

MOOC Machine learning in python with scikit-learn

UN COURS SUR SCIKIT-LEARN PAR SCIKIT-LEARN



Inria



QUELQUES MOTS SUR MOI

Doctorat en physique des particules
enseignant les sciences depuis 2012

"core" contributeur à scikit-learn

2 ans en charge du MOOC "Machine
learning in python with scikit-learn"



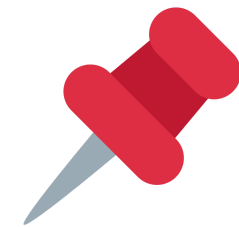
@ArturoAmorQ

UN BREF TOPO

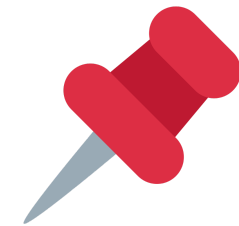
- 📌 gratuit
- 📌 vous recevez un badge
- 📌 rien à installer
- 📌 maintenu par des développeurs de scikit-learn
- 📌 garantie de tourner sur la dernière version



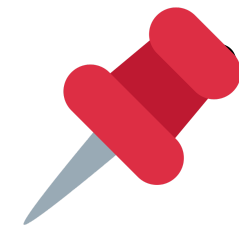
LA VISION DU MOOC



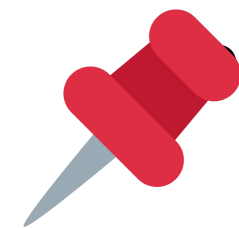
accessible avec des exigences limitées
(connaissance de base de Python, anglais)



intuitions sans entrer dans les mathématiques



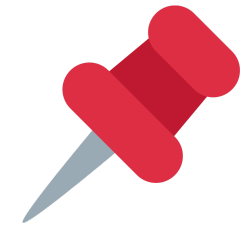
licence CC-BY : rendre le matériel réutilisable
autant que possible



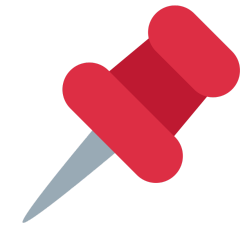
matériel réutilisable dans différents contextes :
tutoriels, enseignement en direct (Inria Academy,
cours universitaires)



EN QUOI ÇA CONSISTE?



7 modules + 1 module d'introduction



15 leçons vidéo

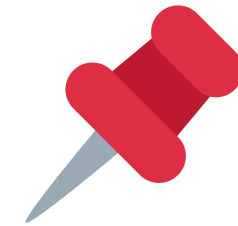


70 cahiers + 21 exercices




26 quizzes + 7 wrap-up quizzes

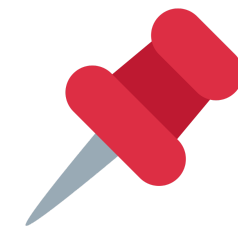
EN QUOI ÇA CONSISTE?



7 modules + 1 module d'introduction



15 leçons vidéo  Convivialité vs
facilité de maintenance



70 cahiers + 21 exercices



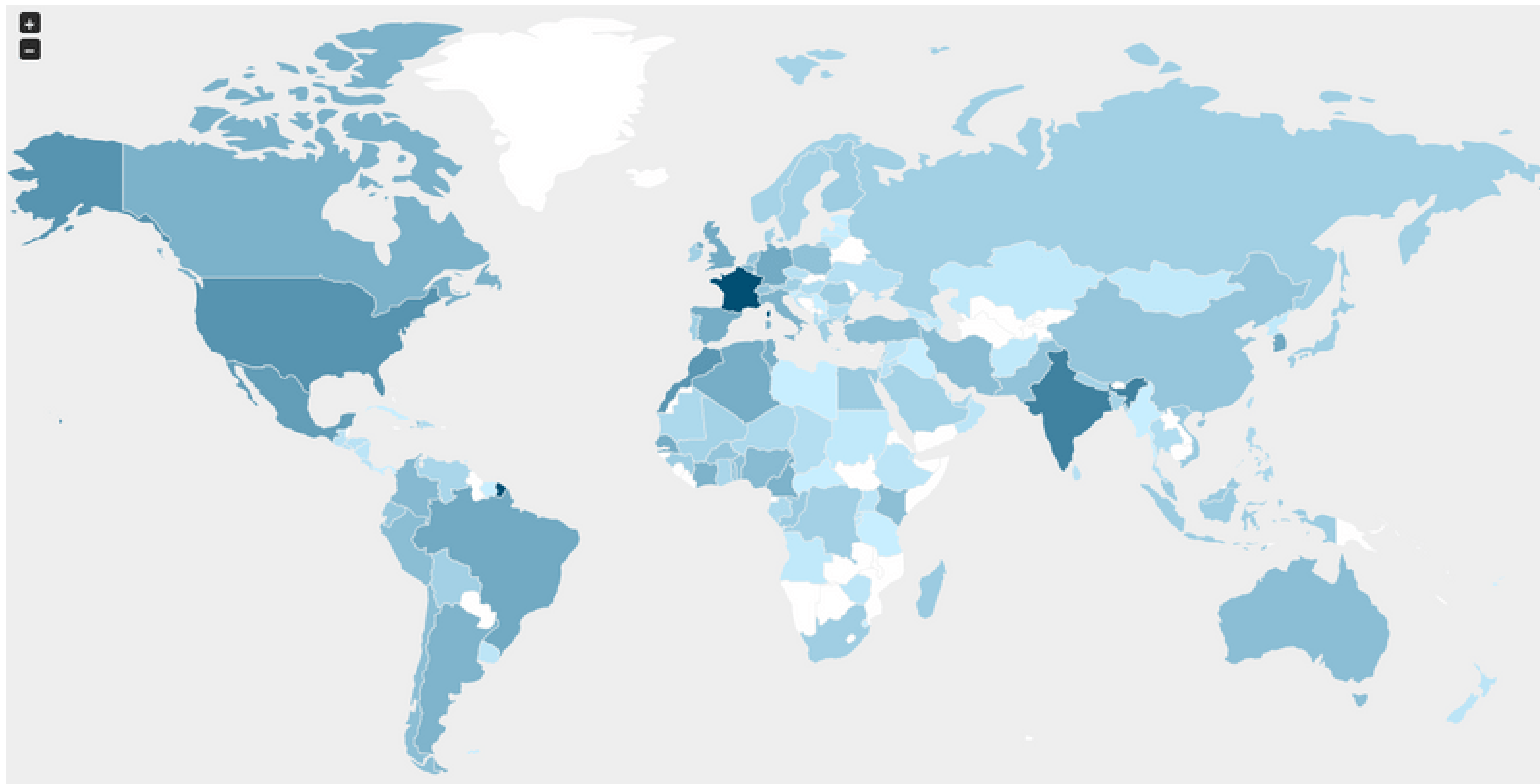
26 quizzes + 7 wrap-up quizzes



3 sessions (durée entre 2-3 mois)

~10k souscriptions/session

~140 pays



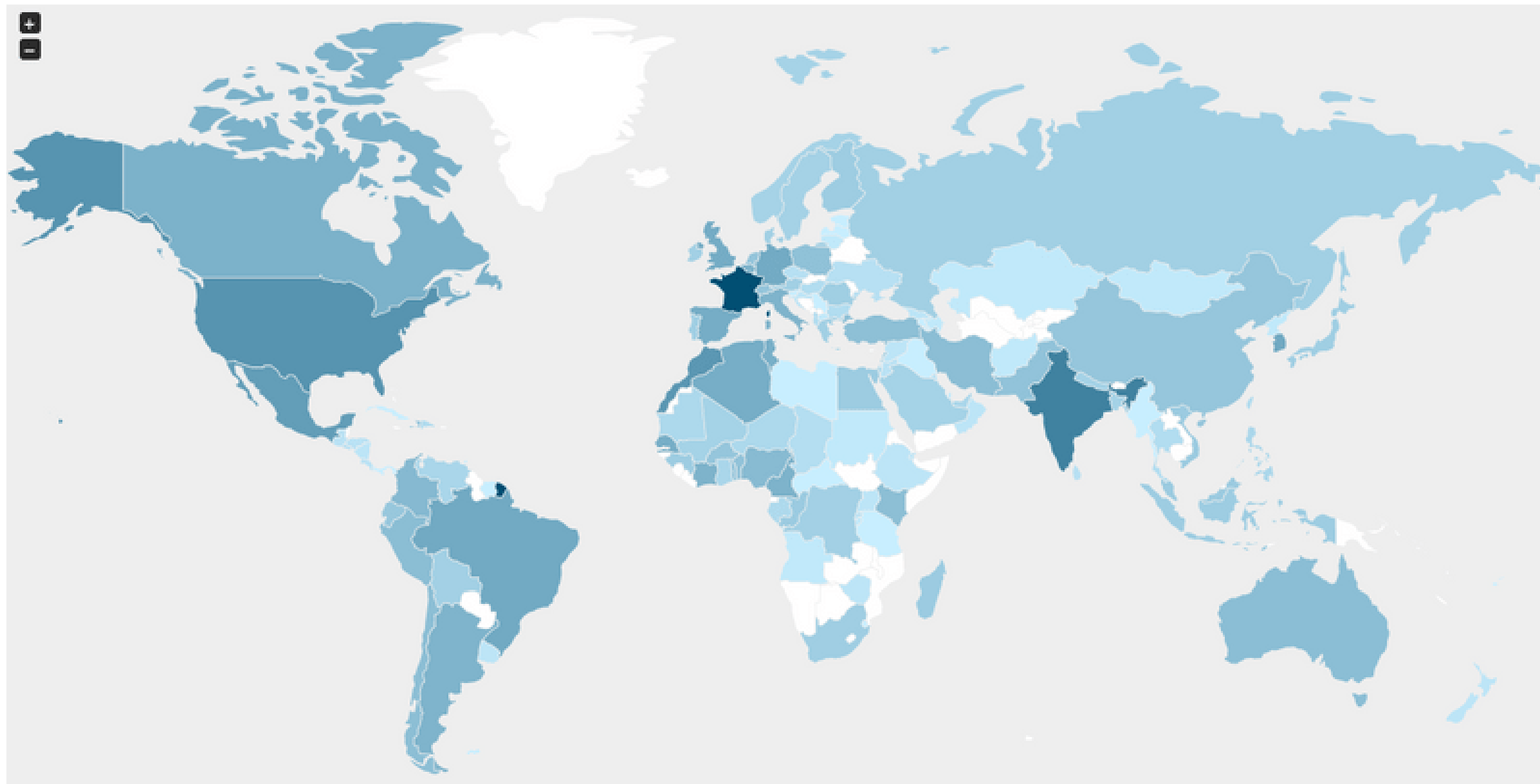
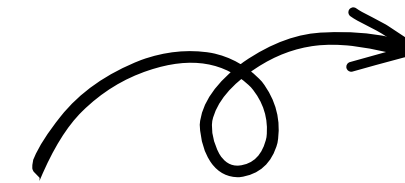


