Partage et réutilisation de données de santé pour la recherche : Principes méthodologiques et applications

Dr Jean-Charles DUFOUR

Mise à jour : février 2021

SESSTIM, Faculté des Sciences Médicales et Paramédicales, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr/

Points abordés

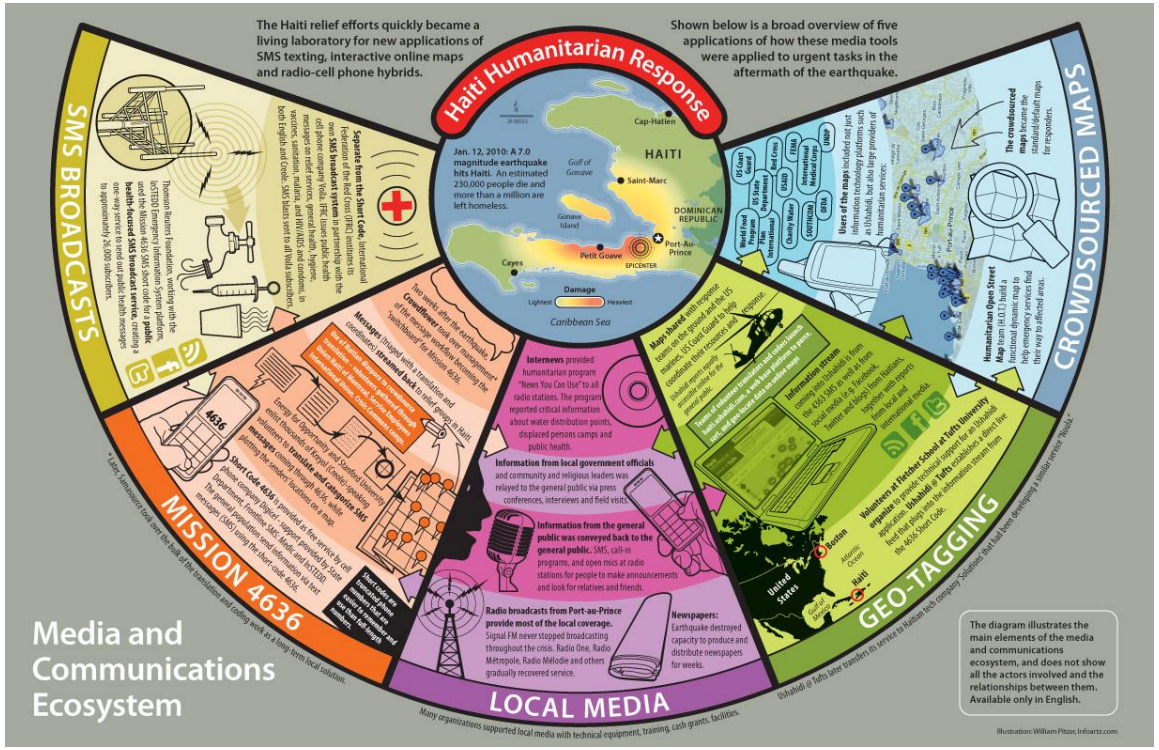
- Position du problème et définitions
- Caractéristiques des données (de santé)
- Intérêts et contraintes de la réutilisation des données
- Principes de constitution des entrepôts de données santé (EDS)
- Perspectives

Introduction

- Réutilisation (ou utilisation secondaire) des données = traiter/analyser des données avec une autre finalité que celle prévue initialement et pour laquelle elles ont déjà été collectées.

« Collect once, use many »

Introduction



https://www.theatlantic.com/technology/archive/2011/01/the-tech-used-to-help-in-haitis-earthquake-recovery/69327/

