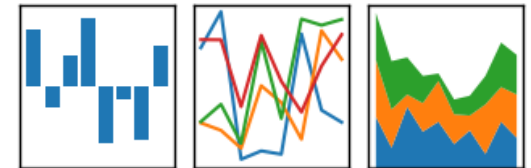


# Midis de l'info scientifique

## 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) / CODIS

09 et 10 avril 2019

# 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

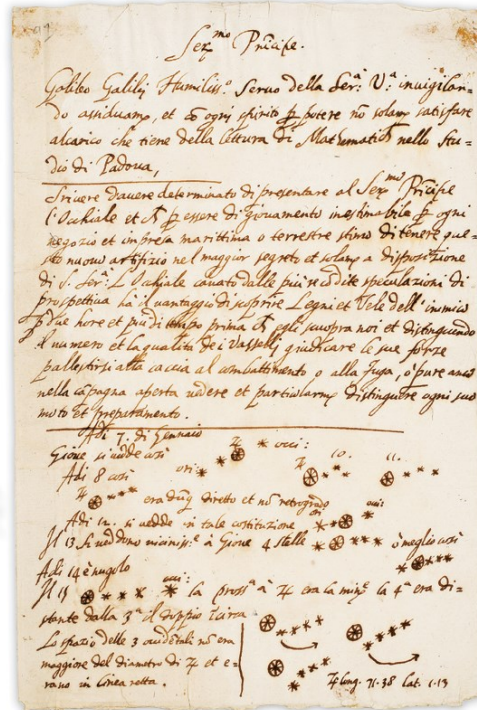
# Introduction

## Historique

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

# Introduction

# Historique

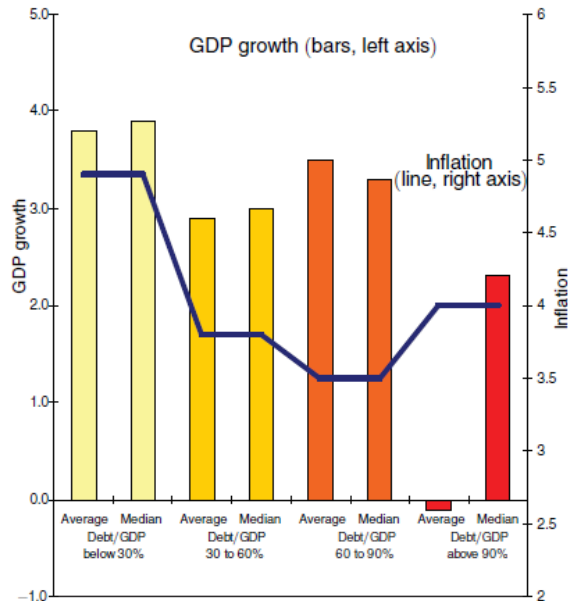


[https://commons.wikimedia.org/wiki/File:Galileo\\_manuscript.png](https://commons.wikimedia.org/wiki/File:Galileo_manuscript.png)

# Introduction

## 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)



Opinion

OP-ED COLUMNIST

### The Excel Depression

The Excel Depression



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](#),

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



# Introduction

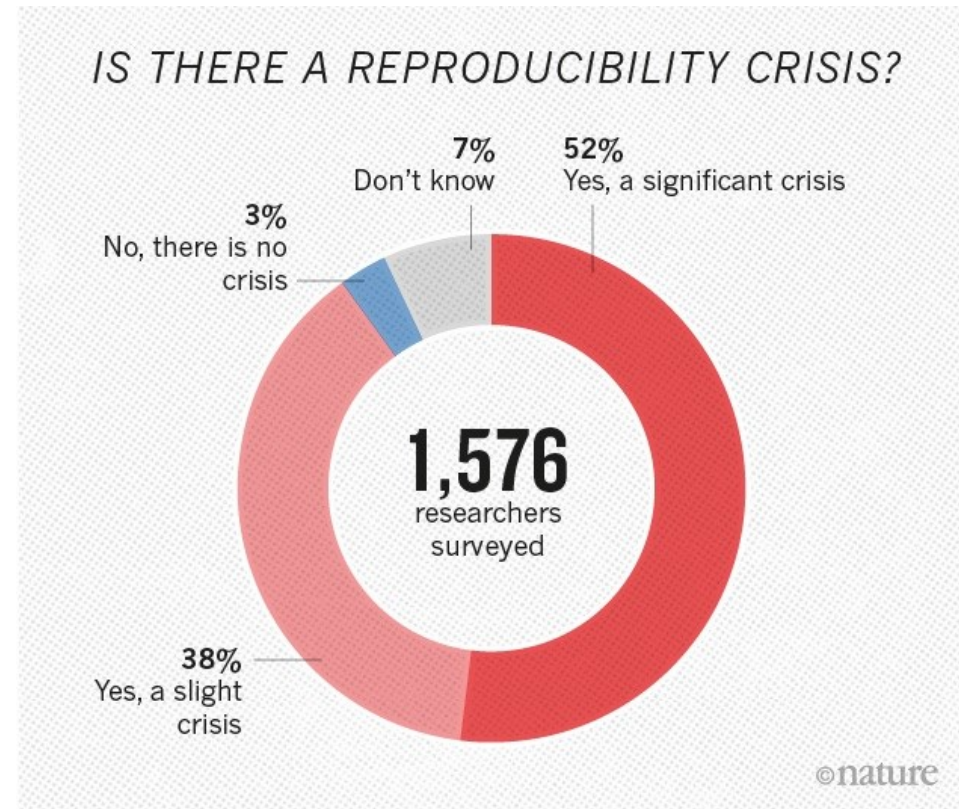
## Reproductibilité et Open Science

La science en crise?

