

# Reproducible Science in Bioinformatics: Scientific workflows, Provenance and beyond

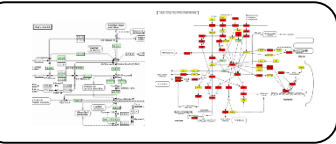
**Sarah Cohen-Boulakia**

*On leave at INRIA Virtual Plants & Zenith, IBC Montpellier*  
Université Paris Sud, LRI CNRS UMR 8623

*Center for Data Science, October 26<sup>th</sup> 2015*

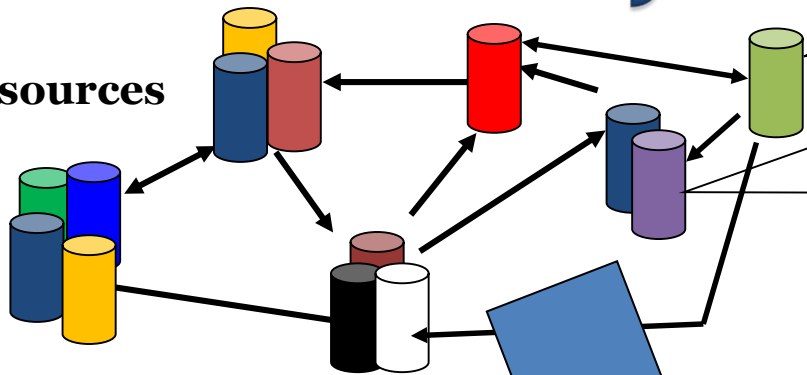


# Bioinformatics analysis



## Data from public sources

- Heterogeneous
  - Distributed
- >1,500 databases



```
CCCTTTCCGTGT
G TCCCGTCTCCG
G T
TGCCGTGTGGC
TAAATGTCTGTG
...
GTCTGTGC...
```

(NAR databases issue)

Tools

Scripts

Python



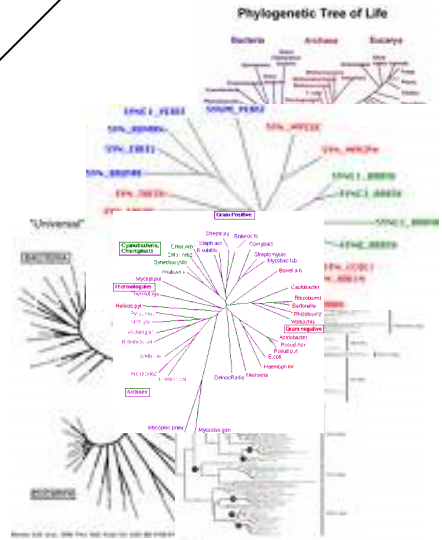
JAVA, Perl

Web services

...

How these data have been generated?  
With which input data? Which tools? Which parameters?

- ### Tools
- Heterogeneous
  - To be Combined



Workspace

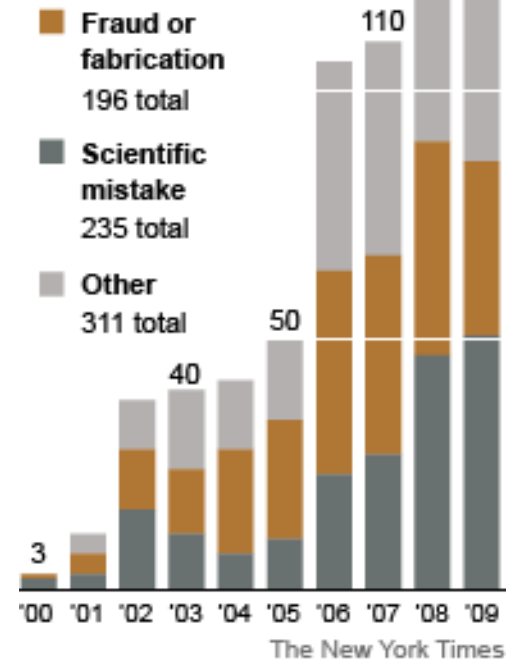
# Take Home Message

Compared to 15 years ago...

- ▶ The number and **diversity of the sources** has increased a lot
- ▶ The **complexity of the pipelines to be designed** has increased a lot
  - increasing difficulties to reproduce experiments!
- ▶ Studies demonstrated low reproducibility of several major scientific results...
- ▶ **Huge impacts** (paper retractions, preclinical studies...)
- ▶ Having access to reproducible results also help **assessing data quality**

## Retractions On the Rise

A study of the PubMed database found that the number of articles retracted from scientific journals increased substantially between 2000 and 2009.

