

Python pour les SHS : Pourquoi ? Comment ?

MATE-SHS

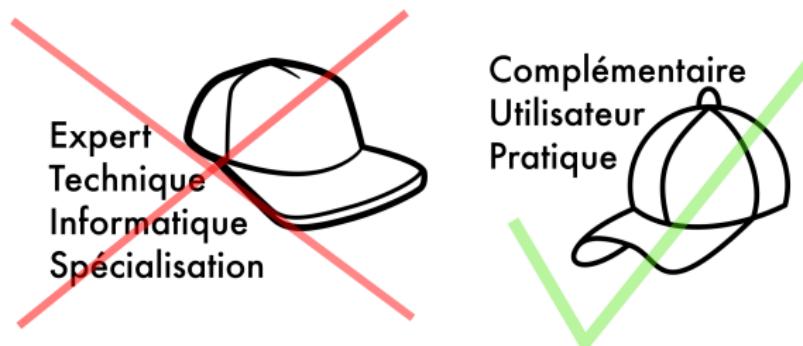
Émilien Schultz (CEPED/SESSTIM)

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

4 janvier 2022

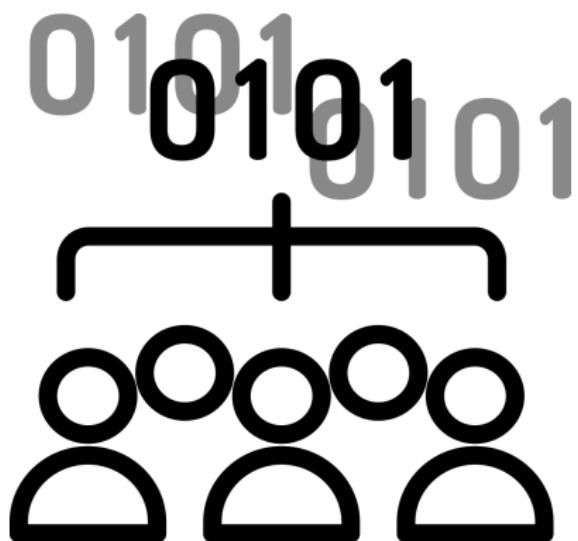
Ce que j'aimerais faire :

- ▶ Discuter de l'intérêt du langage Python pour les SHS
- ▶ Vous montrer un usage pratique [code exécutable pour ceux.illes qui souhaitent ici :
<https://github.com/pyshs/mateshs2022>]
- ▶ Échanger !



D'abord : répondre à 3 questions

1. Pourquoi programmer ?
2. Pourquoi Python ?
3. Pourquoi penser les usages spécifiques aux SHS ?



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
 - ▶ *manipulation données*
- ▶ Courant profond et puissant de la science ouverte
 - ▶ *reproductibilité traitements*
- ▶ Apparition d'objets/méthodes liés aux pratiques numériques
 - ▶ *nouveaux terrain(s)*

Programmer ou quoi ? Ouvrir nos perspectives

Programmer ≠ Construire un logiciel



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

Cinquante nuance de programmation

- ▶ Des *styles* de programmation différentes
- ▶ Un usage spécifique pour la recherche : **la programmation scientifique**
 - ▶ Orientation **script** : réaliser des petites tâches spécifiques
 - ▶ Orientation **interactive** : tester et expérimenter
 - ▶ Orientation **recherche** : des outils spécifiques
- ▶ Usage compatible avec des logiciels et le reste des pratiques
- ▶ Associé à un ensemble d'outils dédiés (en vedette : les Notebooks)

Script scientifique et *literate programming*

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

Une pratique largement orientée data science

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

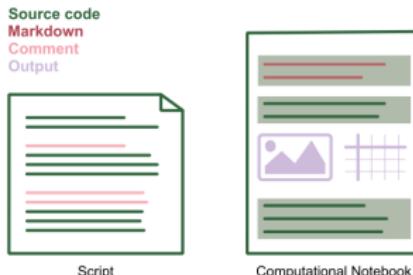


Fig. 1. Current scripting language IDEs support writing and executing code via two programming modalities: *scripts* (left) and *computational notebooks* (right). In this paper we investigate how these modalities are used in data

