Special Interest Group Workflow Tools
03.03.2022
CC-41 Community Meeting

Dennis Gläser, Bernd Flemisch, Jörg F. Unger, Philipp Diercks

BETTY'S WISH: FAIR RESEARCH SOFTWARE

- **F** FINDABLE
- **A** ACCESSIBLE
- **I** INTEROPERABLE
- **R** REUSABLE

BETTY'S WISH: FAIR RESEARCH SOFTWARE

- **F** FINDABLE
- **A** ACCESSIBLE
- **I** INTEROPERABLE / PORTABLE
- **R** REUSABLE / REPRODUCIBLE / COMPOSABLE

BETTY'S WISH: FAIR RESEARCH SOFTWARE

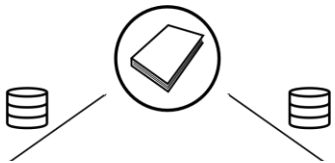
- **F** FINDABLE
- **A** ACCESSIBLE
- **I** INTEROPERABLE / PORTABLE
- **R** REUSABLE / REPRODUCIBLE / COMPOSABLE
- **Q** QUALITY-ASSURED

What is research software?

What is Research Software?

Research Output: scientific publication

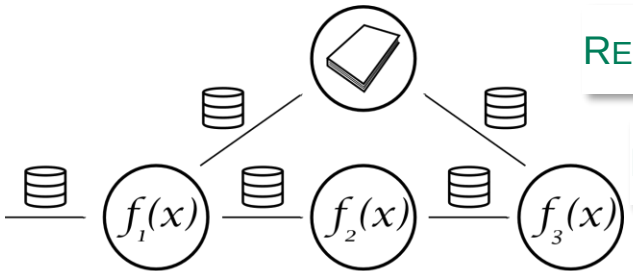
What is Research Software?



Research Output: scientific publication

Based on Generated Data

What is Research Software?

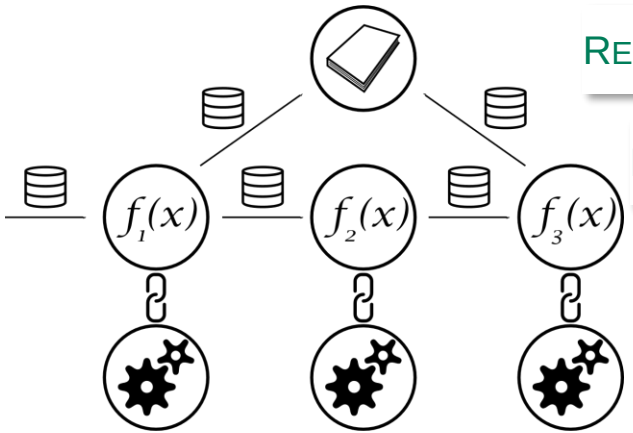


Research Output: scientific publication

Based on Generated Data

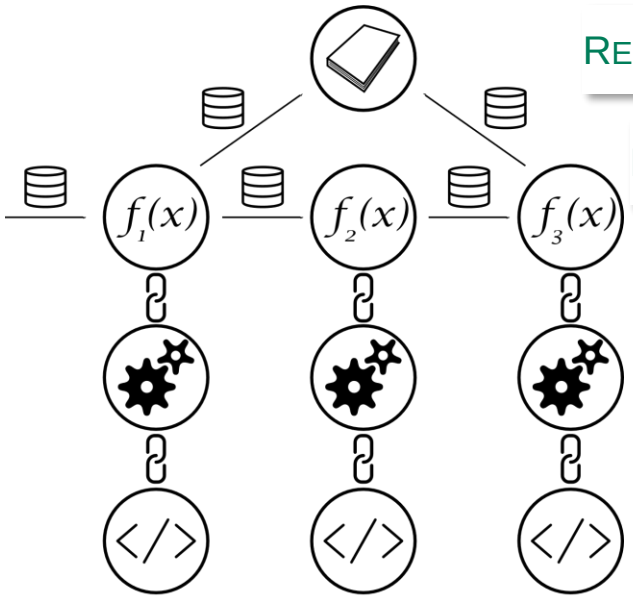
Generated With a Set of Methods

What is Research Software?



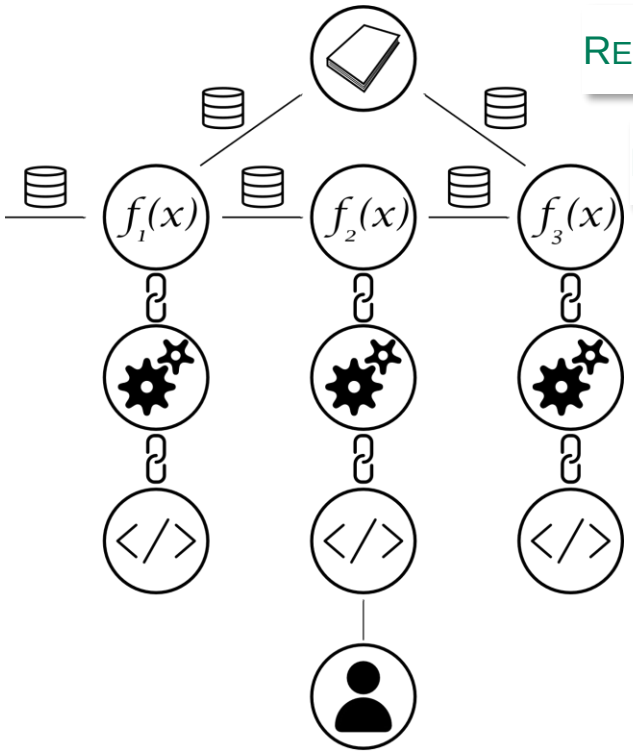
- Research Output: scientific publication
- Based on Generated Data
- Generated With a Set of Methods
- Represented in executable Programs

What is Research Software?



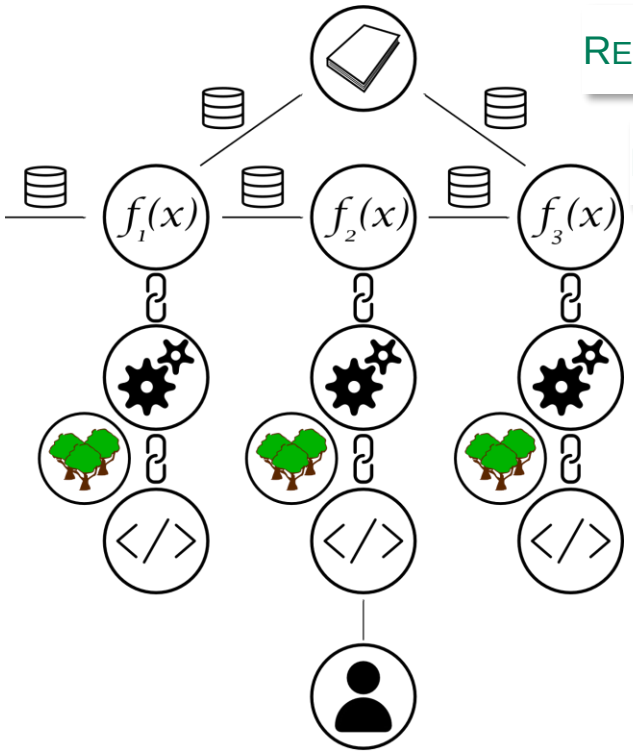
- Research Output: scientific publication
- Based on Generated Data
- Generated With a Set of Methods
- Represented in executable Programs
- Compiled from Source Code

What is Research Software?



