KD44103 Big Data Analytics

KD44103 Big Data Analytics
Faculty of Computing & Informatics,
Universiti Malaysia Sabah

UMS
UNIVERSITI MALAYSIA SABAH

Learning outcomes

- Understand the concept of data, data analysis and Terminology
- Understand the basic activities in data analytics and Business Drivers for Analytics
- Understand the four types of Analytics and data analytics framework
- Understand the Statistical Analysis

Outline

- Data and data sources
- Data Analytics, concepts and terminology
- activities in data analytics
- Examples of Big Data Analytics
- Data analytics framework
- Statistical Analysis

Data

- Data sources
  - Service delivery statistics
  - Census
  - Surveys, evaluations, research studies
  - Sentinel surveillance
  - Budget information
- Data vs. information =

  unsynthesized vs. synthesized

What is Data analysis?

- The way information and results are interpreted and assessed
  - Assigning meaning to figures, stories, observations, etc that have been gathered and recorded.
  - Conceptual frameworks guide data analysis.
  - Data analysis possible by hand or computer (various packages, e.g., SPSS; etc.)

Data Analytics

- **Data Analytics (DA):** is the science of examining row data with the purpose of drawing conclusions about the information.
- **Often involves studying past historical data to:**
  - ➢ Research potential trends
  - ➢ Analysis the effects of certain decisions or events, or
  - ➢ Evaluate the performance of a given tools or scenario
- **Goal:** To improve the business by gaining knowledge which can be used to make improvements

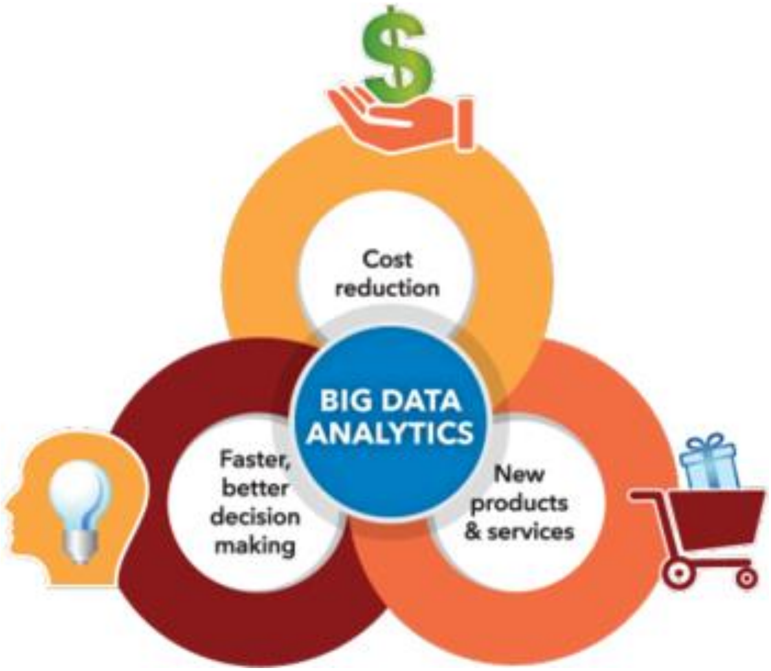
Concepts and Terminology

- ***Data Science****:* Is the professional field that deals with turning data into value, such as new insights or predictive models.
- ***Data Mining***: defined as the process of discovering patterns in data.
- ***Machine Learning***: Is a type of artificial intelligent (AI) that provides computers with the ability to learn without being explicitly programmed.
- ***Datasets***
  - Collections or groups of related data are generally referred to as datasets.
  - Each group or dataset member shares the same set of attributes or properties as others in the same dataset.
  - Examples:
    - ➢ Tweets stored in a flat file
    - ➢ A collection of image files in a directory
    - ➢ An extract of rows from a database table stored in a CSV formatted file
    - ➢ Historical weather observations that are stored as XML files

Concepts and Terminology- cont’d

- ***Algorithms*****:** Algorithm is a set of steps for a computer program to accomplish a tasks.
- **Business Intelligence** (BI): is a technology-driven process for analyzing data and presenting actionable information to help make informed business decisions.
- ***NoSQL database***: provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.
- ***Data Warehouse***: is a collection of corporate information and data derived from operational systems and external data sources.

WHY DO WE NEED DATA ANALYTICS



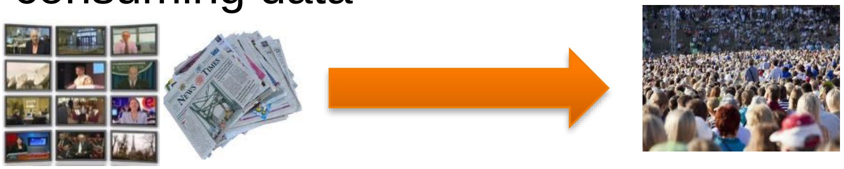
ACTIVITIES IN DATA ANALYTICS

- Foremost, bring the data in the environment.
- **Tidy** the data, such that each column is a variable and each row is an observation.
- **Transform** the data, includes narrowing observations of interest.
- **Visualize** the data, to explore possible relationships
- **Models** are complementary tools to visualization.
- **Communicate** the results

