Découvrabilité et réutilisation de données produites par des workflows (en génomique)

IC 2021

Alban Gaignard[1], Hala Skaf-Molli[2], Khalid Belhajjame[3]

02 juillet 2021

[1] l'institut du thorax, INSERM, CNRS, Université de Nantes

[2] LS2N - CNRS, Université de Nantes, France

[3] LAMSADE – PSL, Université Paris-Dauphine, France

A. Gaignard, H. Skaf-Molli, K. Belhajjame Findable and reusable workflow data products: A genomic workflow case study. Semantic Web Journal 11 (5), 751-763

Context

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.548 | 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.

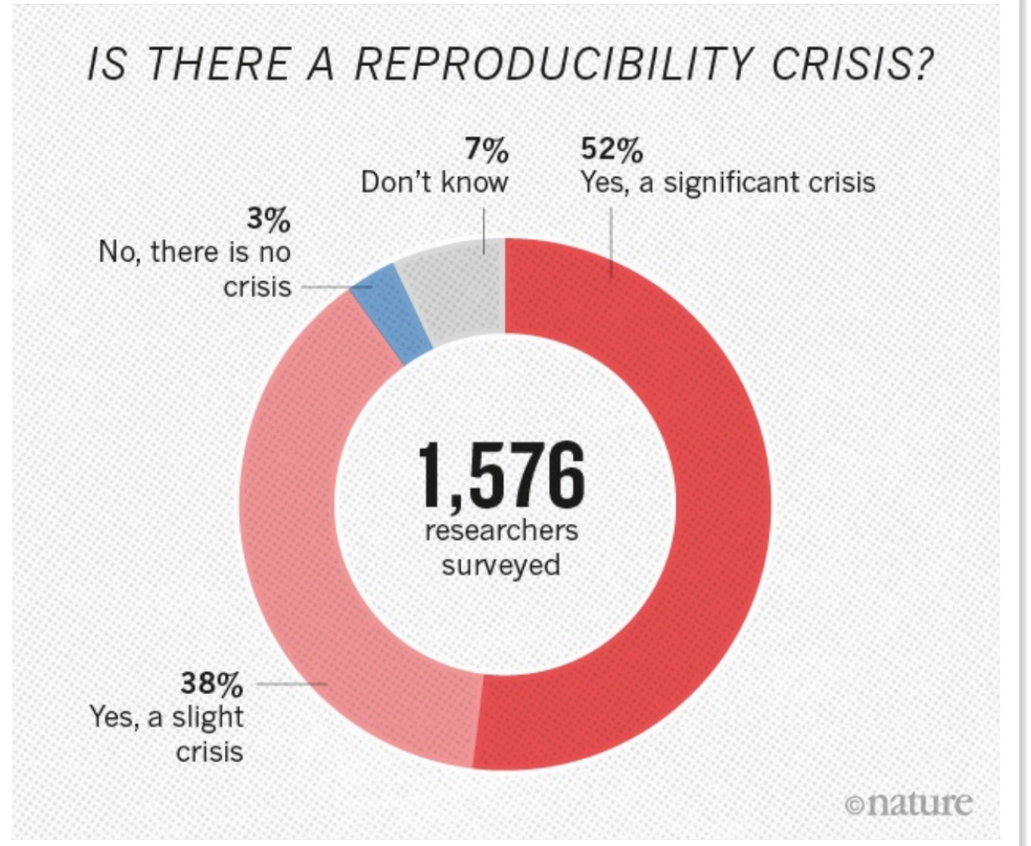
Raise standards for preclinical cancer research

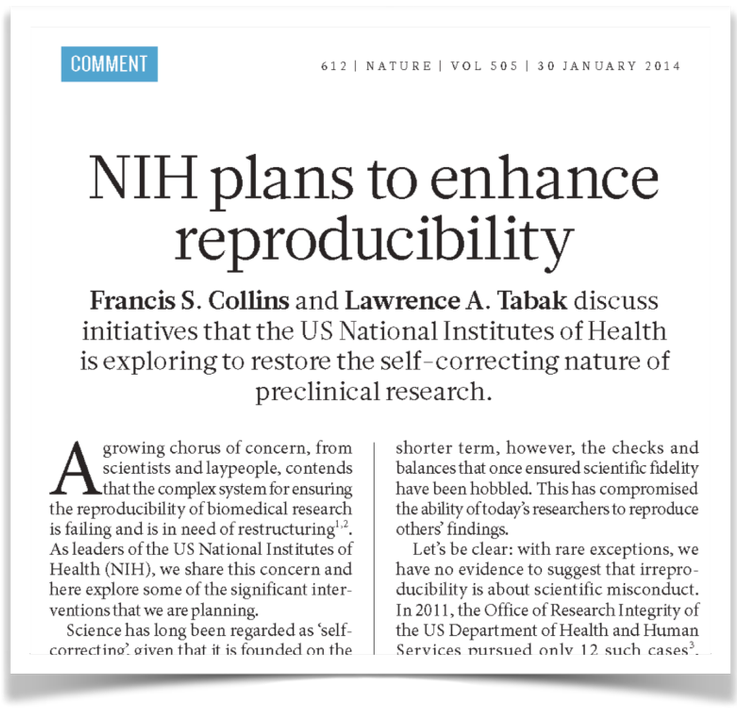
C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical 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 development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and 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[2] 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].

« researchers made headlines when they declared that they had been **unable to reproduce the findings in 47 of 53 'landmark' cancer papers** »

Baker, Monya. « **Biotech Giant Publishes Failures to Confirm High-Profile Science** ». *Nature*, vol. 530, nº 7589, février 2016, p. 141-141. *DOI.org (Crossref)*, doi:10.1038/nature.2016.19269.

"Reusing" is challenging

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

Reuse *new* ideas, *some* commonalities, *new* experiment

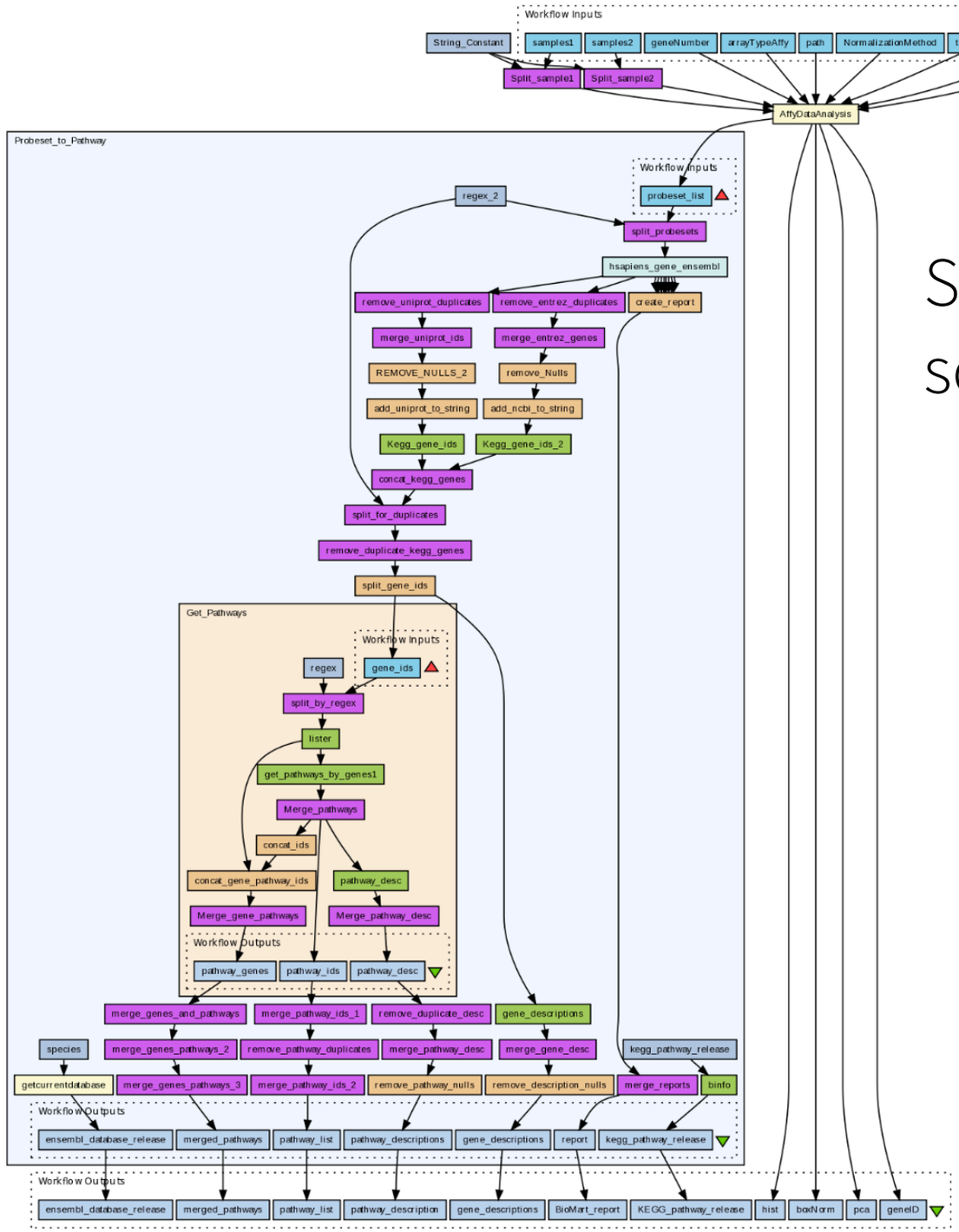
S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. Le Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal, C. Blanchet, **Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities**, Future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

Scientific **workflows** to the rescue …

« 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 :

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

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

→ **provenance** (~transparency)

