Reproducibility and reuse in data-driven sciences: from provenance to summaries

Franck Michel
Univ. Côte d'Azur, CNRS, Inria, I3S

Alban Gaignard
Institut du thorax, CNRS, Univ. de Nantes

Journées APSEM 2019
15 au 18 Octobre 2019, Toulouse

Repeat > Replicate > Reproduce > Reuse

S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. Le Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal, C. Blanchet, ***Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities****, Future Generation Computer Systems, Volume 75, 2017,* [*https://doi.org/10.1016/j.future.2017.01.012*](https://doi.org/10.1016/j.future.2017.01.012) *.*

Repeat > Replicate > Reproduce > Reuse

Same experiment

Same setup

Same lab

S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. Le Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal, C. Blanchet, ***Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities****, Future Generation Computer Systems, Volume 75, 2017, https://doi.org/10.1016/j.future.2017.01.012 .*

Repeat > Replicate > Reproduce > Reuse

Same experiment	Same experiment
Same setup	Same setup
Same lab	~~Same lab~~

S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. Le Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal, C. Blanchet, ***Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities****, Future Generation Computer Systems, Volume 75, 2017, https://doi.org/10.1016/j.future.2017.01.012 .*

Repeat > Replicate > Reproduce > Reuse

Same experiment	Same experiment	Same experiment
Same setup	Same setup	~~Same setup~~
Same lab	~~Same lab~~	~~Same lab~~

*S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. Le Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal, C. Blanchet, **Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities**, Future Generation Computer Systems, Volume 75, 2017, https://doi.org/10.1016/j.future.2017.01.012 .*

Repeat > Replicate > Reproduce > Reuse

Same experiment	Same experiment	Same experiment
Same setup	Same setup	~~Same setup~~
Same lab	~~Same lab~~	~~Same lab~~

new ideas,
new experiment,
some commonalities

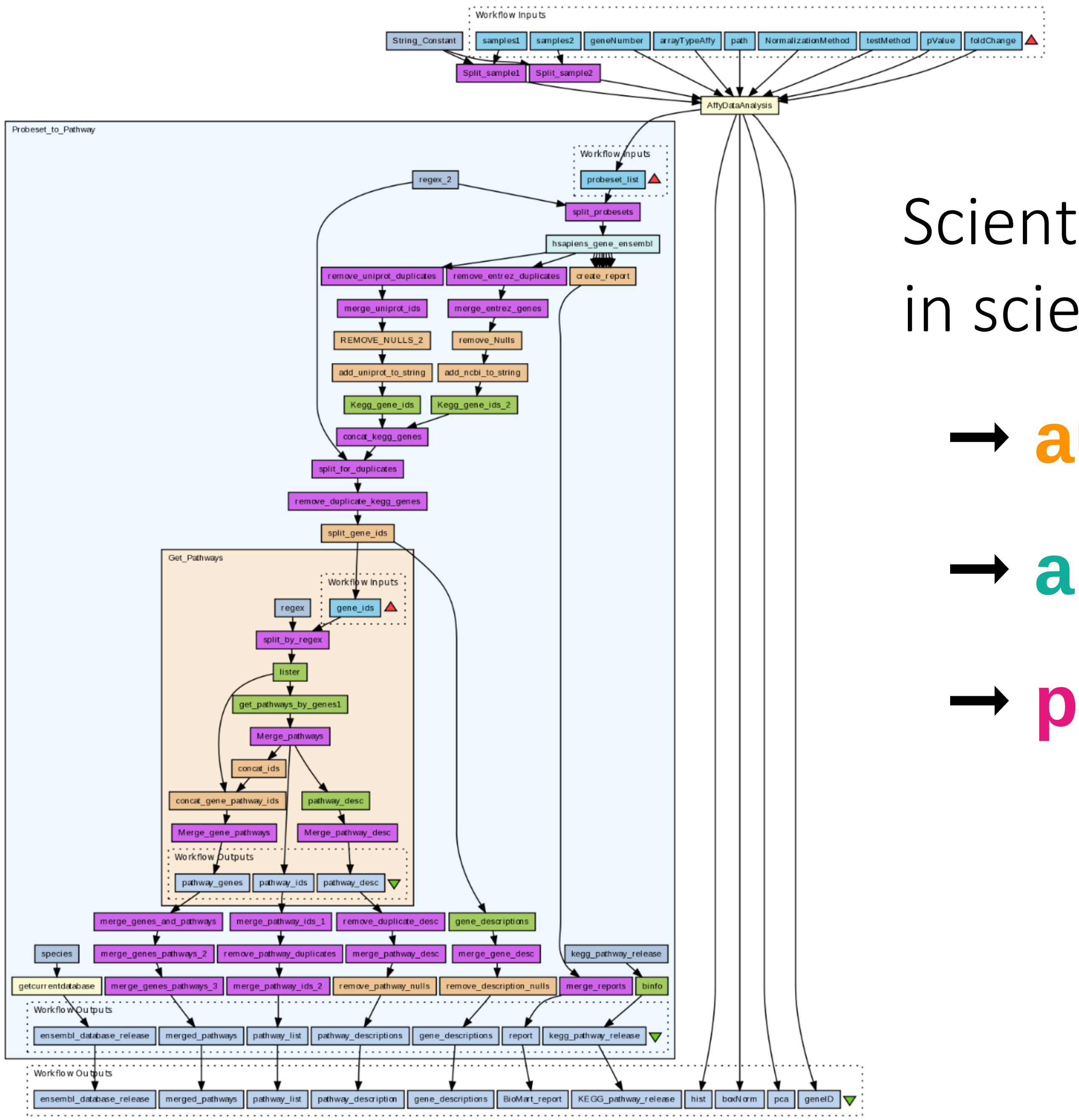
*S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. Le Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal, C. Blanchet, **Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities**, Future Generation Computer Systems, Volume 75, 2017, https://doi.org/10.1016/j.future.2017.01.012 .*

Scientific workflows to the rescue

What is a workflow ?

« a systematic way of **describing the methods** needed and provide the **interface** between **domain specialists** and **computing infrastructures**. »

Malcolm Atkinson, Sandra Gesing, Johan Montagnat, Ian Taylor. ***Scientific workflows: Past, present and future****. Future Generation Computer Systems, Elsevier, 2017, 75, pp.216 - 227. <10.1016/j.future.2017.05.041>*



Scientific workflows to enhance **trust** in scientific results:

- → **automation** of data analysis (at scale)
- → **abstraction** (describe/share methods)
- → **provenance** (~tracability, trust, transparency)

Galaxy COMMUNITY HUB

TensorFlow

Taverna

COMMON WORKFLOW LANGUAGE

Provenance

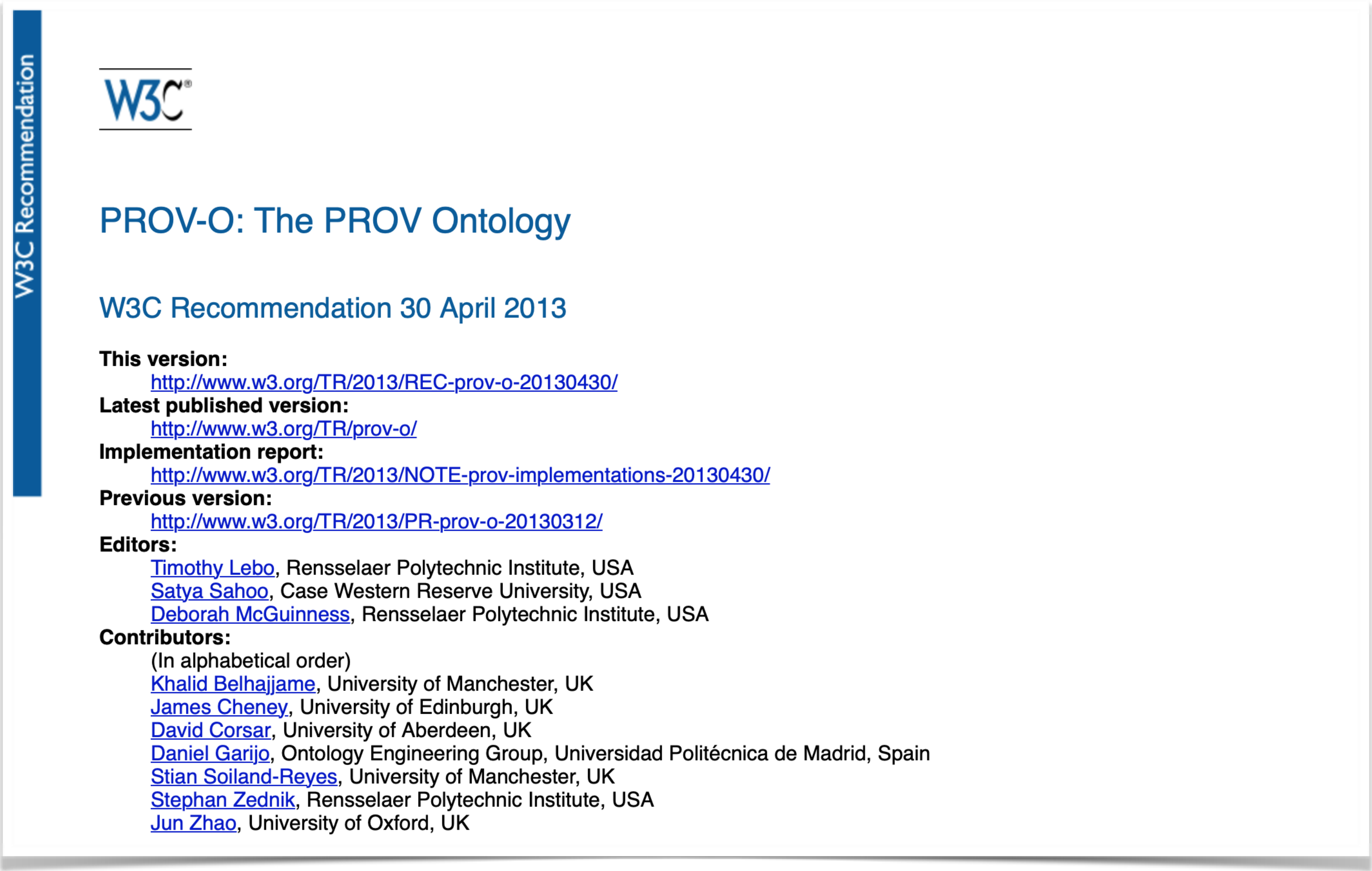
Definition in Computer Science

« Provenance information describes the **origins** and the **history of data in its life cycle**. »

« Today, (…) data is constantly being created, copied, moved around, and combined indiscriminately. Because information sources (…) vary widely in terms of quality, it is essential to provide **provenance and other context information** which can **help end users judge** whether query results are **trustworthy**. »

James Cheney, Laura Chiticariu, and Wang-Chiew Tan. 2009. ***Provenance in Databases: Why, How, and Where****. Found. Trends databases 1, 4 (April 2009), 379-474. DOI=http://dx.doi.org/10.1561/1900000006*

