Data Science

Course code- DAI 101

Shalini Priya
shalinipriyaniki@gmail.com

COURSE DETAILS

Credit: 4
Contact Hours: L: 3 **T:** 1
Mid semester Date: 26/02/2026 – 03/03/2026
End Semester Date: 02/05/2026 – 11/05/2026

Relative weightage
Mid Term Examination: 25 marks
End Term Examination: 50 marks
CWS in MTE : 10 marks
CWS in ETE : 15 marks

S. No.	Contents	Contact hours
I	Introduction to Data Science: Latest and greatest in data science, Real world applications	2
II	Python: Data types (bool, int, float, string, range, list, tuple, set/frozenset, dictionary, string/list operators), Loops (if, elif, else, for, while, for-else, while-else, pass continue, break), Functions (def, lamda, map), Scoping, Modules, File Handling (Reading, Writing, Appending), Graph Plotting (Line/bar graphs, scatter plots), Object Oriented Programming (Abstract data types, Object oriented design, Classes)	6
III	Data Analysis Foundation: Types of data (data matrix, numeric, categorical datasets), Data preparation (data cleaning, data reduction and transformation)	5
IV	Exploratory Data Analysis and Visualization: Univariate and bivariate analysis (Chi-square, t-test, ANOVA, Correlation coefficient), data visualization (scatterplot, segmented bar plot, mosaic plot, contingency table, line chart, pie charts, histogram, box-whisker plot, violin plot, funnel chart, Venn plot, PCA, volcano plot, heatmap)	5
V	Statistical Analysis: Measures of central tendency (arithmetic/geometric mean, median, mode, range, variance, SD, IQR) Confidence Intervals, Hypothesis Testing, Parametric and non-parametric tests (one/two sample tests, Mann-Whitney), p-values, Bias and Variance trade-off	4
VI	Machine Learning: Supervised, unsupervised and reinforcement learning, Linear algebra (transpose, vector-vector product, Euclidean plane/distance, norm, determinant, transformation, Eigenvector, Concave/convex function, matrix calculus), Probability distribution (Binomial, Poisson, Gaussian), Conditional probability (Bayes theorem), Model training, Overfitting and underfitting, Bias and variance, Supervised methods: Linear classification (PLA, pocket algorithm), Linear regression, Gradient descent, Hyperparameter optimization, Logistic regression, Decision trees, SVM, Clustering, K-means, PCA	12
VII	Deep learning and Big Data: Gradient Descent, Neural nets, Convolutional Neural Networks, Big Data technologies (MapReduce, HDFS)	8
Total		42

References

Name of Authors/Books/ Publisher Year of Publication/Reprint

OUTLINE

- Data, Big Data and Challenges
- Data Science
  - Introduction
  - Why Data Science
- Data Scientists
  - What do they do?
- Major/Concentration in Data Science
  - What courses to take.

DATA ALL AROUND

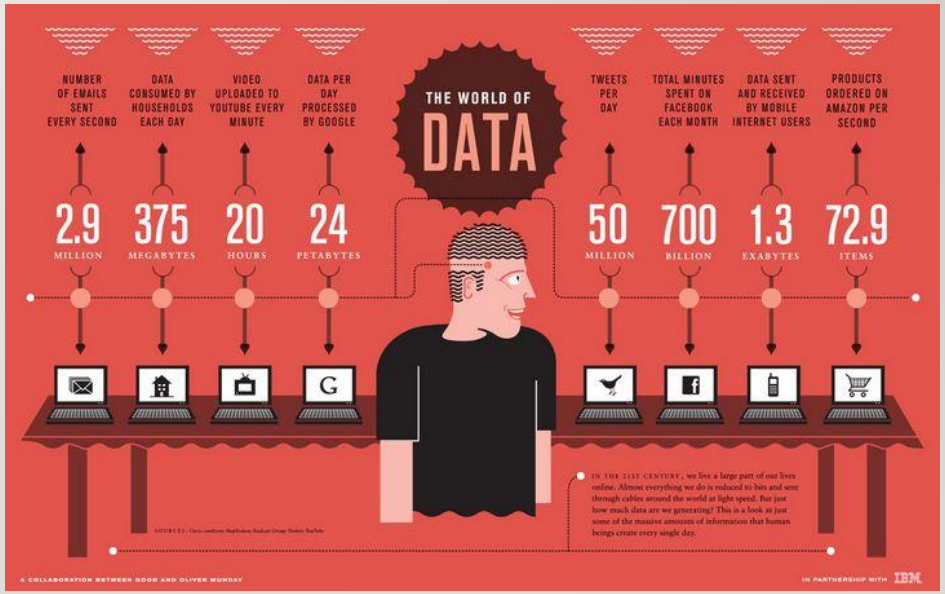
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network

HOW MUCH DATA DO WE HAVE?

- Google processes 20 PB a day (2008-2020)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB

- Cost of 1 TB of disk: $35
- Time to read 1 TB disk: 3 hrs

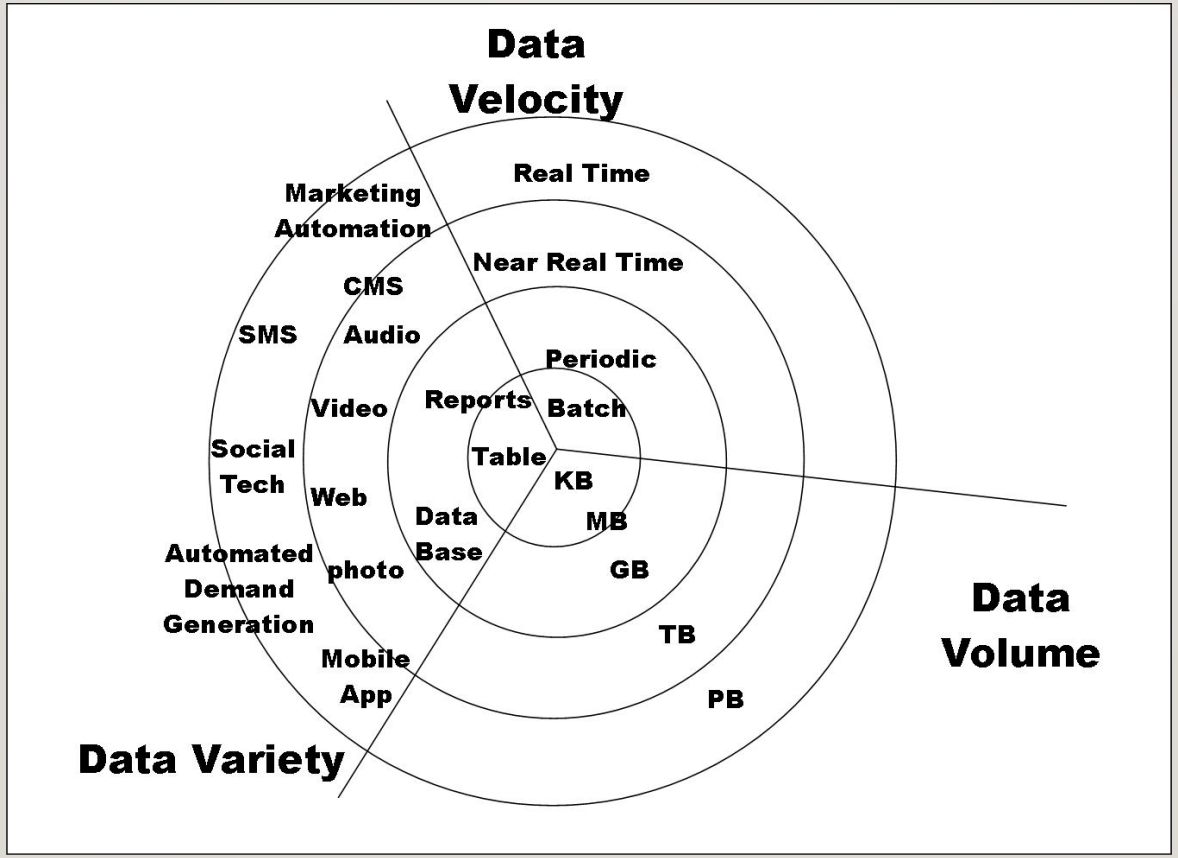
  (100 MB/s)



BIG DATA

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

BIG DATA


Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Table
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Social
Tech
Web
Data
Base
KB
MB
GB
TB
PB
Automated
Demand
Generation
photo
Mobile
App
Data
Volume
Data Variety

TYPES OF DATA WE HAVE

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data

WHAT TO DO WITH THESE DATA?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

WHAT IS DATA SCIENCE?

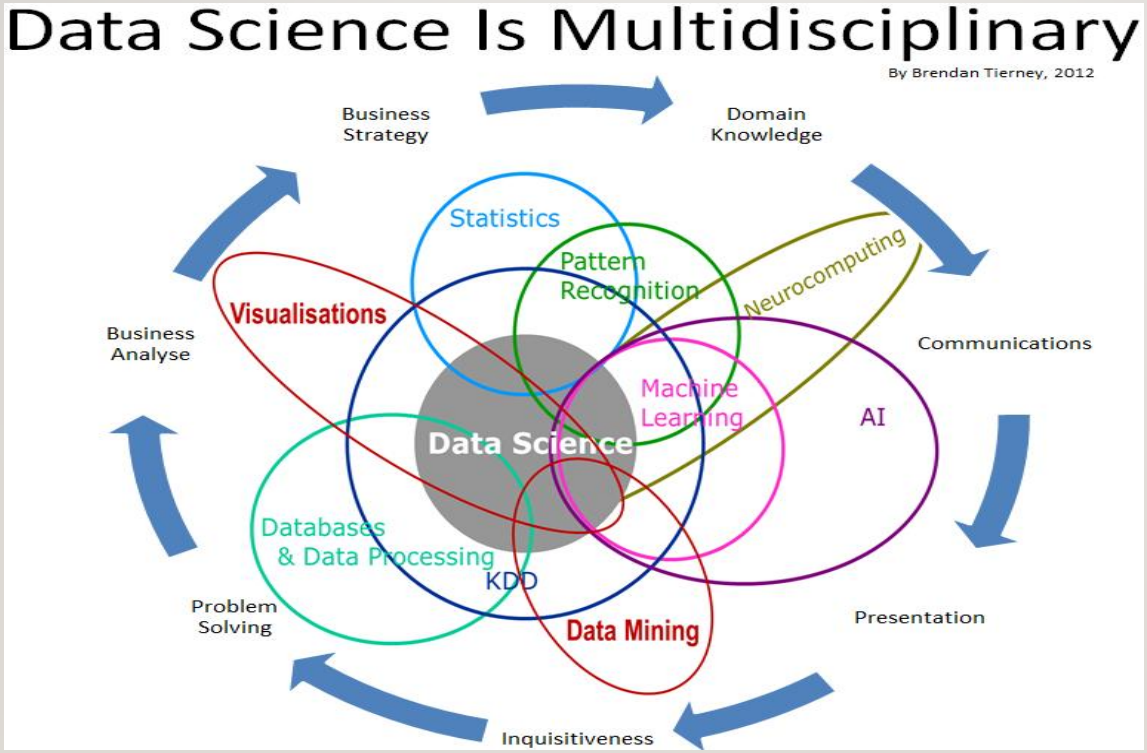
- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

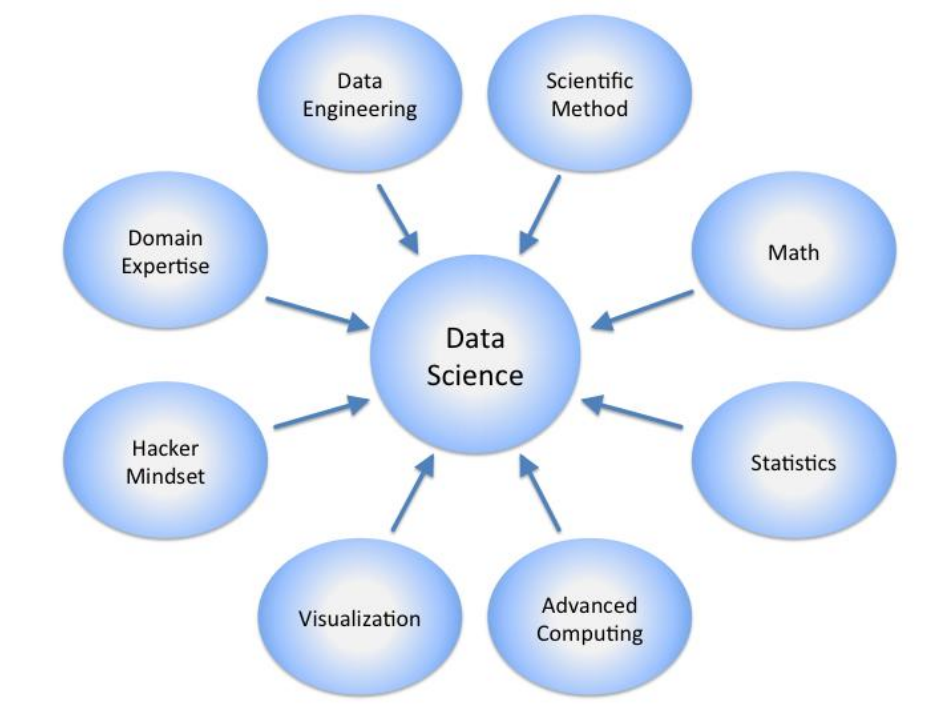
WHAT IS DATA SCIENCE?

- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

DATA SCIENCE



DATA SCIENCE



REAL LIFE EXAMPLES

- Companies learn your secrets, shopping patterns, and preferences
  - For example, can we know if a woman is pregnant, even if she doesn't want us to know? Target case study
- Data Science and election (2008, 2012)
  - 1 million people installed the Obama Facebook app that gave access to info on "friends"

DATA SCIENTISTS

- Data Scientist
  - Most demanding Job of the 21st Century
- They find stories, extract knowledge. They are not reporters

DATA SCIENTISTS

- Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions

WHAT DO DATA SCIENTISTS DO?

- National Security
- Cyber Security
- Business Analytics
- Engineering
- Healthcare
- And more ….

CONCENTRATION IN DATA SCIENCE

- Mathematics and Applied Mathematics
- Applied Statistics/Data Analysis
- Solid Programming Skills (R, Python, Julia, SQL)
- Data Mining
- Data Base Storage and Management
- Machine Learning and discovery

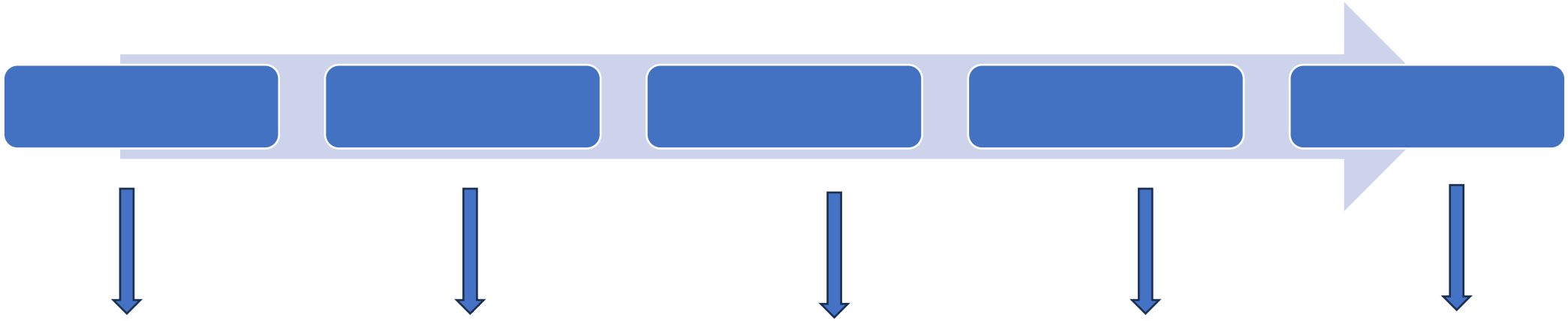
End of Lecture 1

What is Data Analysis

- What is Data
  - Unprocessed Information
- Why is it important
  - Meaning information

- Data analysis is a process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, informing conclusions and supporting decision-making.

Data Analysis Process



Asking a right question

Data wrangling/munging /pre-processing

Exploratory Data Analysis

Data Conclusion

Communicating Results

Step 1: Asking Questions

- What features will contribute to my analysis.
- What features are not important for my analysis.
- Which of the features have a strong correlation.
- Do I need data pre-processing ?
- What kind of feature manipulation/engineering is required.

How can I ask better questions?

Subject Matter Expertise

Experience

Data Reading

Types of data

Step 2: Data Wrangling/Munging

- Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one “raw” data from into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

- Gathering data
- Assessing Data
- Cleaning Data

2a: Gathering Data

- CSV FILES
- API
- WEB SCRAPING
- DATABASES

2b: Assessing Data

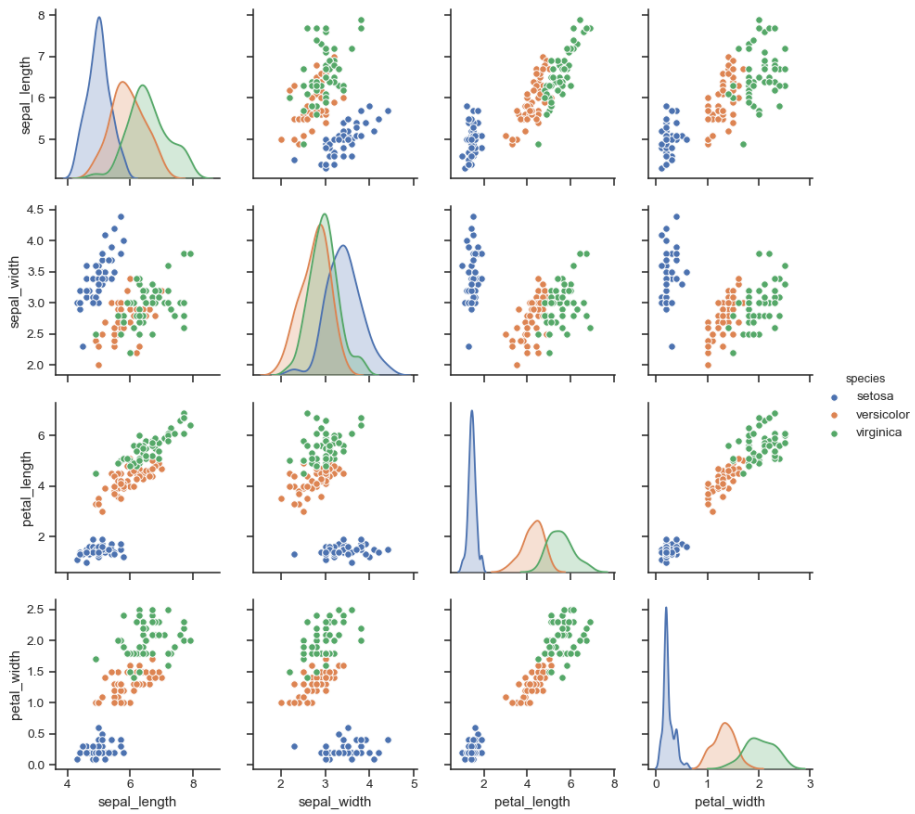
- Finding the number of rows/columns (shape)
- Data types of various columns (info())
- Checking for missing values (info())
- Check for duplicate data (is_unique)
- Memory occupied by the dataset (info)
- High level mathematical overview of the data (describe)

2c: Cleaning Data

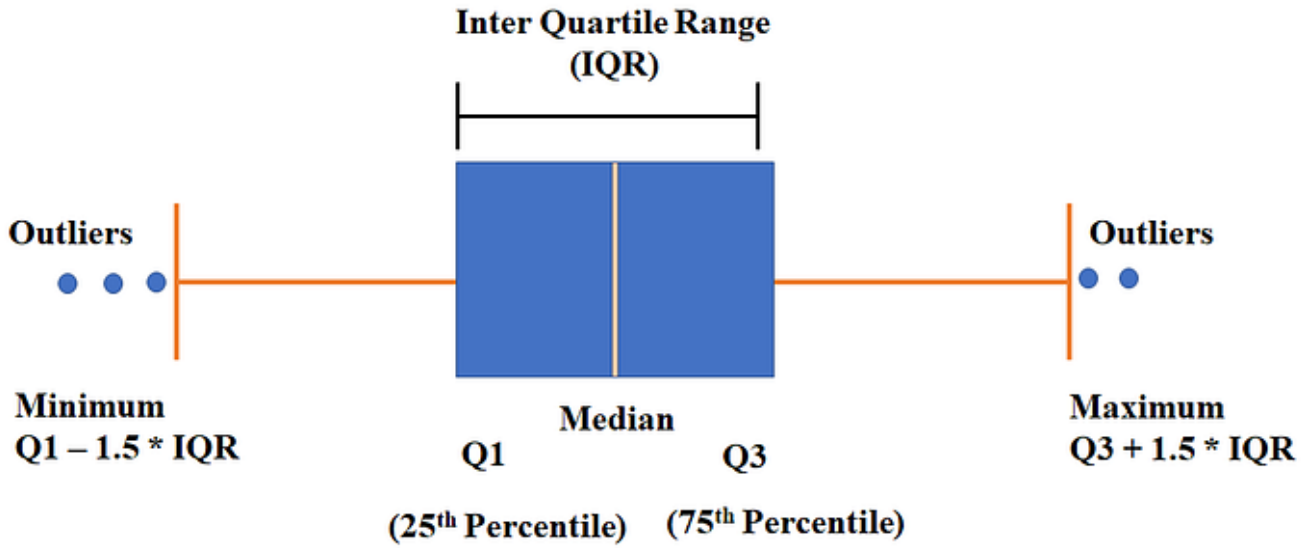
- Missing Data (e.g. mean)
- Remove duplicate data (drop duplicates)
- Incorrect data type (astype)

Step 3 : Exploratory Data Analysis

- (a) Explore
  - Finding correlation and covariance
  - Doing univariate and multivariate analysis
  - Plotting graphs (data visualization)



3b: Augmenting Data



THE ASSIGN METHOD ADDS NEW VARIABLES TO A DATAFRAME

sales_data.assign()

name	region	sales	expense	profit
William	East	50000	42000	8000
Emma	North	52000	43000	9000
Sofia	East	90000	50000	40000
Markus	South	34000	44000	-10000
Edward	West	42000	38000	4000

left

	A	B	key2
K0	A0	B0	K0
K0	A1	B1	K1
K1	A2	B2	K0
K2	A3	B3	K1

right

	C	D	key2
K0	C0	D0	K0
K1	C1	D1	K0
K2	C2	D2	K0
K2	C3	D3	K1

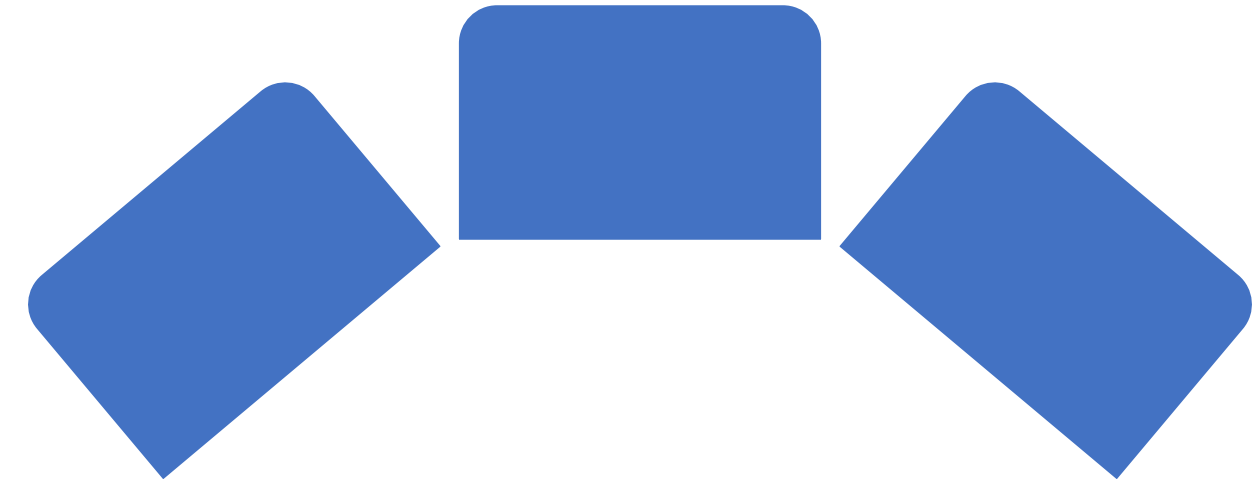
right2

	v
K1	7
K1	8
K2	9

Result

	A	B	key2_x	C	D	key2_y	v
K0	A0	B0	K0	C0	D0	K0	NaN
K0	A1	B1	K1	C0	D0	K0	NaN
K1	A2	B2	K0	C1	D1	K0	7.0
K1	A2	B2	K0	C1	D1	K0	8.0
K2	A3	B3	K1	C2	D2	K0	9.0
K2	A3	B3	K1	C3	D3	K1	9.0

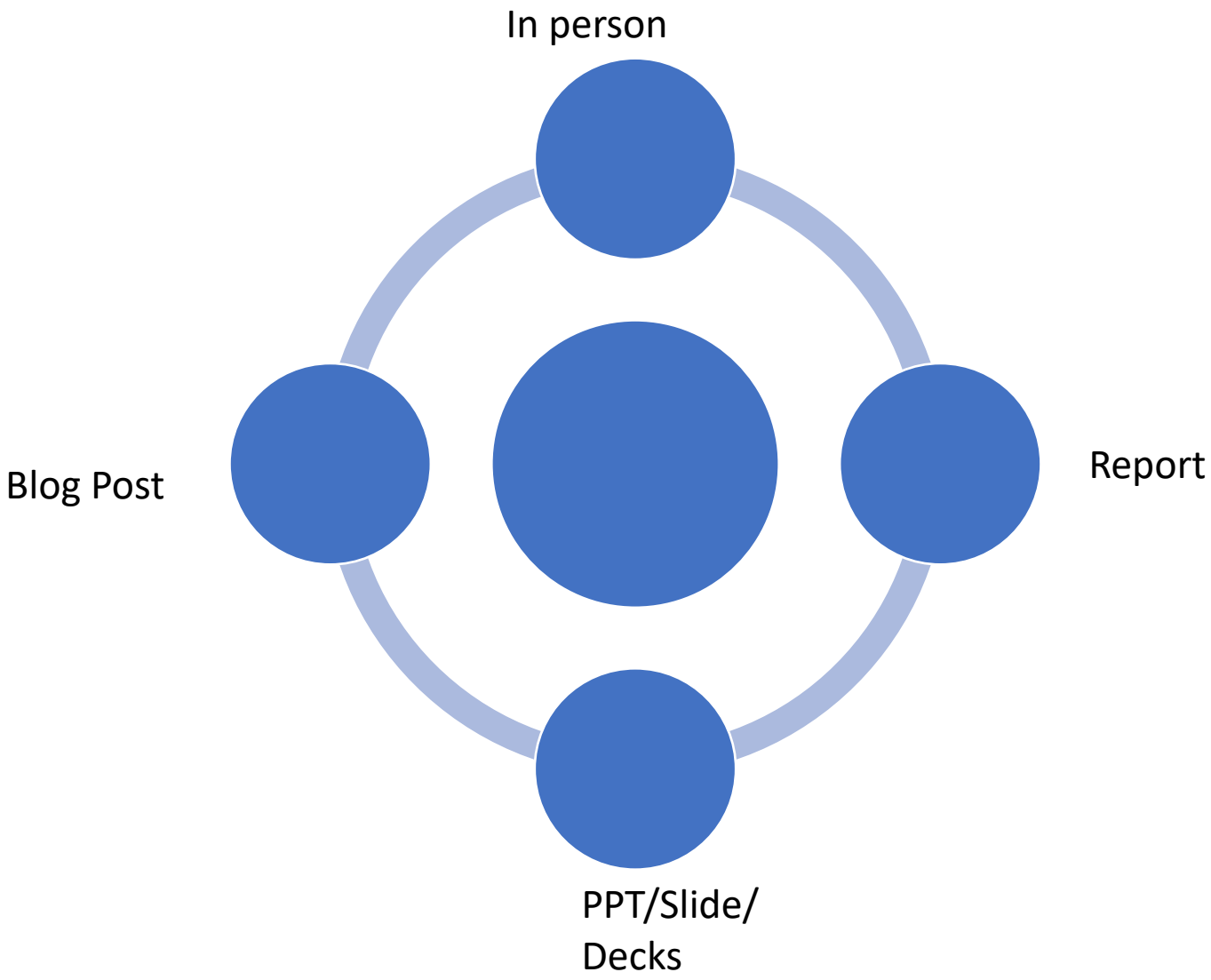
Step 4: Drawing Conclusions



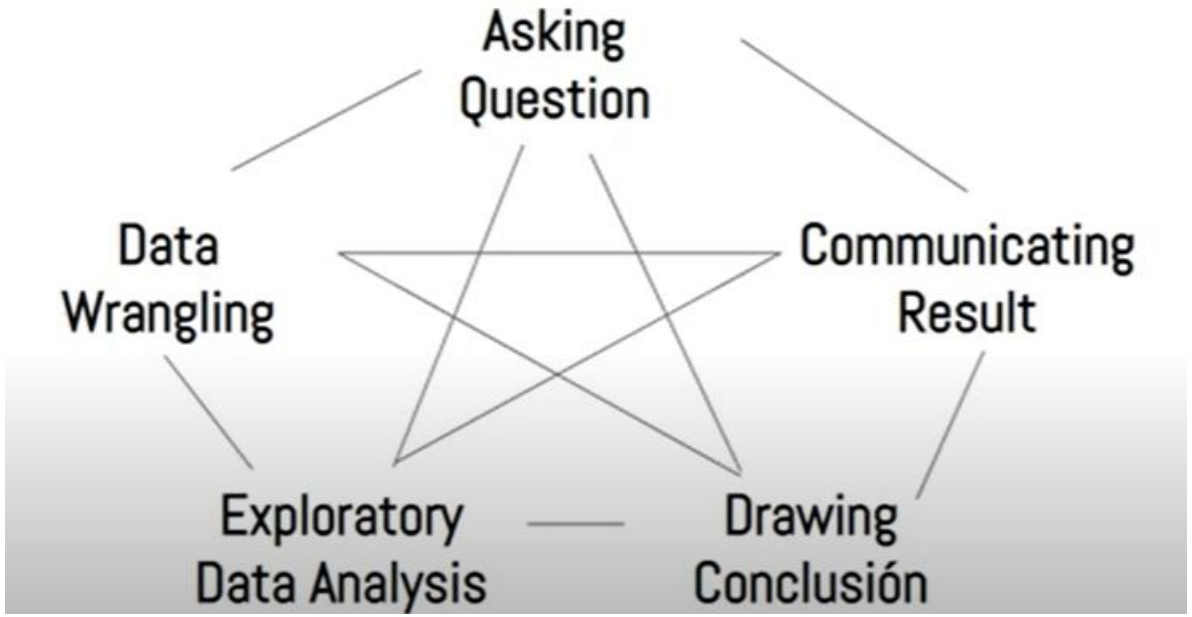
Some example conclusion based on Descriptive Statistics

1. Is Rohit Sharma a better batsman a better batsman in 2nd innings (IPL Dataset) ?
2. Does being a female increases your chances of Survival (Titanic Dataset) ?
3. Is Delhi the most costly place for eating out (Zomato Dataset) ?

