Day 3: Data tools and pipelines

George Githinji, Aquillah Kanzi, Stanford Kwenda,

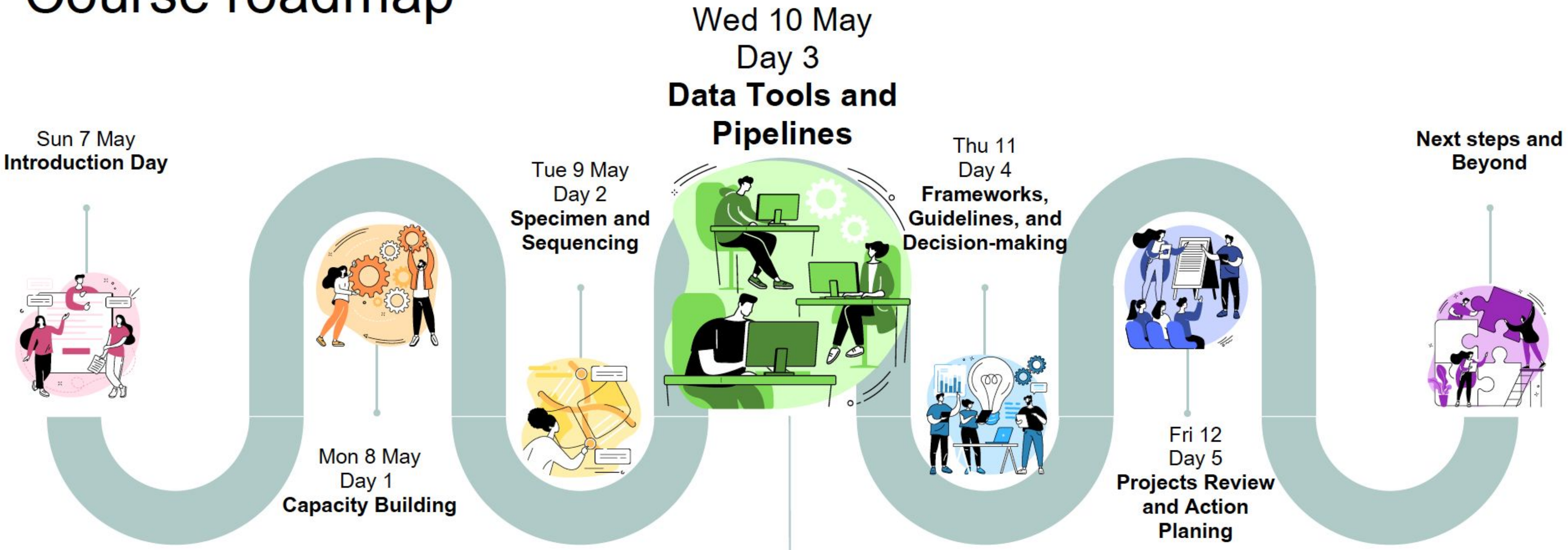
Fatma Guerfali, Kirsty Lee Garson, Alice Matimba, Leigh Jackson, Amadou Diallo

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Addis Ababa, Ethiopia, May 2023

Course roadmap



Sun 7 May
Introduction Day

Mon 8 May
Day 1
Capacity Building

Tue 9 May
Day 2
Specimen and Sequencing

Wed 10 May
Day 3
Data Tools and Pipelines

Thu 11
Day 4
Frameworks, Guidelines, and Decision-making

Fri 12
Day 5
Projects Review and Action Planing

Next steps and Beyond

From raw sequencing output to biological information

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Day 3 Session 1: Setting up Data Infrastructure and Processes

George Githinji, Aquillah Kanzi, Kirsty Lee Garson, Stanford Kwenda, Amadou Diallo

