Midis de l’info scientifique

Traitement de données avec Pandas & Jupyter notebooks

Pablo Iriarte / CODIS

14 et 15 mars 2018

Programme

- Introduction
- Installer Jupyter Notebooks et Pandas via la distribution Anaconda

Jupyter Notebooks

- Créer, organiser et partager des notebooks
- Se familiariser avec les notebooks

Pandas

- Importer et exporter des données
- Analyser des données
- Travailler avec différents types de données et des données manquantes
- Manipuler les données
- Créer des graphiques et des visualisations

Introduction

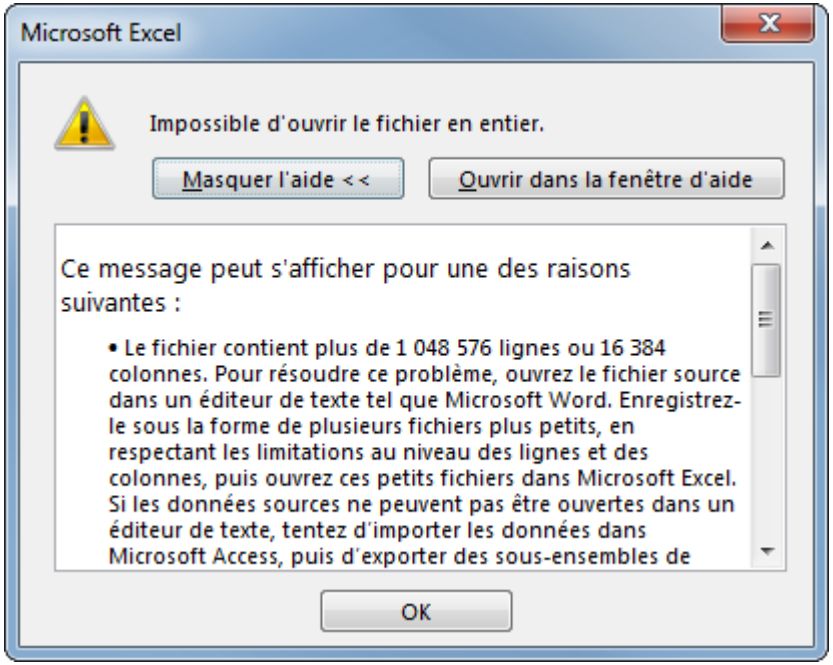
- Excel : limitations
- Excel : erreurs scientifiques
- Reproductibilité et Open Science
- Big Data et Open Data

Introduction

Excel : limitations

Liste complète :

https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3

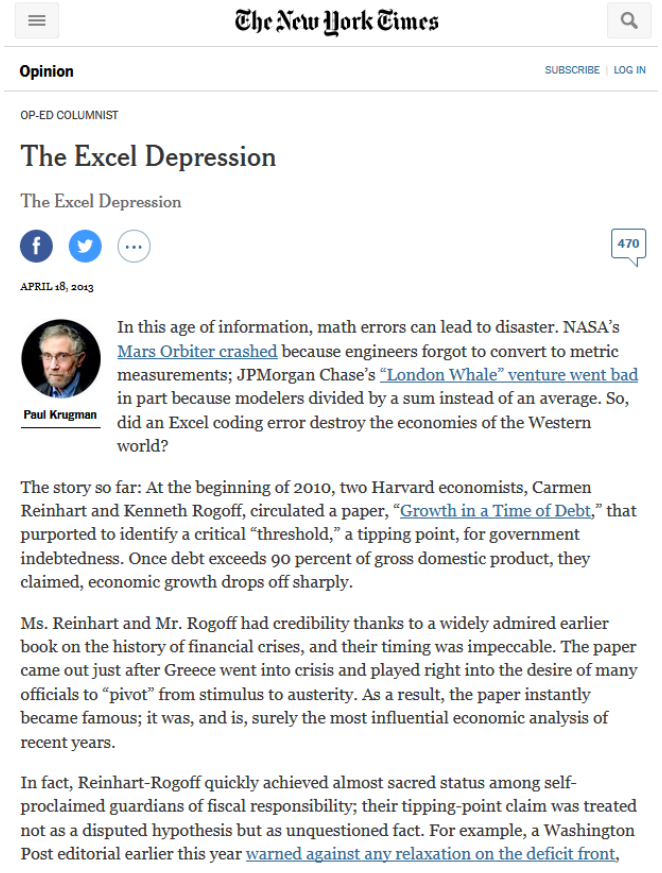


Introduction

Excel : erreurs scientifiques

L'exemple du «Reinhart-Rogoff error»

https://mobile.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html



The New York Times

Opinion

SUBSCRIBE | LOG IN

OP-ED COLUMNIST

The Excel Depression

The Excel Depression

470

APRIL 18, 2013

Paul Krugman

In this age of information, math errors can lead to disaster. NASA's Mars Orbiter crashed because engineers forgot to convert to metric measurements; JPMorgan Chase's "London Whale" venture went bad in part because modelers divided by a sum instead of an average. So, did an Excel coding error destroy the economies of the Western world?

