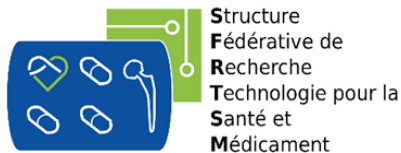


# Génération de données structurées de santé fictives : la plateforme e-Santé du CPER Tec'Santé

Emmanuel Chazard




# Plan de la présentation

1. Diffusion de données et risque de réidentification
2. Cadre réglementaire français
3. Constat : le cadre réglementaire fait obstacle à la recherche
4. La plateforme e-Santé du CPER Tec'Santé
5. Le projet « Health Dataset twins »

# **(1) Diffusion de données et risque de réidentification**

LIVE



 TODAY EXCLUSIVE

# CHARLIE SHEEN SPEAKS OUT

ACTOR REVEALS HE IS HIV-POSITIVE



According to journalist Matt Lauer, M.S. has paid 10 millions dollars to silence those who wanted to spread the information (November 17<sup>th</sup> 2015)

# Quel est le risque ?

- Motivations malhonnêtes :
  - Chantage pour obtenir de l'argent, une action ou une abstention
  - Médisance, ragots
  - Transactions immobilières, viager
  - Caducité des contrats d'assurance
  - Etc.
- Comment ? Bien différencier :
  - Les bases de données anonymes
  - Les bases de données nominatives
  - Les bases de données indirectement nominatives (très fréquentes !)

# Rappelez-vous « Qui est-ce ? »

Je ne sais  
encore rien...





# Rappelez-vous « Qui est-ce ? »

Je sais que :  
- c'est un monsieur



# Rappelez-vous « Qui est-ce ? »

Je sais que :

- c'est un monsieur
- Il porte un chapeau





# Rappelez-vous « Qui est-ce ? »

Je sais que :

- c'est un monsieur
- Il porte un chapeau
- On ne voit pas son nez



# Rappelez-vous « Qui est-ce ? »

Je sais que :

- c'est un monsieur
- Il porte un chapeau
- On ne voit pas son nez
- Il est orange





# Rappelez-vous « Qui est-ce ? »

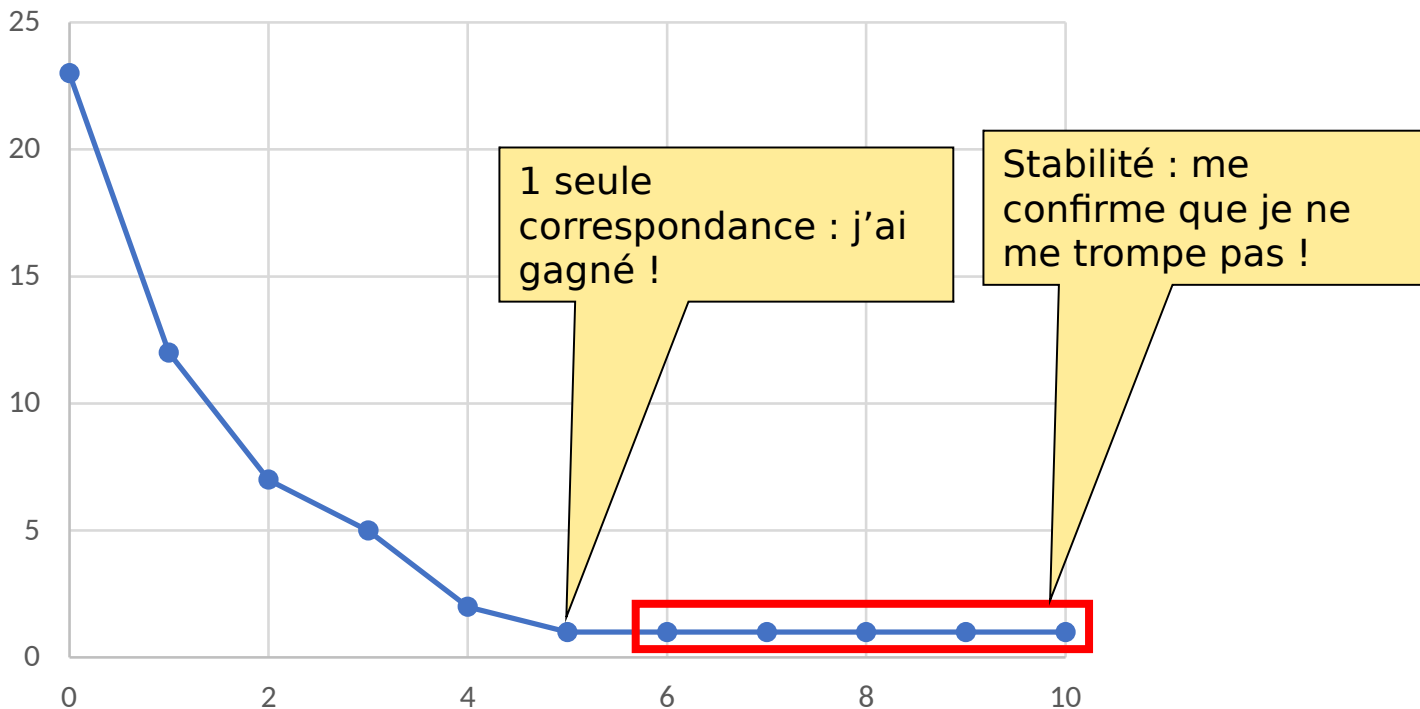
Je sais que :

- c'est un monsieur
- Il porte un chapeau
- On ne voit pas son nez
- Il est orange
- Il n'a pas de canne
- C'est Monsieur Chatouille !



# Rappelez-vous « Qui est-ce ? »

- Nombre de correspondances, en fonction du nombre d'informations





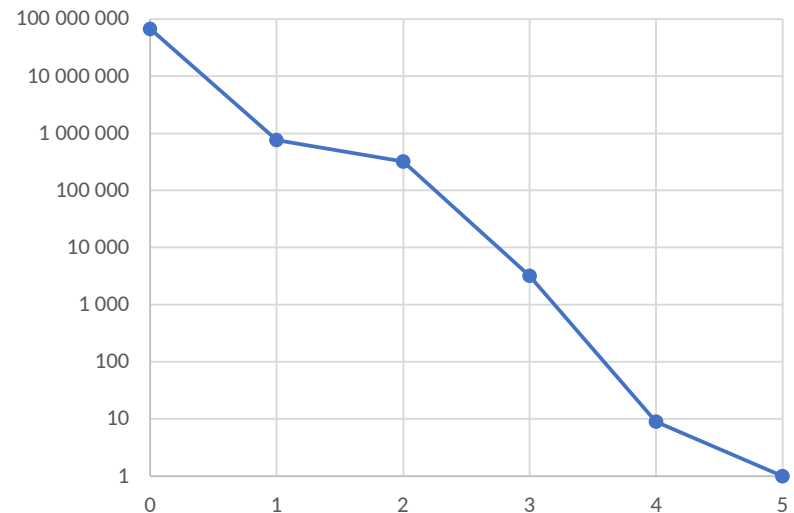
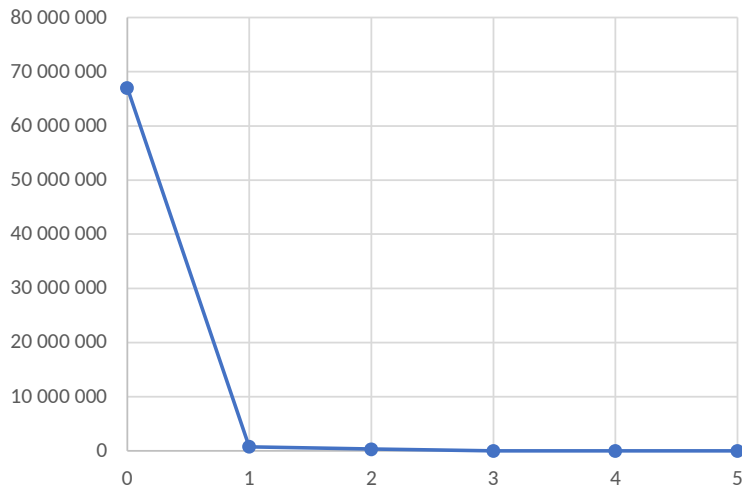
# Scénario type de réidentification

- On s'intéresse à un individu en particulier
  - Un voisin, un client d'assurance, un vendeur immobilier, un personnage public...
  - On dispose d'informations « de notoriété publique » : âge, sexe, date d'hospitalisation...
- On a accès à une base de données
  - De grand effectif, voire exhaustive (ex : Assurance Maladie)
  - Réputée non-nominative
  - Mais contenant des informations exactes
- On peut retrouver l'individu, on peut en apprendre sur lui !

