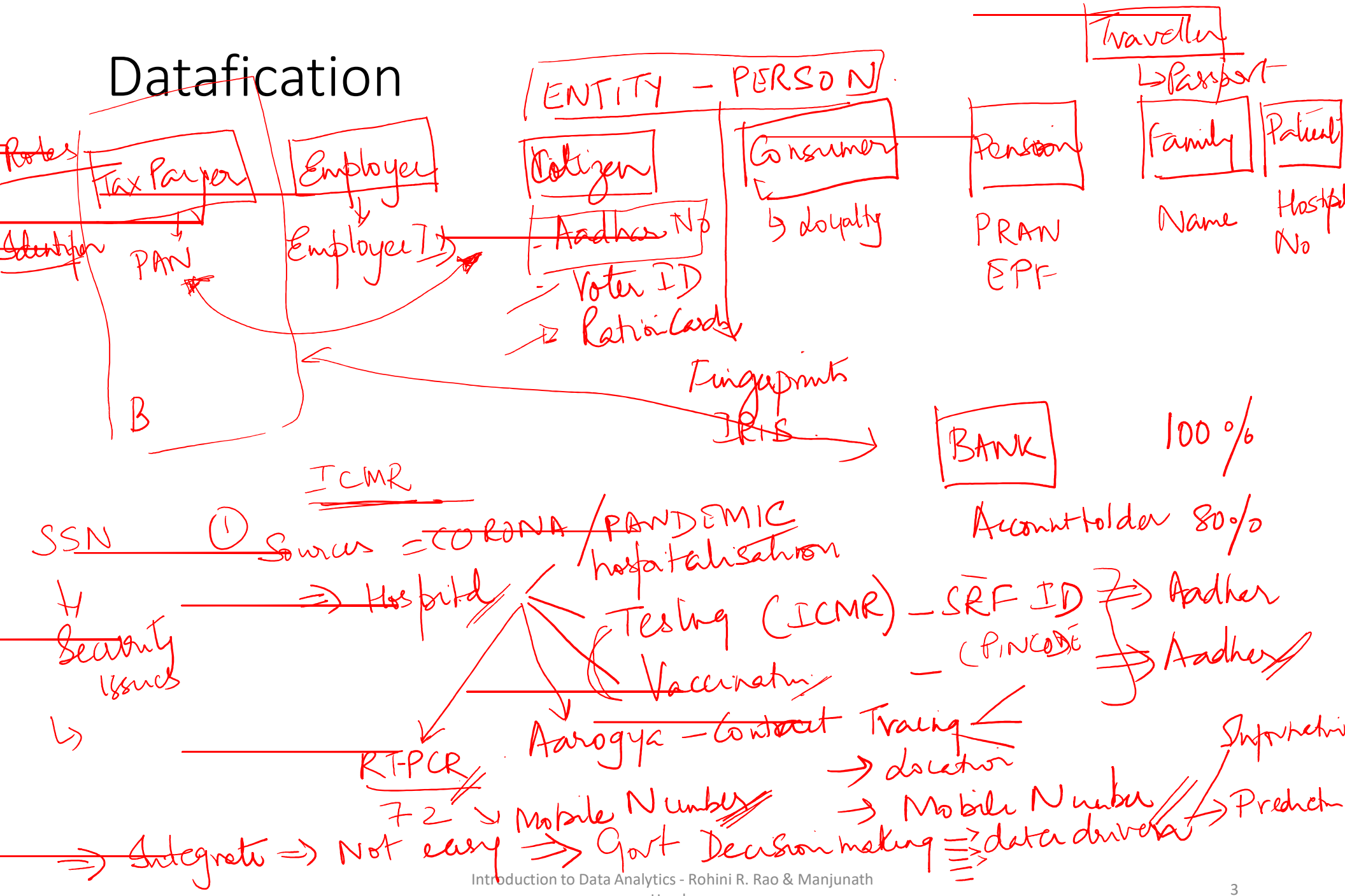
Introduction to Data Analytics
DSE-2151

# B.tech Data Science & Engineering
# 3rd Semester
# Introduction to DA

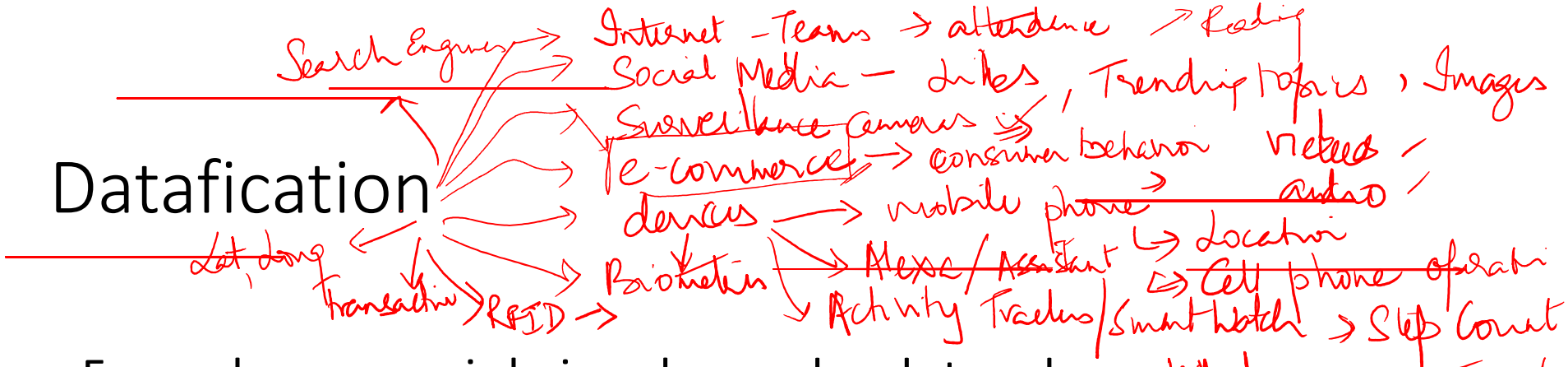
Rohini R. Rao & Manjunath Hegde
Dept of Computer Applications
Sept 2021
(Slide set 1 out of 5)

# Contents

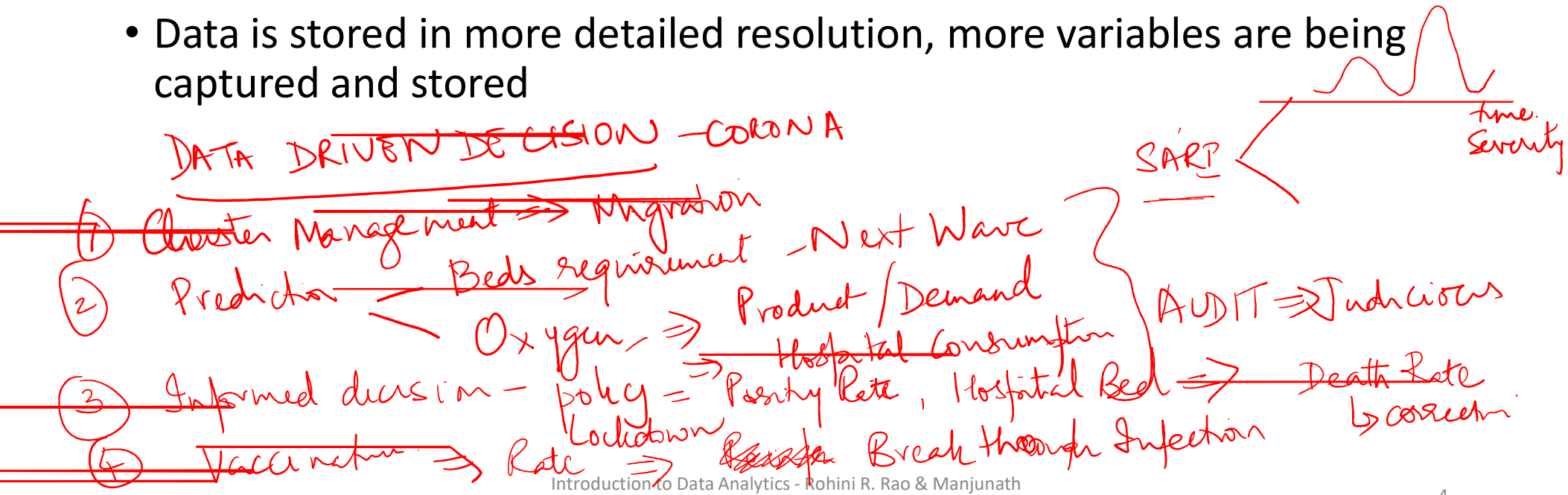
# Datafication



# Datafication




- Every phenomena is being observed and stored
- People and devices are connected to the grid
- Internet of things is growing faster than Internet of people
- Kryder's law says that the density and capacity of hard drive storage media will double every 18 months
- Data is stored in more detailed resolution, more variables are being captured and stored

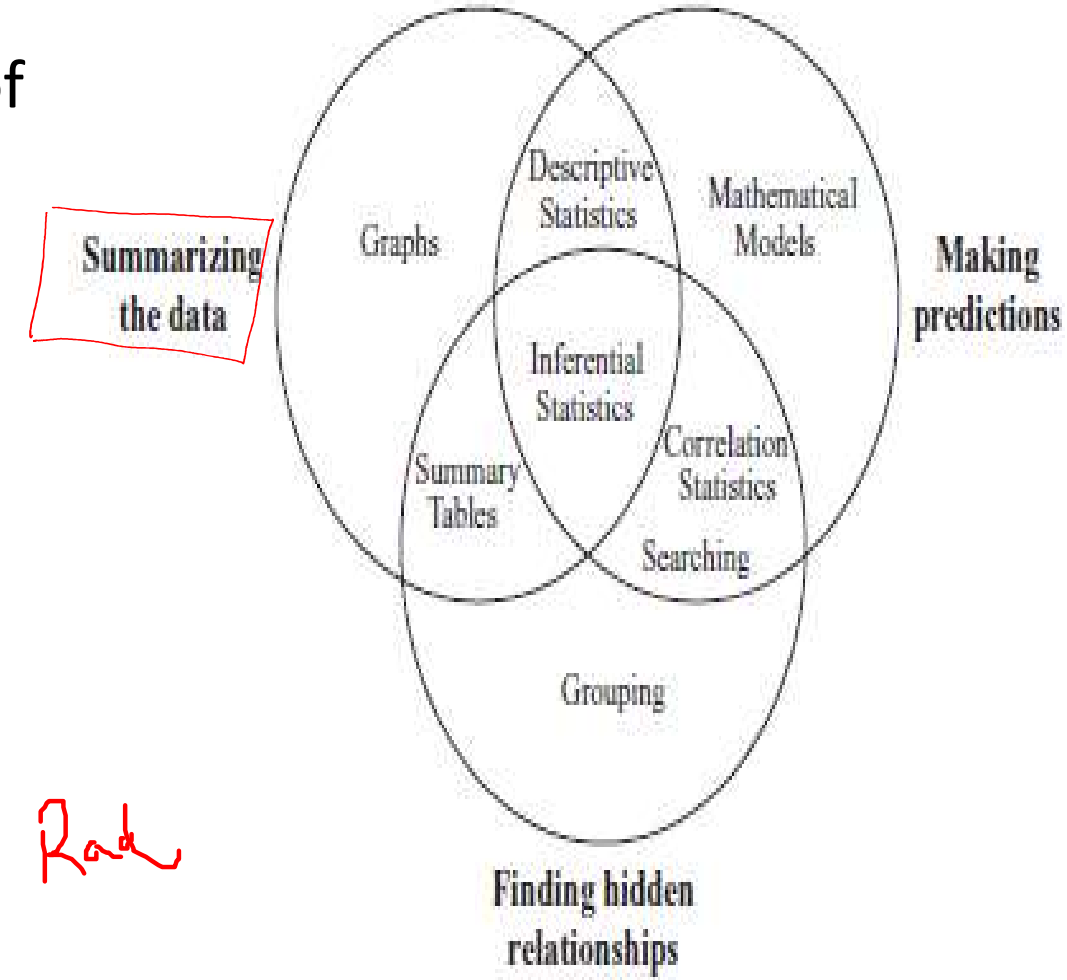


# Need for making sense of data

- unprecedented amount of data is being generated
  - Data rich but information poor situation
  - Data repositories becomes Data tombs
  - So decision making was based on intuition rather than information
  - Expert systems rely on domain experts to manually input system knowledge to knowledge bases
- Leads to information overload
- Need for making sense of the data
- the analysis of data includes
  1. Summarizing and interpret the data
  2. how to identify nontrivial facts, patterns, and relationships in the data
  3. how to make predictions from the data.

