Solutions for Reproducible Research: Data Analytics

Sergio Maffioletti
S3IT: Service and Support for Science IT
Zentrale Informatik
University of Zurich

What is Open Science ?

"Open science is about the way researchers work, collaborate, interact, share resources and disseminate results."

Amsterdam Call for Action on Open Science

Open Science: principles

Open Access	Results
Open Data	Content
Open Workflows	Processes

Reproducible research: Data Analytics

Develop container-based data analytics,

Provide validation procedure and data

Deploy and validate on the supporting infrastructures.

Reproducible research: Data Analytics

Develop container-based data analytics,

Provide validation procedure and data

Deploy and validate on the supporting infrastructures.

Reproducible research: Data Analytics

Develop container-based data analytics,

Provide validation procedure and data

Deploy and validate on the supporting infrastructures.

Reproducible research: Data Analytics

Develop container-based data analytics,

Provide validation procedure and data

Deploy and validate on the supporting infrastructures.

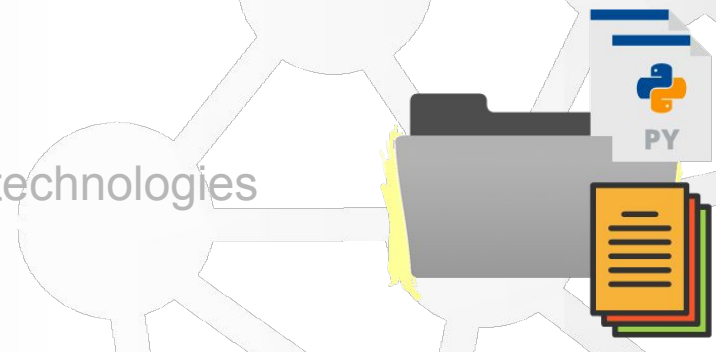
Reproducible research: Data Analytics

Technical challenges

- Heterogeneous infrastructure and technologies

Policy/Human challenges

- New practises for develop and validate reproducibility and portability



Reproducible research: Data Analytics

Technical challenges

- Heterogeneous infrastructure and technologies

Policy/Human challenges

- New practises for develop and validate reproducibility and portability

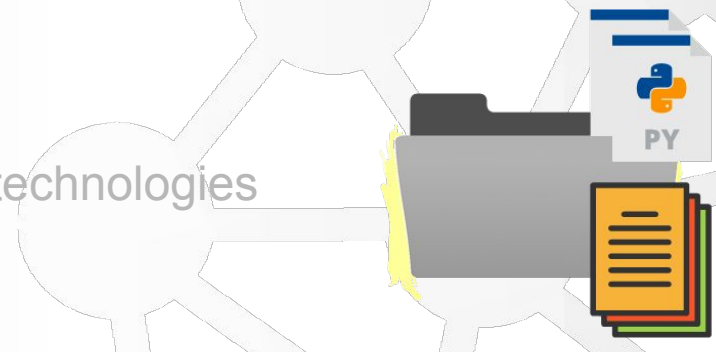
Reproducible research: Data Analytics

Technical challenges

- Heterogeneous infrastructure and technologies

Policy/Human challenges

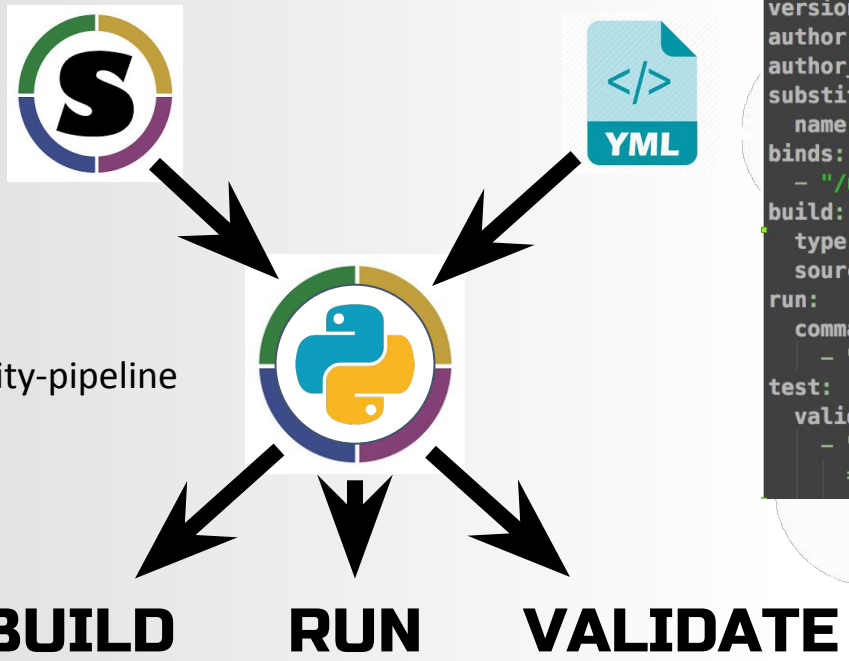
- New practises for develop and validate reproducibility and portability