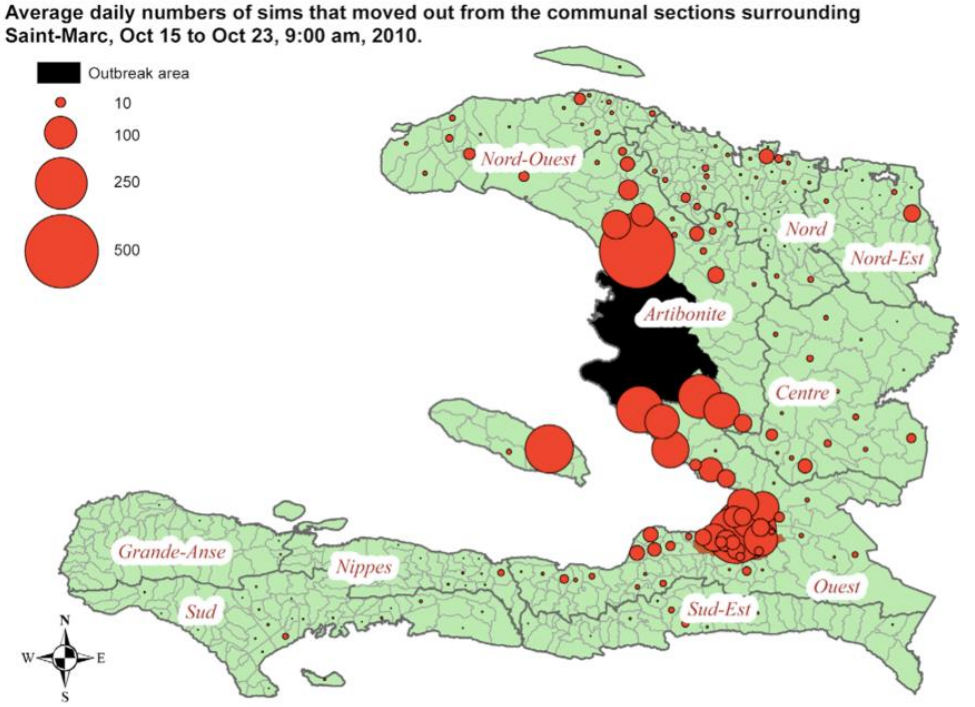
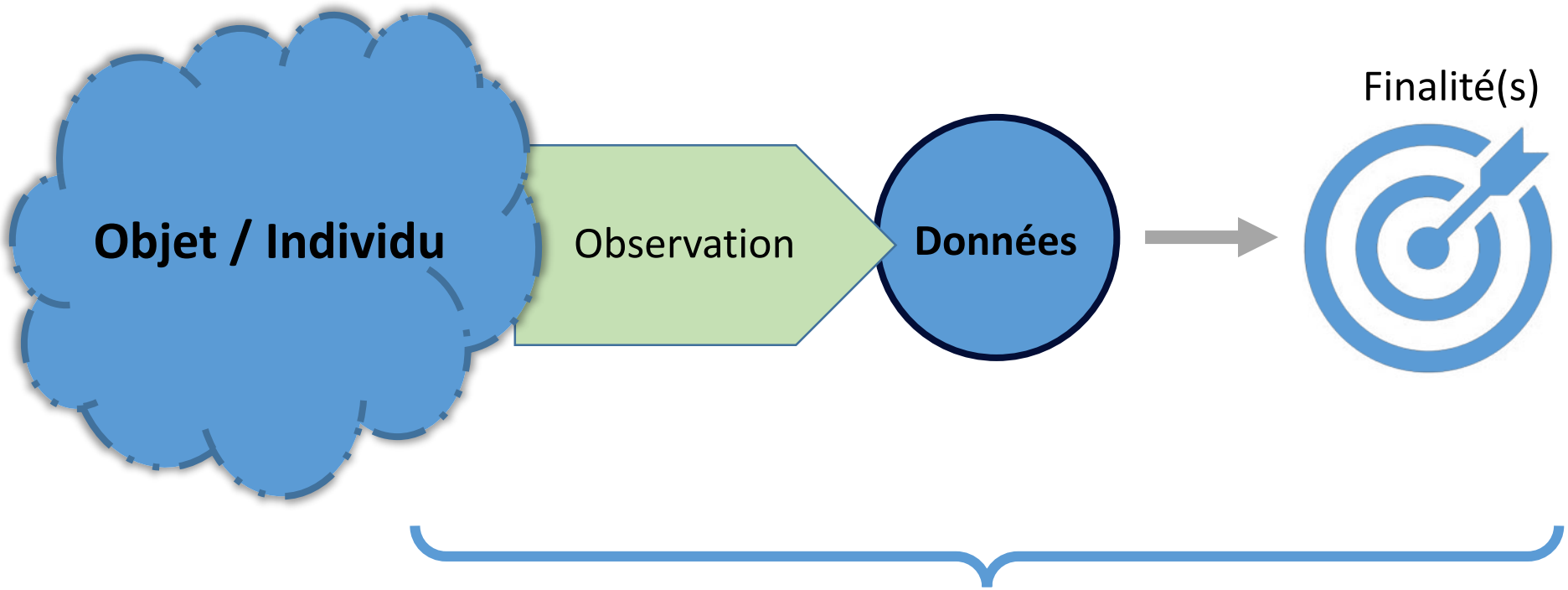
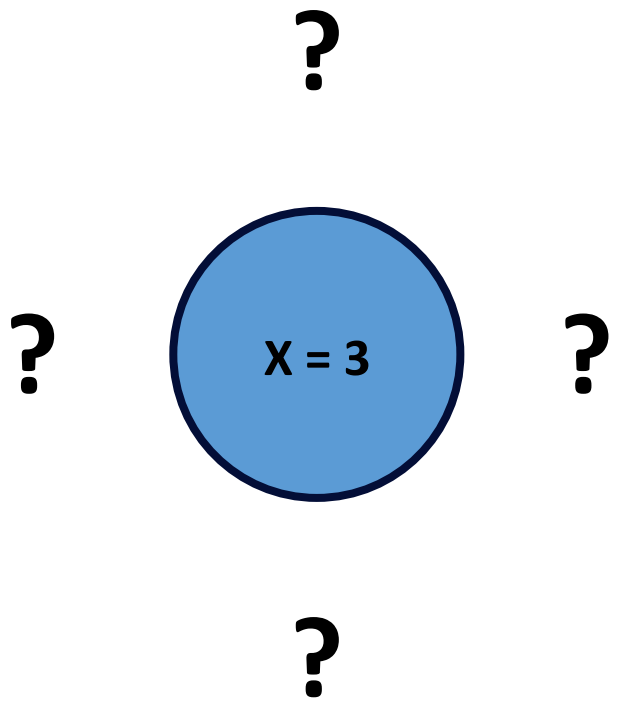


