Des annotations de provenance, aux résultats de workflows "Linked Data"

Alban Gaignard

Nantes Academic Hospital
(CHU de Nantes), CNRS,
France

Hala Skaf-Molli

LINA,
Nantes University, CNRS,
France

Audrey Bihouée

Institut du Thorax, Nantes
University, INSERM, CNRS,
France

Travaux publiés à "Theory and Practice of Provenance" (TaPP'16 workshop), Washington DC

CENTRE HOSPITALIER
UNIVERSITAIRE DE NANTES

UNIVERSITÉ DE NANTES

cnrs

Inserm

Needs for linked experiment reports

Motivations: reusing (massive) RNA-seq data

TopHat: algorithm to align multiple sequence reads to a reference genome (known genes).

Motivations: reusing (massive) RNA-seq data

TopHat: algorithm to align multiple sequence reads to a reference genome (known genes).

	1 sample
Input data	2 x 17 Gb
1-core CPU	170 hours
32-cores CPU	32 hours
Output data	12 Gb

Motivations: reusing (massive) RNA-seq data

TopHat: algorithm to align multiple sequence reads to a reference genome (known genes).

	1 sample	300 samples
Input data	2 x 17 Gb	10.2 Tb
1-core CPU	170 hours	5.9 years
32-cores CPU	32 hours	14 months
Output data	12 Gb	3.6 Tb

Motivations: reusing (massive) RNA-seq data

TopHat: algorithm to align multiple sequence reads to a reference genome (known genes).

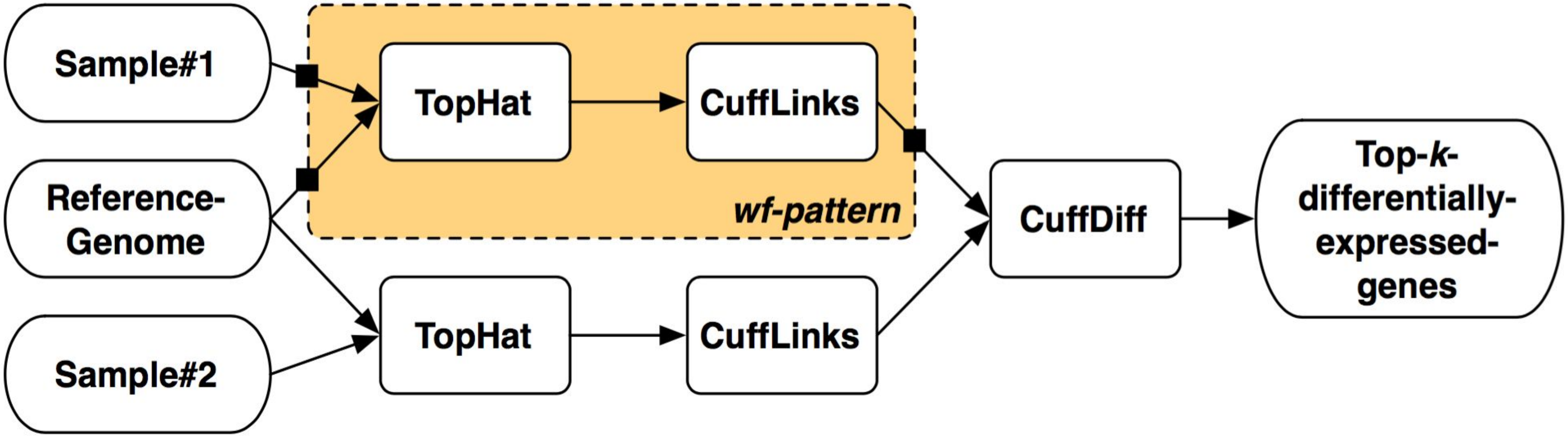
	1 sample	**300 samples**
Input data	2 x 17 Gb	10.2 Tb
1-core CPU	170 hours	5.9 years
32-cores CPU	32 hours	14 months
Output data	12 Gb	3.6 Tb

Challenges

Algorithmic performance, storage, preservation, **reuse (limit recompute) & share**.

