Introduction aux bonnes pratiques pour des analyses reproductibles

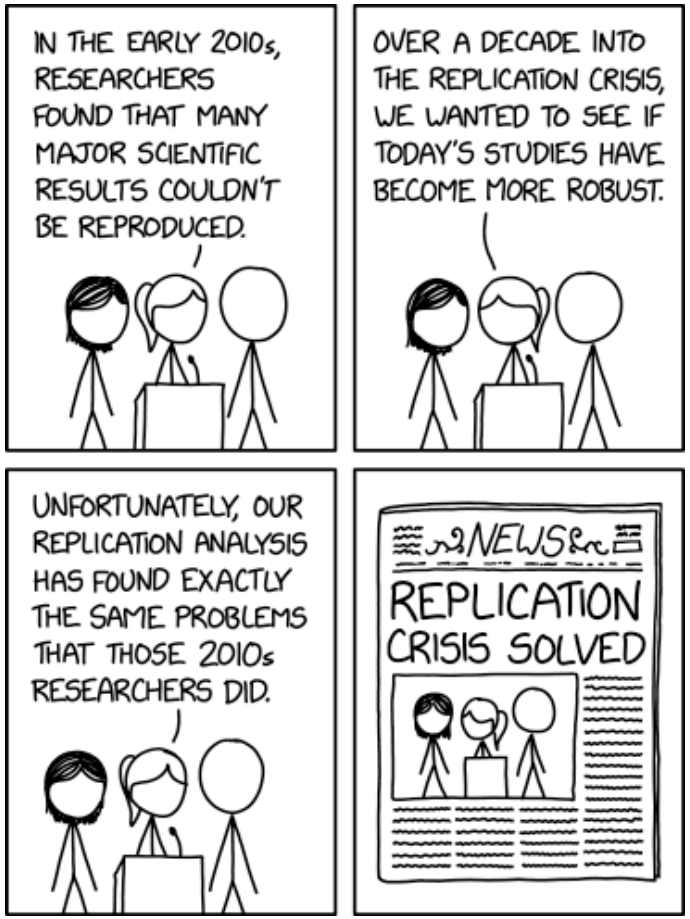
Cédric Midoux
Migale - INRAE

Nadia Goué
AuBi - Université Clermont Auvergne

September 29, 2025

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Reproductibilité et Science Ouverte



https://xkcd.fyi/3117

Tout le monde a déjà eu cette expérience

Un article intéressant

OPEN ACCESS Freely available online PLoS one

The *Arthrobacter arilaitensis* Re117 Genome Sequence Reveals Its Genetic Adaptation to the Surface of Cheese

Christophe Monnet[1,2]*, Valentin Loux[3], Jean-François Gibrat[3], Eric Spinnler[1,2], Valérie Barbe[4], Benoit Vacherie[4], Frederick Gavory[4], Edith Gourbeyre[5], Patricia Siguier[5], Michaël Chandler[5], Rayda Elleuch[6], Françoise Irlinger[1,2☯], Tatiana Vallaeys[7☯]

Un Mat & Meth décevant

collaboration with the user community. Genome comparisons were performed using Origami, an in-house tool developed for microbial genome comparison. Orthologs were defined as reciprocal best hits with an e-value lower than 10^{-3}. Transposases were excluded from the analysis. Core genes were defined as orthologs shared between the four *Arthrobacter* strains. Synteny was studied using an in-house developed tool, Align, using dynamic programming to search conserved gene trains allowing gaps and "mismatches" (homology relation instead of orthology). Circular representation of the genome was produced using the Circos software [27].

Crise de la reproductibilité

Problème général, “Reproducibility Crisis”

- Remis en avant par les sciences sociales, notamment en psychologie
- Étendu à l’ensemble des disciplines scientifiques

Mais un problème qui n’est pas nouveau

Expériences de la pompe à vide au XVIIe siècle (*von Guericke* et *Boyle*)

Un déluge de données

Evolutions des sciences

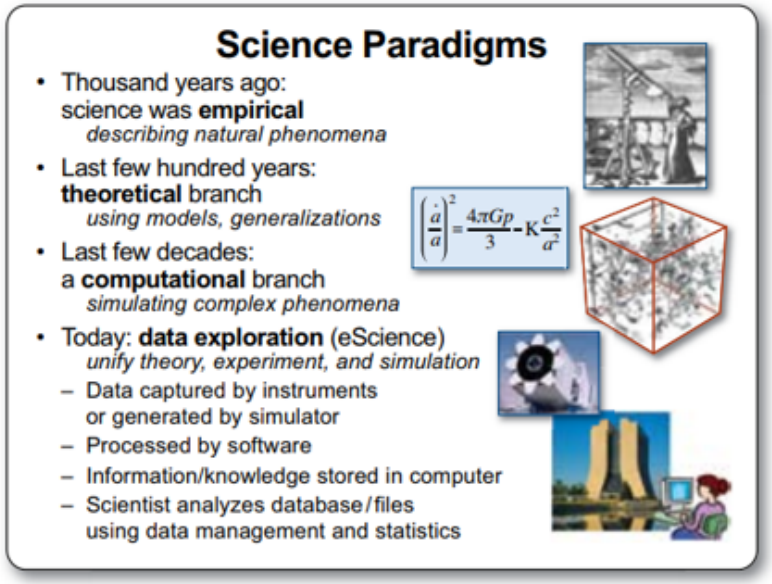
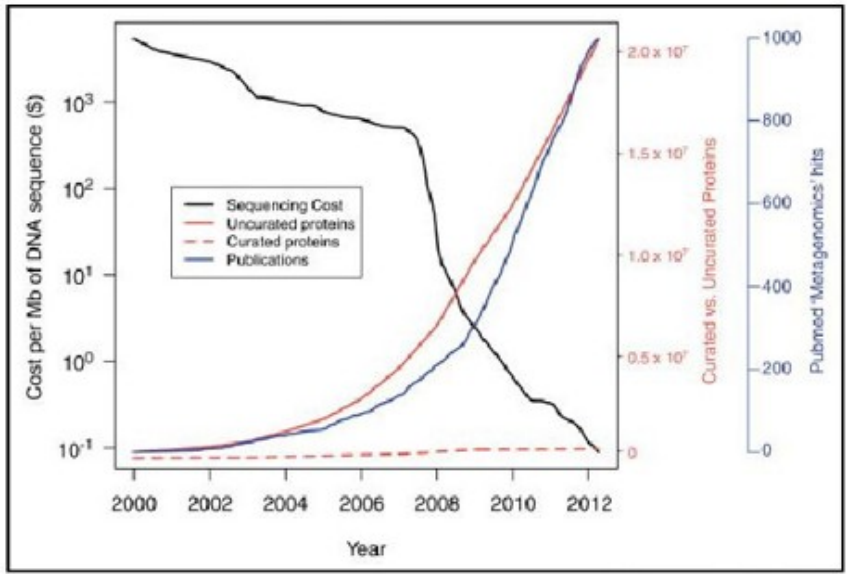


