Data Institute
Univ. Grenoble Alpes

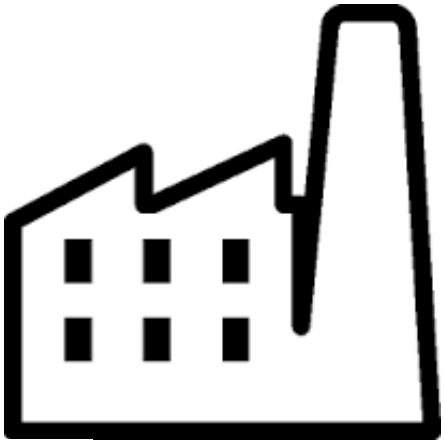
Grenoble Alpes Data Institute

An interdisciplinary research Institute on how data change science and society

https://data-institute.univ-grenoble-alpes.fr/

@grenobledata

Promouvoir la culture des données



Organisation

Dir: M Blum	**WP0** **Grenoble Alpes Data Institute**	Dir. adj.: T Christakis C Noille

WP1	**WP2**	**WP3**	**WP4**	**WP5**
Sciences des données pour les sciences de la Terre, de l'espace, de l'environnement	**Sciences des données pour les sciences de la vie**	**Données riches et massives pour les humanités**	**Sciences des données, sciences sociales et media sociaux**	**Gouvernance des données, protection des données et vie privée**
PO Amblard J Le Sommer	F Forbes J Labarère	L Albaret T Lebarbé	G Bastin E Gaussier	C Castelluccia K Bannelier

Relations avec le monde socio-économique	**Formation**
K Salamatian	A Leclercq Samson

Moyens

~300 K€/WP soit 1700 K€ en tout

Data challenges

Diffuser la culture des données par l’expérimentation

Open seminar & workshops

Présenter des travaux en cours qui soulèvent des enjeux liés aux données (visualisation).

Apparier des présentations de chercheurs de laboratoires différents.

Présentation autour de logiciels (R, Python…).

Commence fin mars avec 1 séminaire sur la reproductibilité.

Un outil de visibilité

Journée annuelle des partenaires avec mise en relation des étudiants et d'entreprises.

Nous solliciter pour des appels à projets en lien avec les sujets « données » (IDEX, région, Europe,…).

Data Institute
Univ. Grenoble Alpes

Grenoble Alpes Data Institute

An interdisciplinary research Institute on how data change science and society

https://data-institute.univ-grenoble-alpes.fr/

@grenobledata

Data Institute
Univ. Grenoble Alpes

WP1: Science des données pour les Sciences de la Terre, de l'Espace, et de l'Environnement

Julien Le Sommer (IGE)
Pierre-Olivier Amblard (gipsa-lab)

Science des données pour les Sciences de la Terre, de l'Espace, et de l'Environnement

Science des données ?

- algorithmes et outils statistiques descriptifs permettant
  - la réduction de dimension de jeux de données,
  - la détection de structures récurrentes (patterns)
  - l'inférence causale

recouvre pour une large part les méthodes d'apprentissage statistique (classification/régression)

- un champ de recherche en plein explosion sous l'effet conjoint :
  - de l'évolution des capacités de calcul, de collecte et de stockage des données
  - du fort besoin dans de nombreux secteurs d'activité privés
  - de nouvelles pratiques et approches de développement logiciel

Science des données pour les Sciences de la Terre, de l'Espace, et de l'Environnement (STEE)

Science de la Terre de l'Espace et de l'Environnement ?

- géosciences (climat/ocean/atmosphere/glace/hydro), planétologie, astrophysique, écologie, physique des particules
- des champs disciplinaires historiquement très présents/visibles à l'UGA (cf: OSUG)
- des champs disciplinaires dans lesquels collecte et traitement de données sont intensifs, systématisés et partagés depuis plusieurs décennies (grands instruments)
- des champs disciplinaires souvent très liés à l'observation spatiale

Science des données, Sciences de la Terre, de l'Espace, et de l'Environnement dans la COMUE UGA

Pôle MSTIC

Participants au WP1

Pôle PAGE
(dont OSUG)

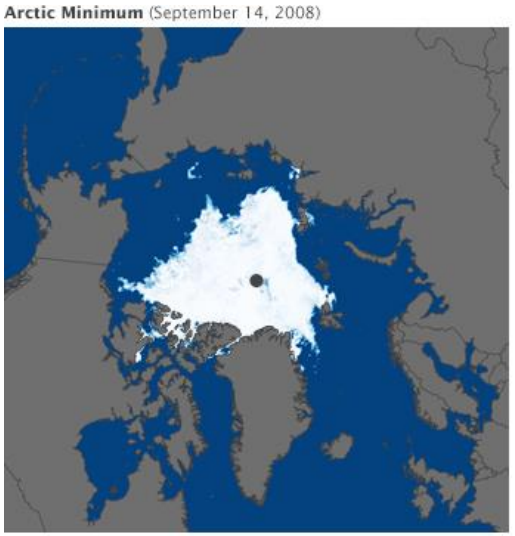
Enjeux de la science des données pour les STEE
potentiel de transformation scientifique

Pourquoi est-il opportun et nécessaire de renforcer les liens entre ces communautés ?

- enjeux communs (data deluge, reproducibilité, brain drain, gap technologique)
- particularités des champs STEE : historique big data, rapport à la causalité
- un fort potentiel de leadership sur le site grenoblois
- des démonstrations d'apport substantiels de nouvelles pratiques d'organisation

Exemple de nouvelle pratique: challenges collaboratifs

- rassemblent sur un groupe de chercheur / étudiants pour « craquer » un problème ou améliorer les performance d'un algorithme
- suppose de disposer de :
  - 1 problème, 1 base de données, 1 solution, 1 critère
- ce mécanisme : amélioration incrémentale mais substantielle
- et permet d'identification des partenaires pour collaborations

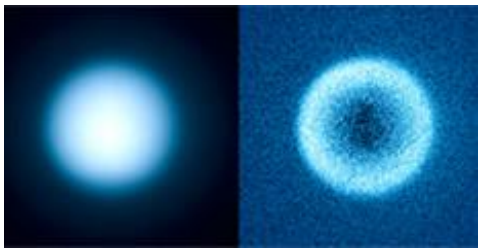


voir par exemple : http://www.ramp.studio/events/sea_ice_colorado

Science des données pour les STEE objectifs du projet

Objectif :

Accélérer le développement et l'appropriation des méthodes/outils de la science des données dans les champs STEE dans les laboratoires de la COMUE UGA

