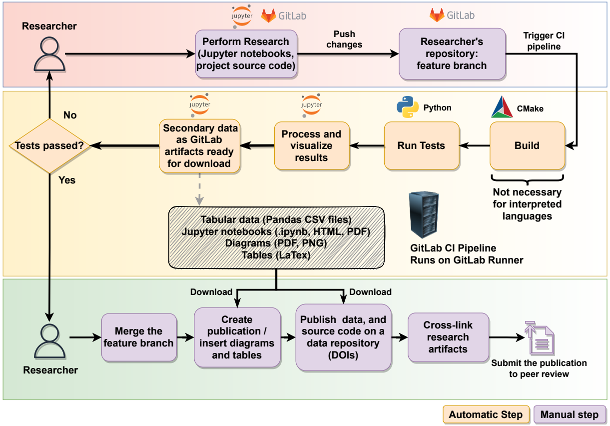
Engineering Research Software in Computational Science and Engineering Fundamentals

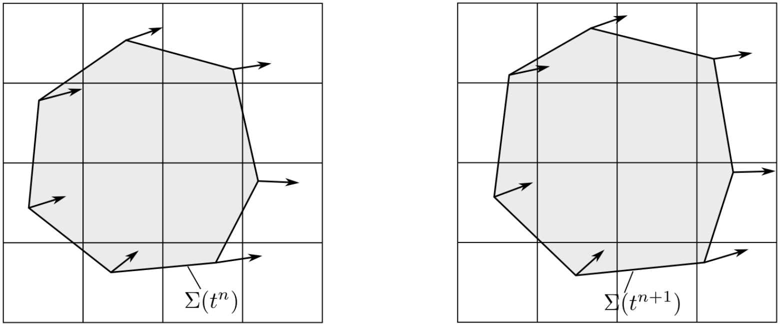
NHR@FAU 2025-05-20



Outline

Motivation: multiphase flow simulation methods

Lagrangian / Eulerian Interface Advection (LEIA) methods



- Fluids that do not mix are separated by an interface $\Sigma(t)$ (surface in 3D).
- Goal: track $\Sigma(t)$ as it moves in time t and changes its topology.

Motivation: multiphase flow simulation software

Lagrangian / Eulerian Interface Advection (LEIA) Methods

TECHNISCHE UNIVERSITÄT DARMSTADT

LEIA methods [1, 2, 3, 4, 5] require **thorough testing**:

- **Verification** cases: evolution of $\Sigma(t)$ and two-phase flows with exact solutions.
- **Validation** with respect to experiments.
- Testing **serial and parallel computational efficiency**.

---

[1] Marić, T., Marschall, H., & Bothe, D. (2015). lentFoam–A hybrid Level Set/Front Tracking method on unstructured meshes. Computers & Fluids, 113, 20-31.

[2] Tolle, T., Bothe, D., & Marić, T. (2020). SAAMPLE: A Segregated Accuracy-driven Algorithm for Multiphase Pressure-Linked Equations. Computers & Fluids, 200, 104450.

[3] Marić, T., Kothe, D. B., & Bothe, D. (2020). Unstructured un-split geometrical Volume-of-Fluid methods–A review. Journal of Computational Physics, 420, 109695.

[4] Marić, T. (2021). Iterative Volume-of-Fluid interface positioning in general polyhedrons with Consecutive Cubic Spline interpolation. Journal of Computational Physics: X, 11, 100093.

[5] Tolle, T., Gründing, D., Bothe, D., & Marić, T. (2021). Computing volume fractions and signed distances from triangulated surfaces immersed in unstructured meshes. arXiv preprint arXiv:2101.08511.

Computational Science and Engineering software in university research groups

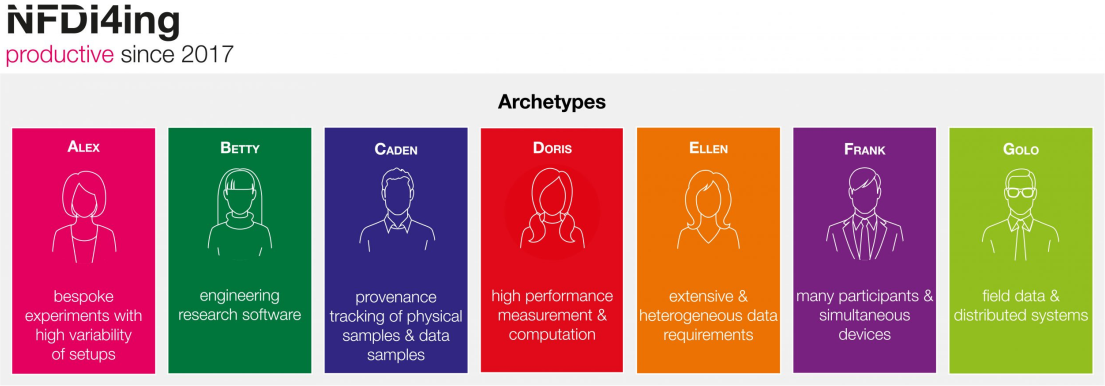
Boundary and initial conditions

- Publish or perish 🎓[6] prioritizes publications over scientific software.
- Dedicated resources for increasing software quality are usually not available.
- Ph.D. students rotate every 3-5 years, postdocs every 1-2 years.
  - Little or no overlap between successors and predecessors.
- Large-scale software design is not a mandatory part of the CSE curriculum.
  - Different CSE background: (Applied) Mathematics, Mechanical Engineering, Physics, Informatics.

[6] Symbol of a publish-or-perish simplification of the workflow :)

NFDI4Ing to the rescue!

Resources for engineering research software



NFDI4Ing resources.

Computational Science and Engineering software in university research groups

The chaos scientific legacy code

Betty is a CSE researcher, working with a legacy research code.
Why is Betty so (rightfully) angry?

- Betty inherited a **research software that is only partially tested**.
- Betty inherited a **research software that isn't automatically tested**.
  - Betty changes one part of the code and gets her model running, only to see 10 other things fail, after days of manually running tests.
- Betty's software has **no documentation of the scientific workflow**.
  - Betty doesn't know how to use existing scripts to run simulations and analyze (reproduce) results.
- Betty's software has **disjoint (diverging) versions** - that she can't integrate.
- **Betty can't even find code versions used to generate results in the publications from her research group.**

> "Après moi, le déluge" - "After me, the flood"
>
> Louis XV of France

Research software generally does not matter, as long as papers are published (🎓).

Missed opportunities - Industrial Career

- DevOps - Development Operations is gaining traction in Engineering.
- Companies buy software based on internal benchmarks - automatic testing.
- Re-use of software implementations within the company - version-control.
- Re-use of research data within the company - cross-linking digital artifacts.

TECHNISCHE UNIVERSITÄT DARMSTADT

"Après moi, le déluge" - "After me, the flood"

Louis XV of France

Research software generally does not matter, as long as papers are published (🎓).

Missed opportunities - Acadmic Career

- *Finding results* made easy by cross-linking code versions, data and publications.
- *Faster extension / combination of existing ideas* if their respective versions are integrated.
- *Faster comparison of results* with previous ideas automating verification / validation.
- *Automatic reproducibility* of results using automated testing and version control.
- *Faster onboarding* with documented scientific verification and validation workflows.

Computational Science and Engineering software in university research groups

Continuous integration and cross-linking to the rescue

Automated testing (verification and validation), **version control**, and **cross-linking** reports, source code and research data increase Findability, Accessibility and Reproducibility (FAIR) and **speed up research.**

- **Continuous Integration (CI) = automatic testing + version control.**
- CSE research requires **scientific workflows**: initialize simulations, run parameter variations, agglomerate data, visualize, and check results.
- CI can be used to **automate and document scientific workflows**.
- CI ensures that the **integration of new changes does not break existing functionality**.
- Once the changes are integrated, the publication, the source code and the data are published on pre-print and data repositories and **cross-linked** using git tags and DOIs.

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Software engineering: version control

What is version control?