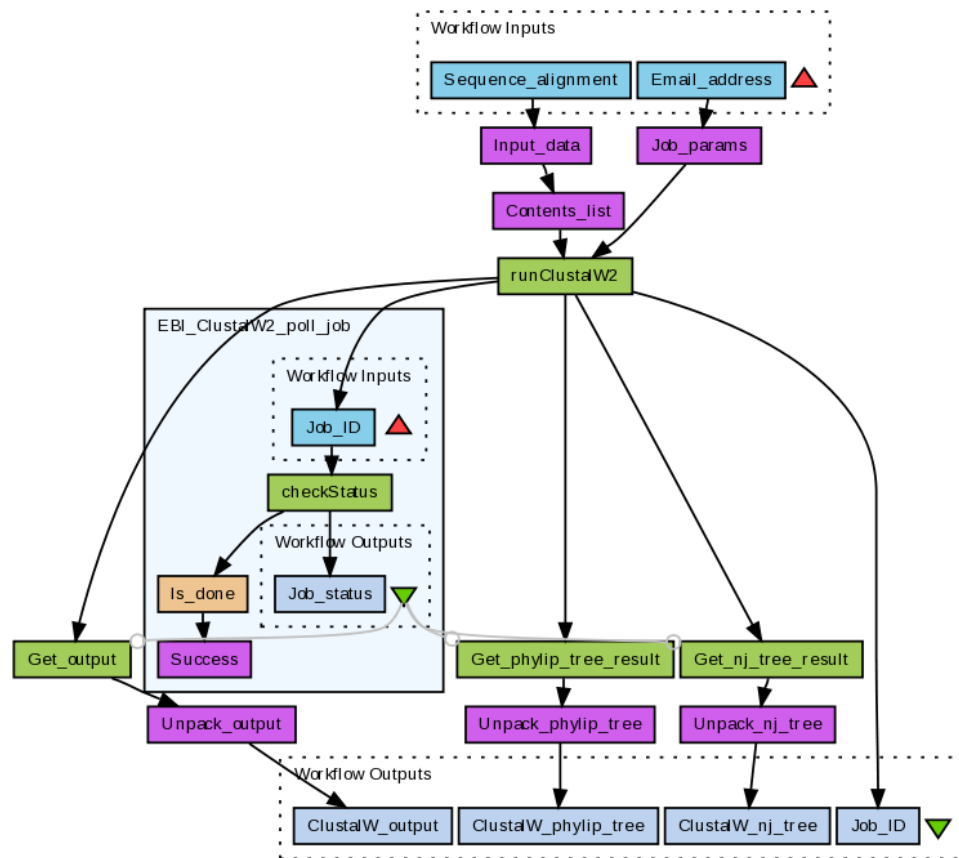


# Outline

- ▶ Context
  - ▶ Scientific workflows, Provenance...
    - Scientific workflow reuse
    - (Re)designing workflows
  - ▶ ... and beyond
    - NoteBooks
    - Virtual machines
  - ▶ Conclusions
- } Newer  
(ongoing work)

# Scientific Workflow (Management System)

- ▶ Experiments
    - “Data analysis pipeline”
    - Chained tools (modules)
    - Graph structure
    - Data flow driven
  - ▶ SWFS manage
    - Workflow design (GUI)
    - Scheduling
    - Logging (provenance)
      - Recursive history of the data
    - Recovery, ...
  - ▶ Several systems
    - Taverna, Kepler, OpenAlea, Vistrails, Galaxy...
- Workflows are born to be reused!



Phylogenetic workflow from Taverna

# Scientific Workflow Repositories

 myexperiment

 Galaxy

 crowdLabs

 Kepler



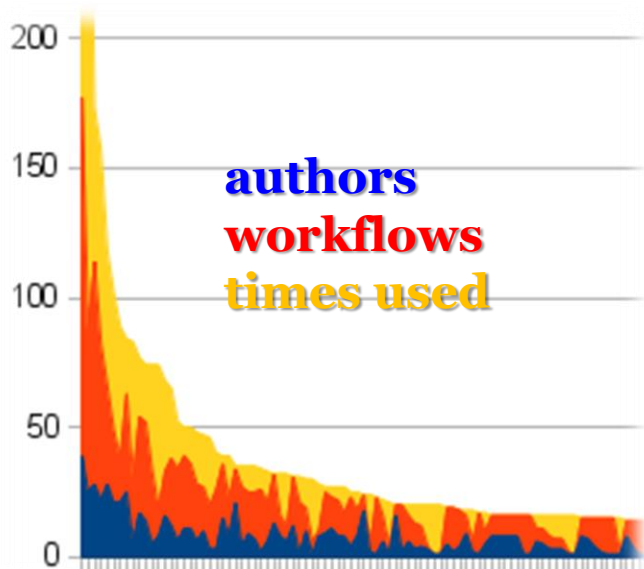
- Upload & annotate scientific workflows
- Search, download & use existing workflows
- Today repositories contain several thousand workflows

# Study on workflow reuse

[SSDBM 2010]

With Ulf Leser &  
Johannes Starlinger

- Based on 1,700 Taverna workflows (myExperiment)
- 36% of elements are re-used
  - connect workflows quite densely
- True cross-author re-use is low: 3%



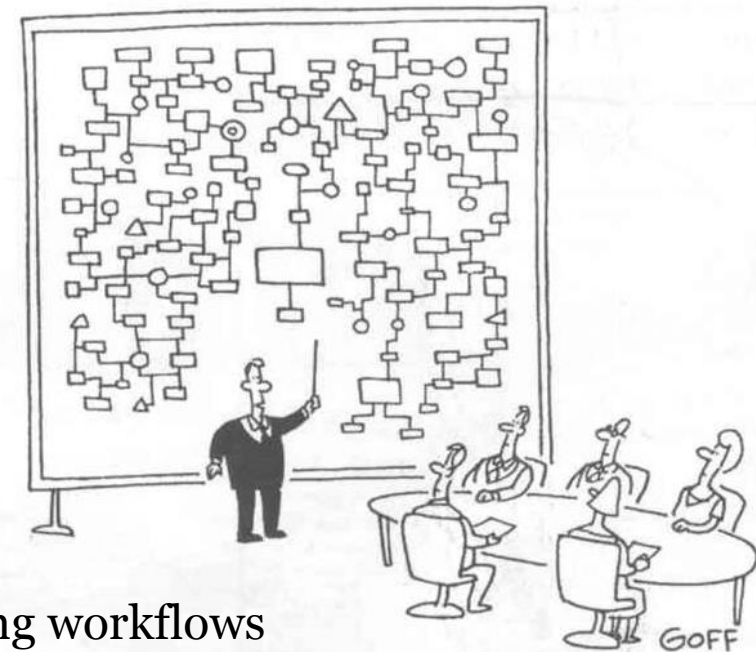
Distinct modules

- Re-use rates have a **Zipf-like distribution**
  - Using information about types of processors
  - **Local** : High re-use rates as-is
  - **Web-Service** : Authors have favorite services, unshared

# How to improve reuse?

Help finding  
*similar*  
*workflows*

Make  
workflow  
structures  
less complex!



