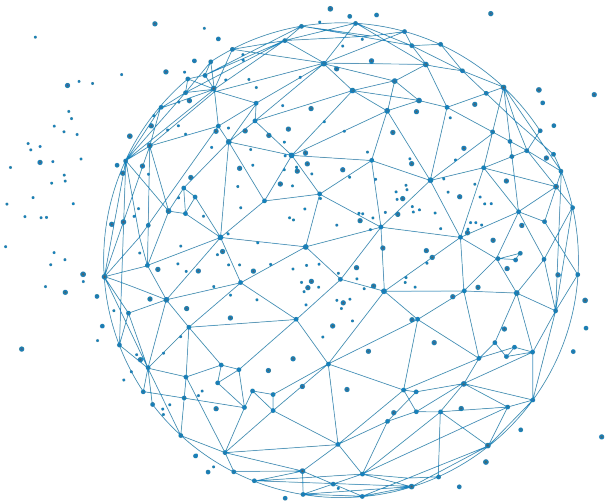


Science ouverte et principes FAIR dans un projet de bioinformatique

Comment rendre un projet bioinformatique plus reproductible ?

Roscoff, 14 mars 2023

Thomas Denecker

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea

Une présentation sous le signe de l'Open

Attribution - Partage dans les Mêmes Conditions 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Avertissement.

Vous êtes autorisé à :

Partager — copier, distribuer et communiquer le matériel par tous moyens et sous tous formats

Adapter — remixer, transformer et créer à partir du matériel pour toute utilisation, y compris commerciale.

L'Offrant ne peut retirer les autorisations concédées par la licence tant que vous appliquez les termes de cette licence.

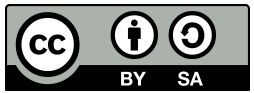
Selon les conditions suivantes :

Attribution — Vous devez créditer l'Œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Oeuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Oeuvre.

Partage dans les Mêmes Conditions — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'Oeuvre originale, vous devez diffuser l'Oeuvre modifiée dans les même conditions, c'est à dire avec la même licence avec laquelle l'Oeuvre originale a été diffusée.

Pas de restrictions complémentaires — Vous n'êtes pas autorisé à appliquer des conditions légales ou des mesures techniques qui restreindraient légalement autrui à utiliser l'Oeuvre dans les conditions décrites par la licence.

=



https://creativecommons.org/licenses/by-sa/4.0/deed.fr

Un contenu trouvable simplement, accessible, décrit et réutilisable

zenodo

DOI 10.5281/zenodo.7729199

ou

slideshare

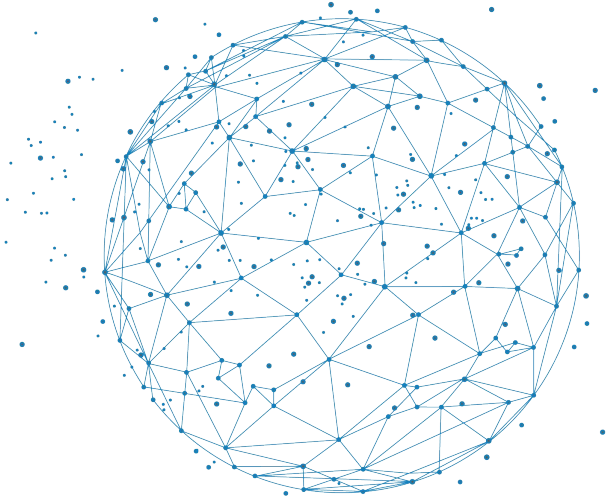
ou

HAL open science

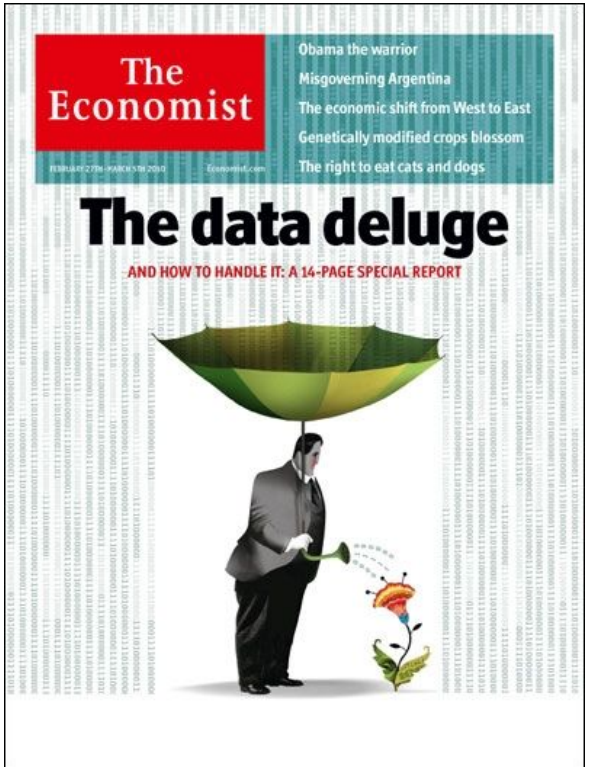
IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

Contexte



De plus en plus de données



Data is the new oil

Clive Humby

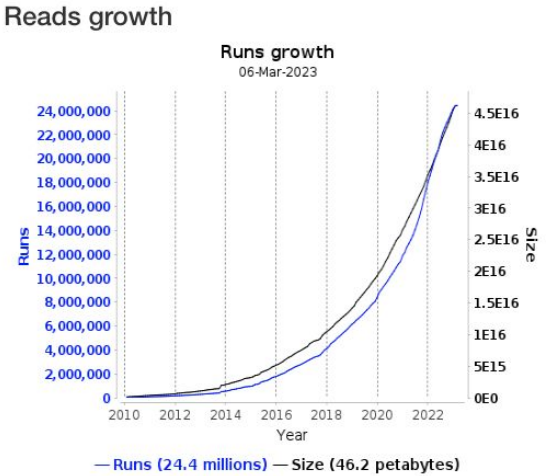
Data is the new oil? No: Data is the new soil.

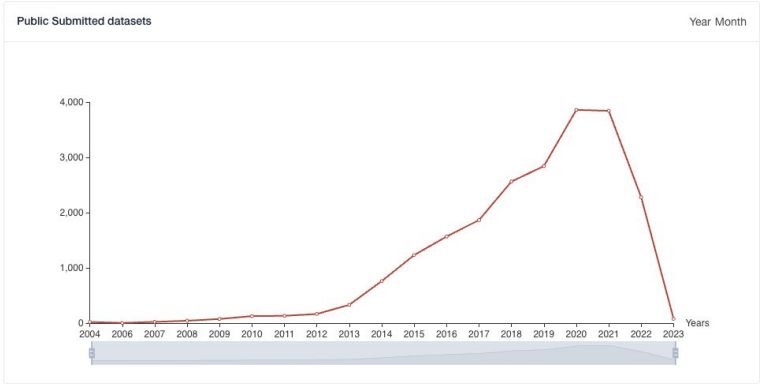
David Mccandless

Les données ne dorment jamais !

https://www.domo.com/data-never-sleeps#

En biologie, c’est pareil !





