

Quel(s) intérêt(s) à utiliser Python pour la 'science ouverte' ?

URFIST Occitanie

Émilien Schultz

9 novembre 2021

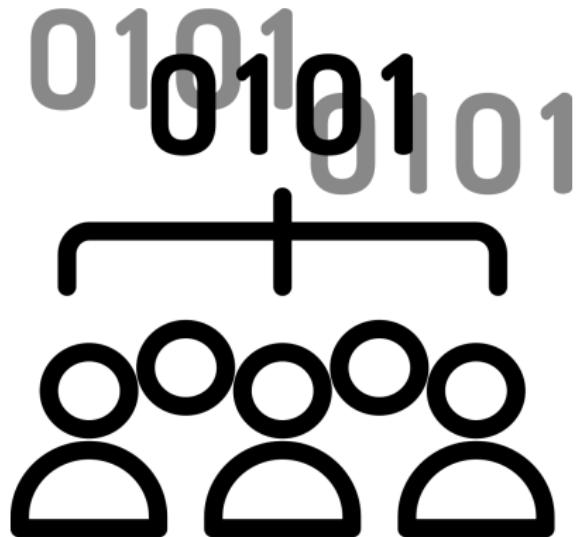
Qu'est-ce qui vous amène ici ?

Objectifs de la séance

- ▶ Nos objectifs généraux :
 - ▶ Voir avec vous les bases du langage
 - ▶ Développer votre autonomie pour avancer
 - ▶ Échanger sur les usages possibles
- ▶ Nos objectifs pour aujourd'hui :
 - ▶ L'environnement pour programmer
 - ▶ Débuter avec les bases du langage
 - ▶ Installer des outils au-delà du langage
 - ▶ Voir au passage certaines notions comme l'API

Avant de se lancer : essayer de répondre à 3 questions

1. Pourquoi programmer ?
2. Pourquoi Python ?
3. Lien avec la science ouverte ?



Pourquoi programmer ?

La numérisation de la recherche

- ▶ Traitement numérique comme point de passage obligé des activités de recherche - *digital turn*
- ▶ Explosion de la disponibilité des données et usages secondaires
- ▶ Courant profond et puissant de la science ouverte

Programmer ou quoi ?

Programmer[Définition pratique] : utiliser un ensemble de commandes (code) pour faire réaliser (exécuter) à l'ordinateur des tâches

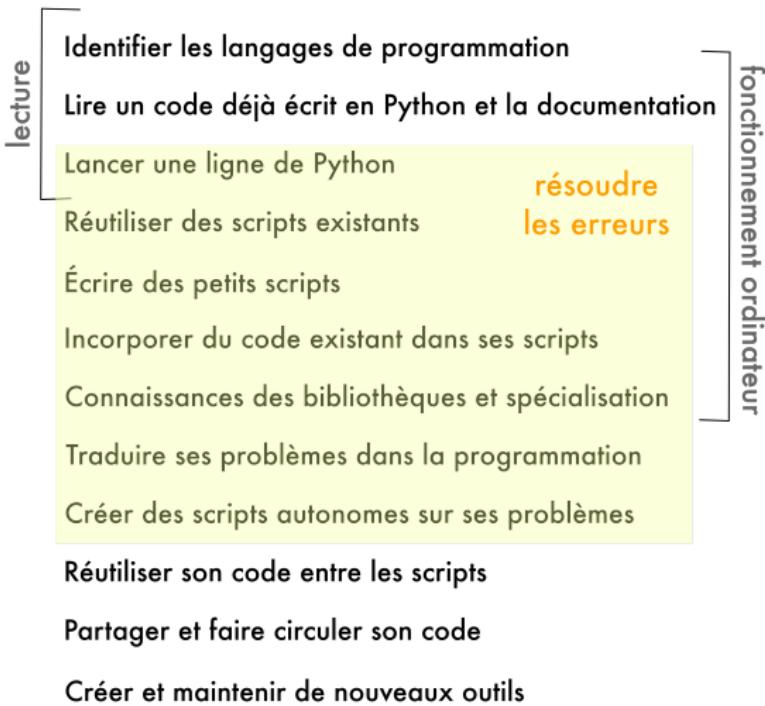


Cinquante nuance de programmation

- ▶ Programmer != Développer des logiciels
- ▶ Un usage spécifique : **la programmation scientifique**
 - ▶ Orientation **script** : réaliser des petites tâches spécifiques
 - ▶ Orientation **interactive** : tester et expérimenter
 - ▶ Orientation **recherche** : des outils spécifiques
- ▶ Pas incompatible avec des logiciels
- ▶ Un effet **oignon** : pour programmer, il faut se familiariser avec la superposition des structures numériques
 - ▶ Format de fichier : csv ou xls ?
 - ▶ Stockage : mémoire vive ou disque dur ?
 - ▶ ...