The story so far: At the beginning of 2010, two Harvard economists, Carmen Reinhart and Kenneth Rogoff, circulated a paper, "Growth in a Time of Debt," that purported to identify a critical "threshold," a tipping point, for government indebtedness. Once debt exceeds 90 percent of gross domestic product, they claimed, economic growth drops off sharply.

Ms. Reinhart and Mr. Rogoff had credibility thanks to a widely admired earlier book on the history of financial crises, and their timing was impeccable. The paper came out just after Greece went into crisis and played right into the desire of many officials to "pivot" from stimulus to austerity. As a result, the paper instantly became famous; it was, and is, surely the most influential economic analysis of recent years.

In fact, Reinhart-Rogoff quickly achieved almost sacred status among self-proclaimed guardians of fiscal responsibility; their tipping-point claim was treated not as a disputed hypothesis but as unquestioned fact. For example, a Washington Post editorial earlier this year warned against any relaxation on the deficit front,

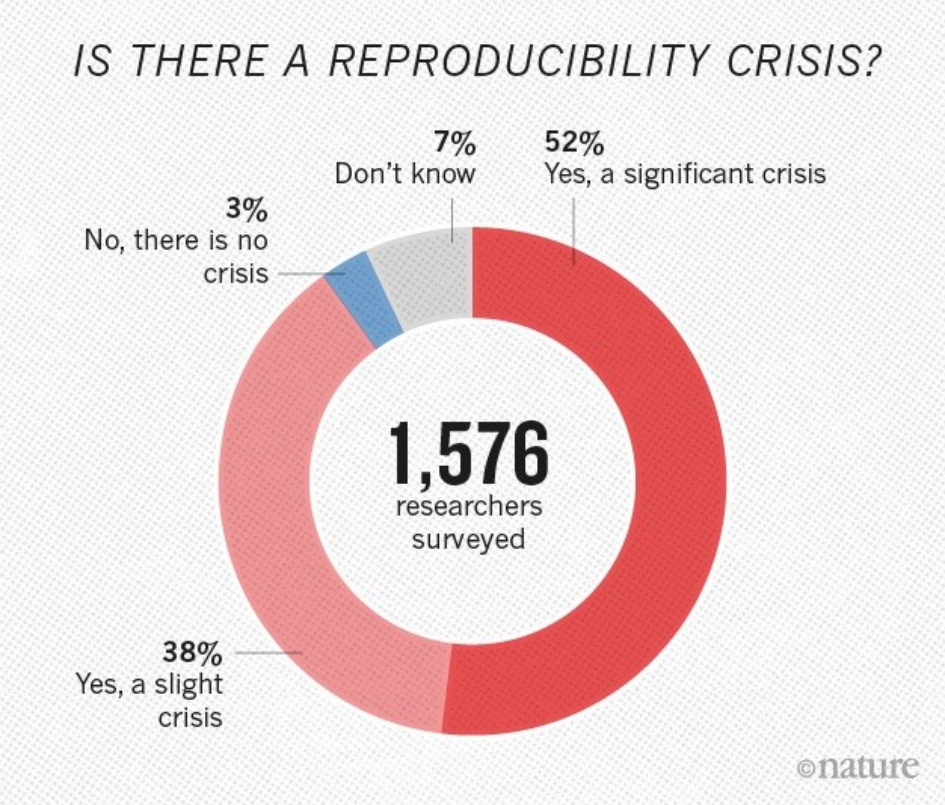
Introduction

Reproductibilité et Open Science

La science en crise?