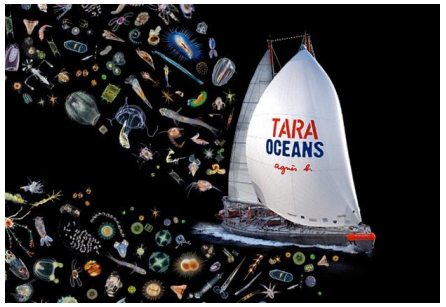


SPHERE

Characterizing exoplanets

SWOT

Inferring vertical exchange in the ocean



Tara-Oceans

Mapping complex ecological networks

Science des données pour les STEE : levier et accélérateurs

- ***clef #1 : le rapport à la production logicielle***
  - la production logicielle est l'un des vrais levier de la révolution en cours.
  - favorise l'assemblage de pipelines d'analyse très composites mais robustes
  - les chaines de traitement ne sont pas ou peu appuyées sur ces technologies

- ***clef #2 : le questionnement sur les ressorts de motivation individuelle***
  - avec la science des données émergent de nouvelles pratiques collaboratives (bootcamps, hackatons, data challenges, code sprints…)
  - qui sont en fait des réponses pragmatiques pour motiver/attirer les acteurs
  - traduisent aussi la difficulté à mettre en oeuvre une démarche interdisciplinaire

Instruments mis en oeuvre dans le cadre du projet (1/2)

Chaire postdoctorale : « Innovative data exploration strategies in Earth, Space and Environmental Sciences »

but : démontrer le potentiel de l'approche « science des données » dans les champs disciplinaires des sciences de la terre de l'environnement et de l'espace

- **modalités** :
  - un appel à candidature portant sur l'ouverture d'une chaire de 3 ans
  - un(une) chercheur(se) dans un laboratoire sur un sujet clairement identifié
  - procédure et calendrier conjoints avec le WP2
- **calendrier** : annonce fin mars 2017, début d'activité en octobre 2017
- **critères** : profil du candidat, projet de recherche, impact sur le data institute
- **profils attendus** :
  - jeune chercheur « disciplinaire » motivé par la mise en oeuvre d'approche de la science des données dans son champ de recherche (reproducibilite + logicielle)
  - « data scientist » motivé par l'implication dans un champ disciplinaire des STE

Instruments mis en oeuvre dans le cadre du projet (2/2)

Data challenges

but : tester de nouveaux modes de rencontre et de collaboration scientifique

- **modalités :**
  - soutien financier et technique à l'organisation de challenges collaboratifs
  - sur des questions bien identifiées dans le cadre thématique STEE
  - approche et outil conjoints avec le WP2

Projets interdisciplinaires

but : aider au montage de nouvelles collaborations structurées entre équipe

- **modalités :**
  - financement de projets collaboratifs incitatifs (~5-10k€)
  - à l'interface entre sciences des données et disciplines STEE
  - procédure et calendrier conjoints avec le WP2
- **calendrier :** un appel à projet au printemps 2017, 2018 et 2019 (30k€/an)

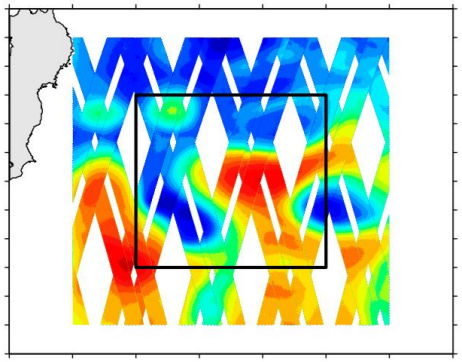
Exemple de challenge collaboratif dans le cadre du projet

<u>Améliorer l'estimation par altimétrie satellitaire des échanges verticaux dans l'océan</u>

<u>question</u> : estimation des échanges verticaux par la mission altimétrique SWOT (2021,PIA)

vitesse verticale : échanges de chaleur et de carbone entre ocean / atmosphère

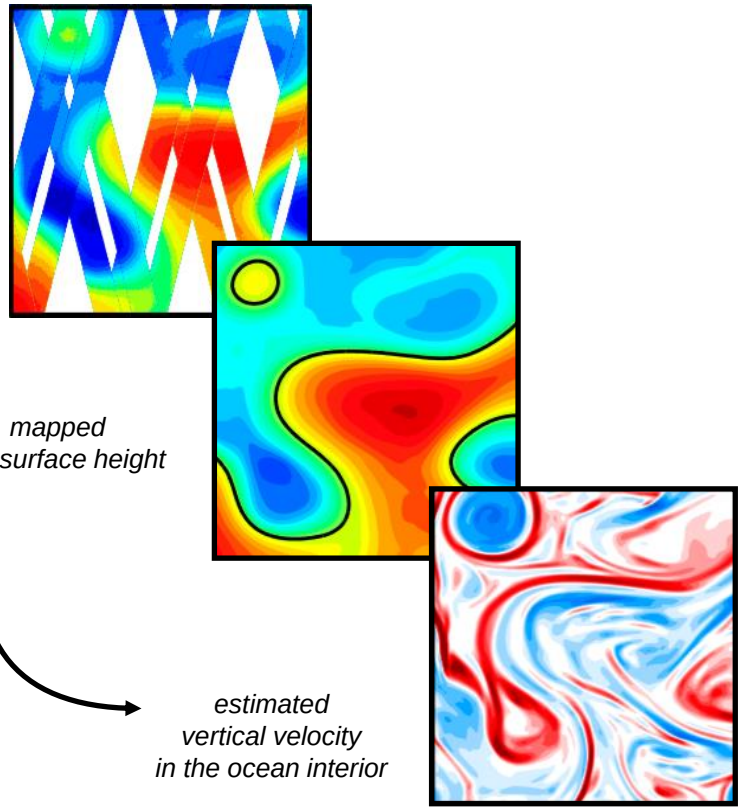
SWOT satellite : observation of sea surface height at high resolution



observed sea surface height

mapped sea surface height

estimated vertical velocity in the ocean interior



WP2: **Science des données pour les sciences de la vie**

Florence.Forbes@inria.fr, JLabarere@chu-grenoble.fr, Michael.Blum@imag.fr

Collaboration/Participants: BGE, CHUG, GIN, GIPSA, IAB, INRIA, LIG, LJK, TIMC

Objectif: Analyse de données biomédicales à grande échelle

- Exploiter la **quantité de données** disponibles pour le bien des patients

- **Technologie de l'information** et pratiques médicales: **Monitoring** de mesures physiologiques individuelles et **facteurs environnementaux**

- **Prédiction** (prédisposition génétique), détection d'anomalies, alertes, **diagnostic, prise de décision, assistance**

Données biomédicales

Human genome/biological data
600GB per full genome
15PB+ in databases of leading institutes

Human proteome
160M data points (2.4GB) per sample
>3TB raw proteome data in ProteomicsDB

Hospital information systems
Often more than 50GB

PubMed database
>23M articles

Medical sensor data
Scan of a single organ in 1s
creates 10GB of raw data

Cancer patient records
>160k records at NCT

Prescription data
1.5B records from 10,000 doctors and
10M Patients (100 GB)

Clinical trials
Currently more than 30k
recruiting on ClinicalTrials.gov

Difficultés & Challenges transverses: éviter sous-utilisation, gaspillage d'information

- **Fusionner** des données multivariées, **hétérogènes et distribuées**, asynchrones
  eg. séries temporelles, images, EHR statique, etc.
- **Multimodales**, multicapteurs et de **qualité inégale (controle qualité)**
- Données et méta-données **structurées et non structurées**
- **Données massives** (omics)
- Données **manquantes**
- Intégrer **connaissances d'experts** (cliniciens)
- Faire émerger de **nouveaux biomarqueurs**
- **Controle des fausses découverte**s, tests multiples, etc.

