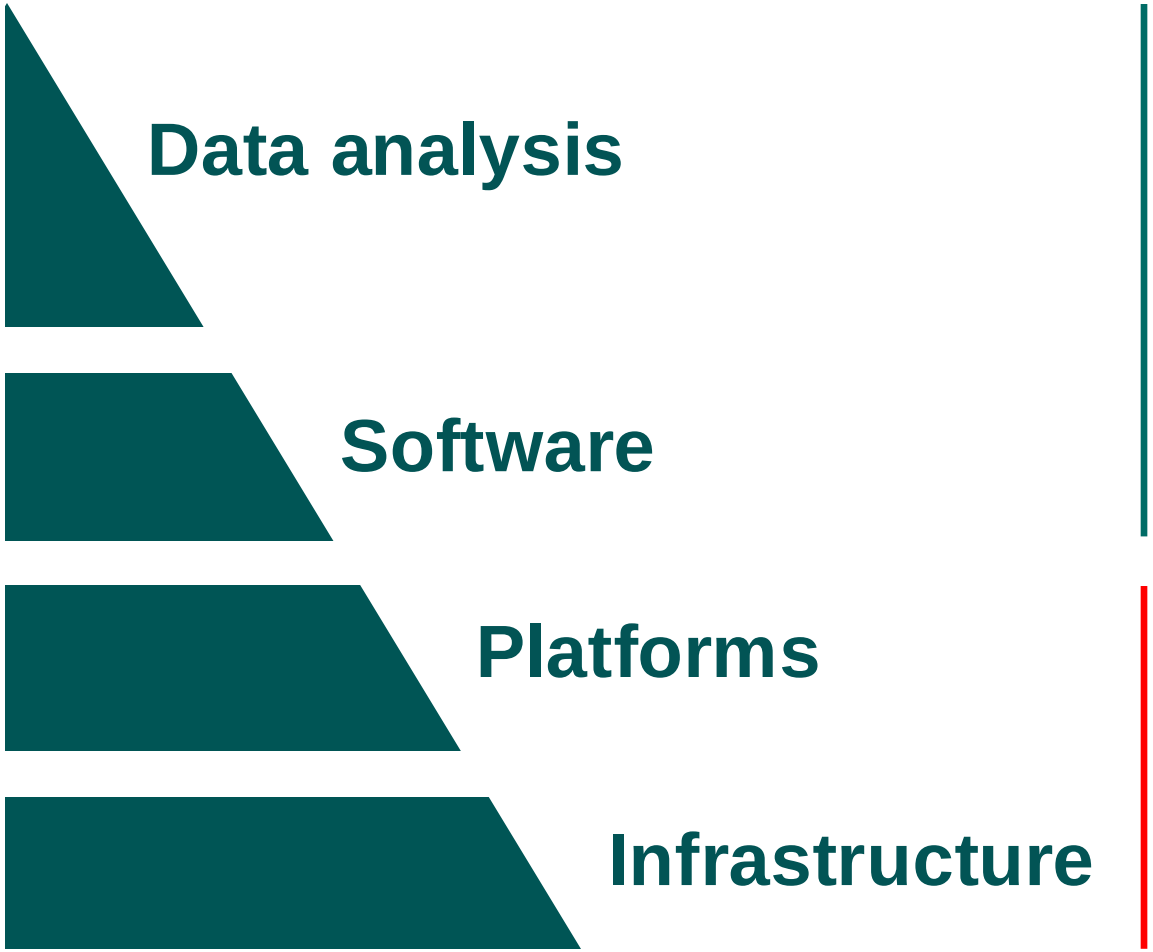
ON CLOUD FULL STACK BIOINFORMATICS

JORGE.BOUCAS@AGE.MPG.DE

HEAD OF BIOINFORMATICS

CLOUD ?



Bioinformaticians (2+1)

Systems Administrator (1)

CLOUD ?

Data analysis

Software

Platforms

Infrastructure

DevOPs
- openstack
- kubernetes
- swarm
- docker
- databases
- reverse proxy
- python

Platform-as-Code

- outsource
- down scale
- scale out

Enterprise Level Support

CLOUD ?

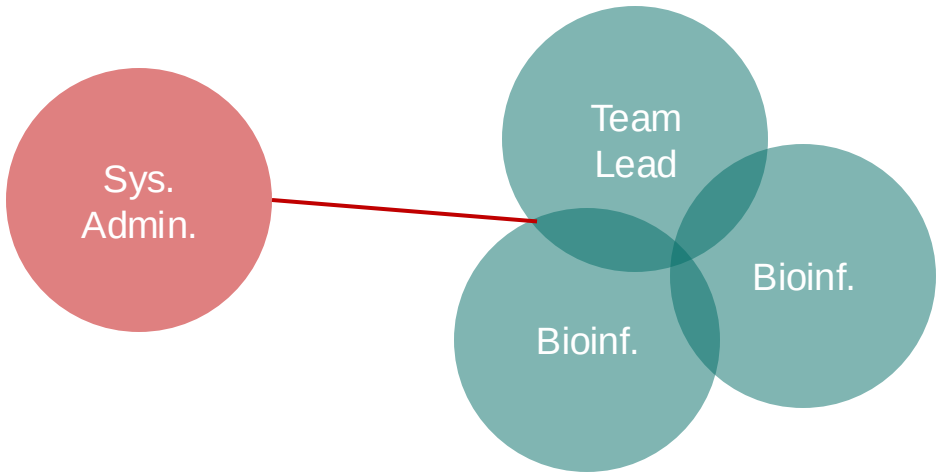
Bioinformaticians (2 + 1)

Cloud Dev Ops (1 + 1)

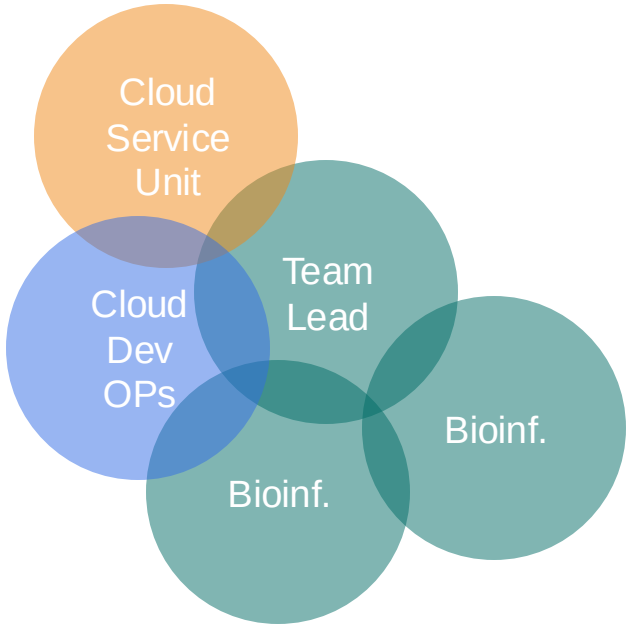
Cloud @ MPCDF

CLOUD ?

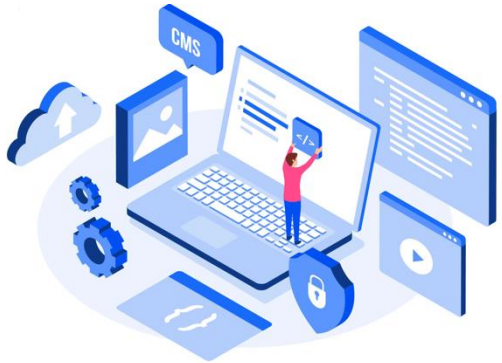
BEFORE



AFTER

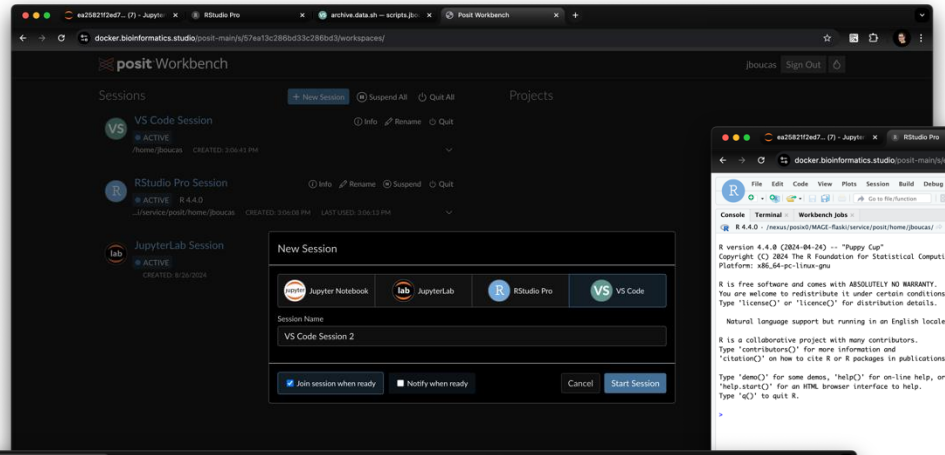


THE BIOINFORMATICS STACK

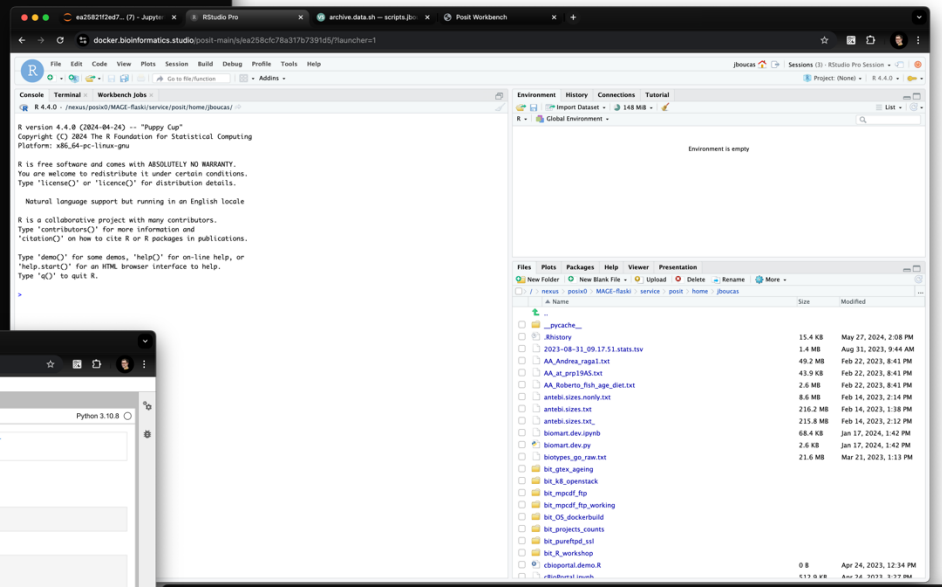
custom data analysis	*software development*	*automated pipelines*
		
Python, R	Python, R	Python, R, bash, Perl, (C/C++)
posit™ docker	kubernetes	nextflow APPTAINER slurm workload manager HPC