# Definition - Data Analytics

- is the science of examining raw data with the purpose of drawing conclusions about that information.
- To make better business decisions
- in the sciences to verify or disprove existing models or theories.
- Consists of
  - exploratory data analysis (EDA)
  - confirmatory data analysis (CDA)
  - Qualitative data analysis (QDA)




Data Analysis tasks & methods

- Your Inputs
  - dfsdafjasdlk

# *To become a Data Analyst:*

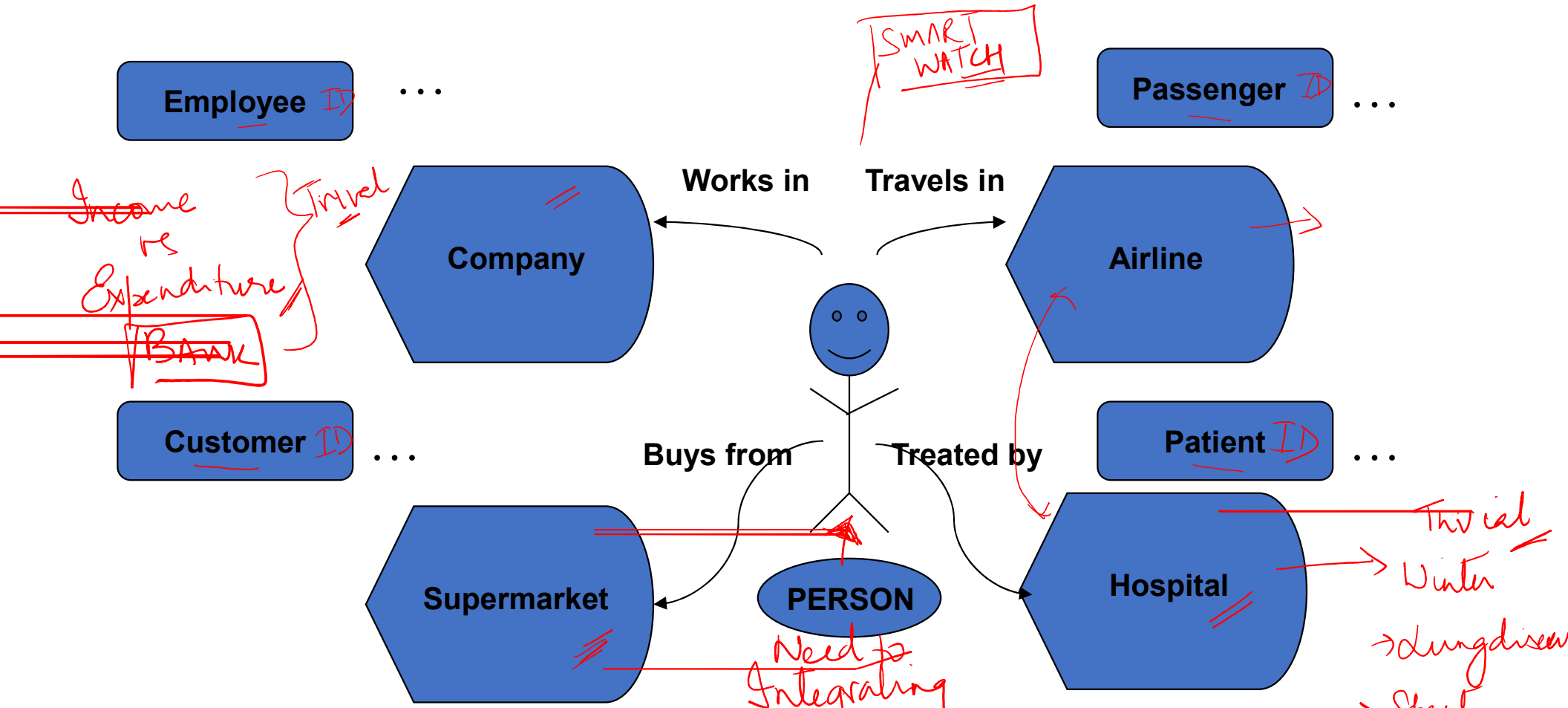
(Expected industry requirements)

- Programming skills:
  - Knowing programming languages are R and Python
- Statistical skills and mathematics:
  - Descriptive and inferential statistics and experimental designs are a must for data scientists.
- Machine learning skills
- Data wrangling skills:
  - The ability to map raw data and convert it into another format that allows for a more convenient consumption of the data.
- Communication and Data Visualization skills
- Data Intuition:
  - it is extremely important for professional to be able to think like a data analyst.

# Definition - Data science includes

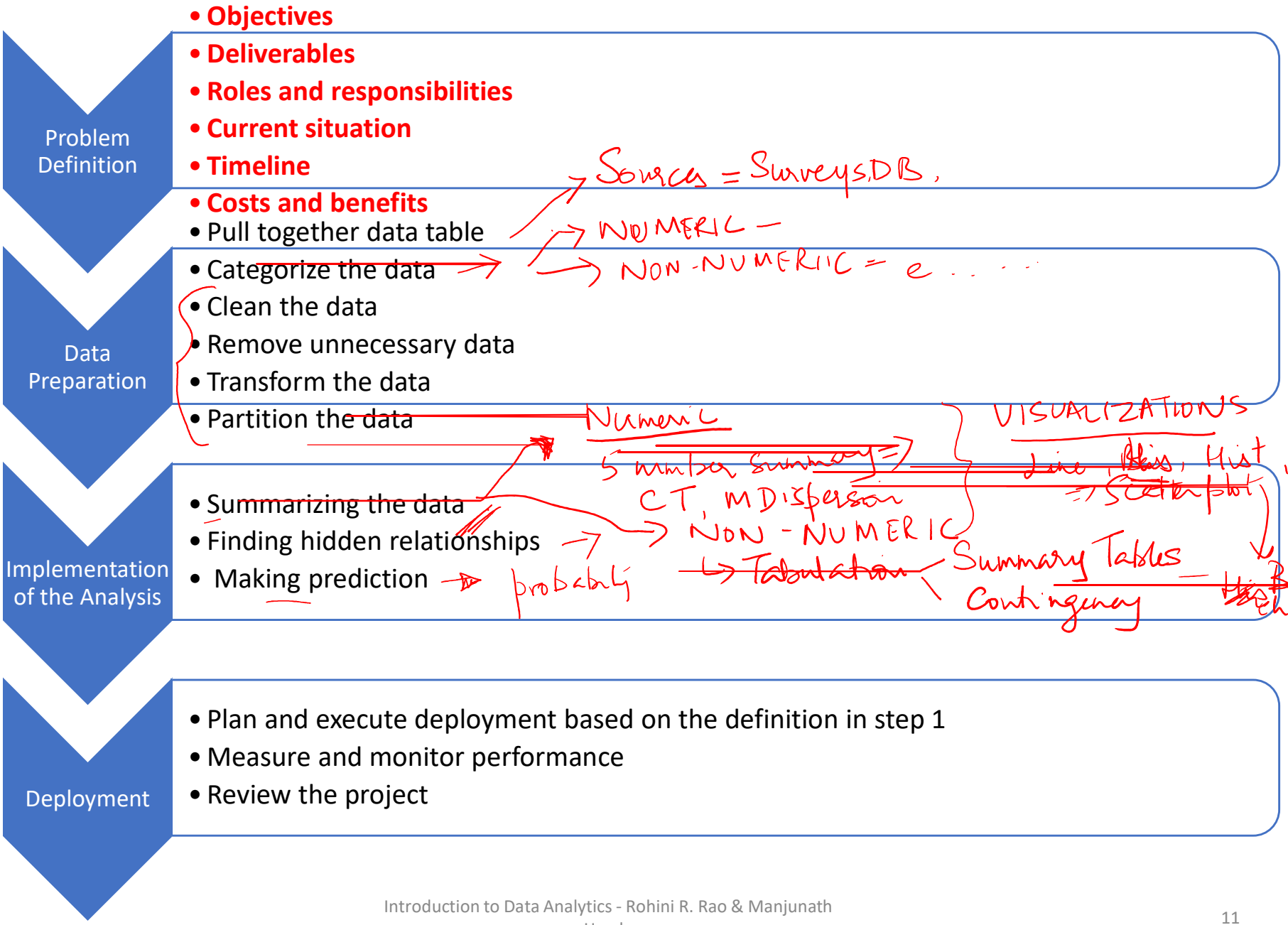
- refers to analytics and data mining in every applicable domains
- Includes
  - Computer science:
    - Internet technology , graph theory,  distributed architectures such as Hadoop, computer programming (Python, Perl, R) and processing sensor and streaming data
  - Statistics:
    - design of experiments including multivariate testing, cross-validation, sampling,
  - Machine learning and data mining
  - Operations research
  - Business intelligence, OLAP

# MOTIVATING EXAMPLE



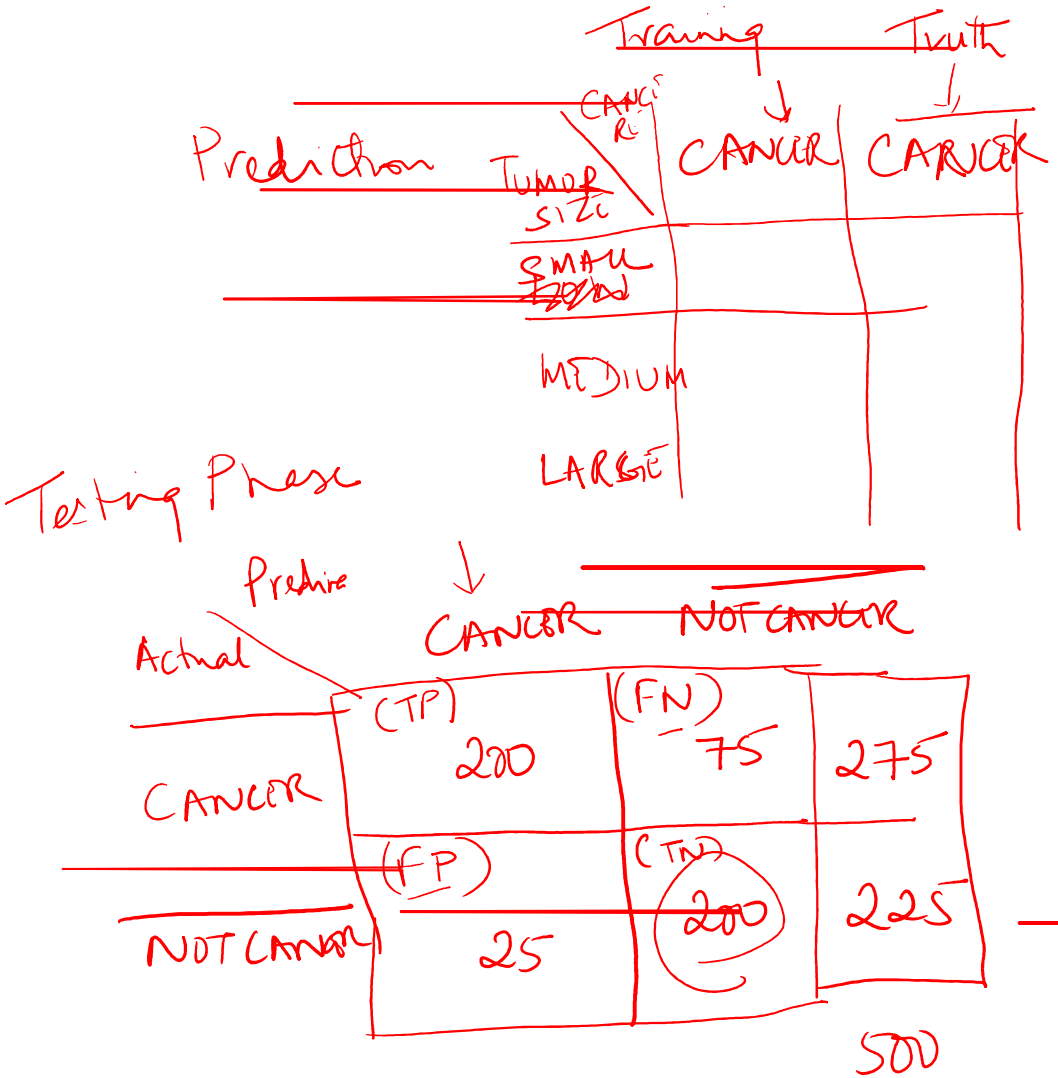

QUESTION : ????