1,500 scientists lift the lid on reproducibility

https://doi.org/10.1038/533452a

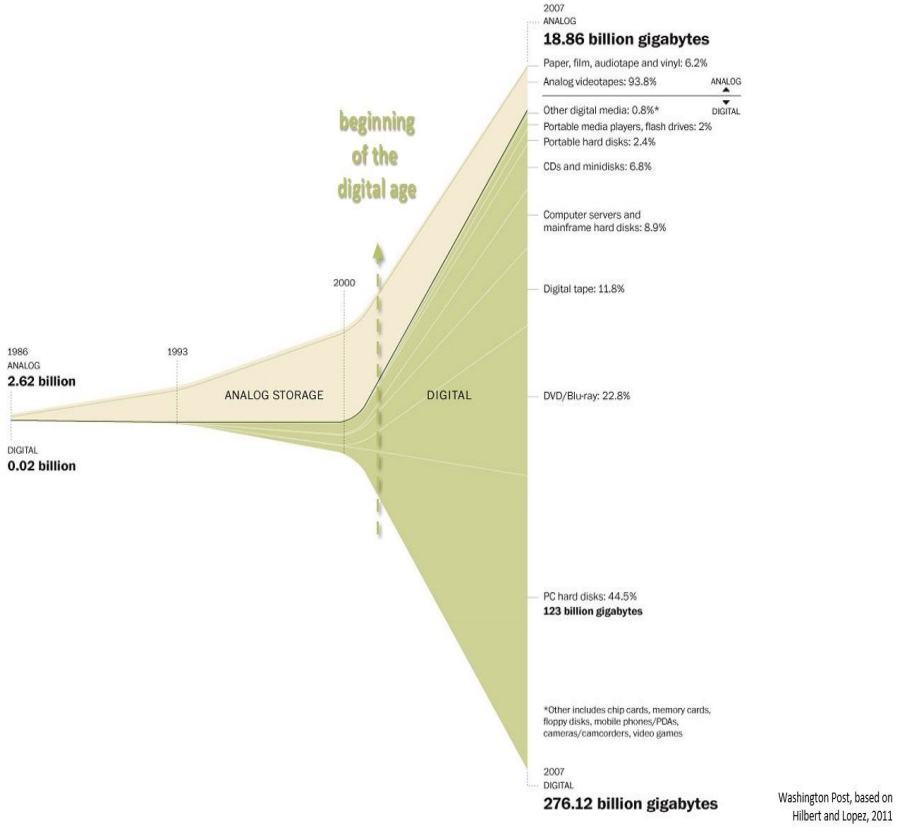


Introduction

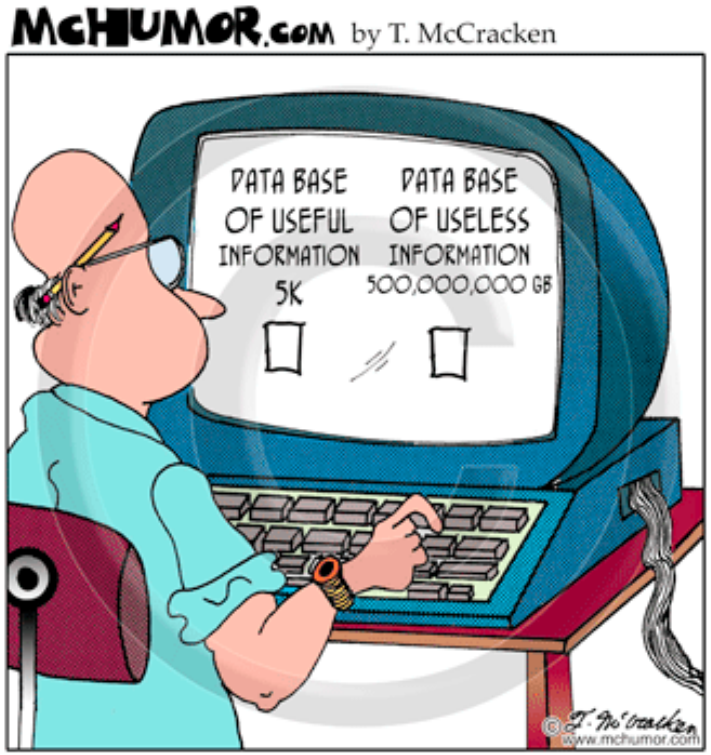
Big Data et Open Data

Quantifying the Data Deluge and the Data Drought
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2984851

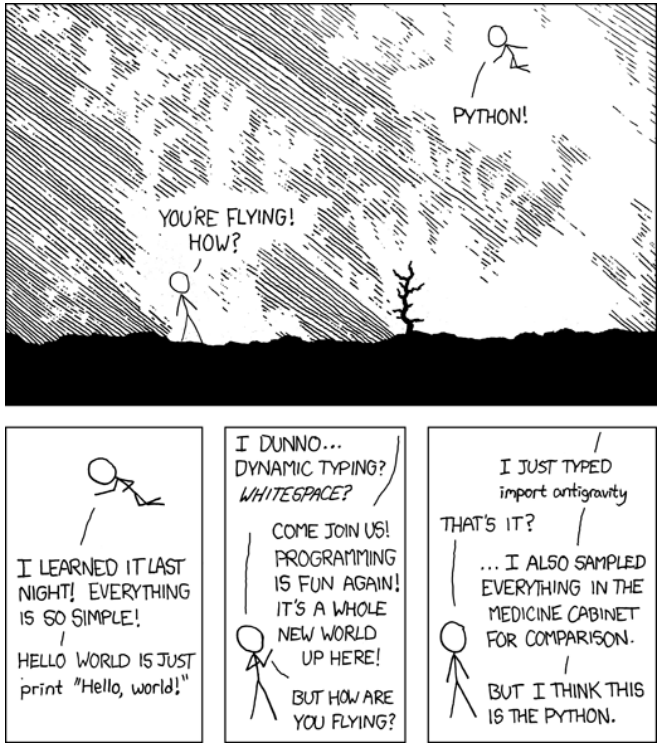
Nombreux réservoirs ouverts
Kaggle : https://www.kaggle.com
Data Hub : http://datahub.io
WikiData : https://www.wikidata.org



