

# **Reproducible research**

**Dr. Tomasz Miksa**

TU Wien & SBA Research

[tmiksa@sba-research.org](mailto:tmiksa@sba-research.org)

# Agenda

- eScience
- Reproducibility challenges
- Auditability and accountability
- Improving reproducibility
- Process preservation
- Summary

**DATA MANAGEMENT**

**REPRODUCIBILITY**

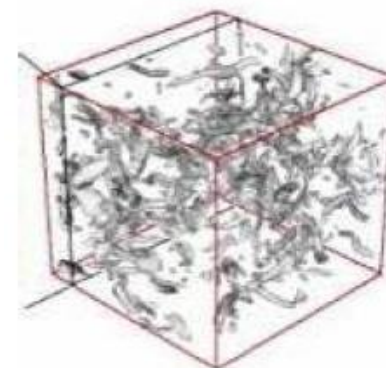
# ESCIENCE

# The Fourth Paradigm

- Empirical Science
  - Description of observed phenomena
- Theoretical Science
  - Model building, generalization
- Computational Science
  - Simulation of complex phenomena
- eScience

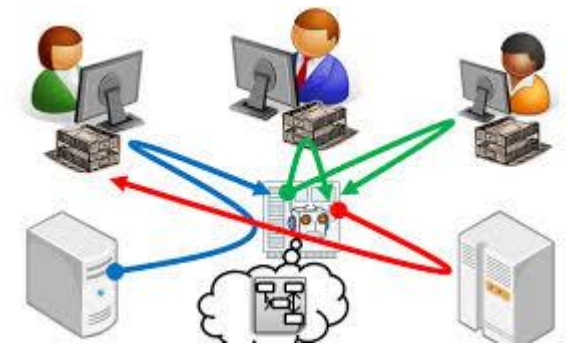


$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



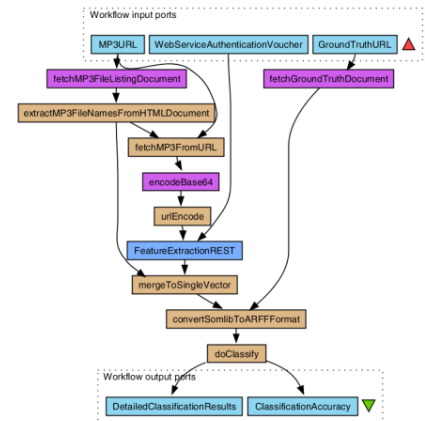
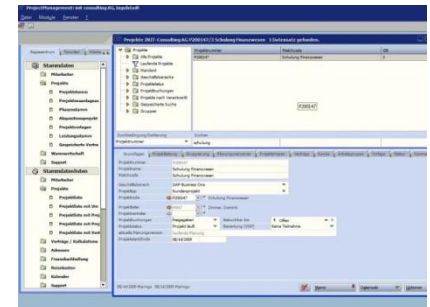
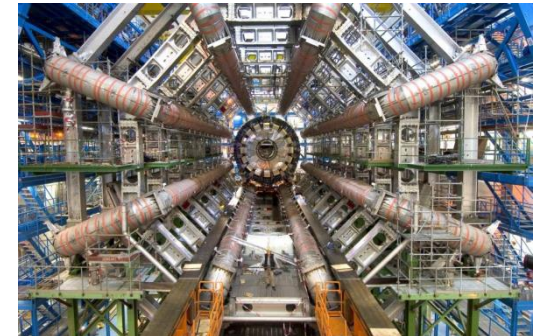
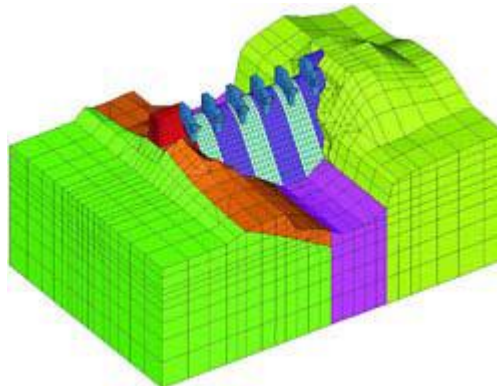
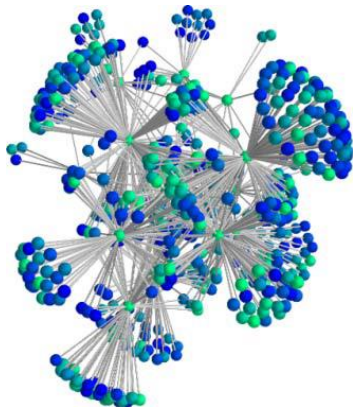
# eScience and Research Infrastructures

- Scientists exchange
  - facilities
  - resources
  - services
  - datasets
- Research requires
  - special tooling and software
  - processes to
    - transform
    - visualize
    - interpret the data



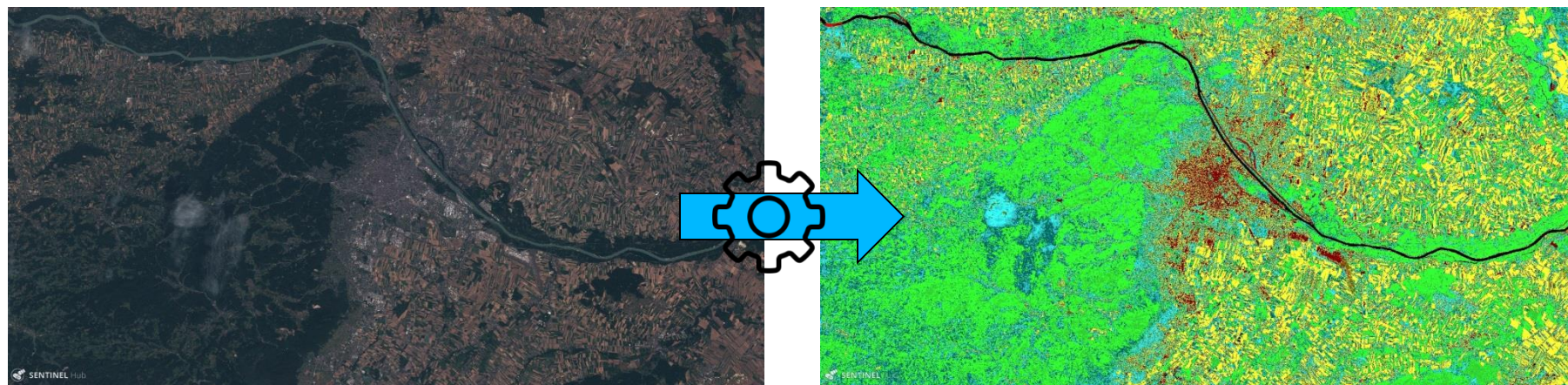
# Related domains

- E-Health
- Civil engineering
- Finance
- Insurance
- Aviation industry
- More examples – see the DMP lecture intro



# e-Science example – Earth Observation

- Measuring vegetation and monitoring fluctuations
- Normalized Difference Vegetation Index (NDVI)
  - calculated from the visible and near-infrared light reflected by vegetation



source: <https://www.sentinel-hub.com/explore/sentinel-playground>



# The Fourth Paradigm

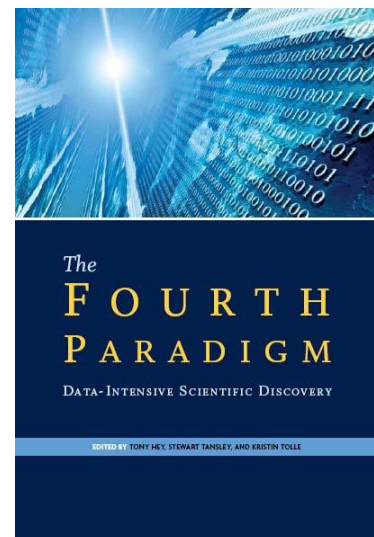
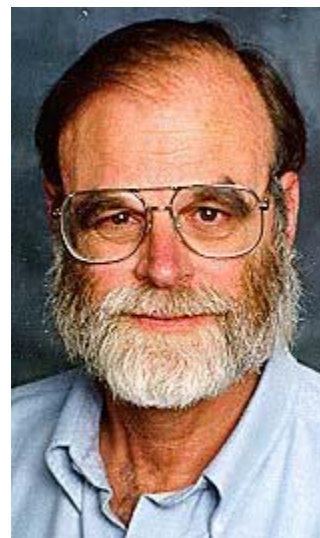
- **The Fourth Paradigm**

Tony Hey, Stewart Tansley, and Kristin Tolle (Eds.),  
Oct. 2009, Microsoft Research

<http://research.microsoft.com/en-us/collaboration/fourthparadigm/>

- **Jim Grey (1944-2007)**

Turing Award Winner 1998

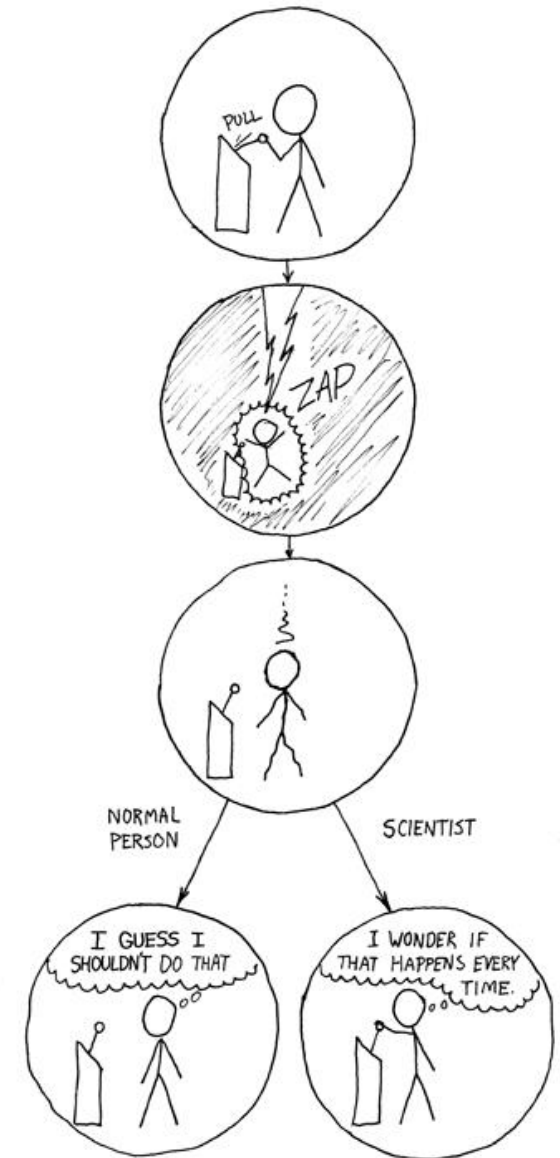




# REPRODUCIBILITY

# Reproducibility

- Core to the scientific method
- Focus
  - NOT on detecting misconduct
  - but on the will to produce good work



<https://xkcd.com/242/>

# Reproducibility

- Related terms often used interchangeably (which is not correct)

Term	Environment	Researcher	Complete workflow	V&V of workflow re-execution
replicate	same	different	yes	yes
repeat	same	same	yes	yes
reproduce	different	same/different	yes	NA
re-execute	same/different	same/different	yes	no
rerun	same/different	same/different	yes	no
reuse	same/different	same/different	no	(yes)

Check 'Terminology' section: <https://www.acm.org/publications/policies/artifact-review-badging>

# Reproducibility challenges – simple example

- Image conversion from *jpg* to *tiff* using *ImageMagick*

	<i>View Path #1</i>	<i>View Path #2</i>
<b>Data formats</b>	Raw JPEG Stream (fmt/41);Portable Network Graphics (fmt/13)	Raw JPEG Stream (fmt/41);Portable Network Graphics (fmt/13)
<b>Application</b>	ImageMagick 6.8.9-7 Q16 Microsoft Visual C++ 2010	ImageMagick 6.8.9-7
<b>JVM</b>	Java SE 6 Update 45	Java SE 7 Update 10
<b>Operating System</b>	Windows 7 Enterprise SP1	OS X 10.9.4
<b>Hardware</b>	3,3GHz Intel Core i3 8GB 1600MHz DDR3 NVIDIA GT630 2GB	2,3GHz Intel Core i5 4GB 1333MHz DDR3 Intel HD Graphics 3000 384MB

# Reproducibility challenges – simple example



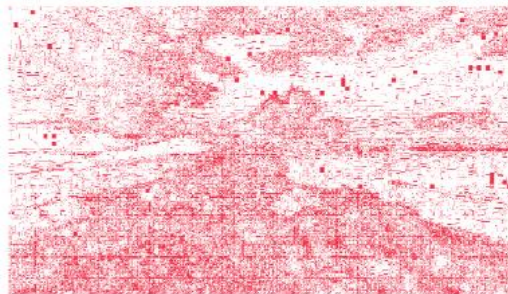
Original jpg



TIFF  
Migration on Windows7

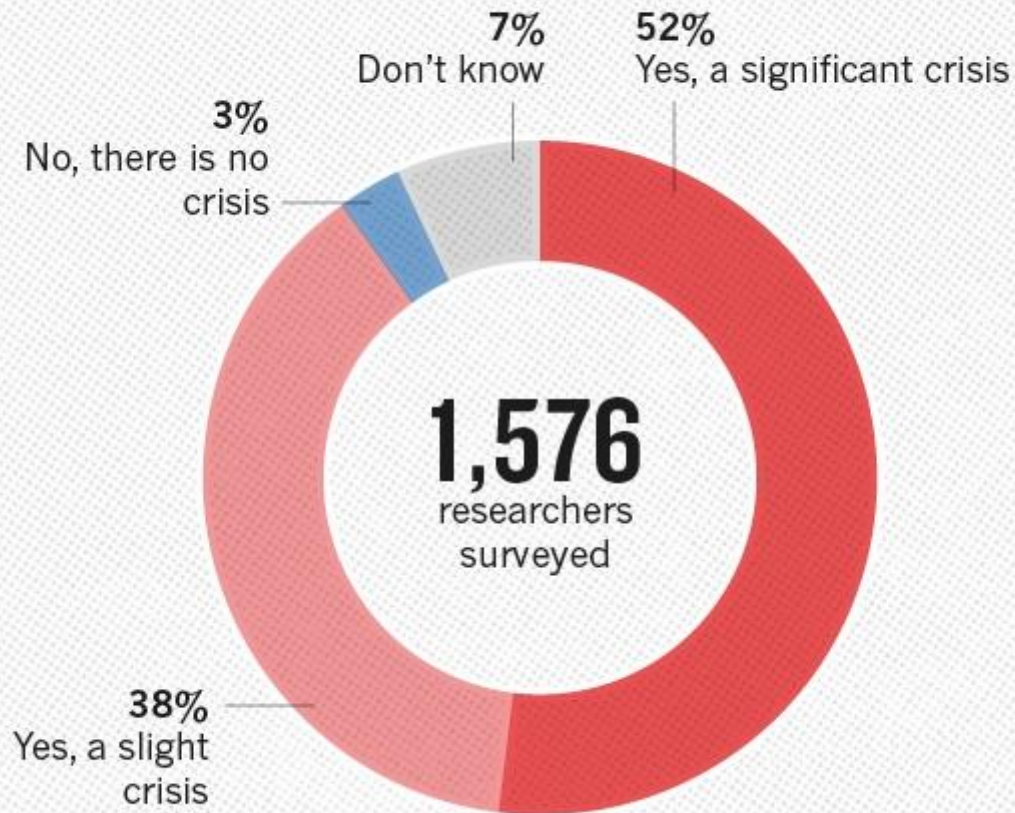


TIFF  
Migration on OSX



Diff

## IS THERE A REPRODUCIBILITY CRISIS?



©nature

[https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970?WT.mc\\_id=FBK\\_NatureNews](https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970?WT.mc_id=FBK_NatureNews)



