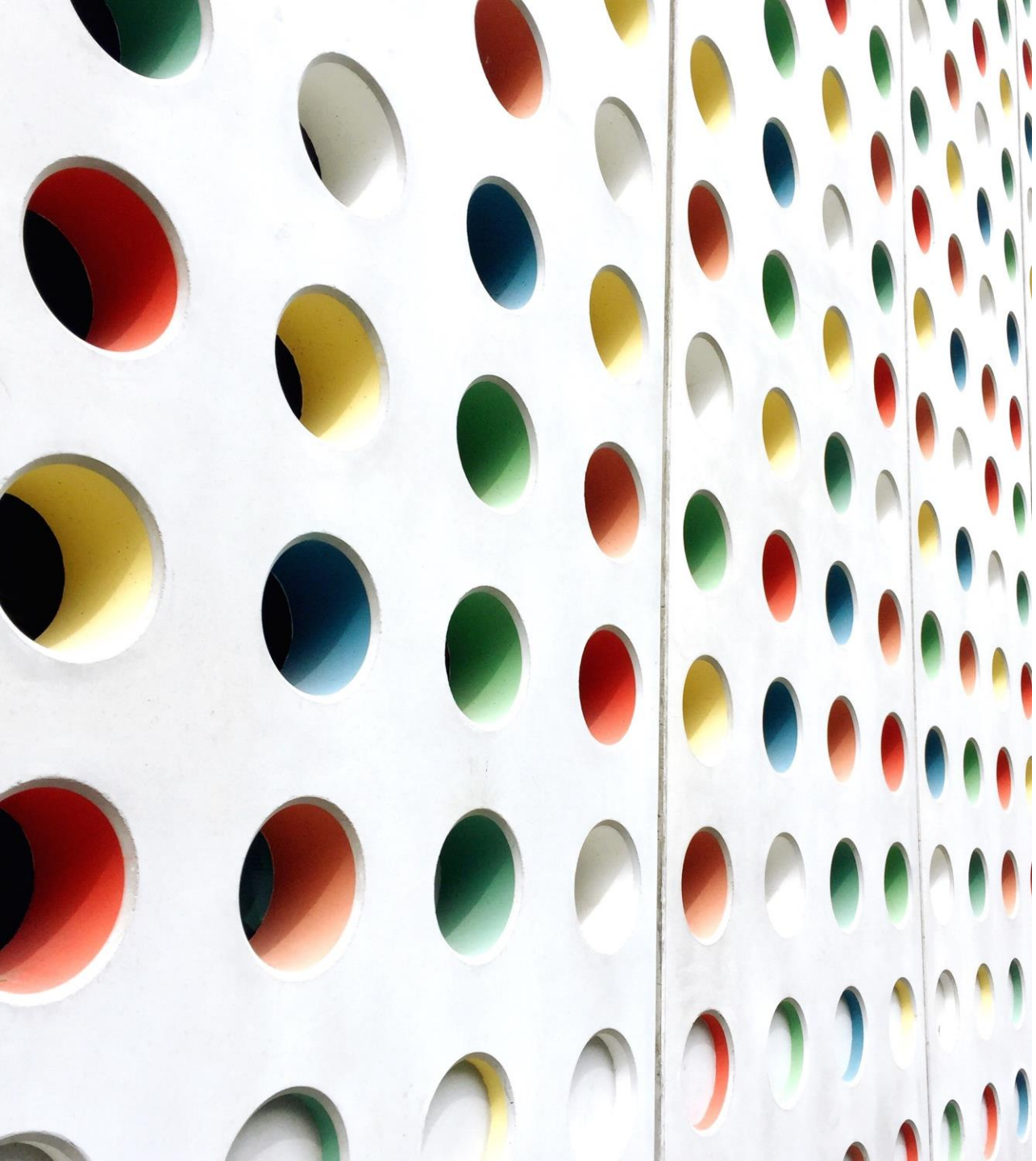


# Database Analytics

**BM 4419**

**BUSINESS ANALYTICS**

# Outline of Database Analytics

By the end of the course, you will:

- Differentiate between a data set and a database and apply Excel range names in data files.
- Construct Excel tables and be able to sort and filter data.
- Apply the Pareto principle to analyze data.
- Use database functions to extract records.

# Difference between a data set and a database.

## Data Set

- Dataset is a collection of data organized for analysis. It supports business analytics activities and Decision-making task.

## Database

- An organized grouping of information within a specific structure that needs to be retrieved frequently. Databases are used for storing and managing data for various applications.

# Rational Database

| Owner_ID | Owner Name |
|---|---|
| 1 | Jim |
| 2 | Joan |

| Pet_ID | Pet_Name | Owner_ID |
|---|---|---|
| 1 | Fifi | 2 |
| 2 | Butch | 1 |
| 3 | Clover | 2 |
| 4 | Animal | 1 |
| 5 | Tank | 1 |

