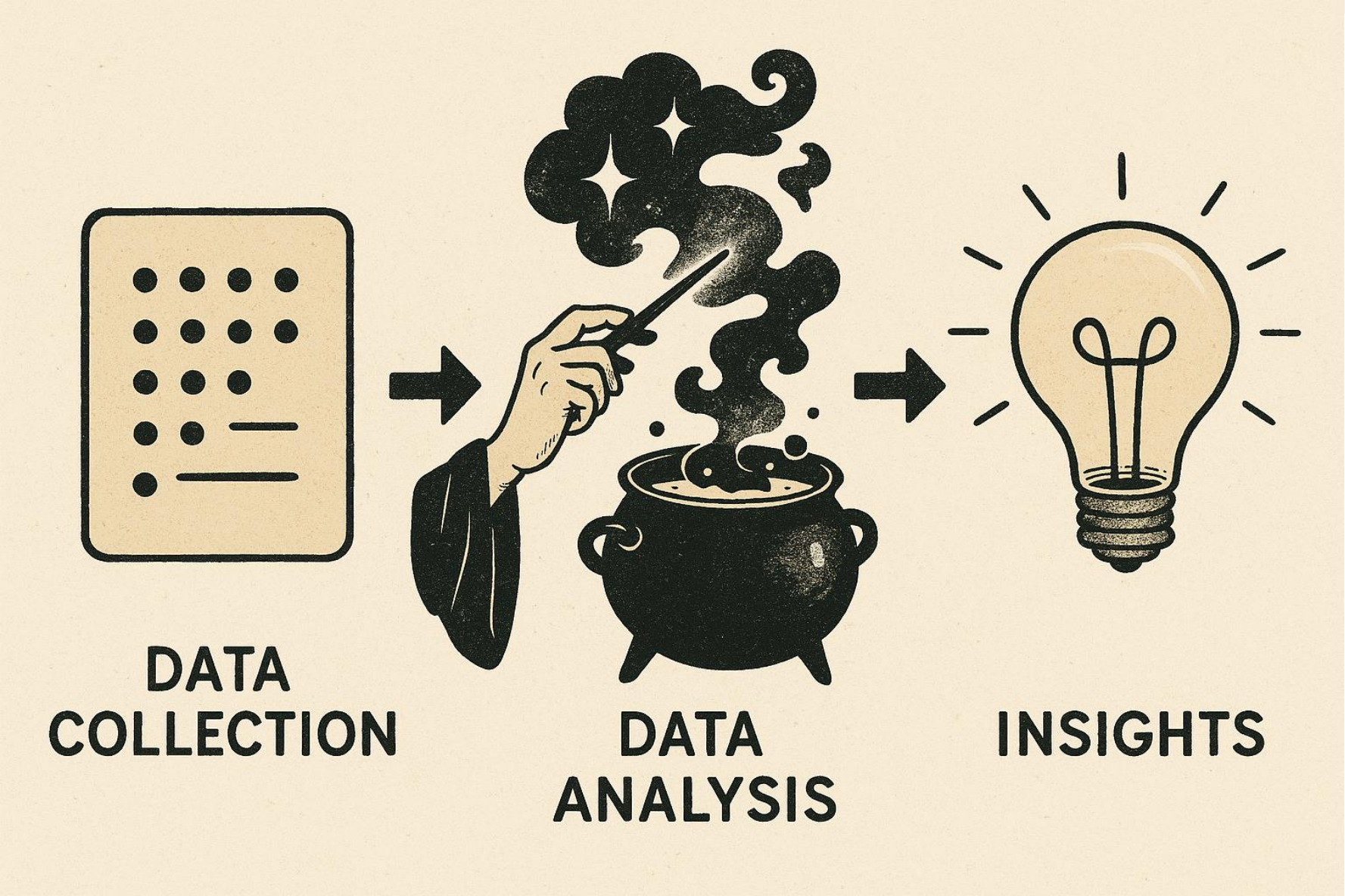
Data Analysis

Tuğrulcan Elmas / Tj

THE UNIVERSITY of EDINBURGH
informatics

Data Analysis Pipeline



THE UNIVERSITY of EDINBURGH
informatics

Data Collection Tips

- Collect as much as you can
  - You can always choose what you need later
- Always collect data
- Document your data
  - How did you collect it e.g., which keywords
  - What is the collection time?
- NEVER OVERWRITE THE RAW DATA

