

Programmer en Python pour les SHS ?

Quoi ? Comment ? Pourquoi ?

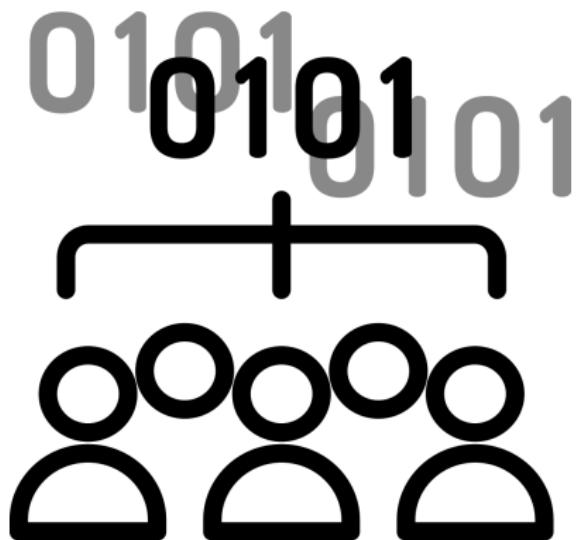
Émilien Schultz

emilien.schultz@sciencespo.fr

médialab - SESSTIM

Avant tout : répondre à 3 questions

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



Pourquoi programmer ?

Programmer !

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

- ▶ Donner des instructions à un ordinateur
- ▶ Automatiser les traitements

Cela se fait en écrivant des instructions dans un langage reconnu par l'ordinateur.

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



J.ME-CLARK.TUMBLR

Cinquante nuance de programmation

- ▶ Des *styles* de programmation différentes (paradigmes)
- ▶ 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

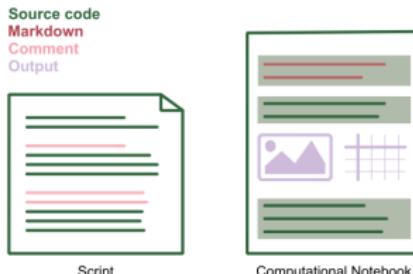


Fig. 1. Current scripting language IDEs support writing and executing code via two programming modalities: *scripts* (left) and *computational notebooks* (right). In this paper we investigate how these modalities are used in data

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

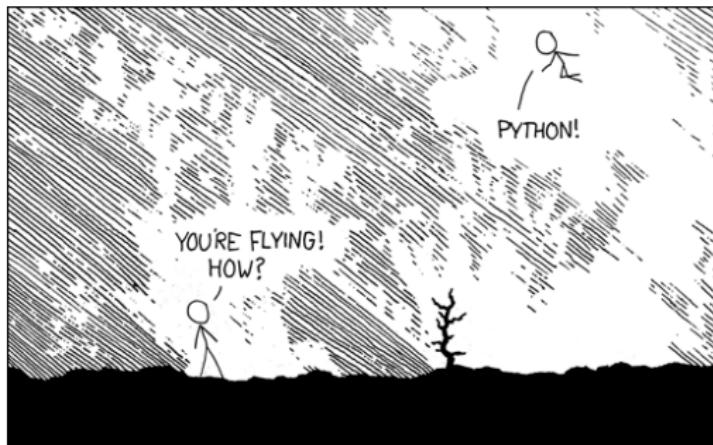
Un effet **oignon** :

- ▶ 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

Un exemple : découvrir qu'une image est en fait un tableau de points, chaque point décrit par trois valeurs (rouge, vert, bleu), et qu'on peut manip

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 (e.g. documentation)
- ▶ En croissance d'usage (recherche et privé)
- ▶ 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.