Vers de nouvelles pratiques d'analyse de données

Approche

- **Analyse d'image et du signal avancée:**

  Statistique bayesienne, algorithme de fusion probabiliste, modèles à variables latentes, modèles graphiques, etc.

- **Apprentissage machine et optimisation**

- Calcul parallèle, **HPC et Big data.**

Validation et Impact

- **Collaboration Plateformes grenobloises,** ex. GRICAD

- **Contacts autres CDP**, ex. LIFE

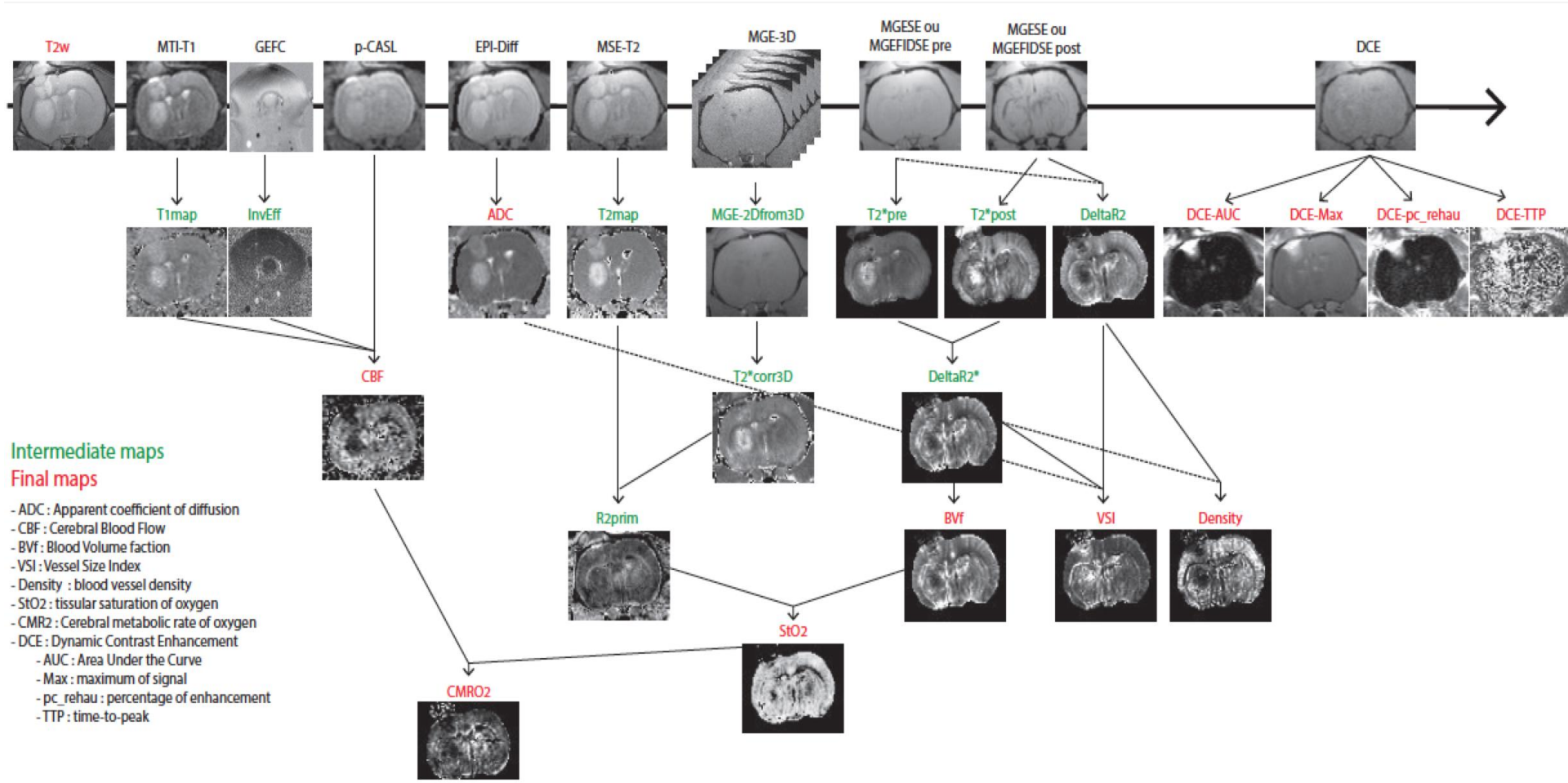
- **Use cases:** ex. UK biobank data, prédisposition génétique et facteurs environnementaux

