

Python Fundamentals.

Voor Data Scientists en Data Engineers

wortell

Onderwerpen

- Welkom, introductie, kennismaking
- Hoogover uitleg: wat is Python? Wat is Pandas?
- Jupyter Notebook
- Inlezen
- Inspectie en verkennen van data
- Selectie en filteren
- Data wrangling
- Joins
- Visualisatie
- Eigen functies
- Installatie en opzet van Jupyter, Python en Pandas



wortell

Introductie.

Waar sta je nu?

- Waar werk je?
- Welke software?
- Gebruik je al Python?



Wat is je leerdoel?

wortell

Waarom Python + Pandas?

Eating the World!

- Leesbare code
- Snel te schrijven
- Efficiënte libraries

Data-analyse, ML, wrangling

- Maar ook handig om snel een API aan te spreken
- .. of een website te scrapen

Spark, Databricks, ...

- Koalas / PySpark

Leuk om iets nieuws te leren

Waarom **GEEN** Python / Pandas?

Higher-level language

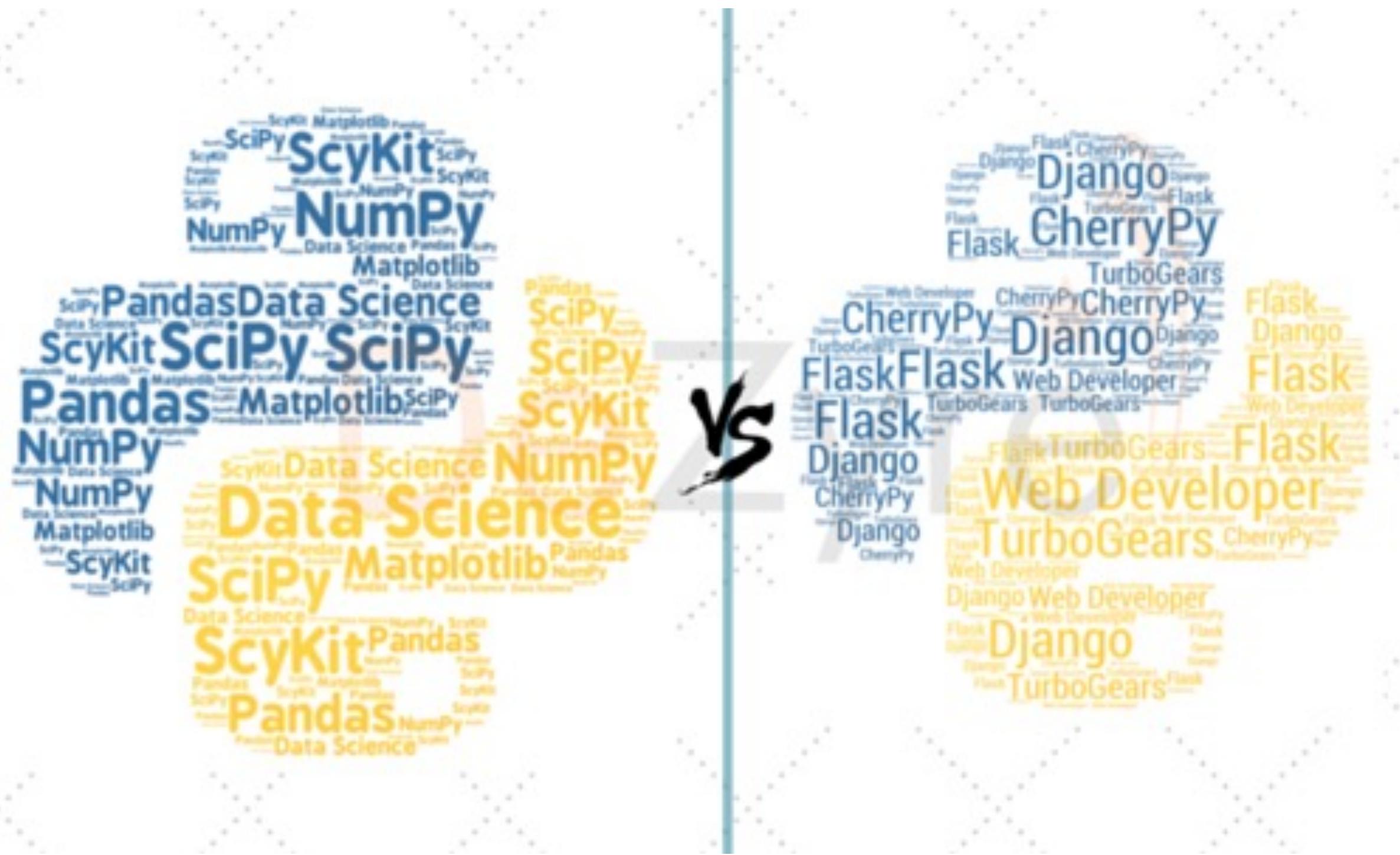
- Dus ook langzamer
- ... "premature optimization is the root of all evil" (Donald Knuth)

