

# CSE 2027-Fundamental of Data Analysis

## Module 1-Introduction to Data Analysis

Introducing Data, overview of data analysis: Data in the Real World, Data vs. Information, Many “Vs” of Data, Structured Data and Unstructured Data, Types of Data, Data Analysis Defined, Types of Variables, Central Tendency of Data, Scales of Data, Sources of Data, Data preparation: Cleaning the data, Removing variables, Data Transformations.



# Introducing Data

- Facts and statistics collected together for reference or analysis
- Data has to be transformed into a **form** that is efficient for movement or processing.



# Over view of Data Analysis



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



- **Data analysis** is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making.
- The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.
- A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision.



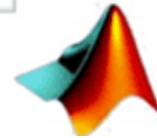
- This is nothing but analyzing our past or future and making decisions based on it.
- For that, we gather memories of our past or dreams of our future.
- So that is nothing but data analysis. Now same thing analyst does for business purposes, is called Data Analysis.



# Data Analysis Tools



SQL



MATLAB



**PRESIDENCY  
UNIVERSITY**  
Private University Estd. in Karnataka State by Act No. 41 of 2013



# Data in the Real World



## Clinical

Demographics, EHR Data, Lab Test Results, Diagnoses, Procedures, Pathology/ Histology Data, Radiology Images, Microbiology Data, Provider Notes, Admission/ Discharge and Progress Reports, Performance Status



## Medication

Medication Orders, Administration (Dose, Route, NDC/RxNorm codes), Concomitant Therapies, Point of Sale Data, (Prescription & OTC) Prescription Refill, Allergies



## Claims

Medical Claims, Prescription Drug Claims, Other Drug and Treatment Use Data



## Molecular Profiling

Genomic and Genetic Testing Data (SNPs/Panels), Multi-Omics Data (Proteomics, Transcriptomics, Metabonomics, Lipidomics), Other Biomarker Status



## Family History

Historical Data on Health Conditions and Allergies Relating to Patient and Extended Family, Smoking Status, Alcohol Use



## Mobile Health

Fitness Trackers, Wearable Devices, Other Health Apps Measuring Activity and Body Function



## Environmental

Climate Factors, Pollutants, Infections, Lifestyle Factors (diets, stress), Other Environmental and Occupational Sources



## Patient Reported

Patient Reported Outcomes, Surveys, Diaries (diets, habits), Personal Health Records, Adverse Event Reporting, Quality of Life Measures



## Social Media

Patient Communities, Twitter, Facebook, Blogs



## Literature

Disease Burden, Clinical Characteristics, Prevalence/Incidence, Rates of Treatment, Resource Use and Costs, Disease Control, Quality of Life Measures



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# DATA VS. INFORMATION

## Examples of Data



120/80 blood pressure reading

Date on an employee application

Number of cesarean sections  
in May

Vendor address



## Examples of Information

John Doe's blood pressure  
reading on 9/15/15

Employee application record

ABC Hospital cesarean  
section rate for May

Vendor record

## Example Data Governance Functions



- » Data Quality Control and Management
- » Identity Management
- » Data Cleansing
- » Metadata Management
- » Master Data Management



## Example Information Governance Functions

- » Enterprise-wide Governance Policies
  - » Life Cycle Management
- » Information Use, Exchange, and Preservation
  - » Physical and Electronic Systems Governance
    - » Privacy and Security
  - » Information Risk Management
- » Legal and Regulatory Response
- » Information Vendor Services Management