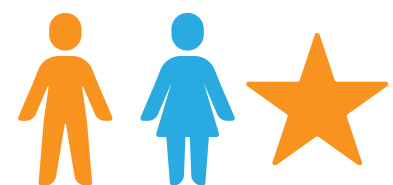
3 sessions (durée entre 2-3 mois)

~10k souscriptions/session

~140 pays

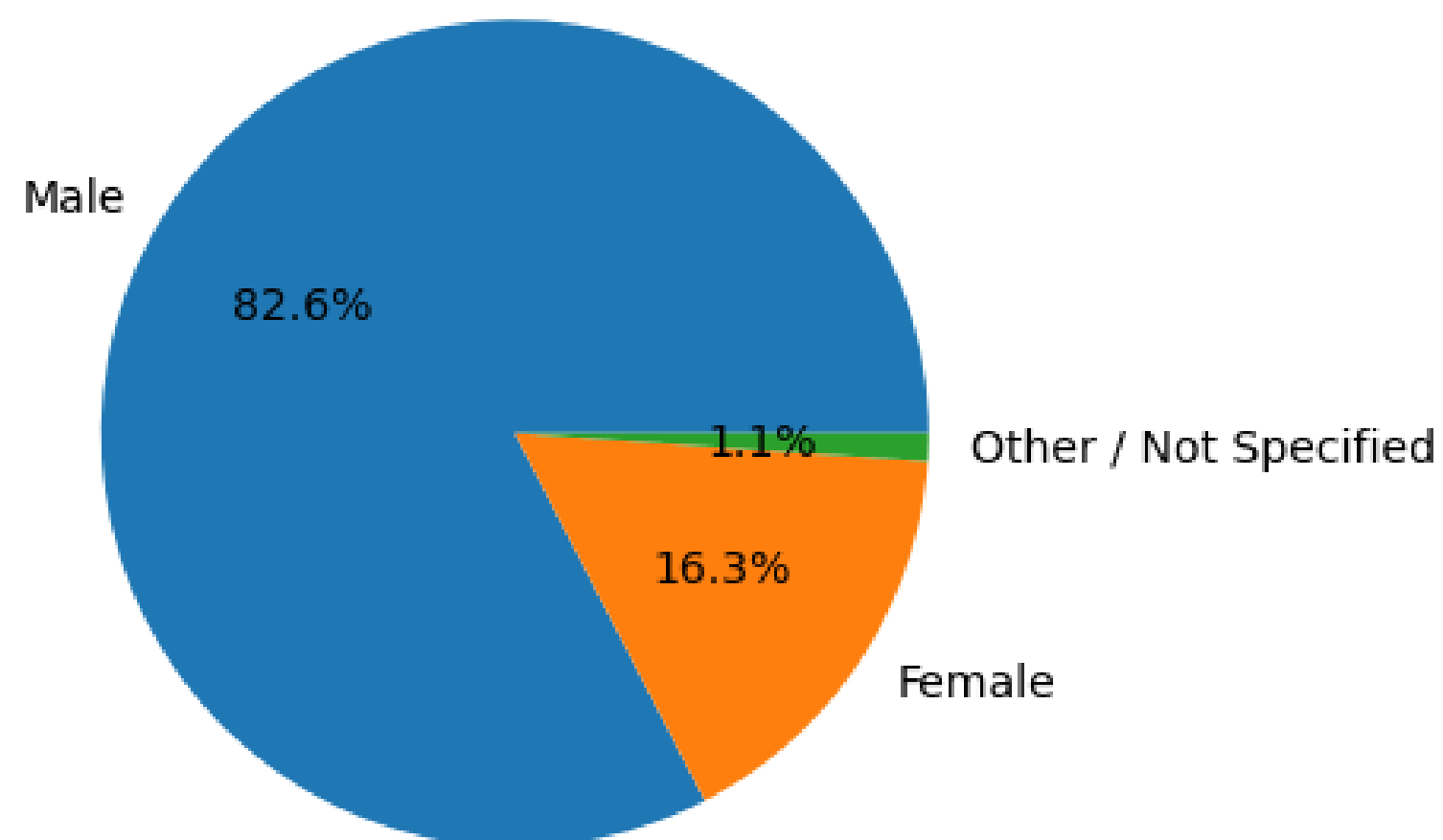
Principalement pays
francophones



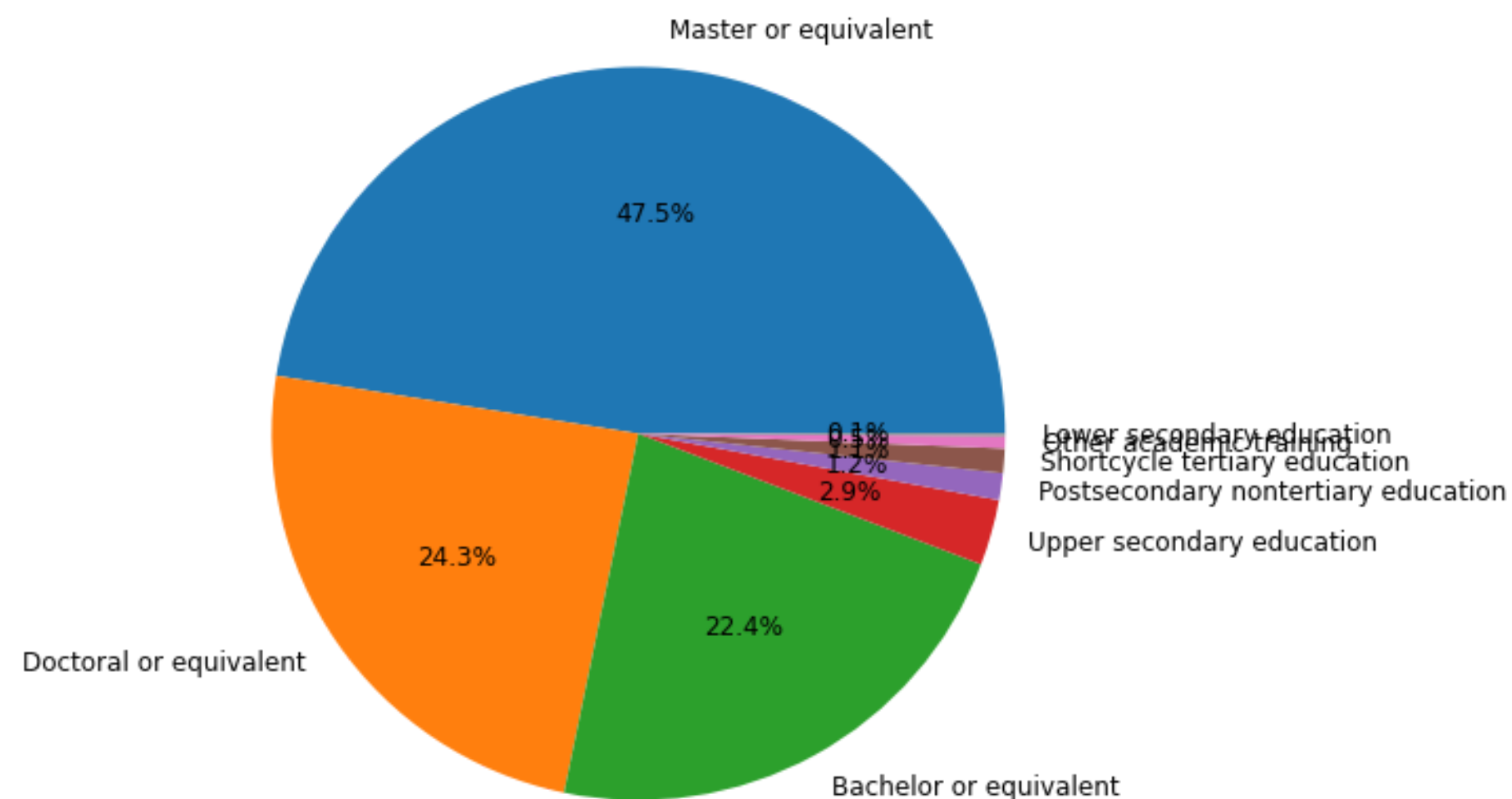


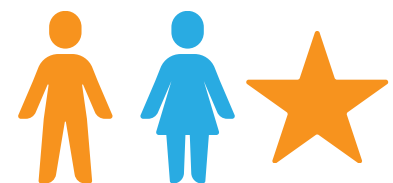
