

# Data Mining

# INDEX UNIT 1



## Introduction:

Data Mining and Analysis,  
Challenges  
Types of Data and Patterns, Application,



## Understanding the Data

Measuring the Central Tendency:  
• Mean, Median, and Mode, Measuring the



## Dispersion of Data:

Range, Quartiles, Variance, Standard Deviation, Data Visualization Techniques.

# Discussion on HBS Case Study



## **Case Synopsis: Netflix Leading with Data: The emergence of data driven video**

By 2009 Netflix had all but trounced its traditional bricks-and-mortar competitors in the video rental industry. Since its founding in the late 1990s, the company had changed the face of the industry and threatened the existence of such entrenched giants as Blockbuster, in large part because of its easy-to-understand subscription model, policy of no late fees, and use of analytics to leverage customer data to provide a superior customer experience and grow its e-commerce media platform. Netflix's investment in data collection, IT systems, and advanced analytics such as proprietary data mining techniques and algorithms for customer and product matching played a crucial role in both its strategy and success. However, the explosive growth of the digital media market presents a serious challenge for Netflix's business going forward. How will its analytics, customer data, and customer interaction models play a role in the future of the digital media space? Will it be able to stand up to competition from more seasoned players in the digital market, such as Amazon and Apple? What position must Netflix take in order to successfully compete in this digital arena?

## Discussion on HBS Case Study – Cont..

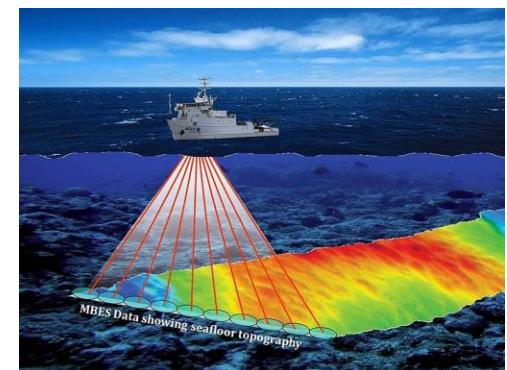
### Learning Objectives

- To examine the benefits and risks of investment in analytical technology as a means for mining customer data for business insights.
- Students will develop a strategy position for Netflix's investment in technology and its digital media business. Students must also consider how new corporate partnerships and changes to the customer channel model will allow the company to prosper in the highly competitive digital space.

# Into the Digital Era

## ➤ People's daily lives

- > 4 billion internet users
- Social media, smart devices, ...



## ➤ Scientific discovery

- Rubin Observatory: 20TB/night



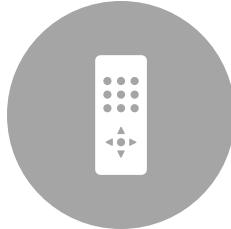
## ➤ Many application domains...



# Digital Era



a lot of digital information.



We are interacting with various types of digital devices.



We are creating a lot of digital content and also many different types of apps, and behind the thing, there's more data involved.



**More than four billion Internet users nowadays, and they are a wide variety of social media or smart devices that we interact on a daily basis.**



Another example, think about the *scientific discovery*. In those domains, we are seeing increasingly a number of sensing devices and also increasing sensing capabilities.



As a result, We are continuously generating a lot of *information* for scientific discovery.

# Why Data Mining?



Explosive data growth

KB, MB, GB, TB, PB, EB, ZB

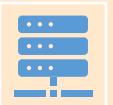
Data creation, transmission, storage, sharing, processing



Ø Drowning in data & starving for knowledge



Need automated analysis of massive data



Its a power to identify patterns and relationships in large volumes of data from multiple sources.

# Why Data Mining

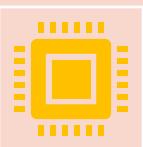


**We are dealing with a lot of data.**



Enabling capabilities -that's the creation of data

we have the capability to collect a lot of data  
we have the way to transmit them so it's easy for us to share or access the different types of data.



We have the storage spaces, so we can now handle a lot of digital data, and it's the computation capabilities

# What is Data Mining?



**Knowledge discovery from data**

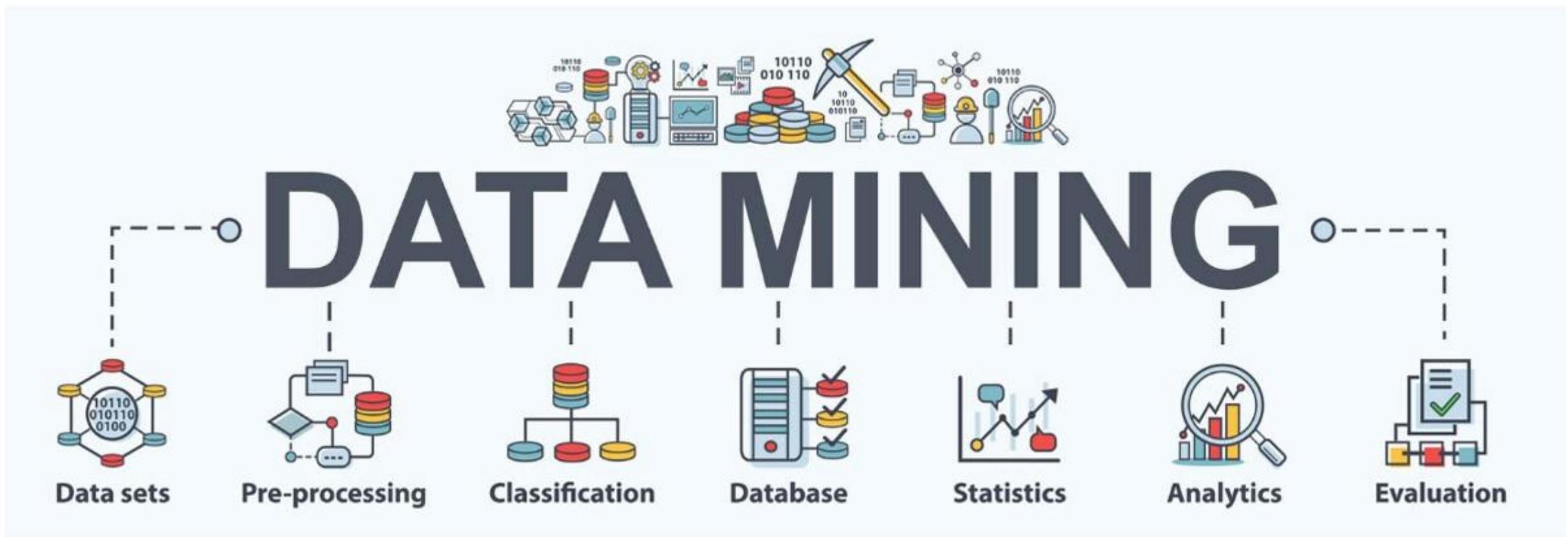
Extraction of interesting patterns or knowledge  
from huge amounts of data

Interesting: valid, previously unknown, potentially  
useful, ultimately understandable by human

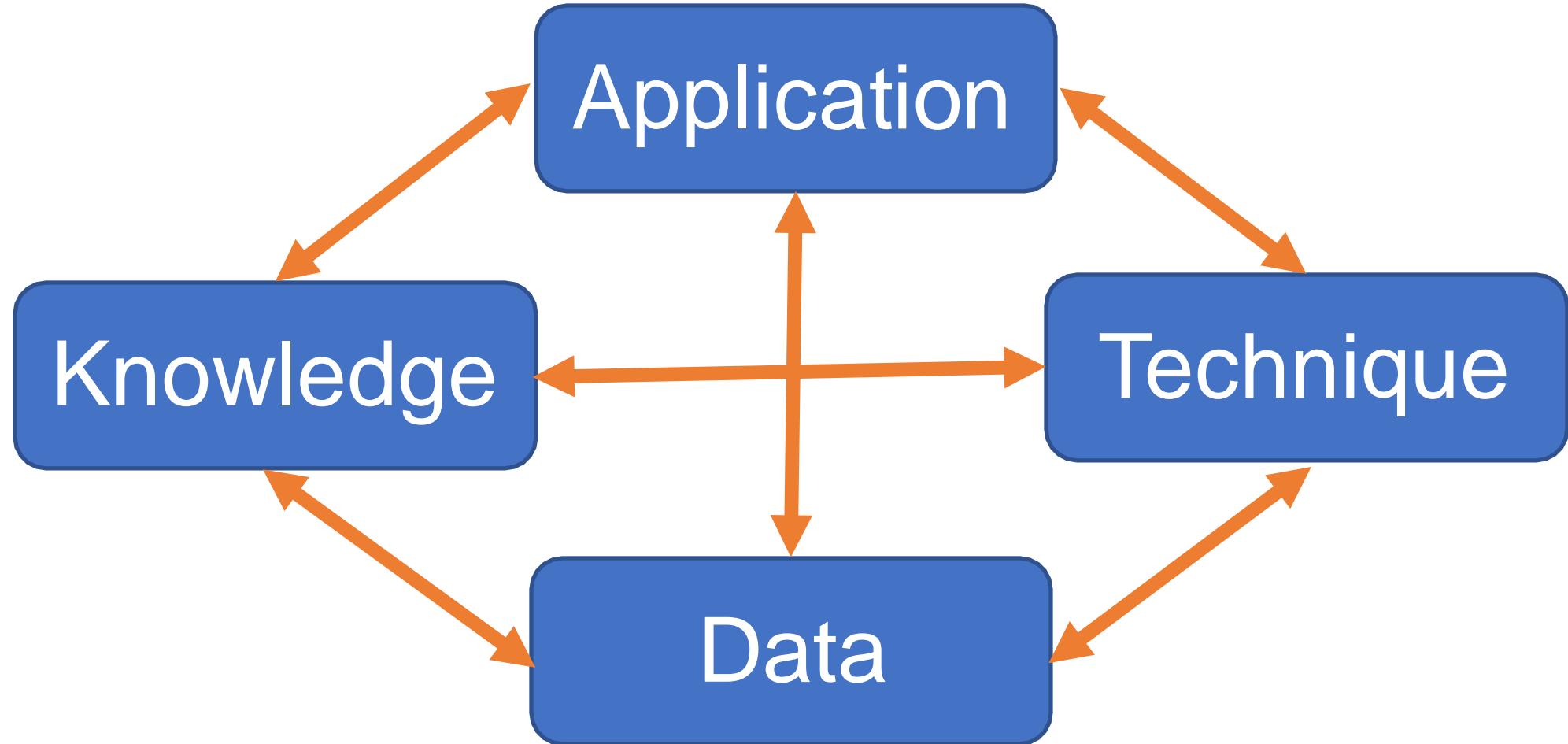


**Huge amounts of data: scalability, efficiency**

# What is Data Mining?



# Data Mining: Four Views



- Arrows really demonstrate how interconnected and also how important it is for you to have this integrated view, because you don't just take one piece and then just focus on that one, you do need to have a reasonably good understanding of all those components and how they come together.

# Data View (1)

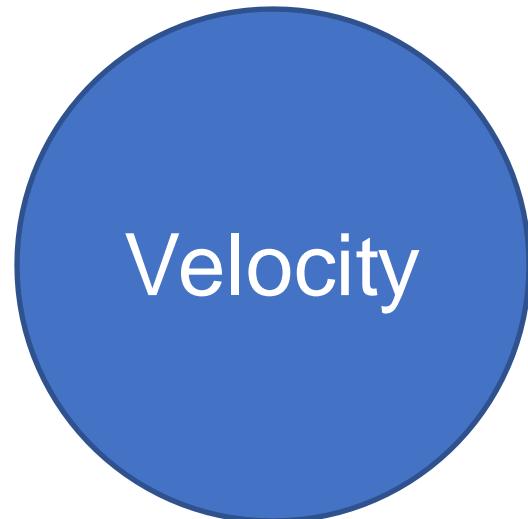
- The 3Vs, 4Vs, 5Vs



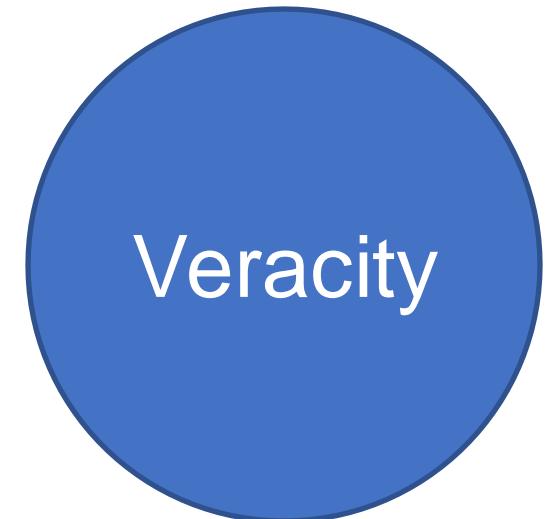
Volume



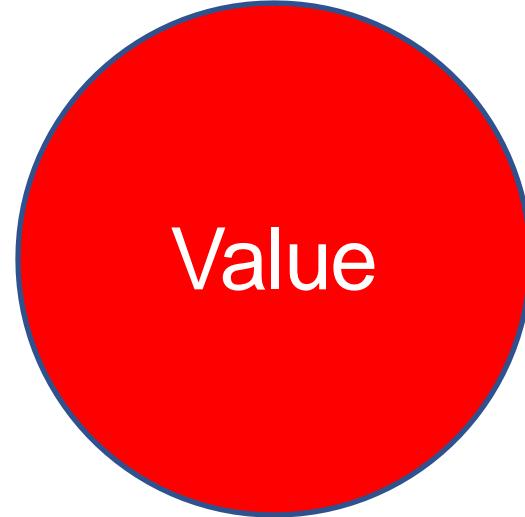
Variety



Velocity



Veracity



Value

# 1<sup>st</sup> V is about volume



It is estimated that we create 2.3 trillion gigabytes of data every day. And that will only increase.



This increase is of course partly caused by the gigantic mobile telephone network.



six of the seven billion people in the world now have a mobile phone.

Text and WhatsApp messages, photos, videos and many apps ensure that the amount of data increases significantly.



As the volume grows so rapidly, so does the need for new database management systems and IT employees.

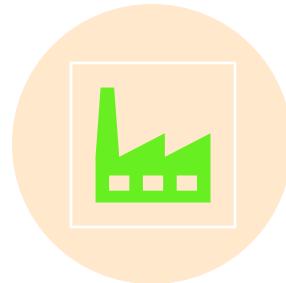
# 2<sup>nd</sup> V is about Verity



The high speed and considerable volume are related to the variety of forms of data.



Consider, for example, the electronic patient records in healthcare, which contribute to many trillions of gigabytes of data



Smart IT solutions are available today for all sectors, from the medical world to construction and business.



the videos we watch on Youtube, the posts we share on Facebook and the blog articles we write.

# 3<sup>rd</sup> V is Velocity



**The third V is about velocity.** That'll basically talk about changes or dynamics of your data.



There are traditional or even nowadays in many other scenarios, starting with static data.



That means you just get one copy of your dataset and then you can work with it, but many other times you're talking about things that change



The velocity really speaks to the point about how quickly data are being generated, but also related to that is about how quickly your analysis needs to be done, how you react to the dynamic data

## 4<sup>th</sup> v is about Veracity

“ ”

**This V is about veracity.** What it talks about is really the quality of your data. To what extent do you trust your data?

To actually the point your dataset is actually hopefully free of errors, or any potential issues.



# That's how we have 5<sup>th</sup> V



V really lead to the 5th V. **That is about a value.**



Whatever you do with the data mining, we want to bring out value from the data.

## Data View (2)



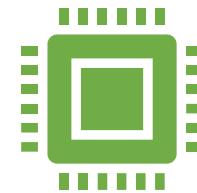
### **Relational, transactional data**

E.g., student records, bank accounts, store purchases



### **Sequential, temporal, streaming data**

E.g., gene sequences, stock prices, sensor readings



### **Spatial, spatial-temporal data**

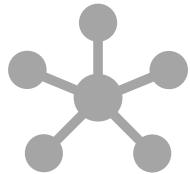
E.g., land use, bird migration, traffic condition

## Data View (3)



**Text, multimedia, Web data**

E.g., news articles, audio/video/image, hypertext



**Graph, network data**

E.g., social network, power grid, co-authorship



**Single or mixture of multiple data types**

# Textual Data



Textual data is widely used nowadays, and also actually big part of many data mining tasks.



**Think about news articles, think about customer reviews, so those could actually be very useful in many scenarios.**



**Also think about the multimedia data where you could have audio, video, images,**

**how you can extract information  
learn useful patterns where I found those  
margin media into our data**

# Application View (1)



## Market analysis, target advertisement

E.g., customer profiling, product recommendation



## Healthcare, medical research

E.g., disease diagnosis, patient care, drug discovery



## Science and engineering

E.g., air pollution, marine life, electric vehicles

## Application View (2)



Ø Security



E.g.,  
surveillance,  
intrusion/crim  
e, fraud,  
cyberattack



Ø  
Government,  
nonprofit



E.g., urban  
planning,  
traffic control,  
education



Ø And many  
many more ...

# Knowledge View (1)



**Frequent pattern,  
association,  
correlation**

E.g., songs listened together or in certain sequence

E.g., A is (more/less) likely to happen given B



**Categorization**

E.g., similarity among users with certain purchases

E.g., differences between two patient groups



**Anomaly,  
outliers**

E.g., sensor errors, fraud activities, extreme events



**Changes over time**

E.g., emerging new patterns, shift of user interest



**Descriptive,  
predictive,  
prescriptive**

Technique View

---

Frequent pattern analysis

---

Classification, prediction

---

Clustering

---

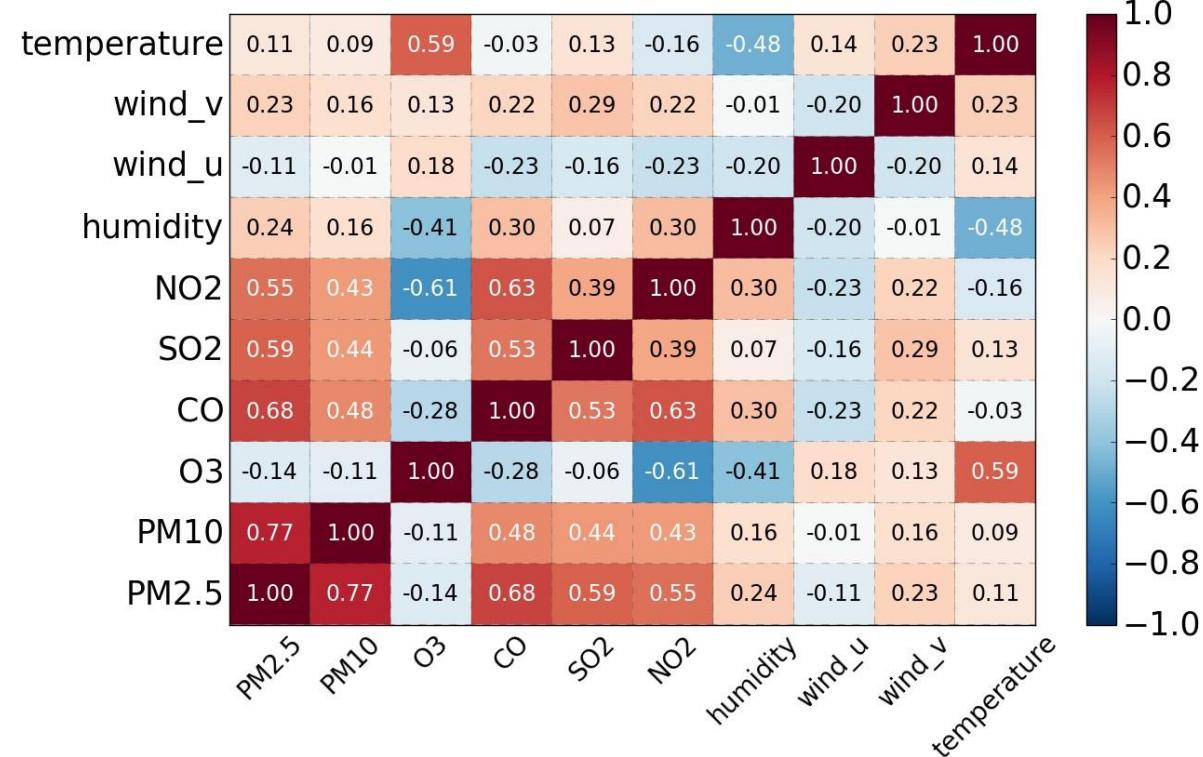
Anomaly detection

---

Trend and evolution analysis

# Frequent Pattern Analysis

- Frequent itemset
- Frequent sequence
- Frequent structure
- Association rules
- Correlation analysis





# Frequent Pattern Analysis

## **First one, frequent of pattern analysis.**

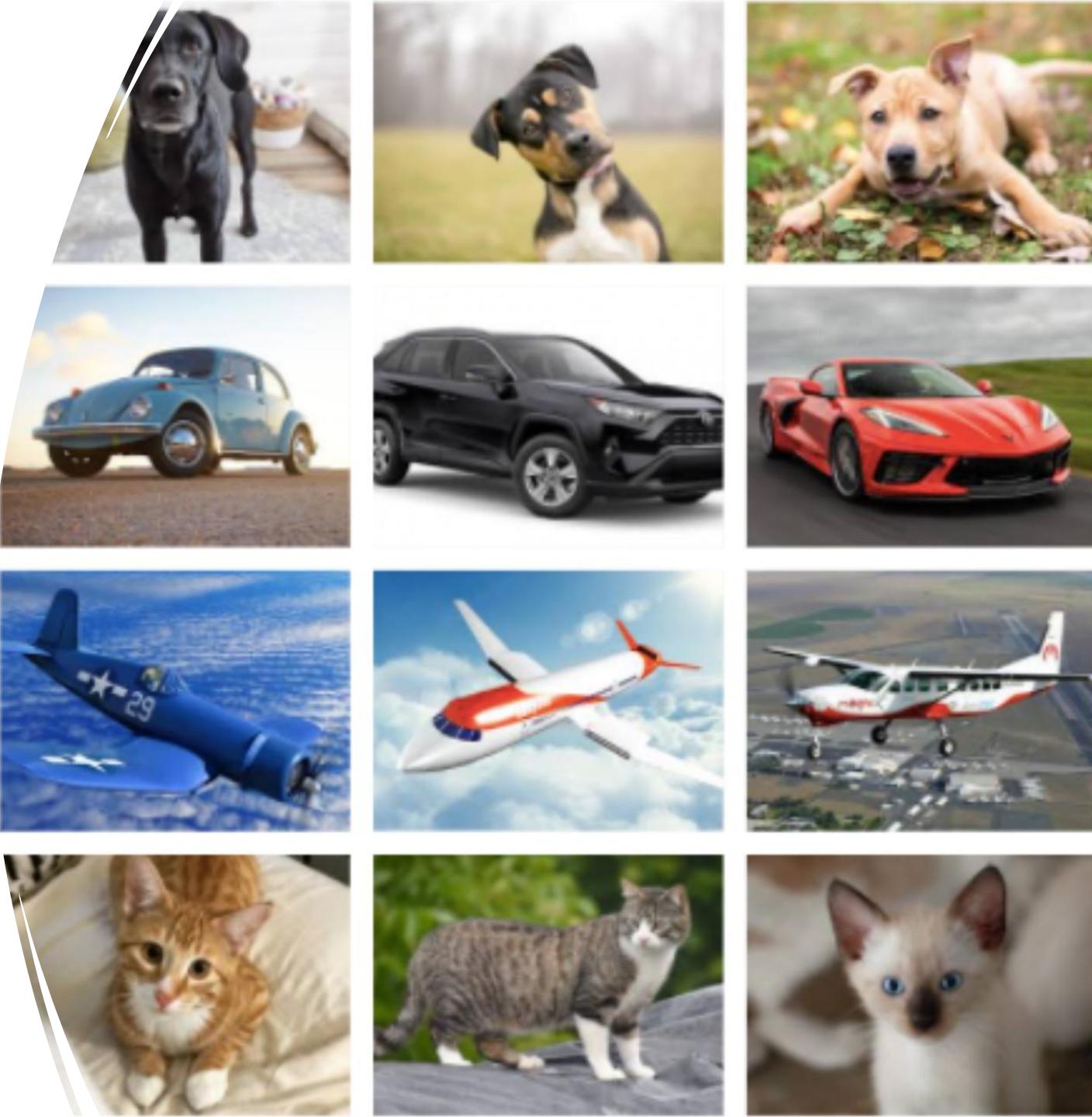
- In terms of knowledge view, we want to be able to identify what happens frequently, so depending on what type of information you're looking at, so you may be looking at frequent itemset, so that display a set of items or events that tend to occur together.