The Model Has Changed…

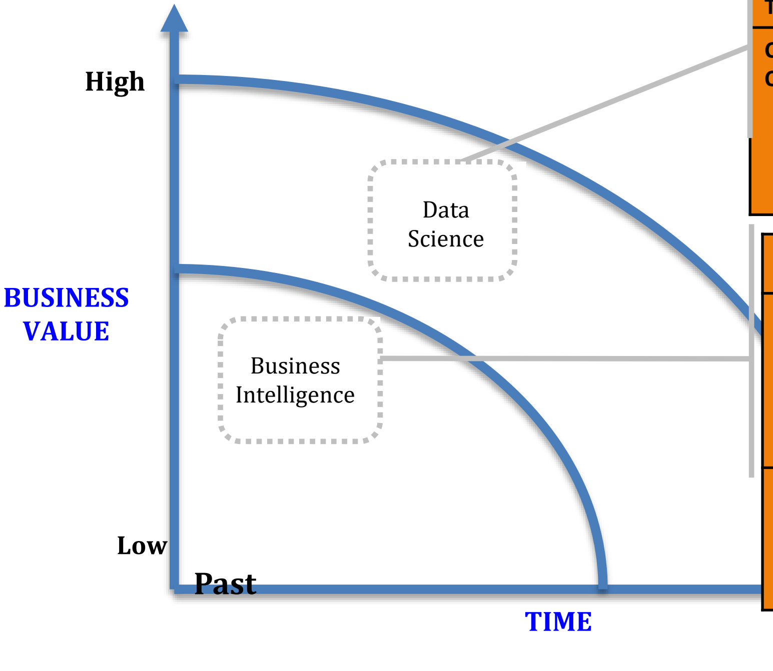
- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data

Business Intelligence vs. Data Science



Predictive Analytics & Data Mining (Data Science)	
Typical Techniques & Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large data sets
Common Questions	• What if…..? • What's the optimal scenario for our business ? • What will happen next? What if these trends continue? Why is this happening?

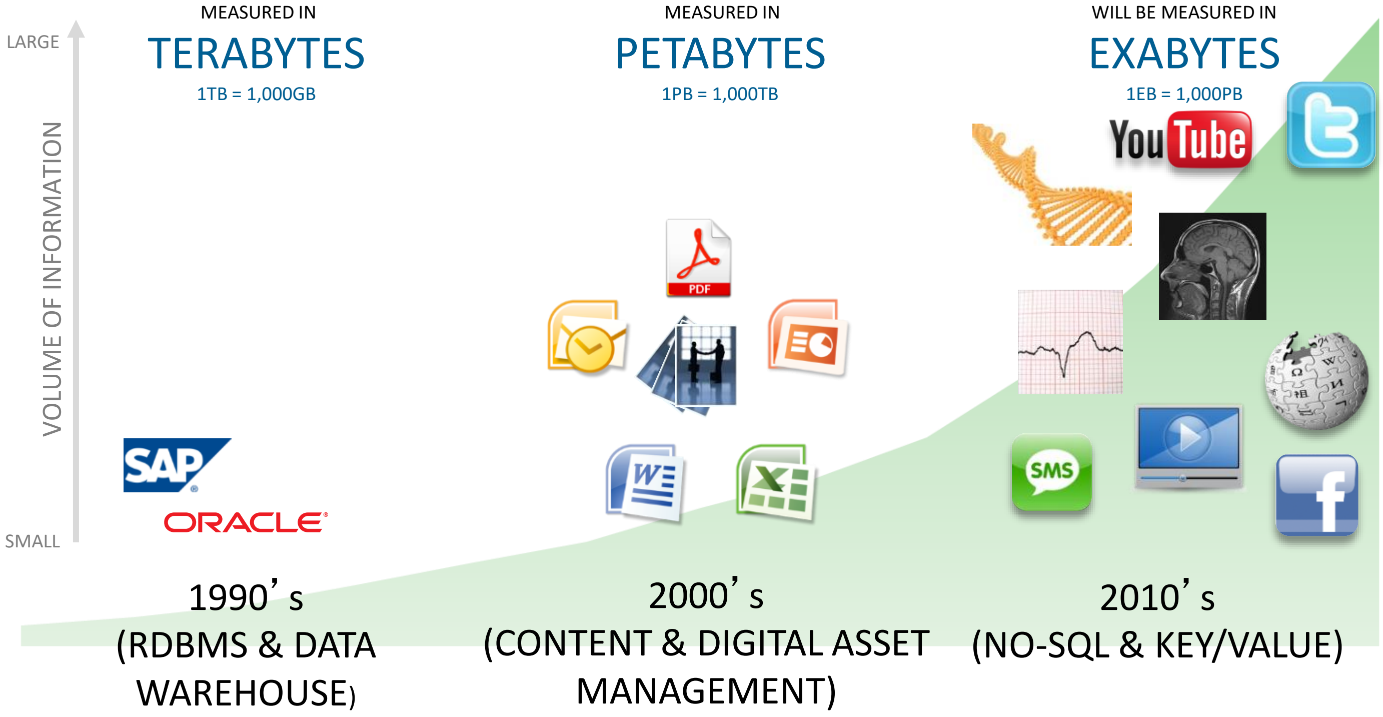
Business Intelligence	
Typical Techniques & Data Types	• Standard and ad hoc reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable data sets
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

