eurac research

SFSCON

WWW.SFSCON.IT

NOI TECHPARK SOUTH TYROL

Embracing CI/CD workflows for building ETL pipelines

how we will gather and monitor multi-source spatially-interpolated meteorological parameters in near-real time

Elena Maines

11.11.23

Who

ELENA MAINES

Junior Researcher

PIERO CAMPALANI

Software Engineer

ALICE CRESPI

Senior Researcher

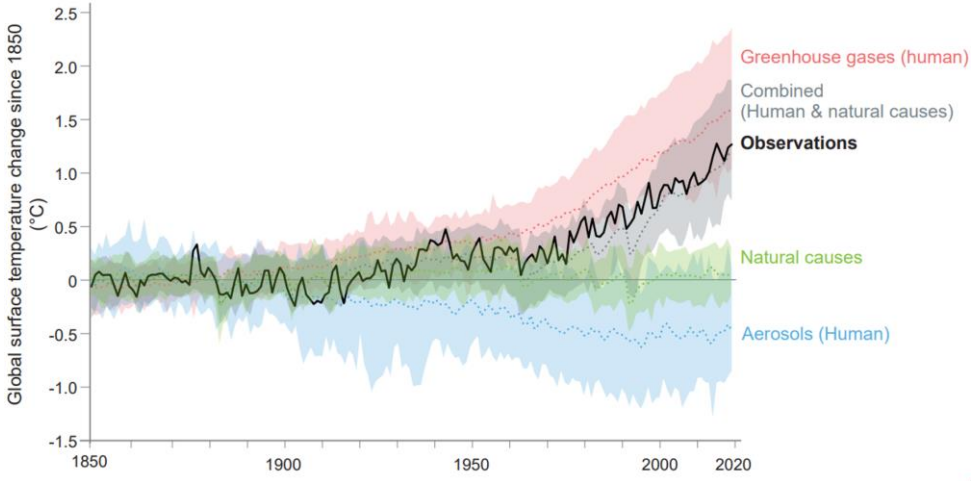
Pipeline: motivation

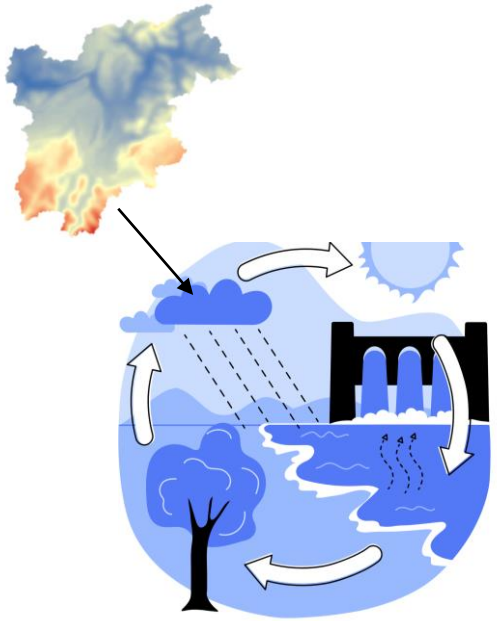
Having **reliable and up-to-date long-term meteorological data** is crucial for

Monitoring current meteorological conditions at regional level

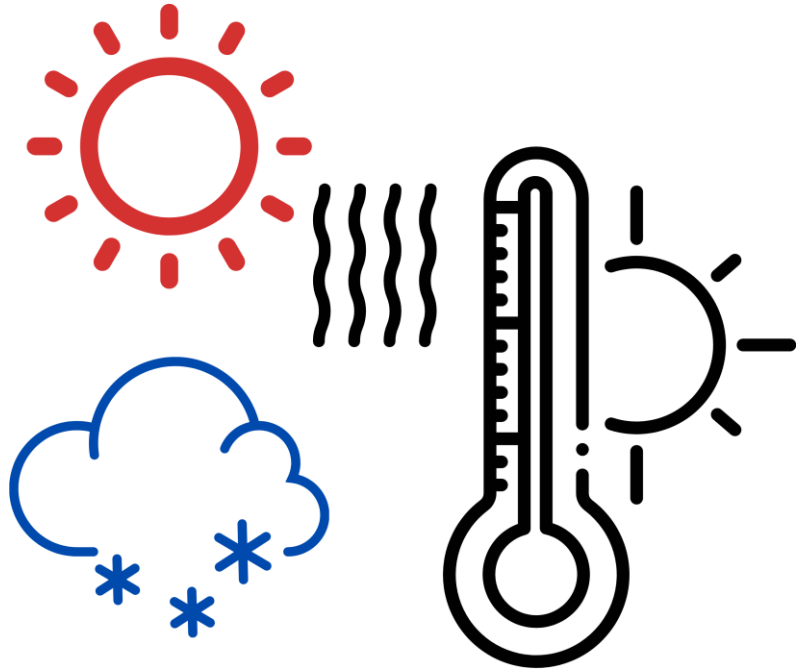
Monitoring **climate variability** at local level

Providing spatially distributed meteorological inputs for **impact models** (e.g., for hydrological simulations)

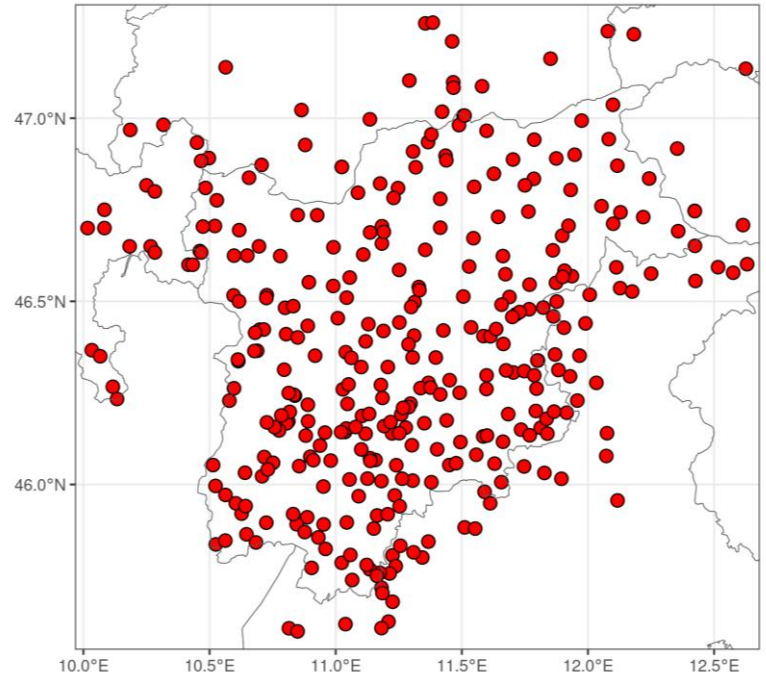




Meteorological data and data preparation

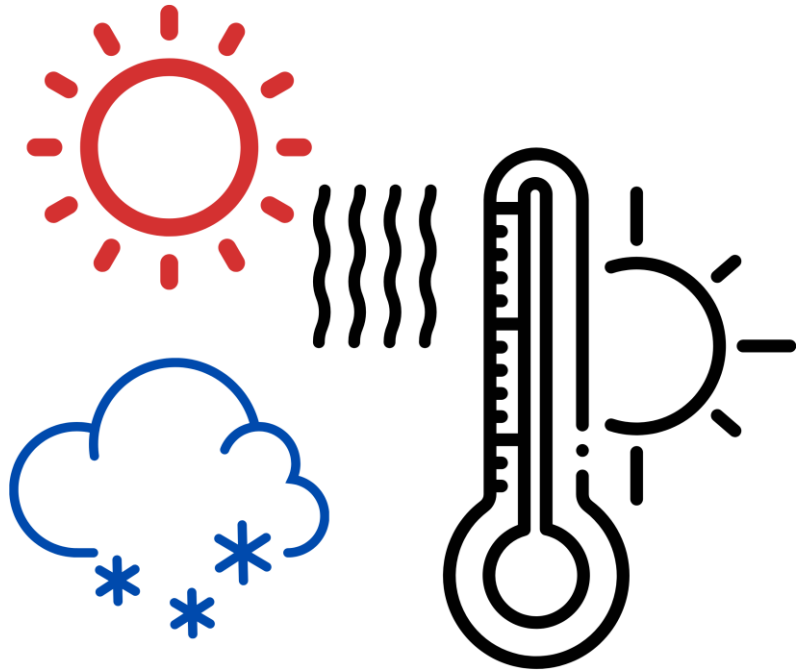


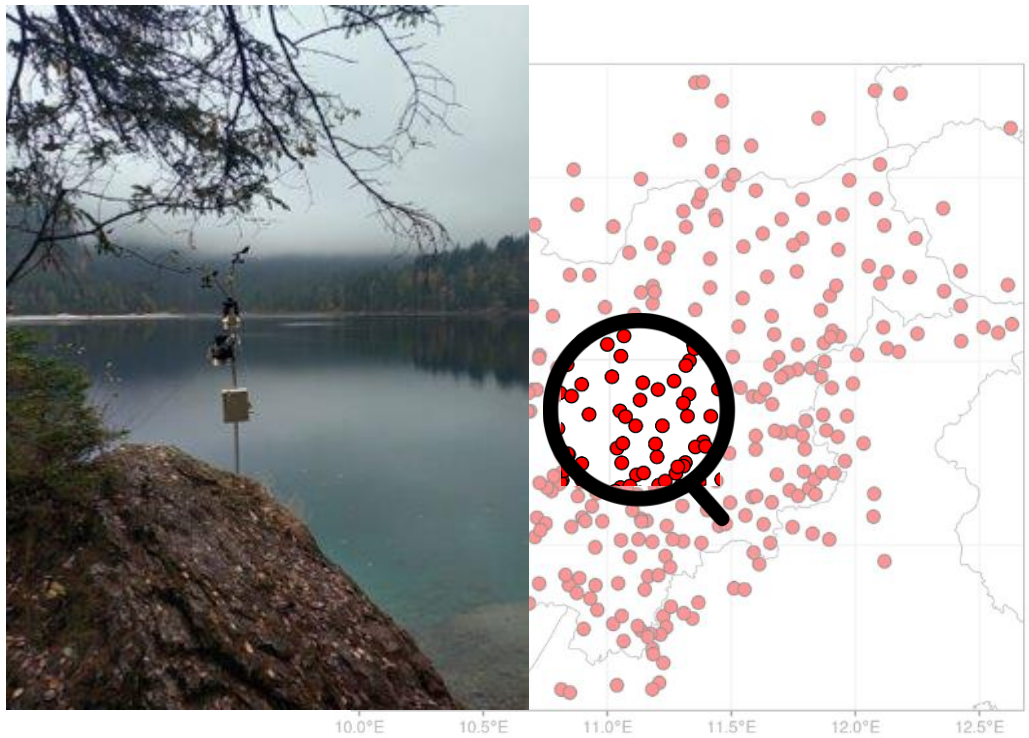
Time series of key weather variables: e.g., precipitation, min/max temperatures.



These data are recorded at daily and sub-daily resolution by weather stations

Meteorological data and data preparation

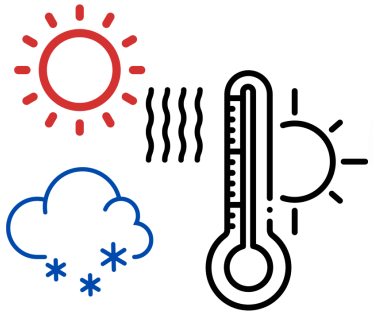




Time series of key weather variables: e.g., precipitation, min/max temperatures.

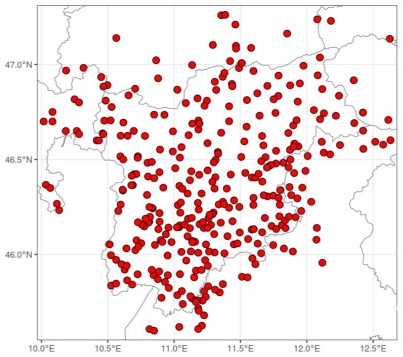
These data are recorded at daily and sub-daily resolution by weather stations

Meteorological data and data preparation



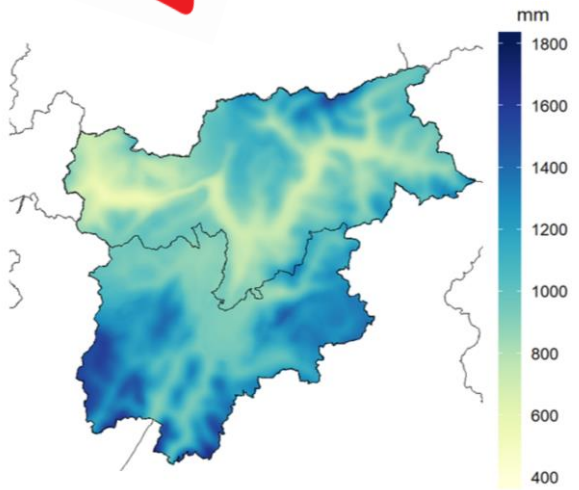
Comprehensive and updated time-series dataset: a robust compilation of at-site meteorological conditions

PostgreSQL



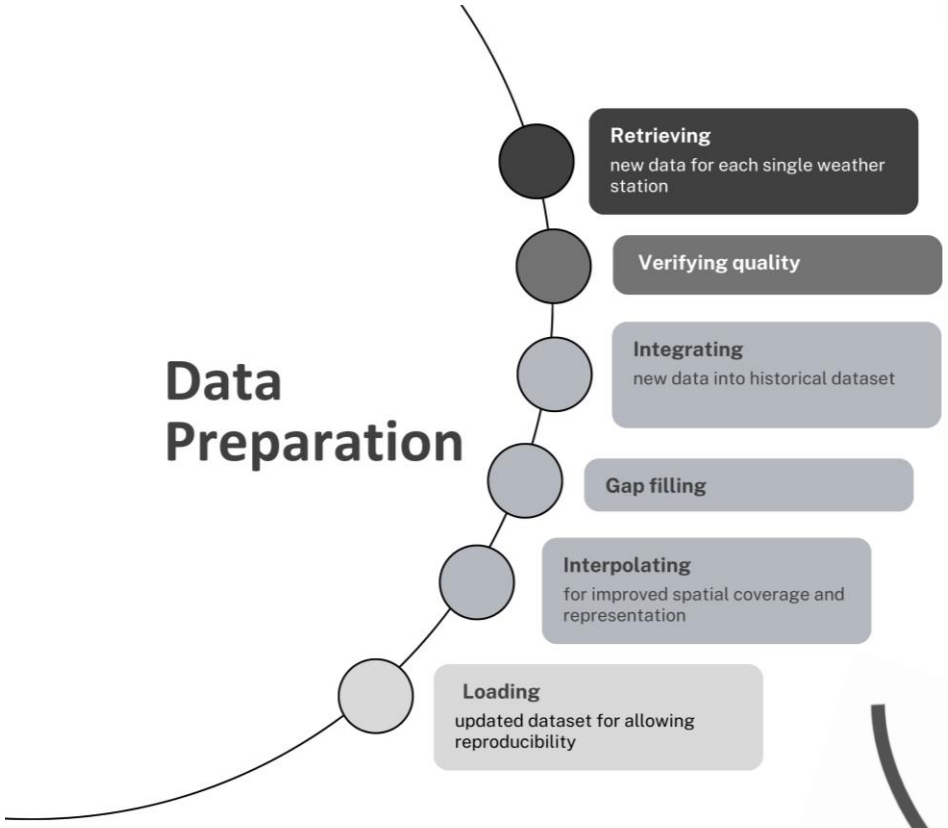
Irregular spatial coverage of weather stations

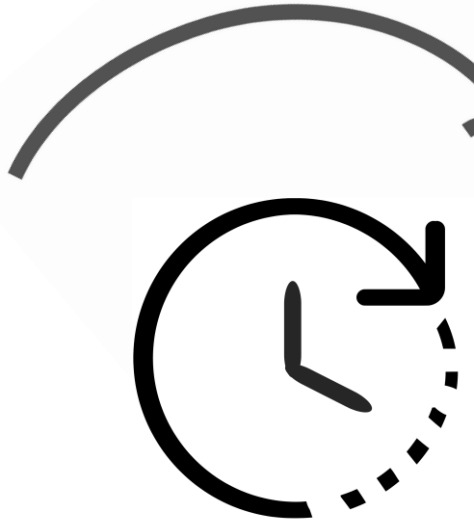
Interpolation for obtaining a gridded dataset



open EO

Pipeline: motivation



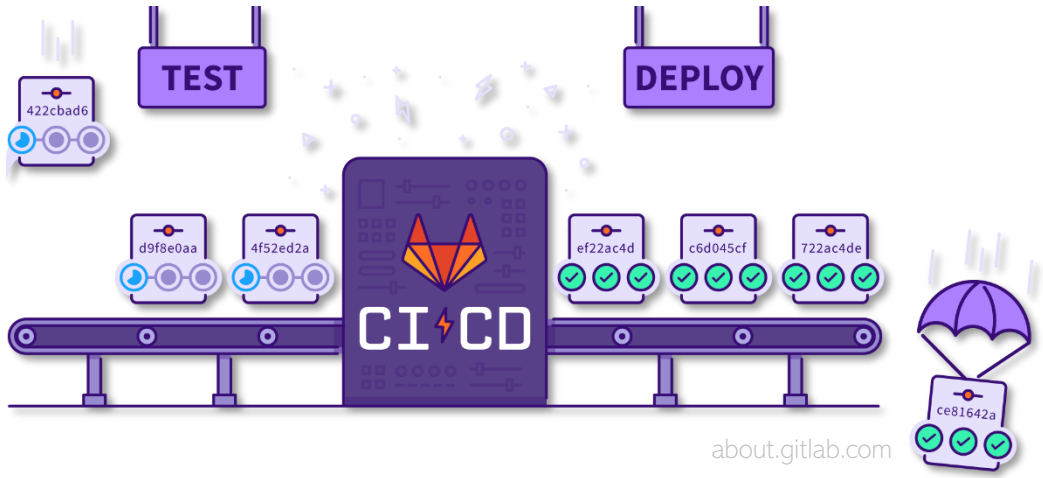


Focus on analysis

Accurate data

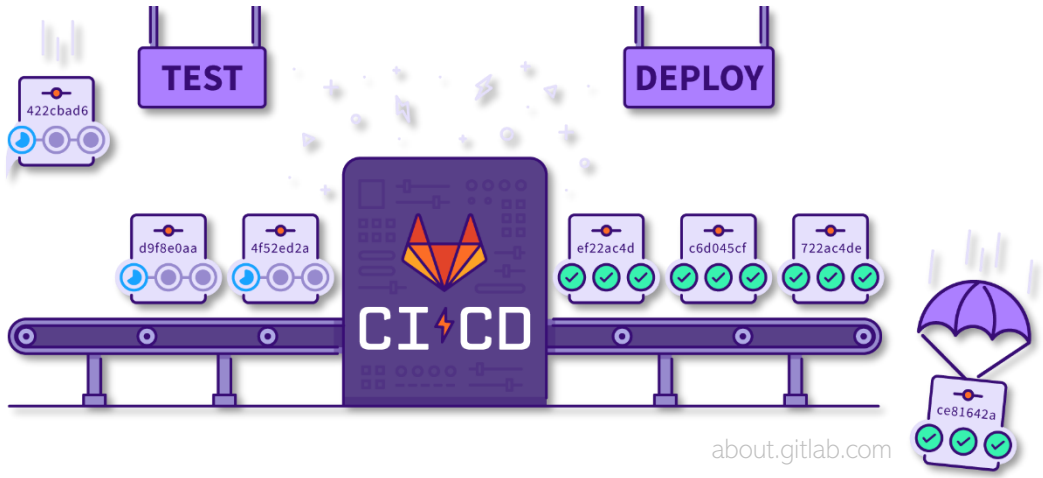
Pipeline: implementation

- **GitLab** CI/CD pipeline



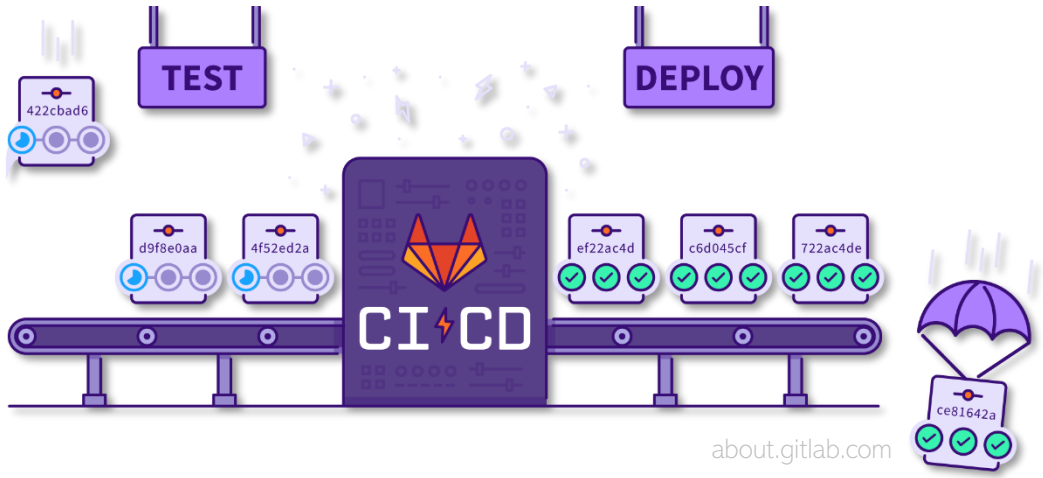
Pipeline: implementation

- **GitLab** CI/CD pipeline
- **Dockerized** jobs runners



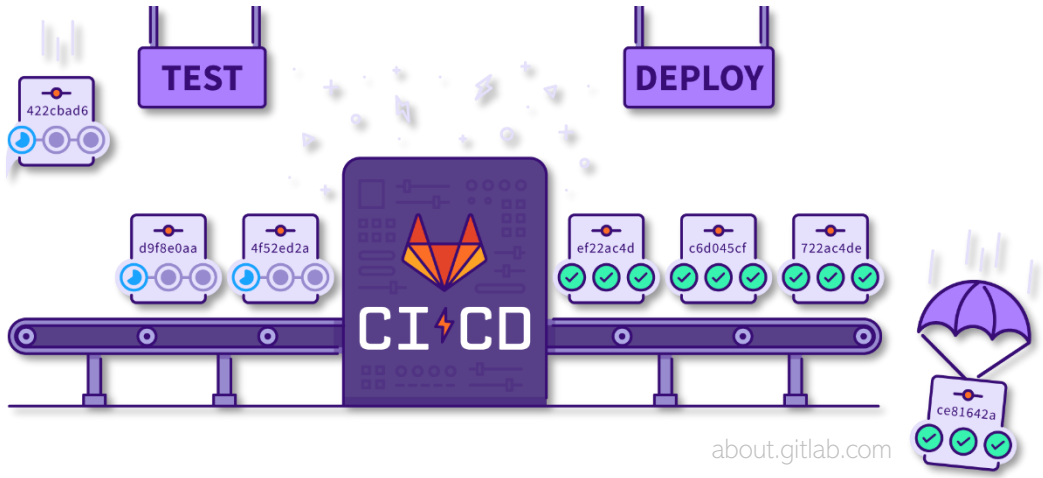
Pipeline: implementation

- **GitLab** CI/CD pipeline
- **Dockerized** jobs runners
- R modules + geo-spatial libraries