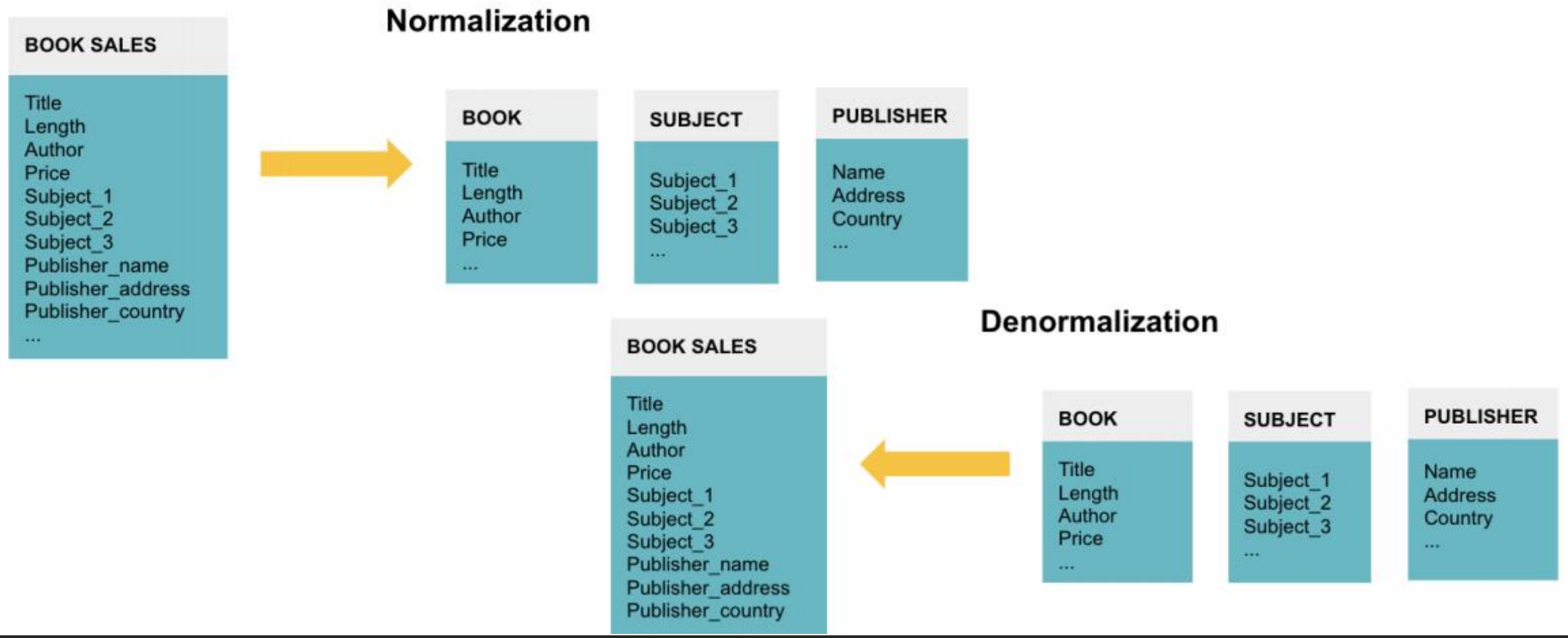
- A relational database organizes data into tables which can be linked—or related—based on data common to each.
- Table 1 contains information about pet owners
- Table II contains information about pets
- The tables are related by the column Owner_ID.
- By relating tables to one another, we can reduce data redundancy and improve database performance.

# DATA SET

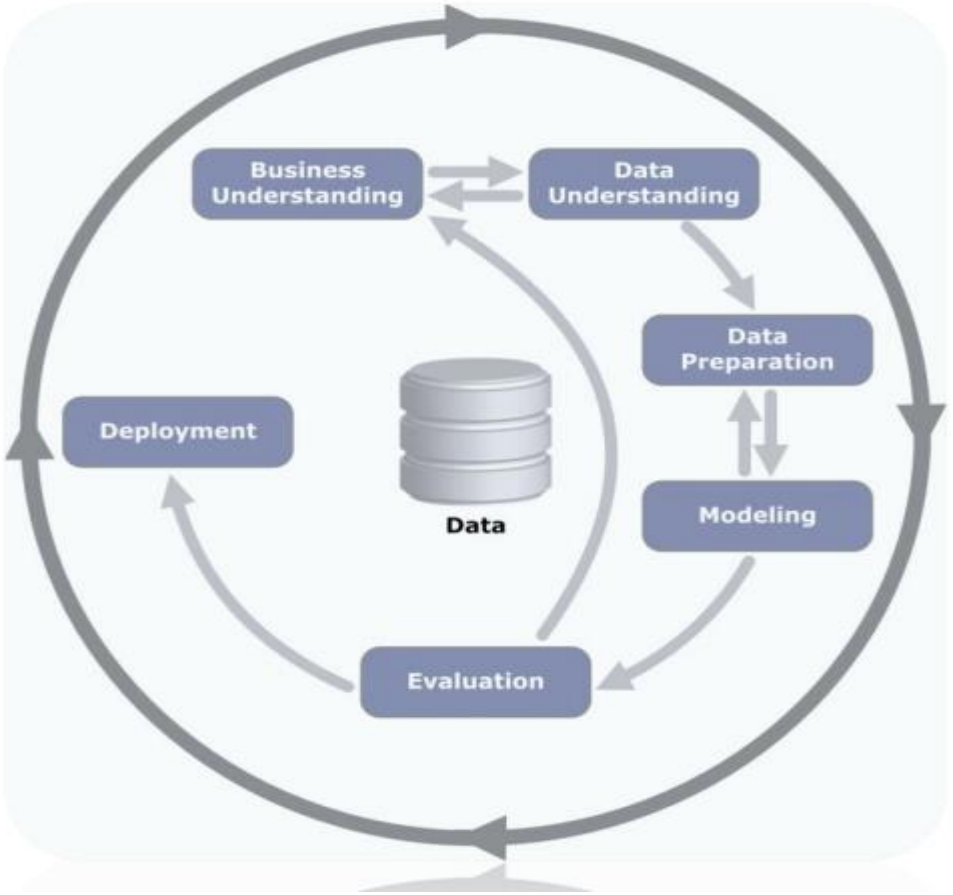
A data set is a subset of a database or a data warehouse.

- It is usually denormalized so that only one table is used.
- The creation of a data set may contain several steps, including appending or combining tables from source database tables, or simplifying some data expressions.
- Data sets may be made up of a representative sample of a larger set of data, or they may contain all observations relevant to a specific group.

