Imperial College London

Introduction to data processing and cleaning

Valentina Giunchiglia and Dragos Gruia

Imperial College London

Schedule of today

Morning

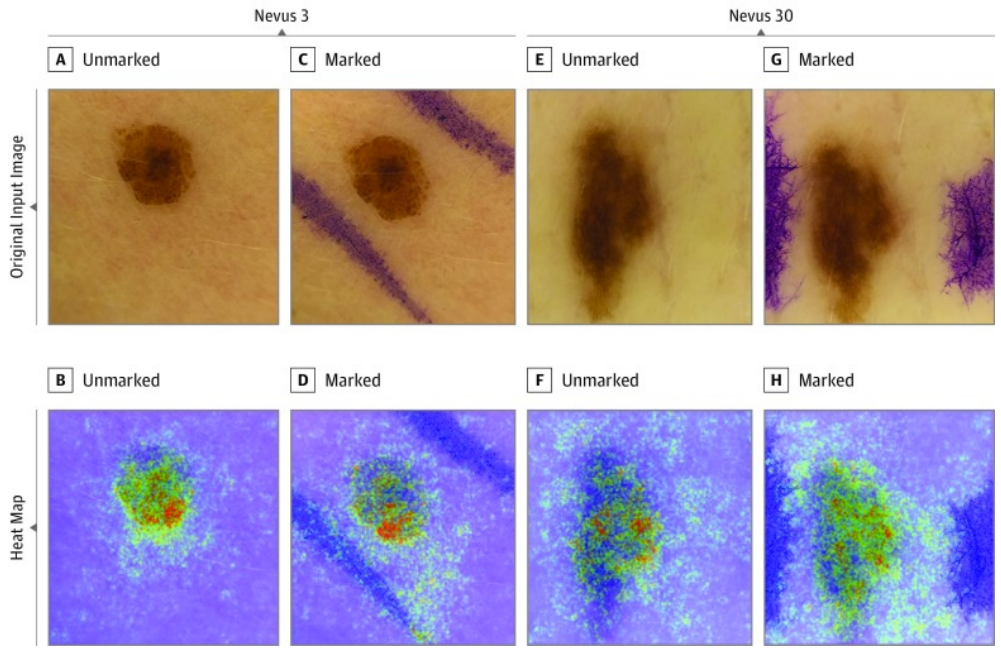
- *PRESENTATION:* What is data cleaning and processing?
- *GUIDED WORKSHOP:* COVID and Cognition

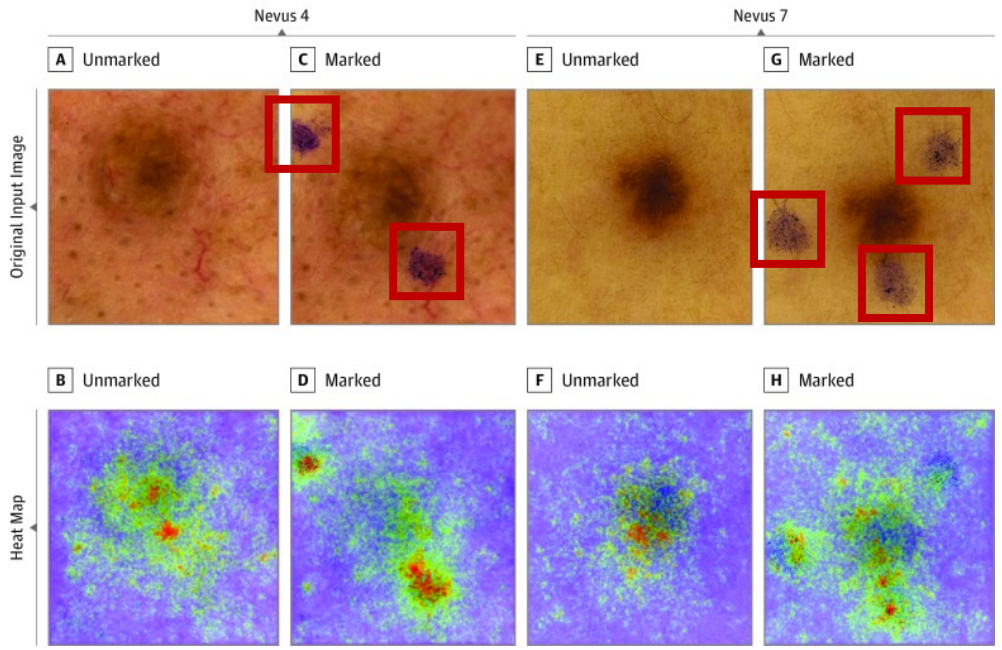
Afternoon

- *INDIVIDUAL WORKSHOP*: Dementia and Cognition
- *PRESENTATION*: How to write a report?

Imperial College London

Why is it important?