Figure 5. Average daily numbers of SIMs moving out of the cholera outbreak area. October 15 to October 23, 2010, divided per communal section of destination. The data were disseminated to relief agencies at the outset of the outbreak (October 24, 2010).
doi:10.1371/journal.pmed.1001083.g005

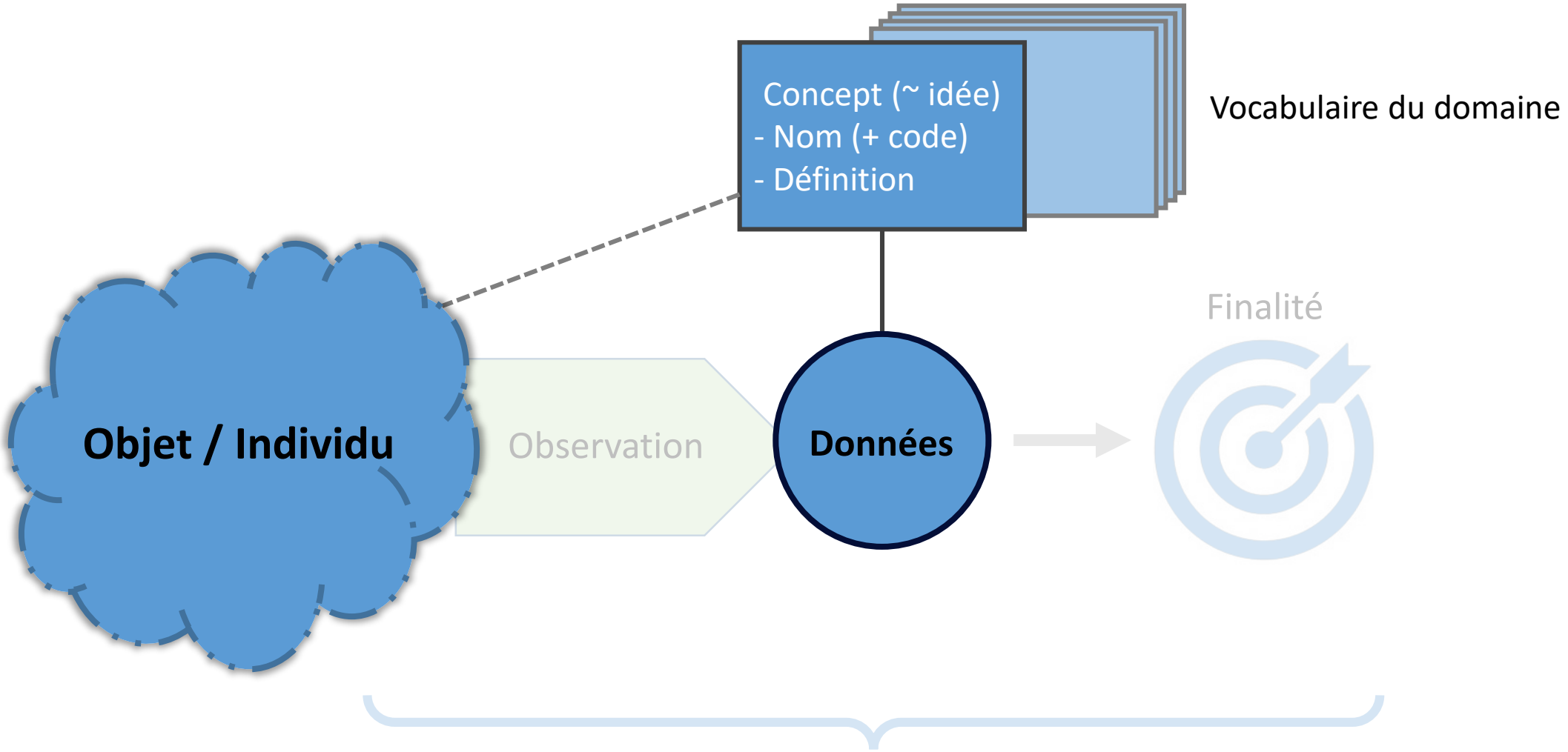
Les données sont des observations sélectives