1,500 scientists lift the lid on reproducibility

Baker 2016, Nature 533

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



# Introduction

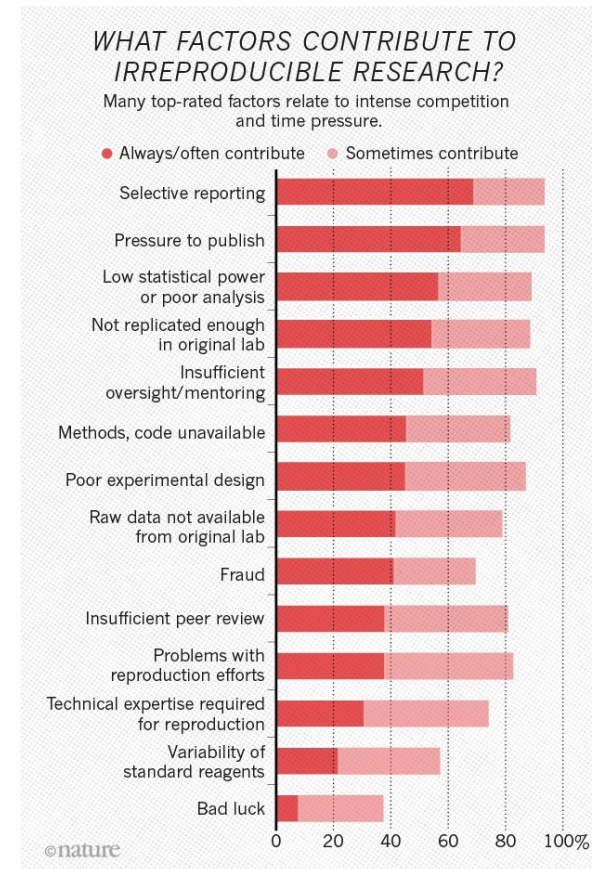
## Reproductibilité et Open Science

La science en crise?

1,500 scientists lift the lid on reproducibility

Baker 2016, Nature 533

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



# Introduction

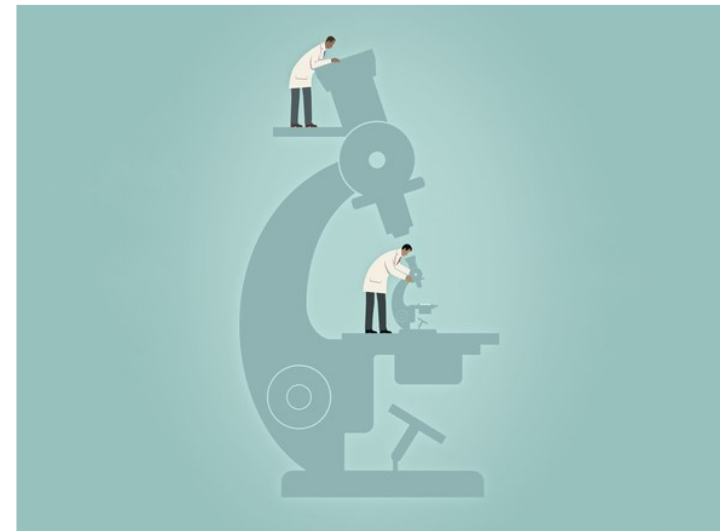
## Reproductibilité et Open Science

Wired

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

MEGAN MEYER OPINION 04.19.17 07:30 AM

## WANT TO FIX SCIENCE'S REPLICATION CRISIS? THEN REPLICATE





# Introduction

## Big Data et Open Data

### Quantifying the Data Deluge and the Data Drought

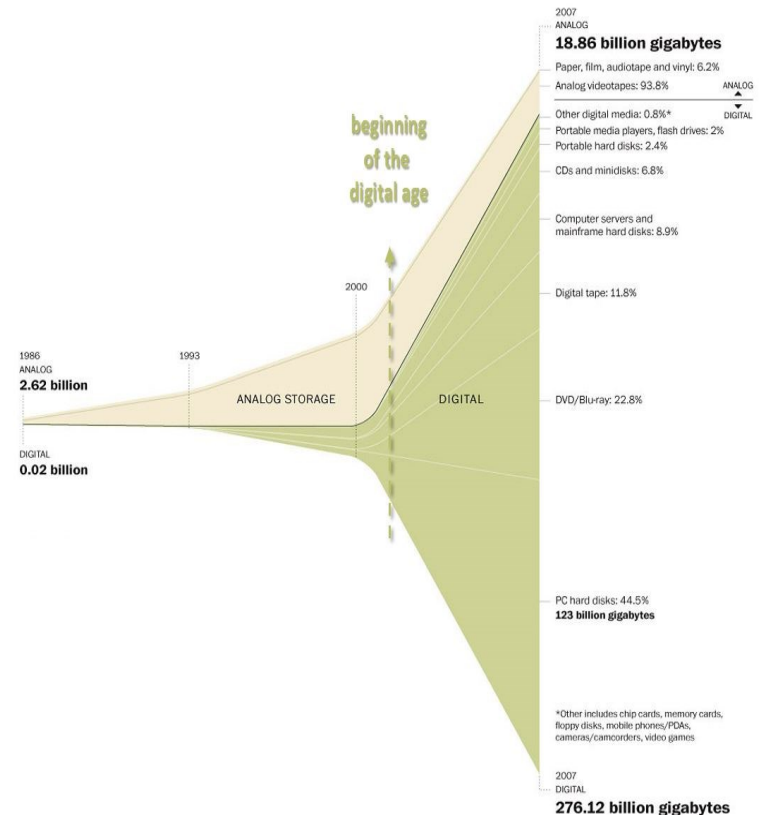
[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2984851](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2984851)

