



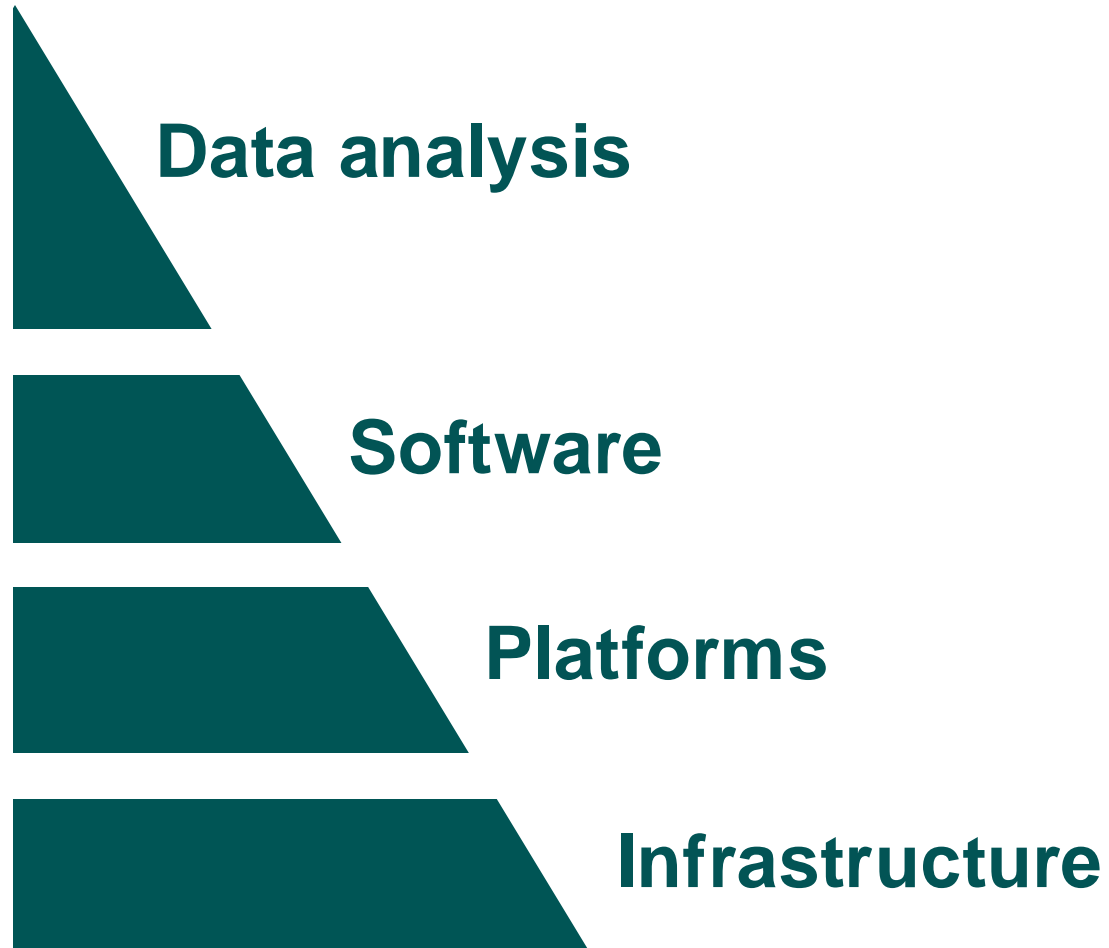
**ON CLOUD
FULL STACK
BIOINFORMATICS**

JORGE.BOUCAS@AGE.MPG.DE

HEAD OF BIOINFORMATICS



CLOUD ?

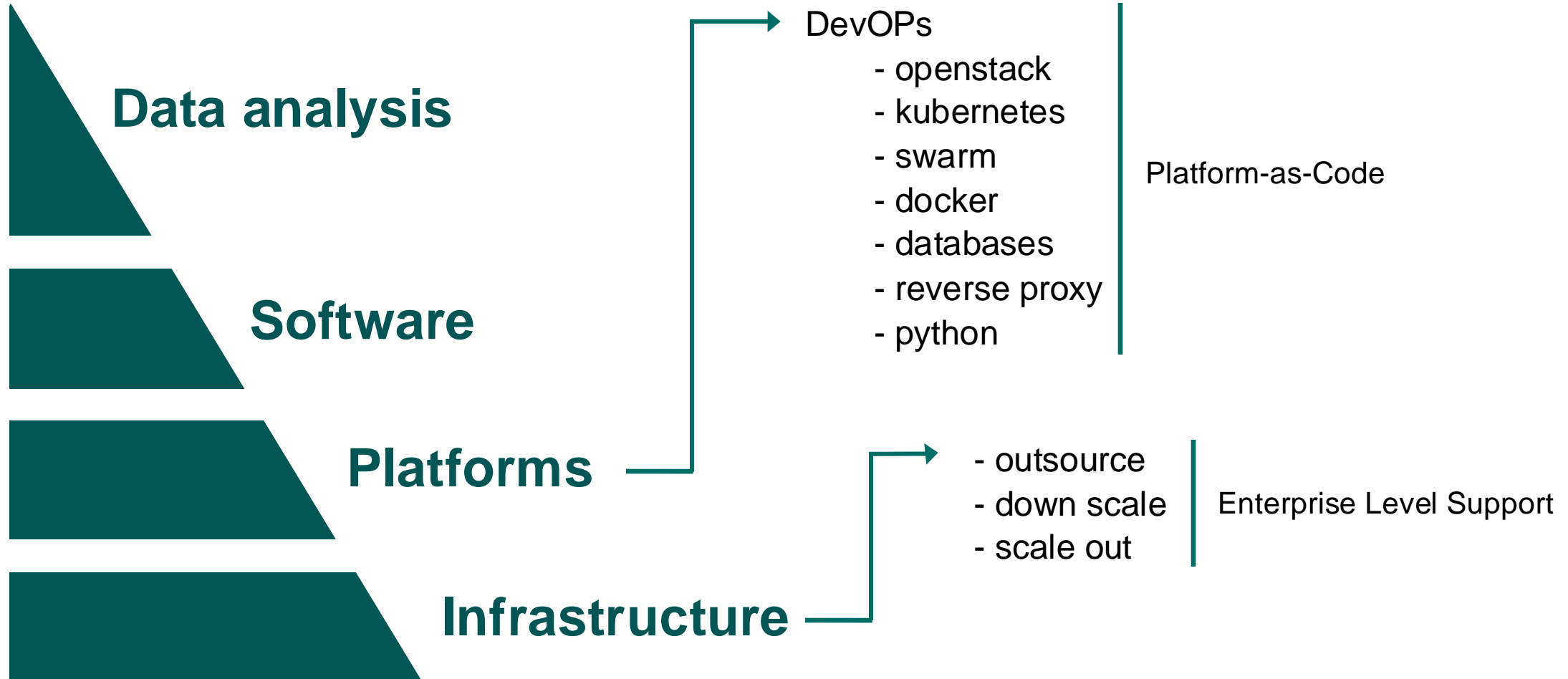


Bioinformaticians (2+1)

Systems Administrator (1)

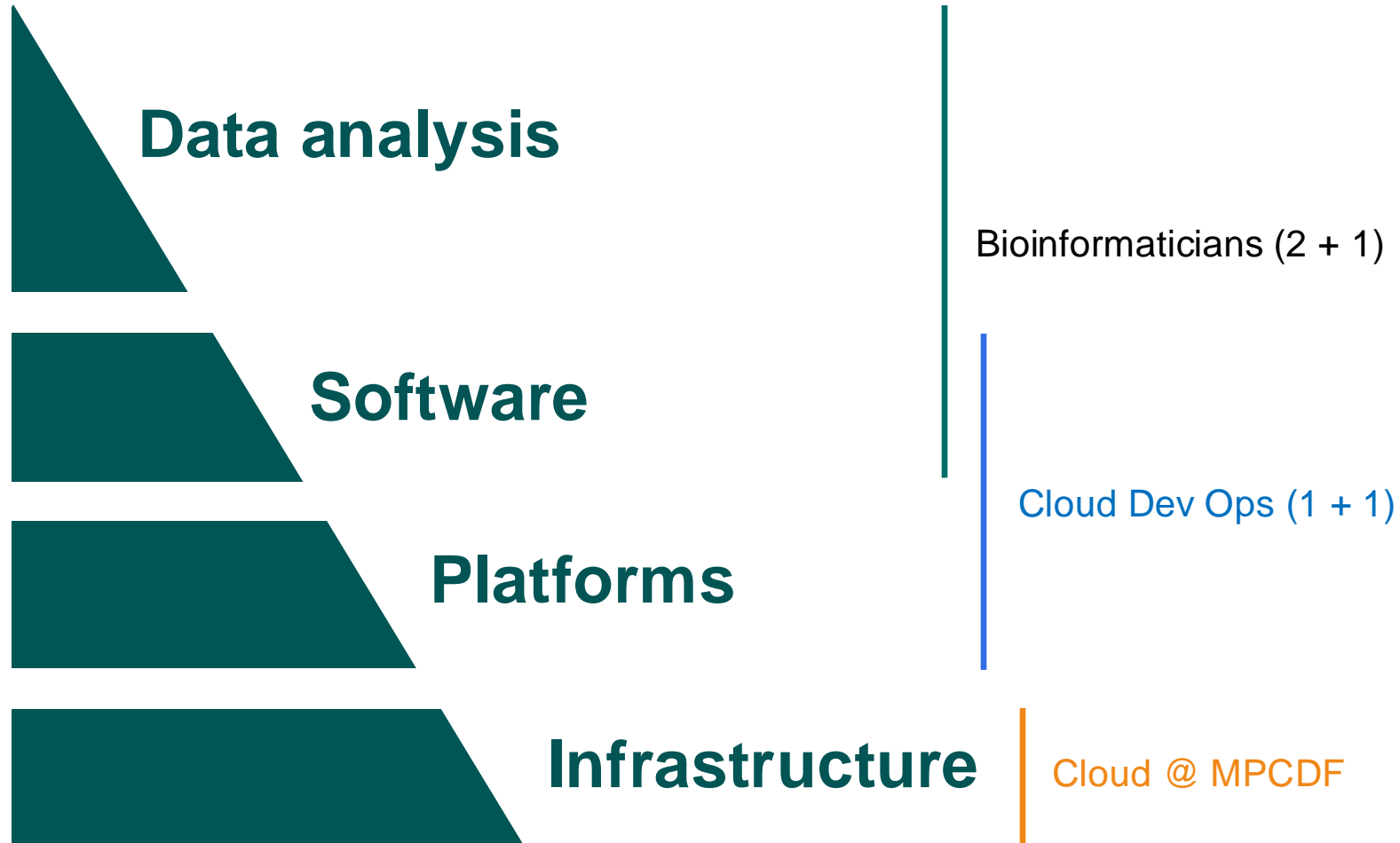


CLOUD ?





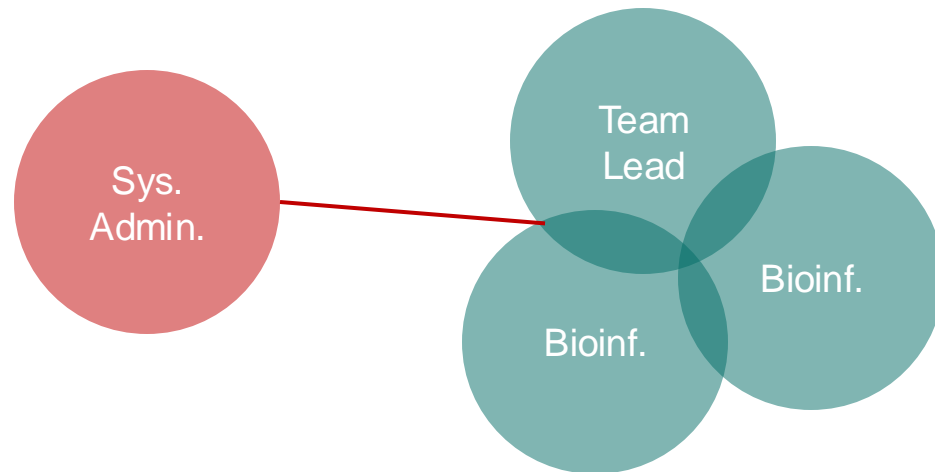
CLOUD ?



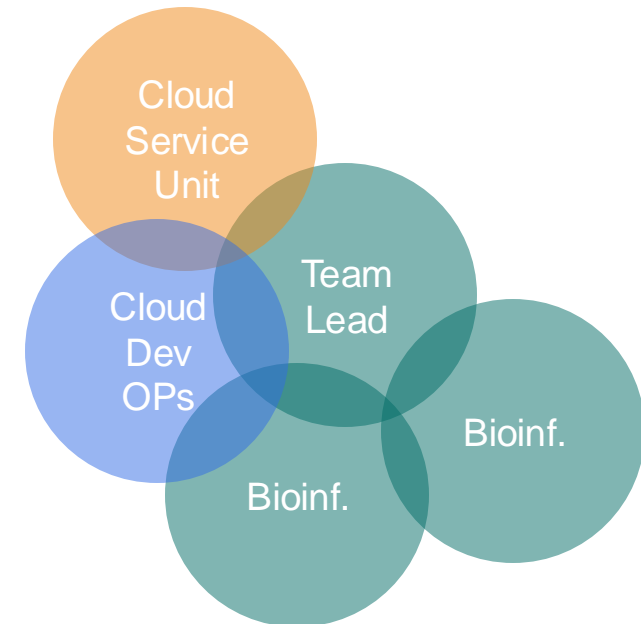


CLOUD ?

