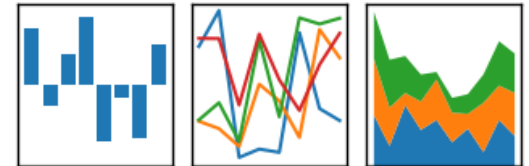


# Rendez-vous de l'info scientifique

## Traitement de données avec Pandas & Jupyter notebooks



pandas  
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Pablo Iriarte – [pablo.iriarte@unige.ch](mailto:pablo.iriarte@unige.ch) / DIS

25 septembre 2024

# Programme

## Introduction

- Historique
- Excel et les erreurs scientifiques
- Reproducibility Crisis & Data deluge

## Jupyter Notebooks

- Famille d'outils
- Accès au JupyterHub du cours ou installation via la distribution Anaconda
- Créer, organiser et partager des notebooks

## Pandas

- Importer et exporter des données
- Manipuler et analyser les données
- Générer des graphiques

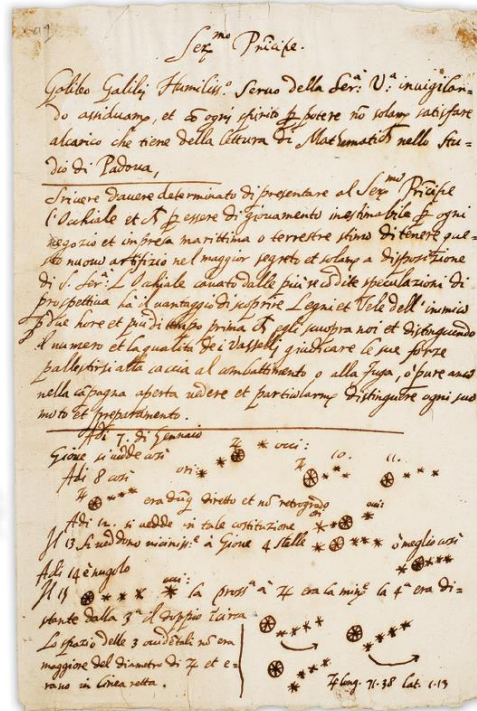
# Introduction

## Historique

- **iPython** (2001->) <https://ipython.org/>
- **Jupyter** (2014 ->) <https://jupyter.org/>
- **Famille d'outils**
  - **Jupyter Hub** : <https://jupyterhub.readthedocs.io/>
  - **Jupyter Lab** : <https://jupyterlab.readthedocs.io/>
  - **NB viewer** : <https://nbviewer.jupyter.org/>
  - **Binder** : <https://mybinder.org/>

# Introduction

## Historique : Jupiter

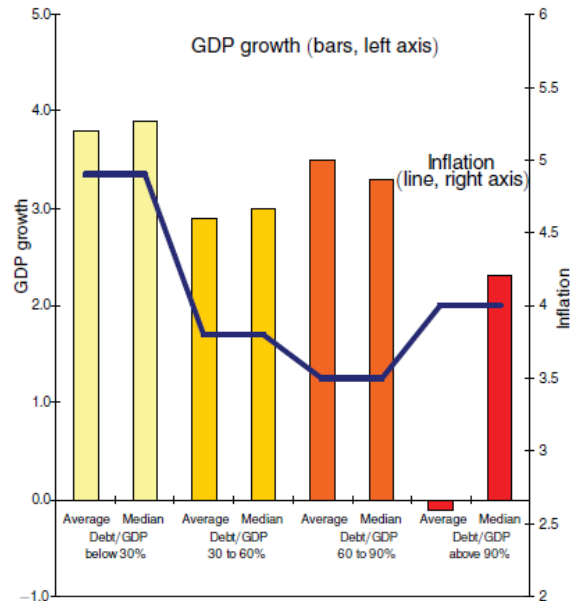


[https://commons.wikimedia.org/wiki/File:Galileo\\_manuscript.png](https://commons.wikimedia.org/wiki/File:Galileo_manuscript.png)

# Introduction

## Excel et les erreurs scientifiques

### L'exemple du «Reinhart-Rogoff error»



Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. [DOI:10.1257/aer.100.2.573](https://doi.org/10.1257/aer.100.2.573)



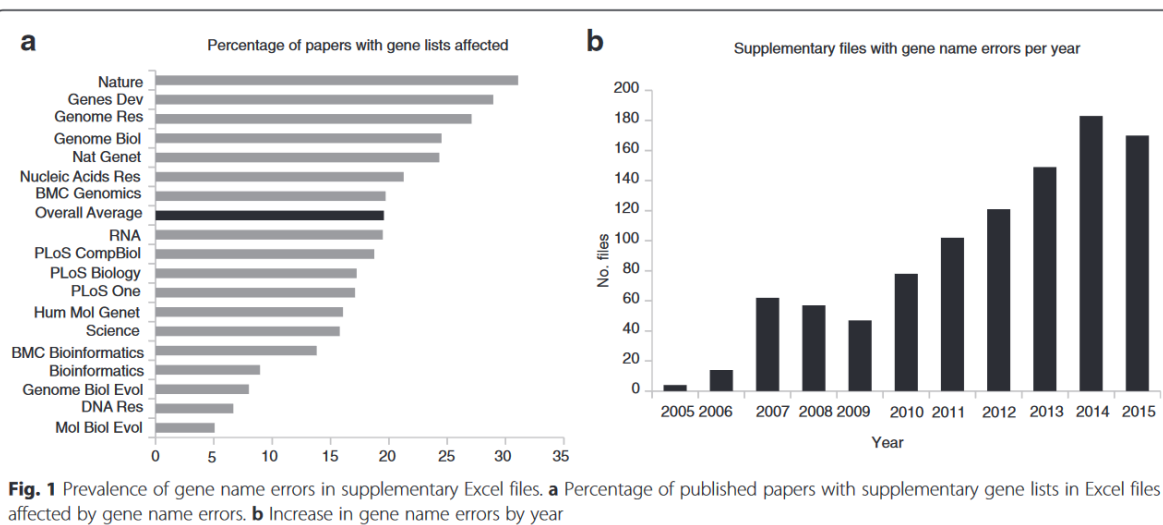
<https://www.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html>



# Introduction

## Excel et les erreurs scientifiques

### L'exemple de la conversion des données



Ziemann et al. 2016. [DOI:10.1186/s13059-016-1044-7](https://doi.org/10.1186/s13059-016-1044-7)

Ziemann et al. *Genome Biology* (2016) 17:177  
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann<sup>1</sup>, Yotam Eren<sup>1,2</sup> and Assam El-Osta<sup>1,3\*</sup>

**Abstract**  
The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

**Keywords:** Microsoft Excel, Gene symbol, Supplementary data

**Abbreviations:** GEO, Gene Expression Omnibus; JIF, journal impact factor

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* (Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase) are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '231009E13' to '2.31E+13'). Since that report, we have uncovered further instances where gene symbols were converted to dates in supplementary data of recently published papers (e.g. *SEPT2* converted to '2006/09/02'). This suggests that gene name errors continue to be a problem in supplementary files accompanying articles. Inadvertent gene symbol conversion is problematic because these supplementary files are an important resource in the genomics community that are frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and .xlsx suffixes) were converted to tabular separated files (.tsv) with *ssconvert* (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Oryza sativa* and *Saccharomyces cerevisiae* [2]. The regex search used was similar to that described previously by Zeeberg and colleagues [1], with the added screen for dates in other formats (e.g. DD/MM/YY and MM-DD-YY). To expedite analysis of supplementary files from multi-disciplinary journals, we limited the articles screened to those that have the keyword 'genome' in the title or abstract (*Science*, *Nature* and *PLoS One*). Excel files (.xls and .xlsx) deposited in NCBI Gene Expression Omnibus (GEO) [3] were also screened in the same way (files released 2005–2015). All URLs screened, results and scripts used in this study are currently available at SourceForge (<https://sourceforge.net/projects/genomewerrenscreen/>). Scripts were run on Ubuntu v14.04 LTS with GNU bash version 4.3.11. These findings were verified manually by downloading and checking Excel files from every paper and GEO file suspected to include gene name errors.

