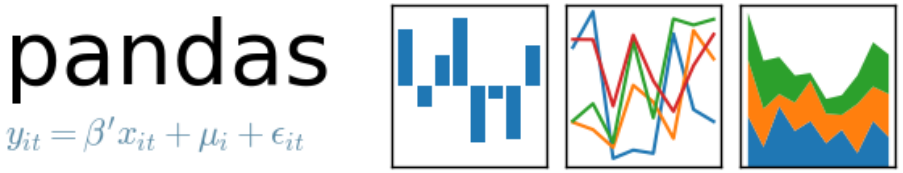
Rendez-vous de l’info scientifique

Traitement de données avec Pandas & Jupyter notebooks



Pablo Iriarte – pablo.iriarte@unige.ch / DIS

16 avril 2025

Programme

Introduction

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

Jupyter Notebooks

- Famille d’outils
- Installation via la distribution Anaconda ou test en ligne via https://jupyter.org/try
- Créer, organiser et partager des notebooks

Pandas

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

Introduction

Historique de deux outils *Open Source*

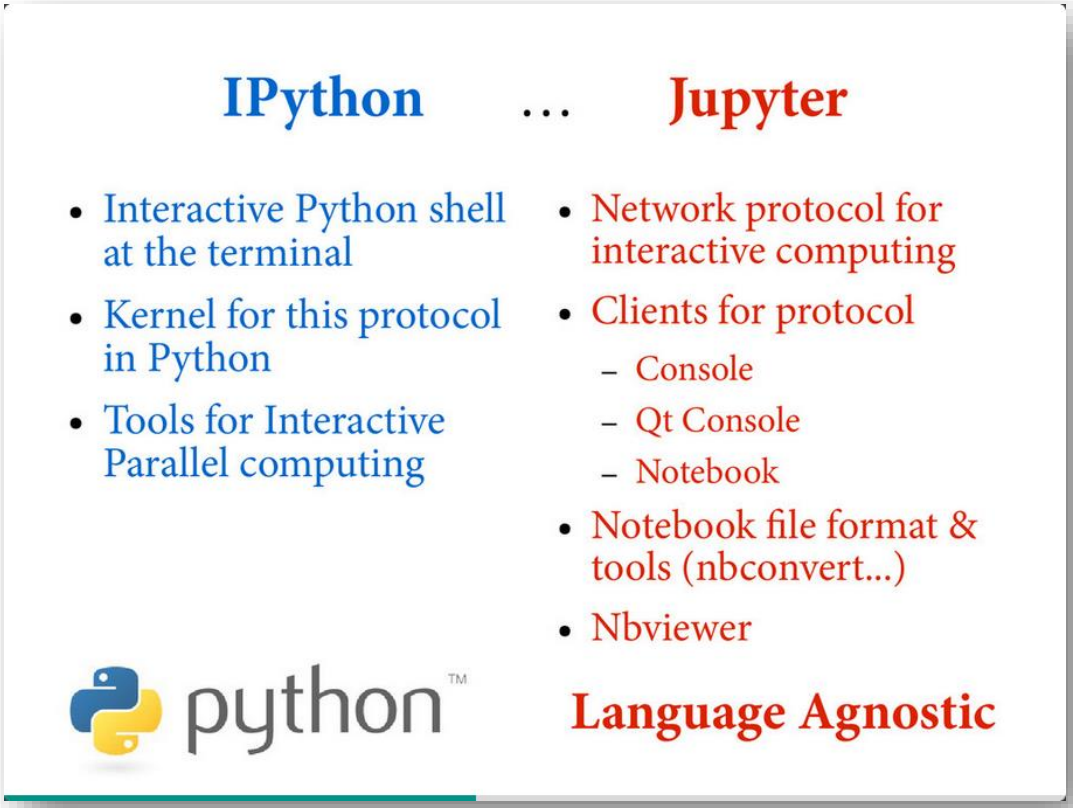
- **iPython** (2001->) https://ipython.org/

  “ When I found out about IPP and LazyPython I tried to join all three into a unified system. I thought this could provide a very nice working environment, both for regular programming and scientific computing: shell-like features, IDL/Matlab numerics, Mathematica-type prompt history and great object introspection and help facilities. I think it worked reasonably well, though it was a lot more work than I had initially planned.” Fernando Perez, graduate student at the University of Colorado, Boulder

- **Jupyter** (2014 ->) https://jupyter.org/

  Spin-off de IPython appellé “Project Jupyter”. IPython continue en tant que Python shell et kernel sur Jupyter mais il n’est plus le seul langage supporté.

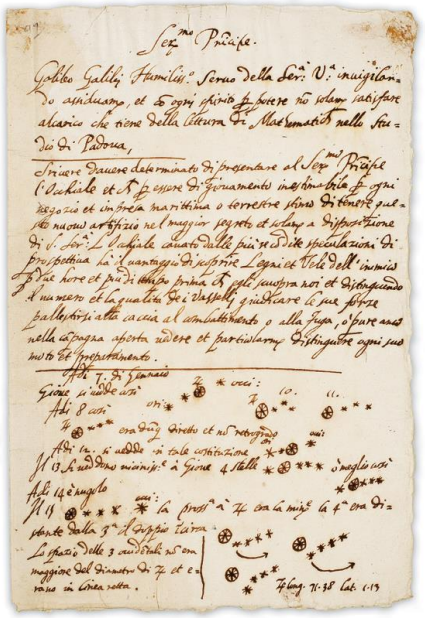
Introduction



https://speakerdeck.com/fperez/project-jupyter?slide=5

Introduction

Historique : Julia + Python + R & Hommage à Galileo



https://en.wikipedia.org/wiki/Project_Jupyter

Introduction

Famille d'outils

- **Jupyter Notebook** (le classique) : https://jupyter.org/
- **Jupyter Lab** (interface étendue) : https://jupyterlab.readthedocs.io/
- **Jupyter Hub** (version pour labos, etc.) : https://jupyterhub.readthedocs.io/
- **Voilà** (transforme le notebook en application Web) : https://voila-gallery.org/
- **NB viewer** (partage et visualisation) : https://nbviewer.jupyter.org/
- **Binder** (partage, visualisation et exécution) : https://mybinder.org/
- **NotebookSharing.space** (partage, visualisation, exécution et annotations) **:** https://notebooksharing.space/
- **GitHub** (visualisation et versioning) : https://github.com
- Etc.

Introduction

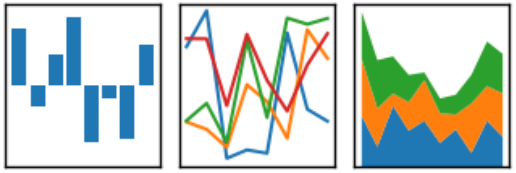
Pléthore de « librairies » python

- **NumPy** (scientific computing) : https://numpy.org/
- **Pandas** (analyse et manipulation de données) : https://pandas.pydata.org/
- **Matplotlib** (graphiques et visualisations) : https://matplotlib.org/
- **Dask** (parallélisme) : https://docs.dask.org
- **Vaex** (données massives, 10^9 lignes par second!) : https://vaex.readthedocs.io
- **Scikit-learn** (Machine Learning) : https://scikit-learn.org
- **TensorFlow** (Deep Learning) : https://www.tensorflow.org/
- **PyTorch** (Deep Learning) : https://pytorch.org/
- Etc.

Introduction

pandas

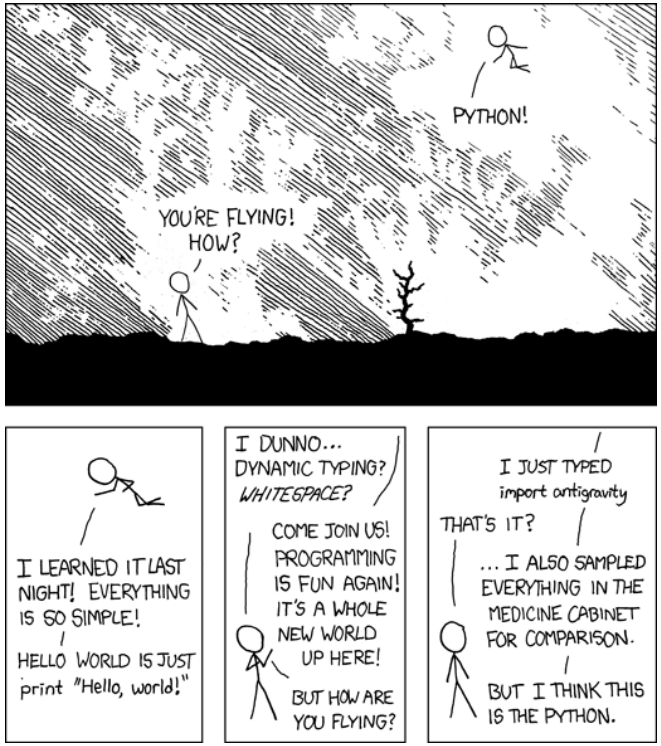
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