Step5: Communicating Results/ Data Stealing



The fun part


Asking
Question
Data
Wrangling
Communicating
Result
Exploratory
Data Analysis
Drawing
Conclusión

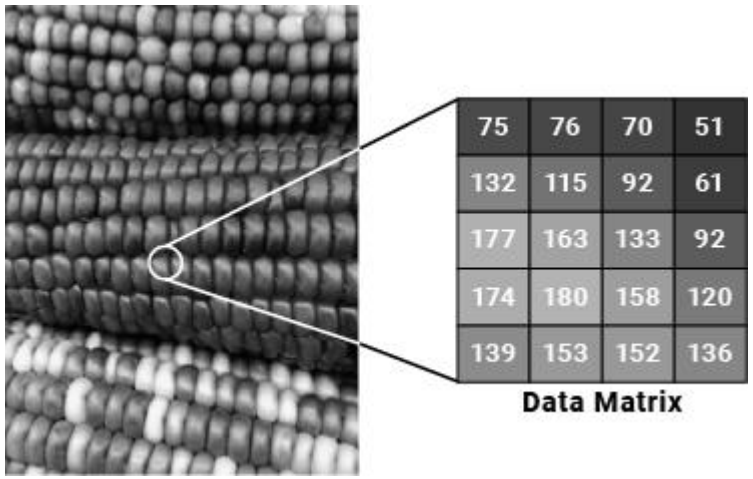
Matrix theory and linear algebra

- Matrix can be used to represent samples with multiple attributes in a compact form.
- Matrix can also be used to represent linear equations in a compact and simple fashion.
- Linear algebra provides tools to understand and manipulate matrics to drive useful knowledge from data

Matrices for data science: Data representation

- Usually matrices are used to store and represent the data on machines
- Matrix is a very natural approach for organizing data
- In general, data is organized in the following fashion.
    - Rows represent samples
    - Columns represent the values of the variables (or attributes)
    - It is also possible to use rows for variable and columns for samples
    - However, we will stick to rows as samples and columns as variables in all of the material that will be presented.

Data representation: Examples



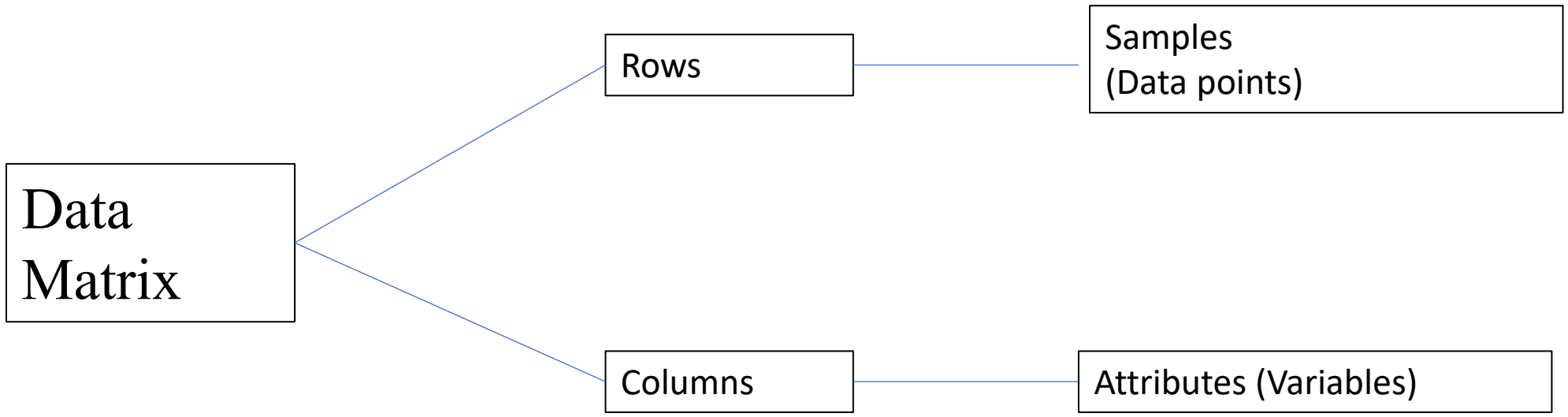
Storing

- The image is stored in the machine as a large matrix of pixel values across the image.
- Thus, storing the pixel value matrix is equivalent to storing the image for the machine.

Identification

- Several machine learning algorithms are deployed in order to teach the machine how to identify a particular image.
- Linear algebra and matrix operations are at the heart of these machine learning.

Data as matrix: Summary



Data Matrix

Data can often be represented or abstracted as an $n \times d$ *data matrix*, with n rows and d columns, given as

$$\boldsymbol{D} = \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \boldsymbol{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \boldsymbol{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

- **Rows**: Also called *instances*, *examples*, *records*, *transactions*, *objects*, *points*, *feature-vectors*, etc. Given as a d-tuple

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

- **Columns**: Also called *attributes*, *properties*, *features*, *dimensions*, *variables*, *fields*, etc. Given as an n-tuple

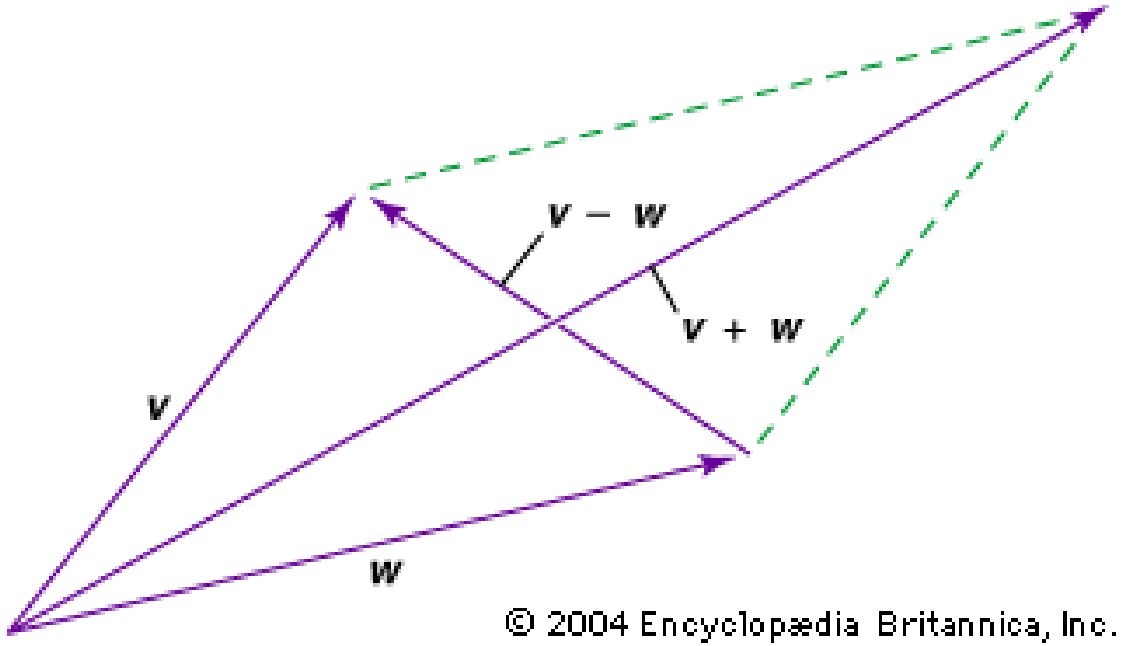
$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Vector and Operations

Row vector
Column vectors

Addition
Multiplication

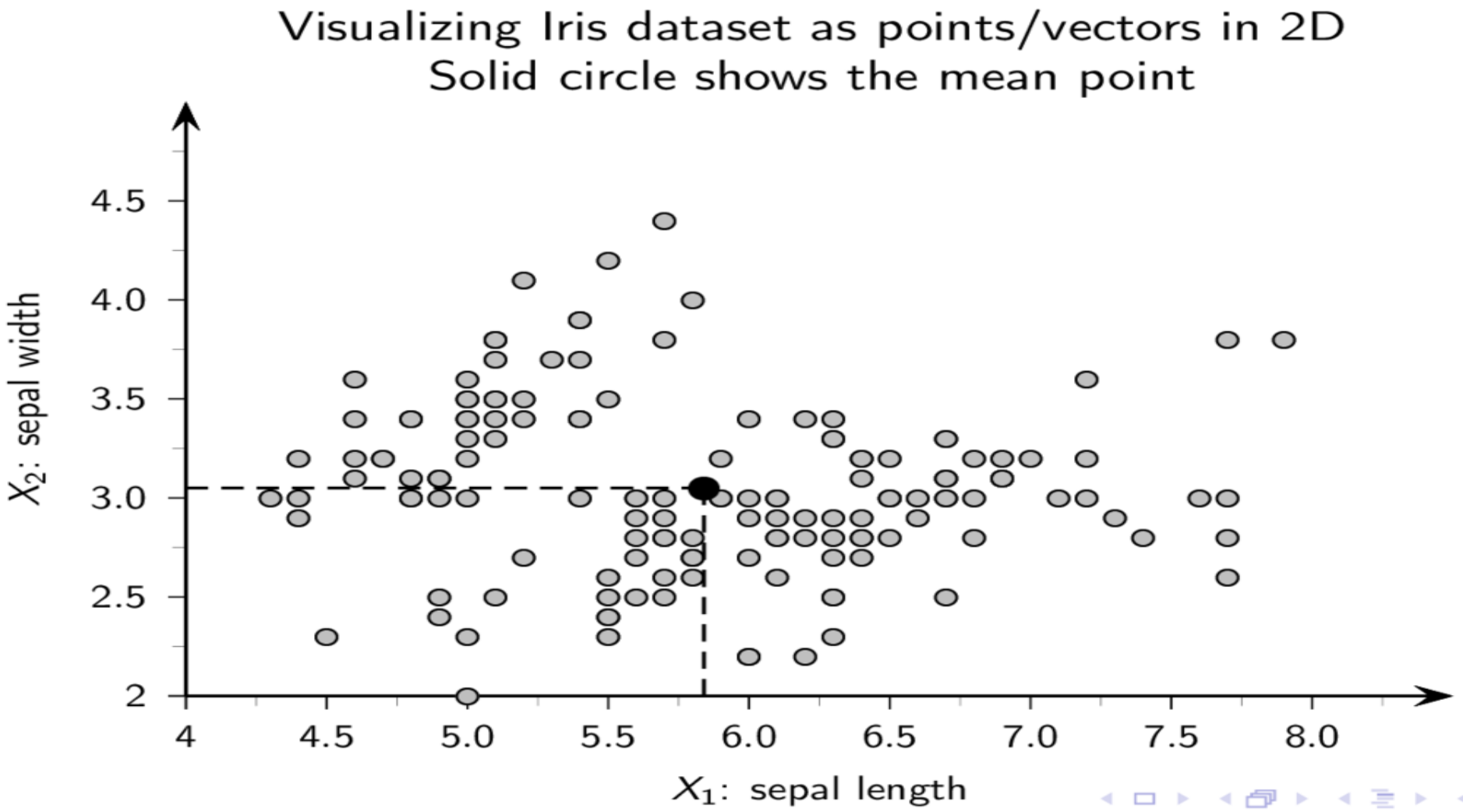
Mean centric



Iris Dataset Extract

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\boldsymbol{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\boldsymbol{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\boldsymbol{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\boldsymbol{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\boldsymbol{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\boldsymbol{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\boldsymbol{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\boldsymbol{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\boldsymbol{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Scatterplot: 2D Iris Dataset
sepal length versus sepal width.



Norm, Distance and Angle

Given two points $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, their *dot product* is defined as the scalar

$$\boldsymbol{a}^T \boldsymbol{b} = a_1 b_1 + a_2 b_2 + \cdots + a_m b_m = \sum_{i=1}^{m} a_i b_i$$

The *Euclidean norm* or *length* of a vector $\boldsymbol{a}$ is defined as

$$\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^T \boldsymbol{a}} = \sqrt{\sum_{i=1}^{m} a_i^2}$$

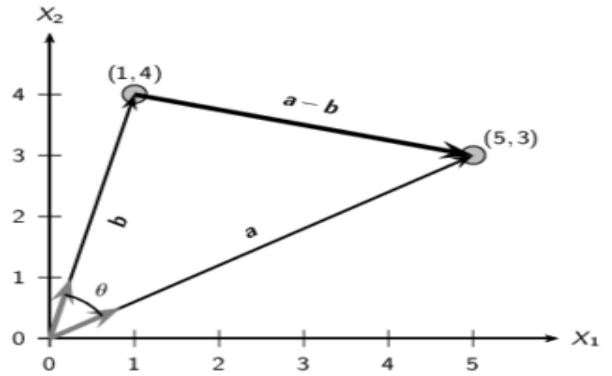
The *unit vector* in the direction of $\boldsymbol{a}$ is $\boldsymbol{u} = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}$ with $\|\boldsymbol{a}\| = 1$.

Distance between $\boldsymbol{a}$ and $\boldsymbol{b}$ is given as

$$\|\boldsymbol{a} - \boldsymbol{b}\| = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2}$$

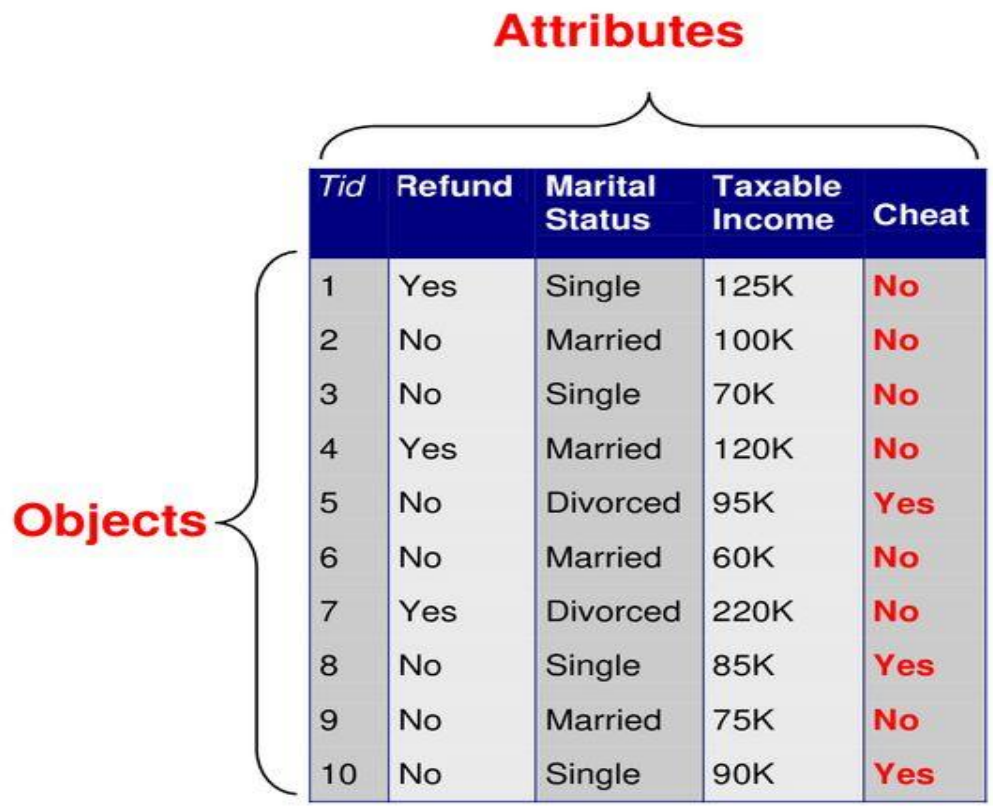
Angle between $\boldsymbol{a}$ and $\boldsymbol{b}$ is given as

$$\cos\theta = \frac{\boldsymbol{a}^T \boldsymbol{b}}{\|\boldsymbol{a}\| \, \|\boldsymbol{b}\|} = \left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}\right)^T \left(\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right)$$



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, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance



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

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- 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
      - ID has no limit but age has a maximum and minimum value

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.
  - Different evaluation measures are used in conjunction with continuous attributes

Types of data sets

The remainder of the slides in this file will not be discussed in Fall 2019!

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Important Characteristics of Structured Data

- Dimensionality
  - Curse of Dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale

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 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

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	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

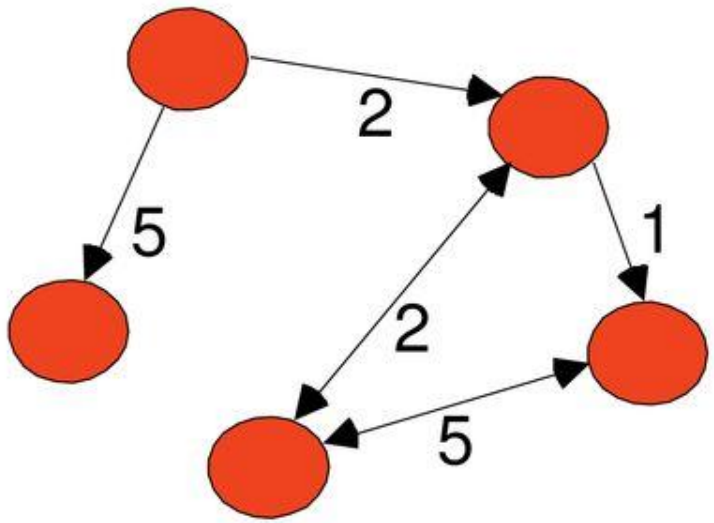
Transaction Data

- 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.

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

Graph Data

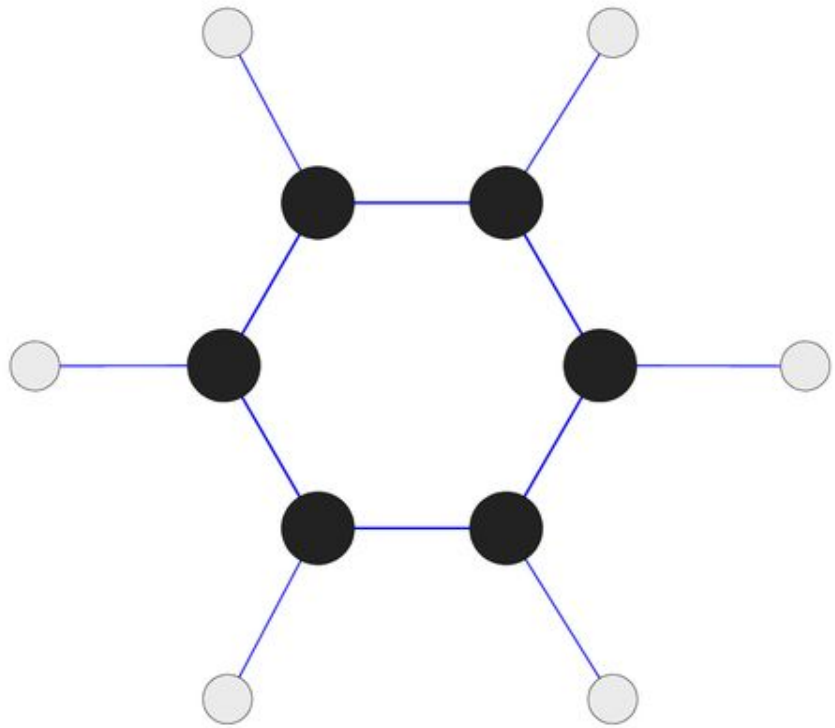
- Examples: Generic graph and HTML Links




```
<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
```

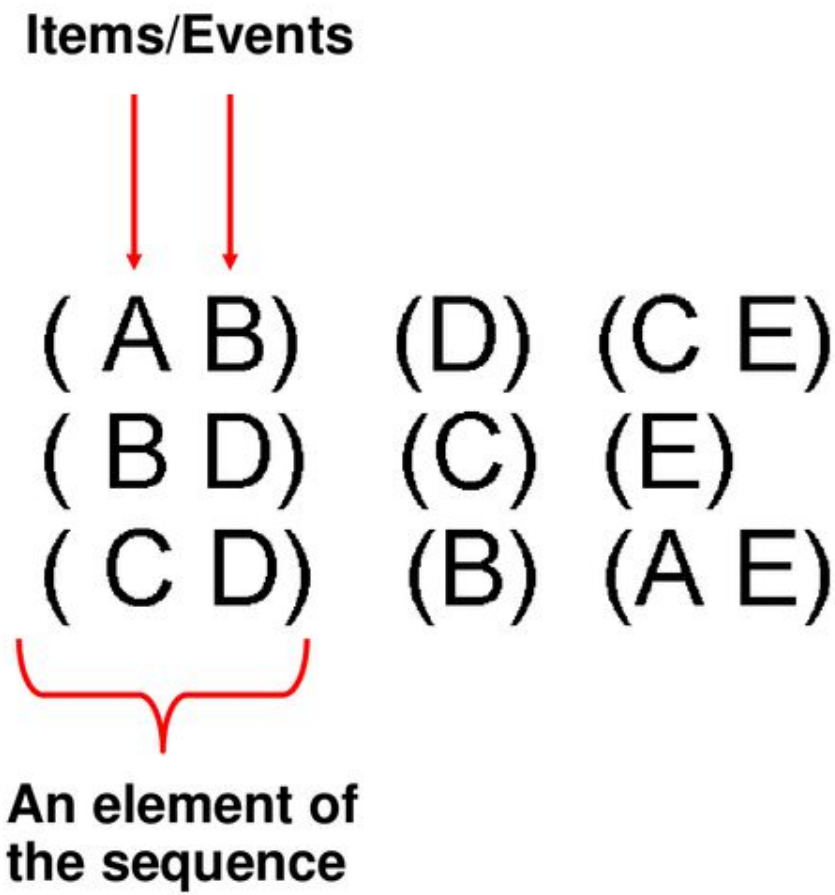
Chemical Data

- Benzene Molecule: C_6H_6



Ordered Data

- Sequences of transactions



Ordered Data

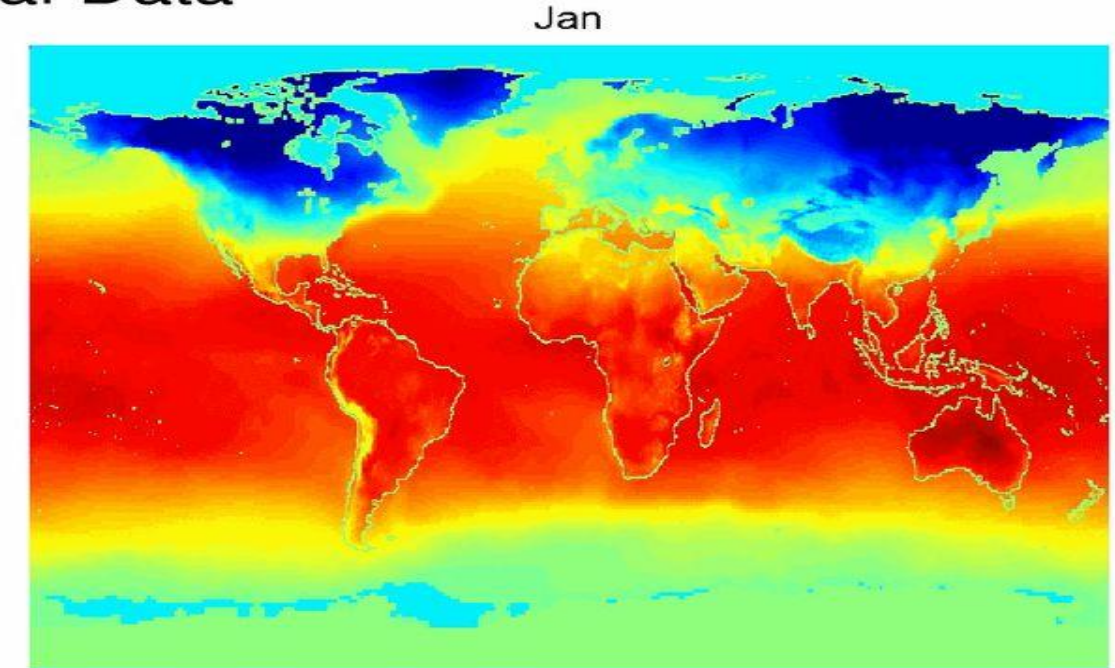
- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Reading data

Type of file formats

Common file formats, such as **CSV, XLSX, ZIP, TXT** etc.

CSV: the CSV is stand for Comma-separated values. as-well-as this name CSV file is use comma to separated values. In CSV file each line is a data record and Each record consists of one or more than one data fields, the field is separated by commas.

XLSX: The XLSX file is Microsoft Excel Open XML Format Spreadsheet file. This is used to store any type of data but it's mainly used to store financial data and to create mathematical models etc.

ZIP: ZIP files are used an data containers, they store one or more than one files in the compressed form. it widely used in internet After you downloaded ZIP file, you need to unpack its contents in order to use it.

TXT: TXT files are useful for storing information in plain text with no special formatting beyond basic fonts and font styles. It is recognized by any text editing and other software programs

JSON: JSON is stand for JavaScript Object Notation. JSON is a standard text-based format for representing structured data based on JavaScript object syntax

HTML: HTML is stand for stands for Hyper Text Markup Language is use for creating web pages. we can read html table in python pandas using read_html() function.

Below are the methods by which we can read text files with Pandas:

- Using read_csv()
- Using read_table()
- Using read_fwf()

CSV

	A	B	C	D
1	ID	Gender	City	Monthly_I
2	ID000002C	Female	Delhi	20000
3	ID000004E	Male	Mumbai	35000
4	ID000007H	Male	Panchkula	22500
5	ID000008I	Male	Saharsa	35000
6	ID000009J	Male	Bengaluru	100000
7	ID000010K	Male	Bengaluru	45000
8	ID000011L	Female	Sindhudu	70000
9	ID000012M	Male	Bengaluru	20000
10	ID000013N	Male	Kochi	75000
11	ID000014C	Female	Mumbai	30000
12	ID000016C	Male	Mumbai	25000
13	ID000018S	Female	Surat	25000
14	ID000019T	Female	Pune	24000
15	ID000021V	Male	Bhubanes	27000
16	ID000022V	Female	Howrah	28000

JSON

```
{
   "Employee": [

      {

         "id":"1",

         "Name": "Ankit",

         "Sal": "1000",

      },
      {

         "id":"2",

         "Name": "Faizy",
```

```
<?xml version="1.0"?>

<contact-info>

<name>Ankit</name>

<company>Anlytics Vidhya</company>

<phone>+9187654321</phone>

</contact-info>
```

CHAPTER 3: DATA PREPROCESSING

- DATA PREPROCESSING: AN OVERVIEW 
  - DATA QUALITY
  - MAJOR TASKS IN DATA PREPROCESSING
- DATA CLEANING
- DATA INTEGRATION
- DATA REDUCTION
- DATA TRANSFORMATION AND DATA DISCRETIZATION
- SUMMARY

42

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?

MAJOR TASKS IN DATA PREPROCESSING

- **DATA CLEANING**
  - FILL IN MISSING VALUES, SMOOTH NOISY DATA, IDENTIFY OR REMOVE OUTLIERS, AND RESOLVE INCONSISTENCIES
- **DATA INTEGRATION**
  - INTEGRATION OF MULTIPLE DATABASES, DATA CUBES, OR FILES
- **DATA REDUCTION**
  - DIMENSIONALITY REDUCTION
  - NUMEROSITY REDUCTION
  - DATA COMPRESSION
- **DATA TRANSFORMATION AND DATA DISCRETIZATION**
  - NORMALIZATION
  - CONCEPT HIERARCHY GENERATION

CHAPTER 3: DATA PREPROCESSING

- DATA PREPROCESSING: AN OVERVIEW
  - DATA QUALITY
  - MAJOR TASKS IN DATA PREPROCESSING
- DATA CLEANING
- DATA INTEGRATION
- DATA REDUCTION
- DATA TRANSFORMATION AND DATA DISCRETIZATION
- SUMMARY

45

DATA CLEANING

- DATA IN THE REAL WORLD IS DIRTY: LOTS OF POTENTIALLY INCORRECT DATA, E.G., INSTRUMENT FAULTY, HUMAN OR COMPUTER ERROR, TRANSMISSION ERROR
  - INCOMPLETE: LACKING ATTRIBUTE VALUES, LACKING CERTAIN ATTRIBUTES OF INTEREST, OR CONTAINING ONLY AGGREGATE DATA
    - E.G., *OCCUPATION*=" " (MISSING DATA)
  - NOISY: CONTAINING NOISE, ERRORS, OR OUTLIERS
    - E.G., *SALARY*="−10" (AN ERROR)
  - INCONSISTENT: CONTAINING DISCREPANCIES IN CODES OR NAMES, E.G.,
    - *AGE*="42", *BIRTHDAY*="03/07/2010"
    - WAS RATING "1, 2, 3", NOW RATING "A, B, C"
    - DISCREPANCY BETWEEN DUPLICATE RECORDS
  - INTENTIONAL (E.G., *DISGUISED MISSING* DATA)
    - JAN. 1 AS EVERYONE'S BIRTHDAY?

Challenges in Data Cleaning:

- **Volume of Data:** Large datasets can be challenging to clean due to their sheer size.
- **Complexity of Data:** Data from diverse sources may have different structures and formats,
- **Continuous Process:** Data cleaning is not a one-time task but an ongoing process..

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

Strategies for Missing Data

1. Deletion Methods (Removal)

- **Listwise/Case Deletion:** Drop rows (cases) with any missing values. Simple but can cause significant data loss and bias if data isn't MCAR.
- **Column Deletion:** Drop entire columns if they have a very high percentage (e.g., >80%) of missing values

2. Imputation Methods (Filling)

- **Simple Imputation:**
  - **Mean/Median:** Fill numerical missing values with the column's mean (sensitive to outliers) or median (more robust).
  - **Mode:** Fill categorical missing values with the most frequent category (mode).
  - **Constant Value:** Replace with a specific number (e.g., 0, -1, 999) or a new category for categorical data.
- **Time-Series Imputation:**
  - **Forward Fill :** Use the last known value.
  - **Backward Fill :** Use the next known value.

- **MODEL-BASED IMPUTATION:**
  - **REGRESSION IMPUTATION:** PREDICT MISSING VALUES USING OTHER FEATURES IN A REGRESSION MODEL.
  - **KNN IMPUTATION:** USE VALUES FROM THE K-NEAREST NEIGHBORS IN THE DATASET TO ESTIMATE THE MISSING VALUE.
  - **MULTIPLE IMPUTATION (MICE):** MORE ADVANCED, CREATES MULTIPLE IMPUTED DATASETS TO ACCOUNT FOR UNCERTAINTY.