Supplementary files in Excel format from 18 journals published from 2005 to 2015 were programmatically screened for the presence of gene name errors. In total, we screened 35,175 supplementary Excel files, finding 7467 gene lists attached to 3597 published papers. We

\* Correspondence: [Assam.El-Osta@monash.edu](mailto:Assam.El-Osta@monash.edu)  
<sup>1</sup>Yotam Eren & Diabetes Institute, The Alfred Medical Research and Education Precinct, Melbourne, Victoria 3004, Australia  
<sup>2</sup>Central Clinical School, Faculty of Medicine, Monash University, Clayton, Victoria 3168, Australia  
Full list of author information is available at the end of the article

BioMed Central

© 2016 The Author(s). Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

# Introduction

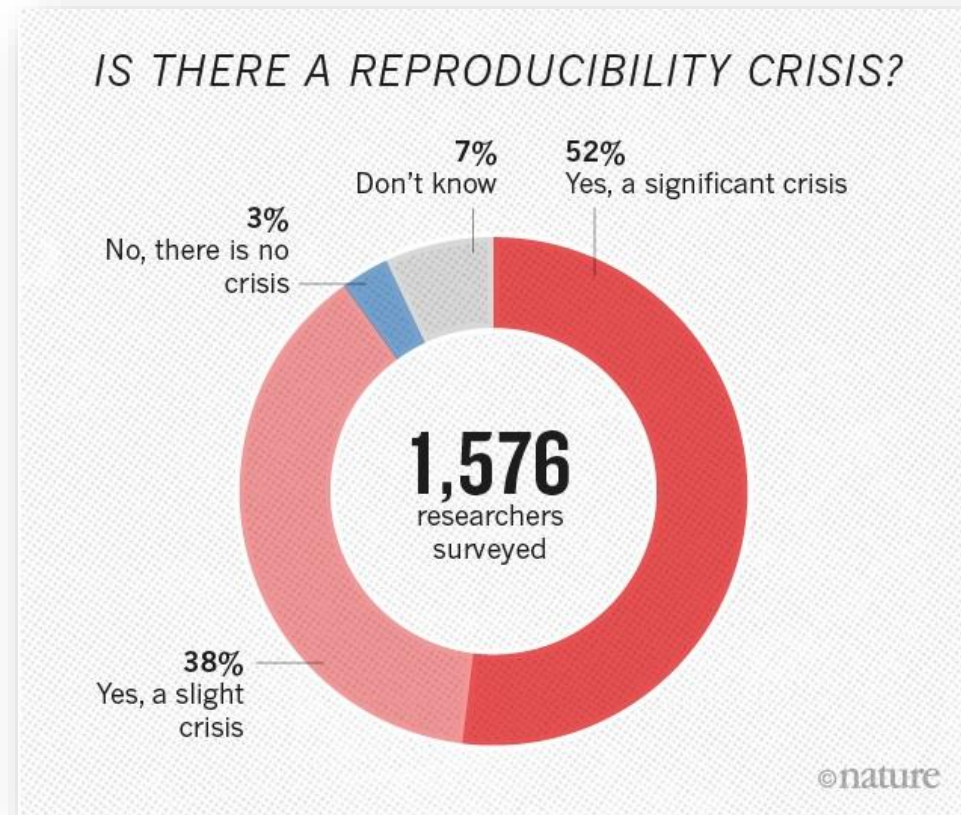
## Reproductibilité et Open Science

La science en crise?

1,500 scientists lift the lid on reproducibility

Baker 2016, Nature 533

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



# Introduction

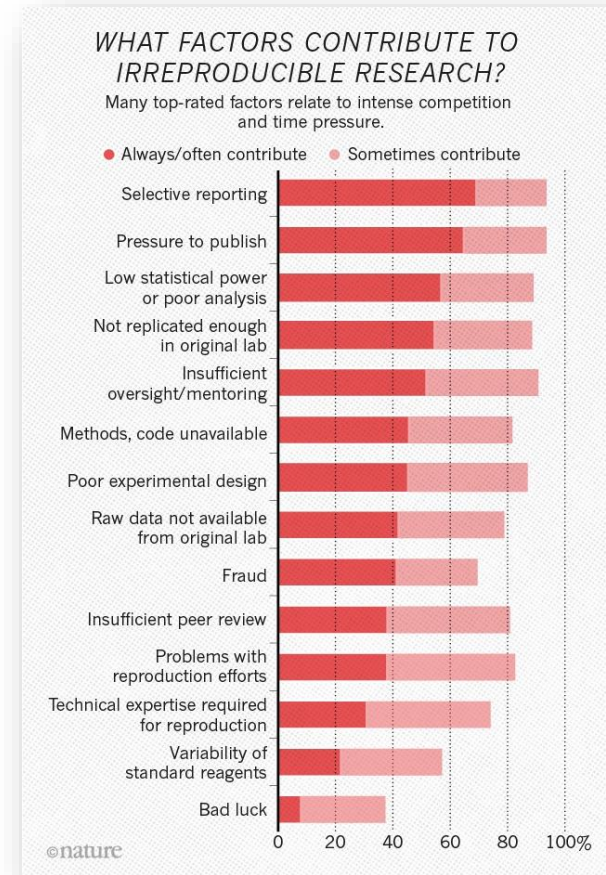
## Reproductibilité et Open Science

La science en crise?