BEFORE









AFTER





THE BIOINFORMATICS STACK

<i>custom data analysis</i>	<i>software development</i>	<i>automated pipelines</i>
		
Python, R	Python, R	Python, R, bash, Perl, (C/C++)
		

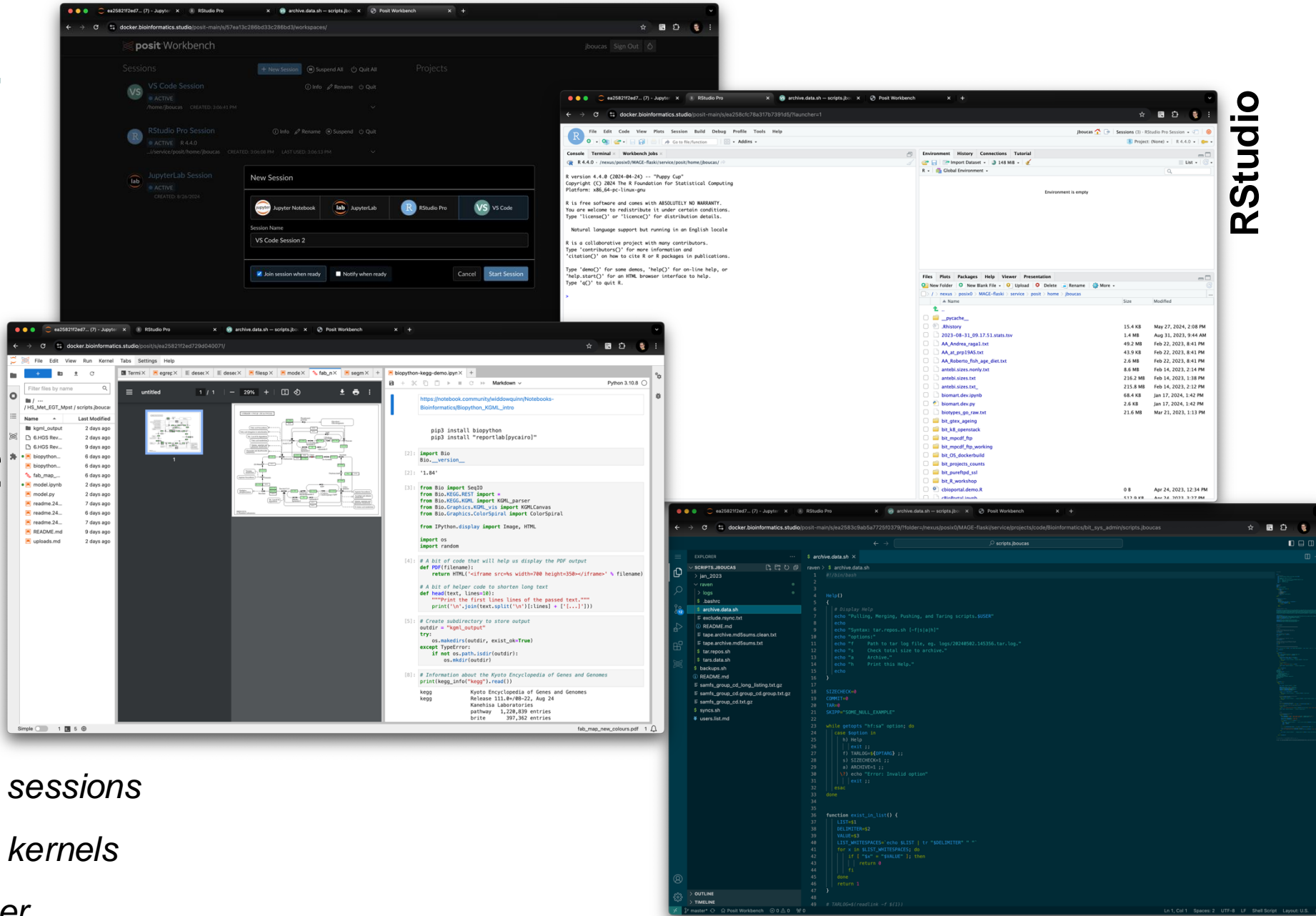


POSIT

JupyterLab

RStudio

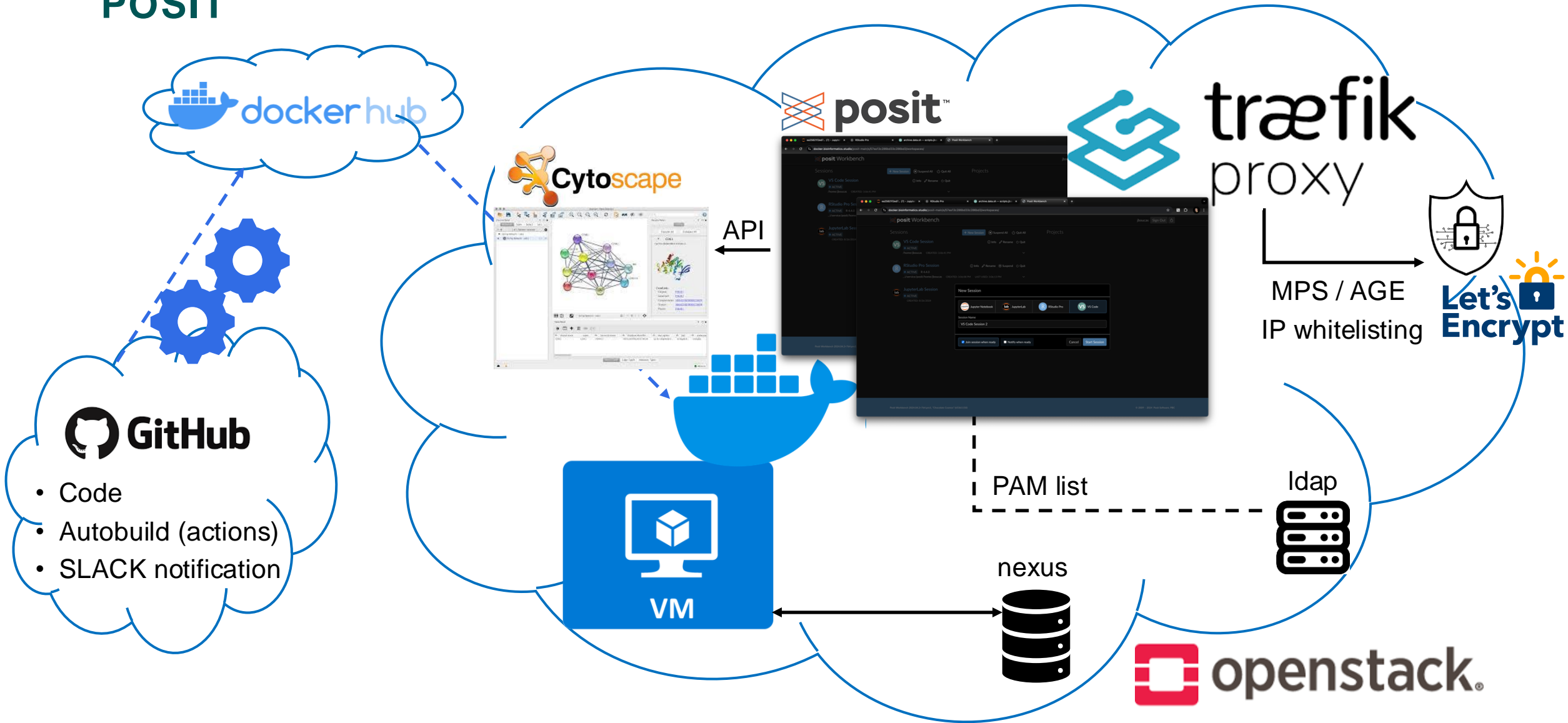
VS Code



- multiple sessions
- multiple kernels
- multi user



POSIT

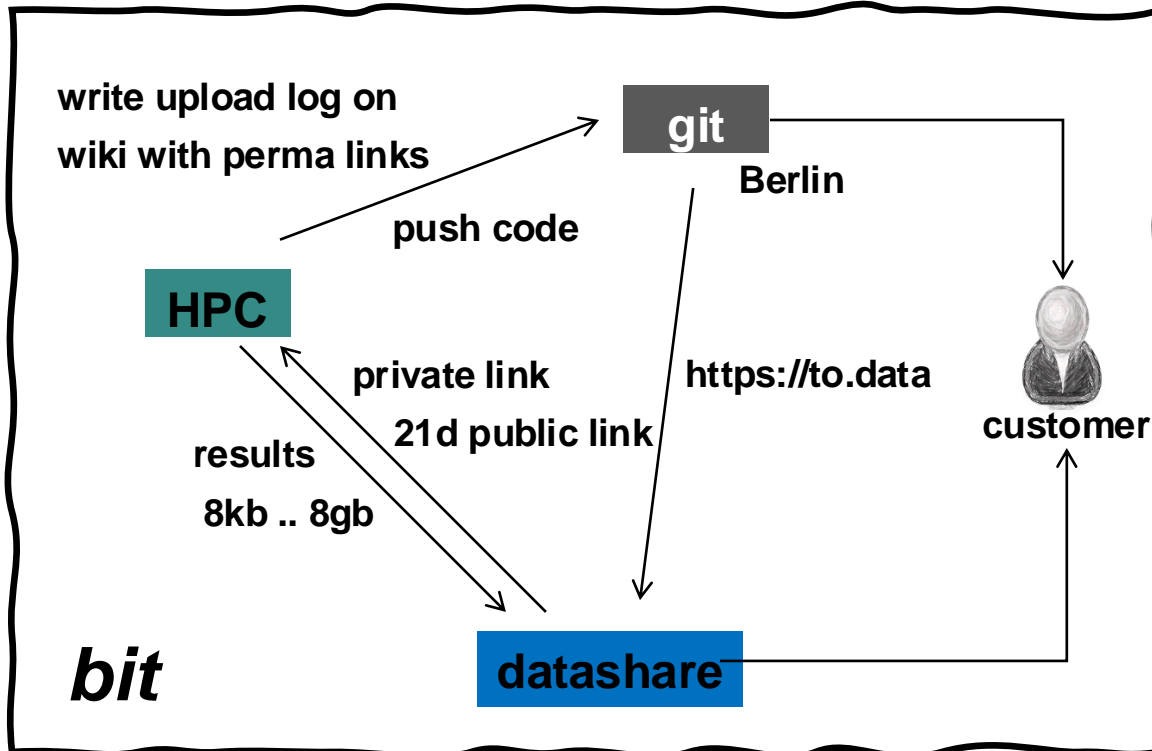
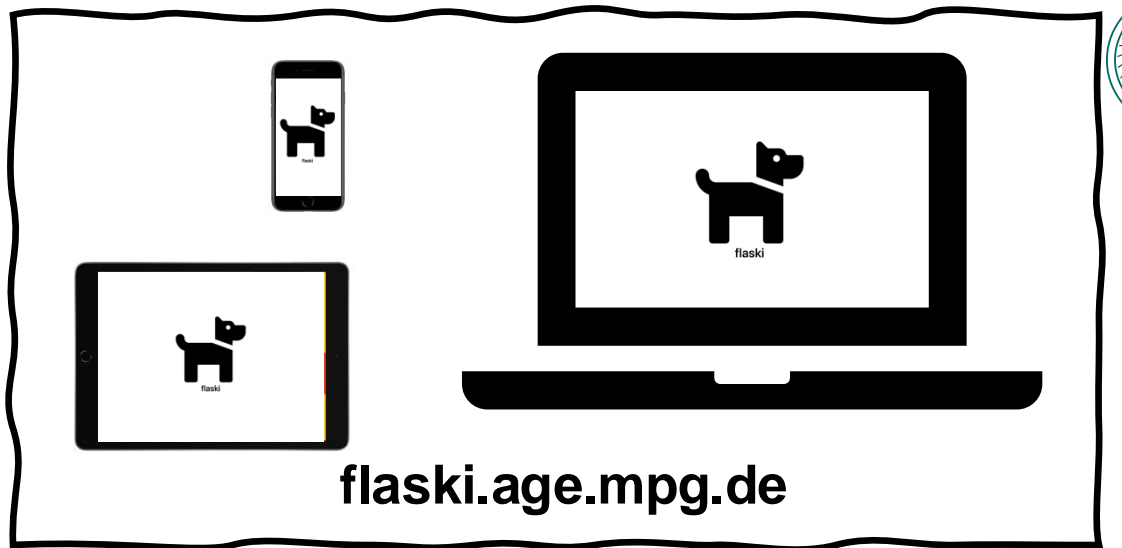


- Code
- Autobuild (actions)
- SLACK notification



SOFTWARE

- Operational
- Scientific
- WebApps



Otasek et al. *Genome Biology* (2019) 20:185
<https://doi.org/10.1186/s13059-019-1758-4>

Genome Biology

SOFTWARE Open Access

Check for updates

Cytoscape Automation: empowering workflow-based network analysis

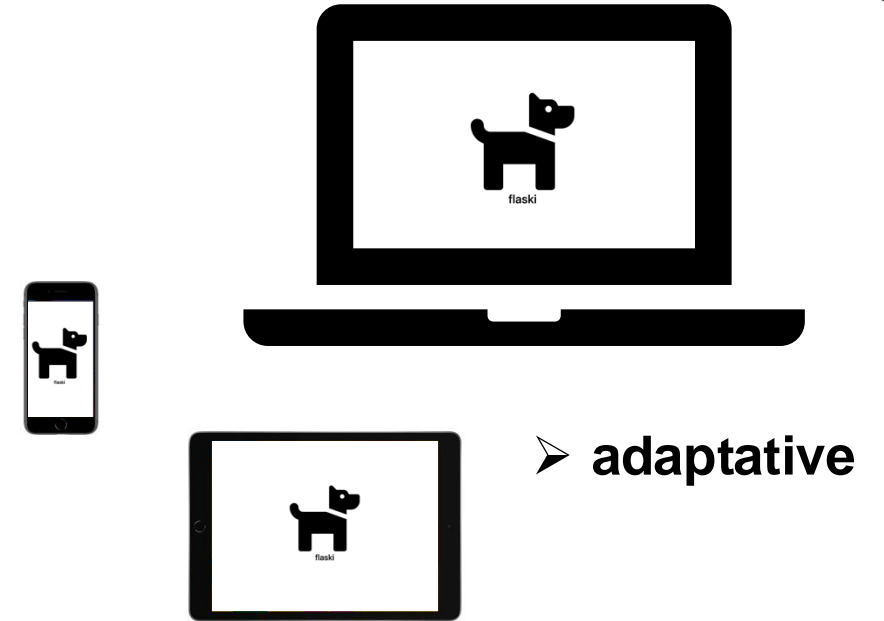
David Otasek¹, John H. Morris², Jorge Bouças³, Alexander R. Pico⁴ and Barry Demchak^{1*}