https://pandas.pydata.org

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

Introduction

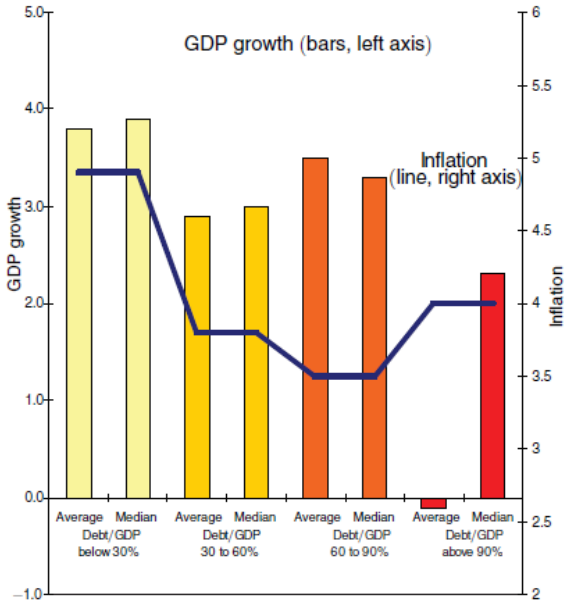


https://www.xkcd.com/353/

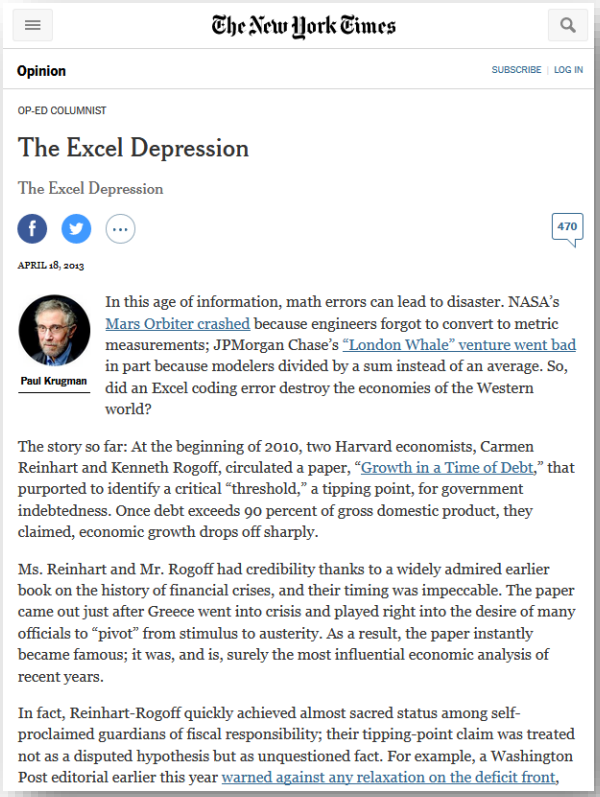
Introduction

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



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,

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

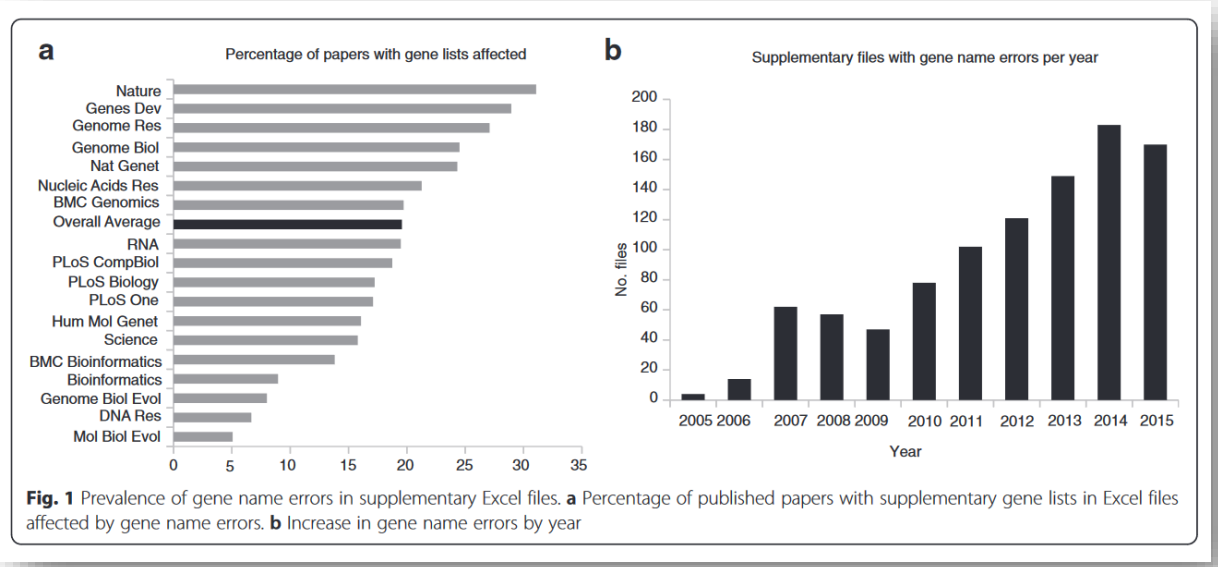
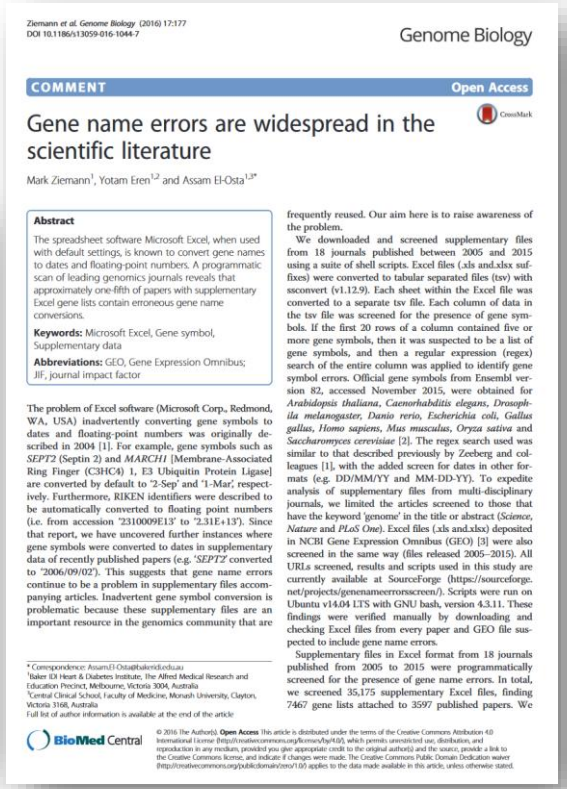


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year



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[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

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 '2310009E13' 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/genenameerrorsscreen/). 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@baker.edu.au
[1]Baker IDI Heart & Diabetes Institute, The Alfred Medical Research and Education Precinct, Melbourne, Victoria 3004, Australia
[3]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