Nombreux réservoirs ouverts

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

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

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

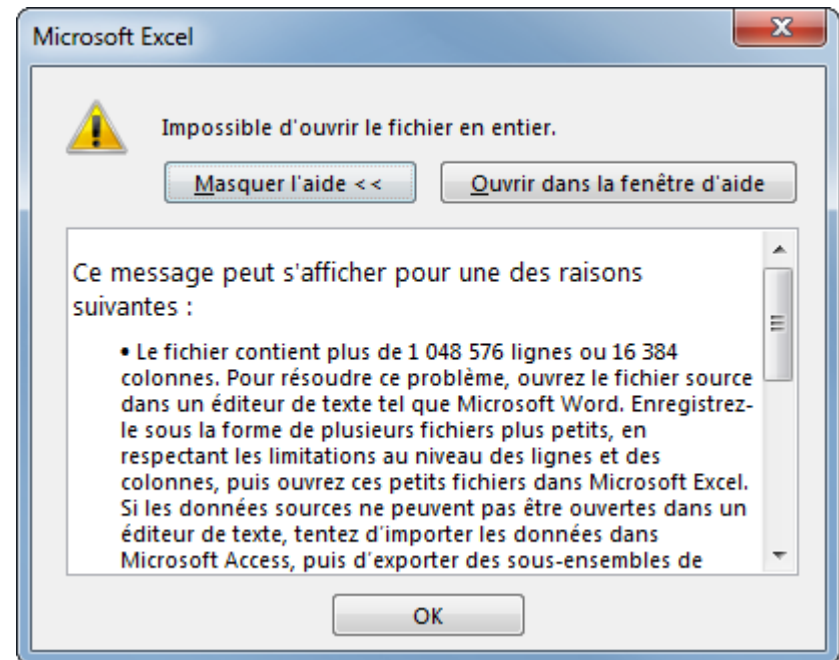


# Introduction

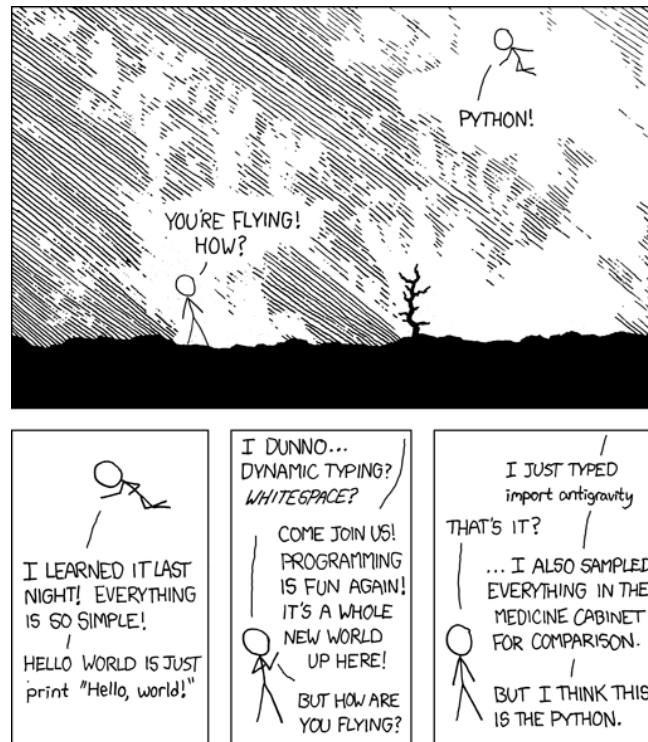
## Excel : limitations

### Liste complète :

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



# Introduction

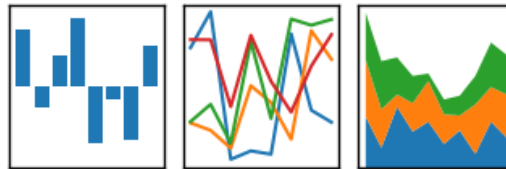


<https://www.xkcd.com/353/>

# 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>

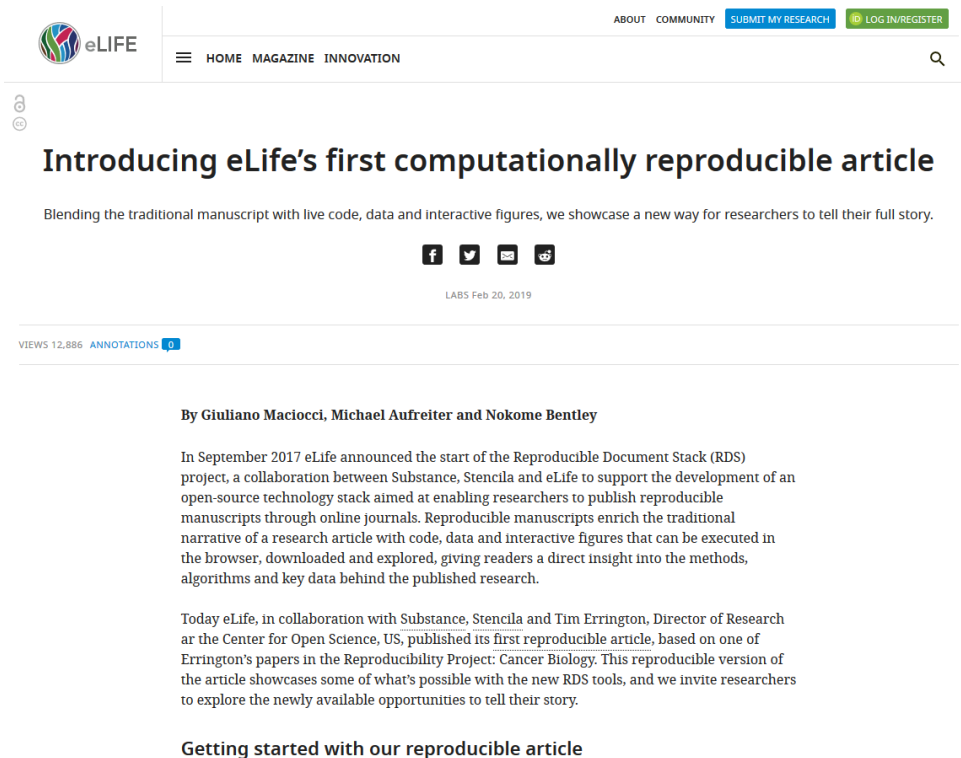


# Introduction

## Reproductibilité et Open Science

eLife

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



The screenshot shows the eLife website header with navigation links: HOME, MAGAZINE, INNOVATION. The article title is "Introducing eLife's first computationally reproducible article". Below the title is a subtitle: "Blending the traditional manuscript with live code, data and interactive figures, we showcase a new way for researchers to tell their full story." Social media icons for Facebook, Twitter, Email, and LinkedIn are present. The article is dated "LABS Feb 20, 2019". The authors are listed as "By Giuliano Maciocci, Michael Aufreiter and Nokome Bentley". The article text describes the start of the Reproducible Document Stack (RDS) project in September 2017, a collaboration between Substance, Stencila, and eLife. It mentions that the project aims to enable researchers to publish reproducible manuscripts through online journals. The article also mentions that the first reproducible article was published by Tim Errington, Director