## ■ FreeSurfer Software

- cortical thickness and volume of neuroanatomical structures

## ■ Different

### - FreeSurfer Versions

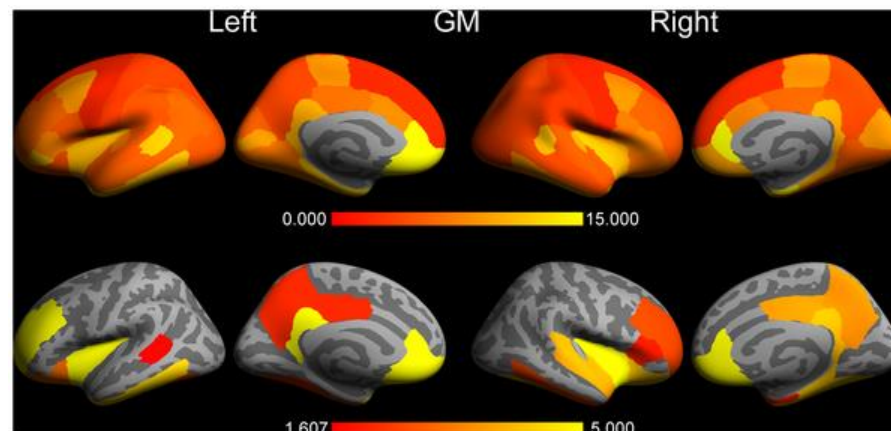
- v4.3.1, v4.5.0, v5.0.0

### - Workstation

- Mac, Hewlett-Packard

### - Operating system version

- OSX 10.5, OSX 10.6



E. Gronenschild, P. Habets, H. I. L. Jacobs, R. Mengelers, N. Rozendaal, J. van Os, and M. Marcelis, "The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements," 2012.

# Reproducibility in Machine Learning

- 30 research papers analysed
- Papers lack information on
  - datasets used
  - parameters (randomisation)
  - software environment
- Code is usually not provided
  - only generic method explanation

Submitted to Reproducibility in ML Workshop at the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia.

<https://openreview.net/pdf?id=By4l2PbQ->

## Reproducibility in Machine Learning-Based Studies: An Example of Text Mining

**Babatunde K. Olorisade**  
School of Computing and Mathematics  
Keele University  
Keele, ST5 5BG  
b.k.olorisade@keele.ac.uk

**Pearl Brereton**  
School of Computing and Mathematics  
Keele University  
Keele, ST5 5BG  
o.p.brereton@keele.ac.uk

**Peter Andras**  
School of Computing and Mathematics  
Keele University  
Keele, ST5 5BG  
p.andras@keele.ac.uk

### Abstract

Reproducibility is an essential requirement for computational studies including those based on machine learning techniques. However, many machine learning studies are either not reproducible or are difficult to reproduce. In this paper, we consider what information about text mining studies is crucial to successful reproduction of such studies. We identify a set of factors that affect reproducibility based on our experience of attempting to reproduce six studies proposing text mining techniques for the automation of the citation screening stage in the systematic review process. Subsequently, the reproducibility of 30 studies was evaluated based on the presence or otherwise of information relating to the factors. While the studies provide useful reports of their results, they lack information on access to the dataset in the form and order as used in the original study (as against raw data), the software environment used, randomization control and the implementation of proposed techniques. In order to increase the chances of being reproduced, researchers should ensure that details about and/or access to information about these factors are provided in their reports.

### 1 Introduction

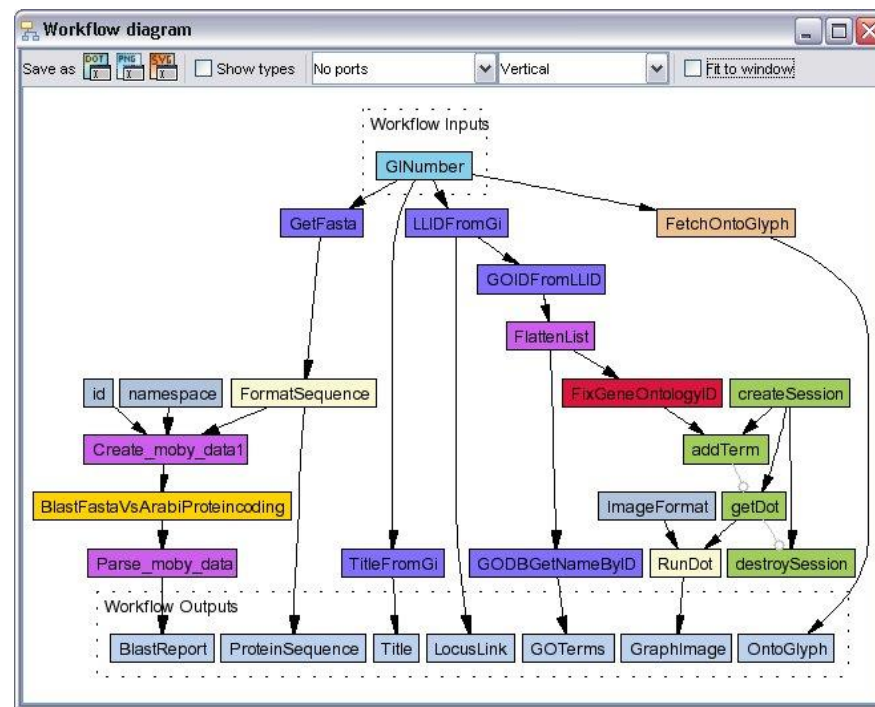
Independent verification of published claims for the purpose of credibility confirmation, extension and building a 'body of knowledge' is a standard scientific practice [13]. Machine learning methods based research are not excluded from this strict scientific research requirement. However, it may sometimes be hard or even impossible to replicate computational studies of this nature [12]. This is why the minimum standard expected of any computational study is for it to be reproducible [11].

In order for a study to be reproduced, an independent researcher will need at least full information and artefacts of the experiment - datasets, experiment parameters, similar software and hardware environment etc., as used in the original study. However, the experience in studies today shows a lack of sufficient information that can enable an independent researcher reproduce majority of the studies successfully.

# Reproducibility of Taverna workflows

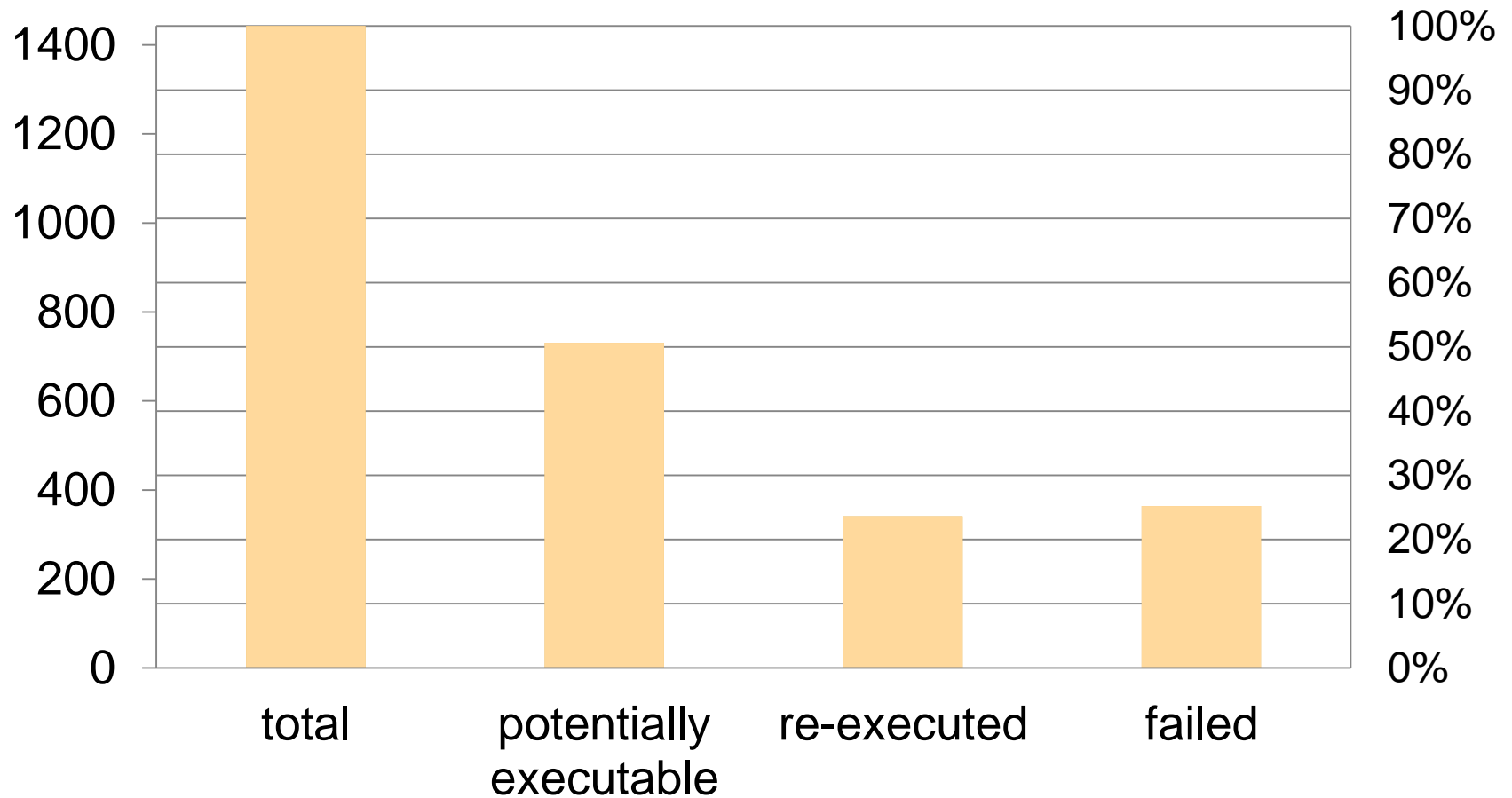
## ■ Workflows at *myExperiment* [1]

- 1443 workflows analysed
- 75% are Taverna workflows
- 43% use web services
- 50% use Beanshells  
for Java code snippets



[1] Rudolf Mayer and Andreas Rauber. A quantitative study on the re-executability of publicly shared scientific workflows. Proc. of IEEE *e-Science* 2015.

# Reproducibility of Taverna workflows



- Reasons for workflows to break
  - Missing input data
  - Data 'was not pure'
    - available data cannot be attributed automatically to a given input
  - External services
    - unavailable
    - password protected
  - Command line tools
    - not available in the environment by default

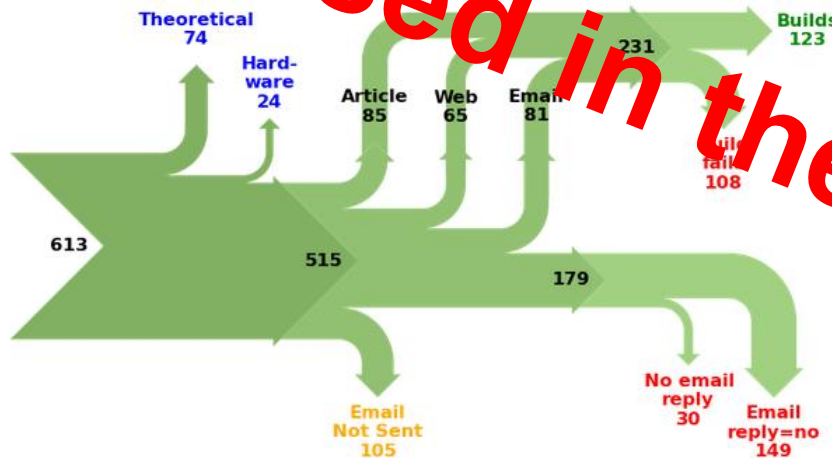
- 613 papers in 8 ACM conferences

- E-mail responses from authors

- Wrong version
- Code will be available soon
- Programmer left
- Bad backup practices
- Commercial code

Proprietary academic code

- Intellectual property
- No intention to release
- ...

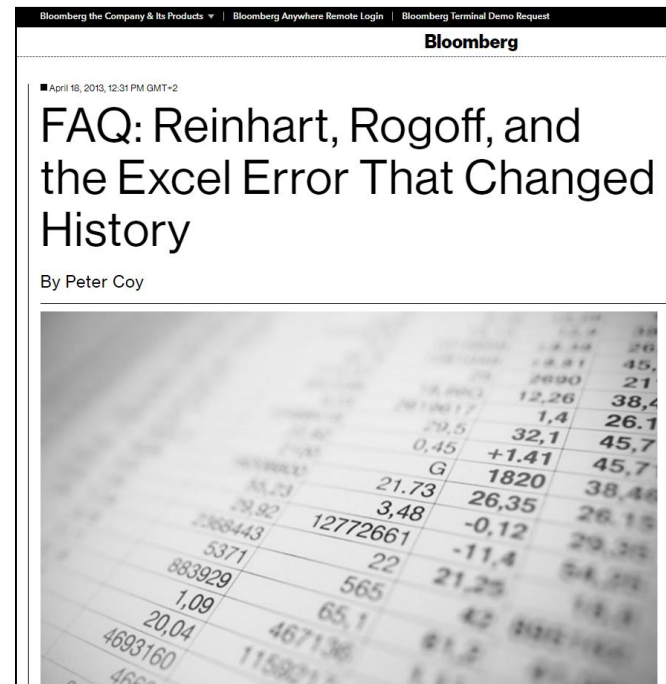


C. Collberg and T. Proebsting, "Measuring reproducibility in computer systems research," 2014. [Online]. Available: <http://reproducibility.cs.arizona.edu/tr.pdf>



# Reproducibility and politics

- *Growth in a Time of Debt*
  - Published after the Greek crisis
  - Analysis supporting budget cuts!
    - Stimulus vs austerity
- Others could not reproduce the result
- Original spreadsheet investigated
  - Some data excluded on purpose
  - Questionable statistical procedures
  - Excel error
    - accidentally excluded 5 rows
    - Average Annual Growth changed from -0.1 to 2.2 after correction



*'In politics, academic studies are used as weapons.'*

<https://www.bloomberg.com/news/articles/2013-04-18/faq-reinhart-rogoff-and-the-excel-error-that-changed-history>  
<https://www.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html>