FIGURE 1

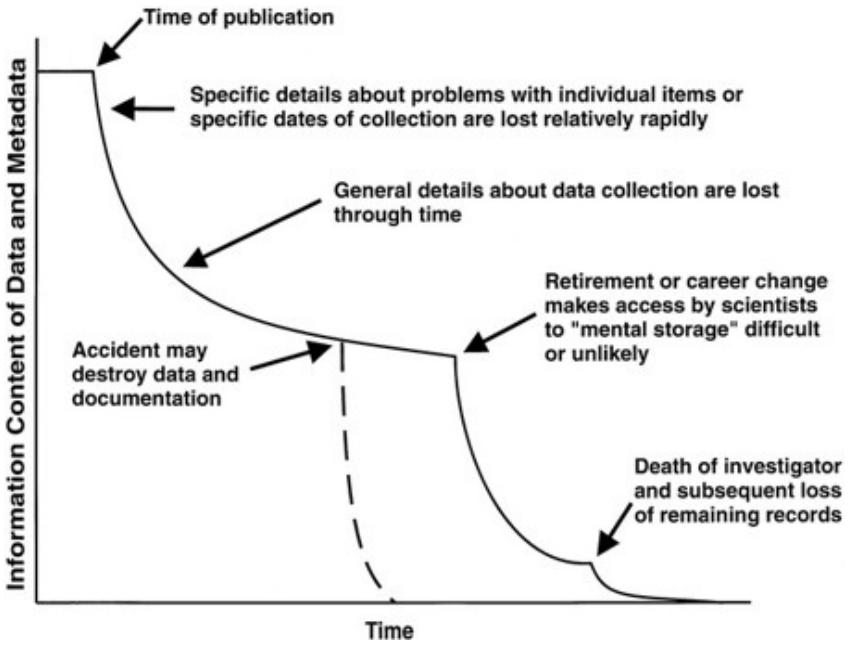
Hey (2009)

De plus en plus de données



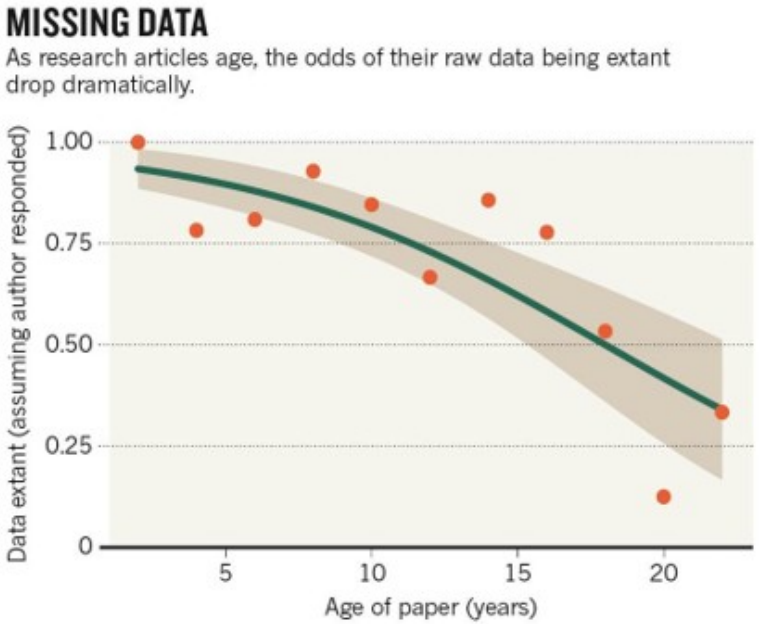
Murphy (2014)

Les ravages du temps …



Michener et al. (1997)

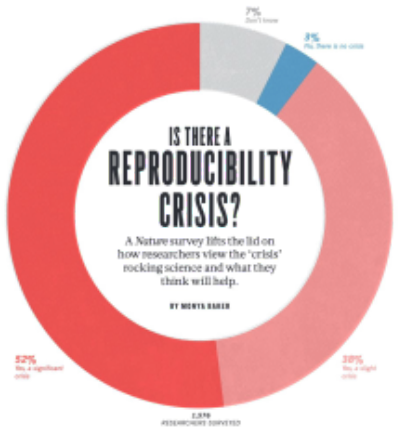
L'érosion



Gibney and Van Noorden (2013)

La reproductibilité vue depuis les laboratoires

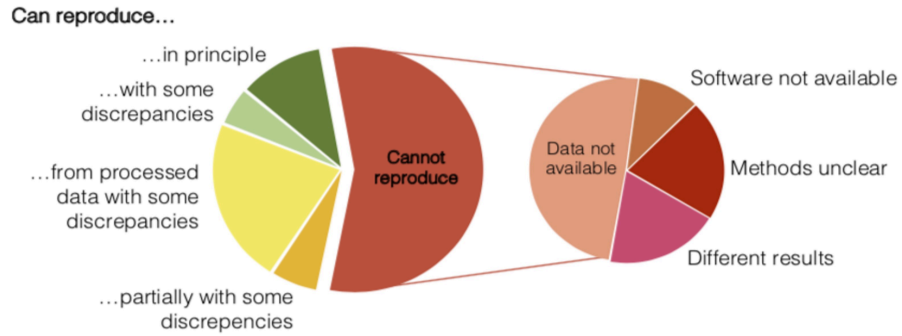
Selon un sondage mené en 2016 auprès de plus de 1 500 scientifiques plus de 70% ont déjà éprouvé des difficultés à reproduire une analyse