Taverna

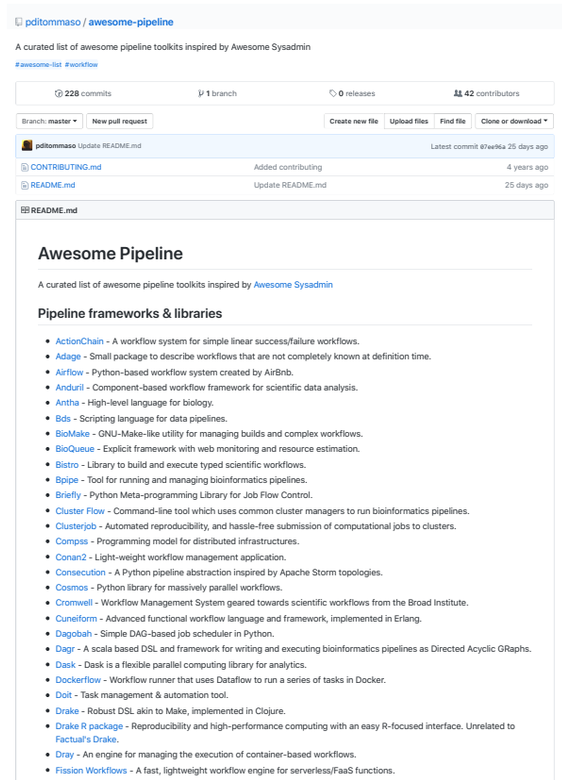
TensorFlow

Galaxy COMMUNITY HUB

nextflow

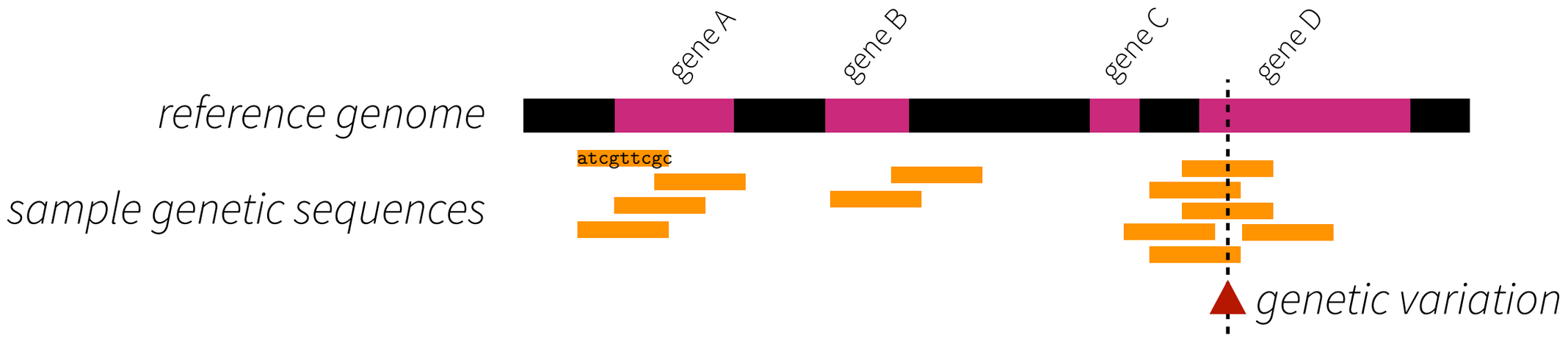
snakemake

COMMON WORKFLOW LANGUAGE

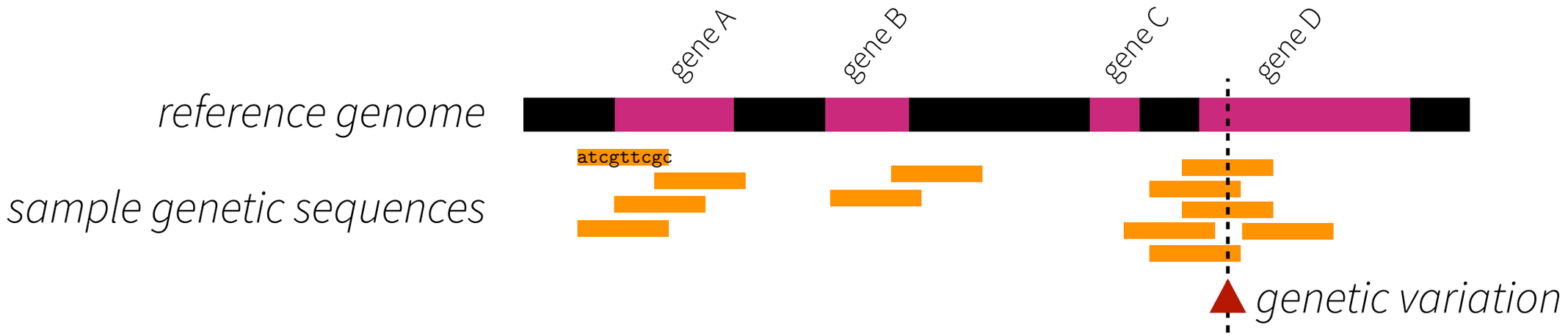


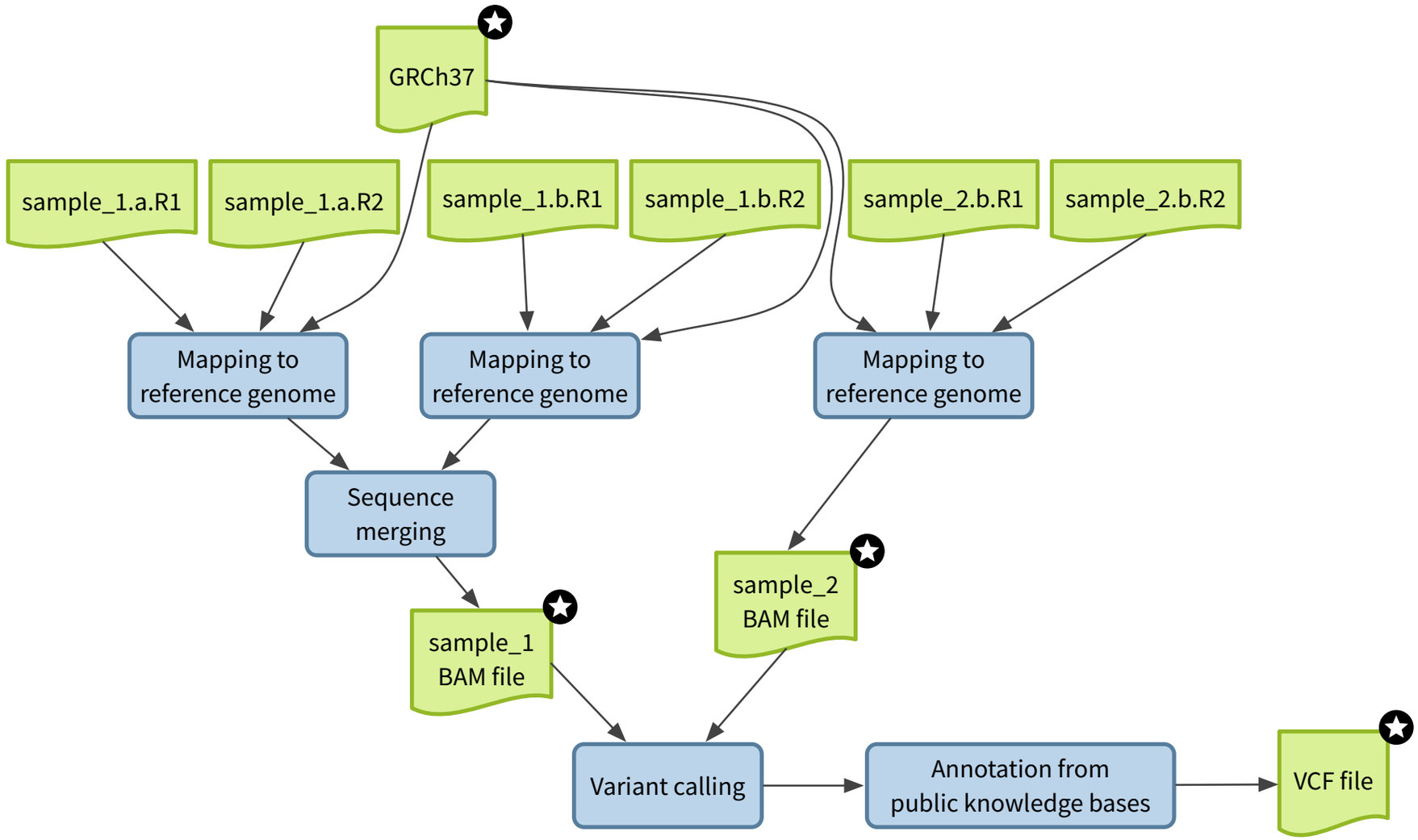
Problem statement

Typical genomic variant detection workflow



Typical genomic variant detection workflow





Computational costs

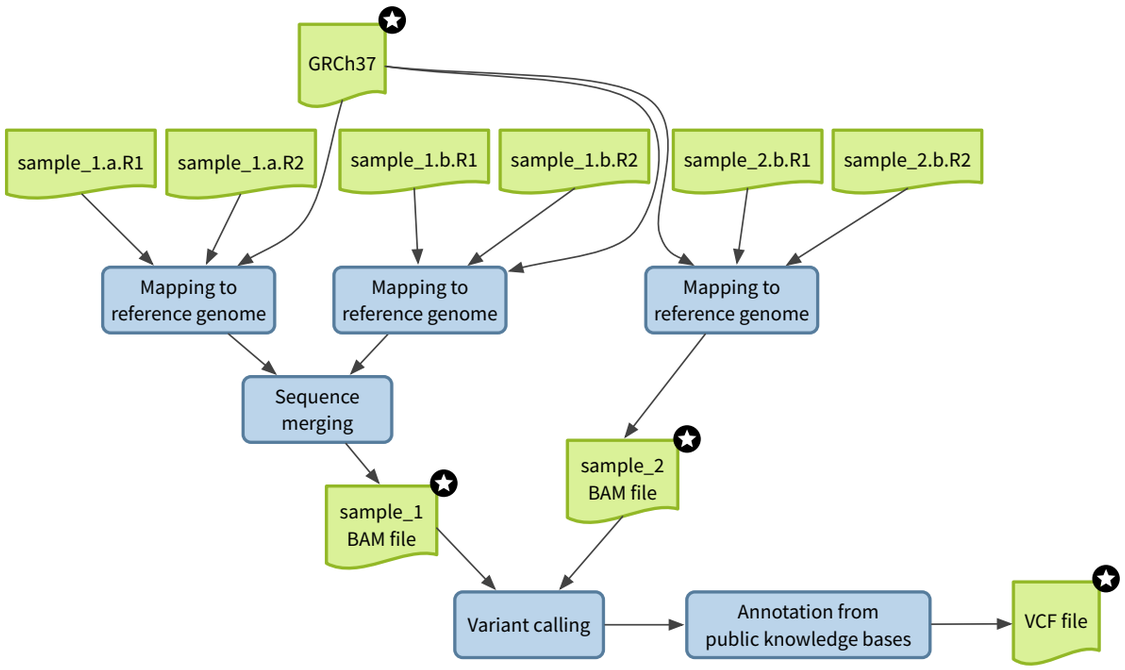
as estimated by CNRGH, Evry

	1 sample	*100 samples*	*5000 samples*
Storing processed exome data (compressed)	18.6 GB	1.8 TB	93 TB
Computating genomic variations (HPC, 196 cores)	2.5 hours	10+ days	1+ year
Computing (single CPU core)	20+ days / CPU	5+ years / CPU	250+ years / CPU

Re-computing: waste of computing time and storage

Can we better **reuse data ?**

Issues when reusing bioinformatics data



« A new tool is available, which data subset should I reprocess ? »

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

→ need for an overall **tracking of provenance**

→ need for **domain-specific** contextual metadata

Objectives

Limit the duplication of computing and storage efforts through better **data reuse**

