Re-use in data-driven sciences: from provenance to (linked) data summaries

Alban Gaignard, PhD, CNRS

Paris, 12 septembre 2019 - France

RDA RESEARCH DATA ALLIANCE

Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES l'institut du thorax

Knowledge production


COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

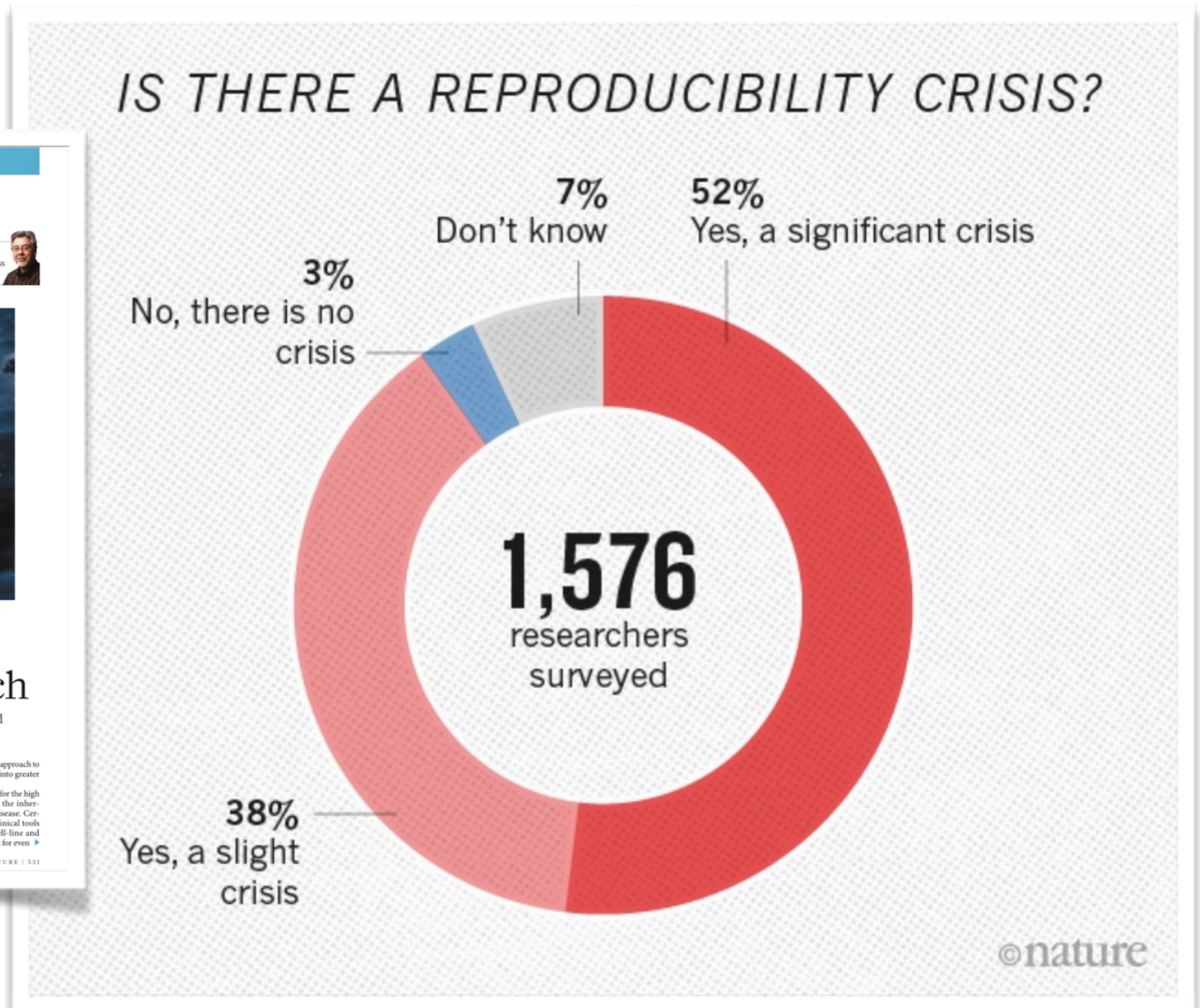
Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this

trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical

investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models' make it difficult for even ▸

29 MARCH 2012 | VOL 483 | NATURE | 531



COMMENT

612 | NATURE | VOL 505 | 30 JANUARY 2014

NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring[1,2]. As leaders of the US National Institutes of Health (NIH), we share this concern and here explore some of the significant interventions that we are planning.

Science has long been regarded as 'self-correcting', given that it is founded on the

shorter term, however, the checks and balances that once ensured scientific fidelity have been hobbled. This has compromised the ability of today's researchers to reproduce others' findings.

Let's be clear: with rare exceptions, we have no evidence to suggest that irreproducibility is about scientific misconduct. In 2011, the Office of Research Integrity of the US Department of Health and Human Services pursued only 12 such cases[3].

« In 2012, Amgen researchers made headlines when they declared that they had been **unable to reproduce the findings in 47 of 53 'landmark' cancer papers** » (doi:10.1038/nature.2016.19269)

Repeat > **Replicate** > **Reproduce** > **Reuse**

Repeat	Replicate	Reproduce
Same experiment	Same experiment	Same experiment
Same setup	Same setup	~~Same setup~~
Same lab	~~Same lab~~	~~Same lab~~

new ideas,
new experiment,
some comonalities

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 .*

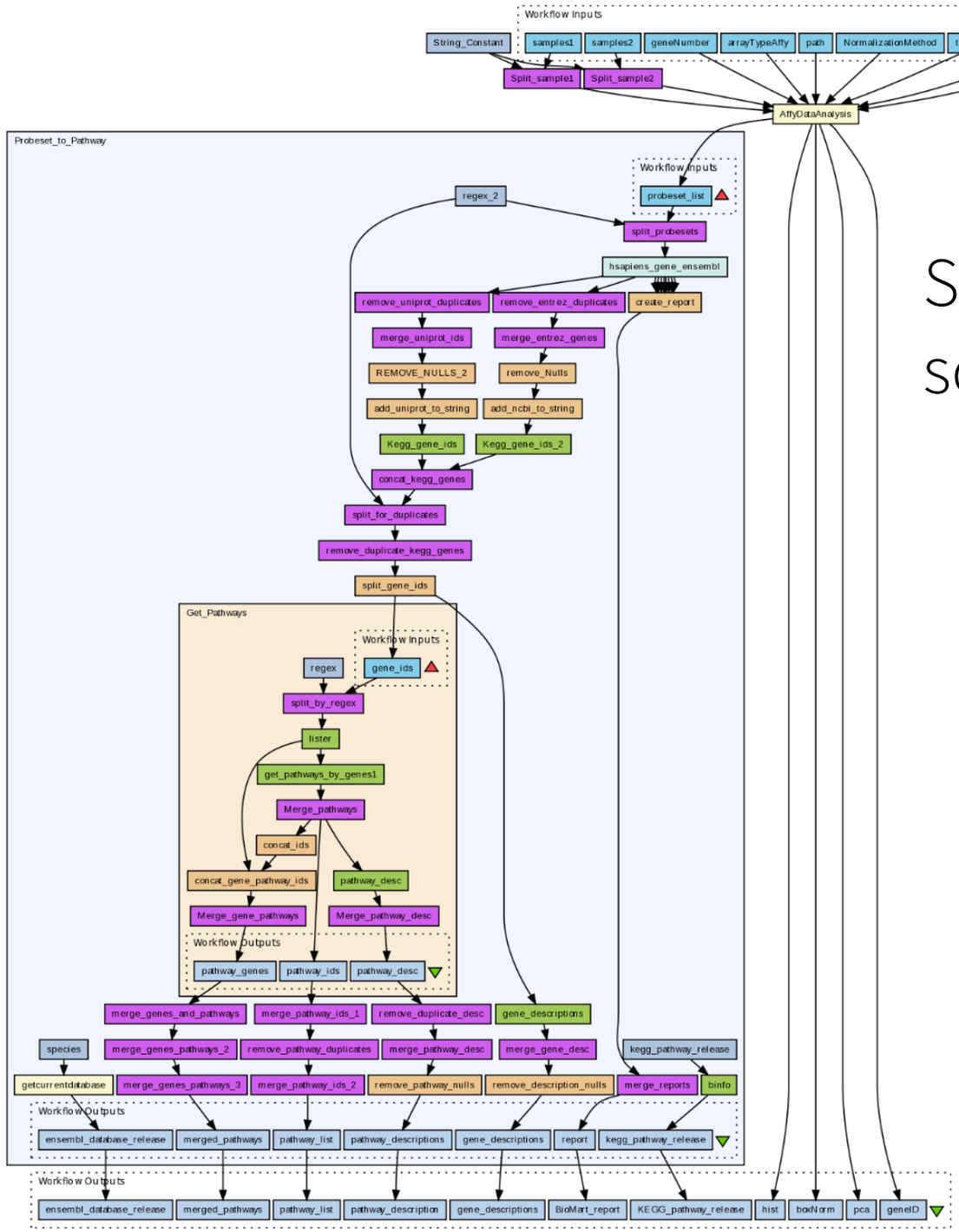
Scientific **workflows** to the rescue …

What is a workflow ?

