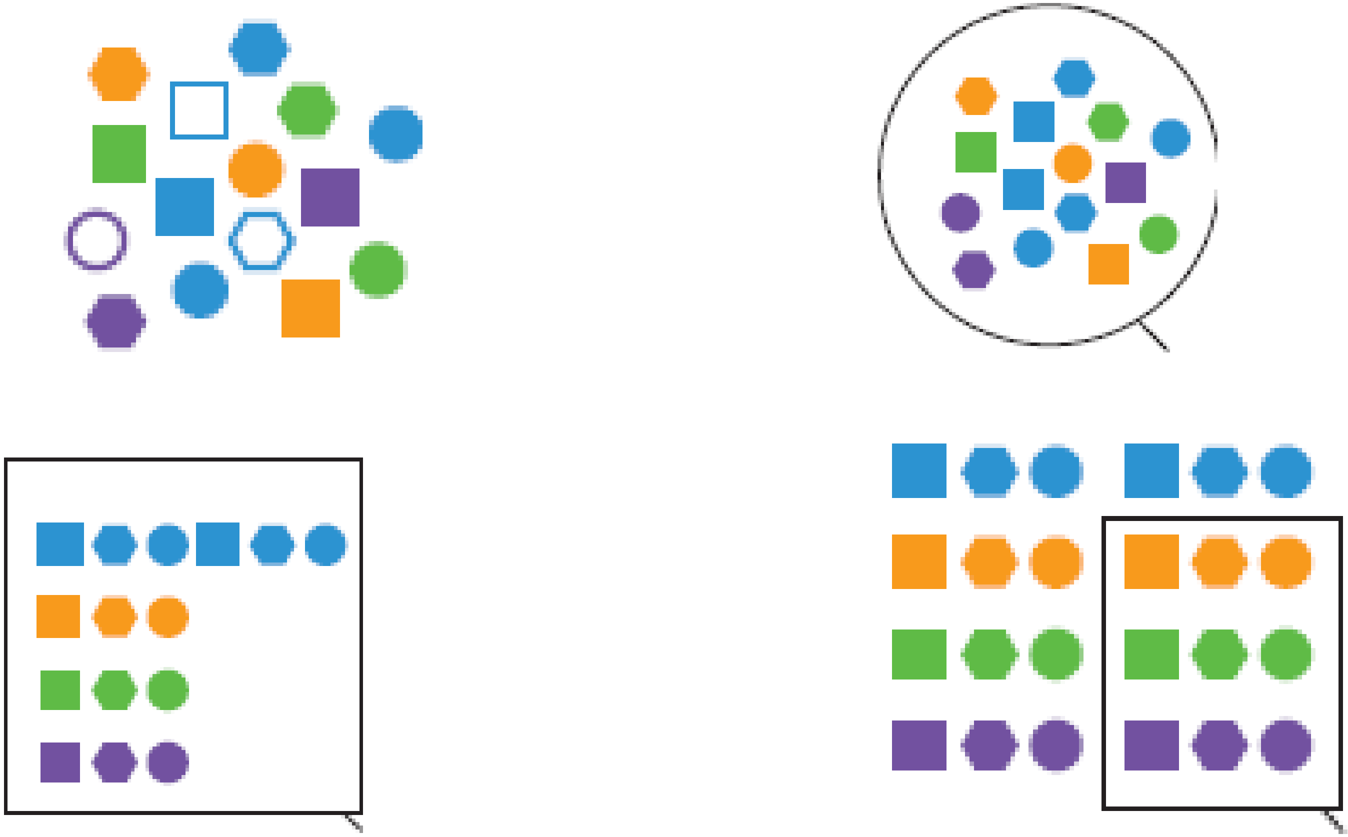
20CSE02 - DATA SCIENCE

What you infer from this?



Unit 1

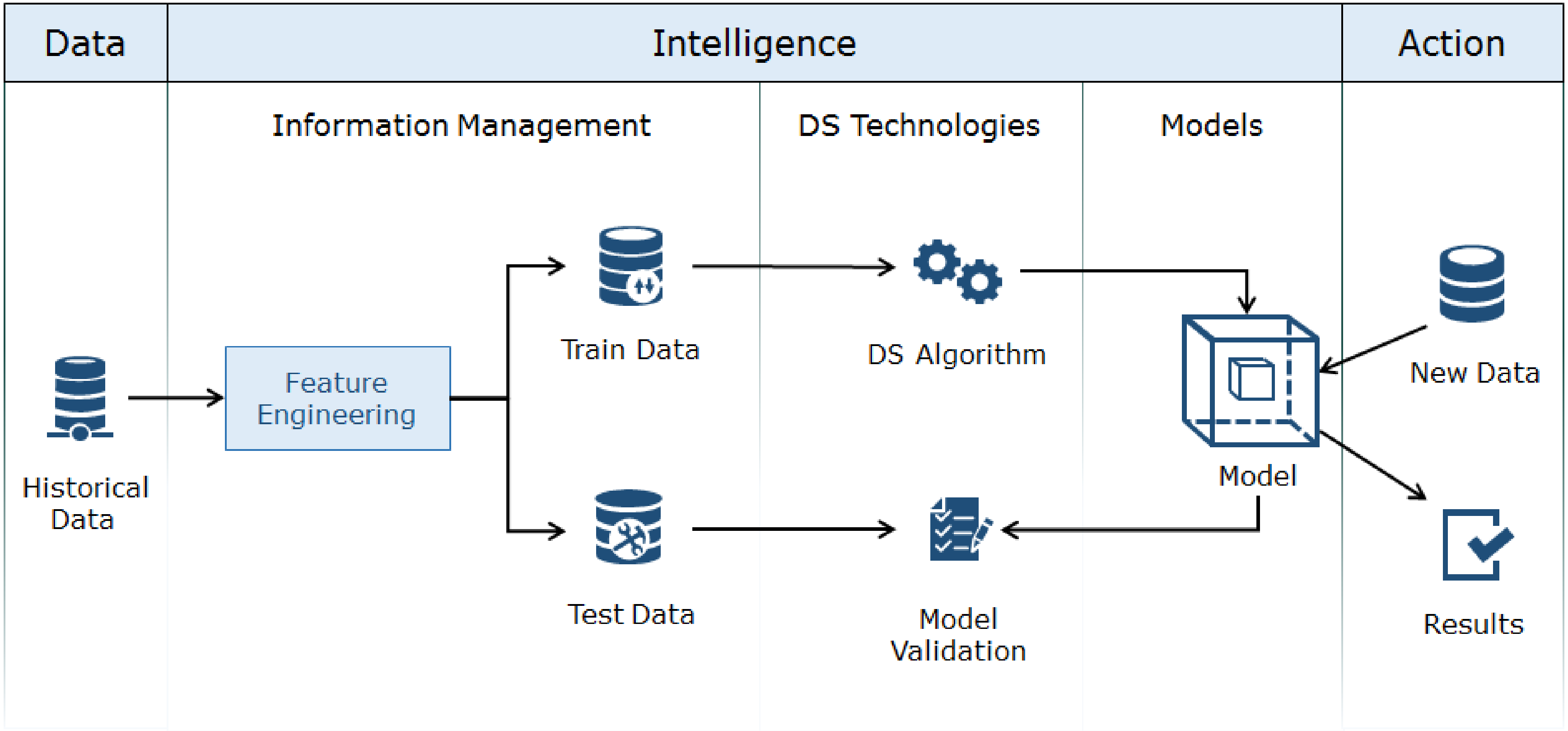
- Introduction
- Data Science
- Data Science Relate to Other Fields
- The Relationship between Data Science and Information Science
- Computational Thinking
- Issues of Ethics, Bias, and Privacy in Data Science

- Data Types
- Data Collections
- Data Pre-processing Techniques
- Data Analysis & Analytics
- Descriptive Analysis
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics
- Exploratory Analysis
- Mechanistic Analysis

Data Science

- Frank Lo, the Director of Data Science at Wayfair, "Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems."
- data science as a field of study and practice that involves the collection, storage, and processing of data in order to derive important insights into a problem or a phenomenon.

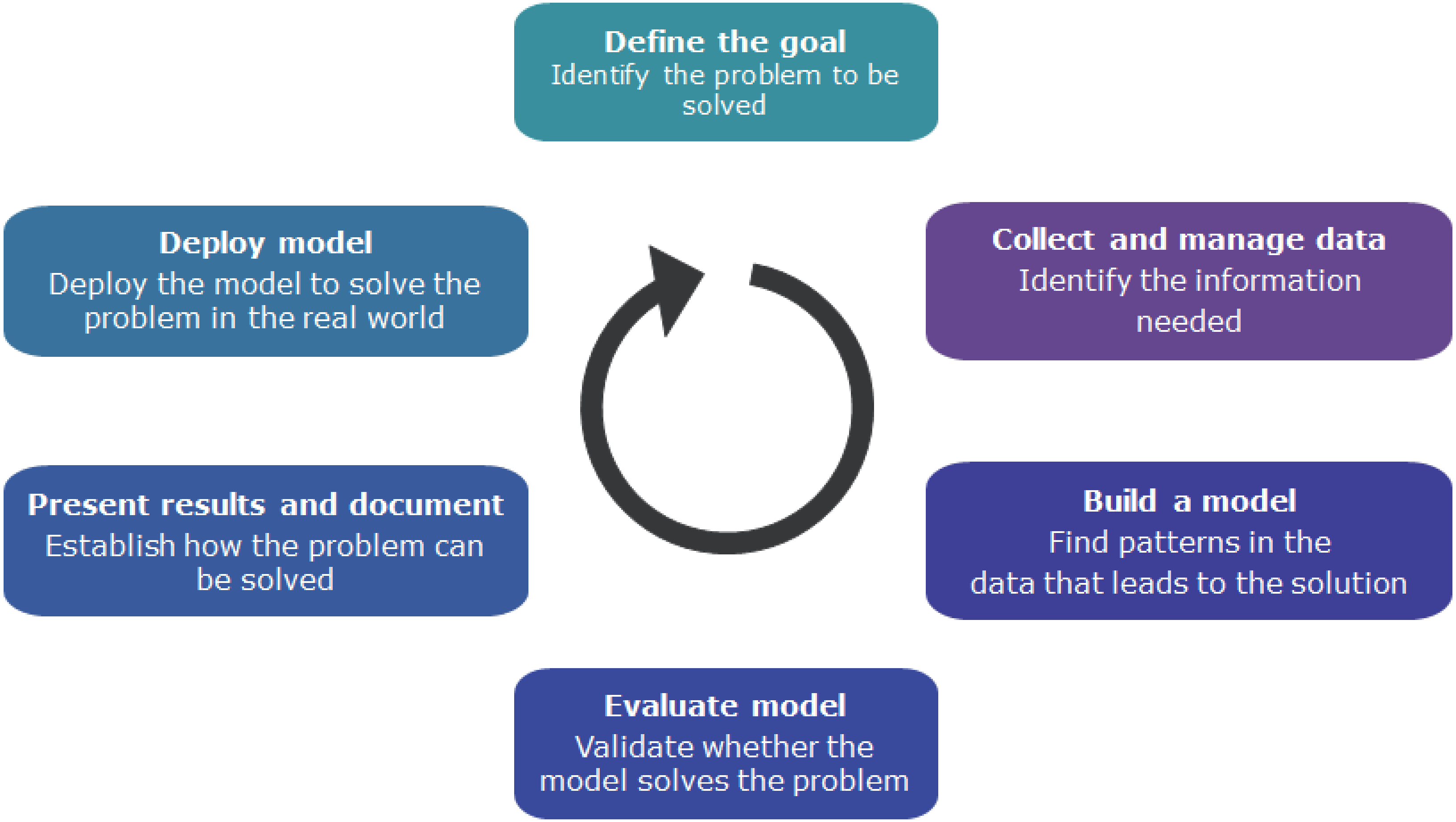
Data Science Process



DS Architecture

Data	Intelligence			Action
	Information Management	DS Technologies	Models	
Data Sources Apps Sensors and Devices	• Databases • Spark • Hadoop	• Exploratory Data Analysis • Statistical Inference • Association Analysis • Regression Analysis • Time Series Analysis • Machine learning		People Web Apps Bots Mobile Automated Systems
• Historic Business data that can be used to gain business insights. • It can lead to gain in efficiency	• Construct and select features for the problem • Cleaning the data • Distributed/parallel computing environment for handling massive amounts of data	• Contains components that create a model based on the patterns in the training data.	• Is the pattern/knowledge extracted from the data • Given new data, it produces the desired result (Classification, regression etc)	• Layer that uses the model to gain business insights.

DS Life Cycle



Define the goal
Identify the problem to be solved

Collect and manage data
Identify the information needed

Build a model
Find patterns in the data that leads to the solution

Evaluate model
Validate whether the model solves the problem

Present results and document
Establish how the problem can be solved

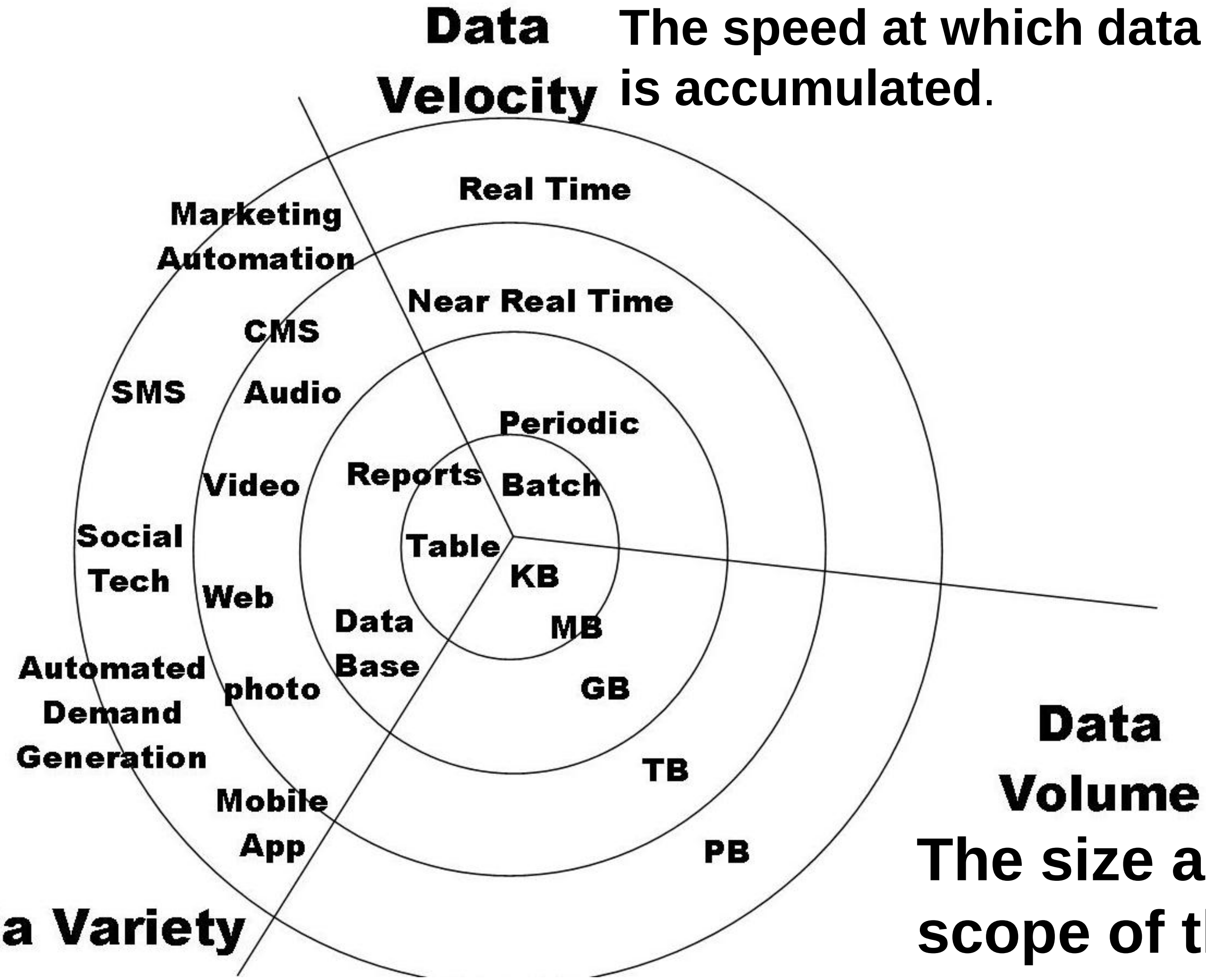
Deploy model
Deploy the model to solve the problem in the real world

Data Science

- data may be generated by humans (surveys, logs, etc.) or machines (weather data, road vision, etc.), and
- could be in different formats (text, audio, video, augmented or virtual reality, etc.).
- Why is data science so important now?

