

Python pour les SHS

Atelier méthode - Centre Émile Durkheim

Émilien Schultz (CEPED/SESSTIM)

<http://eschultz.fr> - emilien.schultz@ird.fr

9 décembre 2021

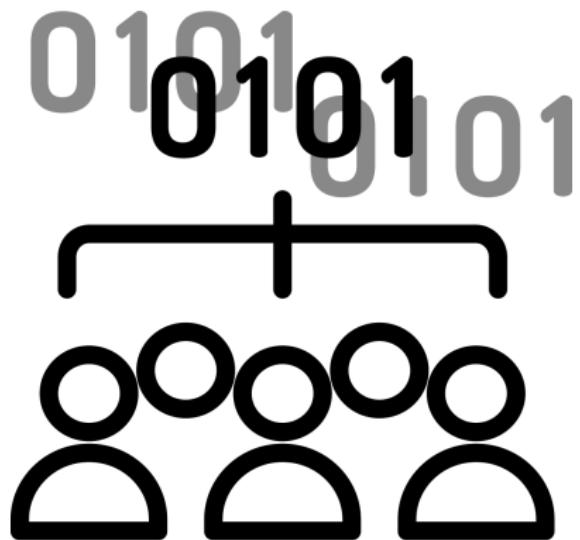
Qu'est-ce qui vous amène ici ?

Objectifs de la séance

- ▶ Nos objectifs aujourd’hui :
 - ▶ Faire un petit point sur la programmation Python en SHS
 - ▶ Voir les bases de la programmation Python
 - ▶ Comprendre l’univers des bibliothèques Python
 - ▶ Balayer des usages plus avancés

Avant de se lancer : essayer de répondre à 3 questions

1. Pourquoi programmer ?
2. Pourquoi Python ?
3. Pourquoi distinguer les SHS ?



(caveat : on va pas épuiser ces questions)

Pourquoi programmer ?

La numérisation de la recherche

- ▶ Traitement numérique comme point de passage obligé du•de la chercheur•se - *digital turn*
- ▶ Explosion de la disponibilité des données et usages secondaires
- ▶ Courant profond et puissant de la science ouverte
- ▶ Apparition de nouveaux objets liés aux pratiques numériques + de nouvelles méthodologies

Programmer ou quoi ?

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



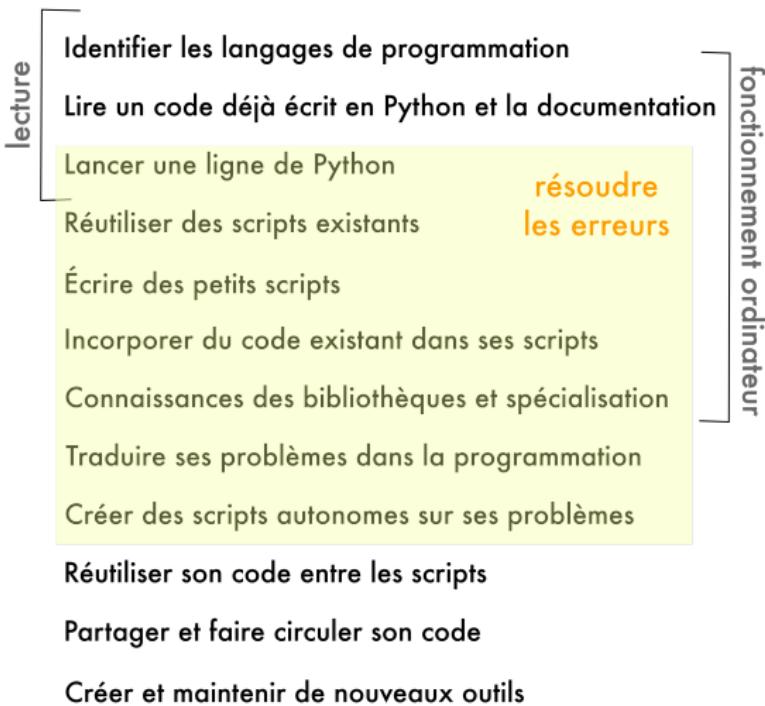
JAKE-CLARK.TUMBLR

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



Script scientifique et *literate programming*

Intégration du code et du texte (Knuth, 1992) puis des résultats dans la *literate computing*.

