Reproductibilité Computationnelle d'analyses de données biologiques massives

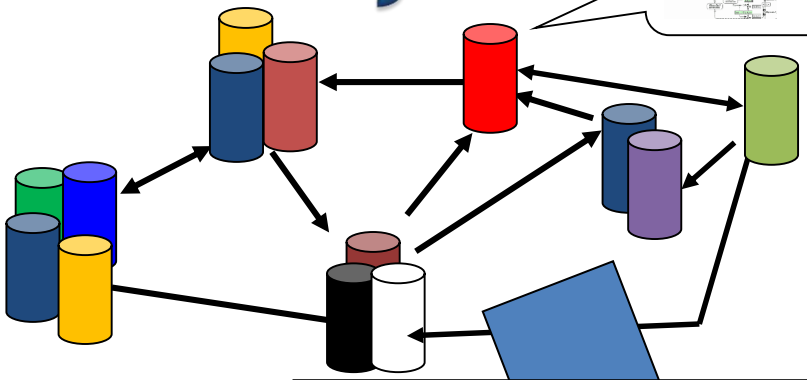
Sarah Cohen-Boulakia

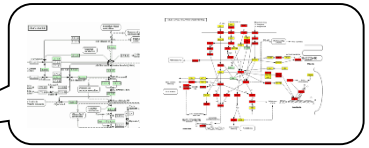
Université Paris-Sud, Laboratoire de Recherche en Informatique
CNRS UMR 8623, Université Paris-Saclay, Orsay, France

Biological analysis

Public sources

- Distributed
- Heterogeneous
- Network





CCCTTTCCGTGT
TGCCGTGTGGC
TAAATGTCTGTG
GTCTGTGC...

Tools
Scripts
Python

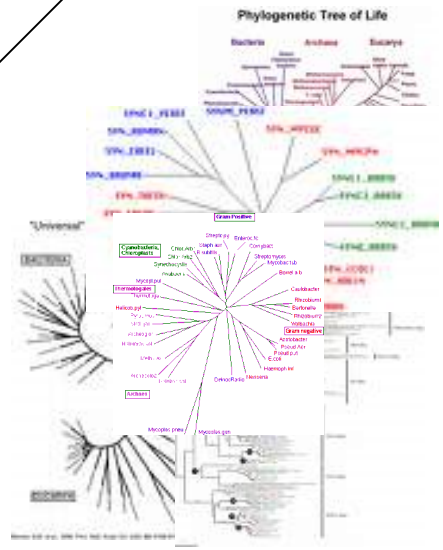
JAVA, Perl
Web services
...

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

Tools

- Distributed
- Heterogeneous
- Chained

What is the difference between these two experiments?



Workspace

Take Home Message

Compared to 20 years ago…

- The number and diversity of the sources has increased a lot
  - > 1,500 databases (NAR databases issue)
  - Need for data provenance to determine data quality
- The complexity of the pipelines to be designed has increased a lot
  - Need for process provenance to determine data quality

→Increase in the heterogeneity of data
+ Increase in the complexity of analysis pipelines
+ *Increase in the need to publish…*
= increasing difficulties to reproduce experiments!

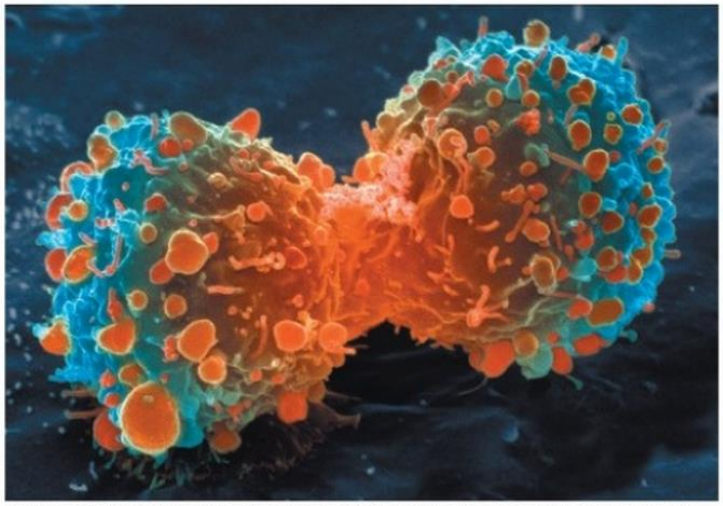
Studies on reproducibility

- Nekrutenko & Taylor, Nature Genetics (2012)
  - 50 papers published in 2011 using the Burrows-Wheeler Aligner for Mapping Illumina reads.
  - 31/50 (62%) provide no information
    - no version of the tool + no parameters used + no exact genomic reference sequence
  - 7/50 (14%) provide all the necessary details