Baker (2016) - *Ten Years Reproducibility Challenge*

Un problème vieux comme la bioinfo

- En 2009, moins de la moitié des 18 expériences de transcriptomique publiées entre 2005 et 2006 parues dans Nature Genetics ont pu être reproduites :



Ioannidis et al. (2009)

- Sur 50 articles citant BWA en 2011, 31 ne citent ni version, ni paramètres, 26 ne donnent pas accès aux données sous-jacentes

Années 2020 : entre avancées et obstacles

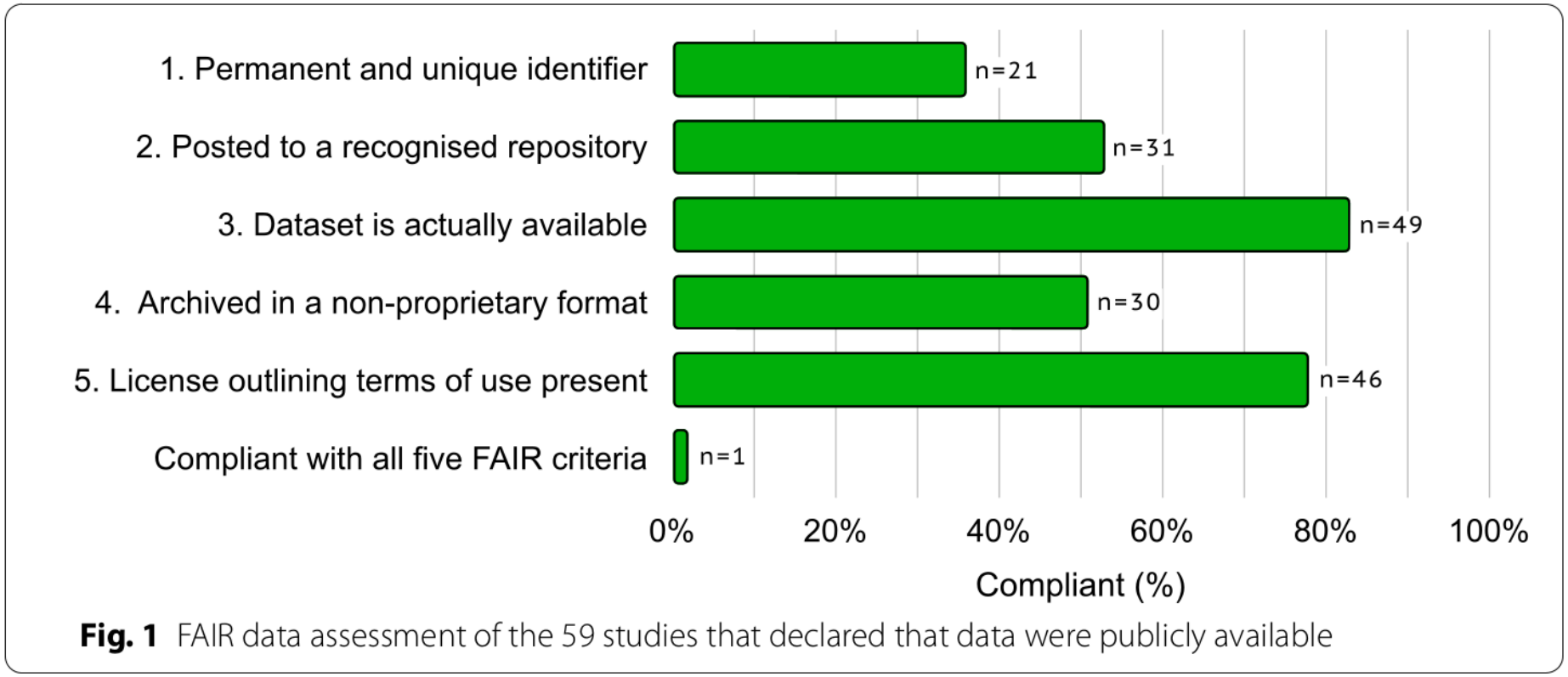


Fig. 1 FAIR data assessment of the 59 studies that declared that data were publicly available

Hamilton et al. (2022)

Quelles sont les difficultés (1/2) ?

- Problèmes **d'accès aux données** :
  - le fameux "data available upon request"
  - données brutes disponibles, mais méta-données inexistantes ou insuffisantes
- Problèmes **d'accès aux outils** :
  - outils anciens ou obsolètes
  - difficultés à installer
- Problèmes de **paramétrage de l'analyse**
  - version des outils
  - paramètres des outils
  - enchaînement des outils
- Problème d'accès aux **ressources nécessaires**
  - calcul
  - stockage

Quelles sont les difficultés (2/2) ?

- **Les pratiques scientifiques**
  - p-hacking : manipulation des données pour atteindre le seuil statistique espéré
  - HARKing : reformulation des hypothèses après l'obtention des résultats
- **Le biais de publication**
  - On ne publie "que" ce qui est nouveau
  - On ne publie "que" les résultats positifs
- **La pression de publication**
  - La culture du "Publish or Perish" incite à privilégier la quantité au détriment de la qualité

Reproductibilité & Réplicabilité (1/2)

Un résultat expérimental n'est pleinement établi que s'il peut être reproduit de manière indépendante.