“3V model”

3V model /Big Data

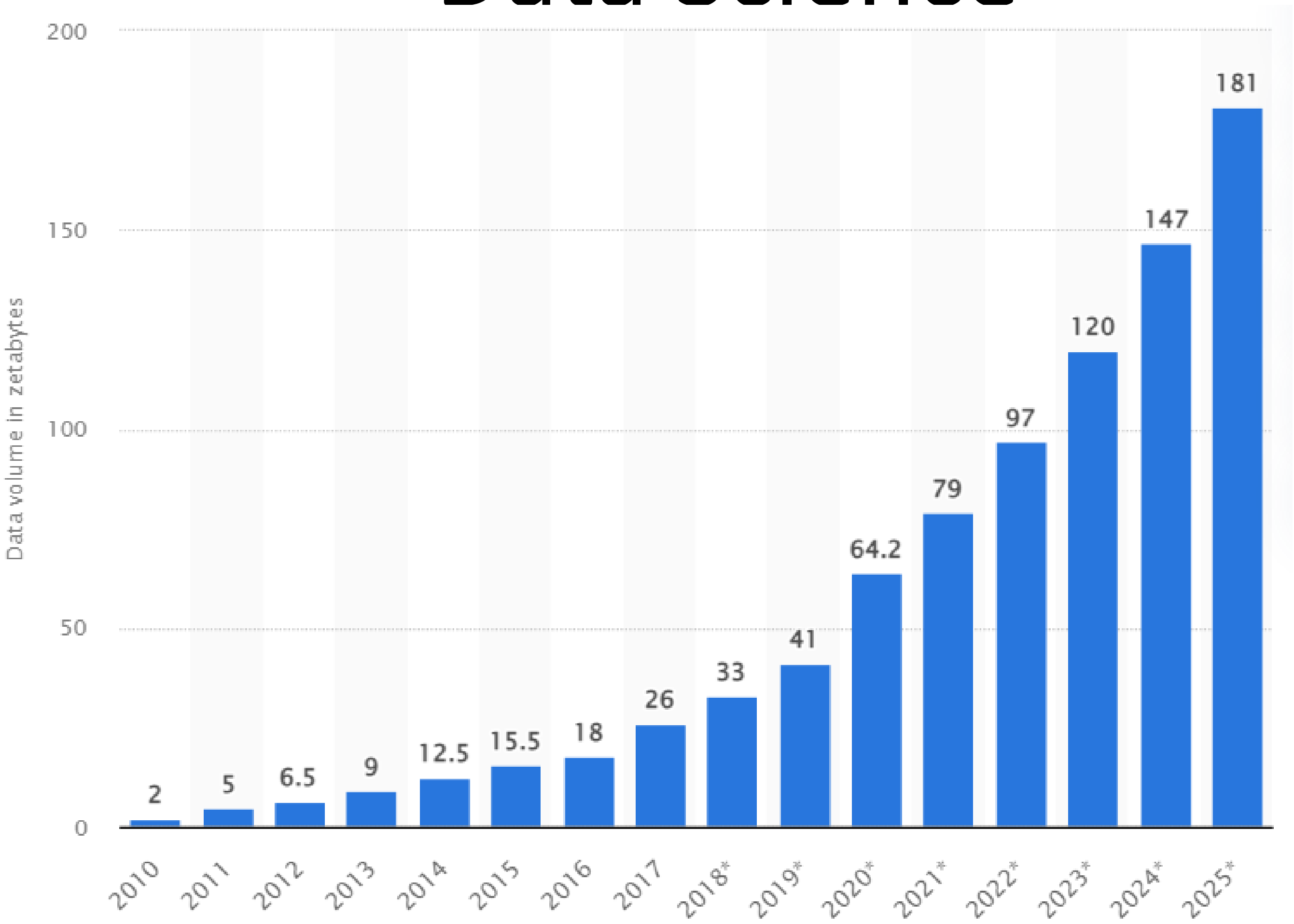


The speed at which data is accumulated.

The size and scope of the data

structured and unstructured

Data Science



Increase of data volume 2010 – 2025 (as on July 2022)
https://www.statista.com/statistics/871513/worldwide-data-created/

Data Science

Sources for exponential growth of data

1. Social media activity,
2. mobile interactions,
3. server logs,
4. Realtime market feeds,
5. customer service records,
6. transaction details, and
7. information from existing databases combine to create a rich and complex conglomeration of information …………..

Data Science

Where Do We See Data Science?

- Finance
- Public Policy -gain insights into citizen behaviours that affect the quality of public life, including traffic, public transportation, social welfare, community wellbeing, etc.
- Politics
- Healthcare
- Urban Planning
- Education
- Libraries - Online Public Access Catalogues (OPACs)

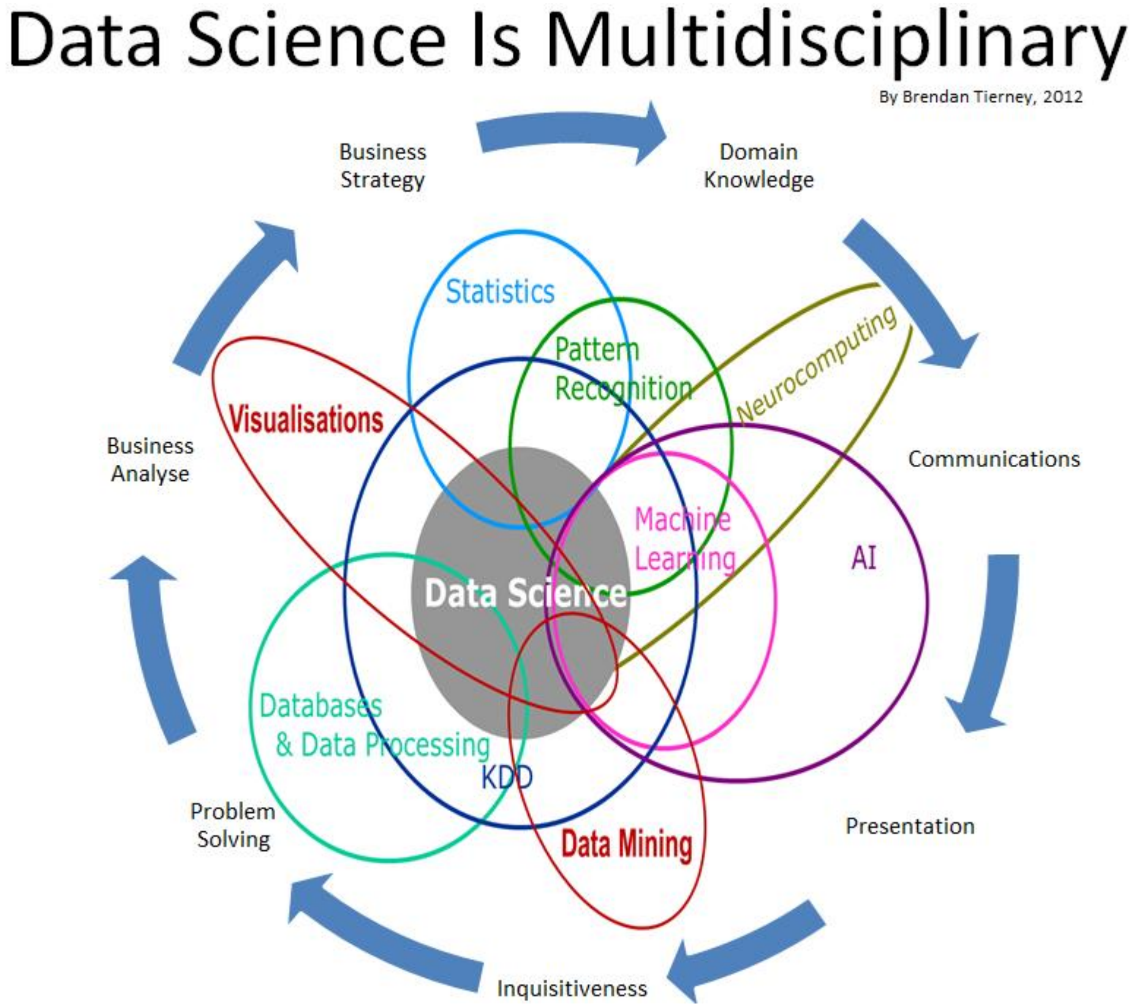
Data Science

What do financial data scientists do?

- Through capturing and analyzing new sources of data, building predictive models and running real-time simulations of market events, they help the finance industry obtain the information necessary to make accurate predictions
- banks and other loan sanctioning institutions

  => can minimize the chance of loan defaults via information such as customer profiling, past expenditures, other essential variables that can be used to analyze the probabilities of risk and default

How Does Data Science Relate to Other Fields?



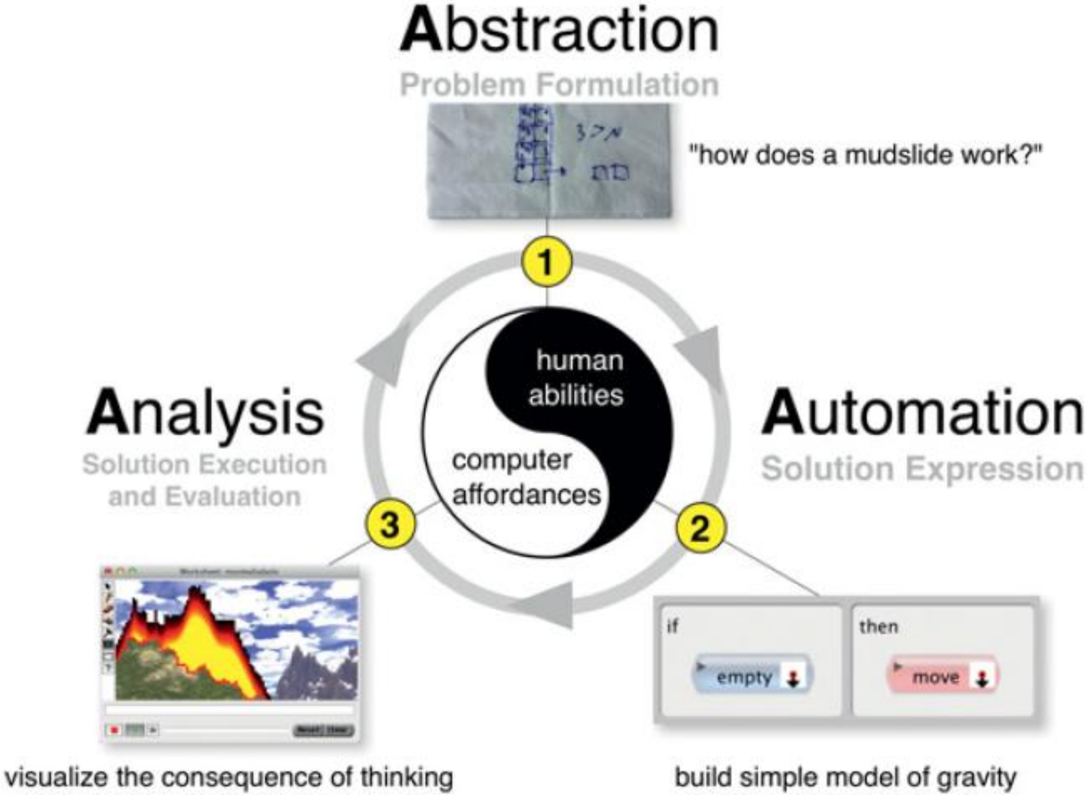
The Relationship between Data Science & Information Science

- Information vs. Data
- Users in Information Science => usefullness
- Data Science in Information Schools (iSchools)

iSchool curriculum helps students acquire diverse perspectives on data and information.

Computational Thinking

Computational thinking is using abstraction and decomposition when attacking a large complex task or designing a large complex system



Abstraction

Problem Formulation

"how does a mudslide work?"