Des usages à différents niveaux

Découvre la programmation



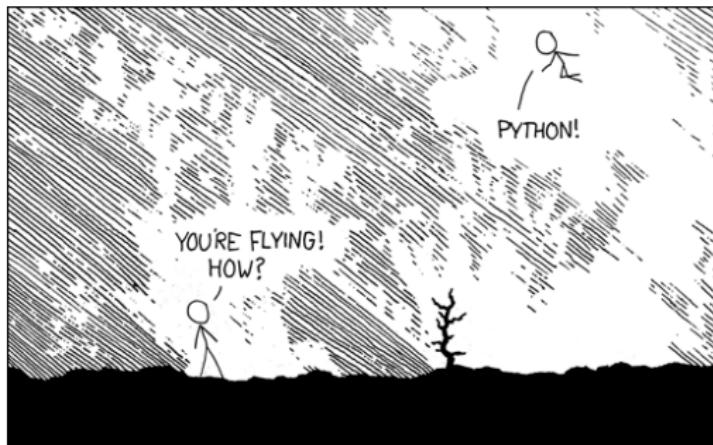
Contributeur•rice Open Source accompli•e

Les obstacles

- ▶ Un outil parmi d'autres : **pas une baguette magique**
- ▶ Courbe d'apprentissage potentiellement longue (mais...)
- ▶ Avoir une idée de quoi en faire : quel imaginaire pratique ?
- ▶ Trouver des ressources locales : importance de la pratique
- ▶ Les bases de programmation permettent d'automatiser des tâches, pas de remplacer les savoirs spécifiques nécessaires à leur mise en oeuvre.

Pourquoi Python ?

Tout est possible avec Python (sur un ordinateur)



I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
/ HELLO WORLD IS JUST
print "Hello, world!"

I DUNNO...
DYNAMIC TYPING?
WHITESPACE?
/ COME JOIN US!
PROGRAMMING IS FUN AGAIN!
IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?

I JUST TYPED
import antigravity
/ THAT'S IT?
/ ... I ALSO SAMPLED
EVERYTHING IN THE
MEDICINE CABINET
FOR COMPARISON.
/ BUT I THINK THIS
IS THE PYTHON.

Propriétés de Python

- ▶ Libre et interopérable
- ▶ Pédagogique
- ▶ En croissance d'usage
- ▶ Enseigné dès le lycée
- ▶ Favorise les bonnes pratiques de programmation

```
(p37) iMac-de-Emilien:~ emilien$ ipython
Python 3.7.7 (default, Mar 26 2020, 10:32:53)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.13.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]: print("La somme est : ",sum([10,12,8]))
La somme est : 30

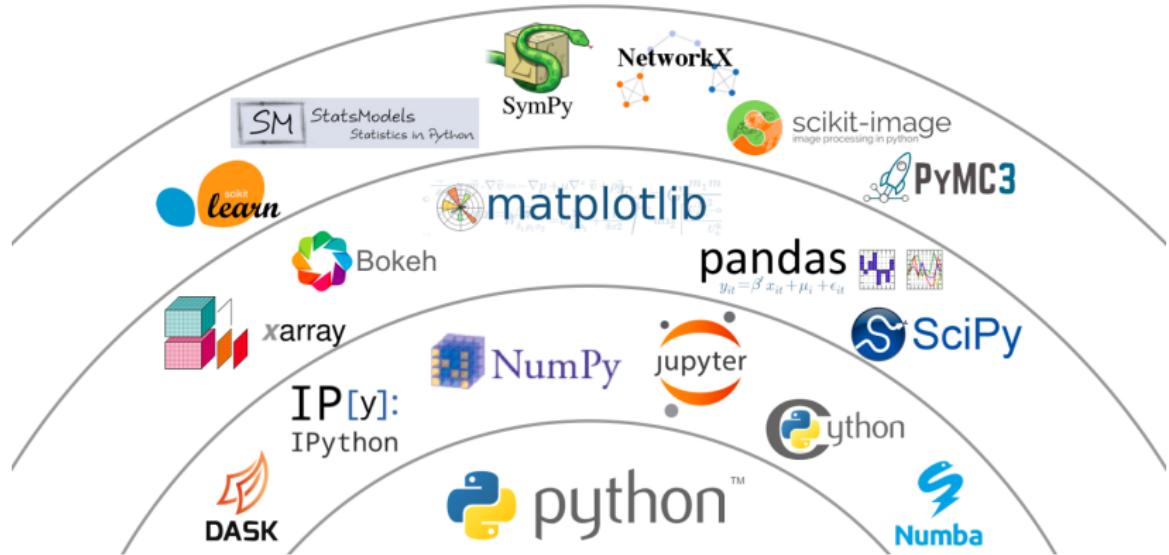
In [2]: 
```