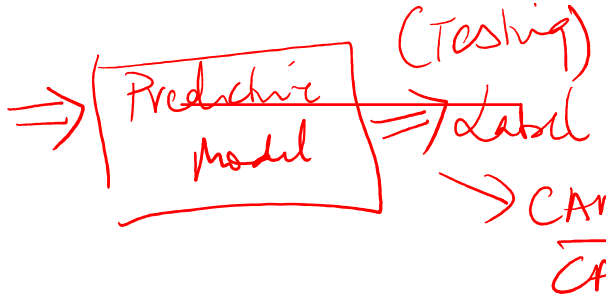
# Steps in data analysis projects



# Steps in a data analysis project
# Step 1– Problem Definition

- **1.1. Objectives**
  - to define the business or scientific problem to be solved
  - Helps to create a focused plan to execute
  - Success criteria for the project should be defined and measurable
  - Collection of suitable information must be available
  - **For example**: Make recommendations to improve sales on the web site by a specific amount. Sub Objectives are:
    1. Identify categories of web site users (on the basis of demographic information) that are more likely to purchase from the web site.
    2. Categorize users of the web site on the basis of usage information.
    3. Determine if there are any relationships between buying patterns and web site usage patterns.
- **1.2. Deliverables**
  - Will the solution be a report, a computer program to be used for making predictions, a new workflow or a set of business rules
  - When developing predictive models, it is useful to understand any required level of accuracy
  - It is also important to understand the consequences of answering questions incorrectly.
  - In many situations, the time to create a model can have an impact on the success of the project




Accuracy of Classifier

$= \frac{TP + TN}{N} = \frac{400}{500} = 80\%$

Error rate = 20% 90%

# Steps in a data analysis project
# Step 1 – Problem Definition

- **1.3. Roles and responsibilities**
  - **Project leader**: responsible for putting together a plan and ensuring the plan is executed.
  - **Subject matter experts and/or business analysts**: have specific knowledge of the subject matter or business problems including
    - (1) how the data was collected,
    - (2) what the data values mean,
    - (3) the level of accuracy of the data,
    - (4) how to interpret the results of the analysis,
    - (5) the business issues being addressed by the project.
  - **Data analysis/data mining expert**: familiar with statistics, data analysis methods and data mining approaches as well as issues of data preparation.
  - **IT expert**: expertise in pulling data sets together (e.g., accessing databases, joining tables, pivoting tables, etc.) as well as knowledge of software and hardware issues important for the implementation and deployment steps.
  - **Consumer:** will use the information derived from the data in making decisions, either as a one-off analysis or on a routine basis.

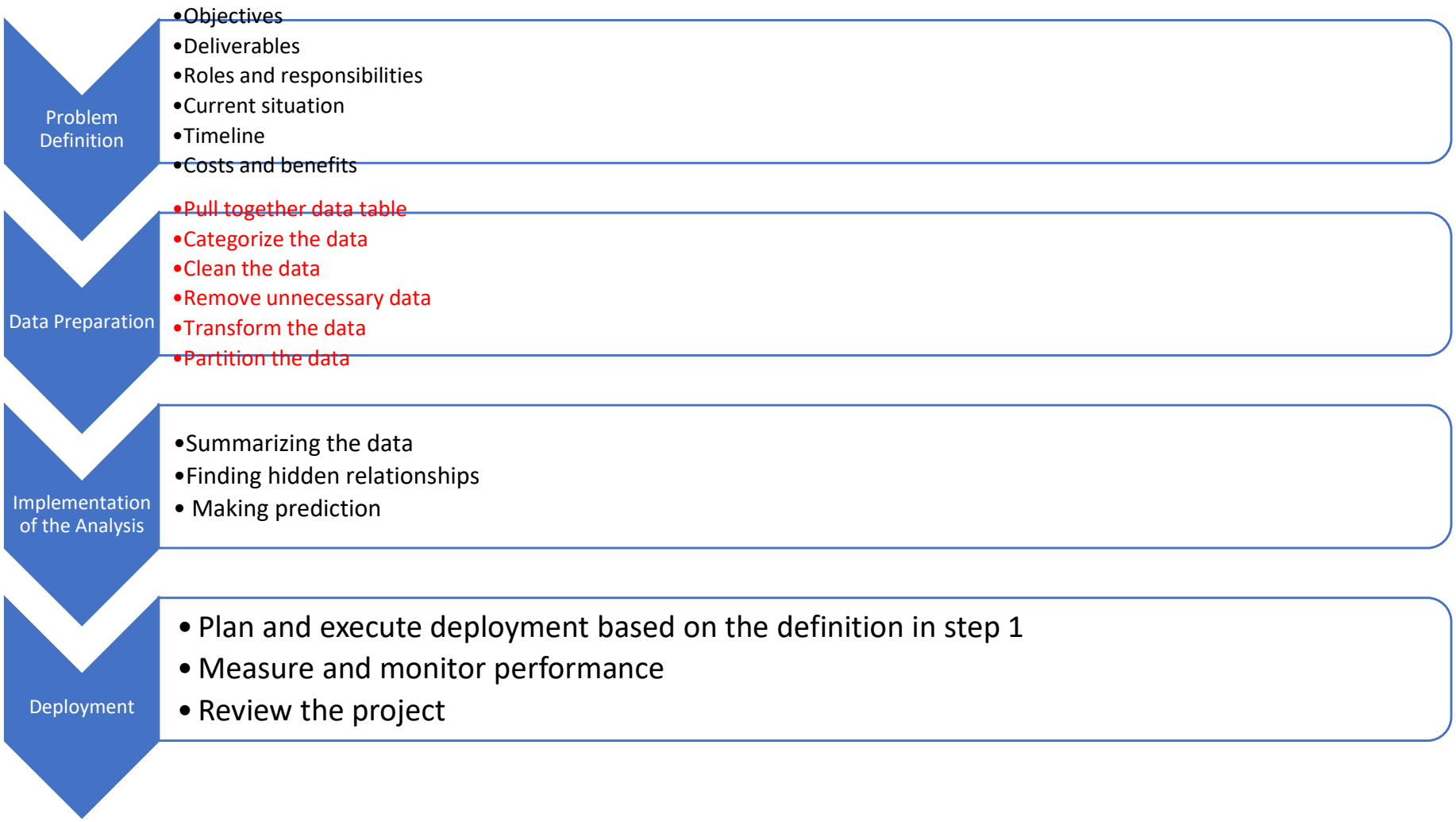
# Steps in a data analysis project
# Step 1 – Problem Definition

- **1.4. Current situation**
  - Define the constraints on the project
  - The sources and locations of the data to be identified.
  - Privacy or legal issues to be listed.
- **1.5. Timeline**
  - Preliminary implementation plan should be put together.
  - Time to be set aside for iteration of activities as the solution is optimized
- **1.6. Costs and benefits**
  - budget based on the plan could be used, alongside the business success criteria, to understanding the cost/benefits for the project

# Sample Project timeline

| Steps | Tasks | Parties involved | Deliverables | Timeline | Budget |
|---|---|---|---|---|---|
| | | | | Apr May Jun Jul Aug Sep Oct Nov | |
| Preparation | Kickoff meeting | All | | | $500 |
| | Create data tables | Lee | Data tables | | $4,000 |
| | Prepare data for analysis | Pam, Tony | Prepared data for analysis | | $10,000 |
| | Meeting to review data | All | Plan for implementation | | $500 |
| | | | | | |
| Implementation | Find hidden relationships | Pam, Tony | Key facts and trends | | $5,000 |
| | Build and optimize model | Pam, Tony | Model with 70% accuracy | | $3,000 |
| | Create a report | All | Report | | $1,000 |
| | Meet to review | All | Plan for next steps | | $500 |
| | | | | | |
| Deployment | Double blind test | Pam, Lee | Business impact projections | | $5,000 |
| | Production rollout | Pam, Lee | Ongoing review | | $5,000 |
| | Assess project | All | Project assessment report | | $1,000 |
| | | | | | $35,500 |

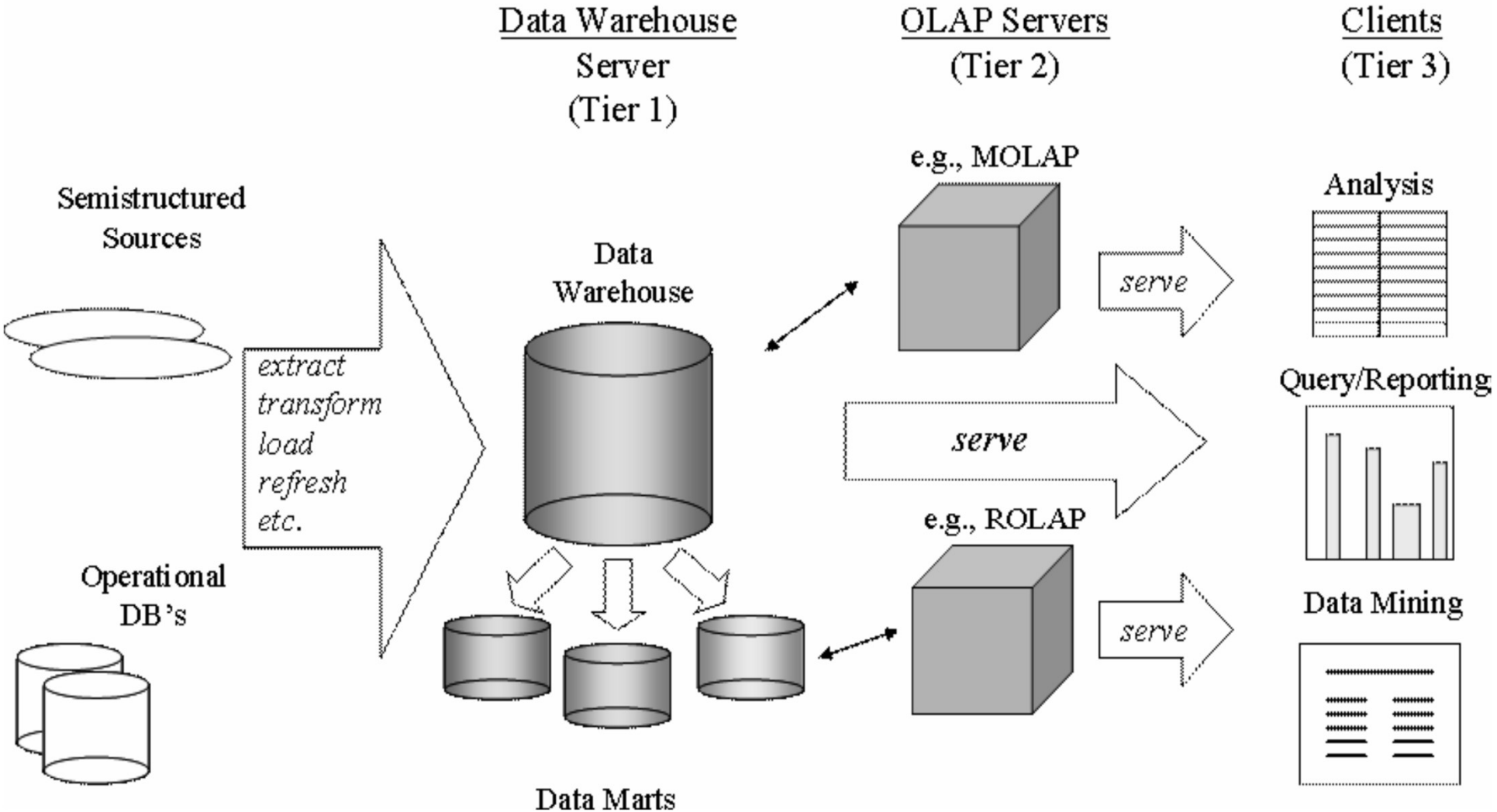
# Steps in data analysis projects



# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.1. Pull together data table**
  - **2.1.1. Query databases to access data**
  - **2.1.2. Integrate multiple data sets and format as a data table**
    - **Data sources include:**
      - Surveys or polls – sampling & bias
      - Experiments - For example, when studying the effects of a new drug, a double blind study is usually used
      - Observational and other studies – when unethical to get experimental data
      - Operational database
      - Data warehouse or Historical data base
      - Purchased data base

