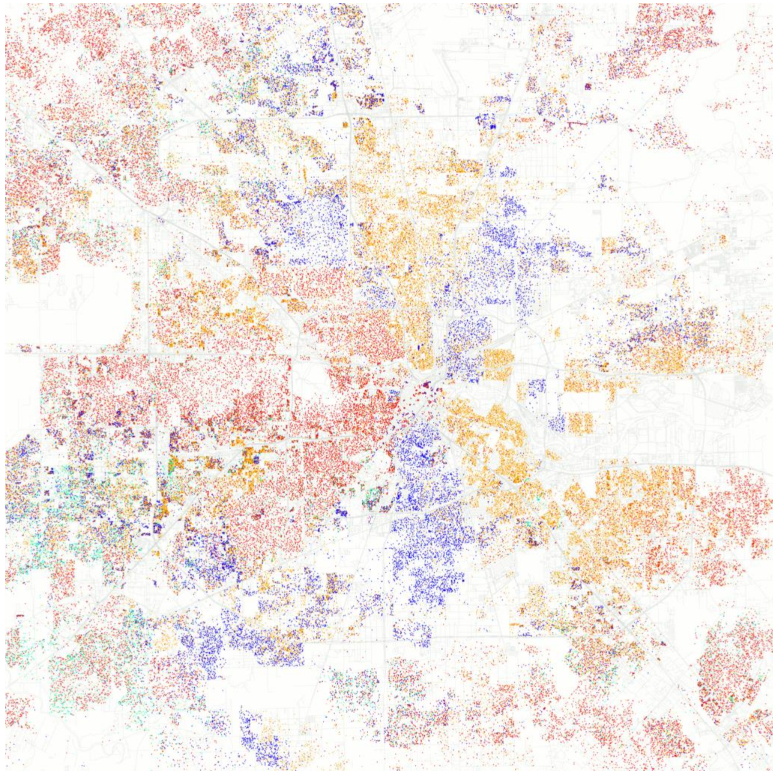
IBM Watson



Data Mining Algorithms

Clustering

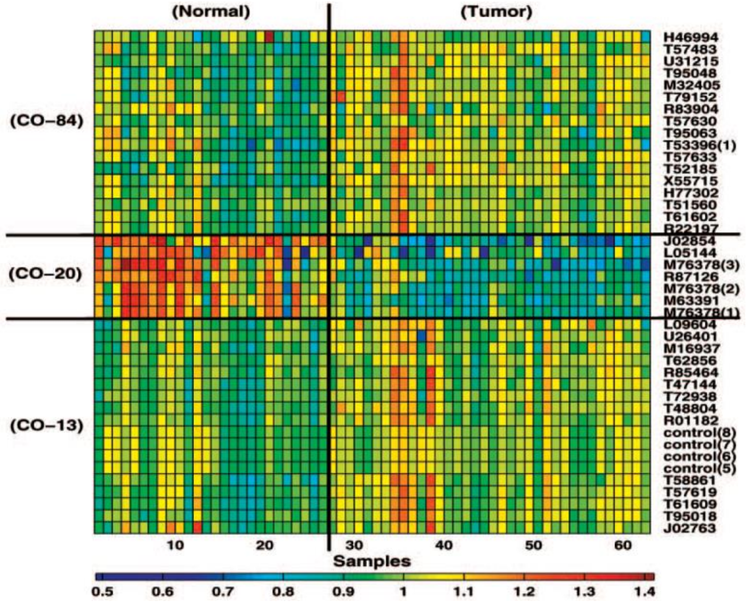
- Divide the data in meaningful partitions
- Needs a goodness measure
- Tool: [Weka](), Matlab

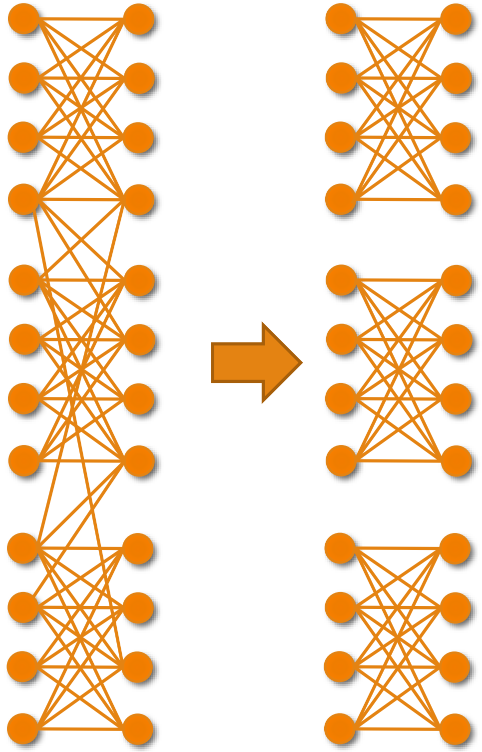


Houston, Ethnic Distribution

Graph Clustering

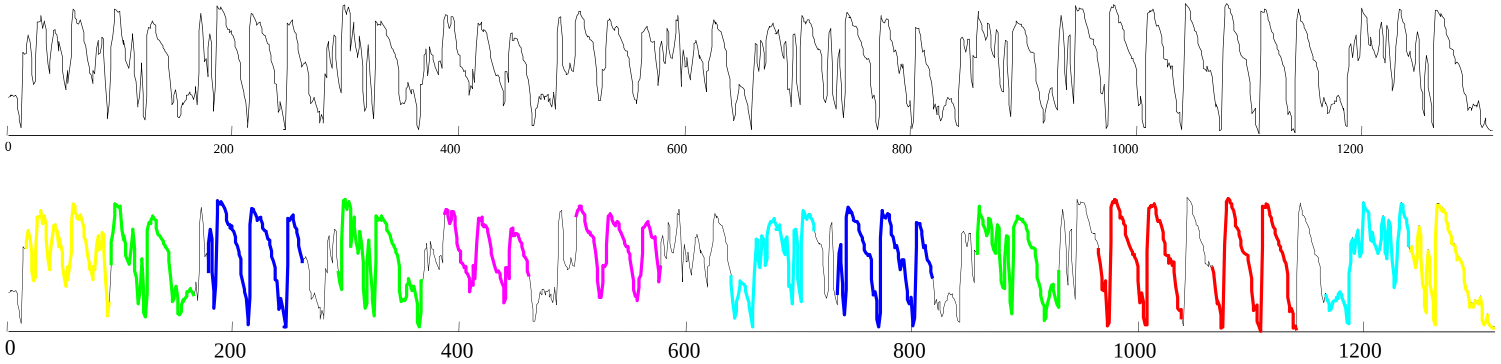
- Neighborhood based similarity
- Co-Clustering is a way to find the heavily connected components of a bipartite graph.
- Tool: [cocluster]





Co-clustering

Signal Clustering



- Clusters the subsequences of the signal
- Ignores unnecessary segments
- Tool: [Epenthesis]

== Poem (original order)==

In a sort of Runic rhyme,
To the throbbing of the bells--
Of the bells, bells, bells,
To the sobbing of the bells;
Keeping time, time, time,
As he knells, knells, knells,
In a happy Runic rhyme,
To the rolling of the bells,--
Of the bells, bells, bells--
To the tolling of the bells,
Of the bells, bells, bells, bells,
Bells, bells, bells,--
To the moaning and the
groaning of the bells.

==Poem (grouped by clusters)==

bells, bells, bells,
Bells, bells, bells,

Of the bells, bells, bells,
Of the bells, bells, bells--

To the throbbing of the bells--
To the sobbing of the bells;
To the tolling of the bells,

To the rolling of the bells,--
To the moaning and the groan-

time, time, time,
knells, knells, knells,

sort of Runic rhyme,
groaning of the bells.

