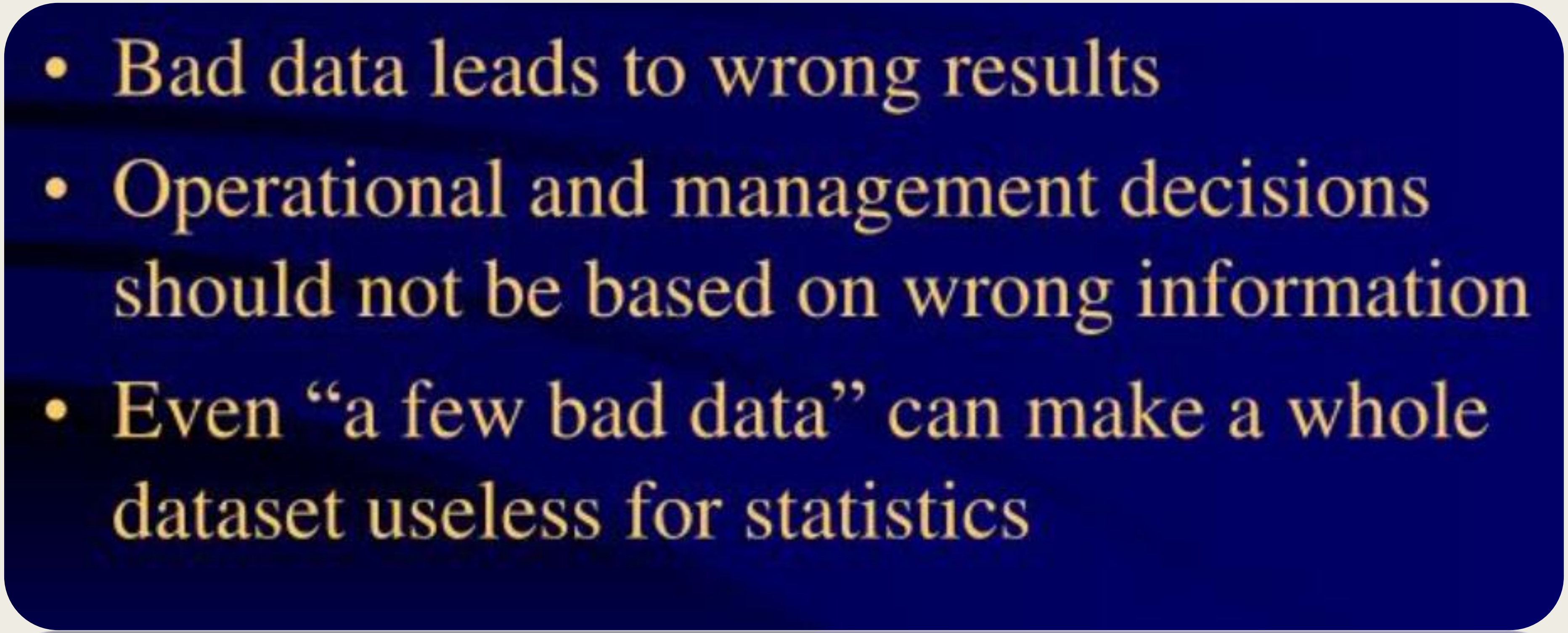
DATA CLEANING TECHNIQUES

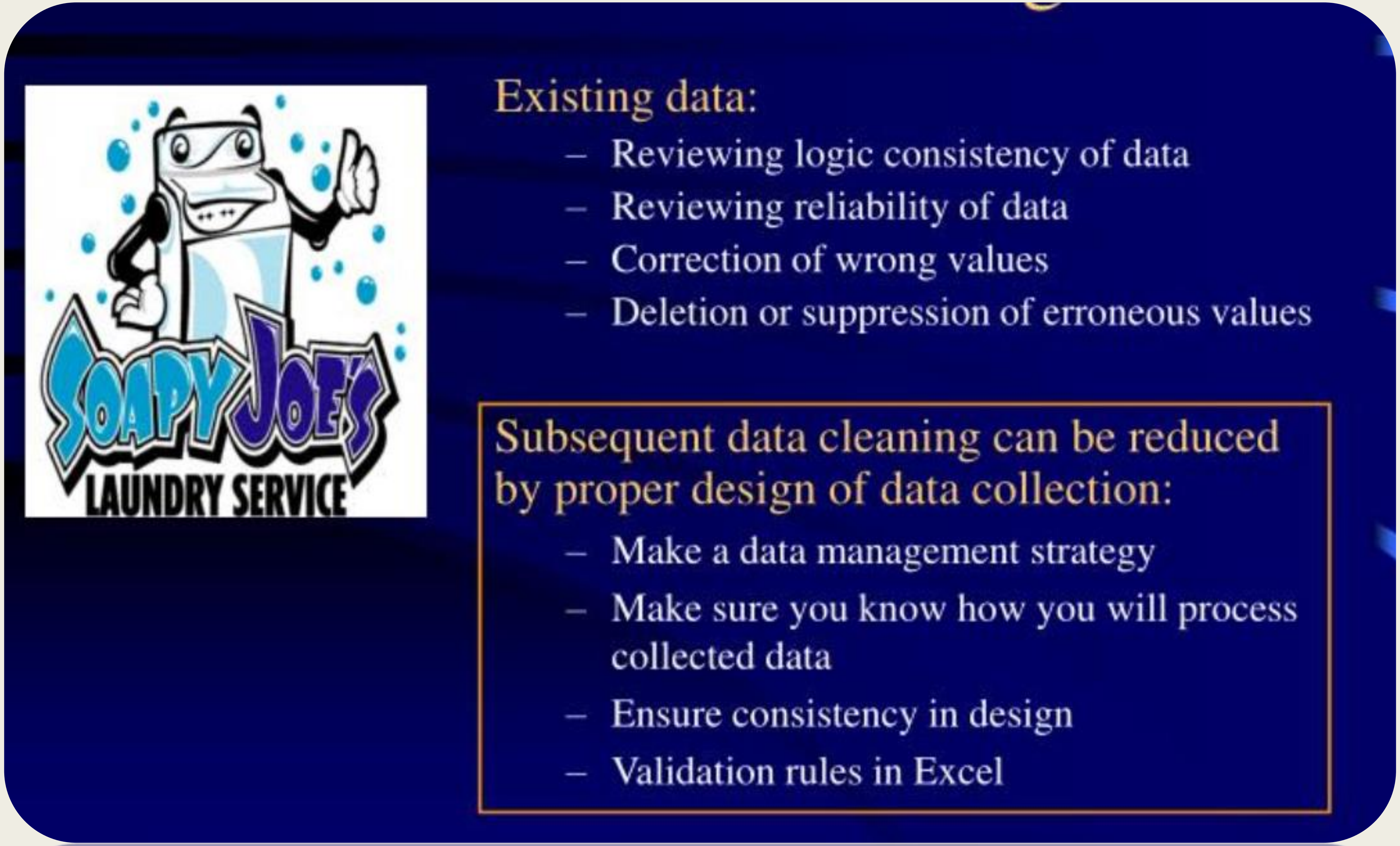
Prof. Sarala Mary

Data Cleaning – Why is it important?



- Bad data leads to wrong results
- Operational and management decisions should not be based on wrong information
- Even “a few bad data” can make a whole dataset useless for statistics

What is Data Cleaning?