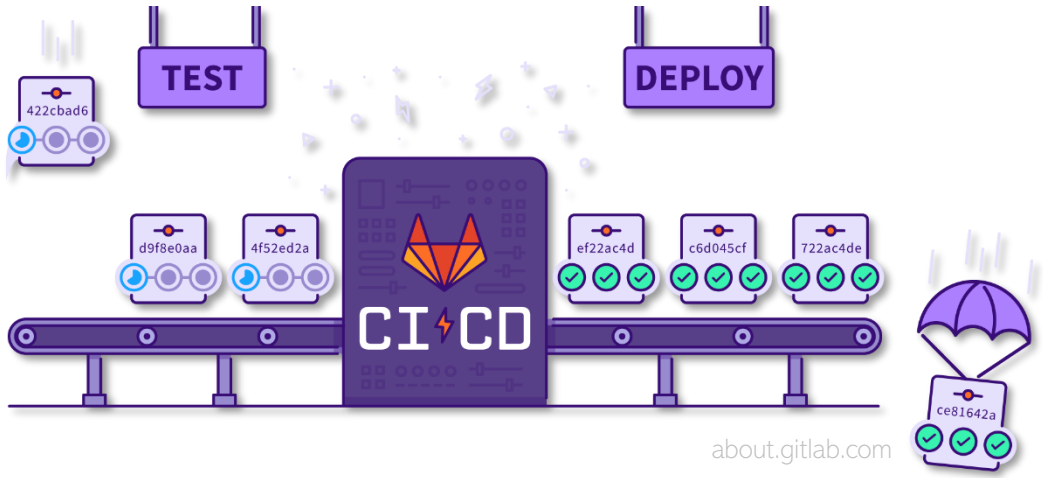
Pipeline: implementation

- **GitLab** CI/CD pipeline
- **Dockerized** jobs runners
- R modules + geo-spatial libraries
- **testthat** + shell-based unit test suites



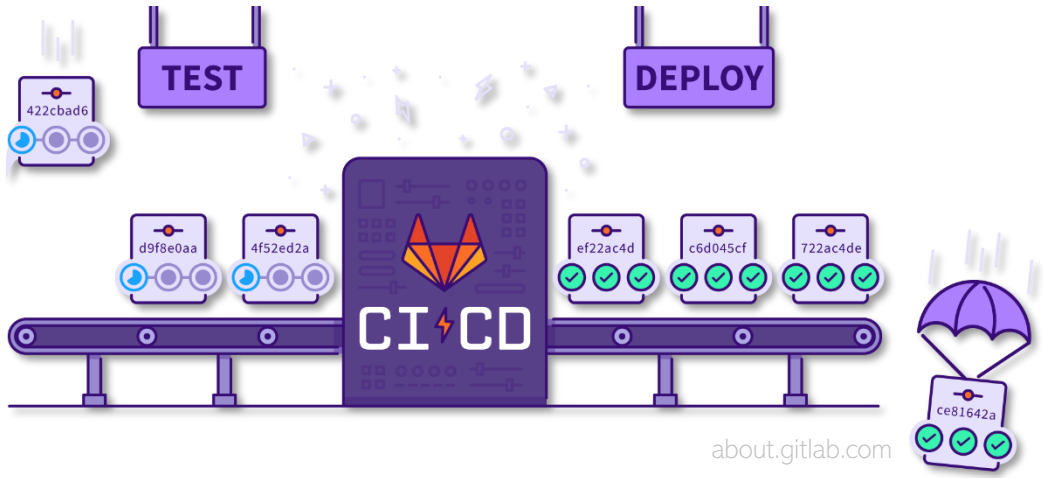
Pipeline: implementation

- **GitLab** CI/CD pipeline
- **Dockerized** jobs runners
- R modules + geo-spatial libraries
- **testthat** + shell-based unit test suites
- **renv** for reproducibile environment

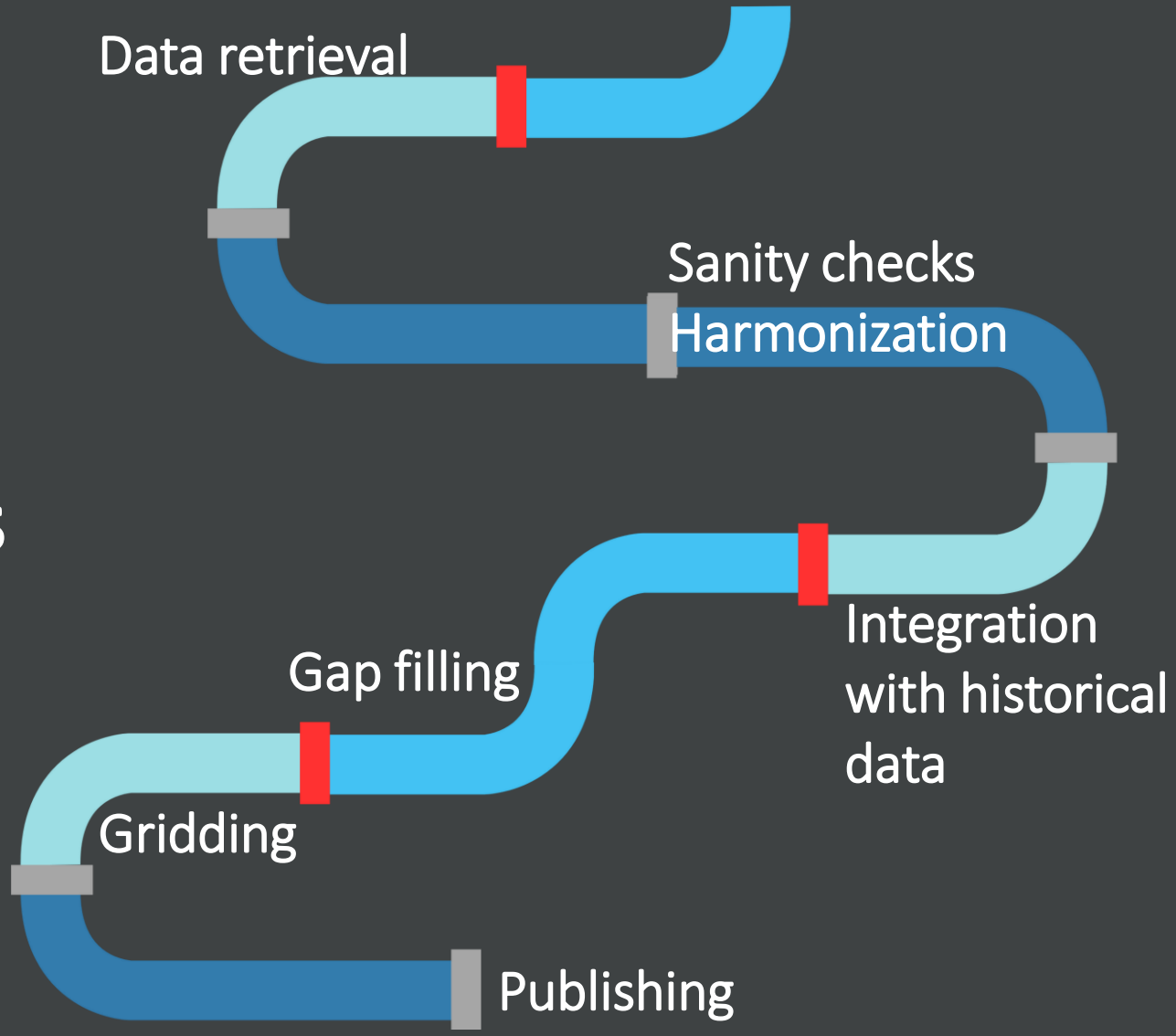


Pipeline: implementation

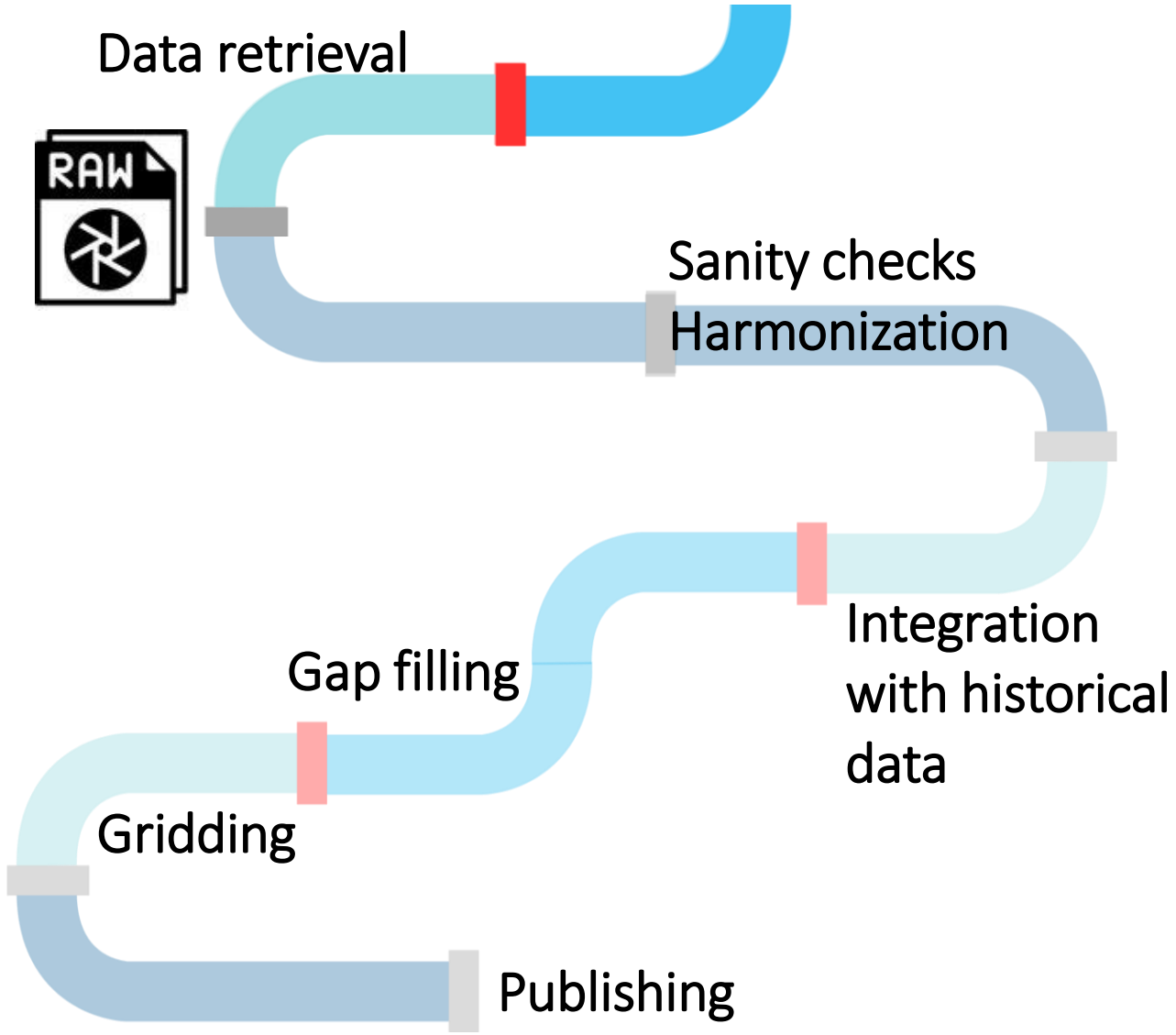
- **GitLab** CI/CD pipeline
- **Dockerized** jobs runners
- R modules + geo-spatial libraries
- **testthat** + shell-based unit test suites
- **renv** for reproducibile environment
- **bash** scripts and wrappers



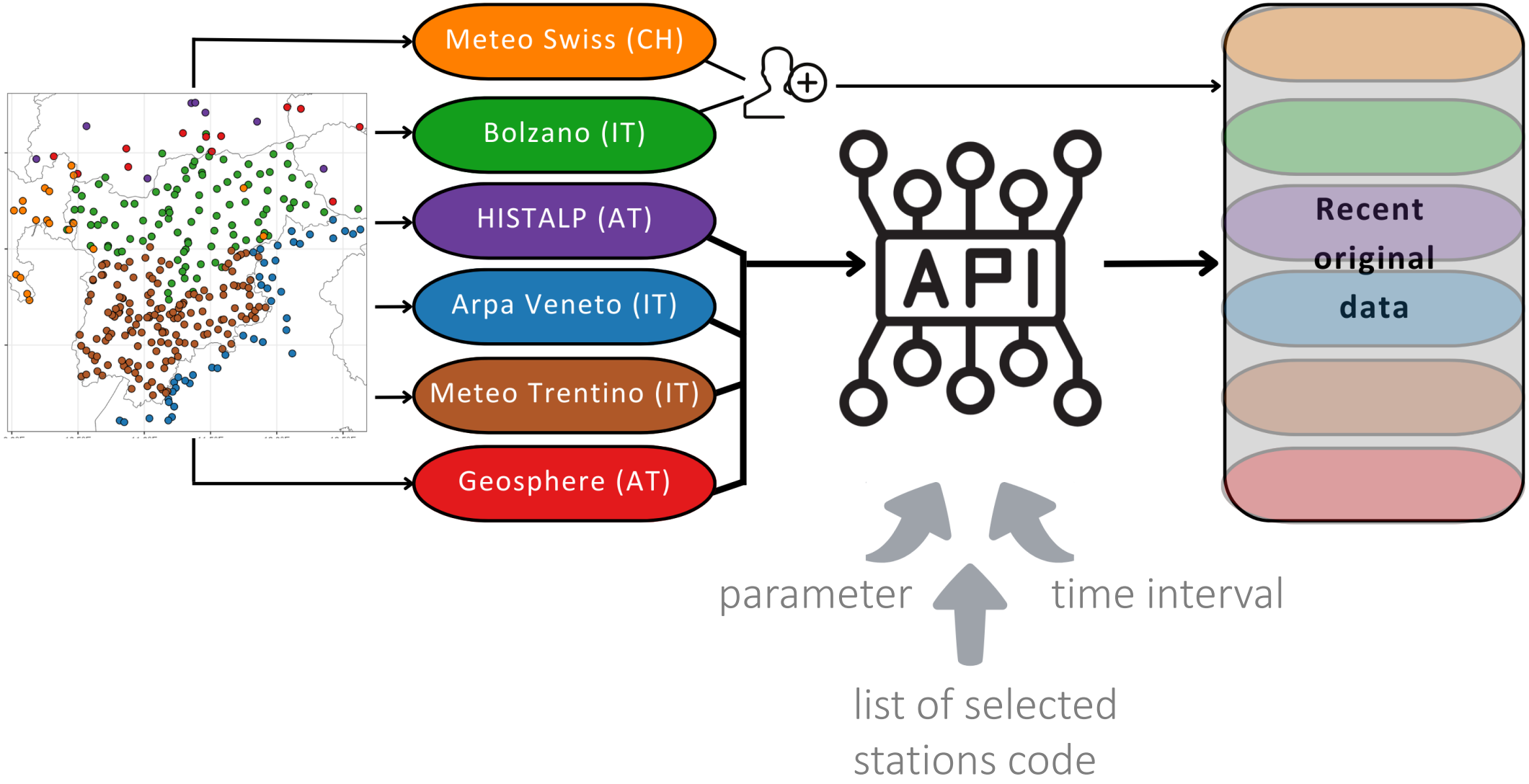
ETL pipeline for updating meteorological datasets in near real-time



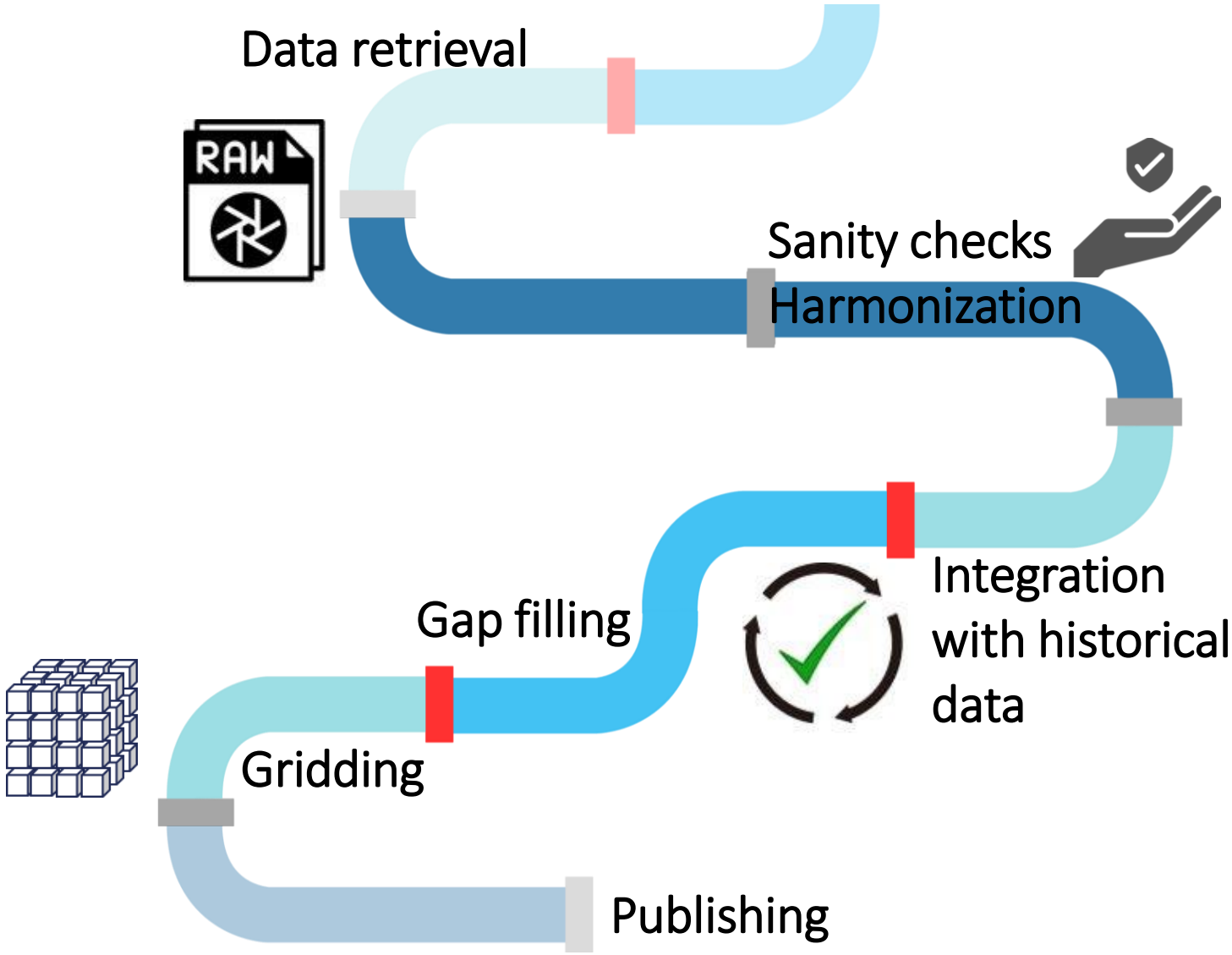
ETL pipeline for updating meteorological datasets in near real-time



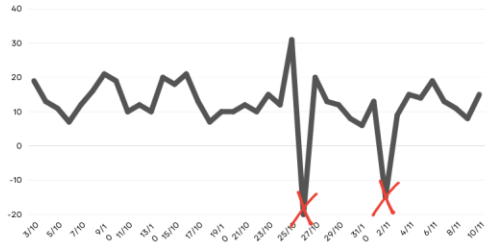
01 : Data extraction



ETL pipeline for updating meteorological datasets in near real-time



02 : Sanity checks and harmonization



REMOVE INCONSISTENCIES

CLEANING USING QUALITY FLAGS

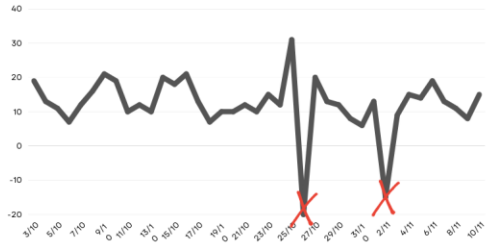
Date	Value	Quality flag
3/10/2023	~~13~~	bad!
4/10/2023	~~15~~	nope!
5/10/2023	11	ok!