# Steps in a data analysis project
## Step 2 – Data Preparation
## Data Sources



# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.1. Pull together data table**

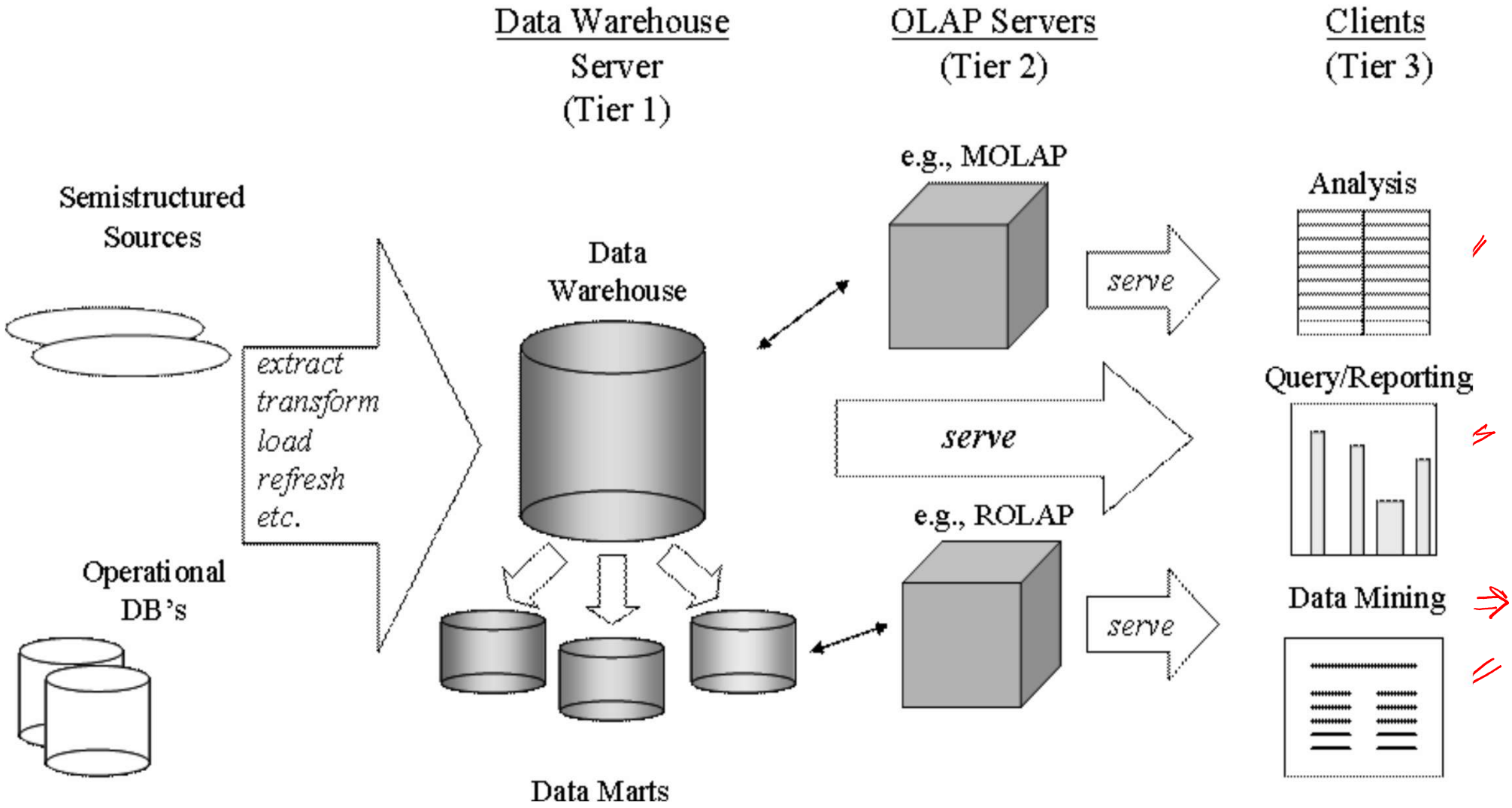
- **Data sets :**
  - contain data objects which represents an entity
  - Data object is described by data fields called attributes
  - Attributes are also called as dimension, feature , variable
  - Values for a given attribute is observation
  - A set of attribute values are also called as attribute vector

| VIN | Manufacturer | Weight | Number of cylinders | Fuel efficiency |
|---|---|---|---|---|
| IM8GD9A_KP042788 | Ford | 2984 | 6 | 20 |
| IC4GE9A_DQ1572481 | Toyota | 1795 | 4 | 34 |

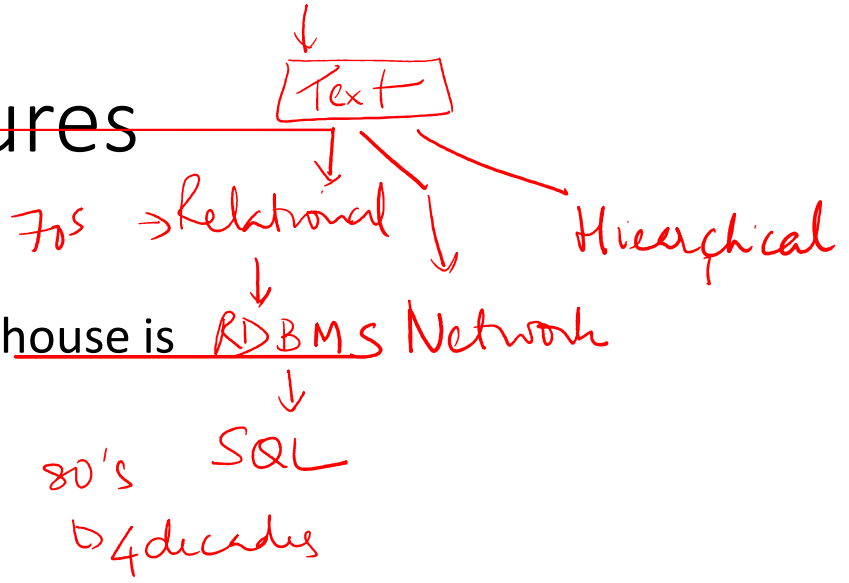
# What is a Data Warehouse ?

- provides architectures and tools for business executives to systematically organize, understand and use their data to make strategic decisions.
- Separate from organization’s operational databases.
- Supports information processing by providing a solid platform of consolidated historical data for analysis.
- A semantically consistent data store that serves as a physical implementation of a decision support data model and stores that information on which an enterprise needs to make strategic decisions.

# Data Warehouse Architecture



# Data warehouse key features




- According to Willian H. Inmon , a Data Warehouse is
  - Subject- oriented
  - Integrated
  - Time-variant
  - Non-volatile

- Data warehousing is the process of constructing and using data warehouses.

- Information from data ware houses are used to support business decision-making activities such as
  - (i) increasing customer focus – customer buying patterns
  - (ii) repositioning products and managing product portfolios – production strategies
  - (iii) analyzing operations and looking for sources of profit
  - (iv) managing the customer relationships , making environmental corrections and managing the cost of corporate assets.

# Characteristics of Data Warehouse:

- **Subject-oriented:**
  - means that all data pertinent to a subject/ business area are collected and stored as a single unit
- **Integrated:**
  - means that data from multiple disparate sources are transformed and stored in a globally accepted fashion
- **Static/non-volatile:**
  - means data once entered into the warehouse does not change frequently. It is periodically updated if required
- **Time variant:**
  - Data warehouse maintains historical data which are used to analyse the business or market trends and facilitate future predictions

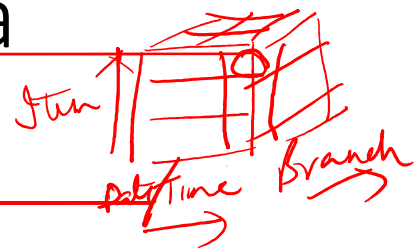
# Data Warehousing Terminology

- **Data sources:**
  - An organization has many functional units with their own data which has to be consolidated and put into a consistent form that would reflect the business of an organization as a whole..
- **Metadata:**
  - Metadata is the information about the data.
  - This is the layer of the data warehouse which stores the information like the source data, transformed data, date and time of data extraction, target databases, data and time of data loading, etc.
- **Measure attributes:**
  - A numerical value that can be summarized or can be aggregated upon.
- **Dimension attributes:**
  - Dimensions can be defined as the perspectives used for looking at the data.

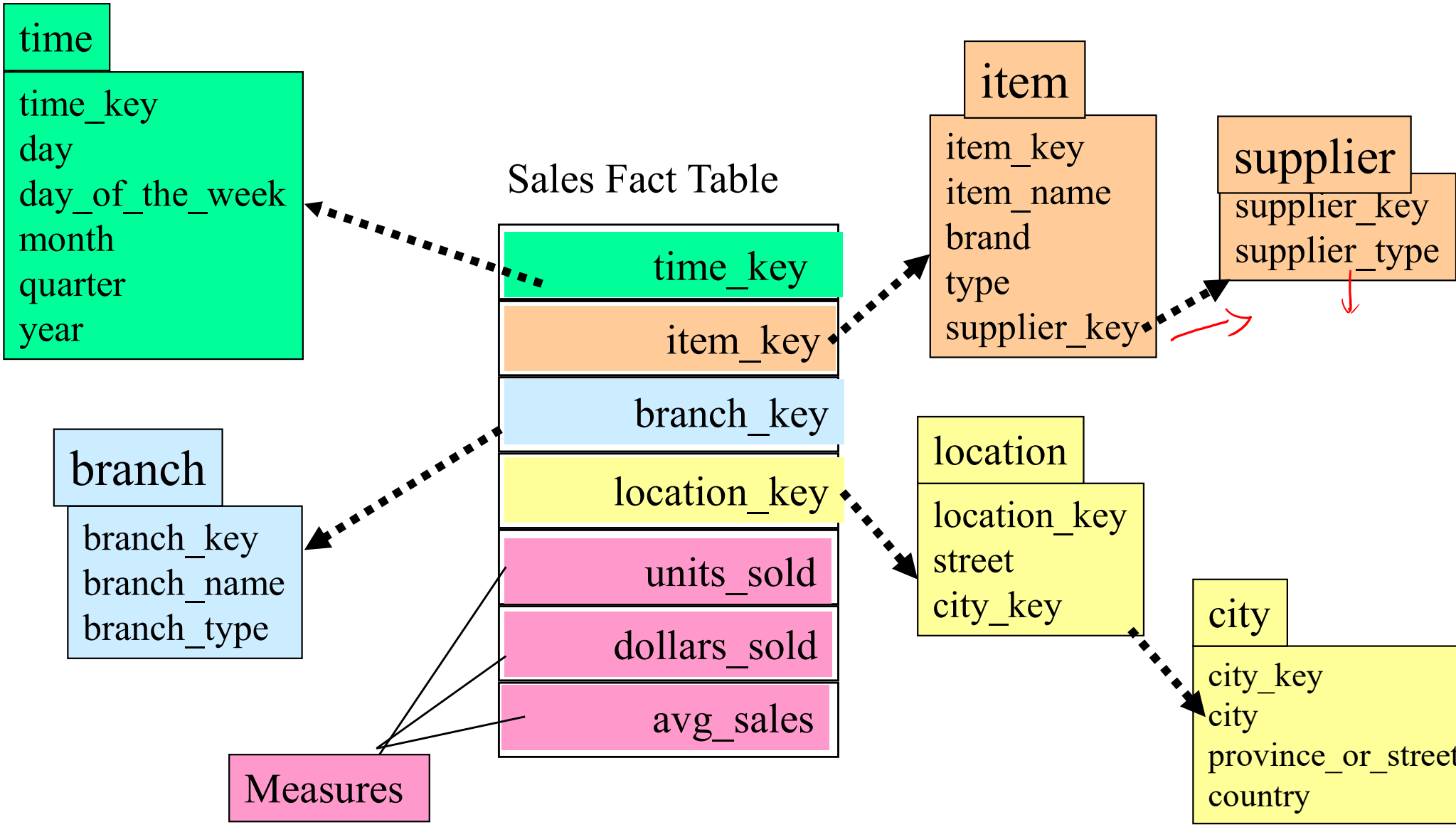
# Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

