Institut de Recherche pour le Développement FRANCE

MiVEGEC
Maladies Infectieuses et Vecteurs
Écologie, Génétique, Évolution et Contrôle

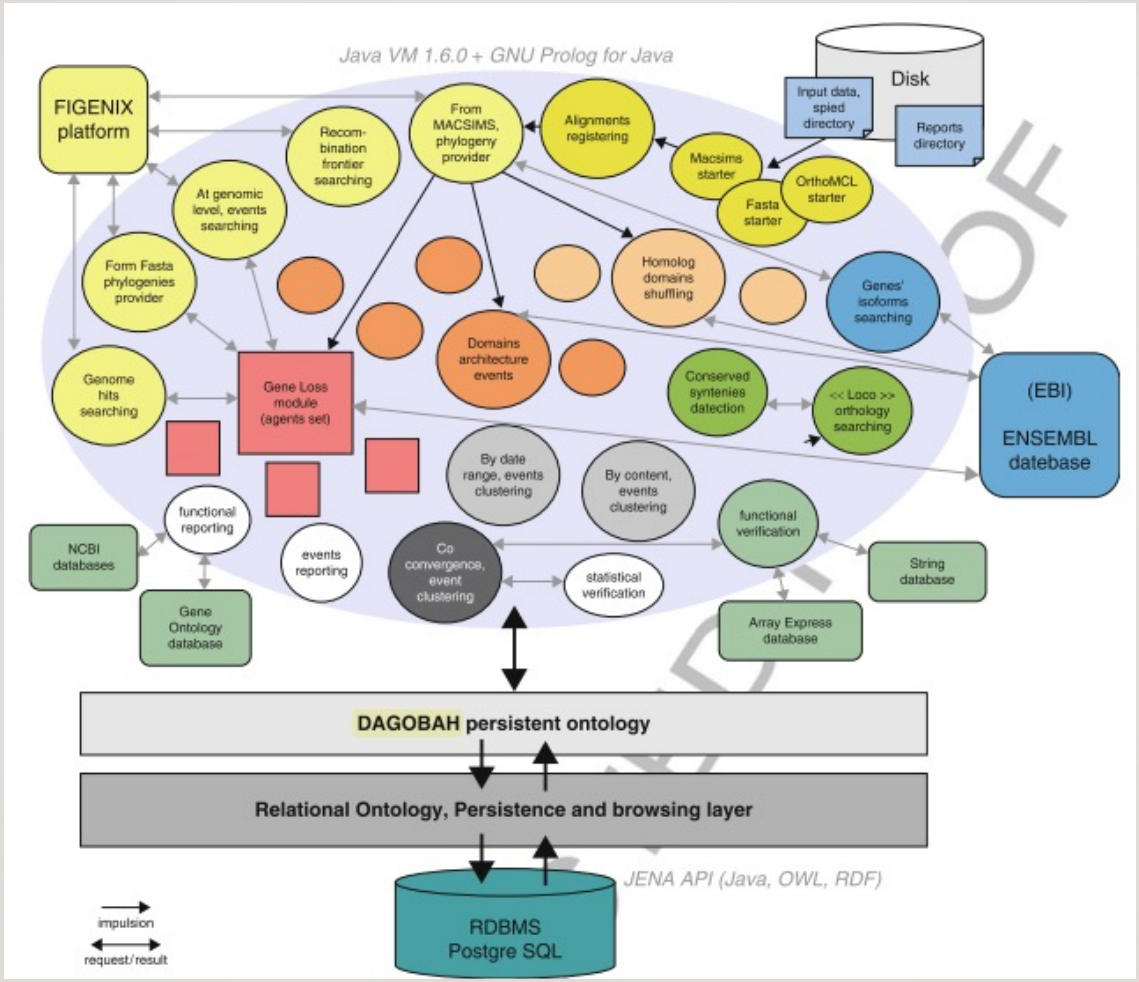
# REPRODUCIBILITY

## TESTIMONIAL

Juke34
@JacquesDainat
Jacques Dainat Ph.D.