Data Formats: CSV

- Csv: comma separated values
  - bla,bla,bla
- Does not need to be comma separated
  - ; is another common delimiter
- Breaks if the delimiter is in the data
  - tj,30,tugrulcan,elmas – how many columns?
  - tj,30,”tugrulcan,elmas” - 3 columns
- May break due to \n’s, \r’s
  - If the reader does not support quoted multi-lines fields

Data Formats: CSV

- NEVER CREATE MANUALLY
  - NO csv = “tj,30,tuğrulcan,elmas”
  - Need quotes, also encoding may be an issue (wtf is “ğ”?)
- Use wrappers when creating (e.g., csv library for Python)
- Convert from other data types
  - E.g., create a pandas dataframe then save to csv
- If in doubt, save in JSON
  - Less likely to break

Data Formats: JSON (APIs Favourite)

{"name":"tj","location":"Edinburgh","friend":{"name":"Björn","location":"Edinbur
}}

{
  "name": "tj",
  "location": "Edinburgh",
  "friend": {
    "name": "Björn",
    "location": "Edinburgh"
  }
}

- Use https://jsonlint.com to ”prettify”
- Again, never create manually
  - but it’s hard to do so anyway

THE UNIVERSITY of EDINBURGH informatics

Data Analysis

- Learn Fundamentals by coding
  - Do not learn coding
- AI can code for you
  - But AI cannot design for you

Too much to learn?

pandas Getting started User Guide API reference Development Release notes Search ⌘+K 2.2 (stable)

User Guide > Reshaping...

Reshaping and pivot tables

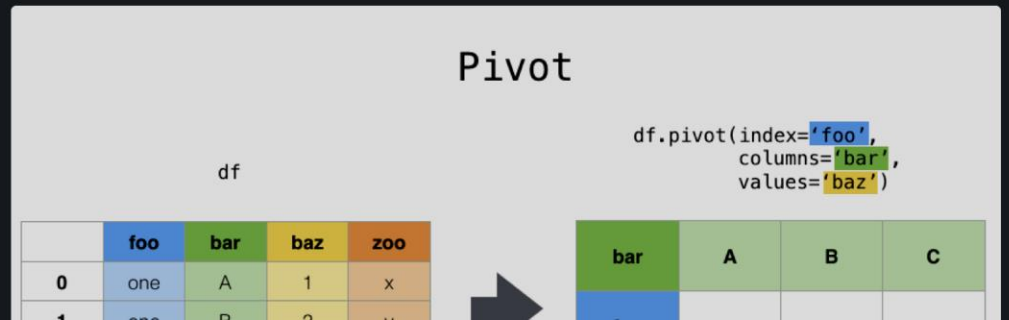
pandas provides methods for manipulating a `Series` and `DataFrame` to alter the representation of the data for further data processing or data summarization.

- `pivot()` and `pivot_table()`: Group unique values within one or more discrete categories.
- `stack()` and `unstack()`: Pivot a column or row level to the opposite axis respectively.
- `melt()` and `wide_to_long()`: Unpivot a wide `DataFrame` to a long format.
- `get_dummies()` and `from_dummies()`: Conversions with indicator variables.
- `explode()`: Convert a column of list-like values to individual rows.
- `crosstab()`: Calculate a cross-tabulation of multiple 1 dimensional factor arrays.
- `cut()`: Transform continuous variables to discrete, categorical values
- `factorize()`: Encode 1 dimensional variables into integer labels.

