UNIVERSITÉ
DE GENÈVE
FACULTÉ DES SCIENCES
Département d'astronomie

Reusable cross-calibration workflows

Dynamically composed unit tests for data and data analysis

Volodymyr SAVCHENKO

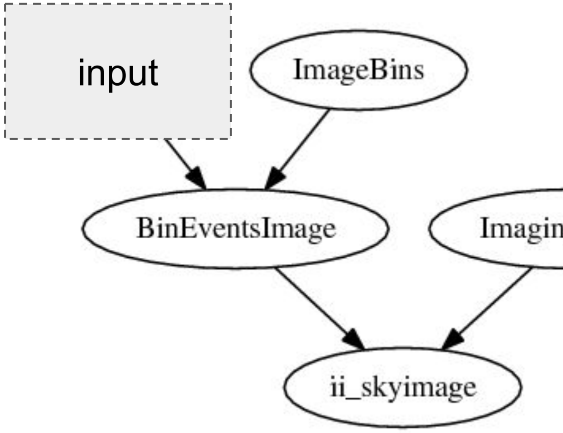
IACHEC 2019

Outline of the problem

- Database of analysis results (IACHECdb), *the objects*
- Scientific software available and documented, *the arrows*:
    - Github, gitlab (almost no metadata)
    - ASCL (metadata includes domains)

- Goal - simplify **finding software**, feeding it with data and **executing** it, by adopting formal models for formulating and evaluating **data analysis workflows** *(~pipelines)*, to*:*
    - Allow to find (also **automatically**), define and execute **some calibration and verification workflows**
    - Help automating execution over diverse **distributed resources**
        - bring code to the data storage or data stream
        - code and data to the CPUs (cloud and grid)

Workflow with an input



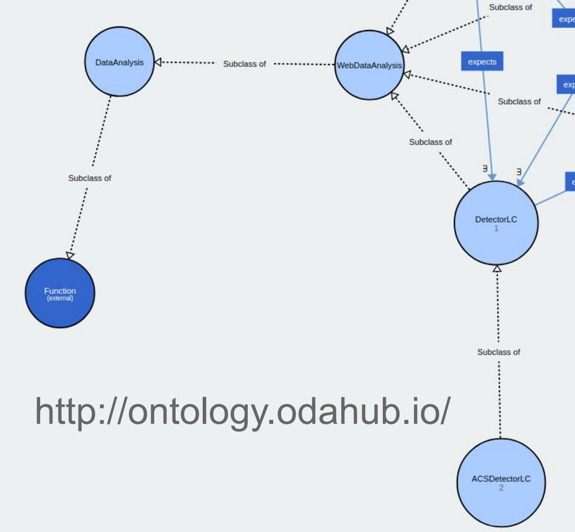
Example

- Crab : may be integrated into routine analysis pipeline
    a. INTEGRAL yields new spectrum
        - New observations
        - New software or calibration
        - New reference models
    b. The workflow composition engine identifies Crab cross-calibration workflow
    c. Fetches last available cross-calibration data e.g. from the **IACHEC db**
    d. Tries different workflows with different methods (**xspec, spex, 3ml**) and instruments, proliferating sharing and **re-use of good methods**
    e. Summarize and allow for review, public or private

- Vela X-1, Her X-1: variable sources with complex spectra, in addition, require adoption of source knowledge to extrapolate non-simultaneous data

What is needed to enable this

- Develop interfaces for adopting standards (VO, etc) wrapping data analysis methods in process model
- Define input/output type ontology: classify data entities
- Astronomers embed source knowledge in verification workflows
- Experts in methods (statistical, etc) help to define process