Une donnée isolée n’a aucun sens

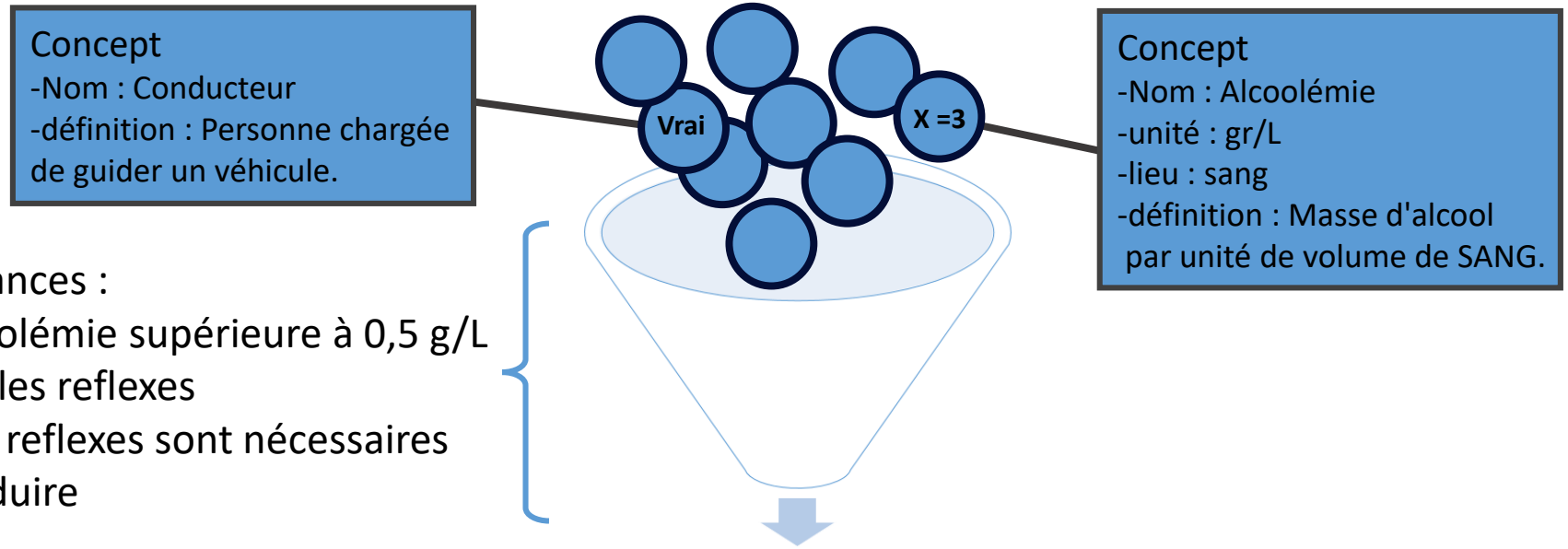


Les données sont associées à des concepts



Données + connaissances => action pertinente

- Connaissances : relations/lois/règles générales applicables à un ensemble de données



Décision
=
« ne pas prendre le volant ! »

Structure des données informatisées

- Structure « simple » (attribut : valeur)

  Exemples :