# Scénario type de réidentification

On s'intéresse à un individu : Nb. corresp. dans SNDS

- Je ne sais rien 67 000 000
- Il a 22 ans 760 000
- C'est un homme 320 000
- Il vit en Indre-et-Loire (37) 3200
- Il est né le 22 janvier 9
- Il a été hospitalisé cette année 1

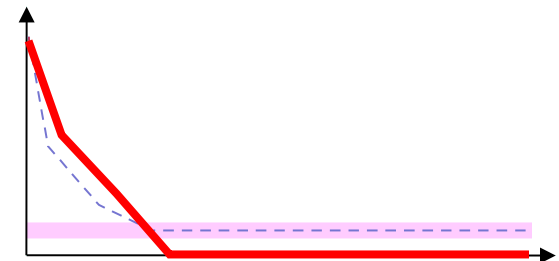
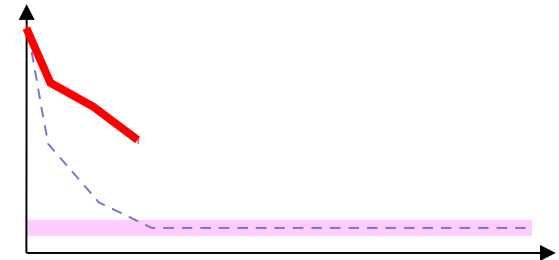
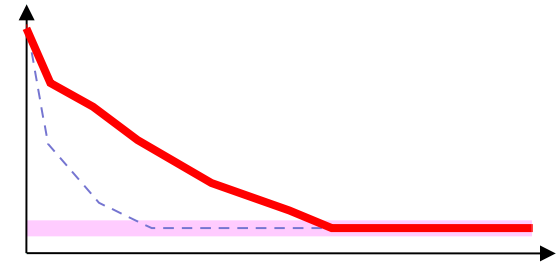


# Scénario type de réidentification

- Et où est le problème ? Exemples d'informations dans le SNDS :
- Codes diagnostiques CIM10
  - F522 Échec de la réponse génitale [insuffisance érectile]
  - F524 Éjaculation précoce
- Codes thérapeutiques CCAM
  - JHLA002 Pose d'une prothèse pénienne hydraulique sans composant extracaverneux
  - JHAA004 Allongement du pénis par section du ligament suspenseur
- Codes de dispositifs médicaux LPP
  - 3174580 Implant pénien expansible par mécanisme hydraulique
- Codes de médicaments ATC
  - G04BE08 Tadalafil
- Et bien entendu tout un tas de choses nettement plus graves !

# Solutions envisagées

- Limiter l'information disponible
  - Ex : âge = 50-60 ans
  - ... il suffit alors de croiser plus de critères !
- Diffuser des informations "anonymes" selon les juristes, i.e. agrégées
  - Ex : 430 000 hommes ont 61 ans
  - Fausses *open data* : même pas des *data* ! Inutilisable par des chercheurs.
- Notre proposition : perturber les données
  - Ex : âge = 53.567 ans
  - Perturbation aléatoire, imprédictible et irréversible, empêche toute réidentification
  - Préserve les associations statistiques (OK pour la recherche)