Plumbing workflows



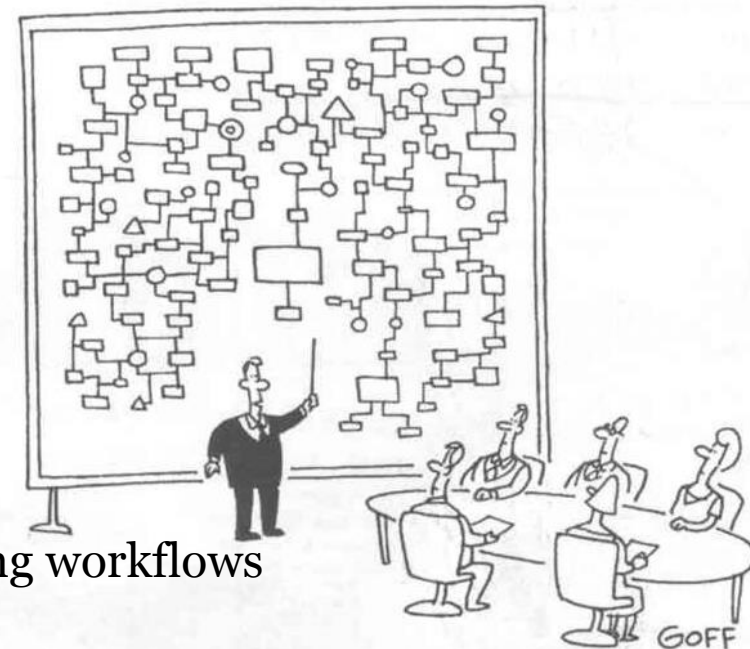
# How to improve reuse?

Help finding  
*similar*  
*workflows*



2 projects  
→ ZOOM  
→ DistillFlow

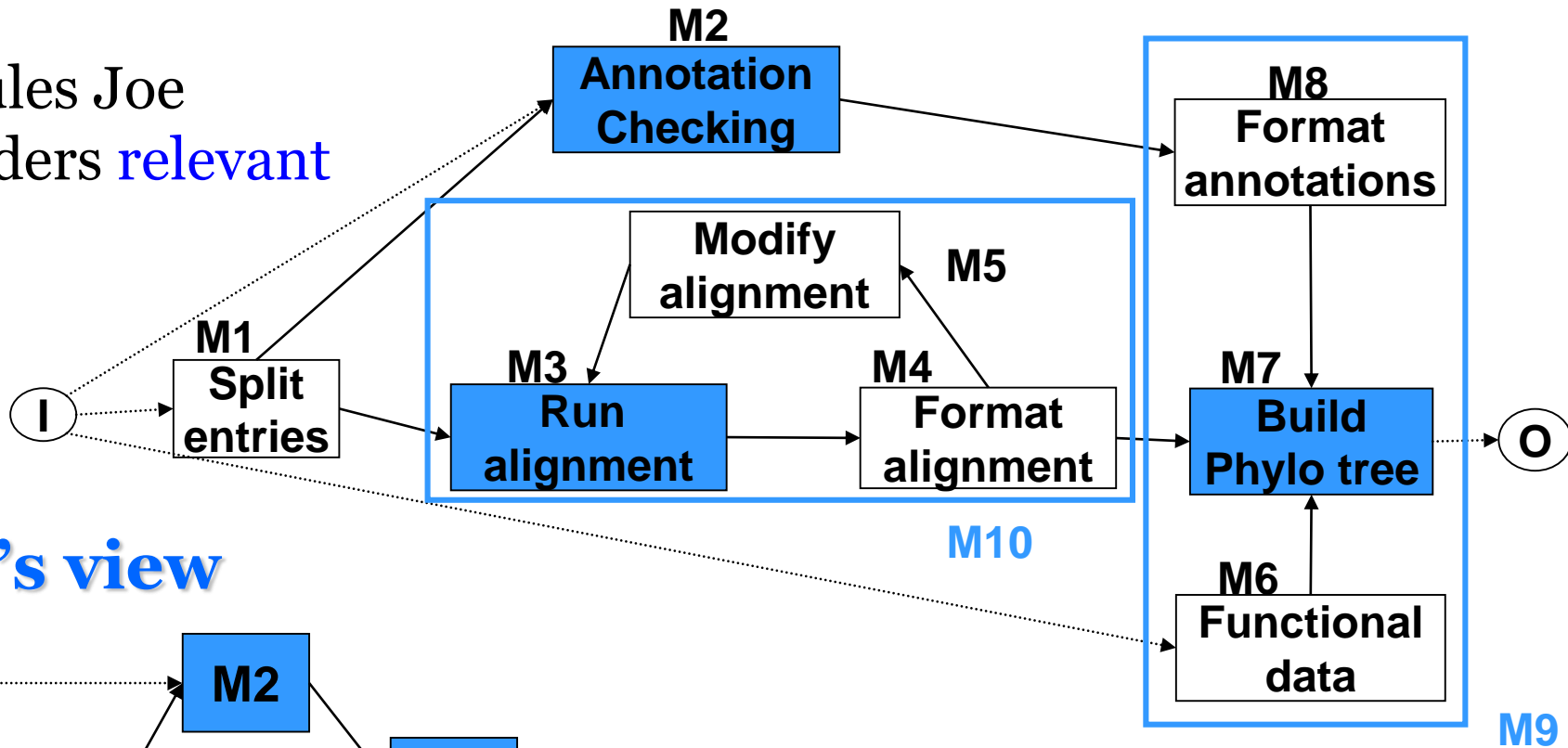
Make  
workflow  
structures  
less complex!