On this page

Show Source

`pivot()` and `pivot_table()`



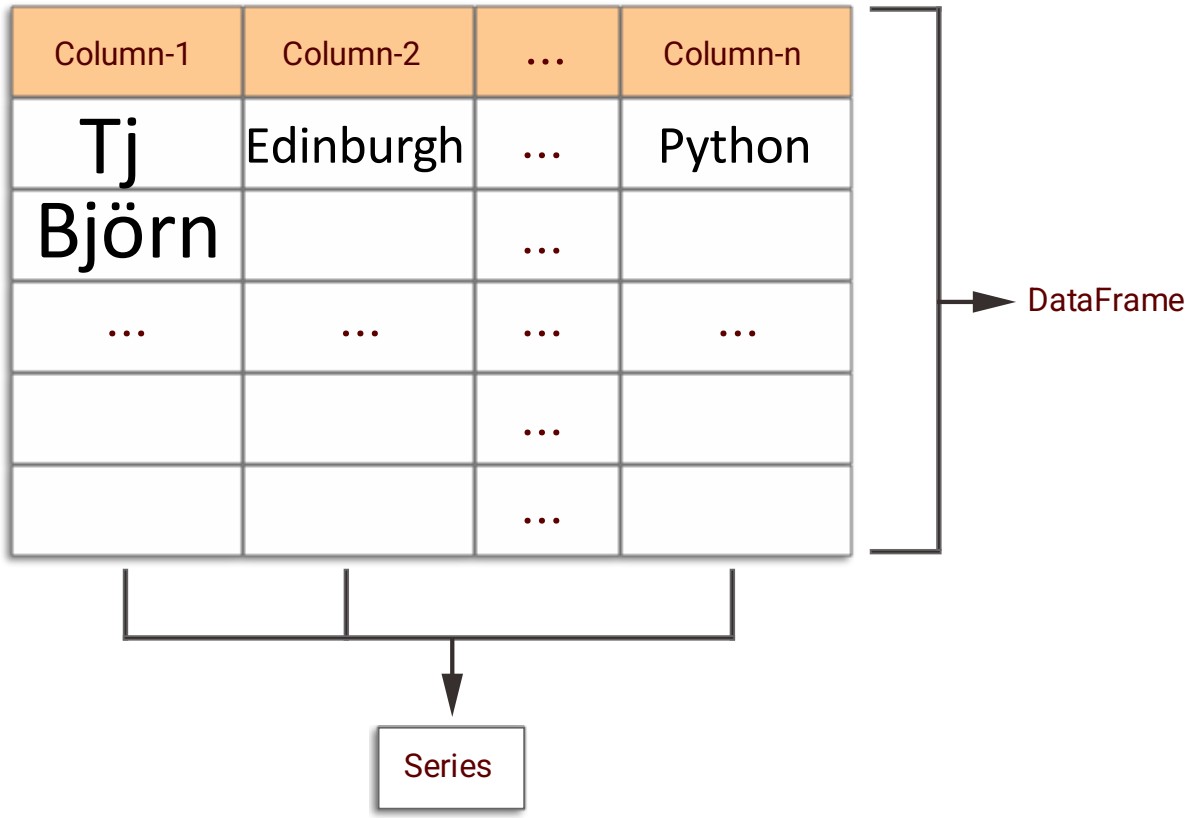
THE UNIVERSITY of EDINBURGH informatics

Data Structure

Column-1	Column-2	…	Column-n
Tj	Edinburgh	…	Python
Björn		…	
…	…	…	…
		…	
		…	

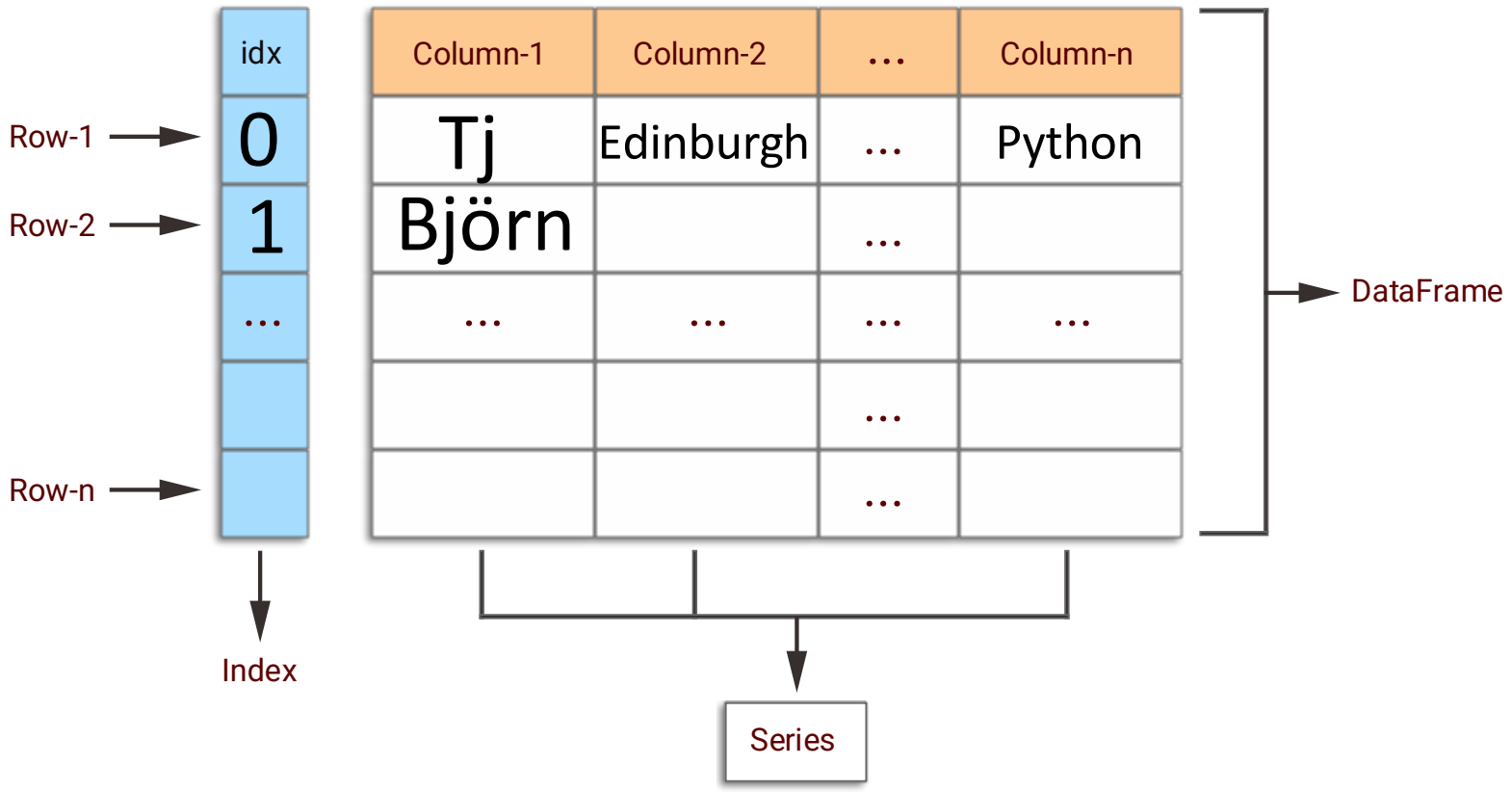
Data Structure

Pandas Data structure



Data Structure

Pandas Data structure



R does not have index, but R library ”data.point” has it

Data Partitioning

Name	City	Country
Tj	Edinburgh	UK
Björn	Edinburgh	UK
Donald Trump	DC	USA

Name	City
Tj	Edinburgh
Björn	Edinburgh
Donald Trump	DC

City	Country
Edinburgh	UK
DC	USA

Data Partitioning

Id	Text	Like count
123123	SICSS is fun!	10
34214	SICSS is fun!	5
234324	SICSS is yay	3

Id	Text
123123	SICSS is fun!
34214	SICSS is fun!
234324	SICSS is yay

Id	Like count
123123	10
34214	5
234324	3

Merging Tables

On Pandas

- Join operation if the common column is the index
  - Faster
- Merge operation if not the index

