



# DATA ANALYTICS

## Course Overview

---

**Gowri Srinivasa**

Department of Computer Science  
and Engineering

# DATA ANALYTICS

Perhaps you would like to know...

---



- What is data analytics?
- How is this different from analysis?
- What does it involve?
- Why is this important?
- What will we learn as a part of this course?
- What is the evaluation policy for this course?

## What is data analytics?

---

The *science* of examining raw data

- To elicit patterns and
- Develop insights
- To interpret relationships between variables
- And draw conclusions
- That support decision-making

## How is analytics different from analysis?

---

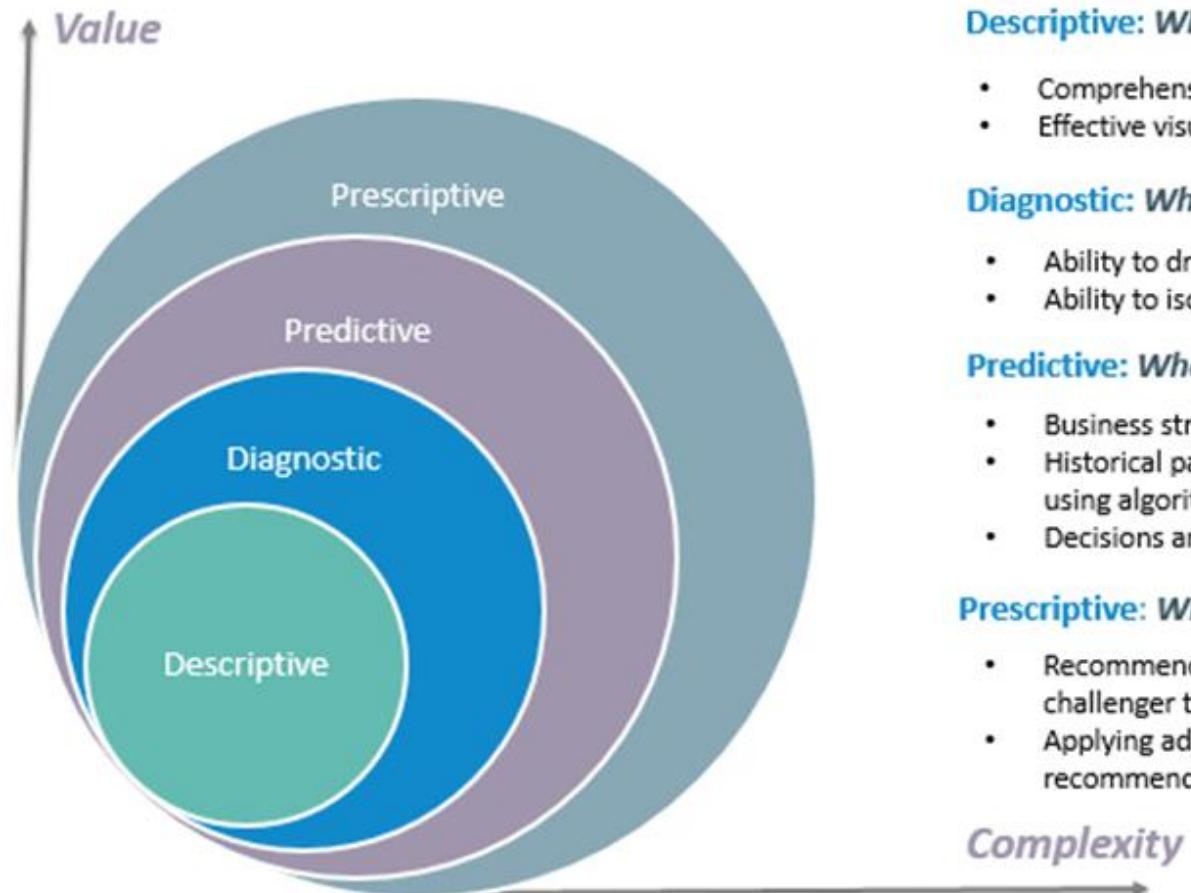
Data Analytics involves analysis

- Analysis – examining and interpreting past data
- Analytics = Analysis + Predictive modeling  
or Forecasting

Use an understanding of the past  
to be able to say something about the future

## What does it involve?

### 4 types of Data Analytics



### What is the data telling you?

#### Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

#### Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

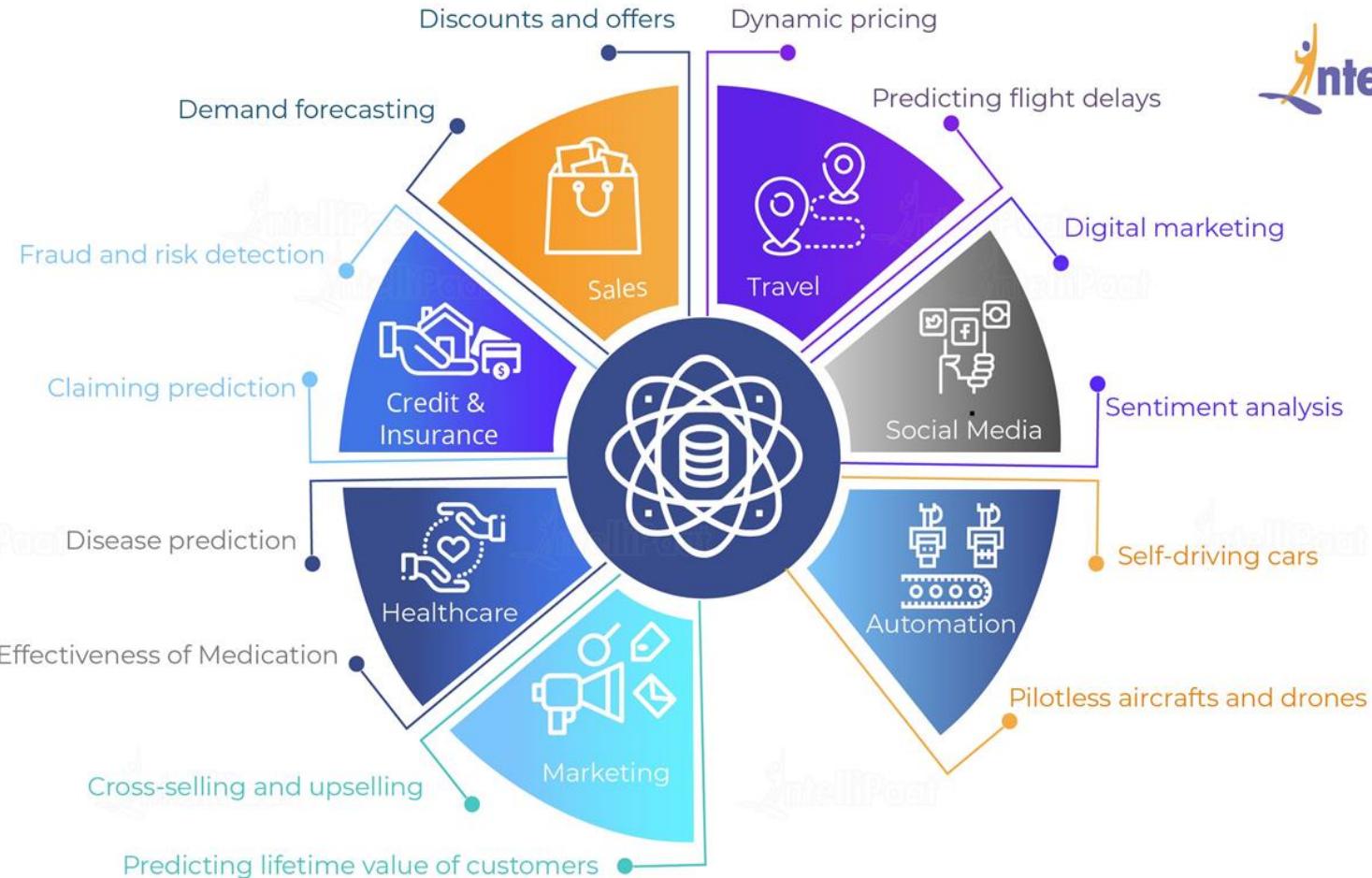
#### Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

#### Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

## Why is it important?



## Top stories

---



**Big Data Analytics Software Market 2020 with COVID-19 ...**

3rd Watch News - 4 hours ago

Big Data Analytics Software Market 2020 with COVID-19 Impact Analysis, Industry Growth, Future Trends, Key Players and Regional Outlook ...

Bandera County ...



**Brazil adopts real-time analytics to present Covid-19 data**

ZDNet - 22 hours ago

The Brazilian government has launched a new portal powered by real-time data analytics to report the evolving picture on the Covid-19 ...



**Big Data Analytics In Banking Market with Imapct of COVID19 ...**

3rd Watch News - 12-Jun-2020

Big Data Analytics In Banking Market with Imapct of COVID19 By Top Keyplayers IBM, Oracle, SAP SE, Microsoft, HP. June 12, 2020 ...

**Global Big Data Analytics in Healthcare Market with COVID-19 ...**

Bulletin Line - 11-Jun-2020

[View all](#)

## Top stories-contd.,

ETCIO

### How data analytics can be leveraged to realize personalized care management

Analytics will, in the near future, aid physicians to effectively identify patients who are most likely to benefit from such care programs, to make informed decisions and to manage their care and lower costs



ETCIO

### Shriram General Insurance's journey from analysis to analytics

The insurance firm is looking at ways to expand the business, improve claim processing by going deeper into machine learning and neural networks.

ETCIO

### Banking analytics trends to look in 2020

The future challenge is to develop analytic strengths that span the organisation and not just areas of expertise.



ETCIO

### Religare Health Insurance looks at AI for claim settlement

Suresh Kolla, EVP, Head-IT at Religare Health Insurance, is in the midst of deploying solutions based on AI and advanced analytics to take business to the next level.

Reuters

### China rolls out fresh data collection campaign to combat coronavirus

China's local governments are ramping up surveillance efforts with new data collection campaigns to better trace residents' moves in public areas, seeking to curb the coronavirus outbreak but heightening privac...



ETCIO

### How RailYatri's innovative technology is revolutionising bus, train travel

Manish Rathi, CEO and Co-founder of RailYatri, provides insights into how the company maximises data analytics to enhance customer experience and enable new revenue streams.

ETCIO

### How Apollo Hospitals use data analytics for infection control surveillance

Arvind Sivaramakrishnan, CHCIO, Apollo Hospitals, used data analytics for infection control surveillance. He also talks about the implementation of personal health records and navigating the challenges around it.



ETCIO

### How technology can predict Bollywood's box-office collection

With more than 300 commercially and critically acclaimed movies in its portfolio, Reliance Entertainment is working on a platform that can predict the business volume with respect to a particular

# DATA ANALYTICS

## Top stories-contd.,



TNN

### India Inc uses data analytics to ensure fair appraisal

Given that objectivity is more important during appraisals than human emotions, many organisations have adopted data analytics to ensure the process of appraisal is fair and transparent



PTI

### Navy to focus on big data analytics, artificial intelligence

"The new transition cycle has also resulted in the overall improvement of operational logistics, spares management and forecasting, refit planning and expenditure management," the Navy said.



ET Bureau

### I-T Department plans to use data analytics tools for examining investment and cash deposits

The I-T dept has started sending missive to banks and post offices to submit details of high-value cash deposits in banks and post office accounts.



ETCIO

### United Nations leverages Qlik's data analytics for humanitarian missions globally

United Nations and Qlik has announced partnership to utilize data analytics to improve efficiency and efficacy of humanitarian works across the world

tells ETCIO.



ETCIO

### And It's a match! OkCupid CTO redefines the dating game with Data science

With one of the most high-tech matching algorithms, OkCupid helps those looking for a meaningful relationship find their kind.



ETCIO

### Inside the data driven model of Ola with Sanjay Kharb, VP- Engineering, Ola

"Deriving insights from data is at the heart of everything we do at Ola," exclaims Sanjay Kharb, VP- Engineering, Ola. In an interview with ETCIO, he provides insights into the data science model

### Indian Student In US Uses Big Data Analytics To Tackle Parking Problem

Indians Abroad | Press Trust of India | Wednesday October 31, 2018



An Indian student in the US has created a space-detecting algorithm that can help tackle the problem of finding a parking spot by using big data analytics and save a person's time and money.

### Using Algorithms, This New Tool Can Predict When A Building Will Collapse

Science | Indo-Asian News Service | Thursday August 16, 2018



Australian researchers today said they have developed a software tool to predict when a building will crack or its foundation will move, even when a dam could break or a mudslide occur. The tool uses applied mathematics and big data analytics to analyse intricate ground motion patterns and track location and time of landslides to forecast them up ...

# DATA ANALYTICS

## Top stories-contd.,

ET Bureau

### Budget 2016: Data analytics helps government collect Rs 10,000 crore from non-filers

The government said it collected more than Rs 10,000 crore by leveraging data analytics in the non-filers monitoring system.

ET Bureau

### Now politicians use data analytics to improve decision-making

Among the techniques and tools being used are social media analysis, predictive algorithms, data analytics and forecasting based on primary data.

### Government to use data analytics to go after 18 lakh suspect depositors

According to revenue secretary Adhia, missives will be sent to tax payers on deposits of over Rs 5 lakh in the first phase of the crackdown.

'Data Analytics' - 11 Video Result(s)



2:22



29:32

["Data Is The New Oil, The New Gold": PM At 'Howdy, Modi!' In Houston](#)

[Exclusive: 'Whistleblower' Reveals Cambridge Analytica's India Link](#)

## Job Trends

ET Bureau

### India should become a data analytics hub: Ravi Shankar Prasad

Ravi Shankar Prasad said the report is not just about digital inclusion but also an opportunity to do business in India.



### Infosys invests additional \$1.5 mn in data discovery firm

Earlier in January 2016, the Bengaluru-headquartered company had spent \$4 million to pick up stake in the US-based firm.



ET Bureau

### India to lead in data science job creation: Biocon Chief Kiran Mazumdar-Shaw

Kiran Mazumdar-Shaw pointed out how startup founders in India, who are data scientists, functioned very differently from their counterparts in the US.



ETCIO

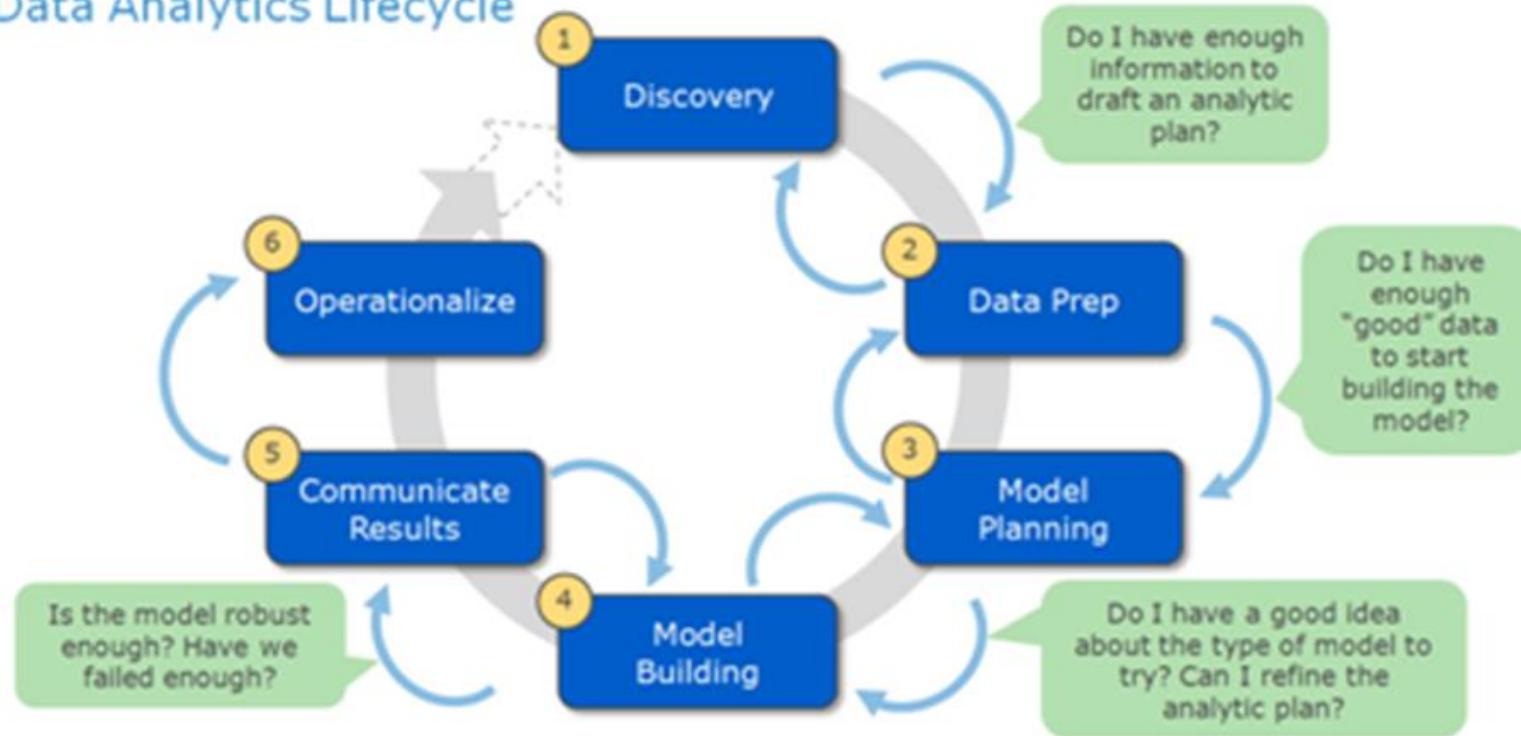
### Women leaders tend to be more collaborative: Nimilita Chatterjee, SVP- Data and Analytics, Equifax

In an interview with ETCIO.in, Nimilita Chatterjee, SVP- Data and Analytics, Equifax shares her views on how IT industry is adapting the women leadership and mentoring their skills for better



## What does it involve?

Data Analytics Lifecycle



# DATA ANALYTICS

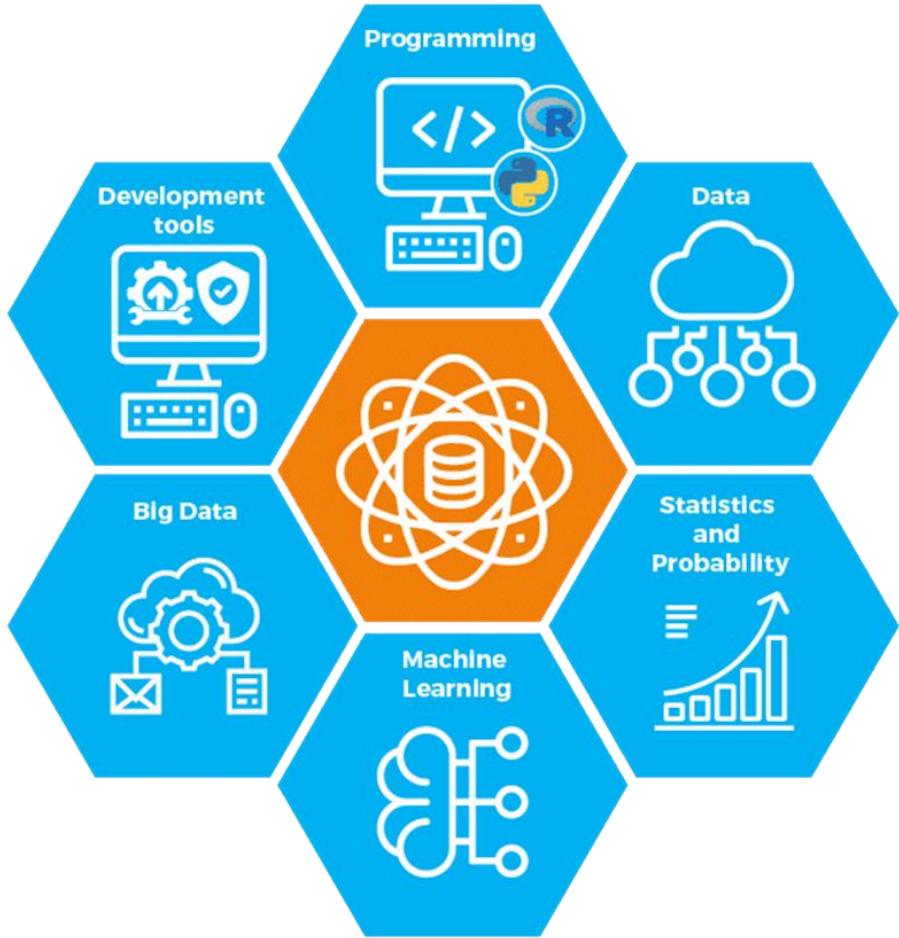
## - A Computer Science Perspective

---



- **Data Collection (Big data and data engineering)**
- **Data Pre-Processing (Cleaning, SQL)**
- **Analysis (EDA, etc.)**
- **Insights (Machine Learning/Deep Learning)**
- **Visual Reports (Visualisations)**

## Skills Required



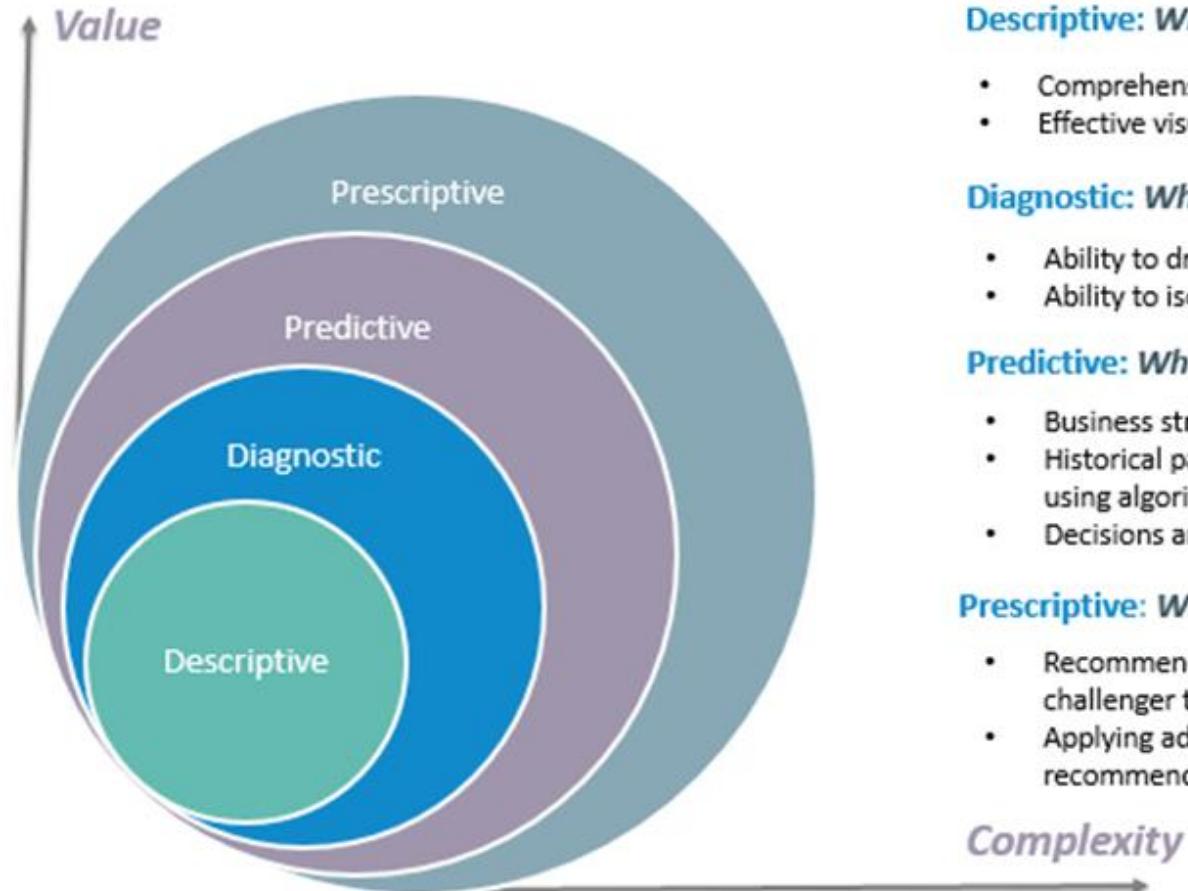
And...  
a secret ingredient



Intuition or  
deductive reasoning and  
domain knowledge

## What do we study in this course?

### 4 types of Data Analytics



### What is the data telling you?

#### **Descriptive:** *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

#### **Diagnostic:** *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

#### **Predictive:** *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

#### **Prescriptive:** *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

# DATA ANALYTICS

# Jargon, jargon, jargon...



- Pattern Recognition
  - Machine Learning
  - Data Mining
  - Information Retrieval
  - Business Intelligence
  - Informatics
  - Data Sciences

In this course, we will learn to use

- Statistics
- Some algorithms to model data
- Visualization tools to analyze data and make predictions



# DATA ANALYTICS

## Teaching Team



Prof. Mamatha H. R.



Prof. Gowri Srinivasa



Prof. Jyothi R.



Bharani U.  
Kempaiah



Ruben John



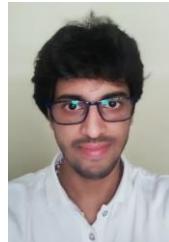
Bhavya  
Charan



Richa



Mainaki



Shreyas B. S.



Pradyumna Y. M.



Mayank Agarwal



Amit Kumar



Tanay Gangey



Greeshma Karanth



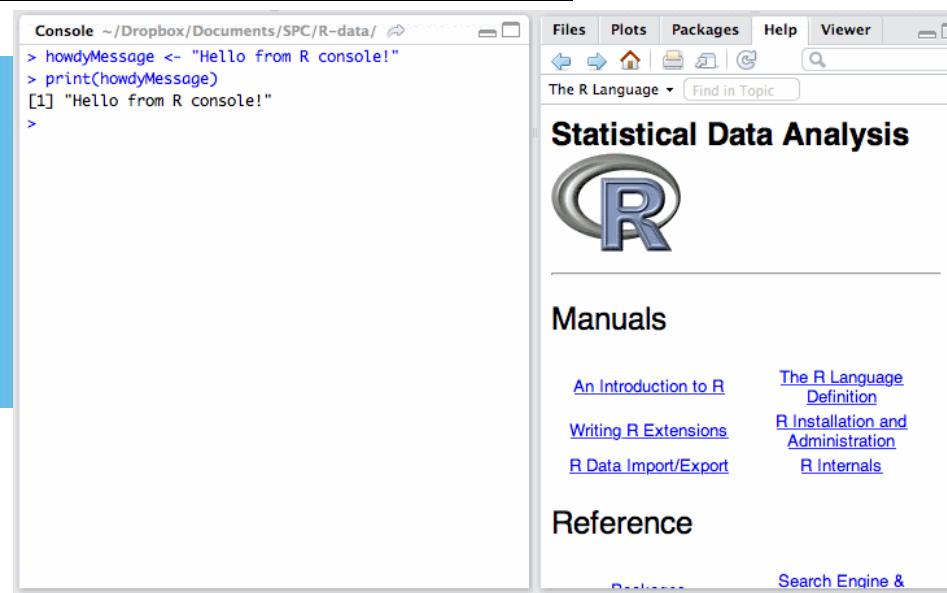
Diya Sateesh



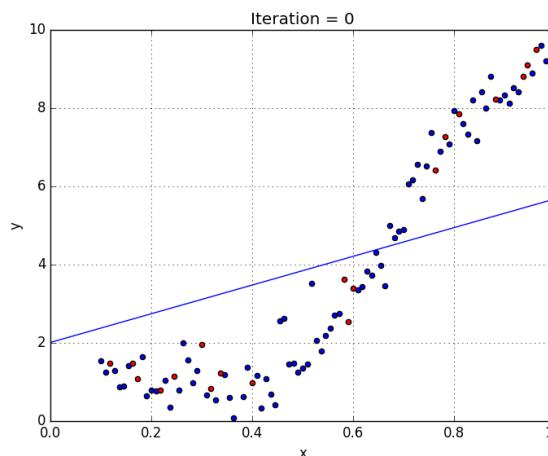
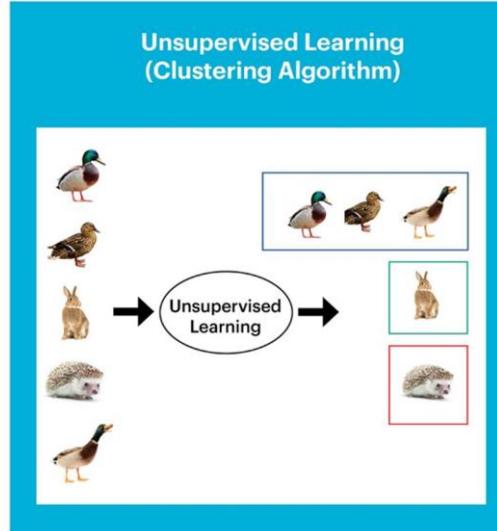
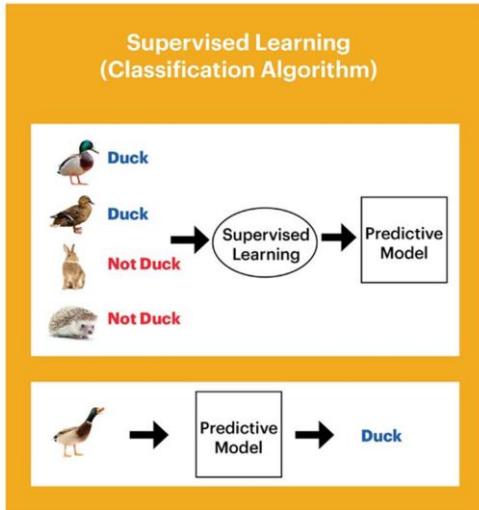
Chiranth Jawahar

# DATA ANALYTICS

## Course Coverage



## Course Outcome



## Course References

---

### Text Book:

“Business Analytics, The Science of Data-Driven Decision Making”,  
U. Dinesh Kumar, Wiley 2017 [\[Chapter 1\]](#)

### Reference Books:

- 1: “Data Mining: Concepts and Techniques” by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3<sup>rd</sup> Edition.
- 2: “The Elements of Statistical Learning”, Trevor Friedman, Robert Tibshirani and Jerome Hastie, Data Mining, Inference and Prediction, Springer 2001.
- 3: “Practical Data Science with R”, Nina Zumel and John Mount, Manning Publications, 2014.

# DATA ANALYTICS

## Image Courtesy

---

<https://gfycat.com/smugscratchyiaerismetalmark-accionadata>

<https://later.com/instagram-marketing/>

<https://www.omnisci.com/technical-glossary/geospatial-analytics>

<http://siteanalystiot.com/>

<https://giphy.com/search/>



**THANK YOU**

---

**Dr. Mamatha H R**

Professor, Department of Computer Science

**[mamathahr@pes.edu](mailto:mamathahr@pes.edu)**

+91 80 2672 1983 Extn 834

**Dr. Gowri Srinivasa**

Professor, Department of Computer Science

**[gsrinivasa@pes.edu](mailto:gsrinivasa@pes.edu)**



# DATA ANALYTICS

## Unit 1:Introduction

---

**Mamatha.H.R**

Department of Computer Science and Engineering

**Gowri Srinivasa**

Department of Computer Science and Engineering

# DATA ANALYTICS

---

## Unit 1: Introduction

**Mamatha H R, Gowri Srinivasa**

Department of Computer Science and Engineering

## What is Data Analytics?

---

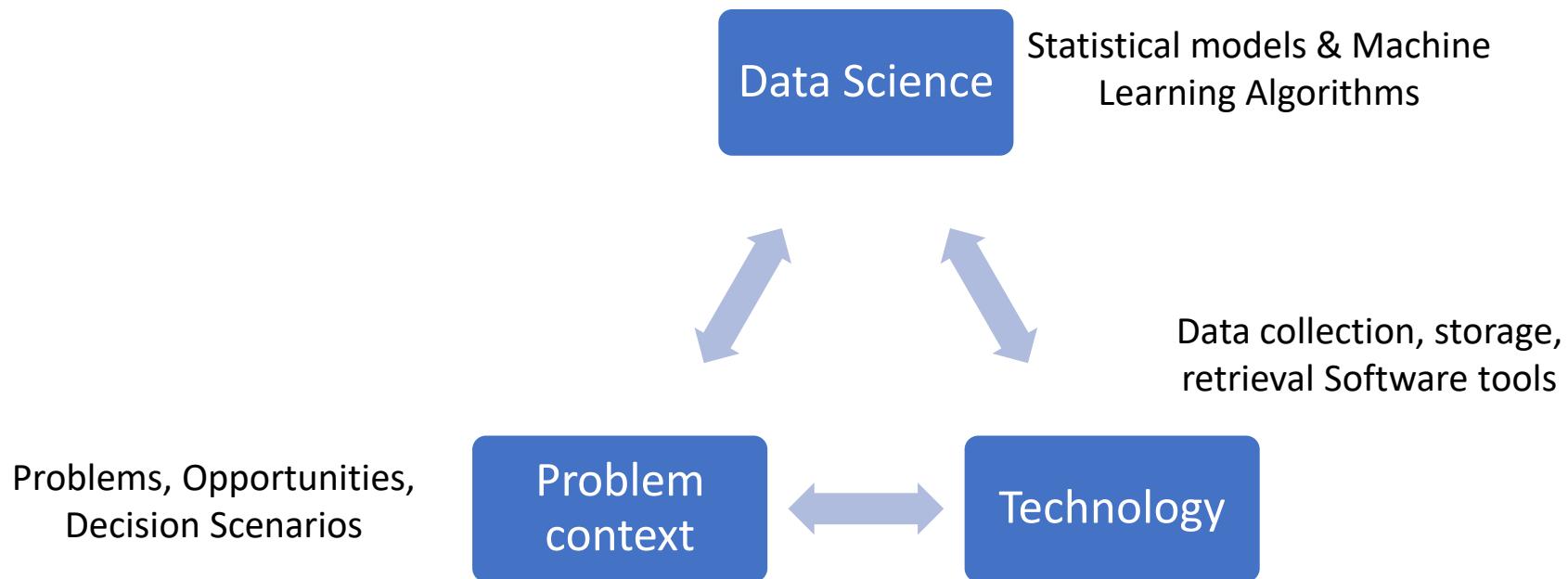
Data analytics (DA) is the process of examining data in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software.

<https://searchdatamanagement.techtarget.com/definition/data-analytics>

## What is Data Analytics?

Data analytics is a set of statistical and operation research techniques, artificial intelligence, information technology and management strategies used for framing the problem, collecting data, and analyzing the data to create value to organizations.

It can be broken into three components



## Applications – search engines

Google

bottle cap challenge



All

Videos

News

Images

Maps

More

Settings

Tools

About 8,02,00,000 results (1.40 seconds)

### Top stories



'Better Late Than Never': Virat Kohli Takes Up The Bottle Cap Challenge

HuffPost India

23 hours ago



Watch: After Shikhar Dhawan and Yuvraj Singh, Virat Kohli aces the bottle cap challenge

Times of India

2 days ago



Watch: Virat Kohli's Bottle Cap Challenge Comes With A Unique Twist

NDTV Sports

23 hours ago



→ More for bottle cap challenge

## Even less ‘trendy’ queries online

Google

sensitivity conjecture



All News Videos Images Shopping More Settings Tools

About 91,60,000 results (0.63 seconds)

[Shtetl-Optimized » Blog Archive » Sensitivity Conjecture resolved](#)

<https://www.scottaaronson.com/blog/?p=4229> ▾

Jul 2, 2019 - The Sensitivity Conjecture, which I blogged about here, says that, for every Boolean function  $f:\{0,1\}^n \rightarrow \{0,1\}$ , the sensitivity of  $f$ —that is, the ...

[Amazing: Hao Huang Proved the Sensitivity Conjecture! - Gil Kalai](#)

<https://gilkalai.wordpress.com/.../amazing-hao-huang-proved-the-sensitivit...> ▾

Jul 2, 2019 - Today's arXived amazing paper by Hao Huang Induced subgraphs of hypercubes and a proof of the Sensitivity Conjecture Contains an ...

[Induced subgraphs of hypercubes and a proof of the Sensitivity ...](#)

[arxiv.org](#) ▾ math ▾

by H Huang - 2019

Jul 1, 2019 - As a direct consequence, we prove that the sensitivity and degree of a ... computer science, the Sensitivity Conjecture of Nisan and Szegedy.

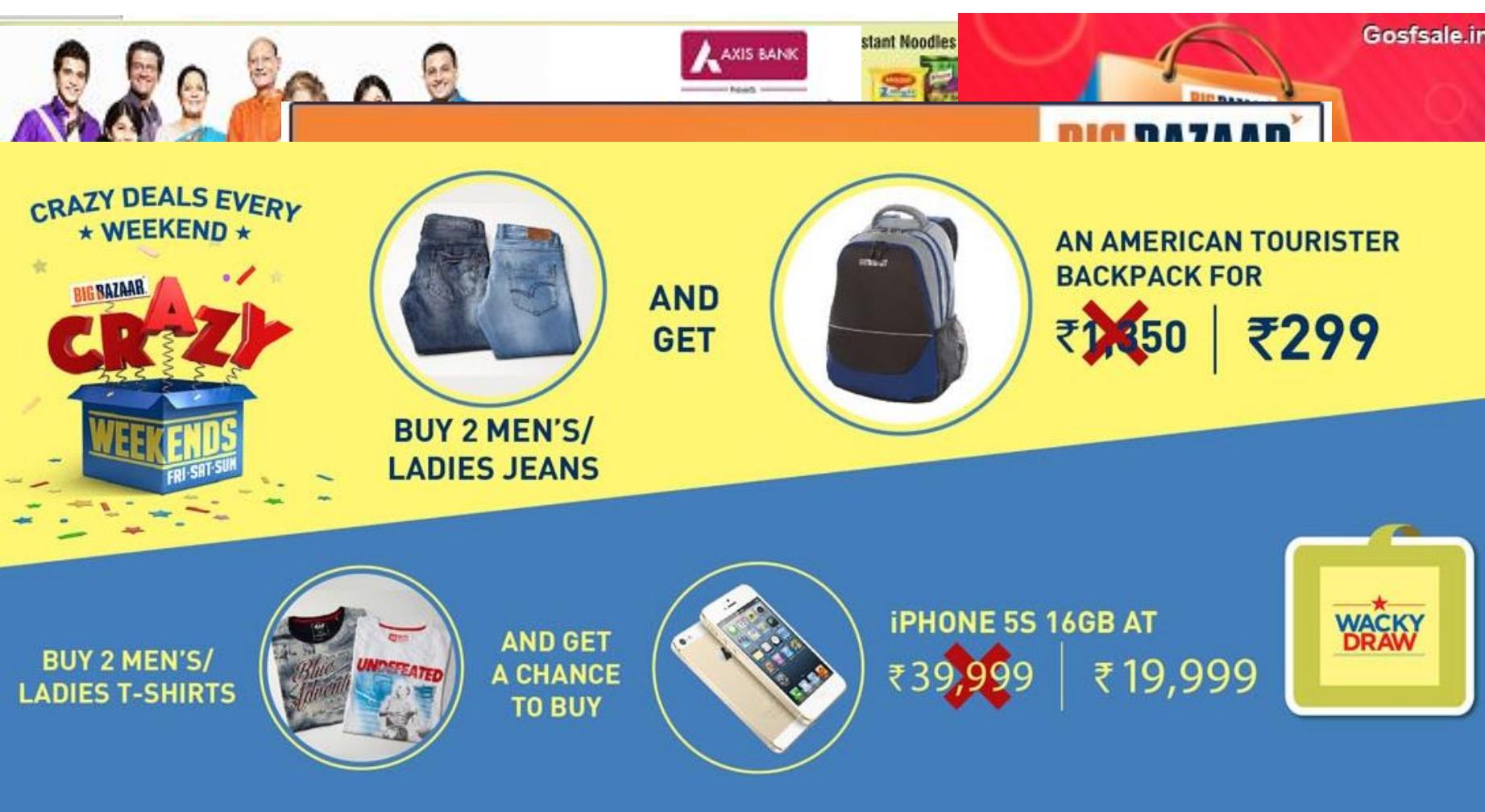
[Mathematician to present a proof of the Sensitivity Conjecture - P...](#)

<https://phys.org> ▾ Other Sciences ▾ Mathematics ▾

Jul 10, 2019 - The Sensitivity Conjecture has stood as one of the most important, and baffling, open problems in theoretical computer science for nearly three ...

# DATA ANALYTICS

## Strategies to increase sales



The banner features several promotional offers:

- CRAZY DEALS EVERY WEEKEND**: Buy 2 Men's/Ladies Jeans and get an American Tourister Backpack for ₹299 (from ₹1,350).
- WACKY DRAW**: A chance to win an iPhone 5S 16GB at ₹19,999 (from ₹39,999).
- HDFC BANK**: Offers 5% cashback.
- Axis Bank**, **stant Noodles**, and **Gosfsale.in** are also mentioned.

## History

---

MIT students were asked to sign up for

- Web Subscription – \$59 (68 students)
- Print Subscription – \$125 (32 students)

Total revenue: **\$8,012**

- ▶ Web Subscription – \$59 (16 students)
- ▶ Print Subscription – \$125 (0 students)
- ▶ Web and Print Subscription – \$125 (84 students)

Total revenue: **\$11,444**

# DATA ANALYTICS

## Targeted advertisements



Gmail by Google

Compose Mail → Free Google Local Listing - Google.com/LocalBusinessCenter - Put your location and more for free on Google Sponsored Link < >

Inbox (516) « Back to Inbox Archive Report spam Delete Move to ▾ Labels ▾ More actions ▾ 1 of 529 Older >

Starred ★ Sent Mail Drafts (4)

Coaches 4 more ▾ Contacts Tasks

+ John Jantsch Search, add, or invite

Invite a friend Give Gmail to: Send Invite 100 left Preview Invite

**Small business marketing** Inbox | X

John Jantsch to me show details 8:03 AM (1 minute ago) Reply ↴  
I need to hire a marketing consultant, any ideas?  
John Jantsch  
Duct Tape Marketing  
4806 Bellevue Ave.  
Kansas City, MO 64112  
866-DUC-TAPE (382-8273)

How do *they* know?

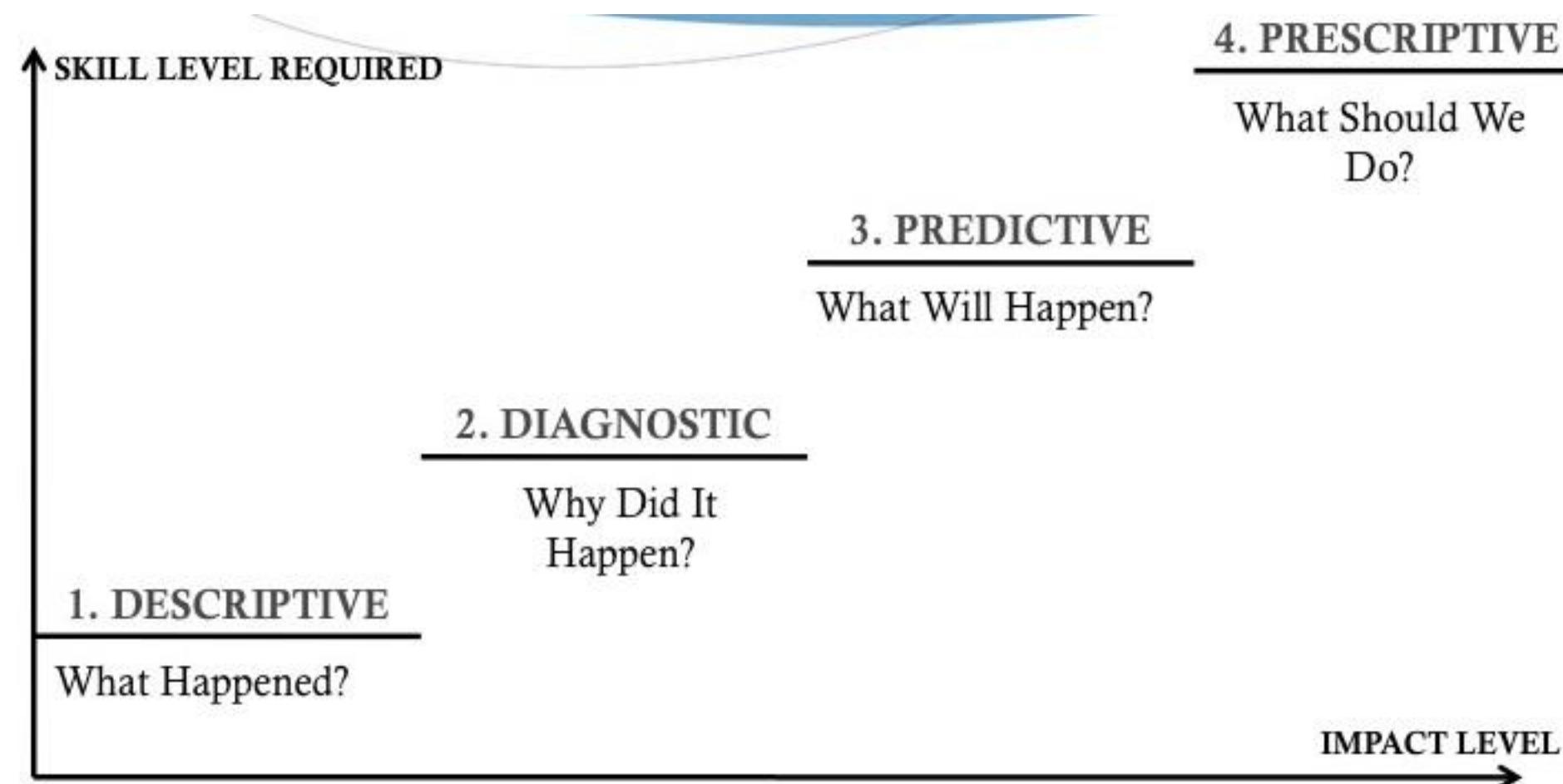
New window Print all Map this 4806 Bellevue Ave Kansas City, MO 64112

Sponsored Links Advertise Small Business Boost your business, target your audience and more with PRWeb! www.prweb.com

Vending Machine Business A Cash Cow! 10 Soda/Snack Vending Machines-Locations Secured For You. www.1800VENDING.com

Entrepreneurship Ideas Join Our Small Business Community. Get Advice from Experts & Peers Now www.BankofAmerica.com/YourBusiness

More about



## Performance analytics

<http://bigmarketresearchh.tumblr.com/post/132095171242/delve-into-sports-analytics-market>



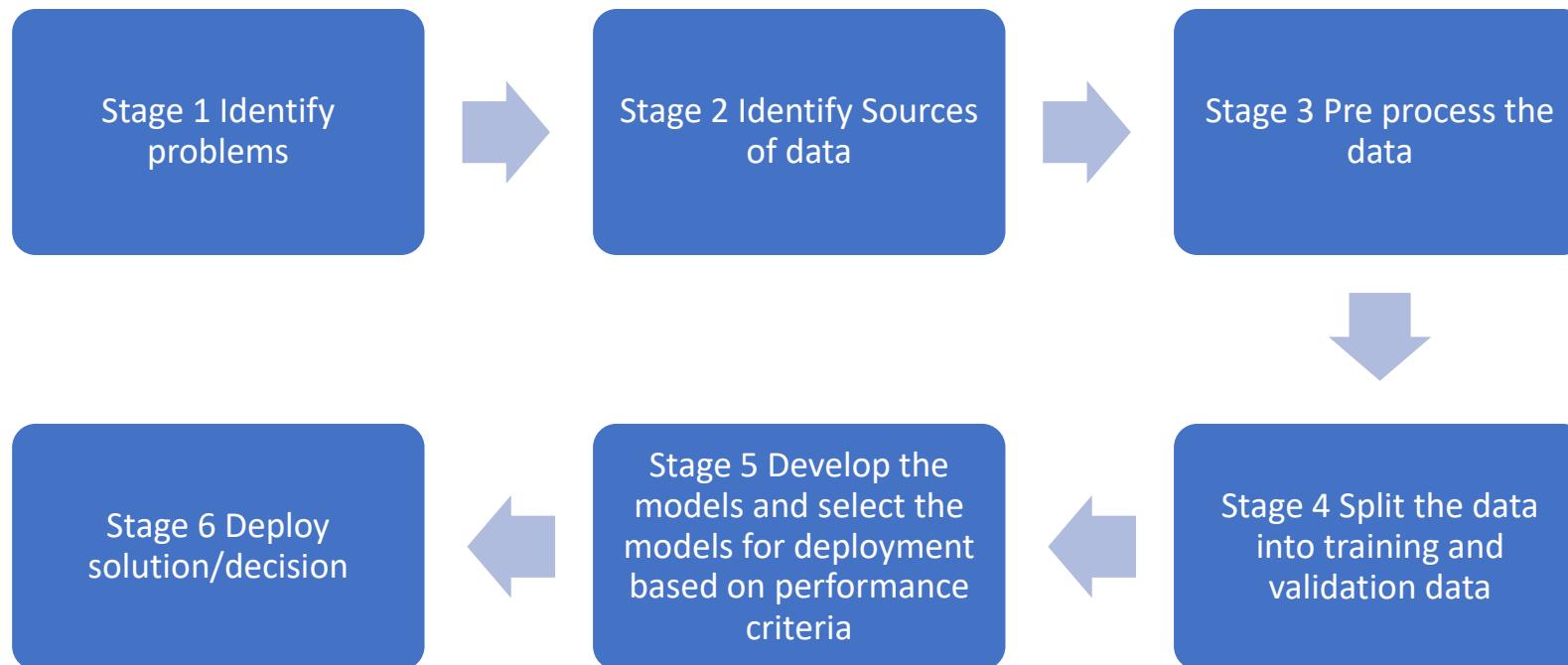
## Other applications

---

- Fraud and risk detection
  - Anomaly detection
- Scheduling and logistics
  - Predicting delays, jams and re-routing vehicles
  - Timetables, surgeries,...
  - Can the bridge support vehicle X carrying load Y?
- Compare prices, features,...
- Analyze feedback
- Image recognition (automated tagging/ retrieval)
- Speech recognition (Siri, Google voice,...)

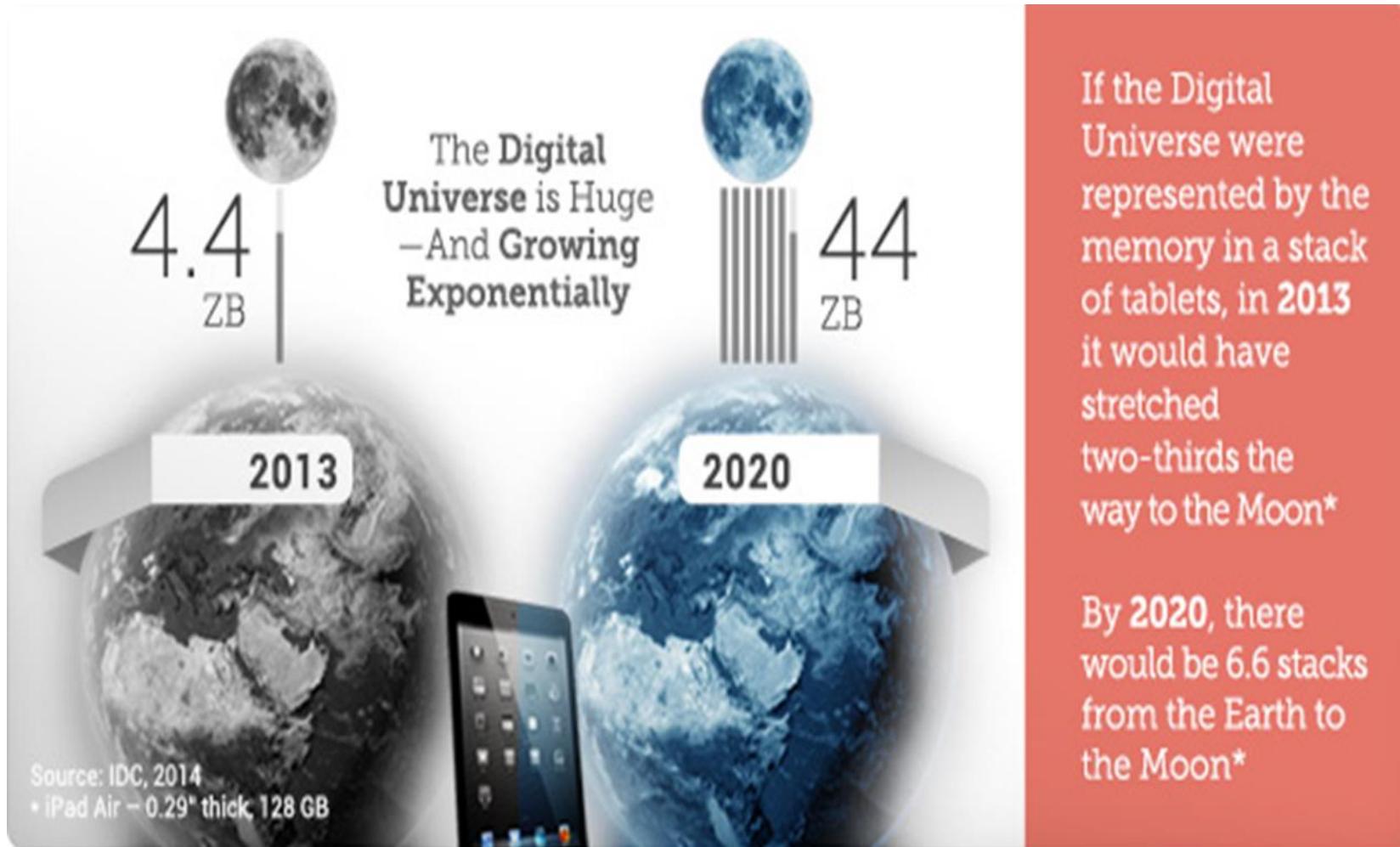
## What is Data Analytics?

