NHR-NORD@GÖTTINGEN

Workflow Tools

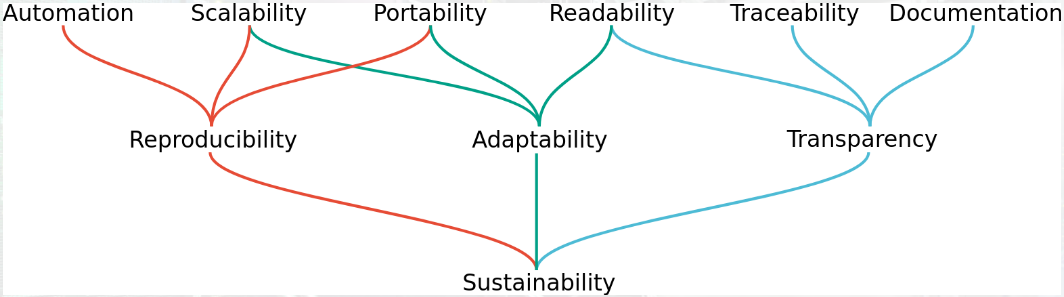
Snakemake, Galaxy, Aiida

Martin Paleico
martin-leandro.paleico@gwdg.de

June 11, 2024

hpc@gwdg.de
GWDG – Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Why Workflows?



(From: "Sustainable data analysis with Snakemake", 2021, official Snakemake Paper; and Snakemake Teaching Alliance repo)

More Workflows?

- Snakemake paper has a good overview and categorization of different workflow approaches: `f1000research.com/articles/10-33/v2`
- Five ”niches”, three today:
  - Snakemake → Domain Specific Language
  - Galaxy → GUI based
  - Aiida → Gneric Programming Language based (Python)
- `github.com/pditommaso/awesome-pipeline`

Note: Snakemake and Galaxy are popular for bioinformatics, AiiDA for materials science and simulations, but any tool can in principle be used for any topic

Snakemake

Concept

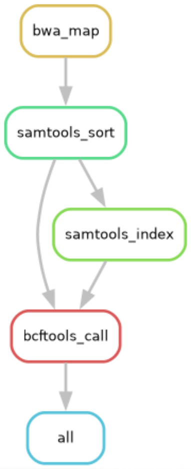
- Snakefile scripts with defined tasks, inputs and outputs
- Snakemake identifies order of tasks and which tasks need to be rerun (*a lá* make)

Example

```
SAMPLES = ["A", "B"]

rule all:
    input:
        "calls/all.vcf"

rule bwa_map:
    input:
        "data/genome.fa",
        "data/samples/{sample}.fastq"
    output:
        "mapped_reads/{sample}.bam"
    shell:
        "bwa mem {input} | samtools view -Sb - > {output}"
# more rules follow
```

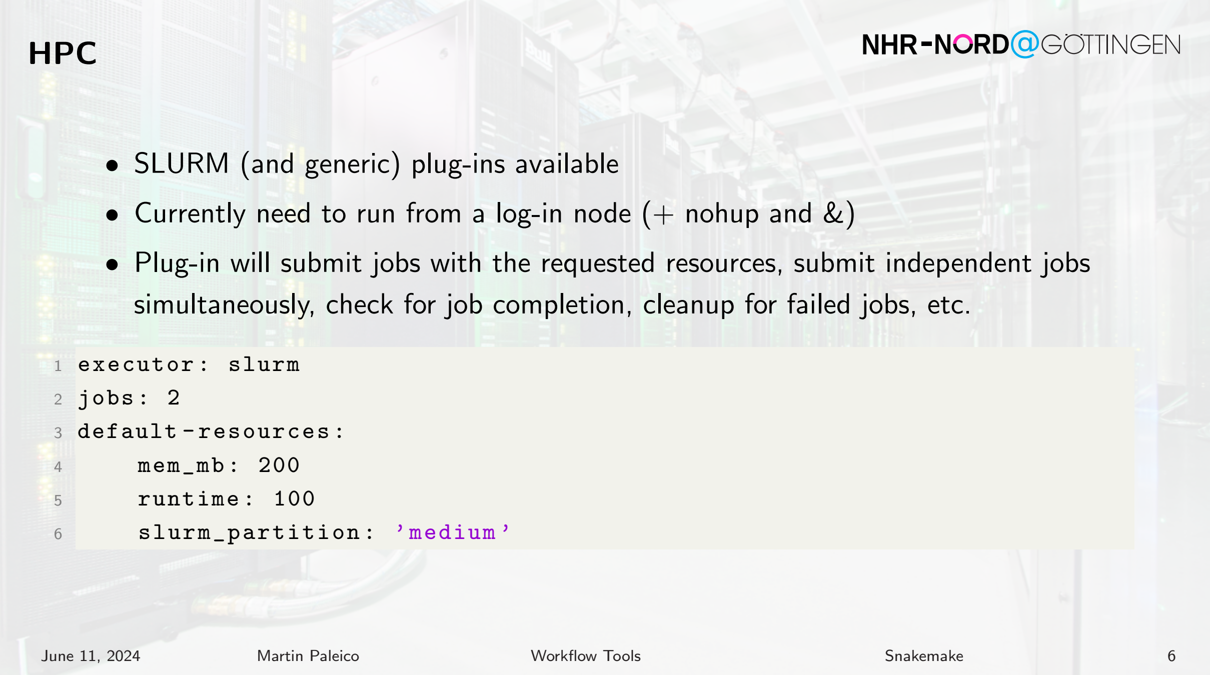
- inputs and outputs are ”promises/placeholders”
- Snakemake uses the chain of IOs to reconstruct what jobs to run and in which order
- {wildcards} are placeholders for text, variables, file names, etc.
- notice the Python syntax!