- Research Output: scientific publication
- Based on Generated Data
- Generated With a Set of Methods
- Represented in executable Programs
- Compiled from Source Code
- Developed by Researchers

What is Research Software?



Research Output: scientific publication

Based on Generated Data

Generated With a Set of Methods

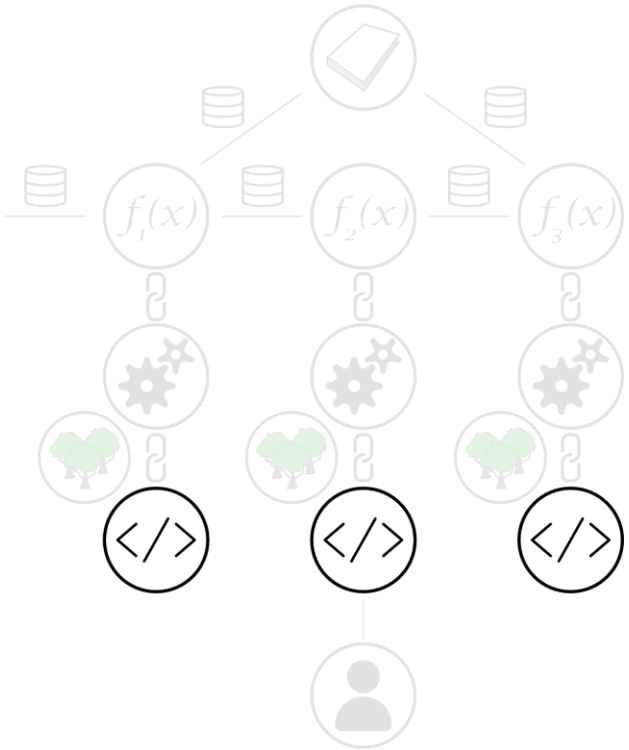
Represented in executable Programs

Compiled from Source Code

Posing Requirements on the Compute environment

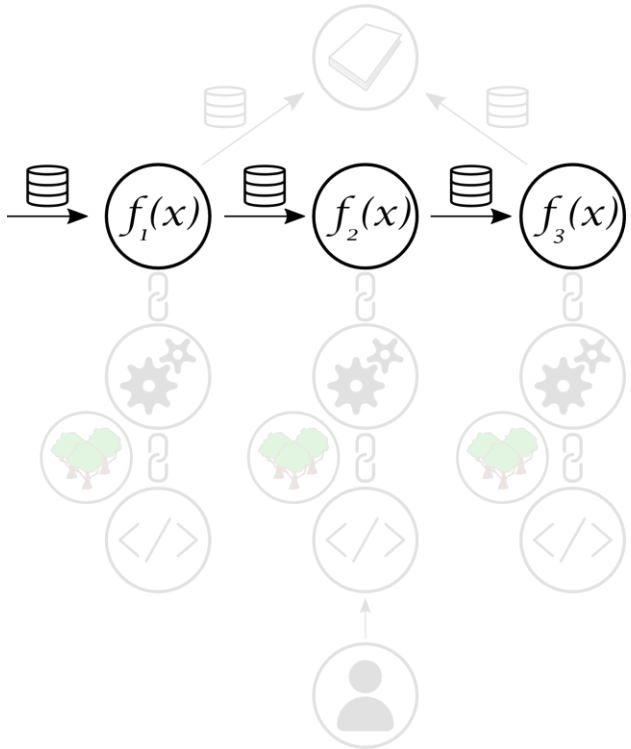
Developed by Researchers

What is Research Software?



Implementation of concrete “methods”

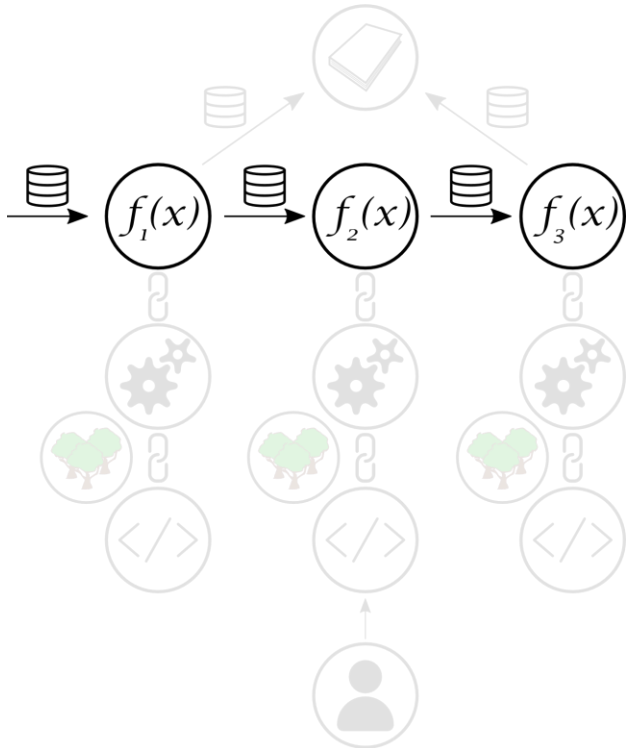
What is Research Software?



Orchestration of the “methods” to address research question

Implementation of concrete “methods”

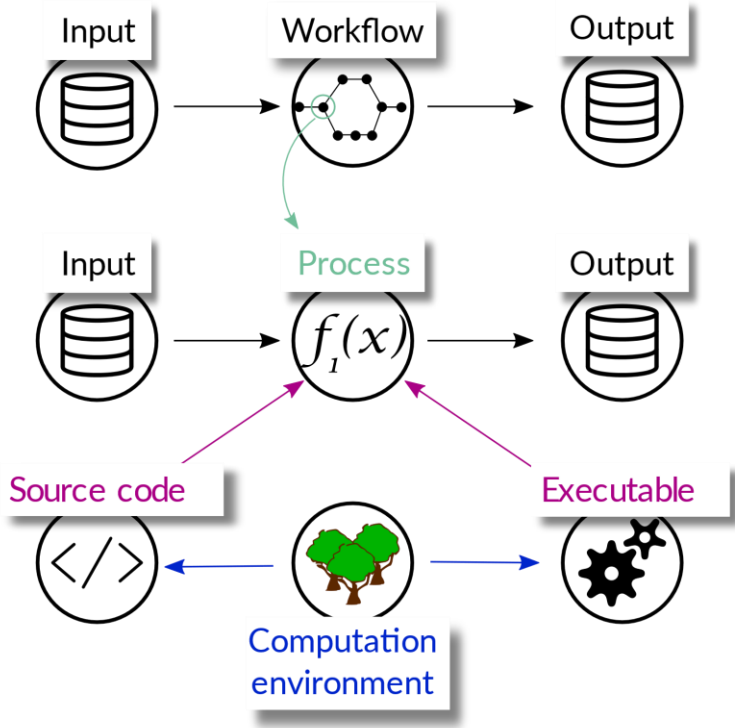
WHAT IS RESEARCH SOFTWARE?