- Organisation de **challenges, animation, formation, publications**

IRM multiparamètrique (Radiomics): GIN, INRIA, LJK

Protocoles IRM multiparamétrique étendus (>6) en routine

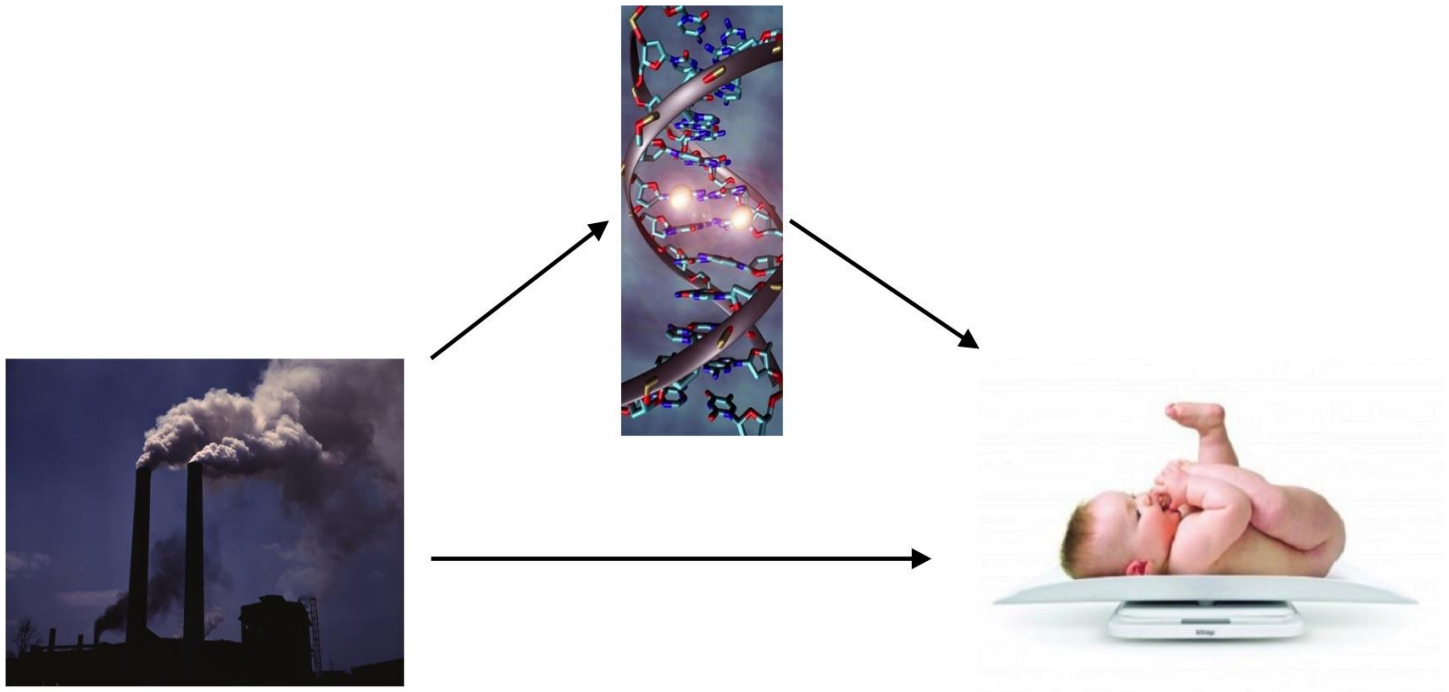
Aide au diagnostic: Caractériser dans l'espace (localisation) et dans le temps (suivi) des lésions **de manière non invasive. Comment extraire de l'information?**



Organisation de Data Challenges

L'effet de la pollution de l'air sur la santé des nourissons est-il modulé par l'épigénome?

http://mediation-data-challenge.imag.fr/



Evaluation de modèles **statistiques et algorithmes** pour une analyse de la médiation **multidimensionelle** en **épidémiologie**

Dépenses de WP2

Junior chair (2017-2020)

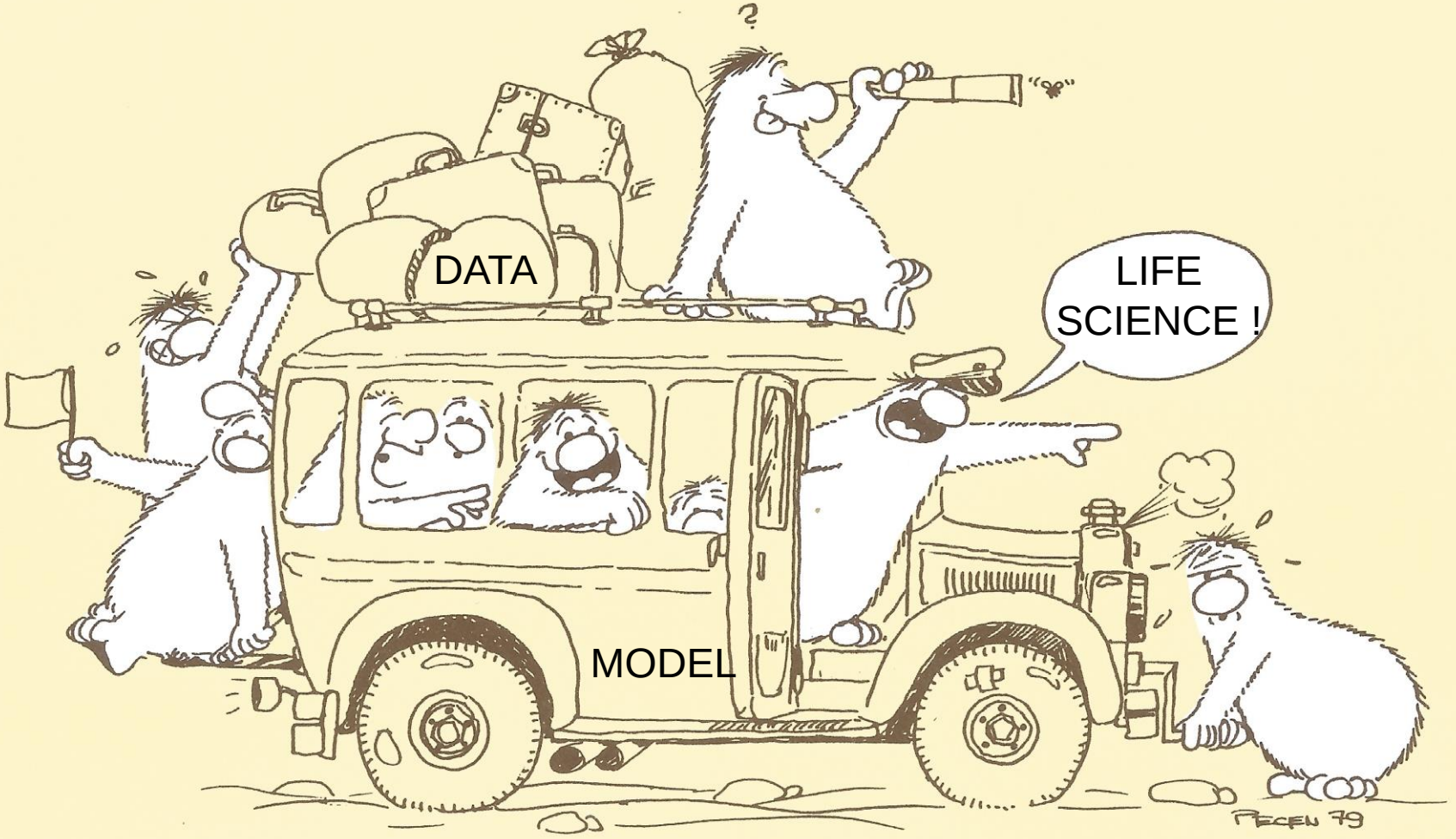
- Spécialiste en data science qui vise à faire des recherches interdisciplinaires en biologie.

- Biologiste impliqué dans le développement d'outils numériques pour la biologie.

Projets interdisciplinaires

- Financement de projets interdisciplinaires (5-10 k€) à raison de 30K€/an distribué tous les ans pendant 3 ans.