http://ontology.odahub.io/

Why this is becoming feasible now

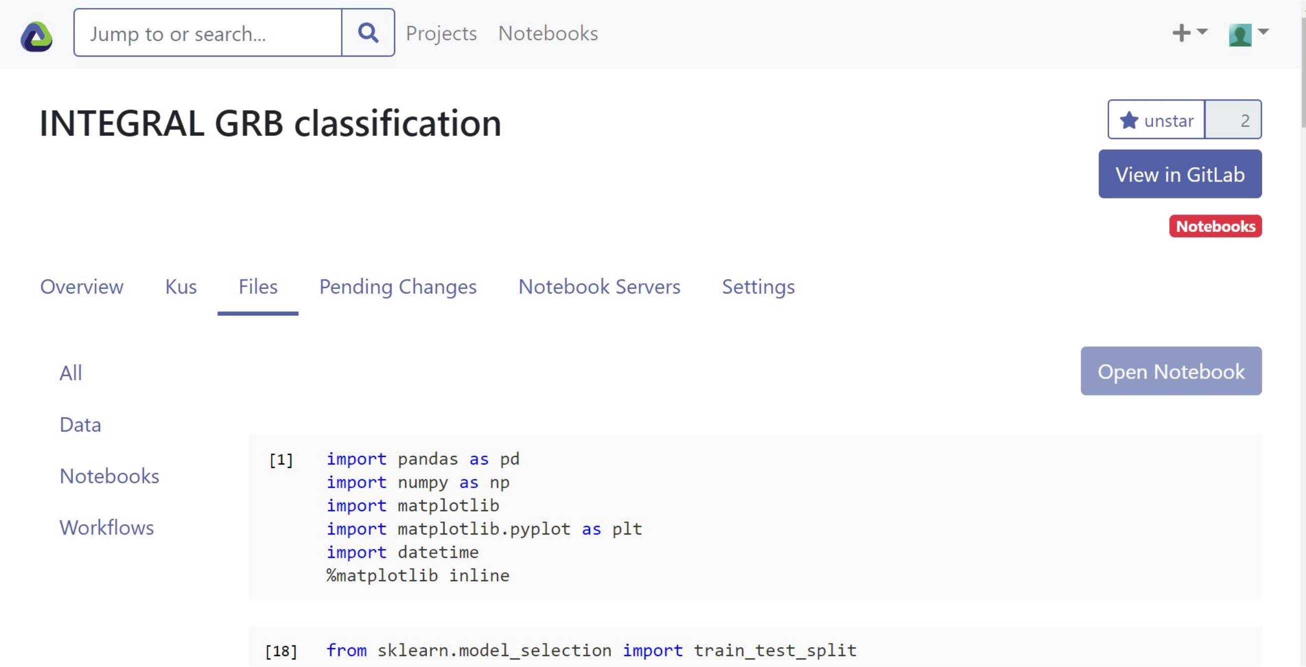
- Development of portable cloud-native technologies allows to execute easily
- Needs in systematic workflow and data management promoted new standards across the industry (CWL, OWL, ...)
- Process as first-class entity is becoming popular: serverless, cloud, etc

Platforms for sharing and exploiting Data Analysis

github/binder, Renku (SDSC/EPFL), KNIME, SEPP (ESA), ...