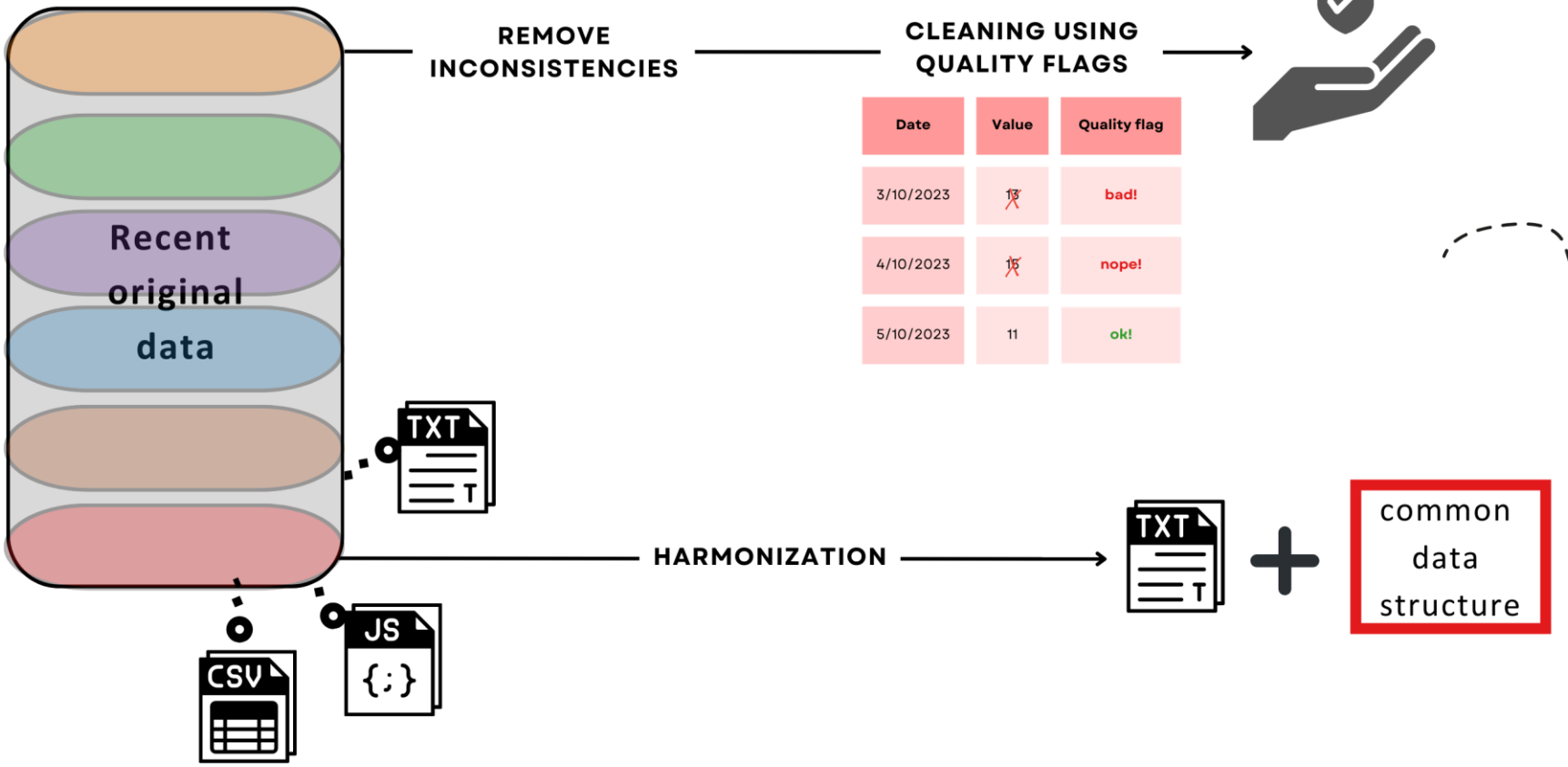
Recent original data

HARMONIZATION

common data structure

02 : Sanity checks and harmonization

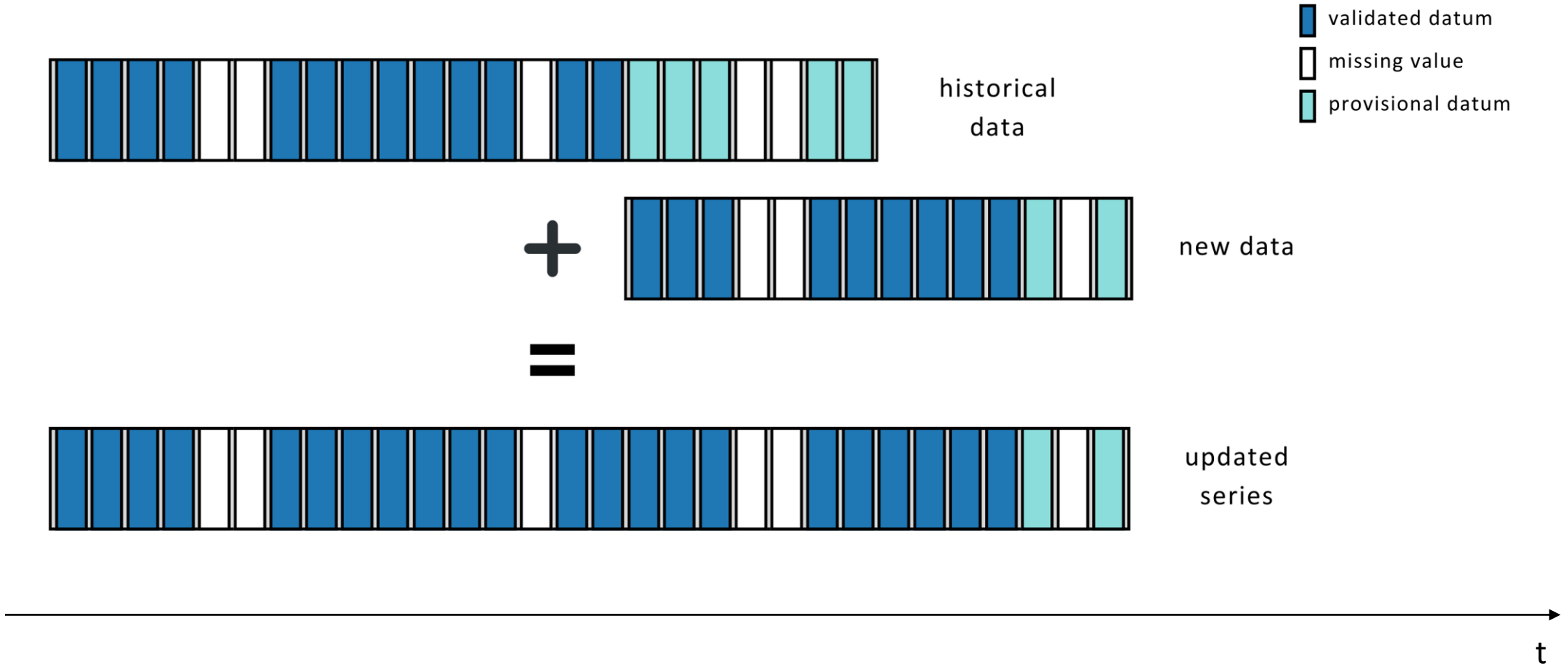




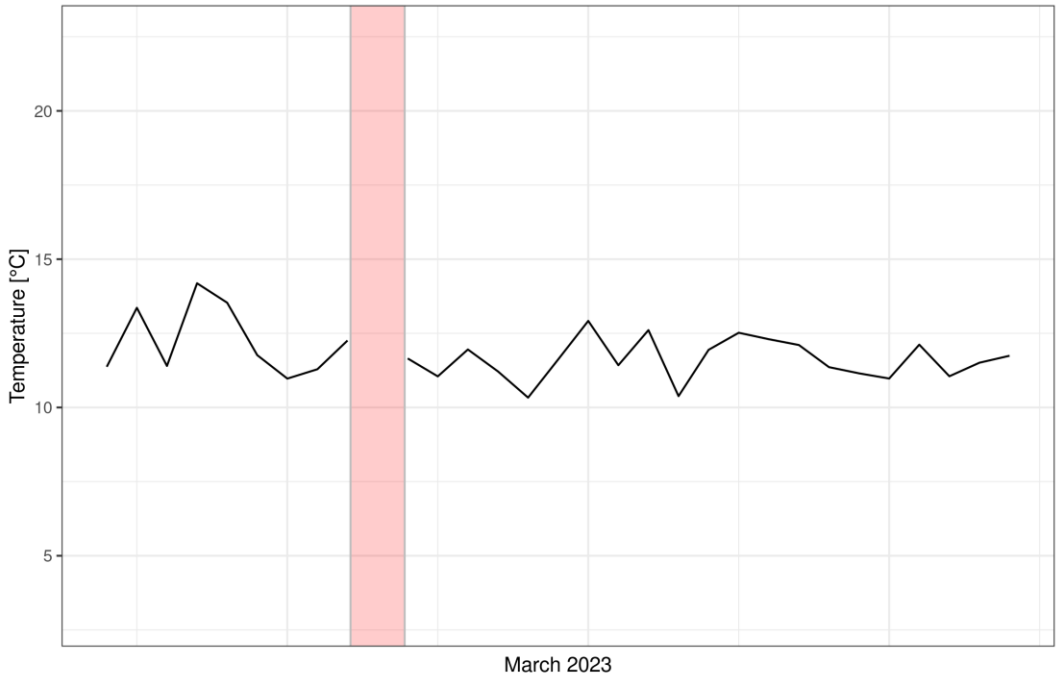
Date	Value	Quality flag
3/10/2023	~~13~~	bad!
4/10/2023	~~15~~	nope!
5/10/2023	11	ok!

Daily mean temperature calculated as mean(min, max)