Key Considerations Before Choosing

- **Proportion of Missing Data:** Small amounts might allow deletion; large amounts need imputation or dropping columns.
- **Type of Data:** Numerical (mean/median/KNN) vs. Categorical (mode/constant).
- **Nature of Missingness:** Is it random or related to other variables?.
- **Domain Knowledge:** Talk to experts to understand *why* data is missing

NOISY DATA

- NOISE: RANDOM ERROR OR VARIANCE IN A MEASURED VARIABLE
- INCORRECT ATTRIBUTE VALUES MAY BE DUE TO
  - FAULTY DATA COLLECTION INSTRUMENTS
  - DATA ENTRY PROBLEMS
  - DATA TRANSMISSION PROBLEMS
  - TECHNOLOGY LIMITATION
  - INCONSISTENCY IN NAMING CONVENTION
- OTHER DATA PROBLEMS WHICH REQUIRE DATA CLEANING
  - DUPLICATE RECORDS
  - INCOMPLETE DATA
  - INCONSISTENT DATA

HOW TO HANDLE NOISY DATA?

- BINNING
  - FIRST SORT DATA AND PARTITION INTO (EQUAL-FREQUENCY) BINS
  - THEN ONE CAN SMOOTH BY BIN MEANS, SMOOTH BY BIN MEDIAN, SMOOTH BY BIN BOUNDARIES, ETC.
- REGRESSION
  - SMOOTH BY FITTING THE DATA INTO REGRESSION FUNCTIONS
- CLUSTERING
  - DETECT AND REMOVE OUTLIERS
- COMBINED COMPUTER AND HUMAN INSPECTION
  - DETECT SUSPICIOUS VALUES AND CHECK BY HUMAN (E.G., DEAL WITH POSSIBLE OUTLIERS)

Outliers

Outliers are data points that are stand out from the rest. It can be either much higher or much lower than the other data points, and its presence can have a significant impact on the results of machine learning algorithms.

They are unusual values that don't follow the overall pattern of our data. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data.

For example, in a group of 5 students the test grades were 29, 38, 39, 27, and 2. The last value seems to be an outlier because it falls below the main pattern of the other grades.

27 - mean (with outlier)

33.25 - mean (without outlier)

Handling Outliers

1. **Removal:**

This involves identifying and removing outliers from the dataset before training the model. Common methods include:

- **Thresholding:** Outliers are identified as data points exceeding a certain threshold (e.g., Z-score > 3).
- **Distance-based methods:** Outliers are identified based on their distance from their nearest neighbors.
- **Clustering:** Outliers are identified as points not belonging to any cluster or belonging to very small clusters.

Handling Outliers

2. Transformation

This involves transforming the data to reduce the influence of outliers. Common methods include:

- **Scaling:** Standardizing or normalizing the data to have a mean of zero and a standard deviation of one.
- **Winsorization:** Replacing outlier values with the nearest non-outlier value.
- **Log transformation:** Applying a logarithmic transformation to compress the data and reduce the impact of extreme values.

Handling Outliers

3. Robust Estimation

This involves using algorithms that are less sensitive to outliers. Some examples include:

- **Robust Statistics:** Choose statistical methods less influenced by outliers like median, mode and interquartile range instead of mean and standard deviation.
- **Outlier-insensitive clustering algorithms:** Algorithms like DBSCAN are less susceptible to the presence of outliers than K-means clustering.

Handling Outliers

4. **Modeling Outliers**

This involves explicitly modeling the outliers as a separate group. This can be done by:

- **Adding a separate feature:** Create a new feature indicating whether a data point is an outlier or not.
- **Using a mixture model:** Train a model that assumes the data comes from a mixture of multiple distributions, where one distribution represents the outliers.

Handling Outliers

5. **Visualization for Outlier Detection**

We can use the box plot, or the box and whisker plot, to explore the dataset and visualize the presence of outliers. The points that lie beyond the whiskers are detected as outliers.

6. **Impute Outliers**

For outliers caused by missing or erroneous values, we can estimate replacements using the mean, median, mode or more advanced imputation methods.

Duplicates

Data points in a dataset that have the same values for all, or part of the characteristics are said to have duplicate values. Due to problems with data input, data collecting, or other circumstances, duplicate values may appear.

These duplicate values add redundancy to our data and can make our calculations go wrong.

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode of a an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Percentiles

- For continuous data, the notion of a percentile is more useful.

Given an ordinal or continuous attribute x and a number p between 0 and 100, the pth percentile is a value x_p of x such that p% of the observed values of x are less than x_p.

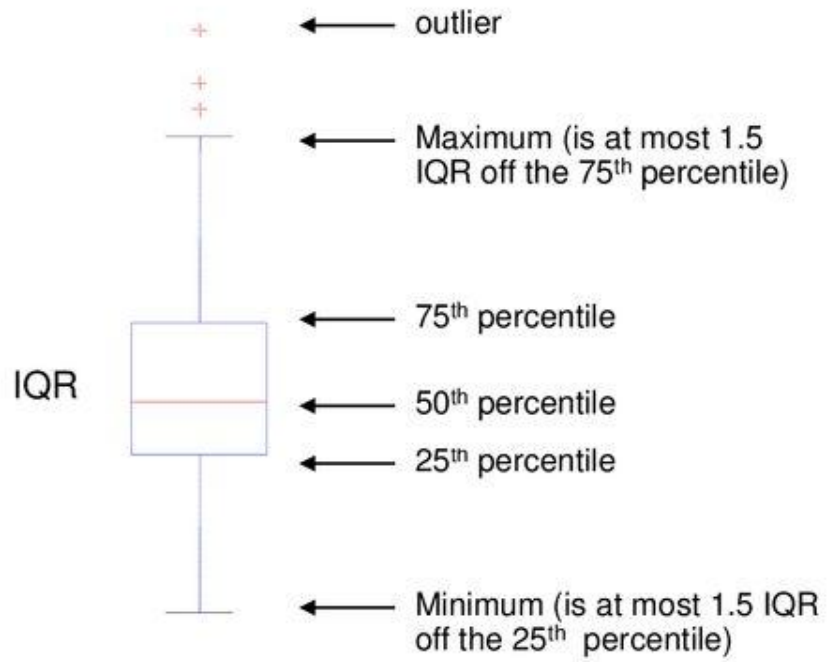
- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.
- Key Idea: associate attribute value with a rank!!

R Box Plots (different from textbook)

- R Box Plots

http://chartsgraphs.wordpress.com/2008/11/18/boxplots-r-does-them-right/

  - Invented by J. Tukey
  - displaying the distribution of data based on percentiles
  - Following figure shows the basic parts of a box plots



Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- Range is the difference between the max and min

0, 2, 3, 7, 8

- The variance or standard deviation

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

11.5

standard_deviation(x)= s_x

3.3

- However, this is also sensitive to outliers, so that other measures are often used.

2.8

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

(Mean Absolute Deviation) [Han]
(Absolute Average Deviation) [Tan]

3

$$\text{MAD}(x) = median\left(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\right)$$

(Median Absolute Deviation)

5

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

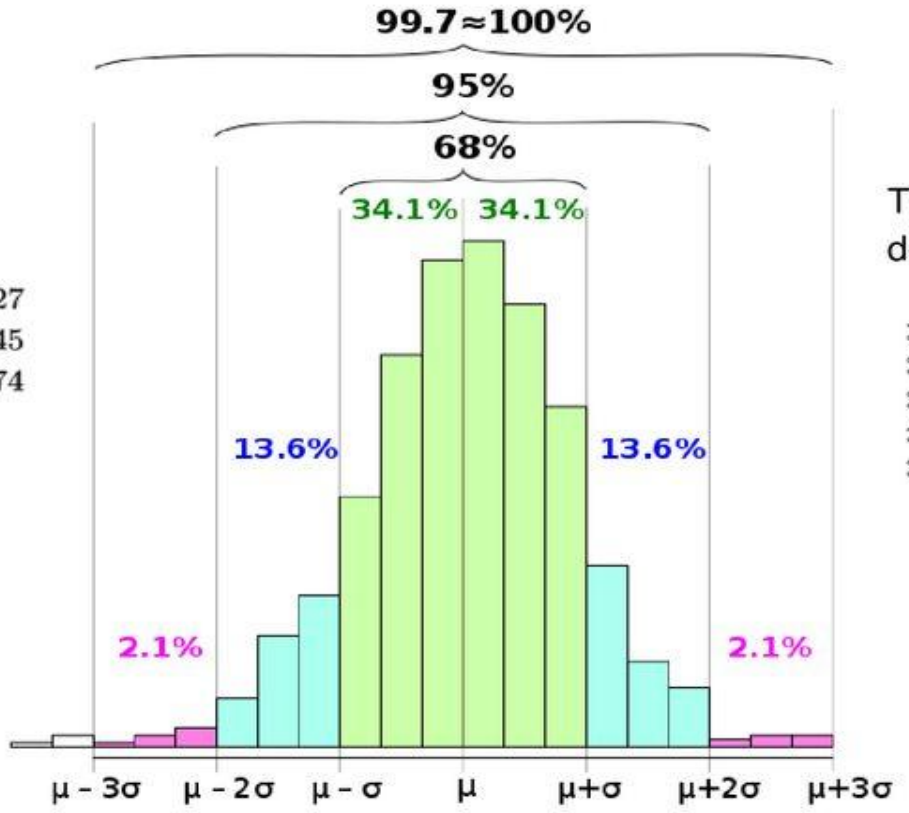
68-95-99.7 Rule

For more details see: https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule

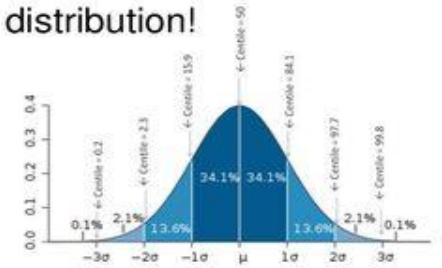
$$\Pr(\mu - \sigma \le X \le \mu + \sigma) \approx 0.6827$$
$$\Pr(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 0.9545$$
$$\Pr(\mu - 3\sigma \le X \le \mu + 3\sigma) \approx 0.9974$$

Remark:

- μ is mean value and
- σ is standard deviation



This rule assumes a Gaussian distribution!



68-95-99.7 Rule continued

Range	Expected fraction of population inside range	Approximate expected frequency outside range	Approximate frequency for daily event
$\mu \pm 0.5\sigma$	0.382924922548026	2 in 3	Four times a week
$\mu \pm \sigma$	0.682689492137086	1 in 3	Twice a week
$\mu \pm 1.5\sigma$	0.866385597462284	1 in 7	Weekly
$\mu \pm 2\sigma$	0.954499736103642	1 in 22	Every three weeks
$\mu \pm 2.5\sigma$	0.987580669348448	1 in 81	Quarterly
$\mu \pm 3\sigma$	0.997300203936740	1 in 370	Yearly
$\mu \pm 3.5\sigma$	0.999534741841929	1 in 2149	Every six years
$\mu \pm 4\sigma$	0.999936657516334	1 in 700415787000 0000000♠1578 7	Every 43 years (twice in a lifetime)

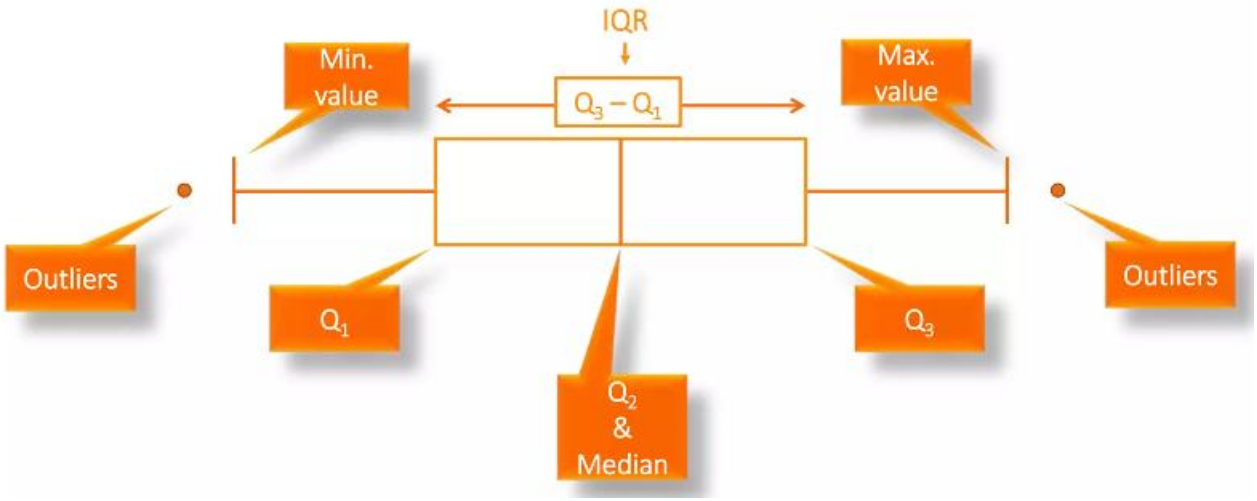
Tan,Steinbach, Kumar: Exploratory Data Analysis (with major modifications and additions by Ch. Eick)

Steps To Find Outliers

- Arrange the data in order from lowest to highest and find Q_1 and Q_3.
- Find the interquartile range (IQR) $Q_3 - Q_1$.
- Multiply IQR by 1.5.
- Subtract step 3 from Q_1 and add in Q_3.
- Check the data set for any data value that is smaller than $Q_1 - 1.5(IQR)$ or larger than $Q_3 + 1.5(IQR)$

How To Identify Outliers ?

A data value less than $Q_1 - 1.5(IQR)$ or greater than $Q_3 + 1.5(IQR)$ can be considered an outlier.



1. Arrange the data & find Q_1, Q_3 .

   7, 10, 11, 15, 25, 30, 35, 68

   Q_1. = 10.5 Q_3 = 32.5

2. Find the IQR $(Q_3 - Q_1)$

   = 32.5 – 10.5

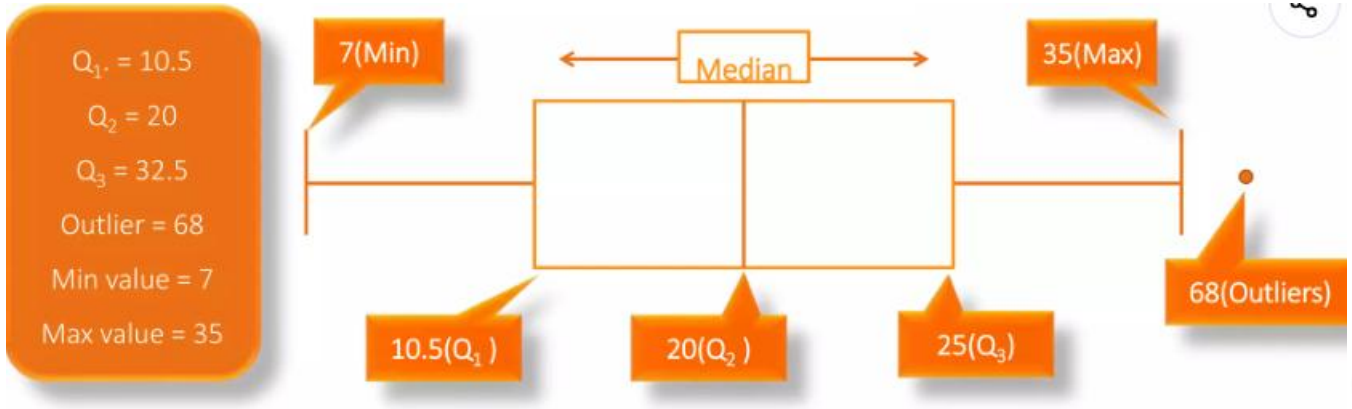
   = 22

3. Multiply IQR by 1.5

   = 33

4. Subtract IQR from Q_1 & add in Q_3 10.5 – 33 = -22.5
   32.5 + 33 = 65.5

Example
10, 11, 15, 25, 35, 30, 7, 68



DATA CLEANING AS A PROCESS

- DATA DISCREPANCY DETECTION
  - USE METADATA (E.G., DOMAIN, RANGE, DEPENDENCY, DISTRIBUTION)
  - CHECK FIELD OVERLOADING
  - CHECK UNIQUENESS RULE, CONSECUTIVE RULE AND NULL RULE
  - USE COMMERCIAL TOOLS
    - DATA SCRUBBING: USE SIMPLE DOMAIN KNOWLEDGE (E.G., POSTAL CODE, SPELL-CHECK) TO DETECT ERRORS AND MAKE CORRECTIONS
    - DATA AUDITING: BY ANALYZING DATA TO DISCOVER RULES AND RELATIONSHIP TO DETECT VIOLATORS (E.G., CORRELATION AND CLUSTERING TO FIND OUTLIERS)
- DATA MIGRATION AND INTEGRATION
  - DATA MIGRATION TOOLS: ALLOW TRANSFORMATIONS TO BE SPECIFIED
  - ETL (EXTRACTION/TRANSFORMATION/LOADING) TOOLS: ALLOW USERS TO SPECIFY TRANSFORMATIONS THROUGH A GRAPHICAL USER INTERFACE
- INTEGRATION OF THE TWO PROCESSES
  - ITERATIVE AND INTERACTIVE (E.G., POTTER'S WHEELS)

CHAPTER 3: DATA PREPROCESSING

- DATA PREPROCESSING: AN OVERVIEW
  - DATA QUALITY
  - MAJOR TASKS IN DATA PREPROCESSING
- DATA CLEANING
- DATA INTEGRATION 
- DATA REDUCTION
- DATA TRANSFORMATION AND DATA DISCRETIZATION
- SUMMARY

70

DATA INTEGRATION

- **DATA INTEGRATION**:
  - COMBINES DATA FROM MULTIPLE SOURCES INTO A COHERENT STORE
- SCHEMA INTEGRATION: E.G., A.CUST-ID ≡ B.CUST-#
  - INTEGRATE METADATA FROM DIFFERENT SOURCES
- ENTITY IDENTIFICATION PROBLEM:
  - IDENTIFY REAL WORLD ENTITIES FROM MULTIPLE DATA SOURCES, E.G., BILL CLINTON = WILLIAM CLINTON
- DETECTING AND RESOLVING DATA VALUE CONFLICTS
  - FOR THE SAME REAL WORLD ENTITY, ATTRIBUTE VALUES FROM DIFFERENT SOURCES ARE DIFFERENT
  - POSSIBLE REASONS: DIFFERENT REPRESENTATIONS, DIFFERENT SCALES, E.G., METRIC VS. BRITISH UNITS

71

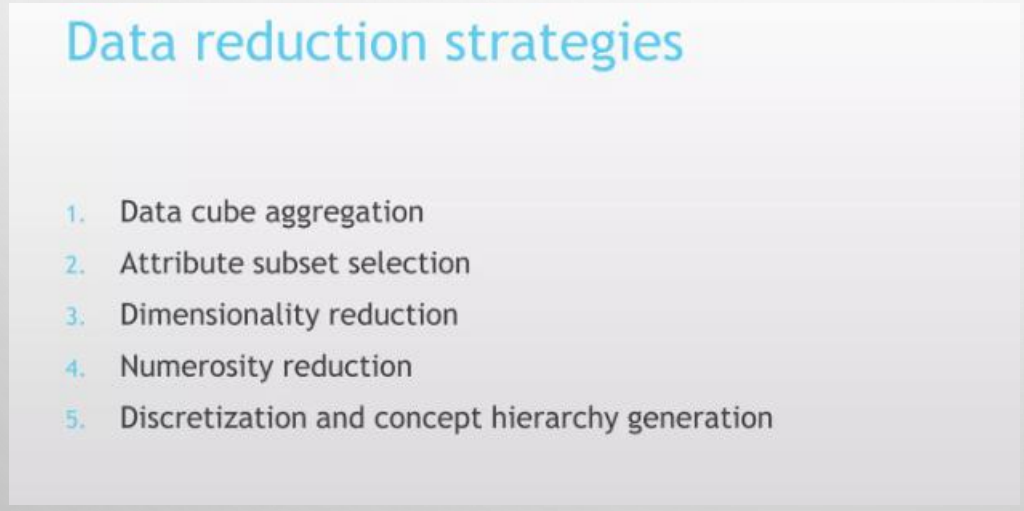
CHAPTER 3: DATA PREPROCESSING

- DATA PREPROCESSING: AN OVERVIEW
  - DATA QUALITY
  - MAJOR TASKS IN DATA PREPROCESSING
- DATA CLEANING
- DATA INTEGRATION
- DATA REDUCTION
- DATA TRANSFORMATION AND DATA DISCRETIZATION
- SUMMARY

72

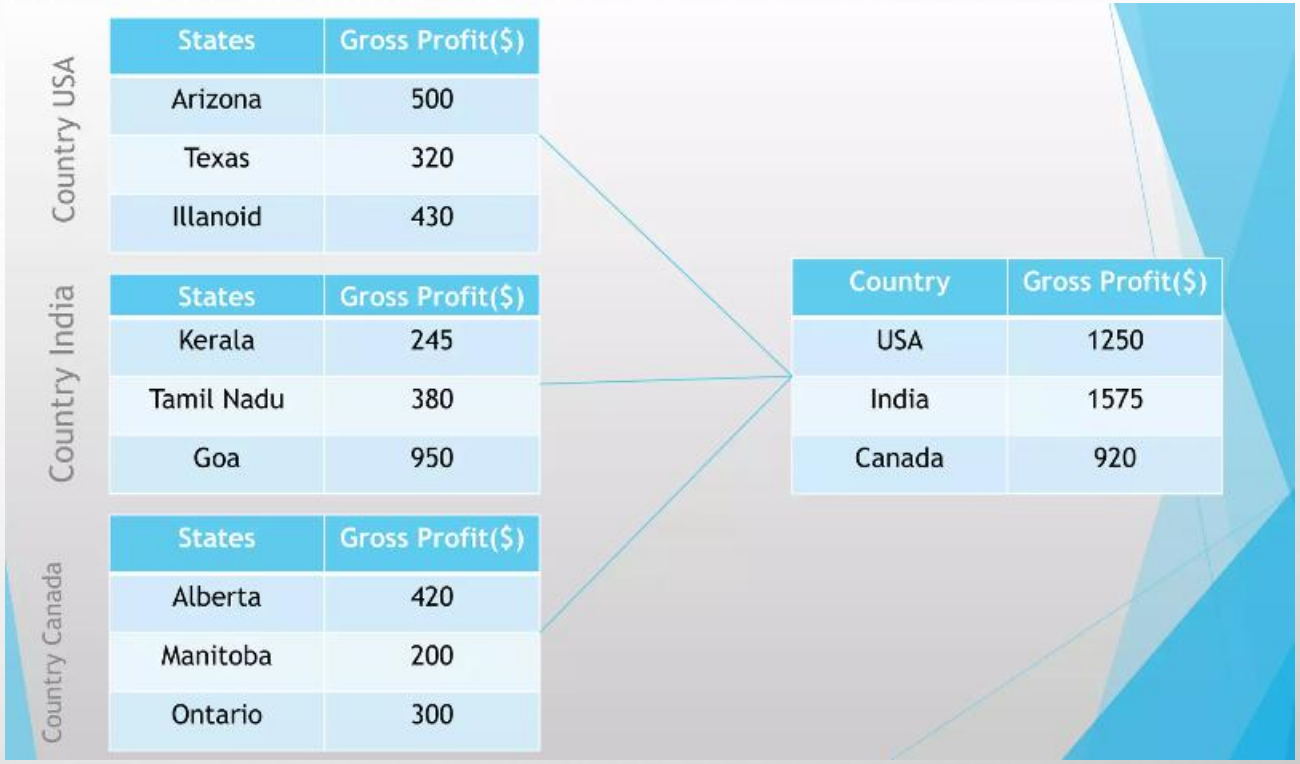
DATA REDUCTION STRATEGIES

- **DATA REDUCTION:** OBTAIN A REDUCED REPRESENTATION OF THE DATA SET THAT IS MUCH SMALLER IN VOLUME BUT YET PRODUCES THE SAME (OR ALMOST THE SAME) ANALYTICAL RESULTS
- WHY DATA REDUCTION? — A DATABASE/DATA WAREHOUSE MAY STORE TERABYTES OF DATA. COMPLEX DATA ANALYSIS MAY TAKE A VERY LONG TIME TO RUN ON THE COMPLETE DATA SET.



Data Cube Aggregation

This technique is used to aggregate (combine) data in a simpler form. So we can summarize the data in such a way that the data is used as result



FEATURE SELECTION AND FEATURE ELIMINATION

STEP-WISE METHODS AND TECHNIQUES

ATTRIBUTE SUBSET SELECTION

- ANOTHER WAY TO REDUCE DIMENSIONALITY OF DATA
- 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 MINING TASK AT HAND
  - E.G., STUDENTS' ID IS OFTEN IRRELEVANT TO THE TASK OF PREDICTING STUDENTS' GPA

HEURISTIC SEARCH IN ATTRIBUTE SELECTION

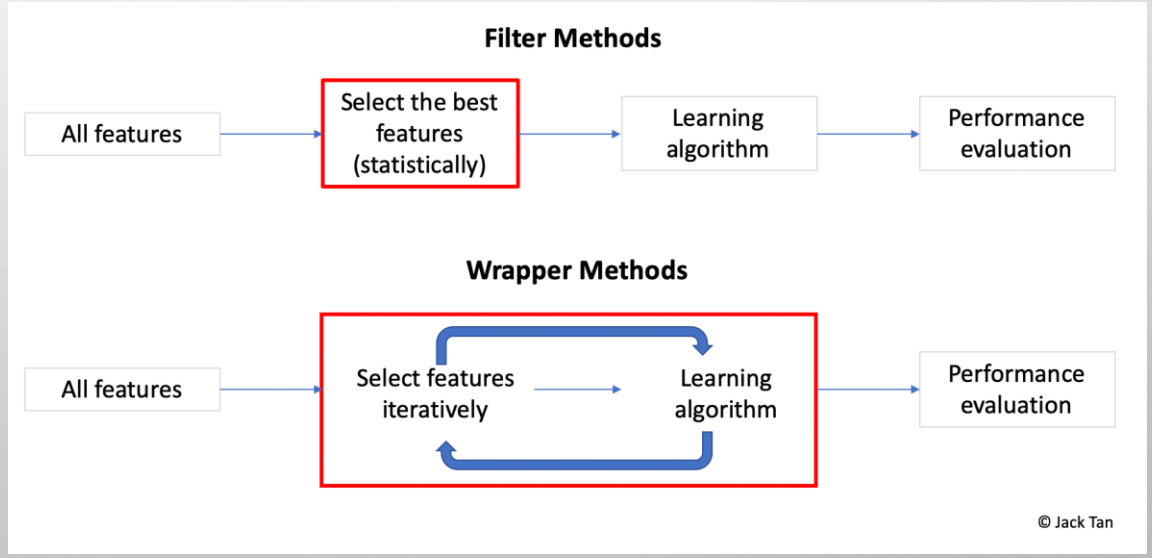
- THERE ARE 2^D POSSIBLE ATTRIBUTE COMBINATIONS OF D ATTRIBUTES
- TYPICAL HEURISTIC ATTRIBUTE SELECTION METHODS:
  - BEST SINGLE ATTRIBUTE UNDER THE ATTRIBUTE INDEPENDENCE ASSUMPTION: CHOOSE BY SIGNIFICANCE TESTS
  - BEST STEP-WISE FEATURE SELECTION:
    - THE BEST SINGLE-ATTRIBUTE IS PICKED FIRST
    - 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

WHY FEATURE SELECTION?

- REDUCES OVERFITTING
- IMPROVES ACCURACY
- REDUCES TRAINING TIME
- REMOVES IRRELEVANT FEATURES

TYPES OF FEATURE SELECTION

- WRAPPER METHODS
- FILTER METHODS
- EMBEDDED METHODS



STEP-WISE FEATURE SELECTION

- ITERATIVE PROCESS
- BASED ON MODEL PERFORMANCE
- USED IN REGRESSION AND CLASSIFICATION

FORWARD SELECTION

- 1. START WITH NO FEATURES
- 2. ADD ONE FEATURE AT A TIME
- 3. EVALUATE MODEL
- 4. SELECT BEST FEATURE
- 5. REPEAT UNTIL NO IMPROVEMENT

BACKWARD ELIMINATION

- 1. START WITH ALL FEATURES
- 2. TRAIN THE MODEL
- 3. REMOVE LEAST SIGNIFICANT FEATURE
- 4. RE-TRAIN MODEL
- 5. REPEAT

STEP-WISE SELECTION

- COMBINATION OF FORWARD AND BACKWARD METHODS
- ADDS SIGNIFICANT FEATURES
- REMOVES INSIGNIFICANT FEATURES

FEATURE ELIMINATION

- REMOVING IRRELEVANT OR REDUNDANT FEATURES
- IMPROVES GENERALIZATION
- REDUCES MODEL COMPLEXITY

FILTER-BASED ELIMINATION

- BASED ON STATISTICAL MEASURES
- MODEL INDEPENDENT
- EXAMPLES: CORRELATION, CHI-SQUARE, MUTUAL INFORMATION

CORRELATION ANALYSIS (NOMINAL DATA)

- **X^2 (CHI-SQUARE) TEST**

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

- THE LARGER THE X^2 VALUE, THE MORE LIKELY THE VARIABLES ARE RELATED
- THE CELLS THAT CONTRIBUTE THE MOST TO THE X^2 VALUE ARE THOSE WHOSE ACTUAL COUNT IS VERY DIFFERENT FROM THE EXPECTED COUNT
- CORRELATION DOES NOT IMPLY CAUSALITY
  - # OF HOSPITALS AND # OF CAR-THEFT IN A CITY ARE CORRELATED
  - BOTH ARE CAUSALLY LINKED TO THE THIRD VARIABLE: POPULATION

CHI-SQUARE CALCULATION: AN EXAMPLE

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (CHI-SQUARE) CALCULATION (NUMBERS IN PARENTHESIS ARE EXPECTED COUNTS CALCULATED BASED ON THE DATA DISTRIBUTION IN THE TWO CATEGORIES)

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

- IT SHOWS THAT LIKE_SCIENCE_FICTION AND PLAY_CHESS ARE CORRELATED IN THE GROUP

CORRELATION ANALYSIS (NUMERIC DATA)

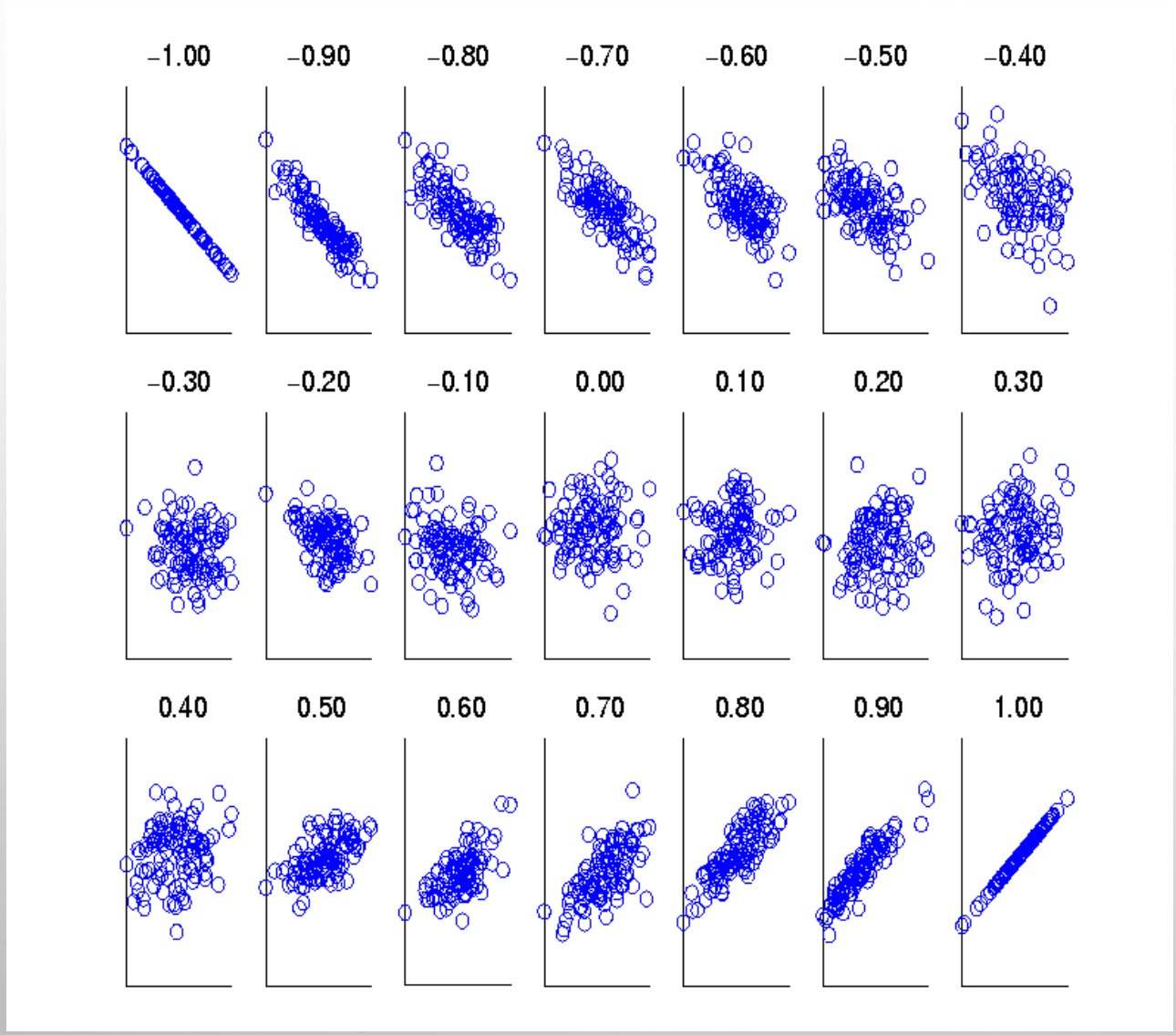
- CORRELATION COEFFICIENT (ALSO CALLED PEARSON'S PRODUCT MOMENT COEFFICIENT)

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

WHERE N IS THE NUMBER OF TUPLES, $\bar{A}$ AND $\bar{B}$ ARE THE RESPECTIVE MEANS OF A AND B, Σ_A AND Σ_B ARE THE RESPECTIVE STANDARD DEVIATION OF A AND B, AND $\Sigma(A_lB_l)$ IS THE SUM OF THE AB CROSS-PRODUCT.