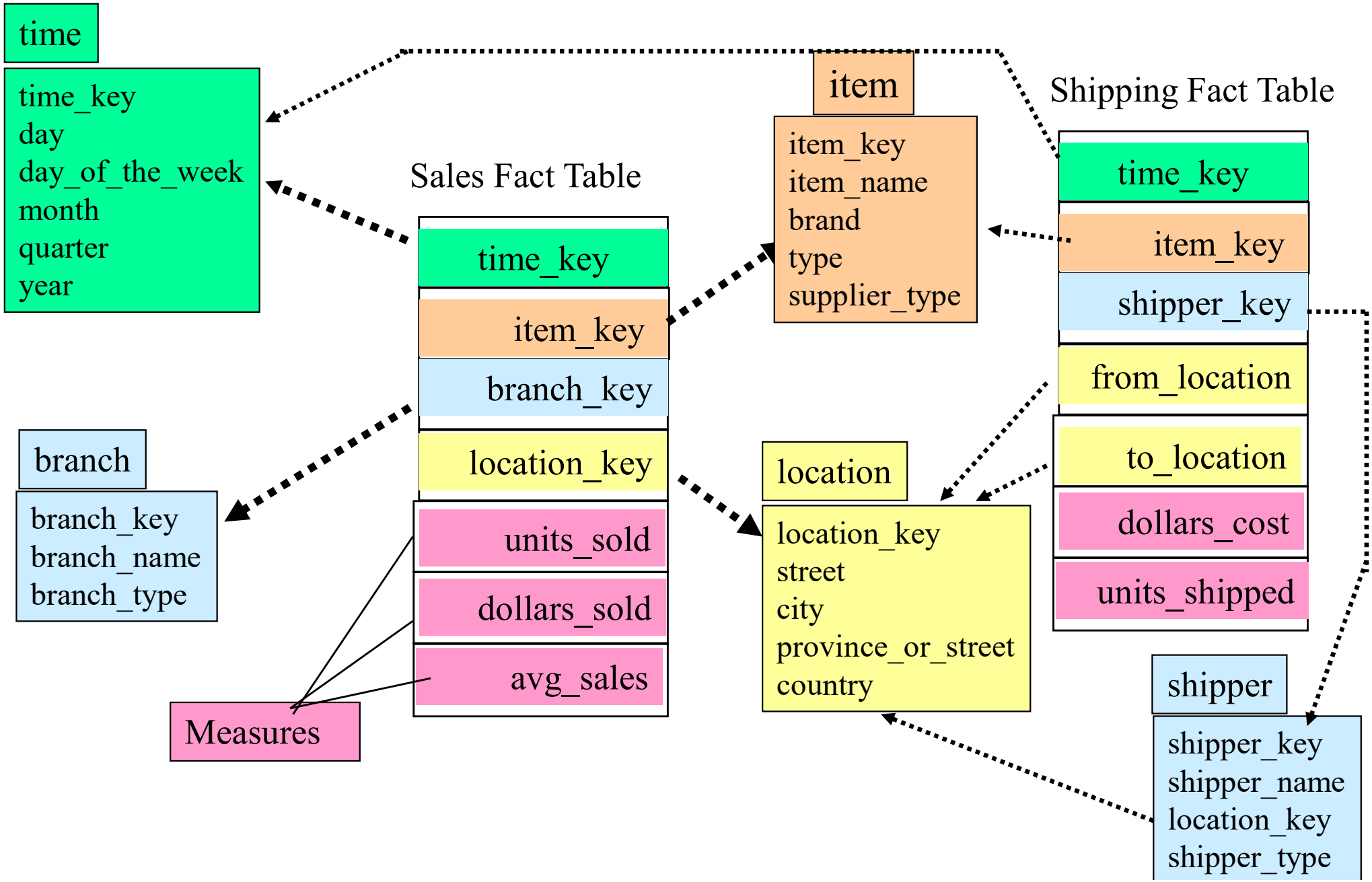
# Example of Star Schema



# Example of Snowflake Schema



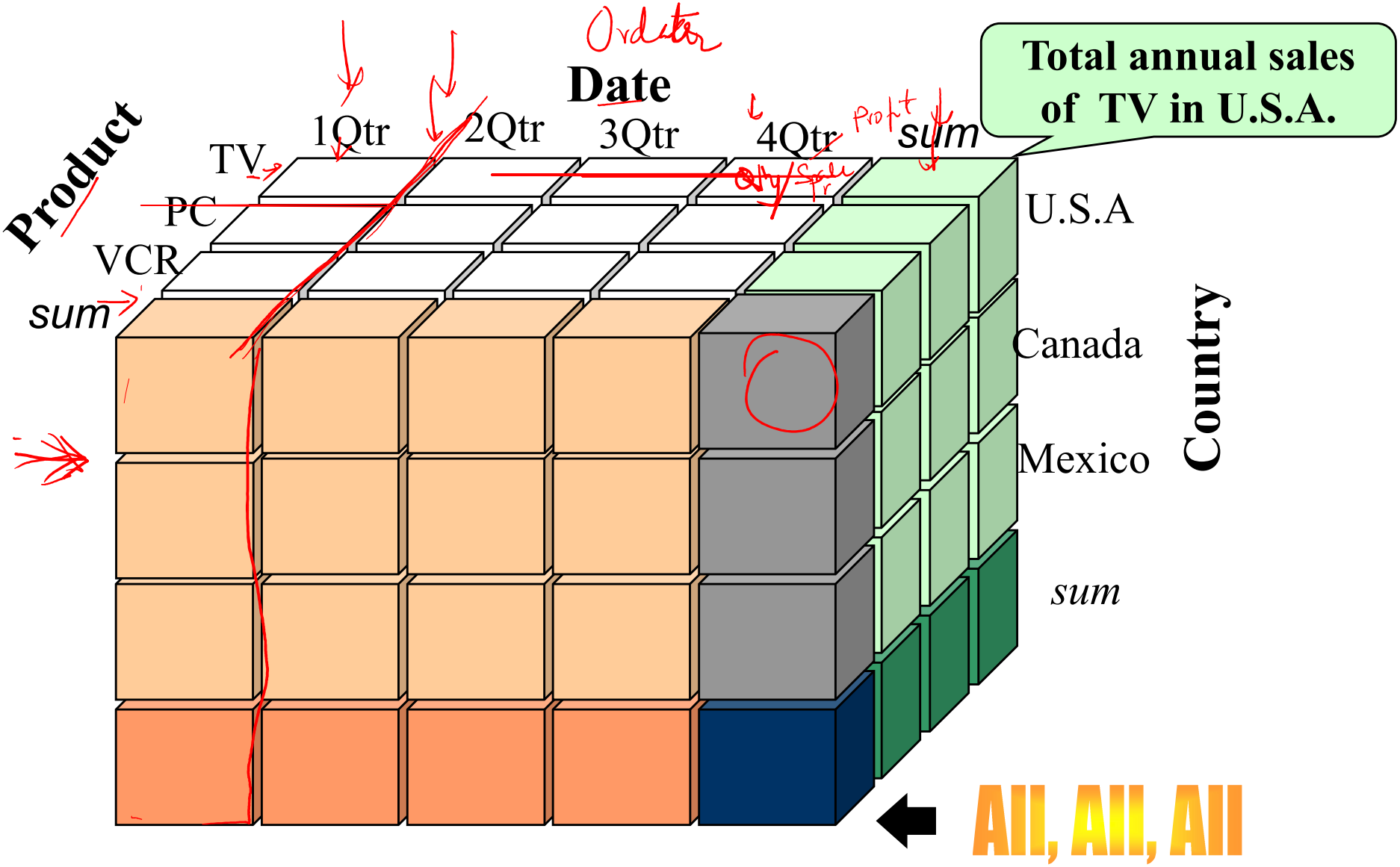
# Example of Fact Constellation



# From Tables and Spreadsheets to Data Cubes

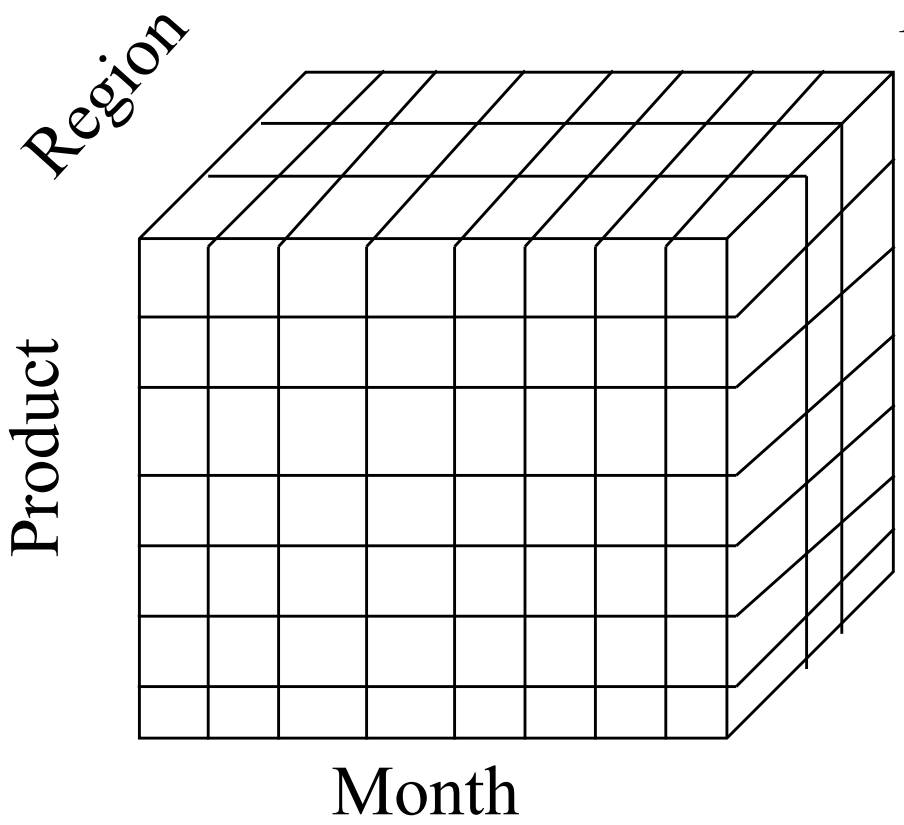
- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