- Management of versions of (usually) textual data, like publications and scientific codes.
- Nowadays version control is **essential** for scientific codes of all shapes and sizes.
- A basis for productive research in teams and increasing the quality of scientific software[7].

[7] Maric, Tomislav, Lehr, Jan-Patrick, Papagiannidis, Ioannis, Lambie, Benjamin, Bothe, Dieter, & Bischof, Christian. (2021, April). A Workflow for Increasing the Quality of Scientific Software (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.4668439

TECHNISCHE UNIVERSITÄT DARMSTADT

Why use version control in scientific codes?

- The ability to work with others (colleagues or students) on your research project.
  - Work together faster.
  - Re-use an interpolation method of a colleague in the group.
- The ability to trivially try out new ideas and switch back if they don't work.
  - Speeds up research!
- The ability to easily recover versions of your project in the same folder.
  - Recovering a specific version in a predecessor project code.
- The ability to understand the motivation behind changes via comments.
  - Crucial for continuing existing research projects.
- The ability to increase the reproducibility of scientific results.
  - Basis for cross-linking of data, source code and publications / reports.

Software engineering: version control

Git version control system (VCS)

- An effective and easy to use software with a set of commands for version control.

Software engineering: version control

Git basics on a single slide

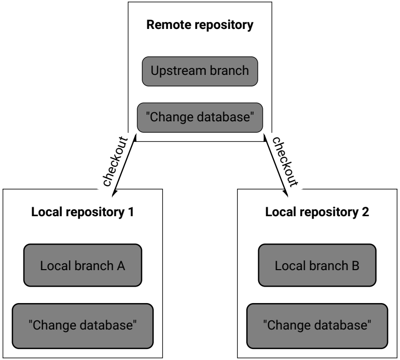
TECHNISCHE UNIVERSITÄT DARMSTADT

- The code/text folder is called a **repository**.
- An online folder shared with the team is the **remote repository** (short: remote).
- Create a new version: **checkout** a new **branch**.
- Integrate with another version: **merge** with a **branch**.
- Add changes in a branch: **add** changes.
- Integrate changes into a branch: **commit** changes.
- Share changes with others: **push** to **upstream repository**.
- Get latest changes: **pull** from the **upstream repository**.

Learn basic git *concepts*, they are the same everywhere.

- Git in 15 minutes
- Git within Matlab
- Feature branch workflow
- GitLab for beginners

Software engineering
Decentralized version control



- **Although git tracks only changes, every repository is still a complete copy of the project.**
- Offline work is supported!

Software engineering: version control

Version control "enforces" modularity

TECHNISCHE UNIVERSITÄT DARMSTADT

Git conflicts

- A file is changed differently on two branches and a merge is needed.
- Two team members edit the same file at once.

Modularity reduces conflicts and speeds up teamwork

- Book chapters as separate files vs. book chapters as folders and sections as separate files.

Modularity via Separation of Concerns and Single Responsibility

TECHNISCHE UNIVERSITÄT DARMSTADT

- University research teams (like our LEIA lecture team!) are generally small (2 - 5 members).
- ▸ Separation of Concerns (SC) and ▸ Single Responsibility Principle (SRP) significantly simplify the branching model.
- **Separation of Concerns**: code is organized in non-overlapping layers and sections.
- **Single Responsibility**: functions or classes perform single clear tasks.
- SC and SRP can be applied to any software.
- Dogmatism should be avoided: single responsibility vs less responsibilities.

Simple version-control branching model

Separation of Concerns and Single Responsibility

- University research teams *working on the same project* are generally small (2 - 5 members).
- ▸ Separation of Concerns (SC) and ▸ Single Responsibility Principle (SRP) significantly simplify the branching model.
- **Separation of Concerns**: code is organized in non-overlapping layers and sections.
- **Single Responsibility**: functions or classes perform single clear tasks.
- SC and SRP can be applied to any software.
- Dogmatism should be avoided: single responsibility vs less responsibilities.
- OpenFOAM already uses object-oriented and generic software design patterns.

Simple version-control branching model

Change integration

TECHNISCHE UNIVERSITÄT DARMSTADT

Maintainers (postdocs, experienced Ph.D. students) manage the integration.

- Keep the branching model as simple as possible.
- Main and development branches are protected and managed by Maintainers.
- Maintainers are responsible for git tags and cleanup:
  - **Main**: integrations from *accepted publications* and *development branch*.
  - **Development**: integration of *(CI)-tested improvements*.
  - **Feature**: SRP reduces git-conflicts with researchers working on different files.
- Complex branching workflow $\Rightarrow$ complications with onboarding new members.

Outline

TDD[8] for CSE

- Define verification and validation tests at the start.
- Focus placed the final result: interpolation, integration, discretization, PDE solution, physics.
- Top-down, instead of bottom-up test coverage.
- Don't go overboard with unit-tests 🎓: extend unit-tests when debugging a failing CSE test.
- Focus kept on tests with real-world (publication) input.

[8] Freeman, Steve, and Nat Pryce. Growing object-oriented software, guided by tests. Pearson Education, 2009.

Verification and validation tests define the Application Programming Interface

- **New code**: it is easier to program the API you wish for, if you are its first user.
  - Make the class interface easy to use correctly and difficult to use incorrectly[9].
  - Reduce number of function arguments, single responsibility, clear naming, ...
- **Legacy code**: extend existing API without modification.
  - OpenFOAM: understanding class hierarchies, *finding a base class with Runtime Type Selection and a virtual function to overload.*
- **The test application is the solver application with a different input.**
  - If possible, testing and solution is done by the same code.
  - This prevents code duplication.
  - Data output and additional checks can be disabled by (compile-time) options.

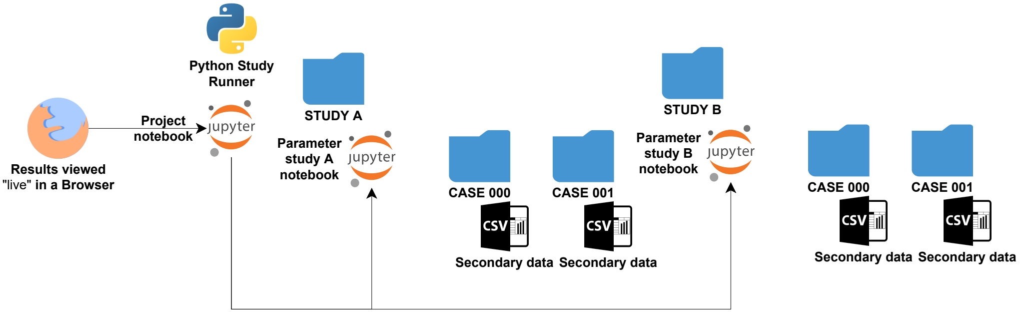
[9] Scott Meyers. 2014. Effective Modern C++: 42 Specific Ways to Improve Your Use of C++11 and C++14 (1st. ed.). O'Reilly Media, Inc.

Jupyter notebooks[10]

- **Documentation**: geometry, initial and boundary conditions, error norms, comparison data.
- **Processing**: verification errors (conservation, convergence, stability), validation errors.
- **Result analysis**: very straightforward, interactive, remote.

[10] https://jupyter.org/

Test Driven Development
(Parameter tests)



Test Driven Development

Parameter tests: primary data (simulation results) organization

- The quality of CSE software is measured using verification and validation data.
- Effective comparison with others (previous versions) hinges on data organization.

- **Legacy code**:
  - use the existing folder structure and parameterization tools 🎓,
  - The mapping (case000) $\rightarrow$ (parameter vector) must be stored (YAML, ...)
- **New code**:
  1. Simple folder and file structure 🎓
  2. HDF5[11] or other open data format.
  3. Alternative to HDF5: **ExDir**[12]

[11] https://www.hdfgroup.org/solutions/hdf5

[12] Dragly, Svenn-Arne, et al. "Experimental Directory Structure (Exdir): An alternative to HDF5 without introducing a new file format." Frontiers in neuroinformatics 12 (2018): 16.