Data analytics (DA) is a typical data-driven decision making process



# DATA ANALYTICS

## Data, data everywhere...



## Stage 1: Identify the problem(s)

Can we reduce the cost of decision making?

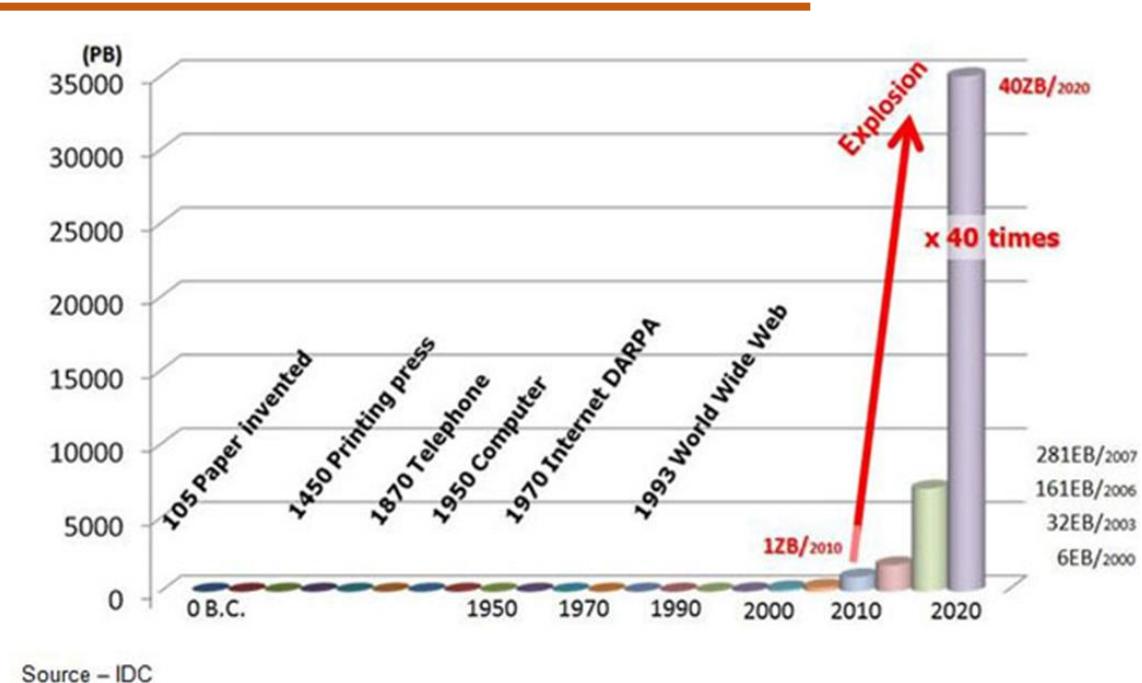
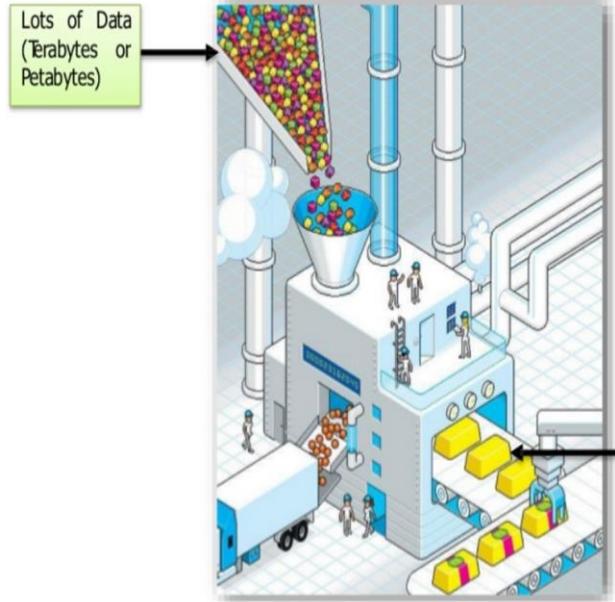
### Cost of decision making fall into three categories

Cost of producing a decision: Cost of reaching a decision with the help of a decision maker or procedure

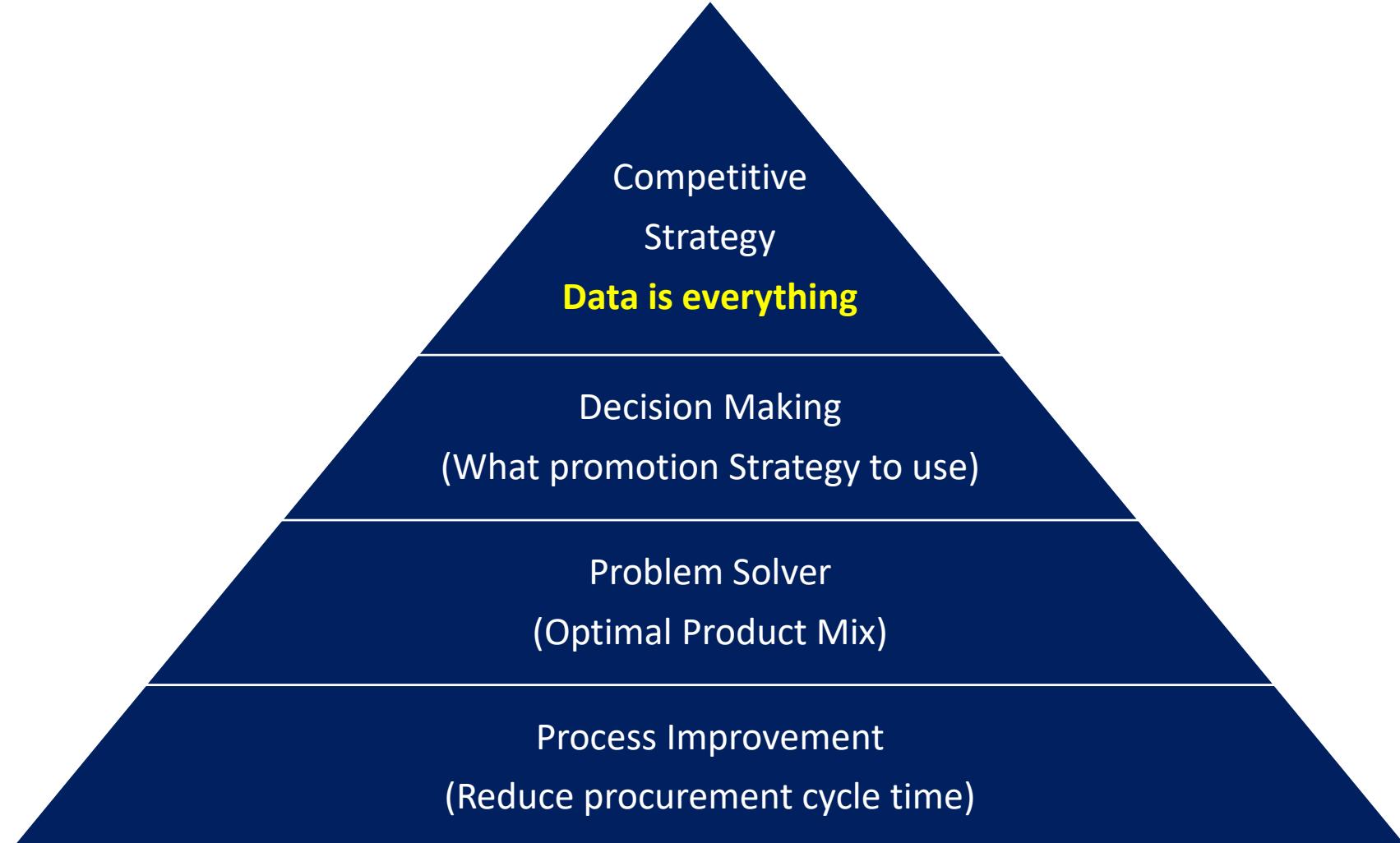
Implementation cost: Cost of actions based on decisions produced

Failure costs: Costs that account for failure of an organization's efforts on production and implementation.

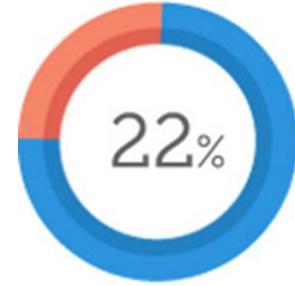
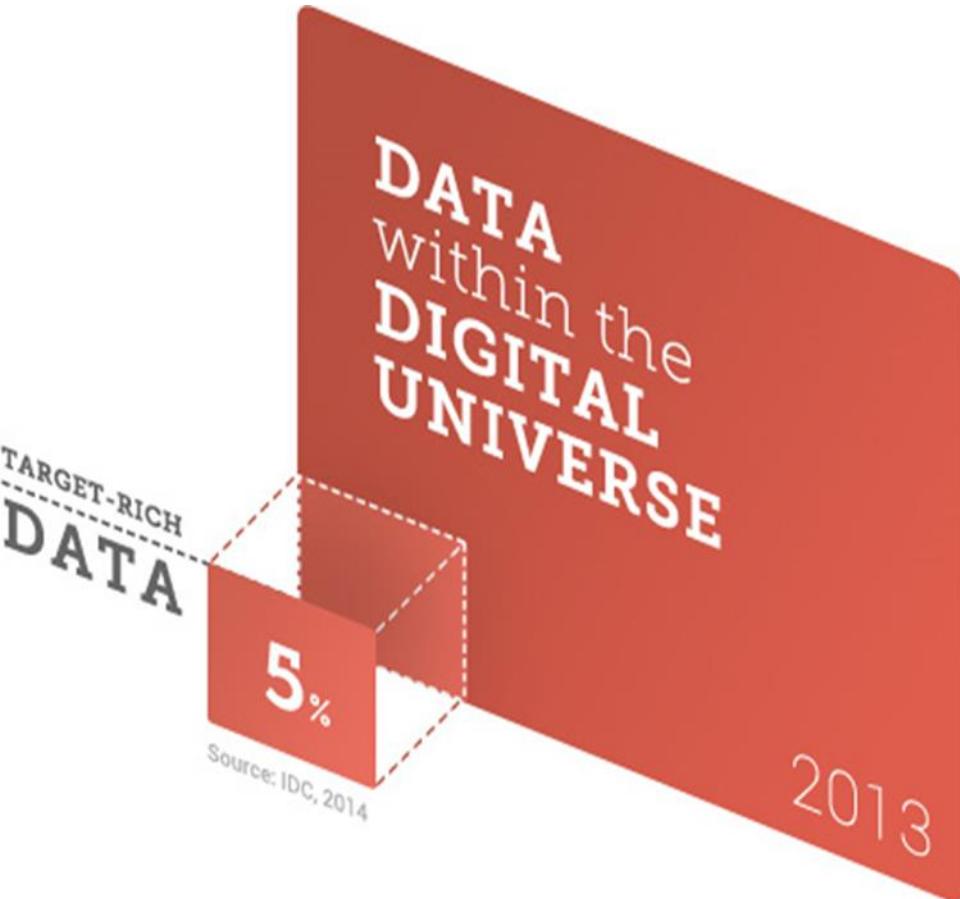
## Stage 2: Identify the source(s) of data



## Why Data Analytics?

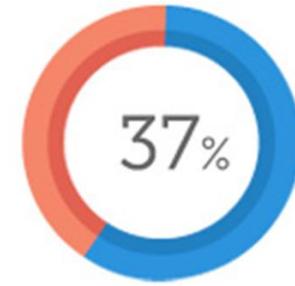


## Why Data Analytics?



Data that is  
Useful if  
Tagged &  
Analyzed

Source IDC, 2014



- Banking – Cheque clearance time
- Healthcare – Patient discharge time
- Manufacturing – Waste minimization
- Retail – Waiting time at check out counters
- E-commerce – Time to deliver the customer order

## Analytics for Problem Solving

---

- Banking – Reduce non-performing assets, Predict Fraud
- Healthcare – Improve net promoter's score (NPS)
- Manufacturing – Reduce inventory management cost
- Retail – Assortment planning and shelf space allocation
- E-commerce – Predict customer cancellations and Fraud

## Analytics for Decision Making

---

- Banking – Loan approval and the interest rate
- Healthcare – Introducing new specialties
- Manufacturing – Whether to introduce a new product
- Retail – Markdown Pricing
- E-commerce – Promotions

## Why Data Analytics?

---

### Example: Akshaya Patra Foundation(TAPF)

Mid-day meal programme in South Bangalore

84000 school children

650 schools

35 vehicles

**Problem:** Vehicle routing –minimize the cost of distribution

1 vehicle-20 schools

Solution space will have  $20!(2.4329 \times 10^{18})$

If a computer can evaluate one million routes per second, it would take more than **77,146** years to evaluate all possible routes

For Akshaya Patra, every rupee saved would enable them to add more children to their mid-day meal programme

## DATA ANALYTICS

Analytics is necessary for survival

---

**Problems faced by E-commerce companies such as Amazon and Flipkart**

- Forecast demand for each Stock keeping unit(SKU).
- Predict customer cancellations and returns.
- Predict customer contacts at the customer service.
- Predict what a customer is likely to purchase in the future?
- How to optimize the delivery system?



## The Game Changers...

---

- **Google**
  - Used Markov chains to rank pages
- **Proctor and Gamble**
  - Analytics as competitive strategy.
- **Target**
  - Predicts customer pregnancy.
- **Capital One**
  - Identifies the most profitable customer.
- **Hewlett Packard**
  - Developed “flight risk score” for 3,30,000 employees.
- **Obama’s 2012 presidential campaign.**
  - Persuasion Modelling.

- **OKCupid:** Predicts which online dating message is most likely to get a response
- **Polyphonic HMI:** Uses “hit song science” to predict commercial success of a song
- **Netflix:** Predicts movie ratings by customers (RMSE is 1%)
- **Amazon.com:** 35% of sales come from product recommendations
- **Divorce360.com:** Predicting success of a marriage!

## Case Study

---

### Indian online grocery store [bigbasket.com](http://bigbasket.com)

**Problem context driving analytics** : “did you forget feature”

The ability to predict the items that a customer may have forgotten to order can have a significant impact on the profits of online grocers such as bigbasket.com

The ability to ask right questions is an important success criteria for analytics projects.

## Case Study

---

### Indian online grocery store [bigbasket.com](http://bigbasket.com)

#### Technology:

To find out whether a customer has forgotten to place an order for an item

Information technology is used for data capture, data storage, data preparation, data analysis, data share and to deploy solution

An important output of analytics is automation of actionable items derived from analytical models which is usually achieved using IT

## Case Study

---

### Indian online grocery store bigbasket.com

#### Data Science:

Data science is the most important component of analytics, it consists of statistical and operations research techniques, machine learning and deep learning algorithms.

The objective of the data science component of analytics is to identify the most appropriate statistical model/machine learning algorithm that is best based on a measure of accuracy.

Example: did you forget prediction is a classification problem in which customers are classified into

1. Forget
2. Not forget

## References

---

### Text Book:

“Business Analytics, The Science of Data-Driven Decision Making”,  
U. Dinesh Kumar, Wiley 2017 [Chapter 1]

## DATA ANALYTICS

Image Courtesy

---

<https://gfycat.com/smugscratchyiaerismetalmark-accionadata>

<https://later.com/instagram-marketing/>

<https://www.omnisci.com/technical-glossary/geospatial-analytics>

<http://siteanalystiot.com/>

<https://giphy.com/search/>

<https://www.epam.com/insights/blogs/mapping-the-analytics-continuum-in-life-sciences>

# DATA ANALYTICS

Coming up next...

---

- Sources of data
- Data types and formats





**THANK YOU**

---

**Dr.Mamatha H R**

Professor, Department of Computer Science

**mamathahr@pes.edu**

+91 80 2672 1983 Extn 834



## DATA ANALYTICS

### Unit 1: Data Sources and Representations

---

**Mamatha.H.R**

Department of Computer Science and  
Engineering

# DATA ANALYTICS

---

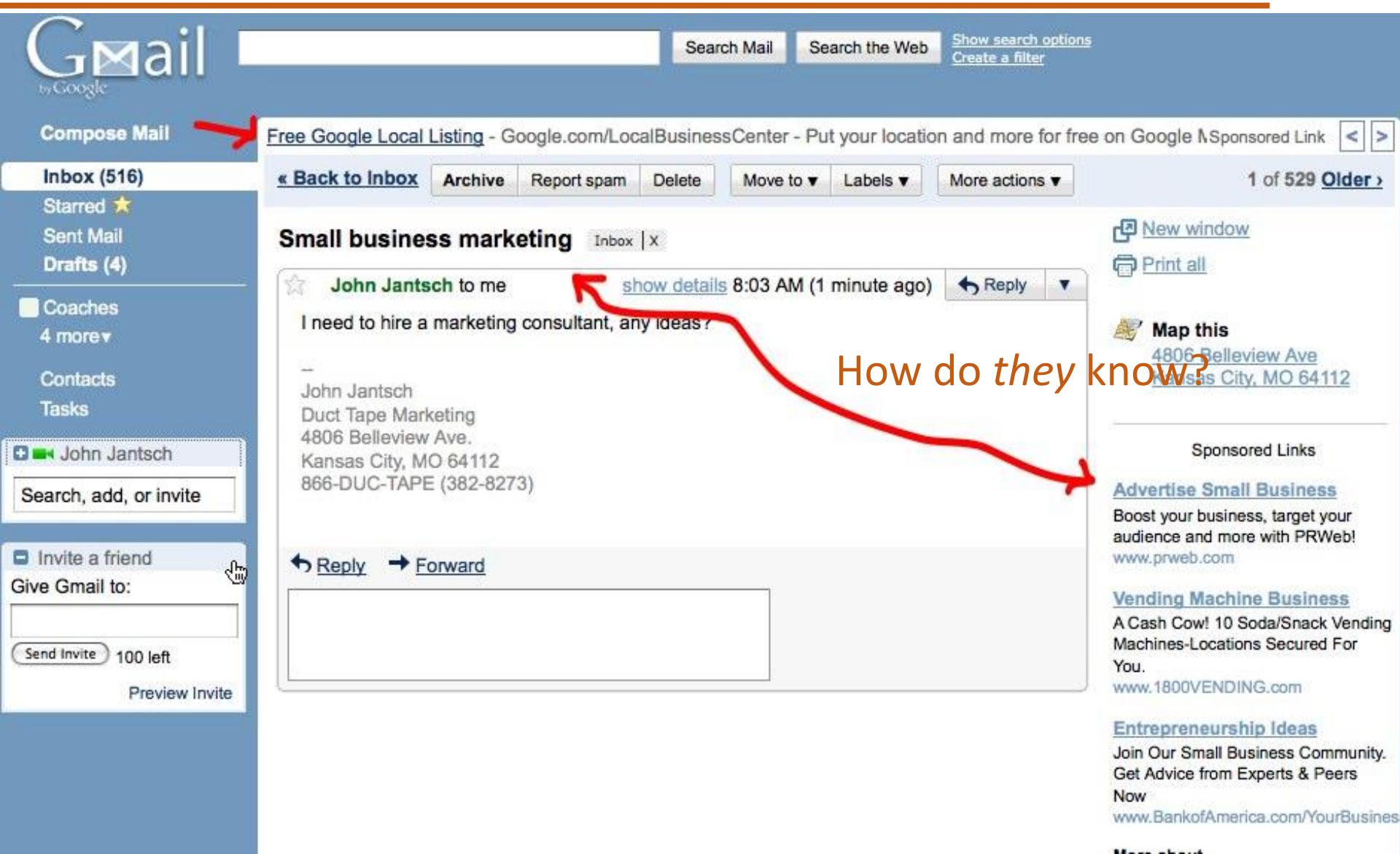
## Unit 1: Data Sources and Representations

Mamatha H R

Department of Computer Science and Engineering

# DATA ANALYTICS

## Targeted advertisements



The screenshot shows a Gmail inbox with the following details:

- Compose Mail** button (highlighted with a red arrow).
- Inbox (516)**
- Starred**
- Sent Mail**
- Drafts (4)**
- Coaches** (4 more)
- Contacts**
- Tasks**
- + John Jantsch** (highlighted with a red arrow)
- Search, add, or invite**
- Invite a friend** button
- Give Gmail to:** input field
- Send Invite** button (100 left)
- Preview Invite** button

**Email Thread:**

- From:** John Jantsch to me
- Date:** 8:03 AM (1 minute ago)
- Subject:** show details
- Message:** I need to hire a marketing consultant, any ideas?
- Reply and Forward buttons**

**Right-click context menu (highlighted with a red arrow):**

- New window
- Print all
- Map this
  - 4806 Bellevue Ave
  - Kansas City, MO 64112

**Advertisement Content:**

- Sponsored Links:** Advertise Small Business, Boost your business, target your audience and more with PRWeb! [www.prweb.com](http://www.prweb.com)
- Vending Machine Business:** A Cash Cow! 10 Soda/Snack Vending Machines-Locations Secured For You. [www.1800VENDING.com](http://www.1800VENDING.com)
- Entrepreneurship Ideas:** Join Our Small Business Community. Get Advice from Experts & Peers Now [www.BankofAmerica.com/YourBusiness](http://www.BankofAmerica.com/YourBusiness)

**Text Overlay:** How do *they* know?

## Case Study

---

### Indian online grocery store [bigbasket.com](#)

**Problem context driving analytics** : “did you forget feature”

The ability to predict the items that a customer may have forgotten to order can have a significant impact on the profits of online grocers such as [bigbasket.com](#)

The ability to ask right questions is an important success criteria for analytics projects.

## Case Study

---

### Indian online grocery store [bigbasket.com](http://bigbasket.com)

#### Technology:

To find out whether a customer has forgotten to place an order for an item

Information technology is used for data capture, data storage, data preparation, data analysis, data share and to deploy solution

An important output of analytics is automation of actionable items derived from analytical models which is usually achieved using IT

## Case Study

---

### Indian online grocery store bigbasket.com

#### Data Science:

Data science is the most important component of analytics, it consists of statistical and operations research techniques, machine learning and deep learning algorithms.

The objective of the data science component of analytics is to identify the most appropriate statistical model/machine learning algorithm that is best based on a measure of accuracy.

Example: did you forget prediction is a classification problem in which customers are classified into

1. Forget
2. Not forget

# DATA ANALYTICS

## Data Sources

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies



*Cyber Security*



*E-Commerce*



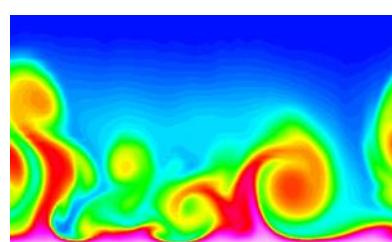
*Traffic Patterns*



*Social Networking: Twitter*



*Sensor Networks*



*Computational Simulations*

## Data Sources

---

- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions



## How large is *big* (data)?

---

- 1 bit
- 1 byte = 8 bits
- 1 KB = 1024 bytes
- 1 MB = 1024 KB (kilobytes)
- 1 GB = 1024 MB (megabytes)
- 1 TB = 1024 GB (gigabytes)  $\approx 10^{12}$  bytes
- 1 PB = 1024 TB (terabytes)  $\approx 10^{15}$  bytes

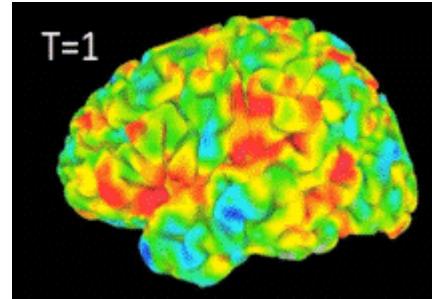
20 PB = amt of data processed by Google per day!

- 1 EB = 1024 PB (petabytes)
  - 1 ZB = 1024 EB (exabytes)
  - 1 YB = 1024 ZB (zettabytes)
- 
- What is a Domezemegrottebyte?

# DATA ANALYTICS

## Data Sources

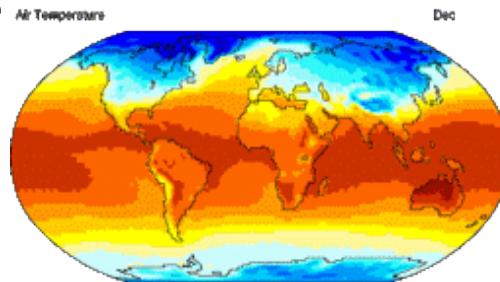
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours



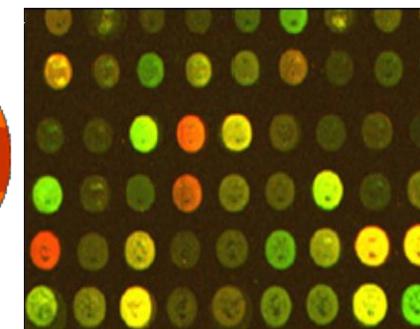
*fMRI Data from Brain*



*Sky Survey Data*



*Surface Temperature of Earth*



*Gene Expression Data*

## What is Data?

- Collection of *data objects* and their *attributes*
- An *attribute* is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an *object*
  - Object is also known as record, point, case, sample, entity, or instance

**Attributes**

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

## A More Complete View of Data

---

- Data may have parts
- Attributes (objects) may have relationships with other attributes (objects)
- More generally, data may have structure
- Data can be incomplete

## Attribute Values

---

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different

## Types of Attributes

---

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

## Discrete and Continuous Attributes

---

- Discrete Attribute

- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: **binary attributes** are a special case of discrete attributes

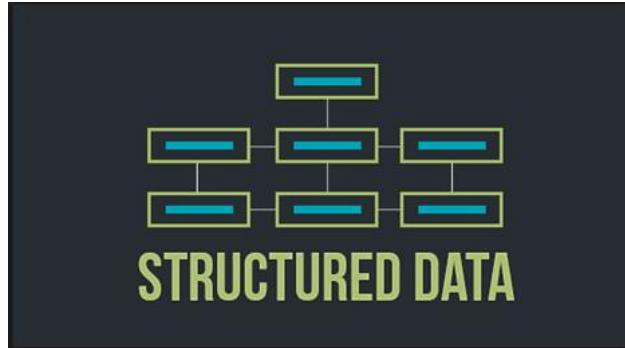
- Continuous Attribute

- Has real numbers as attribute values
- Examples: 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.

## Data Representations

---

- Structured
  - Unstructured
  - Semi structured
- 
- Structured data means that the data is described in a matrix form with labelled rows and columns.
  - Any data that is not originally in the matrix form with rows and columns is an unstructured data.



## Data Representations

---

- relational databases and spreadsheets.
- text and multimedia content. photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.
- XML documents and NoSQL databases.
- For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text.

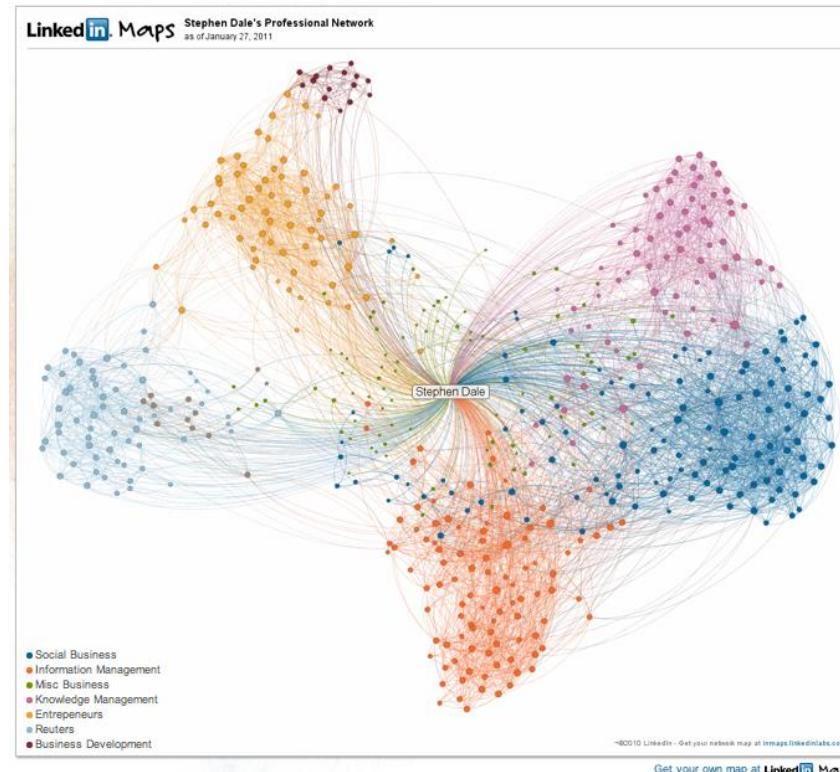
## Data Representations

### 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



## Data Representations

---

- 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:

## Data Representations-Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

## Data Representations-Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

## Data Representations-Document Data

- Each document becomes a 'term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

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

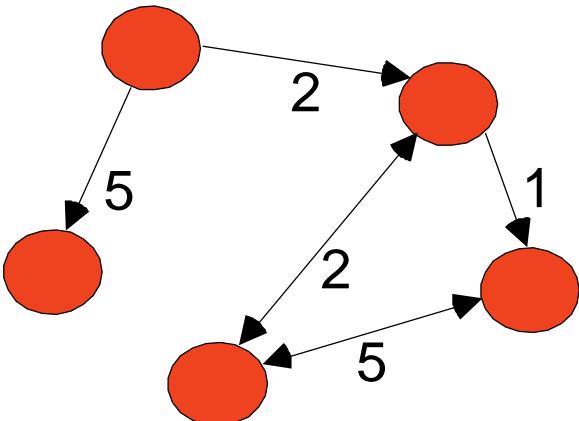
## Data Representations-Transaction data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

ID	Items
1	Bread, Cake, Milk
2	Rice, Bread
3	Rice, Cake, Paper, Milk
4	Rice, Bread, Paper, Milk
5	Cake, Paper, Milk