SOAPY JOE'S
LAUNDRY SERVICE

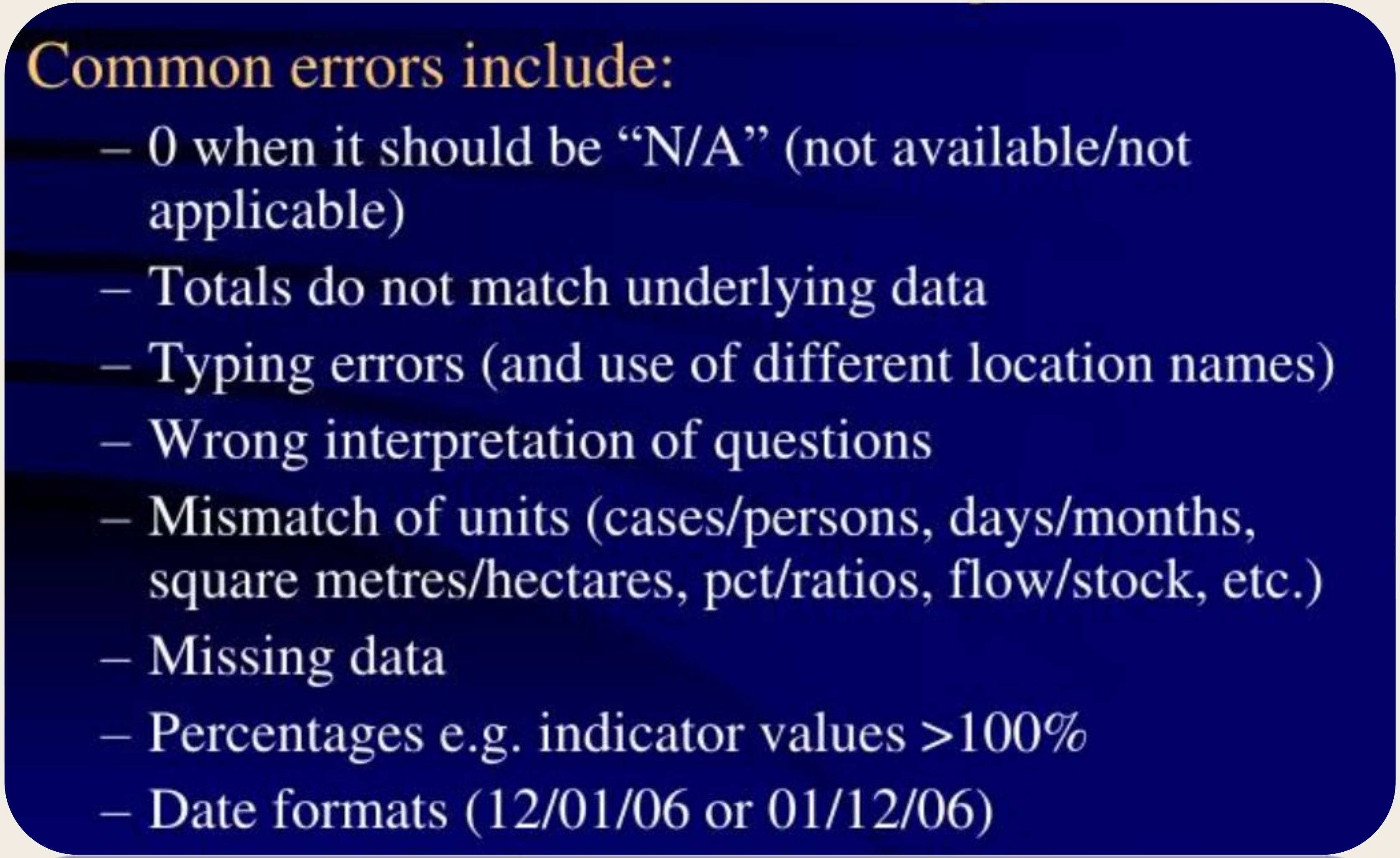
Existing data:

- Reviewing logic consistency of data
- Reviewing reliability of data
- Correction of wrong values
- Deletion or suppression of erroneous values

Subsequent data cleaning can be reduced by proper design of data collection:

- Make a data management strategy
- Make sure you know how you will process collected data
- Ensure consistency in design
- Validation rules in Excel

