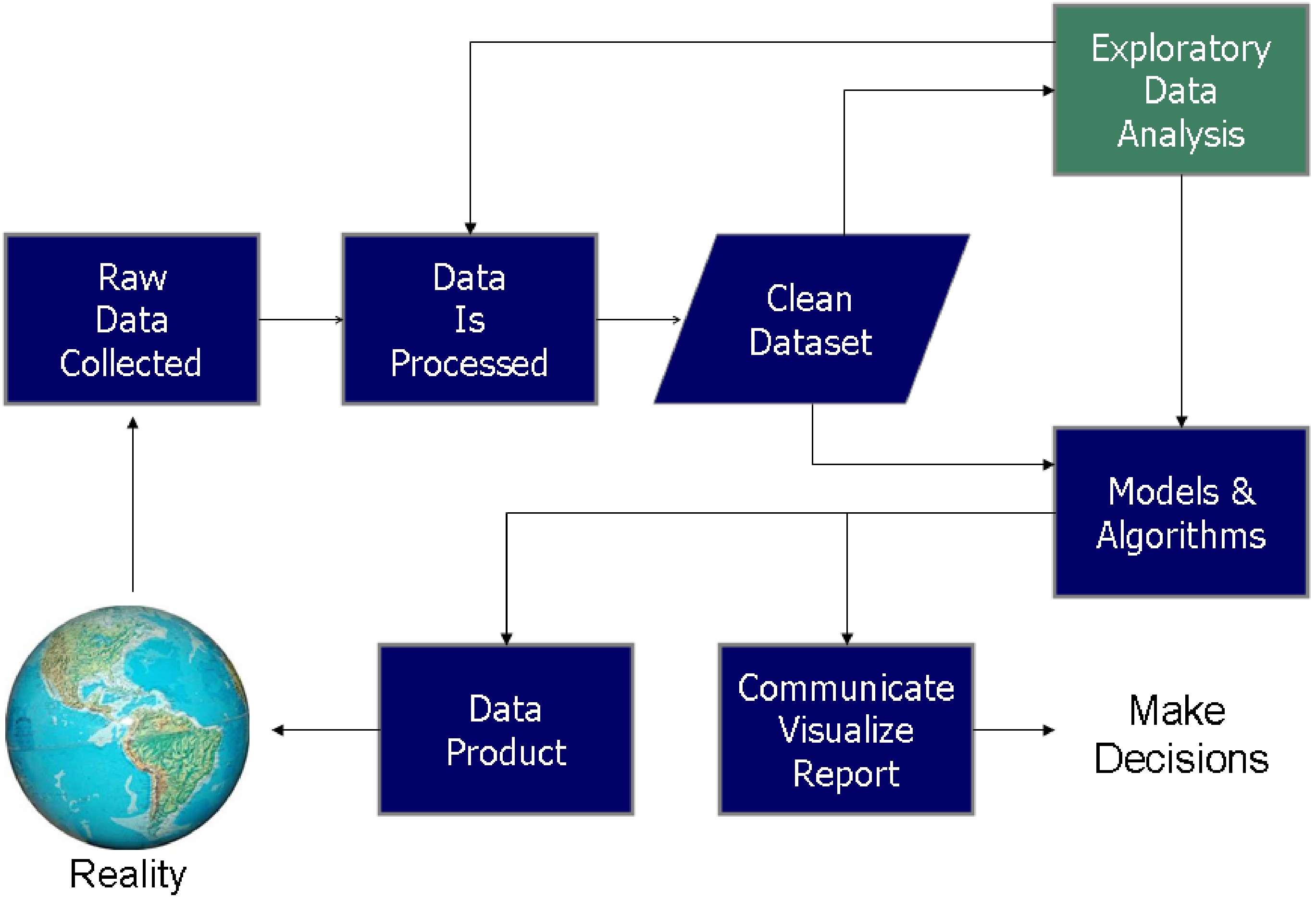
Machine Learning - Introduction

Exploratory Data Analysis

Data Science Process

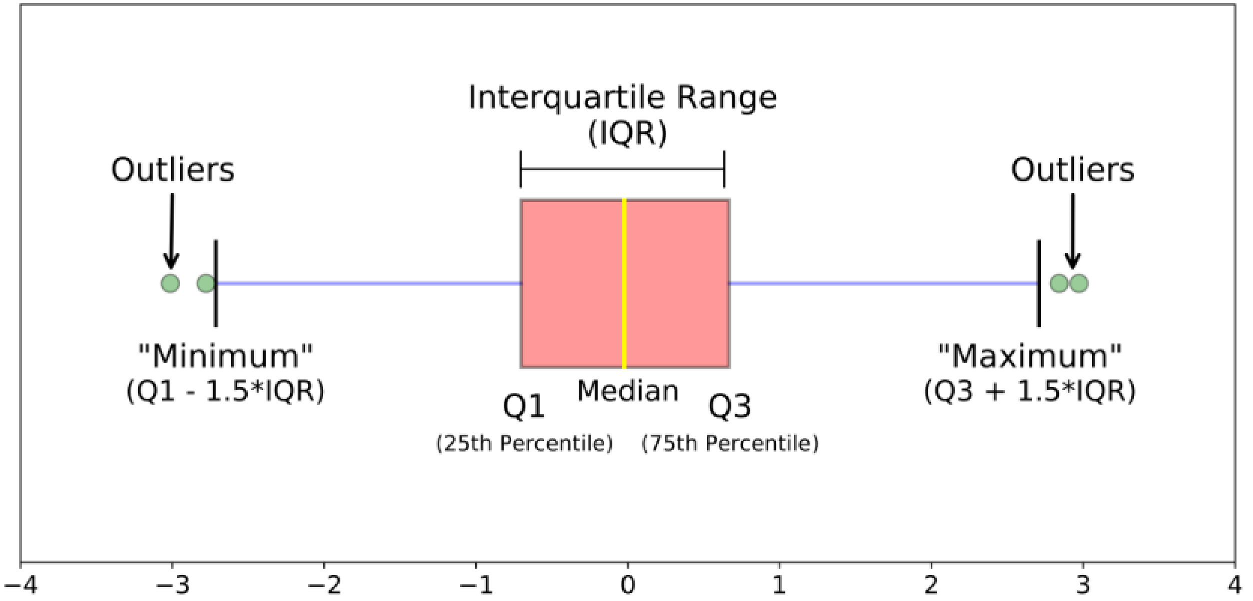


Source: Wikipedia

Exploratory Data Analysis

EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods

Boxplots



Different parts of a boxplot

https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Quartile, IQR, Whisker, Outlier

Boxplots are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”)

- **median (Q2/50th Percentile)**: the middle value of the dataset. (Be careful: a median separates the higher half from the lowe half of the dataset)
- **first quartile (Q1/25th Percentile)**: the middle number between the smallest number (not the “minimum”) and the median of the dataset.
- **third quartile (Q3/75th Percentile)**: the middle value between the median and the highest value (not the “maximum”) of the dataset.
- **interquartile range (IQR)**: 25th to the 75th percentile.
- **whiskers (shown in blue)**
- **outliers (shown as green circles)**
- **“maximum”**: Q3 + 1.5*IQR
- **“minimum”**: Q1 -1.5*IQR

https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51