# Testimonial as a learner

# During my thesis (2008-2012)



# What I thought I made

# What I actually did

# What I actually did


Difficulty

- to track bug
- to update tools
- to collaborate on the code
- to share (image >20Go)
- etc

Reproducibility is nothing withtout comprehensiveness

Version control system was quiet new. (Few around was using it, not learnt at school)
Conda / Container didn’t exist.

# Programming / Development < 2013

=> naming/ versioning

```
script.py
scriptFinal.py
scriptFinalV2.py
scriptUSE_THIS_ONE.py
scriptV2.py
```

=> comment

```
#################
# Script v1. 07/04/2007
# Author: Jacques Dainat
# Modified by JD 15/02/2008 Add feature1
# Modified by JD  27/11/2008 Add feature2
# Modified by Colleague1  11/08/2009 modify feature1 to fix error
```

=> readme

```
---- Script ----
Purpose
Installation
Usage
```

# Programming / Development > 2013

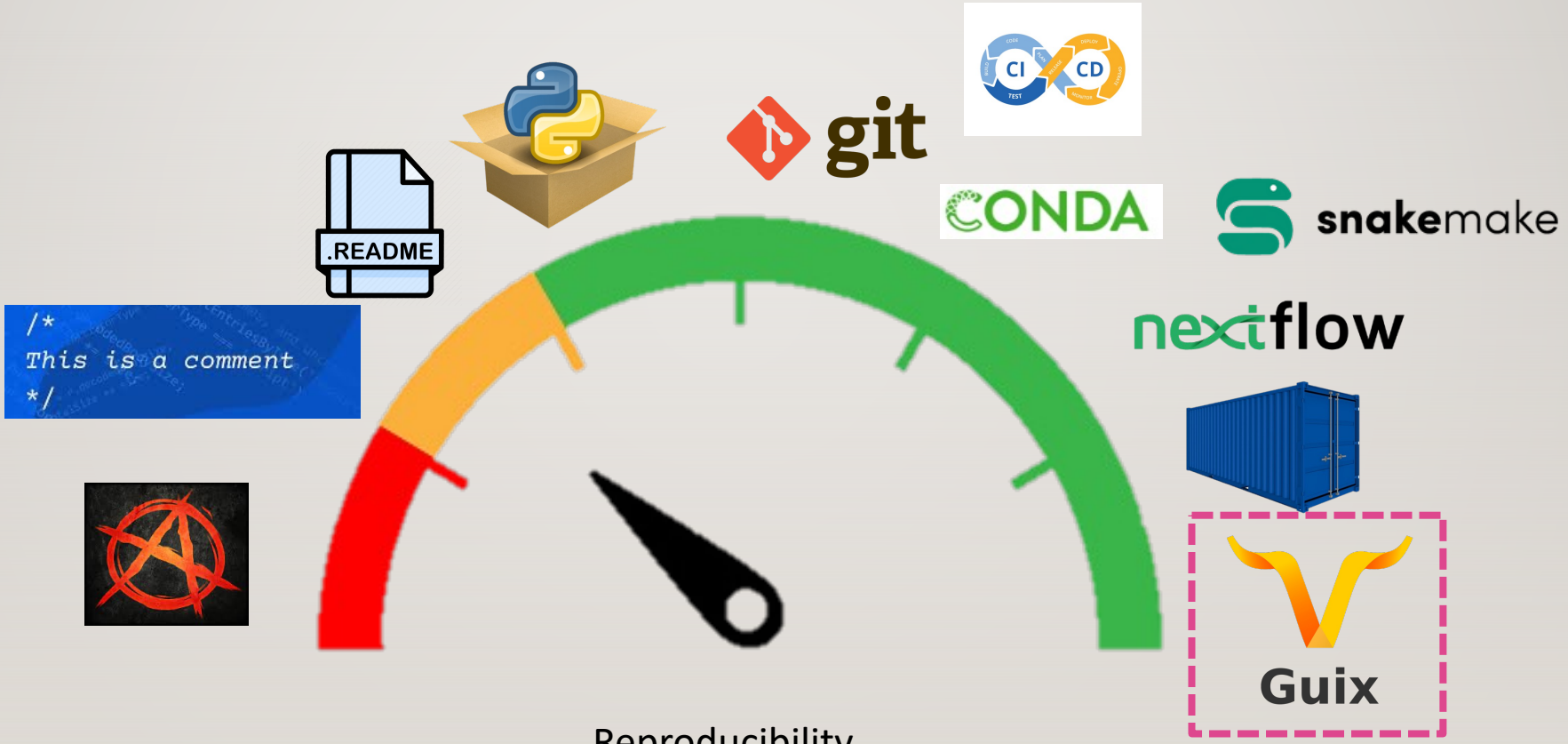

- naming
- comment => 
- Readme
  Doc

MARKDOWN

Wiki

# Programming / Development

- Many tool available depending your needs (level of reproducibility)
- Keep an eye on new tools (future game changer), and decide when to make the step