On R

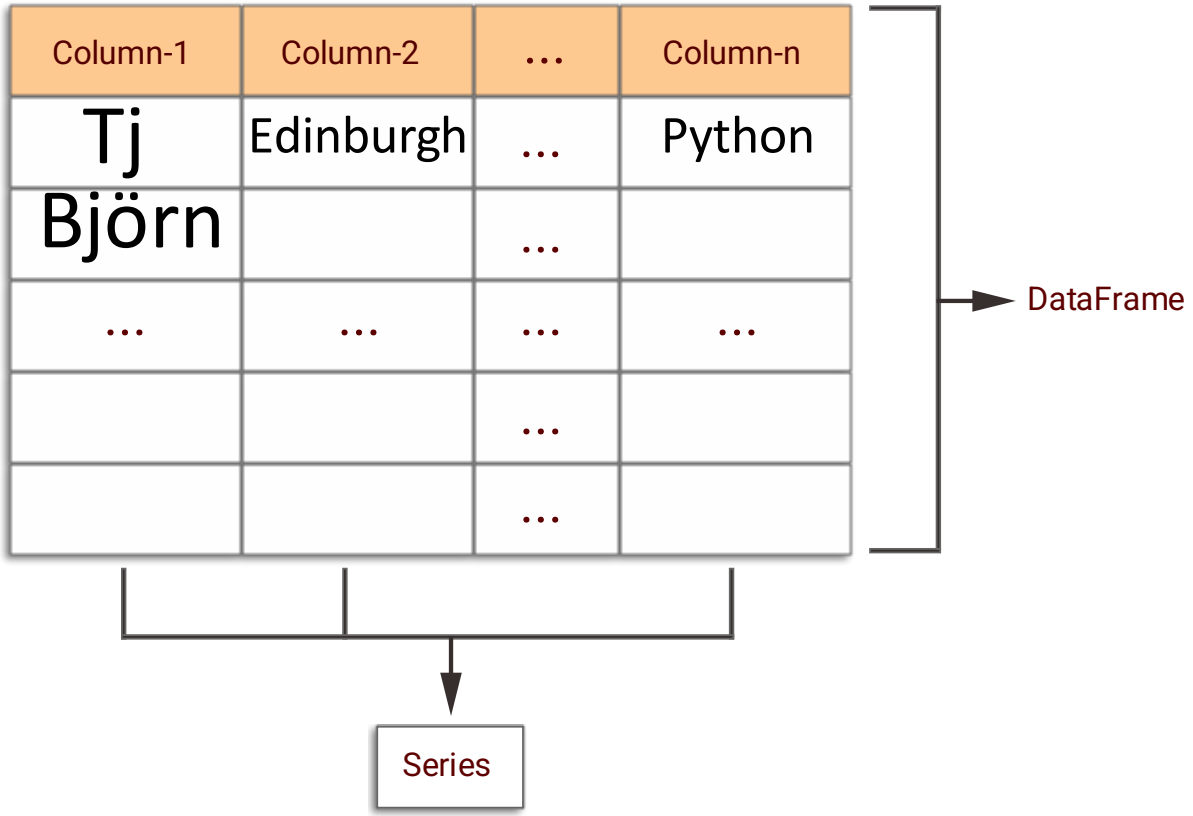
- Use merge but if speed is concern:
- Dplyr package has join operations
- data.point -> set key -> dataframe1[[dataframe2]

Filtering

Both Pandas & R:

- No for loops!
- dataframe[filtering logic]