1

human abilities

computer affordances

Automation

Solution Expression

2

Analysis

Solution Execution and Evaluation

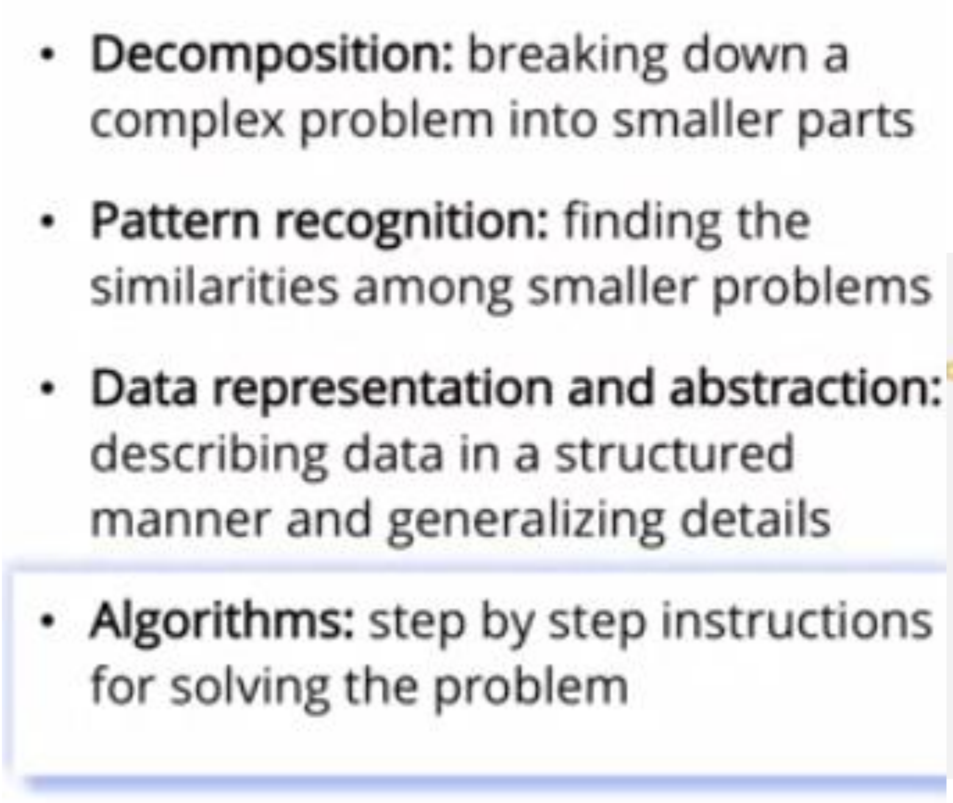
3

visualize the consequence of thinking

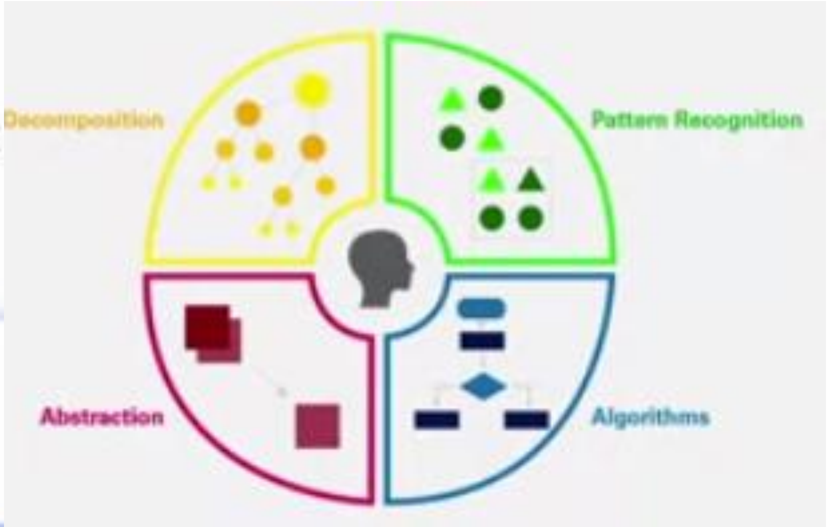
build simple model of gravity

Three-stage process describing computational thinking.

Pillars of Computational Thinking



- **Decomposition:** breaking down a complex problem into smaller parts
- **Pattern recognition:** finding the similarities among smaller problems
- **Data representation and abstraction:** describing data in a structured manner and generalizing details
- **Algorithms:** step by step instructions for solving the problem



Computational Thinking - Example

- We are given the following numbers and are asked to find the largest of them:
- 7, 24, 62, 11, 4, 39, 42, 5, 97, 54
- Second largest number?

- https://www.youtube.com/watch?v=qbnTZCj0ugI

Computational Thinking - Example

Find the sum of all numbers between 1 and 100

- Decompose
- Identify the patterns or trends within a problem
- Identify specific similarities and differences among similar problems to work towards the solution
- Algorithm

Computational Thinking - Example

- Find the sum of all numbers between 1 and 200

Decompose

200+1 =201
199+2=201
198+3=201…
101+100=201

No. of Pairs = 200/2=100

Difference => Last No. – First No.

(Similarity last No. +1

No. of Pairs = 200/2=100

Identify the Pattern

200+1 = 201
(Sum of the pair)

Sum of all numbers = Sum of the pair * No. of Pairs
=(200+1)*(200/2)=20100

https://www.youtube.com/watch?v=qbnTZCj0ugI

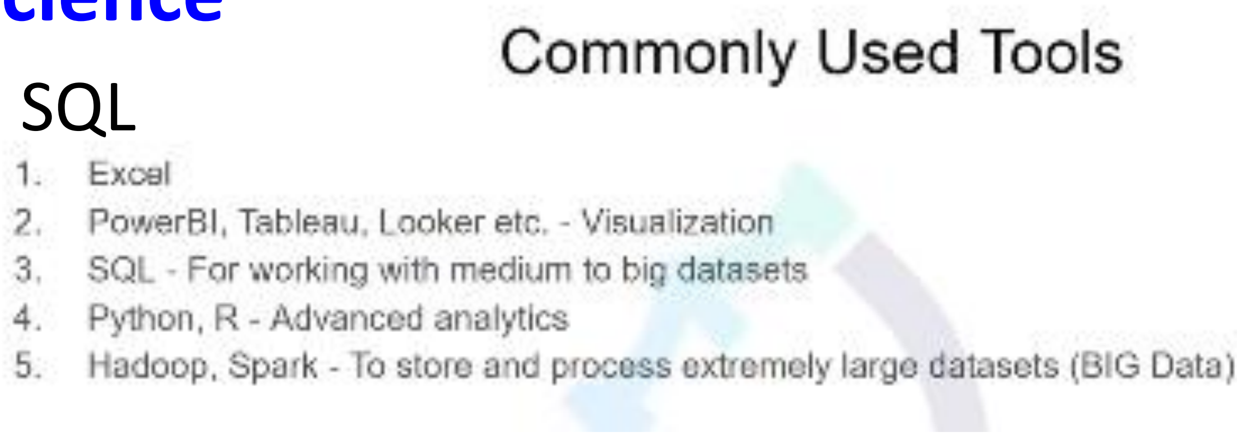
Data Science

Skills for Data Science

1. willing to experiment,
2. proficiency in mathematical reasoning, and
3. data literacy

Tools for Data Science

- Python, R, and SQL
- C, Java, PHP
- MATLAB....



Commonly Used Tools

1. Excel
2. PowerBI, Tableau, Looker etc. - Visualization
3. SQL - For working with medium to big datasets
4. Python, R - Advanced analytics
5. Hadoop, Spark - To store and process extremely large datasets (BIG Data)

Data Science

Issues of Ethics, Bias, and Privacy in Data Science

- how, where, and why was the data collected? Who collected it?
- What did they intend to use it for?
- if the data was collected from people, did these people know?
- Eg Facebook and Google have collected enormous amounts of data about and from their users in order not only to improve and market their products, but also to share and/or sell it to other entities for profit

Data Collections

- Data Types
  - – structured data
  - – unstructured data
    - Challenges with Unstructured Data
- Data Collections

  1.Open Data

    - freely available in a public domain
    - without restrictions from copyright, patents
    - UCI Machine Learning Repository

Data Collections

Principles associated with open data

- Public
- Accessible
- Described
- Reusable
- Complete
- Timely
- Managed Post-Release

Data Collections

2.Social Media Data Application Programming Interface (API) is used to collecting data from social meadia

- Social media data - analyzed for research or marketing purposes

3. Multimodal Data
   - IoT
     - Healthcare Applications
     - Agriculture
     - Industry

Data Collections

Data Storage and Presentation

- comma-separated values (CSV)
- tab-separated values (TSV)
- XML (eXtensible Markup Language)
- RSS (Really Simple Syndication)
  - Information provided by a website in an XML file in such a way is called an RSS feed.
  - Since RSS data is small and fast loading, it can easily be used with services such as mobile phones, personal digital assistants (PDAs), and smart watches.
  - RSS is useful for websites that are updated frequently

Data Collections

Data Storage and Presentation

- JSON (JavaScript Object Notation)
  - Key-value pair = In various languages, this is realized as an object, record, structure, dictionary, hash table, keyed list, or associative array.

Data Pre-processing

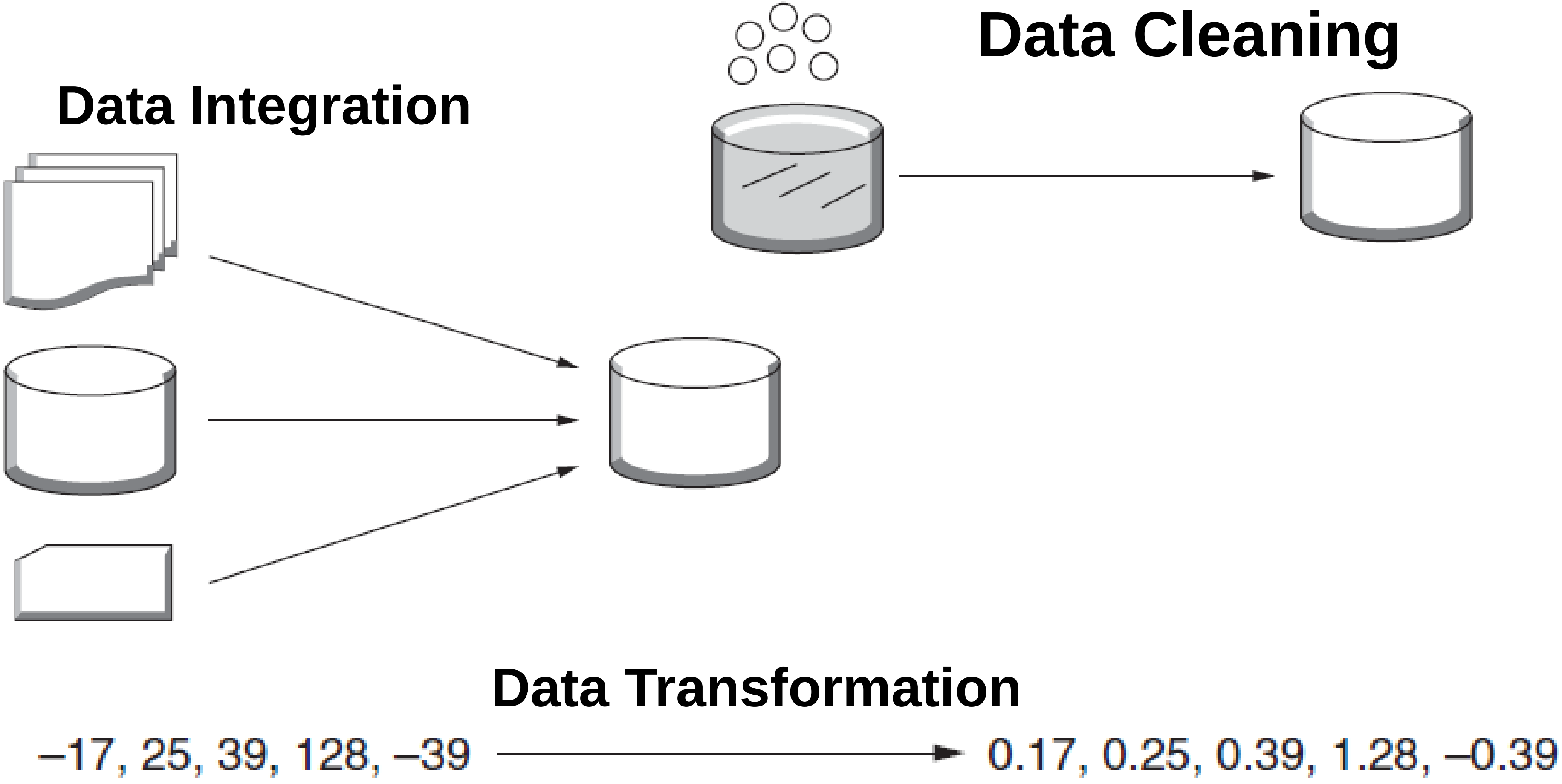
What makes data "dirty"?

- Incomplete
  - lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., Occupation=" " (missing data)
- Noisy
  - e.g., Salary="−10" (an error)
- Inconsistent
  - inconsistent: containing discrepancies in codes or names,
  - Age="42", Birthday="03/07/2018"
  - Grade "S,A,B,C,D,E,RA", now rating "O,A+,A,B+,B,RA"
  - discrepancy between duplicate records

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Data Pre-processing



Data Cleaning

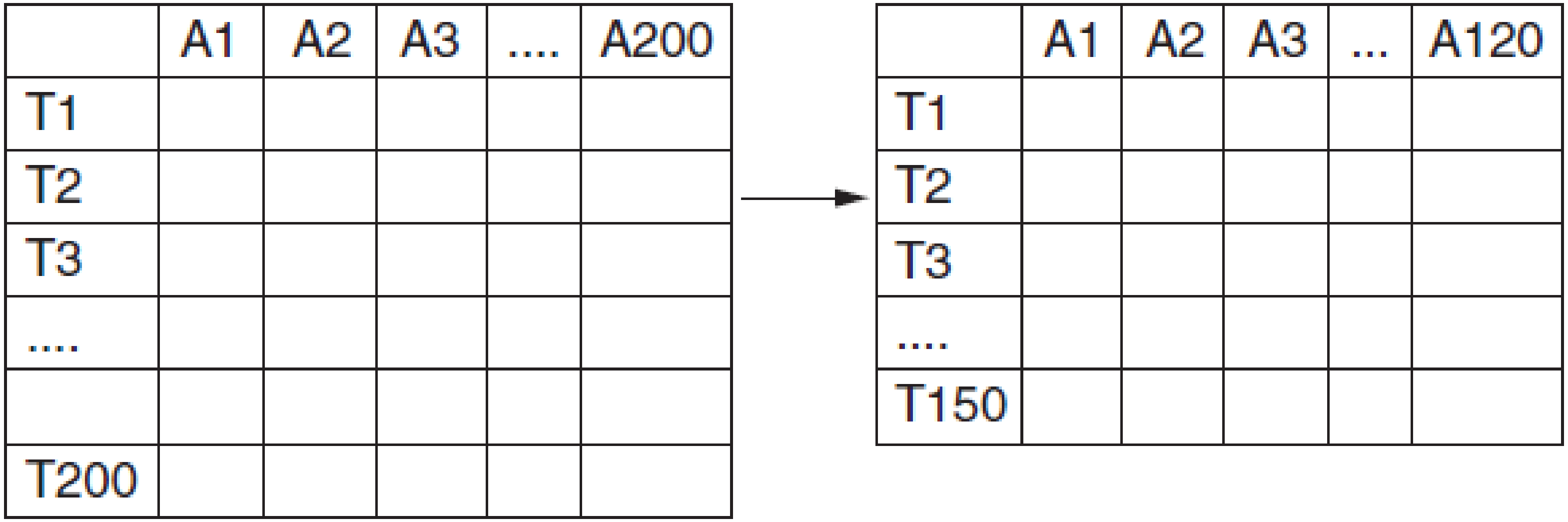
Data Integration

Data Transformation

–17, 25, 39, 128, –39 → 0.17, 0.25, 0.39, 1.28, –0.39

Data Pre-processing

Reduction in number of Columns (Attributes) and No. of rows (instances)



	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

Data Reduction

Data Pre-processing

Data Munging

“Add two diced tomatoes, three cloves of garlic, and a pinch of salt in the mix.”

