



# ReproVirtuFlow (Action MaDICS) Reproductibilité des expériences d'analyse de données scientifiques : bilan et perspectives

**Sarah Cohen-Boulakia**

Université Paris-Sud, Université Paris-Saclay, LRI CNRS INS2I UMR 8623

**Christophe Blanchet**

Institut Français de Bioinformatique (IFB-Core) CNRS INSB UMS 3601



# Contexte, enjeux

Reproductibilité *computationnelle*

Nombre croissant de résultats scientifiques non reproductibles

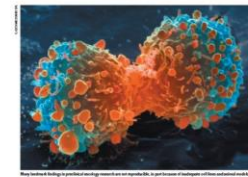
Nombreux domaines concernés

Enjeux économique majeur

- Non reproductibilité des études pré-cliniques évalué à >\$10 milliards annuel pour les USA

Devient une obligation contractuelle

- Projets NSF, certains editeurs



**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 need leading agencies to push for government monitoring 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.

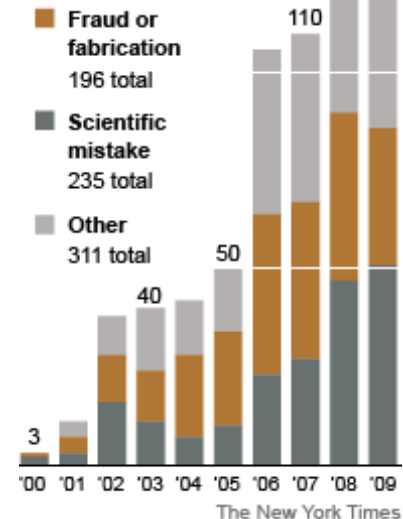
Know when your numbers are significant

Raise standards for preclinical cancer research  
C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives most change if patients are to benefit.

47/53 "landmark" publications could not be replicated  
[Begley, Ellis Nature, 483, 2012]

## 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.



# Objectifs de l'action

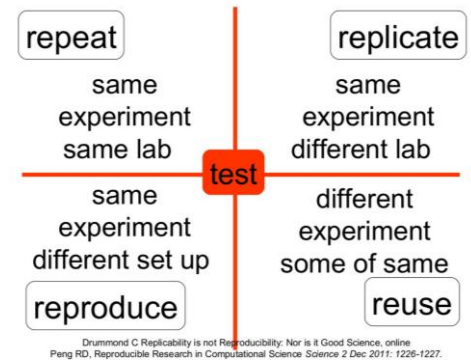
**Sensibiliser** la communauté

**Etat des lieux** et niveaux de reproductibilité

- Solutions ? Niveaux à considérer ?  
Couverture ?

Focus sur trois types d'approches

- Définition de l'analyse
  - **Workflows scientifiques** : outils - ordre
- Données et paramètres d'entrée
  - **Provenance** : Trace des executions
- Environnement d'exécution
  - **Virtualisation, Packaging** : Trace de l'environnement (OS, librairies...)

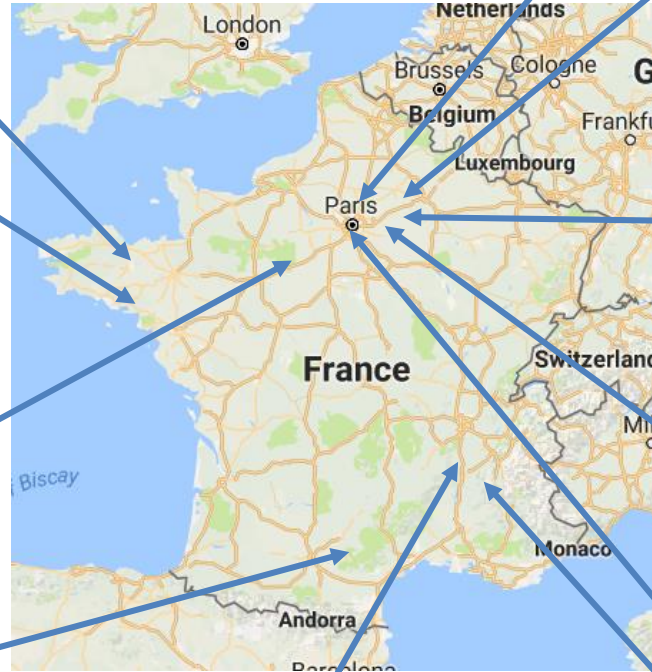


## Communautés :

Base de données, représentation des connaissances, algorithmique (graphes), grilles, systèmes, compilation, langages...