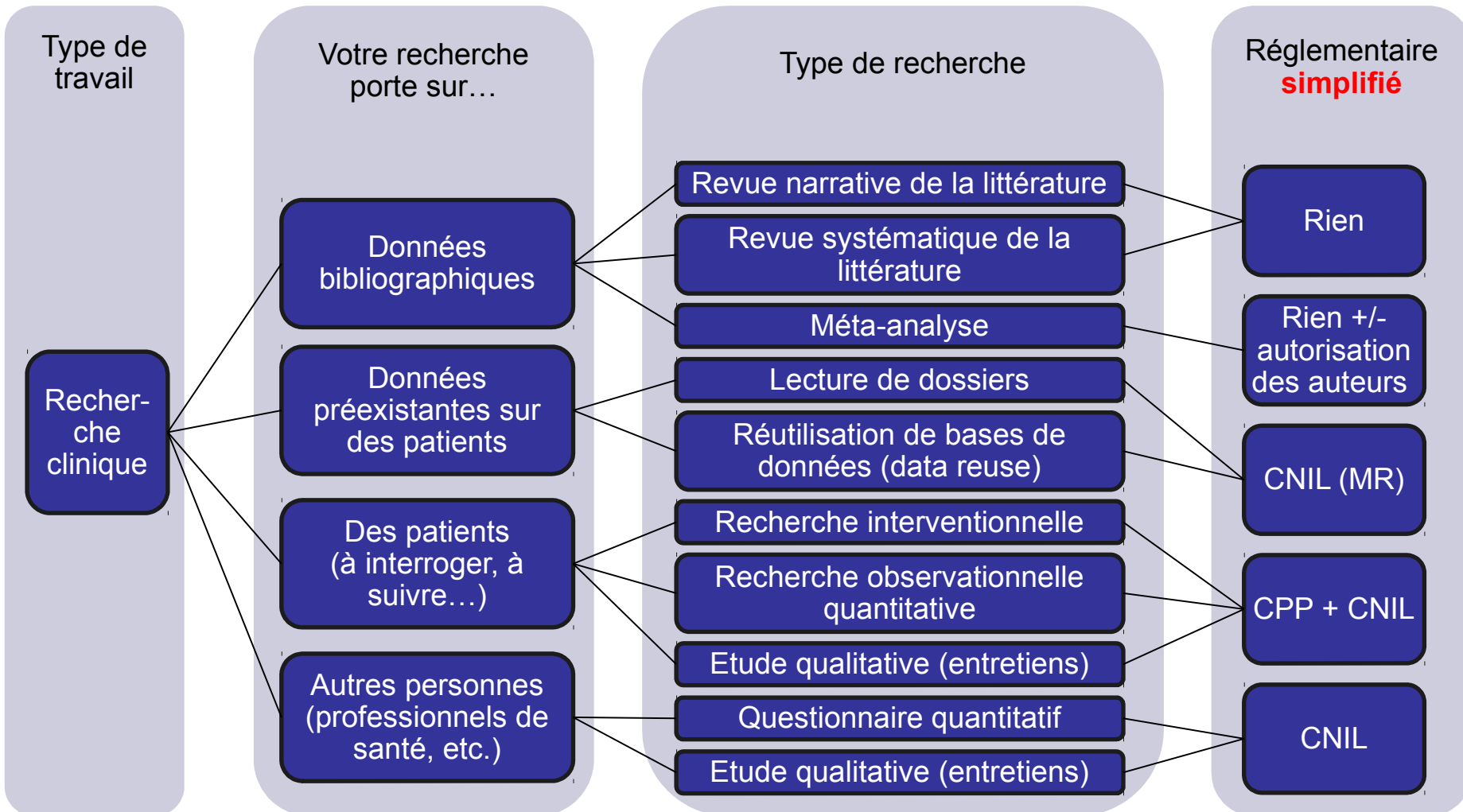




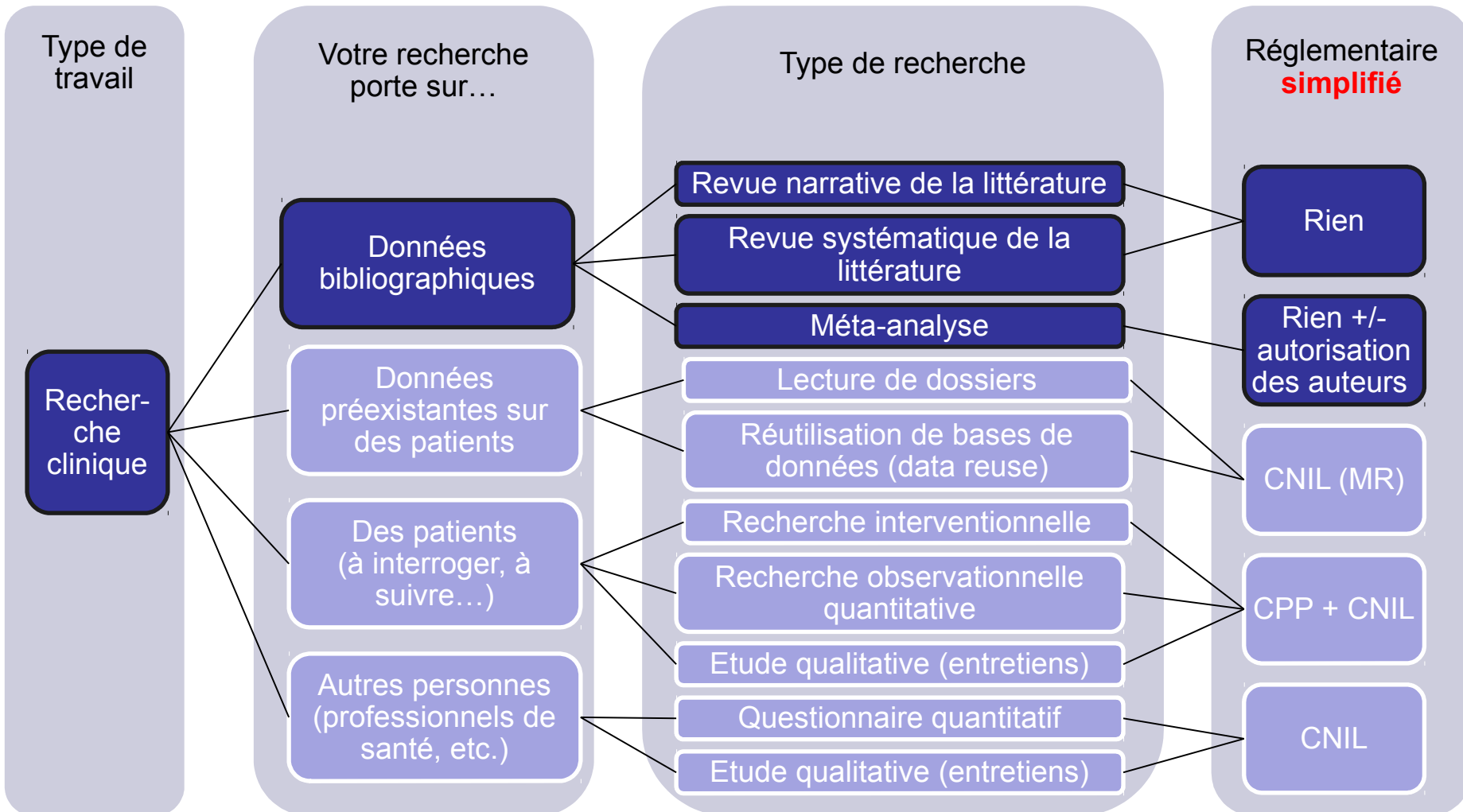
# **(2) Cadre réglementaire français**

**Recherche en santé  
« recherche clinique »**

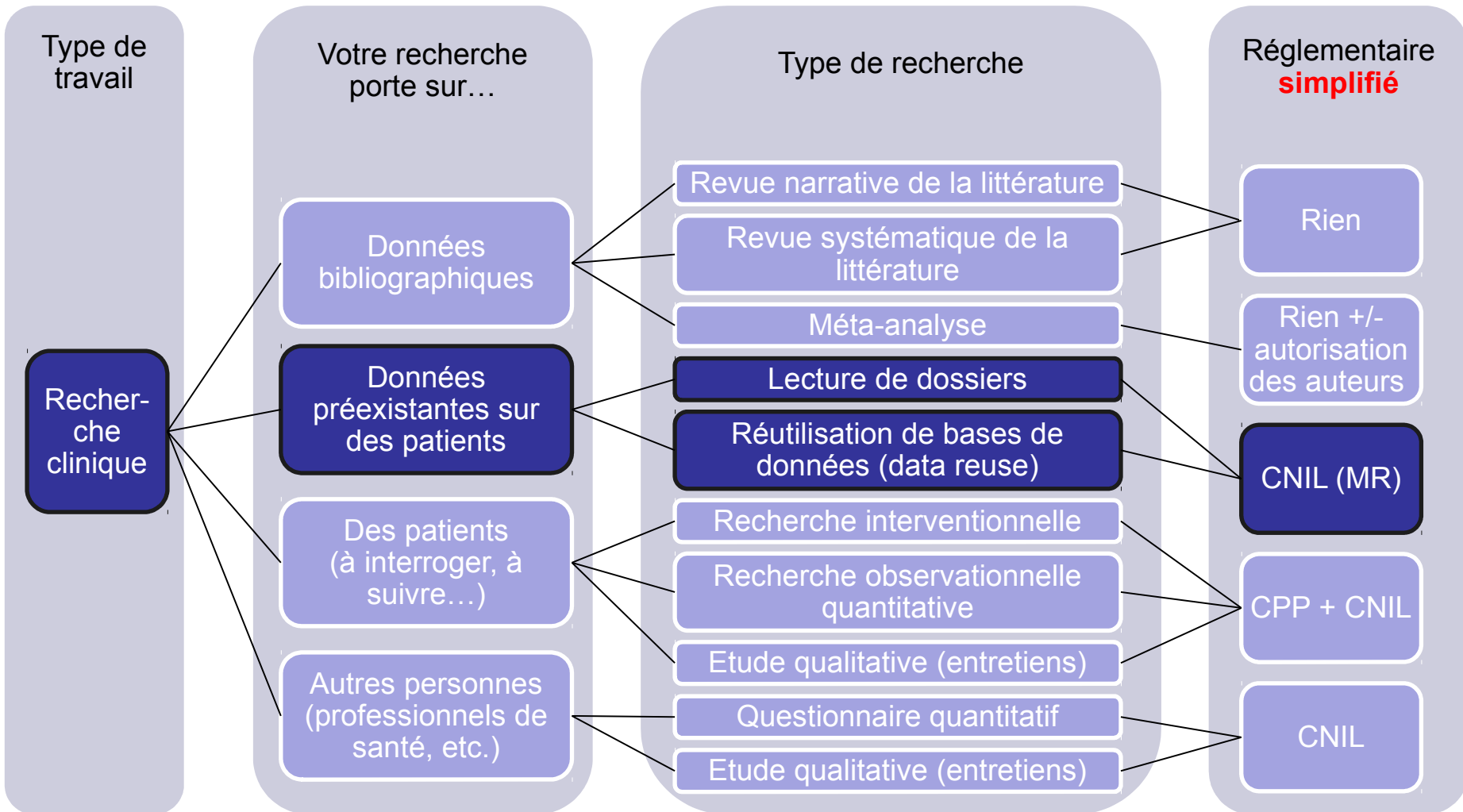
# Typologie de la recherche clinique



# Typologie de la recherche clinique

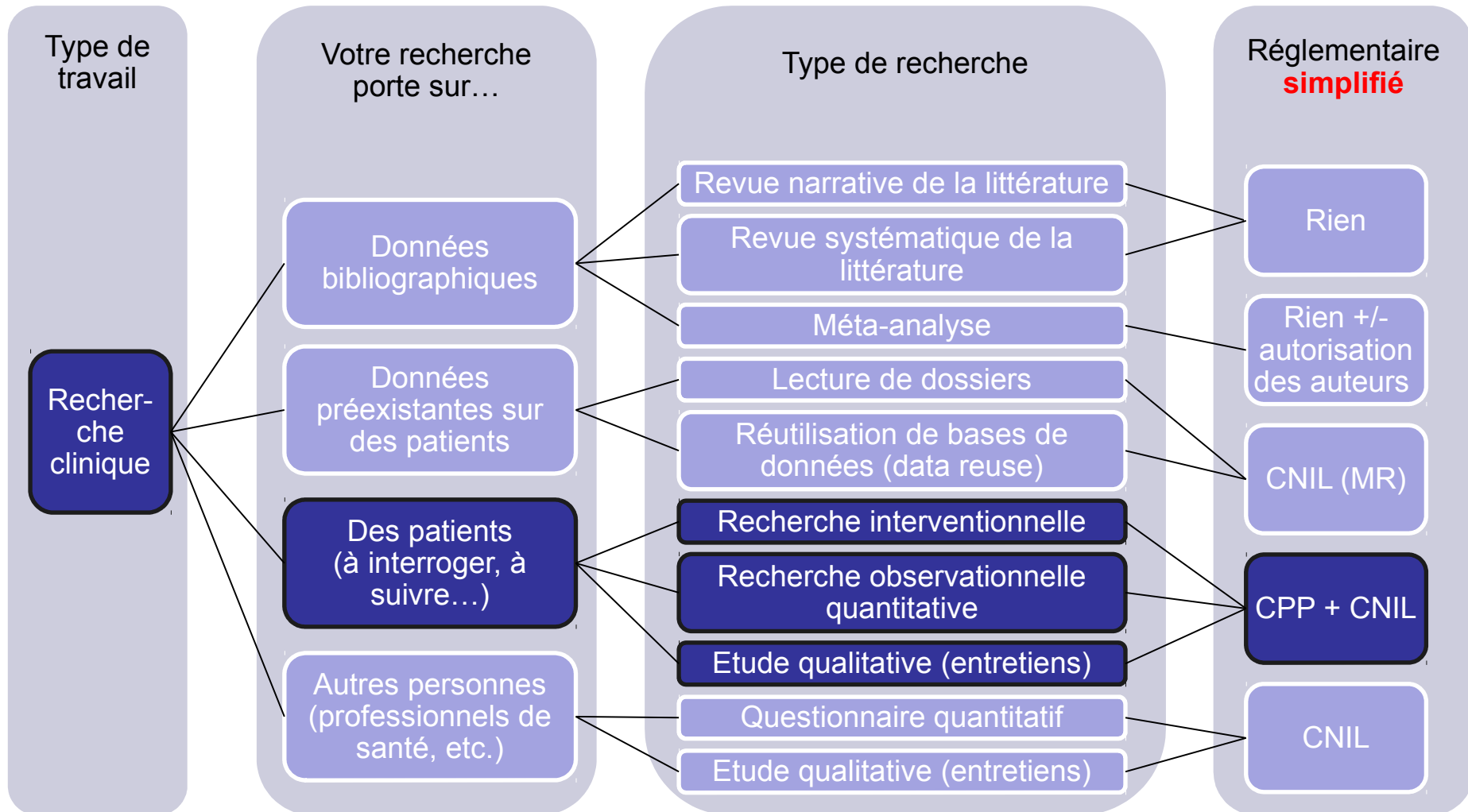


# Typologie de la recherche clinique

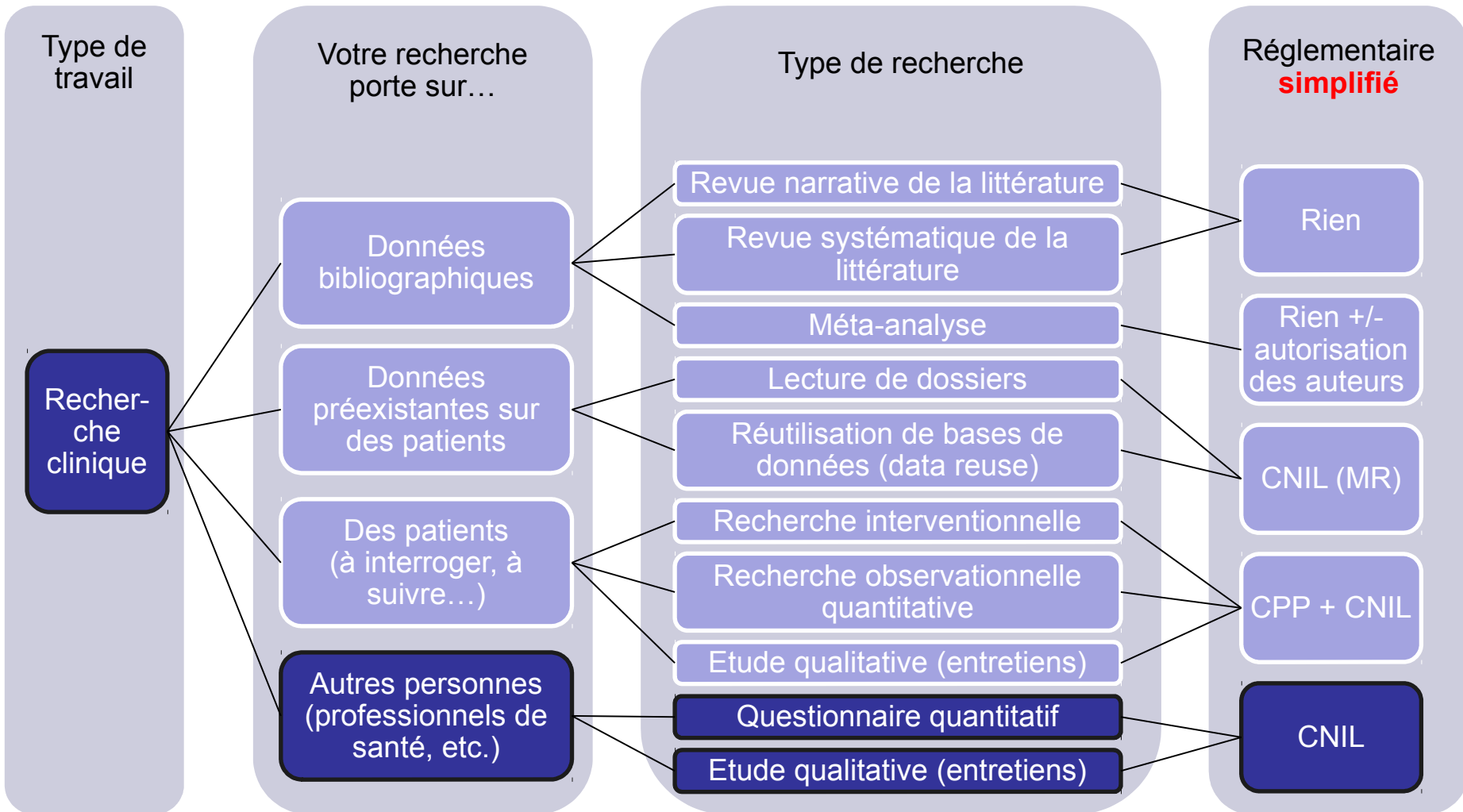




# Typologie de la recherche clinique



# Typologie de la recherche clinique



# Autorisation du CPP, le Comité de Protection des Personnes

- Obligatoire dès qu'on mène une recherche « sur des personnes »
  - Recherches interventionnelles (on modifie la prise en charge)
  - Recherches observationnelles amenant le chercheur au contact des patients :
    - Examens cliniques ou paracliniques
    - Même un questionnaire : « prélèvement psychologique »
- Motivations :
  - Essentiellement éthiques : quel est le risque pris par les patients ? Est-il proportionnel à l'intérêt de la recherche ?
  - Concerne donc : autorisation, protocole, nombre de sujets nécessaires...

