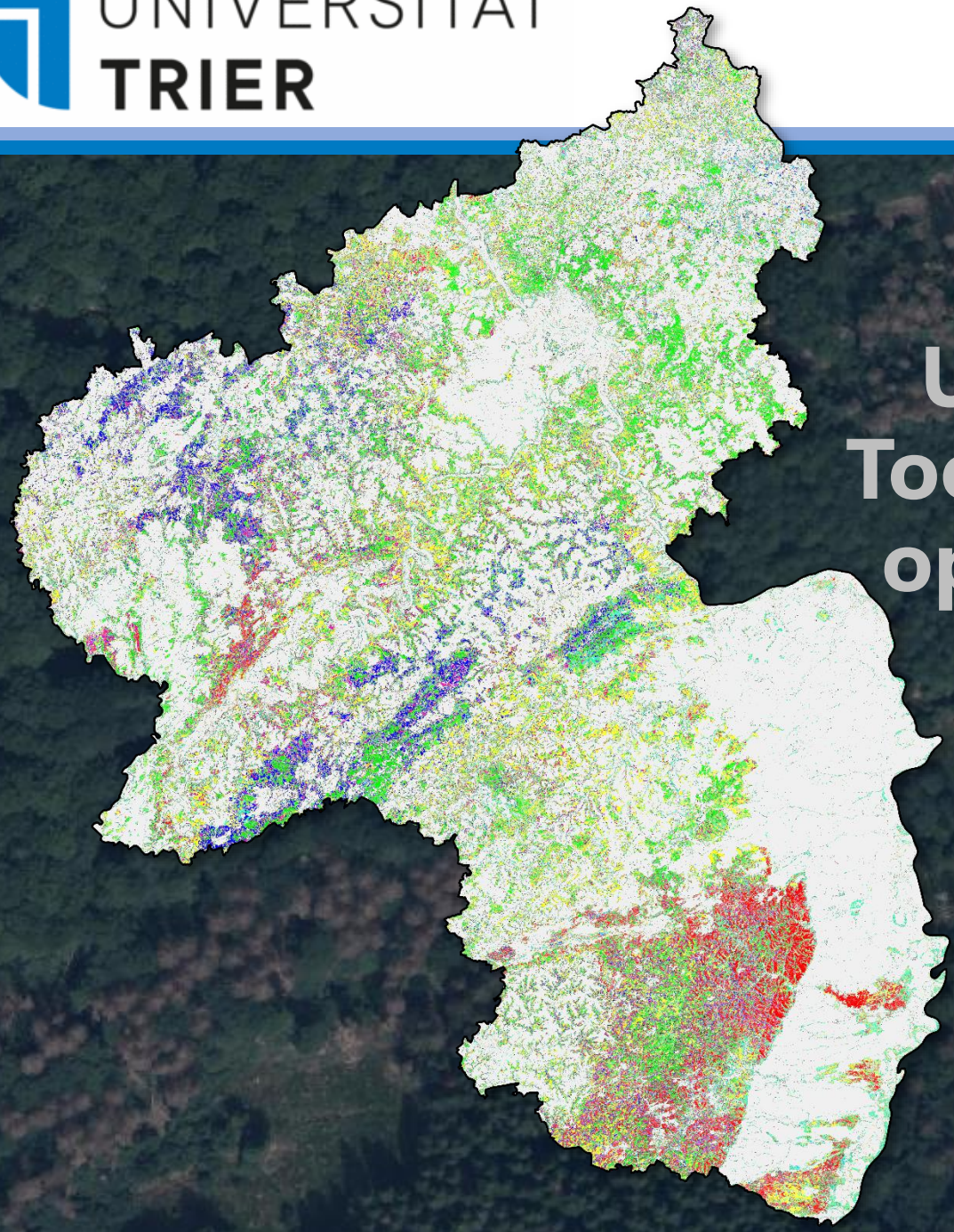
UNIVERSITÄT TRIER

Geoinformatics Spatial Data Science

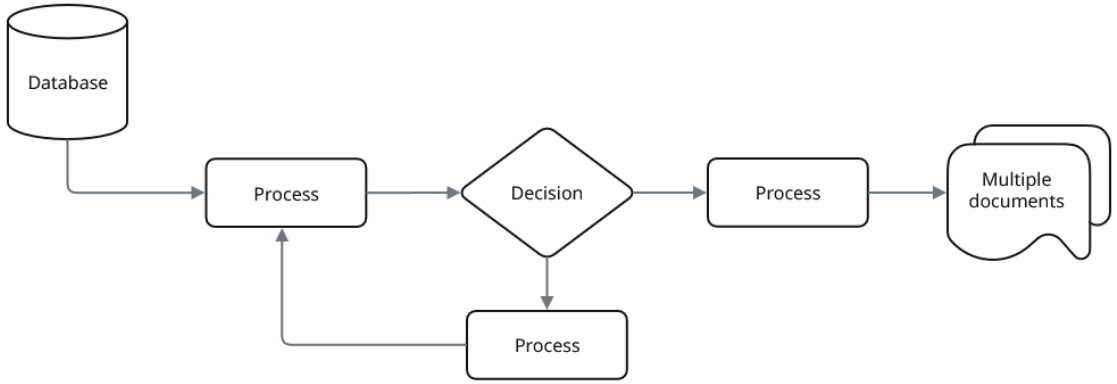


Using Workflow Management Tools for consistent generating of operational EO forestry product

AK Fernerkundung 2025 | 29.09 – 30.09

What is a Workflow?

"A workflow is a structured sequence of computational tasks or activities that achieve a research or analytical objective. Workflows define the flow of work, including the order of steps, the data and control dependencies between them, and the rules governing their execution" [Suter et al. 2025]



Why use Workflow Management Systems?

- Guaranteed consistency, reproducibility and portability
  - To overcome difficulties on replicating experiments
  - Next steps after proofs of concept
- Findability, Accessibility, Interoperability, and Reusability (FAIR)
- Proper handling of resources
  - Efficient alternative to the usage of HPCs
  - More sensible and sustainable

Workflow Management Systems

and many more…

https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems

- Inputs and Outputs are handled as **Channels**, which are then fed to **Workflows** and **Processes**
- Any programming language and algorithm can be wrapped inside a Process, as long as there is a **command line interface** and the right **environment**
- These processes may require their own set of dependencies and libraries, which can be provided and run via the usage of containers like docker.
- Can be run on a personal computer or a HPC, and supports with most execution backends
- Smart caching to avoid constant execution repetition

How does Nextflow work?