Contact: **Florence.Forbes@inria.fr, JLabarere@chu-grenoble.fr, Michael.Blum@imag.fr**



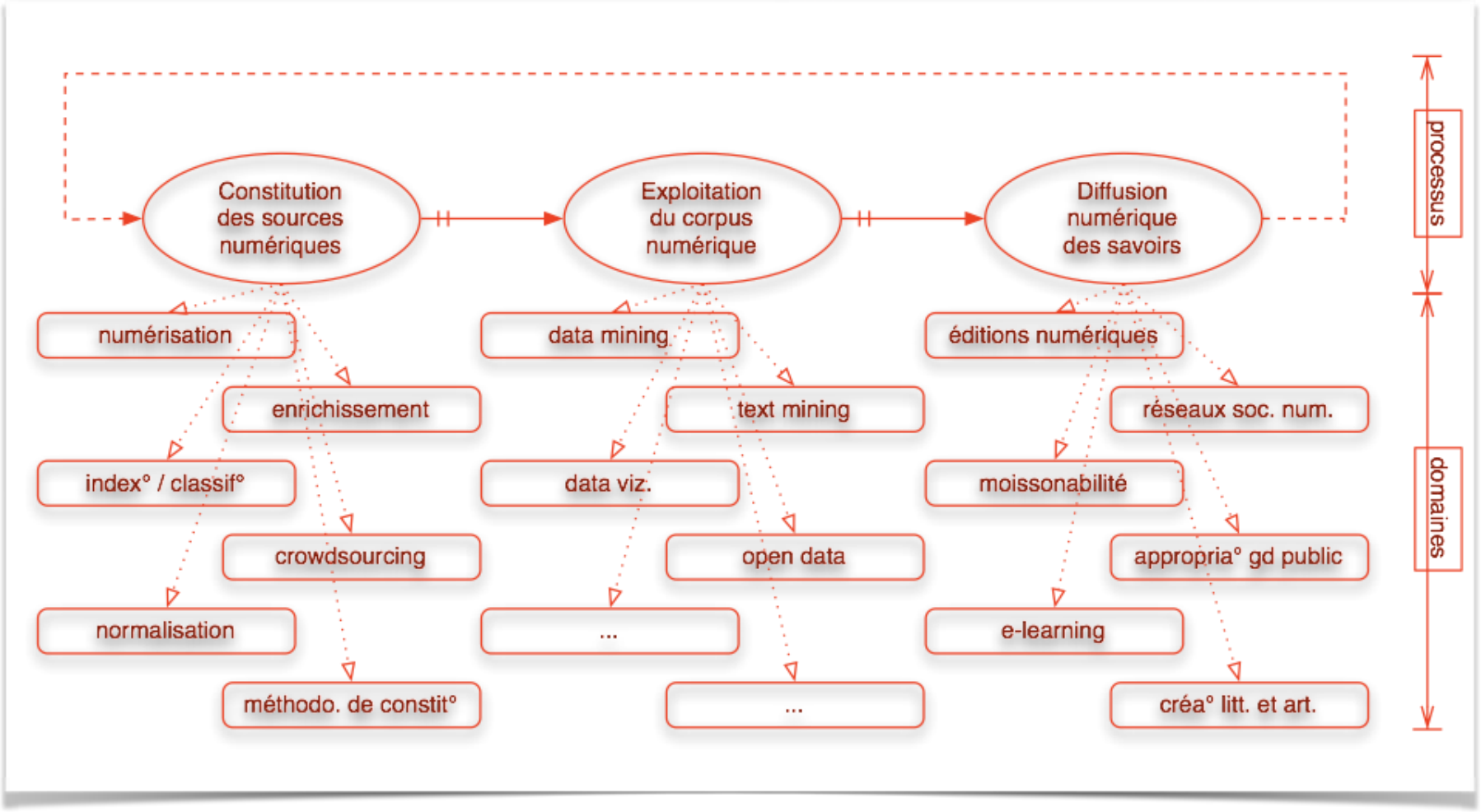
Thomas Lebarbé, Litt&Arts UMR 5316

Démarre SHS !

Data Institute
Univ. Grenoble Alpes

WP3 DonnÉes MAssives & Riches pour la REcherche en SHS

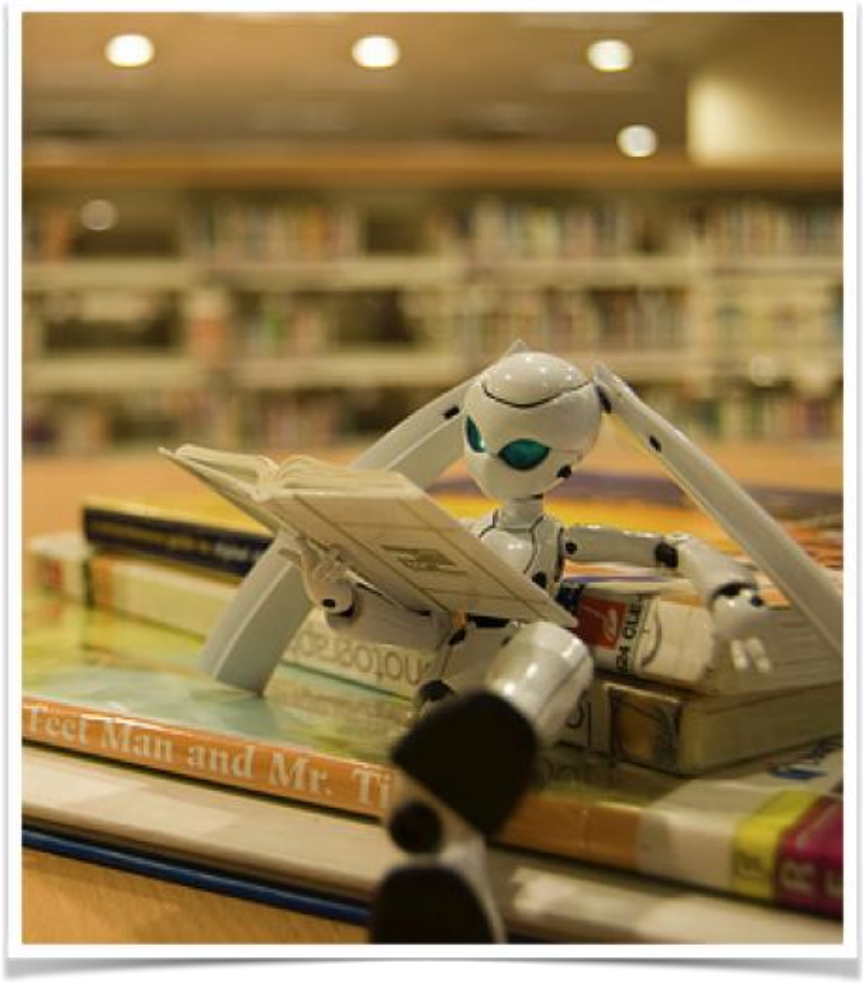
Démarre SHS ! Périmètre



Démarre SHS ! Livrables

- D12: un dépôt complet des données de la recherche patrimoniale et SHS à UGA [mid-term]
- D13: un protocole et un outil de production contributive de données à forte plus-value intellectuelle [mid-term]
- D14: un système d'interrogation des données SHS [end]
- D15: une méthode d'analyse des données hétérogènes riches et massives [end]
- D16: un réseau international des données de la recherche en SHS [end]