ENA
European Nucleotide Archive

https://www.ebi.ac.uk/ena/browser/about/statistics

PRIDE
PRoteomics IDEntifications Database

https://www.ebi.ac.uk/pride/statisticsdetails

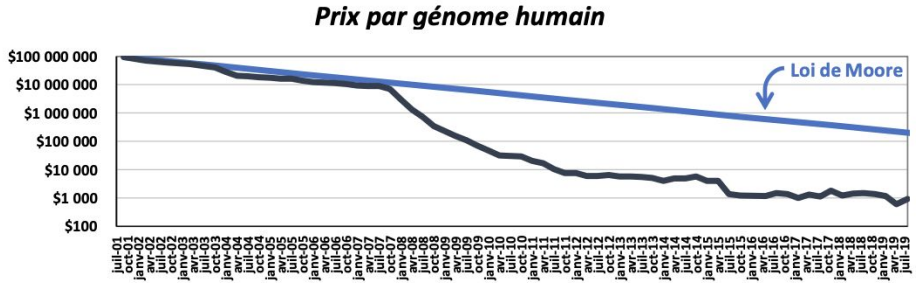
Avec des chiffres

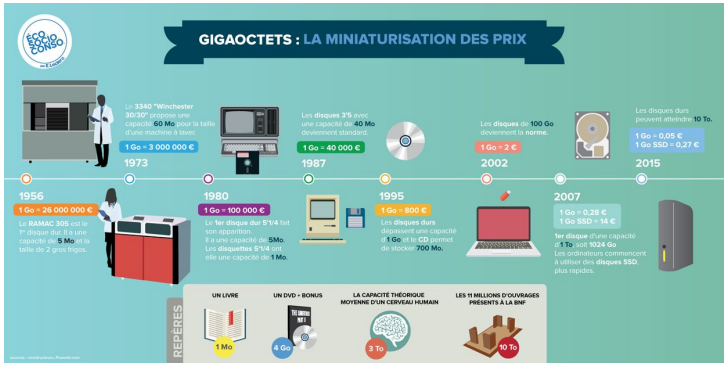
Type de données	Base de données	Volumes de données
Mesures de l'expression des gènes	ArrayExpress	72 938 experiments 2 429 810 assays 56,68 TB of archived data
Mesures de l'expression des gènes	GEO	5 570 704 échantillons
Structure 3D des protéines	PDB	177 009 Structures
Séquence nucléotidiques	GenBank	241 830 635 séquences et 1 731 302 248 418 bases
Données d'identification ou de quantification des protéines	Pride	580 917 268 spectres de masse
Séquence nucléotidiques (COVID-19)	GISAID (EpiCoV)	15 158 725 séquences virales

MAJ : Mars 2023

Comment ?

Cout





Vitesse

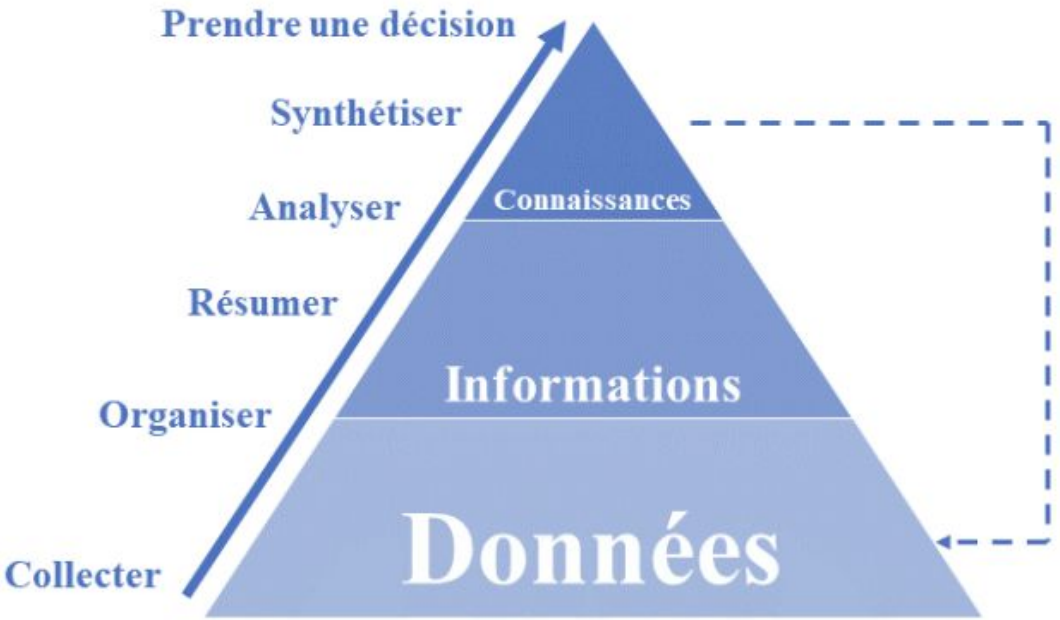
Dans les années 90

4 ans pour le premier milliard de nucléotides du génome humain

Aujourd'hui

Le génome complet en moins de 24h

Pourquoi ? Créer de la connaissance !



Pourquoi générer toujours plus de données ?

Pourquoi ne pas simplement exploiter les données déjà disponibles ?

- La description des données est encore trop souvent incomplète ;
- Les données ne sont pas facilement récupérables ;
- Il n'y a souvent pas de contrôle systématique des erreurs par des experts ;
- Les données ne sont pas générées exactement de la façon souhaitée ;
- Une question de confiance.

Conclusion : Plus simple ? Plus rapide ? Plus sûr ?

Une question de confiance

Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

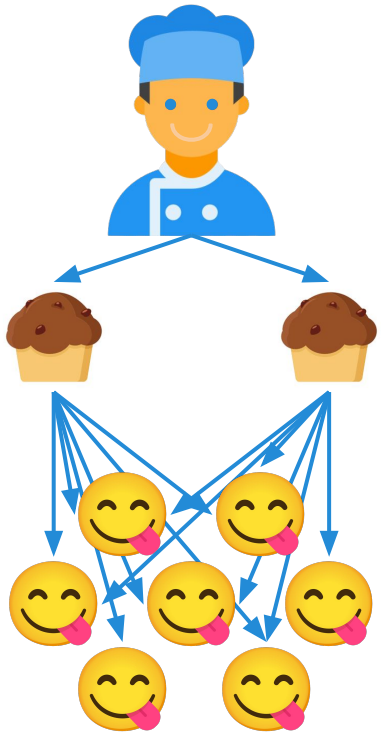
Réplicabilité

Répétabilité

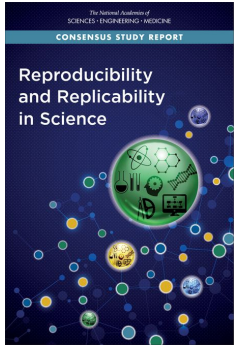
Reproductibilité

Souvent utilisées mais souvent confondues et notamment par la langue (Plesser, 2018)

Réplicabilité



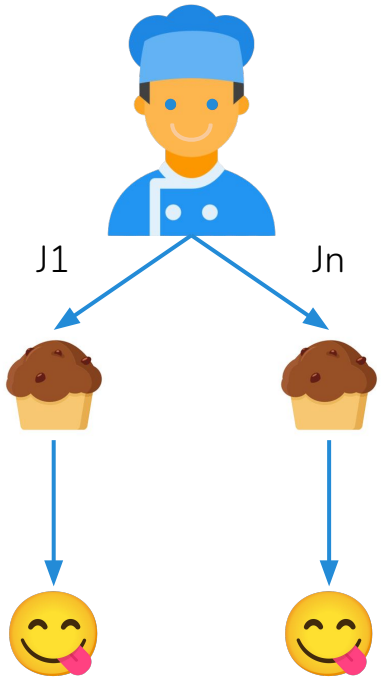
“L’étroitesse de l’accord entre les résultats individuels successifs obtenus sur le même échantillon soumis à l’essai dans le même laboratoire et dans les conditions suivantes : même analyste, même appareil, même jour ”



https://doi.org/10.17226/25303

“*Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data”*

Répétabilité

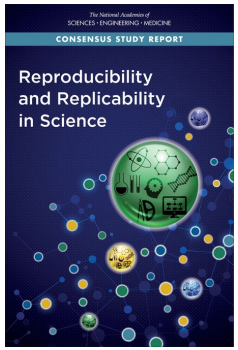



NORME INTERNATIONALE ISO 5725-4

Exactitude (justesse et fidélité) des résultats et méthodes de mesure —

Partie 4: Méthodes de base pour la détermination de la justesse d'une méthode de mesure normalisée

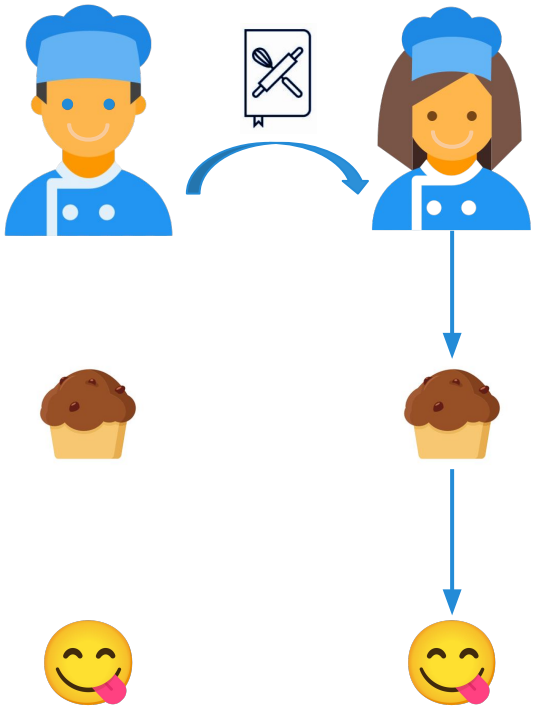
"L'étroitesse de l'accord entre les résultats individuels obtenus sur le même échantillon soumis à l'essai dans le même laboratoire et dont au moins l'un des éléments suivants est différent : l'analyste, l'appareil, le jour"