# IMPROVING REPRODUCIBILITY

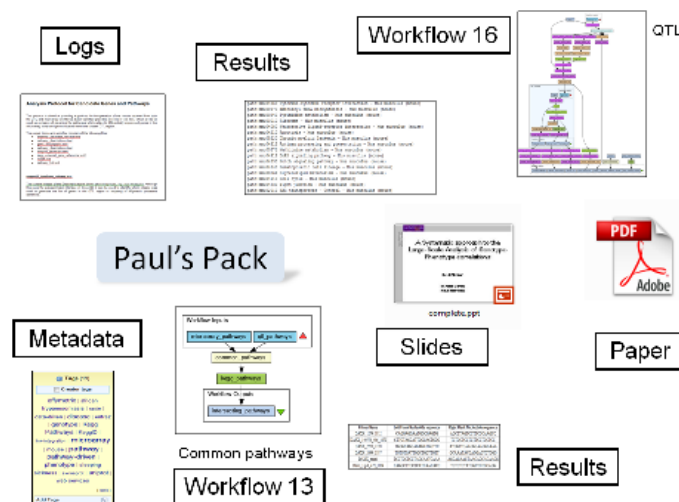
- Aim
  - improve experiment description
  - improve reproducibility of experiment
  - ease sharing of data
- Selected solutions
  - Data Management Plans
  - Research Objects
  - Controlled environments
  - Data sharing platforms
  - Open Access

# Data Management Plans

- Required during grant applications
- Focus on input and output data
- Provide general overview of the experiment
- Are manually created (supported by online tools)



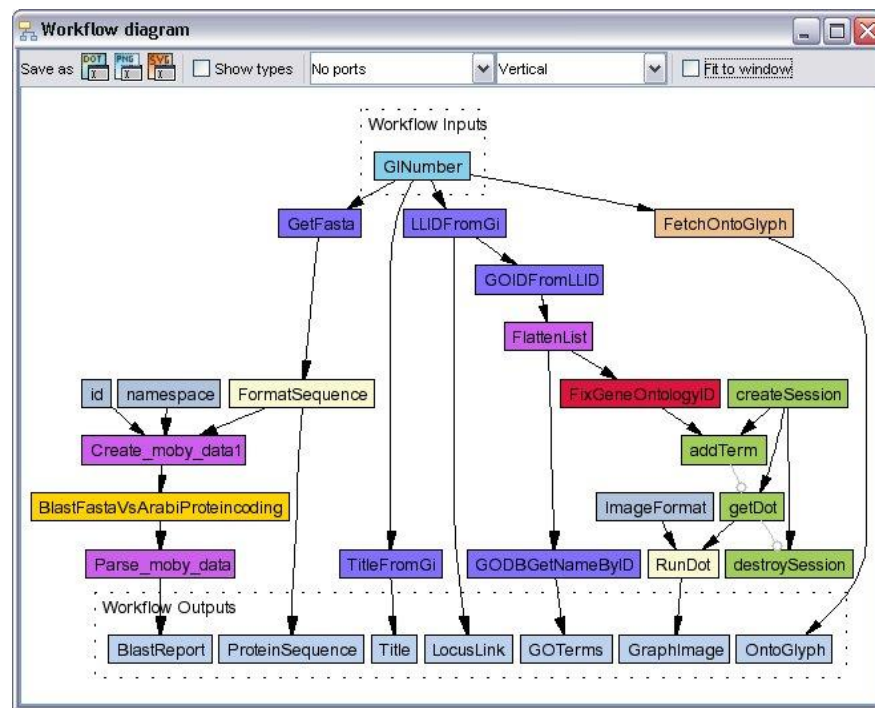
- Aggregate resources describing an experiment
- Contain
  - data and results
  - workflow
  - provenance traces
  - slides, papers, etc.
- Do NOT contain
  - specification of the environment in which the experiment was performed



<http://www.researchobject.org>

# Controlled environments

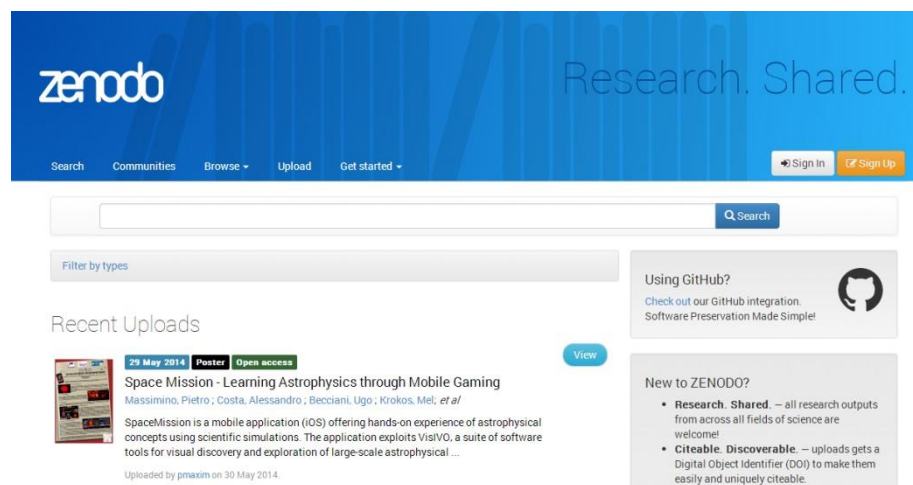
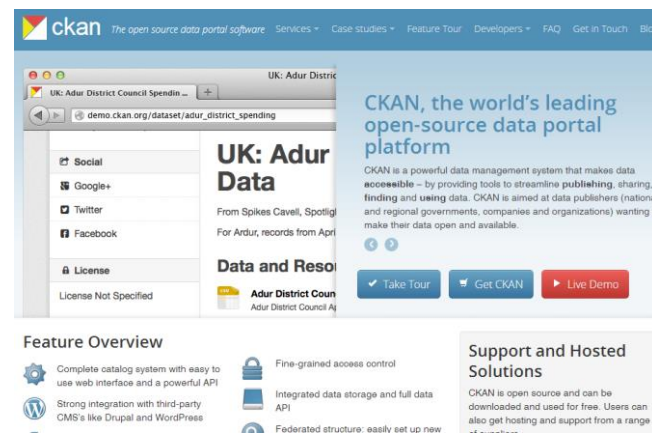
- Standardization of elements
- Automatic provenance collection
- Workflows systems
  - Taverna
  - Kepler





# Data sharing platforms

- Open access
- Unique identifiers for data
- Metadata collection
- Examples
  - Zenodo
  - myExperiment



- Open access (OA)
  - provides immediate, online and free access to research outputs
    - journal articles, conference papers, data....
  - removes *price barriers*
    - e.g. subscriptions, licensing fees, pay-per-view fees
  - removes *permission barriers*
    - most copyright and licensing restrictions
- Costs of publishing still exist
- Intellectual Property Rights (IPR) are not removed
  - not identical with the *public domain*



Open Access Explained: <https://www.youtube.com/watch?v=L5rVH1KGBCY>

## ■ Gold Open Access

- used by open and hybrid journals
- author pays for publishing
- article is available immediately



## ■ Green Open Access

- traditional subscription based journals (reader pays for reading)
- article 'self-archived' in a repository (e.g. of a university)
- embargo period may apply
- branding may be removed
- restrictions of version of manuscript may apply
  - before peer-review vs final version

# ACM Badges

- Artefacts Evaluated
  - Functional
    - basic scripts, data, explanation provided
    - no need to be public
  - Reusable
    - higher quality than functional
- Artefacts Available
  - has a DOI, publically available in a repository
- Results Validated
  - Replicated
  - Reproduced

<https://www.acm.org/publications/policies/artifact-review-badging>



# Public reproducibility review

- Students reproduce papers submitted to a conference in Machine Learning
  - universities, Google, DeepMind, Facebook, Microsoft, and Amazon
- Papers are anonymously published months in advance of the conference
- Comments are public
- Goal: put pressure on authors and improve reproducibility



*„the **software** a computer runs when reproducing an algorithmic experiment, as well as the **configuration** of that software and the **data** used, are comparable to the impact of **gravity** and **temperature** in the physical world. These elements provide **context** for an experiment, and need to be replicated to understand **how and why** the experiment works”*

<https://qz.com/1118671/the-titans-of-ai-are-getting-their-work-double-checked-by-students/>

# Reproducible builds

- Goal
  - **Independently-verifiable path** from **source code** to the **binary code** used by computers.
- **build system needs to be made entirely deterministic**
  - transforming a given source must always create the same result
    - no timestamps, same file order, etc.
- **build environment** should either be **recorded** or **pre-defined**
- User must be able to **validate that the output matches the original build**

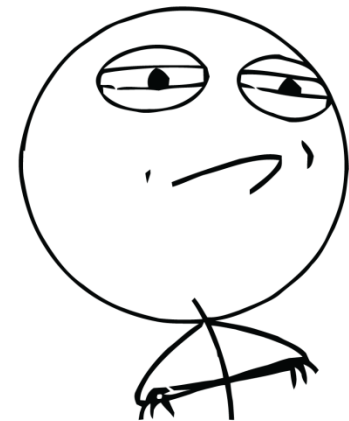
reproducible-builds.org

Provide a verifiable path from source code to binary.



# Challenge

- Aim:
  - preserve the **data analysis process**
  - not only data
  - not only publications
- Reasons:
  - verification of the analysis correctness
  - dismissal of any issues raised by others
  - design of improved analysis processes
  - reuse
  - TRUST



# AUDITABILITY AND ACCOUNTABILITY

# Auditability and Accountability

- Recently discussed in context of AI systems, but not limited to
  - Auditable AI systems
  - Explainable AI systems
- Auditable
  - able to be audited (Wiktionary)
  - Which data was used? Where the data comes from? What tools were used for processing? Who was involved? How the result was computed? Where is the evidence?
  - Common goals with Reproducibility
- Accountable
  - actions can be traced to who did them (ISO/IEC 25010)
- Explainable
  - Explains the black box of an algorithm
  - Why a system made a specific decision?
  - Explainable is not equal to Auditable

- “Uses AI” to provide cancer treatment advice
- Provided unsafe and incorrect treatment recommendations
  - trained on a small number of samples
  - using synthetic patients’ data
- Trained using data from the US
  - offered to customers in Asia that follow different treatment guidelines

Casey Ross, Ike Swetlitz. (2018, July 25) IBM’s Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show. STAT.

<https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrectcancer-treatments-STAT.pdf>

# IBM Watson

**“PHYSICIANS LIKE IT. PHYSICIANS HAVE SAID TO ME, IF I TOOK IT AWAY NOW, I’D HAVE A REVOLT.”**

DEBORAH DISANZO, GENERAL MANAGER OF  
IBM WATSON HEALTH, JUNE 2017

**“THIS PRODUCT IS A PIECE OF S---.  
WE BOUGHT IT FOR MARKETING AND  
WITH HOPES THAT YOU WOULD ACHIEVE THE  
VISION. WE CAN’T USE IT FOR MOST CASES.”**

ONCOLOGIST AT JUPITER MEDICAL CENTER,  
QUOTED IN IBM INTERNAL DOCUMENT, JUNE 26, 2017

# ACM Statement on Algorithmic Transparency and Accountability

1. Awareness

2. Access and redress

3. **Accountability**

Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.

4. Explanation

5. **Data Provenance**

A description of the way in which the training data was collected should be maintained by the builders of the algorithms...

6. **Auditability**

Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.

7. Validation and Testing

[http://www.acm.org/binaries/content/assets/public-policy/2017\\_joint\\_statement\\_algorithms.pdf](http://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf)

- Ai-Now
  - research group representing employees from tech companies like Microsoft and Google
  - issued 10 recommendations for AI development

## 8. Fairness, accountability, and transparency in AI require a detailed account of the “full stack supply chain“

“For meaningful **accountability**, we need to better understand and **track the component parts** of an AI system and **the full supply chain on which it relies**:

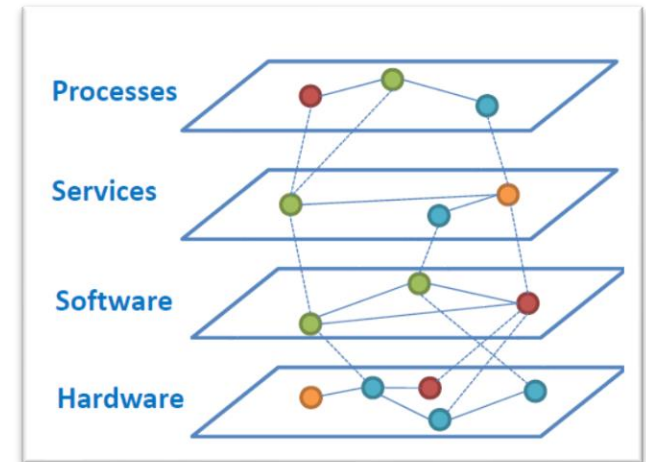
- accounting for the origins and use of training data, test data,
- models,
- application program interfaces (APIs),
- and other infrastructural components over a product life cycle.

We call this accounting for the “full stack supply chain” of AI systems, and it is a necessary condition for a more responsible form of **auditing**. (...)”

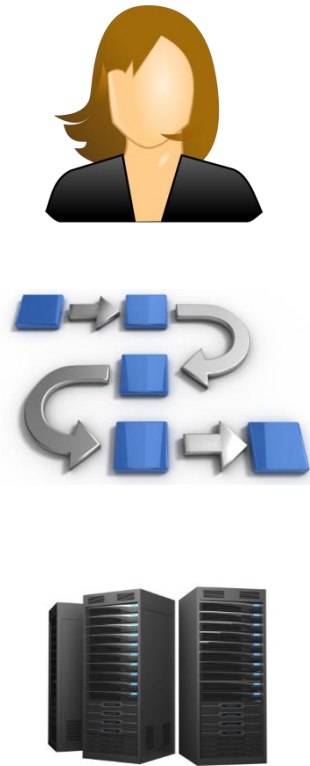
# PROCESS PRESERVATION



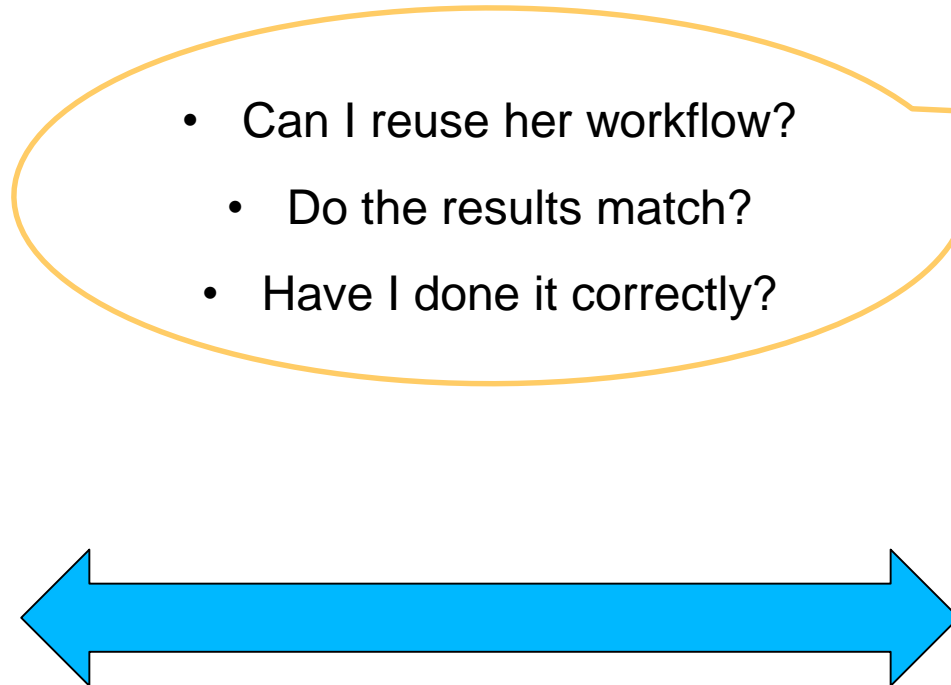
- Processes can be
  - ad hoc commands executed manually
  - well-structured processes within a controlled environment
- Processes
  - share infrastructure with other processes
  - delegate tasks to specific tools installed in the system
  - require specific configurations
  - can use distributed systems



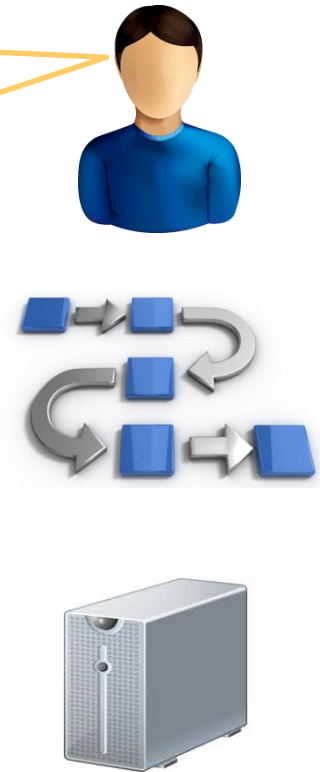
# Challenge



Original  
process



- Can I reuse her workflow?
- Do the results match?
- Have I done it correctly?