03 : Integration with historical data



04 : Gap filling



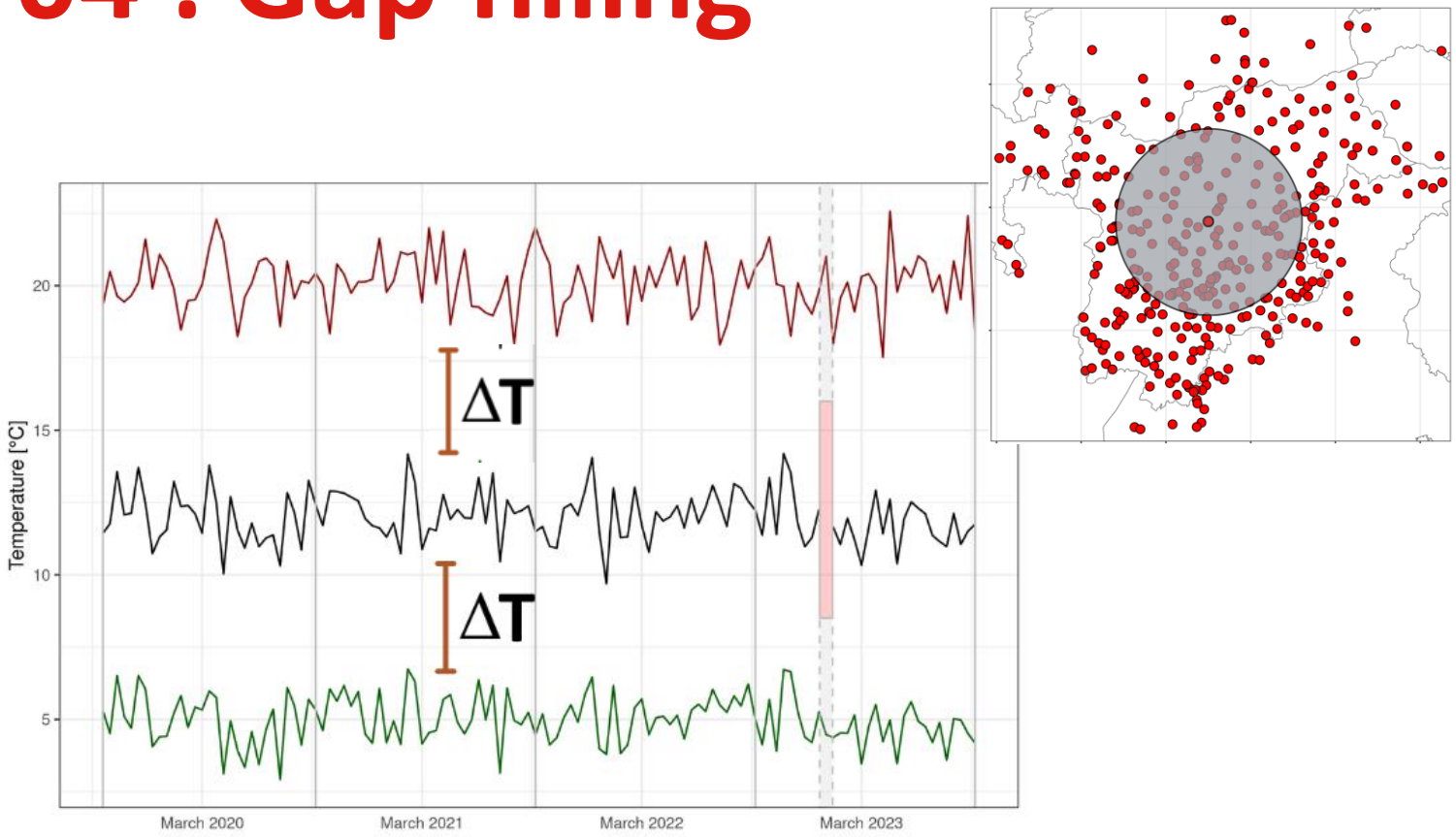
Only isolated gaps

04 : Gap filling

Only isolated gaps

Interpolation based on surrounding stations

04 : Gap filling

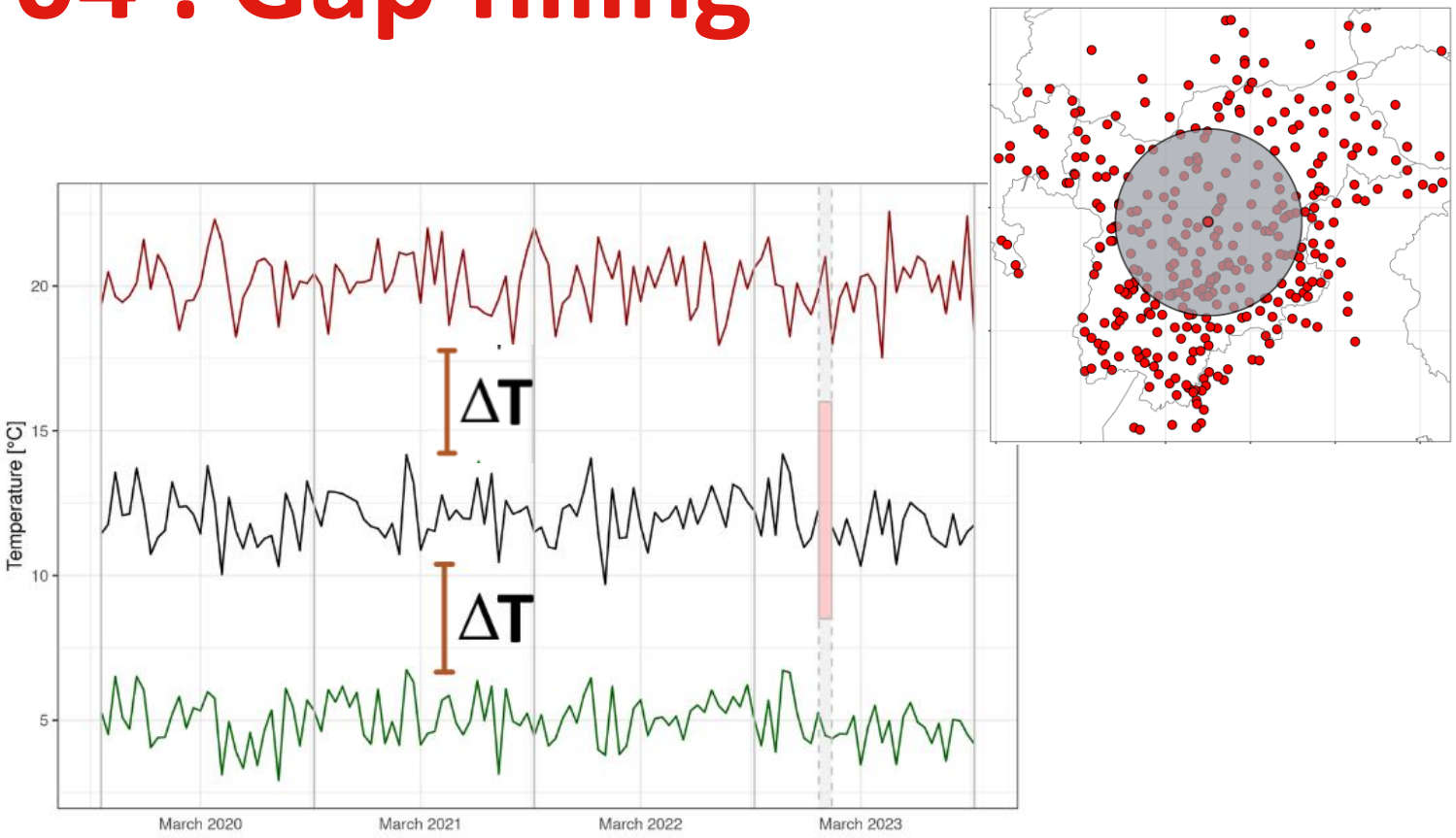


Only isolated gaps

Interpolation based on surrounding stations

Considering common historical data

04 : Gap filling

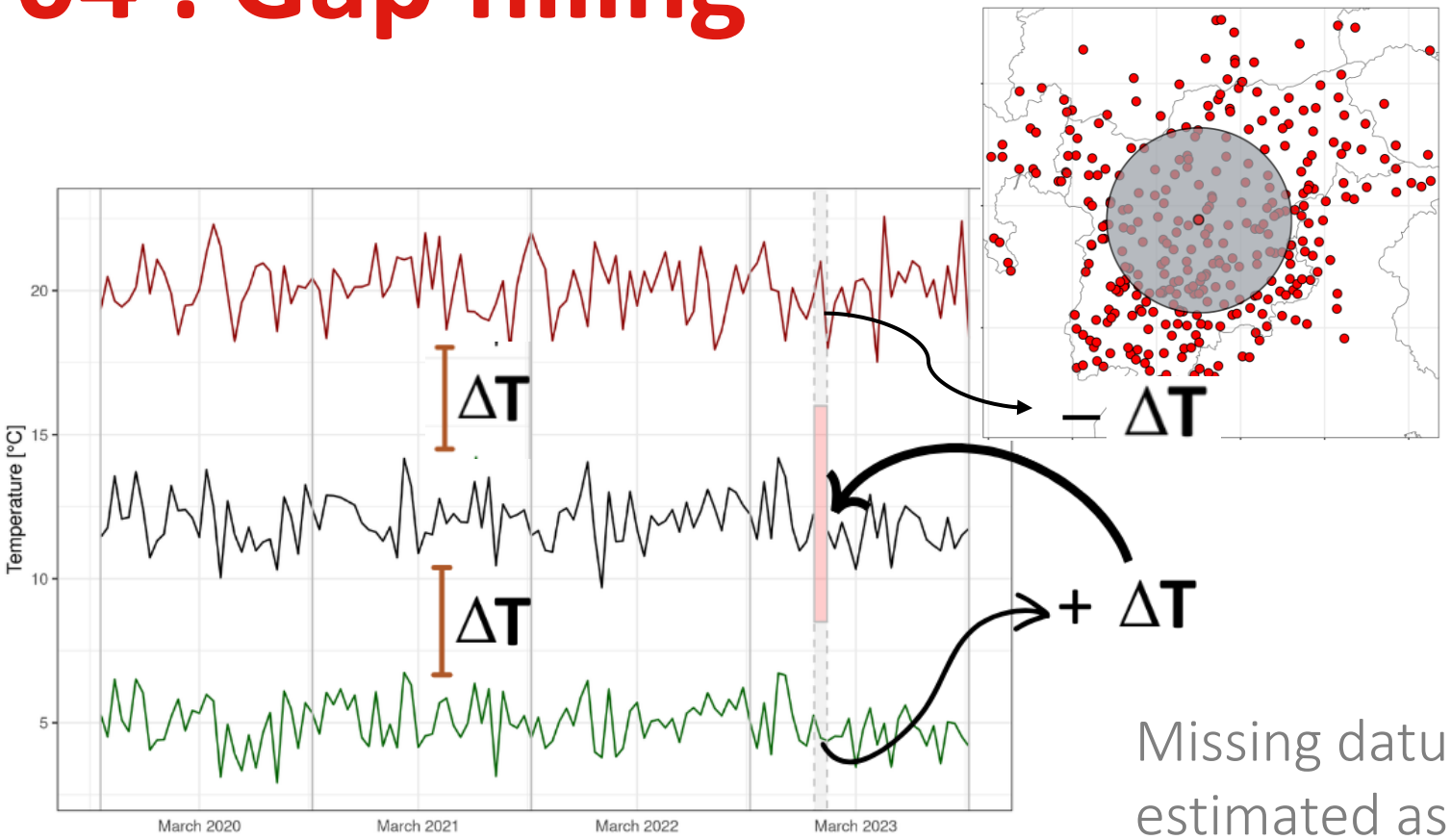


Only isolated gaps

Interpolation based on surrounding **correlated** stations

Considering common historical data

04 : Gap filling



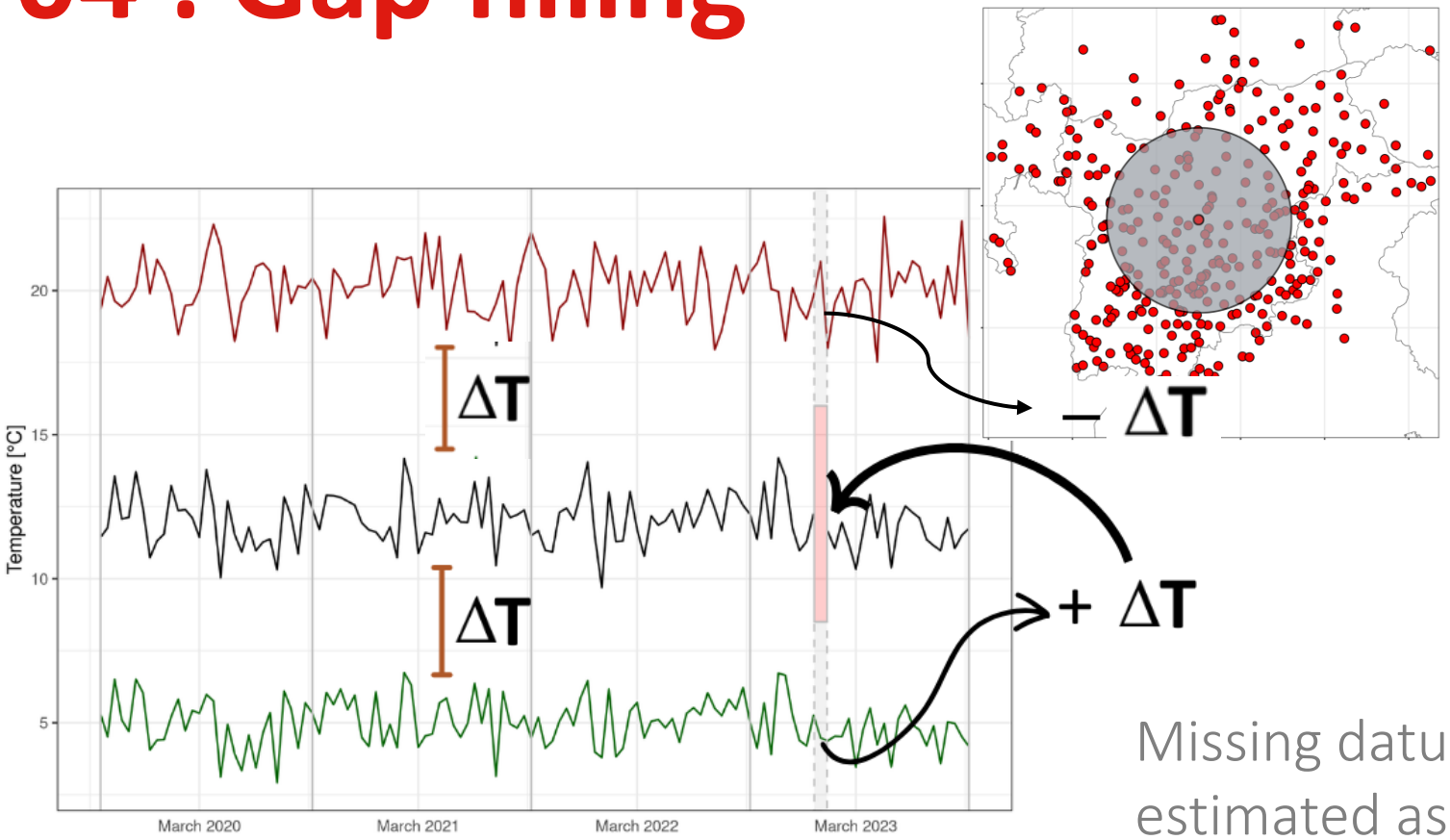
Only isolated gaps

Interpolation based on surrounding **correlated** stations

Considering common historical data

Missing datum estimated as the **weighted average of rescaled values**

04 : Gap filling



Only isolated gaps

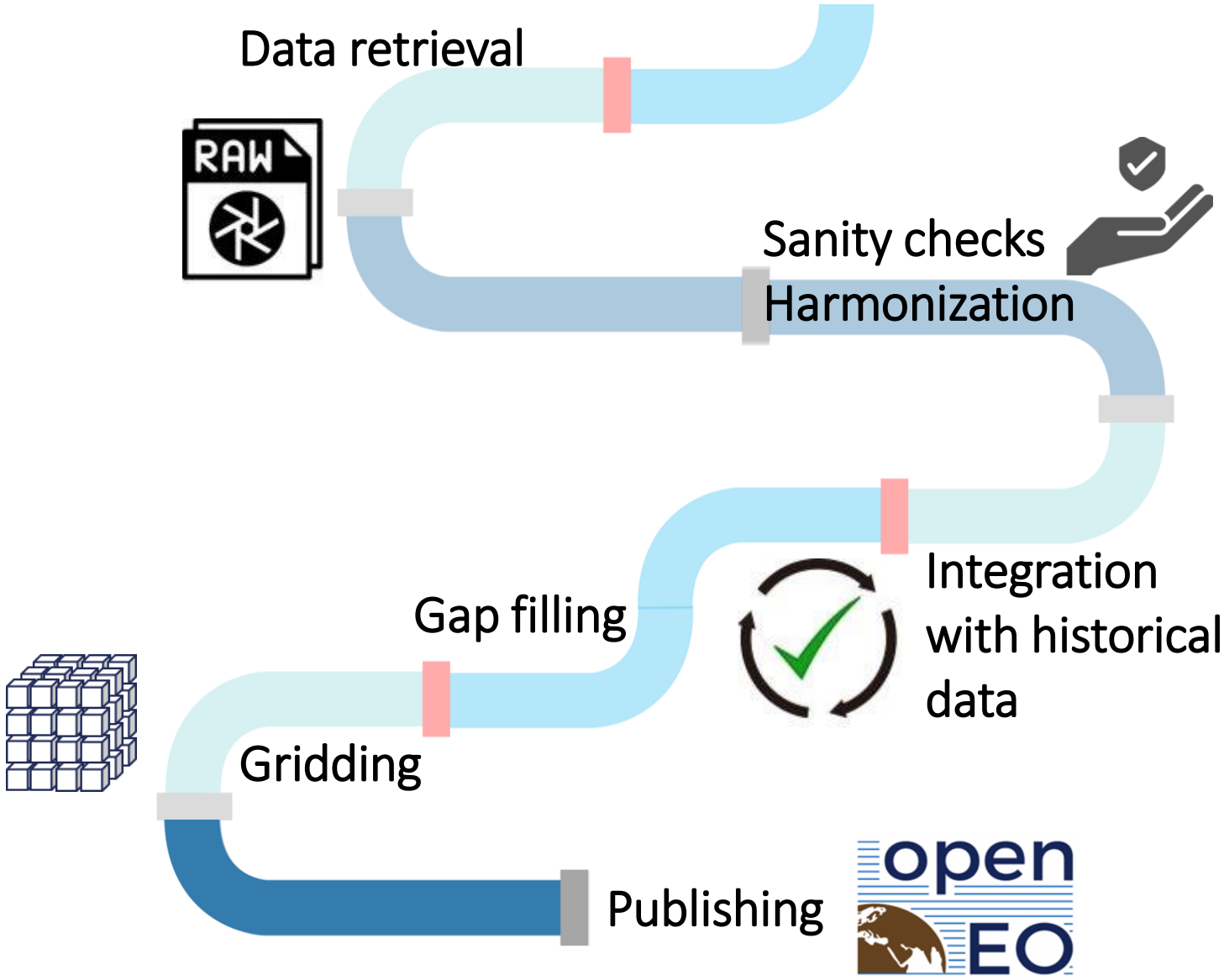
Interpolation based on surrounding **correlated** stations

Considering common historical data

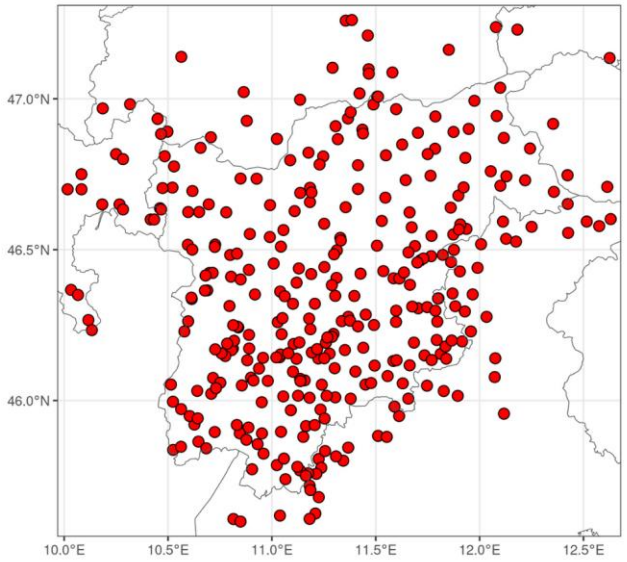
Missing datum estimated as the **weighted average of rescaled values**

Simulated values are **checked for consistency and mean temperature missing values** are determined from reconstructed values of min and max temperatures

ETL pipeline for updating meteorological datasets in near real-time

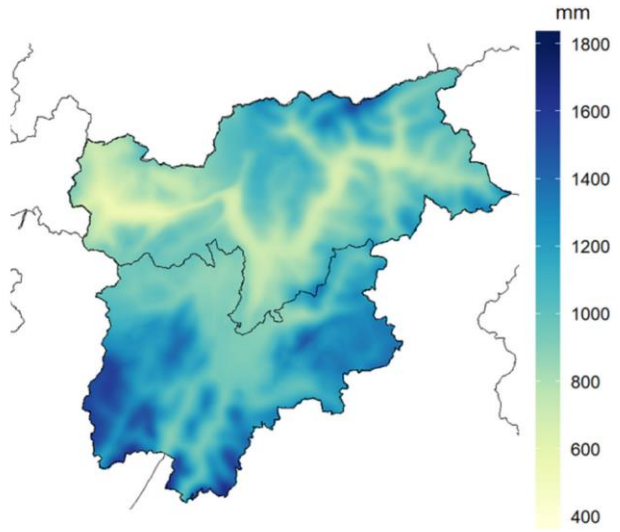


05 : Interpolation/gridding

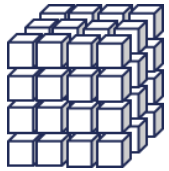


daily at-site time series

interpolation



daily grids



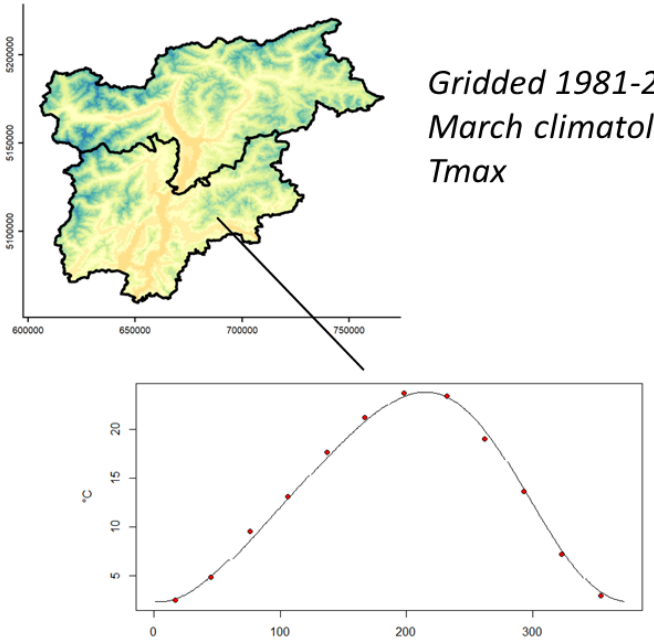
250m x 250m

NetCDF

05 : Interpolation/gridding: temperature

Local weighted linear regression
of temperature normals *vs*
elevation and weights depending
on topographic features

$$\bar{t}_m(x,y) = a_m(x,y) + b_m(x,y) \cdot h(x,y)$$

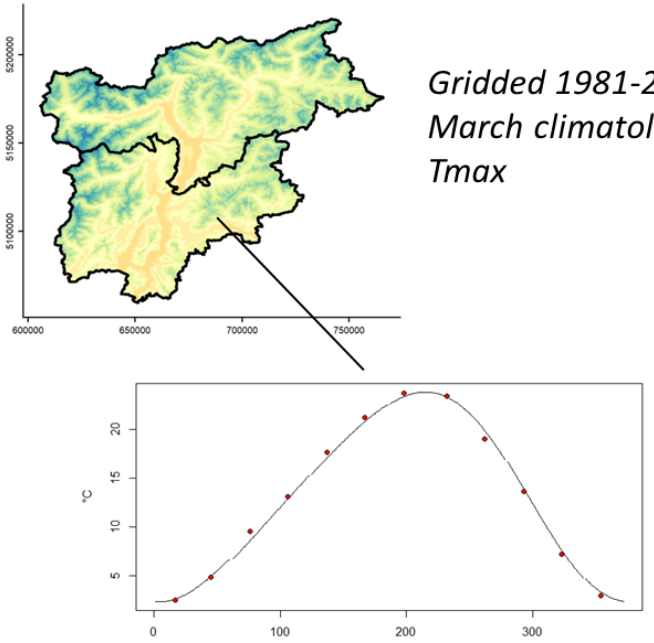


Gridded 1981-2010
March climatologies of
Tmax

05 : Interpolation/gridding: temperature

Local weighted linear regression
of temperature normals *vs*
elevation and weights depending
on topographic features

$$\bar{t}_m(x,y) = a_m(x,y) + b_m(x,y) \cdot h(x,y)$$

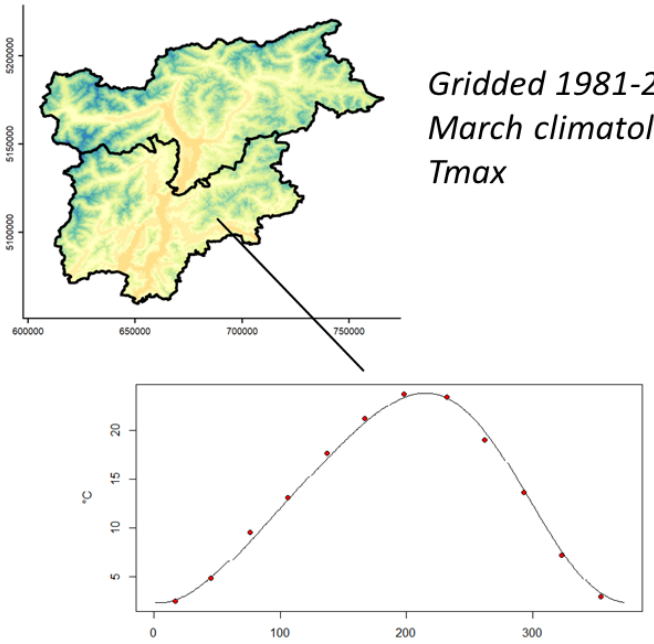


Gridded 1981-2010
March climatologies of
Tmax

05 : Interpolation/gridding: temperature

Local weighted linear regression of temperature normals *vs* elevation and weights depending on topographic features

$$\bar{t}_m(x,y) = a_m(x,y) + b_m(x,y) \cdot h(x,y)$$

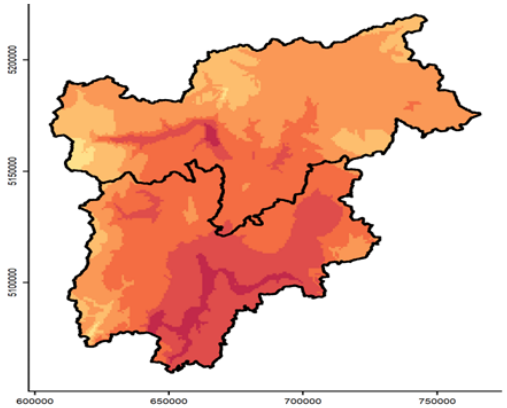


Gridded 1981-2010 March climatologies of Tmax

Daily normals based on a trigonometric fit from monthly normals

Local weighted average of station daily anomalies based on distance and elevation difference

$$a_d(x,y) = \frac{\sum_j w_j(x,y) \cdot a_{d,j}}{\sum_j w_j(x,y)}$$

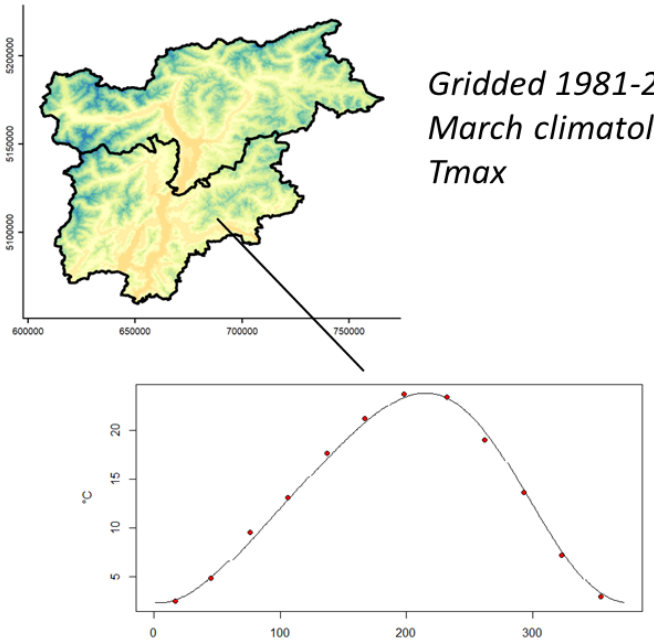


Gridded anomalies on 5th March 2001

05 : Interpolation/gridding: temperature

Local weighted linear regression of temperature normals *vs* elevation and weights depending on topographic features

$$\bar{t}_m(x,y) = a_m(x,y) + b_m(x,y) \cdot h(x,y)$$

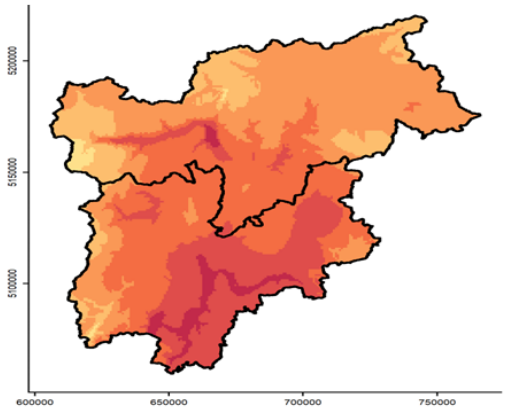


Gridded 1981-2010 March climatologies of Tmax

Daily normals based on a trigonometric fit from monthly normals

Local weighted average of station daily anomalies based on distance and elevation difference