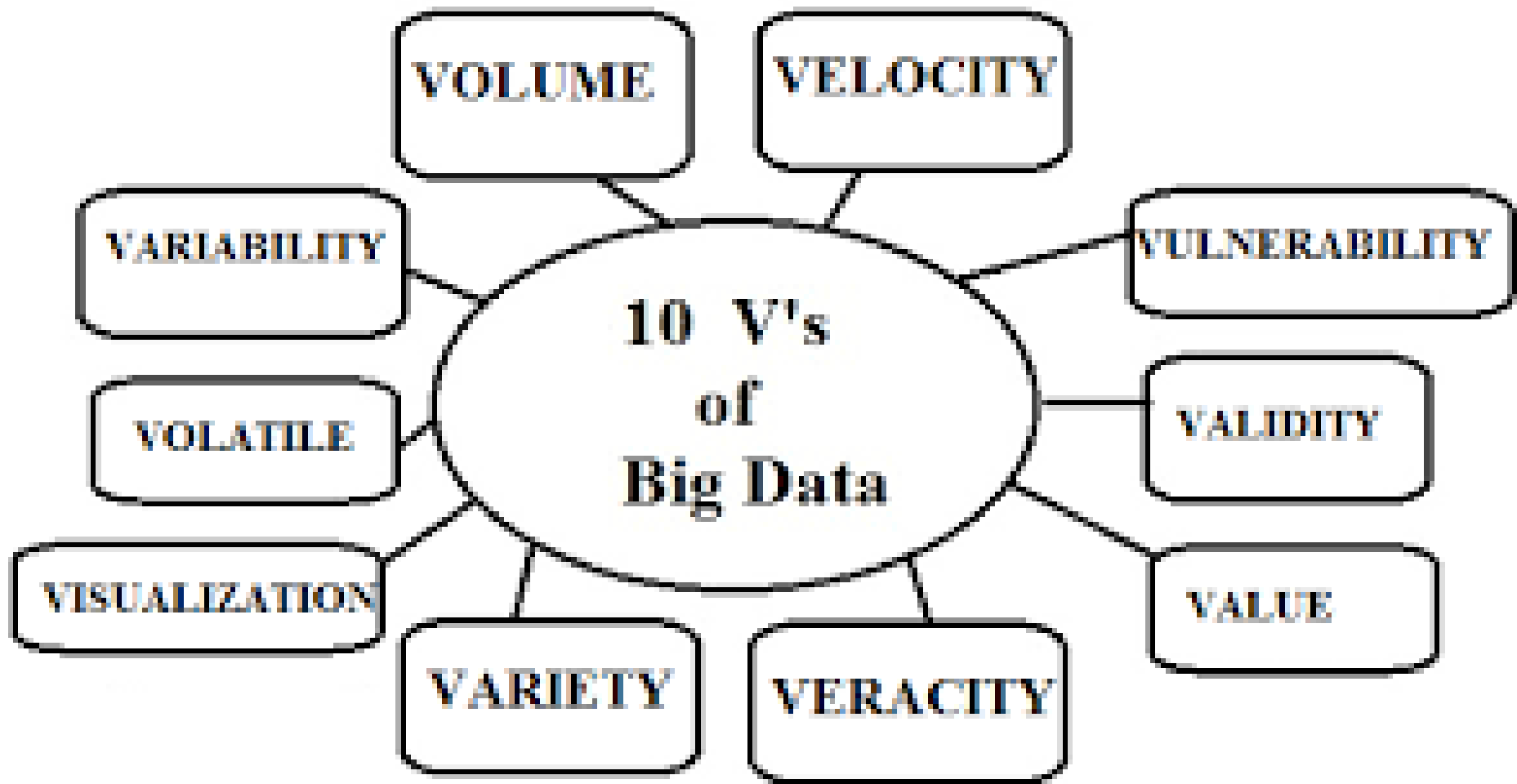
**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013





# Many v's of Data



# A. Volume

- The term **Volume** is meant for the **Magnitude or Scale of data**.
- Massive amounts of data generated from multiple resources are not possible to handle through the traditional ways like a database.
- This large volume data is a composition of multiple data types, which is unstructured in nature.
- This kind of data can be either in the form of audio, video, tweets, likes etc.

# B. Velocity

- Velocity refers to the **speed at which the gigantic amount of data** is being generated, collected and scrutinized.
- With every flip of second, data is being searched on the internet.
- On a day to day basis, social networking sites like Facebook, Twitter, LinkedIn etc, are sharing a large amount of data.
- For easy analysis of this high amount of constantly generating data with keeping an eye on it speed and easy access.