Une diversité de niveaux de compétences utiles

Découvre la programmation

lecture

Identifier les langages de programmation

Lire un code déjà écrit en Python et la documentation

Lancer une ligne de Python

Réutiliser des scripts existants

résoudre
les erreurs

Écrire des petits scripts

Incorporer du code existant dans ses scripts

Connaissances des bibliothèques et spécialisation

Traduire ses problèmes dans la programmation

Créer des scripts autonomes sur ses problèmes

fonctionnement ordinateur

Réutiliser son code entre les scripts

Partager et faire circuler son code

Créer et maintenir de nouveaux outils

Contributeur•rice Open Source accompli•e

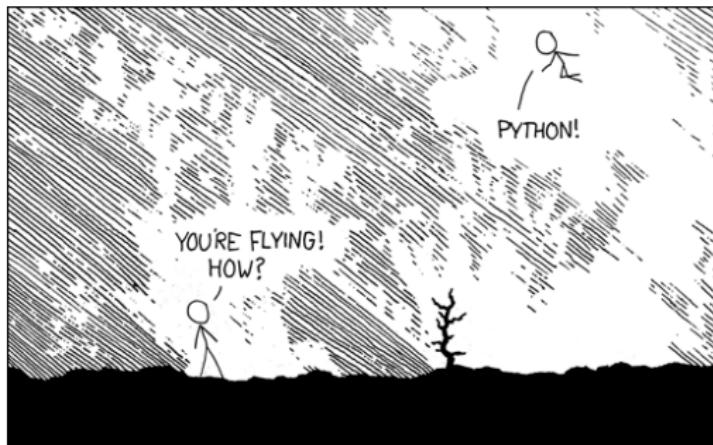
Aussi : programmer comme point d'entrée

Un effet **oignon** :

- ▶ Penser la structures des données et leurs diversité
 - ▶ Format de fichier : csv ou xls ? Passage vers du relationnel ?
- ▶ Penser la matérialité de nos pratiques
 - ▶ Stockage mémoire vive, cloud ou disque dur ?
- ▶ Possibilité d'échanger avec les collaborateurs ressources
 - ▶ Une langue commune entre spécialités

2. 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 *by design*
- ▶ Favorise les bonnes pratiques de programmation
- ▶ En croissance d'usage
- ▶ Un avenir brillant : enseigné dès le lycée

Facile à utiliser comme langage de script

```
(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.
```

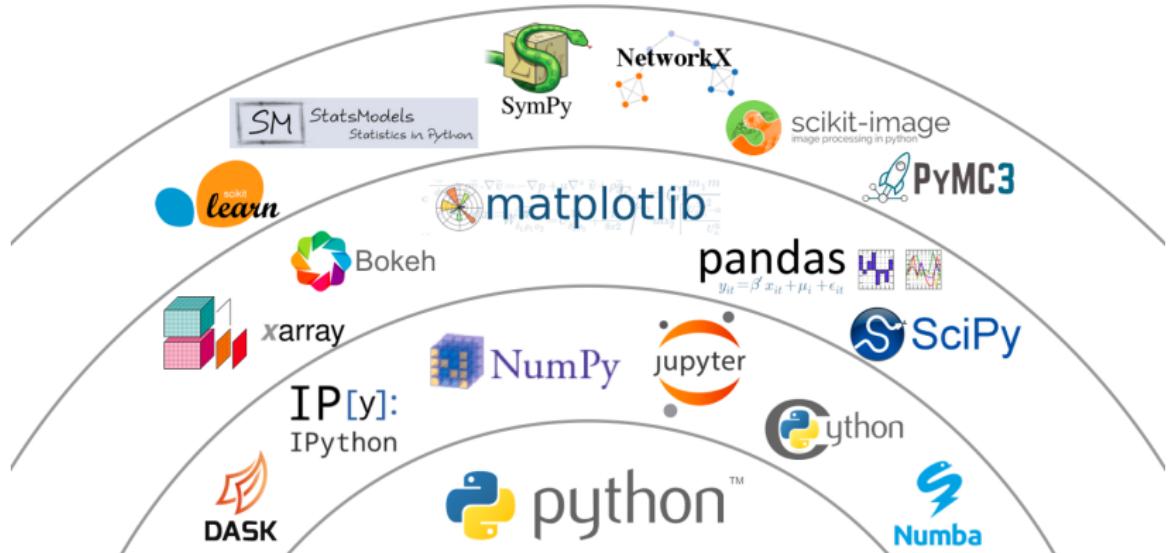
```
In [1]: print("La somme est : ",sum([10,12,8]))
La somme est : 30
```

```
In [2]: 
```

