Genomics England

Running workflows on the HPC and Cloud

Emily Perry and Hamzah Syed

12th November 2024

Data security

- This training session will include data from the GEL Research Environment
- As part of your IG training you have agreed to not distribute these data in any way
- You are not allowed to:
    - Invite colleagues to watch this training with you
    - Take any screenshots or videos of the training
    - Share your webinar link (we will remove anyone who is here twice)
- We will record this training and distribute the censored video afterwards

Presenters

Emily Perry
Research Engagement Manager

Hamzah Syed
Solutions Manager - Lifebit

Questions

All your microphones are muted

Use the Zoom Q&A to ask questions

Upvote your favourite questions: if we are short on time we will prioritise those with the most votes

Questions

Elena Bernabeu
Bioinformatician - Research Services

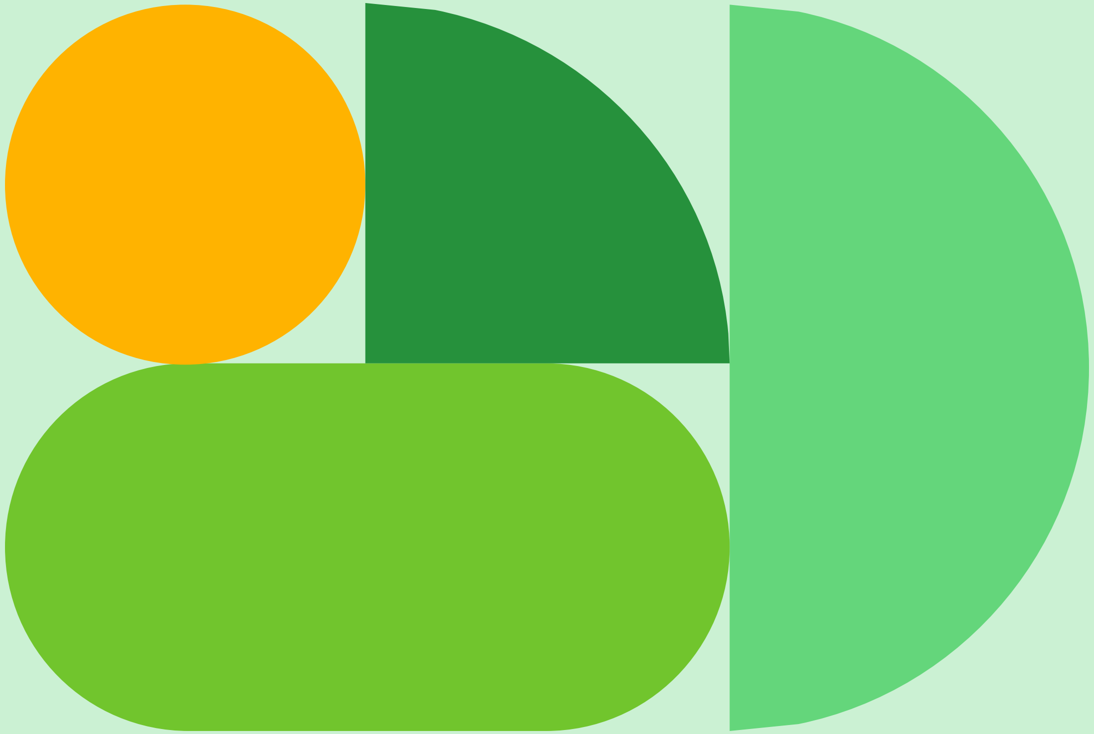
Ian Grant
Platform Engineer (HPCDC & Long Reads)

Eleni Kyriakou
Senior Client Success Manager - LifeBit

Charlotte Maughan
Client Manager - LifeBit

Agenda

1. Introduction and admin
2. What is a High Performance Cluster?
3. Queues available on the HPC
4. How to create and monitor jobs on the HPC
5. Tools and software available and how to load them
6. Interactive coding tools
7. Bringing in your own tools and software
8. CloudOS – batch and interactive jobs on the Cloud
9. Help and questions



2. What is a High Performance Cluster?

What is an HPC?

https://re-docs.genomicsengland.co.uk/hpc_about/

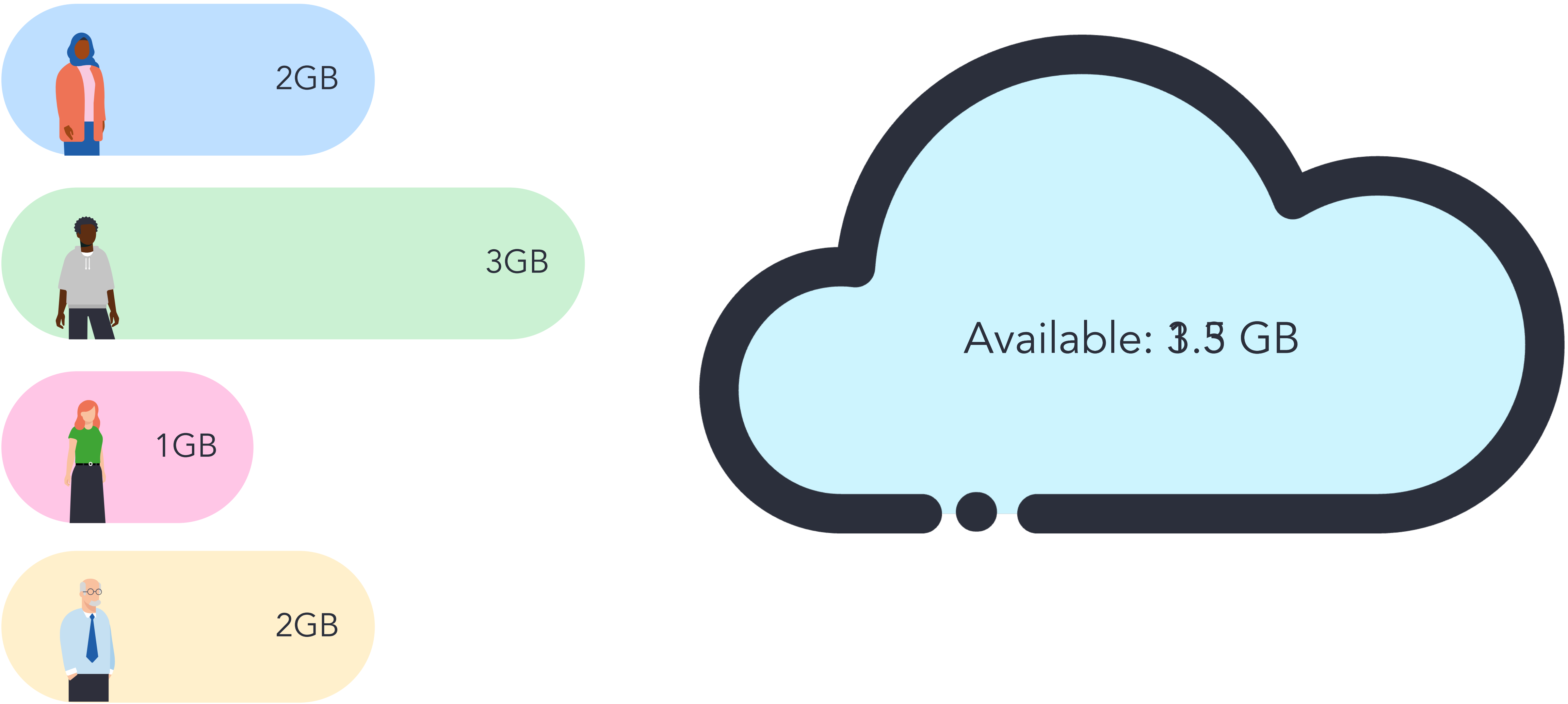
What is an HPC?

Lots of compute power

Shared with other researchers

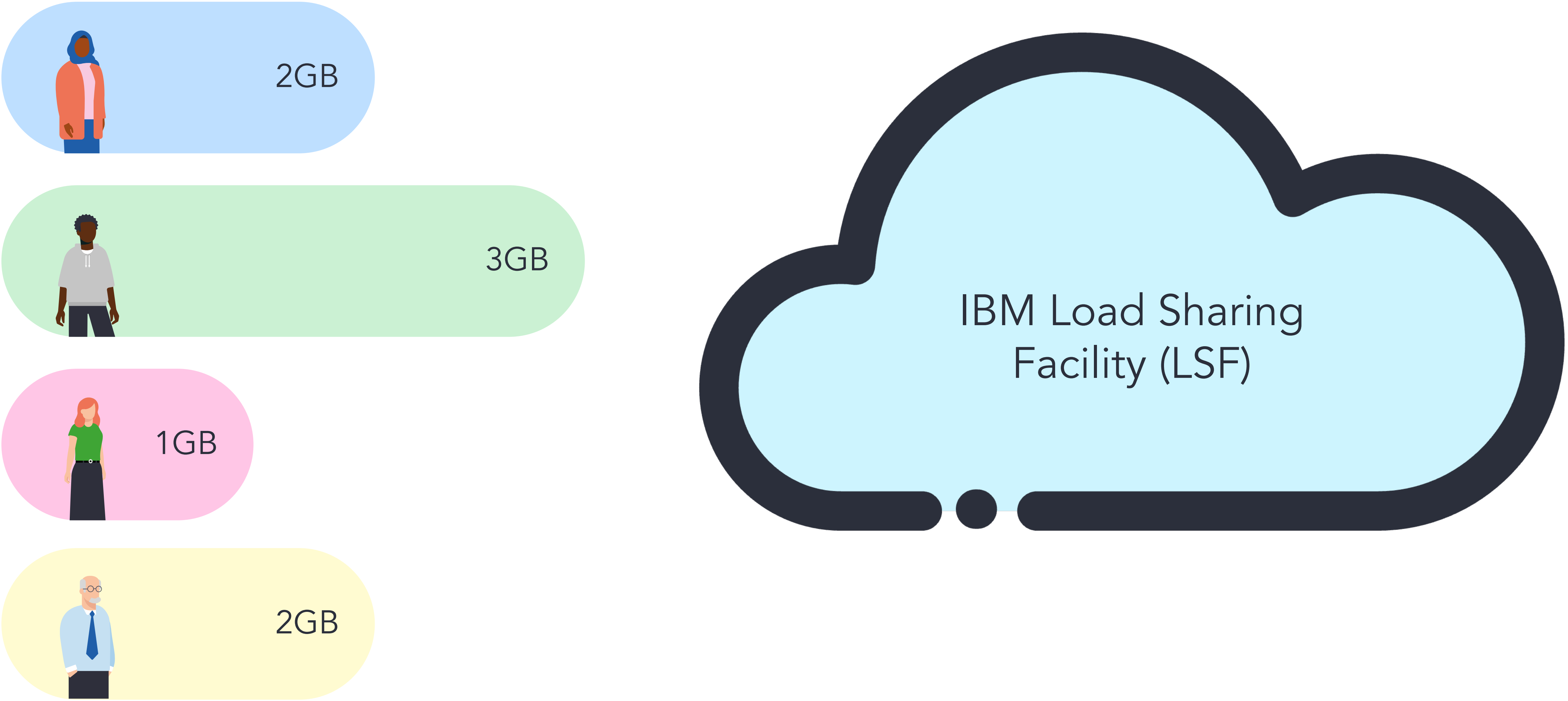
https://re-docs.genomicsengland.co.uk/hpc_about/

How do we share resources?



https://re-docs.genomicsengland.co.uk/hpc_about/

How do we share resources?



2GB

3GB

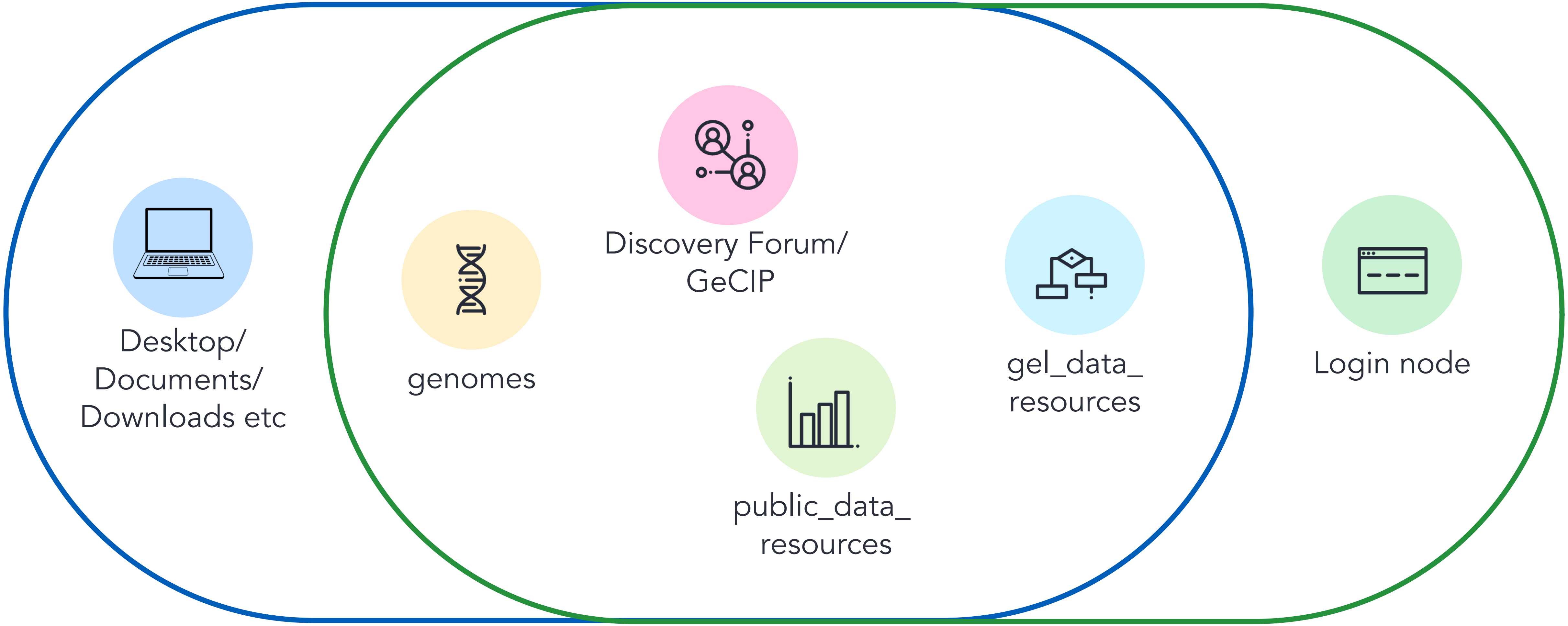
1GB

2GB

IBM Load Sharing Facility (LSF)

https://re-docs.genomicsengland.co.uk/hpc_about/

Files on the HPC



Desktop/ Documents/ Downloads etc

genomes

Discovery Forum/ GeCIP

public_data_ resources

gel_data_ resources

Login node

Access locally

Access on the HPC

Why use the HPC?

The HPC enables you to solve larger, more complex problems in far less time...