Business Drivers for Analytics

Current Business Problems Provide Opportunities for Organizations to Become More Analytical & Data Driven

	Driver	Examples
1	Desire to optimize business operations	Sales, pricing, profitability, efficiency
2	Desire to identify business risk	Customer churn, fraud, default
3	Predict new business opportunities	Upsell, cross-sell, best new customer prospects
4	Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II

Opportunities for a New Approach to Analytics



MEASURED IN
TERABYTES
1TB = 1,000GB

MEASURED IN
PETABYTES
1PB = 1,000TB

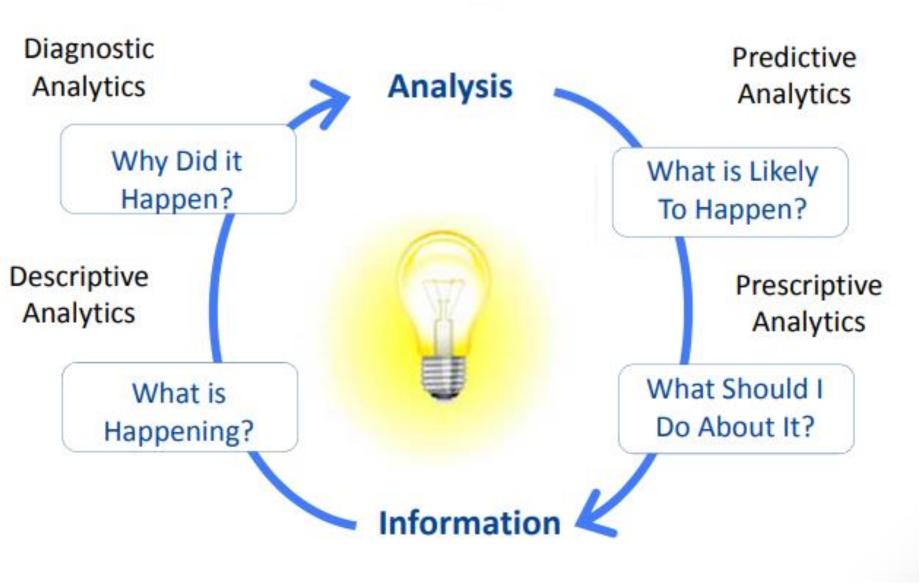
WILL BE MEASURED IN
EXABYTES
1EB = 1,000PB

1990’s
(RDBMS & DATA WAREHOUSE)

2000’s
(CONTENT & DIGITAL ASSET MANAGEMENT)

2010’s
(NO-SQL & KEY/VALUE)

Examples of Big Data Analytics



Gartner

Predictive analytics

(What is likely to happen?)

- To determine the outcome of an event that might occur in the **future**.
- Generate knowledge that conveys how that information is related.
- Form the basis of models that are used to generate future predictions based upon past events.
- Sample questions:
  - ❑ What are the chances that a customer will default on a loan if they have missed a monthly payment?
  - ❑ What will be the patient survival rate if Drug B is administrated instead of Drug A?
  - ❑ If a customer has purchased Product A and B, what are the chances that they will also purchase Product C?
- Predicts the outcomes of events and predictions are made on patterns, trends and exceptions found in historical and current data. This leads to the identification of both risks and opportunities.
- Use of Large datasets comprised internal and external data and various data analysis techniques.
- Greater value and required more skillset.
- Statistical tools with user-friendly front-end interface.

Prescriptive Analytics

What Should I Do About It?

- Build upon the results of predictive analytics by prescribing actions that should be taken.
- Focus is not only on which prescribed option is best to follow, but why.
- Provide results that can be reasoned about because they embed elements of situational understanding.
- Can be used to gain an advantage or mitigate a risk.
- Sample questions:
  - ❑ Among three drugs, which one provides the best results?
  - ❑ When is the best time to trade a particular stock?
- Provide more value and require most advanced skillset and well as specialized software and tools.
- Shift from explanatory to advisory. Simulation of various scenarios.
- Incorporates internal and external data.
- Internal data: customer information, current and historical sales data
- External data: social media data, weather forecasts, and government –produced data.

Descriptive Analytics

What is Happening?