Ziemann et al. 2016. DOI:10.1186/s13059-016-1044-7

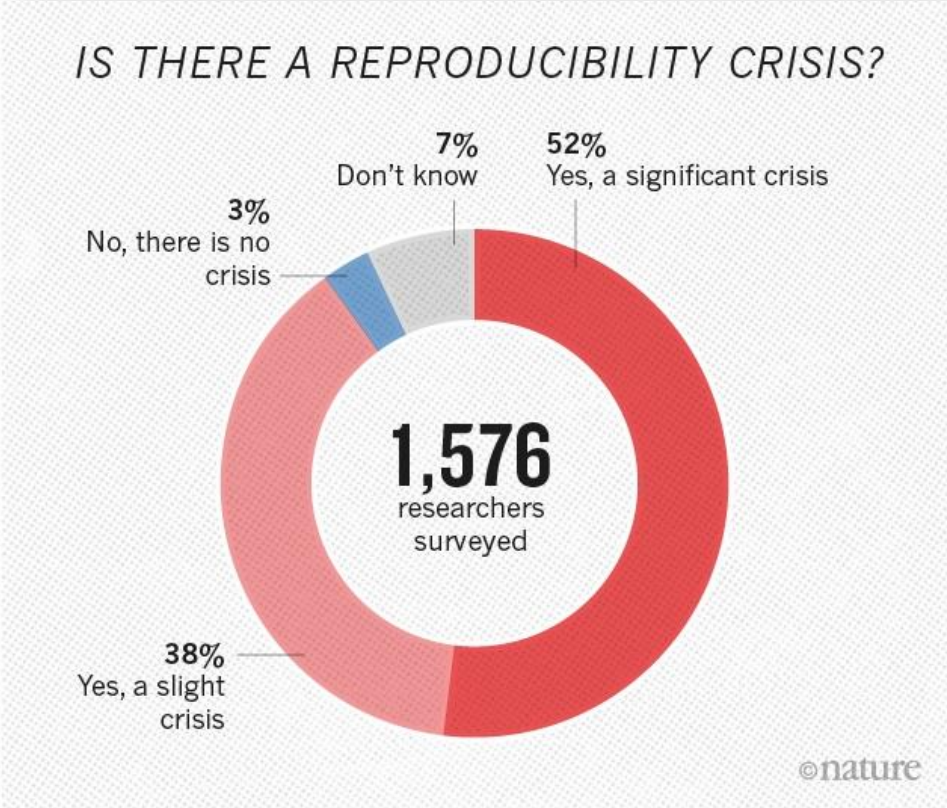
Introduction

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

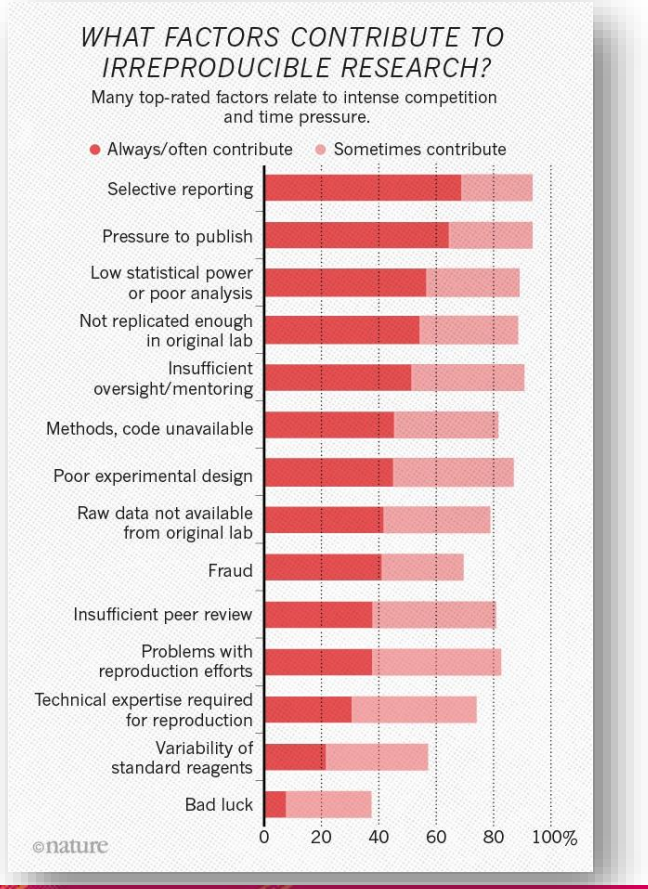
Reproductibilité et Open Science

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

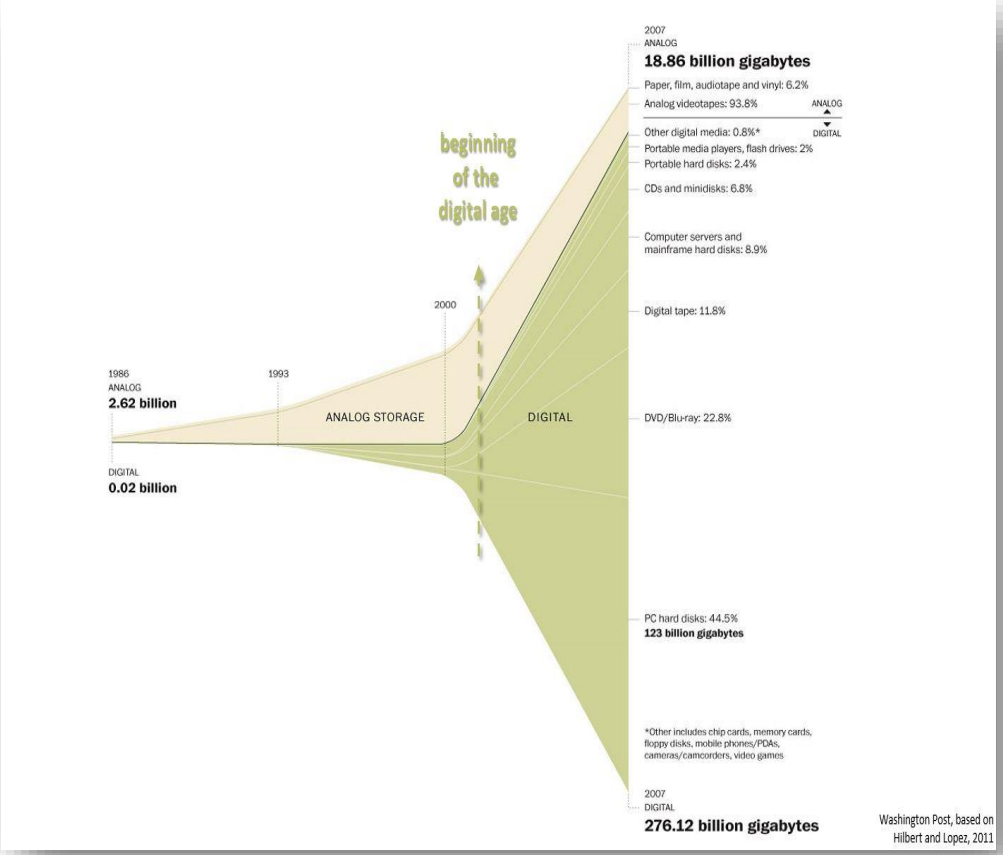
Nombreux réservoirs ouverts

Kaggle : https://www.kaggle.com

Github: http://github.com

WikiData : https://www.wikidata.org

Statistique historique : https://hsso.ch

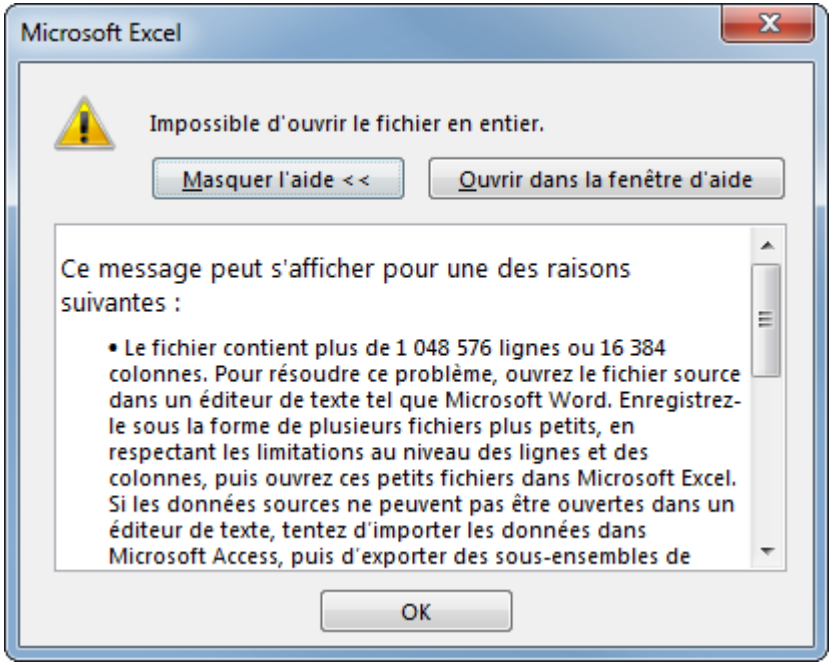


Introduction

Excel : limitations

Worksheet and workbook specifications and limits

Feature	Maximum limit
Open workbooks	Limited by available memory and system resources
Total number of rows and columns on a worksheet	1,048,576 rows by 16,384 columns



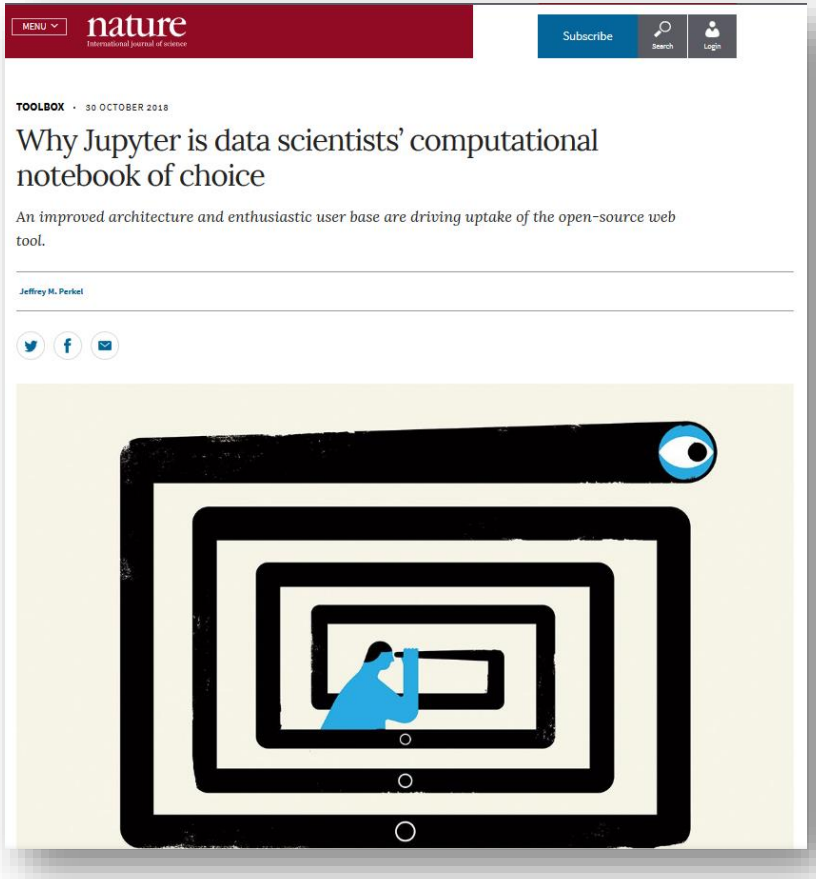
Liste complète des limites : https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3