Test Driven Development

Parameter tests: secondary data (tables and diagrams) organization

pandas.MultiIndex CSV with metadata for secondary data

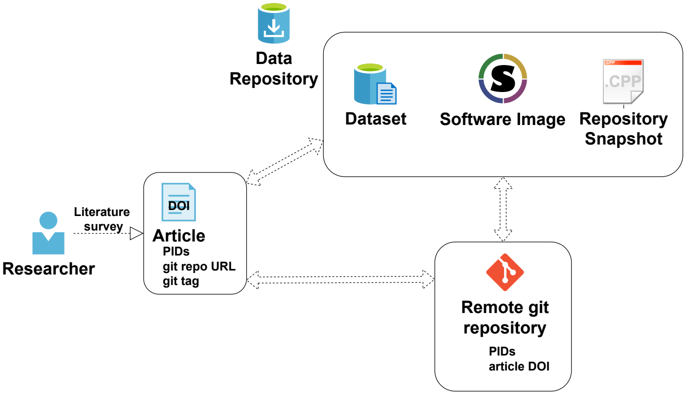
- `pandas.MultiIndex` saved in "metadata columns".
- **Metadata is repeated**: not an issue for the small secondary data!
- Metadata in columns $\rightarrow$ `pandas.MultiIndex` $\rightarrow$ strongly simplified data analysis.
- **Direct readable export of tables to LaTex!**

	H	L_INF	O(L_INF)	EPSILON_R_EXACT_MAX	O(EPSILON_R_EXACT_MAX)
VELOCITY_MODEL					
SHEAR_2D	0.125000	0.032961	1.833407	0.032961	1.833407
SHEAR_2D	0.062500	0.009249	1.955529	0.009249	1.955529
SHEAR_2D	0.031250	0.002385	1.988745	0.002385	1.988745
SHEAR_2D	0.015625	0.000601	1.997178	0.000601	1.997178
SHEAR_2D	0.007813	0.000150	1.999294	0.000150	1.999294
SHEAR_2D	0.003906	0.000038	1.999294	0.000038	1.999294

Outline

Cross-linking data, source code and reports/publications

Schematic diagram



- Whence the Singularity Image[13]?
  - More intuitive than Docker: **Singularity handles images as files.**
  - Built for HPC from the start.
  - Doesn't require root rights.
  - Results as *actual files*, not "data in spinning containers".
  - Maps user folder to the container: result data remains on the host.
- Why not replace Docker with Singularity within GitLab CI?
  - We're learning how to do this using GitLab custom executors.
  - Does the workflow still survive publish-or-perish 🎓 test?
- Why a source-code snapshot on-top of the image and the repository?
  - Repositories get migrated, deleted, and some researchers still fear images.
  - Quick and direct access to source code from the publication.

[13] https://sylabs.io/docs/

(Cross-linking data, source code and reports/publications)

Singularity simplifies reproducibility

TECHNISCHE UNIVERSITÄT DARMSTADT

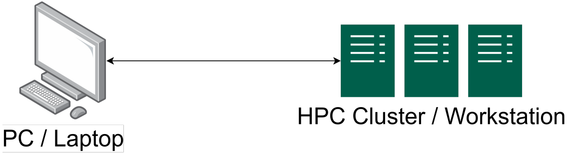
- The source code and the data stored in the image can be quickly reproduced.
- Article reviewers can clone, build, run and visualize easily.

Example: Singularity Image from an active review

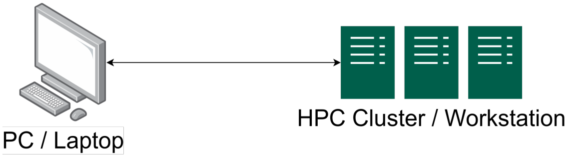
- Clone the code repository from the image:
  `geophase-JCOMP-D-19-01329R2.sif clone geophase`
- Build:
  `geophase-JCOMP-D-19-01329R2.sif build geophase build`
- Run tests:
  `geophase-JCOMP-D-19-01329R2.sif run-tests geophase build`
- Open the jupyter notebook:
  `geophase-JCOMP-D-19-01329R2.sif jupyter-notebook geophase`

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

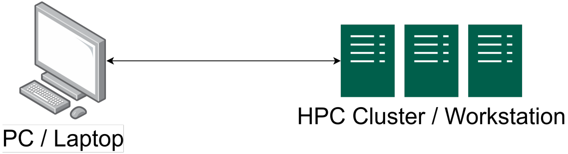


while Results are unsatisfactory **do**
    Work on algorithms.
    (Compile the code.)
    for All studies **do**
        Prepare the study.
        Run the study.
        Analyze results.
        Move results to a report.
    end for
    Compare old and new results.
end while

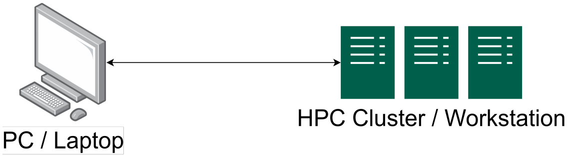


Issues...

- Starting studies takes time.
- Analyzing results takes time.
- Often the results are not checked "live" as the study runs - **waste of research time and CPUh**.
- **Only the researcher knows the details** behind the initialization, running and post-processing scripts - **when this person leaves, the reproducibility is gone.**
- A researcher may forget to run a study and believe all tests have passed.



while Results are unsatisfactory **do**
  Work on algorithms.
  (Compile the code.)
  for All studies **do**
    Prepare the study.
    Run the study.
    Analyze results.
    Move results to a report.
  end for
  Compare old and new results.
end while



while Results are unsatisfactory **do**

  Work on algorithms.

  (Compile the code.)

  Run initialization scripts (jobs).

  Run simulation scripts (jobs).

  (Run postprocessing scripts (jobs)).

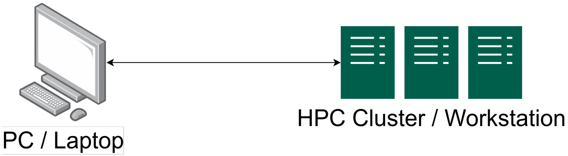
  Visualize results live in Jupyter notebooks.

end while

Manual steps of the research workflow,

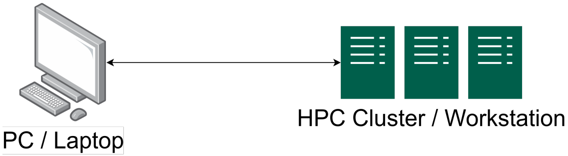
```
(Compile the code.)
for All studies do
    Prepare the study.
    Run the study.
    Analyze results.
    Move results to a report.
end for
Compare old and new results.
```

are now automated using scripts **that do not require additional knowledge / input (metadata).**



1. The **new** results are satisfactory.
2. Similar automated workflows are executed for existing tests.
3. All results are checked.
4. The milestone has been reached, the version can be integrated.

Works well manually when there aren't many previous verification/validation tests and their analysis is relatively simple.
Are we sure that we ran all the tests and examined the results properly?

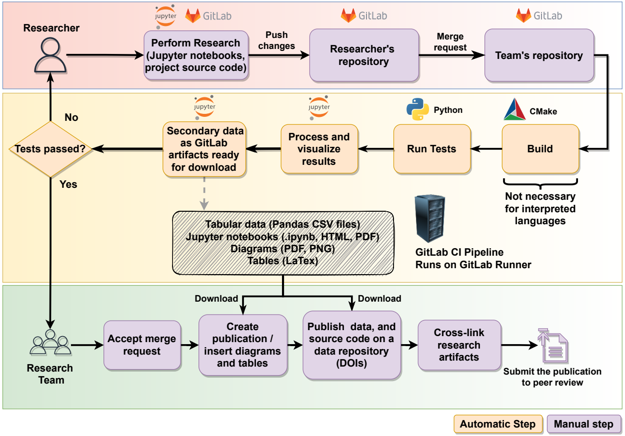


- **Manual testing takes a lot of time**.
- **Manual testing** of all previous tests **is prone to error** - even if V&V scripts do not require metadata.
- Relevant **V&V tests are automated using Continuous Integration (CI)**.
  - Changes are pushed to the upstream version control repository.
  - The remote repository starts the so-called **CI** test pipeline (a sequence of tests).
  - Tests are automatically run, processed and visualized.

