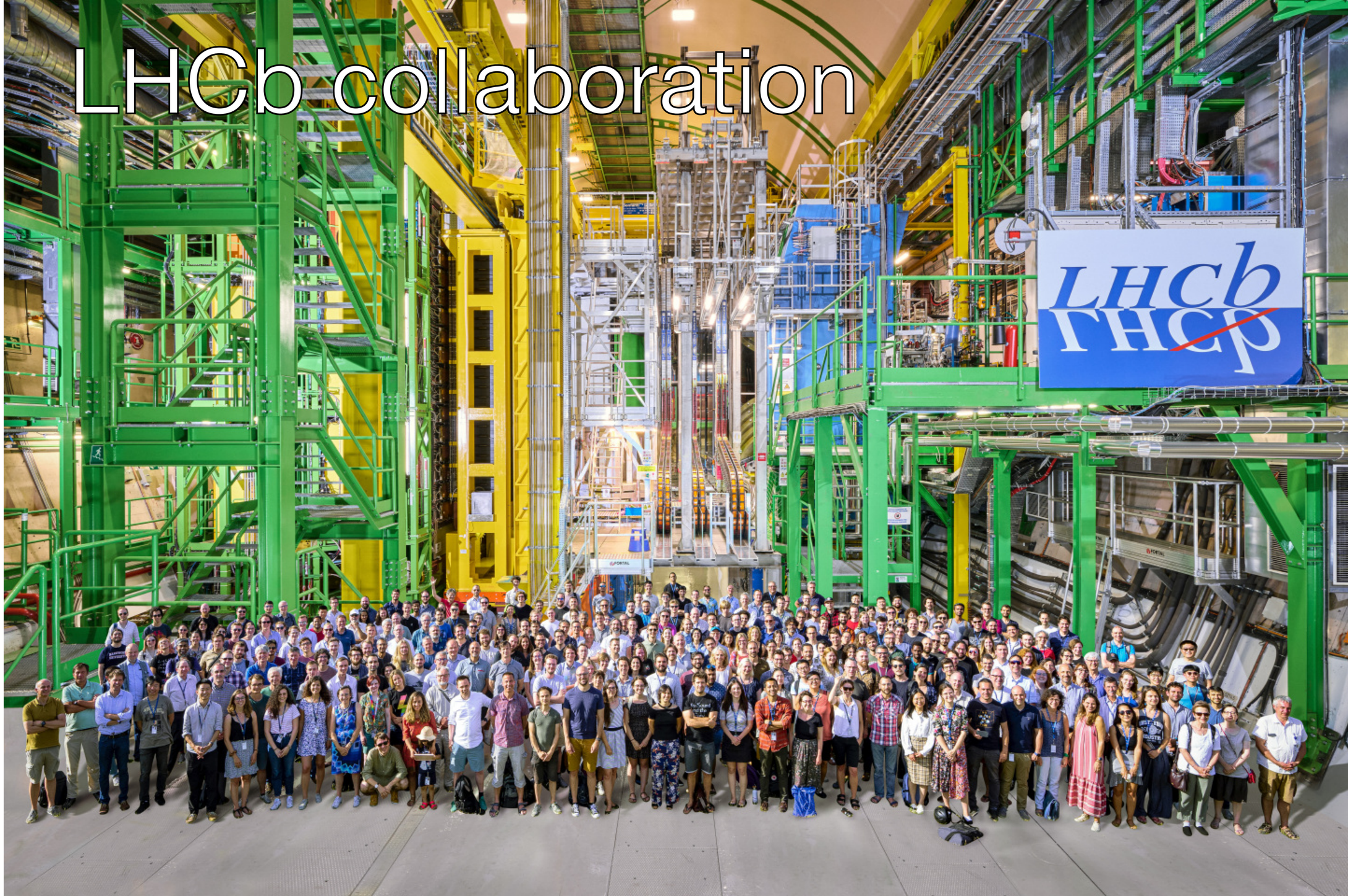
IN LHCb

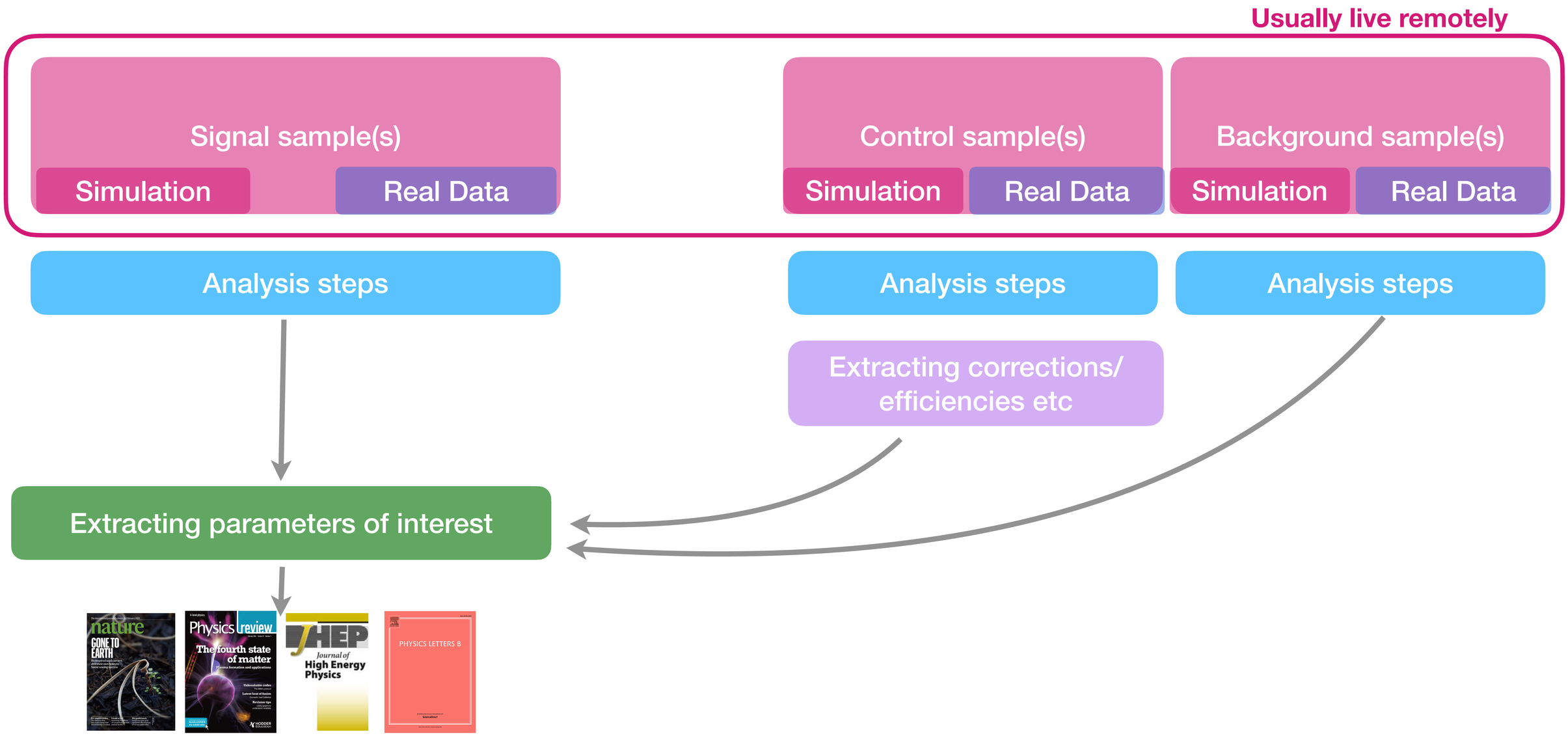
Valeriia (Lera) Lukashenko
UZH

Workshop on workflow languages for HEP analysis, 4 April 2024

LHCb collaboration

~1500 collaborators
~20 countries
100 institutes

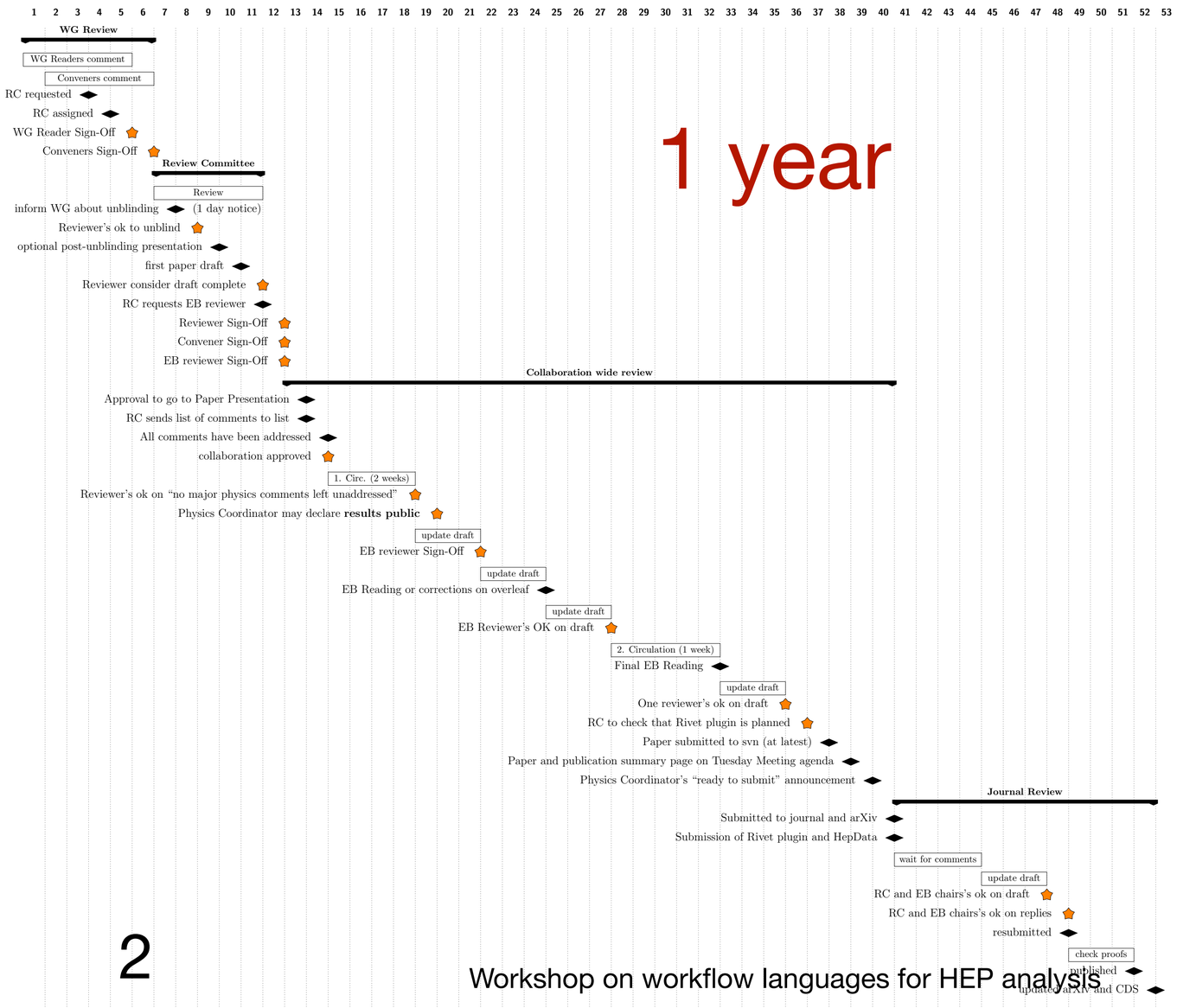



Usually live remotely
Signal sample(s)
Simulation
Real Data
Control sample(s)
Simulation
Real Data
Background sample(s)
Simulation
Real Data
Analysis steps
Analysis steps
Analysis steps
Extracting corrections/ efficiencies etc
Extracting parameters of interest
nature
GONE TO EARTH
Physics review
The fourth state of matter
JHEP
Journal of High Energy Physics
PHYSICS LETTERS B