Connaître notre public

Public plutôt masculin



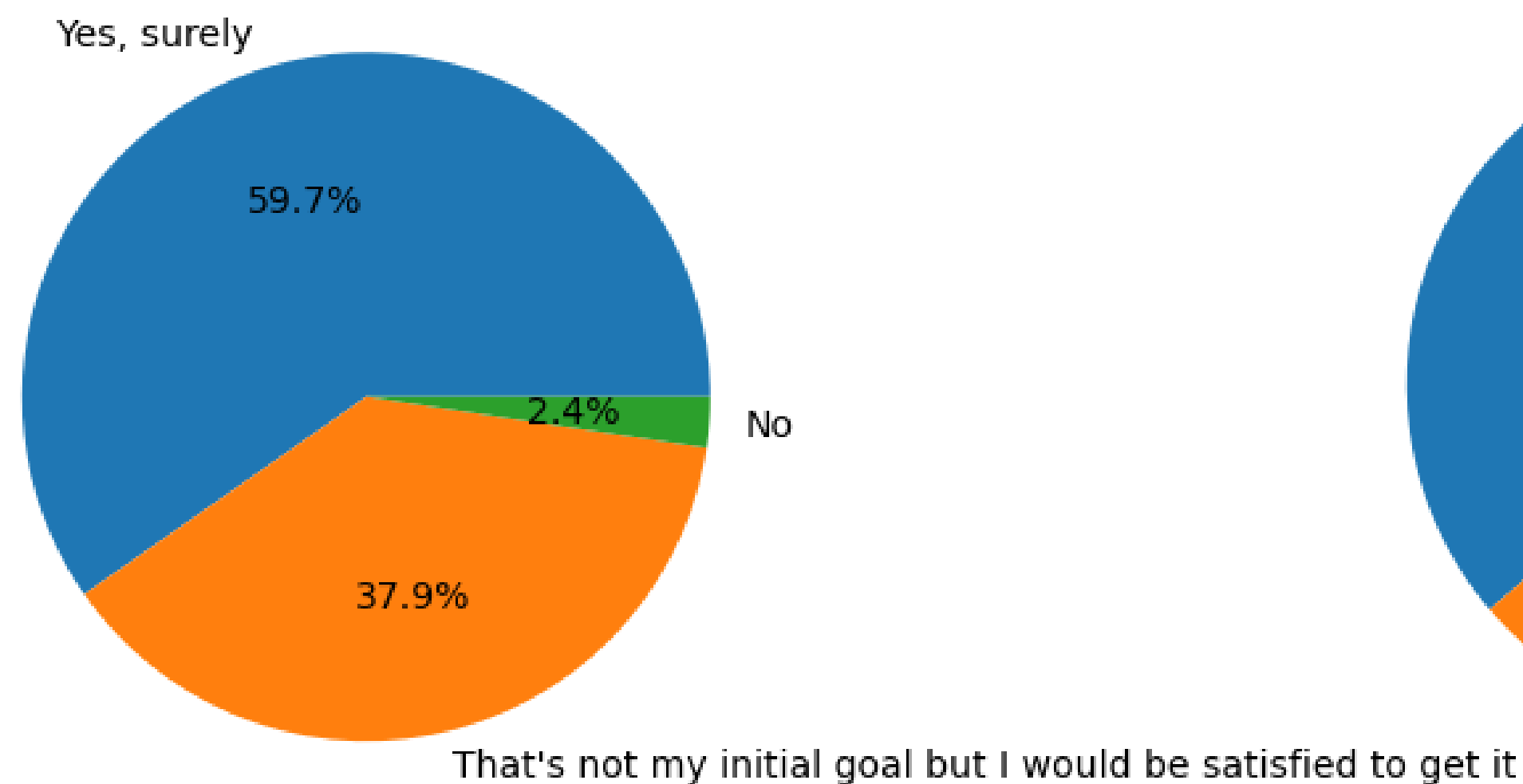
La plupart ayant fait un Master



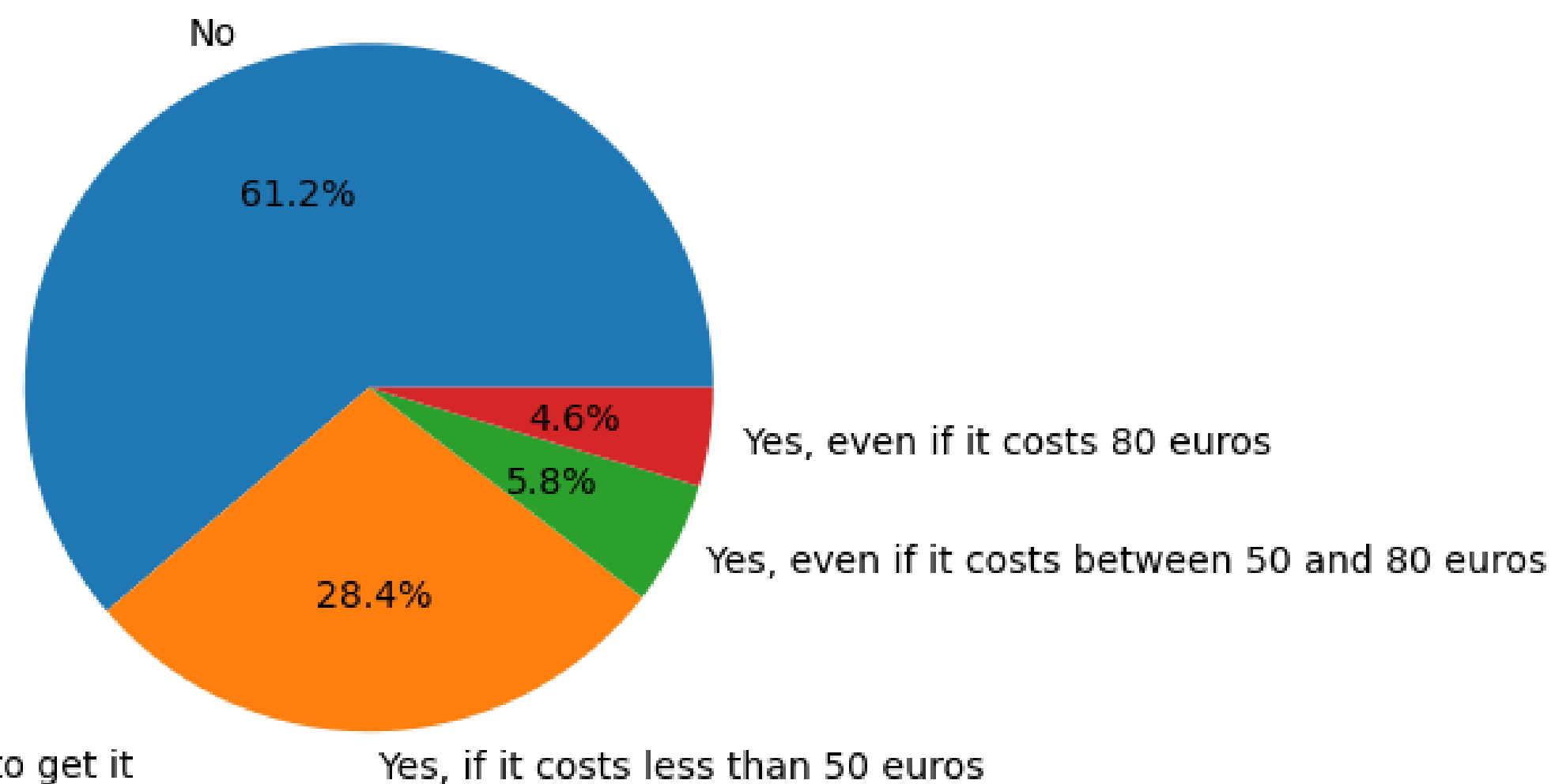


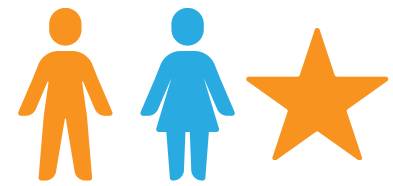
Connaître notre public

Votre objectif est d'obtenir le badge (gratuitement) ?