Representing provenance

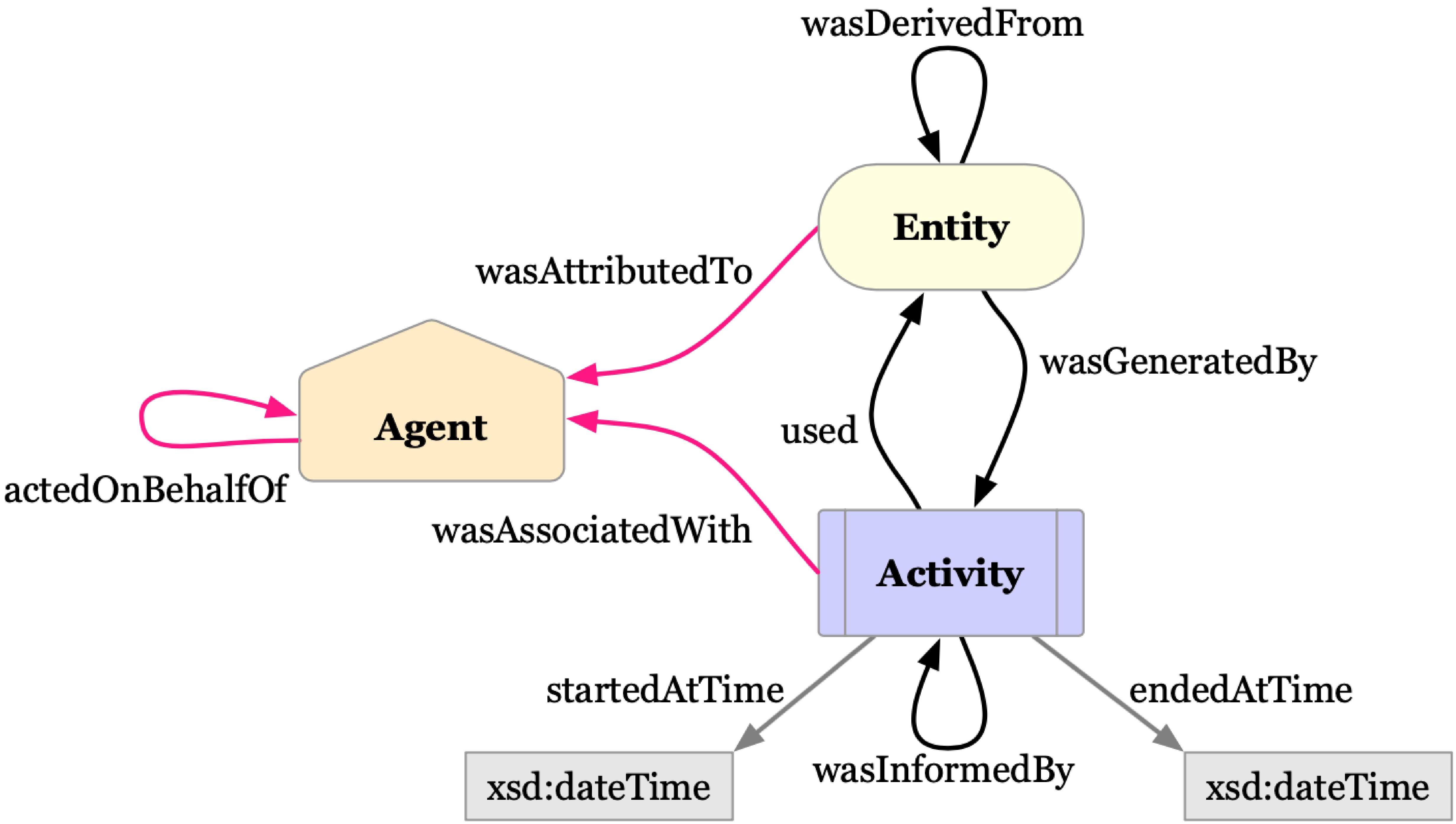


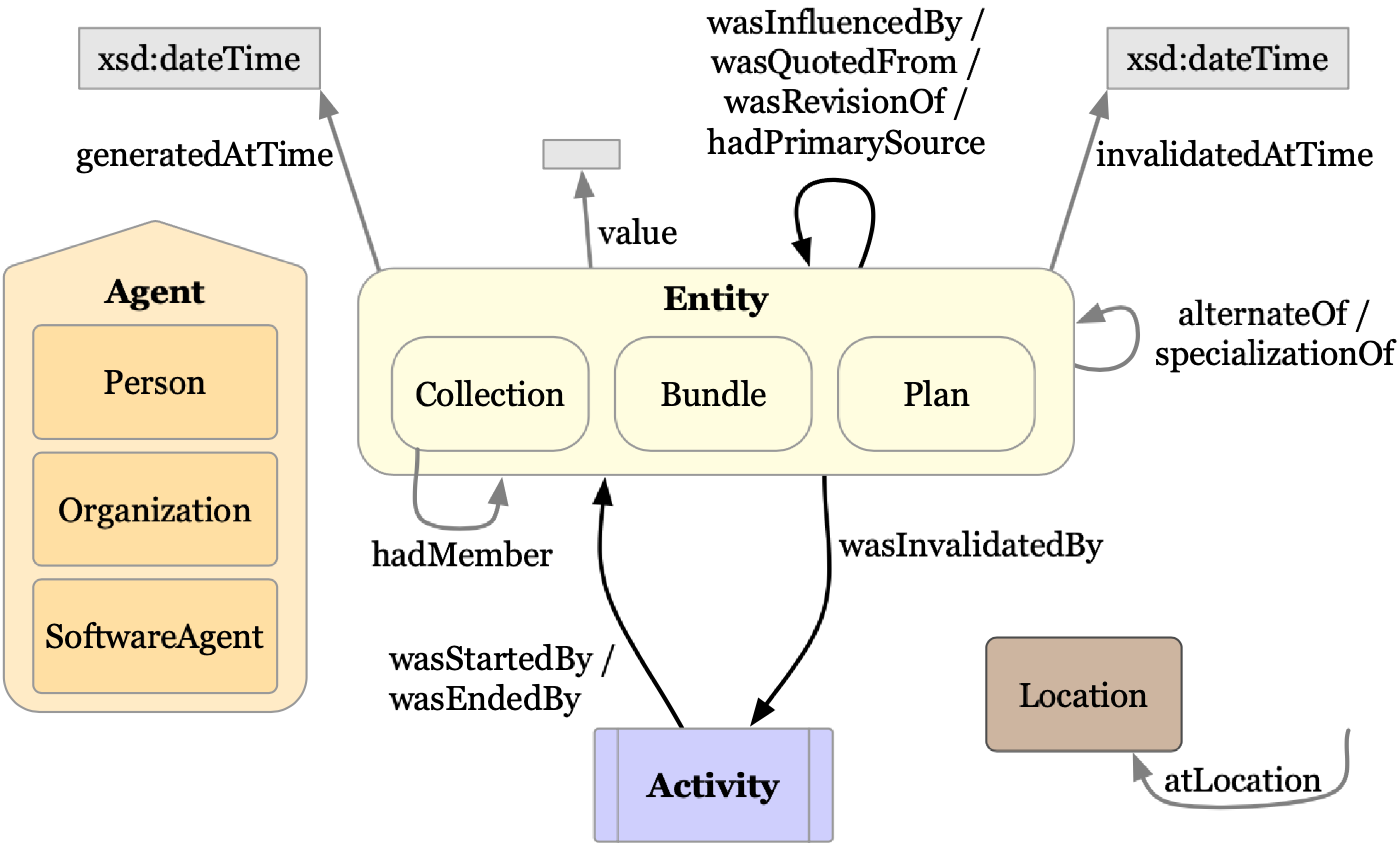
W3C Recommendation

PROV-O: The PROV Ontology

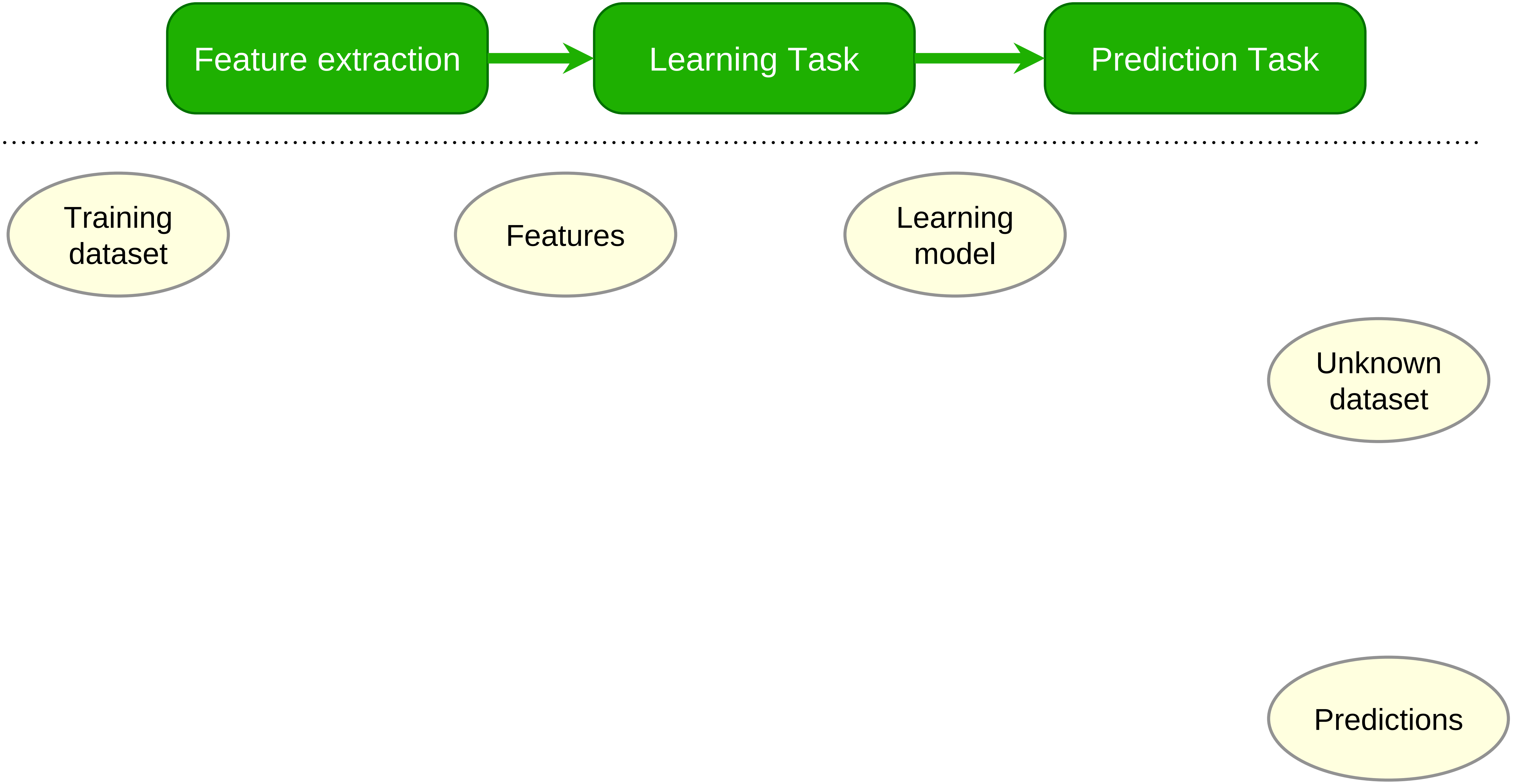
W3C Recommendation 30 April 2013

This version:
http://www.w3.org/TR/2013/REC-prov-o-20130430/
Latest published version:
http://www.w3.org/TR/prov-o/
Implementation report:
http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/
Previous version:
http://www.w3.org/TR/2013/PR-prov-o-20130312/
Editors:
Timothy Lebo, Rensselaer Polytechnic Institute, USA
Satya Sahoo, Case Western Reserve University, USA
Deborah McGuinness, Rensselaer Polytechnic Institute, USA
Contributors:
(In alphabetical order)
Khalid Belhajjame, University of Manchester, UK
James Cheney, University of Edinburgh, UK
David Corsar, University of Aberdeen, UK
Daniel Garijo, Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Stian Soiland-Reyes, University of Manchester, UK
Stephan Zednik, Rensselaer Polytechnic Institute, USA
Jun Zhao, University of Oxford, UK


wasDerivedFrom
Entity
wasAttributedTo
wasGeneratedBy
Agent
used
actedOnBehalfOf
wasAssociatedWith
Activity
startedAtTime
endedAtTime
xsd:dateTime
wasInformedBy
xsd:dateTime


xsd:dateTime
generatedAtTime
value
wasInfluencedBy /
wasQuotedFrom /
wasRevisionOf /
hadPrimarySource
xsd:dateTime
invalidatedAtTime
Agent
Person
Organization
SoftwareAgent
Entity
Collection
Bundle
Plan
alternateOf /
specializationOf
hadMember
wasInvalidatedBy
wasStartedBy /
wasEndedBy
Activity
Location
atLocation

