

Deux MOOCs sur la recherche reproductible

A. Legrand, K. Hinsén

22 octobre 2024

① Recherche reproductible : principes méthodologiques pour une science transparente

- depuis 2018
- bilingue français/anglais
- “grand public” : s’adresse à tout le monde qui travaille avec des données et/ou du code
- licence CC-BY

- ① **Recherche reproductible : principes méthodologiques pour une science transparente**
 - depuis 2018
 - bilingue français/anglais
 - “grand public” : s’adresse à tout le monde qui travaille avec des données et/ou du code
 - licence CC-BY
- ② **Reproducible Research II: Practices and tools for managing computations and data**
 - 1ère session : mai à septembre 2024
 - en anglais
 - très technique : calcul scientifique à l’échelle
 - licence CC-BY-NC-SA

Auteurs

- Arnaud Legrand (informatique)
- Christophe Pouzat (mathématique)
- Konrad Hinsén (physique)

Deux MOOCs sur FUN

Auteurs

- Arnaud Legrand (informatique)
- Christophe Pouzat (mathématique)
- Konrad Hinsén (physique)
- Kim Tâm Huynh
- Ludovic Courtès
- Matthieu Simonin

Deux MOOCs sur FUN

Auteurs

- Arnaud Legrand (informatique)
- Christophe Pouzat (mathématique)
- Konrad Hinsén (physique)
- Kim Tâm Huynh
- Ludovic Courtès
- Matthieu Simonin

Inria Learning Lab

- Deux ingénieurs pédagogiques à chaque instant :
Aurélié Bayle, Marie-Hélène Comte, *Laurence Farhi*,
Tatiana Khomenko, *Madeline Montigny*
- Un informaticien : Benoît Rospars

Introduction

- interviews avec des chercheurs de disciplines différentes

Introduction

1. Cahier de notes, cahier de laboratoire

- prise de note structurée : Markdown, Pandoc
- outils d'indexation : DocFetcher, ExifTool
- gestion de versions : GitLab, Git avec des petits dessins
- exercices :
 - rédiger un document avec une certaine spécification
 - trouver des informations sur qui a fait quoi dans un dépôt git
 - ...

Introduction

1. Cahier de notes, cahier de laboratoire

2. La vitrine et l'envers du décor : le document computationnel

- principes du document computationnel
- principes de l'analyse répliquable
- trois outils, trois parcours :
 - Jupyter et Python (en ligne, dans le MOOC) : 63%
 - RStudio et R : 31%
 - Emacs/Org-mode et Python + R : 7%
- exercices : utilisation des notebooks, visualisation, analyse de données (navette "Challenger")

Introduction

1. Cahier de notes, cahier de laboratoire

2. La vitrine et l'envers du décor : le document computationnel

3. La main à la pâte : une analyse répliquable

- rédaction de documents computationnels
- exemple : données du Réseau Sentinelles sur l'incidence des infections grippales
- travail pratique : rédiger une analyse répliquable (7 sujets)

Contenu MOOC 1

Introduction

1. Cahier de notes, cahier de laboratoire

2. La vitrine et l'envers du décor : le document computationnel

3. La main à la pâte : une analyse répliquable

4. Vers une étude reproductible : la réalité du terrain

- l'enfer des données : taille, formats, pérennité
- l'enfer des logiciels : environnements logiciel
- l'enfer du calcul : flottants, nombres aléatoires

1ère session : 2018

- durée fixe : 2 mois
- 3588 inscrits, 291 attestations (8%)

1ère session : 2018

2ème session : 2019

- durée fixe : 2 mois
- matériel supplémentaire visant les SHS
- traduction du MOOC entier en anglais
- 2192 inscrits, 135 attestations (6%)

Les sessions

1ère session : 2018

2ème session : 2019

3ème session : depuis mars 2020

- encore ouverte
- légère revision des exercices
- début 2024 : 17168 inscrits, 1938 badges/attestations (12%)

Qui sont nos apprenants ?

- 86% en France, 78% francophones

Qui sont nos apprenants ?

- 86% en France, 78% francophones
- 66% entre 19 et 35 ans

Qui sont nos apprenants ?

- 86% en France, 78% francophones
- 66% entre 19 et 35 ans
- 46% femmes, 54% hommes

Qui sont nos apprenants ?

- 86% en France, 78% francophones
- 66% entre 19 et 35 ans
- 46% femmes, 54% hommes
- 50% niveau mastère, 41% niveau doctorat

Qui sont nos apprenants ?

- 86% en France, 78% francophones
- 66% entre 19 et 35 ans
- 46% femmes, 54% hommes
- 50% niveau mastère, 41% niveau doctorat
- 54% doctorants et post-docs

Qui sont nos apprenants ?

- 86% en France, 78% francophones
- 66% entre 19 et 35 ans
- 46% femmes, 54% hommes
- 50% niveau mastère, 41% niveau doctorat
- 54% doctorants et post-docs
- 12% étudiants (dont 42% reçoivent des crédits ECTS)

Qui sont nos apprenants ?

- 86% en France, 78% francophones
- 66% entre 19 et 35 ans
- 46% femmes, 54% hommes
- 50% niveau mastère, 41% niveau doctorat
- 54% doctorants et post-docs
- 12% étudiants (dont 42% reçoivent des crédits ECTS)
- 29% sciences du vivant,
23% informatique,
22% physique/ingénieurs,
10% SHS,
9% mathématiques

Qui sont nos apprenants ?

- 86% en France, 78% francophones
- 66% entre 19 et 35 ans
- 46% femmes, 54% hommes
- 50% niveau mastère, 41% niveau doctorat
- 54% doctorants et post-docs
- 12% étudiants (dont 42% reçoivent des crédits ECTS)
- 29% sciences du vivant,
23% informatique,
22% physique/ingénieurs,
10% SHS,
9% mathématiques
- 72% Windows, 34% Linux, 17% macOS

Satisfaction

50% complètement satisfaits, 46% plutôt satisfaits

Satisfaction

50% complètement satisfaits, 46% plutôt satisfaits

Points de critique

- trop difficile pour les participants sans base solide en informatique
- trop long

Satisfaction

50% complètement satisfaits, 46% plutôt satisfaits

Points de critique

- trop difficile pour les participants sans base solide en informatique
- trop long

Forum d'échange

- outil d'entraide
- plus efficace pour une session courte

- Le réseau Sentinelles change son format de données

- Le réseau Sentinelles change son format de données
- Notre environnement conda devient irreproductible

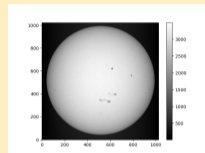
4. Vers une étude reproductible : la réalité du terrain

- l'enfer des données : taille, formats, pérennité
- l'enfer des logiciels : environnements logiciel
- l'enfer du calcul : flottants, nombres aléatoires

Reproducible Research II: Practices and tools for managing computations and data

Introduction

- “fil rouge” : détection de taches solaires
- interviews avec des astrophysiciens
- trois notebooks pour mieux comprendre les données



Introduction

1. Managing data

- Archiving : Zenodo, Software Heritage
- File formats : CSV, JSON, FITS, HDF5
- Project organization
- Versioning : git-annex

Introduction

1. Managing data

2. Managing software

- On the Importance of Software Environments
- Package Management Principles : Debian
- Isolation and Containers : Docker
- Using Containers
- Building and Sharing Containers
- Functional Package Managers : Guix

Introduction

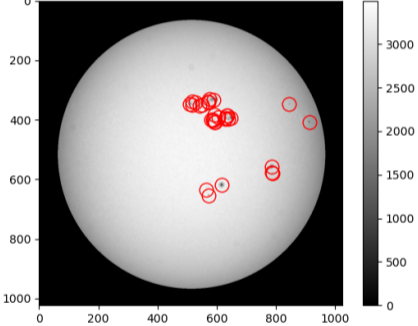
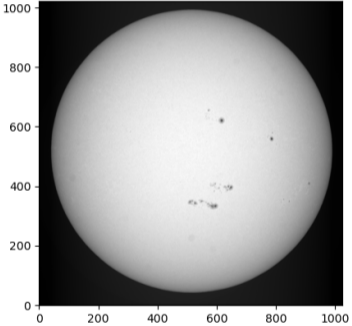
1. Managing data

2. Managing software

3. Managing computations

- Why do we need workflows?
- From notebooks to shell scripts
- Workflows with make
- Workflows with snakemake
- Workflows and environments

Le projet "fil rouge" : détection de taches solaires



Du notebook au workflow

- point de départ : notebook (Jupyter/Python) pour la détection
- transformation en deux tâches d'un workflow

Du notebook au workflow

- point de départ : notebook (Jupyter/Python) pour la détection
- transformation en deux tâches d'un workflow
- tâche de niveau supérieur : série chronologique des moyennes mensuelles entre 2009 et 2020
- gestion des erreurs

Du notebook au workflow

- point de départ : notebook (Jupyter/Python) pour la détection
- transformation en deux tâches d'un workflow
- tâche de niveau supérieur : série chronologique des moyennes mensuelles entre 2009 et 2020
- gestion des erreurs

Nous fournissons le workflow complet.

Les exercices portent sur la compréhension et la modification.

- environ 50.000 images de l'Observatoire Royal de Belgique
- qualité variable au cours des années
- des images non exploitables (éclipses, ...)
- sous-ensemble annexé au workflow par `git annex`

- construit avec Guix
- exporté comme image Docker/Singularity (environ 1 Go)
- instructions pour les trois technologies de conteneurs (Guix, Docker, Singularity)

- outils complexes dont nous avons découvert les détails en route

- outils complexes dont nous avons découvert les détails en route
- environnements imbriqués

- outils complexes dont nous avons découvert les détails en route
- environnements imbriqués
- notre environnement Guix n'est pas compilable pour ARM64

Qui sont nos apprenants ? (68 réponses)

- 1976 inscrits, 298 actifs, 40 éligibles au badge

Qui sont nos apprenants ? (68 réponses)

- 1976 inscrits, 298 actifs, 40 éligibles au badge
- 56% en France

Qui sont nos apprenants ? (68 réponses)

- 1976 inscrits, 298 actifs, 40 éligibles au badge
- 56% en France
- tranches d'âges : $\frac{1}{3}$ 25 à 35, $\frac{1}{3}$ 36 à 45, $\frac{1}{3}$ 46 à 55

Qui sont nos apprenants ? (68 réponses)

- 1976 inscrits, 298 actifs, 40 éligibles au badge
- 56% en France
- tranches d'âges : $\frac{1}{3}$ 25 à 35, $\frac{1}{3}$ 36 à 45, $\frac{1}{3}$ 46 à 55
- 33% femmes, 66% hommes

Qui sont nos apprenants ? (68 réponses)

- 1976 inscrits, 298 actifs, 40 éligibles au badge
- 56% en France
- tranches d'âges : $\frac{1}{3}$ 25 à 35, $\frac{1}{3}$ 36 à 45, $\frac{1}{3}$ 46 à 55
- 33% femmes, 66% hommes
- 48% niveau mastère, 45% niveau doctorat

Qui sont nos apprenants ? (68 réponses)

- 1976 inscrits, 298 actifs, 40 éligibles au badge
- 56% en France
- tranches d'âges : $\frac{1}{3}$ 25 à 35, $\frac{1}{3}$ 36 à 45, $\frac{1}{3}$ 46 à 55
- 33% femmes, 66% hommes
- 48% niveau mastère, 45% niveau doctorat
- 38% sciences du vivant,
32% informatique,
24% physique/ingénieurs

Qui sont nos apprenants ? (68 réponses)

- 1976 inscrits, 298 actifs, 40 éligibles au badge
- 56% en France
- tranches d'âges : $\frac{1}{3}$ 25 à 35, $\frac{1}{3}$ 36 à 45, $\frac{1}{3}$ 46 à 55
- 33% femmes, 66% hommes
- 48% niveau mastère, 45% niveau doctorat
- 38% sciences du vivant,
32% informatique,
24% physique/ingénieurs
- 69% Linux, 54% Windows, 12% macOS

Qui sont nos apprenants ? (68 réponses)

- 1976 inscrits, 298 actifs, 40 éligibles au badge
- 56% en France
- tranches d'âges : $\frac{1}{3}$ 25 à 35, $\frac{1}{3}$ 36 à 45, $\frac{1}{3}$ 46 à 55
- 33% femmes, 66% hommes
- 48% niveau mastère, 45% niveau doctorat
- 38% sciences du vivant,
32% informatique,
24% physique/ingénieurs
- 69% Linux, 54% Windows, 12% macOS
- 54% “plutôt satisfaits”, 39% “complètement satisfaits”

Pour la prochaine session...

- débogage à fond
- exercices révisés, simplifiés
- git-annex et Guix auront leurs propres modules