- **Répétabilité** : même équipe, même conception expérimentale
- **Reproductibilité** : équipes différentes, même conception expérimentale
- **Réplicabilité** : équipes différentes, conceptions différentes

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Whitaker (2017)

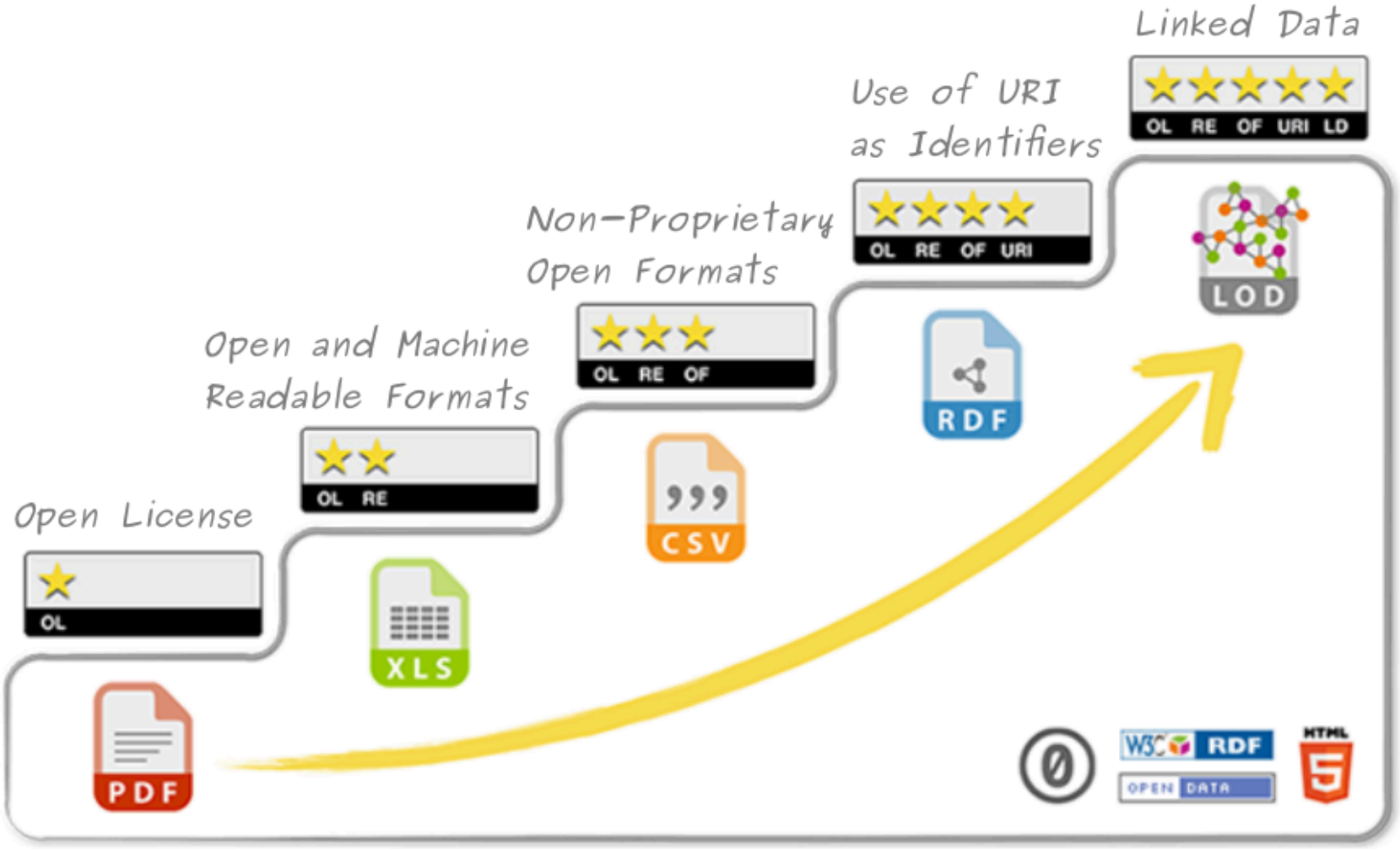
Reproductibilité & Réplicabilité (2/2)

- Il existe une ambiguïté en anglais entre réplication (*replication*) et reproduction (*reproducibility*). Derrière la *reproducibility crisis* on mélange les deux :
  - Impossibilité de répliquer des résultats de façon indépendante (psychologie, médecine, biologie…)
  - Impossibilité de reproduire des analyses à partir des mêmes données de départ

Chacun peut déjà, par la mise en place de pratiques simples et l'utilisation d'outils conviviaux, améliorer la reproductibilité de ses travaux

> Source : Allard (2018)