Timeline of a modern LHCb analysis

x
years of development

+



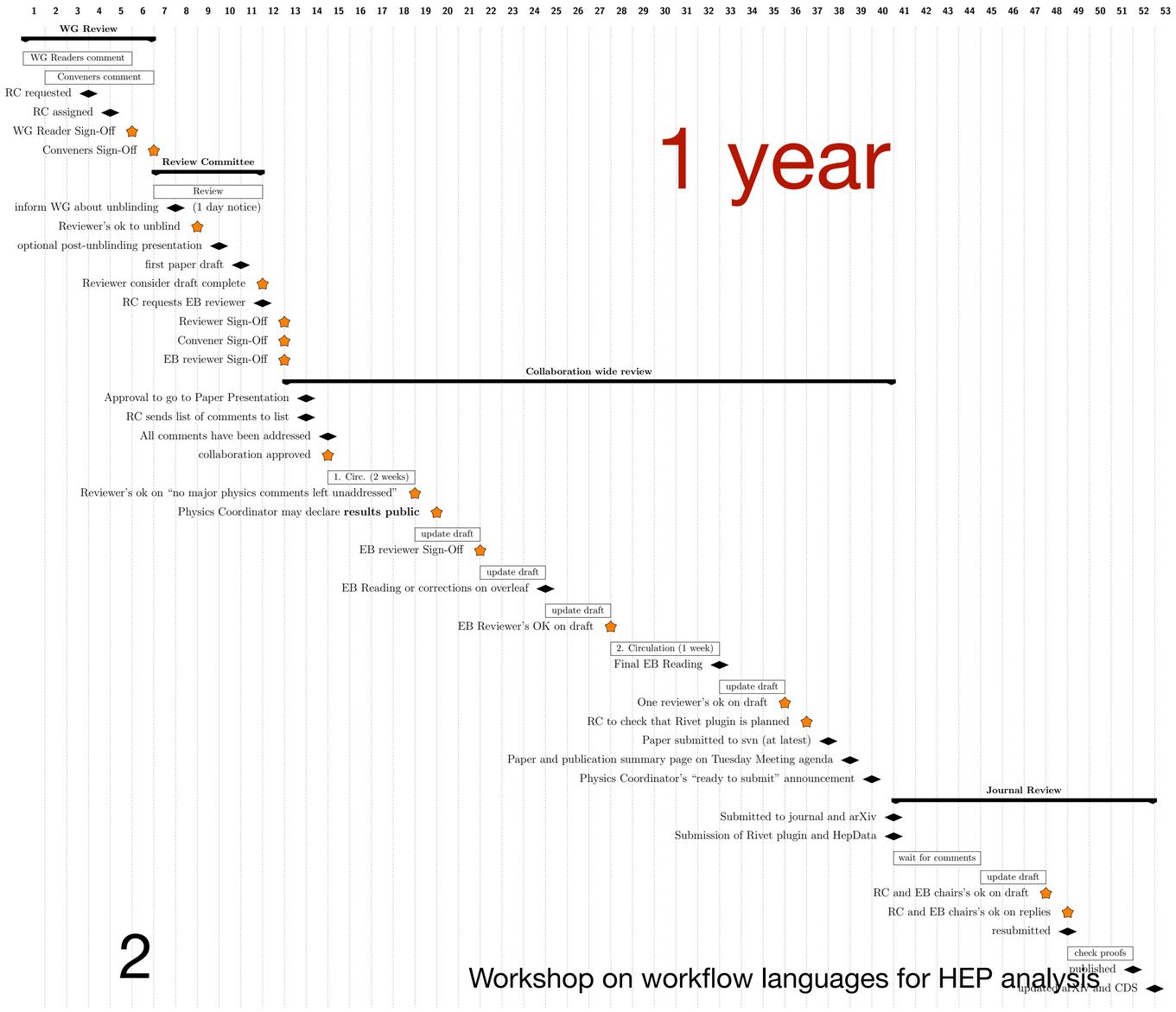
Timeline of a ~~modern~~ LHCb analysis

6 years ago

x

years of development

+



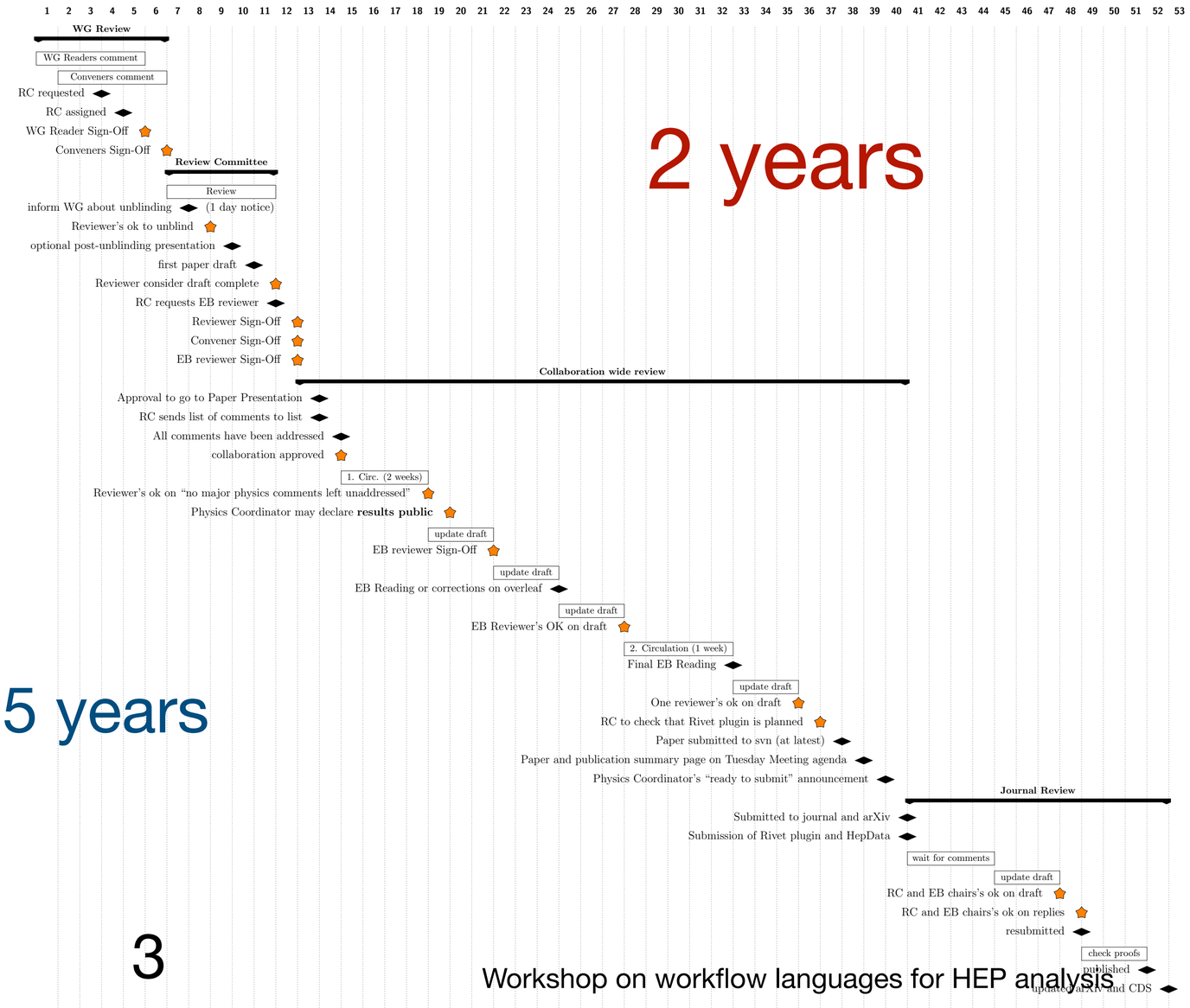
Approximate timeline of a $B_s^0 \to J/\psi\phi$ analysis

~2

years of development

if not including
partial Run 2 analysis,
which another ~ 3 years

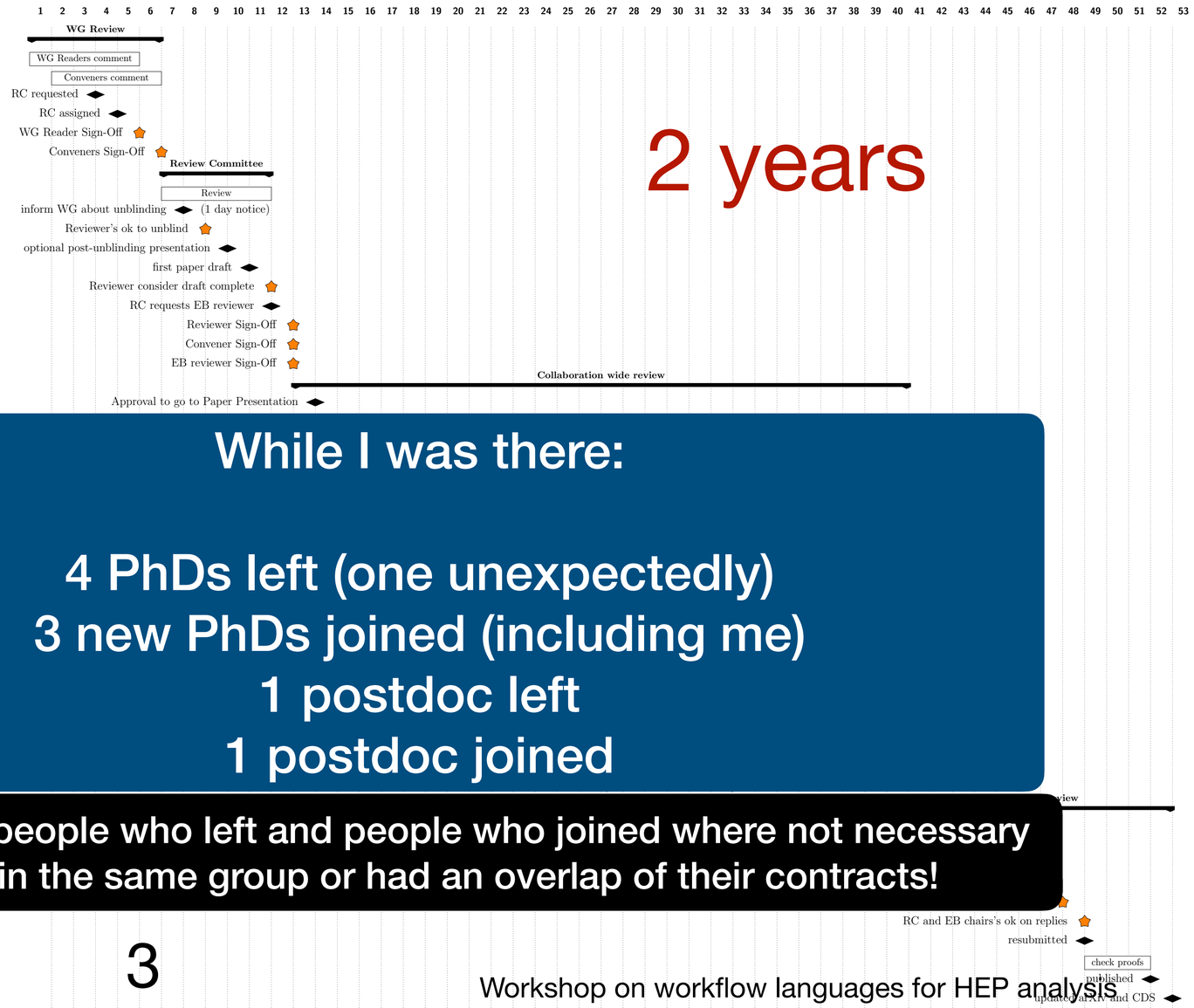
+



2 years

Average PhD lifetime ~ 3.5 years

Approximate timeline of a $B_s^0 \to J/\psi\phi$ analysis



2 years

~2

years of development

if not including
partial Run 2 analysis,
which another ~ 3 years

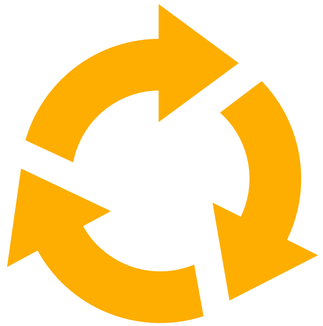
Average PhD lifetir

While I was there:

4 PhDs left (one unexpectedly)
3 new PhDs joined (including me)
1 postdoc left
1 postdoc joined

Note: that people who left and people who joined where not necessary hired in the same group or had an overlap of their contracts!