Introduction



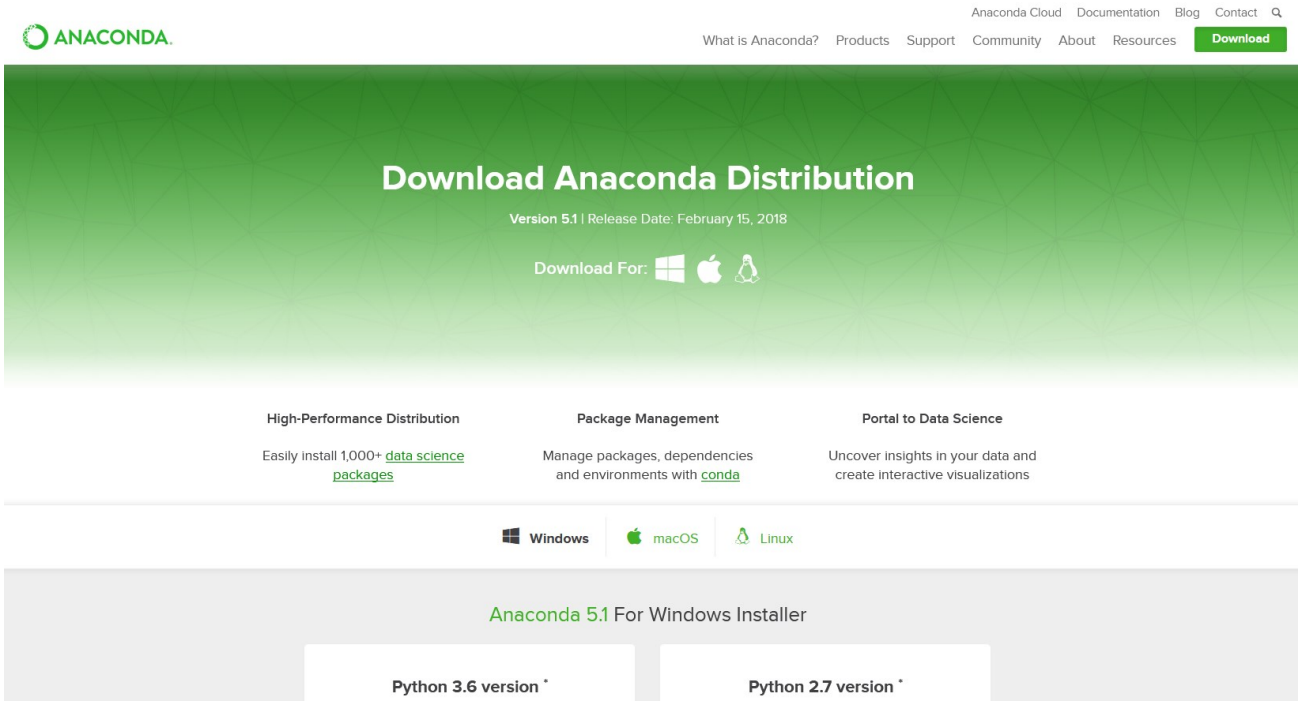
Introduction



https://www.xkcd.com/353/

Installer Jupyter Notebooks et Pandas

Distribution Anaconda : https://www.anaconda.com/download/



Installer Jupyter Notebooks et Pandas

Packages compris dans l'installation :

- Notebook (jupyter)
- Pandas
- NumPy
- Matplotlib
- NLTK
- …

Liste complète :
https://docs.anaconda.com/anaconda/packages/py3.6_win-64

Créer, organiser et partager des notebooks

Lancer Anaconda -> Jupyter Notebook

Se familiariser avec les notebooks

Exercices

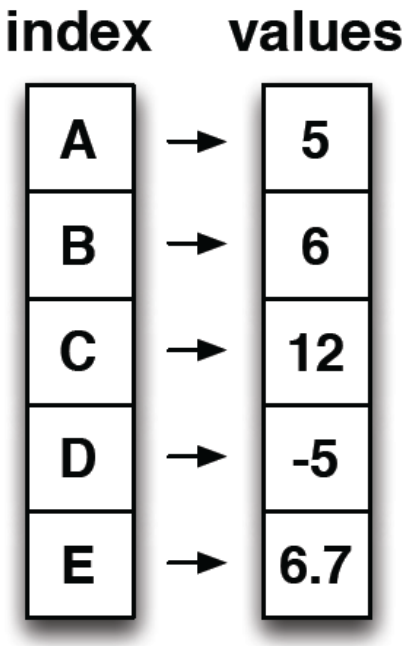
1. Ouvrir un notebook d'exemple (sur le dossier du cours)
2. L'exporter en format HTML
3. Créer un nouveau notebook et le renommer
4. Ajouter une cellule de texte (markdown)
5. Ajouter une cellule de code python (calcul simple)