Les données …

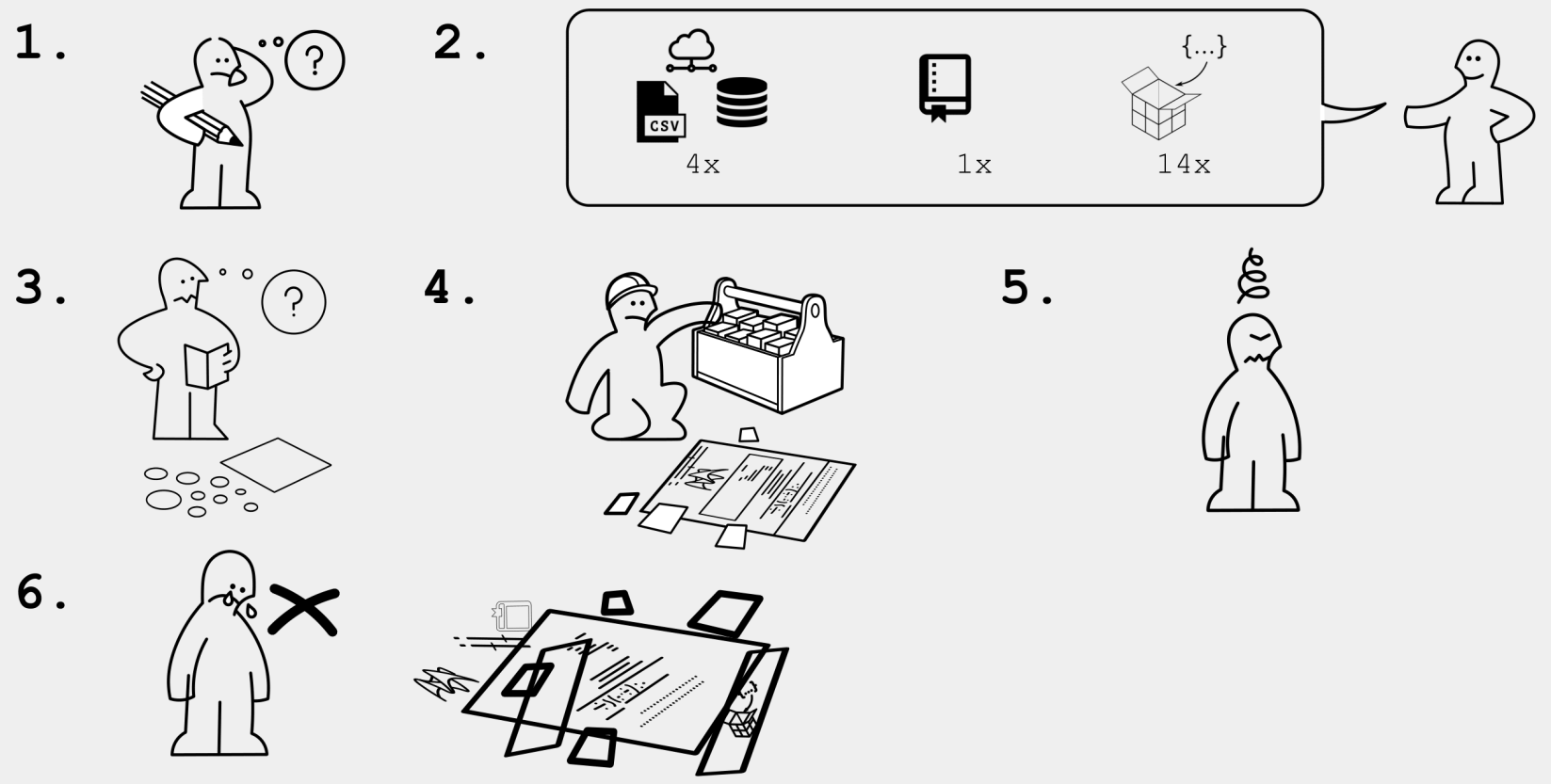


Open Data website (2012)

Les données et … les codes !

Aspect	Données	Code source
Nature	Statique – immuable une fois collectées	Évolutif – peut changer avec les versions
Actions associées	Observations, mesures, collectes	Création de connaissances, transformation d'informations, visualisation…
Modification	Généralement non modifiées,collectées dans un contexte défini	Souvent modifié, adapté, enrichi
Dépendances	Principalement indépendant, ou documenté par un protocole de collecte	S'appuie sur des dépendances et un environnement logiciel et matériel (librairies, OS, …)
Origine	Résultats d'observations ou de faits	Œuvre de l'esprit (résultat d'une création)

En pratique, qu'est- ce qu'être reproductible (1/3) ?



Ram (2019)

En pratique, qu’est- ce qu’être reproductible (2/3) ?

Avoir accès :

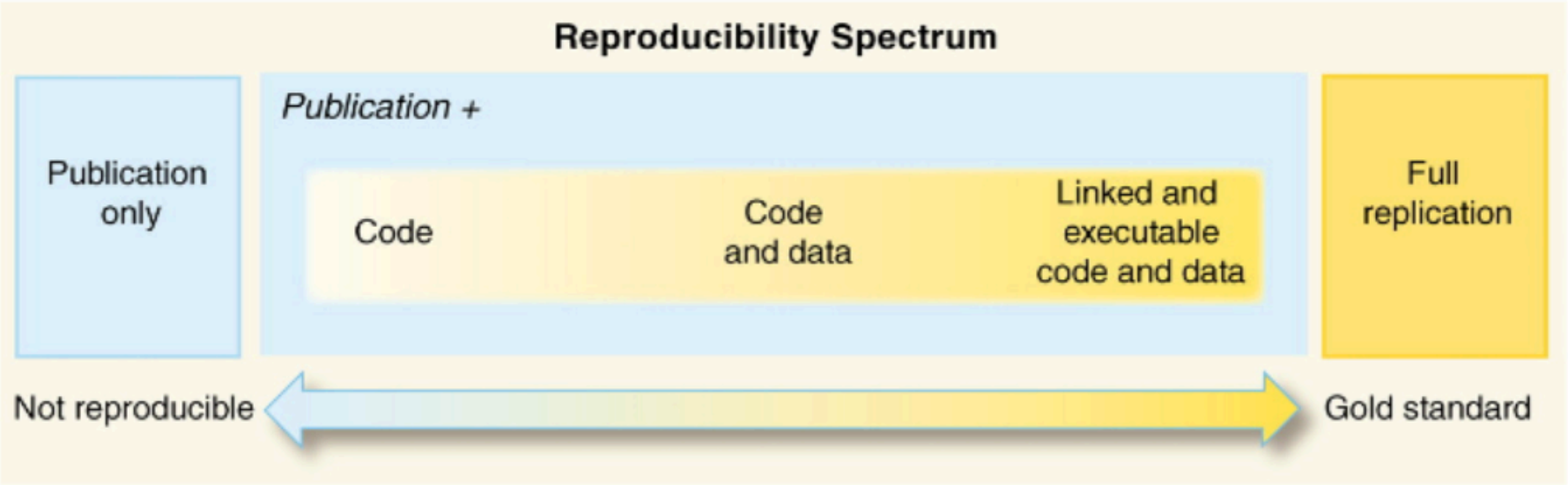
- aux pièces (les **données**)
- aux outils (les **logiciels**)
- au mode d’emploi : **paramètres, workflows d’analyse**

En pratique, qu’est- ce qu’être reproductible (3/3) ?