Plus qu'un langage : un univers d'outils

Python's Scientific Stack

Jake Vanderplas PyCon 2017 Keynote



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

Traitement intégré des données

The screenshot shows a Jupyter Notebook interface. At the top, there's a toolbar with icons for file operations, a search bar, and a "Se déconnecter" button. Below the toolbar is a menu bar with "Fichier", "Édition", "Affichage", "Insérer", "Cellule", "Noyau", "Widgets", and "Aide". A status bar at the bottom indicates "Non flable" and "Python 3". The main area contains two code cells. The first cell, labeled "Entrée [3]:", contains Python code that prints the lengths of five lists: COCONEL1 (~1006), COCONEL2 (~1004), COCONEL3 (~2006), TRACTRUST1 (~1014), and TRACTRUST3 (~1005). The second cell, labeled "Out[136]:", shows the resulting dictionary: {1.0: 'Oui', 2.0: 'Non'}. The notebook has a light gray background with code in black and output in white.

```
Entrée [3]: print("COCONEL1 N=", len(data1))
print("COCONEL2 N=", len(data2))
print("COCONEL3 N=", len(data3))
print("TRACTRUST1 N=", len(data4))
print("TRACTRUST3 N=", len(data5))

COCONEL1 N= 1006
COCONEL2 N= 1004
COCONEL3 N= 2006
TRACTRUST1 N= 1014
TRACTRUST3 N= 1005

Out[136]: {1.0: 'Oui', 2.0: 'Non'}
```