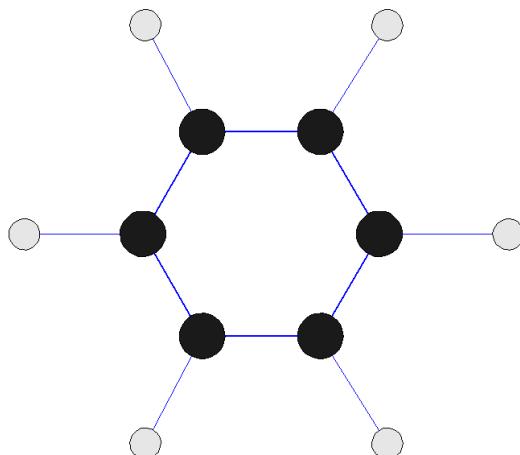
## Data Representations

- Graph Data
- Examples: Generic graph and HTML Links



```
<a href="papers/papers.html#bbbb">  
Data Mining </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>  
<li>  
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

- Chemical Data
- Benzene Molecule: C<sub>6</sub>H<sub>6</sub>



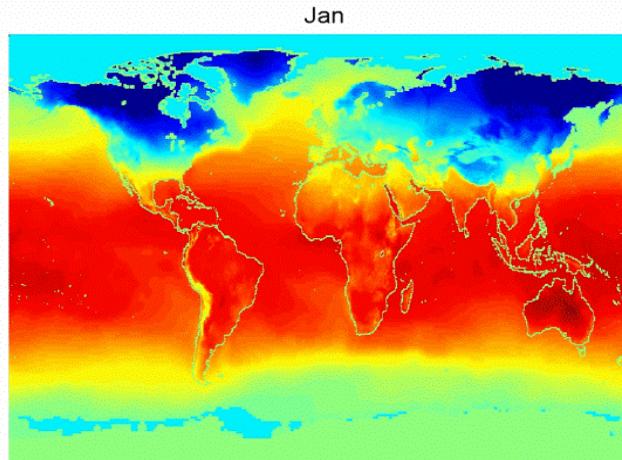
## Data Representations-Ordered Data

- Sequences of transactions
- Spatio-Temporal Data

### Items/Events

( A B) (D) (C E)  
( B D) (C) (E)  
( C D) (B) (A E)

An element of  
the sequence



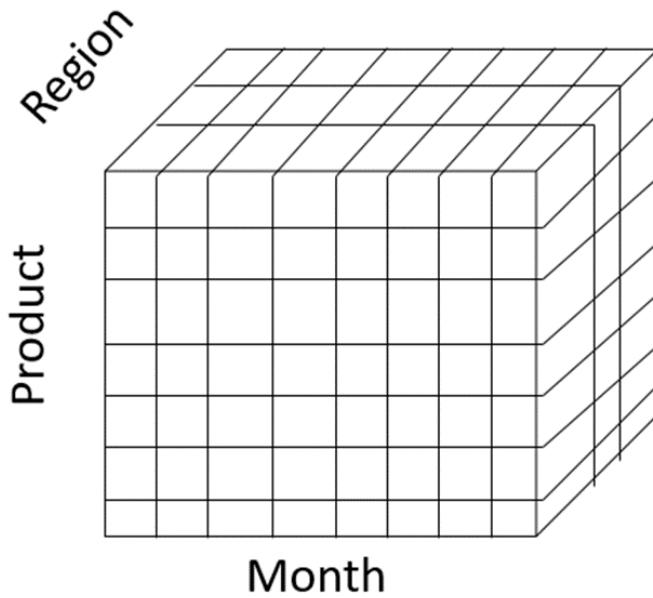
- Genomic sequence data

GGTTCCGCCCTTCAGCCCCGGCGC  
CGCAGGGCCCGCCCCGCGGCCGTC  
GAGAAGGGCCCCTGGCGGGCG  
GGGGGAGGCAGGGCCGCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTGGCCTAGACCTGA  
GCTCATTAGGCAGCAGGGACAG  
GCCAAGTAGAACACCGGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG

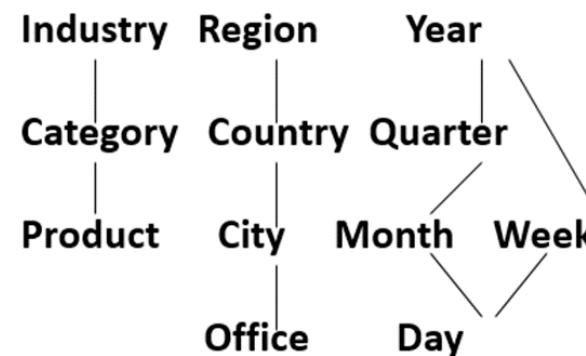
## Data Representations- Data Warehouse

“A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—W. H. Inmon

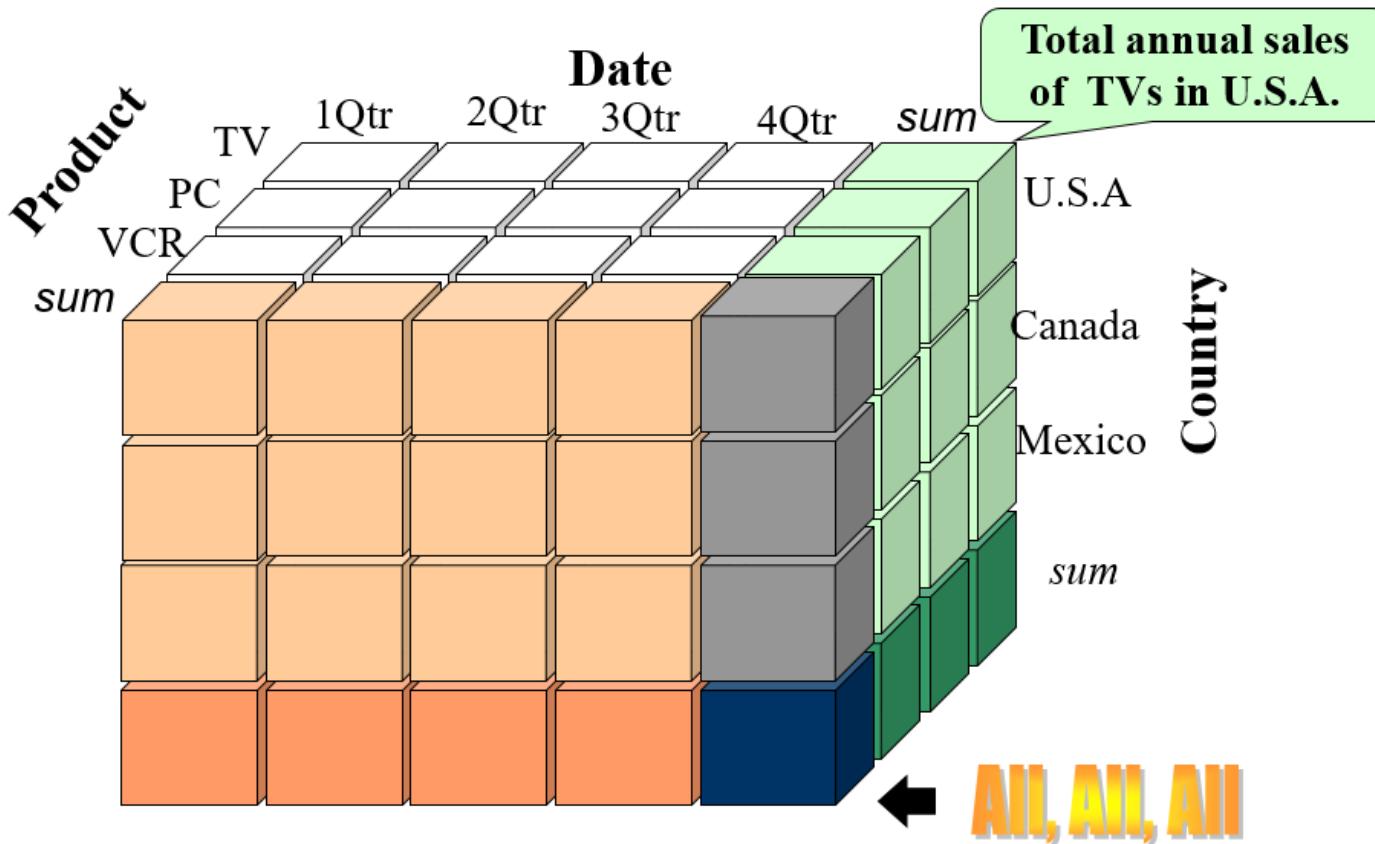
Sales volume as a function of product, month, and region



Dimensions: *Product, Location, Time*  
Hierarchical summarization paths



### A Sample Data Cube



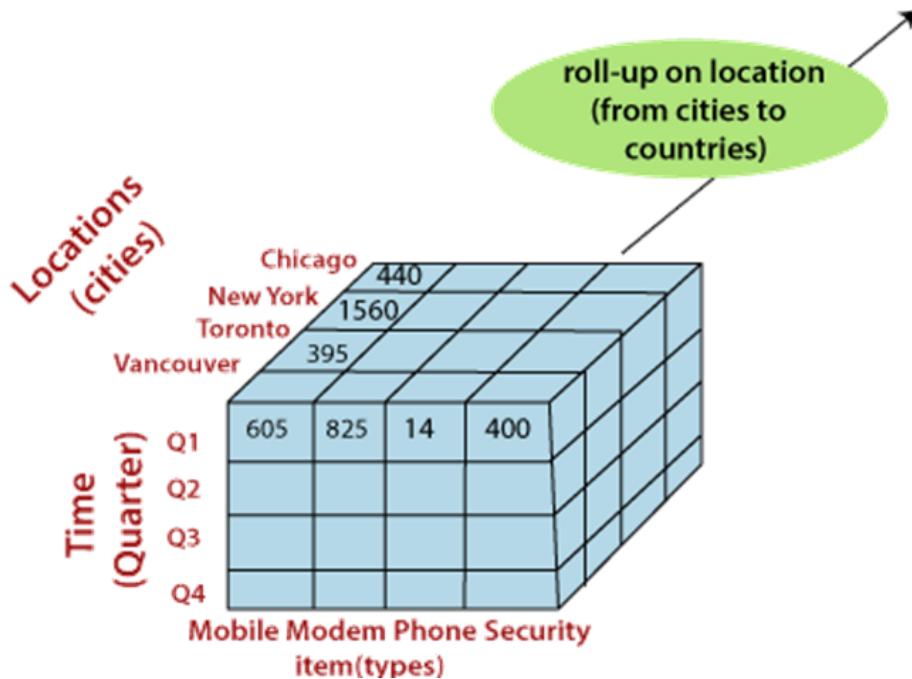
## Typical OLAP Operations

---

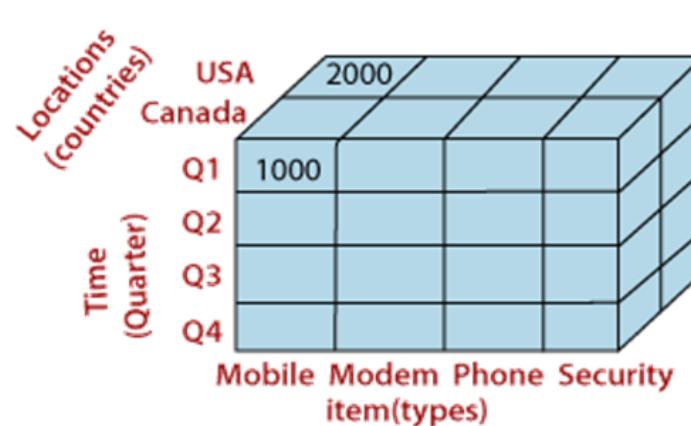
- **Roll up (drill-up):** summarize data
  - by climbing up hierarchy or by dimension reduction
- **Drill down (roll down):** reverse of roll-up
  - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- **Slice and dice:** project and select
- **Pivot (rotate):**
  - reorient the cube, visualization, 3D to series of 2D planes
- Other operations
  - **drill across:** involving (across) more than one fact table
  - **drill through:** through the bottom level of the cube to its back-end relational tables (using SQL)

## Typical OLAP Operations

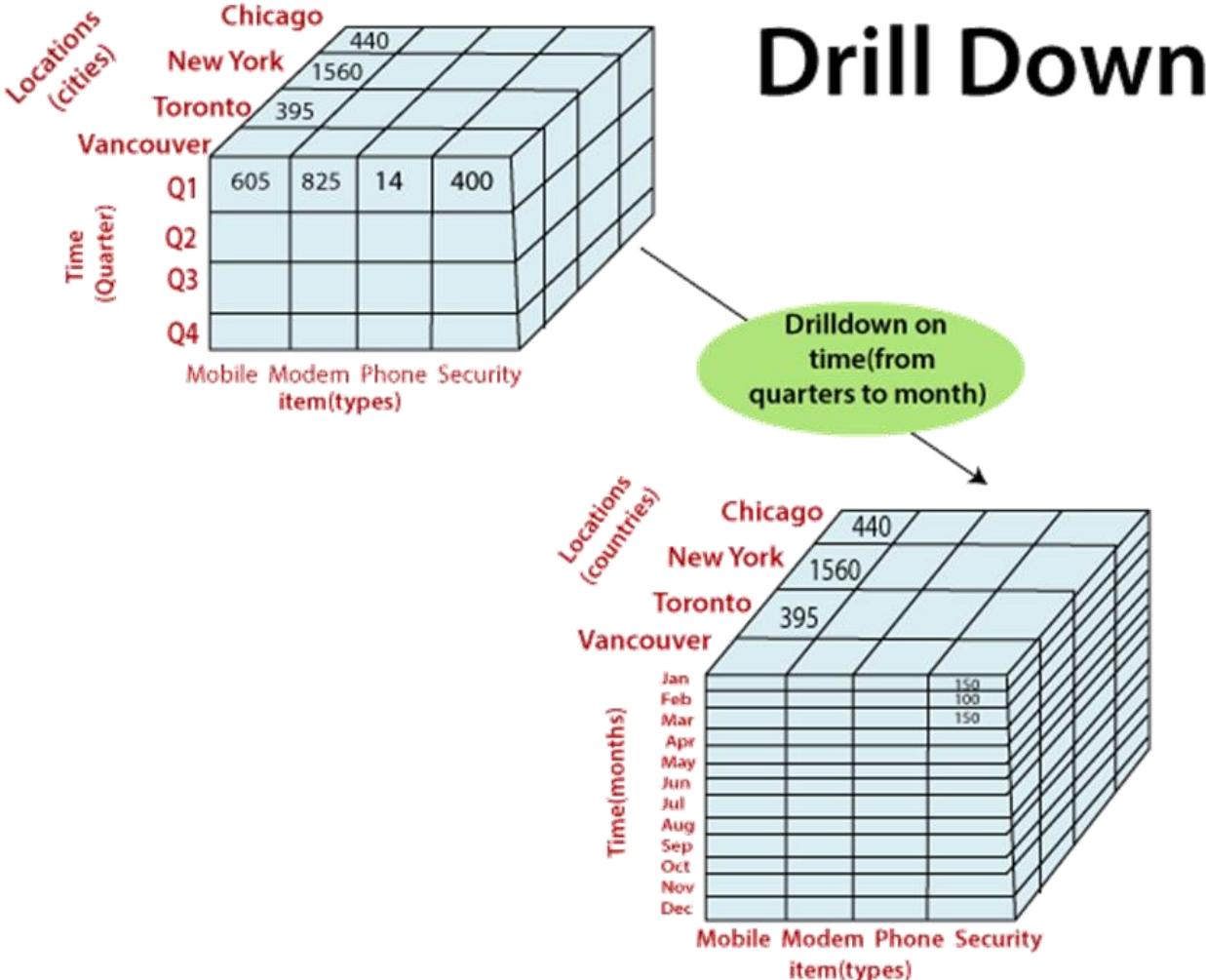
### Roll UP



roll-up on location  
(from cities to countries)

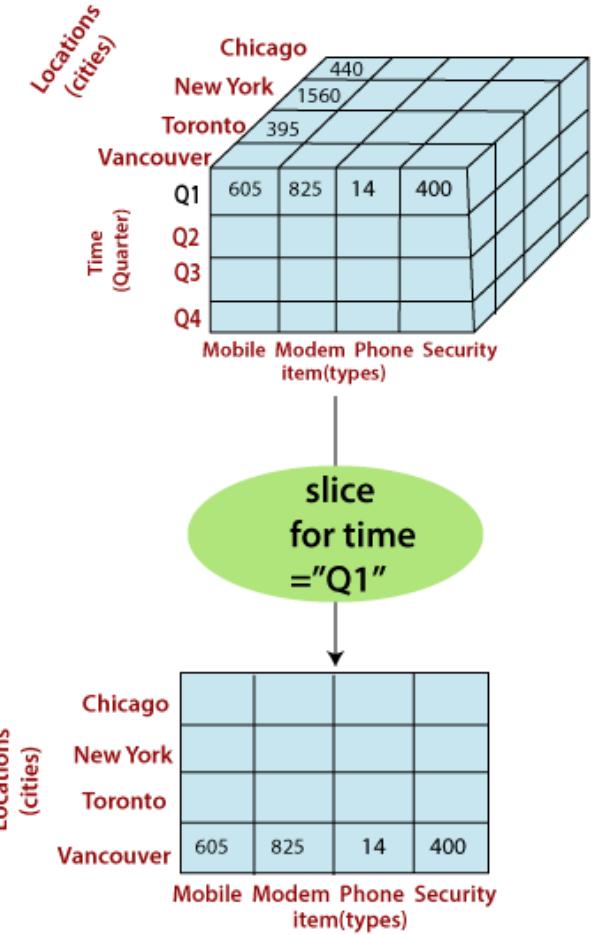


## Typical OLAP Operations

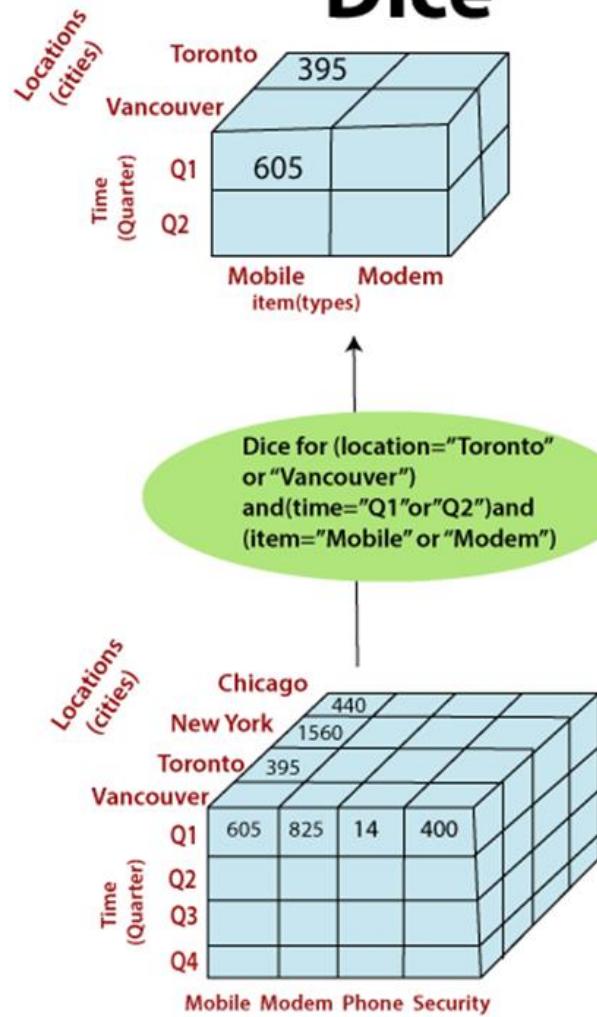


## Typical OLAP Operations

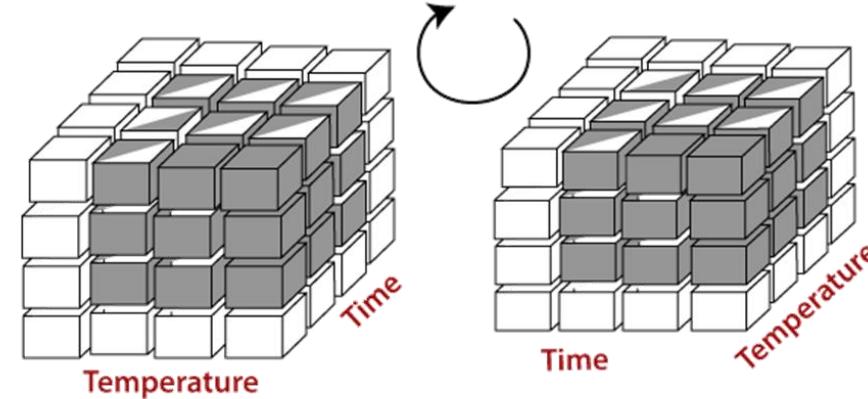
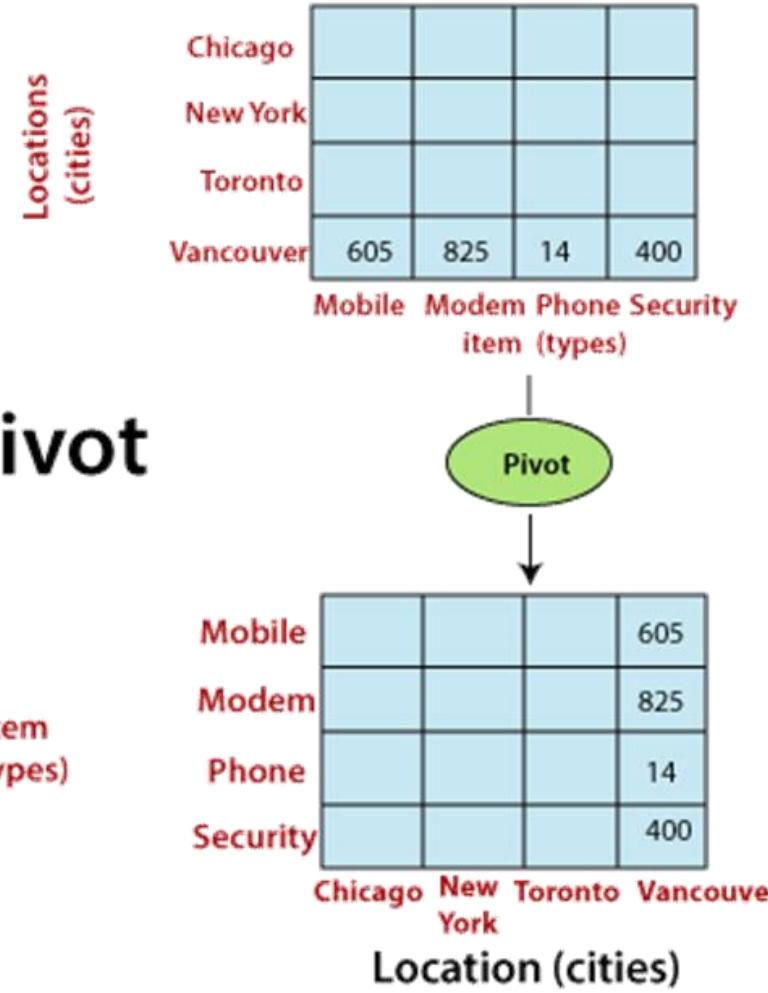
### Slice

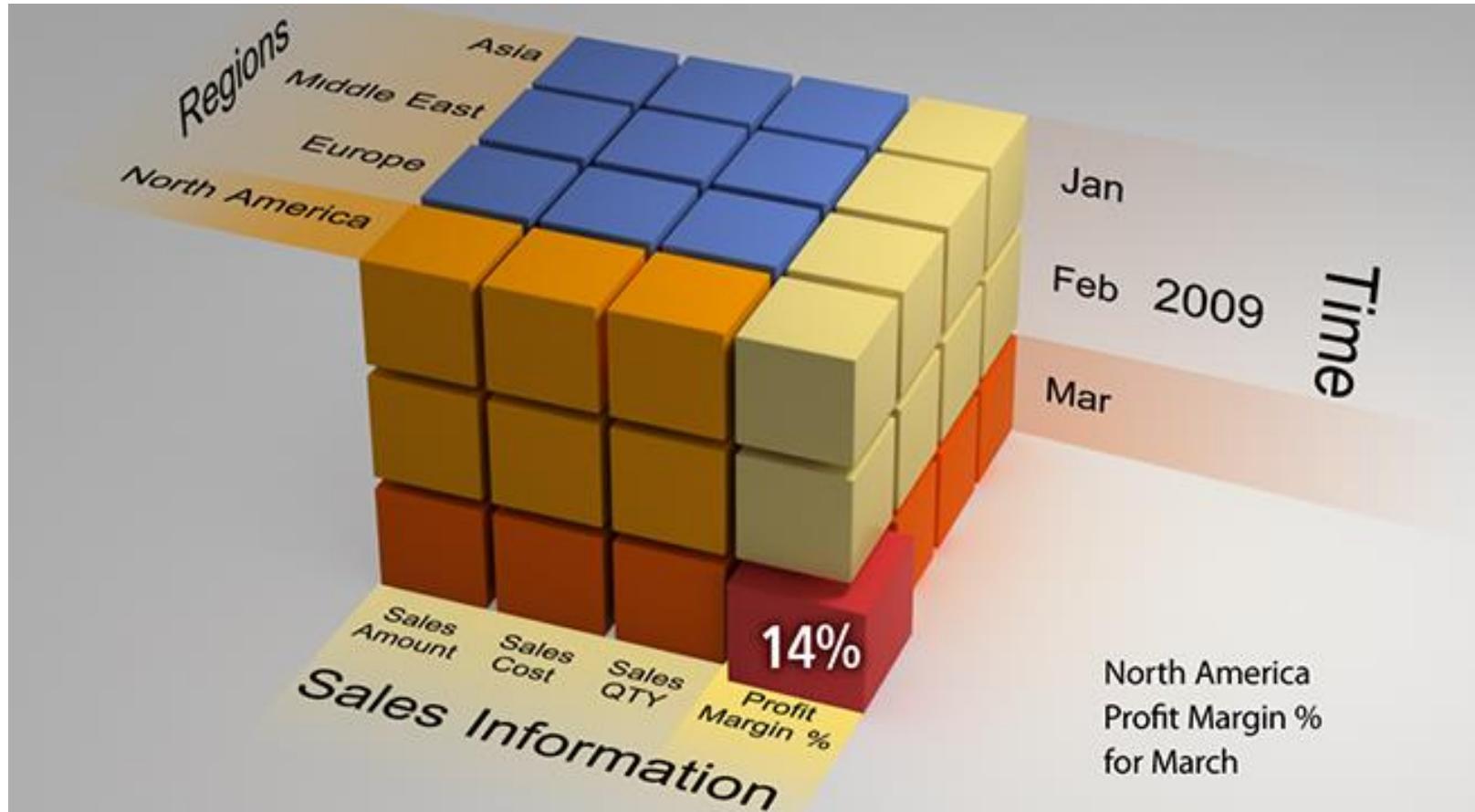


### Dice



## Typical OLAP Operations





## Assignment

---

- Identify an application for each of the data representation you have learnt
- Download a dataset from Kaggle and identify the different types of attributes

## References

---

### Text Book:

- [Data Mining: Concepts and Techniques](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3<sup>rd</sup> Edition
- [Introduction to Data Mining](#), Tan, Steinbach, Kumar, 2<sup>nd</sup> Edition



**THANK YOU**

---

**Dr.Mamatha H R**

Professor,Department of Computer Science

**mamathahr@pes.edu**

+91 80 2672 1983 Extn 834



# DATA ANALYTICS

## Unit 1: Data Exploration

---

**Mamatha.H.R**

Department of Computer Science and  
Engineering

# DATA ANALYTICS

---

## Unit 1: Data Exploration

**Mamatha H R**

Department of Computer Science and Engineering

## What is exploratory data analysis (EDA)?

---

A preliminary exploration of the data to better understand its characteristics.

- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools

## Techniques Used In Data Exploration

---

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization

## Summary Statistics

---

- Summary statistics are numbers that summarize properties of the data
  - Summarized properties include frequency, location and spread
    - Examples: location - mean
    - spread - standard deviation
  - Most summary statistics can be calculated in a single pass through the data

## Data Type

---

- **Cross-Sectional Data:** A data collected on many variables of interest at the same time or duration of time is called cross-sectional data.
- **Time Series Data:** A data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.) is called a time series data.
- **Panel Data:** Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data).

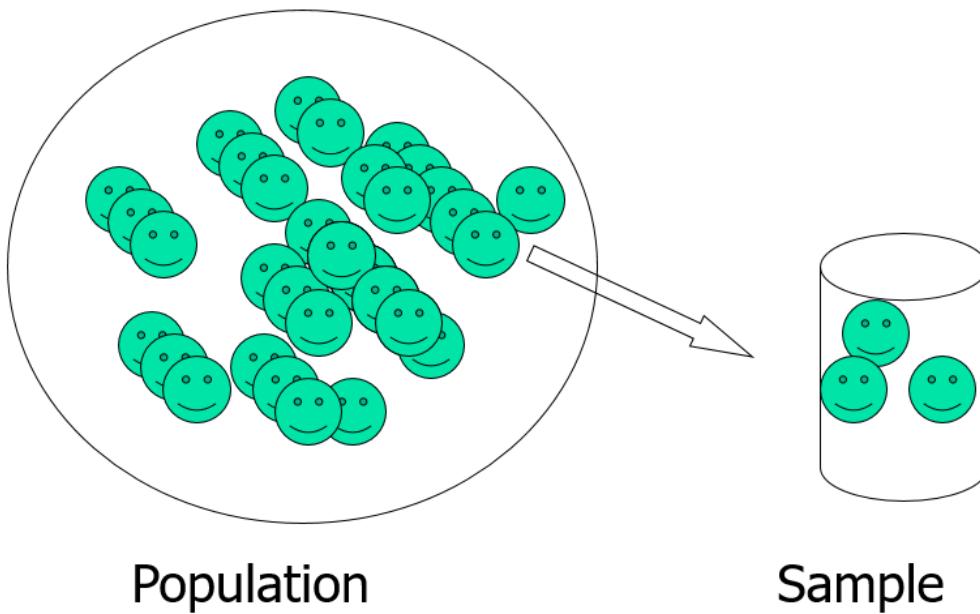
## TYPES OF DATA MEASUREMENT SCALES

---

- **Nominal scale** refers to variables that are basically names (qualitative data) and also known as categorical variables.
- **Ordinal scale** is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.
- **Interval scale** corresponds to a variable in which the value is chosen from an interval set. Variable such as temperature measured in centigrade) or intelligence quotient (IQ) score are examples of interval scale
- Any variable for which the ratios can be computed and are meaningful is called **ratio scale**.

## Population And Sample

- ▶ **Population** is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem.
- ▶ **Sample** is the subset taken from a population.



## TYPES OF SAMPLES

---

- ▶ Systematic sampling – pick every kth sample
  - Time series analysis, sampling pixels from an image, frames in a video,...
- ▶ Simple random sampling with replacement (“infinite population”)
  - Any two values sampled are independent
- ▶ Simple random sampling without replacement (“finite population”)
  - If the sample size is enormous, covariance is nearly zero and the sampling is similar to with replacement
- ▶ Stratified sampling –
  - Example: Analyzing factors that influence performance in high school: Parents / areas divided into strata by income/ occupation
  - Advantages: Samples more likely to be representative (lower sampling error)
  - Disadvantages: Does not work when the population cannot be stratified
- ▶ Cluster sample (two stage cluster sampling – first select cluster, then select samples within the cluster)
  - Example: cluster regions by population size
  - Advantages: feasible, economical
  - Disadvantages: higher sampling error versus simple random sampling (“design effect”); likely biased
- ▶ Convenience sampling
  - Collect the data that is available (eg. survey outside the mall)
- ▶ Sampling in DSP (reconstruction of a signal)
  - Downsampling vs upsampling

## Descriptive Statistics

---

An Illustration: Which Group is Smarter?

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Class B--IQs of 13 Students

127	162
131	103
96	111
80	109
93	87
120	105
109	

*Each individual may be different. If you try to understand a group by remembering the qualities of each member, you become overwhelmed and fail to understand the group.*

## Descriptive Statistics

---

Which group is smarter now?

Class A--Average IQ

110.54

Class B--Average IQ

110.23

They're roughly the same!

With a summary descriptive statistic, it is much easier to answer our question.

## Descriptive Statistics

---

### Types of descriptive statistics:

- Organize Data
  - Tables
  - Graphs
- Summarize Data
  - Central Tendency
  - Variation

## Descriptive Statistics

---

### Types of descriptive statistics:

- Organize Data
  - Tables
    - Frequency Distributions
    - Relative Frequency Distributions
  - Graphs
    - Bar Chart or Histogram
    - Stem and Leaf Plot
    - Frequency Polygon

## Descriptive Statistics

---

### Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode
  
- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

## Measures Of Central Tendency

---

- Mean (or Average) Value

Mean is the arithmetical average value of the data and is one of the most frequently used measures of central tendency.

$$\text{Mean} = \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n}$$

## Mean

---

Symbol  $\bar{x}$  is frequently used to represent the estimated value of the mean from a sample. If the entire population is available and if we calculate mean based on the entire population, then we have the population mean which is denoted by  $\mu$  (population mean).

### Property of Mean

An important property of mean is that the summation of deviation of observations from the mean is zero, that is

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

## Median (or Mid) Value

---

- Median is the value that divides the data into two equal parts, that is, the proportion of observations below median and above median will be 50%.
- Easiest way to find the median value is by arranging the data in the increasing order and the median is the value at position  $(n + 1)/2$  when  $n$  is odd. When  $n$  is even, the median is the average value of  $(n/2)^{\text{th}}$  and  $(n + 2)/2^{\text{th}}$  observation after arranging the data in the increasing order.

## Mode

---

- **Mode** is the most frequently occurring value in the dataset
- Mode is the only measure of central tendency which is valid for qualitative (nominal) data since the mean and median for nominal data are meaningless.
- For example, assume that a customer data with a retailer has the marital status of customer, namely, (a) Married, (b) Unmarried, (c) Divorced Male, and (d) Divorced Female. Mean and median are meaningless when we try to use them on a qualitative data such as marital status. On the other hand, mode will capture the customer type in terms of marital status that occurs most frequently in the database

## Percentile

---

- **Percentile**, decile and quartile are frequently used to identify the position of the observation in the dataset.
- Percentile, denoted as  $P_x$ , is the value of the data at which  $x$  percentage of the data lie below that value

Position corresponding to  $P_x \approx x(n+1)/100$

- $P_x$  is the position in the data calculated , where  $n$  is the number of observations in the data.

## Decile and Quartile

---

- ▶ **Decile** corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on.
  
- ▶ **Quartile** divides the data into 4 equal parts. The first quartile ( $Q_1$ ) contains first 25% of the data,  $Q_2$  contains 50% of the data and is also the median. Quartile 3 ( $Q_3$ ) accounts for 75% of the data

## Example

### Time between failures of wire-cut (in hours)

2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

1. Calculate the mean, median, and mode of time between failures of wire-cuts
2. The company would like to know by what time 10% (ten percentile or  $P_{10}$ ) and 90% (ninety percentile or  $P_{90}$ ) of the wire-cuts will fail?
3. Calculate the values of  $P_{25}$  and  $P_{75}$ .

## Solution

---

- 1) Mean = 57.64, median = 56, and mode = 46
- 2) Note that the data in Table is arranged in increasing order in columns. The position of  $P_{10} = 10 \times (51)/100 = 5.1$ . We can round off 5.1 to its nearest integer which is 5. The corresponding value from table is 21 (10 percentage of observations in Table have a value of less than or equal to 21). That is, by 21 hours, 10% of the wire-cuts will fail. In asset management (and reliability theory), this value is called  $P_{10}$  life.

Instead of rounding the value obtained from Eq, we can use the following approximation:

$$P_{10} = 10 \times (51)/100 = 5.1$$

Value at 5<sup>th</sup> position is 21. Value at position 5.1 is approximated as  $21 + 0.1 \times (\text{value at } 6^{\text{th}} \text{ position} - \text{value at } 5^{\text{th}} \text{ position}) = 21 + 0.1(1) = 21.1$

$$P_{90} = 90 \times 51/100 = 45.9$$

The value at position 45 is 90 and at position 45.9 is

$$90 + 0.9 \times (3) = 92.7$$

That is, 90% of the wire-cuts will fail by 92.7 hours

$P_{25}$  (1<sup>st</sup> Quartile or  $Q_1$ ) =  $25 \times 51/100 = 12.75$  ,  
Value at 12<sup>th</sup> position is 33, so

$P_{25} = 33 + 0.75 (\text{value at } 13^{\text{th}} \text{ position} - \text{value at } 12^{\text{th}} \text{ position}) = 33 + 0.75 (1) = 33.75$

$P_{75}$  (3<sup>rd</sup> Quartile or  $Q_3$ ) =  $75 \times 51/100 = 38.25$

Value at 38<sup>th</sup> position is 86, so

$P_{75} = 86 + 0.25 (\text{value at } 39^{\text{th}} \text{ position} - \text{value at } 38^{\text{th}} \text{ position}) = 86 + 0.25 (0) = 86$

## Measures of Variation

---

- Predictive analytics techniques such as regression attempt to explain variation in the outcome variable (Y) using predictor variables (X)
- Variability in the data is measured using the following measures:
  - Range
  - Inter-Quartile Distance (IQR)
  - Variance
  - Standard Deviation

## Range, IQD and Variance

---

- **Range** is the difference between maximum and minimum value of the data. It captures the data spread.
- **Inter-quartile distance (IQD)**, also called inter-quartile range (IQR) is a measure of the distance between Quartile 1 ( $Q_1$ ) and Quartile 3 ( $Q_3$ )
- **Variance** is a measure of variability in the data from the mean value. Variance for population,  $\sigma^2$ , is calculated using

$$\text{Variance} = \sigma^2 = \sum_{i=1}^n \frac{(X_i - \mu)^2}{n}$$

## Sample Variance

---

- In case of a sample, the Sample Variance ( $S^2$ ) is calculated using