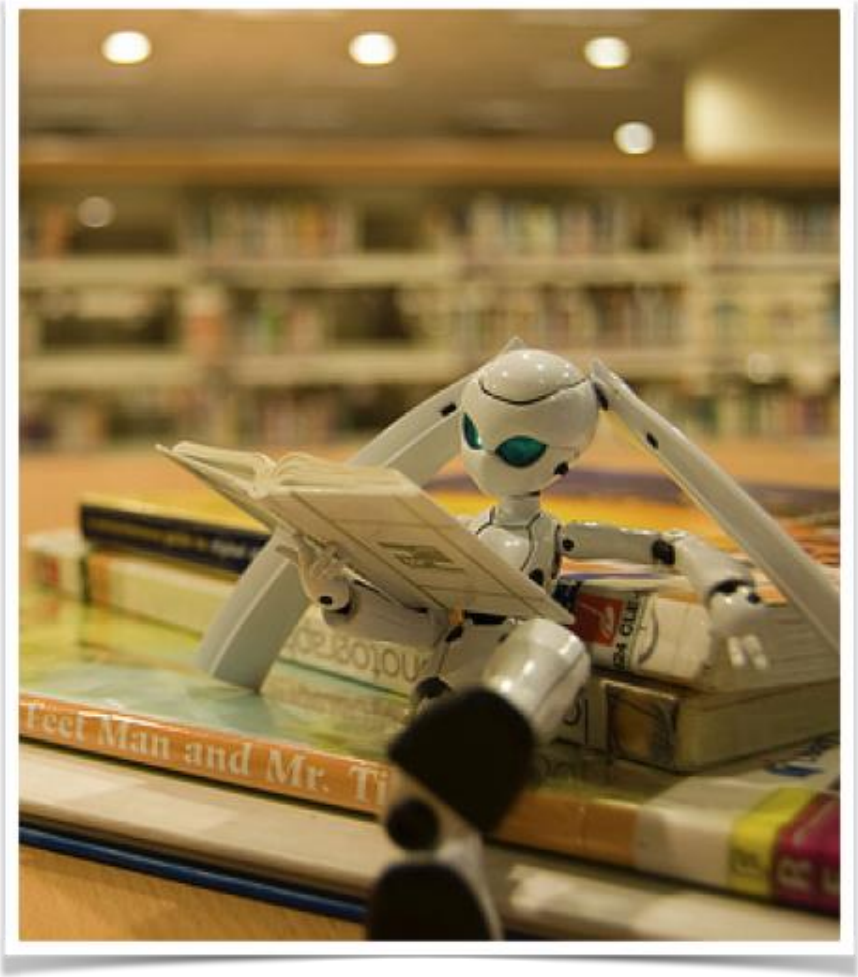
Démarre SHS ! Labellisation



- Labellisation DATA@UGA / Démarre SHS !
- Objectif : kickstart / proof of concept / pierre de touche / prototypage
- Soutien technologique et financier possible
- Contrepartie exigée : libération des données et outils
- Modalités : 2 pages
- Périodicité : annuel, mars

Démarre SHS ! Labellisation

- Critères de sélection formels :
  - Données/outils libres
  - Objet des SHS riches et inter-opérables
  - Labo SHS porteur
  - Ouvert aux doctorants, ingénieurs de recherche, enseignants-chercheurs et chercheurs.



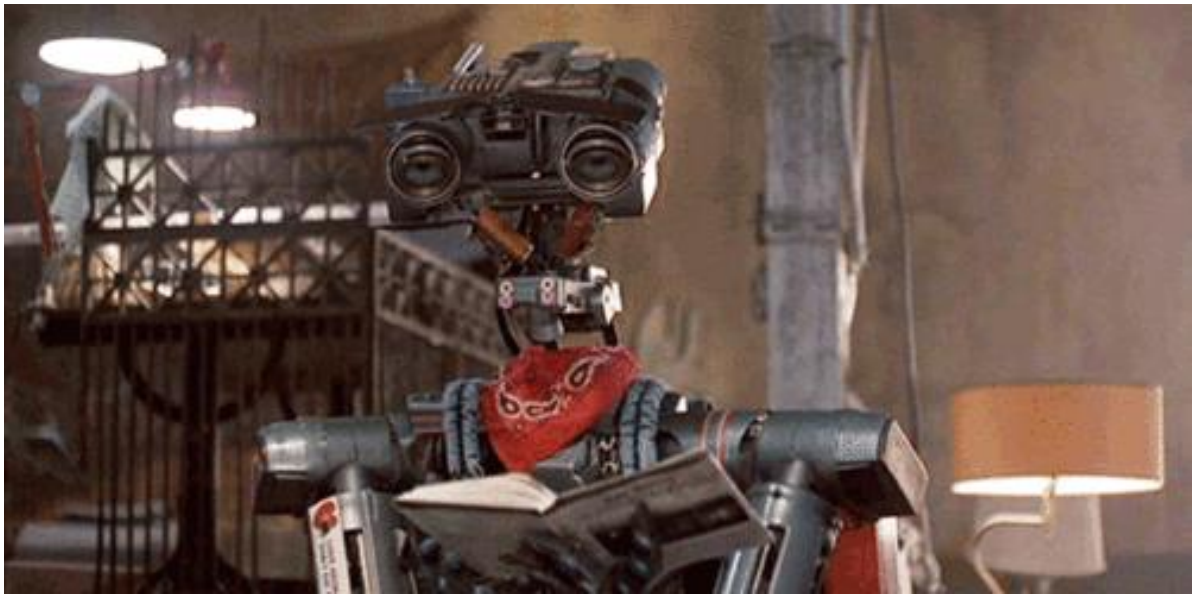
Comité scientifique local

- Coordination :
  - Thomas Lebarbé
  - Lucie Albaret
- Membres :
  - ✓ Anne Dalmasso
  - ✓ Claire Mouraby
  - ✓ Aurélie Nardy
  - ✓ Christine Noille
  - ✓ Elena Pierazzo

Démarre SHS ! formations

- Financements de proposition de formation ou de besoin de formation.
- Ponctuelles :
  - TXM
  - XML / XSLt
  - TEI
  - PHuN
  - …
- Récurrentes :
  - Les « cafés des néophilologues »

Démarre SHS ! Communication



- Carnet d'hypothèses : demarreshs.hypotheses.org
  - suivi du WP, suivi des projets
  - annonces
  - événements

Thomas Lebarbé, Litt&Arts UMR 5316

Démarre SHS !

