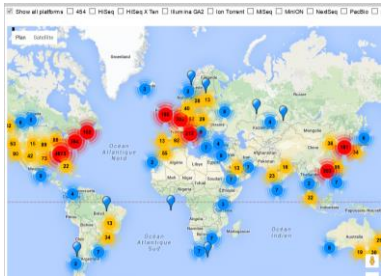
Reproducible Science in Bioinformatics: Current Status, Solutions and Research Opportunities

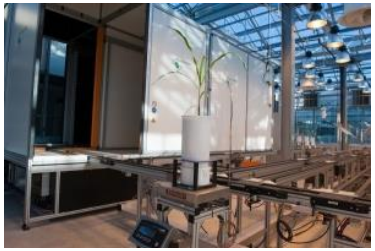
Sarah Cohen-Boulakia

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

Bioinformatics landscape

- Bioinformatics data produced
  - In France (IFB platforms)
  - Europe (Elixir infrastructure)
- Biological knowledge relies on computational experiments
- New technologies producing Big Data sets
  - Sequencing (NGS)
    - 1st Human Genome project: 12 years $10,000/Mbase
    - 2016 : 200 genomes/week $0,03/Mbase
  - Plant Phenotyping
    - Phenotyping Platforms
    - Images, sensor data
      → 11 Tera/year

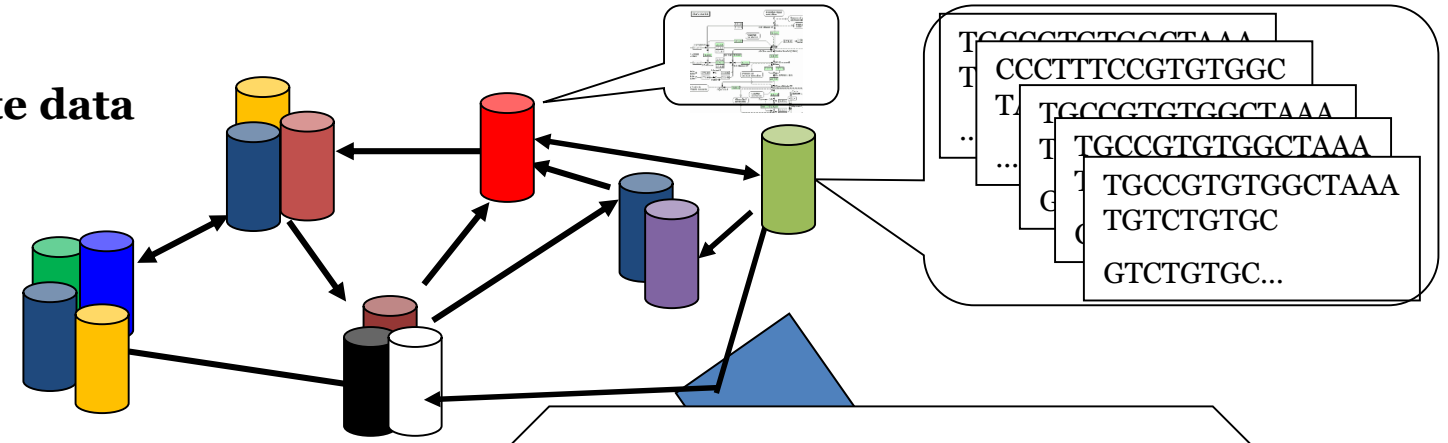




Bioinformatics analysis

Public and private data sources

- Distributed
- Heterogeneous



Binarization Water Use Efficiency

Segmentation RUE ...

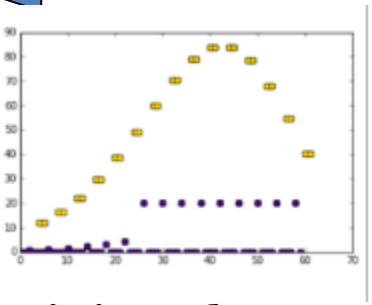
How has this plot been generated?
With which images?
With which binarization algorithms
→ **Provenance**

What is the difference between these experiments?

?

Tools

- Distributed
- Heterogeneous
- To be chained



Biologist's workspace

Take Home Message

Compared to 20 years ago…

- The number and diversity of the data sources has increased a lot
  - > 1,500 public 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 (workflow, tools) 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!