"*The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation* "

https://doi.org/10.17226/25303

Reproductibilité




NORME INTERNATIONALE ISO 5725-4

Exactitude (justesse et fidélité) des résultats et méthodes de mesure —

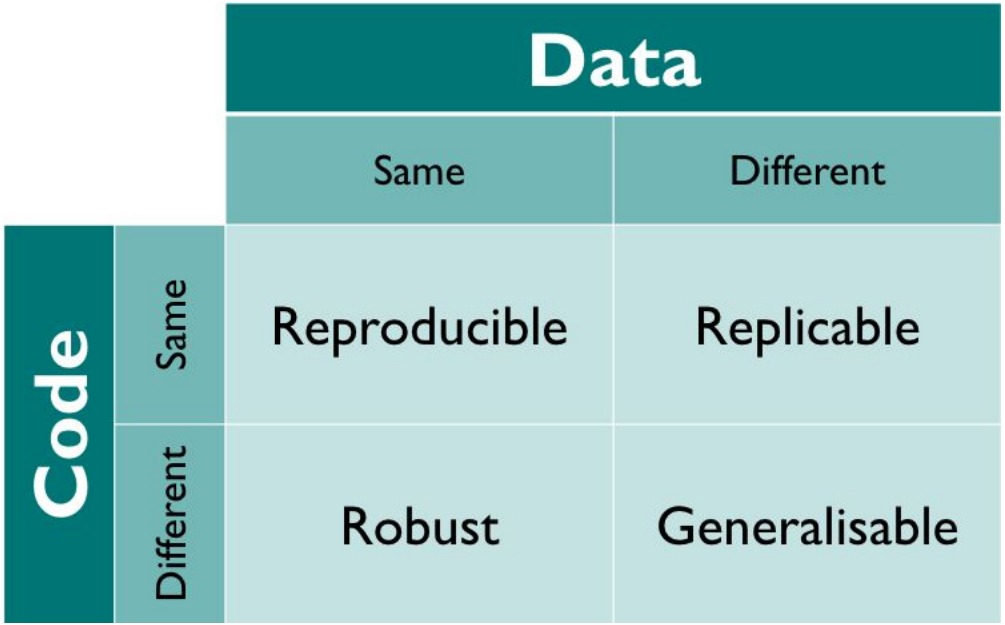
Partie 4: Méthodes de base pour la détermination de la justesse d'une méthode de mesure normalisée

"L'étroitesse de l'accord entre les résultats individuels obtenus sur le même échantillon soumis à l'essai dans des laboratoires différents et dans les conditions suivantes : analyste différent, appareil différent, jour différent ou même jour"

Association for Computing Machinery

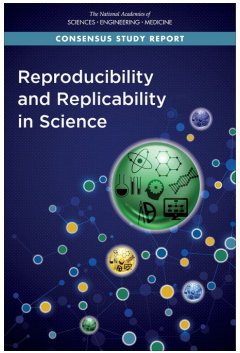
"Obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis"

En résumé



https://doi.org/10.6084/m9.figshare.5443201.v1,

Recommandations pour être reproductible



https://doi.org/10.17226/25303

Description de la partie expérimentale

Méthodes, instruments, procédures, mesures, conditions expérimentales

Description de la partie computationnelle

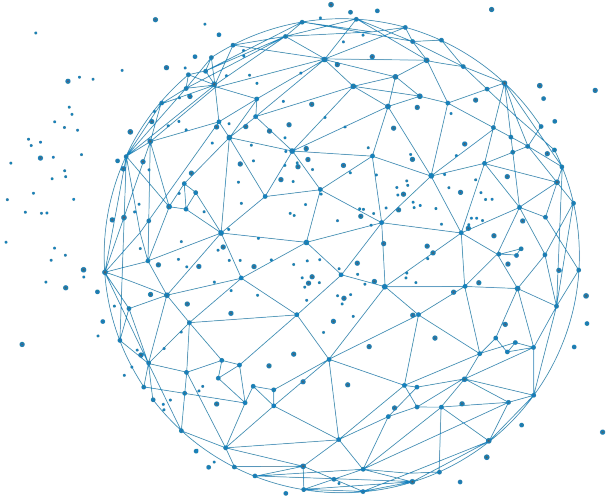
Etapes de l'analyse des données et choix techniques

Description de la partie statistique

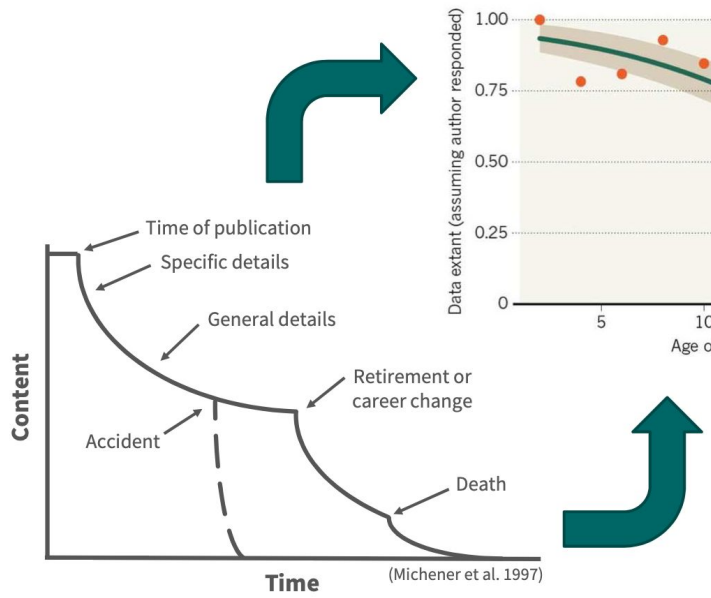
Décisions analytiques : quand, comment, pourquoi

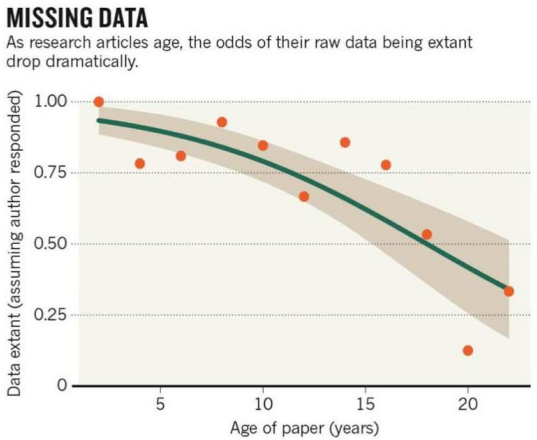
Discussion des choix et des résultats obtenus

Et dans les faits ?



Data Entropy





Vines, T. H. *et al. Curr. Biol.* http://dx.doi.org/10.1016/j.cub.2013.11.014 (2013).