Facile à utiliser comme langage de script

Un univers complet

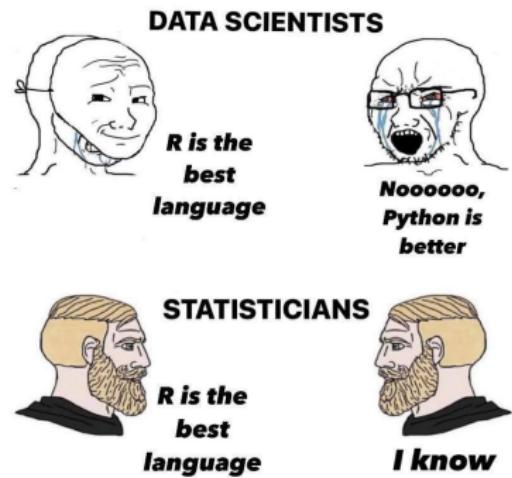
Python's Scientific Stack

Jake Vanderplas PyCon 2017 Keynote



Et Anaconda pour l'installation, ou Google Colab pour le cloud ...

Mais pas le seul choix



Qui mène à la question centrale : dois-je choisir Python ?

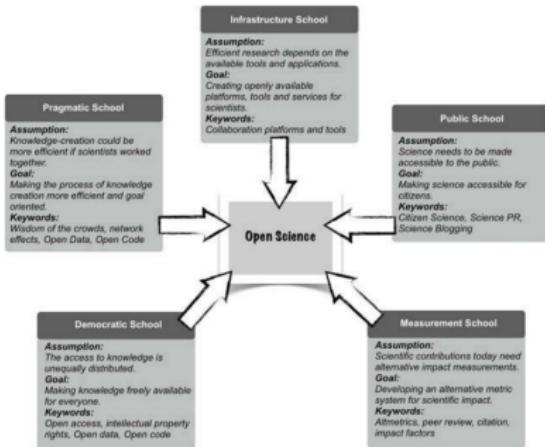
Python & Science ouverte ?

Différentes dimensions de la science ouverte

- ▶ Open source et open science
- ▶ Reproductibilité des résultats
- ▶ Données ouvertes
- ▶ Ouverture des publications
- ▶ ...

Open Science: One Term, Five Schools of Thought

19



Lien Python et Open Science



Install About Us Community Events Documentation NBViewer JupyterHub Widgets Blog Security

About Us

Some information about the Jupyter Project and Community

Project Jupyter is a non-profit, open-source project, born out of the Python Project in 2014 as it evolved to support interactive data science and scientific computing across all programming languages. Jupyter will always be 100% open-source software, free for all to use and released under the liberal terms of the modified BSD license.

Jupyter is developed in the open on GitHub, through the consensus of the Jupyter community. For more information on our governance approach, please see our [Governance Document](#).

All online and in-person interactions and communications directly related to the project are covered by the Jupyter Code of Conduct. This Code of Conduct sets expectations to enable a diverse community of users and contributors to participate in the project with respect and safety.



COMMENT

<https://doi.org/10.1038/s42005-020-00403-4>

OPEN

Creating an executable paper is a journey through Open Science

Jana Lasser^{1,2✉}

Executable papers take transparency and openness in research communication one step further. In this comment, an early career researcher reports her experience of creating an executable paper as a journey through Open Science.

Open Science practices are taking an increasingly central role in the way we conduct research, from accessible research data to transparency in the methodologies used to analyze them. To this end, the novel "executable paper" format offers a transparent and reproducible way to communicate research. Executable papers are dynamic pieces of software that combine text, raw data, and the code used for the analysis, and that a reader can interact with. In this commentary, I introduce the executable paper format and highlight its advantages for research communication. Drawing from my personal experience, I offer practical advice on how to create an executable paper by using Open Source tools such as Python and Jupyter Notebooks, and how to make it accessible by publishing it on open repositories.

La programmation pour le traitement de données : un usage intersticiel

- ▶ Intérêt du script pour interagir avec les données
- ▶ Des usages "discrets" plus que "computationnels"
- ▶ Ne se limite pas aux statistiques
- ▶ Cependant :
 - ▶ Peu d'exemples
 - ▶ Peu d'outils clairement identifiés

En pratique, ça sert à quoi ?