# A Sample Data Cube

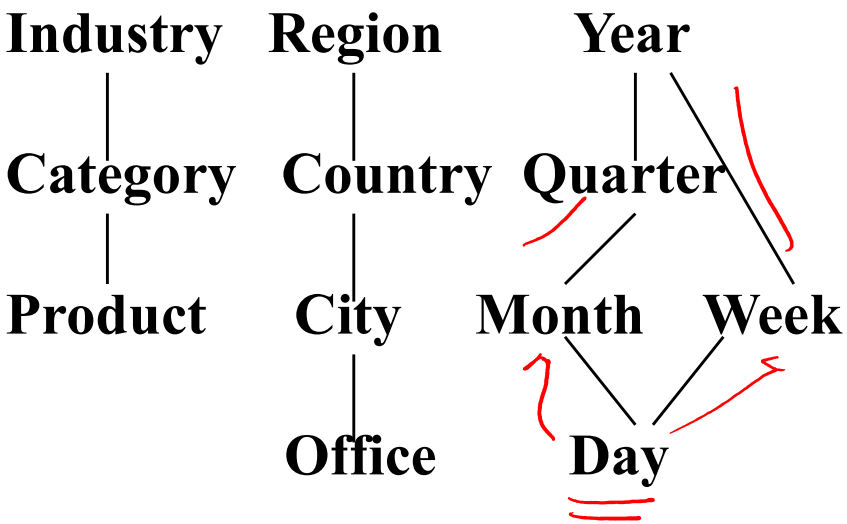


# Multidimensional Data

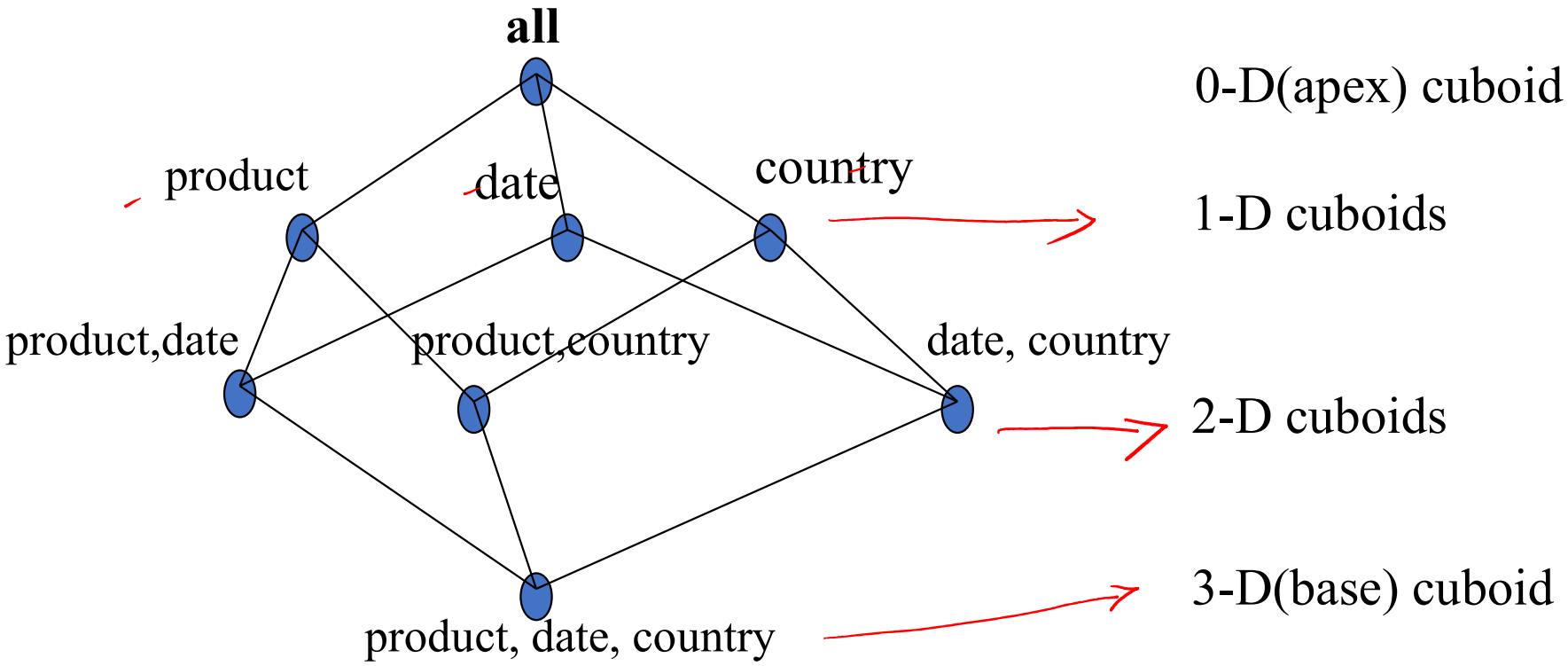
- Sales volume as a function of product, month, and region




**Dimensions: Product, Location, Time**
**Hierarchical summarization paths**



# Cuboids Corresponding to the Cube



# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.2 Categorize the data**
  - **Constant:** A variable where every data value is the same.
    - Example - PI
  - **Dichotomous:** A variable where there are only two values
    - Example- Gender whose values can be male or female.
  - **Discrete, Categoric or Nominal:** A variable that can only take a certain number of values (either text or numbers).
    - Example- the variable Color where values could be black, blue, red etc.
  - **Continuous:** A variable where an infinite number of numeric values are possible within a specific range.
    - Example- temperature between the minimum and maximum temperature, the variable could take any value

# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.2. Categorize the data**

  **Scales of measurement**

  - Nominal: Scale describing a variable with a limited number of different values. The order of these values has no meaning.
  - Ordinal: This scale describes a variable whose values are ordered, the difference between the values does not describe the magnitude of the actual difference.
    - For example, a scale where the only values are low, medium, and high
  - Interval: Scales that describe values where the interval between the values has meaning.
    - For example, when looking at three data points measured on the Fahrenheit scale, 5 F, 10 F, 15 F. Interval scales do not have a natural zero.
  - Ratio: Scales that describe variables where the same difference between values has the same meaning (as in interval) but where a double, tripling, etc.
    - For example of a ratio scale is a bank account balance whose possible values are $5, $10, and $15

# Steps in a data analysis project
# Step 2 – Data Preparation

**2.2 Categorize the data**

**Roles in analysis**

- **Labels:** Variables that describe individual observations in the data.
- **Descriptors:**
  - These variables are almost always collected to describe an observation
  - They are also described as predictors or X variables.
- **Response:**
  - These variables (usually one variable) are predicted from a predictive model (using the descriptor variables as input).

# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.3. Clean the data**
  - in order to help improve the quality of data and consequently of the mining results.
  - data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.
- **Data analyzed may be**
  - Incomplete – no recorded values , lacking attribute values or containing only aggregate values
  - Noisy – containing errors or outliers
  - Inconsistent – containing discrepancies in the department codes used to categorize items

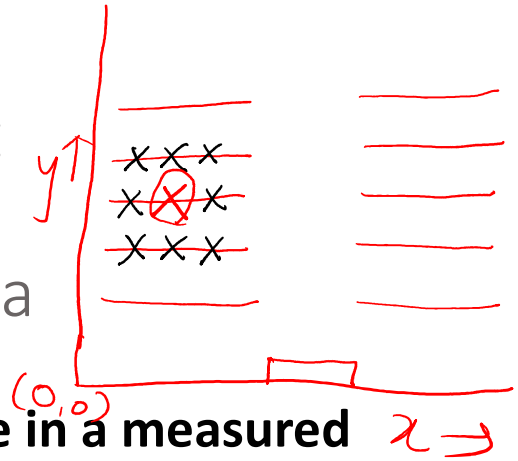
# Steps in a data analysis project
# Step 2 – Data Preparation
# 2.3. Data Cleaning as a process

- **Steps in the Data Cleaning Process are**
- **Discrepancy Detection**
  - **Can be caused by**
    - Poorly designed data entry forms
    - Human error in data entry
    - Deliberate errors
    - Data Decay
    - inconsistent use of codes and any inconsistent data representations
  - **To detect discrepancies**
    - use Meta Data
    - use descriptive data summaries
    - can also be done with Data scrubbing tools which use simple domain knowledge to detect errors
    - Data auditing tools analyze data to discover rules and relationships and detect data that violate these rules
- **Data Transformation**
  - Some inconsistencies can be corrected manually using external references
  - Data Migration tools and ETL tools may have GUIs to perform transformation
- The two step process is iterative

# Steps in a data analysis project
## Step 2 – Data Preparation
## 2.3. Data Cleaning for Missing values

- **Missing Values**
  - Ignore the tuple
  - Fill in the missing value manually
  - Use the global constant to fill in the missing value – ‘unknown’ – can result in overfitting
  - Use the attribute mean to fill in the missing value
  - Use the attribute mean for all samples belonging to the same class as the given tuple.
  - Use the most probable value to fill in the missing value.

# Steps in a data analysis project
## Step 2 – Data Preparation
## 2.3. Data Cleaning for Noisy Data




**Noisy Data- Noise or outlier is a random error or variance in a measured variable.**

- **Rule of thumb**
  - 1.5 * IQR above Q3 or below Q1 is an outlier
  - 2 standard deviations away from mean
- **Binning**
  - Smoothens a sorted data value by consulting its neighborhood.
  - The sorted values are distributed into a number of buckets or bins
  - Also called as local smoothing
    - Smoothing by bin means
    - Smoothing by bin median
    - Smoothing by bin boundaries
- **Regression**
  - Data can be smoothed by fitting the data to a function
- **Clustering**
  - Outliers can be detected by clustering

- **Inconsistent Data - can be detected by analyzing the meta data**

# Example 1 : Data Smoothing

- 4,8,15,21,21,24,25,28,34
- Equi-Depth Partitioning of depth or bin size=3
- Bin 1 : 4, 8, 15
- Bin 2 : 21, 21, 24
- Bin 3 : 25, 28, 34
- Smoothing by Bin Means
- Smoothing by Bin ~~Boundaries~~

# Example 2 – Data Cleaning

- Suppose that the data for analysis includes the attribute 'age'.
- A. Perform data cleaning using 5-number summary
- 13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35, 35, 36,40,45,46,52,70

| Min | Q1 | Q2 | Q3 | Max |
|---|---|---|---|---|
| 13 | | | | 70 |

- Using Rule of thumb - IQR
- Upper Bound:
- Lower Bound:
- Using Rule of thumb - SD =

mean = 29.96 SD = 12.94

29.96 ± (2 × 12.94) UB = 55.84

LB = 4.08

# Example 2 – Data Cleaning

- Suppose that the data for analysis includes the attribute ‘age’.
- A. Perform data cleaning using 5-number summary
- 13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35, 35, 36,40,45,46,52,70

| Min | Q1 | Q2 | Q3 | Max |
|---|---|---|---|---|
| | | | | |

- B. Smoothen the data using bin means into bins of depth 3

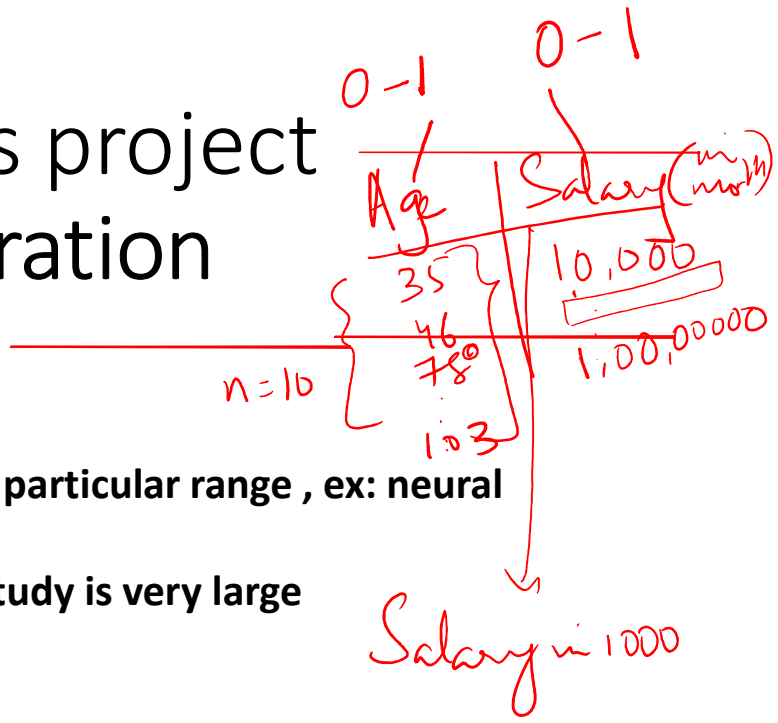
# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.4. Remove the unnecessary data**
  - Constants to be removed
  - Variables with too many missing data points should be considered for removal.
  - Redundant variables to be removed
    - Analysis of the correlations between multiple variables may identify such variables

# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.5. Transform the data**
  - applying mathematical transformations to the data for further analysis or mining
  - **Methods include**
    - **2.5.1. Normalization**
      - process where numeric columns are transformed using a mathematical function to a new range
    - **2.5.2. Value mapping**
      - To convert variables from one form to another
    - **2.5.3. Discretization**
      - where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior).
    - **2.5.4. Aggregation**
      - where summary or aggregation operations are applied to the data
      - For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.