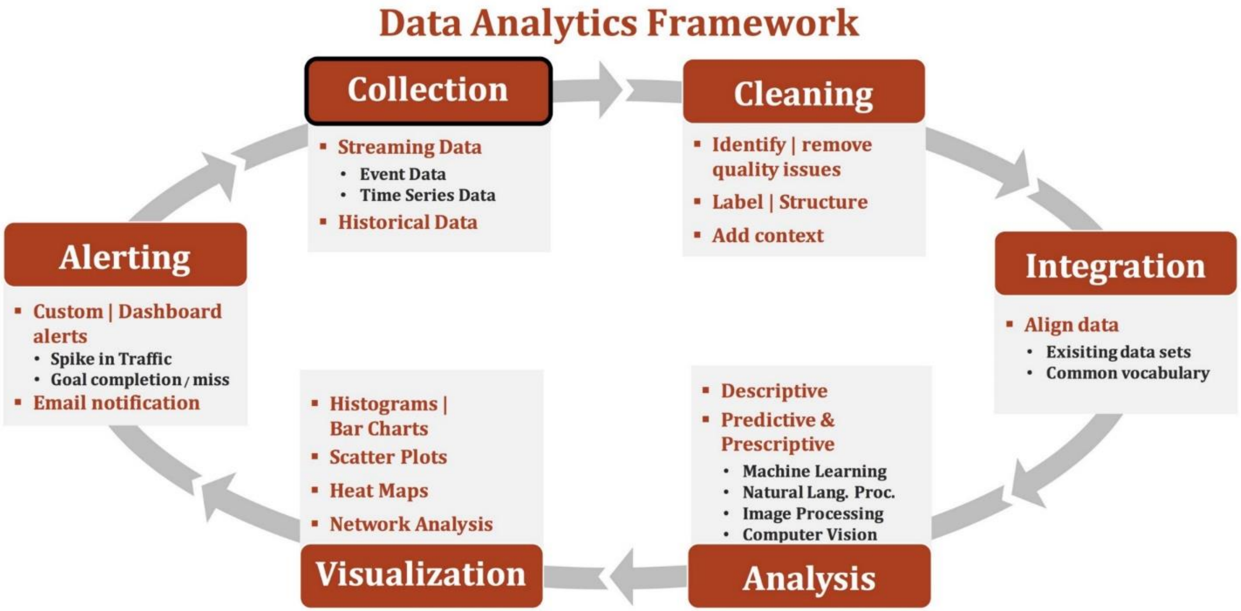
- Answer questions about events that **have already occurred.**
- Generate information
- Sample questions:
  - ❑ What was the sales volume over the 12 months
  - ❑ What is the number of support calls received as categorized by severity and geographic location?
  - ❑ What is the monthly commission earned by each sales agent?
- 80% of generated analytics results are descriptive in nature
- Least worth and require a relatively basic skillset.

Diagnostic Analytics

Why Did it Happen?

- Determine **the cause** of a phenomenon that occurred in the past questions that focus on the reason behind the event.
- Sample questions:
  - ❑ Why were Q2 sales less than Q1 sales?
  - ❑ Why have there been more support calls originating from the Eastern region than from the Western region?
  - ❑ Why was there an increase in patient re-admission rates over the past three months?
- Provide more value than descriptive analytics but require a more advanced skillset.
- Show trends and patterns—**interactive visualization tools**

Data analytics framework



Data Analytics Framework

Collection
- Streaming Data
  - Event Data
  - Time Series Data
- Historical Data

Cleaning
- Identify | remove quality issues
- Label | Structure
- Add context

Integration
- Align data
  - Exisiting data sets
  - Common vocabulary

Analysis
- Descriptive
- Predictive & Prescriptive
  - Machine Learning
  - Natural Lang. Proc.
  - Image Processing
  - Computer Vision

Visualization
- Histograms | Bar Charts
- Scatter Plots
- Heat Maps
- Network Analysis

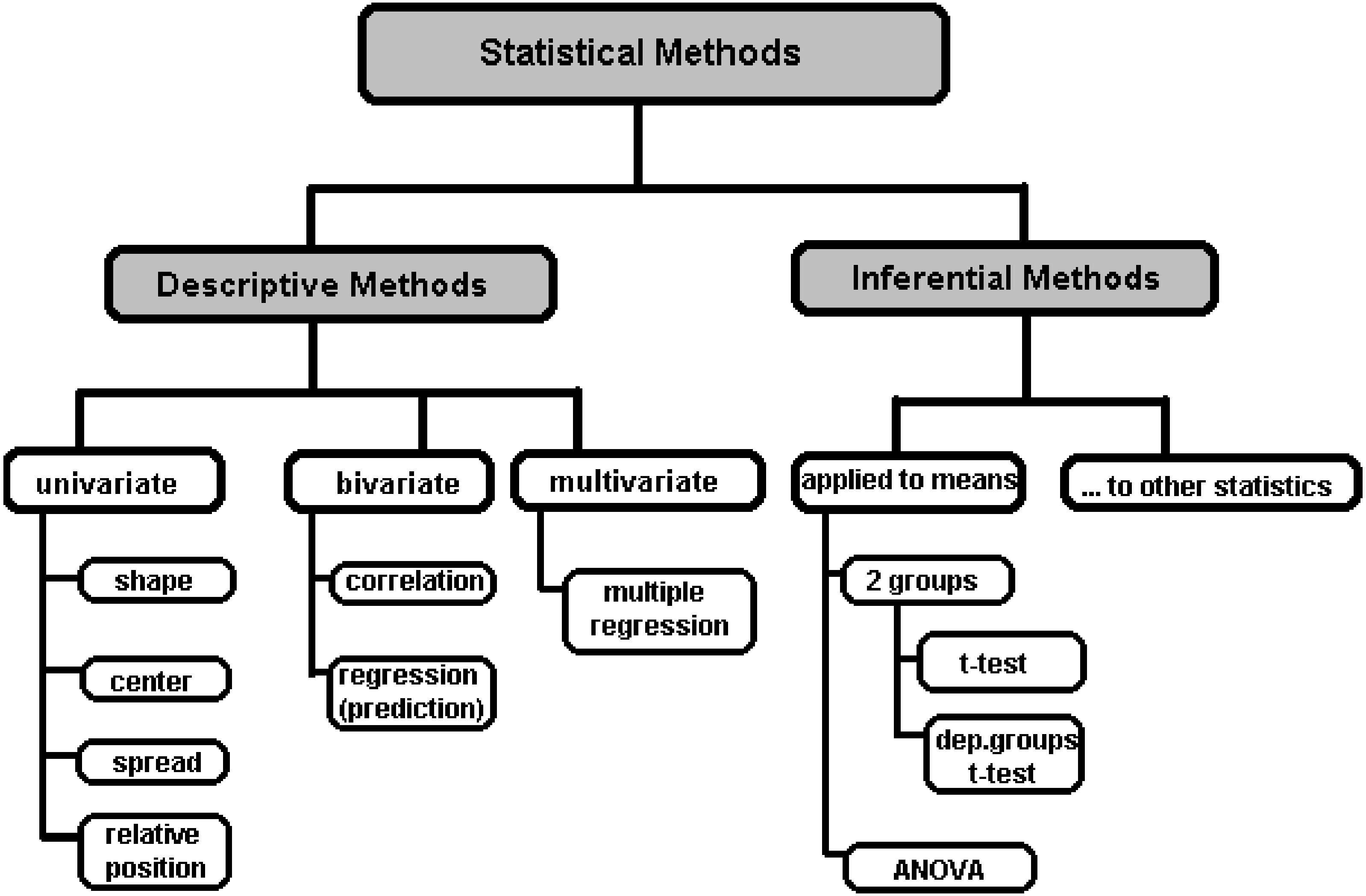
Alerting
- Custom | Dashboard alerts
  - Spike in Traffic
  - Goal completion / miss
- Email notification