Cas : format de données

Passer d'un fichier *.html* à un *.txt* mis en forme pour IRaMuteq

Les Echos, mardi 23 mars
évenement, vendredi 20 mars 2020 813 mots, p. 3

Coronavirus

Aussi paru dans 19 mars 2020 | [lesechos.fr](#)

Les cliniques privées à la rescoussse
SOLVIEG GODELUCK

En Alsace, où les hôpitaux publics sont débordés, les éti

Certaines sont donc la tempête, d'autres l'attendent. Ainsi
Faut de patients atteints du Covid-19. « Nous avons c
directeur général de la Fondation Saint-Vincent à Stras

Des lits transformés pour la réanimation

Ces disponibilités ont pourtant été signa
pouvoir entrer dans le dispositif », plaide Christophe M

« Nous ne sommes pas sollicités à hauteur du service q
Samu : on oriente les malades vers le secteur public. Li
tous les deux jours, on a déprogrammé toutes nos opér

100.000 soins déprogrammés dans le privé lucratif



renforcement » dans d'autres. Le lendemain, le ministre de l
lui-même été infecté, a annoncé l'extension des tests de dép
se lancer dans le ~~déconfinement~~, Sophie Amsili et Tifenn Clir

**** *num_618 *journal_Lefigaro

« Pendant trois heures, Emmanuel Macron a pris connaissance à
résultats obtenus par l'équipe du Pr Raoult », se réjouit la
<acteur>Martine Wonner</acteur>, seule parlementaire LREM à
<url>19-laissons les médecins prescrire</url>. » LIRE AUSSI -
Raoult : les dessous d'une rencontre surprise Cette psych
maladie. Elle s'était aussi engagée avec les écologistes, c
rénouvellement nust de Strasbourg... dont l'Anonyme chantier a



Cas : data science et exploration de données

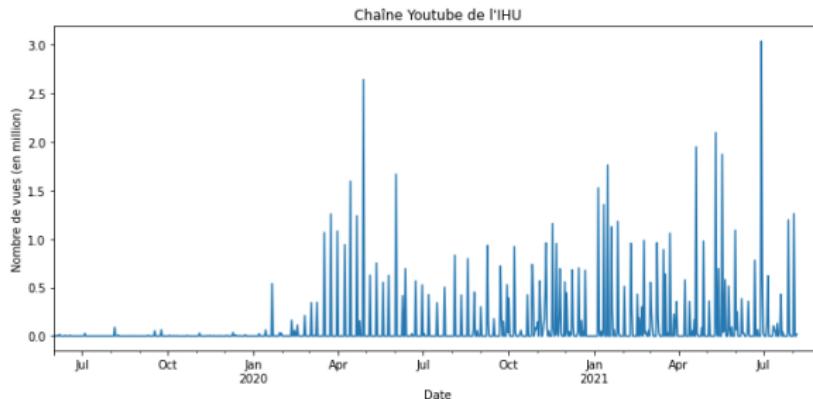
Exploration d'un tableau de données (ici le nombre de vues par vidéos de la chaîne Youtube de l'IHU)

2018-02-12	Les jeudis de l'IHU: 08 février 2018 - 2 - S...	2018-02-12T09:45:09Z	593	0.000593
2018-02-12	Les jeudis de l'IHU: 08 février 2018 - 3 - D...	2018-02-12T09:46:07Z	300	0.000300
2018-02-12	Les jeudis de l'IHU: 08 février 2018 - 4 - D...	2018-02-12T11:24:23Z	629	0.000629
2018-02-12	Les jeudis de l'IHU: 08 février 2018 - 5 - Dr...	2018-02-12T11:24:48Z	276	0.000276
2018-02-12	Les jeudis de l'IHU: 08 février 2018 - 6 - Pr...	2018-02-12T11:25:08Z	553	0.000553

721 rows x 4 columns

```
Entrée [160]: ax = d["vues"].resample("d").sum().plot(figsize=(10,5),style="--")

plt.xlim("2019-06-01", "2021-09-01")
plt.ylabel("Nombre de vues (en million)")
plt.xlabel("Date")
plt.title("Chaine Youtube de l'IHU")
plt.tight_layout()
plt.savefig("ihu_youtube.png",dpi=200)
```



Cas : construire un réseau

Créer la bonne structure relationnelle (ici auteur/auteur) et l'exporter dans un format compatible avec Gephi

AUTHOR	YEAR	ANNEE	AUTHORS	TITLE	JOURNAL
35	1998	LEROLIA A.	LEROLIA A., BRETAGNOVILLE N.	Sea ratio versus Journal of Animal Ecology	
37	1998	LEROLIA A.	RECODER		
44	1998	LEROLIA A.	LEROLIA A.B.A.	Egg and nest record scheme	
47	1998	DE CORNAILLET T.	BERNARD J.	Reproduction of the European Robin	
52	1999	ARROYO E.	BRETAGNOVILLE C.	Breeding bird	Journal of RSPB
55	1999	ALMAGRO M.	MORCILLO I.	Patagonian bird study	
59	2000	ARROYO E.	DECONINCK T.	Reproductive output and age	Condor
60	2000	ARROYO E.	ROUTE B.	PR Activities and Review of Ecological	Ecology
62	2000	ARROYO E.	ROUTE B.	PR Activities and Review of Ecological	Ecology
63	2000	ARROYO E.	ROUTE B.	PR Activities and Review of Ecological	Ecology
66	2000	SALAMANDR M.	BUTET A.	LE Responses or Ecological	Ecology
67	2000	ARROYO E.	ROUTE B.	PR Activities and Review of Ecological	Ecology
70	2001	ROUTE B.	MOUGET F.	BIRD Colonial Breeding Behaviour	Ecology
71	2001	ROUTE B.	MOUGET F.	BIRD Colonial Breeding Behaviour	Ecology
71	2001	ROUTE B.	MOUGET F.	BIRD Colonial Breeding Behaviour	Ecology



