Data Preparation

St. Joseph's University, Bengaluru

Basics of Data Science

Data Cleaning

Redundant White Space

Data Cleaning

Fixing Capital Letter

Case Matters!

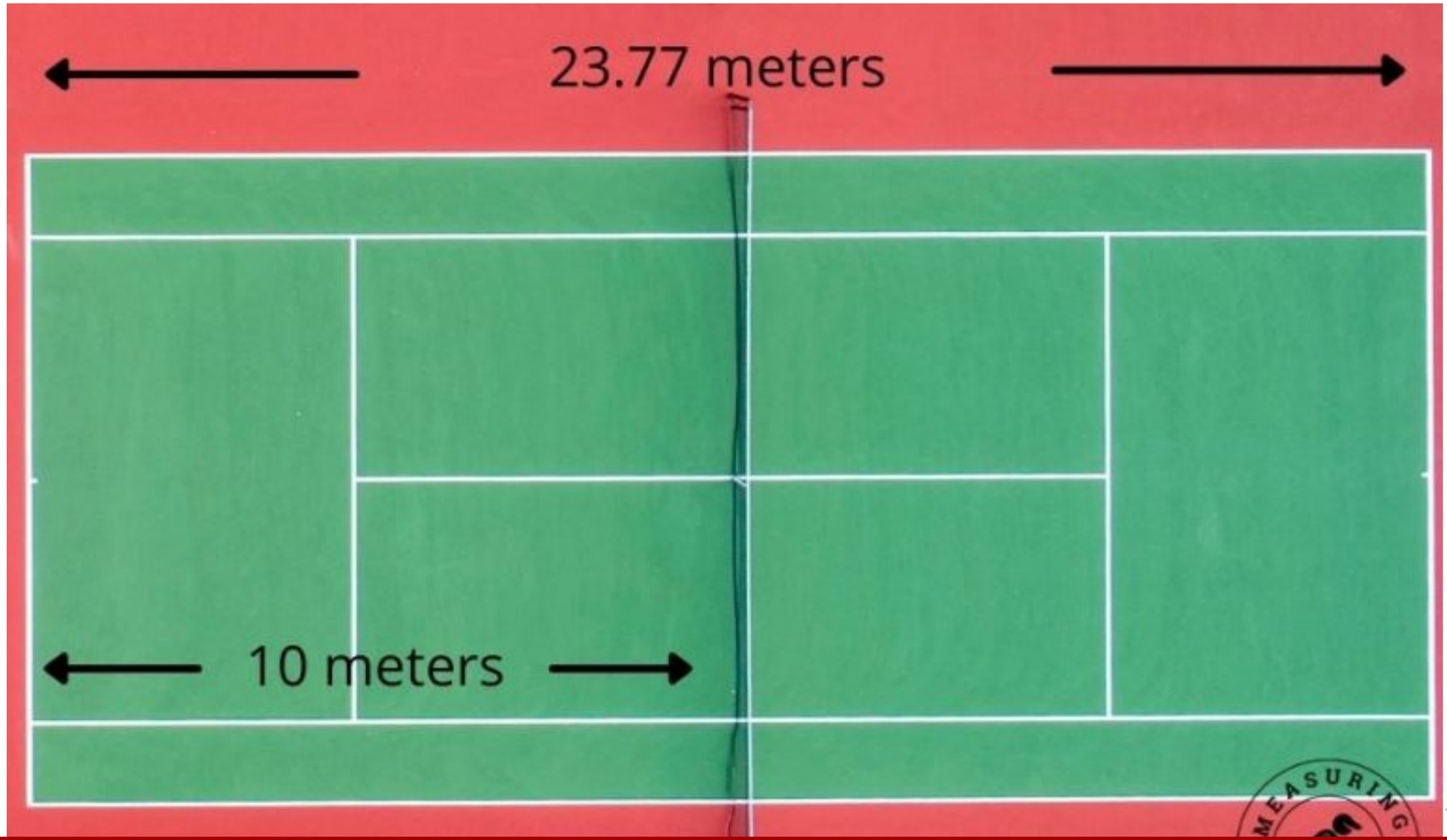
"Brazil" = "brazil"

"HE" = "he" = "He"

Data Cleaning

Impossible Values and Sanity Checks

"My data has a 10m in height, 300yr old man!"



Data Cleaning

Outliers

Data Cleaing

Missing Values

Technique	Advanatge	Disadvantage
Omit the Values	Easy	Loss of infomration
Set Value to Null	Easy	Not applicable to all models
Impute a static value	Easy No loss of information	Can lead to false estimates
Impute a theoretical value	Does not disturb the model	Harder to execute Data assumptions are made

Data Cleaning

Deviations from a Code Book

Features

ATTRIBUTE NAME	ROLE	TYPE	DESCRIPTION	UNITS	MISSING VALUES
sepal length	Feature	Continuous		cm	false
sepal width	Feature	Continuous		cm	false
petal length	Feature	Continuous		cm	false
petal width	Feature	Continuous		cm	false
class	Target	Categorical	class of iris plant		false