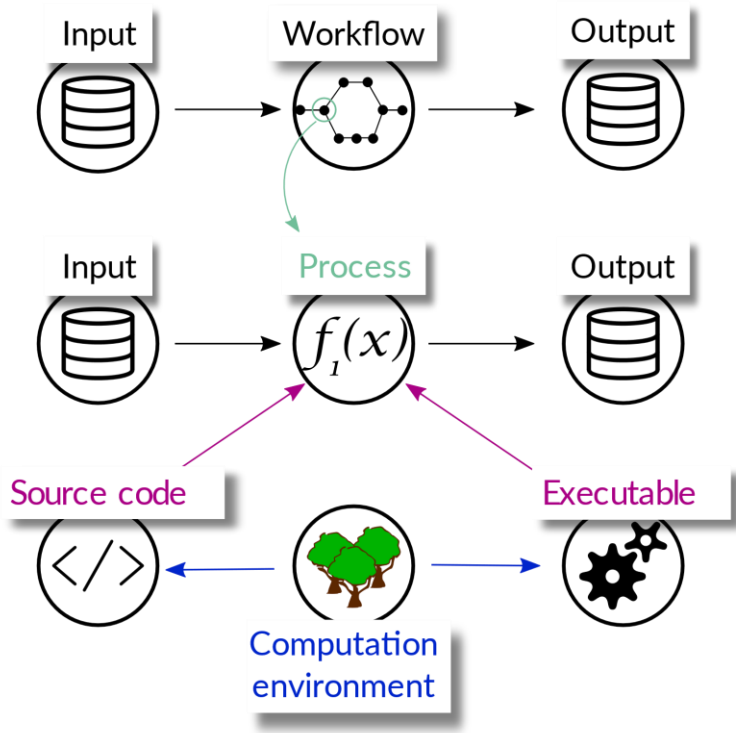
WORKFLOW
TO ADDRESS RESEARCH QUESTION

IMPLEMENTATION OF CONCRETE
“METHODS”

What do we mean by research workflows?



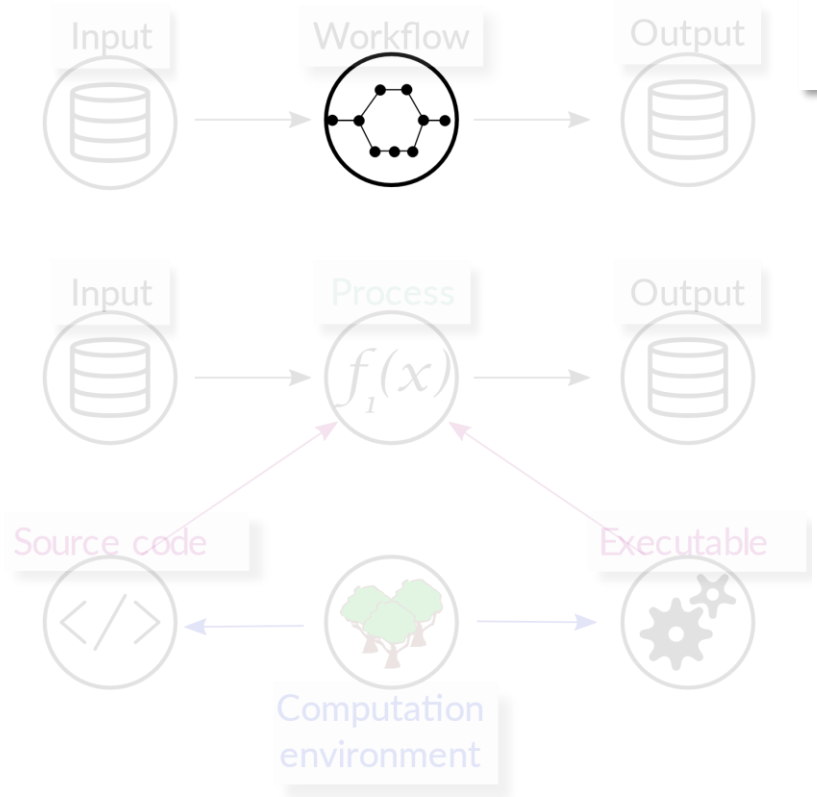
WHY USE A WORKFLOW TOOL?

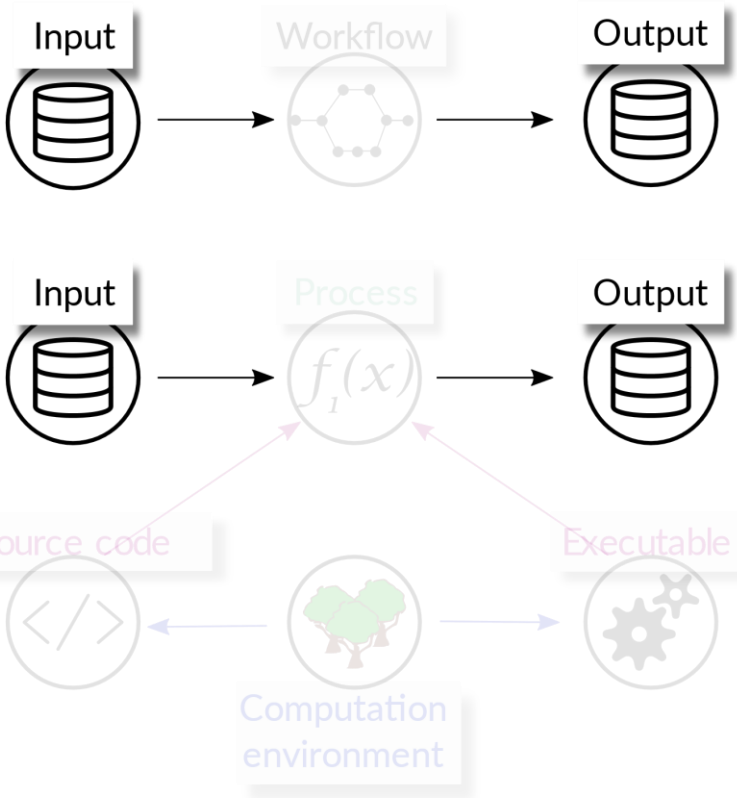


- AVOIDS MANUAL STEPS
- HIGH-LEVEL & MACHINE-READABLE DOCUMENTATION
- MAY HANDLE ENVIRONMENT INSTANTIATION
- MAY BE REUSED IN OTHER WORKFLOWS
- ADDITIONAL WORK (INITIALLY)
- YET ANOTHER TOOL TO LEARN

What requirements may computational research workflows pose on workflow tools?