https://www.capellasolutions.com; https://dianapps.com; https://thinktrends.co

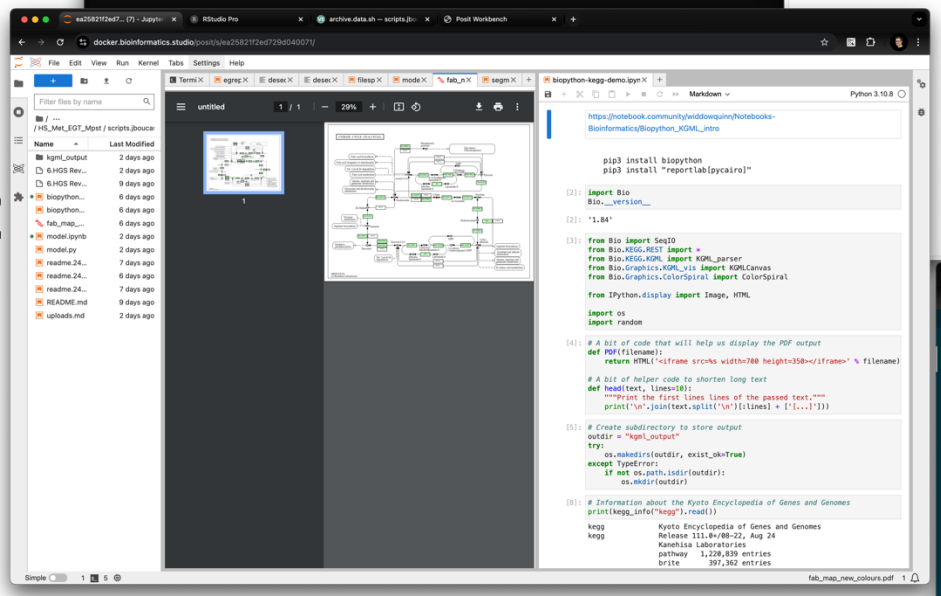
POSIT



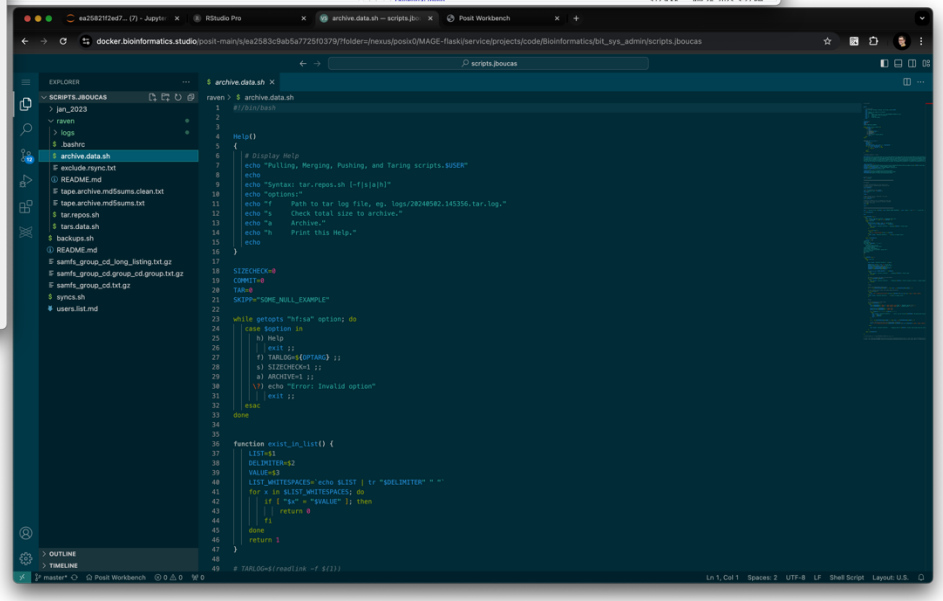
RStudio



JupyterLab



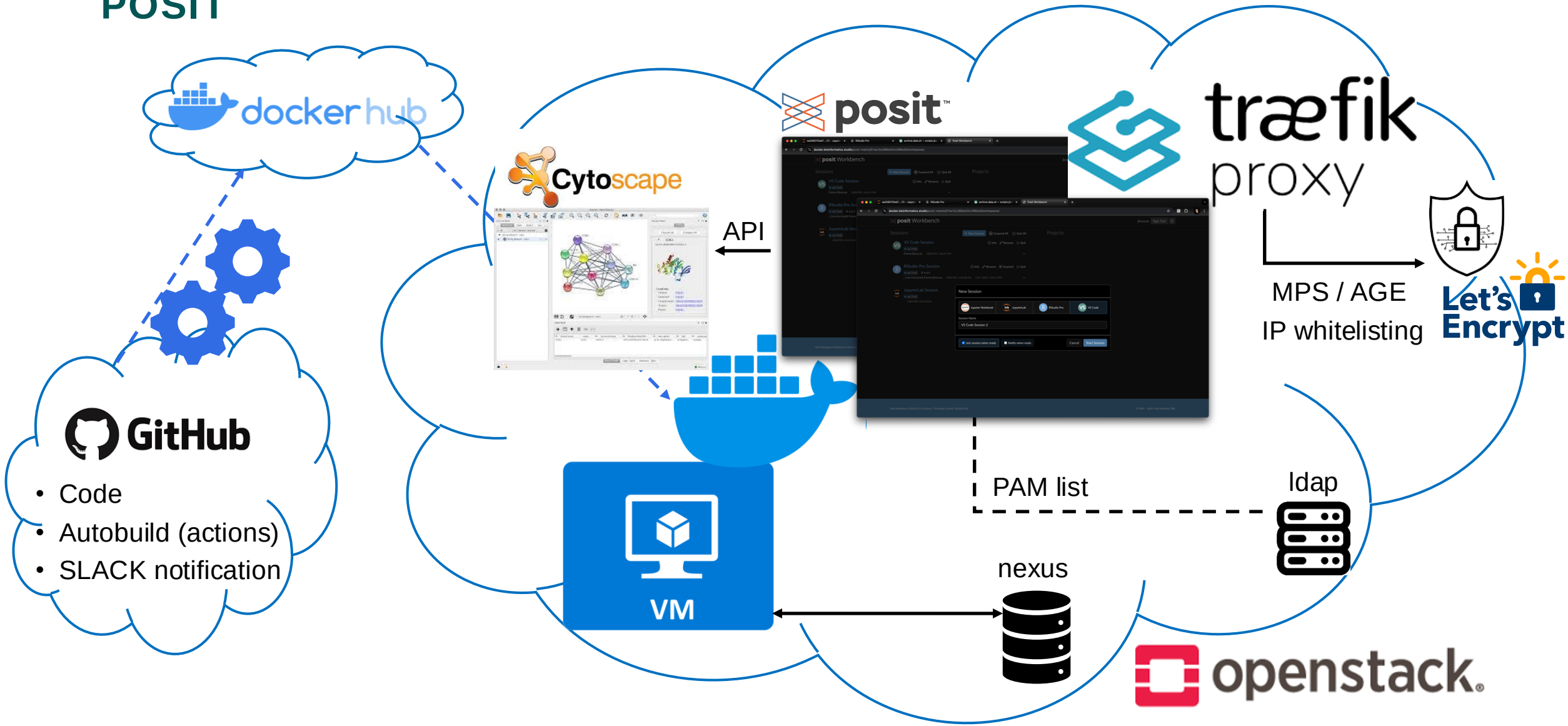
VS Code



- *multiple sessions*
- *multiple kernels*
- *multi user*

https://www.capellasolutions.com; https://dianapps.com; https://thinktrends.co

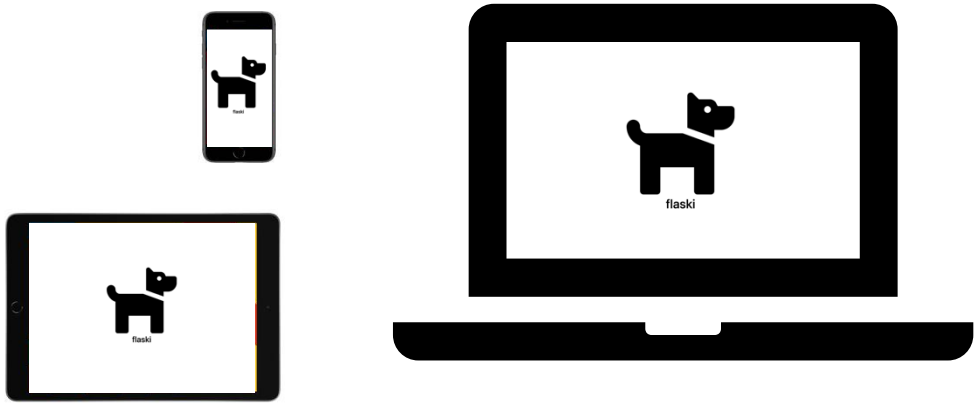
POSIT



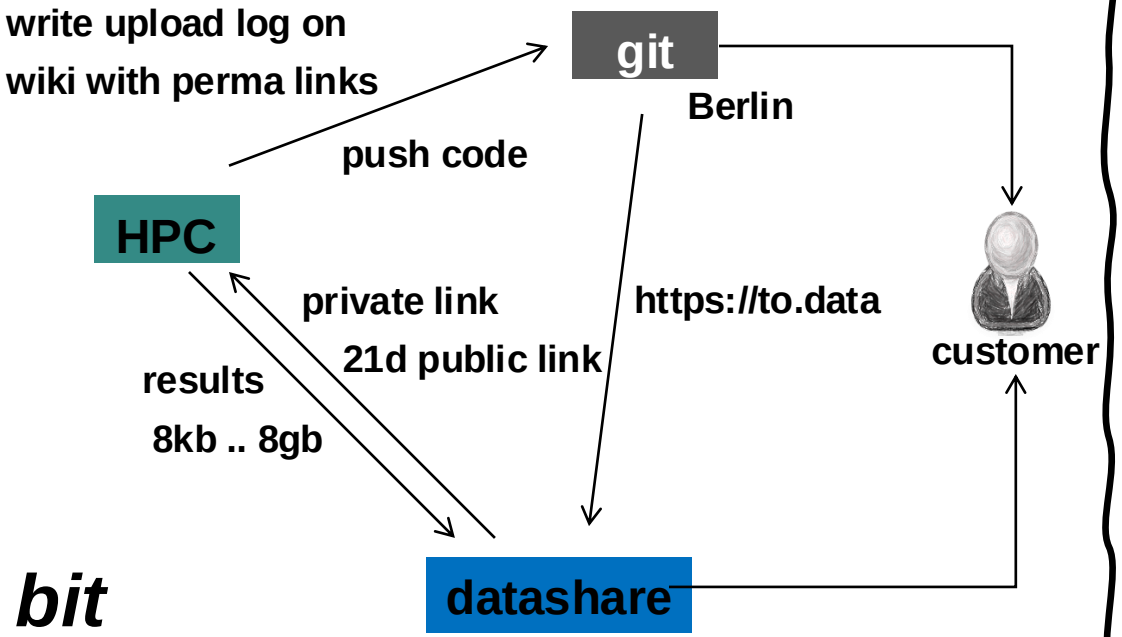
https://www.starwindsoftware.com; https://logixconsulting.com

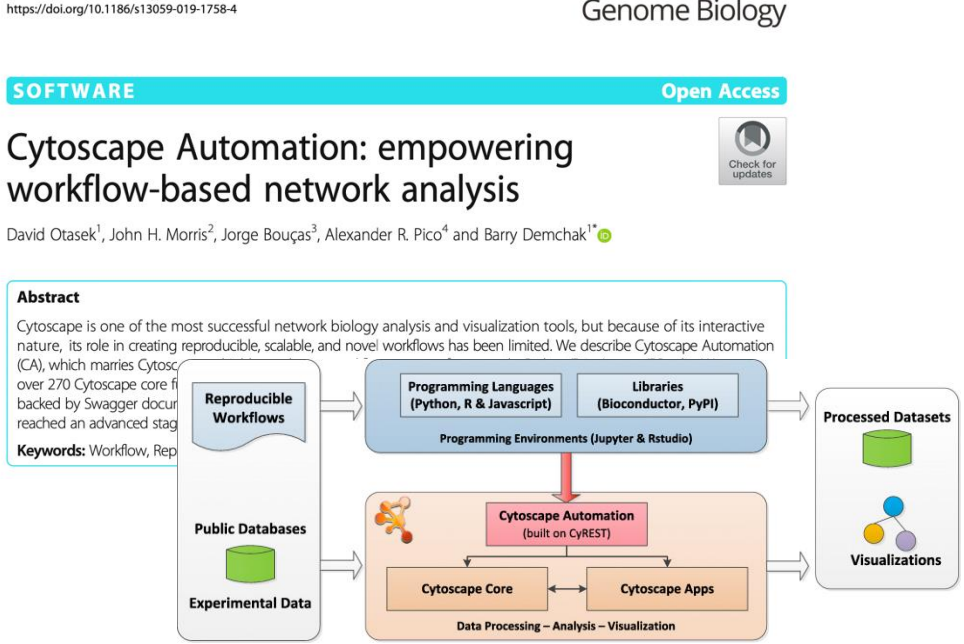
SOFTWARE

- Operational
- Scientific
- WebApps



flaski.age.mpg.de




Otasek et al. Genome Biology (2019) 20:185
https://doi.org/10.1186/s13059-019-1758-4
Genome Biology
SOFTWARE
Open Access

Cytoscape Automation: empowering workflow-based network analysis

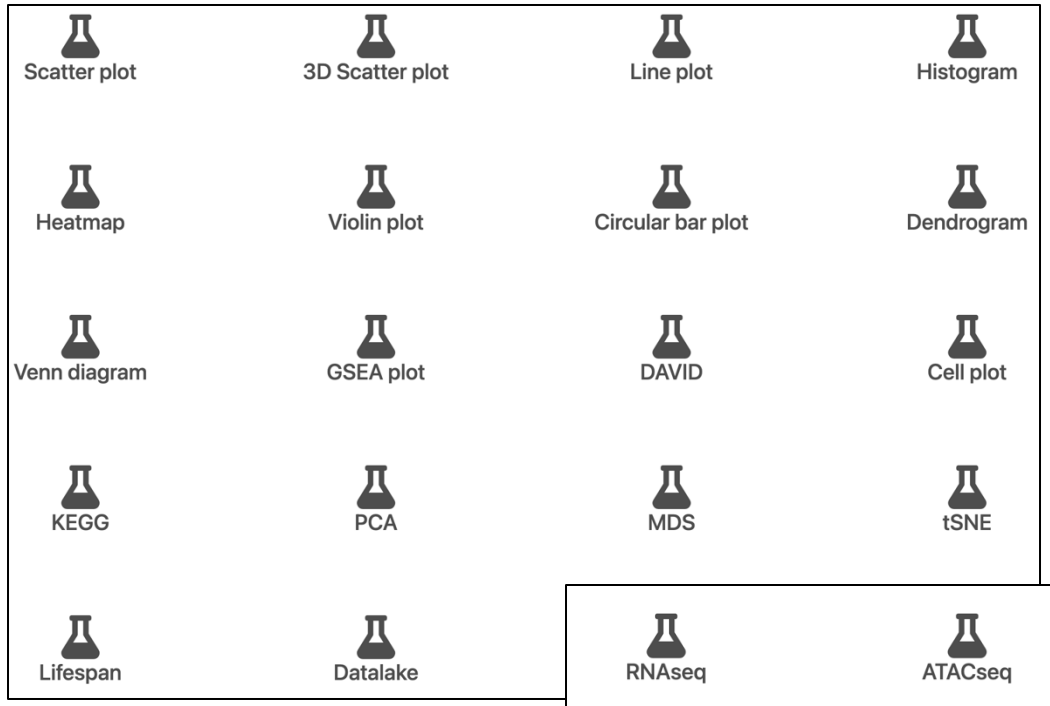
David Otasek[1], John H. Morris[2], Jorge Bouças[3], Alexander R. Pico[4] and Barry Demchak[1*]

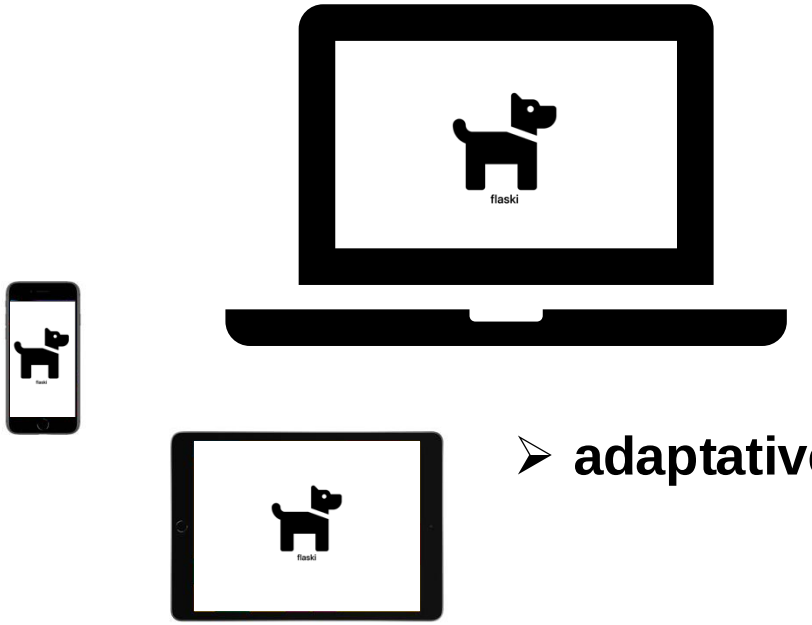
Abstract

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 Cytosc
over 270 Cytoscape core f
backed by Swagger docur
reached an advanced stag