Data Institute
Univ. Grenoble Alpes

DonnÉes
MAssives &
Riches pour la
REcherche en
SHS

Data Institute
Univ. Grenoble Alpes

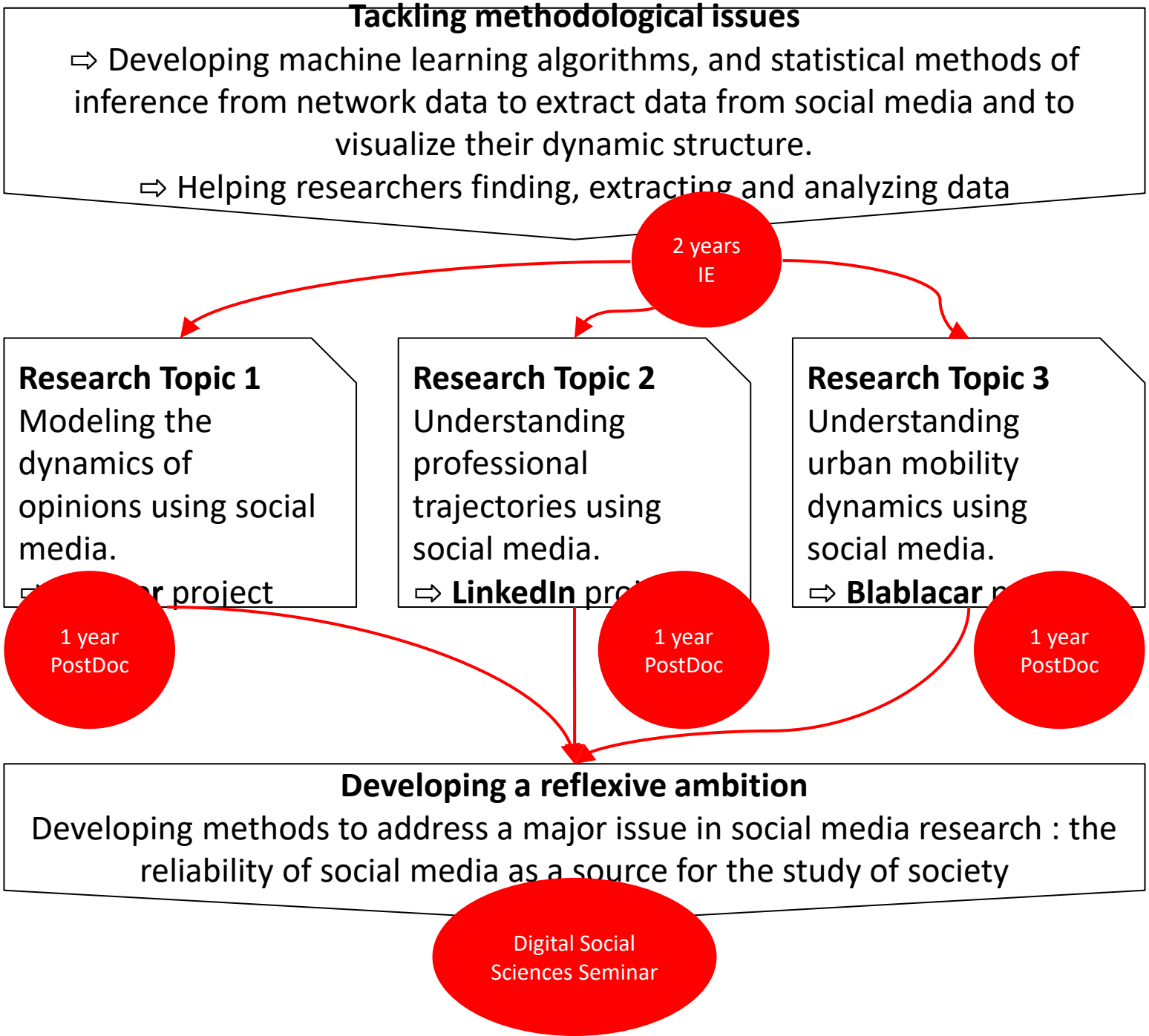
WP4 Data science, Social Media and Social Sciences

- Gilles Bastin (Pacte)
- Eric Gaussier (LIG)

Objectif

Développer des recherches dans lesquelles les média sociaux (où les informations, opinions, expériences, carrières professionnelles sont partagées par les gens) jouent le rôle d'observatoire de la société

⇨ Principaux laboratoires : LIG, LJK, LISTIC and PACTE
⇨ Financement principal : 2 ans IE, 3 ans Post-Doc
⇨ Première réunion le 27 mars


Tackling methodological issues
⇨ Developing machine learning algorithms, and statistical methods of inference from network data to extract data from social media and to visualize their dynamic structure.
⇨ Helping researchers finding, extracting and analyzing data
2 years IE
Research Topic 1
Modeling the dynamics of opinions using social media.
Research Topic 2
Understanding professional trajectories using social media.
Research Topic 3
Understanding urban mobility dynamics using social media.
1 year PostDoc
1 year PostDoc
1 year PostDoc
Developing a reflexive ambition
Developing methods to address a major issue in social media research : the reliability of social media as a source for the study of society
Digital Social Sciences Seminar