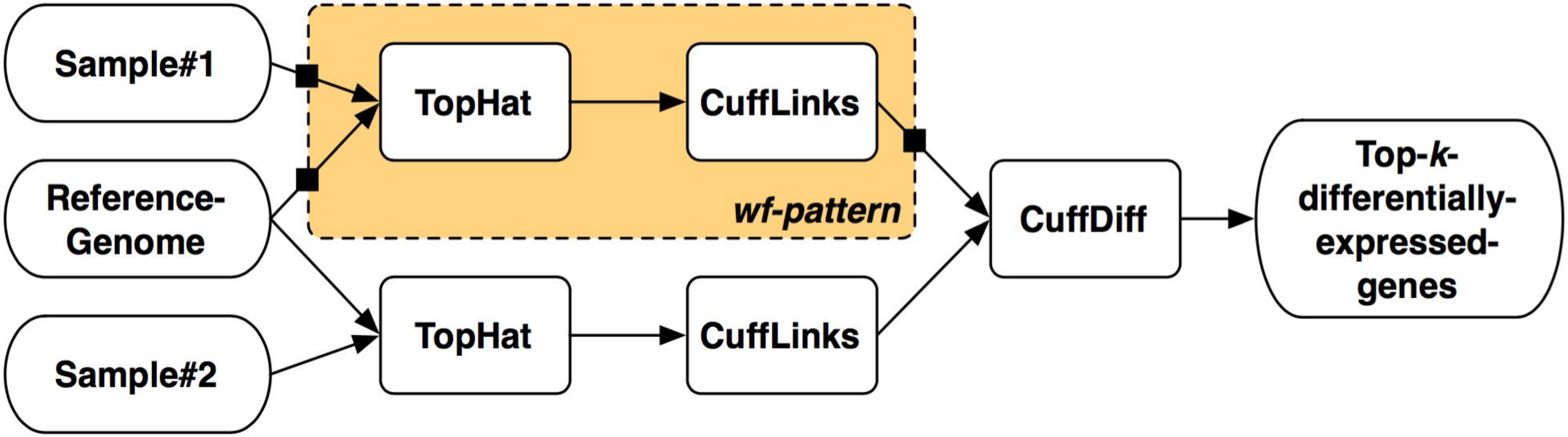
Motivations: reusing experiment results

Scientific experiment: RNA sequencing to quantify gene expression levels under multiple biological conditions.



Motivations: reusing experiment results

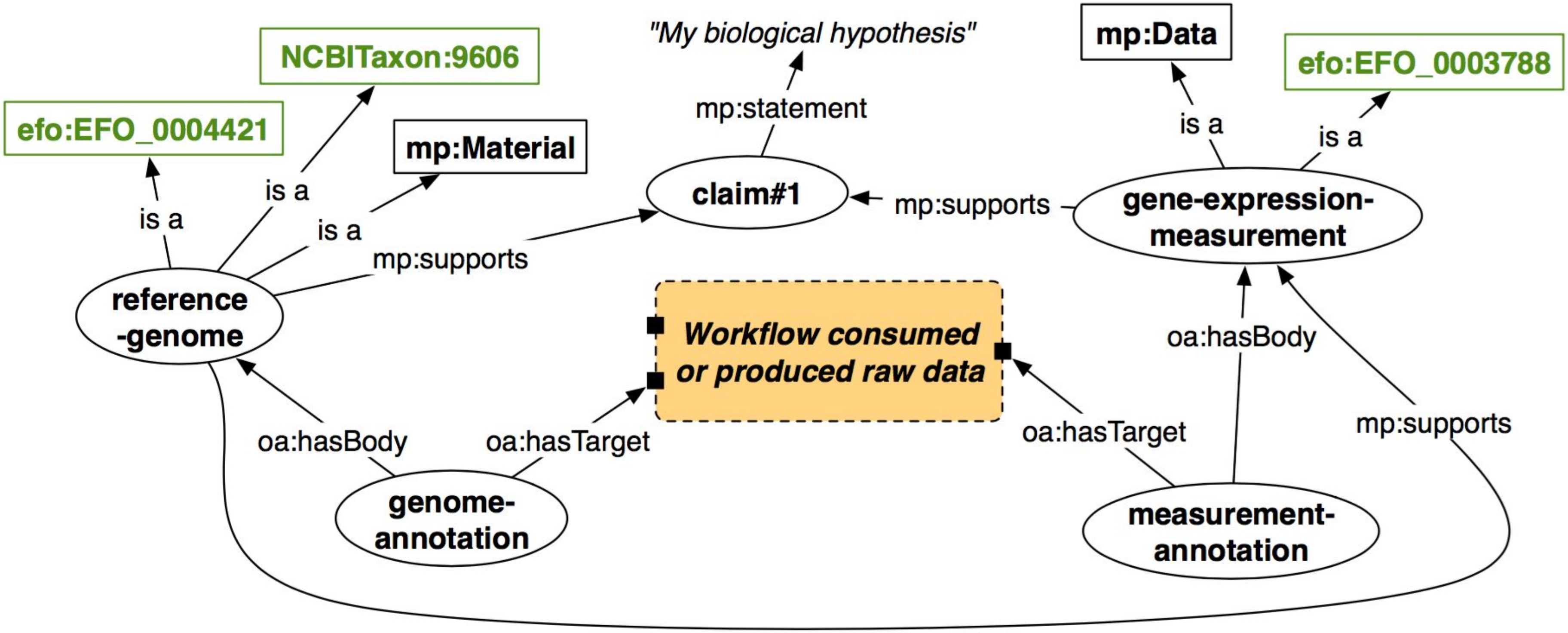
Scientific experiment: RNA sequencing to quantify gene expression levels under multiple biological conditions.



Interpreting and sharing results :

need for scientific context (metadata)

Expected result: human+machine tractable reports



Annotated “Material & Methods”

Links to **some** workflow artifacts (algorithms, data)

5-star Linked Open Data

W3C standards for **machine** and **human** readable data on the web.

★★★★★ : time and expertise !