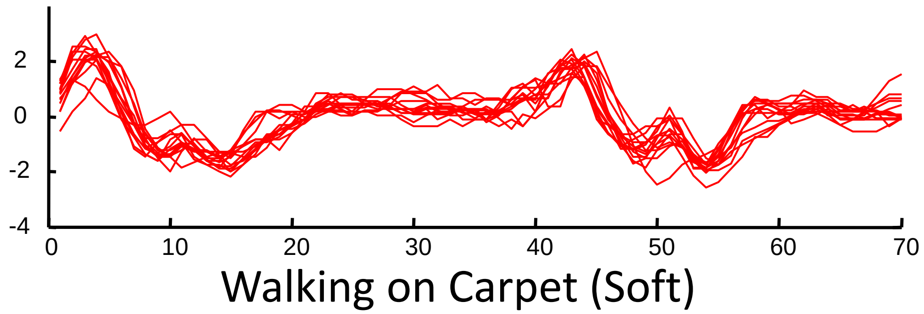
Image Clustering

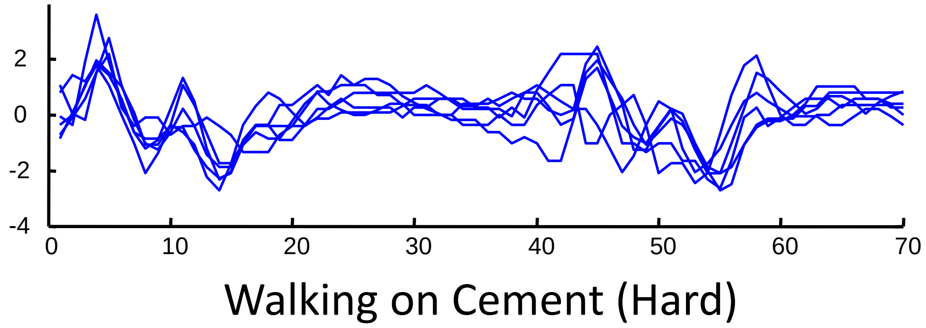
- Clustering based on color, texture, background etc.
- Ranges from small scale to web scale.

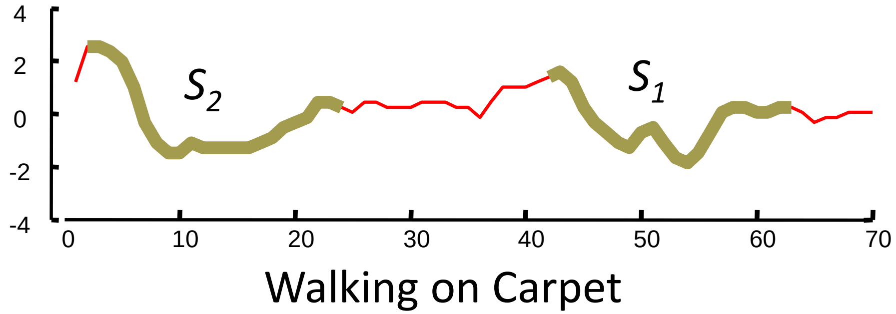
http://www.ulrichpaquet.com/current.html

http://groups.csail.mit.edu/vision/TinyImages/

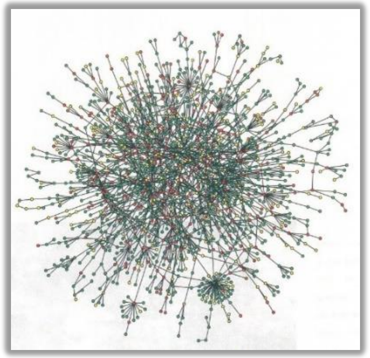
Classification





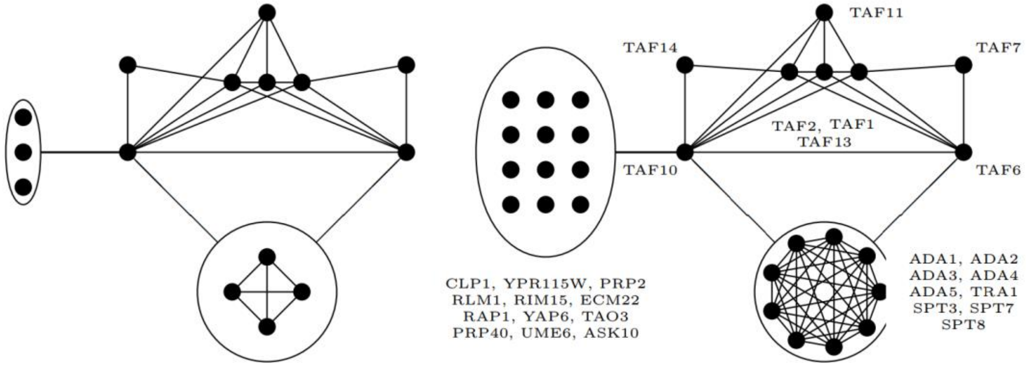


- Intuitive pattern for classification
- Very fast testing
- Tool: [Shapelet]



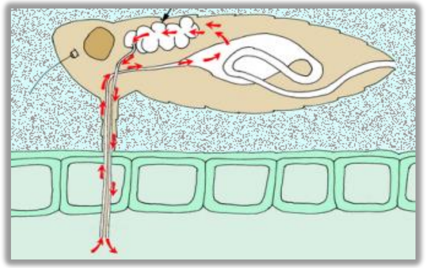
Repetition Detection: Graph

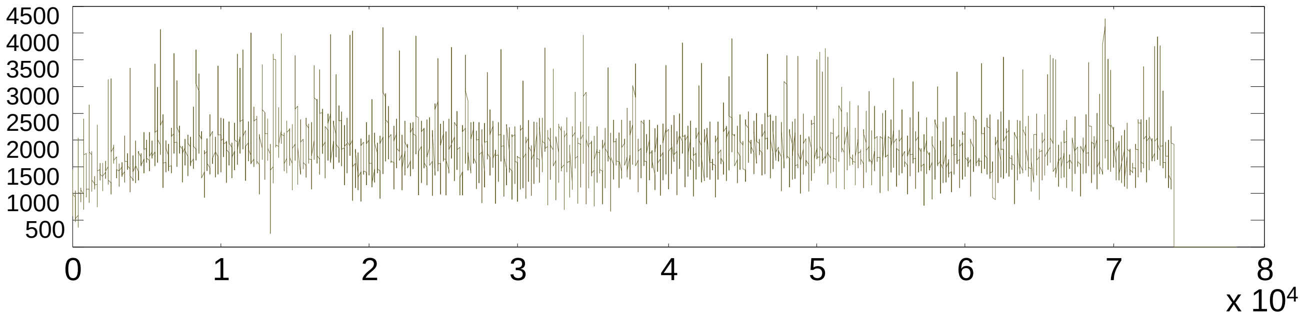
- Frequent Subgraph Mining
- Various Constraints on the Subgraph
- Tool: [gSpan]()

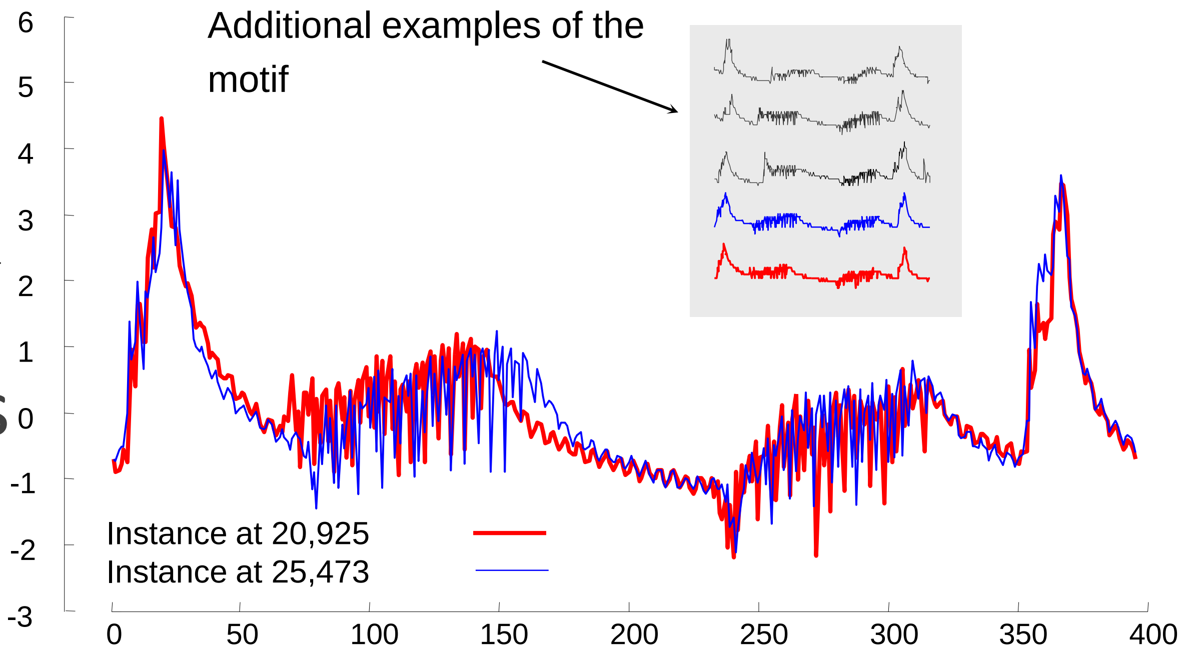


[Reference]()