- Munging is done either manually, automatically, or, in many cases, semi-automatically

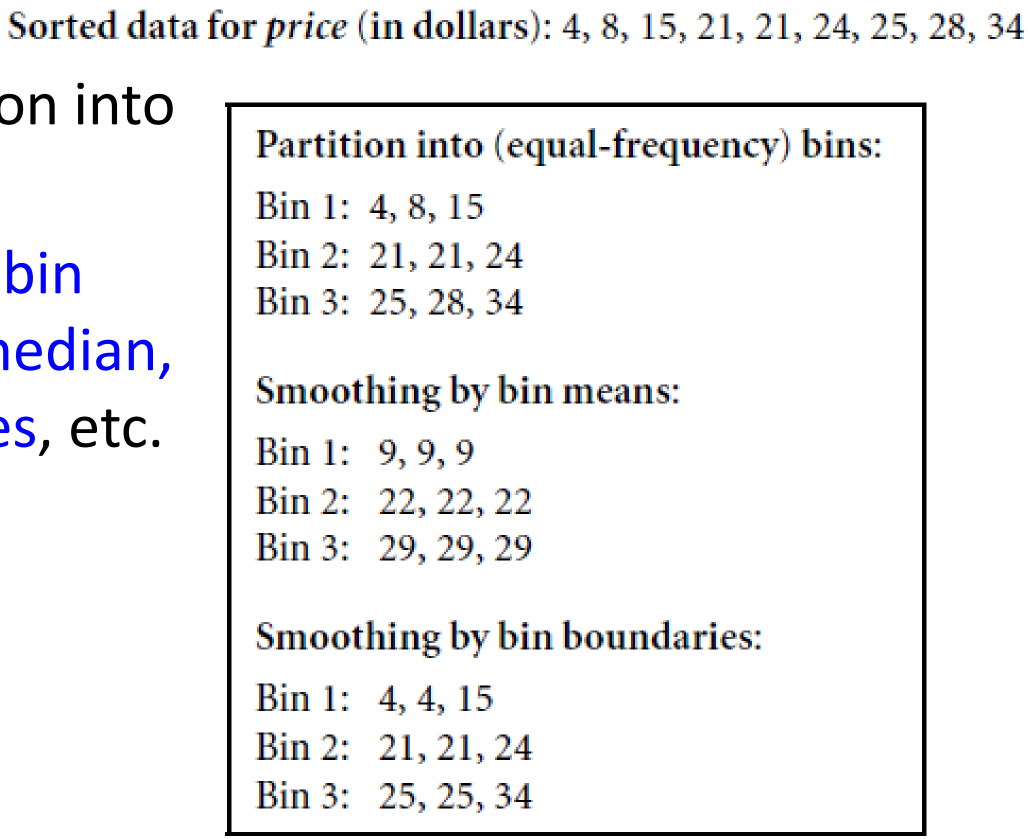
Ingredient	Quantity	Unit/size
Tomato	2	Diced
Garlic	3	Cloves
Salt	1	Pinch

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

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.
  - Binning is used for data discrtization



Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

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
- Address redundant data in data integration

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
  - Generalization : Concept hierarchy climbing

Normalization

Types

- **Min-max normalization**
- **Z-score normalization**
- **Normalization by decimal scaling**

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

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

Normalization

- **Z-score normalization** (μ: mean, σ: standard deviation):
  - Ex. Let μ = 54,000, σ = 16,000. Then

$$v' = \frac{v - \mu_A}{\sigma_A} \qquad \frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Consider the following is the age of 12 persons.
8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34
Normalize the age attribute using min-max, Z-score normalization

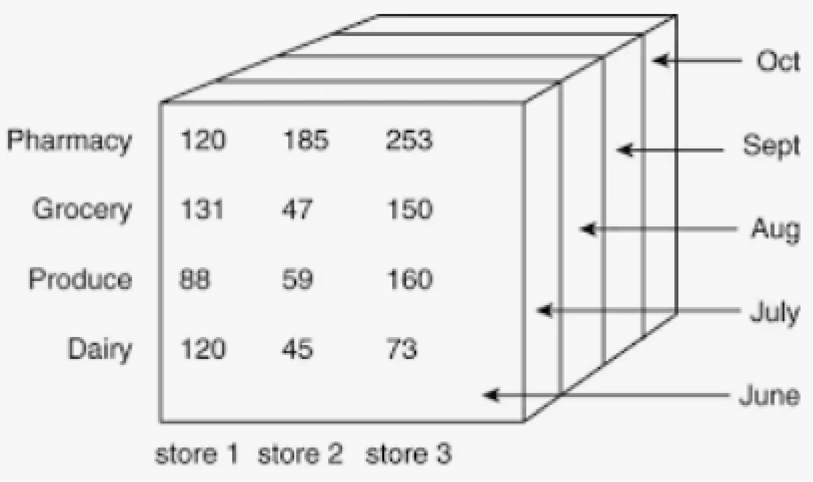
- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$

Where *j* is the smallest integer such that Max(|v'|) < 1

Data Reduction

- Data Cube Aggregation
  - Data cubes: They are multidimensional sets of data that can be stored in a spreadsheet. A data cube could be in two, three, or higher dimensions. Each dimension typically represents an attribute of interest.



	store 1	store 2	store 3
Pharmacy	120	185	253
Grocery	131	47	150
Produce	88	59	160
Dairy	120	45	73

- Dimensionality Reduction

Data Discretization

Divide the range of a continuous attribute into intervals

1. Marks converted to Grade
2. Age mapped to => Young, Adult
3. Range of temperature values => cold, moderate, and hot

Types of attributes

- Categorical variables
  - Nominal: Values from an unordered set (Colors, Blood Groups, Gender)
  - Ordinal: Values from an ordered set (academic rank, Customer satisfaction [Excellent, good...])
- Continuous: Real numbers
- Ratio scaled : No. Male & females in a class 3:4

Data Pre-processing

Data Cleaning

1. Smooth Noisy Data
2. Handling Missing Data

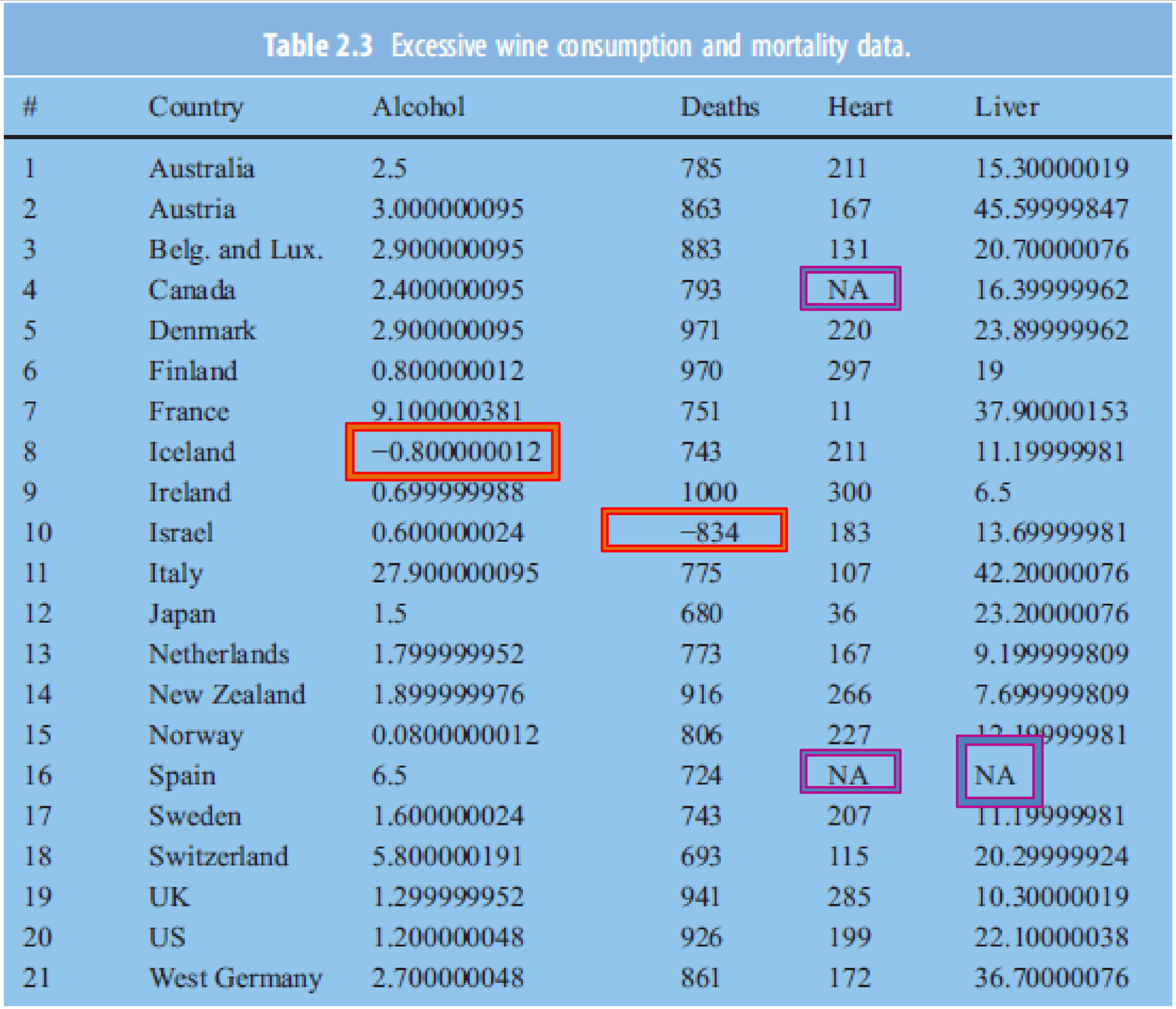


Table 2.3 Excessive wine consumption and mortality data.

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2.5	785	211	15.30000019
2	Austria	3.000000095	863	167	45.59999847
3	Belg. and Lux.	2.900000095	883	131	20.70000076
4	Canada	2.400000095	793	NA	16.39999962
5	Denmark	2.900000095	971	220	23.89999962
6	Finland	0.800000012	970	297	19
7	France	9.100000381	751	11	37.90000153
8	Iceland	−0.800000012	743	211	11.19999981
9	Ireland	0.699999988	1000	300	6.5
10	Israel	0.600000024	−834	183	13.69999981
11	Italy	27.900000095	775	107	42.20000076
12	Japan	1.5	680	36	23.20000076
13	Netherlands	1.799999952	773	167	9.199999809
14	New Zealand	1.899999976	916	266	7.699999809
15	Norway	0.0800000012	806	227	12.19999981
16	Spain	6.5	724	NA	NA
17	Sweden	1.600000024	743	207	11.19999981
18	Switzerland	5.800000191	693	115	20.29999924
19	UK	1.299999952	941	285	10.30000019
20	US	1.200000048	926	199	22.10000038
21	West Germany	2.700000048	861	172	36.70000076

Data Pre-processing

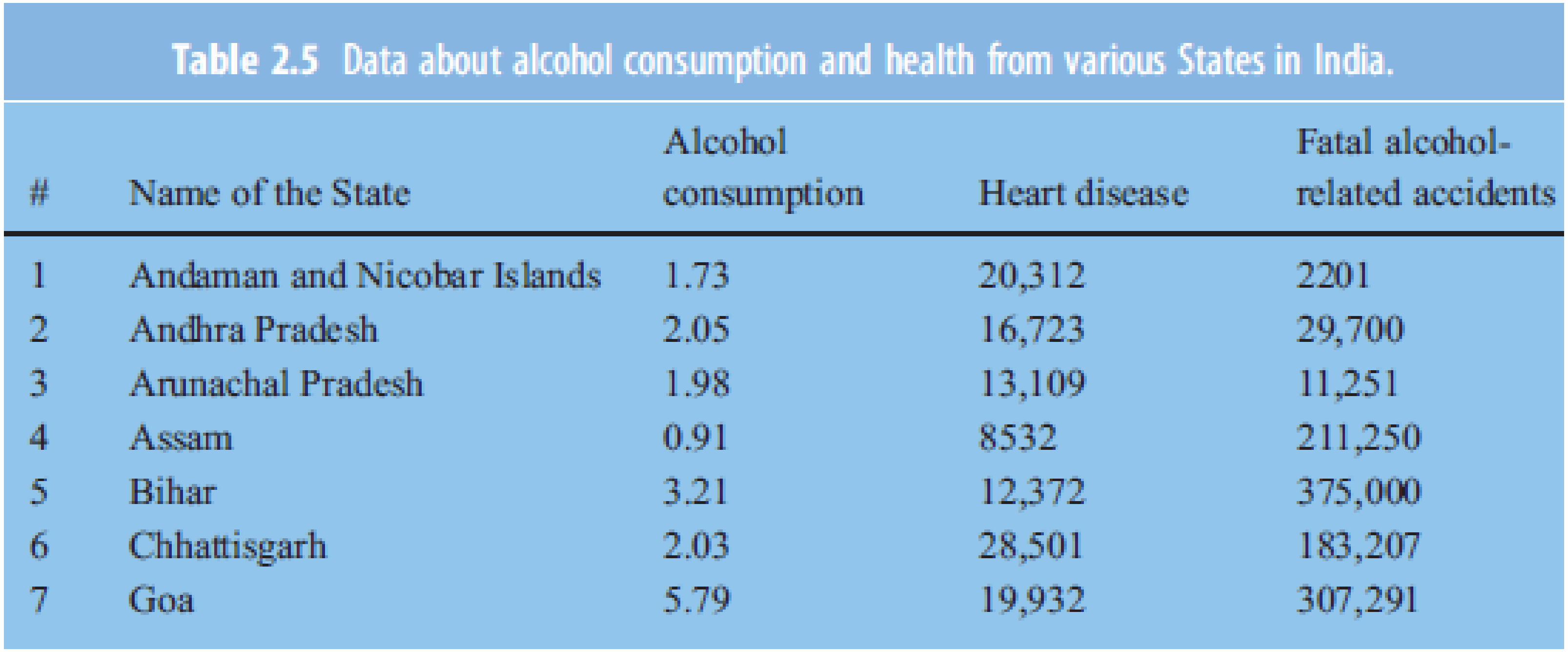


Table 2.5 Data about alcohol consumption and health from various States in India.

#	Name of the State	Alcohol consumption	Heart disease	Fatal alcohol-related accidents
1	Andaman and Nicobar Islands	1.73	20,312	2201
2	Andhra Pradesh	2.05	16,723	29,700
3	Arunachal Pradesh	1.98	13,109	11,251
4	Assam	0.91	8532	211,250
5	Bihar	3.21	12,372	375,000
6	Chhattisgarh	2.03	28,501	183,207
7	Goa	5.79	19,932	307,291

1. Data Cube Aggregation/Concept Hierarchy

Dimensionality Reduction => Sum up all

2. Data Integration from two different sources

Data Pre-processing - Data Discretization

Discretize the wine consumption per capita into **four categories**

1. less than or equal to 1.00 per capita => (represented by 0),
2. more than 1.00 but less than or equal to 2.00 per capita (1),
3. more than 2.00 but less than or equal to 5.00 per capita (2), and
4. more than 5.00 per capita (3).