1. Feed a **knowledge graph** with

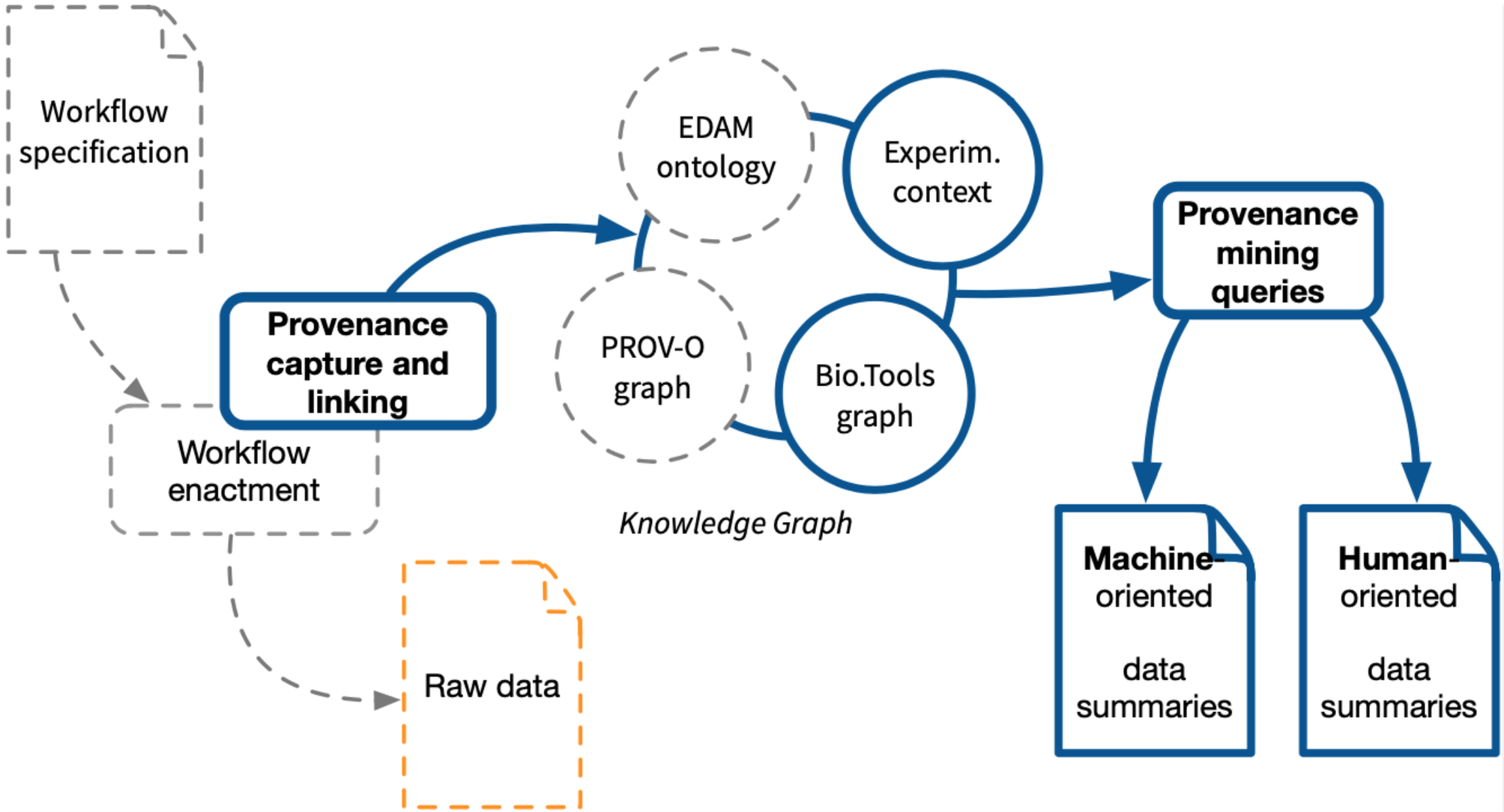
- generic **provenance** metadata
- **bioinformatics** community knowledge

2. Mine this knowledge graph to automatically produce

- **Machine**-oriented **data summaries**
- **Human**-oriented data **summaries**

Tooling a knowledge graph, to produce data summaries

Approach



"Which was the reference genome used to produce this VCF file ?"

"A new tool is available, which raw data should I reprocess ?"

A. Gaignard, H. Skaf-Molli, K. Belhajjame **Findable and reusable workflow data products: A genomic workflow case study**. Semantic Web Journal 11 (5), 751-763

1. Capturing **provenance** at run-time

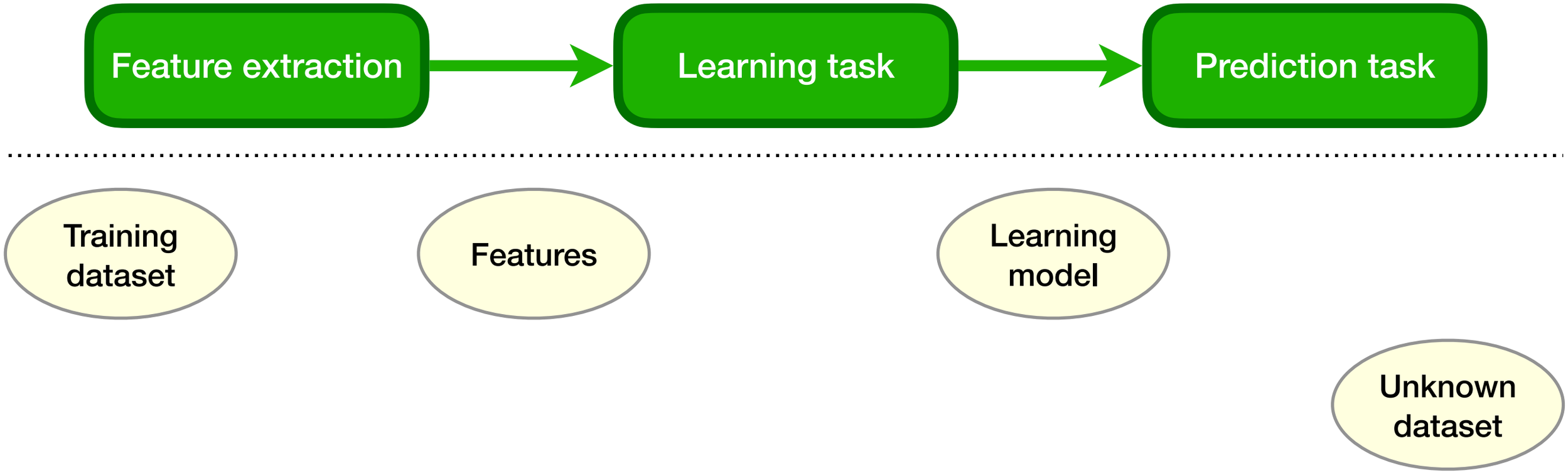
What is provenance metadata

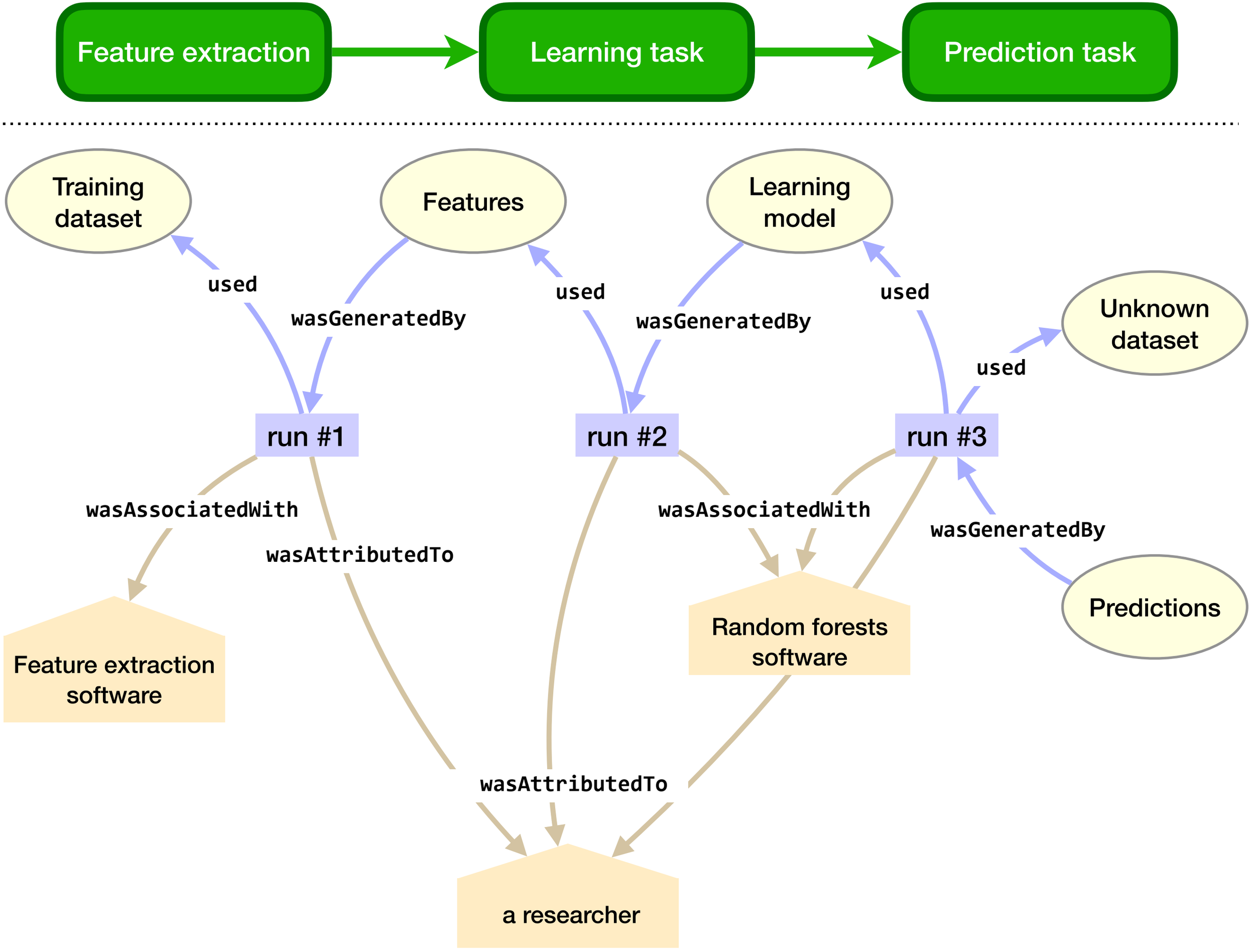
Definitions in Computer Science

« 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
wasAssociatedWith
wasAttributedTo
wasAssociatedWith
wasGeneratedBy
Feature extraction software
Random forests software
Predictions
wasAttributedTo
a researcher