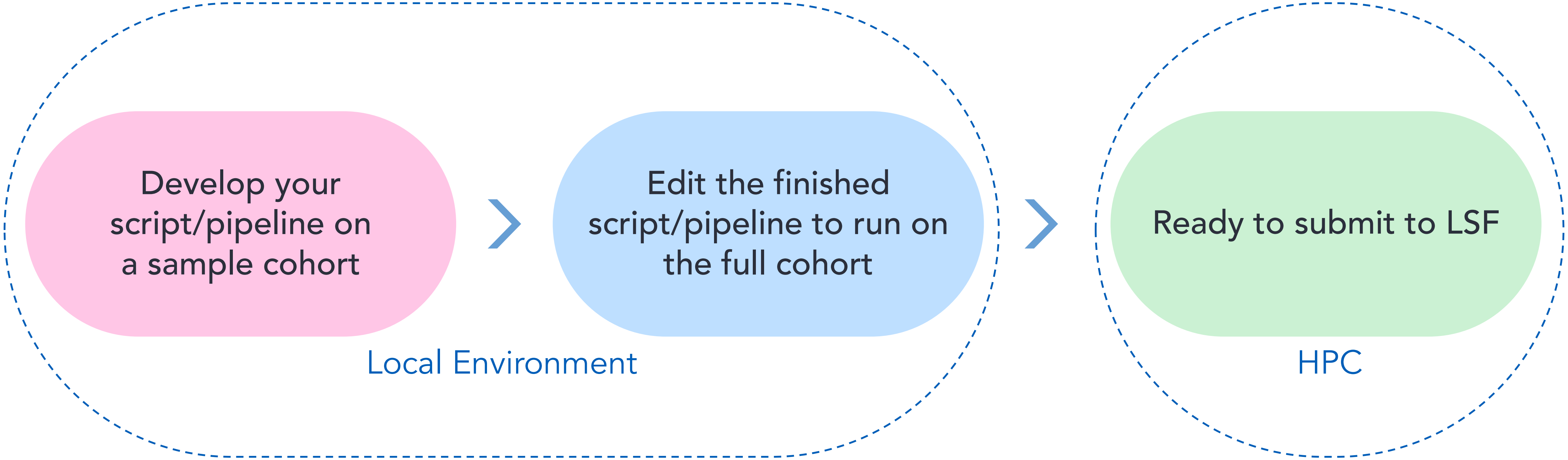
1. Increased performance when running complex scripts and pipelines
2. Avoid RAM cap in the Research Environment desktop

The local environment of the RE is designed for exploratory work.

The HPC is provided for "heavyweight" batch research.

https://re-docs.genomicsengland.co.uk/hpc_about/

When use the HPC?



Develop your script/pipeline on a sample cohort

Edit the finished script/pipeline to run on the full cohort

Ready to submit to LSF

Local Environment

HPC

Why use the HPC?

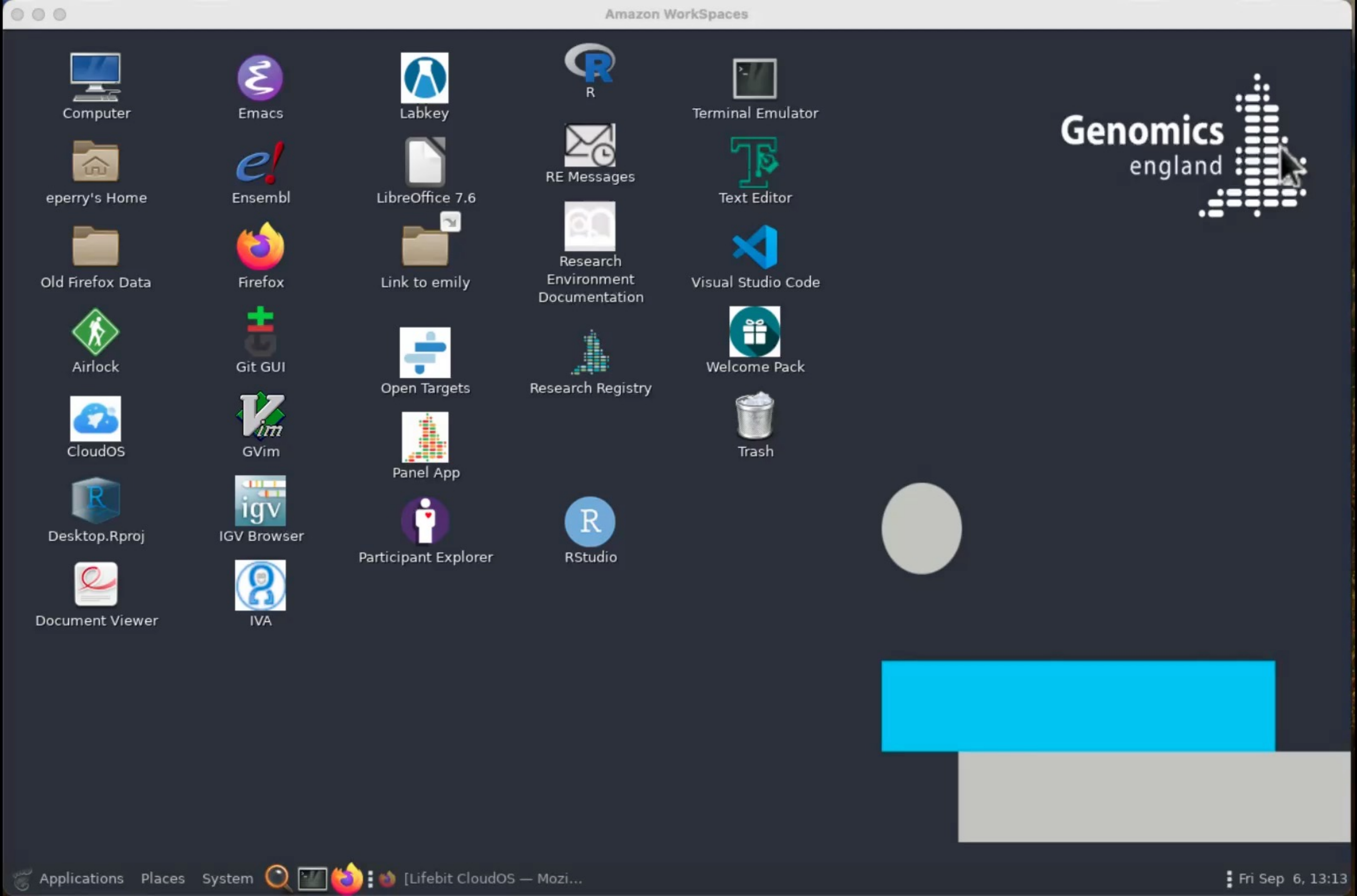
Software only available on the HPC

```
delly/1.2.6                    pindel/0.2.5b8
denovoGear/1.1.1               platypus/0.8.1
discover/0.9.5                 plink/1.9
dotnet/2.0.0                   plink/2.00a3.3LM
dotnet/8.0.1              (D)  plink/2.0                 (D)
drop/1.2.4                     plink_seq/0.10
eigen/3.3.9                    popdel/1.5.0
exomiser/13.3.0                proj/8.2.1
exonerate/2.2.0                prsice-2/2.3.5
fastqc/0.12.1                  pycircos/1.0.2
fetk/1.9.3                     pysam/0.22.0
ffmpeg/6.0                     python/3.8
fribidi/1.0.12                 python/3.8.1
gatk/4.5.0.0                   python/3.11               (D)
gauchian/1.0.2                 readline/8.0
gcc/10.4.0                     regenie/3.4.1
gcta/1.94                      repeatDetector/1.0
gdal/3.7.0                     rtg-tools/3.12.1
geos/3.12.1                    rvtests/2.1.0
gistic/2.0.23                  saige/1.0.9
gmp/6.2.1                      salmon/1.10.0
gnu-parallel/20190222          samtools/1.16.1
gnu/4.4                        shapeit4/4.2.2
gradle/8.5                     sniffles/1.0.11
guppy/3.4.5                    somalier/0.2.19
gvcfgenotyper/2019.02.26       sqlite3/3.40.0
haplocheck/1.3.3               squirls/2.0.1
hipstr/0.7                     stack/2.15.7
hisat2/2.2.1                   star/2.7.2a
hla-la/1.0.3                   star/2.7.11a              (D)
hmftools/2024-02-06            strelka/2.9.10
homer/4.11                     superSTR/1.0.1
htslib/1.18                    svanna/1.0.4
igv/2.17.1                     tabix/1.18
imagemagick/7.1.0              trimmomatic/0.39
java/1.8                       udunits/2.2.28
java/11.0.2                    vcf2maf/1.6.21
java/17.0.2                    vcfanno/0.3.4
java/19.0.2               (D)  vcflib/1.0.9
jq/1.7.1                       vcftools/0.1.16
kallisto/0.50.1                verifyBamID/2.0.1
king/2.3.2                     vt/0.57721
kraken/1.1.1                   xz/5.4.7
kraken2/2.1.3                  zlib/1.3
lapack/3.12.0                  zulu/21.0.1
ldsc/1.0.1                     aws-cli/2.15
singularity/3.8.3              singularity/4.1.1 (D)
```

```
AdapterRemoval/2.3.3               ldstore/2.0
AutoDock_Vina/1.2.5                libdeflate/1.20
BWA/0.7.17                         libgit2/1.6.2
BerkeleyDB/3.01                    libtiff/3.4
Bio-DB-HTS/3.01                    libtiff/4.3.0
CADD/1.6                           libtiff/4.5.0                  (D)
CNView/1.0                         libunwind/1.8.0
CNVnator/0.4.1                     liftover/1.0
CaVEMan/1.15.3                     linasm/1.13
ExpansionHunter/3.2.2              llvm/16.0.6
ExpansionHunter/4.0.2         (D)  locuszoom/1.4
ExpansionHunterDenovo/0.9.0        lollipop/0.3.0
GSL/2.7                            lumpy/0.3.1
MEDICC2/1.0.2                      mafft/7.520
MPFR/4.2.0                         magma/1.10
R/3.6.3                            manta/1.6.0
R/4.2.1                            matlab/8.1
R/4.3.3                       (D)  matlab/24.1                    (D)
REViewer/0.2.7                     maven/3.9.6
aliview/1.28                       meme/5.5.5
ampliconArchitect/1.3.r7           metal/1.0
ampliconClassifier/1.1.1           miniconda3/23.11.0
annotSV/3.3.7                      miniforge3/23.11.0-0
annovar/2019Nov                    minimap2/2.26
annovar/2024-03-14            (D)  mosaicHunter/2024-02-14
ant/1.9.16                         mplayer/1.5
apbs/3.4.1                         msisensor-pro/1.2.0
asmc-asmc/2024-02-26               msisensor/0.6
automake/1.15                      multiqc/1.19
bamtools/2.5.2                     music2/0.2
bcftools/1.16                      mutserve/2.0.0-rc15
beagle/5.4                         mutsig2cv/3.11
bedops/2.4.41                      ncurses/6.4
bedtools/2.30.0                    new_fugue/2010-06-02
bedtools/2.31.0               (D)  nextflow/22.10.5
blast+/2.15                        nextflow/23.04
blat/1.0                           nextflow/23.10-with-plugins
bolt-lmm/2.4.1                     nextflow/23.10
boost/1.83                         nextflow/24.04.2-with-plugins (D)
bowtie2/2.5.2                      nf-core/0.3.1
canvas/1.40.0.1613                 nf-test/0.7.3
circos/0.69-9                      nf-test/0.8.2
clang/16.0.6                       nf-test/0.9.0                  (D)
cmake/3.24.3                       nodejs/16.9.0
cpan/1.7047                        openrefine/3.7.4
cromwell/v65                       openssl/1.1.1o
curl/7.81.0                        pandoc/3.3
cython/3.0.8                       perl/5.38.2
cytoscape/3.10.1                   picard/3.1.1
```

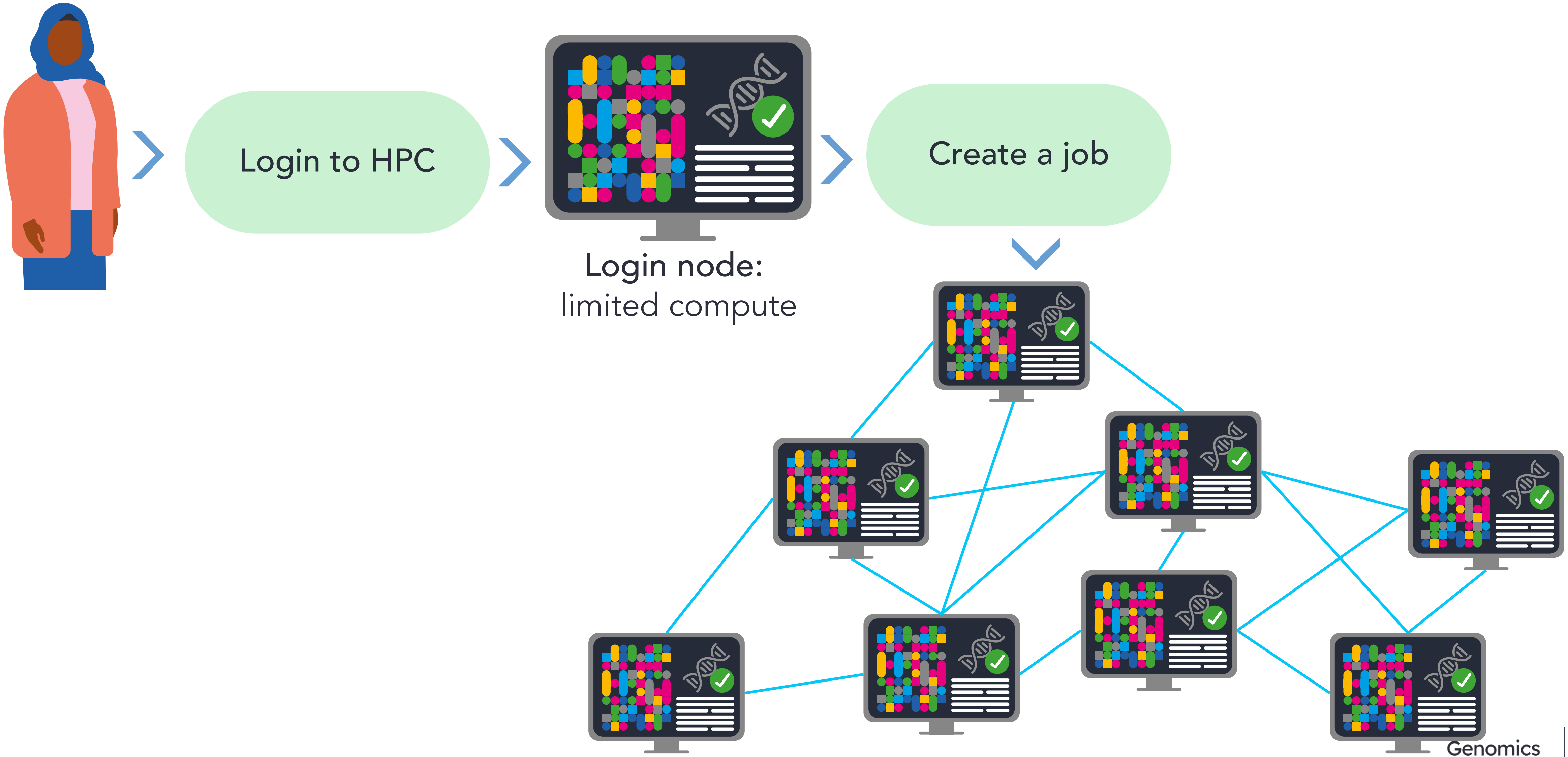
https://re-docs.genomicsengland.co.uk/hpc_using_software/

HPC login demo


Amazon WorkSpaces
Computer
Emacs
Labkey
R
Terminal Emulator
Genomics england
eperry's Home
Ensembl
LibreOffice 7.6
RE Messages
Text Editor
Old Firefox Data
Firefox
Link to emily
Research Environment Documentation
Visual Studio Code
Airlock
Git GUI
Open Targets
Research Registry
Welcome Pack
CloudOS
GVim
Panel App
Trash
Desktop.Rproj
IGV Browser
Participant Explorer
RStudio
Document Viewer
IVA
Applications
Places
System
[Lifebit CloudOS — Mozi...
Fri Sep 6, 13:13

3. Queues available on the HPC

What is a "job"?



