The data preparation process

SHARP SIGHT

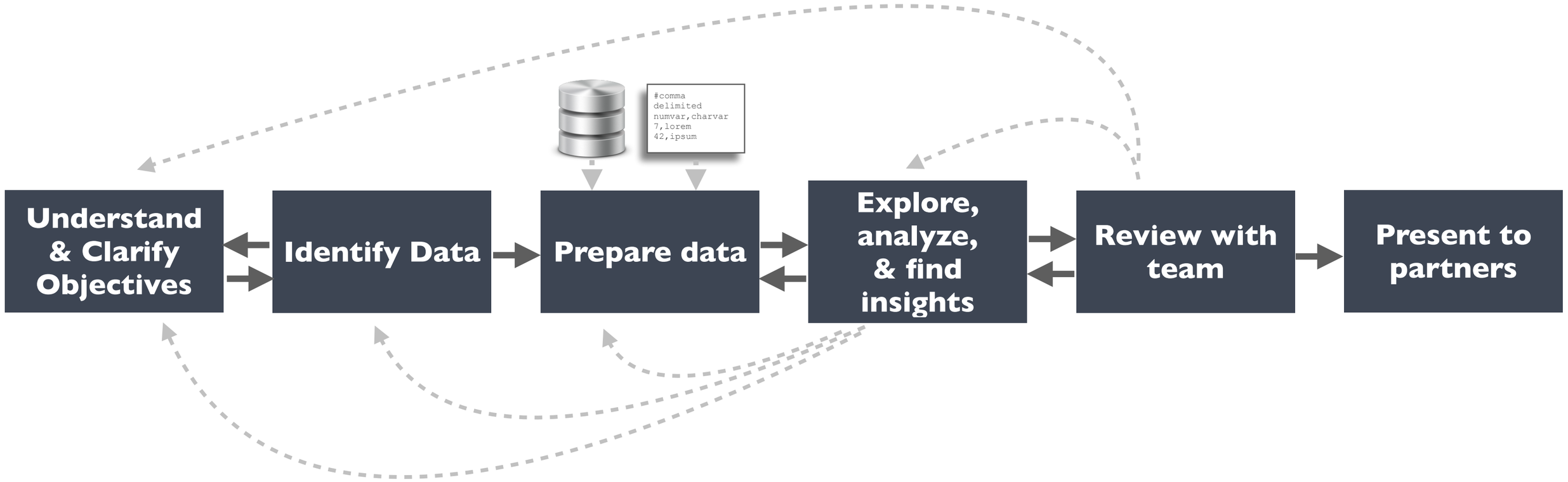
What you'll learn

- The process of data preparation
  - How to approach preparing your data for analysis

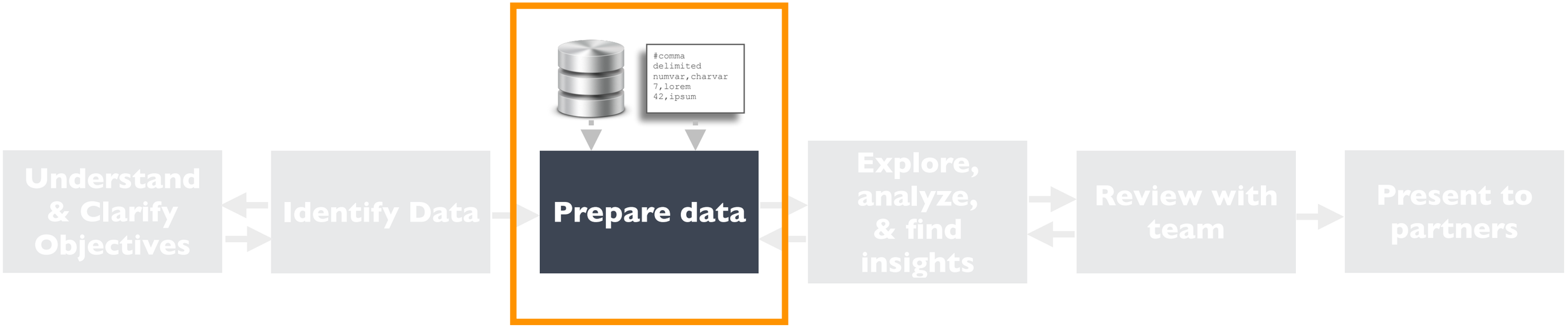
- But keep in mind that the best way to learn this, is to:
  - Read other people's code
  - Work on projects