Cas : construction de tableaux adaptés

Produire des sorties de tableaux adaptés à l'objet (et possibilité ensuite d'aller sur Excel ou Latex)

```
Entrée [64]: var_ind = {"sexe":"1 - Sex","age2":"2 - Age","diplome":"3 - Education", "revenus":"4 - Incomes",  
"PROXPARTI":"5 - Political orientation"}  
  
t = {"COCONEL1":pyshs.tableau_croise_multiple(data1,"HC_c",var_ind,chi2=False)[["1 - HC effective",  
"COCONEL2":pyshs.tableau_croise_multiple(data2,"HC_c",var_ind,chi2=False)[["1 - HC effective",  
"COCONEL3":pyshs.tableau_croise_multiple(data3,"HC_c",var_ind,chi2=False)[["1 - HC effective",  
"TRACTRUST1":pyshs.tableau_croise_multiple(data4,"HC_c",var_ind,chi2=False)[["1 - HC effective",  
"TRACTRUST2":pyshs.tableau_croise_multiple(data5,"HC_c",var_ind,chi2=False)[["1 - HC effective"  
  
t = pd.concat(t,axis=1)  
t.applymap(lambda x : re.findall("\((.*?)%\)",x)[0])
```

Out[64]:

Variable	Modalités	COCONEL1		COCONEL2		COCONEL3		TRACTRUST1	
		1 - HC effective	2 - HC not effective	1 - HC effective	2 - HC not effective	1 - HC effective	2 - HC not effective	1 - HC effective	2 - HC not effective
1 - Sex	Femme	38.3	3.9	34.0	9.1	17.8	9.0	14.2	13.4
	Homme	36.8	7.4	27.2	13.6	21.6	14.7	19.5	19.0
	Total	37.6	5.6	30.8	11.3	19.6	11.7	16.7	16.1
	17-34	36.7	8.9	27.8	15.4	16.8	14.7	14.6	20.4
2 - Age	35-54	41.1	4.5	31.3	10.1	19.9	11.8	18.4	14.2
	55-79	36.8	4.0	33.3	10.2	23.3	8.9	17.7	16.7
	70-100	33.3	4.5	31.0	8.4	19.1	9.6	14.9	11.8
	Total	37.6	5.6	30.8	11.3	19.6	11.7	16.7	16.1
3 - Education	1 - inf bac	33.2	5.3	34.8	8.4	21.3	8.0	18.7	8.3
	2 - bac	42.3	4.7	33.5	9.3	21.4	9.9	17.5	14.0

Cas : collecte automatique de données

Twitter et l'API universitaire

```
Entrée [1]: import json
import pandas as pd
from searchtweets import ResultStream, gen_rule_payload, load_credentials,collect_results
```

Authentification

```
Entrée [2]: creds = load_credentials(filename='./credentials.yaml',
                                     yaml_key='search_tweets_api',
                                     env_overwrite=False)
```

Grabbing bearer token from OAUTH

Requête

```
Entrée [3]: rule = gen_rule_payload("ANR lang:fr", results_per_call=50,
                                    from_date="201101210000",
                                    to_date="201102210000")
print(rule)
tweets = collect_results(rule,
                         max_results=1000,
                         result_stream_args=creds)

{"query": "ANR lang:fr", "maxResults": 50, "toDate": "201102210000", "fromDate": "201101210000"}
```

