

Midis de l'info scientifique

Traitement de données avec Pandas & Jupyter notebooks

Pablo Iriarte – pablo.iriarte@unige.ch / 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

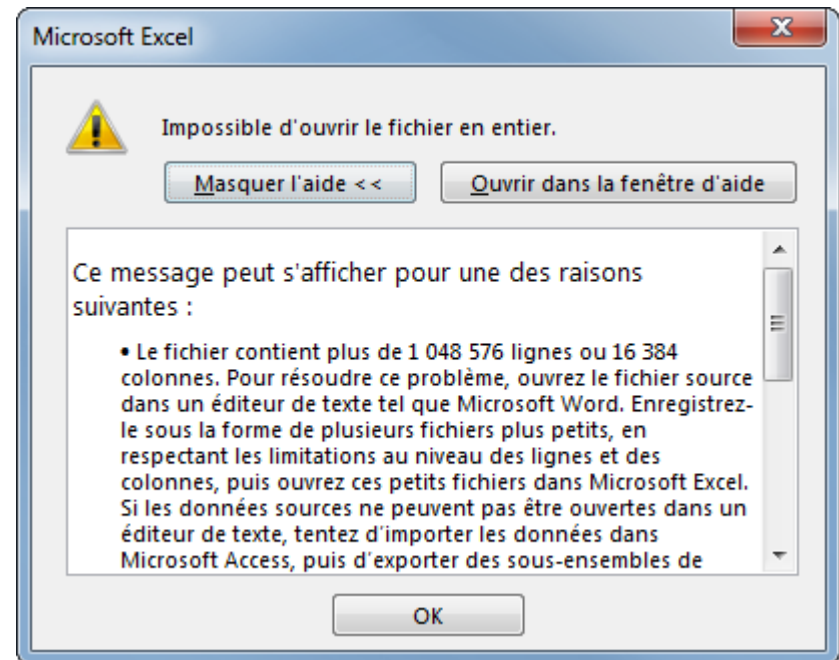
- Excel : limitations
- Excel : erreurs scientifiques
- Reproductibilité et Open Science
- Big Data et Open Data

Introduction

Excel : limitations

Liste complète :

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

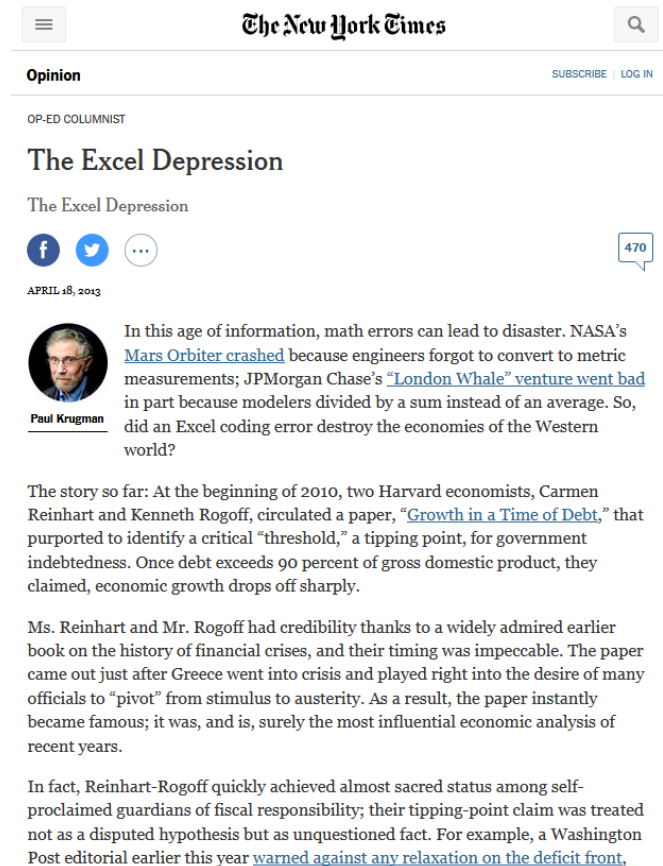


Introduction

Excel : erreurs scientifiques

L'exemple du «Reinhart-Rogoff error»

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



The screenshot shows a web browser displaying a New York Times article. At the top, the New York Times logo is visible on the right, and a search icon is on the far right. Below the logo, the word "Opinion" is centered, with "SUBSCRIBE" and "LOG IN" links to its right. Underneath, "OP-ED COLUMNIST" is written. The article title "The Excel Depression" is prominently displayed. Below the title, the author's name "Paul Krugman" is shown next to a small circular portrait of him. To the right of the author's name is a comment bubble icon with the number "470". The article text begins with "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 text continues with a paragraph about the Reinhart-Rogoff paper and its impact, and another paragraph about the paper's status as a tipping-point claim.