build, run and test pipelines

data-analysis.img

description.yaml




```
name: hello_world_pipeline
version: 1
author: Balazs Laurenczy
author_org: ETHZ
substitutions:
  name: "hello_world_pipeline"
binds:
  - "/usr/lib64:/usr/lib64"
build:
  type: pull
  source: docker://blaurenczy/hello_world:latest
run:
  commands:
    - "{exec} python3 test.py > {name}.out 2> {name}.err"
test:
  validate_commands:
    - "[[ \"$(md5sum {name}.out | cut -f1 -d ' ')\" \
      == \"06c19b37d27dfda293492a4459fe3bc3\" ]] && echo 0"
```

https://github.com/kav2k/singularity-pipeline

Focus on Reproducible research

- Execution environment AND data analytics workflows in containers
- Documented what's for, what goes in, what comes out
- Container's validator
  - reference dataset + expected results
- Registry of supported containers and certified providers (soon available)

Focus on Reproducible research

- Execution environment AND data analytics workflows in containers
- Documented what's for, what goes in, what comes out
- Container's validator
  - reference dataset + expected results
- Registry of supported containers and certified providers (soon available)

Focus on Reproducible research

- Execution environment AND data analytics workflows in containers
- Documented what's for, what goes in, what comes out
- Container's validator
  - reference dataset + expected results
- Registry of supported containers and certified providers (soon available)

Focus on Reproducible research

- Execution environment AND data analytics workflows in containers
- Documented what's for, what goes in, what comes out
- Container's validator
  - reference dataset + expected results
- Registry of supported containers and certified providers (soon available)

Focus on Reproducible research

- Execution environment AND data analytics workflows in containers
- Documented what's for, what goes in, what comes out
- Container's validator
  - reference dataset + expected results
- Registry of supported containers and certified providers (soon available)

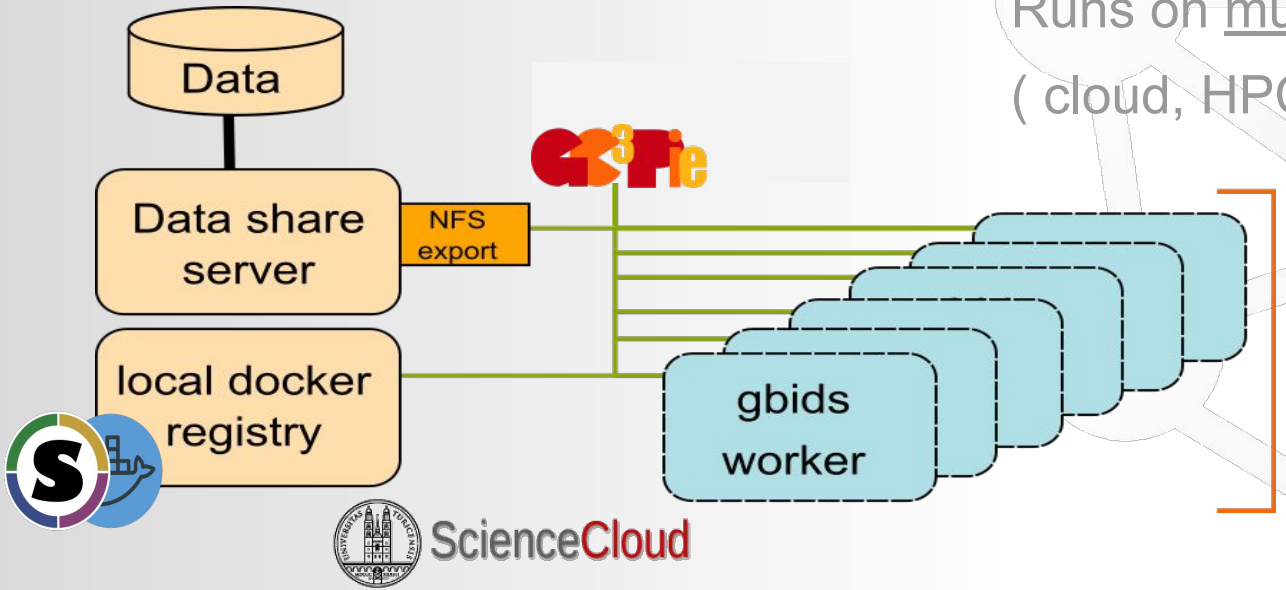
Brain imaging analysis at scale with GC3Pie

