

Chapter 1

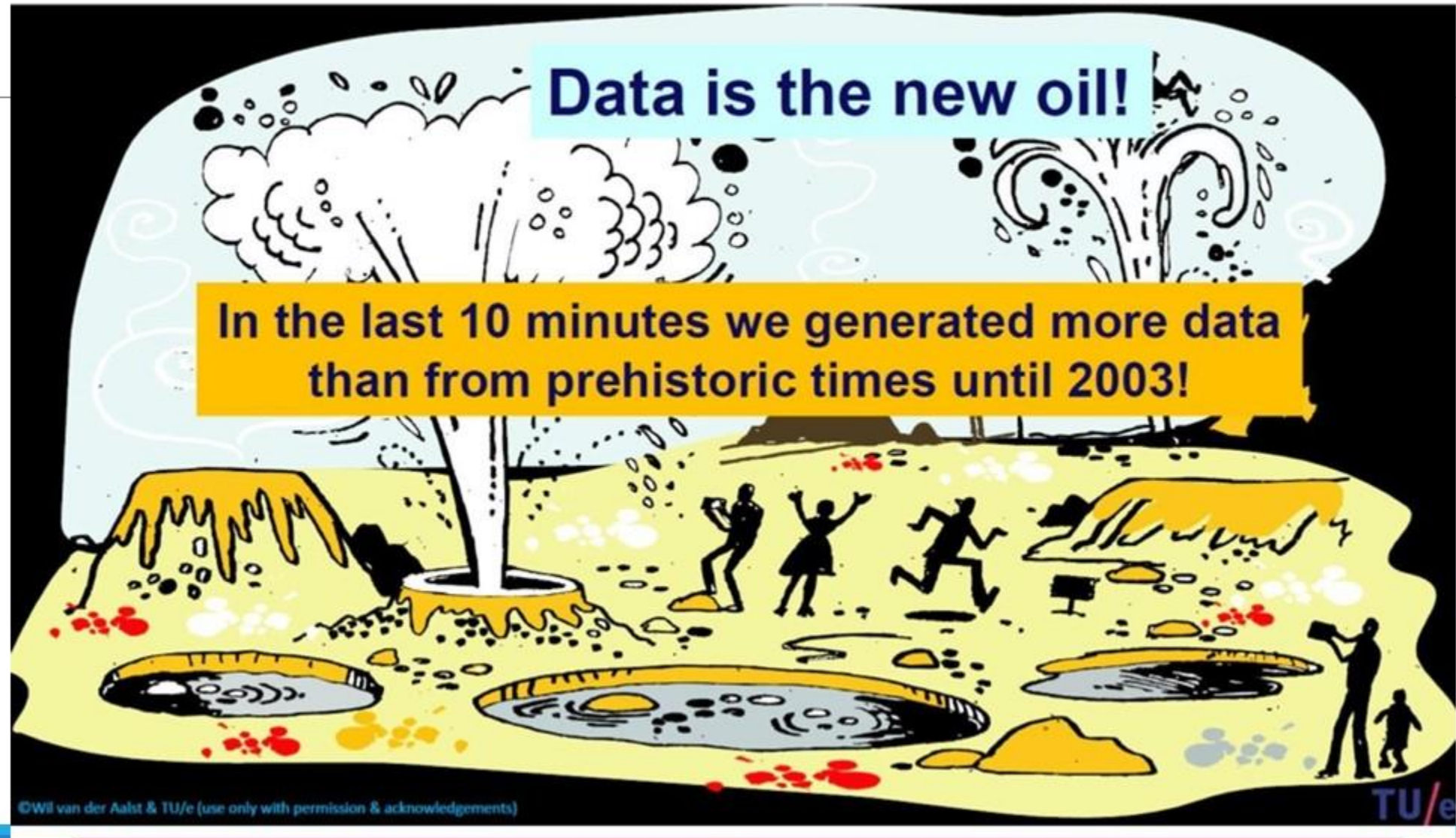
Data Analysis Overview

DATA ANALYSIS IN R

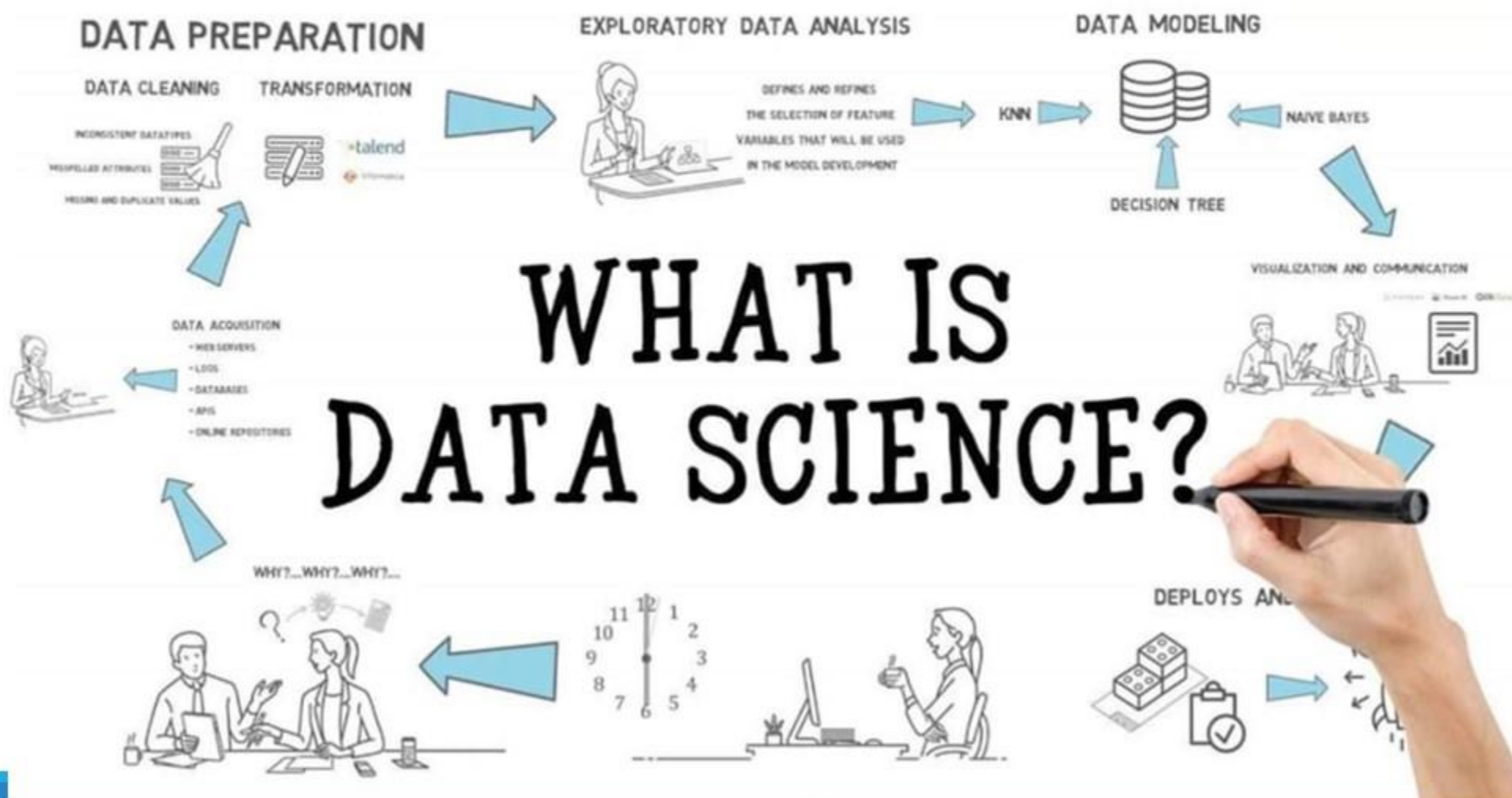
Outline

- Introduction
- Why Data Analysis?
- What is Data Analysis?
- Data Analysis Tool
- Data Analysis: What Kind of Data?
- Types of Data Analysis: Techniques and Methods
- Data Analysis Process