Login to HPC

Login node:
limited compute

Create a job

https://re-docs.genomicsengland.co.uk/hpc_jobs/

Two types of job

- Interactive jobs
- Batch jobs

https://re-docs.genomicsengland.co.uk/hpc_queues/

Interactive jobs – the inter queue

require regular input

make decisions based on the results of the previous command

developing tools that you'll scale up later

exploratory analysis

interactive coding tools

GUI tools

https://re-docs.genomicsengland.co.uk/hpc_queues/

Batch jobs

pipelines

next-step triggered in pipeline

Queues

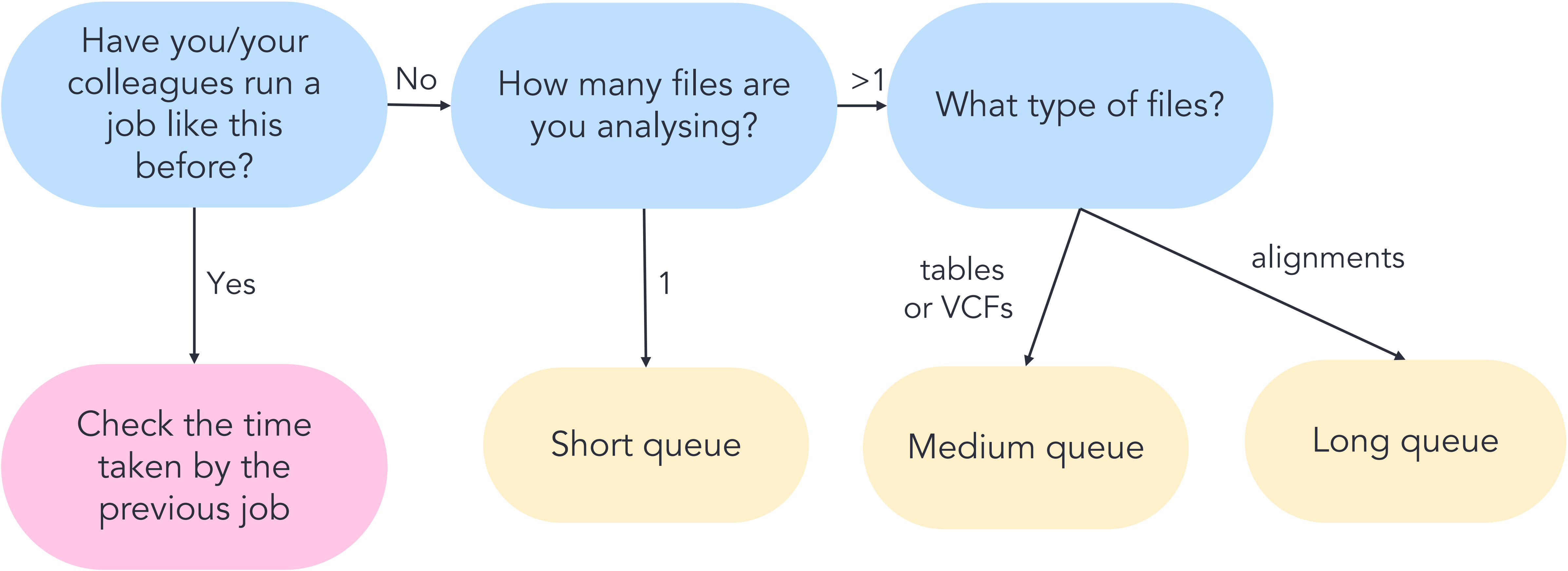
Short <4 hours

Medium 4-24 hours

Long 24 hours +

https://re-docs.genomicsengland.co.uk/hpc_queues/

Choosing a batch queue



Have you/your colleagues run a job like this before?

No

How many files are you analysing?

>1

What type of files?

Yes

Check the time taken by the previous job

1

Short queue

tables or VCFs

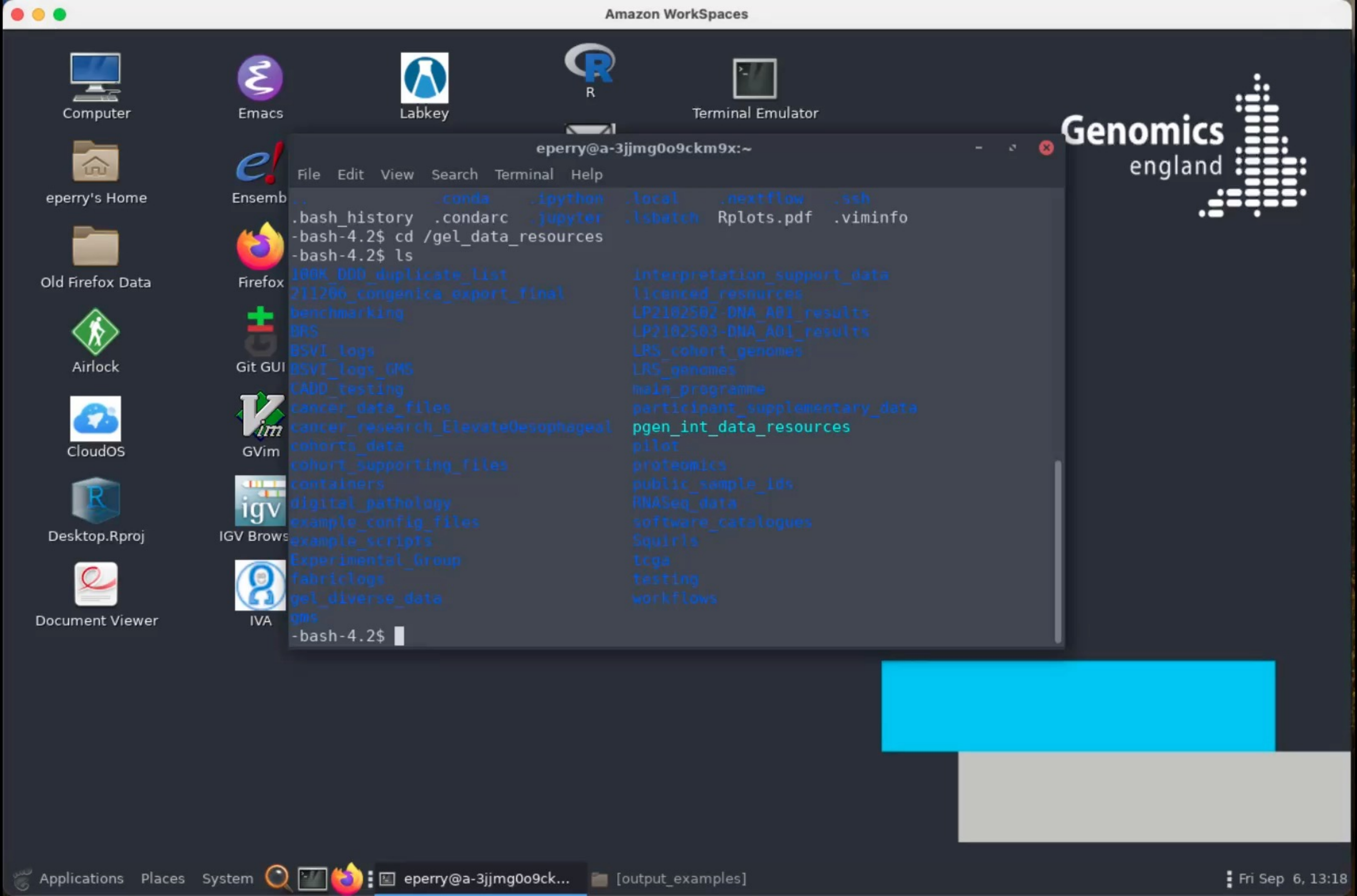
Medium queue

alignments

Long queue

https://re-docs.genomicsengland.co.uk/hpc_queues/

Queues demo



```
..             .conda    .ipython  .local    .nextflow   .ssh
.bash_history  .condarc  .jupyter  .lsbatch  Rplots.pdf  .viminfo
-bash-4.2$ cd /gel_data_resources
-bash-4.2$ ls
100K_DDD_duplicate_list             interpretation_support_data
211206_congenica_export_final       licenced_resources
benchmarking                        LP2102502-DNA_A01_results
BRS                                 LP2102503-DNA_A01_results
BSVI_logs                           LRS_cohort_genomes
BSVI_logs_GMS                       LRS_genomes
CADD_testing                        main_programme
cancer_data_files                   participant_supplementary_data
cancer_research_ElevateOesophageal  pgen_int_data_resources
cohorts_data                        pilot
cohort_supporting_files             proteomics
containers                          public_sample_ids
digital_pathology                   RNASeq_data
example_config_files                software_catalogues
example_scripts                     Squirls
Experimental_Group                  tcga
fabriclogs                          testing
gel_diverse_data                    workflows
gms
-bash-4.2$ 
```

4. How to create and monitor jobs on the HPC

Creating a job - parameters

- -q queue name: inter/short/medium/long
- -P Your project ID
- -o standard output location path_to/job.%J.out
- -e standard error location path_to/job.%J.err
- -J job name: useful for monitoring
- -R Memory usage
- -M Maximum memory
- -n number of nodes
- -cwd current working directory

bsub

https://re-docs.genomicsengland.co.uk/hpc_jobs/

Memory usage

-R Memory usage

- How much memory you’re requesting in MB
- `rusage[mem=1000]`
- There needs to be this much memory available on the queue for your job to start

-M Maximum memory

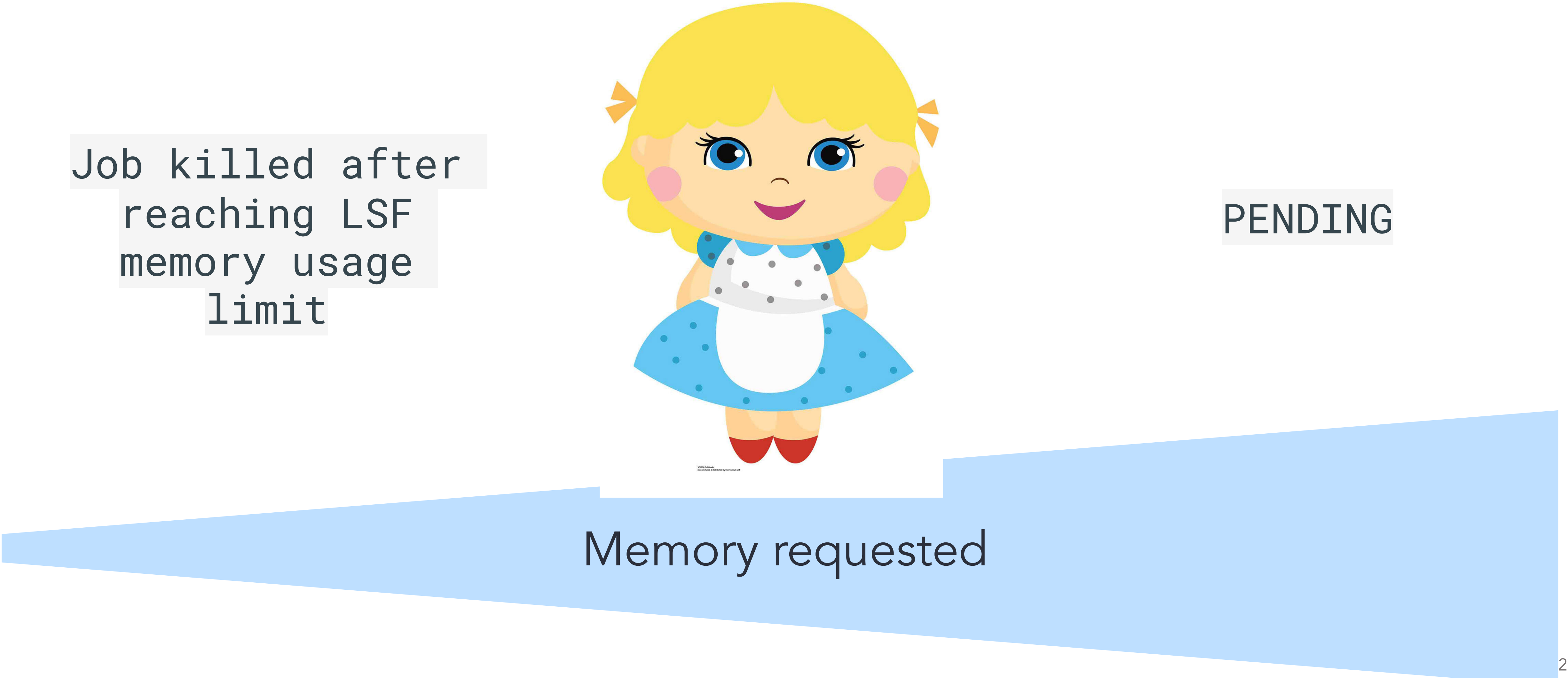
- Maximum memory in MB
- 1000
- Your job will terminate if you exceed this

-n number of nodes

- How many nodes
- 1
- 1GB per node, so increase for more memory

https://re-docs.genomicsengland.co.uk/hpc_memory/

Memory requested



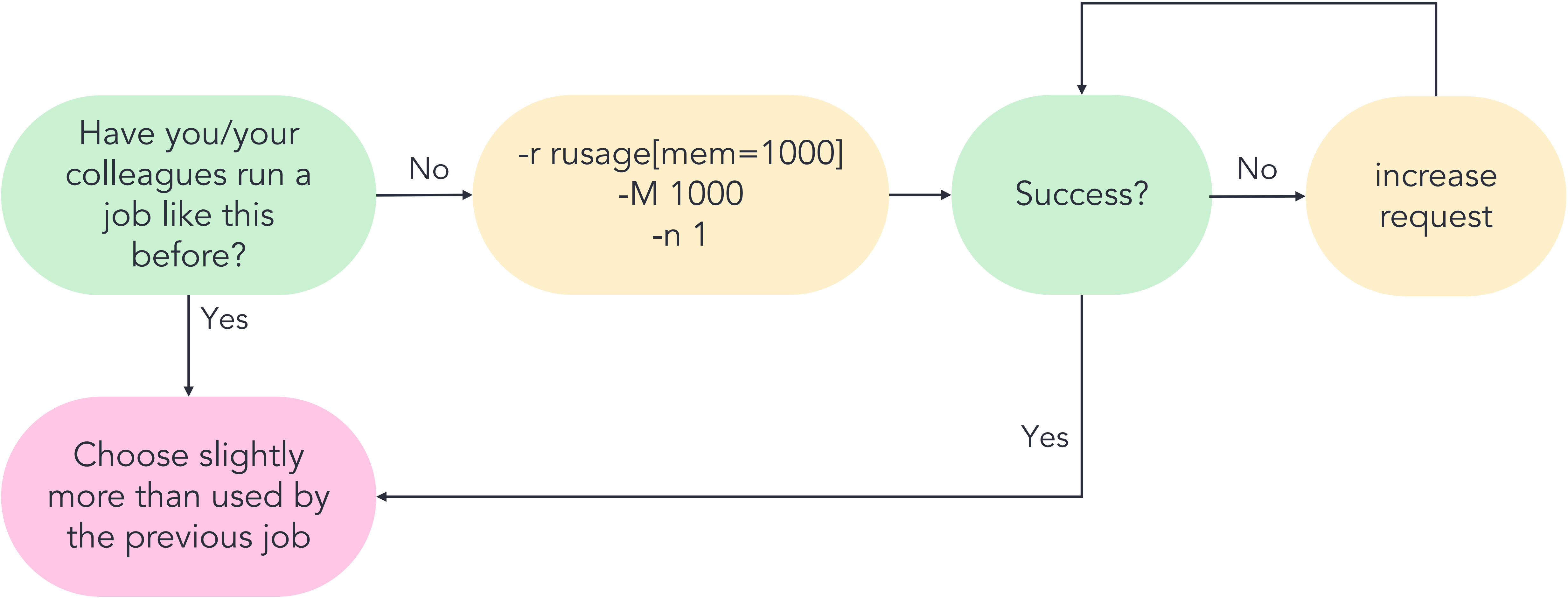
`Job killed after reaching LSF memory usage limit`

`PENDING`

Memory requested

https://re-docs.genomicsengland.co.uk/hpc_memory/

Setting your memory request



Have you/your colleagues run a job like this before?