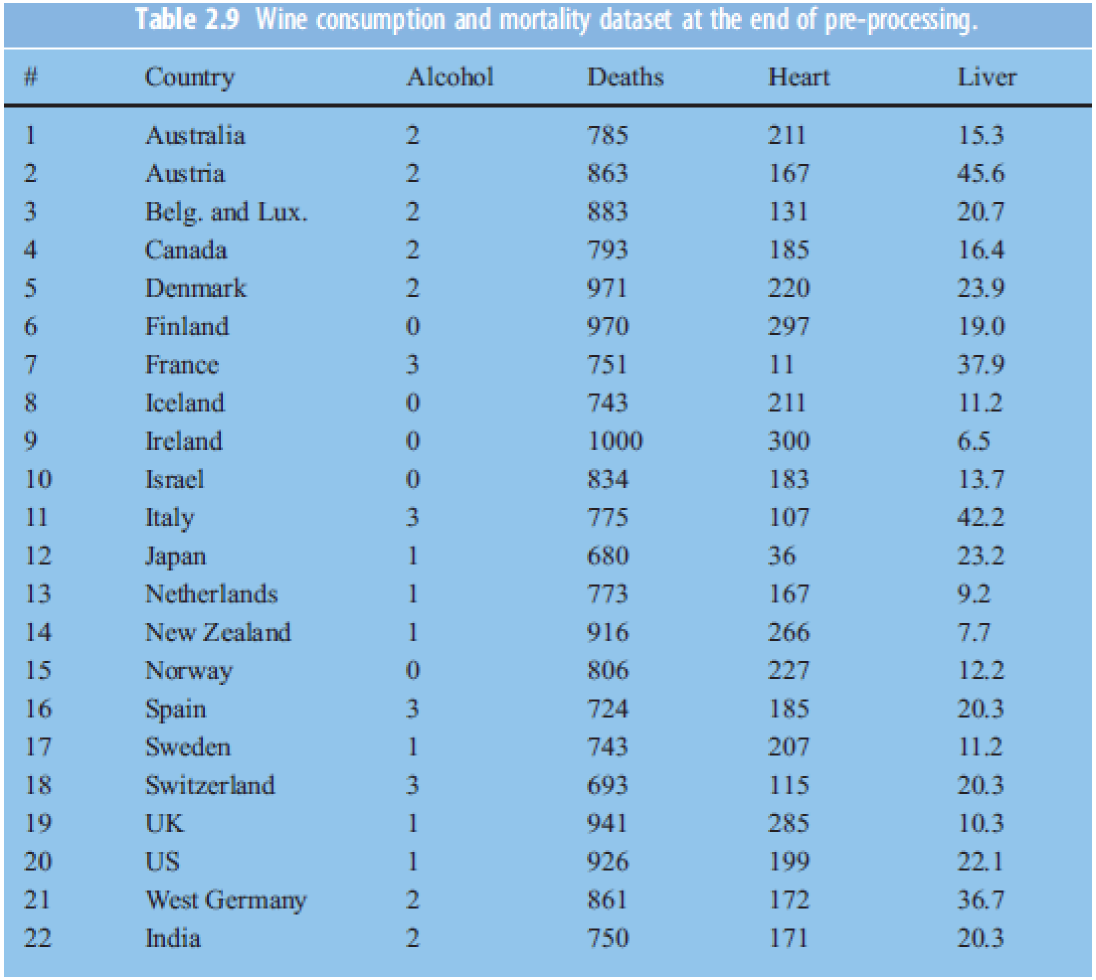


Table 2.9 Wine consumption and mortality dataset at the end of pre-processing.

#	Country	Alcohol	Deaths	Heart	Liver
1	Australia	2	785	211	15.3
2	Austria	2	863	167	45.6
3	Belg. and Lux.	2	883	131	20.7
4	Canada	2	793	185	16.4
5	Denmark	2	971	220	23.9
6	Finland	0	970	297	19.0
7	France	3	751	11	37.9
8	Iceland	0	743	211	11.2
9	Ireland	0	1000	300	6.5
10	Israel	0	834	183	13.7
11	Italy	3	775	107	42.2
12	Japan	1	680	36	23.2
13	Netherlands	1	773	167	9.2
14	New Zealand	1	916	266	7.7
15	Norway	0	806	227	12.2
16	Spain	3	724	185	20.3
17	Sweden	1	743	207	11.2
18	Switzerland	3	693	115	20.3
19	UK	1	941	285	10.3
20	US	1	926	199	22.1
21	West Germany	2	861	172	36.7
22	India	2	750	171	20.3

Data Analysis and Data Analytics

Analysis is the detailed examination of the elements or structure of something.

"Analytics" is the systematic computational analysis of data or statistics.

Data Analysis helps in understanding the data and provides required insights from the past to understand what happened so far

Data Analytics is the process of exploring the data from the past to make appropriate decisions in the future by using valuable insights

Data Analysis & Data Analytics

Descriptive Analysis => reveal what happened in the past

- Typically, it is the first kind of data analysis performed on a dataset.
- Usually it is applied to large volumes of data, such as census data.
- Description and interpretation processes are different steps.

- Eg, to categorize customers by their likely product preferences and purchasing patterns
- social media marketing campaign, use descriptive analytics to assess the number of posts, mentions, followers, fans, page views, reviews, or pins

Data Analysis & Data Analytics

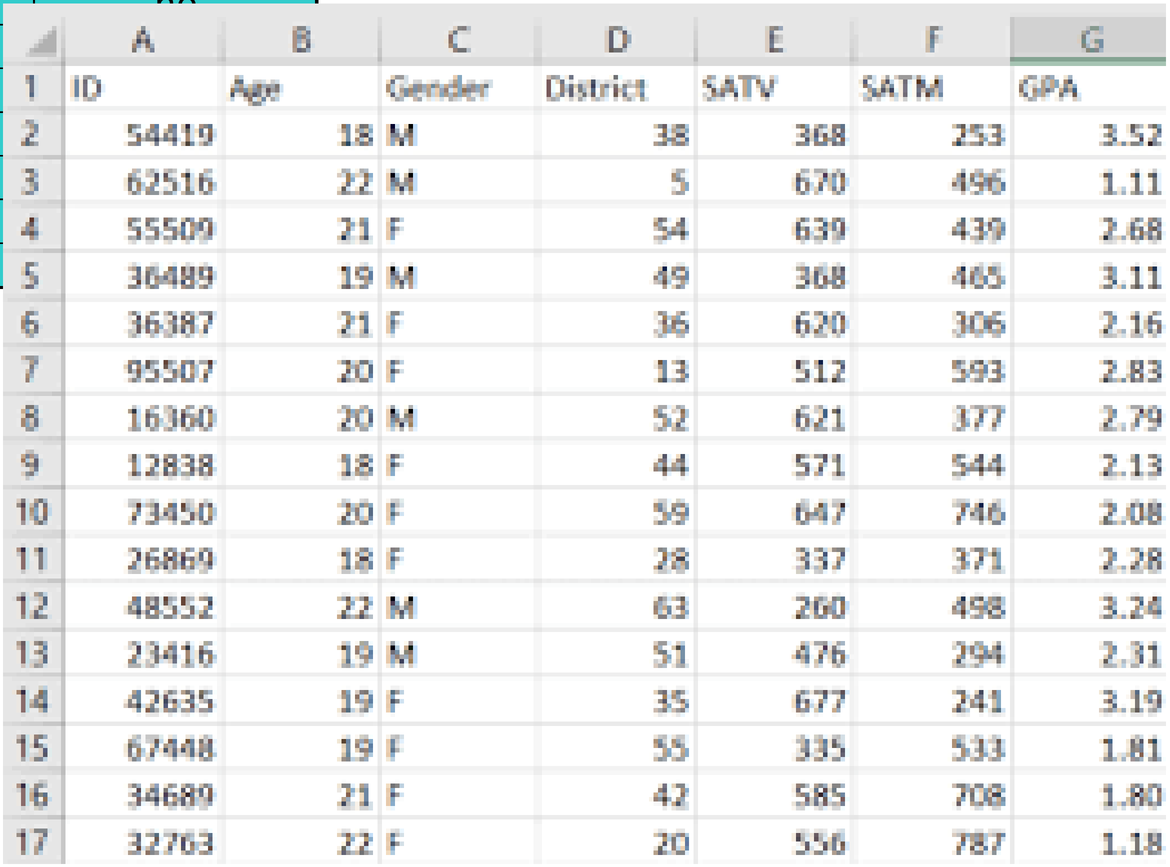
Descriptive Analysis

Type of Variable => categorical variable, Ordinal, continuous variable, ratio ....

- Independent variable/ Predictor variable,
- Dependent variable / Outcome var/ Decision var/ class label/ Target class

Dataset

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	
>40	medium	yes	fair	
<=30	medium	yes	excellent	
31…40	medium	no	excellent	
31…40	high	yes	fair	
>40	medium	no	excellent	



	A	B	C	D	E	F	G
1	ID	Age	Gender	District	SATV	SATM	GPA
2	54419	18	M	38	368	253	3.52
3	62516	22	M	5	670	496	1.11
4	55509	21	F	54	639	439	2.68
5	36489	19	M	49	368	465	3.11
6	36387	21	F	36	620	306	2.16
7	95507	20	F	13	512	593	2.83
8	16360	20	M	52	621	377	2.79
9	12838	18	F	44	571	544	2.13
10	73450	20	F	59	647	746	2.08
11	26869	18	F	28	337	371	2.28
12	48552	22	M	63	260	498	3.24
13	23416	19	M	51	476	294	2.31
14	42635	19	F	35	677	241	3.19
15	67448	19	F	55	335	533	1.81
16	34689	21	F	42	585	708	1.80
17	32763	22	F	20	556	787	1.18

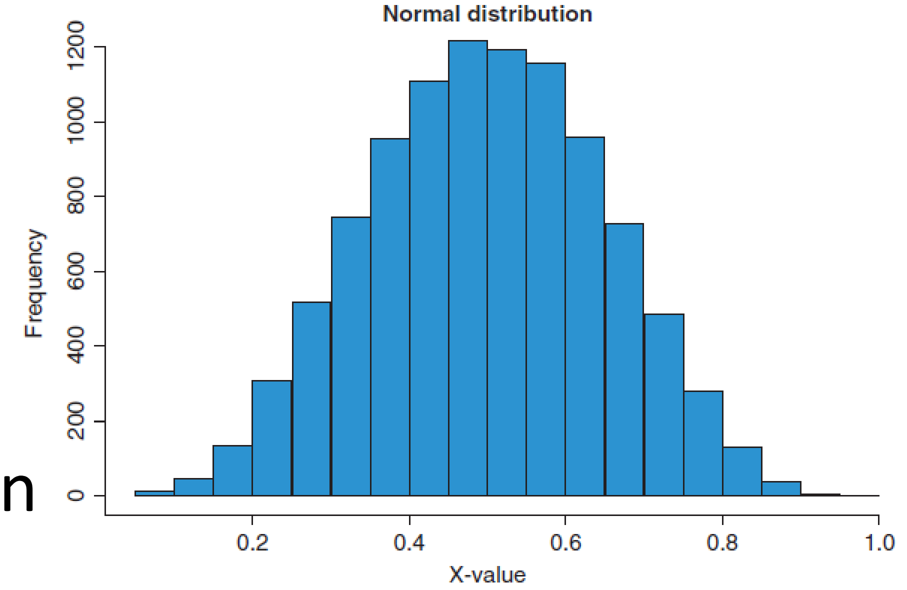
Data Analysis & Data Analytics

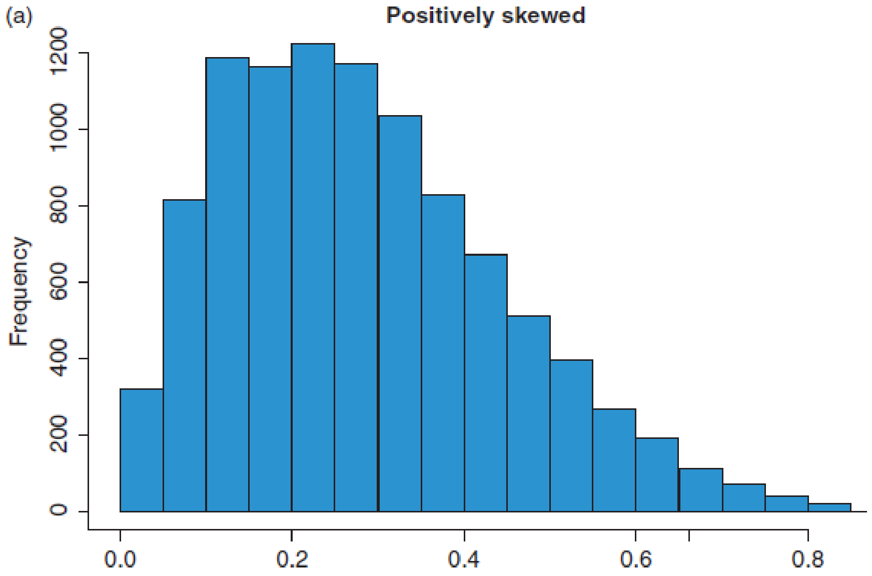
Frequency Distribution

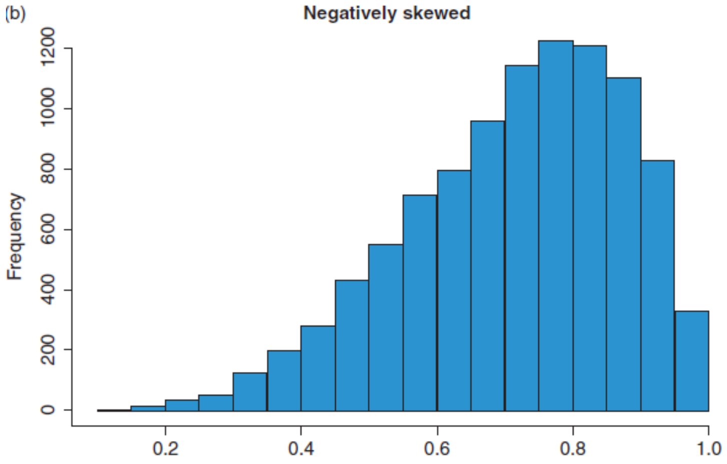
- Histogram
- Pie Chart

Distribution of Data

Normal & Skewed Distribution







Data Analysis & Data Analytics

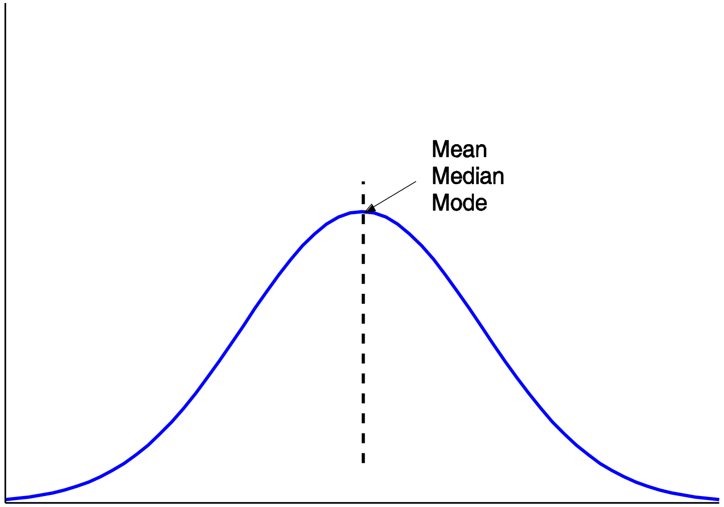
Skewed Distribution

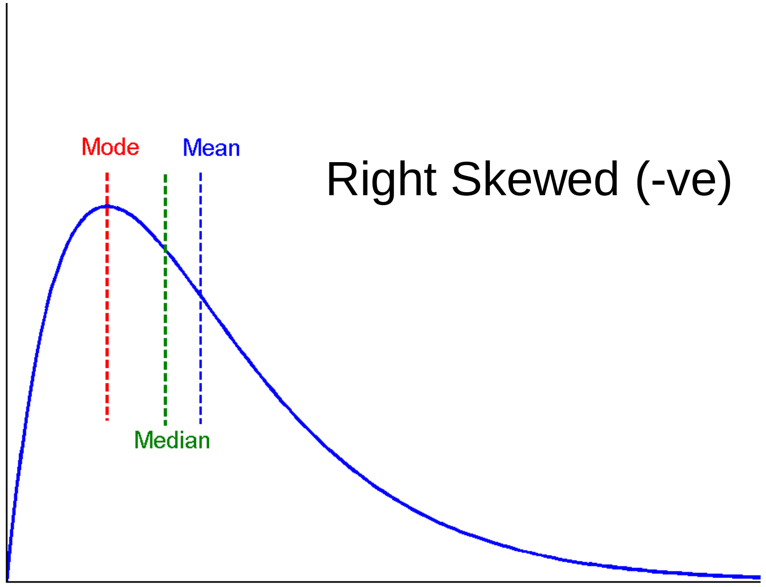
- Cricket Score
- Exam Results – online vs offline; Lab v Theory
- Average Income distribution
- Human Life cycle
- Taxation Regimes
- Record of Long Jumps at a Competition