# C. Variety

- In terms of Big data, term Variety of data pretends to be a composition of structured and unstructured kind of data.
- The data collected from different sources like mobile phones, laptops etc is not homogenous in nature.
- Apart from text, audio ,video files, there may be some log files ,clicks or likes or dislikes etc.



# D. Value

- Value refers to convert **our investigated data into values.**
- Value is one of the most important characteristics of Big data with a composition of collection and analyzing the same **in order to boost the performance of any organization along with a better understanding of customers.**
- With the access to this useful data, one must analyze great values in order to get amazing benefits.



# E. Variability

- Variability refers to **unpredictable changes in the data.**
- It may happen because of multiple data types & the speed with which data is generating and being loaded into the database.



# F. Veracity

- Veracity refers to the **term trustworthiness with reference to accurate data.**
- If the data is accurate, only then you could think of meaningful data.
- For example, consider a dataset of thirty students on which we have to make an analysis about the reason they got distinction.
- Being an analyzer, you can ask questions like:
- what are the methodology you adopted to get good marks in all the subjects?



- How much time you devote to individual subject?
- Do you learn some subjects with the help of daily life activities like sports etc?
- Have you ever been a scholar?
- Be getting answers like this it would be easier to determine the **accuracy of information** which could easily be maintained in statistical form.





# G. Validity

- Two terms of big data veracity and validity seems to be alike but are quite different.
- validity is meant for an accurate analysis in order to get optimized results.



# H. Vulnerability

- **Vulnerability** is one of the major challenge in big data as the data generated from multiple sources with **such an erratic speed has high chances of being harmed by any intruder.**
- Currently, in a case of facebook, where the Belgium court has threatened to fine a high amount on breaking privacy recently.



# I. Volatility

- Volatility refers to how **long the perceived data remains** to be useful for us and how it is to be kept.
- For analyzing the same, it is necessary to develop some new rules and techniques through which rapid access to information is possible.



# J. Visualization

- Data Visualization is one of the most complex challenge in big data.
- In this information age, **data is not only going beyond the limits** but also is composed of different data types.
- So, there is a need of communicate the information by visualizing it through some special ways with special functionalities like a web-based approach, statistical analysis etc.

- Traditional tools of data visualization face severe challenges like low response time, complex methods of scalability, precision in reporting time etc.
- So, it is a challenge to work with the concept which way of communication with data is most suitable in order to make visualization more effective.

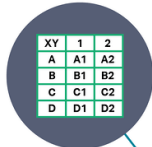


# Structured Data

vs

# Unstructured Data

Can be displayed  
in rows, columns and  
relational databases



XY	1	2
A	A1	A2
B	B1	B2
C	C1	C2
D	D1	D2

Numbers, dates  
and strings



Estimated 20% of  
enterprise data (Gartner)



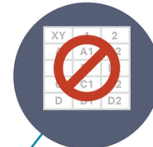
Requires less storage



Easier to manage  
and protect with  
legacy solutions



Cannot be displayed  
in rows, columns and  
relational databases



Images, audio, video,  
word processing files,  
e-mails, spreadsheets



Estimated 80% of  
enterprise data (Gartner)



Requires more storage



More difficult to  
manage and protect  
with legacy solutions



# Typical human-generated unstructured data includes

- **Text files:** Word processing, spreadsheets, presentations, email, logs.
- **Email:** Email has some internal structure thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.
- **Social Media:** Data from Facebook, Twitter, LinkedIn.
- **Website:** YouTube, Instagram, photo sharing sites.
- **Mobile data:** Text messages, locations.
- **Communications:** Chat, IM, phone recordings, collaboration software.
- **Media:** MP3, digital photos, audio and video files.
- **Business applications:** MS Office documents, productivity applications



# Typical machine-generated unstructured data includes:

- **Satellite imagery:** Weather data, land forms, military movements.
- **Scientific data:** Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
- **Digital surveillance:** Surveillance photos and video.
- **Sensor data:** Traffic, weather, oceanographic sensors.