Outline

- Next generation data integration
- Reproducibility: status
- Reproducibility: solutions
  - Demo 1: OpenAlea on plant data sets
  - Demo 2: Galaxy on sequence data sets
- Wrap-up, Conclusion and Challenges

Studies on reproducibility

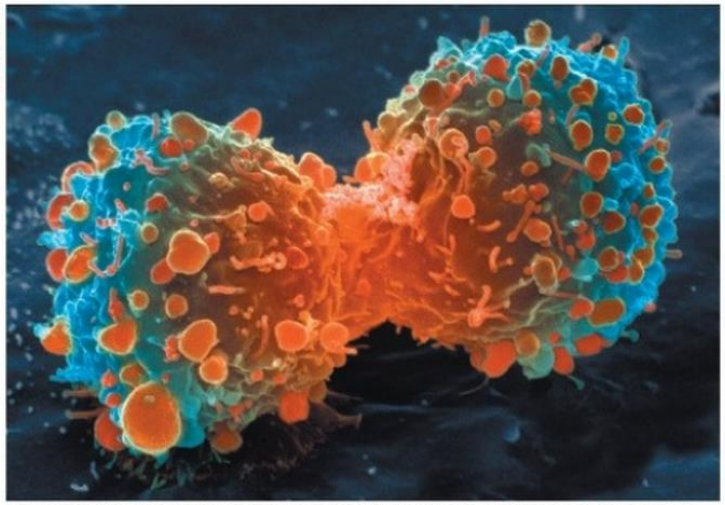
- Nekrutenko & Taylor, Nature Genetics (2012)
  - 50 papers published in 2011 *using the Burrows-Wheeler Aligner for Mapping Illumina reads*.
  - 31/50 (62%) provide no information
    - no version of the tool + no parameters used + no exact genomic reference sequence
  - 7/50 (14%) provide all the necessary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