5-star Linked Open Data

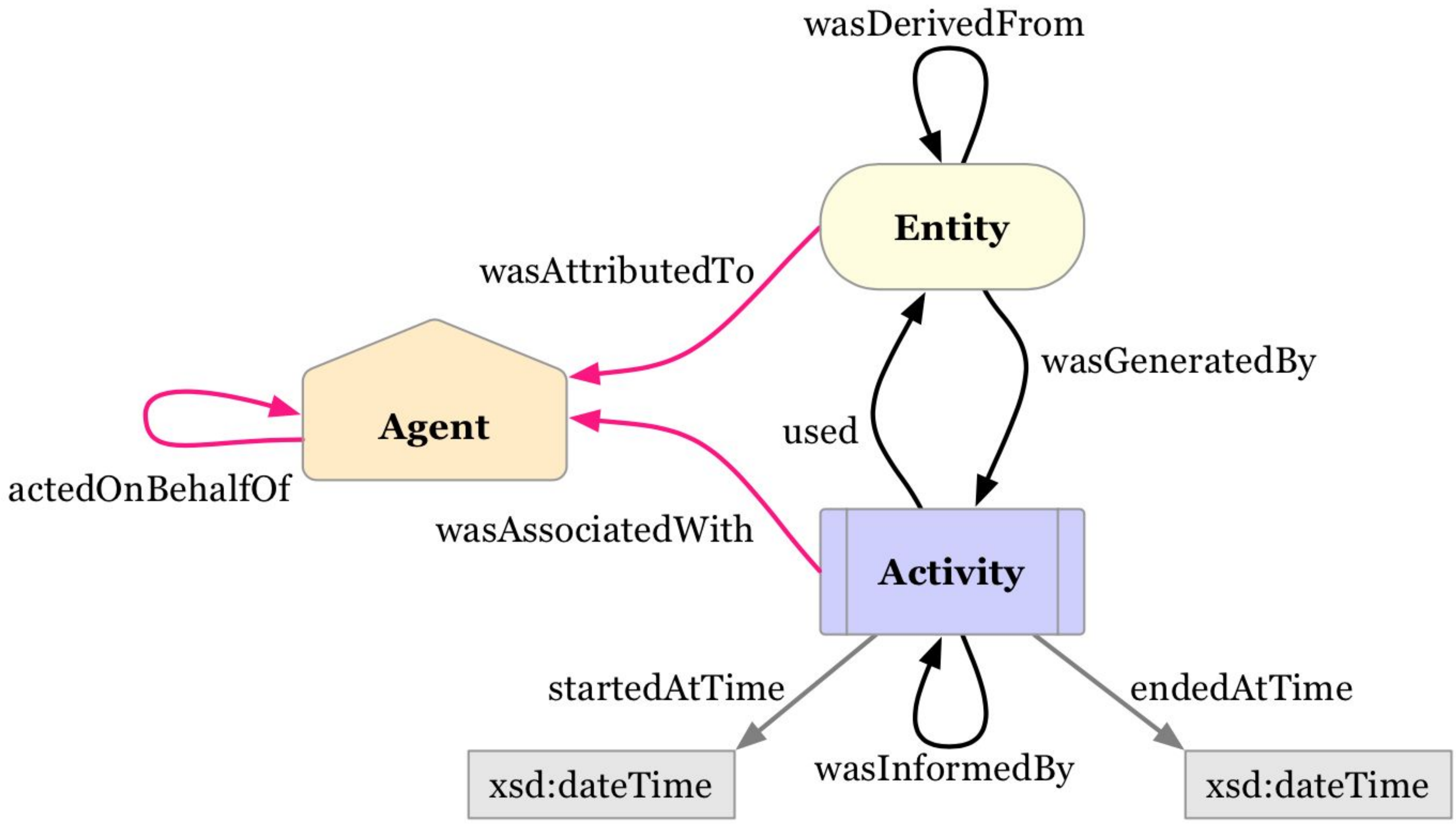
W3C standards for **machine** and **human** readable data on the web.

★★★★★ : time and expertise !

How to ease this process ?

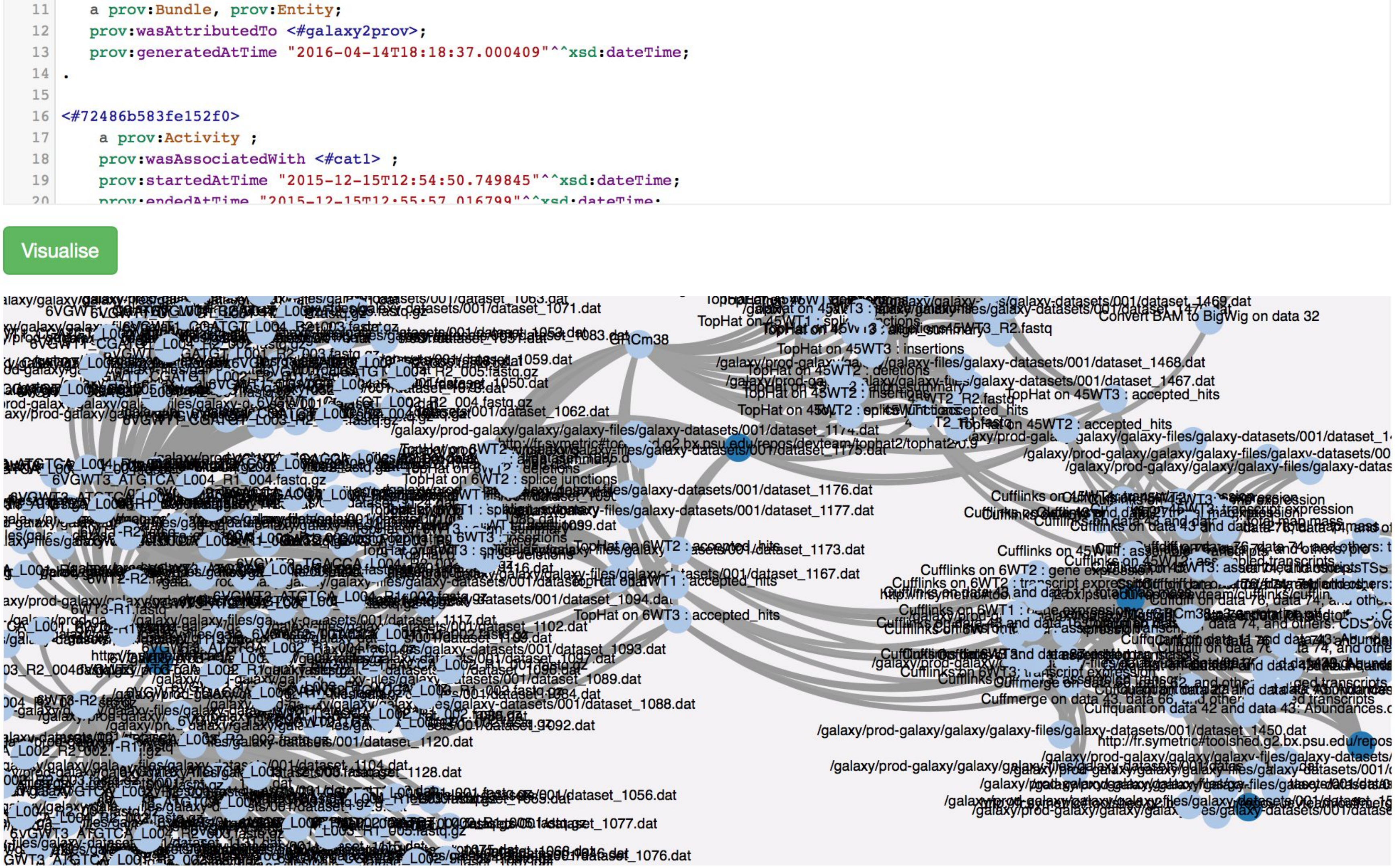
- Workflow engines ⟶ **automation**
- PROV ⟶ **workflow runs as linked data**

