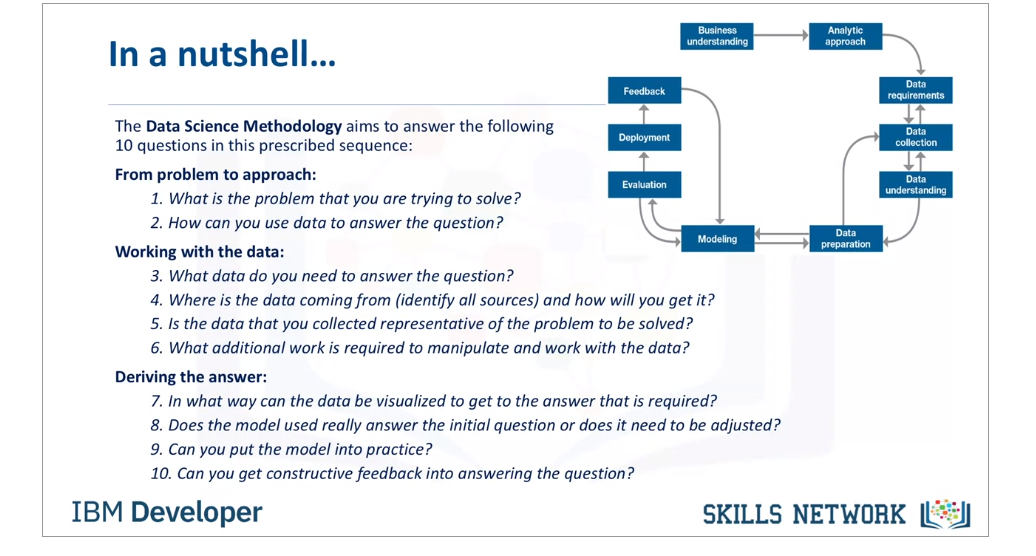
**Data Science Methodologies**

**Week 1**

**Notes:**



A picture containing text, screenshot, web page, font

Description automatically generated

A picture containing text, screenshot, software, computer icon

Description automatically generated

**1-Business Understanding:**

1. Have a strong understanding of what needs to be done.
2. Having a set of clearly defined questions regarding the business statement allows you to establish clarification between the person who has a business problem and the one who is going to provide the solutions to the problem.
3. Breakdown the goal of the business statement into objectives is very crucial part of the data science because it helps one to select the appropriate analytical approach which helps in what kind of data needs to be selected to solve the problem.

**2-Why Analytical Approach?**  
  
Once the business problem has been clearly stated, the data scientist can define the analytic approach to solve the problem. This step entails expressing the problem in the context of statistical and machine-learning techniques, so that the entity or stakeholders with the problem can identify the most suitable techniques for the desired outcome.

**Decision Tree:**

* Decision trees are built using recursive partitioning to classify the data.
* When partitioning the data, decision trees use the most predictive feature (ingredient in this case) to split the data.
* Predictiveness is based on decrease in entropy - gain in information, or impurity.

**A tree stops growing at a node when:**

* Pure or nearly pure.
* No remaining variables on which to further subset the data.
* The tree has grown to a preselected size limit.

**3-Data Requirements (Data Collection Typically)**

Note: Data scientists, essentially, explore the data to:

* understand its content,
* assess its quality,
* discover any interesting preliminary insights, and,
* determine whether additional data is necessary to fill any gaps in the data.

**4- Data Collection:**

In this phase, we will decide what data we need to collect in order to answer business questions and how to organize it so that it is useful. we might use our business task to decide:

* What metrics to measure
* Locate data in your database
* Create security measures to protect that data

Data is a collection of facts that help us provide informed-decision making to the stakeholders. An effective data-driven result can help a business bring a unique change in the business. As the day passes by, the data is gradually and rapidly emerging in a huge amount. A data can be in any form; **structured** (tabular data, rectangular data, speadsheets or relational databases) or **unstructured** (photos, videos, email, social posts, songs, movies etc.). Our **Fitbit Fitness Data** is structured data in spreasheets.

Let's go in the detailed definition of Prepare Phase------

▣ **4.1 Understanding the data collection and why it is collected :**

Before analyzing the data, we need to define what kind of data we need and what are the sources available for the data collection. In accordance with meeting the business tasks's deliverables and achieve the goal of business objective, we need to collect the exact data needed for analysis. In our case, we are using the **Fitbit Fitness tracker data provided by open data source**.

▣ **4.2 Understanding types of data sources and database :**

In order to collect the required data, we need to understand how the data will be collected and from which sources. It could be either internal source or external source.