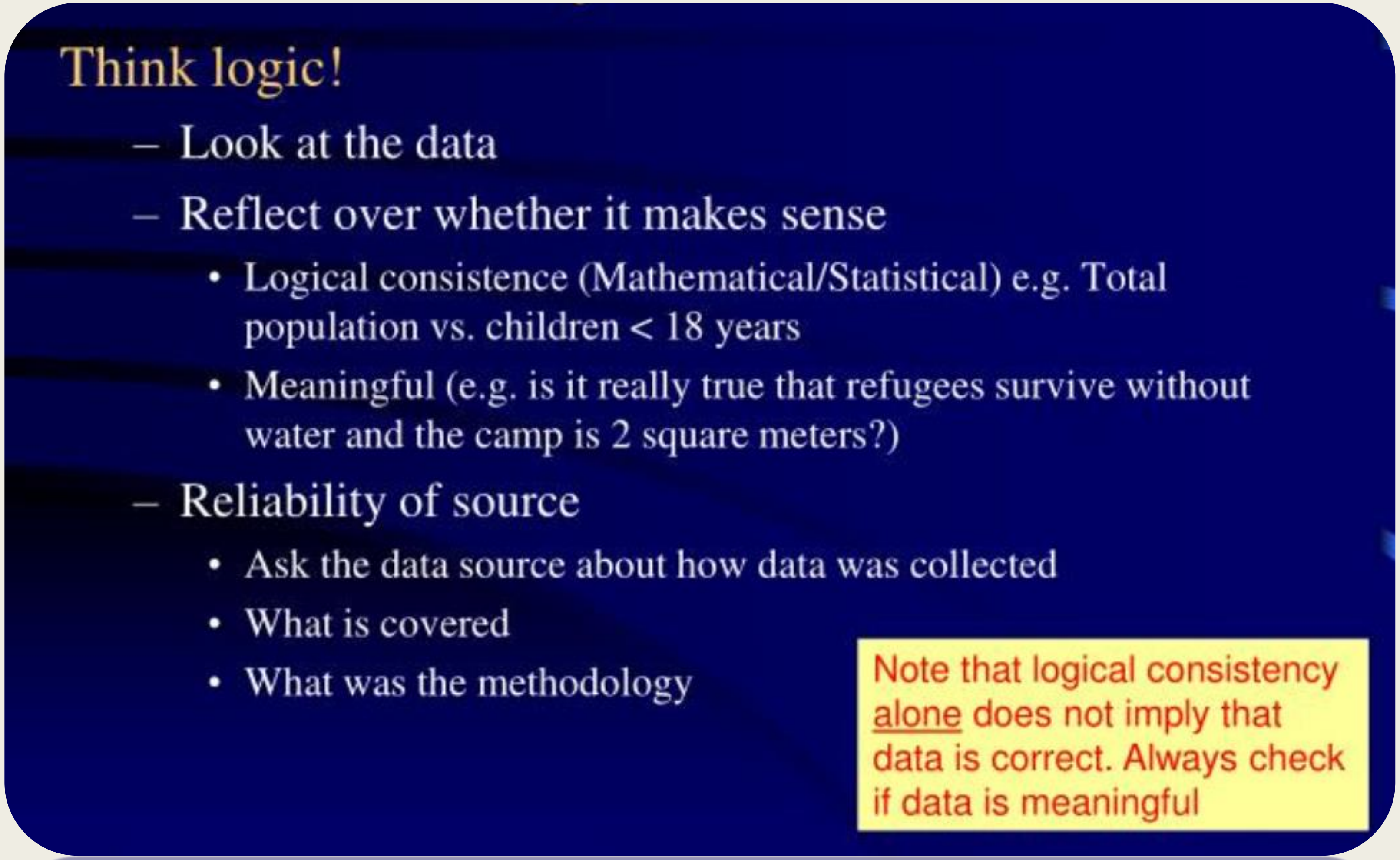
What are we looking for?



Common errors include:

- 0 when it should be “N/A” (not available/not applicable)
- Totals do not match underlying data
- Typing errors (and use of different location names)
- Wrong interpretation of questions
- Mismatch of units (cases/persons, days/months, square metres/hectares, pct/ratios, flow/stock, etc.)
- Missing data
- Percentages e.g. indicator values >100%
- Date formats (12/01/06 or 01/12/06)

How do you clean Data?