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

« Workflow management **systems** (WMS) **perform** the complex analyses on a variety of **distributed resources** »

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 :

→ **automate** data analysis (at scale)

→ **abstraction** (describe/share methods)

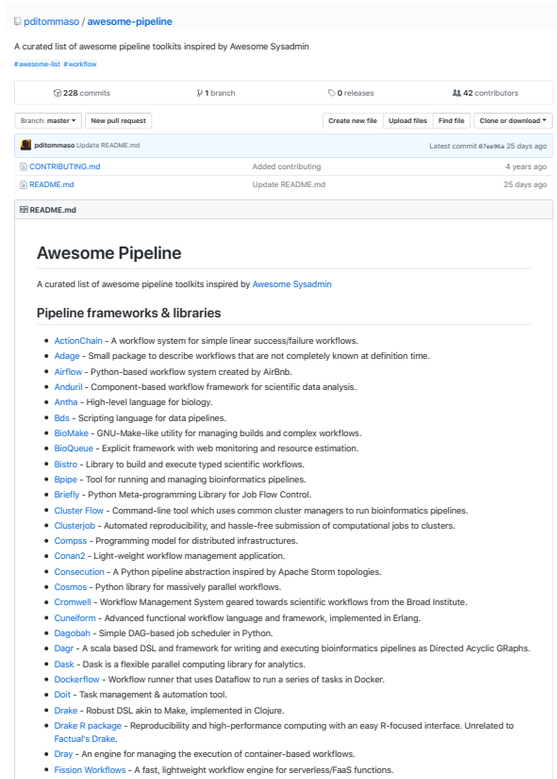
→ **provenance** (~transparency)

Galaxy COMMUNITY HUB

TensorFlow

Taverna

COMMON WORKFLOW LANGUAGE



Provenance : a way to **reuse** produced & analysed data

Definition: Oxford dictionnary

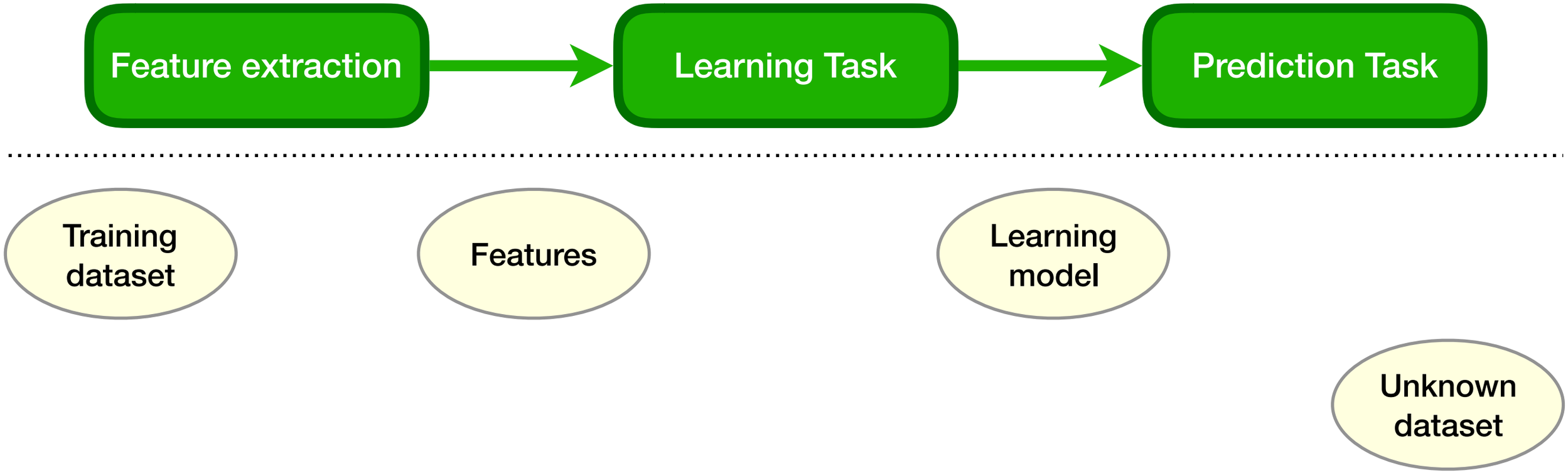
« The beginning of something's existence; something's origin. »

