Reproductibilité(s) et expérimentation(s) numérique(s)

Séminaire recherche reproductible - 2ème session

Pierre-Antoine Bouttier, UAR GRICAD, CNRS

Préambule

- À qui je m'adresse ?
  - Au plus grand nombre
  - Grandes hétérogénéités (de public, de pratiques, d'outils, de niveaux de compétences, de culture numérique, *etc.*)
- Ce que je ne présenterai pas
  - Des outils en détail
  - Des méthodes qui garantissent la reproductibilité dans le contexte numérique

Quelques cas d'usages transverses

- Statistiques sur une enquête
- Nettoyage, normalisation, etc. de données brutes de mesures
- Simulations numériques
- Calcul de quantités résumantes (*e.g.* stats, courbes), visualisation
- ...
- *Points communs de ces **expérimentations numériques** : données numériques & code(s) logiciel(s)*

La reproductibilité dans le cadre numérique

Merci à Konrad Hinsen pour sa présentation

Intérêt de la reproductibilité

Reproductibilité : preuve de **rigueur** qui inspire **confiance**

- Ce qu'un résultat non-reproductible suggère :
  - Une description de la méthodologie incomplète
  - une maîtrise insuffisante des techniques
  - une erreur
  - une fraude
- L'importance de la confiance
  - pour vous-même
  - pour les sciences (résultats solides et donc féconds)
  - pour l'ensemble de la société