80%  
Unstructured

Vs

20%  
Structured



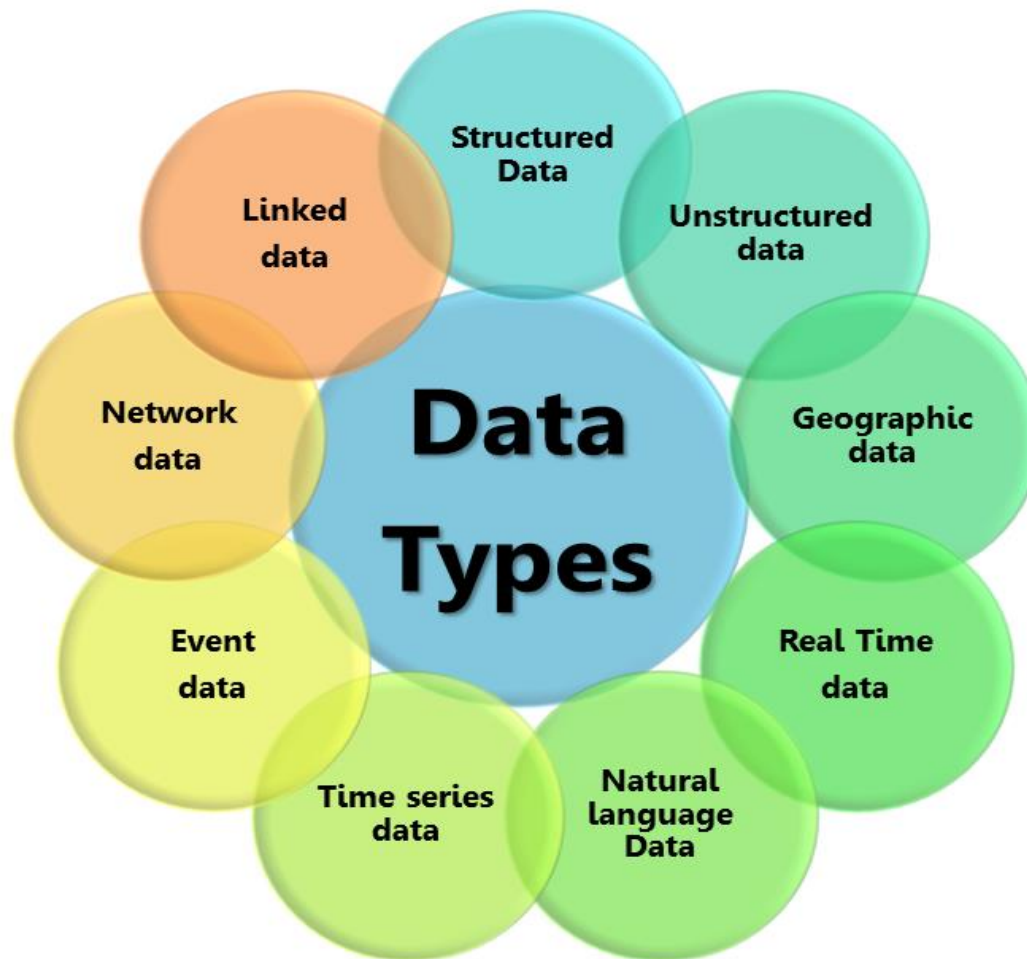
Database



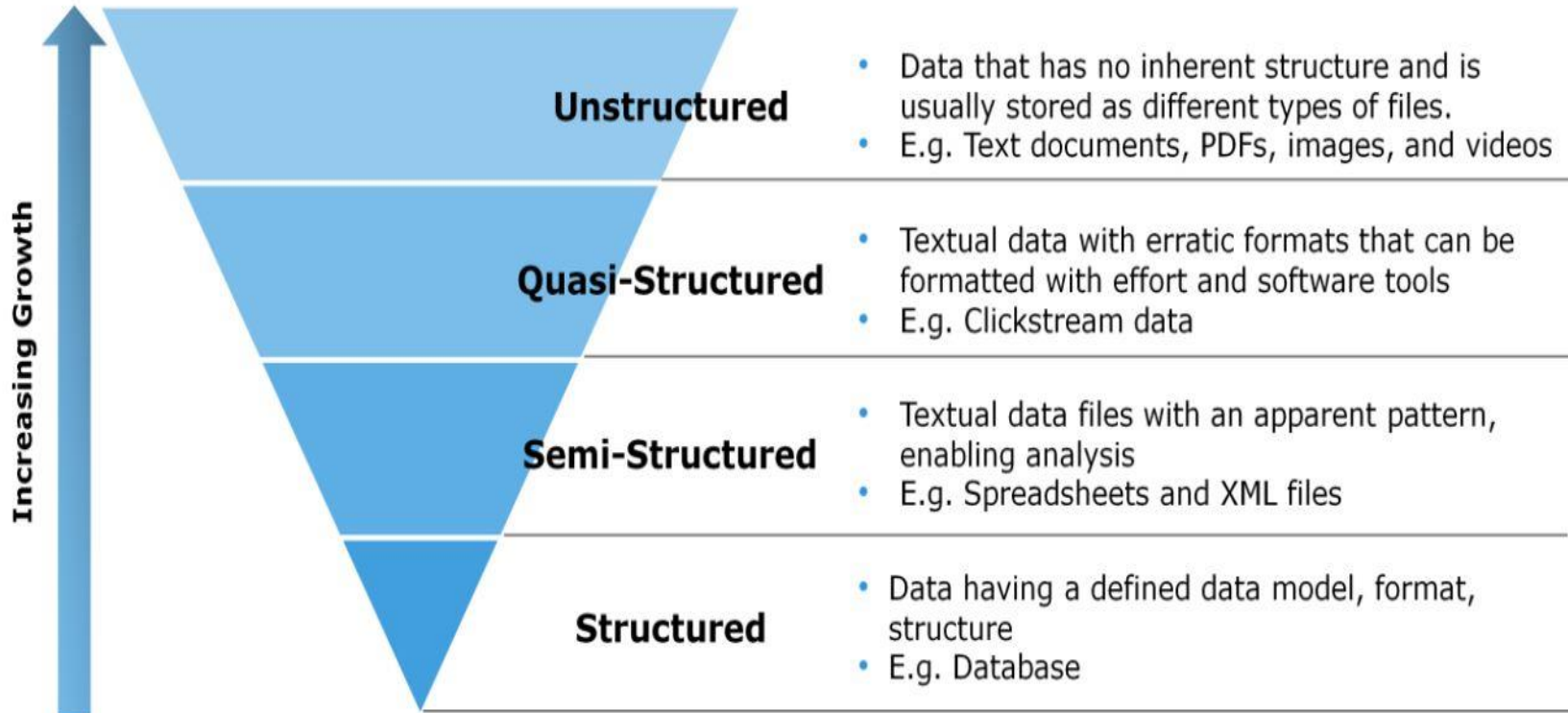
Tables



	Structured Data	Unstructured Data
<b>Characteristics</b>	<ul style="list-style-type: none"> <li>• Pre-defined data models</li> <li>• Usually text only</li> <li>• Easy to search</li> </ul>	<ul style="list-style-type: none"> <li>• No pre-defined data model</li> <li>• May be text, images, sound, video or other formats</li> <li>• Difficult to search</li> </ul>
<b>Resides in</b>	<ul style="list-style-type: none"> <li>• Relational databases</li> <li>• Data warehouses</li> </ul>	<ul style="list-style-type: none"> <li>• Applications</li> <li>• NoSQL databases</li> <li>• Data warehouses</li> <li>• Data lakes</li> </ul>
<b>Generated by</b>	Humans or machines	Humans or machines
<b>Typical applications</b>	<ul style="list-style-type: none"> <li>• Airline reservation systems</li> <li>• Inventory control</li> <li>• CRM systems</li> <li>• ERP systems</li> </ul>	<ul style="list-style-type: none"> <li>• Word processing</li> <li>• Presentation software</li> <li>• Email clients</li> <li>• Tools for viewing or editing media</li> </ul>
<b>Examples</b>	<ul style="list-style-type: none"> <li>• Dates</li> <li>• Phone numbers</li> <li>• Social security numbers</li> <li>• Credit card numbers</li> <li>• Customer names</li> <li>• Addresses</li> <li>• Product names and numbers</li> <li>• Transaction information</li> </ul>	<ul style="list-style-type: none"> <li>• Text files</li> <li>• Reports</li> <li>• Email messages</li> <li>• Audio files</li> <li>• Video files</li> <li>• Images</li> <li>• Surveillance imagery</li> </ul>