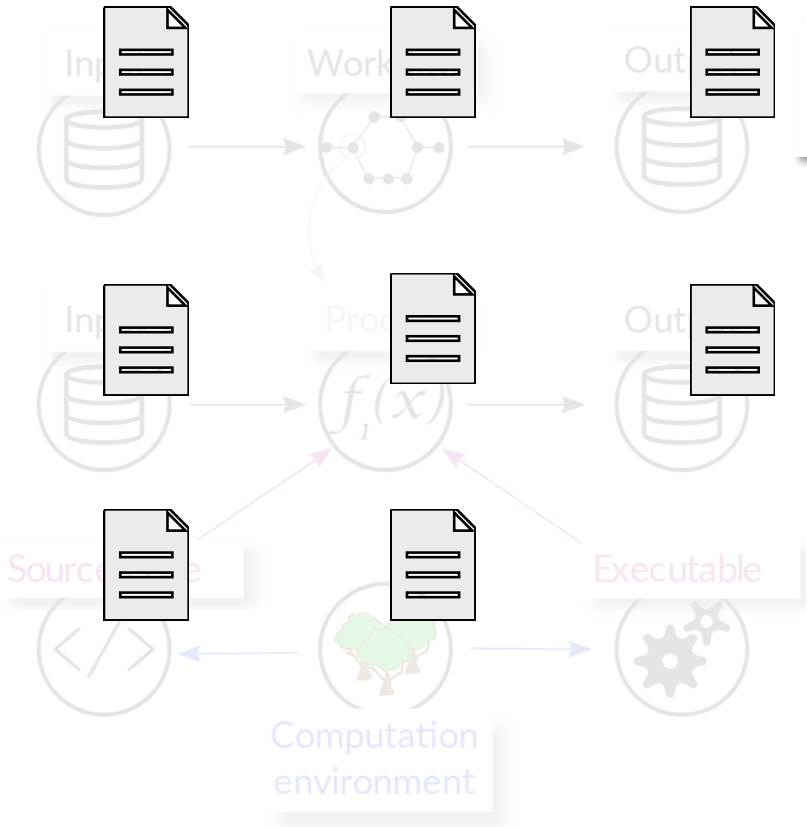
Execution / scheduling / monitoring





Execution / scheduling / monitoring

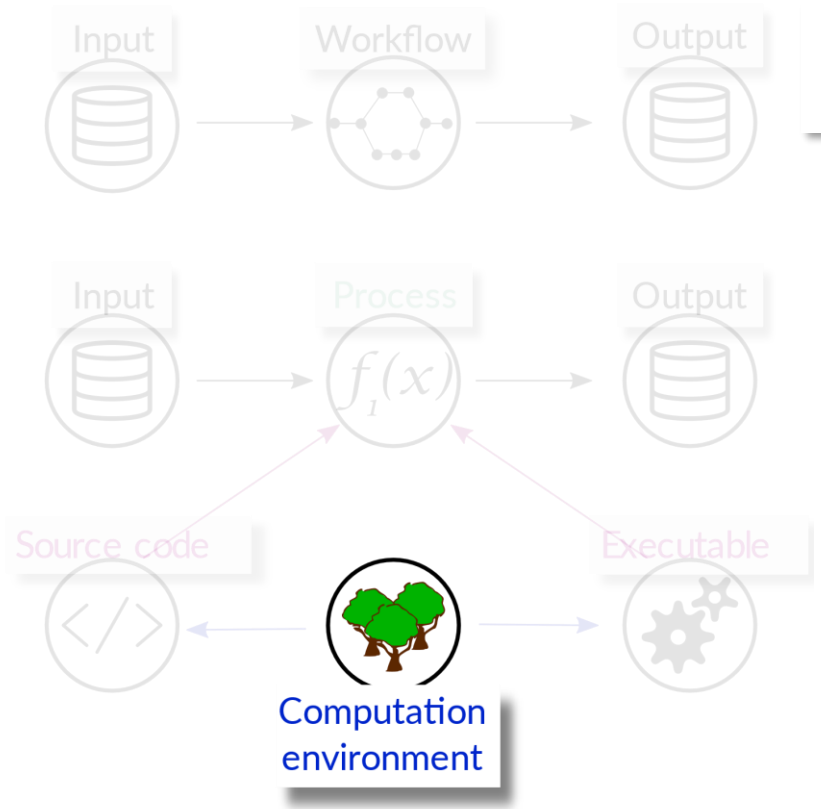
Data provenance tracking



Execution / scheduling / monitoring

Data provenance tracking

Metadata

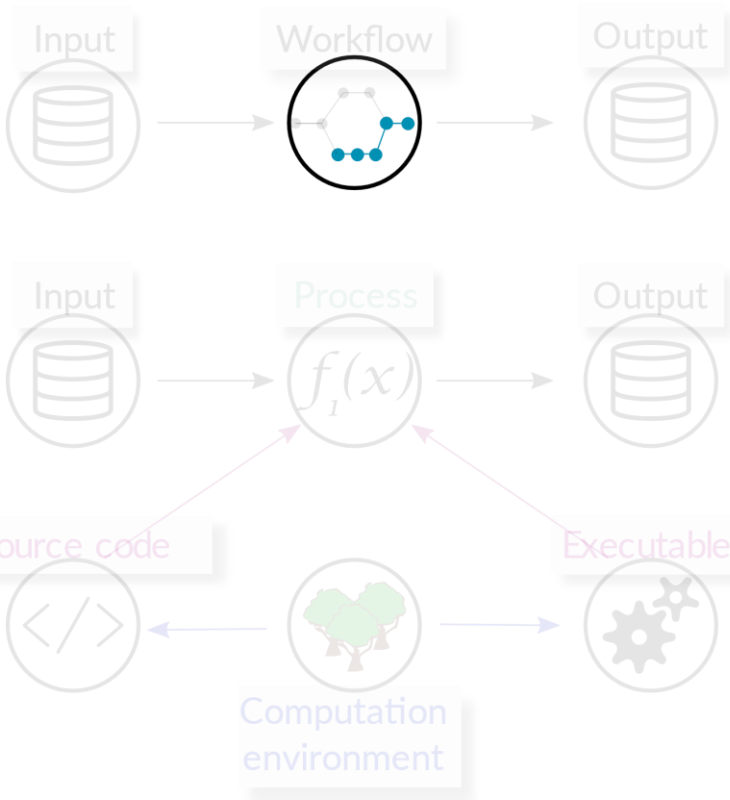


Execution / scheduling / monitoring

Data provenance tracking

Metadata

Compute environment



Execution / scheduling / monitoring

Data provenance tracking

Metadata

Up-to-dateness

Compute environment

NFDi4ing

Execution / scheduling / monitoring

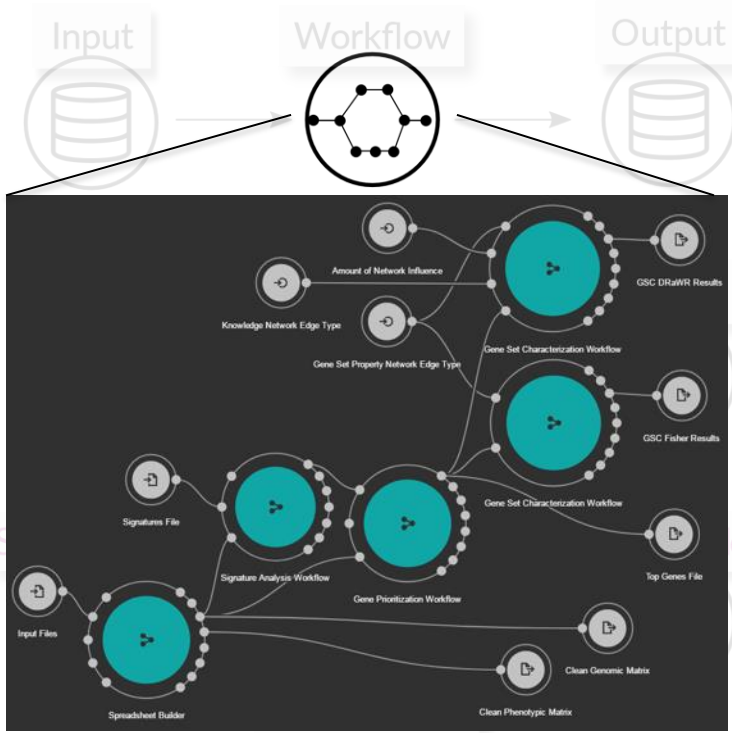
Data provenance tracking

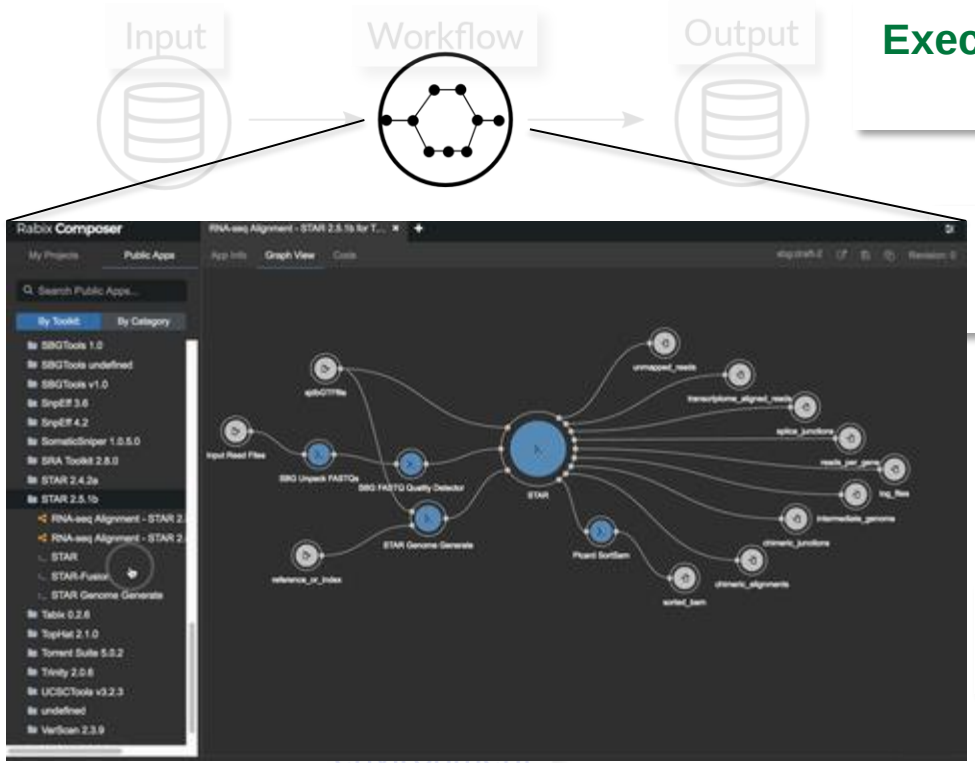
Graphical visualization

Metadata

Up-to-dateness

Compute environment





Graphical User Interface

Graphical visualization

Up-to-dateness

Execution / scheduling / monitoring

Data provenance tracking

Metadata

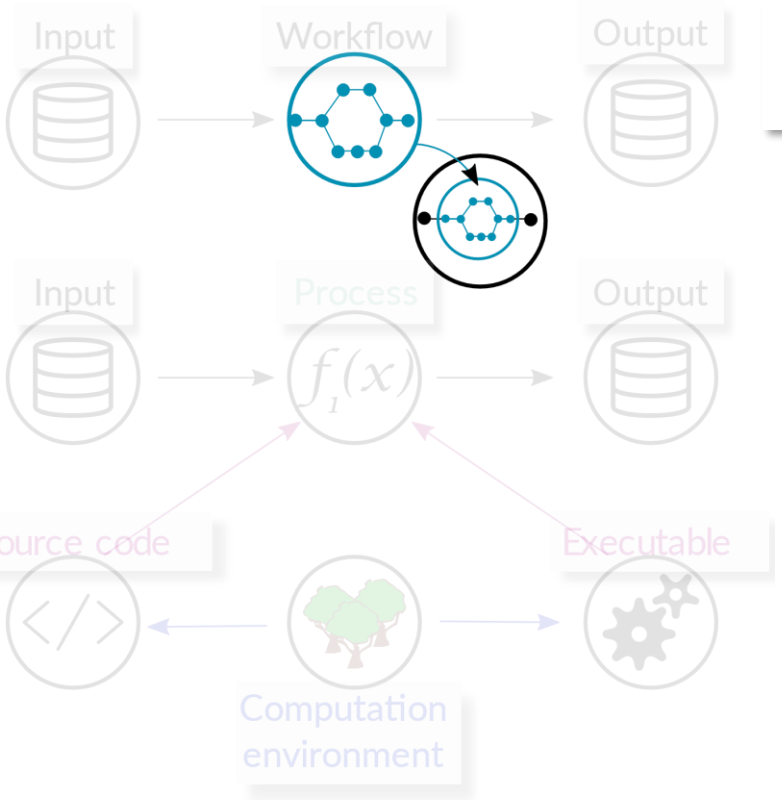
Compute environment

Platform for sharing workflows

Graphical User Interface

Graphical visualization

Up-to-dateness



Execution / scheduling / monitoring