Introduction



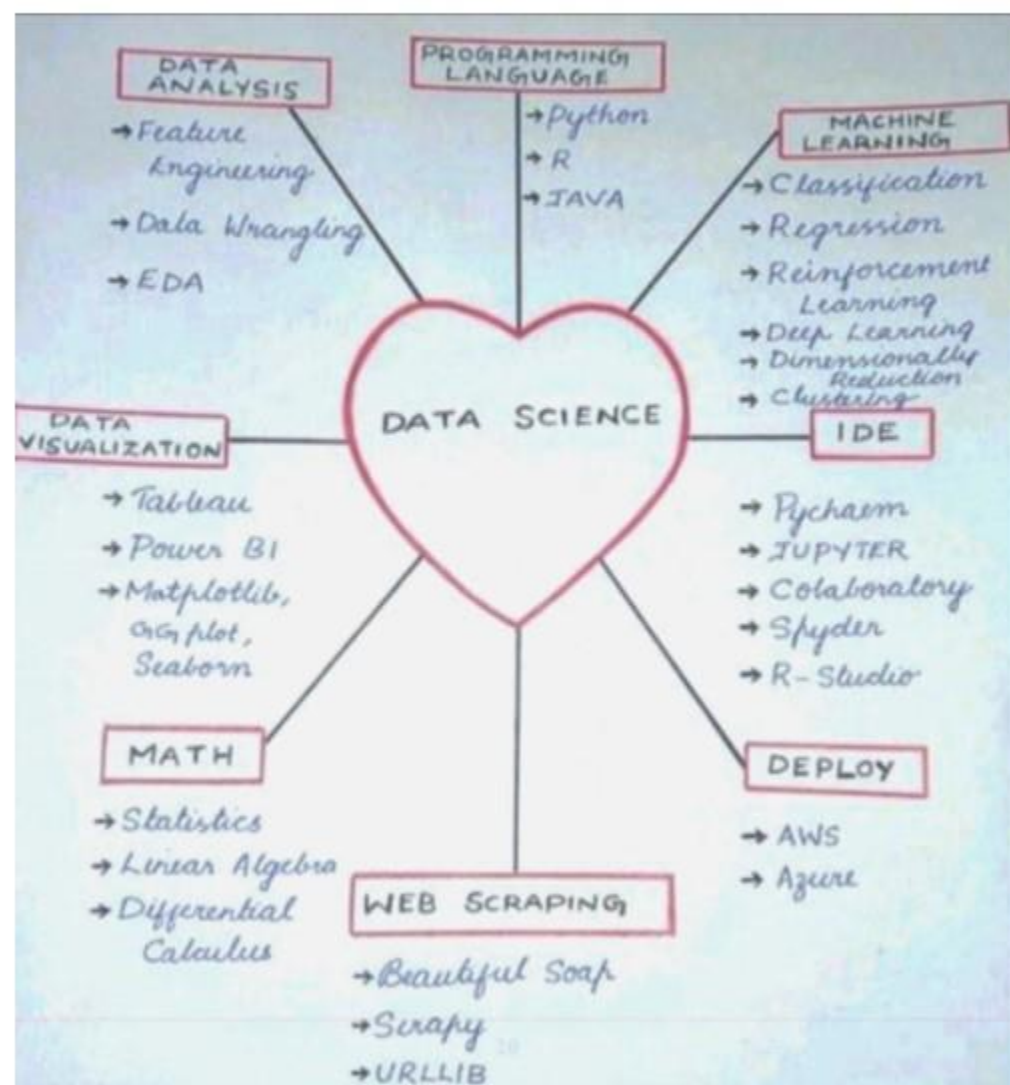
Introduction



Introduction



Introduction



Why Data Analysis?