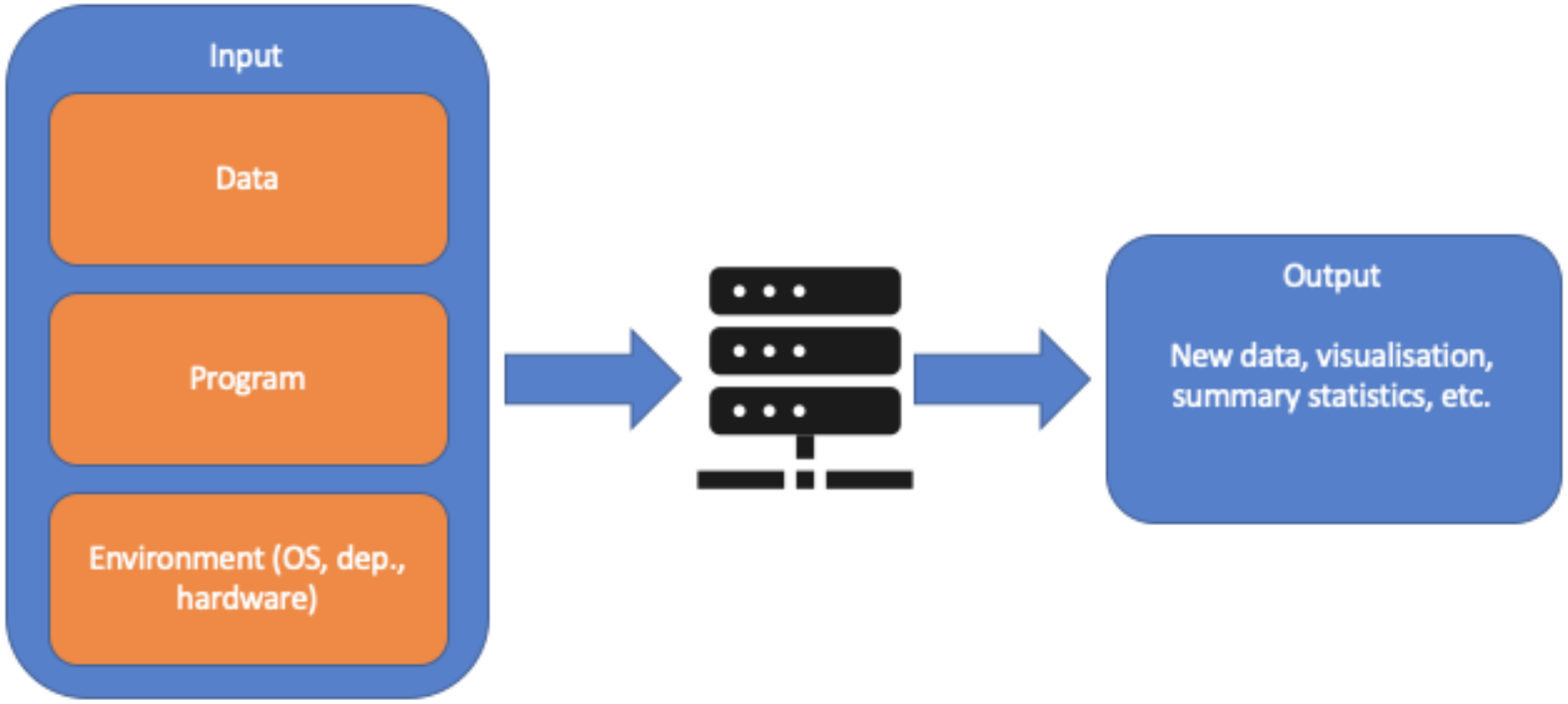
Les reproductibilités

- Reproductibilité **expérimentale**
  - Refaire une expérience d'après la description publiée
  - Obtenir des résultats suffisamment proches
- Reproductibilité **statistique**
  - Refaire une étude avec un autre échantillon ou une autre technique
  - Inférer des conclusions suffisamment proches
- Reproductibilité **computationnelle**
  - Refaire un calcul à l'identique
  - Obtenir des résultats à l'identique

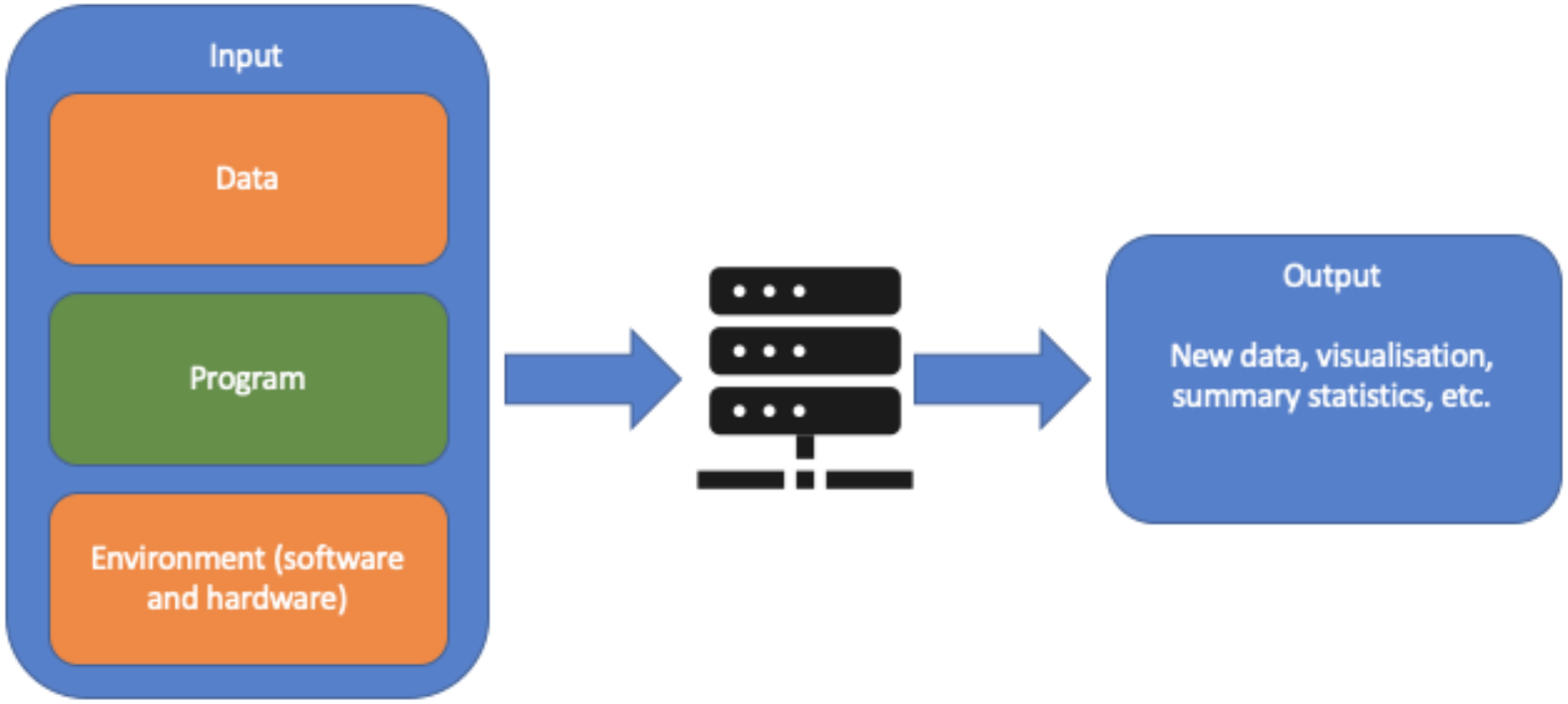
Quelques cas d'usages qui servent pour un papier