Aide markdown : https://guides.github.com/features/mastering-markdown/

Aide python : https://www.stavros.io/tutorials/python/

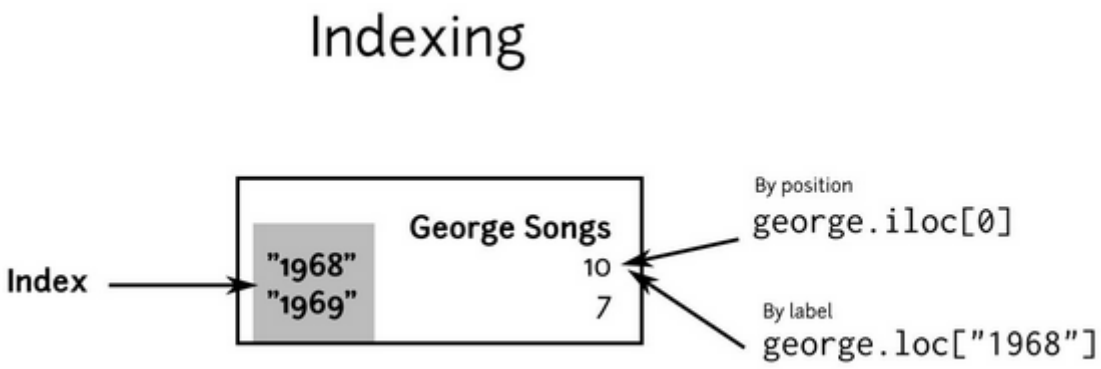
Pandas

Series : 1 dimension



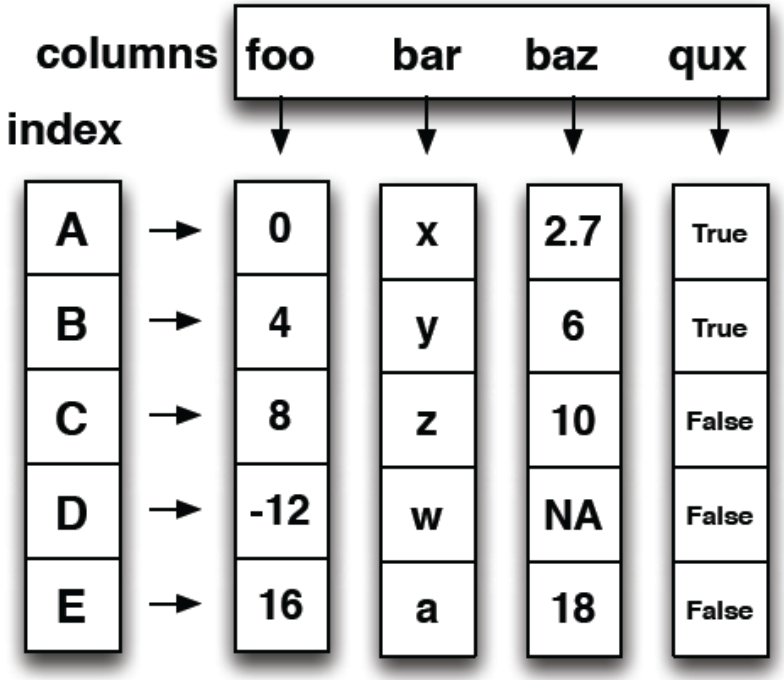
Pandas

Index : afficher des données par la position ou le nom de l'index



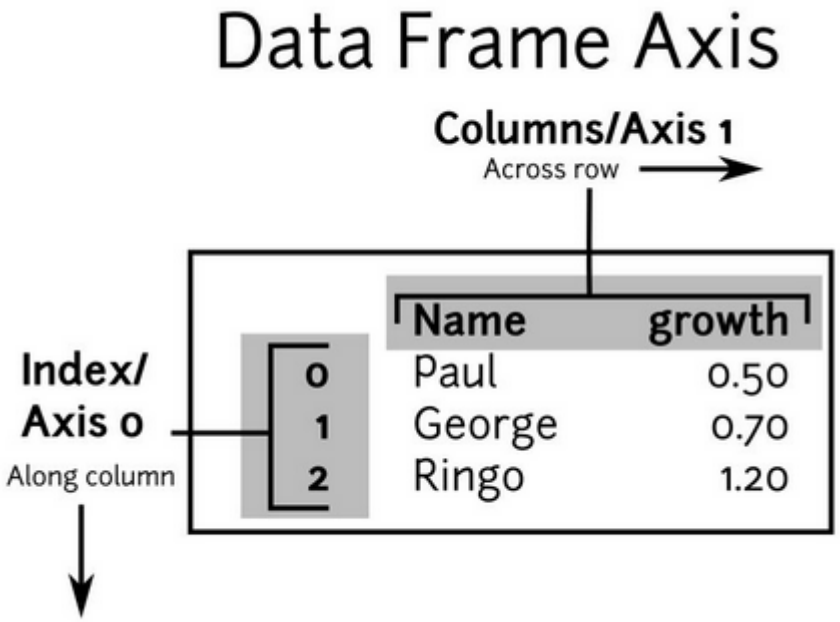
Pandas

DataFrame : 2 dimensions



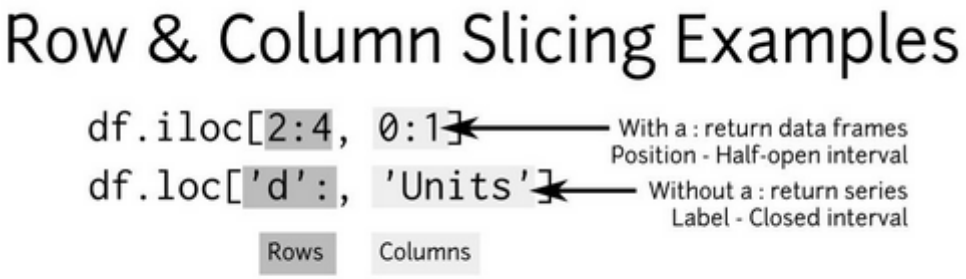
Pandas

DataFrame : axes



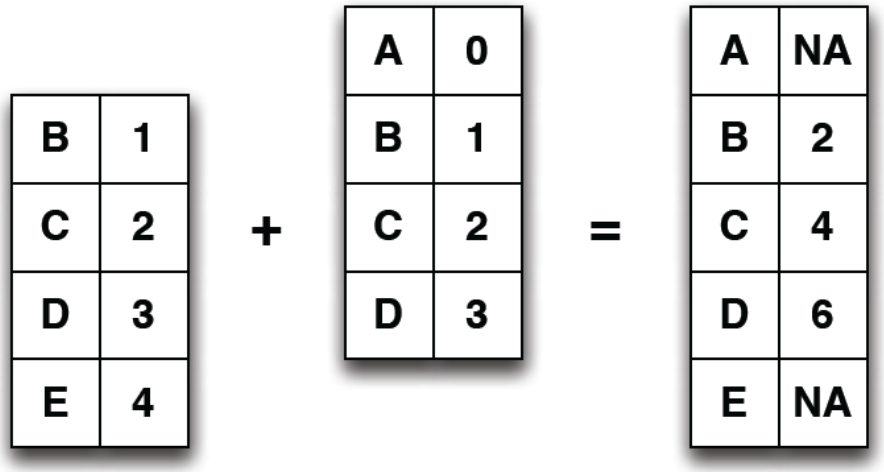
Pandas

DataFrame : slices



Pandas

Opérations facilitées par les index : jointures automatiques



B	1
C	2
D	3
E	4

+

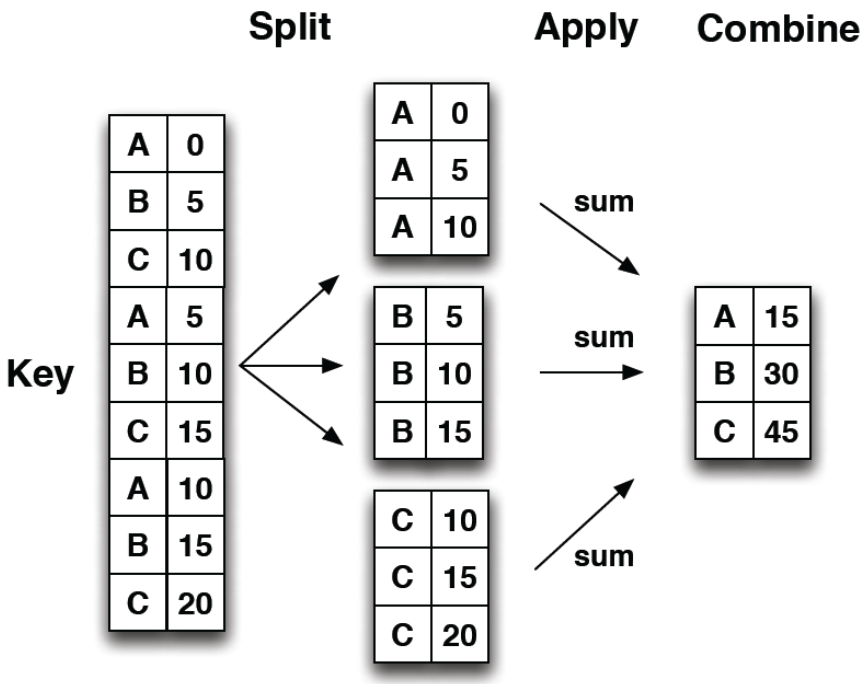
A	0
B	1
C	2
D	3

=

A	NA
B	2
C	4
D	6
E	NA

Pandas

Opérations : GroupBy