Directed Acyclic Graph



- Good for understanding the flow and getting an overview of your job
- Useful for debugging
- Nice for publications

HPC



- SLURM (and generic) plug-ins available
- Currently need to run from a log-in node (+ nohup and &)
- Plug-in will submit jobs with the requested resources, submit independent jobs simultaneously, check for job completion, cleanup for failed jobs, etc.

```
executor: slurm
jobs: 2
default-resources:
    mem_mb: 200
    runtime: 100
    slurm_partition: 'medium'
```

HPC

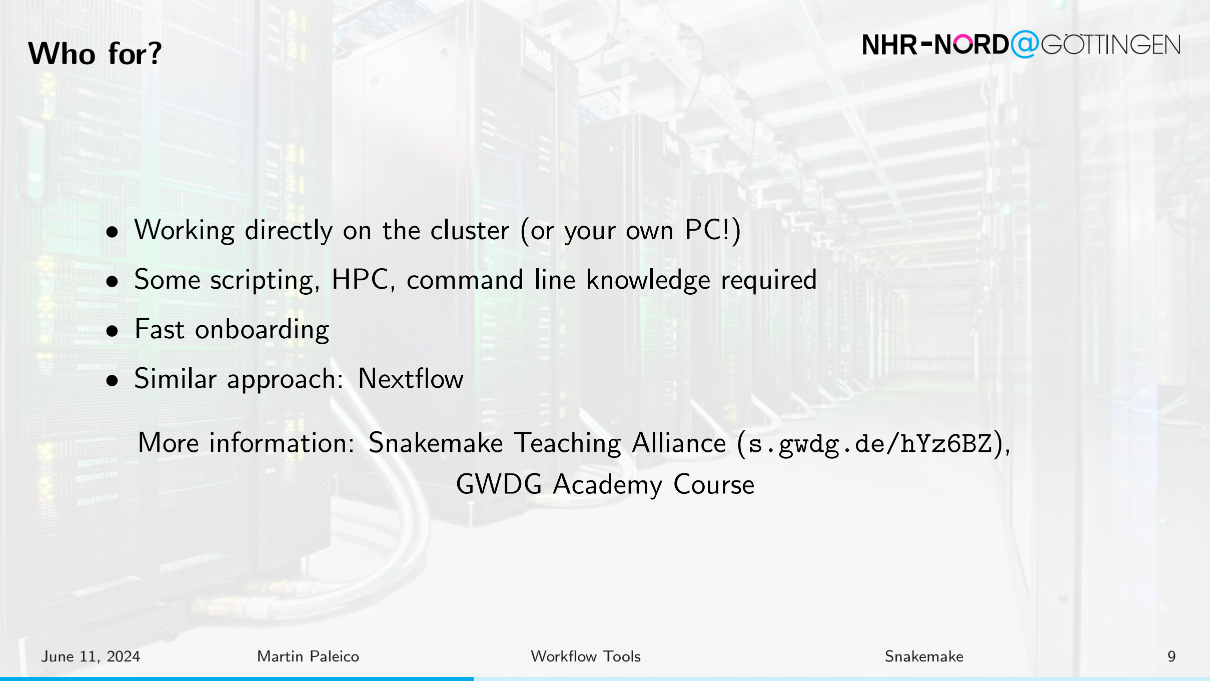
```
> snakemake --executor slurm --default-resources runtime=60 -j 5 -F --snakefile 05_Snakefile

Building DAG of jobs...
SLURM run ID: 09c94694-7362-4fe3-a8c5-b5c5c1eeb2f3
Using shell: /usr/bin/bash
Provided remote nodes: 5
Job stats:
job              count
---------------  -------
all                    1
bcftools_call          1
bwa_map                2
samtools_index         2
samtools_sort          2
total                  8

Select jobs to execute...
Execute 2 jobs...

[Mon Jun 10 12:08:11 2024]
rule bwa_map:
    input: data/genome.fa, data/samples/B.fastq
    output: mapped_reads/B.bam
    jobid: 5
    reason: Forced execution
    wildcards: sample=B
    resources: mem_mb=1000, mem_mib=954, disk_mb=1000, disk_mib=954, tmpdir=<TBD>, runtime=60

Job 5 has been submitted with SLURM jobid 247894 (log: /path/snakemake-tutorial-data-solution/.snakemake/slurm_logs/rule_bwa_map/B/247894.log).

[Mon Jun 10 12:08:12 2024]
rule bwa_map:
    input: data/genome.fa, data/samples/A.fastq
    output: mapped_reads/A.bam
    jobid: 3
    reason: Forced execution
    wildcards: sample=A
    resources: mem_mb=1000, mem_mib=954, disk_mb=1000, disk_mib=954, tmpdir=<TBD>, runtime=60

Job 3 has been submitted with SLURM jobid 247895 (log: /path/snakemake-tutorial-data-solution/.snakemake/slurm_logs/rule_bwa_map/A/247895.log).

[Mon Jun 10 12:11:32 2024]
Finished job 5.
1 of 8 steps (12%) done
[Mon Jun 10 12:11:32 2024]
Finished job 3.
2 of 8 steps (25%) done
Select jobs to execute...
Execute 2 jobs...

#etc.
```