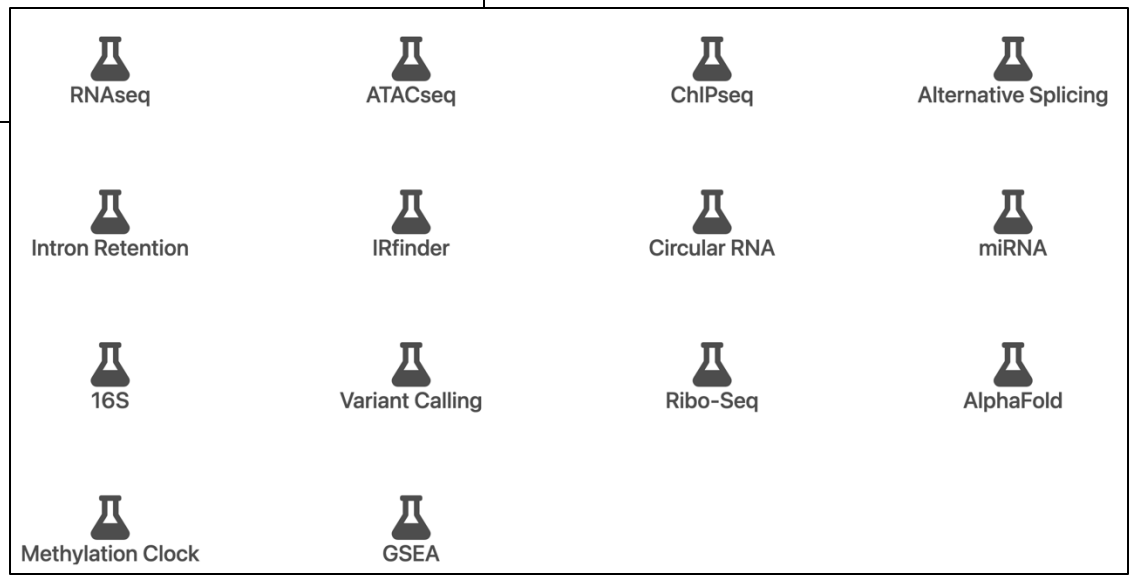
Keywords: Workflow, Rep

FLASKI





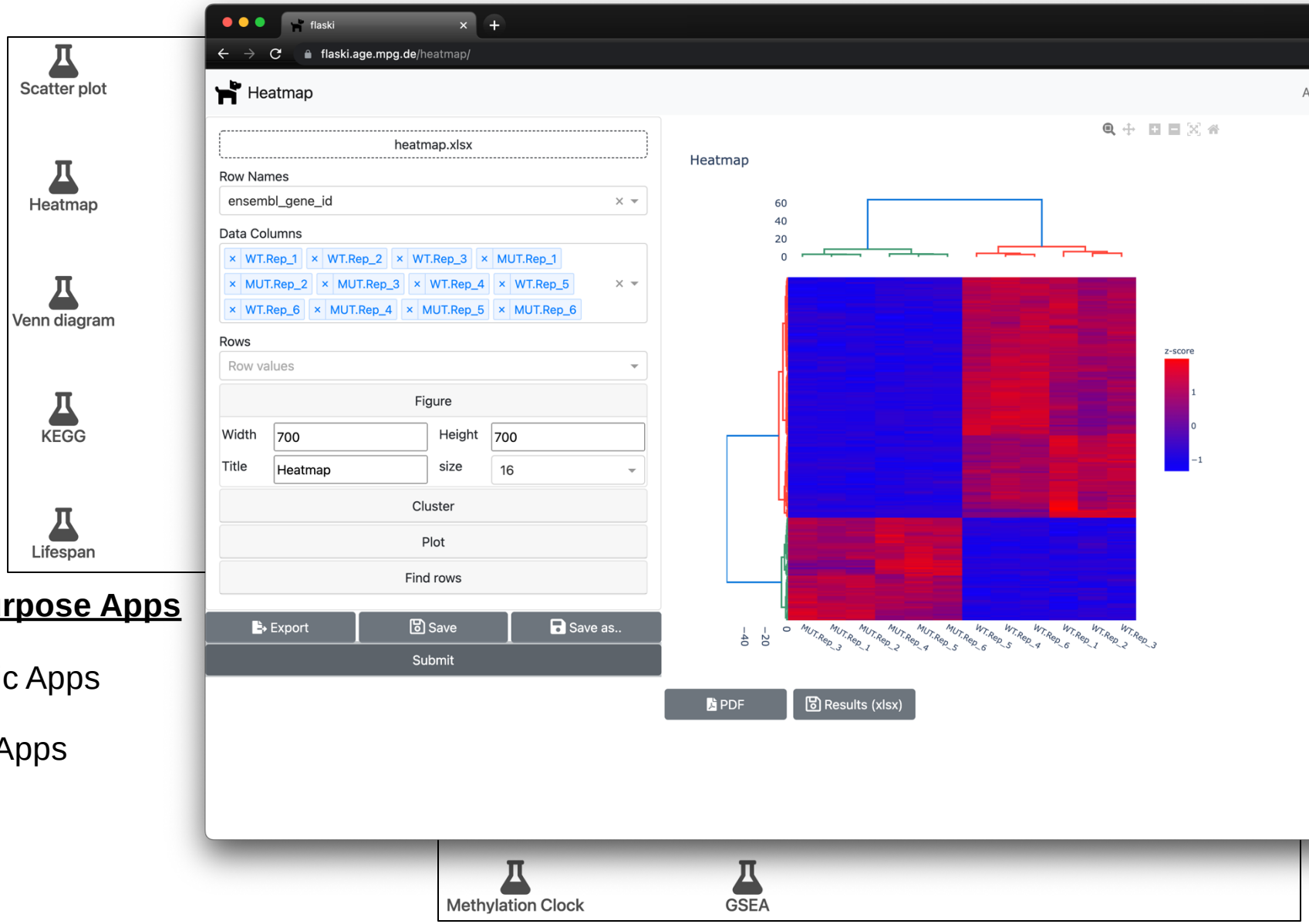
➢ **adaptative**



A collection of web apps for life sciences

image sources: https://fontawesome.com

APPS

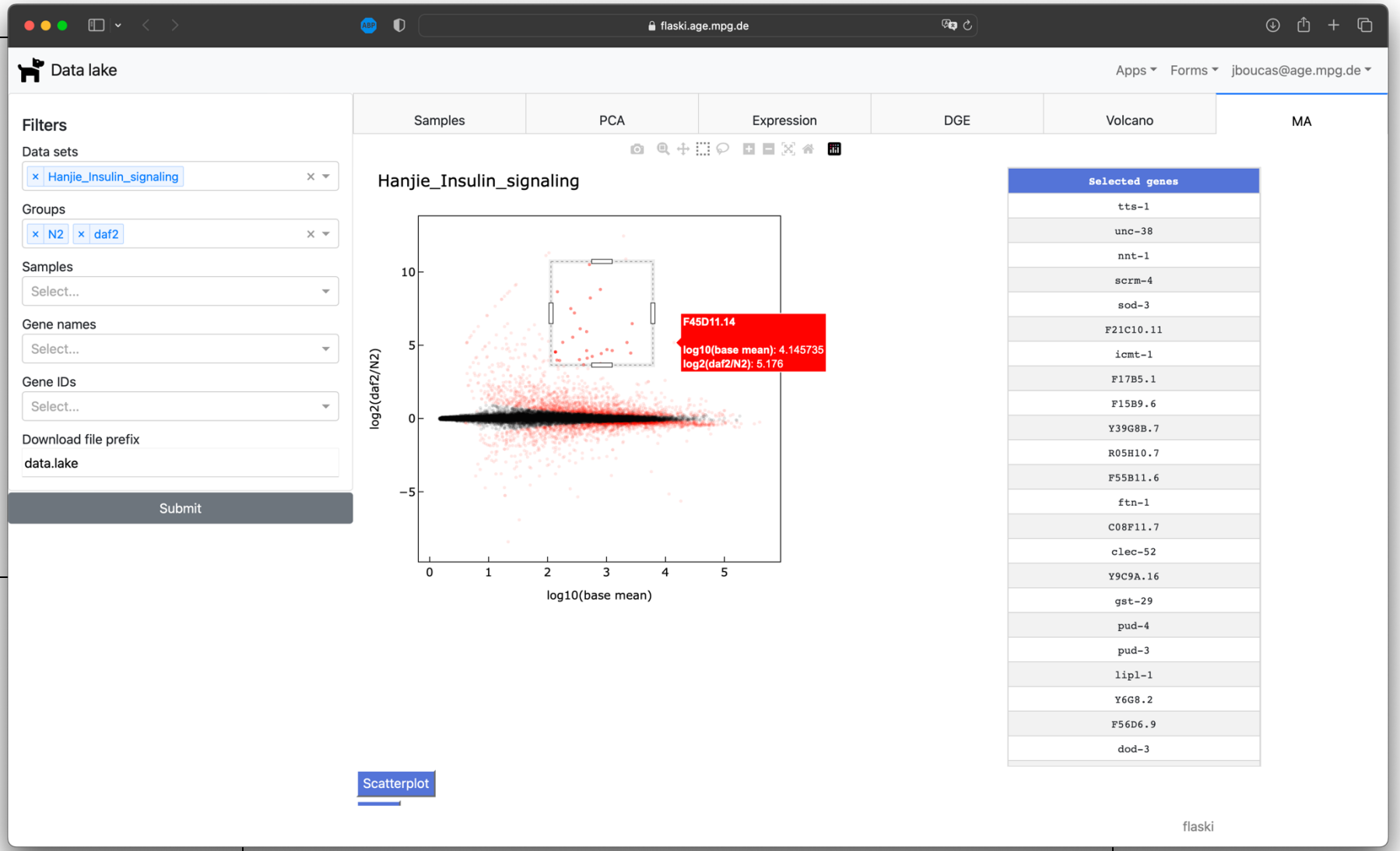


- **general purpose Apps**
- data specific Apps
- interactive Apps
- app2app
- forms