The difference of big analysis - it is not just me!

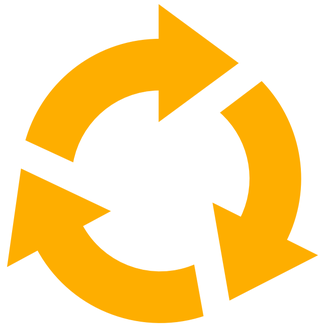


Huge people turn over

No time to train the newcomer

Archeology of others code

The difference of big analysis - it is not just me!



Huge people turn over

No time to train the newcomer

Archeology of others code

Used by all groups

Used by me

Not explored

Portability

Configurability

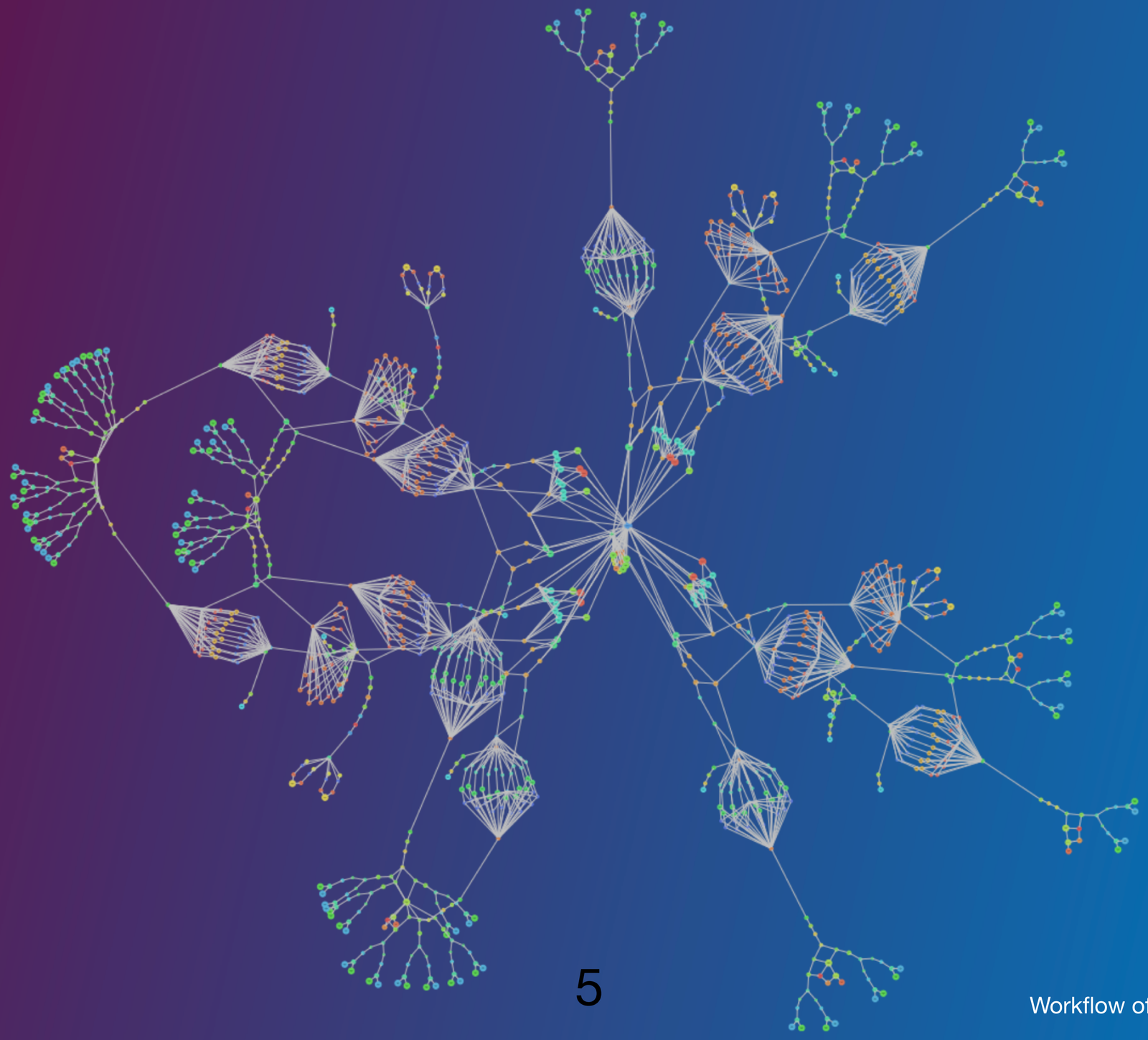
Readability

Modularization

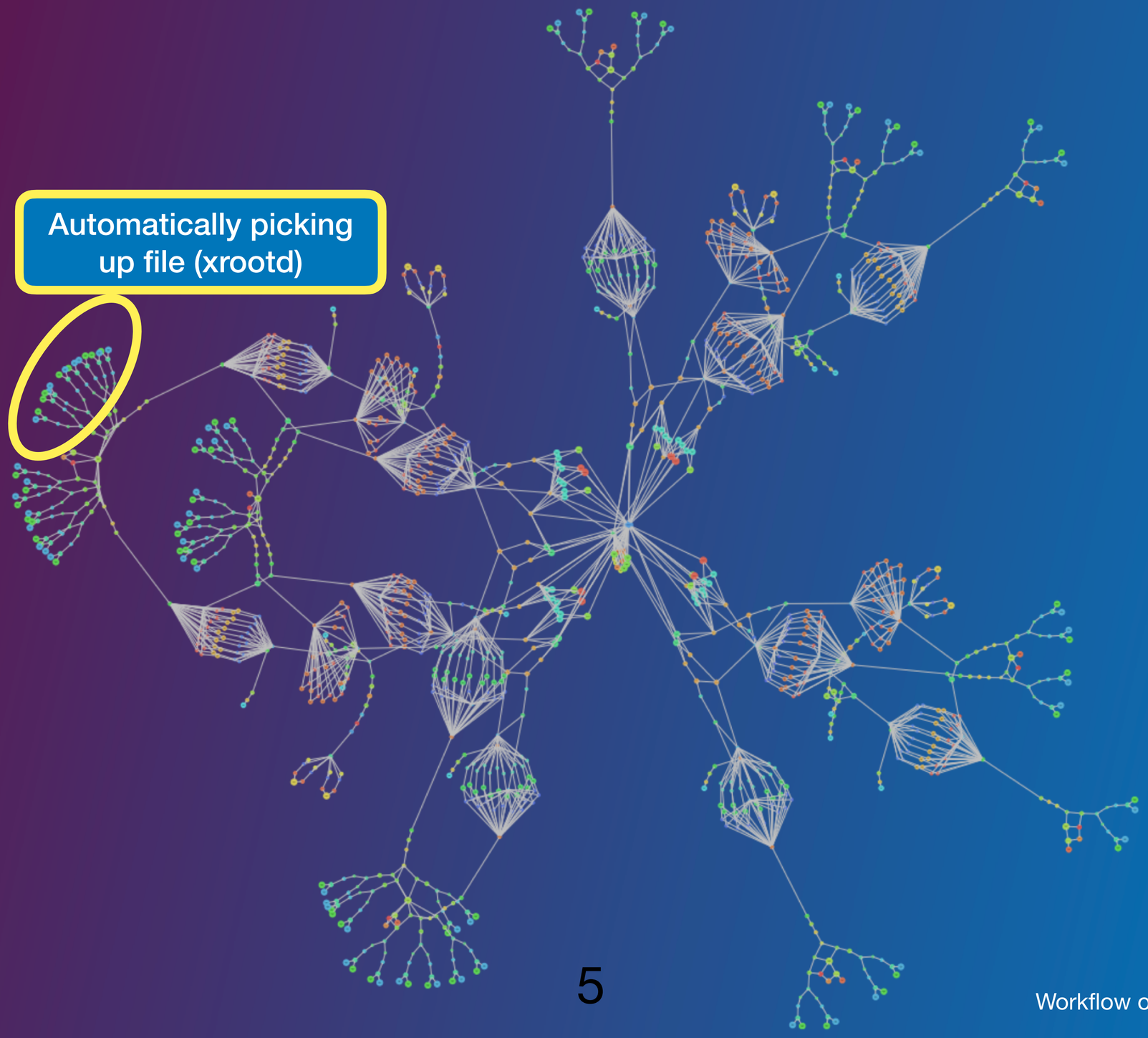
Scalability

snakemake

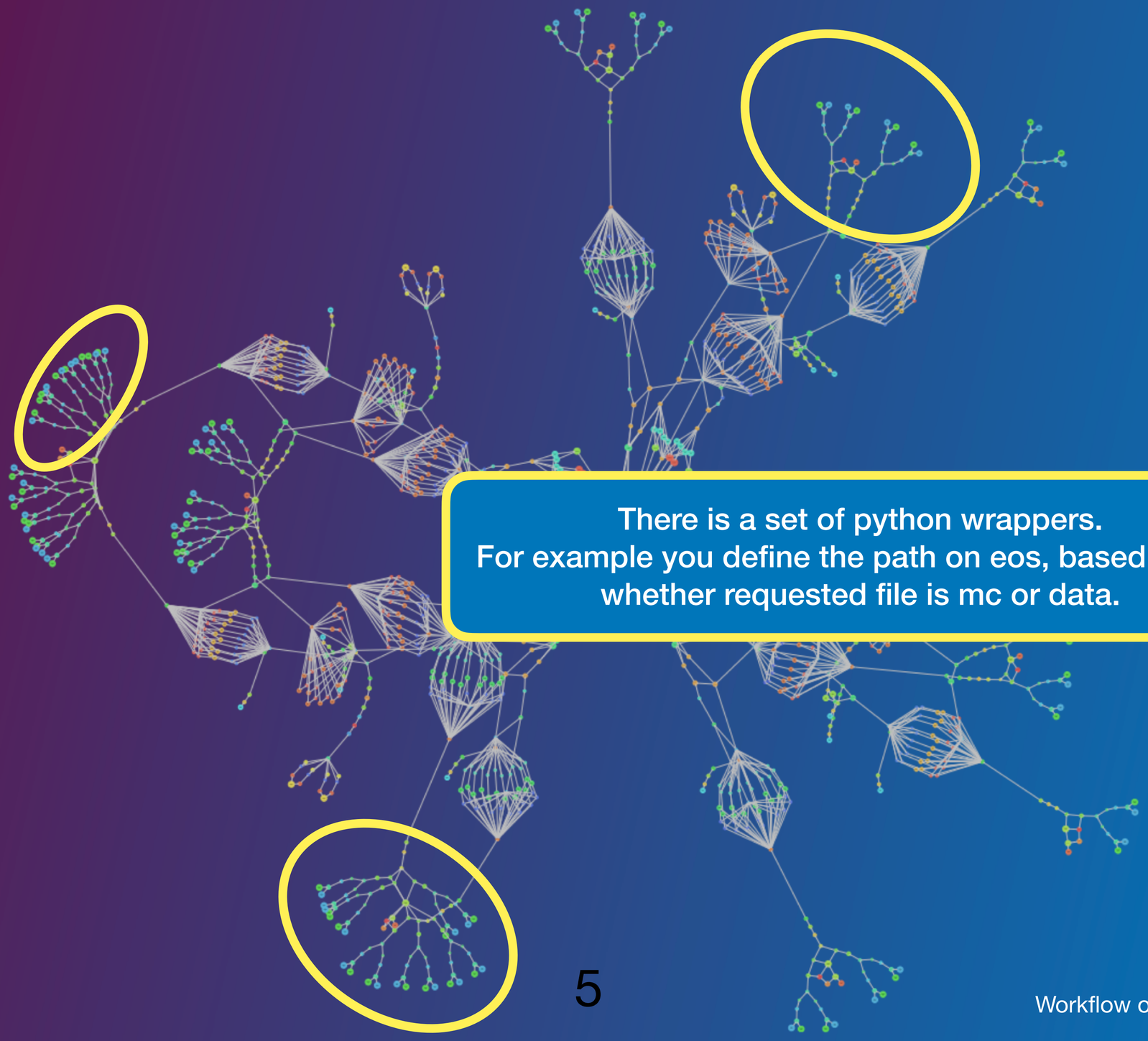
Transparency



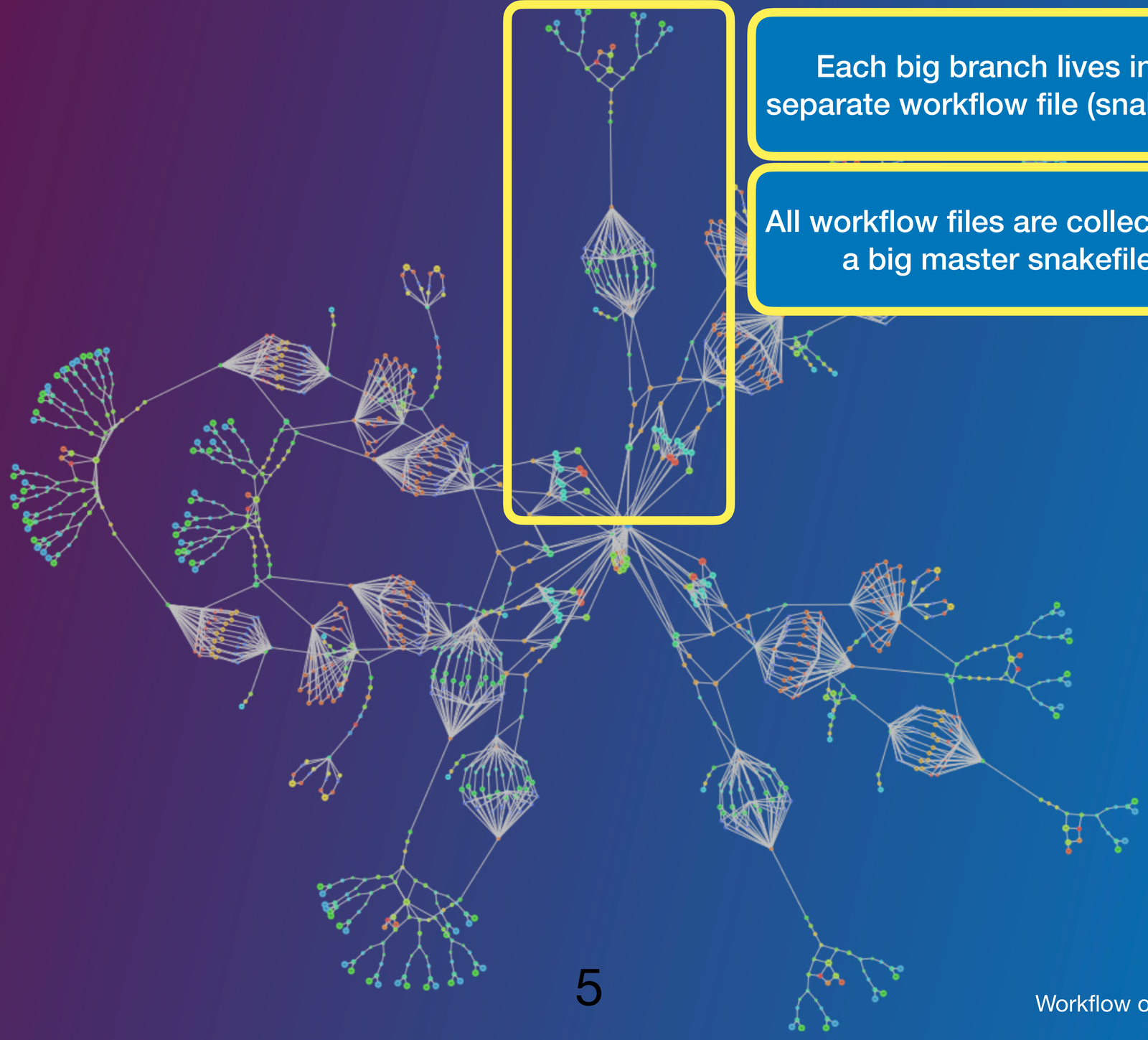
Workflow of Bs2JpsiPhi by V. Chobanova



Workflow of Bs2JpsiPhi by V. Chobanova