Pandas: alles in-memory

- Voor echt grote datasets gedistribueerde oplossingen (Spark / Dask)

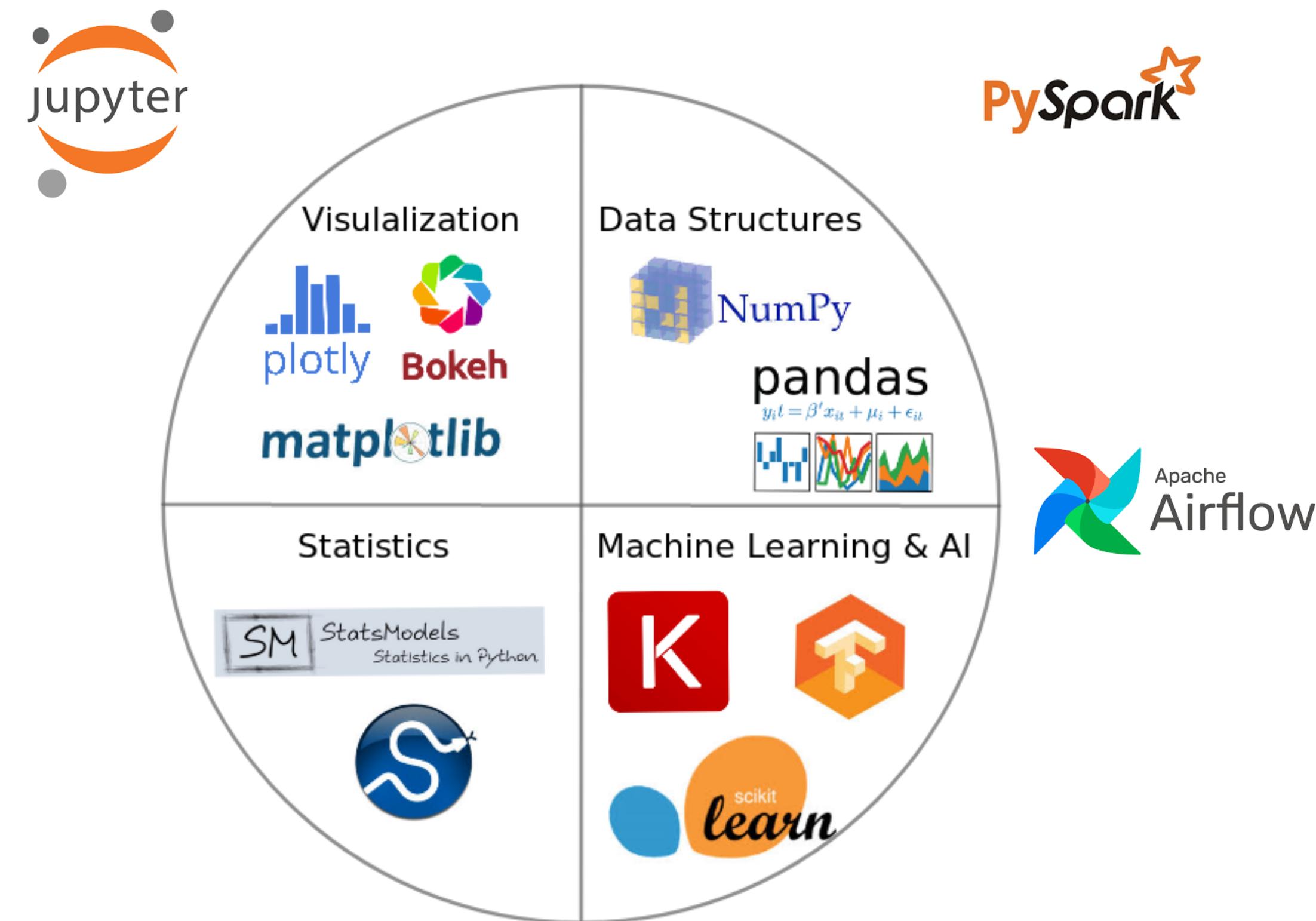
Niet geschikt voor mobile

Wegen naar Python



wortell

Het Python Data-ecosysteem



Jupyter

- Intro Jupyter
- Jupyter Notebook, Jupyter Lab, VS Code, Google Colab, ...

A screenshot of a Jupyter Notebook interface. The left sidebar shows a file tree with notebooks like 'Lorenz.ipynb' (seconds ago), 'Data.ipynb' (an hour ago), 'Fasta.ipynb' (a day ago), 'Julia.ipynb' (a day ago), 'R.ipynb' (a day ago), 'iris.csv' (a day ago), 'lightning.json' (9 days ago), and 'lorenz.py'. The main area displays code and output. A text cell explains the Lorenz system, and a code cell runs 'lorenz.py' which includes sliders for sigma, beta, and rho, and a plot of the Lorenz attractor.

A screenshot of Jupyter Lab. It shows a code editor with 'machineLearning.py' containing data loading and plotting code. Two side panes are visible: 'imports' listing pandas, numpy, and matplotlib imports, and 'graph' showing two scatter plots of wine quality data labeled 'Red Wine' and 'White Wine'.