1,500 scientists lift the lid on reproducibility

Baker 2016, Nature 533

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





# Introduction

## Reproductibilité et Open Science

Wired

<https://www.wired.com/2017/04/want-fix-sciences-replication-crisis-replicate/>



# Introduction

## Big Data et Open Data

Quantifying the Data Deluge and the Data Drought

[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2984851](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2984851)

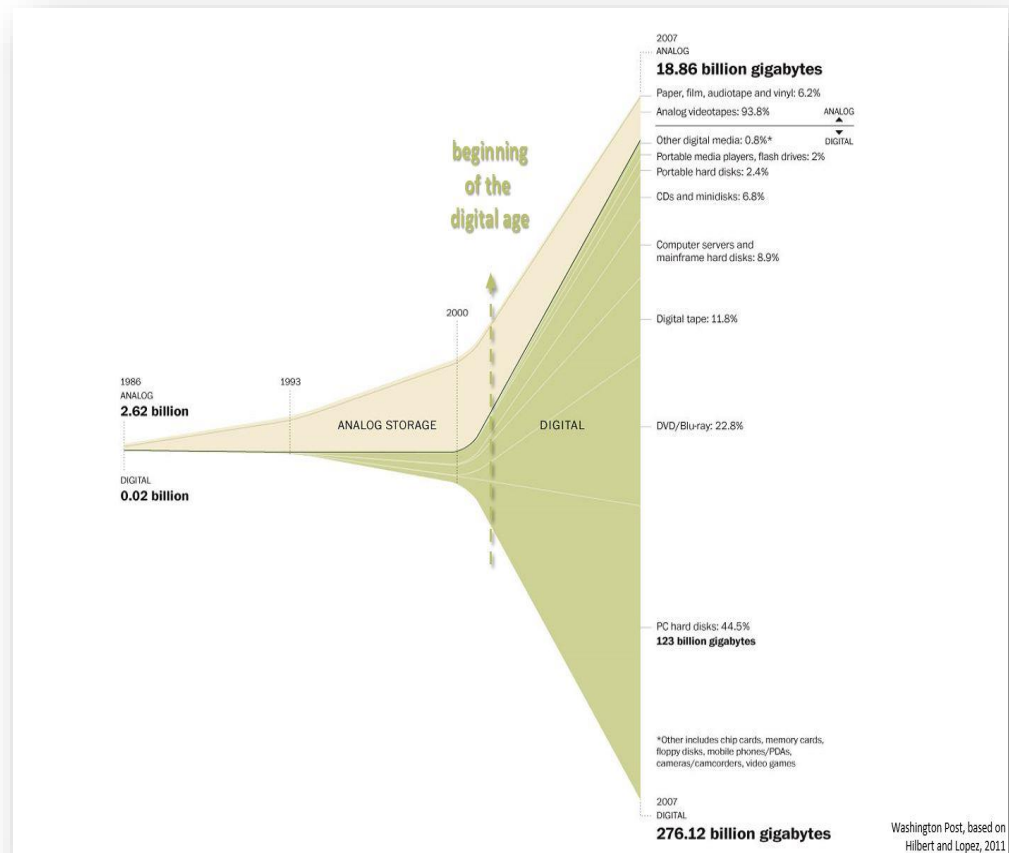
Nombreux réservoirs ouverts

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

Github: <http://github.com>

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

Statistique historique : <https://hssso.ch>

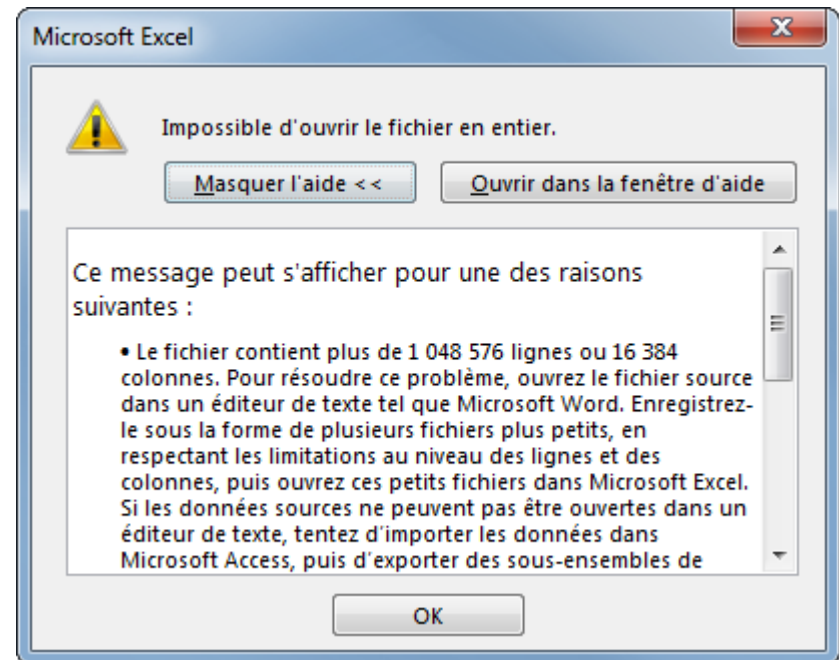


# Introduction

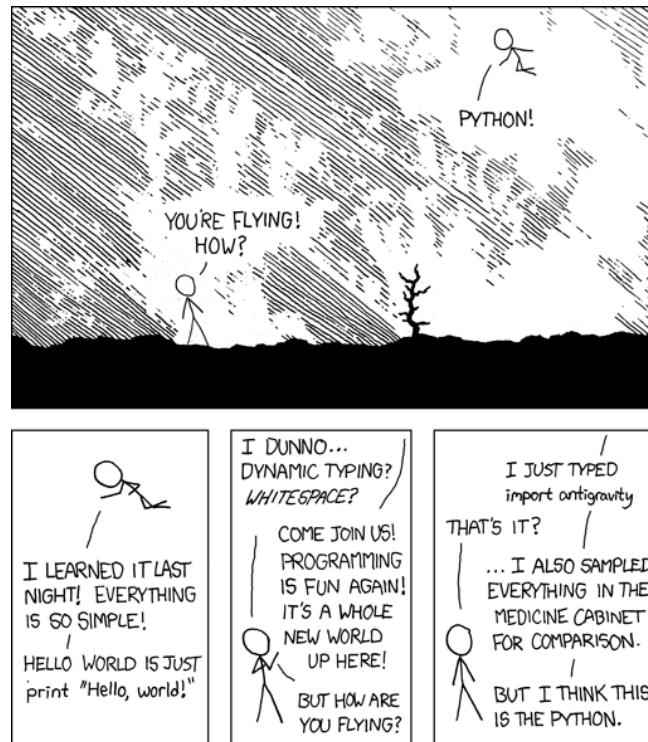
## Excel : limitations

### Liste complète :

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



# Introduction



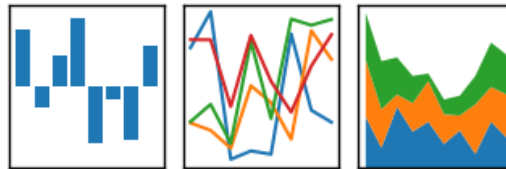
<https://www.xkcd.com/353/>



# Introduction

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



<https://pandas.pydata.org>



<https://www.pinterest.ch/pin/155303887164507907/>

# Introduction

## Reproductibilité et Open Science

Nature

<https://www.nature.com/articles/d41586-018-07196-1>