image sources: https://fontawesome.com

APPS

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

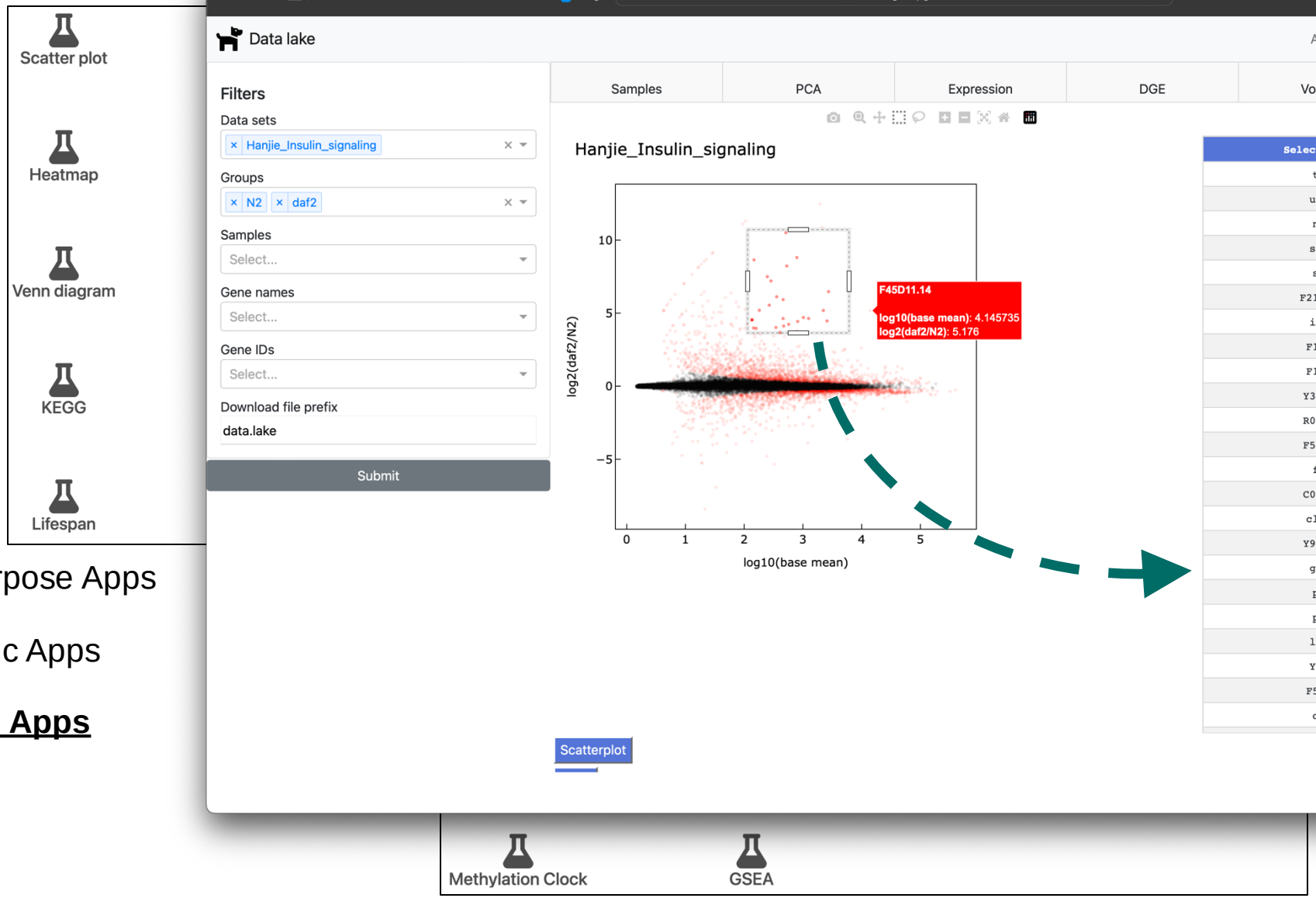


- Methylation Clock
- GSEA

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

image sources: https://fontawesome.com

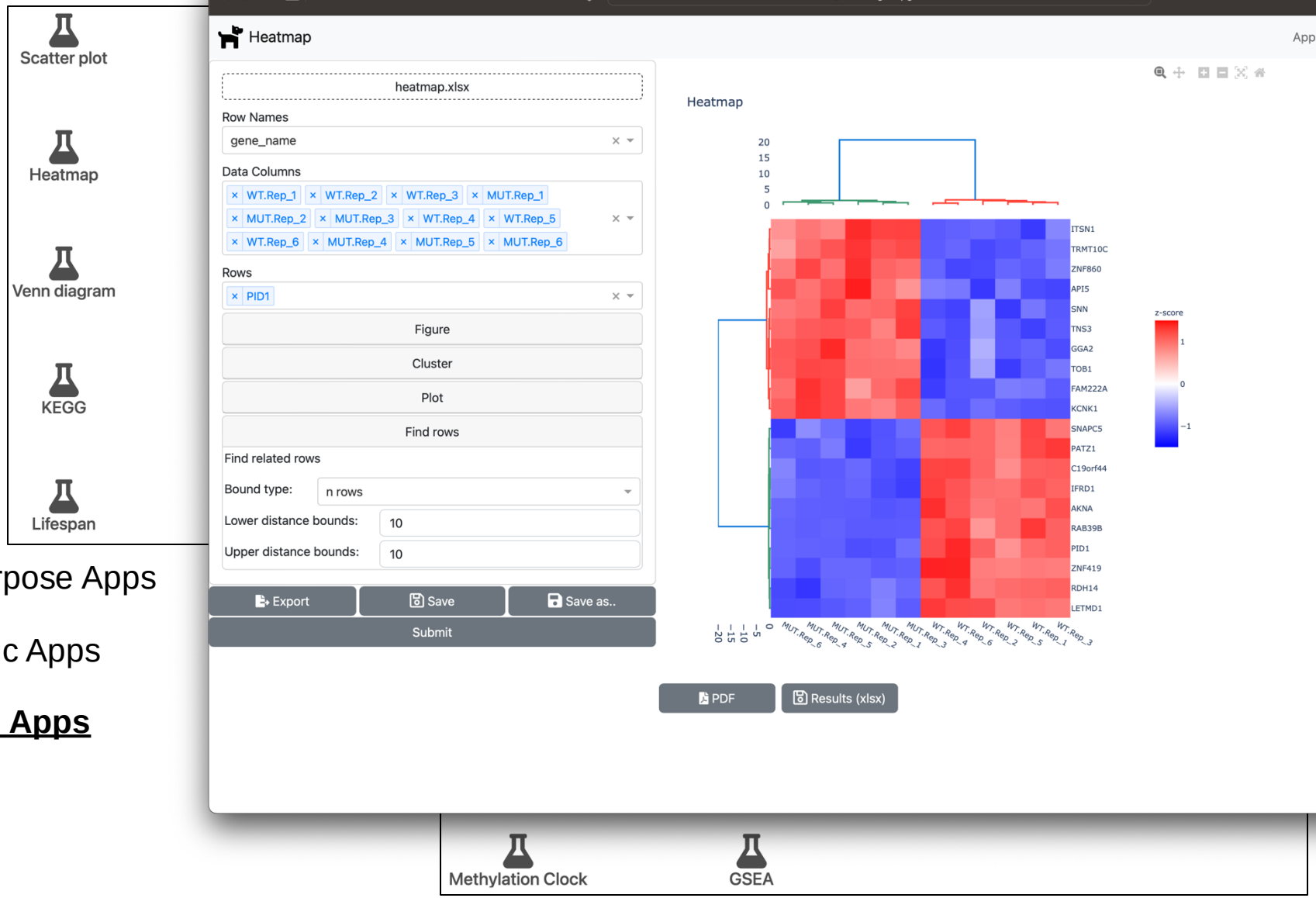
APPS



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

image sources: https://fontawesome.com

APPS



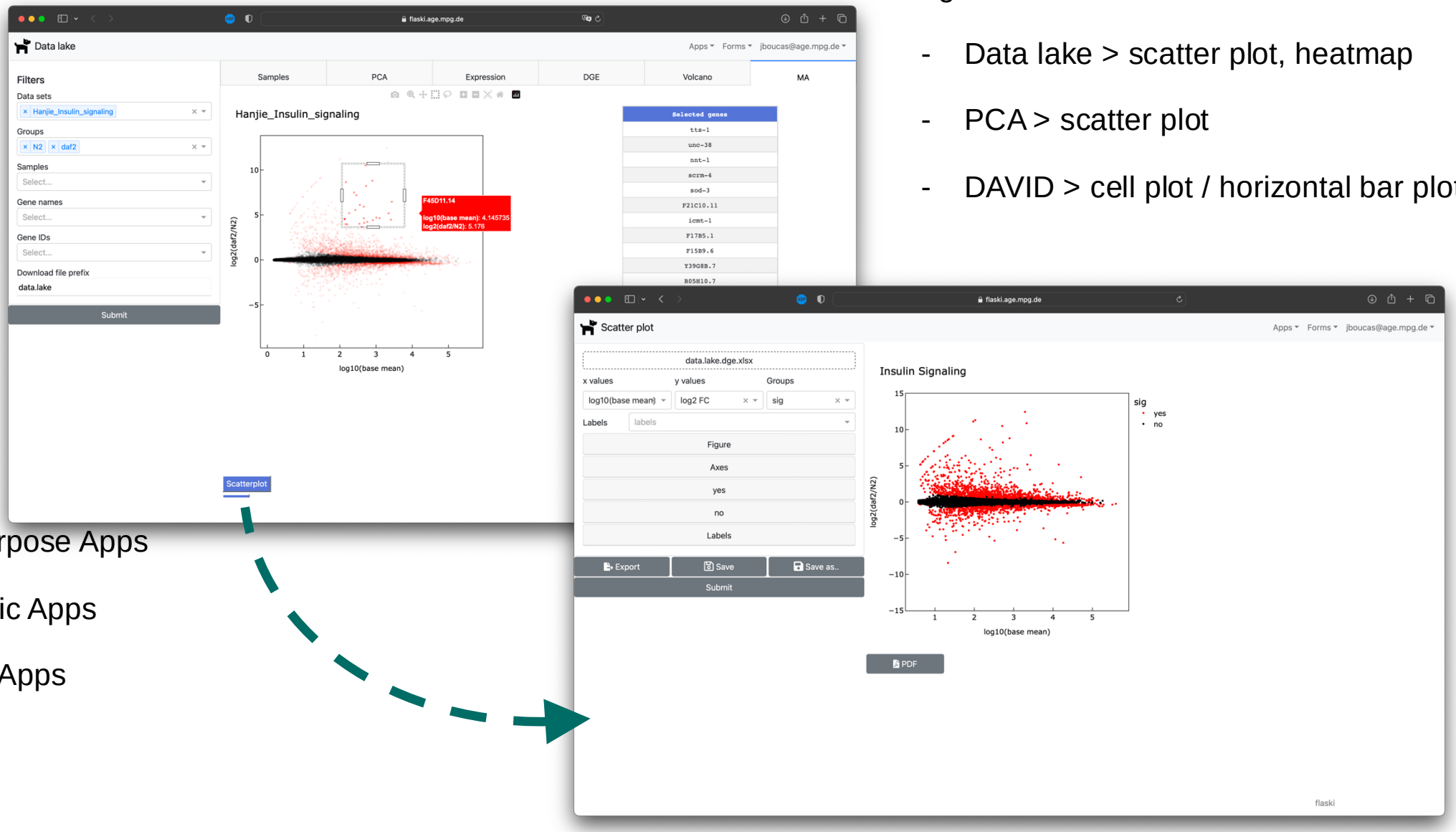
- general purpose Apps
- data specific Apps
- **interactive Apps**
- app2app
- forms

image sources: https://fontawesome.com

APPS

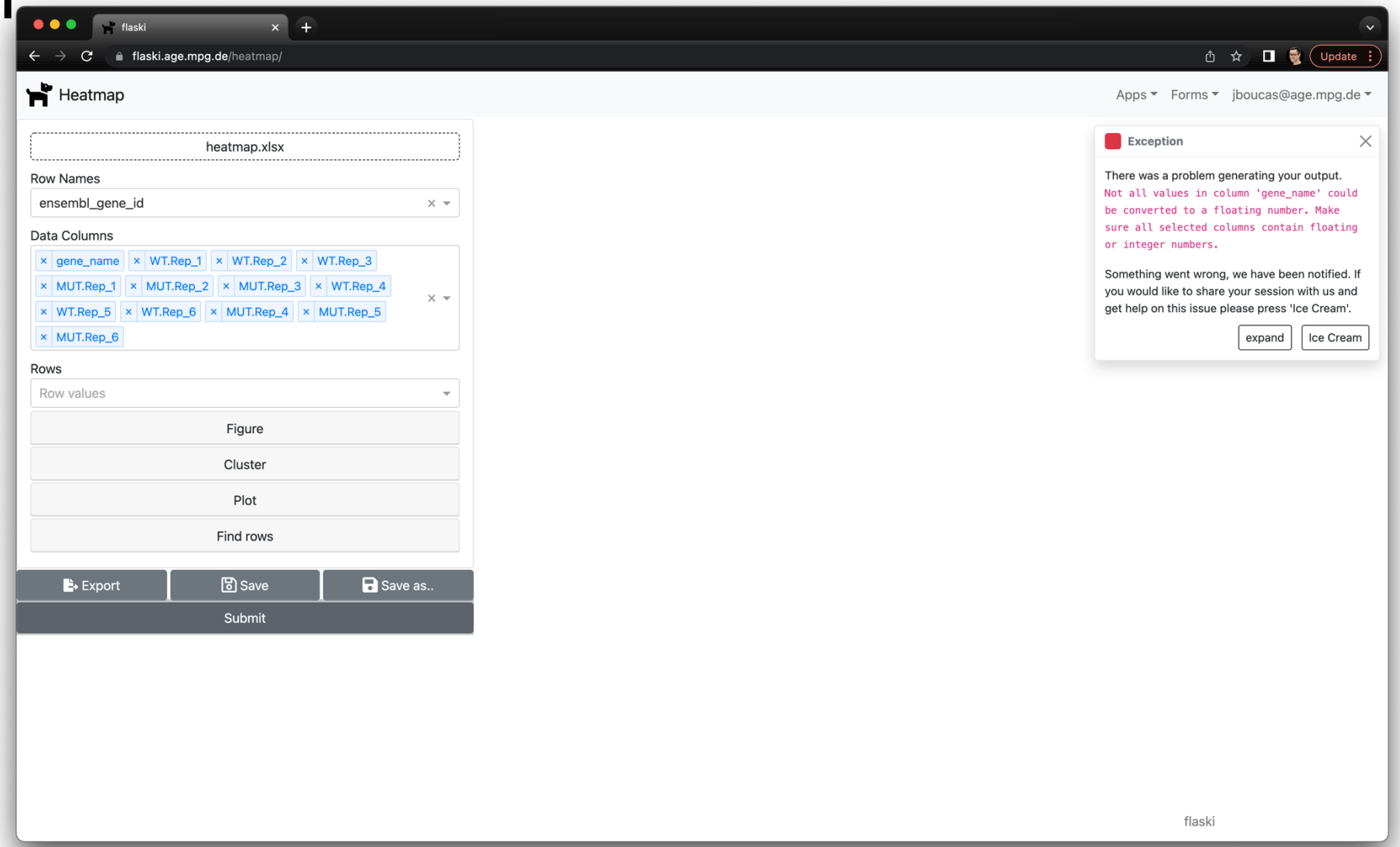
eg.

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



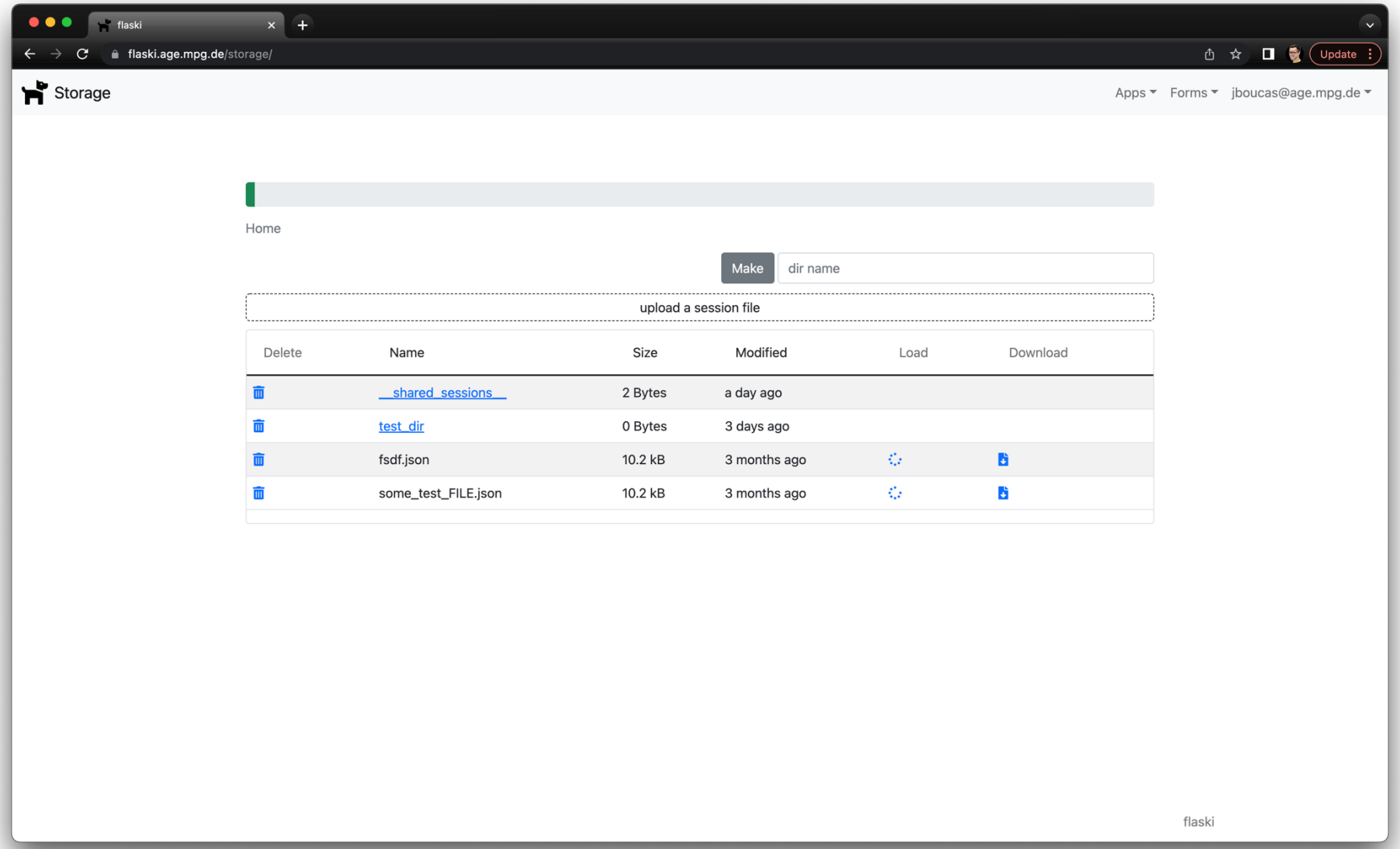
- general purpose Apps
- data specific Apps
- interactive Apps
- **app2app**
- forms