A screenshot of Google Colab. The notebook title is '01_Reading_Data_and_Inspecting_It.ipynb'. The first cell contains text about reading data files and inspecting dataframes. The second cell shows code for reading CSV files and creating scatter plots for red and white wine quality data.

Eerste stappen met Jupyter in Colab

The screenshot shows the Google Colab interface. At the top, there's a header bar with a lock icon, the URL <https://colab.research.google.com>, and icons for A, a document, and a star. Below the header is a dark sidebar with text like "Home To Colaboratory", "View", "Insert", "Runtime", "Tools", "Help", and "Contents". The main area has tabs for "Code" and "Text", and a "Copy to Drive" button. A navigation bar at the top of the main content includes "Examples", "Recent", "Google Drive", "GitHub" (which is underlined in orange), and "Upload". Below this, a search bar asks "Enter a GitHub URL or search by organization or user" with an "Include private repos" checkbox. A search input field contains "wortell-smart-learning" and a magnifying glass icon. Underneath, there are dropdowns for "Repository:" set to "wortell-smart-learning/python-data-fundamentals" and "Branch:" set to "main". A "Path" section is below. Two files are listed: "00_Exercise_Python_and_Jupyter_notebook.ipynb" and "00_Jupyter_Notebooks.ipynb", each with a magnifying glass and a copy icon.