Re-executed  
process

# Challenge

Original environment



Repository



Redeployment environment



**DIRECT COMPARISON NOT POSSIBLE**

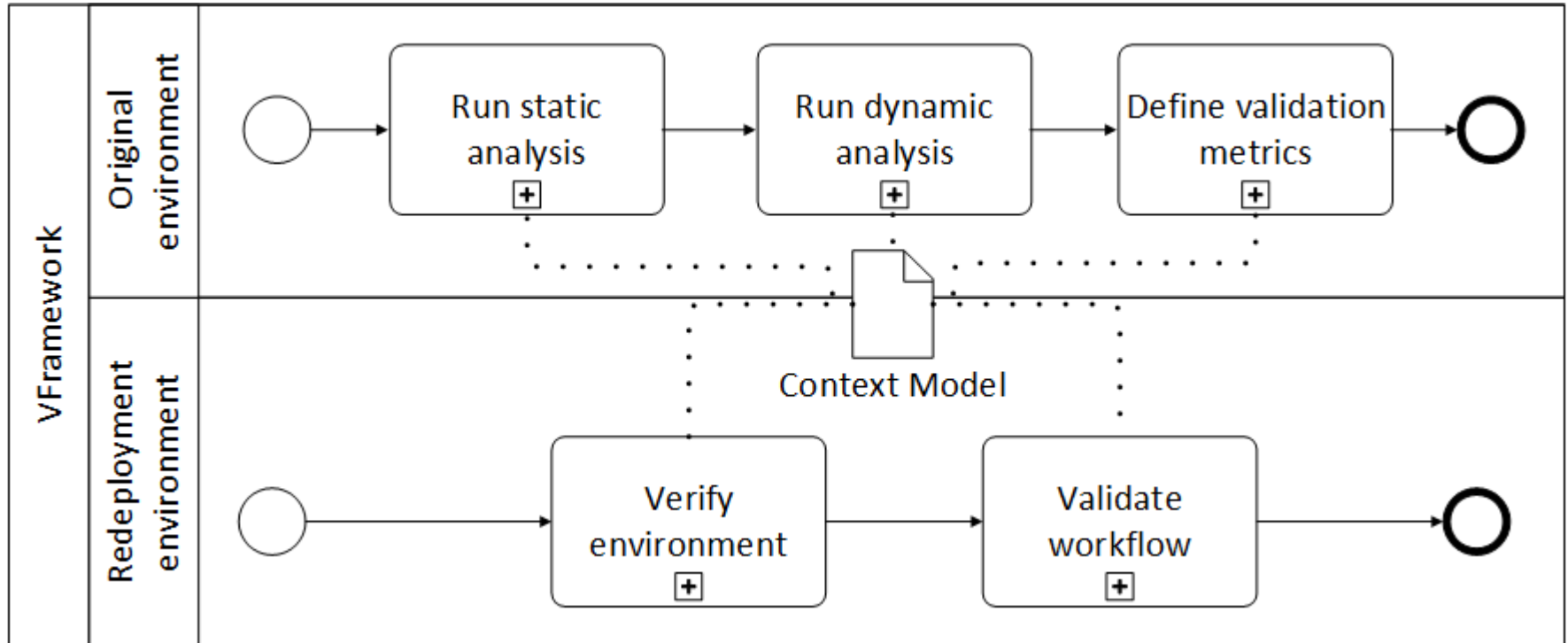


**COLLECT EVIDENCE**



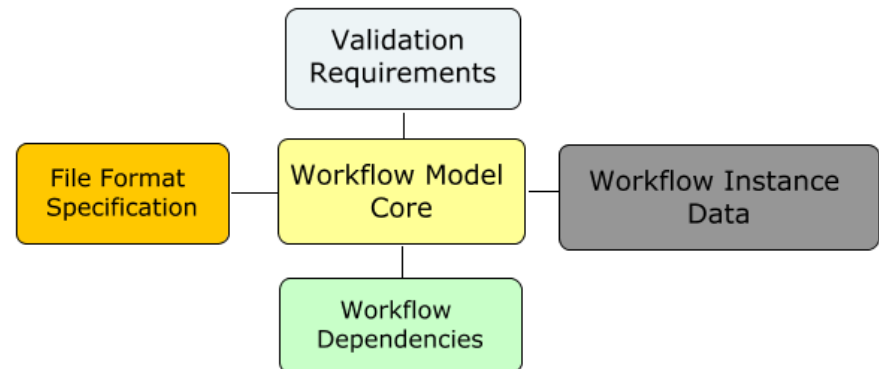
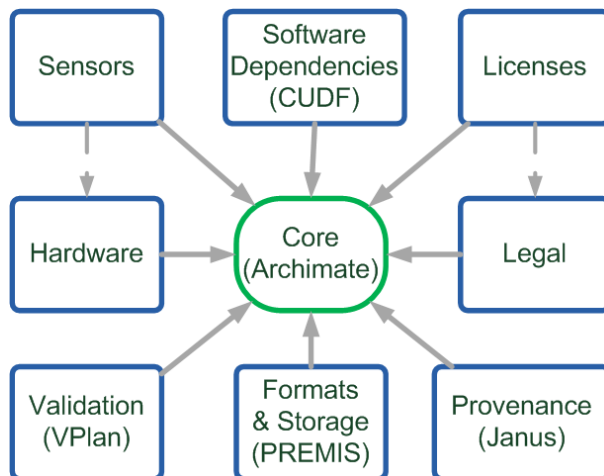
**COMPARE EVIDENCE**

# VFramework



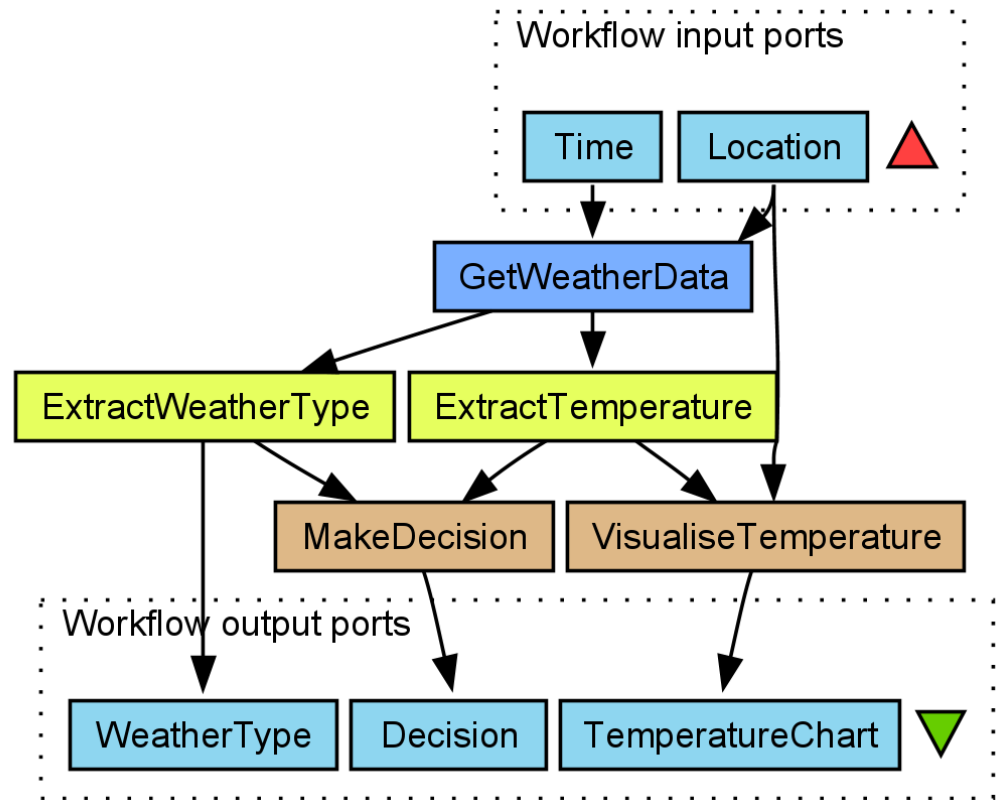
# Context Model

- OWL ontology
- Enables comprehensive process description

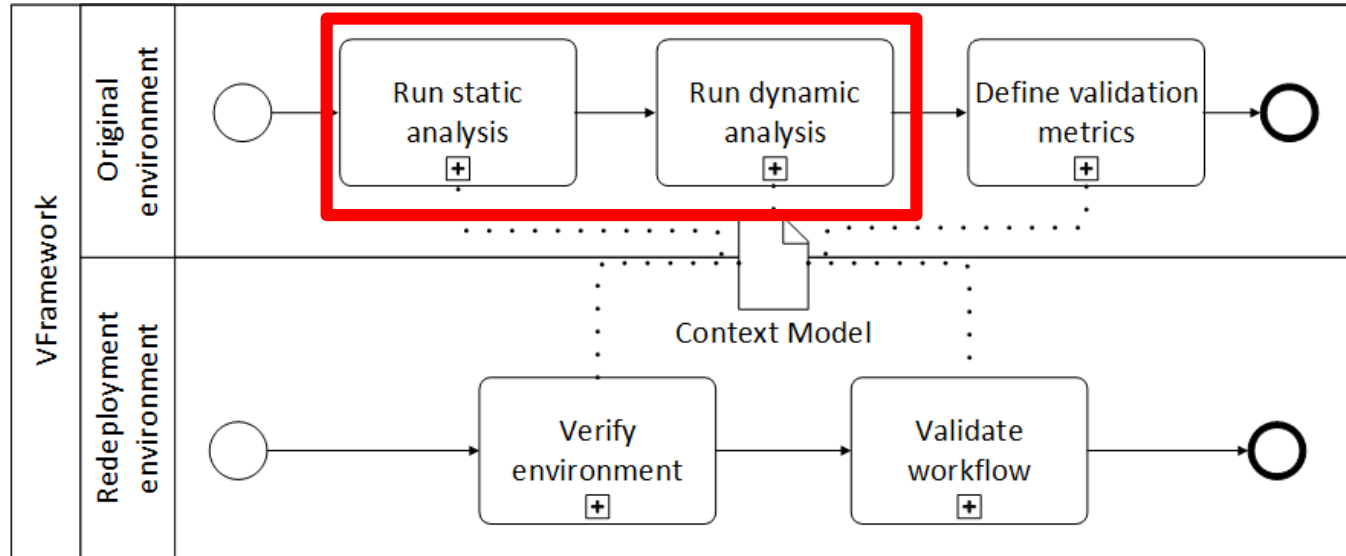


# Running example

- Computational research typically includes [1]
  - data capture
  - data analysis
  - data visualisation



[1] Tony Hey et. al, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.



Capturing original workflow execution

# VFRAMEWORK

- Static analysis
  - describes workflow model
    - steps, platform, services, tools
  - defines boundaries
    - elements inside – monitored during dynamic analysis
    - elements outside – their identity must be ensured
  - support
    - conversion tools
    - SPARQL queries

```
#!/bin/bash  
echo „Hello world”
```





# VFramework - Run static analysis

- Transformation requires
  - mapping between concepts
  - development of conversion tools

```
#!/bin/bash

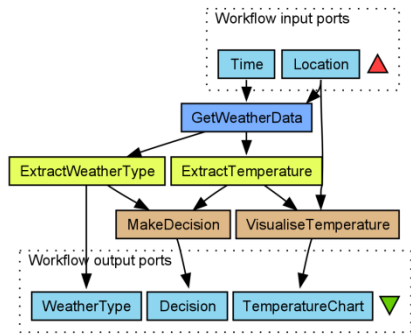
# fetch data
java -jar GestBarragensWSCClientIQData.jar
unzip -o IQData.zip

# fix encoding
#iconv -f LATIN1 -t UTF-8 iq.r > iq_utf8.r

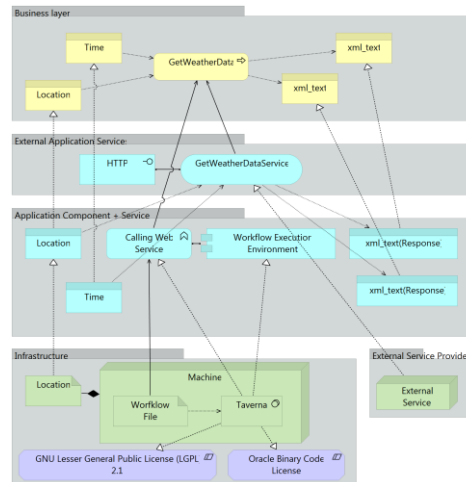
# generate references
R --vanilla < iq_utf8.r > IQout.txt

# create pdf
pdflatex iq.tex
pdflatex iq.tex
```