Studies on reproducibility

- Nekrutenko & Taylor, Nature Genetics (2012)
  - 50 papers published in 2011 using the Burrows-Wheeler Aligner for Mapping Illumina reads.
  - 31/50 (62%) provide no information
    - no version of the tool + no parameters used + no exact genomic reference sequence
  - 7/50 (14%) provide all the necessary details
- Alsheikh-Ali et al, PLoS one (2011)
  - 10 papers in the top-50 IF journals → 500 papers (publishers)
    - 149 (30%) were not subject to any data availability policy
      (0% made their data available)
    - Of the remaining 351 papers
      - 208 papers (59%) did not adhere to the data availability instructions
      - 143 make a statement of *willingness* to share
      - 47 papers (9%) deposited full primary raw data online

Impacts of irreproducibility…



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

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

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

Many factors are responsible for the h failure rate, notwithstanding the inh ently difficult nature of this disease. C tainly, the limitations of preclinical te

47/53 “landmark” publications could not be replicated

[Begley, Ellis Nature, 483, 2012]

Must try harder

Too many sloppy mistakes are creeping into scientific papers. at the data — and at themselves.

Error prone

Biologists must realize the pitfalls massive amounts of data.

If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring that results are reproducible, argues **Jonathan F. Russell**.

The case for open computer programs

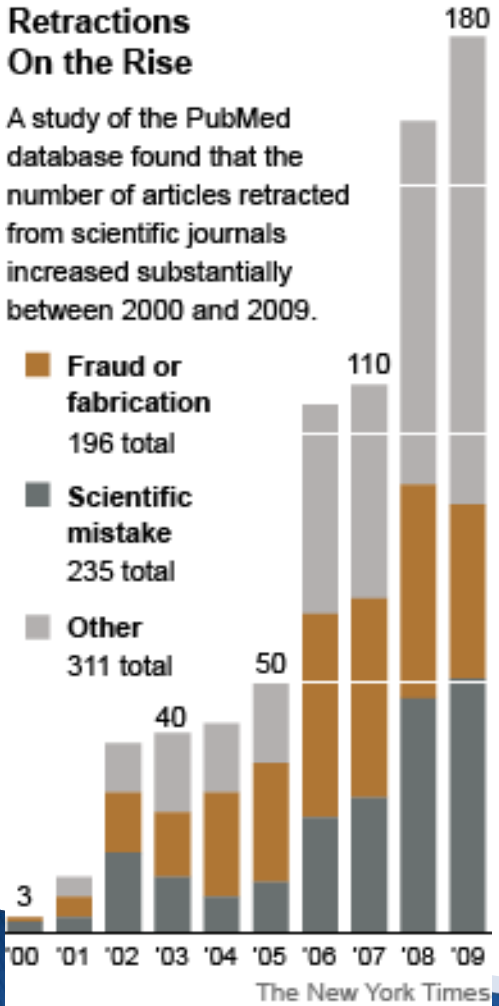
Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won’t stand up.

Know when your numbers are significant

Impacts of irreproducibility (cont.)

- Attacks on authors, editors, reviewers, publishers, funders…



http://www.nature.com/nature/focus/reproducibility/index.html

→ *Nature* checklist
→ *Science* requirements for data and code availability

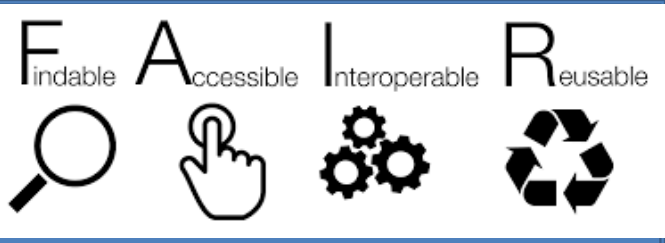
Ten Simple Rules for Reproducible Computational Research (PlosOne)