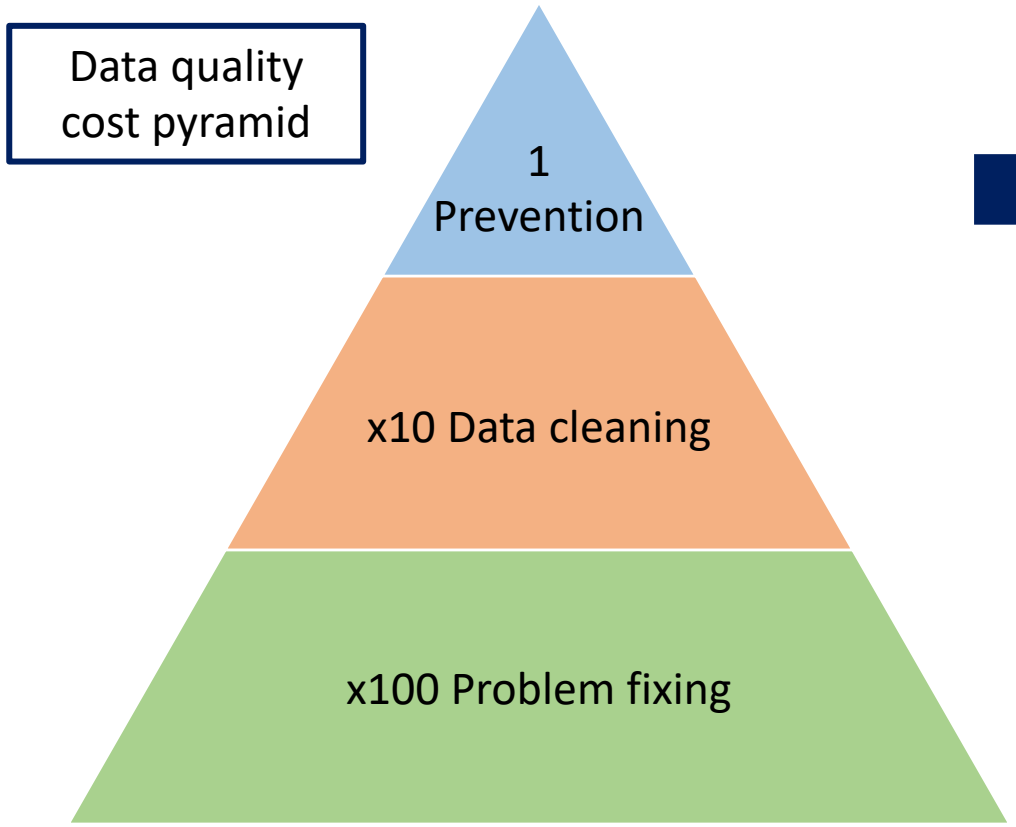


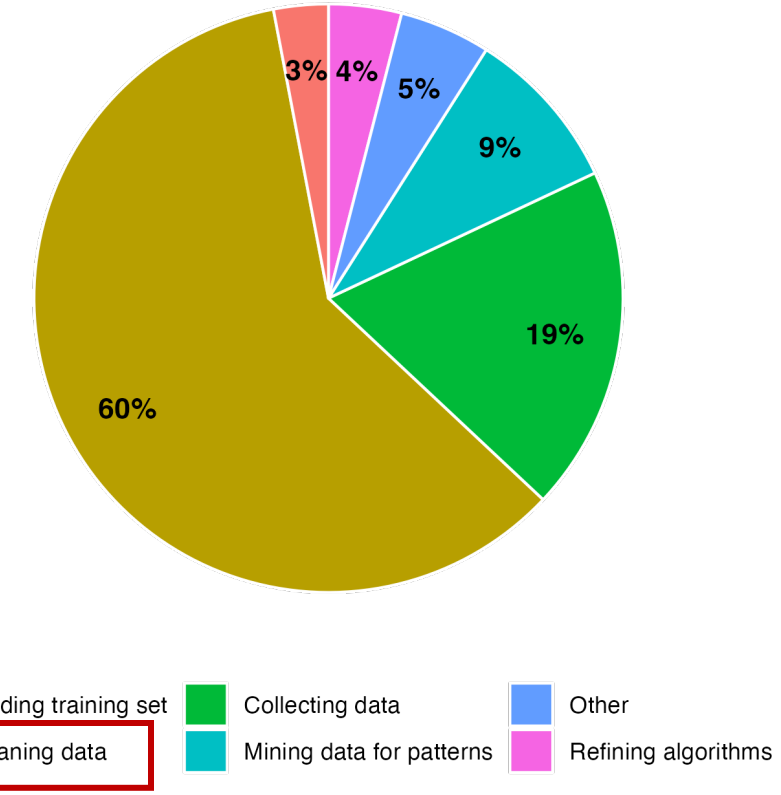
Machine learning algorithm predicts the marking rather than the melanoma

Winkler et al. 2019

Imperial College London

Why is it important?





Sebastianelli et al. 2020

Imperial College London

Data cleaning and processing

DATA CLEANING

Process of fixing and removing incorrect, corrupted, incomplete, duplicated, and incorrectly formatted data from a dataset

VS

DATA PROCESSING

Process of data conversion from a given form to a more easy-to-use one based on the aims of the analysis

Imperial College London

Data cleaning and processing

DATA CLEANING

Process of fixing and removing incorrect, corrupted, incomplete, duplicated, and incorrectly formatted data from a dataset