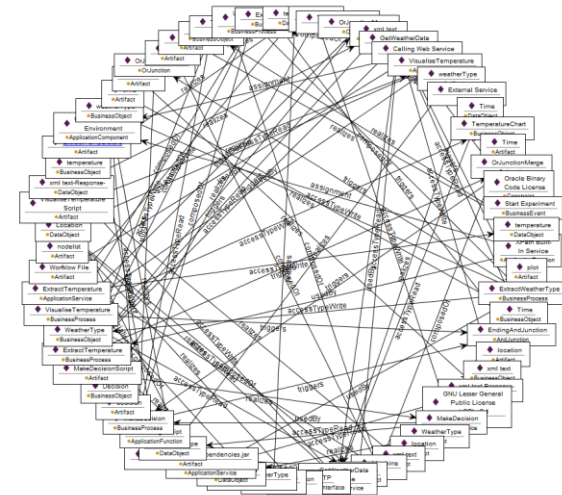
Script



Taverna Workflow



ArchiMate model



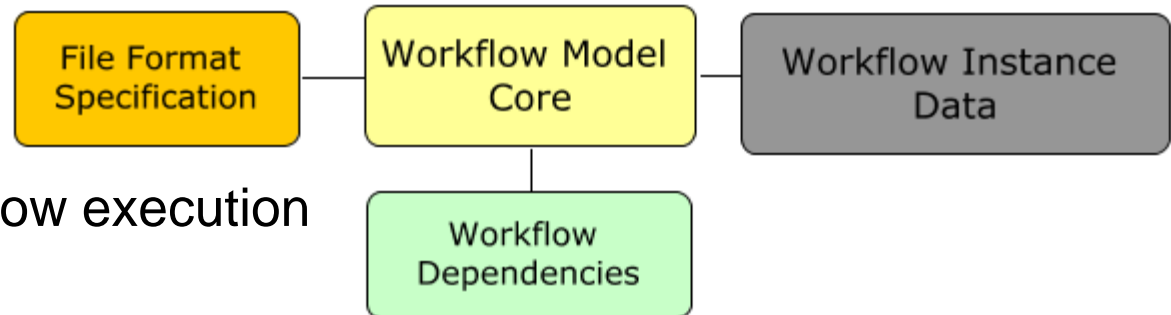
Context Model  
(OWL ontology)

# VFramework - Run dynamic analysis

## Local dependencies

- Test instance selection
- Local dependencies
  - within workflow boundary
  - examples
    - software libraries, scripts, shell tools, packages, operating system

- Capturing
  - manually
  - virtualisation
  - model of a workflow execution



# VFramework - Run dynamic analysis

## Local dependencies

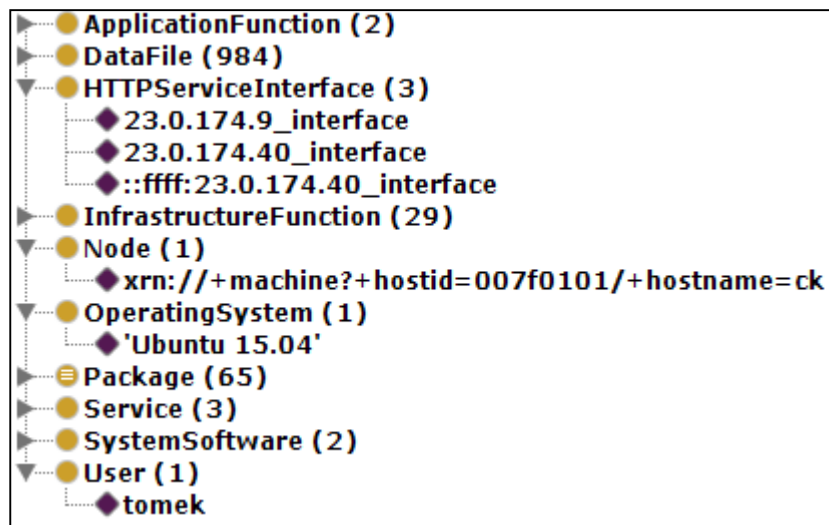
- Process Migration Framework (PMF) [1]
  - designed for automatic redeployments into virtual machines
  - uses *strace* to monitor system calls
  - complete log of all accessed resources
  - creates context model



[1] Johannes Binder. Migration of processes from shared to dedicated systems. Master's thesis, Vienna University of Technology, Wien, Austria, 2014.

# VFramework - Run dynamic analysis

## Local dependencies



Context model  
local dependencies

Dependencies Overview	
Shell calls	0
Remote services	3
Specific debian packages required	48
Specific file dependencies	1
Data files processed during workflow execution	7

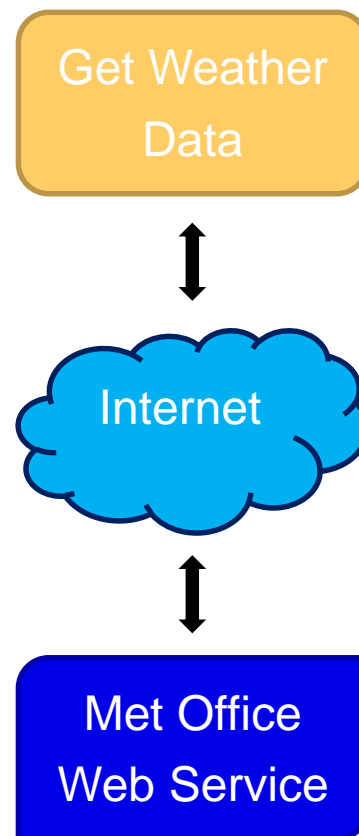
Detailed results	
OS specific command line invocations	
There are no shell calls.	
Workflow communication to external hosts	
23.0.174.40_interface	
23.0.174.9_interface	
::ffff:23.0.174.40_interface	
Required additional files and libraries	
/home/tomek/taverna-commandline-core-2.5.0/lib/chart-1.0-jar-with-dependencies.jar	
Data files used by the workflow	
/home/tomek/Weather/	
/home/tomek/Weather/log	
/home/tomek/Weather/output/Decision/1/1	
/home/tomek/Weather/output/TemperatureChart/1	
/home/tomek/Weather/output/WeatherType/1	
/home/tomek/Weather/workflow/Weather.t2flow	
/home/tomek/Weather/workflowInvocation.sh	
Required additional Debian packages	
base-files	
cups-filters	
fontconfig-config	

Dependency report

# VFramework - Run dynamic analysis

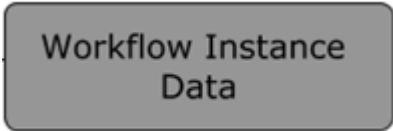
## External dependencies

- External dependencies
  - beyond boundaries
  - workflow exchange data with them
- When re-executing the workflow we can
  - use the original service
  - use a compliant service
  - use a mock-up



## ■ Workflow Instance Data

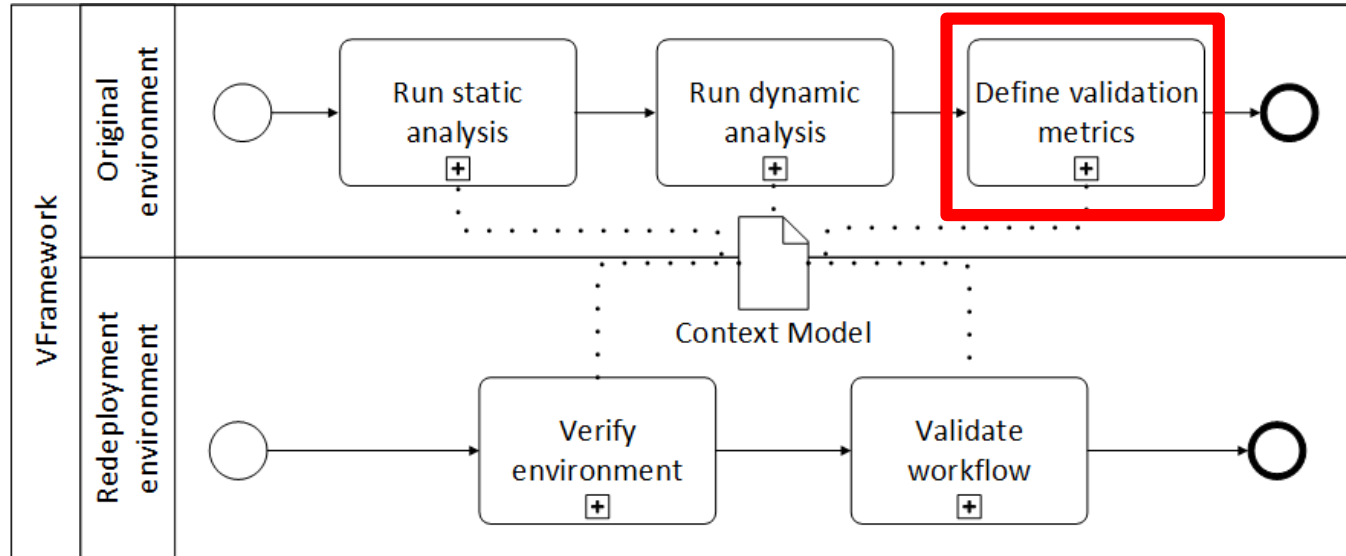
- needed for validation
- contains data processed by workflow
- detected by PMF
- recorded by Taverna in provenance traces

A grey rounded rectangular box containing the text 'Workflow Instance Data' in black.

## ■ File Format Characterisation

- needed for validation
- format detection using PRONOM registry
- modelled using PREMIS ontology

A yellow rounded rectangular box containing the text 'File Format Specification' in black.



Validation requirements

# VFRAMEWORK

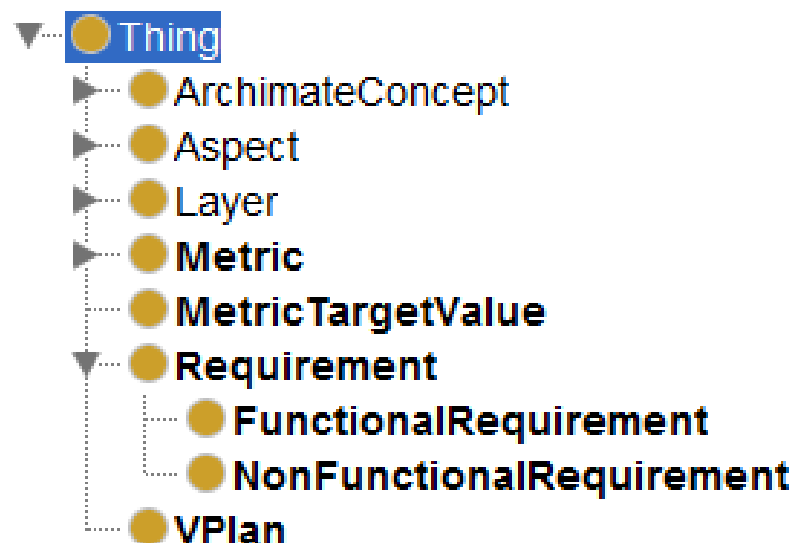
# VFramework – Define validation metrics



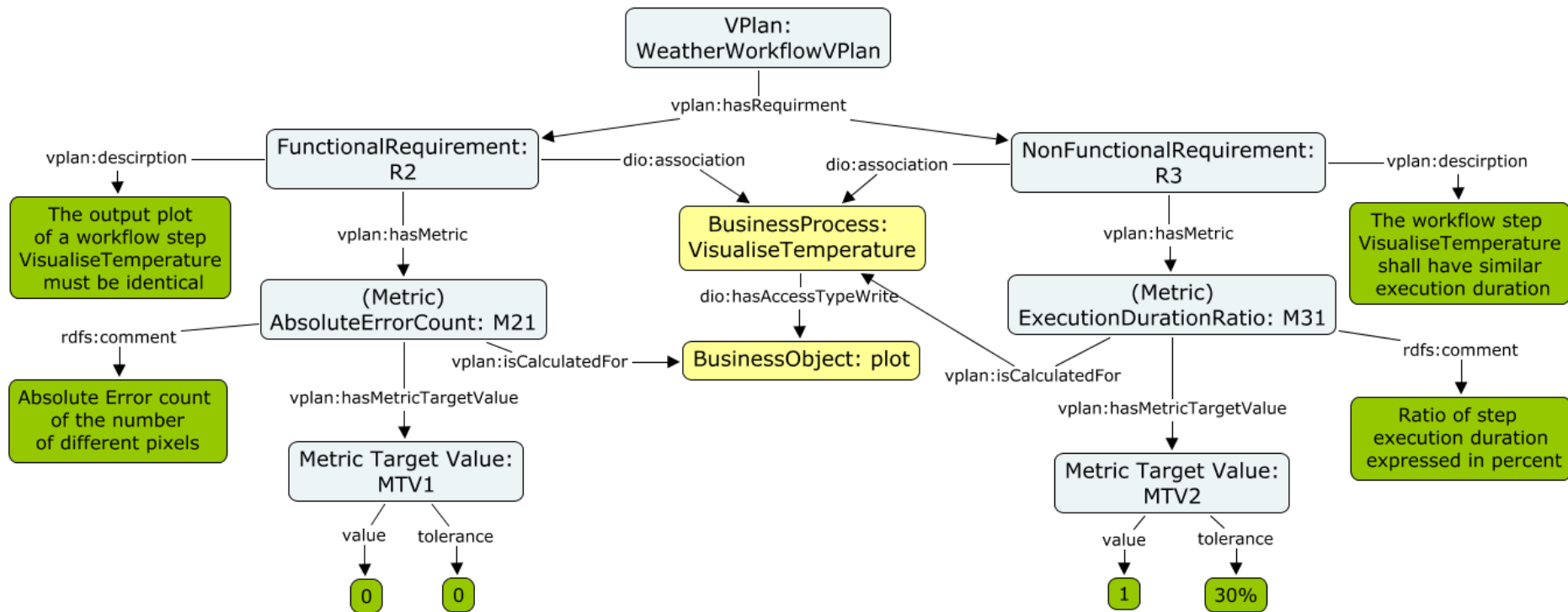
Bob needs to know how to  
compare results.  
How can I tell him that?