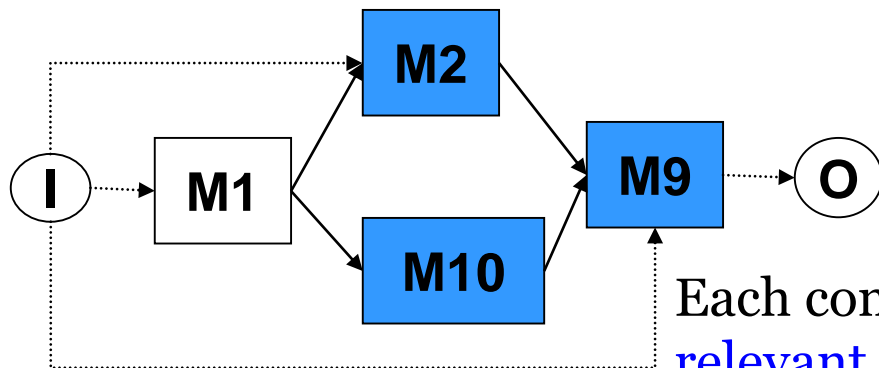
Plumbing workflows

# ZOOM: Using Composite modules

Modules Joe considers **relevant**

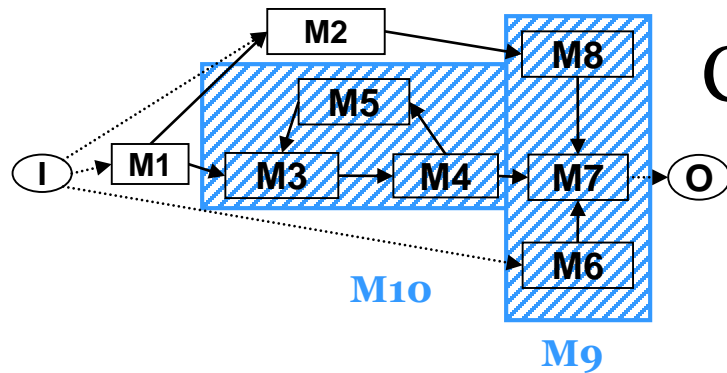


**Joe's view**

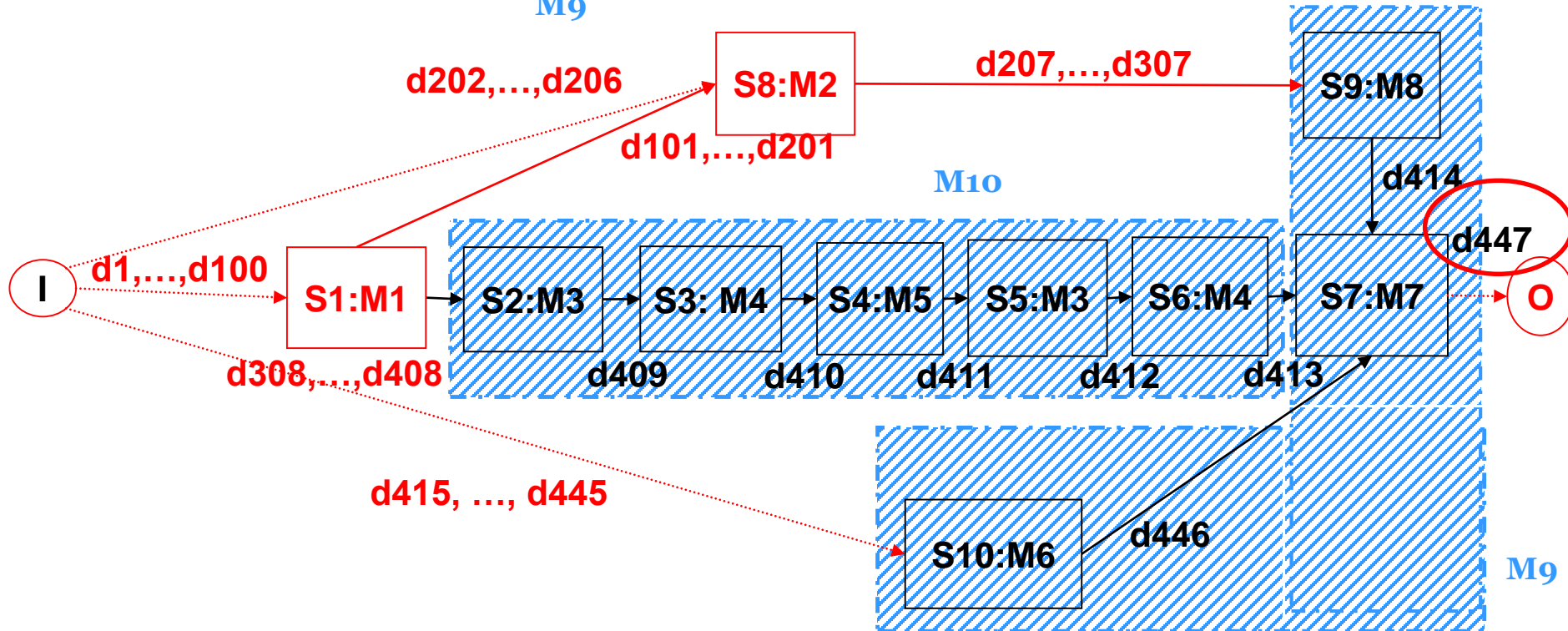


Each composite module takes the **meaning** of the **relevant** module it contains