Première réunion

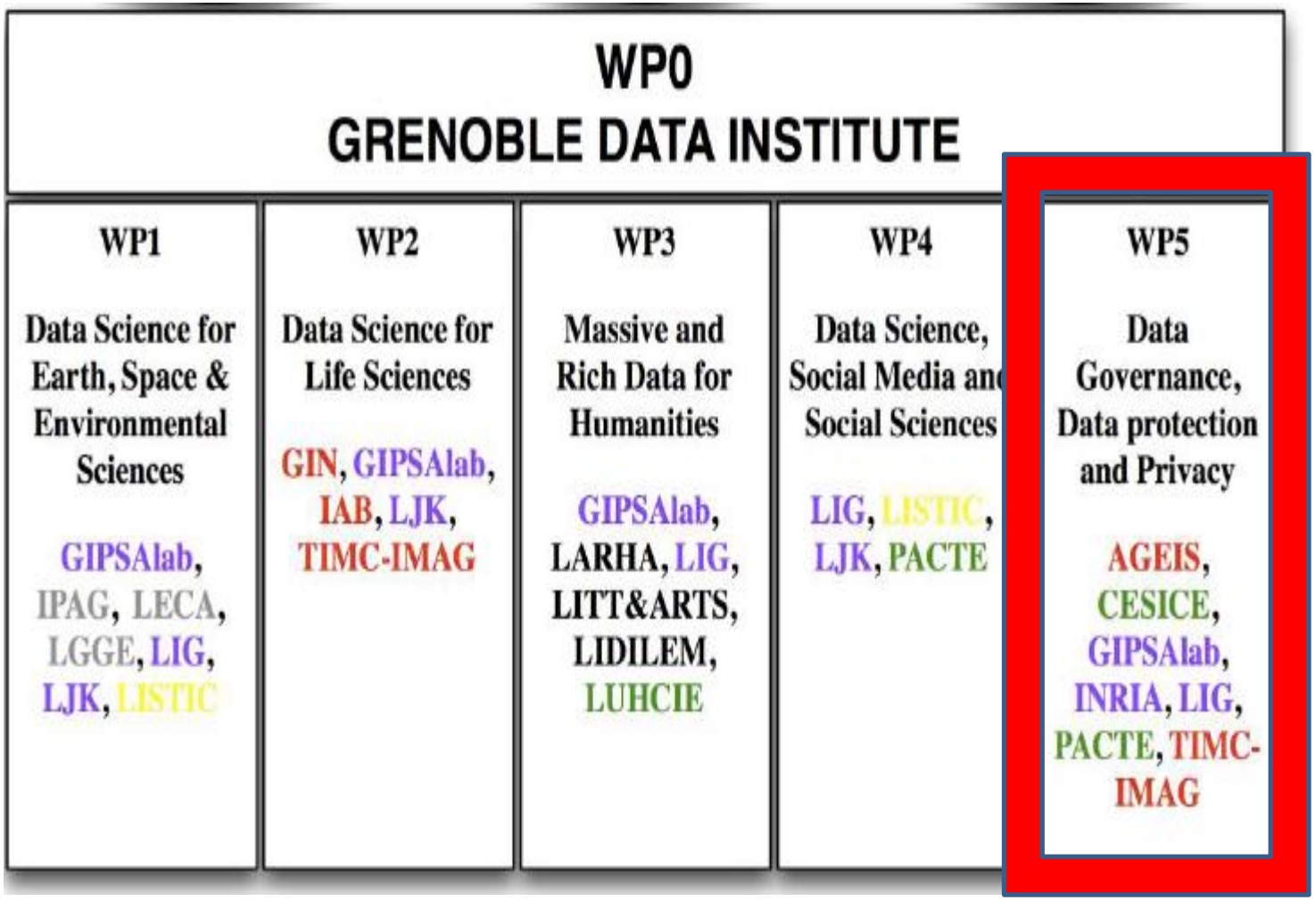
Quand : lundi 27 mars à 14h00
Où : salle séminaire rdc bâtiment IMAG
Qui : toute personne travaillant sur les média sociaux avec une perspective sociale, informatique ou mathématique

Data Institute
Univ. Grenoble Alpes

Work package 5: *Data Governance, Data Security & Protection of Privacy*

Karine Bannelier (pole PSS, CESISE, UGA) & Claude Castelluccia (pole MSTIC, Privatics, INRIA)

March 2017


WP0
GRENOBLE DATA INSTITUTE
WP1
Data Science for Earth, Space & Environmental Sciences
GIPSAlab, IPAG, LECA, LGGE, LIG, LJK, LISTIC
WP2
Data Science for Life Sciences
GIN, GIPSAlab, IAB, LJK, TIMC-IMAG
WP3
Massive and Rich Data for Humanities
GIPSAlab, LARHA, LIG, LITT&ARTS, LIDILEM, LUHCIE
WP4
Data Science, Social Media and Social Sciences
LIG, LISTIC, LJK, PACTE
WP5
Data Governance, Data protection and Privacy
AGEIS, CESICE, GIPSAlab, INRIA, LIG, PACTE, TIMC-IMAG

Data is Everywhere! Data is power!

- Data governance becomes a major issue:
  - For companies, countries, cities, individuals…

- Strong impact on
  - Economy (ex. benefit of big data in the health sector)
  - National/International Security
  - Privacy
  - Ethics

- WP5: Finding solutions to legal, ethical and privacy issues.
  - Using a multi-disciplinary approach
  - lawyers, political scientists, sociologists, economists, IT experts, data specialists and researchers in the health and medicine fields

3 main objectives

1. “Data is power” – **Understanding/Analysing the dynamics of data governance**

2. “Nothing to hide?” - **Proposing legal and technical solutions to data protection and privacy issues**

3. **Making practical recommendations for the health and medical sectors**

A use-case: Medical Data

- A case study will focus on data used in the medical and health domains
- **Legal aspects**: To what extent could hospitals and health actors provide private companies or other stakeholders with massive medical records for research purposes;
- **Technical aspects**: how can medical data be shared/processed? How to perform a risk analysis? What kind of anonymization tools/solutions are necessary?
- Collaboration with the medical researchers, hospitals, CNIL...

Deliverables

- **D21**: Organization of recurrent “open to all” seminars; expert meetings & international workshops; publication of research results in peer-reviewed international journals
- **D22**: Creation of a **comparative dataset** on current data protection and data privacy: laws, regulations, case law on a national, comparative, European and International Law perspective as well as mapping of the actors involved. Longitudinal analysis of data governance structures.
- **D23**: Design **tools** that help analyze current privacy policies and regulations, and tools to improve data transparency (tracking, algorithm transparency).
- **D24**: Define and propose Privacy Impact/Risk assessments **methodology** and software.
- **D25**: Make practical legal and technical **recommendations** on how medical data should be collected and processed.

Collaborations (established or to develop)

- Data Institutes:
  - Oxford Internet Institute (UK)
  - Alexander von Humboldt Institute for Internet and Society (DE)
  - Center for the Internet and Human Rights at the European University Viadrina (DE)…
  - ➢ Grenoble Alpes Data Institute membership to the *Global Network of Internet and Society Research Centers*

- CNIL, ANSSI
- ENISA, OECD
- Industry

Some results already…Amnecys:

Alpine Multidisciplinary NEtwork on CYber-security Studies (https://amnecys.inria.fr/fr/)

- 9 labs (*CESICE, Gipsa-lab, INRIA/Privatics, Institut Fourier, LIG, LISTIC, LJK, TIMA, Vérimag*) and institutions (Université Grenoble-Alpes, le CNRS, l'IEP de Grenoble, l'INRIA, Grenoble INP et l'Université Savoie Mont-Blanc)...working on various aspects (legal, technical, political, economics,...) of cyber-security
- Collaboration on Data surveillance
  - Technical and legal analysis
  - Student co-advising
- ANSSI-Unesco conference on "Construire la Paix et la Securité Internationales dans le monde numérique"
- Teaching...
  - CS prof. Teach in M2 "Securité & Defense"
  - Lawers/Cesise teach in M2 "Cybersecurity"