- Ontology for storing validation requirements
- Integrates with the Context Model
- VPlan model
  - defines classes
  - contains vocabulary of metrics
- VPlan instance
  - describes particular validation requirements

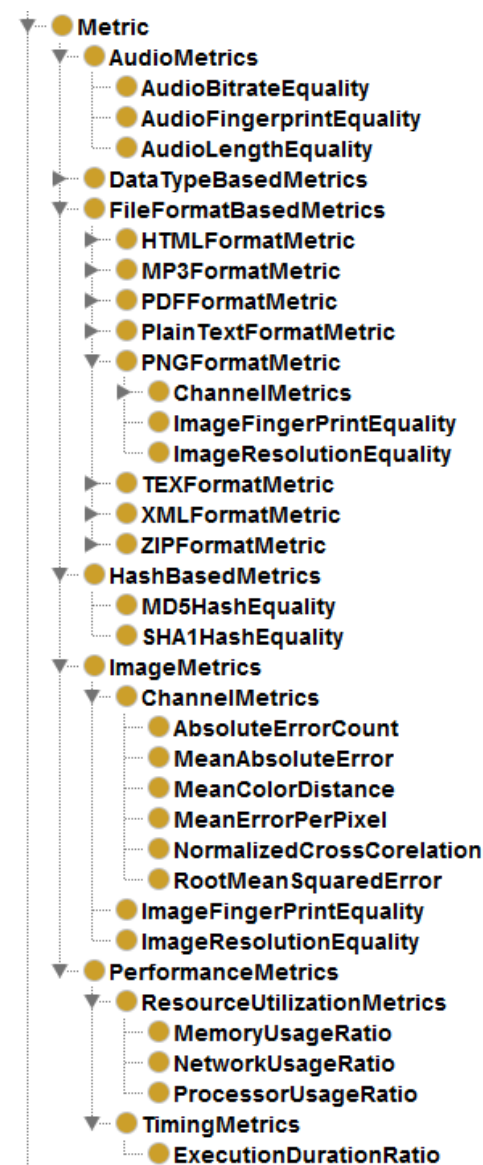


# VPlan – *weather* workflow example

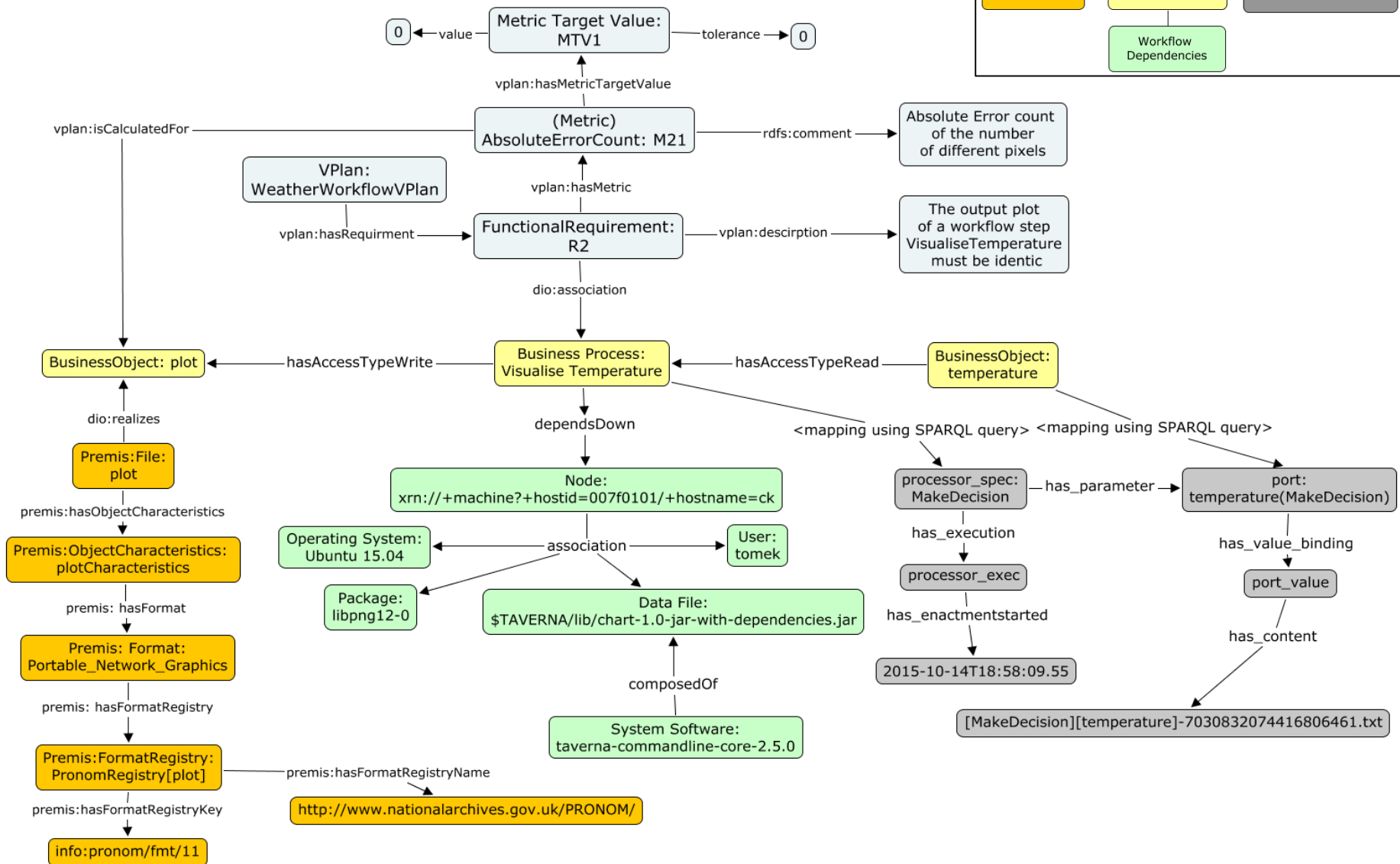


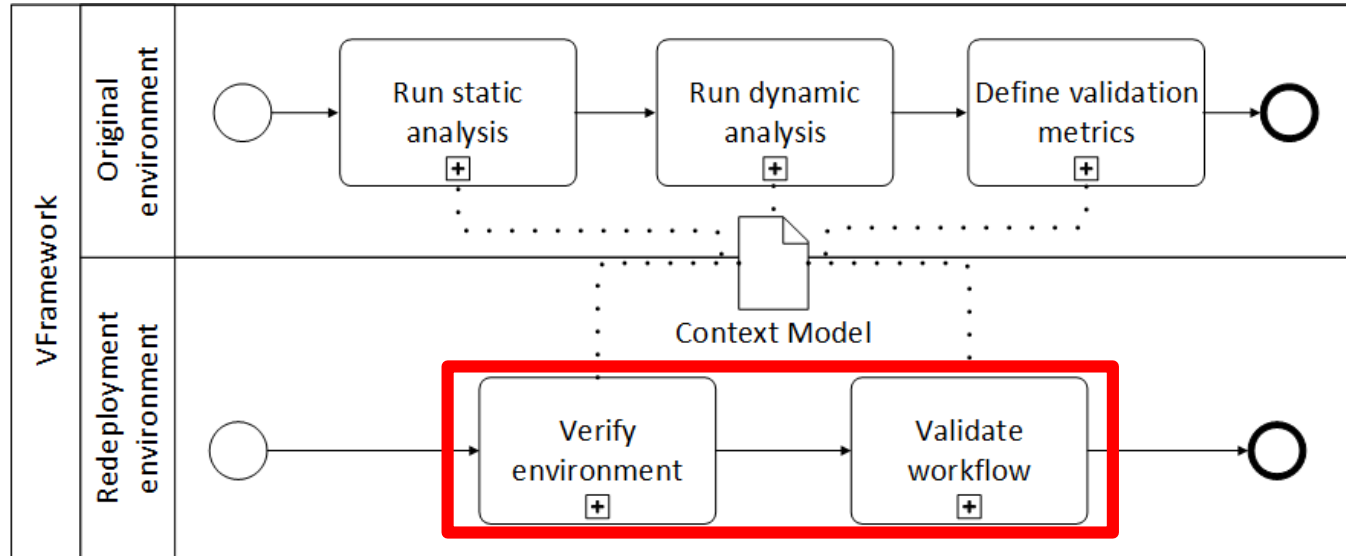
# VPlan – controlled vocabulary

- Based on
  - available tools
  - literature review
- Metrics
  - Audio Metrics
  - Data Type Based Metrics
  - File Format Based Metrics
  - Hash Based Metrics
  - Image Metrics
  - Timing Metrics



# Context model

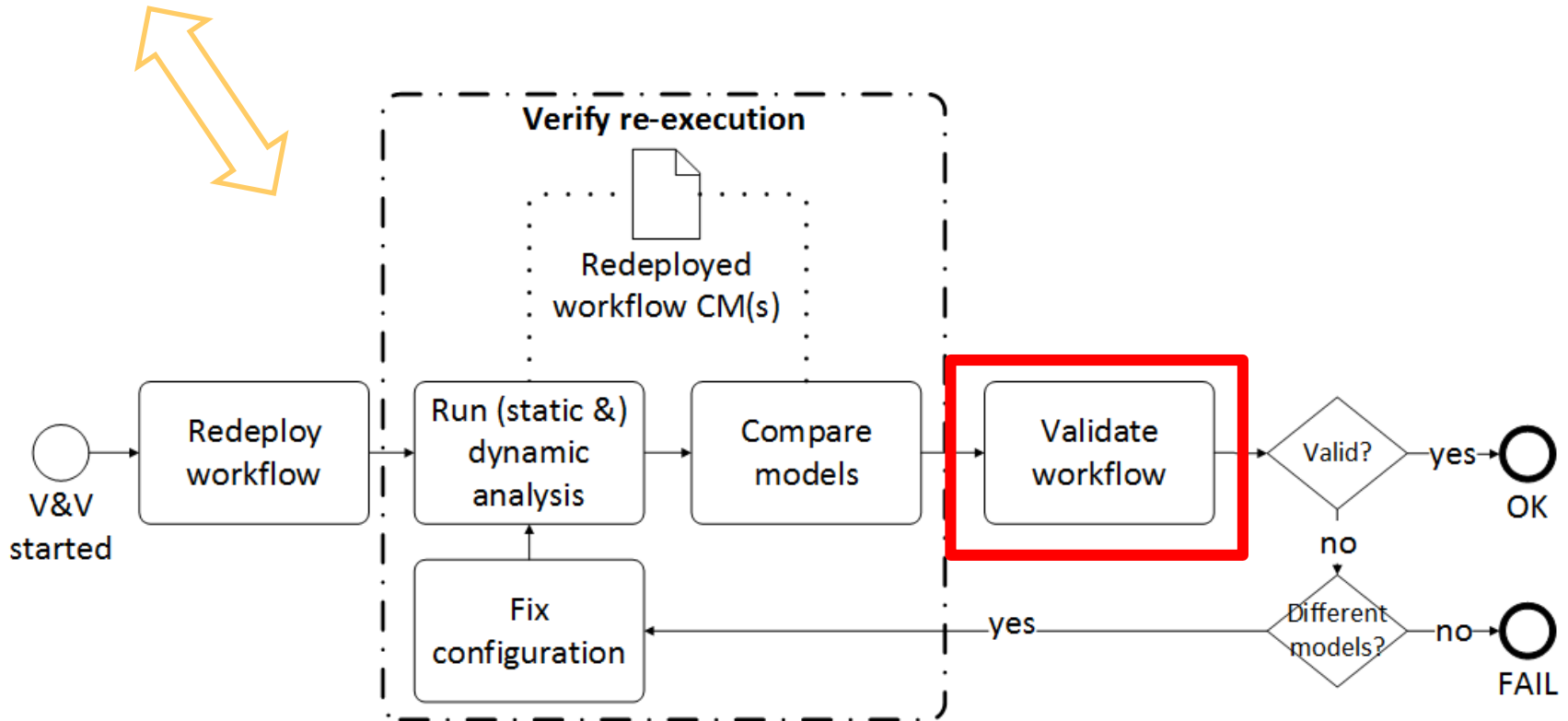
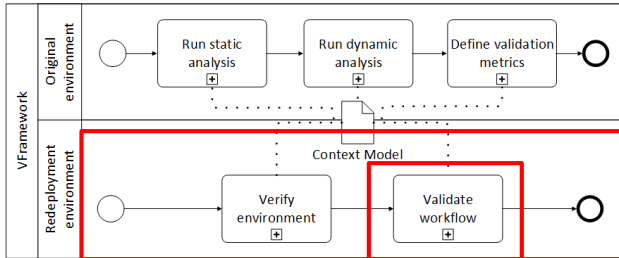




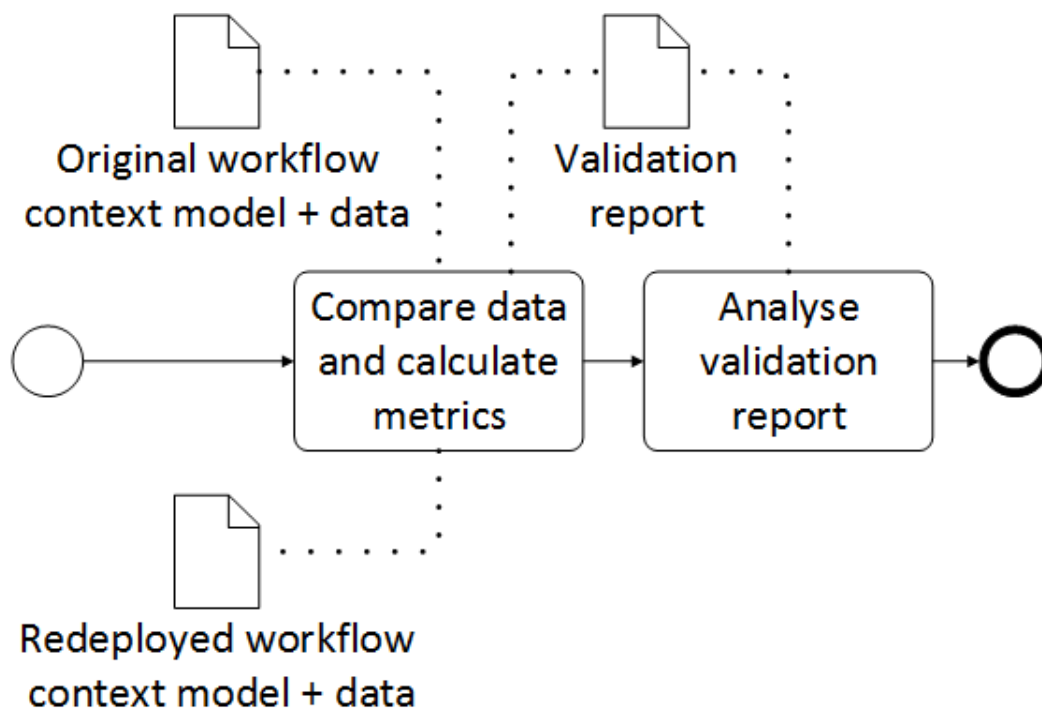
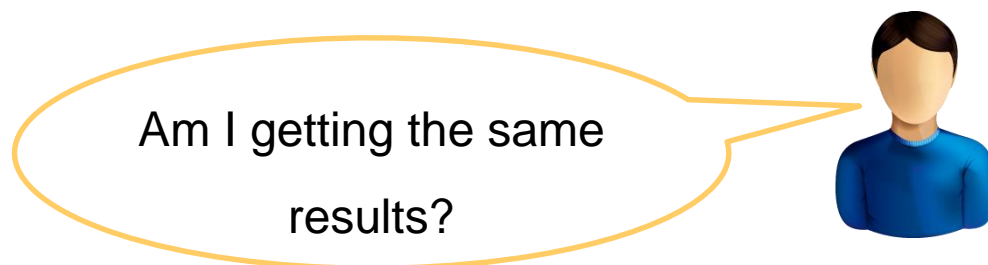
Verification and validation of workflow re-execution

# VFRAMEWORK

# VFramework - Verify environment



# VFramework – Validate workflow



# Validation report for the WeatherExample

Evaluation result: There are 2 not fulfilled metrics. Please see tables below for details.