Abstract
 Cytoscape is one of the most successful network biology analysis and visualization tools, but because of its interactive nature, its role in creating reproducible, scalable, and novel workflows has been limited. We describe Cytoscape Automation (CA), which marries Cytoscape with workflow engines like Nextflow, over 270 Cytoscape core files, and is backed by Swagger documentation. CA has reached an advanced stage of development.
Keywords: Workflow, Reproducible



FLASKI

Scatter plot	3D Scatter plot	Line plot	Histogram
Heatmap	Violin plot	Circular bar plot	Dendrogram
Venn diagram	GSEA plot	DAVID	Cell plot
KEGG	PCA	MDS	tSNE
Lifespan	DataLake		








A collection of web apps for
life sciences

RNAseq	ATACseq	ChIPseq	Alternative Splicing
Intron Retention	IRfinder	Circular RNA	miRNA
16S	Variant Calling	Ribo-Seq	AlphaFold
Methylation Clock	GSEA		



APPS

-  Scatter plot
-  Heatmap
-  Venn diagram
-  KEGG
-  Lifespan






- general purpose Apps
- data specific Apps
- interactive Apps
- app2app
- forms

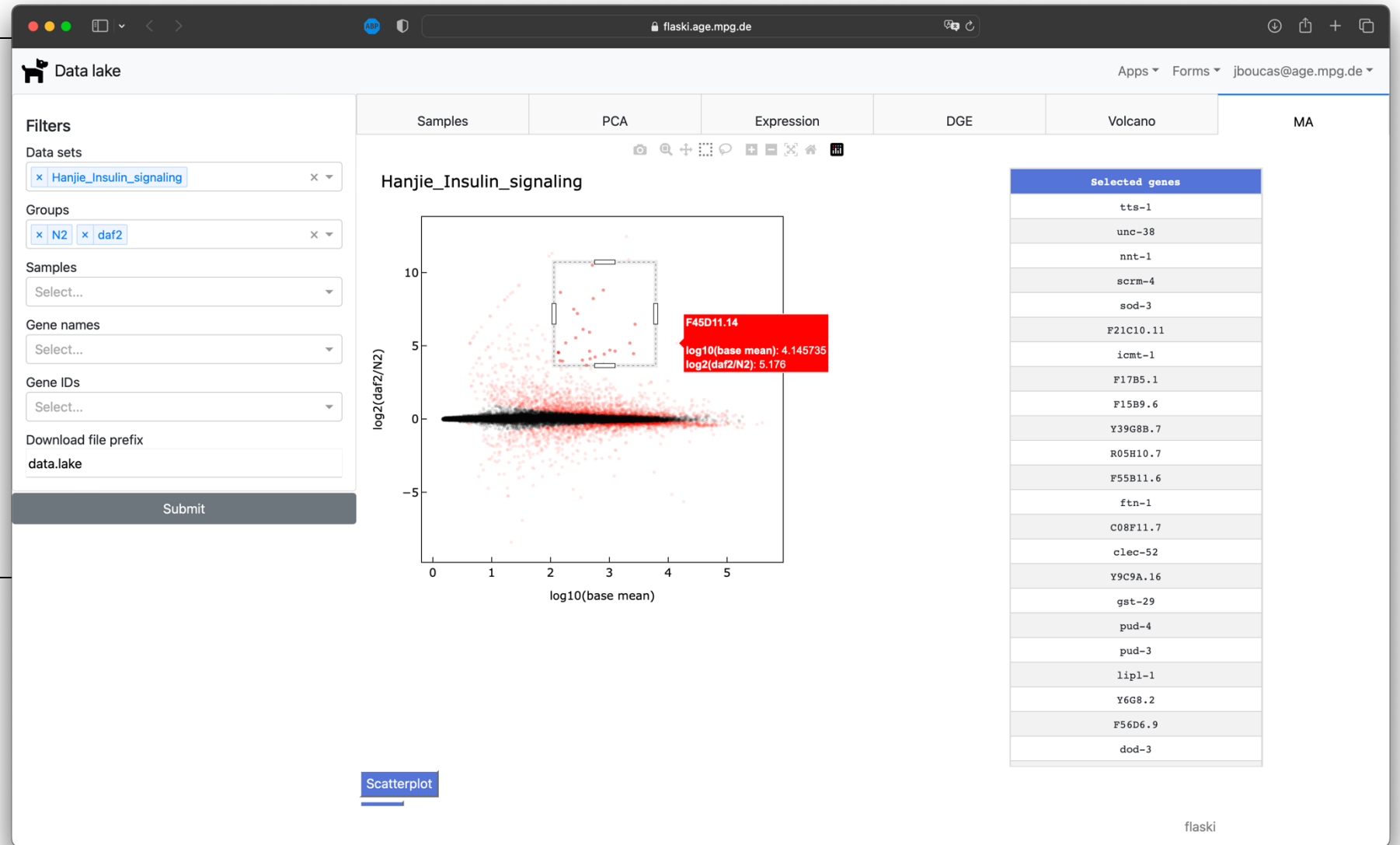
 Methylation Clock

 GSEA



APPS

-  Scatter plot
-  Heatmap
-  Venn diagram
-  KEGG
-  Lifespan



Filters

Data sets: Hanjie_Insulin_signaling

Groups: N2, daf2

Samples: Select...

Gene names: Select...

Gene IDs: Select...

Download file prefix: data.lake

Submit

Hanjie_Insulin_signaling

log₂(daf2/N2) vs log₁₀(base mean)

F45D11.14
log₁₀(base mean): 4.145735
log₂(daf2/N2): 5.176

Selected genes
tts-1
unc-38
nnt-1
scrm-4
sod-3
F21C10.11
icmt-1
F17B5.1
F15B9.6
Y39G8B.7
R05H10.7
F55B11.6
ftn-1
C08F11.7
clec-52
Y9C9A.16
gst-29
pud-4
pud-3
lip1-1
Y6G8.2
F56D6.9
dod-3

Scatterplot






flaski

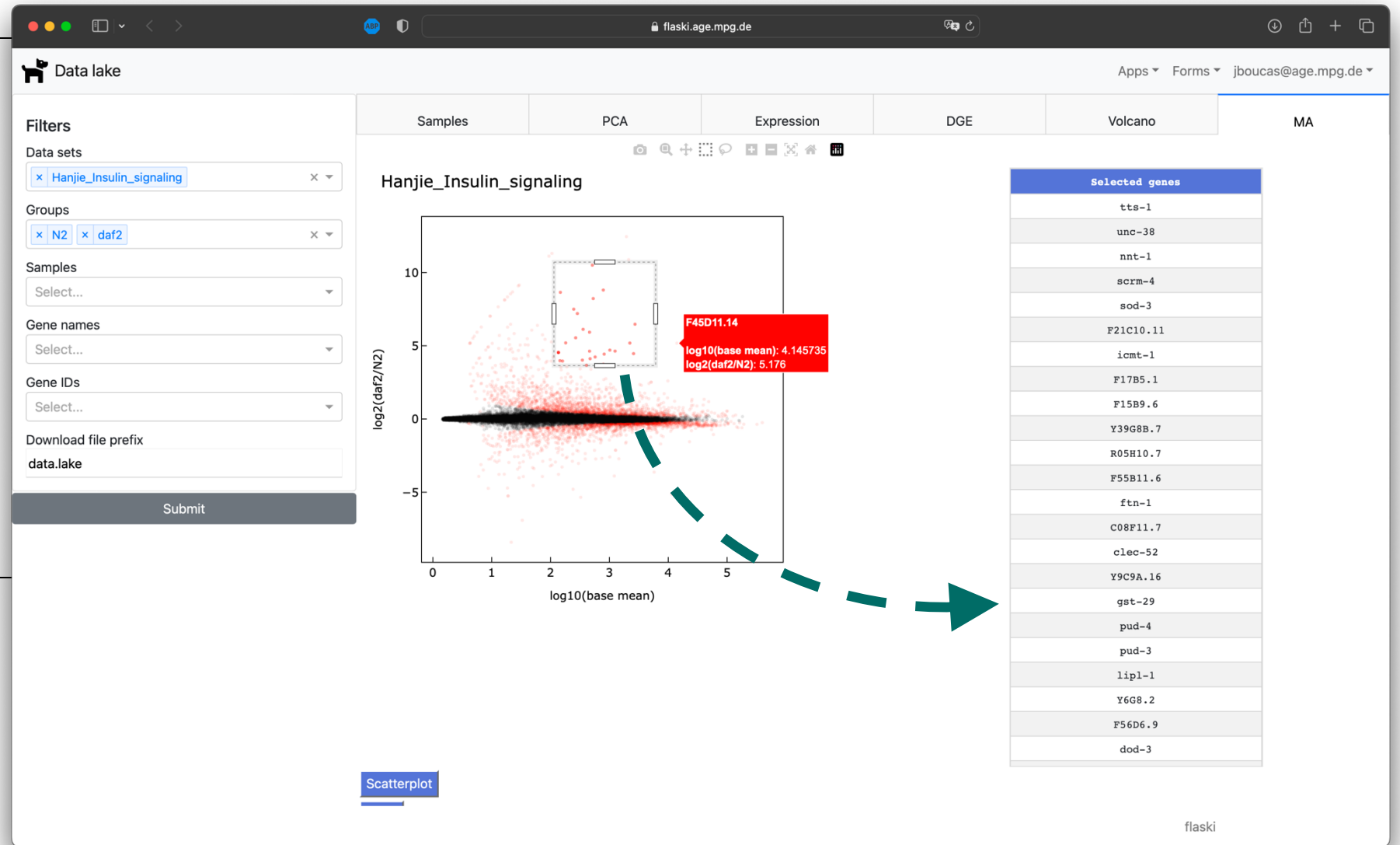
- general purpose Apps
- **data specific Apps**
- interactive Apps
- app2app
- forms

-  Methylation Clock
-  GSEA



APPS

-  Scatter plot
-  Heatmap
-  Venn diagram
-  KEGG
-  Lifespan



Filters

Data sets: Hanjie_Insulin_signaling

Groups: N2, daf2

Samples: Select...

Gene names: Select...

Gene IDs: Select...

Download file prefix: data.lake

Submit

Hanjie_Insulin_signaling

log₂(daf2/N2) vs log₁₀(base mean)

Selected genes

Selected genes
tts-1
unc-38
nnt-1
scrm-4
sod-3
F21C10.11
icmt-1
F17B5.1
F15B9.6
Y39G8B.7
R05H10.7
F55B11.6
ftn-1
C08F11.7
clec-52
Y9C9A.16
gst-29
pud-4
pud-3
lip1-1
Y6G8.2
F56D6.9
dod-3






Scatterplot

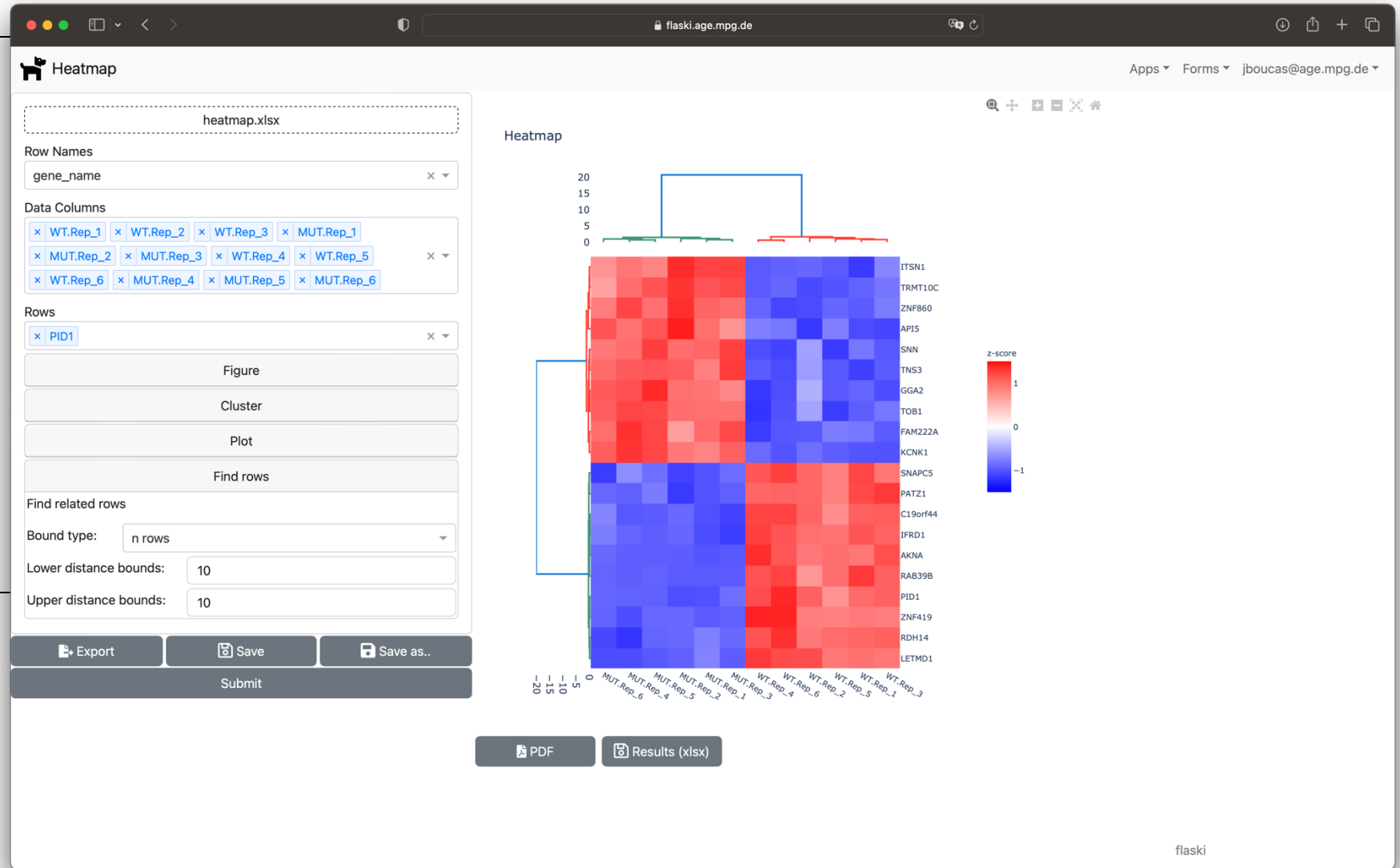
- general purpose Apps
- data specific Apps
- **interactive Apps**
- app2app
- forms

