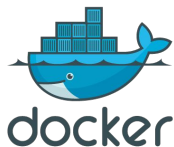


Bioinformatics Skills for Microbial Genomics: Getting started with Linux containers and Nextflow workflows

2nd of February 2022



Course overview



- **One day introduction course**
- **Using containers** – Containers, such as Docker and Singularity.
- **Workflow languages (Nextflow DSL2)** – workflow managers, like NextFlow, provide a framework for running analyses.
- **GNU/Linux command-line** - All other concepts depend on using the command-line.
- **Work in pairs (or small groups)** on a cloud virtual machine provided by CLIMB-BIG DATA (QIB)

```
@16CL0000

Date and time           = Fri Jan 28 15:48:05 GMT 2022
Hostname                = chomp
Uptime                  = 15:48:05 up 227 days
Load avg.               = 0.05, 0.07, 0.02
Release                 = Ubuntu 18.04.4 LTS
Kernel                  = 4.15.0-144-generic GNU/Linux
CPU Usage (Core)        = 0%
Memory                  = 3673/61923MB
Swap                    = 0/0MB
Disk(/)                 = 57G/113G
Conda                   =

-----
Mount(s):
Filesystem              Size  Used Avail Use% Mounted on
udev                    31G   0   31G   0% /dev
udev                    31G   0   31G   0% /dev

If you need any support, please contact the Bioinformatics and Informatics support group
We are at zone Orange level 2, email: bioinformatics@quadram.ac.uk
Slack Chat               https://quadraminstitute.slack.com/archives/CRLG4B5E (#cloud channel)
Bioinformatics helpdesk  https://bioinformatics.quadram.ac.uk
Bioinformatics wiki      https://bioinformatics.quadram.ac.uk/confluence/display/BSUP

You have new mail.
Last login: Fri Jan 28 12:02:26 2022 from 149.155.192.89
ubuntu@chomp:~$
```



Course prerequisites and outcomes

- You will need a basic understanding of navigating the GNU/Linux command line. You should be able to use commands such as *cd*, *ls*, *cat*, *grep*.
- You will need a basic understanding of microbial genomics.
- You will need a stable internet connection and a web browser
- You will need a Two-factor authentication (2FA) application.

See discord
#learning-linux

- You will learn about how bioinformaticians organise their data and analysis.
- You will learn how to deploy bioinformatics software through Linux containers.
- You will be introduced to chaining bioinformatics software to run in a “pipeline” via NextFlow and Snakemake.
- You will be introduced to writing your own workflows using existing NextFlow modules.
- You will learn how to use these frameworks to run regular bioinformatics analyses.



CLIMB-BIG-DATA

Cloud Environment with root access to Virtual Machines with pre-installed Operating System and software.



7 academic partners: Birmingham, Cardiff, Swansea, Warwick, QIB, Leicester, Bath, LSHTM



Data security, even for clinical applications



Graphic Processing Units
Enhanced storage



Research Software Engineers



Tools for enhanced sharing of software
and data



Integration with external
facilities

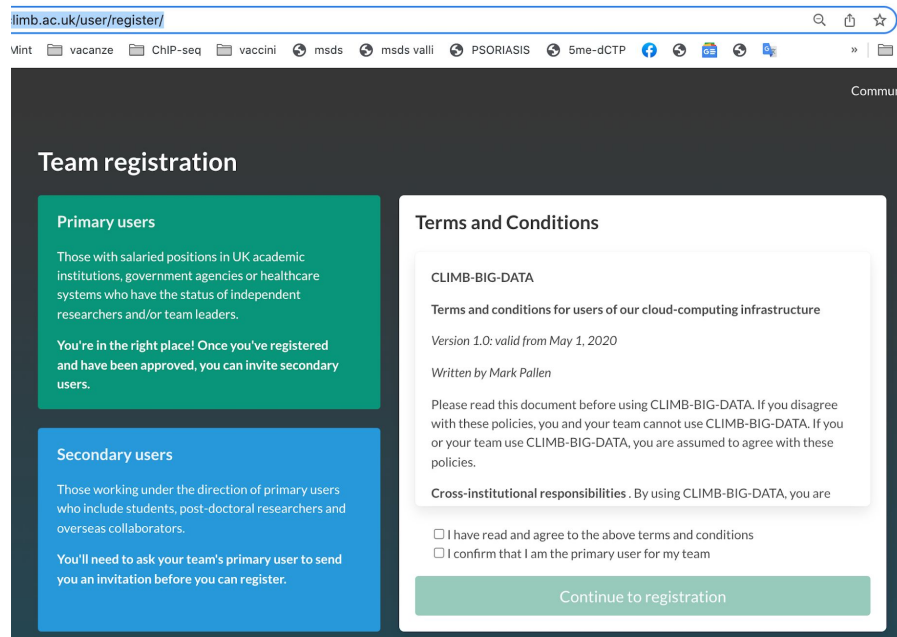
How to apply for an account (free-tier)

- <https://www.climb.ac.uk/>
- <https://bryn.climb.ac.uk/user/register/>

A member of the CLIMB team will verify the application before granting access to the infrastructure

Primary users:

“Those with salaried positions in UK academic institutions, government agencies or healthcare systems who have the status of independent researchers and/or team leaders.”