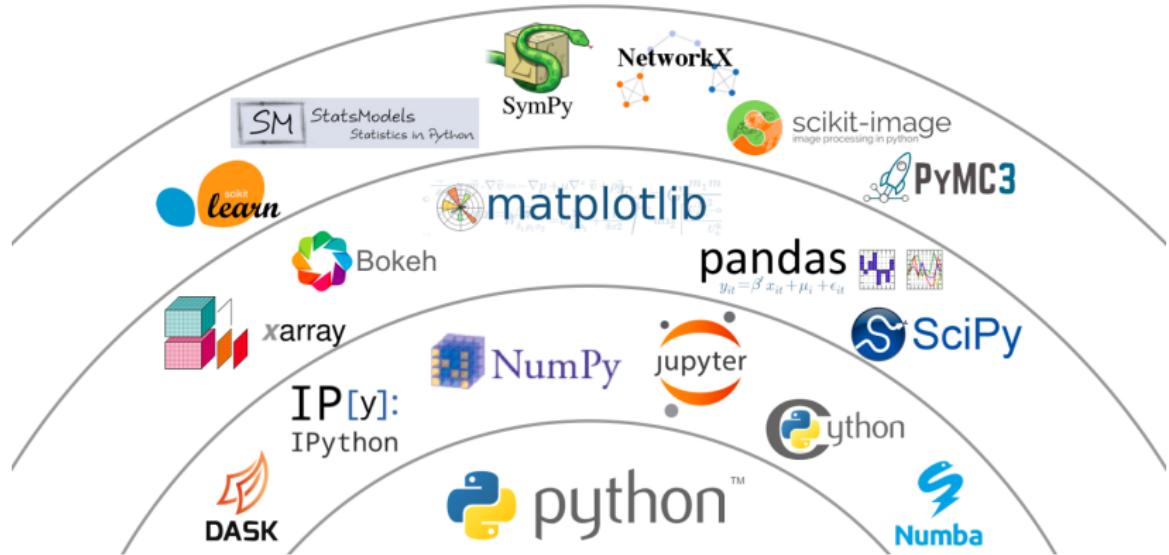
In [1]: print("La somme est : ",sum([10,12,8]))
La somme est : 30

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 ...

De nombreux outils

Free software, open standards, and web services for interactive computing across all programming languages

JupyterLab: A Next-Generat Notebook Interface

JupyterLab is the latest web-based interac environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in

Broken Barb CapStyle Plotting categorical variables Plotting the coherence of two signals

CSD Demo Curve with error band Errorbar limit selection Errorbar subsampling

EventCollection Demo Eventplot Demo Filled polygon Fill Between and Alpha

Lines, bars and markers
Images, contours and fields
Subplots, axes and figures
Statistics
Pie and polar charts
Color
Shapes and collections
Style sheets
axes_grid1
axisartist
Showcase
Animation
Event handling
Front Page
Miscellaneous
3D plotting
Scales
Specialty Plots
Spines
Ticks
Units
Embedding Matplotlib in graphical user interfaces
Userdemo
Widgets

Gallerie Matplotlib

Des bibliothèques puissantes

 [Install](#) [User Guide](#) [API](#) [Examples](#) [Community](#) [More](#) [Go](#)

scikit-learn

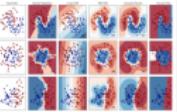
Machine Learning In Python

[Getting Started](#) [Release Highlights for 1.0](#) [GitHub](#)

Classification
Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

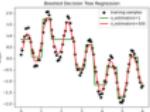
Algorithms: SVM, nearest neighbors, random forest, and more...



Regression
Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...



Clustering
Automatic grouping of samples into sets.

Applications: Customer segmentation, Grouping experimental data.

Algorithms: k-Means, hierarchical clustering, mean-shift, and more...



 [Out now: spaCy v3.3](#)

[USAGE](#) [MODELS](#) [API](#) [UNIVERSE](#)  23,242  [Search docs](#)

Industrial-Strength Natural Language Processing

IN PYTHON

Get things done

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.

Permettant un traitement intégré des données

jupyter Traitement données HC Dernière Sauvegarde : 17/12/2021 (modifié) Se déconnecter

Fichier Édition Affichage Insérer Cellule Noyau Widgets Aide Non flable | Python 3 O