# Types of Digital Data



# Data Analysis-Types

- There are several **types of Data Analysis** techniques that exist based on business and technology. However, the major Data Analysis methods are:
- Text Analysis
- Statistical Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis



# THE FOUR MAIN TYPES OF DATA ANALYSIS

## **Descriptive**

What happened?

## **Diagnostic**

Why did it happen?

## **Predictive**

What is likely to happen in the future?

## **Prescriptive**

What's the best course of action?





# Text Analysis

- Text Analysis is also referred to as Data Mining. It is one of the methods of data analysis to discover **a pattern in large data sets using databases or data mining tools.**
- It used to transform raw data into business information. Business Intelligence tools are present in the market which is used to take strategic business decisions. Overall it offers a way to extract and examine data and deriving patterns and finally interpretation of the data.



# Statistical Analysis

- Statistical Analysis shows "What happen?" by using past data in the form of dashboards. Statistical Analysis includes collection, Analysis, interpretation, presentation, and modeling of data. It analyses a set of data or a sample of data.
- There are two categories of this type of Analysis -  
Descriptive Analysis  
Inferential Analysis.



# Descriptive Analysis

- Analyses complete data or a sample of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data.

# Inferential Analysis

- Analyses sample from complete data. In this type of Analysis, you can find different conclusions from the same data by selecting different samples.

# Diagnostic Analysis

- Diagnostic Analysis shows "Why did it happen?" by finding the cause from the insight found in Statistical Analysis. This Analysis is useful to identify behavior patterns of data. If a new problem arrives in your business process, then you can look into this Analysis to find similar patterns of that problem. And it may have chances to use similar prescriptions for the new problems.

# Predictive Analysis

- Predictive Analysis shows "what is likely to happen" by using previous data. The simplest data analysis example is like if last year I bought two dresses based on my savings and if this year my salary is increasing double then I can buy four dresses. But of course it's not easy like this because you have to think about other circumstances like chances of prices of clothes is increased this year or maybe instead of dresses you want to buy a new bike, or you need to buy a house!
- So here, this Analysis makes predictions about future outcomes based on current or past data. Forecasting is just an estimate. Its accuracy is based on how much detailed information you have and how much you dig in it.