LIVE SUPPORT



- **Error reporting**
- **Custom error msg.**
- **Session sharing**

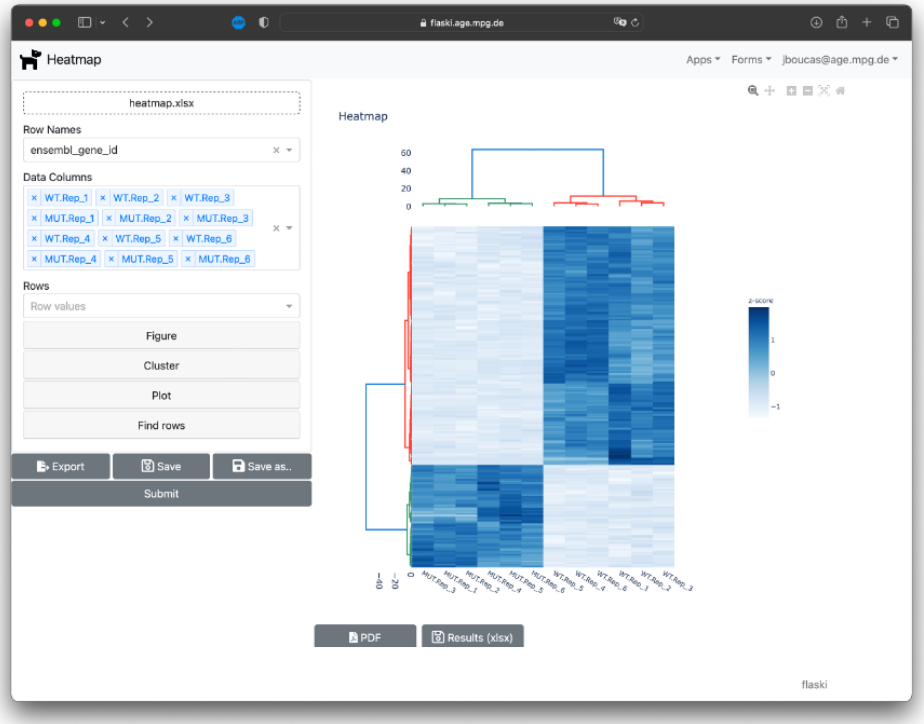
SESSIONS



- **Storage**
- **pyflaski versioning**

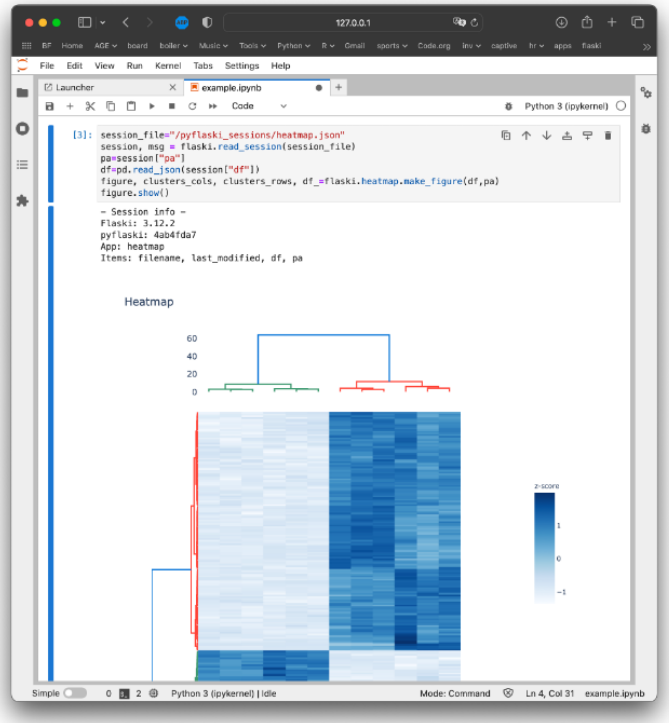
PYFLASKI

web App (graphical user interface)



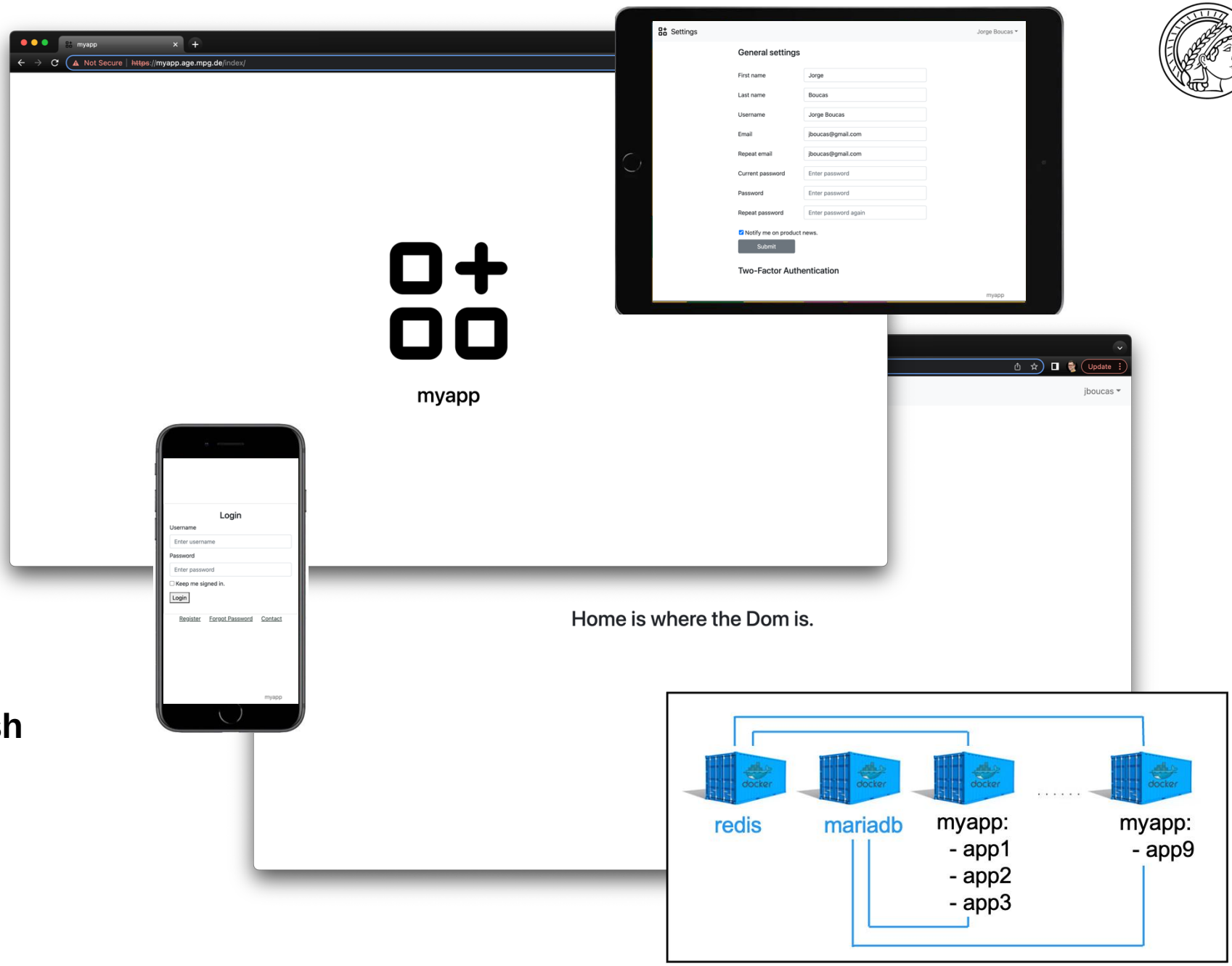
session transfer

Jupyter Notebook (programmatic interface)



MYAPP

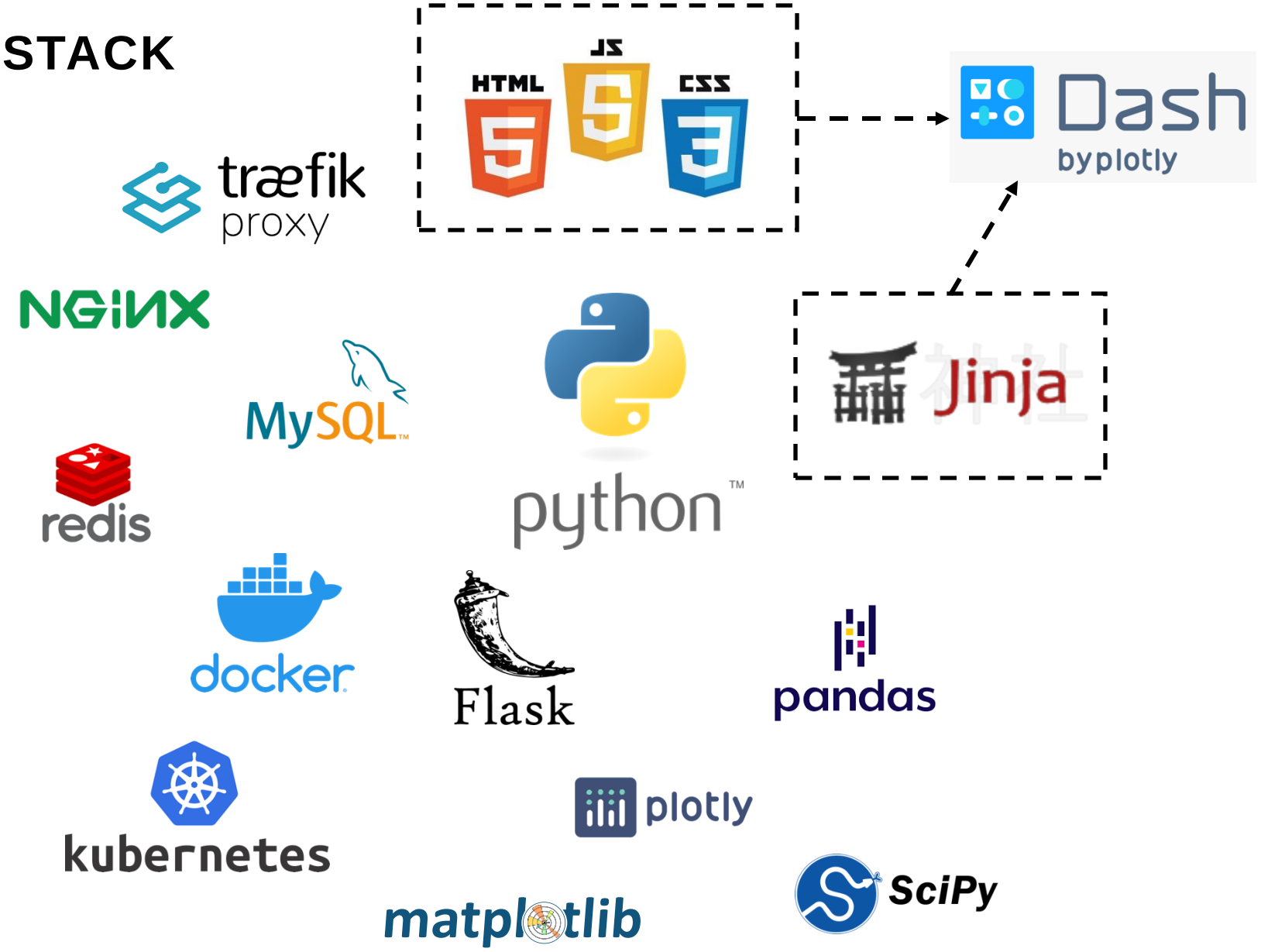
- docker-compose.yml
- kubernetes.yml
- mariadb & redis
- granular user access
- 2FA
- admin auth. tokens for new users
- user & admin dashboard
- Let's Encrypt
- amd64, arm64, arch64
- apps as routes
- apps as containers / plugins