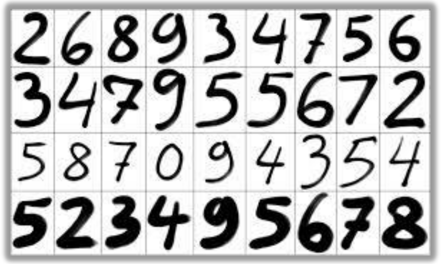
Repetition Detection: Signal





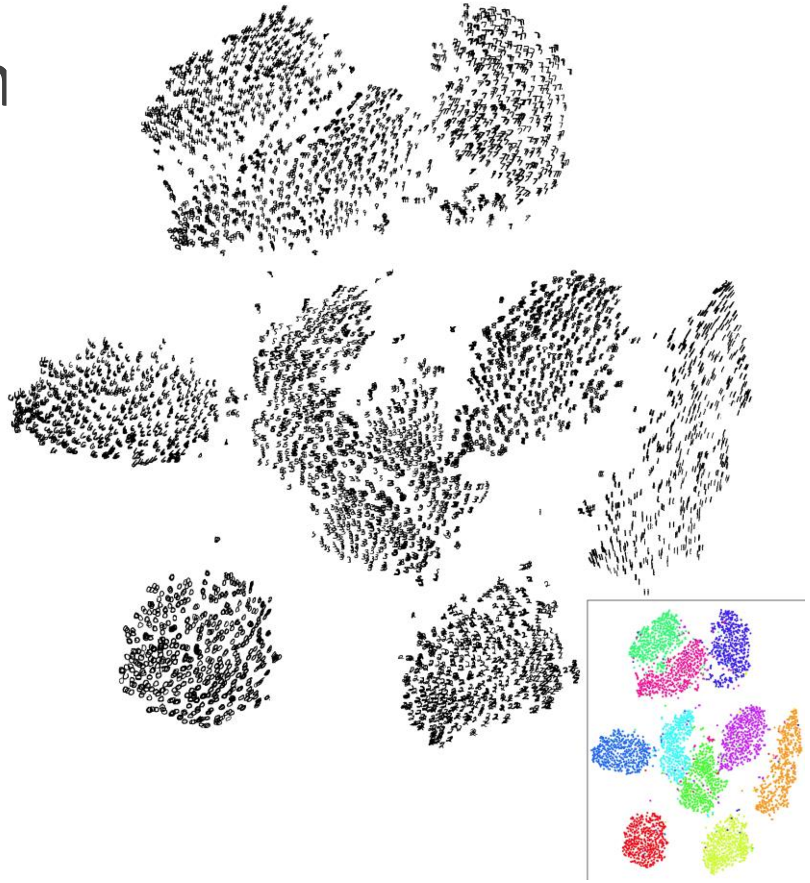


- Motif Discovery in Time Series
- Parameter-free method
- Tool: [MOEN]

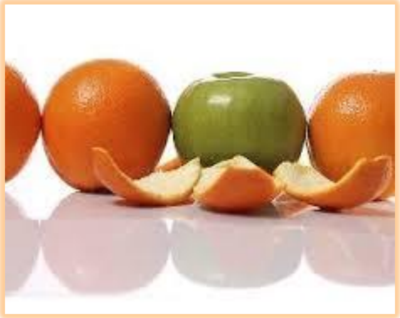


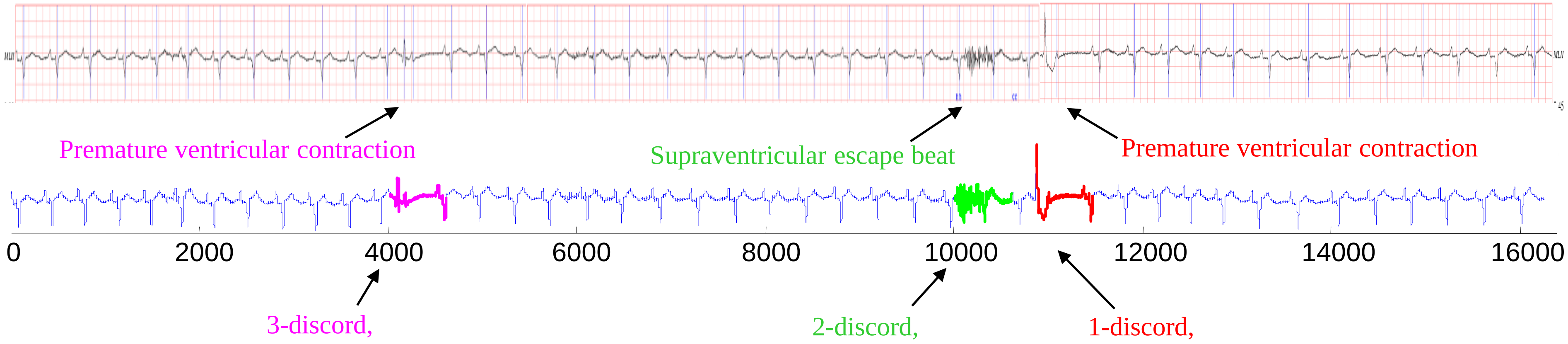
Visualization

- High Dimensional Data Visualization
- 2D and 3D
- Preserving Neighborhood of the points
- Tool: [t-SNE]

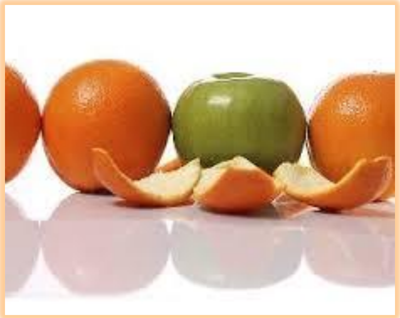


Anomaly Detection: Signal

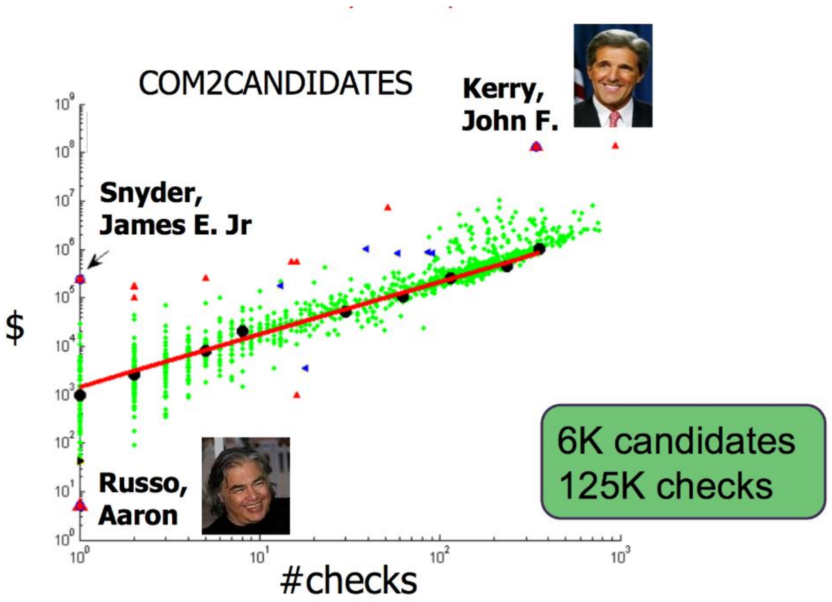


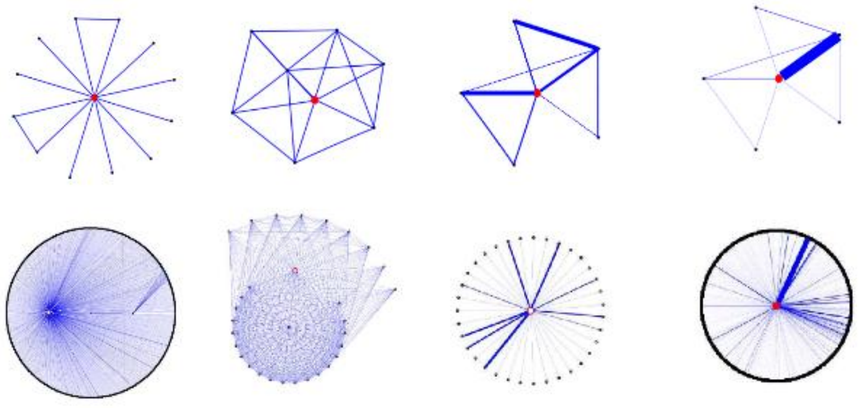


- Most unusual pattern in the signal
- Works in two passes
- Tool: [Discord]