Introduction

Reproductibilité et Open Science

Nature

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

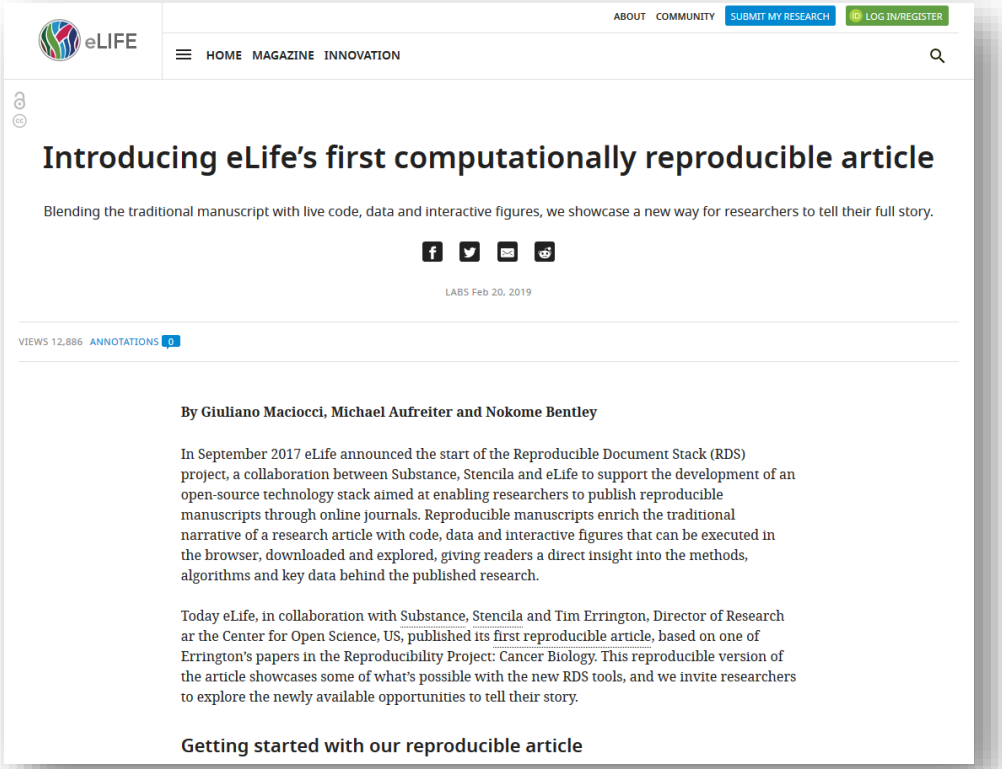


Introduction

Reproductibilité et Open Science

eLife

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



eLIFE ABOUT COMMUNITY SUBMIT MY RESEARCH LOG IN/REGISTER

HOME MAGAZINE INNOVATION

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

VIEWS 12,886 ANNOTATIONS 0

By Giuliano Maciocci, Michael Aufreiter and Nokome Bentley

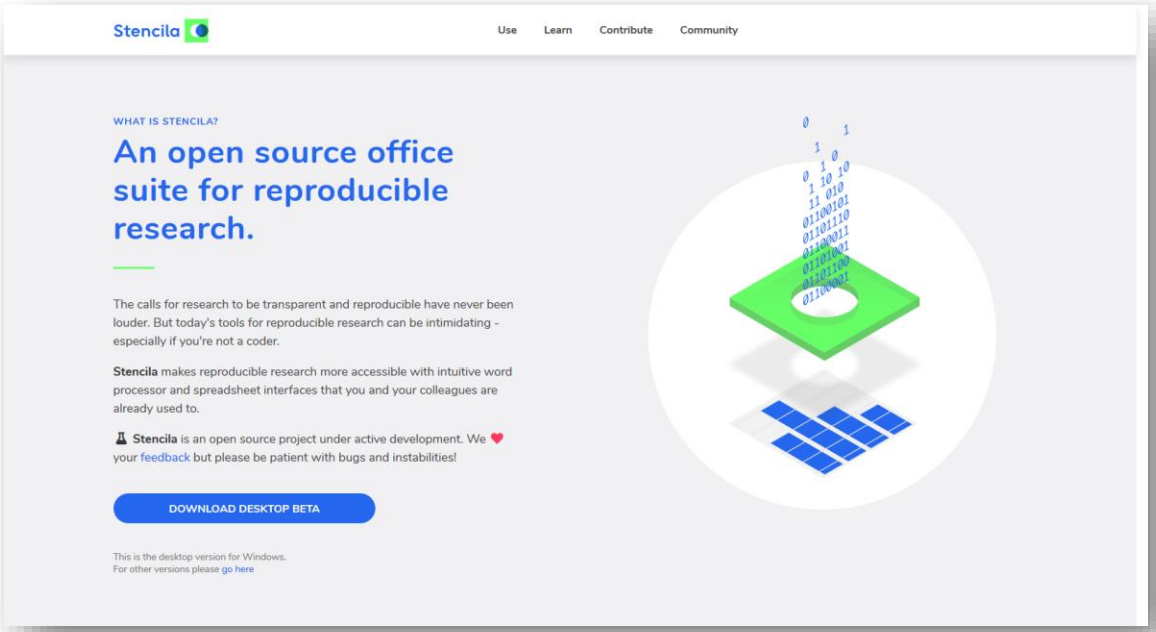
In September 2017 eLife announced the start of the Reproducible Document Stack (RDS) project, a collaboration between Substance, Stencila and eLife to support the development of an open-source technology stack aimed at enabling researchers to publish reproducible manuscripts through online journals. 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 ar 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