# Frequent Pattern Analysis

- Think about social networks.
- If you look at a certain groups or communities, you may see a frequent structural where it's star-shaped, so that means you have one leader who's interacting closely with the individual members or you may have some peer-to-peer graphs where they're smaller.
- Everybody's talking to everybody.
- Those all general structure that may occur regularly.
- But then beyond identified things that occur frequently, we also want to be able to identify relationships, so this auxiliary is about the association, correlation which was spent more time talking about, but here just one example

# Classification

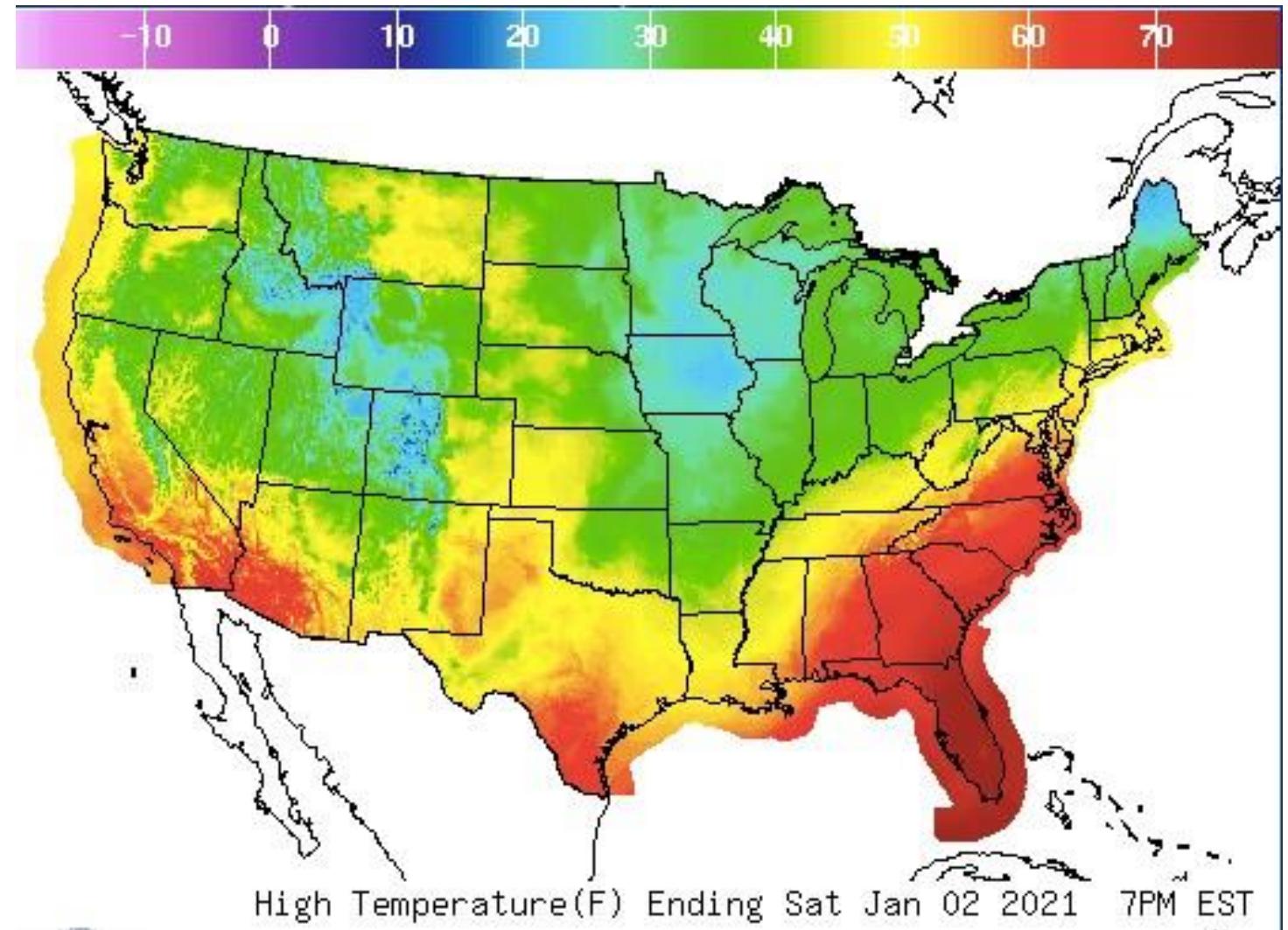
- Pre-defined classes
- Need training data
  - Build model to distinguish classes



# Prediction

Numerical prediction  
(continuous value)

- E.g., weather
- E.g., stock price
- E.g., traffic

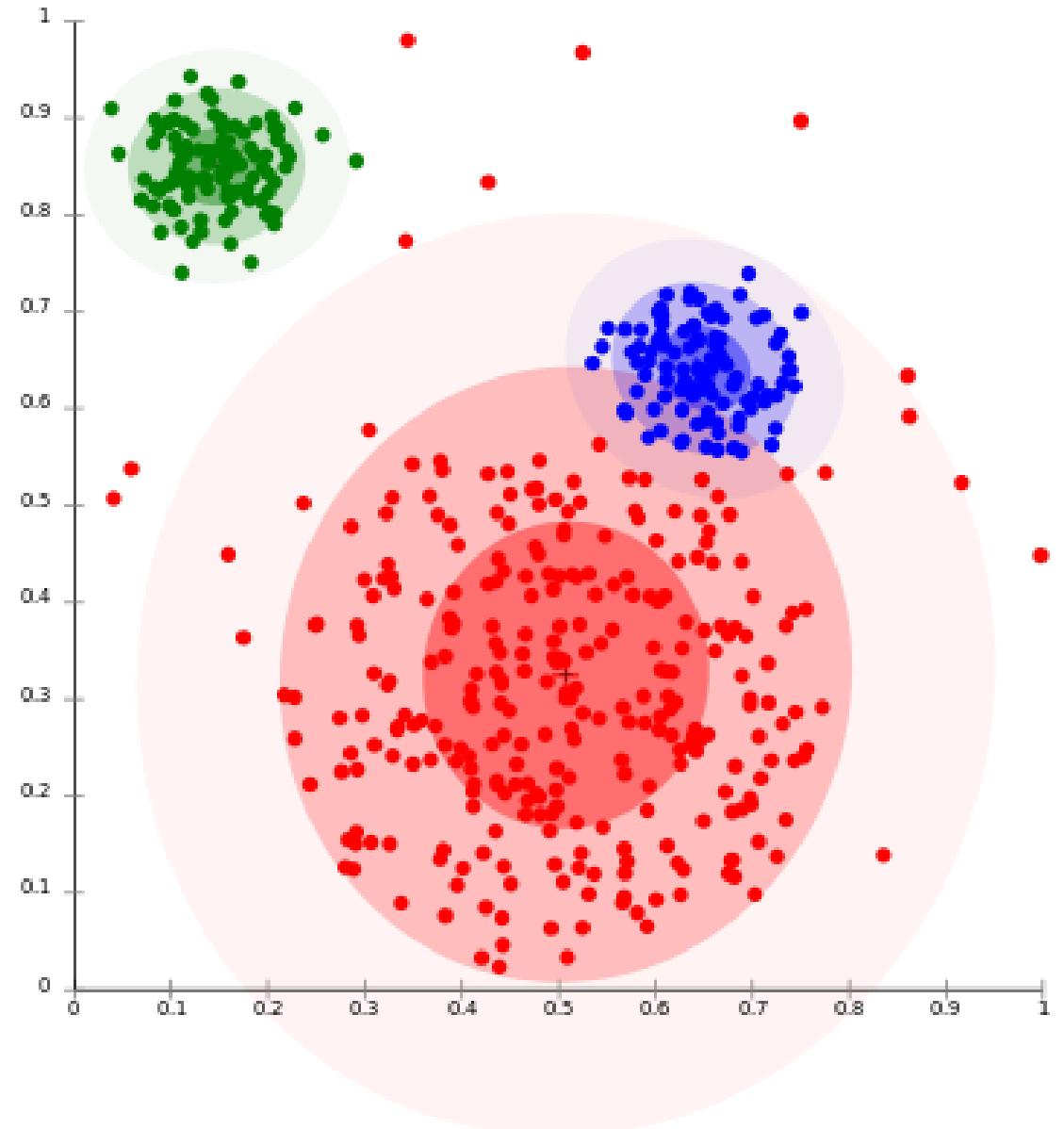


# Prediction Data

- This is generally referred as a numerical prediction in a continuous range.
- Think about whether prediction, it could be predicting the temperature, wind, speed, or amount of precipitation, so those are numerical values you're trying to predict.
- Also stock price, traffic volume, all those are things that are related to being able to differential things, but also then make predictions about what the value would it be.

# Clustering

- No predefined classes
- Intra-cluster similarity
- Inter-cluster dissimilarity

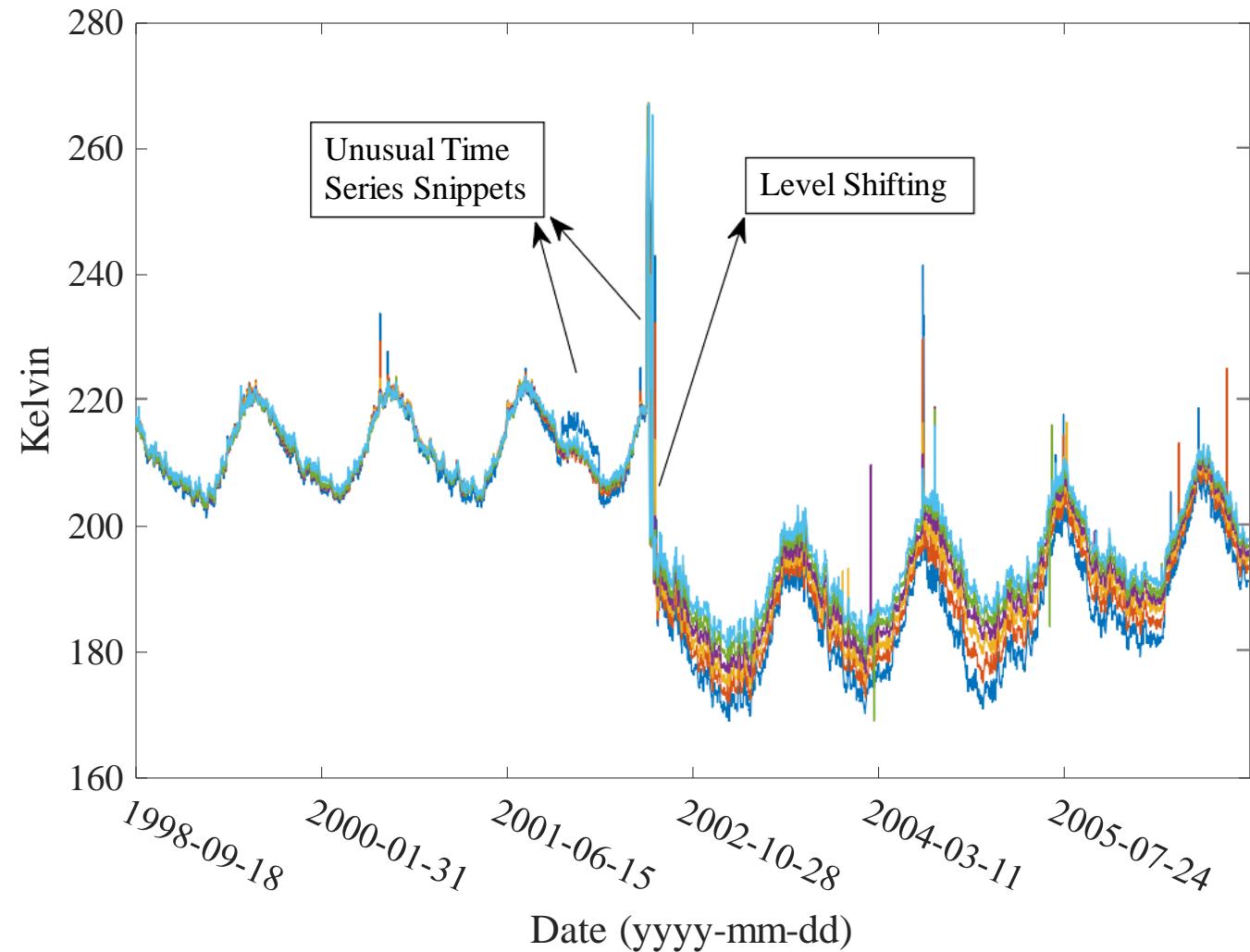


# Clustering

- Another thing related to classification prediction, which is also very useful in many data mining scenarios, is clustering.
- But the key difference between clustering and classification is that here we don't have any predefined class labels.
- The idea is that if you do 2D space, if all the data points are uniformly distributed, then you don't see any classrooms; there's no clusters.
- But many real-world datasets they do have those clustering effect.
- That basically means that you tend to see denser areas where there are more data points, and there are areas that are more sparse and datasets are more separated.

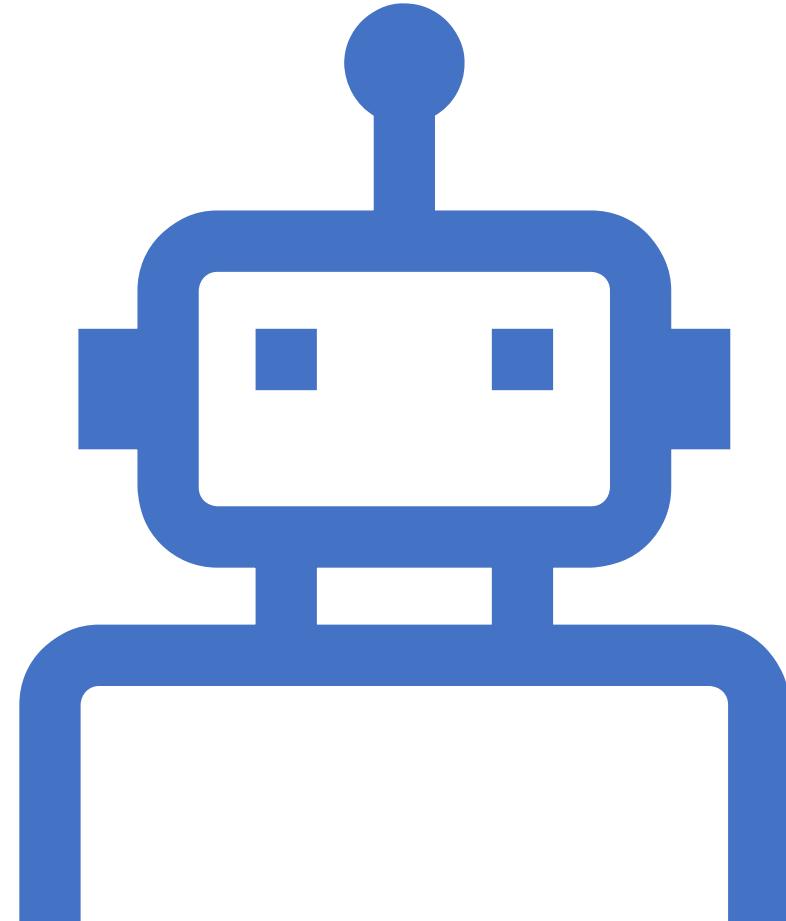
# Anomaly Detection

- Anomaly/outlier
  - Differ from the “norm”
  - E.g., error, noise
  - E.g., fraud
  - E.g., extreme events



# Anomaly

- The general idea with anomalies or outliers is that they're just different.
- But that notion of course is very vague, and the outliers actually turn into, in many real-world scenarios, to identify those