Anomaly Detection: Graph



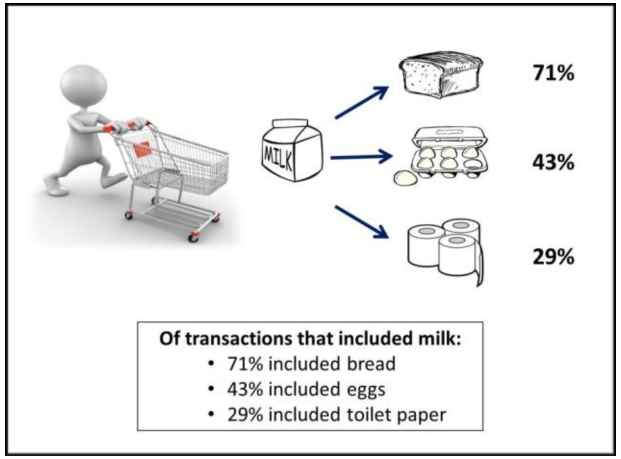


(a) Near-star (b) Near-clique (c) Heavy vicinity (d) Dominant edge

- Neighborhood based features
- Finds extremes in both direction
- Tool: OddBall

Association Detection

- Finds association among items with high support and confidence
- The algorithms are mostly exponential
- Tool: [SPSS Modeler](), Weka



No.	Association Rule	Support	Confidence
1	{Vaginal ultrasound; Surgical pathology; Pregnancy test; Hematology; Induced abortion; Penicillin injection} $\Longrightarrow$ {Legally induced abortion}	173	99.42%
2	{Pulmonary bronchospasm evaluation; Pulmonary vital capacity test; Non-pressurized inhalation treatment for acute airway obstruction; Doctor's office visit } $\Longrightarrow$ {Asthma}	56	91.80%
3	{Debridement of nails, manual, five or less; Debridement of nails, each additional, five or less; Intestine excision: Enteroenterostomy, anastomosis of intestine with or without cutaneous enterostomy; Transurethral surgery (Urethra and bladder)} $\Longrightarrow$ {Dermatophytosis}	619	91.43%

[Reference]()

Data Types

Getting to Know Your Data

Data Objects and Attribute Types

Basic Statistical Descriptions of Data

Data Visualization

Measuring Data Similarity and Dissimilarity

Summary

Types of Data Sets

Record

- Relational records
- Data matrix, e.g., numerical matrix, crosstabs
- Document data: text documents: term-frequency vector
- Transaction data

Graph and network

- World Wide Web
- Social or information networks
- Molecular Structures

Ordered

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data

Spatial, image and multimedia:

- Spatial data: maps
- Image data:
- Video data:

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Important Characteristics of Structured Data

Dimensionality

- Curse of dimensionality

Sparsity

- Only presence counts

Resolution

- Patterns depend on the scale

Distribution

- Centrality and dispersion

Data Objects

Data sets are made up of data objects.

A **data object** represents an entity.

Examples:

- sales database: customers, store items, sales
- medical database: patients, treatments
- university database: students, professors, courses

Also called *samples , examples, instances, data points, objects, tuples*.

Data objects are described by **attributes**.

Database rows -> data objects; columns ->attributes.

Attributes

Attribute (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.

- *E.g., customer _ID, name, address*

Types:

- Nominal
- Binary
- Ordinal
- Numeric: quantitative
  - Interval-scaled
  - Ratio-scaled

Attribute Types

Nominal: categories, states, or "names of things"

- *Hair_color = {auburn, black, blond, brown, grey, red, white}*
- marital status, occupation, ID numbers, zip codes

Binary

- Nominal attribute with only 2 states (0 and 1)
- <u>Symmetric binary</u>: both outcomes equally important
  - e.g., gender
- <u>Asymmetric binary</u>: outcomes not equally important.
  - e.g., medical test (positive vs. negative)
  - Convention: assign 1 to most important outcome (e.g., HIV positive)

Ordinal

- Values have a meaningful order (ranking) but magnitude between successive values is not known.
- *Size = {small, medium, large}*, grades, army rankings

Numeric Attribute Types

Quantity (integer or real-valued)

Interval

- Measured on a scale of **equal-sized units**
- Values have order
  - E.g., *temperature in C˚or F˚, calendar dates*
- No true zero-point

Ratio

- Inherent **zero-point**
- We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
  - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

Discrete Attribute

- Has only a finite or countably infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

Continuous Attribute

- Has real numbers as attribute values
  - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables

Getting to Know Your Data

Data Objects and Attribute Types

Basic Statistical Descriptions of Data

Data Visualization

Measuring Data Similarity and Dissimilarity

Summary

Basic Statistical Descriptions of Data

Motivation

- To better understand the data: central tendency, variation and spread

Data dispersion characteristics

- median, max, min, quantiles, outliers, variance, etc.

Measuring the Central Tendency

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

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

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

$$\mu = \frac{\sum x}{N}$$

- Weighted arithmetic mean:

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

- 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*):

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

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Median interval → 21–50

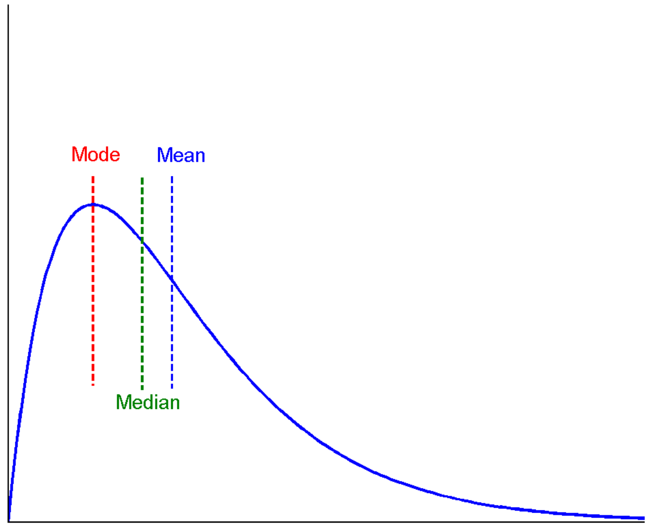
Mode

- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula:

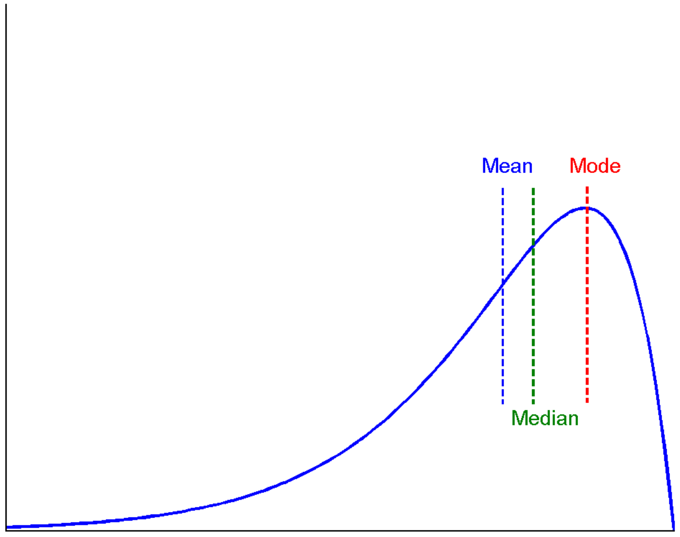
$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