No

-r rusage[mem=1000]
-M 1000
-n 1

Success?

No

increase request

Yes

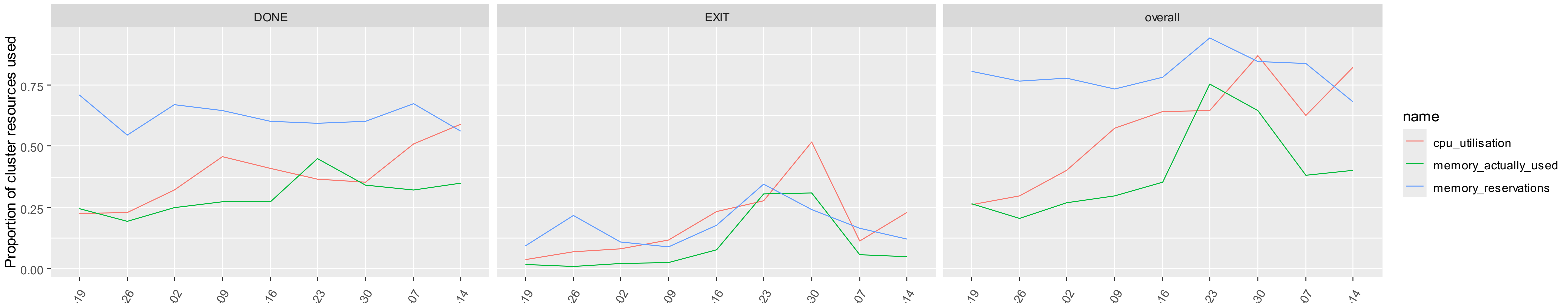
Choose slightly more than used by the previous job

Yes

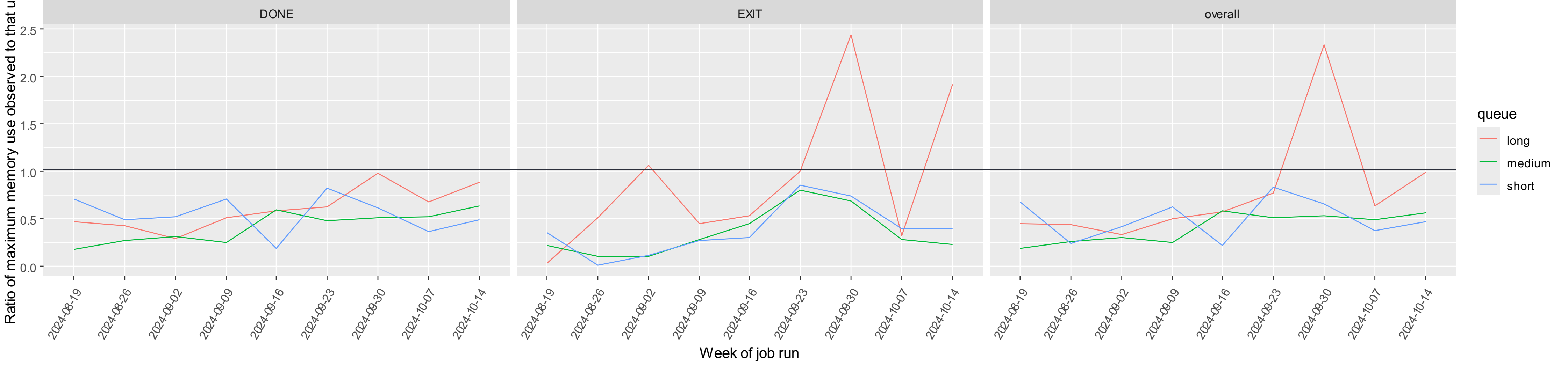
https://re-docs.genomicsengland.co.uk/hpc_memory/

Over-requested memory

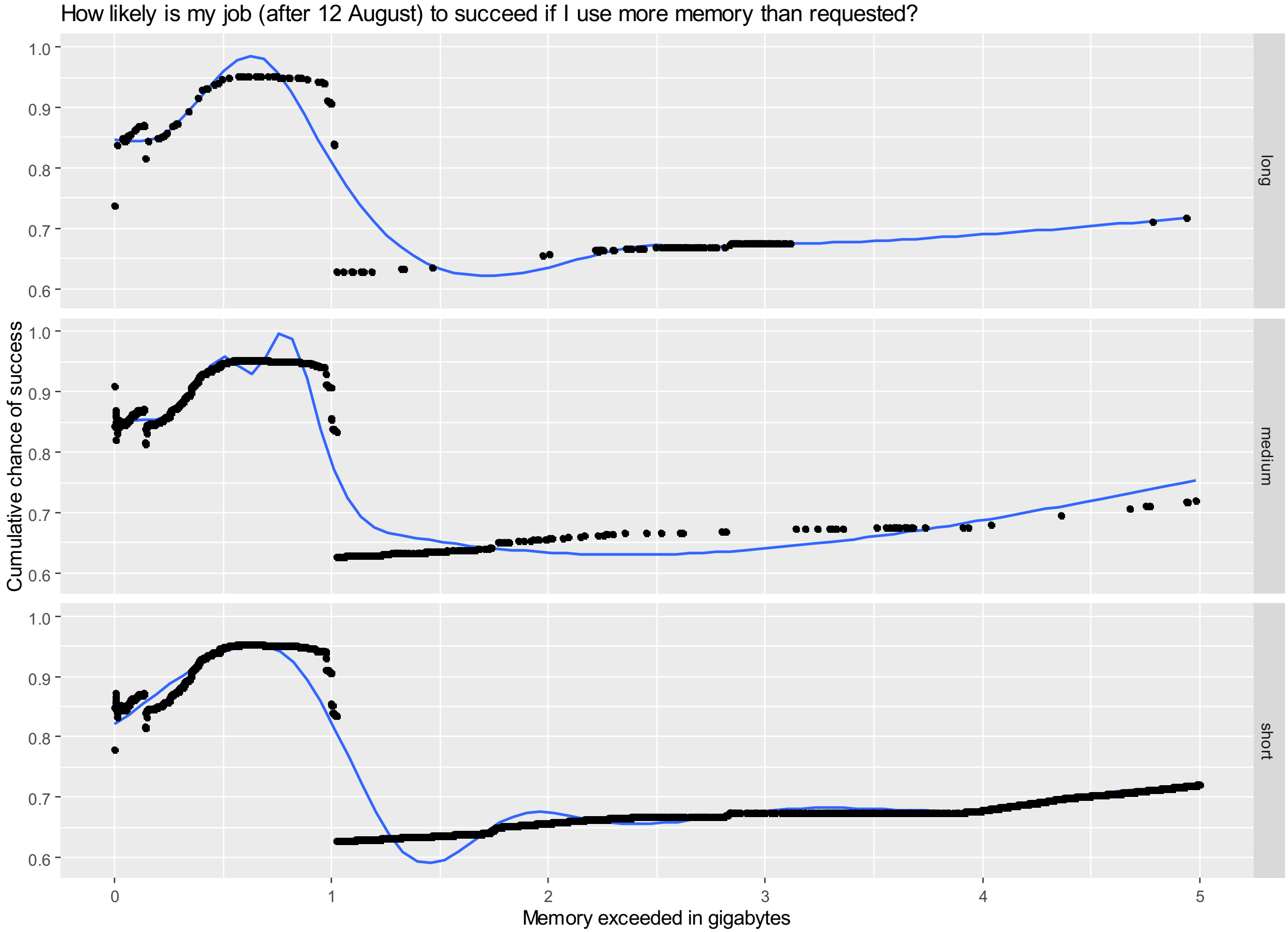
What is taking up cluster resources?



How much of the requested memory is being used?



How likely is my job to succeed if I use more memory than requested?

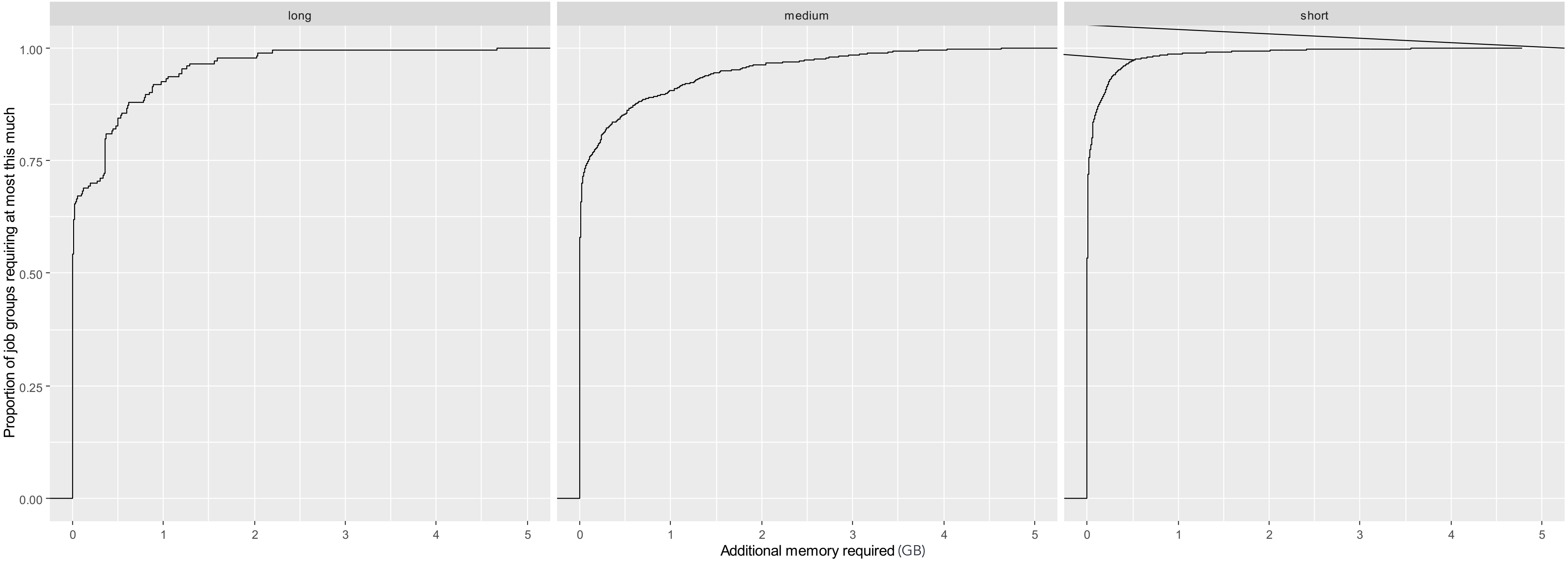


2.8% (62680) of jobs exceeded requested memory 19 Aug-20 Oct, of which 74% completed

Estimating memory for large job groups

Run the job for the first three samples → Check output for memory usage → Set memory request for 1.2 x maximum

Will my job groups need much more memory than if I look at the first 3?



About 2/3 (1,449,665/2,249,870) of jobs in that period were run as part of groups of 5+ jobs

Creating a job - parameters

-q	queue name: inter/short/medium/long
-P	Your project ID
-o	standard output location path_to/job.%J.out
-e	standard error location path_to/job.%J.err
-J	job name: useful for monitoring
-R	Memory usage
-M	Maximum memory
-n	number of nodes
-cwd	current working directory

bsub

https://re-docs.genomicsengland.co.uk/hpc_scripts/

A job on the inter queue

```
bsub -q inter -P <your_project_code> -R
rusage[mem=1000] –M 1000 –n 1 -Is bin/bash
```

Do some work

`bjobs` (to find job number)

```
bkill <job number>
```

https://re-docs.genomicsengland.co.uk/hpc_queues/#the-inter-queue

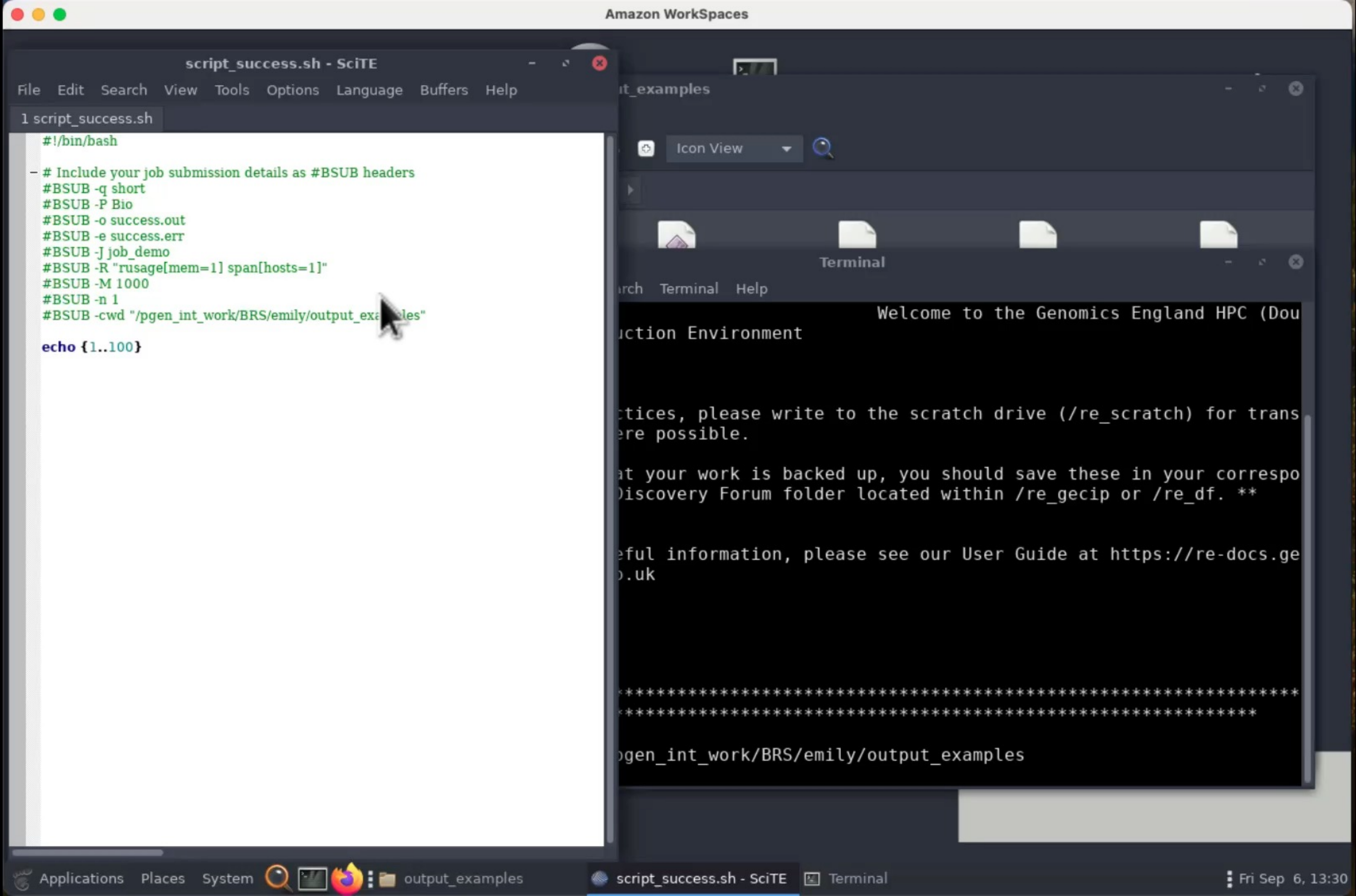
A job on a batch queue – shell script

```
#!/bin/bash
# Include your job submission details as #BSUB headers
#BSUB -q <your_queue>
#BSUB -P <yourProject>
#BSUB -o <path_to/job.%J.out>
#BSUB -e <path_to/job.%J.err>
#BSUB -J <jobName>
#BSUB -R "rusage[mem=1000] span[hosts=1]"
#BSUB -M <max_memory_in_MB>
#BSUB -n <number_of_cores>
#BSUB -cwd <"your_dir">

# Set your temp directory as the re_scratch folder export
TMPDIR=/re_scratch/re_gecip/<your_GECIP>/<your_username> export
TMPDIR=/re_scratch/re_discovery_forum/<your_discovery_forum_folder>/<your_username>

# Load any required modules from the HPC
 module load <moduleName>

# The actual script you want to run
```

```
bsub <
my_submission_script.sh
```