Pros and Cons

- ✓ Simple to understand
- ✓ Repository of workflows
- ✓ Widely used
- ✓ Plain text → easy to read at a glance
- ✓ Shell commands → easy to reconstruct
- ✓ Easy to install
- ✓ Easy to start with
- ✓ Many other features: mark temp and permanent files, mark local non-node tasks, various levels of config files, per task resource assignment, generate reports, etc.

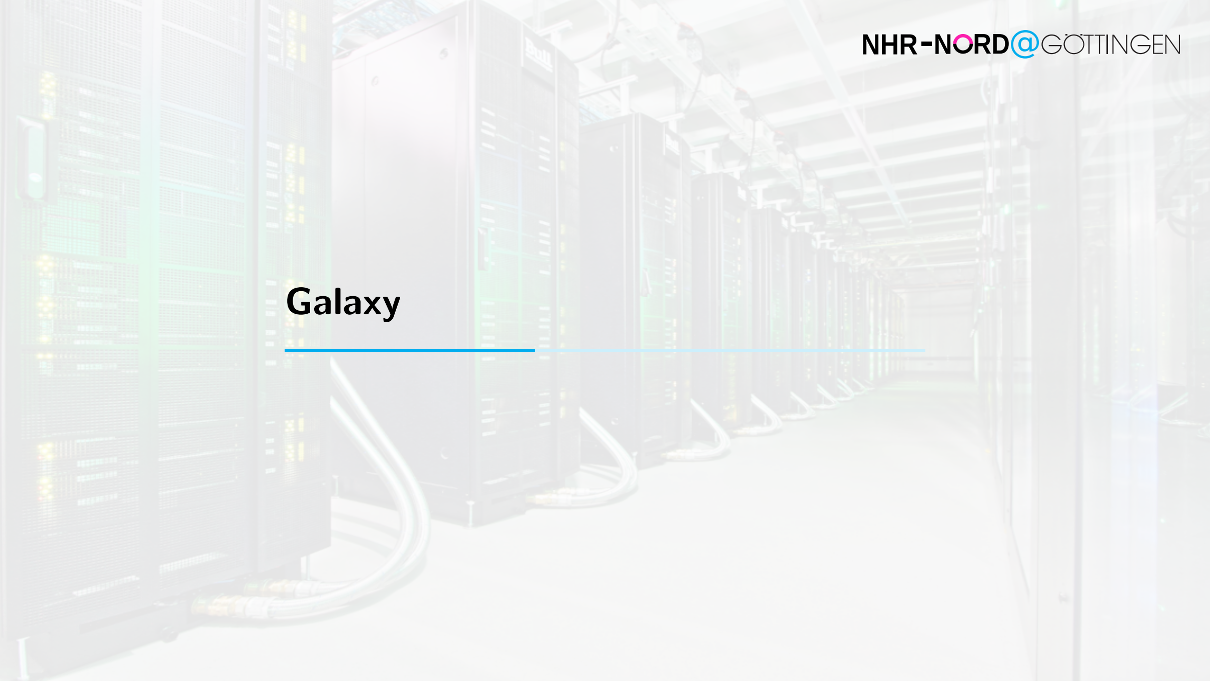
- × Hard to master
- × Confusing formats (Python vs. YAML, : vs =, quote marks or not)
- × Can be hard to debug

Who for?



- Working directly on the cluster (or your own PC!)
- Some scripting, HPC, command line knowledge required
- Fast onboarding
- Similar approach: Nextflow

More information: Snakemake Teaching Alliance (`s.gwdg.de/hYz6BZ`),
GWDG Academy Course