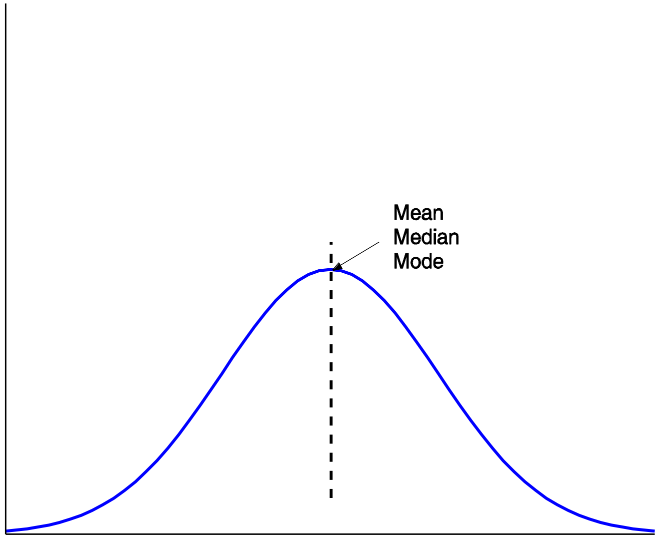
Median, mean and mode of symmetric, positively and negatively skewed data



positively skewed



negatively skewed



symmetric

Measuring the Dispersion of Data

Quartiles, outliers and boxplots

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th 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

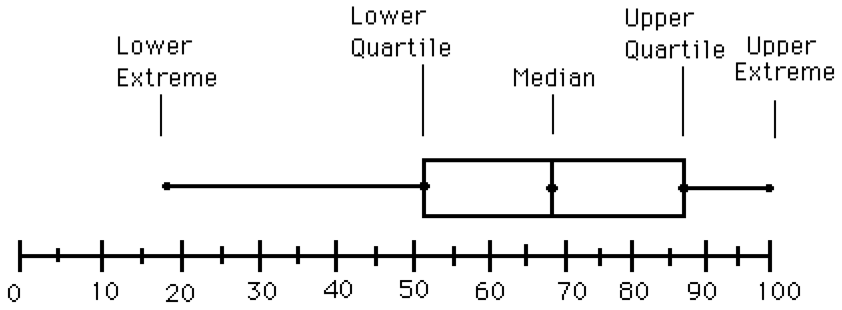
Variance and standard deviation (*sample: s, population: σ*)

- **Variance**: (algebraic, scalable computation)
- **Standard deviation** *s (or σ)* is the square root of variance s^2 (or σ^2)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

Boxplot Analysis

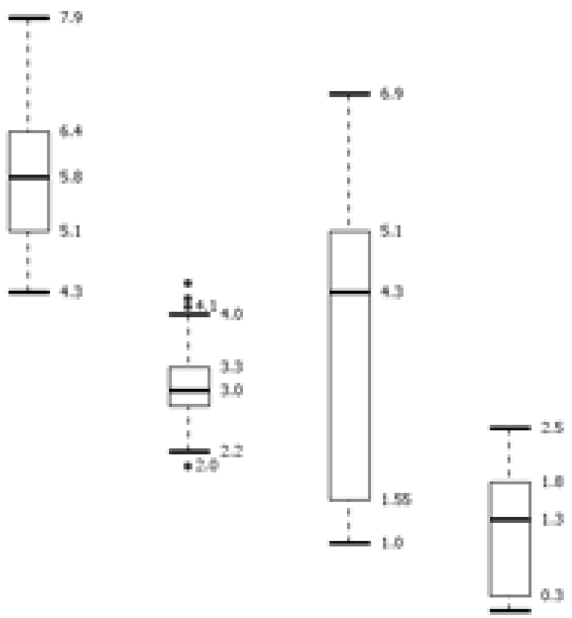


Five-number summary of a distribution

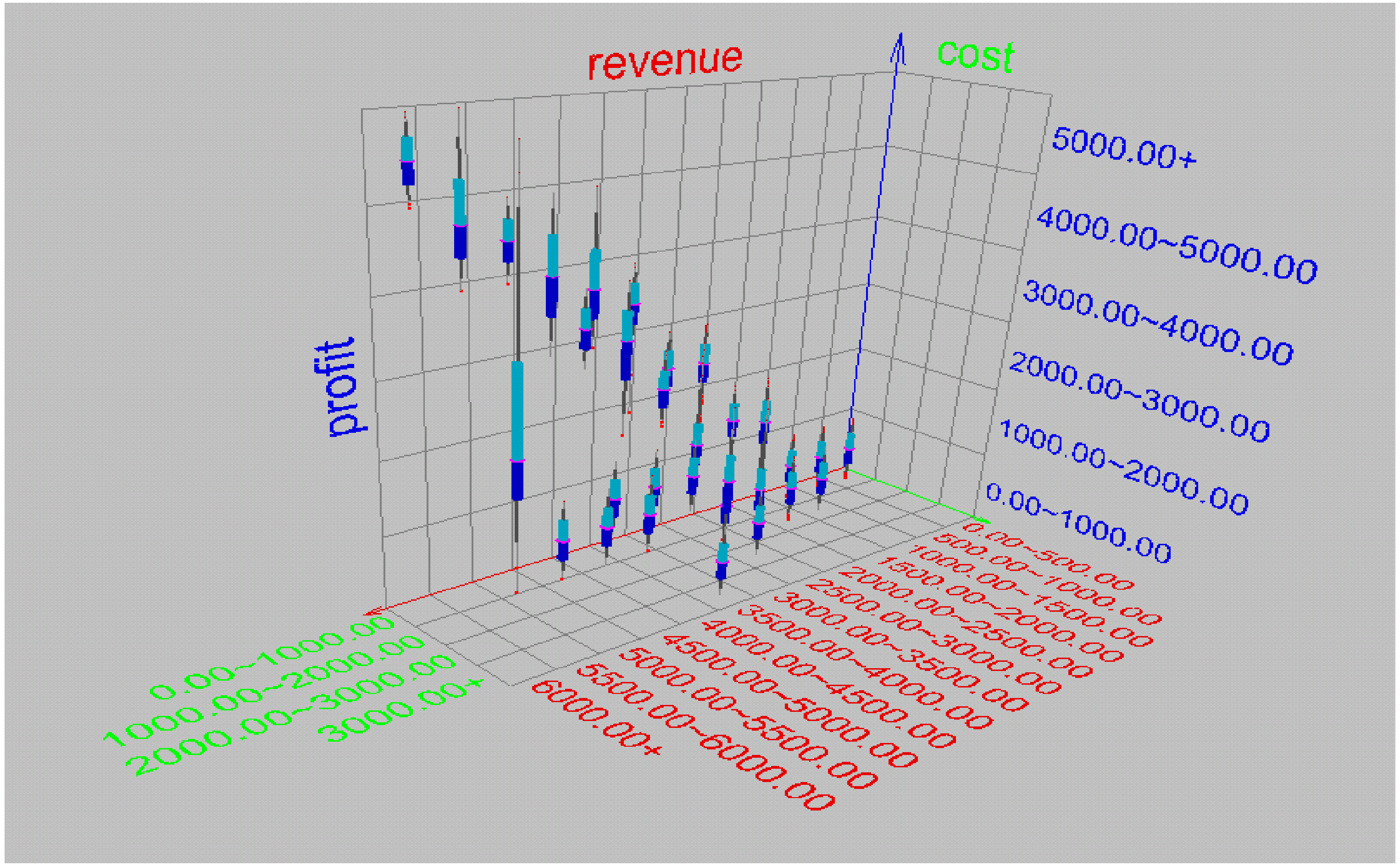
- Minimum, Q1, Median, Q3, Maximum

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



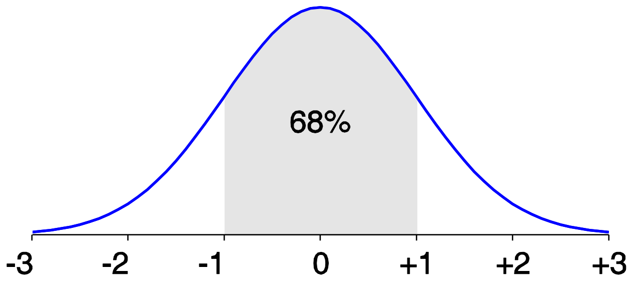
Visualization of Data Dispersion: 3-D Boxplots