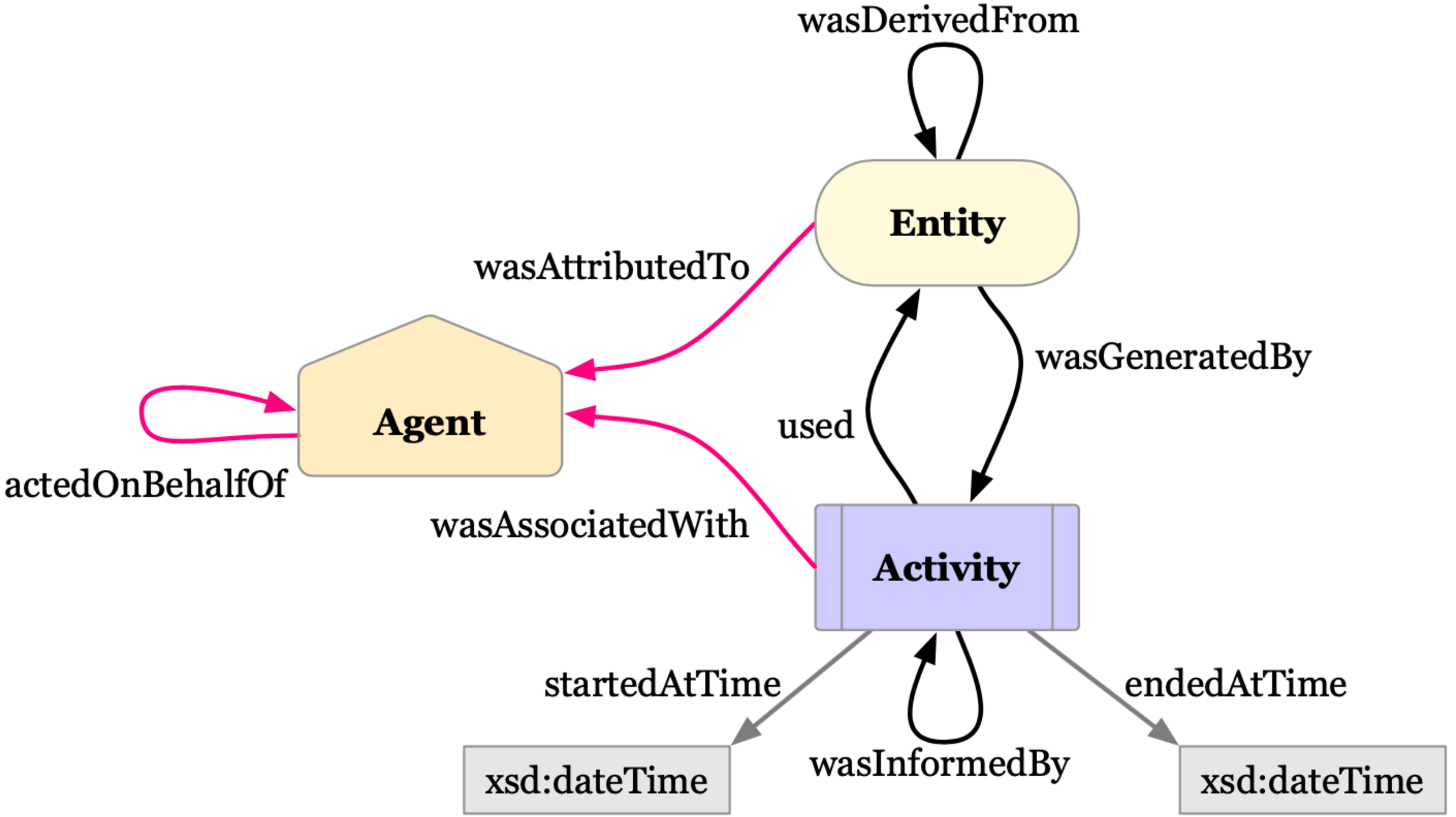
Representing provenance

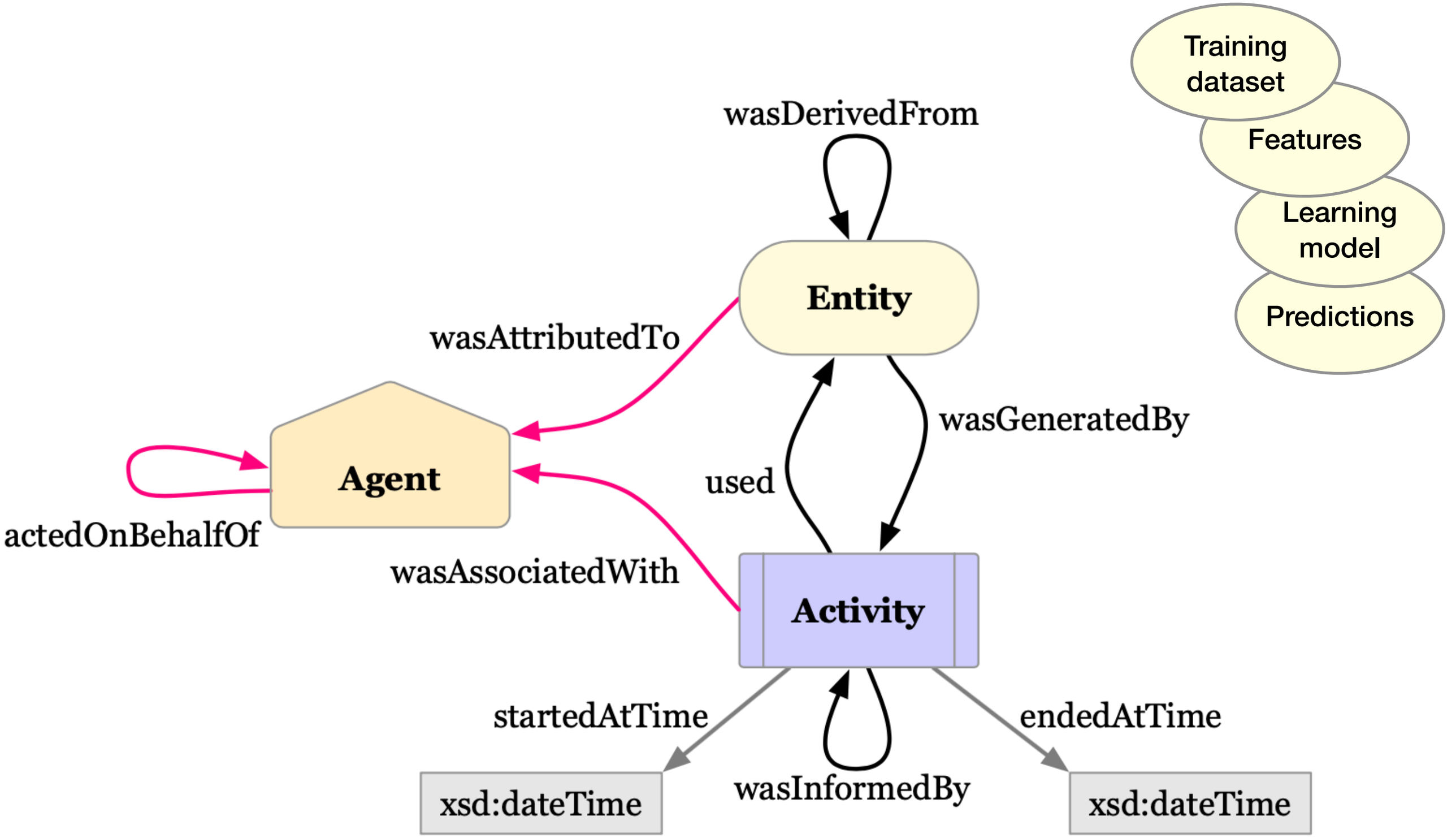
W3C Recommendation

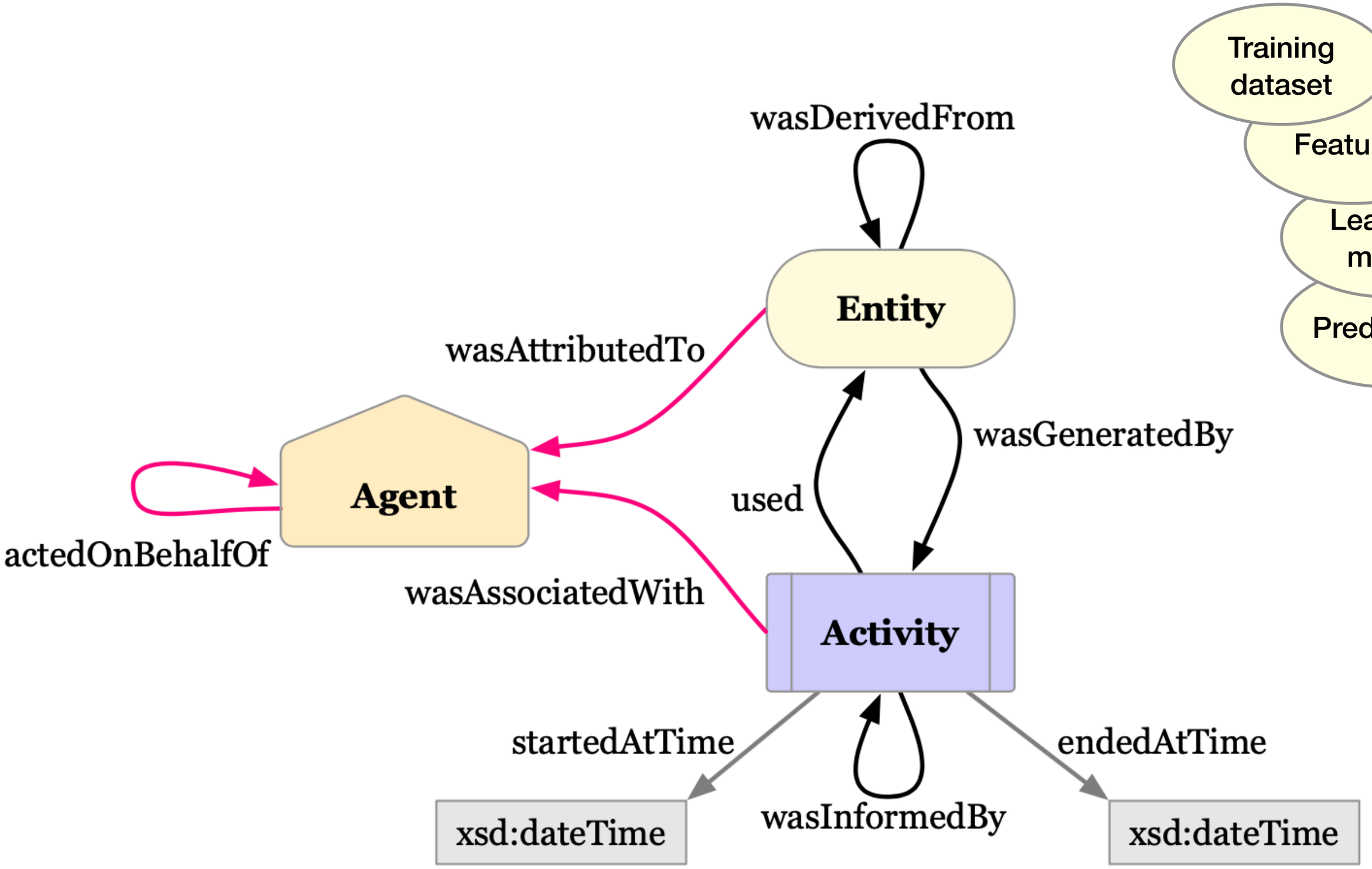
PROV-O: The PROV Ontology

W3C Recommendation 30 April 2013

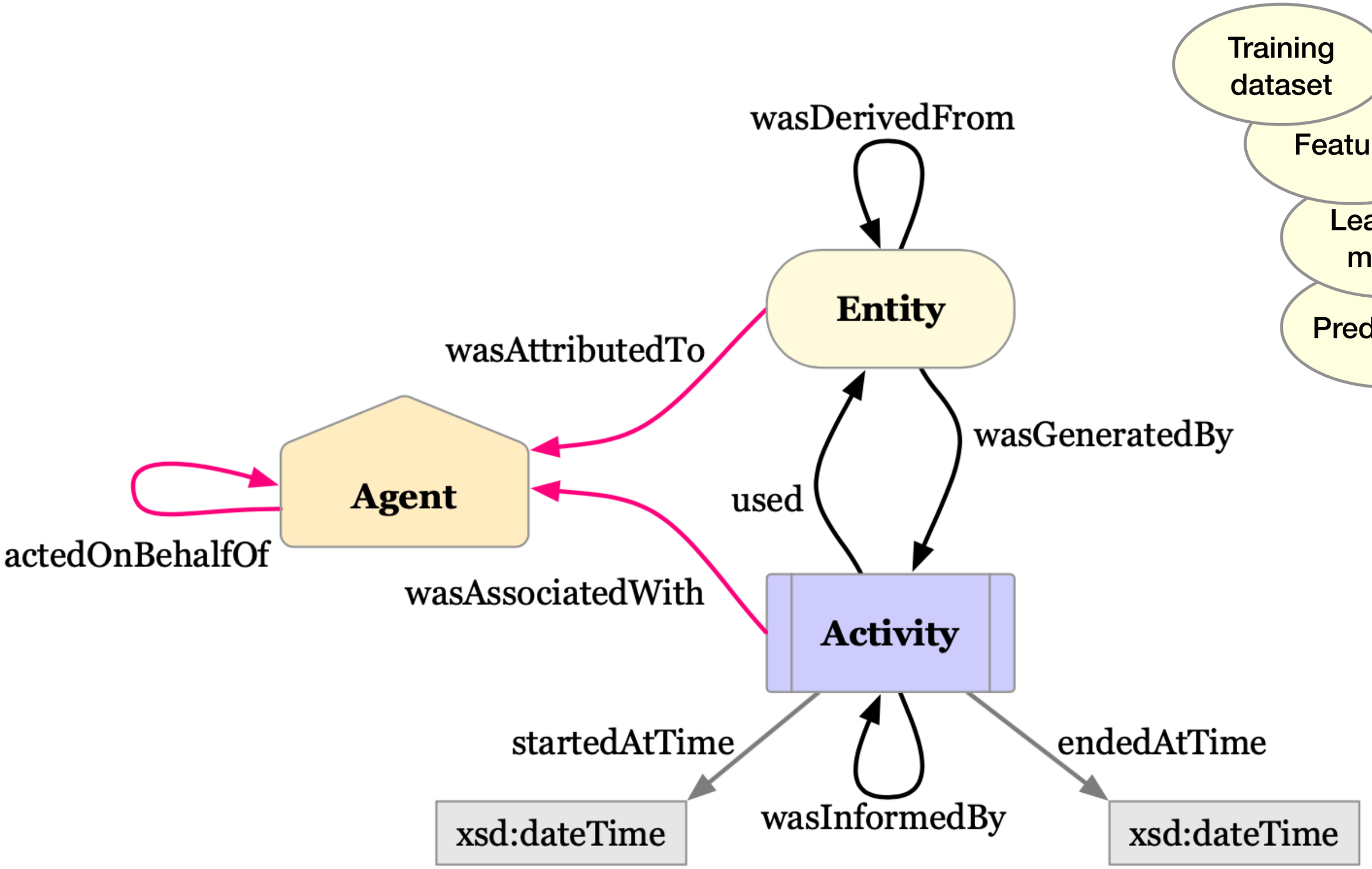
This version:
http://www.w3.org/TR/2013/REC-prov-o-20130430/
Latest published version:
http://www.w3.org/TR/prov-o/
Implementation report:
http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/
Previous version:
http://www.w3.org/TR/2013/PR-prov-o-20130312/
Editors:
Timothy Lebo, Rensselaer Polytechnic 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
wasInformedBy
endedAtTime
xsd:dateTime
xsd:dateTime


Training dataset
Features
Learning model