Mais aussi :

- à la description des pièces, de la façon dont elles ont été produites (**méta-données**)
- à la documentation technique (**choix techniques explicites**)
- au savoir faire du monteur (**formations**)
- Éventuellement à un atelier équipé pour le montage (**ressources informatiques**)

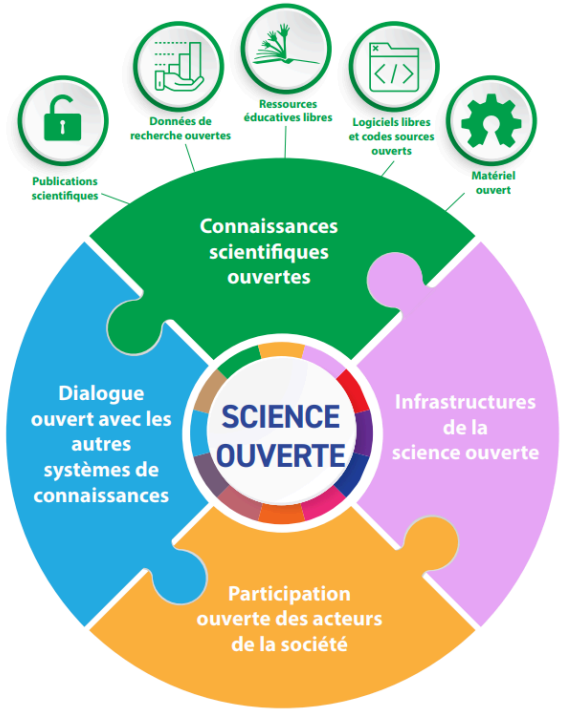
Le spectre de la reproductibilité



Piazzi et al. (2018)

Science Ouverte

Contexte



Recommandation de l'UNESCO Sur Une Science Ouverte (2021)

Rendre la recherche accessible et transparente

- Accès à la connaissance
- Accès aux méthodes
- Accès à la dissémination
- Engagement de et vers la société

Cadre juridique

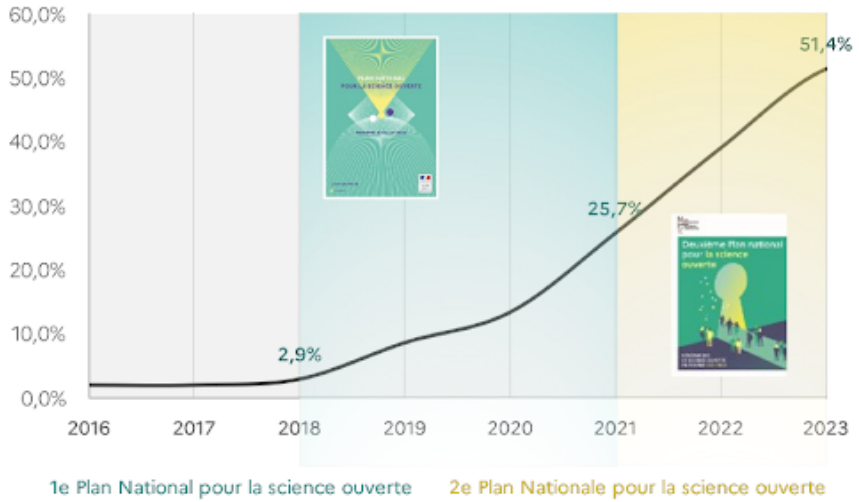
Les données de la recherche sont des informations publiques si financement public à 50% et plus.

- Principe d'ouverture par défaut et de libre utilisation :
  - 2015 **Loi Valter** : constitution d'une liste fermée d'administration pouvant fixer des redevances (IGN, Météo France, …)
  - 2016 **Loi Lemaire** : Loi pour une République Numérique

> **« aussi ouvert que possible, aussi fermé que nécessaire»**

Deux Plans Nationaux pour la Science Ouverte (2018-2021 puis 2022-2024)

- **diffusion sans entrave** des **publications** et des **données de la recherche**.
- **mobilisation du personnel** pour un **accompagnement** des équipes de recherche.



105 établissements ont répondu à l'enquête (2023-24)

ouvrirlascience.fr

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

FAIR : pré-requis à la reproductibilité

The Turing Way Community (2025)

Source : Wilkinson et al. (2016)

En pratique : Quels outils utiliser ? (1/2)

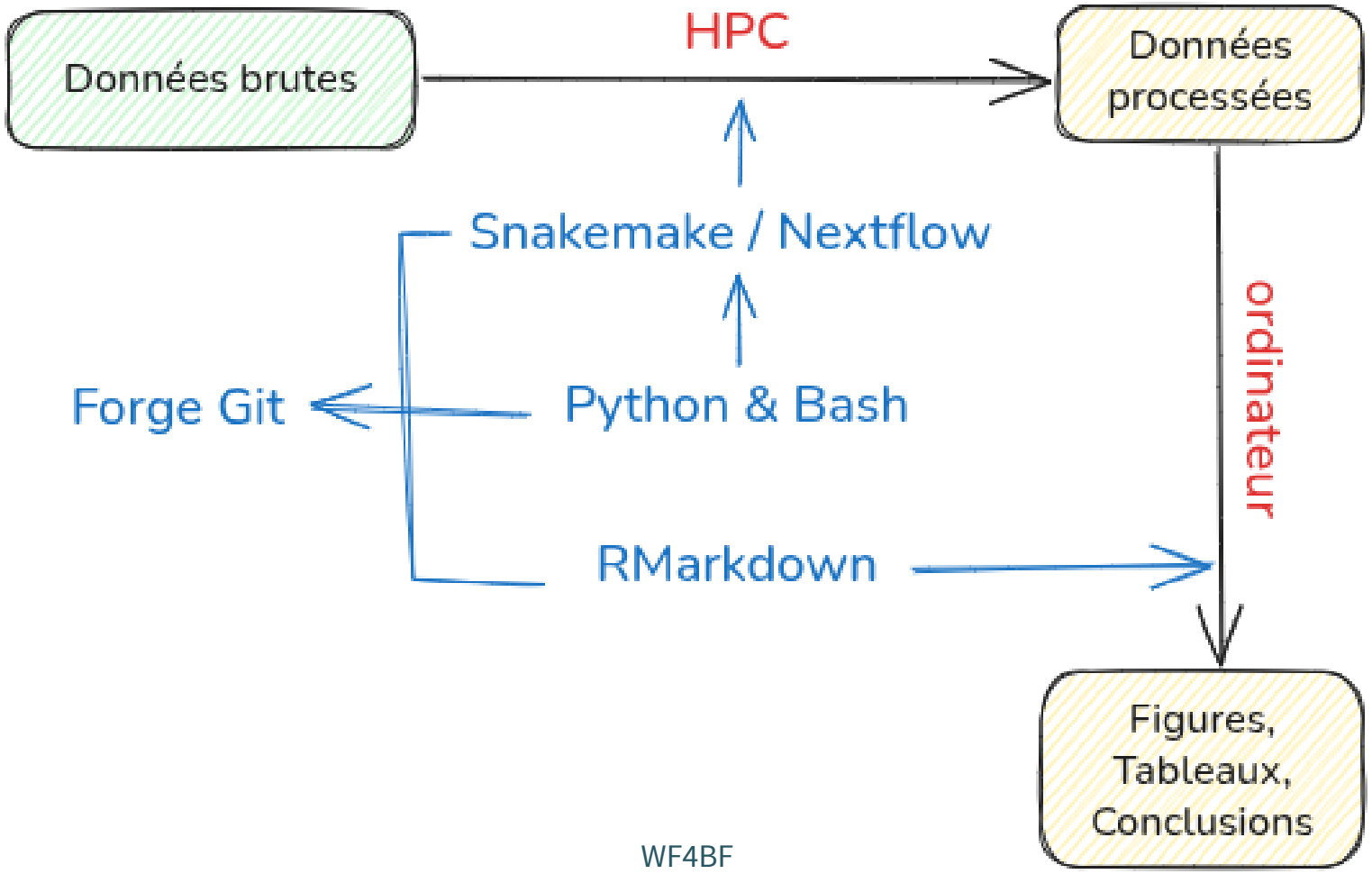
Documentation et partage des données de **manière pragmatique**