[FIGURE 1] Evolution de l'attitude en France

The screenshot shows a Jupyter Notebook cell with Python code. The code reads Google Trends data from a CSV file, processes it to calculate percentages, and then creates a line graph. The graph plots the intensity of Google searches for terms related to hydroxychloroquine in France over time, from April 2020 to June 2021. The graph includes three data series: 'HC is effective' (red circles), 'HC is ineffective' (green circles), and 'Uncertain' (blue circles). The 'Intensity of Google searches using Google Trends' is shown as a gray line. The graph has a sharp peak for the 'HC is effective' series in early 2021, while the other series remain low. The notebook has a light green background for the code cell.

```
Entrée [9]: # Tableau par enquête
d = {"04-07-2020": pyhs.tri_a_plat(data1,"HC_c","RED")["Pourcentage (%)"],
"04-19-2020": pyhs.tri_a_plat(data2,"HC_c","RED")["Pourcentage (%)"],
"06-23-2020": pyhs.tri_a_plat(data3,"HC_c","RED")["Pourcentage (%)"],
"11-03-2020": pyhs.tri_a_plat(data4,"HC_c","RED")["Pourcentage (%)"],
"06-08-2021": pyhs.tri_a_plat(data5,"HC_c","RED")["Pourcentage (%)"]}
t = pd.concat(d,axis=1).drop("Total").T

# Données Google Trends
hc = pd.read_csv("./multiTimeline.csv").replace({'<\xa0>':0})
hc["chloroquine (France)"] = hc["chloroquine (France)"].apply(int)
hc["hydroxychloroquine (France)"] = hc["hydroxychloroquine (France)"].apply(int)
hc["Semaine"] = pd.to_datetime(hc["Semaine"])
hc = hc.set_index("Semaine")["chloroquine (France)"]

# Graphique
t.index = pd.to_datetime(t.index)
ax = t.plot(color=['r','g','b'], figsize=(10,5), marker='o', linestyle='--')
pd.DataFrame(hc.resample("w").sum()).plot(ax=ax,color="gray")
plt.xlim("2020-02-01","2021-06-20")
plt.xlabel("Date (per week)")
plt.ylabel("Percentage (%)")
plt.legend(["HC is effective","HC is ineffective","Uncertain","Intensity of Google searches using Google Trends"])
plt.title("Figure 1. Evolution of attitudes toward HC in France and media coverage between April 2020 and June 2021")

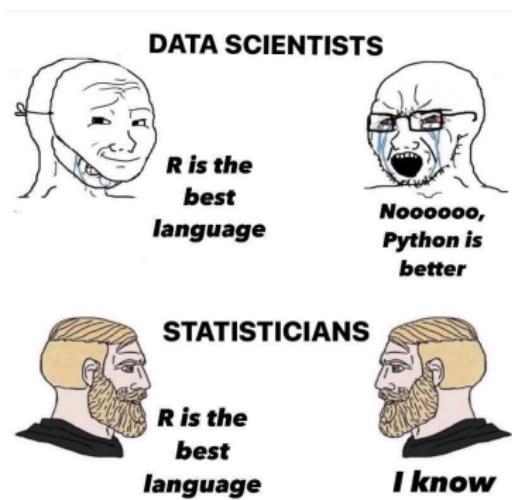
plt.tight_layout()
plt.savefig("./figures/Figure 1 - evolution.png",dpi=1000)
```

Figure 1. Evolution of attitudes toward HC in France and media coverage between April 2020 and June 2021



Mais pas le seul choix...

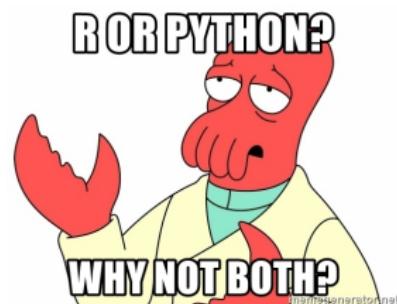
Convergence et divergences avec d'autres langages, R en premier lieu



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

Python ou R ? Python et R ? Ou quoi encore ?

- ▶ Python et R permettent la majorité des traitements associés à la collecte des données, au traitement, et à la visualisation, et évoluent en permanence.
- ▶ Python est davantage compris par les informaticiens et assimilés + secteur privé
- ▶ R excellent pour les statistiques
- ▶ Python est en avance pour les applications en machine learning
- ▶ Python permet de déployer
- ▶ Python semble avoir une meilleure logique de documentation

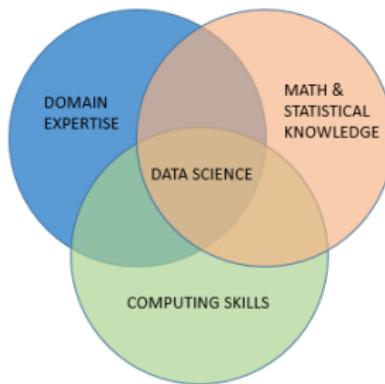


Dans tous les cas, importance des ressources disponibles pour apprendre : collègues, etc.