Example: Provenance for a ML workflow


Feature extraction

Learning Task

Prediction Task

Training dataset

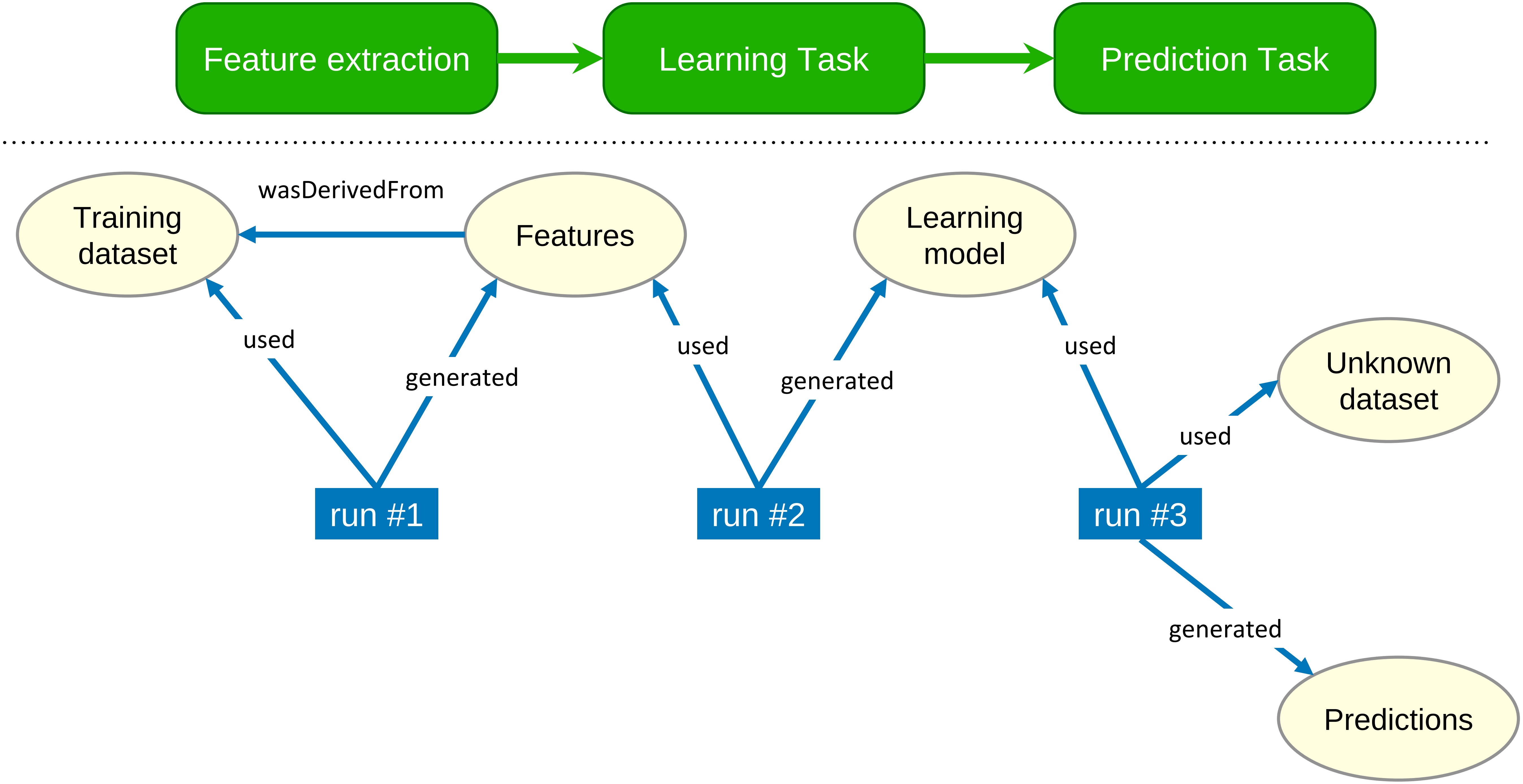
Features

Learning model

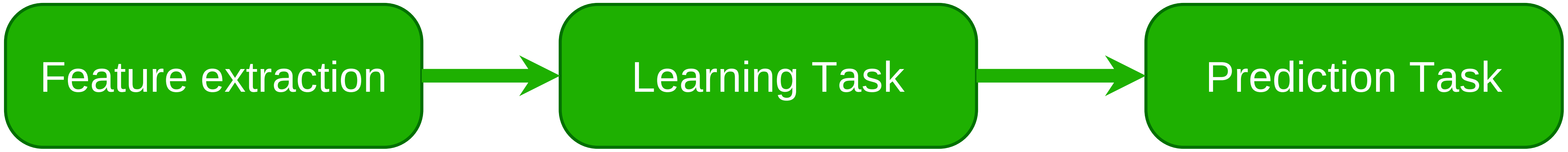
Unknown dataset

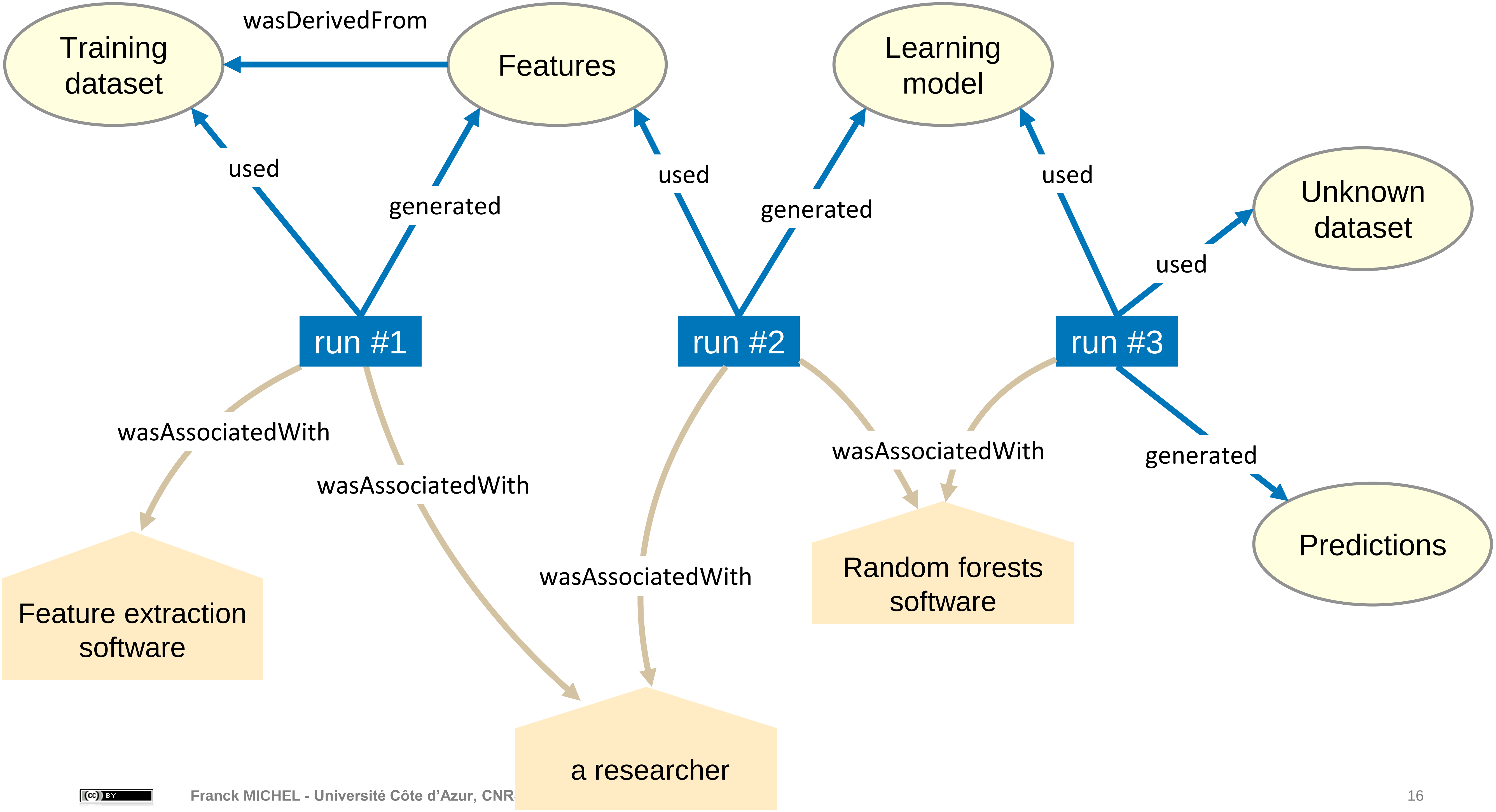
Predictions

Example: Provenance for a ML workflow

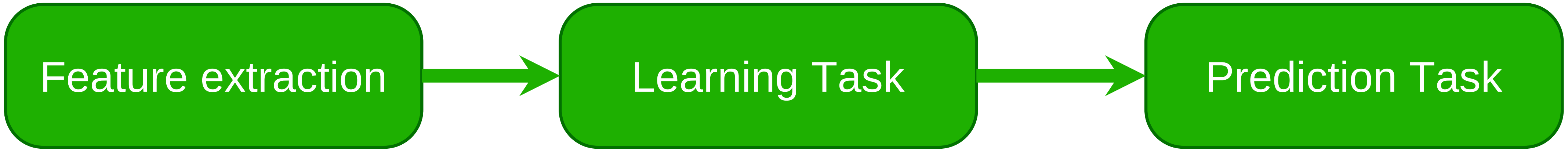


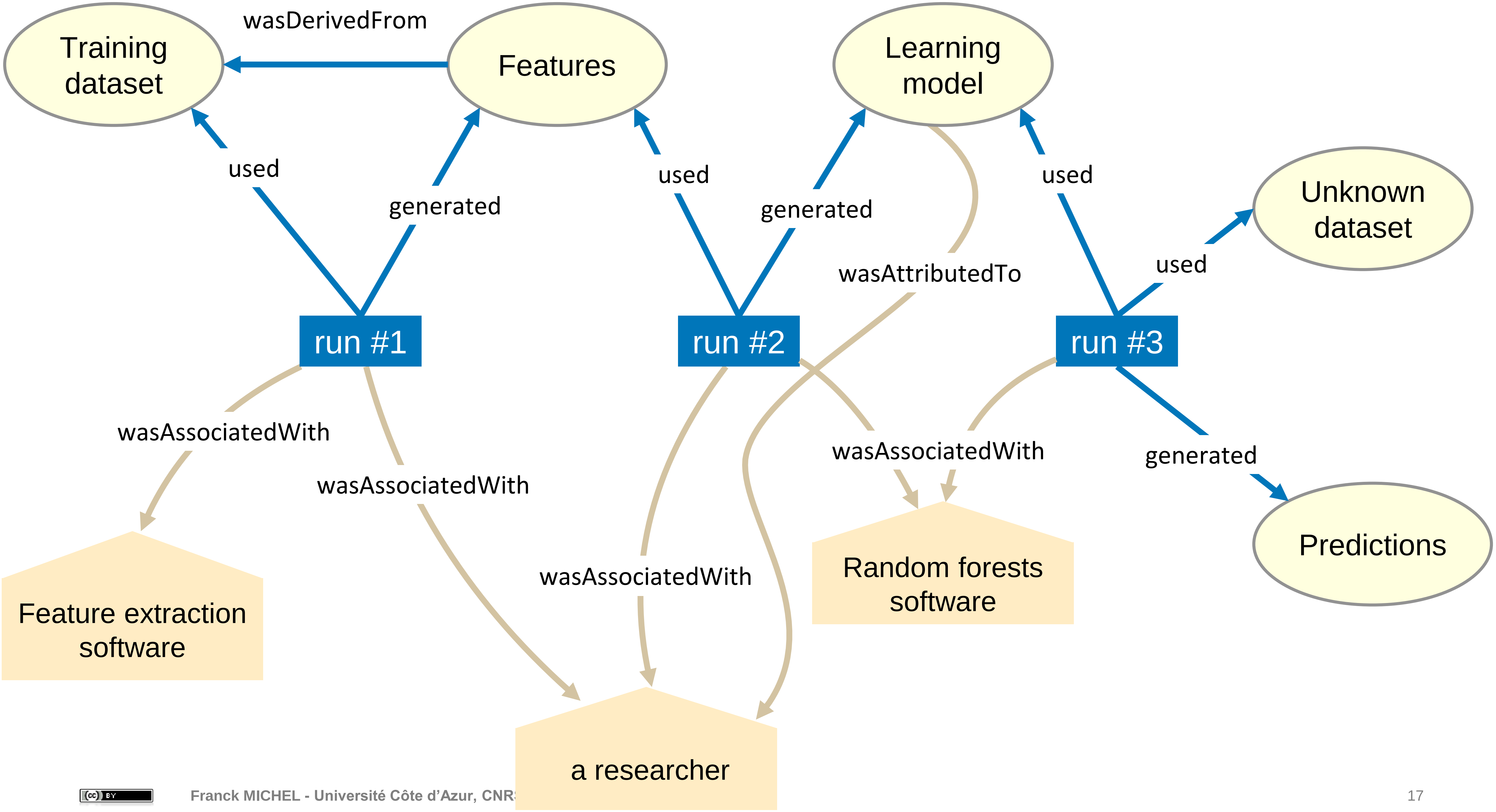
Example: Provenance for a ML workflow



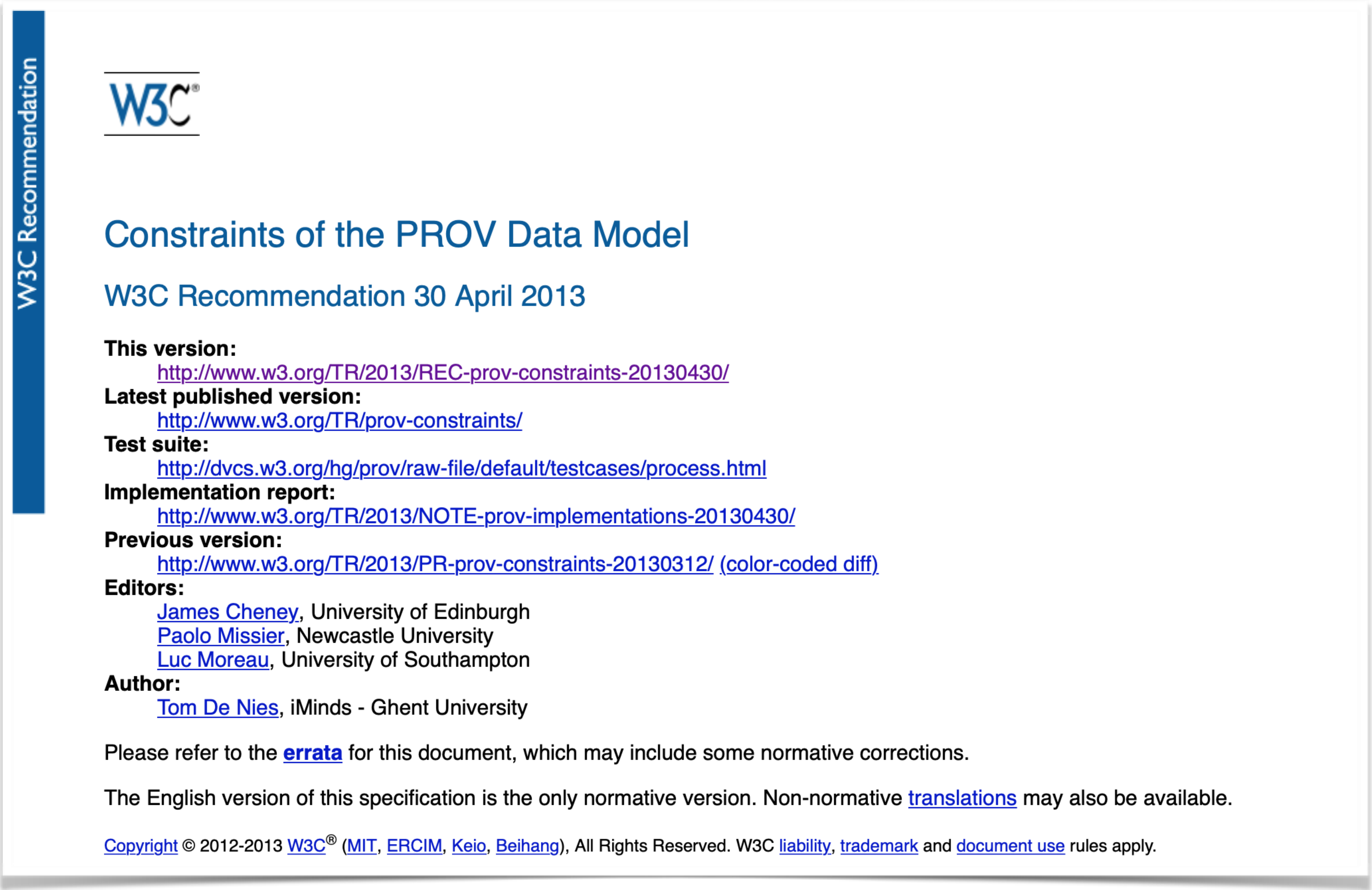


Example: Provenance for a ML workflow





Reasoning with provenance



W3C Recommendation

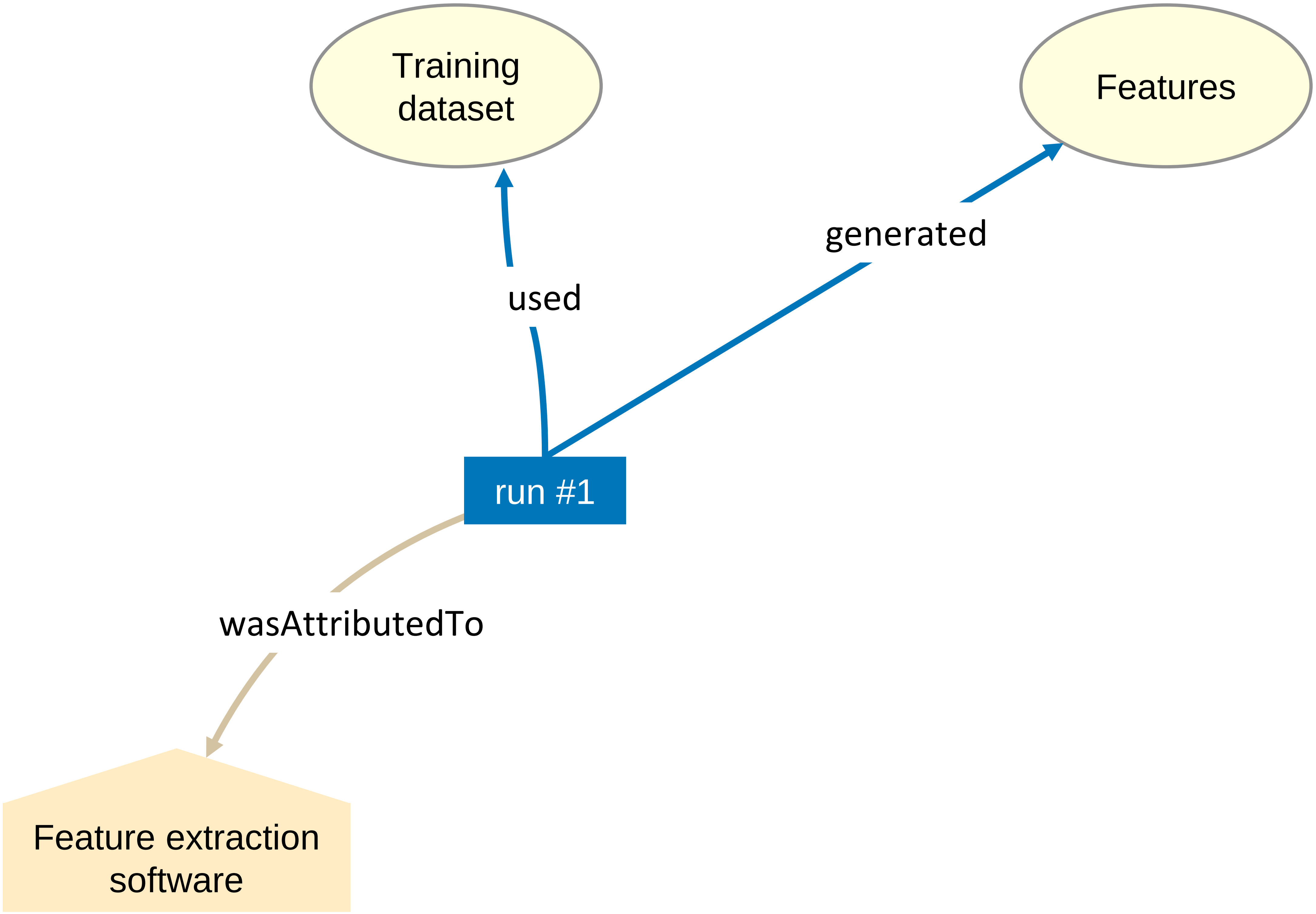
Constraints of the PROV Data Model

W3C Recommendation 30 April 2013

This version:
http://www.w3.org/TR/2013/REC-prov-constraints-20130430/
Latest published version:
http://www.w3.org/TR/prov-constraints/
Test suite:
http://dvcs.w3.org/hg/prov/raw-file/default/testcases/process.html
Implementation report:
http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/