The screenshot shows a web browser window with the URL limb.ac.uk/user/register/. The page is titled "Team registration" and contains two main sections: "Primary users" and "Secondary users".

Primary users

Those with salaried positions in UK academic institutions, government agencies or healthcare systems who have the status of independent researchers and/or team leaders.

You're in the right place! Once you've registered and have been approved, you can invite secondary users.

Secondary users

Those working under the direction of primary users who include students, post-doctoral researchers and overseas collaborators.

You'll need to ask your team's primary user to send you an invitation before you can register.

Terms and Conditions

CLIMB-BIG-DATA

Terms and conditions for users of our cloud-computing infrastructure

Version 1.0: valid from May 1, 2020

Written by Mark Pallen

Please read this document before using CLIMB-BIG-DATA. If you disagree with these policies, you and your team cannot use CLIMB-BIG-DATA. If you or your team use CLIMB-BIG-DATA, you are assumed to agree with these policies.

Cross-institutional responsibilities. By using CLIMB-BIG-DATA, you are

☐ I have read and agree to the above terms and conditions

☐ I confirm that I am the primary user for my team

[Continue to registration](#)

How to apply for an account (paid-tier)

Intensive and sustained use of CLIMB-BIG-DATA:

- 1% - 2% of grant budget
- £10/week per training VM
- £1,000/week for Research Software Engineer
- Training events (fee to attend)
- PhD student subscriptions available
- Personalised quotations available (email climb-big-data@quadram.ac.uk)



Meet your course coordinators



Anna Price



Mavis
Foster-Nyarko



Lisa Marchioreto



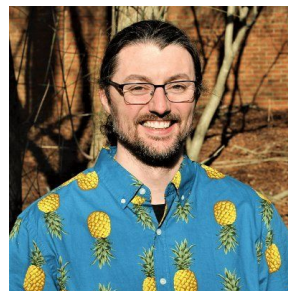
Andrea Telatin



Thanh Le Viet



Nabil-Fareed Alikhan



Robert Petit III

So where you are dialling
in from ? Where are you
based? Let us know (in
discord
#modern-bioinformatics)



Schedule

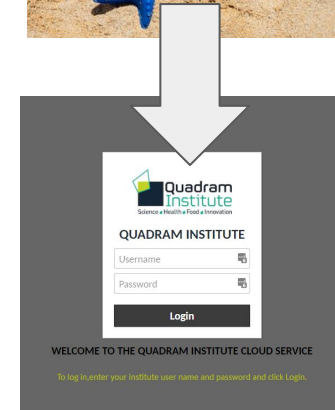
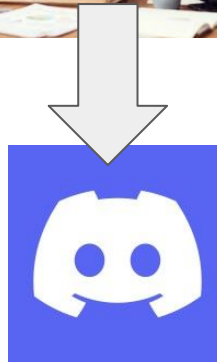
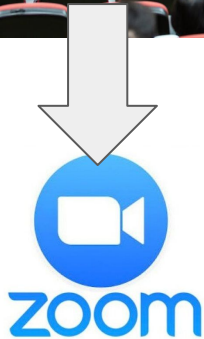
Difficulty



Time (GMT)	Item	Speaker/Chair
9:00	Orientation and testing virtual machines	
10:00	Formal welcome	Organising committee
10:10	Lecture: How does a modern bioinformatician organise their work?	Nabil-Fareed Alikhan
10:50	Lecture: Getting things done with Conda and Snakemake	Anna Price
11:30	Lecture: The value and use of containers	Anna Price
12:00	Lunch time break	
13:00	Practical session 1 - Assemble and examine a microbial genome using containers	Anna Price
14:30	Lecture: Provenance and portability through Nextflow	Andrea Telatin
15:00	Practical session 2 - Basic bioinformatics using Nextflow	Andrea Telatin + Nabil-Fareed Alikhan
16:20	Afternoon break	
16:50	Lecture: Working with Nextflow, DSL2 modules and Bactopia	Robert Petit
17:20	Discussion panel	All
18:00	Final remarks	Organising committee
18:10	End of workshop	