https://renkulab.io/

- Sharing
- Searching
- Executing
- reusing (building from) workflows

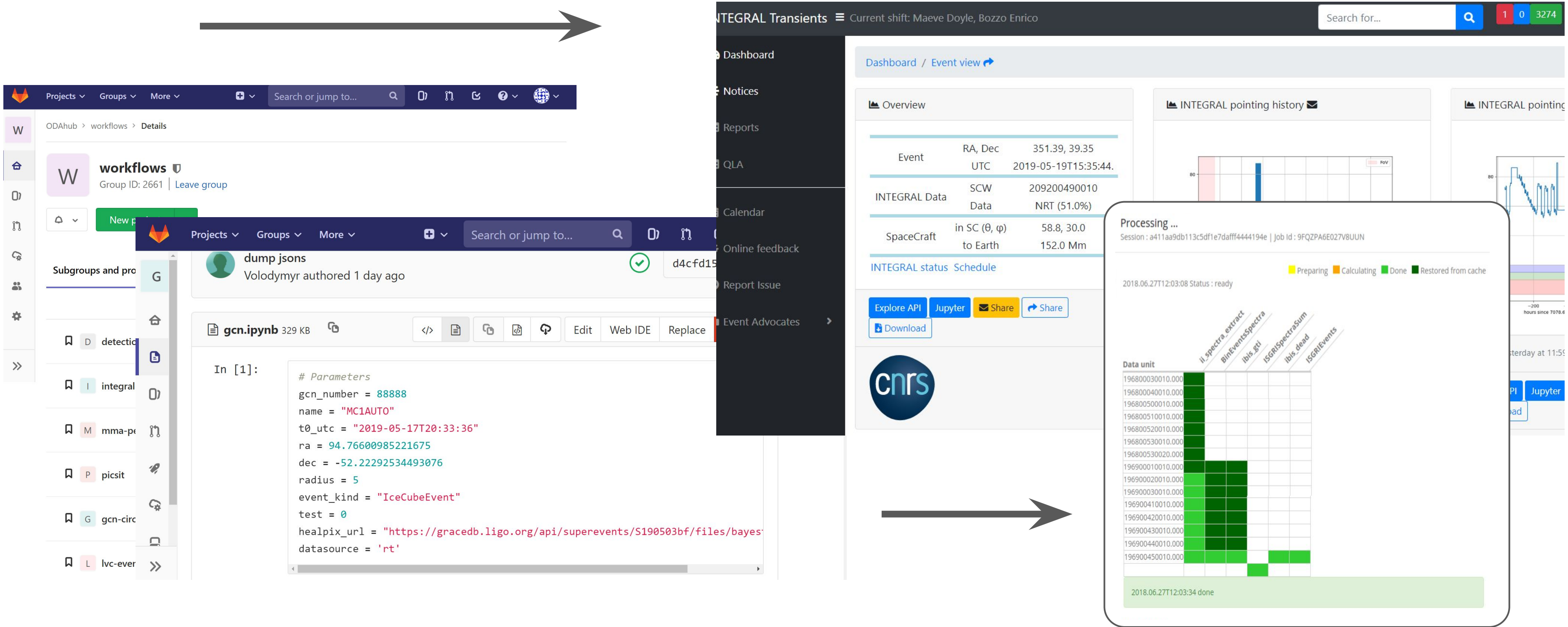


```
[1]   import pandas as pd
      import numpy as np
      import matplotlib
      import matplotlib.pyplot as plt
      import datetime
      %matplotlib inline
```

```
[18]  from sklearn.model_selection import train_test_split
```

Simple example: INTEGRAL transient analysis

A collection of transient analysis workflows is defined by instrument and domain experts, single interface allows shift to get the best results fast



```
# Parameters
gcn_number = 88888
name = "MC1AUTO"
t0_utc = "2019-05-17T20:33:36"
ra = 94.76600985221675
dec = -52.22292534493076
radius = 5
event_kind = "IceCubeEvent"
test = 0
healpix_url = "https://gracedb.ligo.org/api/superevents/S190503bf/files/bayes
datasource = 'rt'
```

Event	RA, Dec	351.39, 39.35
	UTC	2019-05-19T15:35:44.
INTEGRAL Data	SCW	209200490010
	Data	NRT (51.0%)
SpaceCraft	in SC (θ, φ)	58.8, 30.0
	to Earth	152.0 Mm

Processing ...