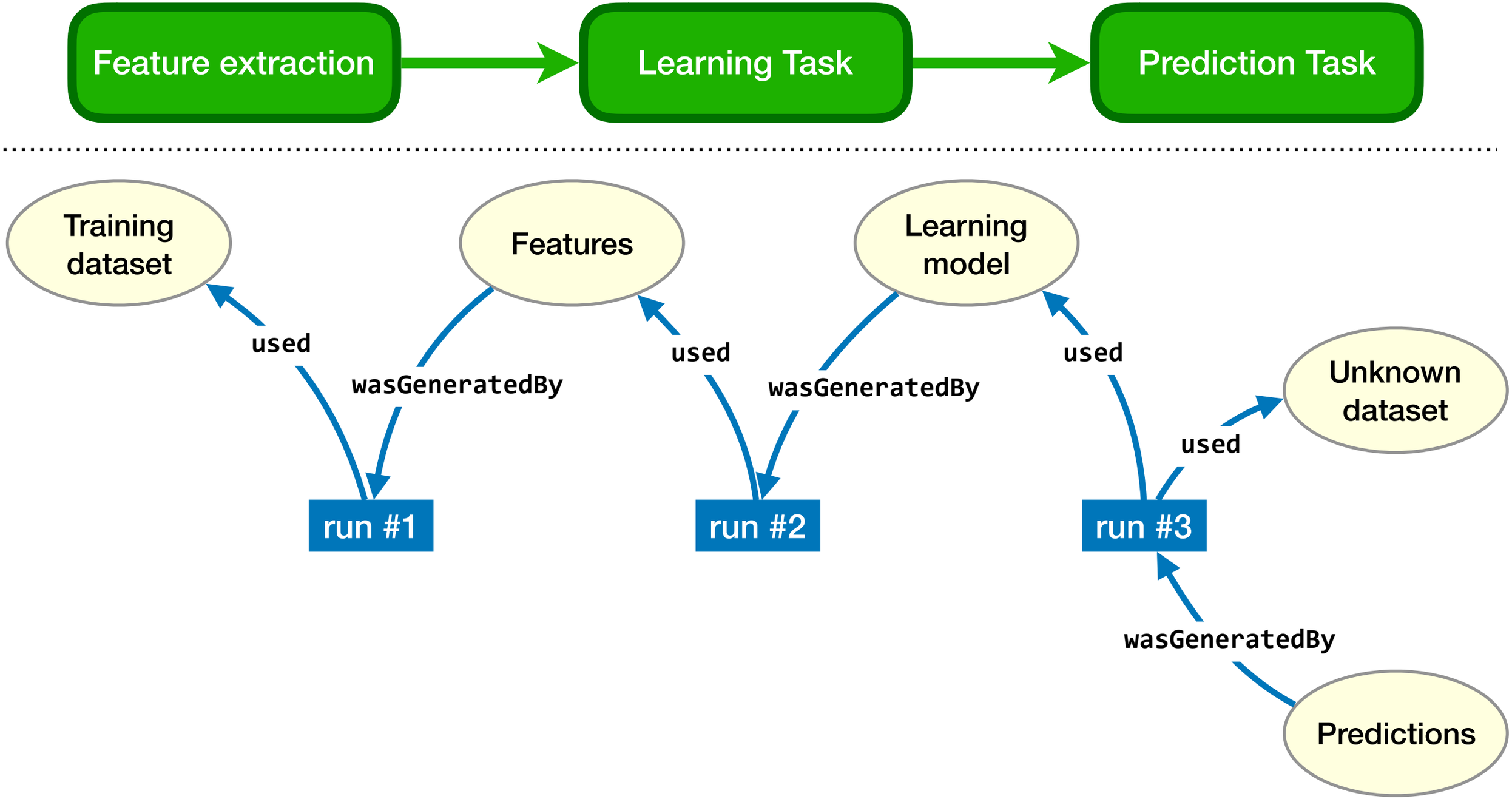
Definition: Computer Science

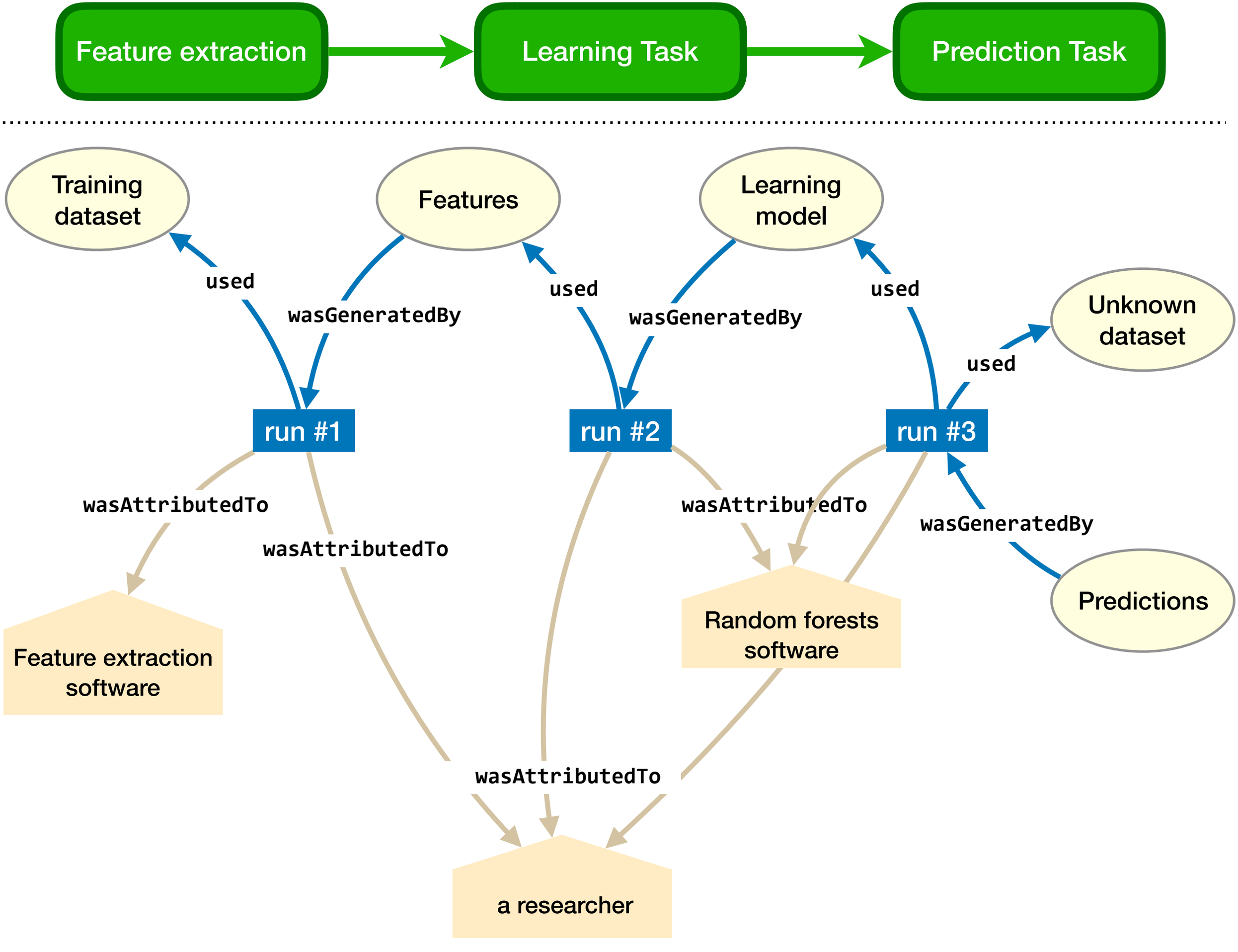
« Provenance information describes the **origins** and the **history of data in its life cycle**. »

« Today, data is often made **available on the Internet** with **no centralized control over its integrity**: data is constantly being created, copied, moved around, and combined indiscriminately. Because information sources (or different parts of a single large source) may 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*


Feature extraction
Learning Task
Prediction Task
Training dataset
Features
Learning model
Unknown dataset
Predictions


Feature extraction
Learning Task
Prediction Task
Training dataset
Features
Learning model
Unknown dataset
used
wasGeneratedBy
used
wasGeneratedBy
used
used
run #1
run #2
run #3
wasGeneratedBy
Predictions


Feature extraction
Learning Task
Prediction Task
Training dataset
Features
Learning model
Unknown dataset
used
wasGeneratedBy
used
wasGeneratedBy
used
used
run #1
run #2
run #3
wasAttributedTo
wasAttributedTo
wasAttributedTo
wasGeneratedBy
Feature extraction software
Random forests software
Predictions
wasAttributedTo
a researcher