Data provenance tracking

Metadata

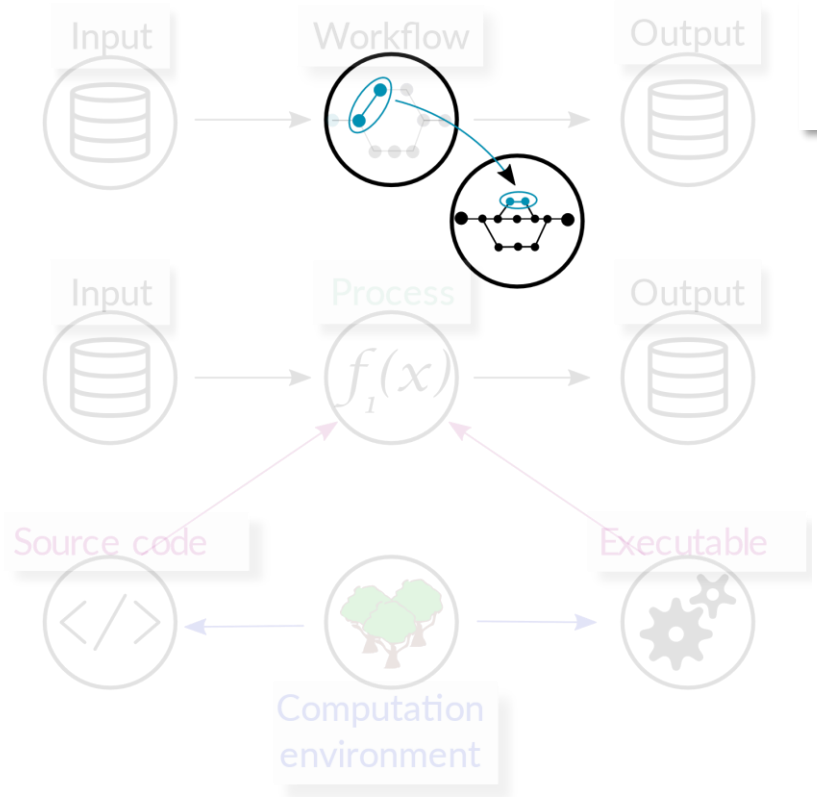
Compute environment

Platform for sharing workflows

Graphical User Interface

Graphical visualization

Up-to-dateness



Execution / scheduling / monitoring

Data provenance tracking

Metadata

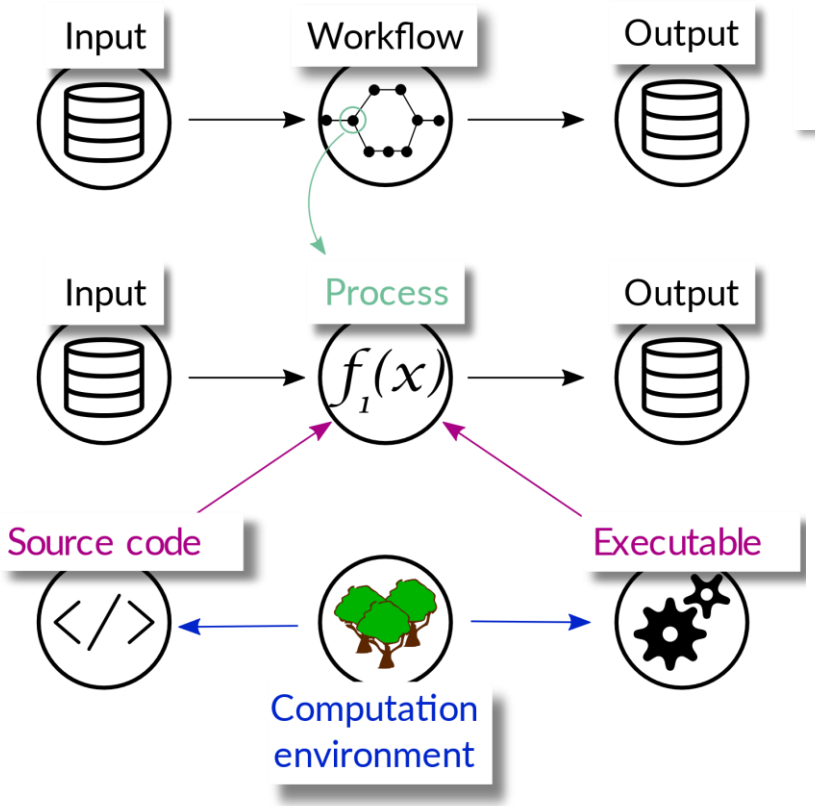
Compute environment

Platform for sharing workflows

Graphical User Interface

Graphical visualization

Up-to-dateness



Execution / scheduling / monitoring

Data provenance tracking

Metadata

Compute environment

Project Concept

- Phase 1
  - Define a simple exemplary use case
  - Implement the required components (processes)
  - Realize the workflow with different workflow tools
- Phase 2
  - Implement a more complex/representative use case
  - Realize with a subset of the tools of phase 1, filtered by requirements
- Phase 3
  - Evaluate the tools regarding the formulated requirements
  - Derive recommendations for suitable tools depending on user story

User stories

User Story A

Reproducible paper

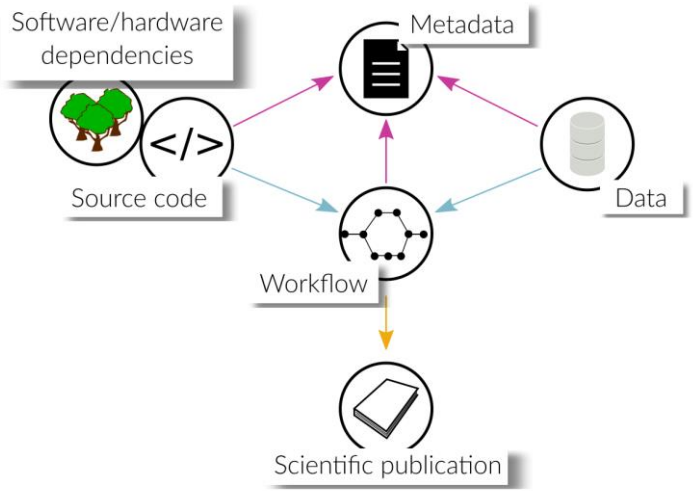