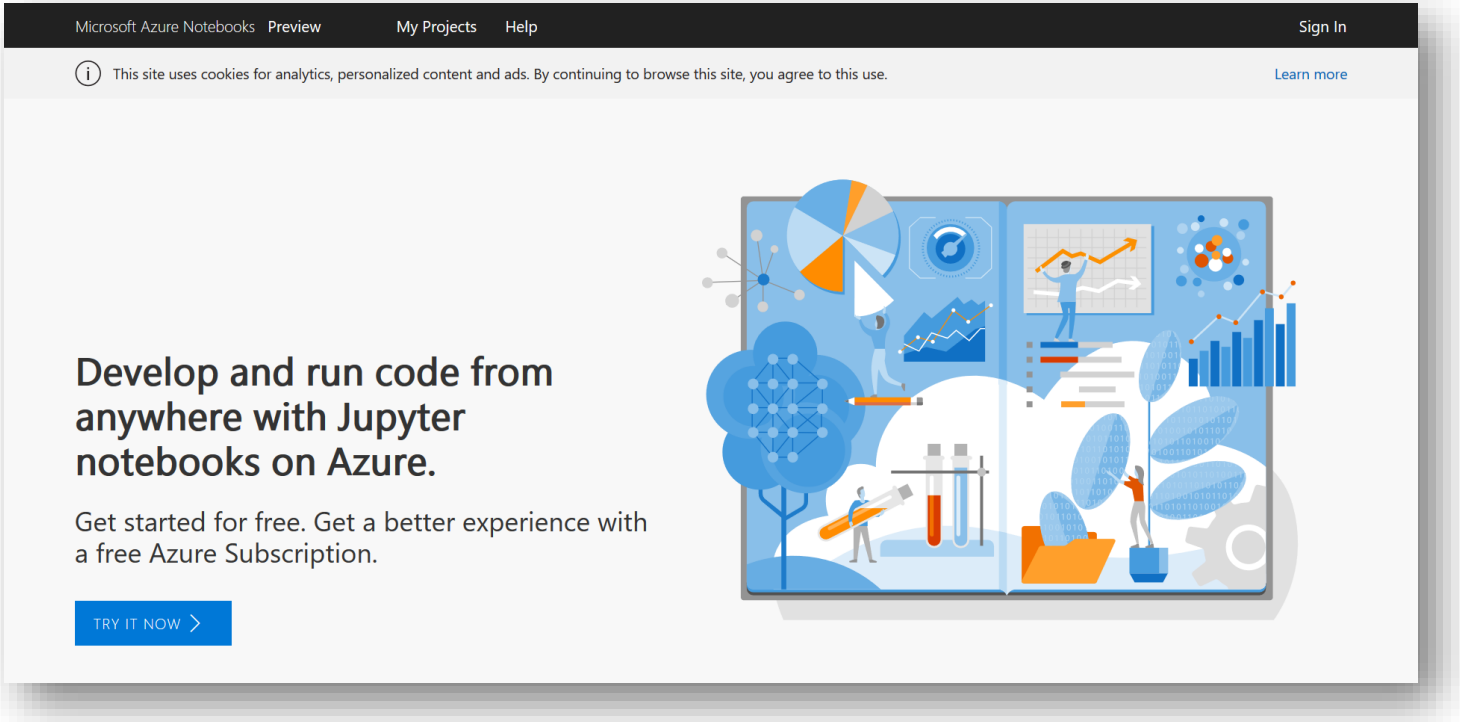
Introduction

Reproductibilité et Open Science



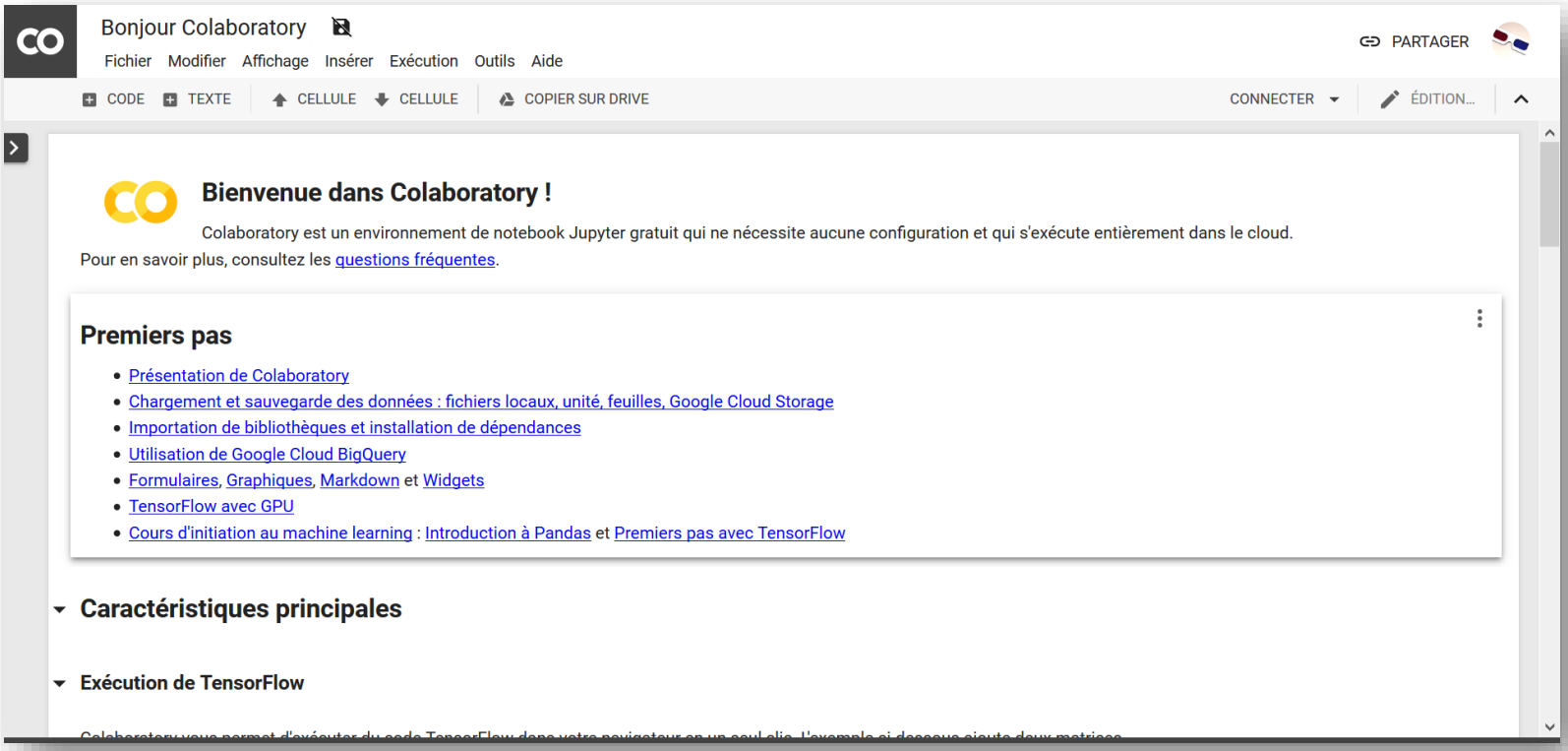
https://stenci.la

Introduction



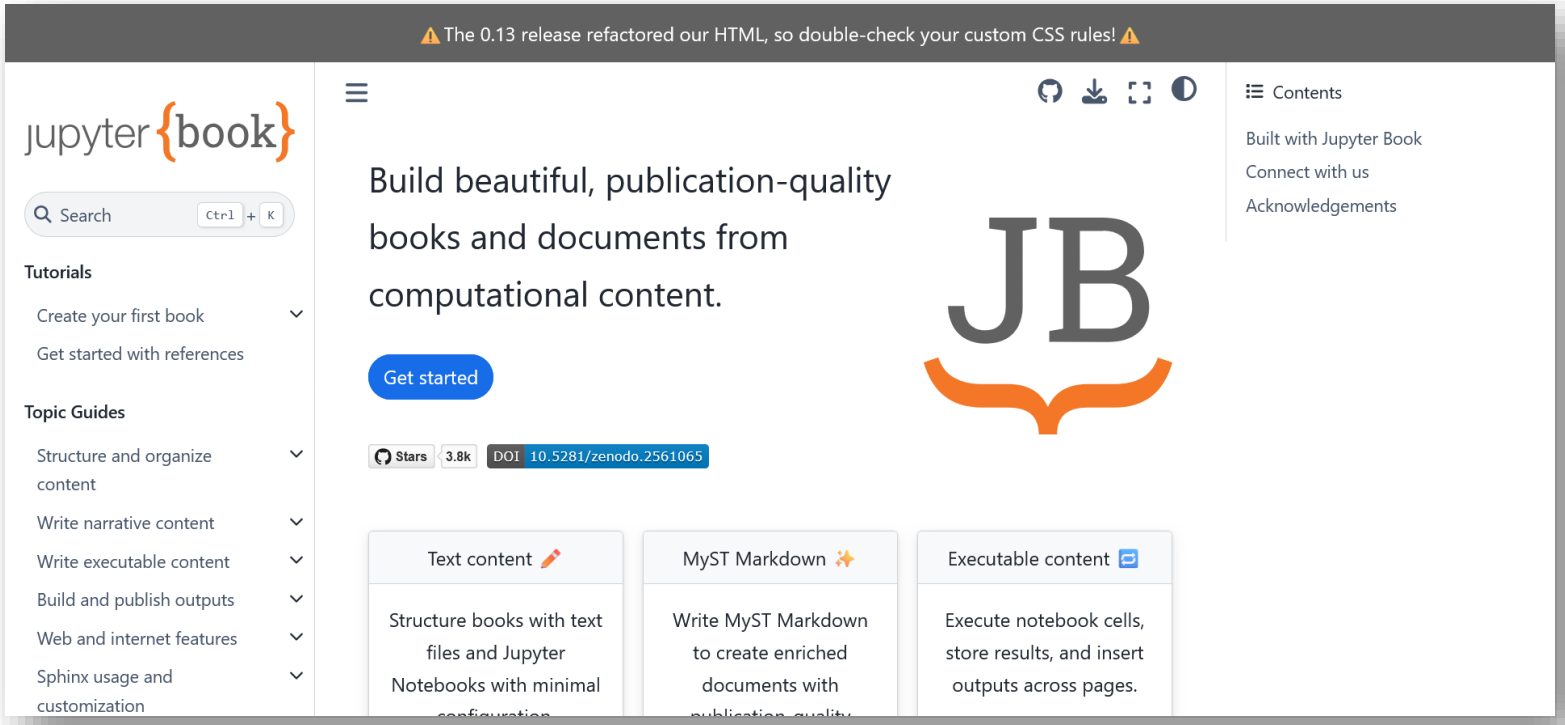
https://notebooks.azure.com/

Introduction



https://colab.research.google.com

Introduction

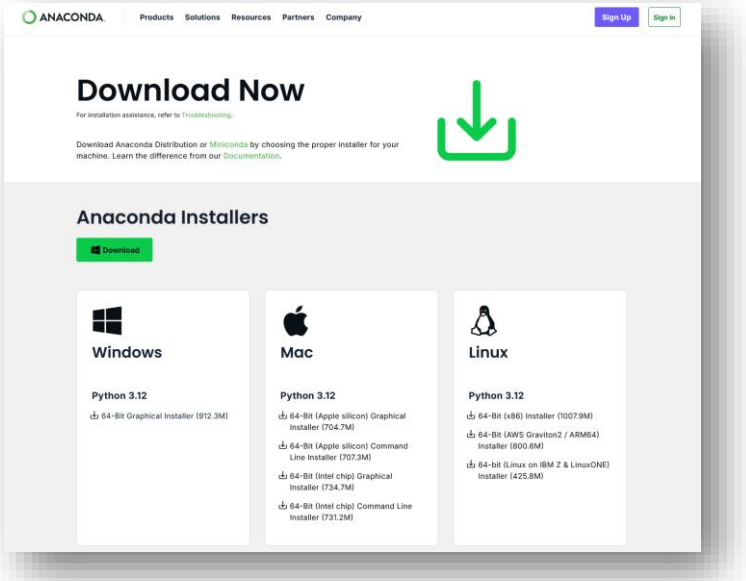


https://jupyterbook.org

Jupyter Notebooks

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

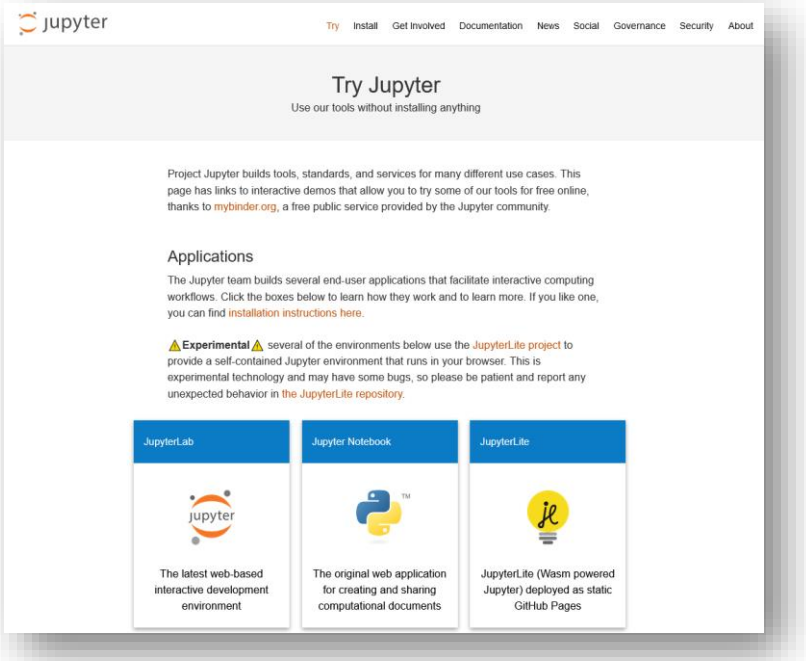
https://www.anaconda.com/download/



Jupyter Notebooks

Ou utilisation sans installation possible sur «Try Jupyter»

https://jupyter.org/try



Jupyter Notebooks

Packages importants (compris avec l'installation Anaconda) :

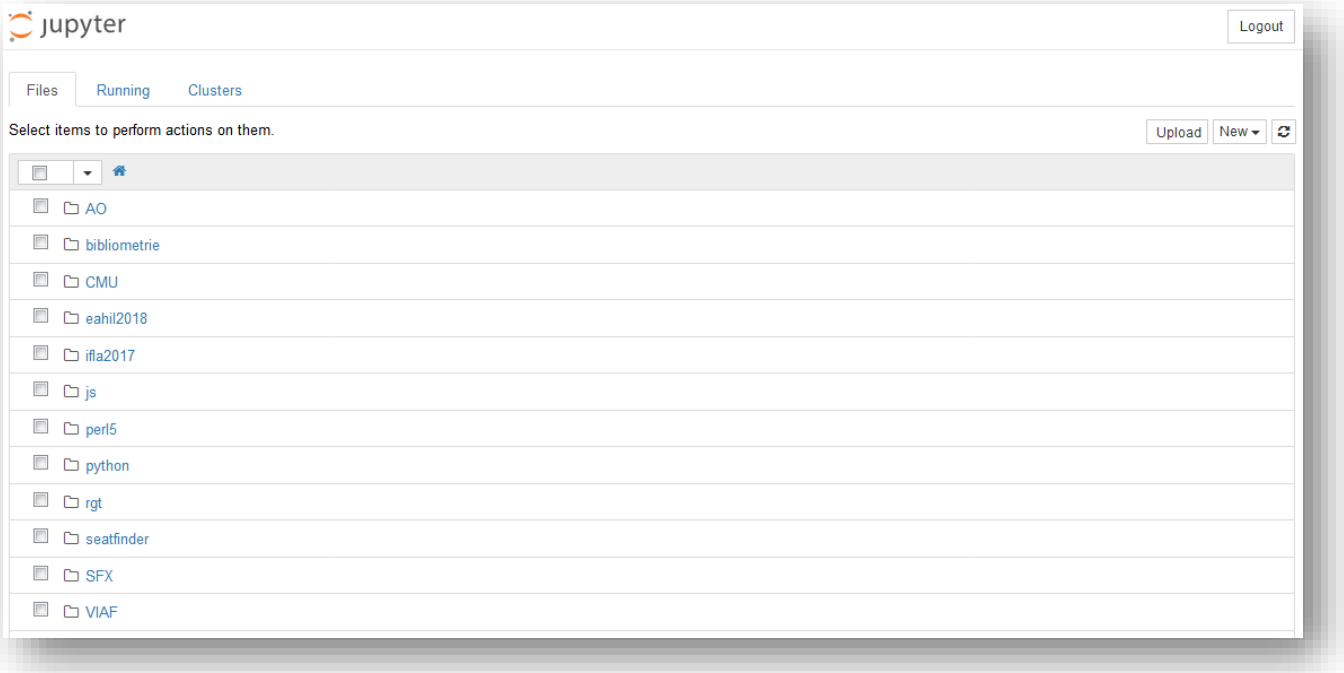
- Notebook (jupyter)
- Pandas
- NumPy
- Matplotlib
- NLTK
- …

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

Jupyter Notebooks

Créer, organiser et partager des notebooks

Lancer Anaconda -> Jupyter Notebook



Jupyter Notebooks

Attention, le dossier par défaut à l'ouverture de Jupyter est le «home» du poste. Sur Windows c'est le dossier : **C:\Users\[votre_id]**

Ensuite on peut naviguer dans les sous-dossier mais pas en dehors !