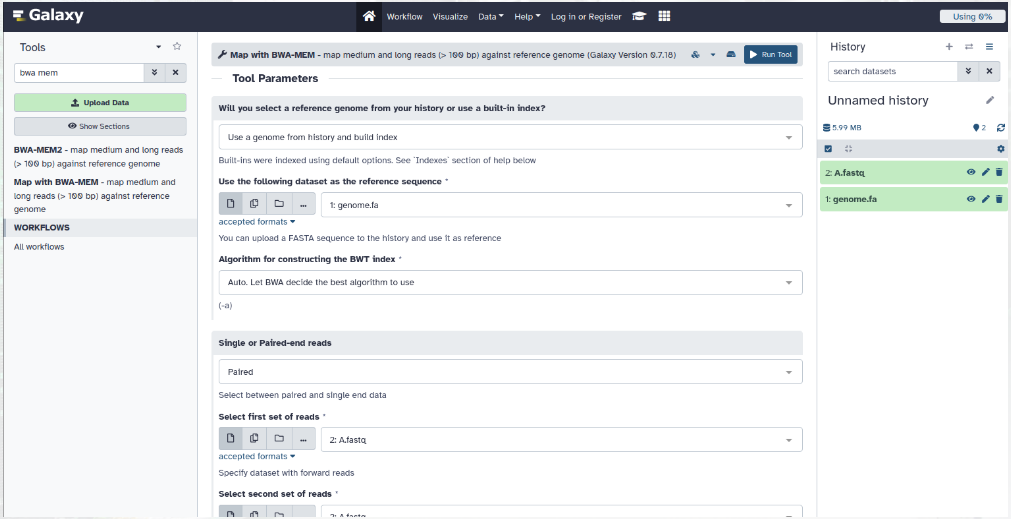
Galaxy

Concept

- Browser, GUI based
- Build a workflow by adding individual tool-based steps on a browser
- Internal DB for files (no on-system files directly)
- Any command line based tool can be turned into a Galaxy tool with a rich XML tool definition language
- Support for user quotas, workflow sharing, executing containers, HPC, etc.

Galaxy PROJECT

Example: Building a history



bwa mem

BWA-MEM2 - map medium and long reads (> 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.18)

Tool Parameters

Will you select a reference genome from your history or use a built-in index?

Use a genome from history and build index

Built-ins were indexed using default options. See `Indexes` section of help below

Use the following dataset as the reference sequence *

1: genome.fa

accepted formats

You can upload a FASTA sequence to the history and use it as reference

Algorithm for constructing the BWT index *

Auto. Let BWA decide the best algorithm to use

(-a)

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads *

2: A.fastq

accepted formats

Specify dataset with forward reads

Select second set of reads *

2: A.fastq

Unnamed history

5.99 MB

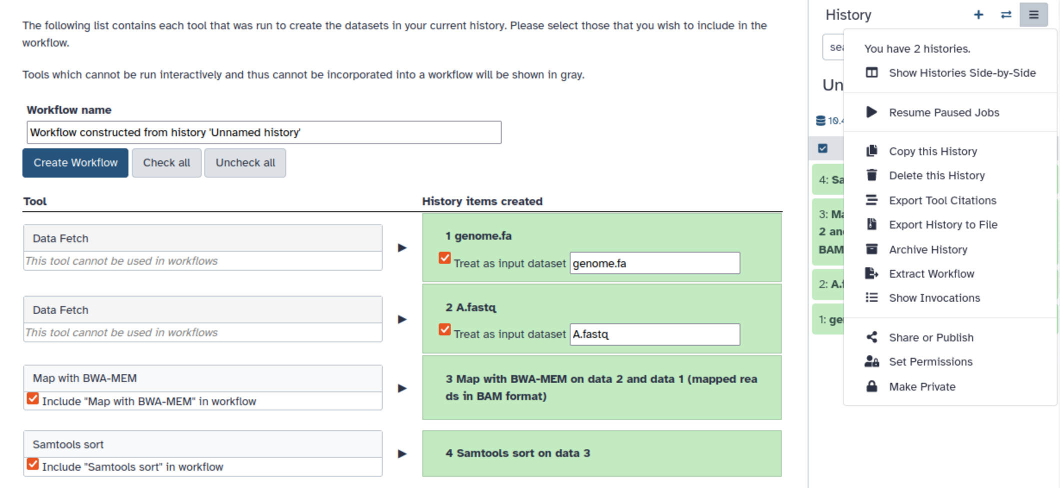
2

2: A.fastq

1: genome.fa

Example: Exporting to a workflow

NHR-NORD@GÖTTINGEN



The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

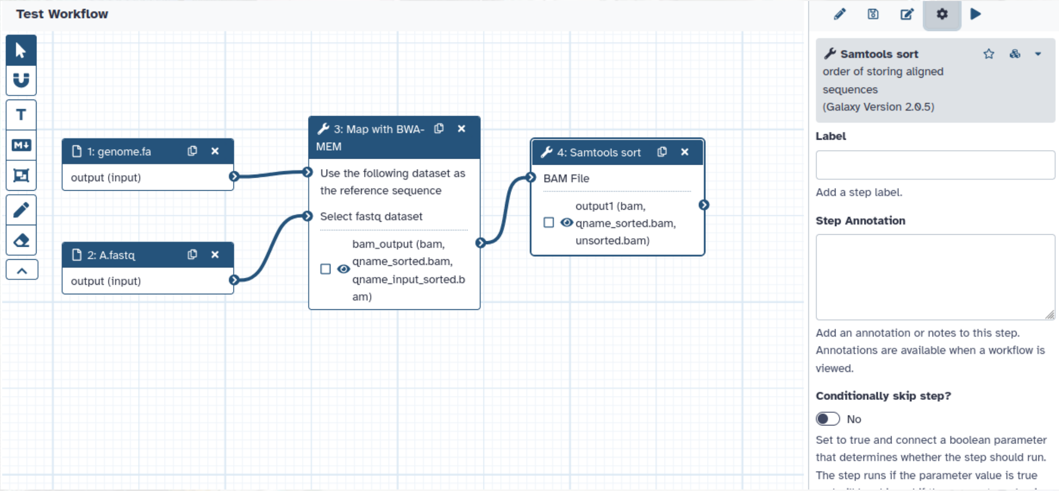
Workflow name

Workflow constructed from history 'Unnamed history'