- **The Explosive Growth of Data: from terabytes to petabytes or even brontobytes**
 - Data collection and data availability
 - Automated data collection tools, database systems, Web, computerized society
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: Remote sensing, bioinformatics, scientific simulation, ...
 - Society and everyone: news, images, video, documents, ...
- **We are drowning in data, but starving for knowledge!**

Data Analysis &... Data Art



From Data to Wisdom



The Information Hierarchy

- **Data**
The raw material of information
- **Information**
Data organized and presented by someone
- **Knowledge**
Information read, heard or seen and understood and integrated
- **Wisdom**
Distilled knowledge and understanding which can lead to decisions

What is Data Analysis?

- **The process of:**
 - Cleaning
 - Transforming
 - Modeling data to discover useful information for business decision-making.
- **Purpose:**

Extract useful information from data and taking the decision based upon the data analysis.

Data Analysis Tool



SOL



sas



MATLAB

Evolution of Sciences

- **Before 1600, empirical science**
- **1600-1950s, theoretical science:**

Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.

- **1950s-1990s, computational science**

-Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology or physics, or linguistics.)

-Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.

- **1990-now, data science**

-The flood of data from new scientific instruments and simulations

-The ability to economically store and manage petabytes of data online

-The Internet and computing Grid that makes all these archives universally accessible

-Scientific info. management, acquisition, organization, query and visualization tasks scale almost linearly with data volumes. Data Analysis and Data Mining is a major new challenge!

Evolution of Data Analysis

Evolutionary Step	Business Question	Enabling Technologies
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive

Data Analysis: On What Kinds of Data?

- **Database-oriented data sets and applications**

- Relational database, data warehouse, transactional database
- Object-relational databases, Heterogeneous databases and legacy databases

- **Advanced data sets and advanced applications**

- Data streams and sensor data
- Time-series data, temporal data, sequence data (incl. bio-sequences)
- Structure data, graphs, social networks and information networks
- Spatial data and spatiotemporal data
- Multimedia database
- Text data and other semi-structured data
- The World-Wide Web

Data Analysis: What Kind of Data?

- **Structured Databases**

- relational object-relational, etc.
- can use SQL to perform parts of the process

e.g., `SELECT count(*) FROM Items WHERE
type=video GROUP BY category`

Customer								Borrow				
customerID	name	address	password	birthdate	family_income	group	...	customerID	date	itemID	#	...
C1234	John Smith	120 main street	Marty	1965/10/10	\$45000	A	...	C1234	99/09/06	98765	1	...
...								...				

Items							
itemID	type	title	media	category	Value	#	...
98765	Video	Titanic	DVD	Drama	\$15.00	2	...
...							

Data Analysis: What Kind of Data?

- **Flat Files**

- most common data source
- can be text (or HTML) or binary
- may contain transactions, statistical data, measurements, etc.

- **Transactional databases**

- Set of records each with a transaction id, time stamp, and a set of items
- may have an associated "description" file for the items
- typical source of data used in market basket analysis

Rentals				
transactionID	date	time	customerID	itemList
T12345	99/09/06	19:38	C1234	{12, 16, 110, 145 ...}
...				

Data Analysis: What Kind of Data?

- **Other Types of Databases**

- legacy databases
- multimedia databases (usually very high-dimensional)
- spatial databases (containing geographical information, such as maps, or satellite imaging data etc.)
- Tune Series Temporal Data (time dependent information such as stock market data; usually very dynamic)

- **World Wide Web**

- basically, a large, heterogeneous, distributed database
- need for new or additional tools and techniques
 - information retrieval, filtering and extraction
 - agents to assist in browsing and filtering
 - Web content usage, and structure (linkage) mining tools
- The "social Web"
 - User generated meta-data, social networks, shared resources, etc.