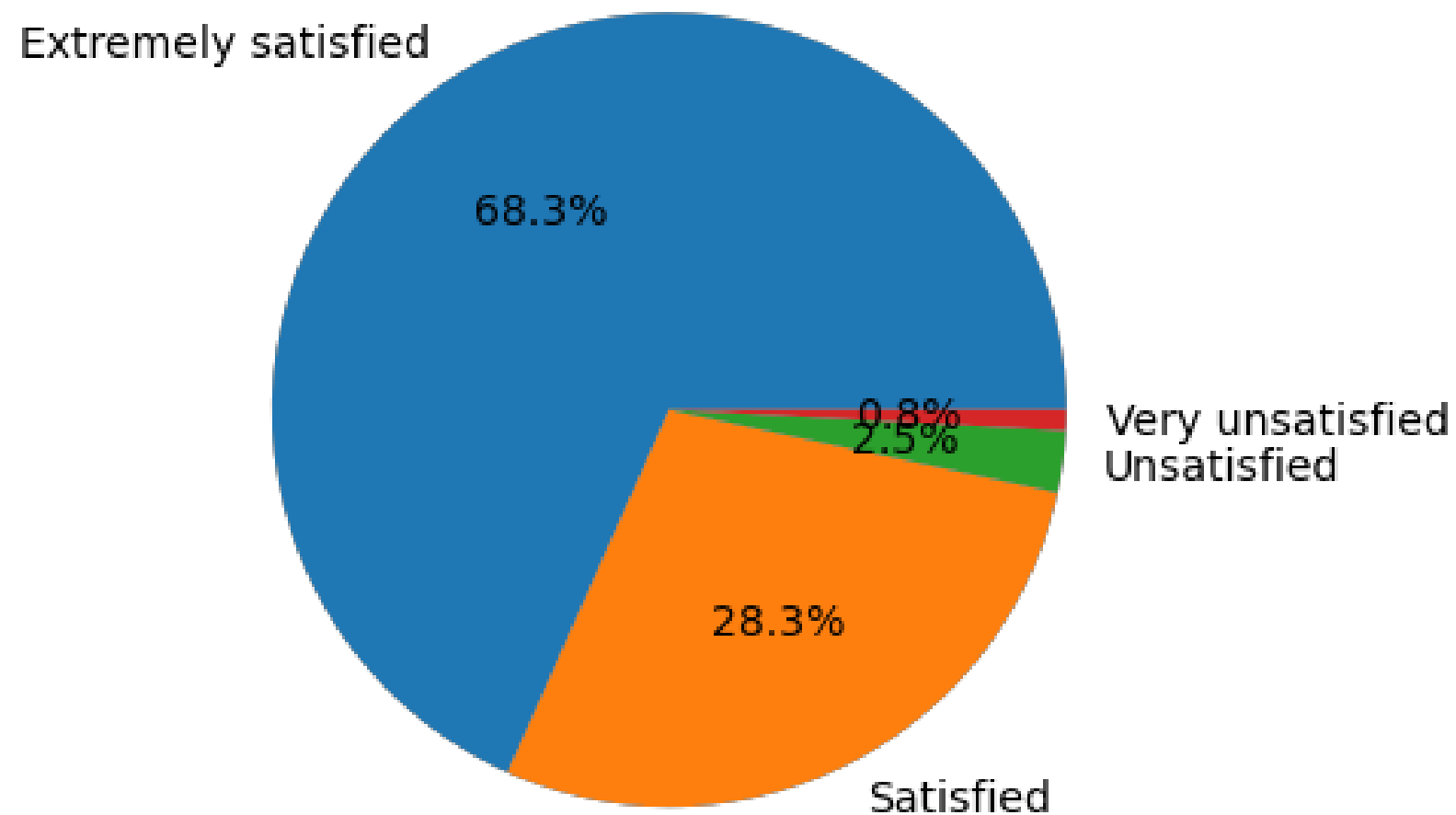
Seriez-vous prêts à payer pour cela ?



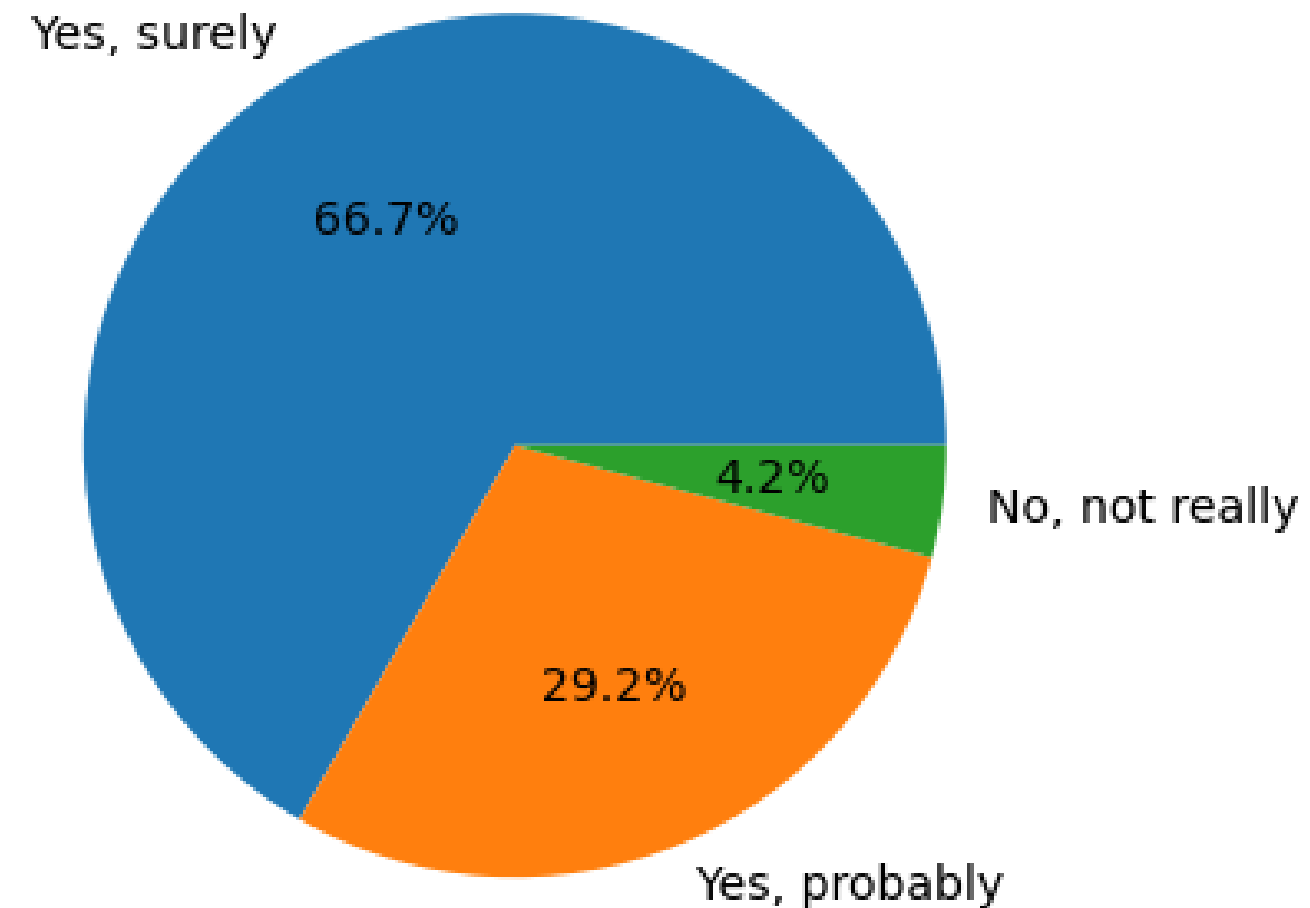


Connaître notre public

Quel est votre degré de satisfaction sur ce MOOC ?



Serez-vous en mesure de pratiquer ou d'appliquer ce que vous avez appris ?





Selection de retours apprenants

"A long time ago, I used to read Descartes, and it blew my mind that this guy literally re-framed the world with his ideas. I think of each of you contributing to this MOOC and to scikit-learn in the same vein."

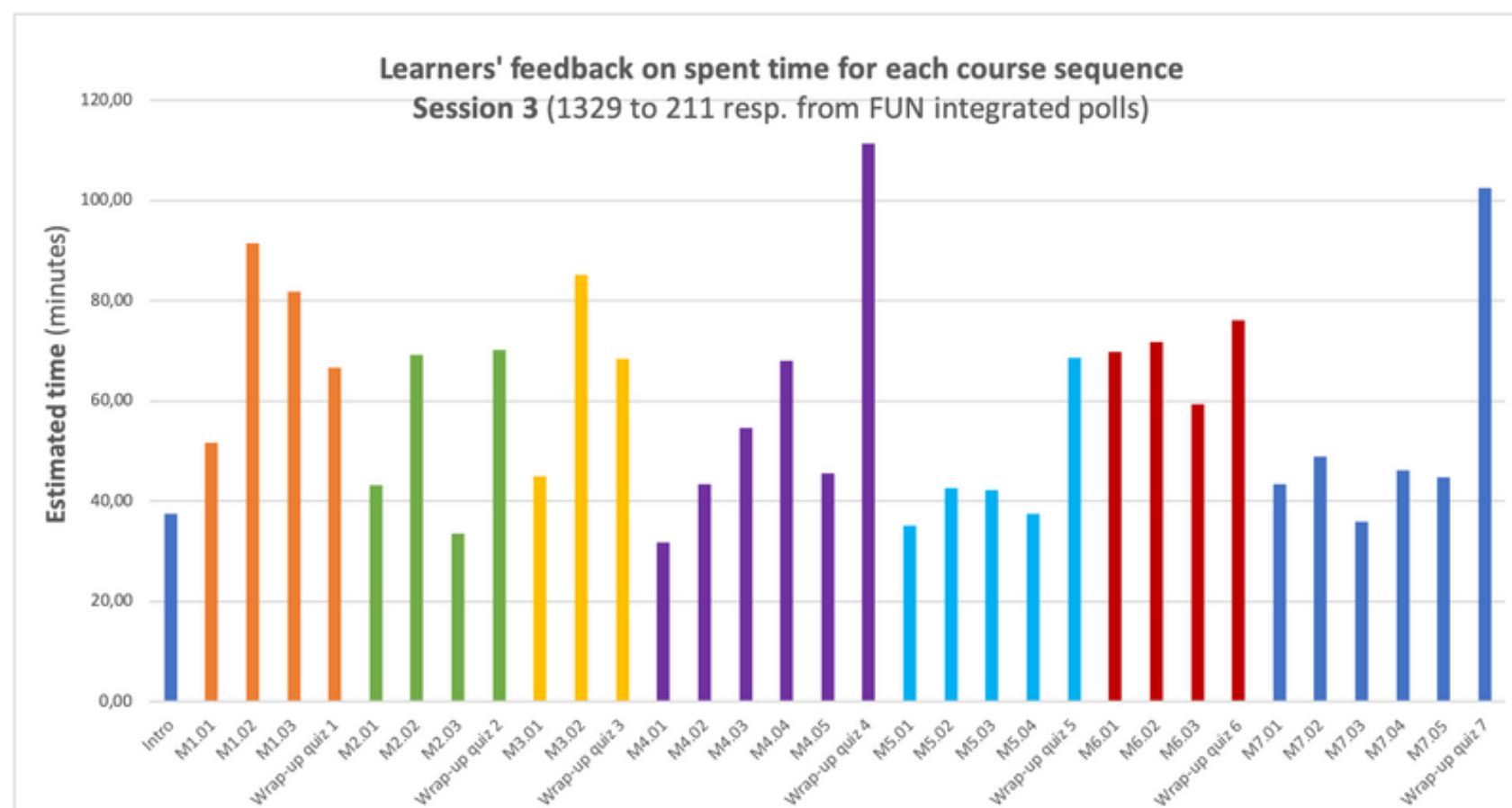
Participant feedback

"As a trainer, I really appreciate the quality of the pedagogical approach and really liked the balance between explanation (notebook oriented) and practice"