https://re-docs.genomicsengland.co.uk/hpc_scripts/

Job submission demo



```
#!/bin/bash

# Include your job submission details as #BSUB headers
#BSUB -q short
#BSUB -P Bio
#BSUB -o success.out
#BSUB -e success.err
#BSUB -J job_demo
#BSUB -R "rusage[mem=1] span[hosts=1]"
#BSUB -M 1000
#BSUB -n 1
#BSUB -cwd "/pgen_int_work/BRS/emily/output_examples"

echo {1..100}
```

```
                    Welcome to the Genomics England HPC (Dou
uction Environment



ctices, please write to the scratch drive (/re_scratch) for trans
ere possible.

at your work is backed up, you should save these in your correspo
Discovery Forum folder located within /re_gecip or /re_df. **


eful information, please see our User Guide at https://re-docs.ge
o.uk






*****************************************************************
*************************************************************

pgen_int_work/BRS/emily/output_examples
```

5. Tools and software available and how to load them

Software on the HPC

R, Python, Java, Perl

Nextflow

bioinformatics tools

singularity

```
delly/1.2.6
denovoGear/1.1.1
discover/0.9.5
dotnet/2.0.0
dotnet/8.0.1 (D)
drop/1.2.4
eigen/3.3.9
exomiser/13.3.0
exonerate/2.2.0
fastqc/0.12.1
fetk/1.9.3
ffmpeg/6.0
fribidi/1.0.12
gatk/4.5.0.0
gauchian/1.0.2
gcc/10.4.0
gcta/1.94
gdal/3.7.0
geos/3.12.1
gistic/2.0.23
gmp/6.2.1
gnu-parallel/20190222
gnu/4.4
gradle/8.5
guppy/3.4.5
gvcfgenotyper/2019.02.26
haplocheck/1.3.3
hipstr/0.7
hisat2/2.2.1
hla-la/1.0.3
hmftools/2024-02-06
homer/4.11
htslib/1.18
igv/2.17.1
imagemagick/7.1.0
java/1.8
java/11.0.2
java/17.0.2
java/19.0.2 (D)
jq/1.7.1
kallisto/0.50.1
king/2.3.2
kraken/1.1.1
kraken2/2.1.3
lapack/3.12.0
ldsc/1.0.1
singularity/3.8.3

pindel/0.2.5b8
platypus/0.8.1
plink/1.9
plink/2.00a3.3LM
plink/2.0 (D)
plink_seq/0.10
popdel/1.5.0
proj/8.2.1
prsice-2/2.3.5
pycircos/1.0.2
pysam/0.22.0
python/3.8
python/3.8.1
python/3.11 (D)
readline/8.0
regenie/3.4.1
repeatDetector/1.0
rtg-tools/3.12.1
rvtests/2.1.0
saige/1.0.9
salmon/1.10.0
samtools/1.16.1
shapeit4/4.2.2
sniffles/1.0.11
somalier/0.2.19
sqlite3/3.40.0
squirls/2.0.1
stack/2.15.7
star/2.7.2a
star/2.7.11a (D)
strelka/2.9.10
superSTR/1.0.1
svanna/1.0.4
tabix/1.18
trimmomatic/0.39
udunits/2.2.28
vcf2maf/1.6.21
vcfanno/0.3.4
vcflib/1.0.9
vcftools/0.1.16
verifyBamID/2.0.1
vt/0.57721
xz/5.4.7
zlib/1.3
zulu/21.0.1
aws-cli/2.15
singularity/4.1.1 (D)

AdapterRemoval/2.3.3
AutoDock_Vina/1.2.5
BWA/0.7.17
BerkeleyDB/3.01
Bio-DB-HTS/3.01
CADD/1.6
CNView/1.0
CNVnator/0.4.1
CaVEMan/1.15.3
ExpansionHunter/3.2.2
ExpansionHunter/4.0.2 (D)
ExpansionHunterDenovo/0.9.0
GSL/2.7
MEDICC2/1.0.2
MPFR/4.2.0
R/3.6.3
R/4.2.1
R/4.3.3 (D)
REViewer/0.2.7
aliview/1.28
ampliconArchitect/1.3.r7
ampliconClassifier/1.1.1
annotSV/3.3.7
annovar/2019Nov
annovar/2024-03-14 (D)
ant/1.9.16
apbs/3.4.1
asmc-asmc/2024-02-26
automake/1.15
bamtools/2.5.2
bcftools/1.16
beagle/5.4
bedops/2.4.41
bedtools/2.30.0
bedtools/2.31.0 (D)
blast+/2.15
blat/1.0
bolt-lmm/2.4.1
boost/1.83
bowtie2/2.5.2
canvas/1.40.0.1613
circos/0.69-9
clang/16.0.6
cmake/3.24.3
cpan/1.7047
cromwell/v65
curl/7.81.0
cython/3.0.8
cytoscape/3.10.1

ldstore/2.0
libdeflate/1.20
libgit2/1.6.2
libtiff/3.4
libtiff/4.3.0
libtiff/4.5.0 (D)
libunwind/1.8.0
liftover/1.0
linasm/1.13
llvm/16.0.6
locuszoom/1.4
lollipop/0.3.0
lumpy/0.3.1
mafft/7.520
magma/1.10
manta/1.6.0
matlab/8.1
matlab/24.1 (D)
maven/3.9.6
meme/5.5.5
metal/1.0
miniconda3/23.11.0
miniforge3/23.11.0-0
minimap2/2.26
mosaicHunter/2024-02-14
mplayer/1.5
msisensor-pro/1.2.0
msisensor/0.6
multiqc/1.19
music2/0.2
mutserve/2.0.0-rc15
mutsig2cv/3.11
ncurses/6.4
new_fugue/2010-06-02
nextflow/22.10.5
nextflow/23.04
nextflow/23.10-with-plugins
nextflow/23.10
nextflow/24.04.2-with-plugins (D)
nf-core/0.3.1
nf-test/0.7.3
nf-test/0.8.2
nf-test/0.9.0 (D)
nodejs/16.9.0
openrefine/3.7.4
openssl/1.1.1o
pandoc/3.3
perl/5.38.2
picard/3.1.1
```

https://re-docs.genomicsengland.co.uk/hpc_using_software/

Loading software

```
module avail myfavouritesoftware
```

list of available software including the string: "myfavouritesoftware"

```
module load myfavouritesoftware/3.2
```

https://re-docs.genomicsengland.co.uk/hpc_using_software/

R libraries

R/3.6.3
R/4.2.1
R/4.3.3

Library availability dependent on version of R

```
library(myfavouriteRlibrary)
```

https://re-docs.genomicsengland.co.uk/r_packages/

Conda environments

Pre-built environments

pandas

numpy

Labkey

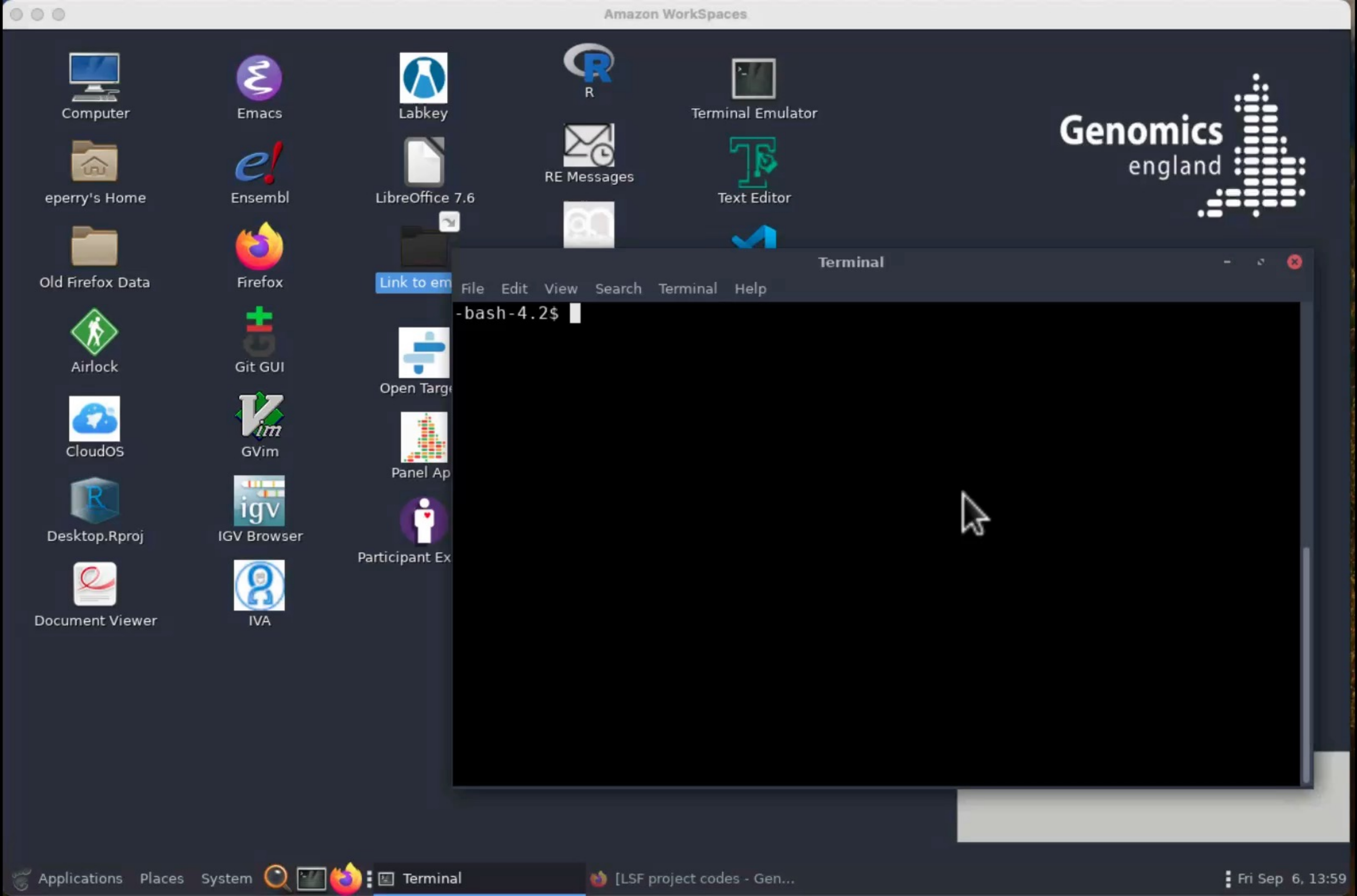
matplotlib

Jupyter

+ make your own

https://re-docs.genomicsengland.co.uk/hpc_conda/

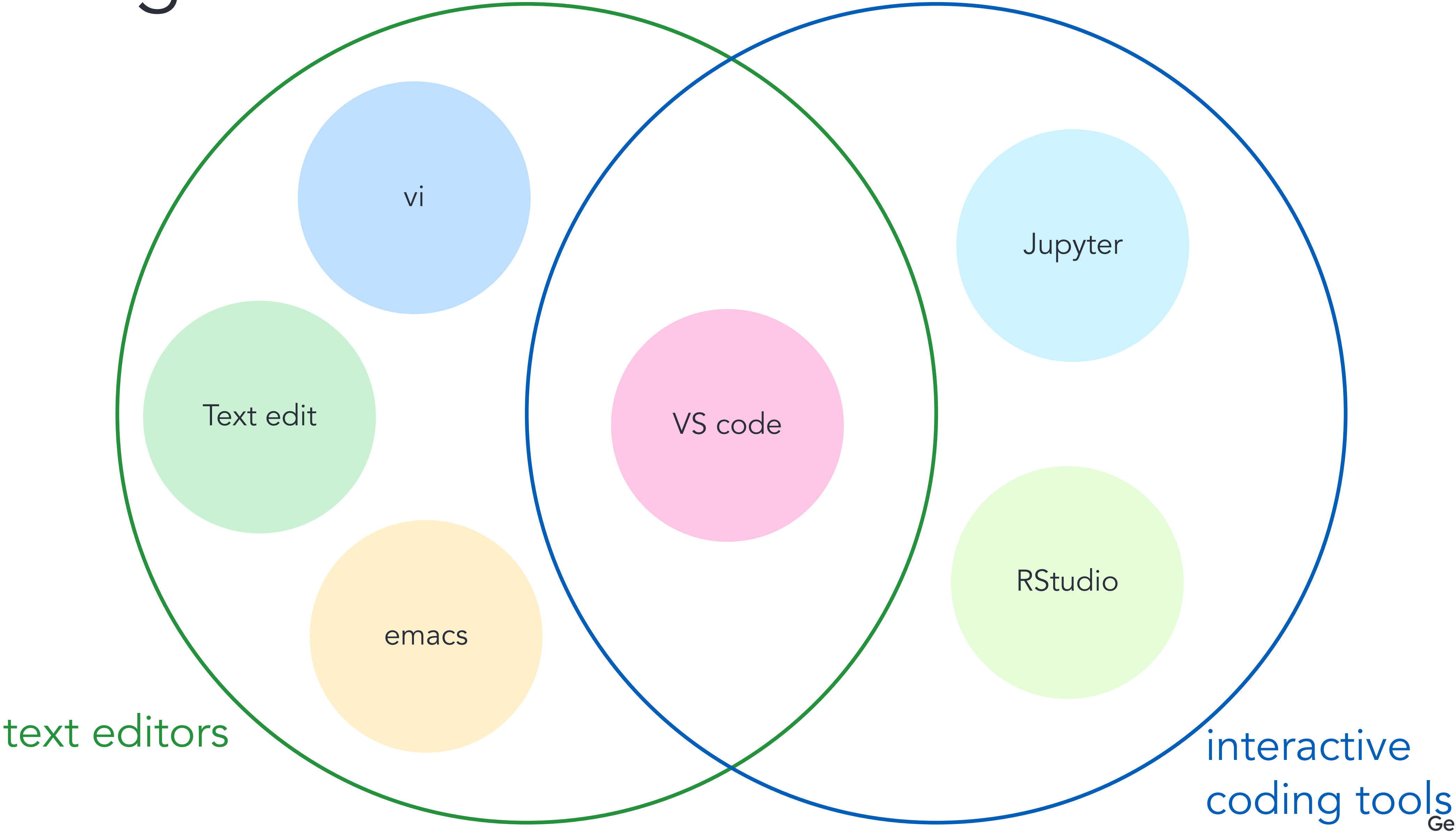
Loading software demo


Amazon WorkSpaces
Computer
Emacs
Labkey
R
Terminal Emulator
eperry's Home
Ensembl
LibreOffice 7.6
RE Messages
Text Editor
Old Firefox Data
Firefox
Link to em
Airlock
Git GUI
Open Targe
CloudOS
GVim
Panel Ap
Desktop.Rproj
IGV Browser
Participant Ex
Document Viewer
IVA
Genomics england
Terminal
File
Edit
View
Search
Terminal
Help
Applications
Places
System
Terminal
[LSF project codes - Gen...
Fri Sep 6, 13:59

```
-bash-4.2$
```

