Introduction de nouveaux outils (learnr, bookdown, machine virtuelle, Github Classroom, . . .) dans un cours de Science des Données Biologiques

Guyliann Engels a et Philippe Grosjean a

a Laboratoire d'Écologie numérique des Milieux aquatiques

A-Institut de Recherche Complexys

A-8 avenue du Champ de Mars, 7000 Mons, Belgique

`guyliann.engels@umons.ac.be`

`philippe.grosjean@umons.ac.be`

Mots clefs : science des données, biologie, apprentissage, classe renversée.

Les statistiques classiques sont de plus en plus insuffisantes dans le contexte actuel : masse de données en croissance exponentielle, nécessitant des outils spécialisés pour les aborder, crise de la reproductibilité en science [1], Open Science, Open Data, Open Knowledge (Directives européennes). Aujourd'hui, et encore plus demain, nos étudiants devront pouvoir maitriser les outils adéquats dans ce nouveau contexte. Cela dépasse les statistiques, dans un ensemble plus large nommé Science des Données [2].

De plus, les outils conventionnels d'apprentissage que sont les livres de référence, les syllabus avec des professeurs omniscients déversant leurs savoirs n'intéressent plus les étudiants. Les étudiants sont toujours plus connectés et cherchent l'information sur Internet via Google, Youtube,. . .

Dans ce contexte, nous présentons ici un exemple de cursus complètement remanié en Science des Données Biologiques à l'Université de Mons, en Belgique (http://biodatascience-course.sciviews.org). Celui-ci s'étale sur 4 années. La partie pratique et interactive y est prépondérante. Elle se base sur R, R Markdown, RStudio, Git, Github Classroom, bookdown, blogdown, learnr, et se focalise sur le développement de bonnes pratiques en science reproductible [3].

Nous avons fait le choix d'employer une machine virtuelle sous VirtualBox complètement configurée avec R Studio Server et Jupyter comme interfaces, appelée "SciViews Box" (http://www.sciviews.org/blog/The-SciViews-Box/) pour nos enseignements. Cette machine virtuelle simplifie la configuration de l'ensemble des outils logiciels que les étudiants manipulent, et permet un travail reproductible sans se soucier des problèmes d'installation, de mises à jour, de compatibilités,. . . Elle est accompagnée d'un installeur simplissime sous Windows ou MacOS. Elle est actualisée chaque année afin d'y ajouter les nouveautés proposées par R, RStudio, les nouveaux packages intéressants. L'expérience montre qu'elle simplifie grandement l'approche de R par nos étudiants qui continuent encore à l'utiliser au delà de leur formation.

L'apprentissage du language R n'est pas aisé pour des non-informaticien légèrement déstabilisé lorsque l'on prononce la mot : "code". La mise à dispositions d'une boite à outils comprennant une succession de menus (`snippet RStudio`) dans la SciViews Box leurs facilite grandement la tâche. Par exemple, les snippets dans le menu "dataframes" va contenir des templates d'instructions permettant de réaliser du remaniement des données (sélection des colonnes, filtre sur les lignes, calcul de nouvelles variables,. . .).

Le nouveau matériel pédagogique se compose d'un ouvrage en ligne de type bookdown (http://biodatascience-course.sciviews.org/sdd-umons/) renvoyant vers des tutoriaux vidéo (http://go.sciviews.org/BioDataScience-videos) et des documents interactifs learnr (https://github.com/BioDataScience-Course/BioDataScience). Les étudiants utilisent des projets RStudio et des R Notebooks pour rédiger les rapports de leurs analyses de manière reproductible. Enfin, nous prévoyons d'utiliser également Github Classroom (https://github.com/BioDataScience-Course) dès l'an prochain.

Références

[1] Baker, M. 2016. "1,500 Scientists Lift the Lid on Reproducibility." Nature 533 (7604): 452–54. doi: 10.1038/533452a.

[2] Cleveland, W.S. 2001. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." ISI Review 69: 21–26. doi:10.1111/j.1751-5823.2001.tb00477.x.

[3] Leek, Jeffrey T. & Peng, Roger D. 2015. "Opinion: Reproducible research can still be wrong: Adopting a prevention approach". Proceedings of the National Academy of Sciences: 112-6.