VS

DATA PROCESSING

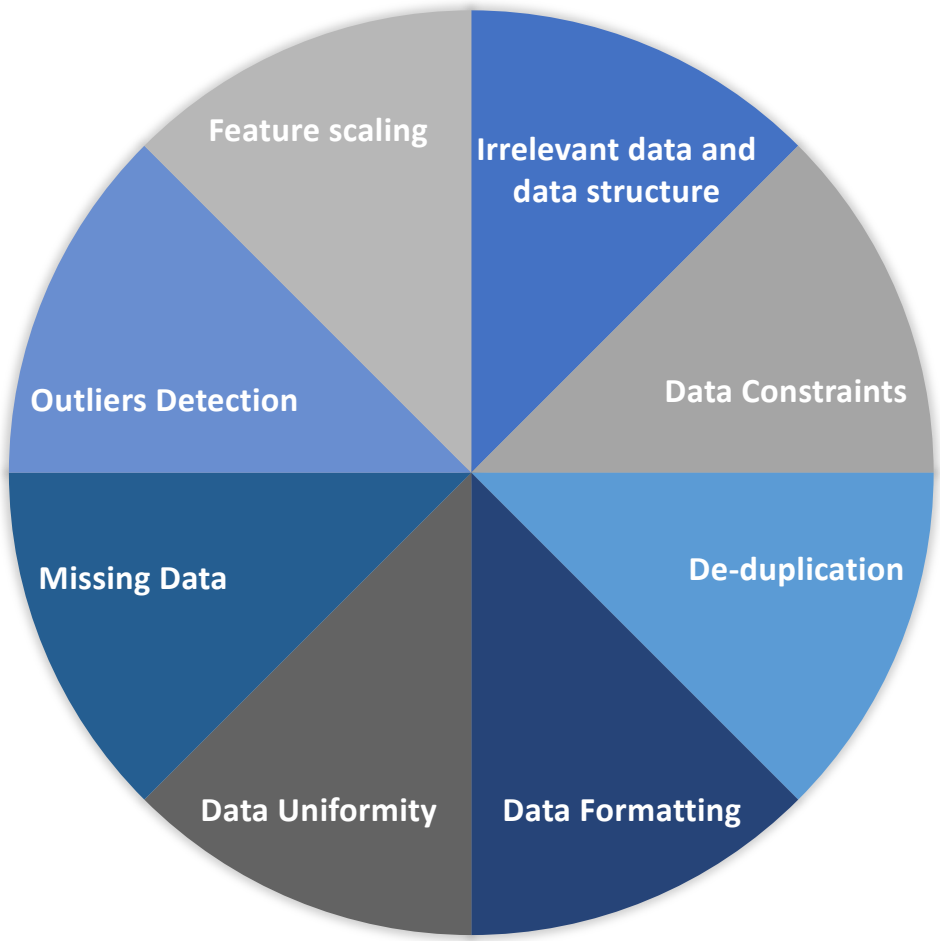
Process of data conversion from a given form to a more easy-to-use one based on the aims of the analysis

Imperial College London

Data cleaning

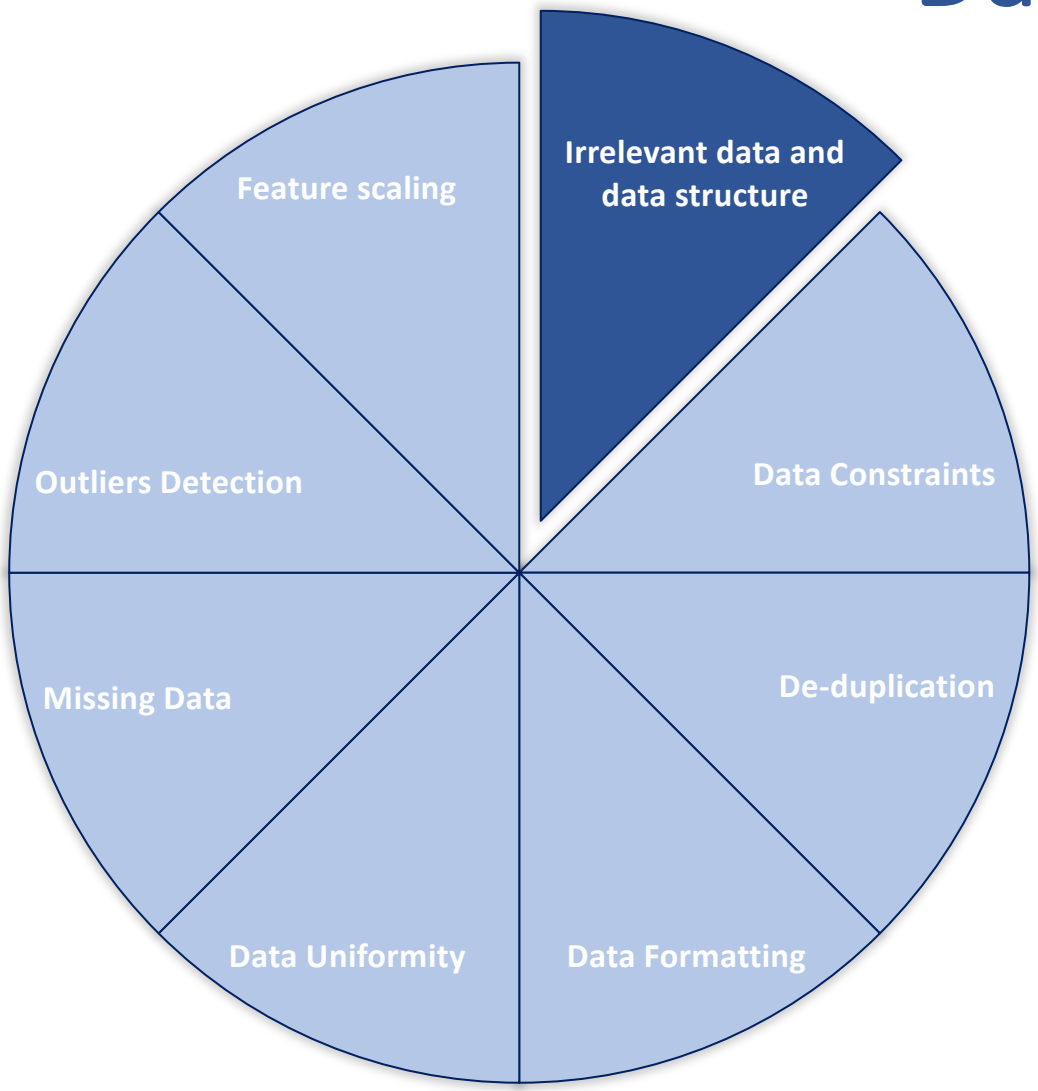
DATA CLEANING

Process of fixing and removing incorrect, corrupted, incomplete, duplicated, and incorrectly formatted data from a dataset



Imperial College London

Data cleaning: irrelevant data



Irrelevant data and data structure

?

1. Do you have row and column names?
2. Are column and row names interpretable?
3. Which data do you need for your analysis?