http://gc3pie.readthedocs.io

Support BIDS data format and apps

Runs on multiple Infrastructures

(cloud, HPC, local server, ...)

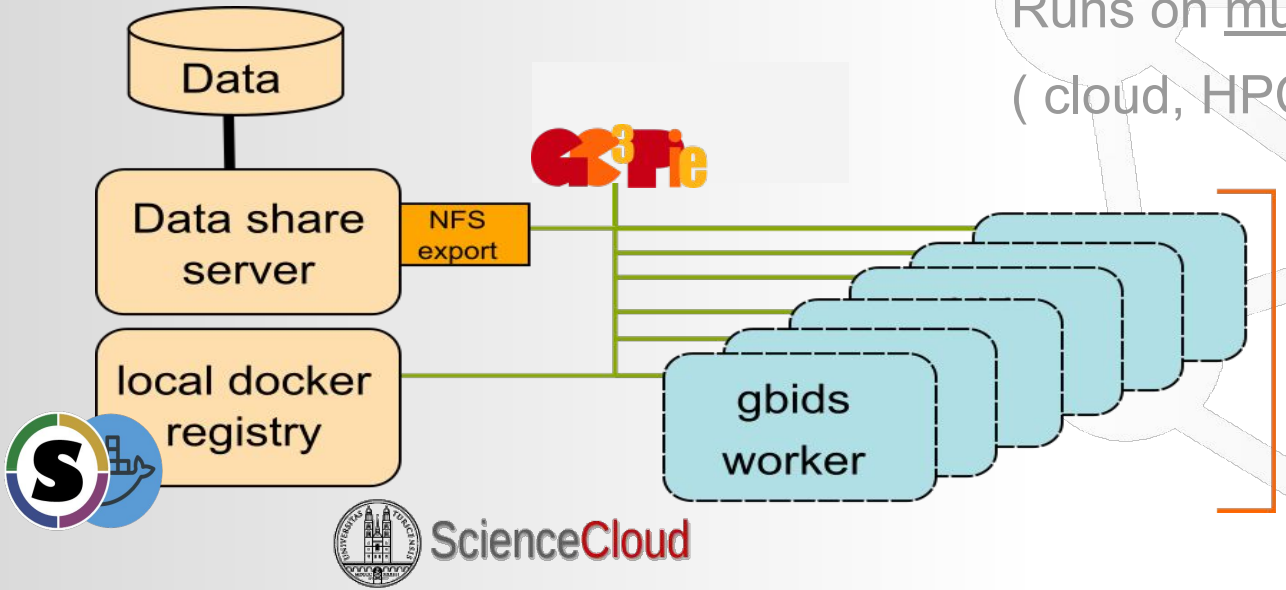


Autoscale on subjects

Brain imaging analysis at scale with GC3Pie

http://gc3pie.readthedocs.io

Support BIDS data format and apps



Runs on multiple Infrastructures

(cloud, HPC, local server, ...)