# NORMALISED VS DENORMALISED

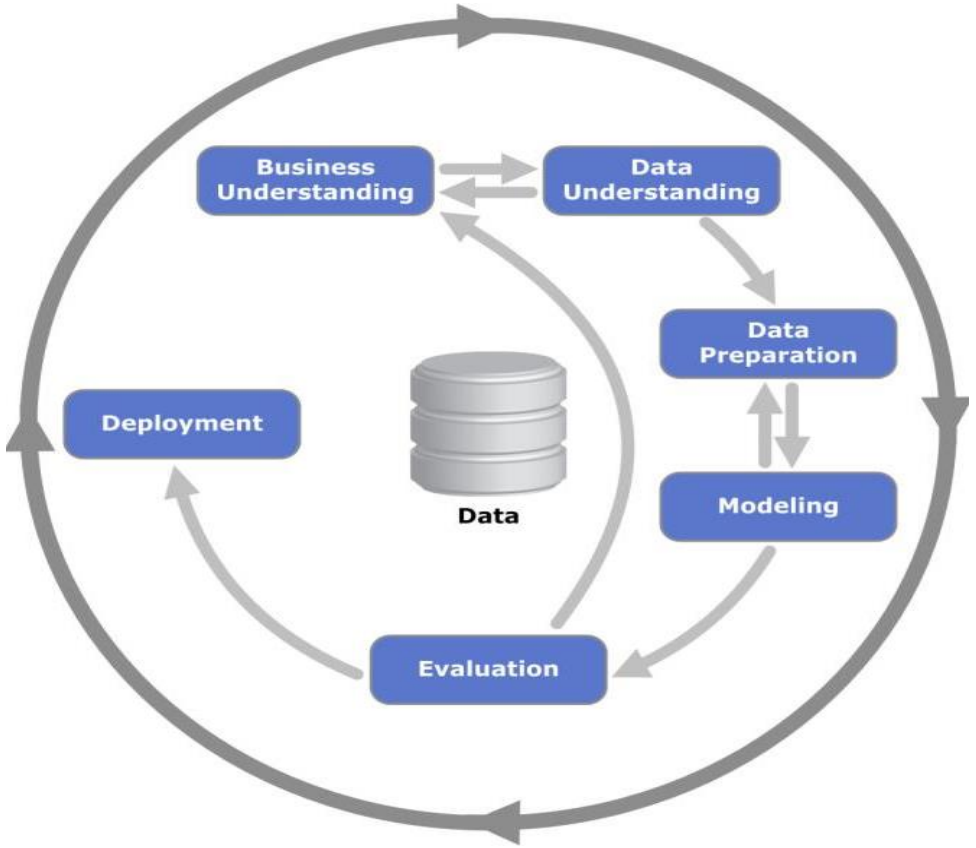


# Life Cycle of Data




- The life cycle of a data mining project consists of six phases.
- Not rigid process moving back and forth between each phase determines which phases has to be performed next.