# Autorisation de la CNIL, Commission Nationale de l'Informatique et des Libertés

- Procédure mise en conformité avec le RGPD
- Obligatoire pour les recherches « sur des personnes »
- Également obligatoire pour les recherches « sur des données »
  - Lecture de dossier patient papier
  - Réutilisation de données
- Objectifs et portée de l'autorisation :
  - Recueil de l'autorisation expresse des sujets « opt in », (ou de l'absence d'opposition si finalité de santé publique, « opt out »)
  - Interdiction de la collecte de certaines données (ethniques, religieuses...)
  - Limite au maximum de la collecte des données utiles à la recherche
  - Sécurité du stockage
  - Absence ou sécurisation des échanges de données
  - Absence ou encadrement des recoupements de données
  - Absence de diffusion des données natives
  - Intensité de l'agrégation des données en résultats publiables
  - Destruction planifiée des données



# Autorisation de la CNIL, Commission Nationale de l'Informatique et des Libertés

- Soit Autorisation ad hoc
- Soit Méthodologies de référence « pour simplifier et accélérer ».  
Exemples :
- MR004 :
  - Avec information des patients et « opt out » si finalité de santé publique
  - Pour l'analyse de données de santé internes à un établissement par exemple
- MR005
  - Avec contraintes gérées par l'ATIH et l'Assurance Maladie
  - Pour l'analyse de données de santé du SNDS (système national des données de santé, incluant le SNIIRAM) et de la base nationale du PMSI (programme de médicalisation des systèmes d'information)
  - Traitements à distance sur des plateformes mises à disposition par l'ATIH et l'Assurance Maladie

**(3) La Loi protège les personnes mais fait obstacle à la recherche...**

# Obstacle à la diffusion de données de santé réelles

- Préjudice pour les recherches méthodologiques
- Préjudice pour les recherches appliquées, et donc pour les patients français

# (1) Préjudice pour les recherches méthodologiques

- Les équipes méthodologiques se cantonnent aux mêmes jeux d'entraînement
  - Ex : iris, mtcars, tournois de Wimbledon, élection des Présidents des USA, diverses banques d'images
- Sur-implication dans les champs plus homogènes
  - Exemples : imagerie médicale, textes en langage naturel
  - Explications :
    - schéma de données constant
    - jeux d'entraînement internationaux
    - paradoxalement, requiert moins d'interaction avec le monde de la Santé
- Peu de lien entre recherches méthodologiques et recherches appliquées
  - Difficile et peu utile d'appliquer des méthodes développées sur des données peu réalistes
  - Voire déconnexion entre les champs investigués par les chercheurs, et les besoins de la population

# (2) Préjudice pour les recherches appliquées, donc pour la population

- Impossible d'accès aux données pour les travaux académiques
  - En France, sur-investigation des méthodologies peu contributives (ex : études qualitatives, questionnaires soumis à des professionnels, etc.)
  - A l'étranger, jamais d'analyse de données françaises
- Difficultés d'accès aux données pour les travaux de recherche financés
  - Gaspillage de temps et de moyens
  - Recherches menées sur des données USA => meilleure connaissance des patients USA
  - Obstacle à la recherche sur des données française => obstacle à l'acquisition de meilleures connaissances sur les patients français

# Nos objectifs

- Objectif opérationnel général :
  - Diffuser des jeux de données « légalement »
  - Diffuser des services et formations liés
- Objectifs stratégiques : contribuer à...
  - Dynamiser la recherche (française notamment) sur les données de santé, fondamentale ou appliquée
    - En diffusant des données fidèles sur la forme
    - En formant les équipes méthodologiques aux données et problèmes de santé
  - Améliorer la connaissance des problèmes de santé affectant les Français, notamment à travers des recherches menées à l'Etranger
    - En diffusant des données fidèles sur le fond


# **(4) La plateforme e-Santé du CPER Tec'Santé**



# Structuration

- CPER Tec'Santé (contrat de plan Etat-Région)
  - Initié par Régis Beuscart
  - Porté par Nicolas Blanchemain et Anthony Trezeibre
- 3 plateformes dont...
- La plateforme e-Santé
  - Portée par Emmanuel Chazard et Jean-Baptiste Beuscart

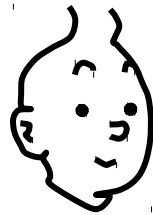
# Diffuser légalement des données de santé pour la recherche et les mémoires académiques



**Données de santé**

Fidèles en forme et en fond  
Intéressantes pour tout

- Diffusion libre : illégale
- Diffusion autorisée : complexe. Comment savoir si cela vaut le coup ?



## Données fictives fidèles en forme

Données éventuellement de haute résolution  
Intérêt : recherches méthodologiques, applications opérationnelles d'interopérabilité



## Données agrégées exactes en fond

Intérêt : certaines recherches appliquées en Santé



## Données perturbées

Données fidèles en forme, et irréelles mais très réalistes en fond  
Intérêt : preuve de concept simple et rapide  
Possibilité d'acquérir secondairement le véritable jeu de données



## Formations

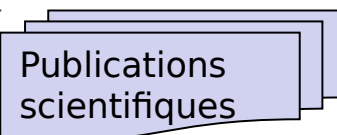
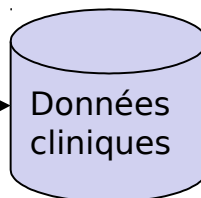
Acculturation en santé pour les équipes méthodologiques

# Exemple de workflow

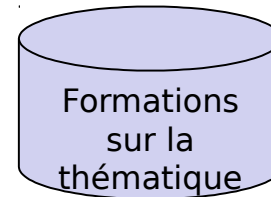
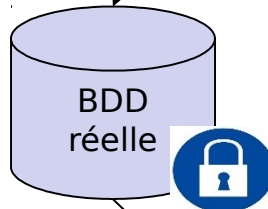
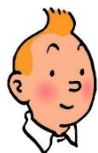
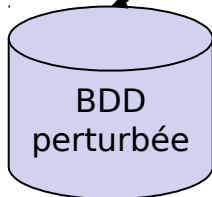
## Projet tiers de recherche clinique



Chercheurs cliniciens



## Plateforme e-santé



## Nouveau projet de recherche

Accès libre et gratuit

Si besoin et si essai concluant, accès payant et sur autorisation

Accès libre et gratuit

# Structuration

**WP1 – Coordination et management**

**WP2 – Structuration et organisation**  
(plateforme de partage, niveaux de sécurité, pré-travail données)

**WP3 – Juridique**  
Réglementation

**WP4 – Simulation**  
Simulation sur lois  
Avatars

**WP5 – Haute résolution**  
données individuelles  
fictives de très haute  
résolution technique  
(logiciels)

**WP6 – Facteurs humains**  
Facteurs humains  
Données qualitatives

**WP7 – Spatial**  
Mise en continuité  
de données  
discrètes

**WP8 – Application**  
Mise en service  
Logiciels  
Modèle économique

**WP9 – Formation**  
Aux problèmes de santé,  
aux données de santé,  
à l'analyse de données

**WP10 – Valorisation, communication,  
publications**

# Consortium

- ULR 2694 – METRICS
  - Beuscart Jean-Baptiste, Chazard Emmanuel, Génin Michaël, Pelayo Sylvia, Marot Guillemette, Canon Valentine, Benjamin Guinhouya, Dhifli Wajdi, Djamel Zitouni, Ficheur Grégoire, Vilhelm Christian
- UAR 2014 – US41 - PLBS
  - Pericard Pierre
- OSL/CRISTAL
  - Hammadi Slim, Hayfa Zgaya, Ben Othman Sarah
- LAMIH CNRS UMR 8201
  - Chaabane Sondes
- ULR7365 - GRITA
  - Odou Pascal, Décaudin Bertrand, Masse Morgane, Cousein Etienne
- CERI SN
  - Fleury Anthony, Lecomte Sylvain, Desertot Mikael, Allaert Benjamin