Casual Notebooks and Rigid Scripts: Understanding Data Science Programming

Krishna Subramanian, Nur Hamdan, Jan Borchers

RWTH Aachen University

52074 Aachen, Germany

{krishna, hamdan, borchers}@cs.rwth-aachen.de

Abstract—Data workers are non-professional data scientists who often use scripting languages like R, Python, or MATLAB, and employ an exploratory programming workflow. Current IDEs offer them two main programming modalities: script files and computational notebooks. To understand how these modalities impact work practice, we conducted a study with 21 data workers, and a subsequent larger survey with 62 respondents. Through interviews, walkthroughs, and screen recordings, we collected information about their workflows. Our analysis shows a tension between scripts and computational notebooks. Scripts are more common, better support storage and execution of previous analyses, but hinder experimentation. Notebooks better suit the actual data science workflow, but can become easily unorganized. We discuss how this dual nature of modality usage leads to several issues that affect data workers' workflows, and discuss implications for the design of programming IDEs.

Index Terms—scripting languages, exploratory programming, programming interfaces, data science, notebooks

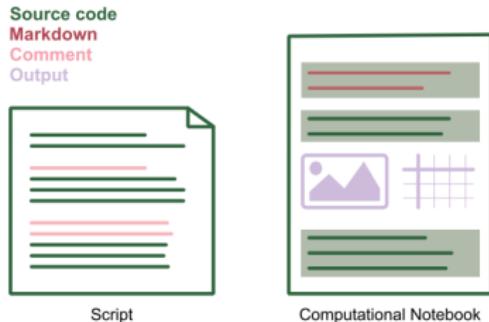
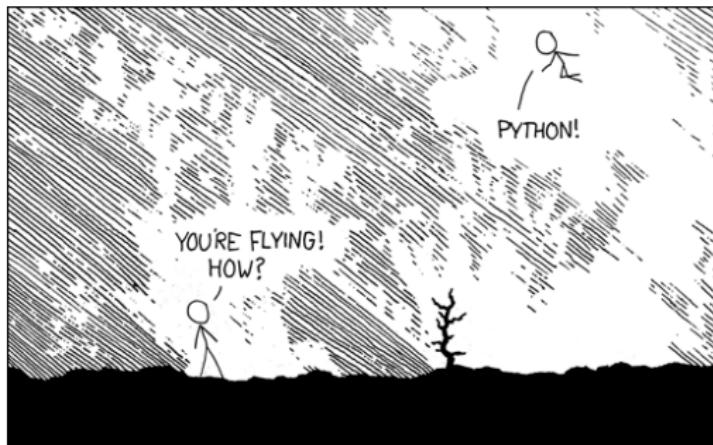


Fig. 1. Current scripting language IDEs support writing and executing code via two programming *modalities*: *scripts* (left) and *computational notebooks* (right). In this paper we investigate how these modalities are used in data