of Research at the Center for Open Science, US, based on one of his papers in the Reproducibility Project: Cancer Biology. The article concludes by stating that the new RDS tools will allow researchers to explore the newly available opportunities to tell their story.

**Introducing eLife's first computationally reproducible article**

Blending the traditional manuscript with live code, data and interactive figures, we showcase a new way for researchers to tell their full story.

LABS Feb 20, 2019

By Giuliano Maciocci, Michael Aufreiter and Nokome Bentley

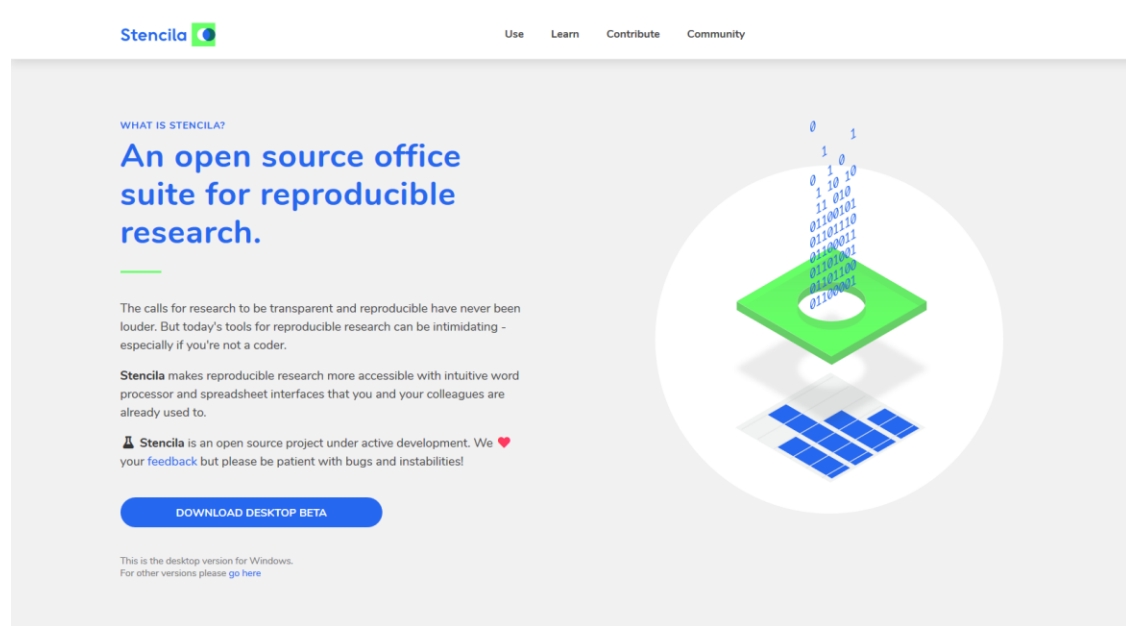
In September 2017 eLife announced the start of the Reproducible Document Stack (RDS) project, a collaboration between Substance, Stencila and eLife to support the development of an open-source technology stack aimed at enabling researchers to publish reproducible manuscripts through online journals. Reproducible manuscripts enrich the traditional narrative of a research article with code, data and interactive figures that can be executed in the browser, downloaded and explored, giving readers a direct insight into the methods, algorithms and key data behind the published research.

Today eLife, in collaboration with Substance, Stencila and Tim Errington, Director of Research at the Center for Open Science, US, published its first reproducible article, based on one of Errington's papers in the Reproducibility Project: Cancer Biology. This reproducible version of the article showcases some of what's possible with the new RDS tools, and we invite researchers to explore the newly available opportunities to tell their story.