Previous version:
http://www.w3.org/TR/2013/PR-prov-constraints-20130312/ (color-coded diff)
Editors:
James Cheney, University of Edinburgh
Paolo Missier, Newcastle University
Luc Moreau, University of Southampton
Author:
Tom De Nies, iMinds - Ghent University

Please refer to the **errata** for this document, which may include some normative corrections.

The English version of this specification is the only normative version. Non-normative translations may also be available.



Training dataset

Features

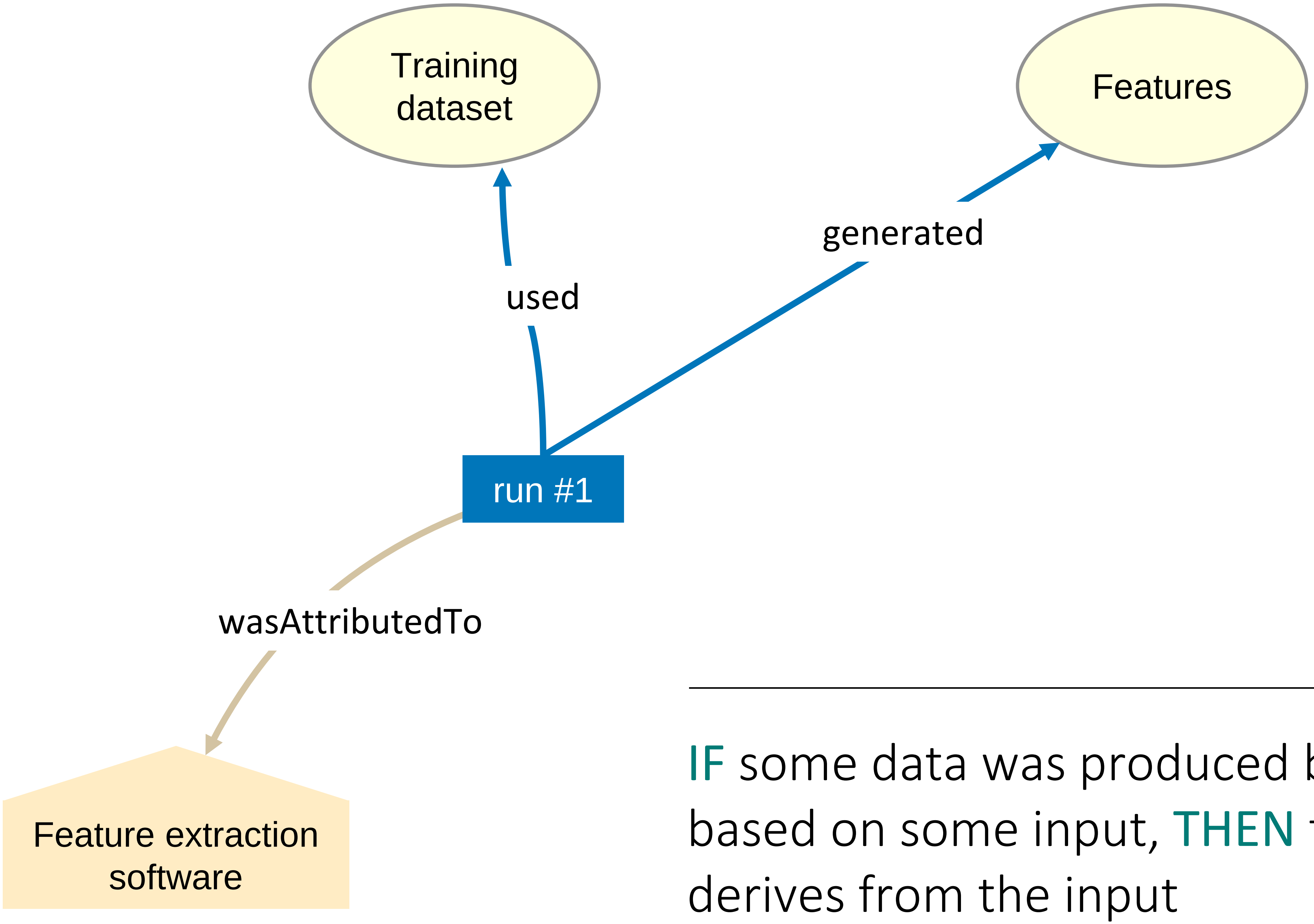
used

generated

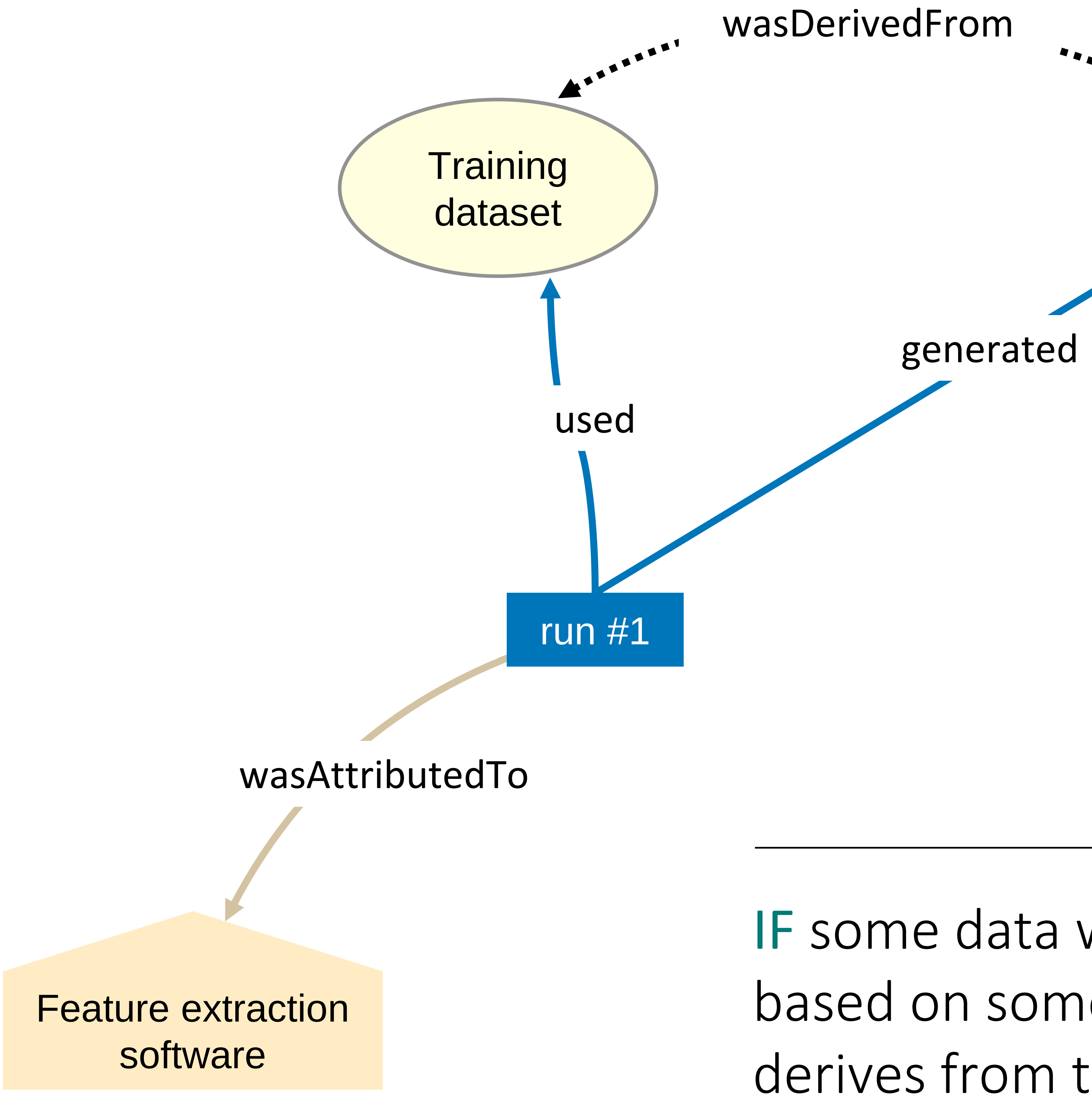
run #1

wasAttributedTo

Feature extraction software



IF some data was produced by a tool based on some input, THEN this data derives from the input



IF some data was produced by a tool based on some input, THEN this data derives from the input

Is provenance enough for reproducibility?

“An activity used an entity”. What to denote inputs’ roles?