- Statistiques sur une enquête - *Repr.* ***statistique*** *et* ***computationnelle***
- Nettoyage, normalisation, etc. de données brutes de mesures - *Repr* ***computationnelle***
- Simulations numériques - *Repr.* ***expérimentale***, ***statistique*** *et* ***computationnelle***
- Calcul de quantités résumantes (stats, courbes), visualisation - *Repr.* ***computationnelle***
- ...
- *Quand le numérique intervient, la reproductibilité computationnelle est, a minima, recherchée*

Expérimentations numériques



Expérimentation(s) numérique(s)



Quels outils et quelles pratiques indispensables pour les expérimentations numériques ?

Objectif : dans le **contexte numérique**, nous aider **faire montre de rigueur** à travers une pratique **transparente**, **lisible** et **accessible** dans la méthodogie employée pour (re)produire de la connaissance

Open source (et libre, si possible)

- Exigence de **transparence** (et d'**accessibilité**)
- Perennité dans le temps *plus* assurée (*software heritage*)
- La plupart des logiciels closed source ont des **alternatives** (e.g. matlab vs python, intel-compiler vs. gcc)
- **Linux** (et, UNIX) : point focal de l'open source ; Environnement logiciel **aussi** open source/libre
- *N'oubliez pas d'aposer une licence logicielle sur votre code source...*

Transparent ≠ Lisible

```
          main(l
      ,a,n,d)char**a;{
  for(d=atoi(a[1])/10*80-
 atoi(a[2])/5-596;n="@NKA\
CLCCGZAAQBEAADAFaISADJABBA^\
SNLGAQABDAXIMBAACTBATAHDBAN\
ZcEMMCCCCAAhEIJFAEAAABAfHJE\
TBdFLDAANEfDNBPHdBcBBBEA_AL\
 H E L L O,    W O R L D! "
   [l++-3];)for(;n-->64;)
      putchar(!d+++33^
           l&1);}
```

Documentation (au sens large)

- Exigence de **lisibilité**
- Du logiciel que vous développez ou que vous utilisez
- Plusieurs formes : description des commandes utilisées, des algorithmes, commentaires dans le code, code lui-même explicite, notebooks, etc.
- *Tout ce qui est **indispensable** pour comprendre et rééxecuter (au niveau de votre logiciel) votre **méthodologie** doit être **explicitée**.*

Développement de code : Forge logicielle

- Ensemble d'outils, le plus souvent accessible sur le web, pour gérer et diffuser des codes sources : e.g. **gitlab**, github, bitbucket, etc.
- Basée sur un **gestionnaire de version** (*e.g.* **git**, svn, mercurial)
- Permet de :
  - gérer son code proprement, de façon collaborative si besoin, le sauvegarder
  - De publier son code
  - De publier de la documentation (doc proprement dite, accès au code, site web)
  - De mettre en place, entre autres, des mécanismes des tests automatiques (*intégration continue*, à utiliser avec parcimonie)

Remarques en vrac

Le choix des outils logiciels

- Pratiques et outils standards :
  - Compilation : *e.g.* make, cmake
  - Distribution de votre code : CRAN, pypi ;
  - Respectez les normes ! (codage, empaquetage, etc.)
- Privilégiez les outils qui ont une communauté active…
- …mais pas au détriment du cahier des charges !