Getting started with our reproducible article

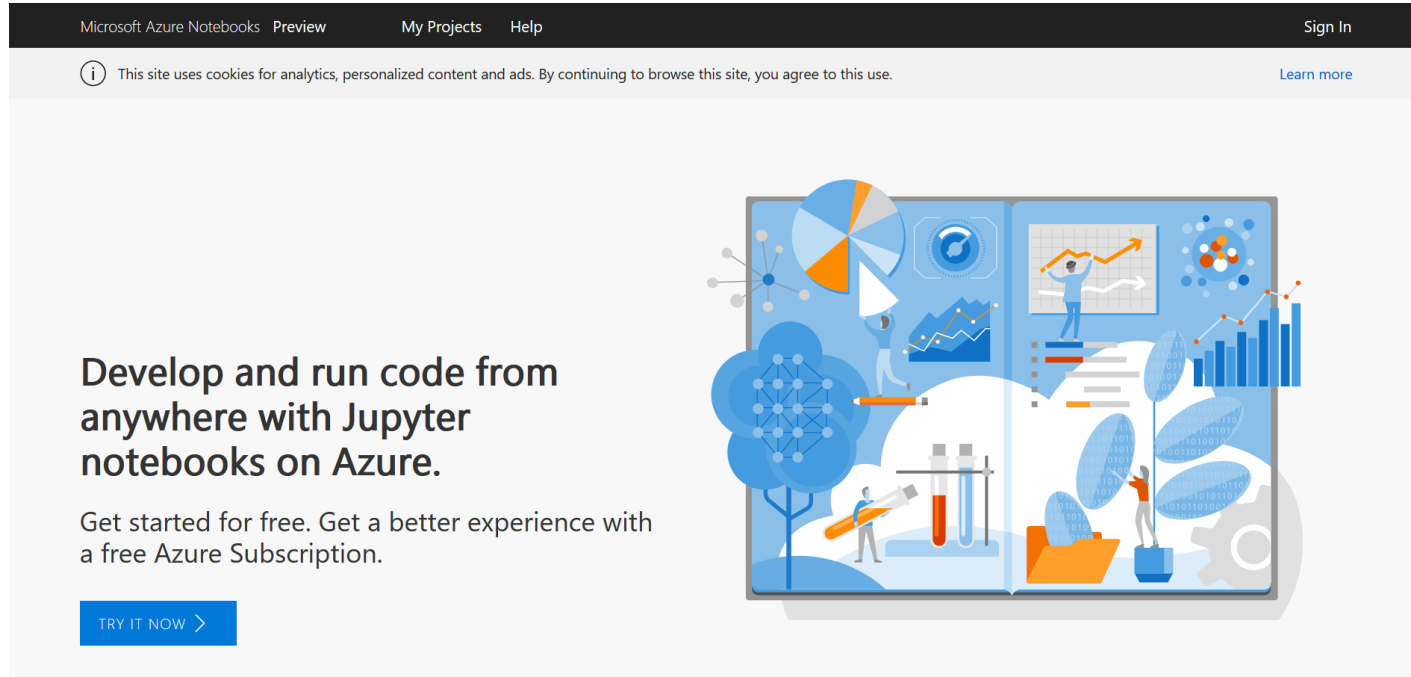
# Introduction

## 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 banner 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 with a magnifying glass, and a person with a pencil. 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 >".

Microsoft Azure Notebooks Preview My Projects Help Sign In

*i* This site uses cookies for analytics, personalized content and ads. By continuing to browse this site, you agree to this use. [Learn more](#)

**Develop and run code from anywhere with Jupyter notebooks on Azure.**

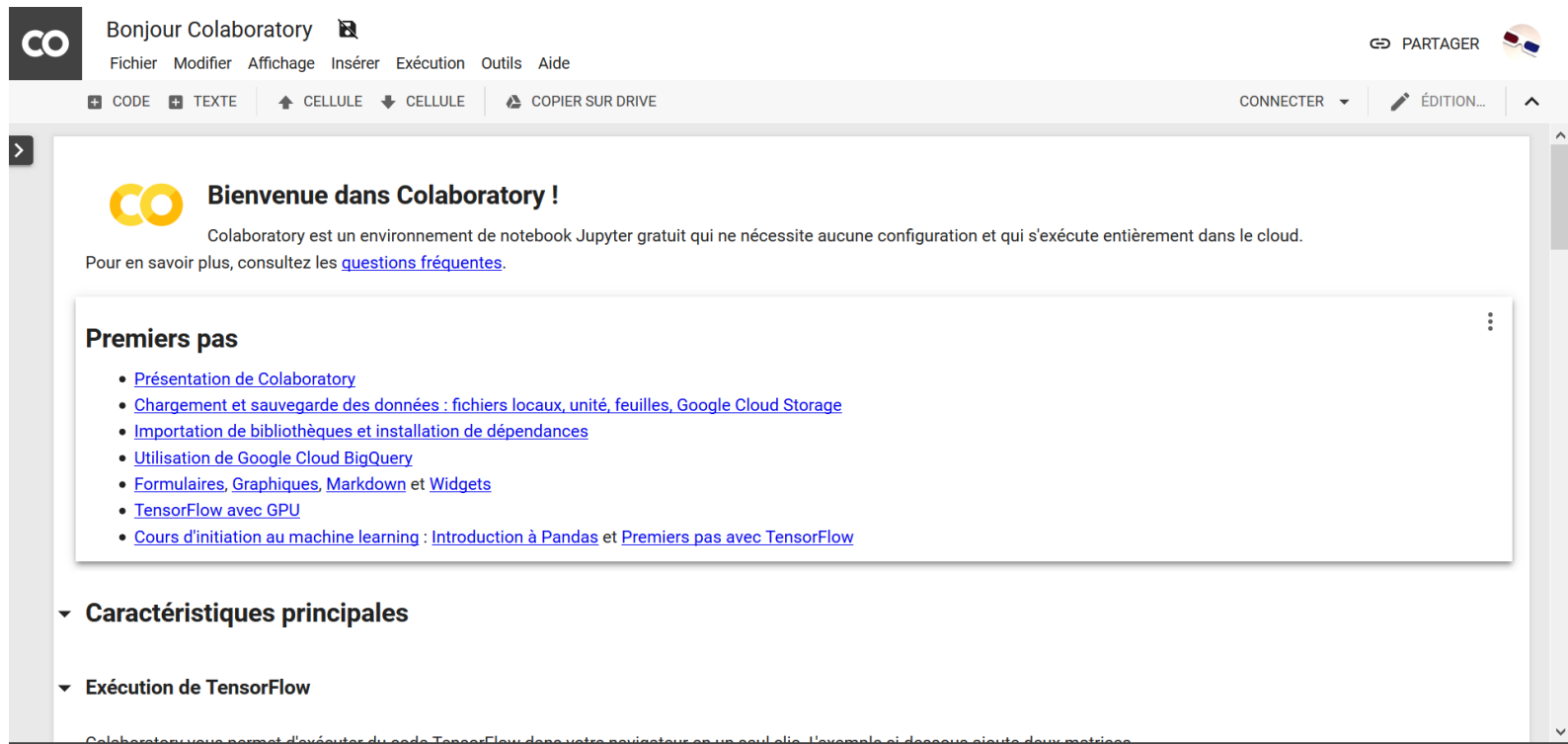
Get started for free. Get a better experience with a free Azure Subscription.