- **Documentation accrue**
  - Penser gestion de données : responsabilités, formats, cycle de vie, …
  - Penser Plan de Gestion de données : OPIDoR, DSW, DAISY, …
- **Partage des données**
  - Dépôts spécialisés internationaux : ENA, NCBI, ensembl, …
  - Plateformes généralistes : DataVerse, Figshare, Zenodo, …
- **Standardisation des outils**
  - Conda, Bioconda
  - Singularity, Docker, Apptainer
  - Machines Virtuelles

En pratique : Quels outils utiliser ? (2/2)

- Décrire son workflow d'analyse, le rendre portable :
  - Galaxy : Grüning et al. (2018)
  - Snakemake : Mölder et al. (2021)
  - Nextflow : Di Tommaso et al. (2017)
- Gérer les versions de ses codes, les publier :
  - git
  - GitHub / GitLab
  - Software Heritage & HAL
- Tracer les analyses dans des documents computationnels partageables et réutilisables :
  - Rmarkdown
  - Jupyter Notebooks

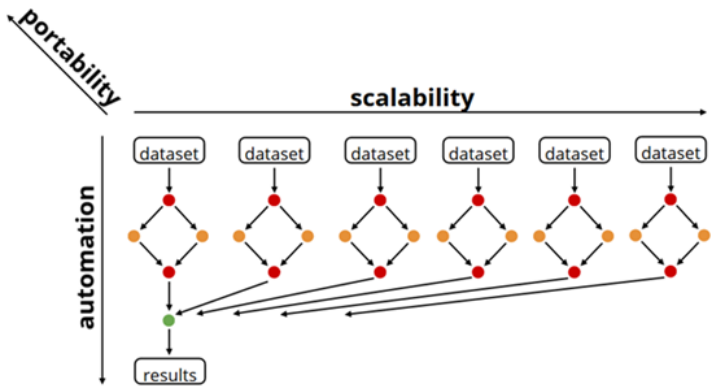
Des workflows pour une science ouverte et reproductible



Gestionnaires de workflows

Snakemake ou Nextflow pour définir de façon “simple” et modulaire des workflows d’analyse :

- **Parallélisables** : les étapes indépendantes peu vent être jouées en parallèle.
- **Reprise sur erreur** : si on refait une analyse, change un paramètre, seul ce qui doit être rejoué est relancé.
- **Portables** : un même script peut être joué en local, sur des clusters différents en changeant le fichier de configuration.
- **Gestion des dépendances** : installation des outils avec conda, apptainer, …



Exemple de `Snakefile`

Bash

```
for sample in `ls *.fastq.gz` do
    fastqc ${sample}
done
```

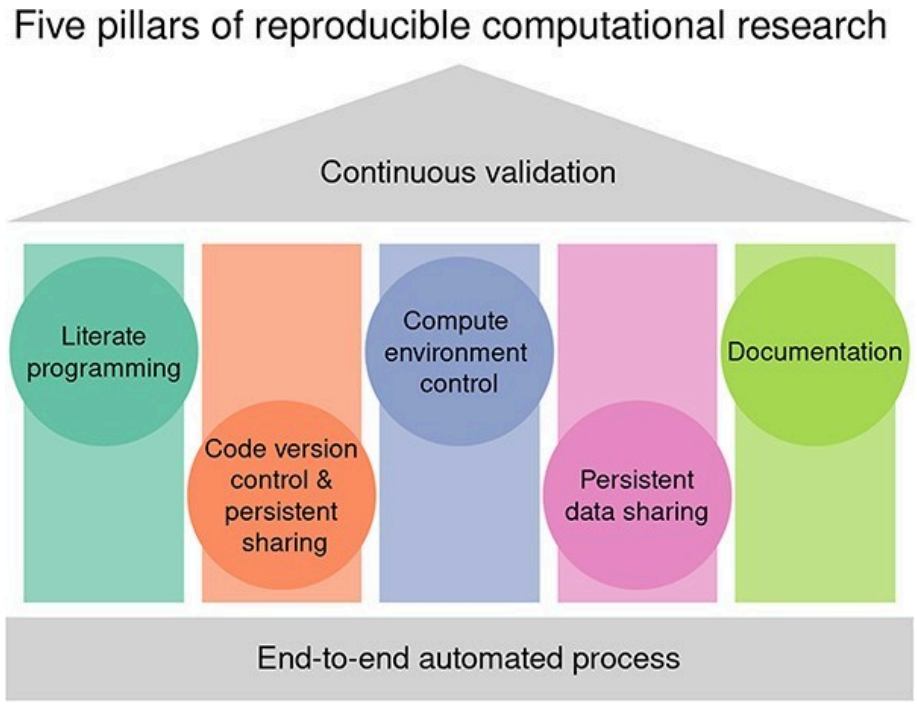
Snakefile

```
SAMPLES = glob_wildcards("./{sample}.fastq.gz")

rule final:
    input:expand("fastqc/{sample}/{sample}_fastqc.zip", sample=SAMPLES)

rule fastqc:
    input: "{sample}.fastq.gz"
    output: "fastqc/{sample}/{sample}_fastqc.zip"
    conda: "fastqc.yaml"
    message: """Quality check"""
    shell: """fastqc {input} --outdir fastqc/{wildcards.sample}"""
```