The New York Times

Opinion

OP-ED COLUMNIST

The Excel Depression

The Excel Depression

Paul Krugman

APRIL 18, 2013

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](#),

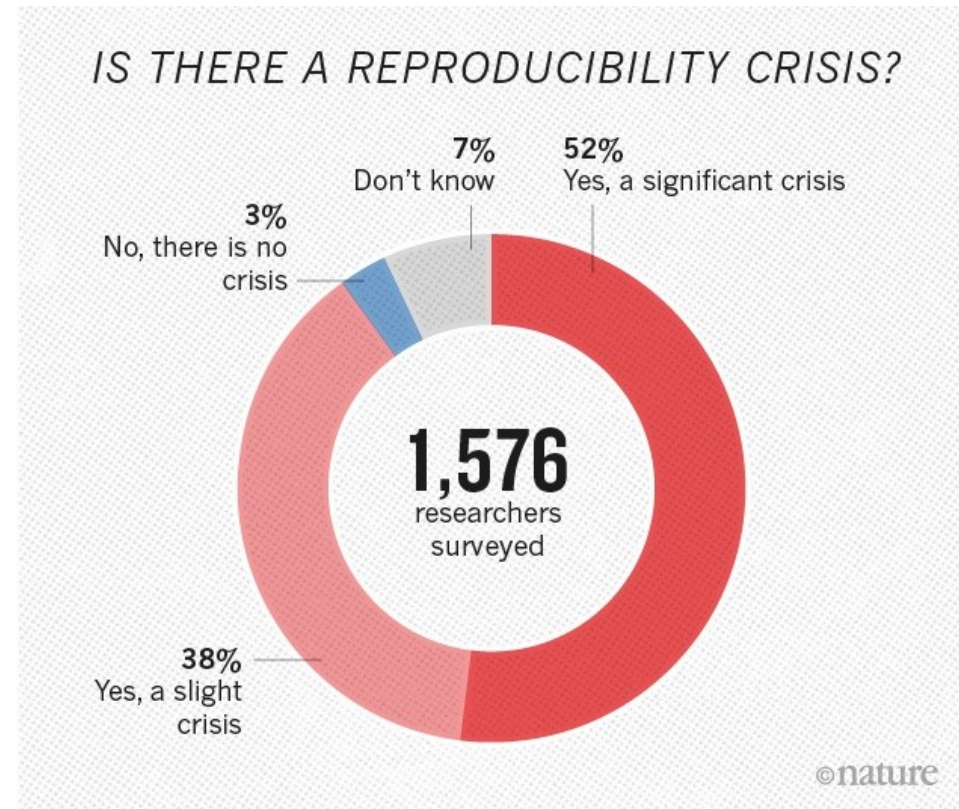
Introduction

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

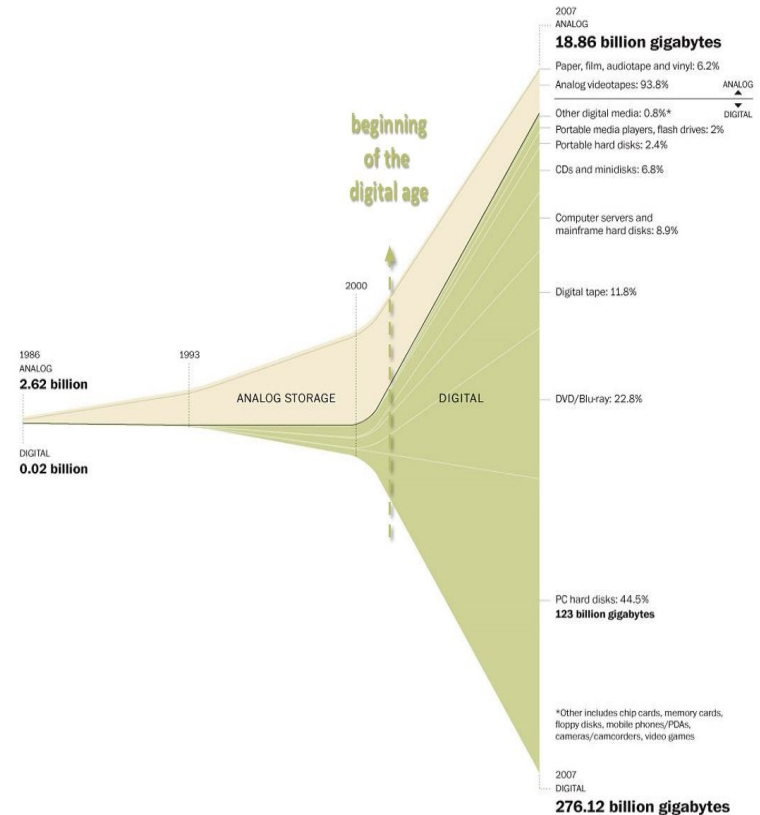
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2984851