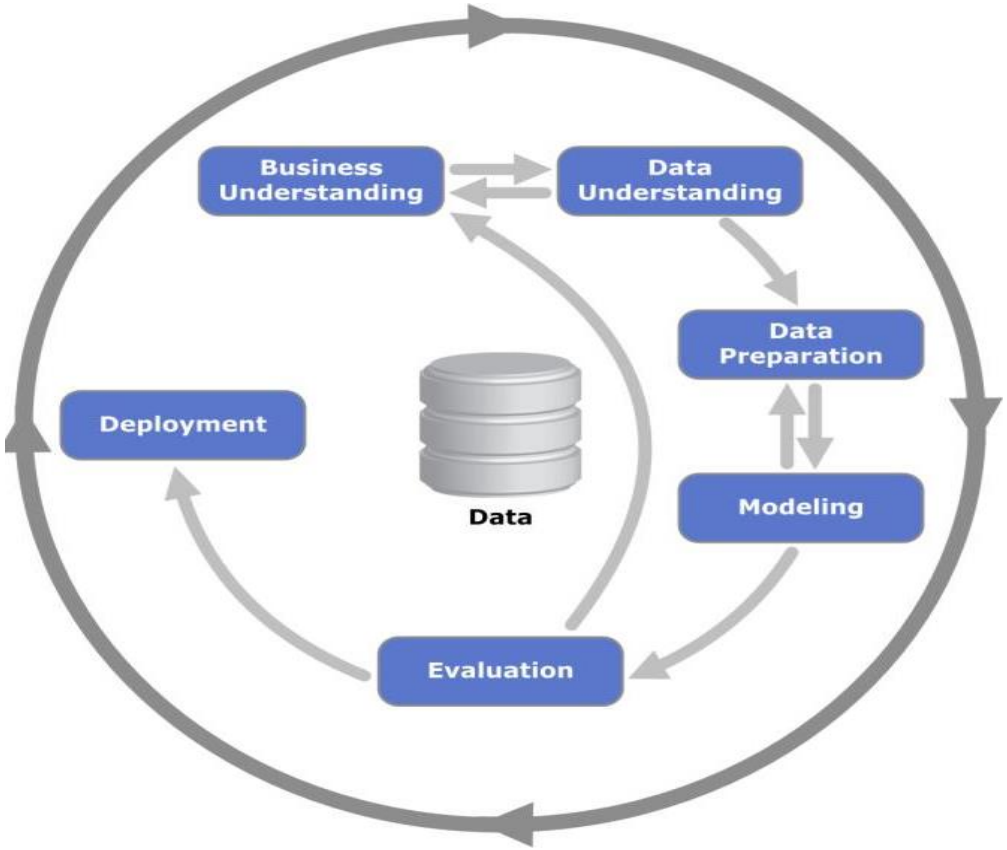
# Phase 1-Business Understanding

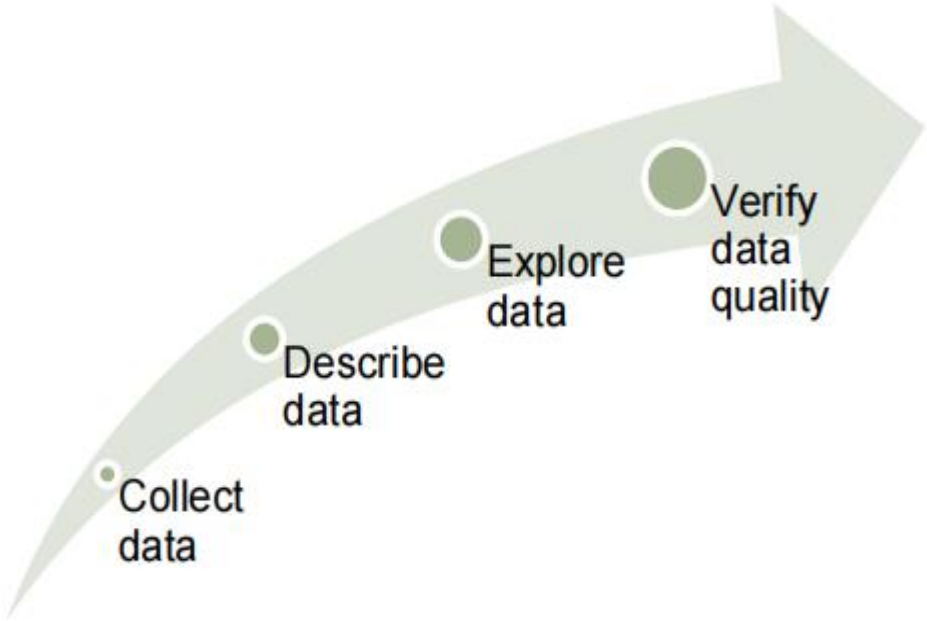



## 4 TASKS

1. Determine business objective
1. Assess situation
1. Determine data mining goals
1. Produce project plan

# Phase 2-Data Understanding



# Phase 2-Data Understanding

1. Collect initial data

· acquire the data listed in the project resources

· includes data loading if necessary, for data understanding

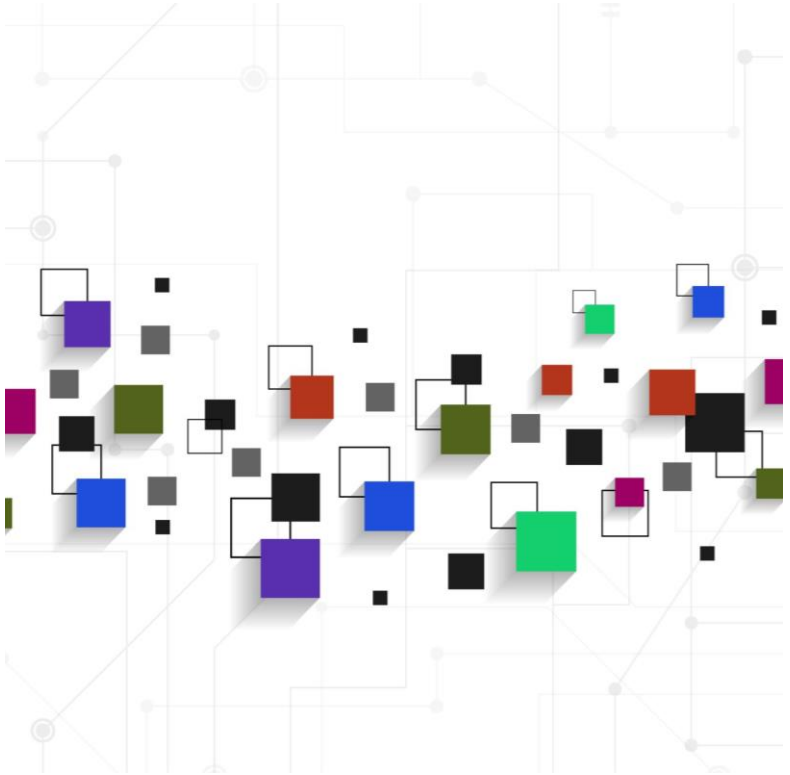
· possibly leads to initial data preparation steps

· Notes: if acquiring multiple data sources, integration is an additional issue, either here or in the later data preparation phase

1. Describe data

· examine the "gross" or "surface" properties of the acquired data

· report on the results



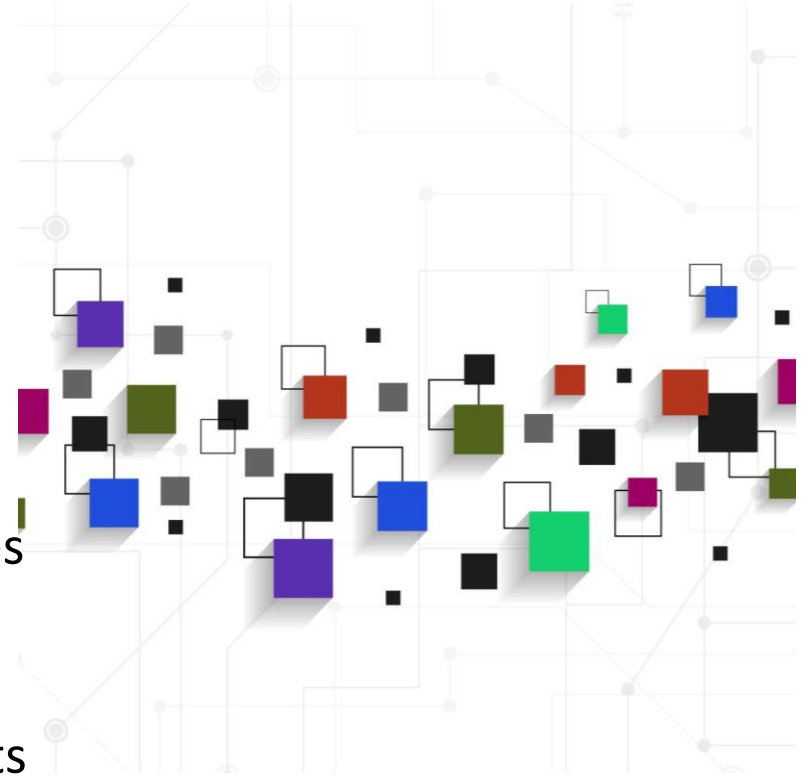
# Phase 2-Data Understanding

**3. Explore data**

- tackles the data mining questions, which can be addressed using

querying, visualization and reporting including:

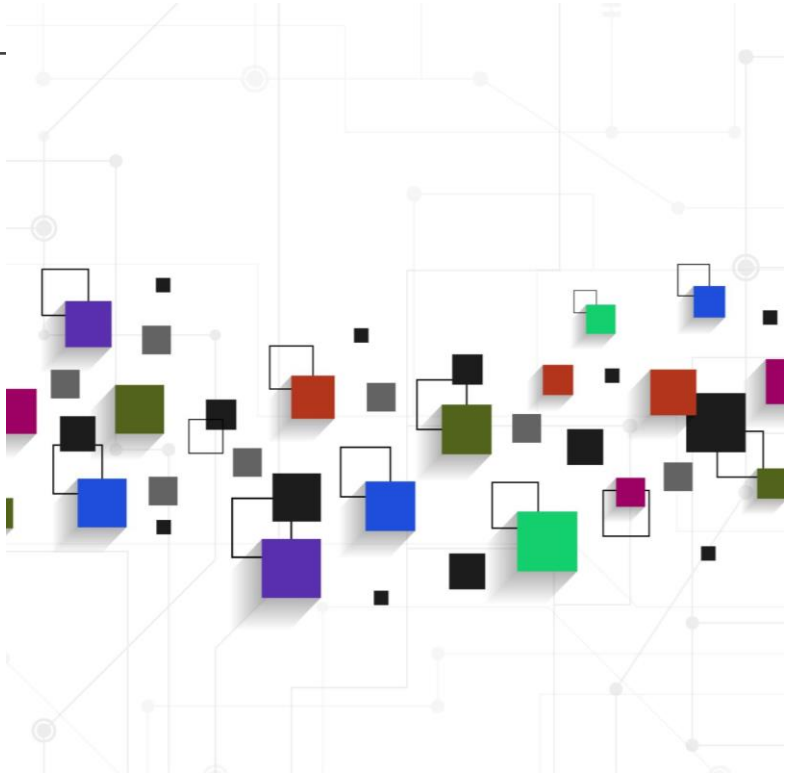
- distribution of key attributes, results of simple aggregations
- relations between pairs or small numbers of attributes
- properties of significant sub-populations, simple statistical analyses
- may address directly the data mining goals
- may contribute to or refine the data description and quality reports
- may feed into the transformation and other data preparation needed



# Phase 2-Data Understanding

**4. Verify data quality**

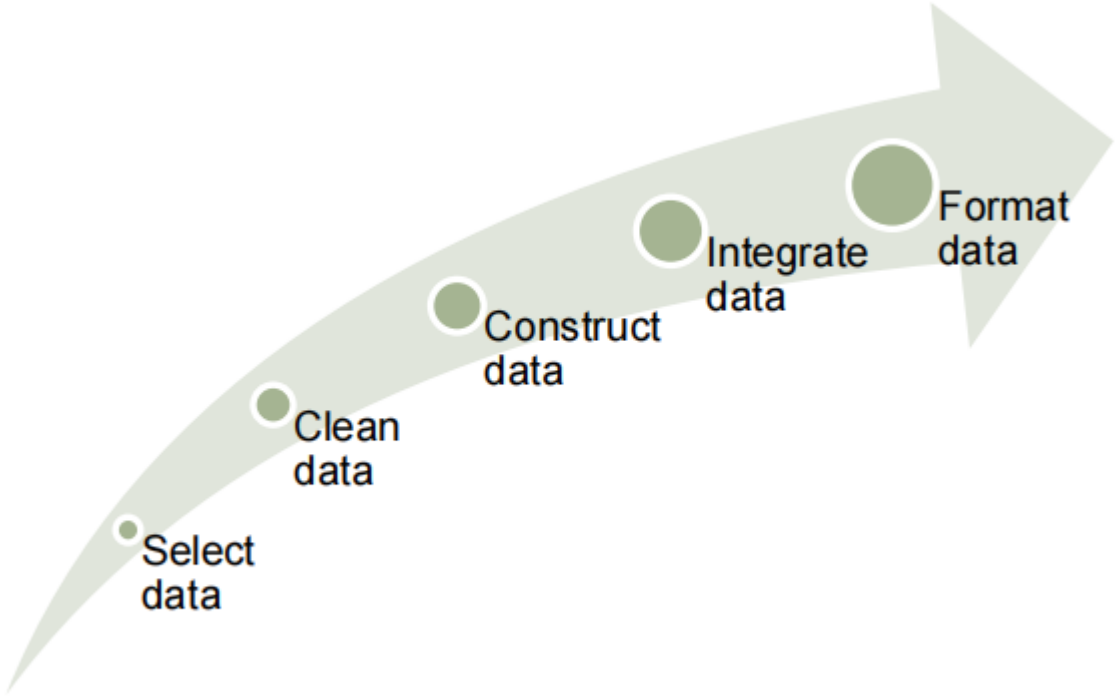
- examine the quality of the data, addressing questions such as:
- “Is the data complete?”,
- Are there missing values in the data?”



# Phase 3-Data Preparation







# Phase 3-Data Preparation

- 1. Select data
  - decide on the data to be used for analysis criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types covers selection of attributes as well as selection of records in a table
- 2. Clean data
  - raise the data quality to the level required by the selected analysis techniques may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modeling.

# Phase 3-Data Preparation

- 3. Construct data

constructive data preparation operations such as the production of derived attributes, entire new records or transformed values for existing attributes

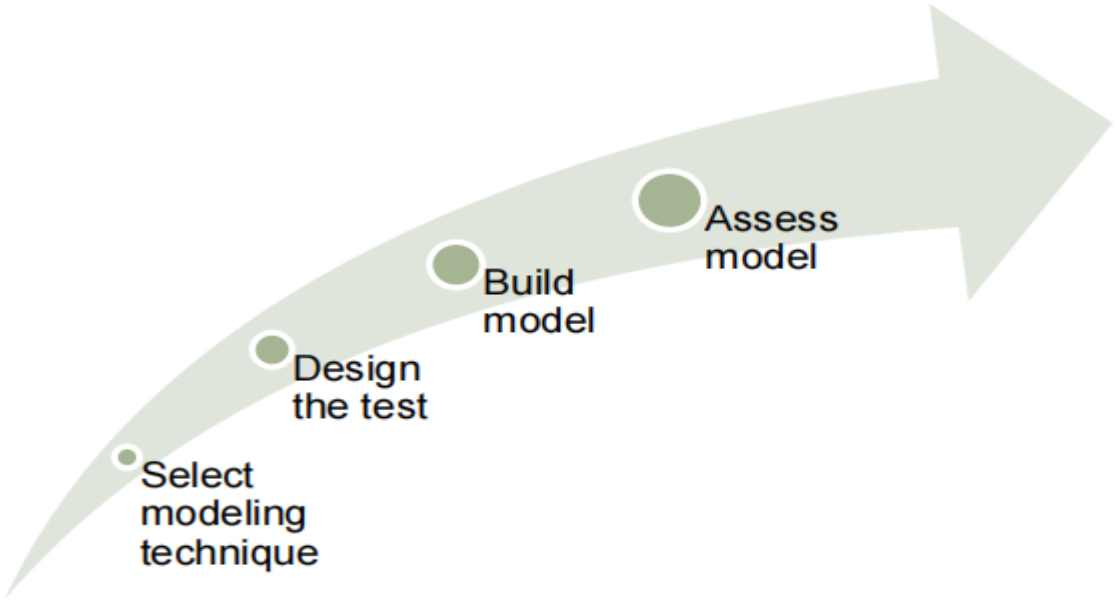
- 4. Integrate data

methods whereby information is combined from multiple tables or records to create new records or values