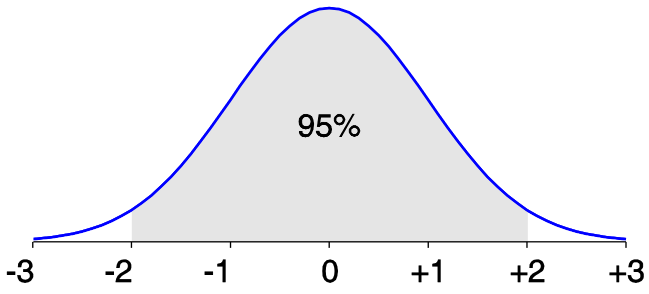


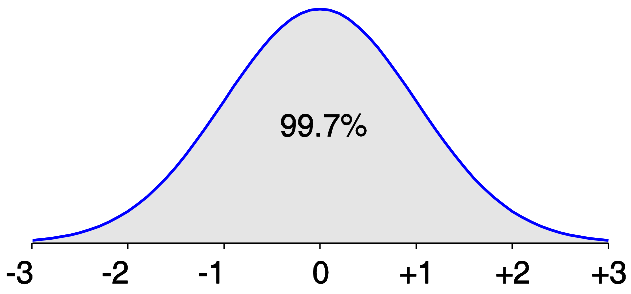
Properties of Normal Distribution Curve

The normal (distribution) curve

- From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
- From μ–2σ to μ+2σ: contains about 95% of it
- From μ–3σ to μ+3σ: 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 represents frequencies

Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$

Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another

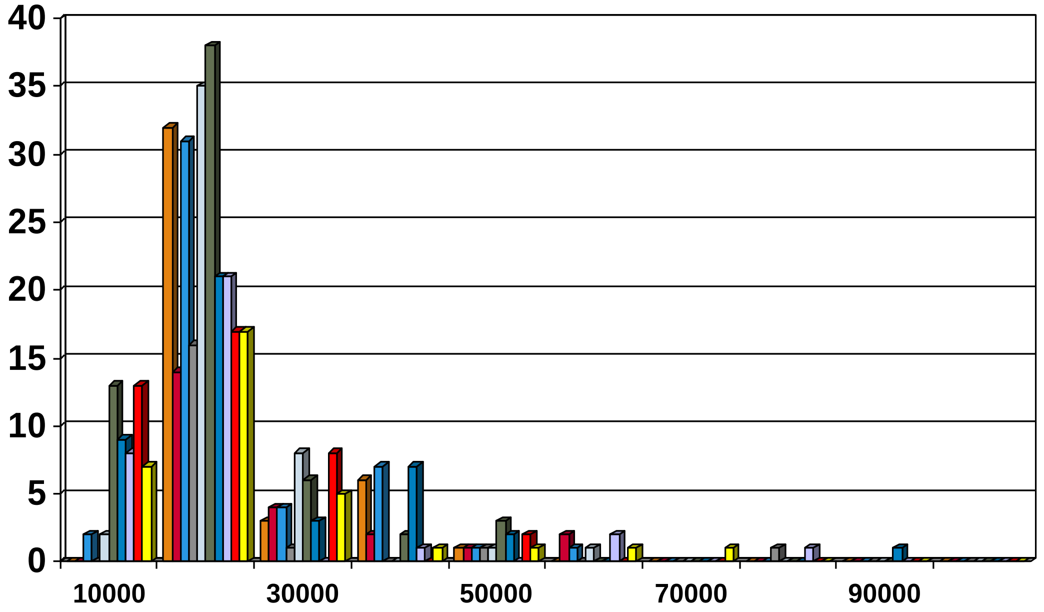
Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

Histogram Analysis

Histogram: Graph display of tabulated frequencies, shown as bars

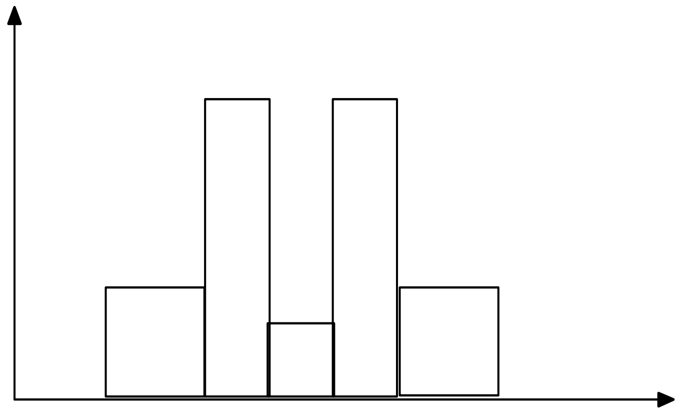
It shows what proportion of cases fall into each of several categories

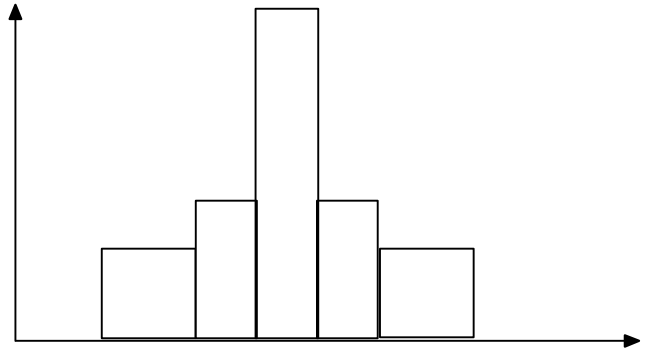
Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width



The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

Histograms Often Tell More than Boxplots





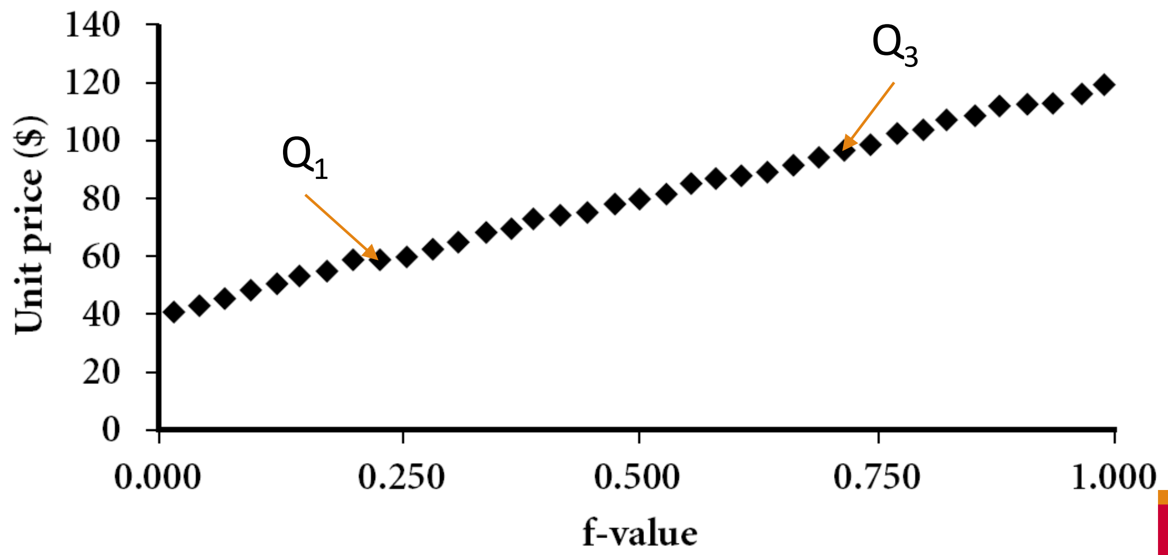
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Quantile Plot

Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

Plots **quantile** information

- For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i

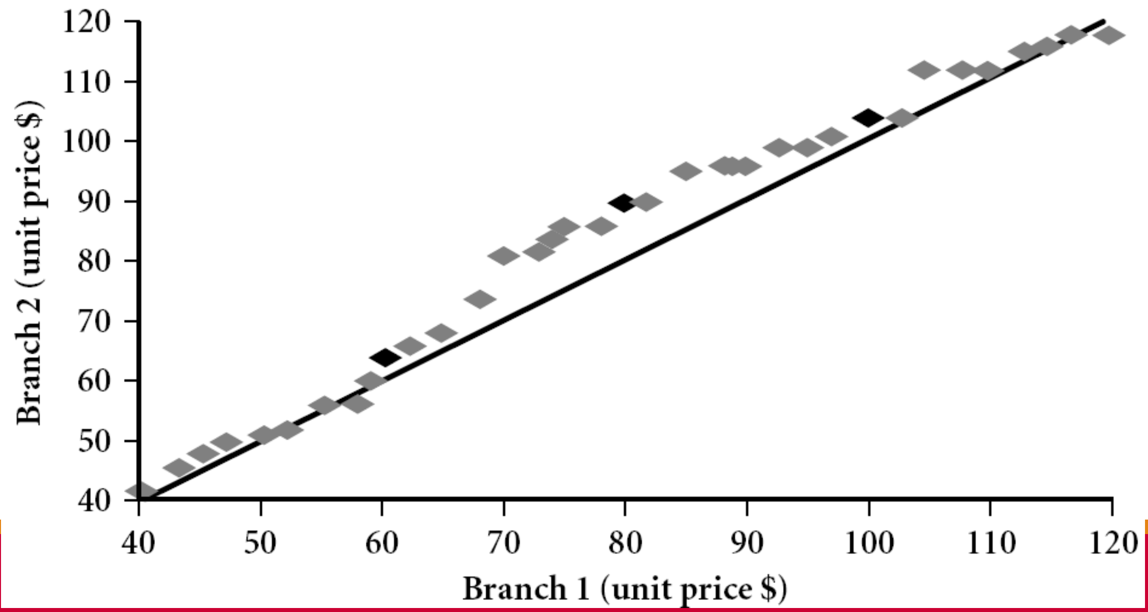


Quantile-Quantile (Q-Q) Plot

Graphs the quantiles of one univariate distribution against the corresponding quantiles of another

View: Is there is a shift in going from one distribution to another?