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

image sources: https://fontawesome.com

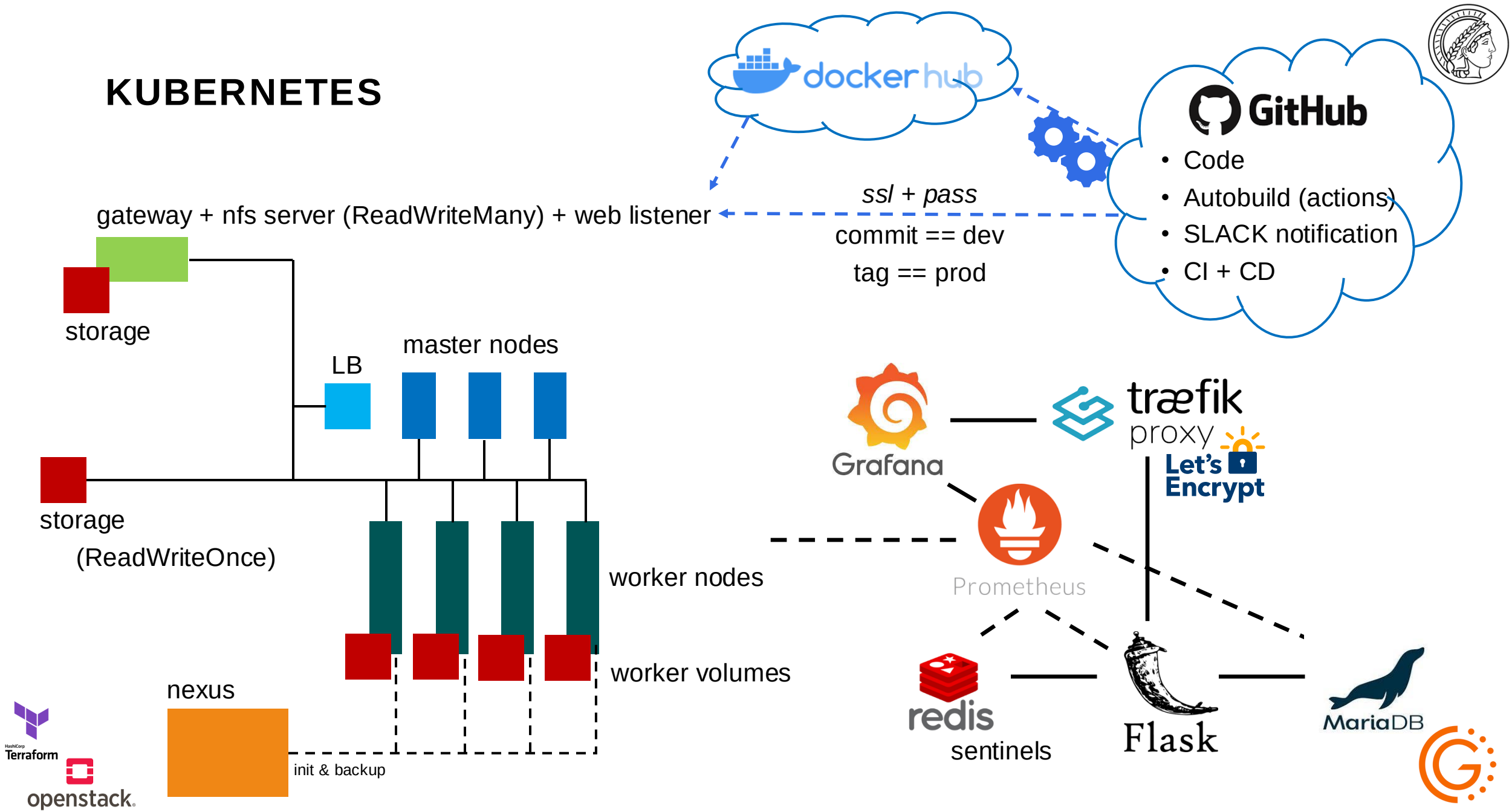
WEB DEV STACK



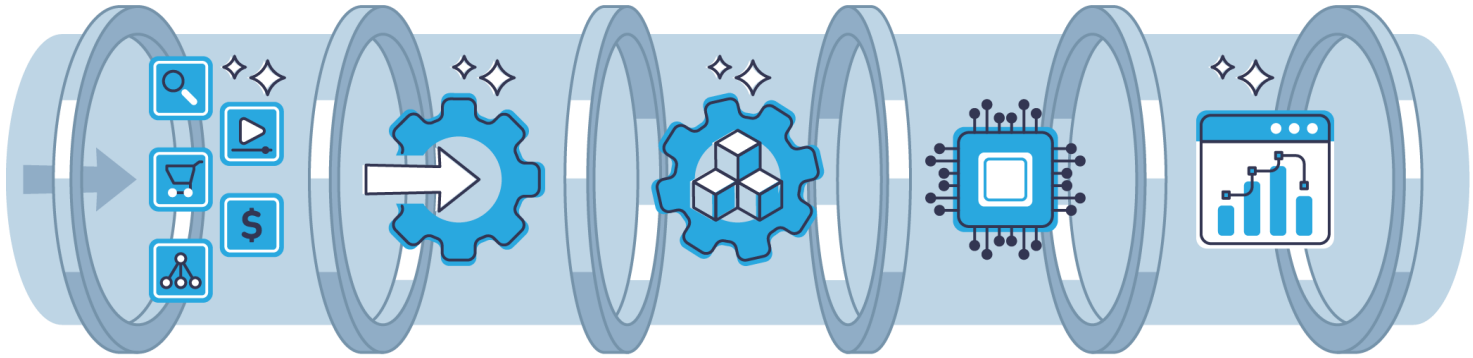
SUMMARY

- ✓ 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

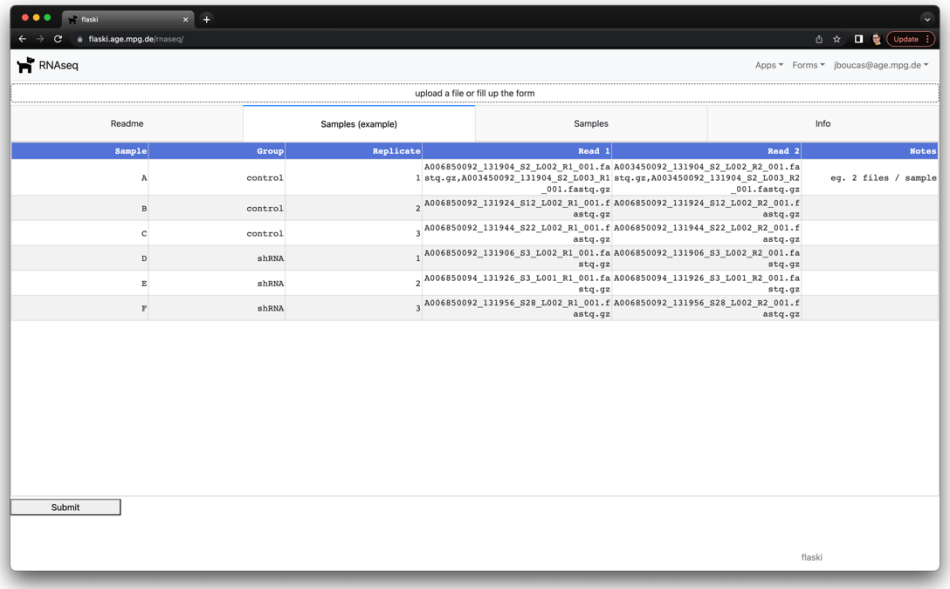

docker hub
GitHub
• Code
• Autobuild (actions)
• SLACK notification
• CI + CD
ssl + pass
commit == dev
tag == prod
gateway + nfs server (ReadWriteMany) + web listener
storage
LB
master nodes
storage
(ReadWriteOnce)
worker nodes
worker volumes
nexus
init & backup
Grafana
træfik
proxy
Let's Encrypt
Prometheus
redis
sentinels
Flask
MariaDB
HashiCorp
Terraform
openstack.

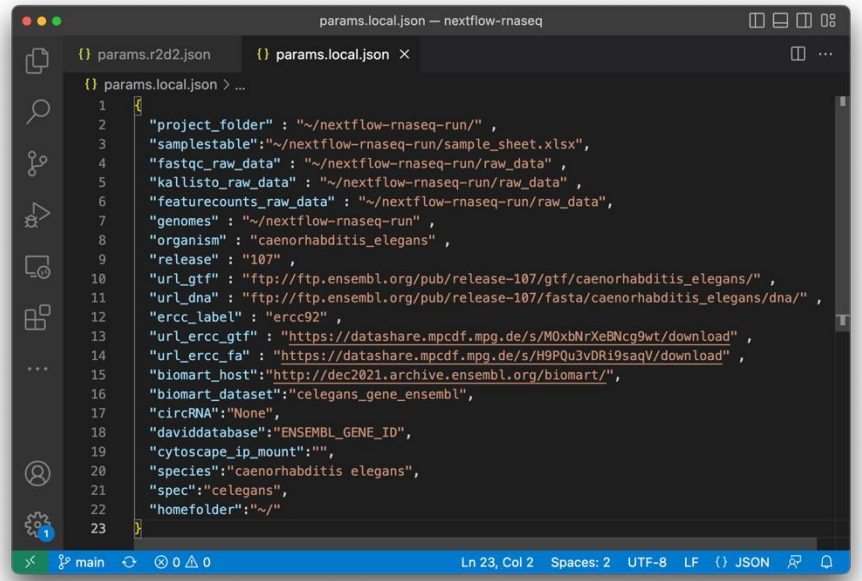
AUTOMATION



https://www.capellasolutions.com; https://dianapps.com; https://thinktrends.co

AUTOMATION





nextflow

docker hub

HPC MPCDF / RAVEN

HPC MPI-AGE / CLOUD

AWS Batch

Local

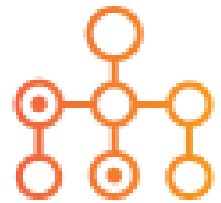
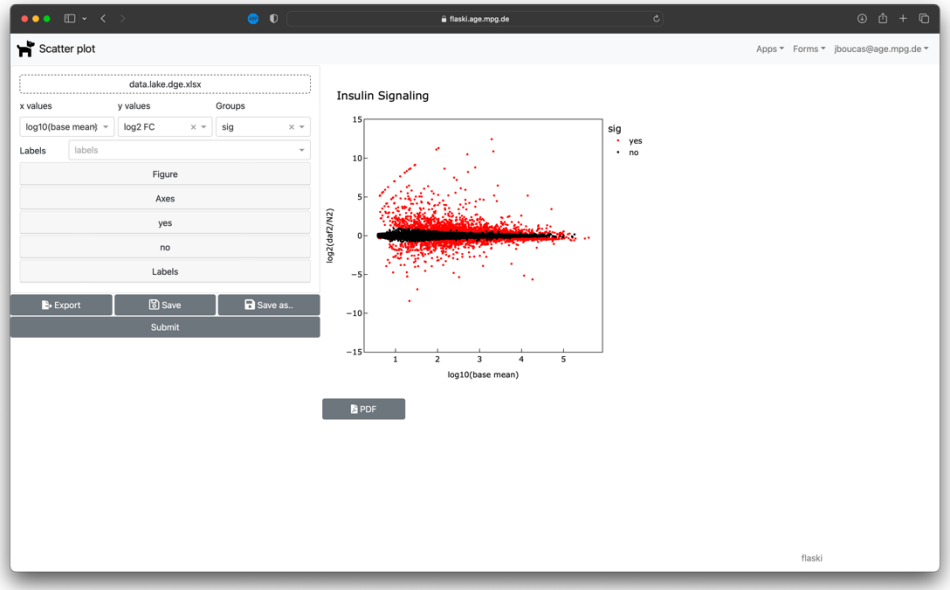


image sources: https://dareplanet.tech

AUTOMATION



session.json

owncloud

owncloud API

HPC MPCDF / RAVEN

HPC MPI-AGE / CLOUD

Link to results

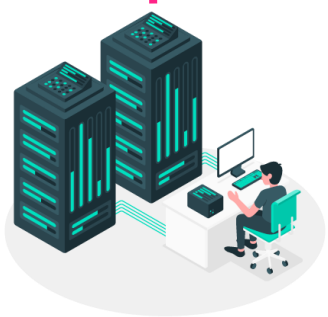
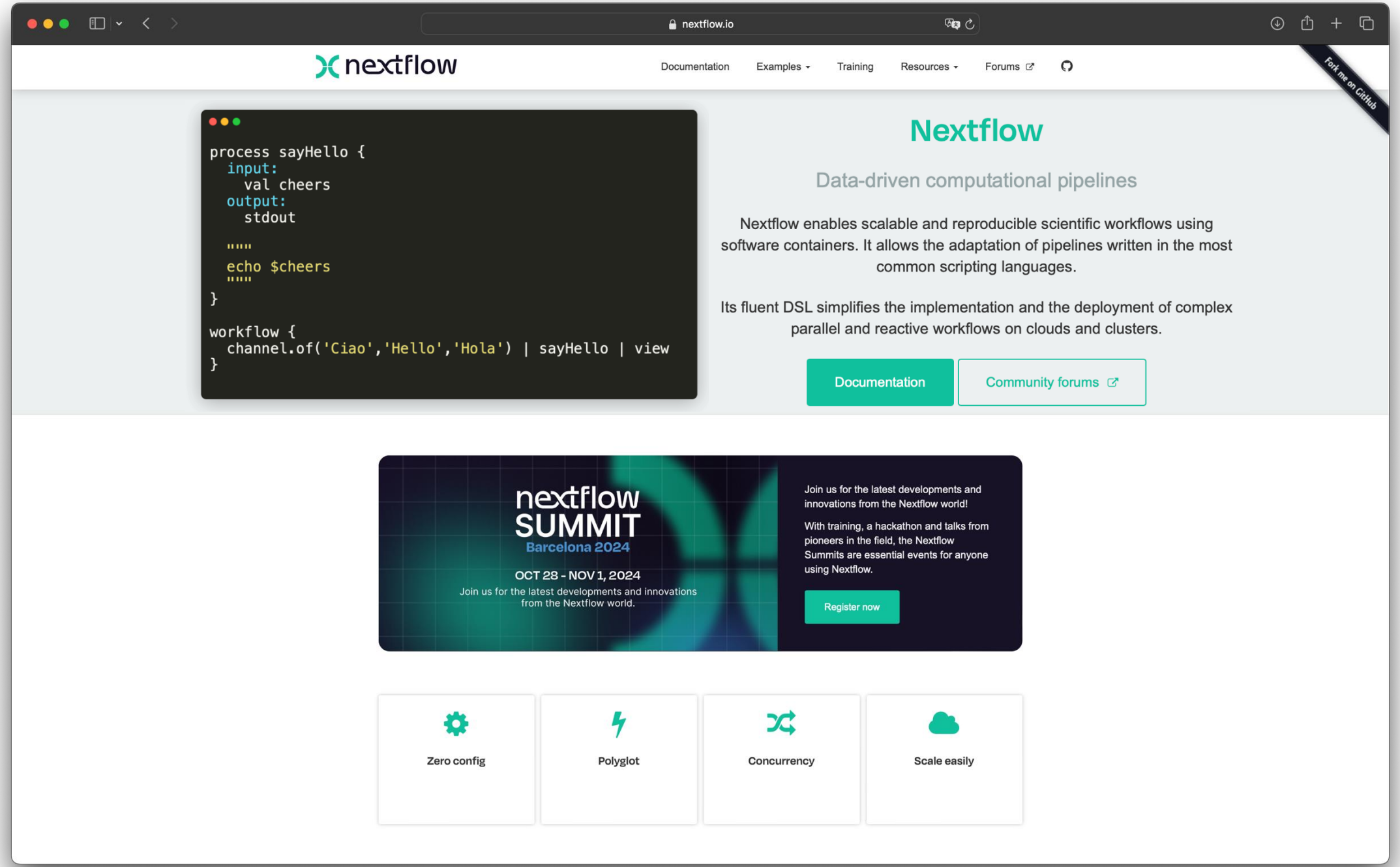
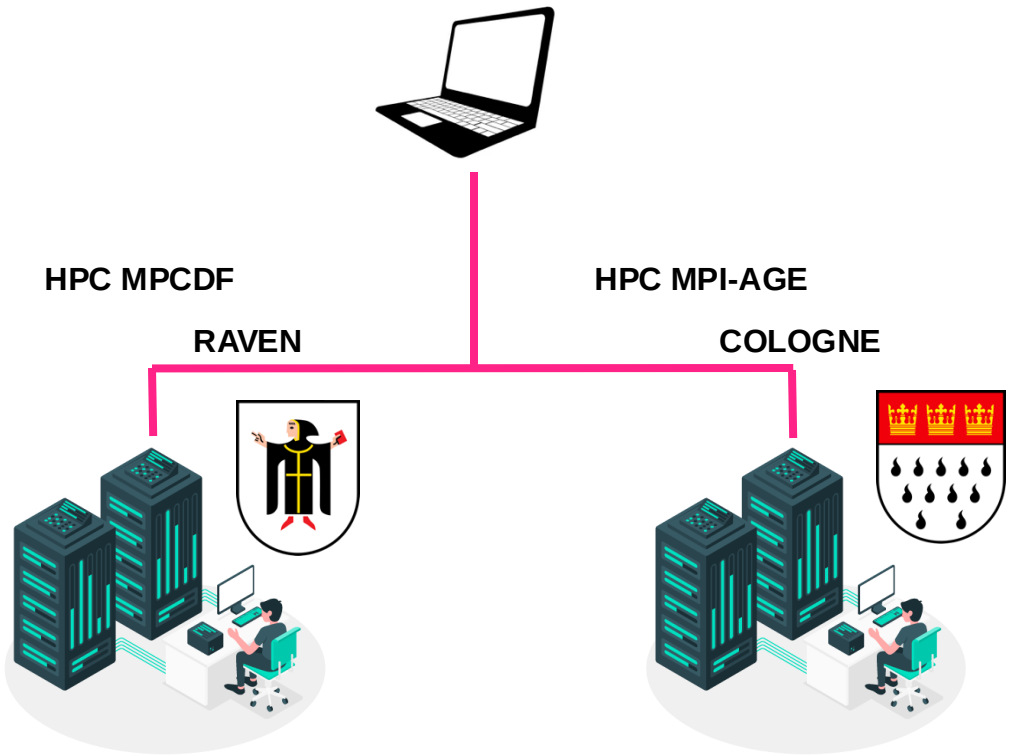