Drug research -- Life-saving therapies – Cures…



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 to

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

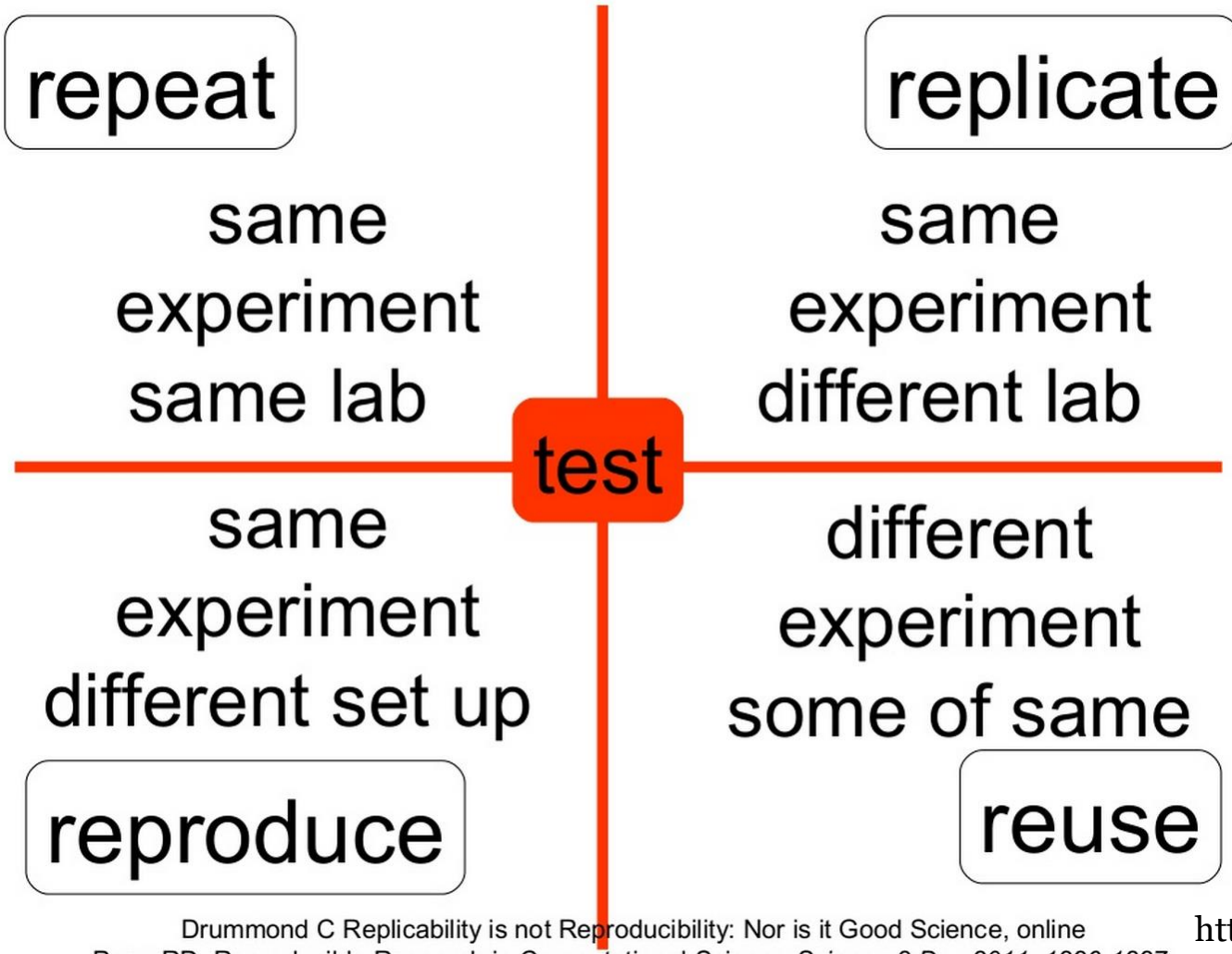
Six red flags for suspect work

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

Know when your numbers are significant

http://www.slideshare.net/carolegoble/ismb2013-keynotecleangoble

Repeat, reproduce, replicate, reuse…



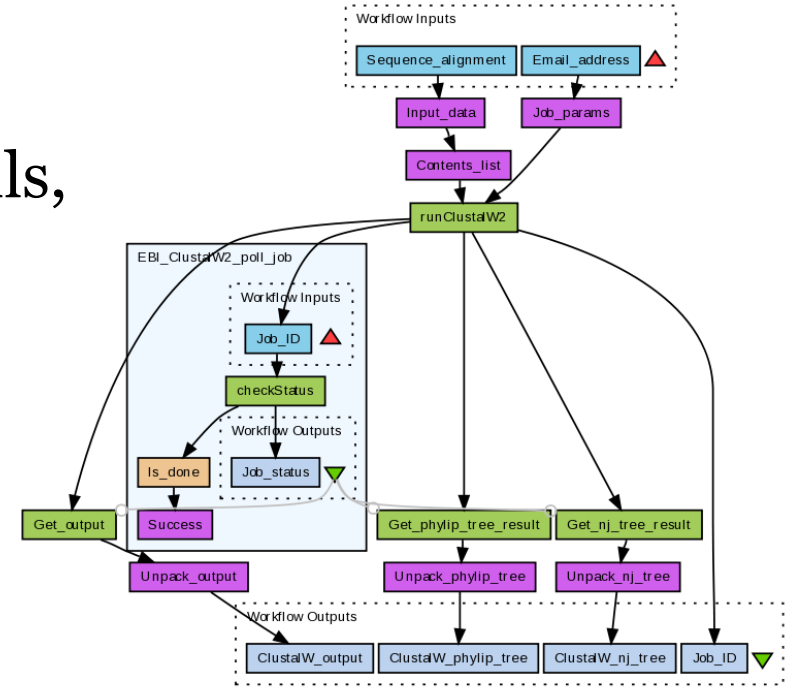
Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

http://www.slideshare.net/carolegoble/ismb2013-keynotecleangoble

Outline

- Next generation data integration
- Reproducibility: status
- Reproducibility: solutions
- Wrap-up, Conclusion and Challenges

Scientific workflow systems and Provenance
[Element of solution 1 & 2]

