# Reproducibility for Bioinformatic tools and workflows

Authors: Peter Crisp, UMN, USA (0000-0002-3655-0130); Diep Ganguly, ANU, AUS (0000-0001-6746-0181)

# Topic →

# Challenges →

## Solutions →

#### Resources -

#### Resources & further reading

- → Reproducible computational research Sandve et al. (2013), PloS Comp Bio, https://doi.org/10.1371/journal.pcbi.1003285
- → **Software**: reproducible installation Mangul S et al. (2018), BioRxiv,

https://doi.org/10.1101/452532

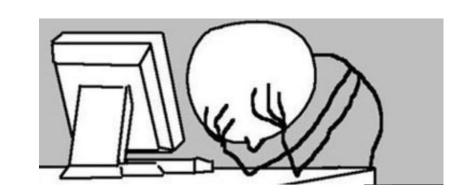
- → Documentation
- ◆ 'Bioinformatic data skills' Buffalo http://shop.oreilly.com/product/0636920 <u>030157.do</u>
- Software carpentry <a href="https://software-">https://software-</a> carpentry.org/
- → Version control:

http://smutch.github.io/VersionControlTutor

- → Containers:
  - ◆ Docker <a href="https://docs.docker.com/">https://docs.docker.com/</a>
  - ◆ Biocontainers <a href="http://biocontainers.pro">http://biocontainers.pro</a>
- → Reviewing:
  - Computation checklist github.com/vivekbhr/reproChecklist

#### Documentation

#### Why did I do this?



How can I record the steps of a bioinformatic analysis?

#### Notebooks

Notebooks are an essential tool for documenting analyses, enabling reproducibility and sharing.

- Keep track of analysis
- Interactive coding
- Interactive data exploration
- Imbedded visualization

Name -

Data.ipynb

Fasta.ipynb

Julia.ipynb

R.ipynb

iris.csv

(:) lightning.json

lorenz.py

Last Modified

a day ago

a day ago

a day ago

9 days ago

3 minutes ago

- Easy access to docstrings Mix of code and documentation
- Over 40 programming languages
- Easily shared
- Widgets
- Interactive plots
- Run remotely on server

### Version Control

What version of the program, data etc... did I use?



Version conflict

#### Git

Git tracks changes to files over time, enabling documentation of development and reversion to prior versions.

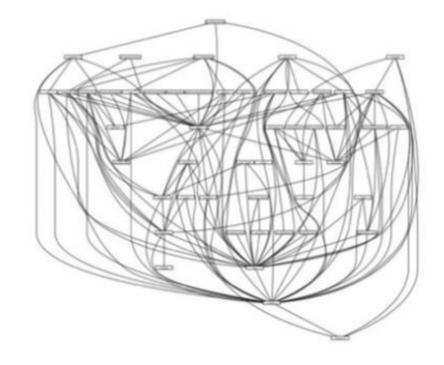
- Records and illustrates changes between versions
- Lets you share code and collaborate easily
- Active community with lots of tutorials and help pages
- Provides a backup for all your code to be accessed remotely as needed

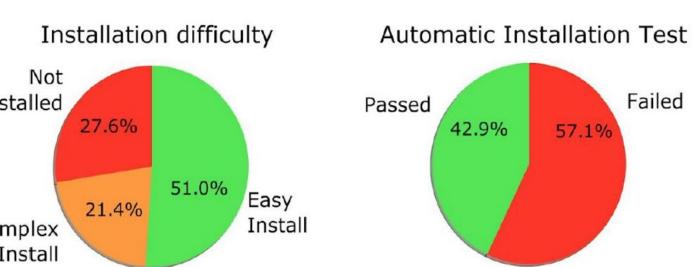
git

#### Software

#### Dependency hell

# install Installation difficulty





28% of the tools are impossible to

https://www.biorxiv.org/content/early/2018/10/25/452532

### Containers

Managers and containers enable consistent use of software when emulating or reproducing an analysis.

- Handles installs and dependencies
- Allows for multiple independent environments
- Easily configurable
- Allows for manual installs as well
- Runs on all three major systems
- Open source
- You can package your own work and contribute
- Docker runs images as containers that are
- self contained with all code, programs, libraries included. No subsequent installation required.
- Isolated
- Portable including dissemination
- Lightweight



https://jupyter.org/documentation

B + % □ □ ▶ ■ C Code

beta 2.67

rho \_\_\_\_\_\_ 28.00

from lorenz import solve\_lorenz



https://www.rstudio.com/



Python 3

Lorenz.ipynb X ■ Terminal 1 X ■ Console 1 X ■ Data.ipynb X ■ README.md X

In this Notebook we explore the Lorenz system of differential equations

 $\dot{x} = \sigma(y - x)$ 

 $\dot{z} = -\beta z + xy$ 

 $\dot{y} = \rho x - y - xz$ 

Let's call the function once to view the solutions. For this set of parameters, we see the trajectories swirling around two points,

# prepare the axes limit

ax.set\_xlim((-25, 25)) ax.set\_ylim((-35, 35)) ax.set\_zlim((5, 55))

def solve\_lorenz(N=10, max\_time=4.0, sigma=10.0, beta=8./3, rho=28.0):

def lorenz\_deriv(x\_y\_z, t0, sigma=sigma, beta=beta, rho=rho):

return [sigma \* (y - x), x \* (rho - z) - y, x \* y - beta \* z]

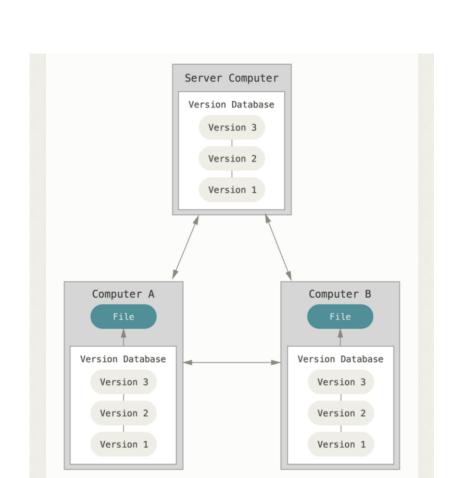
# Choose random starting points, uniformly distributed from -15 to 15

ax = fig.add\_axes([0, 0, 1, 1], projection='3d')

 $x\theta = -15 + 30 * np.random.random((N, 3))$ 



https://git-scm.com/doc









Turns a GitHub repo with data and notebooks into a collection of interactive notebooks run in the cloud



CO CODE OCEAN

Configuration, preservation, & reuse of executable code using containers for researchers