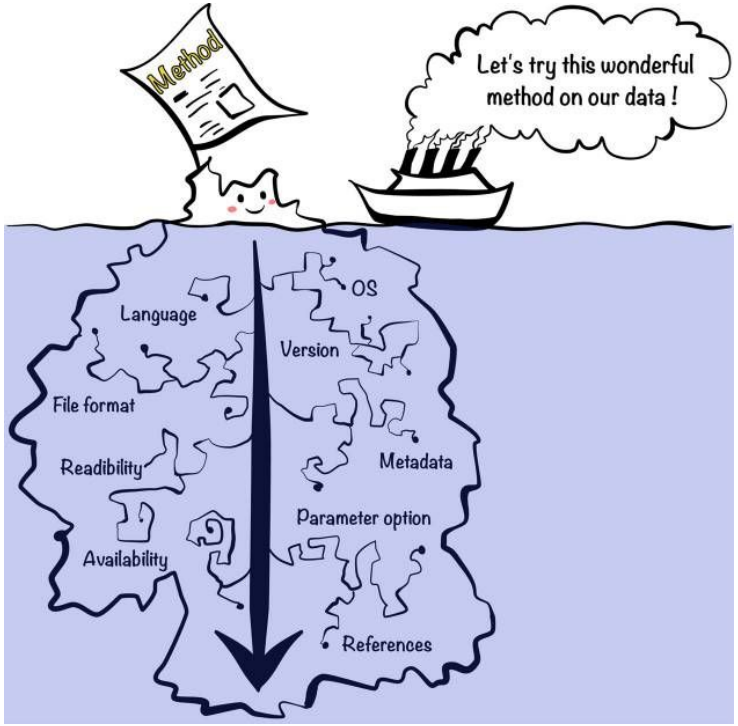
DataONE

3

Récupérer les données

https://youtu.be/66oNv_DJuPc

Reproduire les données



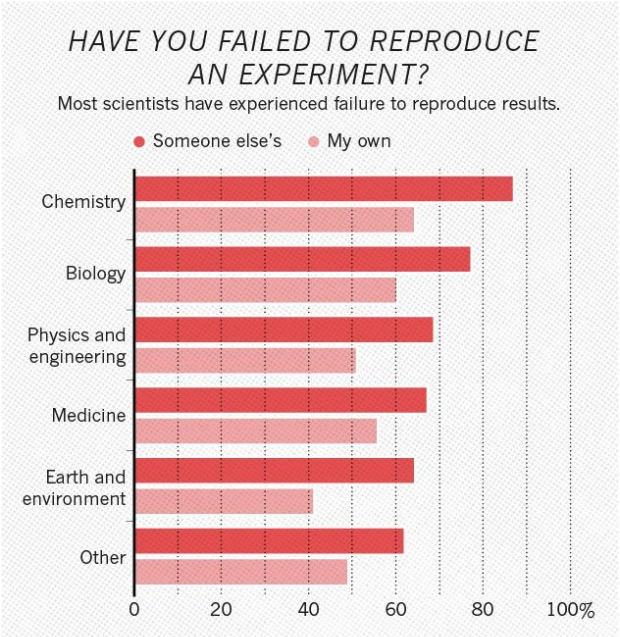
Kim et al, 2018

https://dx.doi.org/10.1093%2Fgigascience%2Fgiy077

70 %

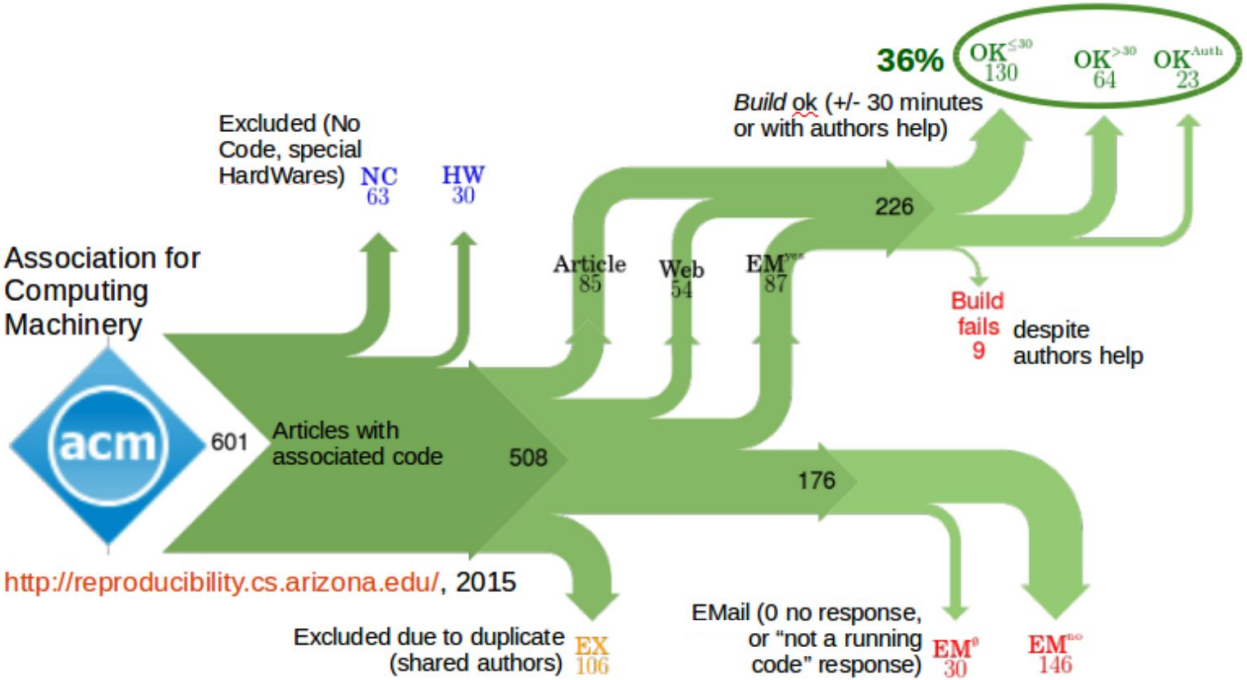
des analyses en biologie expérimentale ne sont

pas reproductibles



Monya Baker, 2016

La crise de la reproductibilité : en informatique

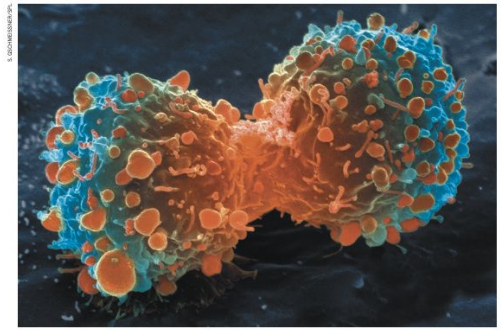


(Collberg et al. 2015)

La crise de la reproductibilité : en bioinformatique

COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even ▸

29 MARCH 2012 | VOL 483 | NATURE | 531

Nekrutenko & Taylor, Nature Genetics (2012)
Alsheikh-Ali et al. PLoS ONE (2011)
Begley & Ellis Nature (2012)

Mais pourquoi ?

Impossibilité d'installer des outils

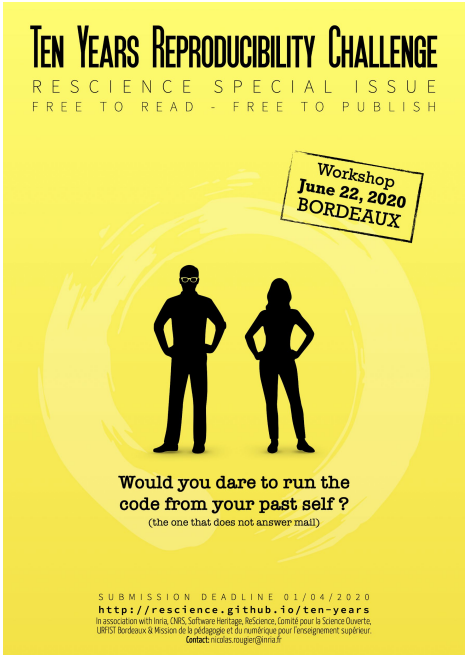
- OS non compatible
- Dépendance plus disponible / plus valide

Mise à jour de l'outil rendant inutilisable les codes

- Python 2 et Python 3 !
- Changement des arguments des fonctions utilisées (R)

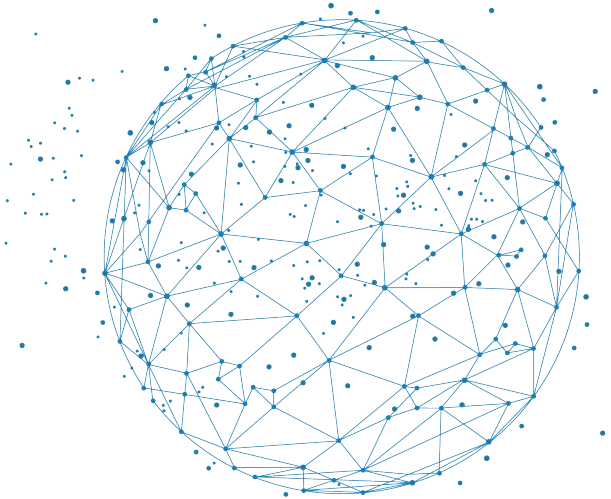
Impossibilité de reproduire les résultats de l'analyse computationnelle

- IDE : version stable du langage différente selon l'OS (Rstudio)
- Version des packages

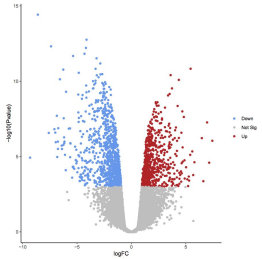


(Perkel, Nature, 2020)

Comment rendre un projet bioinformatique plus reproductible ?