Session : a411aa9db113c5df1e7dafff4444194e | Job Id : 9FQZPA6E027V8UUN

🟨 Preparing 🟧 Calculating 🟩 Done ■ Restored from cache

2018.06.27T12:03:08 Status : ready

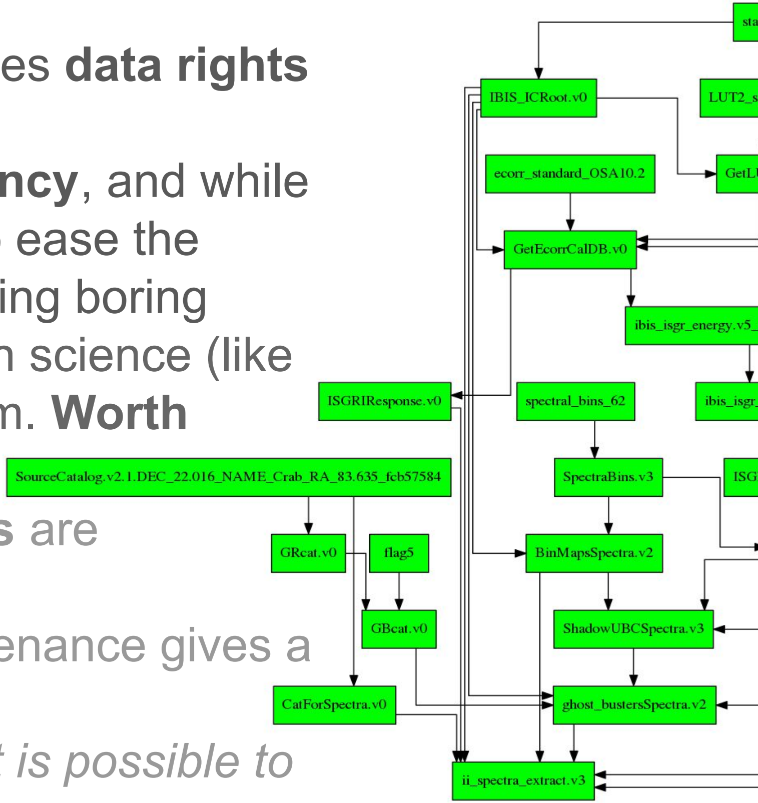
Data unit	ii_spectra_extract	BinEventsSpectra	ibis_gti	ISGRISpectraSum	ibis_dead	ISGRIEvents
196800030010.000	■					
196800040010.000	■					
196800500010.000	■					
196800510010.000	■					
196800520010.000	■					
196800530010.000	■					
196800530020.000	■					
196900010010.000	■	■	■			
196900020010.000	🟩	■	■			
196900030010.000	🟩	■	■			
196900410010.000	🟩	■	■			
196900420010.000	🟩	■	■			
196900430010.000	🟩	■	■			
196900440010.000	🟩	■	■			
196900450010.000	🟩	🟩	🟩		🟩	🟩
				🟩		

2018.06.27T12:03:34 done

Many benefits of managed workflows

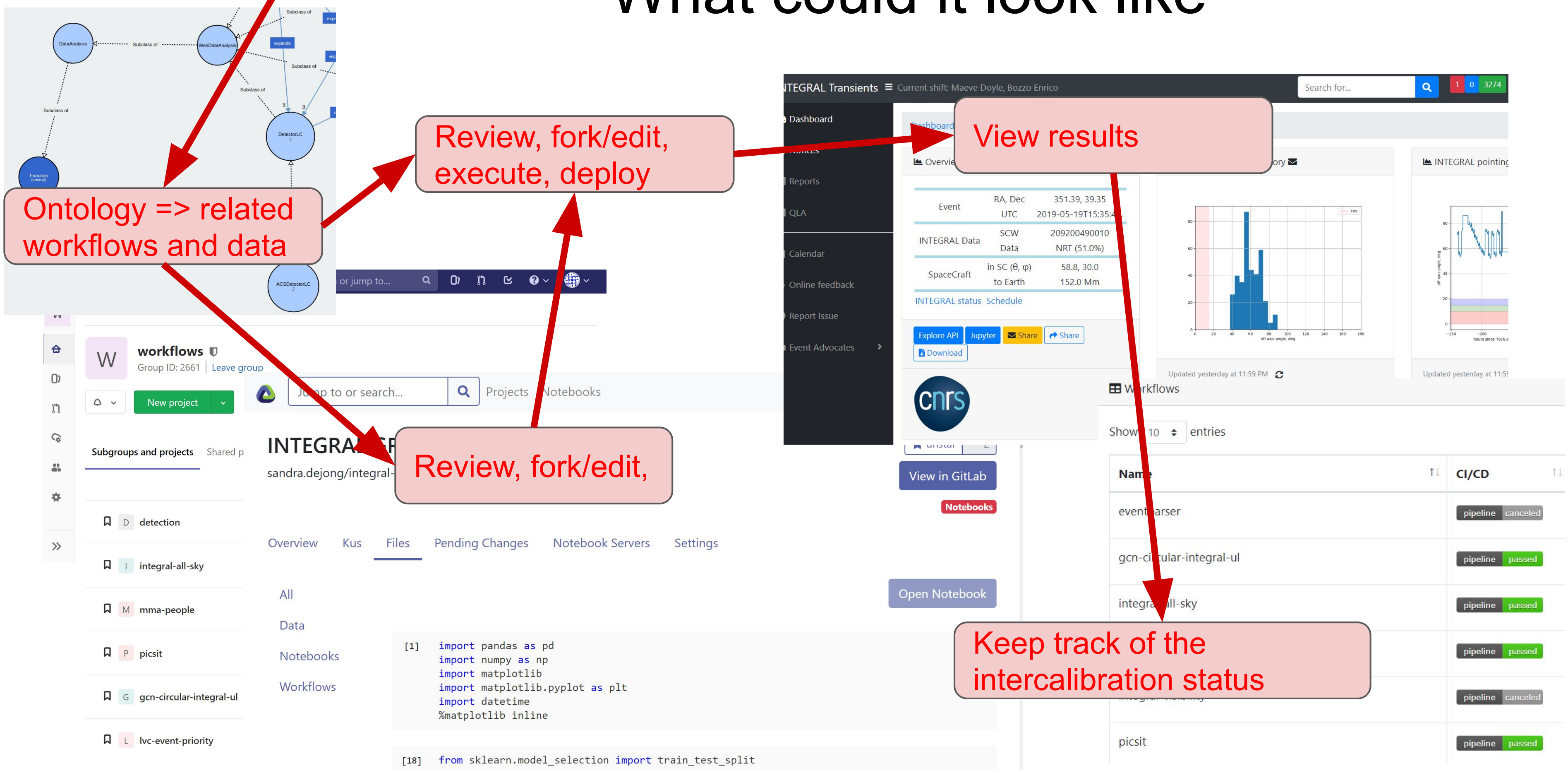
- **Provenance derived from the workflows** induces **data rights and credits** for data and calibrations
- **Automation of verification promotes consistency**, and while it does not replace specialist analysis, it allows to ease the routine and enable processes discouraged by being boring
- Failed regular automated check might even mean science (like variable hard X-ray Crab), but more likely problem. **Worth checking if it costs almost no man-hours**.
- **Adapters between data formats and interfaces** are implemented with workflow
- Meaningful sharing/open impies explaining, provenance gives a perspective on explaining calibration
- *If **calibration** itself can be embedded workflow, it is possible to track impact of **cross-calibration impact on the informativeness** of the results*