- 5. Format data

formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool

# Phrase 4-**Modelling**

# Phrase 4- Modelling

- ❑ Select the modeling technique (based upon the data mining objective)
- ❑ Build model (Parameter settings)
- ❑ Assess model (rank the models)

Various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
Some techniques have specific requirements on the form of data.
Therefore, stepping back to the data preparation phase is often necessary.

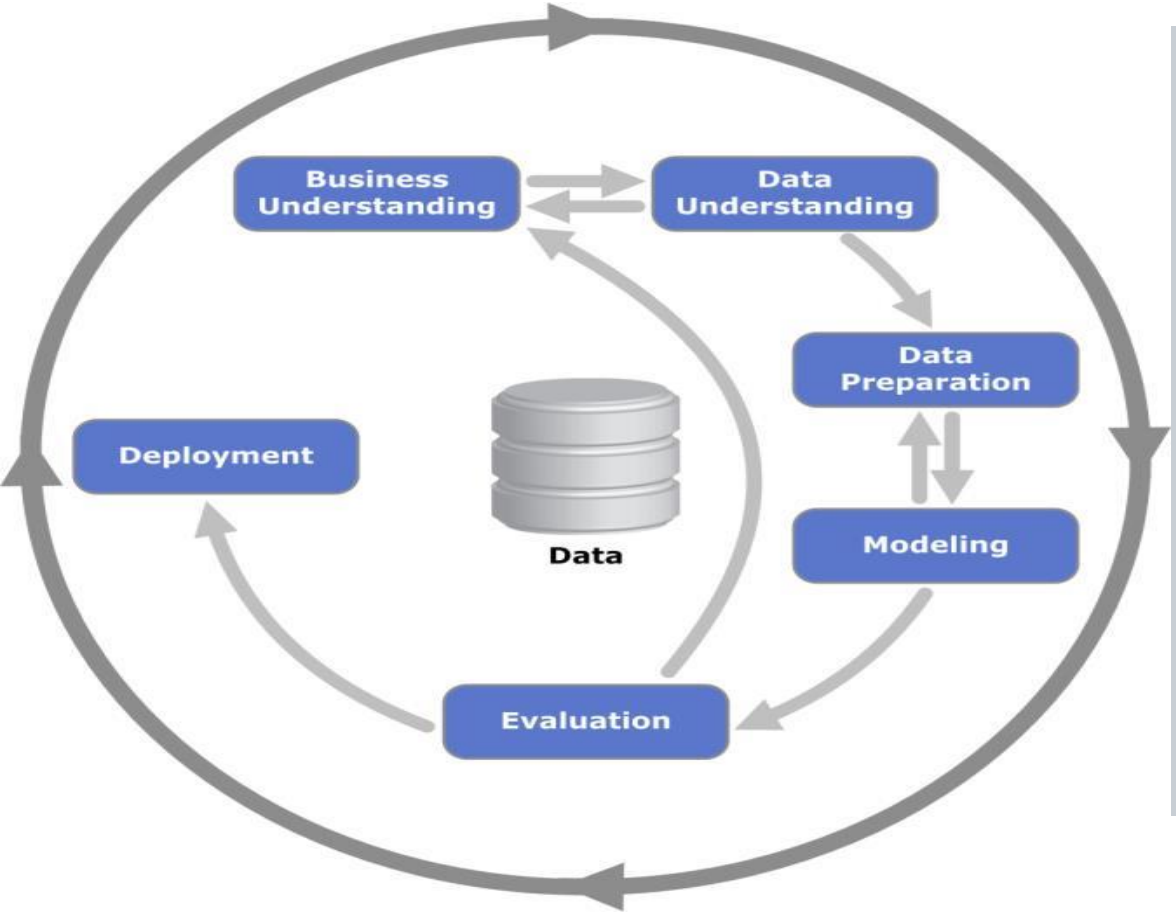
# Phrase 4- Modelling

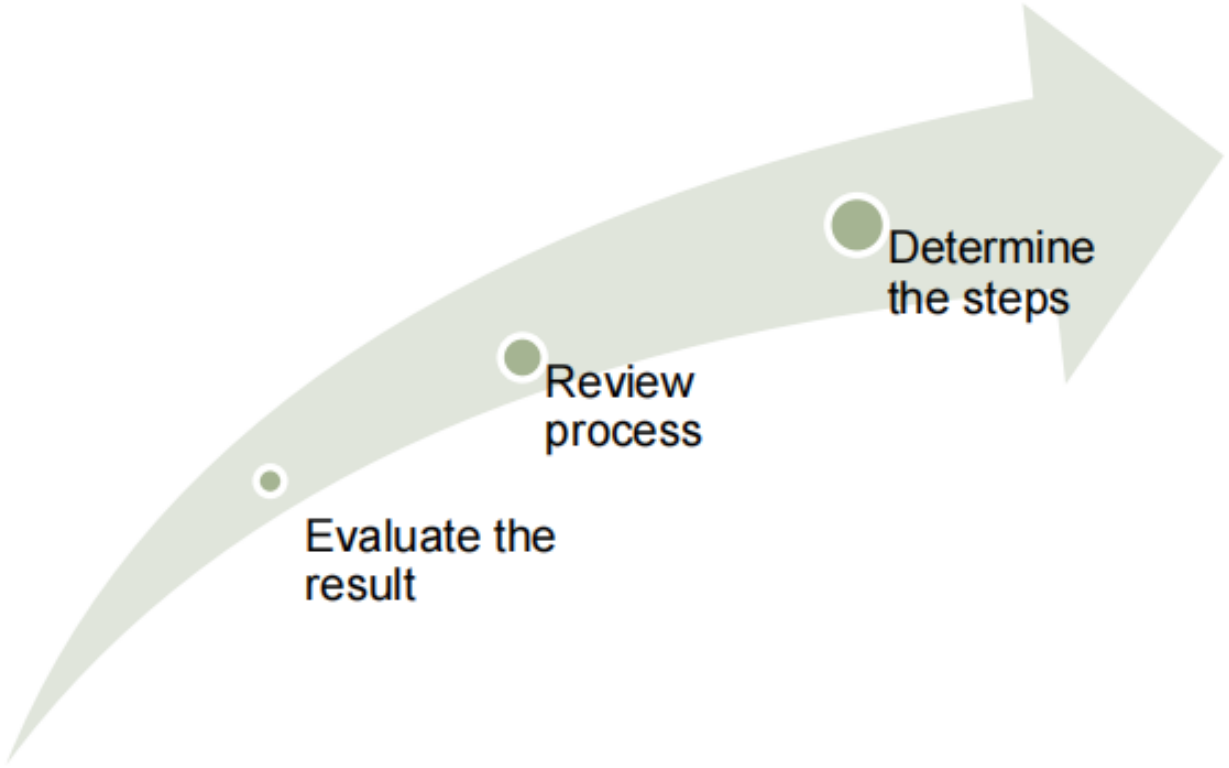
- interprets the models according to his domain knowledge, the data mining

success criteria and the desired test design

- judges the success of the application of modeling and discovery techniques more technically
- contacts business analysts and domain experts later in order to discuss the data mining results in the business context
- only consider models whereas the evaluation phase also takes into account all other results that were produced in the course of the project