-  Methylation Clock
-  GSEA



APPS

-  Scatter plot
-  Heatmap
-  Venn diagram
-  KEGG
-  Lifespan



- general purpose Apps
- data specific Apps
- **interactive Apps**
- app2app
- forms



Methylation Clock



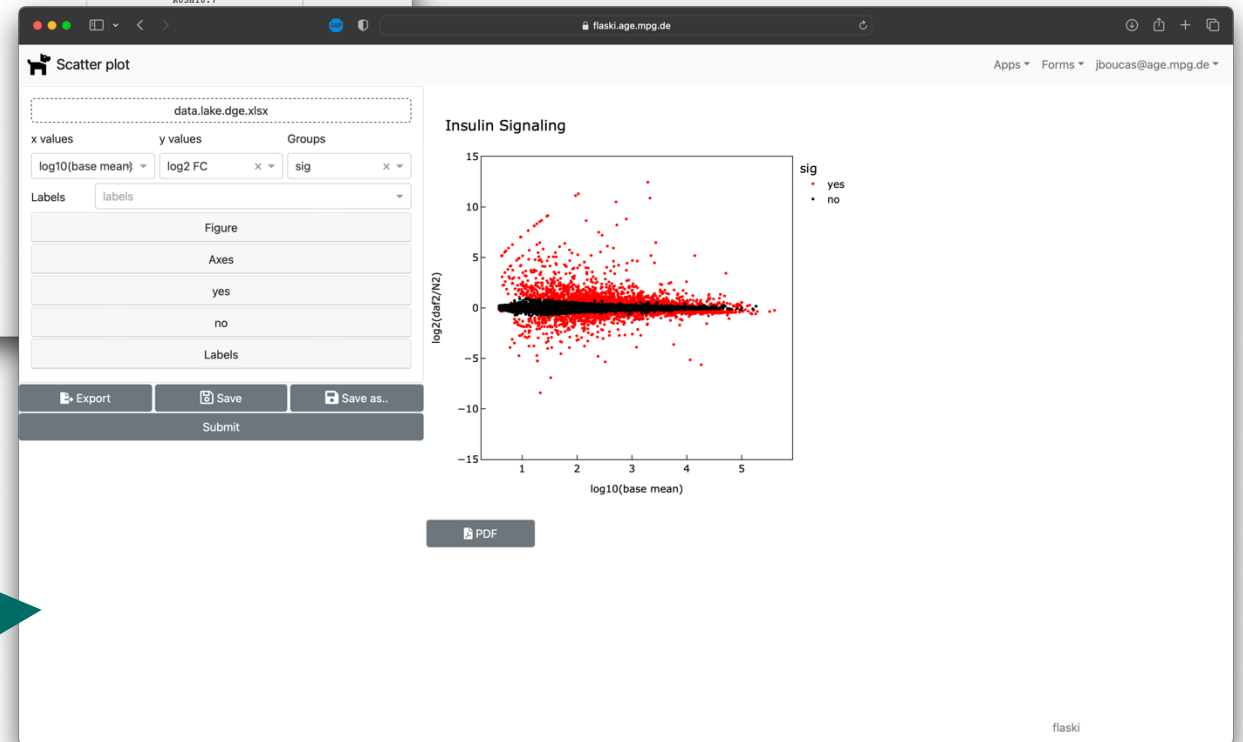
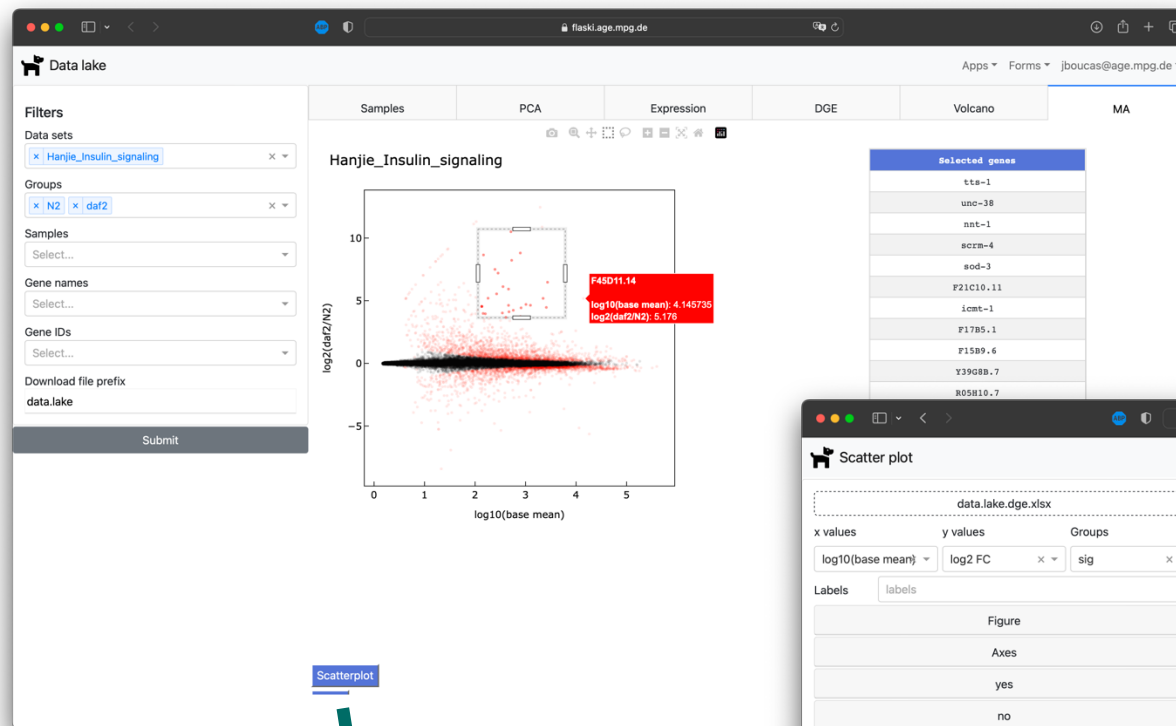
GSEA



APPS

eg.

- Data lake > scatter plot, heatmap
- PCA > scatter plot
- DAVID > cell plot / horizontal bar plots



- general purpose Apps
- data specific Apps
- interactive Apps
- **app2app**
- forms



LIVE SUPPORT

The screenshot shows a web browser window with the URL `flaski.age.mpg.de/heatmap/`. The page title is "Heatmap". The interface includes a file upload field containing "heatmap.xlsx". Below it, there are sections for "Row Names" (set to "ensembl_gene_id") and "Data Columns" (containing a list of columns like "gene_name", "WT.Rep_1", "WT.Rep_2", "WT.Rep_3", "MUT.Rep_1", "MUT.Rep_2", "MUT.Rep_3", "WT.Rep_4", "WT.Rep_5", "WT.Rep_6", "MUT.Rep_4", "MUT.Rep_5", and "MUT.Rep_6"). There are also buttons for "Figure", "Cluster", "Plot", and "Find rows". At the bottom, there are buttons for "Export", "Save", "Save as..", and "Submit".

An "Exception" dialog box is open on the right side of the screen. It contains the following text:

Exception

There was a problem generating your output.
Not all values in column 'gene_name' could be converted to a floating number. Make sure all selected columns contain floating or integer numbers.

Something went wrong, we have been notified. If you would like to share your session with us and get help on this issue please press 'Ice Cream'.

Buttons: expand, Ice Cream

- Error reporting
- Custom error msg.
- Session sharing



SESSIONS

- Storage
- pyflaski versioning

The screenshot shows a web browser window with the URL `flaski.age.mpg.de/storage/`. The page title is "Storage" and the user is logged in as `jboucas@age.mpg.de`. The interface includes a "Home" link, a "Make" button, and a text input field containing "dir name". Below this is a dashed box labeled "upload a session file". The main content is a table listing files and directories:

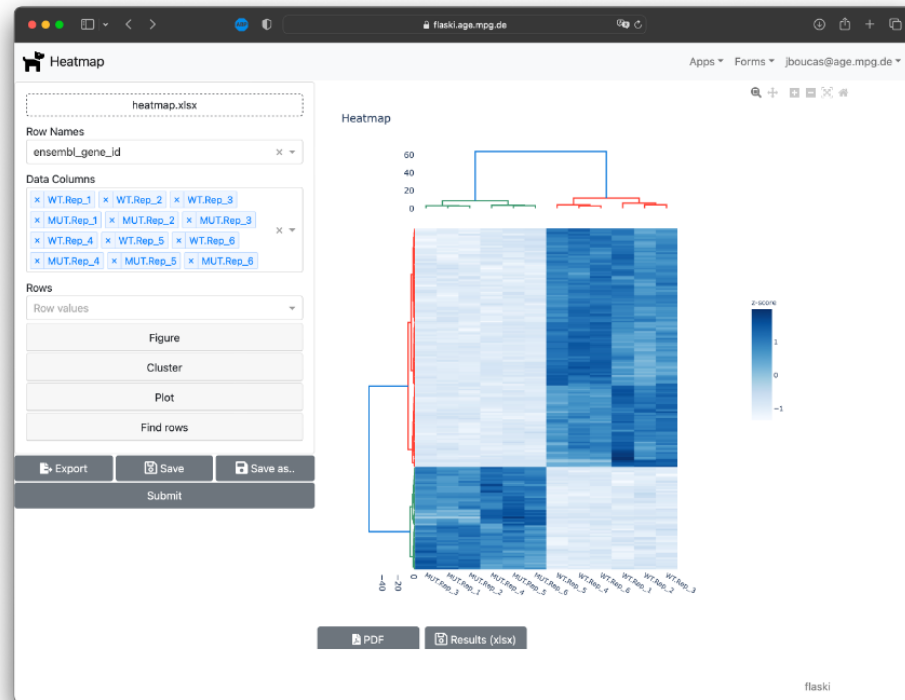
Delete	Name	Size	Modified	Load	Download
	_shared_sessions_	2 Bytes	a day ago		
	test_dir	0 Bytes	3 days ago		
	fsdf.json	10.2 kB	3 months ago		
	some_test_FILE.json	10.2 kB	3 months ago		

The flaski logo is visible in the bottom right corner of the page.

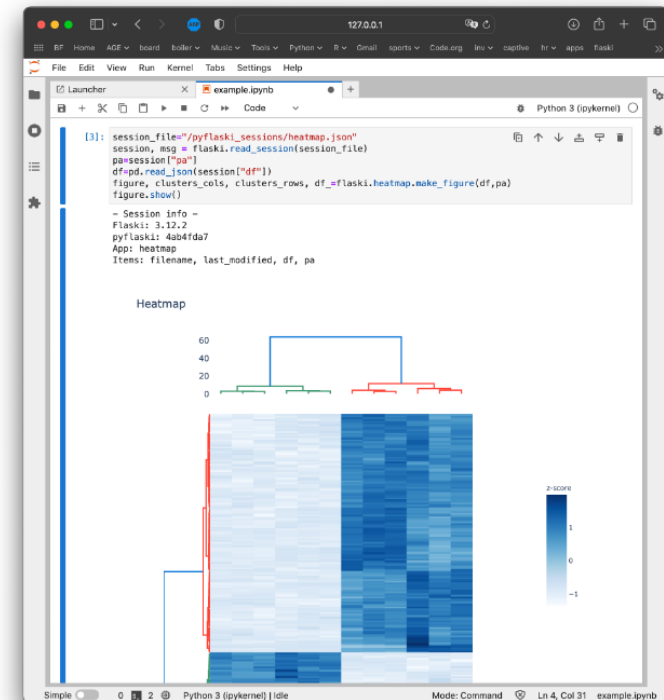


PYFLASKI

web App
(graphical user interface)



Jupyter Notebook
(programmatic interface)



session
transfer



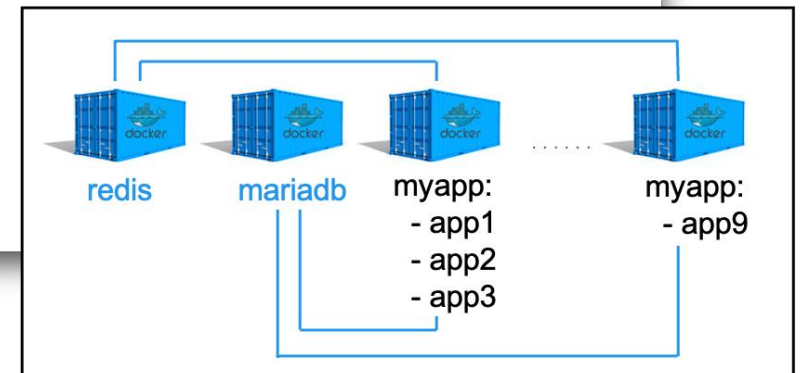
MYAPP

- docker-compose.yml
- kubernetes.yml
- mariadb & redis
- granular user access
- 2FA
- admin auth. tokens for new users
- user & admin dashboard
- Let's Encrypt
- amd64, arm64, arch64
- apps as routes
- apps as containers / plugins

- **Base framework: flask + dash**
- **Agile App development**
- **Daily deployed with Flaski**