$$S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$$

- While calculating sample variance  $S^2$ , the sum of squared deviation  $\sum_{i=1}^n (x_i - \bar{x})^2$  is divided by  $(n-1)$ , this is known as Bessel's correction.

## Standard Deviation

---

- The population standard deviation ( $\sigma$ ) and sample standard deviation ( $S$ ) are given by

$$\sigma = \sqrt{\sum_{i=1}^n \frac{(X_i - \mu)^2}{n}} \quad S = \sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}}$$

## Degrees of Freedom

---

- **Degrees of freedom** is equal to the number of independent variables in the model (Trochim, 2005). For example, we can create any sample of size  $n$  with mean value of  $\bar{x}$  by randomly selecting  $(n - 1)$  values. We need to fix just one out of  $n$  values. Thus the number of independent variables in this case is  $(n - 1)$
- Degrees of freedom is defined as the difference between the number of observations in the sample and number of parameters estimated (Walker 1940, Toothaker and Miller, 1996). If there are  $n$  observations in the sample , then the degrees of freedom is  $(n - k)$ .

## Measures of Shape – Skewness and Kurtosis

---

- **Skewness** is a measure of symmetry or lack of symmetry. A dataset is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. That is, the proportion of data between  $\mu$  and  $\mu - k\sigma$  is same as  $\mu$  and  $\mu + k\sigma$ , where  $k$  is some positive constant.
- **Pearson's moment coefficient of skewness** for a dataset with  $n$  observations is given by

$$g_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3 / n}{\sigma^3}$$

- The value of  $g_1$  will be close to 0 when the data is symmetrical. A positive value of  $g_1$  indicates a positive skewness and a negative value indicates **negative skewness**.

## Skewness

---

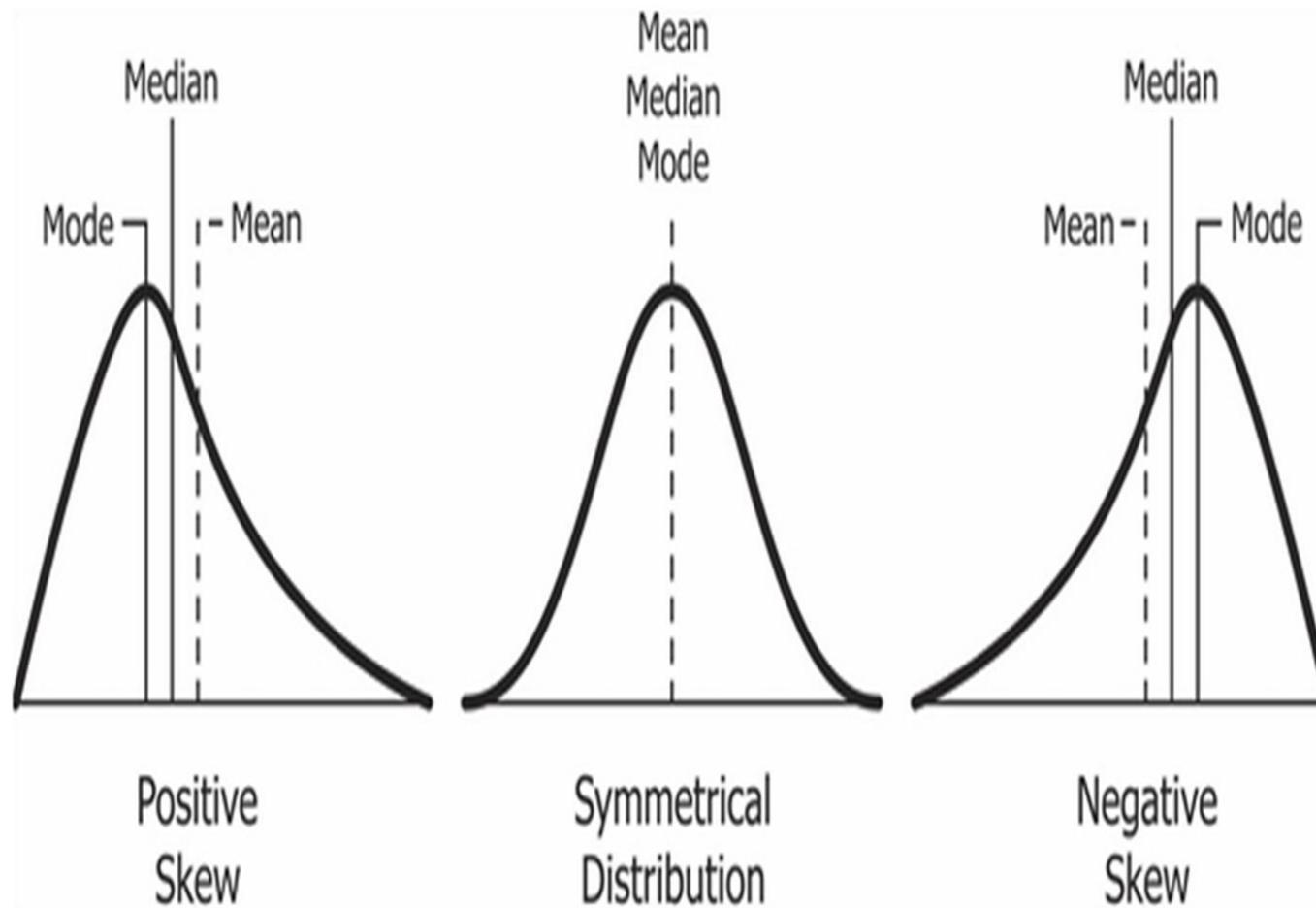
- The following formula is used usually for a sample with  $n$  observations (Joanes and Gill, 1998):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

- The value of  $\frac{\sqrt{n(n-1)}}{n-2}$  will converge to 1 as the value of  $n$  increases.

## Symmetric vs. Skewed Data

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



## Kurtosis

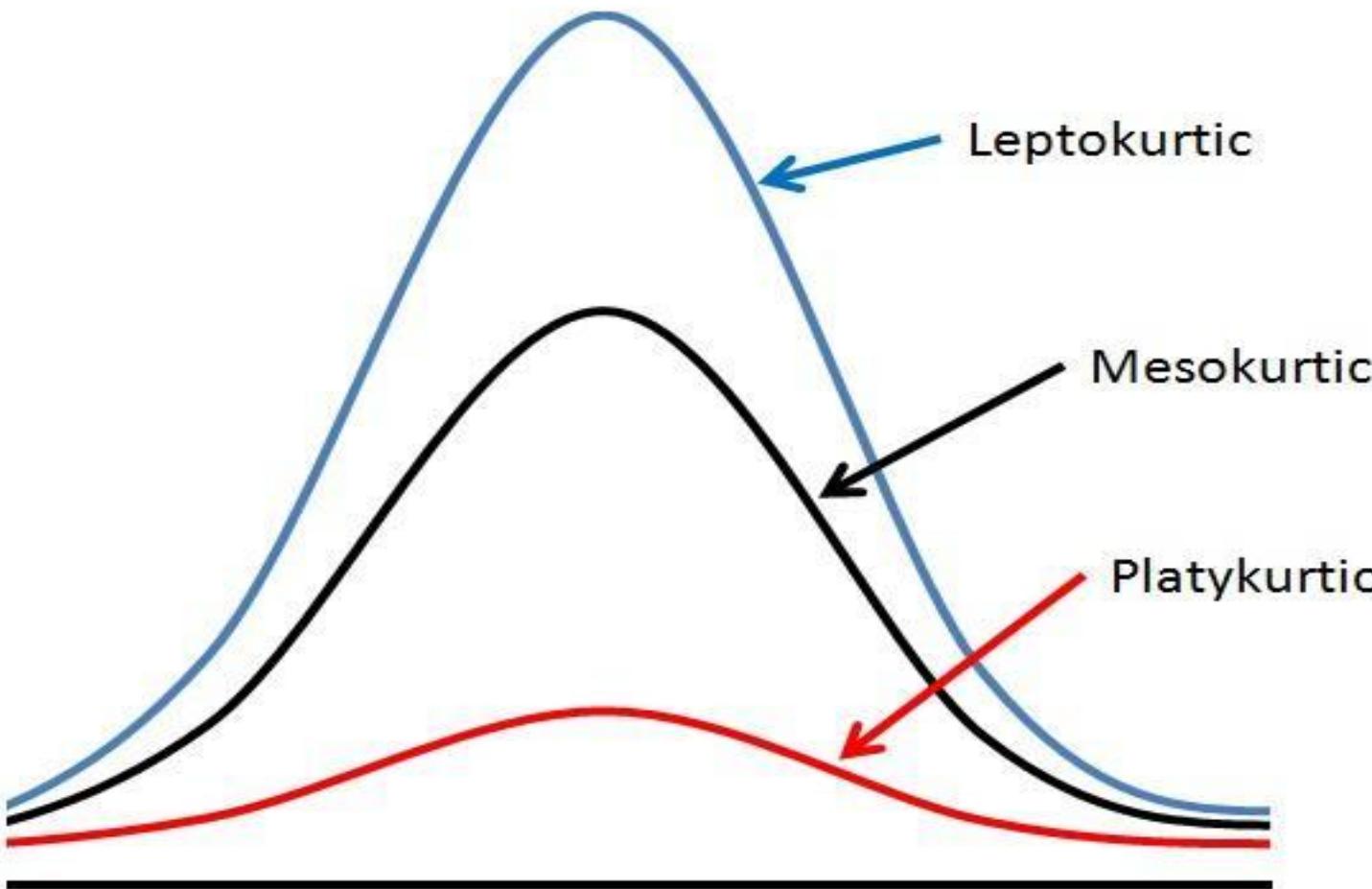
---

- **Kurtosis** is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis is measured using the following equation:

$$\text{Kurtosis} = \frac{\sum_{i=1}^4 (X_i - \bar{X})^4 / n}{\sigma^4}$$

- Kurtosis value of less than 3 is called **platykurtic distribution** and greater than 3 is called **leptokurtic distribution**. The kurtosis value of 3 indicates standard normal distribution (also called **mesokurtic**)

## Leptokurtic, mesokurtic, and platykurtic distributions



## Excess Kurtosis

---

- The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is given by:

$$\text{Excess Kurtosis} = \frac{\sum_{i=1}^4 (X_i - \bar{X})^4 / n}{\sigma^4} - 3$$

## Exercise

---

The daily footfall at a retail store in Bangalore over the last 30 days is shown in Table 1. calculate the Mean, Median , Mode and Standard Deviation.

**Table 1.Footfall data**

232	277	261	173	283	197	251	212	213	213
229	164	219	196	186	247	244	269	216	272
252	314	161	165	221	260	219	290	225	251

For the data in Table 1, calculate the skewness and kurtosis. what can you infer from the skewness and kurtosis of the football data?

For the data in Table 1, calculate the values of first quartile and third quartile. Are there any outliers in the data?

## References

---

### Text Book:

- [“Business Analytics, The Science of Data-Driven Decision Making”](#), U. Dinesh Kumar, Wiley 2017
- [Data Mining: Concepts and Techniques](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3<sup>rd</sup> Edition.
- [Introduction to Data Mining](#), Tan, Steinbach, Kumar, 2<sup>nd</sup> Edition



**THANK YOU**

---

**Dr.Mamatha H R**

Professor,Department of Computer Science

**mamathahr@pes.edu**

+91 80 2672 1983 Extn 834



## DATA ANALYTICS

### Unit 1: Data Exploration (contd.)

---

**Mamatha.H.R**

Department of Computer Science and  
Engineering

**Gowri Srinivasa**

Department of Computer Science and  
Engineering

# DATA ANALYTICS

---

## Unit 1: Exploratory data analysis

**Mamatha H R, Gowri Srinivasa**

Department of Computer Science and Engineering

## TYPES OF DATA MEASUREMENT SCALES

---

- **Nominal scale** refers to variables that are basically names (qualitative data) and also known as categorical variables.
- **Ordinal scale** is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.
- **Interval scale** corresponds to a variable in which the value is chosen from an interval set. Variable such as temperature measured in centigrade) or intelligence quotient (IQ) score are examples of interval scale
- Any variable for which the ratios can be computed and are meaningful is called **ratio scale**. Concept of a **true zero**.

# DATA ANALYTICS

## TYPES OF DATA

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ( $=, \neq$ )	zip codes, employee ID numbers, eye color, sex: { <i>male, female</i> }	mode, entropy, contingency correlation, $\chi^2$ test
Ordinal	The values of an ordinal attribute provide enough information to order objects. ( $<, >$ )	hardness of minerals, { <i>good, better, best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
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, <i>t</i> and <i>F</i> 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

## Operations on Data

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it 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}.
Interval	$new\_value = a * old\_value + b$ where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	$new\_value = a * old\_value$	Length can be measured in meters or feet.

What is the data type for each of these?

---

- Qualitative/ Quantitative, Ordinal/Interval/ Ratio, Continuous/ Discrete
- Time in terms of AM or PM
  - Binary/ Qualitative/ Ordinal
- Brightness as measured by a light meter
  - Continuous, quantitative, ratio
- Brightness as measured by people's judgments
  - Discrete, qualitative, ordinal
- Angles as measured in degrees between 0 and 360
  - Continuous, quantitative, ratio
- Bronze, Silver, and Gold medals as awarded at the Olympics
  - Discrete, qualitative, ordinal
- Height above sea level
  - Continuous, quantitative, interval/ratio  
(depends on whether sea level is regarded as an arbitrary origin)
- Number of patients in a hospital
  - Discrete, quantitative, ratio
- ISBN numbers for books
  - Discrete, qualitative, nominal (ISBN numbers do have order information, though)

## Degrees of Freedom

- Degrees of freedom is equal to the number of independent variables in the model (Trochim, 2005). For example, we can create any sample of size  $n$  with mean value of  $\bar{x}$  by randomly selecting  $(n - 1)$  values. We need to fix just one out of  $n$  values. Thus the number of independent variables in this case is  $(n - 1)$

Values	6	8	5	9	6	8	4	11	7	X
Average	6.9									
Sum	69									

$$\begin{aligned}X &= 69 - \text{sum(values)} \\&= 69 - 64 \\&= 5\end{aligned}$$

## Chebyshev's Theorem

---

- **Chebyshev's theorem** (also known as Chebyshev's inequality) is an empirical rule that allows us to predict proportion of observations that is likely to lie between an interval defined using mean and standard deviation. Probability of finding a randomly selected value in an interval defined by  $\mu \pm k\sigma$  is  $1 - \frac{1}{k^2}$  that is

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

## Example

---

- Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000?

- **Solution:**

$$P(8000 \leq X \leq 16000)$$

$$= P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75$$

That is, the proportion of customers spending between 8000 and 16000 is at least 0.75 (or 75%)

# DATA ANALYTICS

---

## Unit 1: Data Visualization

**Mamatha H R, Gowri Srinivasa**

Department of Computer Science and Engineering

## Histogram

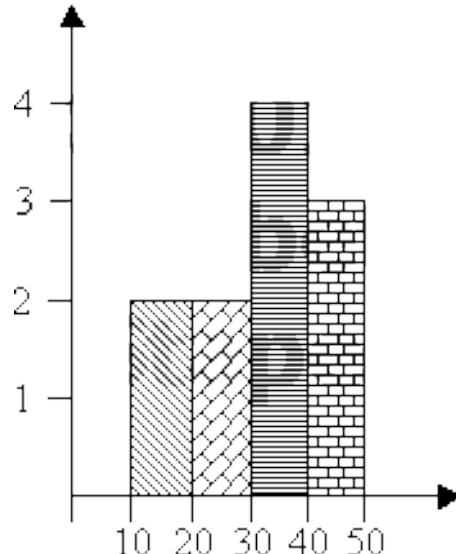
---

- **Histogram** is the visual representation of the data which can be used to assess the probability distribution (frequency distribution) of the data
- Histograms are created for continuous (numerical) data.
- It is a frequency distribution of data arranged in consecutive and non-overlapping intervals

# Histograms and stem-and-leaf plots

Frequency Class	Frequency
-----------------	-----------

10 - 19	2
20 - 29	2
30 - 39	4
40 - 49	3



There are two samples in the first two bins, but what are the exact values of each?

Details are obscured

stem	leaf
1	2 3
2	1 7
3	3 4 5 7
4	0 0 1

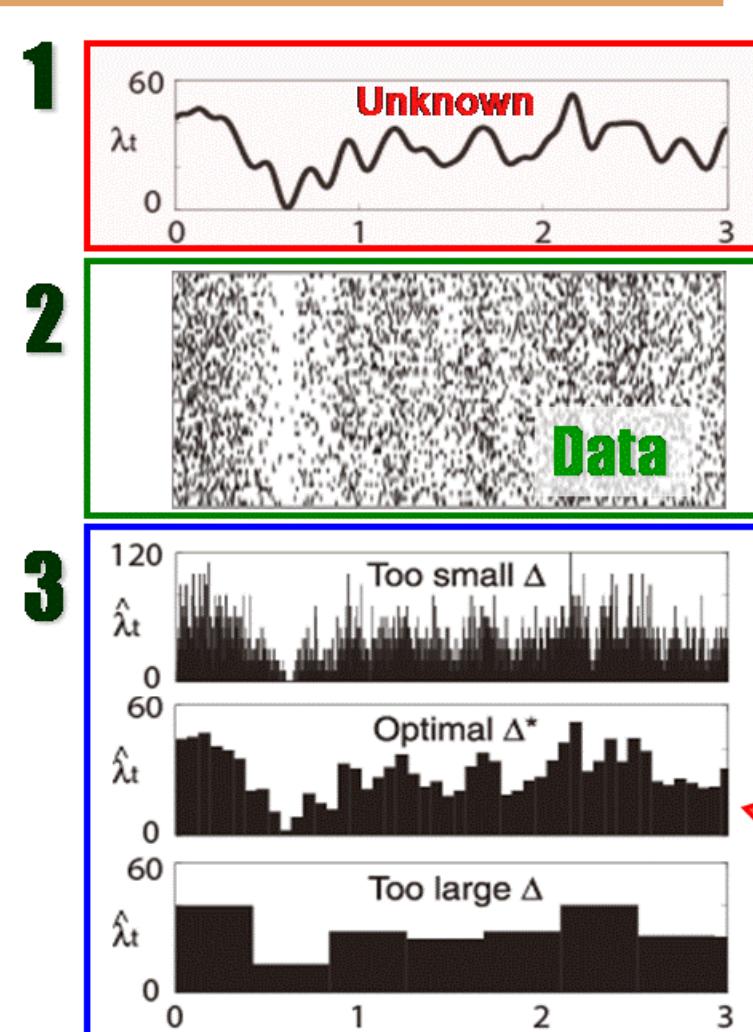
12, 13, 21, 27, 33, 34, 35, 37, 40, 40, 41

# Histogram

- Effective for showing both skew and kurtosis

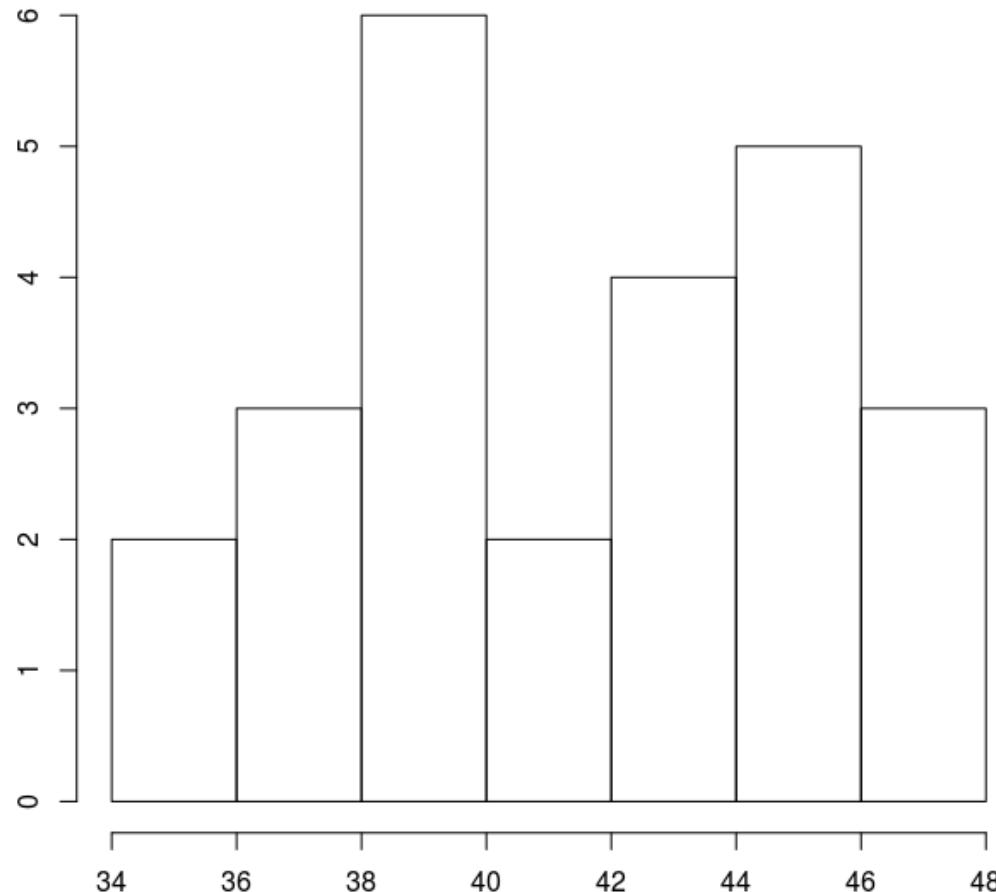
Optimal bin size?

- Bin size too small – noisy
- Bin size too large – details are smoothed out

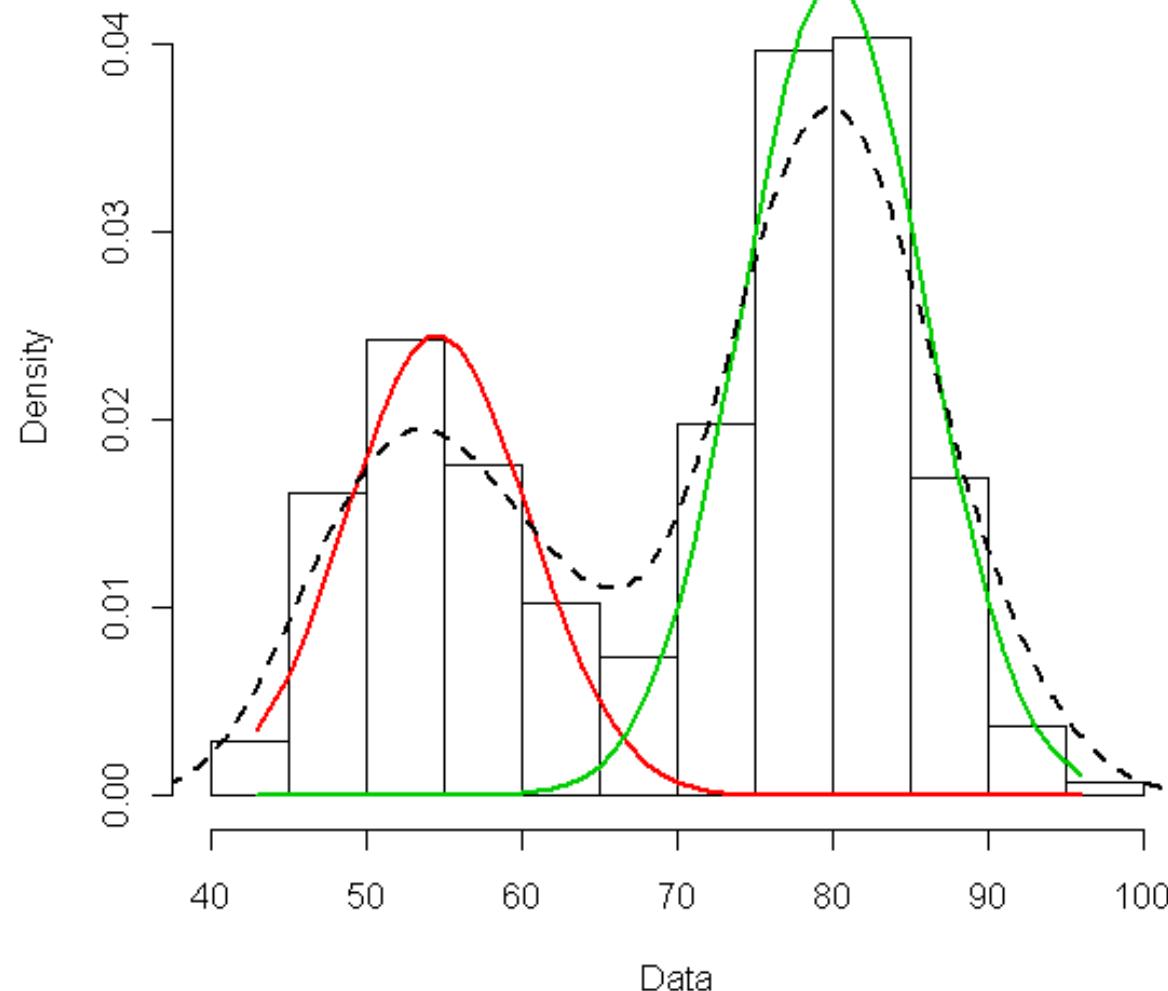


# Is this histogram unimodal or bimodal?

---



# A biomodal function – mixture of 2 Gaussians

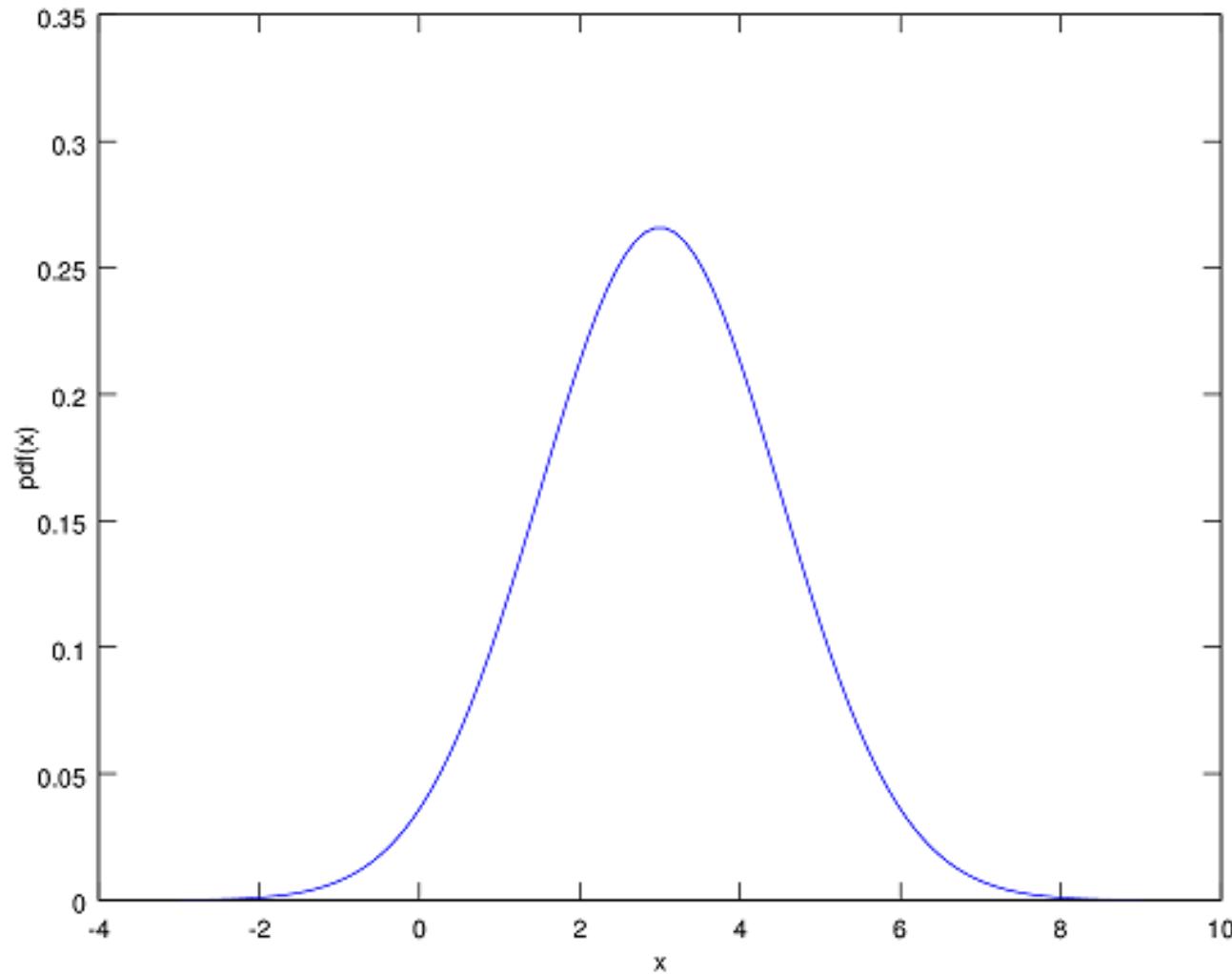


Mixtools  
package in R

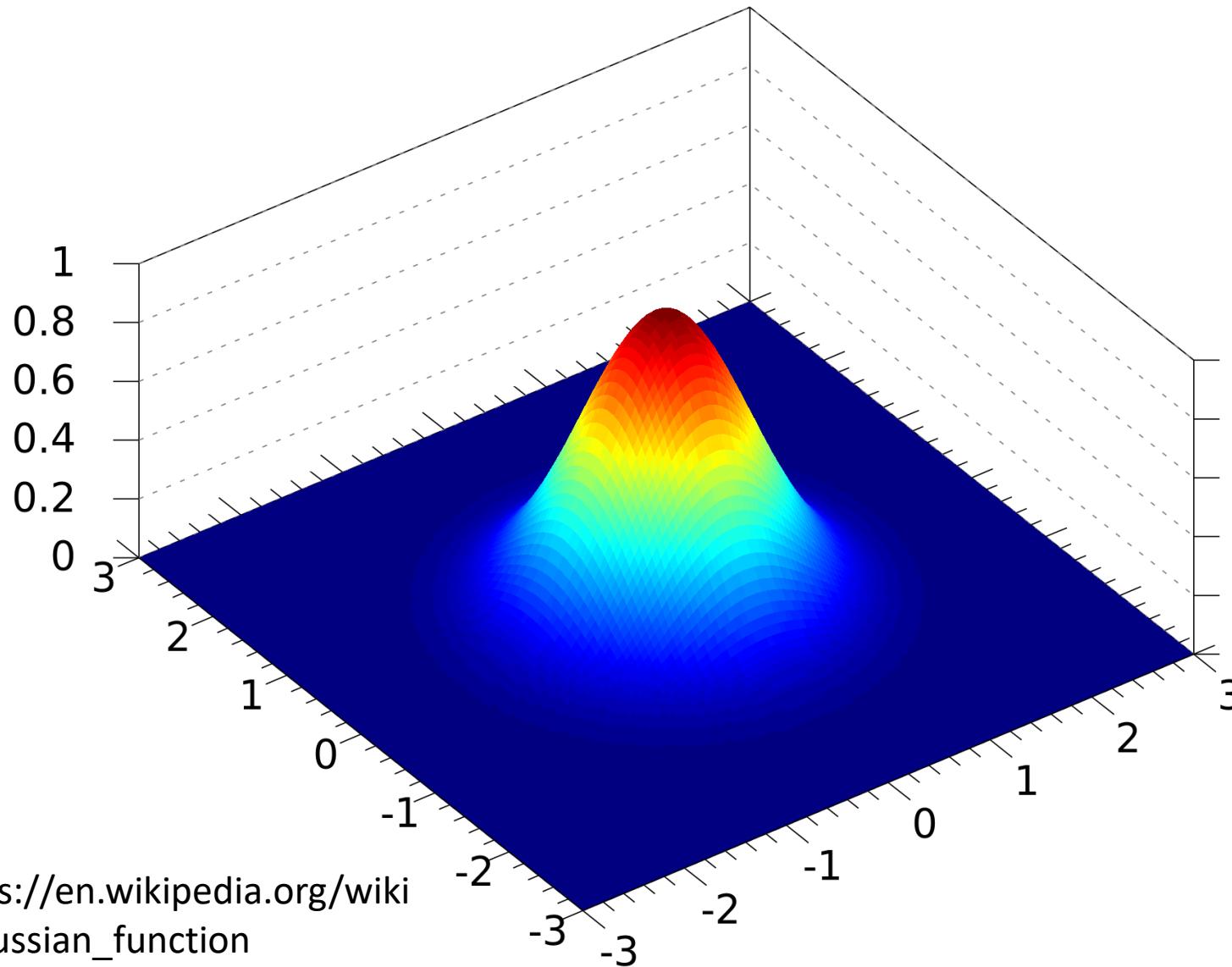
Bin size (size of the interval) affects the number of ‘modes’ we see in the mixture

# Univariate Gaussian

---

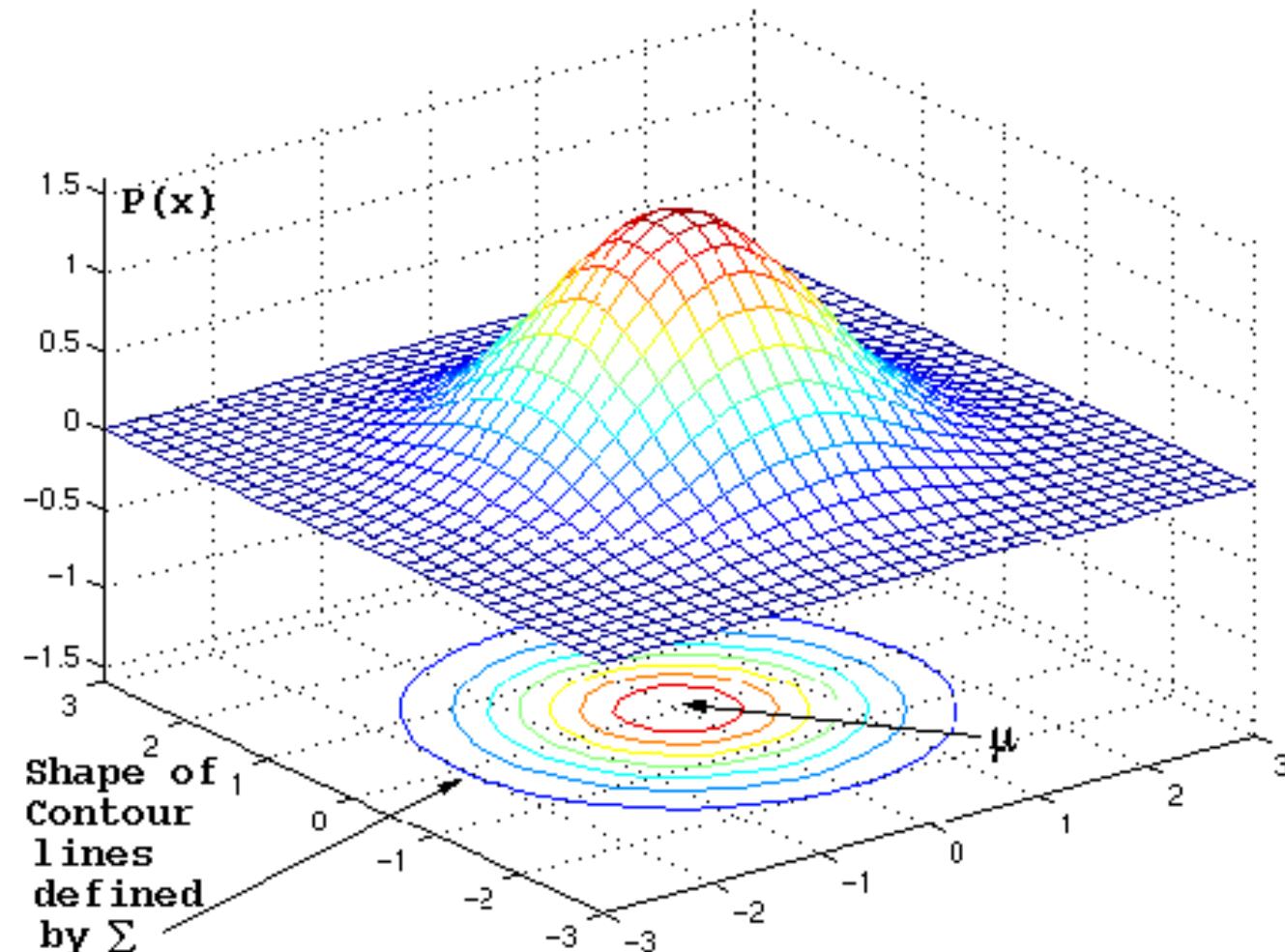


# Bivariate Gaussian



[https://en.wikipedia.org/wiki/Gaussian\\_function](https://en.wikipedia.org/wiki/Gaussian_function)

# Sections of a Bivariate Gaussian



# Covariance matrix for the Bivariate

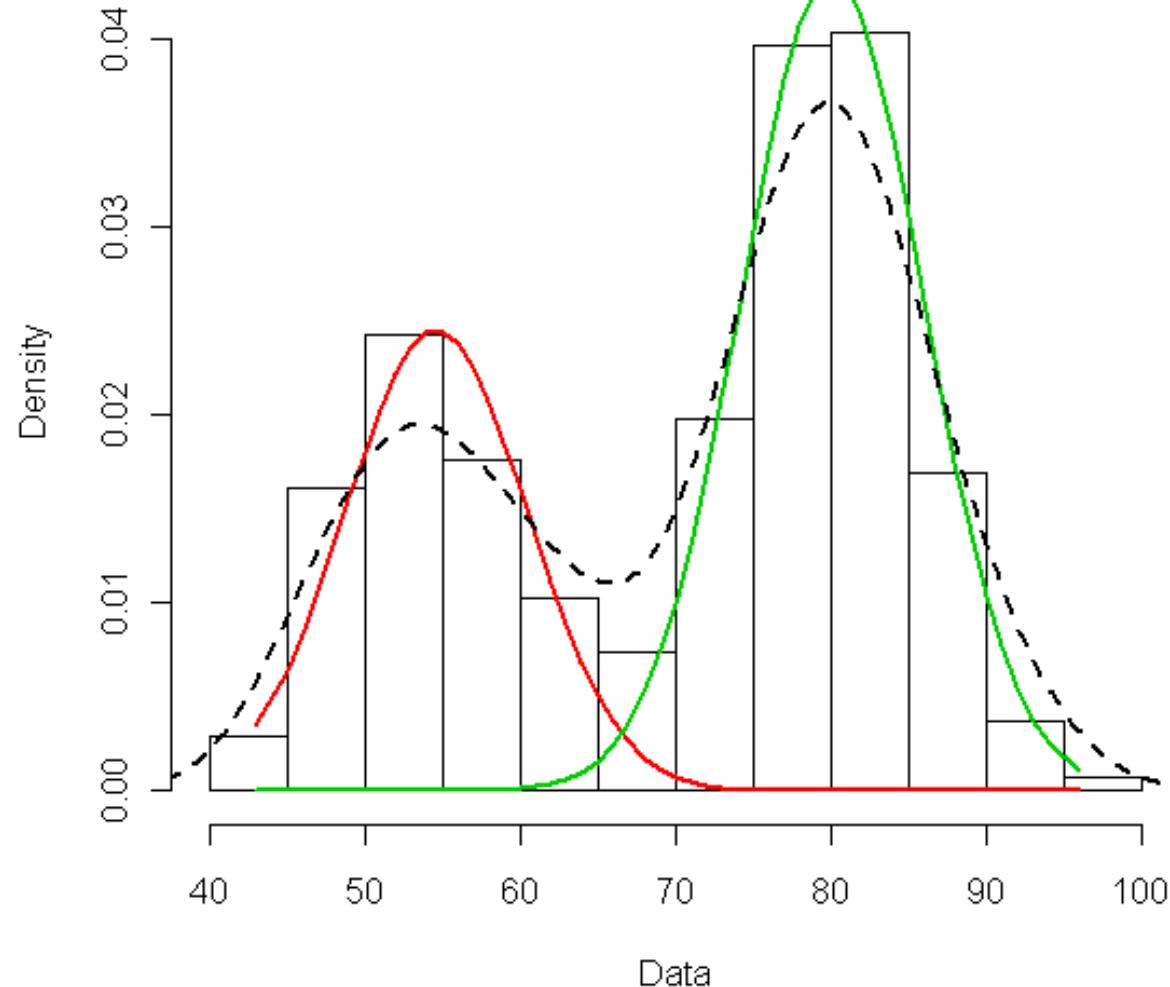
---

- Expectation of the outer product of  $\mathbf{x}-\boldsymbol{\mu}$ , where  $\mathbf{x}$  and  $\boldsymbol{\mu} \in \mathbb{R}^d$

$$\Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]$$

Equation of the pdf etc., on the board

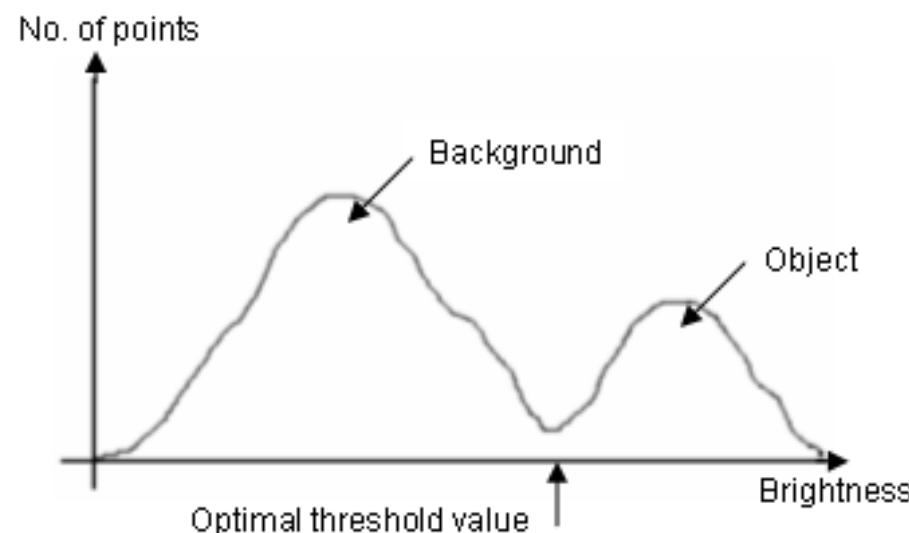
# A biomodal function – mixture of 2 Gaussians



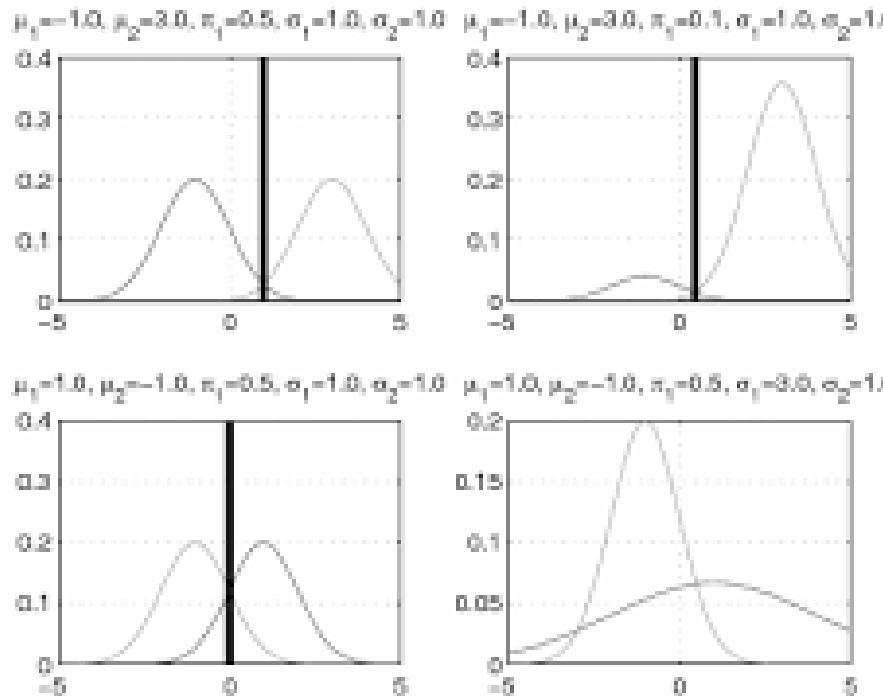
Mixtools  
package in R

# Otsu method

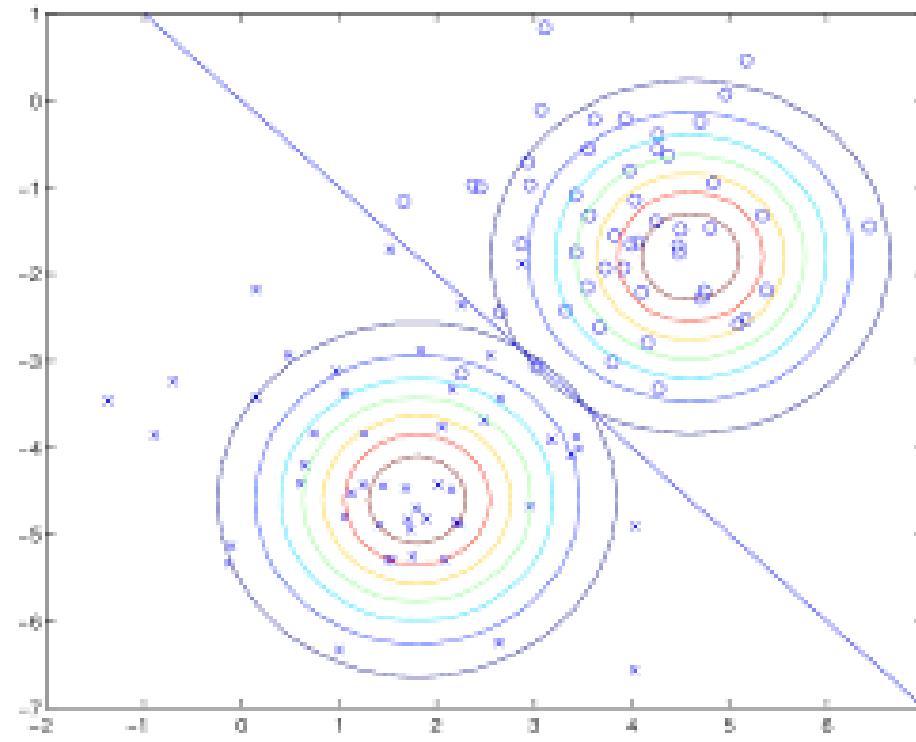
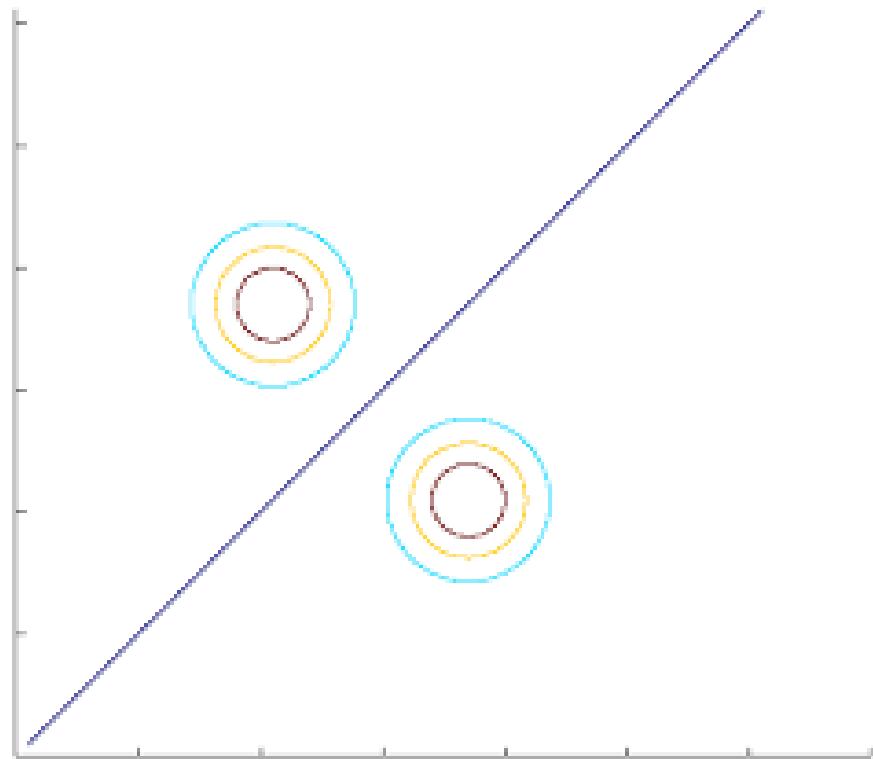
---



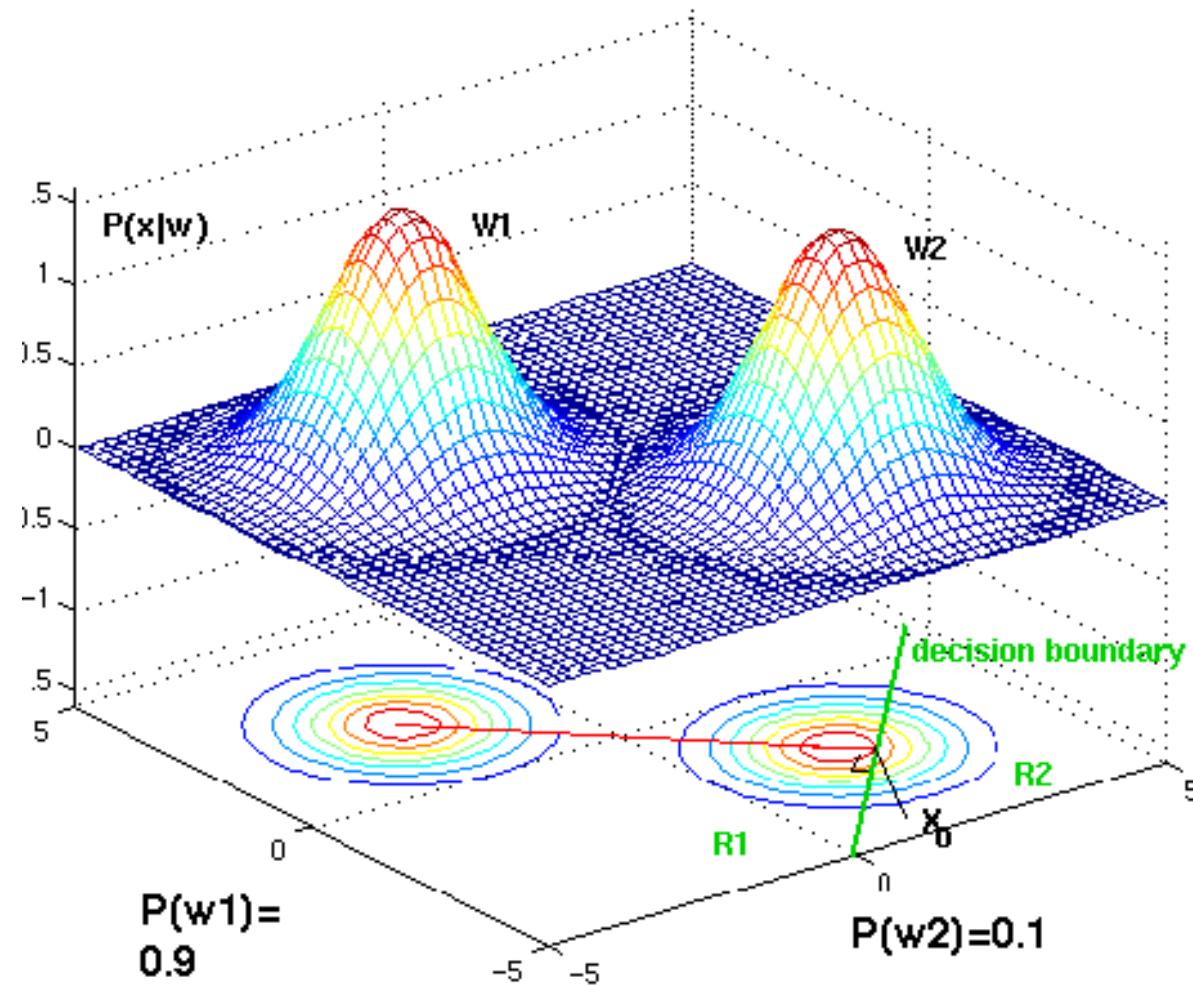
# Separating mixtures or decision boundaries



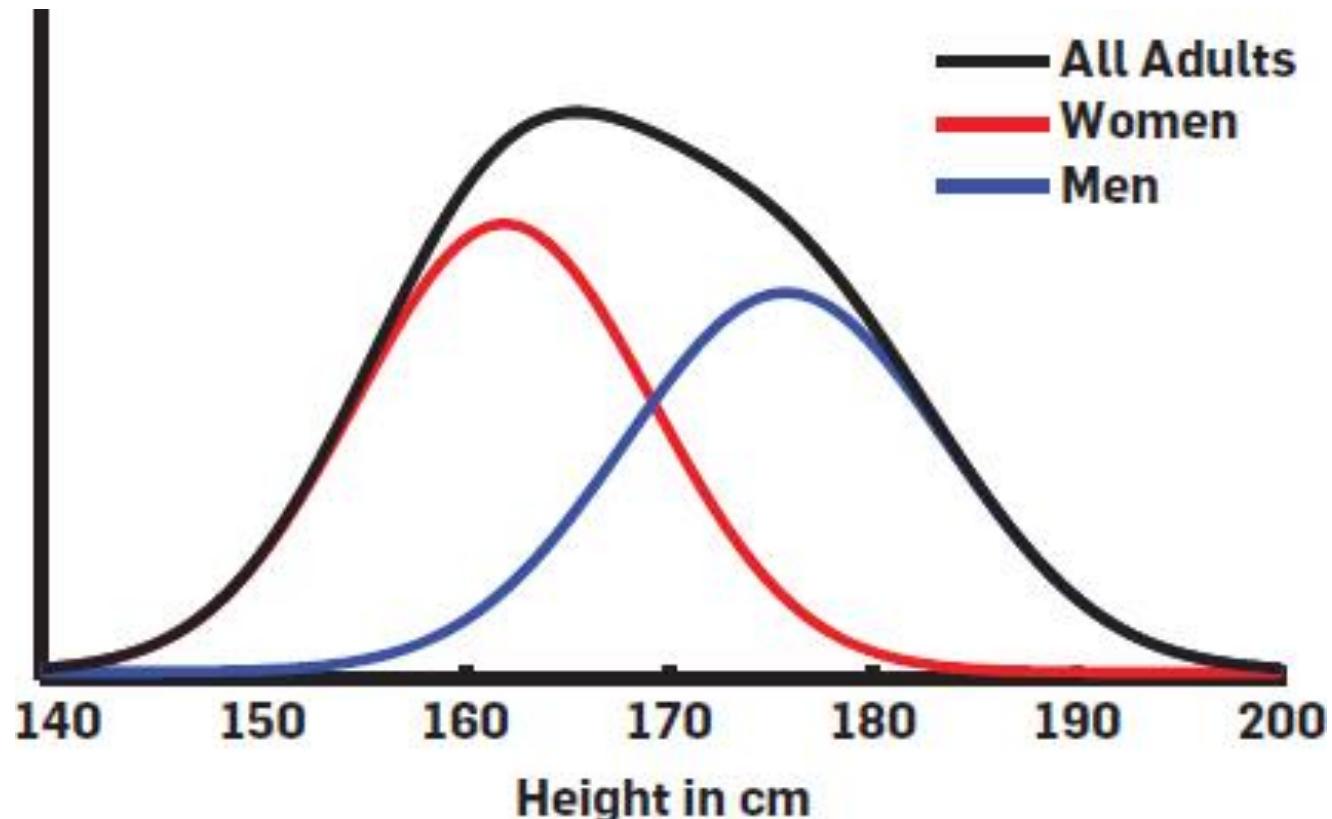
# Boundaries in higher dimensions



# Bivariate Gaussians – not equiprobable



# Indistinguishable mixture



# Tests for bimodality

---

- Necessary condition:
  - Kurtosis – (skewness)<sup>2</sup> ≤ 1 (Pearson's criterion)
  - Equality holds in the extreme case of two diracs
- Is the distribution anything other than Unimodal?
  - Dip test (Hartigan's dip test statistic or HDS)
    - $p < 0.05$  significant multimodality,
    - $0.05 < p < 0.10$  multimodality with marginal significance
  - Mixtools, flexmix, mcclust, mcdist
- Silverman's mode detection method

<http://adereth.github.io/blog/2014/10/12/silvermans-mode-detection-method-explained/>

## Use of Histogram

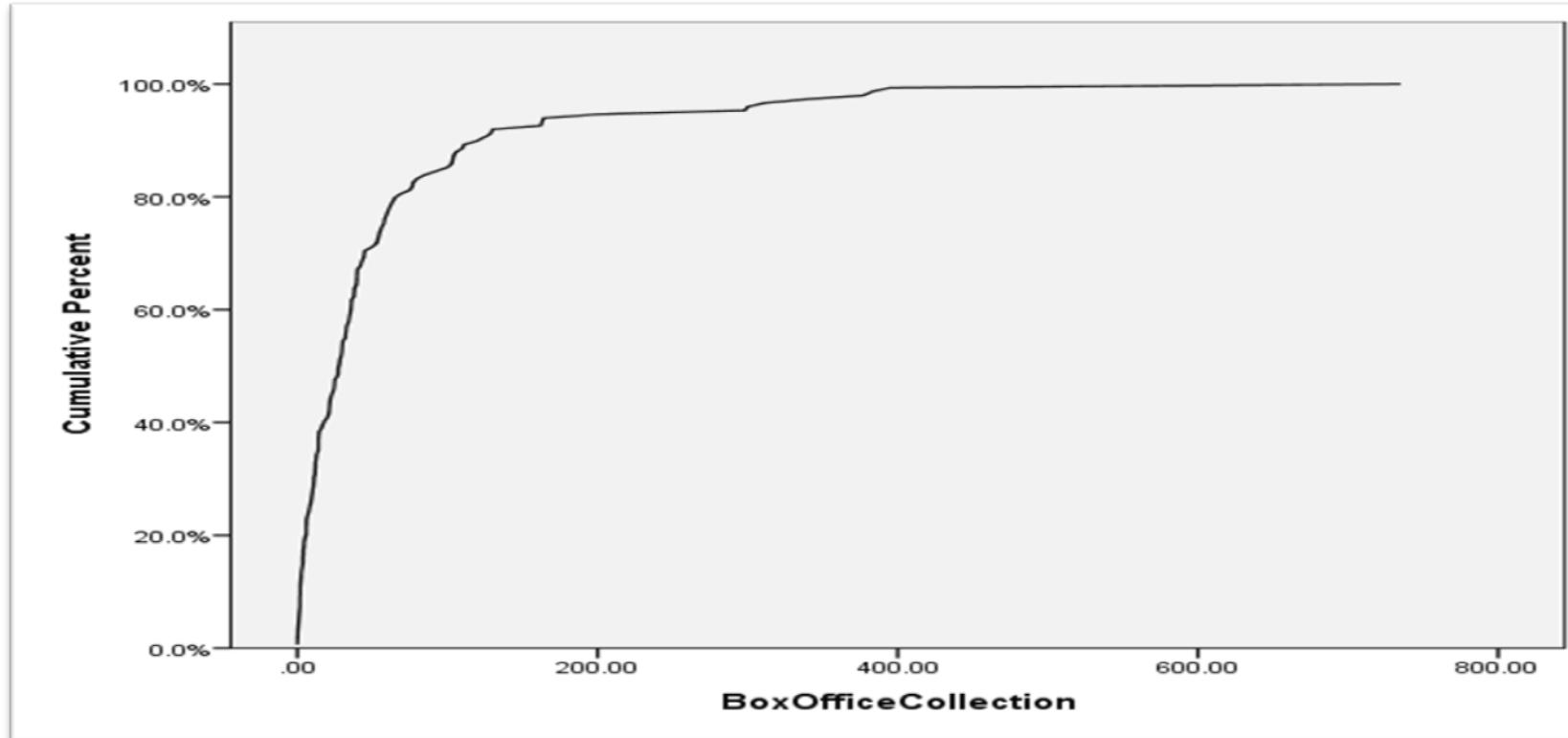
---

Histogram is very useful since it assists data scientist to identify the following:

- The shape of the distribution and to assess the probability distribution of the data.
- Measures of central tendency such as median and mode.
- Measures of variability such as spread.
- Measure of shape such as skewness

## Ogive Curves

- The cumulative histograms are called **Ogive curves**. The Ogive curve for Bollywood box-office collection is shown below:

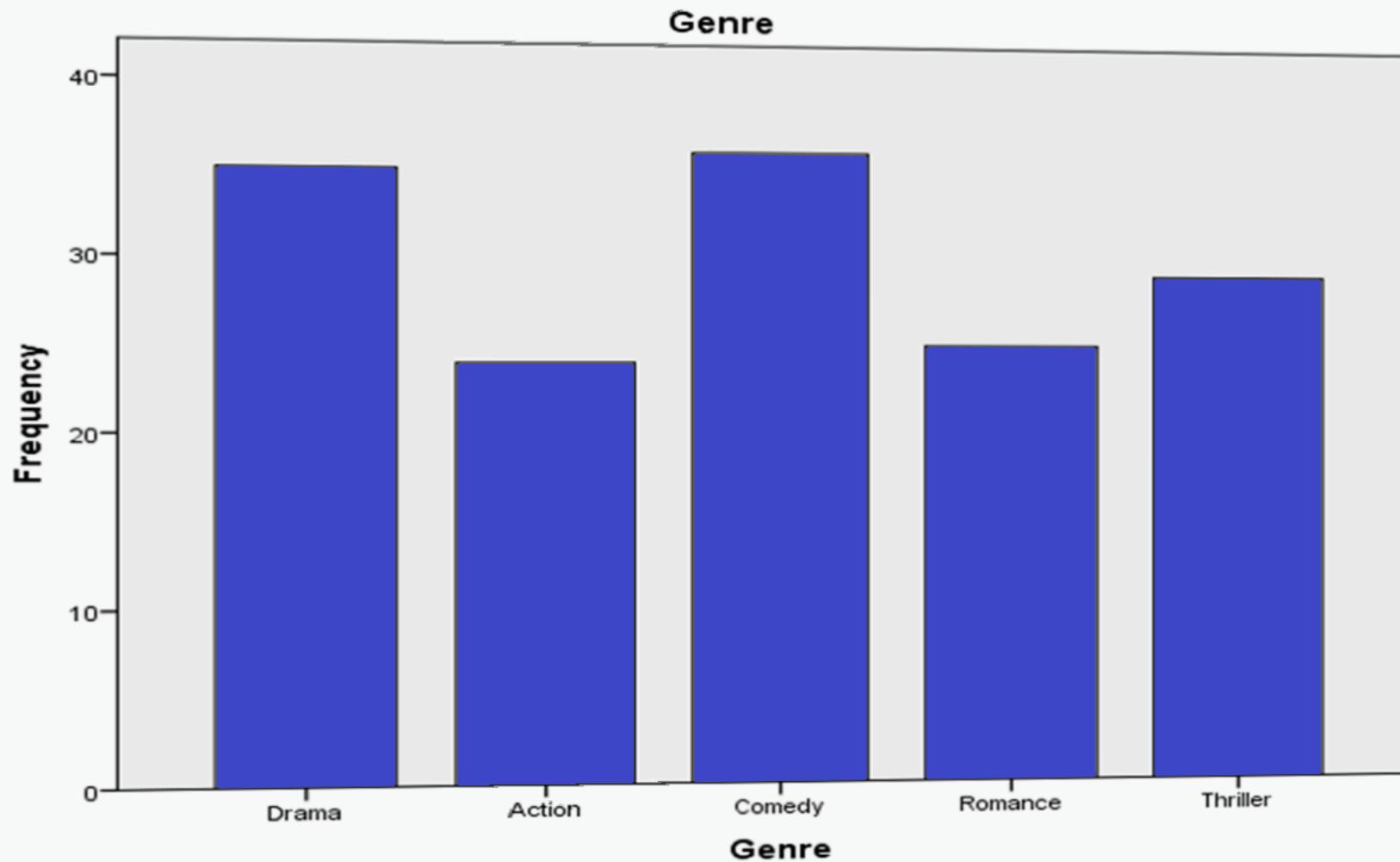


## Bar Chart

---

- **Bar chart** is a frequency chart for qualitative variable (or categorical variable)
- Bar chart can be used to assess the most-occurring and least-occurring categories within a dataset
- Histograms cannot be used when the variable is qualitative

## Bar chart for movie genre



# Collecting data – an example

---

- Problem:  
What are the factors that drive academic excellence?
- Case Study:  
What influences VIII standard marks?

What we came up with in class apart from the list published by NCERT

- Average income per household in the area
- Basic utilities in the area (#hrs of uninterrupted power/ water supply)
- #hours spent on study, #hours spent on social media
- Average #hours spent by educated an educated guardian at home
- Average time spent commuting to school and back/ mode of transport



## What influences Class VIII marks?

Impact ▾ Across ▾ Subject ▾

These results are based on 185,348 students across India.

The factors that influence the marks the **most** (and their impact in marks) are:

1	Father edu	+11.0%
2	Father occupation	+10.6%
3	Mother edu	+10.5%
4	Mother occupation	+9.1%
5	Help in household	+5.2%

The factors that influence the marks the **least** (and their impact in marks) are:

1	Gender	+0.7%
2	Distance	+0.8%
3	Private tuition	+1.3%
4	Watch TV	+2.1%
5	Siblings	+2.4%

Gender ▾ State ▾ Below poverty ▾ Siblings ▾

Factor	Total %	Maths %	Reading %	Science %	Social %