3. Pourquoi réfléchir les usages spécifiques aux SHS ?

L'autonomisation de la "data science"

- ▶ De plus en plus autonome comme littérature (manuels dédiés, beaucoup tournés vers l'opérationnel)
- ▶ Toujours relatif à des domaines spécifiques



Hétérogénéité des SHS

- ▶ Rôle central de la problématique (perspectivisme)
- ▶ Méthodologies très variées
- ▶ Données plus ou moins accessibles et normalisées
- ▶ Culture du numérique variable



Des identités en transformation autour du numérique

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 (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

Un état des lieux encore à faire...

4. 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

IRaMUTEQ

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	LEROLA A.	LEROLA A., BRETAGNOVILLE N.	Sea rats visit Jardine Islands	Marine Biology
37	1998	A RECODER			
44	1998	LEROLA A.	LEROLA A.B.A.	Egg and nest record card	Marine Biology
47	1998	DE CORNAILLET T., BERNARD C.			
52	1999	ARROYO E.	ARROYO E., BRETAGNOVILLE N.	Breeding bird	Journal of R
55	1999	ALMAMOLARD M.	ALMAMOLARD M., MORISON G.	Patagonian seal study	Marine Biology
59	2000	ARROYO E.	ARROYO E.	Reproductive behaviour	Marine Biology
59	2000	ARROYO E.	DECONINCK T.R., ARROYO E.	Age and gender	Marine Biology
62	2000	ARROYO E.	ARROYO E., BRETAGNOVILLE N.	PF Activities and Review of E	Marine Biology
63	2000	ARROYO E.	ARROYO E.	PF Activities and Review of E	Marine Biology
66	2000	ALMAMOLARD M.	ALMAMOLARD M., BUTET A.	LE Responses or Ecologie	Ecologie
69	2001	ARROYO E., MOUGET F.	ARROYO E., MOUGET F.	BRED Colonial breeder behaviour	Marine Biology
70	2001	LEROLA A.	LEROLA A.	Colonial breeder behaviour	Marine Biology
71	2001	JOUET F.	JOUET F., BRETAGNOVILLE N.	Colonial breeder behaviour	Marine Biology



Cas : data science et exploration de données

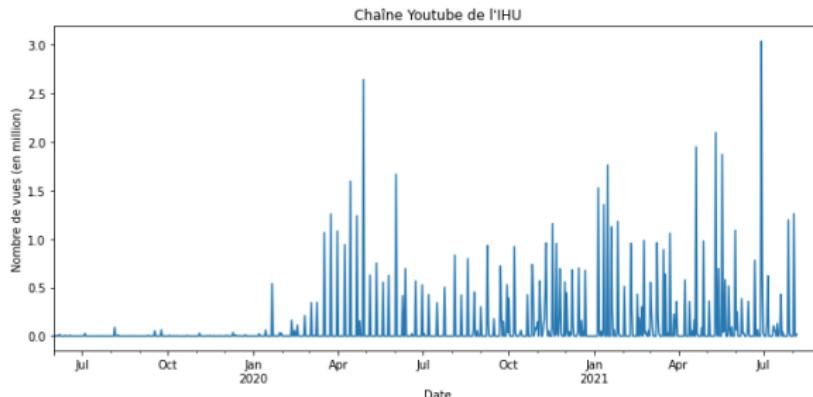
Exploration d'un tableau de données (ici le nombre de vues par vidéos de la chaîne Youtube de l'IHU)