  - Alcoolémie sanguine : 3,2 g/L
  - Asthénie : Vrai
  - Couleur des yeux : Bleu

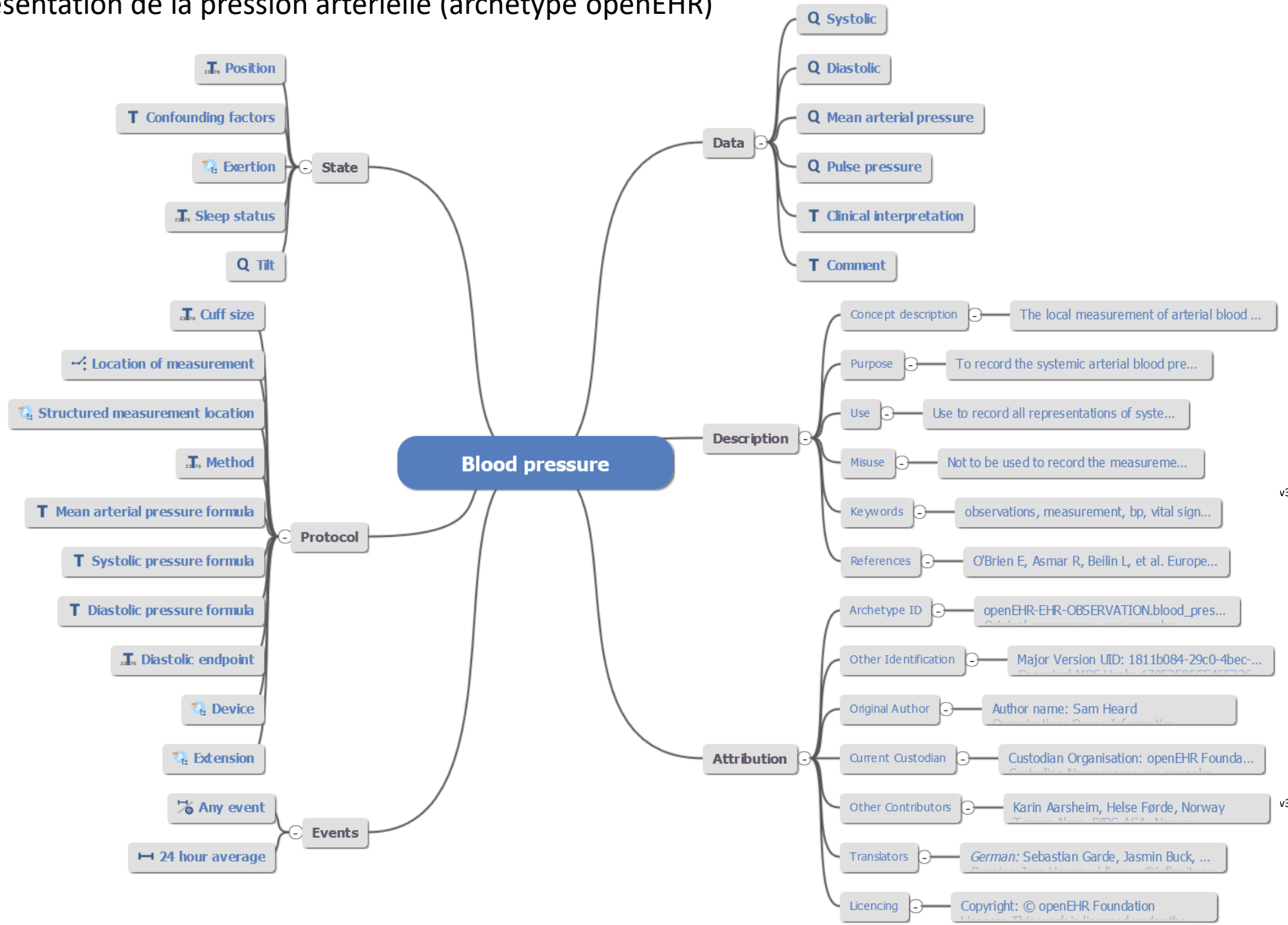
- Structure « composée »

  Exemple :

  - Pression artérielle
    - Pression systolique : 12 mmHg
    - Pression diastolique : 8 mmHg