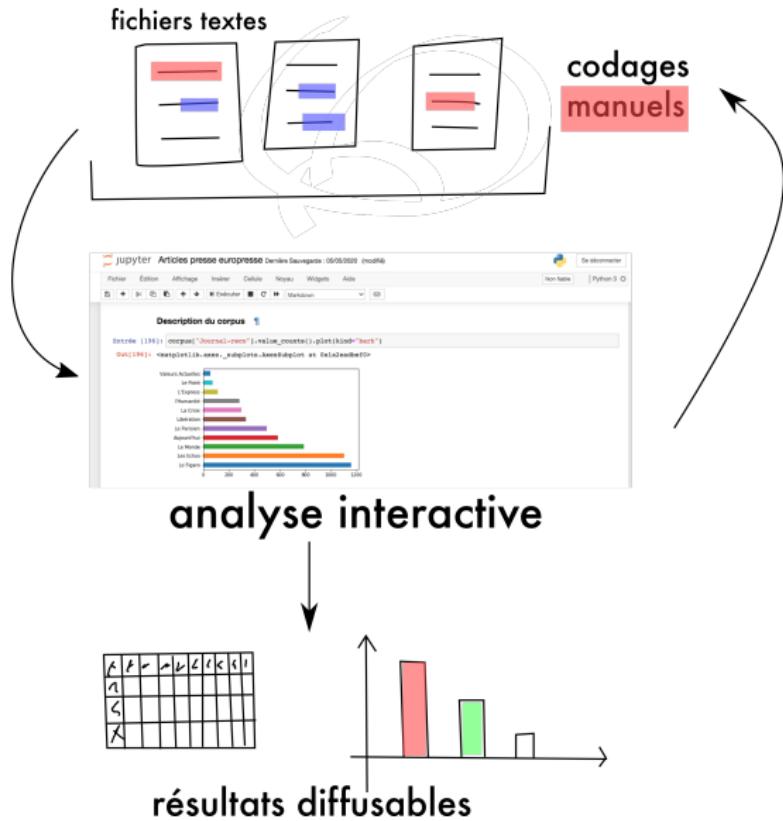
```
Entrée [4]: print(len(tweets))
pd.DataFrame([(i.created_at_datetime,i.all_text) for i in tweets])
```

136

Out[4]:

	0	1
0	2011-02-20 18:21:50	"ANR Estée Lauder Advanced Night Repair sérum ...
1	2011-02-20 10:53:33	Recherches Partenariales et Innovation Biomédi...
2	2011-02-19 11:38:04	L'ANR propose une boîte à idées pour préparer ...
3	2011-02-18 10:28:41	A lire RT @CollectifPAPER La Cour des Comptes...
4	2011-02-18 10:26:09	La Cours des Comptes rappelle à l'ordre l'ANR ...
...
131	2011-01-25 07:52:30	Chaires d'excellence de l'ANR: accueil des che...

Cas : codage de matériel qualitatif



Cas : figures d'un article faciles à reproduire

Production des statistiques et des figures facile à relancer en cas de révision de l'article.

Open Access Article

French Public Familiarity and Attitudes toward Clinical Research during the COVID-19 Pandemic

by  Émilien Schultz 1,2,*   Jeremy K. Ward 3,4   Laëtitia Atlani-Duault 1,5,6   Seth M. Holmes 2,7,8   and  Julien Mancini 2,9  

* Author to whom correspondence should be addressed.

¹ CEPED (UMR 196), Université de Paris, IRD, 75006 Paris, France

² SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, CANBIOS Team (Équipe Labelisée LIGUE 2019), Aix-Marseille University, INSERM, IRD, 13009 Marseille, France

³ CERME3, INSERM, CNRS, EHESS, Université de Paris, 94801 Villejuif, France

⁴ VITROME, Aix-Marseille University, IRD, AP-HM, SSA, 13005 Marseille, France

⁵ Institut COVID-19 Add Memoriam, University of Paris, 75006 Paris, France

⁶ WHO Collaborative Center for Research on Health and Humanitarian Policies and Practices, IRD, Université de Paris, 75006 Paris, France

⁷ Society and Environment, Medical Anthropology, and Public Health, University of Berkeley, Berkeley, CA 94720, USA

⁸ Mediterranean Institute for Advanced Study IMéRA, Institut Paoli Calmettes, Aix-Marseille University, 13004 Marseille, France

⁹ BioSTIC, APHM, Timone, 13005 Marseille, France

Academic Editor: Roy McConkey

Int. J. Environ. Res. Public Health **2021**, *18*(5), 2611; <https://doi.org/10.3390/ijerph18052611>

Received: 2 February 2021 / **Revised:** 2 March 2021 / **Accepted:** 2 March 2021 / **Published:** 5 March 2021

(This article belongs to the Section [Global Health](#))

[View Full-Text](#) [Download PDF](#) [Browse Figures](#) [Citation Export](#)

Abstract

The COVID-19 pandemic put clinical research in the media spotlight globally. This article proposes a first measure of familiarity with and attitude toward clinical research in France. Drawing from the “Health Literacy Survey 2019” (HLS19) conducted online between 27 May and 5 June 2020 on a sample of the French adult population (N = 1003), we show that a significant proportion of the French population claimed some familiarity with clinical trials (64.8%) and had positive attitudes (72%) toward them. One of the important findings of this study is that positive attitudes toward clinical research exist side by side with a strong distancing from the pharmaceutical industry. While respondents acknowledged that the pharmaceutical industry plays an important role in clinical

Cas : diffuser ses outils à la communauté

The screenshot shows a project page for "pyshs 0.1.12". At the top, there's a search bar with "Search projects" and a magnifying glass icon. To the right are links for "Help", "Sponsors", "Log in", and "Register". Below the header, the project name "pyshs 0.1.12" is displayed in large blue text. Underneath it, there's a button with the command "pip install pyshs" and a small icon. To the right of the project name, there's a green button with a checkmark and the text "Latest version". Below the main title, the text "Module PySHS - Faciliter le traitement statistique en SHS" is visible. On the far right, the release date "Released: Aug 8, 2021" is shown.