# Impact on provenance overload

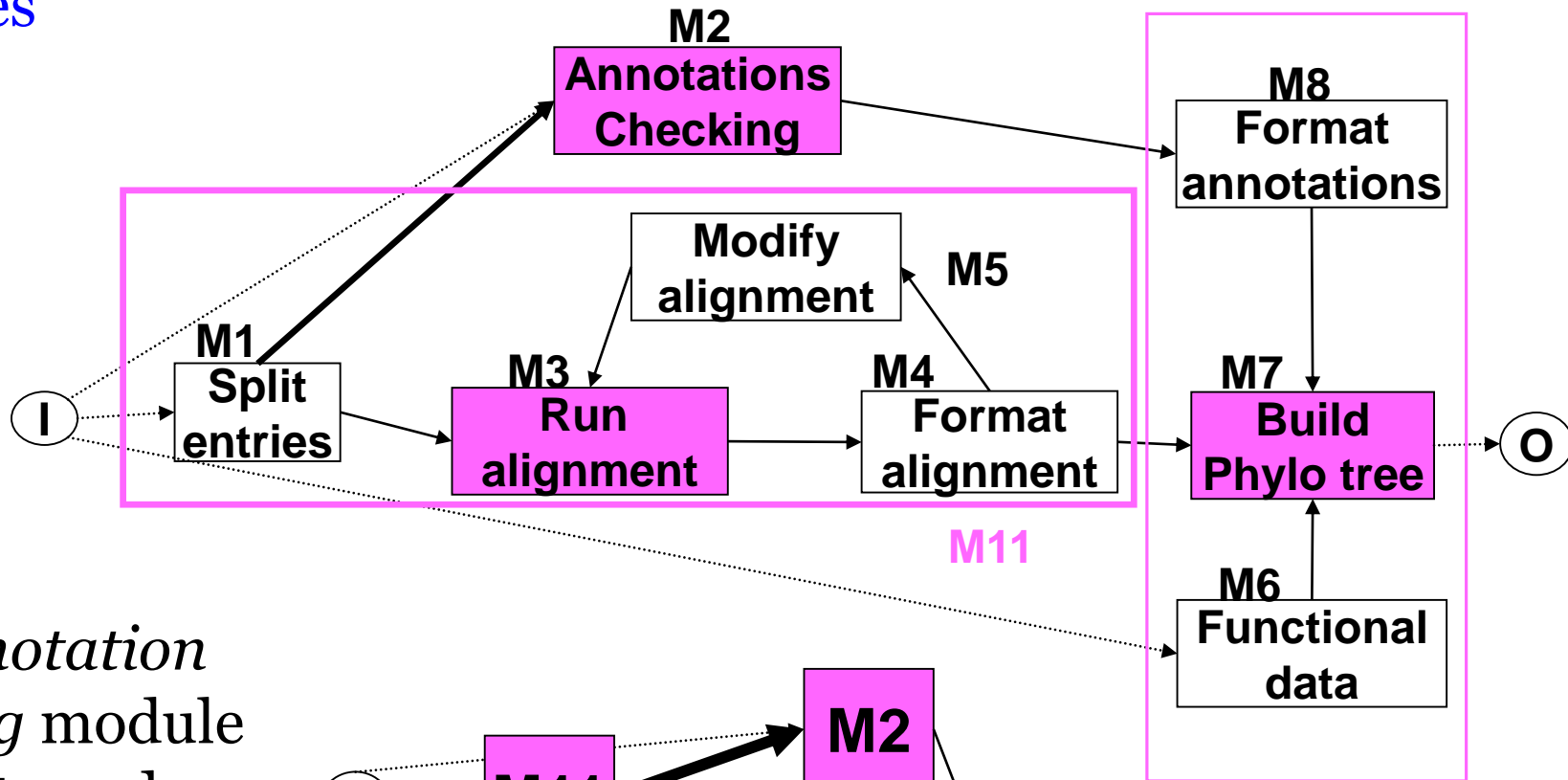


Composition simplifies provenance!

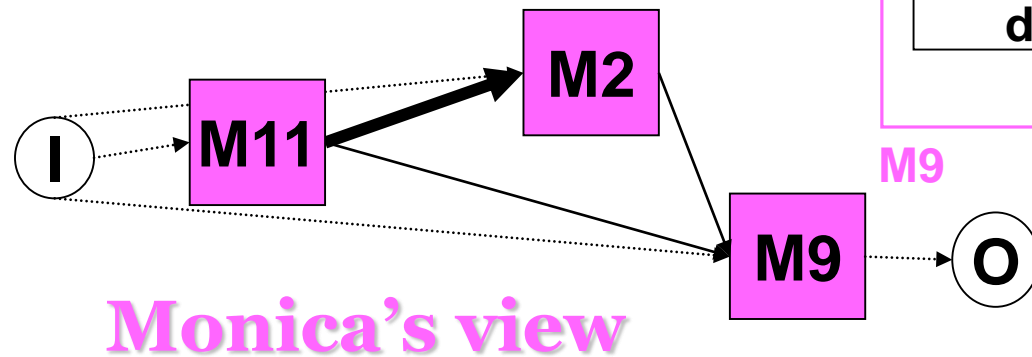


# Grouping may be error-prone!

Grouping should **preserve the relationships** between **relevant modules**



The *annotation checking* module does not need input from the *run alignment* module!



## ▶ **Results [ICDE 08]**

- **Formalization** of the set of properties to be preserved
  - **Property 1:** Given  $G_w$  and  $R \subseteq N$  relevant modules,  **$U$  is well-formed** iff every composite module in  $U$  contains at most one element of  $R$ .
  - **Property 2:** A user view  $U$  **preserves dataflow** iff every edge in  $G_w$  that induces an edge on an  $nr$ -path from  $C(r)$  to  $C(r')$  in  $U(G_w)$  lies on an  $nr$ -path  $r$  to  $r'$  in  $G_w$ .
  - **Property 3:** A user view  $U$  **is complete w.r.t dataflow** iff for every edge  $e$  on an  $nr$ -path from  $r$  to  $r'$  in  $G_w$  that induces an edge  $e'$  in  $U(G_w)$ ,  $e'$  lies on an  $nr$ -path from  $C(r)$  to  $C(r')$ .
- **Theorem:** *ZOOM* is a **polynomial-time** which *preserves Properties 1- 3 and produces a minimal user view*

## ▶ **Implementation** of ZOOM

- Taverna, BerkeleyDB/Oracle **[VLDB 07 (demo)]**
- Used within the **1<sup>st</sup> Provenance Challenge [CCPE Journ. 07]**

## ▶ **Numerous reuses** of ZOOM

- 200+ citations

# DistillFlow: removing redundancy

- ▶ Collaboration with Taverna & BioVel
- ▶ BioVel (FP7)