As a researcher, I want to share my paper, such that others are able to reproduce the results.

Goal:

- Reproducible and transparent computational research

Challenges:

- Publication of the (self-written) code
- Publication of (measurement) data
- Sufficient documentation of the code/workflow
- Execution of the workflow by other researchers



User Story B

Research group collaboration

As part of a research group, I want to be able to interconnect and reuse components of several different workflows so that everyone may benefit from their colleagues' work.

monitoring

simulation

Goal:

- Reusability of components/modules in an interdisciplinary workflow

Challenges:

- The workflow may consist of heterogeneous models
- Machine-independent execution
- High computational costs
- Caching of results

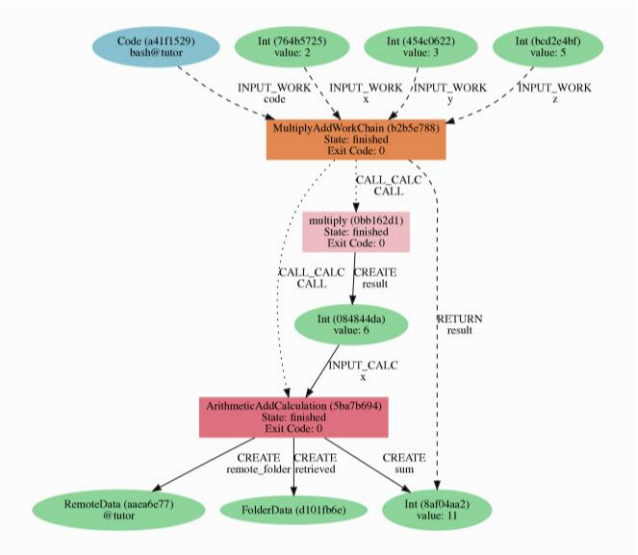
model updating

model prognosis

User Story C

High-throughput simulations

As a materials scientist, I want to be able to automate and manage complex workflows so I can keep track of all associated data.