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

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
```

[FIGURE 1] Evolution de l'attitude en France

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({'<\xa01':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")

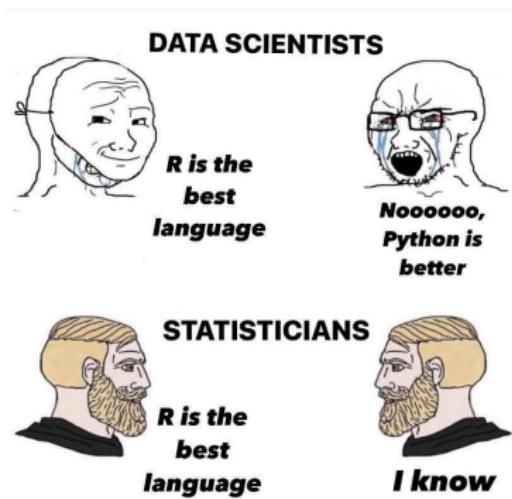
plt.tight_layout()
plt.savefig("./figures/Figure 1 - evolution.png",dpi=1000)
```

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

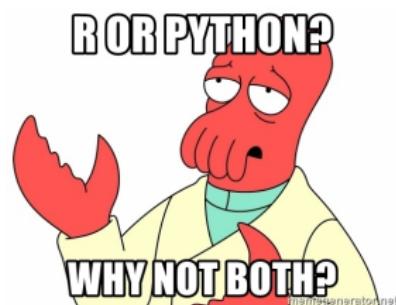


Qui mène à la question centrale : dois-je choisir Python ?

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

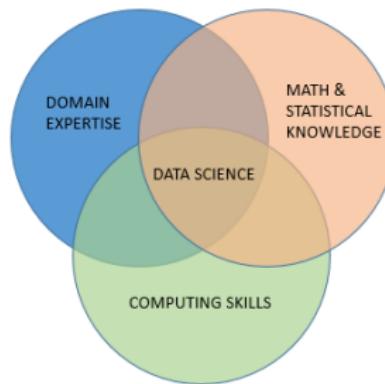
Dans tous les cas, importance des ressources disponibles pour apprendre : collègues, etc.



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>

Des dynamiques en cours

4. FOCUS SUR 3 OUTILS NUMÉRIQUES ET 3 LOGIQUES D'INNOVATION

Nous procédons à une analyse plus approfondie des 3 premiers outils les plus cités: Excel, R et Python. Leurs caractéristiques propres en font à la fois des « concurrents » et des outils complémentaires. Notre analyse tente d'évaluer si l'on peut trouver des profils de chercheurs, qui par leurs caractéristiques propres peuvent être associés à chacun de ces trois outils. Nous constatons que nous rencontrons trois configurations. Nous rencontrons l'innovation : en voie d'institutionnalisation (N. Alter, 2015) symbolisée par R; le logiciel institutionnalisé représenté par Excel; et la pratique en émergence avec Python.

Les utilisateurs de R (n = 244): la voie de l'institutionnalisation

Une moyenne d'âge des utilisateurs de R plus jeune

Les utilisateurs de R se caractérisent par une moyenne d'âge et un âge médian inférieur d'environ 4 ans à la POP. L'usage de R est lié à des chercheurs parmi les plus jeunes, les écarts étant sensibles pour les 35-45 ans et nettement plus marqués pour les chercheurs de moins de 35 ans.

Constats (à discuter)

- ▶ Une division persistante quanti/quali que la programmation permet de dépasser
- ▶ Des usages "discrets" plus que "computationnels" à identifier
- ▶ Constat d'une limite des exemples disponibles : que faire ?
- ▶ Programmation souvent ramenée aux statistiques (et à R)
- ▶ Encore peu de bibliothèques Python dédiées SHS (donc de la place pour en développer de nouvelles)
- ▶ Des usages encore peu stabilisés (Notebooks, etc.)
- ▶ Division du travail vs. culture partagée

Un gros potentiel d'interface entre les pratiques. Mais un état des lieux en cours.