# Dependencies HELL

1995 CPAN

2007 Virtualbox

2008 Homebrew

2011 Pip

2012 Conda

2013 Docker / Guix

2015 Singularity

2006 CPAN

2011 Virtualbox

2014 Homebrew

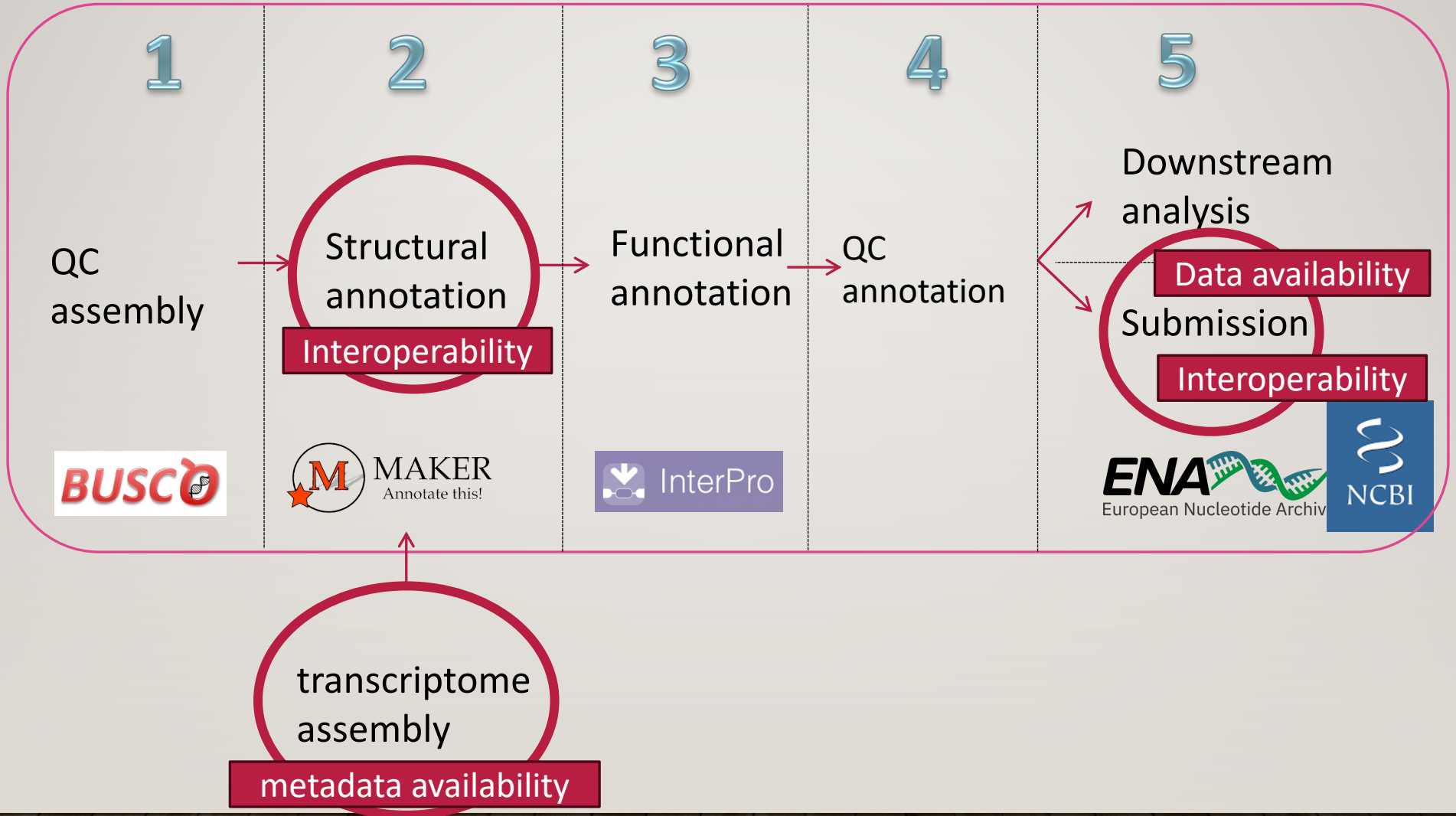
2016 Conda/Pip

2018 Docker / Singularity

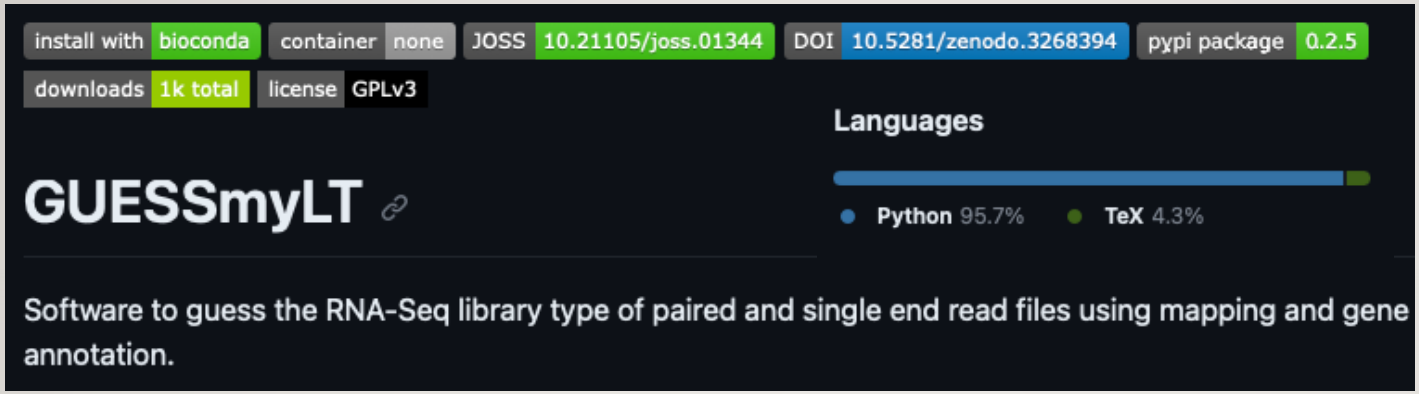
# DATA

- on demand
- on hard drive