W3C Recommendation : the PROV Ontology



https://www.w3.org/TR/prov-o/

PROV traces for a WF run



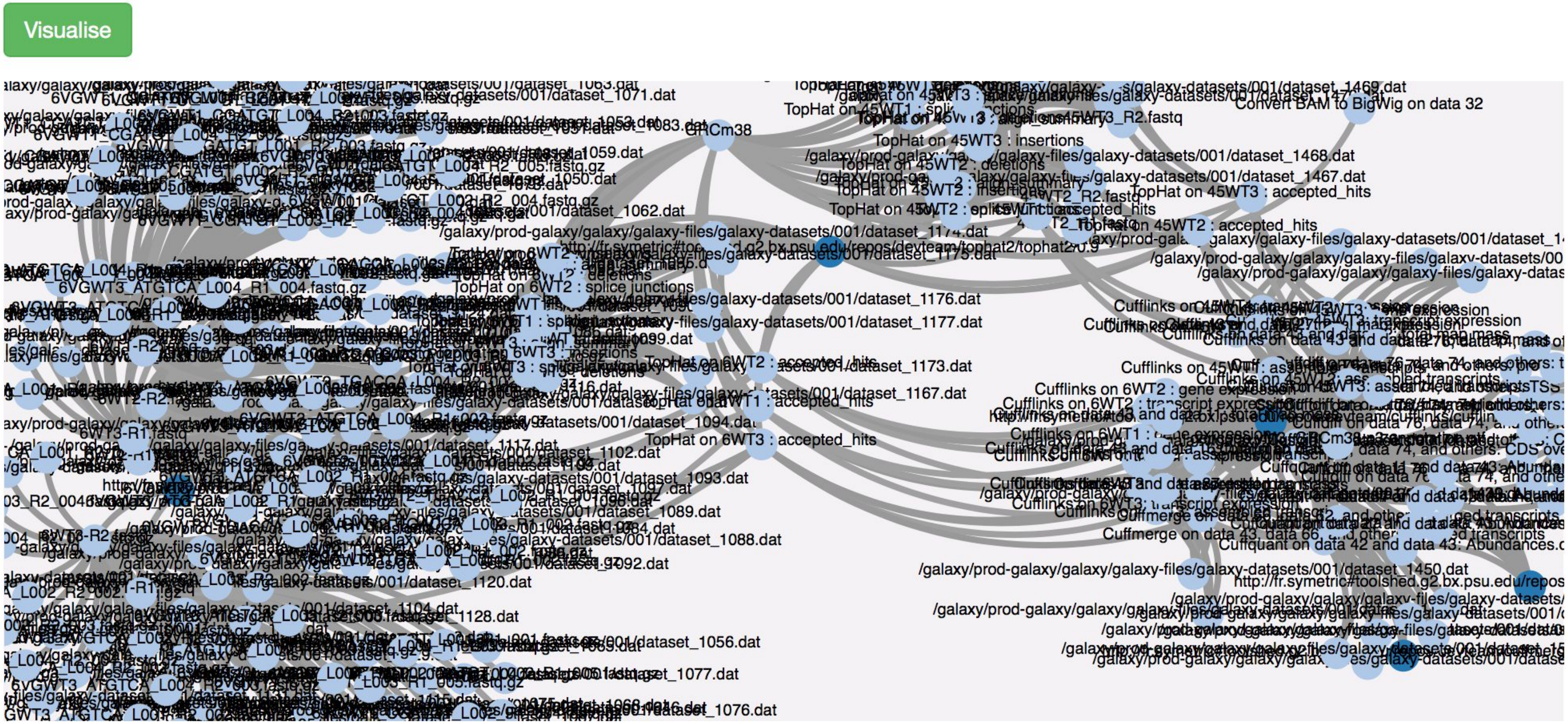
```
    a prov:Bundle, prov:Entity;
    prov:wasAttributedTo <#galaxy2prov>;
    prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
.

<#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;
```