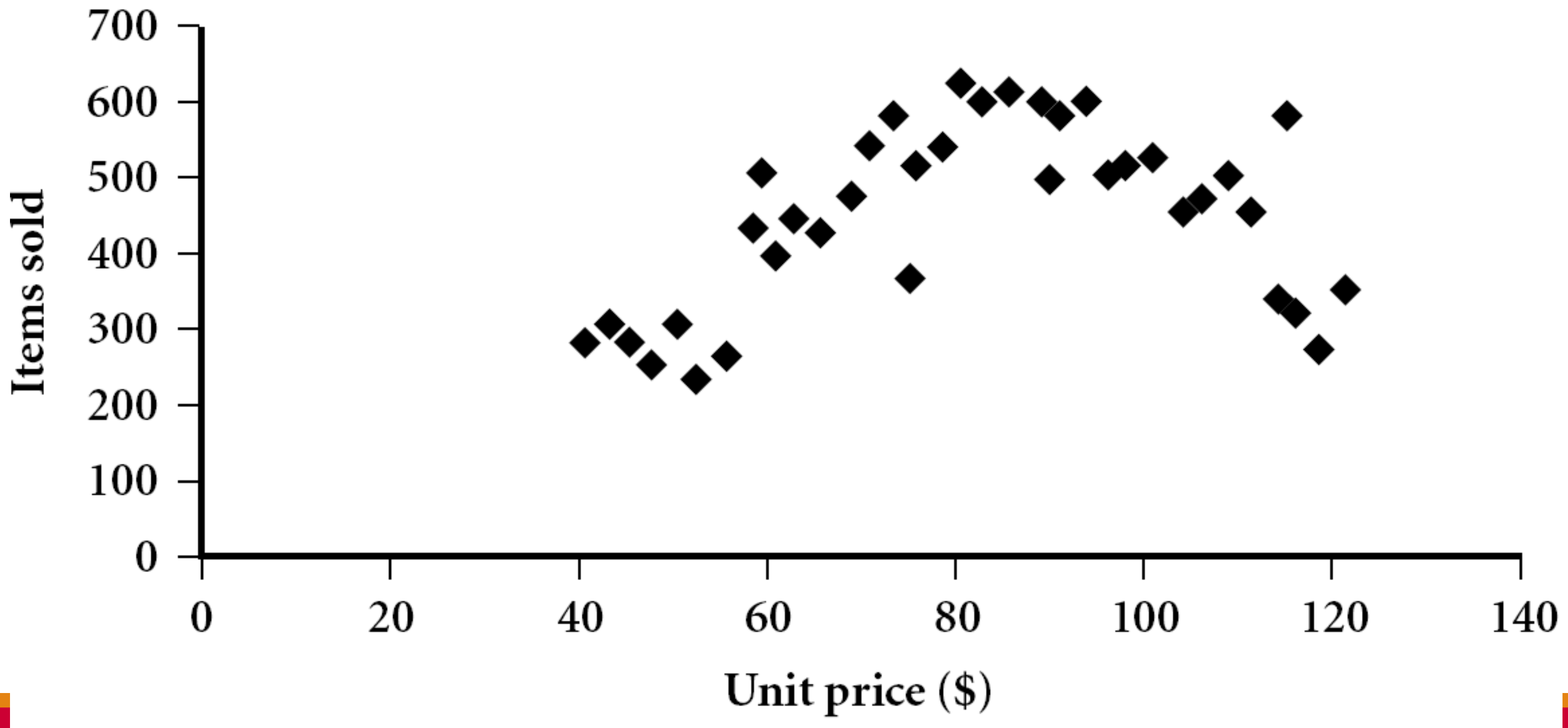
Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



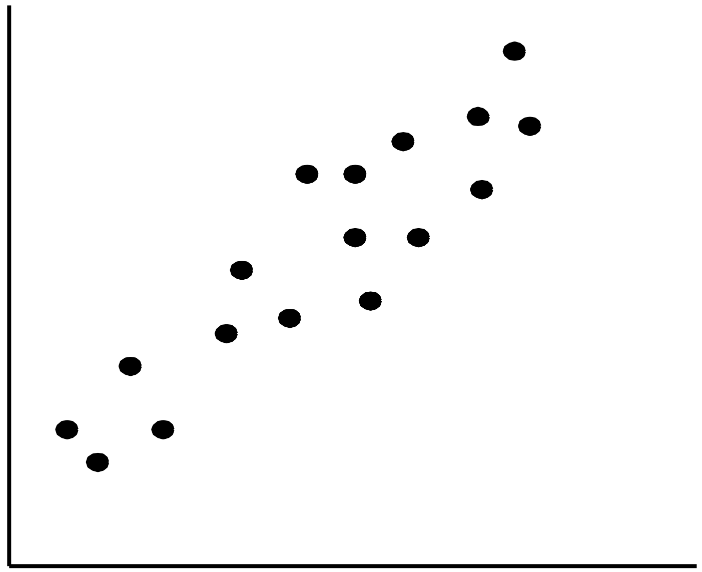
Scatter plot

Provides a first look at bivariate data to see clusters of points, outliers, etc

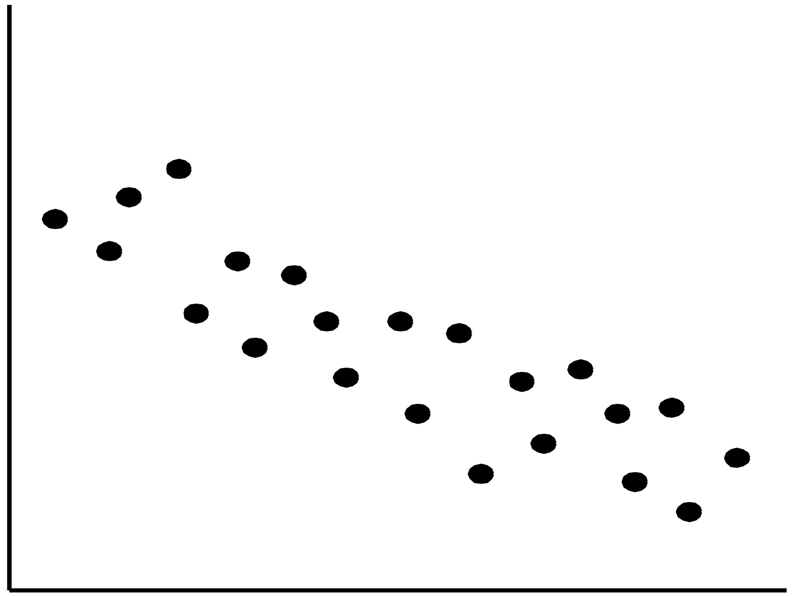
Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