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 : 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"}
```

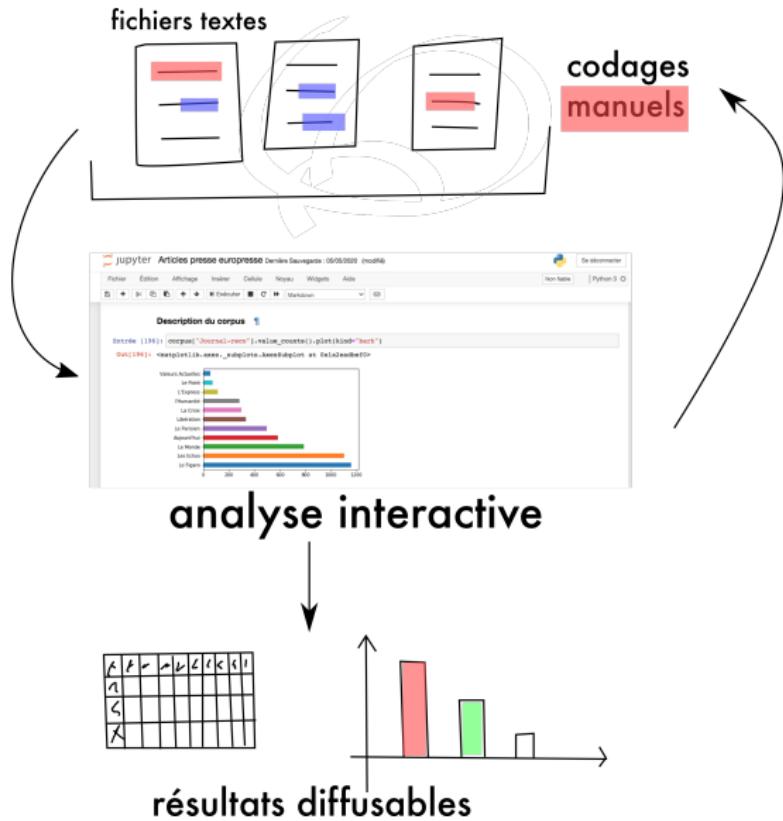
```
Entrée [4]: print(len(tweets))
pd.DataFrame([(i.created_at_datetime,i.all_text) for i in tweets])
136
```

```
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 10:26:41	A lire RT @CollectifPAPER La Cour des Comptes...
4	2011-02-18 10:26:09	La Cour des Comptes rappelle à l'ordre l'ANR
...
131	2011-01-25 07:52:30	Chaires d'excellence de l'ANR: accueil des che...

Ca peut aussi être la transformation de pdf en texte, mobilisant potentiellement de la reconnaissance de caractères.

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 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
* 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
- ▶ Traitement des images...

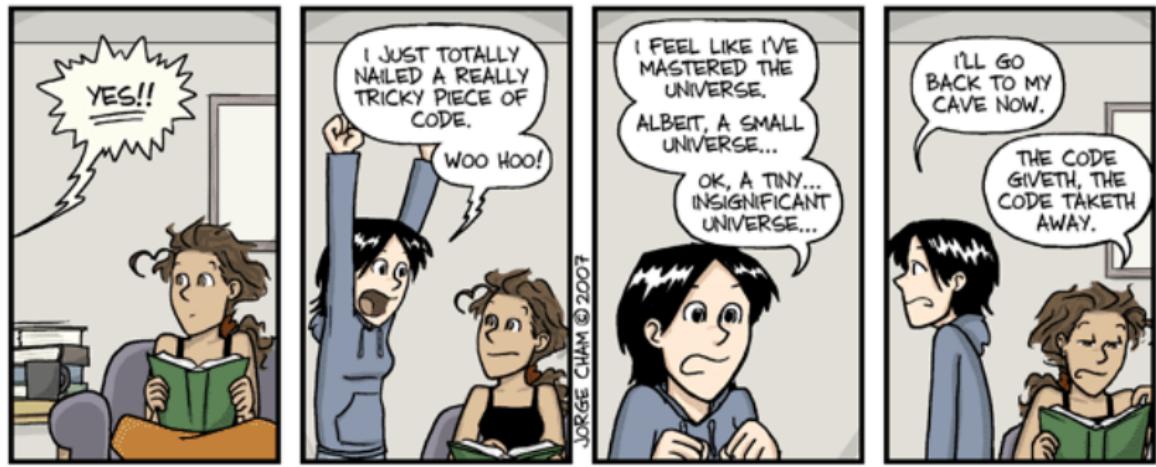
5. S'y mettre !