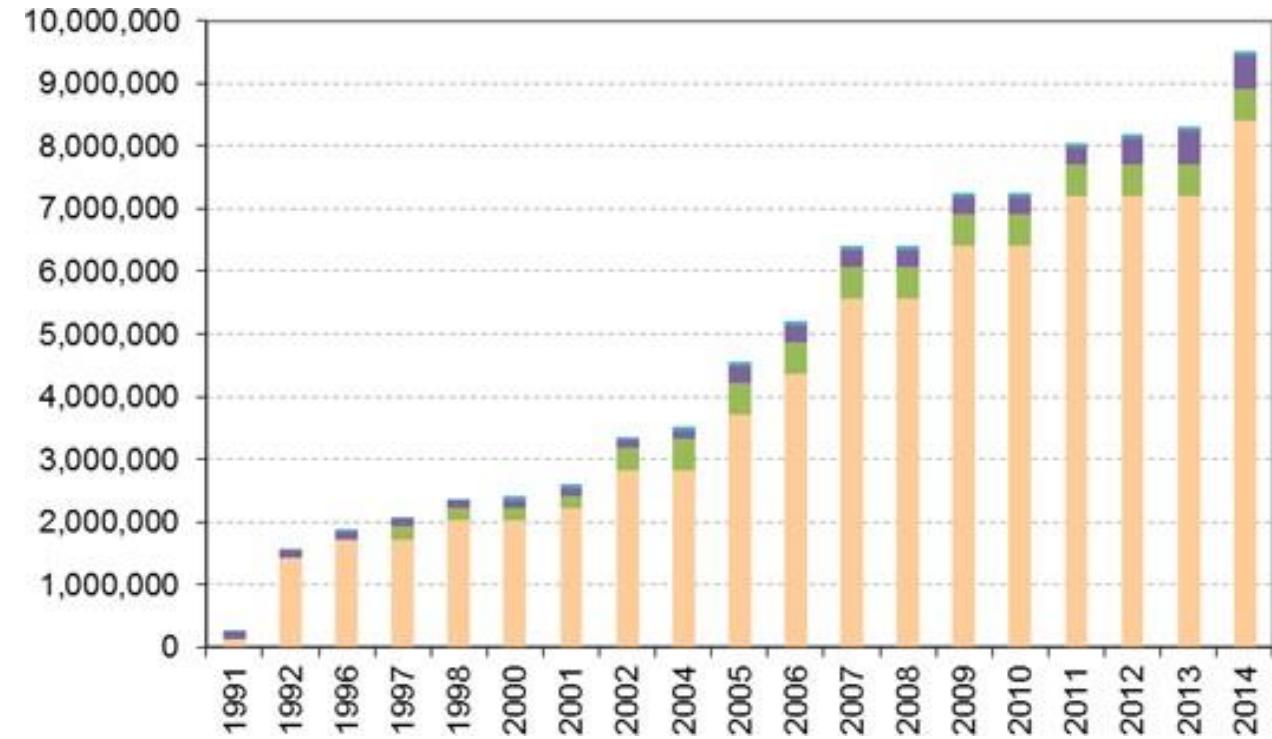


# Trend and Evolution Analysis

## ➤ Changes over time

- Overall trend
- Periodical patterns
- Anomalies
- E.g.,

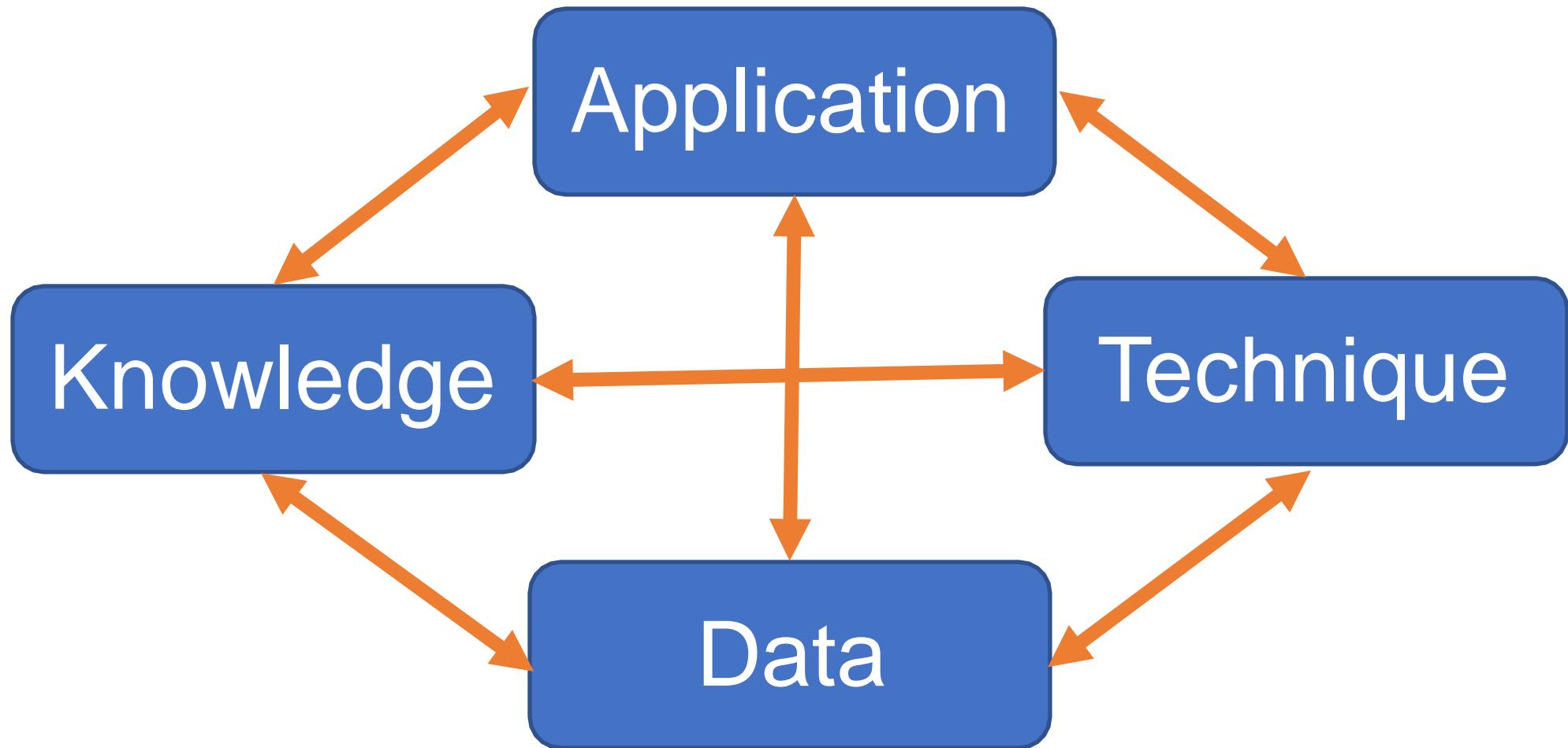
**Google Trends**



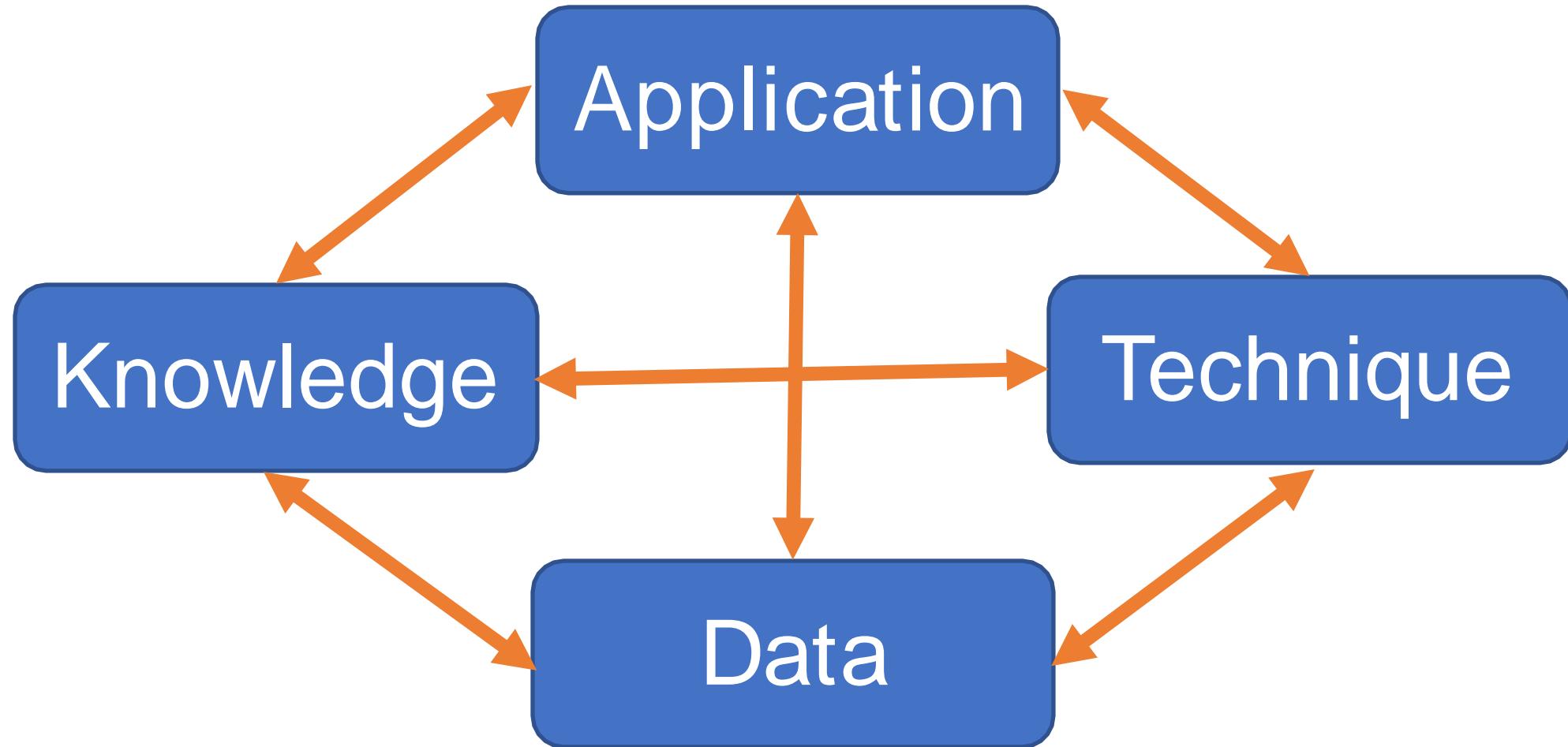
# Trend Analysis

- When you look at your dataset, they're dynamic, they change over time.
- So you need a way to understand how things change over time.
- Again, you can use some visualization capabilities or you can use quantifiable calculations.
- That means the general trend could be a general upward trend even though there could be fluctuations, or downward, or some seasonal patterns.
- Those are general patterns, but of course, you would still be looking for some anomalies.
- Many times when you look at time series, it may be a compensation of the various patterns, so it's not just one single pattern.
- Here of course there are many ways to look at the trends.
- **Also Google Trends is actually a very nice tool.**
- If you go to Google Trends website and you can type in any particular keyword.
- **This is based on the search queries.**

# Data Mining: Four Views



# Data Mining: Four Views

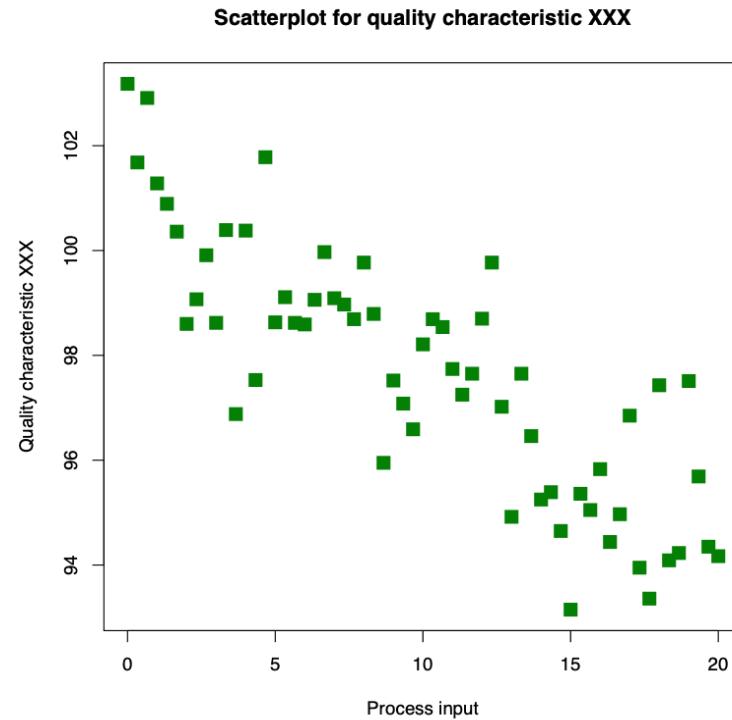


# Four Views

- What data you're dealing with, what applications you're working with, and also what knowledge you would like to learn, and of course, what techniques you can use

# Data Understanding

- What types of data?
- What do they look like?
- Statistics & visualization
- Similarity vs. dissimilarity
- General patterns vs. anomalies

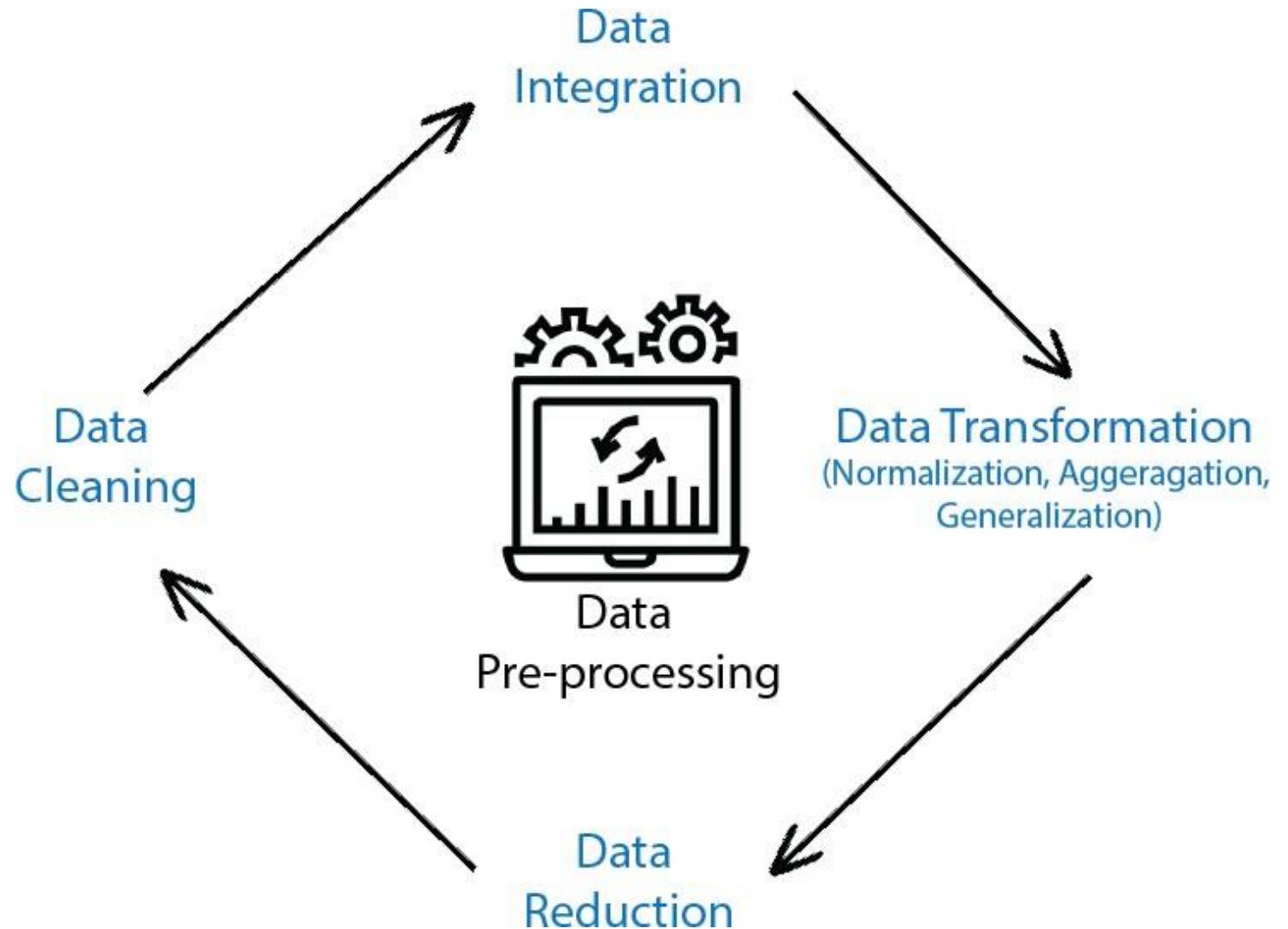


# Data Understanding

- That basically means that, well, are you talking about what types of the attributes, what do they mean?
- **What kind of values do they have?**
- There are various kinds of can do statistical analysis or visualization that will allow you to gain a better understanding of your dataset, okay?
- **There is another very important angle when you have your data or data objects, it's about measuring similarity and dissimilarity.**
- That's important because, with most of the data and mining tasks when you are looking for either general patterns, you'll say, those patterns or those objects tend to be similar

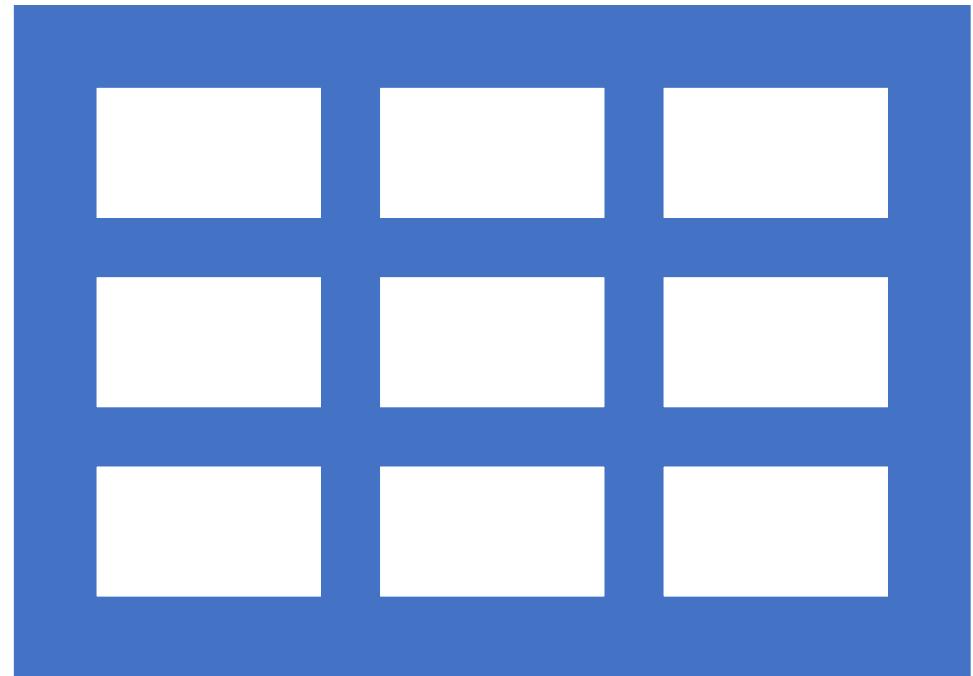
# Data Preprocessing

- Potential issues with data
- E.g., missing data, errors, inconsistency
- Preparing data for the mining process
- Data cleaning, integration, transformation, reduction
- No good data, no good data mining!



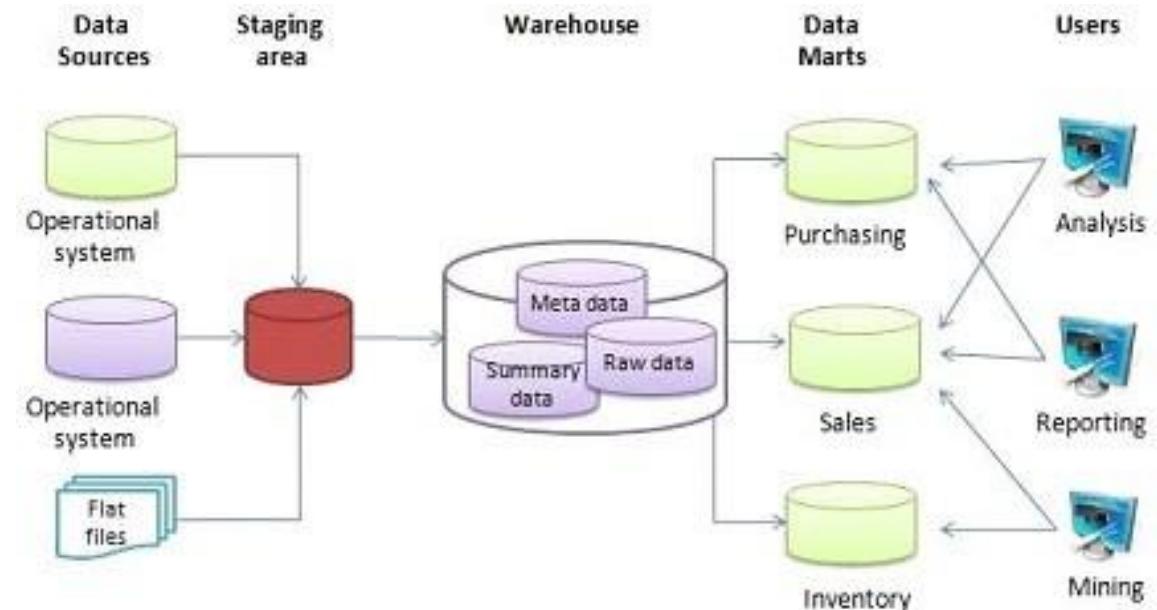
# Data Pre-processing

- There are multiple pieces in it, but the starting point is really that there may be various issues in your dataset.
- The raw data, especially in the real world, are usually nugget.
- They have various types of quality issues or limitations.
- Think about potential errors or missing values, inconsistency
- Those are actually typical in many real-world datasets.
- Need to understand what kind of issues it has, but then continue to then prepare your dataset.



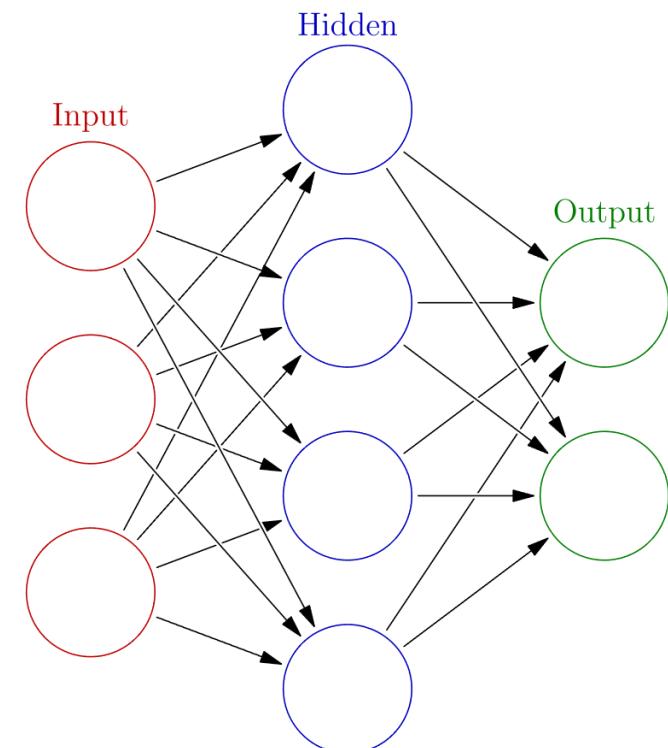
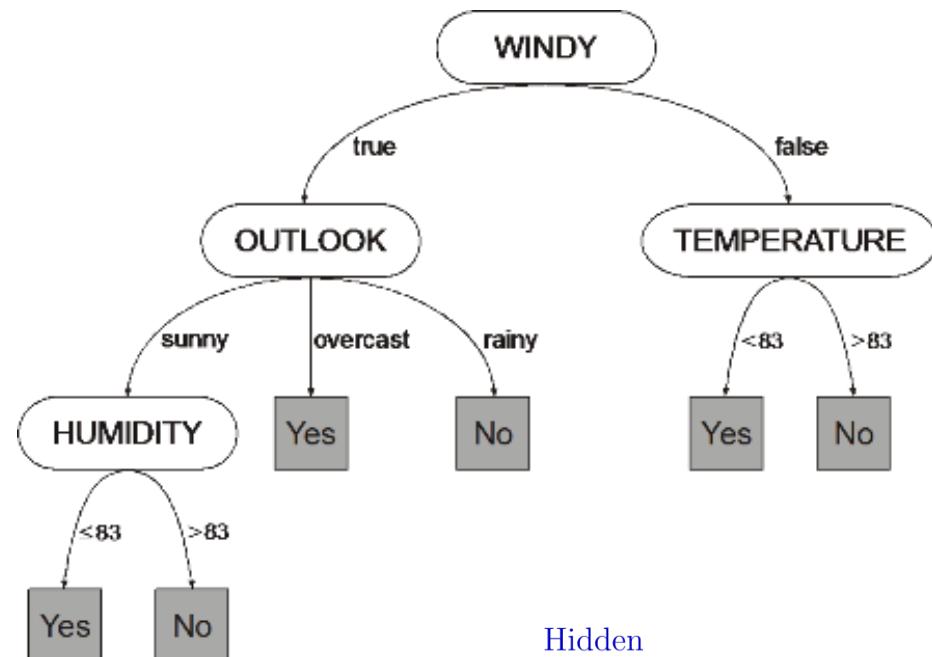
# Data Warehousing

- Data warehouse
- vs. operational data
- Data cube & OLAP
- Multi-dimensional data management
- Data warehouse architecture



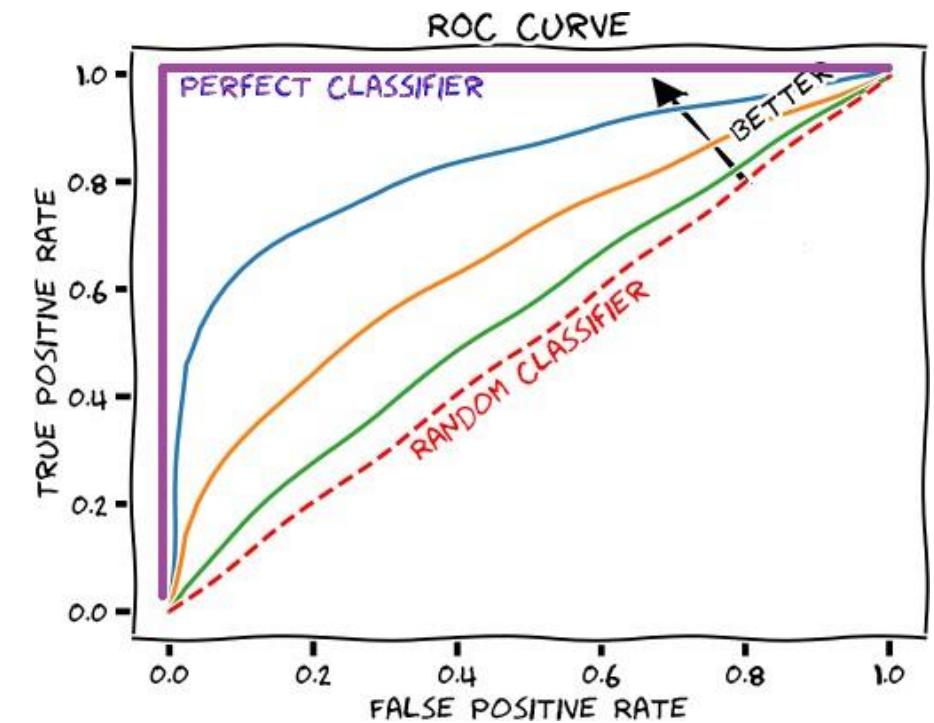
# Data Modeling

- Frequent pattern analysis
- Classification, prediction
- Clustering
- Anomaly detection
- Trend and evolution analysis



# Pattern Evaluation

- Finding interesting patterns from data
  - New, valid, generalizable, useful, explainable
- Evaluation metrics
  - Accuracy, error rate
  - False positive/negative rate
  - Efficiency, latency, ...
- Model selection



# Pattern Evaluation

- To find interesting patterns for your data.
- When you say, "Interesting." You should have some metrics in terms of why do I call interesting?
- You need to be this, there would be a new pattern, it is valid,
- it is a generalized support to certain settings.
- It's useful, or valuable, it can be applied to certain cases, and automatically would like to explain your model.