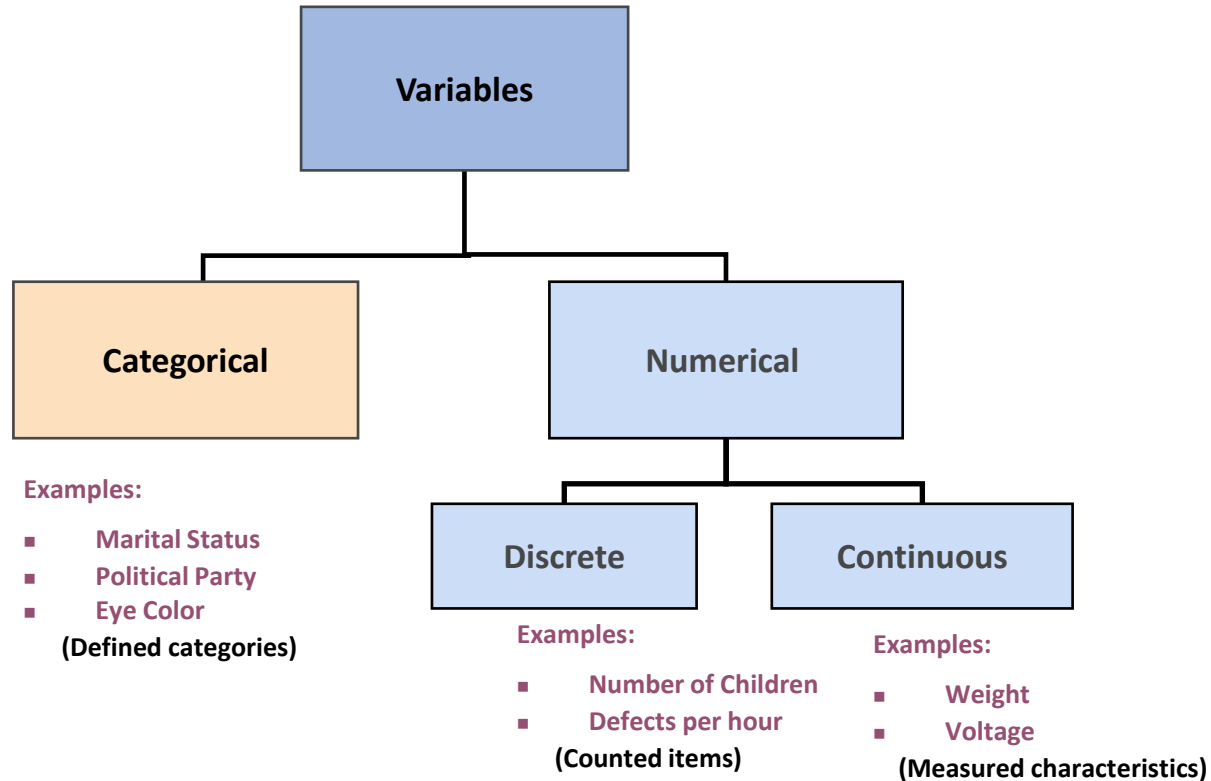
# Prescriptive Analysis

- Prescriptive Analysis combines the insight from all previous Analysis to determine which action to take in a current problem or decision. Most data-driven companies are utilizing Prescriptive Analysis because predictive and descriptive Analysis are not enough to improve data performance. Based on current situations and problems, they analyze the data and make decisions.

# Types of Variable

- **Categorical** (*qualitative*) variables have values that can only be placed **into categories**, such as “yes” and “no.”
- **Numerical** (*quantitative*) variables have values that represent quantities.
  - **Discrete** variables arise from a *counting process*
  - **Continuous** variables arise from a *measuring process*

# Types of Variables



# Central Tendency-Mode

- The mode is the most commonly reported value for a particular variable.
- It is illustrated using the following variable whose values are: 3, 4, 5, 6, 7, 7, 7, 8, 8, 9
- The mode would be the value 7 since there are three occurrences of 7 (more than any other value).
- The following values, both 7 and 8 are reported three times: 3, 4, 5, 6, 7, 7, 7, 8, 8, 8, 9 The mode may be reported as {7, 8} or 7.5.

# Median

- The median is the middle value of a variable once it has been sorted from low to high. For variables with an even number of values, the mean of the two values closest to the middle is selected (sum the two values and divide by 2).
- The following set of values will be used to illustrate: 3, 4, 7, 2, 3, 7, 4, 2, 4, 7, 4.
- Before identifying the median, the values must be sorted: 2, 2, 3, 3, 4, 4, 4, 4, 7, 7, 7



# Mean

Mean

$$\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$$

Problem: 3, 4, 5, 7, 7, 8, 9, 9, 9

Soln

$$\begin{aligned}\bar{x} &= \frac{3 + 4 + 5 + 7 + 7 + 8 + 9 + 9 + 9}{9} \\ &= 61/9 = 6.78\end{aligned}$$



# Source of Data

- **Surveys or polls**

A survey or poll can be useful for gathering data to answer specific questions

- **Experiments:**

Experiments measure and collect data to answer a specific question in a highly controlled manner. The data collected should be reliably measured, that is, repeating the measurement should not result in different values. Experiments attempt to understand cause and affect phenomena by controlling other factors that may be important.