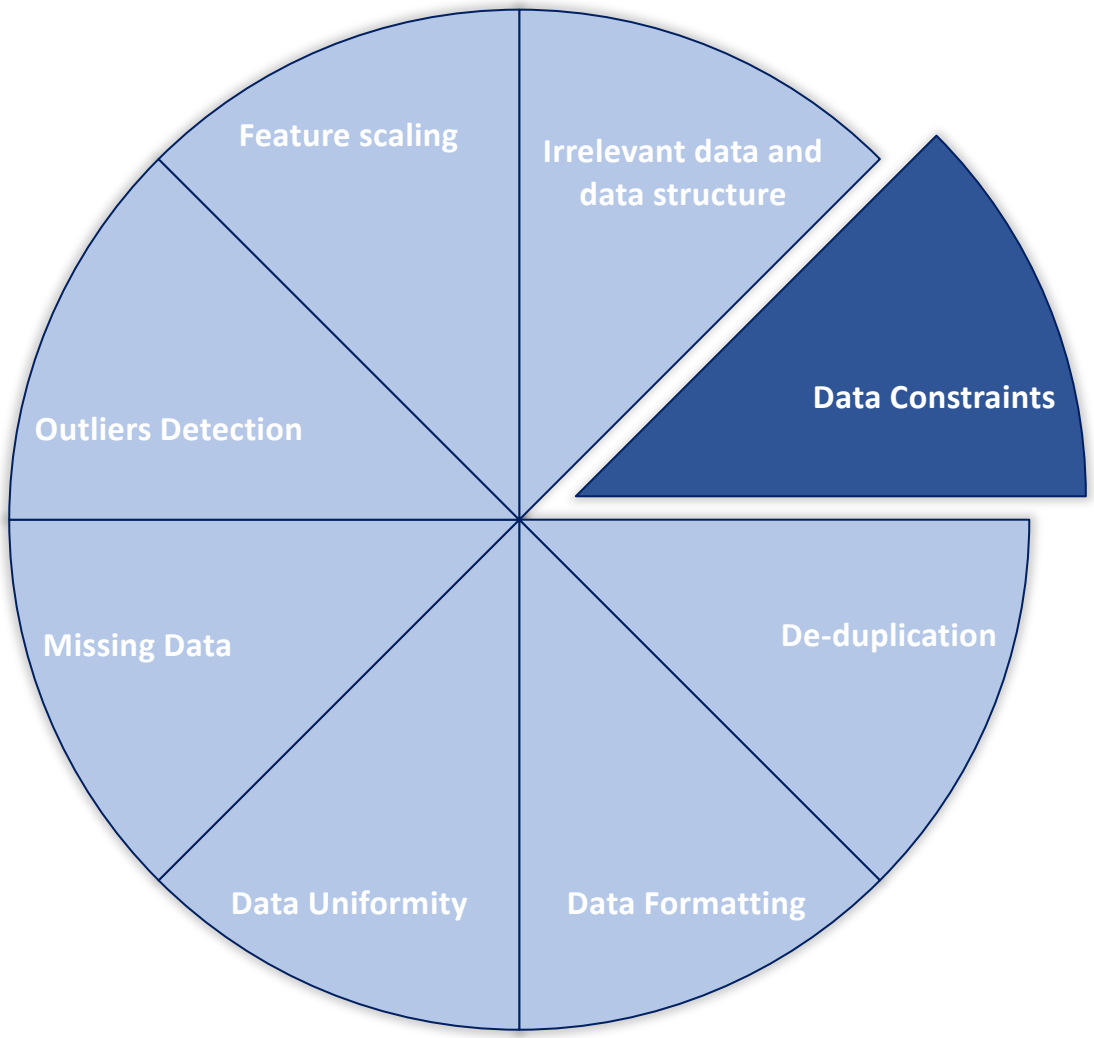
✓

1. Add column and row names
2. Modify column and row names
3. Filter data

Computational resources and interpretability

Imperial College London

Data cleaning: data constraints



Data Constraints

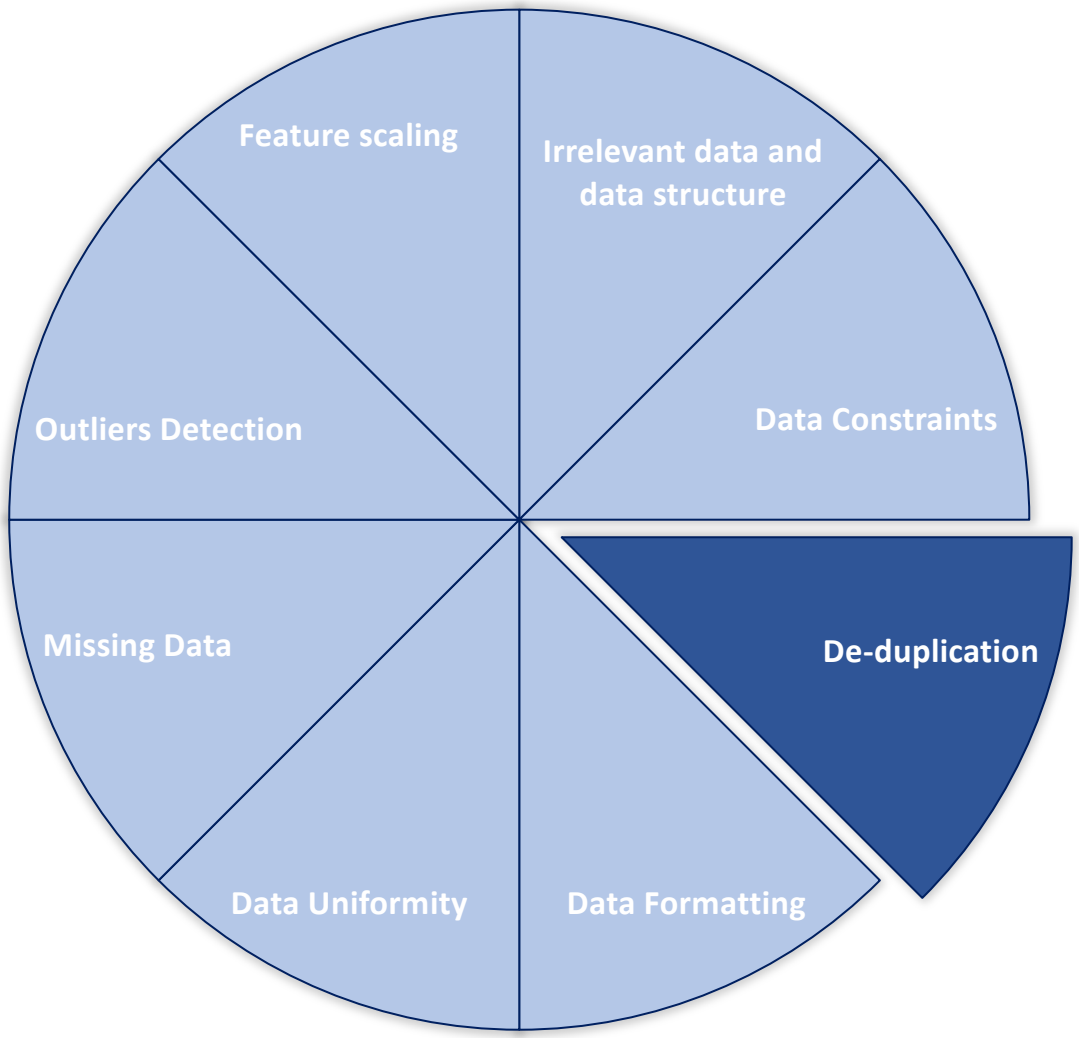
1. Data Type:
   - Are numerical columns *int* or *float*?
   - Are categorical variables *str*?