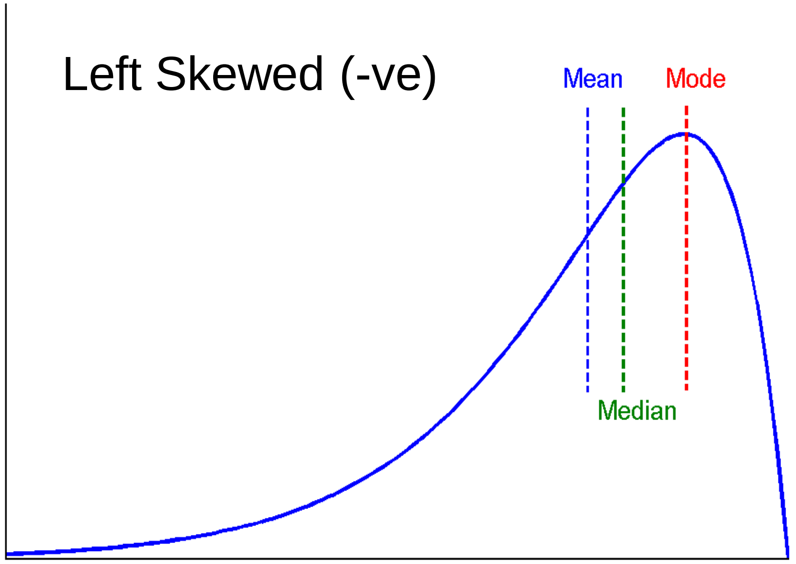
Data Analysis & Data Analytics

Measures of Centrality

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




Right Skewed (-ve)


Left Skewed (-ve)

Data Analysis & Data Analytics

Dispersion of a Distribution

- Range => largest score - smallest score.

  Disadvantage: it uses only the highest and lowest values, extreme scores or outliers tend to result in an inaccurate picture of the more likely range.

- Interquartile range is defined as the difference between the 25th and 75th percentile

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots

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

Inter-quartile range: IQR = $Q_3 - Q_1$

Five number summary:

min, Q_1, median, Q_3, max

Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually

Outlier: usually, a value higher/lower than 1.5 x IQR

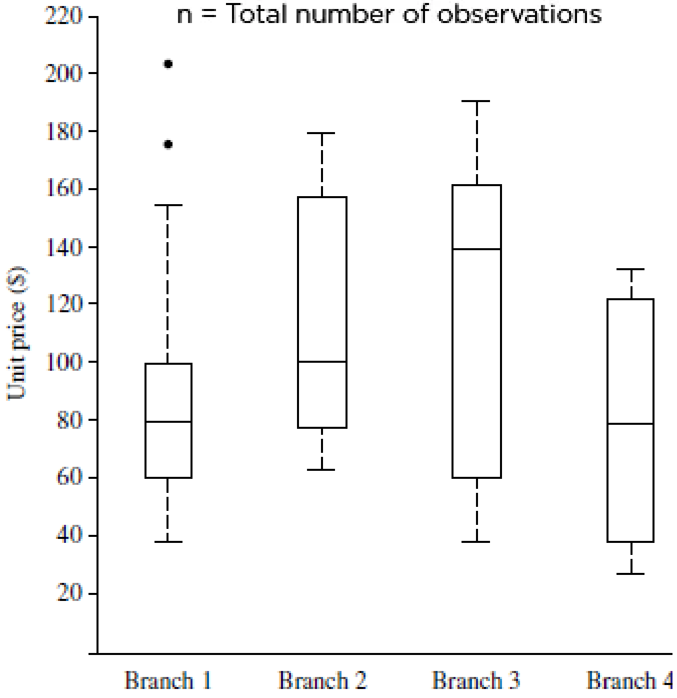
Beyond this range - Outlier

Q1-1.5*IQR to Q3+1.5*IQR

$$P_x = \frac{x\,(n + 1)}{100}$$

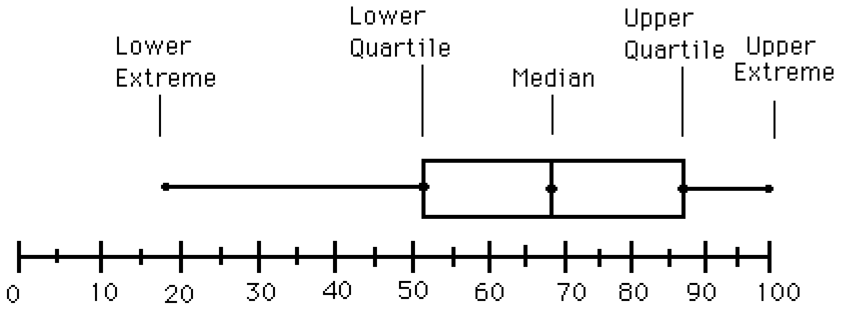
P_x = The value at which x percentage of data lie below that value

n = Total number of observations



Boxplot Analysis

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum



The following are the scores of Coding test of 12 members. Draw the box plot & also find out is there any outliers, according to our rule of thumb?

3,40,41,45,40,60,61,62,63,65,70,99

Data Analysis & Data Analytics

- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)
  - variance of a population (σ^2)

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

  - variance of a sample (s^2)

$$s^2 = \frac{\sum (x_i - x)^2}{(n-1)}$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

$$s = \sqrt{\frac{\sum (x_i - x)^2}{(n-1)}}$$

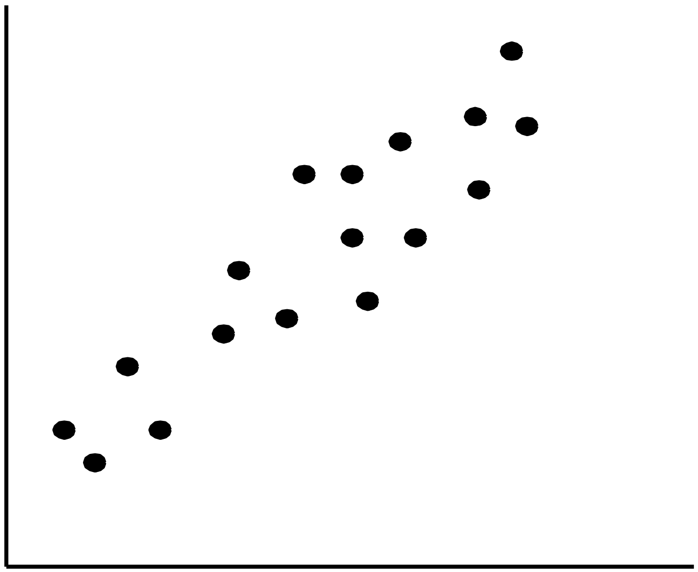
Data Analysis & Data Analytics

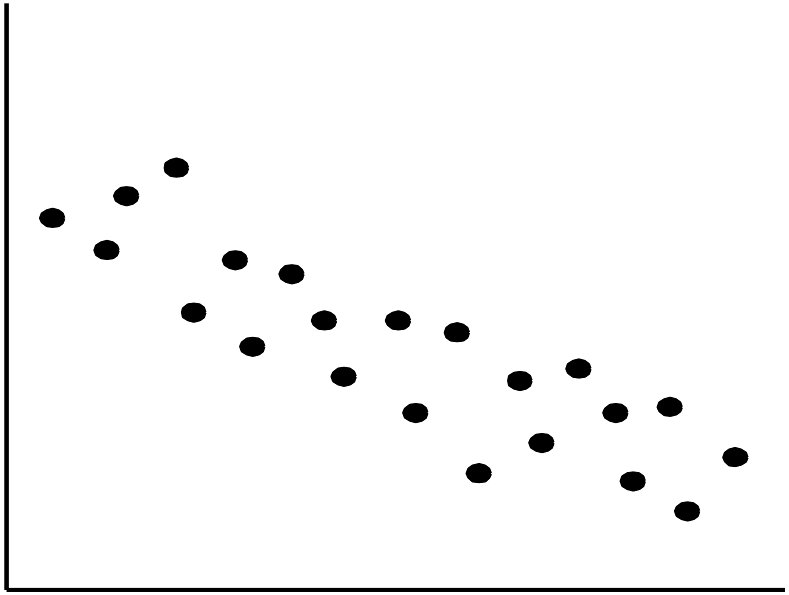
Diagnostic Analytics

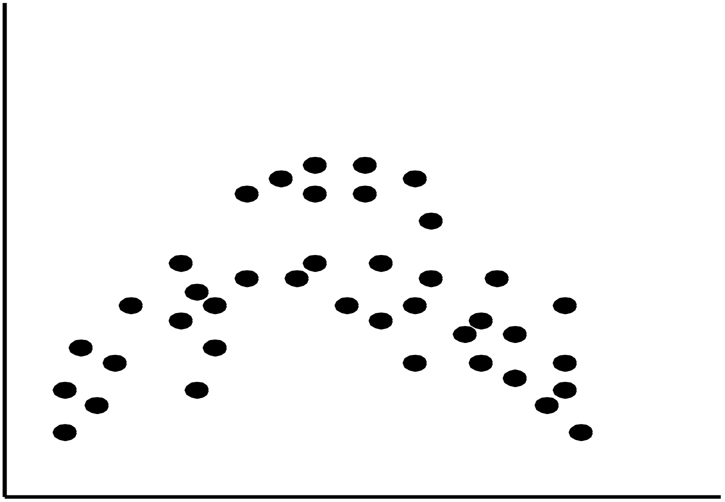
- used for discovery, or to determine why something happened Eg “rain” vs “umbrella”
- Correlations - statistical analysis that is used to measure and describe the strength and direction of the relationship between two variables.