(Continuous) Integration of scientific software

Schematic diagram for the team workflow

TECHNISCHE UNIVERSITÄT DARMSTADT



Working in a team.

TECHNISCHE UNIVERSITÄT DARMSTADT

- A **text (YAML) file** is added to a repository, that **specifies the tests** (jobs) in a CI pipeline.
- When the YAML file is pushed to an upstream git repository (GitLab), **GitLab creates a CI pipeline from the YAML file**.
- The CI pipeline needs a **machine for running tests - the GitLab runner**.
  - Shared runners on gitlab.com have limited capacity.
  - We can install and register our own GitLab runner.
- **A Docker image encapsulates the computing environment**.
  - **Virtualization/Containerisation** increases reproducibility and simplifies testing.
- The Docker image must be publicly accessible for it to be used by a shared runner.

(Continuous) Integration of scientific software

CI in a nutshell II

```yaml
initialization_param_study:
  stage: running
  dependencies:
    - build_release
  script:
    # run the parameter variation tests
    - cd cases/initialization/3dinit
    - ./create_and_run_levelset.sh
    - ./reproduce_publication_results.sh
  artifacts:
    paths:
        - cases/initialization/3dinit/*.csv
        - cases/initialization/3dinit/*.pdf
```

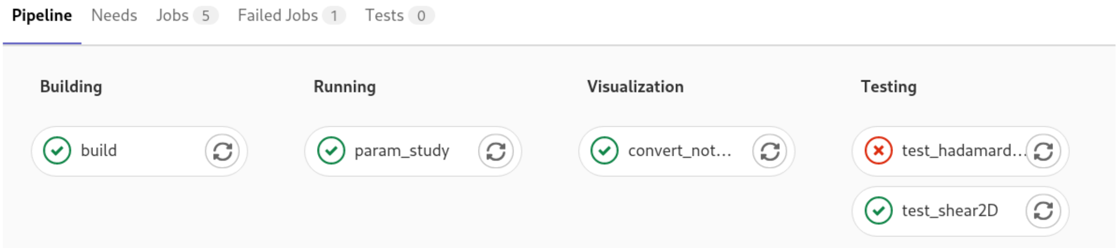
Example YAML file

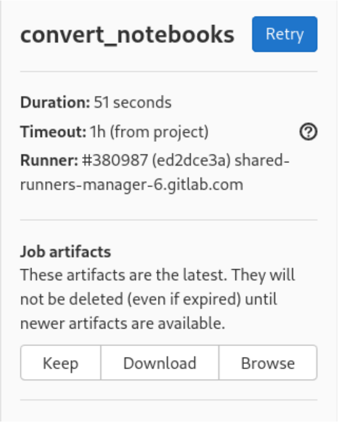
- The **CI pipeline starts the right scripts in the right order**: it documents the research workflow.
- A click of a button in a web browser reproduces results for any version of the research software.
- Continuous **integration** is used to **integrate** only those changes that improve the software and don't break existing tests.

(Continuous) Integration of scientific software

CI in a nutshell I

An example CI pipeline

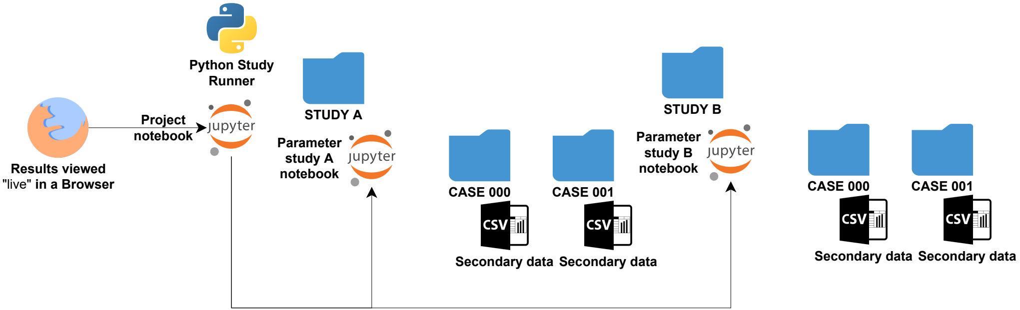




- Files created within a CI job are gone when the job ends.
- GitLab uses **job artifacts** to pass on data from one job to the next.
- **Job artifacts can only be files stored in project's sub-folders.**
- Libraries and applications are passed to other jobs as artifacts.
- **Artifacts can be downloaded on the GitLab project website.**

(Continuous) Integration of scientific software

Running tests I



Organize your simulation studies.

(Continuous) Integration of scientific software

Running tests II

TECHNISCHE UNIVERSITÄT DARMSTADT

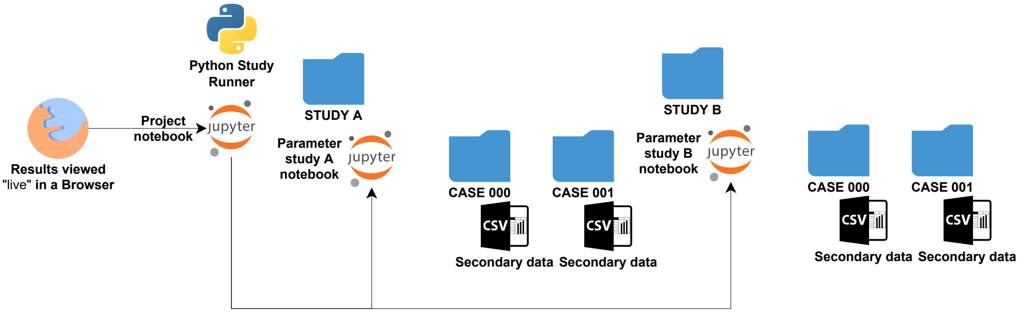
- Success of CSE methods is measured using verification and validation data.
- Effective comparison with others (previous versions) hinges on data organization.
- **Goal:** easily **programmatically identify** parameters used in a simulation case.
- **Legacy code**:
  - use the existing folder structure and parameterization tools
  - The mapping (case000) $\rightarrow$ (parameter vector) must be stored (YAML, ...)
- **New code**:
  1. Simple folder and file structure
  2. HDF5[14] or other open data format.
  3. Alternative to HDF5: **ExDir**[15]

[14] https://www.hdfgroup.org/solutions/hdf5

[15] Dragly, Svenn-Arne, et al. "Experimental Directory Structure (Exdir): An alternative to HDF5 without introducing a new file format." Frontiers in neuroinformatics 12 (2018): 16.

(Continuous) Integration of scientific software

Running tests I



- Associate simulation cases with their metadata.
- `{case000 : {N_CELLS: 32, MODEL : shear2D}}`
- Store this information using a standard open-source format (**Interoperability** in FAIR).

Use Jupyter notebooks[16] and pandas[17] for

- **Documentation**: geometry, initial and boundary conditions, error norms, comparison data.
- **Data processing**: verification errors (conservation, convergence, stability), validation errors
- **Result analysis**: interactive and remote, while simulations are running!

[16] https://jupyter.org/
[17] https://pandas.pydata.org/

(Continuous) Integration of scientific software

Processing and visualizing tests

TECHNISCHE UNIVERSITÄT DARMSTADT

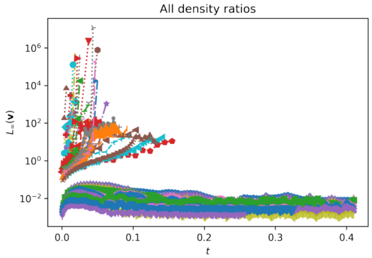
```
jupyter nbconvert notebook.ipynb --execute --to FORMAT
```

- Agglomerate secondary data into `pandas.MultiIndex` CSV files.
- Run each jupyter notebook in the repository.
- Export secondary data and notebooks in different formats as artifacts.
- **Visualization**
  - Download the artifact and open the notebook 🎓.
  - Notebooks contain information on failing tests.
  - Mapping "caseID" $\rightarrow$ "parameters" is crucial for re-starting failed parameter variations!

