



ReproHackathon : un projet né dans le cadre de l'action ReproVirtuFlow (GDR Madics)

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



Le GDR MaDICS



- ▶ GDR du CNRS

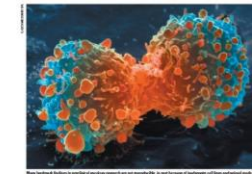
Masses de Données, Informations et Connaissances en Sciences

- ▶ Fonctionnement de MaDICS sous la forme d'actions
 - **Animation scientifique** pluri-disciplinaire autour d'un thème
 - 9 actions en cours
- ▶ **Action ReproVirtuFlow**
Reproductibilité des expériences d'analyse de données scientifiques



Contexte, enjeux

- ▶ Reproductibilité *computationnelle*
- ▶ Nombre croissant de résultats scientifiques non reproductibles
 - Y compris dans les revues à fort facteur d'impact
 - Pas (toujours) volontairement
- ▶ Nombreux domaines concernés
 - Certains plus critiques que d'autres...
- ▶ **Enjeux économique** majeur
 - Non reproductibilité des études pré-cliniques évalué à >\$10 milliards annuel pour les USA
- ▶ Devient une obligation contractuelle
 - Projets NSF, certains éditeurs



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.

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

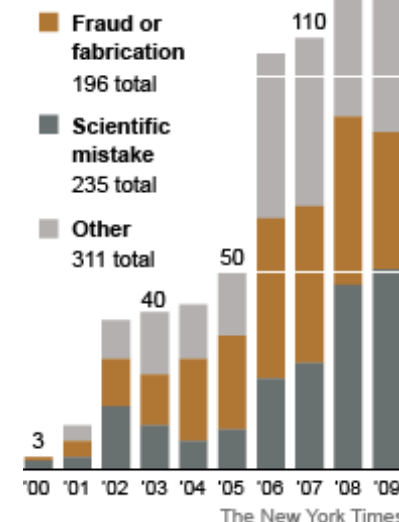
Six red flags for suspect work

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

Know when your numbers are significant

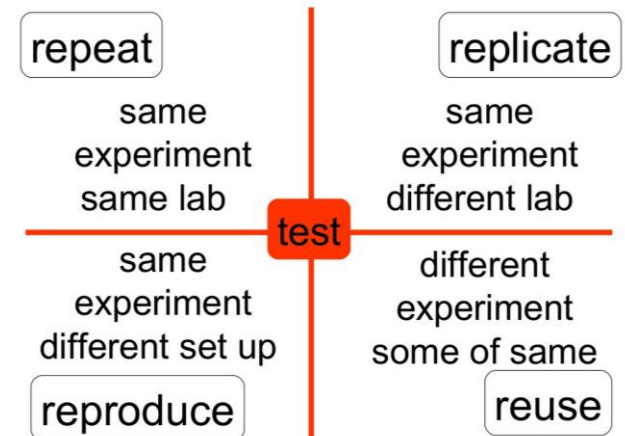
Retractions On the Rise

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

- **Etat des lieux** vis-à-vis des différents niveaux de reproductibilité
 - Quelles solutions existent ?
 - Quels niveaux considérer ?
 - Quels niveaux sont *couverts* ?



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.

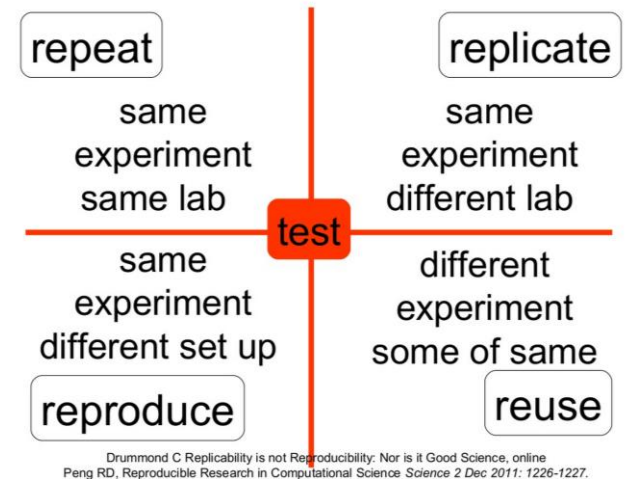
Objectifs de l'action

► **Etat des lieux** vis-à-vis des différents niveaux de reproductibilité

- Quelles solutions existent ?
- Quels niveaux considérer ?
- Quels niveaux sont *couverts* ?