Nombreux réservoirs ouverts

Kaggle : <https://www.kaggle.com>

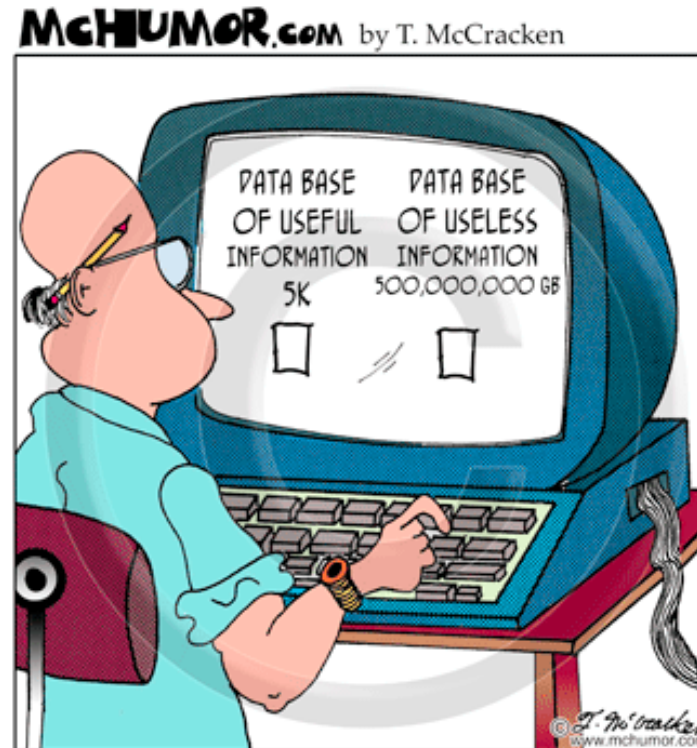
Data Hub : <http://datahub.io>

WikiData : <https://www.wikidata.org>

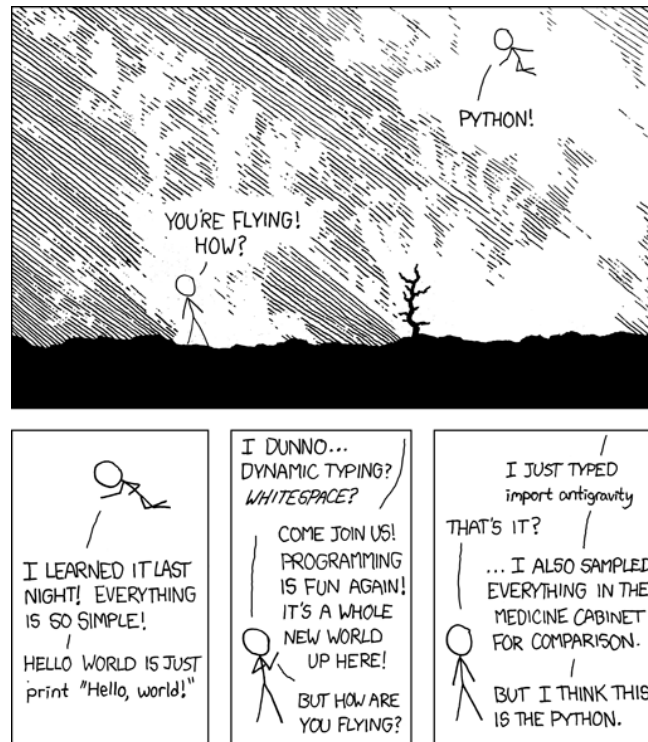


Washington Post, based on Hilbert and Lopez, 2011

Introduction



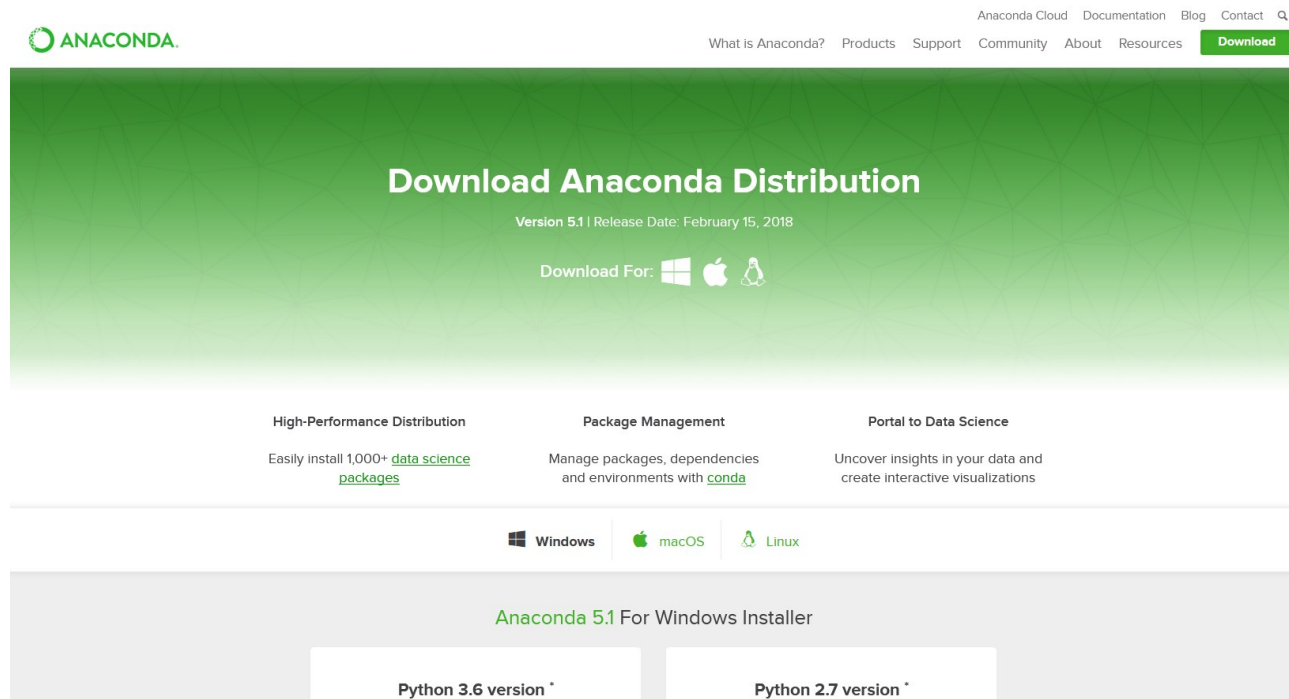
Introduction



<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

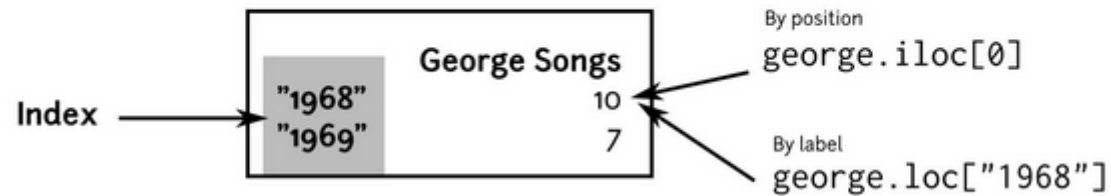
Series : 1 dimension

index		values
A	→	5
B	→	6
C	→	12
D	→	-5
E	→	6.7