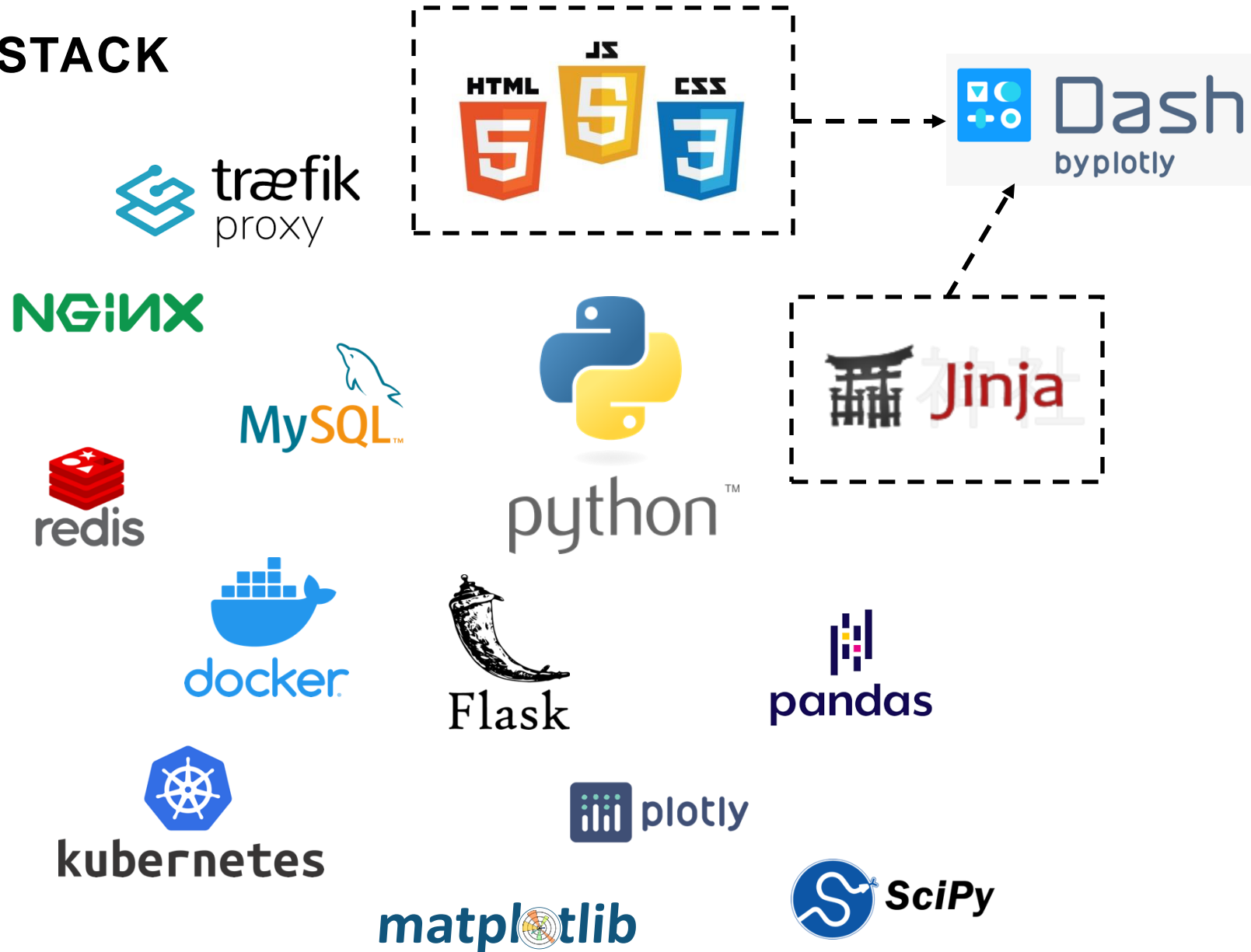
myapp

Home is where the Dom is.





WEB DEV STACK

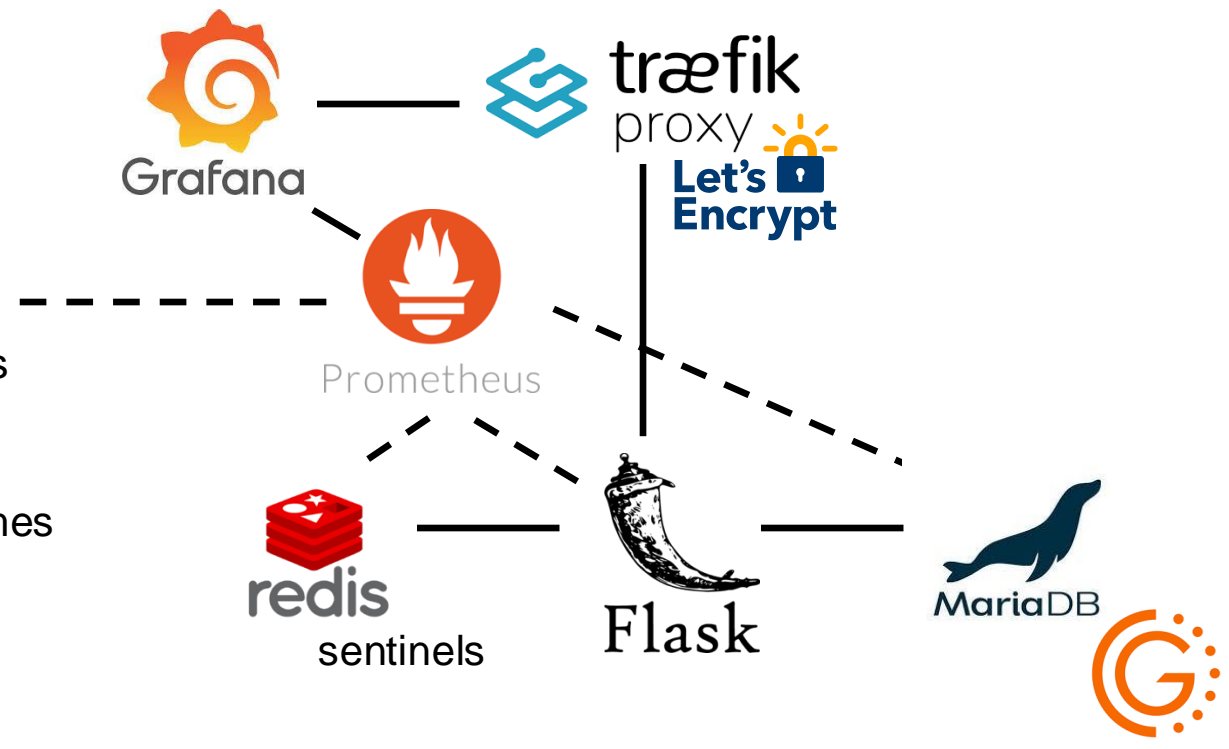
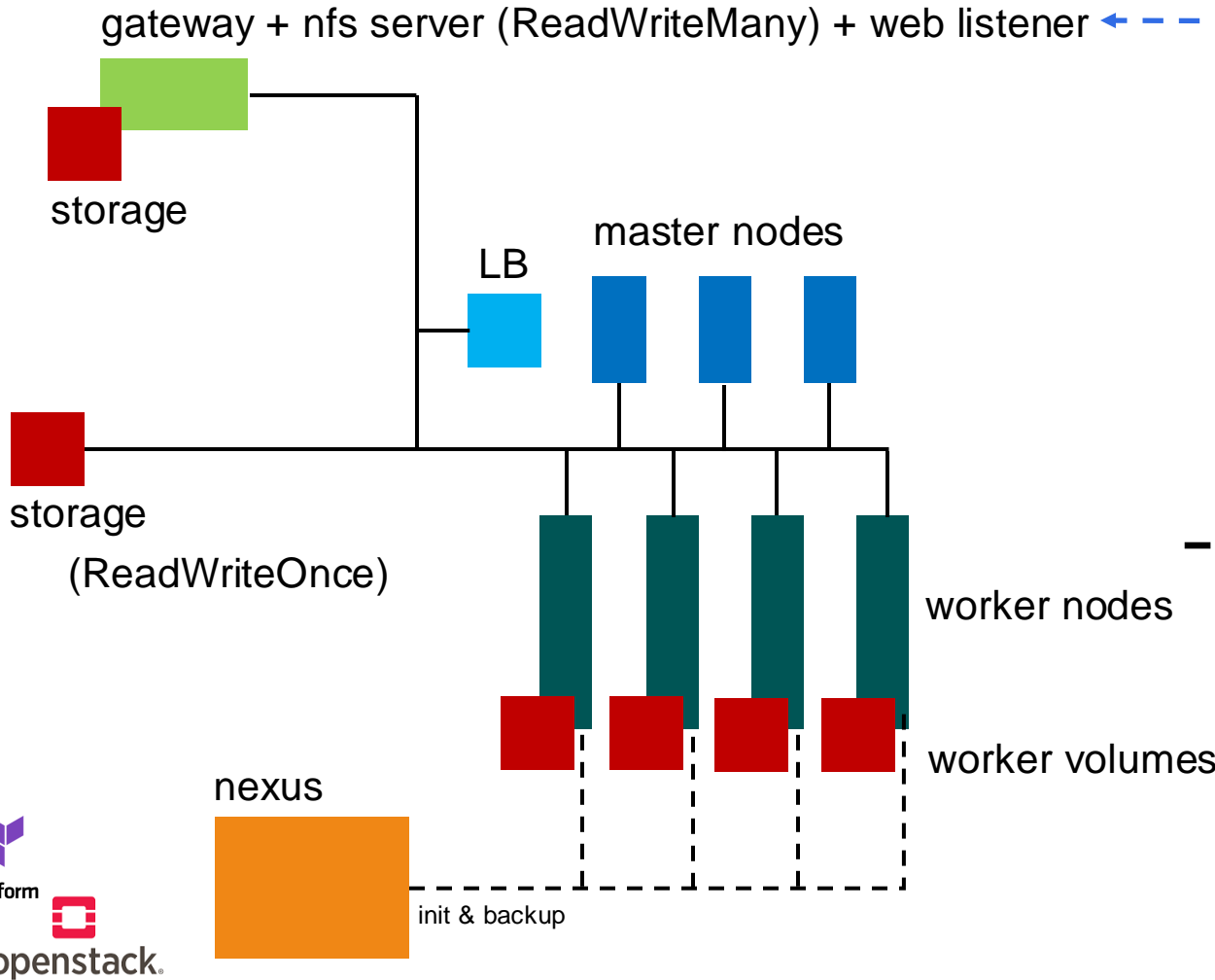
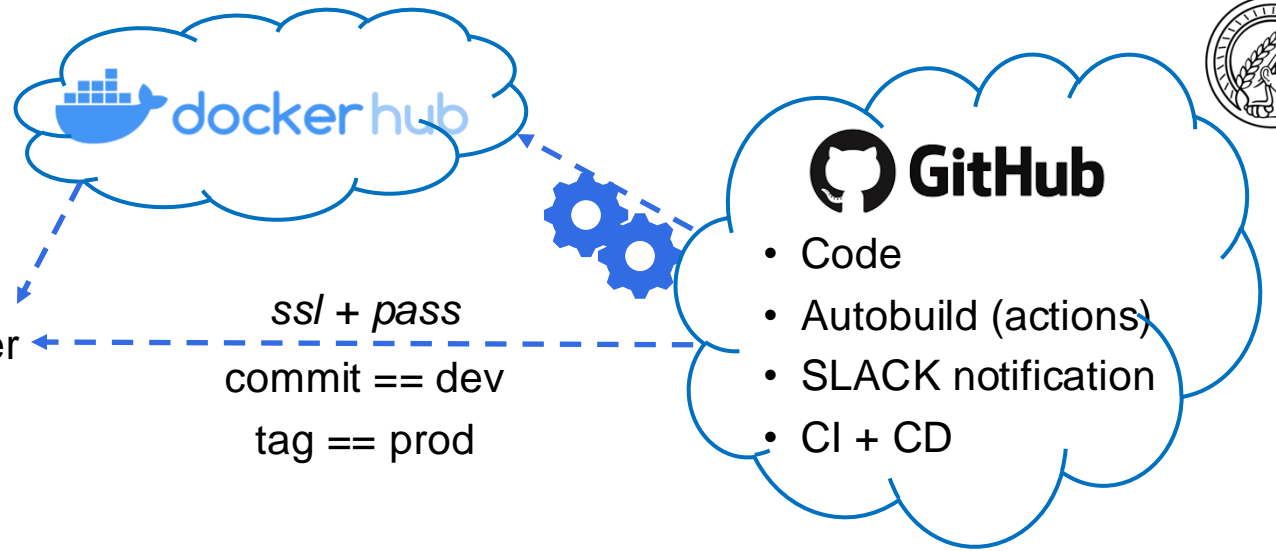




SUMMARY

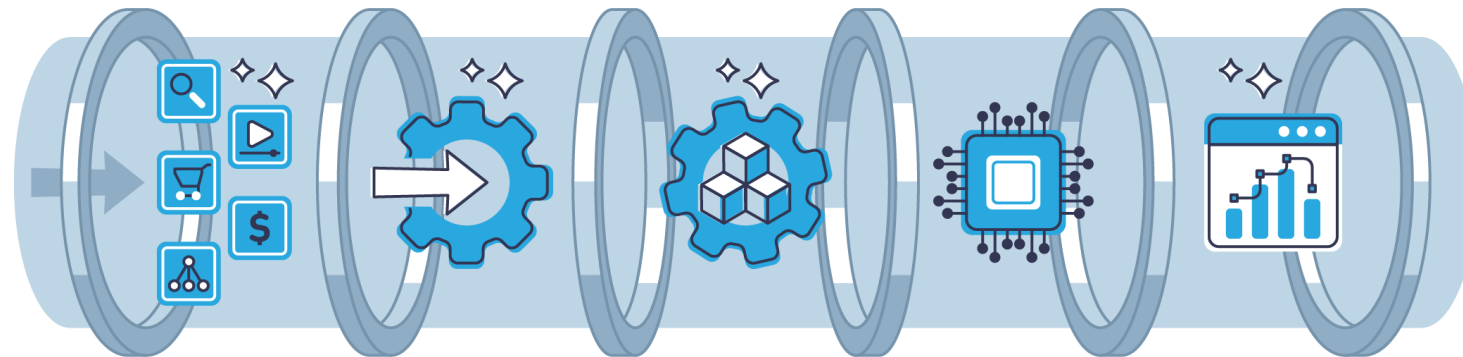
- ✓ full power of data science libraries to a graphical user interface
- ✓ web / server based: continuous deployment
- ✓ session management and tool versioning – reproducible science
- ✓ graphical (GUI) to programmatic interface
- ✓ granular user access: who can see and use which app and function
- ✓ open source, open formats, democratic: no code nor data sequestering