Gender	0.7%	0.2%	1.9%	0.1%	0.5%
Age	4.1%	3.1%	8.0%	3.3%	2.9%
Siblings	2.4%	1.5%	8.3%	2.2%	0.6%
Father edu	11.0%	6.6%	18.8%	9.9%	7.9%
Mother edu	10.5%	4.3%	18.4%	10.3%	7.7%
Father occupation	10.6%	8.7%	17.4%	10.3%	8.5%
Mother occupation	9.1%	5.7%	14.9%	7.3%	7.9%
Below poverty	3.3%	1.8%	5.7%	2.6%	3.0%
Use calculator	2.7%	0.9%	5.0%	2.7%	2.1%
Use dictionary	3.3%	1.3%	6.4%	3.1%	2.2%
Read other books	3.1%	0.8%	6.4%	2.5%	2.6%
# Books	4.7%	2.5%	8.2%	4.1%	3.8%
Distance	0.8%	2.2%	1.2%	0.9%	0.9%
Computer use	3.4%	6.5%	5.9%	3.2%	4.1%
Library use	2.8%	2.9%	5.6%	2.7%	3.3%

These are results for the 7,344 students  
that match these conditions:

- State = Karnataka

The factors that influence the marks the  
most (and their impact in marks) are:

1	# Books	+6.8%
2	Computer use	+6.7%
3	Age	+6.7%
4	Help in household	+6.6%
5	Mother occupation	+6.1%

The factors that influence the marks the  
least (and their impact in marks) are:

1	Distance	+0.2%
2	Gender	+0.8%
3	Below poverty	+1.0%
4	Private tuition	+1.4%
5	Use calculator	+2.2%

The top influencers by subject are:

- **Maths %:** Computer use, Age, Library use, Help in household, ...
- **Reading %:** Father occupation, Mother edu, Father edu, # Books, ...
- **Science %:** Mother occupation, Mother edu, Age, Computer use, ...
- **Social %:** Mother occupation, Father occupation, Help in household, Age, ...

Factor	Total %	Maths %	Reading %
Gender	0.8%	0.1%	2.7%
Age	6.7%	8.0%	8.3%
Siblings	2.4%	3.3%	4.6%
Father edu	5.0%	3.6%	12.2%
Mother edu	4.8%	3.1%	12.4%
Father occupation	5.6%	4.9%	12.6%
Mother occupation	6.1%	3.9%	7.8%
Below poverty	1.0%	0.4%	2.9%
Use calculator	2.2%	1.3%	2.9%
Use dictionary	4.9%	3.4%	7.3%
Read other books	3.6%	3.1%	5.9%
# Books	6.8%	5.6%	12.0%
Distance	0.2%	1.5%	1.8%
Computer use	6.7%	9.3%	5.9%
Library use	4.4%	6.3%	4.1%
Private tuition	1.4%	1.3%	2.7%
Watch TV	2.5%	2.6%	5.2%
Read magazine	4.3%	3.9%	5.7%
Read a book	5.1%	4.8%	7.7%
Play games	5.0%	5.9%	5.2%
Help in household	6.6%	6.1%	9.3%

# What else can we answer?

---

- Economic health/ literacy of various areas in the state
  - Based on average income per household/ based on education or occupation of people
- Impact of technology on grades in various courses
  - Based on whether a majority of students know how to use the calculator, computer, etc., hours spent viewing television
- Typical occupations of people in various areas of a state
  - Based on mother's occupation and father's occupation
- Are families mostly nuclear or joint?
  - Based on number of people in the household, help available in the house, number of siblings, etc.
- Health of public transportation/ reachability of various areas in a state
  - Based on distances travelled (and mode of transport) to get to various schools in the area
- What impact (if any) do Government schemes have on performance of students?
  - Based on correlation between midday meals and performance in a certain school
- A study of gender bias in various places: For eg) Do boys help with household chores or help take care of younger siblings?
  - Filtered based on the number of pupils who help with household chores and gender

# Who owns this data?

- Seeking permission
- Privacy, confidentiality, security



**THANK YOU**

---

**Dr. Mamatha H R**

Professor, Department of Computer Science

**mamathahr@pes.edu**

+91 80 2672 1983 Extn 834



## DATA ANALYTICS

### Unit 1: Data Visualization

---

**Mamatha.H.R**

Department of Computer Science and  
Engineering

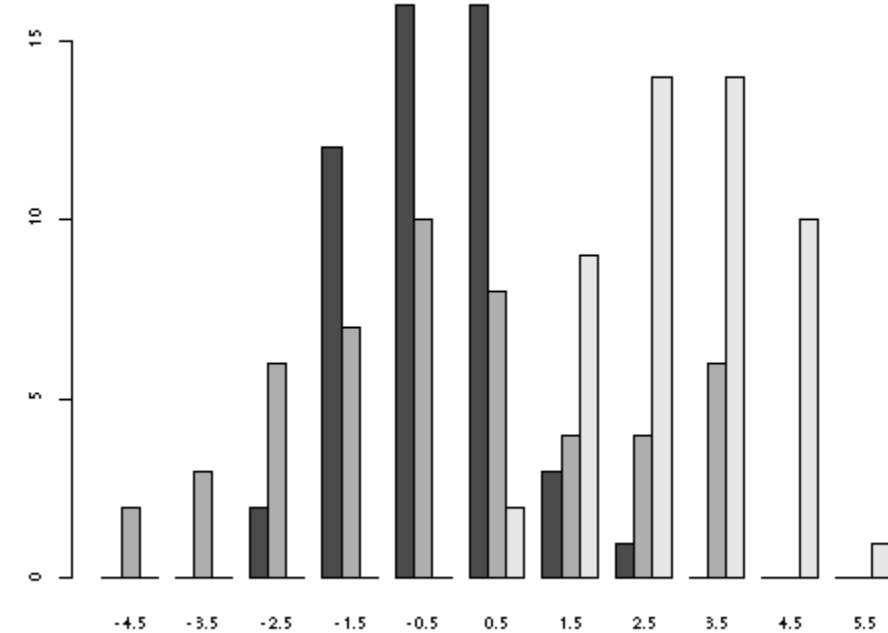
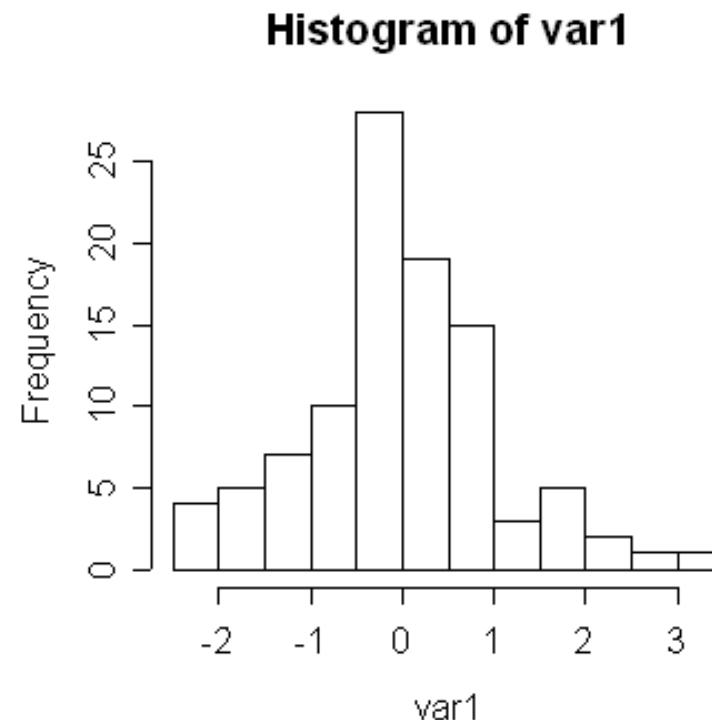
# DATA ANALYTICS

---

## Unit 1: Data Visualization contd.

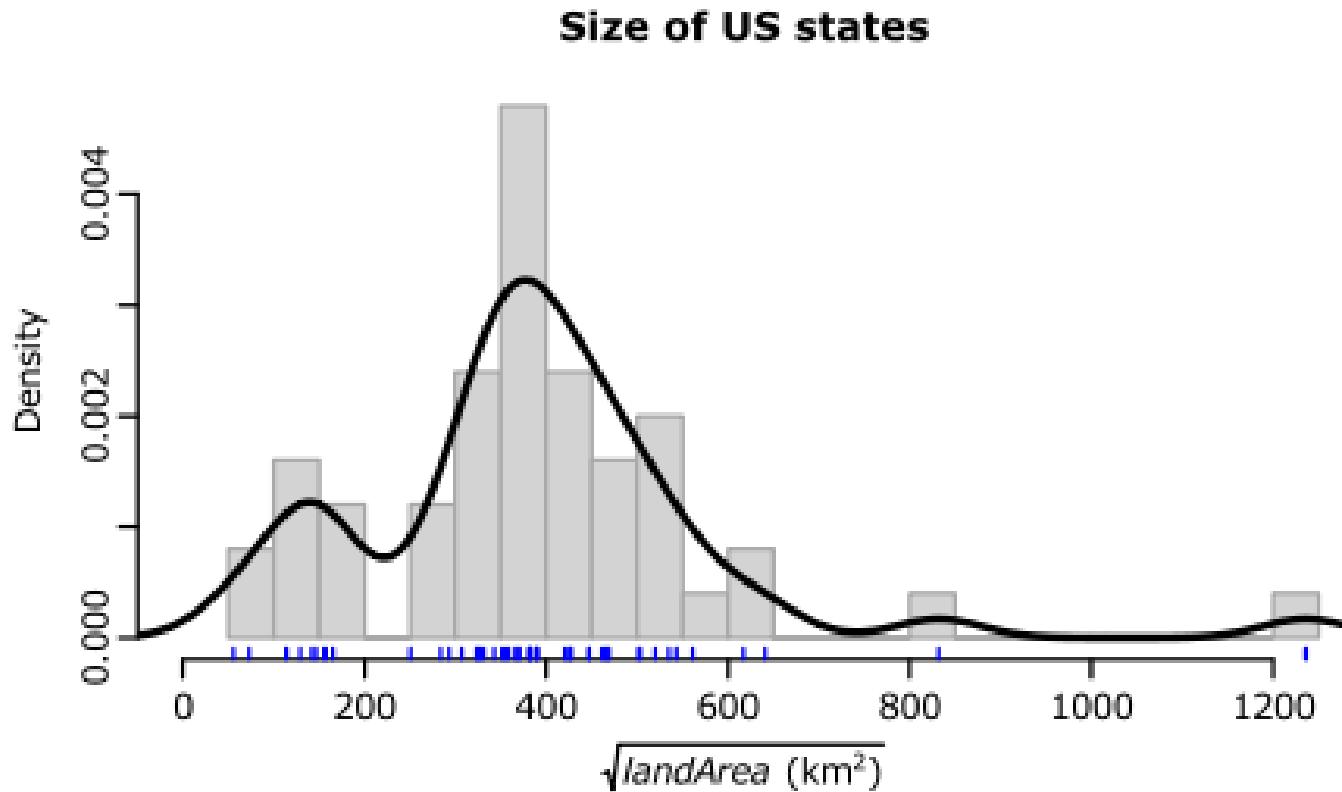
**Mamatha H R, Gowri Srinivasa**  
Department of Computer Science and Engineering

## hist and multihist

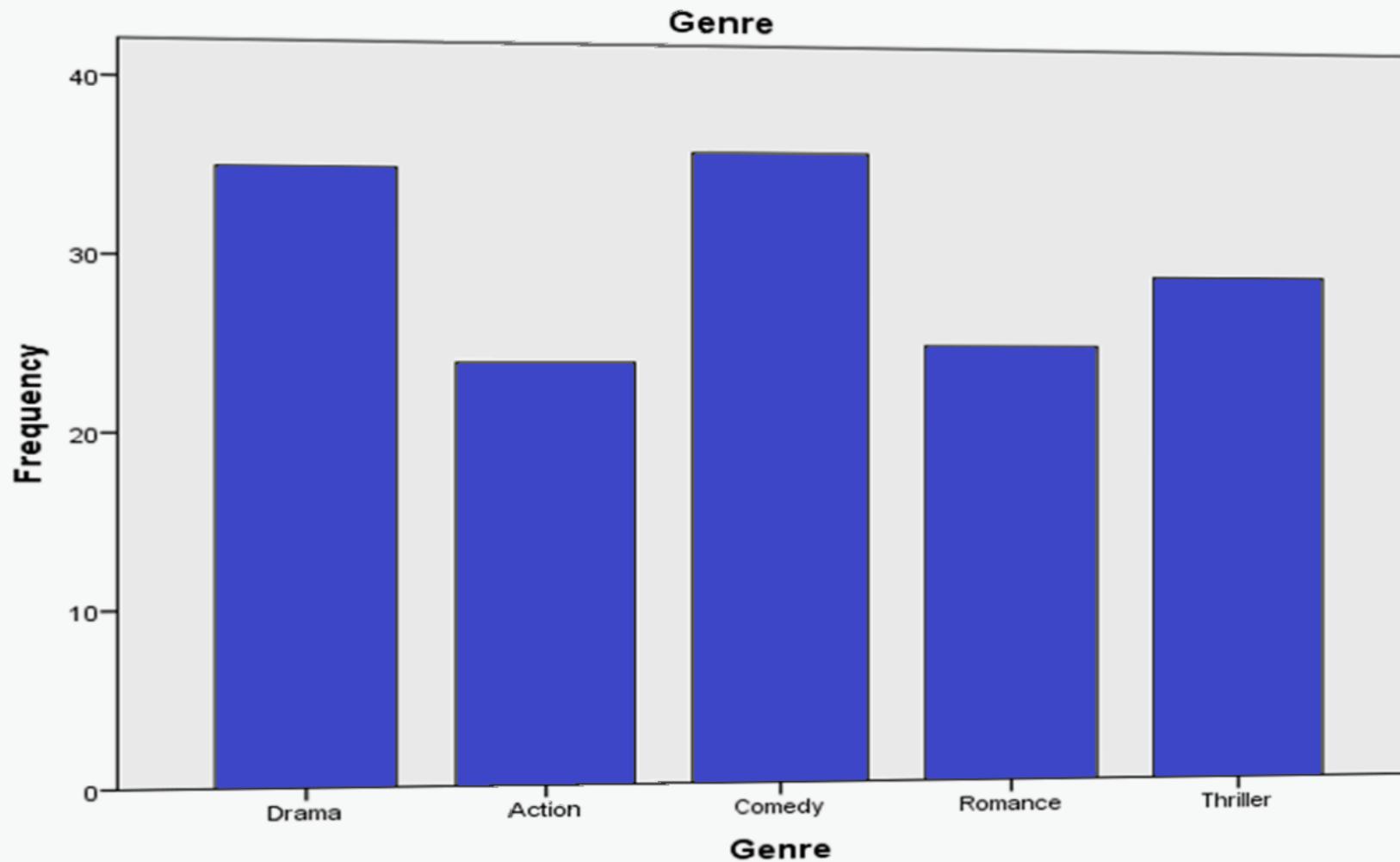


Requires package plotrix

# hist and density

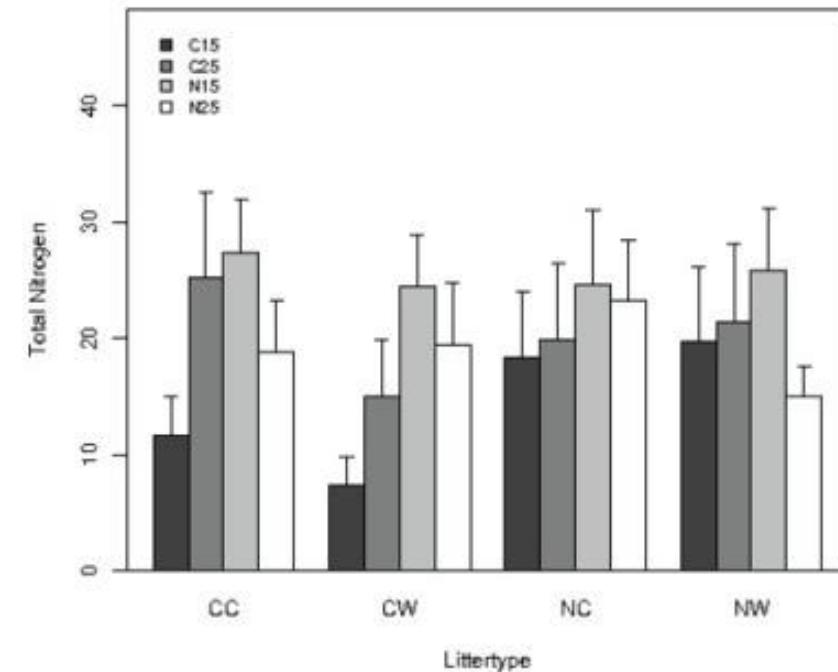
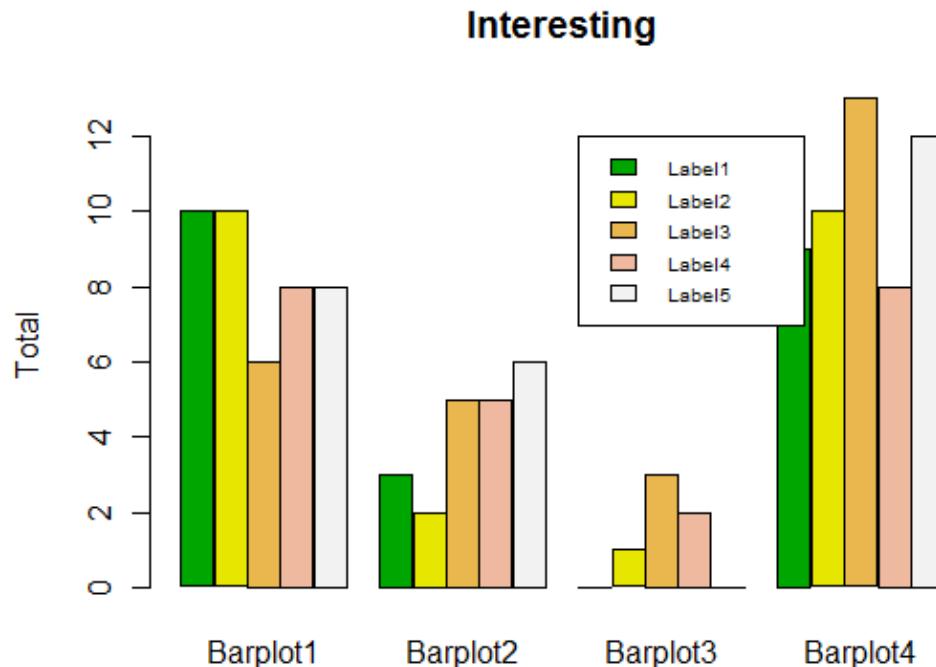


## Bar chart for movie genre



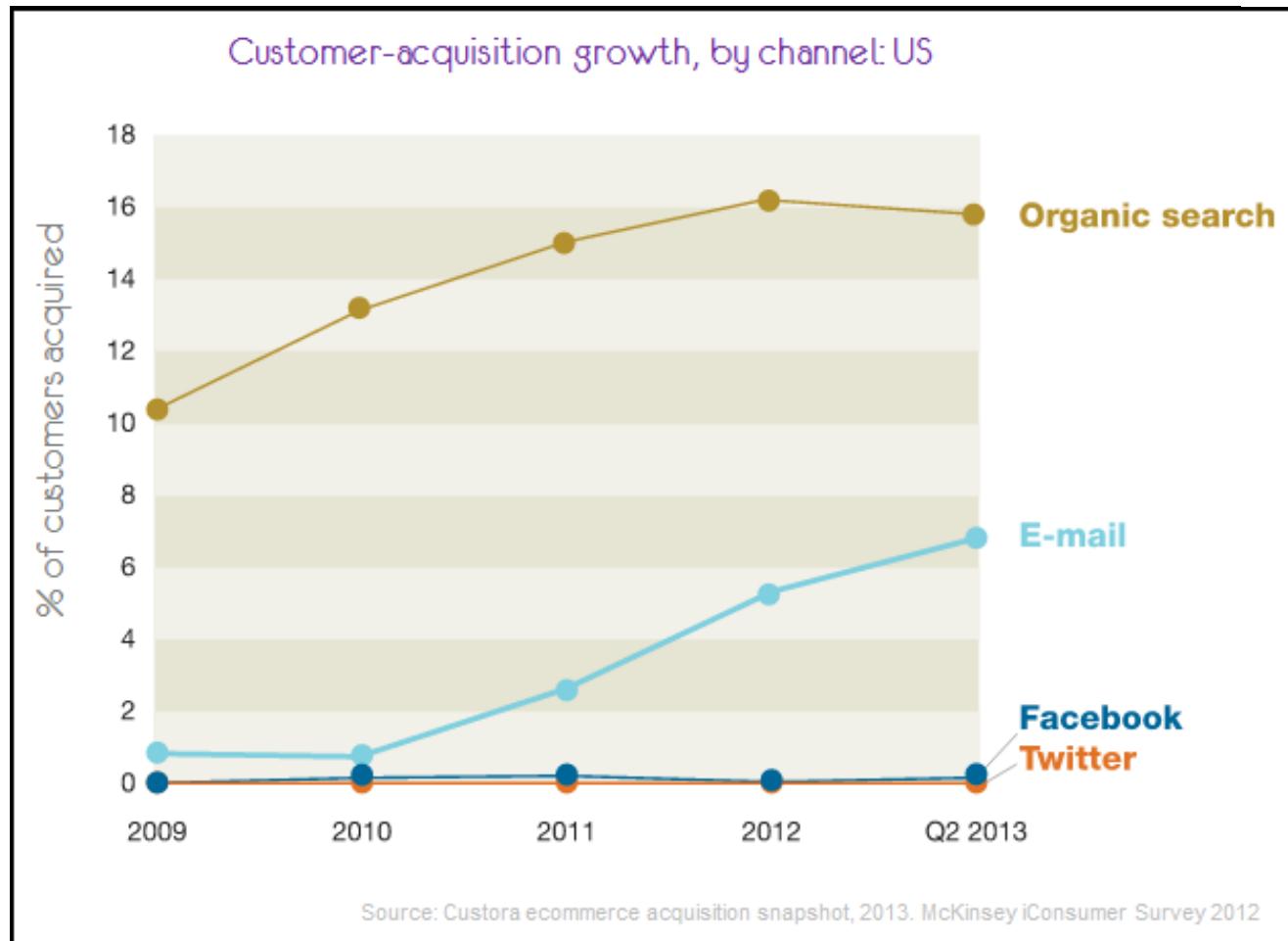
# barplot

---



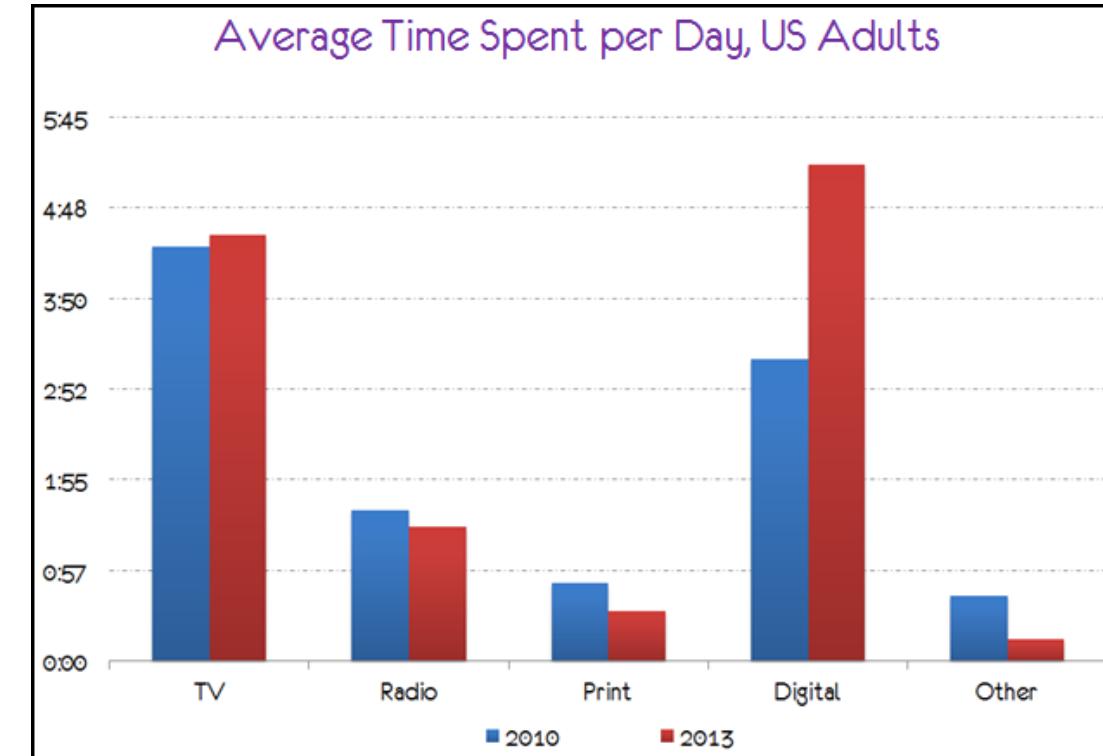
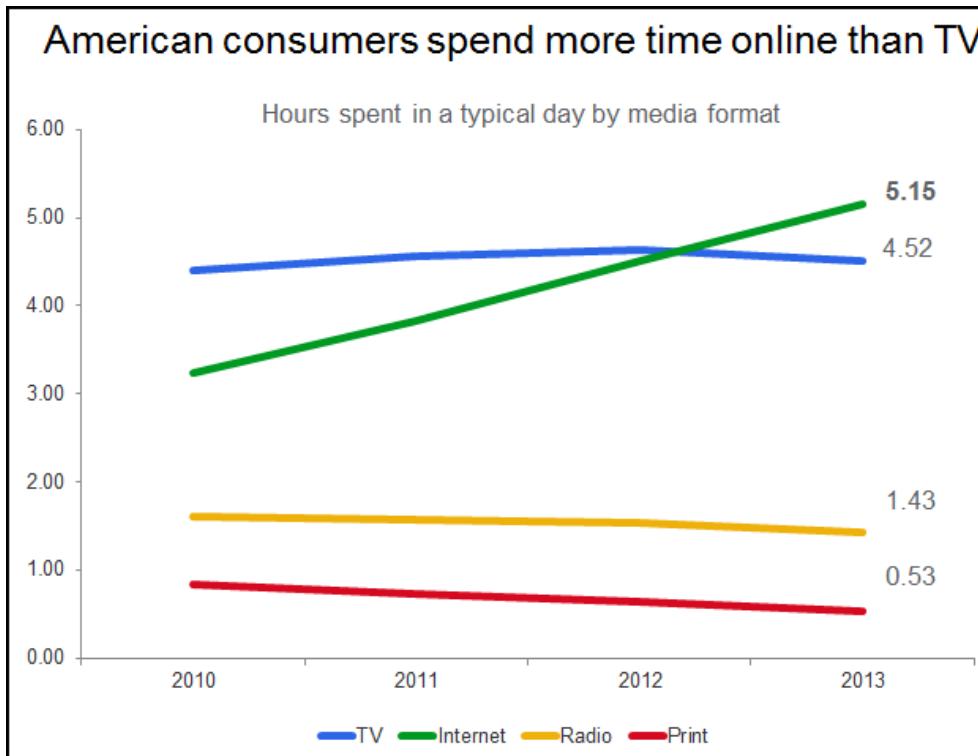
Box plots with error bars  
Uses: arrows and anova

# Labels and legends

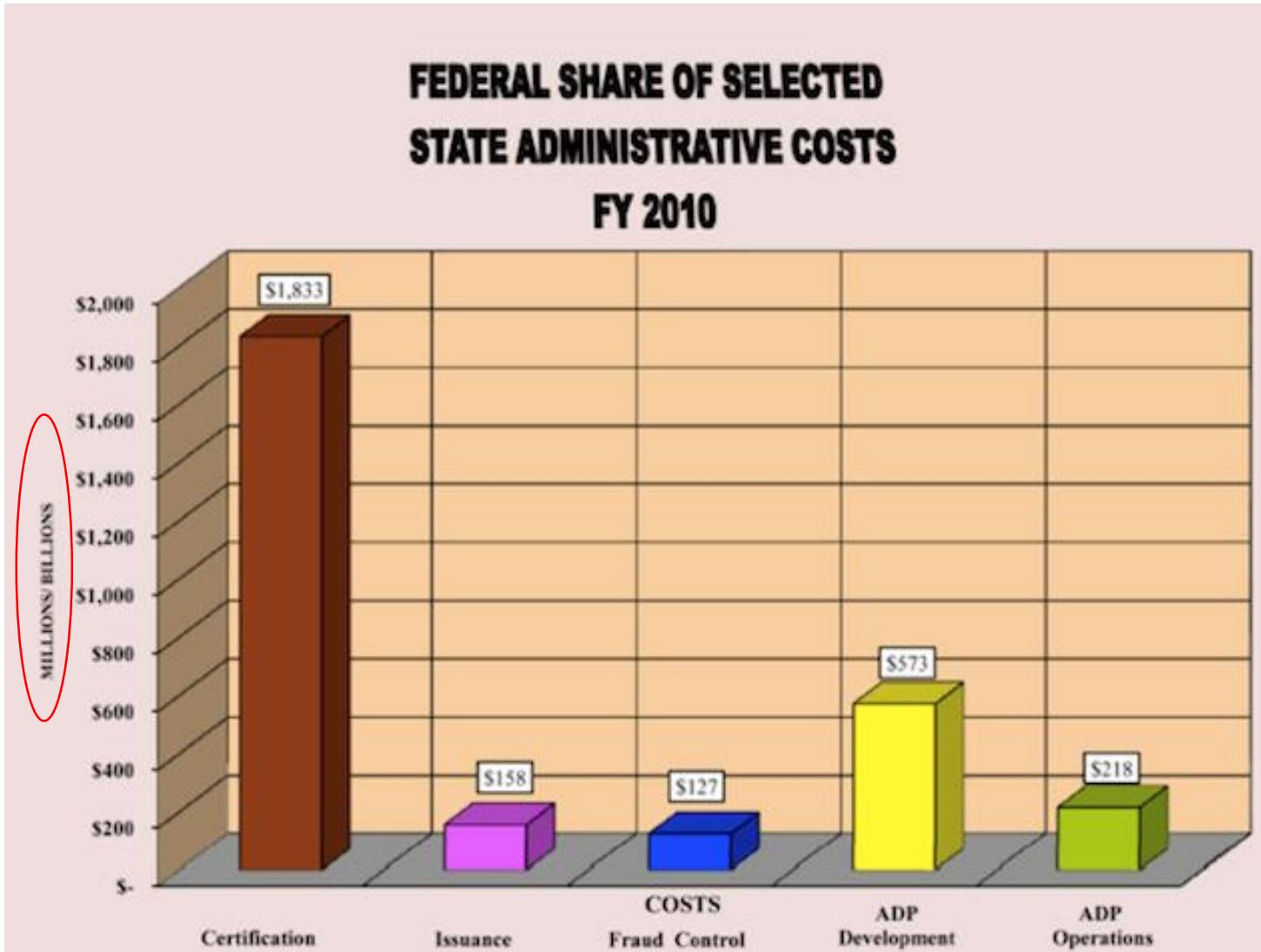


Avg time spent/ day on  
various entertainment options

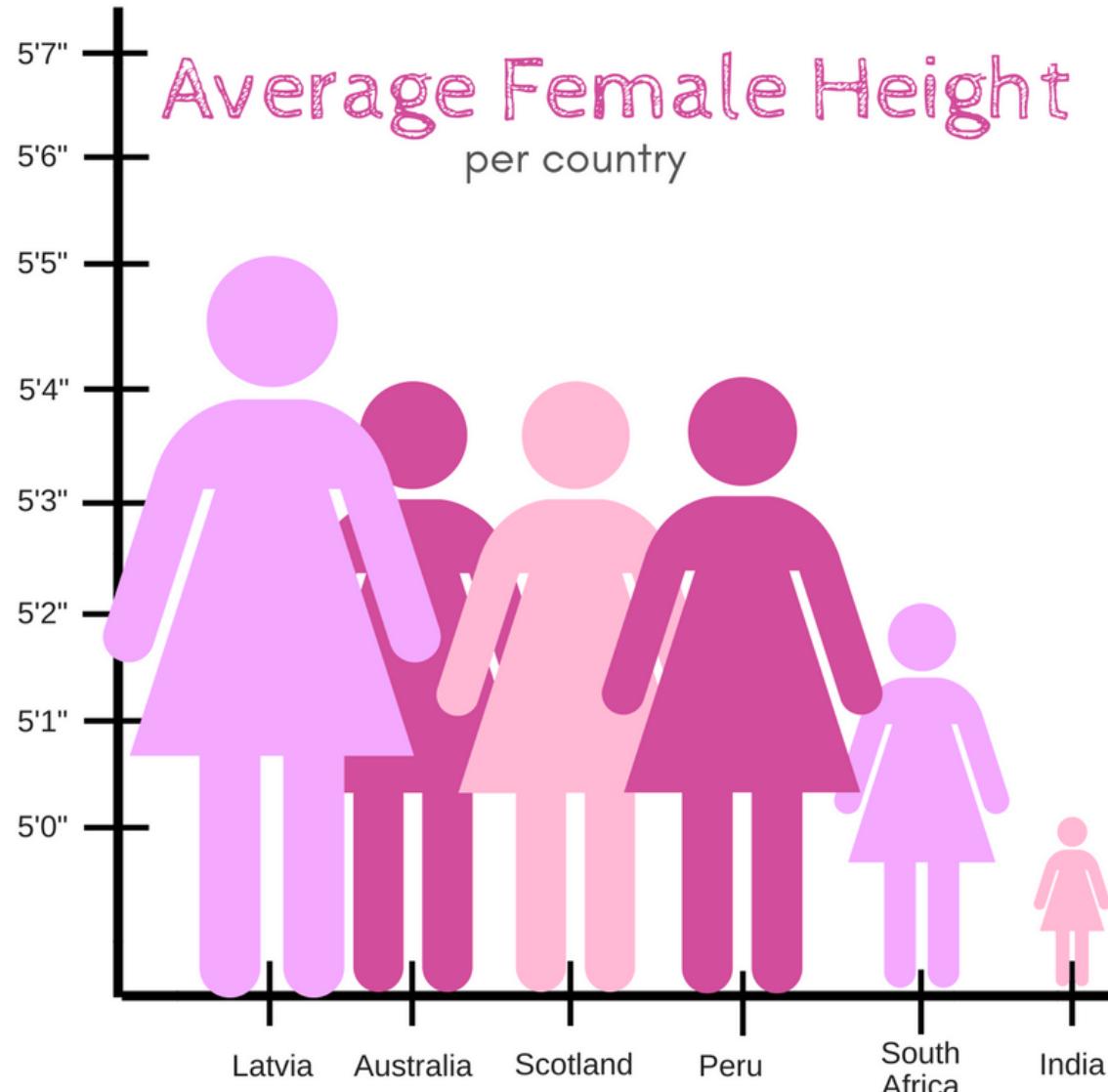
# Horizontal vs Vertical



# Use colors... but carefully!



# What can we say about this graph?

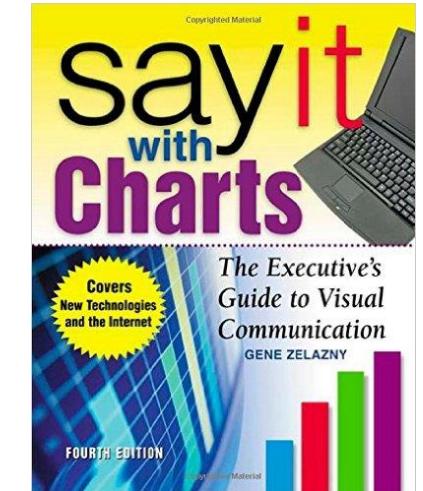


## Some dos and don'ts

- Use the full axis
- **Avoid distortion**
- Sort the data for ease of comparison
- Use consistent intervals on any axis or indicate a break
- Use the chart type wisely
- Don't use colors and effects without reason

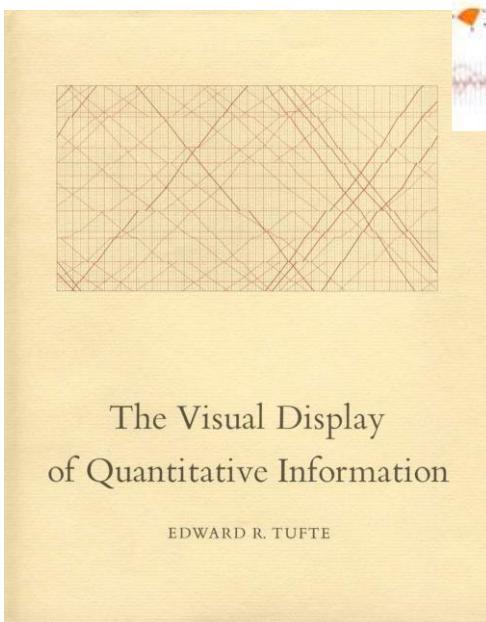
<https://www.eea.europa.eu/data-and-maps/daviz/learn-more/chart-dos-and-donts>

# Do's and do not's

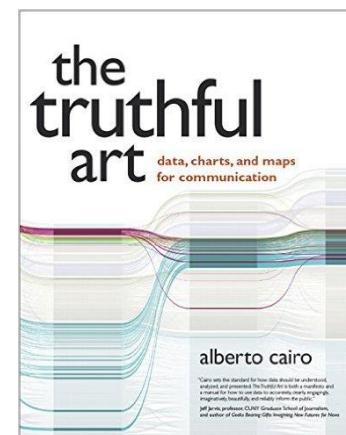
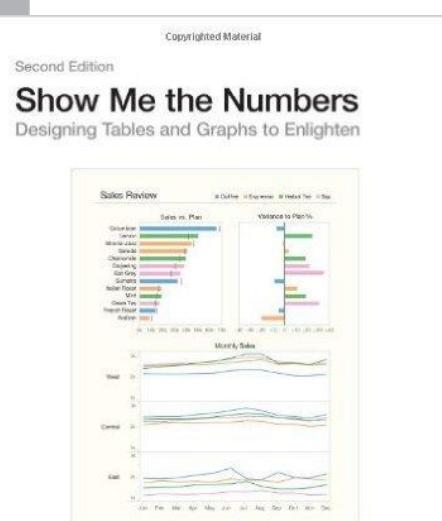
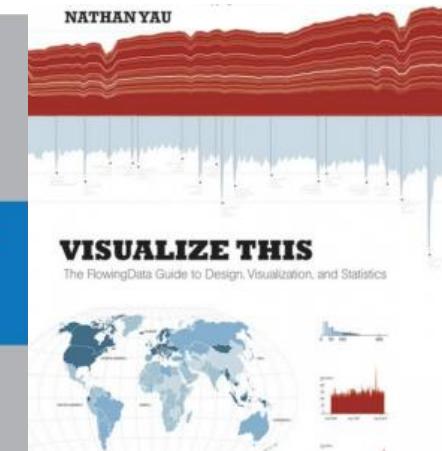
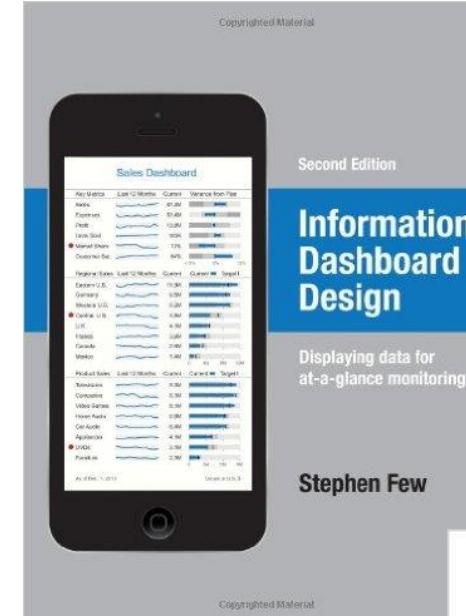


Edward R. Tufte

## Envisioning Information



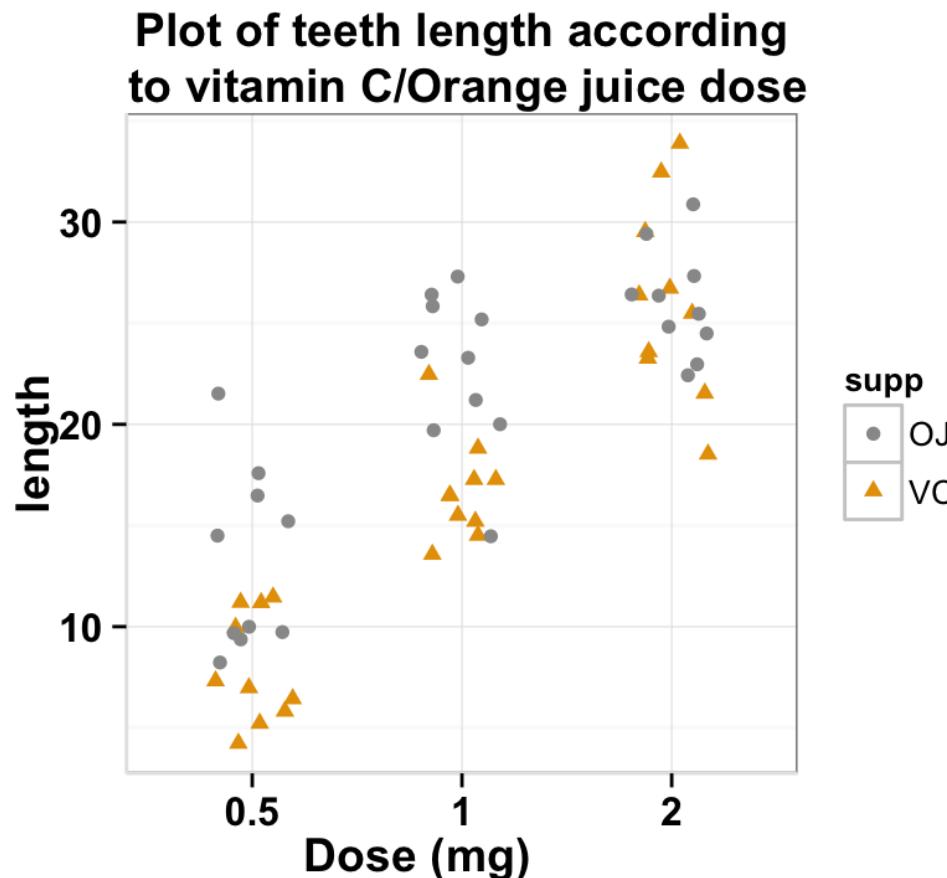
THE WALL STREET JOURNAL  
**GUIDE TO INFORMATION GRAPHICS**  
THE DOS & DON'TS OF PRESENTING DATA, FACTS, AND FIGURES  
DONA M. WONG



# Length of teeth

- Different doses of orange juice versus vit C

# stripchart



# wordcloud



# Packages in R

---

- classInt: univariate class intervals
- ggplot2: graphical features
- gpclib: polygon clipping
- hexbin: bivariate data manipulation
- latticist: Interface between R and Latticist
- mapdata: has data that can be added to maps
- maps: maps of various geographical areas
- maptools: access mechanisms to use maps

# Packages in R

---

- Lattice
  - xyplot (relationship between two attributes)
    - Display data points as text
  - histogram (visualize by month)
  - barchart (stacked barchart of data)
  - dotplot ()
  - bwplot ()
  - cloud ()
  - parallel ()

# Other packages in R

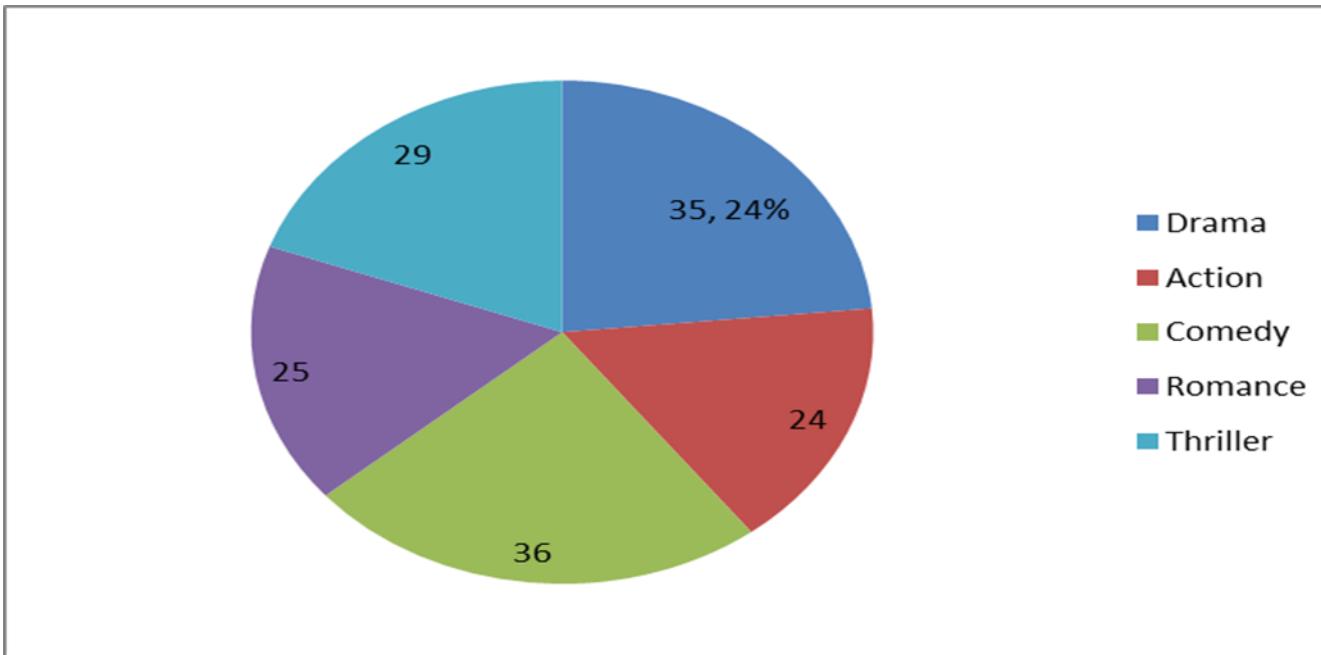
---

- car: companion to applied regression
- gclus: scatter plots
  - cpairs (very similar to the pairs function)
- hexbin: density scatter plots
  - hexagonal bins
- MASS: support functions and data sets for ‘Modern Applied Statistics with S’ by Venables and Ripley

## Pie Chart

- **Pie chart** is mainly used for categorical data and is a circular chart that displays the proportion of each category in the dataset

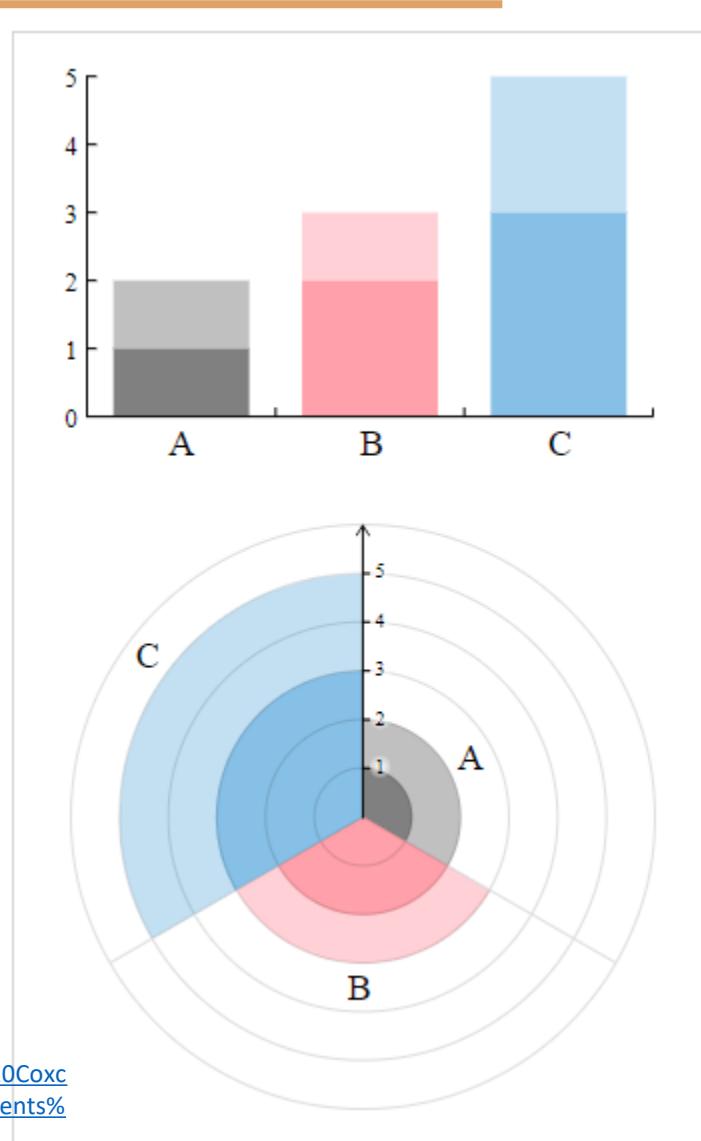
Pie chart for movie genre



## Coxcomb Chart

- Coxcomb chart (also known as polar area chart or roses) is an extension of pie chart made popular by Florence Nightingale (Lewi, 2006)
- Nightingale rose chart or Polar area diagram
- In a Coxcomb chart, each area represents a magnitude of the category
- In a pie chart the radius of each sector is same, whereas, in a Coxcomb chart the radius of the sector is adjusted to create the magnitude of the area

[https://datavizcatalogue.com/methods/nightingale\\_rose\\_chart.html#:~:text=Also%20known%20as%20a%20Coxcomb,soldiers%20during%20the%20Crimean%20war.&text=Each%20category%20or%20interval%20in,segments%20on%20this%20radial%20chart.](https://datavizcatalogue.com/methods/nightingale_rose_chart.html#:~:text=Also%20known%20as%20a%20Coxcomb,soldiers%20during%20the%20Crimean%20war.&text=Each%20category%20or%20interval%20in,segments%20on%20this%20radial%20chart.)



# DATA ANALYTICS

## Coxcomb chart on causes of mortality in the army prepared by Florence Nightingale



PES  
UNIVERSITY  
ONLINE

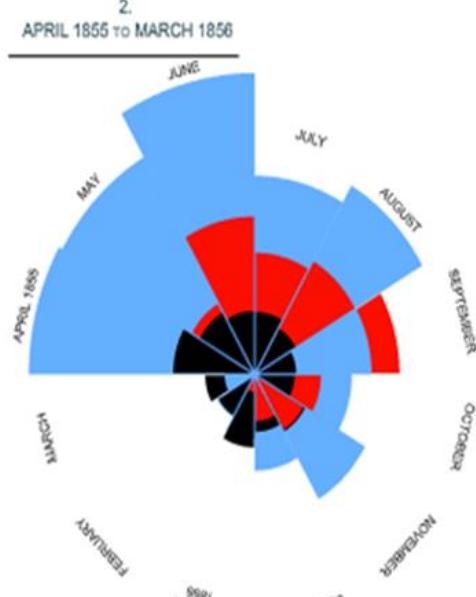
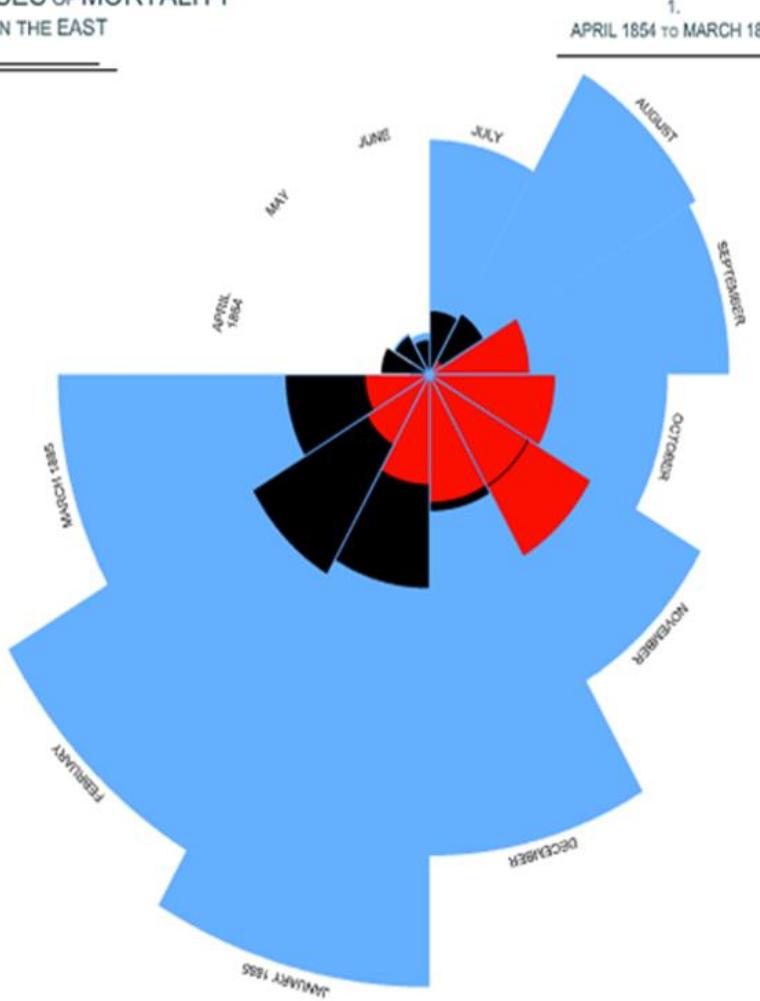


DIAGRAM OF THE CAUSES OF MORTALITY  
IN THE ARMY IN THE EAST



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes

The black line across the red triangle in Nov' 1854 marks the boundary of the deaths from all other causes during the month

In October 1854, & April 1855, the black area coincides with the red, in January & February 1856 the blue coincides with the black

The entire areas may be compared by following the blue, the red & the black enclosing lines.

## Scatter Plot

---

- Scatter plot is a plot of two variables that will assist data scientists to understand if there is any relationship between two variables
- The relationship could be linear or non-linear
- scatter plot is also useful for assessing the strength of the relationship and to find if there are any outliers in the data

## Scatter Plot

---

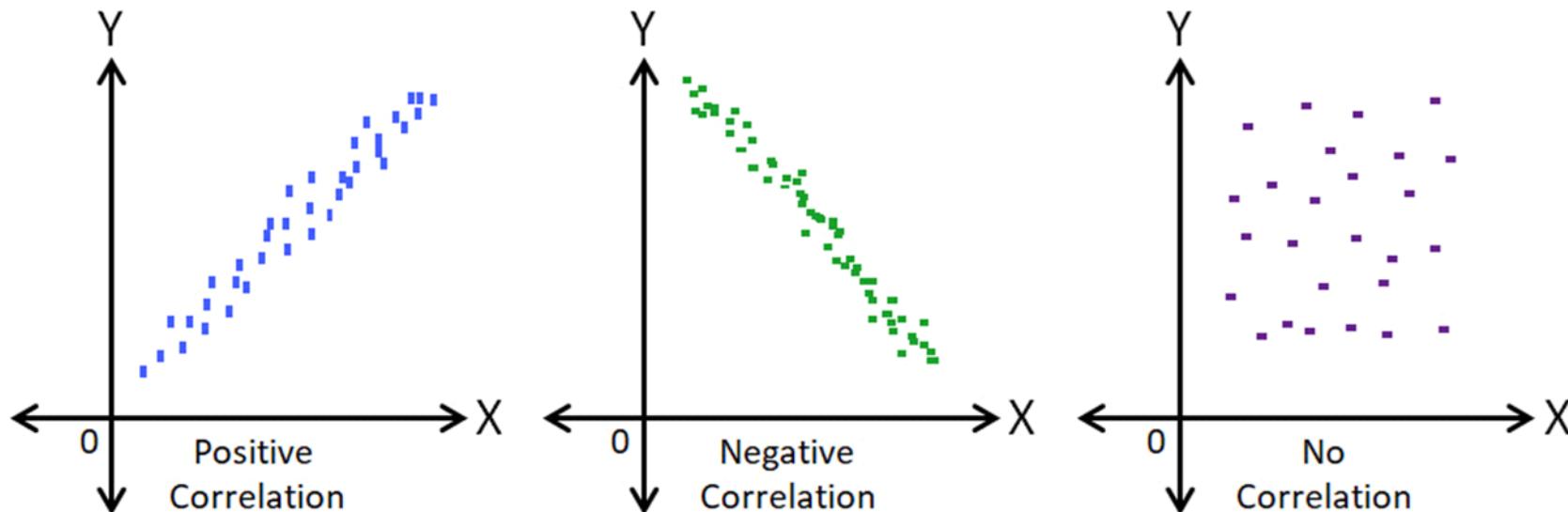
- There are many types of coefficients of correlation in scatter points, most popular one is

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

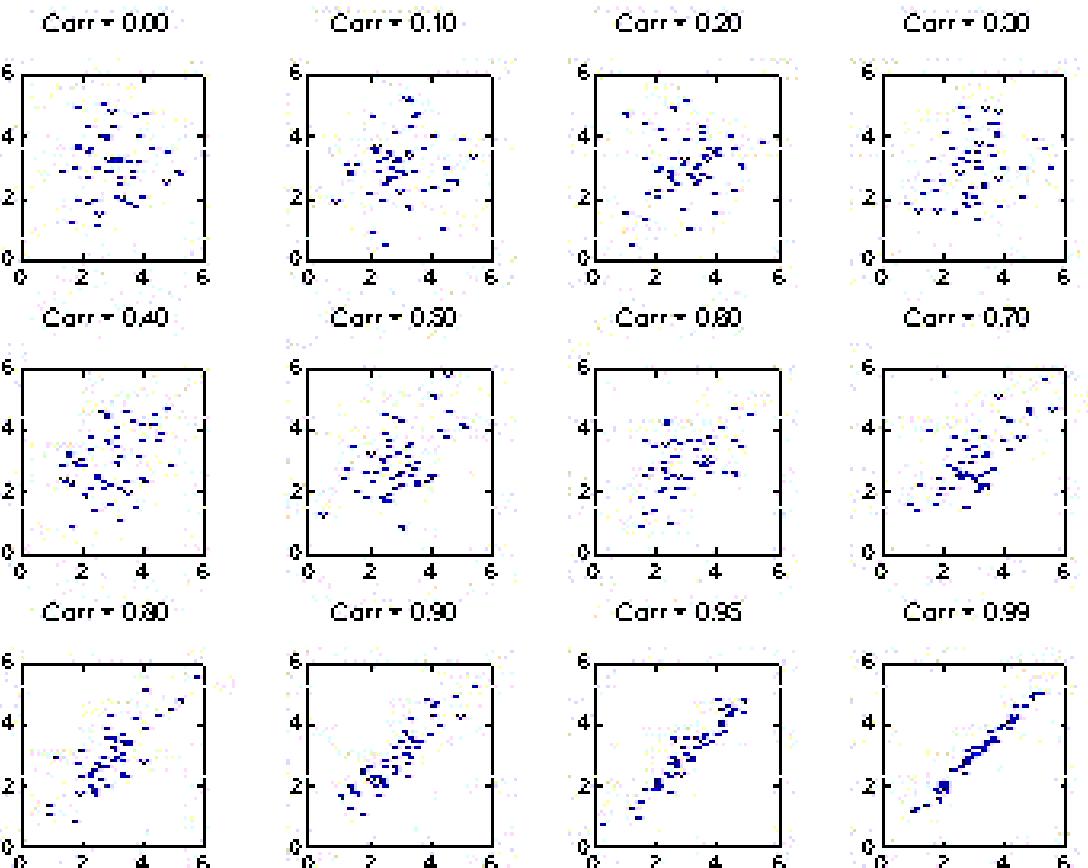
- Pearson's Co-efficient of correlation
- x – value of data point on x-axis
- y – value of data point on y-axis
- n – no of datapoints

## Scatter Plot

### Scatter Plots & Correlation Examples



# Notion of correlation



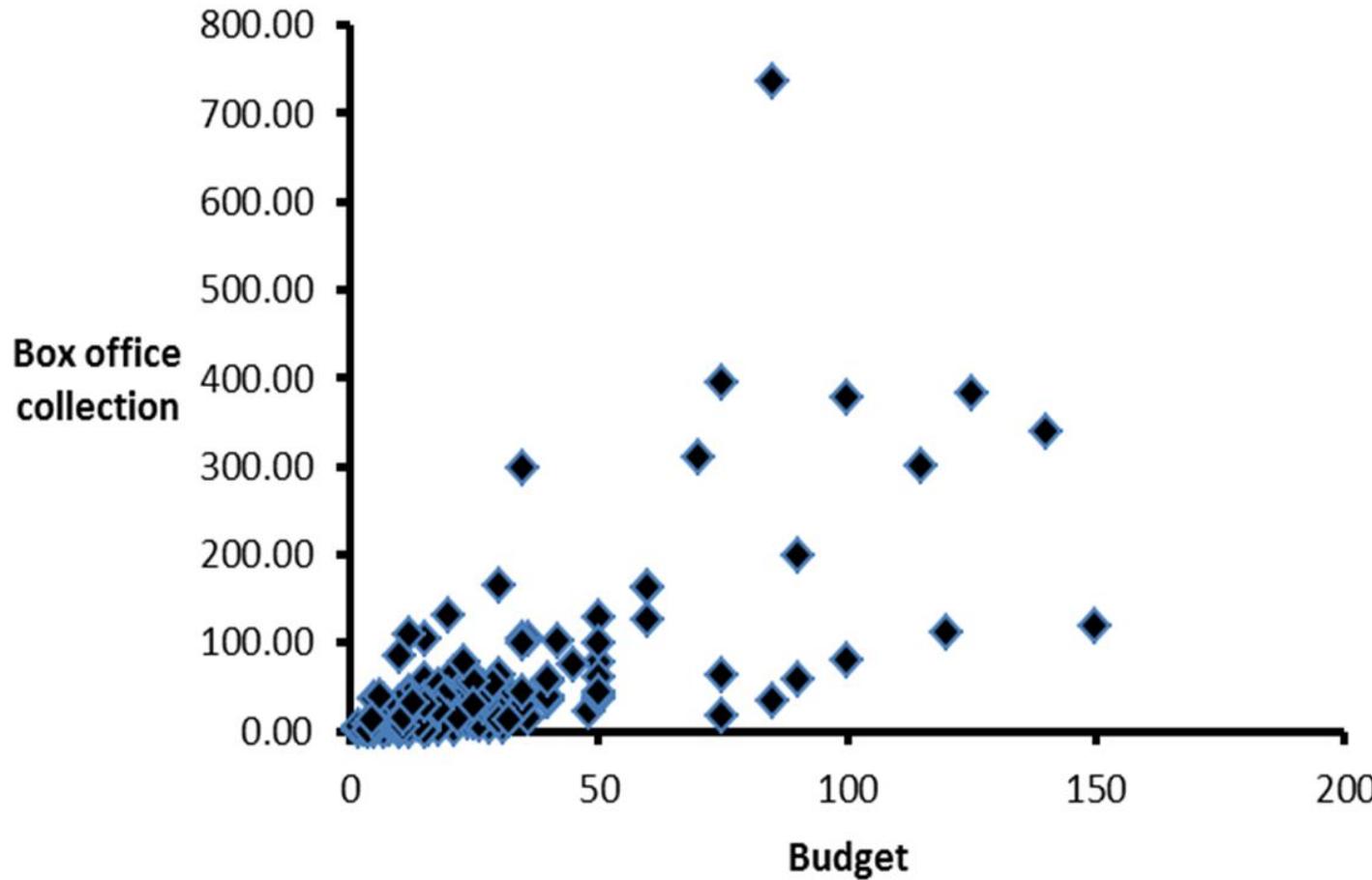
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

## Scatter Plot

---

- Pearson's co-efficient of correlation:
  - $\text{co-eff} > 0$  : positively correlated
  - $\text{co-eff} < 0$  : negatively correlated
  - $\text{co-eff} = 0$  : no correlation
- +1 or -1, mean perfect correlation between the data points

## Scatter plot between movie budget and box office collection

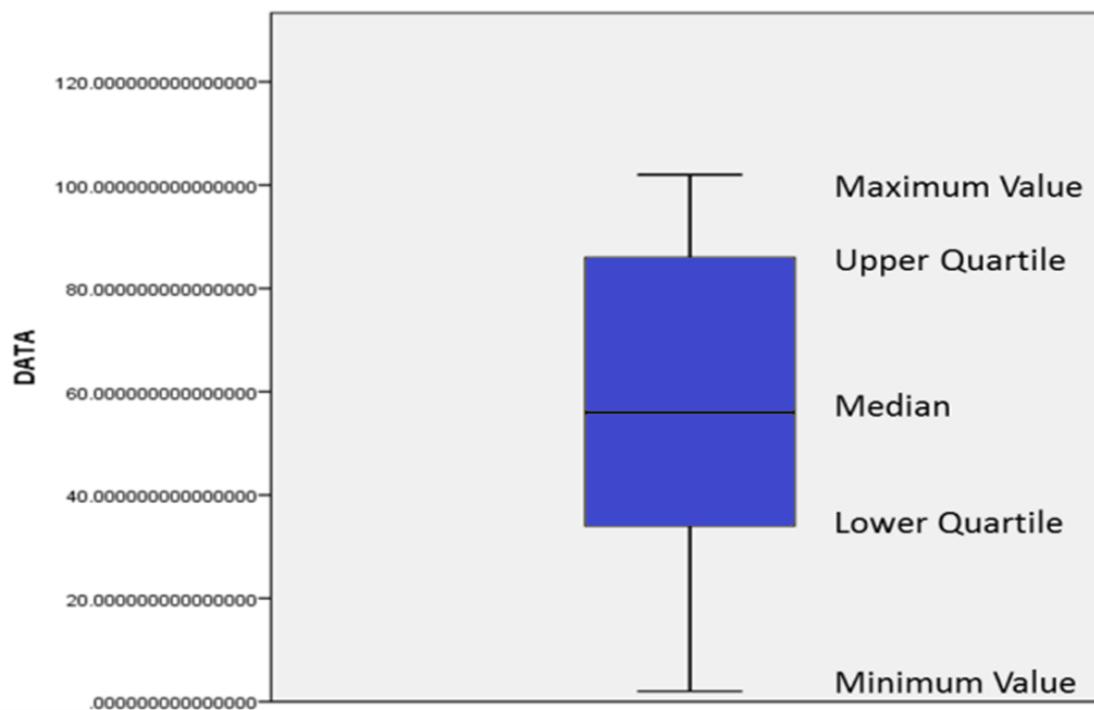


## Box Plot (or Box and Whisker Plot)

---

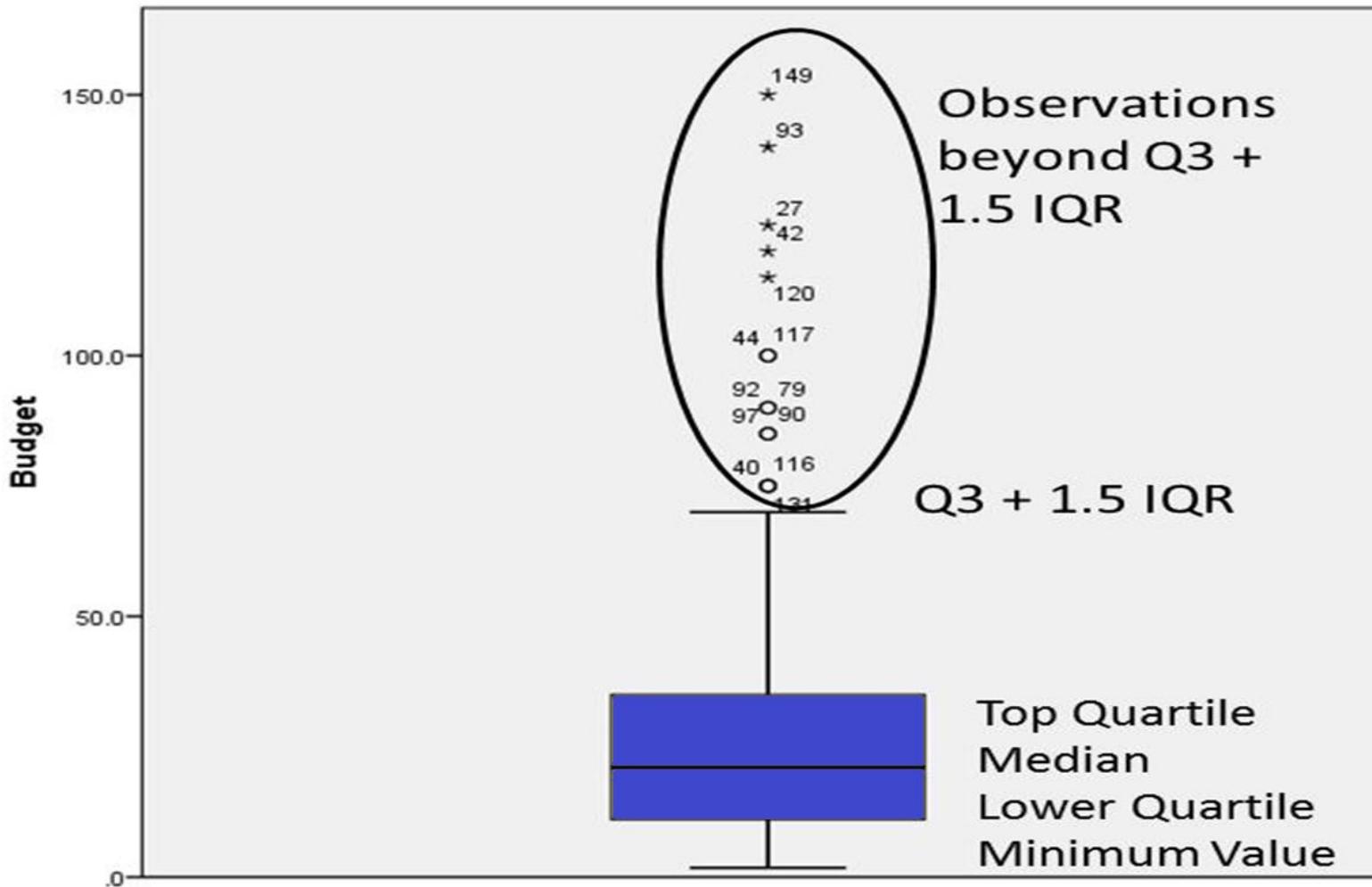
- Box plot (aka Box and Whisker plot) is a graphical representation of numerical data that can be used to understand the variability of the data and the existence of outliers
- Box plot is designed by identifying the following descriptive statistics:
  - Lower quartile (1<sup>st</sup> Quartile), median and upper quartile (3<sup>rd</sup> Quartile).
  - Lowest and highest value
  - Inter-quartile range (IQR).

- The box plot is constructed using IQR, minimum and maximum values



## Bollywood movie Budget Boxplot

- The box plot for the Bollywood movie budget

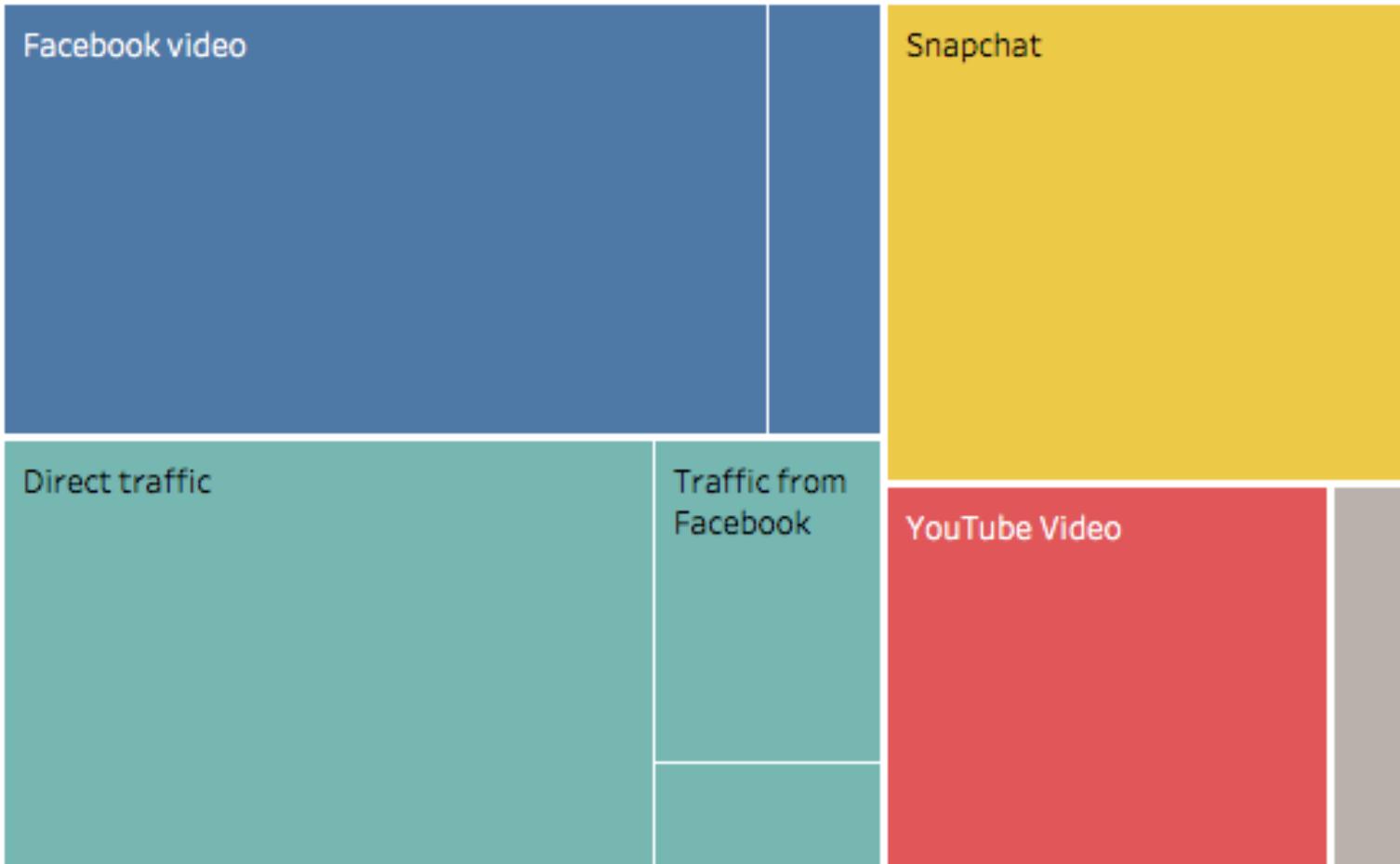


## Treemap

---

- **Treemap** is a hierarchical map made up of nested rectangles frequently used as part of business intelligence reports which helps organizations to understand the data hierarchically
- The size of rectangle and colour are used for describing/differentiating the characteristics of the data.

Where people consumed BuzzFeed content in 2015

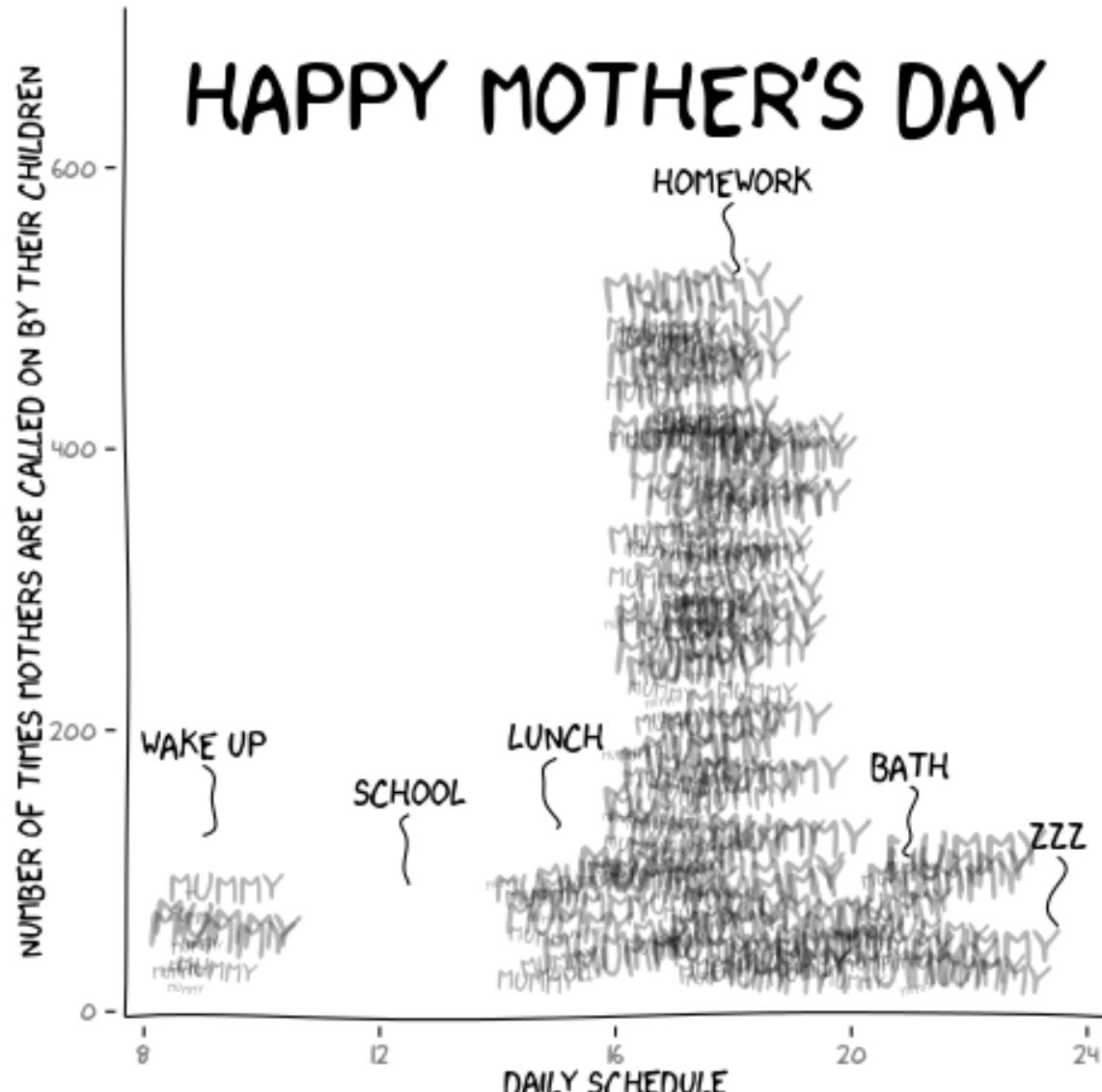


# Some interesting examples

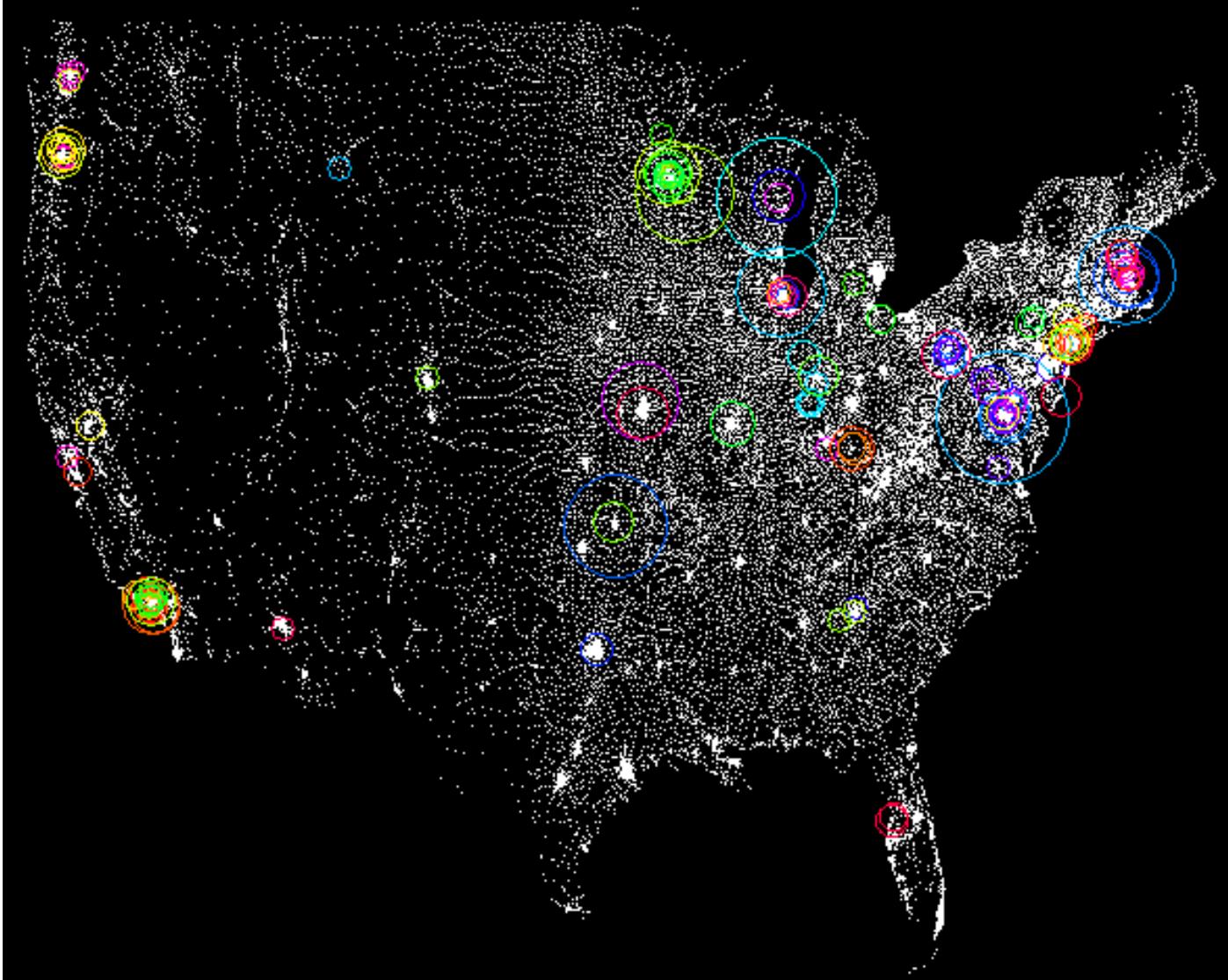
---

- Special packages
- Credit card fraud
- Maps
- Heat maps
- How can we tell a good story using visualization?

# xkcd package



# A picture is worth a thousand words – Credit card fraud: an example

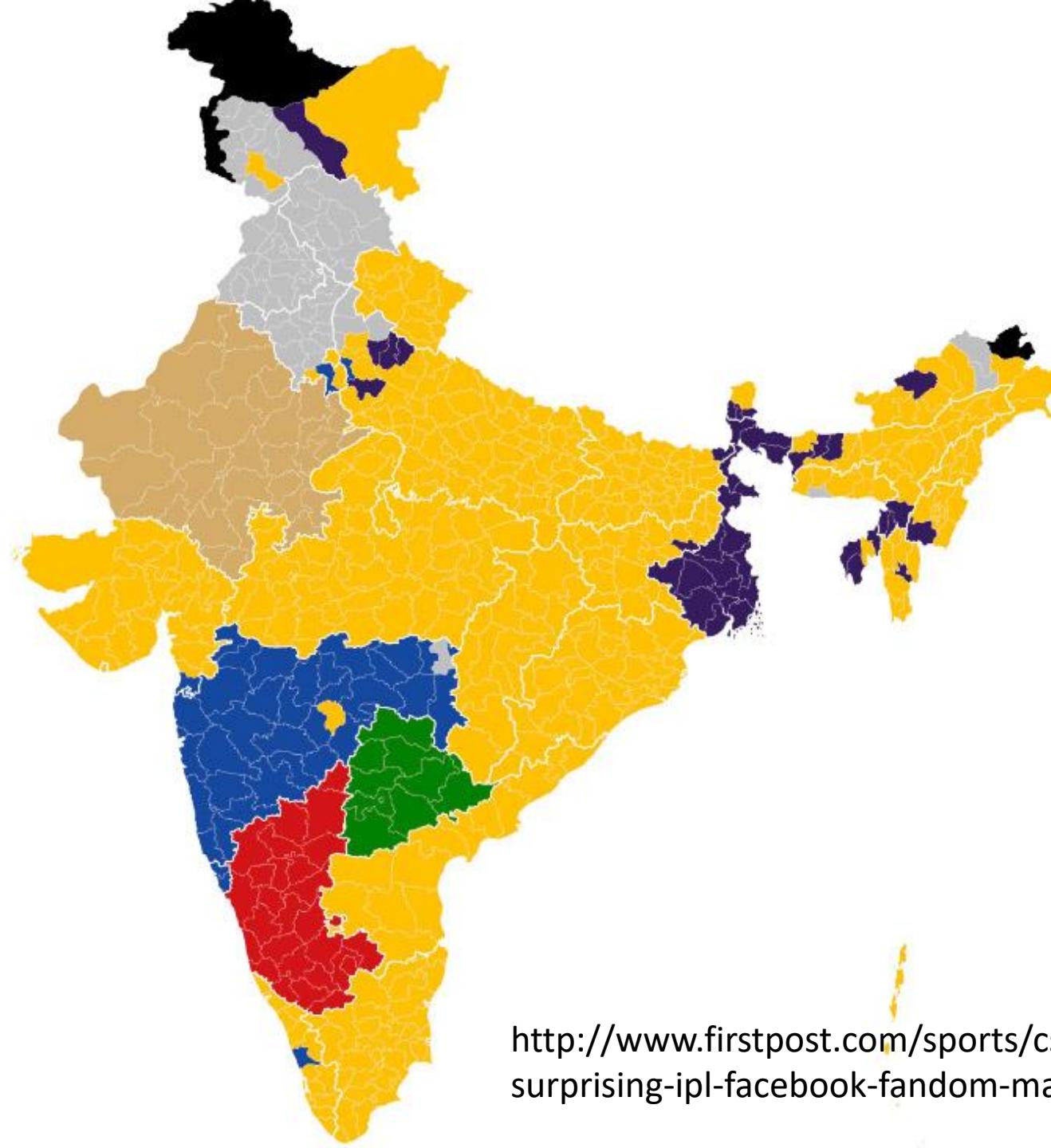


Color:  
Blue: old  
Red: recent

Radius:  
Dollar amounts

# Use of maps





Facebook Fandom Map

## Indian Premier League

This 2015 map displays Facebook fans of all of the IPL teams. Each district is color-coded based on which official Facebook team page has the most likes from people who live there.

- Mumbai Indians
- Chennai Super Kings
- Royal Challengers Bangalore
- Kolkata Knight Riders
- Rajasthan Royals
- Sunrisers Hyderabad
- Kings XI Punjab
- Delhi Daredevils

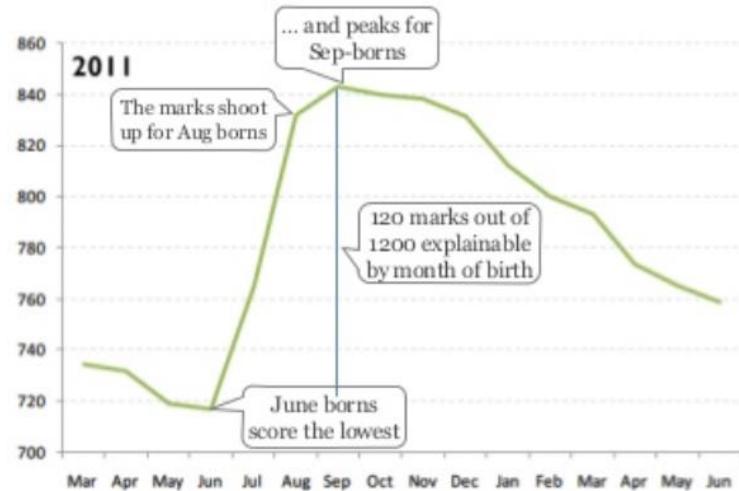
<http://www.firstpost.com/sports/csk-rocks-incredibly-surprising-ipl-facebook-fandom-map-2249912.html>

# Use of heat maps

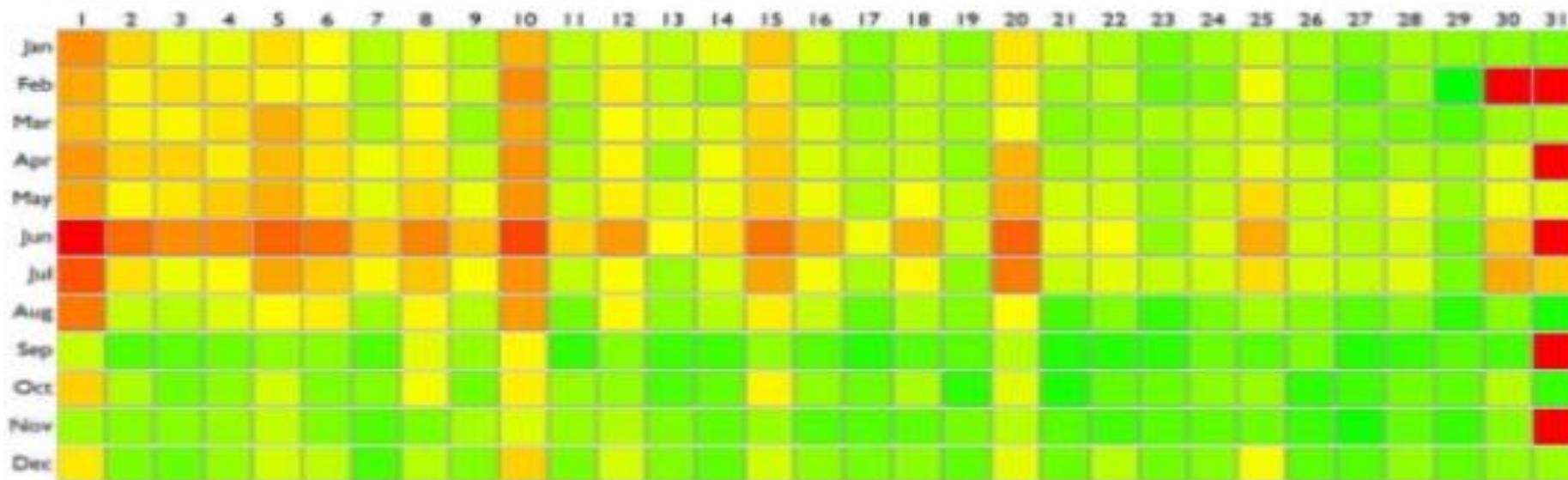


# Results of students in X Standard (TN Board exams)

Based on the results of the 20 lakh students taking the Class XII exams at Tamil Nadu over the last 3 years, it appears that the month you were born in can make a difference of as much as 120 marks out of 1,200.



- Birthdays of students registered for the exam



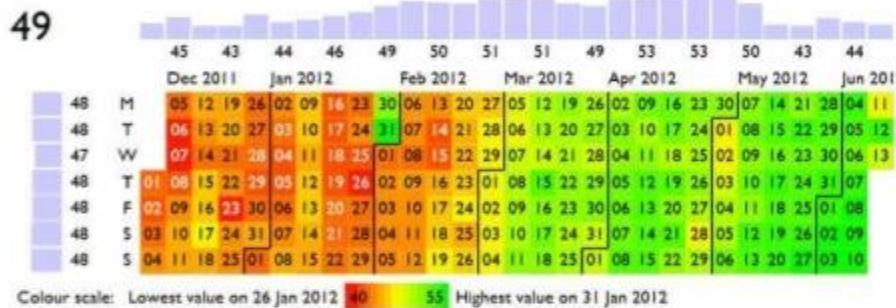
# How can good visualization be used to tell a story?

- Restaurant had collected years of data

[https://www.slideshare.net/gramener/making-big-data-relevant-importance-of-data-visualization-and-analytics?next\\_slideshow=1](https://www.slideshare.net/gramener/making-big-data-relevant-importance-of-data-visualization-and-analytics?next_slideshow=1)

## DAILY SALES CALENDAR MAP

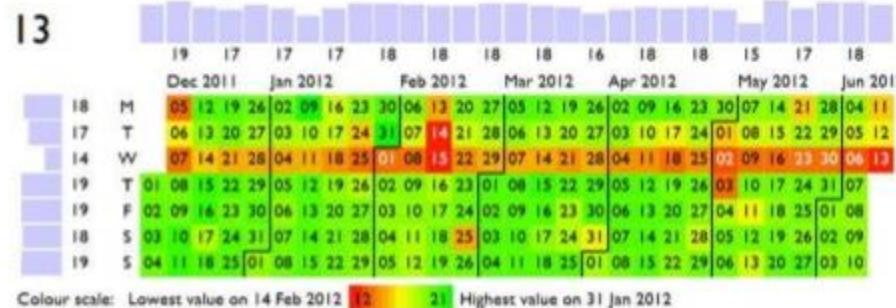
FOR PoS: S1 (IN THOUSANDS OF Rs.)



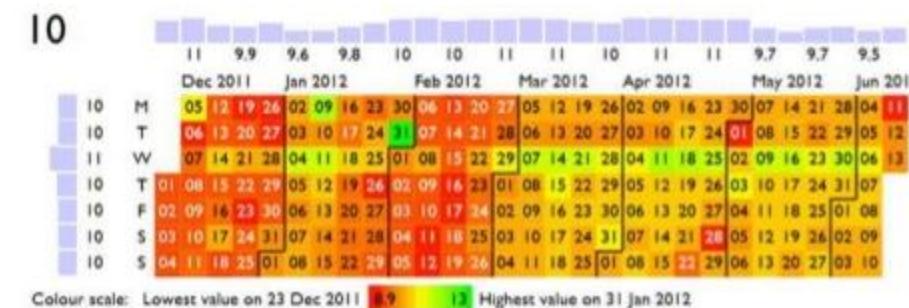
FOR PoS: S2 (IN THOUSANDS OF Rs.)



FOR PoS: R1 (IN THOUSANDS OF Rs.)



FOR PoS: R2 (IN THOUSANDS OF Rs.)



## Exercises

---

What are the ideal use cases that warrant the use of a Treemap chart and Coxcomb chart?

## Summary

---

- Descriptive analytics is beginning of any analytics project that uses data summarization, descriptive statistics, visualization and queries to gain insights about what happened in the past
- Measures of central tendency, measures of variation and measures of shape assist data scientists to understand the data for characteristics such as variability and skewness.