image sources: https://dareplanet.tech


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

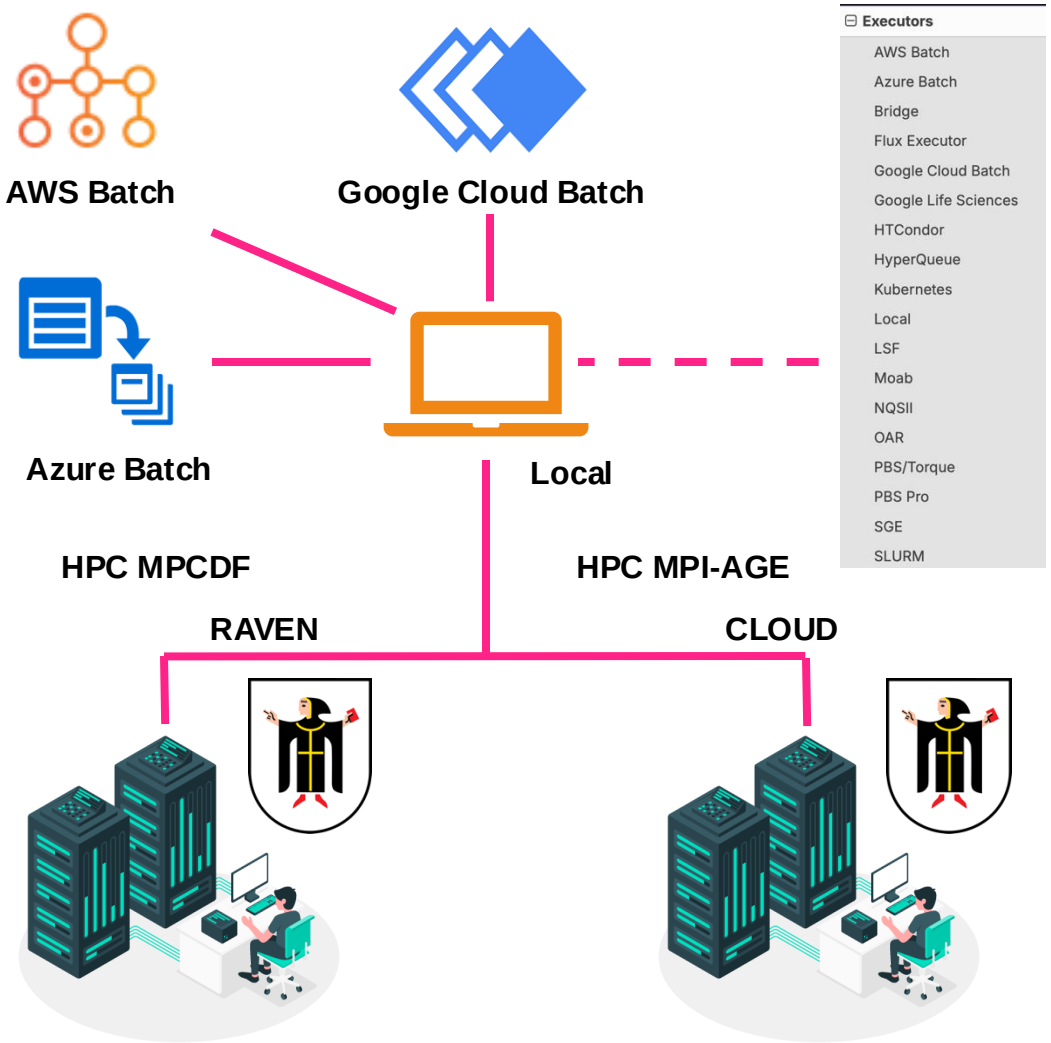
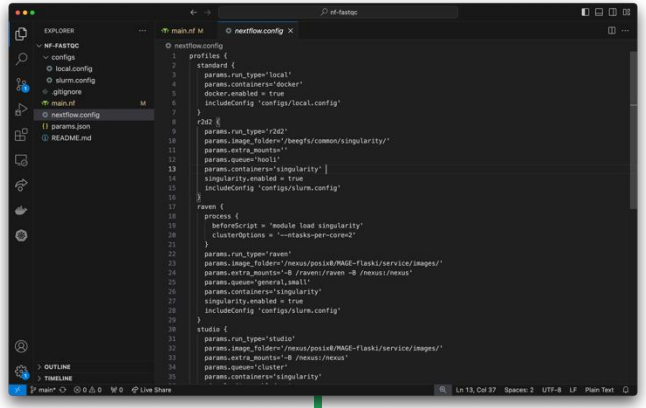
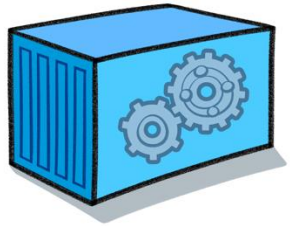


image sources: https://dareplanet.tech

Executor profile: slurm / local / ..



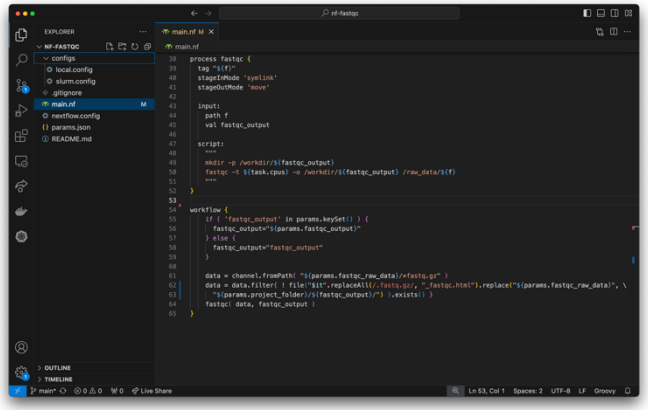
Modular



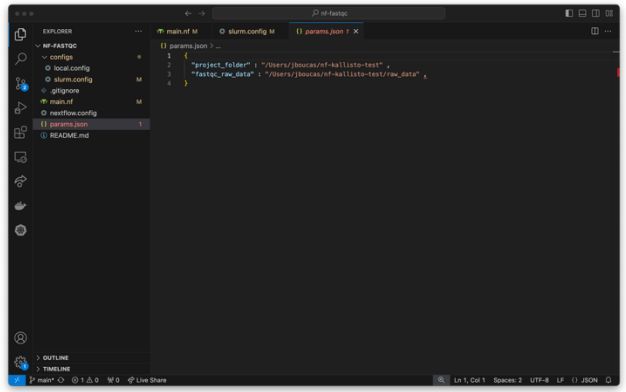
Docker / Singularity / ..

nextflow

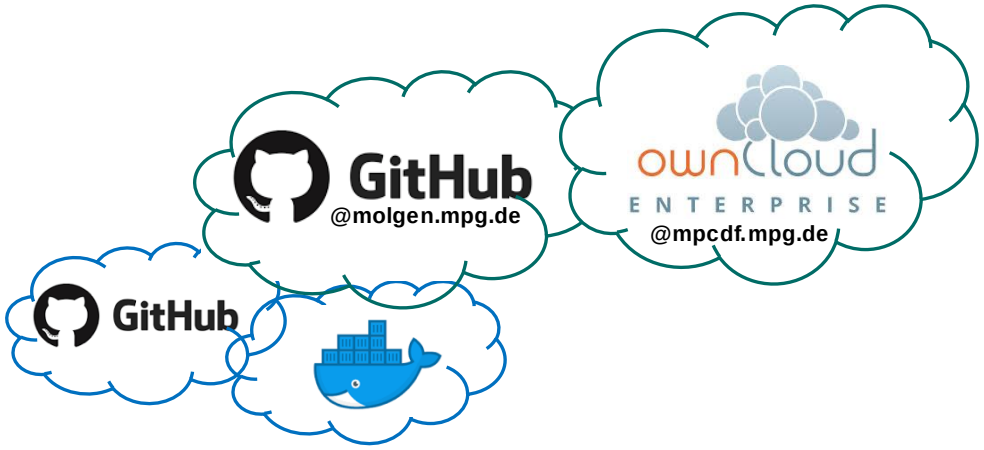
Universal code: bash / Python / R / ..