[TRY IT NOW >](#)

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



# Introduction



Bonjour Colaboratory

Fichier Modifier Affichage Insérer Exécution Outils Aide

CODE TEXTE CELLULE CELLULE COPIER SUR DRIVE

CONNECTER ÉDITION...

**Bienvenue dans Colaboratory !**

Colaboratory est un environnement de notebook Jupyter gratuit qui ne nécessite aucune configuration et qui s'exécute entièrement dans le cloud. Pour en savoir plus, consultez les [questions fréquentes](#).

**Premiers pas**

- [Présentation de Colaboratory](#)
- [Chargement et sauvegarde des données : fichiers locaux, unité, feuilles, Google Cloud Storage](#)
- [Importation de bibliothèques et installation de dépendances](#)
- [Utilisation de Google Cloud BigQuery](#)
- [Formulaires, Graphiques, Markdown et Widgets](#)
- [TensorFlow avec GPU](#)
- [Cours d'initiation au machine learning : Introduction à Pandas et Premiers pas avec TensorFlow](#)

▼ **Caractéristiques principales**

▼ **Exécution de TensorFlow**

Colaboratory vous permet d'exécuter du code TensorFlow dans votre navigateur en un seul clic. L'exemple ci-dessous ajoute deux matrices.

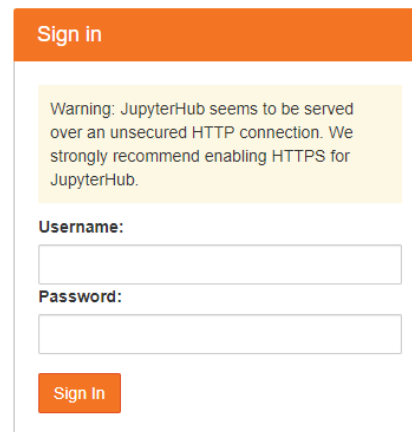
<https://colab.research.google.com>

# Jupyter Notebooks

## Travail sur le JupyterHub du cours

Se connecter sur cette adresse avec le login/pwd qui figure sur le post-it de l'ordinateur :

<http://68.183.213.32>



Sign in

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

Username:

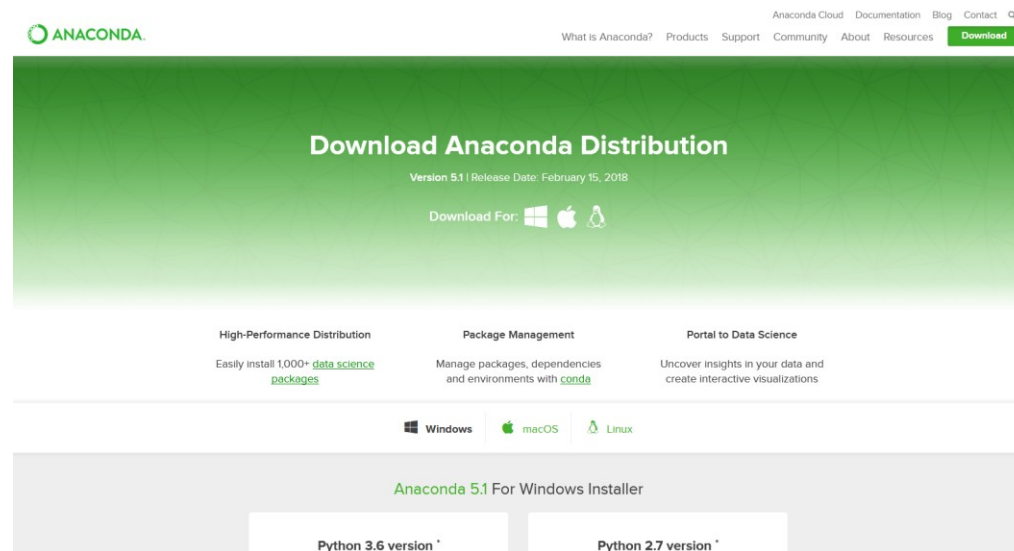
Password:

Sign In

# Jupyter Notebooks

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

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



# Jupyter Notebooks

Packages compris dans l'installation :

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

Liste complète :

[https://docs.anaconda.com/anaconda/packages/py3.6\\_win-64](https://docs.anaconda.com/anaconda/packages/py3.6_win-64)