En étant plus FAIR !



Données brutes

Principes FAIR data

&

Plan de gestion de données

Analyse de données

Codes

Algorithmes

Workflow

...

Communications

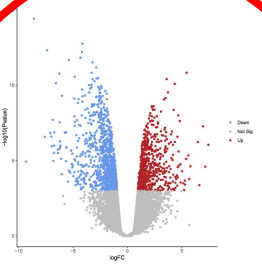
Articles

Thèse

Poster

...

En étant plus FAIR !



Données brutes

Principes FAIR data

&

Plan de gestion de données

Analyse de données

Codes

Algorithmes

Workflow

...

Communications

Articles

Thèse

Poster

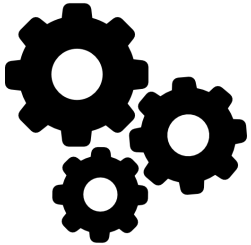
...

Des principes hérités des principes FAIR data

Findable

Accessible

Interoperable



Reusable

Des principes hérités des principes FAIR data

Facile à trouver (pour l'Homme et la machine)

- Identifiant unique (un DOI par exemple)
- Métadonnées décrivant l'analyse et les outils
- Les métadonnées sont FAIR, consultables et indexables

Accessible

- Ouvert, gratuit et universellement implémentable
- Les métadonnées sont accessibles, même lorsque le logiciel n'est plus disponible

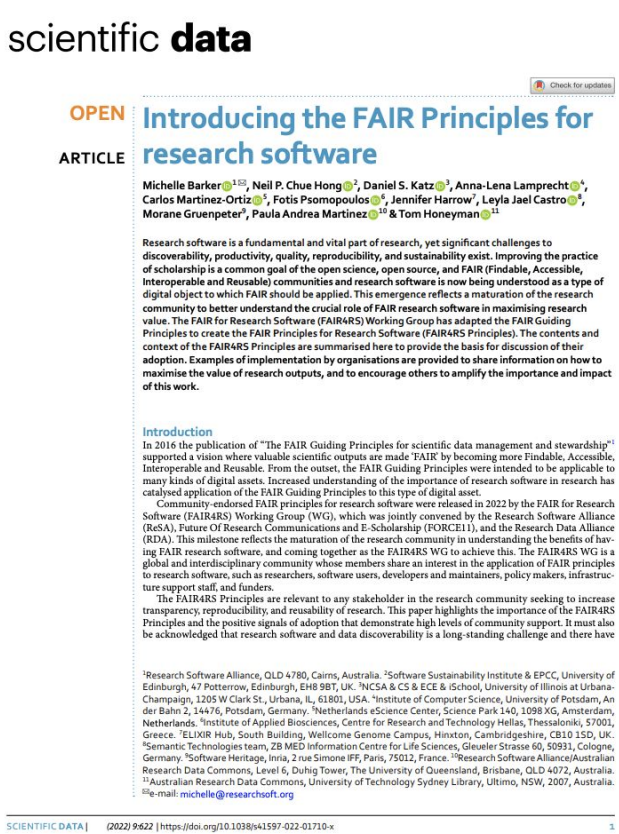
Interopérable

- Coopération des outils aussi bien en local que sur serveurs

Réutilisable

- Licence claire
- Suivi les normes communautaires

Pour aller plus loin



scientific **data**

Check for updates

OPEN

ARTICLE

Introducing the FAIR Principles for research software

Michelle Barker[1✉], Neil P. Chue Hong[2], Daniel S. Katz[3], Anna-Lena Lamprecht[4], Carlos Martinez-Ortiz[5], Fotis Psomopoulos[6], Jennifer Harrow[7], Leyla Jael Castro[8], Morane Gruenpeter[9], Paula Andrea Martinez[10] & Tom Honeyman[11]

Research software is a fundamental and vital part of research, yet significant challenges to discoverability, productivity, quality, reproducibility, and sustainability exist. Improving the practice of scholarship is a common goal of the open science, open source, and FAIR (Findable, Accessible, Interoperable and Reusable) communities and research software is now being understood as a type of digital object to which FAIR should be applied. This emergence reflects a maturation of the research community to better understand the crucial role of FAIR research software in maximising research value. The FAIR for Research Software (FAIR4RS) Working Group has adapted the FAIR Guiding Principles to create the FAIR Principles for Research Software (FAIR4RS Principles). The contents and context of the FAIR4RS Principles are summarised here to provide the basis for discussion of their adoption. Examples of implementation by organisations are provided to share information on how to maximise the value of research outputs, and to encourage others to amplify the importance and impact of this work.

Introduction

In 2016 the publication of "The FAIR Guiding Principles for scientific data management and stewardship"[1] supported a vision where valuable scientific outputs are made 'FAIR' by becoming more Findable, Accessible, Interoperable and Reusable. From the outset, the FAIR Guiding Principles were intended to be applicable to many kinds of digital assets. Increased understanding of the importance of research software in research has catalysed application of the FAIR Guiding Principles to this type of digital asset.

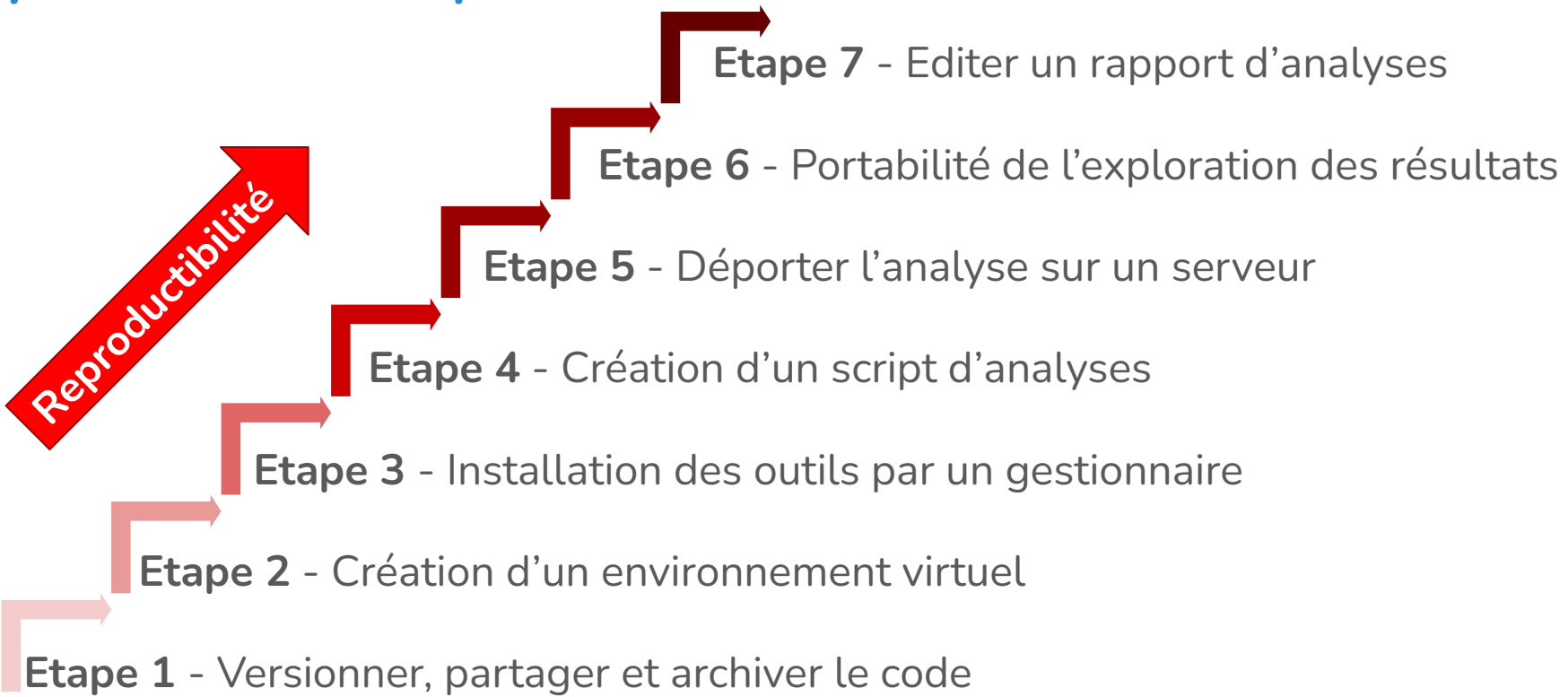
Community-endorsed FAIR principles for research software were released in 2022 by the FAIR for Research Software (FAIR4RS) Working Group (WG), which was jointly convened by the Research Software Alliance (ReSA), Future Of Research Communications and E-Scholarship (FORCE11), and the Research Data Alliance (RDA). This milestone reflects the maturation of the research community in understanding the benefits of having FAIR research software, and coming together as the FAIR4RS WG to achieve this. The FAIR4RS WG is a global and interdisciplinary community whose members share an interest in the application of FAIR principles to research software, such as researchers, software users, developers and maintainers, policy makers, infrastructure support staff, and funders.