Participant feedback

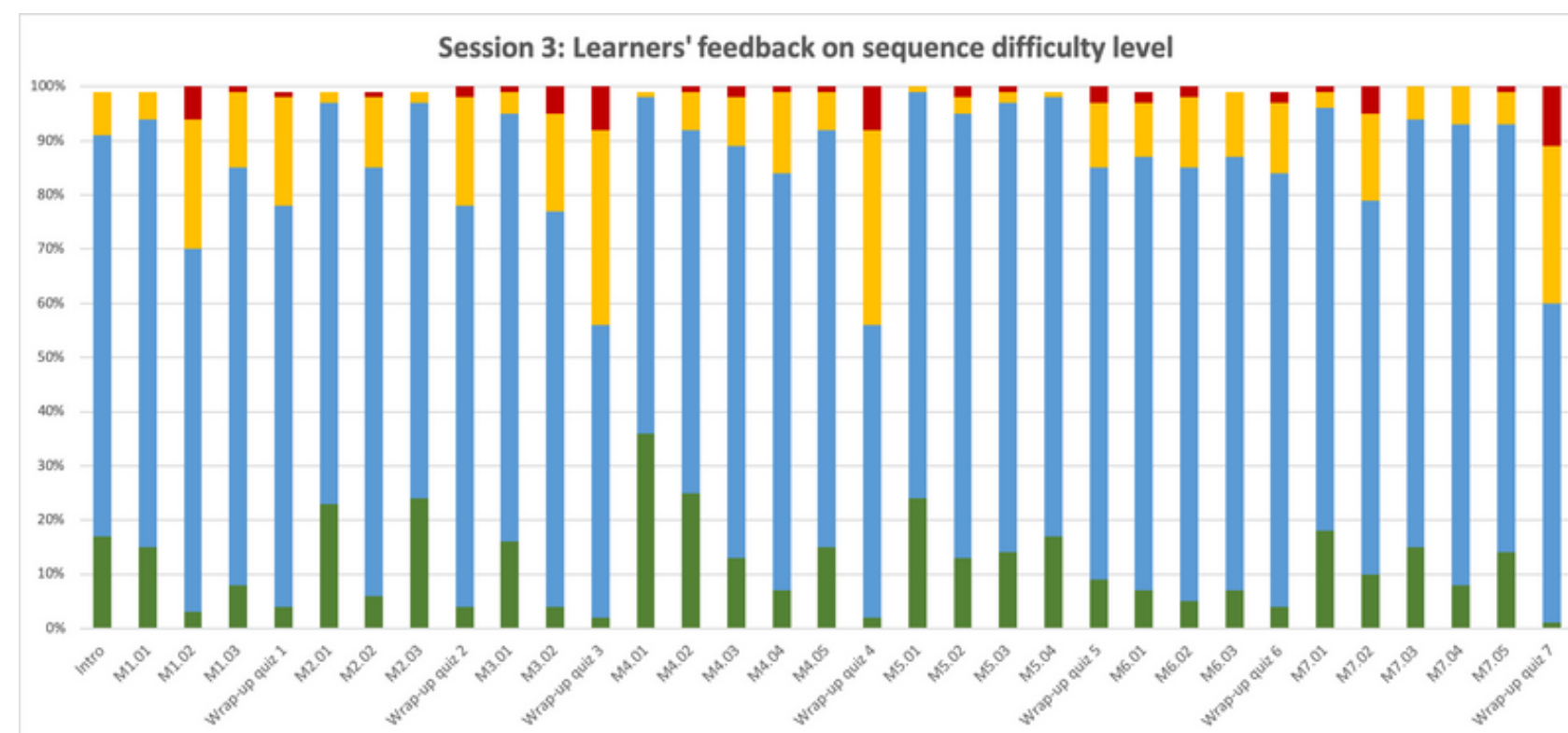


Nous améliorons chaque session grâce aux retours



Créer des équivalences
avec des crédits de
cours

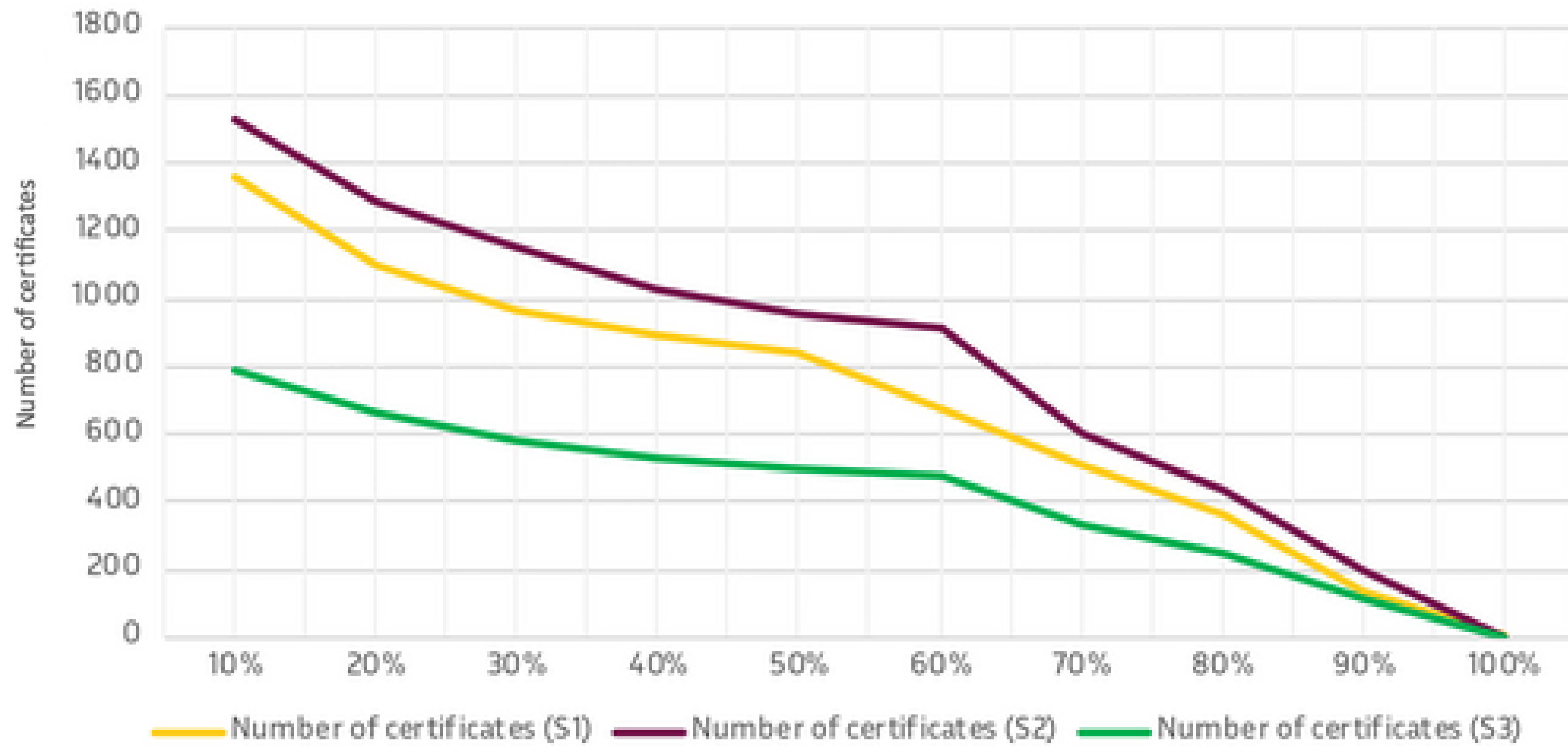
Identifier les concepts ou
les questions difficiles