# Jupyter Notebooks

**Créer, organiser et partager des notebooks**

Lancer Anaconda -> Jupyter Notebook



# 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

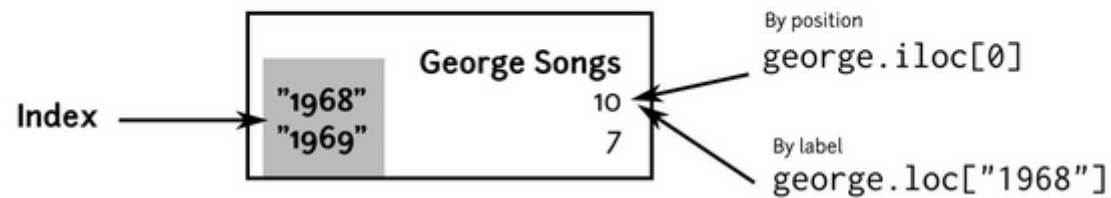
Series : 1 dimension

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





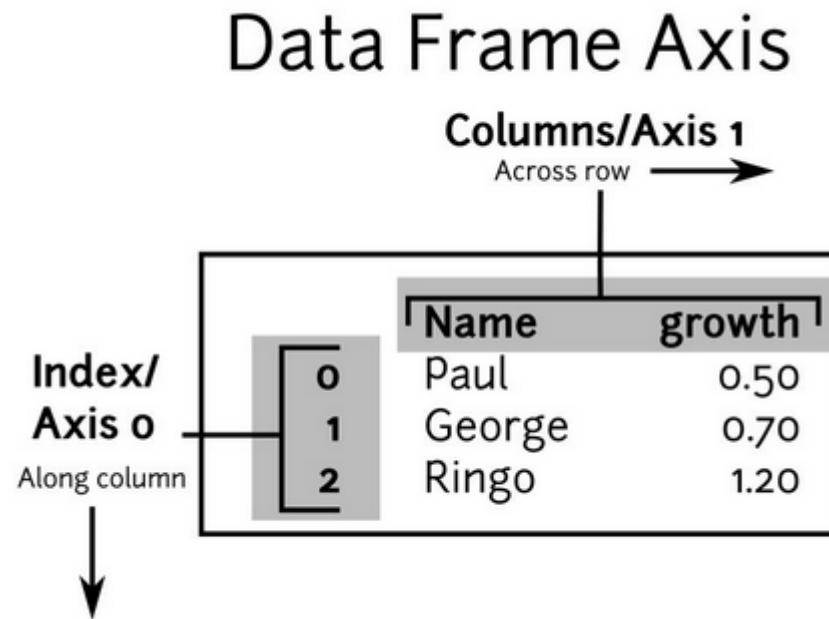
# Pandas

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

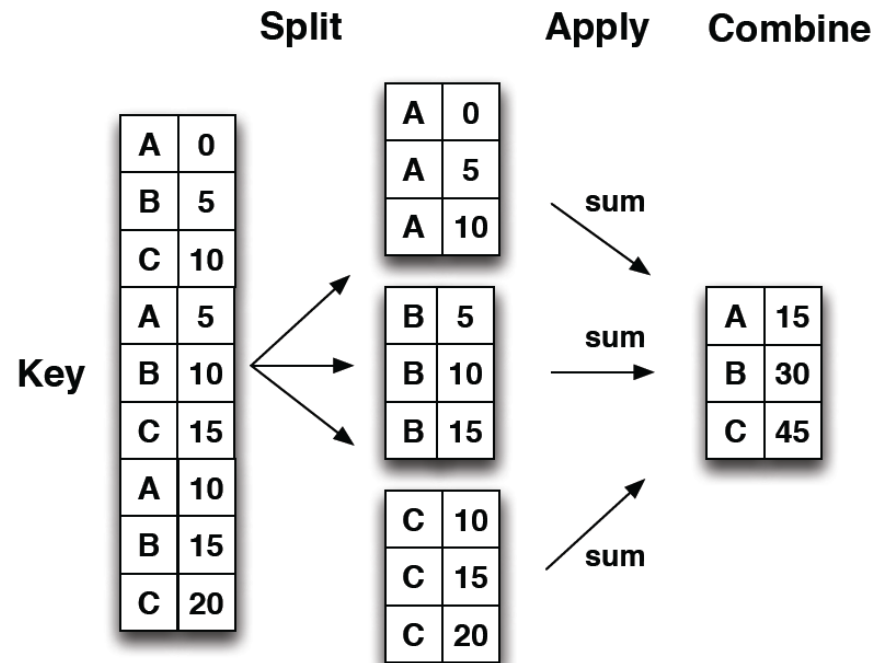
# Pandas

Opérations facilitées par les index : jointures automatiques

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

# Pandas

## Opérations : GroupBy





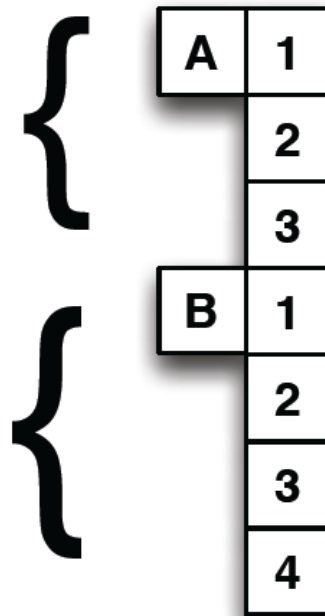