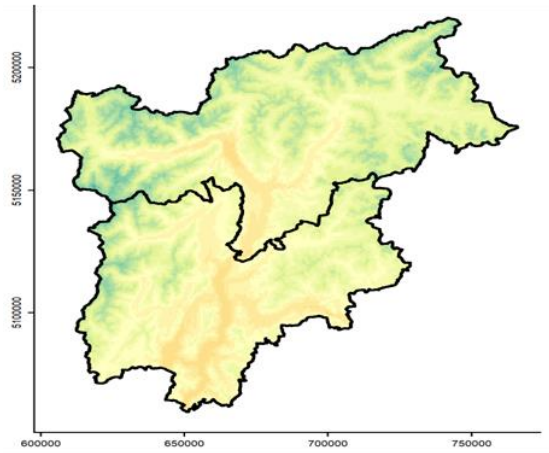
$$a_d(x,y) = \frac{\sum_j w_j(x,y) \cdot a_{d,j}}{\sum_j w_j(x,y)}$$



Gridded anomalies on 5th March 2001

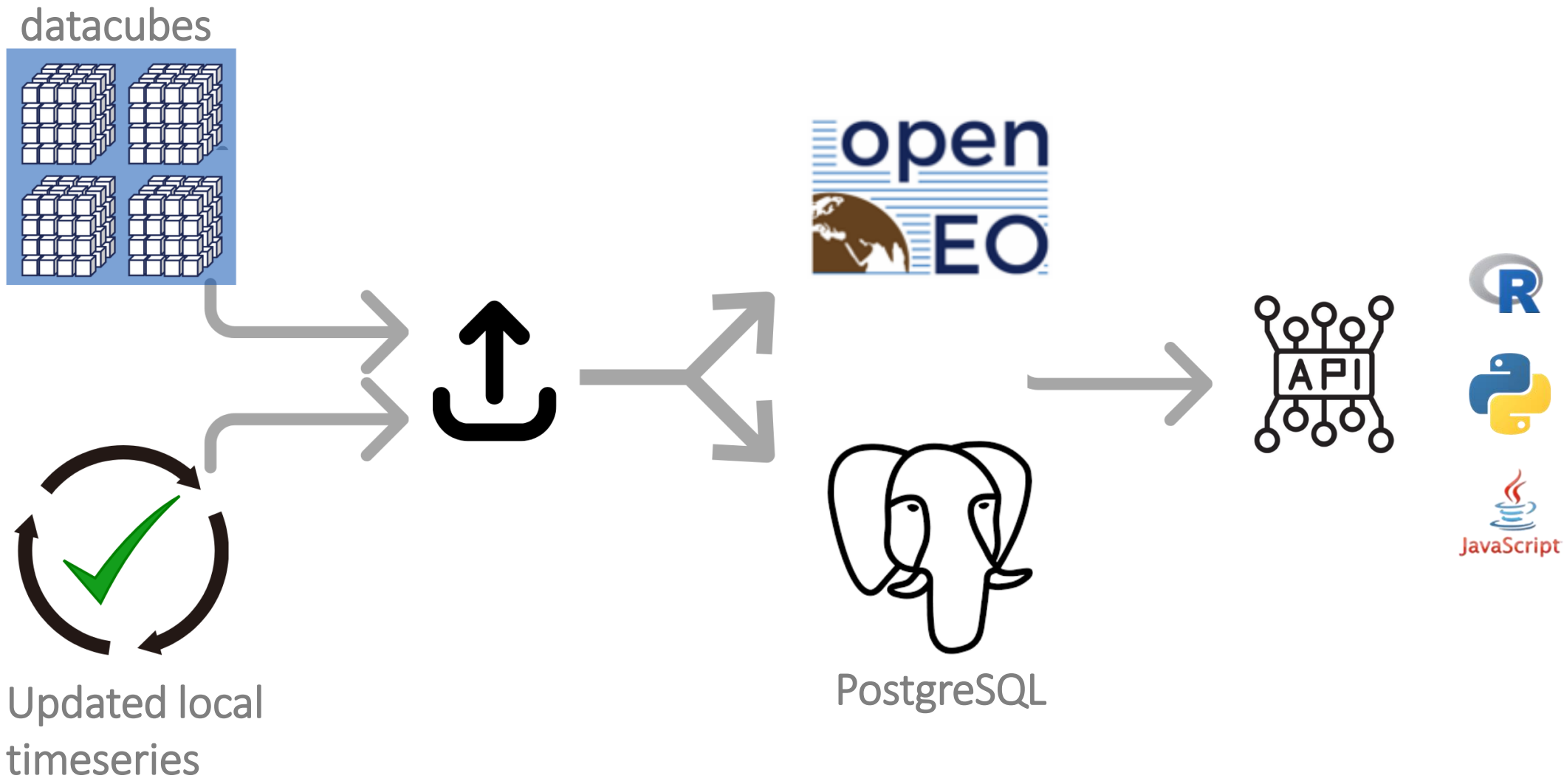
Superimposition (sum) of gridded anomalies and climatologies