Data preparation in the broader data science process

The data science process



Data preparation is one of the steps in the data science process



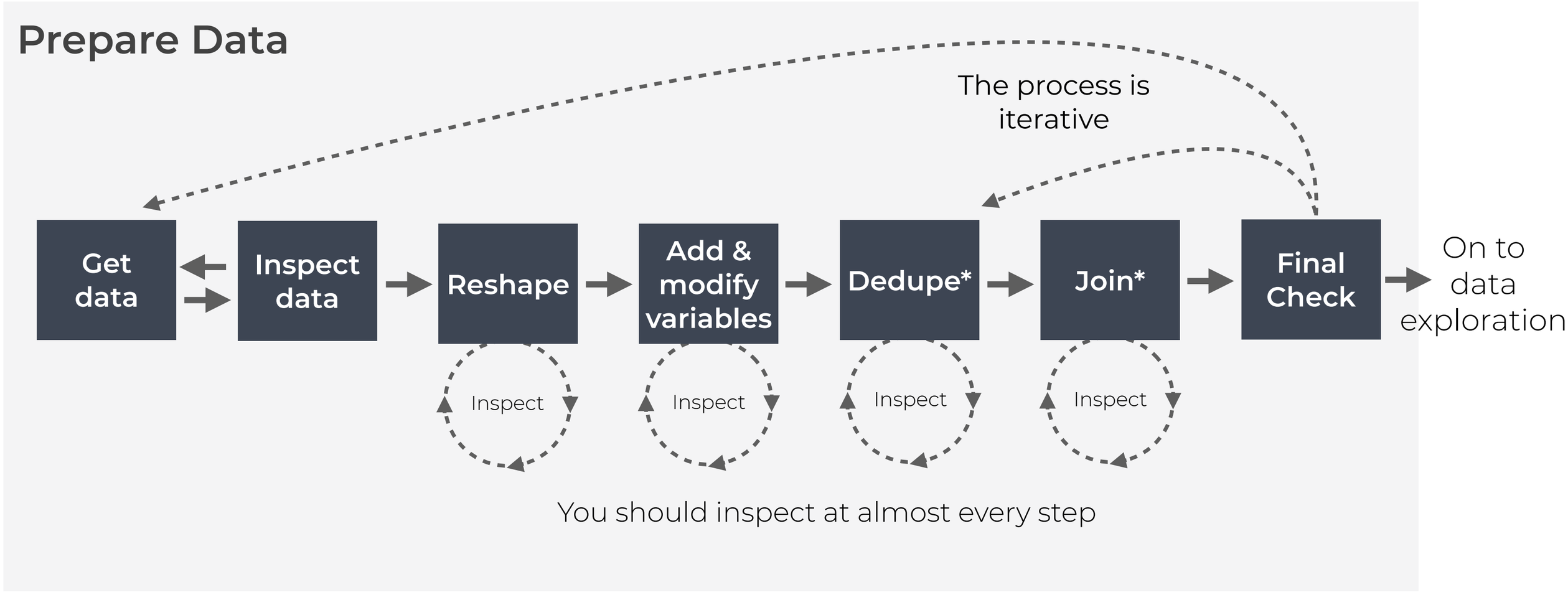
Data preparation: an overview

The data preparation process

This step has sub-steps ...

We're going to zoom in to examine the process more deeply

The data preparation process



* You'll dedupe and join only if you have multiple source files

Data preparation, step-by-step

Get data

- You need to obtain your data from one of several sources:
  - Flat files: read them in with `read_csv`
  - Databases: query the database with SQL (if you use SQL)
  - Scraping: Python has data scraping tools that you can use

- You need to read you data into Pandas DataFrames

Inspect data

- Inspect all of your source files:
    1. Do the files contain the data you need?
    2. Do the records look "OK"?
    3. Are categorical values correct?
    4. Are dates formatted properly?
    5. Is data in "tidy" shape?