- 1: For Every Result, Keep Track of How It Was Produced
- 2: Avoid Manual Data Manipulation Steps
- 3: Archive the Exact Versions of All External Programs Used
- 4: Version Control All Custom Scripts
- 5: Record All Intermediate Results, When Possible in Standardized Formats
- 6: For Analyses That Include Randomness, Note Underlying Random Seeds
- 7: Always Store Raw Data behind Plots
- 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- 9: Connect Textual Statements to Underlying Results
- 10: Provide Public Access to Scripts, Runs, and Results

→ Several ways to follow them

→ More or less complex (from manually to fully automatically)

→ More or less time-consuming (repeat, reproduce, …., reuse)



- Findable
  - (meta)data assigned a globally unique and eternally persistent identifier.
  - (meta)data registered or indexed in a searchable resource.
- Accessible
  - (meta)data retrievable by their identifier using a standardized communications protocol.
- Interoperable
  - (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - (meta)data use vocabularies that follow FAIR principles.
- Re-usable
  - (meta)data are released with a clear and accessible data usage license.
  - (meta)data are associated with their provenance.
  - (meta)data meet domain-relevant community standards.

Aims of our action

- **Action in the GDR MaDICS** (since 2015)
- **Concepts, Needs/solutions**
  - Which *levels* of reproducibility can we consider?
  - Which are the solutions (methods and tools) currently available for *reproducibility*?
- **Opportunities, challenges**
  - What is missing?
  - Which are the *research* (vs technical) *open issues*?
- **Evaluation of solutions** based on **practice and state-of-the-art**
  - Experience of developers in using solutions in real contexts
  - ReproHackathon

    → Real use cases from the Bioinformatics Domain

Biological Data Analysis

From Data to Knowledge

Data

Distributed, **Heterogeneous**

Tools

Different kinds, various parameters

Analysis pipelines (*workflows*)