KEMRI | Wellcome Trust

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

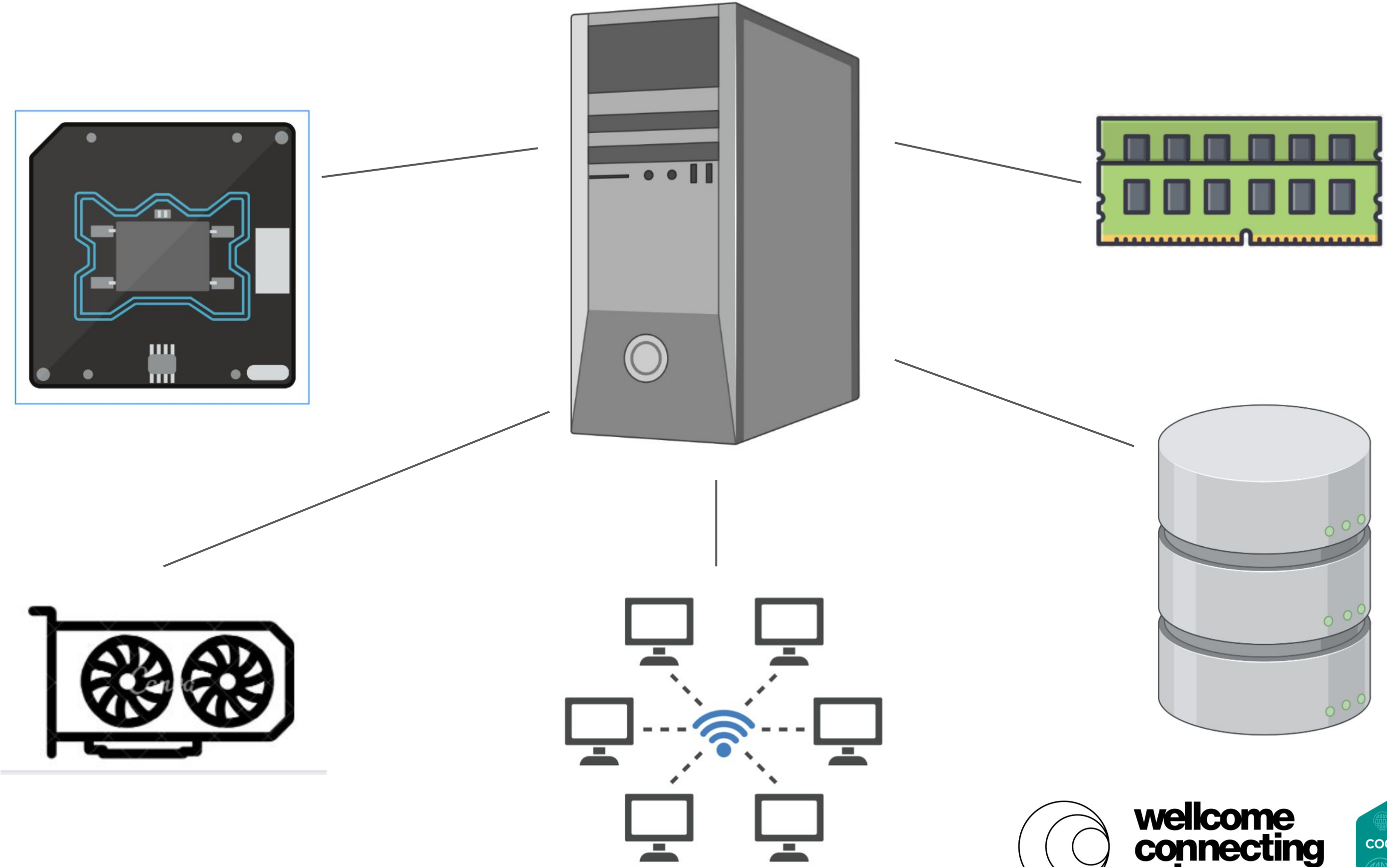
Session outline

- Choice of computing hardware
- Single machines vs HPC vs Cloud
- Operating systems for bioinformatics
- Use cases for genomic analysis
- Outline the use of containers and conda environments

Session outcomes

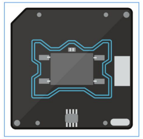
- List the basic components of a data analysis system
- Define the meaning of cpu, hard disk, bandwidth, etc.
- Compare the costs of computing and analysis platforms
- List the differences between local and cloud infrastructure
- Identify the resources needed to setup and maintain computing infrastructure
- Identify operating systems used for bioinformatics analysis