# Introduction

## Reproductibilité et Open Science

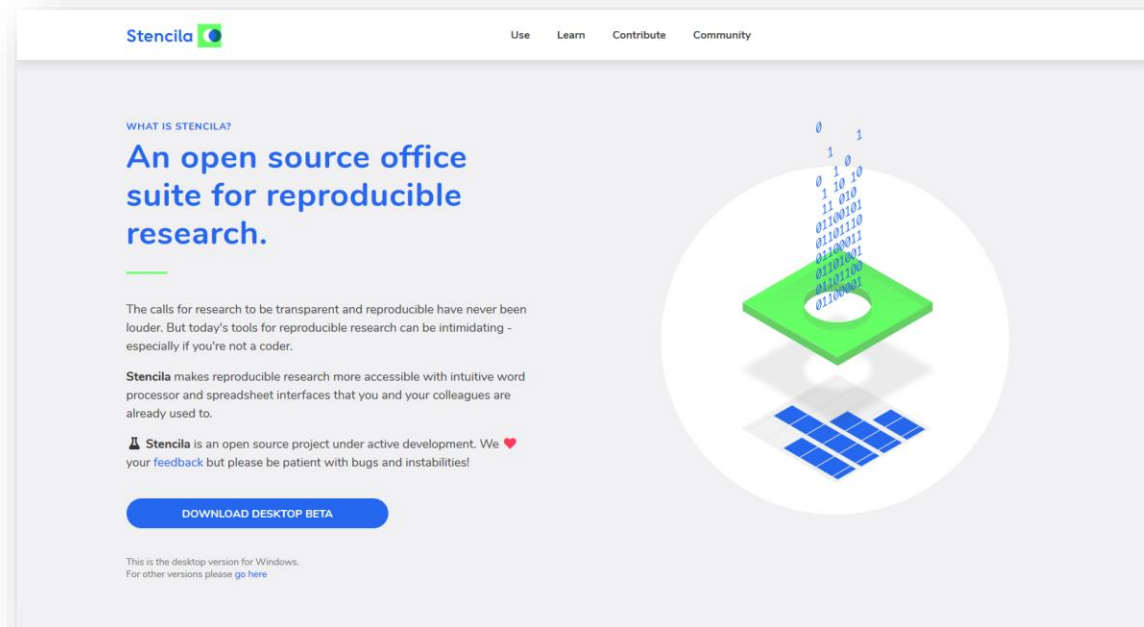
eLife

<https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article>



# Introduction

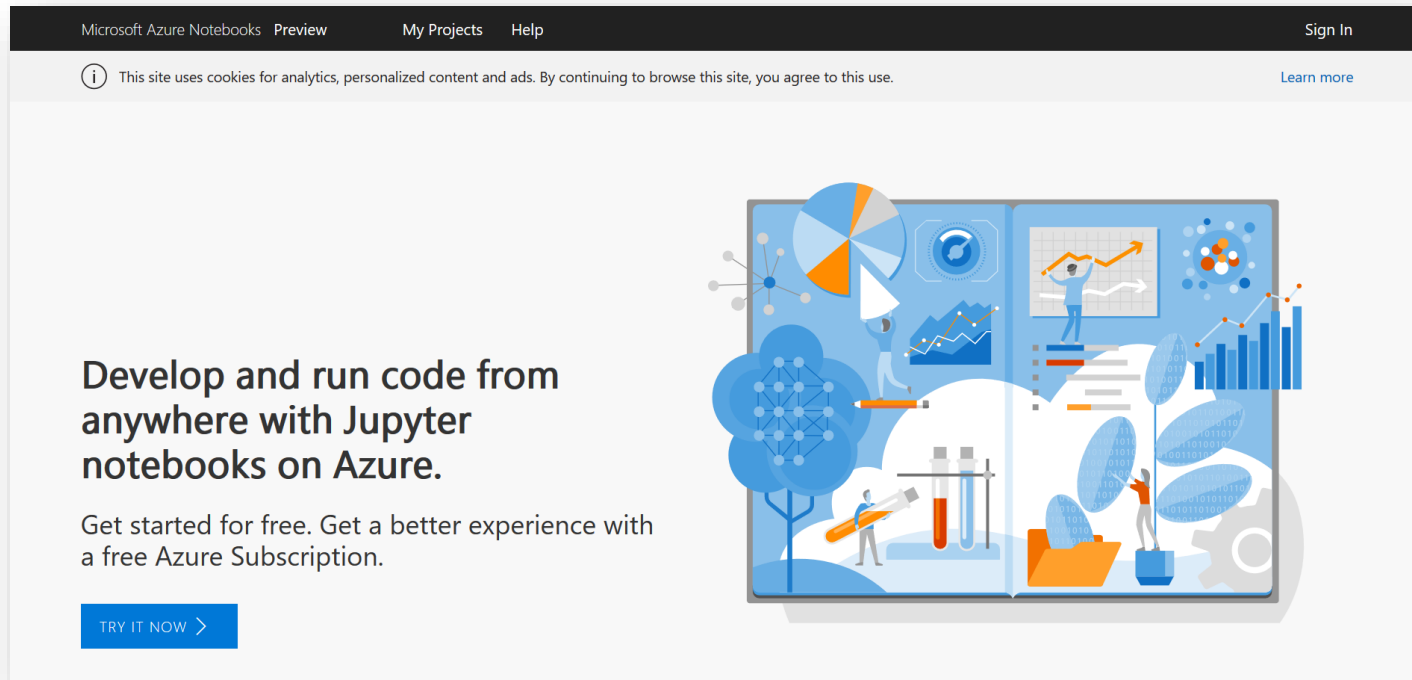
## Reproductibilité et Open Science



<https://stenci.la>



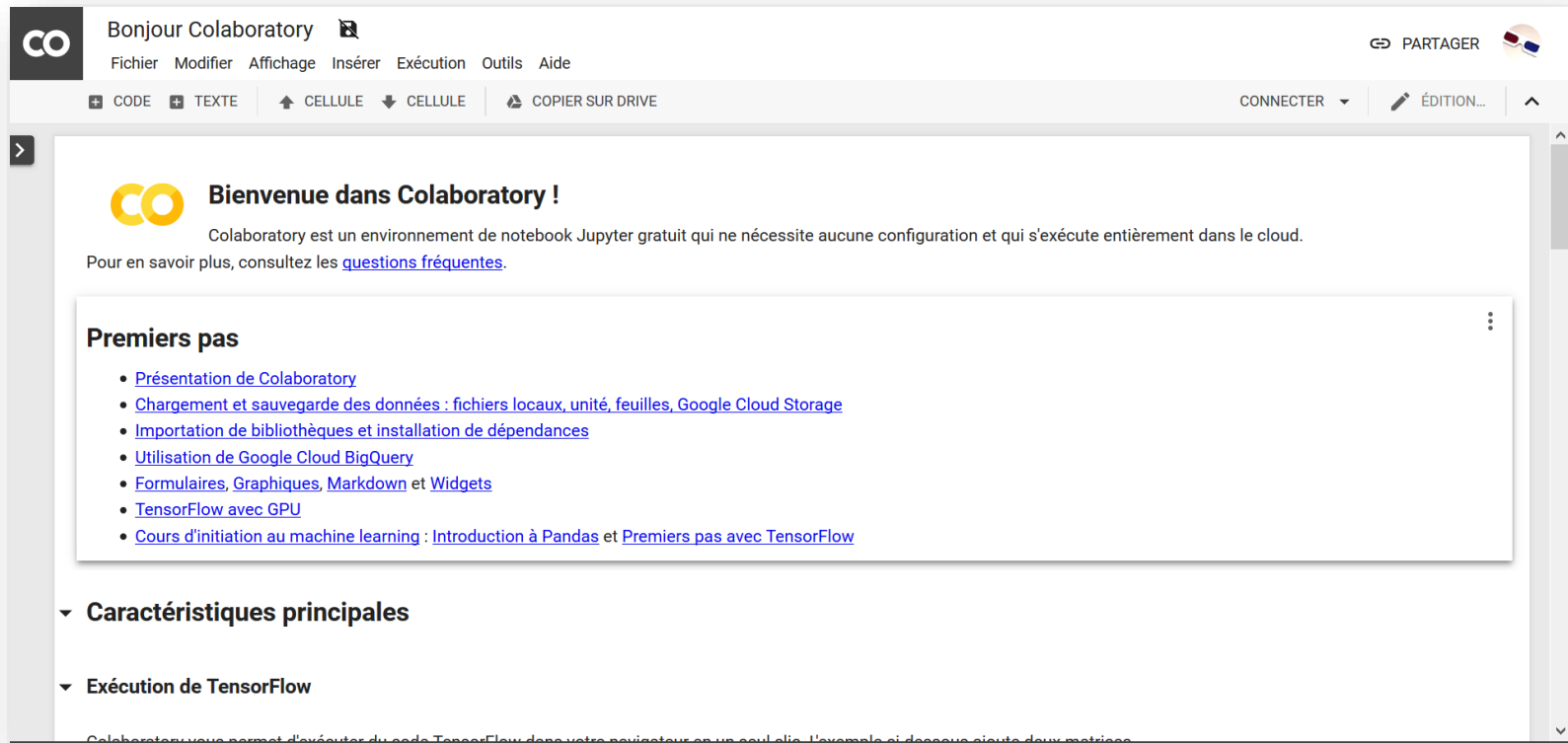
# Introduction



The screenshot shows the Microsoft Azure Notebooks website. At the top, there is a navigation bar with links for "Microsoft Azure Notebooks", "Preview", "My Projects", "Help", and "Sign In". Below the navigation bar, a cookie consent message states: "This site uses cookies for analytics, personalized content and ads. By continuing to browse this site, you agree to this use." with a "Learn more" link. The main content area features a large illustration of an open book filled with various data science and technology icons, including a pie chart, a line graph, a bar chart, a network diagram, a person holding a magnifying glass, and a person standing next to a large gear. To the left of the illustration, the text reads: "Develop and run code from anywhere with Jupyter notebooks on Azure." followed by "Get started for free. Get a better experience with a free Azure Subscription." and a blue button labeled "TRY IT NOW >".