Autoscale on subjects

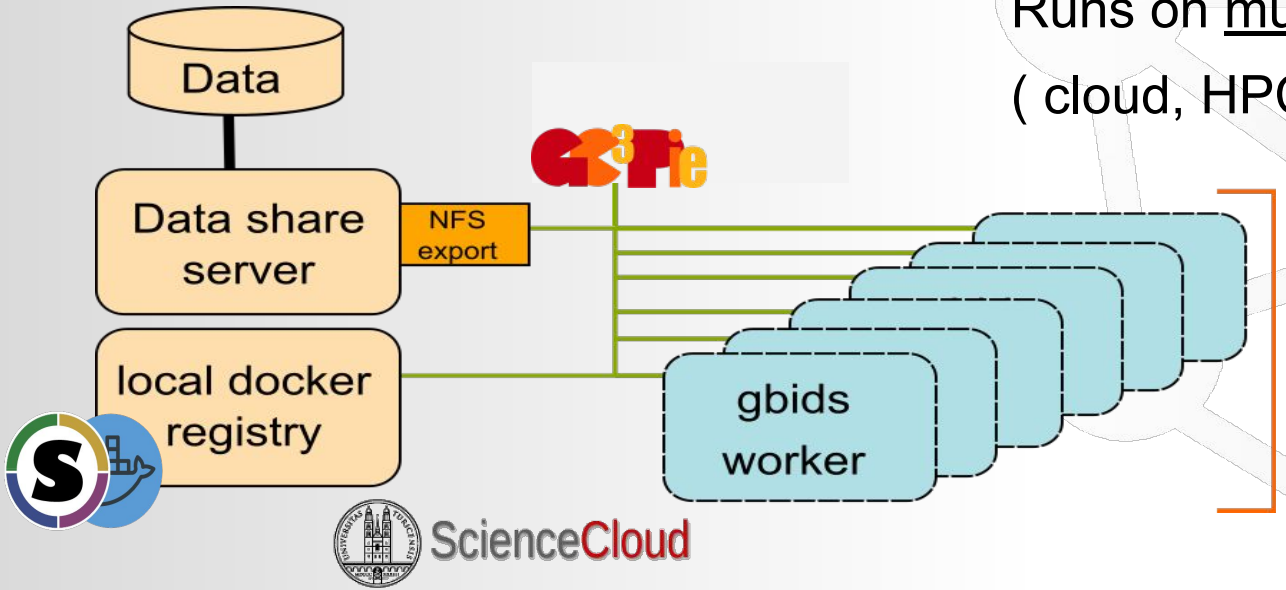
Brain imaging analysis at scale with GC3Pie

http://gc3pie.readthedocs.io

Support BIDS data format and apps

Runs on multiple Infrastructures

(cloud, HPC, local server, ...)



Autoscale on subjects

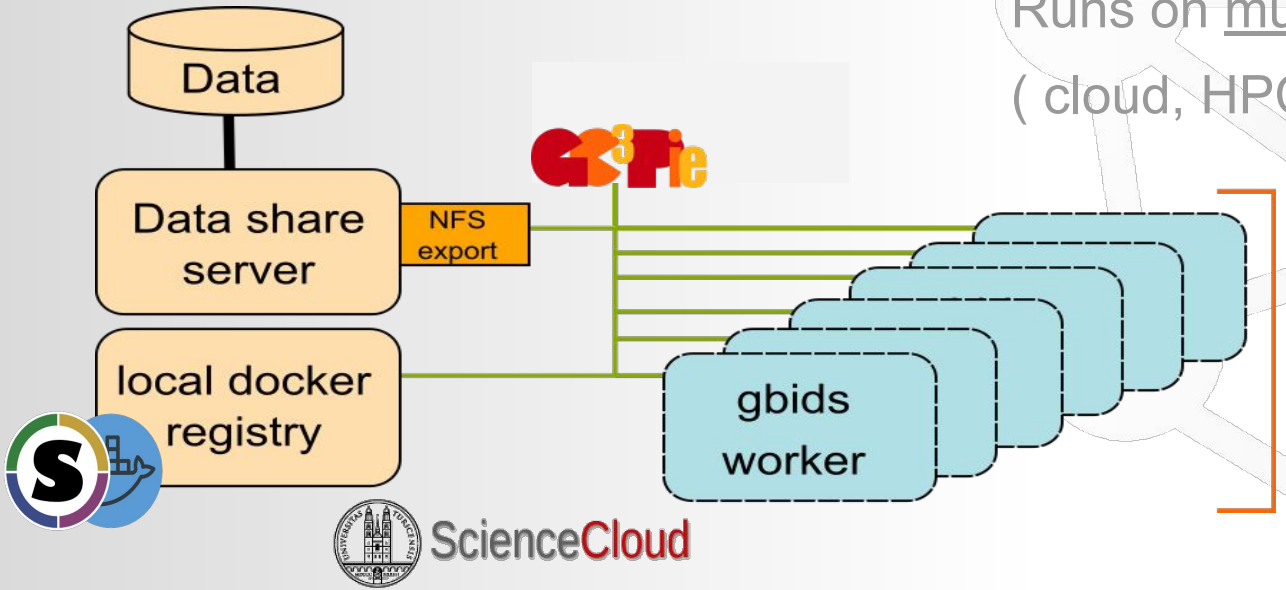
Brain imaging analysis at scale with GC3Pie