6. Interactive coding tools

Coding tools



vi

Text edit

emacs

VS code

Jupyter

RStudio

text editors

interactive coding tools

VScode

Syntax highlighting

Python debugging

Integrated terminal

curated set of extensions `~/public_data_resources/vscode_extensions/vsix`

https://re-docs.genomicsengland.co.uk/vscode/

Jupyter

- Combine text and code
- Interactive GUI
- available in conda environments

https://re-docs.genomicsengland.co.uk/hpc_jupyter/

RStudio

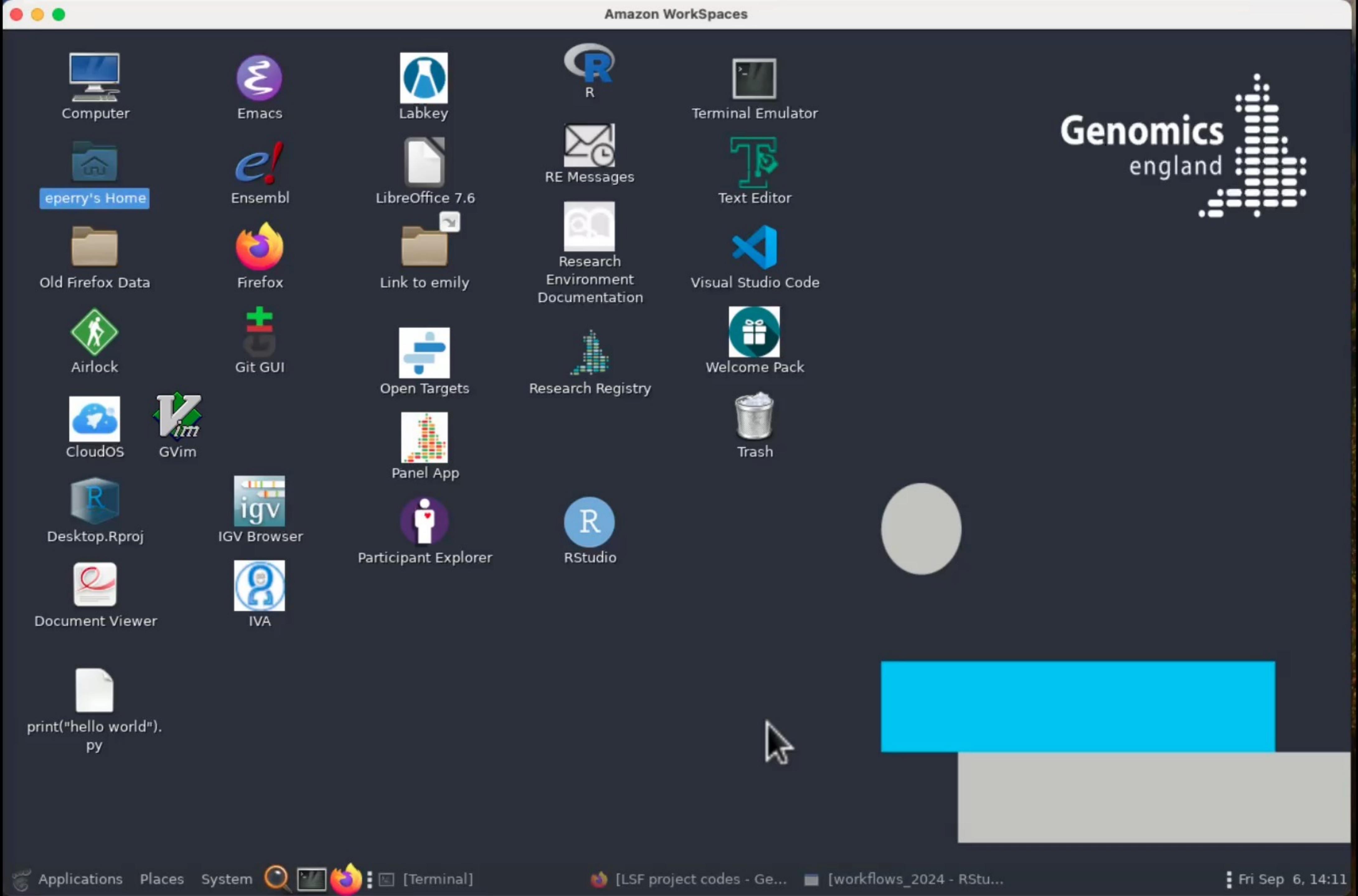
Interactive GUI

launch an HPC job from terminal

send code to terminal

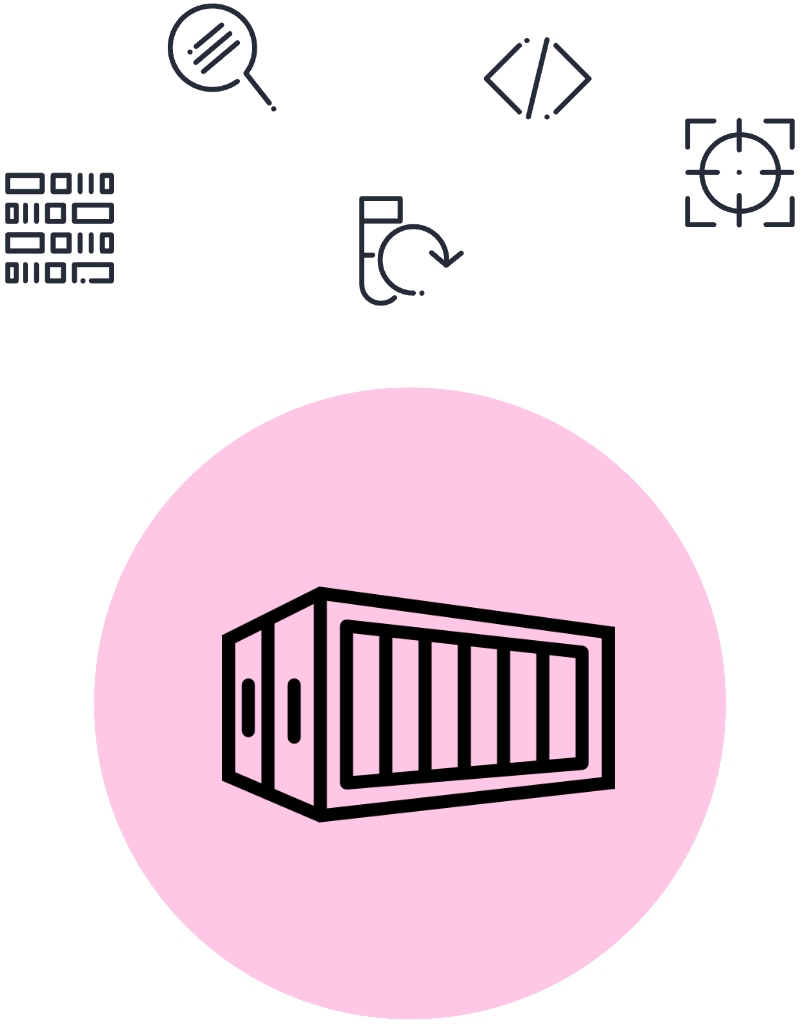
https://re-docs.genomicsengland.co.uk/enable_rstudio/

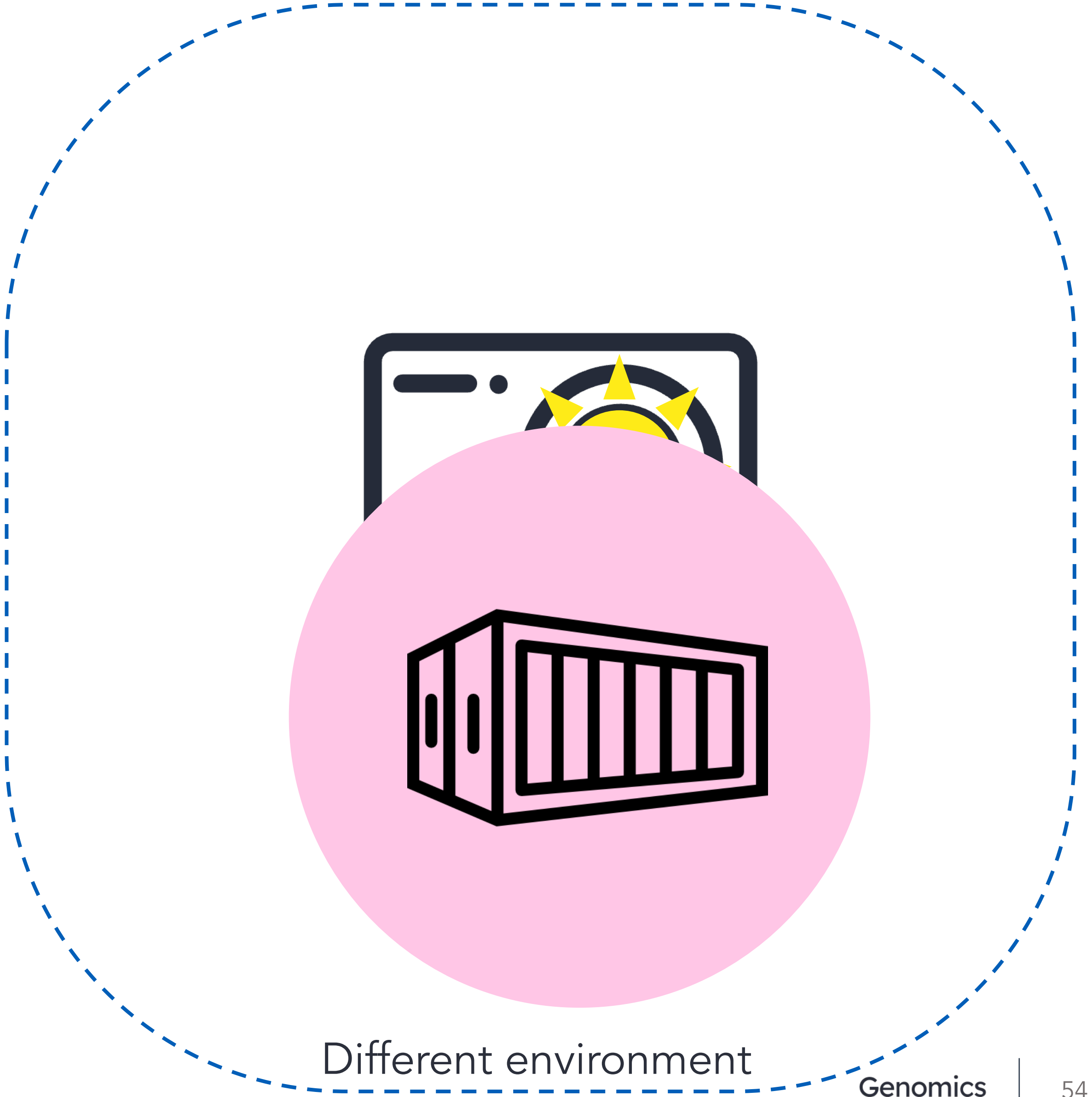
Coding tools demo


Amazon WorkSpaces
Computer
Emacs
Labkey
R
Terminal Emulator
eperry's Home
Ensembl
LibreOffice 7.6
RE Messages
Text Editor
Old Firefox Data
Firefox
Link to emily
Research Environment Documentation
Visual Studio Code
Airlock
Git GUI
Open Targets
Research Registry
Welcome Pack
CloudOS
GVim
Panel App
Trash
Desktop.Rproj
IGV Browser
Participant Explorer
RStudio
Document Viewer
IVA
print("hello world").py
Genomics england
Applications
Places
System
[Terminal]
[LSF project codes - Ge...
[workflows_2024 - RStu...
Fri Sep 6, 14:11