UNIVERSITÄT TRIER

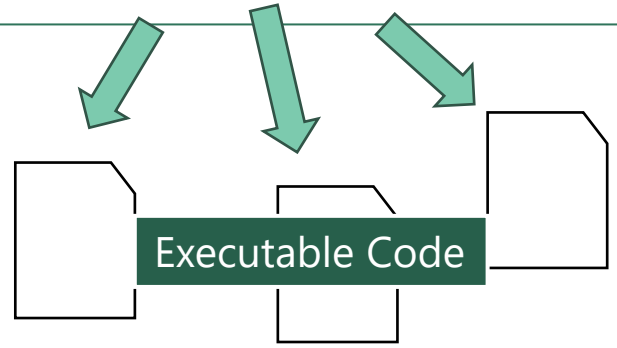
Process Samples
- Curated Samples

Generate Libraries
- Synthetic Libraries

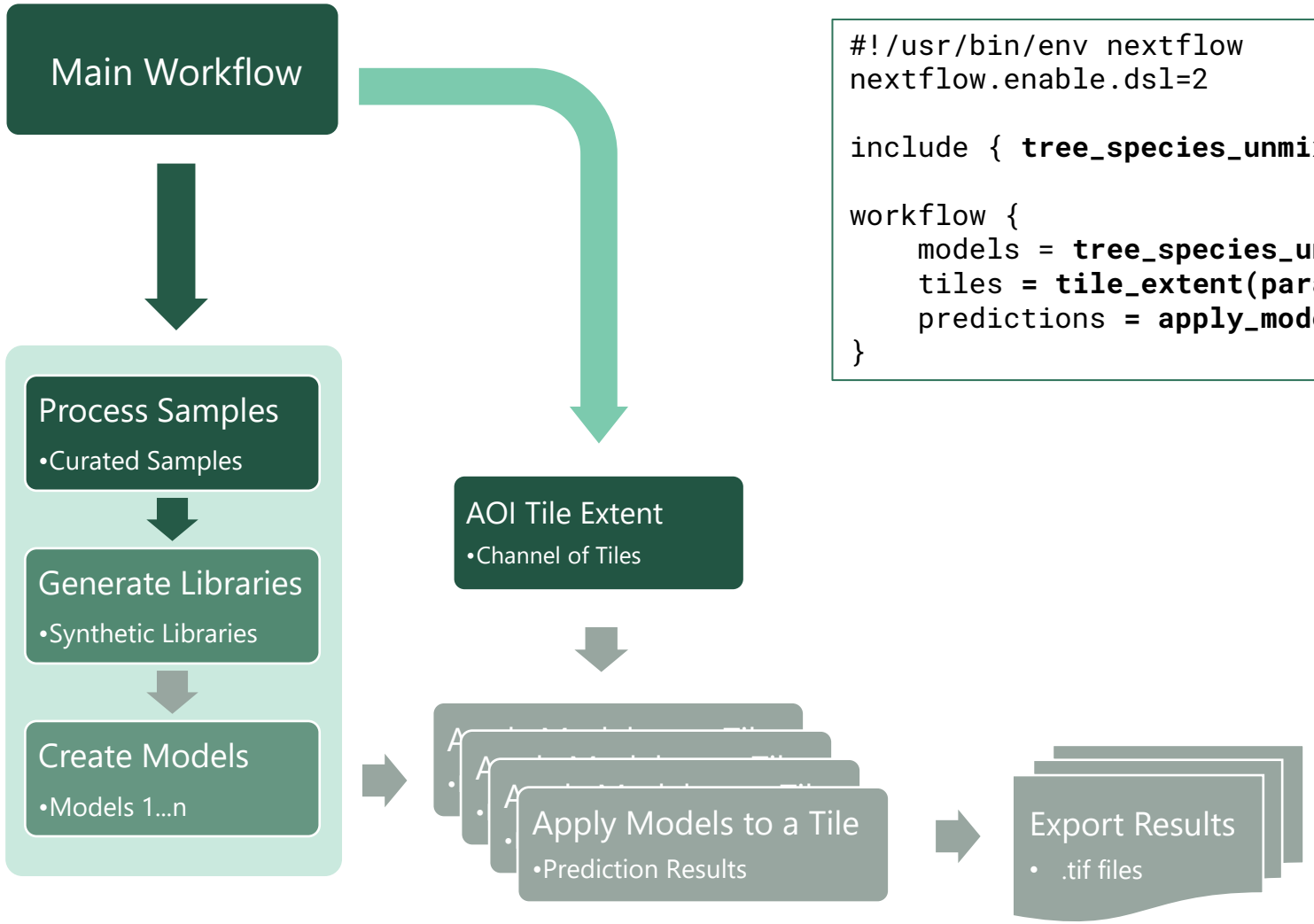
Create Models
- Models 1...n

```
#!/usr/bin/env nextflow
nextflow.enable.dsl=2

// Can and should be placed in a configuration file
params.datacube = 'path/to/datacube'
params.points = 'path/to/training_points'
params.year = 2020

workflow tree_species_unmixing {
    take:
    working_dir

    main:
    models = process_samples(working_dir)
    | generate_synthetic_libraries
    | train_ANN

    emit:
    models
}
```

```
process process_samples {
    label 'tree-species' //assign a container

    input:
    path working_dir

    output:
    path "${working_dir}/*"

    script:
    """
    python extract_pure_samples.py \
      --dc_folder ${params.datacube} \
      --training_points ${params.points} \
      --year ${params.year} \
      --working_directory ${files}
    """
}
```



How does Nextflow work?




```
#!/usr/bin/env nextflow
nextflow.enable.dsl=2

include { tree_species_unmixing; apply_model } from './some/module'

workflow {
    models = tree_species_unmixing()
    tiles = tile_extent(params.aoi) //returns a Channel
    predictions = apply_model(models, tiles) //iterates over a Channel
}
```

```
process apply_model {
    input:
    path model_directory
    path tile

    publishDir ${params.output_folder}

    output:
    path "prediction/*"

    script:
    """
    python apply_models.py \
      --tile ${tile} \
      --models ${model_directory} \
      --output "./prediction/"
    """
}
```

How can it be adapted?