# Steps in a data analysis project
# Step 2 – Data Preparation

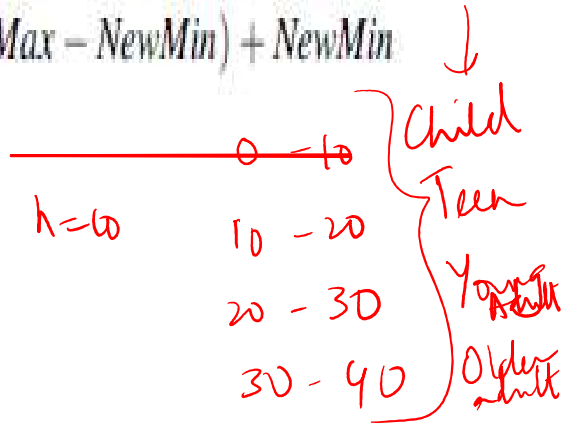



- **2.5.1. Normalization**
  - **Used when some algorithms require data to be in a particular range , ex: neural networks**
  - **When the difference in range of 2 variables under study is very large**
- **Methods include**
  - **Mean Normalization**

$$\boldsymbol{Value'} = \frac{\boldsymbol{Value} - \boldsymbol{average(x)}}{\boldsymbol{Max(x)} - \boldsymbol{Min(x)}}$$

  - **Min-max Normalization**

$$Value' = \frac{Value - OriginalMin}{OriginalMax - OriginalMin}(NewMax - NewMin) + NewMin$$



# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.5.1. Normalization**
  - **Used when some algorithms require data to be in a particular range , ex: neural networks**
  - **When the difference in range of 2 variables under study is very large**
- **Methods include**
  - **Z-score Normalization**

$$Value' = \frac{Value - \bar{x}}{s}$$

  - **Decimal Scaling**

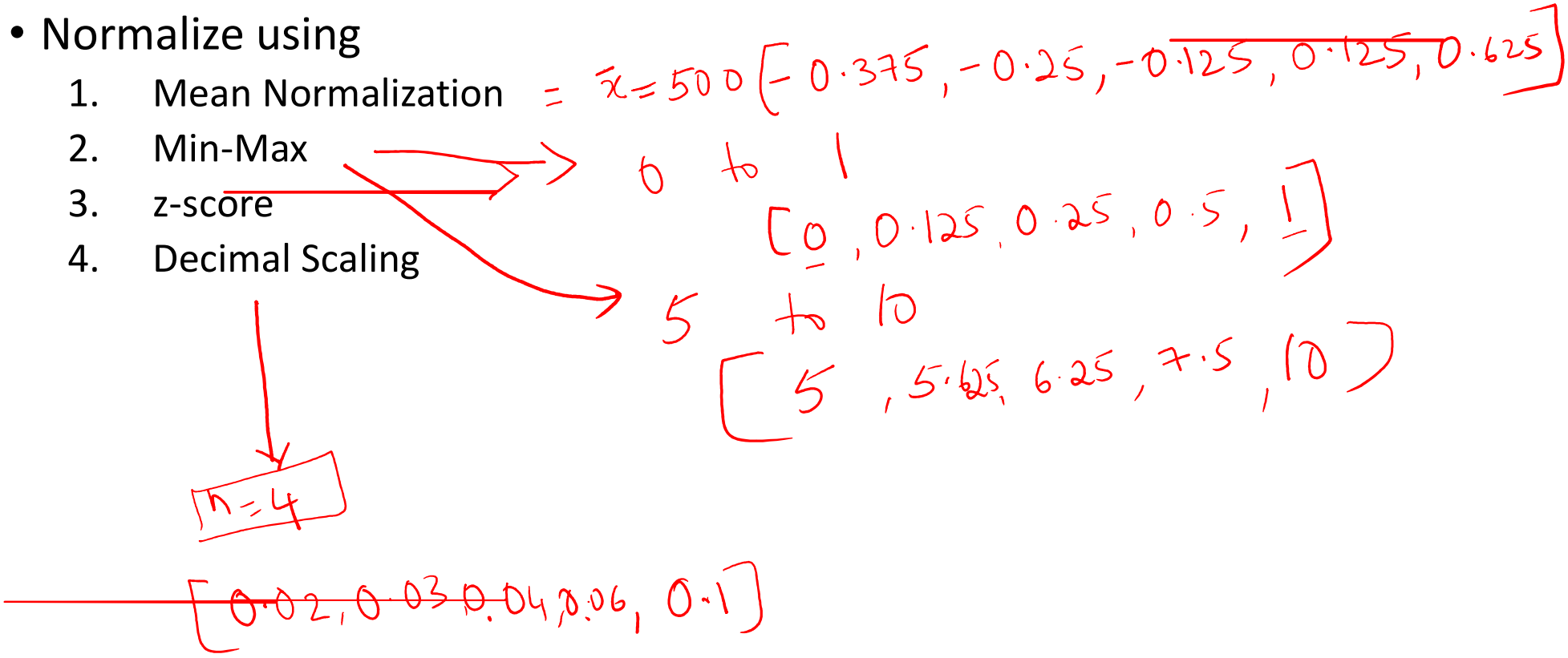
    **(-1 to 1)**

$$Value' = \frac{Value}{10^n}$$

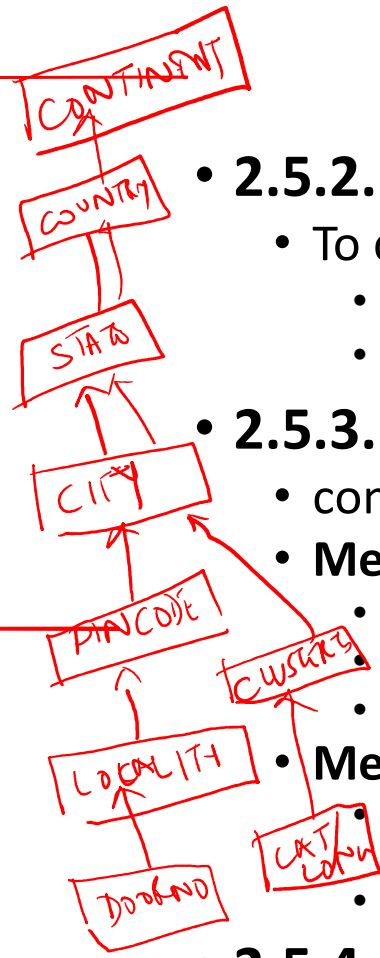
n is the number of digits of the maximum absolute value

# Example - Normalization

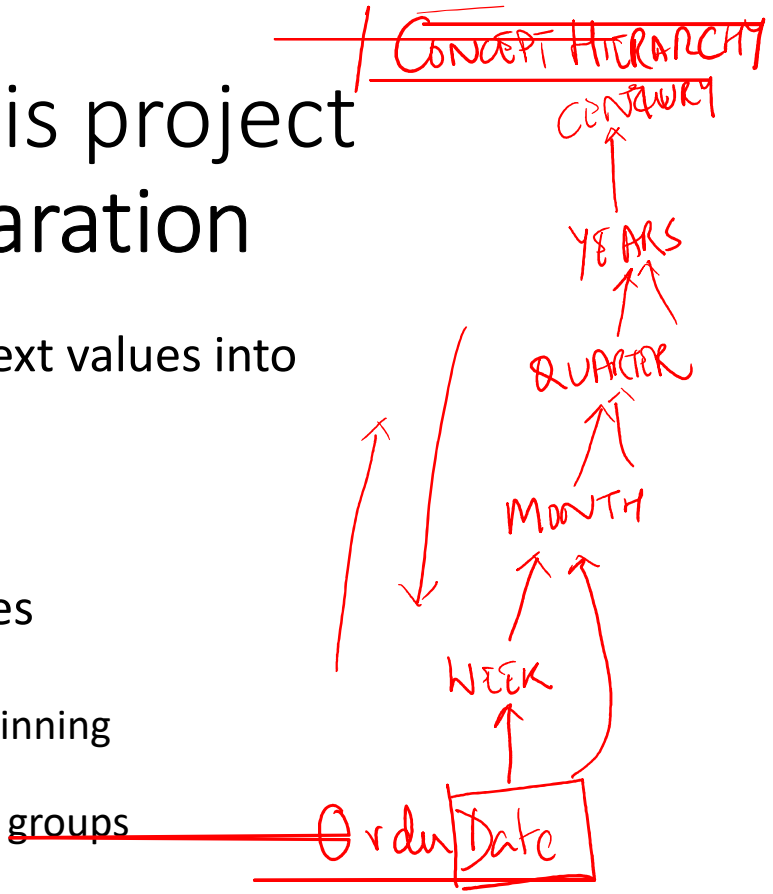
- 200,300,400,600,1000
- Normalize using
  1. Mean Normalization
  2. Min-Max
  3. z-score
  4. Decimal Scaling



# Steps in a data analysis project
# Step 2 – Data Preparation




- **2.5.2. Value Mapping**
  - To convert ordinal variables described using text values into
    - Numeric
    - Binary
- **2.5.3. Discretization**
  - converting continuous data into discrete values
  - **Method used for numeric attributes**
    - Equal width binning or Equal depth or frequency binning
    - Cluster Analysis
    - Example: age into age groups, income into income groups
  - **Methods used for nominal attributes**
    - Specification of a partial ordering of attributes explicitly at the schema level by users or experts
    - Example : date or location into concept hierarchy




- **2.5.4. Aggregation**
  - The variable derived from other variables present.
  - Usually has a mathematical operation
  - Example : age from DOB

# Value mapping

**Table 3.7.** Mapping nominal data onto a series of dummy variables

| *Original column* | *New variables (value-mapping)* | | | | |
|---|---|---|---|---|---|
| Color | Color = Red | Color = Green | Color = Blue | Color = Orange | Color = Yellow |
| red | 1 | 0 | 0 | 0 | 0 |
| green | 0 | 1 | 0 | 0 | 0 |
| blue | 0 | 0 | 1 | 0 | 0 |
| red | 1 | 0 | 0 | 0 | 0 |
| blue | 0 | 0 | 1 | 0 | 0 |
| orange | 0 | 0 | 0 | 1 | 0 |
| yellow | 0 | 0 | 0 | 0 | 1 |
| red | 1 | 0 | 0 | 0 | 0 |

# Steps in a data analysis project
# Step 2 – Data Preparation

- **2.6. Partition the data**
  - larger data sets take more computational time to analyze
  - Can create random subset
    - It should be sufficient to answer the analytical question
    - Should closely matches the target population
  - important to note the criteria for subset the data
  - Models for different data subsets can be created.
  - A method to consolidate data and results should also be defined