Pandas

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

Indexing



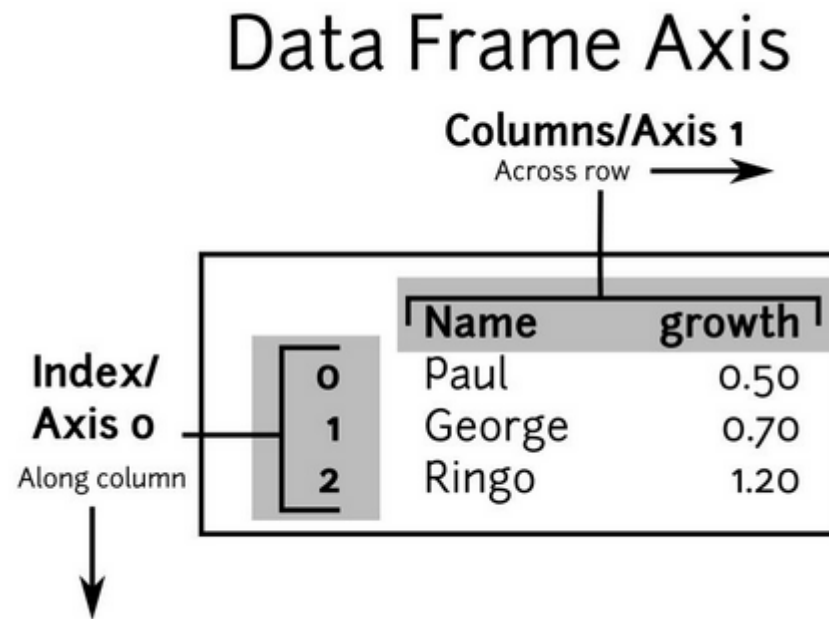
Pandas

DataFrame : 2 dimensions

columns		foo	bar	baz	qux
index					
A	→	0	x	2.7	True
B	→	4	y	6	True
C	→	8	z	10	False
D	→	-12	w	NA	False
E	→	16	a	18	False

Pandas

DataFrame : axes



Pandas

DataFrame : slices

Row & Column Slicing Examples

`df.iloc[2:4, 0:1]` ← With a : return data frames
Position - Half-open interval
`df.loc['d':, 'Units']` ← Without a : return series
Label - Closed interval