Components of a Computer:



Components of a Computer:

Central Processing Unit (**CPU**)

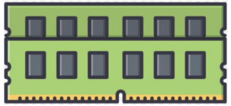


- Reduced Instruction Set Computing (RISC) CPUs

-Complex Instruction Set Computing (CISC) CPUs

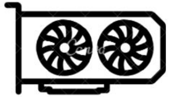
Vendors
Intel, AMD, ARM,IBM, Qualcom, Apple

Random Access Memory (**RAM**)



- Dynamic RAM (DRAM)
- Static RAM (SRAM)

Graphical Processing Unit (**GPU**)



- Integrated GPU’s (laptops)
- Dedicated GPU’s (HPCs)

Vendors
NVIDIA, Apple, AMD, Intel, Qualcom, Imaging technologies

Disk Storage

- Direct Attached Storage (DAS)
-Network Attached Storage (NAS)
-Storage Area Network (SAN)
-Object Storage
-Parallel File System

File system considerations

- Lustre
- XFS
- ZFS

Vendors
HPE, DELL,NetApp, IBM,HITACHI Vantara, Western Digita, PureStorage

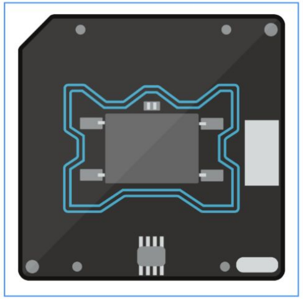
wellcome connecting science

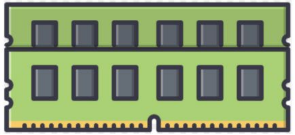
COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Created in BioRender.com bio

Canva

Components of a Computer:





CPU:

- Number will determine speed of analysis
- Strength will determine complexity and speed
- “Threads” determine how many data streams can be processed at the same time

RAM:

- Fast memory used to feed into CPU from the Disk
- Size in Gigabytes ranges between 1 - 512GB
- Does not store data long term

Disk Storage:

- Slow memory used to keep data for long terms
- Size in ranges between Megabytes and Terabytes
- Stores all input and output files for processing

Operating systems

The operating system manages computer hardware, software resources, and provides common services for computer programs

UNIX®
An Open Group Standard

Linux

macOS

Windows 11

Commercial vs Open source OS

- How do I choose an OS ?
- How do I choose which Linux?

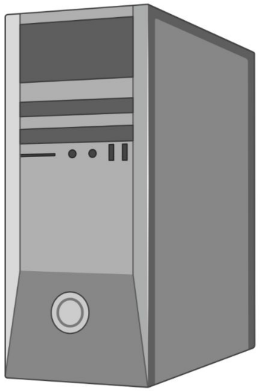
RedHat, SUSE,UNIX,

- Redhat Family - Centos, Rocky Linux, Fedora
- Debian family - Debian, Ubuntu
- OpenSUSE

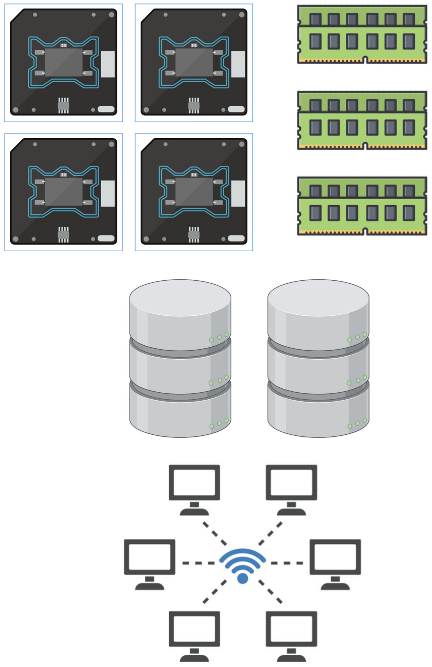
Created in BioRender.com bio
Canva

Computing Environments:

Single Desktop Computer



High Performance Computer



Cloud Computing

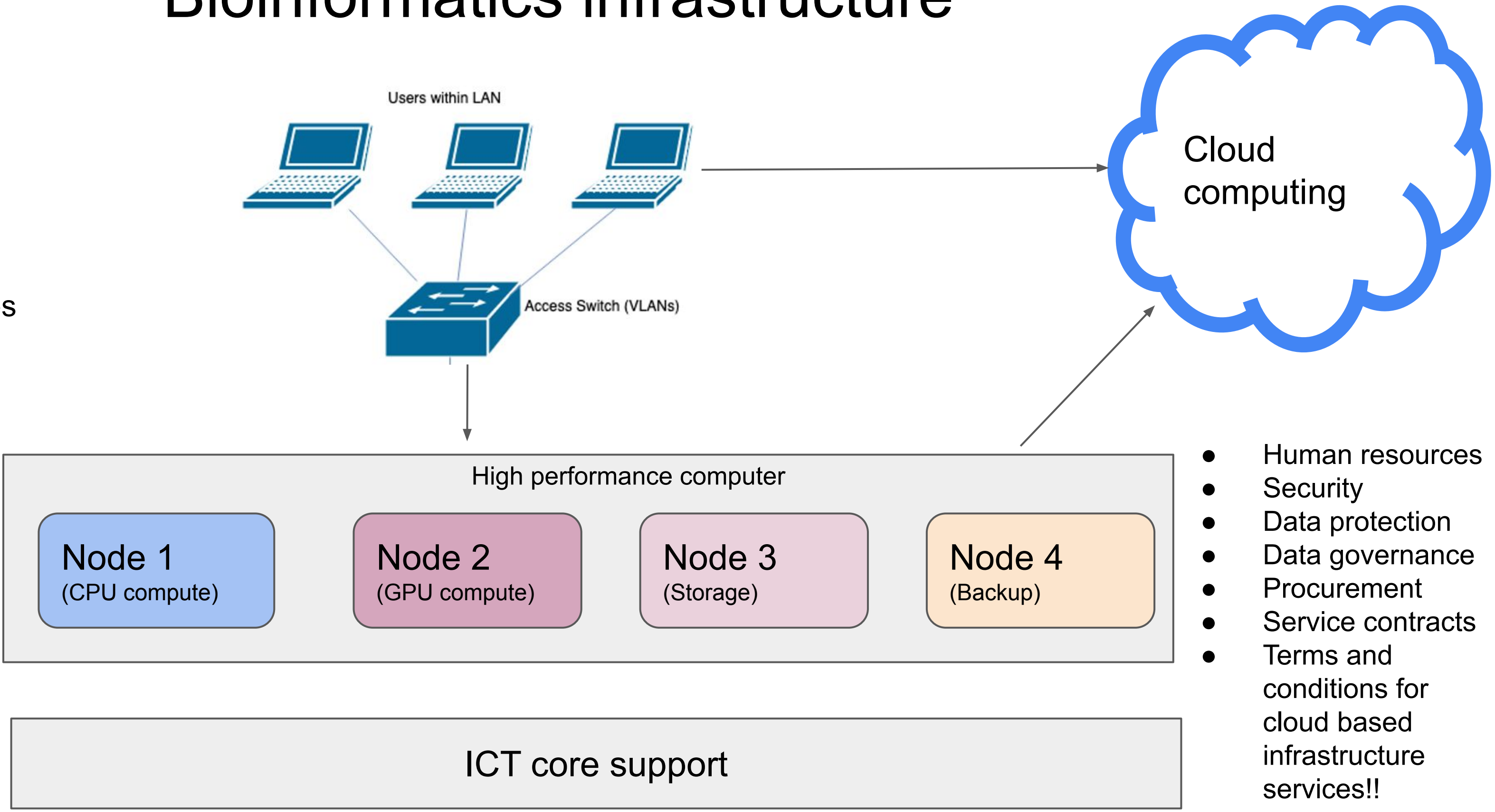
wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Created in BioRender.com bio Canva

Bioinformatics infrastructure

High throughput devices



Users within LAN

Access Switch (VLANs)

Cloud computing

High performance computer

Node 1 (CPU compute)

Node 2 (GPU compute)

Node 3 (Storage)

Node 4 (Backup)

ICT core support

- Human resources
- Security
- Data protection
- Data governance
- Procurement
- Service contracts
- Terms and conditions for cloud based infrastructure services!!