2. Data Range:
   - Are the numerical columns within the expected range? (Ex: Age)

1. Convert to proper data types
2. Different solutions:
   1. Set to *NA*
   2. Set equal to *max* or *min*
   3. Remove participants
   4. Check for typing error

Imperial College London

Data cleaning: duplicates



De-duplication

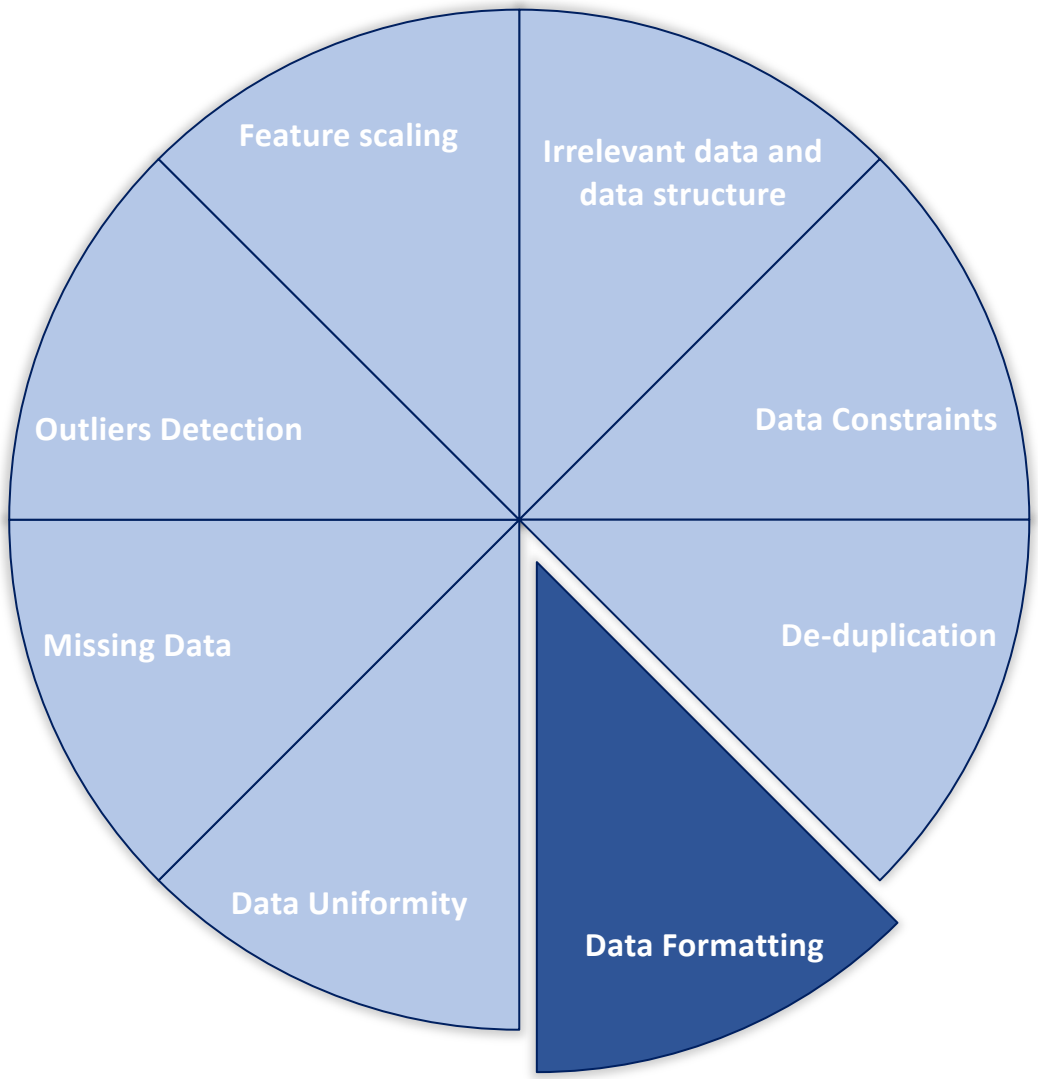
-> Why do you have duplicates?