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, no. 23169
évenement, vendredi 20 mars 2020 813 mots, p. 3

Coronavirus

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

Les cliniques privées à la rescousse
SOLVIEG GODELUCK

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

Certains sont dans la tempête; d'autres l'attendent. Ainsi
Faute 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 sign
pouvoir entrer dans le dispositif », plaide Christophe M

« Nous ne sommes pas autorisés 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 Ansill et Tiphaine Clin

**** *enum_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
«vid-19-Laissons les médecins prescrire». » LIRE AUSSI -
Rappelez-vous les dessous d'une rencontre surprise... Cette psych
mais... Elle s'était aussi engagée avec les écologistes, c
coursuivraient ouest de Strasbourg... dont l'énorme chantier a

IRaMuTeQ

Ou encore : passer d'un fichier *.pdf* à un *.txt* pour faire du TAL

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 : exploration de données de l'API aux statistiques

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

jupyter Canal Youtube IHU Dernière Sauvegarde : 02/04/2022 (modifié)

Fichier Édition Affichage Insérer Cellule Nouvel Widgets Aide

Se déconnecter Non faire Python 3 (ipykernel).ipynb

Intrée [24]:

```
df = [i['snippet']['title'], i['snippet']['publishedAt'], i['statistics']['viewCount'], i['id']] for i in corpus]
df = pd.DataFrame(df)
df['date'] = pd.to_datetime(df[1].apply(lambda x: x[0:10]))
df.set_index('date')
df[2] = df[2].apply(int)
df.columns = ['Titre', 'Date', 'Vues', 'ID']
df['Vues_M'] = (df['Vues']/1000000)
```

Outil 24:

Date	Titre	Date	Vues	ID	Vues_M
2022-04-01	Les Jeuds de l'IHU - Sophie Barre	2022-04-01T08:10:49Z	7406	uHJ_uHRTA	0.007406
2022-04-01	Les Jeuds de l'IHU - Maxime Bernard Godin	2022-04-01T08:14:12Z	9642	bhGCoGvGLAv	0.009642
2022-04-26	Les Jeuds de l'IHU - Origine et fonction des ...	2022-01-26T09:16:46Z	26148	sICnQXWqDmf	0.026148
2022-01-26	Les Jeuds de l'IHU - Origine et fonction des ...	2022-01-26T21:16:17Z	19453	vrCohPspGbz	0.019453
2022-01-27	Les Jeuds de l'IHU - Origine et fonction des ...	2022-01-27T23:29:37Z	11530	XH89f7-WNGE	0.011530
...
2018-09-12	Les Jeuds de l'IHU: DB Morier 2018 - 2 - ...	2018-09-12T09:49:09Z	663	xyM29RQHw	0.000663
2018-09-12	Les Jeuds de l'IHU: DB Morier 2018 - 3 - ...	2018-09-12T09:46:07Z	340	vAg2Zenfc_Bc	0.000340
2018-09-12	Les Jeuds de l'IHU: DB Morier 2018 - 4 - ...	2018-09-12T11:24:03Z	736	xhYnKnpMyk	0.000736
2018-09-12	Les Jeuds de l'IHU: DB Morier 2018 - 5 - ...	2018-09-12T11:24:48Z	325	uPrz6B810w	0.000325
2018-09-12	Les Jeuds de l'IHU: DB Morier 2018 - 6 - ...	2018-09-12T11:29:08Z	677	MvWigXALM	0.000677

903 rows x 5 columns

Faire une représentation graphique en regroupant par jours les vidéos

Intrée [34]:

```
ax = df['Vues_M'].resample('d').sum().plot(figsize=(15,5),style="--")  
# Fenêtre temporelle  
plt.xlim("2018-04-01", "2022-03-30")  
  
# Mise en forme et sauvegarde  
plt.ylabel("Nombre de vues (en million)")  
plt.xlabel("Date")  
plt.title("Chaine Youtube de l'IHU")  
plt.savefig('ihu_youtube.png',dpi=200,bbox_inches="tight")
```