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

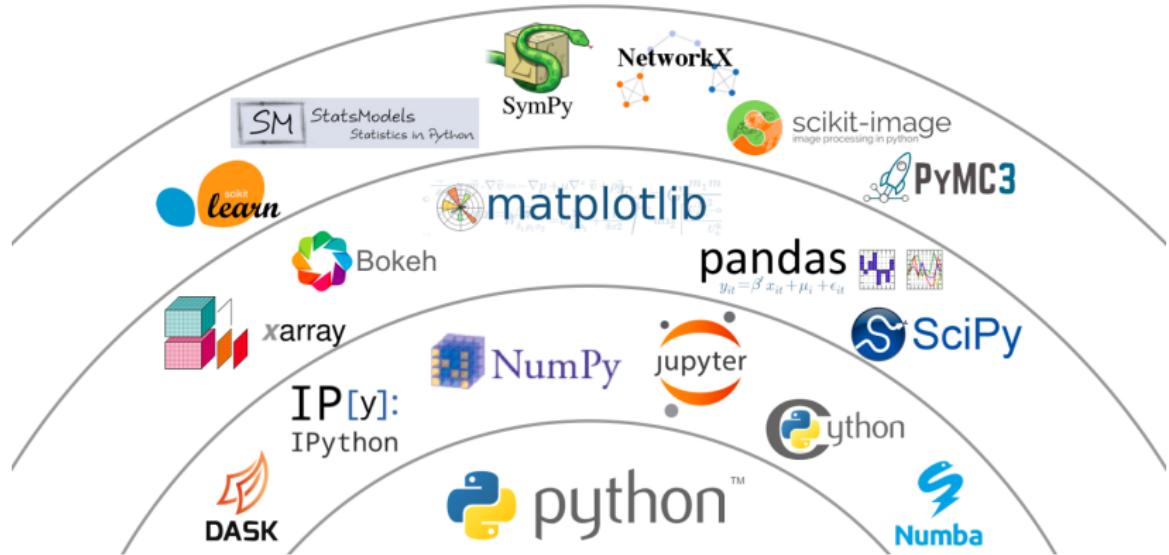
In [2]: 
```

Facile à utiliser comme langage de script

Un univers complet

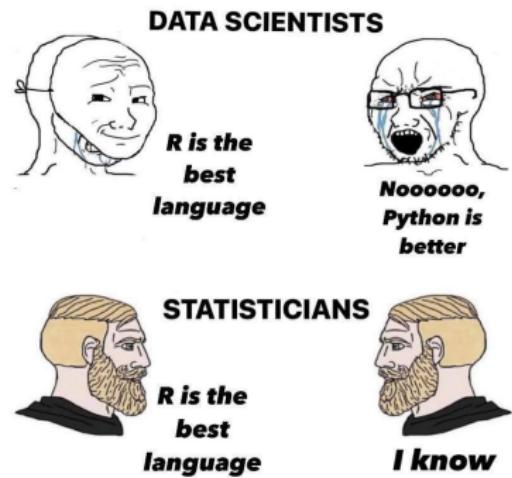
Python's Scientific Stack

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 ?

Pourquoi distinguer les SHS ?

Pluralité des approches

Les SHS loin d'être homogènes :

- ▶ Méthodologies très variées
- ▶ Données plus ou moins accessibles et normalisées
- ▶ Culture du numérique différenciée
- ▶ Problématisation centrale



Une tension entre numérisation et pratiques

Revenir à la poussière ? L'identité professionnelle des historiens et historiennes

Le livre d'Arlette Farge (1989) a connu un tel succès national et international qu'il semble avoir contribué à stabiliser la définition même du métier d'historien et d'historienne autour de celui ou celle qui noircit ses mains de poussière, qui « descend aux archives », etc. C'est la raison pour laquelle les médiations numériques sont très peu évoquées dans les remerciements de thèse, les blogs ou, plus simplement, les livres : historiens et historiennes seraient prisonniers de « faux récits de l'archive » qui le conduisent à valoriser la mise en scène du contact physique au document plutôt que la réalité du travail derrière l'écran ou la fouille via les moteurs de recherche⁸. Un certain « récit de l'archive », déphasé par rapport aux pratiques réelles, reste central dans la construction de l'identité professionnelle. La numérisation du métier est pourtant bien avancée : rares sont les gestes qui ne sont pas médiés par l'ordinateur ou l'instrument, scanner, téléphone ou encore appareil photo. Comment expliquer ce décalage entre récit de l'archive et pratiques concrètes ? Le déni de la numérisation du métier dans la présentation des coulisses des enquêtes historiques révèle la force des représentations qui lient empathie, imprégnation du passé et immersion dans des cartons de documents physiques. Quels seraient des récits d'archive plus proches des pratiques ?

Caroline Muller et Frédéric Clavert, « De la poussière à la lumière bleue », Signata [En ligne], 12 — 2021
<https://journals.openedition.org/signata/3136>

Constats (personnels, à discuter)

- ▶ Une division persistante quanti/quali
- ▶ Des usages "discrets" plus que "computationnels"
- ▶ Limite des exemples disponibles
- ▶ Programmation souvent ramenée aux statistiques (et à R)
- ▶ Encore peu de bibliothèques Python dédiées SHS

Importance de créer des communautés **#pyshs**

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 établissements privés ont pris en charge des patients atteints du Covid-19. « Nous avons été sollicités à plusieurs reprises », assure le directeur général de la Fondation Saint-Vincent à Strasbourg.

Certains sont donc la tempête, d'autres l'attendent. Ainsi, certains patients atteints du Covid-19 « nous avons été sollicités à plusieurs reprises », assure le directeur général de la Fondation Saint-Vincent à Strasbourg.

Des lits transformés pour la réanimation

Ces disponibilités ont pourtant été signalées au ministère de la Santé. « Nous ne sommes pas sollicités à hauteur du service que Samu : on oriente les malades vers le secteur public. Si tous les deux jours, on a déprogrammé toutes nos opérations, nous avons 100.000 soins déprogrammés dans le privé lucratif ».



renforcement » dans d'autres. Le lendemain, le ministre de l'Intérieur a été infecté, a annoncé l'extension des tests de dépistage dans le déconfinement. Sophie Amsili et Tifenn Clir