7. Bringing in your own tools and software

Containers

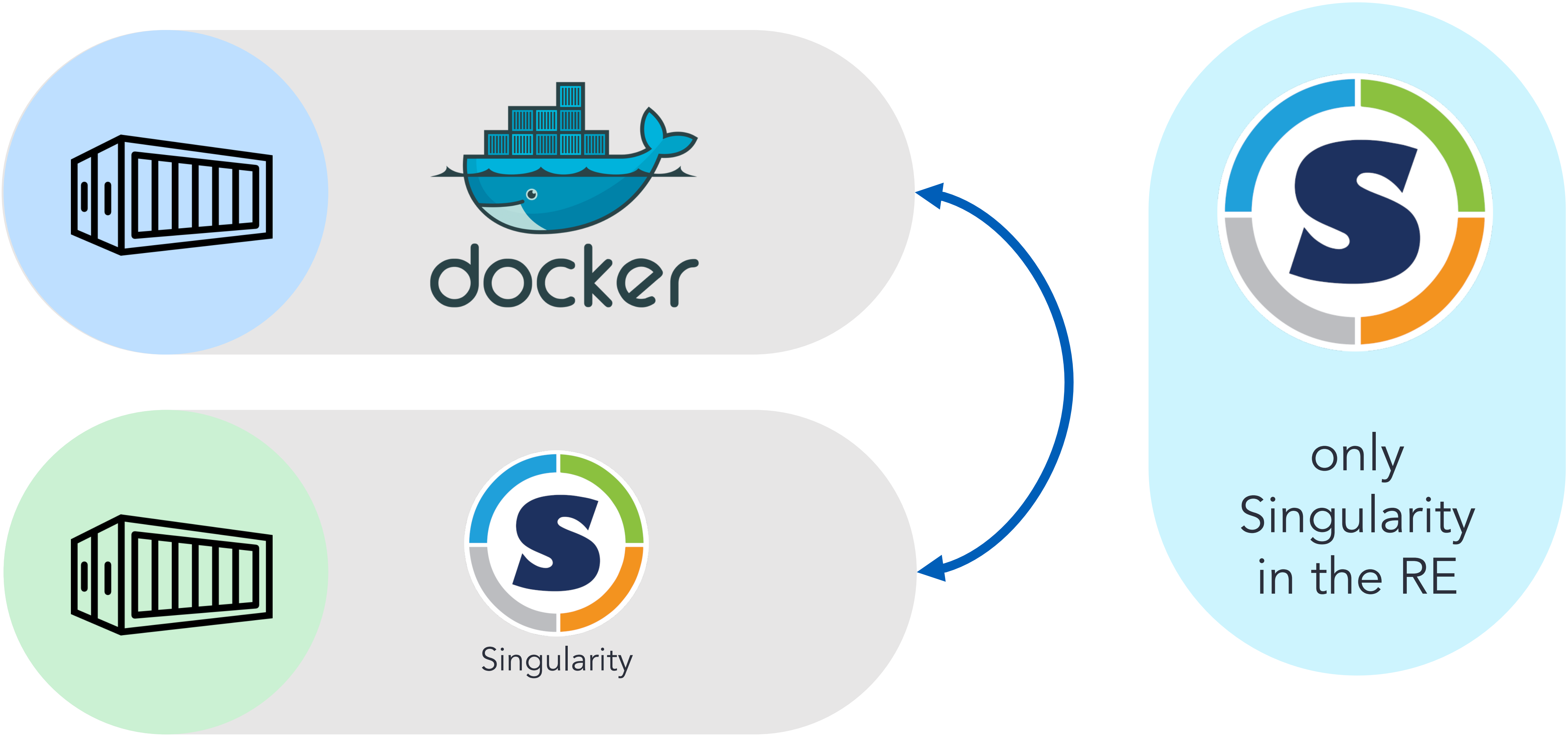




Different environment

https://re-docs.genomicsengland.co.uk/hpc_containers/

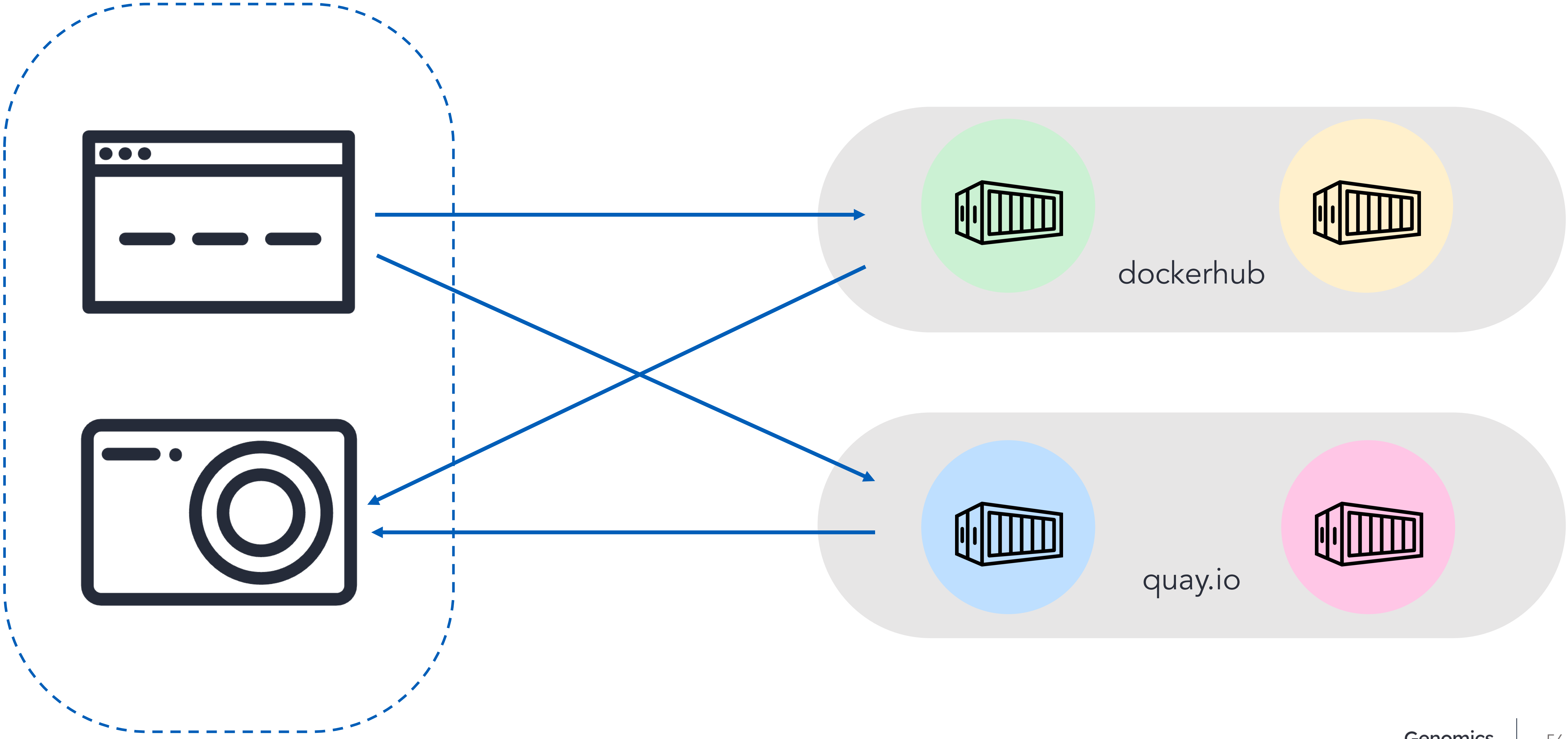
Container types



only
Singularity
in the RE

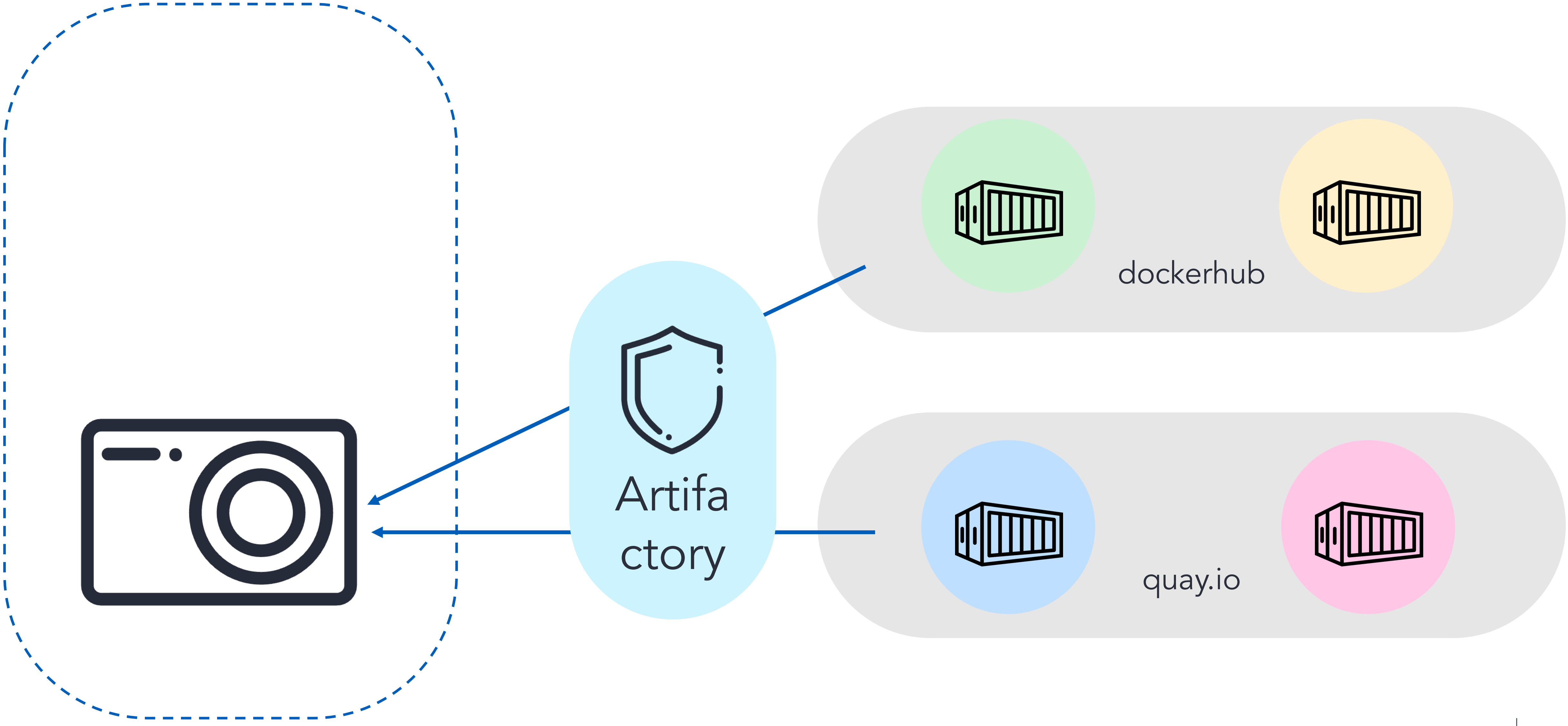
https://re-docs.genomicsengland.co.uk/hpc_containers/

Container repositories



https://re-docs.genomicsengland.co.uk/hpc_containers/

Containers in the RE

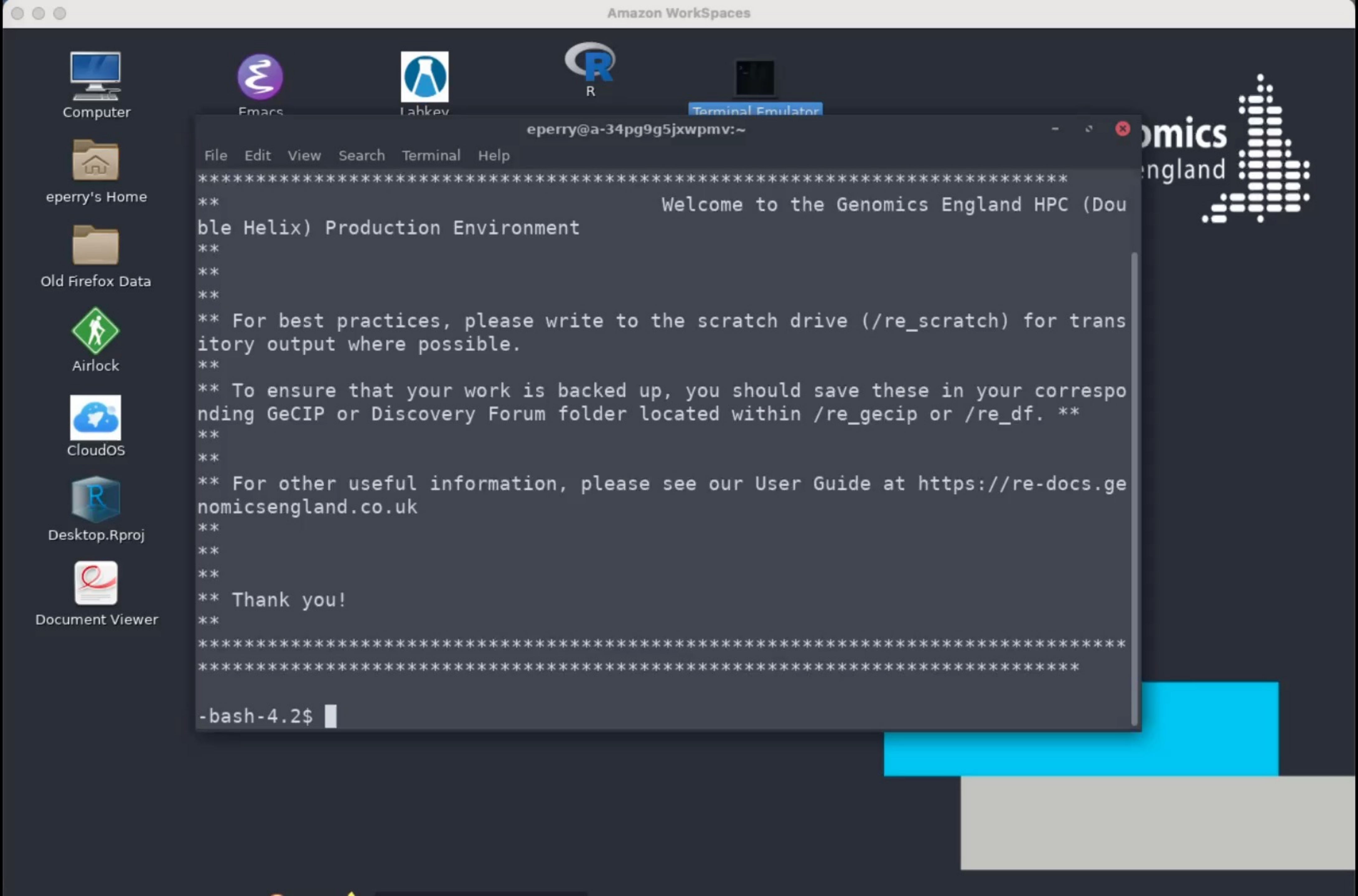


https://re-docs.genomicsengland.co.uk/hpc_containers/

Steps to import container

1. Find container in repository: identify docker or quay.io container location
2. modify import location to include artifactory redirect
3. pull container from an interactive job on the HPC
4. mount data and run analysis

Singularity demo



```
***************************************************************************
**                                 Welcome to the Genomics England HPC (Dou
ble Helix) Production Environment
**
**
**
** For best practices, please write to the scratch drive (/re_scratch) for trans
itory output where possible.
**
** To ensure that your work is backed up, you should save these in your correspo
nding GeCIP or Discovery Forum folder located within /re_gecip or /re_df. **
**
**
** For other useful information, please see our User Guide at https://re-docs.ge
nomicsengland.co.uk
**
**
**
** Thank you!
**
********************************************************************************
****************************************************************************

-bash-4.2$ 
```

8. CloudOS – batch and interactive jobs on the Cloud

What is CloudOS?