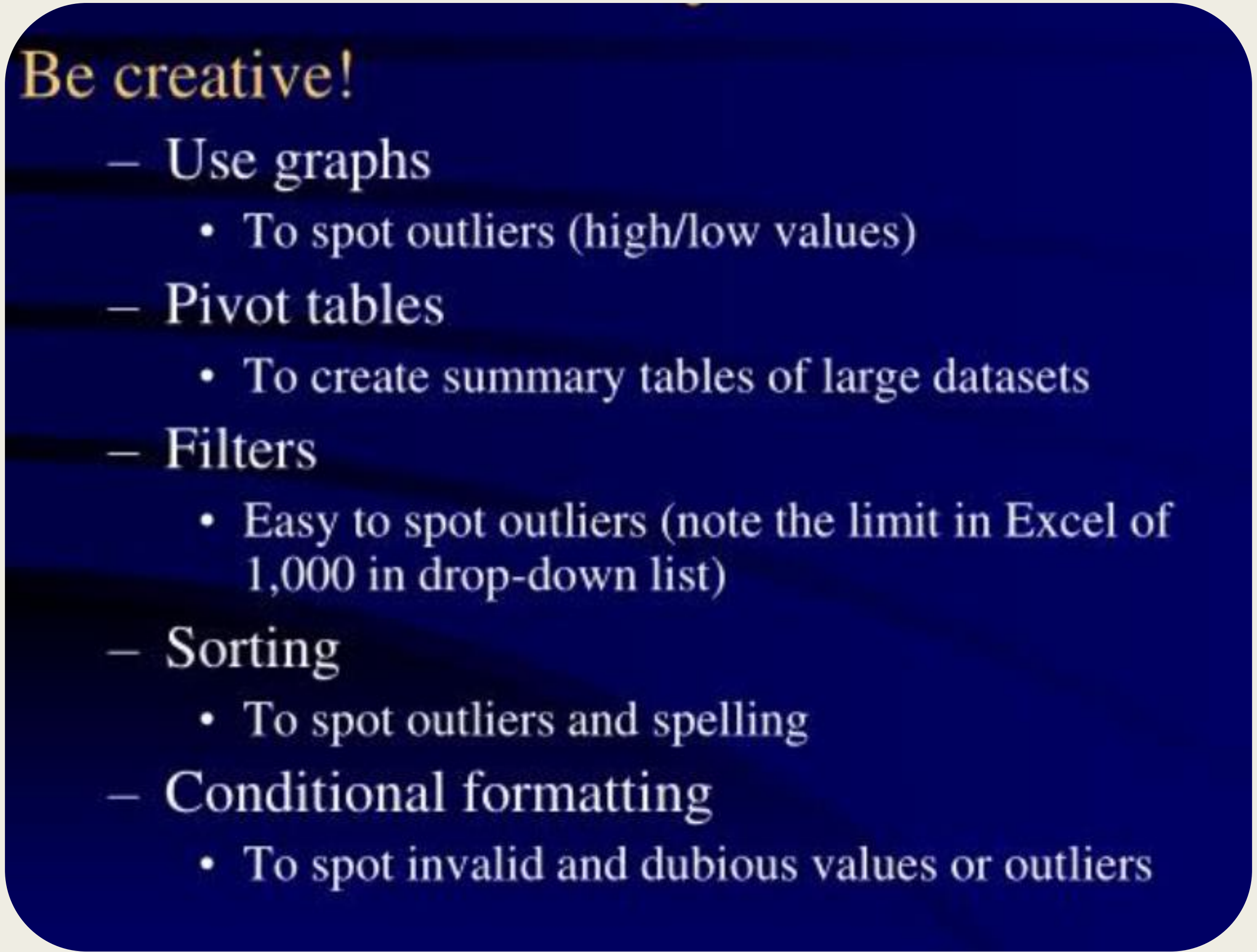


Think logic!

- Look at the data
- Reflect over whether it makes sense
  - Logical consistence (Mathematical/Statistical) e.g. Total population vs. children < 18 years
  - Meaningful (e.g. is it really true that refugees survive without water and the camp is 2 square meters?)
- Reliability of source
  - Ask the data source about how data was collected
  - What is covered
  - What was the methodology

Note that logical consistency alone does not imply that data is correct. Always check if data is meaningful