Navigation

☰ Project description

⌚ Release history

💾 Download files

Project links

🏡 Homepage

Statistics

View statistics for this project via
[Libraries.io](#), or by using our public
dataset on [Google BigQuery](#)

Project description

Bibliothèque PySHS

La bibliothèque PySHS a pour but de réunir des outils utiles à un public de praticiens des sciences humaines et sociales francophones pour traiter des données. Elle a pour but de s'enrichir progressivement pour permettre à Python de devenir une alternative (réaliste) à R avec des fonctions facilement utilisable sur les opérations habituelles.

La version actuelle est la 0.1.8

Contenu

Traiter des données d'enquête par questionnaire

- Description d'un tableau de données
- Tri à plat et tableau croisé avec pondération
- Tableau croisant une variable dépendante avec une série de variables indépendantes, avec pondération
- Wrapper pour la régression logistique binomiale pondérée

Autres usages

- ▶ Traitement massif de données : parallélisation, déploiement sur des grandes infrastructures, recours aux outils du machine learning
- ▶ Collaboration autour des données
- ▶ Formalisation des étapes de traitement
- ▶ Traitement des images qui arrive...

Avant de se lancer

Ne pas hésiter à chercher...

Un bon code est un code qui fonctionne. Ensuite on l'améliore.

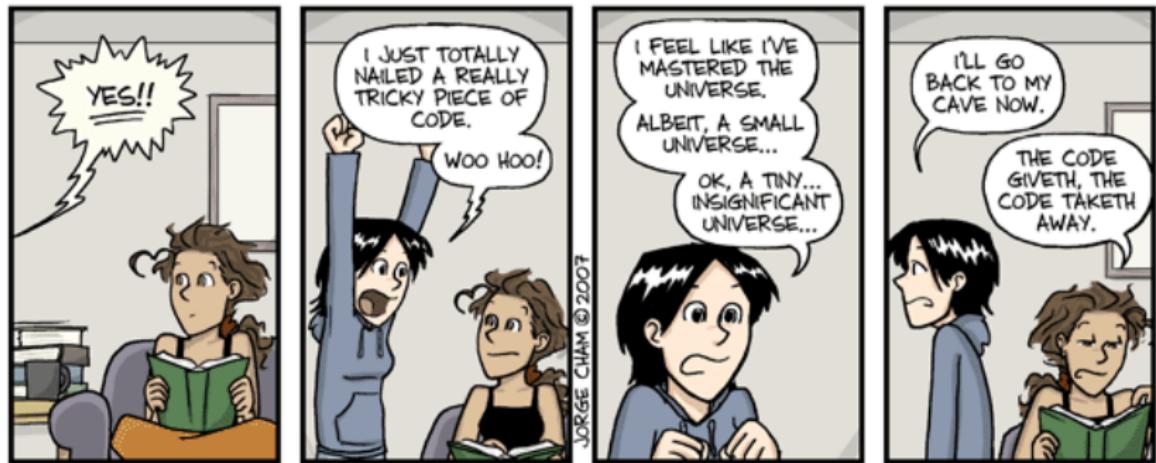


Les obstacles

- ▶ Un outil parmi d'autres : **pas une baguette magique**
- ▶ Courbe d'apprentissage potentiellement longue (mais...)
- ▶ Avoir une idée de quoi en faire : quel imaginaire pratique ?
- ▶ Trouver des ressources locales : importance de la pratique
- ▶ Les bases de programmation permettent d'automatiser des tâches, pas de remplacer les savoirs spécifiques nécessaires à leur mise en oeuvre.