$$t_d(x,y) = a_d(x,y) + \bar{t}_d(x,y)$$



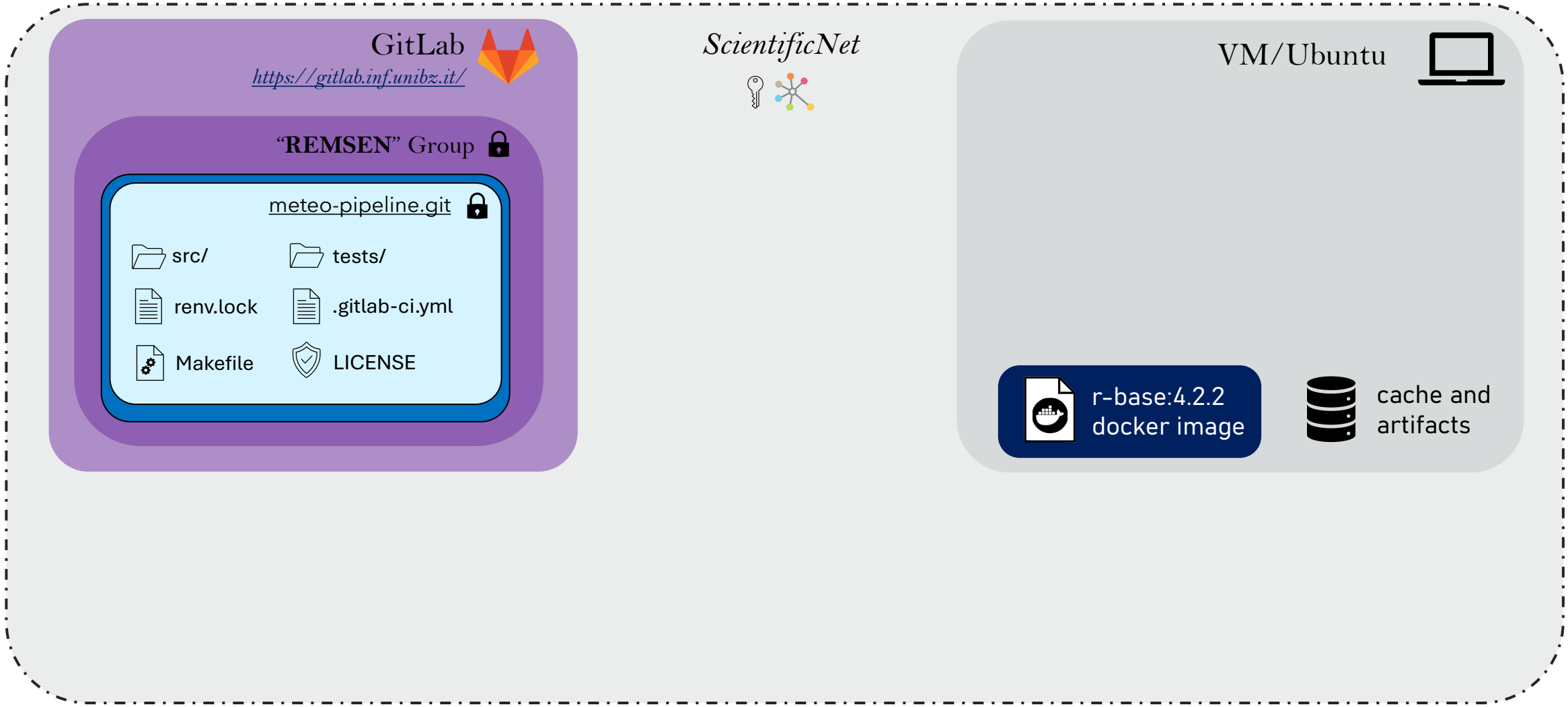
Gridded Tmax on 5th July 2001

06 : Publishing

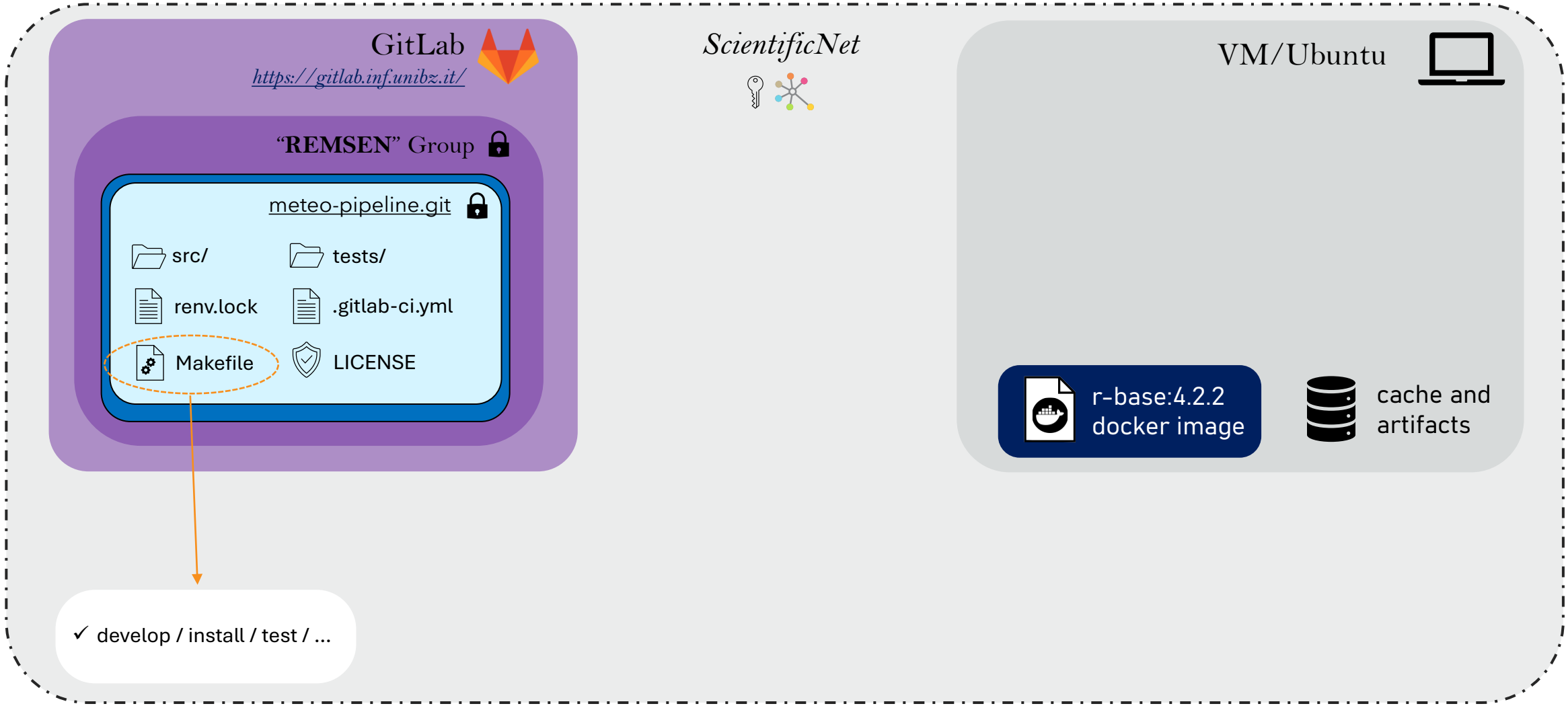


Automating the pipeline

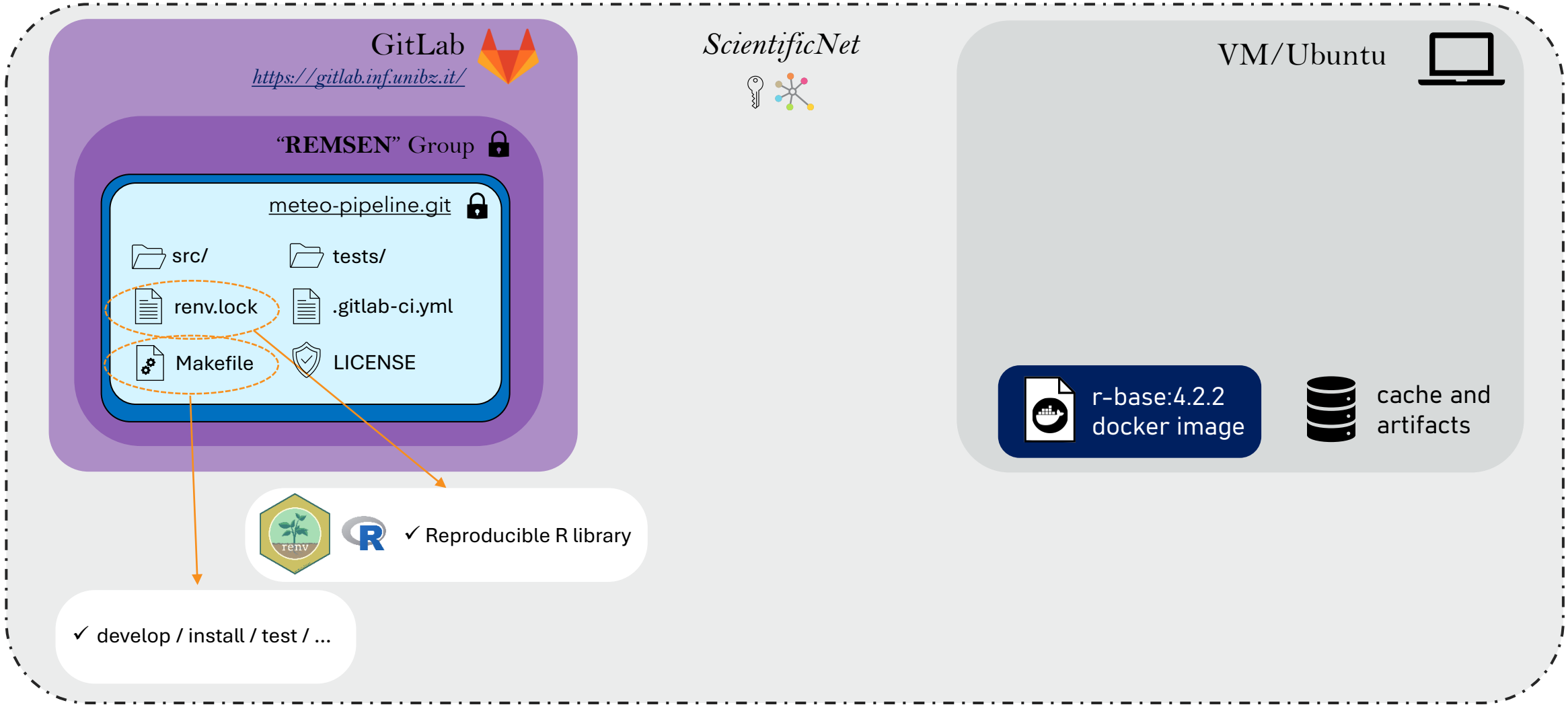
Automating the pipeline



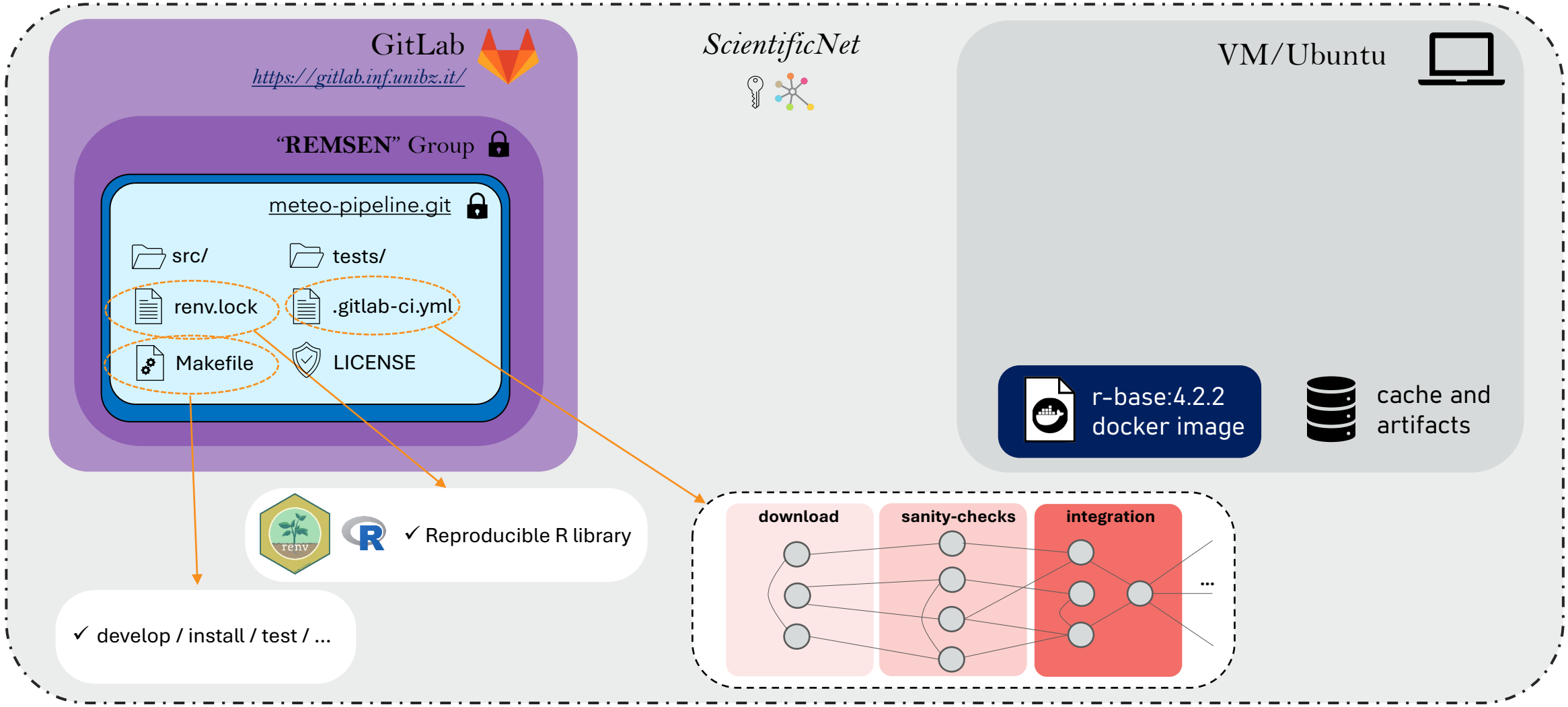
Automating the pipeline



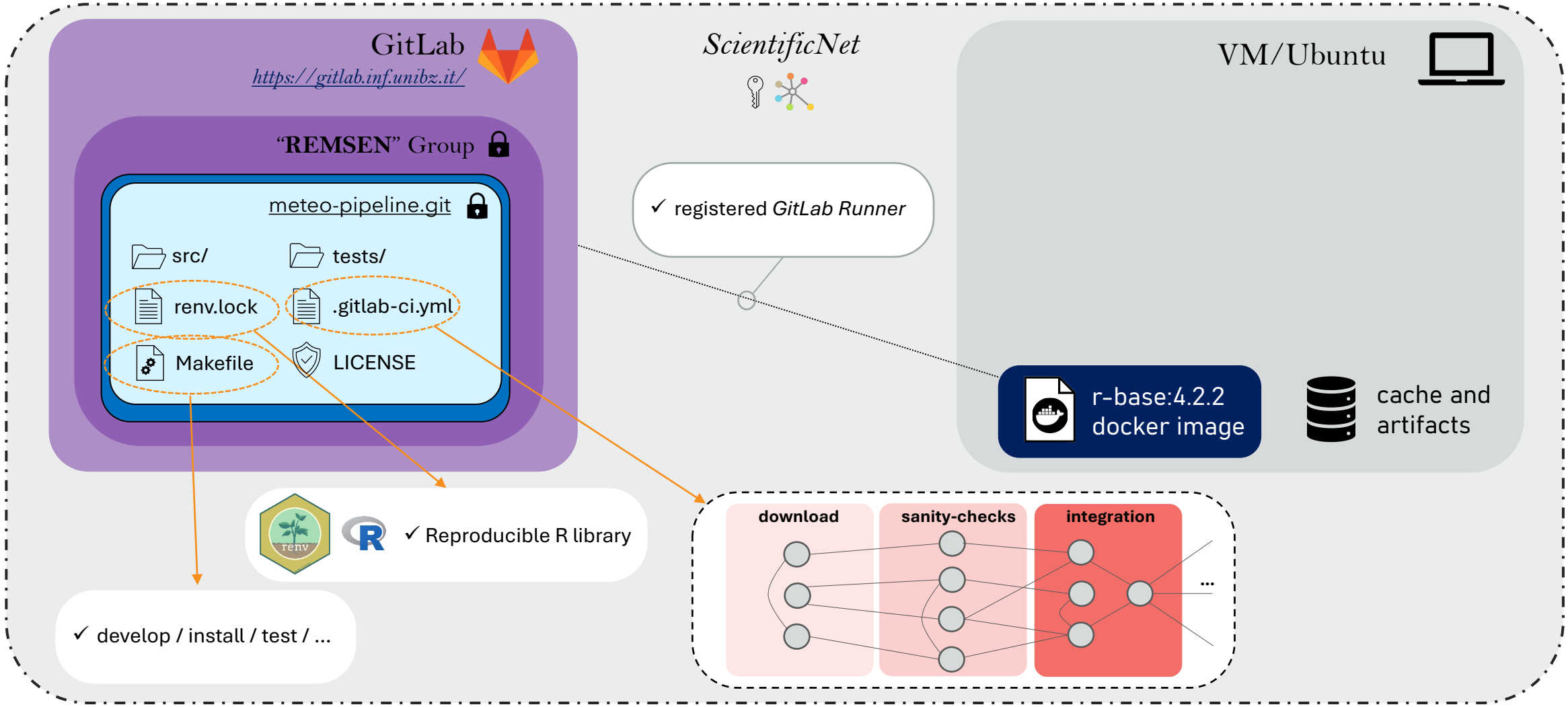
Automating the pipeline



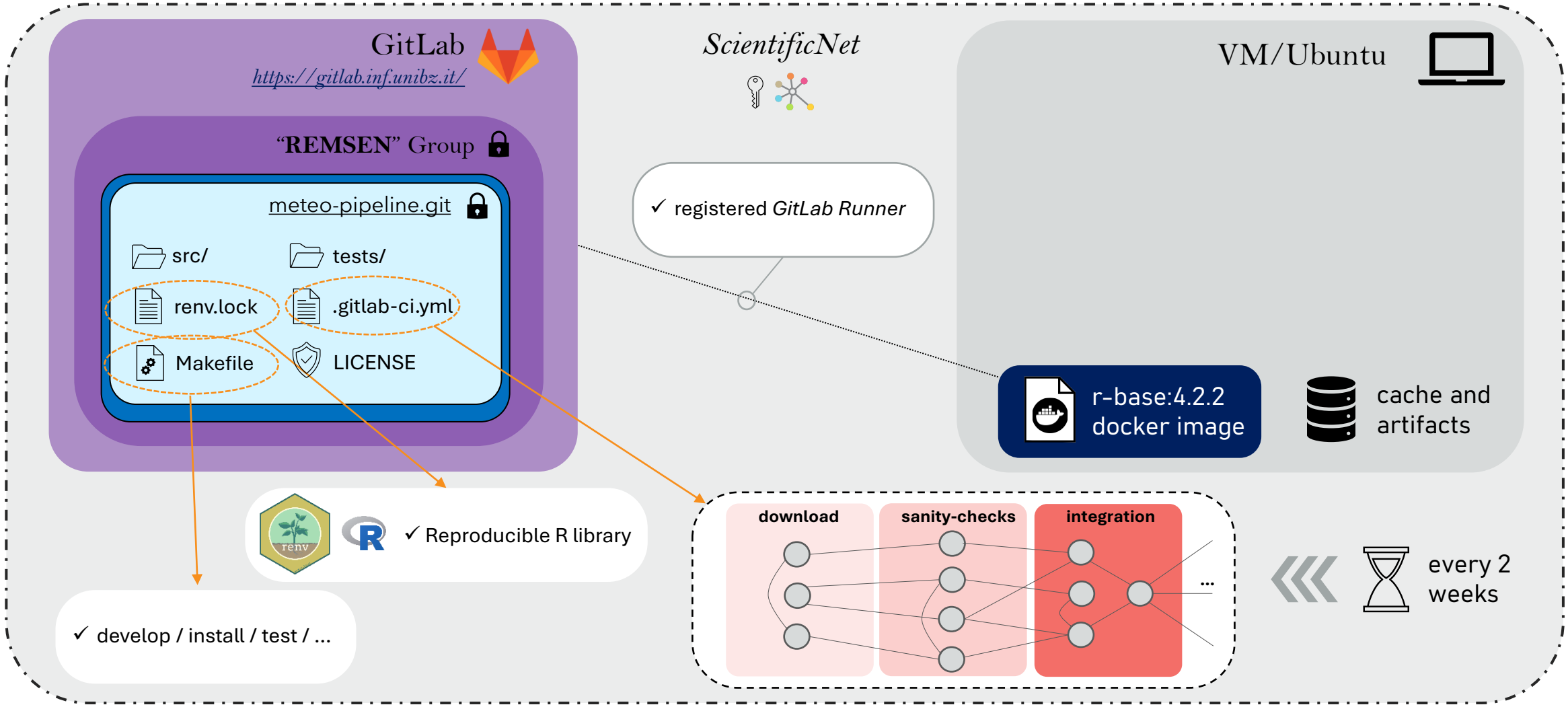
Automating the pipeline



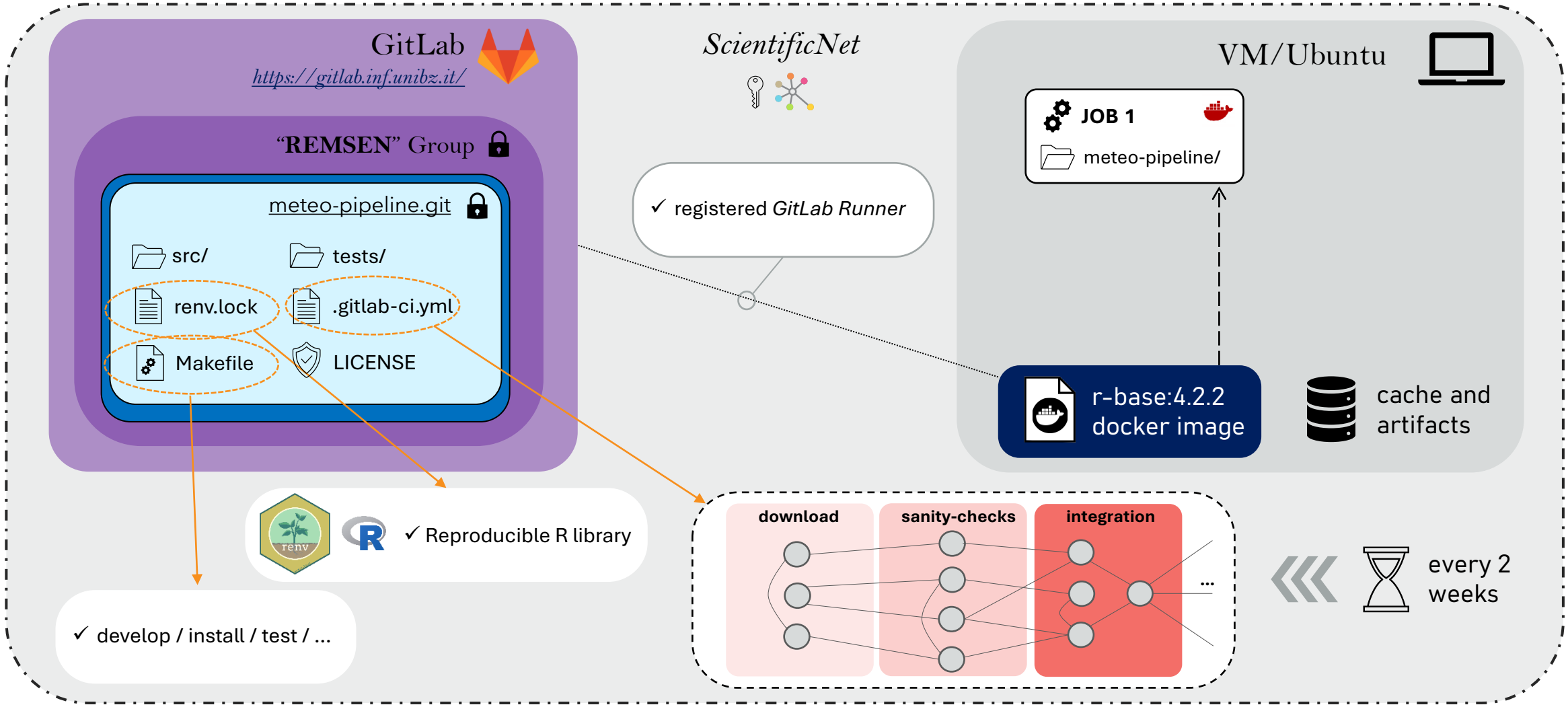
Automating the pipeline



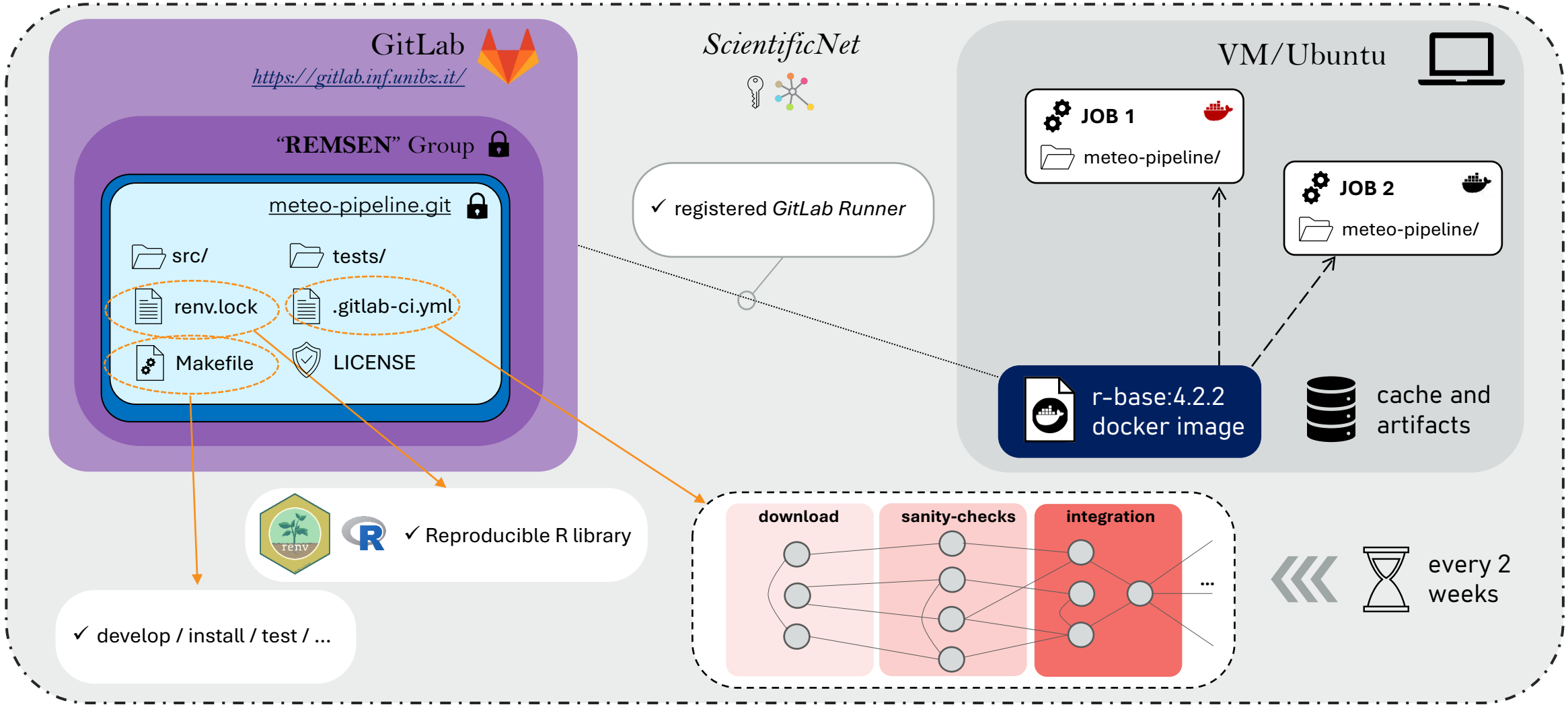
Automating the pipeline



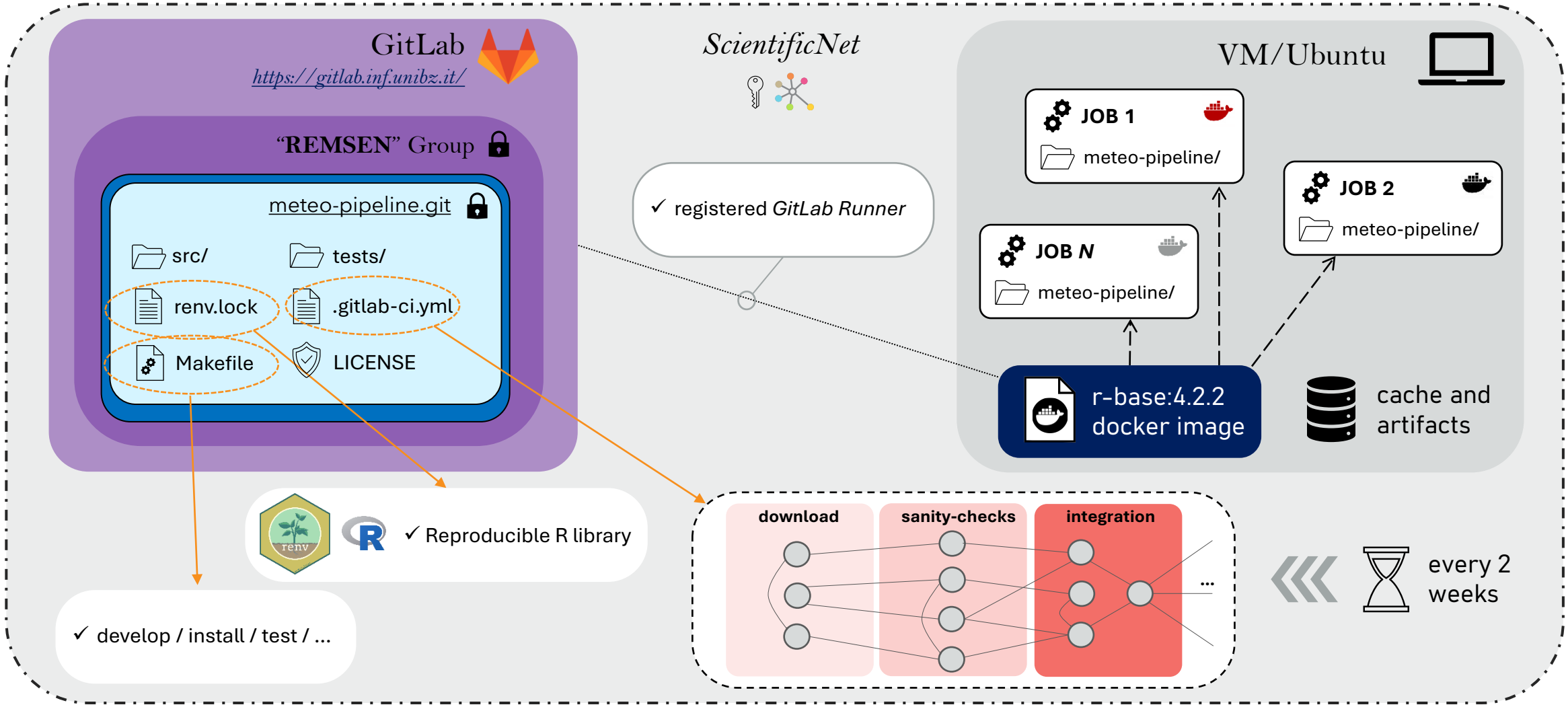
Automating the pipeline



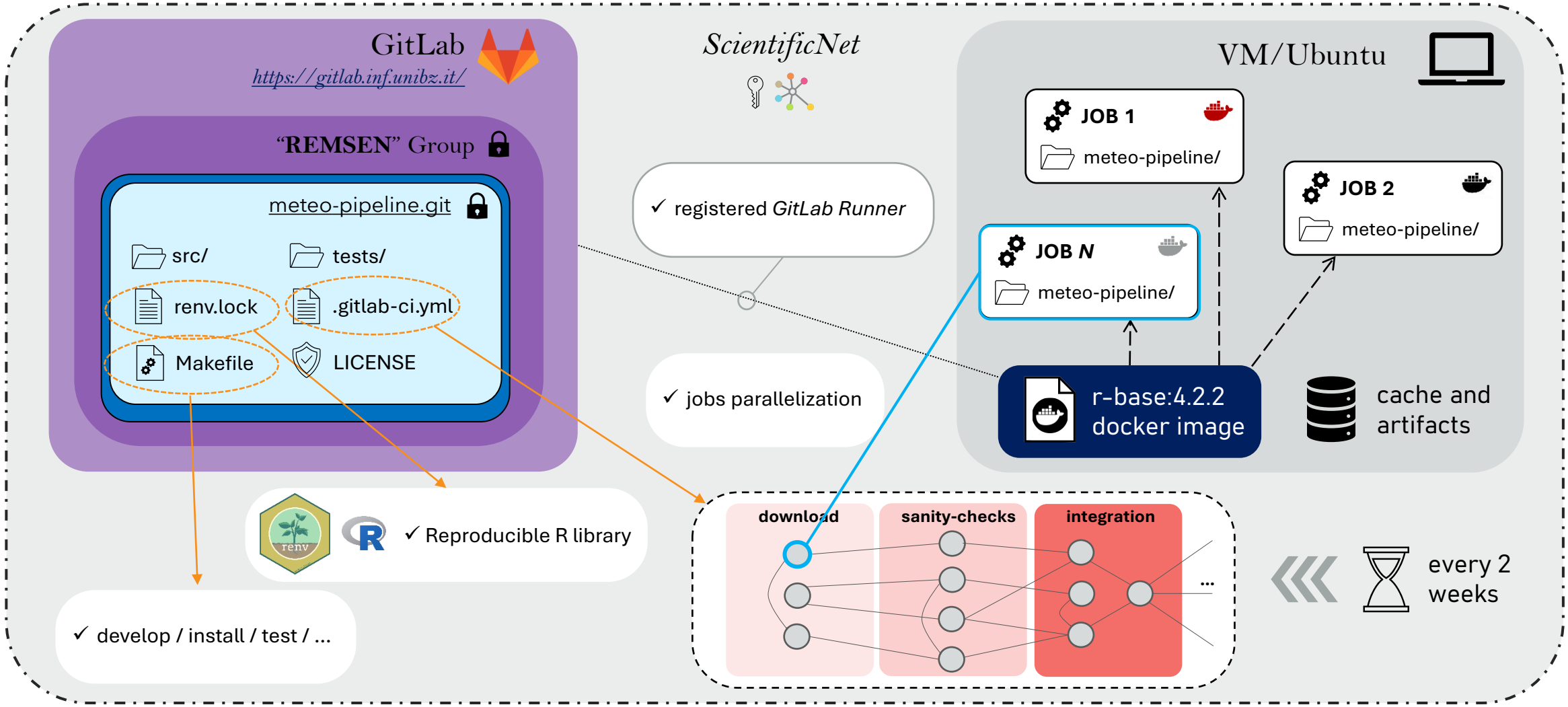
Automating the pipeline



Automating the pipeline



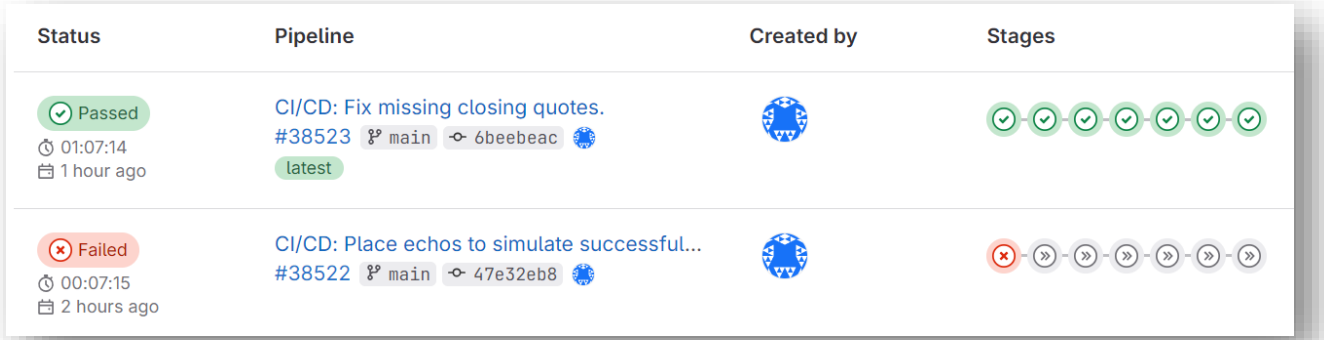
