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

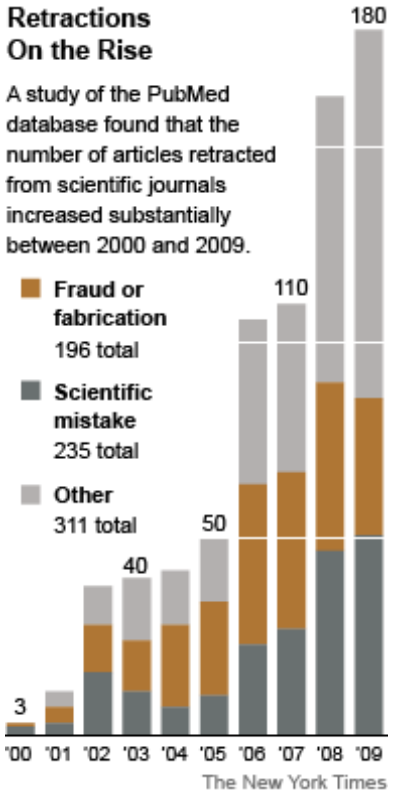
- Fonctionnement de MaDICS sous la forme d'actions
  - Animation scientifique pluri-discplinaire autour d'un thème
  - 9 actions en cours
- Action ReproVirtuFlow

  Reproductibilité des expériences d'analyse de données scientifiques

Contexte, enjeux

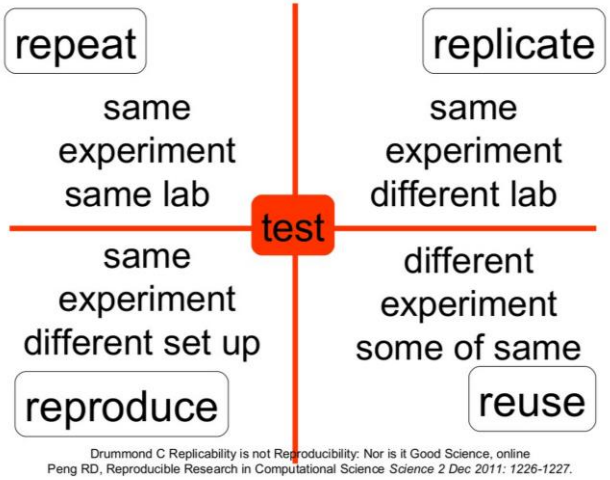
- Reproductibilité *computationnelle*
- Nombre croissant de résultats scientifiques non reproductibles
  - Y compris dans les revues à fort facteur d'impact
  - Pas (toujours) volontairement
- Nombreux domaines concernés
  - Certains plus critiques que d'autres…
- Enjeux économique majeur
  - Non reproductibilité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

[Begley, Ellis Nature, 483, 2012]



Objectifs de l'action

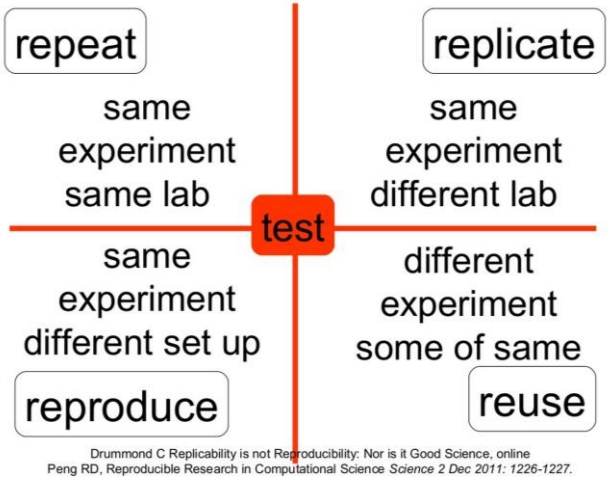
- **Etat des lieux** vis-à-vis des différents niveaux de reproductibilité
  - Quelles solutions existent ?
  - Quels niveaux considérer ?
  - Quels niveaux sont *couverts* ?



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* ?



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.*

- Focus sur trois types d'approches
  - Capturer la definition de l'experience
    - Workflows scientifiques
      - Trace des outils avec ordre d'enchaînement
  - Capturer les données et paramètres d'entrée
    - Provenance
      - Trace des executions
  - Capturer l'environement d'exécution
    - Virtualisation, Packaging
      - Trace de l'environment (OS, librairies…)

Communautés :
Base de données, représentation des connaissances, algorithmique (graphes), systèmes, compilation, langages…

Membres

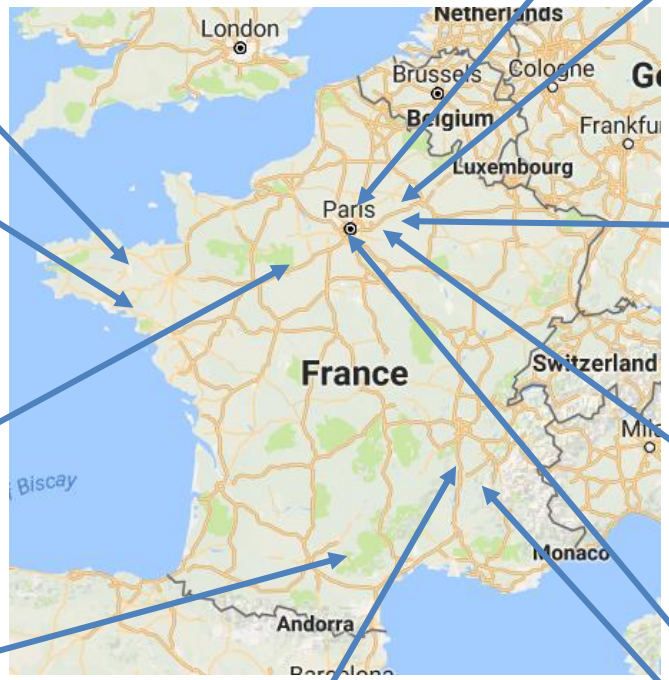
UMR et UMS du CNRS

IRISA Univ. Rennes

Univ. CHU Nantes

Centre de Biophysique Moléculaire, CNRS Orléans

IRD, CIRAD, INRA, Inria, Univ. Montpellier



Univ. Lyon 1 LIRIS

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)

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

Données d'intêret

- 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 !

MaDICS

université PARIS-SACLAY

Groupe CompBio

Rejoignez nous : Inscrivez-vous sur madics !