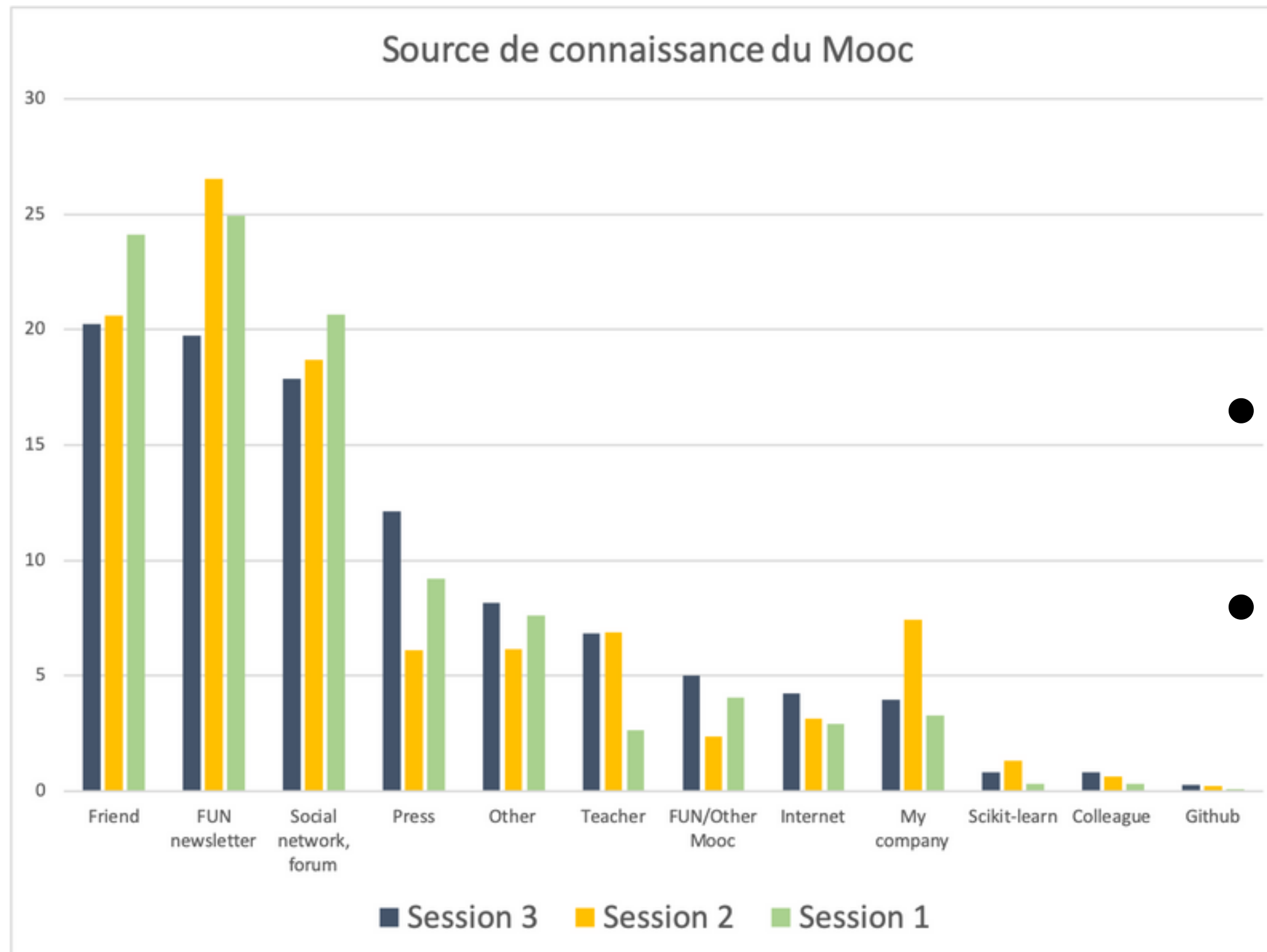
~7% des participants obtiennent une attestation

Number of certificates / passing threshold





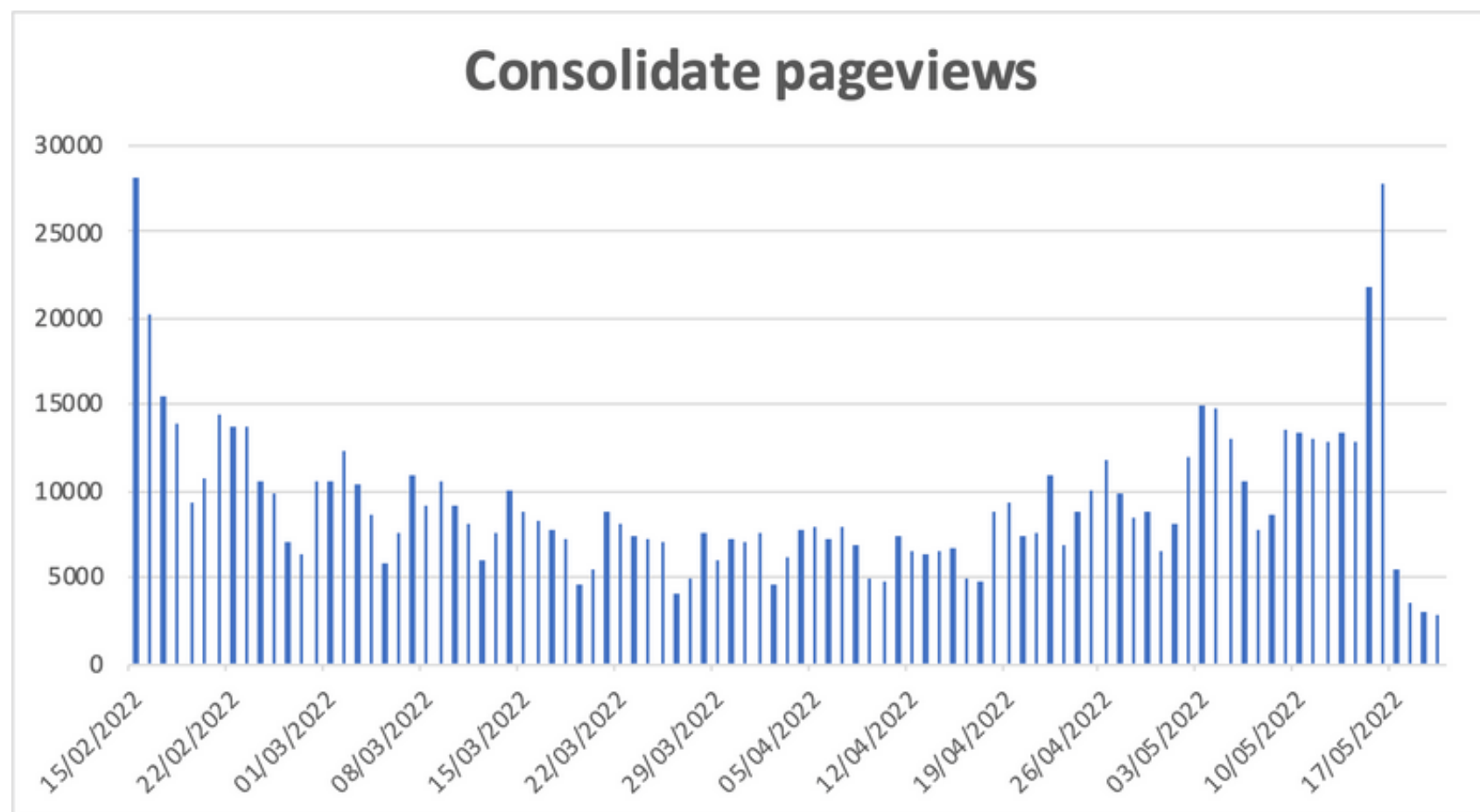
Les canaux de diffusion



- Le canal "Friend" dépasse légèrement le canal "FUN"
- Augmentation de la visibilité dans la presse



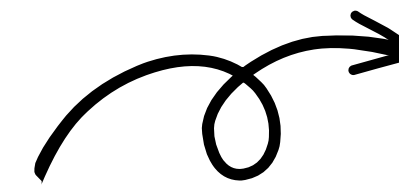
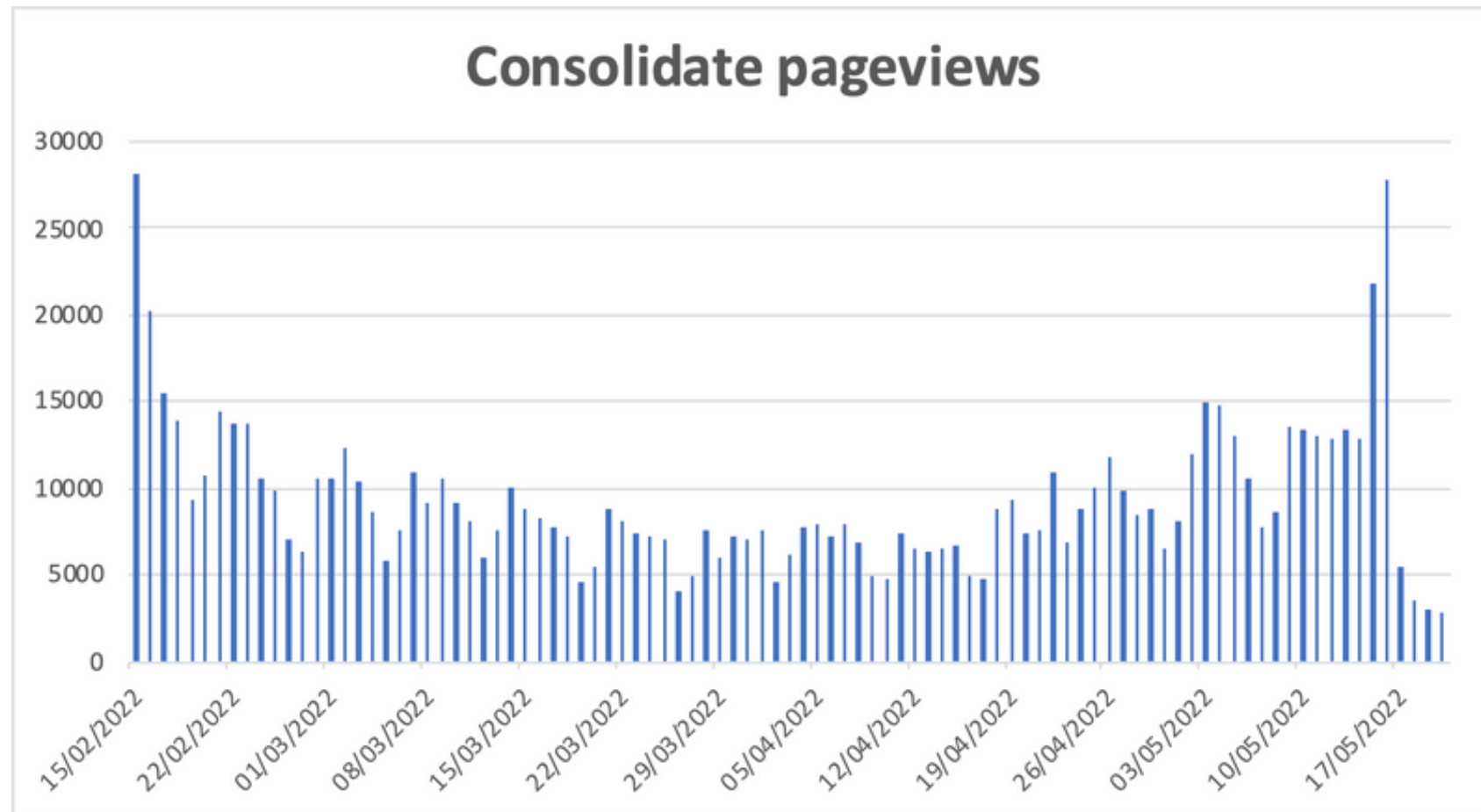
Le forum sur Discourse (Inria Learning Lab Benoit)