(Continuous) Integration of scientific software

Secondary data I

- Data used for diagrams and tables in a publication.
- Data we compare our results with.
- Data we waste time scanning from (sometimes our own) publications in CSE.



(a) Old inconsistent method: interface stable only for cases with density ratio $\rho^-/\rho^+ = 1$

Imagine scanning this diagram.

Preprint: https://arxiv.org/abs/2109.01595

Data: https://doi.org/10.48328/tudatalib-627

TECHNISCHE UNIVERSITÄT DARMSTADT

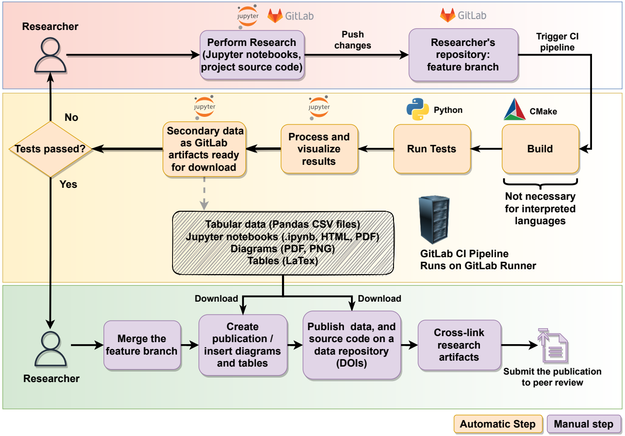
pandas.MultiIndex CSV with metadata for secondary data

- `pandas.MultiIndex` saved in "metadata columns".
- **Metadata is repeated**: not an issue for the small secondary data!
- Metadata in columns → `pandas.MultiIndex` → strongly simplified data analysis.
- **Direct readable export of tables to LaTeX!**

	H	L_INF	O(L_INF)	EPSILON_R_EXACT_MAX	O(EPSILON_R_EXACT_MAX)
VELOCITY_MODEL					
SHEAR_2D	0.125000	0.032961	1.833407	0.032961	1.833407
SHEAR_2D	0.062500	0.009249	1.955529	0.009249	1.955529
SHEAR_2D	0.031250	0.002385	1.988745	0.002385	1.988745
SHEAR_2D	0.015625	0.000601	1.997178	0.000601	1.997178
SHEAR_2D	0.007813	0.000150	1.999294	0.000150	1.999294
SHEAR_2D	0.003906	0.000038	1.999294	0.000038	1.999294

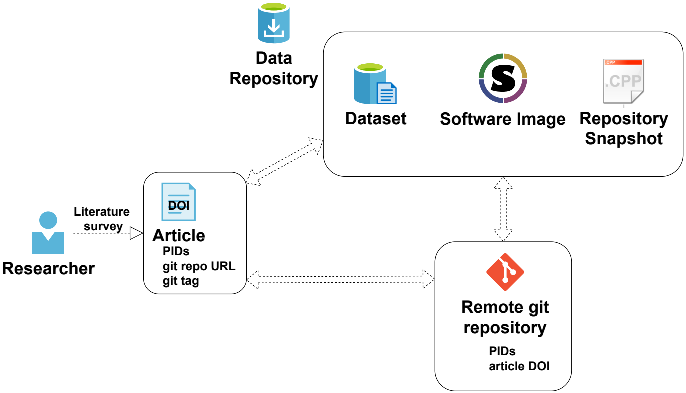
(Continuous) Integration of scientific software

Schematic diagram for the individual workflow



Working alone.

(Continuous) Integration of scientific software
Cross-linking I



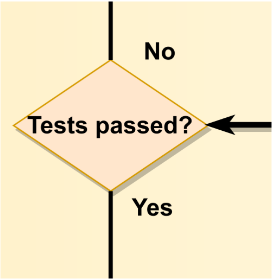
TECHNISCHE UNIVERSITÄT DARMSTADT

Cross-linking is done manually.

- Place whatever you can under version control.
- **When a set of milestones is reached (*release*)** , use git-tags as version snapshots, and upload the research data to a data repository, e.g. TUDatalib at TU Darmstadt, or Zenodo.
    - Secondary data (diagrams, tables), raw data (simulations, experiments), archive of the research software, ...
- Data uploaded to a data repository is associated with Persistent Identifiers (PIDs), e.g. DOIs.
- Cite the research data using DOIs in the report (article, preprint).
- Upload the report to a pre-print repository, e.g. ArXiv.
- Edit the data on the data repository and mention the arXivID.
- Submit the pre-print to a journal for peer-review.

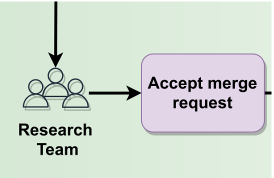
TECHNISCHE UNIVERSITÄT DARMSTADT

- **Research software is compared with existing publications.**
- A major milestone are improved results for a set of verification / validation tests.
- The cross-linking therefore revolves around the publication (pre-print, report, ...).
- The cross-linking makes it possible to find the version of research software used to generate the results in the publication: repository link + git tag, repository snapshot, software image.
- Once the version is found, CI automatically reproduces all results from the publication with a click of a button.

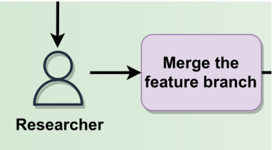


Straightforward for easily quantifiable errors

- **Examples:** volume conservation, order of convergence, total wall clock time, weak scaling, ...
- **Python scripts test secondary data** agglomerated by Jupyter notebooks from simulation results.

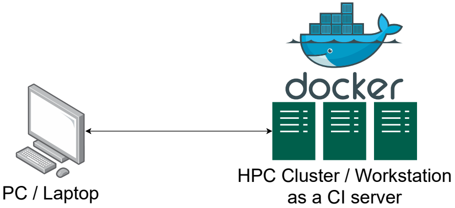


or

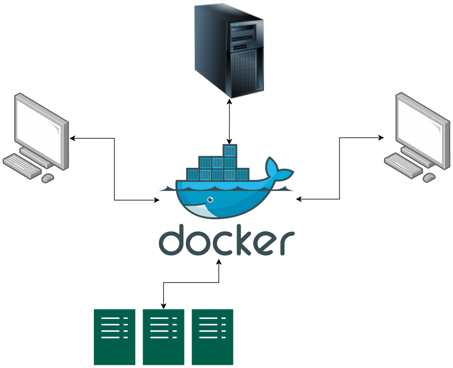


Difficult for errors that cannot be quantified easily

- **Examples:**
  - Is is the difference between simulation and experiment data $\leq 4\%$?
  - How to quantify the difference for complex signals?
- **Option 1: Researchers evaluate the test results even if all CI jobs pass.**
  - A simple and efficient solution 🎓.
- **Option 2: Use statistics to quantify the difference.**



- Instead of installing the research software only on the laptop/PC and the HPC cluster / workstation, we install it in a virtual environment - **a Docker image**.
- The Docker image then works on any machine that runs Docker.
- Sharing research software becomes trivial - if our colleague wants to use our software, no installation (besides Docker) is required.



- Instead of installing the research software only on the laptop/PC and the HPC cluster / workstation, we install it in a virtual environment - **a Docker image**.
- The Docker image then works on any machine that runs Docker.
- Sharing research software becomes trivial - if our colleague wants to use our software, no installation (besides Docker) is required.