The FAIR4RS Principles are relevant to any stakeholder in the research community seeking to increase transparency, reproducibility, and reusability of research. This paper highlights the importance of the FAIR4RS Principles and the positive signals of adoption that demonstrate high levels of community support. It must also be acknowledged that research software and data discoverability is a long-standing challenge and there have

[1]Research Software Alliance, QLD 4780, Cairns, Australia. [2]Software Sustainability Institute & EPCC, University of Edinburgh, 47 Potterrow, Edinburgh, EH8 9BT, UK. [3]NCSA & CS & ECE & iSchool, University of Illinois at Urbana-Champaign, 1205 W Clark St., Urbana, IL, 61801, USA. [4]Institute of Computer Science, University of Potsdam, An der Bahn 2, 14476, Potsdam, Germany. [5]Netherlands eScience Center, Science Park 140, 1098 XG, Amsterdam, Netherlands. [6]Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, 57001, Greece. [7]ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. [8]Semantic Technologies team, ZB MED Information Centre for Life Sciences, Gleueler Strasse 60, 50931, Cologne, Germany. [9]Software Heritage, Inria, 2 rue Simone IFF, Paris, 75012, France. [10]Research Software Alliance/Australian Research Data Commons, Level 6, Duhig Tower, The University of Queensland, Brisbane, QLD 4072, Australia. [11]Australian Research Data Commons, University of Technology Sydney Library, Ultimo, NSW, 2007, Australia. ✉e-mail: michelle@researchsoft.org

https://doi.org/10.1038/s41597-022-01710-x

Proposition en 7 étapes

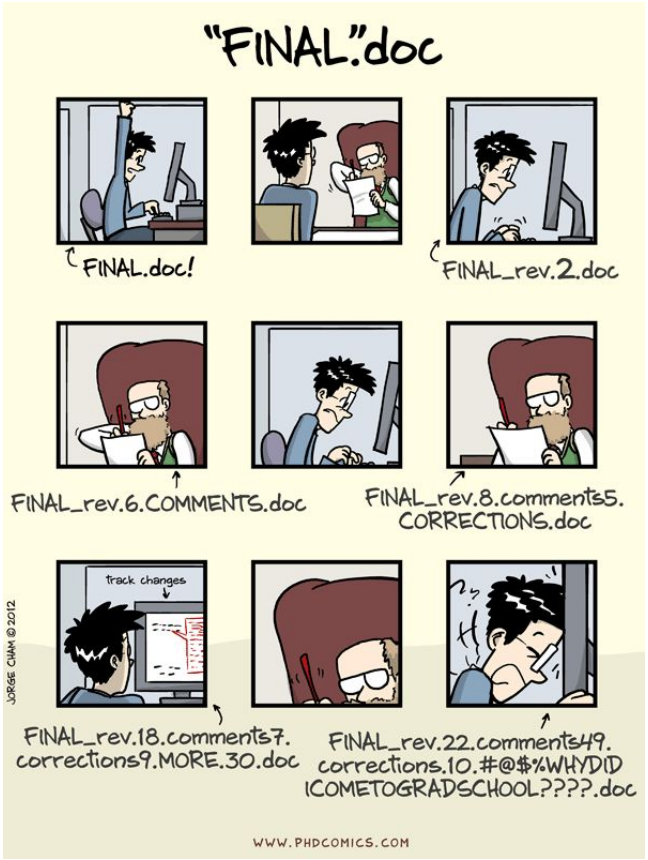


Débat sur les choix techniques

Les principes sont totalement indépendant des outils présentés

Pourquoi ?

- Avoir la bonne version du code
- Vision dans le temps
- Ouverture à la communauté



Étape 1 - Versionner le code, le partager et l'archiver

Verisonner → **Versionner / Partager** → **Partager** → **Archiver**

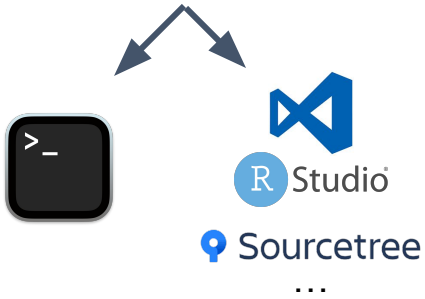
GitLab

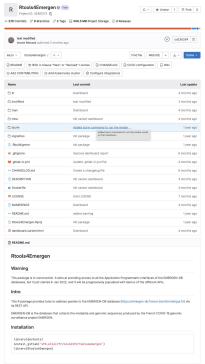
GitHub

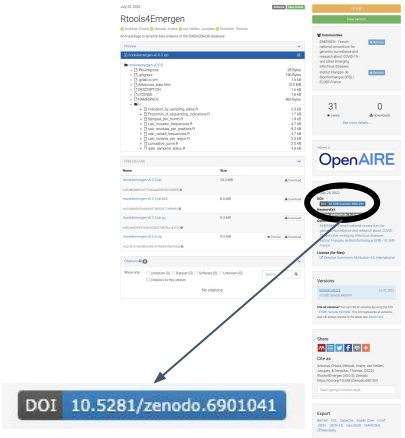
Atlassian Bitbucket

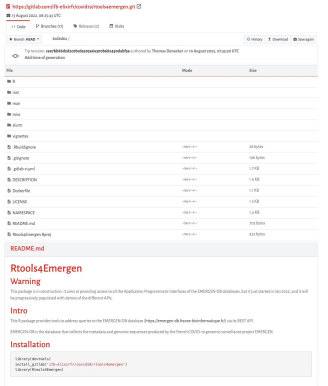
zenodo

Software Heritage









IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

Étape 1 - Versionner le code, le partager et l'archiver

Pourquoi ?

- Figer l'environnement
- Partager l'environnement

Comment

Avantages / Inconvénients

+ Sauvegarde du code
+ Simple pour partager
+ Gestion automatique des versions

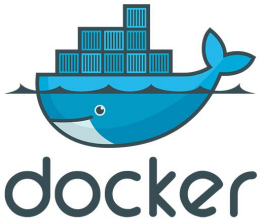
- Pas simple pour les novices

Étape 2 - Création d'un environnement virtuel

Pourquoi ?

- Figer l'environnement
- Partager l'environnement

Comment



Avantages / Inconvénients

+ Rapide et léger
+ Portable
+ Simple à partager et déployer

- Avec un système à jour
- [docker] Accepté dans votre structure ?

Étape 3 - Installation des outils par un gestionnaire

Pourquoi ?

- Avoir la bonne version des outils utilisés
- Les installer simplement

Comment

Avantages / Inconvénients

+ Gestionnaire simple à installer
+ Installation simple des paquets
+ Gestion des versions

- Peut être lourd (solution miniconda)
- Paquets manquants (R)

Étape 4 - Création d'un script d'analyse

Pourquoi ?

- Avoir un script d'analyse reproductible
- Ne pas refaire ce qui est déjà fait
- Paralléliser

Comment

Avantages / Inconvénients

+ Gestion des jobs
+ Puissant et rapide
+ Capable d'utiliser des environnements Conda
+ Parallélisable sur un cluster

- Une logique à apprendre
- Syntaxe moins simple que le script shell

Étape 5 - Déporter l'analyse sur un serveur

Pourquoi ?

- Environnement contrôlé
- Déport de l'analyse

Comment

Avantages / Inconvénients

+ Simple à mettre en place
+ Augmentation de la puissance (cloud ou cluster)
+ Pour tout le monde

- Pas simple pour les novices
- Attention aux données sensibles

Étape 6 - Portabilité de l'exploration des résultats

Pourquoi ?

- Rendre simple l'exploration
- Simple à partager

Comment

Avantages / Inconvénients

+ Portable (HTML)
+ Accessible partout
+ Interactif (paramétrable, graphes dynamiques, ...)

- Mélange de langage

Étape 7 - Editer un rapport d'analyse

Pourquoi ?