Predictions
wasDerivedFrom
Entity
wasAttributedTo
wasGeneratedBy
Agent
used
actedOnBehalfOf
wasAssociatedWith
Activity
startedAtTime
endedAtTime
wasInformedBy
xsd:dateTime
xsd:dateTime

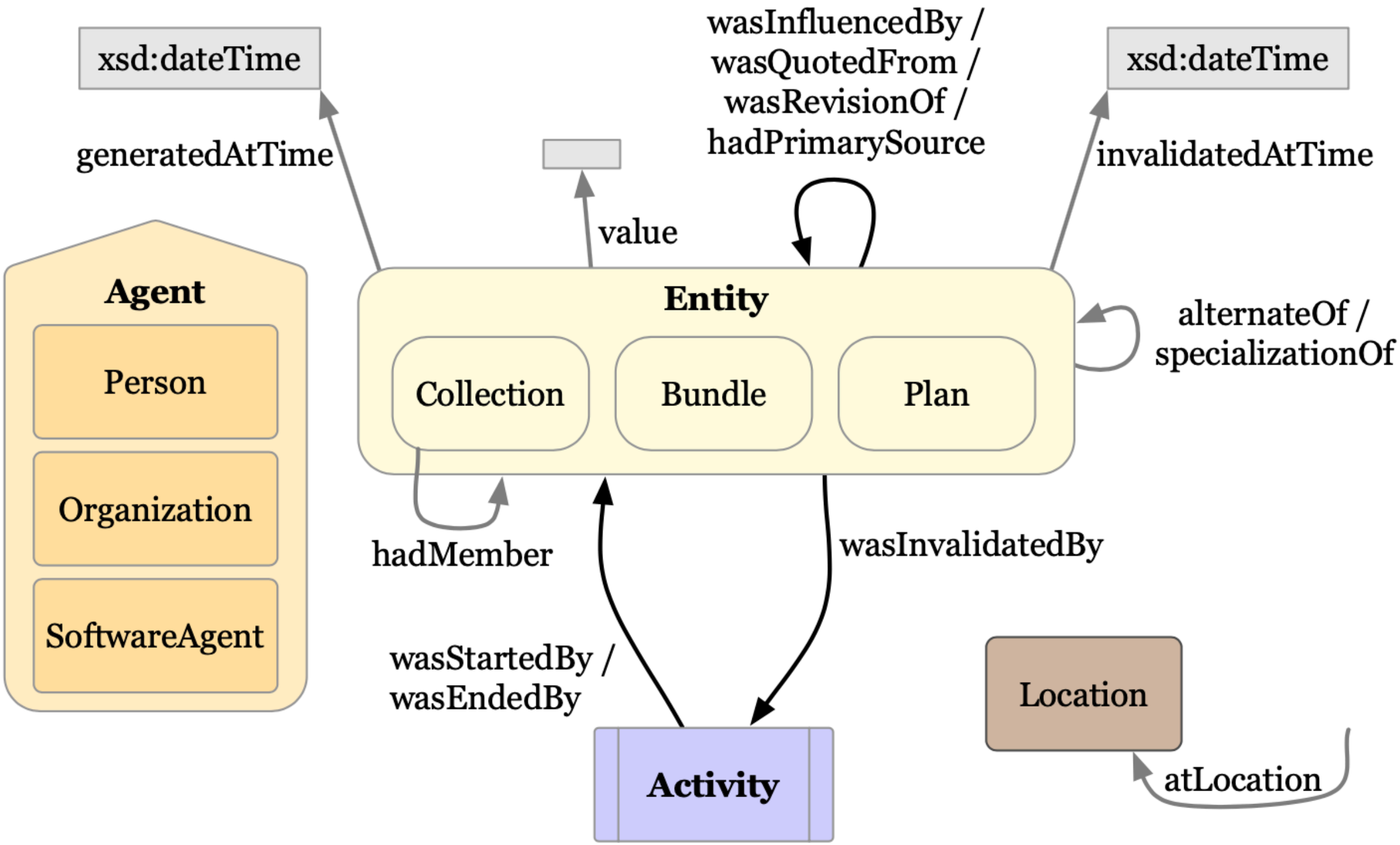


Scikit-learn run #1



Feature extraction software

Scikit-learn run #1


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.

Reasoning with provenance

W3C Recommendation

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,[])`.

A revision admits the following inference, stating that the two entities linked by a revision are also alternates.

Inference 12 (revision-is-alternate-inference)

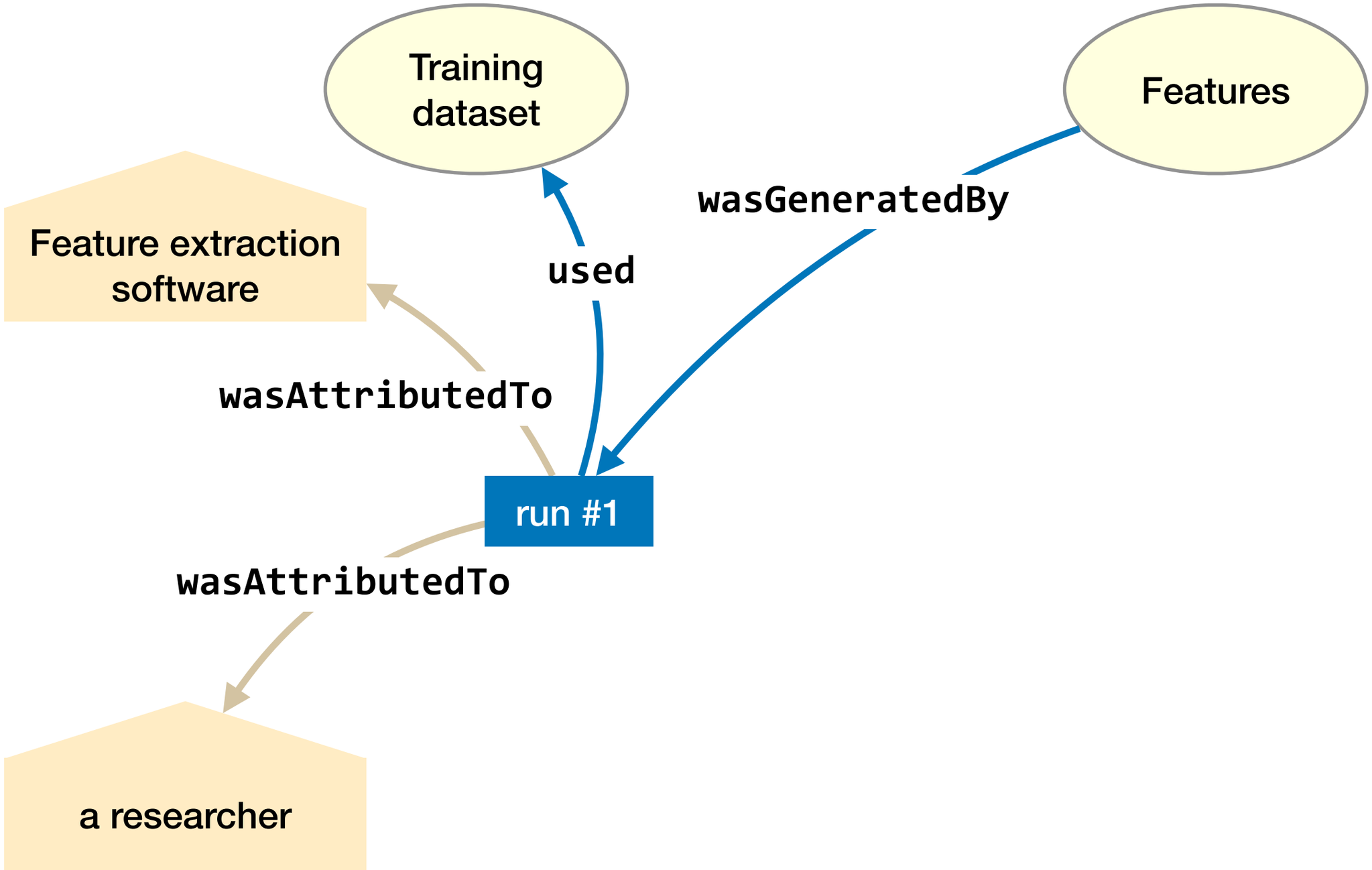
In this inference, any of `_a`, `_g` or `_u` MAY be placeholders.