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Use case 1

Single computer

Applicable workflow options

- Terra and Galaxy workflows
- Web-accessible, software as a service solutions

Minimum specs

- i7, >16Gb RAM, >1TB disk
- M2, >16Gb RAM, >1TB disk

Use case 2

High-Performance Computing (HPC)

Applicable workflow options

- Command-line workflows
- Batch processing Workflows
- Memory intensive workflows

Minimum specs

- CPU - Multicore (2.6GHz-3GHz)
- RAM >2Gb
- Storage > 1Tb per node
- Network >10Gbits/S

Use case 3

Cloud computing

Applicable workflow options

- Performance-intensive workflows e.g. machine learning algorithms, eukaryotic genome analysis
- Algorithms requiring parallel computing

Minimum specs

- Similar to HPC but could be scaled depending on workflow requirements

Terra and Galaxy workflows

Examples include:

COVID-19 Galaxy Workflows
https://covid19.galaxyproject.org/artic/

Theiagen's Public Health Viral Genomics WDL Workflows [**Terra**]
https://dockstore.org/organizations/Theiagen/collections/PublicHealthViralGenomics

Infrastructure and personnel requirements/advantages/disadvantages

Web-accessible, software as a service solutions

Examples include:

EnteroBase
https://enterobase.warwick.ac.uk/

Pathogenwatch
https://pathogen.watch/

Chan Zuckerberg ID
(formerly known as IDseq)
https://czid.org/

Infrastructure and personnel requirements/advantages/disadvantages

Command-line interface tools

Examples include:

- Nextflow

  Nextflow workflows repositories (https://nf-co.re/)

- Snakemake

Infrastructure and personnel requirements/advantages/disadvantages

Virtualization

Containerisation

- Docker
- Singularity (HPC)