- Data inspection is also done in the "data exploration" phase
    - When you begin plotting your data, you may find things out of place, unusual features, etc

Reshape data

- You want your data in "tidy" format
  - every observation has it's own row
  - every variable has it's own column
  - every value has it's own cell

- Use `pivot` and `melt` to reshape data

Add and modify variables

- Create new variables:
  - Create day, month, year from date data
  - Create "calculated" variables

- Modify variables:
  - Recode categorical values
  - Change variable names
  - etc

Look for duplicate records *

- To join multiple datasets, the "join" variable should be unique

- So, you need to identify duplicates
  - create a counter variable with Pandas `assign`
  - aggregate on the join variable using Python `groupby` and `agg`

Remove duplicate records

- To remove records, we can use the `drop_duplicates` function from Pandas

Join source data files

- Once your source files are prepped, join them
  - use `join` or `merge`

- At every step, make sure to inspect your data
  - Count the number of records
  - Look at the data using `head`

Final data check

- When you're finished with these steps, check the final dataset:
  - Does it have the right number of records?
  - Are the variable names correct?
  - Did we create the new variables we will need?
  - Are the categories coded correctly?

- If you find something amiss, go back to an earlier step

- Note: you may find errors in later phases of the analysis
  - If so, go back and do more data prep

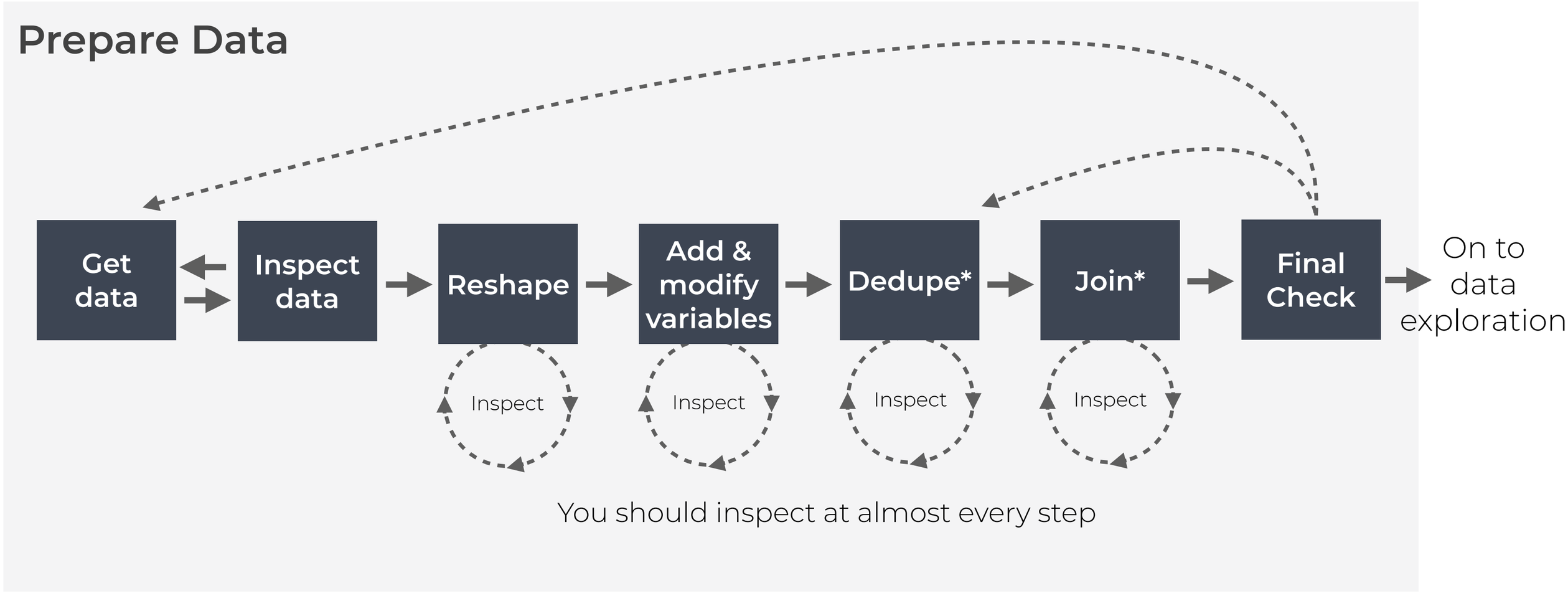
Note: inspect your data throughout the process

- You need to inspect your data constantly

- At every step, make sure that your code executed correctly
  - Did it do what you wanted it to?
  - Is there something else you need to do?

RECAP

The data preparation process



Remember ....

- The process of data preparation is highly iterative
  - check your work
  - if you find something amiss, go back, fix your code, and rebuild file

- The best way to learn data preparation:
  - Read other people's code
  - Work on projects
  - (assuming that you've already mastered the essential functions)