Complex

Use cases

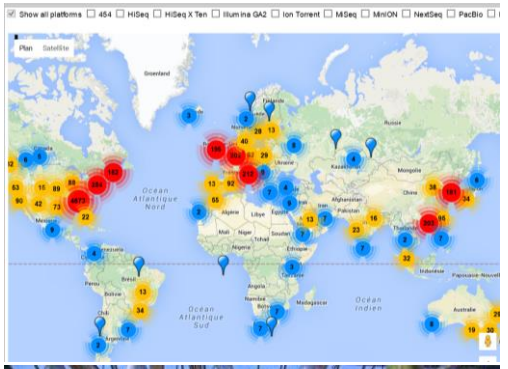
NGS (cancer), Plant Phenotyping

Big data sets

European Research Infrastructure

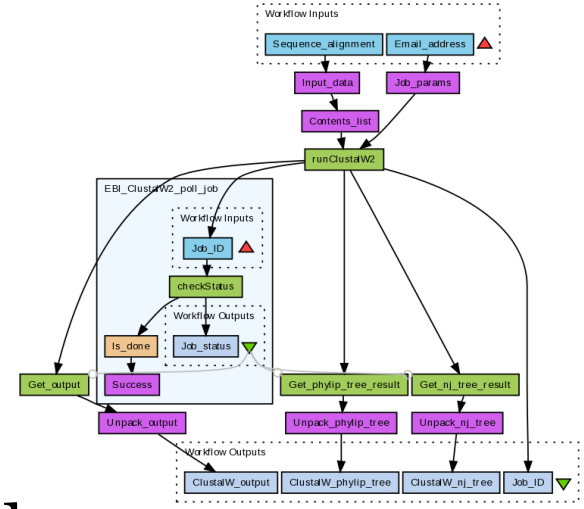
21 countries, 180 partners

➔ Analyses with scientific workflow systems



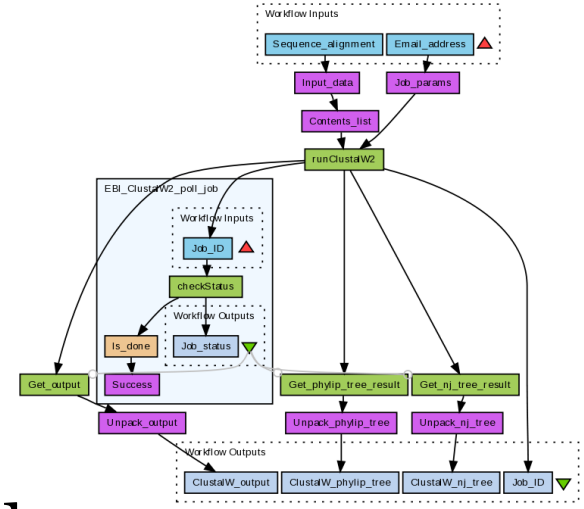
Scientific workflow systems

- SWFS = WFS for scientific tasks
  - “Data analysis pipeline”
- Complex pipelines are broken into tasks and their connection
- Data flow driven
- Tasks can be executed locally or distributed
- SWFS manages scheduling, process control, logging, recovery, reproducibility, …
- Equipped with graphical workflow designer
- Several systems available (Galaxy, NextFlow, SnakeMake, OpenAlea…)



Scientific workflow systems

- SWFS = WFS for scientific tasks
  - “Data analysis pipeline”
- Complex pipelines are broken into tasks and their connection
- Data flow driven
- Tasks can be executed locally or distributed
- SWFS manages scheduling, process control, logging, recovery, reproducibility, …
- Equipped with graphical workflow designer
- Several systems available (Galaxy, NextFlow, SnakeMake, OpenAlea…)

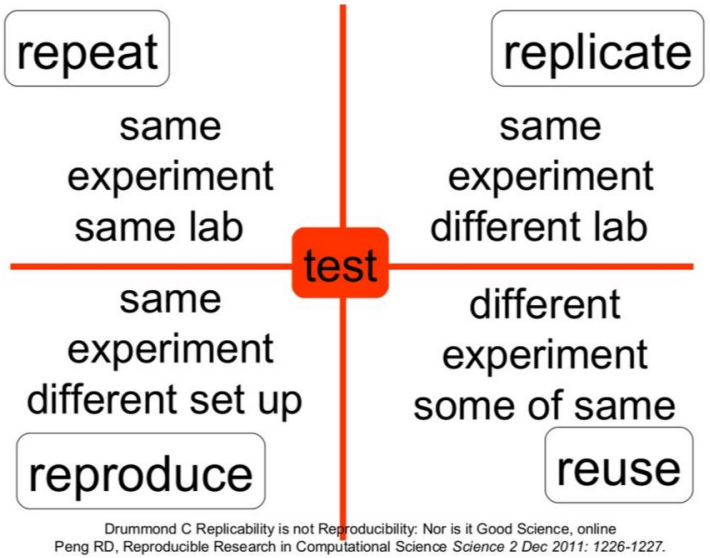


Which reproducibility levels when using workflow systems?
Which features for a *reproducibility-friendly* workflow system?

Outline

- Context
- Levels of reproducibility in scientific workflow systems
- Reproducibility-friendly features
- Open challenges

A continuum of possibilities



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

- Repeat
  - *Redo*: exact same context
  - Same workflow, execution setting, environement
  - Identical *output*

→Aim = proof for reviewers ☺

3 ingredients

Workflows Specification

Chained Tools

Workflow Execution

Input data and parameters

Environment

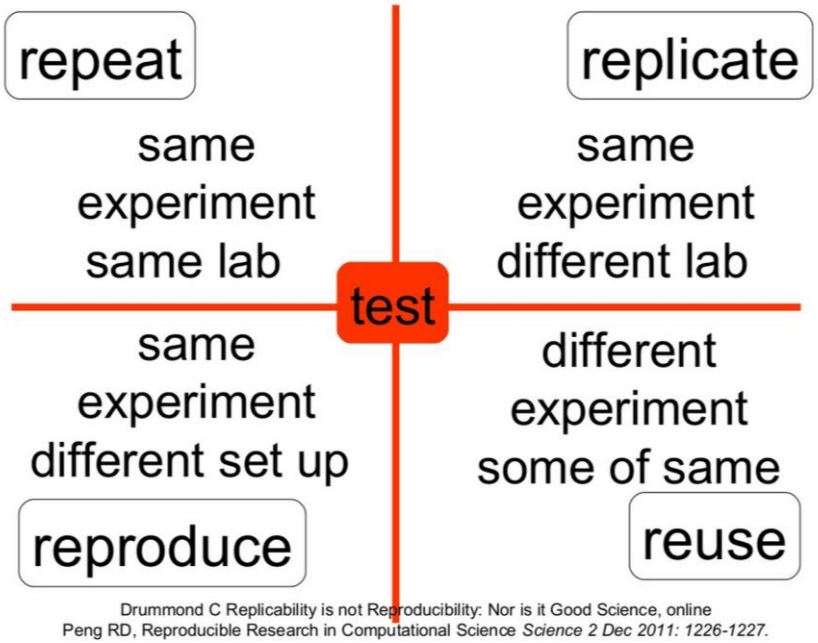
OS/librairies installed…

- Replicate
  - Variation allowed in the workflows, execution setting, environement
  - Similar *output*