KUBERNETES





AUTOMATION





AUTOMATION

flaski

flaski.age.mpg.de/rnaseq/

Apps Forms jboucas@age.mpg.de

upload a file or fill up the form

Readme	Samples (example)	Samples	Info		
sample	group	replicate	read 1	read 2	notes
A	control	1	A006850092_131904_S2_L002_R1_001.fq.gz	A003450092_131904_S2_L002_R2_001.fq.gz	eg. 2 files / sample
B	control	2	A006850092_131924_S12_L002_R1_001.fq.gz	A006850092_131924_S12_L002_R2_001.fq.gz	
C	control	3	A006850092_131944_S22_L002_R1_001.fq.gz	A006850092_131944_S22_L002_R2_001.fq.gz	
D	shRNA	1	A006850092_131906_S3_L002_R1_001.fq.gz	A006850092_131906_S3_L002_R2_001.fq.gz	
E	shRNA	2	A006850094_131926_S3_L001_R1_001.fq.gz	A006850094_131926_S3_L001_R2_001.fq.gz	
F	shRNA	3	A006850092_131956_S28_L002_R1_001.fq.gz	A006850092_131956_S28_L002_R2_001.fq.gz	

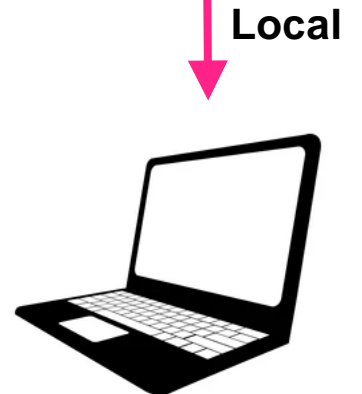
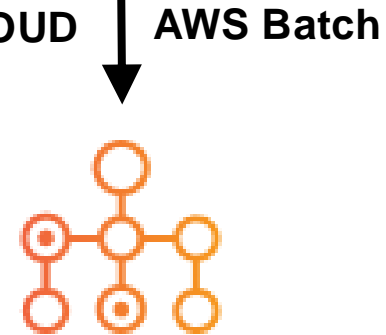
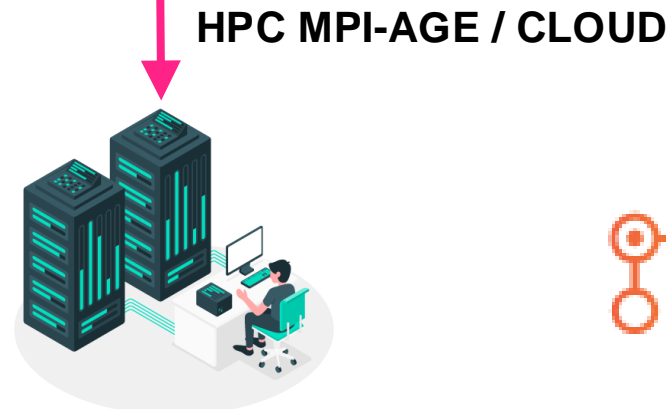
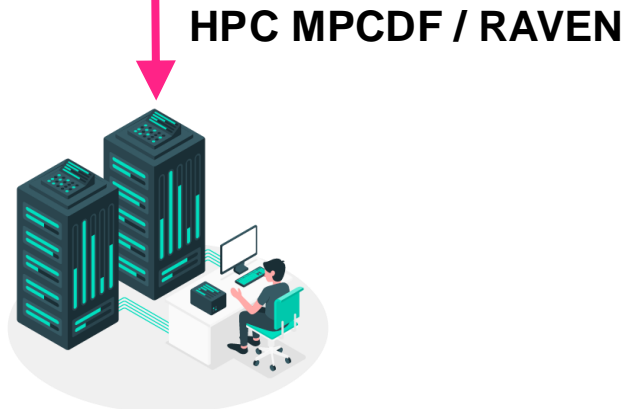
Submit



```

params.local.json — nextflow-rnaseq
{} params.r2d2.json {} params.local.json x
{} params.local.json > ...
1
2 "project_folder": "~/nextflow-rnaseq-run/",
3 "samplestable": "~/nextflow-rnaseq-run/sample_sheet.xlsx",
4 "fastqc_raw_data": "~/nextflow-rnaseq-run/raw_data",
5 "kallisto_raw_data": "~/nextflow-rnaseq-run/raw_data",
6 "featurecounts_raw_data": "~/nextflow-rnaseq-run/raw_data",
7 "genomes": "~/nextflow-rnaseq-run/",
8 "organism": "caenorhabditis_elegans",
9 "release": "107",
10 "url_gtf": "ftp://ftp.ensembl.org/pub/release-107/gtf/caenorhabditis_elegans/",
11 "url_dna": "ftp://ftp.ensembl.org/pub/release-107/fasta/caenorhabditis_elegans/dna",
12 "ercc_label": "ercc92",
13 "url_ercc_gtf": "https://datashare.mpcdf.mpg.de/s/M0xbNrXeBNcg9wt/download",
14 "url_ercc_fa": "https://datashare.mpcdf.mpg.de/s/H9PQu3vDRi9saqV/download",
15 "biomart_host": "http://dec2021.archive.ensembl.org/biomart/",
16 "biomart_dataset": "celegans_gene_ensembl",
17 "circRNA": "None",
18 "daviddatabase": "ENSEMBL_GENE_ID",
19 "cytoscape_ip_mount": "",
20 "species": "caenorhabditis_elegans",
21 "spec": "celegans",
22 "homefolder": "~/",
23

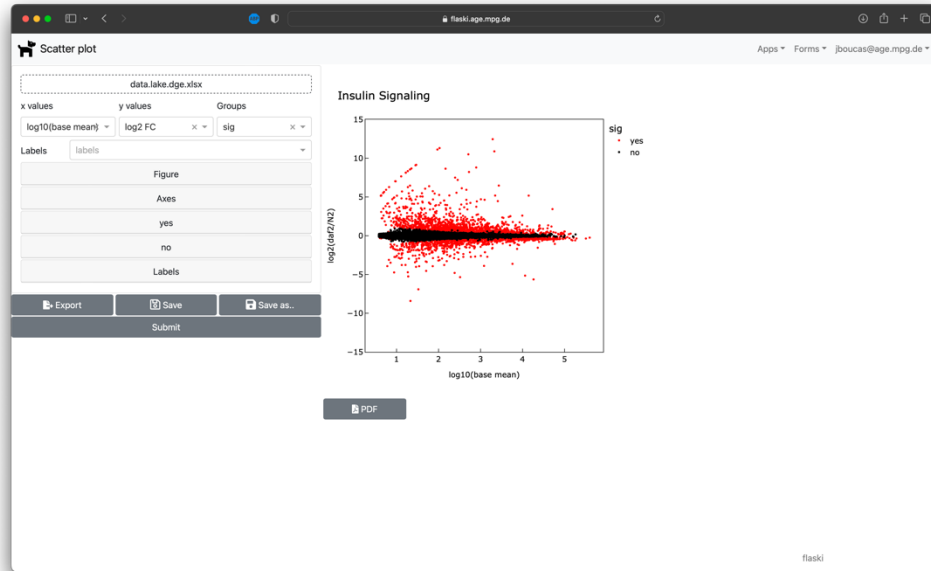
```



nextflow



AUTOMATION



session.json



owncloud API

HPC MPCDF / RAVEN



HPC MPI-AGE / CLOUD



Link to results





nextflow.io

nextflow

Documentation Examples Training Resources Forums

Fork me on GitHub

```
process sayHello {
  input:
  val cheers
  output:
  stdout

  """
  echo $cheers
  """
}

workflow {
  channel.of('Ciao','Hello','Hola') | sayHello | view
}
```

Nextflow

Data-driven computational pipelines

Nextflow enables scalable and reproducible scientific workflows using software containers. It allows the adaptation of pipelines written in the most common scripting languages.

Its fluent DSL simplifies the implementation and the deployment of complex parallel and reactive workflows on clouds and clusters.

[Documentation](#) [Community forums](#)

nextflow SUMMIT

Barcelona 2024

OCT 28 - NOV 1, 2024

Join us for the latest developments and innovations from the Nextflow world.

Join us for the latest developments and innovations from the Nextflow world!

With training, a hackathon and talks from pioneers in the field, the Nextflow Summits are essential events for anyone using Nextflow.

[Register now](#)

Zero config

Polyglot

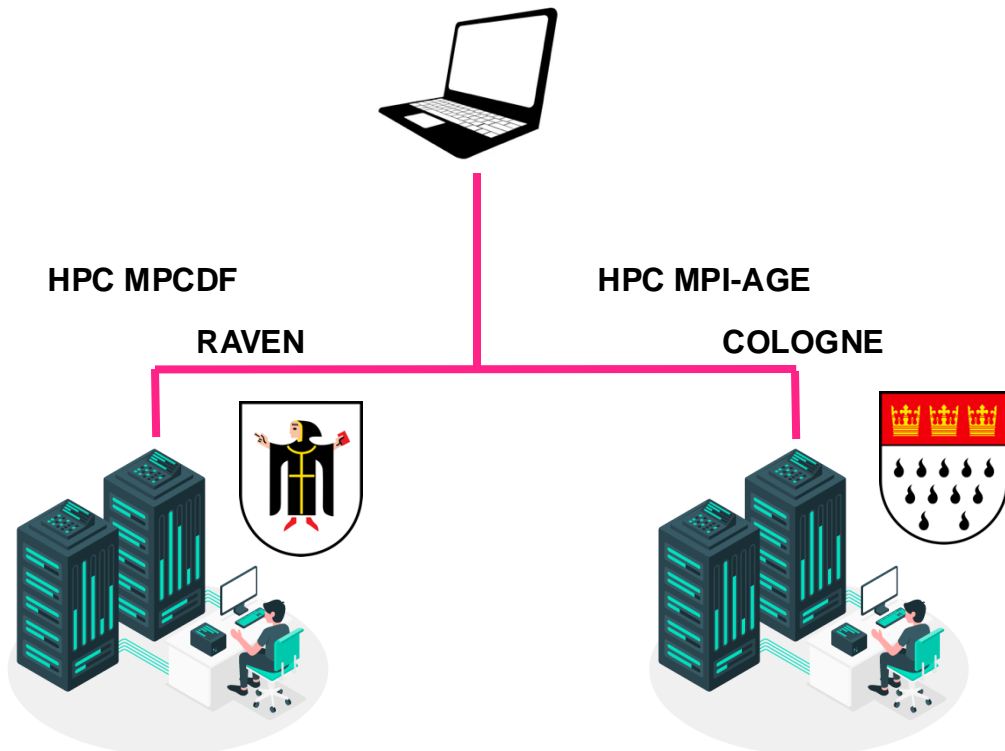
Concurrency

Scale easily

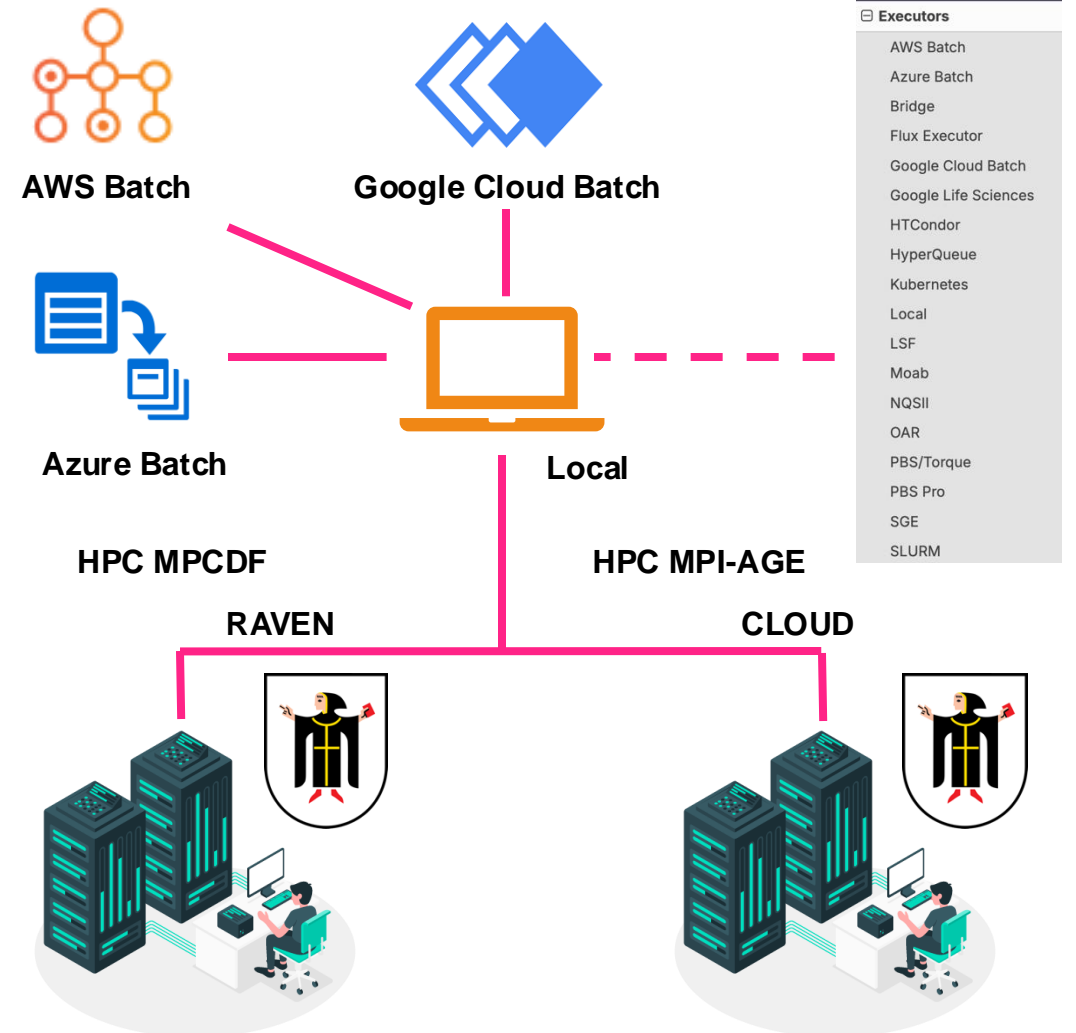


NEXTFLOW ?

BEFORE



AFTER





Executor profile: slurm / local / ..