With Ch.  
Froidevaux, C.  
Goble, P. Missier,  
A. Williams, **J.  
Chen**

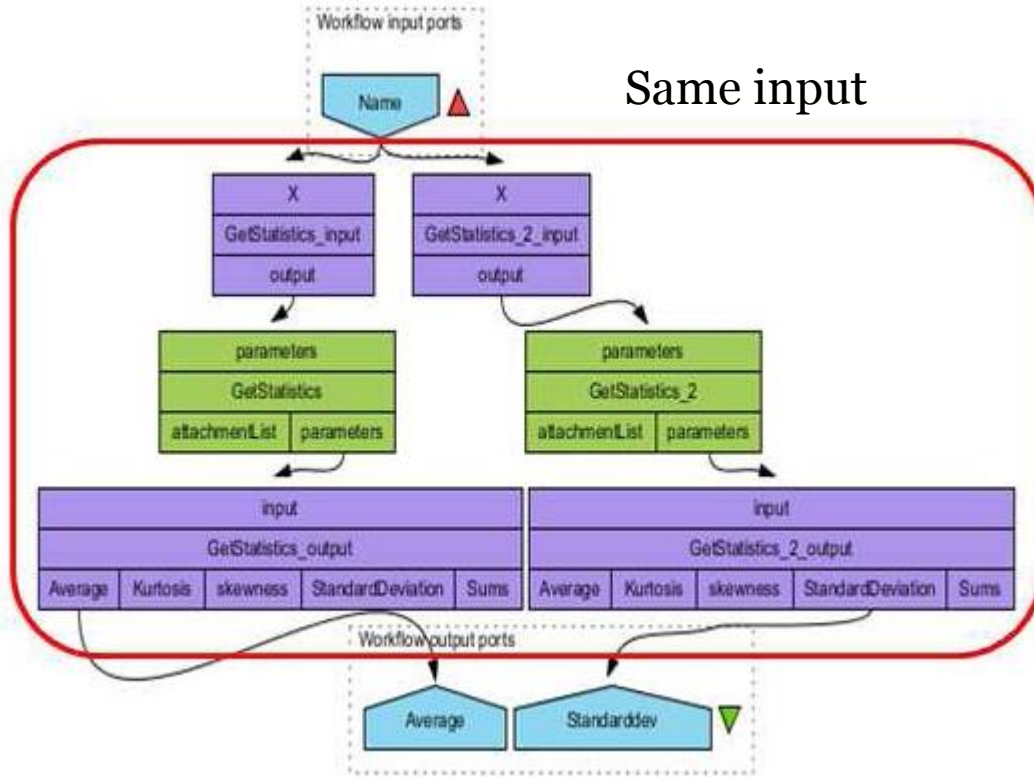


- Virtual laboratory: Libraries of workflows for research on biodiversity
- Consortium of 15 partners (9 countries)

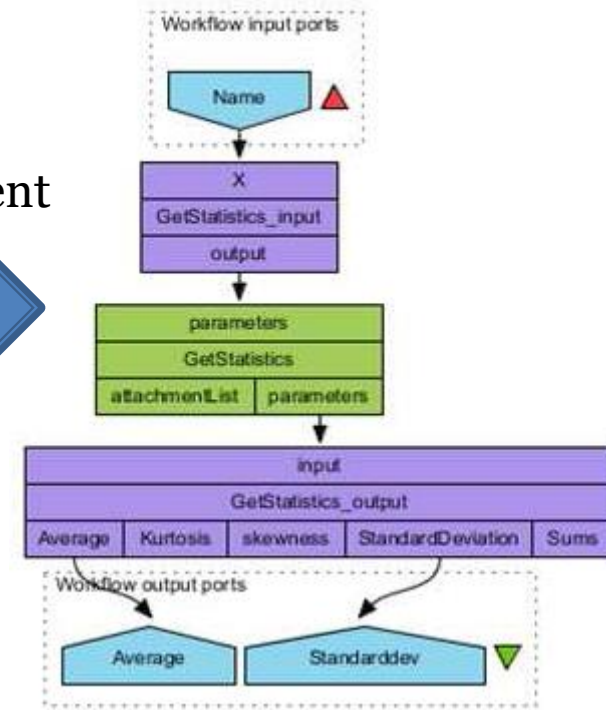
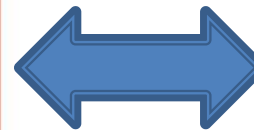
- Understanding reuse based on BioVel workflows
- More generally: improving reuse in Taverna

**Distilling workflow structures: Removing redundancy**

# Example of use case



Equivalent



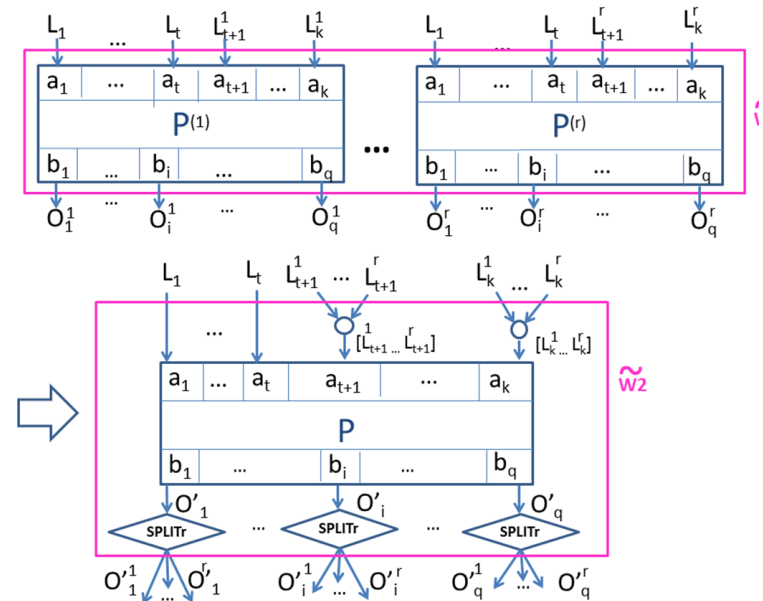
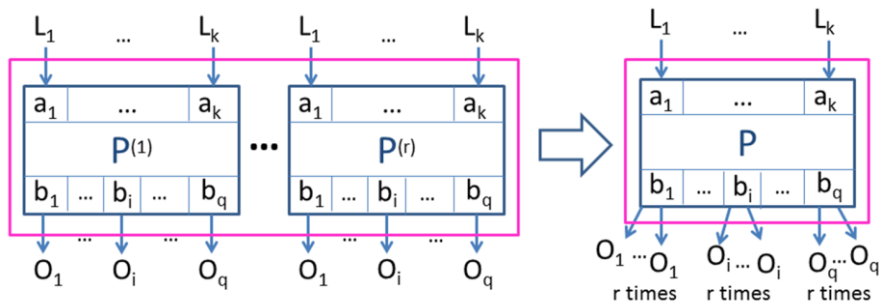
3 processors duplicated!  
→ Pure redundancy