Pour palier à ça il y a deux solutions :

1. définir l'emplacement du dossier avec la commande :
   > **jupyter notebook --notebook-dir=C:\[dossier souhaité]**
2. créer un lien symbolique pour ajouter le dossier souhaité sur le «home» :
   > **mklink /D Nom-du-lien Dossier-de-destination**

Exercices

Se familiariser avec les notebooks

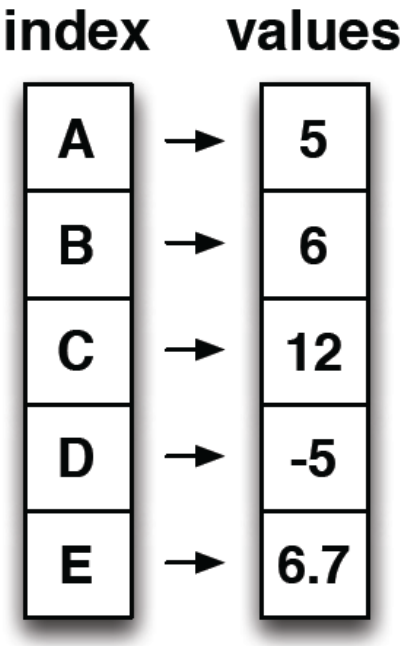
1. Ouvrir un notebook d'exemple (sur le dossier du cours)
2. Créer un nouveau notebook et le renommer
3. Ajouter une cellule de texte (markdown)
4. Ajouter une cellule de code python (calcul simple)
5. L'exporter en format HTML