Introduction to Statistical Analysis

Basics of Statistics

- Definition: Science of collection, presentation, analysis, and reasonable interpretation of data.
- Statistics presents a rigorous scientific method for gaining insight into data. For example, suppose we measure the weight of 100 patients in a study. With so many measurements, simply looking at the data fails to provide an informative account. However, statistics can give an instant overall picture of data based on graphical presentation or numerical summarization irrespective to the number of data points. Besides data summarization, another important task of statistics is to make inference and predict relations of variables.

A Taxonomy of Statistics



Statistical Description of Data

- Statistics describes a numeric set of data by its
  - Center
  - Variability
  - Shape
- Statistics describes a categorical set of data by
  - Frequency
  - percentage or proportion of each category

Some Definitions

Variable - any characteristic of an individual or entity. For several people, a variable can have distinct values. Variables might be quantitative or categorical. Per S. S. Stevens.

- **Nominal** - Categorical variables, such as names or classes, without a natural hierarchy or ranking (e.g., gender). Value may have a numerical component, although not necessarily (e.g., I, II, III). Nominal variables can only be subjected to enumeration as an operation.
- **Ordinal** - Variables having a natural hierarchy, such as mild, moderate, and severe. It is possible to compare things for equality, more or less value, but not by how much.
- **Interval** - Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely anchored. Calendar dates and temperatures on the Fahrenheit scale are examples. Addition and subtraction, but not multiplication and division are meaningful operations.
- **Ratio** - Variables with all properties of Interval plus an absolute, non-arbitrary zero point, e.g. age, weight, temperature. Addition, subtraction, multiplication, and division are all meaningful operations.

Some Definitions

Distribution - (of a variable) tells us what values the variable takes and how often it takes these values.

- **Unimodal** - having a single peak (only a single highest value)
- **Bimodal** - having two distinct peaks
- **Symmetric** - left and right half are mirror images.

Frequency Distribution

Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable ‘age’ can be tabulated as follows:

Frequency Distribution of Age

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

Grouped Frequency Distribution of Age:

Age Group	1-2	3-4	5-6
Frequency	8	12	6

Cumulative Frequency

Cumulative frequency of data in previous page

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2
Cumulative Frequency	5	8	15	20	24	26

Age Group	1-2	3-4	5-6
Frequency	8	12	6
Cumulative Frequency	8	20	26

Data Presentation

Two types of statistical presentation of data - graphical and numerical.

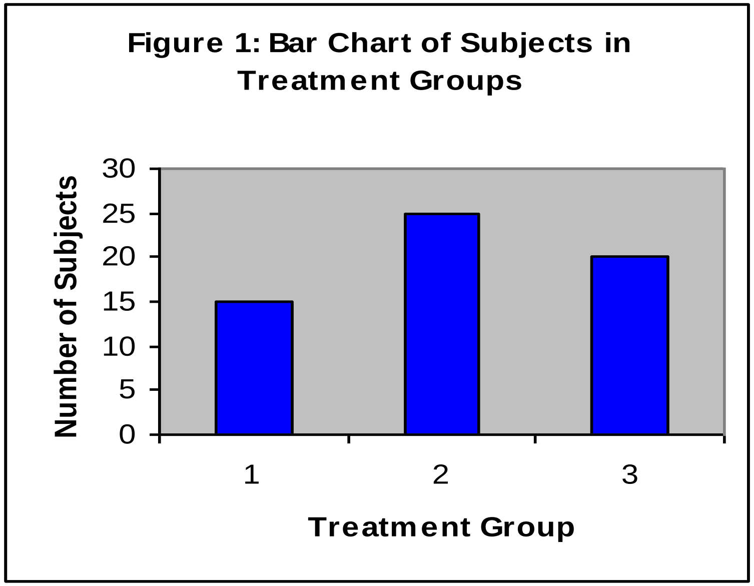
Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Overall pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an ***outlier***.