- 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

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A' \bullet B'$$

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, Σ_A AND Σ_B ARE THE RESPECTIVE STANDARD DEVIATION OF A AND B.

- **POSITIVE COVARIANCE**: IF $COV_{A,B} > 0$, THEN A AND B BOTH TEND TO BE LARGER THAN THEIR EXPECTED VALUES.
- **NEGATIVE COVARIANCE**: IF $COV_{A,B} < 0$ THEN IF A IS LARGER THAN ITS EXPECTED VALUE, B IS LIKELY TO BE SMALLER THAN ITS EXPECTED VALUE.
- **INDEPENDENCE**: $COV_{A,B} = 0$ 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 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

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

- 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
  - COV(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- THUS, A AND B RISE TOGETHER SINCE COV(A, B) > 0.

RECURSIVE FEATURE ELIMINATION (RFE)

1. TRAIN MODEL WITH ALL FEATURES
2. RANK FEATURE IMPORTANCE
3. REMOVE LEAST IMPORTANT FEATURE
4. REPEAT

EMBEDDED METHODS

- FEATURE SELECTION DURING TRAINING
- EXAMPLES:

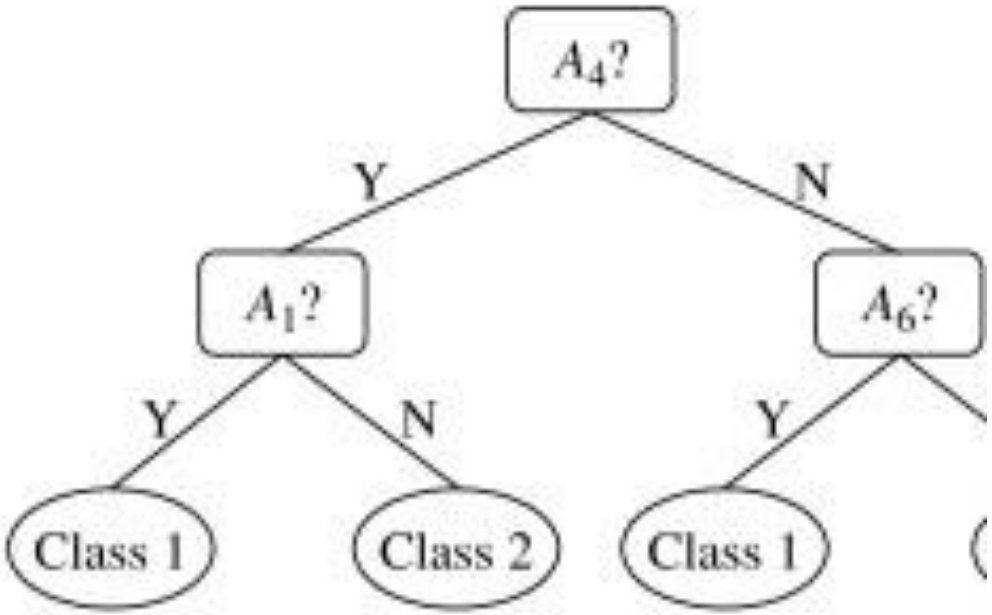
  - LASSO REGRESSION

  - DECISION TREES

  - RANDOM FOREST

COMPARISON OF METHODS

- • FILTER: FAST, MODEL INDEPENDENT
- • WRAPPER: ACCURATE, COMPUTATIONALLY EXPENSIVE
- • EMBEDDED: EFFICIENT, MODEL SPECIFIC

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: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => 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 (NOT COVERED)
  - ATTRIBUTE CONSTRUCTION
    - COMBINING FEATURES (SEE: DISCRIMINATIVE FREQUENT PATTERNS IN CHAPTER 7)
    - DATA DISCRETIZATION

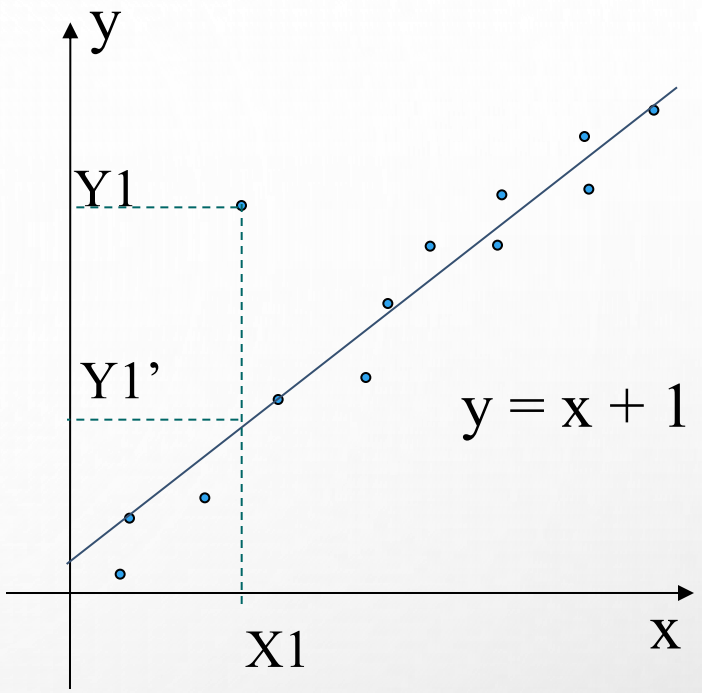
DATA REDUCTION 2: NUMEROSITY REDUCTION

- 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



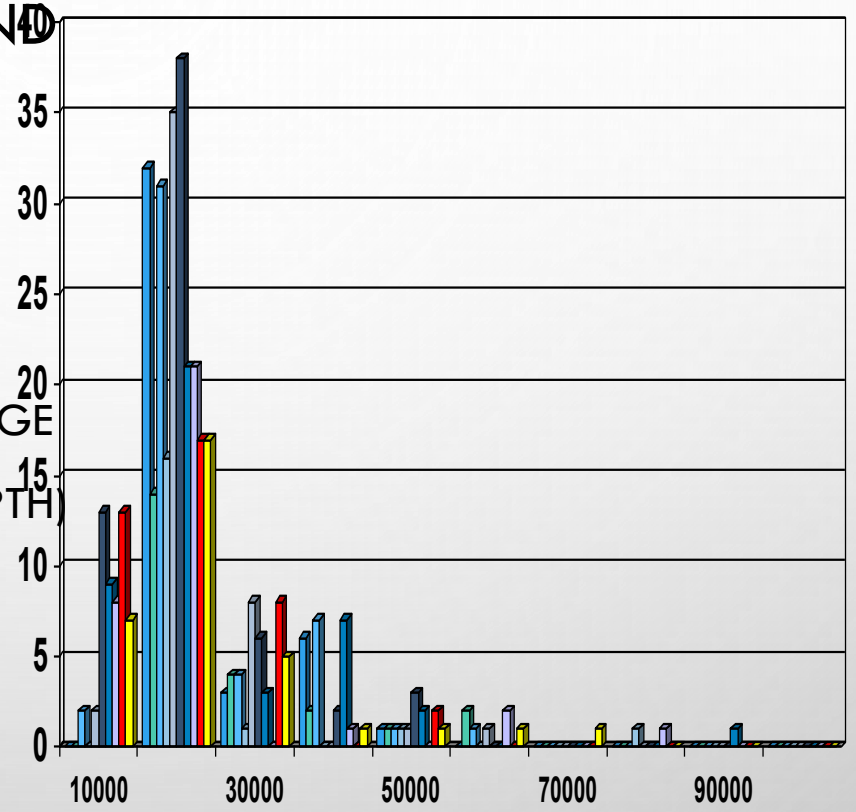
- REGRESSION ANALYSIS: A COLLECTIVE NAME FOR TECHNIQUES FOR THE MODELING AND ANALYSIS OF NUMERICAL DATA CONSISTING OF VALUES OF A ***DEPENDENT VARIABLE*** (ALSO CALLED ***RESPONSE VARIABLE*** OR *MEASUREMENT*) AND OF ONE OR MORE *INDEPENDENT VARIABLES* (AKA. ***EXPLANATORY VARIABLES*** OR ***PREDICTORS***)
- 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
- USED FOR PREDICTION (INCLUDING FORECASTING OF TIME-SERIES DATA), INFERENCE, HYPOTHESIS TESTING, AND MODELING OF CAUSAL RELATIONSHIPS

REGRESS 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, …, X_1, X_2, ….
- MULTIPLE REGRESSION: $Y = B_0 + B_1 X_1 + B_2 X_2$
  - MANY NONLINEAR FUNCTIONS CAN BE TRANSFORMED INTO THE ABOVE
- 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)



CLUSTERING

- 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
- CLUSTER ANALYSIS WILL BE STUDIED IN DEPTH IN CHAPTER 10

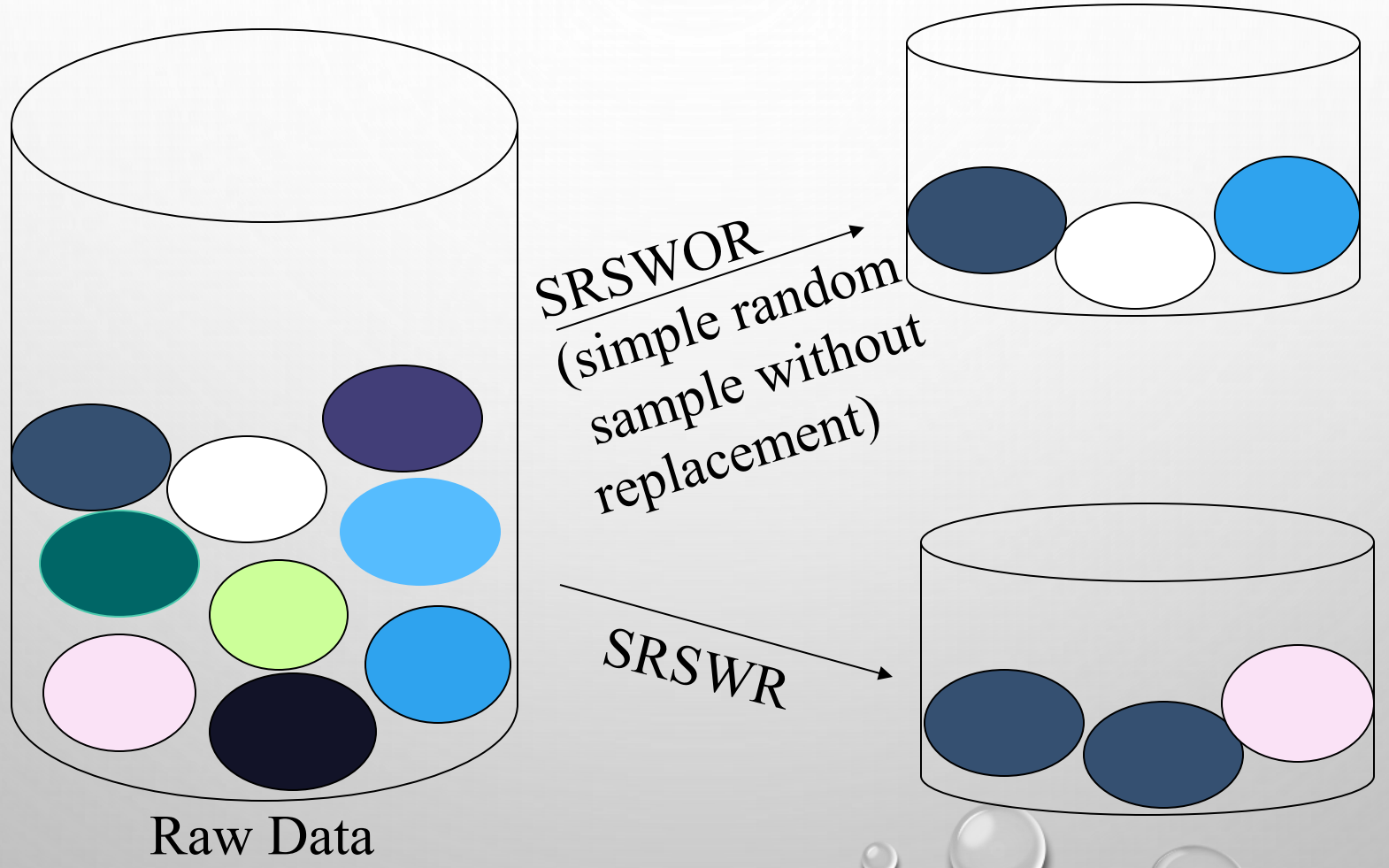
SAMPLING

- SAMPLING: OBTAINING A SMALL SAMPLE *S* TO REPRESENT THE WHOLE DATA SET *N*
- ALLOW A MINING 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)

TYPES OF SAMPLING

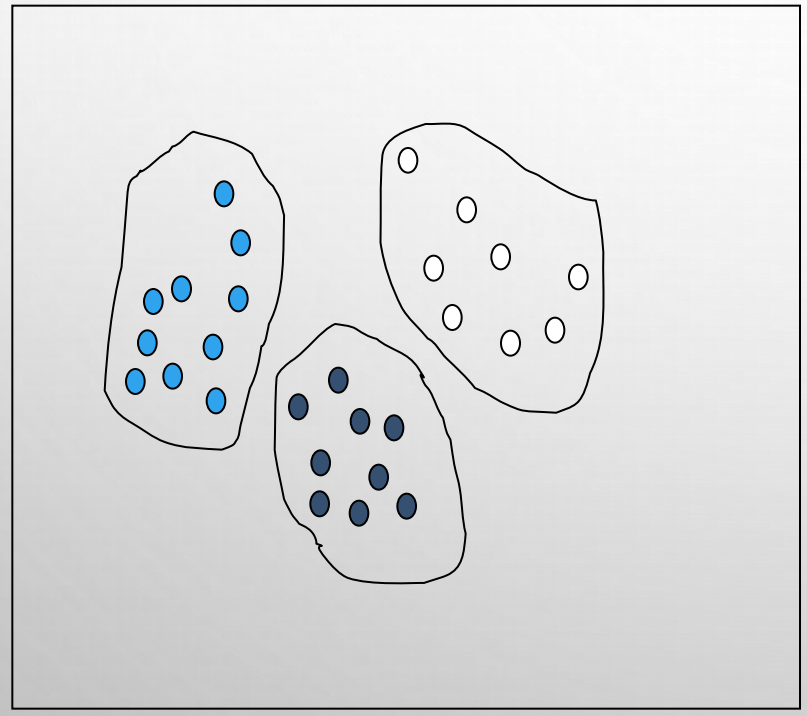
- **SIMPLE RANDOM SAMPLING**
  - THERE IS AN EQUAL PROBABILITY OF SELECTING ANY PARTICULAR ITEM
- **SAMPLING WITHOUT REPLACEMENT**
  - ONCE AN OBJECT IS SELECTED, IT IS REMOVED FROM THE POPULATION
- **SAMPLING WITH REPLACEMENT**
  - A SELECTED OBJECT IS NOT REMOVED FROM THE POPULATION
- **STRATIFIED SAMPLING:**
  - PARTITION THE DATA SET, AND DRAW SAMPLES FROM EACH PARTITION (PROPORTIONALLY, I.E., APPROXIMATELY THE SAME PERCENTAGE OF THE DATA)
  - USED IN CONJUNCTION WITH SKEWED DATA

Sampling: With or without Replacement

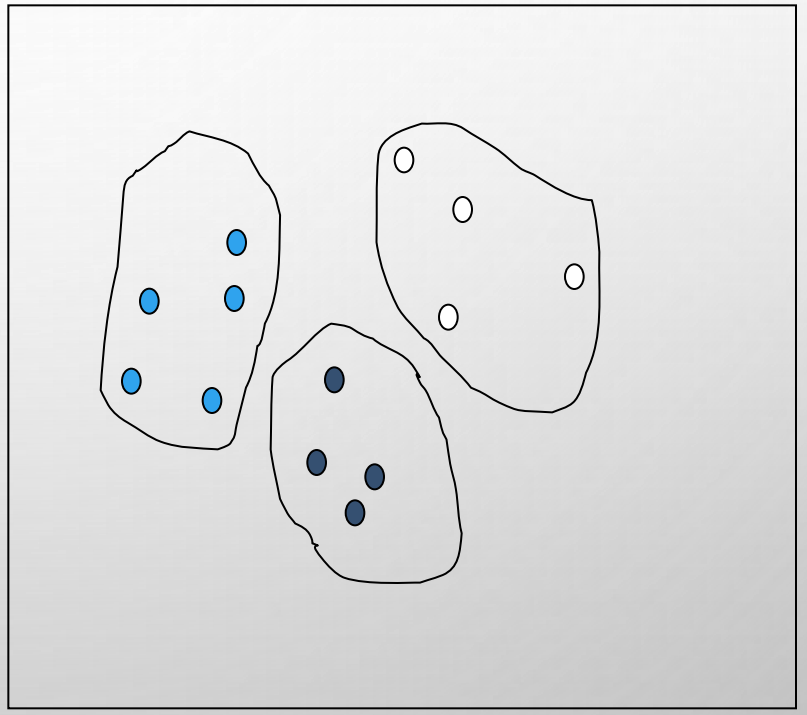


SAMPLING: CLUSTER OR STRATIFIED SAMPLING

Raw Data



Cluster/Stratified Sample



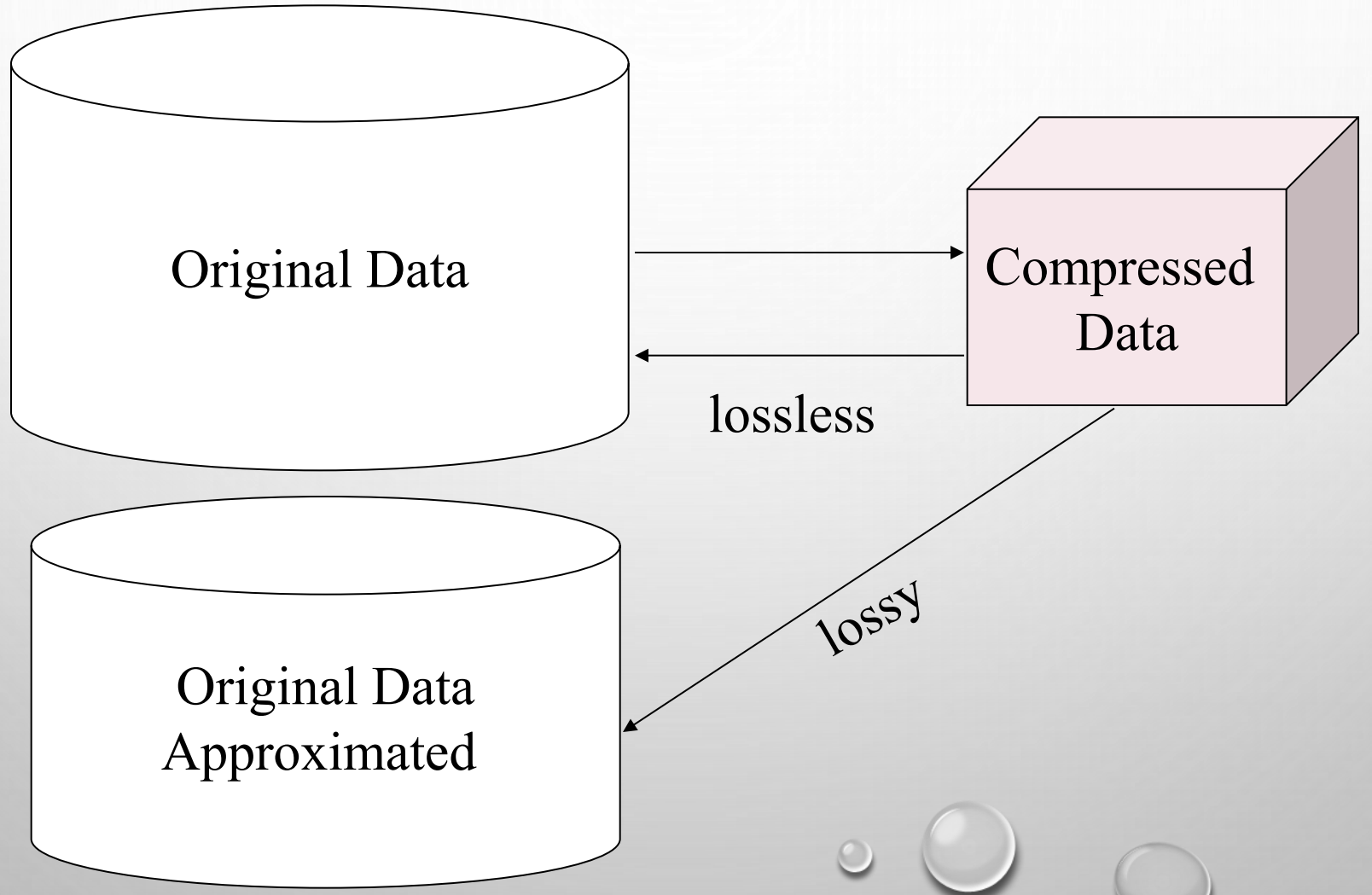
DATA CUBE AGGREGATION

- 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 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

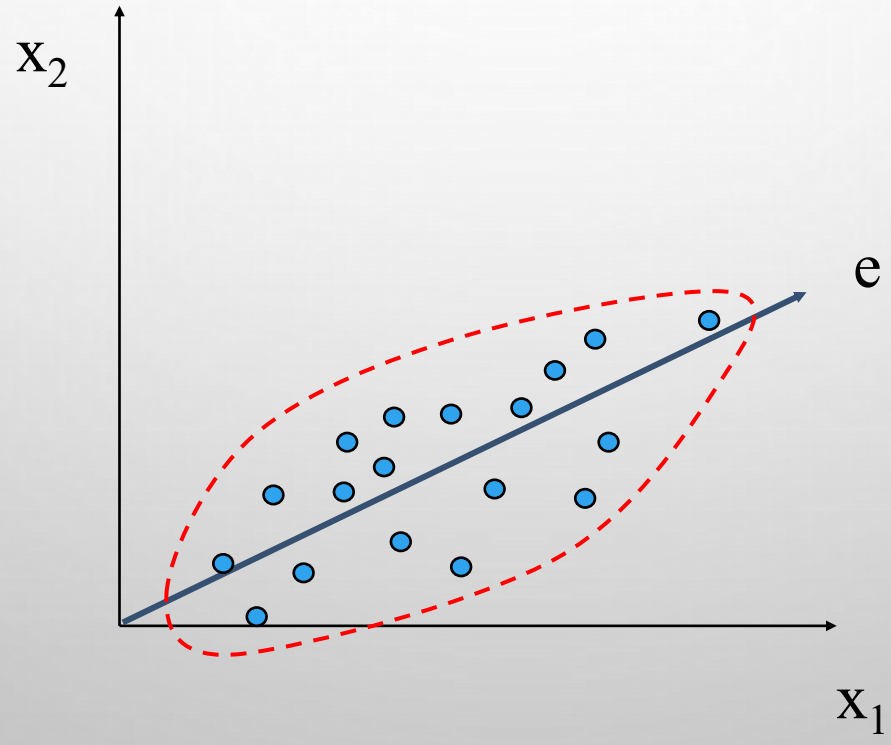


DATA REDUCTION 1: DIMENSIONALITY REDUCTION

- **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
- **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)

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

DATA TRANSFORMATION

- A FUNCTION THAT MAPS THE ENTIRE SET OF VALUES OF A GIVEN ATTRIBUTE TO A NEW SET OF REPLACEMENT VALUES S.T. EACH OLD VALUE CAN BE IDENTIFIED WITH ONE OF THE NEW VALUES
- METHODS
  - SMOOTHING: REMOVE NOISE FROM DATA
  - ATTRIBUTE/FEATURE CONSTRUCTION
    - NEW ATTRIBUTES CONSTRUCTED FROM THE GIVEN ONES
  - 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

CHAPTER 3: DATA PREPROCESSING

- DATA PREPROCESSING: AN OVERVIEW
  - DATA QUALITY
  - MAJOR TASKS IN DATA PREPROCESSING
- DATA CLEANING
- DATA INTEGRATION
- DATA REDUCTION
- DATA TRANSFORMATION AND DATA DISCRETIZATION
- SUMMARY

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 \$73,000 IS MAPPED TO $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- **Z-SCORE NORMALIZATION** (M: MEAN, Σ: STANDARD DEVIATION):

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

  - EX. LET M = 54,000, Σ = 16,000. THEN $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **NORMALIZATION BY DECIMAL SCALING**

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

Example: Salary Data

Assume a dataset of employee salaries: **$100, $500, $1000, $5000**.

1. **Find the maximum absolute value**: $\max(|v|) = 5000$.
2. **Determine j (number of digits in max value)**: 5000 has 4 digits, so $j = 4$.
3. **Calculate 10^j**: $10^4 = 10,000$.
4. **Divide each value by $10,000$**:
    1. $100/10,000 = \mathbf{0.01}$
    2. $500/10,000 = \mathbf{0.05}$
    3. $1000/10,000 = \mathbf{0.1}$
    4. $5000/10,000 = \mathbf{0.5}$

Normalized Data: 0.01, 0.05, 0.1, 0.5.

DISCRETIZATION

- THREE TYPES OF ATTRIBUTES
  - NOMINAL—VALUES FROM AN UNORDERED SET, E.G., COLOR, PROFESSION
  - ORDINAL—VALUES FROM AN ORDERED SET, E.G., MILITARY OR ACADEMIC RANK
  - NUMERIC—REAL NUMBERS, E.G., INTEGER OR REAL NUMBERS
- DISCRETIZATION: DIVIDE THE RANGE OF A CONTINUOUS ATTRIBUTE INTO INTERVALS
  - INTERVAL LABELS CAN THEN BE USED TO REPLACE ACTUAL DATA VALUES
  - REDUCE DATA SIZE BY DISCRETIZATION
  - 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., χ^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
- 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

Example of Equi-width

- *Data:* {5, 10, 15, 20, 25, 30}
- *Bins ($k = 3$):* Range is 25, width is $25/3 \approx 8.33$.
- *Result:* $[5, 13.3), [13.3, 21.6), [21.6, 30]$.

Example of Equi-depth

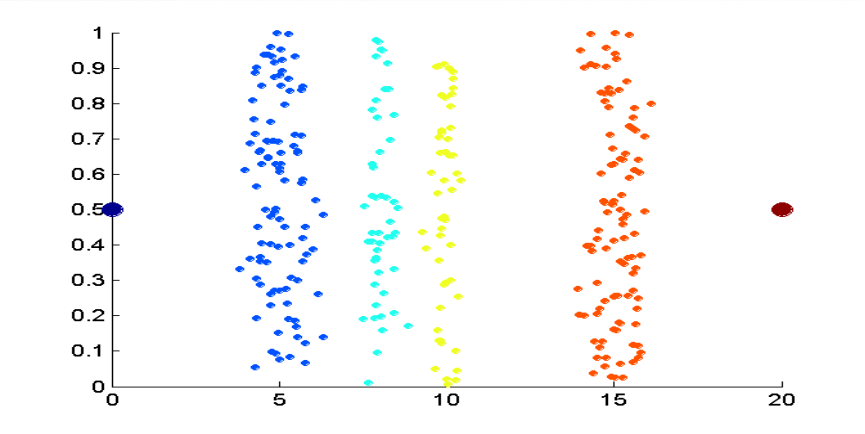
- *Data:* {5, 10, 15, 20, 25, 30}
- *Bins ($k = 3$):* Each bin gets $6/3 = 2$ values.
- *Result:* Bin 1: {5, 10}, Bin 2: {15, 20}, Bin 3: {25, 30}.