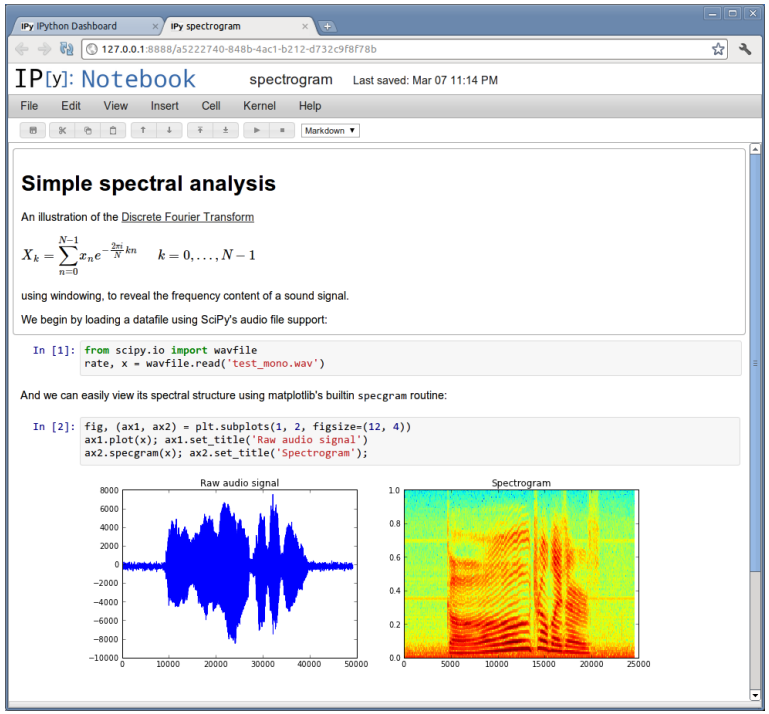


- Numerous systems: Galaxy, VisTrails, Taverna, WINGS, OpenAlea …
- Visual programming
  - Chaining tools
- Specification vs Executions
  - Specification
    - Designed by end-users
    - Describes the tools to be used /programs to be called, in which order
    - The workflow and its components can be annotated (meta-data)
  - Execution (*Provenance module*)
    - The specification run with a given input dataset + parameter setting
    - Tracking, logging data produced and consumed
    - (Pieces of) executions can be annotated (meta-data)

Notebooks [Element of solution 3]

IP[y]: IPython Interactive Computing

jupyter



- Web-based interactive computational environment
- Combination of code execution, text, mathematics, plots and rich media into a single document
- Some systems export workflows as executable IPython/Jupyter papers

Packaging the context, runtime environment [Element of Solution 4]

- 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)
- Such solutions can be used to capture the workflow systems context and runtime environments

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)

FORCE11
The Future of Research Communications and e-Scholarship

FAIR principles
(fragments)

elixir

- 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.

Illustration on two systems and use cases

- Use case and Demo 1 (Jérôme Chopard)
  - Kind of data sets: Plants
  - Workflow system: OpenAlea
  - Companion tools: Conda…
- Use case and Demo 2 (Yvan Le bras)
  - Kind of data sets: sequences
  - Workflow system: Galaxy
  - Companion tools: Docker…

Outline

- Next generation data integration
- Reproducibility: status
- Reproducibility: solutions
- Wrap-up, Conclusion and Challenges

OpenAlea and Galaxy: *reproducible-friendly* workflow systems?

System	Specification	Execution	Environment
OpenAlea	Nested wf: +++ Format : - Expressivity : +++ Source code : ++ Annotation : + *Package manager*	*Provenance* Format: PROV-one, PROV Interactive GUI (Jupyter) Annotation : -	++ Conda
Galaxy	Nested wf : ~ Format: +++ JSON (CWL) Expressivity: ~ Source code : ++ Annotation : EDAM *myexperiment*	*Histories* Format: JSON Browsing GUI Interactive GUI (Jupyter) Annotation : EDAM	+++ Conda, Docker

Research opportunities

1. From repeat to replicate

- Automatically finding 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

- Reproducible papers (Notebooks)
  - Interactive computational environment
  - Combination of code execution, text, mathematics, plots and rich media into a single document

➔To be formalized

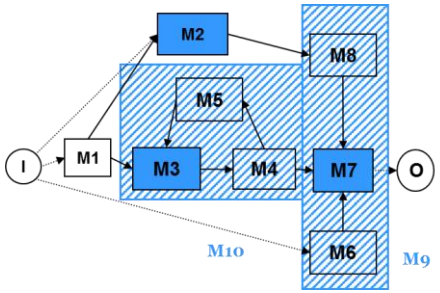
➔Efficiently reusing (searching for) notebooks is an open question

2. Finding the Right Workflow

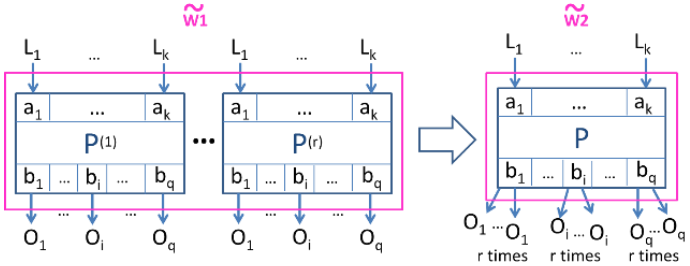
- Repositories queried (IR-style)
- Open question: Query languages for repositories
  - Given a high-level description of a (integration) task – a sketch
  - Given a input and/or and output format/type
  - *Given a workflow – find similar workflows*
  - Search across workflow models (Galaxy, Taverna…) …
- Core of the problem: Workflow similarity
  - State-of-the-art [SCB+14]
  - Need to design hybrid and efficient solutions
- Becomes a practical topic
  - Large repositories are available + Smaller provenance repositories
- Relationships with Business workflows
  - BPQL, BPMN-Q, BP-QL, … considering logs