1. Are there fully duplicated rows?
2. Are there partially duplicates rows?

1. Removal of duplicates
2. Different solutions (depend on WHY):
   - Merge the partial duplicates
   - Remove the partial duplicates
   - Select one of the partial duplicates
   - Keep entry that are duplicates and set rest as *NA*

Imperial College London

Data cleaning: formatting



Data formatting

?

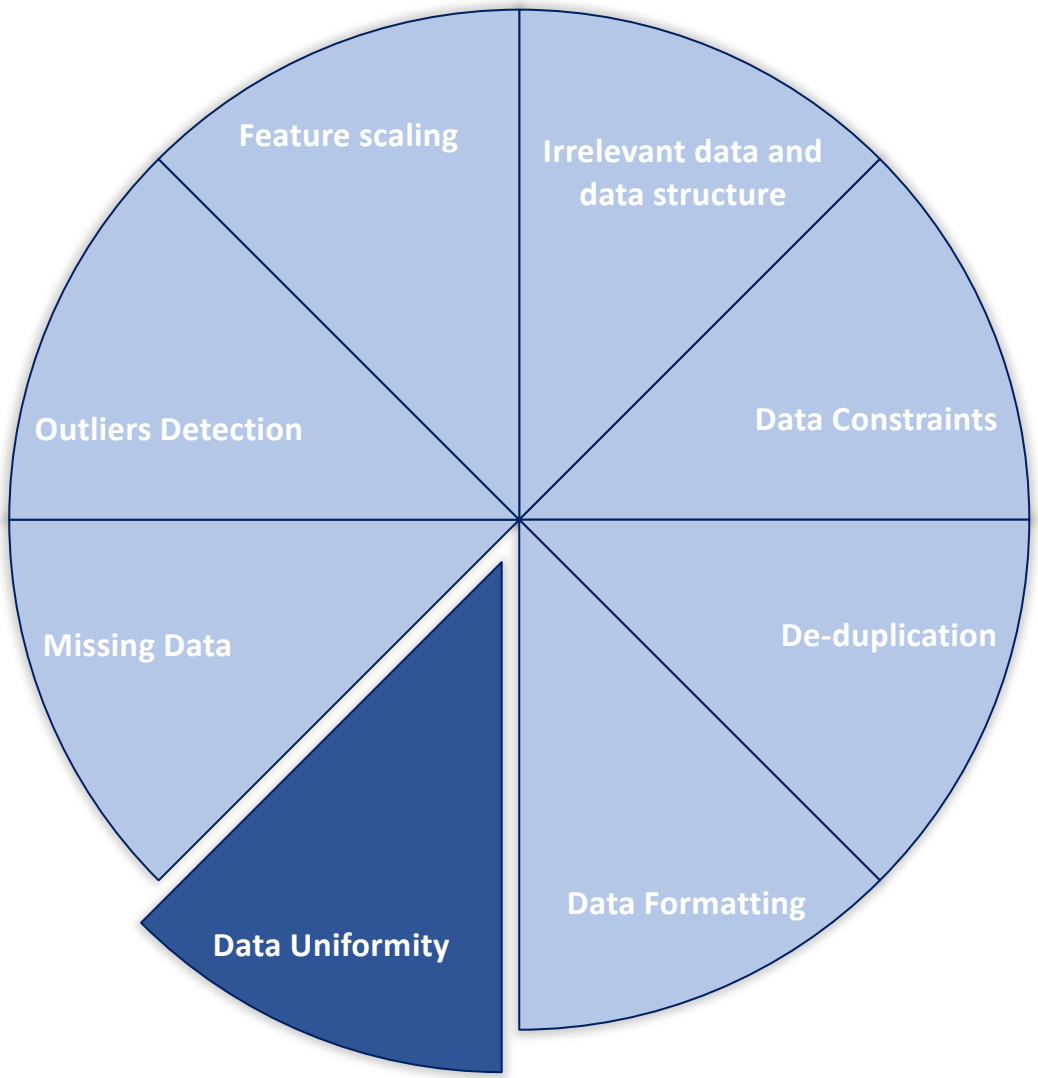
1. Are there unique names for each category (e.g. Education?)
2. Are text data consistently formatted? (Ex: dates)
3. Are there special characters?

✓

1. Different solutions based on data:
   - Map different categories to unique names
   - Set as NA categories with different formatting (if right category not clear)
   - Infer right category from different data entries
2. Different solutions:
   1. Match right formatting
   2. Set as NA (if not interpretable)
   3. Infer from other data points
3. Remove special characters (iMPORTANT for free text)