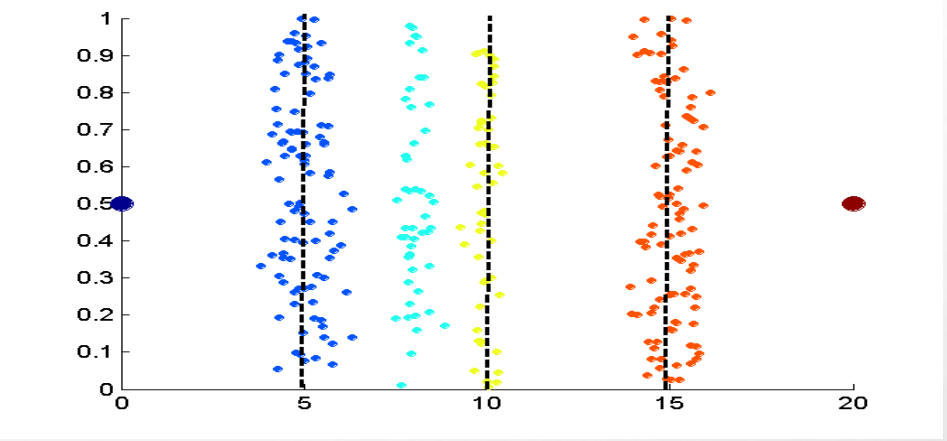
BINNING METHODS FOR DATA SMOOTHING

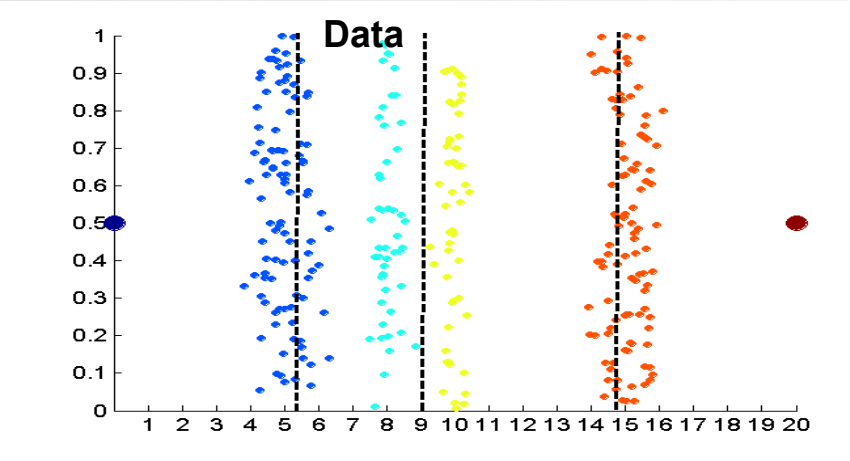
- SORTED DATA FOR PRICE (IN DOLLARS): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

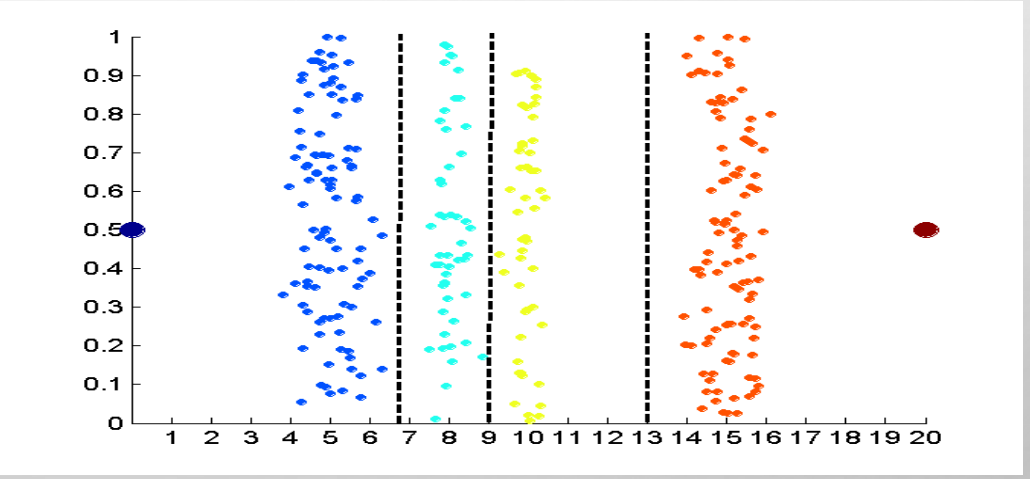
* PARTITION INTO EQUAL-FREQUENCY (**EQUI-DEPTH**) BINS:
  - BIN 1: 4, 8, 9, 15
  - BIN 2: 21, 21, 24, 25
  - BIN 3: 26, 28, 29, 34
* SMOOTHING BY **BIN MEANS**:
  - BIN 1: 9, 9, 9, 9
  - BIN 2: 23, 23, 23, 23
  - BIN 3: 29, 29, 29, 29
* SMOOTHING BY **BIN BOUNDARIES**:
  - BIN 1: 4, 4, 4, 15
  - BIN 2: 21, 21, 25, 25
  - BIN 3: 26, 26, 26, 34

DISCRETIZATION WITHOUT USING CLASS LABELS (BINNING VS. CLUSTERING)









Equal frequency (binning)

K-means clustering leads to better results

DISCRETIZATION BY CLASSIFICATION & 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
  - DETAILS TO BE COVERED IN CHAPTER 7
- CORRELATION ANALYSIS (E.G., CHI-MERGE: X^2-BASED DISCRETIZATION)
  - SUPERVISED: USE CLASS INFORMATION
  - BOTTOM-UP MERGE: FIND THE BEST NEIGHBORING INTERVALS (THOSE HAVING SIMILAR DISTRIBUTIONS OF CLASSES, I.E., LOW X^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 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 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
  - {URBANA, CHAMPAIGN, CHICAGO} < 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: {*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

CHAPTER 3: DATA PREPROCESSING

- DATA PREPROCESSING: AN OVERVIEW
  - DATA QUALITY
  - MAJOR TASKS IN DATA PREPROCESSING
- DATA CLEANING
- DATA INTEGRATION
- DATA REDUCTION
- DATA TRANSFORMATION AND DATA DISCRETIZATION
- SUMMARY

SUMMARY

- **DATA QUALITY**: ACCURACY, COMPLETENESS, CONSISTENCY, TIMELINESS, BELIEVABILITY, INTERPRETABILITY
- **DATA CLEANING**: E.G. MISSING/NOISY VALUES, OUTLIERS
- **DATA INTEGRATION** FROM MULTIPLE SOURCES:
  - ENTITY IDENTIFICATION PROBLEM
  - REMOVE REDUNDANCIES
  - DETECT INCONSISTENCIES
- **DATA REDUCTION**
  - DIMENSIONALITY REDUCTION
  - NUMEROSITY REDUCTION
  - DATA COMPRESSION
- **DATA TRANSFORMATION AND DATA DISCRETIZATION**
  - NORMALIZATION
  - CONCEPT HIERARCHY GENERATION

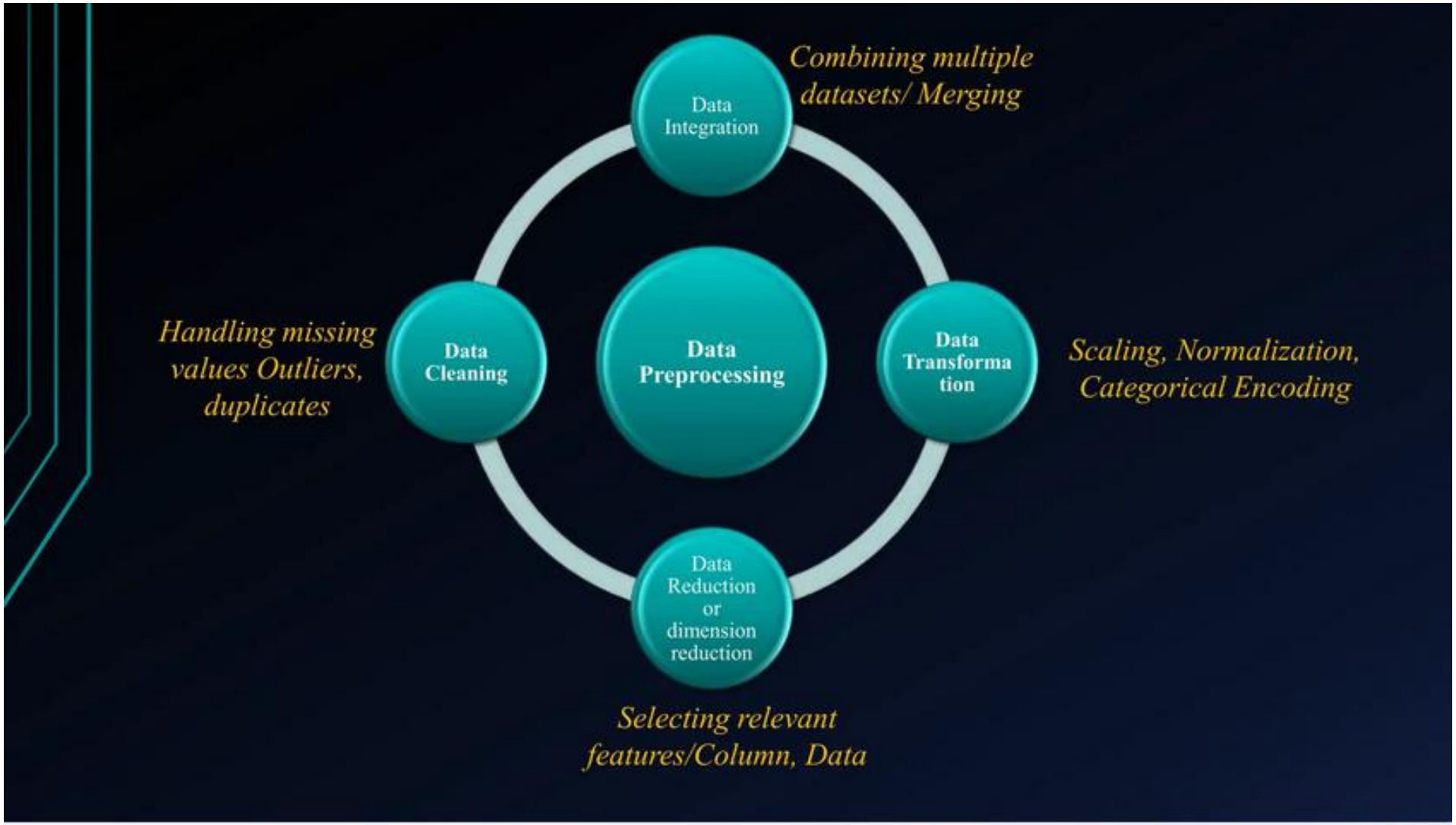
REFERENCES

- D. P. BALLOU AND G. K. TAYI. ENHANCING DATA QUALITY IN DATA WAREHOUSE ENVIRONMENTS. COMM. OF ACM, 42:73-78, 1999
- A. BRUCE, D. DONOHO, AND H.-Y. GAO. WAVELET ANALYSIS. *IEEE SPECTRUM*, OCT 1996
- T. DASU AND T. JOHNSON. EXPLORATORY DATA MINING AND DATA CLEANING. JOHN WILEY, 2003
- J. DEVORE AND R. PECK. *STATISTICS: THE EXPLORATION AND ANALYSIS OF DATA*. DUXBURY PRESS, 1997.
- H. GALHARDAS, D. FLORESCU, D. SHASHA, E. SIMON, AND C.-A. SAITA. DECLARATIVE DATA CLEANING: LANGUAGE, MODEL, AND ALGORITHMS. *VLDB'01*
- M. HUA AND J. PEI. CLEANING DISGUISED MISSING DATA: A HEURISTIC APPROACH. *KDD'07*
- H. V. JAGADISH, ET AL., SPECIAL ISSUE ON DATA REDUCTION TECHNIQUES. BULLETIN OF THE TECHNICAL COMMITTEE ON DATA ENGINEERING, 20(4), DEC. 1997
- H. LIU AND H. MOTODA (EDS.). *FEATURE EXTRACTION, CONSTRUCTION, AND SELECTION: A DATA MINING PERSPECTIVE*. KLUWER ACADEMIC, 1998
- J. E. OLSON. *DATA QUALITY: THE ACCURACY DIMENSION*. MORGAN KAUFMANN, 2003
- D. PYLE. DATA PREPARATION FOR DATA MINING. MORGAN KAUFMANN, 1999
- V. RAMAN AND J. HELLERSTEIN. POTTERS WHEEL: AN INTERACTIVE FRAMEWORK FOR DATA CLEANING AND TRANSFORMATION, VLDB'2001
- T. REDMAN. *DATA QUALITY: THE FIELD GUIDE*. DIGITAL PRESS (ELSEVIER), 2001
- R. WANG, V. STOREY, AND C. FIRTH. A FRAMEWORK FOR ANALYSIS OF DATA QUALITY RESEARCH. IEEE TRANS. KNOWLEDGE AND DATA ENGINEERING, 7:623-640, 1995

Data Preprocessing

Data Preprocessing can be defined as a process of converting raw data into a format that is understandable and usable for further analysis. It is an important step in the Data Preparation stage. It ensures that the outcome of the analysis is **accurate**, **complete**, and **consistent**.

Data preprocessing refers to the cleaning, transforming and integrating of data to make it ready for analysis.


Data Integration
Combining multiple datasets/ Merging
Data Cleaning
Handling missing values Outliers, duplicates
Data Preprocessing
Data Transformation
Scaling, Normalization, Categorical Encoding
Data Reduction or dimension reduction
Selecting relevant features/Column, Data

Data Cleaning

Data cleaning involves the systematic identification and correction of errors, inconsistencies, and inaccuracies within a dataset, encompassing tasks such as ***handling missing values, removing duplicates, and addressing outliers.***

Data cleaning is essential because raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of the insights derived from it.

Handling Missing Values

Missing values can be handled by many techniques, such as removing rows/columns containing NULL values and imputing NULL values using mean, mode etc.

There are several useful functions for detecting, removing, and replacing null values in Pandas DataFrame :

- isnull()
- notnull()

Both function help in checking whether a value is NaN or not

- dropna()
- fillna()
- replace()
- Interpolate()

To fill null values in a datasets, we use fillna(), replace() and interpolate() function these function replace NaN values with some value of their own.
Interpolate() function is basically used to fill NA values in the dataframe but it uses various interpolation technique to fill the missing values rather than hard-coding the value.

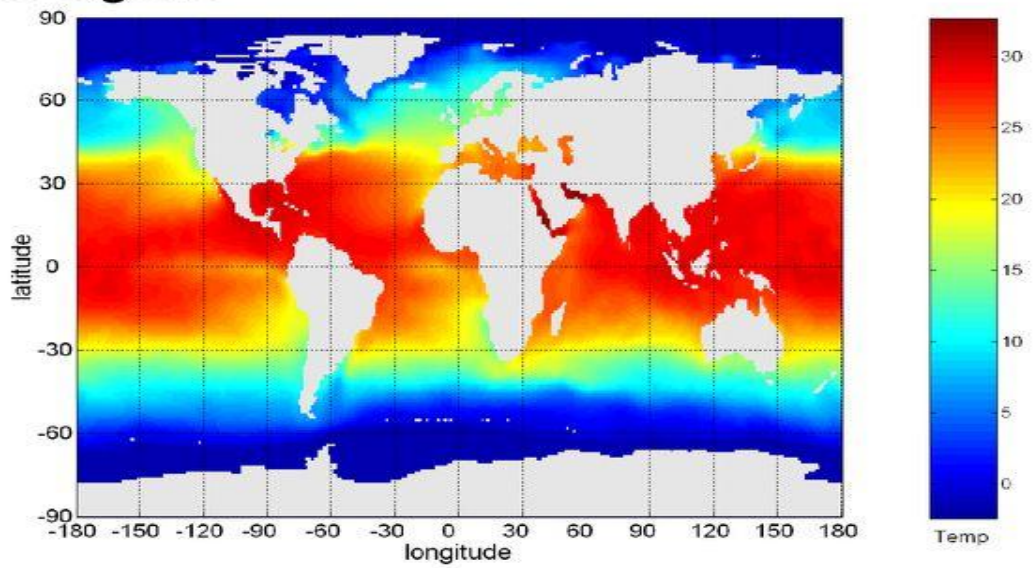
3. Visualization

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

- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Example: Sea Surface Temperature

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

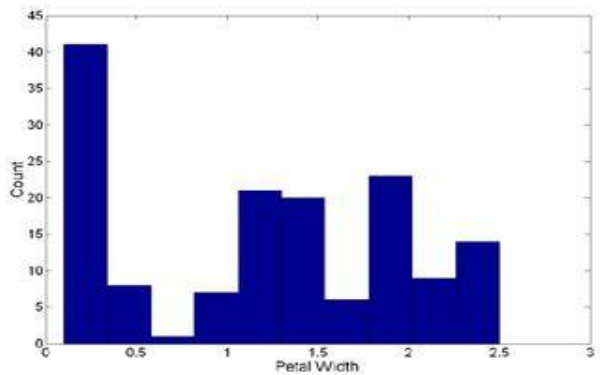


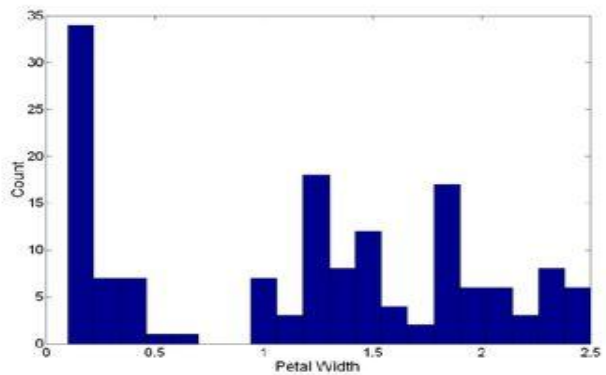
Representation (more on Vis. Sept. 9!)

- Is the mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

Visualization Techniques: Histograms

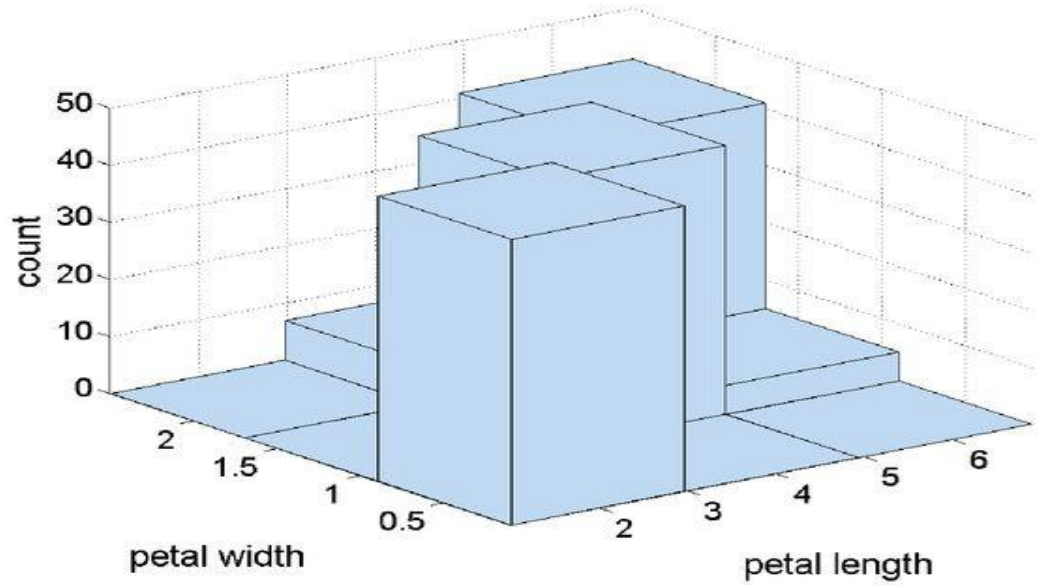
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width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





Two-Dimensional Histograms

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

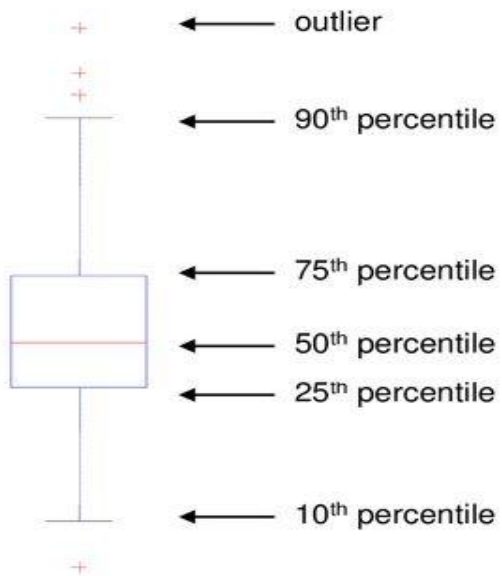


Visualization Techniques: Histograms

- Several variations of histograms exist: equi-bin(most popular), other approaches use variable bin sizes…
- Choosing proper bin-sizes and bin-starting points is a non trivial problem!!
- Example Problem from the midterm exam 2009: Assume you have an attribute A that has the attribute values that range between 0 and 6; its particular values are: 0.62 0.97 0.98 1.01. 1.02 1.07 2.96 2.97 2.99 3.02 3.03 3.06 4.96 4.97 4.98 5.02 5.03 5.04. Assume this attribute A is visualized as a equi-bin histogram with 6 bins: [0,1), [1,2), [2,3],[3,4), [4,5), [5,6]. Does the histogram provide a good approximation of the distribution/density function of attribute A? If not, provide a better histogram for attribute A. Give reasons for your answers! [7]
- https://en.wikipedia.org/wiki/Histogram

Visualization Techniques: Box Plots

- Box Plots (*we do not use the version depicted below!*)
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



Also see: https://en.wikipedia.org/wiki/Box_plot

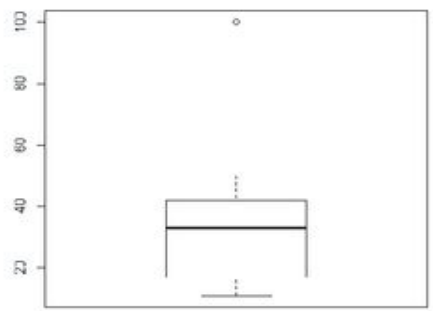
Boxplots in R (*we use those!!*)

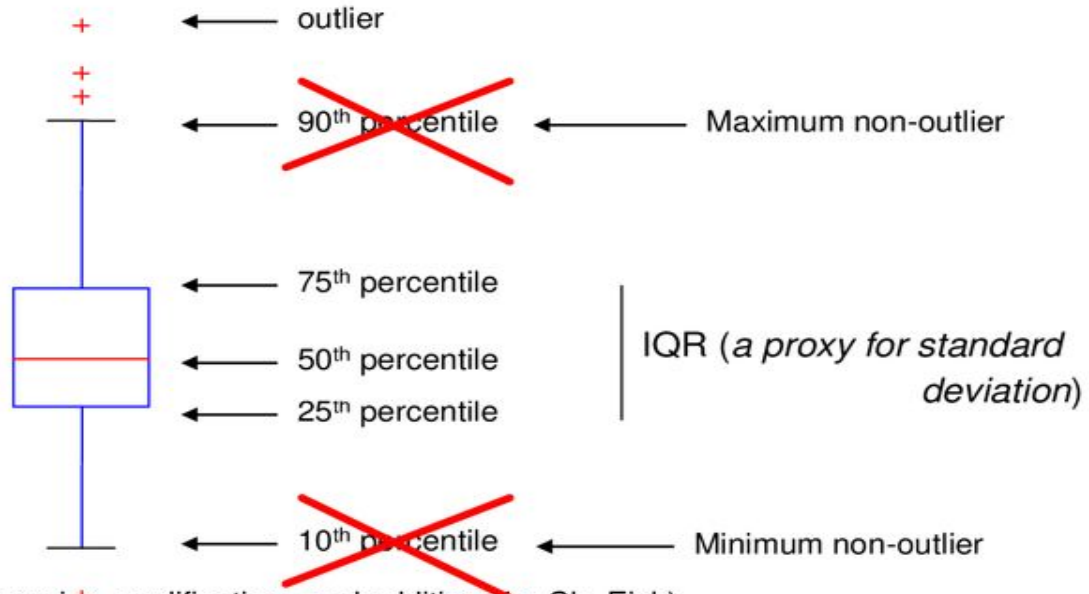
By default, boxplot() in R plots the maximum and the minimum non-outlying values instead of the 10th and 90th percentiles as the book describes. Outliers in BPs are values that are 1.5*IQR or more away from the box, where IQR is the height of the box!

See:

```
> a<-c(11,12, 22, 33, 34, 100)
> boxplot(a)
> b<-c(11,12, 22, 33, 34, 65)
> boxplot(b)
```

- a<-c(11,12,22, 33, 34, 50, 100)
- boxplot(a)

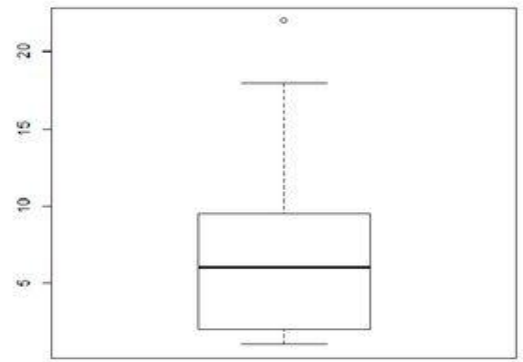




Example of R Box Plots *(Mid1 Question)*

b) The following boxplot has been created using the following R-code for an attribute x:

```
> x<-c(1,2,2,2,4,4,8,9,9,10,18,22)
> boxplot(x)
```



R version 3.4.3: *Kite-Eating Tree*

What is the median for the attribute x? What is the IQR for the attribute x? The lower whisker of the boxplot as at 1; what does this tell you? According the boxplot 18 is not an outlier and 22 as an outlier; why do you believe this is the case? [5]

Median is 6=(4+8)/2 [1]

IQR=9.5-2=7.5 [1]

1 is the lowest value in the dataset that is not an outlier [1] Every value that is 1.5∗IQR above the 75th percentile is an outlier; that is, for the particular boxplots values above 9.5+1.5*7.5=20.75 and below the 25th percentile -9.25 are outliers; consequently, 22 is an outlier and 1 and 18 is not, and the whiskers are therefore at 1 and 18! [2]

Attibute Standardization: Z-scores

- Attribute Standardization/Normalization→makes attributes equally import, alleviates impact of attribute scale
- Z-score standardization:
  - Calculate the mean m_f, the standard deviation s_f:
  - Calculate the standardized measurement (*z-score*)

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

- Result of Z-score standardization is a dataset in which each attribute has a mean of 0 and a standard deviation of 1.
- The obtained attribute values allow for statistical interpretation: e.g. if a person's z-scored age is -1 her age is on standard deviation below the average age…
- Z-scores can be interpreted based on the 68-95-99.7 Rule!

http://en.wikipedia.org/wiki/Standard_score

[0,1] Attribute Standardization

Approach: Normalize interval-scaled variables using

$$z_{if} = (x_{if} - \min_f) / ((\max_f - \min_f))$$

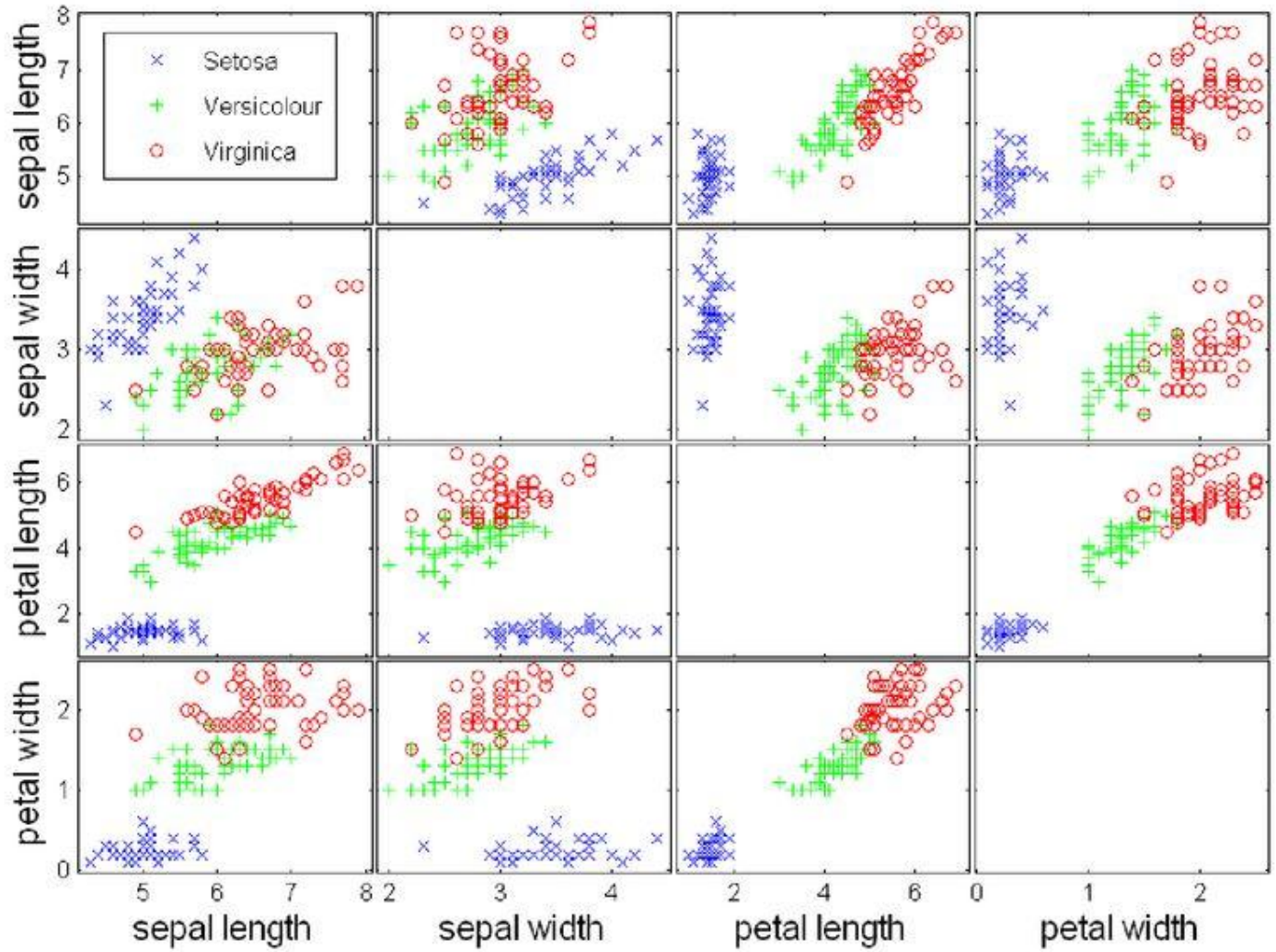
where $\min_f$ denotes the minimum value and $\max_f$ denotes the maximum value of the f-th attribute in the data set; that is, all values of the normalized dataset are numbers in [0,1].

Question: if the normalized value of an attribute is 0; what does this mean?

