Data Science Masterclass

Data Cleaning

H.M. Samadhi Chathuranga Rathnayake

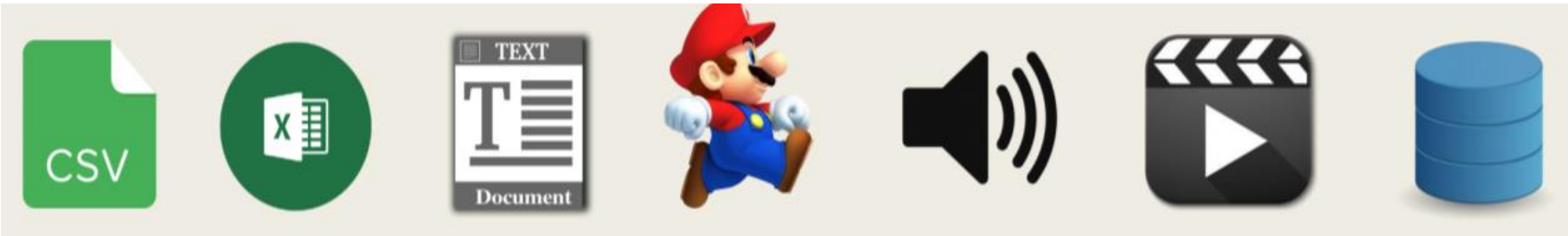
M.Sc in CS (SU), PG.Dip in SML (Othm), PG.Dip in HRM (LRN), B.Sc (Hons) in IS (UOC), B.Eng (Hons) in SE (LMU), P. Dip EP & SBO (ABE), Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

Data Sources

Where does data come from?

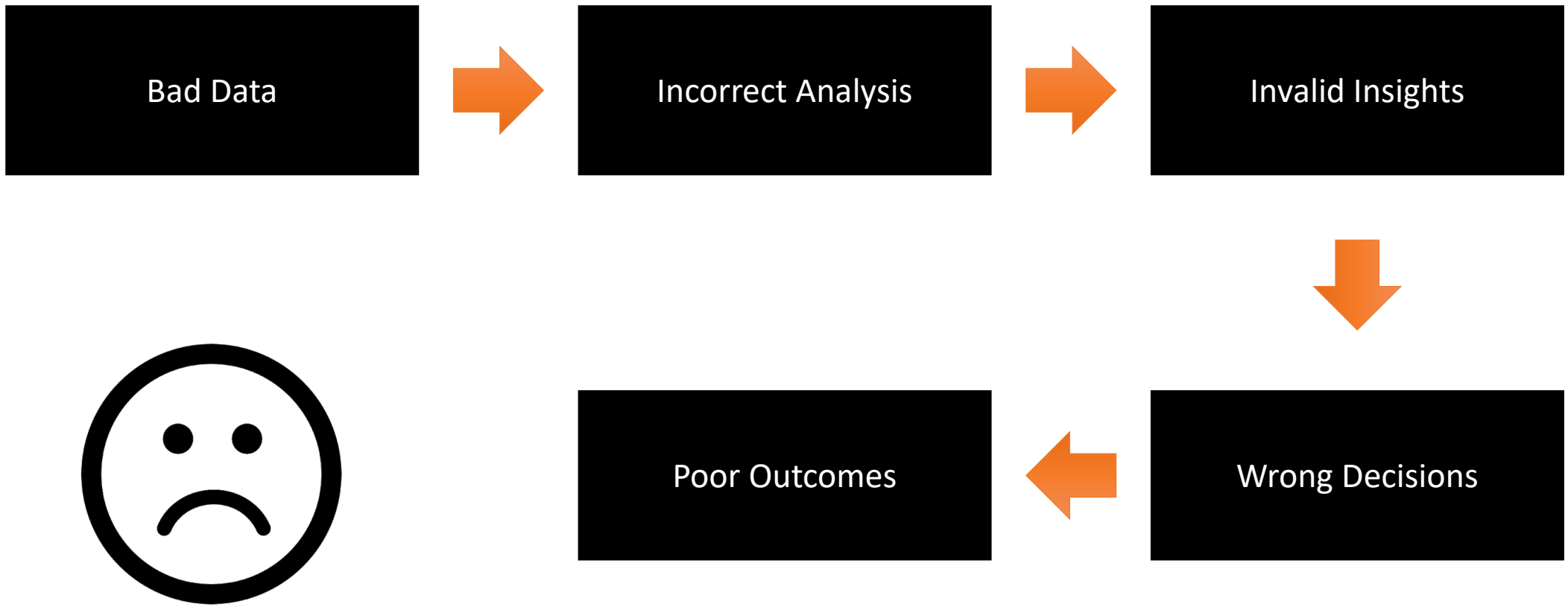
- Proprietary data sources
- Government data sets
- Academic data sets
- Web search
- Sensor data
- Crowdsourcing
- By researcher (Creating own datasets)