Tool	History items created
Data Fetch *This tool cannot be used in workflows*	**1 genome.fa** ☑ Treat as input dataset genome.fa
Data Fetch *This tool cannot be used in workflows*	**2 A.fastq** ☑ Treat as input dataset A.fastq
Map with BWA-MEM ☑ Include "Map with BWA-MEM" in workflow	**3 Map with BWA-MEM on data 2 and data 1 (mapped reads in BAM format)**
Samtools sort ☑ Include "Samtools sort" in workflow	**4 Samtools sort on data 3**

You have 2 histories.

Example: Editing a Workflow


Test Workflow
1: genome.fa
output (input)
2: A.fastq
output (input)
3: Map with BWA-MEM
Use the following dataset as the reference sequence
Select fastq dataset
bam_output (bam, qname_sorted.bam, qname_input_sorted.bam)
4: Samtools sort
BAM File
output1 (bam, qname_sorted.bam, unsorted.bam)

Samtools sort order of storing aligned sequences (Galaxy Version 2.0.5)

Label

Add a step label.

Step Annotation

Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

Conditionally skip step?

No

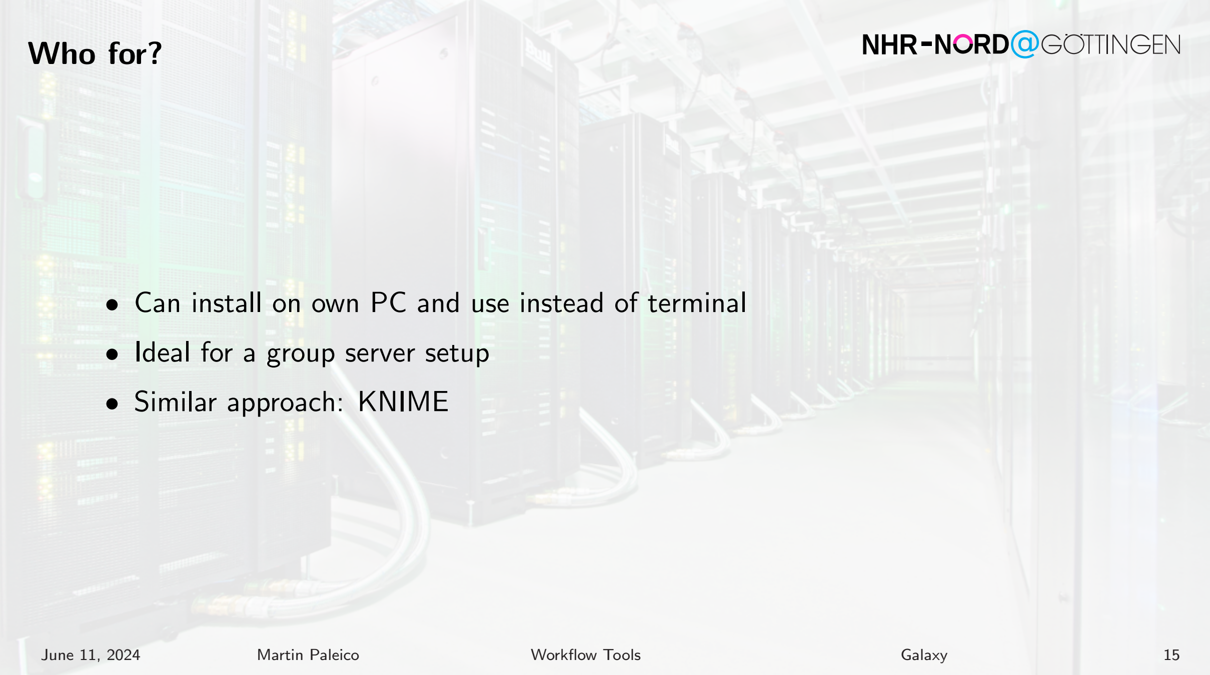
Set to true and connect a boolean parameter that determines whether the step should run. The step runs if the parameter value is true