Comparison performed using following workflow execution traces

Original Workflow

ID: 37b4d2fb-e71c-4b67-b7b3-17888ee82977

Timestamp: 2015-10-14 18:58:06.475

Compared Workflow

ID: ede04b87-5f58-4a89-b0c3-e179957cbad0

Timestamp: 2015-11-13 13:59:56.443

Table 1: Overview of requirements

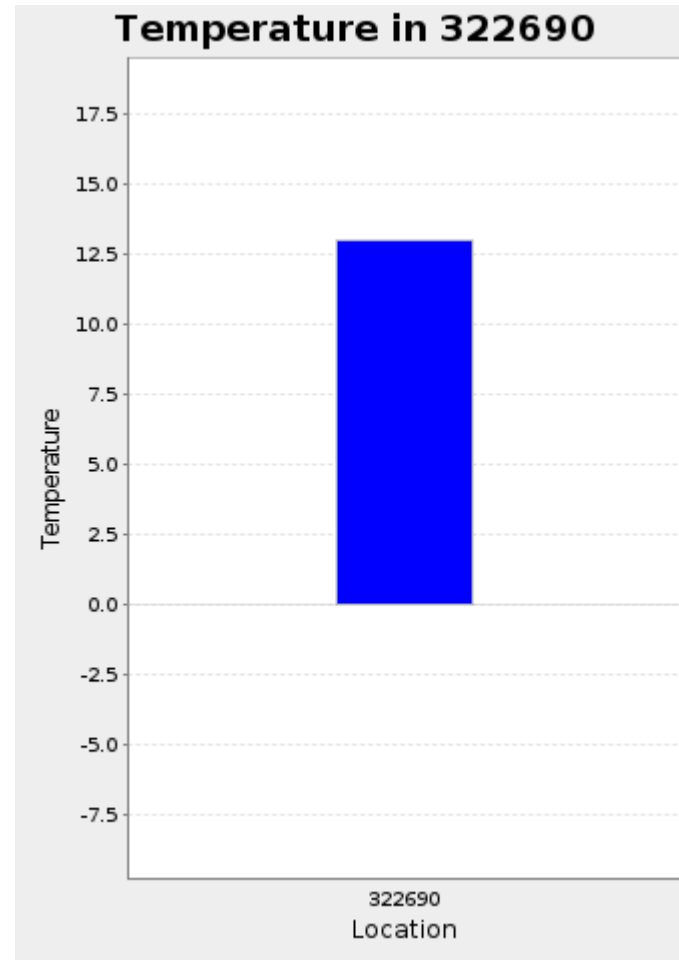
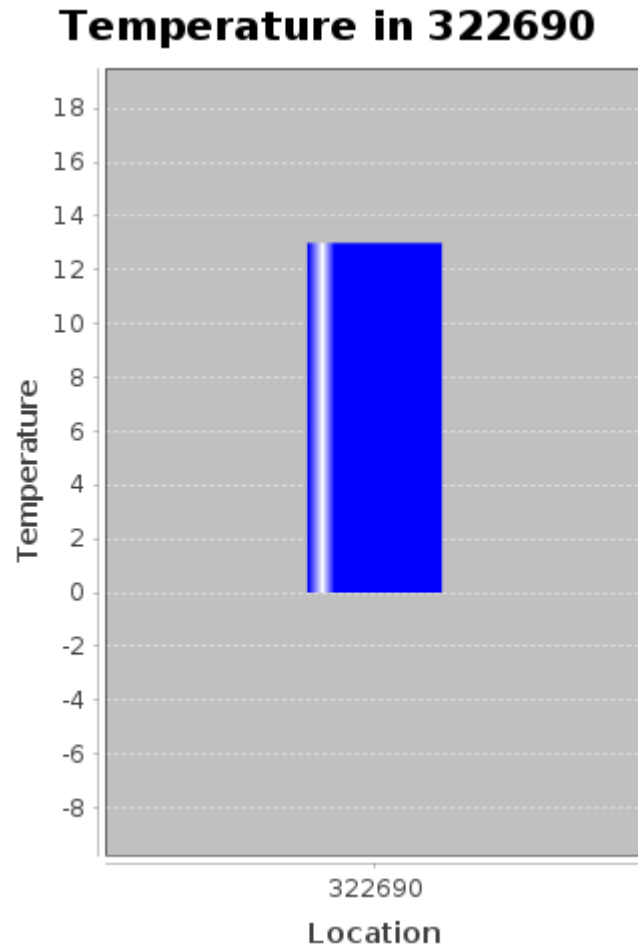
Requirement	Description	Is Fulfilled
R1	The inputs to the workflow are the same	true
R2	The outputs of the workflow are the same	false
R3	The workflow step ExtractTemperature must have identical outputs	true
R4	The workflow step GetWeatherData must have identical outputs	true
R5	The workflow step MakeDecision must have identical outputs	true
R6	The workflow step ExtractWeatherType must have identical outputs	true
R7	The workflow step VisualiseTemperature must have identical outputs	false
R8	Execution duration of each of the workflow steps shall be similar	true

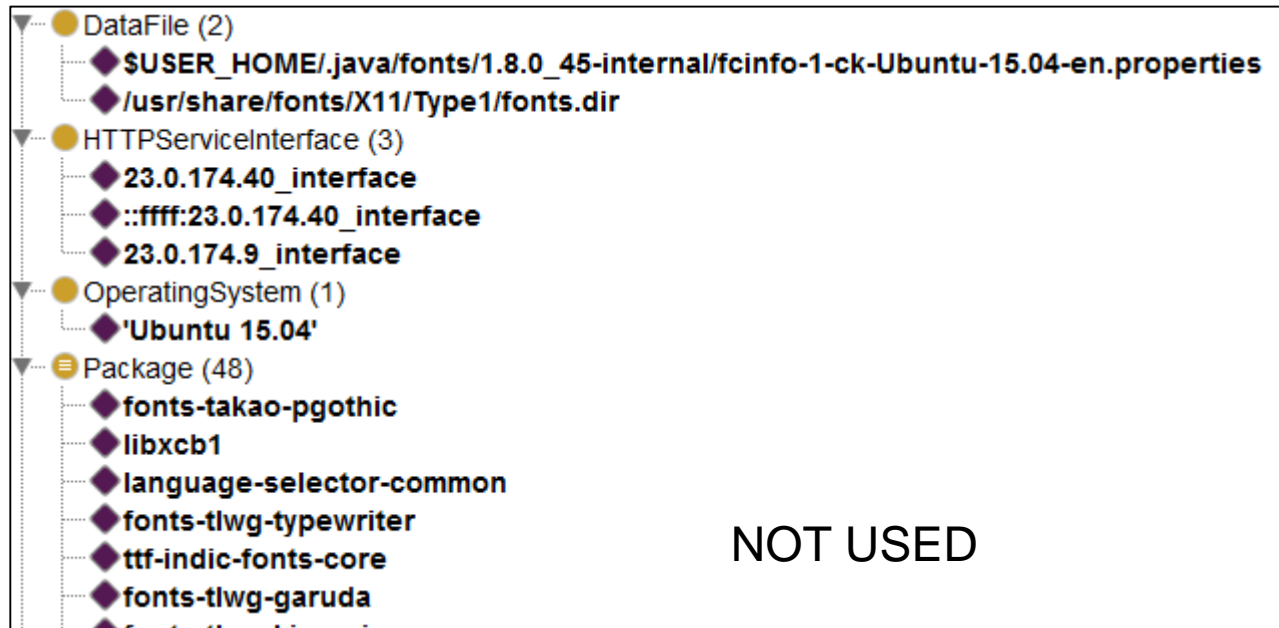
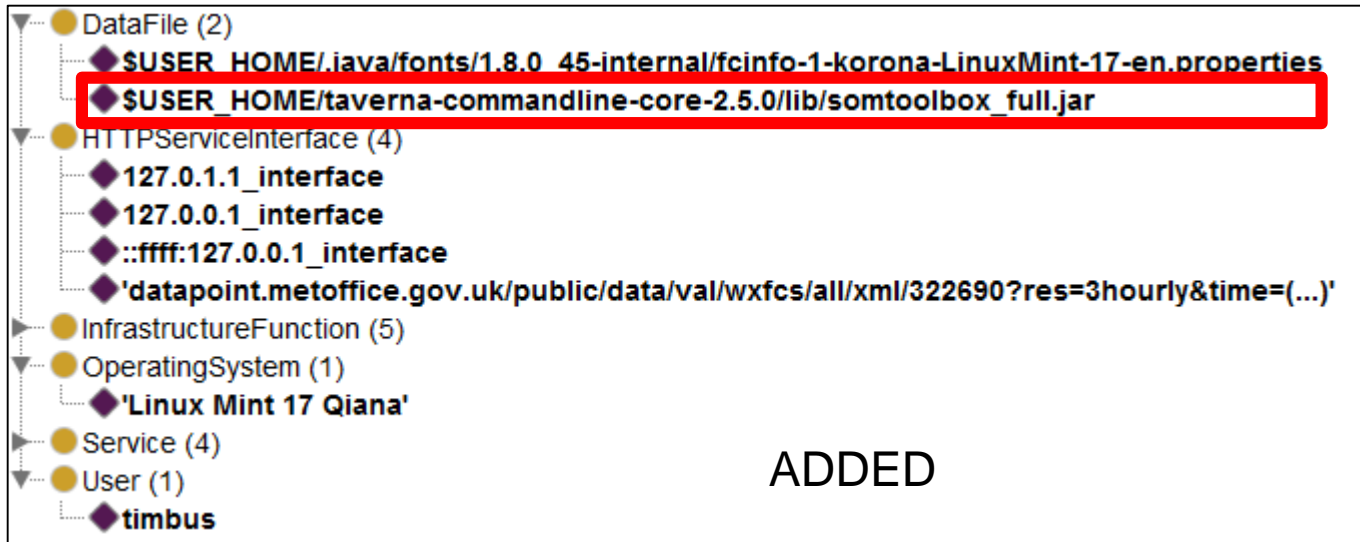
Table 2: List of requirements and metrics that failed.

Req	Sub-req	Sub-requirement description	Measure ment point	Metric	Validity
R2	R7.1	The output plot of workflow step VisualiseTemperature must be identical	plot	ImageFingerPrintEquality	false
				ImageResolutionEquality	true
				AbsoluteErrorCount	false
R7	R7.1	The output plot of workflow step VisualiseTemperature must be identical	plot	ImageFingerPrintEquality	false
				ImageResolutionEquality	true
				AbsoluteErrorCount	false

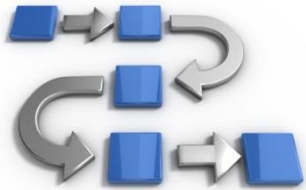


# Same experiment different library

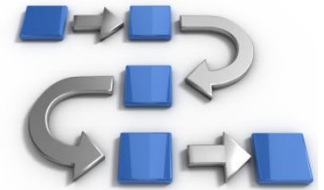
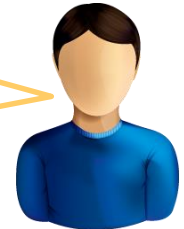
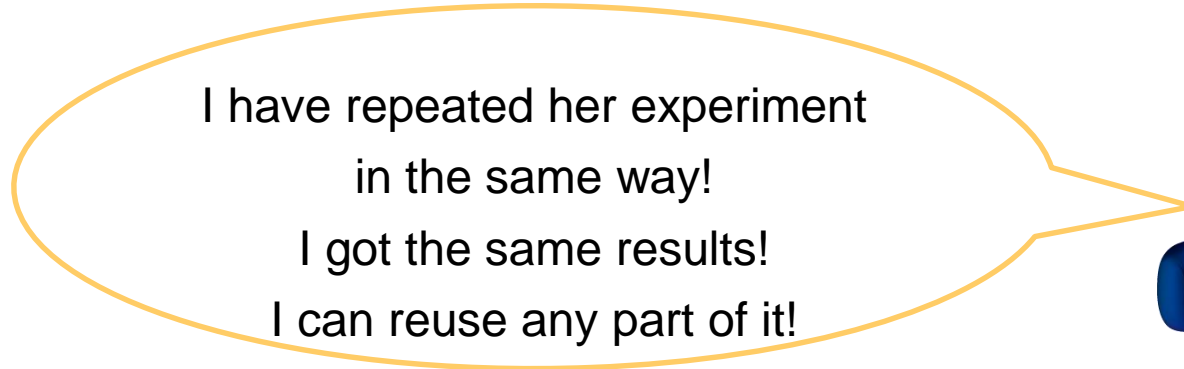




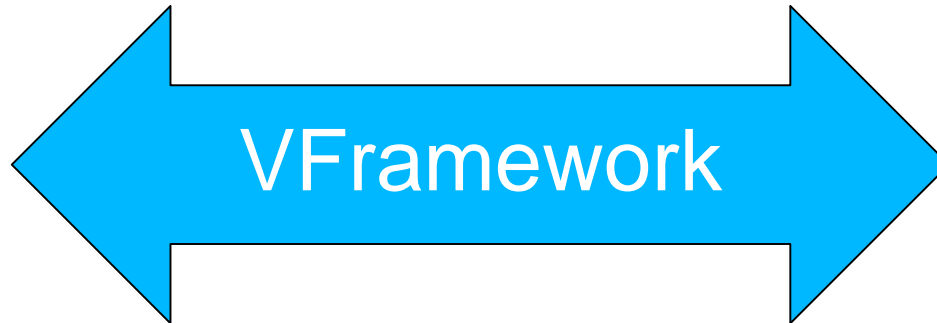
# Conclusions



Original  
experiment

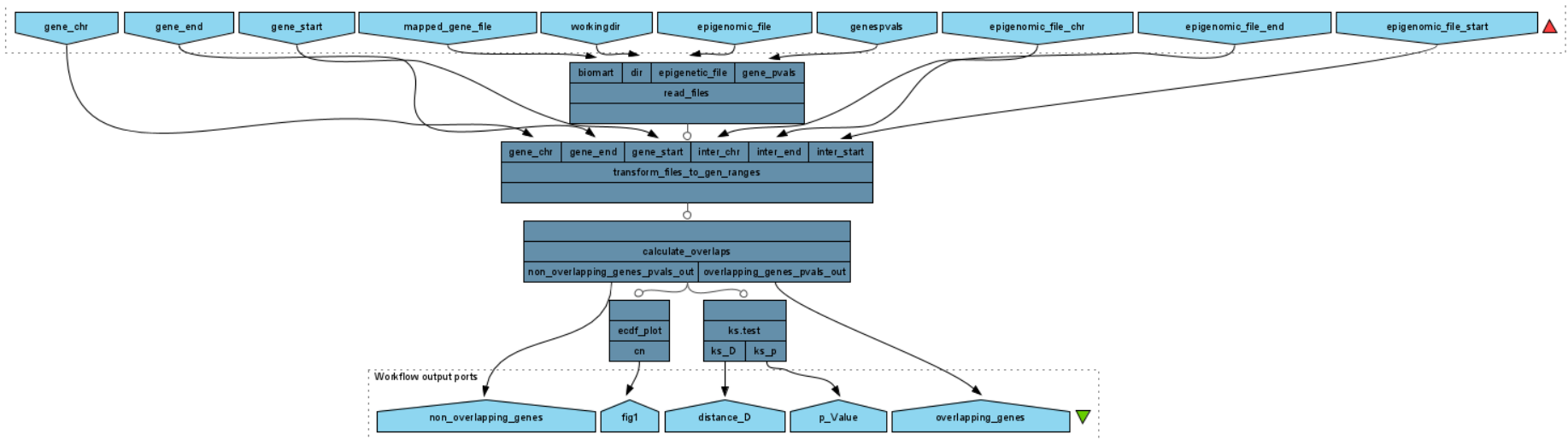


Re-executed  
experiment



# Interesting cases

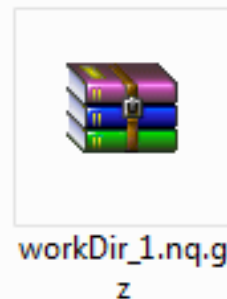
- All external communications detected (web services)
- Limited validation for external services (Rserve scripts)
  - global variables
  - steps with no outputs
  - only final result of workflow computation is validated
    - black box testing