- Collaborative Environment
- All in one Platform

Using the Cloud

CloudOS uses multiple AWS resources

Wide-range of CPU and GPU Instances available to select from

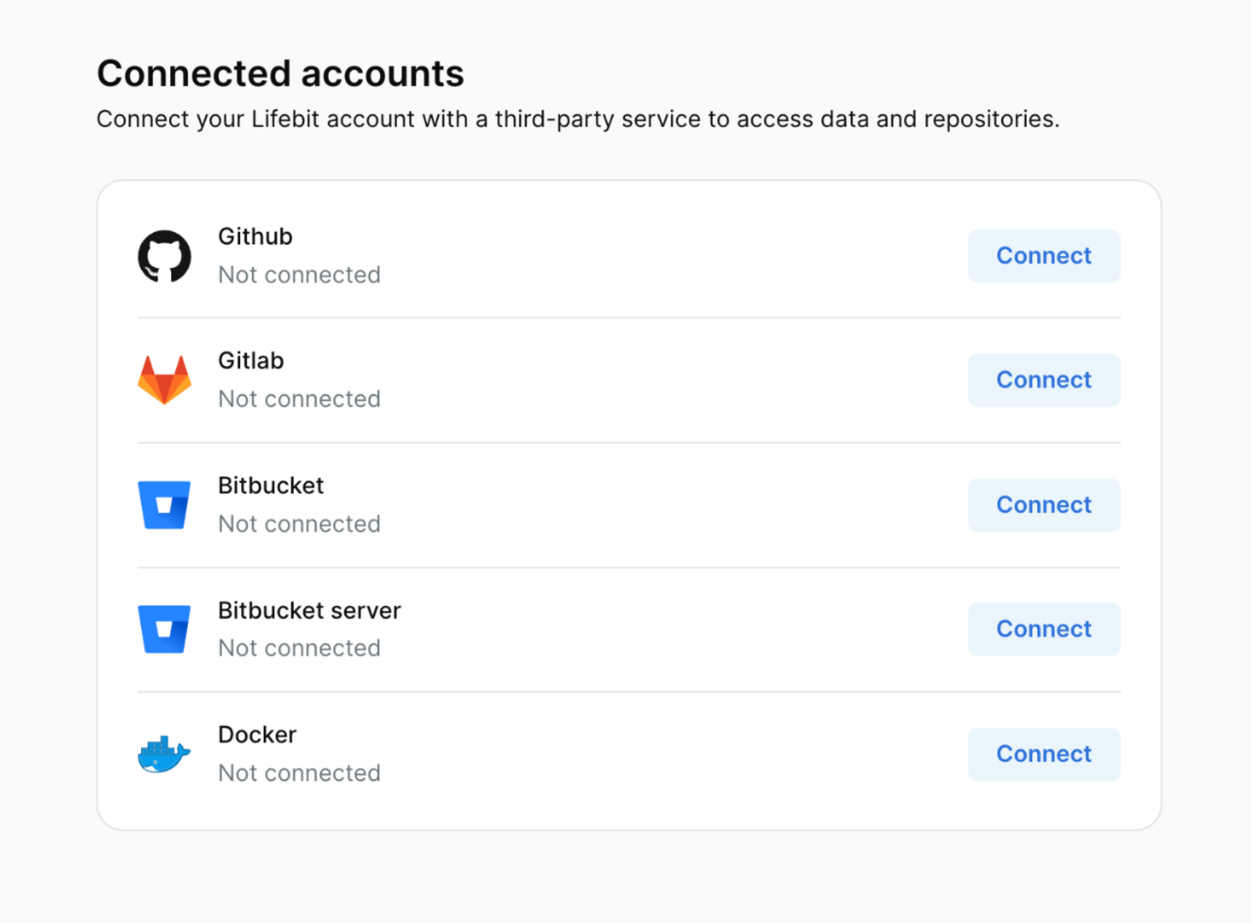
Interactive Sessions on Cloud allow you to spin up an EC2 instance on demand to run RStudio, Jupyter Notebook or Spark Notebook sessions

You have flexibility in creating environments, managing packages and bringing in data from repositories

Working interactively is limited by instance, designed for smaller analyses

Nextflow/WDL Batch Jobs can utilise multiple AWS instances

Setting up your Account



Connected accounts

Connect your Lifebit account with a third-party service to access data and repositories.

Github	Not connected	Connect
Gitlab	Not connected	Connect
Bitbucket	Not connected	Connect
Bitbucket server	Not connected	Connect
Docker	Not connected	Connect

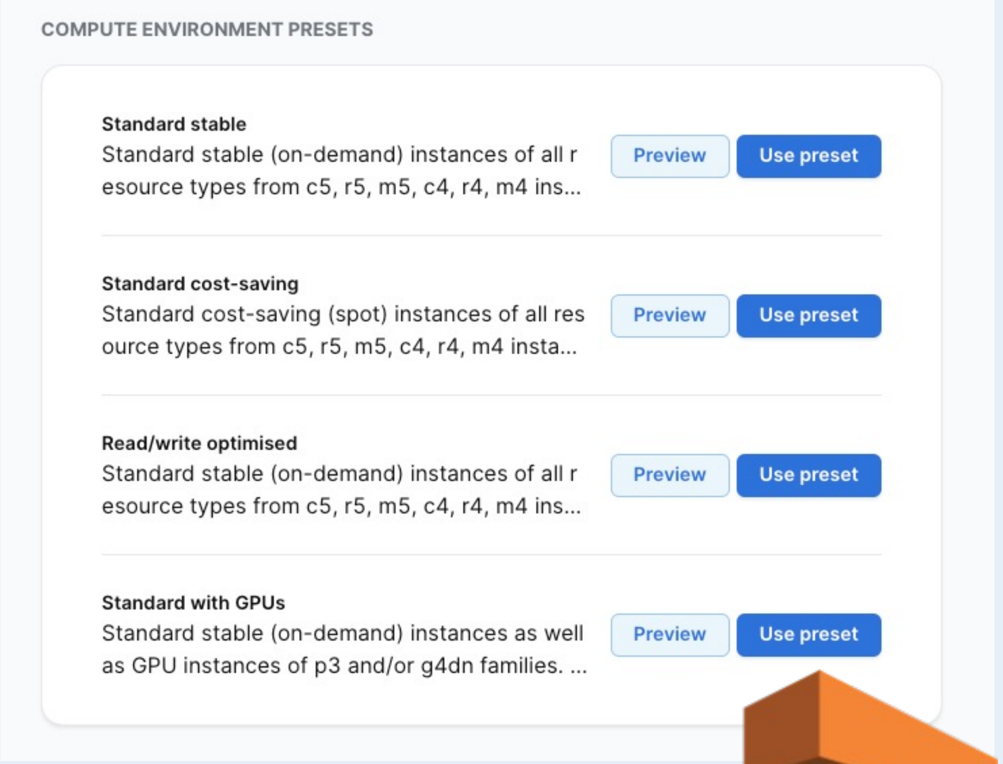
Link your private repository accounts

Access your pipelines and code securely

https://lifebit.atlassian.net/wiki/spaces/CD/pages/814941321/Link+and+remove+my+repository+accounts

Batch Queues

- CloudOS enables you to configure the compute resources and create job queues optimised for specific tasks
- Choose from a list of pre-configured environments
- Complete flexibility to customise vCPUs, Volume type, IOPS, On-demand or Spot instance types



COMPUTE ENVIRONMENT PRESETS

Standard stable
Standard stable (on-demand) instances of all r
esource types from c5, r5, m5, c4, r4, m4 ins...

Standard cost-saving
Standard cost-saving (spot) instances of all res
ource types from c5, r5, m5, c4, r4, m4 insta...

Read/write optimised
Standard stable (on-demand) instances of all r
esource types from c5, r5, m5, c4, r4, m4 ins...

Standard with GPUs
Standard stable (on-demand) instances as well
as GPU instances of p3 and/or g4dn families. ...

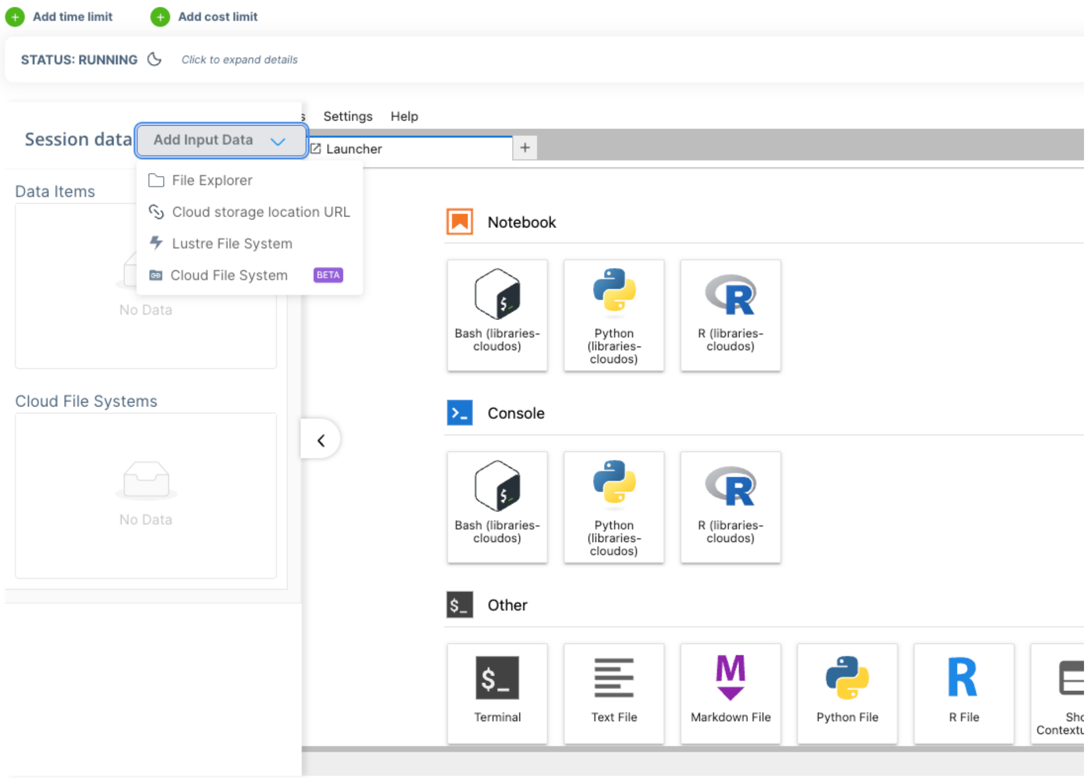
AWS Batch

https://lifebit.atlassian.net/wiki/spaces/CD/pages/816448106/Create+and+Monitor+Batch+Queu

Account setup demo

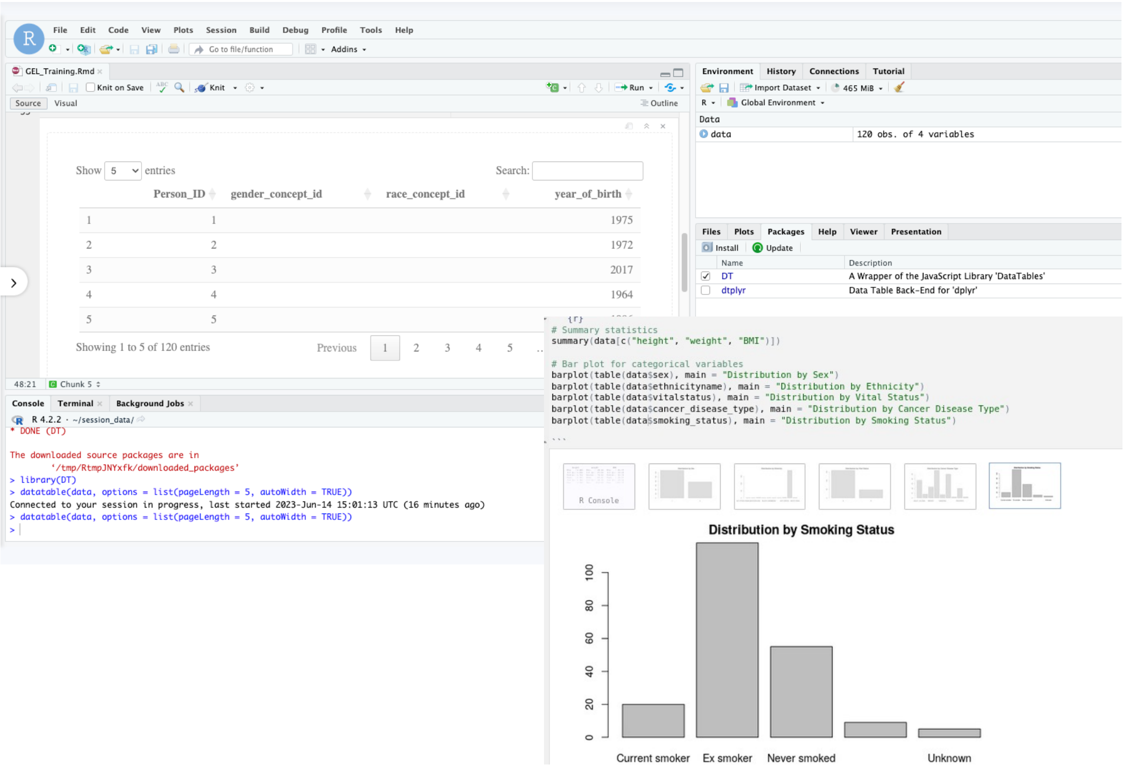
Jupyter notebooks

- Set time and cost limits
- Choose between Bash, Python and R Notebooks
- Install the packages you want using Conda
- Run Docker containers and Nextflow pipelines
- Save Snapshots of your session
- Access data programmatically
- Cloud File System for fast linking of large data files



https://lifebit.atlassian.net/wiki/spaces/CD/pages/813105781/Run+a+Jupyter+Notebook+session

RStudio



Show 5 entries Search:

	Person_ID	gender_concept_id	race_concept_id	year_of_birth
1	1			1975
2	2			1972
3	3			2017
4	4			1964
5	5			

Showing 1 to 5 of 120 entries

Previous 1 2 3 4 5 …

48:21 Chunk 5

Data

data 120 obs. of 4 variables

	Name	Description
☑	DT	A Wrapper of the JavaScript Library 'DataTables'
☐	dtplyr	Data Table Back-End for 'dplyr'

```
{r}
# Summary statistics
summary(data[c("height", "weight", "BMI")])

# Bar plot for categorical variables
barplot(table(data$sex), main = "Distribution by Sex")
barplot(table(data$ethnicityname), main = "Distribution by Ethnicity")
barplot(table(data$vitalstatus), main = "Distribution by Vital Status")
barplot(table(data$cancer_disease_type), main = "Distribution by Cancer Disease Type")
barplot(table(data$smoking_status), main = "Distribution by Smoking Status")
```

```
R 4.2.2 · ~/session_data/
* DONE (DT)

The downloaded source packages are in
        '/tmp/RtmpJNYxfk/downloaded_packages'
> library(DT)
> datatable(data, options = list(pageLength = 5, autoWidth = TRUE))
Connected to your session in progress, last started 2023-Jun-14 15:01:13 UTC (16 minutes ago)
> datatable(data, options = list(pageLength = 5, autoWidth = TRUE))
>
```

Build and Share
RMarkdowns and Shiny Apps

Install packages using
CRAN and Bioconductor

Save Snapshots of
the environment

https://lifebit.atlassian.net/wiki/spaces/CD/pages/813105278/Run+a+RStudio+session

Interactive Analysis demo

Pipelines and tools

GEL community tools

Small Variant
Workflow
Structural Variant

Lifebit system tools

GWAS
VEP
PRS
Burden Test
Burden Test Helper
Genomic Data QC

Add your own pipelines

NF-Core
Github
Gitlab
Bitbucket
Docker

- Easy to Setup and Configure (no command line needed)
- Monitor, Clone & Resume analyses
- Easy to debug errors
- Optimise your analysis e.g. GWAS on 100k participants ~3 hour runtime

Batch Analysis demo

9. Help and questions

Getting help

Check our documentation:
https://re-docs.genomicsengland.co.uk/
Click on the documentation icon in the environment

Contact our Service Desk:
https://jiraservicedesk.extge.co.uk/plugins/servlet/desk

Questions

All your microphones are muted

Use the Zoom Q&A to ask questions

Upvote your favourite questions: if we are short on time we will prioritise those with the most votes

Training sessions 2024

10/12 Introduction to the RE

↑ Materials from past training all online

https://re-docs.genomicsengland.co.uk/upcoming/

Training sessions 2025

3rd Tuesday every month

Introduction to the RE

21/1

18/2

18/3

15/4

20/5

17/6

https://re-docs.genomicsengland.co.uk/upcoming/

Materials from past training all online

Training sessions 2025

14/1	Using the Research Environment for clinical diagnostic discovery
11/2	Importing data and tools to use in the RE
11/3	Working with R in the RE
8/4	Working with python in the RE
13/5	Building cancer cohorts and survival analysis
10/6	Building rare disease cohorts with matching controls

https://re-docs.genomicsengland.co.uk/upcoming/

Materials from past training all online

In-person training day

- For any registered RE users
- 20th November 2024
- Held in our offices at Canary Wharf
- Hybrid attendance
- Hands-on practicals and exercises

https://www.genomicsengland.co.uk/events/research-environment-training-event-november-2024

Feedback

Thank you

Visit: https://re-docs.genomicsengland.co.uk/