Imperial College London

Data cleaning: uniformity



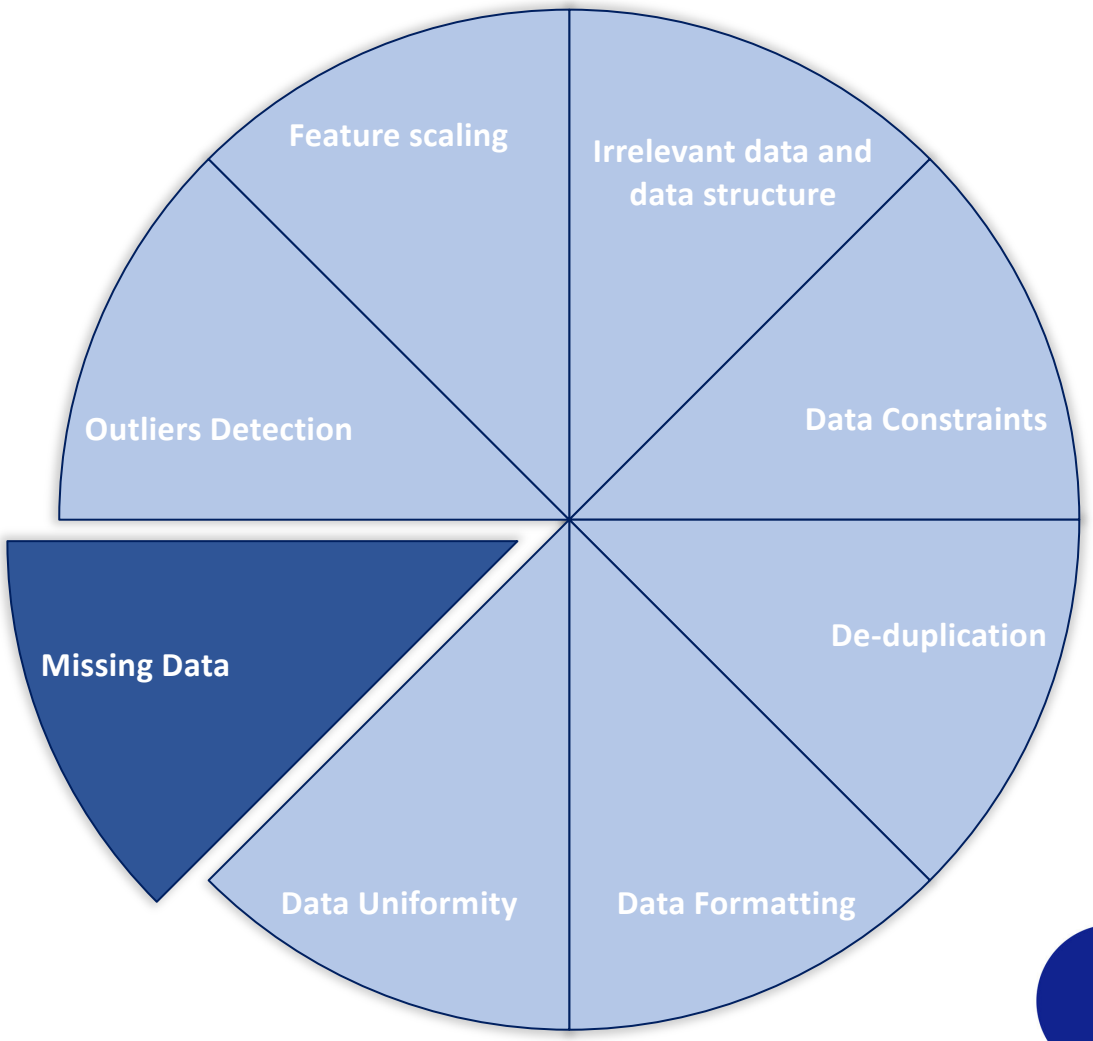
Data Uniformity

1. Do all numerical values have the same units (Ex: ms)?
2. Are columns with similar information consistent (Ex: age and age in decade)?

1. Different solutions:
   - Standardise units (if easy to understand)
   - Set as NA
2. Different solutions:
   - Correct the inconsistent values
   - Set as NA

Imperial College London

Data cleaning: missing data



Missing data

-> Are the data missing at random or not? Why are they missing?

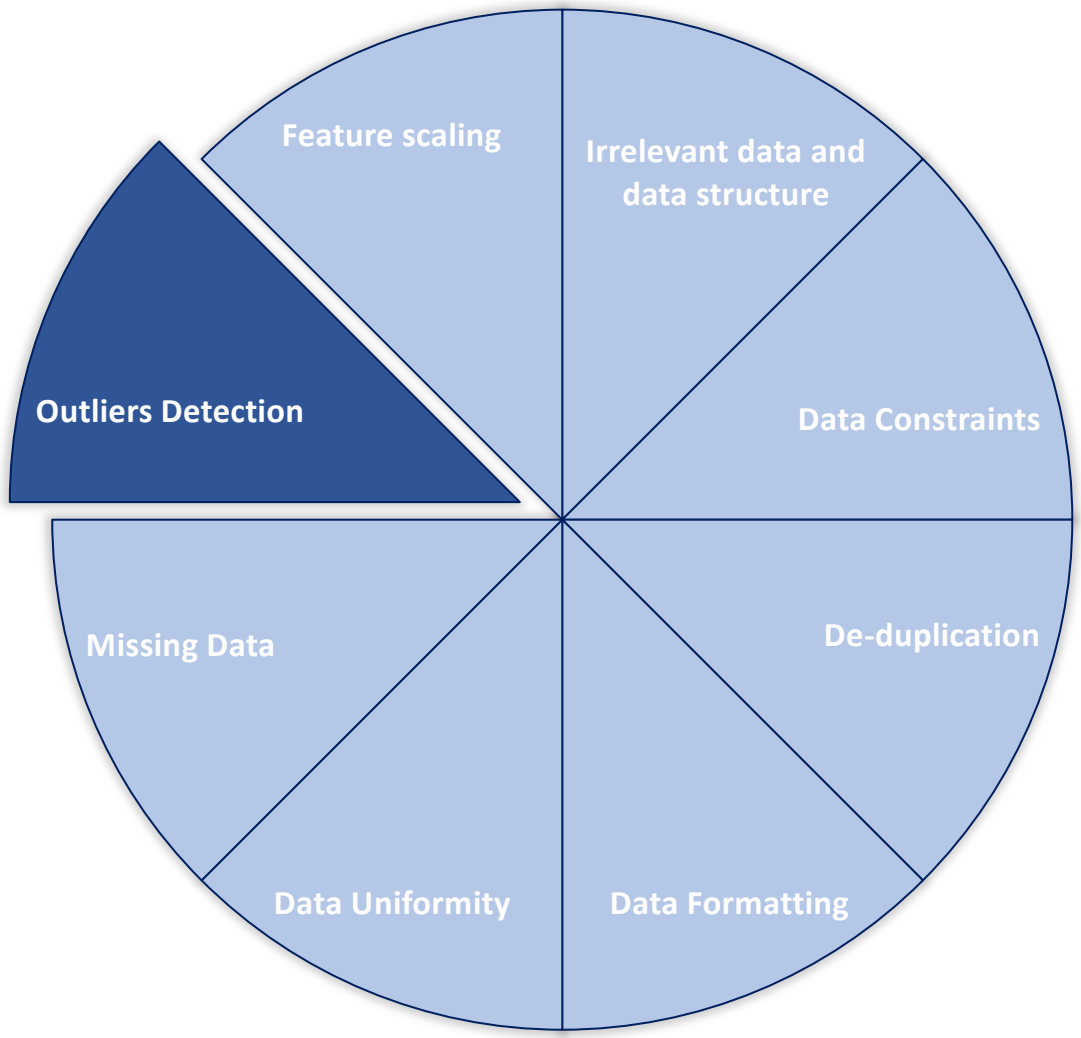
1. Are there missing data? How many? (IMPORTANT: Always check the number of missing data)