- E.g. image registration: 2 inputs image, image to register + atlas

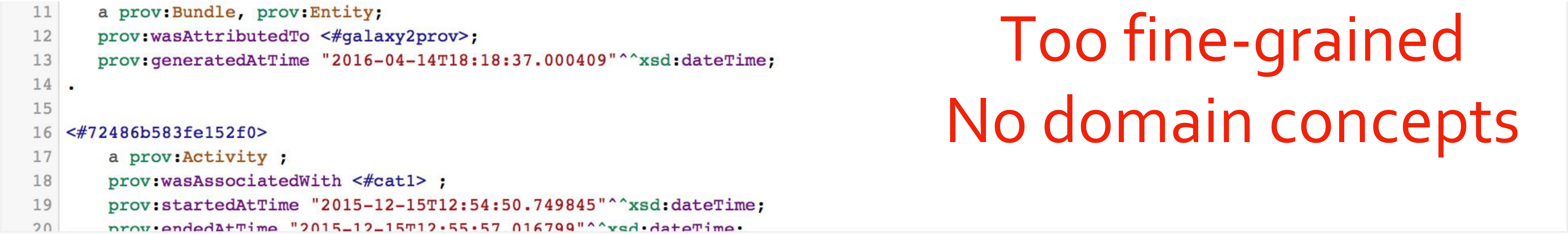
Problem of hidden parameters

- Configuration parameters passed inline in a script
- Hyper-parameters of a DNN

Hence the need for **PROV-O extensions**

- ProvONE: PROV Extension for Scientific Workflow Provenance
- Sensor Data Provenance: SSNO and PROV-O
- PAV: Provenance, Authoring and Versioning ontology
- …

Is provenance enough for reuse?

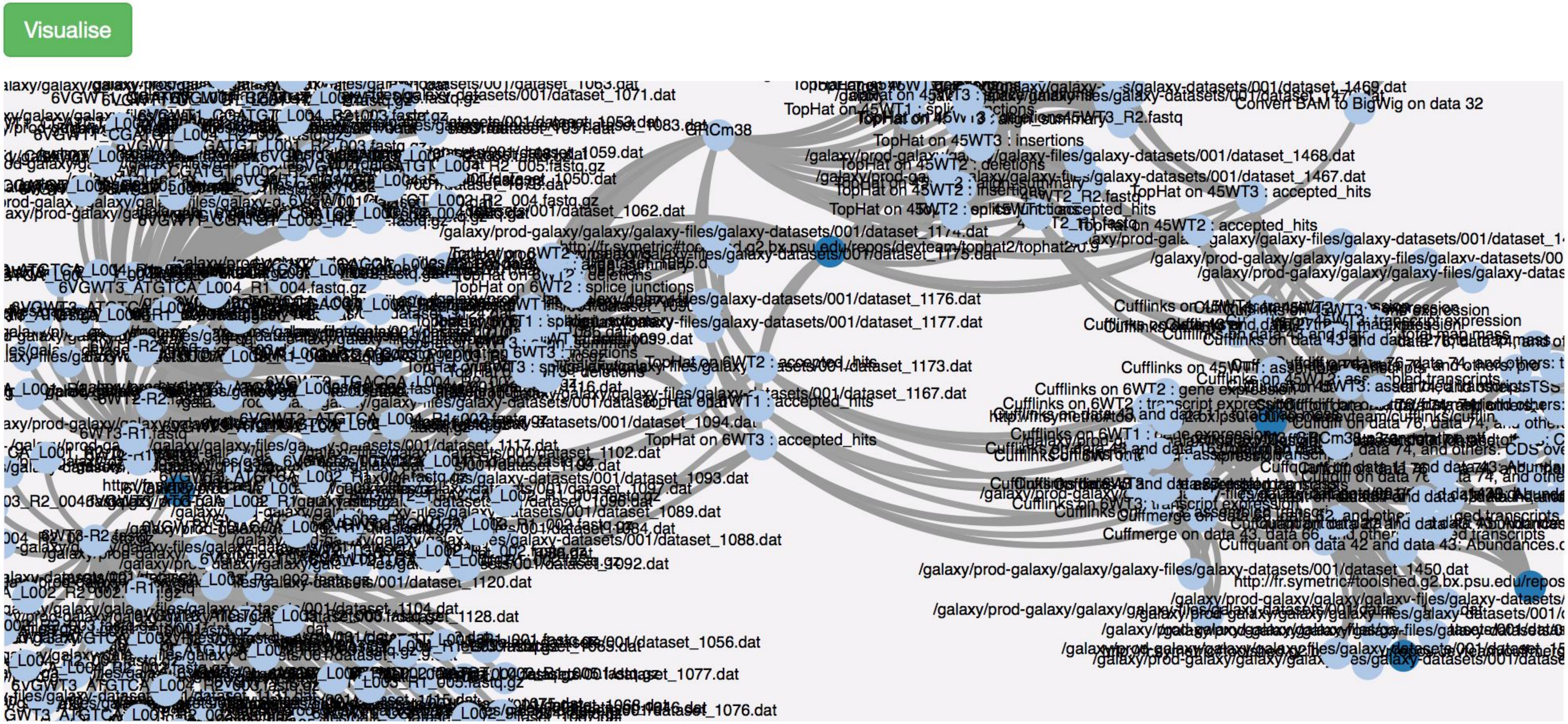


```
    a prov:Bundle, prov:Entity;
    prov:wasAttributedTo <#galaxy2prov>;
    prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
.

<#72486b583fe152f0>
     a prov:Activity ;
     prov:wasAssociatedWith <#cat1> ;
     prov:startedAtTime "2015-12-15T12:54:50.749845"^^xsd:dateTime;
     prov:endedAtTime "2015-12-15T12:55:57.016799"^^xsd:dateTime;
```

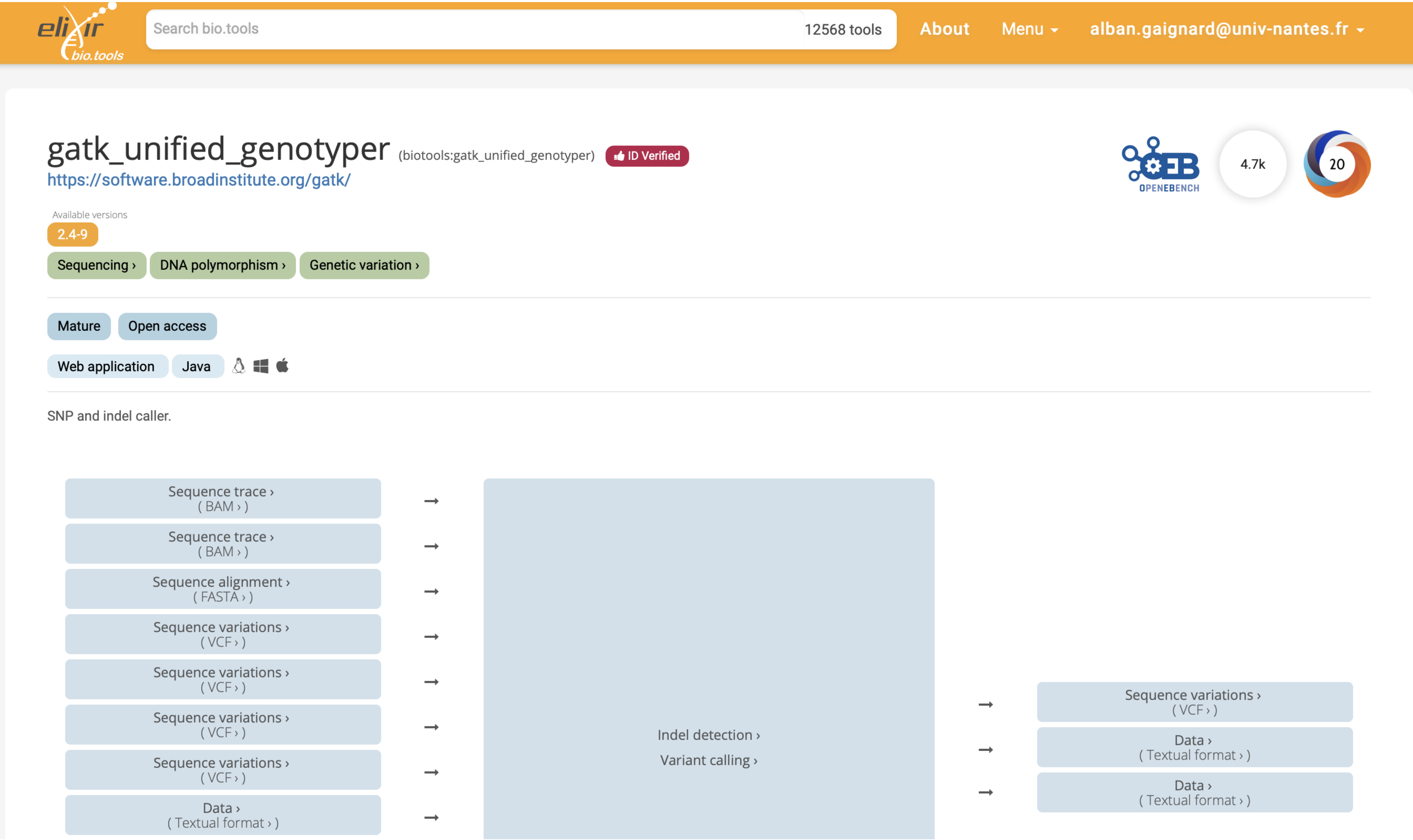
Too fine-grained

No domain concepts



Visualise
6VGWT3_ATGTCA_L004_R1_004.fastq.gz
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1176.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1177.dat
TopHat on 6WT2 : splice junctions
TopHat on 6WT2 : deletions
TopHat on 6WT2 : accepted_hits
TopHat on 6WT3 : accepted_hits
TopHat on 6WT1 : accepted_hits
TopHat on 45WT3 : insertions
TopHat on 45WT3 : accepted_hits
TopHat on 45WT2 : accepted_hits
TopHat on 45WT2 : deletions
TopHat on 45WT2 : insertions
Convert BAM to BigWig on data 32
GRCm38
Cufflinks on 6WT2 : gene expression
Cufflinks on 6WT3 : transcript expression
Cuffmerge on data 43, data 66, and others
Cuffquant on data 42 and data 43: Abundances
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1450.dat
http://fr.symetric#toolshed.g2.bx.psu.edu/repos
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1468.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1467.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1469.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1173.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1167.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1094.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1093.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1089.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1088.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1120.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1056.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1077.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1076.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1062.dat
/galaxy/prod-galaxy/galaxy/galaxy-files/galaxy-datasets/001/dataset_1071.dat

Semantic tools catalogs



elixir bio.tools

Search bio.tools 12568 tools

About Menu alban.gaignard@univ-nantes.fr

gatk_unified_genotyper (biotools:gatk_unified_genotyper) ID Verified

https://software.broadinstitute.org/gatk/

OPENEBENCH 4.7k 20

Available versions

2.4-9

Sequencing › DNA polymorphism › Genetic variation ›

Mature Open access

Web application Java

SNP and indel caller.

Sequence trace › (BAM ›) →

Sequence trace › (BAM ›) →

Sequence alignment › (FASTA ›) →

Sequence variations › (VCF ›) →

Sequence variations › (VCF ›) →

Sequence variations › (VCF ›) →

Sequence variations › (VCF ›) →

Data › (Textual format ›) →

Indel detection ›

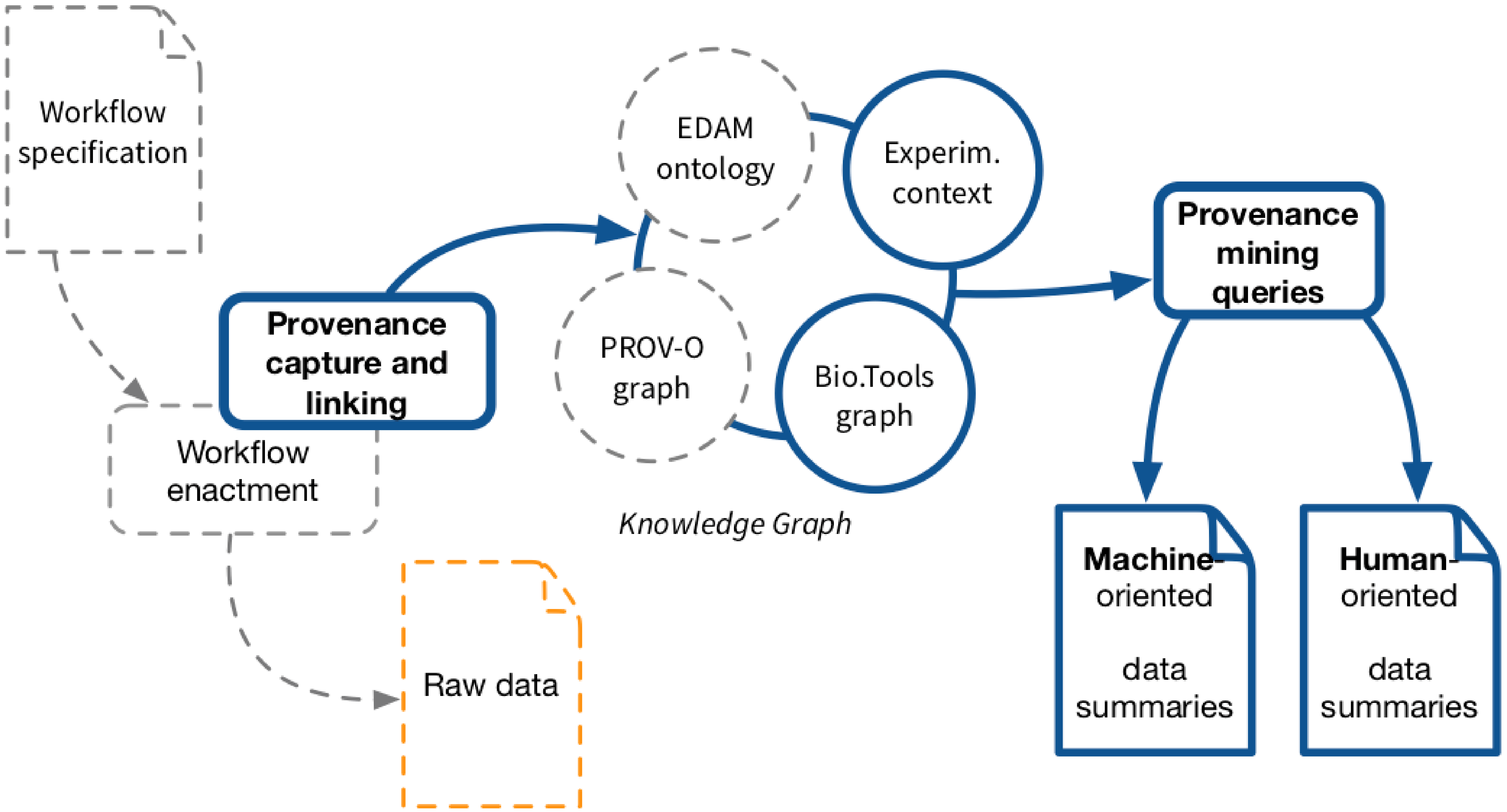
Variant calling ›

→ Sequence variations › (VCF ›)

→ Data › (Textual format ›)

→ Data › (Textual format ›)

Domain-centric provenance summary



What about ML tools?