Visualization Techniques: Scatter Plots

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
    - For prediction scatter plots see:
      http://en.wikipedia.org/wiki/Scatter_plot
      http://en.wikipedia.org/wiki/Correlation (Correlation)
    - See example for classification, also called *supervised scatter plots*, on the next slide

Scatter Plot Array of Iris Attributes

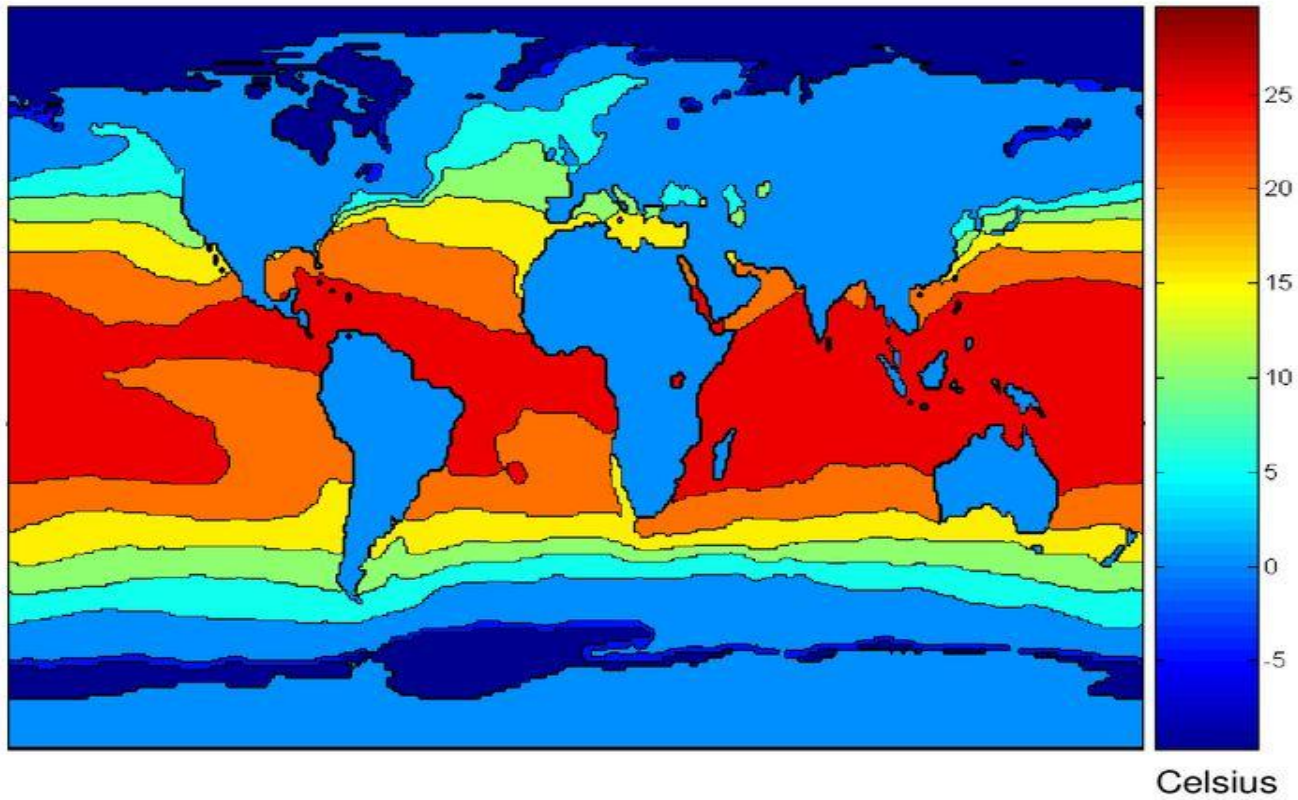


Visualization Techniques: Contour Plots

- Contour plots
  - Useful when a continuous attribute is measured on a spatial grid
  - They partition the plane into regions of similar values
  - The contour lines that form the boundaries of these regions connect points with equal values
  - The most common example is contour maps of elevation
  - Can also display temperature, rainfall, air pressure, etc.
    - An example for Sea Surface Temperature (SST) is provided on the next slide

https://plot.ly/r/contour-plots/

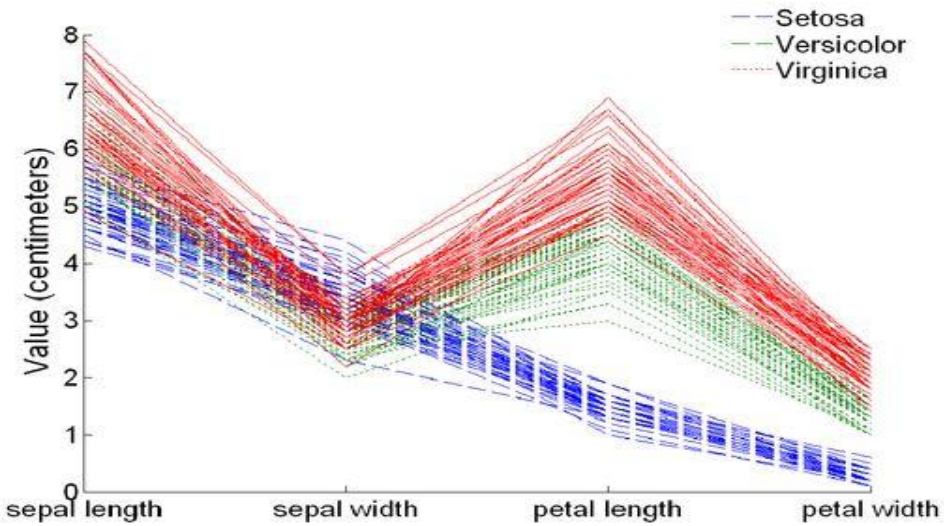
Contour Plot Example: SST Dec, 1998

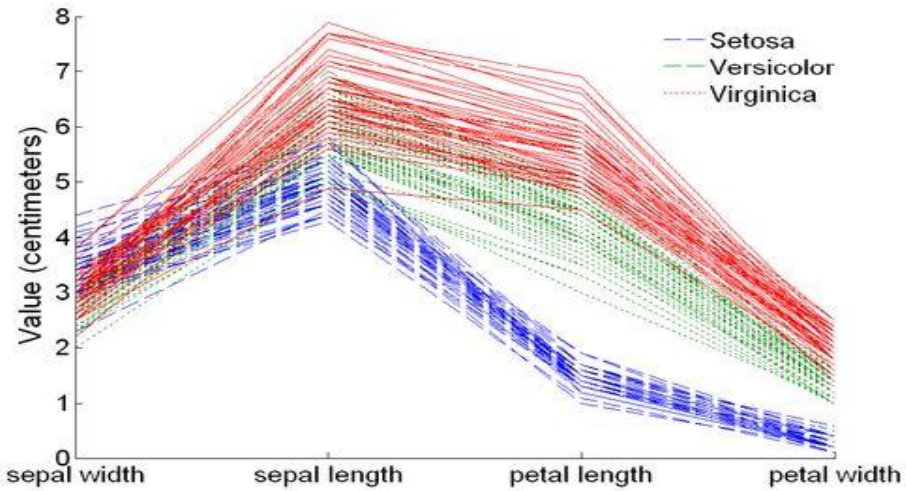


Visualization Techniques: Parallel Coordinates

- Parallel Coordinates
  - Used to plot the attribute values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus, each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some attributes
  - Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots for Iris Data

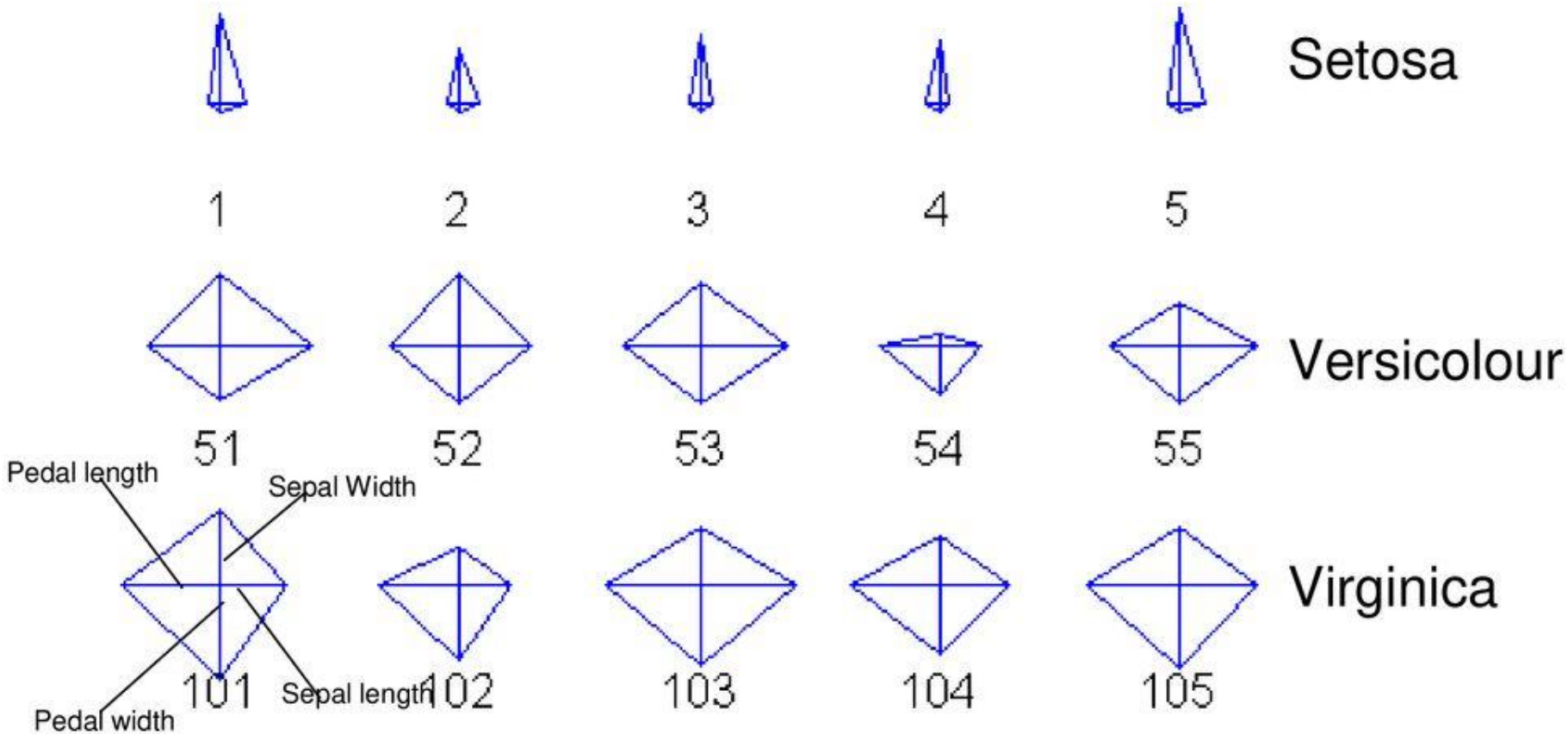




Other Visualization Techniques

- Star Coordinate Plots
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon
- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human's ability to distinguish faces
  - http://people.cs.uchicago.edu/~wiseman/chernoff/
  - http://kspark.kaist.ac.kr/Human%20Engineering.files/Chernoff/Chernoff%20Faces.htm#

Star Plots for Iris Data



Confidence Intervals

A **point estimate** of a parameter is the value of a statistic that estimates the value of the parameter.

The sample mean, $\bar{x}$, is the **best point estimate** of the population mean, μ.

A **confidence interval estimate** of a parameter consists of an interval of numbers along with a probability that the interval contains the unknown parameter.

The **level of confidence** in a confidence interval is a probability that represents the percentage of intervals that will contain if a large number of repeated samples are obtained. The level of confidence is denoted

For example, a 95% level of confidence would mean that if 100 confidence intervals were constructed, each based on a different sample from the same population, we would expect 95 of the intervals to contain the population mean.

The construction of a confidence interval for the population mean depends upon three factors

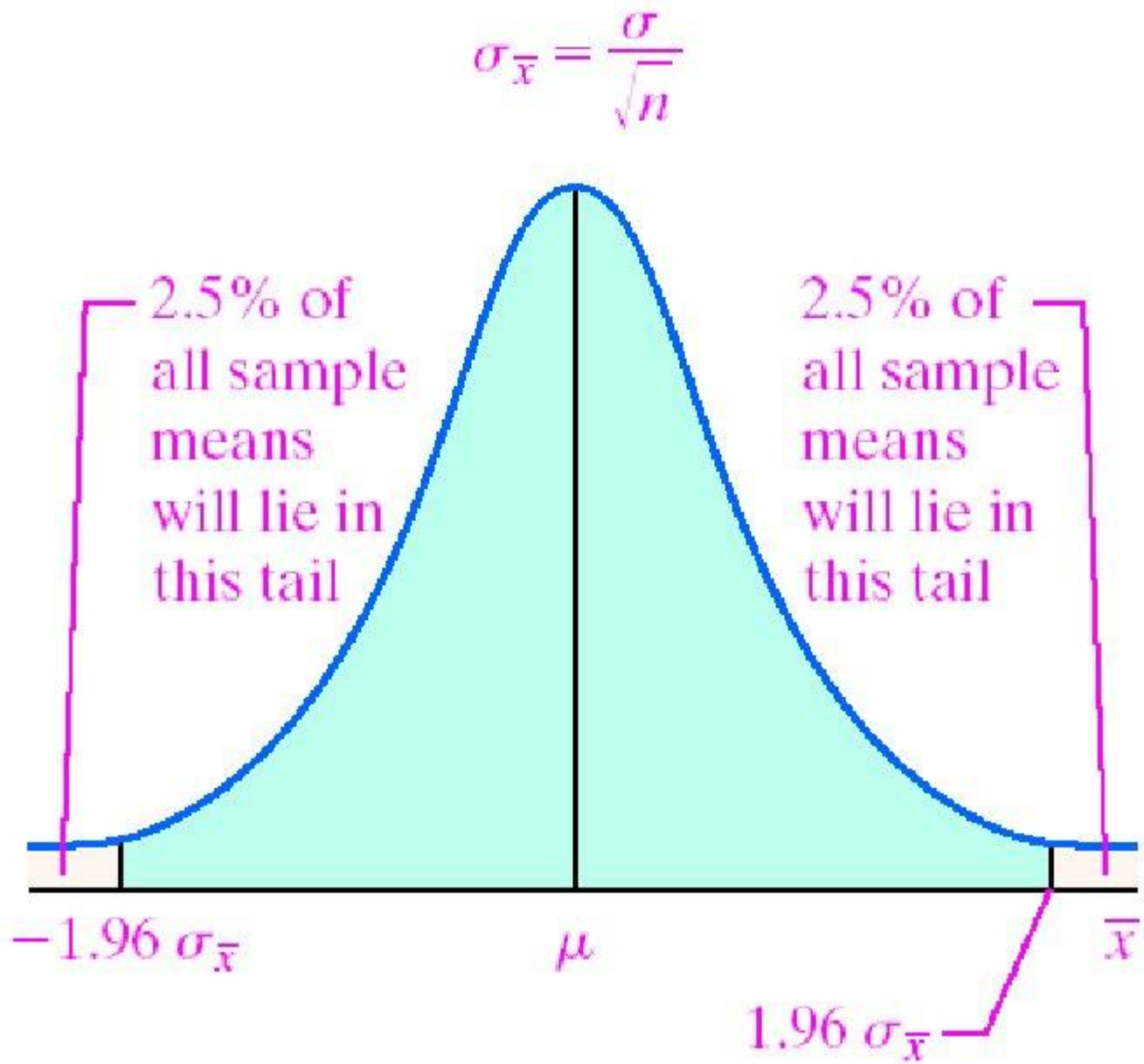
- The point estimate of the population
- The level of confidence
- The standard deviation of the sample mean $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Suppose we obtain a simple random sample from a population. Provided that the population is normally distributed or the sample size is large, the distribution of the sample mean will be normal with

$$\text{mean} = \mu$$

$$\text{and standard deviation} = \frac{\sigma}{\sqrt{n}}$$

Because $\overline{x}$ is normally distributed, we know that 95% of all sample means should lie within 1.96 standard deviations of the population mean, μ , and 2.5% will lie in each tail.


$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
2.5% of all sample means will lie in this tail
2.5% of all sample means will lie in this tail
$-1.96\ \sigma_{\bar{x}}$
μ
$\bar{x}$
$1.96\ \sigma_{\bar{x}}$

95% of all sample means are in the interval

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} < \overline{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

With a little algebraic manipulation, we can rewrite this inequality and obtain:

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

$$P(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Interpretation of a Confidence Interval

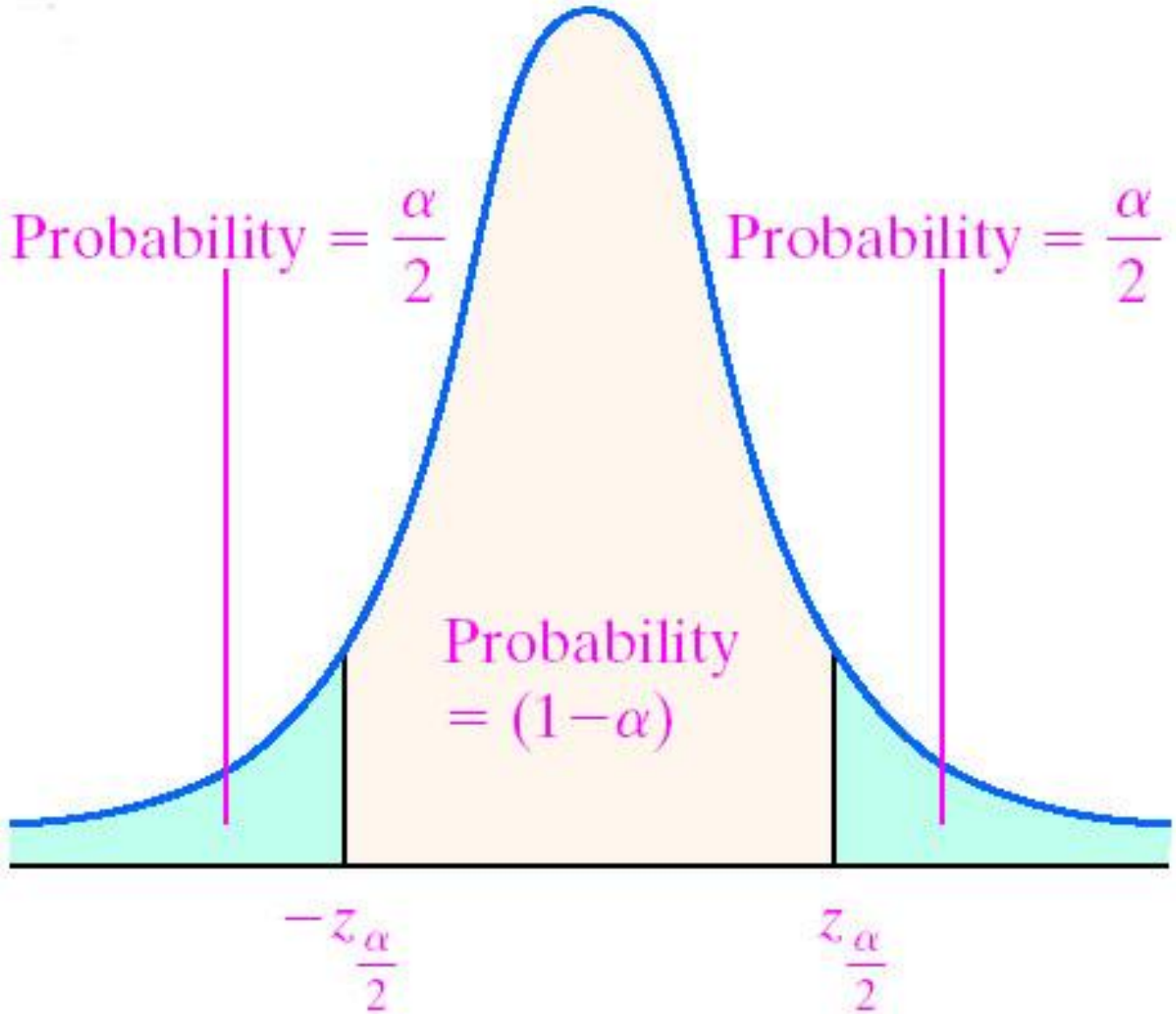
A $(1 - \alpha)100\%$ confidence interval means that if we obtained many simple random samples of size n from the population whose mean, μ, is unknown, then approximately $(1 - \alpha)100\%$ of the intervals will contain μ.

Constructing a (1 - α)100% Confidence Interval about μ, σ Known

Suppose a simple random sample of size n is taken from a population with unknown mean μ and known standard deviation σ. A (1 - α)100% confidence interval for is given by

Lower Bound:	$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	Upper Bound:	$\bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

NOTE: The size, n, of the population must be greater than or equal to 30 or the population must be normally distributed.


Probability = $\frac{\alpha}{2}$
Probability = $\frac{\alpha}{2}$
Probability
= $(1-\alpha)$
$-z_{\frac{\alpha}{2}}$
$z_{\frac{\alpha}{2}}$

The Margin of Error

The margin of error, E, in a $(1 - \alpha)100\%$ confidence interval in which σ is known is given by

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where n is the sample size.

NOTE: We require the population from which the sample was drawn be normally distributed or the sample size n be greater than or equal to 30.

What is a Hypothesis?

A hypothesis is an assumption about the population parameter.

A parameter is characteristic of the population, like its mean or variance.

The parameter must be identified before analysis.

Testing of Hypothesis

A hypothesis is an assumption about the population parameter (say population mean) which is to be tested.

For that we collect sample data , then we calculate sample statistics (say sample mean) and then use this information to judge/decide whether hypothesized value of population parameter is correct or not.

- To test the *validity of assumed or hypothetical value of population* , we gather sample data and determine the difference between hypothesized value and actual value of the sample mean.

- Then we judge whether the difference is significant or not.

- **The smaller the difference, the greater the likelihood that our hypothesized value for the mean is correct. The larger the difference, the smaller the likelihood.**

- In hypothesis testing **the first step** is to state the assumed or hypothesized(numerical) value of the population parameter.

- The assumption we wish/ want to test is called the *null hypothesis.* The symbol for *null hypothesis* is H_0

The Null Hypothesis, H_0

- State the Assumption (numerical) to be tested
- e.g. The average weight of the semester 2 student is 58kgs (H_0: $\mu = 58$)
- Begin with the assumption that the null hypothesis is TRUE.

(Similar to the notion of innocent until proven guilty)

The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis

eg. The average weight of the students is not equal to 58kgs. (H_1: $\mu \neq 58$)

Procedure of Hypothesis Testing

The Hypothesis Testing comprises the following steps:

Step 1

Set up a hypothesis.

Step 2

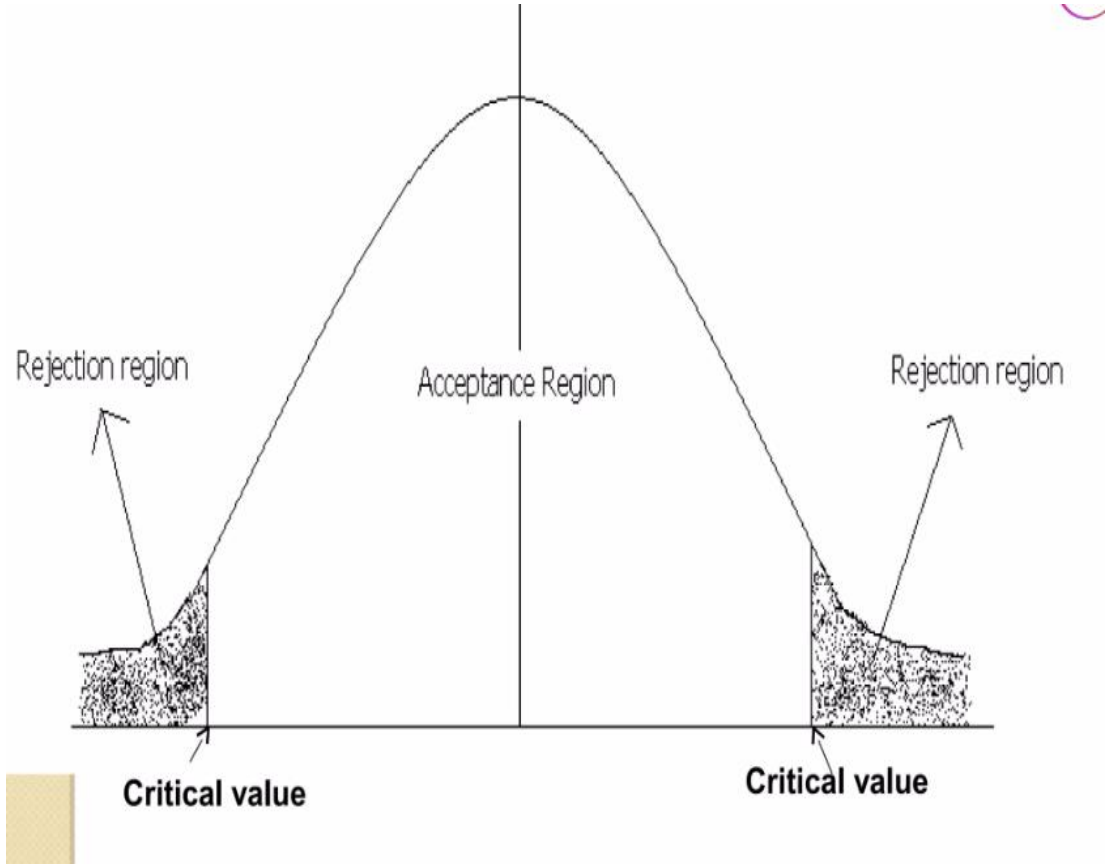
Set up a suitable significance level.

The confidence with which an experimenter rejects or accepts Null Hypothesis depends on the significance level adopted**.** Level of significance is the **rejection region (** which is outside the confidence or acceptance region).The level of significance, usually denoted by the **α.**

Selecting a significance level

Though any level of significance can be adopted, **in practice we either take 5% or 1% level of significance .**

When we take 5% level of significance(α= .05), then there are about 5 chances out of 100 that we would reject the null hypothesis. In other words out of 100, 95% chances are there that the null hypothesis will be accepted i.e. we are about 95% confident that we have made the right decision.



If our sample statistic(calculated value) fall in the non-shaded region(acceptance region), then it simply means that there is no evidence to reject the ***null hypothesis.***

It proves that null hypothesis (H_0) is true. Otherwise, it will be rejected.

Step 3

Determination of suitable test statistic: For example Z, t Chi-Square or F-statistic.

Step 4

Determine the critical value from the table.

Step 5

After doing computation, check the sample result.

Compare the calculated value(sample result) with the value obtained from the table.(tabulated or critical value)

Step 6

Making Decisions

Making decisions means either accepting or rejecting the null hypothesis.

If computed value(absolute value) is more than the tabulated or critical value, then it falls in the critical region. In that case, reject null hypothesis, otherwise accept.

Type I and Type II Errors

When a statistical hypothesis is tested, there are 4 possible results:

(1)The hypothesis is true but our test accepts it.
(2)The hypothesis is false but our test rejects it.
(3)The hypothesis is true but our test rejects it.
(4)The hypothesis is false but our test accepts it.

Obviously, the last 2 possibilities lead to errors.

Rejecting a null hypothesis when it is true is called a ***Type I error.***
Accepting a null hypothesis when it is false is called ***Type II error.***

Example 1 - Court Room Trial

In court room, a defendant is considered not guilty as long as his guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough charging evidence the defendant is condemned. In the start of the procedure, there are two hypotheses H_0: "the defendant is not guilty", and H_1: "the defendant is guilty". The first one is called *null hypothesis*, and the second one is called *alternative (hypothesis)*.

	Null Hypothesis (H_0) is true He is not guilty	Alternative Hypothesis (H_1) is true He is guilty
Accept Null Hypothesis	Right decision	Wrong decision Type II Error
Reject Null Hypothesis	Wrong decision Type I Error	Right decision

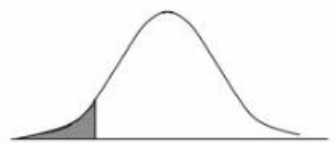
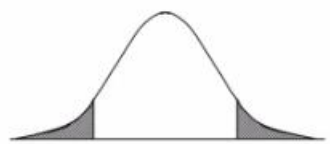
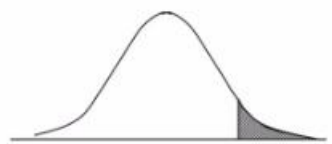
One-Tailed and Two-Tailed Tests

Two-Tailed Test is that where the hypothesis about the population parameter is rejected for the value of sample statistic failing into **either tail** of the distribution.(fig3)

When the hypothesis about the population parameter is rejected for the value of sample statistic failing into **one side tail** of the distribution, then it is known as **one-tailed test.**

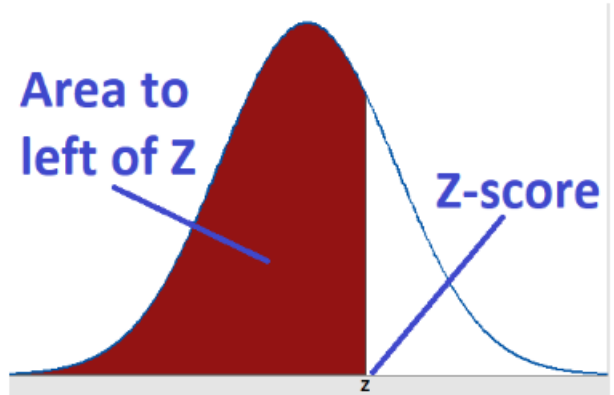
If the rejection area falls on the right side, then it is called right-tailed test.(fig 2) On the other hand If the rejection area falls on the right side, then it is called left-tailed test.(fig 1)

Summary of One- and Two-Tail Tests…

One-Tail Test (left tail)	Two-Tail Test	One-Tail Test (right tail)
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$
		

The following table gives critical values of Z for both one-tailed and two- tailed tests at various levels of significance.

level of significance	0.10	.05	.01
critical value of z for one-tailed test	1.28 or -1.28	1.645 or -1.645	2.33 0r -2.33
critical value of z for two-tailed test	1.645 or -1.645	1.96 or -1.96	2.58 or -2.58



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
-3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
-3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
-3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
-3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
-2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
-2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
-2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
-2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
-2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
-2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
-2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
-1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
-1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
-1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
-1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
-1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
-1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
-1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
-1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
-1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
-1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
-0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
-0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
-0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
-0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995

Where Hypothesis Testing can be applied?

1. Testing the effectiveness of interventions or treatments.
2. Comparing means or proportions.
3. Analysing relationships between variables.
4. Evaluating the goodness of fit
5. Testing the independence of categorical variables.
6. A/B testing

Hypothesis Testing ML Applications

1. **Model comparison**
2. **Feature selection**
3. **Hyperparameter tuning**
4. **Assessing model assumptions**

Hypothesis Testing

- Performing a Z test Example 1 (Rejection Region Approach)

- Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employee before implementing the training program. The average productivity was 50 units per day with a known population standard deviation of 5 units. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average productivity of 53 units per day. The company wants to know if the new training program has significantly increased productivity.

Example 2

- Suppose a snack food company claims that their lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from claimed 50 grams. The organization collects a random sample of 40 lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a known population standard deviation of 4 grams.

P-value

•**Definition:** It is the probability of obtaining a test statistic as extreme or more extreme than the observed result, assuming $H0$ is true.

•**Decision Rule:** Compare the p-value to a pre-determined significance level (α), usually 0.05.
p≤α: Reject the null hypothesis (statistically significant)
p>α : Fail to reject the null hypothesis (not statistically significant).