→ Aim = robustness

A continuum of possibilities

- Reproduce
  - Same *scientific result*
  - But the means used may be changed
  - Different workflows, execution setting, environment
  - Different output but in accordance with the result
- Reuse
  - Different scientific result
  - Use of tools/… designed in another context



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

Outline

- Context
- Levels of reproducibility
- Reproducibility-friendly features
- Open challenges

Reproducibility-friendly features in scientific workflows

5 Systems: Galaxy, VisTrails, Taverna, OpenAlea, NextFlow

Workflow specification

- Language (XML, Python…) → repeat … reuse
- Interoperability (CWL…) → replicate … reuse
- Description of steps
  - Remote services → repeat
  - Command line → repeat … reuse
  - Access to source code → replicate
- Modularity (nested workflows?) → reuse
- Annotation (tags, ontologies…) → reuse

Execution

- Language and standard (PROV…,) → repeat … reuse
- Presentation (interactivity with the results/provenance, notebooks) → replicate … reuse
- Annotations → reuse

Reproducibility-friendly features in scientific workflows (cont.)

Environment (companion tools)

Ability to run workflows within a given environment → repeat (… reuse)

Virtual machines capture the programming environment

- Package, *freeze,* and expose the environment
- VMWare, KVM, VirtualBox, Vagran,…

Lighter solutions (containers)

- Only capture software dependencies
- Docker, Rocket, OpenVZ, LXC, Conda

Capturing the command-line history, input/output, specification

CDE, ReproZip (NewYork University)

Our new concept: ReproHackathon

- Hackathon
  - Several developers in the same room
  - Same goal to achieve (e.g., predicting plants images)
  - Create useable software in a short amount of time
  - Aim: Demonstrating feasability
- ReproHackathon
  - A hackathon where
    - Given a scientific publication + input data (+ possibly contacts with authors)
    - Several (groups of) developers reimplement the methods to try to get the same result
  - Aim : Ability of current tools to reproduce a scientific result

The first edition of ReproHackathon

- RNA-Seq data from patients with uveal melanoma: genes involved
- Divergent published results…
- June 1-2, Gif s/Yvette, 25 participants (IGRoussy, Curie, Pasteur, Saclay, Paris, Nantes, Lyon, …)

https://ifb-elixirfr.github.io/ReproHackathon/hackathon_1.html

Systems : SnakeMake, NextFlow, iPython notebooks, Galaxy, scripts…
Executed in the Cloud@IFB