ML-Schema

Community Group

Goal: *Improve **interoperability**, **reproducibility** and **interpretability** of DM/ML experiments*

How:

- Define a schema to represent/share information on DM/ML **algorithms**, **datasets** and **experiments**
- Align existing DM/ML ontologies to this schema, develop ontologies for specific purposes/applications
- Turn DM/ML algorithms and results into **Linked Open Data**

ML-Schema

Community Group

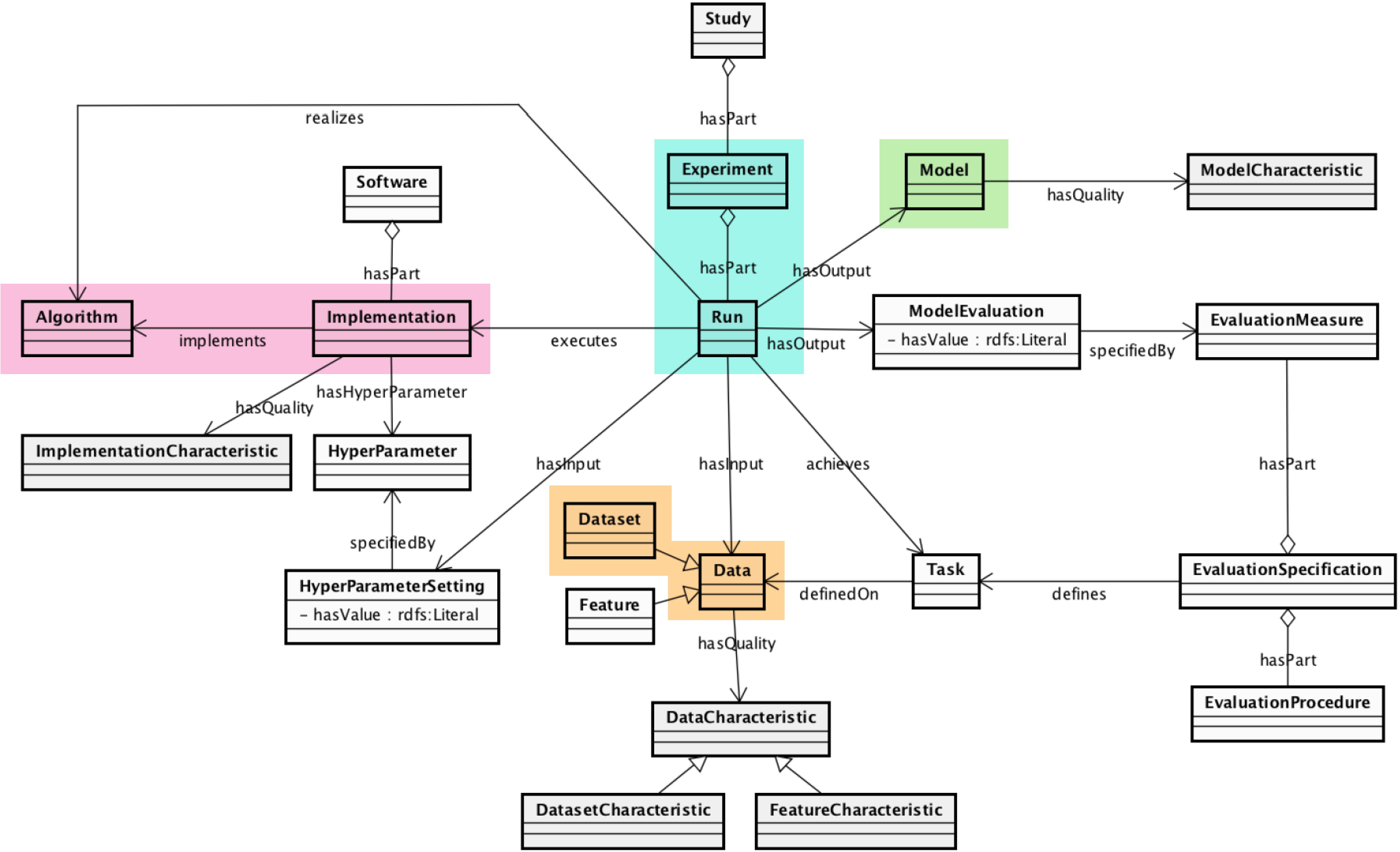
Inspired by previous works on ML/DM vocabularies

- **OntoDM**-core ontology: DM, based on BFO
- **Exposé** ontology: ML experiments, on top of OntoDM. Used in OpenML
- **D**ata **M**ining **OP**timization ontology: taxonomy of DM algos and ML models
- **MEX**: lightweight vocabulary to exchange basic ML metadata

Publio G. C., Esteves D., Ławrynowicz A., Panov P., Soldatova L., Soru T., Vanschoren J. & Zafar H. (2018). **ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies**. In *Proc. of the 2nd Reproducibility in Machine Learning*, p. 5. Stockholm, Sweeden.

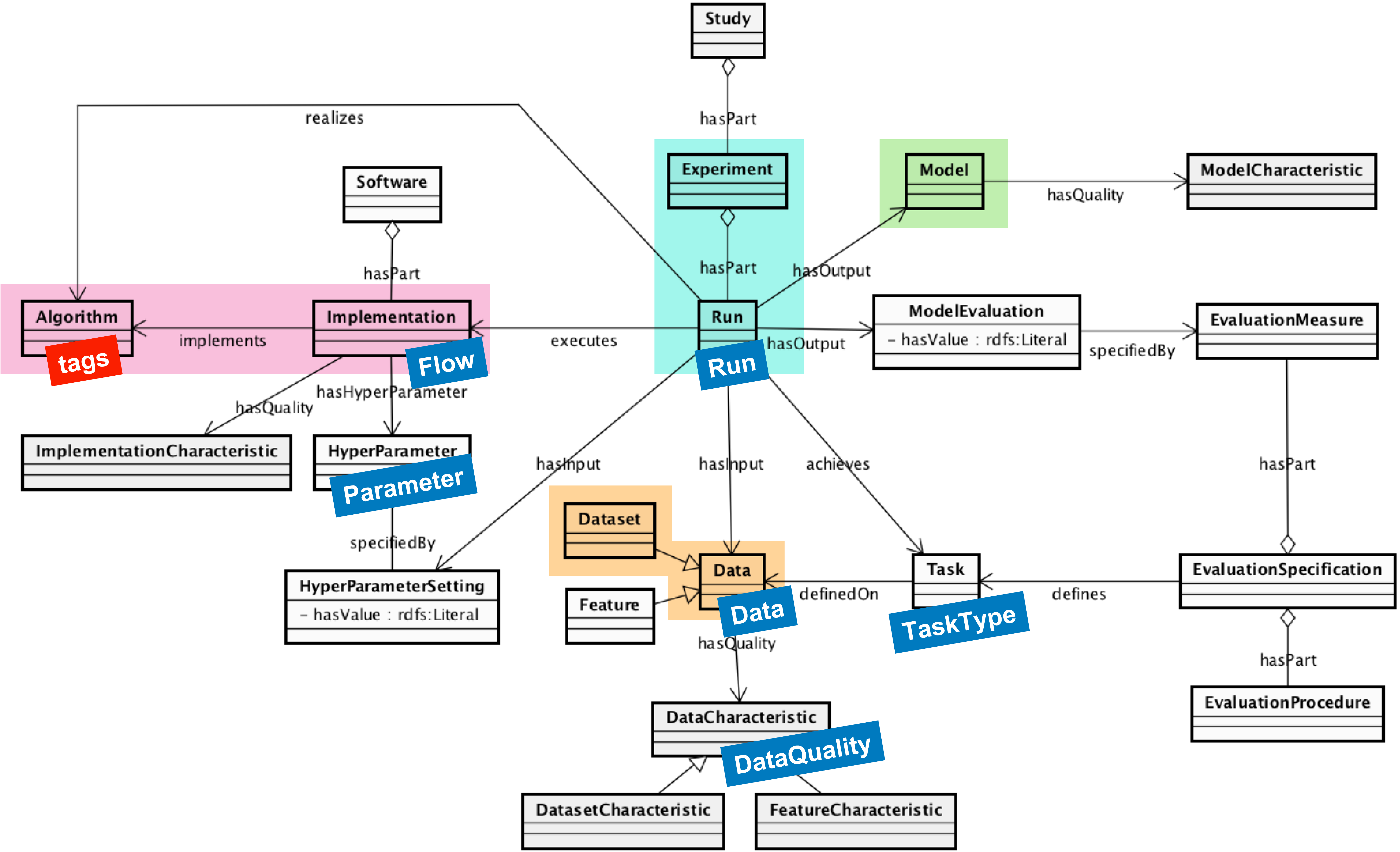
ML-Schema

W3C Community Group



Publio G. C., Esteves D., Ławrynowicz A., Panov P., Soldatova L., Soru T., Vanschoren J. & Zafar H. (2018). **ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies**. In *Proc. of the 2nd Reproducibility in Machine Learning*, p. 5. Stockholm, Sweeden.

Imperfect mapping of ML-Schema to OpenML



Source: https://github.com/ML-Schema/core/wiki/Vocabulary-(mappings)

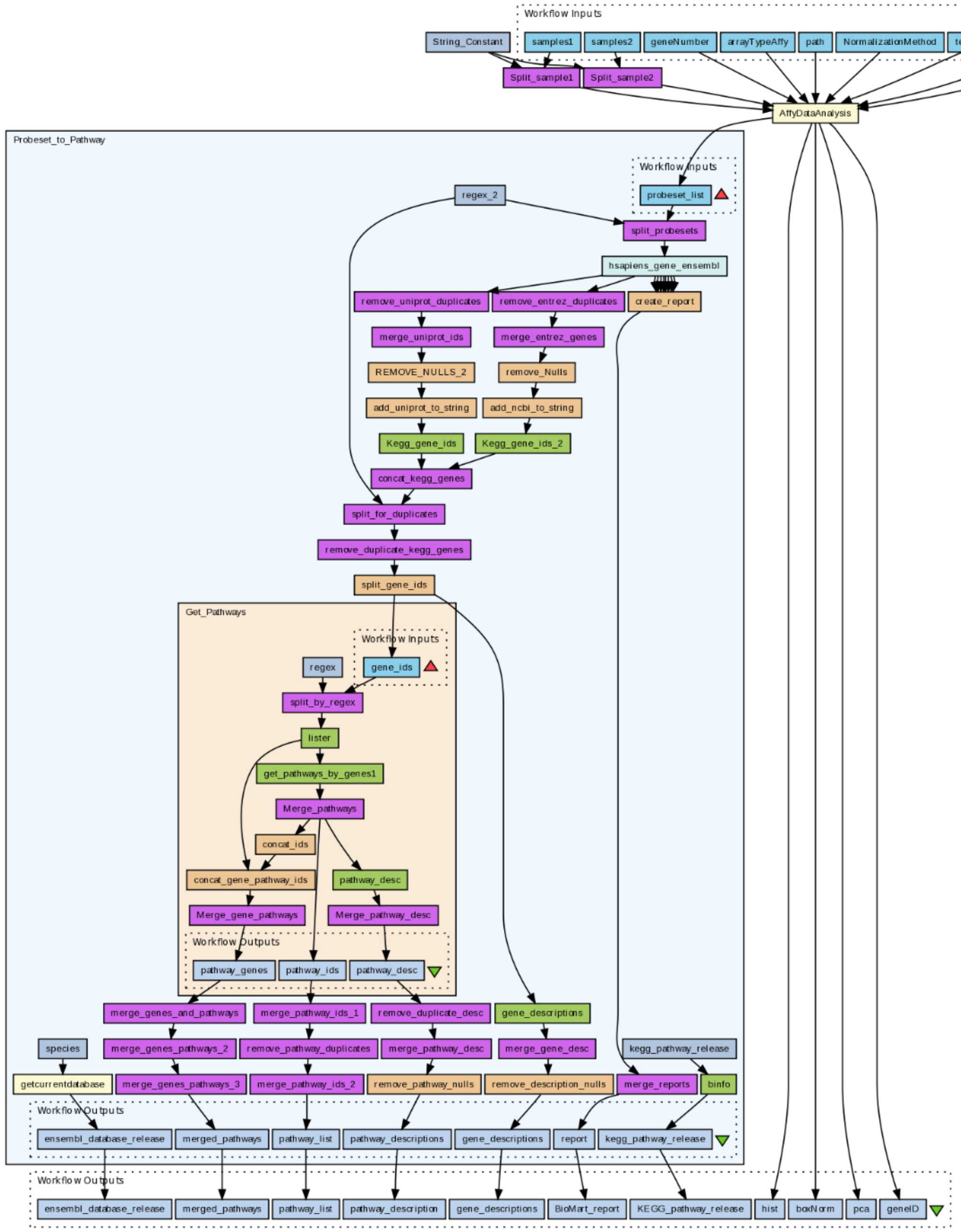
ML-Schema

Community Group

Current status (personal feeling)