PROV traces for a WF run

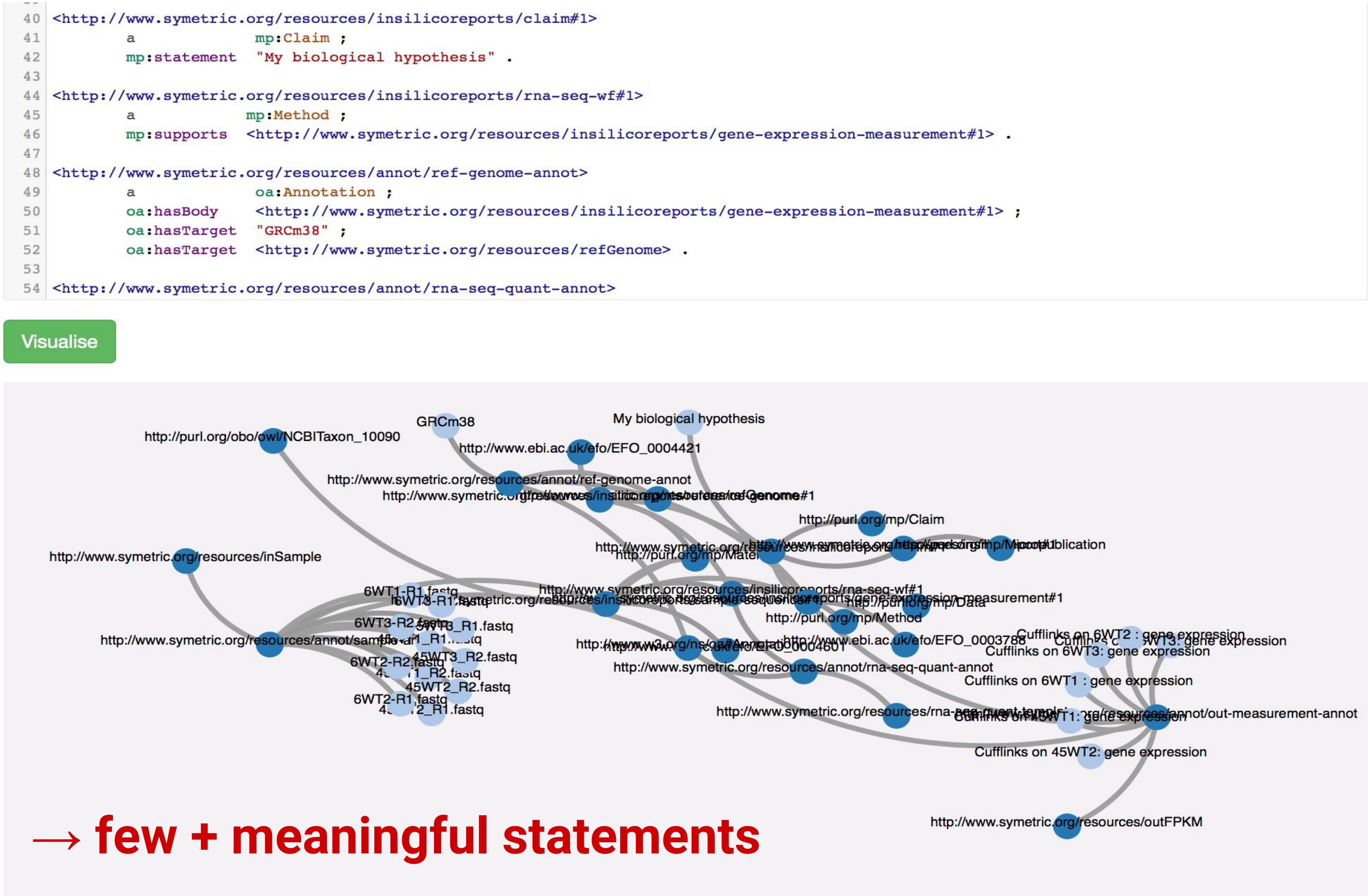
```
    a prov:Bundle, prov:Entity;
    prov:wasAttributedTo <#galaxy2prov>;
    prov:generatedAtTime "2016-04-14T18:18:37.000409"^^xsd:dateTime;
.

<#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;
```

granularity ?

domain concepts ?



Provenance as a Linked Experiment Report



```
<http://www.symetric.org/resources/insilicoreports/claim#1>
        a               mp:Claim ;
        mp:statement  "My biological hypothesis" .

<http://www.symetric.org/resources/insilicoreports/rna-seq-wf#1>
        a              mp:Method ;
        mp:supports   <http://www.symetric.org/resources/insilicoreports/gene-expression-measurement#1> .

<http://www.symetric.org/resources/annot/ref-genome-annot>
        a               oa:Annotation ;
        oa:hasBody      <http://www.symetric.org/resources/insilicoreports/gene-expression-measurement#1> ;
        oa:hasTarget  "GRCm38" ;
        oa:hasTarget  <http://www.symetric.org/resources/refGenome> .

<http://www.symetric.org/resources/annot/rna-seq-quant-annot>
```

→ **few + meaningful statements**

Problem statement & objectives

Problem statement

Scientific workflows produce massive raw results. Their publication into curated query-able linked data repositories requires lot of time and expertise.

Can we exploit provenance traces to ease the publication of scientific results as Linked Data ?

Problem statement & objectives

Problem statement

Scientific workflows produce massive raw results. Their publication into curated query-able linked data repositories requires lot of time and expertise.

Can we exploit provenance traces to ease the publication of scientific results as Linked Data ?