IF `wasDerivedFrom(_id; e2,e1,_a,_g,_u,[prov:type='prov:Revision'])`, THEN `alternateOf(e2,e1)`.

Remark

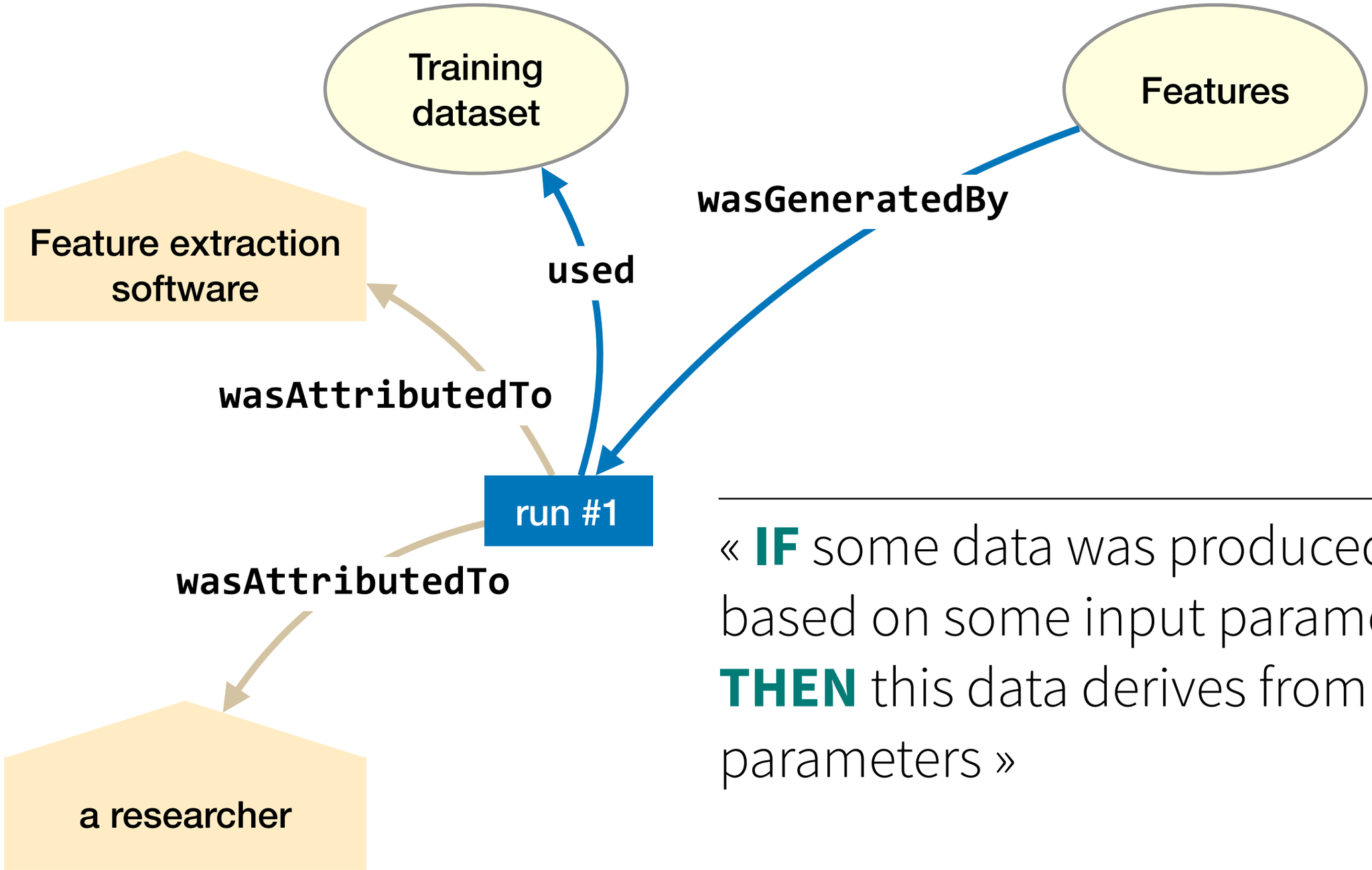
There is no inference stating that `wasDerivedFrom` is transitive.

Reasoning with provenance



Gaignard, A., et al. « **SHARP: Harmonizing and Bridging Cross-Workflow Provenance** ». *The Semantic Web: ESWC 2017 Satellite Events*, édité par Eva Blomqvist et al., vol. 10577, Springer International Publishing, 2017, p. 219-34. *DOI.org (Crossref)*, doi:10.1007/978-3-319-70407-4_35

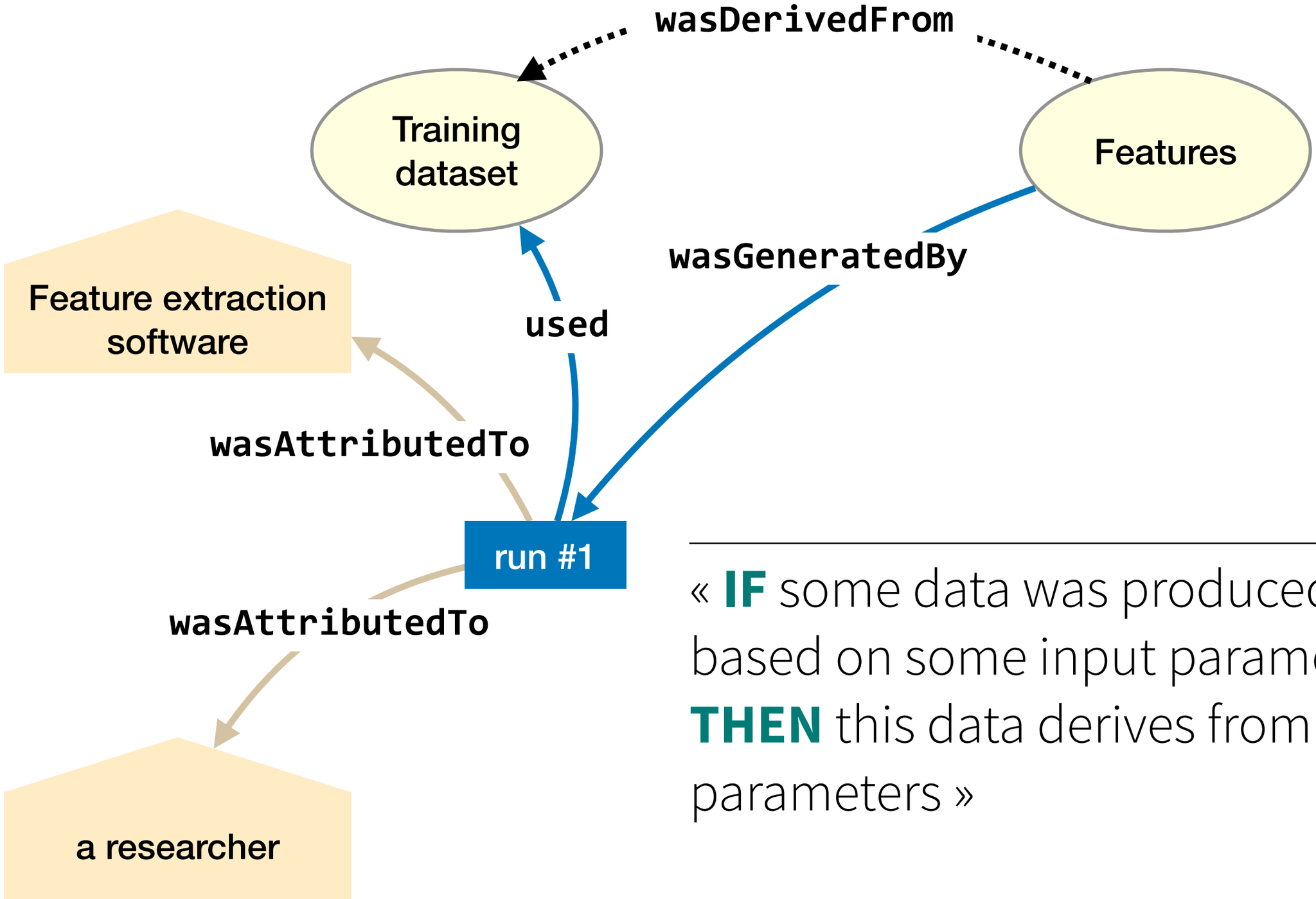
Reasoning with provenance



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

Gaignard, A., et al. « **SHARP: Harmonizing and Bridging Cross-Workflow Provenance** ». *The Semantic Web: ESWC 2017 Satellite Events*, édité par Eva Blomqvist et al., vol. 10577, Springer International Publishing, 2017, p. 219-34. *DOI.org (Crossref)*, doi:10.1007/978-3-319-70407-4_35