Virtual platforms for the course



Discord and Zoom

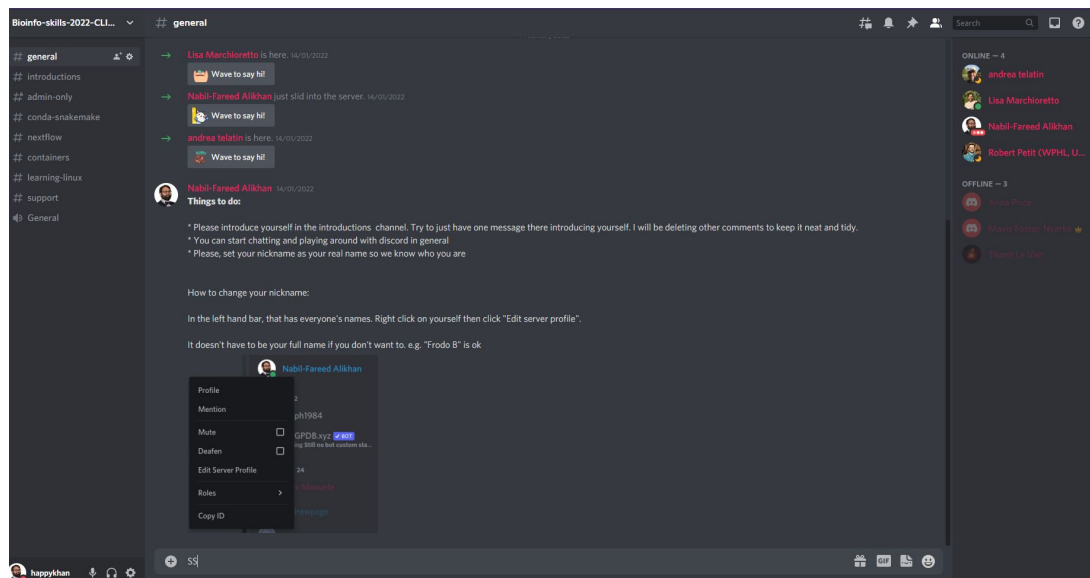
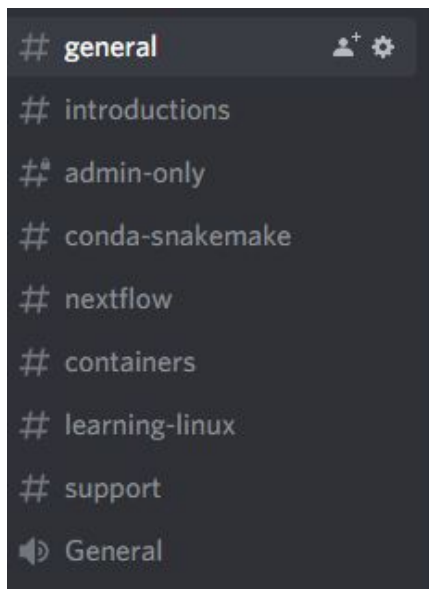
- Orientation session - You should already have access to both platforms
- Keep discussion and questions on Discord (Avoid Zoom chat). Don't be shy: ask questions!
- Moderators will take your questions and ask them during lecture sessions
- Be respectful
- Talks will be recorded
- Material will be made available online
- Contact coordinators for any issues



Discord

Voice over IP, instant messaging platform - A mix of IRC, chat room, Skype

Channels




CLIMB BIG DATA - Virtual machines

- Orientation session - Coordinators should be assigning you into groups
- You should have the “Access QIB Cloud User Guide” document
- How to access:
 - Sign-in
 - 2FA
 - Screenshare

ALL CONNECTIONS

- 112CPU1.6TB
- AndrewTest
- bio-comp-workshop-2
- bio-skills-1
- bio-skills-1-old
- bio-skills-10
- bio-skills-11
- bio-skills-12
- bio-skills-13
- bio-skills-14
- bio-skills-15
- bio-skills-16
- bio-skills-17
- bio-skills-18
- bio-skills-19
- bio-skills-2
- bio-skills-20
- bio-skills-21
- bio-skills-22
- bio-skills-23



```
@16CL0000

Date and time      = Fri Jan 28 15:59:13 GMT 2022
Hostname           = bio-skills-11
Uptime             = 15:59:13 up 2 days
Load avg.          = 0.00, 0.00, 0.00
Release            = Ubuntu 20.04.3 LTS
Kernel             = 5.4.0-94-generic GNU/Linux
CPU Usage (Core)   = 0%
Memory             = 263/31356MB
Swap               = 0/0MB
Disk(/)            = 5.2G/75G
conda               = conda 4.11.0

-----
(s):

You can use mamba in place of conda to install bioinformatics software, for example: mamba install minimap2
If you need any support, please contact the Bioinformatics and Informatics support group
We are at zone Orange level 2, email: bioinformatics@quadrant.ac.uk
Slack Chat          https://quadrant.ac.uk/slack/ (channel)
Bioinformatics helpdesk https://bioinformatics.quadrant.ac.uk
Bioinformatics wiki  https://bioinformatics.quadrant.ac.uk/confluence/display/BSUP

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@bio-skills-11:~$
```



CLIMB BIG DATA - 2FA

Multi-factor authentication has been enabled on your account.

To complete the enrollment process, scan the barcode below with the two-factor authentication app on your phone or device.



► Details: [Show](#)

After scanning the barcode, enter the 6-digit authentication code displayed to verify that enrollment was successful.