- A Workflow management system focuses on distributing tasks on their adequate environments, while letting the algorithms run its intended scopes (i.e. for each cell)

- If a cell calculation “fails”, results from the other cells and processes are still cached and usable for the next iteration

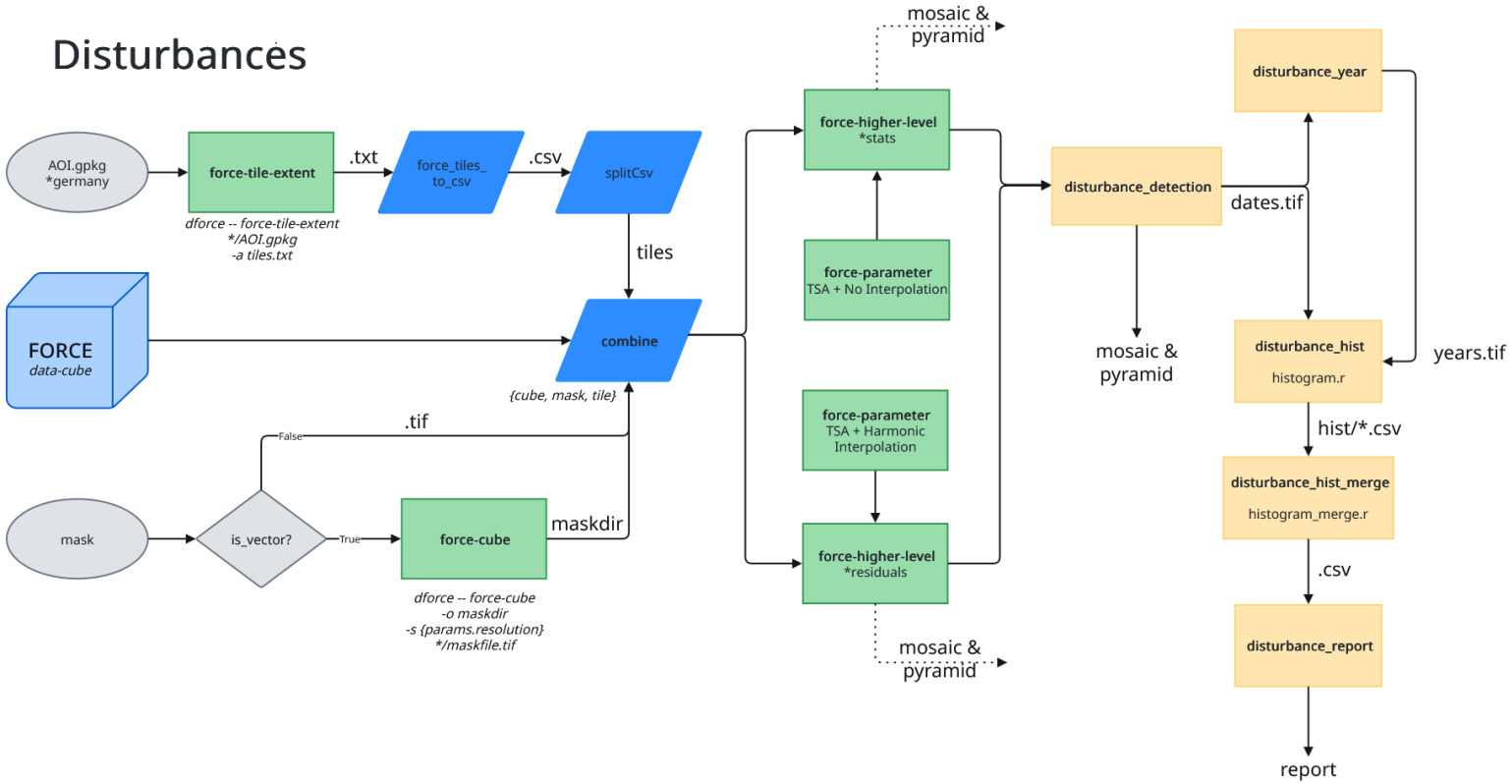
ForestPulse

precision data in sync with the forest's heartbeat

https://forestpulse.thuenen.de/

Inside the Project

- One repository for each package/algorithms
- Containers will be created from Dockerfile(s) found in the repositories
- Nextflow files wrap scripts and tools (e.g. FORCE or gdal)
- Said files will call each step of the algorithms via command line
- Process Analysis will be performed to identify redundant tasks



Why Containers?