# Quelques catégories de données

- Données structurées (en tables) de soins
  - Hospitalières
  - Extrahospitalières
  - Issues de bases nationales (PMSI, SNDS)
- Données d'imagerie médicale
- Données -omiques
- Données d'études qualitatives liées aux soins (ou aux soignants) : verbatims et concepts formalisés
- Données de santé non-liées au soin
  - Données de réseaux sociaux
  - Environnementales
  - Open data

# Plateforme du CERN pour le partage de données

The screenshot shows the OpenData CERN search interface. At the top left is the 'opendata CERN' logo. A search bar contains the word 'electron' with a magnifying glass icon to its right. On the top right, there are links for 'Help' and 'About'. Below the search bar, on the left side, is a filter panel. It includes a checkbox for 'include on-demand datasets'. Under 'Filter by type', there are several categories with counts: Dataset (1168), Documentation (8), Environment (15), and Software (6). Each category has sub-items with their own counts. At the bottom of the filter panel is the text 'Filter by experiment'. On the right side, there are sorting and display options: 'Sort by: Best match' and 'asc.', and 'Display: detailed' and '20 results'. Below these options, it says 'Found 1203 results.'. Two search results are shown. The first is titled 'W to electron and neutrino from 2011' and includes a description and three tags: 'Dataset', 'Derived', and 'CMS'. The second is titled 'OPERA electron neutrino event 12092012487' and includes a description and three tags: 'Dataset', 'Derived', and 'OPERA'.

opendata  
CERN

electron

Help About

include on-demand datasets

Filter by type

- Dataset 1168
  - Collision 182
  - Derived 892
  - Simulated 94
- Documentation 8
  - About 1
  - Guide 7
- Environment 15
  - Validation 15
- Glossary 4
- News 2
- Software 6
  - Analysis 4
  - Tool 2
  - Workflow 2

Filter by experiment

Sort by: Best match asc.  
Display: detailed 20 results

Found 1203 results.

W to electron and neutrino from 2011

This document contains 100k events where one electron candidate was observed and missing energy implies to the presence of a neutrino. The data was selected from the primary dataset SingleElectron ...

Dataset Derived CMS

OPERA electron neutrino event 12092012487

This OPERA detector event is a electron neutrino interaction with the lead target. The event data consist of Electronic Detector files (such as Drift Tube, RPC, and Target Tracker files) and Emulsion ...

Dataset Derived OPERA



# **(5) Le projet « Health Dataset twins »**

# Cadre

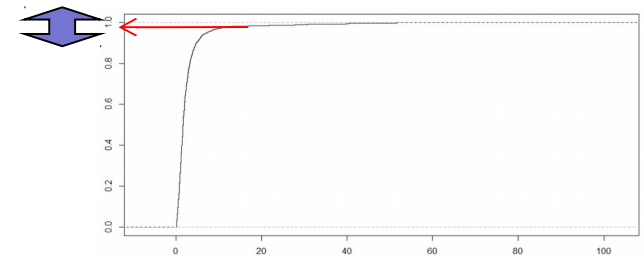
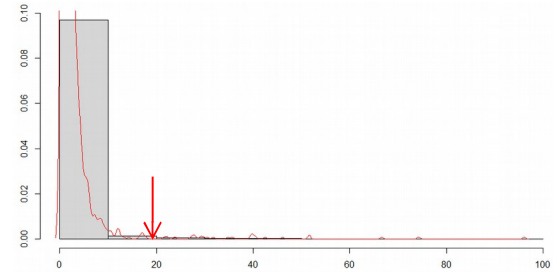
- Réponse à un appel à projet interne au CPER Tec'Santé
- Financement 90k€
- 3 partenaires :
  - ULR 2694 – METRICS : Evaluation des technologies de santé et des pratiques médicales.
  - CERI SN – IMT Nord Europe
  - UMR CNRS 7338 Biomécanique et Bioingénierie – Equipe C2MUST, UTC, Compiègne

# Structuration

- WP1 : perturbation de données
  - procédés unitaires de perturbation univariée de données réelles, génériques et auto-paramétrables
- WP2 : simulation de jeux de données
  - Une tâche par jeu de données
  - Exemple ici : T2.3 : entrepôt de données intrahospitalier incluant les mouvements du patient, les résultats d'analyse biologiques et les médicaments administrés
- WP3 : évaluation des jeux de données simulés
  - Qualité : un expert peut-il distinguer les données réelles des données simulées ?
  - Utilisabilité : les résultats d'analyse statistique obtenus sur données réelles sont-ils reproductibles sur données simulées ?
  - Sécurité ressentie : un « hacker » qui exploite les données simulées peut-il croire identifier des personnes réelles ?
- WP4 : test de chargement en vue de la diffusion des jeux de données
  - Sur la plateforme du CERN
  - Documentation du jeu de données

# WP1 : procédés unitaires de perturbation de données

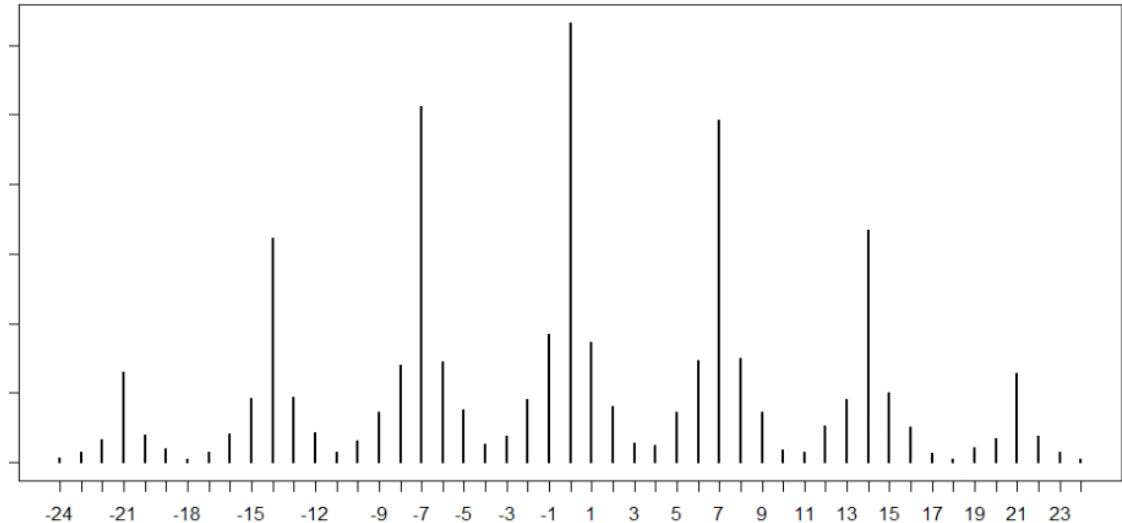
- Exemple pour une variable quantitative continue (ex : TSH) :
  - Entrée : 1 valeur et la série (distribution empirique)
  - Déduire le quantile
  - Perturber le quantile avec une loi uniforme [gérer à proximité des bornes]
  - Déduire la nouvelle valeur
- => approche totalement agnostique, auto-paramétrée
- Démarche du même genre pour quantitative discrète, binaire, qualitative, et qualitative ordonnée



# WP1 : procédés unitaires de perturbation de données

- Exemple de la date. Perturbation :

- Préserve la saison (épidémiologie)
- Préserve le jour de la semaine (logistique et réalisme)
- Pas vraiment agnostique
- Néanmoins générique



# WP2 : simulation de jeux de données

- Approche classique :
  - Modéliser d'abord, pour ensuite générer aléatoirement des mesures
  - Si distribution quelconque : la distribution empirique devient le modèle
  - OK pour simulation univariée, OK pour bivariée
  - ... mais données de santé trop complexes !
- Approche proposée :
  - Utiliser une vraie BDD (notre « loi empirique de distributions jointes »)
  - Tirage au sort avec remise de séjours (ou de patients)
  - Perturbation des séjours (ou patients) => nouveaux individus de même forme, très complets, et réalistes au fond