Automating the pipeline



GitLab's out of the box:
Monitoring & Troubleshooting

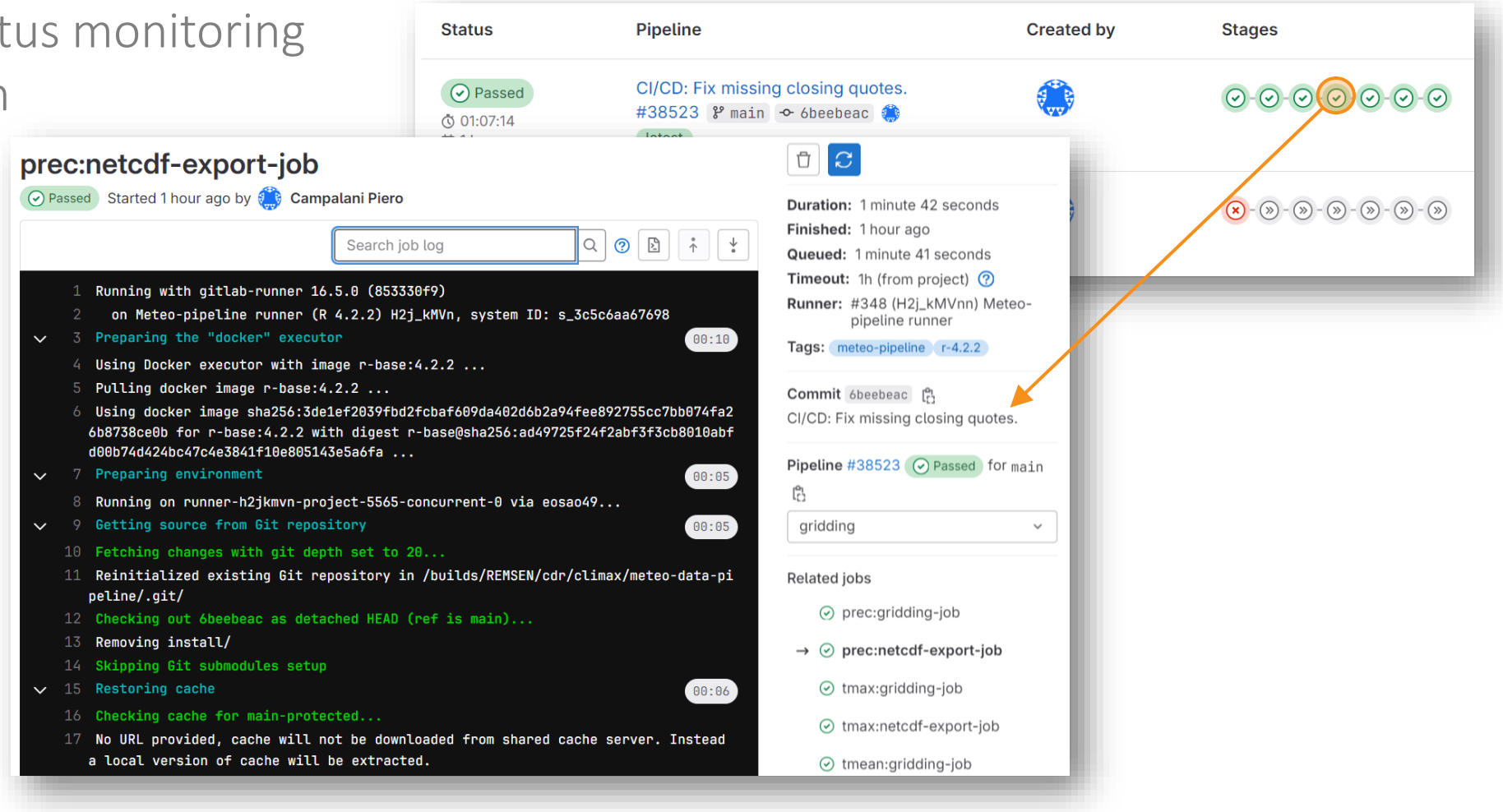
GitLab’s out of the box:
Monitoring & Troubleshooting

✓ UI-based pipeline status monitoring



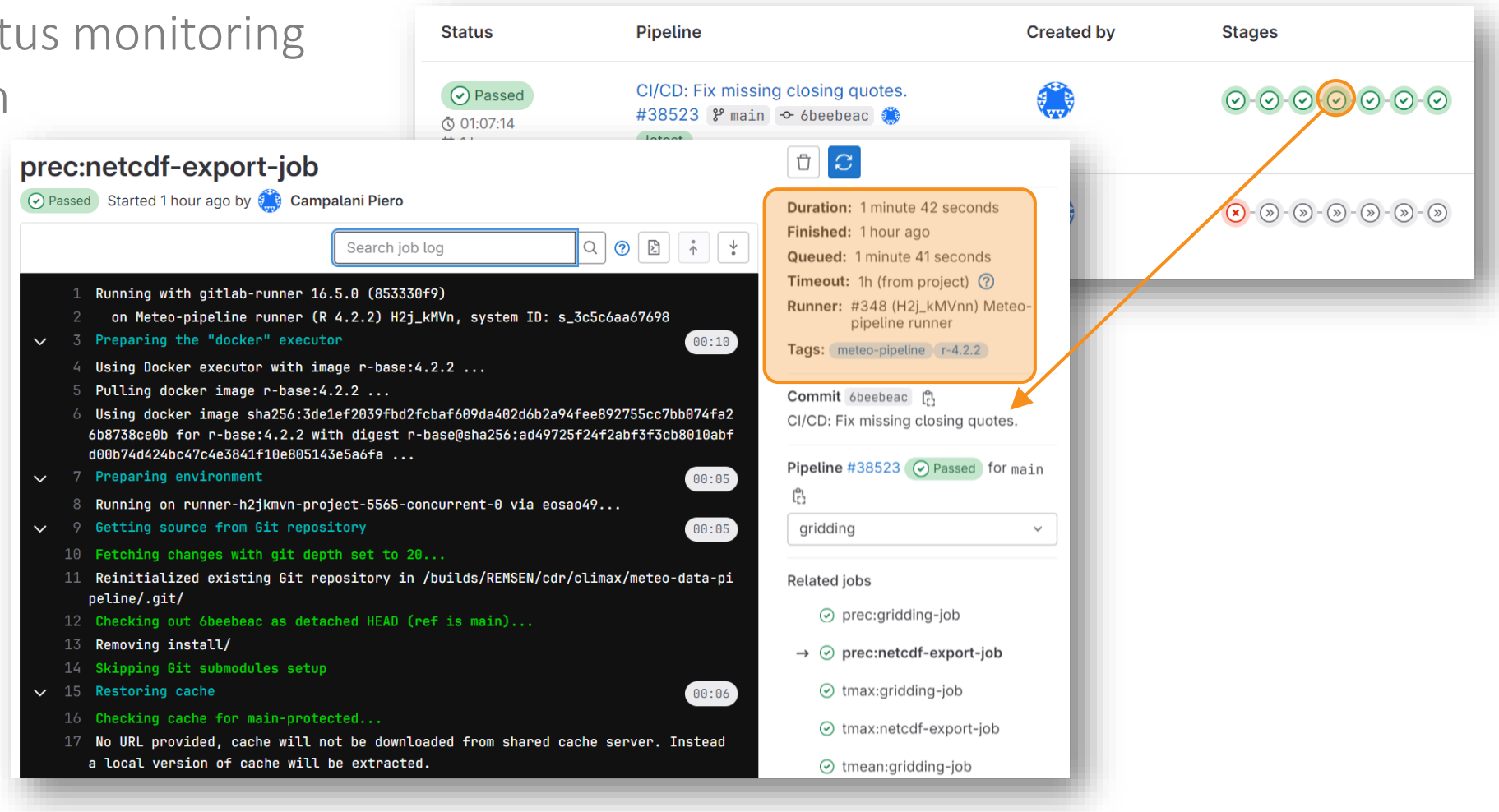
GitLab's out of the box:
Monitoring & Troubleshooting

- ✓ UI-based pipeline status monitoring
- ✓ Jobs logs visualization



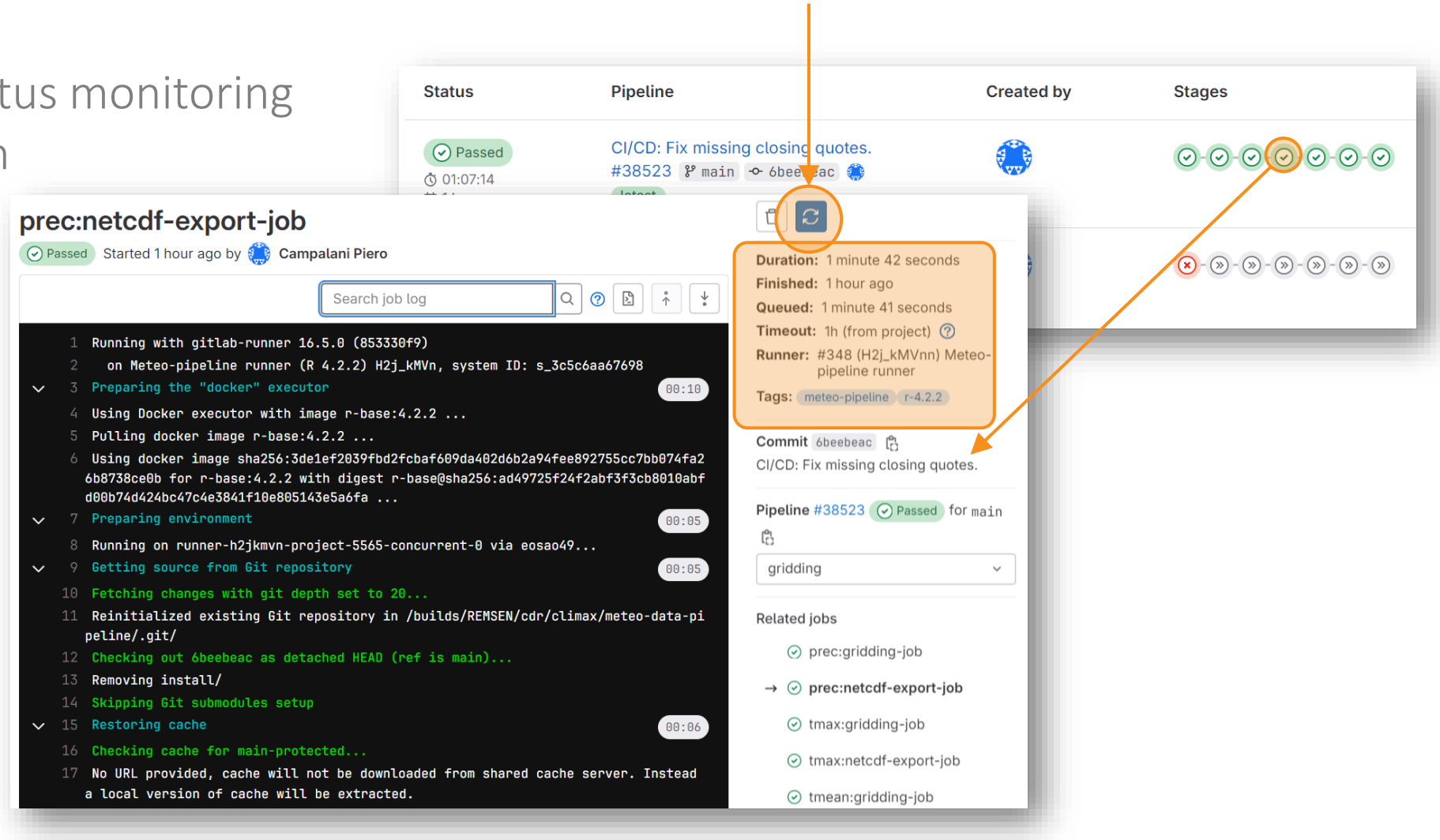
GitLab's out of the box:
Monitoring & Troubleshooting

- ✓ UI-based pipeline status monitoring
- ✓ Jobs logs visualization
- ✓ Timing information



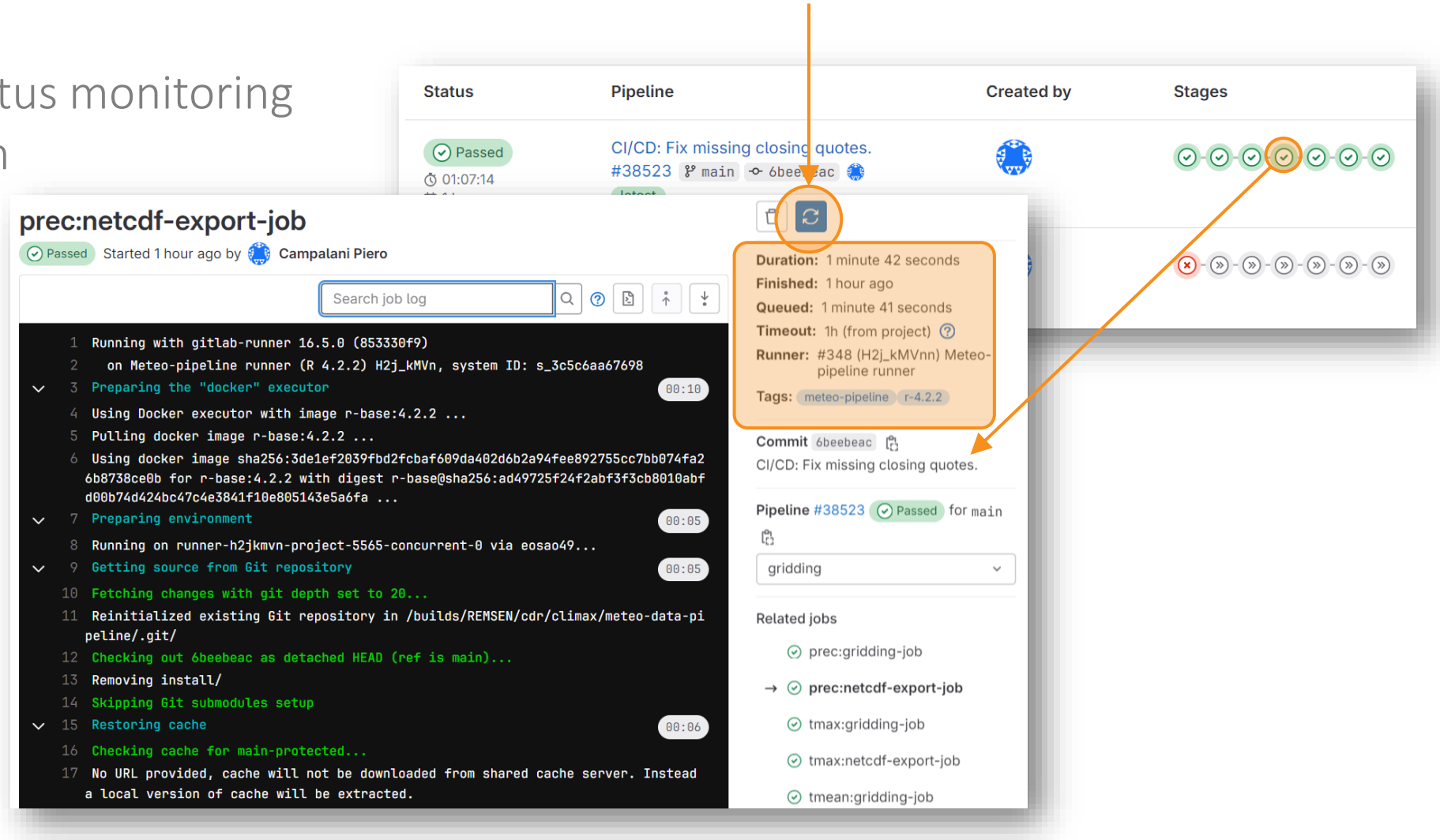
GitLab's out of the box:
Monitoring & Troubleshooting

- ✓ UI-based pipeline status monitoring
- ✓ Jobs logs visualization
- ✓ Timing information
- ✓ Job's manual re-runs