Datatypes in Python

Name	Type	Description
Integers	int	Whole numbers, such as: 3 300 200
Floating point	float	Numbers with a decimal point: 2.3 4.6 100.0
Strings	str	Ordered sequence of characters: "hello" 'Sammy' "2000" "楽しい"
Lists	list	Ordered sequence of objects: [10,"hello",200.3]
Dictionaries	dict	Unordered Key:Value pairs: {"mykey": "value", "name": "Frankie"}
Tuples	tup	Ordered immutable sequence of objects: (10,"hello",200.3)
Sets	set	Unordered collection of unique objects: {"a","b"}
Booleans	bool	Logical value indicating True or False

Jupyter: tips & tricks

- Shift + Enter to run code
- Tab completion
- Nieuwe cell: Escape gevolgd door a (above) of b (below) of dd (delete)
- Shift Tab to see arguments and information about methods, functions or classes
- Magic commands, such as ls
- ? or ?? to get extra help and info



Pandas intro

- Data inlezen `pd.read_csv()`
- Data inspectie `df.info()` `df.head()` `df.describe()`
- Data selectie `df[df.column == 'value']` `df.loc[df.column == 'value', :]`
- Data wrangling `df['column'].fillna()` `df.drop_duplicates()`
- Data joinen `df.merge(df2, how='inner', on='column_name')`
- Data visualisatie `px.scatter(df, x, y)`

pandas

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,
built on top of the [Python](#) programming language.

wortell

Pandas: the DataFrame (df)

The diagram illustrates a Pandas DataFrame representing NBA player statistics. The columns are labeled: Name, Team, Number, Position, Age, Height, Weight, College, and Salary. The rows are indexed from 0 to 6. A purple box highlights the first row (index 0). A blue arrow points to the 'Name' column header. A green box highlights the entire row at index 2. A red box highlights the 'Age' value for player 3 (Jordan Mickey), which is 'NaN'. An orange box highlights the 'Weight' value for player 5 (Jared Sullinger), which is '190.0'. A pink box highlights the 'Age' value for player 6 (Evan Turner), which is '27.0'. A purple arrow points to the 'Index label' at index 0. A green arrow points to the 'Index axis=0' at index 0. A blue arrow points to the 'Columns axis=1'. A red arrow points to the 'Missing value' 'NaN'. An orange arrow points to the 'Data' value '2569260.0'.

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0	6-8	235.0	LSU	1170960.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN	6-9	260.0	Ohio State	2569260.0
6	Evan Turner	Boston Celtics	11.0	SG	27.0	6-7	220.0	Ohio State	3425510.0

Pandas: the Series

	Name	Team	Number
0	Avery Bradley	Boston Celtics	0.0
1	John Holland	Boston Celtics	30.0
2	Jonas Jerebko	Boston Celtics	8.0
3	Jordan Mickey	Boston Celtics	NaN
4	Terry Rozier	Boston Celtics	12.0
5	Jared Sullinger	Boston Celtics	7.0
6	Evan Turner	Boston Celtics	11.0

ser = pd.Series(df ['Name'])

ser = pd.Series(df ['Team'])

ser = pd.Series(df ['Number'])

DG

wortell

Next slides contains elements you can **copy and paste with the 'format' tool.**

Photo Side.

Big photo covering the side

A photo doesn't always have to be circular or rounded. It can also contain 50% or less/more on the side as a separator.



Photo Circular.

How do you do this?

To get the circular photo, make sure that you have a square photo and simply copy the formatting of the sample to your required photo. If the photo isn't squared yet, read below on how to do that.

How to crop on 1:1 ratio

- Paste your photo
 - Ideal situation would be a 1:1/square photo
 - If not, you can crop it with the instructions below
- Go to Picture Format > Crop > Square 1:1
- Copy and paste the formatting from the sample on the right to the new one.



wortell

Drop Shadow Element.

Drop Shadow

Voor vlakken achter een tekst, of andere elementen die kadering nodig hebben. (Niet verplicht)



Just a sample

- Bullet 1
- Bullet 2
- Bullet 3

- Just a big bullet
- Another one
 - Sub bullet
 - Another sub bullet