* **Internal Data Source** (Primary Data Source)
  + First-Party Data Source (Data collected by an individual or group using their own resources or data collected in their own business system)
* **External Data Source** (Secondary Data Source)

**Garamond**

* + Second-Party Data Source (Data collected by a group directly from its audience and then sold it) **Ex; survry monkeys**
  + Third-Party Data Source (Data collected from outside sources who did not collect it directly) **Ex; google platform**
  + Open Data Source (Data provided by government institutions and asks for permission to collect their data, **Ex; WHO**)
  + Public Data Source (Data available for anyone on the platforms such as **Google Cloud, Kaggle, Github etc.)**

Databases are a collection of data stored in the computer system that follows a range of certain business rules that we need to follow while collecting the data. Metadata is a great way of observing the information about other data. It works as glossary for other data. It helps us intepret the contents of the data within the databases, combines the data with other data and evaluate the quality of data. It can be three types of metadata; Descriptive, Structured, and Administrative

Available data can be collected through many ways such as API, Databases, Webscrapping.

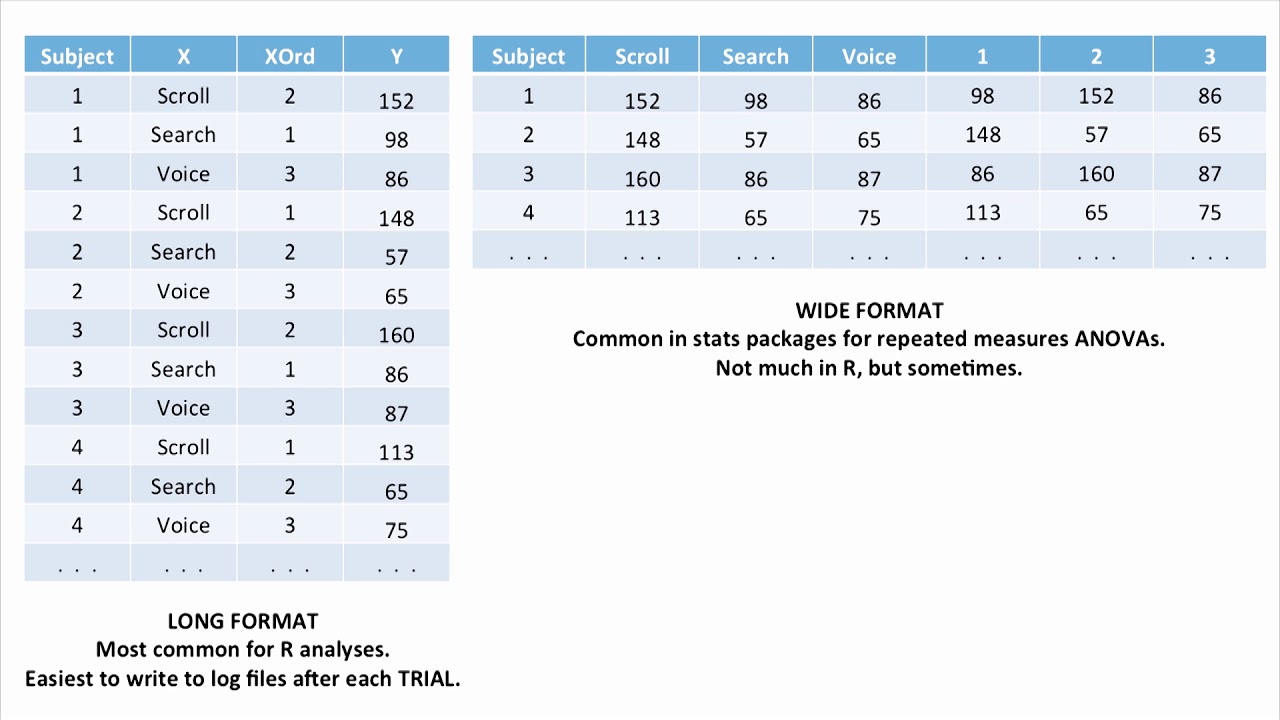
In our case study, we will collect the data from **Fitbit Fitbit Tracker Data (a open source dataset)** given below and we can verify that the metadata of our dataset we can confirm it is open-source. The owner has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

⋇ **Our Data Source**

[FitBit Fitness Tracker Data](https://www.kaggle.com/arashnic/fitbit) (CC0: Public Domain, dataset made available through [Mobius](https://www.kaggle.com/arashnic)): This Kaggle data set contains personal fitness tracker from **30** fitbit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users’ habits.

▣ **4.3 Ensuring the data modeling and type of data :**

Data modeling is very useful and it is based on business rules that allow us to retrieve the needed information in an organized way. It provides connections between a wide range of databases and it provides the process of creating diagrams that visually represent how the data is organized in a defined structured. Understaing the data modeling enlightens the way of data collection. When analyzing the data, data can come with two basic types; **Long Data** & **Wide Data**. According to the business task, we can use any of them. In this project, it is a **Long Data**.



▣ **4.4 Defining the data bias and fairness :**

Data bias is a type of error that systematically skews the results in a certain direction. It occurs when data points at a certain preference in a favour of audience or against audience, a person or thing. It must be showing the equal amount of the people no matter what. It doesn't need to reinforce the partiality. It can be categorised into three types: **Observer Bias**, **Interpretation Bias**, and **Confirmation Bias**.