► Focus sur trois types d'approches

- Capturer la définition de l'expérience
 - **Workflows scientifiques**
 - Trace des outils avec ordre d'enchaînement
- Capturer les données et paramètres d'entrée
 - **Provenance**
 - Trace des executions
- Capturer l'environnement d'exécution
 - **Virtualisation, Packaging**
 - Trace de l'environnement (OS, librairies...)

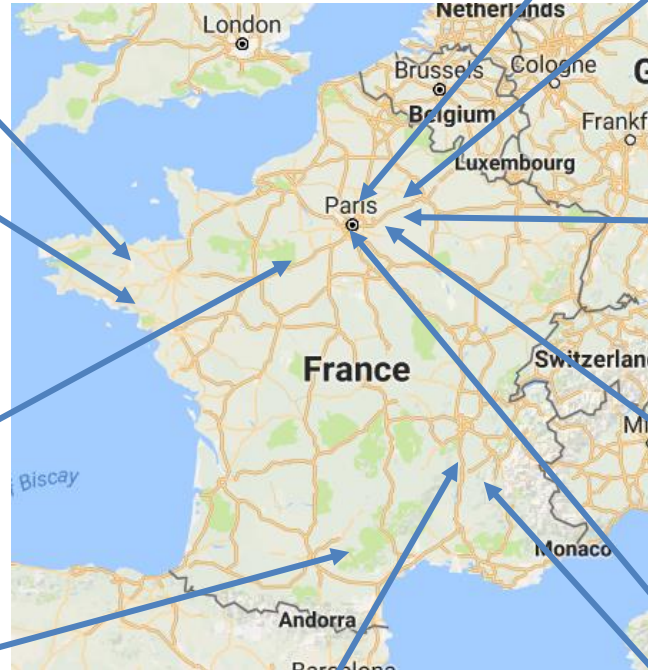


Communautés :

Base de données,
représentation des
connaissances,
algorithmique
(graphes),
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, Gpes 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)

Données d'intérêt

- ▶ Données bioinformatiques
 - Française (plateformes IFB)
 - Européennes
 - Complexes et très hétérogènes
 - ▶ Données très volumineuses
 - Séquencage (NGS)
 - 1st Human Genome project: 12 ans \$10,000/Mbase
 - 2016 : 200 genomes humains/semaine \$0,03/Mbase
 - Phenotypage de plantes
 - Plateformes de phenotypage
 - Images, données sensor (arrosage, lumière...)
- 11 Tera/an



Activités

► Objectifs

- **Evaluer** les capacités des systèmes de workflows pour la reproductibilité
- **Considérer un large panel** de systèmes
- **Considérer** des cas réels (données et analyses)

► Activités

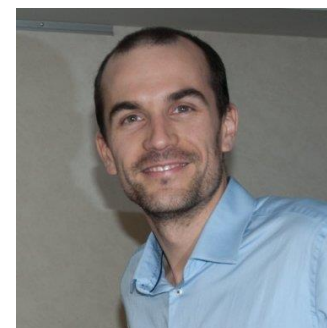
- **Etat de l'art** [Cohen-Boulakia et al., 2017, FGCS]
- **Webinar** : 6 décembre 2016, Grenoble (A. Legrand)
- Démarrage d'une **série de ReproHackathons** !
- **Reprohackathon 1** : aujourd'hui ☺
 - Sur le Cloud de l'IFB (Christophe Blanchet)
 - Sur un cas d'utilisation précis (Frédéric Lemoine)



Merci à nos sponsors !



université
PARIS-SACLAY
Groupe CompBio



**Rejoignez nous :
Inscrivez-vous sur madics !**