No redundancy

## Other use cases have been considered

# Rewriting workflows

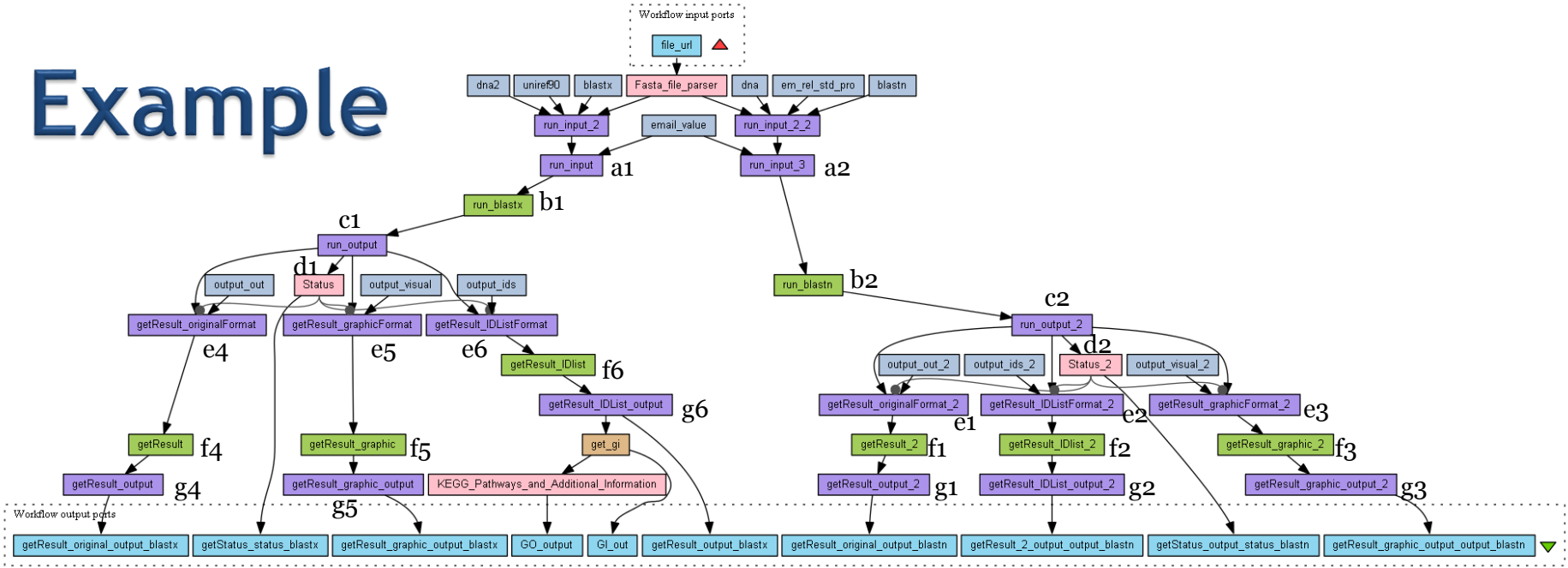
- ▶ Exploiting the **implicit iteration** feature of Taverna
  - ▶ Assumptions before merging copies of a processor
    - exact same code, do not depend on each other, **deterministic** processors
- **Anti-patterns** and the corresponding rewritings  
(concept from the software engineering community)



**[BMC Bioinformatics 2014]**  
 With Carole Goble, Paolo Missier, Jiuqiang Chen  
 (PhD student co-supervised), Christine Froidevaux

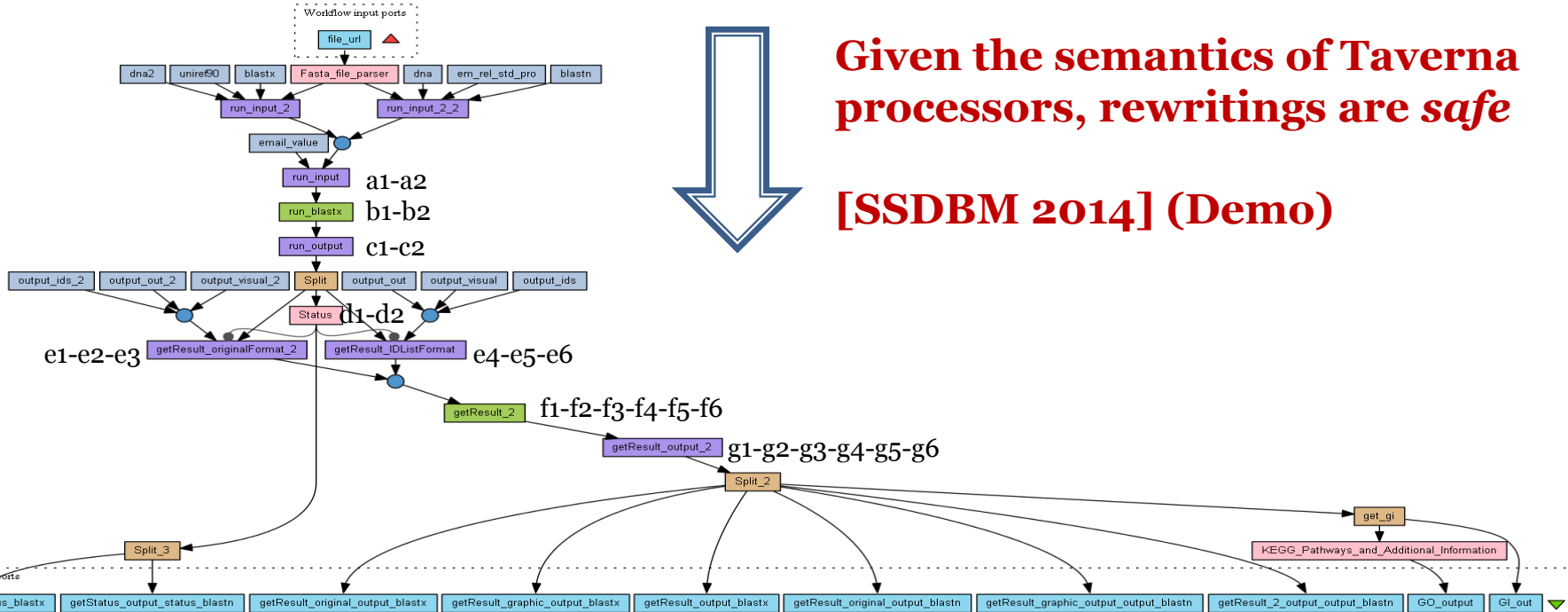


# Example



**Given the semantics of Taverna processors, rewritings are *safe***

**[SSDBM 2014] (Demo)**



# Outline

- ▶ Context
  - ▶ Scientific workflows and Provenance
    - Scientific workflow reuse
    - (Re)designing workflows
  - ▶ ... and beyond
    - NoteBooks
    - Virtual machines
  - ▶ Conclusions
- } Ongoing work

# NoteBooks and Workflows

## ▶ Notebook

- Web-based **interactive computational environment**
- Combination of code execution and rich media **into a single document**