# Pandas

## 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

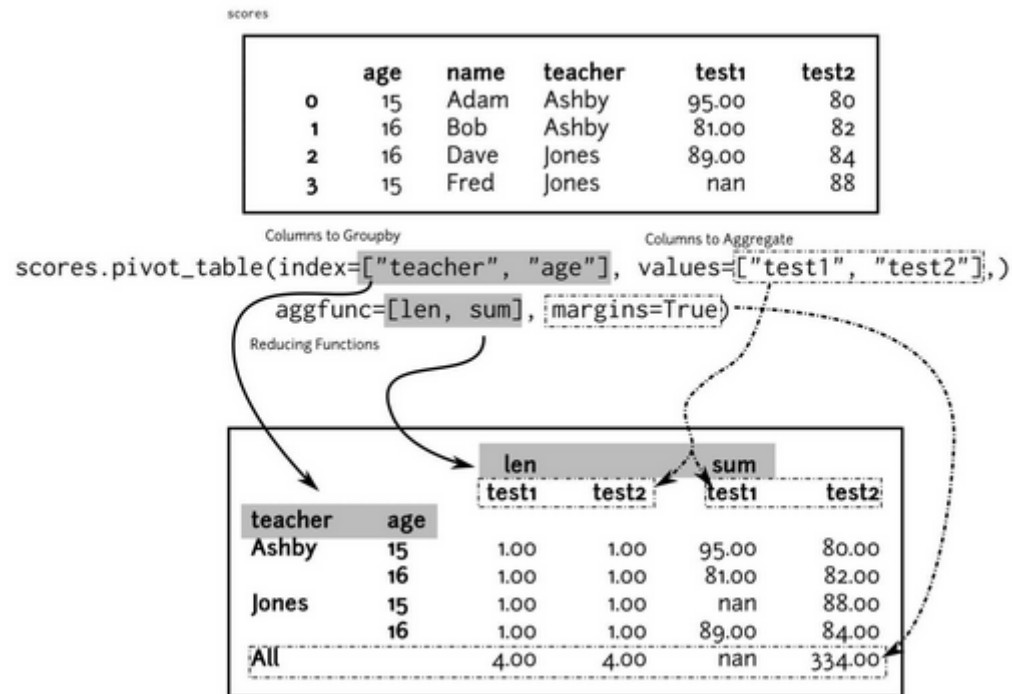
# Pandas

## Index multidimensionnels



# Pandas

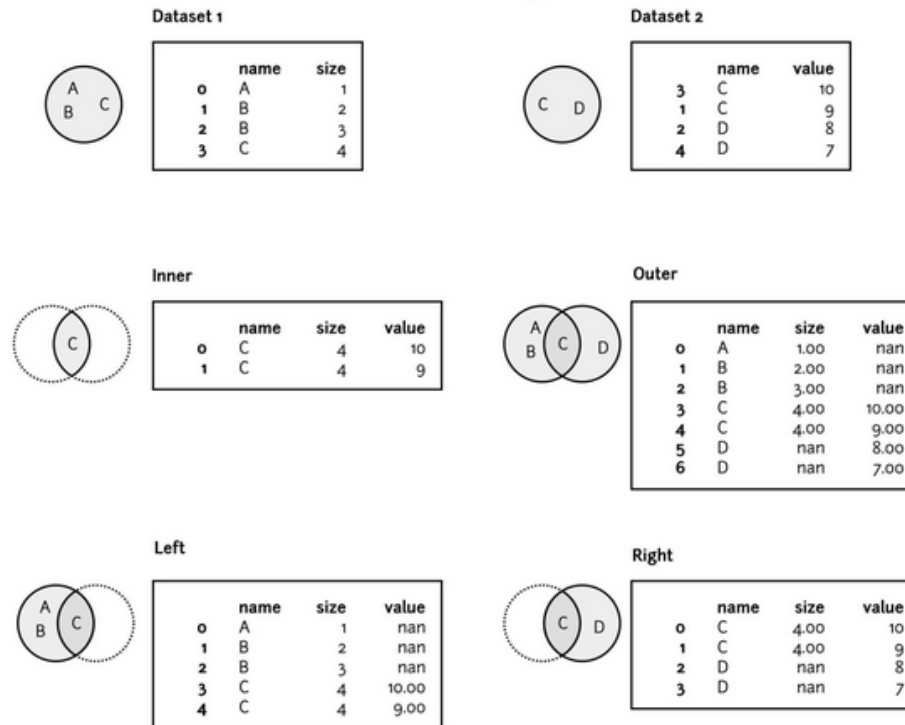
## Tables pivot



# 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

## 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

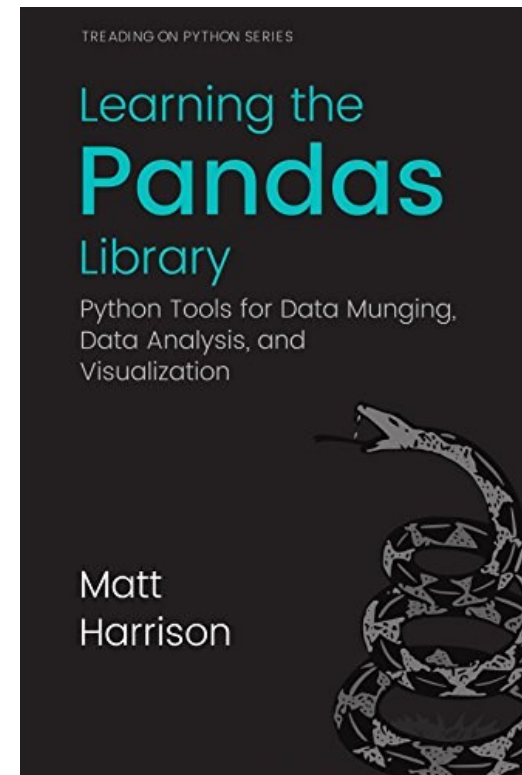
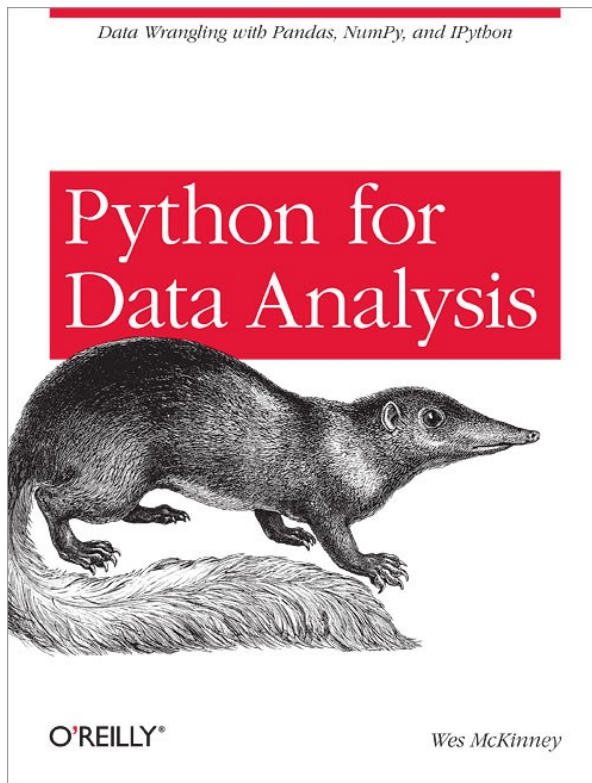
<https://github.com/jupyter/>

## 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



# 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)