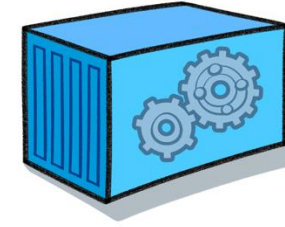


Modular

```

1 nextflow.config
2 profiles {
3   standard {
4     params_run_type='local'
5     params_containers='docker'
6     docker-enabled = true
7     includeConfig 'configs/local.config'
8   }
9   P2D {
10    params_run_type='P2D'
11    params_image_folder='nexus/common/singularity'
12    params_extra_mounts=''
13    params_queue='hpc1'
14    params_containers='singularity'
15    singularity-enabled = true
16    includeConfig 'configs/slurm.config'
17  }
18  reven {
19    beforeScript = 'module load singularity'
20    clusterOptions = '-mfastq-get-corex'
21  }
22  params_run_type='reven'
23  params_image_folder='nexus/poslib/MAGE-flaski/service/iaages/'
24  params_extra_mounts='-B /reven/reven -B /nexus/nexus'
25  params_queue='hpc1'
26  params_containers='singularity'
27  singularity-enabled = true
28  includeConfig 'configs/slurm.config'
29  }
30  studio {
31    params_run_type='studio'
32    params_image_folder='nexus/poslib/MAGE-flaski/service/iaages/'
33    params_extra_mounts='-B /nexus/nexus'
34    params_queue='Cluster'
35    params_containers='singularity'
36  }
37 }

```



Docker / Singularity / ..

nextflow

Universal code: bash / Python / R / ..

```

39 process fastqc {
40   tag "${if}"
41   stepMode 'serial'
42   stepOutMode 'none'
43 }
44 input
45 path f
46 val fastqc_output
47 scripts
48 ----
49 mkdir -p /workdir/${fastqc_output}
50 fastqc -z $task.cpus -o /workdir/${fastqc_output} /raw_data/${if}
51 ----
52 }
53 workflow {
54   [f, fastqc_output in params.keySet()] {
55     fastqc_output="${params.fastqc_output}"
56     > fastqc {
57       fastqc_output="${fastqc_output}"
58     }
59   }
60   data = channel.fromPath("${params.fastqc_raw_data}/${fastqc}")
61   data = data.filter { file "${it.replace("${fastqc}", "${fastqc.html}").replace("${params.fastqc_raw_data}", "${params.project_folder}/${fastqc_output}/${it}").exists() }
62   fastqc_data, fastqc_output
63 }

```

Run specific parameters: variables

```

1 nextflow.config
2 profiles {
3   standard {
4     params_run_type='local'
5     params_containers='docker'
6     docker-enabled = true
7     includeConfig 'configs/local.config'
8   }
9   P2D {
10    params_run_type='P2D'
11    params_image_folder='nexus/common/singularity'
12    params_extra_mounts=''
13    params_queue='hpc1'
14    params_containers='singularity'
15    singularity-enabled = true
16    includeConfig 'configs/slurm.config'
17  }
18  reven {
19    beforeScript = 'module load singularity'
20    clusterOptions = '-mfastq-get-corex'
21  }
22  params_run_type='reven'
23  params_image_folder='nexus/poslib/MAGE-flaski/service/iaages/'
24  params_extra_mounts='-B /reven/reven -B /nexus/nexus'
25  params_queue='hpc1'
26  params_containers='singularity'
27  singularity-enabled = true
28  includeConfig 'configs/slurm.config'
29  }
30  studio {
31    params_run_type='studio'
32    params_image_folder='nexus/poslib/MAGE-flaski/service/iaages/'
33    params_extra_mounts='-B /nexus/nexus'
34    params_queue='Cluster'
35    params_containers='singularity'
36  }
37 }

```

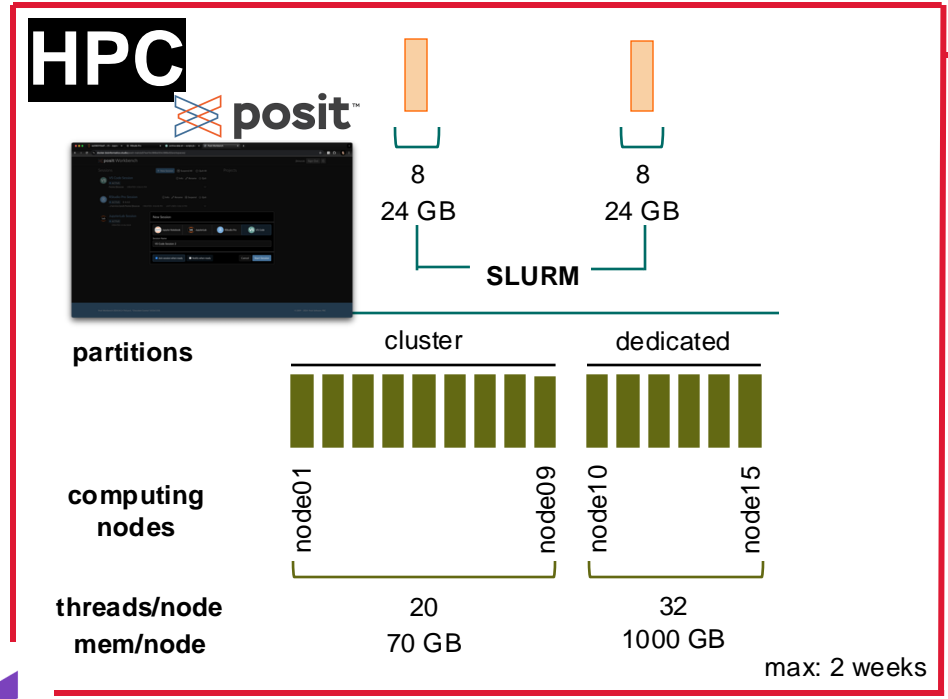
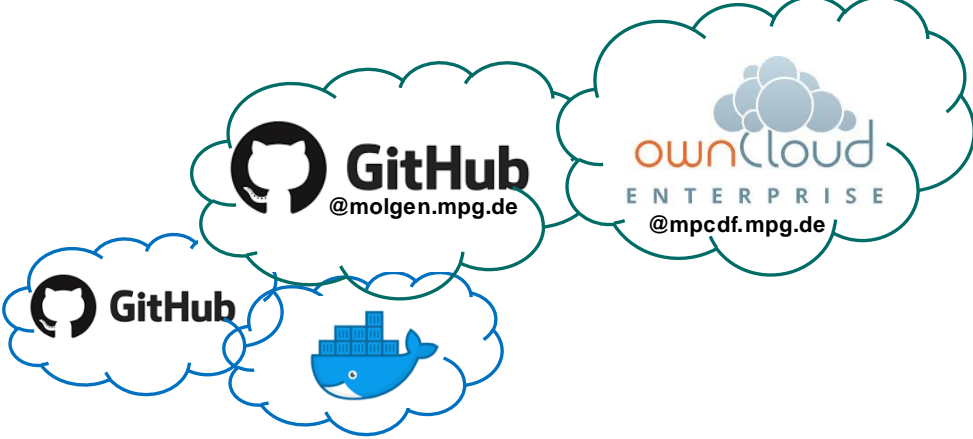
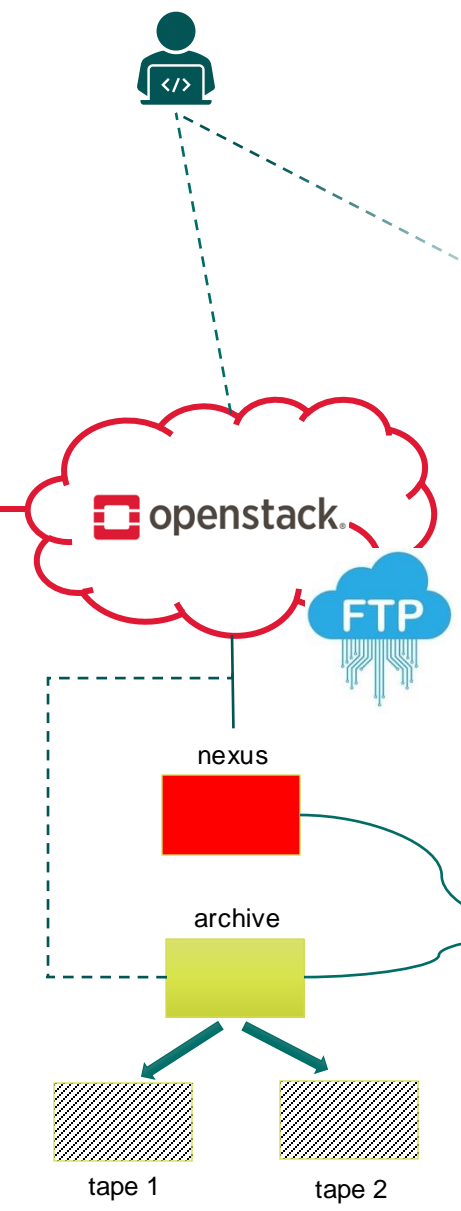
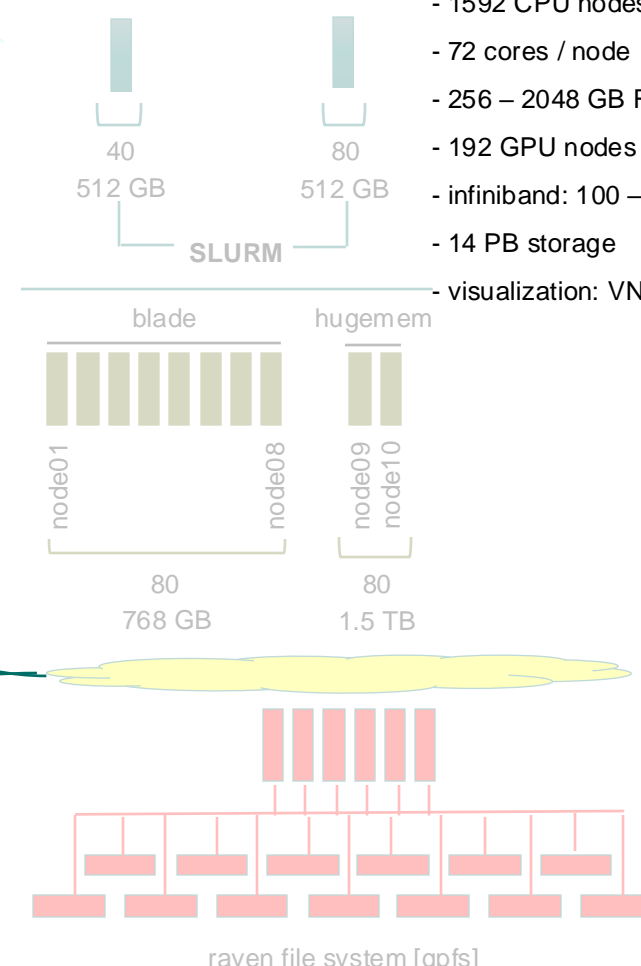


HPC INFRASTRUCTURE

HPC

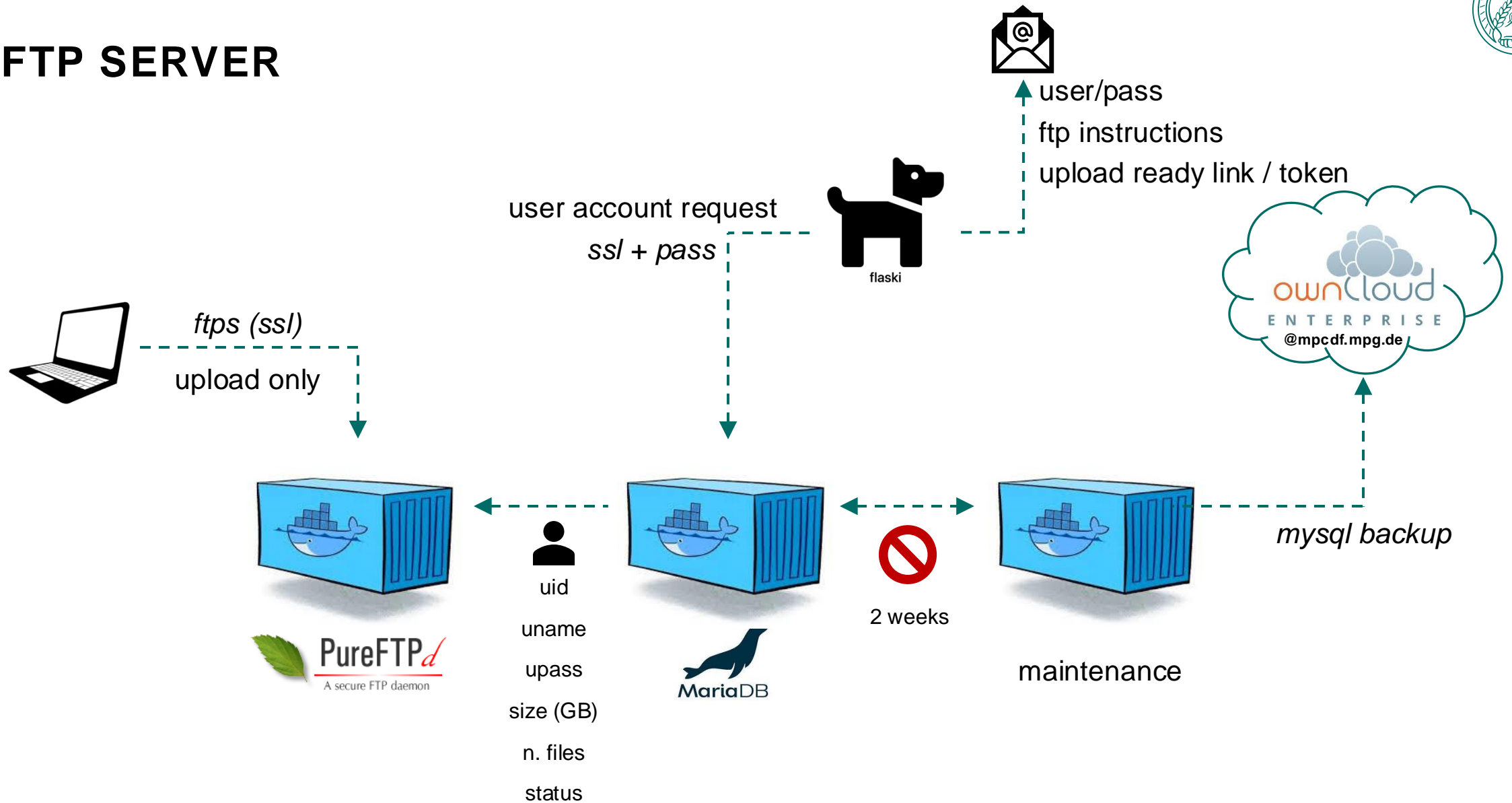
raven

- 1592 CPU nodes
- 72 cores / node
- 256 – 2048 GB RAM / node
- 192 GPU nodes
- infiniband: 100 – 200 Gbit/s
- 14 PB storage
- visualization: VNC + jupyter





FTP SERVER





NEXUS: FUNCTIONAL USER AS ROOT

/nexus/posix0/MAGE-flaski

```
/tmp          d rwx rwx r-x  flaski mage
/hpc          d rwx r-x ---  flaski mage

/group       d rwx r-x ---  flaski mage //  A:fdi:flaski@mpcdf.mpg.de:rwaDxtncy
Adam_Antebi d rwx --- ---  flaski mage //  A::<uname>@mpcdf.mpg.de:rwaDxtTnNcCy
Bioinformatics d rwx --- ---  flaski mage //  A::<uname>@mpcdf.mpg.de:rwaDxtTnNcCy
.
/home        d rwx r-x ---  flaski mage //  A:fdi:flaski@mpcdf.mpg.de:rwaDxtncy
jboucas     d rwx --- ---  jboucas  mage
.
```



— nfs4_setfac! != setfac!