Rows Columns

Pandas

Opérations facilitées par les index : jointures automatiques

			A	0		A	NA
			B	1		B	2
			C	2		C	4
			D	3		D	6
						E	NA

B	1
C	2
D	3
E	4

		+

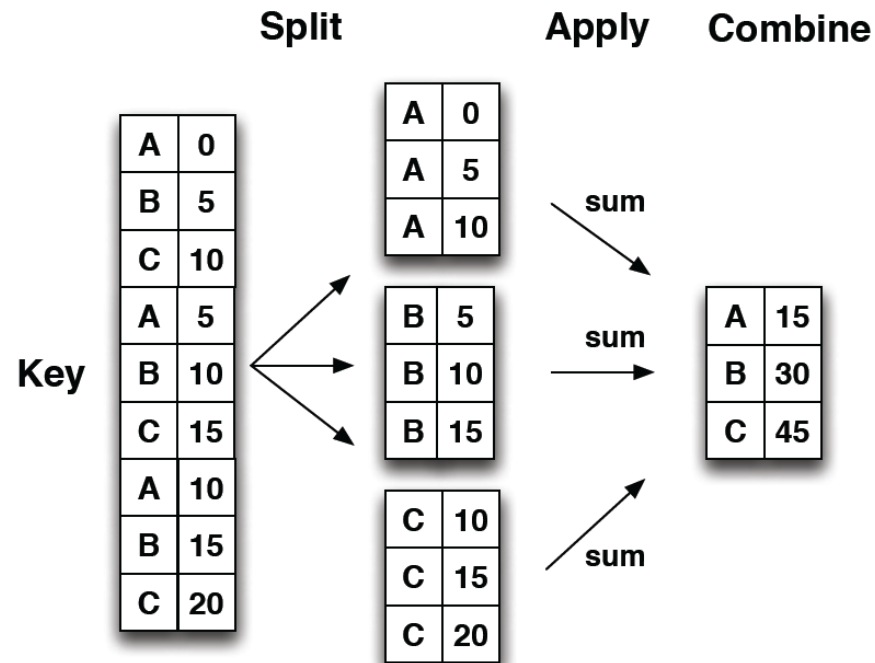
A	0
B	1
C	2
D	3

		=

A	NA
B	2
C	4
D	6
