University of Essex

Pandas data input/output

Akitaka Matsuo

Pandas

University of Essex

- Data storage and manipulation solution for Python
  - Using *DataFrame* object
  - Read/write data
    - Understand most of common data formats
    - Read them into DataFrame (then do this and that in Pandas)
  - Data manipulation
    - Sophisticated indexing
    - Data wrangling
      - Reshaping/pivoting (wide-to-long conversion)
      - Missing data handling
      - Combining/merging datasets

Various data formats

University of Essex

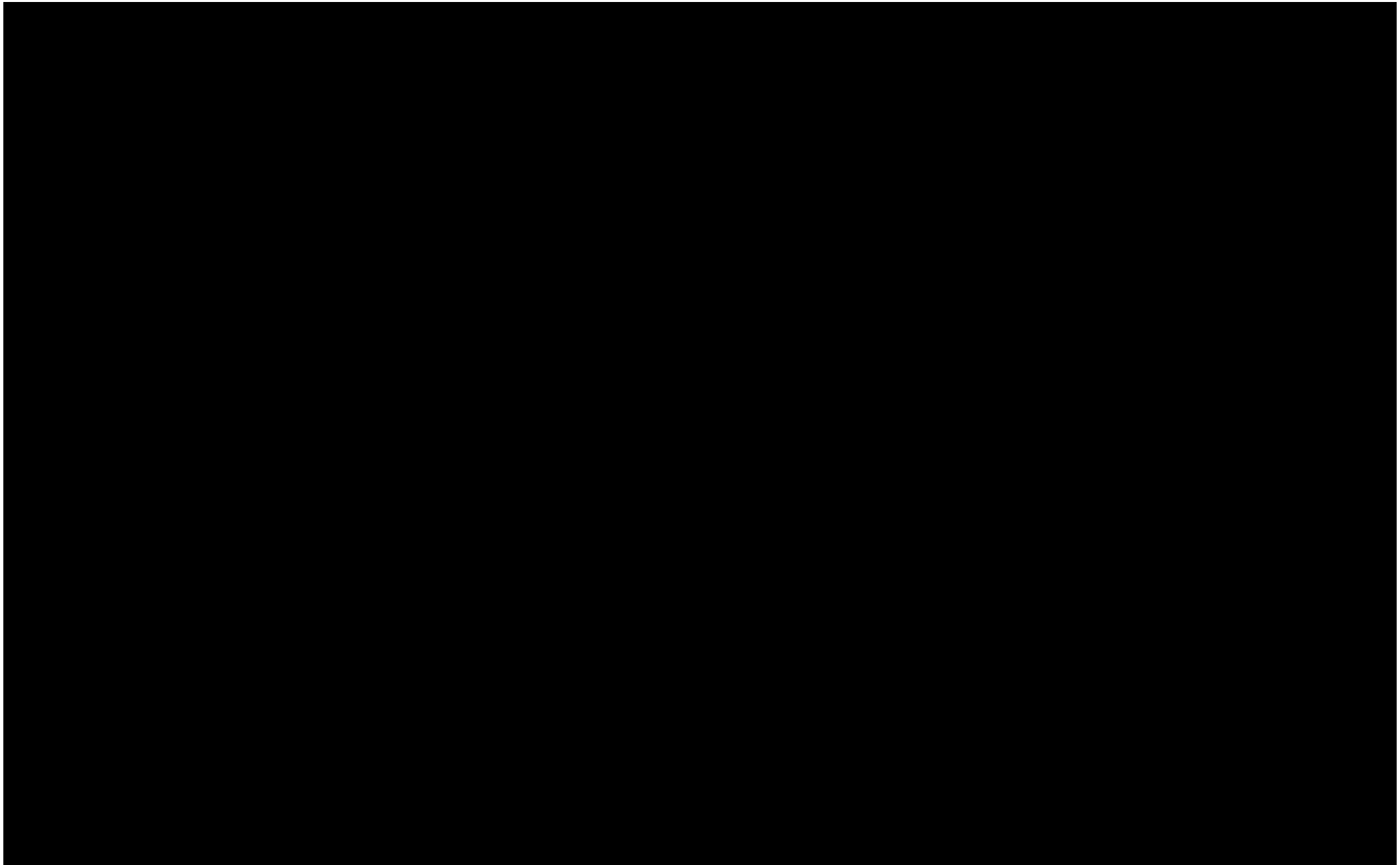
- There are various data formats you have to work with
- Typical examples:
  - csv: comma-separated value
    - Text-based, the most common data format for distributing data
  - xlsx: Excel file
  - Statistical software specific data files
    - sps (SPSS), dta (STATA), rda (R)
  - Common data formats on the Internet
    - html, xml, json
  - Database
    - SQL
- Most of them can be read by Python via pandas, using read_**()

Getting the data files on Colab

University of Essex

- There are multiple ways to do that
  - Using terminal command
    - `!wget`
      - Direct download from the web
    - `!git clone`
      - Get the data placed on github repository
  - Using Google Drive
    - Colab has access to Google Drive
      1. Download file to your computer
      2. Upload on Google Drive
    - Simplest, but sometimes a bit inefficient

Pandas I/O demo



University of Essex

University of Essex