## ▶ Coupling scientific workflows & notebooks

- Notebooks to **document workflows**
- VisTrails, Taverna, Galaxy

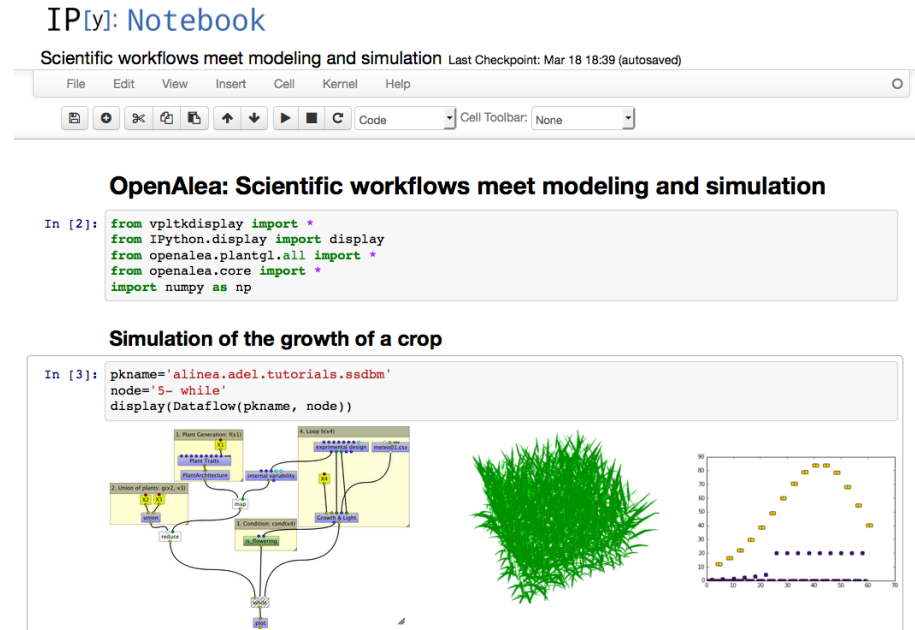
## ▶ **OpenAlea**

- Scientific workflow system

## Workflow **executions** saved into **notebooks**

- Actors of the workflow → cells in the notebook
- Data produced and used (execution) can be visualized

With **Ch. Pradal**, Ch. Fournier, P. Valduriez



IP[y]: Notebook

Scientific workflows meet modeling and simulation Last Checkpoint: Mar 18 18:39 (autosaved)

File Edit View Insert Cell Kernel Help

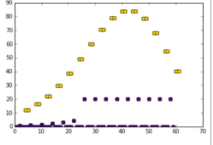

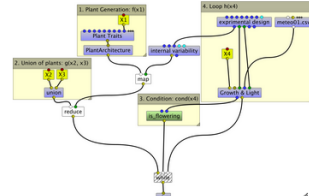
Cell Toolbar: None

### OpenAlea: Scientific workflows meet modeling and simulation


```
In [2]: from vpltkdisplay import *
from IPython.display import display
from openalea.plantgl.all import *
from openalea.core import *
import numpy as np
```

### Simulation of the growth of a crop

```
In [3]: pkname='alinea.adel.tutorials.ssdsm'
node='5- while'
display(Dataflow(pkname, node))
```

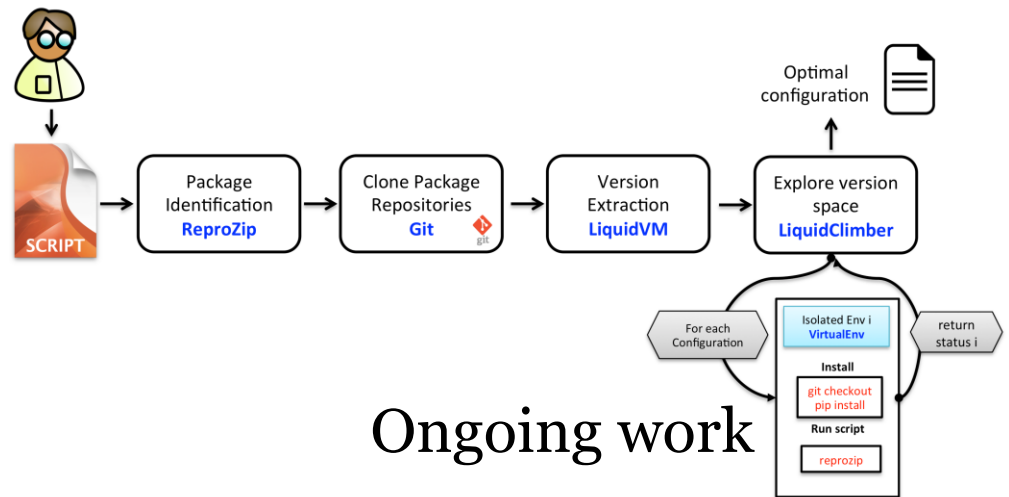


# Virtual machines

- ▶ VM capture the programming environment
- ▶ Docker  docker
- ▶ ReProZip (NewYork University)
- ▶ Can be used to capture the **workflow system** environment

- ▶ Toward *Liquid VM* ?

- *Defrost* your VM
  - Update packages as much as possible (stop when it does not work anymore)



With **Ch. Pradal**, **D. Shasha**, **P. Valduriez**

# Conclusion

- ▶ Too many scientific results are not reproducible
- ▶ Mature solutions exist and are able to solve a large number of cases, now need to combine them (several communities)
- ▶ Same problematics in other domains (e.g., astronomy... and even computer science!)
- ▶ Several Initiatives: Force 11, Data and Software Carpentry



Please contact me  
[cohen@lri.fr](mailto:cohen@lri.fr)

# Thanks!



Workflow reuse and workflow similarity  
with U. Humboldt, Berlin  
(U. Leser, J. Starlinger)



Workflow refactoring  
Ch. Froidevaux, J. Chen  
with U. Manchester (Taverna group,  
C. Goble)



Reducing workflow complexity (ZOOM)  
with Univ. Pennsylvania  
(S. Davidson, S. Khanna)

OpenAlea development group  
& Notebooks (Ch. Pradal, Ch.  
Fournier)  
with P. Valduriez (Inria) & P.  
Neveu (INRA)  
with D. Shasha (NYU/Inria)