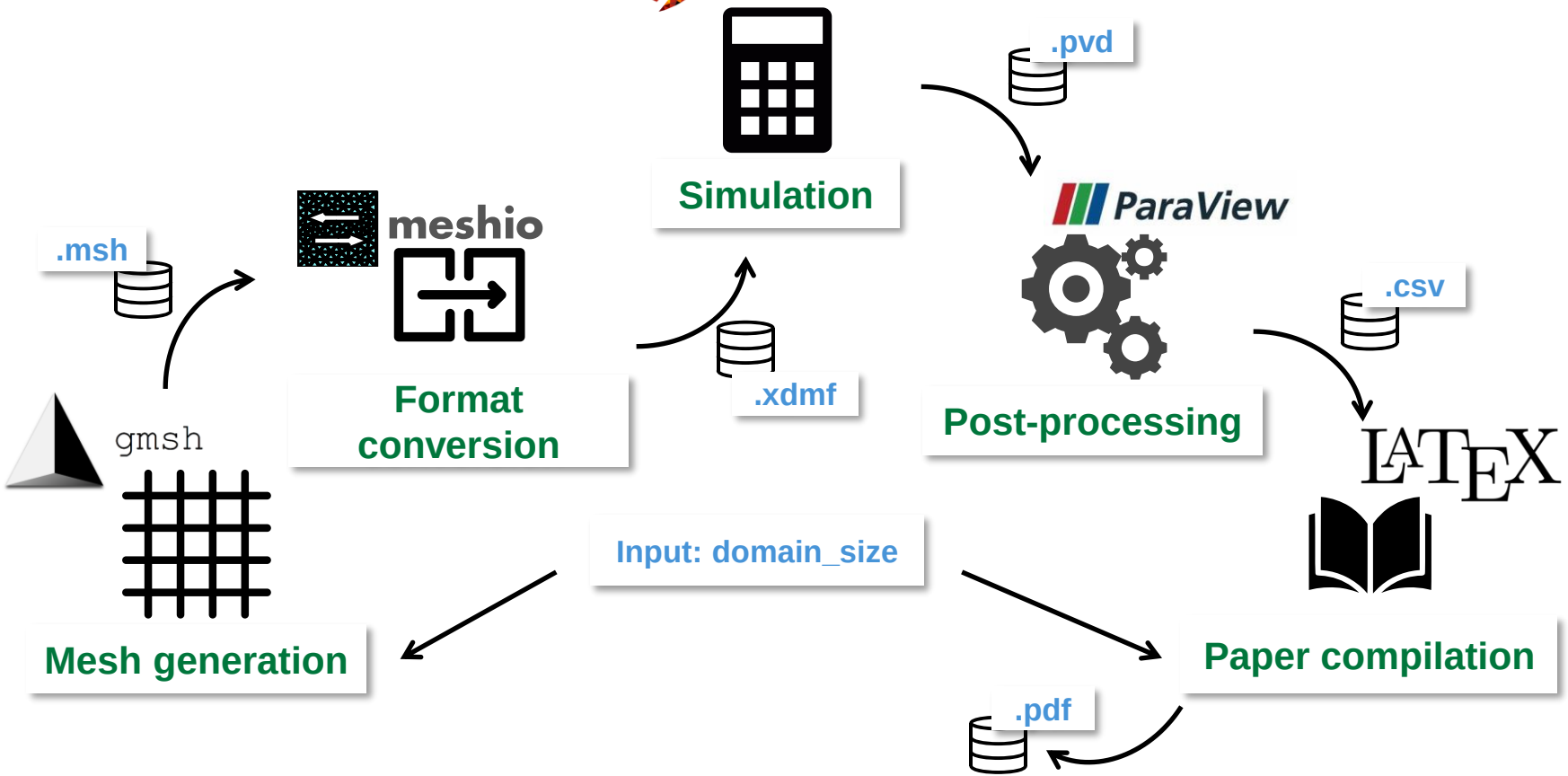
Goal:

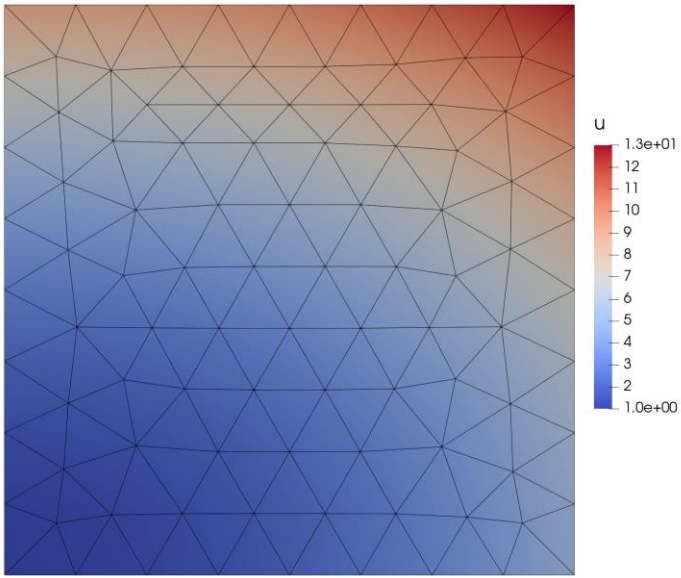
- Automatically record and keep track of all inputs, outputs and metadata of all calculations

Challenges:

- Data provenance (preserve the full lineage of all data)
- Managing of thousands of processes per hour
- Fast graph/database queries
- Graphical visualization
- Dealing with high computational cost