Bar diagram and Pie charts are used for categorical variables.

Histogram, stem and leaf and Box-plot are used for numerical variable.

Data Presentation –Categorical Variable

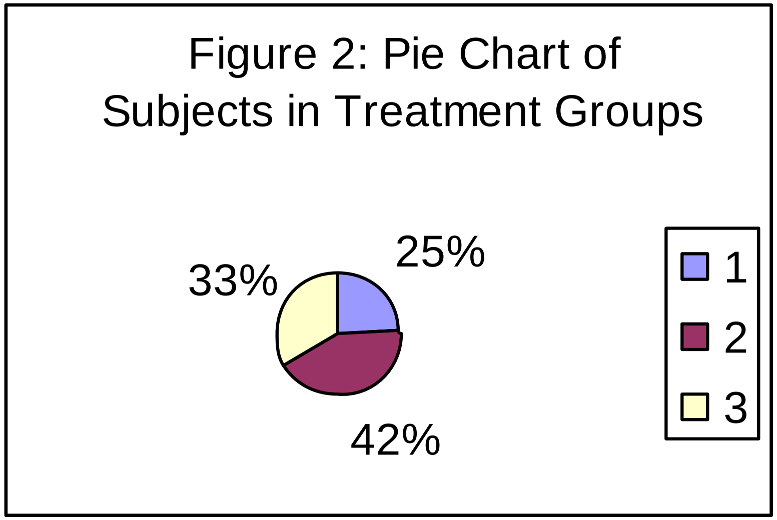
Bar Diagram: Lists the categories and presents the percent or count of individuals who fall in each category.



Treatment Group	Frequency	Proportion	Percent (%)
1	15	(15/60)=0.25	25.0
2	25	(25/60)=0.333	41.7
3	20	(20/60)=0.417	33.3
Total	60	1.00	100

Data Presentation –Categorical Variable

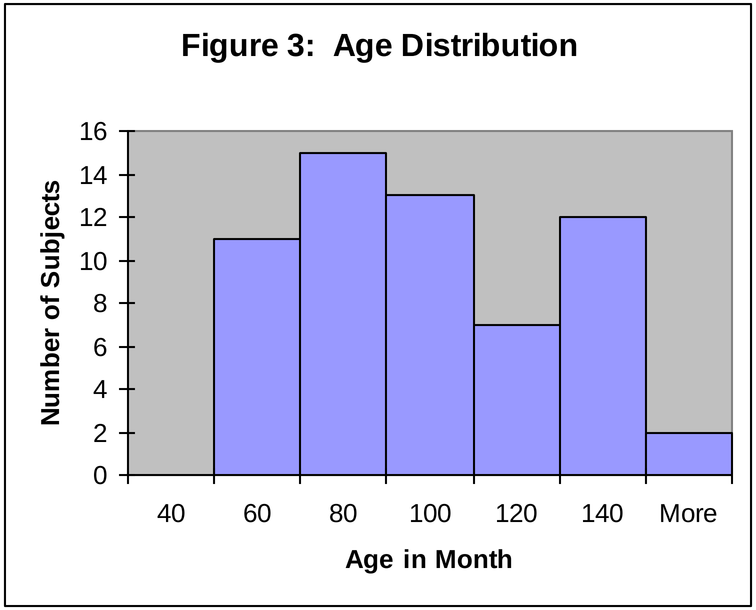
Pie Chart: Lists the categories and presents the percent or count of individuals who fall in each category.



Treatment Group	Frequency	Proportion	Percent (%)
1	15	(15/60)=0.25	25.0
2	25	(25/60)=0.333	41.7
3	20	(20/60)=0.417	33.3
Total	60	1.00	100