Reasoning with provenance



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

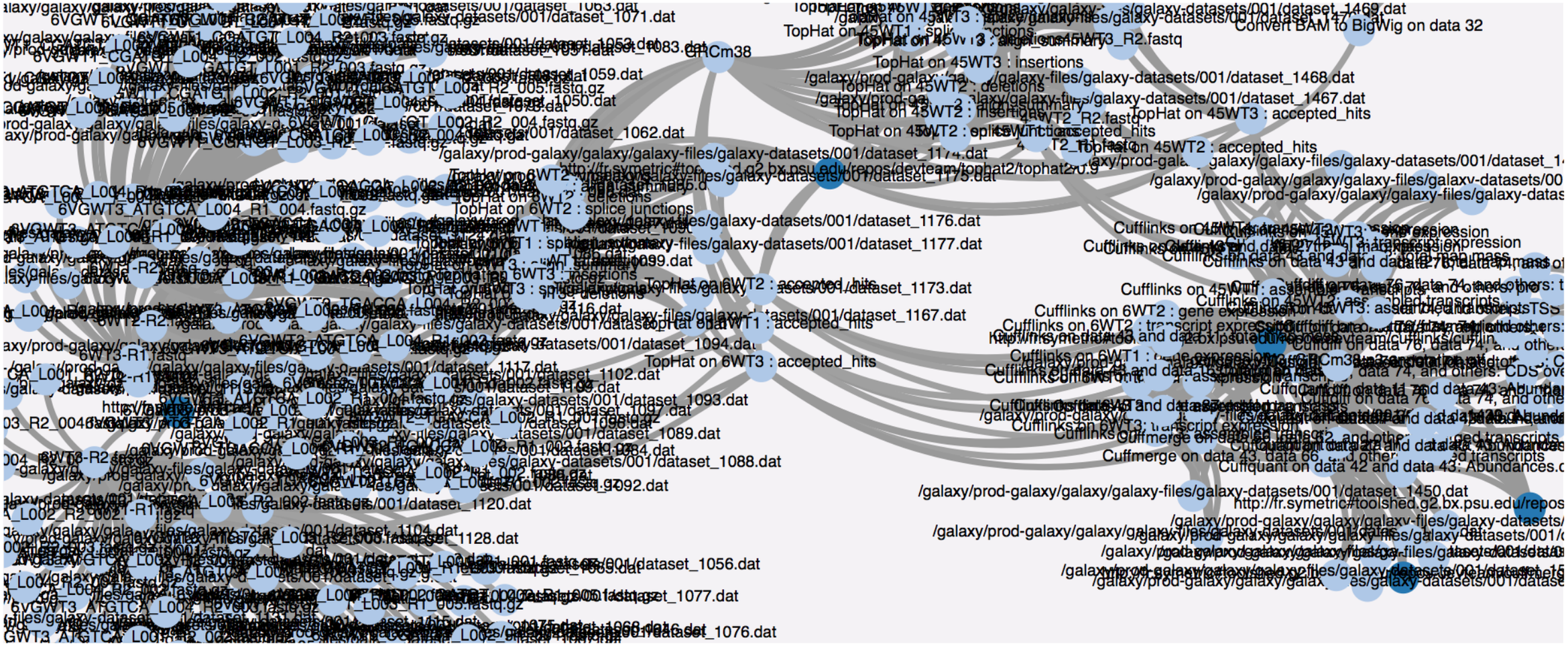
Gaignard, A., et al. « **SHARP: Harmonizing and Bridging Cross-Workflow Provenance** ». *The Semantic Web: ESWC 2017 Satellite Events*, édité par Eva Blomqvist et al., vol. 10577, Springer International Publishing, 2017, p. 219-34. *DOI.org (Crossref)*, doi:10.1007/978-3-319-70407-4_35

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



Is provenance enough for reuse ?

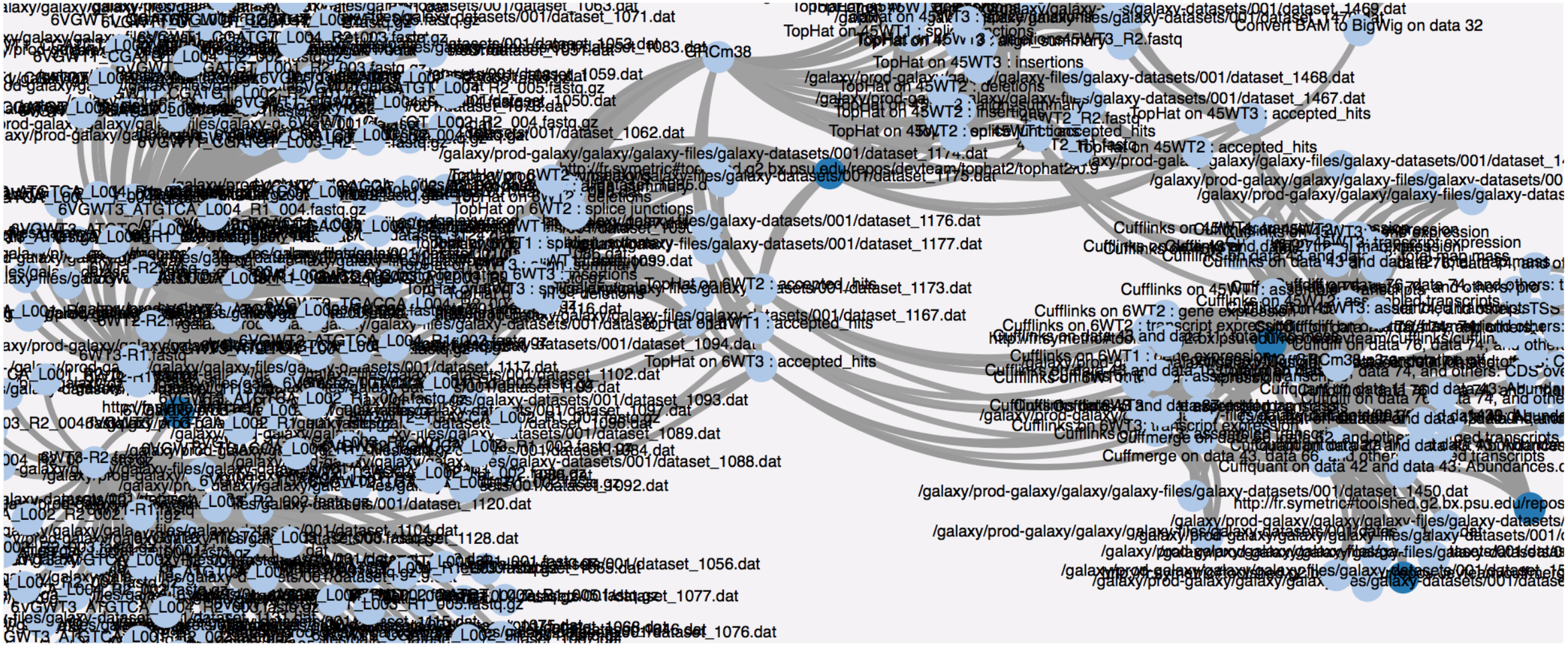
```
   a prov:Bundle, prov:Entity;
   prov:wasAttributedTo <#galaxy2prov>;
   prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
.

<#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

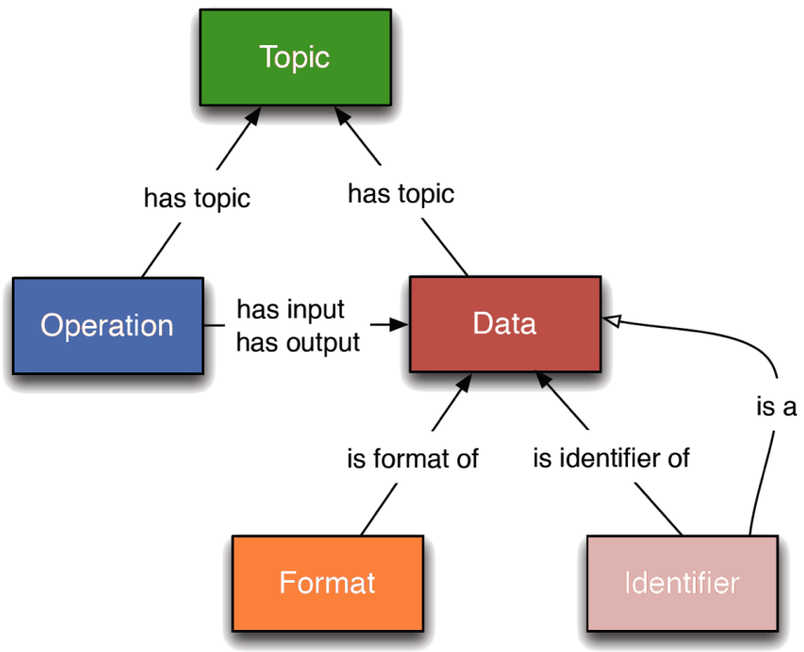
Too fine-grained

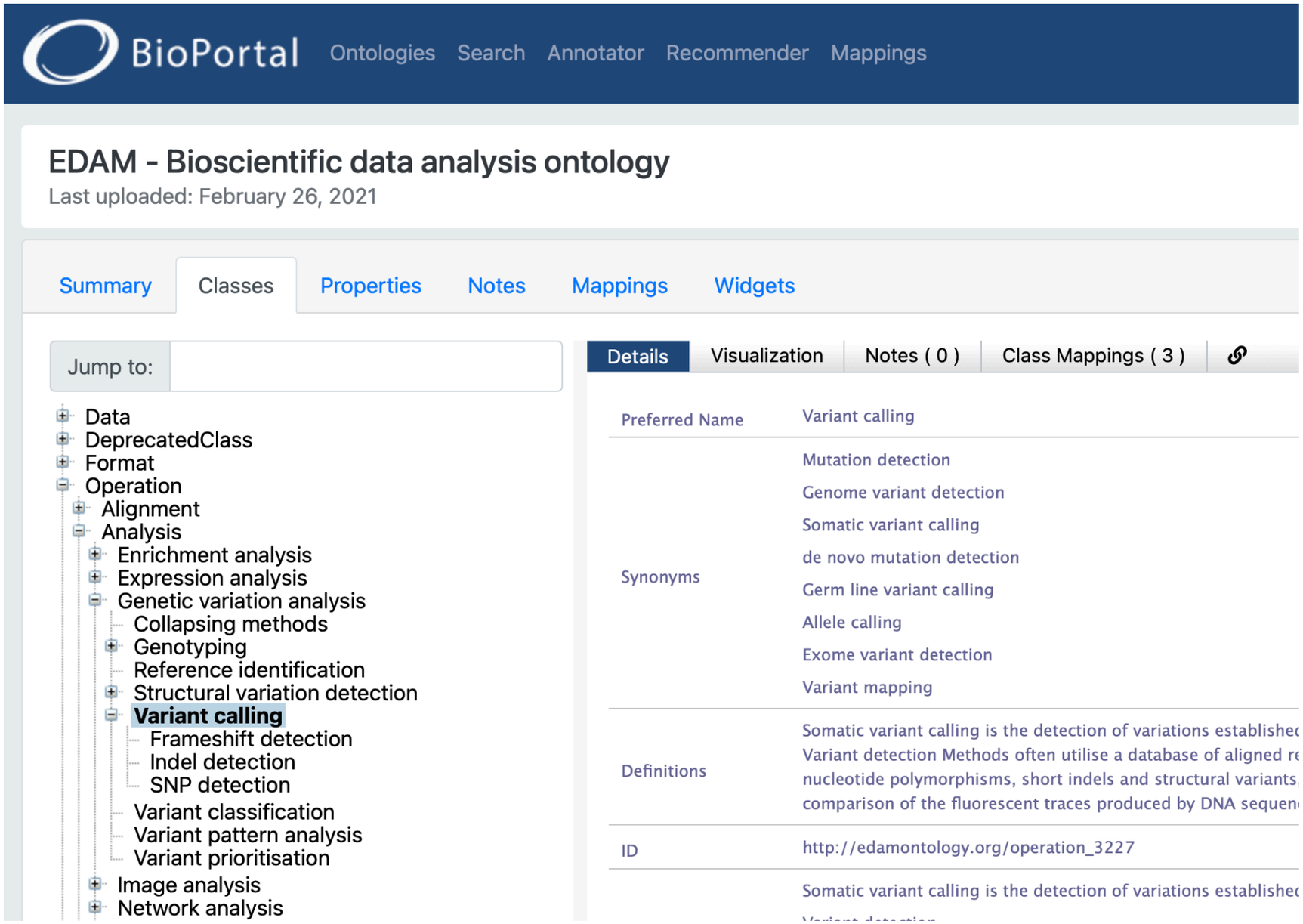
No domain concepts



2. Bioinformatics **domain-specific** concepts

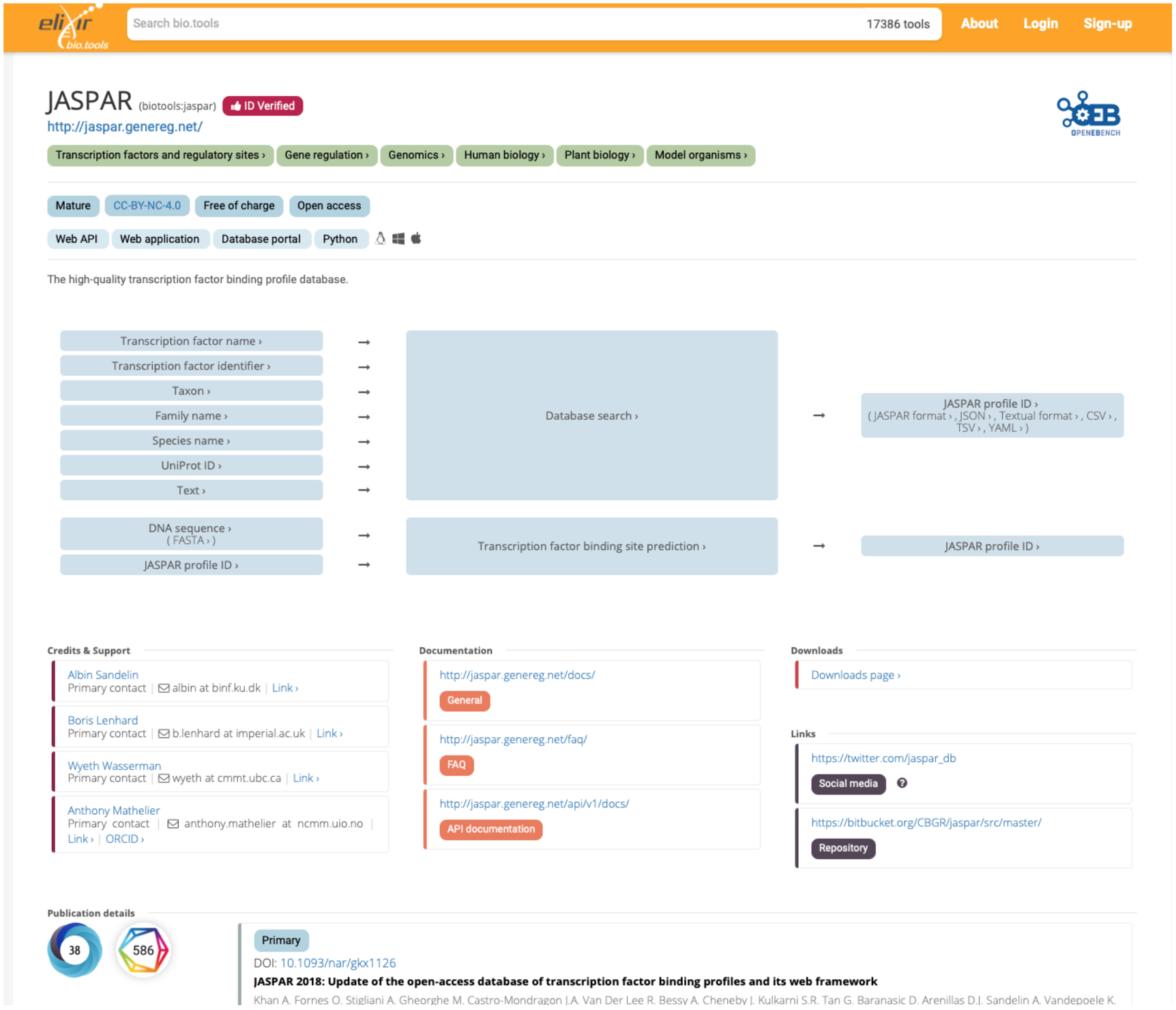
EDAM ontology





Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S. and Rice, P. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics, **29**(10): 1325-1332. DOI: 10.1093/bioinformatics/btt113 PMID: 23479348

bio.tools (semantic) catalog



Annotate workflow data with EDAM concepts ?

3. From knowledge graph to **data summaries**

SPARQL queries to produce data summaries

jupyter FRESH-notebook (autosaved)

Visit repo | Copy Binder link

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

Run | Code | Download | GitHub | Binder | **Memory:** 351.3 MB / 2 GB

2. 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']) ))
```