Exemple de représentation de la pression artérielle (archétype openEHR)

https://ckm.openehr.org



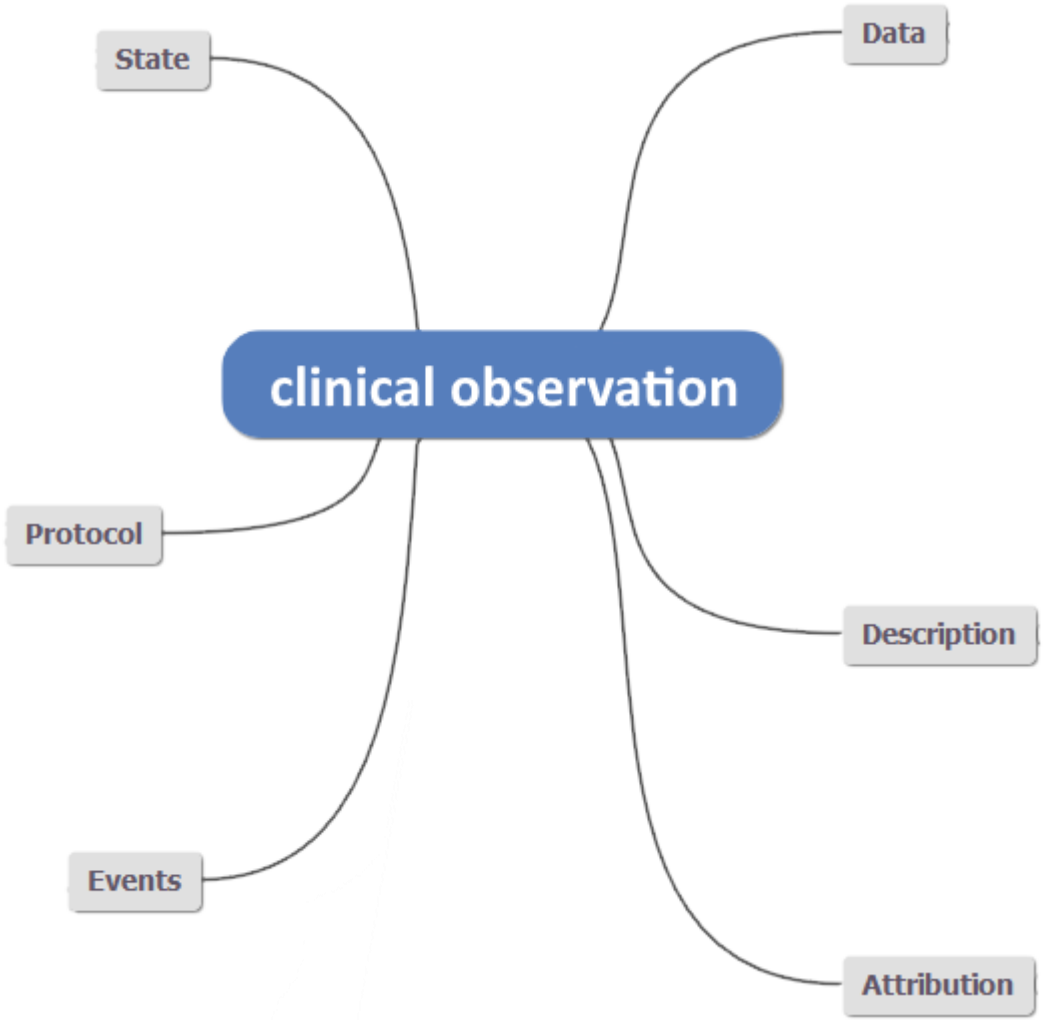
v3-ObservationInterpretation",

v3-ObservationInterpretation",

Structure <-> Modèle

Modèle

Représentation et organisation d'éléments considérés essentiels d'une réalité observée dans un contexte donné et une/des finalité(s) donnée(s)