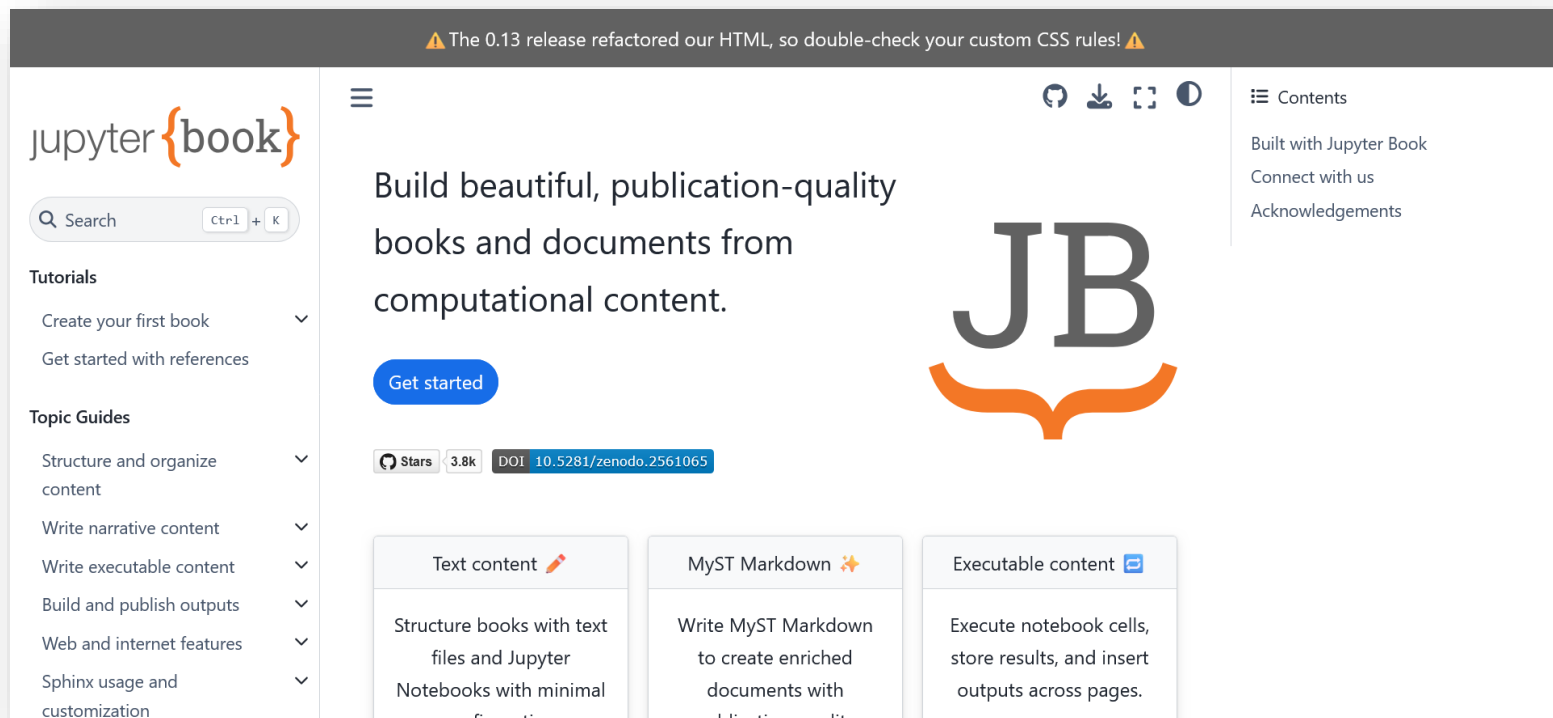
<https://notebooks.azure.com/>

# Introduction



<https://colab.research.google.com>

# Introduction

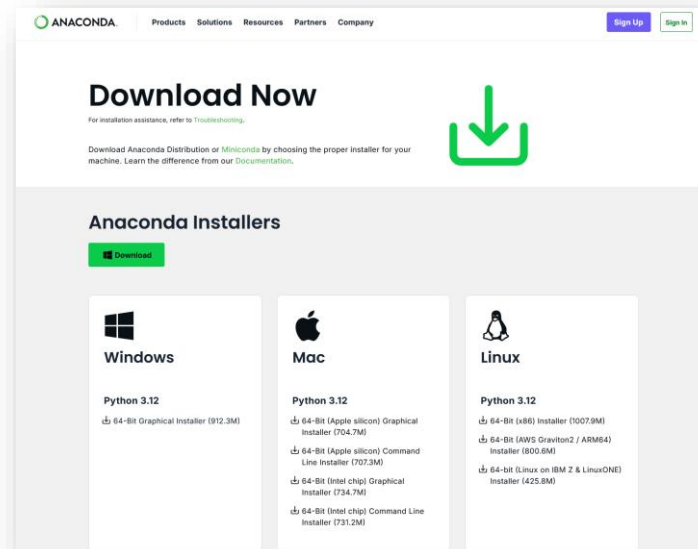


<https://jupyterbook.org>

# Jupyter Notebooks

**Installer Jupyter Notebooks et Pandas sur son poste personnel avec la distribution « Anaconda » :**

<https://www.anaconda.com/download/>



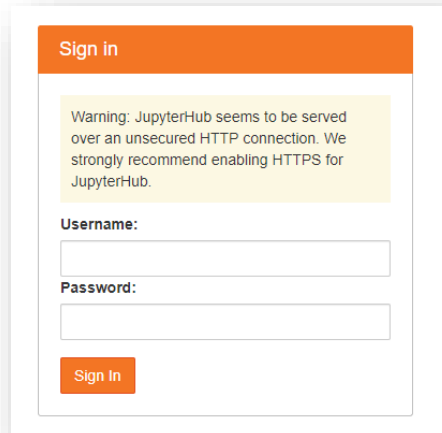


# Jupyter Notebooks

## Travail sur le JupyterHub du cours

Se connecter sur cette adresse avec le login/pwd fourni pendant le cours :

<https://learning.jupyterhub.unige.ch>

A screenshot of a JupyterHub sign-in form. The form has an orange header with the text "Sign in". Below the header is a yellow warning box with the text: "Warning: JupyterHub seems to be served over an unsecured HTTP connection. We strongly recommend enabling HTTPS for JupyterHub." Below the warning box are two input fields: "Username:" and "Password:". At the bottom of the form is an orange button with the text "Sign in".

Sign in

Warning: JupyterHub seems to be served over an unsecured HTTP connection. We strongly recommend enabling HTTPS for JupyterHub.

Username:

Password:

Sign in

# Jupyter Notebooks

Packages compris dans l'installation :

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

Liste complète : <https://docs.anaconda.com/anaconda/pkg-docs/>

# Jupyter Notebooks

**Créer, organiser et partager des notebooks**

Lancer Anaconda -> Jupyter Notebook



# Jupyter Notebooks

**Si besoin** : créer un lien symbolique entre le « home » et le dossier avec les notebooks

1. Avec le shell se positionner sur le «home»
2. Créer le lien avec la commande :  
`mklink /D Nom-du-lien Dossier-de-destination`

Aide : <https://www.howtogeek.com/howto/16226/complete-guide-to-symbolic-links-symlinks-on-windows-or-linux/>