Chaine Youtube de l'IHU

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"}
```

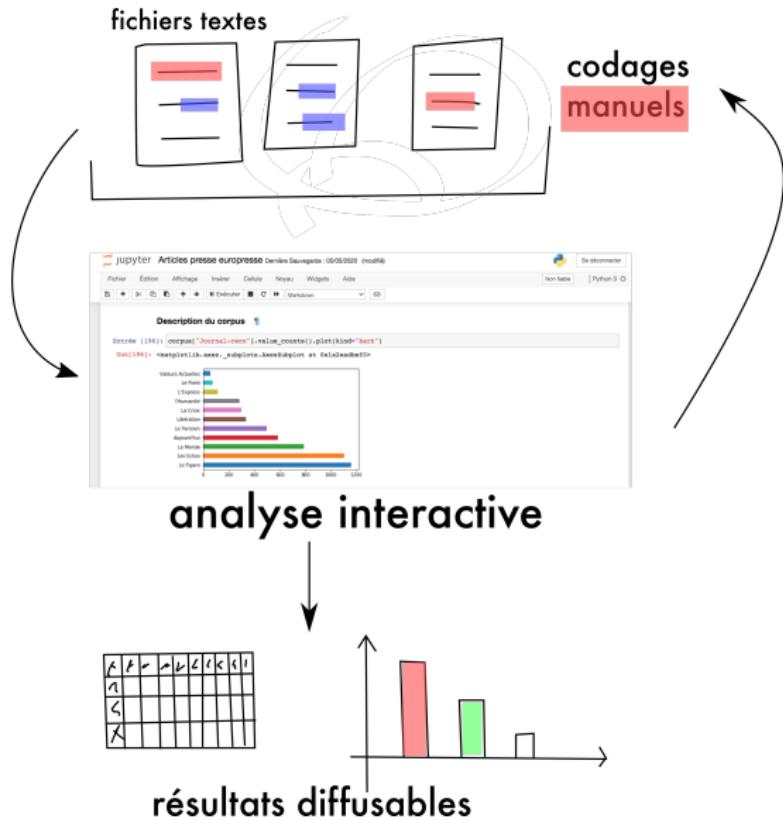
```
Entrée [4]: print(len(tweets))
pd.DataFrame([(i.created_at_datetime,i.all_text) for i in tweets])
```

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 @CollectifPAPERA 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...

Cas : codage de matériel qualitatif



Outils dédiés facilement interfaçable



doccano

code quality CI passing

dooccano is an open source text annotation tool for humans. It provides annotation features for text classification, sequence labeling and sequence to sequence tasks. So, you can create labeled data for sentiment analysis, named entity recognition, text summarization and so on. Just create a project, upload data and start annotating. You can build a dataset in hours.

Demo

You can try the [annotation demo](#).

A screenshot of the doccano annotation interface. The main area shows a text document about Donald John Trump with various entities highlighted in boxes. The sidebar on the right displays project details: File: webPage0, Value: 484872; Date: 1946; Political party: Republican; Spouse: Melania Knauss; Parents: Fred Trump, Mary Anne MacLeod. The top navigation bar shows the URL as 127.0.0.1:3000/projects/sequence-labeling.

Cas : toponyme et cartographie

A partir d'un texte, identifier les lieux géographies et produire des cartes (potentiellement interactives)



Cartes interactives

A partir d'outils de cartographies, nous avons obtenu une visualisation des noms géographiques cités dans les romans, ce qui permet de se faire une idée des zones du monde qui faisaient partie de l'univers des enfants français sous la troisième République. Deux titres, Petite-Pierre ou le bon cultivateur, Maurice ou le travail, sont antérieurs à la troisième République mais sont encore présents dans les listes de manuels de la troisième République.