- Avoir une trace de l'analyse (date, heure, paramètres, ...)
- Stocker les versions des outils

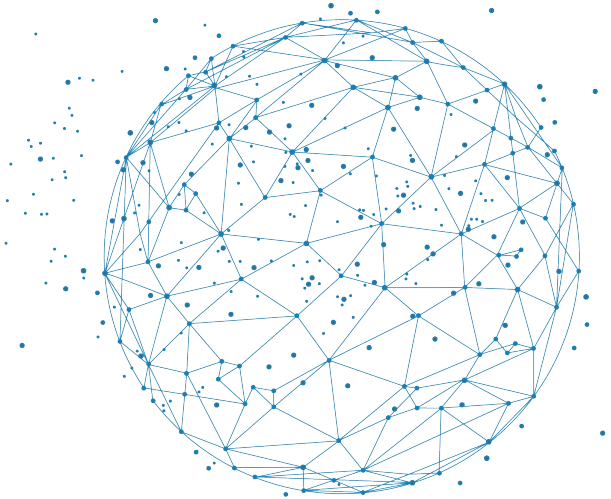
Comment

Avantages / Inconvénients

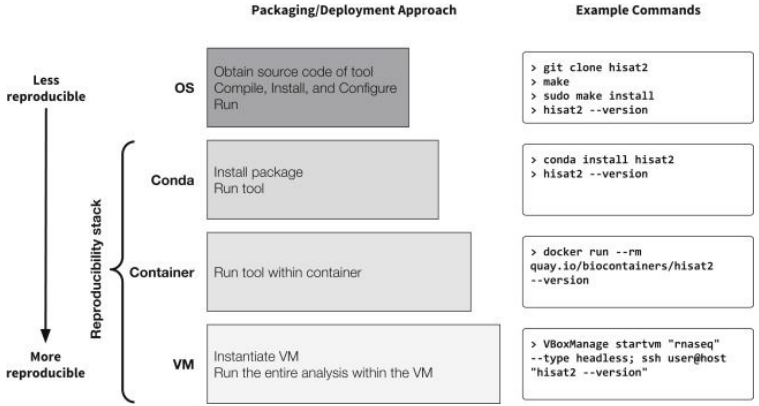
+ Syntaxe simple (Markdown)
+ Partage (PDF, HTML, ...)

- Rares problèmes de visualisation en LaTeX

Conclusion

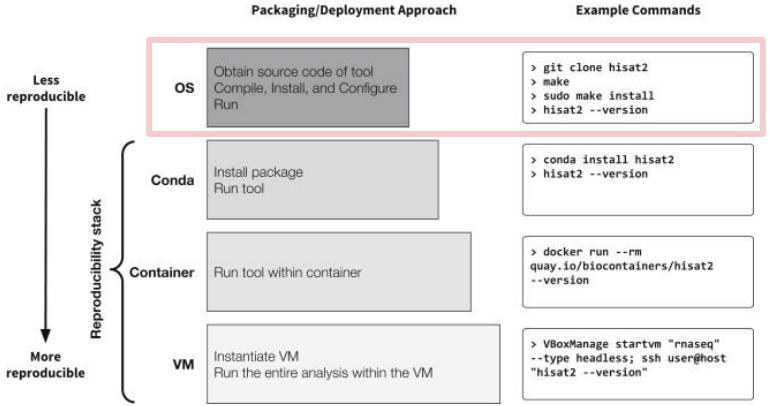


Quel est notre niveau de reproductibilité?



Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018

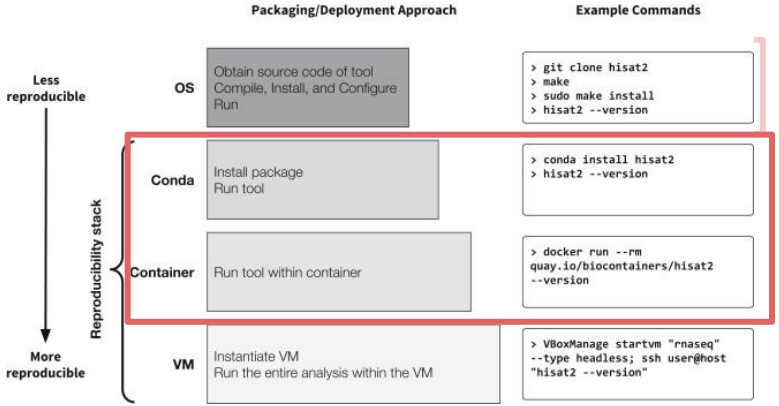
Quel est notre niveau de reproductibilité?



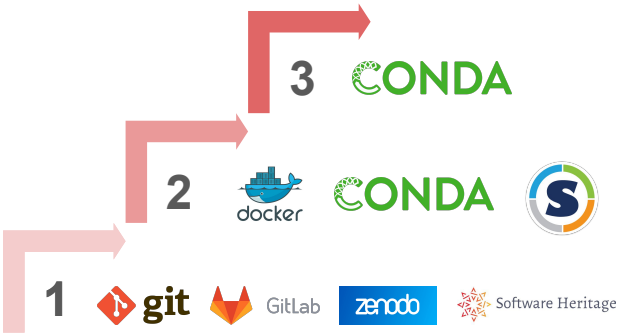
Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018

1 git GitLab zenodo Software Heritage

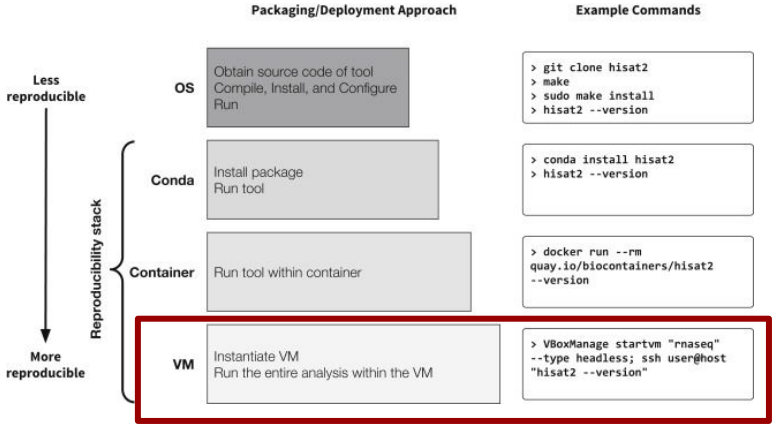
Quel est notre niveau de reproductibilité?



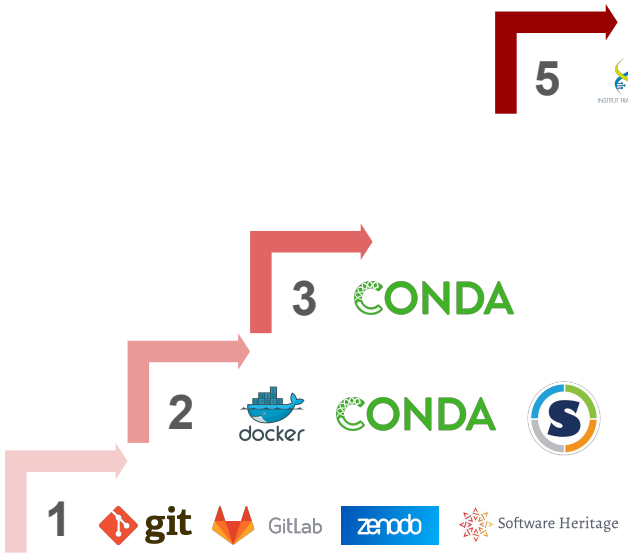
Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018



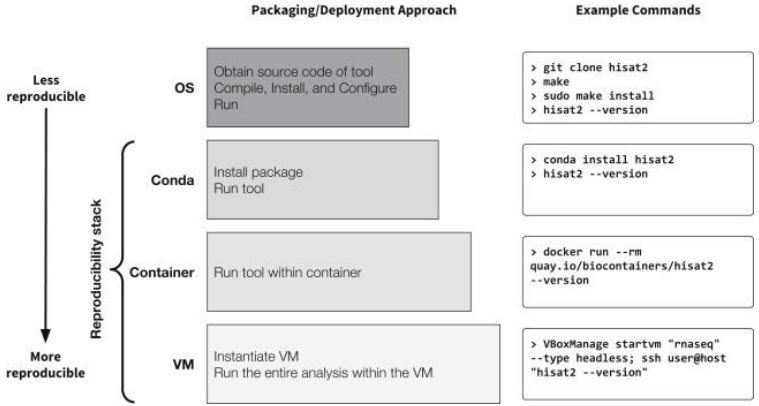
Quel est notre niveau de reproductibilité?