# Phrase 5-Evaluation



# Phrase 5-Evaluation

**1. Evaluate results**

**• assesses the degree to which the model meets the business objectives• seeks to determine if there is some business reason why this model is deficient**

**• test the model(s) on test applications in the real application if time and budget constraints permit**

**• also assesses other data mining results generated**

**• unveil additional challenges, information or hints for future directions**

# Phrase 5-Evaluation

**2. Review process**

**• do a more thorough review of the data mining engagement in**

**order to determine if there is any important factor or task that has somehow been overlooked**

**• review the quality assurance issues**

**• i.e. “Did we correctly build the model?”**

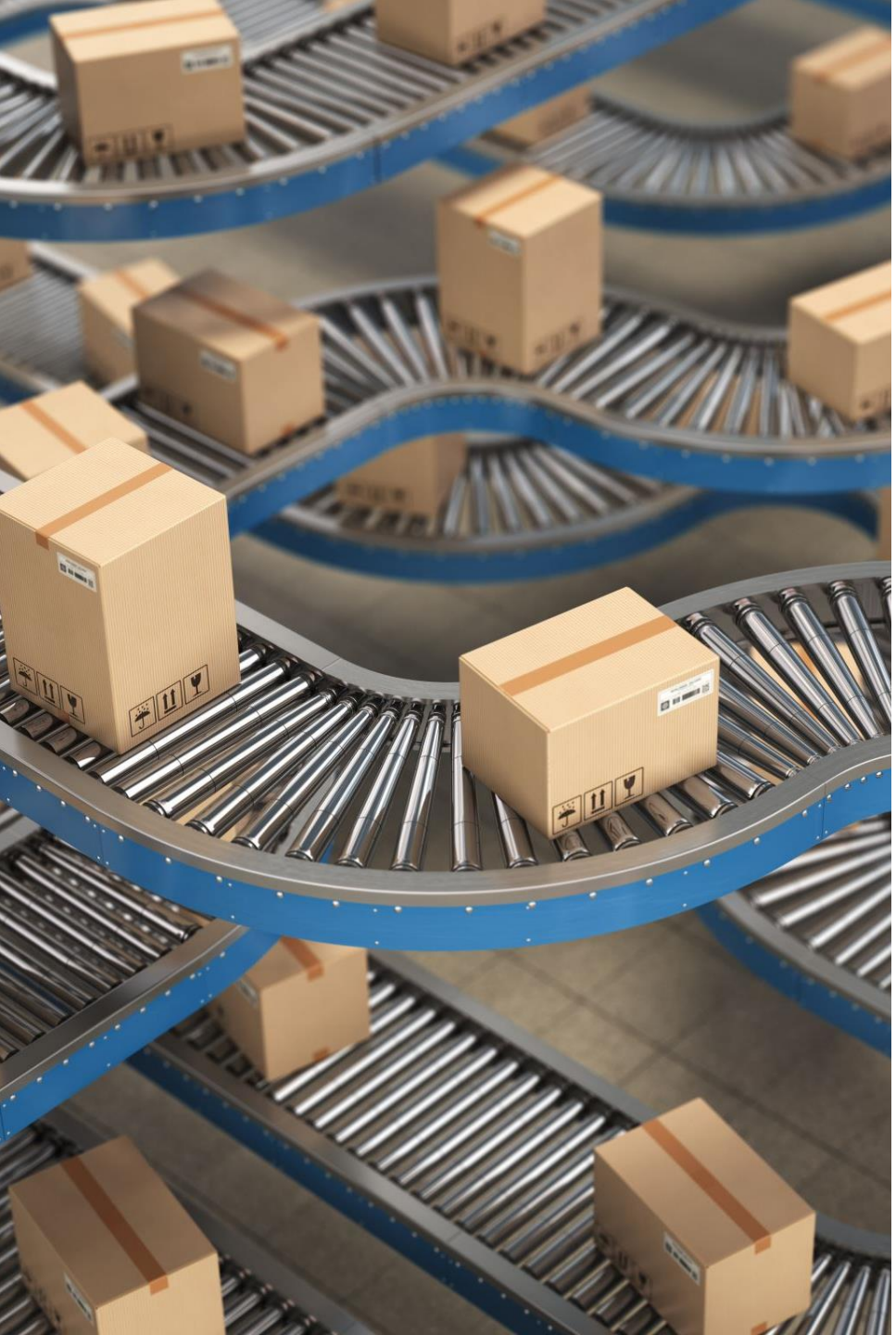
# Phrase 5-Evaluation

**3. Determine next steps**

**• decides how to proceed at this stage**

**• decides whether to finish the project and move on to deployment if appropriate or whether to initiate further iterations or set up new data mining projects**