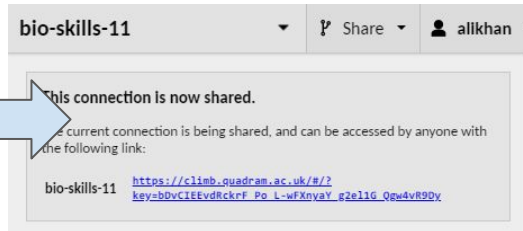
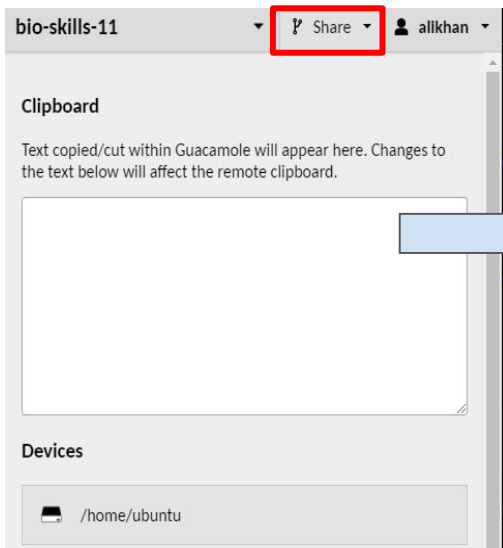
Continue

For the first time you log on to the service, you will be asked to enrol the second authentication factor (2FA) for your account. You need to have a 2FA client installed on your phone or computer for enrollment. We recommend Authy [<https://authy.com/download/>], this app is available on iOS, Android, macOS, Windows, and Linux.



CLIMB BIG DATA - Configuration and screen sharing

Most of the features described below are accessed via the Guacamole menu. Press `Ctrl + Alt + Shift` (MacOS: `Ctrl + Option + Shift`) to show or hide the Guacamole menu within the Web Shell browser tab.

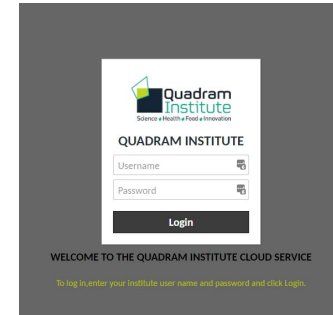


```
ubuntu@bio-skills-11:~$ mkdir nabil
ubuntu@bio-skills-11:~$ ls -alg
total 32
drwxr-xr-x 5 ubuntu 4096 Jan 28 16:05 .
drwxr-xr-x 3 root   4096 Jan 26 12:15 ..
-rw-r--r-- 1 ubuntu 220  Feb 25  2020 .bash_logout
-rw-r--r-- 1 ubuntu 3771 Feb 25  2020 .bashrc
drwx----- 2 ubuntu 4096 Jan 28 15:59 .cache
-rw-r--r-- 1 ubuntu 807  Feb 25  2020 .profile
drwx----- 2 ubuntu 4096 Jan 26 12:15 .ssh
drwxrwxr-x 2 ubuntu 4096 Jan 28 16:05 nabil
ubuntu@bio-skills-11:~$
```

Remember you are sharing: Make your own directory for your work in the home directory



Questions?



All set?



How does a modern bioinformatician organise their work?

Nabil-Fareed Alikhan



@happy_khan



CLIMB
BIG DATA

Agenda

- Background, motivation & theory - while everyone settles!
- Data management for genomics
- Our data cascades through multiple stages
- Our analysis should be reproducible
- Our analysis is iterative
- Containers and workflows fit into your projects
- What is a “container” and why should I use it
- What is a “workflow language” and why should I use it