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

Positively and Negatively Correlated Data

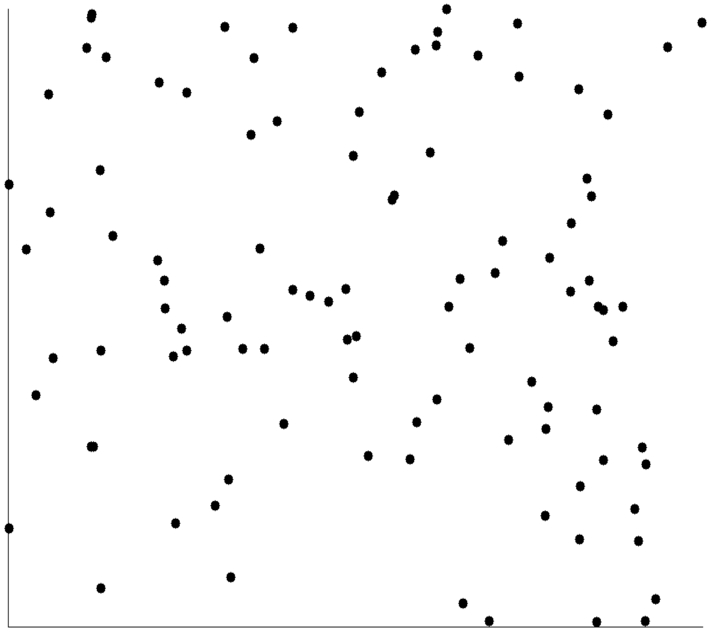


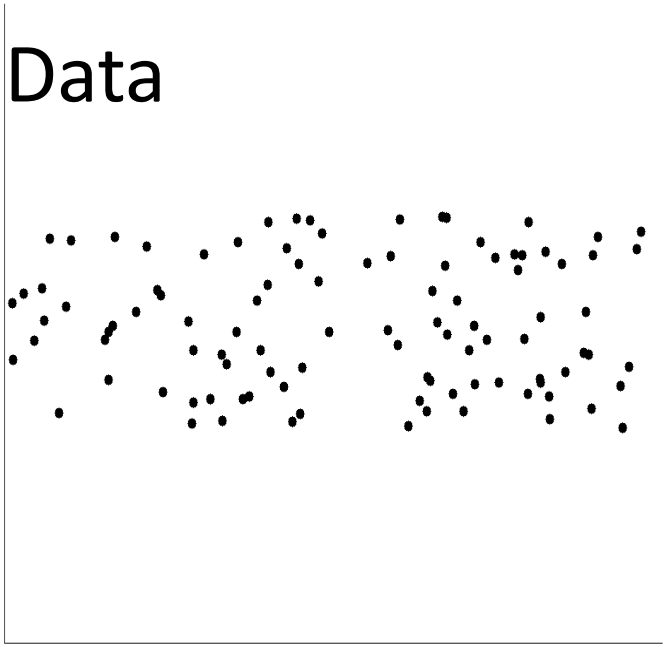


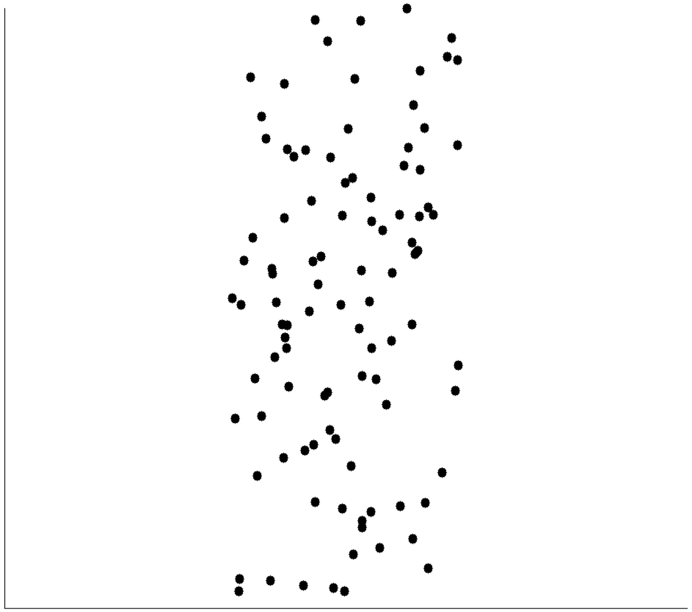


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data

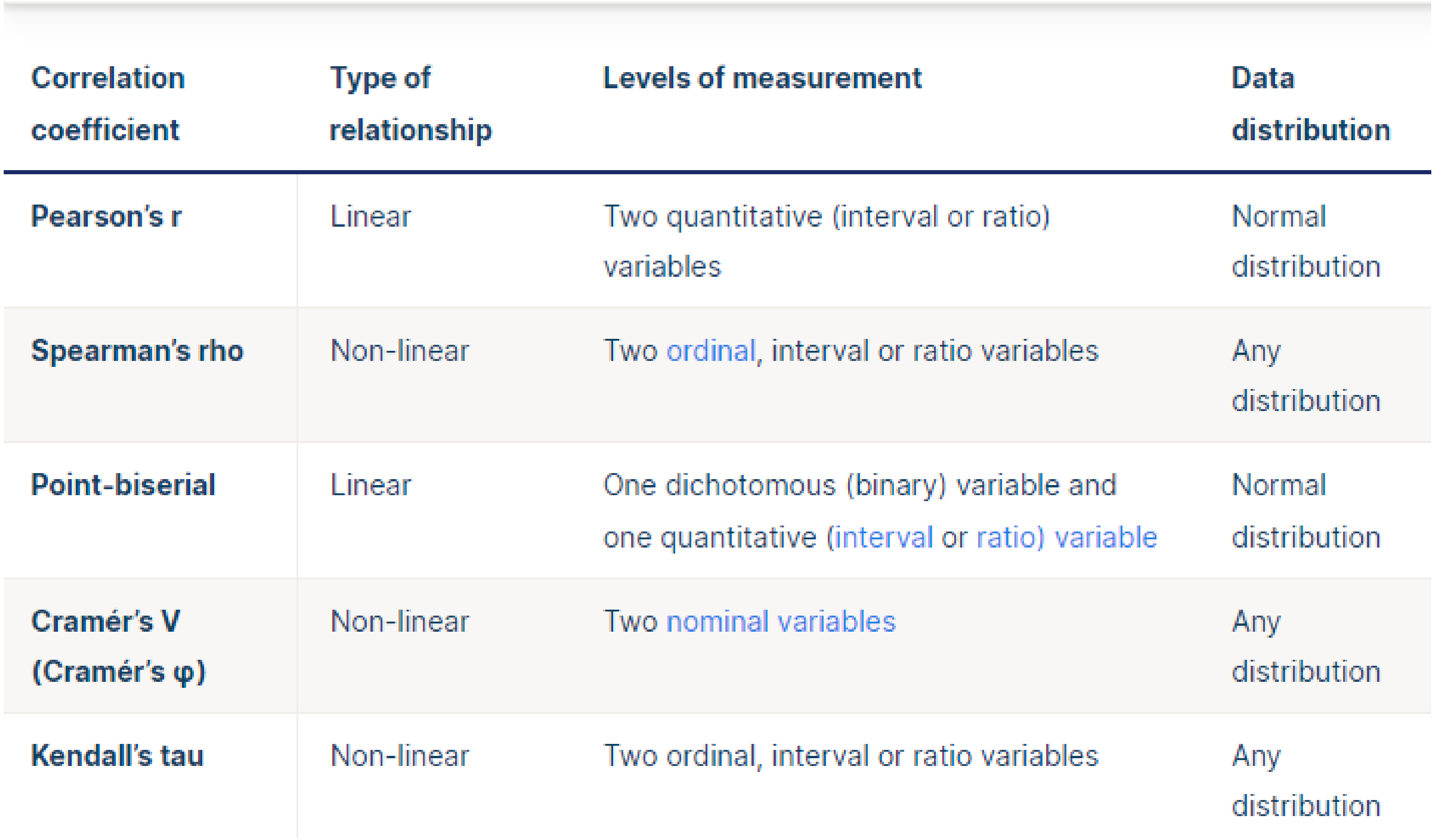






Data Analysis & Data Analytics

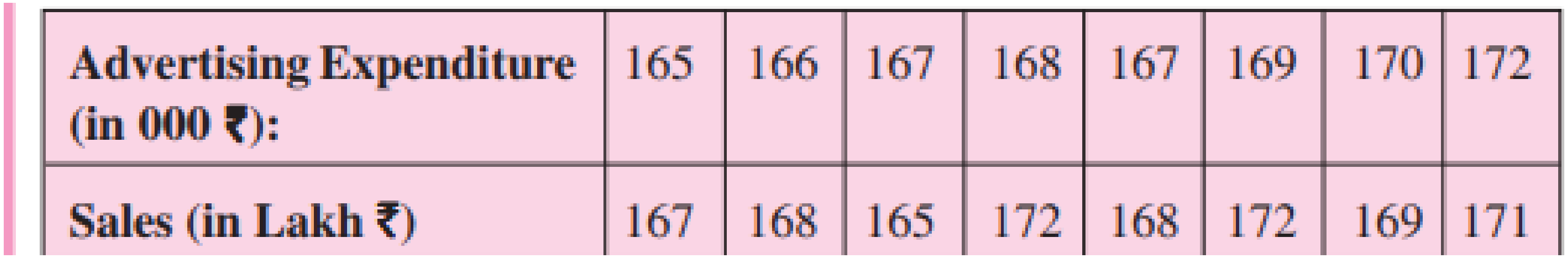
Diagnostic Analytics



Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

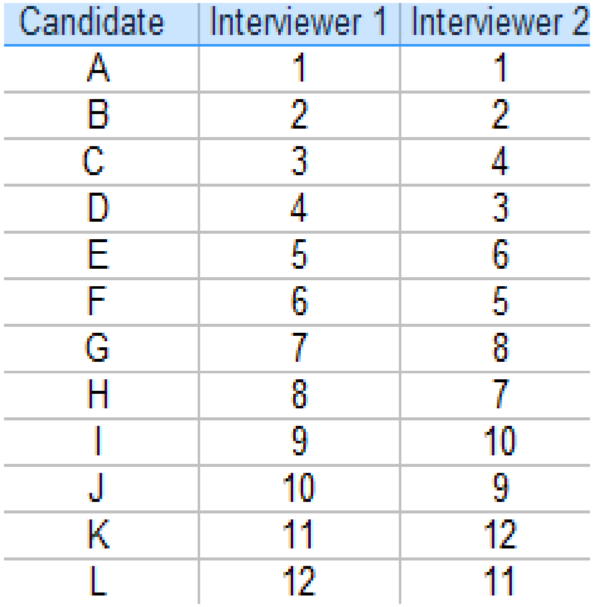
Data Analysis & Data Analytics

Find Correlation between the attributes



Advertising Expenditure (in 000 ₹):	165	166	167	168	167	169	170	172
Sales (in Lakh ₹)	167	168	165	172	168	172	169	171

Two interviewers ranked 12 candidates (A through L) for a position. Find Correlation among

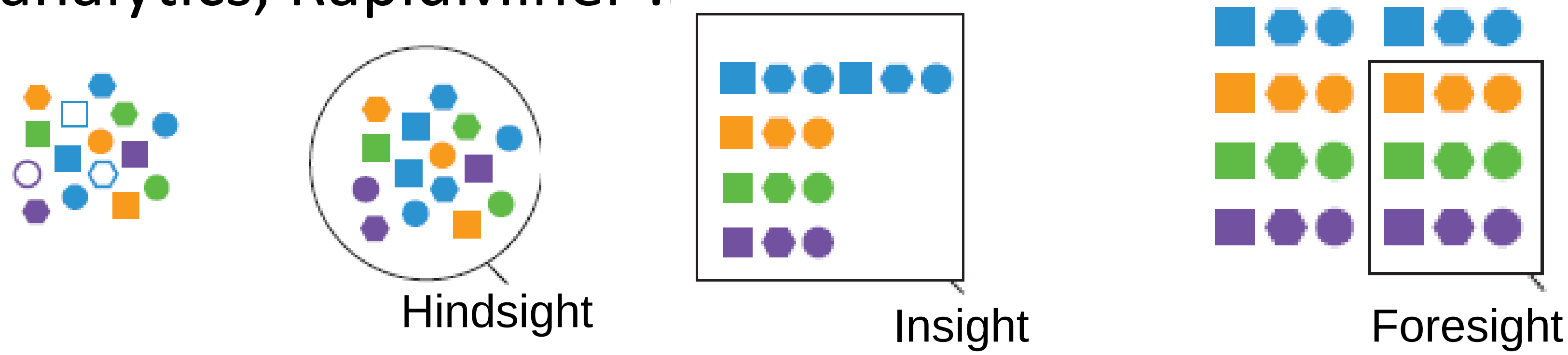


Candidate	Interviewer 1	Interviewer 2
A	1	1
B	2	2
C	3	4
D	4	3
E	5	6
F	6	5
G	7	8
H	8	7
I	9	10
J	10	9
K	11	12
L	12	11

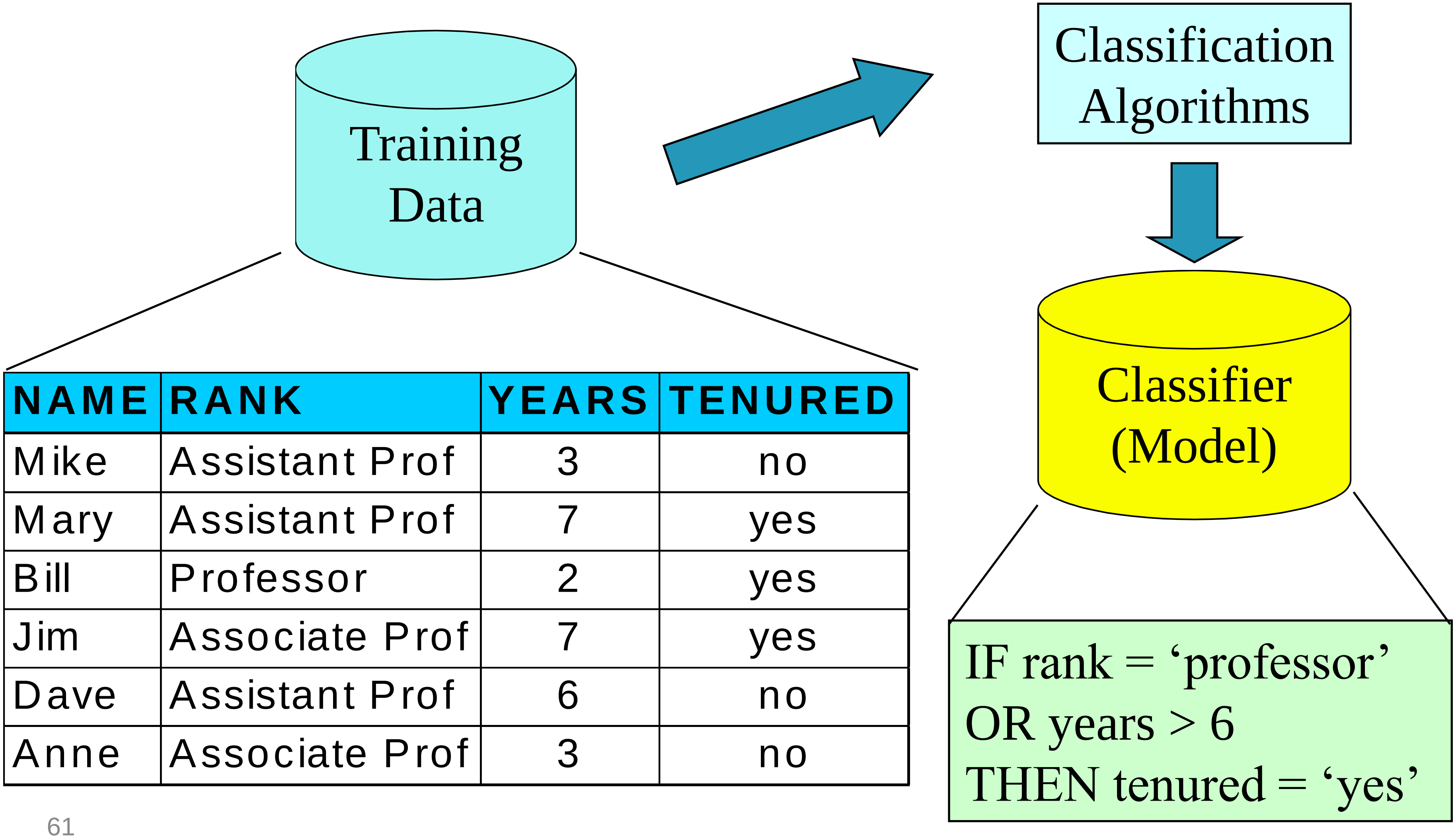
Data Analysis & Data Analytics

Predictive Analytics

- understanding the future using the data and the trends we have seen in the past
- no statistical algorithm can “predict” the future with 100% certainty because the foundation of predictive analytics is based on probability
- predictive analytics software : SAS, IBM predictive analytics, RapidMiner .



Predictive model Construction



Training Data

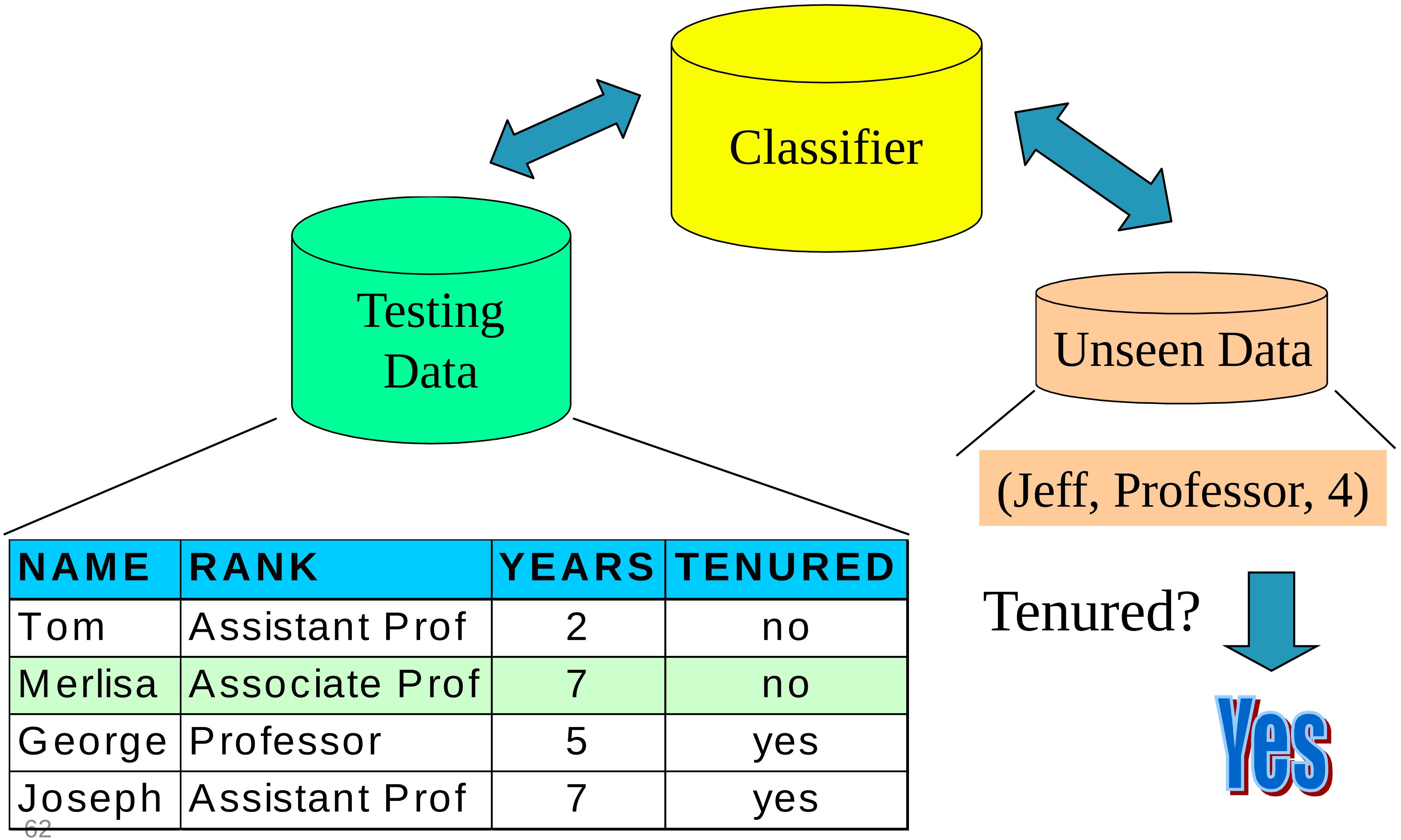
Classification Algorithms

Classifier (Model)

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Using the Model in Prediction



Classifier

Testing Data

Unseen Data

(Jeff, Professor, 4)

Tenured?