# Jupyter Notebooks

## Se familiariser avec les notebooks

### Exercices

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

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

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

# Pandas

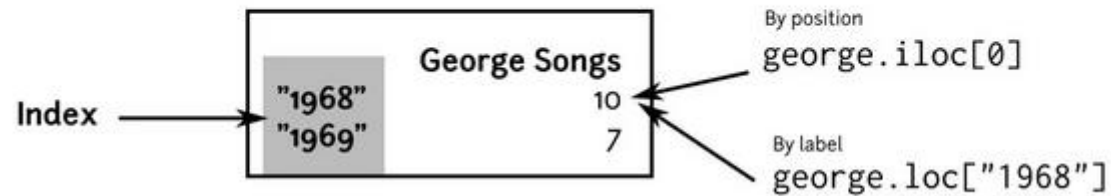
Series : 1 dimension

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



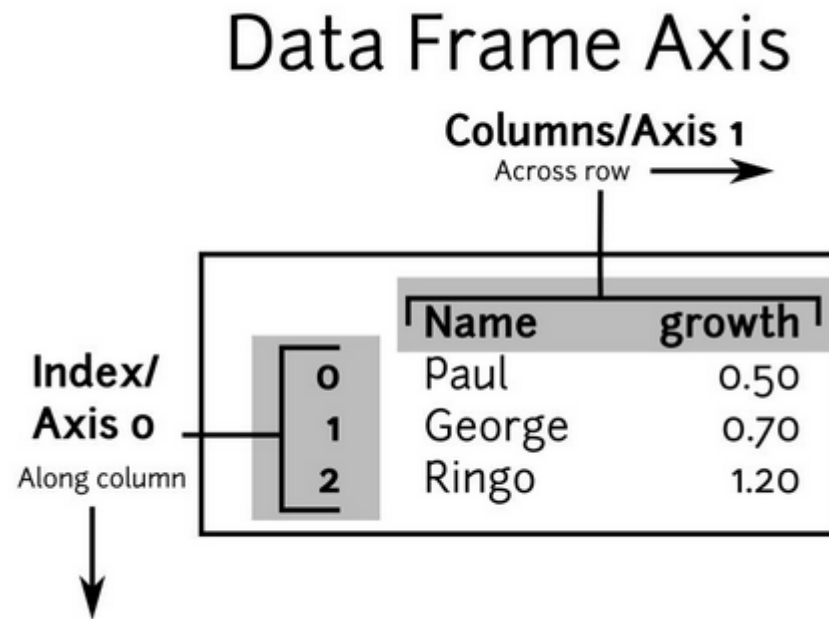
# Pandas

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

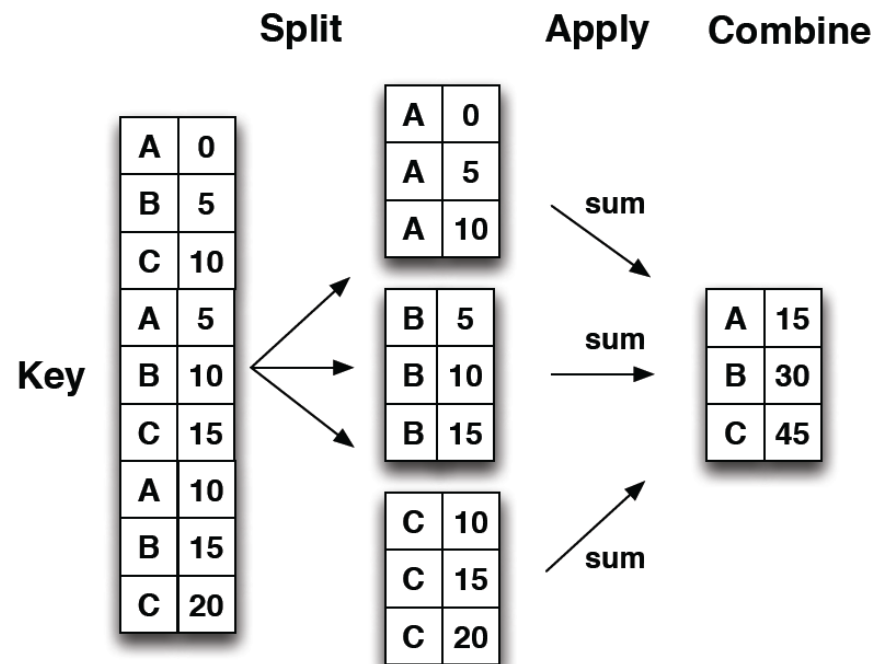
B	1
C	2
D	3
E	4

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

B	1
C	2
D	3
E	4

# Pandas

## Opérations : GroupBy



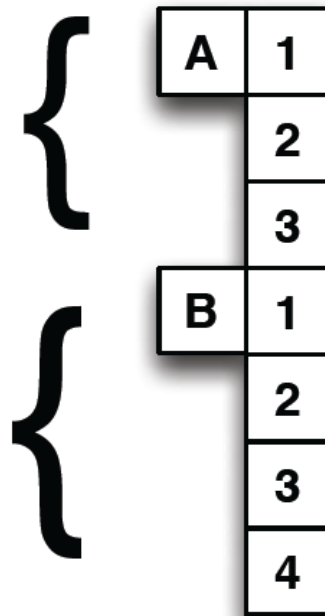
# Pandas

## 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

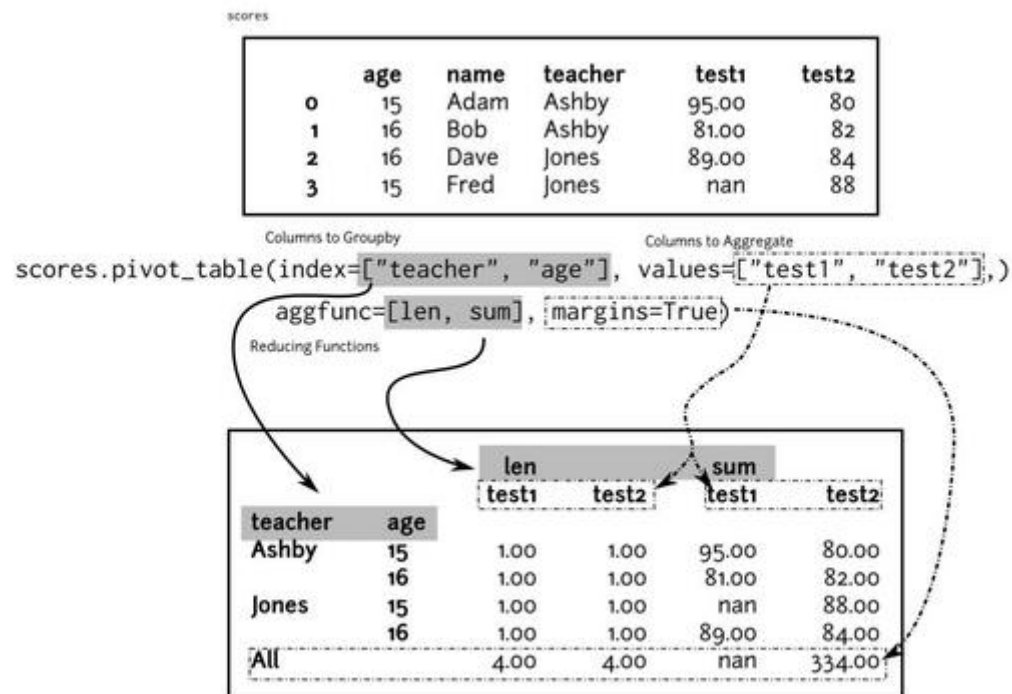
## Index multidimensionnels





# Pandas

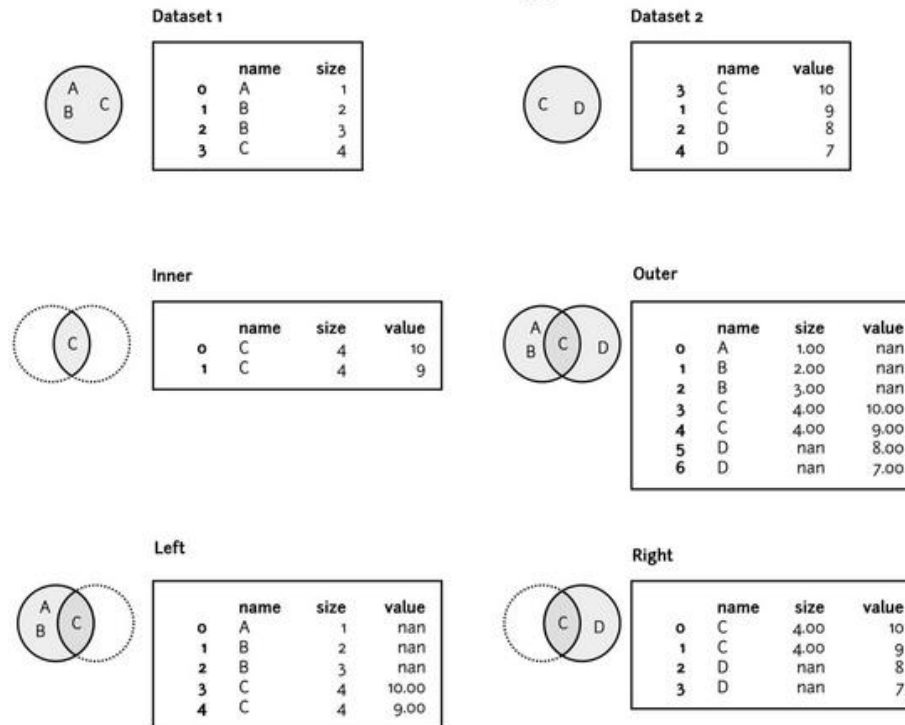
## Pivoter les tables



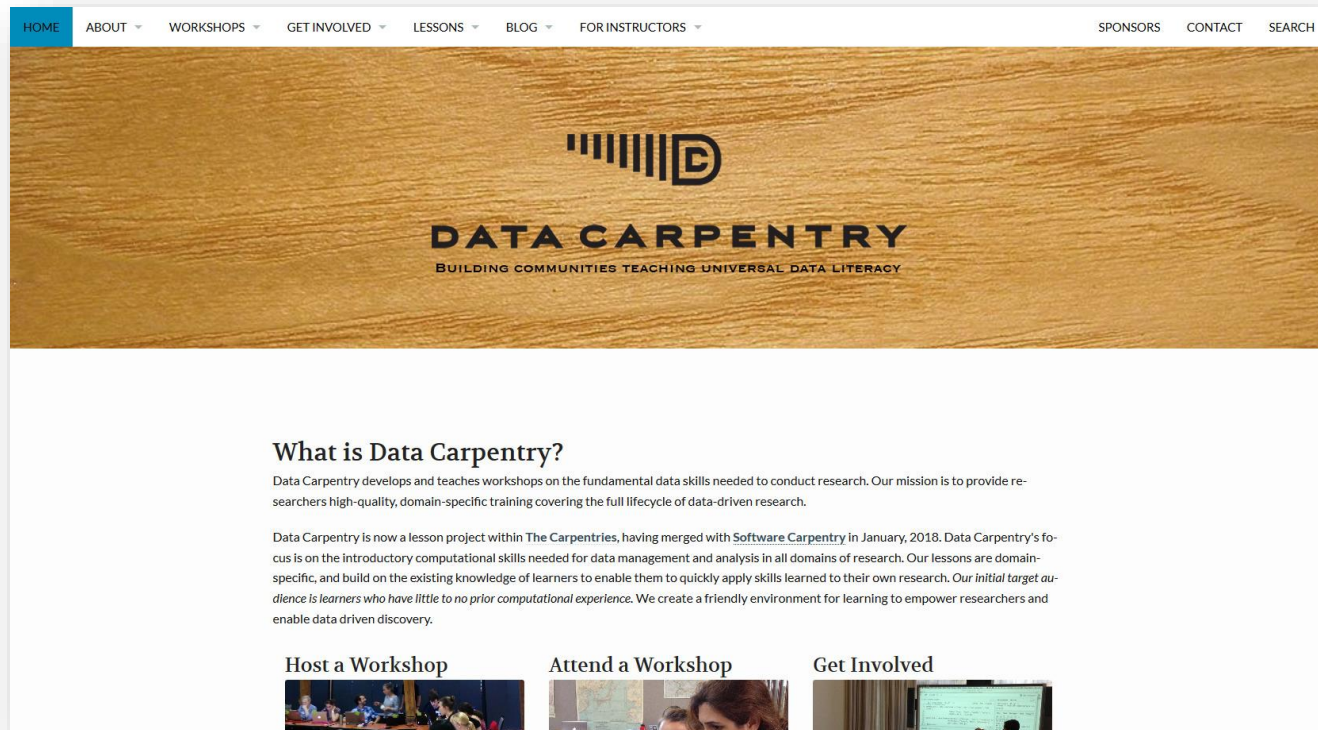
# Pandas

## Jointures (merge)

### Visualizing Joins



# Pour avancer « pas à pas »



<https://datacarpentry.org/python-socialsci/>

# Pour aller plus loin

## A gallery of interesting Jupyter Notebooks

Hans Fangohr edited this page 12 days ago · 74 revisions

[Edit](#)
[New Page](#)

This page is a curated collection of Jupyter/IPython notebooks that are notable. Feel free to add new content here, but please try to only include links to notebooks that include interesting visual or technical content; this should *not* simply be a dump of a Google search on every ipynb file out there.