What are the formats?



If they contain bad data

What happens then?



Data Wrangling

Data wrangling is a process which should be conducted before starting an analysis or model fitting. Some popular preprocessing steps are,

- Data Manipulation
- Data Cleaning (Missing Values, Duplicates & Outliers)

There are many more.

Data Cleaning

Common problems with data

- Missing values
- Outliers
- Duplicates

Data Cleaning

Missing values

Name	Age	Height (cm)	Weight (kg)
Jane	23	167	50
David	24	168	70
Scott	21	170	
Harry		182	50
Anne	20	153	38

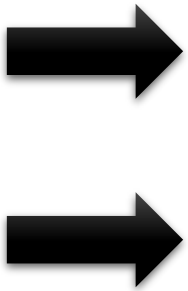
Data Cleaning

Dealing with missing values

- Removing missing value columns & rows
- Filling missing values

Data Cleaning

Duplicate values



Name	Age	Height (cm)	Weight (kg)
Jane	23	167	50
David	24	168	70
Scott	21	170	68
Harry	22	182	50
Scott	21	170	68

Data Cleaning

Dealing with duplicate values

- Removing duplicate rows

Data Cleaning

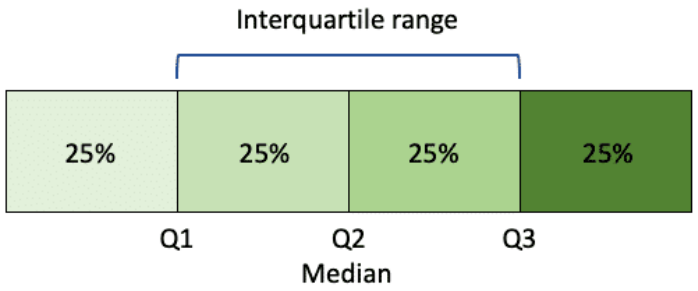
Outliers

Name	Age	Height (cm)	Weight (kg)
Jane	23	167	50
David	24	168	150
Scott	21	170	68
Harry	22	182	50
Anne	20	153	38