Répartition géographique des lieux cités dans le corpus

Cliquer sur une carte pour afficher la version pleine page interactive

[Carte de chaleur reprenant tous les lieux cités dans l'ensemble du corpus de romans scolaires](#)



<https://baoia.huma-num.fr/contact/>

tutoriel-complet-de-lextraction-documentaire-a-la-cartographie

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 CERMEES3, 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 : étendre des traitements existants

- ▶ Travail de littérature italienne sur un corpus et thématique du doute
- ▶ Codage des passages dans le corpus
- ▶ Entrainement d'un modèle de NER SpaCy
- ▶ Extension de la réflexion sur l'ensemble des écrits de l'auteur

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

- ▶ Garder une mémoire de ses traitements.
- ▶ Collaboration autour des données : partager son code, faire relire ses résultats intermédiaires
- ▶ Traitement massif de données : parallélisation, déploiement sur des grandes infrastructures, recours aux outils du machine learning
- ▶ Créer une interface utilisateur pour accéder à vos 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



Programmer ≠ Tout savoir

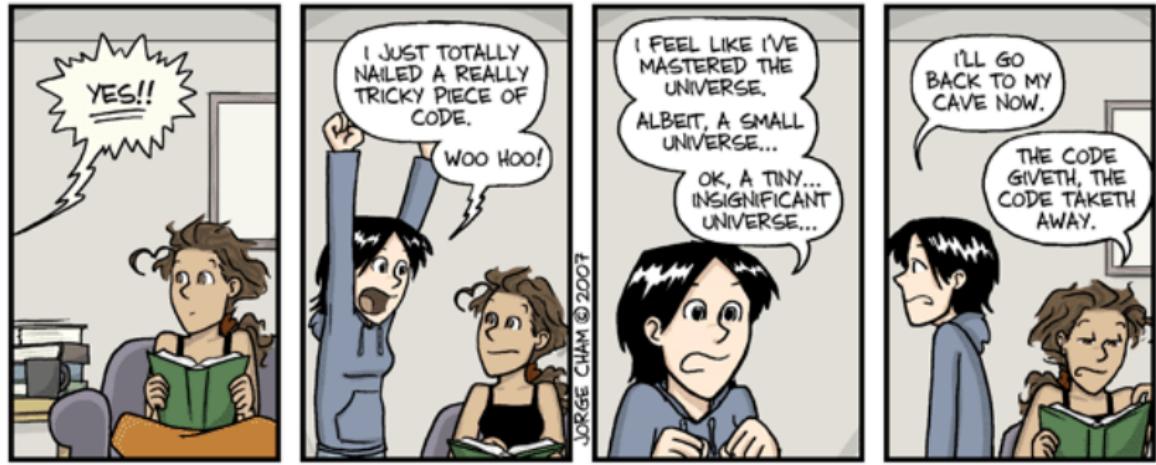
Apprendre à programmer signifie apprendre à potentiellement pouvoir utiliser de nombreux outils développés par des chercheurs.

Mais chaque domaine a ses savoirs spécifiques : *machine learning*, analyse de réseaux, textométrie, ...

La frontière peut être difficile à tracer.

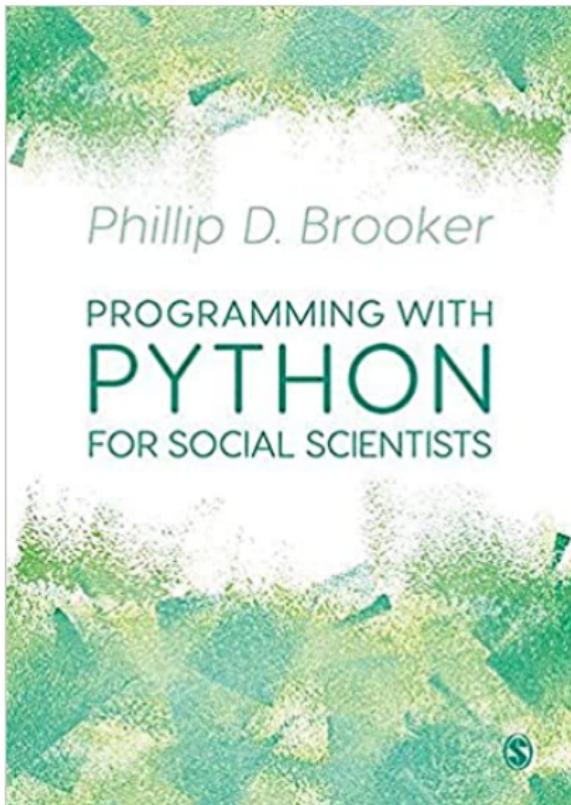
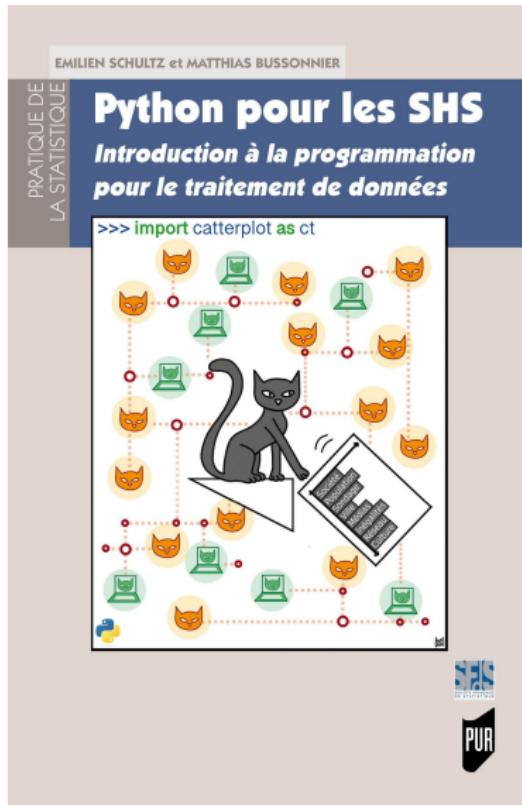
- ▶ Réutilisation d'outils facilité
- ▶ Mais cela ne replace pas une connaissance experte

Important de valoriser les petites victoires