- Still a preliminary work: a simple Turtle file
- Need for more complete recommendations (e.g. wrt. use of 3rd party vocabularies: DC Terms, DCAT etc.)
- Need for richer mapping descriptions
- Need for real interoperability tests

But a key to instrument ML algos/datasets/experiments together with PROV-O-instrumented workflow engines



ML workflows:

- → **automation** of data analysis
- → **abstraction**: ML-Schema + extensions
- → **provenance**: PROV-O + extensions

Take-aways & open questions

Scientific Workflows ➞ automation, abstraction, provenance

Standards for **provenance representation** and **reasoning**

Reproducibility/reuse requires **domain-specific tools' description** and **provenance-enabled workflow engines**

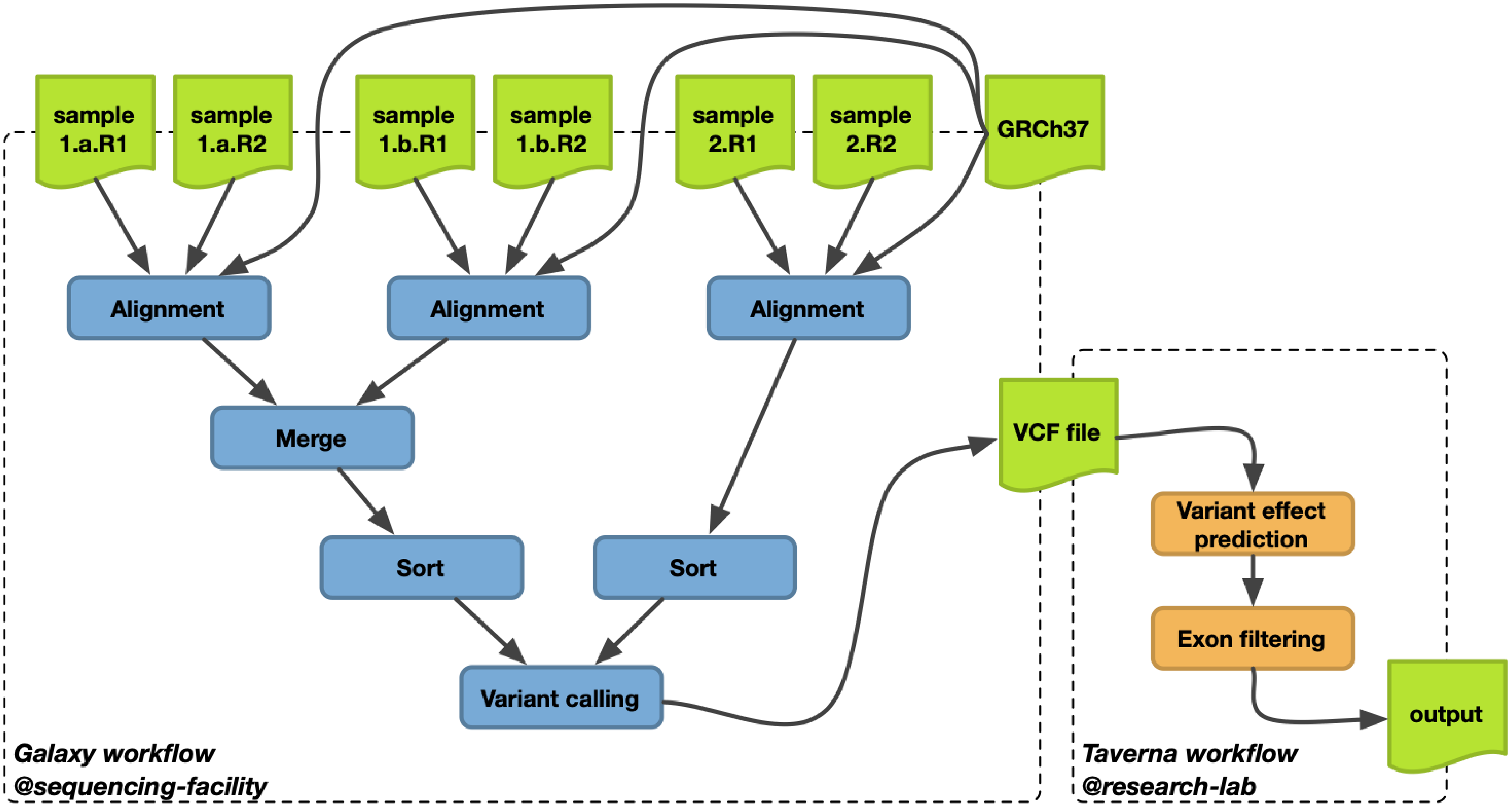
ML-Schema + PROV-O ➞ the future winning couple?

Distributed data analysis ➞ **Distributed provenance, reasoning**?

Backup slides

Provenance in **multi-site** studies ?

Multi-site studies → ≠ workflow engines !

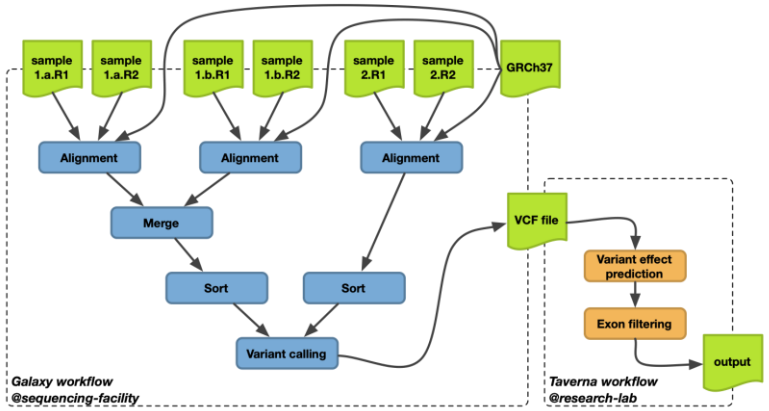


Galaxy

Taverna

Scattered provenance capture ?

Provenance issues

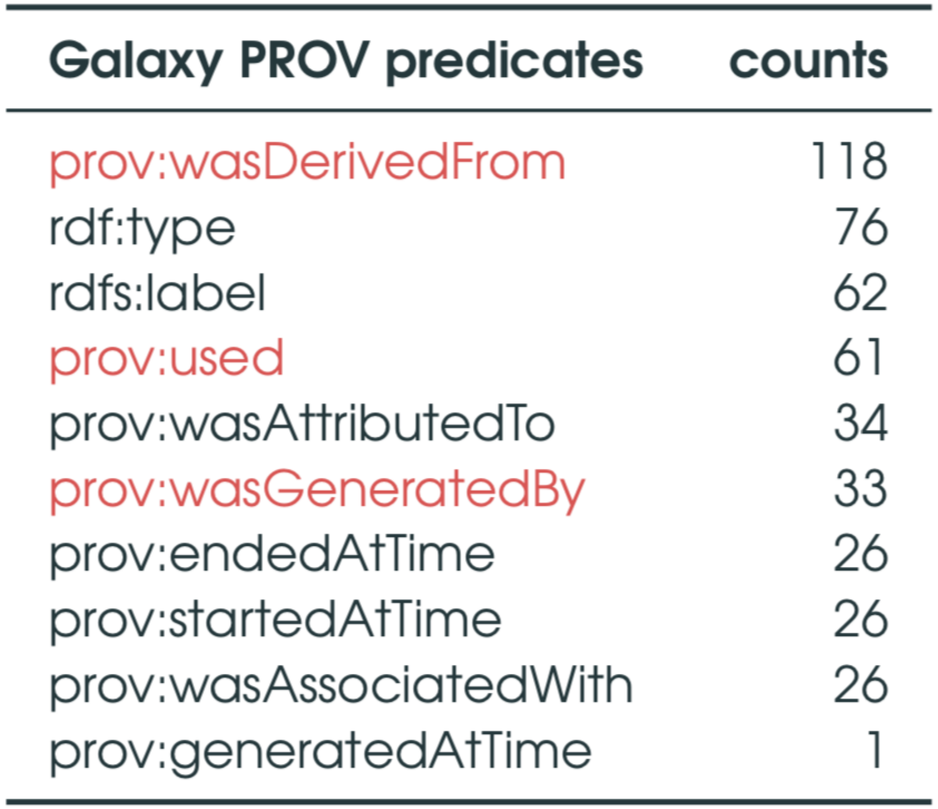


« Which alignment algorithm was used
when predicting these pathogenic store? »

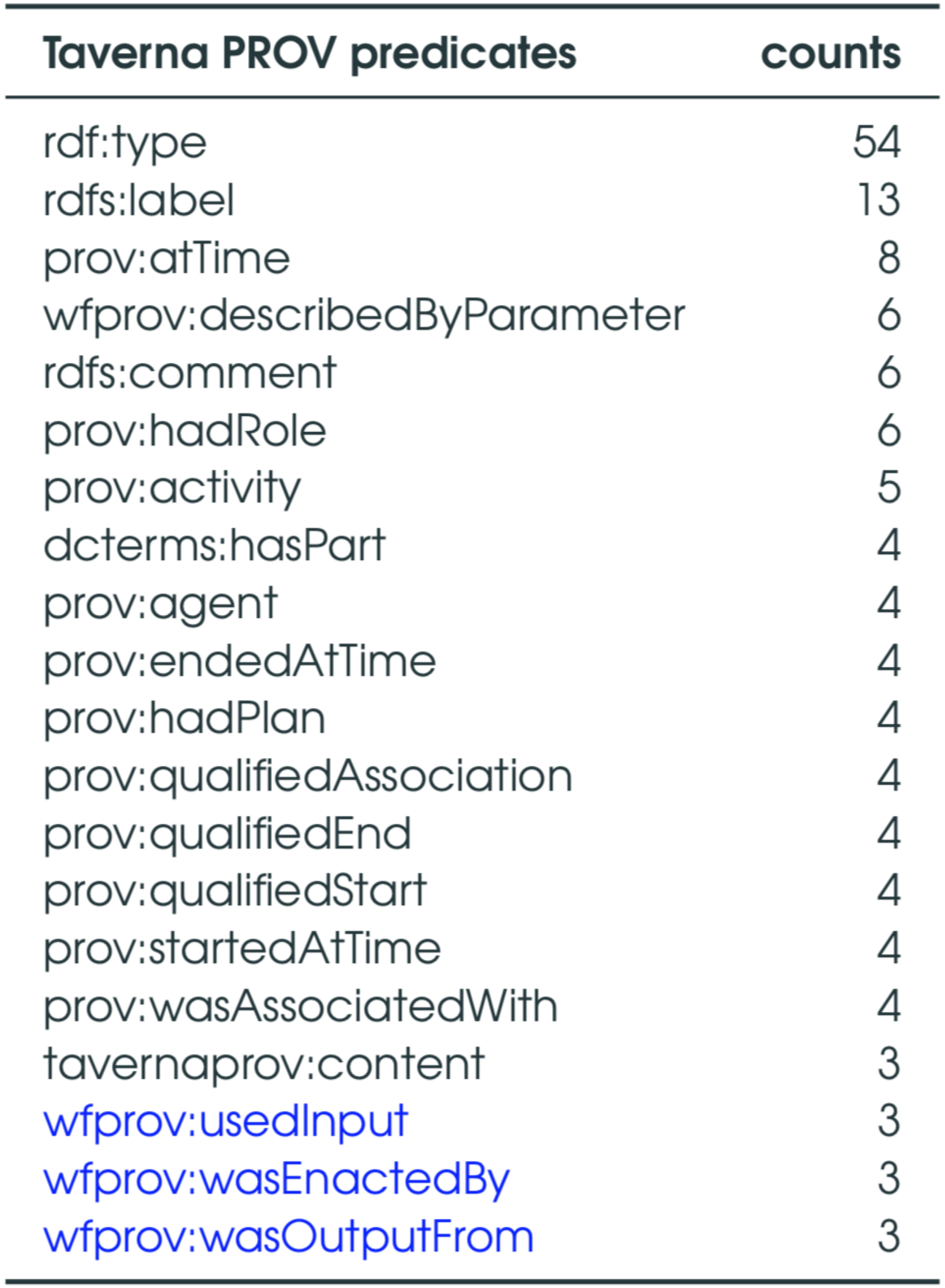
« A new version of a reference genome is available, which
genome was used when predicting these phenotypes ? »

Need for an overall tracking of provenance
over both Galaxy and Taverna workflows !