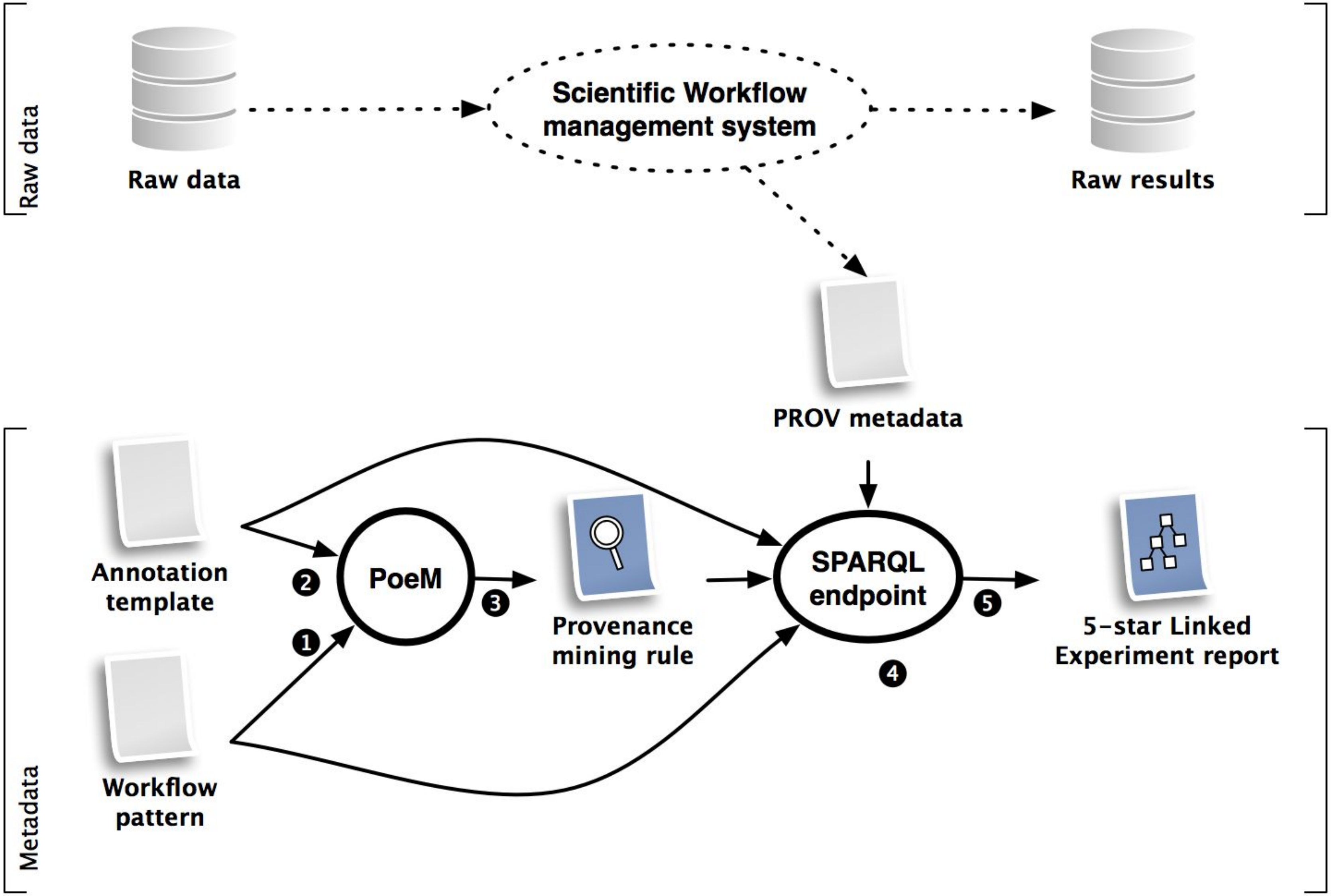
Objectives

(1) Leverage **annotated workflow patterns** to generate **provenance mining rules**.

(2) Refine **provenance** traces into **linked experiment reports**.

Rules generation

Approach


Raw data
Raw data
Scientific Workflow management system
Raw results
PROV metadata
Metadata
Annotation template
Workflow pattern
❷
❶
PoeM
❸
Provenance mining rule
SPARQL endpoint
❹
❺
5-star Linked Experiment report

Input domain-specific annotations (➊,➋)

Workflow patterns ➊

Sequence patterns, with possibly intermediate steps

- P-PLAN ontology: *Step, Variable, hasInputVar, hasOutputVar*
- EDAM ontology: *hasFunction, RNA sequence, Genome map*

Input domain-specific annotations (❶,❷)