# Data Mining In the Real World



Integrated views



Data, application, knowledge, technique



Data mining pipeline

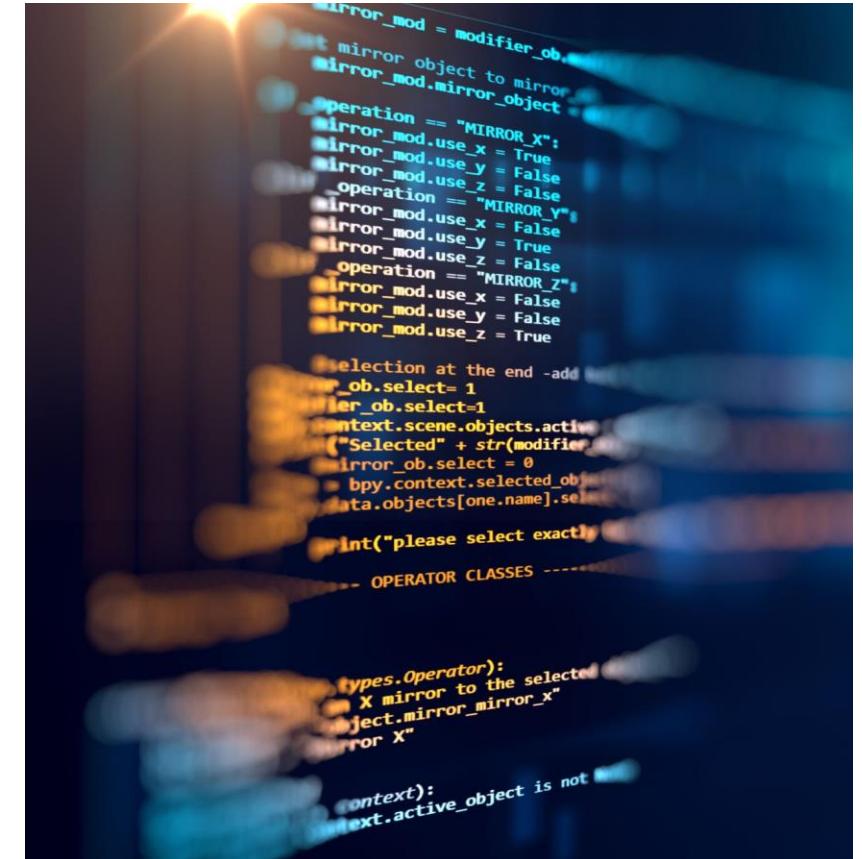


Data understanding,  
preprocessing, warehousing,  
modeling, evaluation

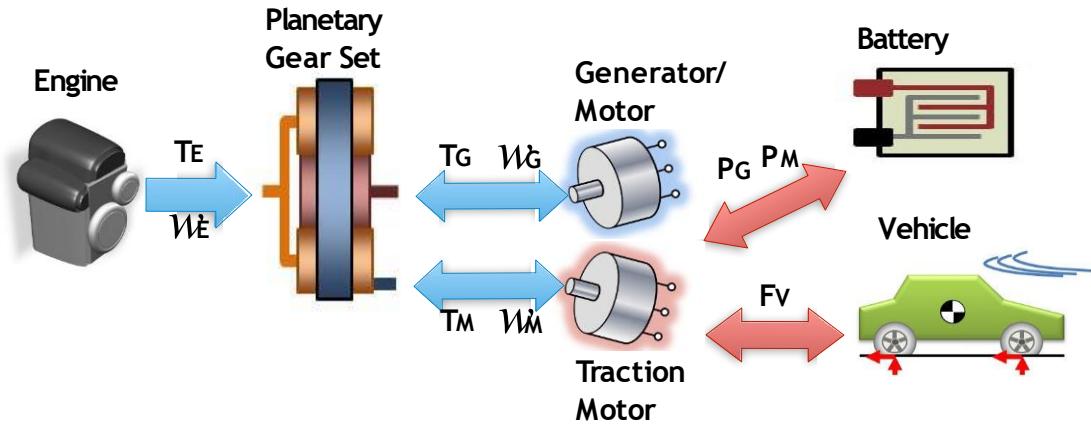
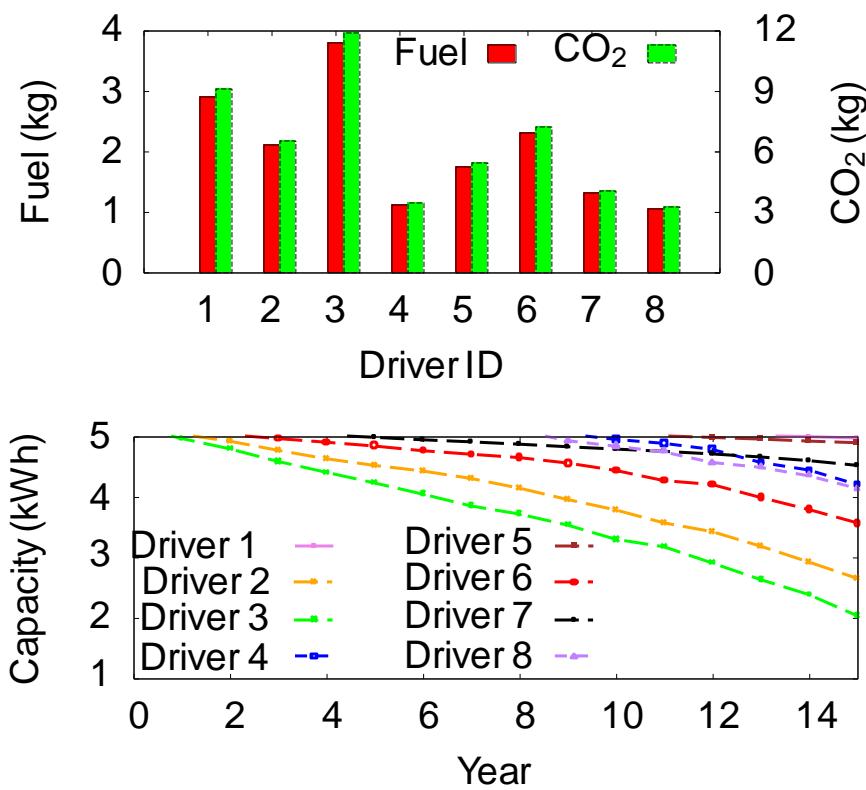
Analytical reasoning!

# Data Mining Examples

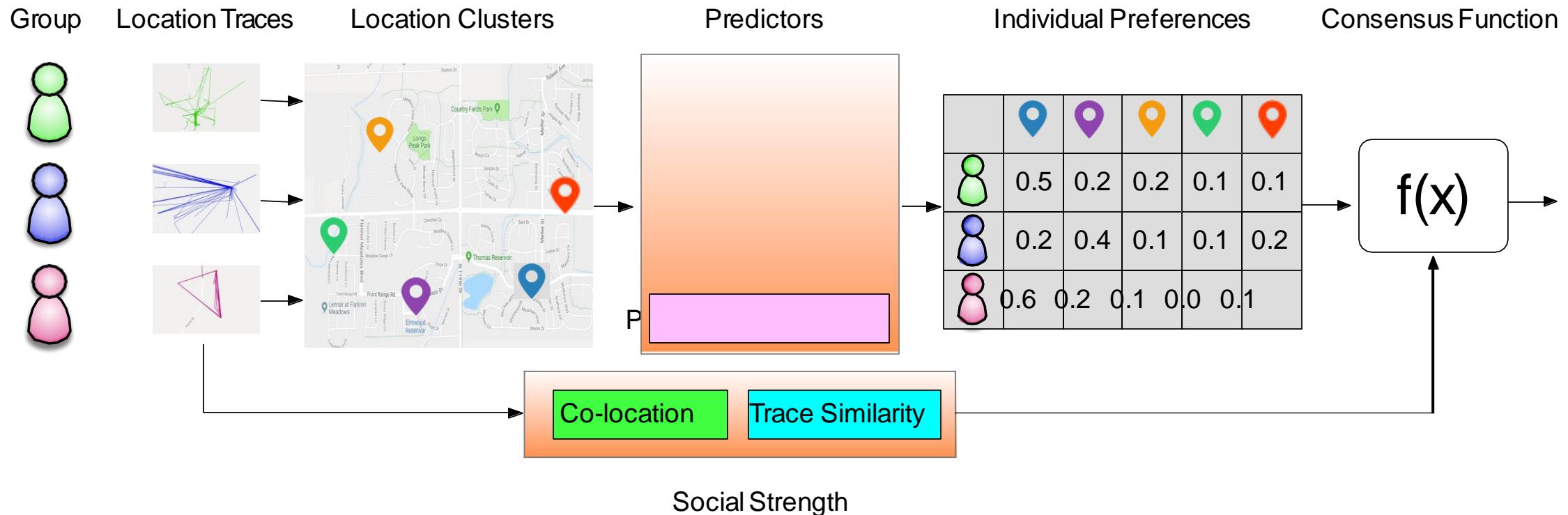
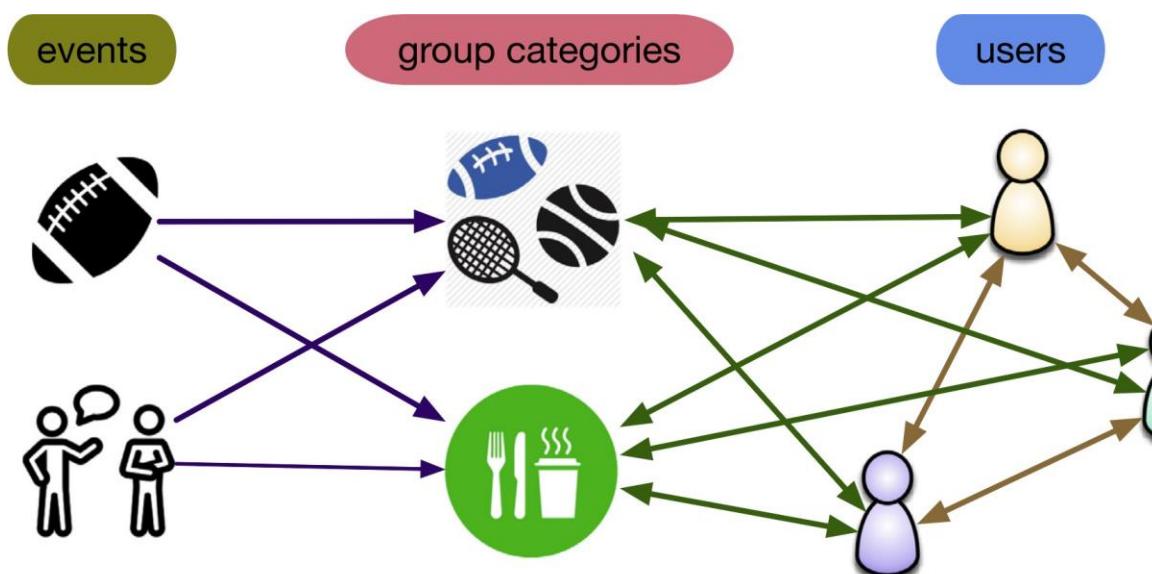
- Business intelligence
- E.g., customers, products, logistics, promotion, fraud
- Cyberspace
- E.g., service providers, online social media, security
- Pick examples of your interest



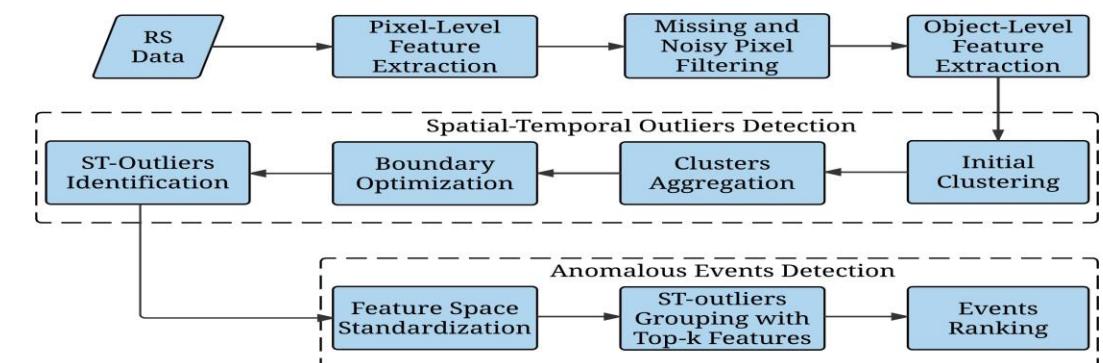
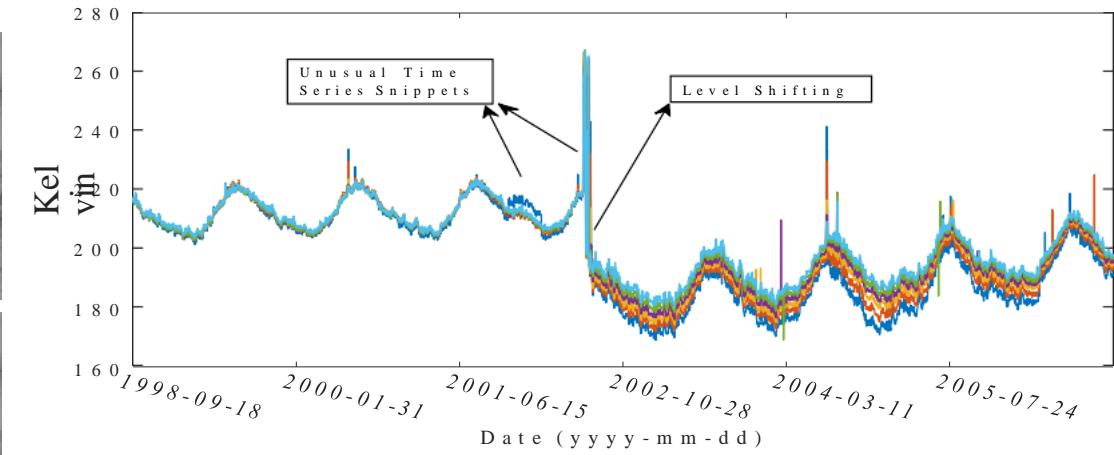
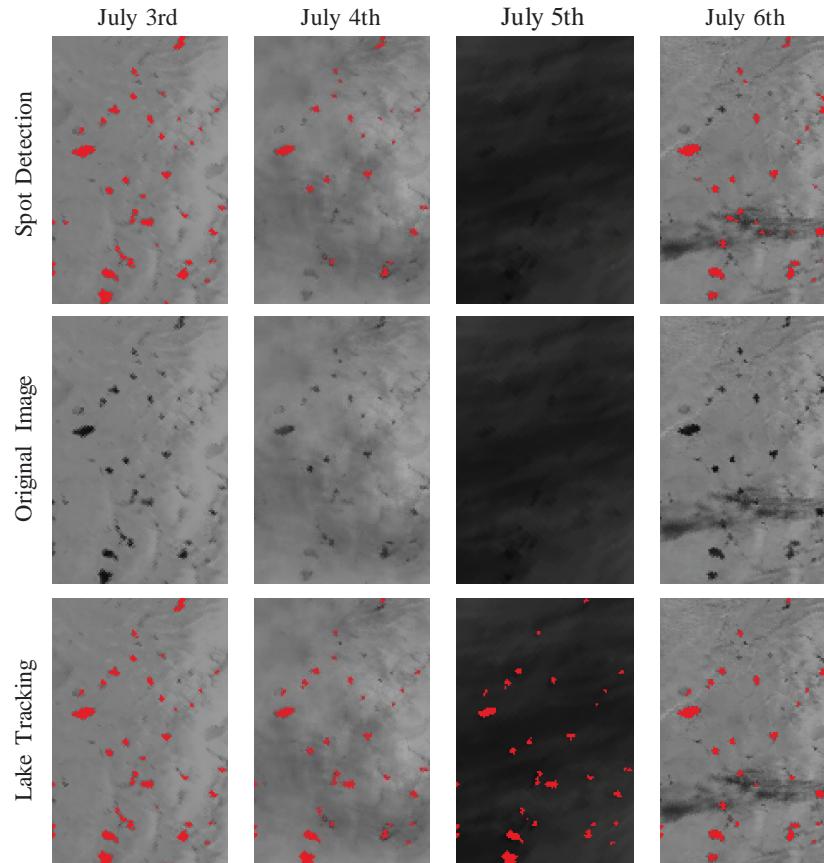
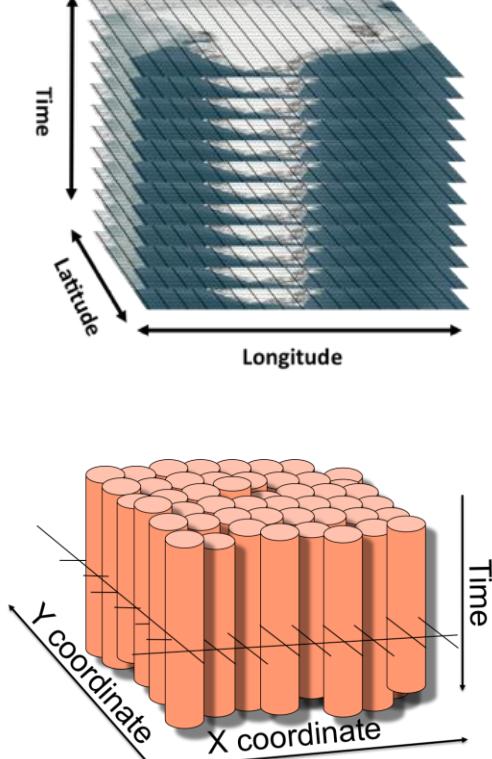
# Transportation Electrification



# Group Event Scheduling



# Remote Sensing Data Analysis





DIVERSE DATA =>  
DIVERSE  
KNOWLEDGE



DATA QUALITY  
ISSUES



SUPERVISED VS.  
UNSUPERVISED  
LEARNING



PERFORMANCE  
EVALUATION



EFFECTIVENESS VS.  
EFFICIENCY



INCREMENTAL,  
INTERACTIVE  
MINING



INTEGRATION OF  
DOMAIN  
KNOWLEDGE



VISUAL ANALYTICS



PRIVACY-  
PRESERVING  
MINING



...

# **Data Mining**

## **Understanding the data**

---

# Types of data and patterns



Data objects & attributes



Data statistics



Data visualization



Data similarity

# Dataset



A collection of data objects



E.g., employee records, product catalog, online posts



Each described by several attributes



Also referred to as features, dimensions, variables



E.g., employee: name, gender, age, salary, job title



E.g., online post: user, time, content, #likes, responses

# Dataset



A dataset is a collection of data objects.



objects could be employee records.



Say the individual objects would be each individual inquiry or this could be a particular set of products you are interested in, or any online posts, so each one would be one of the object that you're interested in.



To describe each object, There are number of attributes.



Here attributes can also refer to as features, variables, dimensions.

# What is an attribute



A property or characteristics of an object that may vary either from one object to another or from one time to another.



Ex. Eye color may vary person to person

Temperature of an object varies over time



To analyze more precisely the characteristics of object we need a measuring scale



Measurement scale: A rule (function) that associates a numerical or symbolic values with an attribute of an object

# Attribute Types

Categorical

Nominal

Binary

ordinal

- E.g., major, CS major, academic ranks

Numeric: discrete or continuous

Interval-scale or ratio-scaled (true zero)

- E.g., year 2000, number of users, annual income

## Different types of Attribute

Attribute Type	Description	Examples	Operations
Categorical (Qualitative)	Nominal  The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	zip codes, employee ID numbers, eye color, gender	mode, entropy, contingency correlation, $\chi^2$ test
	Ordinal  The values of an ordinal attribute provide enough information to order objects. $(<, >)$	hardness of minerals, $\{good, better, best\}$ , grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric (Quantitative)	Interval  For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, $t$ and $F$ tests
	Ratio  For ratio variables, both differences and ratios are meaningful. $(*, /)$	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

# TRANSFROMATION IN ATTRIBUTES



Permissible transformations



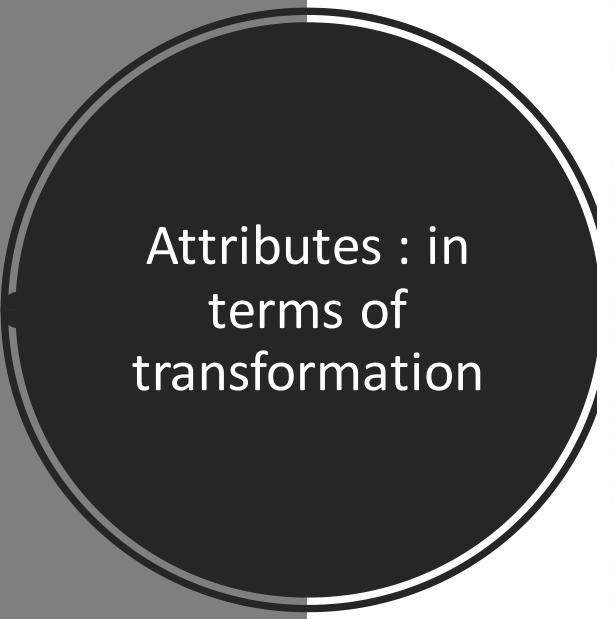
For example, the meaning of a length attribute is unchanged if it is measured in meters instead of feet.



The statistical operations that make sense for a particular type of attribute are those that will yield the same results when the attribute is transformed by using a transformation that preserves the attribute's meaning.



To illustrate, the average length of a set of objects is different when measured in meters rather than in feet, but both averages represent the same length.



## Attributes : in terms of transformation

Table 2.3. Transformations that define attribute levels.

Attribute Type		Transformation	Comment
Categorical (Qualitative)	Nominal	Any one-to-one mapping, e.g., a permutation of values	If all employee ID numbers are reassigned, it will not make any difference.
	Ordinal	An order-preserving change of values, i.e., $new\_value = f(old\_value)$ , where $f$ is a monotonic function.	An attribute encompassing the notion of good, better, best can be represented equally well by the values $\{1, 2, 3\}$ or by $\{0.5, 1, 10\}$ .
Numeric (Quantitative)	Interval	$new\_value = a * old\_value + b$ , $a$ and $b$ constants.	The Fahrenheit and Celsius temperature scales differ in the location of their zero value and the size of a degree (unit).
	Ratio	$new\_value = a * old\_value$	Length can be measured in meters or feet.

# QUESTION

- Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity

# QUESTIONS

1. Time in terms of AM or PM.
2. Brightness as measured by a light meter.
3. Brightness as measured by people's judgments.
4. Angles as measured in degrees between  $0^\circ$  and  $360^\circ$ .
5. Bronze, Silver, and Gold medals as awarded at the Olympics.
6. Height above sea level.
7. Number of patients in a hospital.
8. ISBN numbers for books. (Look up the format on the Web.)

CONTD.