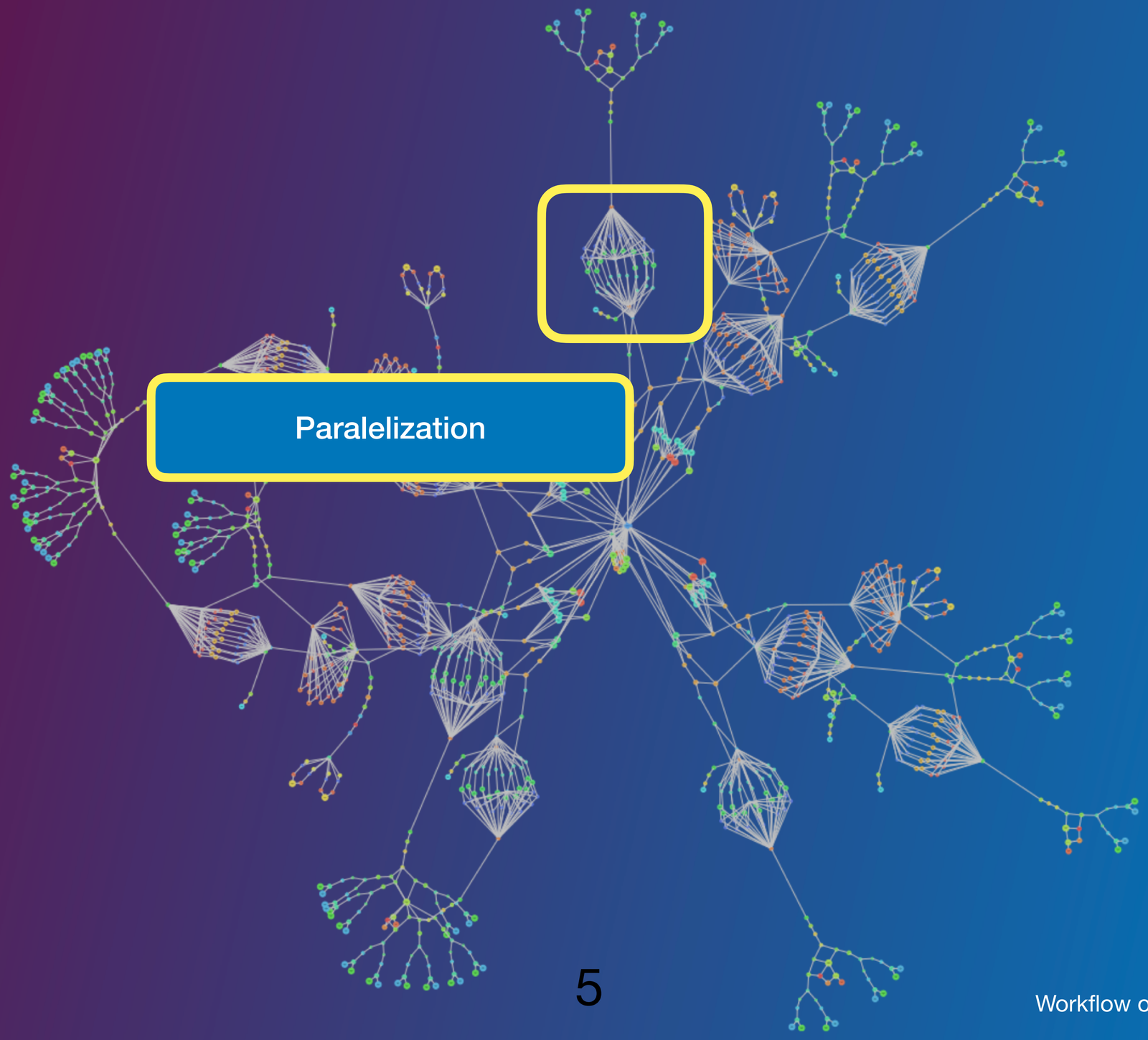
There is a set of python wrappers.
For example you define the path on eos, based on the whether requested file is mc or data.



Each big branch lives in a separate workflow file (snakefile)

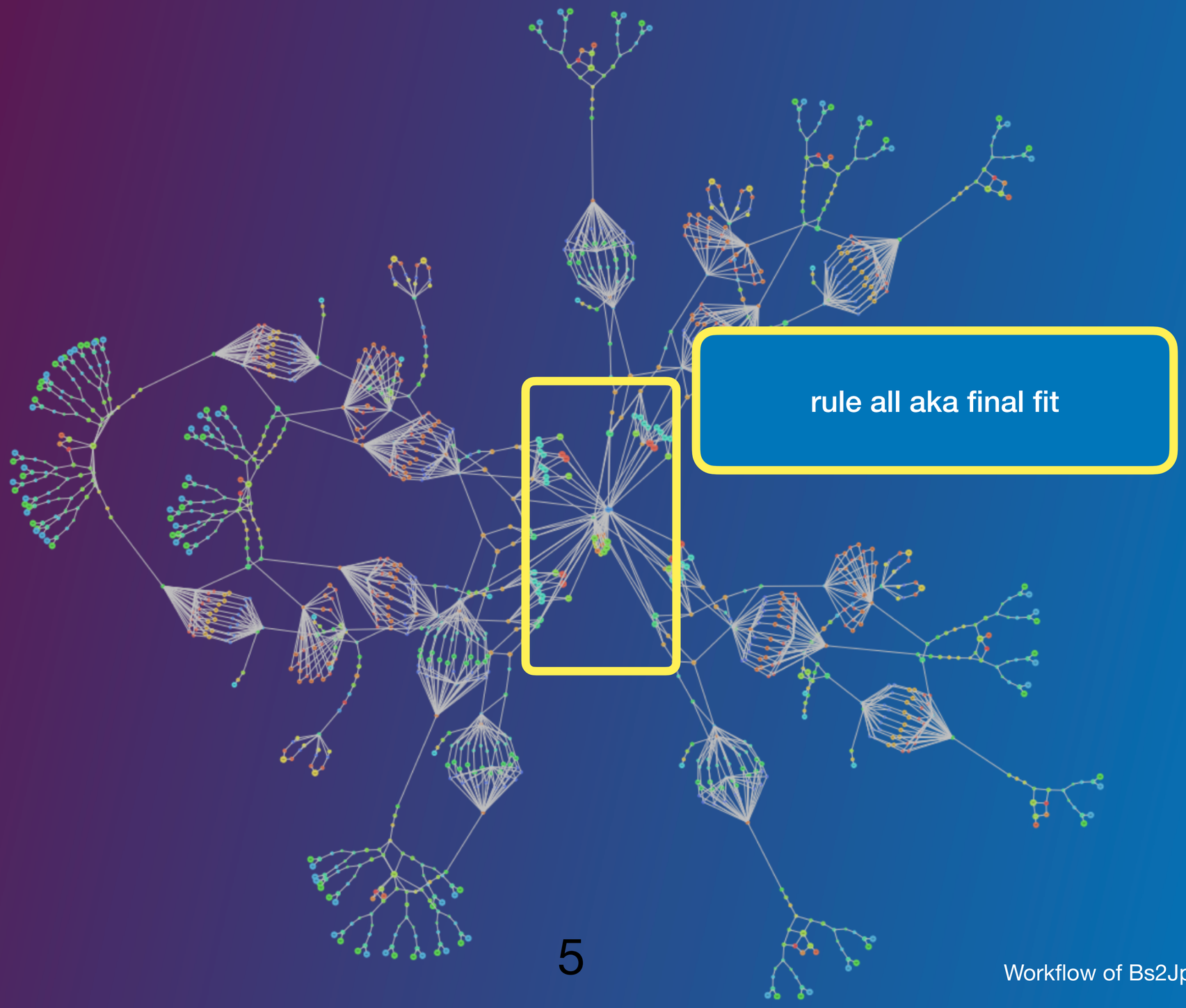
All workflow files are collected in a big master snakefile