•**Interpretation:** A p-value of 0.03 means there is a 3% chance of seeing the observed data (or more extreme) if the null hypothesis is true.

•**Purpose:** It quantifies the strength of evidence against the null hypothesis, helping determine if results are due to a real effect or random variation.

•**Limitations:** A low p-value does not prove the null hypothesis is false, only that it is unlikely given the data. It also does not measure the size of the effect.
Common Thresholds:
• **p<0.05:** Statistically significant.
• **p<0.01:** Highly significant.
• **p>0.05:** Not significant; fail to reject null hypothesis.

P-value in context of Z-test

- Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was 50 units per day. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average productivity of 53 units per day and the sample std is 4. The company wants to know if new training program has significantly increased productivity.

Example 2

- Suppose a snack company claims that their lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They finds that the sample has an average weight of 49 grams, with a population standard deviation of 5 grams.

Single Sample t-test

- It checks whether a sample mean differs from the population mean.

- Assumptions for a single sample t-test
- 1. Normality – Population from which the sample is drawn is normally distributed.
- 2. Independence- The observation in the sample must be independent, which means that the value of the observation should not influence the value of another observation.
- 3. Random sampling – The sample must be a random and representative subset of the population.
- 4. Unknown population std
- Smaller sample size than 30.

One-way ANOVA Table

Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{n}$ $SSB = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X}_t)^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups (Error)	n-k	$SSW = \sum_{j-1}^{K}\sum_{i-1}^{n} X_{ij}^2 - \sum_{j-1}^{K}\left(\frac{T_j^2}{n_j}\right)$ $SSW = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$	$MSW = \frac{SSW}{n-k}$	
Total	n-1	$SST = \sum_{j=1}^{K}\sum_{i=1}^{n} X_{ij}^2 - \frac{T^2}{n}$ $SST = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_t)^2$		

- SST = SSB + SSW

k: number of groups n: number of samples

df: degree of freedom

Example 1: Three types of fertilizers are used on three groups of plants for 5 weeks. We want to check if there is a difference in the mean growth of each group. Using the data given below apply a one way ANOVA test at 0.05 significant level.

Fertilizer 1	Fertilizer 2	Fertilizer 3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Solution:

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1: The means are not equal

Fertilizer 1	Fertilizer 2	Fertilizer 3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12
$\overline{X}_1 = 5$	$\overline{X}_1 = 9$	$\overline{X}_1 = 10$

Total mean, $\overline{\overline{X}} = 8$

$n_1 = n_2 = n_3 = 6, k = 3$

$SSB = 6(5 - 8)^2 + 6(9 - 8)^2 + 6(10 - 8)^2$

$= 84$

$df_1 = k - 1 = 2$

Fertilizer 1	$(X - 5)^2$	Fertilizer 2	$(X - 9)^2$	Fertilizer 3	$(X - 10)^2$
6	1	8	1	13	9
8	9	12	9	9	1
4	1	9	0	11	1
5	0	11	4	8	4
3	4	6	9	7	9
4	1	8	1	12	4
$\overline{X}_1 = 5$	Total = 16	$\overline{X}_1 = 9$	Total = 24	$\overline{X}_1 = 10$	Total = 28

$SSE = 16 + 24 + 28 = 68$

$N = 18$

$df_2 = N - k = 18 - 3 = 15$

$MSB = SSB / df_1 = 84 / 2 = 42$

$MSE = SSE / df_2 = 68 / 15 = 4.53$

ANOVA test statistic, $f = MSB / MSE = 42 / 4.53 = 9.33$

Using the f table at $\alpha = 0.05$ the critical value is given as $F(0.05, 2, 15) = 3.68$

As $f > F$, thus, the null hypothesis is rejected and it can be concluded that there is a difference in the mean growth of the plants.

Answer: Reject the null hypothesis

ANOVA

- Disadvantages
- **Assumption of Normality**: ANOVA assumes that the data for each group follow a normal distribution. This assumption may not hold true for all datasets, especially those with skewed distributions
- **Assumption of Homogeneity of Variance**: ANOVA assumes that the variances of the different groups are equal. This is the assumption of homogeneity of variance (also known as homoscedasticity). If this assumption is violated, it may lead to incorrect results.
- **Independence of Observations**: ANOVA assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (e.g., time series data, nested data).
- **Effect of Outliers**: ANOVA is sensitive to outliers. A single outlier can significantly affect the F-statistic leading to a potentially erroneous conclusion.
- **Doesn't Account for Interactions**: Just like other univariate feature selection methods, ANOVA does not consider interactions between features.

Example 2: A trial was run to check the effects of different diets. Positive numbers indicate weight loss and negative numbers indicate weight gain. Check if there is an average difference in the weight of people following different diets using an ANOVA Table.

Low Fat	Low Calorie	Low Protein	Low Carbohydrate
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3

F-table of Critical Values of $\alpha = 0.05$ for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Chi-square Test

- A non-parametric test that is used to measure the association between two categorical variables.
- We use it when we have observed frequencies of a categorial variable (e.g., male vs female, healthy vs sick).

Chi-square Test

There are 2 types of chi-square tests:

- The **chi-square goodness of fit test** is used to compare the observed frequencies in a data sample with the frequencies based on some prior expectation – either empirical or theoretical.
- The **chi-square test of independence** assesses whether observed frequencies are dependent on (i.e., contingent on) certain conditions.

- **Step 1: Define Hypothesis**

- **Null Hypothesis (H_0):** The relationship between categorical variables is determined by the use of statistical analysis. This means the researcher assumes that there is no relationship between the two variables under study no matter the differences or patterns identified are as a result of random chance. Observing this hypothesis helps us protect our analysis from possible prejudices hence ensuring it is just.

- **Alternative Hypothesis (H_1):** The hypothesis suggests that there is a relation between the two categorical independent variables which are under study, therefore showing that there is an actual relationship instead of mere coincidence.

Chi-square frequency tables

- A frequency distribution table / contingency table shows how observations are distributed between different groups (i.e., the number of observations in each group).

Example: Bird species at a bird feeder

Frequency of visits by bird species at a bird feeder during a 24-hour period

Bird species	Frequency
House sparrow	15
House finch	12
Black-capped chickadee	9
Common grackle	8
European starling	8
Mourning dove	6

https://www.scribbr.com/statistics/chi-square-tests/

A **chi-square goodness of fit test** can test whether these observed frequencies are significantly different from what was expected, such as equal frequencies.

- **Null hypothesis (H_0):** The bird species visit the feeder in the **same** proportions as the average over the past five years.
- **Alternative hypothesis (H_1):** The bird species visit the feeder in **different** proportions from the average over the past five years.

Chi-square frequency tables

- A frequency distribution table / contingency table shows how observations are distributed between different groups (i.e., the number of observations in each group).

Example: Handedness and nationality

Contingency table of the handedness of a sample of Americans and Canadians

	Right-handed	Left-handed
American	236	19
Canadian	157	16

A **chi-square test of independence** can test whether these observed frequencies are significantly different from the frequencies expected if handedness is unrelated to nationality.

- **Null hypothesis (H_0):** The proportion of people who are left-handed is **the same** for Americans and Canadians.
- **Alternative hypothesis (H_1):** The proportion of people who are left-handed **differs** between nationalities.

https://www.scribbr.com/statistics/chi-square-tests/

Assumptions of the Chi-square

Since we only require that the two variables are categorical, there are no assumptions required on the variables in the Chi-Square test.

Chi Square Table

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Problem 1: Suppose we want to know if gender has anything to do with political party preference. So, we poll 440 voters in a simple random sample to find out which political party they prefer. The results of the survey are provided in the table below.

–	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	440

H_0 = There is no link between gender and political party preference.

H_1 = There is a link between gender and political party preference.

Now we calculate the expected frequency.

$$Expected\ Value = \frac{(Row\ Total) \times (Column\ Total)}{Total\ Number\ of\ Observations}$$

For one case such as,

Expected Value of Male Republican = $\frac{240 \times 200}{440} = 109$

Similarly, we calculate the expected value for each of the cells such as,

–	Republican	Democrat	Independent	Total
Male	109	59	22.72	200
Female	120	65	25	220
Total	240	130	50	440

- Use the chi-square formula: $X^2 = \Sigma ((O - E)^2 / E)$
- Here, O is the observed frequency and E is the expected frequency.

–	Republican	Democrat	Independent	Total
Male	0.7431197	2.050847	2.332676056	200
Female	3.3333333	0.384615	1	220
Total	240	130	50	440

Finally we calculate the test statistics X^2 which is the sum of the cell values from the above table.

$$X^2 = 0.743 + 2.05 + 2.33 + 3.33 + 0.384 + 1 = 9.837$$

Next we have to calculate the degrees of freedom such as,

$$(r - 1) \times (c - 1) = (3 - 1) \times (2 - 1) = 2$$

Where, r = number of column items, and

c = number of row items.

Now we compare our obtained statistic to the critical statistic found in the Chi Square table. As we can see, for an alpha level of 0.05 and 2 degrees of freedom, the critical statistic shown is 5.991, which is less than our obtained statistics of 9.83.

Therefore we can reject the null hypothesis because the critical statistics is higher than the obtained statistic.

Chi-square test

- Disadvantages
- **Categorical Data Only**: The chi-square test can only be used with categorical variables. It is not suitable for continuous variables unless they have been discretized into categories, which can lead to loss of information.
- **Independence of Observations:** The chi-square test assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (e.g., time series data, nested data).
- **Sufficient Sample Size:** Chi-square test requires a sufficiently large sample size. The results may not be reliable if the sample size is too small or if the frequency count in any category is too low (typically less than 5).
- **No Variable Interactions:** Chi-square test, like other univariate feature selection methods, does not consider interactions between features. It might miss out on identifying important features that are significant in combination with other features.

Bias-variance Trade-off

Generalization

Training set (labels known)

Test set (labels unknown)

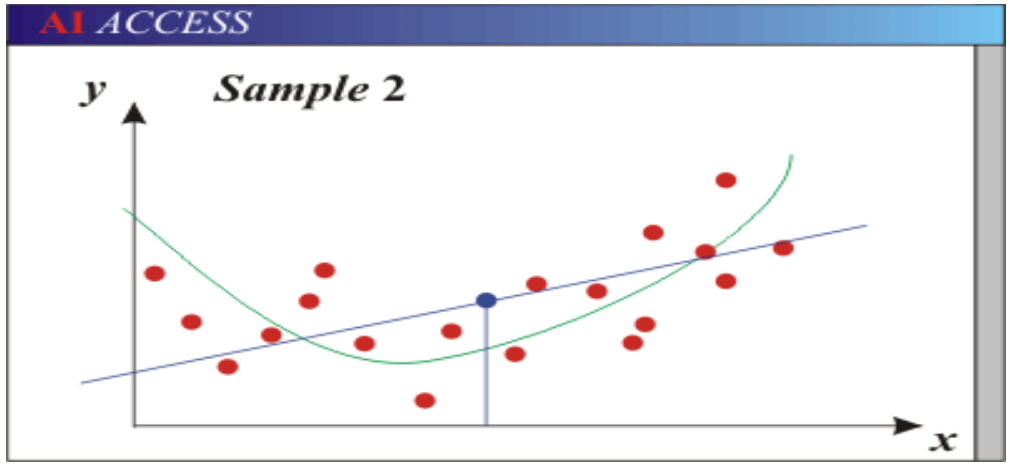
- How well does a learned model generalize from the data it was trained on to a new test set?

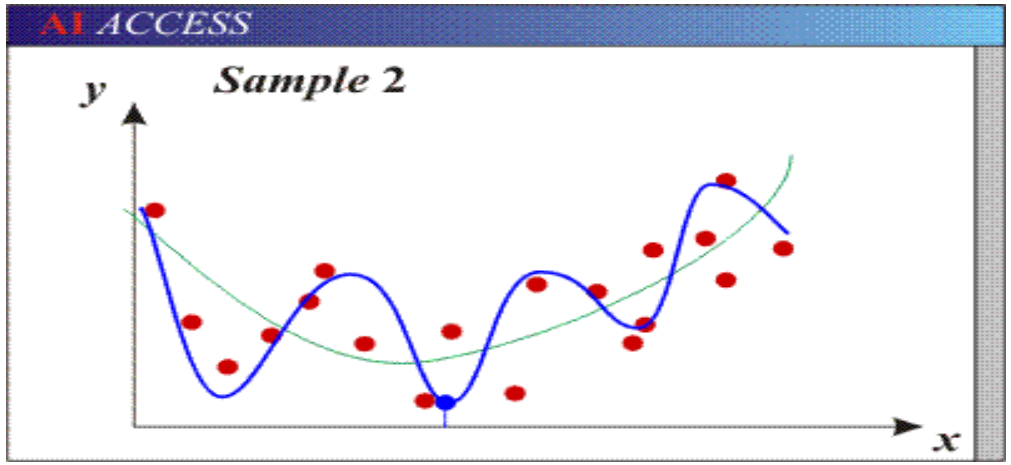
Generalization

- Components of generalization error
  - **Bias:** how much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - **Variance:** how much models estimated from different training sets differ from each other
- **Underfitting:** model is too “simple” to represent all the relevant class characteristics
  - High bias and low variance
  - High training error and high test error
- **Overfitting:** model is too “complex” and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error

Slide credit: L. Lazebnik

Bias-Variance Trade-off



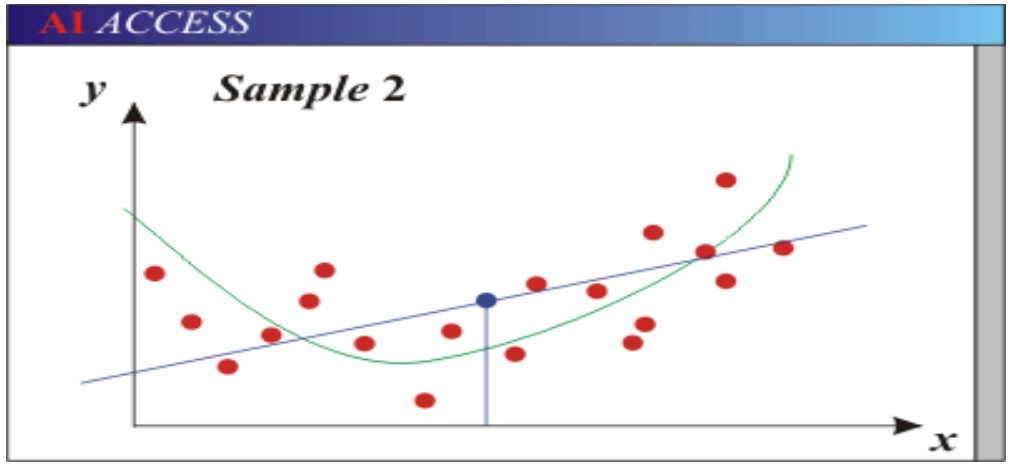


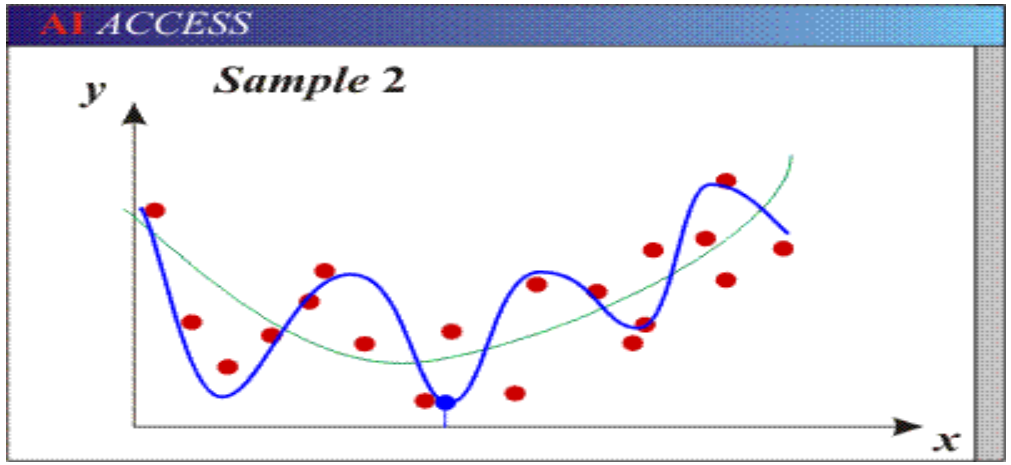
Models with too few parameters are inaccurate because of a large bias (not enough flexibility).

Models with too many parameters are inaccurate because of a large variance (too much sensitivity to the sample).

Slide credit: D. Hoiem

Bias-Variance Trade-off





- Models with too few parameters are inaccurate because of a large bias (not enough flexibility).
- Models with too many parameters are inaccurate because of a large variance (too much sensitivity to the sample).

Slide credit: D. Hoiem

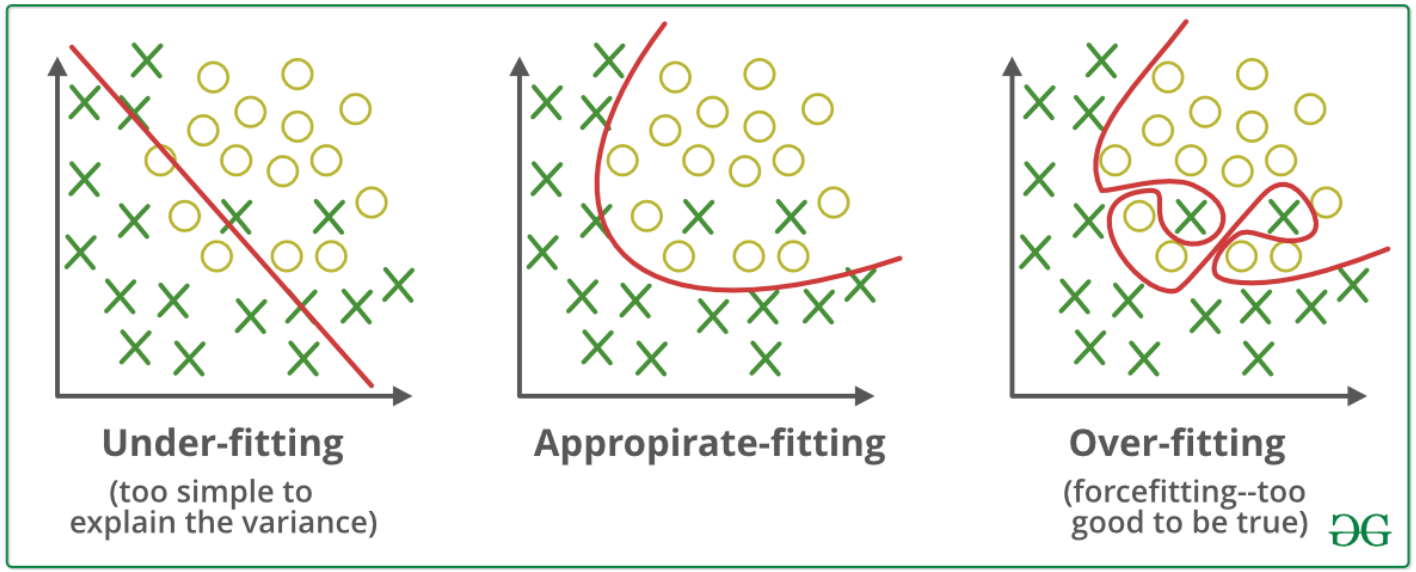
Bias-Variance Trade-off

$$E(MSE) = noise^2 + bias^2 + variance$$

Unavoidable error

Error due to incorrect assumptions

Error due to variance of training samples

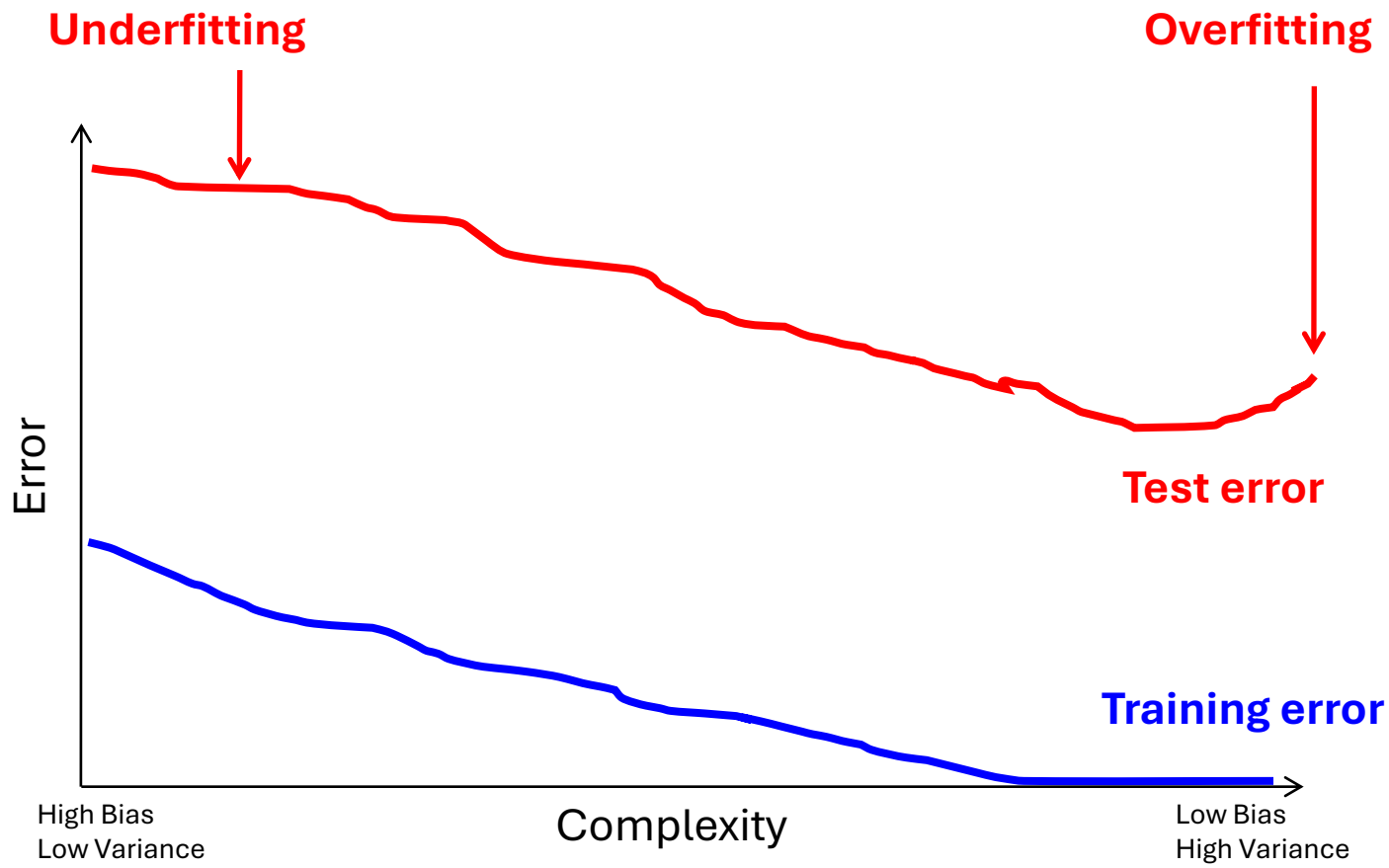


See the following for explanations of bias-variance (also Bishop's "Neural Networks" book):

•http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture4/BiasVariance.pdf

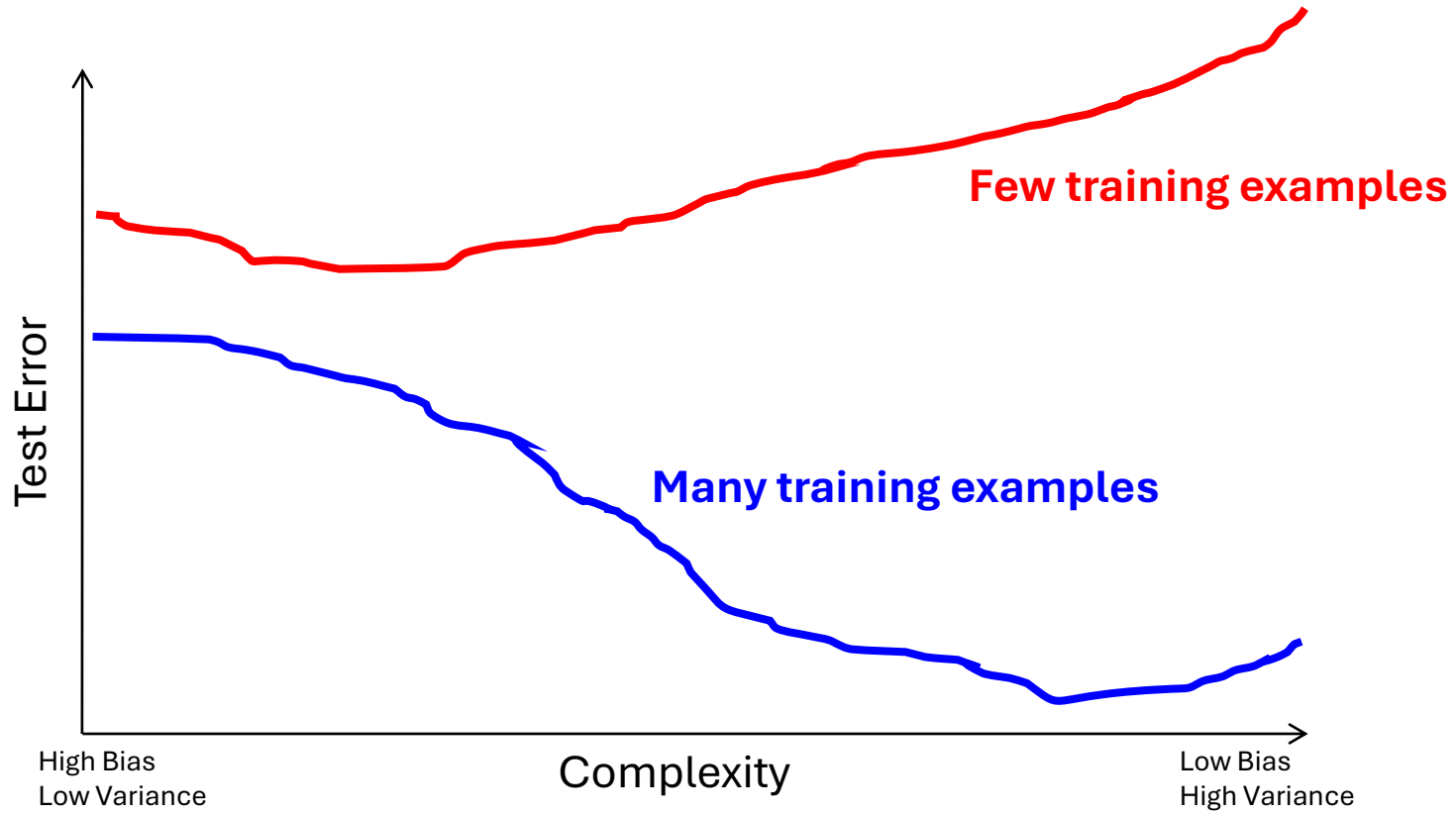
Image credit: geeksforgeeks.com
Slide credit: D. Hoiem

Bias-variance tradeoff



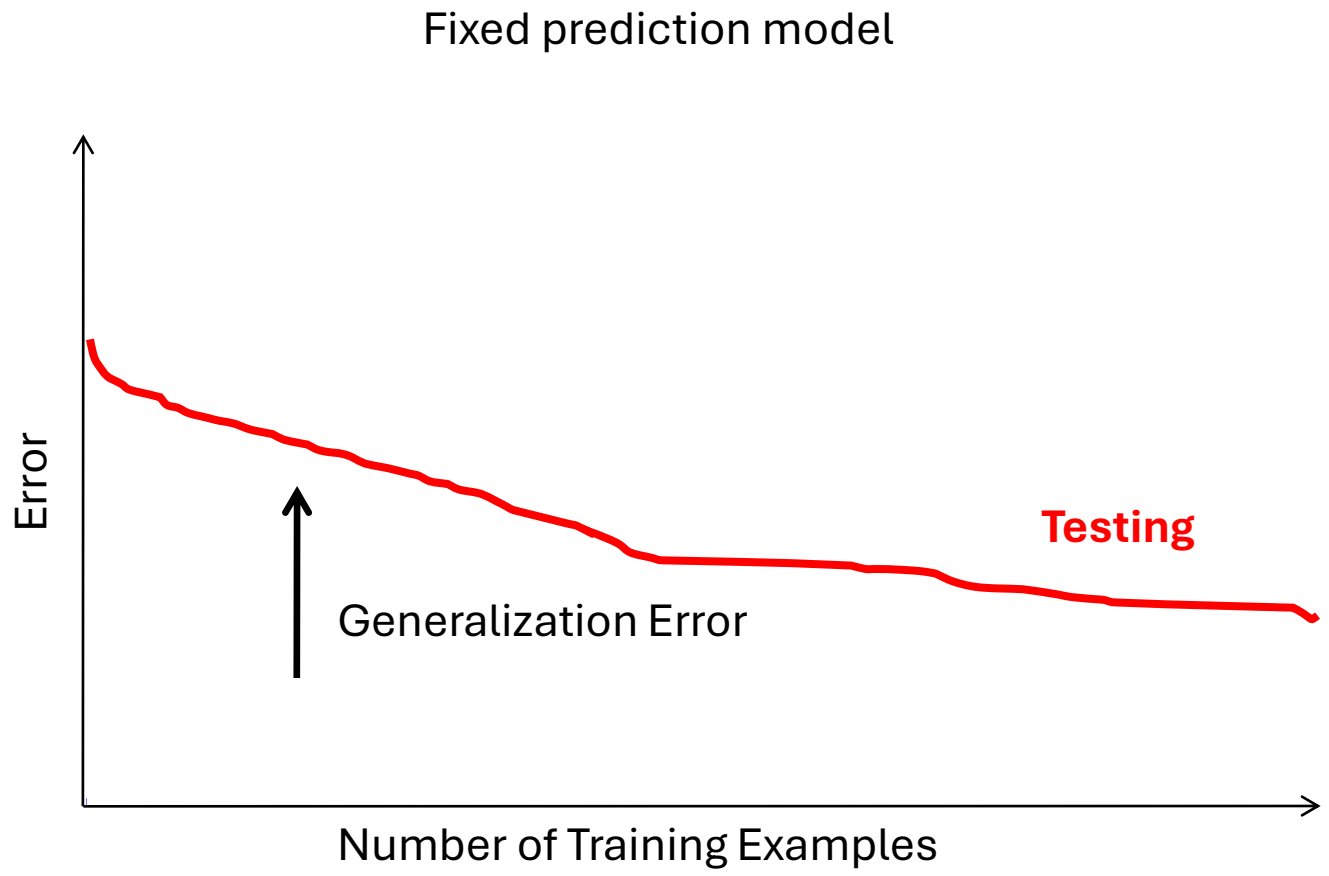
Slide credit: D. Hoiem

Bias-variance tradeoff



Slide credit: D. Hoiem

Effect of Training Size



Slide credit: D. Hoiem

The perfect classification algorithm

- Objective function: encodes the right loss for the problem
- Parameterization: makes assumptions that fit the problem
- Regularization: right level of regularization for amount of training data
- Training algorithm: can find parameters that maximize objective on training set
- Inference algorithm: can solve for objective function in evaluation

Slide credit: D. Hoiem

Remember…

- No classifier is inherently better than any other: you need to make assumptions to generalize

- Three kinds of error
  - Inherent: unavoidable
  - Bias: due to over-simplifications
  - Variance: due to inability to perfectly estimate parameters from limited data

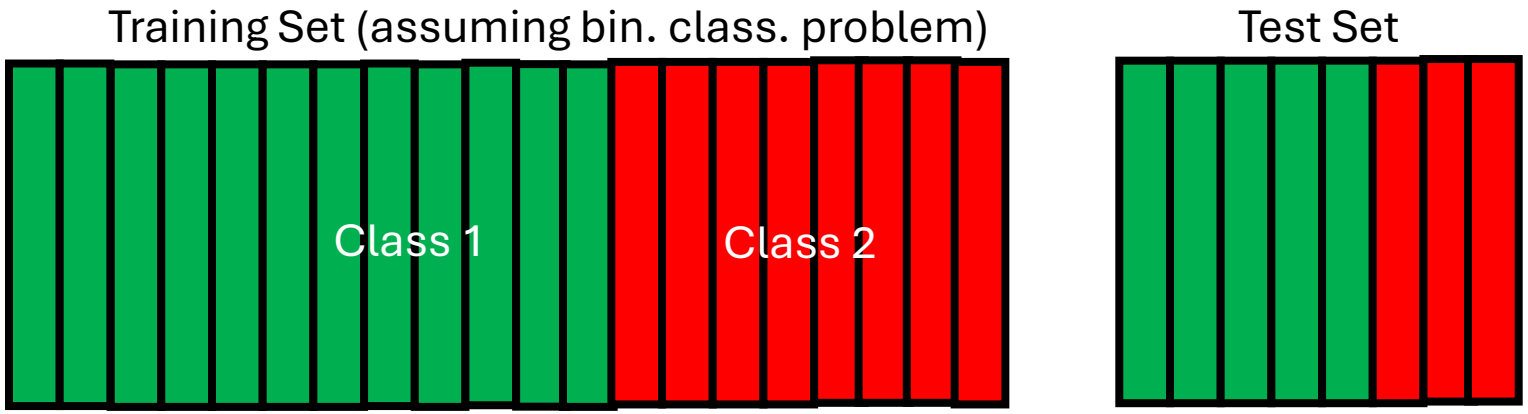
Slide credit: D. Hoiem

How to reduce variance?