# WP2 : simulation de jeux de données

## ■ Exemple d'un séjour hospitalier :

### Diagnosics :

```
diag      diag_lib
E6601    Obésité due à un excès
K566     Occlusions intestinales
T8138    Désunions d'une plaie
```

### Actes :

```
act      delay_act  act_lib
YYYY600 0        Supplément pour archivage numérique d'une m
```

```
ZCQH001
ZCQK002
```

### Médicaments administrés :

```
name      atc      delay_drug  dose      unit      route
ESOMEPR... 20 MG ARBO... CÉLUIE... RÉTOMANT...
METOCL...
PERFALC...
```

### Résultats d'analyse de biologie médicale :

delay_bio	kind	value	unit	inf_bound	sup_bound
0	100% de la formule	102.182640134735		\N	\N \N
0	AGE 51.1725584131283	ans	0	50	
0	Bilirubine totale	4.01954742842377	mg/L		0 10
0	C.C.M.H.	33.3808430646576	g/dL		31 35
0	Chlore	97.1283246030867	mEq/l	98	107
0	Créatinine	7.76596534879122	mg/L		5 9
0	Erythrobl/100 leuco	0	/100leuco	\N	\N
0	Erythroblastes /mm3	0	/mm3	\N	\N
0	G.G.T.	109.256287385609	UI/L		5 36
0	Hématocrite	39.5593302899573	%	37	47
0	Hémoglobine	13.8824924997465	g/dL		11.5 15
0	I.N.R.	0.998261916404654	\N	0	3.5
0	Lactate	285.843229277622	mg/L		45 198
0	LEUCOCYTES/mm3	9519.03729208286	/mm3		4000 10000
0	Lipase	28.6785326767218	UI/L		13 60
0	Lympho /mm3	1255.50794275653	/mm3		1500 4000
0	Lymphocytes	12.7793584206755	%	\N	\N
0	millions de GR/mm3	4.31161105393238		10 6	/mm3 0 5

```
Age 47.39
Sexe 0
Date entrée 2027-08-12
Via urgences 1
Date sortie 2027-08-14
```

```
Durée 2
Décès 0
Transfert 0
Diag principal K439
```

```
Libellé Hernie de la paroi abdominale
antérieure autre et non précisée,
sans occlusion ni gangrène
```

```
GHM 06M091
```

```
Libellé Autres affections digestives,
âge supérieur à 17 ans, niveau 1
```



# WP2 : simulation de jeux de données

- Stratégie de perturbation :
  - Normalisation poussée
    - Ex : date de sortie supprimée, sera recalculée depuis date entrée et durée
  - Perturbation des variables, conditionnellement
    - Ex : âge et sexe perturbés conditionnellement à la racine de groupe homogène de malade
    - Ex : valeur de biologie perturbée conditionnellement au type de mesure
  - Pour les diagnostics :
    - Suppression des codes sans intérêt médical. Ex : codes diagnostiques portant sur les circonstances d'accidents
    - Transformation des codes diagnostiques trop rares Ex : XXX.1 => XXX.9 (et recalcul des libellés)
  - Pour les dates : transformation, puis recalcul des autres dates

# WP3 : Qualité : un expert peut-il distinguer les données réelles des données simulées ?

- Interface soumettant des paires de séjour, un réel et un simulé. Un expert médical peut-il trouver le séjour simulé ?

```

age sex admission_date from_emergency_01 discharge_date duration deat
47.3891475404997 0 2027-08-12 1 2027-08-14 2 0 0 K439 Herr

Diagnostics :

diag diag_lib
E6601 Obésité due à un excès calorique de l'adulte avec indice de masse co
K566 Occlusions intestinales, autres et sans précision
T8138 Désunions d'une plaie opératoire non classées ailleurs, autres et no

Actes :

act delay_act act_lib
YYYY600 0 Supplément pour archivage numérique d'une mammographie ou d'un e
ZCQH001 0 Scanographie de l'abdomen et du petit bassin [pelvis], avec inje
ZCQK002 0 Radiographie de l'abdomen sans préparation

Médicaments administrés :

name atc delay_drug dose unit route
ESOMEPRAZOLE 20 MG ARROW, GÉLULE GASTRO-RÉSISTANTE A02BC05 0 1 gel \N
METOCLOPRAMIDE 10 MG/2 ML ADULTE RENAUDIN, SOL INJ = PRIMPERAN 10 MG/2ML INJ
PERFALGAN 1 G/100 ML, SOL PR PERF, POCHE N02BE01 0 1 poche \N

Résultats d'analyse de biologie médicale :

delay_bio kind value unit inf_bound sup_bound
0 100% de la formule 102.182640134735 \N \N \N
0 AGE 51.1725584131283 ans 0 50
0 Bilirubine totale 4.01954742842377 mg/L 0 10
0 C.C.M.H. 33.3808430646576 g/dL 31 35
0 Chlore 97.1283246030867 mEq/l 98 107
0 Créatinine 7.76596534879122 mg/L 5 9
0 Erythrobl/100 leuco 0 /100leuco \N \N
0 Erythroblastes /mm3 0 /mm3 \N \N
0 G.G.T. 109.256287385609 UI/L 5 36
0 Hématocrite 39.5593302899573 % 37 47
0 Hémoglobine 13.8824924997465 g/dL 11.5 15
0 I.N.R. 0.998261916404654 \N 0 3.5
0 Lactate 285.843229277622 mg/L 45 198
0 LEUCOCYTES/mm3 9519.03729208286 /mm3 4000 10000
    
```

```

age sex admission_date from_emergency_01 discharge_date duration deat
47.3891475404997 0 2027-08-12 1 2027-08-14 2 0 0 K439 Herr

Diagnostics :

diag diag_lib
E6601 Obésité due à un excès calorique de l'adulte avec indice de masse co
K566 Occlusions intestinales, autres et sans précision
T8138 Désunions d'une plaie opératoire non classées ailleurs, autres et no

Actes :

act delay_act act_lib
YYYY600 0 Supplément pour archivage numérique d'une mammographie ou d'un e
ZCQH001 0 Scanographie de l'abdomen et du petit bassin [pelvis], avec inje
ZCQK002 0 Radiographie de l'abdomen sans préparation

Médicaments administrés :

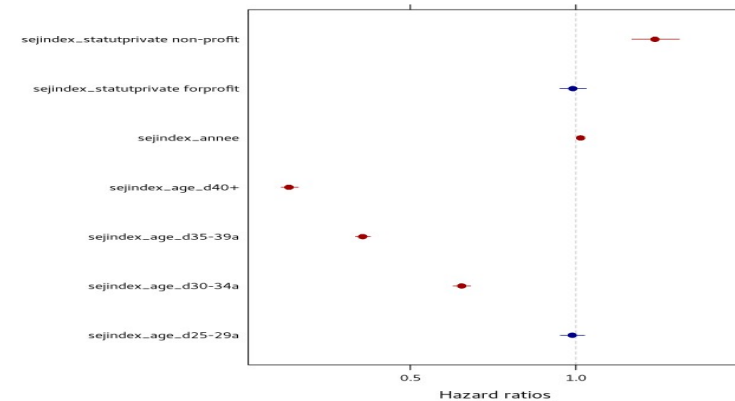
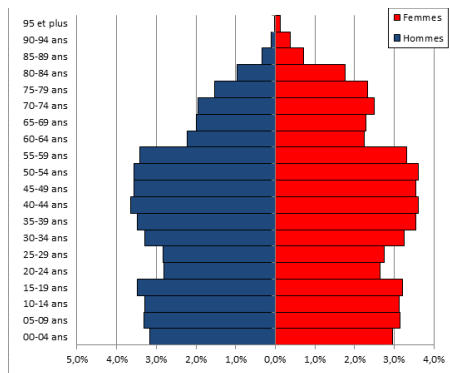
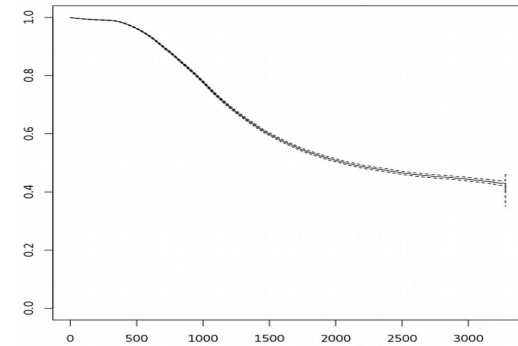
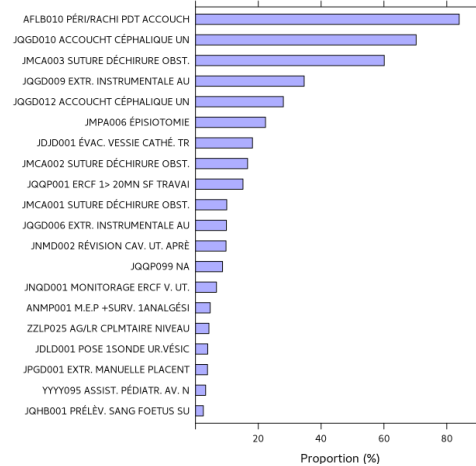
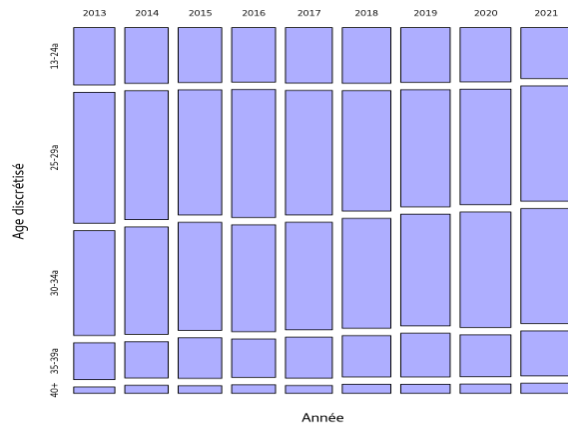
name atc delay_drug dose unit route
ESOMEPRAZOLE 20 MG ARROW, GÉLULE GASTRO-RÉSISTANTE A02BC05 0 1 gel \N
METOCLOPRAMIDE 10 MG/2 ML ADULTE RENAUDIN, SOL INJ = PRIMPERAN 10 MG/2ML INJ
PERFALGAN 1 G/100 ML, SOL PR PERF, POCHE N02BE01 0 1 poche \N

Résultats d'analyse de biologie médicale :

delay_bio kind value unit inf_bound sup_bound
0 100% de la formule 102.182640134735 \N \N \N
0 AGE 51.1725584131283 ans 0 50
0 Bilirubine totale 4.01954742842377 mg/L 0 10
0 C.C.M.H. 33.3808430646576 g/dL 31 35
0 Chlore 97.1283246030867 mEq/l 98 107
0 Créatinine 7.76596534879122 mg/L 5 9
0 Erythrobl/100 leuco 0 /100leuco \N \N
0 Erythroblastes /mm3 0 /mm3 \N \N
0 G.G.T. 109.256287385609 UI/L 5 36
0 Hématocrite 39.5593302899573 % 37 47
0 Hémoglobine 13.8824924997465 g/dL 11.5 15
0 I.N.R. 0.998261916404654 \N 0 3.5
0 Lactate 285.843229277622 mg/L 45 198
0 LEUCOCYTES/mm3 9519.03729208286 /mm3 4000 10000
    
```