Provenance



What could it look like

New software, method, data, idea



Ontology => related workflows and data

Review, fork/edit, execute, deploy

View results

Review, fork/edit,

Keep track of the intercalibration status

Event	RA, Dec	351.39, 39.35
	UTC	2019-05-19T15:35:4
INTEGRAL Data	SCW	209200490010
	Data	NRT (51.0%)
SpaceCraft	in SC (θ, φ)	58.8, 30.0
	to Earth	152.0 Mm

- D detection
- I integral-all-sky
- M mma-people
- P picsit
- G gcn-circular-integral-ul
- L lvc-event-priority

```
[1]   import pandas as pd
      import numpy as np
      import matplotlib
      import matplotlib.pyplot as plt
      import datetime
      %matplotlib inline

[18]  from sklearn.model_selection import train_test_split
```

Name	CI/CD
event arser	pipeline canceled
gcn-ci ular-integral-ul	pipeline passed
integra ll-sky	pipeline passed
	pipeline passed
	pipeline canceled
picsit	pipeline passed

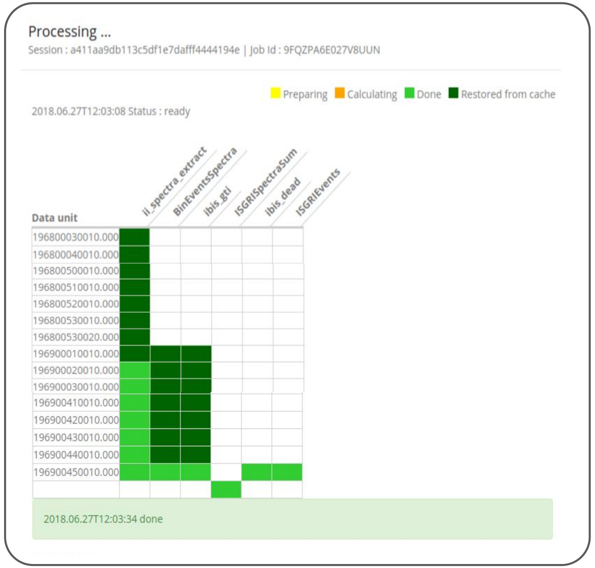
Open (Online) Data Analysis / CDCI

Online analysis s implemented for INTEGRAL (IBIS, JEM-X) and POLAR: serves as source of INTEGRAL data for the calibration as well as for executing (cross)calbration workflows stored in github