▼ Moderator	▼ Flags reviewed	▼ Time reading	▼ Topics created	▼ PMs created	▼ Posts created	▼ Revisions
arturoamorq	—	23h	—	2	249	27
glemaitre58	—	16h	—	—	283	16
ogrisel	—	9h	—	1	97	15
mariecollin	—	5h	1	—	20	5
lesteve	—	4h	—	—	34	13
brospars	—	2h	—	—	39	1
lfarhi	—	2h	—	—	5	3
glemaitre	—	12m	—	—	3	—
gaelvaroquaux	—	4m	—	—	1	—
mhc	—	< 1m	—	—	—	—



Le forum sur Discourse (Inria Learning Lab Benoit)

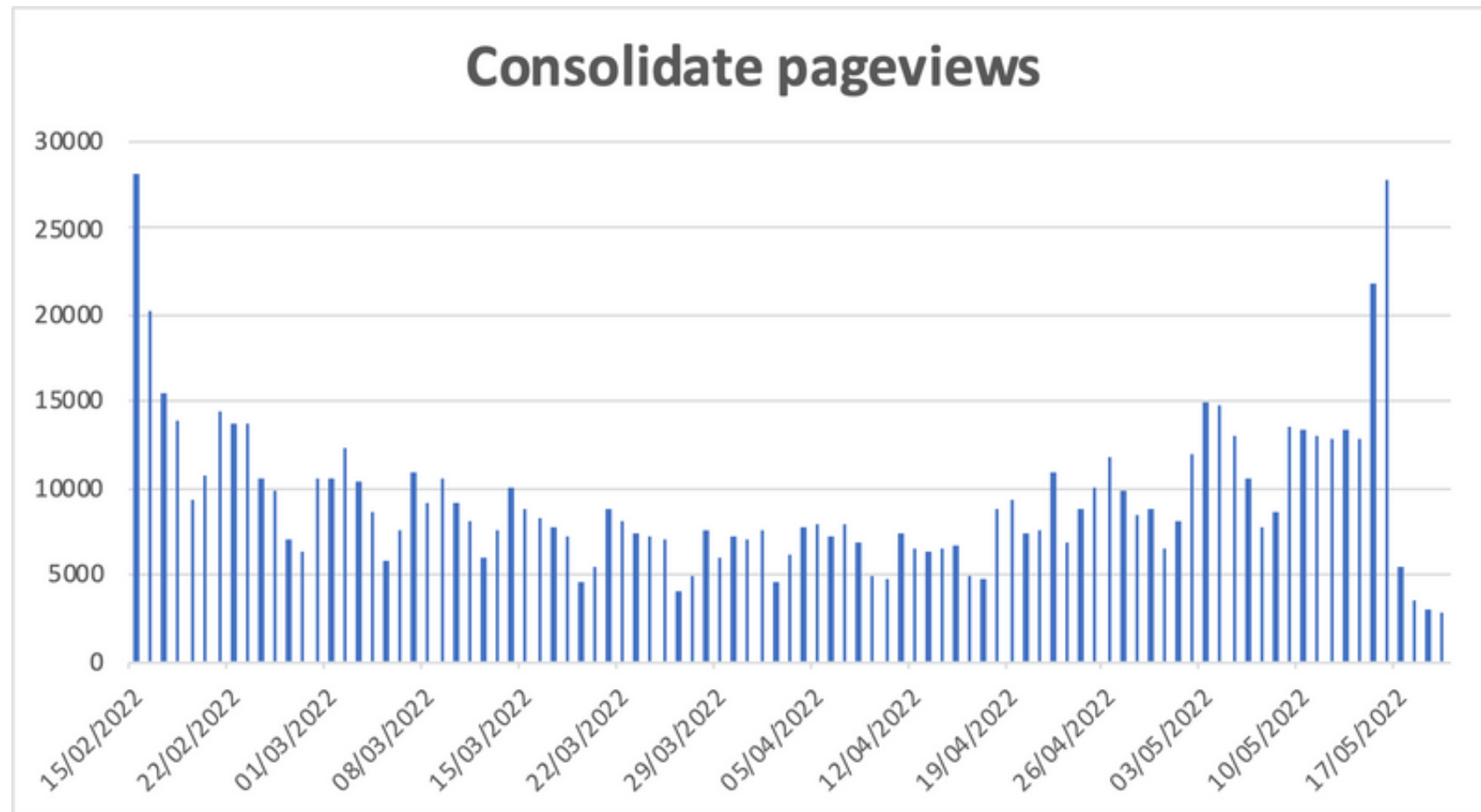


utilisation constante
du forum

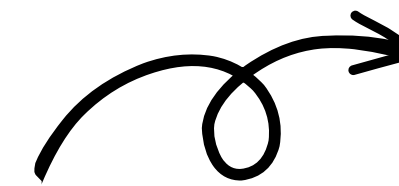
▼ Moderator	▼ Flags reviewed	▼ Time reading	▼ Topics created	▼ PMs created	▼ Posts created	▼ Revisions
arturoamorq	—	23h	—	2	249	27
glemaitre58	—	16h	—	—	283	16
ogrisel	—	9h	—	1	97	15
mariecollin	—	5h	1	—	20	5
lesteve	—	4h	—	—	34	13
brospars	—	2h	—	—	39	1
lfarhi	—	2h	—	—	5	3
glemaitre	—	12m	—	—	3	—
gaelvaroquaux	—	4m	—	—	1	—
mhc	—	< 1m	—	—	—	—



Le forum sur Discourse (Inria Learning Lab Benoit)



soutien constant
du forum



utilisation constante
du forum

Moderator	Flags reviewed	Time reading	Topics created	PMs created	Posts created	Revisions
arturoamorq	—	23h	—	2	249	27
glemaitre58	—	16h	—	—	283	16
ogrisel	—	9h	—	1	97	15
mariecollin	—	5h	1	—	20	5
lesteve	—	4h	—	—	34	13
brospars	—	2h	—	—	39	1
lfarhi	—	2h	—	—	5	3
glemaitre	—	12m	—	—	3	—
gaelvaroquaux	—	4m	—	—	1	—
mhc	—	< 1m	—	—	—	—



Le forum sur Discourse (Inria Learning Lab Benoit)

- ajustement important :
un forum au bas de
chaque page avec un
sujet Discourse associé

[Module 7. Evaluating model performance](#)
[Conclusion](#)
[Appendix](#)

Separate the data and the target

```
In [3]: target_name = "class"
        target = adult_census[target_name]
        target
```

Out[3]: 0 <=50K

STAFF DEBUG INFO

FORUM (EXTERNAL RESOURCE)

☒ Latest ☐ Bookmarks

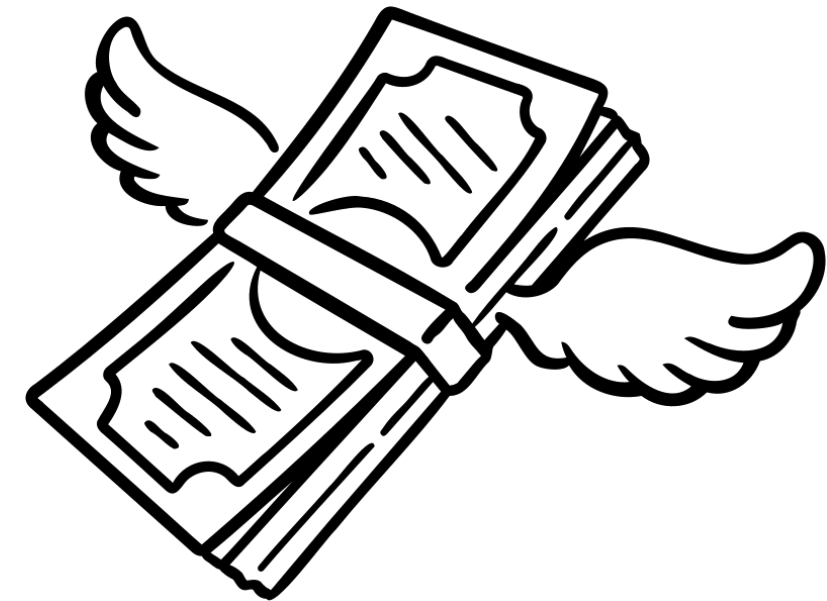
Home > M1. The Predictive Modeling Pipeline > M1. First model with scikit-learn