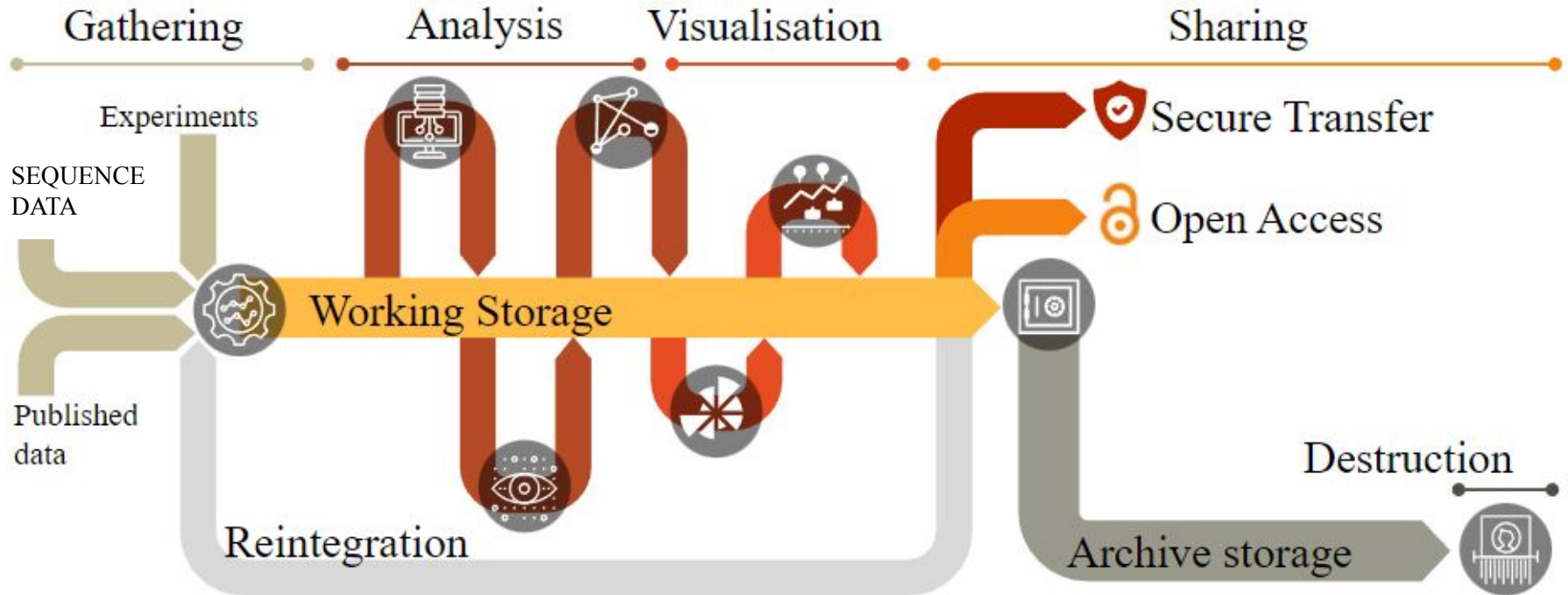
This course is actually about data management

- Who decides how the data will be used?
- How do we secure our data?
- How do we use our data ethically (e.g. privacy)?
- How can I know how data was generated (software versions, database versions)?

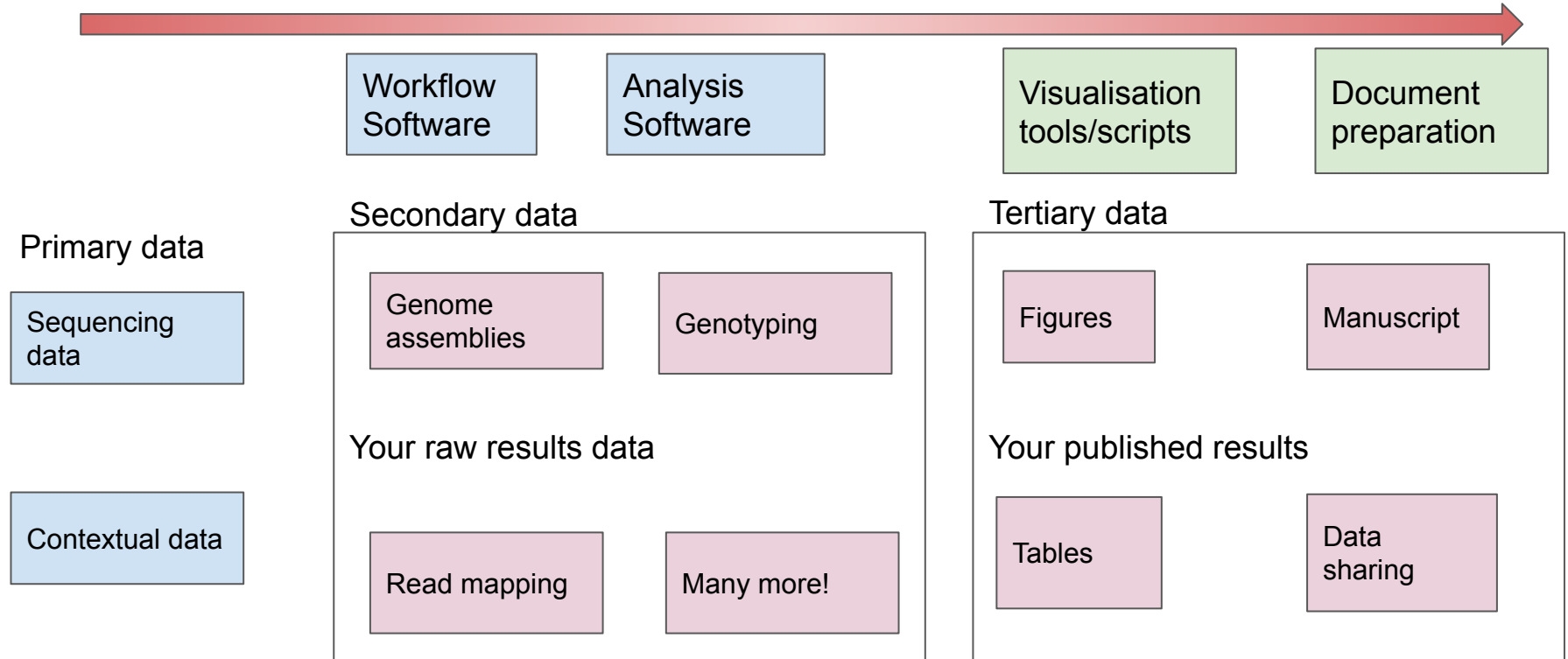
Tell us (in discord #modern-bioinformatics) about a situation where you have had to deal with data management