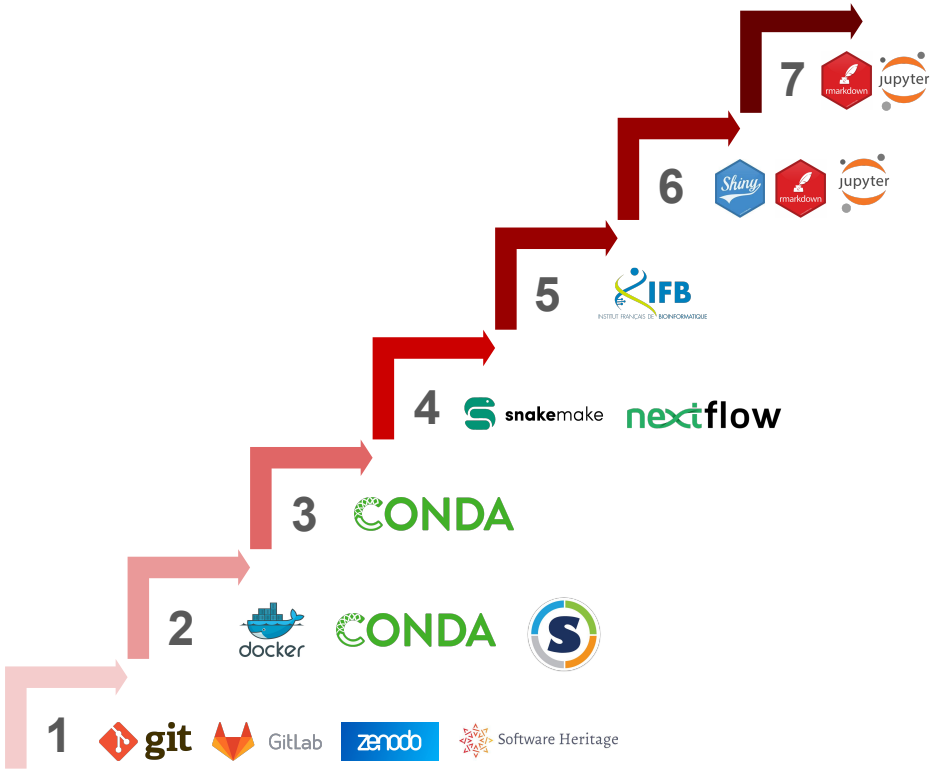
Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018



Quel est notre niveau de reproductibilité?

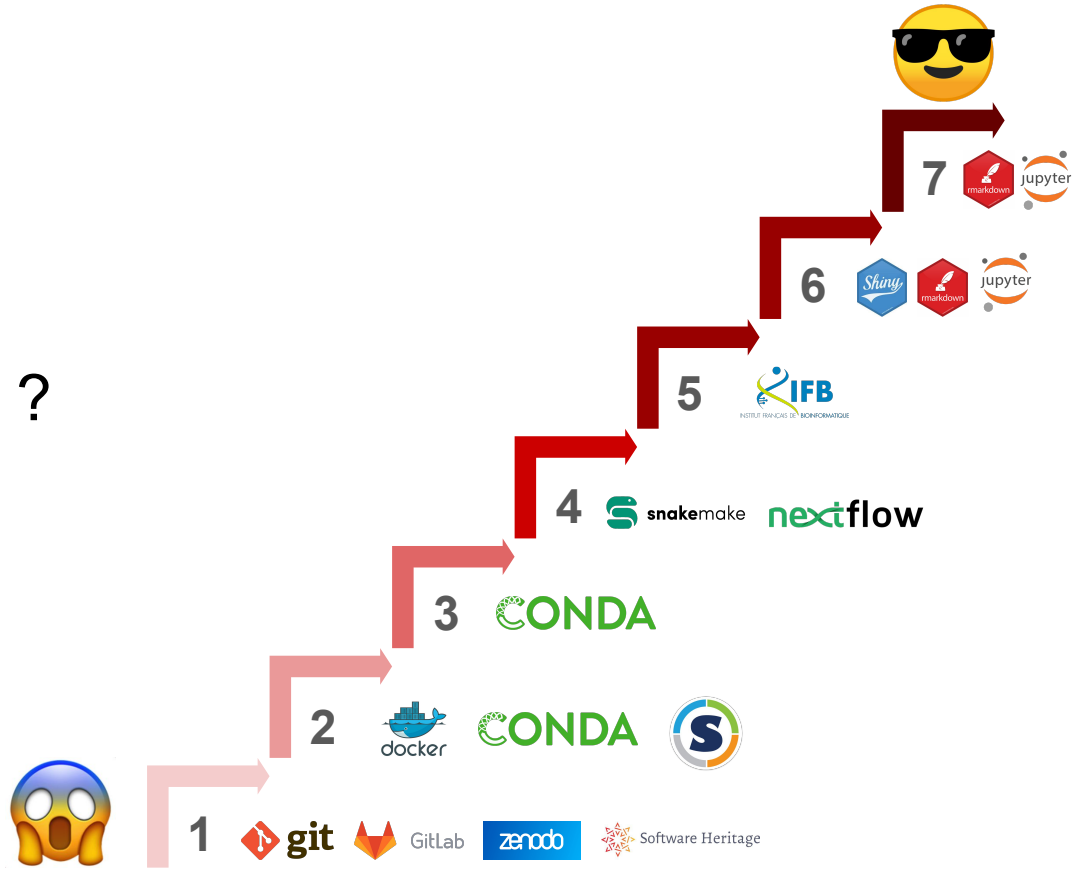


Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018



Quel est notre niveau de reproductibilité?

Et vous ?
Où vous situez-vous ?



Pour aller encore plus loin dans la reproductibilité

Guix

Une question toujours ouvertes

1. **Données**
   a. Le faire si les données ne sont pas FAIR ?
   b. Comment gérer les gros volumes des données ?
   c. Comment gérer les mises à jour des données et métadonnées ?
2. **Code**
   a. Comment être sûr que le code sera toujours accessible ?
   b. Acceptable s'il faut faire des adaptations ? A quel point ?
   c. Fournir tout le code ? (valorisation, création d'une start-up,...)
3. **Temps de calcul** (jour, mois, années)
4. **Compétence et sensibilité**
   a. Volonté mais incapacité technique à le faire
   b. "Pourquoi faire, Ca ne sert à rien"
   c. Trop long
5. **La couverture**
   a. Faut-il tout rendre reproductible ?
6. **Quand le faire ?**
   a. Au début ? Mais si ça ne marche pas ?
   b. A la fin ?

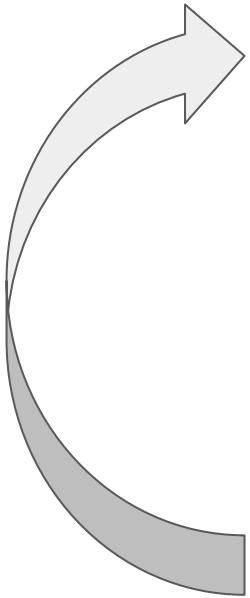
Une vraie réflexion sur la reproductibilité de ses analyses en Bioinformatique

(équilibre, temps, périmètre, …)

Proposition d'une solution qui aide à rendre reproductible n'importe quel protocole d'analyse

La reproductibilité est une plus value pour la Bioinformatique !

Un cercle vertueux



FAIR raw data

+

FAIR_bioinfo scripts/protocols

=

FAIR processed data

Formations IFB sur le thème du FAIR

1. Les principes FAIR pour la gestion des données de recherche en sciences de la vie
2. Les principes FAIR dans un projet de Bioinformatique

https://moodle.france-bioinformatique.fr/

Comité pour la Science ouverte

Ambition posée par la loi pour une république numérique (2016)

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

2018-2021

2021-2024

Ouvrir l'ensemble des produits et des méthodes de la recherche

- Les publications
- Les données
- Les codes sources

Une science plus cumulative, plus robuste, **plus reproductible**, **transparente** et **accessible à toutes et tous**

Le premier bénéficiaire de la science ouverte est le chercheur lui-même

Isabelle Blanc

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE
elixir FRANCE

Merci pour votre attention

Et n’hésitez pas à me contacter pour continuer à en discuter

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea INVESTIR L’AVENIR