http://gc3pie.readthedocs.io

Support BIDS data format and apps

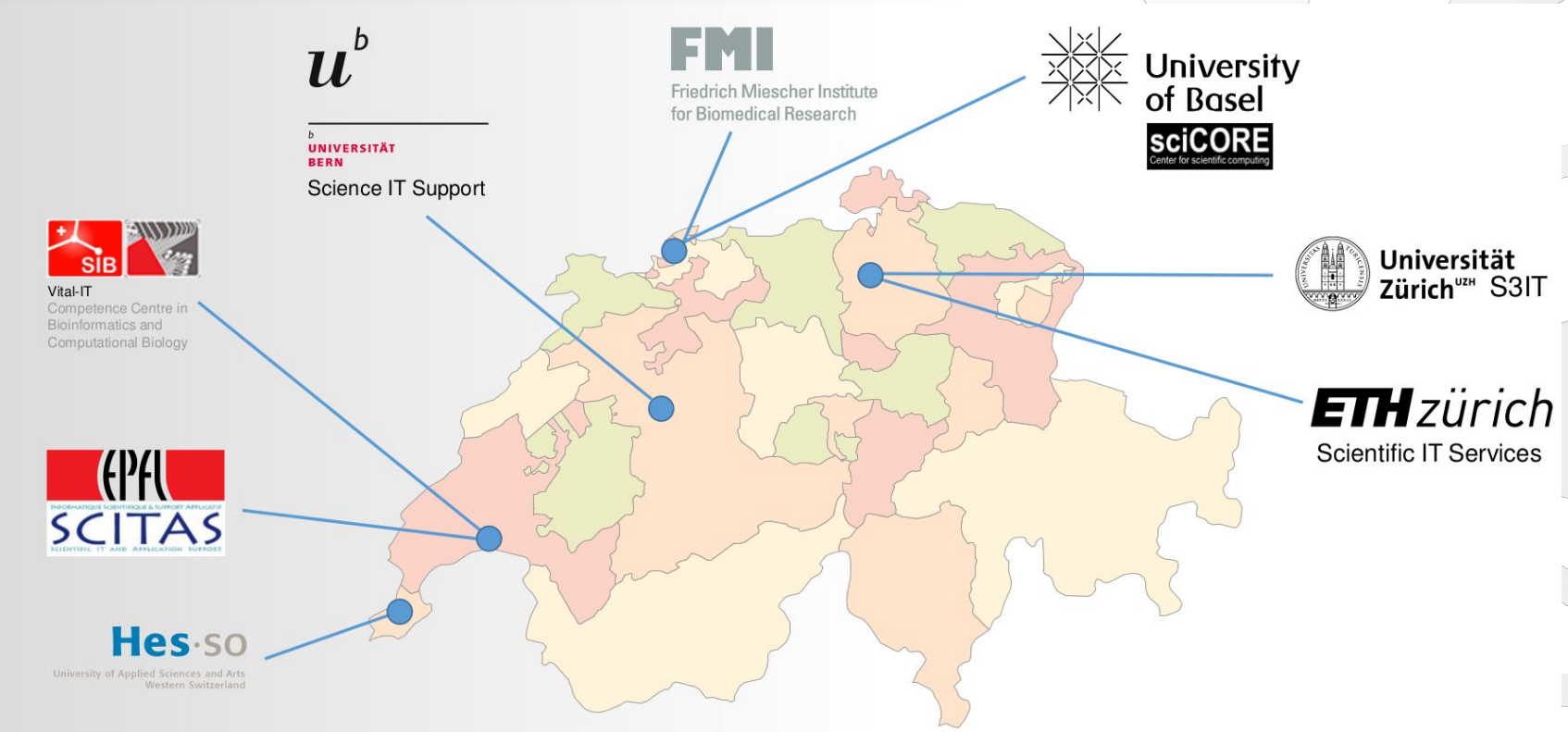
Runs on multiple Infrastructures

(cloud, HPC, local server, ...)



Autoscale on subjects

EnhanceR: enhancing Research through IT expertise



https://www.enhancer.ch

swissuniversities