Structure <-> Syntaxe

Syntaxe

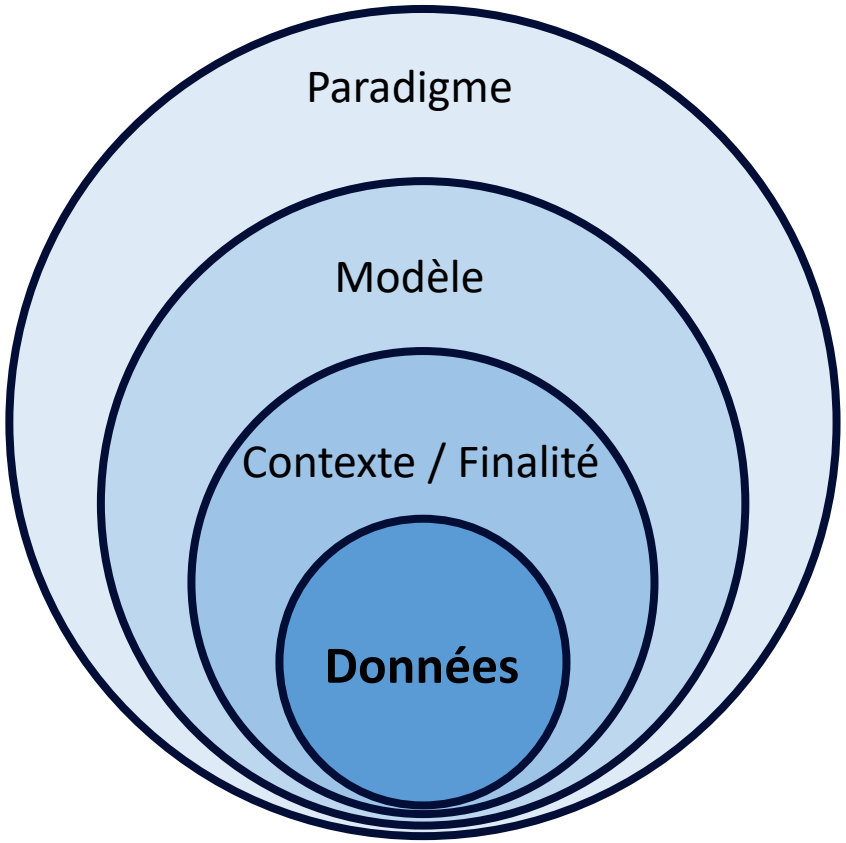
Règles précisant la manière d'écrire et/ou de disposer des informations/données

Exemple syntaxe JSON

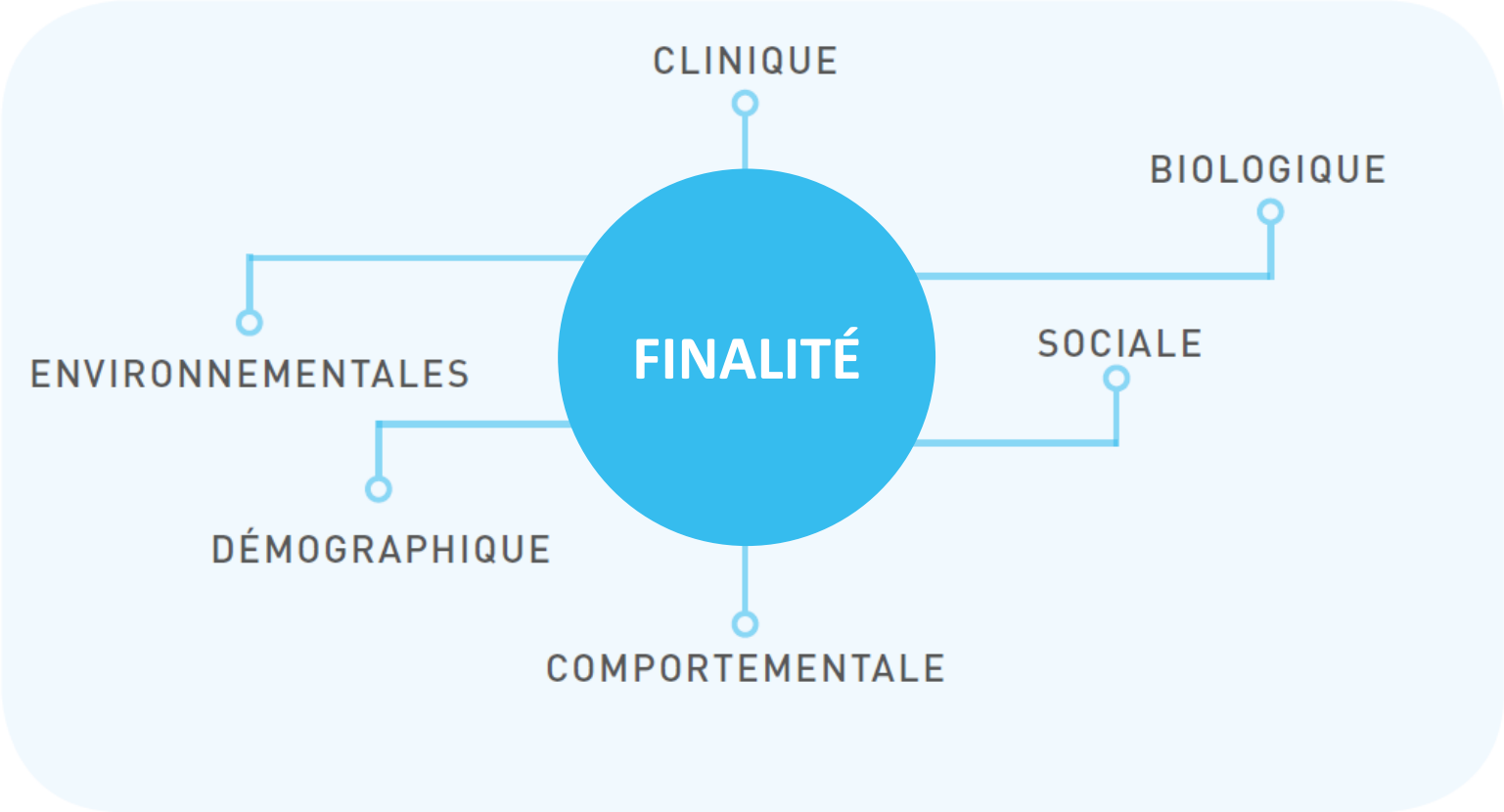
```
 "resourceType": "Observation",
 "id": "blood-pressure",
…
 "bodySite": {
  "coding": [
   {
    "system": "http://snomed.info/sct",
    "code": "368209003",
    "display": "Right arm"
…
 "component": [
  {
   "code": {
    "coding": [
     {
      "system": "http://loinc.org",
      "code": "8480-6",
      "display": "Systolic blood pressure"
…
    "valueQuantity": {
     "value": 107,
     "unit": "mmHg",
     "system": "http://unitsofmeasure.org",
     "code": "mm[Hg]"
    },
…
```

```
"display": "normal"
  {
   "code": {
    "coding": [
     {
      "system": "http://loinc.org",
      "code": "8462-4",
      "display": "Diastolic blood pressure"
…
    "valueQuantity": {
     "value": 60,
     "unit": "mmHg",
     "system": "http://unitsofmeasure.org",
     "code": "mm[Hg]"
    },
    "interpretation": [
     {
      "coding": [
       {
        "system": "http://terminology.hl7",
        "code": "L",
        "display": "low"
…
```