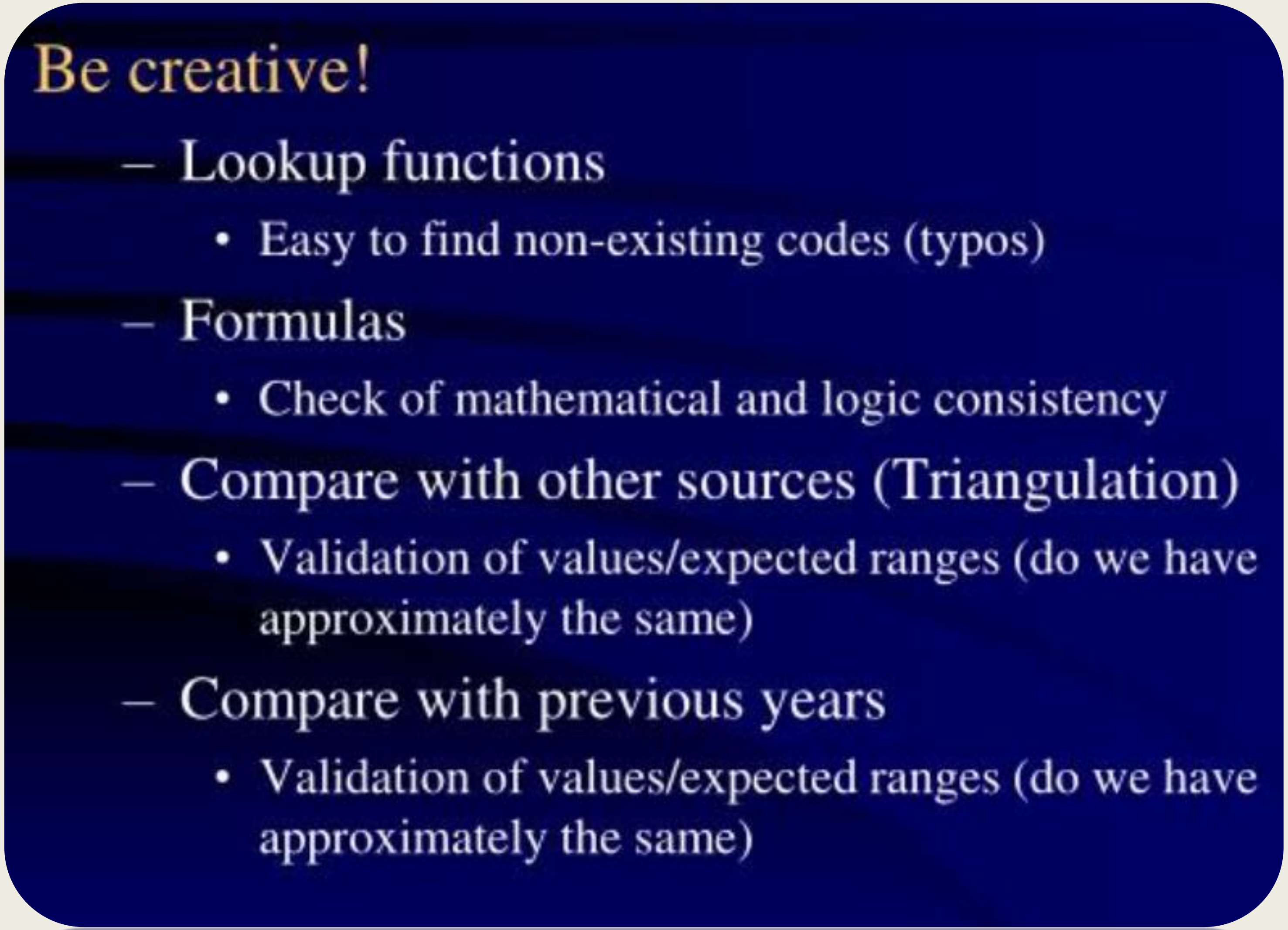
How do you clean Data?



Be creative!