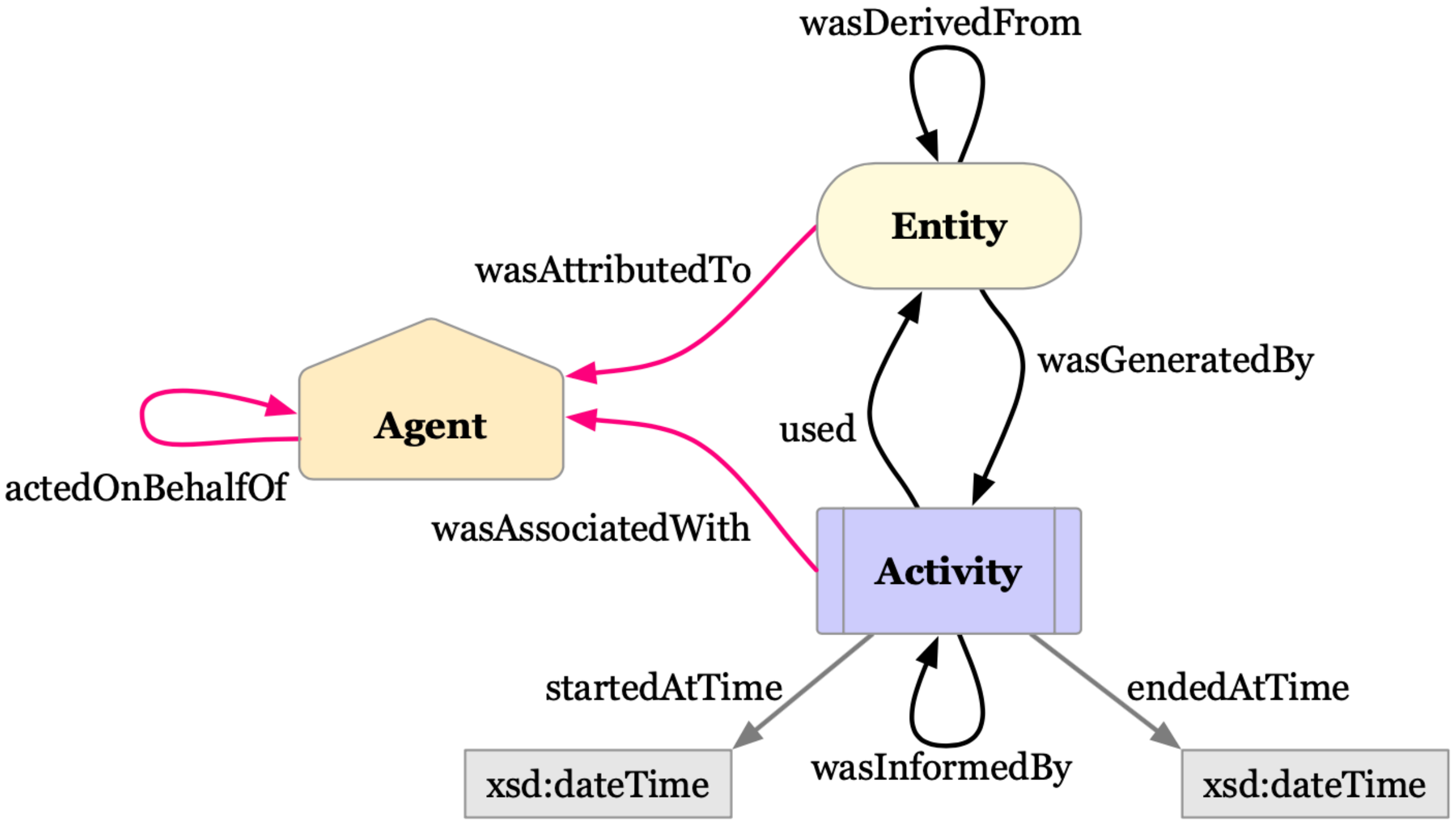
Representing provenance

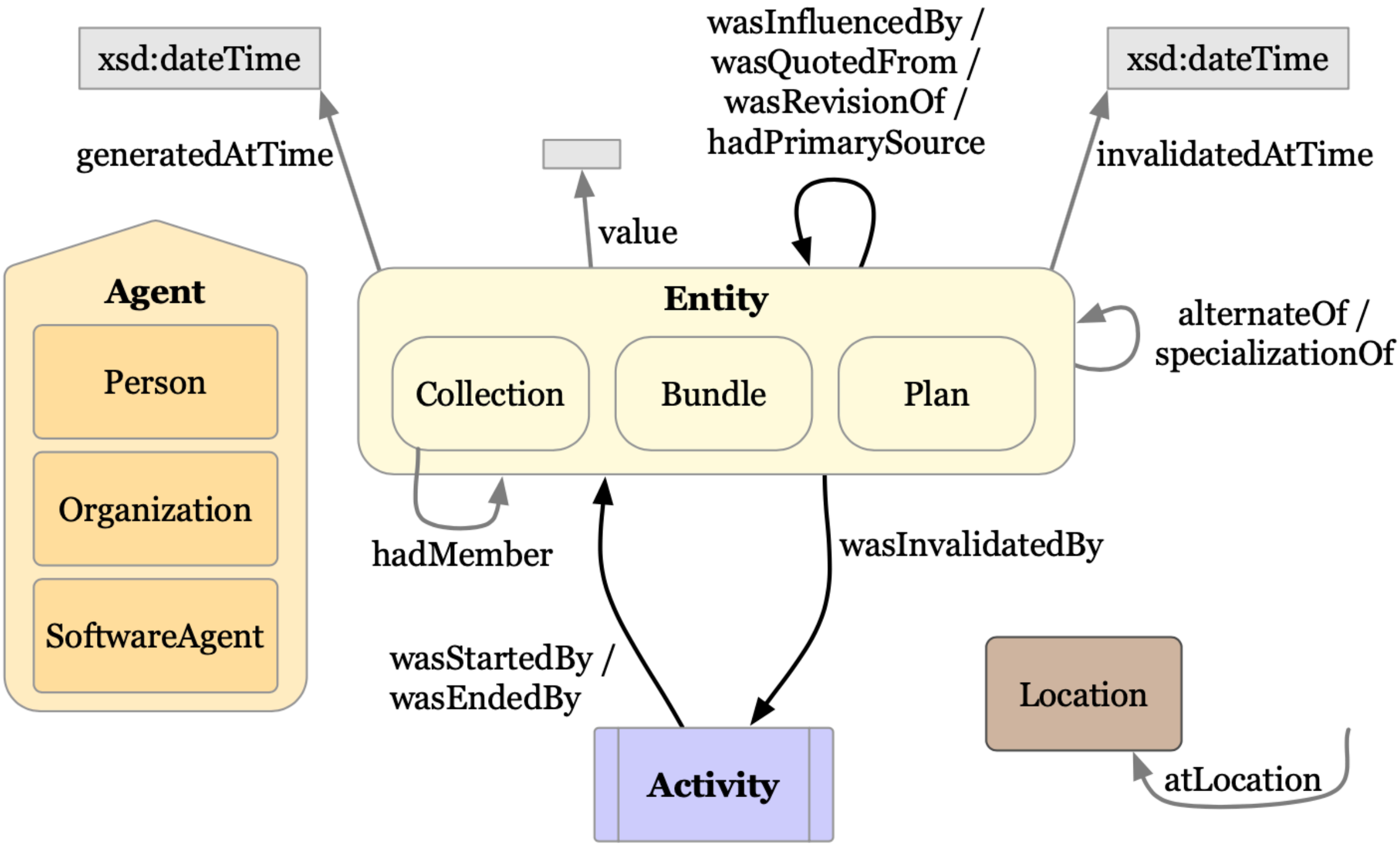
W3C Recommendation

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
wasInformedBy
xsd:dateTime
xsd:dateTime


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

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.

5.3 Derivations

Derivations with explicit activity, generation, and usage admit the following inference:

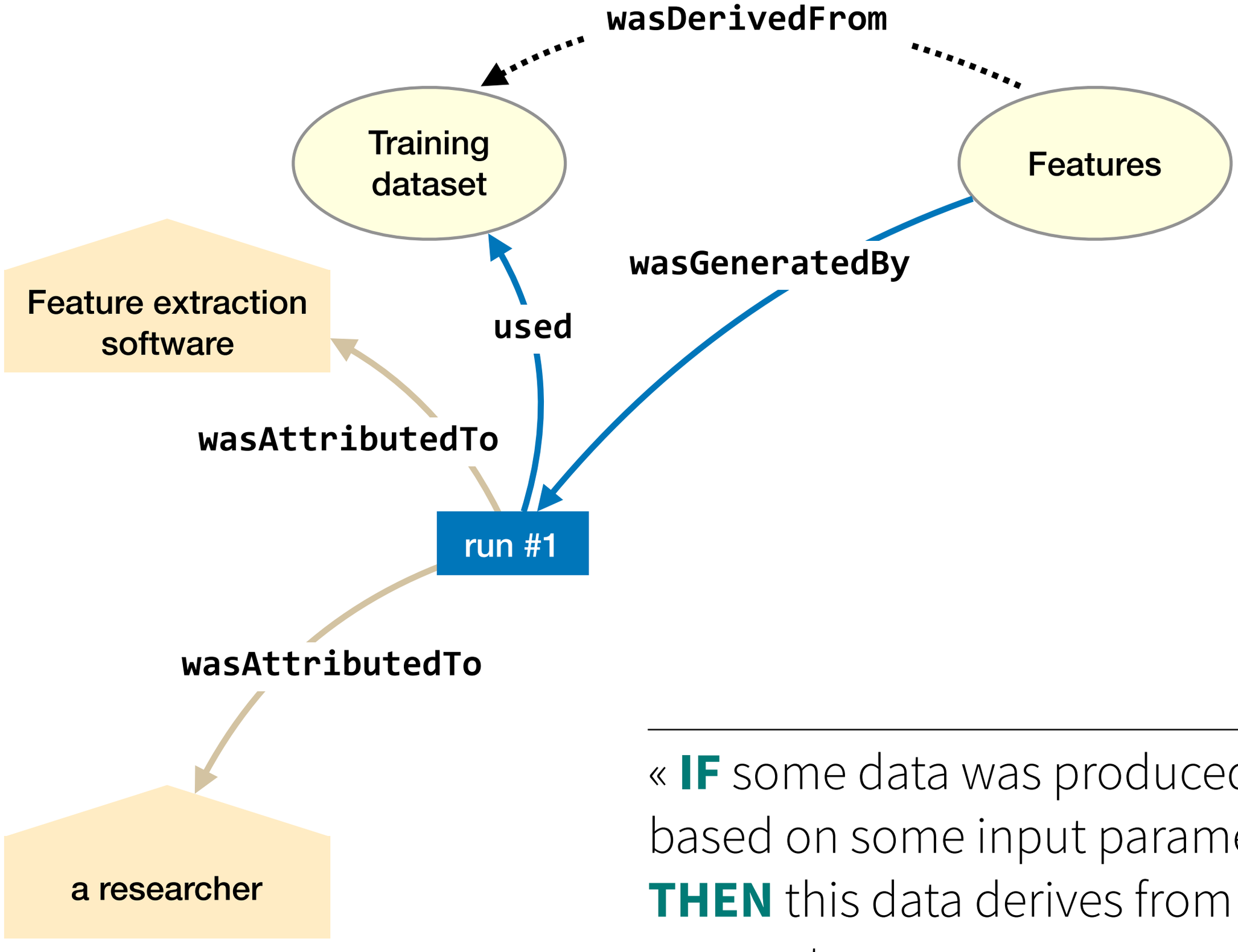
Inference 11 (derivation-generation-use-inference)