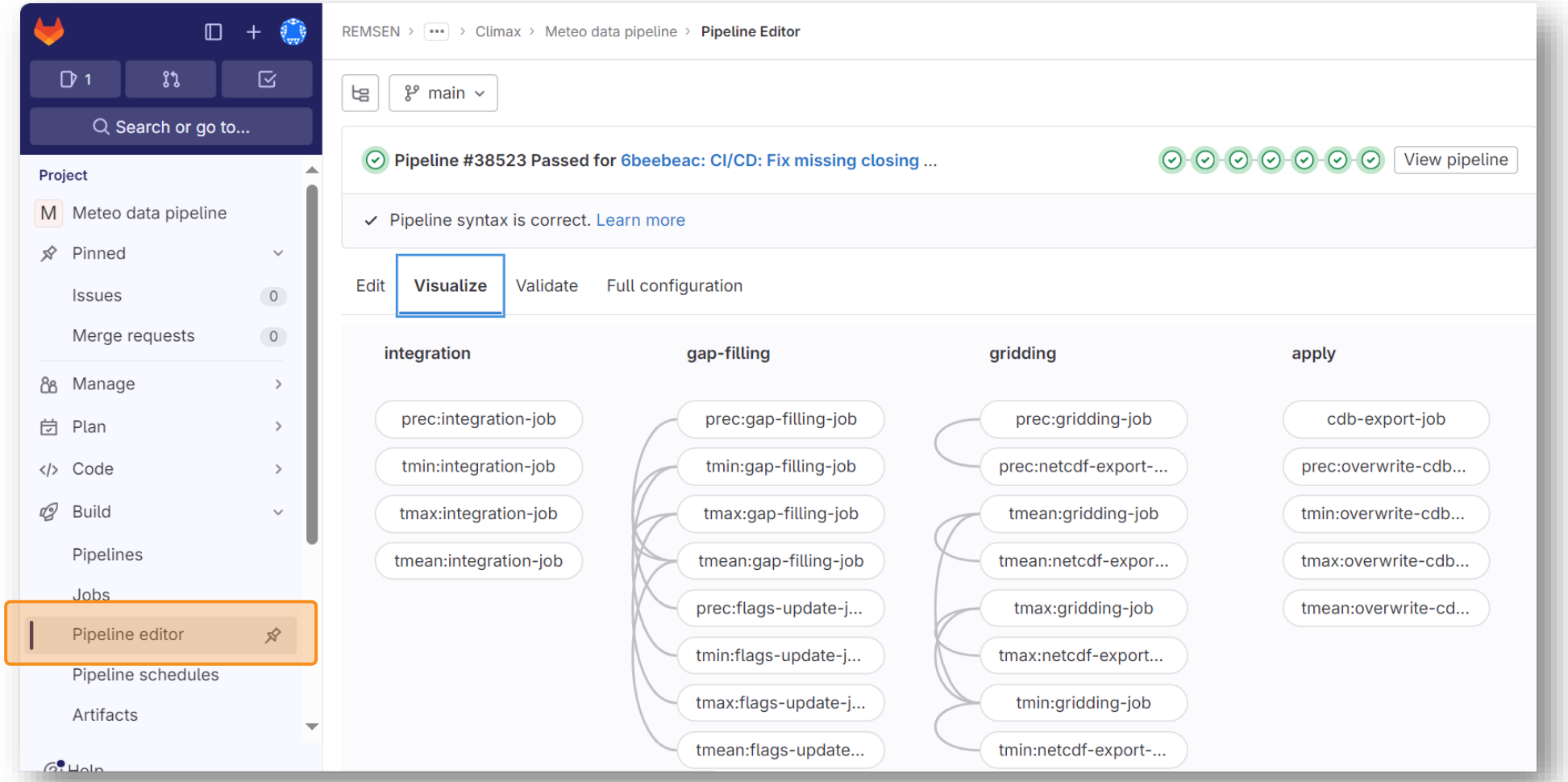


GitLab's out of the box:
Monitoring & Troubleshooting

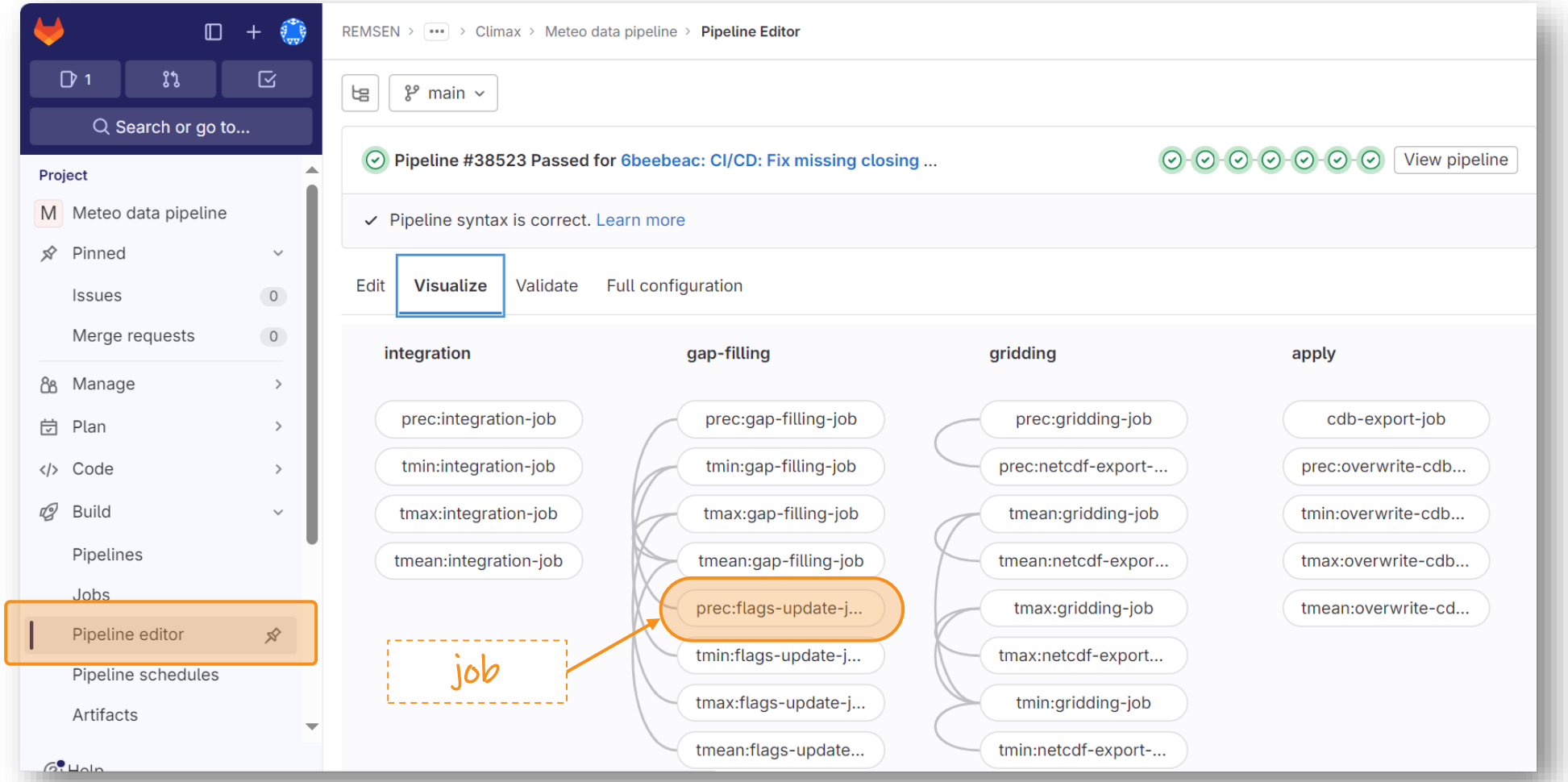
- ✓ UI-based pipeline status monitoring
- ✓ Jobs logs visualization
- ✓ Timing information
- ✓ Job's manual re-runs
- ✓ API also available



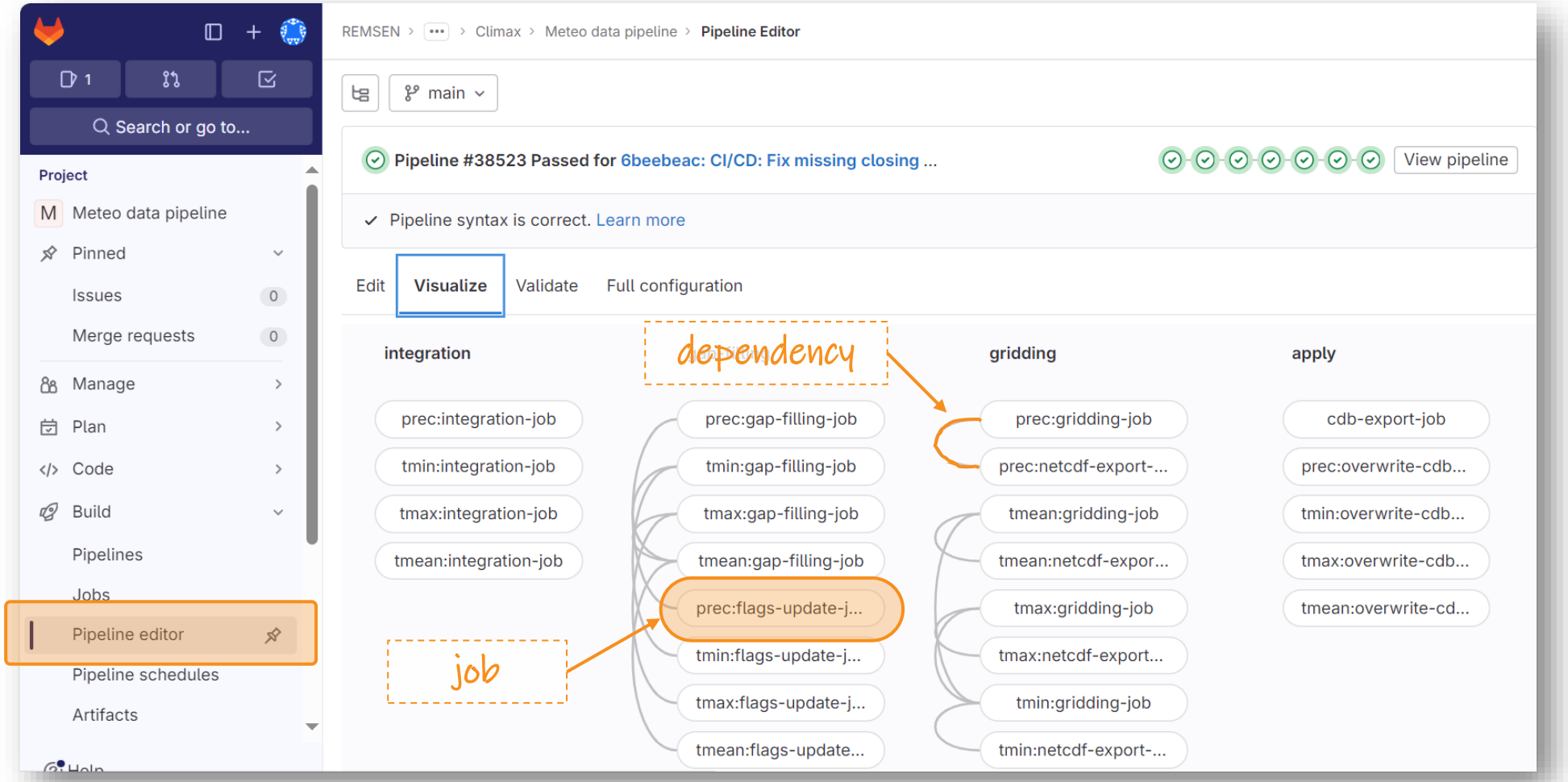
GitLab’s out of the box:
The Pipeline Editor



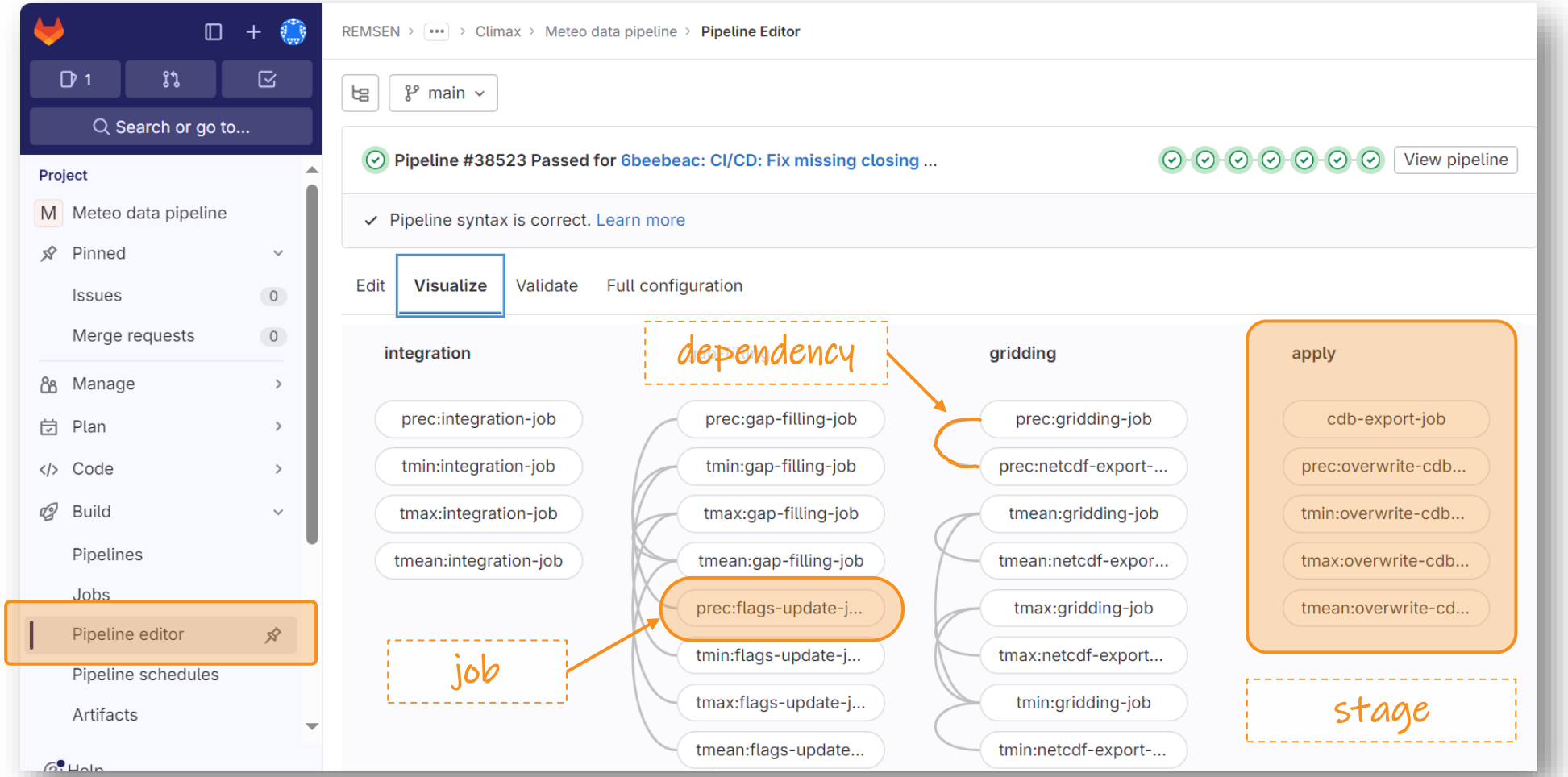
GitLab's out of the box:
The Pipeline Editor



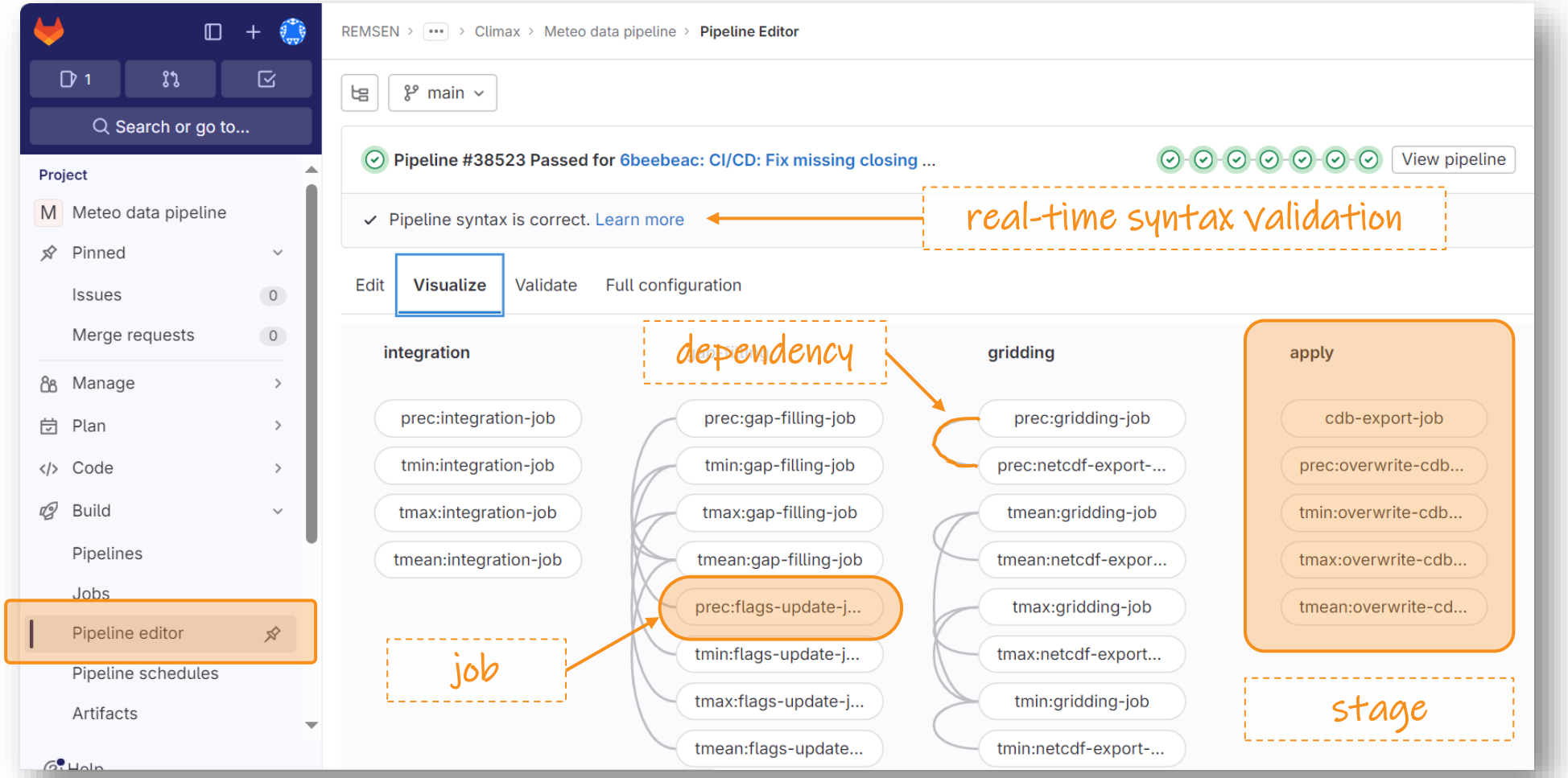
GitLab's out of the box:
The Pipeline Editor



GitLab's out of the box:
The Pipeline Editor



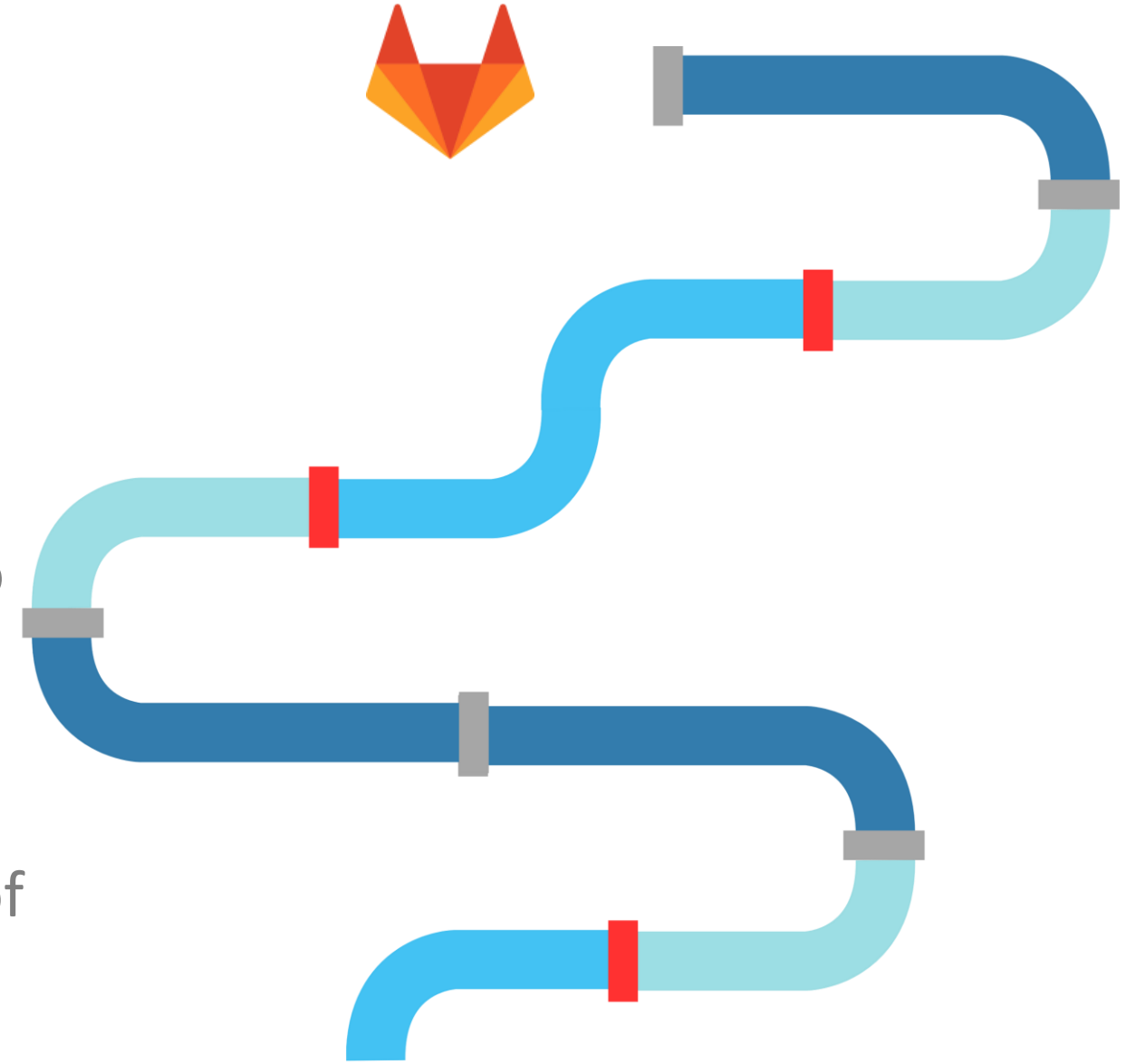
GitLab's out of the box:
The Pipeline Editor



What's next

What's next

- Finalized implementation of the GitLab pipeline
- Automating the connectors with online catalogues
- Set up of webhooks for reporting to MS Teams channel(s)
- Implementation of a monitoring dashboard with GIS-based visualization of the data store and of near-real time updated fields



Thank you!

Contact us:

alice.crespi@eurac.edu

elena.maines@eurac.edu

piero.campalani@eurac.edu