The file `VCF/hapcaller.snv.recal.filter.vcf.gz.tbi` **results from tool** gatk2_variant_filtration-IP **which** Analyse a genetic variation, for example to annotate its location, alleles, classification, and effects on individual transcripts predicted for a gene model.

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

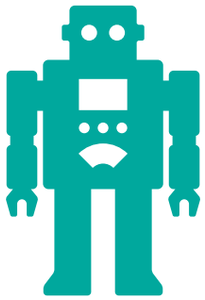
The file `VCF/hapcaller.snv.recal.filter.vcf.gz.tbi` **results from tool** gatk2_variant_filtration-IP **which** Filter a set of files or data items according to some property.

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

launch binder

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

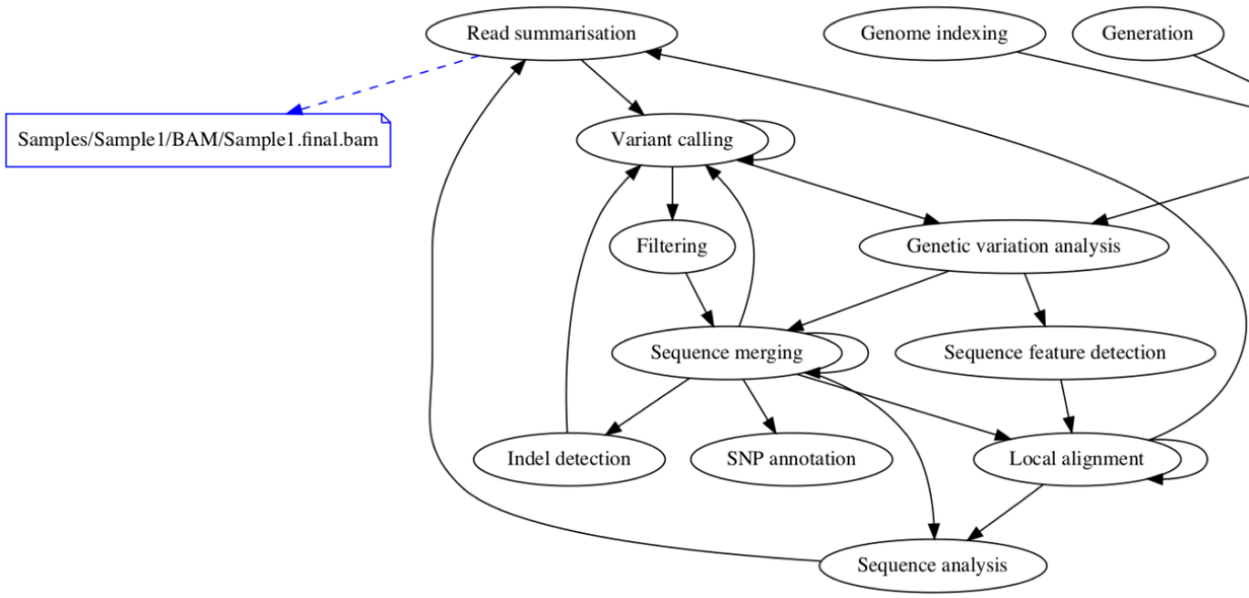
Data summaries

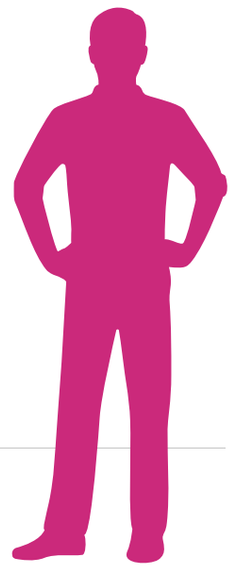


```
[...]
:head {
    _:np1 a np:Nanopublication .
    _:np1 np:hasAssertion :assertion .
    _:np1 np:hasProvenance :provenance .
    _:np1 np:hasPublicationInfo :pubInfo .
}

:assertion {
  <http://snakemake-provenance/Samples/Sample1/
  BAM/Sample1.merged.bai> rdfs:seeAlso
  <http://edamontology.org/operation_3197> .

  <http://snakemake-provenance/VCF/hapcaller.
  indel.recal.filter.vcf.gz> rdfs:seeAlso
  <http://edamontology.org/operation_3695> .
}
[...]
```





```
...
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
...
```

1. It’s possible to automatically produce **machine-oriented nanoPublications**
2. It’s possible to automatically display the **typical bioinformatics tasks** data originate from
3. It’s possible to document data with **text** leveraging ontology definitions (EDAM)

Perspectives

Take home message & perspectives

Scientific Workflows → automation, abstraction, provenance

Standards for **provenance representation** and **reasoning** (PROV-O)

Contributions

- feeding a **knowledge graph** with generic provenance metadata captured at runtime and domain knowledge (EDAM)
- generating **domain-specific** machine and human-oriented **data summaries**
- in line with **F - - R principles** : machine-readable data (F2), community standards (R1.3), provenance (R1.2)

Future works

- distributed data sources → **distributed provenance, reasoning**
- application to ML workflows → better interpret/explain predictions ?
- **evaluation** → large bioinformatics communities through the MuDiS4LS Equipex+ project federating HPC clusters, workflow developers and biologists

Acknowledgments

Hala Skaf-Molli, LS2N, University of Nantes

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

contact: alban.gaignard@univ-nantes.fr