- Conda / Miniconda
- Virtual env

Bioinformatics workflow managers

- Nextflow
- Snakemake
- Common workflow language

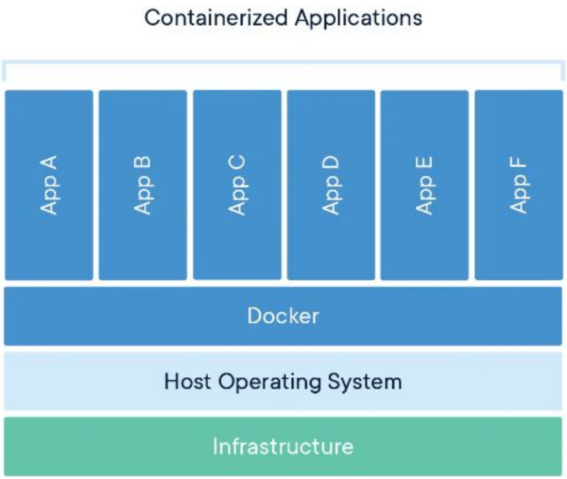
Package Managers

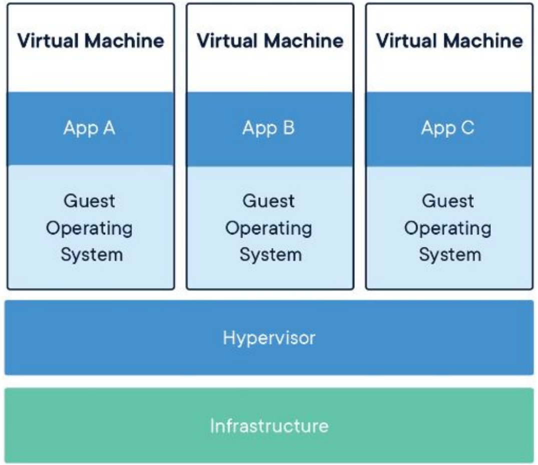
- Mamba
- Apt
- Yum
- System modules

Version Control

- Git
- Mercurial

Containers vs Virtual Machines

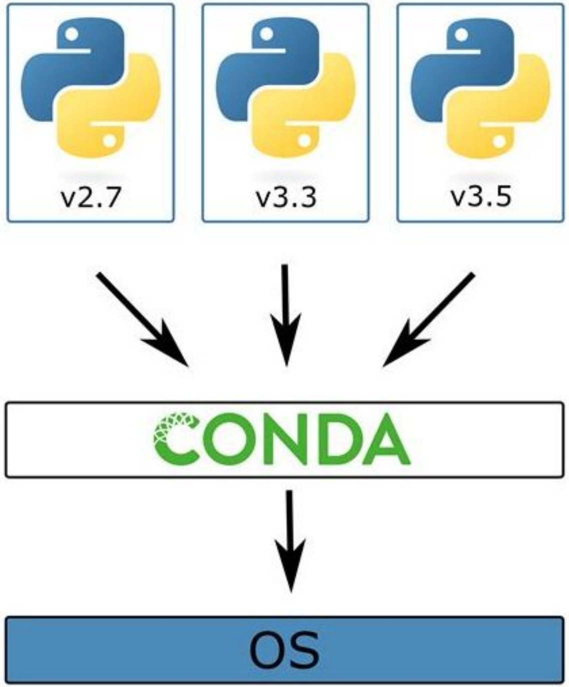




https://www.docker.com/resources/what-container/
https://docs.docker.com/get-started/docker_cheatsheet.pdf

Environment Managers

- Allows running multiple versions of the same software
- Resolve dependency issues between softwares



https://docs.conda.io/projects/conda/en/latest/user-guide/
https://docs.conda.io/projects/conda/en/latest/user-guide/cheatsheet.html

Relevant resources

PHA4GE Bioinformatics Solutions for SARS-CoV-2 Genomic Analysis:
https://github.com/pha4ge/pipeline-resources/blob/main/docs/bioinfo-solutions.md

Module 1
SARS-CoV-2 Bioinformatics Training, October-November 2021
Peter van Heusden & George Githinji
https://uct-cbio.github.io/ngs-academy/uploads/sars-cov_analysis_workflows.pdf

Human Resources for Data Analysis:

- Bioinformaticians
- Data Scientists
- Systems Administrators
- Data Administrators
- Server managers
- Support Teams
- Programming teams
- Network Administrators

Activity: Manage computer infrastructure for analysis

You are the **Lead Data Manager** for your analysis team:

You have been given a defined dataset of sequencing data to analyse

1. Purchasing equipment and setup is instant in this scenario!

1. Process as many samples in a 24hr day as possible

1. Create a system to process the data so your team can begin analysis!

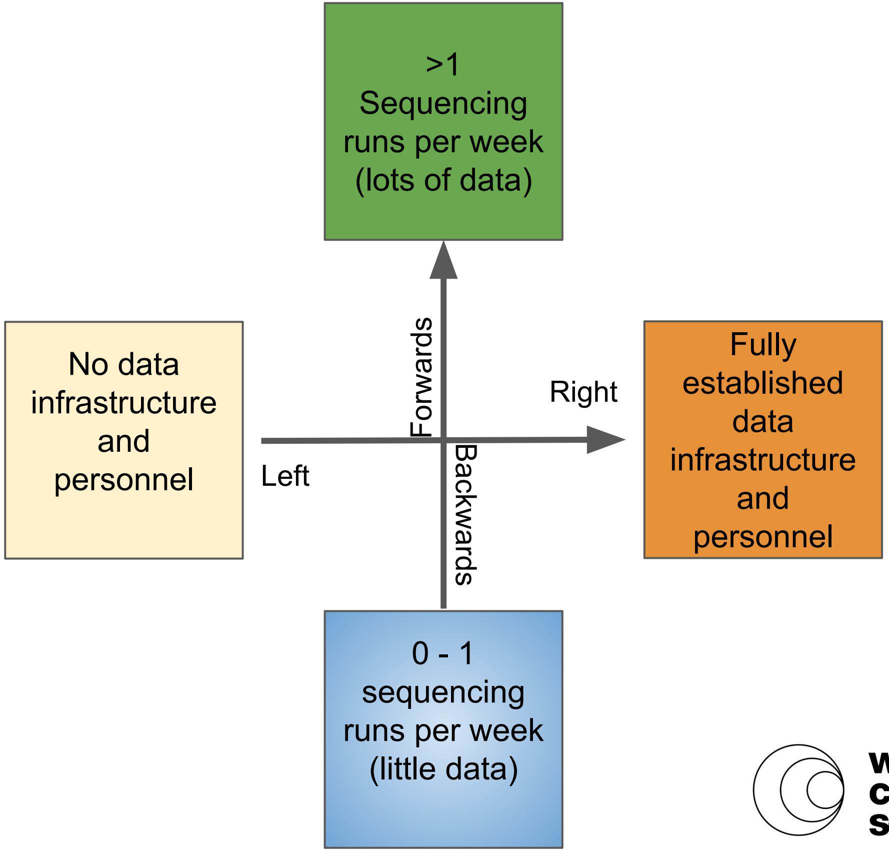
Activity: Data Analysis Needs Assessment

Move around the room based on your needs and environment:

Form

5 groups

Based on who are closest together!



>1 Sequencing runs per week (lots of data)

No data infrastructure and personnel

Fully established data infrastructure and personnel

0 - 1 sequencing runs per week (little data)

wellcome connecting science

COG-TRAIN COVID-19 GENOMICS GLOBAL TRAINING

Group Tasks - 15 minutes

Group 1:
Budget =
$500

Number of Samples:
24

Condition:
All data must be backed up once

Group 2:
Budget =
$750

Number of Samples:
48

Condition:
Input files must be deleted

Group 3:
Budget =
$1200

Number of Samples:
72

Condition:
Power cuts 12hrs per day

Group 4:
Budget =
$2000

Number of Samples:
100

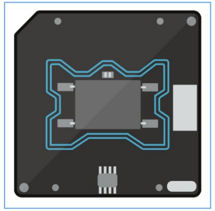
Condition:
The data is sovereign

Group 5:
Budget =
$400

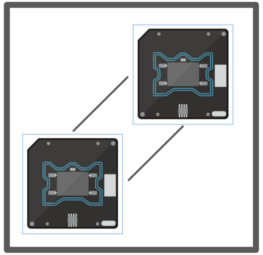
Number of Samples:
10

Condition:
Outputs must be backed up once

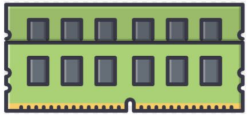
Costs and Performance Sheet:



Single CPU:
1 Sample per hour
Costs $200

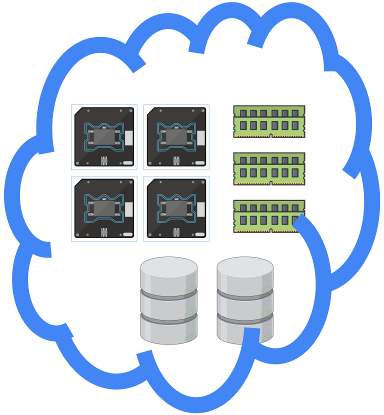


Dual core CPU:
2 samples per hour
Costs $300



RAM:
Each sample running through the workflow requires 4GB of RAM
Cost: $50 for 8GB RAM

Storage:
Each sample is 10GB in sizes
The output of each analysis is 5GB in size
Cost: $50 for 500GB storage



Rent Cloud Computing:
The cloud can process 3 samples per hour
Costs: $50 per hour for running
Costs: $20 per 1000GB of data stored
The cloud is hosted in USA

wellcome connecting science

COG-TRAIN
COVID-19 GENOMICS GLOBAL TRAINING