NEXUS: FUNCTIONAL USER AS ROOT

functional user cron jobs:

1. check and correct own permissions to all files and folders (using `su` on VM; ! conda !)
2. check sizes of users and group folders (email report)
3. inform users that exceed size limits (email)
4. block users / tar home folders that have not corrected folder size for > xx days
5. tar home folders for users that have not logged in for longer than xx days
6. archive home folders that were tared for longer than xx days



NEXUS: SPACE MANAGEMENT

Filesystem	Inodes	IUsed	IFree	IUse%
10.186.16.28:/nexus/posix0/MAGE-flaski	6.0M	5.1M	939K	85%

Filesystem	Size	Used	Avail	Use%
10.186.16.28:/nexus/posix0/MAGE-flaski	100T	37T	64T	37



sources of small files:

1. libraries / packages
2. code files / repos
3. some projects here and there



NEXUS: GROUP SELF CLEANING

```
/nexus/posix0/MAGE-flaski/group/Bioinformatics
```

```
  /raw_data
```

```
    /project_x
```

```
  /code
```

```
    /project_x
```

```
  /data
```

```
    /project_x
```

cron jobs:

1. Archive projects in raw_data
2. Remove projects in raw_data older than 14 days
3. Tar folders in projects for code and data not accessed for more than 14 days
4. Archive and remove projects in data and code not accessed for more than 14 days



SOFTWARE

One image, multiple software (modules system)

```

jbcucas — ssh -XY jbcucas@raven02i.mpcdf.mpg.de — Solarized Dark ansi — 109x59
~ -- -zsh ... ~ -- -zsh ... ~ -- -zsh ... ~ -- -zsh ... ~ -- -zsh ... ~ -- -zsh ... +
raven02: ~.0/MAGE-flaski/service/images$ singularity exec -B /nexus:/nexus bioinformatics_software.v4.0.4.sif
/bin/bash
INFO: fuse2fs not found, will not be able to mount EXT3 filesystems
container: ~.0/MAGE-flaski/service/images$ module avail
----- /modules/modulefiles/general -----
jdk/18.0.2(default) perl/5.32.1(default) rlang/4.2.1(default)
jupyterhub/2.3.1(default) python/3.9.13(default)

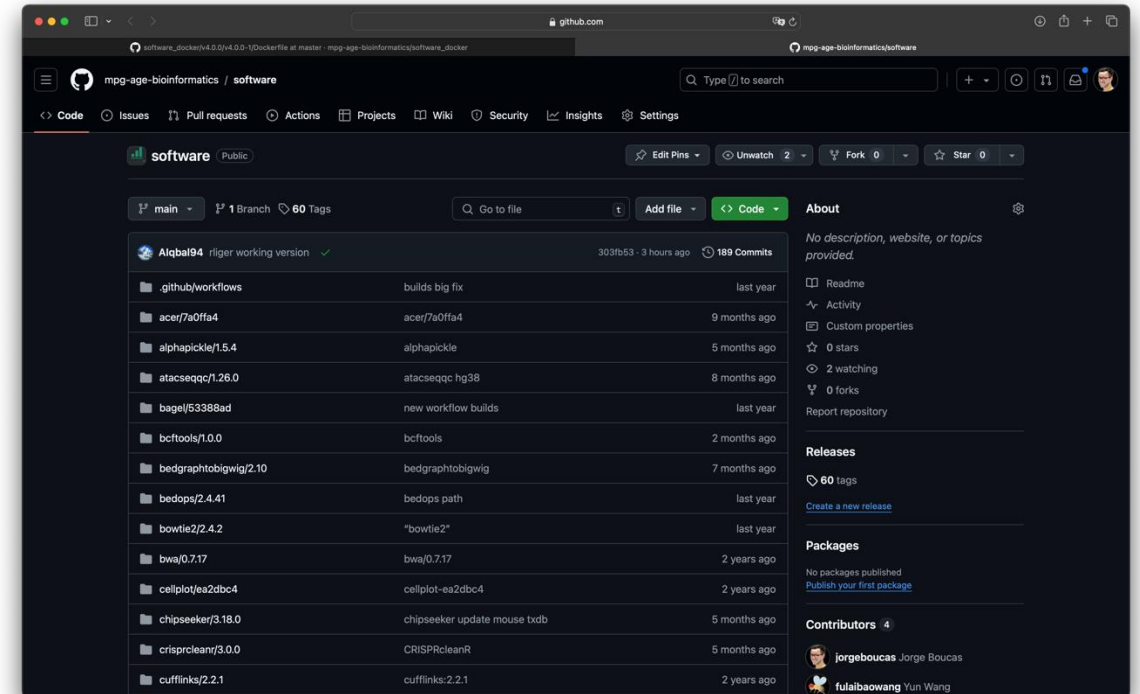
----- /modules/modulefiles/libs -----
bzip2/1.0.8(default) htlib/1.16(default) openblas/0.3.21(default)
gsl/2.7.1(default) imagemagick/7.1.0-47(default) xz/5.2.5(default)

----- /modules/modulefiles/bioinformatics -----
abismal/3.0.0(default) expat/2.4.8(default) mitools/1.5.0(default) spades/3.15.4(default)
bamutil/1.0.15(default) fastqc/0.11.9(default) near/1.0.0(default) sratoolkit/2.11.3(default)
bedtools/2.30.0(default) flexbar/3.5.0(default) ngsutils/0.5.9(default) star/2.7.10a(default)
bismark/0.24.0(default) gatk/4.2.6.1(default) nlopt/2.7.1(default) star/2.7.10b
blast/2.13.0(default) gsea/4.3.2(default) picard/2.27.4(default) stringtie/2.2.1(default)
bowtie/1.3.1 hisat/2.1.0(default) primer3/2.6.1(default) subread/2.0.3(default)
bowtie2/2.4.5(default) homer/4.11.0(default) quast/5.2.0(default) tophat/2.1.1(default)
bwa/0.7.17(default) igrec/3.1.1(default) rsem/1.3.3(default) trimalore/0.6.7(default)
bwtool/face01(default) iseerna/1.2.2(default) samtools/1.15.1(default) trimmomatic/0.39(default)
cufflinks/2.2.1(default) kallisto/0.48.0(default) segemehl/0.3.4(default) vcftools/0.1.16(default)
cytoscape/3.9.1(default) kenttools/435(default) seqtk/1.3.0(default) vjtools/1.2.1(default)
emboss/6.6.0(default) lofreq/2.1.5(default) skewer/0.2.2(default) walt/1.1.0(default)
epiteome/1.0.0(default) meme/5.4.0(default) snpeff/4.3.t(default)

```

https://github.com/mpg-age-bioinformatics/software_docker

One image per software



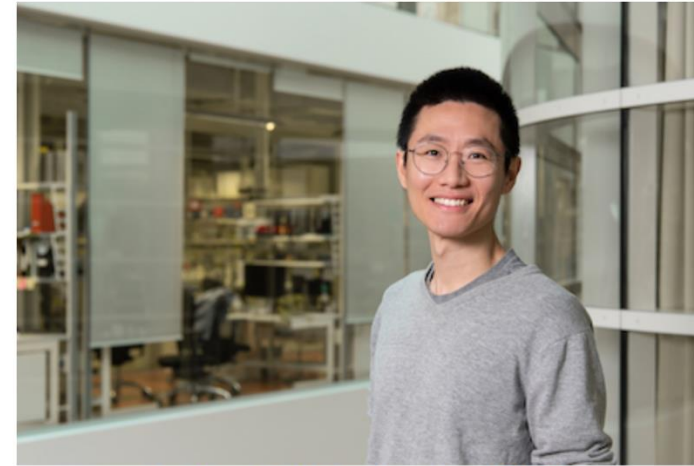
<https://github.com/mpg-age-bioinformatics/software>

<https://hub.docker.com/u/mpgagebioinformatics>

HEROES



Jorge Bouças, Ph.D. | Head of Bioinformatics | Tel.: 312



Yun Wang, Ph.D. | Bioinformatician | Tel.: 313



Ayesha Iqbal | Data Scientist | Tel.: 316



Hossain Md Al Amin | cloud DevOPs engineer | Tel.: 257

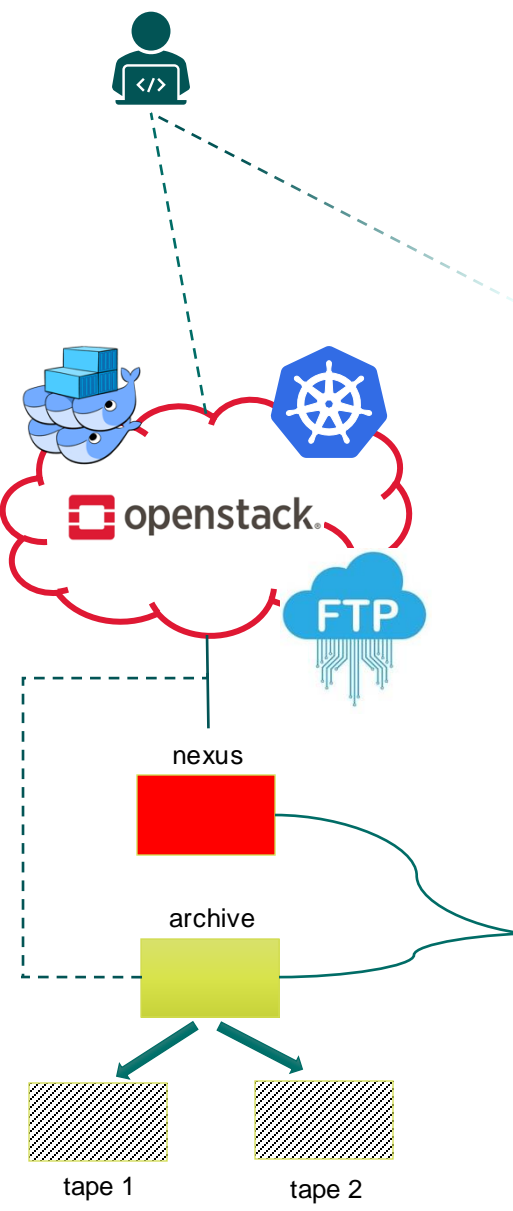
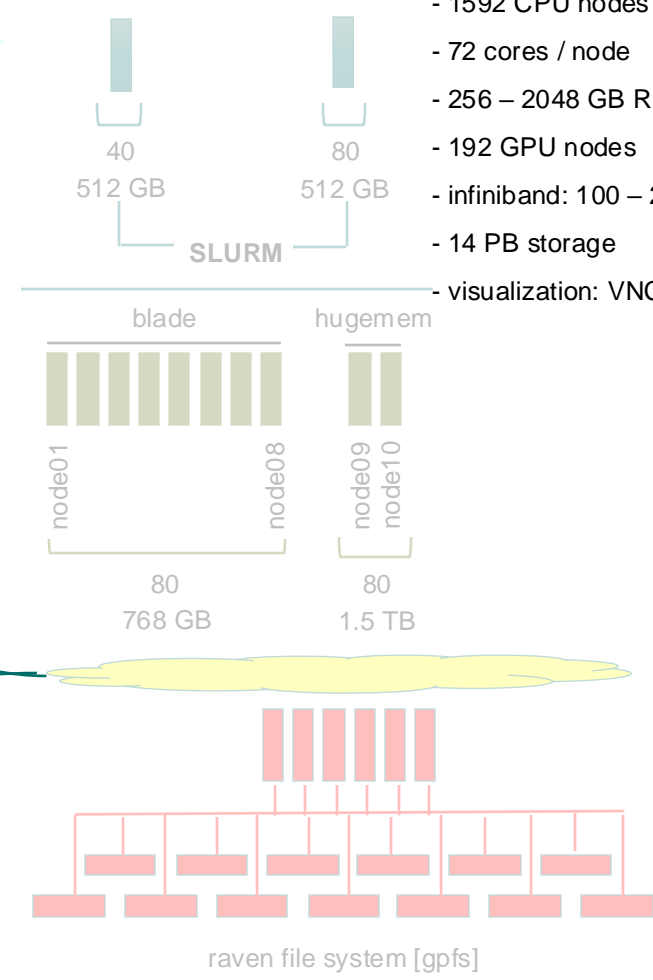




HPC

raven

- 1592 CPU nodes
- 72 cores / node
- 256 – 2048 GB RAM / node
- 192 GPU nodes
- infiniband: 100 – 200 Gbit/s
- 14 PB storage
- visualization: VNC + jupyter



Prometheus

openstack

FTP

nexus

archive

tape 1

tape 2

HPC