# Steps in data analysis projects –Example 1

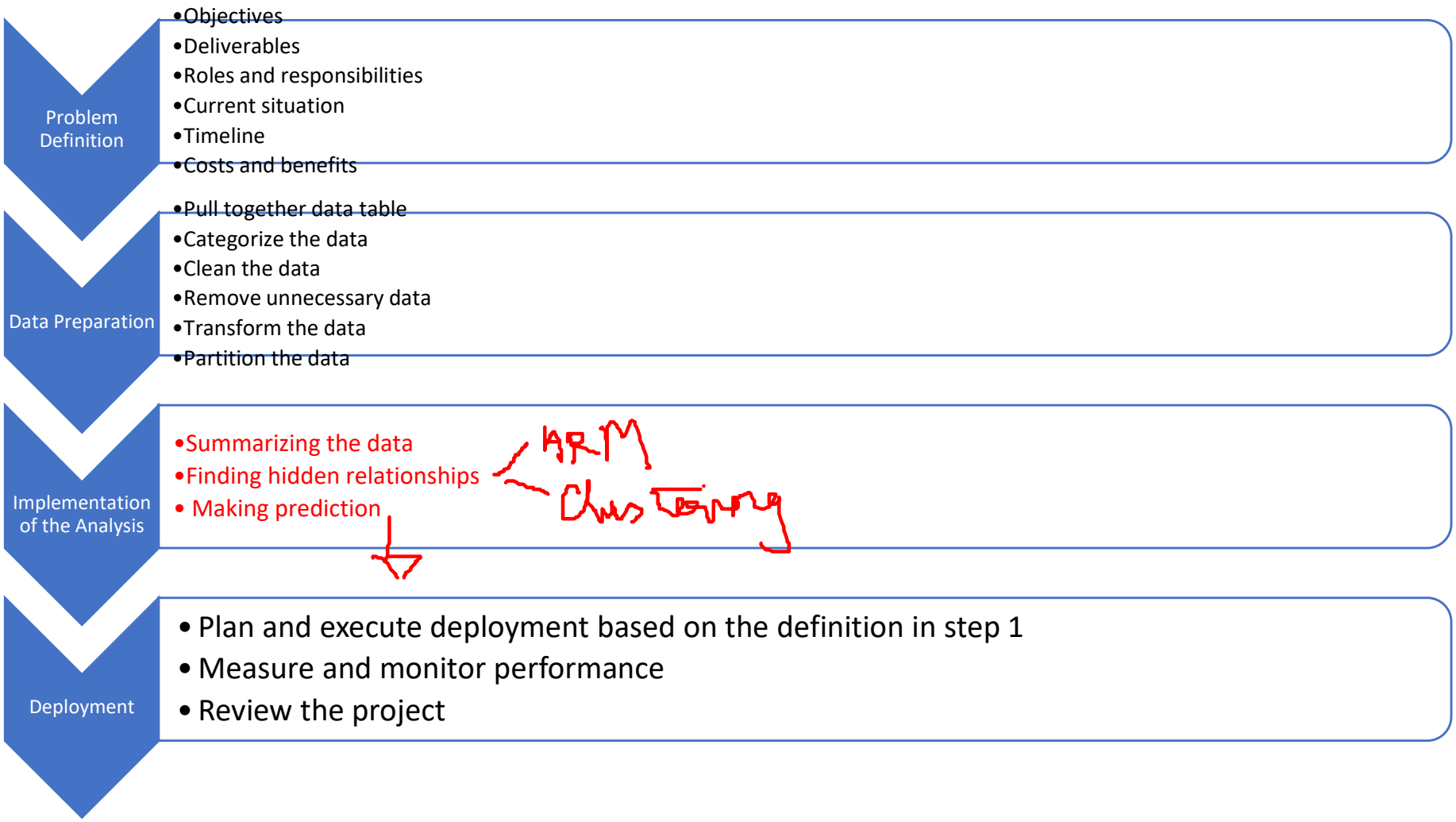
**Table 3.9.** Table of patient records

| Name | Age | Gender | Blood group | Weight (kg) | Height (m) | Systolic blood pressure | Diastolic blood pressure | Temperature (°F) | Diabetes |
|---|---|---|---|---|---|---|---|---|---|
| P. Lee | 35 | Female | A $Rh^+$ | 50 | 1.52 | 68 | 112 | 98.7 | 0 |
| R. Jones | 52 | Male | O $Rh^-$ | 115 | 1.77 | 110 | 154 | 98.5 | 1 |
| J. Smith | 45 | Male | O $Rh^+$ | 96 | 1.83 | 88 | 136 | 98.8 | 0 |
| A. Patel | 70 | Female | O $Rh^-$ | 41 | 1.55 | 76 | 125 | 98.6 | 0 |
| M. Owen | 24 | Male | A $Rh^-$ | 79 | 1.82 | 65 | 105 | 98.7 | 0 |
| S. Green | 43 | Male | O $Rh^-$ | 109 | 1.89 | 114 | 159 | 98.9 | 1 |
| N. Cook | 68 | Male | A $Rh^+$ | 73 | 1.76 | 108 | 136 | 99.0 | 0 |
| W. Hands | 77 | Female | O $Rh^-$ | 104 | 1.71 | 107 | 145 | 98.3 | 1 |
| P. Rice | 45 | Female | O $Rh^+$ | 64 | 1.74 | 101 | 132 | 98.6 | 0 |
| F. Marsh | 28 | Male | O $Rh^+$ | 136 | 1.78 | 121 | 165 | 98.7 | 1 |

# Example – Data transformation

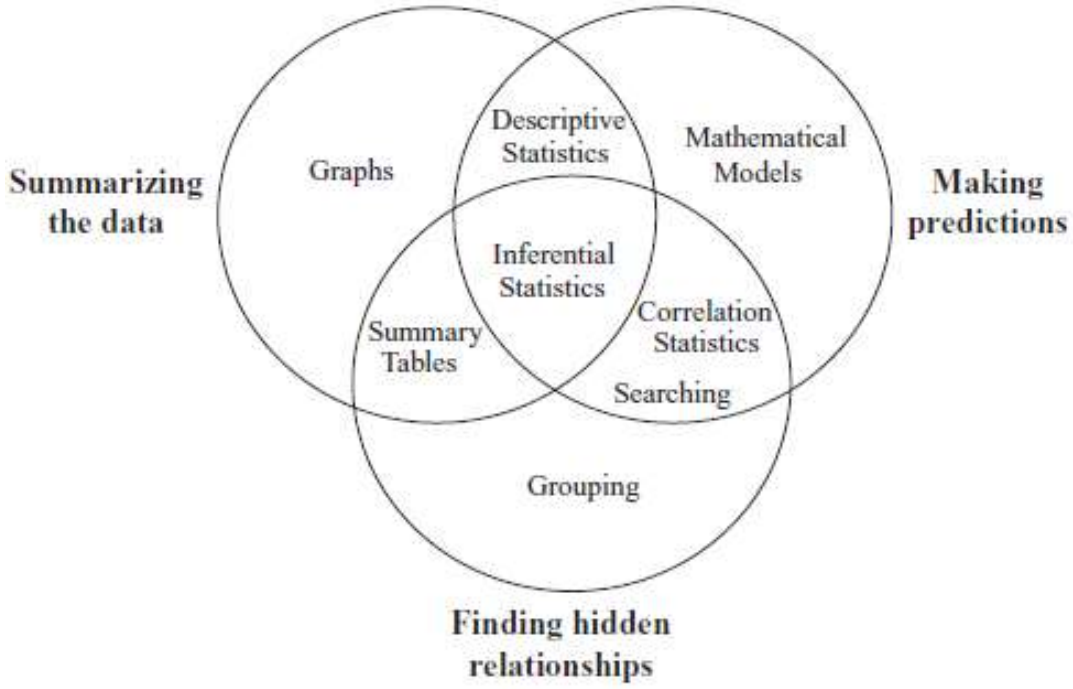
1. Create a new column by normalizing the Weight (kg) variable into the range 0 to 1 using the min-max normalization.
2. Create a new column by binning the Weight variable into three categories: low (less than 60 kg), medium (60–100 kg), and high (greater than 100 kg).
3. Create an aggregated column, body mass index (BMI)
4. Segment the data into data sets based on values for the variable Gender

# Steps in data analysis projects

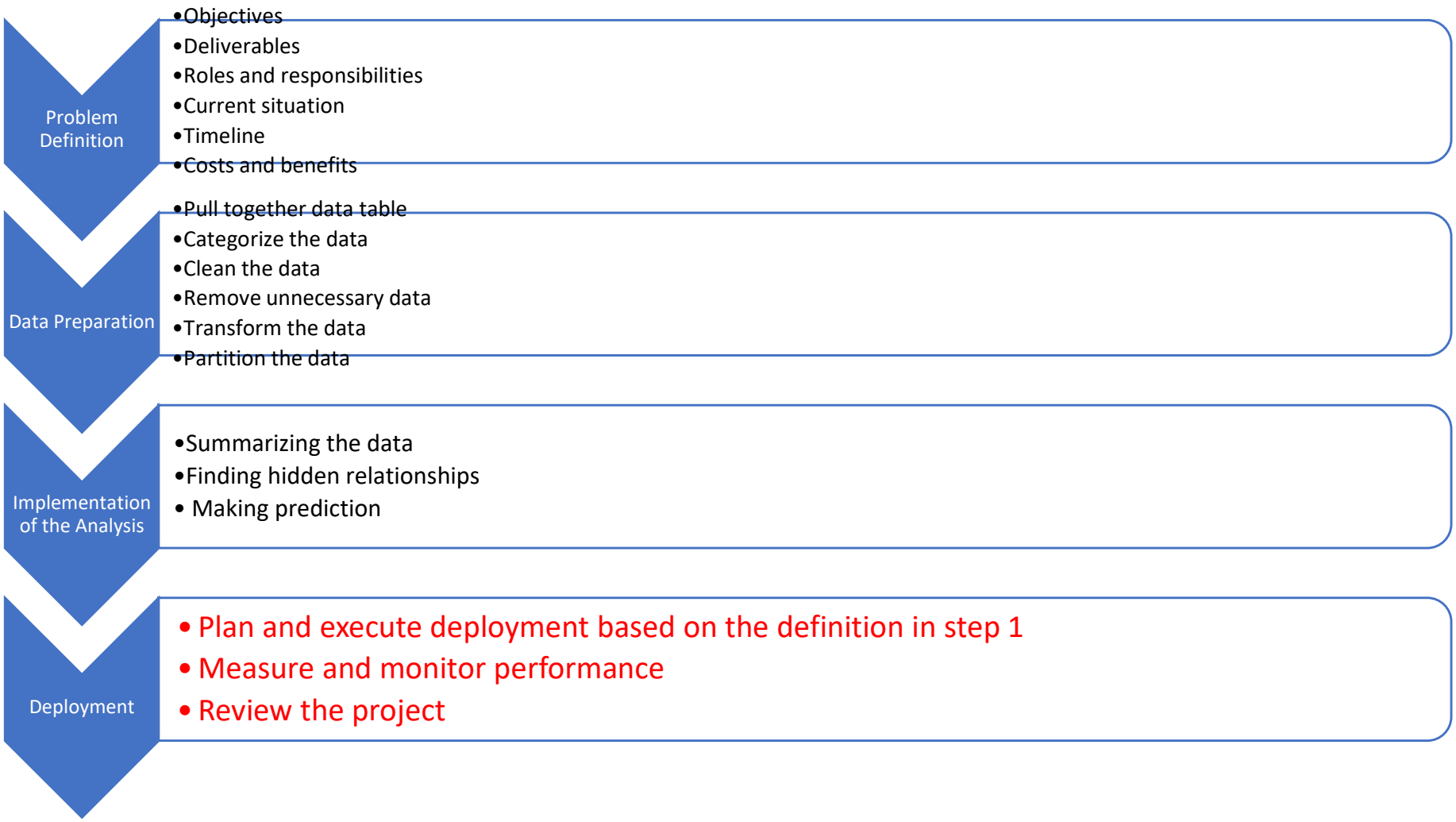


# Steps in a data analysis project

- **Step 3 – Implementation of the analysis**
  - Summarizing the data
    - Methods include
      - Summary tables
      - Graphs
      - Descriptive statistics
      - Inferential statistics
      - Correlation statistics
  - Finding hidden relationships
    - Methods include data mining
      - Market Basket Analysis
      - Clustering
  - Making prediction



# Steps in data analysis projects



# Steps in a data analysis project

- **Step 4 – Deployment**
  - analysis is translated into a benefit to the business
  - Plan and execute deployment based on the definition in step 1
  - Deliverables could be
    - Could be static report of the analysis to management or to the customer
    - Could be a predictive model deployed as standalone or integrated with other software such as spreadsheets or web pages.
  - Measure and monitor performance
  - Review the project

# References


- Glenn J. Myatt, Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining, John Wiley, November 2006.
- Dr Anil Maheshwari, “Data Analytics Made Accessible, 2021 Edition
- G. Shmueli, N. R. Patel, and P.C. Bruce, Data Mining for Business Intelligence, John Wiley and Sons, 2010