1. Different solutions:
   - Drop participants with more than N *NA*
   - Drop columns with more than N *NA*
   - Impute missing data with mean/median
   - Impute missing data with ML algorithms
   - Infer missing data from other data points

You cannot analyse data if there are missing values

Imperial College London

Data cleaning: outliers



Outliers detection

1. Are there outliers? (WHY?)

?

Are they really outliers?

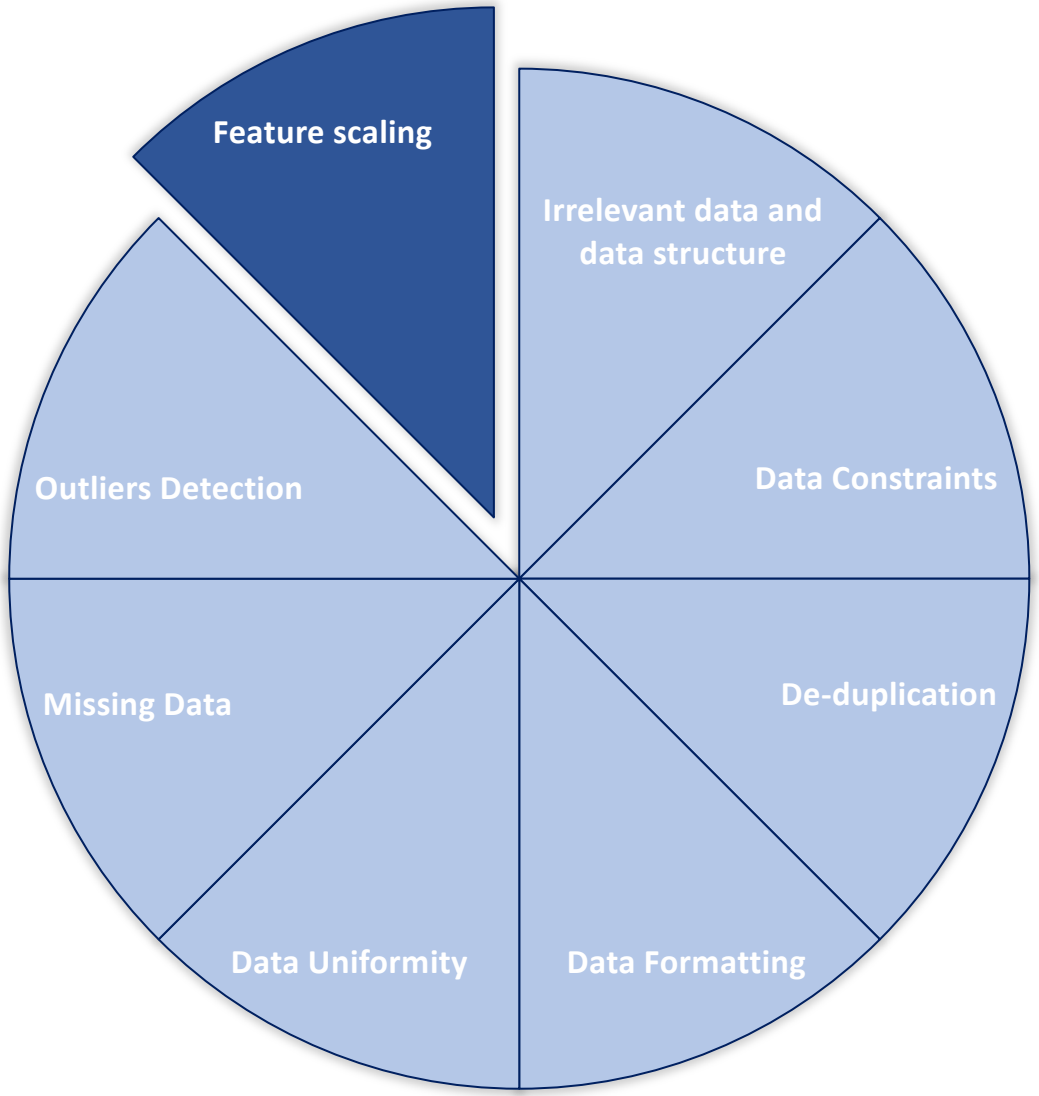
1. Different solutions:
   - Filter out the outliers (e.g. set as NA)
   - Winsorize the outliers
   - Set as mean/median

✓

It really depends on the type of outliers!!

Imperial College London

Data cleaning: feature scaling



Feature scaling

1. Are the numerical data properly scaled?

1. Different solutions:
   - Data stadardisation: rescale data to have a mean of 0 and standard deviation of 1
   - Data normalisation: rescale values in a range between 0 and 1

Scaling is really important in Machine learning

Imperial College London

Let's apply what you have learnt to real data!