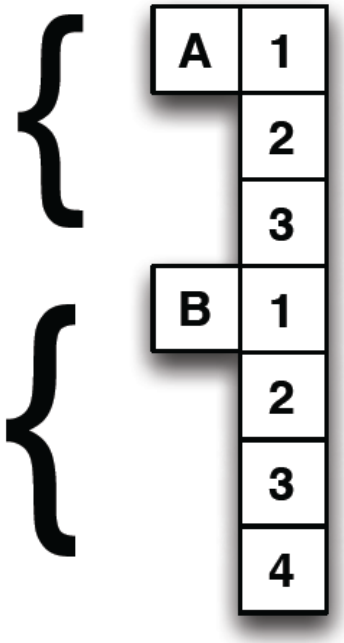
Pandas

Opérations : GroupBy

Method	Result
`.all`	Boolean if all cells in group are `True`
`.any`	Boolean if any cells in group are `True`
`.count`	Count of non null values
`.size`	Size of group (includes null)
`.idxmax`	Index of maximum values
`.idxmin`	Index of minimum values
`.quantile`	Quantile (default of .5) of group
`.agg(func)`	Apply `func` to each group. If `func` returns scalar, then reducing
`.apply(func)`	Use split-apply-combine rules
`.last`	Last value
`.nth`	Nth row from group
`.max`	Maximum value
`.min`	Minimum value
`.mean`	Mean value
`.median`	Median value
`.sem`	Standard error of mean of group
`.std`	Standard deviation
`.var`	Variation of group
`.prod`	Product of group
`.sum`	Sum of group