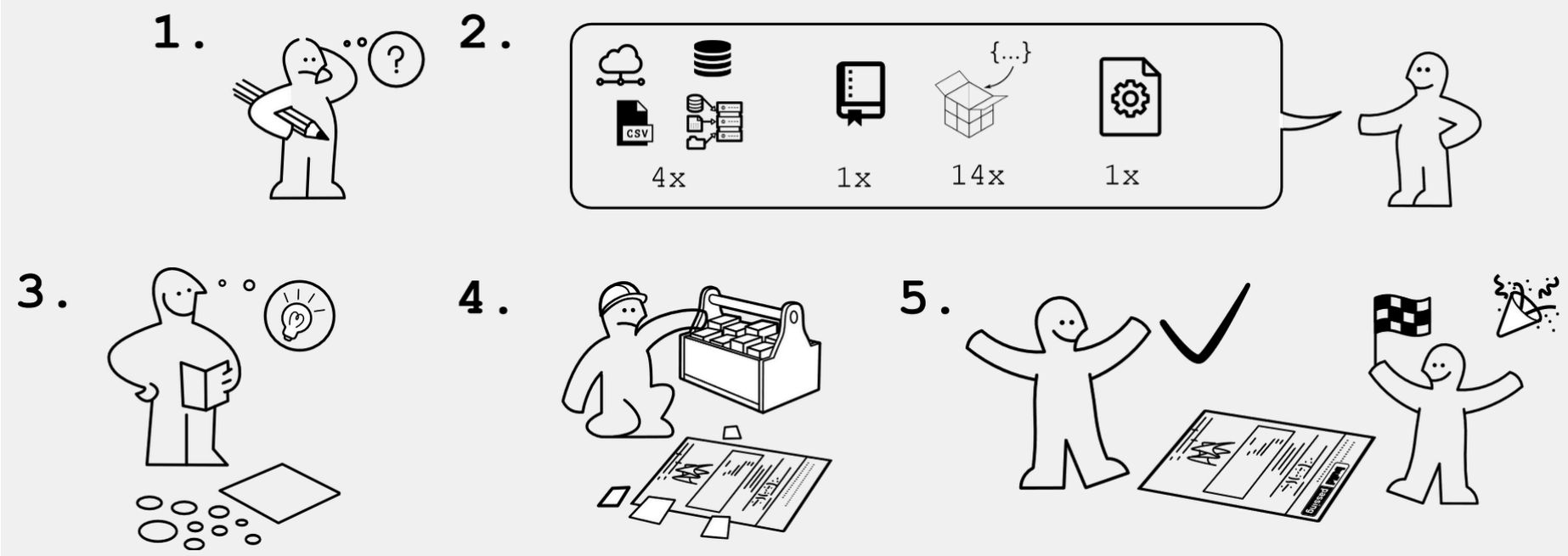
Cinq piliers de la reproductibilité



Ziemann, Poulain, and Bora (2023)

- L'irréproductibilité des études en bioinformatique demeure un problème majeur et toujours d'actualité.
- Ces cinq piliers sont un ensemble de bonnes pratiques permettant de mettre en place des flux de travail hautement reproductibles.
- L'adoption généralisée de ces principes renforcera la fiabilité de la recherche et accélérera la traduction des découvertes fondamentales en bénéfices concrets.

Épilogue



Ressources

- FUN MOOC Recherche Reproductible
- FAIR Bioinfo
- Cours Git et Github
- Github pages
- Rmd the definitive Guide
- Snakemake
- NextFlow et nf-core
- Les mémo présentés dans ce cours :
  - markdown

Sources

Allard, A. 2018. “La Crise de La Réplicabilité.” https://laviedesidees.fr/La-crise-de-la-replicabilite.html.

Baker, Monya. 2016. “1, 500 Scientists Lift the Lid on Reproducibility.” *Nature* 533 (7604): 452–54. https://doi.org/10.1038/533452a.

Di Tommaso, Paolo, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. “Nextflow Enables Reproducible Computational Workflows.” *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Gibney, Elizabeth, and Richard Van Noorden. 2013. “Scientists Losing Data at a Rapid Rate.” *Nature*, December. https://doi.org/10.1038/nature.2013.14416.

Grüning, Björn, John Chilton, Johannes Köster, Ryan Dale, Nicola Soranzo, Marius van den Beek, Jeremy Goecks, Rolf Backofen, Anton Nekrutenko, and James Taylor. 2018. “Practical Computational Reproducibility in the Life Sciences.” *Cell Systems* 6 (6): 631–35. https://doi.org/10.1016/j.cels.2018.03.014.

Hamilton, Daniel G., Matthew J. Page, Sue Finch, Sarah Everitt, and Fiona Fidler. 2022. “How Often Do Cancer Researchers Make Their Data and Code Available and What Factors Are Associated with Sharing?” *BMC Medicine* 20 (1): 438. https://doi.org/10.1186/s12916-022-02644-2.