Graphical Presentation –Numerical Variable

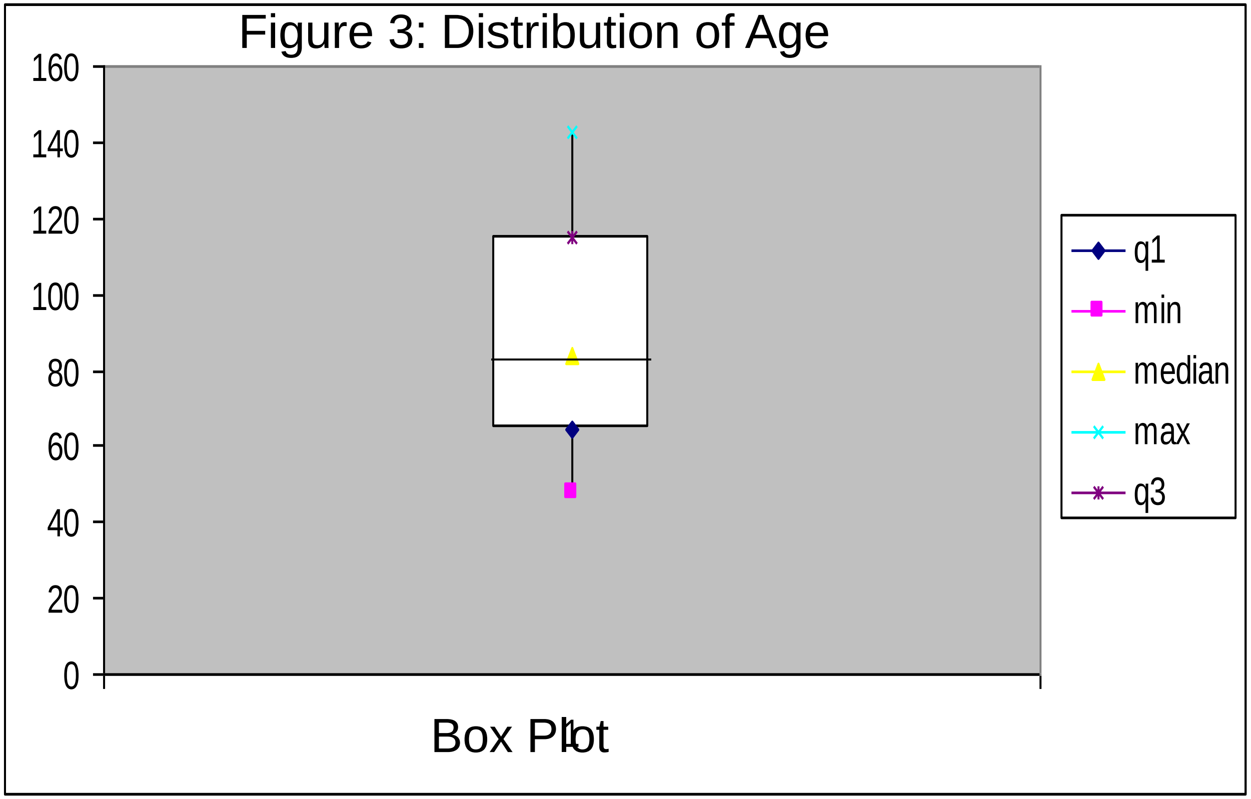
Histogram: Overall pattern can be described by its shape, center, and spread. The following age distribution is right skewed. The center lies between 80 to 100. No outliers.



Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60

Graphical Presentation –Numerical Variable

Box-Plot: Describes the five-number summary



Types of data

- **Univariate/Multivariate**

  •**Univariate**: Use one variable to describe a person, place, or thing.

  •**Multivariate**: Use two or more variables to measure a person, place or thing. Variables may or may not be dependent on each other.

- **Cross-sectional data/Time-ordered data (business, social sciences)**

  •**Cross-Sectional**: Measurements taken at one time period

  •**Time-Ordered:** Measurements taken over time in chronological sequence**.**

The type of data will dictate (in part) the appropriate data-analysis method.

Measurement Scales

- **Nominal or Categorical Scale**
  - Classification of people, places, or things into categories (e.g. age ranges, colors, etc.).
  - Classifications must be mutually exclusive (every element should belong to one category with no ambiguity).
  - Weakest of the four scales. No category is greater than or less (better or worse) than the others. They are just different.

- **Ordinal or Ranking Scale**
  - **Classification of people, places, or things into a ranking such that the data is arranged into a meaningful order (e.g. poor, fair, good, excellent).**
  - **Qualitative classification only**

Univariate Analysis

- Min/Max
- Average
- Median
- Mode
- Variance
- Standard Deviation

Univariate Analysis/Descriptive Statistics

- **The Average (Mean)**
  - Sum of all values divided by the number of values in the data set.
  - One measure of central location in the data set.

Average = $$\frac{1}{N}\sum_{i=1}^{N} m_i$$

Average=(73+66+69+67+49+60+81+71+78+62+53+87+74+65+74+50+85+45+63+100)/20 = 68.6

Univariate Analysis/Descriptive Statistics

- **The Median**
  - The middle value in a sorted data set. Half the values are greater and half are less than the median.
  - Another **measure of central location in the data set**.

(45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)

Median: 68

(1, 2, 4, 7, 8, 9, 9)

- Excel function: MEDIAN()

Univariate Analysis/Descriptive Statistics

- **The Mode**
  - Most frequently occurring value.
  - Another measure of central location in the data set.
  - (45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)
  - Mode: 74

  - Generally not all that meaningful unless a larger percentage of the values are the same number.

Univariate Analysis/Descriptive Statistics

- **Variance**
  - **One measure of dispersion** (deviation from the mean) of a data set. The larger the variance, the greater is the average deviation of each datum from the average value.

Variance = $$\frac{1}{N}\sum_{i=1}^{N}(m_i - \overline{m})^2$$

$$\overline{m} =$$

Variance = $[(45 – 68.6)^2 + (49 – 68.6)^2 + (50 – 68.6)^2 + (53 – 68.6)^2 + …]/20 = 181$

Univariate Analysis/Descriptive Statistics

- **Standard Deviation**
  - **how much scores deviate from the mean .Most commonly used measure of spread**
  - Square root of the variance. Can be thought of as the average deviation from the mean of a data set.
  - The magnitude of the number is more in line with the values in the data set.