# The main steps in genome annotation

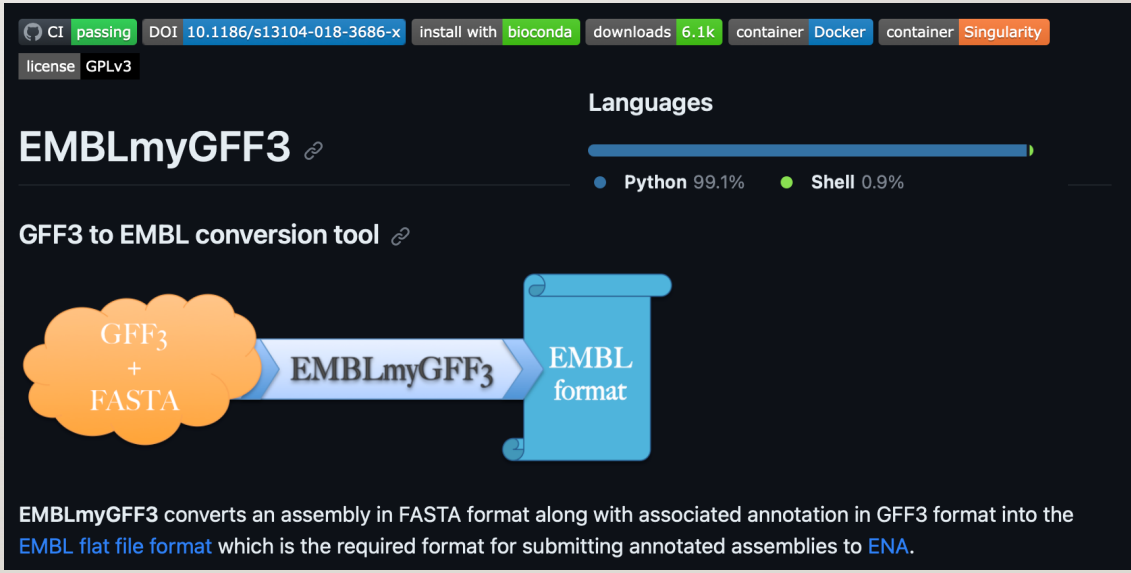


# DATA

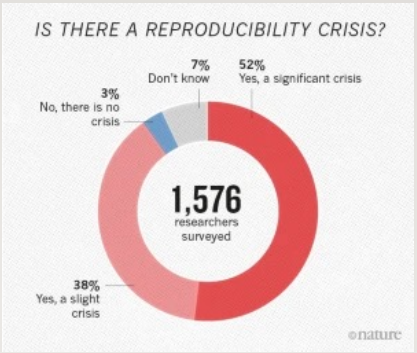



! Data submitted to ENA was not exactly what ended up in the DB…

Discrepancies between publication and data in public archive for the functional annotation.




# DATA




Guiding Principles for scientific data management and stewardship