Provenance « heterogeneity »



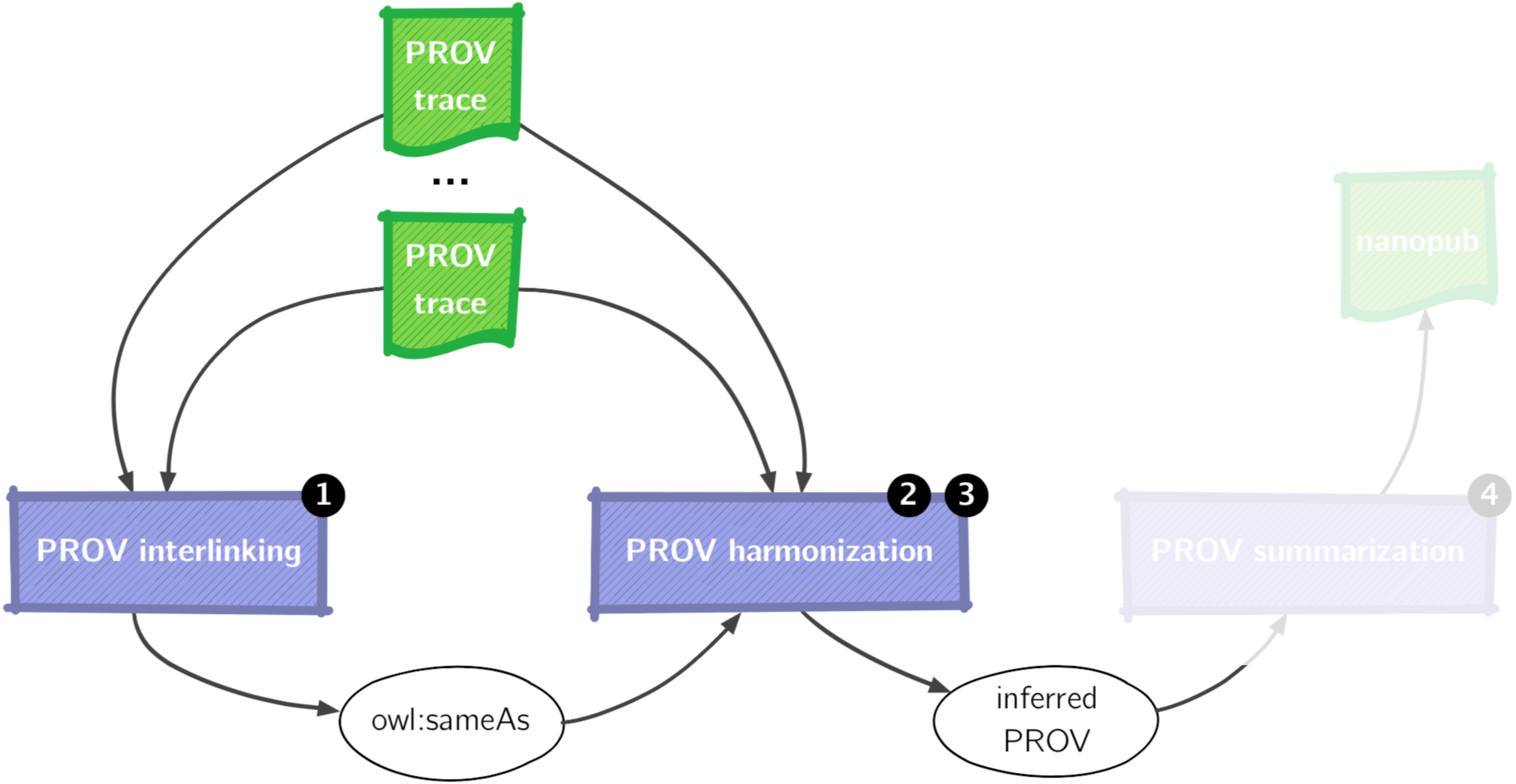
Galaxy PROV predicates	counts
prov:wasDerivedFrom	118
rdf:type	76
rdfs:label	62
prov:used	61
prov:wasAttributedTo	34
prov:wasGeneratedBy	33
prov:endedAtTime	26
prov:startedAtTime	26
prov:wasAssociatedWith	26
prov:generatedAtTime	1



Taverna PROV predicates	counts
rdf:type	54
rdfs:label	13
prov:atTime	8
wfprov:describedByParameter	6
rdfs:comment	6
prov:hadRole	6
prov:activity	5
dcterms:hasPart	4
prov:agent	4
prov:endedAtTime	4
prov:hadPlan	4
prov:qualifiedAssociation	4
prov:qualifiedEnd	4
prov:qualifiedStart	4
prov:startedAtTime	4
prov:wasAssociatedWith	4
tavernaprov:content	3
wfprov:usedInput	3
wfprov:wasEnactedBy	3
wfprov:wasOutputFrom	3

How to reconcile these provenance traces?

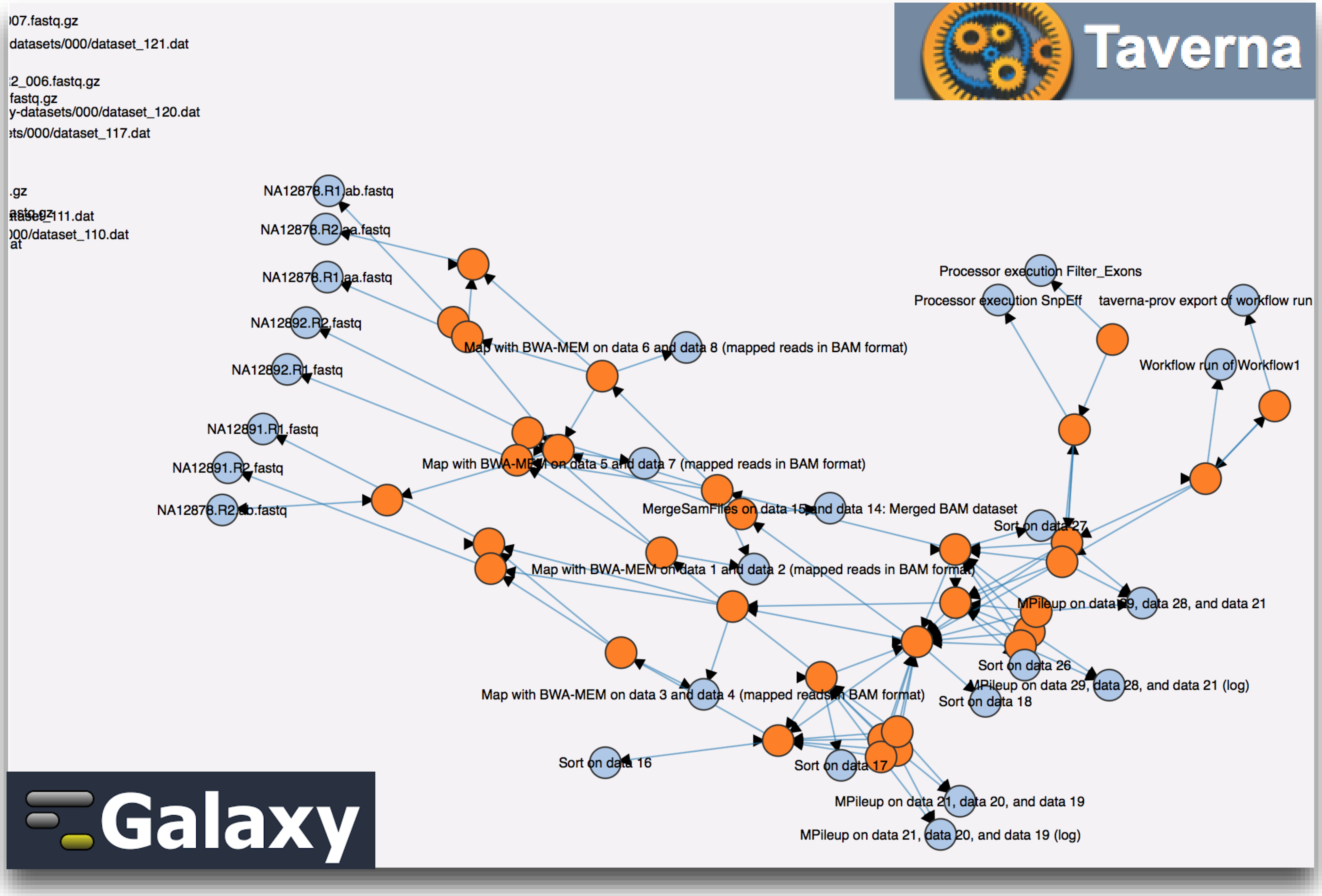
Approach



A. Gaignard, K. Belhajjame, H. Skaf-Molli. ***SHARP: Harmonizing and Bridging Cross-Workflow Provenance****. The Semantic Web: ESWC 2017 Satellite Events Portorož, Slovenia, May 28 – June 1, 2017, Revised Selected Papers, 2017*

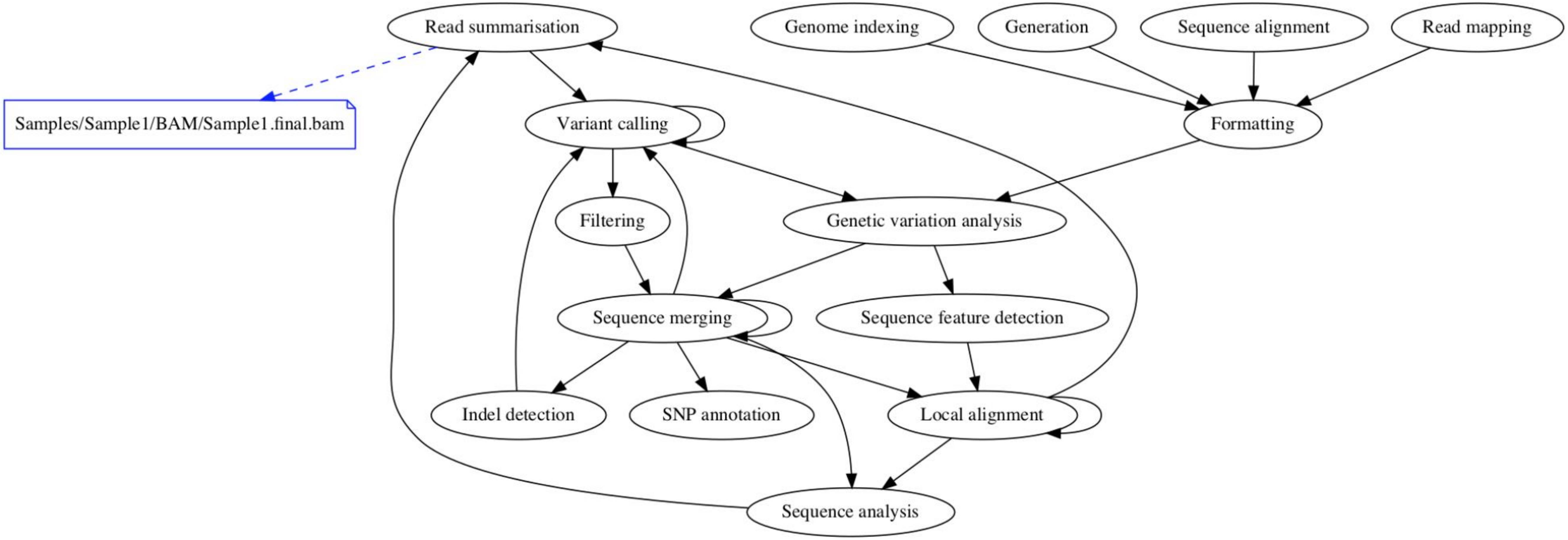
Results

Reconciled provenance as an « influence graph »


Taverna
Galaxy
)07.fastq.gz
datasets/000/dataset_121.dat
2_006.fastq.gz
fastq.gz
y-datasets/000/dataset_120.dat
ets/000/dataset_117.dat
.gz
ataset_111.dat
)00/dataset_110.dat
NA12878.R1.ab.fastq
NA12878.R2.aa.fastq
NA12878.R1.aa.fastq
NA12892.R2.fastq
NA12892.R1.fastq
NA12891.R1.fastq
NA12891.R2.fastq
NA12878.R2.ab.fastq
Map with BWA-MEM on data 6 and data 8 (mapped reads in BAM format)
Map with BWA-MEM on data 5 and data 7 (mapped reads in BAM format)
MergeSamFiles on data 15 and data 14: Merged BAM dataset
Map with BWA-MEM on data 1 and data 2 (mapped reads in BAM format)
Map with BWA-MEM on data 3 and data 4 (mapped reads in BAM format)
Sort on data 16
Sort on data 17
Sort on data 18
Sort on data 26
Sort on data 27
MPileup on data 29, data 28, and data 21
MPileup on data 29, data 28, and data 21 (log)
MPileup on data 21, data 20, and data 19
MPileup on data 21, data 20, and data 19 (log)
Processor execution Filter_Exons
Processor execution SnpEff
taverna-prov export of workflow run
Workflow run of Workflow1

https://github.com/albangaignard/sharp-prov-toolbox

Provenance summary



```
...
The file Samples/Sample1/BAM/Sample1.realign.bai results from
tool gatk2_indel_realigner-IP which Locally align two or more molecular
sequences.

It was produced in the context of Rare Coding Variants in ANGPTL6 Are
Associated with Familial Forms of Intracranial Aneurysm
...
```

https://github.com/albangaignard/fresh-toolbox