In this inference, none of `a`, `gen2` or `use1` can be placeholders `-`.

IF `wasDerivedFrom(_id; e2,e1,a,gen2,use1,_attrs)`, THEN there exists `_t1` and `_t2` such that `used(use1; a,e1,_t1,[])` and `wasGeneratedBy(gen2; e2,a,_t2,[])`.

Inference 15 (influence-inference)

1. IF `wasGeneratedBy(id; e,a,_t,attrs)` THEN `wasInfluencedBy(id; e, a, attrs)`.
2. IF `used(id; a,e,_t,attrs)` THEN `wasInfluencedBy(id; a, e, attrs)`.
3. IF `wasInformedBy(id; a2,a1,attrs)` THEN `wasInfluencedBy(id; a2, a1, attrs)`.
4. IF `wasStartedBy(id; a2,e,_a1,_t,attrs)` THEN `wasInfluencedBy(id; a2, e, attrs)`.
5. IF `wasEndedBy(id; a2,e,_a1,_t,attrs)` THEN `wasInfluencedBy(id; a2, e, attrs)`.
6. IF `wasInvalidatedBy(id; e,a,_t,attrs)` THEN `wasInfluencedBy(id; e, a, attrs)`.
7. IF `wasDerivedFrom(id; e2, e1, _a, _g, _u, attrs)` THEN `wasInfluencedBy(id; e2, e1, attrs)`. Here, `_a`, `_g`, `_u` MAY be placeholders `-`.
8. IF `wasAttributedTo(id; e,ag,attrs)` THEN `wasInfluencedBy(id; e, ag, attrs)`.
9. IF `wasAssociatedWith(id; a,ag,_pl,attrs)` THEN `wasInfluencedBy(id; a, ag, attrs)`. Here, `_pl` MAY be a placeholder `-`.
10. IF `actedOnBehalfOf(id; ag2,ag1,_a,attrs)` THEN `wasInfluencedBy(id; ag2, ag1, attrs)`.

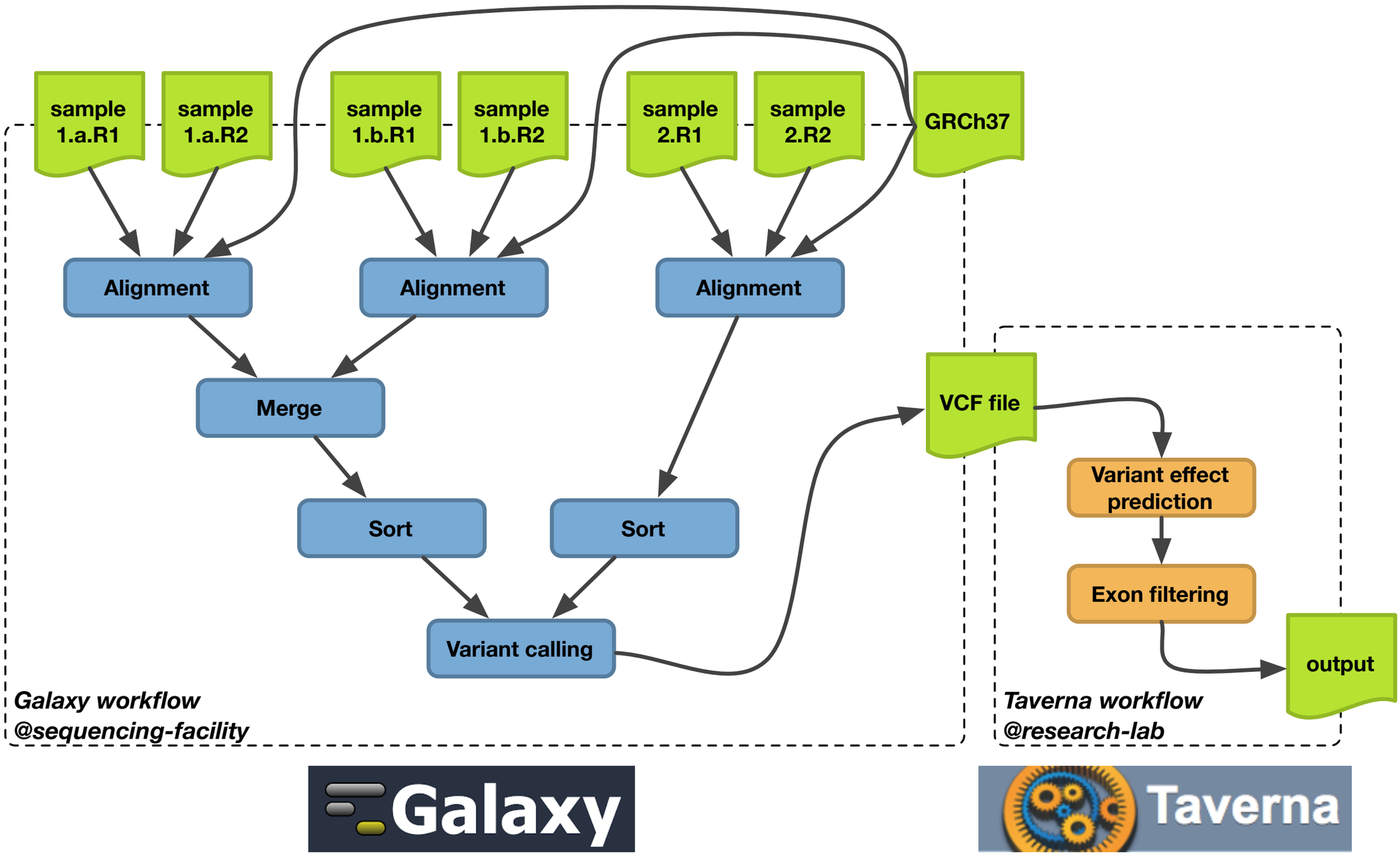


« **IF** some data was produced by a tool based on some input parameters, **THEN** this data derives from the input parameters »

Provenance

in **multi-site** studies ?

Multi-site studies → ≠ workflow engines !