Testing several levels of reproducibility: repeat and replicate

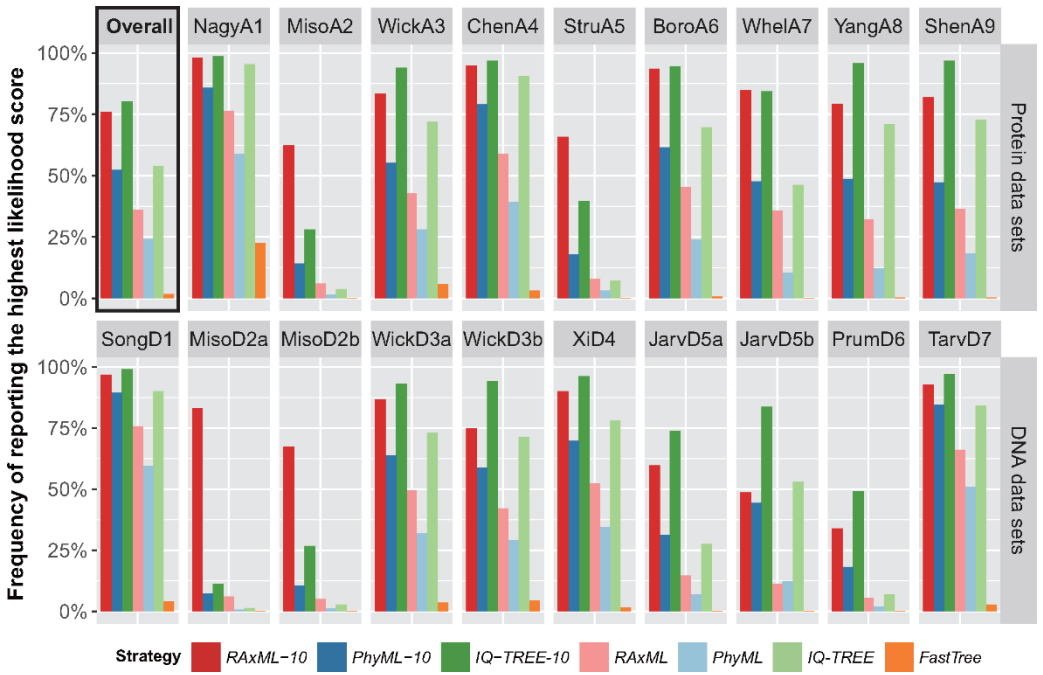
Next editions of Reprohackathon

- Reprohackathon 2 in Lyon, 9-10 July 2018
- Phylogeny data
- Based on a publication on comparision of single-gene trees inference
- In collaboration

with the GDR BIM

- ReproHackathon 3

coming next October!



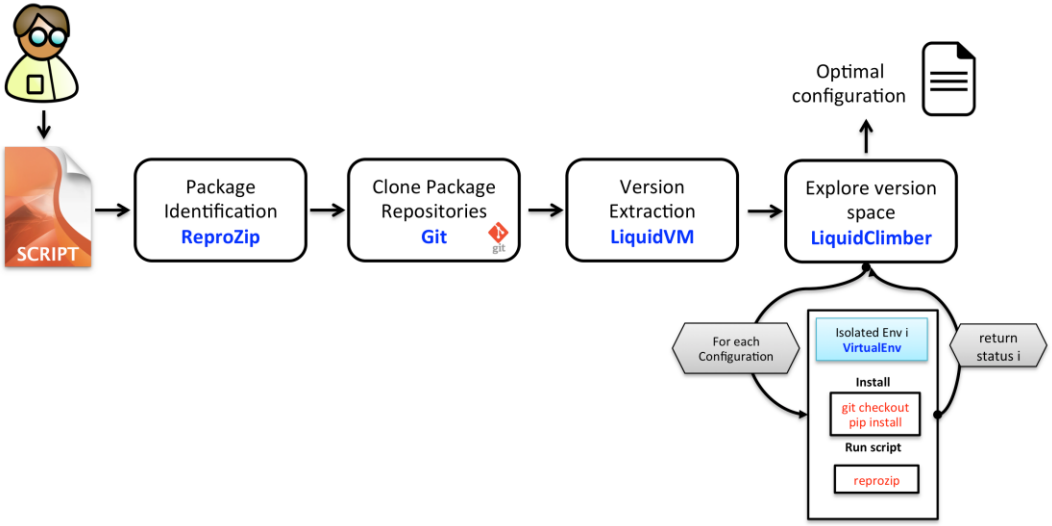
Outline

- Context
- Levels of reproducibility
- Reproducibility-friendly features
- Open Challenges

1. From repeat to replicate

Automatically find the right set *of compatible libraries*

- Docker, VM allows to freeze the environment → Need to liquefy!
- Given a program P that can be repeated in an environment E…
- … Find an environment E' (E' uses more recent versions of libraries than E) where P still *works*

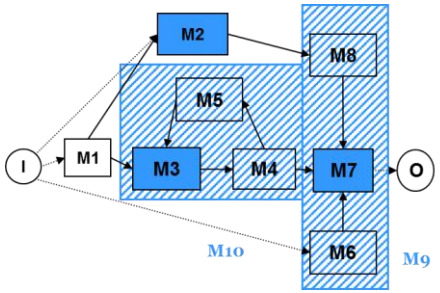


Ongoing work

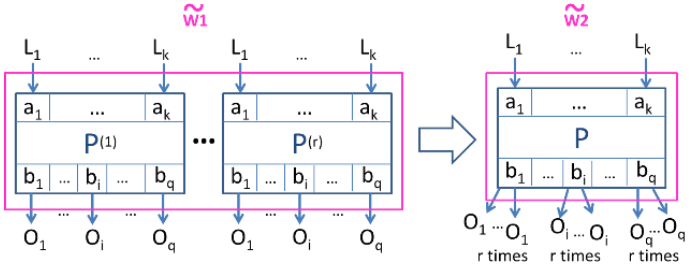
With **Ch. Pradal**, **D. Shasha**, P. Valduriez