## Interesting cases

- Dynamic analysis detected
  - data created through shell calls
    - such files are not a part of provenance traces in Taverna
  - 'real' workflow outputs
    - Taverna traces can contain paths only, but not the content
    - Taverna workflows can create files not linked to any output

```
I, [2015-11-30T15:44:14.448859 #8787] INFO -- : row 1
skipped. I, [2015-11-30T15:44:15.159204 #8787] INFO -- :
===== running time total: 0.710367914 =====
No of statements in a file 29590 I,
[2015-11-30T15:44:15.159824 #8787] INFO -- : 0.708272803
```



# Interesting cases

- Metadata included in data
  - generation timestamp

```
% latex table generated in R 3.0.2 by xtable 1.7-1 package
% Tue Apr 21 13:37:05 2015
\begin{table}[ht]\centering
\begin{tabular}{rrrrrr}
\hline & Min & 1Q & Mediana & Media & 3Q & Max \\
\hline Residuos (mm) & -2.42 & -1.23 & -0.52 & -0.00 & 0.45 & 5.52 \\
\hline
\end{tabular}
\caption{Residuos do modelo}
\end{table}
```

Tabela 1: Dados gerais do modelo de IQ

Utilizador admin	
Data da execu $\tilde{t}_{\frac{1}{2}}\tilde{t}_{\frac{1}{2}}\tilde{o}$	21-04-2015 12:37
Barragem	275
Fim da constru $\tilde{t}_{\frac{1}{2}}\tilde{t}_{\frac{1}{2}}\tilde{o}$	1943-04-01
In $\tilde{t}_{\frac{1}{2}}\tilde{t}_{\frac{1}{2}}\tilde{c}$ io do primeiro enchimento	1960-01-01
Tipo de instrumento	125
Grandezas	Desl. radial (mm) Desl. tangencial (mm)
N $\tilde{t}_{\frac{1}{2}}\tilde{t}_{\frac{1}{2}}\tilde{i}$ instrumentos	4
Designa $\tilde{t}_{\frac{1}{2}}\tilde{t}_{\frac{1}{2}}\tilde{o}$	FP1 FP1 FP2 FP2
Elementos da matriz X	CTE H, H4 COSD, SEND T
Per $\tilde{t}_{\frac{1}{2}}\tilde{t}_{\frac{1}{2}}\tilde{o}$ do an $\tilde{t}_{\frac{1}{2}}\tilde{t}_{\frac{1}{2}}\tilde{i}$ lise	[2005-01-01,2013-01-01]
Fim da IQ	2008-01-01

# Recommendations

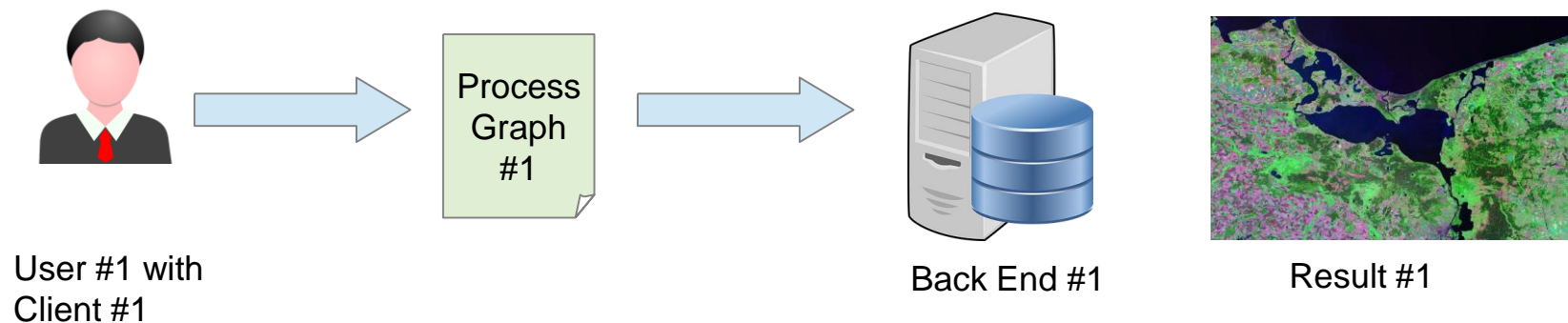
- Analyse dependencies and evade shell calls
  - e.g. use scripting mechanisms provided by the workflow engine
- Write code that runs on all platforms
  - e.g. do not encode specific paths
- Publish experiment setup and context
  - e.g. exact versions of tools used
- Publish validation data
  - e.g. provenance but also other files created during execution
- Test the replicability on your own
  - e.g. try rerunning your experiment in a clean virtual machine

# DOMAIN SPECIFIC DEPLOYMENTS

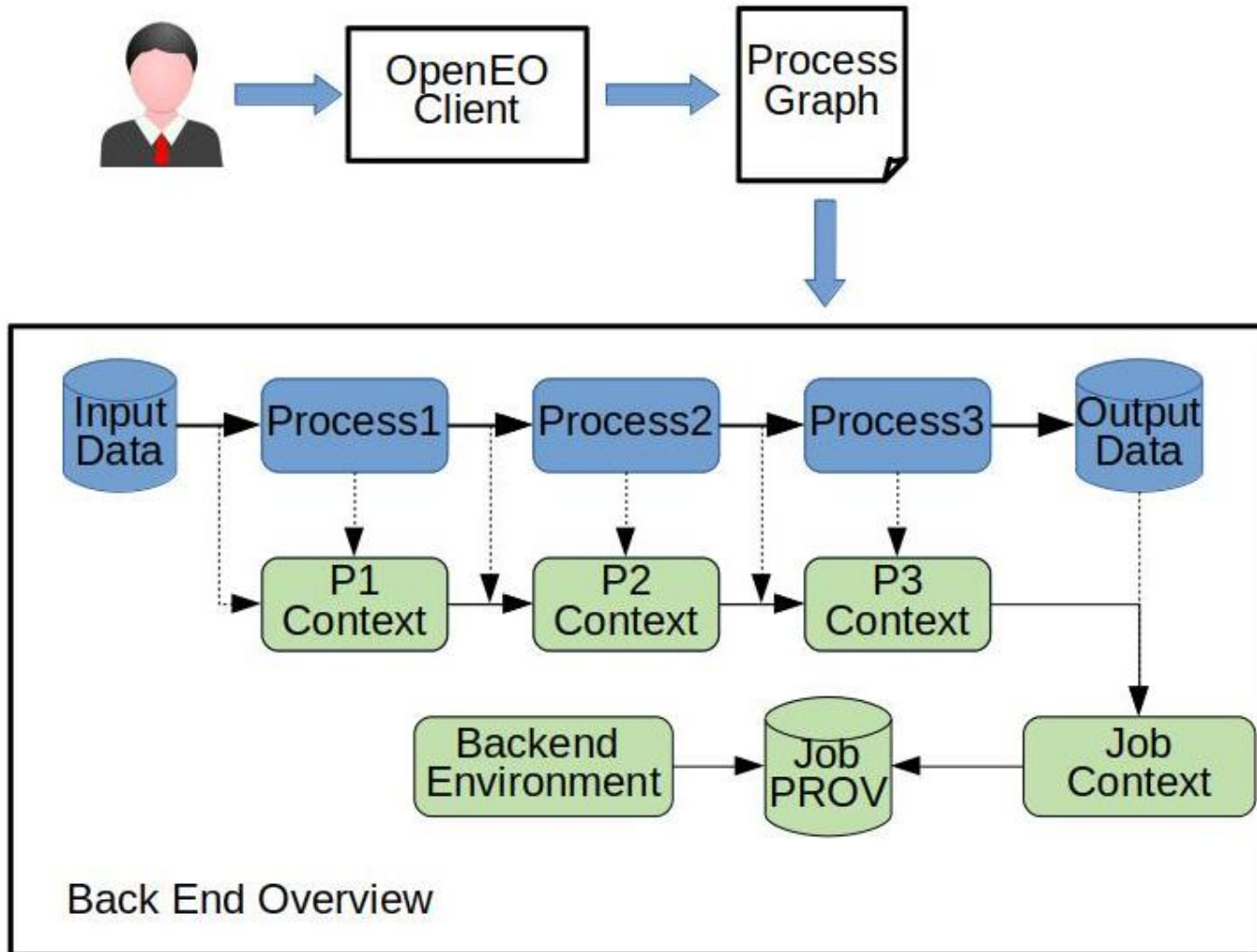


- User defines a process graph using a client library and sends it to the back end of his choice.
- Back end calculates the result and returns it with a download link for the user.

```
{
  "process_graph":{
    "imagery":{
      "imagery":{
        "extent":{
          "2017-01-01",
          "2017-01-31"
        },
        "imagery":{
          "extent":{
            "north":49.041469,
            "east":17.171631,
            "west":9.497681,
            "south":46.517296,
            "crs":"EPSG:32632"
          },
          "imagery":{
            "process_id":"get_collection",
            "name":"s2a_prd_msilic"
          },
          "process_id":"filter_bbox"
        },
        "process_id":"filter_daterange"
      },
      "nir":"B08",
      "process_id":"NDVI",
      "red":"B04"
    },
    "process_id":"min_time"
  }
}
```



# openEO - solution



# openEO – context model

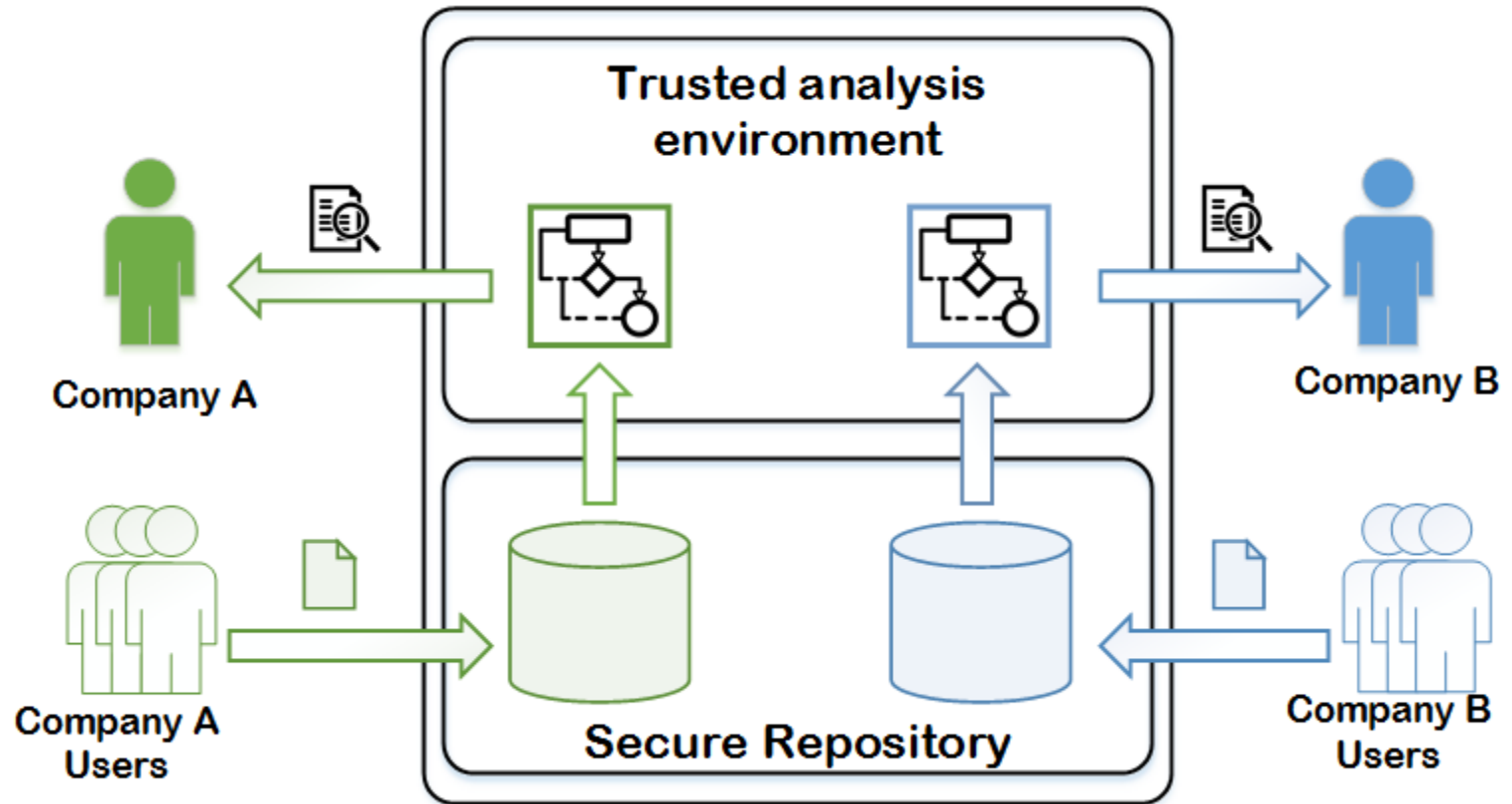
## ■ Static

- Backend version
- Code identifier
- API version
- Publication timestamp

## ■ Dynamic

- Input data PID
- Programming language
- Software dependencies
- Result checksum

```
{
  "backend_env": {
    "backend_version": 2,
    "openeo_api": "0.0.2",
    "git_repos": [
      {
        "branch": "master",
        "commit": "05f4765de578467fef8e1a24404bbc",
        "diff": null,
        "url": "https://github.com/Open-EO/openec
      ]
    },
    "code_env": [
      "surlex==0.2.0",
      "typing==3.6.4",
      ...
    ],
    "interpreter": "Python 3.7.1",
    "input_data": "Q-bc4f15cc-3790-4612-9b29-0a2eb48falce",
    "job_id": "Test",
    "start_time": "2018-10-21 11:27:47,653",
    "end_time": "2018-10-21 11:28:44,644",
    "output_data": "c448e6f8cd027bf0af70bab8d8372b339e708
  }
```



# SUMMARY

# Summary

- Reproducibility and its impact
- Auditability of systems
- Data alone is insufficient
- Process to generate, analyse and render data is needed
- VFramework
  - Documents system set-up and process execution
  - Represents data using context model
  - Can be used as provenance documentation
  - Can be used to verify re-execution
  - Can be used to trace causes for differing behaviour

## If you're looking for a topic...

- We offer topics for theses (or other projects) related to reproducibility and data management
- Various application domains
- No synthetic problems
- Contact us!

# References

- Links included on specific slides
- Tomasz Miksa, Andreas Rauber, and Eleni Mina. Identifying impact of software dependencies on replicability of biomedical workflows. *Journal of Biomedical Informatics*, 64:232 -- 254, 2016.  
<http://dx.doi.org/10.1016/j.jbi.2016.10.011>
- Tomasz Miksa and Andreas Rauber. Using ontologies for verification and validation of workflow-based experiments. *Web Semantics: Science, Services and Agents on the World Wide Web*, 43(Supplement C):25 -- 45, 2017. <https://doi.org/10.1016/j.websem.2017.01.002>