- To encapsulate running environments for reproducibility and portability
- Should contain all required dependencies + repository code

```
FROM python:3.12-slim

ENV PYTHONDONTWRITEBYTECODE=1 \
    PYTHONUNBUFFERED=1 \
    CPLUS_INCLUDE_PATH=/usr/include/gdal \
    C_INCLUDE_PATH=/usr/include/gdal

# Install only runtime system dependencies
RUN apt-get update && apt-get install -y --no-install-recommends \
    gdal-bin \
    libgdal-dev
    && apt-get clean && rm -rf /var/lib/apt/lists/*

# Install Python dependencies
RUN pip install --no-cache-dir -r requirements.txt

# Set working directory and add app
WORKDIR /app
COPY . .

# Ensure scripts are executable
RUN find . -name "*.py" -exec chmod +x {} \;

# Add app src to PATH
ENV PATH="/app/src:$PATH"

# Default command
ENTRYPOINT ["/bin/bash", "-c"]
CMD ["sample.py"]
```

Running ForestPulse

UNIVERSITÄT TRIER

```
eouser@compute-node-1:~/Workflow$ nextflow run -resume main.nf
Nextflow 25.09.0-beta is available - Please consider updating your version to it

 N E X T F L O W   ~  version 25.04.2

Launching `main.nf` [infallible_hirsch] DSL2 - revision: 9bb2ae9ae9

[47/c08f1c] treed_mask:force_get_tiles:force_tile_extent (1)          [100%] 1 of 1, cached: 1 ✔
[d0/51f143] treed_mask:force_get_tiles:force_tiles_to_csv (1)         [100%] 1 of 1, cached: 1 ✔
[6e/cffdd4] treed_mask:fold_TSA_labels:force_parameter                [100%] 1 of 1, cached: 1 ✔
[ee/37848d] treed_mask:fold_TSA_labels:fill_parameter_labels (532)    [100%] 535 of 535, cached: 535
[31/f9e98a] treed_mask:fold_TSA_labels:force_higher_level_chain (518) [100%] 519 of 519, cached: 519
[e2/a9e890] treed_mask:obtain_samples:force_parameter                 [100%] 1 of 1, cached: 1 ✔
[a8/beb02d] treed_mask:obtain_samples:fill_parameter_stats (494)      [100%] 494 of 494, cached: 494
[a6/9f6ad7] treed_mask:obtain_samples:force_higher_level_chain (476)  [100%] 479 of 479, cached: 479
[-        ] treed_mask:concatenateFiles                               -
[-        ] treed_mask:split_samples                                  -
[-        ] treed_mask:augment                                        -
[-        ] treed_mask:aggregate_weekly                               -
[-        ] treed_mask:train                                          -
[69/f91f9f] treed_mask:fold_TSA_aoi:force_parameter                   [100%] 1 of 1, cached: 1 ✔
[75/14fdcb] treed_mask:fold_TSA_aoi:fill_parameter_aoi (525)          [100%] 527 of 527, cached: 527
[01/1ea189] treed_mask:fold_TSA_aoi:force_higher_level_chain (511)    [100%] 511 of 511, cached: 511
[-        ] treed_mask:predict                                        -
[-        ] treed_mask:merge_masks                                    -
```

Potential Drawbacks

- Misconfigured WMS may overload the host computer
  - Pairing it with another manager like Kubernetes is a sensible option
- Investigating into the working directory to detect artifacts and erroneous by-products can be complicated
- Steep entry barrier in terms of configuration of environments and learning the native operators and operations.
- There’s plenty of already tried and tested use cases for Bioinformatics (nf-core), such endeavor is still not present in regards to Earth Observation

Questions?

UNIVERSITÄT TRIER

Contact
Juan Rodrigo Velarde Jara
jara@uni-trier.de

Project Webpage
https://forestpulse.thuenen.de/

Open Development
https://github.com/ForestPulse

Gefördert durch:

Bundesministerium
für Digitales
und Verkehr