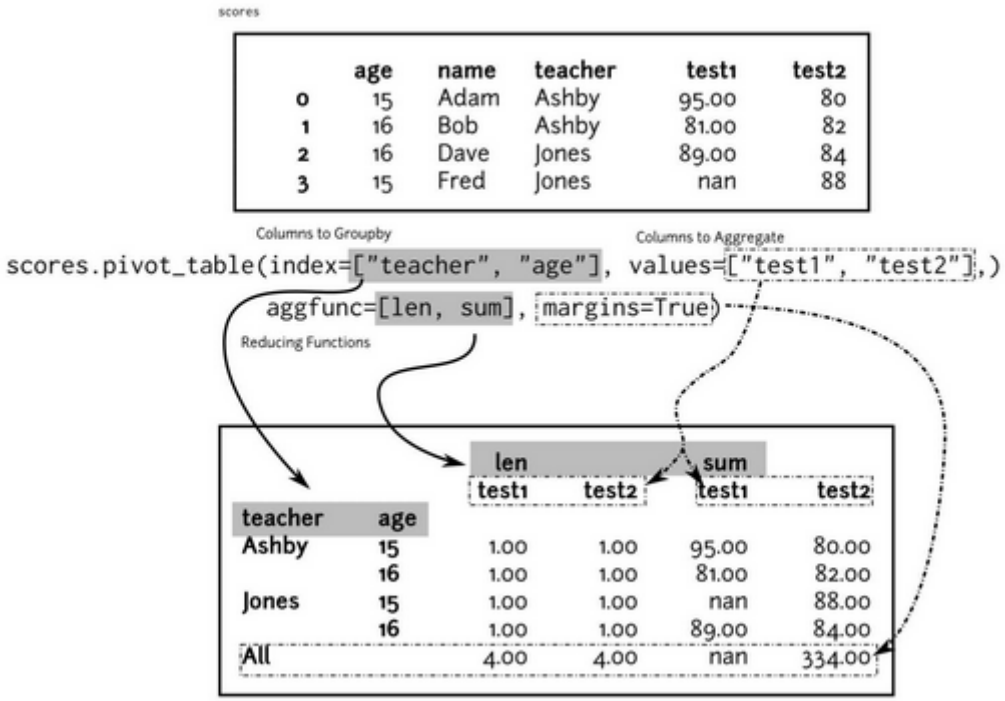
Pandas

Index multidimensionnels



Pandas

Tables pivot

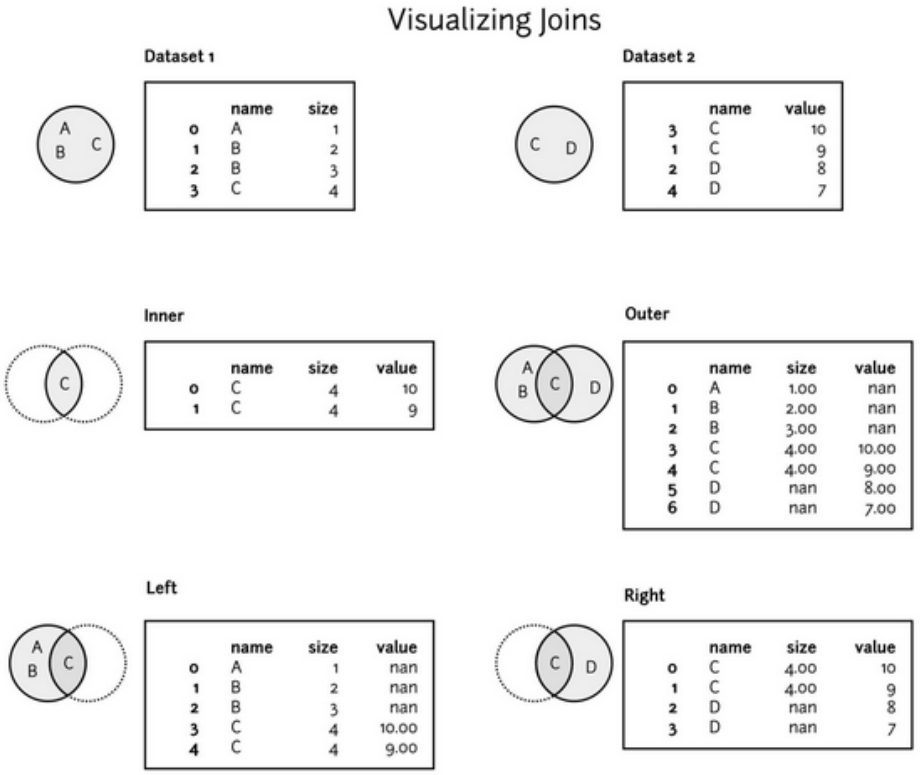


	age	name	teacher	test1	test2
0	15	Adam	Ashby	95.00	80
1	16	Bob	Ashby	81.00	82
2	16	Dave	Jones	89.00	84
3	15	Fred	Jones	nan	88

		len		sum	
		test1	test2	test1	test2
teacher	age				
Ashby	15	1.00	1.00	95.00	80.00
	16	1.00	1.00	81.00	82.00
Jones	15	1.00	1.00	nan	88.00
	16	1.00	1.00	89.00	84.00
All		4.00	4.00	nan	334.00