WWW.PHDCOMICS.COM

Ressources



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

Des espaces collectifs à construire

#CocoPySHS

Les CO ulisses du CO de Python pour les SHS



échanger autour de nos pratiques de **programmation** en **Python** pour les **SHS**

partager nos **expériences**

favoriser la **reproductibilité**

développer de **bonnes pratiques d'ouverture**

*Un jeudi par mois, de 13h à 14h
(en visioconférence)*



17 mars 2022 - Fouille de texte & Ingrédients avec Tristan Salord



7 avril 2022 - Données de questionnaire & S avec Mariannig Le Béchec et Emilien Schulte



12 mai 2022 - Collecte & Nettoyage de données avec Lucie Loubère



9 juin 2022 - Approches cartographiques et de science ouverte avec Célya Gruson-Danielle Anderson-Gonzalez et Camille Moulin

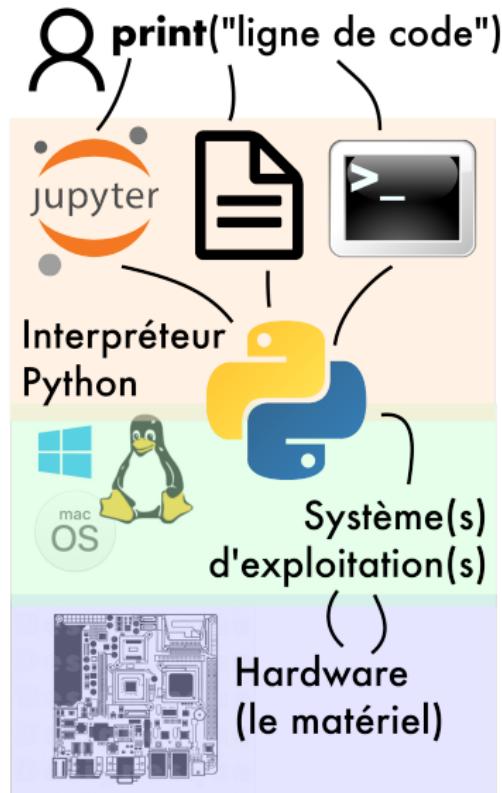


7 juillet 2022 - Collecter des données Twitter pour Ethnographies numériques avec Léo Mignot

Une fois ceci dit : qu'est-ce que ça veut dire concrètement ?

Rendez-vous à la formation le 24 juin avec URFIST Lyon ;)

Prérequis 1 : Où faire du Python



Prérequis 2 : Notre choix



Products

Pricing

Solutions

Resources

Partners

Blog

Company

Contact Sales

Individual Edition is now

ANACONDA DISTRIBUTION

The world's most popular open-source Python distribution platform

Anaconda Distribution

Download

For MacOS

Python 3.9 • 64-Bit Graphical Installer • 515 MB

Get Additional Installers



Open Source

Access the open-source software you need for projects in any field, from data visualization to robotics.



User-friendly

With our intuitive platform, you can easily search and install packages and create, load, and switch between environments.



Trusted

Our securely hosted packages and artifacts are methodically tested and regularly updated.