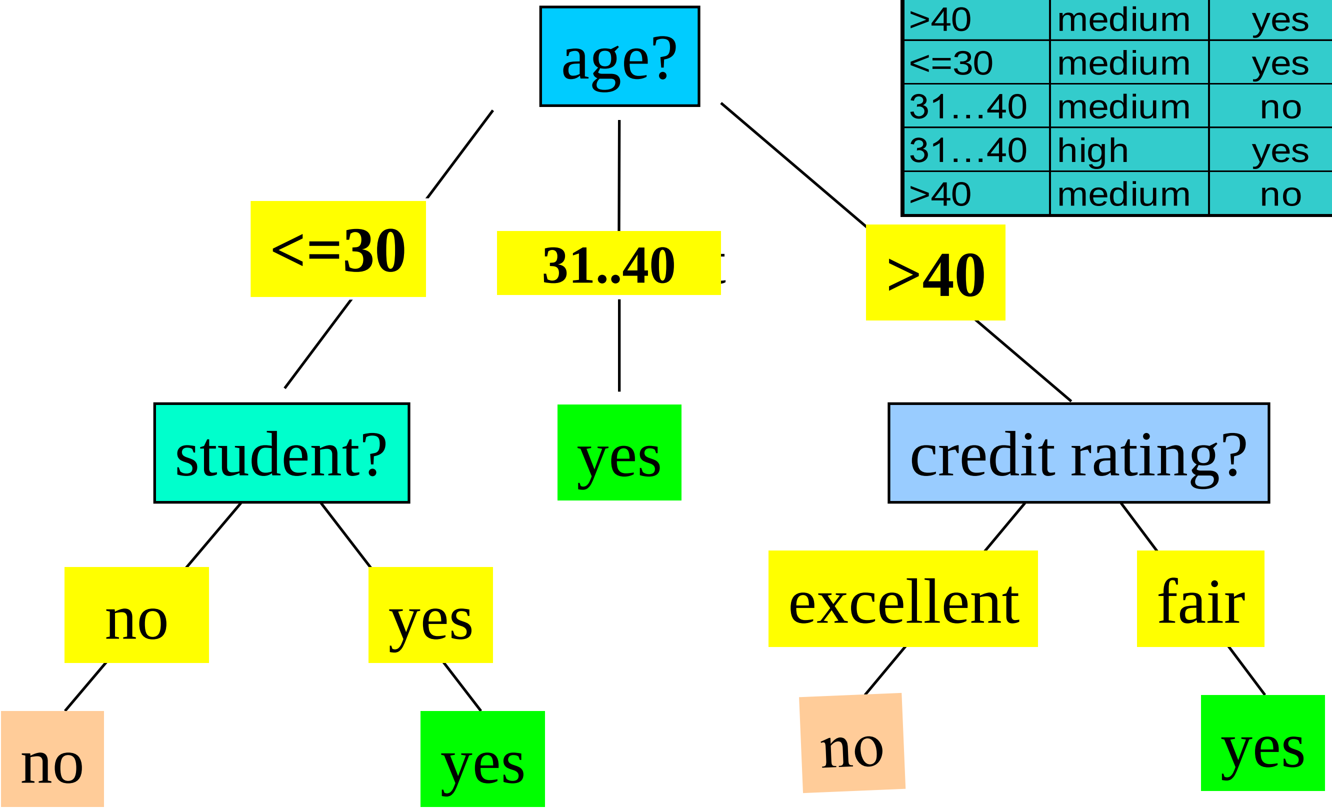
Yes

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Dataset & Model

Predictive Analytics

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Data Analysis & Data Analytics

Prescriptive Analytics

- analyzes potential decisions, the interactions between decisions, the influences that bear upon these decisions, and the bearing all of this has on an outcome to ultimately prescribe an optimal course of action
- **the process of using current and historical data to identify trends and relationships**.
- Techniques include optimization, simulation, game theory, and decision-analysis methods
- [Gartner] 13% of organizations are using predictive analytics, but only 3% are using prescriptive analytics.

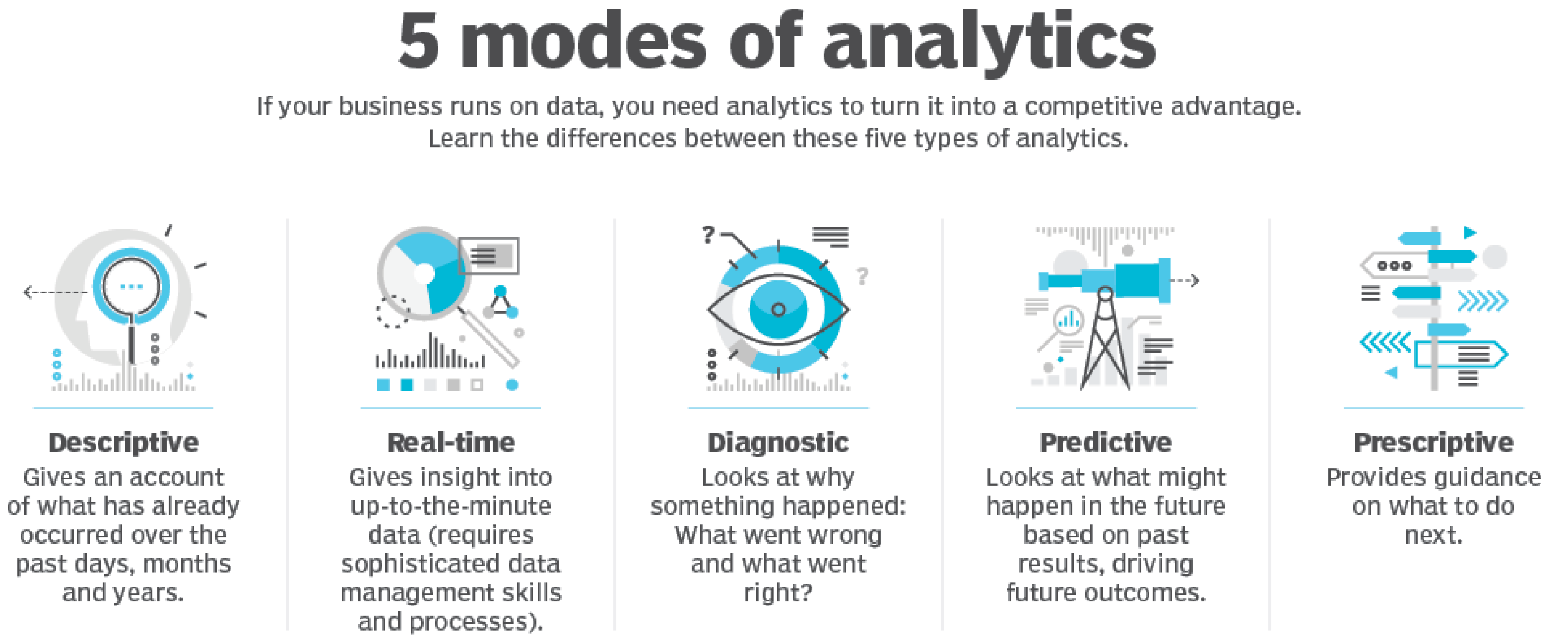
Data Analysis & Data Analytics

- **Exploratory analysis** is an approach to analyzing datasets to find previously unknown relationships.
- involves using various data visualization approaches.
- exploratory analysis is about the methodology or philosophy of doing the analysis, rather than a specific technique

Data Analysis & Data Analytics

Mechanistic Analysis

- understanding the exact changes in variables that lead to changes in other variables for individual objects(studying a relationship between two variables)
- Regression => process for estimating the relationships among variables
- Corelation vs Regression
- Correlation by itself does not provide any indication of how one variable can be predicted from another. But Regression provides



5 modes of analytics

If your business runs on data, you need analytics to turn it into a competitive advantage. Learn the differences between these five types of analytics.

Descriptive
Gives an account of what has already occurred over the past days, months and years.

Real-time
Gives insight into up-to-the-minute data (requires sophisticated data management skills and processes).

Diagnostic
Looks at why something happened: What went wrong and what went right?

Predictive
Looks at what might happen in the future based on past results, driving future outcomes.

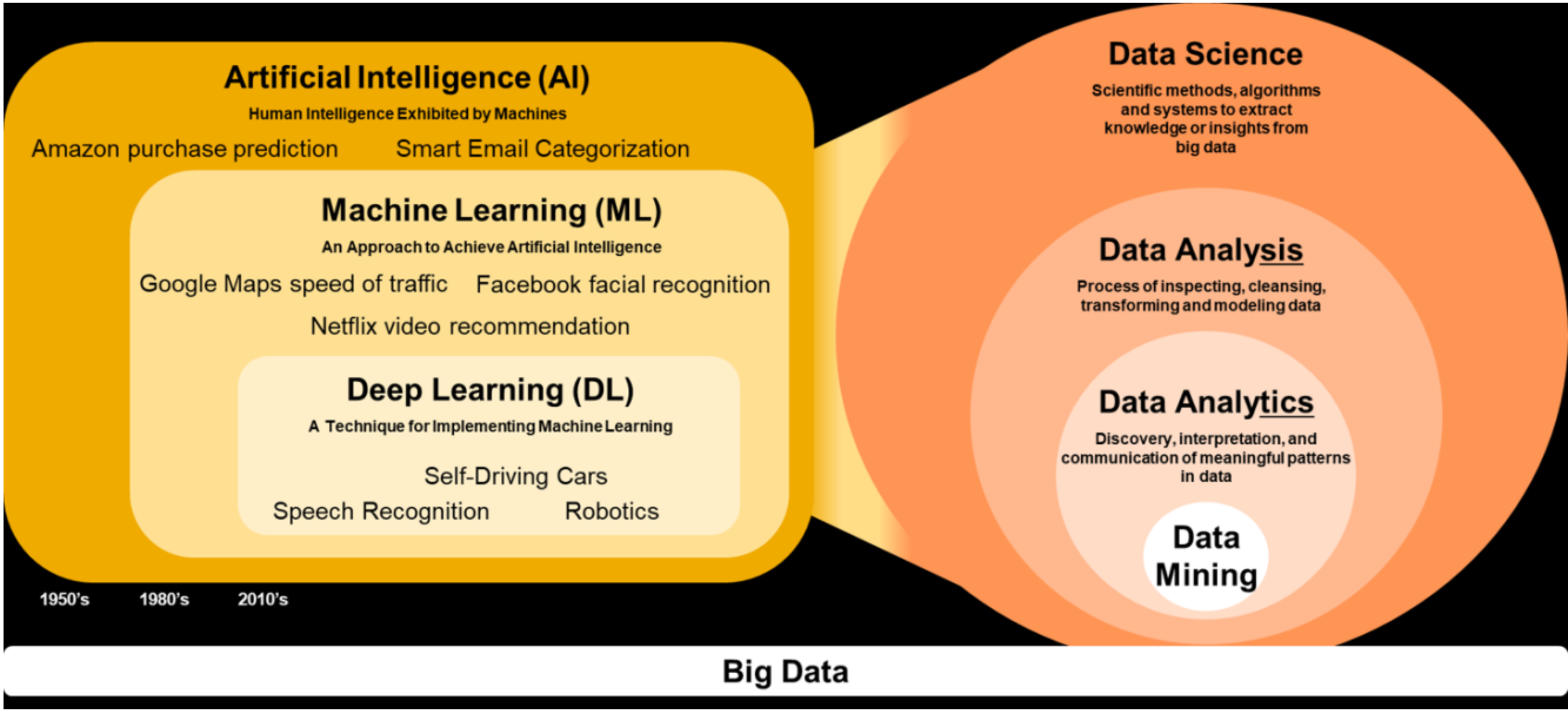
Prescriptive
Provides guidance on what to do next.

Source: https://www.techtarget.com/searchcio/definition/Prescriptive-analytics

Summary - Analytics

- **Descriptive Analytics** tells you what happened in the past.
- **Diagnostic Analytics** helps you understand why something happened in the past.
- **Predictive Analytics** predicts what is most likely to happen in the future.
- **Prescriptive Analytics** recommends actions you can take to affect those outcomes.

AI vs ML vs DL vs DS



Artificial Intelligence (AI)

Human Intelligence Exhibited by Machines

Amazon purchase prediction

Smart Email Categorization

Machine Learning (ML)

An Approach to Achieve Artificial Intelligence

Google Maps speed of traffic

Facebook facial recognition

Netflix video recommendation

Deep Learning (DL)

A Technique for Implementing Machine Learning

Self-Driving Cars

Speech Recognition

Robotics

1950's

1980's

2010's

Data Science

Scientific methods, algorithms and systems to extract knowledge or insights from big data

Data Analysis

Process of inspecting, cleansing, transforming and modeling data

Data Analytics

Discovery, interpretation, and communication of meaningful patterns in data

Data Mining

Big Data

AI, ML, DL and Data Science with Data Analysis, Data Analytics and Data Mining - all based on the foundation of #BigData

Data Science - Scientific methods, algorithms and systems to extract knowledge or insights from big data

- Also known as Predictive or Advanced Analytics
- Algorithmic and computational techniques and tools for handing large data sets
- Increasingly focused on preparing and modeling data for ML & DL tasks
- Encompasses statistical methods, data manipulation and streaming technologies (e.g. Spark, Hadoop)
- Key skill and tools behind building modern AI technologies

Data Analysis - Process of inspecting, cleansing, transforming and modeling data

Data Analytics - Discovery, interpretation, and communication of meaningful patterns in data

Data Mining - Process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems

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
- You can afford to scan the data once

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

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

- https://colab.research.google.com/drive/1kucNxA3sD3A_qyZp9OwRi_V8HVkGsiOl#scrollTo=80zUqqGRuivN