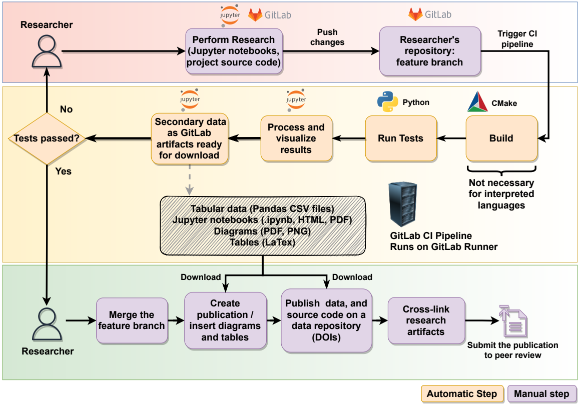
The GitLab CI requires a **GitLab runner: a machine that runs the CI jobs.**

1. **Short few CPU-core tests**: work-PC 🎓.
2. **Short many-core tests**: obtain a workstation with a 64-Core CPU[18]🎓.
3. **HPC tests**: combine 1. or 2. with an HPC cluster.

An HPC cluster is relevant for production tests and performance measurements.

- This workflow uses coarse ("smoke") tests 🎓
  - Unit tests run for 1. and 2.
  - Convergence ensured for 1. and 2.
  - Is efficient in parallel for 1. and 2.
- **Challenge**: Is it possible to combine 1., 2. and 3. and publish instead of perish 🎓?

[18] Thanks to CRC 1194 at TU Darmstadt.




```
Track changes using version-control.
while Milestone not reached do
    for study in studies do                          ▷ On an HPC cluster.
        Automate data processing and visualization.
        Run study.
        Check results and apply code changes.
    end for
    if results are improved on the HPC cluster then
        Push changes to the remote repository.
        if CI pipeline tests pass then
            Milestone reached.
            Add new tests to the CI pipeline,
            Merge feature into development branch.
            Cross-link publication, data, and source code.
        end if
    end if
end while
```

(Continuous) Integration of scientific software

Similarity with other workflows / best practices

TECHNISCHE UNIVERSITÄT DARMSTADT

Our *(subjective)* estimates* of similarity $1 - 5$ (higher is more similar), $-$: aspect not addressed.

DOI	*Branching model*	*TDD*	*Cross-linking*	*CI*	(Meta)data standardization
10.12688/f1000research.11407.1	-	-	-	-	1
10.3934/math.2016.3.261	-	-	-	-	2
10.1371/journal.pbio.1001745	1	2	-	-	-
10.1371/journal.pcbi.1005510	-	-	3	1	3
10.1145/2723872.2723881	1	-	-	1	-
10.1145/3324989.3325719	1	-	-	5	-
10.1371/journal.pone.0230557	1	-	-	1	4
10.1145/3219104.3219147	1	-	-	4	-

**The list may still be incomplete.*

Conclusions

TECHNISCHE UNIVERSITÄT DARMSTADT

- The very basics of version control are essential for teamwork and speed up individual work.
- Merge and prune branches periodically.
- Write top-level tests first; only if those fail, branch out deeper into the implementation.
- Separate concerns and implement single responsibility in your code.
- Format secondary data using the simple and open tabular CSV data/metadata format to simplify data analysis.
- Use the same automatic tests with coarse input for quick checks and with fine-grained input for production-level HPC simulations.
- Archive read-only snapshots of your source code and secondary data on a (TUdatalib) data repository,
- Cross-link digital research data artifacts: publication/report, scripts/code, secondary and primary data.

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Repository preparation

A minimal repository representing an exemplary "status quo".

2. Create a Docker image

Configure a reproducible testing environment.

3. Define CI pipeline through a YAML file

Define tests, how and when they are executed and what results to store.

4. Setup your own GitLab runner

Provide a machine for execution of tests.

TECHNISCHE UNIVERSITÄT DARMSTADT

Create your own copy of the example repository by forking:

- Log in to https://gitlab.com/.
- Go to https://gitlab.com/tmaric/minimal-cse-ci-examples.
- Click **fork** (upper right corner).
- Select a namespace, e.g. your personal one.
- Select either *Private* or *Public* as visibility level, both are fine.
- Click **Fork project**.
- Clone your fork on your machine:

```
?> git clone your-fork-URL
```

(Continuous) Integration of scientific software
Hands on: overview

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Repository preparation

A minimal repository representing an exemplary "status quo".

2. Create a Docker image

Configure a reproducible testing environment.

3. Define CI pipeline through a YAML file

Define tests, how and when they are executed and what results to store.

4. Setup your own GitLab runner

Provide a machine for execution of tests.

Hands on: install Docker I

Specific steps depend on your Linux distribution (Docker documentation)
Here for Ubuntu Focal:

1. `?> sudo apt-get update`
2. `?> sudo apt-get install apt-transport-https ca-certificates curl gnupg lsb-release`
3.
```
?> curl -fsSL https://download.docker.com/linux/ubuntu/gpg \
 | sudo gpg --dearmor -o /usr/share/keyrings/docker-archive-keyring.gpg
```
4.
```
?> echo \
"deb [arch=amd64 signed-by=/usr/share/keyrings/docker-archive-keyring.gpg] \
https://download.docker.com/linux/ubuntu \
$(lsb_release -cs) stable" | sudo tee /etc/apt/sources.list.d/docker.list > /dev/null
```
5. `?> sudo apt-get update`
6. `?> sudo apt-get install docker-ce docker-ce-cli containerd.io`

(Continuous) Integration of scientific software
Hands on: install Docker II

TECHNISCHE UNIVERSITÄT DARMSTADT

Check your Docker installation by running

- `?> sudo docker run hello-world`

The output should look as shown on the right.

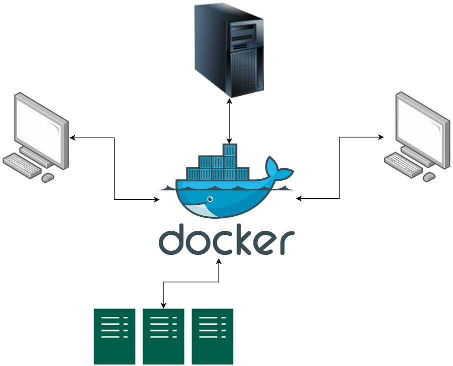
```
CSI\tolle@wmpc82:~$ sudo docker run hello-world

Hello from Docker!
This message shows that your installation appears to be working correctly.

To generate this message, Docker took the following steps:
 1. The Docker client contacted the Docker daemon.
 2. The Docker daemon pulled the "hello-world" image from the Docker Hub.
    (amd64)
 3. The Docker daemon created a new container from that image which runs the
    executable that produces the output you are currently reading.
 4. The Docker daemon streamed that output to the Docker client, which sent it
    to your terminal.

To try something more ambitious, you can run an Ubuntu container with:
 $ docker run -it ubuntu bash

Share images, automate workflows, and more with a free Docker ID:
 https://hub.docker.com/

For more examples and ideas, visit:
 https://docs.docker.com/get-started/

CSI\tolle@wmpc82:~$
```

In the **`minimal-cse-ci-examples`** repository

```
?> git checkout starting-point
?> git checkout -b feature/dockerfile
```

(Continuous) Integration of scientific software

Hands on: creating a Docker image I



- Docker images are computing environments that contain (**dependencies**) needed to build the research software, run simulations and process results.
- Sharing docker images removes the need to install the dependencies on different machines.
- The computing environment in a Docker image is usually based on an existing Linux distribution.
- The Docker image is built from a text file, that specifies installation steps for the dependencies, the so-called **Dockerfile**.

In a file named 'minimal-cse-ci-dockerfile_ubuntu-focal', write

```
FROM ubuntu:focal

# Set timezone
RUN apt-get update --fix-missing && \
    DEBIAN_FRONTEND="noninteractive" apt-get -y install tzdata
```

- We'll use Ubuntu 20.04 (focal) as the base system.
- Steps that are usually done manually (setting the timezone) are automated.

TECHNISCHE UNIVERSITÄT DARMSTADT

Dependency installation

```
# Install packages
RUN apt update && apt-get install --fix-missing -y \
    # Building
    build-essential cmake \
    # Version control
    git \
    # Python
    python3 \
    # Visualization
    python3-matplotlib python3-numpy \
    # Data analysis
    python3-pandas \
    # Test visualization
    jupyter-notebook jupyter-nbconvert \
    # Debugging the image
    vim
```

- **RUN** runs commands in the **Docker container**.
- The **Docker container** is a process spawned using the Docker image as the computing environment.
- Install the software needed for the scientific workflow (dependencies).