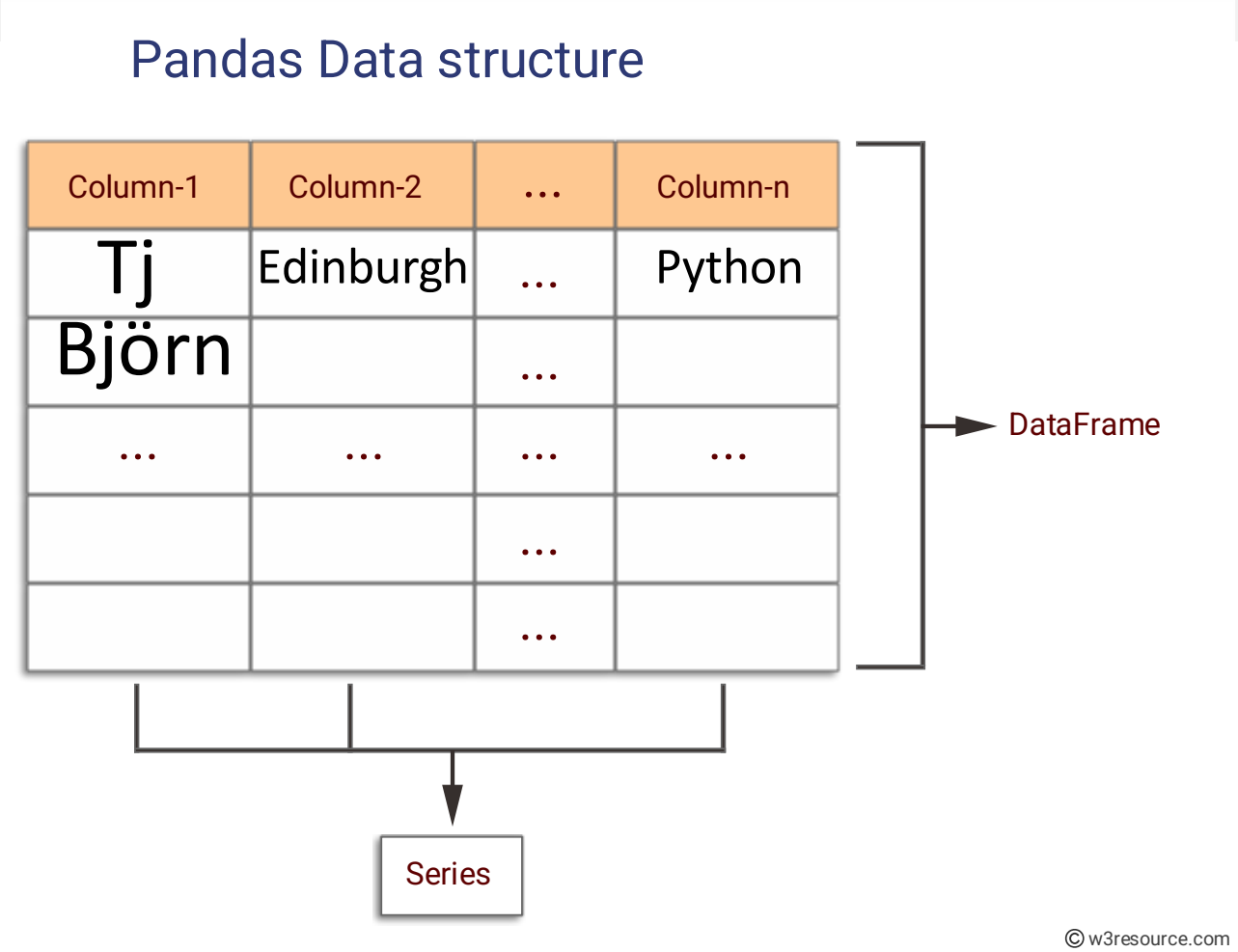
Pandas Data structure



Modifying Values

Both Pandas & R:

- No for loops!
- Get me the rows with some filtering logic, then I modify a column
- dataframe.loc[filtering logic, “column”] = newvalue
- df$column[filtering logic] <- newvalue



Data with missing values

- Drop them
- Or impute
  - By mean or nearest neighbour
- THEY BROKE INTEGERS IN PANDAS!
  - Fix them beforehand or read long integers as strings

Tips

- Do not work with big tables
  - Big if bigger than 500mb uncompressed
  - Partition data into multiple tables
- Work on a small sample initially
  - Partition latter if you like

Visualization

- Matplotlib for Python
  - BUT already embedded in Pandas
  - Seaborn for pretty plots
- R has native support
- Use tools? Tableau
- Use AI?
- Important: keep the code or software files
  - You will reuse them a lot

THE UNIVERSITY of EDINBURGH

informatics