**Important contribution instructions:** If you add new content, please ensure that for any notebook you link to, the link is to the rendered version using [nbviewer](#), rather than the raw file. Simply paste the notebook URL in the nbviewer box and copy the resulting URL of the rendered version. This will make it much easier for visitors to be able to immediately access the new content.

Note that [Matt Davis](#) has conveniently written a set of [bookmarklets and extensions](#) to make it a one-click affair to load a Notebook URL into your browser of choice, directly opening into nbviewer.

► Pages 10

<https://orzota.com/wp-content/uploads/2014/04/Slide2.jpg>

Clone this wiki locally

<https://github.com/jupyter/>

## Table of Contents

1. Entire books or other large collections of notebooks on a topic
  - [Introductory Tutorials](#)
  - [Programming and Computer Science](#)
  - [Statistics, Machine Learning and Data Science](#)
  - [Mathematics, Physics, Chemistry, Biology](#)

<https://github.com/jupyter/jupyter/wiki#a-gallery-of-interesting-jupyter-notebooks>



# Pour aller plus loin

**EPFL** About Education Research Innovation Schools Campus Q

EN

Education / Educational Initiatives / Center for Digital Education / Projects, Training and Support / Jupyter Notebooks For Education / Teaching and learning with Jupyter Notebooks / Projects and labs

Virtual demonstrations