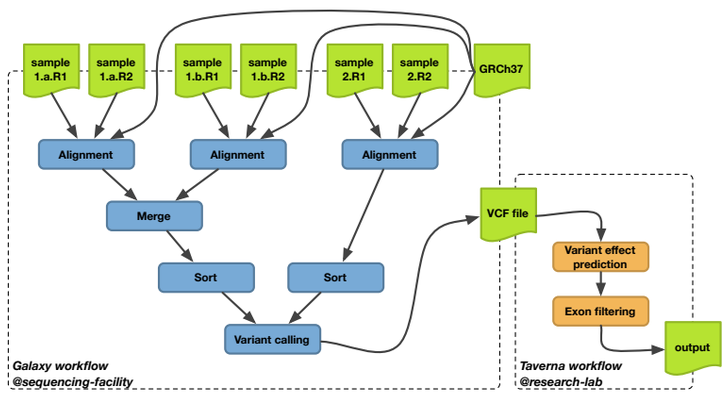


Galaxy

Taverna

Scattered provenance capture ?

Provenance issues



« Which alignment algorithm was used when predicting these effects ? »

« 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 !

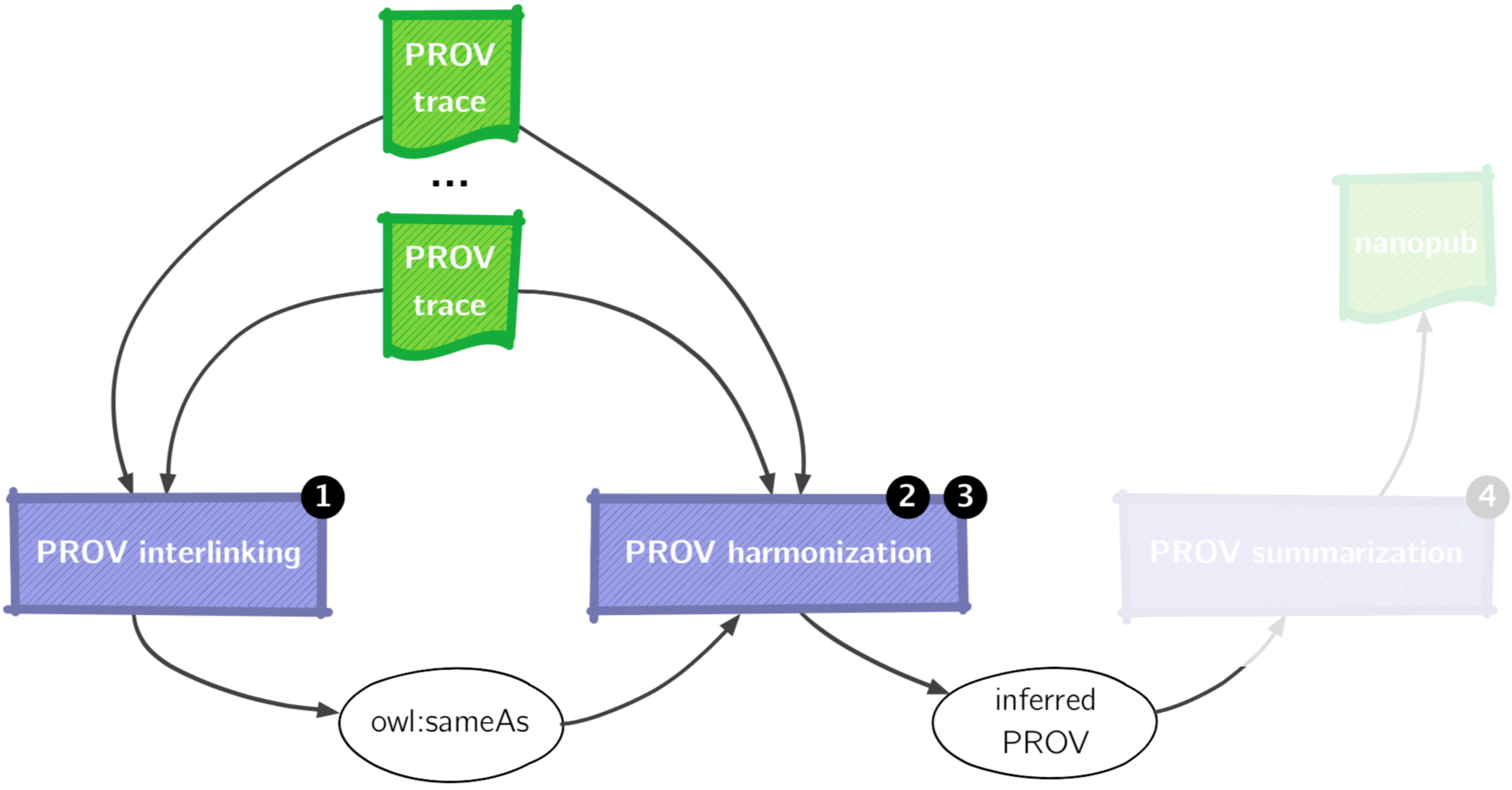
Provenance « heterogeneity »

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 ?

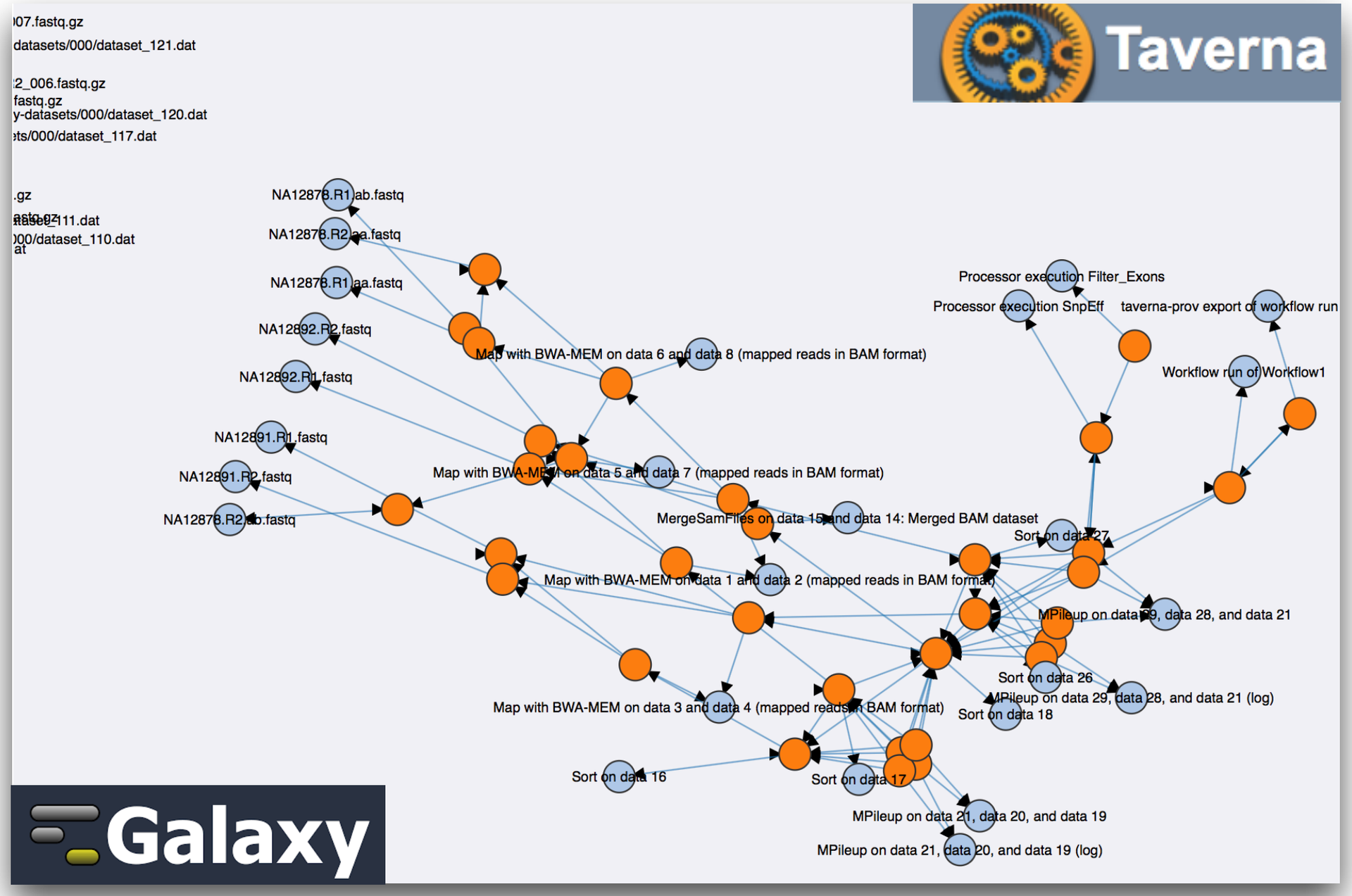
Approach



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 »



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

Reuse instead of re-execution ?