Data Cleaning

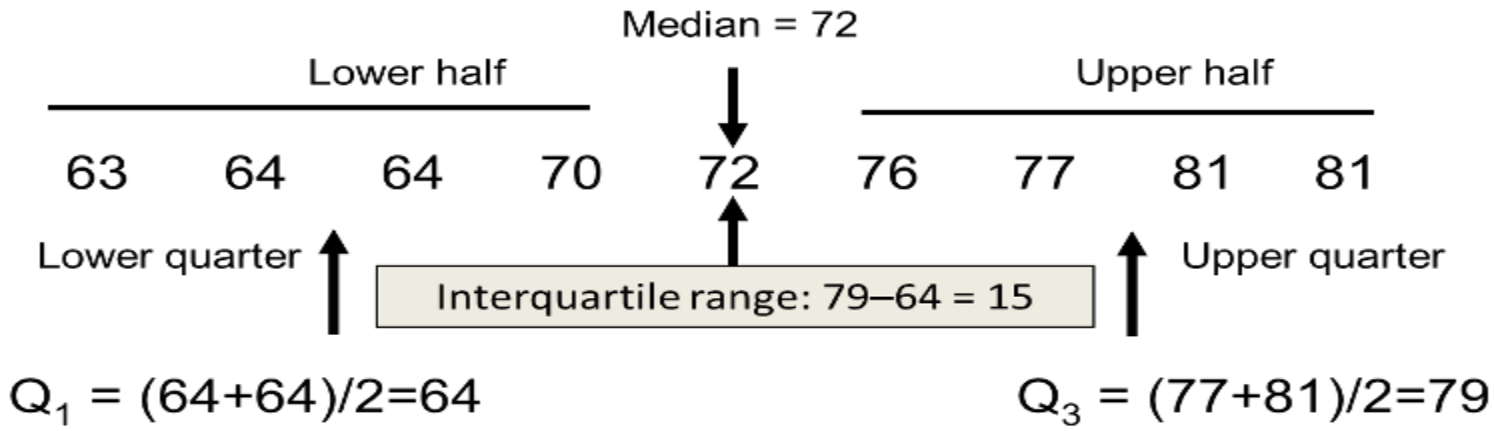
Dealing with outliers

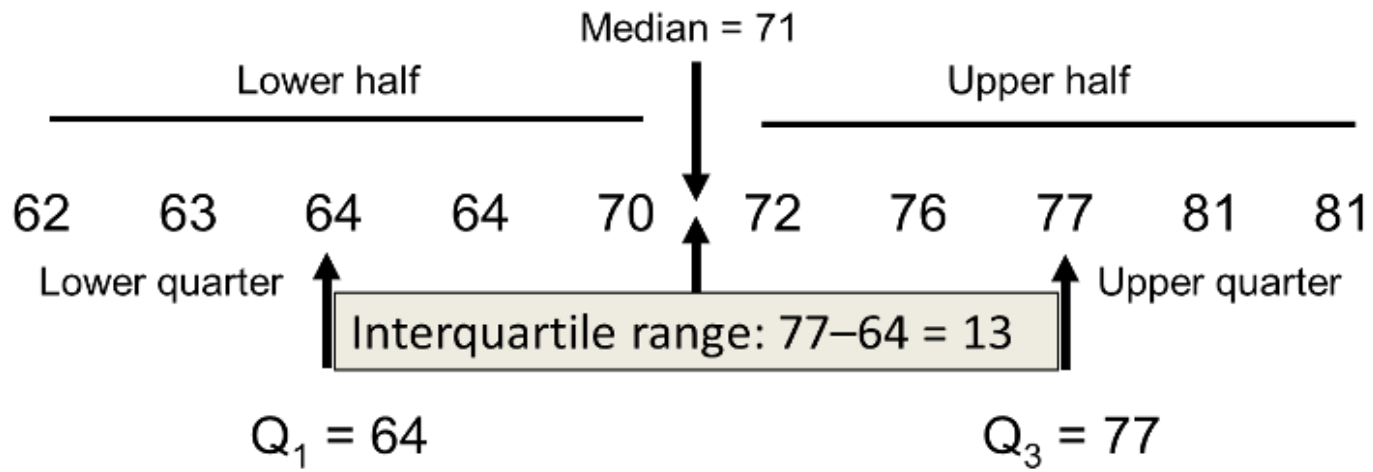
- Detect the outliers
- Remove the outliers
  - Calculate quartiles
  - Calculate the Inter Quartile Range (IQR)
  - Remove the values outside 1.5 times IQR



Data Cleaning

IQR Example





Handling Untidy Data

Uni	2015	2016	2017
UOC	8902	9221	9021
UOK	6789	7834	7634
UOM	5600	5467	6234

Uni	Year	Intake
UOC	2015	8902
UOC	2016	9221
UOC	2017	9021
UOK	2015	6789
UOK	2016	7834
UOK	2017	7634
UOM	2015	5600
UOM	2016	5467
UOM	2017	6234

- Not only these cases, but also there are some situations where the column types are bad.
- Sometimes the existing columns should be changed, or new columns should be created before the analysis.