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

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

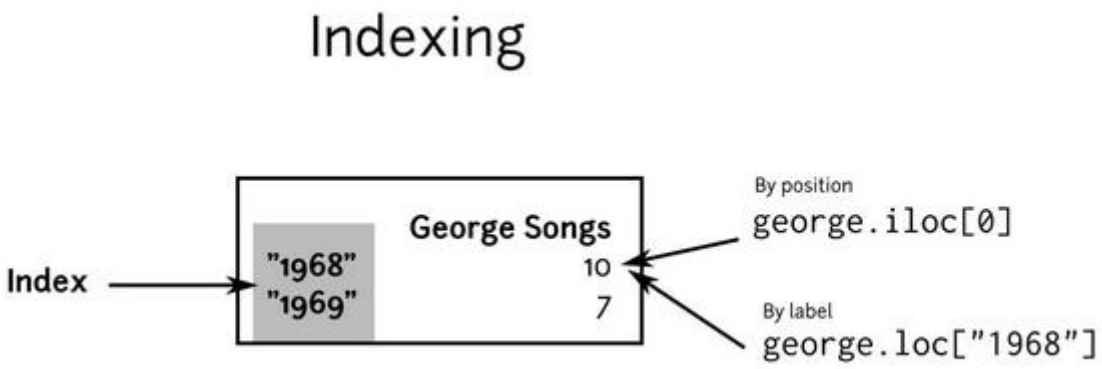
Pandas

Series : 1 dimension



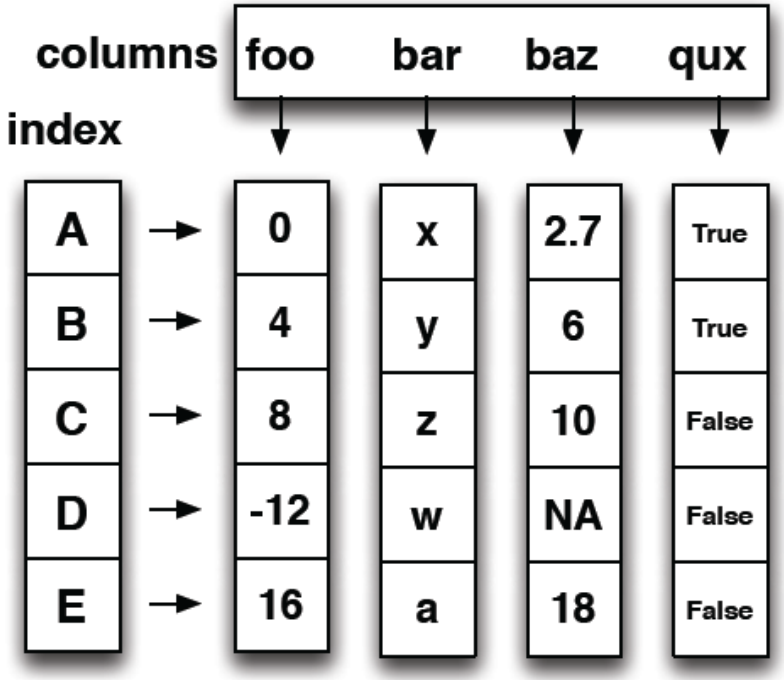
Pandas

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



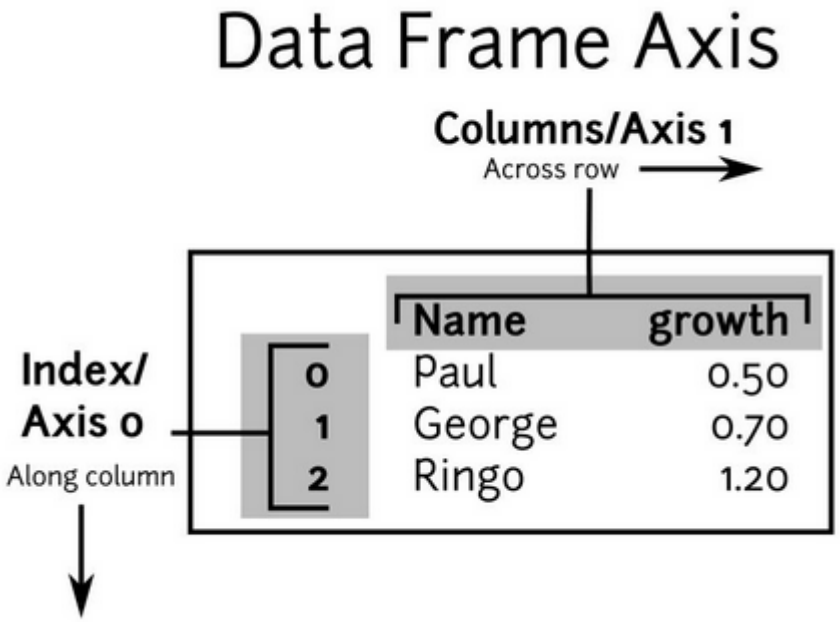
Pandas

DataFrame : 2 dimensions



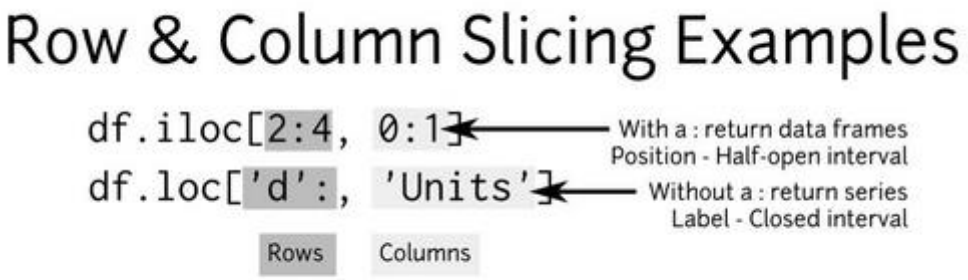
Pandas

DataFrame : axes



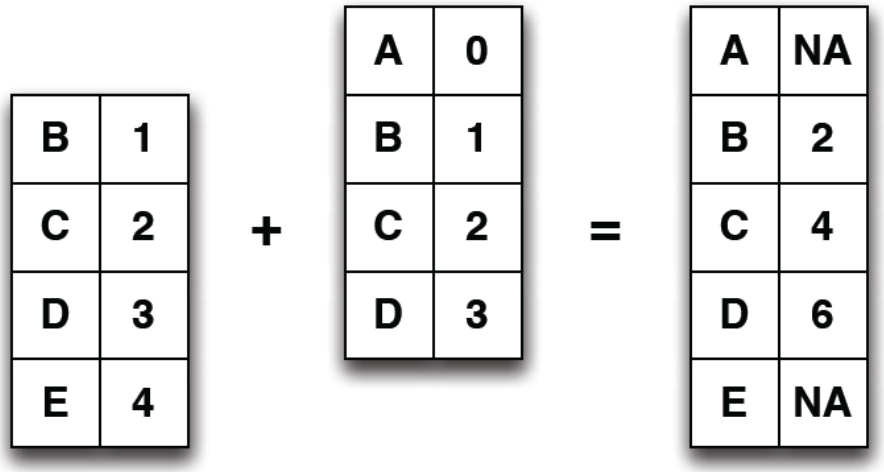
Pandas

DataFrame : slices



Pandas

Opérations facilitées par les index : jointures automatiques



B	1
C	2
D	3
E	4

+

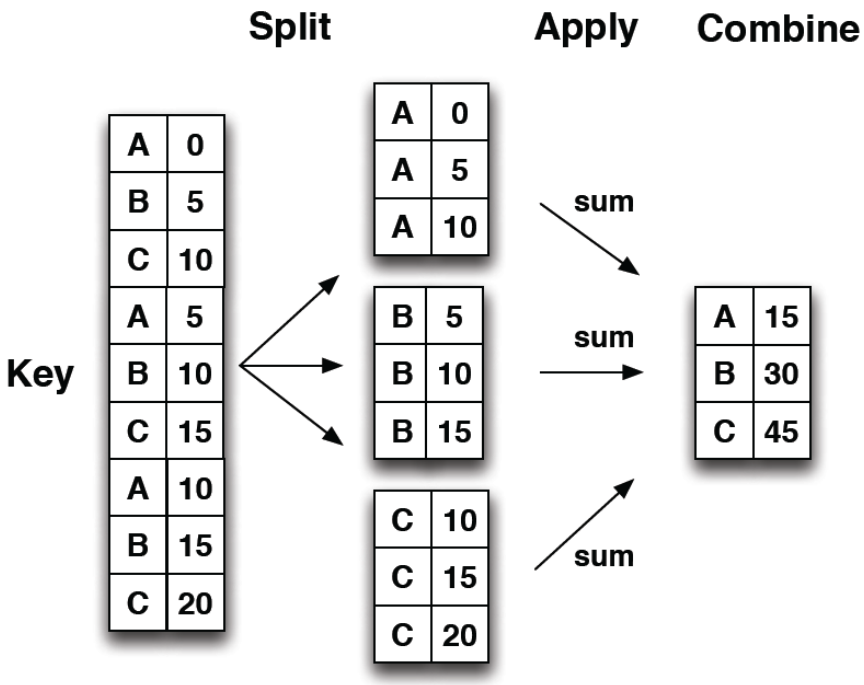
A	0
B	1
C	2
D	3

=

A	NA
B	2
C	4
D	6
E	NA

Pandas

Opérations : GroupBy



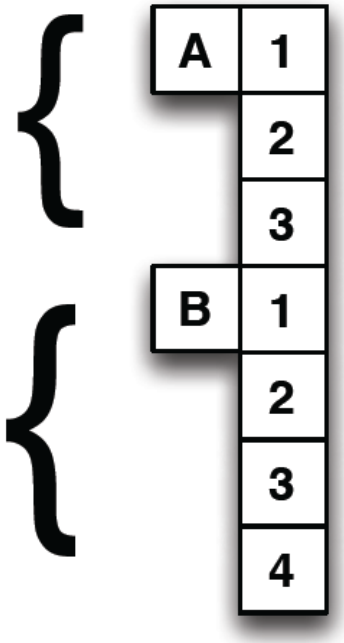
Pandas

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

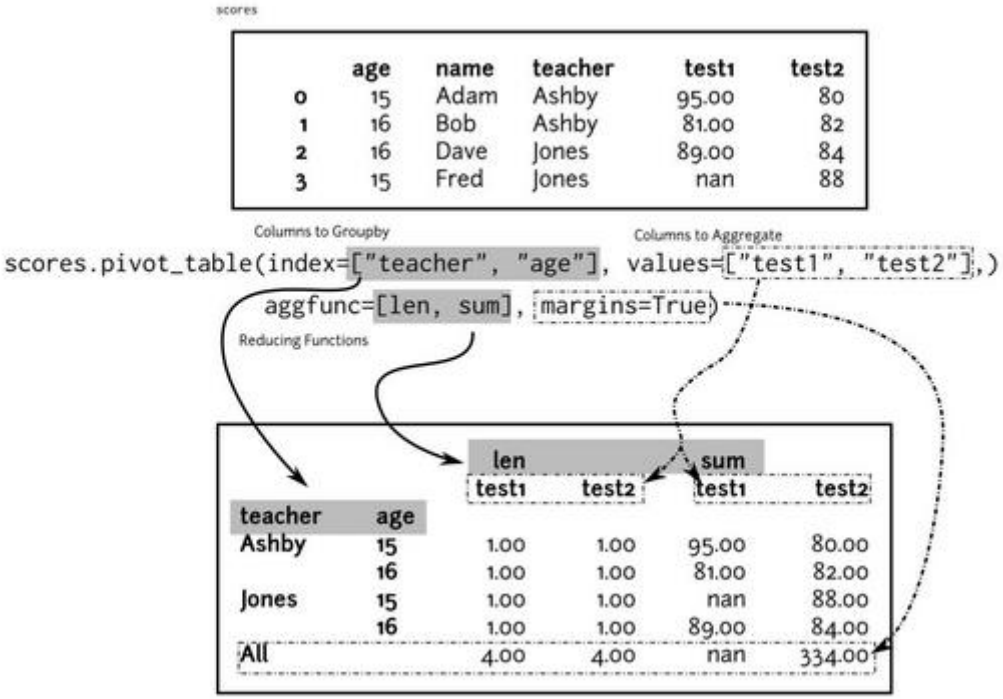
Pandas

Index multidimensionnels



Pandas

Pivoter les tables

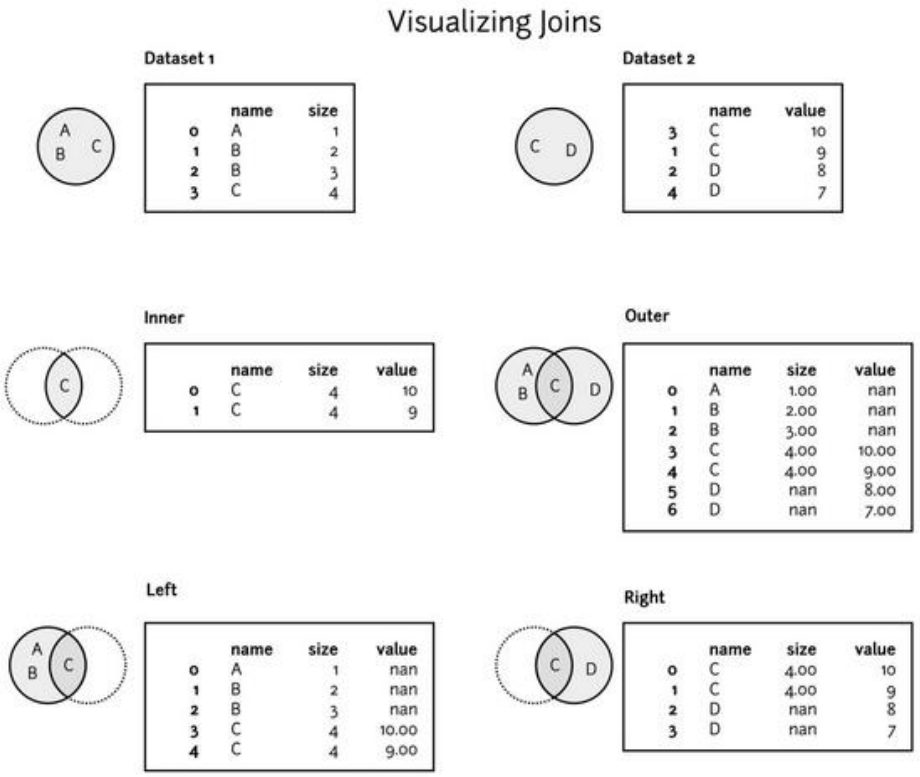


	age	name	teacher	test1	test2
0	15	Adam	Ashby	95.00	80
1	16	Bob	Ashby	81.00	82
2	16	Dave	Jones	89.00	84
3	15	Fred	Jones	nan	88

		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 (merge)