Pros and Cons

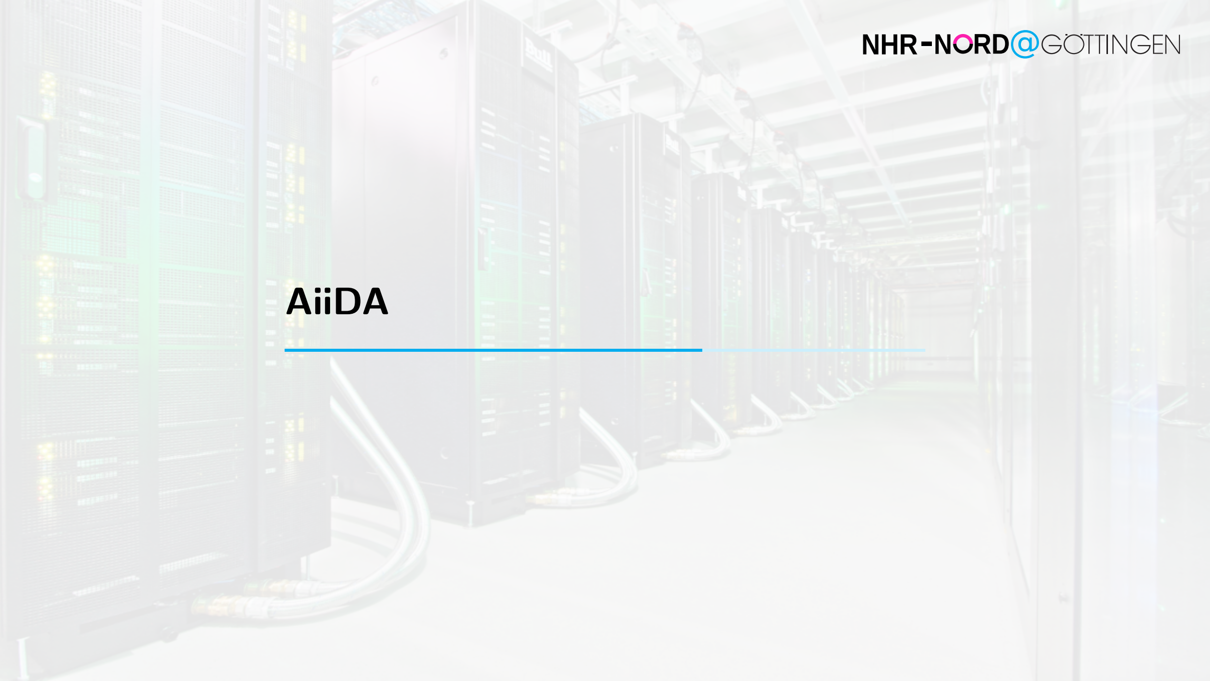
- ✓ GUI
- ✓ Tool repositories
- ✓ Easy to create GUI for your own tool
- ✓ Long development (15+ years)
- ✓ Easy to start with
- ✓ Constellation of Galaxy servers, intercompatible
- ✓ Links for sharing workflows, results, etc.
- ✓ Simple but useful Admin GUI panel

- × GUI
- × "Galaxy" is a very bad name for searches...
- × Getting files to and from server is not so simple
- × Might be hard to reconstruct DB and server if needed...
- × HPC setup possible but not trivial
- × Lots of configuration options
- × Users can consume a lot of storage if not careful (100's of TBs)

Who for?



- Can install on own PC and use instead of terminal
- Ideal for a group server setup
- Similar approach: KNIME



AiiDA

Concept

- Python based
- Works from your local PC or on-server, with its on quasi-scheduler daemon and command line tool (called ”verdi”)
- From there, jobs can run locally or on a remote HPC cluster (or another PC)
- Focused on ”data provenance”: Can query and filter jobs and files, create a single file archive of a group of jobs
- Can pack jobs into a workflow (”work chain”)
- Local DB keeps track of jobs and files
- Use the verdi shell or script directly on Python
- Very complex