Is provenance enough for reuse ?

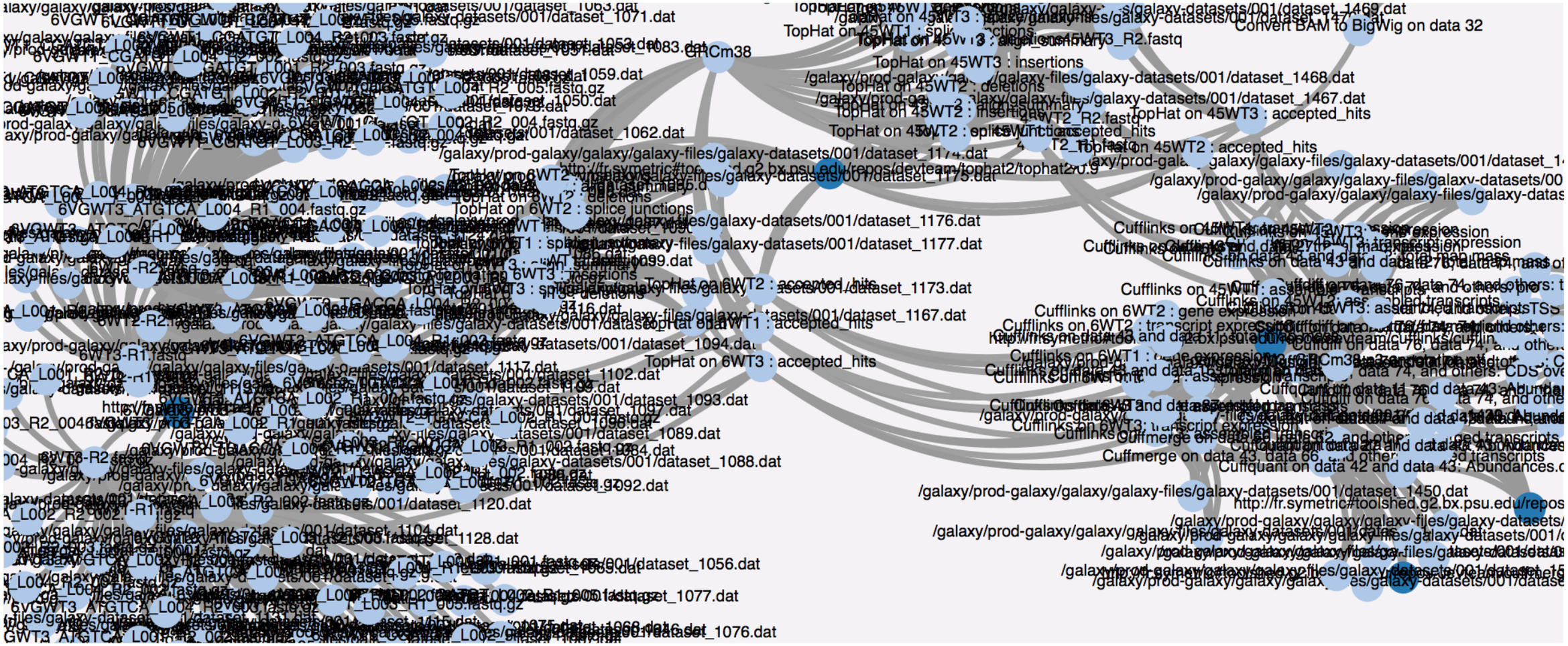
```
   a prov:Bundle, prov:Entity;
   prov:wasAttributedTo <#galaxy2prov>;
   prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
.

<#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;
```