1. Ability to pass light in terms of the following values: opaque, translucent, transparent.
2. Military rank
3. Distance from the center of campus.
4. Density of a substance in grams per cubic centimetre.
5. Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

# Data Statistics

---

#objects, #attributes

---

Distribution of each attribute's values

---

Categorical: % of each value

---

Numeric: central tendency, dispersion

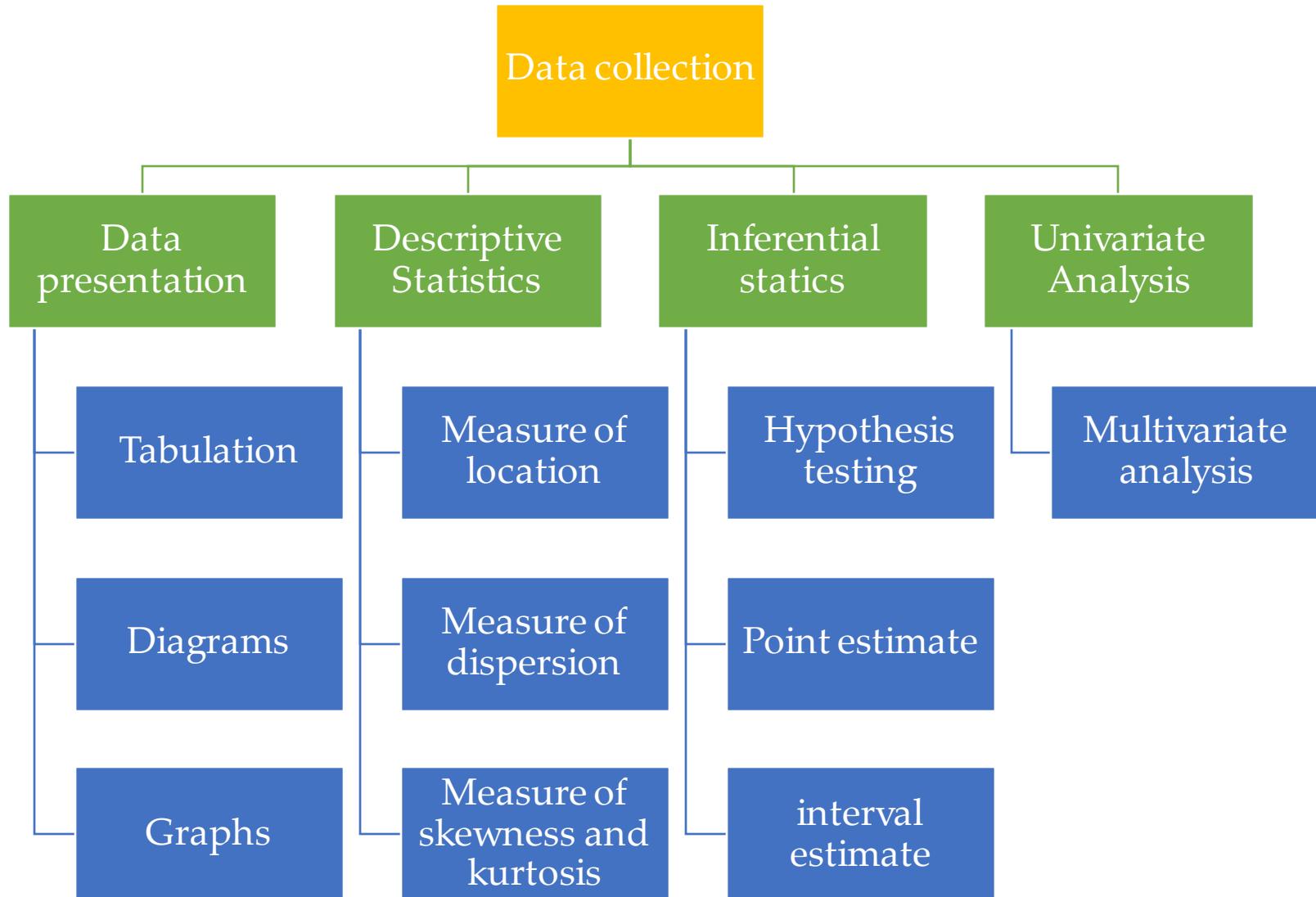
---

Comparison across attributes & datasets



Measures of Central Tendency  
or  
Measures of Location  
or  
Measures of Averages

# INVESTIGATION



# Descriptive Statistics

The goal of descriptive statistics is to summarize a collection of data in a clear and understandable way.

# Measure of Central Tendency

---

- A single summary score that best describes the central location of an entire distribution of scores.
  - The typical score.
  - The center of the distribution.
- One distribution can have multiple locations where scores cluster.
  - Must decide which measure is best for a given situation.

# Central Tendency



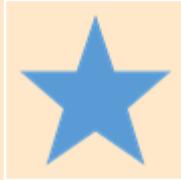
## Mean

The sum of all scores divided by the number of scores



## Median

The value that divides the distribution in half when observations are ordered.



## Mode

The most frequent score.

# Measure of central tendency ***Arithmetic Mean (Mean)***



Sum of all the observations divided by the number of the observations



The arithmetic mean is the most common measure of the central location of a sample.

$$\text{Population } \mu = \frac{\sum_{i=1}^N x_i}{N}$$
$$\text{Sample } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

# Mean

## Population

## Sample

“sigma”, the sum of X, add up all scores

$$\text{“mu”} \rightarrow \mu = \frac{\sum X}{N}$$

“N”, the total number of scores in a population

“sigma”, the sum of X, add up all scores

$$\text{“X bar”} \rightarrow \bar{X} = \frac{\sum X}{n}$$

“n”, the total number of scores in a sample

# Mean: Example

Data:

{ 1,3,6,7,2,3,5 }

- Number of observations: 7
- Sum of observations: 27
- Mean: 3.9



# Simple Frequency Distributions

raw-score distribution

name	X
Student1	20
Student2	23
Student3	15
Student4	21
Student5	15
Student6	21
Student7	15
Student8	20

frequency distribution



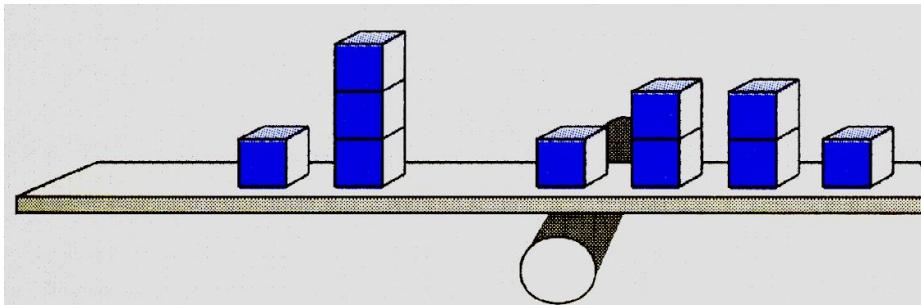
f	X
3	15
2	20
2	21
1	23

Mean

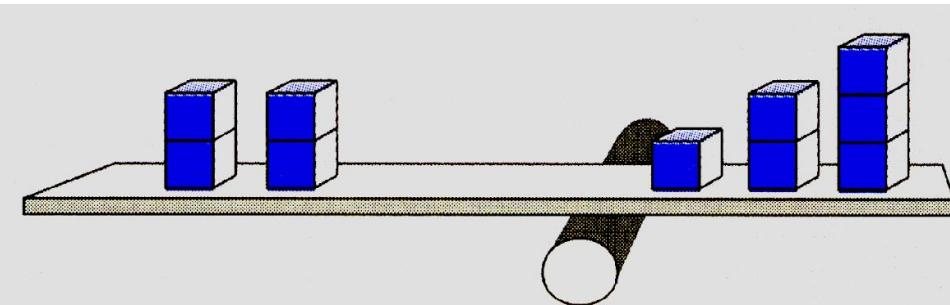
$$\bar{X} = \frac{\sum X}{f_N}$$

# Mean

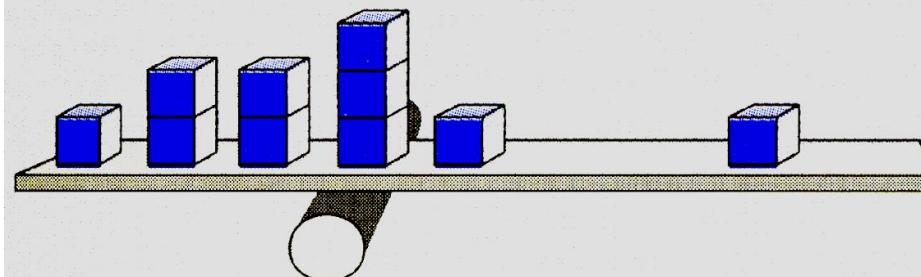
- Is the balance point of a distribution.
- The sum of negative deviations from the mean exactly equals the sum of positive deviations from the mean.



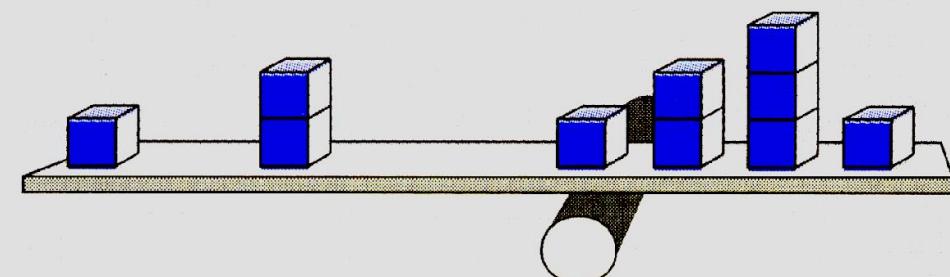
1 2 3 4 5 6 7 8 9



1 2 3 4 5 6 7 8 9



1 2 3 4 5 6 7 8 9



1 2 3 4 5 6 7 8 9

# Pros and Cons of the Mean

## *Pros*

- Mathematical center of a distribution.
- Good for interval and ratio data.
- Does not ignore any information.
- Inferential statistics is based on mathematical properties of the mean.

## *Cons*

- Influenced by extreme scores and skewed distributions.
- May not exist in the data.

# Some Important Properties of the Mean



**Interval-Ratio Level of Measurement**



**Center of Gravity**(the mean balances all the scores).



**Sensitivity to Extremes**

# *Median*

The value that is larger than half the population and smaller than half the population

*n is odd:* the median score

$$5, 8, 9, 10, 28 \quad \text{median} = 9$$

*n is even:* the  $\frac{n+1}{2}$  th score

$$6, 17, 19, 20, 21, 27 \quad \text{median} = 19.5$$

# Pros and Cons of Median

## Pros

- Not influenced by extreme scores or skewed distributions.
- Good with ordinal data.
- Easier to compute than the mean.

## Cons

- May not exist in the data.
- Doesn't take actual values into account.

# Mode

*Most frequently occurring value*

Data

{1,3,7,3,2,3,6,7}

- Mode : 3

Data

{1,3,7,3,2,3,6,7,1,1}

- Mode : 1,3

Data

{1,3,7,0,2,-3, 6,5,-1}

- Mode : none

- 52, 76, 100, 136, 186, 196, 205, 150, 257, 264, 264, 280, 282, 283, 303, 313, 317, 317, 325, 373, 384, 384, 400, 402, 417, 422, 472, 480, 643, 693, 732, 749, 750, 791, 891
- Mode: most frequent observation
- Mode(s) for hotel rates:
  - 264, 317, 384

## Pros

- Good for nominal data.
- Easiest to compute and understand.
- The score comes from the data set.

## Cons

- Ignores most of the information in a distribution.
- Small samples may not have a mode.

## Example: Central Location

Suppose the age in years of the first 10 subjects enrolled in your study are:

34, 24, 56, 52, 21, 44, 64, 34, 42, 46

Then the mean age of this group is 41.7 years

To find the median, first order the data:

21, 24, 34, 34, 42, 44, 46, 52, 56, 64

The median is  $\frac{42 + 44}{2} = 43$  years

The mode is 34 years.

# Comparison of Mean and Median



Mean is sensitive to a few very large (or small) values “outliers” so sometime mean does not reflect the quantity desired.



Median is “resistant” to outliers



Mean is attractive mathematically



50% of sample is above the median, 50% of sample is below the median.

Suppose the next patient enrolls and their age is 97 years.

How does the mean and median change?

To get the median, order the data:

21, 24, 34, 34, 42, 44, 46, 52, 56, 64, 97

If the age were recorded incorrectly as 977 instead of 97, what would the new median be?  
What would the new mean be?

# Calculating the Mean from a Frequency Distribution

# of Children(Y)	Frequency (f)	Frequency*Y(fY)
0	12	0
1	25	25
2	733	1466
3	333	999
4	183	732
5	26	130
6	15	90
7	12	84
Total	1339	3526

$$\bar{Y} = \frac{\sum fY}{N} = \frac{3526}{1339} = 2.6$$

**Table 6.1 Calculation of Arithmetic Mean for a Series of Serum Albumin Levels (g%) of 24 Pre-School Children**

2.90	3.75	3.66
3.57	3.45	3.76
3.73	3.71	3.43
3.55	3.84	3.69
3.72	3.30	3.77
3.88	3.62	3.43
2.98	3.76	3.68
3.61	3.38	3.76

The total of all these values, i.e.  $\Sigma x = 85.93$ .

Total number of observations ( $n$ ) = 24

$$\text{Therefore the arithmetic mean, } \bar{x} = \frac{\Sigma x}{n} = \frac{85.93}{24} = 3.58 \text{ g\%}$$

**Table 6.2 Calculation of Arithmetic Mean of Protein Intake of 400 Families**

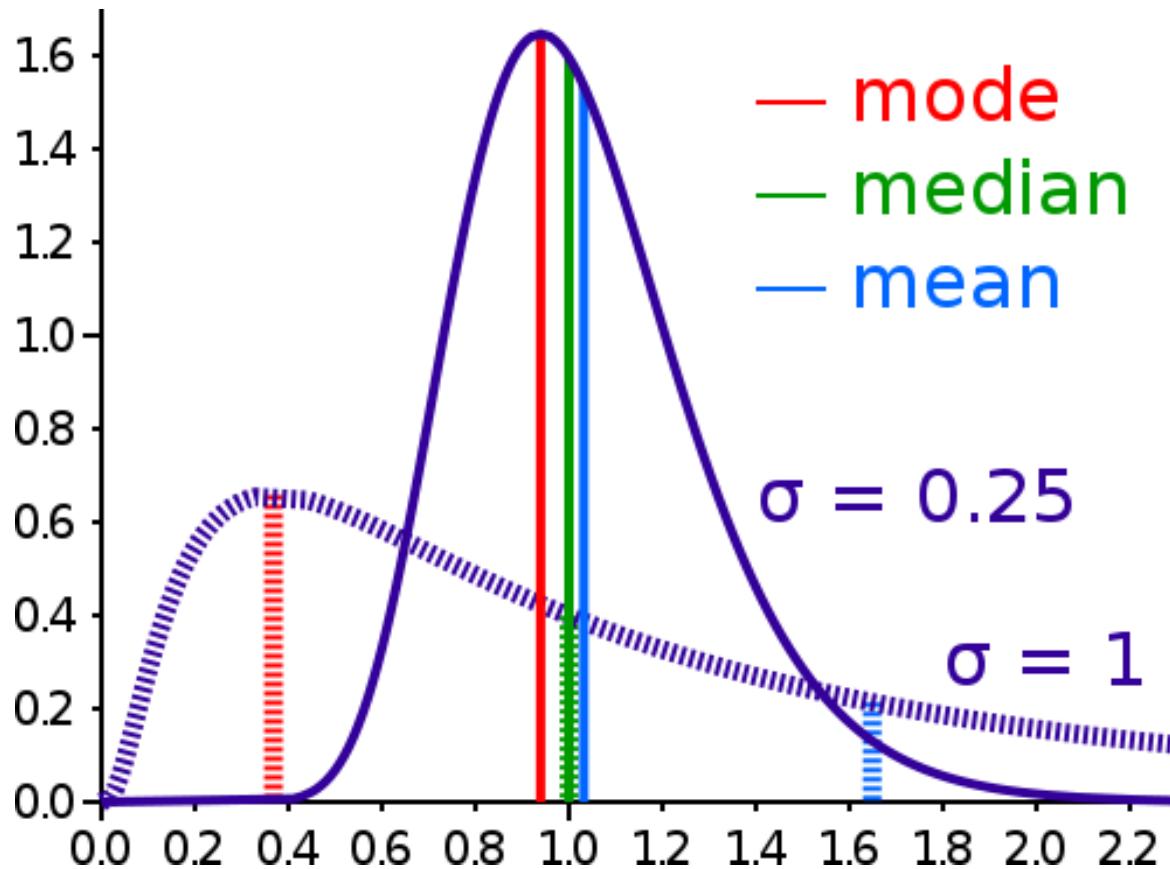
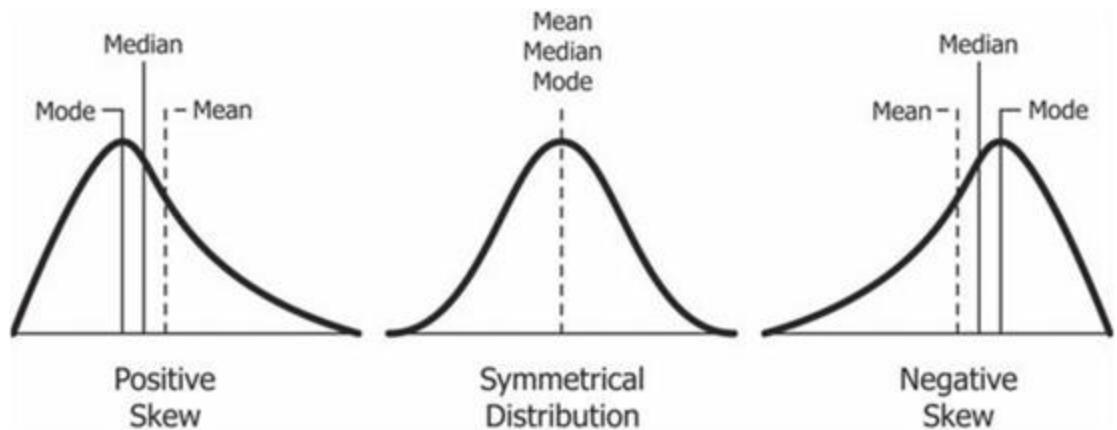
Protein intake/consumption unit/day (g)	No. of families	Midpoint of class interval	Multiply $f$ & $x$
Class interval	$f$	$x$	$fx$
15–25	30	20	600
25–35	40	30	1200
35–45	100	40	4000
45–55	110	50	5500
55–65	80	60	4800
65–75	30	70	2100
75–85	10	80	800
<b>Total</b>	<b>400</b>		<b>19000</b>

Arithmetic mean:

$$\begin{aligned}\frac{\Sigma fx}{n} &= \frac{30 \times 20 + 40 \times 30 + \dots + 10 \times 80}{400} \\ &= \frac{19000}{400} = 47.50 \text{ g}\end{aligned}$$

# Central Tendency

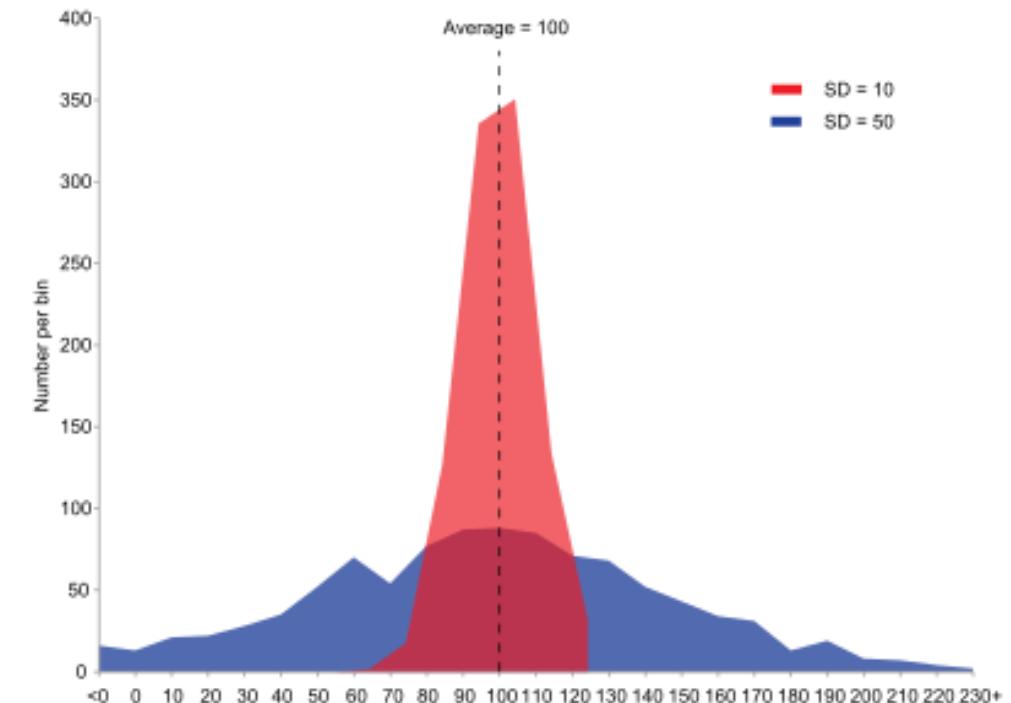
- Mean
- Median
- Mode
- Midrange
- $(\text{Max} - \text{Min})/2$



# *Dispersion*

➤ How much a distribution is stretched or squeezed

1. Range: max – min
2. Quartiles: Q1 (25%), Q3 (75%)
3. IQR (interquartile range): Q3 – Q1
4. Variance
5. Standard deviation



A yellow circular icon resembling a bomb or a sun, positioned on the left side of the slide. It has a yellow circle with a smaller yellow circle inside, and five yellow dashed lines radiating from the top right of the inner circle.