Online interactive textbooks

Exercise worksheets

**Projects and labs**

Graded assignments

## Projects and labs with Jupyter notebooks

25%	893029.000000	1.000000	1.000000	4.000000	3.000000	4.000000	3.000000	2.000000	4.00
50%	901227.000000	1.000000	1.000000	4.000000	3.000000	4.000000	3.000000	2.000000	4.00
75%	902506.500000	1.000000	3.000000	4.000000	3.000000	4.000000	3.000000	2.000000	4.00
max	906177.000000	4.000000	5.000000	5.000000	3.000000	5.000000	5.000000	5.000000	5.00

i rows x 164 columns

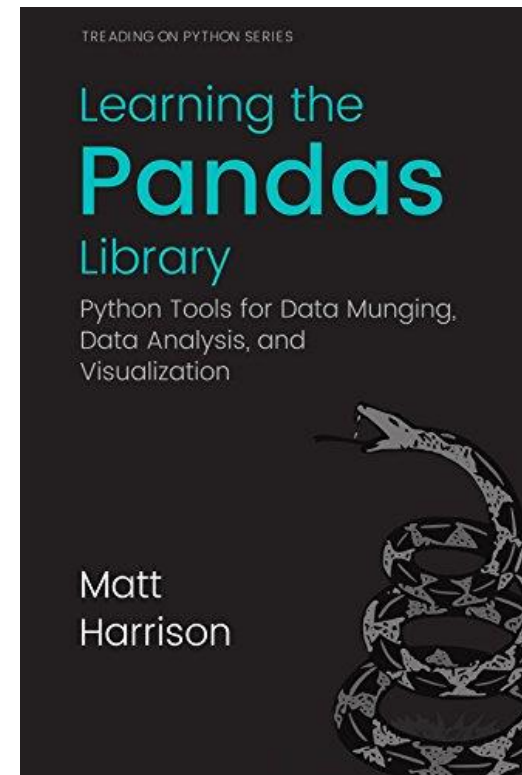
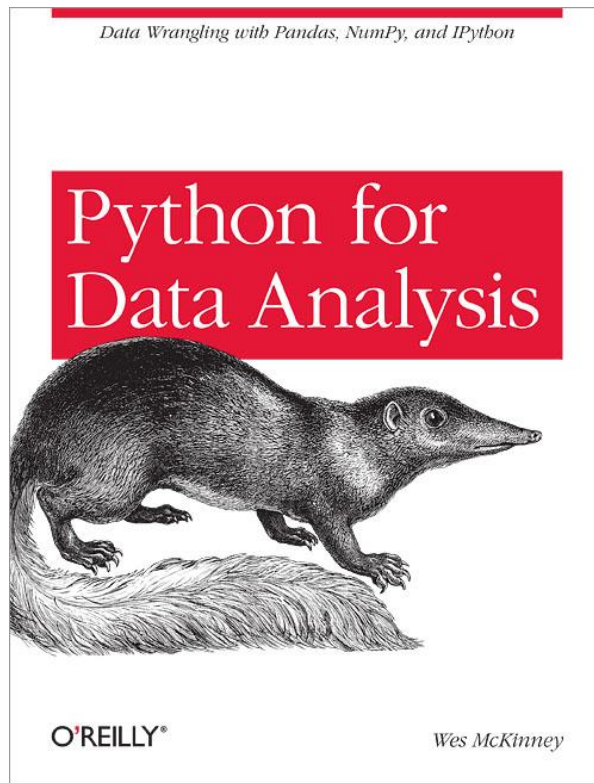
Transformations

Let's transform the variable of interest into categories (factors).

<https://www.epfl.ch/education/educational-initiatives/jupyter-notebooks-for-education/teaching-and-learning-with-jupyter-notebooks/projects-and-labs-with-jupyter-notebooks/>



# Pour aller plus loin



# Pandas

Exercices pratiques disponibles ici :

<https://github.com/dis-unige/formations>

1. Importer des données
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/user\\_guide/10min.html](https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html)

# Sources

Cheat Sheets distribués dans le cours :

- Jupyter notebook :

<https://www.datacamp.com/cheat-sheet/jupyter-notebook-cheat-sheet>

- Markdown :

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

- Pandas :

[https://pandas.pydata.org/Pandas\\_Cheat\\_Sheet.pdf](https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf)