Deux MOOCs sur la recherche reproductible

Konrad Hinsen

Centre de Biophysique Moléculaire, Orléans, France
Synchrotron SOLEIL, Saint Aubin, Frane

28 mars 2024

Deux MOOCs sur FUN

1. Recherche reproductible : principes méthodologiques pour une science transparente
   - depuis 2018
   - bilingue francais/anglais
   - “grand public” : s’adresse à tout le monde qui travaille avec des données et/ou du code
   - licence CC-BY

Deux MOOCs sur FUN

1. Recherche reproductible : principes méthodologiques pour une science transparente
   - depuis 2018
   - bilingue francais/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
   - en test, ouverture prévue pour mai 2024
   - en anglais
   - très technique : calcul scientifique à l’échelle
   - licence CC-BY-NC-SA

Deux MOOCs sur FUN

Auteurs

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

Deux MOOCs sur FUN

Auteurs

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

Deux MOOCs sur FUN

Auteurs

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

Inria Learning Lab

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

Contenu MOOC 1

Introduction

- interviews avec des chercheurs de disciplines différentes

Contenu MOOC 1

Introduction

1. Cahier de notes, cahier de laboratoire

- prise de note structurée : Markdown, Pandoc
- outils d'indexation : DocFetcher, ExifTool
- gestion de versions : GitLab

Contenu MOOC 1

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 réplicable
- trois outils, trois parcours :
  - Jupyter et Python (en ligne, dans le MOOC) : 63%
  - RStudio et R : 31%
  - Emacs/Org-mode et Python + R : 7%

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 réplicable

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

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

Les sessions

1ère session : 2018

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

Les sessions

1ère session : 2018

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

2ème session : 2019

- durée fixe : 2 mois
- matériel supplémentaire visant les SHS
- 2192 inscrits, 135 attestations (6%)

Les sessions

1ère session : 2018

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

2ème session : 2019

- durée fixe : 2 mois
- matériel supplémentaire visant les SHS
- 2192 inscrits, 135 attestations (6%)

3ème session : depuis mars 2020

- encore ouverte
- legère 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% recoivent 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% recoivent 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% recoivent des crédits ECTS)
- 29% sciences du vivant,
  23% informatique,
  22% physique/ingénieurs,
  10% SHS,
  9% mathématiques
- 72% Windows, 34% Linux, 17% macOS

Retours

Satisfaction

50% complètement satisfaits, 46% plutôt satisfaits

Retours

Satisfaction

50% complètement satisfaits, 46% plutôt satisfaits

Points de critique

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

Défis techniques

- Le réseau Sentinelles change son format de données

Défis techniques

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

Sortir des trois enfers

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

Contenu MOOC 2

Introduction

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

Contenu MOOC 2

Introduction

1. Managing data

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

Contenu MOOC 2

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

Contenu MOOC 2

Introduction

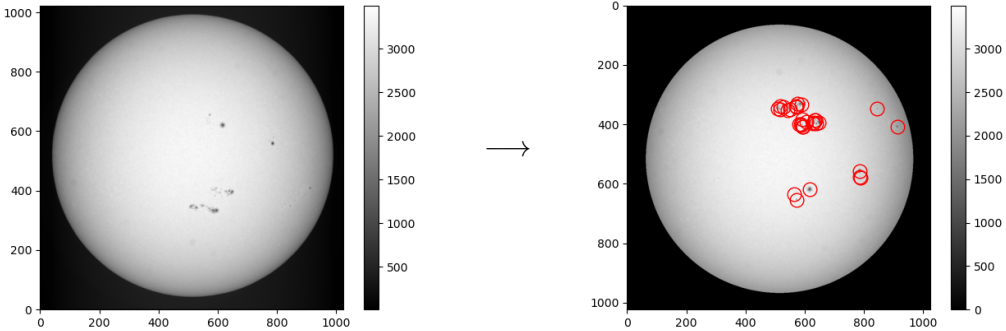
1. Managing data

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 moyens mensuels 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 moyens mensuels entre 2009 et 2020
- Gestion des erreurs

Nous fournissons le workflow complet.
Les exercices portent sur la compréhension et la modification.

Gestion des données

- 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`

Environnement logiciel

- Construit avec Guix
- Exporté comme image Docker/Singularity (environ 1 Go)
- Instructions pour les trois technologies de conteneurs (Guix, Docker, Singularity)

Défis techniques

- Outils complexes dont nous avons découvert les détails en route

Défis techniques

- Outils complexes dont nous avons découvert les détails en route
- Environnements imbriqués

Défis techniques

- Outils complexes dont nous avons découvert les détails en route
- Environnements imbriqués
- Notre environnement Guix n’est pas compilable pour ARM64