Workflow of Bs2JpsiPhi by V. Chobanova

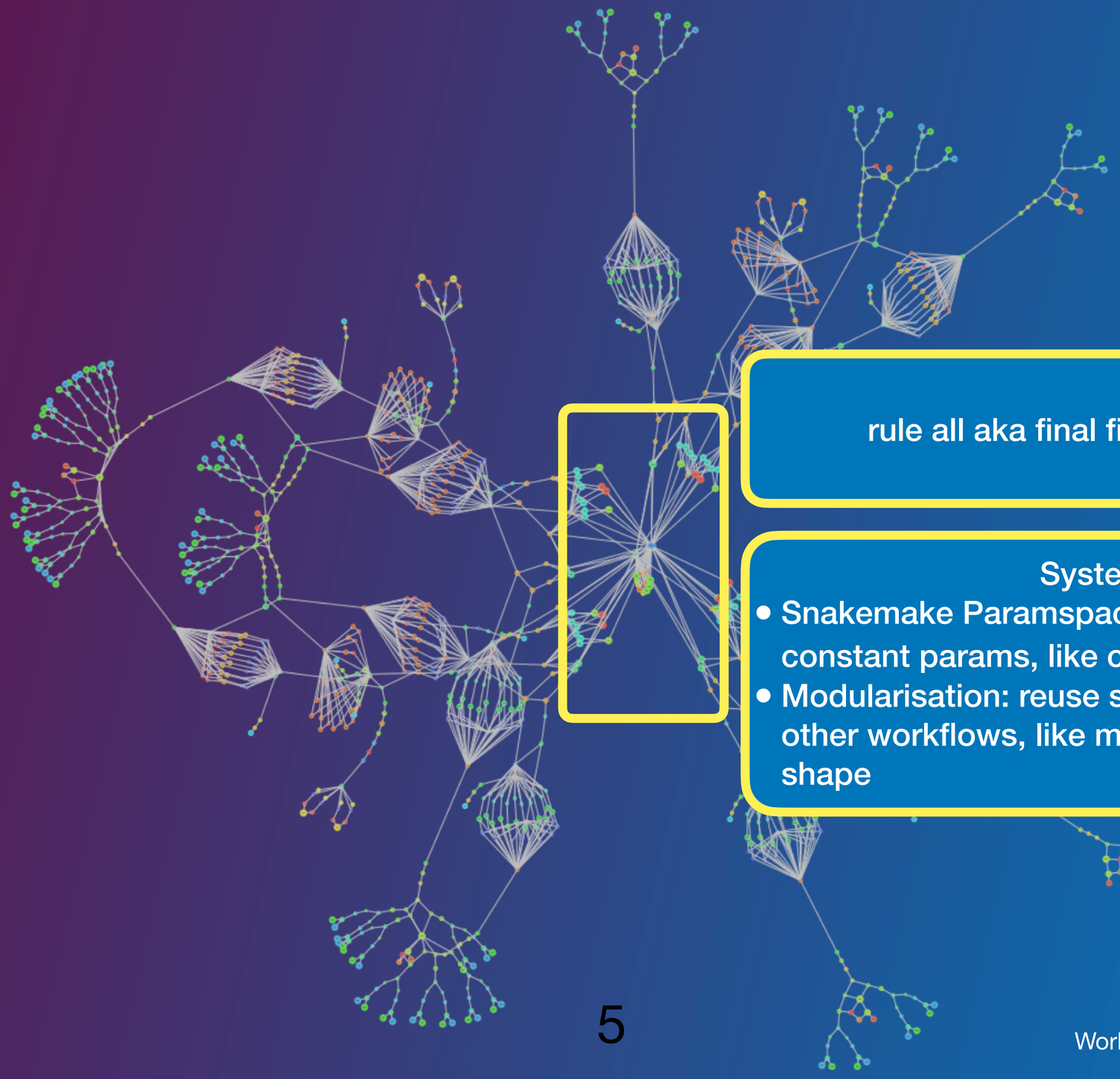


Paralelization

Workflow of Bs2JpsiPhi by V. Chobanova



Workflow of Bs2JpsiPhi by V. Chobanova



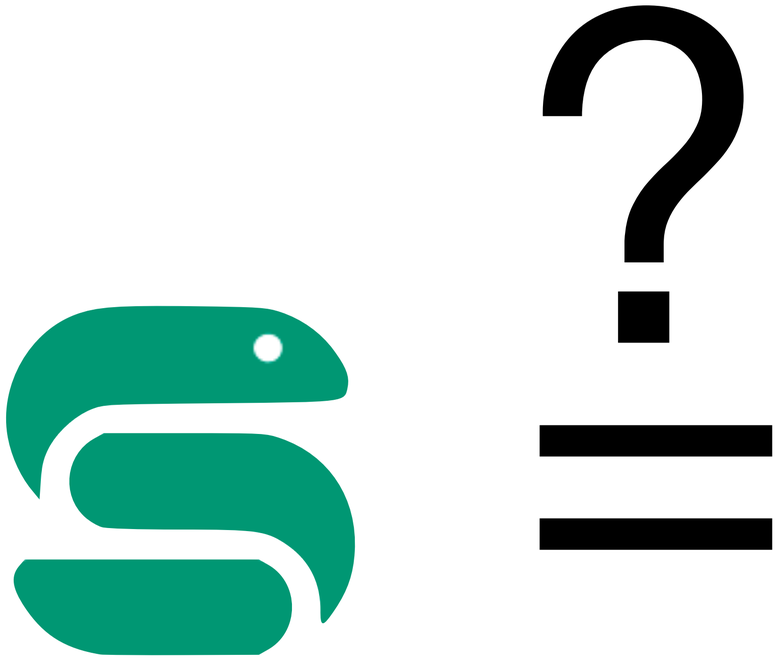
rule all aka final fit

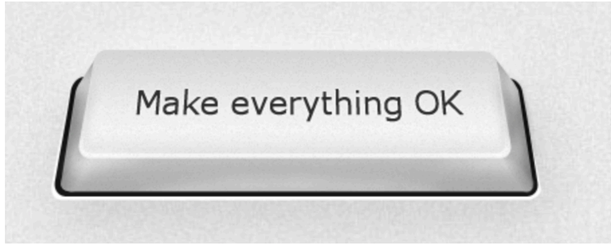
Systematics:

- Snakemake Paramspace for looping over constant params, like conservative +/- σ
- Modularisation: reuse slightly modified rules from other workflows, like mass fit with a different shape

Workflow of Bs2JpsiPhi by V. Chobanova

Danger



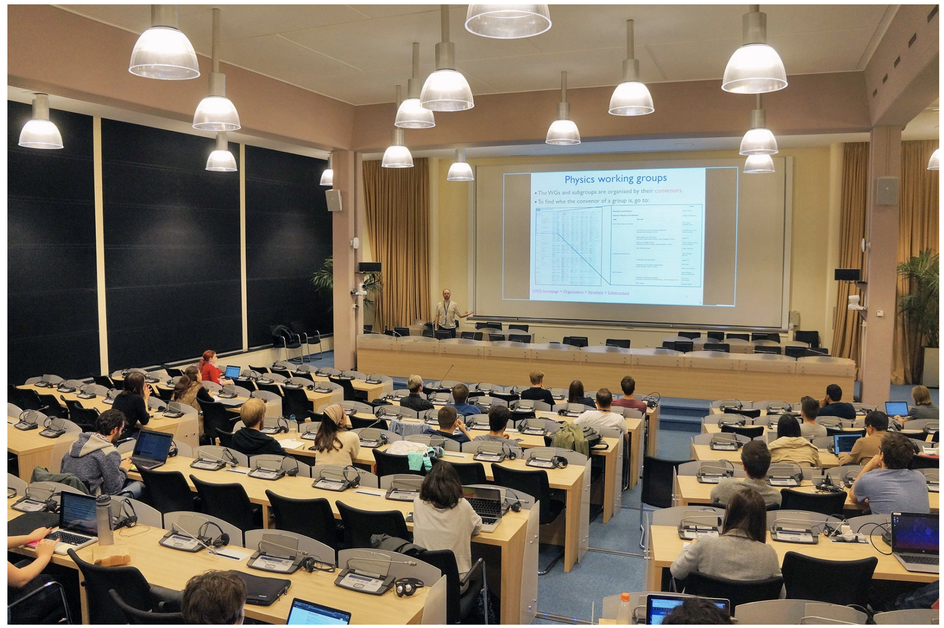


make everything ok button

It is dangerous to rerun the workflow blindly, especially inherited

Combat with gitlab-ci and automated snakemake unit tests

How to increase the number of users?