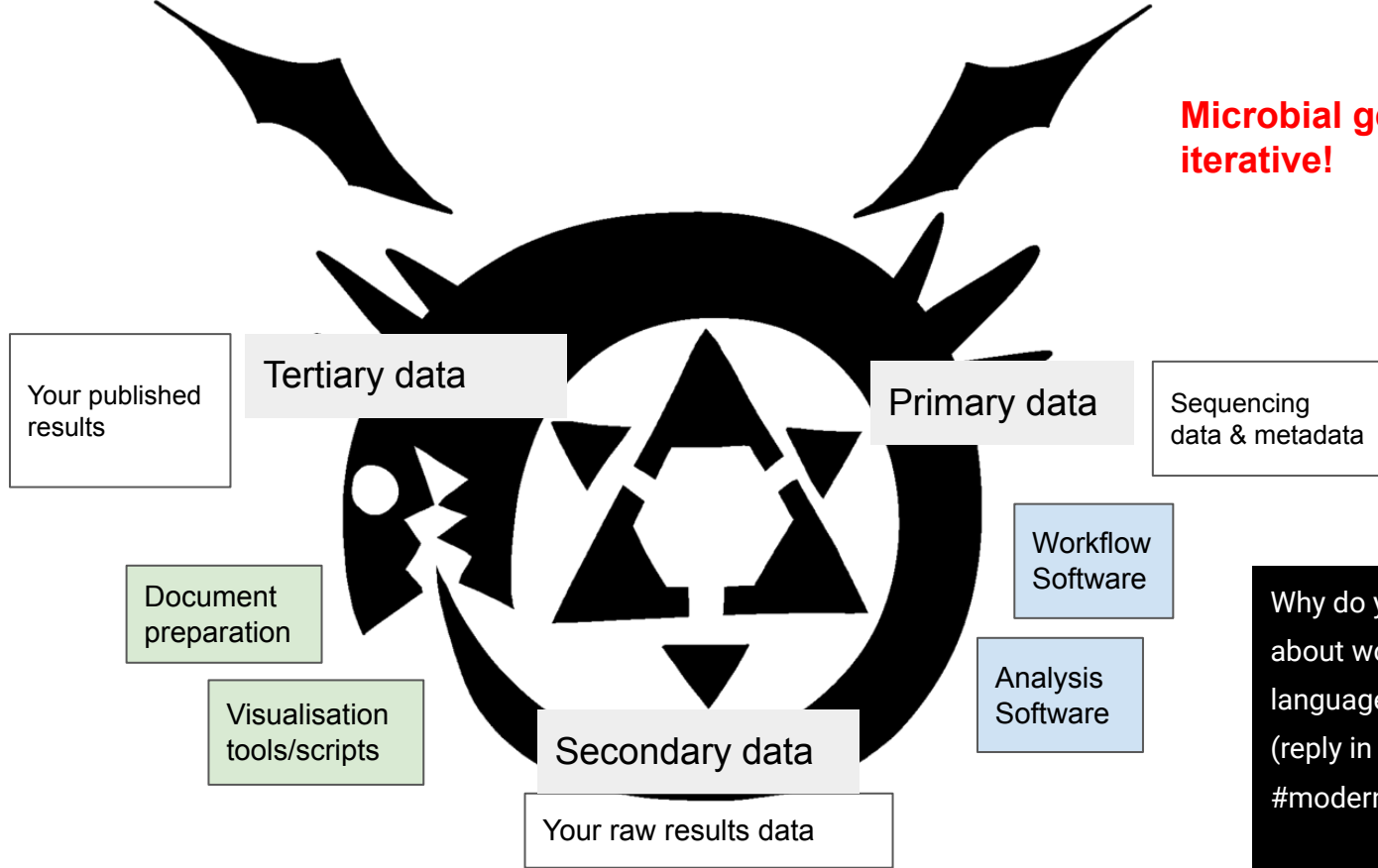
Our data cascades through multiple stages



Our results should be reproducible from primary data



Microbial genomics is iterative!



Why do you want to learn about workflow languages and containers (reply in discord #modern-bioinformatics)

The bioinformatics ouroboros



Containers and workflows fit into your projects

PICK ONE



nextflow

Workflow
Software

Analysis
Software



docker

CONDA®

PICK ONE
(FOR DEPLOYMENT)

Secondary data

Genome
assemblies

Genotyping

Your raw results data

Read mapping

Many more!