Workflow patterns ❶

Sequence patterns, with possibly intermediate steps

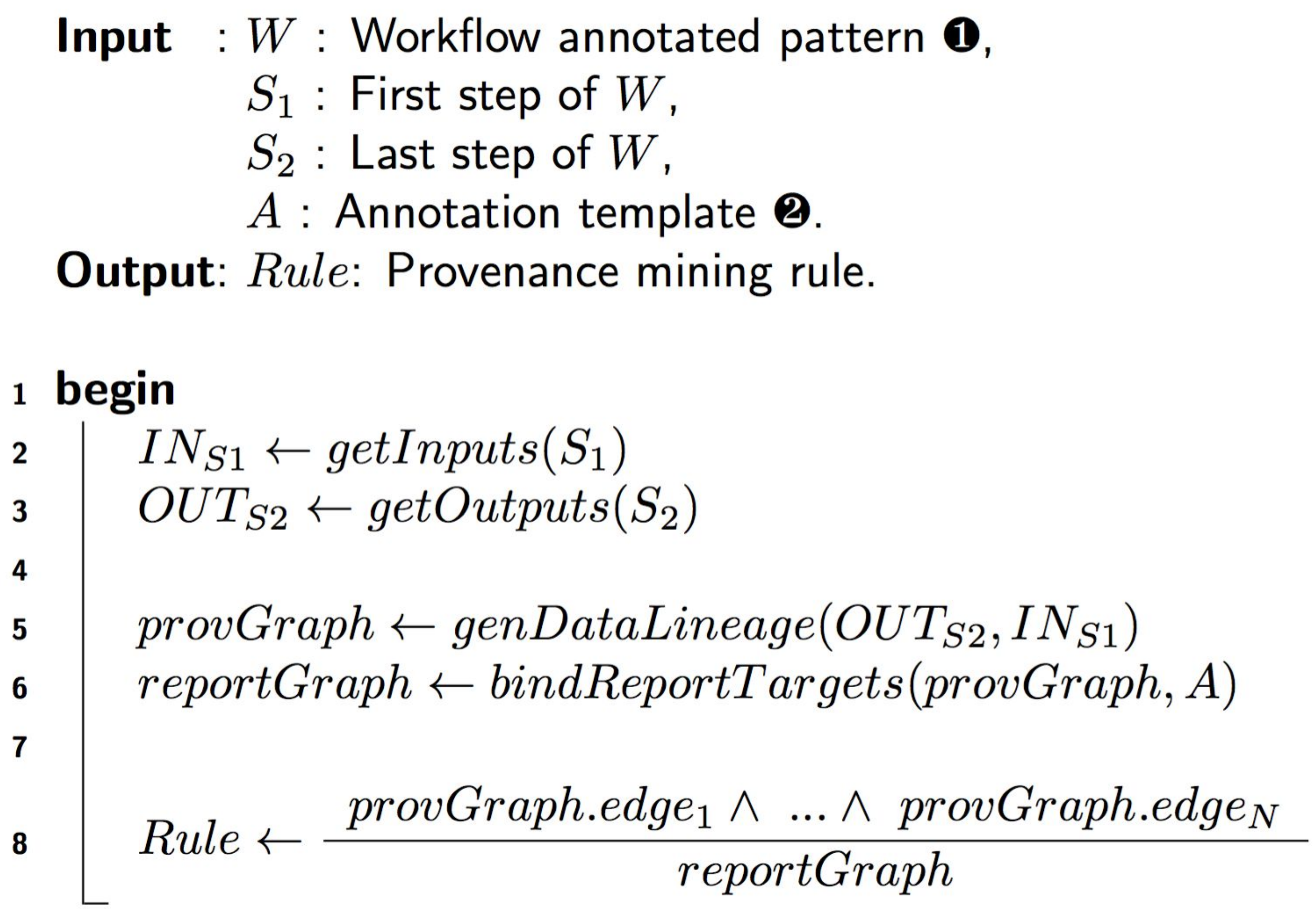
- P-PLAN ontology: *Step, Variable, hasInputVar, hasOutputVar*
- EDAM ontology: *hasFunction, RNA sequence, Genome map*

Experiment report template ❷

Link scientific claims, statements, material and methods

- MicroPublication ontology: *Material, Method, Claims*
- Experimental factor ontology: *Transcriptome, Gene expression*
- NCBI taxonomy: *Homo Sapiens*
- Open Annotation model: *hasBody, hasTarget*

PoeM: generating PrOvEnance Mining rules ❸



```
Input  : $W$ : Workflow annotated pattern ❶,
         $S_1$ : First step of $W$,
         $S_2$ : Last step of $W$,
         $A$ : Annotation template ❷.
Output: $Rule$: Provenance mining rule.

begin
    $IN_{S1} \leftarrow getInputs(S_1)$
    $OUT_{S2} \leftarrow getOutputs(S_2)$

    $provGraph \leftarrow genDataLineage(OUT_{S2}, IN_{S1})$                (SPARQL Property Path)
    $reportGraph \leftarrow bindReportTargets(provGraph, A)$                 (SPARQL Basic Graph Pattern)

    $Rule \leftarrow \frac{provGraph.edge_1 \wedge \ldots \wedge provGraph.edge_N}{reportGraph}$    (SPARQL Construct query)
```

PoeM: sample generated rule ❸

<Then> part

<If> part

```
PREFIX ...

CONSTRUCT {

        <ht     ./insilicoreport#1> rdf:type <http://purl.org/mp/Micropublication> .
# ...
                 ./out-measurement-annot> rdf:type <http://www.w3.org/ns/oa#Annotation> .
                /.../out-measurement-annot> oa:hasTarget ?lout0 .
             p://.../out-measurement-annot> oa:hasBody <http://.../gene-expression-measurement#1> .
         ttp://.../gene-expression-measurement#1> rdf:type <http://purl.org/mp/Data> .
        <http://.../gene-expression-measurement#1> rdf:type <http://www.ebi.ac.uk/efo/EFO_0003788> .
   ..

} WHERE {

# ...
        ?out  1 prov:wasGeneratedBy ?step1 .
        ?i     rdfs:label | ^rdfs:label | prov:wasDerivedFrom)* ?outS0 .

                rov:used ?in2 .
                 prov:wasAssociatedWith ?soft2 .
          anStep2 a p-plan:Step .
        ?planStep2 rdfs:label ?planStep2Label .
        FILTER (contains(lcase(str(?soft2)),lcase(str(?planStep2Label)))) .

        ?out0 prov:wasGeneratedBy ?step2 .
        ?out0 rdfs:label ?lout0 .
        FILTER (contains(lcase(str(?lout0)), lcase("gene expression"))) .
}
```