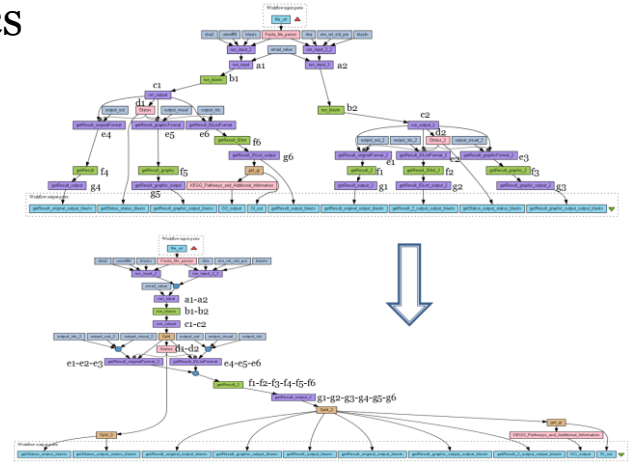
2. From repeat to reuse: Reduce the complexity of workflow structure

- Designing more coarse-grained workflows
  - **Biton *et al.* :** Automatic Design of subworkflows (graph-based)
  - **Alper *et al.*: Abstraction** of provenance traces
  - **Gaignard *et al.*: Summarization** (Web Semantics)



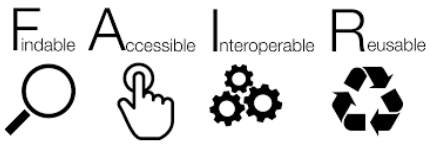
- **Refactoring** workflows
  - Remove redundancies in workflows
    - **DistillFlow (Chen et al.)**: simplifying workflows : Rewritting **Anti-patterns,** Based on Taverna's semantics





Conclusion

- Too many scientific results are not reproducible
- Several Scientific workflow systems and companion tools are mature solutions
  - Repeat is (almost) always reachable
  - Next levels may be more difficult to reach
- Several open challenges are directly related to improvement in research in computer science (graphs, algorithmics…)
- Several Initiatives: Force 11,
  Data and Software Carpentry



FORCE
The Future of Research Communications and e-Scholarship

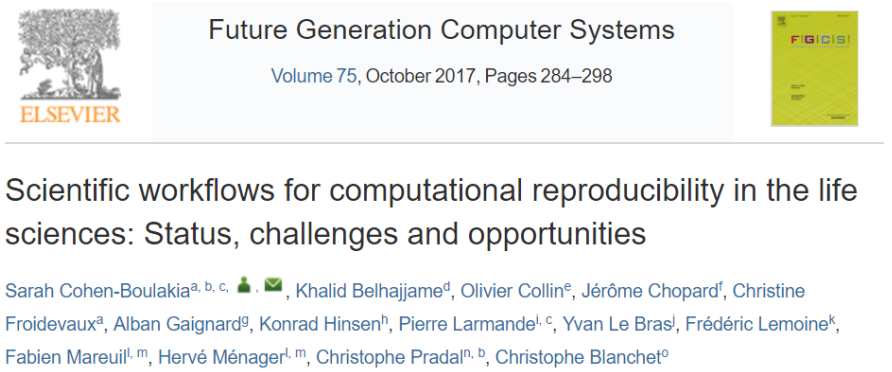
software carpentry

DATA CARPENTRY
MAKING DATA SCIENCE MORE EFFICIENT

Results of our Action

(1) Paper @ FGCS

- Levels of reproducibility
- Criteria of choice
- Open Challenges



(2) 3 hour Webinar : Tutorial + 2 demos

(3) ReproHackathon

- New concept designed
- Second edition on Phylogeny Analysis, in Lyon, 9-10 July 2018














Rejoignez nous !
cohen@lri.fr