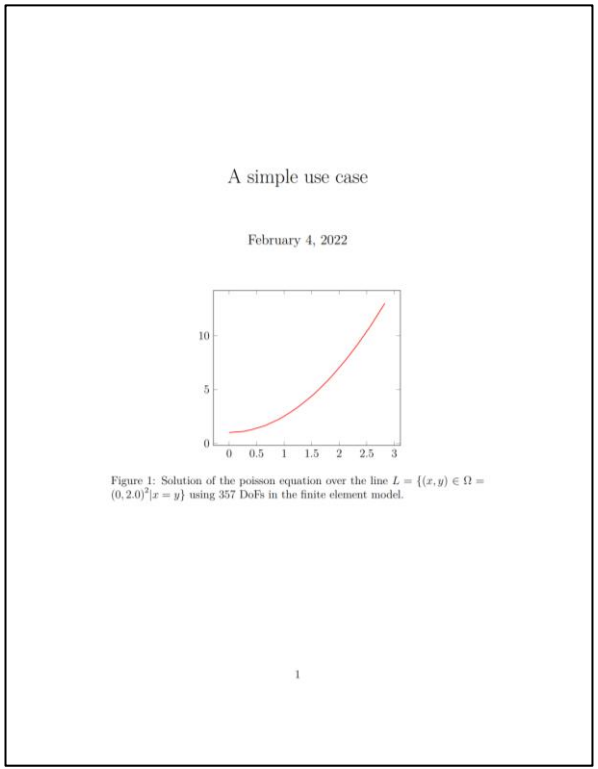
A simple use case

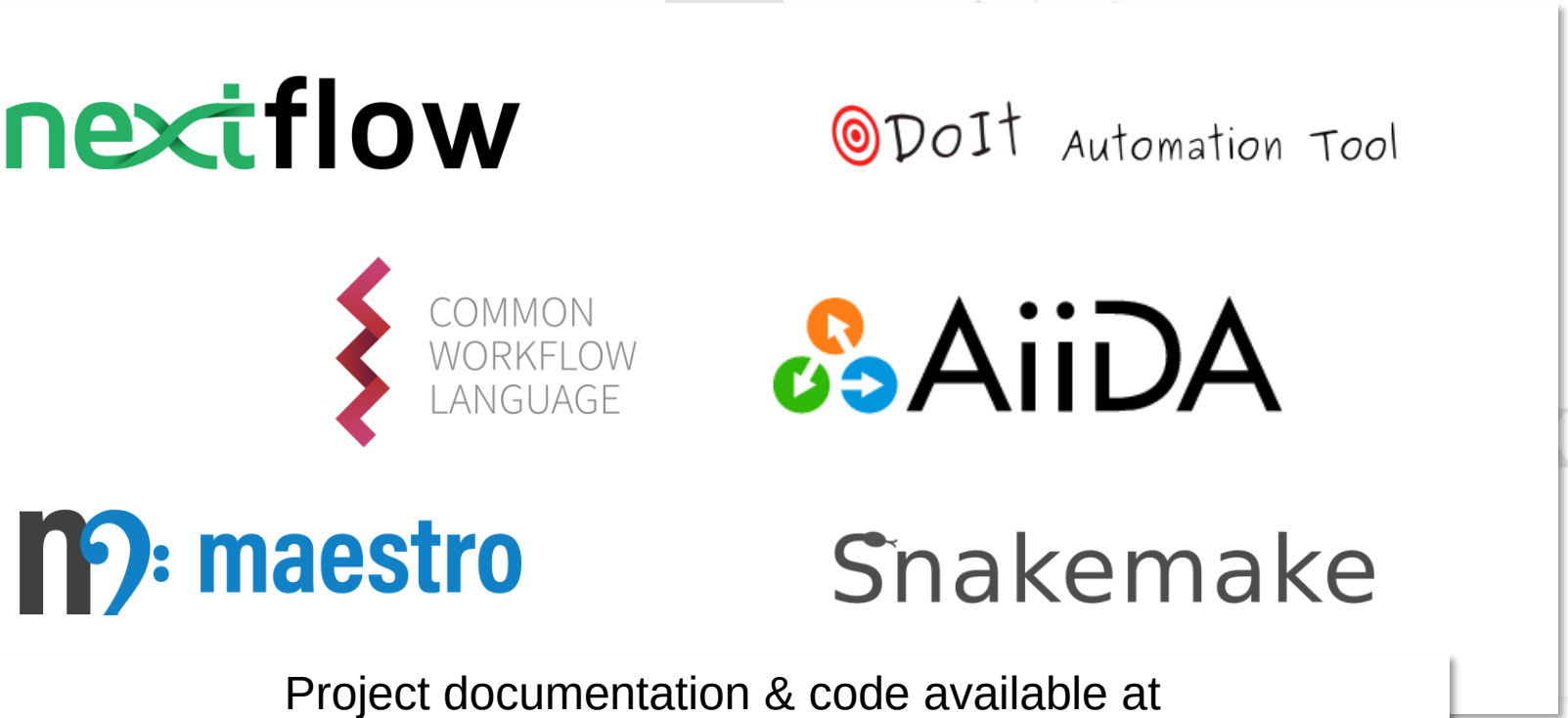

FENICS PROJECT
NFDi4ing
Simulation
.pvd
ParaView
Post-processing
.csv
LaTeX
Paper compilation
.pdf
Input: domain_size
Mesh generation
gmsh
.msh
meshio
Format conversion
.xdmf



FENICS PROJECT

LATEX





Project documentation & code available at
github.com/BAMresearch/NFDI4IngScientificWorkflowRequirements

Example: (part of) the cwl workflow

Process

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: CommandLineTool
doc: |
  Generate the computational mesh with gmsh
baseCommand: ["gmsh"]
arguments: ["-setnumber", "domain_size", "$(inputs.domain_size)",
            "-2", "$(inputs.geofile.path)",
            "-o", "mesh.msh"]
inputs:
  geofile:
    type: File
    doc: "Geometry file in a gmsh-readable format"
    default:
      class: File
      location: ../source/unit_square.geo
  domain_size:
    type: float
    doc: "Specify the size of the domain to be meshed"
outputs:
  mesh:
    type: File
    outputBinding:
      glob: "mesh.msh"
```

Workflow

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: Workflow

outputs:
  paperpdf:
    type: File
    outputSource: compile_paper/pdf

inputs:
  domain_size:
    type: float
    default: 1.0

steps:

  make_mesh:
    run: make_gmsh_mesh.cwl
    in:
      domain_size: domain_size
    out: [mesh]

  convert_mesh:
    run: convert_msh_to_xdmf.cwl
    in:
      inputmesh: make_mesh/mesh
    out: [outputmesh, outputmeshdata]

  run_simulation:
    run: run_dolfin.cwl
    in:
```

Example: (part of) the cwl workflow

Process

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: CommandLineTool
doc: |
  Generate the computational mesh with gmsh
baseCommand: ["gmsh"]
arguments: ["-setnumber", "domain_size", "$(inputs.domain_size)",
            "-2", "$(inputs.geofile.path)",
            "-o", "mesh.msh"]
inputs:
  geofile:
    type: File
    doc: "Geometry file in a gmsh-readable format"
    default:
      class: File
      location: ../source/unit_square.geo
  domain_size:
    type: float
    doc: "Specify the size of the domain to be meshed"
outputs:
  mesh:
    type: File
    outputBinding:
      glob: "mesh.msh"
```

Output

Input

Workflow

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: Workflow

outputs:
  paperpdf:
    type: File
    outputSource: compile_paper/pdf

inputs:
  domain_size:
    type: float
    default: 1.0

steps:

  make_mesh:
    run: make_gmsh_mesh.cwl
    in:
      domain_size: domain_size
    out: [mesh]

  convert_mesh:
    run: convert_msh_to_xdmf.cwl
    in:
      inputmesh: make_mesh/mesh
    out: [outputmesh, outputmeshdata]

  run_simulation:
    run: run_dolfin.cwl
    in:
```

Next steps

- More complex use case
- Support on HPC systems?
- Evaluation of the tools w.r.t. defined requirements
- Next SIG status meeting on the 23rd of March

Interested in joining?
dennis.glaeser@iws.uni-stuttgart.de