Visualisation
tools/scripts

Document
preparation

Tertiary data

Figures

Manuscript

Your published results

Tables

Data sharing

Primary data

Sequencing
data

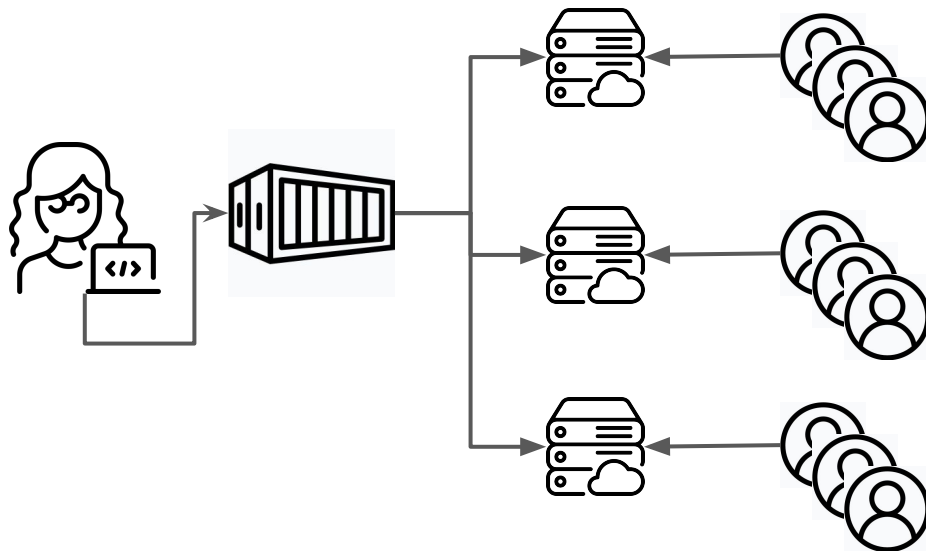
Contextual
data



What is a “container” and what is it for

Software development/deployment

- See: “Docker”
- Testing & continuous integration
- Web application
- Fragile applications
- Pipelines (dependencies)



Doesn't work well with:

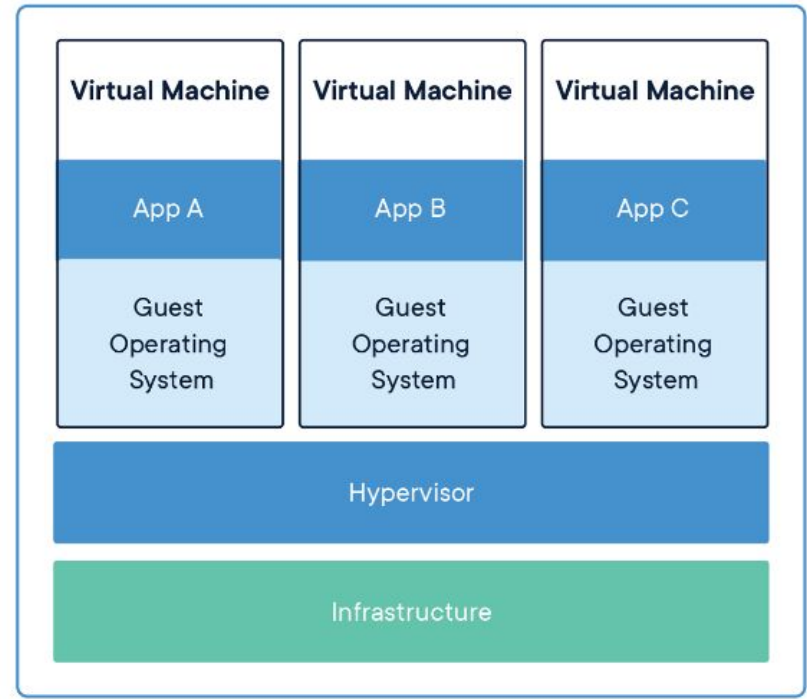
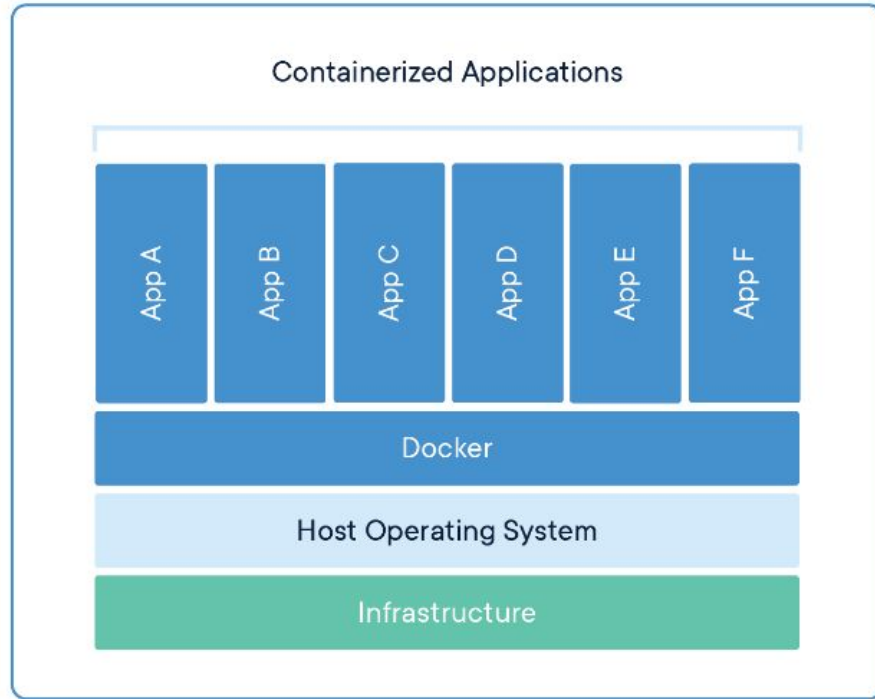
- graphical/GUI applications
- CLI tools that heavily use the filesystem