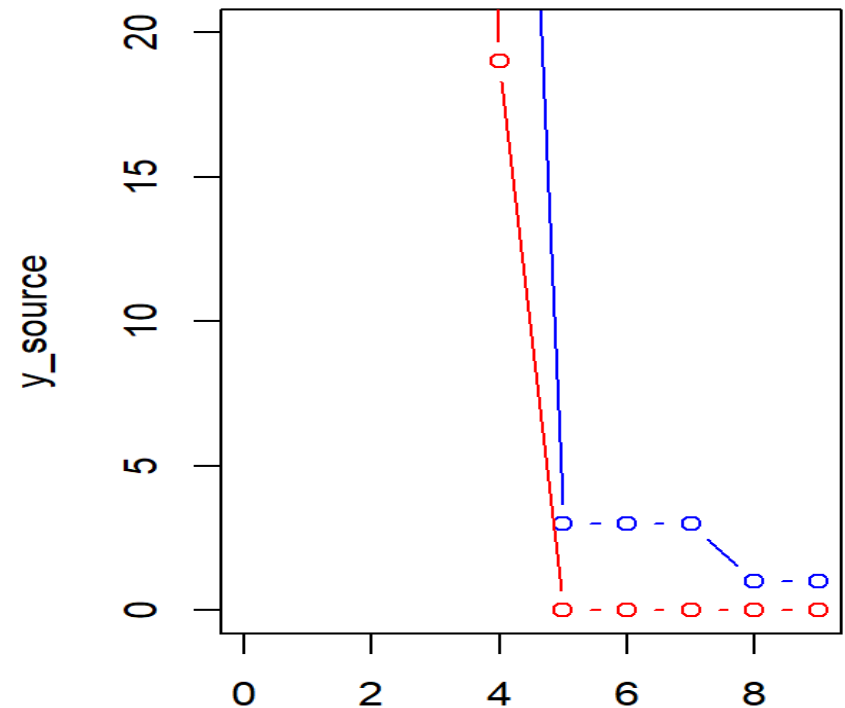
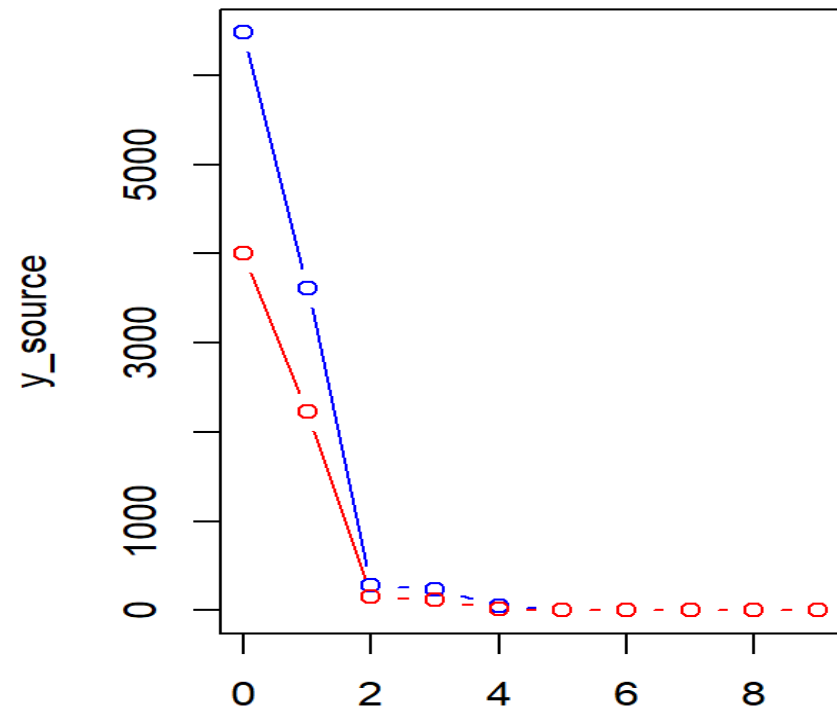
# WP3 Utilisabilité : les résultats d'analyse statistique obtenus sur données réelles sont-ils reproductibles sur données simulées ?

- Mener une analyse statistique sur un sujet médical d'intérêt (univarié, bivarié, reg logistique, reg de Cox, etc.)



# WP3 Sécurité ressentie : un « hacker » qui exploite les données simulées peut-il croire identifier des personnes réelles ?

- Nombre de médian correspondances en fonction du nombre d'observations, en partant de 6480 séjours réels
- (bleu : vraie BDD ; rouge : BDD simulées)



**Merci de votre attention !**

emmanuel.chazard@univ-lille.fr