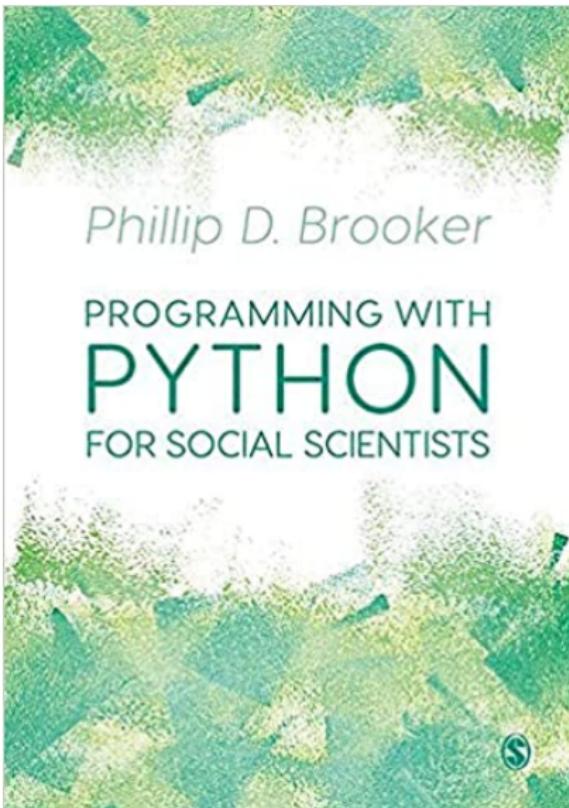
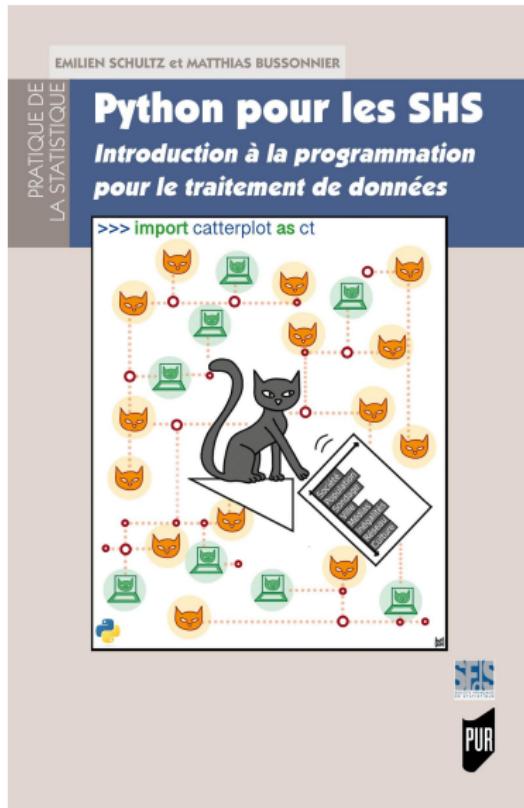


Important de valoriser les petites victoires



WWW.PHDCOMICS.COM

Ressources



<https://github.com/pyshs/ressources-pyshs>

Déroulement de la formation

1. Ce matin : bases du langage
2. Cette après-midi : bibliothèques et API
3. Demain : Pandas et statistiques

Dimension appliquée : prendre le temps de faire les étapes

Tous les documents et les informations à jour sur
<https://github.com/pyshs/>
Formation-URFIST-2021-Toulouse-ScienceOuverte.

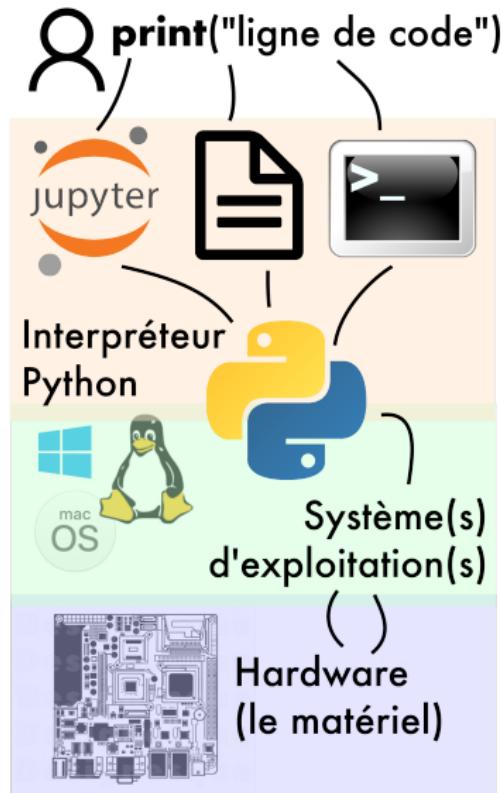
L'environnement nécessaire
pour lancer un script

Première étape : exécuter du code déjà écrit

Où ? Comment ? Quand ?

- ▶ Installer de quoi "faire" du Python
- ▶ Se repérer dans les différentes manières de faire

Où faire du Python



Notre choix : le Notebook Jupyter

- ▶ Des avantages
 - ▶ Ludique et interactif
 - ▶ Avoir tous les éléments au même endroit
 - ▶ Partager son script
- ▶ Quelques limites
 - ▶ Orde d'exécution des cellules
 - ▶ Vite confus

Le plus simple est de voir ensemble

Quelques éléments d'algorithmique

La structure générale

Algorithme