Once the data represent the whole population by all means, it brings the **fairness** to the analysis. However, for this data **(Fitbit data)**, due to the limitation of size (30 users) and not having any demographic information we could encounter a sampling bias. We are not sure if the sample is representative of the population as a whole. Another problem we would encounter is that the dataset is not current and also the time limitation of the survey (2 months long). That is why we will give our case study an operational approach.

▣ **4.5 Defining the clean and bad data source through ROCCC :**

A data source can be very complex and frustrating when it is not organized well. There are many data sources that are not reliable, current, cited and original. To identify the good data source, we need to follow a certain process given below and based on our Fitbit Fitness Tracker Data, we can define whether data is collected from clean or bad source;

* **ROCCC for Fitbit Fitness Data:**
  + **R** : Realiable | biased data, not sure if the data is vetted
  + **O** : Original | yes, can locate the original public data
  + **C** : Comprehensive | no, many informations are missing such as age, weight, height, gender, etc also missing the completeness and accuracy.
  + **C** : Current | no, this is a historical data source (04/12/2016 - 05/12/2016)
  + **C** : Cited | yes

▣ **4.6 Data Ethics, Data Privacy, Data Security :**

The most important part of data collection is ensure the data ethics, data security and data privacy. When collecting the data, keep in mind that data ethics involves the ownership, transaction, transparency, consent, currency, privacy, and openness. Preserving one's data has to be first priority when data trasaction occurs. This will bring the transparency between the data collector and data giver. It will bring smooth process with no hurdles in the future. Securing data from anyone outside is also major step for the company. These days, it is a great concern to secure data because it is the key aspect of a running-business. It can be done through introducing **Encryptions** and **Tokenization**. Once the data is secured, on one from outside can interfere. Since **Fitbit Fitness Tracker Data** is open-source data, it means it is not encrypted or tokenized.

▣ **4.7 Data Organization & Verification :**

Organizing the data with predefined naming conventions, archiving order files, and foldering help us organize the data in a sophisticated way so that collected data can be easily communicated whenever it is needed. Predetermined guidelines that describe date, content or version of a file in its name is always set with a company's business system rules.

Each dataset represents different quantitative data tracked by Fitbit. The data is considered long since each row is one time point per subject, so each subject will have data in multiple rows.Every user has a unique ID and different rows since data is tracked by day and time.

Due to the small size of sample, I created pivot tables, sorted and filtered. I was able to verify attributes and observations of each table and relations between tables. Counted sample size (users) of each table and verified time length of analysis - 31 days.

### ****4.8 Our collected data related to business tasks and importing the datasets here :****

When collecting this data [FitBit Fitness Tracker Data](https://www.kaggle.com/arashnic/fitbit) (CC0: Public Domain, dataset made available through [Mobius](https://www.kaggle.com/arashnic)), We figure out that there are data limitations. To figure out the limitations and ROCCC analysis, we need to upload the files;

While collecting the data, we come up with data limitations given below;

**Data Limitations**

* The **30 users** sample size may greatly underrepresent **Fitbit** users population.
* More importantly, the data is from **2016** which may not represent current trends of using smart fitness devices.
* Besides, Bellabeat products are for woman while Fitbit dataset has not specify the gender of the respondents.
* Weightlog (8 unique ids), heartrate(14 unique ids), and sleepday (24 unique ids) dataset don't have equal amount of unique ids(33) unlike rest of the files do.

Thus, these limitations will impact how should Bellabeat apply the insights getting from this analysis to its products. For a more accurate and comprehensive data representation, Bellabeat should consider additional internal and, or external data sources.

**5-Data Understanding:**

**Feature Engineering**: Feature engineering is the process of using domain knowledge of the data to create features that make the machine learning algorithms work.

Feature Engineering is a critical part when machine learning tools are being applied to analyze the data.

**6-Data Preparation:**

**7-Data Modeling:**

Data modeling focuses on developing models that are either descriptive or predictive.

These models are based on the analytic approach that was taken, either statistically driven or machine learning driven.

The data scientist will use a training set for predictive modelling. A training set is a set of historical data in which the outcomes are already known.

**8-Data Evaluation:**

A model evaluation goes hand-in-hand with model building as such, the modeling and evaluation stages are done iteratively.

Model evaluation is performed during model development and before the model is deployed.

Evaluation allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request.

A picture containing text, screenshot, software, computer icon

Description automatically generated

**Model evaluation can have two main phases.**

The first is the **diagnostic** measures phase, which is used to ensure the model is working as intended. If the model is **a predictive model**, a decision tree can be used to evaluate if the answer the model can output is aligned to the initial design. It can be used to see where there are areas that require adjustments. If the model is a **descriptive model**, one in which relationships are being assessed, then a testing set with known outcomes can be applied, and the model can be refined as needed.

The second phase of evaluation that may be used is statistical significance testing. This type of evaluation can be applied to the model to ensure that the data is being properly handled and interpreted within the model. This is designed to avoid unnecessary second guessing when the answer is revealed.

**9-Model Deployment:**

**10-Feedback:**

A picture containing text, screenshot, software, computer icon

Description automatically generated