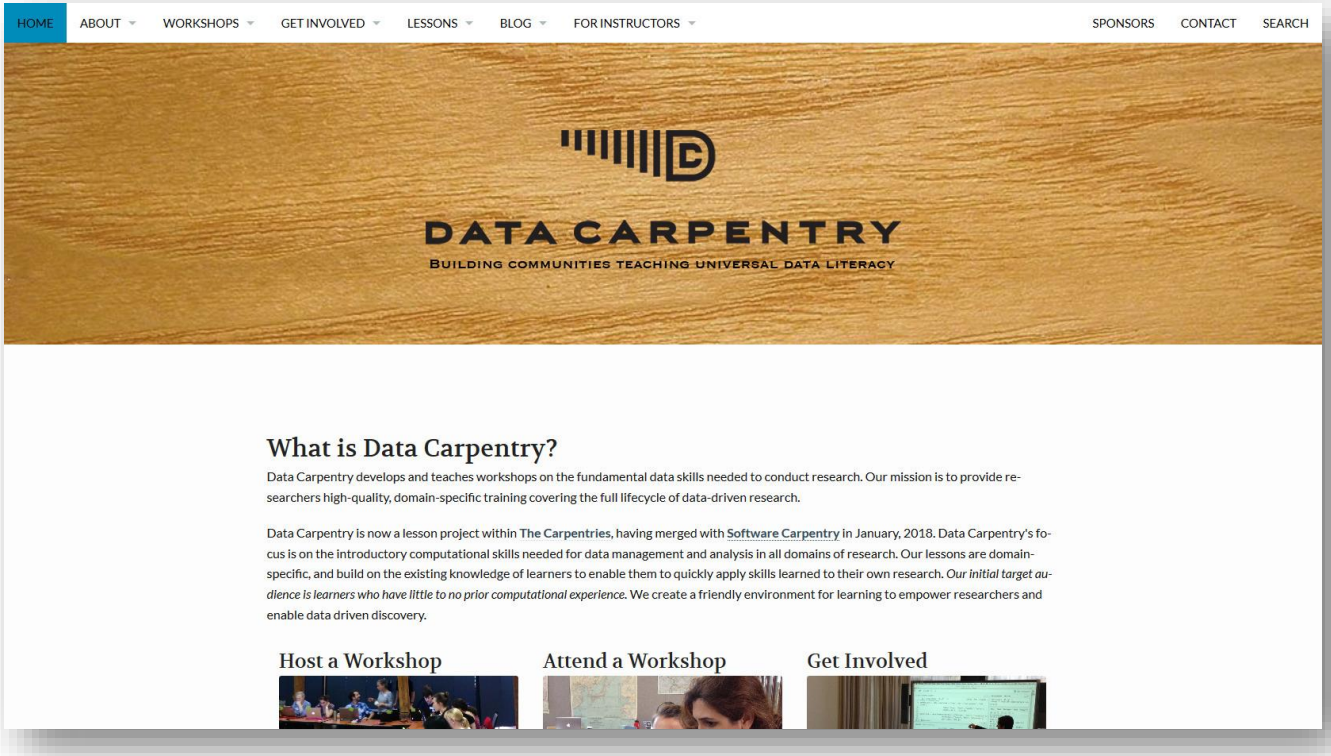
Exercices

Exemples 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

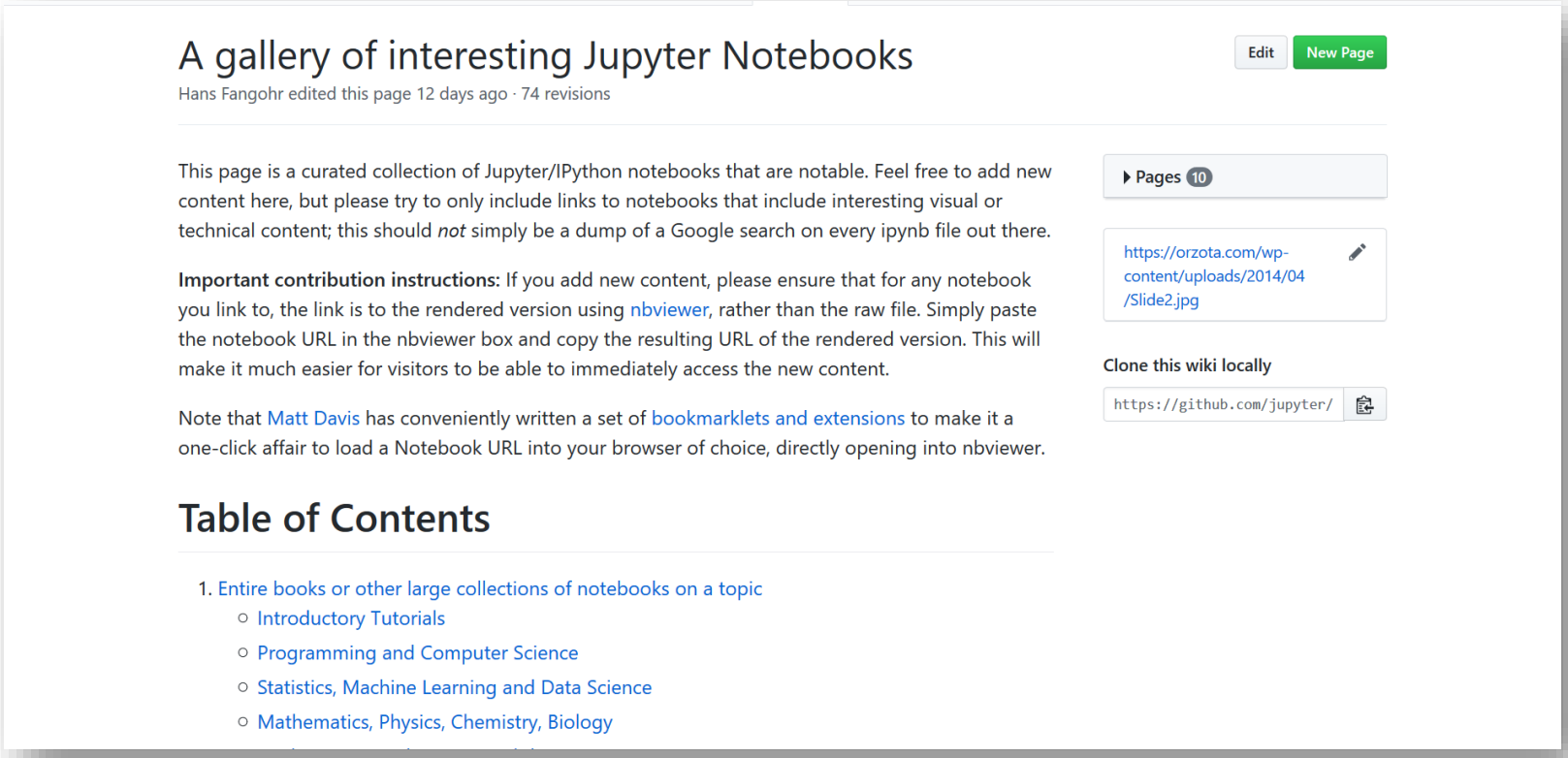
Aide Pandas : https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html

Pour aller plus loin



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

Pour aller plus loin



A gallery of interesting Jupyter Notebooks

Edit | New Page

Hans Fangohr edited this page 12 days ago · 74 revisions

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

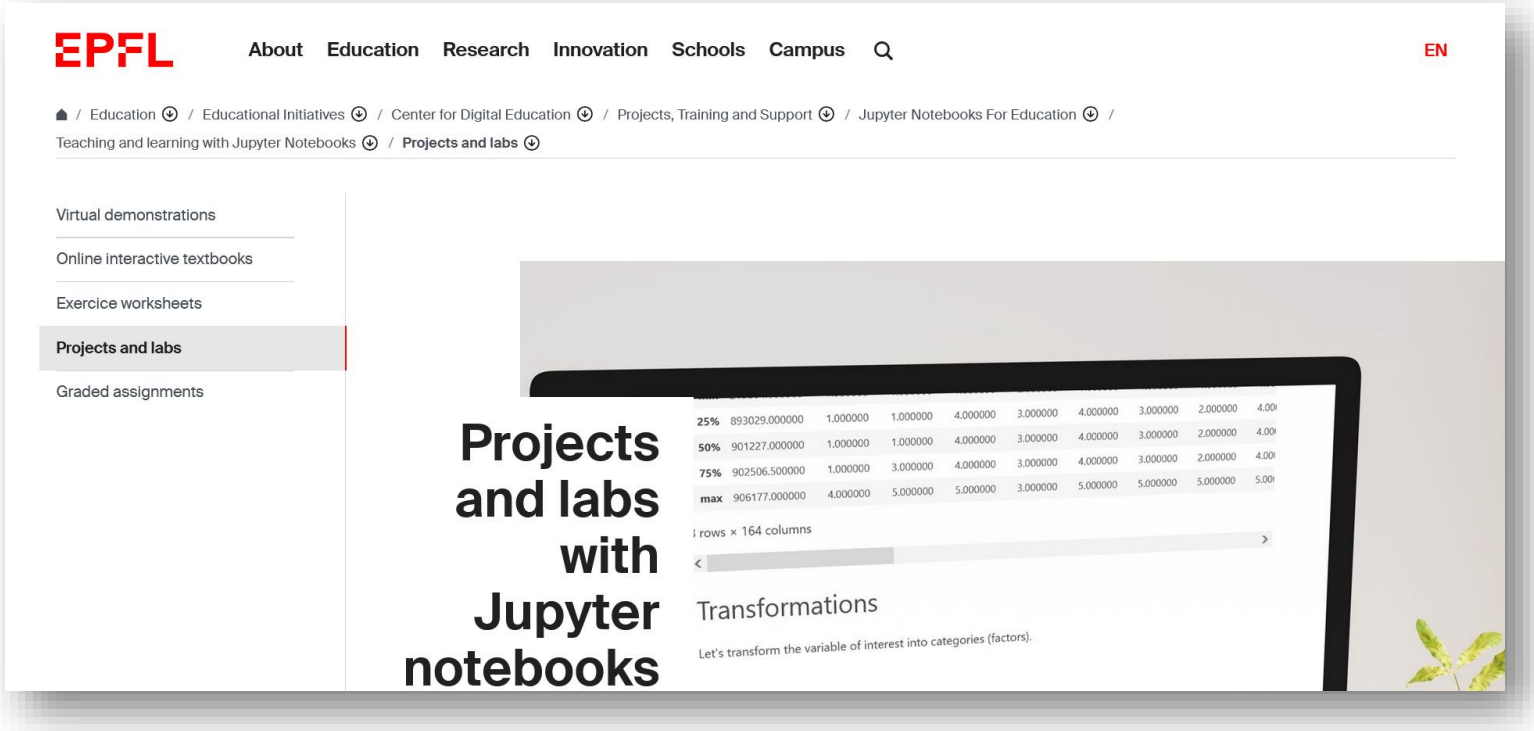
https://github.com/jupyter/

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

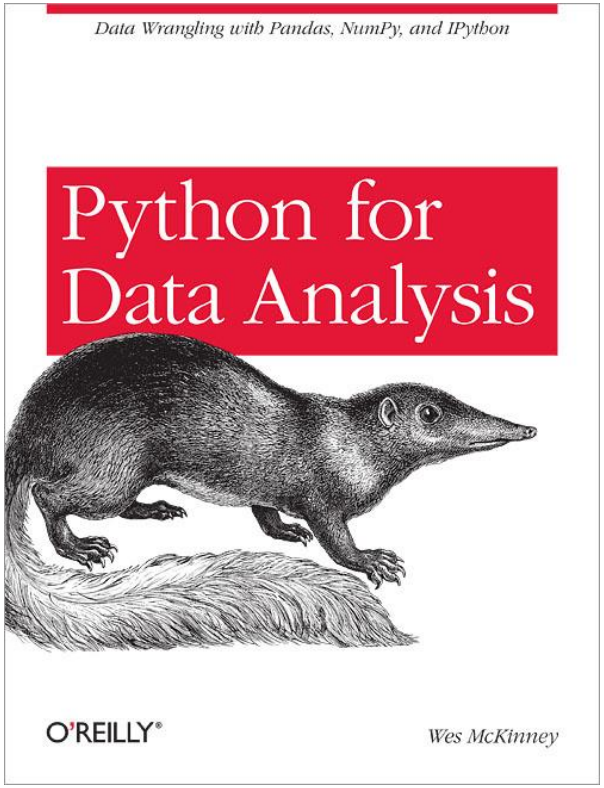
https://github.com/jupyter/jupyter/wiki#a-gallery-of-interesting-jupyter-notebooks

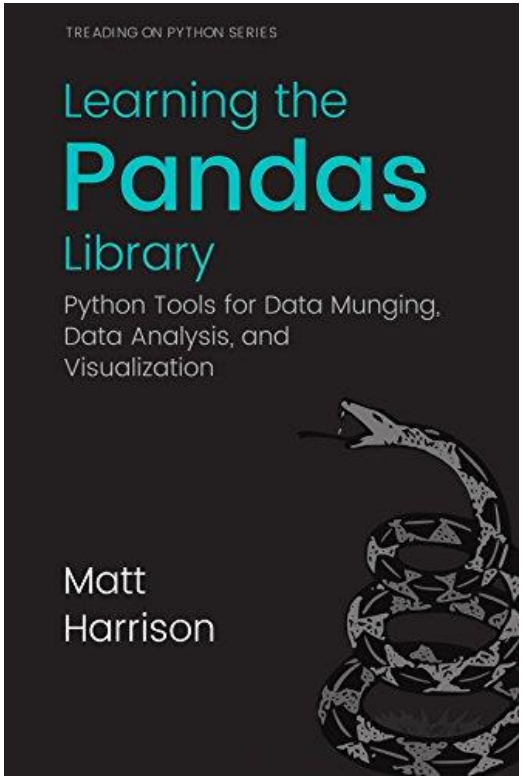
Pour aller plus loin



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





Sources

Cheat Sheets :

- 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