21 September 2011

LHCb Publication Procedure

This document describes the steps that should be followed for the publication of analyses using data from the LHCb experiment.

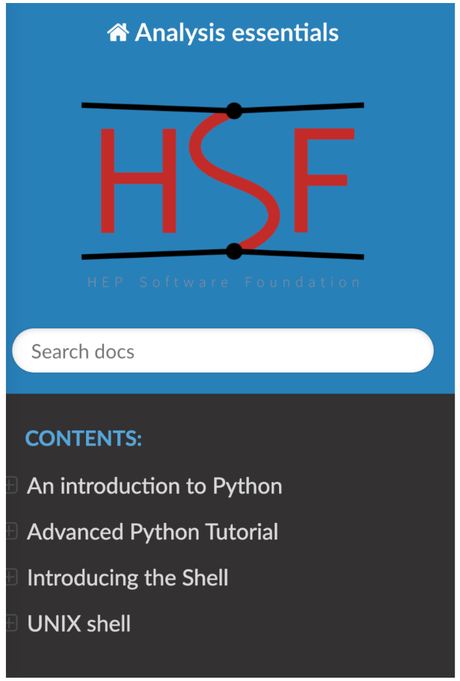
Promote:

training more and more people

Enforce:

no analysis is published without Snakemake

Starterkit: on-boarding training



@

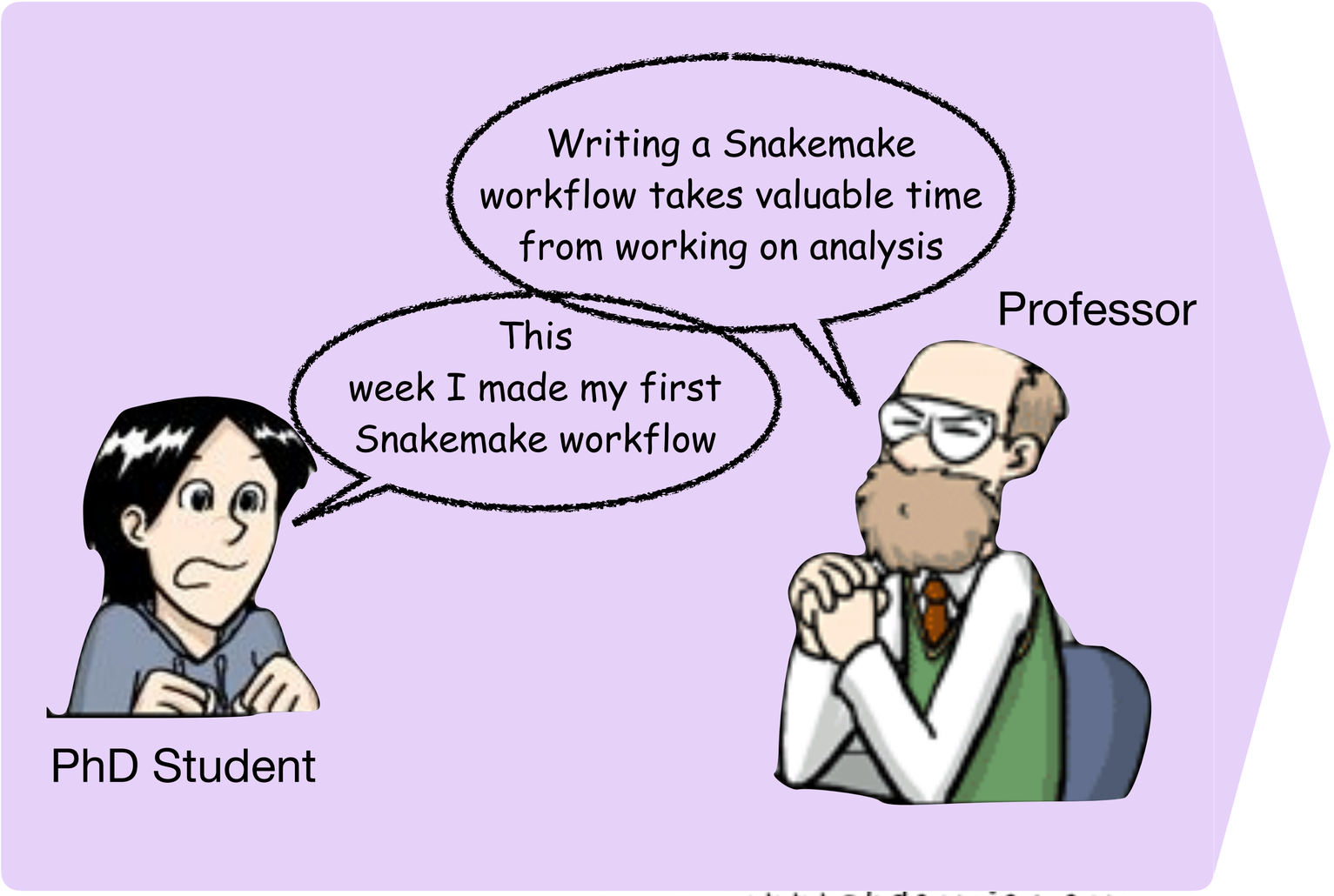
Attendance:
~40 in person
~100 online

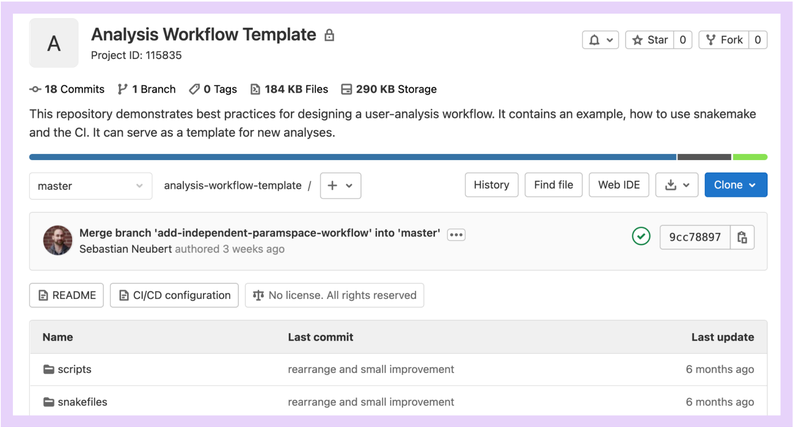
Snakemake lesson is one of the most well received

Starterkit advice : write workflow as soon as you start analysis

The LHCb snakemake training lives **under HEP Software Foundation training umbrella** - available for anyone to use

A common problem after Starterkit





LHCb snakemake template

Discussion points

- It is more than just saving workflow for the purposes of analysis preservation. For big analysis (> X people) having a defined workflow is a necessity to make sure the things are up-to-date and do not get lost. *What should be guidelines for the big analysis groups? Are they different from small ones?*

- There is a huge danger when the workflow is considered “working” it is less likely that individual jobs outputs get checked regularly. I myself relied too much on “this is an automatic procedure and therefore trivial”, which is a logical mistake. *How do we promote more testing in addition to workflow? How to incorporate unit tests in the best way?*

- Promotion among the older generation: this is a waste of time, when you could do physics. *How to change this?*

- *In a view of upcoming upgrades (and the just finished upgrade of LHCb) - should we prepare workflows for the early measurement/data and use it as a monitoring tool?*