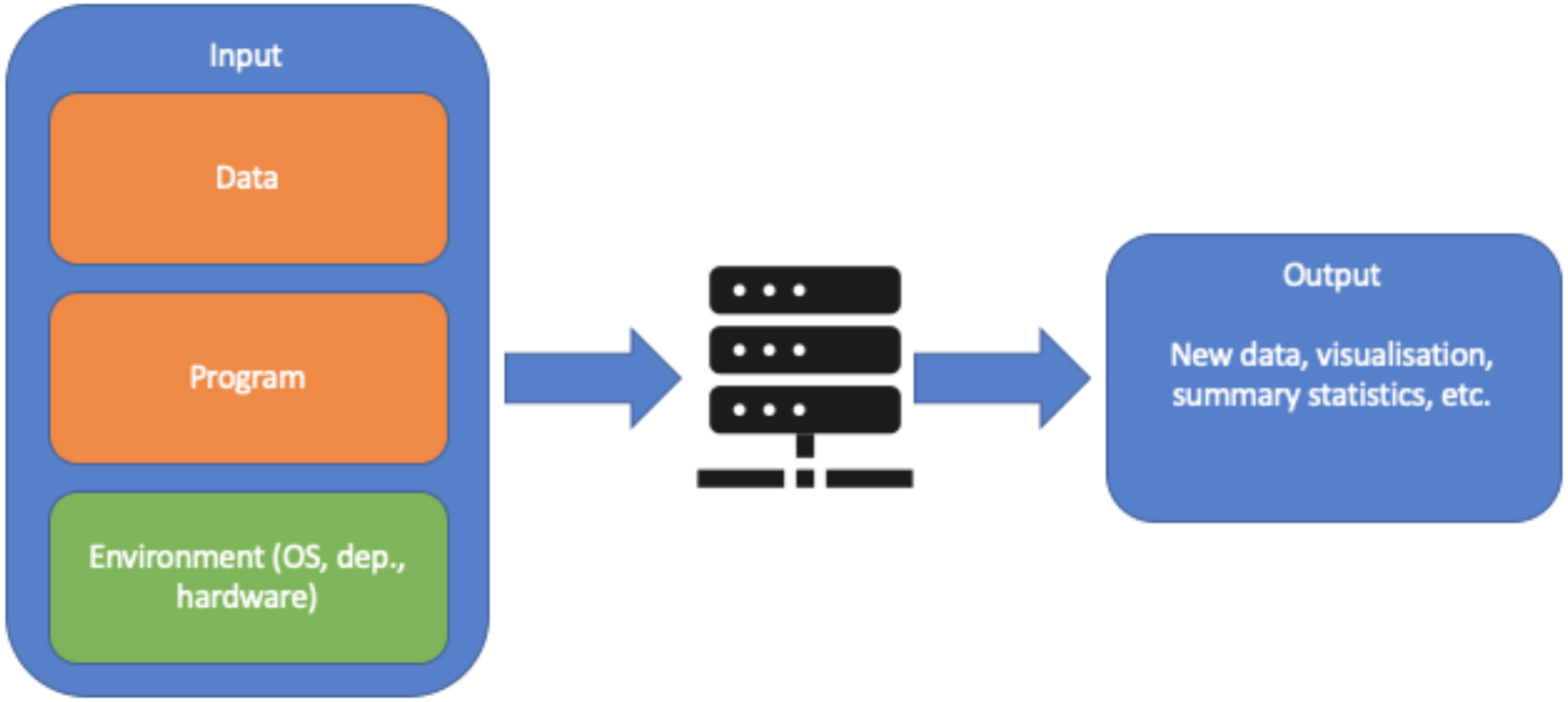
Performance et reproductibilité

- **Utilisez les bons outils :** Par exemple, un langage compilé sera souvent plus adapté au besoin de performance qu'un langage interprété
- **Peut rentrer en conflit** avec l'exigence de lisibilité et, parfois, de transparence et/ou d'accessibilité (*e.g.* compilateur intel, code involontairement obfusqué).
- **Accentue la dépendance** à l'environnement logiciel et matériel
- **Effet rebond** : cf. Prés. L. Bourgès et C. Bonamy
- Doit répondre à un *réel* besoin de performance

Les notebooks

- Un notebook (interface mélangeant texte, support visuel et code logiciel) :
  - Est un bon outil pour **expliquer une méthodologie**, présenter des résultats
  - Peut être un bon outil pour reproduire de simples calculs
  - Peut être une bonne interface pour exécuter des calculs
  - N'est pas souvent un bon outil pour *construire et reproduire* une expérimentation numérique (sauf exploration)

Quelques mots sur l'environnement logiciel (open source)



Quelques mots sur l'environnement logiciel (open source)

- Crucial pour la reproductibilité computationnelle…
- …Mais sujet complexe, notamment pour les néophytes (et pas que).
- **Les conteneurs ne sont pas la panacée. Ni conda !**
- N'hésitez pas à demander de l'aide sur ces sujets (*e.g.* ITA de labos, GRICAD) !
- *Présentation de L. Courtès à l'ANF UST4HPC (Guix inside)*

TL;DR

Pour tendre vers une recherche reproductible dans le contexte numérique, adoptez des pratiques et des outils qui vous aident à respecter :

- L'exigence de **transparence**
- L'exigence de **lisibilité**
- L'exigence d'**accessibilité**

TL;DR bis

Mais ça ne suffit pas !

- Utilisez ces outils rigoureusement : N'hésitez pas à **vous former** !
- Si vous ne savez pas, n'hésitez pas à **demander de l'aide** (GRICAD :)) !
- ***La rigueur prend du temps.***

Merci de votre attention