Example: verdi shell

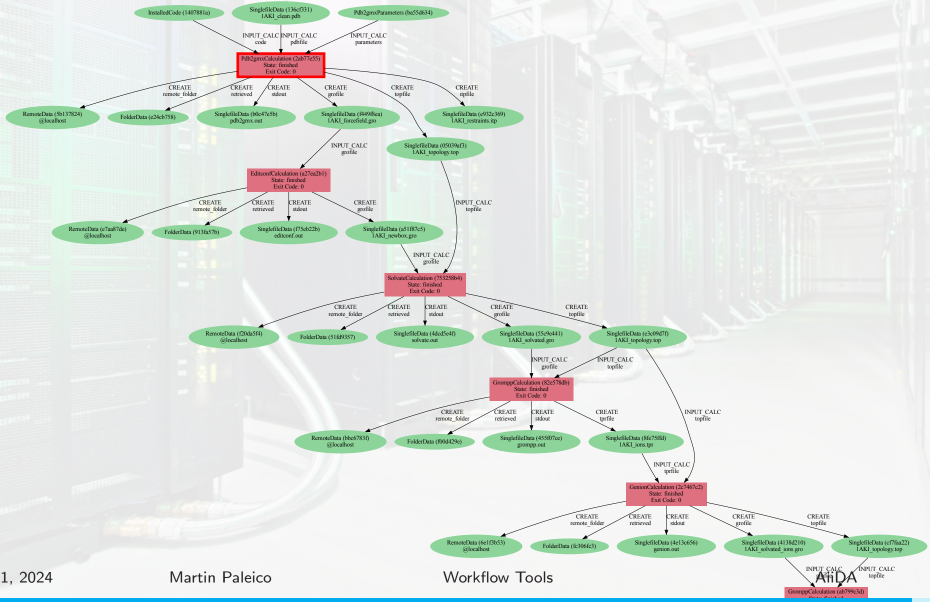
```
# GROMACS plug—in
> gmx_pdb2gmx —f 1AKI_clean.pdb —ff oplsaa —water spce —o 1AKI_forcefield.gro —p 1AKI_topology.top —i 1AKI_restraints.itp

> verdi process list —a
   PK  Created     Process label         Process State     Process status
----  ---------  -------------------  ----------------  -----------------
    4  11s ago     Pdb2gmxCalculation    Finished [0]

Total results: 1

# after more plug—in calls
> verdi process list —a
   PK  Created     Process label          Process State     Process status
----  ---------  --------------------  ----------------  ----------------------------------------
    4  4m ago      Pdb2gmxCalculation     Finished [0]
   12  2m ago      EditconfCalculation    Finished [0]
   18  2m ago      SolvateCalculation     Finished [0]
   26  1m ago      GromppCalculation      Finished [0]
   33  53s ago     GenionCalculation      Finished [0]
   41  9s ago      GromppCalculation      Finished [0]
   47  4s ago      MdrunCalculation       Waiting          Monitoring scheduler: job state RUNNING
Total results: 7
Report: last time an entry changed state: 4s ago (at 12:40:52 on 2024—06—11)
Report: Checking daemon load... OK
Report: Using 0%% of the available daemon worker slots.
```

Example: Inspecting a node

```
> verdi node show 4
Property     Value
-----------  ------------------------------------
type         Pdb2gmxCalculation
state        Finished [0]
pk           4
uuid         2ab77e55-0936-47df-a6cf-9df4619ae831
label
description  record pdb2gmx data provenance via the aiida_gromacs plugin
ctime        2024-06-11 12:36:23.178336+02:00
mtime        2024-06-11 12:36:24.564897+02:00
computer     [1] localhost

Inputs        PK  Type
----------  ----  -----------------
code           1  InstalledCode
parameters     3  Pdb2gmxParameters
pdbfile        2  SinglefileData

Outputs          PK  Type
-------------  ----  --------------
grofile           8  SinglefileData
itpfile          10  SinglefileData
remote_folder     5  RemoteData
retrieved         6  FolderData
stdout            7  SinglefileData
topfile           9  SinglefileData
```

Everything has its own PK (primary key), UUID, and label

Example: Provenance Graph


InstalledCode (1407881a)
SinglefileData (136cf331) 1AKI_clean.pdb
Pdb2gmxParameters (ba55d634)
INPUT_CALC code
INPUT_CALC pdbfile
INPUT_CALC parameters
Pdb2gmxCalculation (2ab77e55) State: finished Exit Code: 0
CREATE remote_folder
CREATE retrieved
CREATE stdout
CREATE grofile
CREATE topfile
CREATE itpfile
RemoteData (5b137824) @localhost
FolderData (e24cb758)
SinglefileData (b0c47e5b) pdb2gmx.out
SinglefileData (f449f8ea) 1AKI_forcefield.gro
SinglefileData (e932c369) 1AKI_restraints.itp
SinglefileData (05039af3) 1AKI_topology.top
INPUT_CALC grofile
EditconfCalculation (a27ea2b1) State: finished Exit Code: 0
CREATE remote_folder
CREATE retrieved
CREATE stdout
CREATE grofile
INPUT_CALC topfile
RemoteData (e7aa87de) @localhost
FolderData (913fa57b)
SinglefileData (f75eb22b) editconf.out
SinglefileData (a51f87c5) 1AKI_newbox.gro
INPUT_CALC grofile
SolvateCalculation (753258b4) State: finished Exit Code: 0
CREATE remote_folder
CREATE retrieved
CREATE stdout
CREATE grofile
CREATE topfile
RemoteData (f20da5f4) @localhost
FolderData (51fd9357)
SinglefileData (4dcd5e4f) solvate.out
SinglefileData (55c9e441) 1AKI_solvated.gro
SinglefileData (e3c09d7f) 1AKI_topology.top
INPUT_CALC grofile
INPUT_CALC topfile
GromppCalculation (82e578db) State: finished Exit Code: 0
CREATE remote_folder
CREATE retrieved
CREATE stdout
CREATE tprfile
INPUT_CALC topfile
RemoteData (bbc6783f) @localhost
FolderData (f00d429e)
SinglefileData (455f07ce) grompp.out
SinglefileData (8fe75ffd) 1AKI_ions.tpr
INPUT_CALC tprfile
GenionCalculation (2c7467c2) State: finished Exit Code: 0
CREATE remote_folder
CREATE retrieved
CREATE stdout
CREATE grofile
CREATE topfile
RemoteData (6e1f3b53) @localhost
FolderData (fc306fc3)
SinglefileData (4e13c656) genion.out
SinglefileData (4138d210) 1AKI_solvated_ions.gro
SinglefileData (cf7faa22) 1AKI_topology.top
INPUT_CALC grofile
INPUT_CALC topfile
GromppCalculation (ab799c3d)