Topic	Replies	Last reply
About the M1. First model with scikit-learn category M1. First model with scikit-learn brospars Jan '22	1	Jan '22 MarieCollin
<input checked="" type="checkbox"/> Optimistic accuracy M1. First model with scikit-learn pronod Nov '22	1	Nov '22 ArturoAmorQ
<input checked="" type="checkbox"/> Fit method and predict function M1. First model with scikit-learn kinuthia Oct '22	2	Nov '22 kinuthia



JupyterHub pour l'environnement live (Inria Learning Lab Benoit)

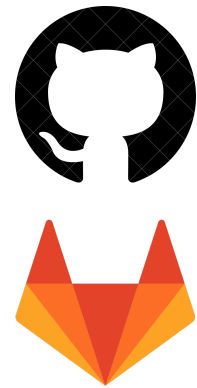
- on-premises
- OVH credits
 - 600€/mois pour session
(10 serveurs de 13 pers max)
 - 270€/mois pour le SPOC
(3 serveurs de 13 pers max)



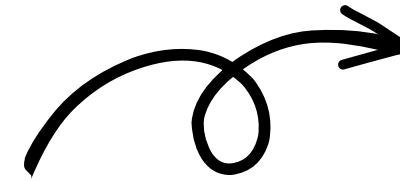
0,2 à 7€/personne/mois



Esprit de collaboration



github + CI setup publique
gitlab privé for quizzes



GHA + Netlify
(Loïc)

The screenshot shows the GitHub repository page for INRIA/scikit-learn-mooc. At the top, the repository name is displayed in large bold letters, followed by the INRIA logo. Below this, a subtitle reads 'Machine learning in Python with scikit-learn MOOC'. A row of statistics is shown: 23 Contributors, 44 Issues, 864 Stars, and 450 Forks. A GitHub logo is also present. Below the statistics is a horizontal bar with an orange segment on the left and a blue segment on the right. The description section follows, stating the repository's purpose and providing instructions on how to contribute. A GitHub logo is at the bottom left of the description.

INRIA/scikit-learn-mooc

Machine learning in Python with scikit-learn MOOC

23 Contributors 44 Issues 864 Stars 450 Forks

INRIA/scikit-learn-mooc: Machine learning in Python with scikit-learn MOOC

Machine learning in Python with scikit-learn MOOC. Contribute to INRIA/scikit-learn-mooc development by creating an account on GitHub.

GitHub

☆ 858 stars
👁 36 watching
🔗 449 forks



Scripts personnalisés (Inria Loïc)

- générer des questions de quiz sur github à partir de solutions de quiz sur gitlab privé
- générer l'exercice à partir de l'exercice + solution sur github repo
- jupytertext pour générer .ipynb à partir de .py



Intégration sur FUN (Inria Learning Lab Laurence & Marie)

- JupyterBook comme source de vérité
- la plupart des pages du MOOC réutilisent directement le contenu du repo : JupyterBook HTML ou notebooks du repo
- gestion des quizzes manuelle

The screenshot displays the FUN MOOC interface. On the left, a sidebar lists course modules: Welcome, Introduction: Machine Learning concepts, Module 1. The Predictive Modeling Pipeline, Module 2. Selecting the best model, Module 3. Hyperparameter tuning, Module 4. Linear Models, and Module 5. Decision tree. The main content area shows the 'Preprocessing for numerical features' notebook. A dropdown menu is open over the notebook title, showing options: 'Preprocessing for numerical features', 'Archive notebook', and 'Forum'. The notebook content includes a Jupyter cell with the code `data_train.describe()` and its output, a table of statistics for the training data. Below the table, there is a text explanation and a 'Tip' box about scaling features.

Module overview
due Jan 15, 2023 at 22:00 UTC

Tabular data exploration
Quiz M1 due Jan 15, 2023 at 22:00 UTC

Fitting a scikit-learn model on numerical data
Quiz M1 due Jan 15, 2023 at 22:00 UTC

Handling categorical data
Quiz M1 due Jan 15, 2023 at 22:00 UTC

Wrap-up quiz
Wrap-up quiz due Jan 15, 2023 at 22:00 UTC

Main take-away
due Jan 15, 2023 at 22:00 UTC

Preprocessing for numerical features

This notebook is read only

Let's start by printing some statistics about the training data.

```
In [6]: data_train.describe()
```

Out[6]:

	age	capital-gain	capital-loss	hours-per-week
count	36631.000000	36631.000000	36631.000000	36631.000000
mean	38.642352	1087.077721	89.665311	40.431247
std	13.725748	7522.692939	407.110175	12.423952
min	17.000000	0.000000	0.000000	1.000000
25%	28.000000	0.000000	0.000000	40.000000
50%	37.000000	0.000000	0.000000	40.000000
75%	48.000000	0.000000	0.000000	45.000000
max	90.000000	99999.000000	4356.000000	99.000000

We see that the dataset's features span across different ranges. Some algorithms make some assumptions regarding the feature distributions and usually normalizing features will be helpful to address these assumptions.

Tip

Here are some reasons for scaling features:

- Models that rely on the distance between a pair of samples, for instance k-nearest neighbors, should be trained on normalized features to make each feature contribute approximately equally to the distance



La prochaine session sera ouverte en mode self-paced

- à partir d'octobre 2023
- petites corrections et moins fréquentes
- support du forum une fois par semaine
- négociations avec OVH



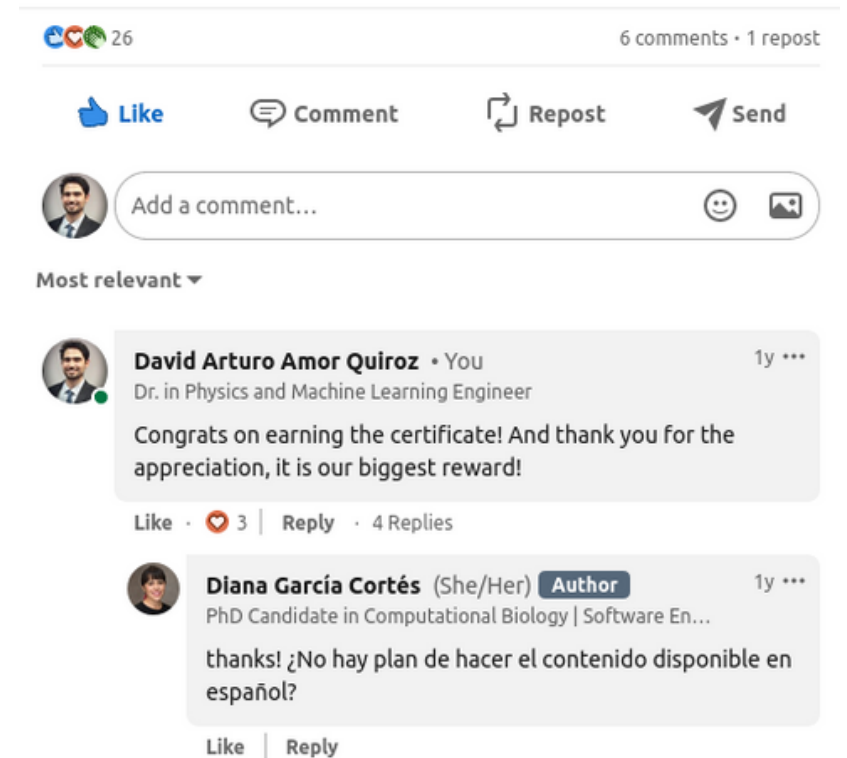
Wish list

- supporter Pyodide + Jupyterlite
- synchronisation automatique des quizzes
- moteur de recherche interne
- badges par module ?



Wish list

- supporter Pyodide + Jupyterlite
- synchronisation automatique des quizzes
- moteur de recherche interne
- badges par module ?
- élargir la communauté des utilisateurs

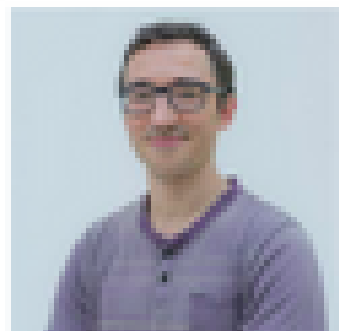


THANK YOU!

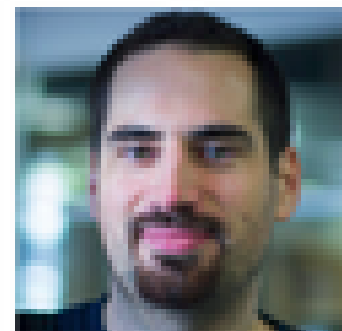
👤 Teachers



Arturo Amor is an engineer at Inria. He is in charge of broadening the scikit-learn documentation's accessibility to all kind of users.



Loïc Estève has been a scikit-learn core developer at Inria since 2016. He has been involved in projects at the heart of the scientific Python ecosystem, such as Scikit-learn, Dask and joblib.



Olivier Grisel has been a core contributor to scikit-learn since 2010. He is an expert in machine learning, teaching deep learning at École Polytechnique and Université Paris Saclay.



Guillaume Lemaître is a research engineer. He has a PhD in computer science and has been a scikit-learn core developer since 2017.



Thomas Schmitt is a machine Learning Engineer at Inria Saclay (Parietal team).



Gaël Varoquaux is a Research Director at Inria. He used Scikit-learn in his research and was active in coordinating the community of developers. In 2018, he became project manager for the Scikit-learn consortium.

👤 The pedagogical and technical team

Inria LearningLab



Aurélie Lagarrigue
Educational Engineer
Inria Learning Lab



Laurence Farhi
Educational Engineer
Inria Learning Lab



Marie Collin
Educational Engineer
Inria Learning Lab



Benoit Raspars
IT engineer
Inria Learning Lab



Madeline MONTIGNY
Educational Engineer
Inria - Service Éducation et Médiation Scientifiques