https://astro.unige.ch/cdci/astrooda_

ISDC Quick-Look Analysis

As INTEGRAL data arrives to ISDC, Quick Look Analysis is performed, including checks of calibration sanity. It could use more elaborate cross-calibration



Processing ...

Session : a411aa9db113c5df1e7dafff4444194e | Job Id : 9FQZPA6E027V8UUN

🟨 Preparing 🟧 Calculating 🟩 Done ■ Restored from cache

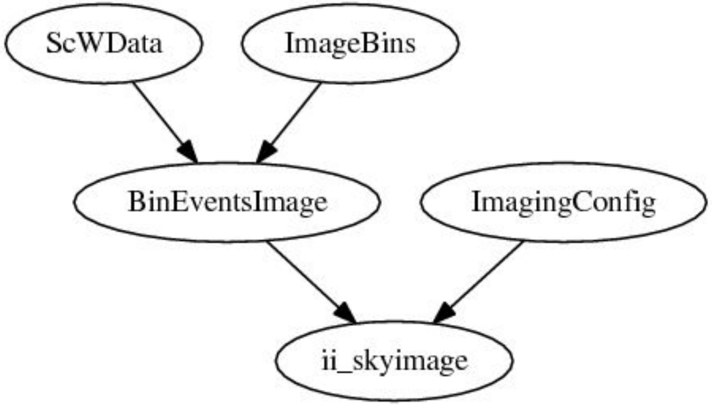
2018.06.27T12:03:08 Status : ready

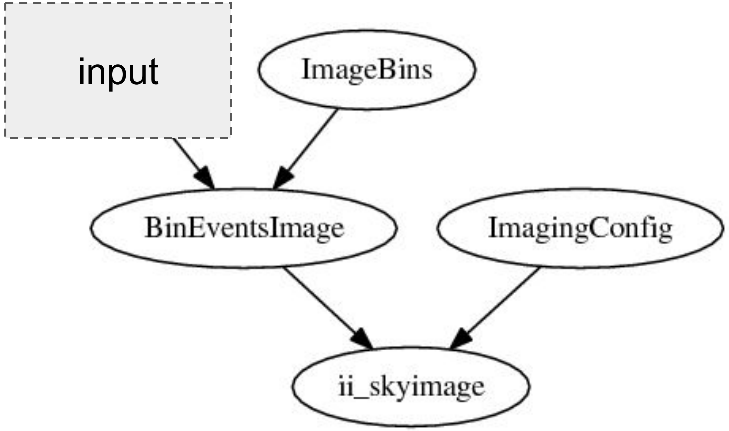
Data unit	ii_spectra_extract	BinEventsSpectra	ibis_gti	ISGRISpectraSum	ibis_dead	ISGRIEvents
196800030010.000	■					
196800040010.000	■					
196800500010.000	■					
196800510010.000	■					
196800520010.000	■					
196800530010.000	■					
196800530020.000	■					
196900010010.000	■	■	■			
196900020010.000	🟩	■	■			
196900030010.000	🟩	■	■			
196900410010.000	🟩	■	■			
196900420010.000	🟩	■	■			
196900430010.000	🟩	■	■			
196900440010.000	🟩	■	■			
196900450010.000	🟩	🟩	🟩		🟩	🟩
				🟩		

2018.06.27T12:03:34 done

What is a workflow

1. Workflow is an arrow in the process category, a morphism of the data
2. Workflow may be a composition of other workflows


ScWData
ImageBins
BinEventsImage
ImagingConfig
ii_skyimage


input
ImageBins
BinEventsImage
ImagingConfig
ii_skyimage