HPC

```
> verdi computer setup
Computer label: scc
Hostname: gwdu101.gwdg.de
Transport plugin: core.ssh
Scheduler plugin: core.slurm
Work directory on the computer [/scratch/{username}/aiida/]: /scratch/users/{username}/aiida/
Success: Computer<3> scc created

> verdi -p quicksetup computer configure core.ssh scc
User name []: myusername
Port number [22]:
Look for keys [Y/n]: Y
SSH key file []: ~/.ssh/mysshkey
...
Success: scc successfully configured for test@test.com


> verdi computer test scc
...
Success: all 6 tests succeeded

> verdi code create core.code.installed
Computer: scc
Absolute filepath executable: /opt/sw/rev/23.12/linux-scientific7-cascadelake/gcc-11.4.0/gromacs-2023.3-4ehgu3/bin/gmx_mpi
Label: gmxmpi
Description: Remote GROMACS
Default ‘CalcJob‘ plugin: gromacs
Success: Created InstalledCode<78>

> gmx_pdb2gmx --code gmxmpi@scc ...
```

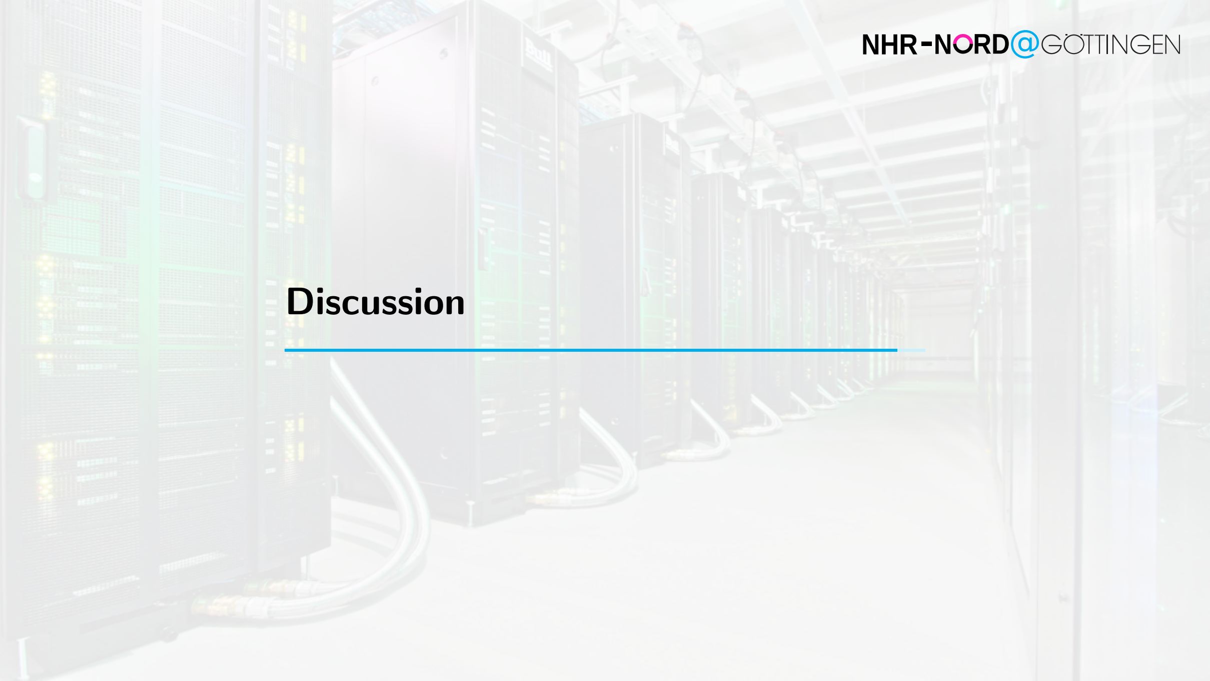
Pros and Cons

- ✓ Extremely flexible
- ✓ Launch jobs remotely
- ✓ Provenance
- ✓ Queryable databases

- × Complex
- × Difficult onboarding
- × Have to set up plug-ins for all your tools, probably
- × Not very popular/active

Who for?

- Advanced users
- Handful of tools
- Long time use
- Similar approach: Covalent `covalent.readthedocs.io`



Discussion

General Discussion

- Takeaways:
  - Many different paradigms
  - Pick the one you like the most
  - Local only or also remote?
  - How easily can I migrate away or retrieve my scripts and data?
  - But keep an eye on practicalities such as onboarding difficulty, availability of support, active development, etc.
- Do you know other unique workflow paradigms?
- Do you know other workflow tools within these example frameworks?
- Would you like to know more about any of these tools (workshop, articles, etc.)?