**Bioinformatics software is
fragile and regularly updated**

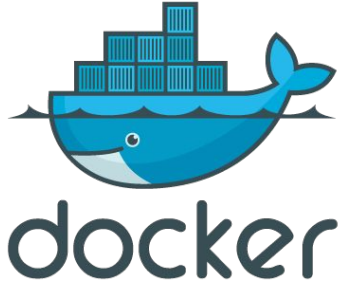


“Containerisation” vs “virtualisation”

The infrastructure can be anything



Docker vs Singularity



- Software development/deployment
- Daemon
- User interface is easier
- Best on your own VM/laptop/desktop



- Specifically HPC
- Newer -> unpolished
- Works on VM/laptop/desktop/HPC
- Can convert Docker to Singularity

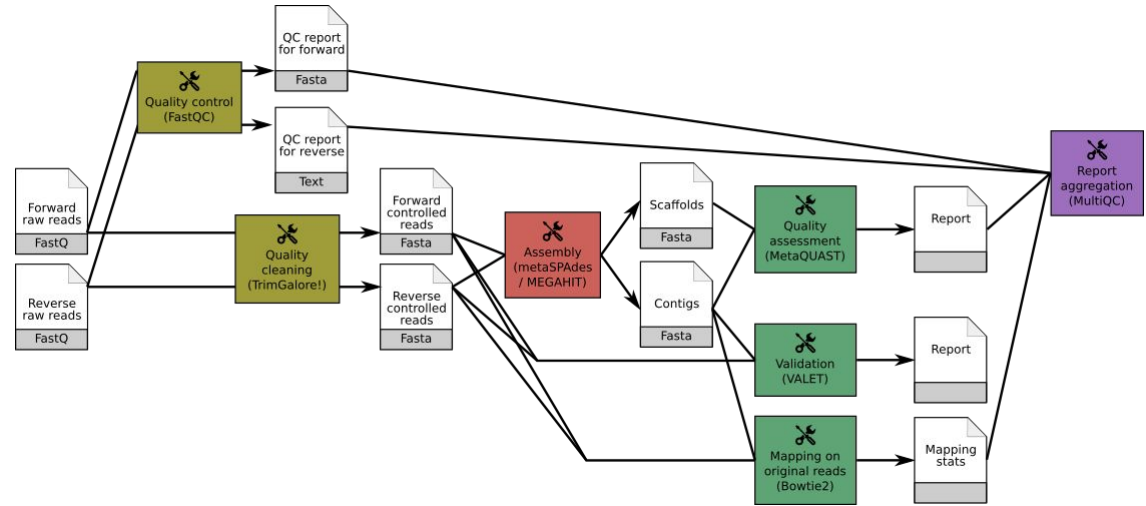


What is a “workflow language”



nextflow

- UNIX philosophy and “pipes”
- Orchestration
- Hides file management
- Handles queues and parallelization
- The “glue” between software



```
samtools view -h in.bam | grep -v "<RG:Z:ERR00001>" | samtools view -bS - > out.bam
```



Why should I use a workflow language?

CAN

- Version tracking
- Portability
- Reproducibility
- Scaling up and out
- Checkpoints - “Resume”

CANT

- Replicability
- Interpretation
- Save time - zero sum game



nextflow



Schedule

Time (GMT)	Item	Speaker/Chair
9:00	Orientation and testing virtual machines	
10:00	Formal welcome	Organising committee
10:10	Lecture: How does a modern bioinformatician organise their work?	Nabil-Fareed Alikhan
10:50	Lecture: Getting things done with Conda and Snakemake	Anna Price
11:30	Lecture: The value and use of containers	Anna Price
12:00	Lunch time break	
13:00	Practical session 1 - Assemble and examine a microbial genome using containers	Anna Price
14:30	Lecture: Provenance and portability through Nextflow	Andrea Telatin
15:00	Practical session 2 - Basic bioinformatics using Nextflow	Andrea Telatin + Nabil-Fareed Alikhan
16:20	Afternoon break	
16:50	Lecture: Working with Nextflow, DSL2 modules and Bactopia	Robert Petit
17:20	Discussion panel	All
18:00	Final remarks	Organising committee
18:10	End of workshop	