- Use graphs
  - To spot outliers (high/low values)
- Pivot tables
  - To create summary tables of large datasets
- Filters
  - Easy to spot outliers (note the limit in Excel of 1,000 in drop-down list)
- Sorting
  - To spot outliers and spelling
- Conditional formatting
  - To spot invalid and dubious values or outliers

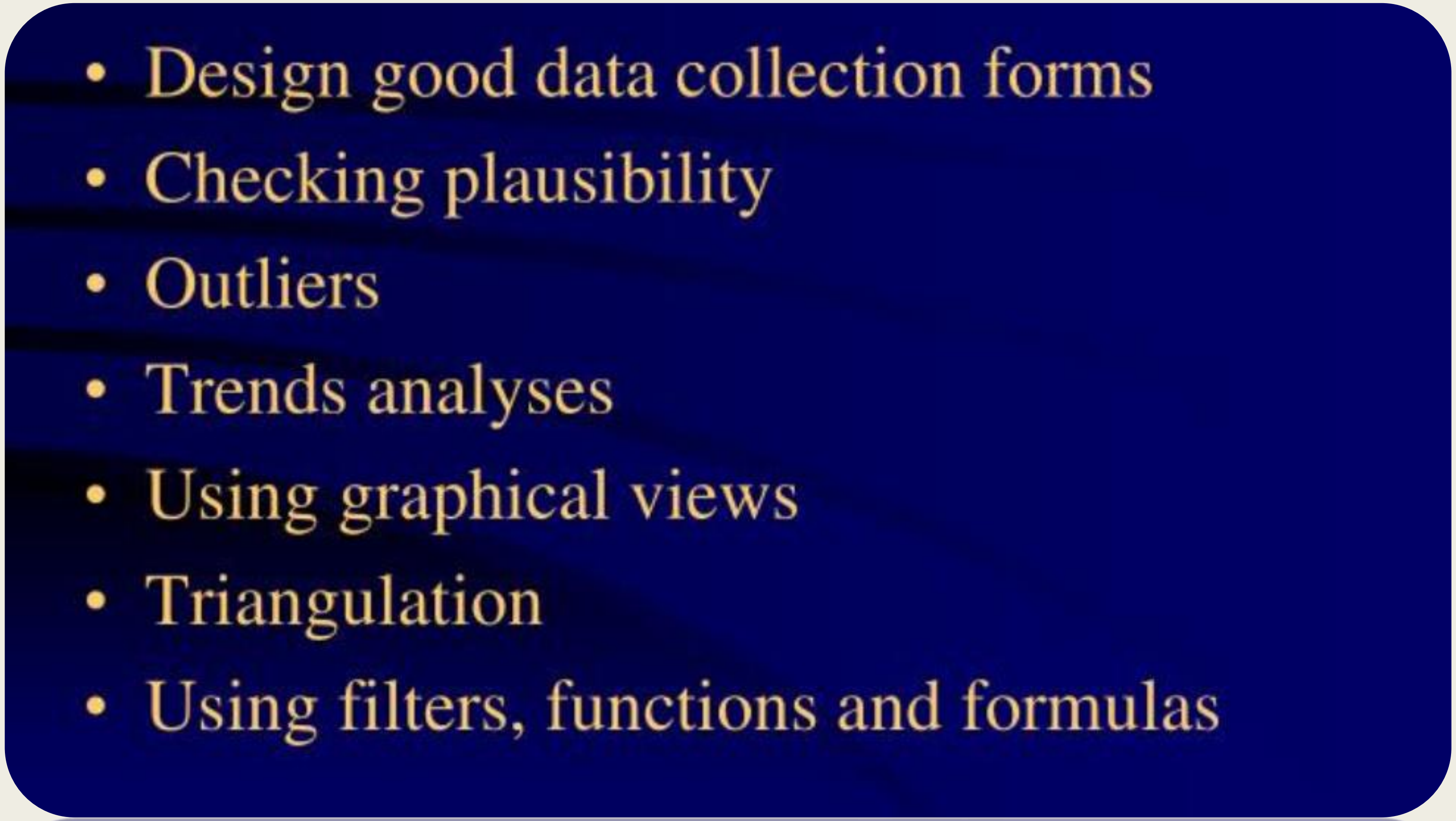
How do you clean Data?



Be creative!

- Lookup functions
  - Easy to find non-existing codes (typos)
- Formulas
  - Check of mathematical and logic consistency
- Compare with other sources (Triangulation)
  - Validation of values/expected ranges (do we have approximately the same)
- Compare with previous years
  - Validation of values/expected ranges (do we have approximately the same)

Data Cleaning : Some Tips



- Design good data collection forms
- Checking plausibility
- Outliers
- Trends analyses
- Using graphical views
- Triangulation
- Using filters, functions and formulas