TECHNISCHE UNIVERSITÄT DARMSTADT

Software setup

```
## Default Ubuntu to python3
RUN update-alternatives --install \
    /usr/bin/python python /usr/bin/python3 10
```

Some specifics

- Alternative (**g++**) compiler.
- Alternative working directory.

(Continuous) Integration of scientific software

Hands on: creating a Docker image V

TECHNISCHE UNIVERSITÄT DARMSTADT

Complete Dockerfile for the minimal example

```
FROM ubuntu:focal

# Set timezone
RUN apt-get update --fix-missing && \
    DEBIAN_FRONTEND="noninteractive" apt-get -y install tzdata

# Install packages
RUN apt update && apt-get install --fix-missing -y \
    # Building
    build-essential cmake \
    # Version control
    git \
    # Python
    python3 \
    # Visualization
    python3-matplotlib python3-numpy \
    # Data analysis
    python3-pandas \
    # Test visualization
    jupyter-notebook jupyter-nbconvert \
    # Debugging the image
    vim

## Default Ubuntu to python3
RUN update-alternatives --install \
    /usr/bin/python python /usr/bin/python3 10
```

- The example Dockerfile installs all dependencies for the minimal example on Ubuntu 20.04.
- The installation commands would be different for another operating system.
- A more complex software (e.g. OpenFOAM) requires a larger Dockerfile.
- This lets us define the computing environments that are supported by the research software.

TECHNISCHE UNIVERSITÄT DARMSTADT

Building the image

```
%?> sudo docker build . \
    %-f minimal-cse-ci-dockerfile_ubuntu-focal \
    %-t minimal-cse-ci-dockerfile_ubuntu-focal
    sudo docker build . -f minimal-cse-ci-dockerfile_ubuntu-focal -t minin
```

- ".": current directory
- "-f" name of the Dockerfile (defaults to "Dockerfile")
- "-t" tag (name) of the Docker image

Listing Docker images

```
?> sudo docker image list
REPOSITORY                              TAG    IMAGE   ID   CREATED       SIZE
minimal-cse-ci-dockerfile_ubuntu-focal latest 921233ec4b44 9 minutes ago 982MB
```

- The image is built on the machine (host) where the **`docker build`** command is called.
- Docker uses a so-called **image registry** to store images.
- For Continuous Integration the images are built on the machine where the tests are run or shared on **Dockerhub**.

"**Spinning a container**" (running a Docker image)

```
?> sudo docker run -it minimal-cse-ci-dockerfile_ubuntu-focal /bin/bash
root@b2c14ee0fd58:/# ls
bin   boot   dev   etc   home   lib   lib32   lib64   libx32   media   mnt
opt   proc   root   run   sbin   srv   sys   tmp   usr   var
root@b2c14ee0fd58:/# cd
root@b2c14ee0fd58:~# pwd
/root
```

- The container behaves just like a "regular" Ubuntu.
- **Jobs (test) commands for the Continuous Integration are checked/debugged inside a running container.**
  - Forgot to install a dependency.
  - The research software does not compile with installed dependencies.
  - ...

Working within the container : compiling the software

```
?> git clone https://gitlab.com/tmaric/minimal-cse-ci-examples.git
?> cd minimal-cse-ci-examples && mkdir build && cd build
?> cmake .. && make
?> ./myapp
```

The same steps will be done in the Docker container by the Continuous Integration

- Clone the repo.
- Build the software.
- Run the tests.

Analyzing the data using Jupyter notebooks

```
?> cd ..
?> jupyter nbconvert --execute mynotebook.ipynb --to html
```

- On the cluster, one would start the Jupyter notebook server and connect to it locally.
- Here the notebook is used to process the results and visualize secondary data as tables and diagrams.

Extracting the data from the container:

- Find the ID of the container you're on (execute on your machine)

```
sudo docker ps
```

- Copy the results from the container onto the local machine (execute on your machine)

```
mkdir container-data
sudo docker cp f2dff55edf7a:/root/minimal-cse-ci-examples \
    container-data/
```

- Examine the data and the Jupyter notebook in a browser.

Note: the sequence f2dff55edf7a is system dependent ID, so it'll be different for you.

(Continuous) Integration of scientific software

Hands on: creating a Docker image XII

TECHNISCHE UNIVERSITÄT DARMSTADT

- Saving the container or an image as a tar file

```
sudo docker commit f2dff55edf7a test:latest
```

- You can exit/close the container by pressing Ctrl+d.
- View the newly create image with 'name:tag' using

```
sudo docker image list

REPOSITORY TAG     IMAGE ID      CREATED             SIZE
test       latest  f2dff55edf7a  About a minute ago  983MB
```

- Save the image into a tar file

```
sudo docker save test:latest -o container-archive.tar
```

- Load an image into Docker's registry to work with it

```
    sudo docker load -i container-archive.tar
```

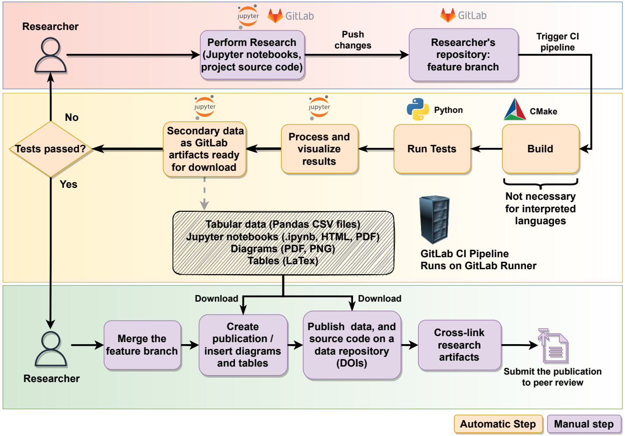
Hands on: creating a Docker image XIII

TECHNISCHE UNIVERSITÄT DARMSTADT

- Usually, the Docker image "lives" locally on the test machine.
- However, it can also be shared publicly on Dockerhub, for example (don't do this now)

```
?> docker login
?> docker tag name:tag username/name:tag
?> docker push username/name:tag
```

- This image can now be used by everyone.
- Note: once you exit/stop a container all data/files created inside the container are discarded.



All the steps done so far manually using Docker, namely,

1. **building** the scientific software,
2. **running tests**,
3. **processing data**
4. **exporting** the data and Jupyter notebooks,

are automated by Continuous Integration, that uses Docker for encapsulating the computing environment.

(Continuous) Integration of scientific software

Hands on: overview

1. Repository preparation

A minimal repository representing an exemplary "status quo".

2. Create a Docker image

Configure a reproducible testing environment.

3. Define CI pipeline through a YAML file

Define tests, how and when they are executed and what results to store.

4. Setup your own GitLab runner

Provide a machine for execution of tests.

TECHNISCHE UNIVERSITÄT DARMSTADT

In the **`minimal-cse-ci-examples`** repository

```
?> git checkout added-dockerfile
?> git checkout -b feature/enable-ci
```

(Continuous) Integration of scientific software

Hands on: enabling CI for a GitLab project I

TECHNISCHE UNIVERSITÄT DARMSTADT

- Adding the **`.gitlab-ci.yml`** file your project and pushing the change to the GitLab remote repo configures the CI pipeline.
- The YAML file specifies the Docker image that is used for testing

```yaml
image: "tmaric/minimal-cse-ci:ubuntu-focal"

stages:
  - building
  - running
  - visualization
```

- and the so-called job **stages**: collections of jobs for building, running tests and visualization.
- For example, the **building** stage may multiple jobs, building the software for
  - production,
  - debugging,
  - performance measurements.

Hands on: enabling CI for a GitLab project II

- The building stage in the YAML file defines build jobs like this one

```
build:
  stage: building
  script:
      - git clone https://gitlab.com/tmaric/minimal-cse-ci-examples.git
      - cd minimal-cse-ci-examples && mkdir build && cd build
      - cmake ..
      - make

  artifacts:
    paths:
        - minimal-cse-ci-examples/mynotebook.ipynb
        - minimal-cse-ci-examples/build/myapp
```

- where the repository is cloned and built with specific options.
- For example **`cmake -DCMAKE_BUILD_TYPE=Debug`** can set up the build for debugging.
- **artifacts** are downloadable files passed on to other jobs.