**• include analyses of remaining resources and budget that influences the decisions**



# Phrase 6-Deployment

## 1. Plan deployment

- in order to deploy the data mining result(s) into the business, takes the evaluation results and concludes a strategy for deployment
- document the procedure for later deployment

## 2. Plan monitoring and maintenance

- important if the data mining results become part of the day-to-day business and its environment
- helps to avoid unnecessarily long periods of incorrect usage of data mining results
- needs a detailed-on monitoring process
- takes into account the specific type of deployment

# Phrase 6-Deployment

3. Produce final report

- the project leader and his team write up a final report may be only a summary of the project and its experiences
- may be a final and comprehensive presentation of the data mining result(s)

4. Review project

- assess what went right and what went wrong, what was done well and what needs to be improved

# Construct Excel tables and be able to sort and filter data.

can quickly **reorganize** a worksheet by **sorting** your data.

Eg: organize a list of contact information by last name.

**Filtering data** can be used to **narrow down** the data in your worksheet, allowing you to view only the information you need.

*HK*

You can sort your Excel data on one column or multiple columns.

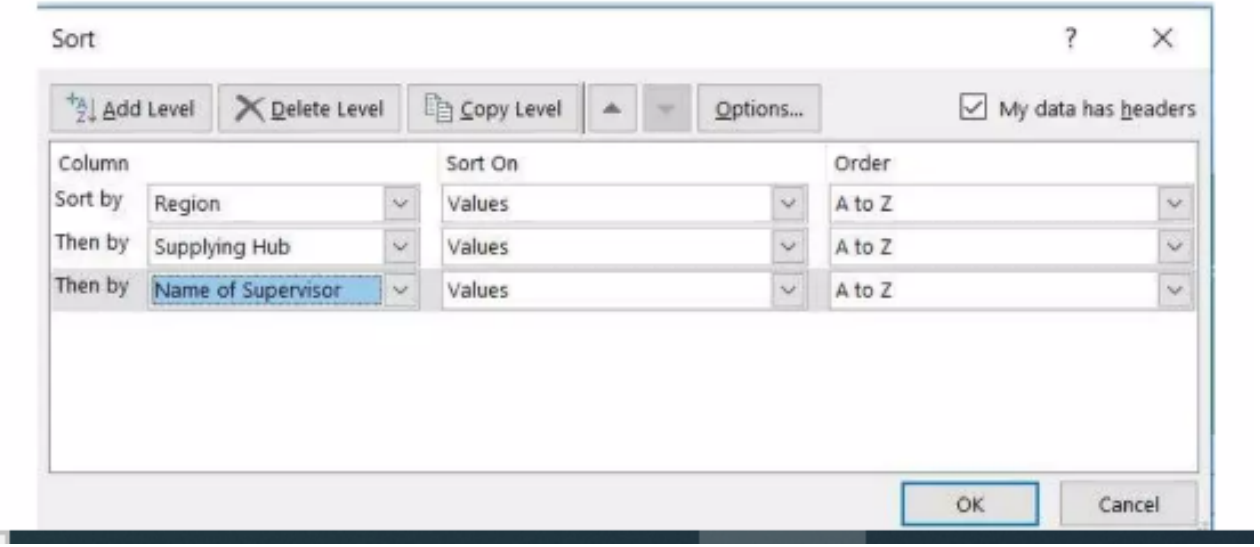
**Data> sort data > select the column**

**Or**

**home > Sort & Filter > Sort**

Lets Sort ss data

- By region,
- by suppling hub
- Name of supervisor

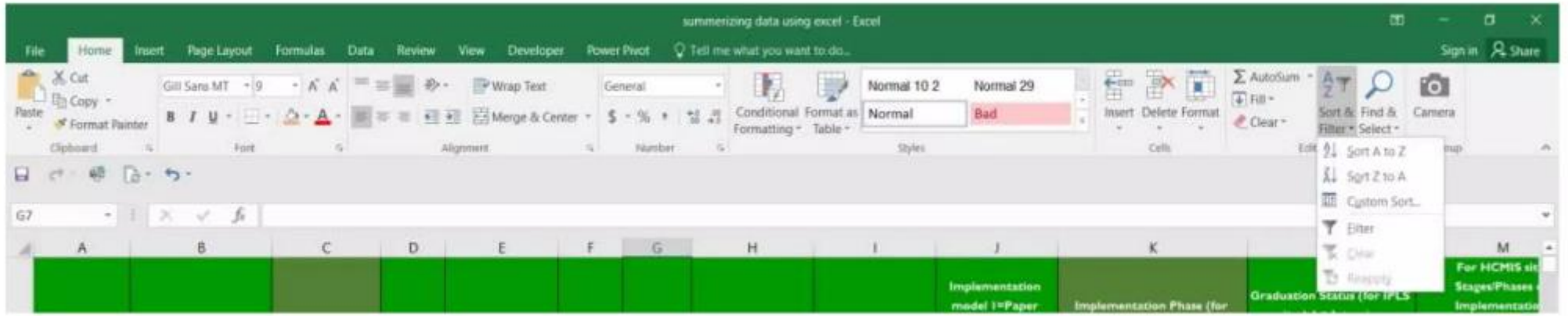


Filter your data if you want to display records that meet certain criteria.

**Data> Filter data**

**Or**

**home > Sort & Filter > Filter**



# Try the Practice

- excel_filtering_practiceWEEK2
- excel_sorting_practiceWEEK2

*May2024 _HK*

# Analyze Data by Principle of Pareto

- **Pareto Chart is a very powerful tool for showing the relative importance of problems.**
- **individual values are represented in descending order by bars, and the cumulative total of the sample is represented by the curved line.**
- **80/20 rule applies**

# How do you interpret a Pareto chart of standardized effects?

- excel_Pareto_practiceWEEK2

*May2024 _HK*