**** *num_618 *journal_Lefigaro

« Pendant trois heures, Emmanuel Macron a pris connaissance des résultats obtenus par l'équipe du Pr Raoult », se réjouit la députée Martine Wonner, seule parlementaire LREM à avoir rencontré les médecins prescripteurs. « LIRE AUSSI - Retour sur les dessous d'une rencontre surprise ». Cette psychologue, qui travaille au sein d'un cabinet de conseil en management à Paris, s'est également engagée avec les écologistes, c

IRaMUTEQ

Cas : data science et exploration de données

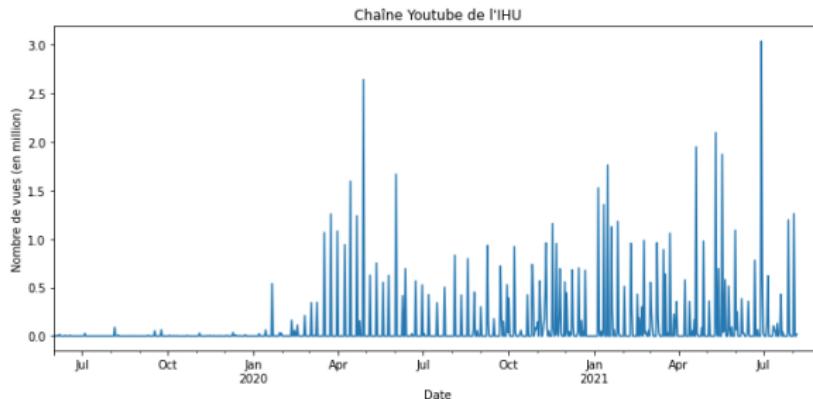
Exploration d'un tableau de données (ici le nombre de vues par vidéos de la chaîne Youtube de l'IHU)

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 rats visit Jard. de la Plante	Journal of Animal Ecology
37	1998	LEROLIA A.	RECODER		
44	1998	LEROLIA A.	LEROLIA A.B.A.	Egg and nest record	Journal of Animal Ecology
47	1998	DE CORNAILLET T.	BERNARD		
52	1999	ARROYO E.	BRETAGNOVILLE C.	Breeding bird	Journal of RSPB
55	1999	ALMAGRO M.	MORCILLO I.	Pastoral sheep Birthing Study	Journal of Animal Ecology
59	2000	ARROYO E.	DECONINCK T.	Reproduction	Journal of Animal Ecology
59	2000	ARROYO E.	DECONINCK T.	Reproduction and age	Condor
62	2000	ARROYO E.	ROUTE B.	PR Activities and Review	Ecology Letters
63	2000	ARROYO E.	ROUTE B.	PR Activities and Review	Ecology Letters
66	2000	SALAMANDRO M.	BUTET A.	LE Responses or Ecologie	Ecology Letters
69	2001	ARROYO E.	MOUGROT C.	BIRD Colonial Breeding Behavior	Ecology Letters
70	2001	LEROLIA A.	ROUTE B.	Colonial Breeding Behaviour	Ecology Letters
71	2001	ROUTE B.	BRETAGNOVILLE C.	Colonial Breeding Behaviour	Ecology Letters



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

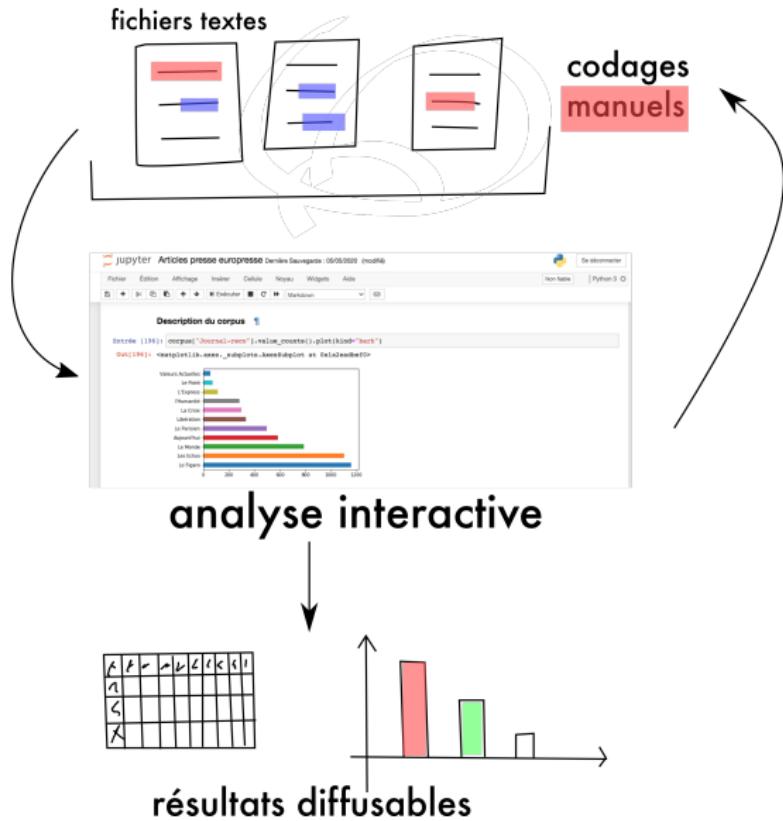