Visualise

Too fine-grained

No domain concepts



Semantic tool 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

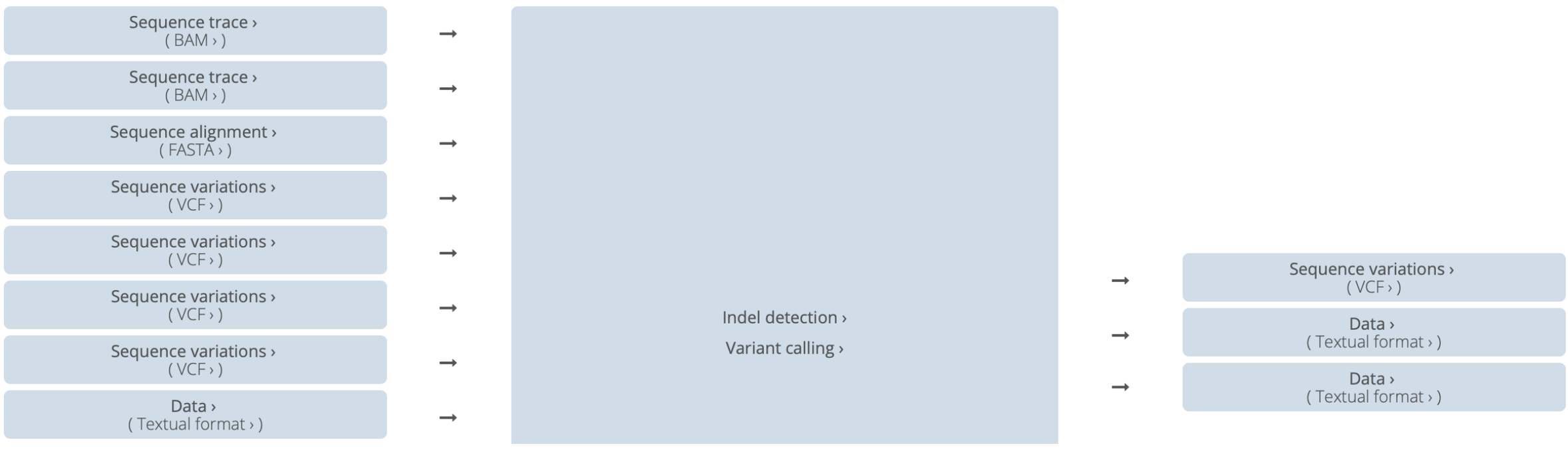
2.4-9

Sequencing › | DNA polymorphism › | Genetic variation ›

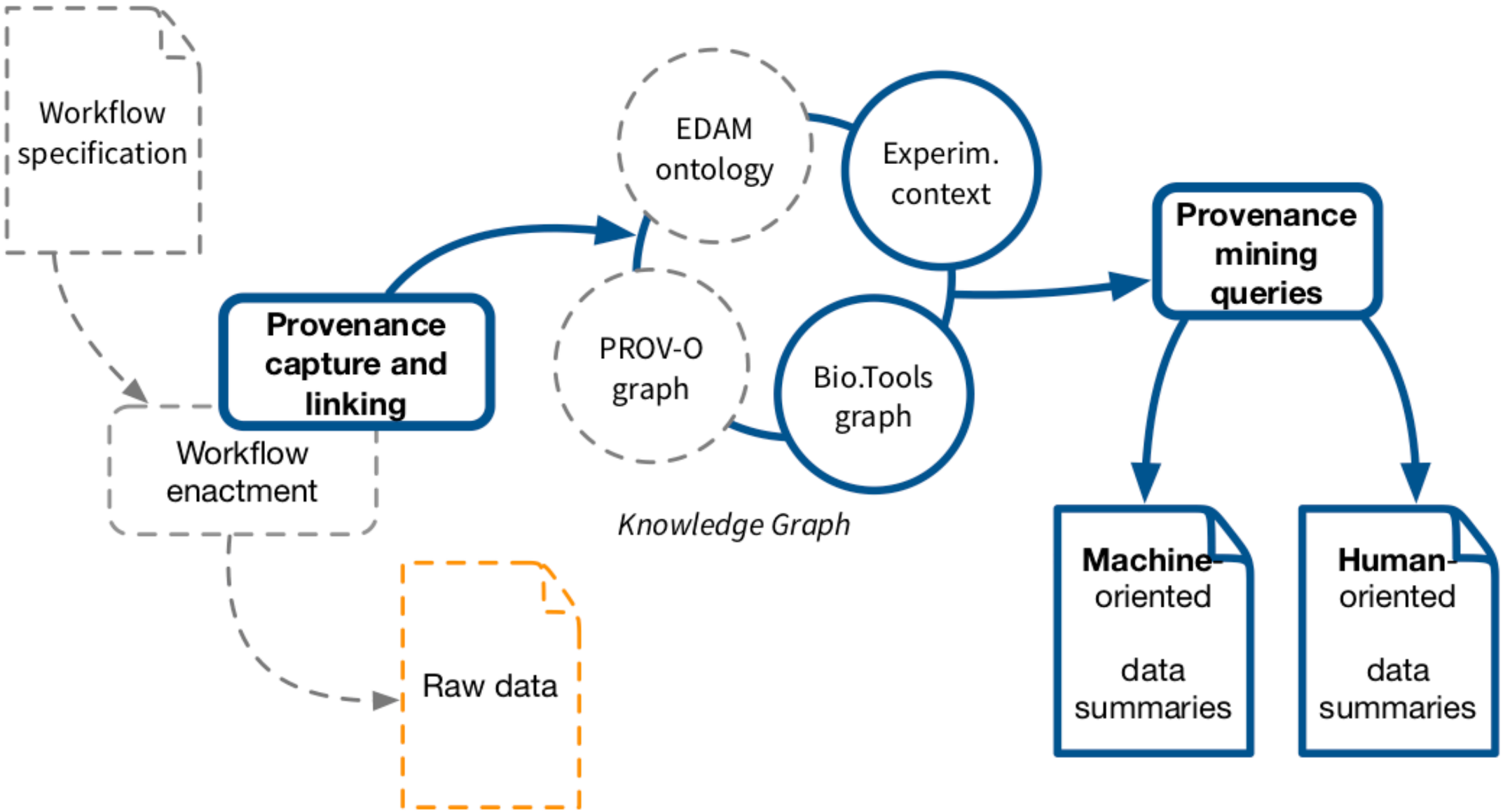
Mature | Open access

Web application | Java

SNP and indel caller.

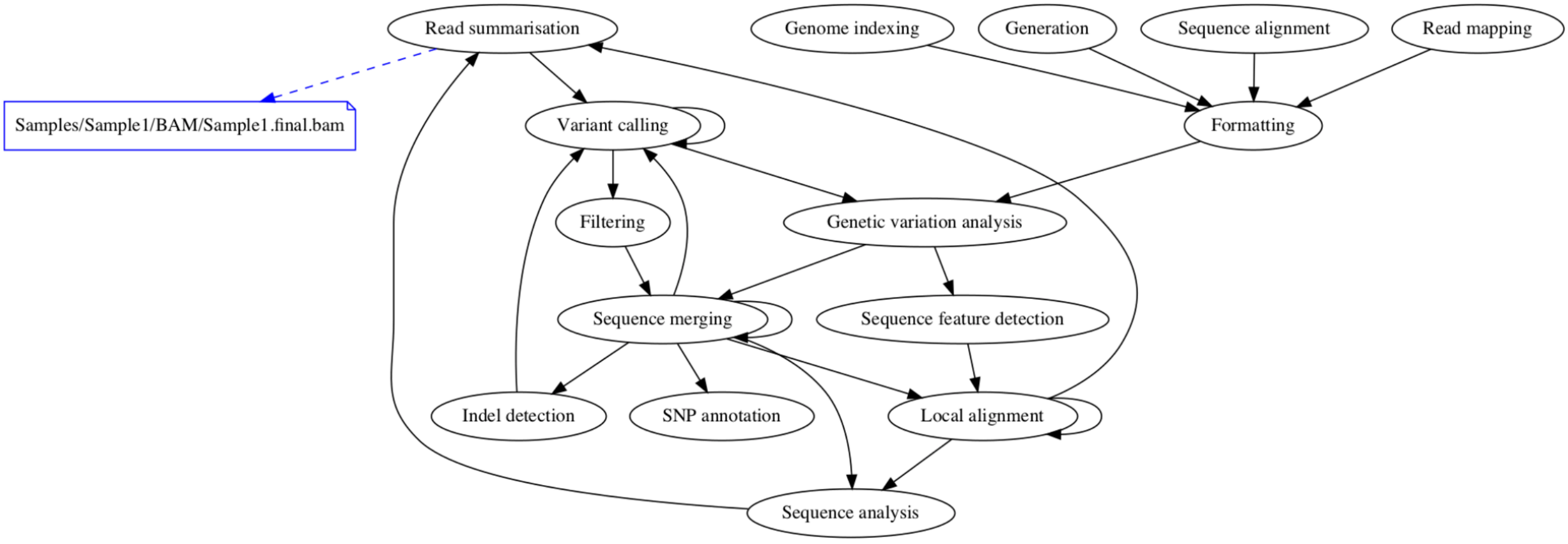


Approach



Methods and tools : graph pattern matching, inference rules, SPARQL, Python, Jupyter

Results




```
...
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

Implementation

jupyter FRESH-notebook (autosaved)

File Edit View Insert Cell Kernel Widgets Help — Not Trusted — Python 3

Run — Code

3. Human-oriented data summaries

Sentence-based data explanations

Here, the goal is to describe a piece of data with the consensual definition of **what** does the tool that generates this piece of data.
Technically, this is done with a SPARQL query that combine the provenance information (*prov:wasGeneratedBy*), the description of the tool (*biotools:has_function*), and the domain knowledge on the nature of the processing (*oboInOwl:hasDefinition*).

In [6]:
```
%%time
query = """
    SELECT ?d_label ?title ?f_def ?st WHERE {
        ?d rdf:type prov:Entity ;
           prov:wasGeneratedBy ?x ;
           prov:wasAssociatedWith ?tool ;
           rdfs:label ?d_label .

        ?tool dc:title ?title ;
            biotools:has_function ?f .

        ?f rdfs:label ?f_label ;
            oboInOwl:hasDefinition ?f_def .

        ?c rdf:type mp:Claim ;
            mp:statement ?st .
    }
"""

results = g.query(query)
for r in results :
    display(Markdown('The file `' + str(r['d_label']) + '`'
                     +' **results from tool** ' + str(r['title'])
                     + ' **which** ' + str(r['f_def'])))
    display(Markdown(' **It was produced in the context of** ' + str(r['st']) ))
```

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

The file `VCF/hapcaller.indel.recal.select.vcf.gz` **results from tool** gatk2_variant_select-IP **which** Identify and map genomic alterations, including single nucleotide polymorphisms, short indels and structural variants, in a genome sequence.

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

The file `VCF/hapcaller.snv.recal.select.vcf.gz.tbi` **results from tool** gatk2_variant_select-IP **which** Identify and map genomic alterations, including single nucleotide polymorphisms, short indels and structural variants, in a genome sequence.

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

Summary

Take home message & open questions

- **Scientific Workflows** → automation, abstraction, provenance
- Standards for **provenance representation** and **reasoning**

- Better handle **multi-site studies** (ESWC'17 satelite event paper)
- Linked experiment reports = **contextualized** and **summarized** provenance (TaPP'16 paper, Semantic Web Journal (in revision))

- Distributed data analysis → **Distributed provenance, reasoning** ?
- **Learning patterns** in provenance graphs ?
- **Predicting domain-specific annotation** for workflow results ? What about trust ?

Acknowledgments

Audrey Bihouée, Institut du Thorax, BiRD Bioinformatics facility, University of Nantes

Hala Skaf-Molli, LS2N, University of Nantes

Khalid Belhajjame, LAMSADE, University of Paris-Dauphine, PSL

GDR MaDICS

action **ReproVirtuFlow**

Inserm cnrs UNIVERSITÉ DE NANTES CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES l'institut du thorax