# Measures of Dispersion

# Definition

- *Measures of dispersion* are descriptive statistics that describe how similar a set of scores are to each other
- The more similar the scores are to each other, the lower the measure of dispersion will be
- The less similar the scores are to each other, the higher the measure of dispersion will be
- In general, the more spread out a distribution is, the larger the measure of dispersion will be

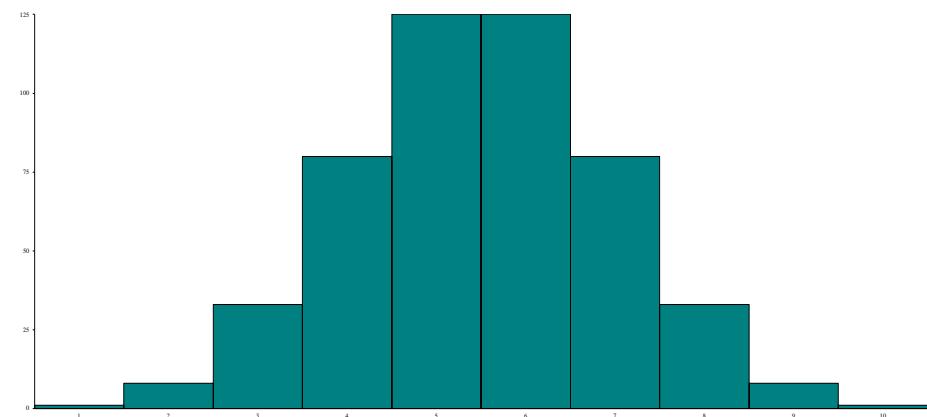
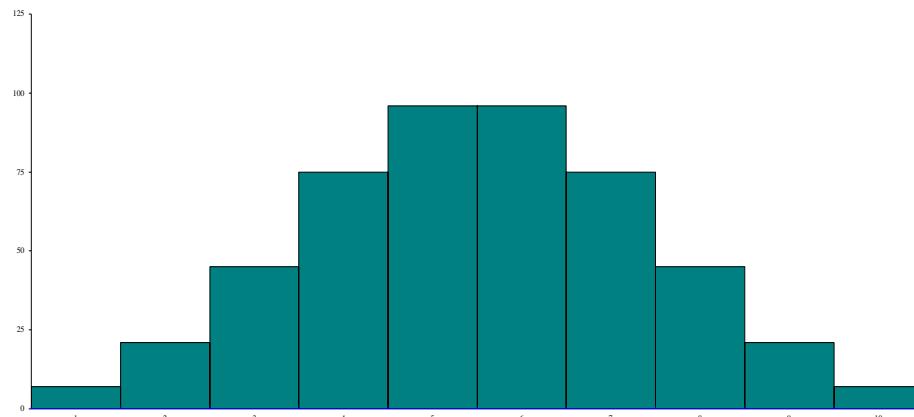
# Two types **Measures of Dispersion**

Absolute  
Measure of  
Dispersion

Relative  
Measure of  
Dispersion

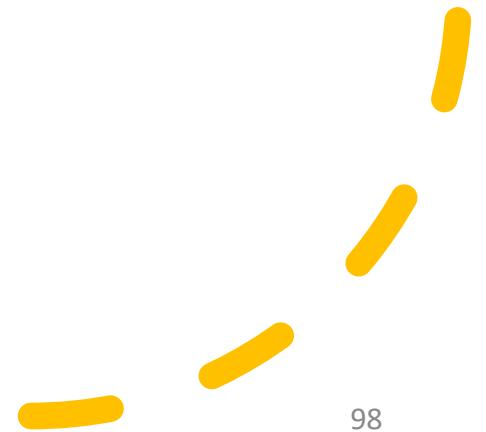
# Absolute Measures of Dispersion

- Which of the distributions of scores has the larger dispersion?
- The upper distribution has more dispersion because the scores are more spread out
- That is, they are less like each other



# Measures of Dispersion

- There are three main measures of dispersion:
  - The range
  - The semi-interquartile range (SIR)
  - Variance / standard deviation



# The Range

- The *range* is defined as the difference between the largest score in the set of data and the smallest score in the set of data,  $X_L - X_S$
- What is the range of the following data:  
4 8 1 6 6 2 9 3 6 9
- The largest score ( $X_L$ ) is 9; the smallest score ( $X_S$ ) is 1; the range is  $X_L - X_S = 9 - 1 = 8$

# When To Use the Range

- The range is used when
  - you have ordinal data or
  - you are presenting your results to people with little or no knowledge of statistics
- The range is rarely used in scientific work as it is fairly insensitive
  - It depends on only two scores in the set of data,  $X_L$  and  $X_S$
  - Two very different sets of data can have the same range:  
1 1 1 1 9 vs 1 3 5 7 9

---

# The Semi-Interquartile Range

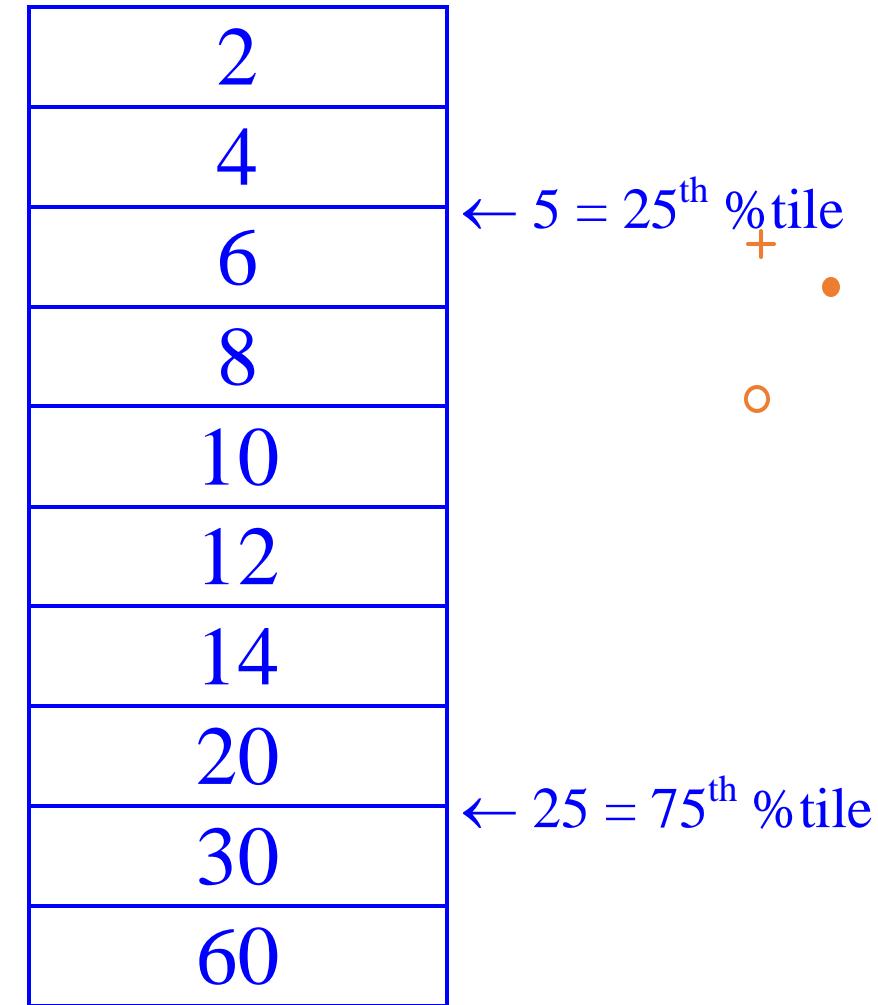
- The *semi-interquartile range* (or *SIR*) is defined as the difference of the first and third quartiles divided by two
  - The first quartile is the 25<sup>th</sup> percentile
  - The third quartile is the 75<sup>th</sup> percentile

$$SIR = (Q_3 - Q_1) / 2$$



# SIR Example

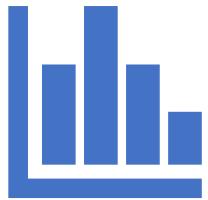
- What is the SIR for the data to the right?
- 25 % of the scores are below 5
  - 5 is the first quartile
- 25 % of the scores are above 25
  - 25 is the third quartile
- $\text{SIR} = (Q_3 - Q_1) / 2 = (25 - 5) / 2 = 10$



# When To Use the SIR

The SIR is often used with skewed data as it is insensitive to the extreme scores





# Variance

*Variance* is defined as the average of the square deviations:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

---

# What Does the Variance Formula Mean?

- First, it says to subtract the mean from each of the scores
  - This difference is called a *deviate* or a *deviation score*
  - The deviate tells us how far a given score is from the typical, or average, score
  - Thus, the deviate is a measure of dispersion for a given score

+ .  
o

---

# What Does the Variance Formula Mean?

- Why can't we simply take the average of the deviates?
- One of the definitions of the *mean* was that it always made the sum of the scores minus the mean equal to 0
- Thus, the average of the deviates must be 0 since the sum of the deviates must equal 0
- To avoid this problem, statisticians square the deviate score prior to averaging them
  - Squaring the deviate score makes all the squared scores positive



---

# What Does the Variance Formula Mean?

- Variance is the mean of the squared deviation scores
- The larger the variance is, the more the scores deviate, on average, away from the mean
- The smaller the variance is, the less the scores deviate, on average, from the mean



---

# Standard Deviation

- When the deviate scores are squared in variance, their unit of measure is squared as well
  - E.g. If people's weights are measured in pounds, then the variance of the weights would be expressed in pounds<sup>2</sup> (or squared pounds)
- Since squared units of measure are often awkward to deal with, the square root of variance is often used instead
  - The standard deviation is the square root of variance



# Standard Deviation

---

Standard deviation =  $\sqrt{\text{variance}}$

Variance = standard deviation<sup>2</sup>

# Computational Formula

---

- When calculating variance, it is often easier to use a computational formula which is algebraically equivalent to the definitional formula:

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N} = \frac{\sum (X - \mu)^2}{N}$$

*$\sigma^2$  is the population variance,  $X$  is a score,  $\mu$  is the population mean, and  $N$  is the number of scores*

# Computational Formula Example

X	X <sup>2</sup>	X-μ	(X-μ) <sup>2</sup>
9	81	2	4
8	64	1	1
6	36	-1	1
5	25	-2	4
8	64	1	1
6	36	-1	1
$\Sigma = 42$	$\Sigma = 306$	$\Sigma = 0$	$\Sigma = 12$

# Computational Formula Example

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$$
$$= \frac{306 - \frac{42^2}{6}}{6}$$
$$= \frac{306 - 294}{6}$$
$$= \frac{12}{6}$$
$$= 2$$

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$
$$= \frac{12}{6}$$
$$= 2$$

# Relative Measure of Dispersion

- The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:
  1. Co-efficient of Range
  2. Co-efficient of Variation
  3. Co-efficient of Standard Deviation
  4. Co-efficient of Quartile Deviation
  5. Co-efficient of Mean Deviation

# Co-efficient of Dispersion

- The coefficients of dispersion are calculated (along with the measure of dispersion) when two series are compared, that differ widely in their averages.
- The dispersion coefficient is also used when two series with different measurement units are compared.
- It is denoted as C.D.
- The common coefficients of dispersion are:

# Co-efficient of Dispersion

C.D. in terms of	Coefficient of dispersion
Range	$C.D. = (X_{\max} - X_{\min}) / (X_{\max} + X_{\min})$
Quartile Deviation	$C.D. = (Q_3 - Q_1) / (Q_3 + Q_1)$
Standard Deviation (S.D.)	$C.D. = S.D. / \text{Mean}$
Mean Deviation	$C.D. = \text{Mean deviation}/\text{Average}$

# Data Visualization

Boxplot

Histogram

Quantile  
plot

Q-Q plot

Scatter  
plot

# Data Visualization Methods



Visualizing complex data & relations



Chart type: line, pie, (stacked) bar,  
bubble, area, heatmap, word cloud,  
network, ...

Color, size, layout, hierarchy, ...

Exploration vs. explanation

Automation, interaction, efficiency & effectiveness

# Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository  
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length



Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

# Visualization



**Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.**



**Visualization of data is one of the most powerful and appealing techniques for data exploration.**

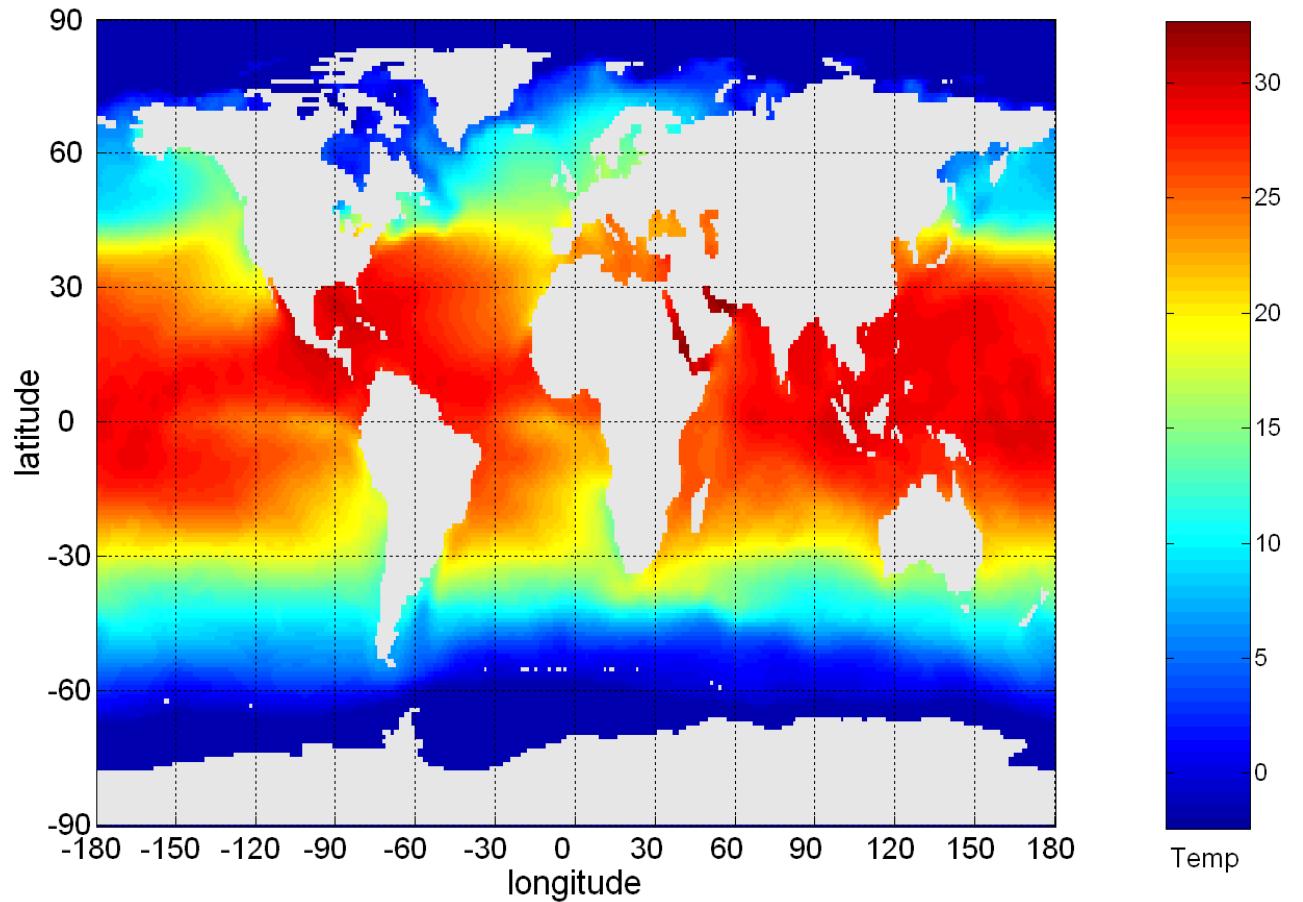
Humans have a well developed ability to analyze large amounts of information that is presented visually

Can detect general patterns and trends

Can detect outliers and unusual patterns

# Example: Sea Surface Temperature

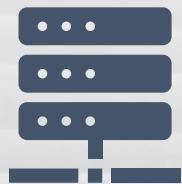
- The following shows the Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points are summarized in a single figure



# Representation



**Is the mapping of  
information to a visual  
format**



**Data objects, their  
attributes, and the  
relationships among  
data objects are  
translated into graphical  
elements such as points,  
lines, shapes, and  
colors.**



**Example:**

Objects are often represented as points

Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape

If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived

# Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:

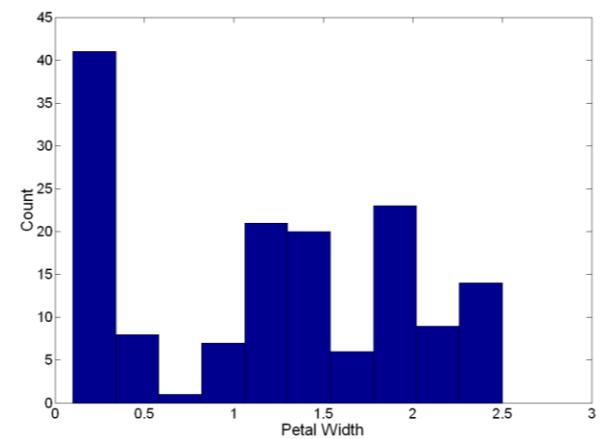
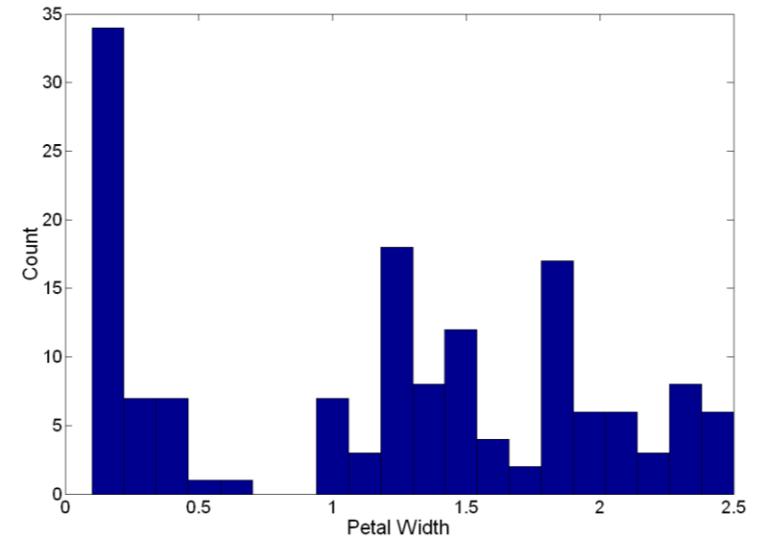
	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

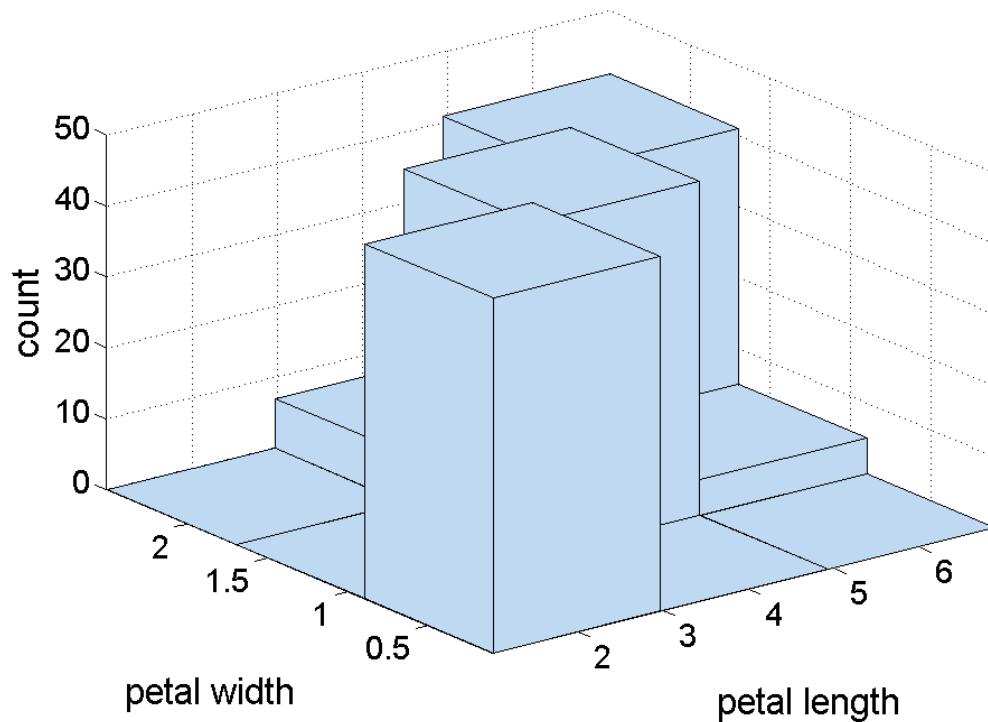
# Visualization Techniques: Histograms

---

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)



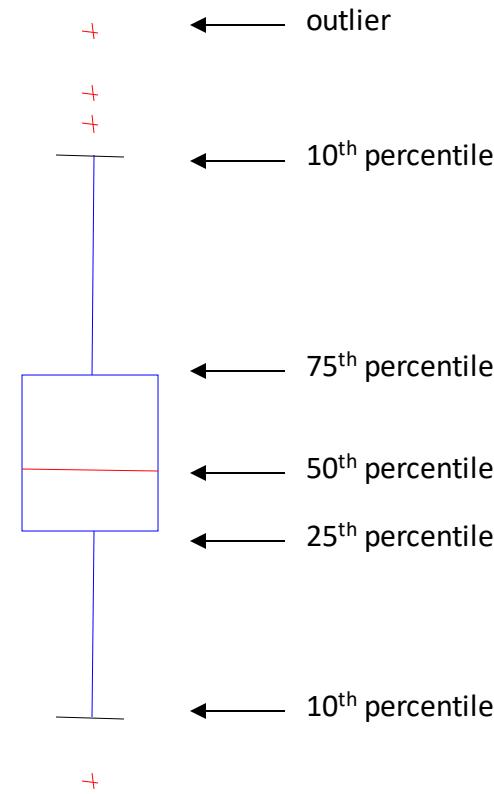
# Two-Dimensional Histograms



- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?

# Visualization Techniques: Box Plots

- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



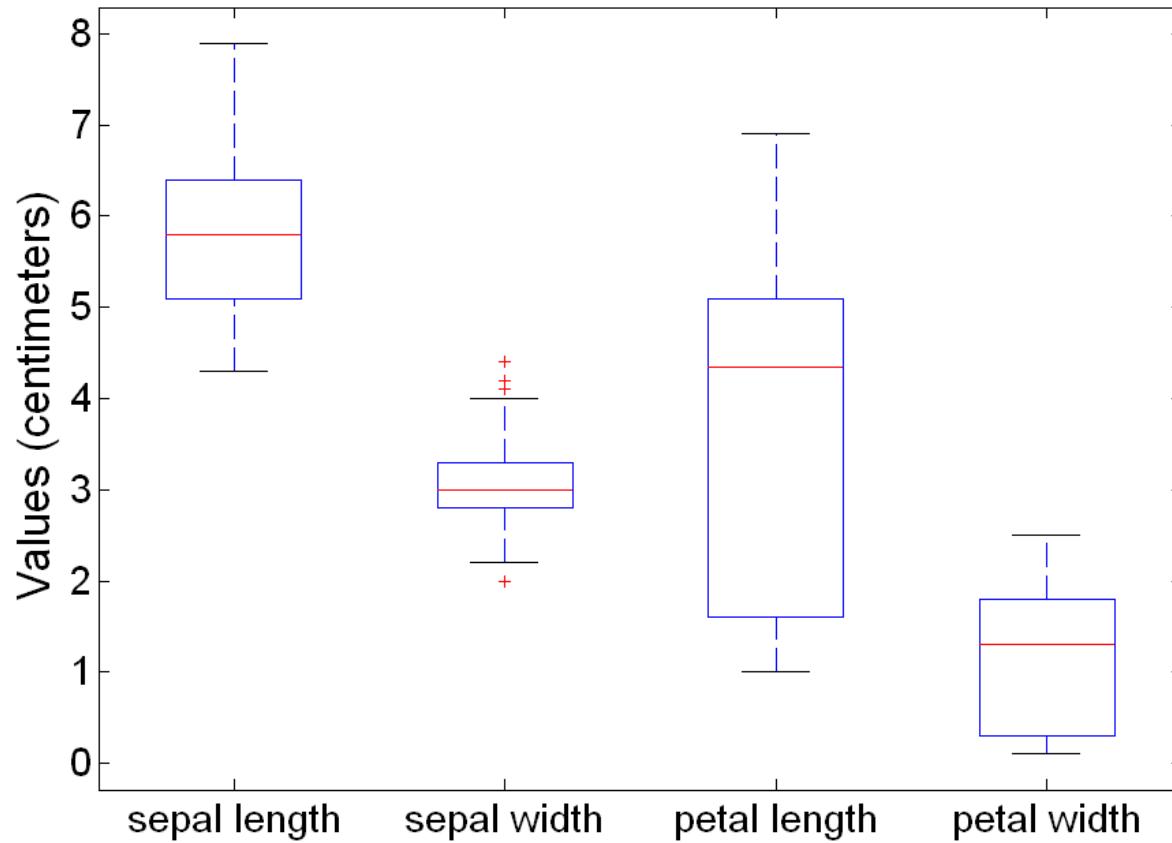
# Boxplot

- **Box**
  - Q1, Q2, Q3, IQR
- **Whiskers**
  - Min, max,
  - $1.5 \times \text{IQR}$
- **Outliers**



# Example of Box Plots

- Box plots can be used to compare attributes



# Visualization Techniques: Scatter Plots



Attributes values determine the position



Two-dimensional scatter plots most common,  
but can have three-dimensional scatter plots