Entrée [3]: rule = gen_rule_payload("ANR lang:fr", results_per_call=50,
                                    from_date="201101210000",
                                    to_date="201102210000")
print(rule)
tweets = collect_results(rule,
                         max_results=1000,
                    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

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 Labellisée 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

* Author to whom correspondence should be addressed.
† Current address: CEPED, 45 Rue des Saints-Pères, 75006 Paris, 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...

Aujourd'hui

- ▶ Ce matin 10h-12h30
 - 1. Les bases du langage Python : mots et structures
 - 2. Python et ses bibliothèques
- ▶ Cette après-midi 13h30-17h
 - 1. La bibliothèque Pandas et ses promesses
 - 2. Traitement statistique et visualisation

Et à tout moment : des échanges sur les usages...

Avant de se lancer

Ne pas hésiter à chercher...

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

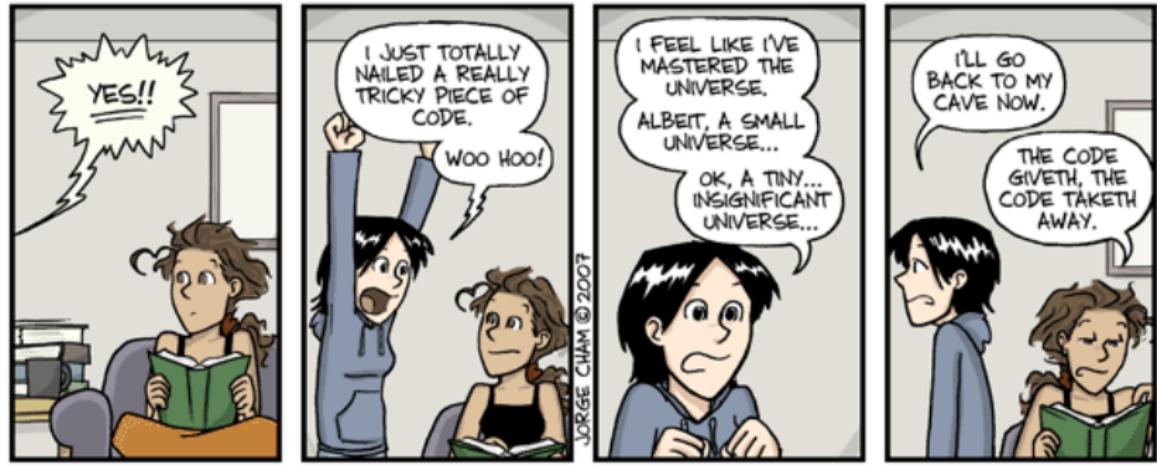


Les obstacles

- ▶ 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.

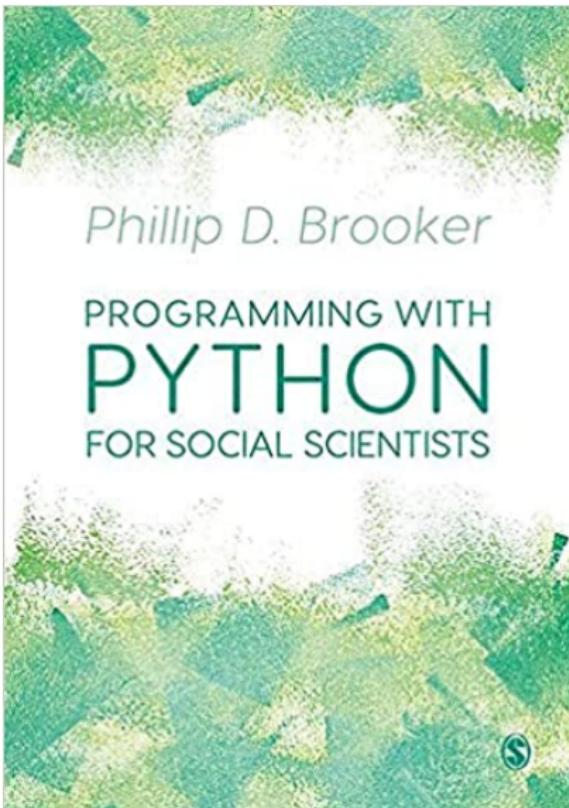
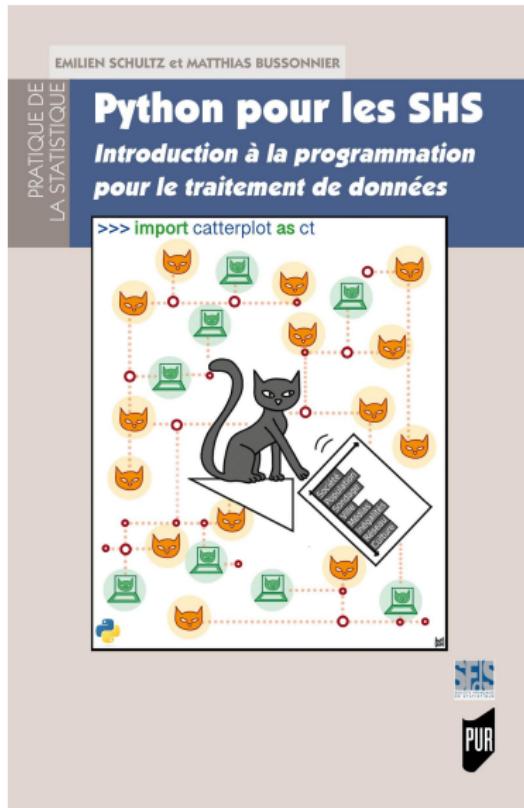


Important de valoriser les petites victoires



WWW.PHDCOMICS.COM

Ressources



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

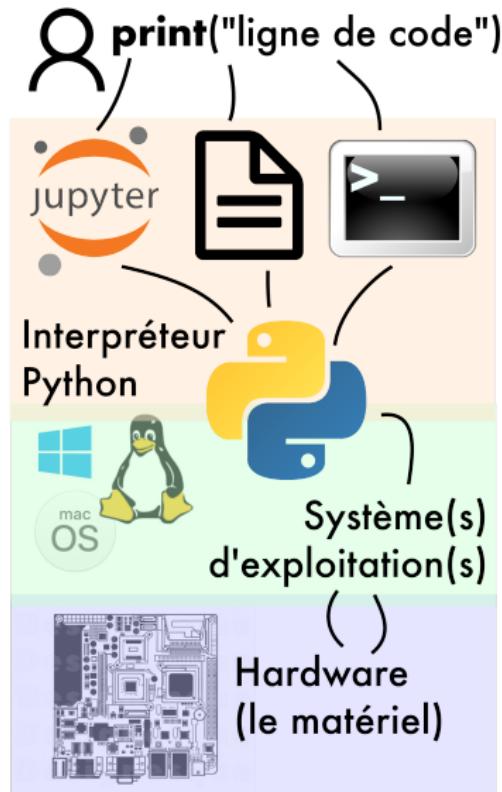
L'environnement nécessaire
pour lancer un script

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

Une philosophie générale : la programmation lettrée (*literate computing*).

- ▶ 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

Petit mémo : structure générale d'un algorithme

Algorithme

