# W1

**Introduction to Data Science**

1. **What is data science?**
   * Data science is a multidisciplinary domain that allows you to turn raw data into understanding, insight, and knowledge. It typically involves solving analytically complex problems using machine learning algorithms.
2. **What compares with traditional algorithms; how is ML different?**
   * Machine learning (ML) algorithms are data-driven, whereas traditional algorithms are not. In ML, the data determines the algorithm's response. For example, in an application to detect faces, a traditional algorithm would define what a face is, while an ML algorithm learns by examples through images grouped into face and non-face categories.
3. **List four Vs of big data:**
   * The four Vs of big data are Volume, Velocity, Variety, and Veracity.
4. **What is the size of small data? Size of big data?**
   * Small data: Hundreds of megabytes (Mb) to 1-2 gigabytes (Gb).
   * Large data: 10-100 gigabytes (Gb) or terabytes (Tb).

**Why are we focusing on small data first?**

* + Focusing on small data is crucial because one can't tackle big data without experience with small data. Additionally, a big data problem could be a small data problem in disguise. The data needed to answer a specific question might be small, and the problem may be decomposable into multiple small data problems.

1. **Why use R (4 reasons)?**
   * R is a great place to start a data science journey.
   * R is designed from the ground up to support data science.
   * R is both a programming language and an interactive environment for data science.
   * R helps data scientists focus on problems, supporting interaction between their brain and the computer.
2. **What is rectangular data?**
   * Rectangular data are collections of values, each associated with a variable and an observation. Rows represent observations, and columns represent variables.

**What is non-rectangular data?**

* + Non-rectangular data includes images, sounds, trees, graphs, sequences, text, videos, etc.

1. **What is a Hypothesis?**
   * A hypothesis is an educated guess or an intelligent guess. In science, it's a starting point for investigation.

**Hypothesis generation?**

* + Hypothesis generation involves looking at data with subject knowledge to generate interesting hypotheses that explain why the data behaves the way it does.

**Hypothesis confirmation?**

* + Hypothesis confirmation involves creating a precise mathematical model to generate falsifiable predictions. This requires preregistering an analysis plan and adhering to it without deviation, even after seeing the data. Warning: it is not data exploration, exploration is without hypothesis.

1. **Why focus on hypothesis generation?**
   * Hypothesis generation is essential for understanding business problems by:
     1. Inferring various factors affecting the target variable.
     2. Identifying major factors responsible for solving the problem.
     3. Guiding data collection to convert business problems into data science problems.
     4. Enhancing domain knowledge and helping approach the problem in a structured manner.

**Data Science Life Cycle**

1. **List 6 stages and questions for each stage:**
   * **Defining the goal:**
     + Why do sponsors want the project?
     + What do they lack or need?
     + What are the resources required?
     + How will the project results be deployed?
   * **Identifying and conditioning data:**
     + What data is available?
     + Will it help solve the problem?
     + Is it enough, and is the quality good?
   * **Modelling:**
     + What insights can be extracted from the data?
     + Is the model representation the best one for the data?
   * **Model Evaluation and Critique:**
     + Is the model accurate and generalizable?
     + Does it outperform current methods or estimates?
   * **Presentation and Documentation:**
     + How to present the results to different stakeholders?
     + How to document the model for various audiences?
   * **Model Deployment and Maintenance:**
     + How to ensure the model runs smoothly and makes sound decisions?
     + How to update the model when the environment changes?
2. **For each stage, why do we need them?**
   * Each stage is crucial for guiding the data science project from understanding the problem to deploying the model effectively, ensuring it meets its goals, remains useful, and adapts to changes over time.

**Introduction to R**

1. **Difference between arrow (<-) and equal (=)?**
   * (<-) is used for assigning values.
   * = is used in function parameter passing.
2. **What are the naming restrictions (4)?**
   * Object names cannot contain special symbols like !, +, -, #.
   * A dot (.) and underscore (\_) are allowed; names can start with a dot.
   * Object names can contain numbers but cannot start with a number.
   * R is case-sensitive (X and x are different objects).
3. **Definitions:**
   * **Vectors:**
     + An ordered collection of data of the same type.
   * **Matrices:**
     + A rectangular table of data of the same type.
   * **Arrays:**
     + A higher-dimensional matrix, turning a vector into a multi-way dimensional matrix.
     + Difference from a vector: An array can have multiple dimensions, whereas a vector is one-dimensional.
   * **List:**
     + An ordered collection of data of arbitrary types.
     + Difference from a vector: A list can contain elements of different types, whereas a vector must contain elements of the same type.
   * **Data Frames:**
     + A rectangular table with rows and columns; data within each column has the same type, but different columns may have different types. It represents a typical data table like a spreadsheet.