(Continuous) Integration of scientific software
Hands on: enabling CI for a GitLab project III

TECHNISCHE UNIVERSITÄT DARMSTADT

- The running stage in the YAML file defines how simulations (studies) run

```yaml
param_study:
  stage: running
  dependencies:
    - build

  script:
    - cd minimal-cse-ci-examples/build && ./myapp

  artifacts:
    paths:
        - minimal-cse-ci-examples/mynotebook.ipynb
        - minimal-cse-ci-examples/build/myapp
        - minimal-cse-ci-examples/build/poly-data.csv
```

- Without a successful **build**, simulations do not run.
- Here the artifacts are the secondary data and the notebooks that visualize them.

(Continuous) Integration of scientific software

Hands on: enabling CI for a GitLab project IV

- The **visualization** stage in the YAML file saves time by converting Jupyter notebooks

```yaml
convert_notebooks:
  stage: visualization
  dependencies:
    - param_study

  script:
    - cd minimal-cse-ci-examples
    - jupyter nbconvert mynotebook.ipynb --execute --to html

  artifacts:
    paths:
        - minimal-cse-ci-examples/mynotebook.*
        - minimal-cse-ci-examples/build/myapp
        - minimal-cse-ci-examples/build/polydata.csv
```

- HTML is easiest, other formats are available (PDF, markdown,...).
- HTML notebooks can be viewed in the browser.

(Continuous) Integration of scientific software

Hands on: enabling CI for a GitLab project V

Final **.gitlab-ci.yml** file.

Lessons learned I

- Defining artifacts path starts at **`your-project/`**.
- YAML files require debugging:
  - syntax: use GitLab's CI Lint tool
  - everything else: the only way to do this effectively is to commit changes and push them upstream.
- It is possible to partially debug locally using

  ```
  gitlab-runner exec docker job-name
  ```

  but this does not work with artifacts and dependencies.

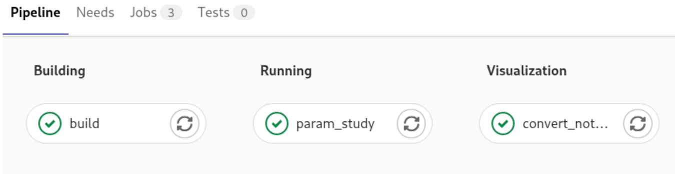
TECHNISCHE UNIVERSITÄT DARMSTADT

Final **.gitlab-ci.yml** file.

Lessons learned II

- Generally, and for the CI, scripts that reproduce data without requiring input for the users speed up work.
  `simulation-directory > ./reproduce-density-ratio-data`
- It takes time to set up the CI, but it pays off in debugging time as problems are found automatically.
- Exporting **`*.ipynb`** jupyter notebooks and their data

The CI pipeline of the Minimal Working Example (MWE) repository



(Continuous) Integration of scientific software

Hands on: overview

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Repository preparation

A minimal repository representing an exemplary "status quo".

2. Create a Docker image

Configure a reproducible testing environment.

3. Define CI pipeline through a YAML file

Define tests, how and when they are executed and what results to store.

4. Setup your own GitLab runner

Provide a machine for execution of tests.

An incomplete comparison:

Self-managed runner

- No shared runners available
- Provided CI/CD minutes of plan insufficient (e.g. 400 min per month for GitLab's free plan)
- Control over hardware and runner configuration

Shared runners

- No need to provide "always on" hardware
- No need for maintenance

Overall: require less of your time

Using instructions from GitLab documentation for Ubuntu:

1. `?> curl -L \ "https://packages.gitlab.com/install/repositories/runner/gitlab-runner/script.deb.sh" \ | sudo bash`
2. `?> sudo apt-get install gitlab-runner`

Check the status of the runner:

- `?> sudo systemctl status gitlab-runner.service`

The output should indicate that it is active:

```
CSI\tolle@wmpc82:~$ sudo systemctl status gitlab-runner.service
● gitlab-runner.service - GitLab Runner
     Loaded: loaded (/etc/systemd/system/gitlab-runner.service; enabled; vendor preset: enabled)
     Active: active (running) since Mon 2021-09-20 15:59:03 CEST; 20h ago
   Main PID: 327927 (gitlab-runner)
      Tasks: 37 (limit: 154341)
     Memory: 15.3M
     CGroup: /system.slice/gitlab-runner.service
             └─327927 /usr/bin/gitlab-runner run --working-directory /home/gitlab-runner --config
```

Follow GitLab documentation on how to register a runner:

- Obtain a token for project-specific runner: go to your fork of the *minimal-cse-ci-examples* on gitlab.com and then to **Settings > CI/CD** and expand the **Runners** sections.
- There you find a section **Specific runners** and aforementioned token.

Register your runner (instructions for Linux):

- `?> sudo gitlab-runner register`

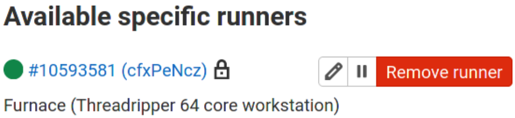
You need to provide some information regarding your runner, e.g. your project's token. See next slide.

(Continuous) Integration of scientific software

Hands on: setup a GitLab runner IV (register GitLab runner)

Option	Value
GitLab instance URL	https://gitlab.com/
Token	Obtained from the project on GitLab, see previous slide
Runner description	describe the machine used as runner, useful to distinguish multiple runners
Tags	leave empty, not required here. Useful for advanced pipelines
Runner executor	docker (see here for comparison of executors.)
Default image	Because we chose docker as executor: name of the default Docker image

You should now see your runner under *Available specific runners*:



Hands-on part

From the course description

- Participants will apply concepts to their own research projects.
- Course content is version-control agnostic; examples use GitLab.
- Participants should have GitLab accounts (gitlab.com) and bring laptops.
- For examples, a working Python environment (e.g., venv, miniconda, or conda) is required.
- Required Python packages: numpy, matplotlib, pandas, pytorch, scikit-learn, jupyter, jupyter notebook.

Bring your own project
My project(s)

TECHNISCHE UNIVERSITÄT DARMSTADT

github.com/tmaric/TwoPhaseFlow

- Merge: feature/non-orthogonality, feature/density-ratio, feature/wetting, with main branch.
- Cross-link the papers in the main branch with the README.md
- Update the compilation for OpenFOAM-v2412.
- Investigate GitHub actions.
- Implement GitHub actions for selected tests.

github.com/bosh/sepMultiPhaseFlow

- Extend the jupyter notebook to easily add results to the benchmark.
- Add results from other groups to the benchmark.

gitlab.com/tmaric/fvcreconstruct

- Defect Correction reconstruction algoritm which re-uses Finite-Volume error estimates.
- Extend the jupyter notebook for the DEC algorithm.
- Extend CI for the DEC algorithm.

Bring your own project
Your project(s)

TECHNISCHE UNIVERSITÄT DARMSTADT

This is a basics session - simple workflow

- Add "git" version-control to your project.
- Design a test application that report scientific results in your paper.
- Make a test application take coarse input and fine input.
- Create a remote repository on gitlab.com
- Push your local repository to the remote repository.
- Create a gitlab CI pipeline for your remote repository.
- Run the pipeline.
- Call the test application in the CI pipeline.

Bring your own project

Dont' have a project?

Project: function approximation with a deep neural network.

https://shorturl.at/kKauw

Acknowledgements

TECHNISCHE UNIVERSITÄT DARMSTADT

CRC 1194 **Interaction between Transport and Wetting Processes**

Funded by the German Research Foundation (DFG) – Project-ID 265191195 – CRC 1194 : Z-INF

NFDI4ing

Funded by the NFDI4Ing consortium: German Research Foundation (DFG) - project number 442146713.