(https://doi.org/10.1038/533452a)

**F**indable

**A**ccessible

**I**nteroperable

**R**eusable


elixir

Institut de Recherche pour le Développement FRANCE

MIVEGEC Maladies Infectieuses et Vecteurs: Ecologie, Génétique, Évolution et Contrôle

D M P

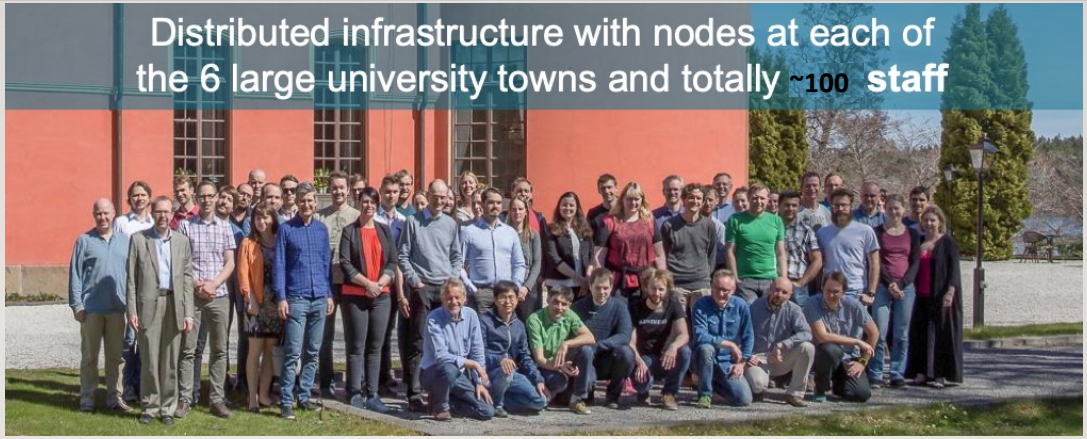
Data Management Planning

(ANR 2019)