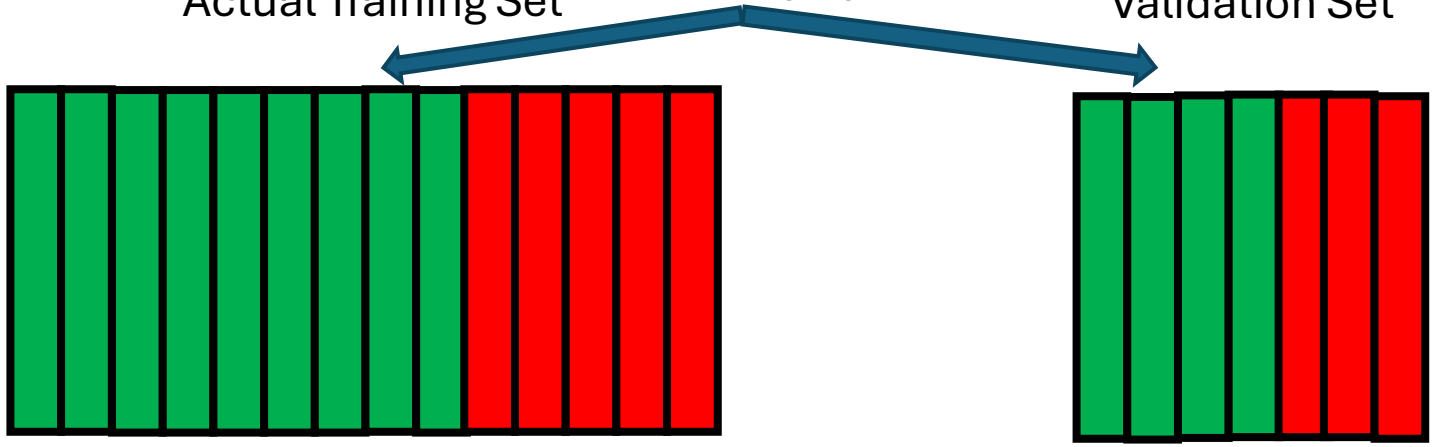
- Choose a simpler classifier
- Cross-validate the parameters
- Get more training data

Cross-Validation

No peeking while building the model



Note: Not just h.p. selection; we can also use CV to pick the best ML model from a set of different ML models (e.g., say we have to pick between two models we may have trained - LwP and nearest neighbors. Can use CV to choose the better one.



Randomly split the original training data into actual training set and validation set. Using the actual training set, train several times, each time using a different value of the hyperparam. Pick the hyperparam value that gives best accuracy on the validation set

What if the random split is unlucky (i.e., validation data is not like test data)?

If you fear an unlucky split, try multiple splits. Pick the hyperparam value that gives the best average CV accuracy across all such splits. If you are using N splits, this is called N–fold cross validation

Parametric and Non-parametric

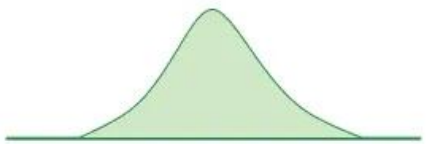
- Parametric and non-parametric tests are named based on whether they require assumptions about the underlying population parameters (like mean and variance). **Parametric tests** rely on specific distributional assumptions (usually normal distribution) and specific parameters, while **non-parametric tests** make no such assumptions, acting as distribution-free methods

Parametric vs. Non-Parametric Tests

Parametric Tests

Assumptions :

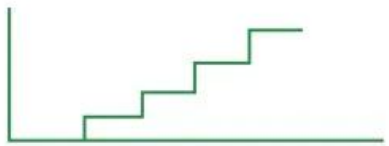
- Normality
- Homogeneity of variance
- Independence



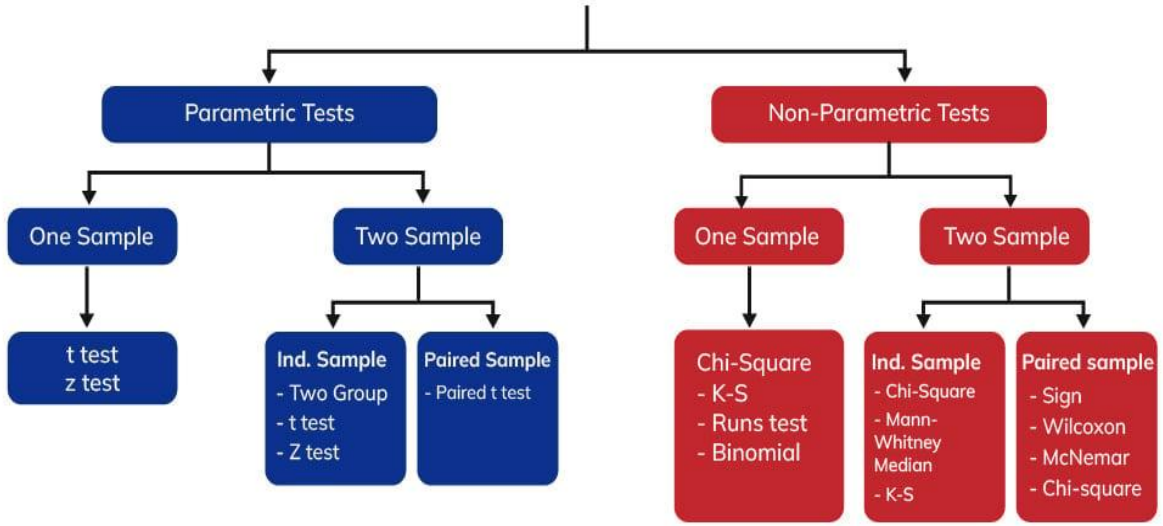
Non-Parametric Tests

When to Use :

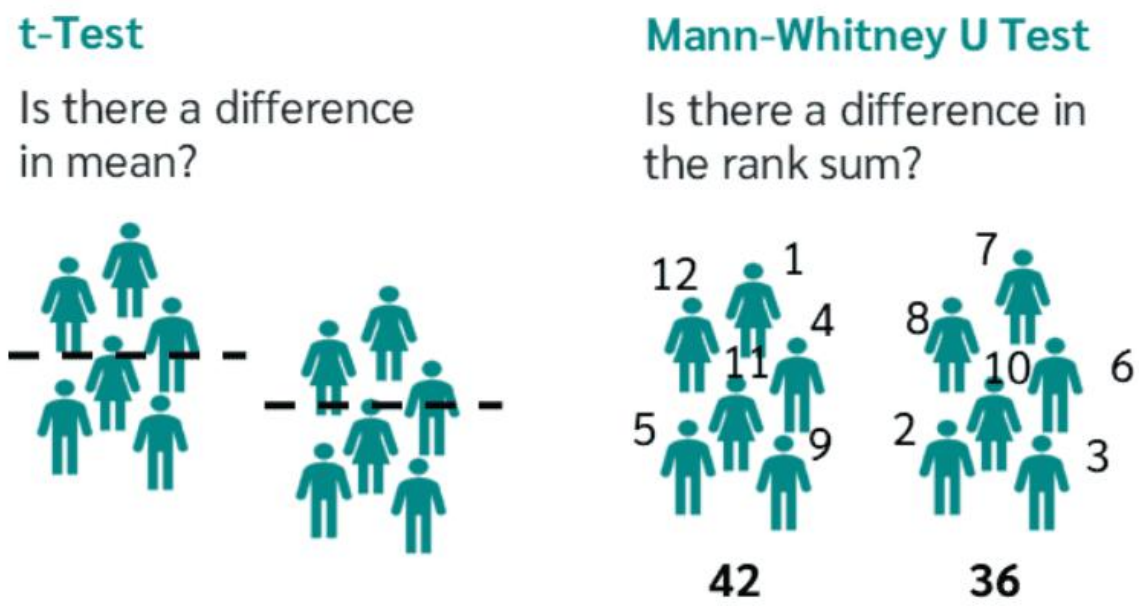
- Data is not normally distributed
- Small sample size
- Ordinal/nominal data
- Presence of outliers



Parametric & Non-Parametric Test



Mann-Whitney U Test



- **Null hypothesis:** There is no difference (in terms of central tendency) between the two groups in the population.
- **Alternative hypothesis:** There is a difference (with respect to the central tendency) between the two groups in the population.

Gender	Reaction time	Rang
female	34	2
female	36	4
female	41	7
female	43	9
female	44	10
female	37	5
male	45	11
male	33	1
male	35	3
male	39	6
male	42	8

Calculation of the rank sums

$T_1 = 2 + 4 + 7 + 9 + 10 + 5 = 37$

$T_2 = 11 + 1 + 3 + 6 + 8 = 29$

Female

Number of cases $n_1 = 6$ Rank sum $T_1 = 37$

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - T_1$$
$$= 6 \cdot 5 + \frac{6 \cdot (6 + 1)}{2} - 37$$
$$= 14$$

Male

Number of cases $n_2 = 5$ Rank sum $T_2 = 29$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - T_2$$
$$= 6 \cdot 5 + \frac{5 \cdot (5 + 1)}{2} - 29$$
$$= 16$$

U-Wert

$$U = min(U_1, U_2) = min(14, 16) = 14$$

Expected value of U

$$\mu_U = \frac{n_1 \cdot n_2}{2} = \frac{6 \cdot 5}{2} = 15$$

Standard error of U

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{6 \cdot 5 \cdot (6 + 5 + 1)}{12}} = 5.4$$

z-value

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{14 - 15}{5{,}4772} = -0.1825$$

P-value = 0.855 (two tail) , can not reject the null hypothesis

t-test

Some more materials on T-test

Steps for Significance Testing

1. Set alpha (p level).
2. State hypotheses, Null and Alternative.
3. Calculate the test statistic (sample value).
4. Find the critical value of the statistic.
5. State the decision rule.
6. State the conclusion.

t-test

- t –test is about means: distribution and evaluation for group distribution
- Withdrawn form the normal distribution
- The shape of distribution depend on sample size and, the sum of all distributions is a normal distribution
- t- distribution is based on sample size and vary according to the degrees of freedom

What is the t -test

- *t* test is a useful technique for comparing mean values of two sets of numbers.
- The comparison will provide you with a statistic for evaluating whether the difference between two means is statistically significant.
- *t* test is named after its inventor, William Gosset, who published under the pseudonym of student.
- *t* test can be used either :
  1. to compare two independent groups (independent-samples *t* test)
  2. to compare observations from two measurement occasions for the same group (paired-samples *t* test).

What is the t -test

- The null hypothesis states that any difference between the two means is a result to difference in distribution.
- Remember, both samples drawn randomly form the same population.
- Comparing the chance of having difference is one group due to difference in distribution.
- ***Assuming that both distributions came from the same population, both distribution has to be equal***.

What is the t -test

- Then, what we intend:

“To find the difference due to chance”

- Logically, The larger the difference in means, the more likely to find a significant *t* test.
- But, recall:

1. Variability

More variability = less overlap = larger difference

2. Sample size

Larger sample size = less variability (pop) = larger difference

Types

1. The ***independent-sample t test*** is used to compare two groups' scores on the same variable. For example, it could be used to compare the salaries of dentists and physicians to evaluate whether there is a difference in their salaries.

2. The ***paired-sample t test*** is used to compare the means of two variables within a single group. For example, it could be used to see if there is a statistically significant difference between starting salaries and current salaries among the general physicians in an organization.

Assumption

1. Dependent variable should be continuous (I/R)
2. The groups should be randomly drawn from normally distributed and independent populations

e.g. Male X Female

Dentist X Physician

Manager X Staff

NO OVER LAP

Assumption

3. the independent variable is categorical with two levels
4. Distribution for the **two independent** variables is normal
5. Equal variance (homogeneity of variance)
6. large variation = less likely to have sig t test = accepting null hypothesis (fail to reject) = Type II error = a threat to power

Sending an innocent to jail for no significant reason

Independent Samples *t*-test

- Used when we have two independent samples, e.g., treatment and control groups.
- Formula is:

$$t_{\bar{X}_1 - \bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{SE_{diff}}$$

- Terms in the numerator are the sample means.
- Term in the denominator is the standard error of the difference between means.

Independent samples *t*-test

The formula for the standard error of the difference in means:

$$SE_{diff} = \sqrt{\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}}$$

Suppose we study the effect of caffeine on a motor test where the task is to keep a the mouse centered on a moving dot. Everyone gets a drink; half get caffeine, half get placebo; nobody knows who got what.

Independent Sample Data (Data are time off task)

Experimental (Caff)	Control (No Caffeine)
12	21
14	18
10	14
8	20
16	11
5	19
3	8
9	12
11	13
	15
$N_1=9$, $M_1=9.778$, $SD_1=4.1164$	$N_2=10$, $M_2=15.1$, $SD_2=4.2805$

Independent Sample Steps(1)

1. Set alpha. Alpha = .05
2. State Hypotheses.

   Null is H_0: $\mu_1 = \mu_2$.

   Alternative is H_1: $\mu_1 \neq \mu_2$.

Independent Sample Steps(2)

3. Calculate test statistic:

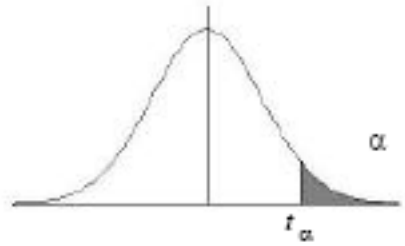
$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE_{diff}} = \frac{9.778 - 15.1}{1.93} = \frac{-5.322}{1.93} = -2.758$$

$$SE_{diff} = \sqrt{\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}} = \sqrt{\frac{(4.1164)^2}{9} + \frac{(4.2805)^2}{10}} = 1.93$$

Independent Sample Steps (3)

4. Determine the critical value. Alpha is .05, 2 tails, and df = N1+N2-2 or 10+9-2 = 17. The value is 2.11.
5. State decision rule. If |-2.758| > 2.11, then reject the null.
6. Conclusion: Reject the null. the population means are different. Caffeine has an effect on the motor pursuit task.

Table 4: Percentage Points of the *t* distribution



	α					
df	0.250	0.100	0.050	0.025	0.010	0.005
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
			⋮			
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

Dependent Samples t-tests

Dependent Samples *t*-test

- Used when we have dependent samples – matched, paired or tied somehow
  - Repeated measures
  - Brother & sister, husband & wife
  - Left hand, right hand, etc.
- Useful to control individual differences. Can result in more powerful test than independent samples *t*-test.

Dependent Samples *t*

Formulas:

$$t_{\bar{X}_D} = \frac{\overline{D}}{SE_{diff}}$$

t is the difference in means over a standard error.

$$SE_{diff} = \frac{SD_D}{\sqrt{n_{pairs}}}$$

The standard error is found by finding the difference between each pair of observations. The standard deviation of these difference is SD_D. Divide SD_D by sqrt (number of pairs) to get SE_{diff}.

Another way to write the formula

$$t_{\bar{X}_D} = \frac{\bar{D}}{SD_D / \sqrt{n_{pairs}}}$$

Dependent Samples *t* example

Person	Painfree (time in sec)	Placebo	Difference
1	60	55	5
2	35	20	15
3	70	60	10
4	50	45	5
5	60	60	0
M	55	48	7
SD	13.23	16.81	5.70

Dependent Samples *t* Example (2)

1. Set alpha = .05
2. Null hypothesis: H_0: $\mu_1 = \mu_2$.
   Alternative is H_1: $\mu_1 \neq \mu_2$.
3. Calculate the test statistic:

$$SE_{diff} = \frac{SD}{\sqrt{n_{pairs}}} = \frac{5.70}{\sqrt{5}} = 2.55$$

$$t = \frac{\overline{D}}{SE_{diff}} = \frac{55-48}{2.55} = \frac{7}{2.55} = 2.75$$

Dependent Samples t Example (3)

4. Determine the critical value of t.
   Alpha =.05, tails=2
   df = N(pairs)-1 =5-1=4.
   Critical value is 2.776
5. Decision rule: is absolute value of sample value larger than critical value?
6. Conclusion. Not (quite) significant. Painfree does not have an effect.

Dependent or Paired t-Test: Output

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Pre	4.7000	10	2.11082	.66750
	Post	6.2000	10	2.85968	.90431

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Pre & Post	10	.968	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Pre - Post	-1.50000	.97183	.30732	-2.19520	-.80480	-4.881	9	.001

Is there a difference between pre & post?

$t(9) = -4.881, p = .001$

Yes, 4.7 is significantly different from 6.2

GKTCS INNOVATIONS

Object Oriented Programming Using Python

Index

GKTCS INNOVATIONS

1. Introduction to Object Oriented Programming in Python

GKTCS INNOVATIONS

Object Oriented Programming is a way of computer programming using the idea of "objects" to represents data and methods. It is also, an approach used for creating neat and reusable code instead of a redundant one.

2. Difference between Object-Oriented and Procedural Oriented Programming

GKTCS INNOVATIONS

Object-Oriented Programming (OOP)	Procedural-Oriented Programming (Pop)
It is a bottom-up approach	It is a top-down approach
Program is divided into objects	Program is divided into functions
Makes use of Access modifiers 'public', private', protected'	Doesn't use Access modifiers
It is more secure	It is less secure
Object can move freely within member functions	Data can move freely from function to function within programs
It supports inheritance	It does not support inheritance

3. What are Classes and Objects?

GKTCS INNOVATIONS

A class is a collection of objects or you can say it is a blueprint of objects defining the common attributes and behavior. Now the question arises, how do you do that?

Class is defined under a "Class" Keyword.

Example:

```
class class1(): // class 1 is the name of the class
```

GKTCS INNOVATIONS

Creating an Object and Class in python:

Example:

```
class employee():
    def __init__(self,name,age,id,salary):   //creating a function
        self.name = name // self is an instance of a class
        self.age = age
        self.salary = salary
        self.id = id

emp1 = employee("harshit",22,1000,1234) //creating objects
emp2 = employee("arjun",23,2000,2234)
print(emp1.__dict__)//Prints dictionary
```

4. Object-Oriented Programming methodologies:

GKTCS INNOVATIONS

- Inheritance
- Polymorphism
- Encapsulation
- Abstraction

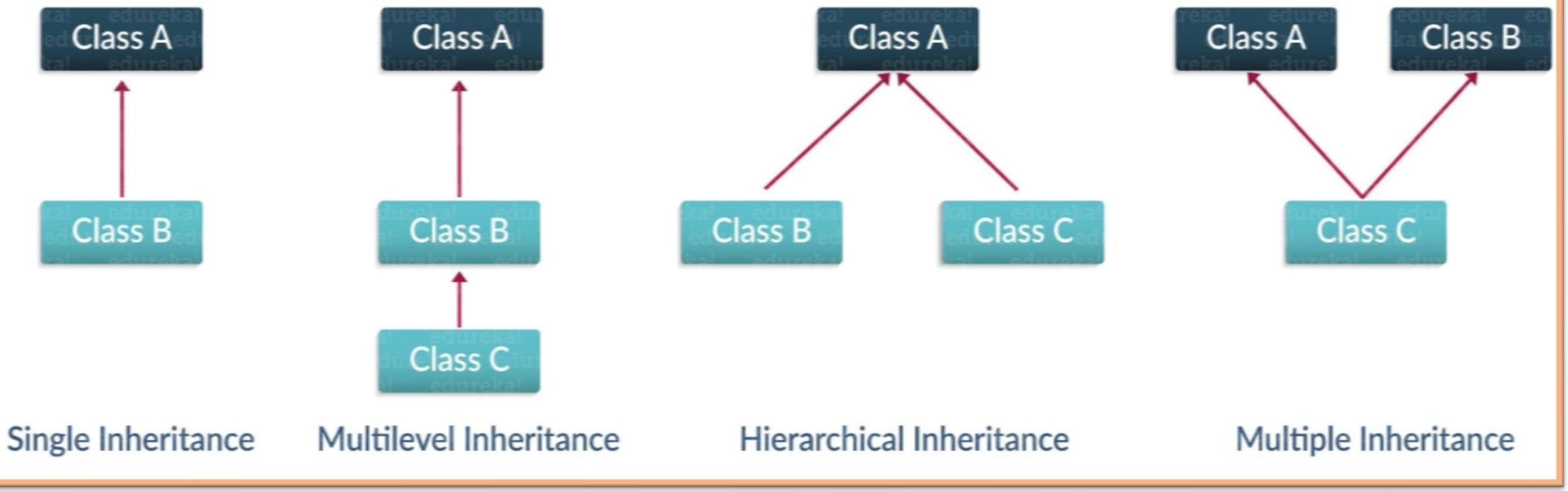
GKTCS INNOVATIONS

Inheritance:

Ever heard of this dialogue from relatives "you look exactly like your father/mother" the reason behind this is called 'inheritance'. From the Programming aspect, It generally means "inheriting or transfer of characteristics from parent to child class without any modification". The new class is called the derived/child class and the one from which it is derived is called a parent/base class.

GKTCS INNOVATIONS

Types Of Inheritance



GKTCS INNOVATIONS

Single Inheritance:

Single level inheritance enables a derived class to inherit characteristics from a single parent class.

GKTCS INNOVATIONS

Example:

```
class employee1()://This is a parent class
    def __init__(self, name, age, salary):
        self.name = name
        self.age = age
        self.salary = salary

class childemployee(employee1)://This is a child class
    def __init__(self, name, age, salary,id):
        self.name = name
        self.age = age
        self.salary = salary
        self.id = id
emp1 = employee1('harshit',22,1000)
print(emp1.age)
```

Output: 22

GKTCS INNOVATIONS

Multilevel Inheritance:

Multi-level inheritance enables a derived class to inherit properties from an immediate parent class which in turn inherits properties from his parent class.

Example:

```
class employee()://Super class
    def __init__(self,name,age,salary):
        self.name = name
        self.age = age
        self.salary = salary
class childemployee1(employee)://First child class
    def __init__(self,name,age,salary):
        self.name = name
        self.age = age
        self.salary = salary
```

```
class childemployee2(childemployee1)://Second child class
    def __init__(self, name, age, salary):
        self.name = name
        self.age = age
        self.salary = salary
emp1 = employee('harshit',22,1000)
emp2 = childemployee1('arjun',23,2000)

print(emp1.age)
print(emp2.age)
```

Output: 22,23

GKTCS INNOVATIONS

Hierarchical Inheritance:

Hierarchical level inheritance enables more than one derived class to inherit properties from a parent class.

Example:

```
class employee():
    def __init__(self, name, age, salary):     //Hierarchical Inheritance
        self.name = name
        self.age = age
        self.salary = salary
```

```
class childemployee1(employee):
    def __init__(self,name,age,salary):
        self.name = name
        self.age = age
        self.salary = salary

class childemployee2(employee):
    def __init__(self, name, age, salary):
        self.name = name
        self.age = age
        self.salary = salary
emp1 = employee('harshit',22,1000)
emp2 = employee('arjun',23,2000)
```

GKTCS INNOVATIONS

Multiple Inheritance:

Multiple level inheritance enables one derived class to inherit properties from more than one base class.

Example:

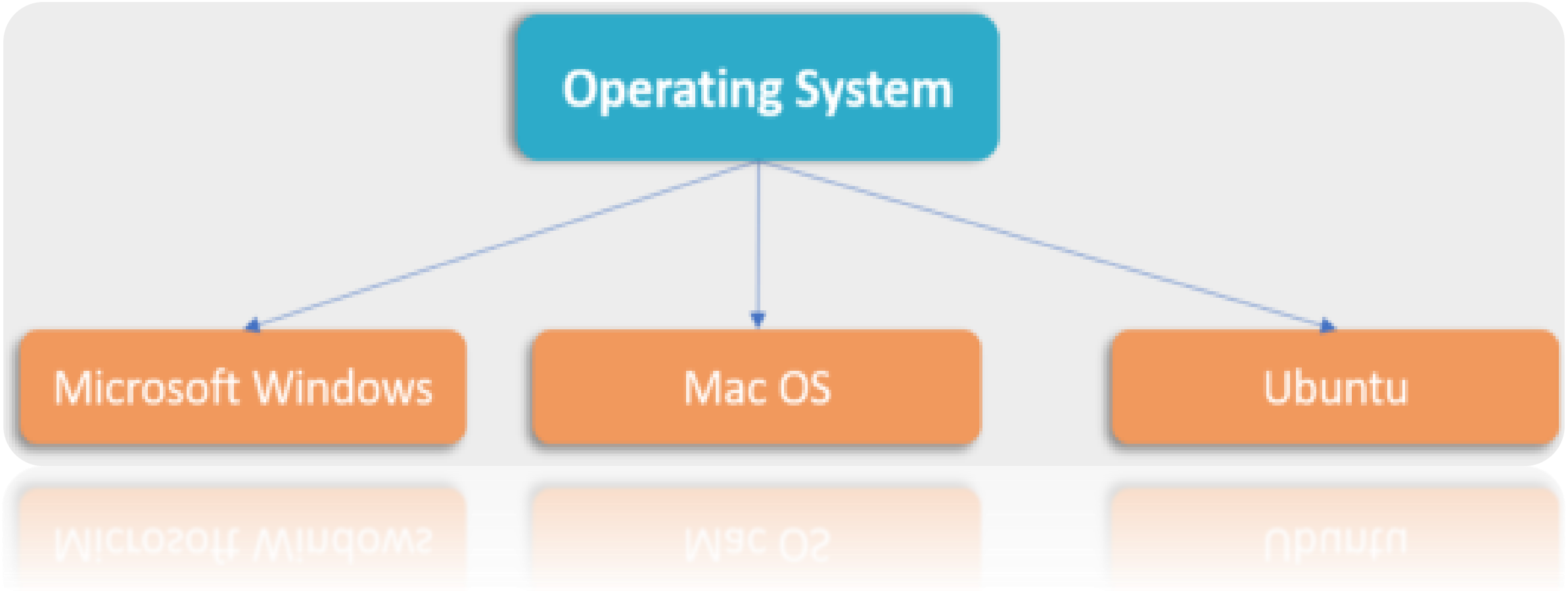
```
class employee1(): //Parent class
    def __init__(self, name, age, salary):
        self.name = name
        self.age = age
        self.salary = salary
```

```
class employee2(): //Parent class
    def __init__(self,name,age,salary,id):
     self.name = name
     self.age = age
     self.salary = salary
     self.id = id

class childemployee(employee1,employee2):
    def __init__(self, name, age, salary,id):
     self.name = name
     self.age = age
     self.salary = salary
     self.id = id

emp1 = employee1('harshit',22,1000)
emp2 = employee2('arjun',23,2000,1234)
```

GKTCS INNOVATIONS

Polymorphism:

You all must have used GPS for navigating the route, Isn't it amazing how many different routes you come across for the same destination depending on the traffic, from a programming point of view this is called 'polymorphism'. It is one such OOP methodology where one task can be performed in several different ways. To put it in simple words, it is a property of an object which allows it to take multiple forms.


Operating System
Microsoft Windows
Mac OS
Ubuntu

GKTCS INNOVATIONS

Polymorphism is of two types:

- Compile-time Polymorphism
- Run-time Polymorphism

GKTCS INNOVATIONS

Compile-time Polymorphism:

A compile-time polymorphism also called as static polymorphism which gets resolved during the compilation time of the program. One common example is "method overloading"

Example:

GKTCS INNOVATIONS

```
class employee1():
    def name(self):
        print("Harshit is his name")
    def salary(self):
        print("3000 is his salary")
    def age(self):
        print("22 is his age")

class employee2():
    def name(self):
        print("Rahul is his name")
    def salary(self):
        print("4000 is his salary")
    def age(self):
        print("23 is his age")
```

```
def func(obj)://Method Overloading
    obj.name()
    obj.salary()
    obj.age()

obj_emp1 = employee1()
obj_emp2 = employee2()
func(obj_emp1)
func(obj_emp2)
```

Output:

```
Harshit is his name
3000 is his salary
22 is his age
Rahul is his name
4000 is his salary
23 is his age
```

GKTCS INNOVATIONS

Run-time Polymorphism:

A run-time Polymorphism is also, called as dynamic polymorphism where it gets resolved into the run time. One common example of Run-time polymorphism is "method overriding".

GKTCS INNOVATIONS

Example:

```
class employee():
    def __init__(self,name,age,id,salary):
        self.name = name
        self.age = age
        self.salary = salary
        self.id = id
    def earn(self):
        pass

class childemployee1(employee):
    def earn(self): //Run-time polymorphism
        print("no money")
```

```
class childemployee2(employee):
    def earn(self):
        print("has money")

c = childemployee1
c.earn(employee)
d = childemployee2
d.earn(employee)
```

Output: no money, has money

Abstraction:

Suppose you booked a movie ticket from bookmyshow using net banking or any other process. You don't know the procedure of how the pin is generated or how the verification is done. This is called 'abstraction' from the programming aspect, it basically means you only show the implementation details of a particular process and hide the details from the user.