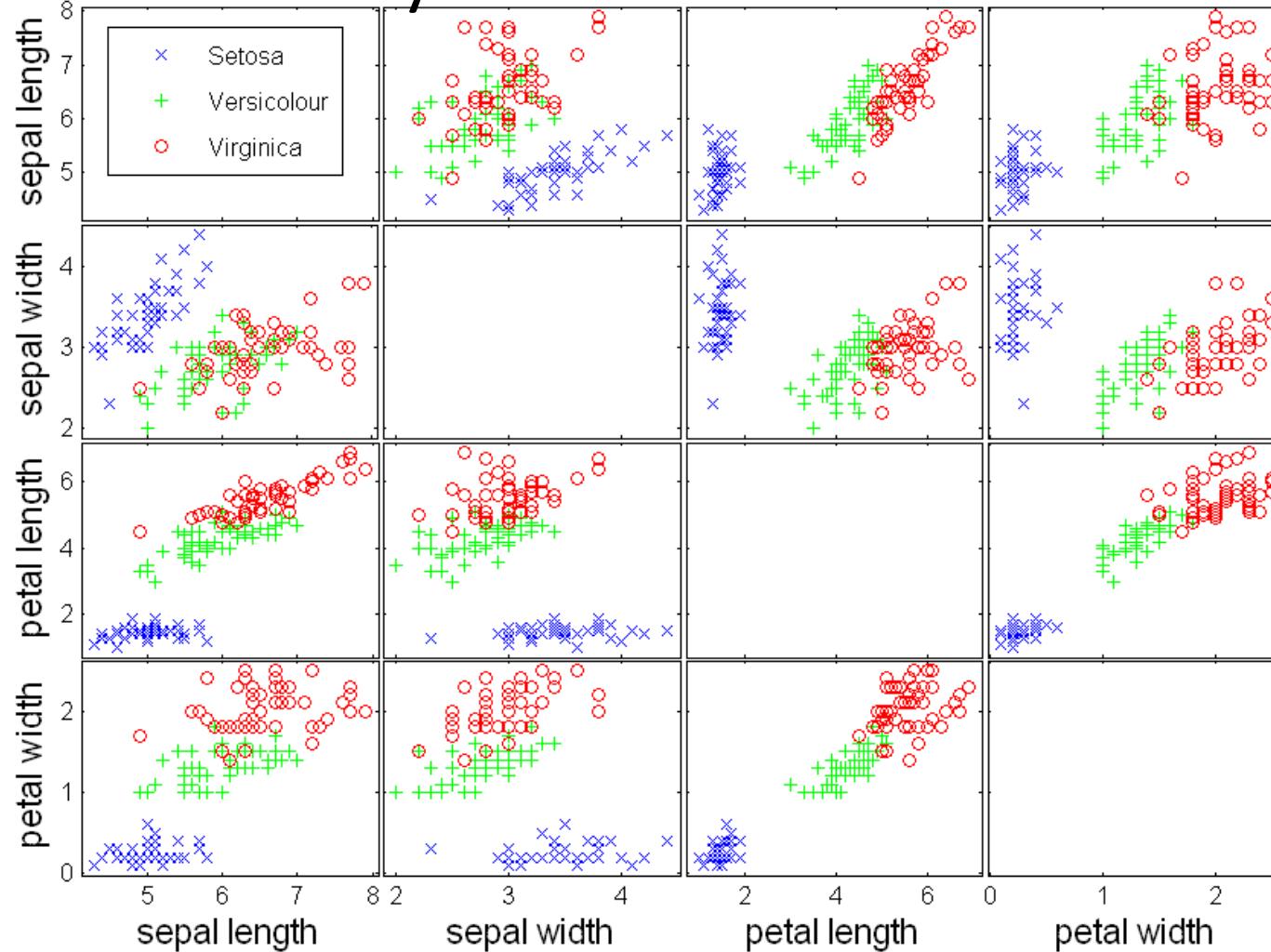


Often additional attributes can be displayed by  
using the size, shape, and color of the markers  
that represent the objects



It is useful to have arrays of scatter plots can  
compactly summarize the relationships of  
several pairs of attributes

# Scatter Plot Array of Iris Attributes



# Visualization Techniques: Contour Plots

---

Useful when a continuous attribute is measured on a spatial grid

---

They partition the plane into regions of similar values

---

The contour lines that form the boundaries of these regions connect points with equal values

---

The most common example is contour maps of elevation

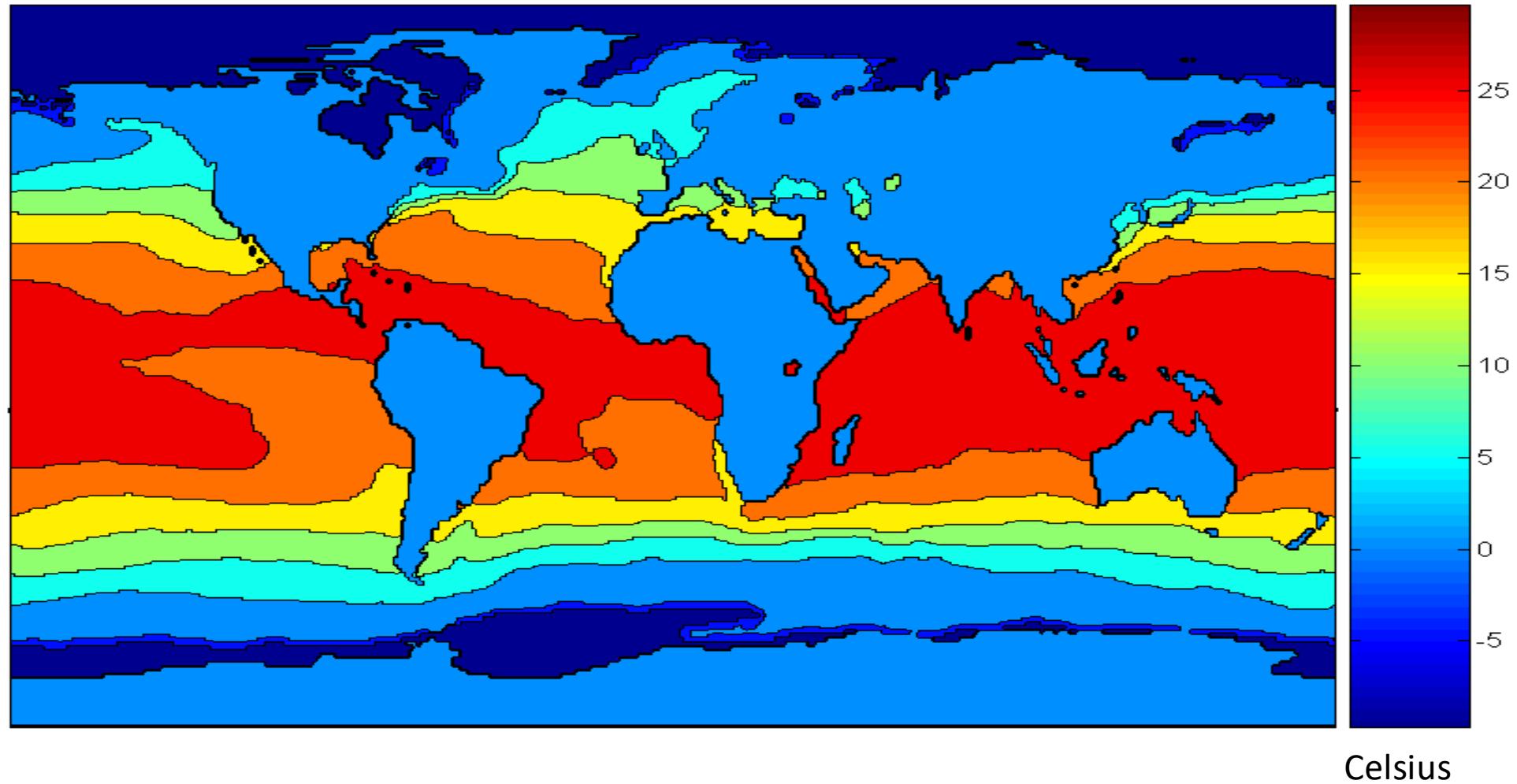
---

Can also display temperature, rainfall, air pressure, etc.

---

An example for Sea Surface Temperature (SST) is provided on the next slide

# Contour Plot Example: SST Dec, 1998



# Visualization Techniques: Matrix Plots

---

Can plot the data matrix

---

This can be useful when objects are sorted according to class

---

Typically, the attributes are normalized to prevent one attribute from dominating the plot

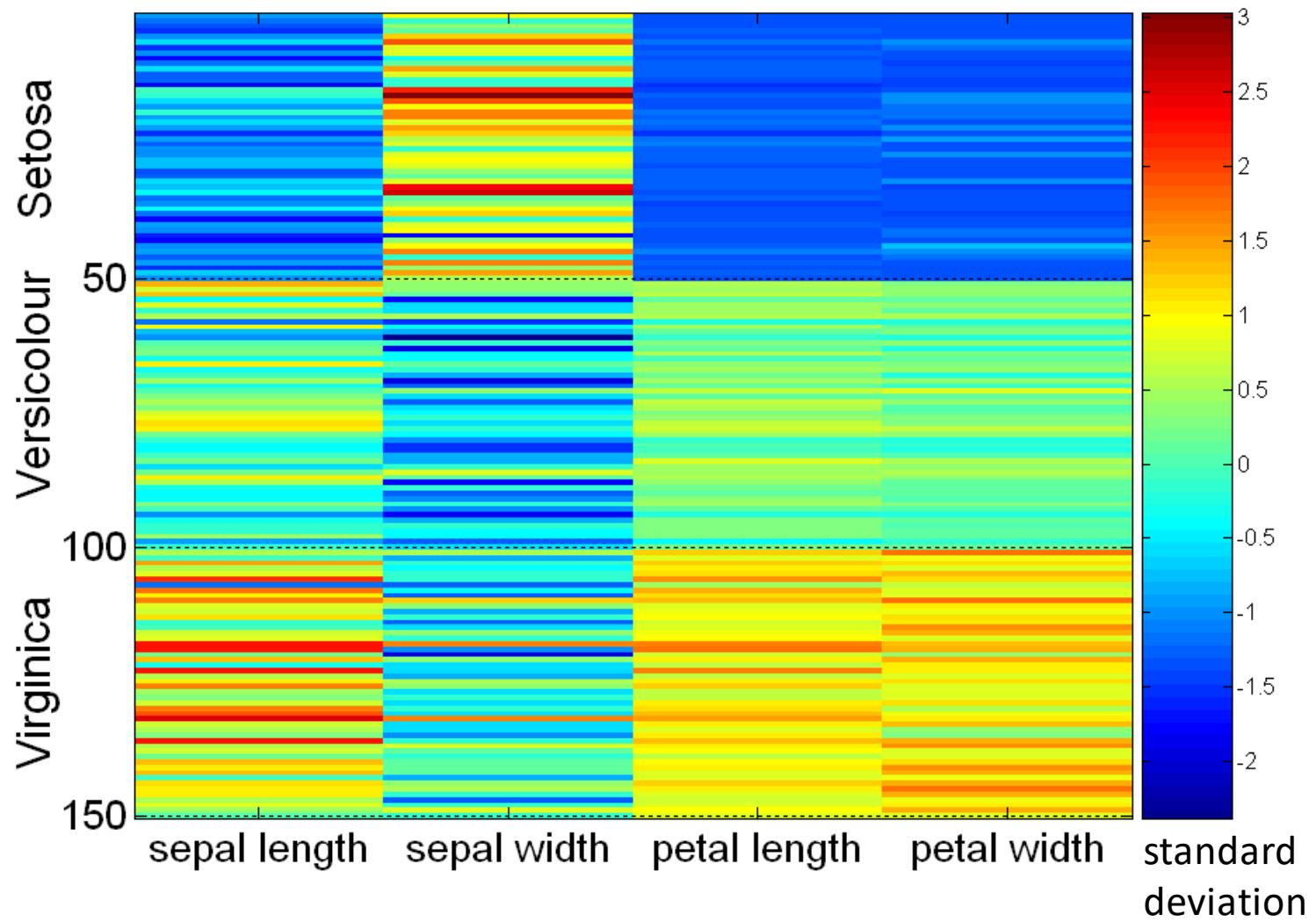
---

Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects

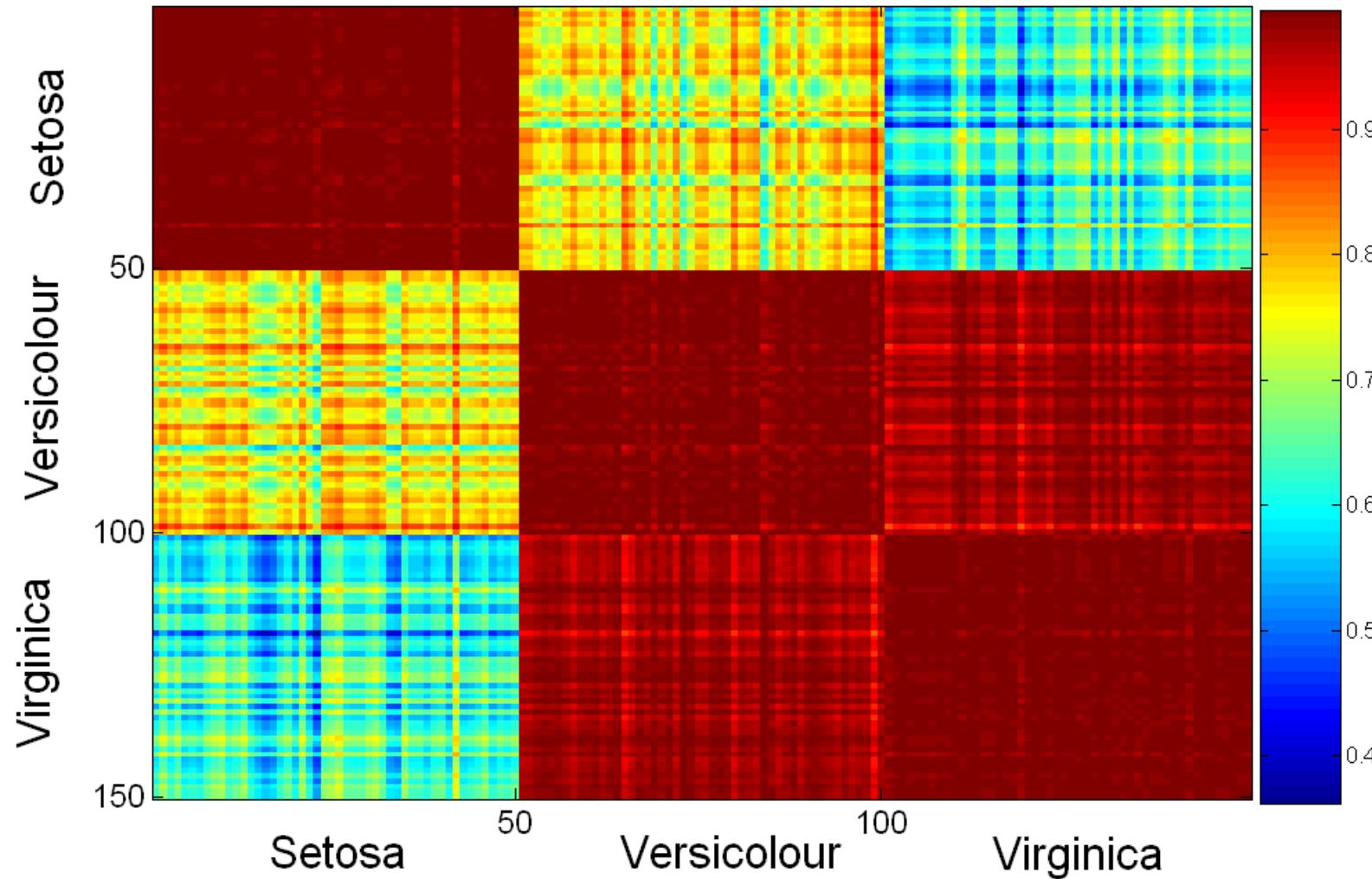
---

Examples of matrix plots are presented on the next two slides

# Visualization of the Iris Data Matrix



# Visualization of the Iris Correlation Matrix



# Visualization Techniques: Parallel Coordinates



Used to plot the attribute values of high-dimensional data



Instead of using perpendicular axes, use a set of parallel axes



The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line



Thus, each object is represented as a line

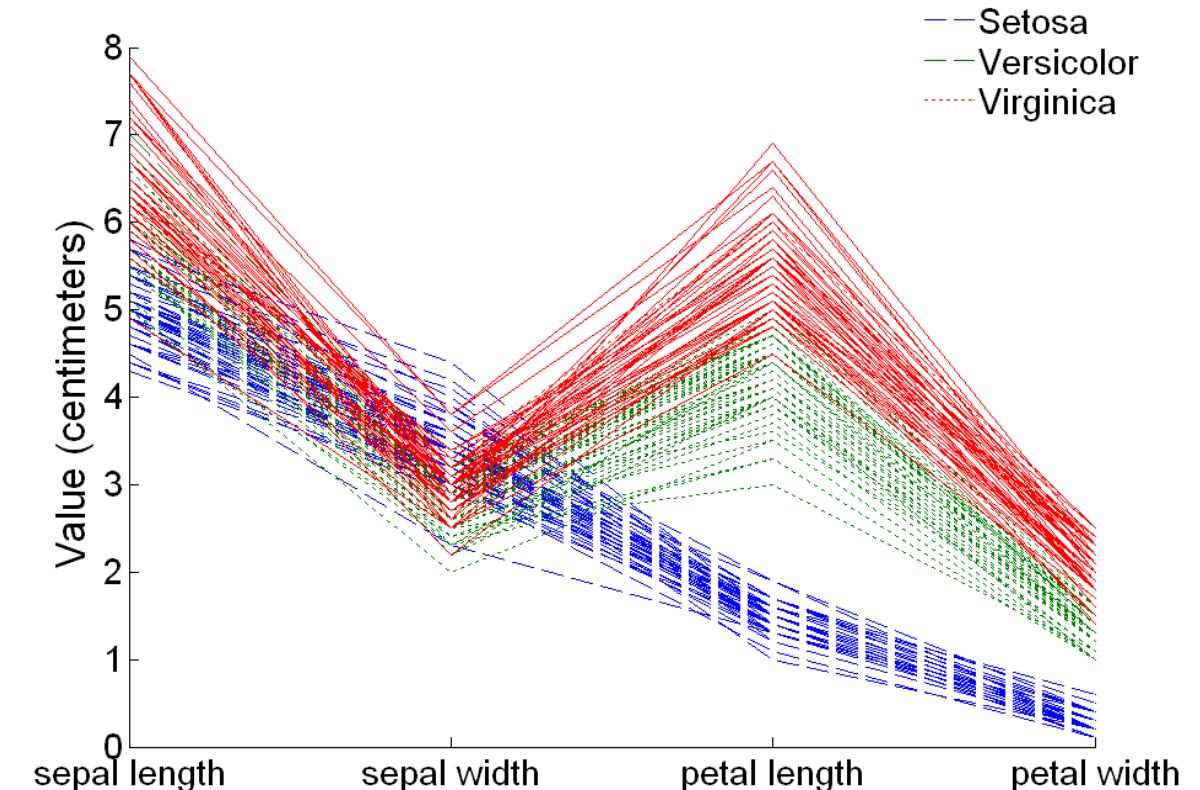
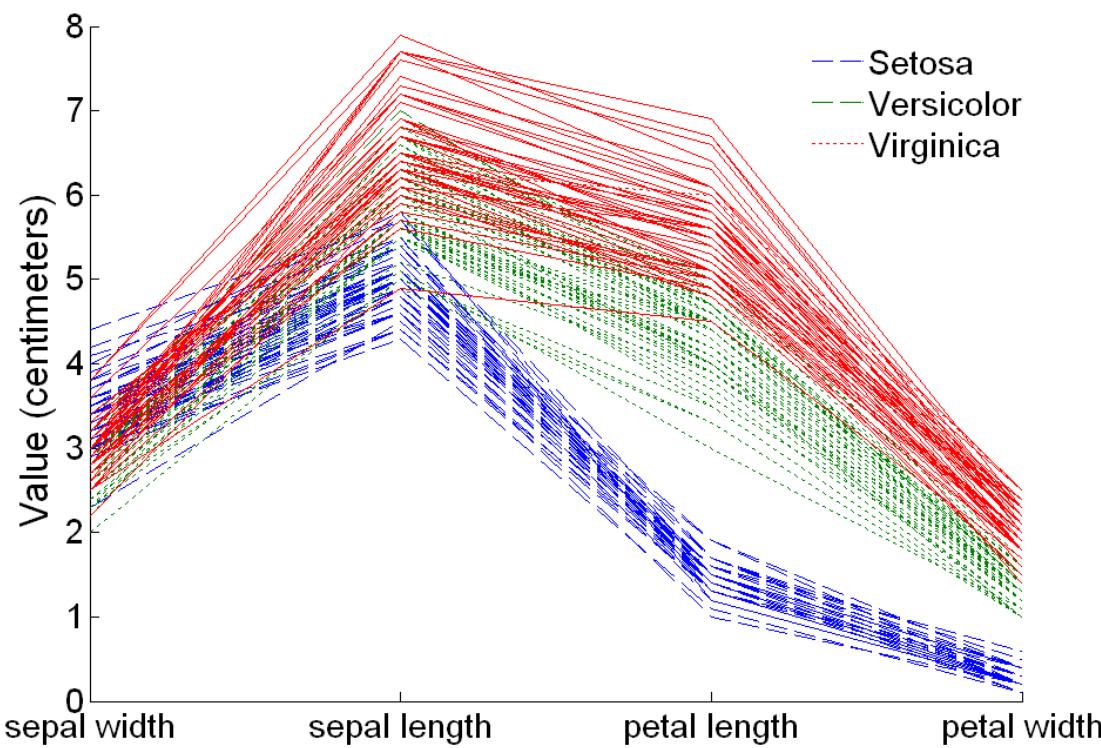


Often, the lines representing a distinct class of objects group together, at least for some attributes



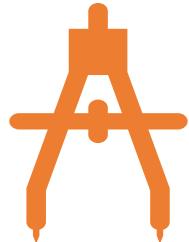
Ordering of attributes is important in seeing such groupings