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

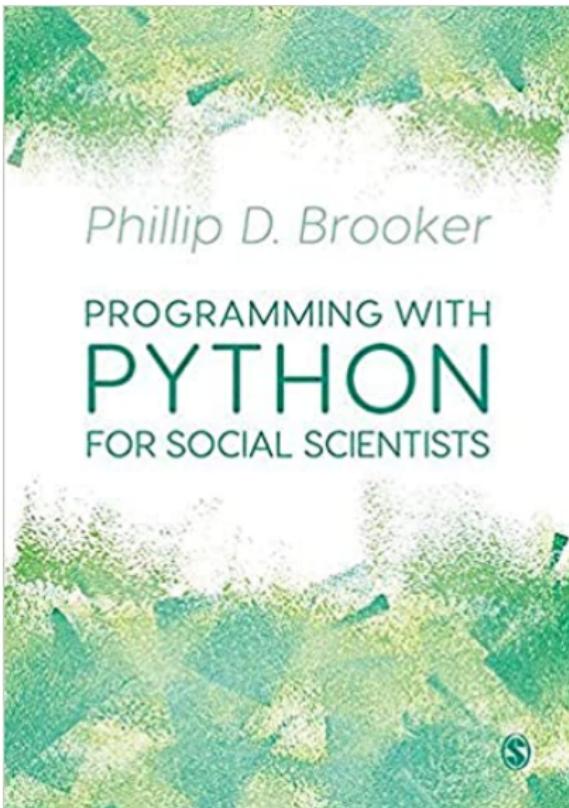
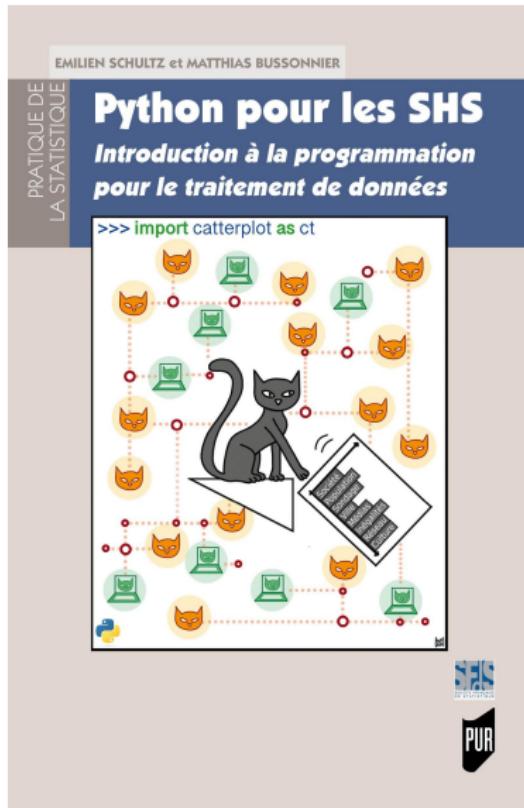


Important de valoriser les petites victoires



WWW.PHDCOMICS.COM

Ressources



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

Créer des espaces collectifs

- ▶ Merci le réseau MATE-SHS de le faire aujourd'hui :)
- ▶ Point de rencontre CocoPySHS (URFIST Lyon) à partir de mars
 - ▶ 17 Mars : Fouille de texte & Ingrédients alimentaires avec Tristan Salord
 - ▶ 7 Avril : Données de questionnaire & Statistiques avec Mariannig Le Béchec et Emilien Schultz
 - ▶ 12 Mai : Collecte & Nettoyage de données avec Lucie Loubère
 - ▶ 9 Juin : Approches cartographiques & science ouverte avec Célya Gruson-Daniel, Maya Anderson-Gonzalez et Camille Moulin -
 - ▶ 7 Juillet : Collecter des données Twitter & Ethnographies numériques avec Léo Mignot