First experiments & results

Experiment

Material & methods

- Real-life RNA-seq workflow to study 3 mice populations
- WF implemented in Galaxy, run on 2 biological samples
- PROV traces exported from Galaxy Histories (API)

Experiment

Material & methods

- Real-life RNA-seq workflow to study 3 mice populations
- WF implemented in Galaxy, run on 2 biological samples
- PROV traces exported from Galaxy Histories (API)

Results (for 1 biological sample)

- 60h CPU (12 cores for genome alignment), 21Gb storage
- 3s to export 81 PROV triples from the Galaxy history
- 2s to apply the rule and produce 35 Micropublication triples

Results (for 4 biological samples)

- 1232 PROV triples from the Galaxy history
- 49 Micropublication triples (4%)

Conclusion & perspectives

Semi-automated approach

(1) PoeM generates **semantic web rules**

(2) PoeM rules applied on **PROV** traces to assemble **linked experiment reports** (MicroPublication)

Limitations:

- Sequence workflow patterns only
- *SPARQL property paths* with complex WF patterns ?
- Syntactic matching between WF patterns and PROV labels

Usage scenarios:

→ **Query** workflow datasets with domain concepts

→ **Populate** RDF repositories with WF results

Conclusion & perspectives

Future works

(1) **WF patterns**: split-merge, “common motifs”

(2) **Genericity**: other domains / other reports (RO, Nanopub.)

(3) **PROV heterogeneity**: multi-systems PROV reconciliation

(4) **Evaluation**: involving biologists, at larger scale

Questions ?

Demo: http://poem.univ-nantes.fr

Contact: alban.gaignard@univ-nantes.fr

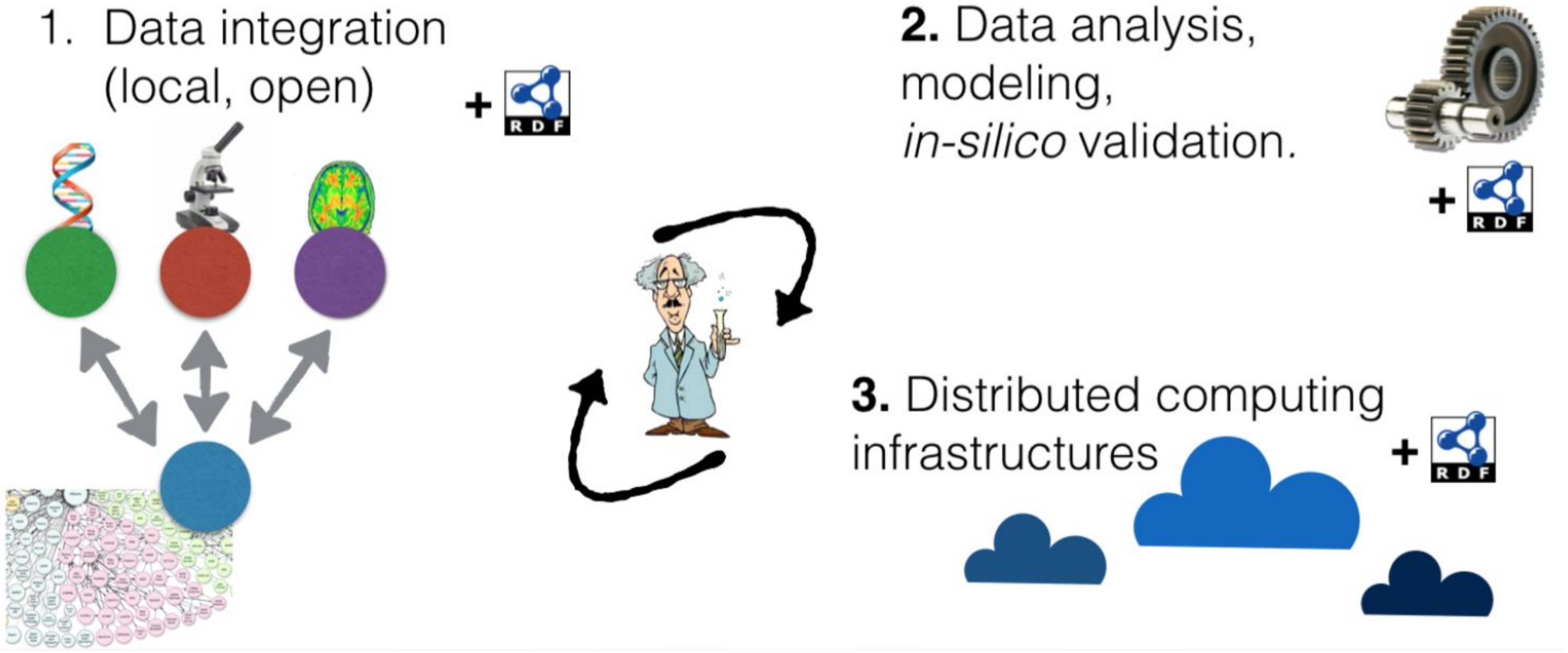
Acknowledgments

BiRD bioinformatics facility

Connect Talent Call

Experiment context

SyMeTRIC: **systems medicine** project (2015-2017, call “Connect Talent”), funded by the french region Pays de la Loire.



1. Data integration (local, open) + RDF

2. Data analysis, modeling, *in-silico* validation. + RDF

3. Distributed computing infrastructures + RDF