- Descriptive analytics can help data scientists with further analysis of the data by identifying relationships that may exist in the data

## Summary

---

- Data visualization is an integral part of descriptive analytics and plays a major role in business intelligence (BI) by displaying data using innovative graphs and dashboards for easy comprehension of data to top management.
- Descriptive analytics will provide hints for developing predictive analytics models.

## References

---

### Text Book:

- [“Business Analytics, The Science of Data-Driven Decision Making”](#), U. Dinesh Kumar, Wiley 2017
- [Data Mining: Concepts and Techniques](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3<sup>rd</sup> Edition.
- [Introduction to Data Mining](#), Tan, Steinbach, Kumar, 2<sup>nd</sup> Edition

## A brief recap

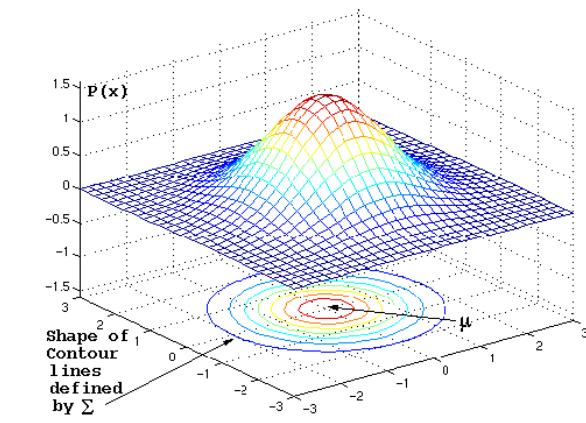
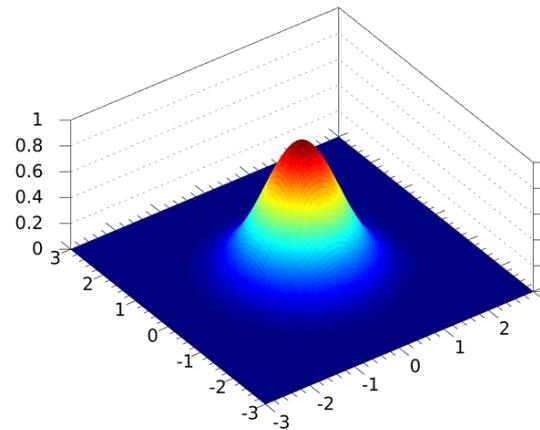
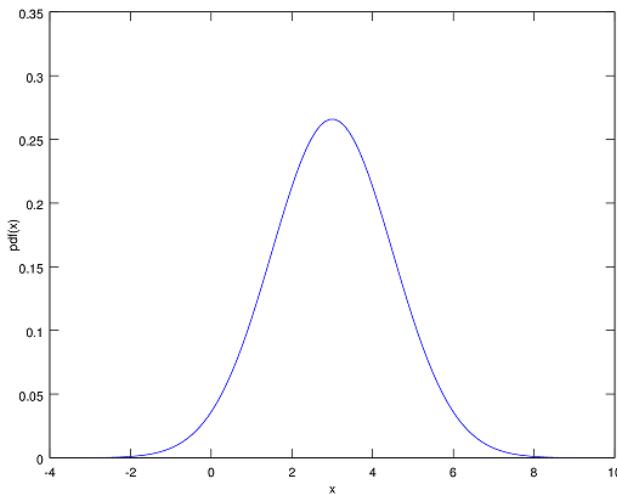
---

Should we memorize all this for ISA/ ESA?

Hopefully, you don't have to...

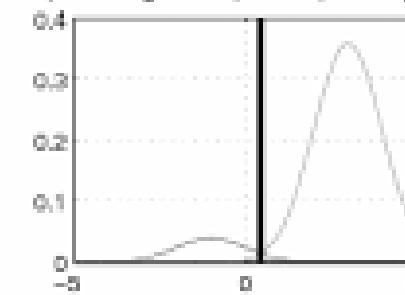
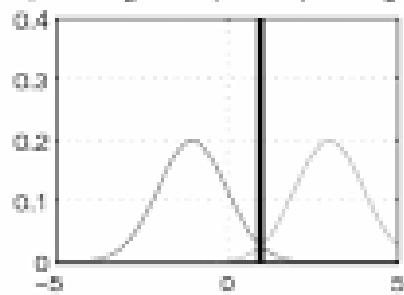
But we do not need to understand  
what the covariance matrix means

# Gaussian – Univariate, bivariate, multivariate

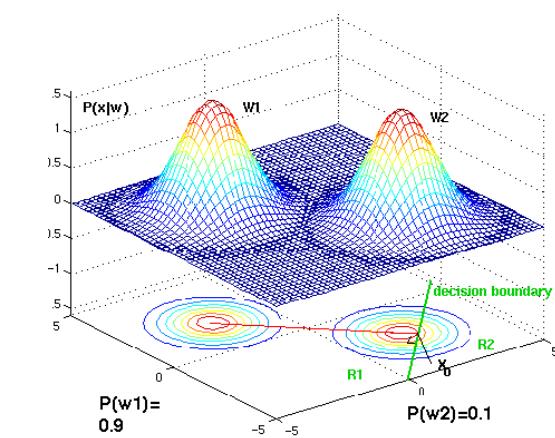
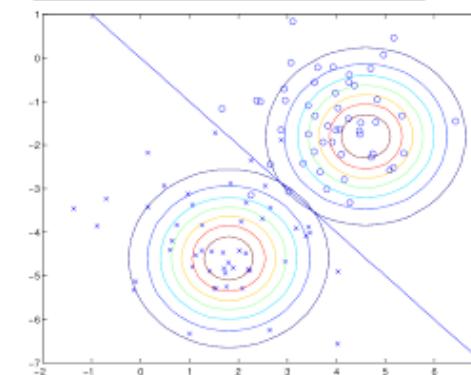
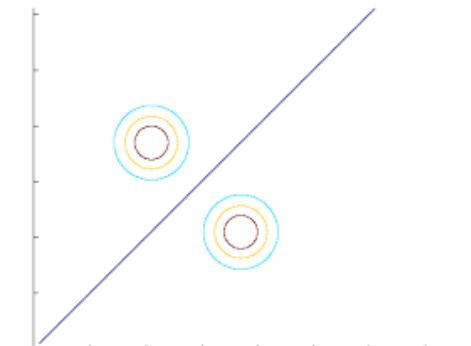
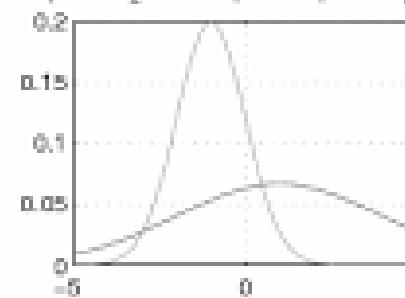
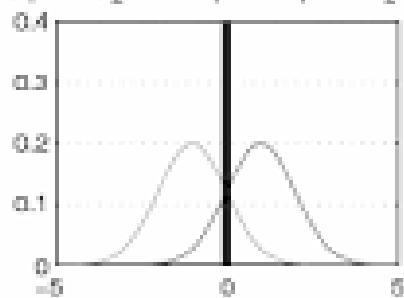


# Separation of a mixture of (two) Gaussians

$$\mu_1=1.0, \mu_2=3.0, \pi_1=0.5, \sigma_1=1.0, \sigma_2=1.0 \quad \mu_1=-1.0, \mu_2=3.0, \pi_1=0.1, \sigma_1=1.0, \sigma_2=1.0$$



$$\mu_1=1.0, \mu_2=-1.0, \pi_1=0.5, \sigma_1=1.0, \sigma_2=1.0 \quad \mu_1=1.0, \mu_2=-1.0, \pi_1=0.5, \sigma_1=0.0, \sigma_2=1.0$$





**THANK YOU**

---

**Dr.Mamatha H R**

Professor,Department of Computer Science

**mamathahr@pes.edu**

+91 80 2672 1983 Extn 834



**PES**  
**UNIVERSITY**  
**ONLINE**

# DATA ANALYTICS

## Unit 1: Data Cleaning

---

**Mamatha.H.R**

Department of Computer Science and  
Engineering

# DATA ANALYTICS

---

## Unit 1: Data Cleaning

**Mamatha H R**

Department of Computer Science and Engineering

## Data Cleaning

---

- Data in the real world is dirty: Plenty 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)

## Data Cleaning

---

- 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?

## 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 should we deal with 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**
  - This results from developers trying to squeeze in new attribute definitions to unused (bit) portions of already defined attributes (eg., trying to use 1 bit of an attribute whose range is only 31 out of 32 bits)
- Check uniqueness rule, consecutive rule and null rule
  - A **unique rule** says that each value of the given attribute must be different from all other values for that attribute
  - A **consecutive rule** says that there can be no missing values between the lowest and highest values for the attribute, and that all values must also be unique (e.g., as in check numbers)
  - A **null rule** specifies the use of blanks, question marks, special characters, or other strings that may indicate the null condition (e.g., where a value for a given attribute is not available), and how such values should be handled.
- 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 Cleaning as a Process

---

- 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)

## Exercise

---

- Explore how binning, clustering and regression are used in handling noisy data.
  
- Is combined computer and human inspection of noisy data a better way of handling the noisy data? Give reasons.
  
- Explain the process of data cleaning.

# DATA ANALYTICS

---

## Unit 1: Data Pre-processing

**Mamatha H R**

Department of Computer Science and Engineering

## Introduction

---

- Low-quality data will lead to low-quality analysis results.
- What can we say about these entries?

- Age = 122 years
- Age = 0 years
- Income = -5000



Jeanne Calment (122 years 164 days)

- How can the data be preprocessed so as to improve the efficiency and ease of the analysis process?

## Important Characteristics of Data

---

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

## 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?

## Data Quality

---

- Poor data quality negatively affects many data processing efforts

“The most important point is that poor data quality is an unfolding disaster.

- Poor data quality **costs the typical company at least ten percent (10%) of revenue**; twenty percent (20%) is probably a better estimate.”

Thomas C. Redman, DM Review, August 2004

- Example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

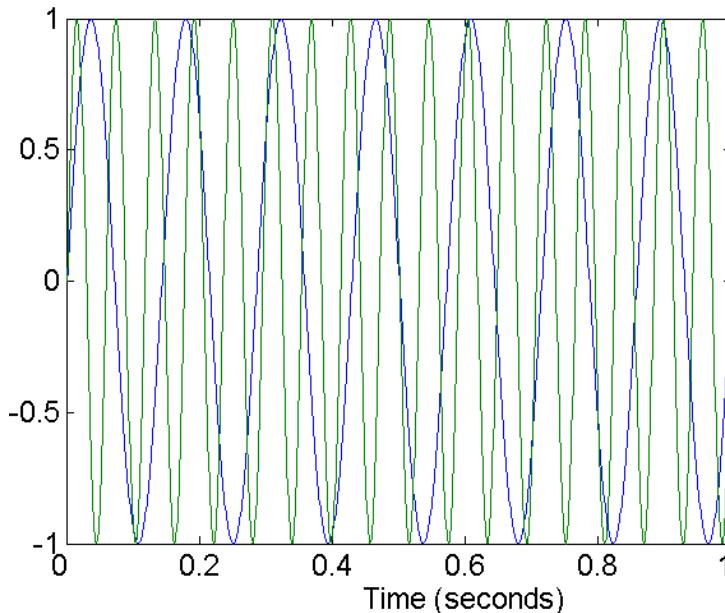
## Data Quality

---

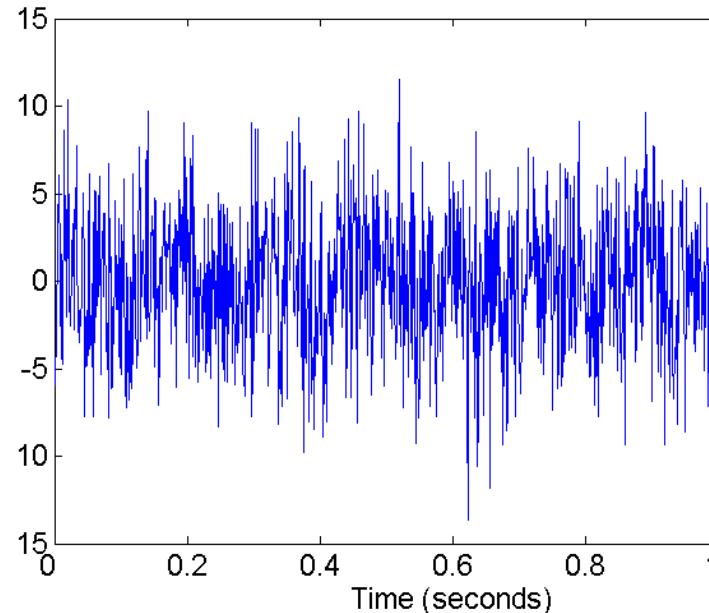
- What kinds of data quality problems are possible?
- How can we detect problems with the data?
- What can we do about these problems?
  
- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data
  - Fake data

## Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



Two Sine Waves



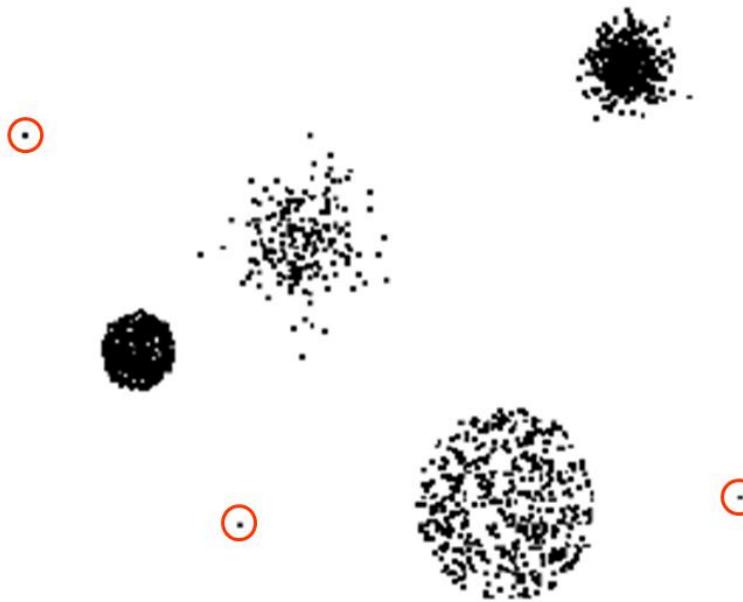
Two Sine Waves + Noise

## Outliers

- **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set

- **Case 1:** Outliers are noise that interferes with data analysis

- **Case 2:** Outliers are the goal of our analysis
  - Credit card fraud
  - Intrusion detection



## Missing Values

---

- Reasons for missing values
  - Information is not collected  
(e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases  
(e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate data objects or variables
  - Estimate missing values
    - Example: time series of temperature
    - Example: census results
  - Ignore the missing value during analysis

## Missing Values ...

---

- Missing completely at random (MCAR)
  - Example: Weighing scale ran out of batteries
  - Missingness of a value is independent of attributes
    - Has nothing to do with entries that are missing (pure chance)
  - Fill in values based on the attribute (often unrealistic)
  - Analysis may be unbiased overall
- Missing at Random (MAR)
  - Missingness is related to other variables
    - Weighing scale does not show a reading on some surfaces
  - Fill in values based on other values (often realistic)
  - Almost always produces a bias in the analysis
- Missing Not at Random (MNAR or NMAR)
  - Missingness is related to unobserved measurements
    - Weighing scale starts to wear out, weight of heavier objects more likely 'NA'
    - Censored data
  - Informative or non-ignorable missingness
    - Find more data
    - 'What if' analysis under varying scenarios
- Not possible to know the situation from the data

## Some solutions...

---

- Missing completely at random (MCAR)
  - Delete rows
    - If it is a small fraction of the data
  - Delete columns
    - If it is a small fraction of the attributes
  - Pairwise deletion
    - Compute the mean, variance and covariance with another variable on available data
    - Works reasonably if the multivariate normal assumption holds
  - Mean imputation (can be considered only for MCAR)
- Missing at Random (MAR)
  - Regression imputation
    - Unbiased estimates of mean under MCAR
    - Unbiased under MAR if factors that influence missingness included
    - Can expect false positives and spurious correlations
  - Stochastic regression imputation: same as linear regression but adds a random residual to the prediction
  - Last Observation Carried Forward (LOCF), Baseline Observation Carried Forward (BOCF) and Worst Observation Carried Forward (WOCF)
    - yield biased estimates even under MCAR
    - Use only if the assumptions that underlie these estimates are scientifically justified
  - Use of multiple imputation (mice, amelia in R)
- Missing Not at Random (MNAR or NMAR)
  - Model the missing values explicitly
    - <http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html>
    - <https://stefvanbuuren.name/fimd/sec-simplesolutions.html>

## Duplicate Data

---

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues
- When should duplicate data not be removed?

## Similarity and Dissimilarity Measures

---

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity measure
  - 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

## Major Tasks in Data Preprocessing

---

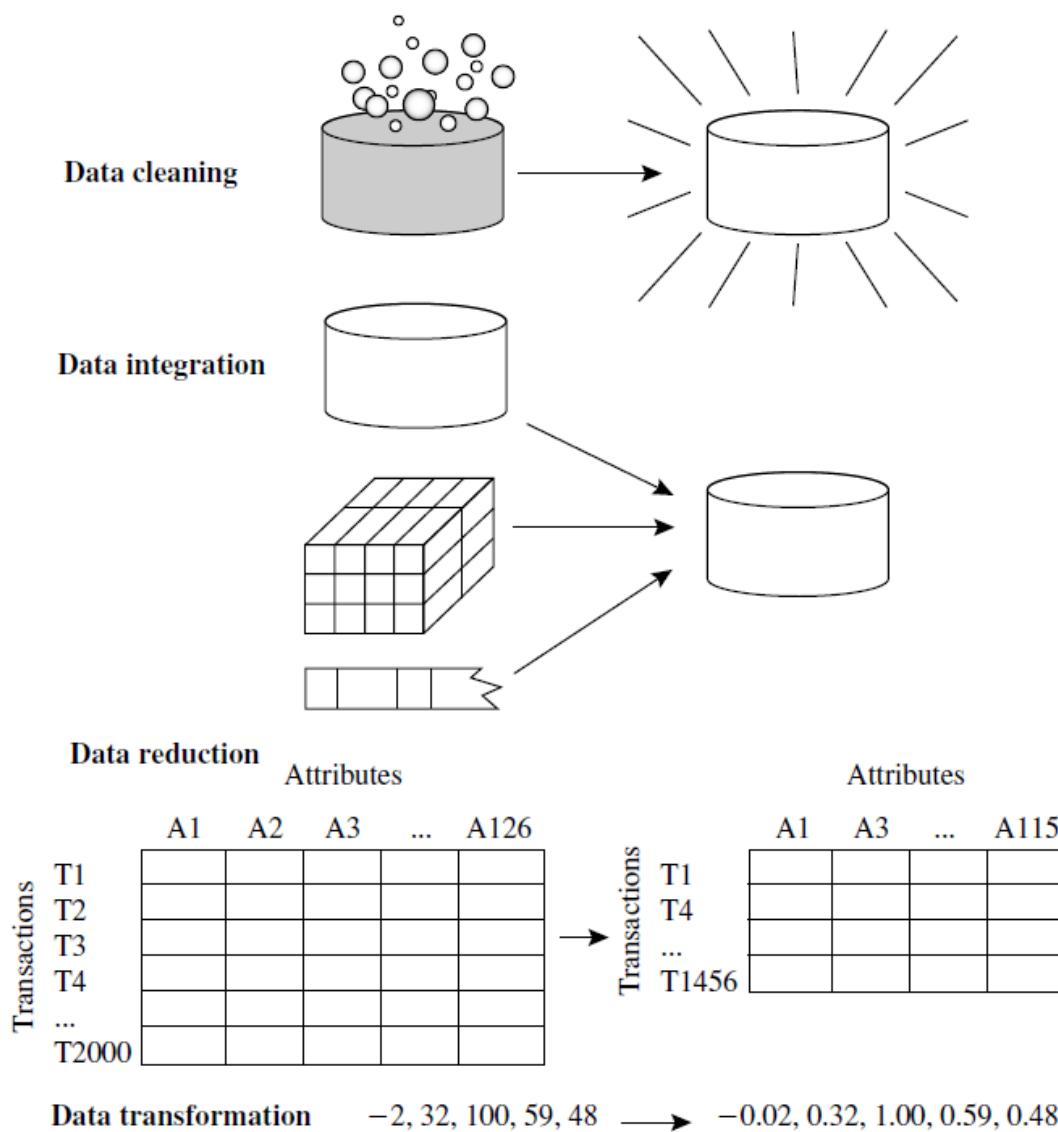
- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files

## Major Tasks in Data Preprocessing

---

- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

## Major Tasks in Data Preprocessing



## Exercise

---

- Mention the important characteristics of the data.
- Why we need to pre-process the data?
- Explain the process of data pre-processing.

## References

---

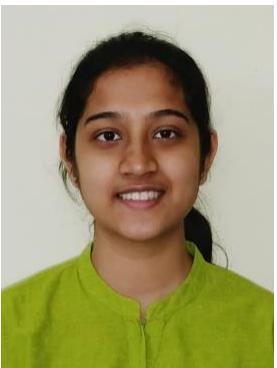
### Text Book:

- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.
- Introduction to Data Mining by Tan, Steinbach, Kumar, 2nd Edition

# DATA ANALYTICS

## “To do” before the next class...

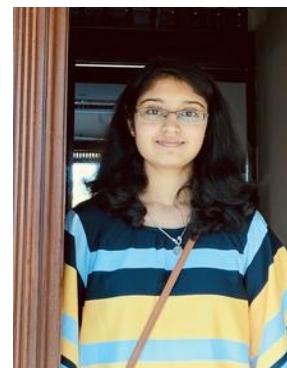
- Review the material discussed so far
  - Slides shared (and references within the slides as needed)
  - Sessions 1-7 on PESU Academy
  - Chapters 1-6 from the prescribed text  
(Business Analytics by U. Dinesh Kumar)
  - Chapter 3 from the prescribed reference  
(Data Mining by Han, Kamber, Pei)
- Install R and work on Worksheet 1
  - Painstakingly prepared by Student Mentors for Unit 1



Bharani Ujjaini Kempaiah



Ruben John



Bhavya Charan

# DATA ANALYTICS

Coming up next week...

---



- Data Integration
- Data (dimensionality) Reduction
- Data Transformations
- Getting started with Regression



**THANK YOU**

---

**Dr.Mamatha H R**

Professor,Department of Computer Science

**mamathahr@pes.edu**

+91 80 2672 1983 Extn 834



## DATA ANALYTICS

### Unit 1: Data Integration, Cleaning and Reduction

---

**Mamatha.H.R**

Department of Computer Science and  
Engineering

# DATA ANALYTICS

---

## Unit 1: Data Integration

**Mamatha H R**

Department of Computer Science and Engineering

## Data Integration

---

- Data analysis often requires data integration—the **merging of data from multiple data stores** into a coherent store.
- Careful integration can help reduce and **avoid redundancies and inconsistencies** in the resulting data set. This can help improve the accuracy and speed of the subsequent data analysis process.
- The semantic heterogeneity and structure of data pose great challenges in data integration.
- **How can we match schema and objects from different sources?**
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources

## Data Integration

---

- 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 can be detected using *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 (for Categorical/ Nominal Data)

- $\chi^2$  (chi-square) test for independence of two variables in a contingency table

- Null hypothesis: the two variables are independent

- Alternate hypothesis: the two variables are not independent

- $\chi^2$  statistic

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

- ‘Expected’= what would we ‘expect’ if the null hypothesis were true?

- Larger the  $\chi^2$  value, the more likely the variables are correlated

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

- Can be used for categorical variables where entries are numbers (counts) and not percentages or fractions (for example, 20% of 200 has to be entered as 40 in the table)

- Correlation does not imply causation

- The number of hospitals and number of car-thefts in a city may *appear to be* correlated

- Both are causally linked to a third variable: population

## Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250	200	450
Not like science fiction	50	1000	1050
Sum(col.)	300	1200	1500

	Play chess	Not play chess	Sum (row)
Like science fiction	90	360	450
Not like science fiction	210	840	1050
Sum(col.)	300	1200	1500

Actual distribution (observed)

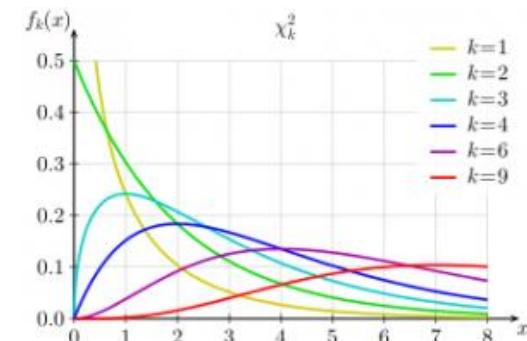
- $\chi^2$  (chi-square) calculation

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

- Degrees of freedom,  $k = (\text{no\_of\_rows}-1)(\text{no\_of\_columns}-1) = 1$
- It shows that like\_science\_fiction and play\_chess are correlated in the group

Expected distribution

$$e_{ij} = \frac{\text{sum}(A = a_i) * \text{sum}(B = b_j)}{N}$$



## Correlation Analysis (Numeric Data)

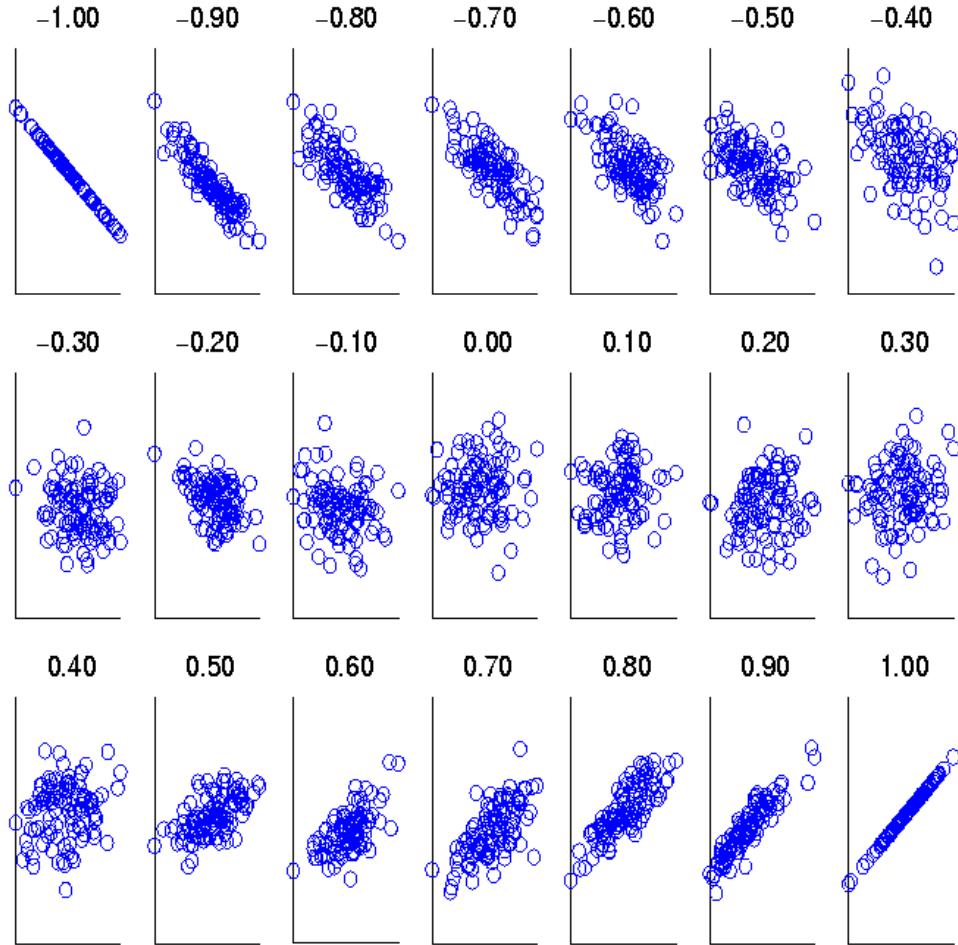
- Correlation coefficient (also called Pearson's product moment coefficient)

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