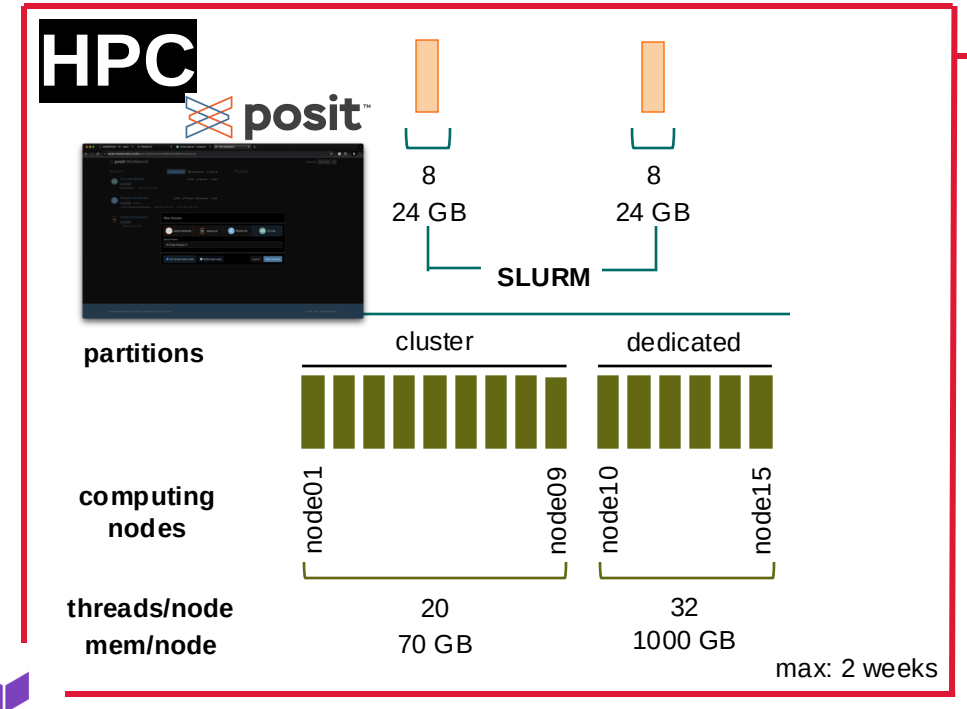


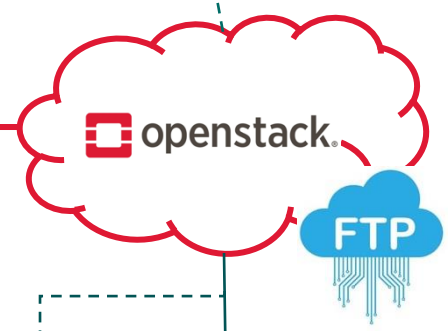
Run specific parameters: variables

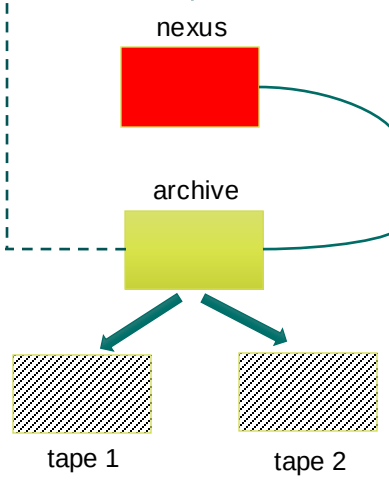


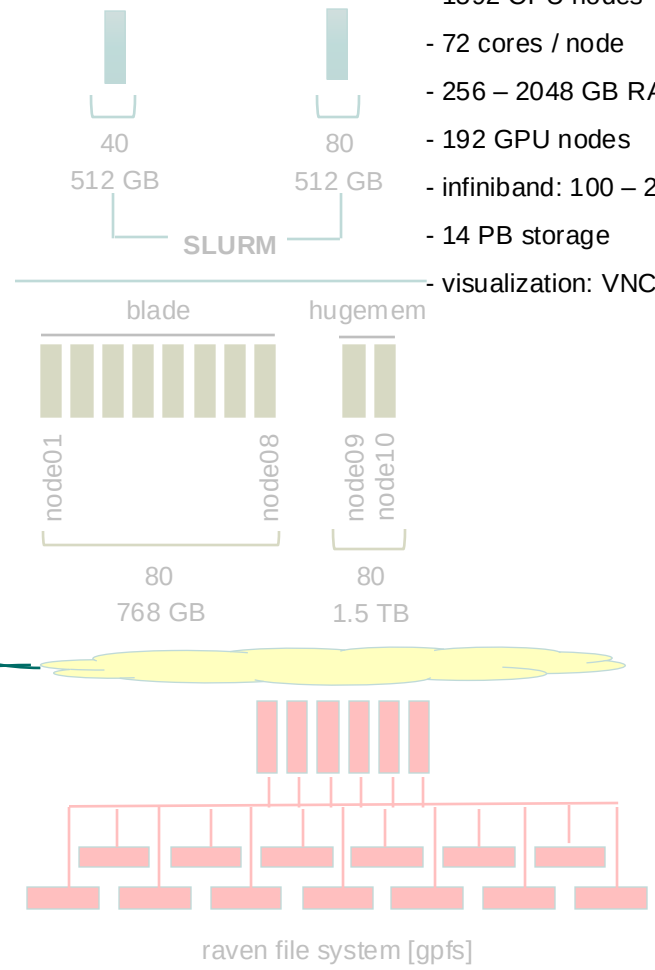
HPC INFRASTRUCTURE











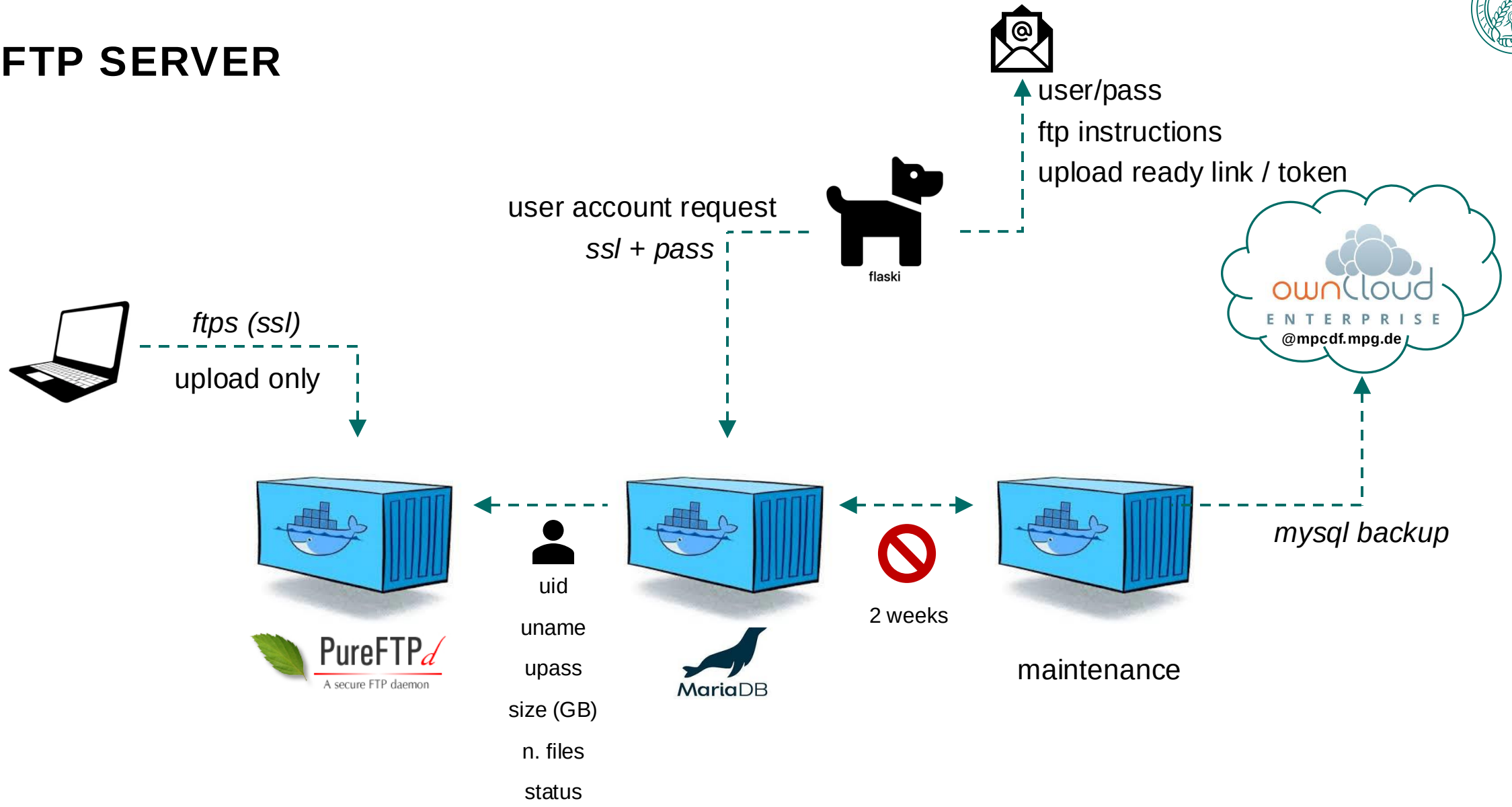
HPC

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

HashiCorp Terraform

FTP SERVER


user/pass
ftp instructions
upload ready link / token
user account request
ssl + pass
flaski
ownCloud
ENTERPRISE
@mpcdf.mpg.de
ftps (ssl)
upload only
uid
uname
upass
size (GB)
n. files
status
2 weeks
mysql backup
PureFTPd
A secure FTP daemon
MariaDB
maintenance

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_setfacl != setfacl

/ RAVEN

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 smal files:

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

NEXUS: GROUP SELF CLEANING

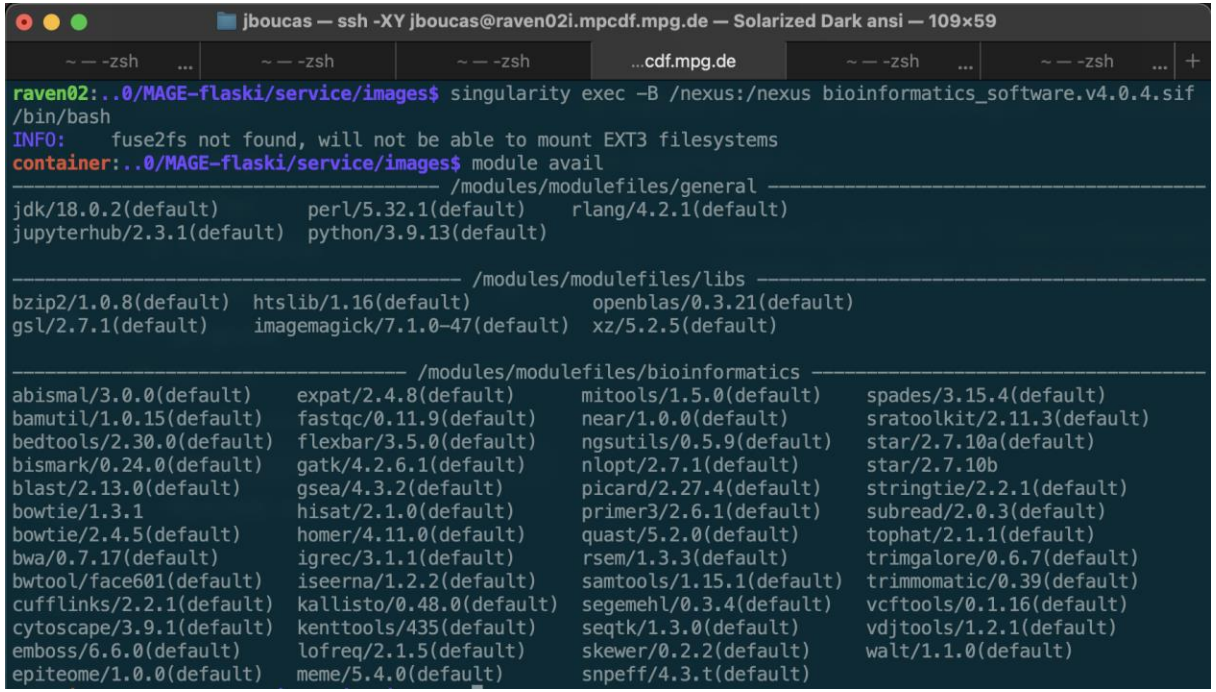
```
/nexus/posix0/MAGE-flaski/group/Bioinformatics
    /raw_data
        /project_x
    /code
        /project_x
    /data
        /project_x
```

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

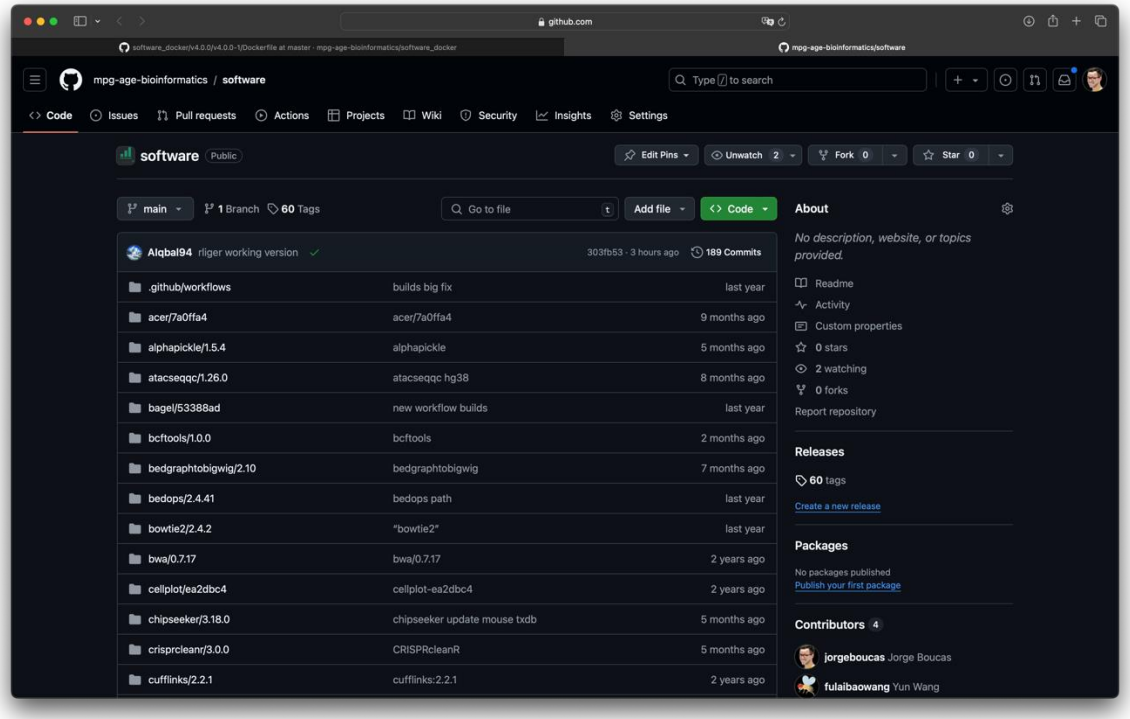
SOFTWARE

One image, multiple software (modules system)



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

One image per software

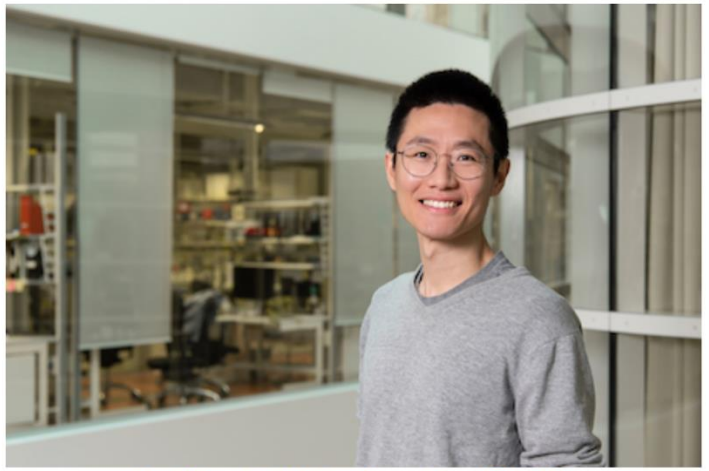


https://github.com/mpg-age-bioinformatics/software

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

HEROES

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



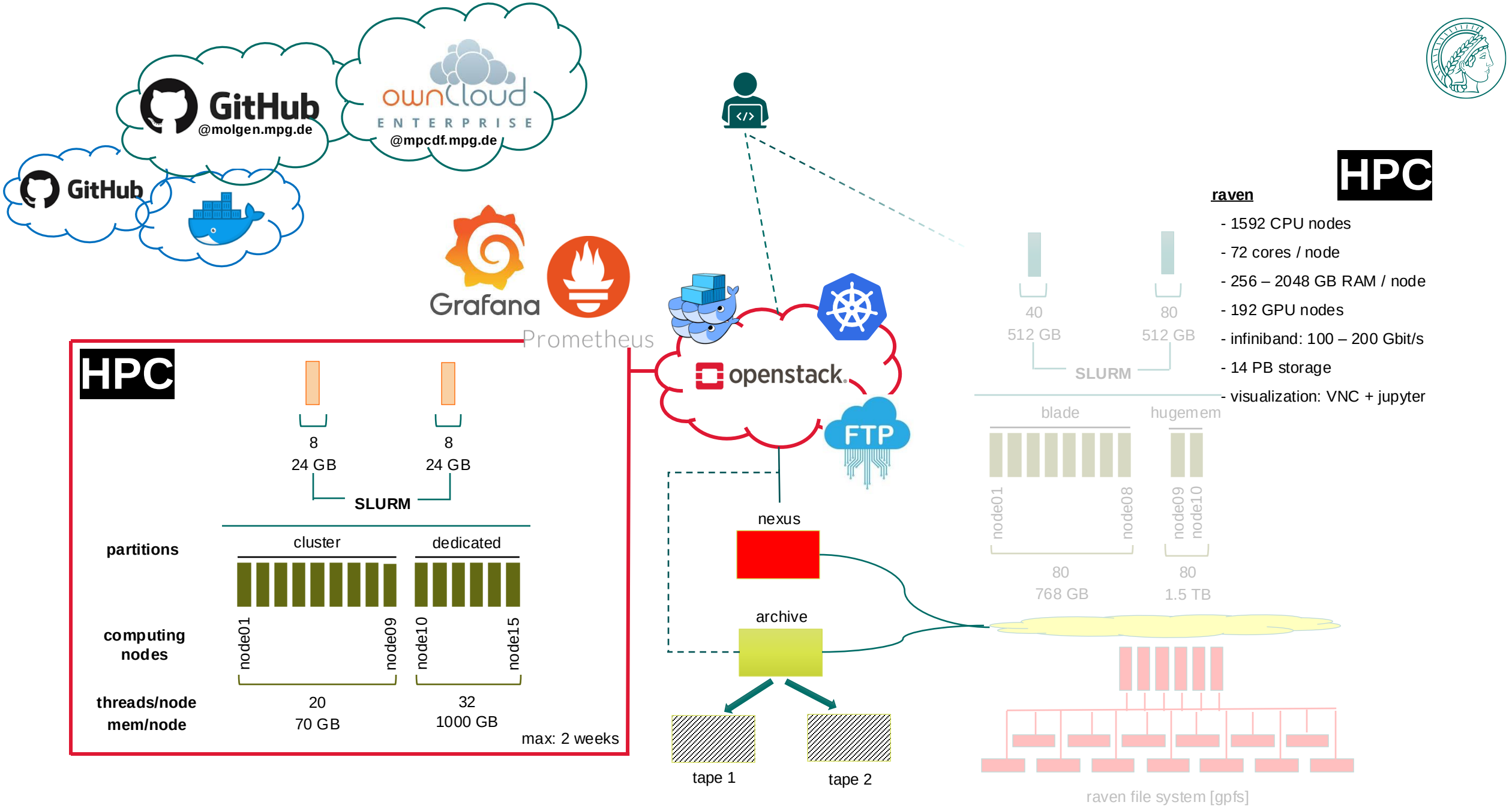
Yun Wang, Ph.D. | Bioinformatician | Tel.: 313

MPCDF

Ayesha Iqbal | Data Scientist | Tel.: 316

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

image sources: https://www.pinterest.com/pin/superman--213217982281863 53/ ; https://www.amazon.in/Festiko-Super-Theme-Cutouts-Multicolour/dp/B0B21K9N71


GitHub
@molgen.mpg.de
ownCloud
ENTERPRISE
@mpcdf.mpg.de
GitHub
Grafana
Prometheus
HPC
8
24 GB
8
24 GB
SLURM
partitions
cluster
dedicated
computing nodes
node01
node09
node10
node15
threads/node
20
32
mem/node
70 GB
1000 GB
max: 2 weeks
openstack.
FTP
nexus
archive
tape 1
tape 2
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
40
512 GB
80
512 GB
SLURM
blade
hugemem
node01
node08
node09
node10
80
768 GB
80
1.5 TB
raven file system [gpfs]