What Can Data analysis Do

- **Many Data analysis Tasks**

- often inter-related
- often need to try different techniques/algorithms for each task
- each tasks may require different types of knowledge discovery

- **What are some of data analysis tasks**

- Classification
- Prediction
- Clustering
- Affinity Grouping / Association discovery
- Sequence Analysis
- Characterization
- Discrimination

Some Applications of Data Analysis

- **Business data analysis and decision support**

- ▶ Marketing focalization

- Recognizing specific market segments that respond to particular characteristics
 - Return on mailing campaign (target marketing)

- ▶ Customer Profiling

- Segmentation of customer for marketing strategies and/or product offerings
 - Customer behavior understanding
 - Customer retention and loyalty
 - Mass customization / personalization

Some Applications of Data analysis

- **Business data analysis and decision support (cont.)**

- ▶ **Market analysis and management**

- Provide summary information for decision-making
 - Market basket analysis, cross selling, market segmentation.
 - Resource planning

- ▶ **Risk analysis and management**

- "What if" analysis
 - Forecasting
 - Pricing analysis, competitive analysis
 - Time-series analysis (Ex. stock market)

Some Applications of Data analysis

- **Fraud detection**

- ▶ Detecting telephone fraud:

- Telephone call model: destination of the call, duration, time of day or week
 - Analyze patterns that deviate from an expected nonn
 - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud scherne

- ▶ Detection of credit-card fraud

- ▶ Detecting suspicious money transactions (money laundering)

- **Text mining:**

- ▶ Message filtering (e-mail, newsgroups, etc.)

- ▶ Newspaperarticles analysis

- ▶ Text and document categorization

- **Web Mining**

- ▶ Mining patterns from the content, usage, and structure of Web resources

Types of Data Analysis: Techniques and Methods

- **Text Analysis**

- ▶ Discover a pattern in large data sets using databases or data mining tools.
- ▶ Used to transform raw data into business information.
- ▶ Business Intelligence tools are used to take strategic business decisions.

- **Statistical Analysis**

- ▶ "What happen?" by using past data in the form of dashboards.
- ▶ includes collection, Analysis, interpretation presentation, and modeling of data. Diagnostic Analysis

- **Diagnostic Analysis:**

- ▶ "Why did it happen?" by finding the cause from the insight found in Statistical Analysis.
- ▶ 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.

Types of Data Analysis: Techniques and Methods

- **Predictive Analysis**

"what is likely to happen" by using previous data.

- **Prescriptive Analysis**

Combines the insight from all previous Analysis to determine which action to take in a current problem or decision.

Data Analysis Process

- Gathering information by using proper application or tool which allows you to explore the data and find a pattern in it. Based on that, you can take decisions, or you can get ultimate conclusions.
- Data Analysis consists of the following phases:
 - ▶ Data Requirement Gathering
 - ▶ Data Collection
 - ▶ Data Cleaning
 - ▶ Data Analysis
 - ▶ Data Interpretation
 - ▶ Data Visualization



Data Requirement Gathering

- Why do you want to do this data analysis?
- Find out the purpose or aim of doing the Analysis.
- Decide which type of data analysis you want to do!

In this phase, you have to decide what to analyze and how to measure it, you have to understand why you are investigating and what measures you have to use to do this Analysis.

Data Collection

- Collect your data based on requirements.
- Once you collect your data, remember that the collected data must be processed or organized for Analysis.
- As you collected data from various sources, you must have to keep a log with a collection date and source of the data.

Data Cleaning

- Check for non useful or irrelevant data to your aim of Analysis.
- Check for duplicate records, white spaces or errors.
- The data should be cleaned and error free.
- This phase must be done before Analysis because based on data cleaning, your output of Analysis will be closer to your expected outcome.

Data Analysis

- Once the data is collected, cleaned, and processed, it is ready for Analysis.
- As you manipulate data, you may find you have the exact information you need, or you might need to collect more data.
- During this phase, use data analysis tools and software which will help you to understand, interpret, and derive conclusions based on the requirements..



Data Interpretation

- After analyzing your data, it's finally time to interpret your results.
- Choose the way to express or communicate your data analysis either you can use simply in words or maybe a table or chart.
- Use the results of your data analysis process to decide your best course of action.

Data Visualization

- Charts and graphs
- Data visualization often used to discover unknown facts and trends.
- By observing relationships and comparing datasets, you can find a way to find out meaningful information.

Questions