where  $n$  is the number of tuples,  $\bar{A}$  and  $\bar{B}$  are the respective means of  $A$  and  $B$ ,  $\sigma_A$  and  $\sigma_B$  are the respective standard deviation of  $A$  and  $B$ , and  $\Sigma(a_i b_i)$  is the inner product  $A^T B$  or sum of the point-wise product of  $A$  and  $B$ .

- If  $r_{A,B} > 0$ ,  $A$  and  $B$  are positively correlated ( $A$ 's values increase as  $B$ 's). The higher, the stronger correlation.
- $r_{A,B} = 0$ : independent;  $r_{AB} < 0$ : negatively correlated

## Visually Evaluating Correlation



**Scatter plots  
showing the  
similarity  
from -1 to 1.**

## Correlation (viewed as linear relationship)

---

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$$

$$b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$$

$$\text{correlation}(A, B) = A^T B$$

## Covariance (Numeric Data)

---

- Covariance is similar to correlation

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

Correlation coefficient:  $r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$

where n is the number of tuples,  $\bar{A}$  and  $\bar{B}$  are the respective mean or **expected values** of A and B,  $\sigma_A$  and  $\sigma_B$  are the respective standard deviation of A and B.

## Covariance (Numeric Data)

---

- **Positive covariance:** If  $\text{Cov}_{A,B} > 0$ , then A and B both tend to be larger than their expected values.
- **Negative covariance:** If  $\text{Cov}_{A,B} < 0$  then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence:**  $\text{Cov}_{A,B} = 0$ , but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply statistical independence

## Co-Variance: An Example

---

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

## Co-Variance: An Example

---

Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$E(A) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$$

$$E(B) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$$

$$\text{Cov}(A,B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$$

Thus, A and B rise together since  $\text{Cov}(A, B) > 0$ .

## Tuple Duplication

---

- In addition to detecting redundancies between attributes, duplication should also be detected at the tuple level (e.g., where there are two or more identical tuples for a given unique data entry case).
- The use of denormalized tables (often done to improve performance by avoiding joins) is another source of data redundancy

## Data Value Conflict Detection and Resolution

---

- Data integration also involves the detection and resolution of data value conflicts.
- For example, for the same real-world entity, attribute values from different sources may differ.
- This may be due to differences in representation, scaling, or encoding.
- For instance, a weight attribute may be stored in metric units in one system and British imperial units in another.

## Exercise

---

- Explain how redundancy is handled in data integration.
- Compare and contrast Correlation and Covariance.

## References

---

### Text Book:

- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.

# DATA ANALYTICS

---

## Unit 1:Data Reduction

**Mamatha H R, Gowri Srinivasa**

Department of Computer Science and Engineering

## Data Reduction

---

**Data reduction:** Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

## Data Reduction Strategies

---

### Data reduction strategies

- Dimensionality reduction, e.g., remove unimportant attributes
  - Wavelet transforms
  - Principal Components Analysis (PCA)
  - Feature subset selection, feature creation
- Numerosity reduction (some simply call it: Data Reduction)
  - Regression and Log-Linear Models
  - Histograms, clustering, sampling
  - Data cube aggregation
- Data compression

## Data Reduction 1: Dimensionality Reduction

---

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially

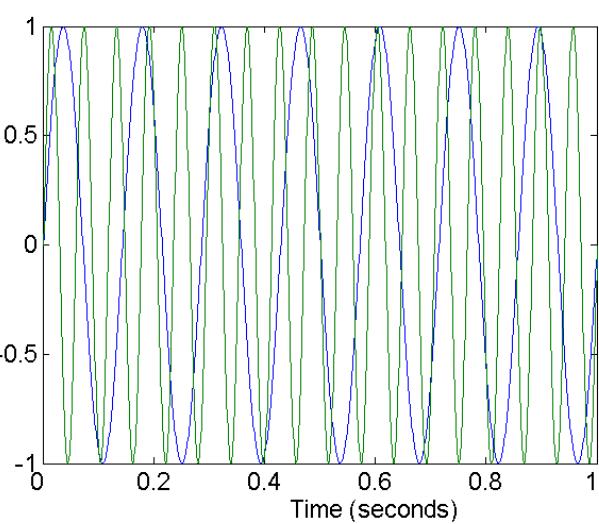
## Data Reduction 1: Dimensionality Reduction

---

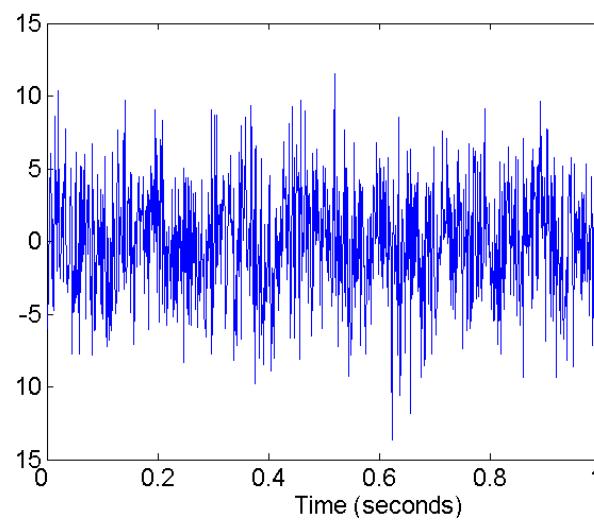
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

## Mapping Data to a New Space

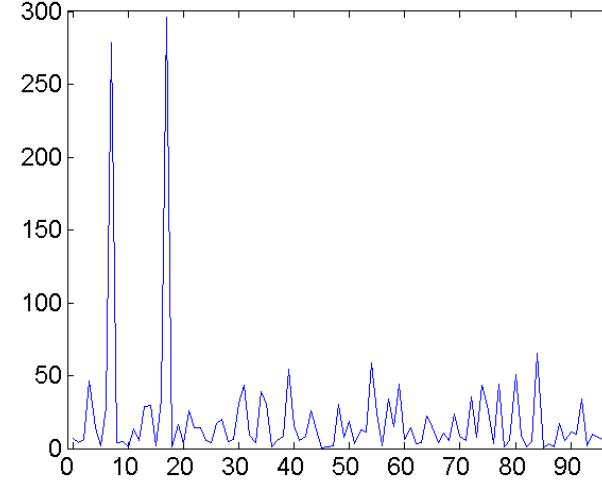
- Fourier transform
- Wavelet transform



Two Sine Waves



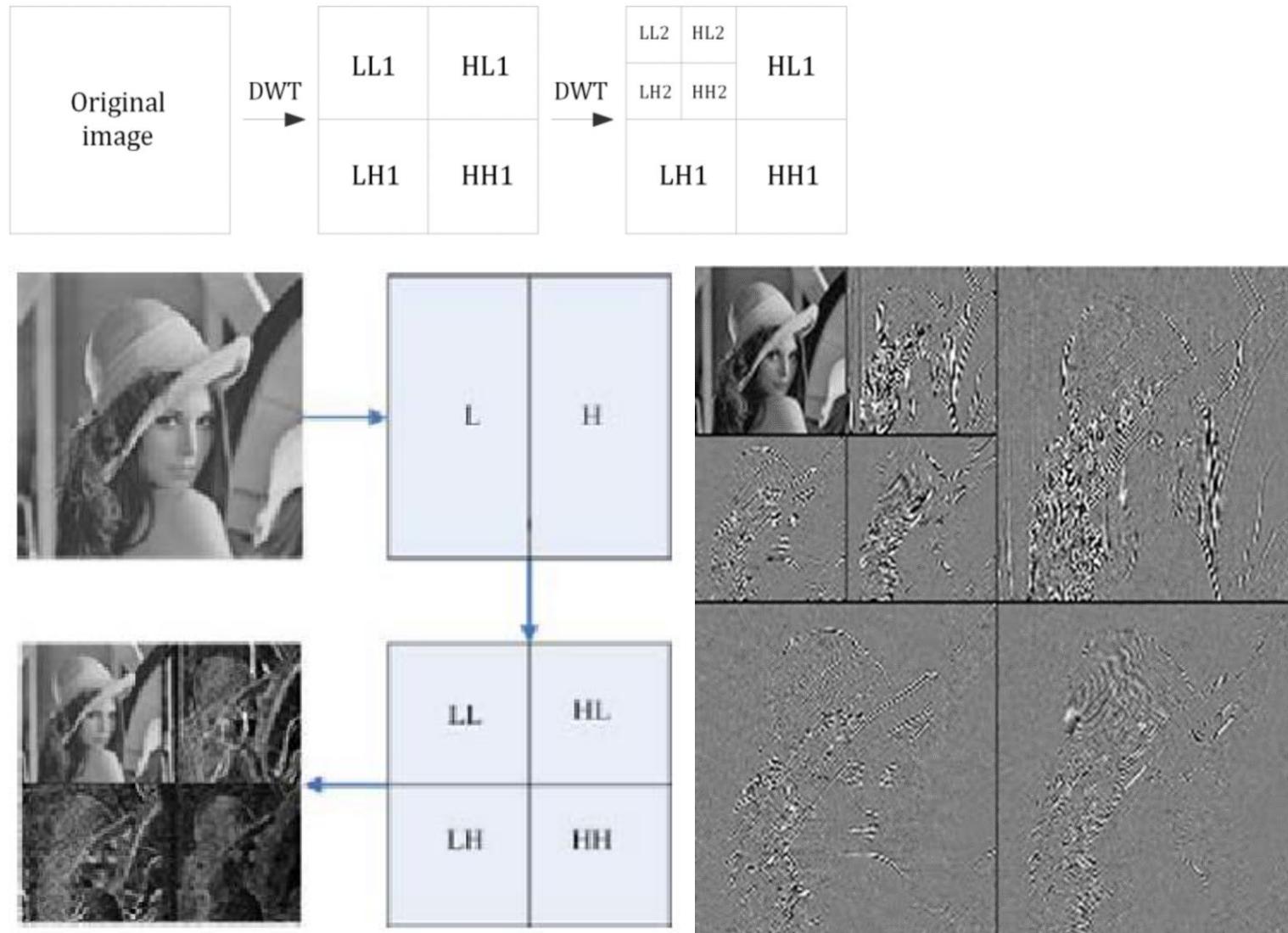
Two Sine Waves + Noise



Frequency

## What Is Wavelet Transform?

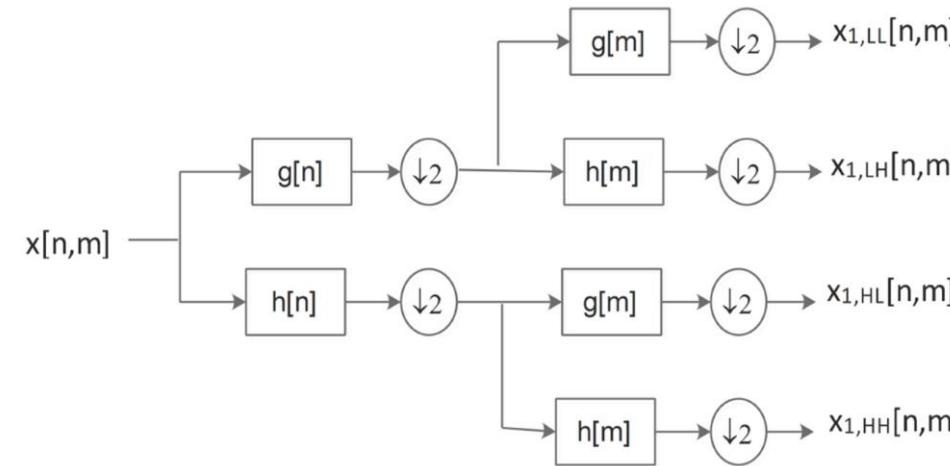
- Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression



## Wavelet Transformation

Method:

- Length, L, must be an integer power of 2 (padding with 0's, when necessary)
- Each transform has 2 functions: smoothing (g), difference (h)
- Applies to pairs of data, resulting in two set of data of length L/2
- Applies the two functions recursively, until the desired level of decomposition is reached



$x(n,1)$	56	40	8	24	48	48	40	16
----------	----	----	---	----	----	----	----	----

$$g(n) = \frac{1}{2}[1,1] \quad 48 \quad 24 \quad 16 \quad 36 \quad 48 \quad 44 \quad 28 \quad 16$$

$$h(n) = [1, -1] \quad 16 \quad 32 \quad 16 \quad 24 \quad 0 \quad 8 \quad 24 \quad 0$$

$$g(n) \downarrow 2 \quad 48 \quad 16 \quad 48 \quad 28$$

$$h(n) \downarrow 2 \quad 32 \quad 24 \quad 8 \quad 0$$

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

## Why Wavelet Transform?

---

- Use hat-shape filters
  - Emphasize region where points cluster
  - Suppress weaker information in their boundaries
- Effective removal of outliers
  - Insensitive to noise, insensitive to input order
- Multi-resolution
  - Detect arbitrary shaped clusters at different scales
- Efficient
  - Complexity  $O(N)$
- Only applicable to low dimensional data

# Principal component analysis

---

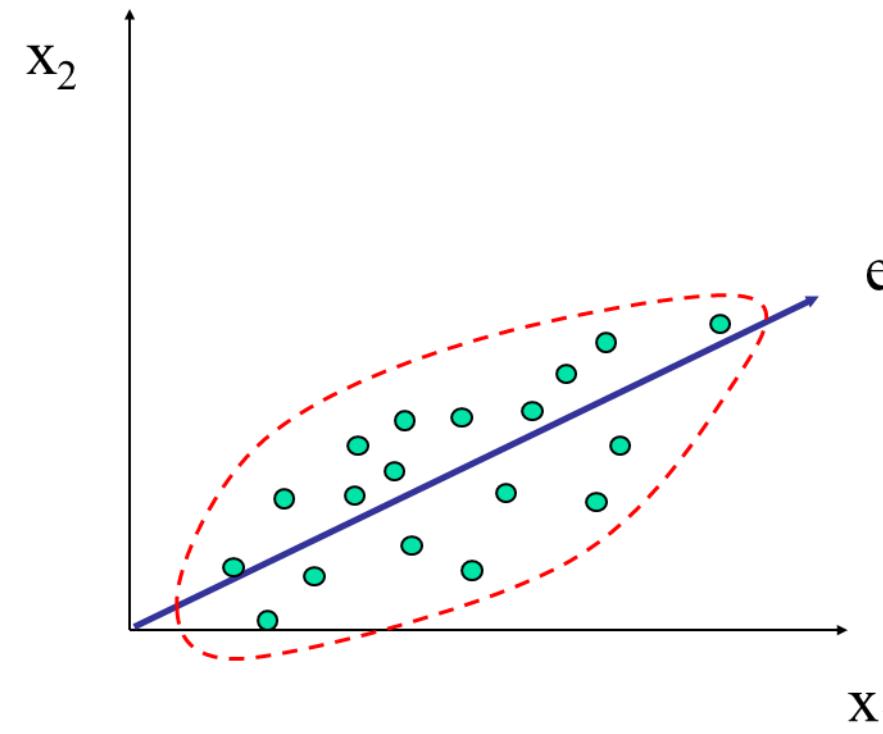
- Simplify data
- Understand relationship between variables
- Get an insight to patterns

## Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data

- The original data are projected onto a much smaller space, resulting in dimensionality reduction.

We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space

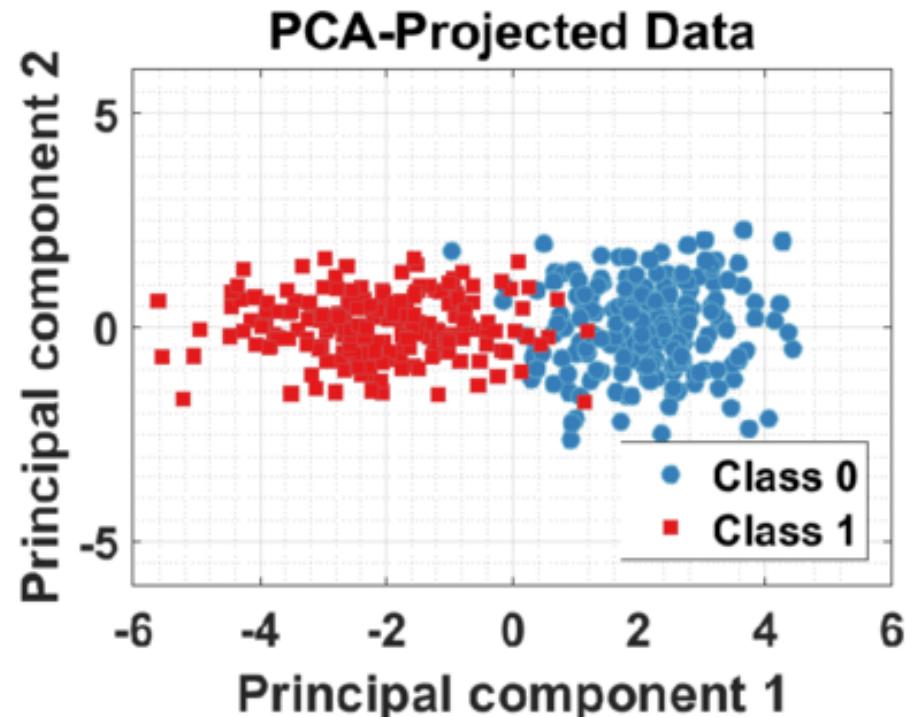
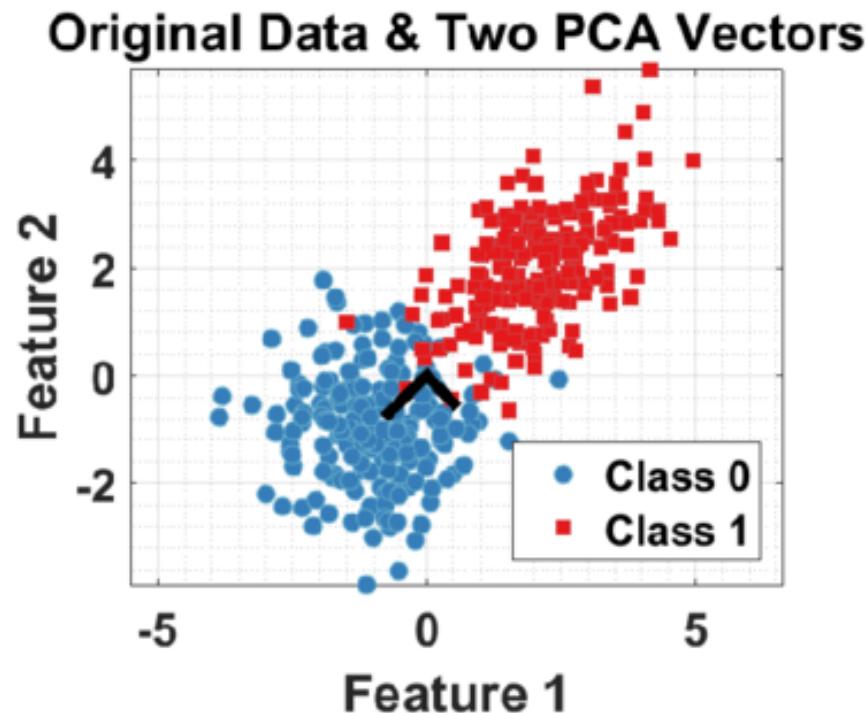


## Principal Component Analysis (Steps)

---

- Given  $N$  data vectors from  $n$ -dimensions, find  $k \leq n$  orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute  $k$  orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the  $k$  principal component vectors
  - The principal components are sorted in order of decreasing “significance” or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
  - Works for numeric data only

# PCA: Data in the Eigen Space – A different representation



[https://www.researchgate.net/publication/320410861\\_Physically\\_Motivated\\_Feature\\_Development\\_for\\_Machine\\_Learning\\_Applications/figures?lo=1](https://www.researchgate.net/publication/320410861_Physically_Motivated_Feature_Development_for_Machine_Learning_Applications/figures?lo=1)

# Principal component analysis

---

- Get data
- Subtract mean
  - (or bring it to zero mean, unit standard deviation form)
- Compute the covariance matrix
- Find Eigen values and Eigen vectors
- Select principal Eigen vectors (PCA)
  - Use proportion of variance retained by an eigen vector (using eigen values)
- Project data onto selected Eigen vectors
- Plot data

## PCA example

	x	y		x	y
Data =	2.5	2.4		.69	.49
	0.5	0.7		-1.31	-1.21
	2.2	2.9		.39	.99
	1.9	2.2		.09	.29
	3.1	3.0	DataAdjust =	1.29	1.09
	2.3	2.7		.49	.79
	2	1.6		.19	-.31
	1	1.1		-.81	-.81
	1.5	1.6		-.31	-.31
	1.1	0.9		-.71	-1.01

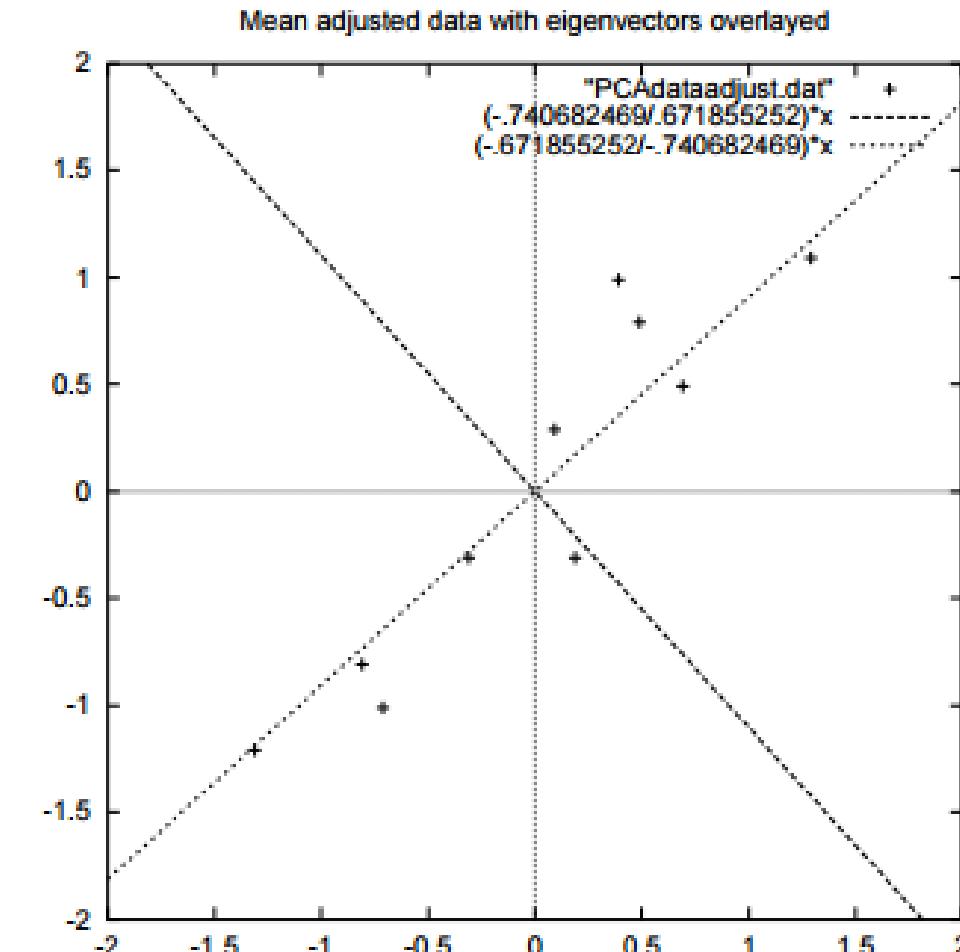
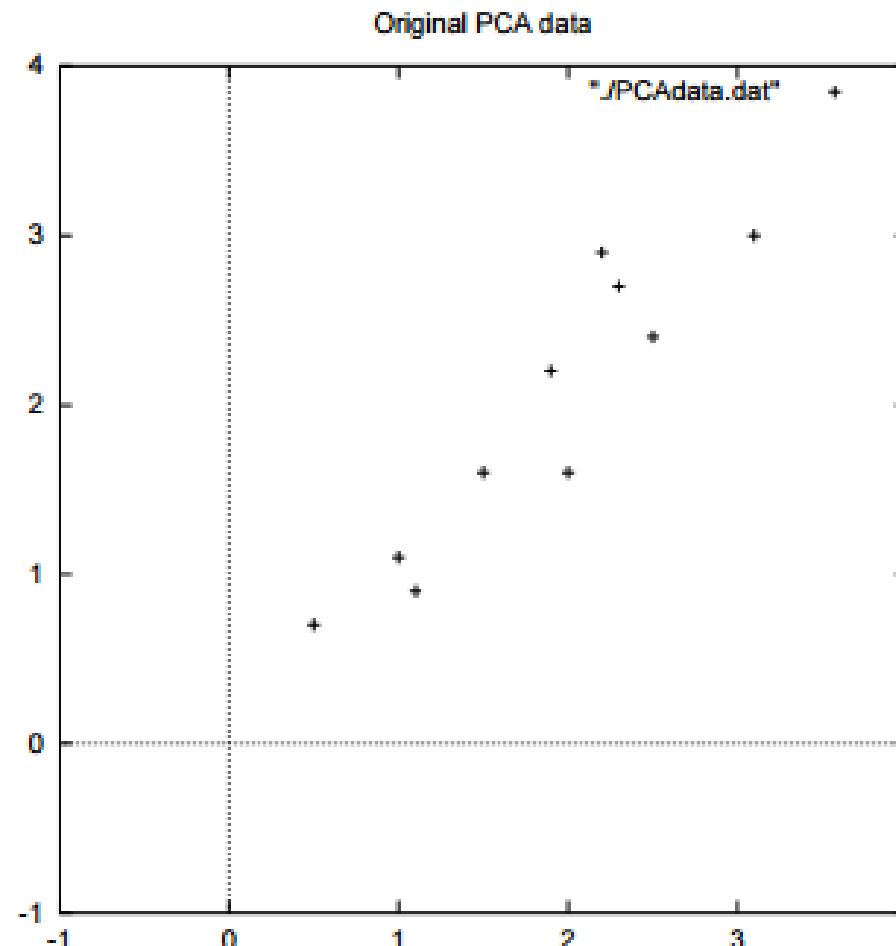
# Covariance, Eigen analysis

$$ccv = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

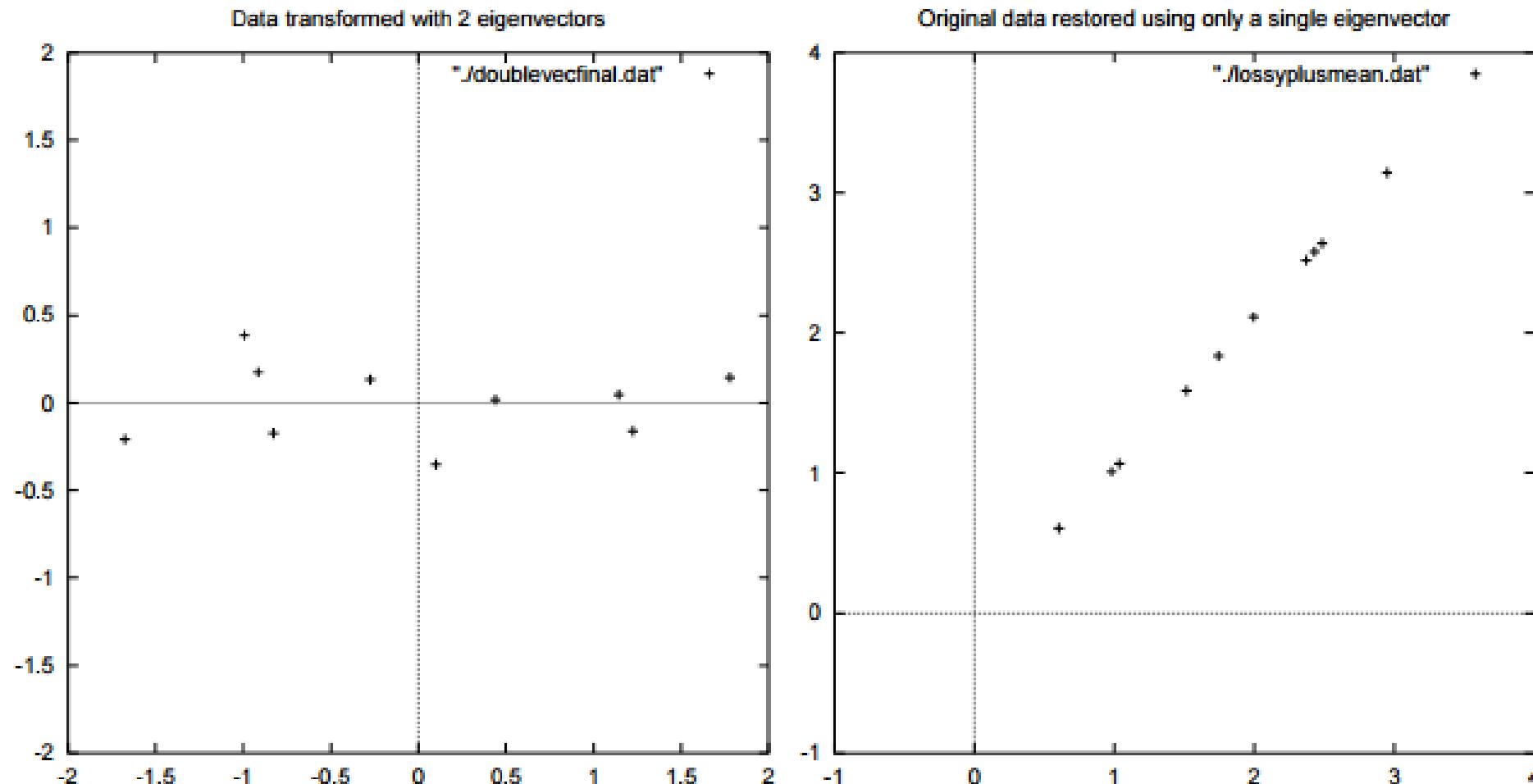
$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

# Choosing an appropriate ‘axis’



# A new representation





**THANK YOU**

---

**Dr.Mamatha H R**

Professor,Department of Computer Science

**mamathahr@pes.edu**

+91 80 2672 1983 Extn 834



## DATA ANALYTICS

### Unit 1: Data Integration, Cleaning and Reduction

---

**Mamatha.H.R**

Department of Computer Science and  
Engineering

# DATA ANALYTICS

---

## Unit 1: Data Reduction (contd.)

**Mamatha H R, Gowri Srinivasa**

Department of Computer Science and Engineering

## Wavelet Decomposition

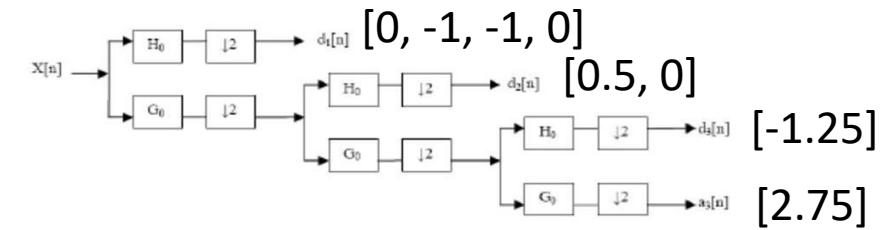
- Wavelets: A math tool for space-efficient hierarchical decomposition of functions

- $S = [2, 2, 0, 2, 3, 5, 4, 4]$  can be transformed to

$$S_\Delta = [2^3/4, -1^1/4, 1/2, 0, 0, -1, -1, 0]$$

- Compression: many small detail coefficients can be replaced by 0's, and only the significant coefficients are retained

Resolution	Averages	Detail Coefficients
8	$[2, 2, 0, 2, 3, 5, 4, 4]$	
4	$[2, 1, 4, 4]$	$[0, -1, -1, 0]$
2	$[1\frac{1}{2}, 4]$	$[\frac{1}{2}, 0]$
1	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$



## Haar Wavelet Coefficients

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

$$H_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

### Coefficient “Supports”

2.75 

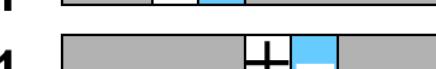
-1.25 

0.5 

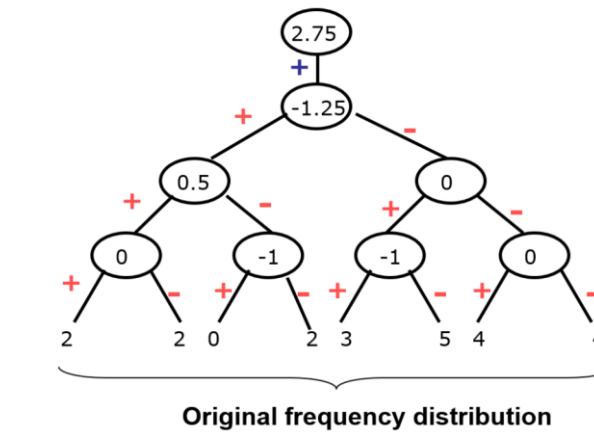
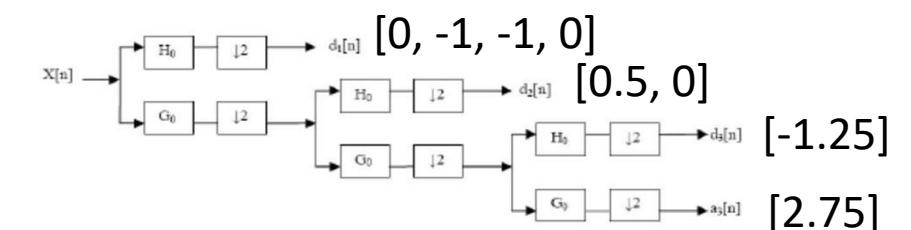
0 

0 

-1 

-1 

0 



Hierarchical decomposition structure (a.k.a. “error tree”)

# DATA ANALYTICS

## Zeroing out detailed coefficients



JPEG



JPEG 2000



JPEG



JPEG 2000

## Why Wavelet Transform?

---

- Use hat-shape filters
  - Emphasize region where points cluster
  - Suppress weaker information in their boundaries
- Effective removal of outliers
  - Insensitive to noise, insensitive to input order
- Multi-resolution
  - Detect arbitrary shaped clusters at different scales
- Efficient
  - Complexity  $O(N)$
- Only applicable to low dimensional data

# Principal component analysis

---

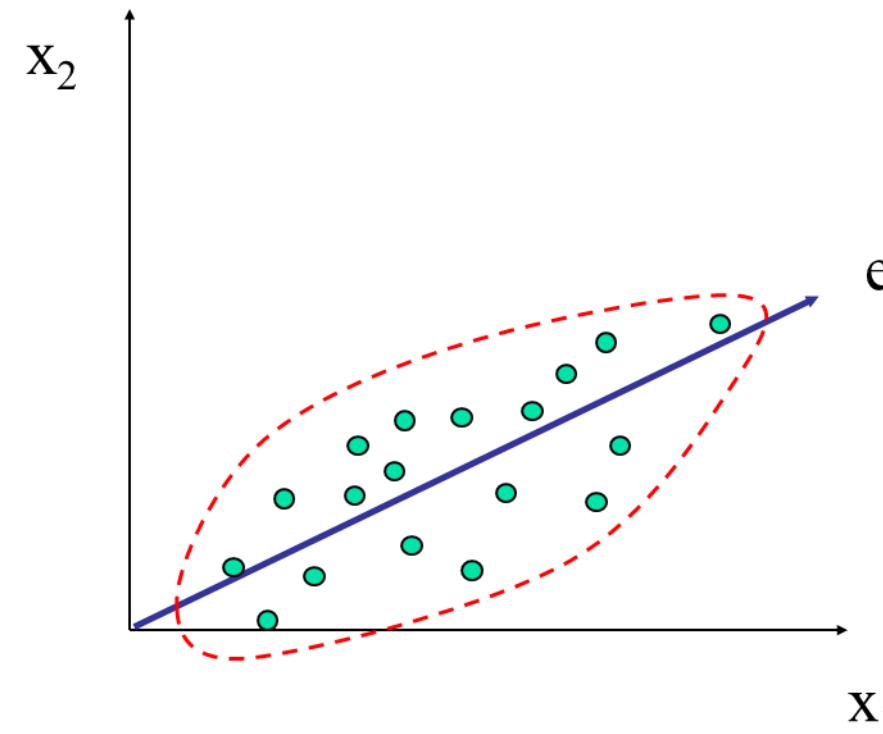
- Simplify data
- Understand relationship between variables
- Get an insight to patterns

## Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data

- The original data are projected onto a much smaller space, resulting in dimensionality reduction.

We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space

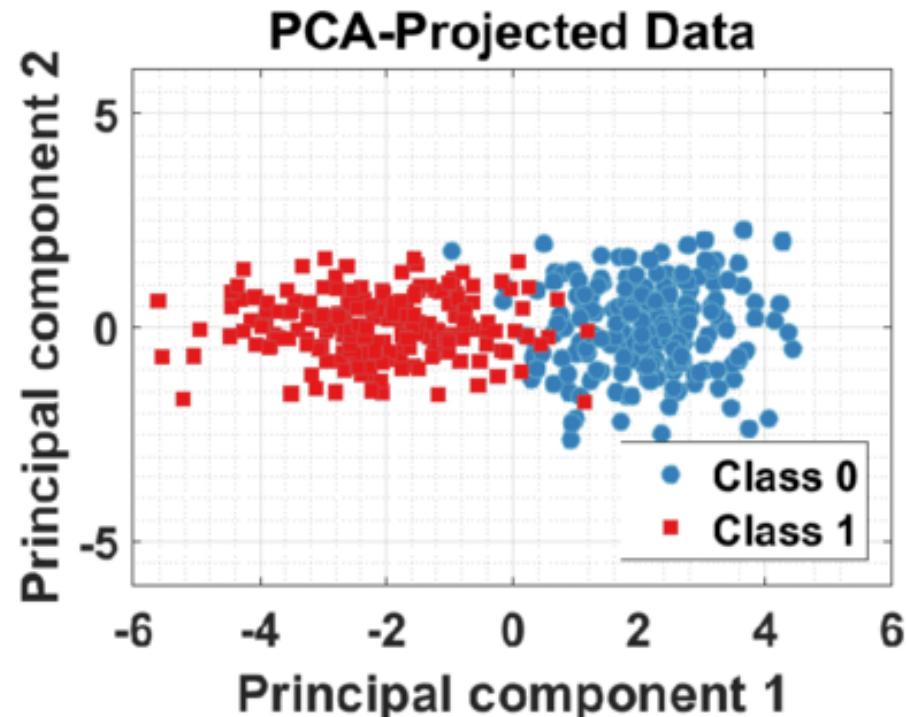
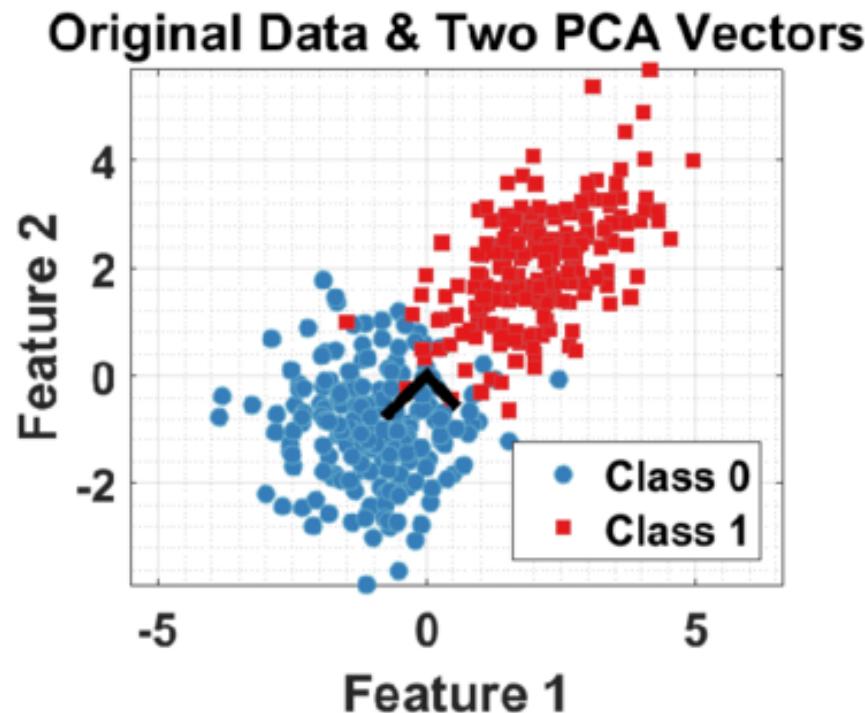


## Principal Component Analysis (Steps)

---

- Given  $N$  data vectors from  $n$ -dimensions, find  $k \leq n$  orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute  $k$  orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the  $k$  principal component vectors
  - The principal components are sorted in order of decreasing “significance” or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
  - Works for numeric data only

# PCA: Data in the Eigen Space – A different representation



[https://www.researchgate.net/publication/320410861\\_Physically Motivated Feature Development for Machine Learning Applications/figures?lo=1](https://www.researchgate.net/publication/320410861_Physically_Motivated_Feature_Development_for_Machine_Learning_Applications/figures?lo=1)

# Principal component analysis

---

- Get data
- Subtract mean
  - (or bring it to zero mean, unit standard deviation form)
- Compute the covariance matrix
- Find Eigen values and Eigen vectors
- Select principal Eigen vectors (PCA)
  - Use proportion of variance retained by an eigen vector (using eigen values)
- Project data onto selected Eigen vectors
- Plot data

## PCA example

	x	y		x	y
Data =	2.5	2.4		.69	.49
	0.5	0.7		-1.31	-1.21
	2.2	2.9		.39	.99
	1.9	2.2		.09	.29
	3.1	3.0	DataAdjust =	1.29	1.09
	2.3	2.7		.49	.79
	2	1.6		.19	-.31
	1	1.1		-.81	-.81
	1.5	1.6		-.31	-.31
	1.1	0.9		-.71	-1.01

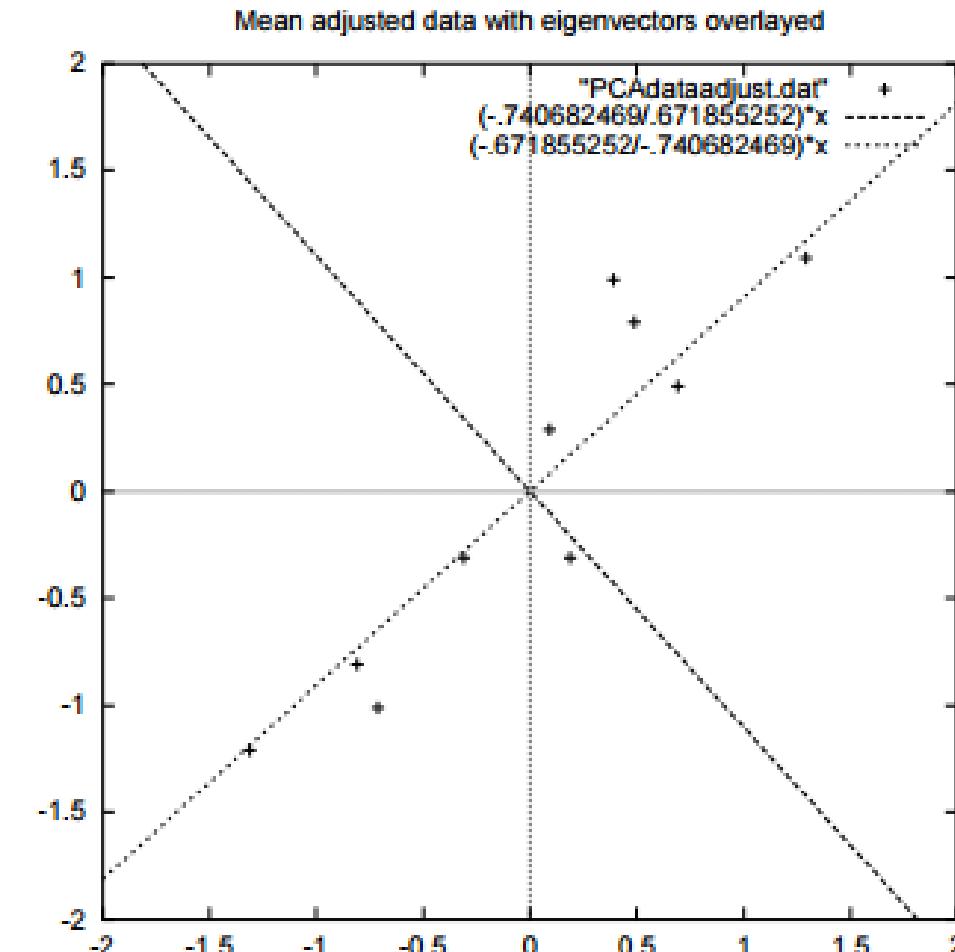
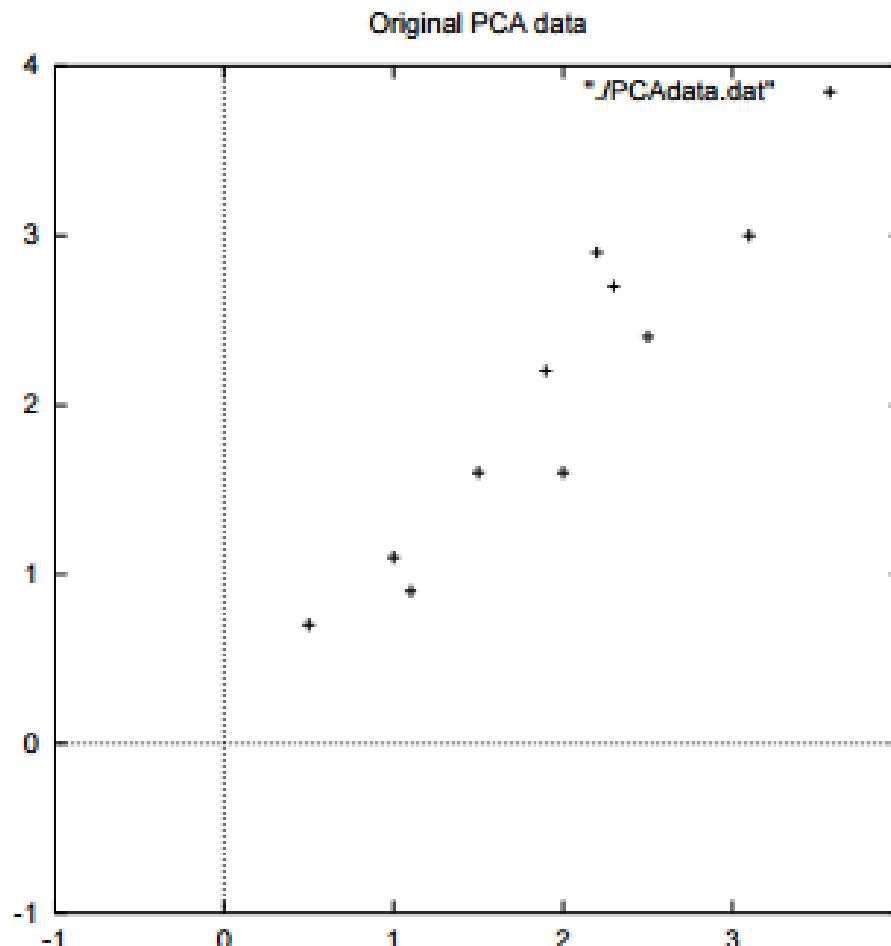
# Covariance, Eigen analysis

$$ccv = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

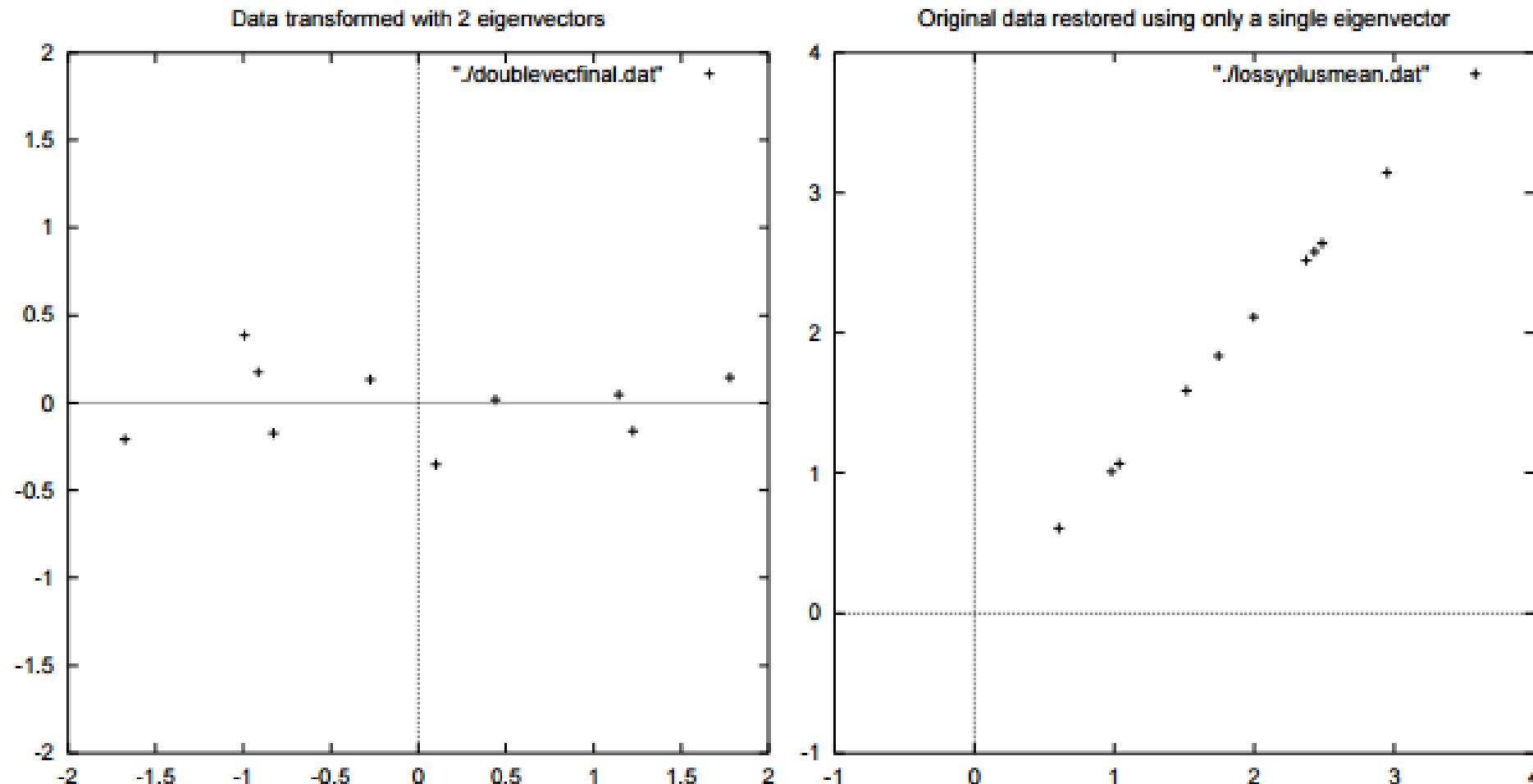
$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

# Choosing an appropriate ‘axis’



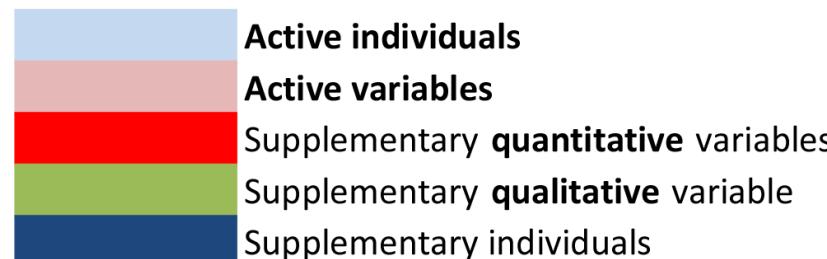
# A new representation



# PCA using R – 1

## (factoMineR, factoextra)

name	100m	Long.jump	//	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58		63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4		60.15	301.5	2	8122	Decastar
Macey	10.89	7.47		58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74		55.39	278.05	5	8343	OlympicG
\\"								
Zsivoczky	10.91	7.14		63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19		57.76	264.35	7	8237	OlympicG
Pogorelov	10.95	7.31		53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3		60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99		64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3		50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6		51.77	278.1	6	8030	Decastar
Nool	10.8	7.53		61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38		51.53	274.21	19	7926	OlympicG



<http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/>

# PCA using R - 2

---

- Selecting the principal components

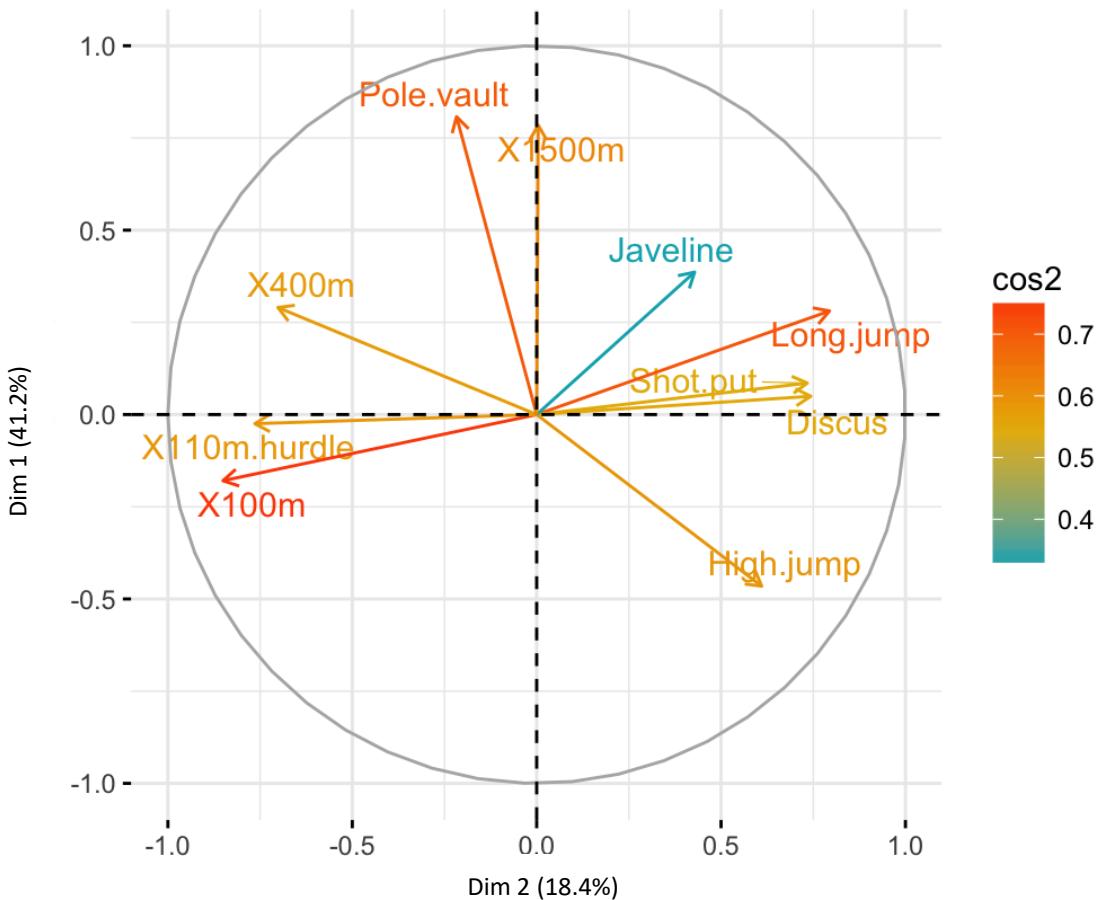
```
library("factoextra") eig.val <-get_eigenvalue(res.pca)  
eig.val
```