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

2014-2020

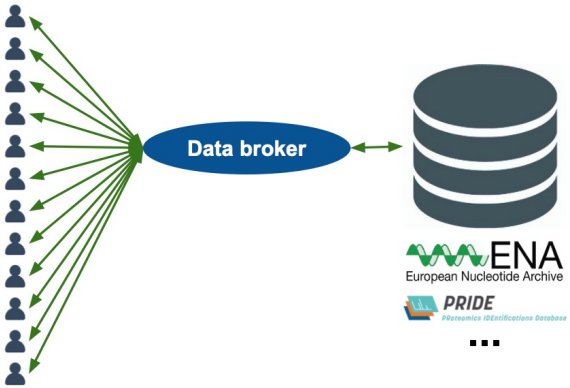
## When I think about it




4 FTE staff dedicated to data managment

13 FTE Systems development

Un *data broker*




Intermédiaire entre le producteur de données et la ressource internationale de stockage/archivage

Rationalisation des échanges

35 FTE Support

12 FTE Training

...

# Testimonial as a teacher


## Tools for Reproducible Research

10 courses since may 2019

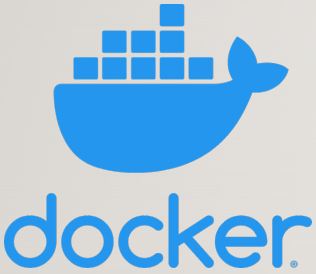
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN


Institut de Recherche pour le Développement FRANCE

MiVEGEC
Maladies Infectieuses et Vecteurs: Ecologie, Génétique, Évolution et Contrôle

# Tools covered









Tools are one thing, but you need to know how to you use them properly…

# Make the course your own

(check license!)

- Git hosted static website
  Rendering: ReadThedocs => Mkdocs
- Lectures: .Rmd compiled => PPTX
- Nextflow Snakemake 50/50 => 20/80 => 0/100
- We choose to ask help from IFB via Thomas Denecker
- We added Quarto
- We choose to install tools during the course (conda, then conda env)

# Make the course your own

It is time expensive

- numerous meeting
- split course according to knowledge of each teacher
- check material
- update slides
- adapt material (command git, problem with conda)

# What we experienced

Surprised by the appeal of data management (Round table)


We did not have expert for that and questions were beyond our expertise.

# What we experienced

Conda was a mess!!!
(Remind: Conda share pre-compiled binaries.)

=> We should have prepared a environment file and test them for each OS (windows, linux, macOS).

<u>architecture</u>

**VS**

**x86**


**ARM (Apple Silicon)**

=> We checked conda was working on ARM but didn't check that all dependencies were available for this architecture.

=> All the course was based on conda environment 😱

# What we experienced


**ARM (Apple Silicon)**

- Version beta but worked fine

- Must install UTM (VM) to install a Rocky Linux (Linux OS ARM architecture) to install singularity

# What we experienced


**ARM (Apple Silicon)**

Images from biocontainer are not all ARM compatible (cross-plartform compilation)

Emulation of x86 platform can make tools very slow
Rosetta2 for emulation works great (but need to be present on the macOS computer)

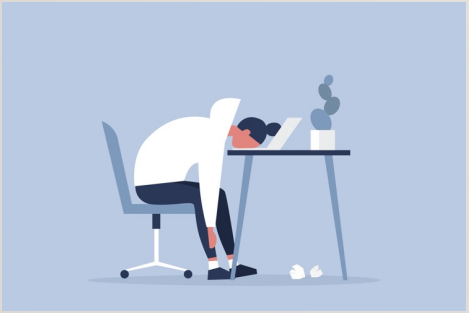
# Avoid the pitfalls

- teach conda but avoid to use it outside this part of the course
- Teach container early in the course
  => Use container for all part of the course
- Introduce computer processors architectures?
- Teach  Cross-Compilation container (or check rosetta2 presence for macOS users)
- Test (CI?) on all

# Overall impression

- Greater cohesion within the team that set up the course
- Networking with students, useful feedback
- Transfer of skills between teachers
- Awareness of reproducibility limits

Rewards



Is setting a second session of the same course as demanding?


Jacques.dainat@ird.fr