E	NA

Pandas

Opérations : GroupBy



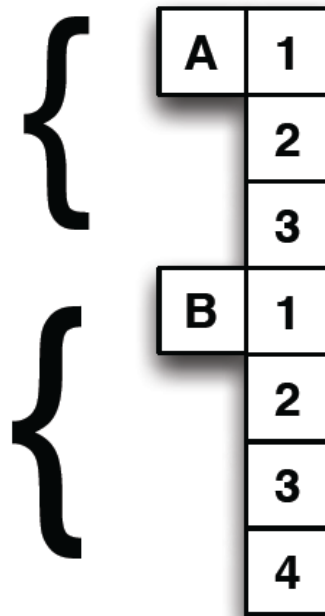
Pandas

Opérations : GroupBy

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

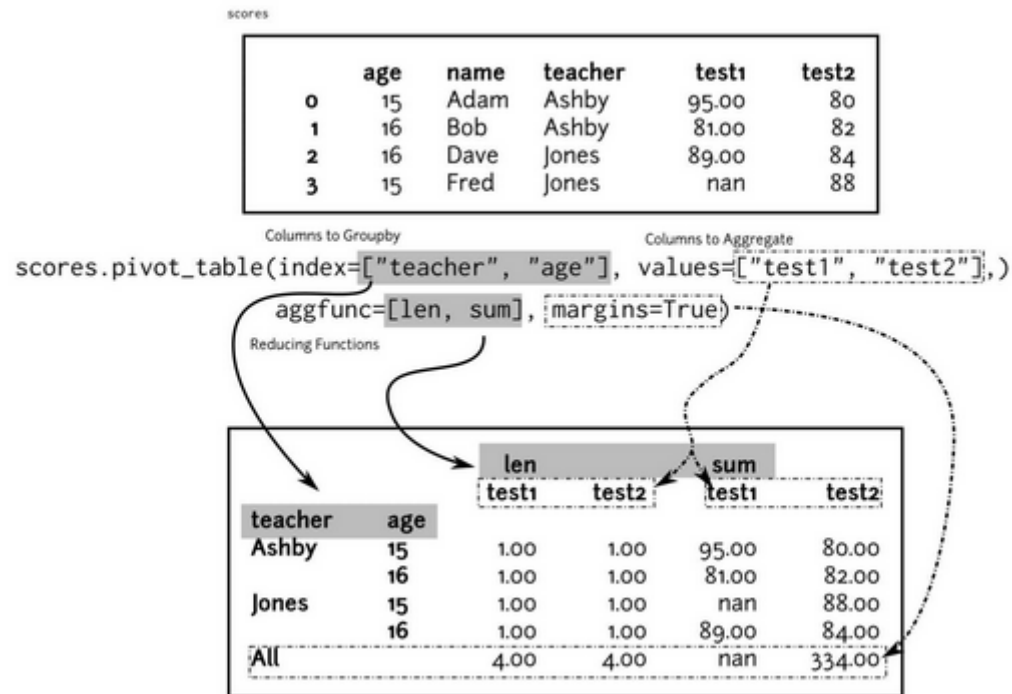
Pandas

Index multidimensionnels