```
##          eigenvalue variance.percent cumulative.variance.percent  
## Dim.1      4.124           41.24                  41.2  
## Dim.2      1.839           18.39                  59.6  
## Dim.3      1.239           12.39                  72.0  
## Dim.4      0.819            8.19                  80.2  
## Dim.5      0.702            7.02                  87.2  
## Dim.6      0.423            4.23                  91.5  
## Dim.7      0.303            3.03                  94.5  
## Dim.8      0.274            2.74                  97.2  
## Dim.9      0.155            1.55                  98.8  
## Dim.10     0.122            1.22                 100.0
```

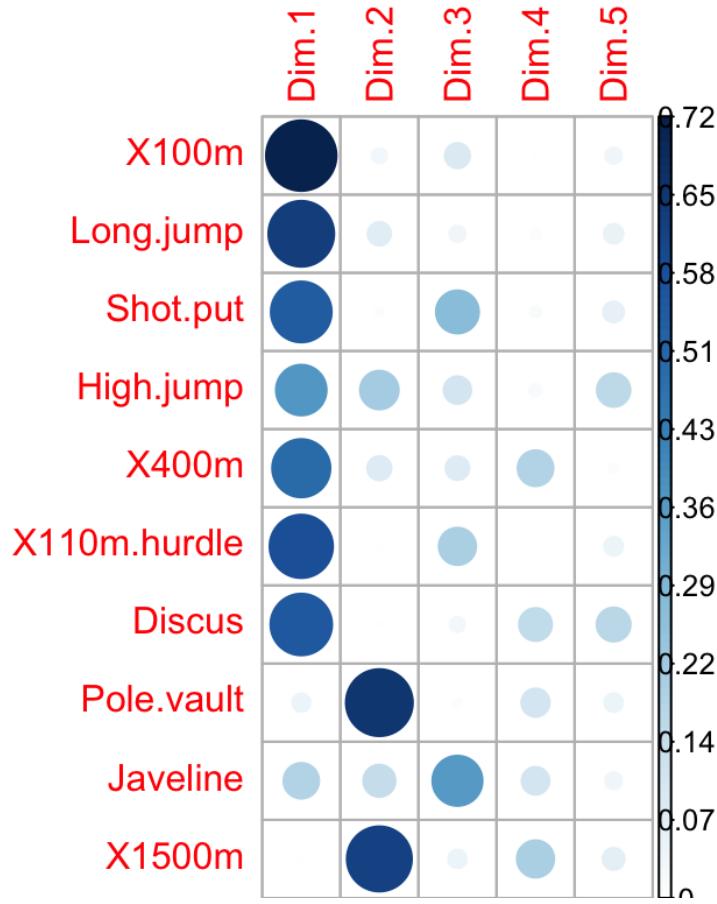
## PCA using R - 3

- Correlation circle

Variables - PCA



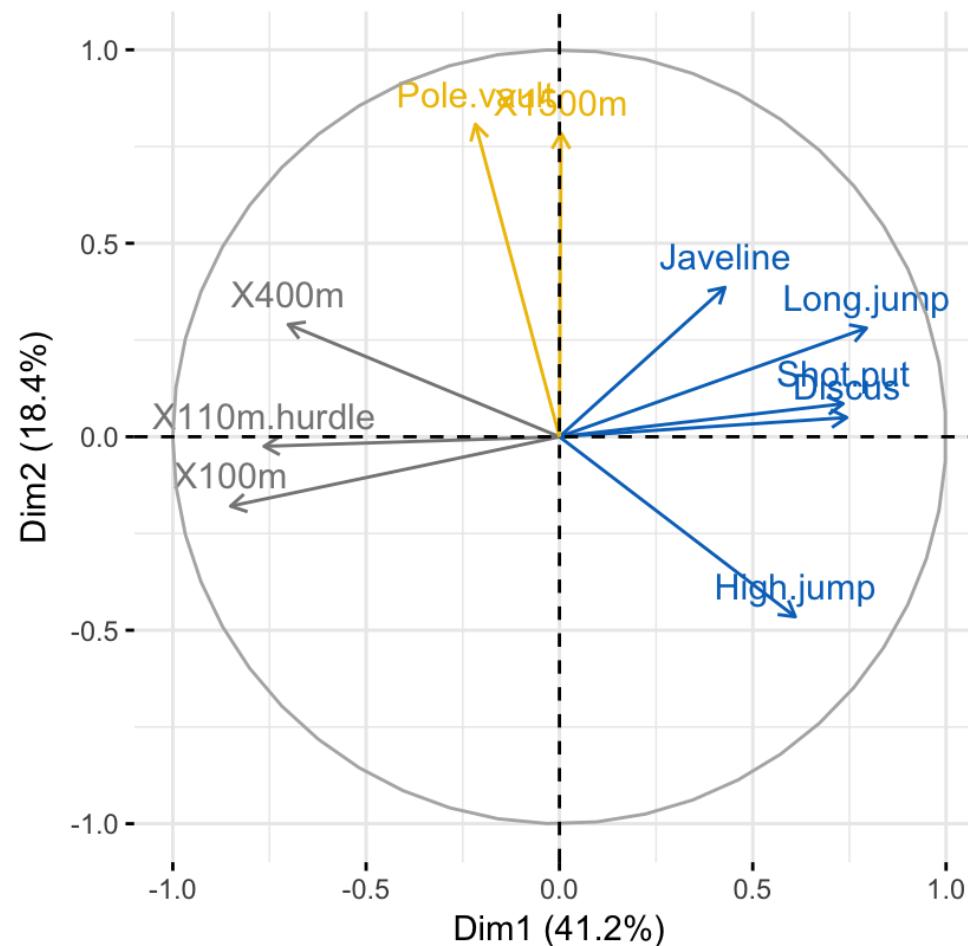
`var.cos2 =  
var.coord * var.coord`



## PCA using R - 4

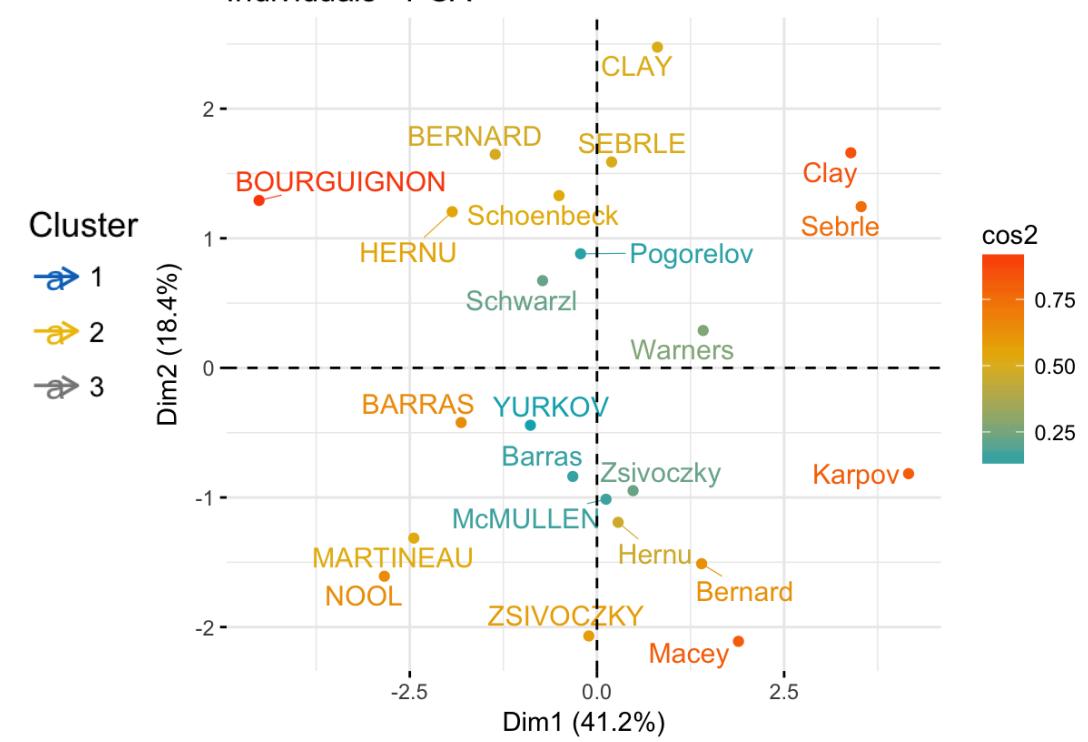
- Which events are similar?

Variables - PCA



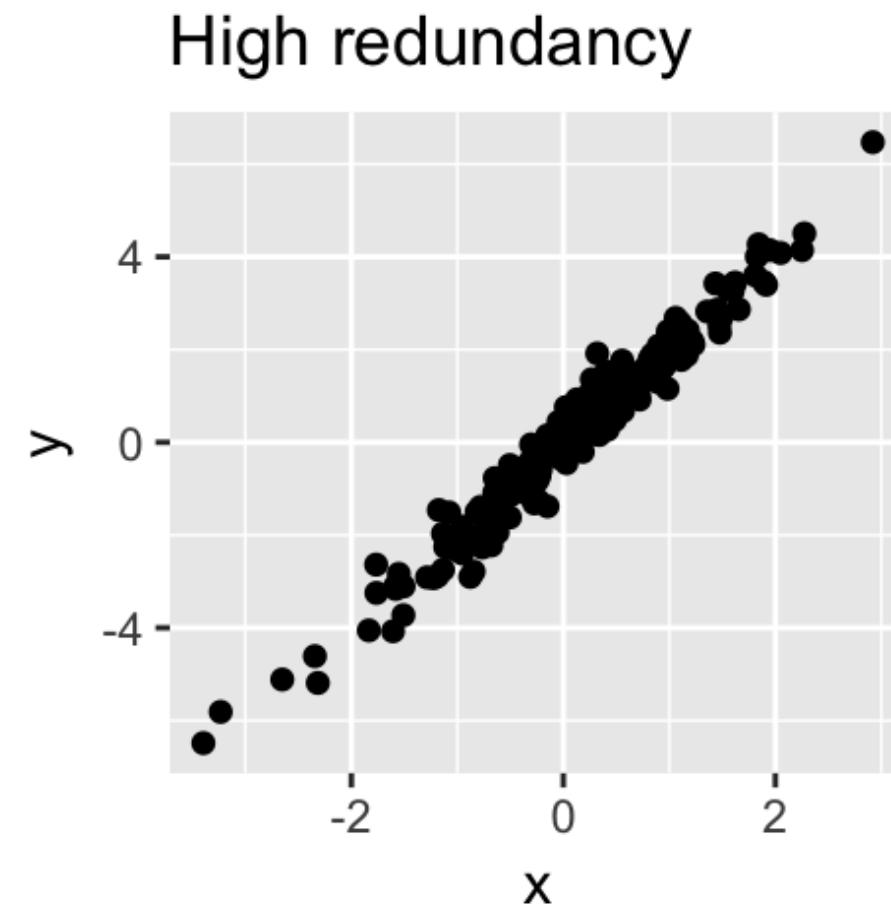
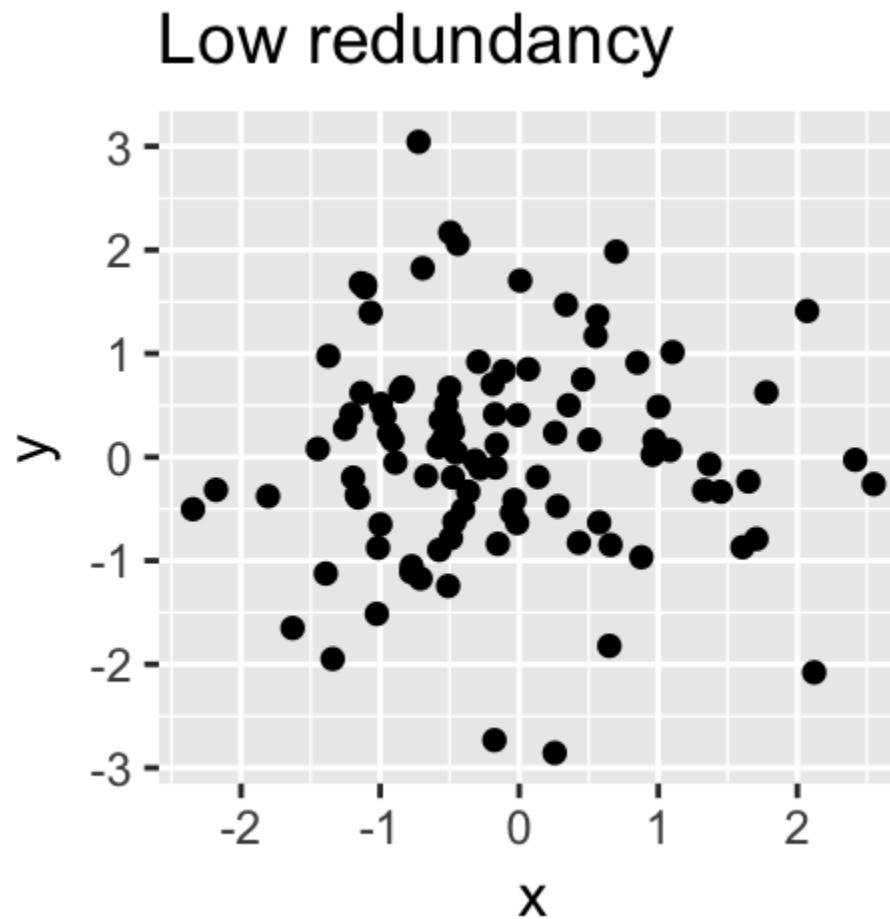
- Which athletes are similar?

Individuals - PCA



# Can PCA be used on all data?

- PCA helps to un-correlate data



## Attribute Subset Selection

---

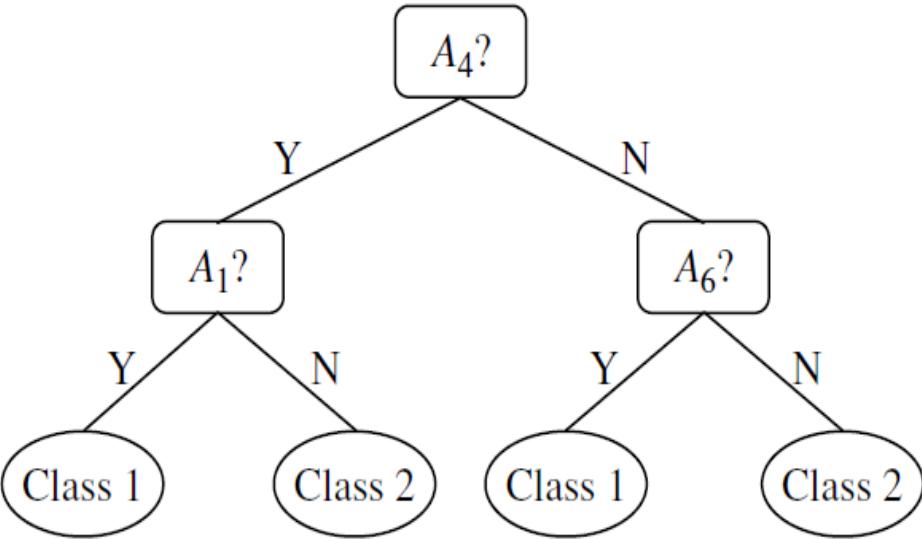
- Another way to reduce dimensionality of data is to eliminate
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data analysis task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

## Heuristic Search in Attribute Selection

---

- There are  $2^d$  possible attribute combinations of  $d$  attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination
  - Optimal branch and bound:
    - Use attribute elimination and backtracking

## Heuristic Search in Attribute Selection

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$   <pre> graph TD     A4[A4?] -- Y --&gt; A1[A1?]     A4 -- N --&gt; A6[A6?]     A1 -- Y --&gt; Class1_1((Class 1))     A1 -- N --&gt; Class2_1((Class 2))     A6 -- Y --&gt; Class1_2((Class 1))     A6 -- N --&gt; Class2_2((Class 2))   </pre> $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$

## Attribute Creation (Feature Generation)

---

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space (see: data reduction)
    - E.g., Fourier transformation, wavelet transformation, manifold approaches
  - Attribute construction
    - Combining features
    - Data discretization

## Exercise

---

- ❑ Mention and explain the different data reduction strategies.**
  
- ❑ Explain how Wavelet transform and Principal Component Analysis are used in the process of data reduction.**

## Data Reduction 2: Numerosity Reduction

---

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- Parametric methods (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Ex.: Log-linear models—obtain value at a point in  $m$ -D space as the product on appropriate marginal subspaces
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling, ...

## Parametric Data Reduction: Regression and Log-Linear Models

---

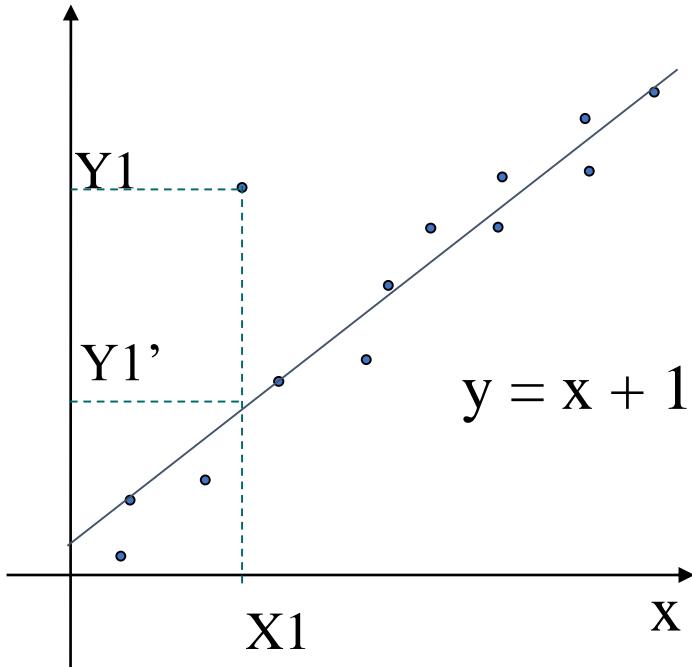
- **Linear regression**
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
- **Multiple regression**
  - Allows a response variable  $Y$  to be modeled as a linear function of multidimensional feature vector
- **Log-linear model**
  - Approximates discrete multidimensional probability distributions

## Regression Analysis

---

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or ***measurement***) and of one or more ***independent variables*** (aka. ***explanatory variables*** or ***predictors***)
- The parameters are estimated so as to give a "**best fit**" of the data
- Most commonly the best fit is evaluated by using the ***least squares method***, but other criteria have also been used

## Regression Analysis



- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

## Regression Analysis and Log-Linear Models

---

- **Linear regression:**  $Y = w X + b$
- Two regression coefficients,  $w$  and  $b$ , specify the line and are to be estimated by using the data at hand
- Using the least squares criterion to the known values of  $Y_1, Y_2, \dots, X_1, X_2, \dots$
- **Multiple regression:**  $Y = b_0 + b_1 X_1 + b_2 X_2$
- Many nonlinear functions can be transformed into the above

## Regression Analysis and Log-Linear Models

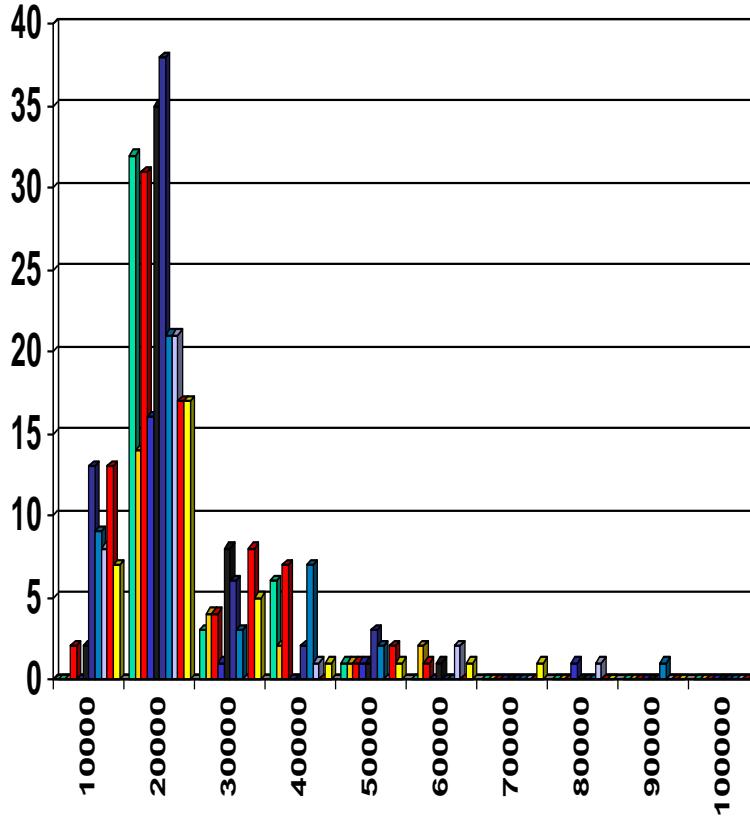
---

### Log-linear models:

- Approximate discrete multidimensional probability distributions
- Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
- Useful for dimensionality reduction and data smoothing

## Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms

The process of identifying a subset from a population of elements (aka observations or cases) is called **sampling process** or **simply sampling**

### Steps used in any Sampling process:

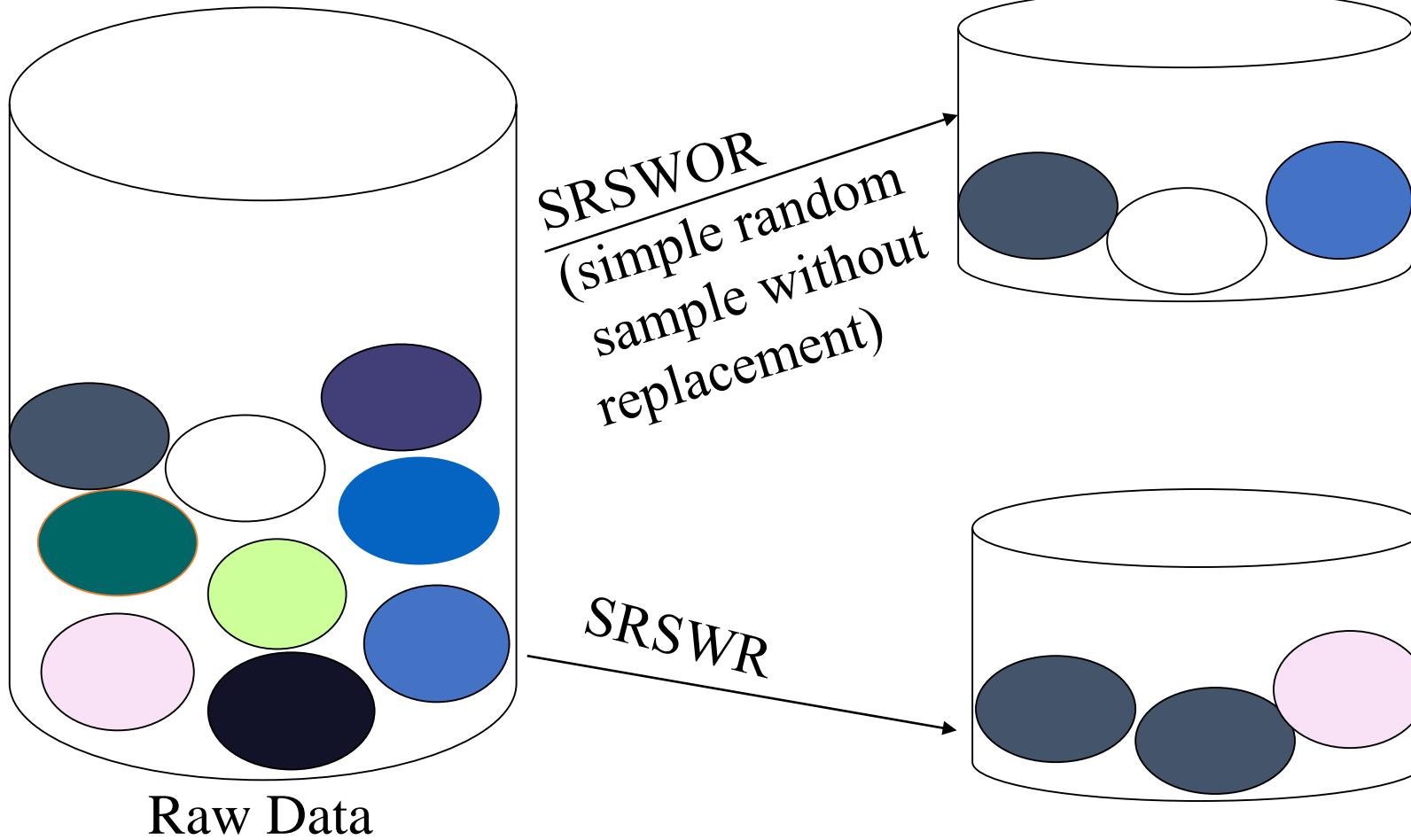
- Identification of target population that is important for a given problem under study
- Decide the sampling frame.
- Determine the sample size
- Sampling method

## Sampling

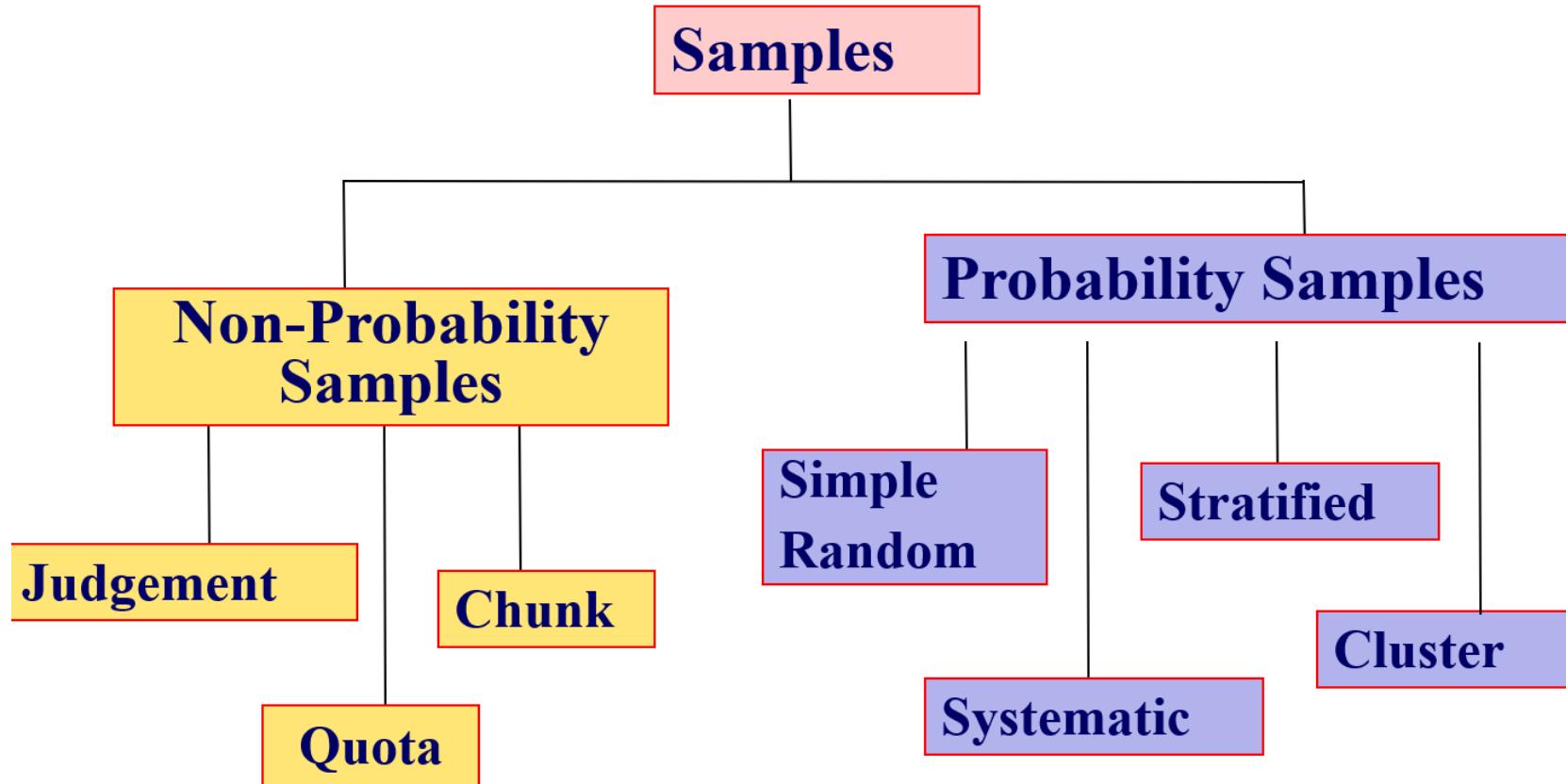
---

- Sampling: obtaining a small sample  $s$  to represent the whole data set  $N$
- Allow an analytics algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a **representative** subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:
- Note: Sampling may not reduce database I/Os (page at a time)

## Sampling: With or without Replacement



## Types of Sampling Methods

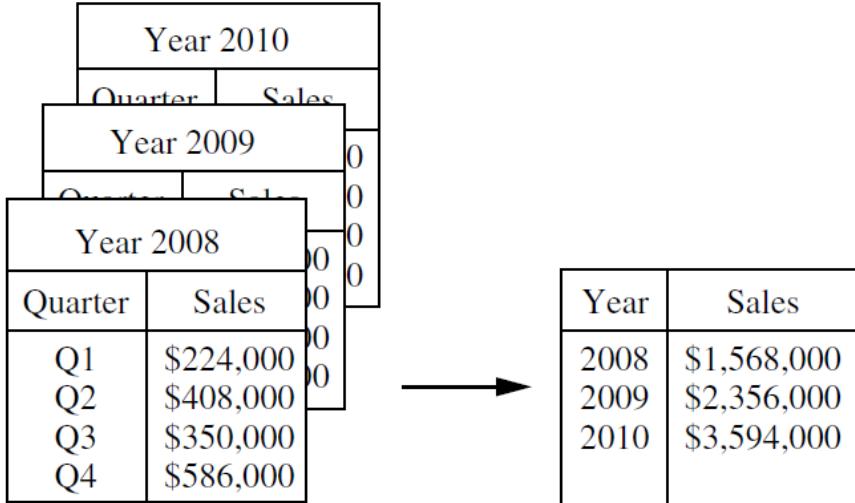


## Data Cube Aggregation

---

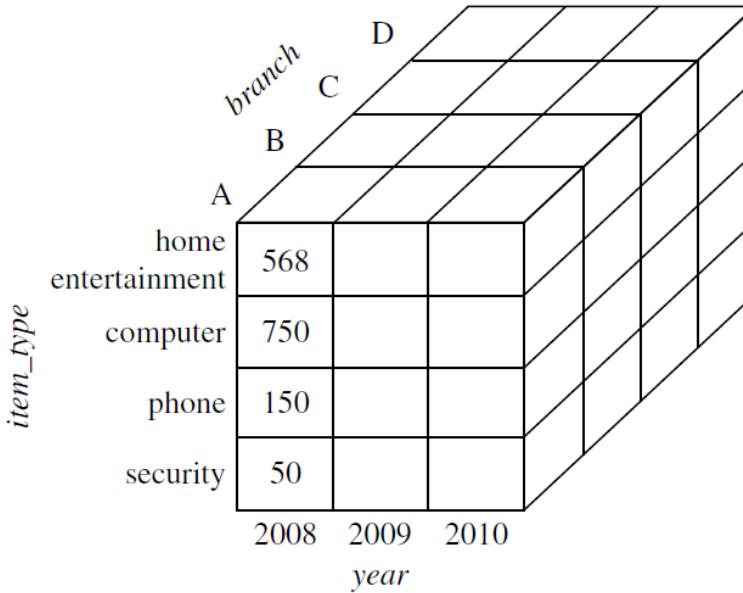
- The lowest level of a data cube (base cuboid)
  - The aggregated data for an **individual entity of interest**
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

## Data Cube Aggregation



The diagram illustrates the process of data cube aggregation. On the left, a 4D data cube is shown with dimensions: Year (2008, 2009, 2010), Quarter (Q1, Q2, Q3, Q4), Sales (\$224,000, \$408,000, \$350,000, \$586,000), and branch (A, B, C, D). An arrow points from this cube to a 2D summary table on the right.

Year	Sales
2008	\$1,568,000
2009	\$2,356,000
2010	\$3,594,000

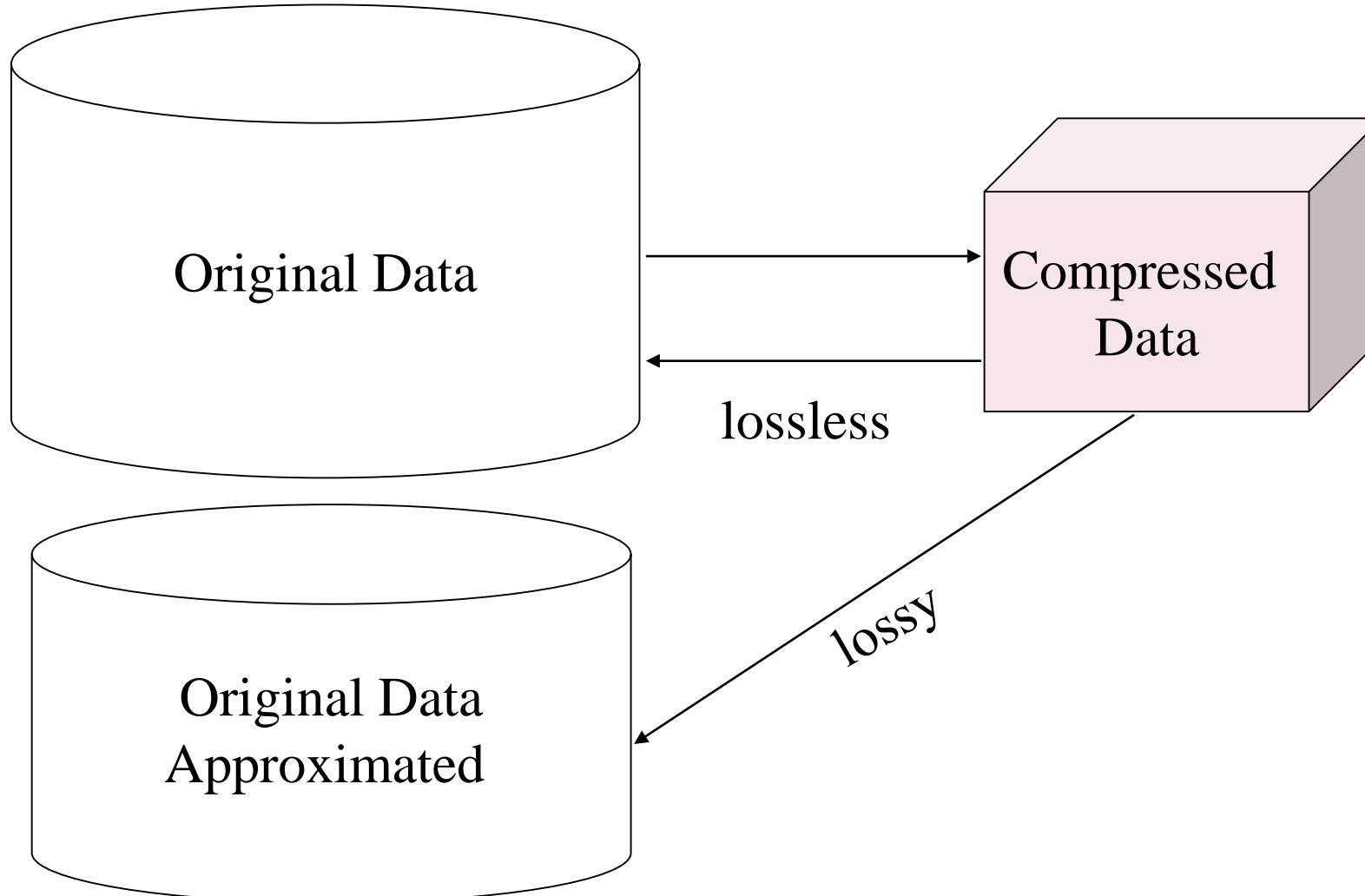


## Data Reduction 3: Data Compression

---

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless, but only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time
- Dimensionality and numerosity reduction may also be considered as forms of data compression

## Data Compression



## Exercise

---

- ❑ Mention and explain the different parametric and non parametric methods used in data reduction.
  
- ❑ Compare and contrast the probability and non probability sampling methods.

## References

---

### Text Book:

- [Data Mining: Concepts and Techniques](#) by Jiawei Han,  
Micheline Kamber and Jian Pei, The Morgan Kaufmann Series  
in Data Management Systems, 3rd Edition.

# DATA ANALYTICS

---

## Unit 1: Data Transformation and Data Discretization

Mamatha H R

Department of Computer Science and Engineering

## Data Transformation

---

A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

## Data Transformation

---

### ■ Methods

- Smoothing: Remove noise from data
  - Simple average or weighted average or Gaussian
- Attribute/feature construction
  - New attributes constructed from the given ones
- Aggregation: Summarization, data cube construction
- Normalization: Scaled to fall within a smaller, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Discretization: Concept hierarchy climbing

## Normalization

---

- **Min-max normalization:** to  $[new\_min_A, new\_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

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

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

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

- **Normalization by decimal scaling**

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

## Discretization

---

- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

## Data Discretization Methods

---

- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split, unsupervised
  - Histogram analysis
    - Top-down split, unsupervised
  - Clustering analysis (unsupervised, top-down split or bottom-up merge)
  - Decision-tree analysis (supervised, top-down split)
  - Correlation (e.g.,  $\chi^2$ ) analysis (unsupervised, bottom-up merge)

## Simple Discretization: Binning

---

- Equal-width (distance) partitioning
  - Divides the range into  $N$  intervals of equal size: uniform grid
  - if  $A$  and  $B$  are the lowest and highest values of the attribute, the width of intervals will be:  $W = (B - A)/N$ .
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well

## Simple Discretization: Binning

---

- Equal-depth (frequency) partitioning
  - Divides the range into  $N$  intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

## Binning Methods for Data Smoothing

---

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- \* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

## Binning Methods for Data Smoothing

---

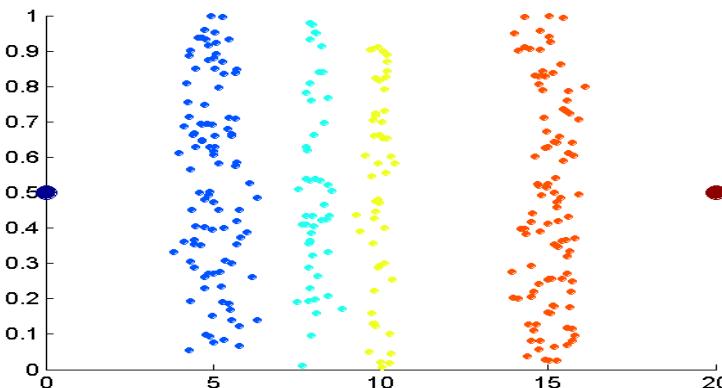
- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

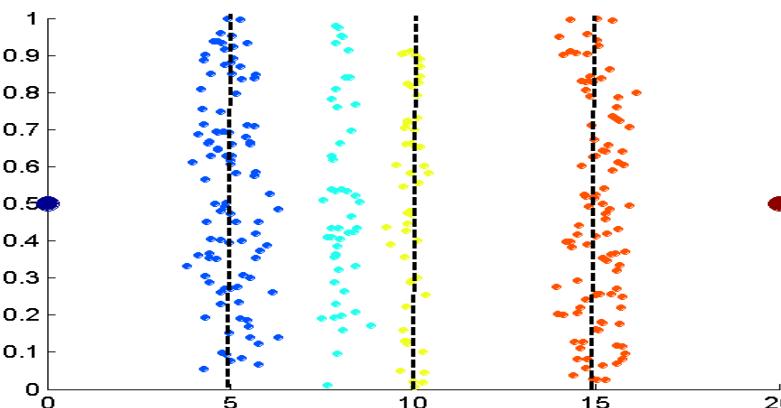
\* Smoothing by **bin boundaries**:

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

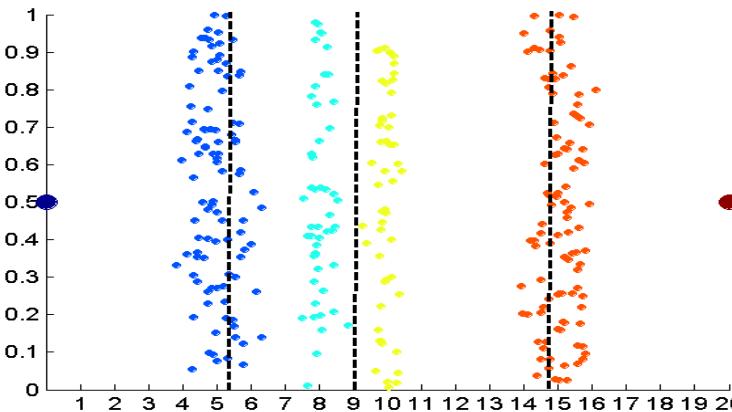
## Discretization Without Using Class Labels (Binning vs. Clustering)



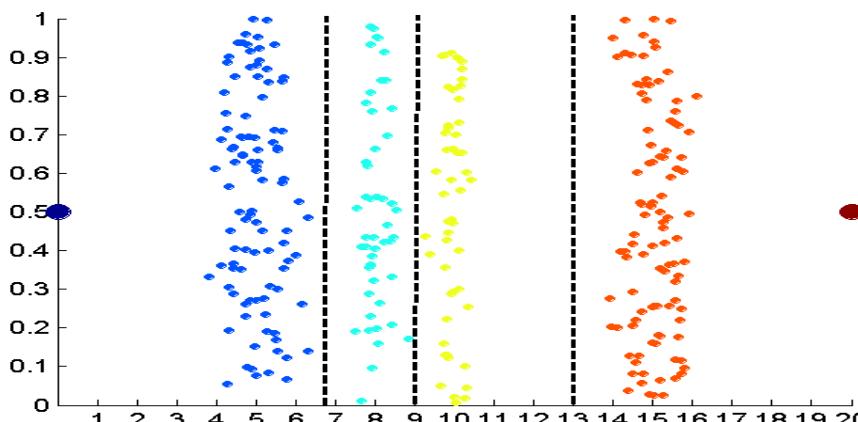
Original Data



Equal Width(binning)



Equal frequency (binning)



K-means clustering leads to better results

## Discretization by Classification &amp; Correlation Analysis

- Classification (e.g., decision tree analysis)

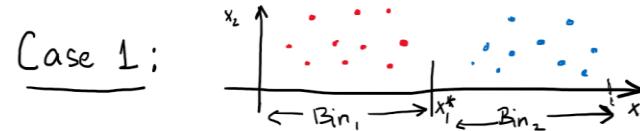
- Supervised: Given class labels, e.g., cancerous vs. benign

- Using *entropy* to determine split point (discretization point)

- Top-down, recursive split

$$\text{Entropy } E = \sum_{i=1}^n P(C_i) \log_2 \left( \frac{1}{P(C_i)} \right), \quad n \text{ is the \# categories}$$

$$= - \sum_{i=1}^n P(C_i) \log_2 (P(C_i))$$

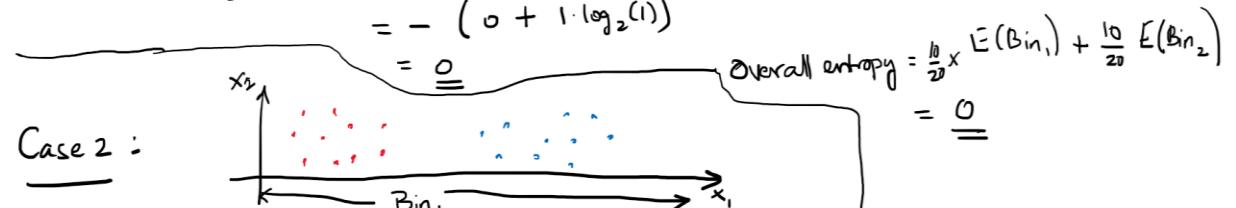


$$\begin{aligned} \text{Entropy}(Bin_1) &= - (P(\text{red}) \log_2 (P(\text{red})) + P(\text{blue}) \log_2 (P(\text{blue}))) \\ &= - (1 \cdot \log_2(1) + 0 \cdot \log_2(0)) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Entropy}(Bin_2) &= - (P(\text{red}) \log_2 (P(\text{red})) + P(\text{blue}) \log_2 (P(\text{blue}))) \\ &= - (0 + 1 \cdot \log_2(1)) \\ &= 0 \end{aligned}$$

Overall entropy =  $\frac{10}{20} E(Bin_1) + \frac{10}{20} E(Bin_2) = 0$

Case 2:



$$\begin{aligned} \text{Entropy}(Bin_1) &= - (P(\text{red}) \log_2 (P(\text{red})) + P(\text{blue}) \log_2 (P(\text{blue}))) \\ &= - (1/2 \log_2(1/2) + 1/2 \log_2(1/2)) \\ &= - (-1/2 - 1/2) \\ &= 1 \end{aligned}$$

Since  $\overset{\text{overall}}{\text{Entropy}}(\text{Case}_1) < \overset{\text{overall}}{\text{Entropy}}(\text{Case}_2)$ ,  $\{(0, x_1^*), (x_1^*, 1)\}$  is a better binning strategy than  $(0, 1)$  as one bin.

## Discretization by Classification & Correlation Analysis

---

- Correlation analysis (e.g., Chi-merge:  $\chi^2$ -based discretization)
  - Supervised: use class information
  - Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low  $\chi^2$  values) to merge
  - Merge performed recursively, until a predefined stopping condition

## Concept Hierarchy Generation

---

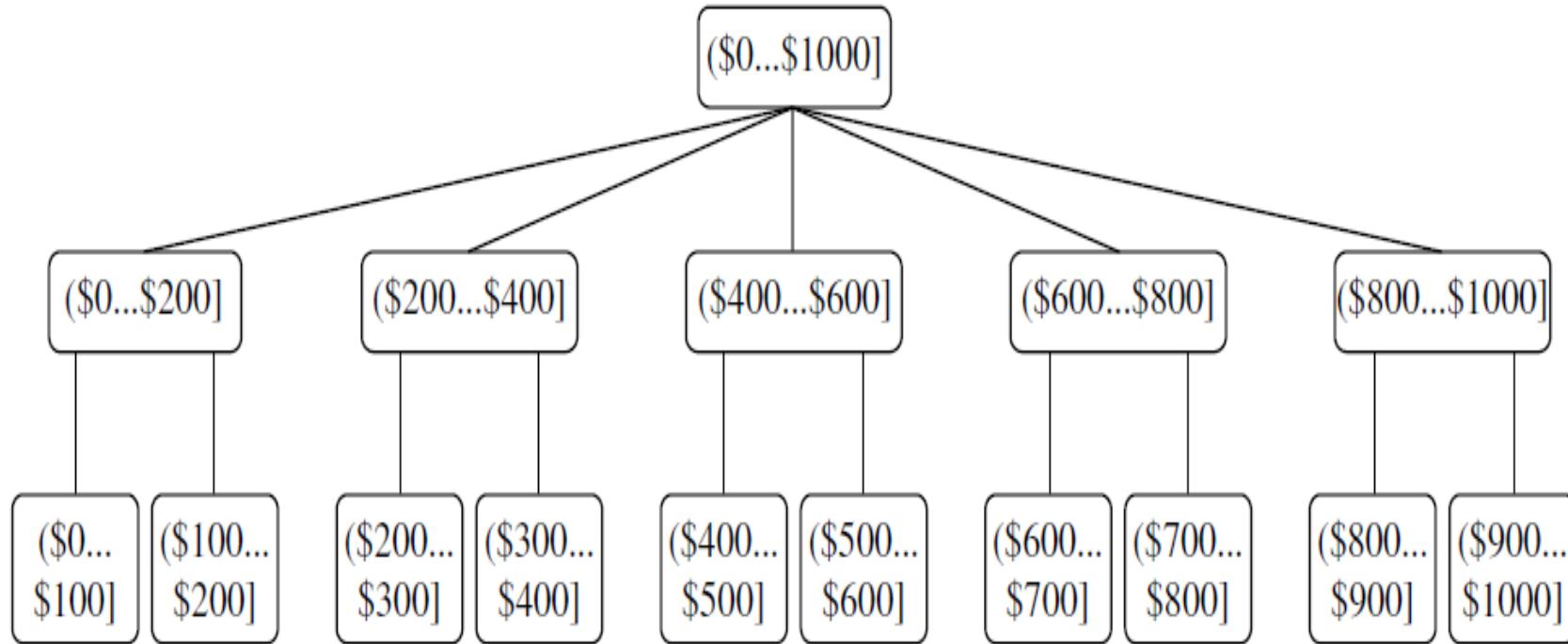
- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity

## Concept Hierarchy Generation

---

- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth*, *adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods shown.

## Concept Hierarchy Generation



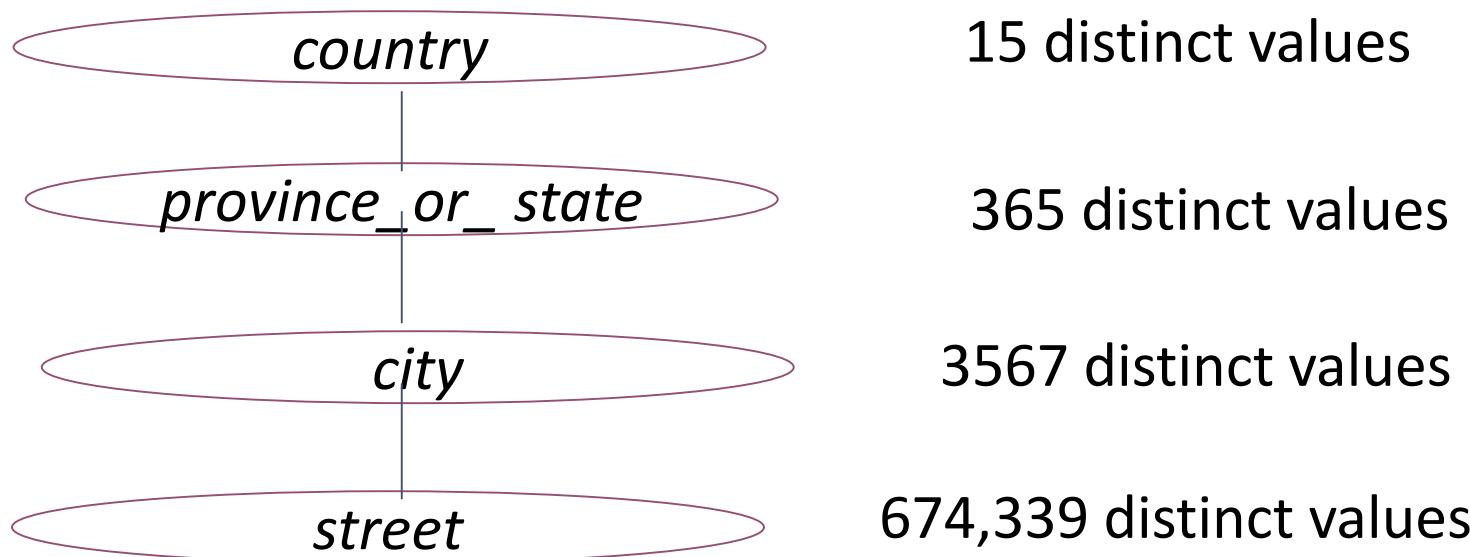
## Concept Hierarchy Generation for Nominal Data

---

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - *street < city < state < country*
- Specification of a hierarchy for a set of values by explicit data grouping
  - $\{\text{Urbana, Champaign, Chicago}\} < \text{Illinois}$
- Specification of only a partial set of attributes
  - E.g., only *street < city*, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes:  $\{\text{street, city, state, country}\}$

## Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



## Exercise

---

- ❑ Mention and explain the different data normalization techniques.
- ❑ How classification and correlation analysis is used in data discretization.

## References

---

### Text Book:

- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.



**THANK YOU**

---

**Dr.Mamatha H R**

Professor,Department of Computer Science

**mamathahr@pes.edu**

+91 80 2672 1983 Extn 834