- **Observational and other studies:** In certain situations it is impossible on either logistical or ethical grounds to conduct a controlled experiment. In these situations, a large number of observations are measured and care taken when interpreting the results.
- **Operational databases:** These databases contain ongoing business transactions. They are accessed constantly and updated regularly. Examples include supply chain management systems, customer relationship management (CRM) databases and manufacturing production databases.

- **Data warehouses:** A data warehouse is a copy of data gathered from other sources within an organization that has been cleaned, normalized, and optimized for making decisions. It is not updated as frequently as operational databases.
- **Historical databases:** Databases are often used to house historical polls, surveys and experiments.
- **Purchased data:** In many cases data from in-house sources may not be sufficient to answer the questions now being asked of it. One approach is to combine this internal data with data from other sources.

# Scales of Data

- **Nominal:** Scale describing a variable with a limited number of different values. This scale is made up of the list of possible values that the variable may take. It is not possible to determine whether one value is larger than another.
- **Ordinal:** This scale describes a variable whose values are ordered; however, the difference between the values does not describe the magnitude of the actual difference.

- **Interval:** Scales that describe values where the interval between the values has meaning.
- **Ratio:** Scales that describe variables where the same difference between values has the same meaning (as in interval) but where a double, tripling, etc. of the values implies a double, tripling, etc. of the measurement.



# Table

	Meaningful order	Meaningful difference	Natural zero
<b>Nominal</b>	No	No	No
<b>Ordinal</b>	Yes	No	No
<b>Interval</b>	Yes	Yes	No
<b>Ratio</b>	Yes	Yes	Yes

# Cleaning the Data

- Since the data available for analysis may not have been originally collected with this project's goal in mind, it is important to spend time cleaning the data.
- It is also beneficial to understand the accuracy with which the data was collected as well as correcting any errors.
- For variables measured on a nominal or ordinal scale (where there are a fixed number of possible values), it is useful to inspect all possible values to uncover mistakes and/or inconsistencies.
- Any assumptions made concerning possible values that the variable can take should be tested.



- For example, a variable Company may include a number of different spellings for the same company such as:
- General Electric Company
- General Elec. Co
- GE
- Gen. Electric Company
- General electric company
- G.E. Company



- These different terms, where they refer to the same company, should be consolidated into one for analysis.
- In addition, subject matter expertise may be needed in cleaning these variables.
- For example, a company name may include one of the divisions of the General Electric Company and for the purpose of this specific project it should be included as the “General Electric Company.”



# Removing Variables

- On the basis of an initial categorization of the variables, it may be possible to remove variables from consideration at this point.
- For example, constants and variables with too many missing data points should be considered for removal.
- Further analysis of the correlations between multiple variables may identify variables that provide no additional information to the analysis and hence could be removed.

# Data Transformation

## Normalization

- Normalization is a process where numeric columns are transformed using a mathematical function to a new range. It is important for two reasons.
- First, analysis of the data should treat all variables equally so that one column does not have more influence over another because the ranges are different.
- For example, when analyzing customer credit card data, the Credit limit value is not given more weightage in the analysis than the Customer's age.
- Second, certain data analysis and data mining methods require the data to be normalized prior to analysis, such as neural networks or k-nearest neighbors

# Problem

Variable

33, 21, 7, 53, 29, 42, 12, 19, 22, 36

$$\text{Value}' = \frac{\text{Initial Value} - \text{original min}}{\text{original max} - \text{original min}} (\text{New max} - \text{New min}) + \text{New min}$$

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)} (\text{new\_max}(A) - \text{new\_min}(A)) + \text{new\_min}(A)$$



# Solution

New max = 1 ; New min = 0

Soln      Value' =  $\frac{33-7}{53-7} (1-0) + 0$   
= 0.565

Thank  
you!



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013