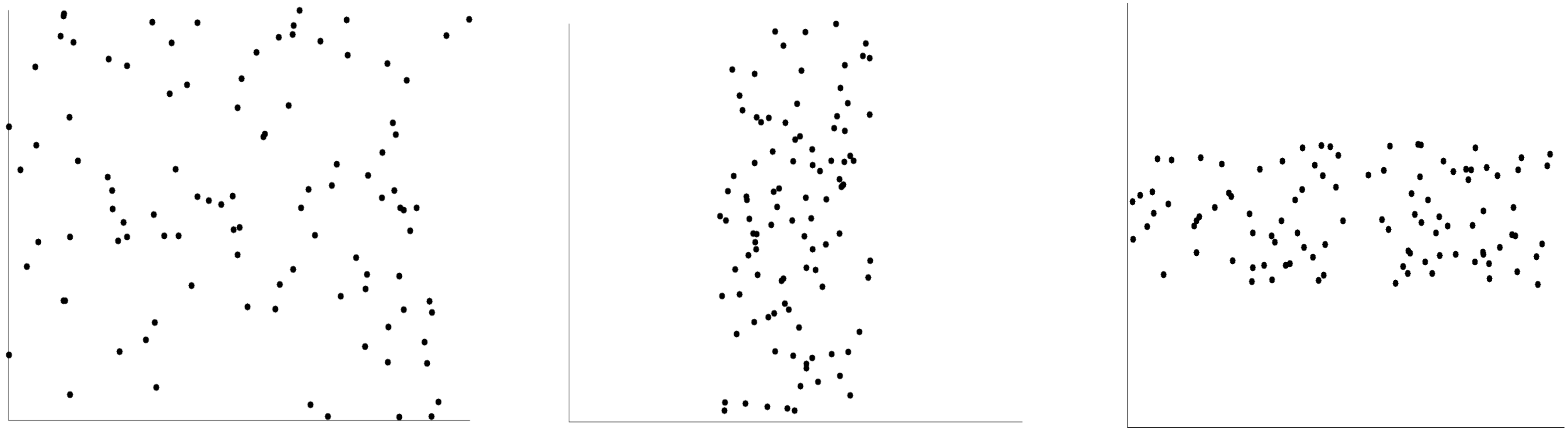


positively correlated



negative correlated

Uncorrelated Data



Getting to Know Your Data

Data Objects and Attribute Types

Basic Statistical Descriptions of Data

Data Visualization

Measuring Data Similarity and Dissimilarity

Summary

Similarity and Dissimilarity

Similarity

- Numerical measure of how alike two data objects are
- Value is higher when objects are more alike
- Often falls in the range [0,1]

Dissimilarity (e.g., distance)

- Numerical measure of how different two data objects are
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

Proximity refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

Data matrix

- n data points with p dimensions

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity matrix

- n data points, but registers only the distance
- A triangular matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Proximity Measure for Nominal Attributes

Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)

Method 1: Simple matching

- m: # of matches, p: total # of variables/features

$$d(i, j) = \frac{p - m}{p}$$

Method 2: Use a large number of binary attributes

- creating a new binary attribute for each of the M nominal states

Proximity Measure for Binary Attributes

A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Distance measure for symmetric binary variables:

$$d(i, j) = \frac{r+s}{q+r+s+t}$$

Distance measure for asymmetric binary variables:

$$d(i, j) = \frac{r+s}{q+r+s}$$

Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$$

Dissimilarity between Binary Variables

Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Standardizing Numeric Data

Z-score:

- X: raw score to be standardized, μ: mean of the population, σ: standard deviation
- the distance between the raw score and the population mean in units of the standard deviation
- negative when the raw score is below the mean, "+" when above

$$z = \frac{x - \mu}{\sigma}$$

An alternative way: Calculate the mean absolute deviation

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where

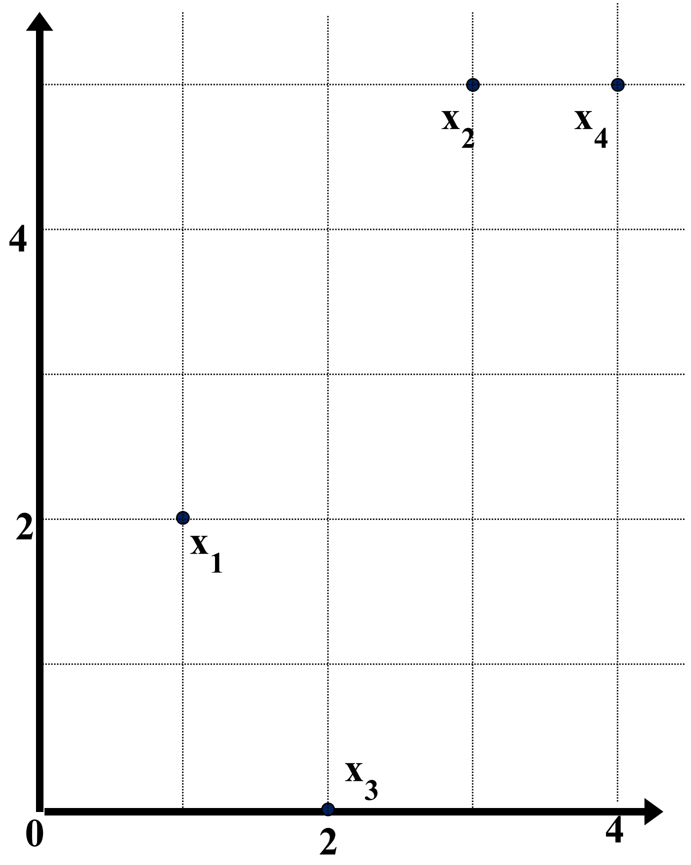
$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

- standardized measure (*z-score*):

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is more robust than using standard deviation

Example:
Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

Minkowski distance: A popular distance measure

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)

Properties

- d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness)
- d(i, j) = d(j, i) (Symmetry)
- d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)

A distance that satisfies these properties is a metric

Special Cases of Minkowski Distance

h = 1: Manhattan (city block, L_1 norm) distance

- E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

h = 2: (L_2 norm) Euclidean distance

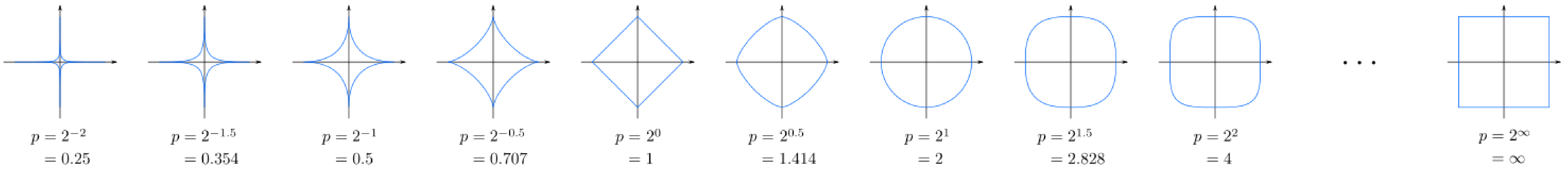
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

$h \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.

- This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if} - x_{jf}|^h\right)^{\frac{1}{h}} = \max_{f}^{p}|x_{if} - x_{jf}|$$

h=p



$p = 2^{-2}$
$= 0.25$
$p = 2^{-1.5}$
$= 0.354$
$p = 2^{-1}$
$= 0.5$
$p = 2^{-0.5}$
$= 0.707$
$p = 2^{0}$
$= 1$
$p = 2^{0.5}$
$= 1.414$
$p = 2^{1}$
$= 2$
$p = 2^{1.5}$
$= 2.828$
$p = 2^{2}$
$= 4$
$\cdots$
$p = 2^{\infty}$
$= \infty$

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Ordinal Variables

An ordinal variable can be discrete or continuous

Order is important, e.g., rank

Can be treated like interval-scaled

$$r_{if} \in \{1, \ldots, M_f\}$$

- replace x_{if} by their rank
- map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

A database may contain all attribute types

- Nominal, symmetric binary, asymmetric binary, numeric, ordinal

One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

- f is binary or nominal:

  $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise

- f is numeric: use the normalized distance
- f is ordinal
  - Compute ranks r_{if} and
  - Treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Cosine Similarity

A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Other vector objects: gene features in micro-arrays, …

Applications: information retrieval, biologic taxonomy, gene feature mapping, ...

Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2|| ,$$

where $\bullet$ indicates vector dot product, $||d||$: the length of vector d

Example: Cosine Similarity

$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||$,
where $\bullet$ indicates vector dot product, $||d|$: the length of vector d

Ex: Find the **similarity** between documents 1 and 2.

d_1 **=** (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
d_2 **=** (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$
$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$
$\cos(d_1, d_2) = 0.94$

Summary

Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled

Many types of data sets, e.g., numerical, text, graph, Web, image.

Gain insight into the data by:

- Basic statistical data description: central tendency, dispersion, graphical displays
- Data visualization: map data onto graphical primitives
- Measure data similarity

Above steps are the beginning of data preprocessing

Many methods have been developed but still an active area of research

References

W. Cleveland, Visualizing Data, Hobart Press, 1993

T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003

U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001

L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.

H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997

D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002

D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999

S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999

E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001

C. Yu et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009