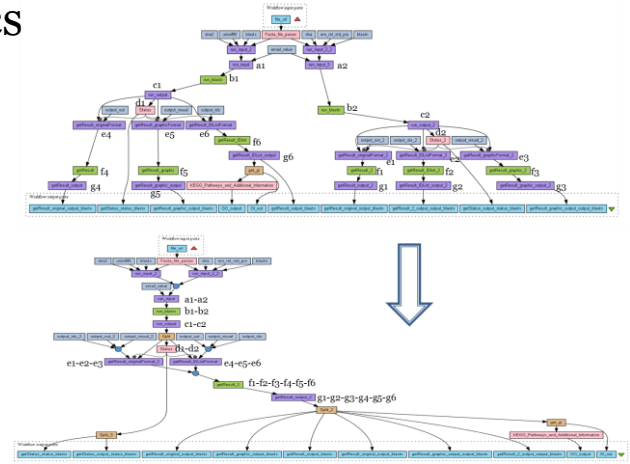
3. Reducing the complexity of workflow stucture

- 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





4. Reconciling workflow and scripting

- A lot of bioinformatics analysis are performed using scripts (instead of workflows)
- Provenance of a script execution?
  - noWorkflow [MBC+14], yesWorkflow [MSK+15]
- Equivalence between scripts and workflows?
  - Provenance-equivalence [CBC+14]? Other kind of equivalence?
- Aim
  - Optimization of workflows (using ZOOM*userviews, DistillFlow…) ↔ Optimization of scripts (refactoring, …)
- New workflow systems based on scripts: NextFlow, SnakeMake…

Conclusion

- Too many scientific results are not reproducible
- Mature solutions exist, not perfect but able to solve a large number of cases, increasingly used in the bioinformatics community
- Several open challenges are directly related to improvement in research in computer science (graphs, algorithmics…)
- Several Initiatives: Force 11,
  Data and Software Carpentry





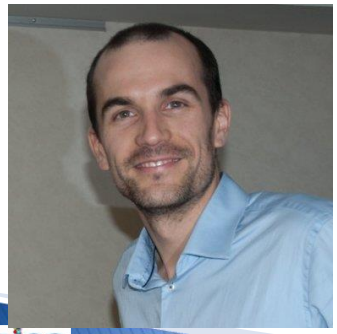



ReproVirtuFlow @ MaDICS

Join us!

https://www.madics.fr/actions/actions-en-cours/reprovirtuflow/





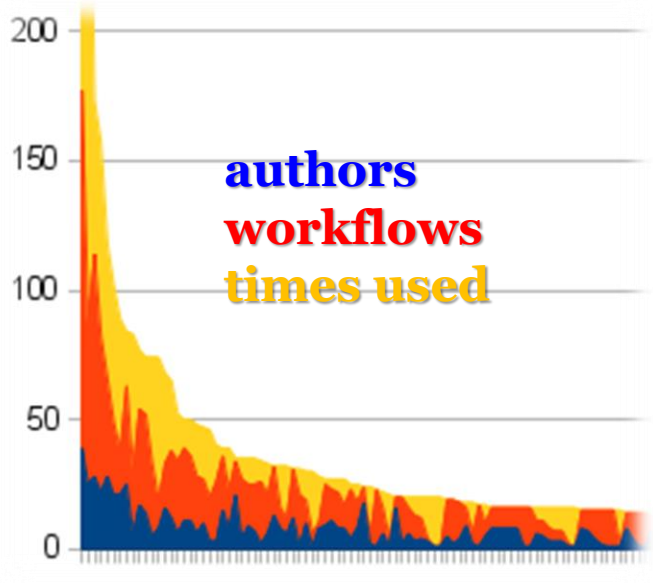


Study on workflow reuse

[SSDBM 2010, SIGMOD Record 2009]

With Ulf Leser & Johannes Starlinger (U. Berlin)

- 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