Standard Deviation = $([(45 - 68.6)^2 + (49 - 68.6)^2 + (50 - 68.6)^2 + (53 - 68.6)^2 + \ldots]/20)^{1/2} = 13.5$

Multivariate Analysis/Descriptive Statistics

- T-Test
- Z-Test
- ANOVA(Analysis of Variance)
- Chi Square

T-Test

- The t-test assesses whether the means of two groups are *statistically* different from each other.

- This analysis is appropriate whenever you want to compare the means of two groups,

Hypotheses

There are two kinds of hypotheses for a one sample *t*-test.

- Null hypothesis
- Alternative hypothesis.

Hypotheses

- Null hypothesis assumes

It assumes that no difference exists.

- The alternative hypothesis :

It assumes that some difference exists between the true mean (µ) and the comparison value (m0),

The purpose of the one sample *t*-test is to determine if the null hypothesis should be rejected

Procedure (T- Test)

- Symbols used :
- Y = Random sample
- yi = The ith observation in YY
- n = The sample size
- m0 = The hypothesized value
- y¯ = The sample mean
- σ^ = The sample standard deviation
- T =The critical value of a *t*-distribution with (n − 1n − 1) degrees of freedom
- t = The *t*-statistic (*t*-test statistic) for a one sample *t*-test
- p = The pp-value (probability value) for the *t*-statistic.

Procedure (T- Test)

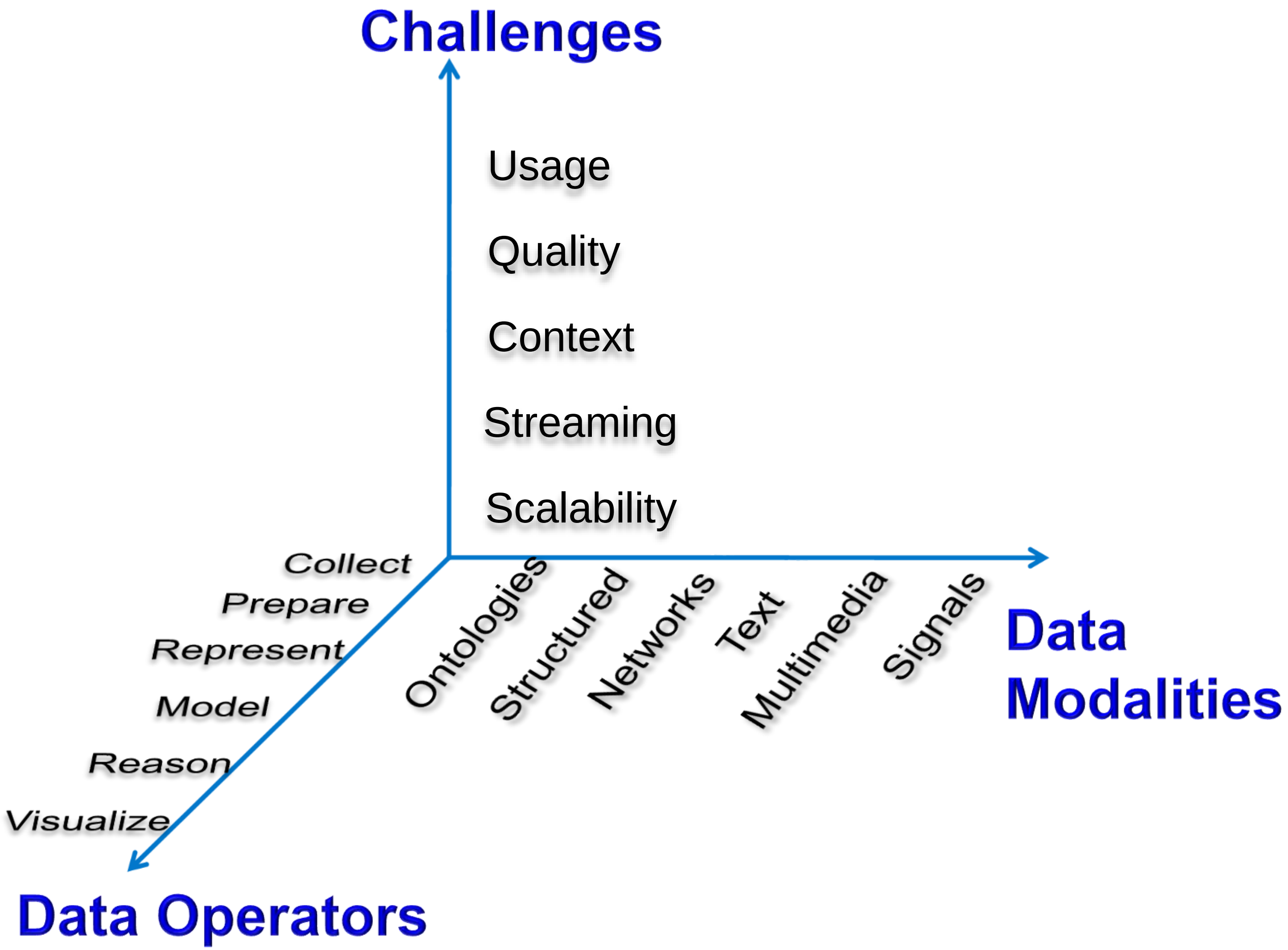
- 1. Calculate the sample mean.

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

- 2. Calculate the sample standard deviation.

$$\hat{\sigma} = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}}$$

What matters when dealing with data?



The End