Pandas

Jointures



Dataset 1

	name	size
0	A	1
1	B	2
2	B	3
3	C	4

Dataset 2

	name	value
3	C	10
1	C	9
2	D	8
4	D	7

Inner

	name	size	value
0	C	4	10
1	C	4	9

Outer

	name	size	value
0	A	1.00	nan
1	B	2.00	nan
2	B	3.00	nan
3	C	4.00	10.00
4	C	4.00	9.00
5	D	nan	8.00
6	D	nan	7.00

Left

	name	size	value
0	A	1	nan
1	B	2	nan
2	B	3	nan
3	C	4	10.00
4	C	4	9.00

Right

	name	size	value
0	C	4.00	10
1	C	4.00	9
2	D	nan	8
3	D	nan	7

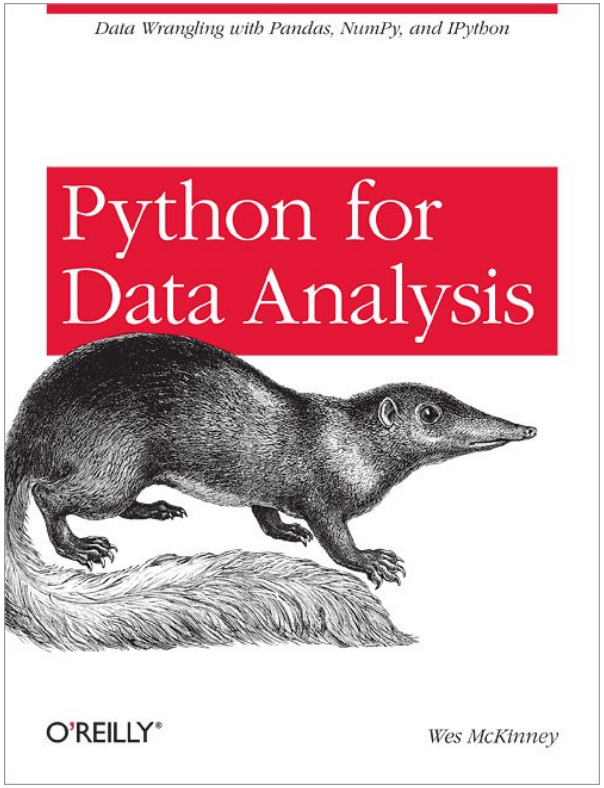
Pandas

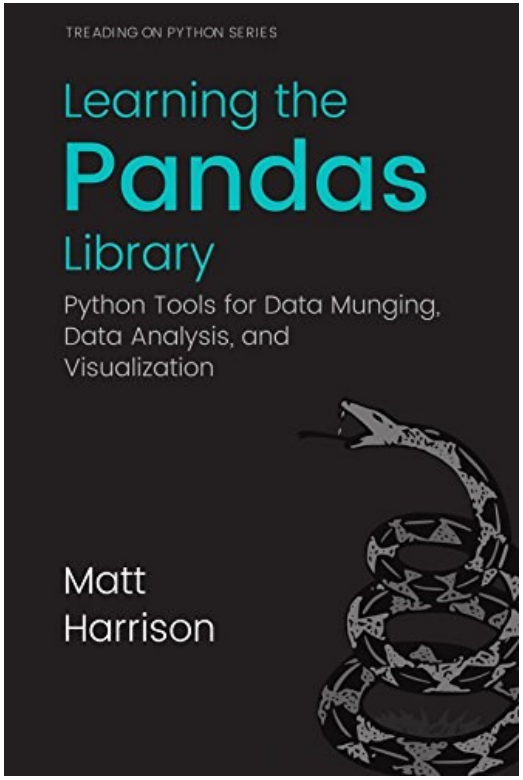
Exercices

1. Importer des données (disponibles sur le dossier du cours)
2. Analyser des données
3. Travailler avec différents types de données et des données manquantes
4. Exporter des données
5. Créer des graphiques simples

Aide Pandas : https://pandas.pydata.org/pandas-docs/stable/10min.html

Pour aller plus loin





Sources

Cheat Sheets distribués dans le cours :

- Jupyter notebook :

  https://www.datacamp.com/community/blog/jupyter-notebook-cheat-sheet

- Markdown :

  http://geog.uoregon.edu/bartlein/courses/geog607/Rmd/MDquick-refcard.pdf

- Pandas :

  https://github.com/pandas-dev/pandas/blob/master/doc/cheatsheet/Pandas_Cheat_Sheet.pdf