# Parallel Coordinates Plots for Iris Data



# Other Visualization Techniques

---



## Star Plots

Similar approach to parallel coordinates, but axes radiate from a central point

The line connecting the values of an object is a polygon



## Chernoff Faces

Approach created by Herman Chernoff

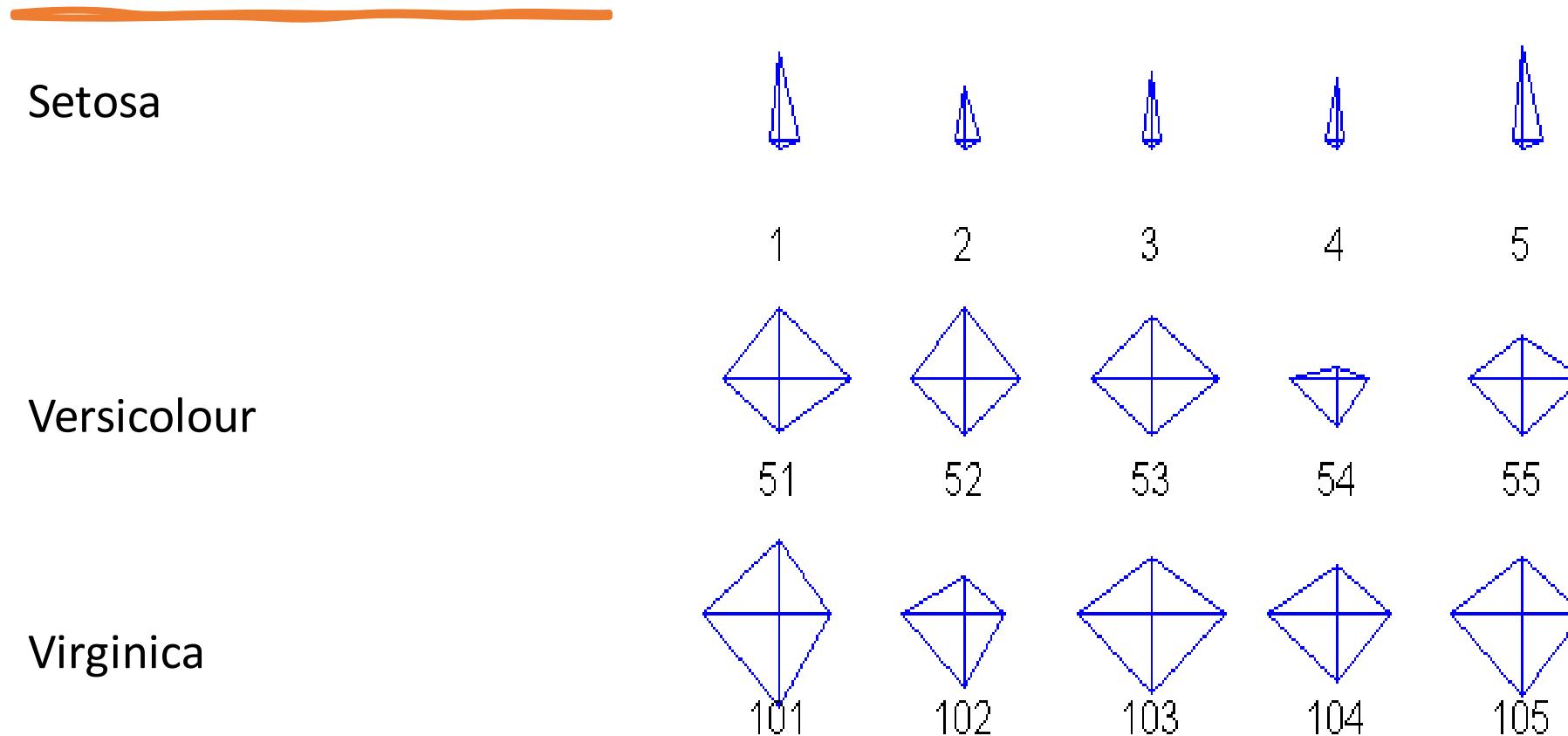
This approach associates each attribute with a characteristic of a face

The values of each attribute determine the appearance of the corresponding facial characteristic

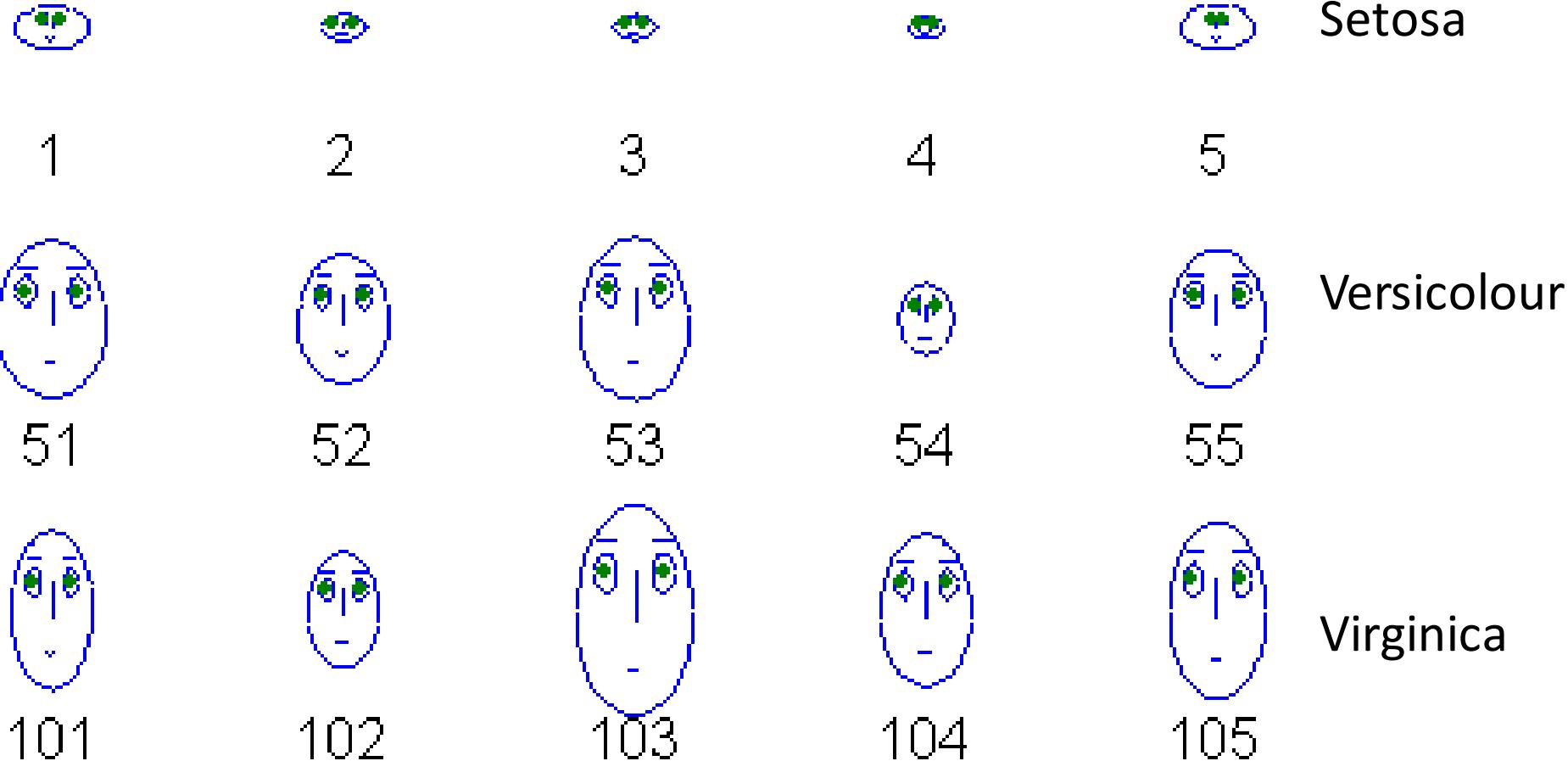
Each object becomes a separate face

Relies on human's ability to distinguish faces

# Star Plots for Iris Data



# Chernoff Faces for Iris Data



# Data Pre-Processing

# Data Pre-Processing

Data Objects and Attribute Types

Measuring Data Similarity and Dissimilarity,

Why Pre-process the Data

Data Cleaning

Data Integration

Data Reduction

Data Transformation

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, data cubes, or files



## Data reduction

Dimensionality reduction  
Numerosity reduction  
Data compression

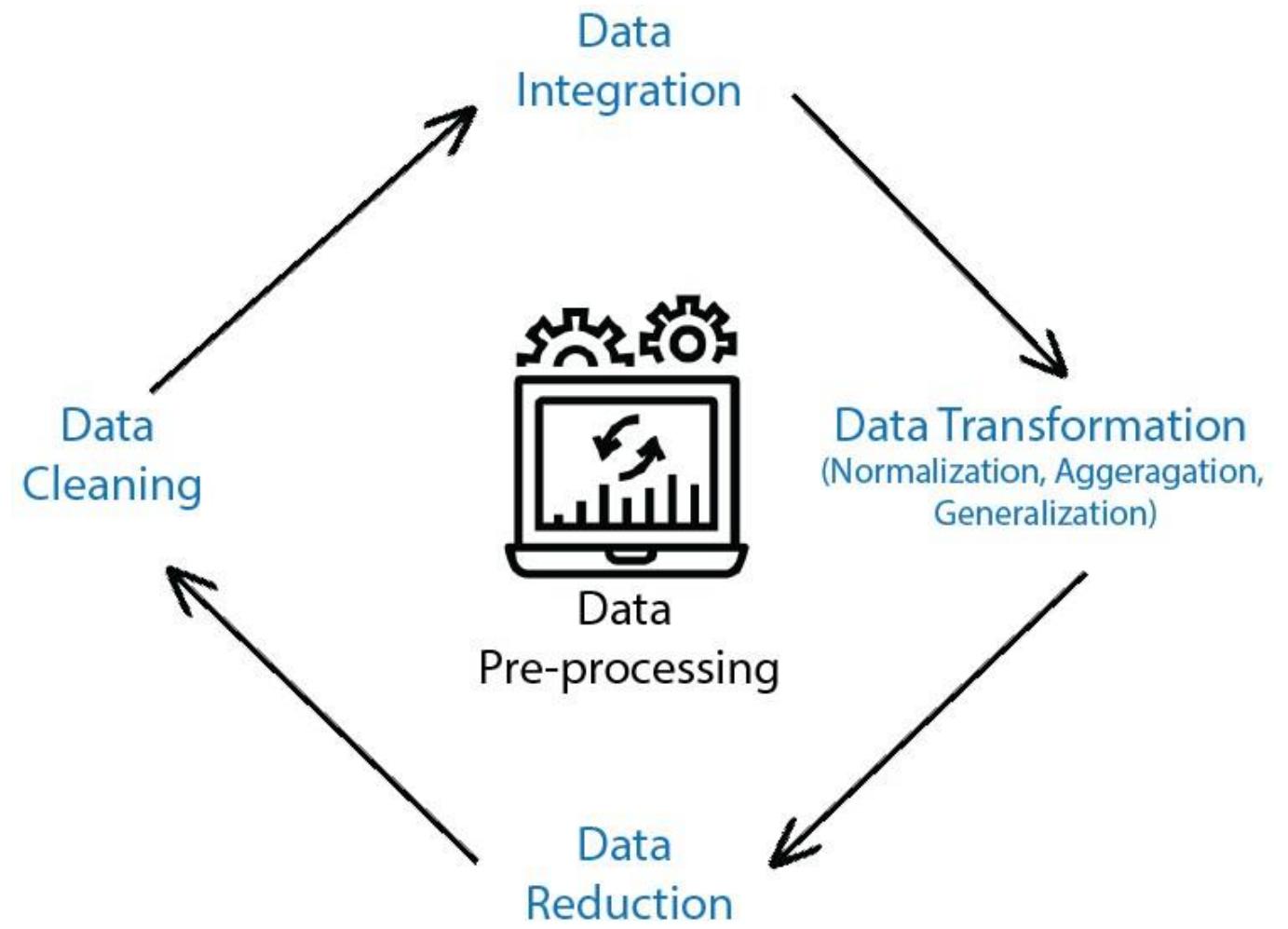


## Data transformation and data discretization

Normalization  
Concept hierarchy generation

# Data Preprocessing

- Potential issues with data
- E.g., missing data, errors, inconsistency, availability
- Preparing data for the mining process
- Data cleaning, integration, transformation, reduction
- No good data, no good data mining!



# Data Quality



# Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, ...
  - Consistency: some modified but some not, dangling, ...
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

# Issues in Real-world Data

Incomplete

Missing values,  
missing  
attributes

Noisy

Imprecision,  
errors, outliers:  
e.g., age = “-10”

Inconsistent

E.g., age vs.  
birthday, rating  
scale

# Causes of Data Issues



Data collection/transmission/processing



Human, hardware, and software



Limitations, errors, multiple sources



Changes over time



Updated survey, new sensing capabilities

# 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

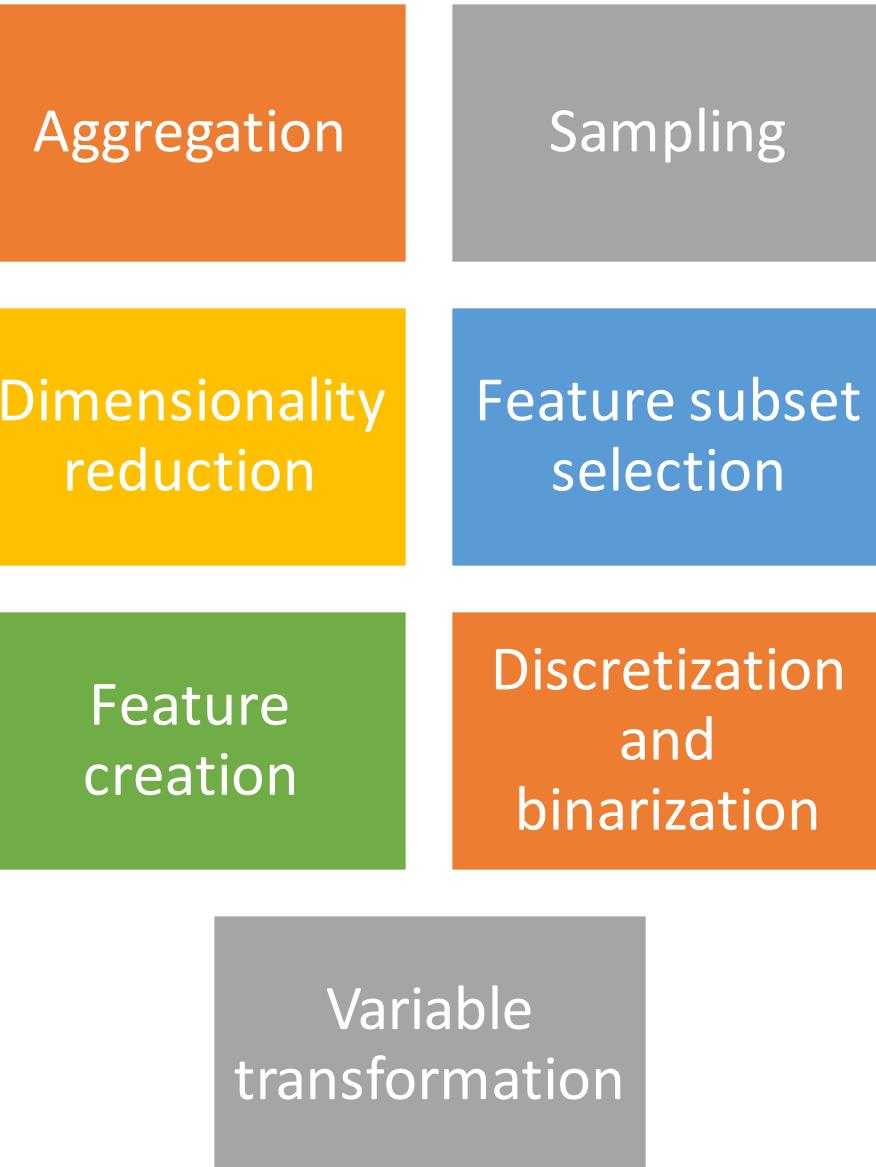
**Intentional** (e.g., *disguised missing data*)

- Jan. 1 as everyone’s birthday?

# DATA PREPROCESSING

In two aspects:

1. selecting data objects and attributes for the analysis
2. for creating/changing the attributes.



# Aggregation

- The combining of two or more objects into a single object.
- One way to aggregate transactions for this data set is to replace all the transactions of a single store with a single storewide transaction.
- This reduces the hundreds or thousands of transactions that occur daily at a specific store to a single daily transaction, and the number of data objects per day is reduced to the number of stores.

**Table 2.4.** Data set containing information about customer purchases.

Transaction ID	Item	Store Location	Date	Price	...
:	:	:	:	:	
101123	Watch	Chicago	09/06/04	\$25.99	...
101123	Battery	Chicago	09/06/04	\$5.99	...
101124	Shoes	Minneapolis	09/06/04	\$75.00	...
:	:	:	:	:	

# Interpretation

The data can also be viewed as a multidimensional array, where each attribute is a dimension.

From this viewpoint, aggregation is the process of eliminating attributes,

Such as the type of item, or reducing the number of values for a particular attribute

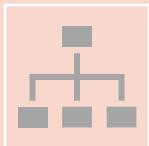
e.g., reducing the possible values for date from 365 days to 12 months.

This type of aggregation is commonly used in Online Analytical Processing (OLAP).

# Advantages of aggregation



The smaller data sets resulting from data reduction require less memory and processing time



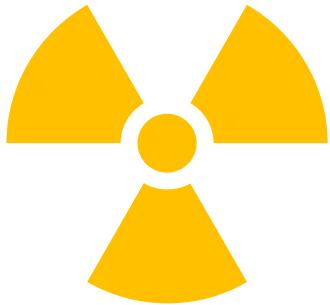
Aggregation can act as a change of scope or scale by providing a high-level view of the data instead of a low-level view.

In the previous example, aggregating over store locations and months gives us a monthly, per store view of the data instead of a daily, per item view.

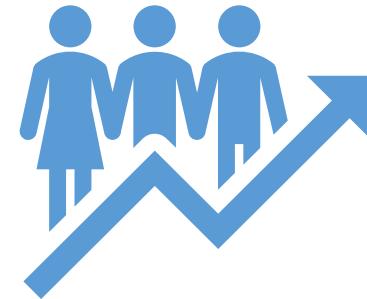


the behaviour of groups of objects or attributes is often more stable than that of individual objects or attributes.

# Disadvantage of aggregation



The potential loss of interesting details.



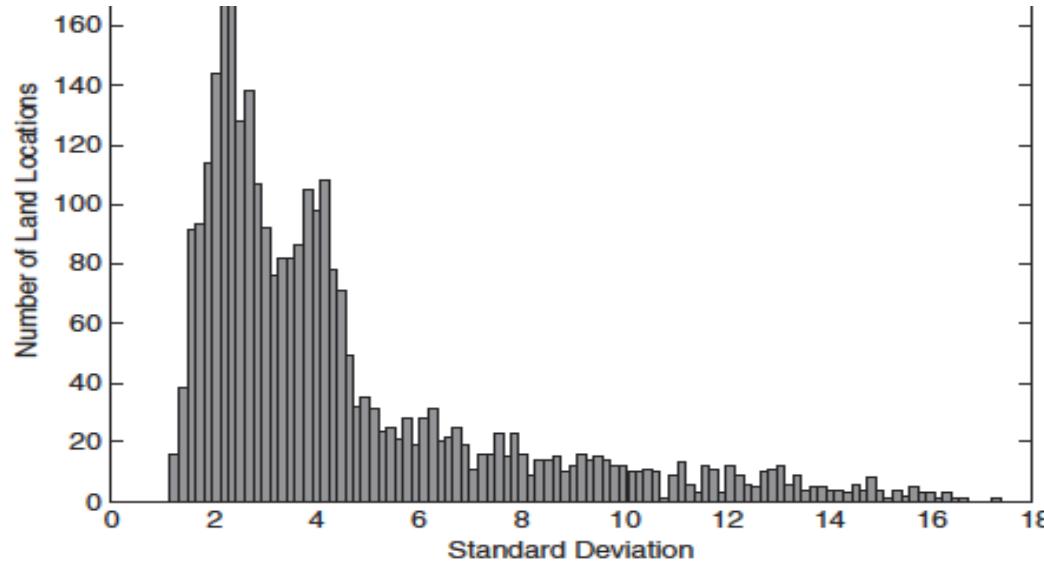
In the store example, aggregating over months loses information about which day of the week has the highest sales.

# Example- Australian Precipitation

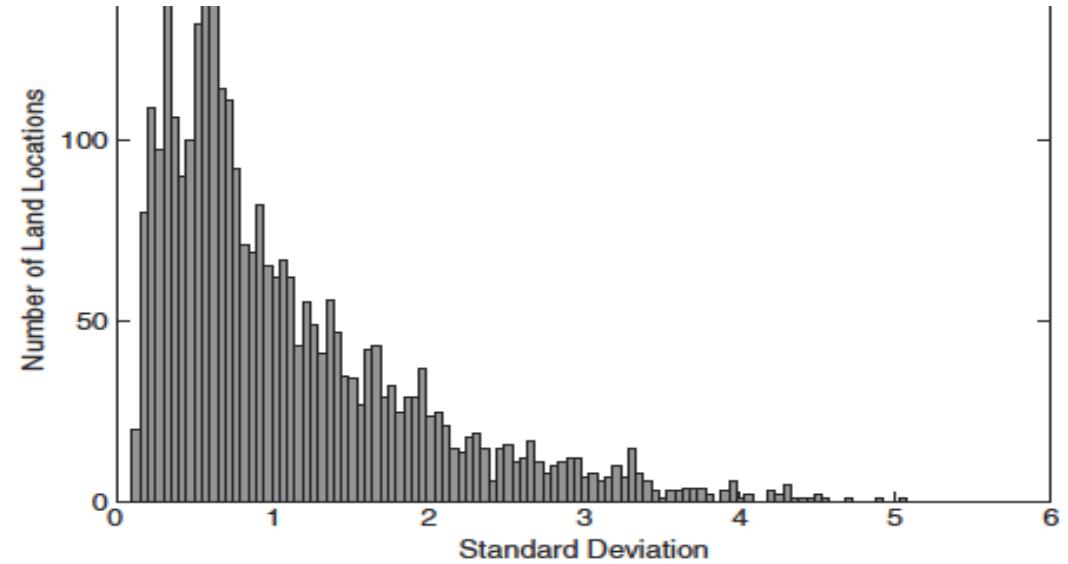
- This example is based on precipitation in Australia from the period 1982–1993.
- Figure 2.8(a) shows a histogram for the standard deviation of average monthly precipitation for 3,030  $0.5^\circ$  by  $0.5^\circ$  grid cells in Australia
- Figure 2.8(b) shows a histogram for the standard deviation of the average yearly precipitation for the same locations. The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimetres.



# Example- Australian Precipitation



(a) Histogram of standard deviation of average monthly precipitation



(b) Histogram of standard deviation of average yearly precipitation

**Figure 2.8.** Histograms of standard deviation for monthly and yearly precipitation in Australia for the period 1982–1993.

# Incomplete (Missing) Data

Data is not always available

- E.g., 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

# 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*

# Noisy Data

Noise: random error or variance 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 require data cleaning

- duplicate records
- incomplete data
- inconsistent data

# 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)

# Data Cleaning as a Process

## Data discrepancy detection

- Use metadata (e.g., domain, range, dependency, distribution)
- Check field overloading
- Check uniqueness rule, consecutive rule and null rule
- 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 migration and integration

- Data migration tools: allow transformations to be specified
- ETL (Extraction/Transformation>Loading) tools: allow users to specify transformations through a graphical user interface

## Integration of the two processes

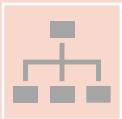
- Iterative and interactive (e.g., Potter's Wheels)

# Data Integration



## Data integration:

Combines data from multiple sources into a coherent store



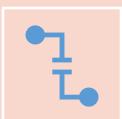
Schema integration: e.g.,  
 $A.cust-id \equiv B.cust\#$

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

# Handling Redundancy in Data Integration

Redundant data occur often when integration of multiple databases

- *Object identification:* The same attribute or object may have different names in different databases
- *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., annual revenue

Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*

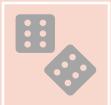
Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

# Correlation Analysis (Nominal Data)



## $\chi^2$ (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$



The larger the  $\chi^2$  value, the more likely the variables are related



The cells that contribute the most to the  $\chi^2$  value are those whose actual count is very different from the expected count



Correlation does not imply causality

# of hospitals and # of car-theft in a city are correlated  
Both are causally linked to the third variable: population