Pandas

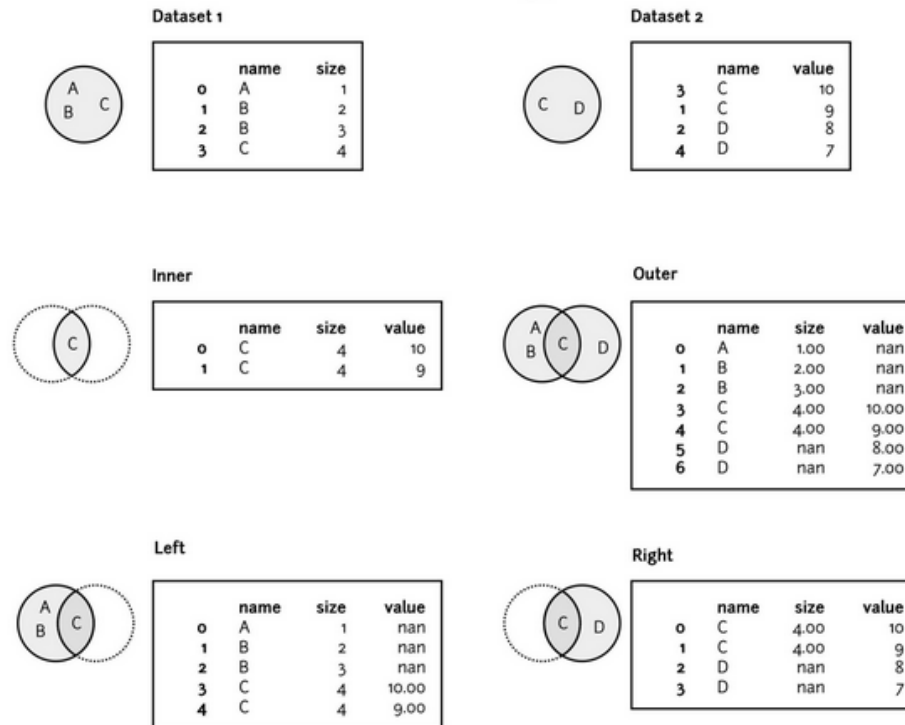
Tables pivot



Pandas

Jointures

Visualizing Joins



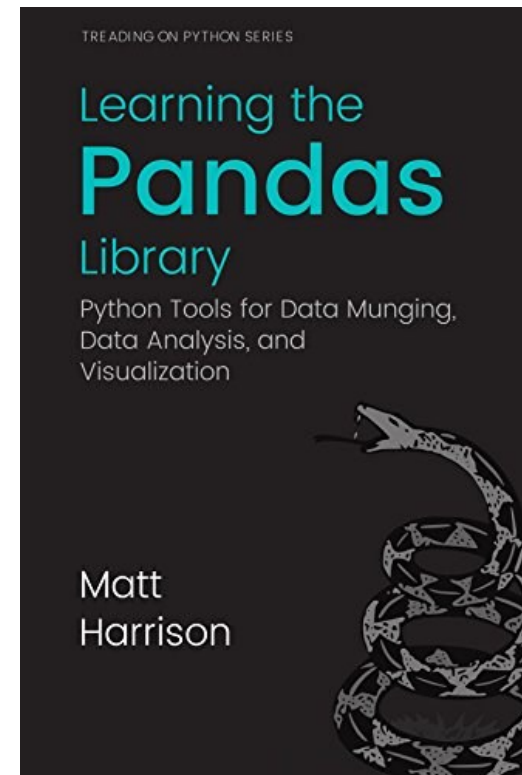
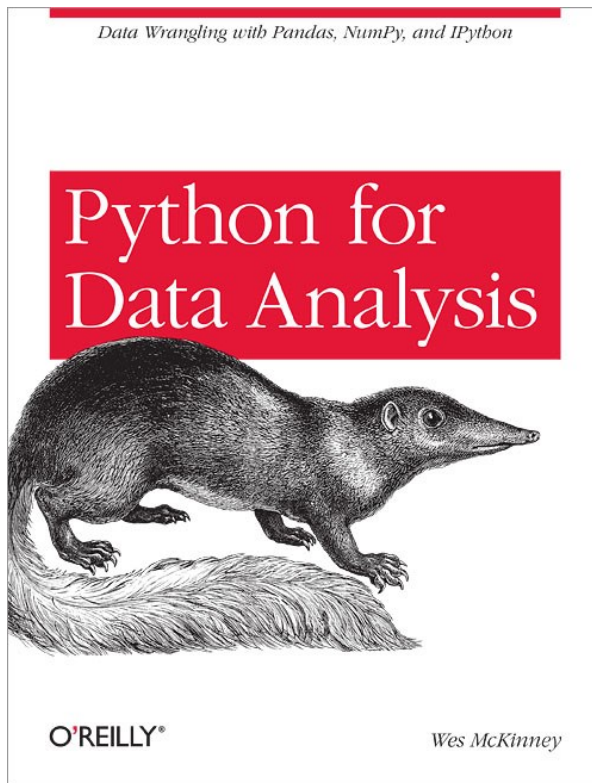
Pandas

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