# Membres

## UMR et UMS du CNRS



IRISA Univ.  
Rennes

Univ. CHU  
Nantes

Centre de  
Biophysique  
Moléculaire,  
CNRS Orléans

IRD, CIRAD,  
INRA, Inria,  
Univ.  
Montpellier

Univ. Lyon 1 LIRIS

GDR Bioinfo, Groupes de travail IFB  
Centres *Data Sciences internationaux*

LRI Univ.  
Paris Sud

CDS, Center for  
Data Science  
Saclay

Institut Francais  
Bioinformatique  
Gif s/Yvette  
Institut Pasteur,  
Paris

Lamsade Univ.  
Paris Dauphine

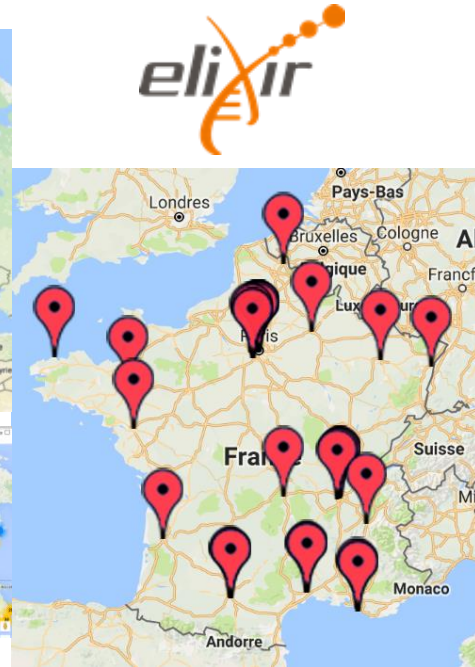
LIG  
(Grenoble)

## Des Données aux connaissances

- **Données**  
Distribuées, **Hétérogènes**
- **Outils**  
Different types, nb paramètres
- **Pipelines d'analyses (*workflows*)**  
Complexes

## Cas d'utilisation cibles

- NGS (cancer), Phenotypage Plantes  
**Données massives**
- ELIXIR Europ. Research Infra.  
21 Pays, 180 partenaires  
➔ Analyses avec **workflows scientifiques**

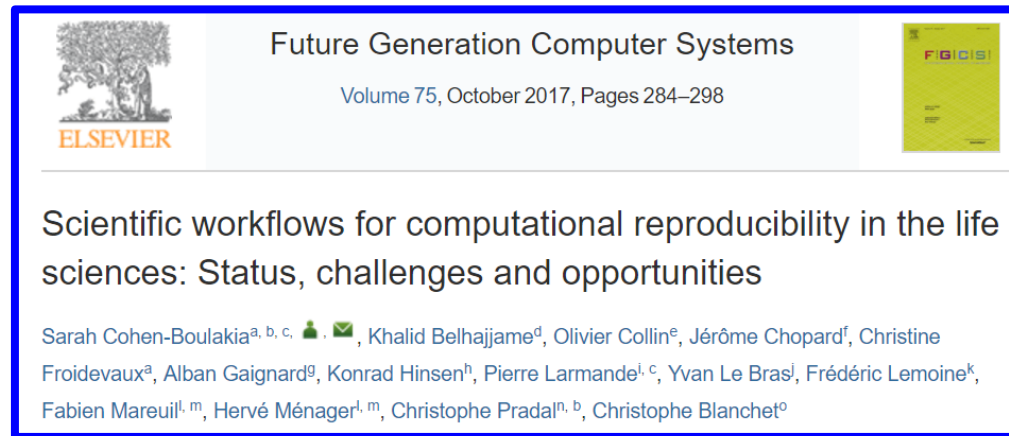


# Résultats

(1) Etat des lieux sur les solutions existantes à base de workflows + verrous

(2) Webinar en ligne tutoriel et 2 démonstrations

(3) Conception et organisation du premier ReproHackathon



ReproHackathon 1 : Reproduction d'un pipeline d'analyse de données RNA-Seq publié (mélanome Uvéal)

1-2 Juin 2017  
Gif s/Yvette  
25 participants

IGRoussy,  
Curie, Pasteur,  
Saclay, Paris,  
Nantes, Lyon...



# Perspectives

## 2015-2017: Reproductibilité Computationnelle ...

- Traitements de données à l'aide de **systemes de workflows**
- Etapes d'analyses **déterministes**
- Données de **biologie moléculaire**

## Généralisation, nouveaux problèmes ? Solutions ?

- Workflows scientifiques → **scripts**
- Analyses **non déterministes**, apprentissage (statistique)
- Nouveaux **types de données** à analyser (images, graphes...) et nouveaux **domaines** (robotique, reproductibilité et éthique)

## Journées de rencontre et montage d'au moins **deux nouveaux ReproHackathon**

- Images, machine learning, traitement du signal...
- Collaboration MaDICS : EADM, ATLAS, GRAMINEE...
- **Numéro spécial** revue IEEE CiSE (Computing in Science Engineering)

## Journée **éthique et reproductibilité**

- CERNA - Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene
- Groupe travail « Recherche et innovation responsables : Éclairages interdisciplinaires » (Maison des sciences de l'Homme Paris-Saclay)





# ReproVirtuFlow @



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