Les données résultent de choix

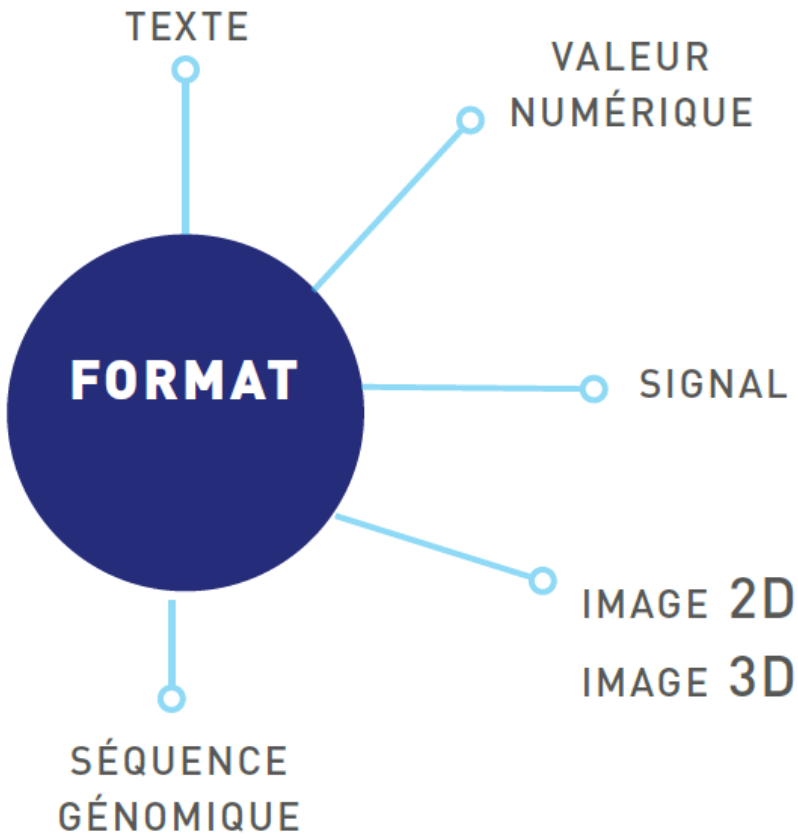


Hétérogénéité des données de santé



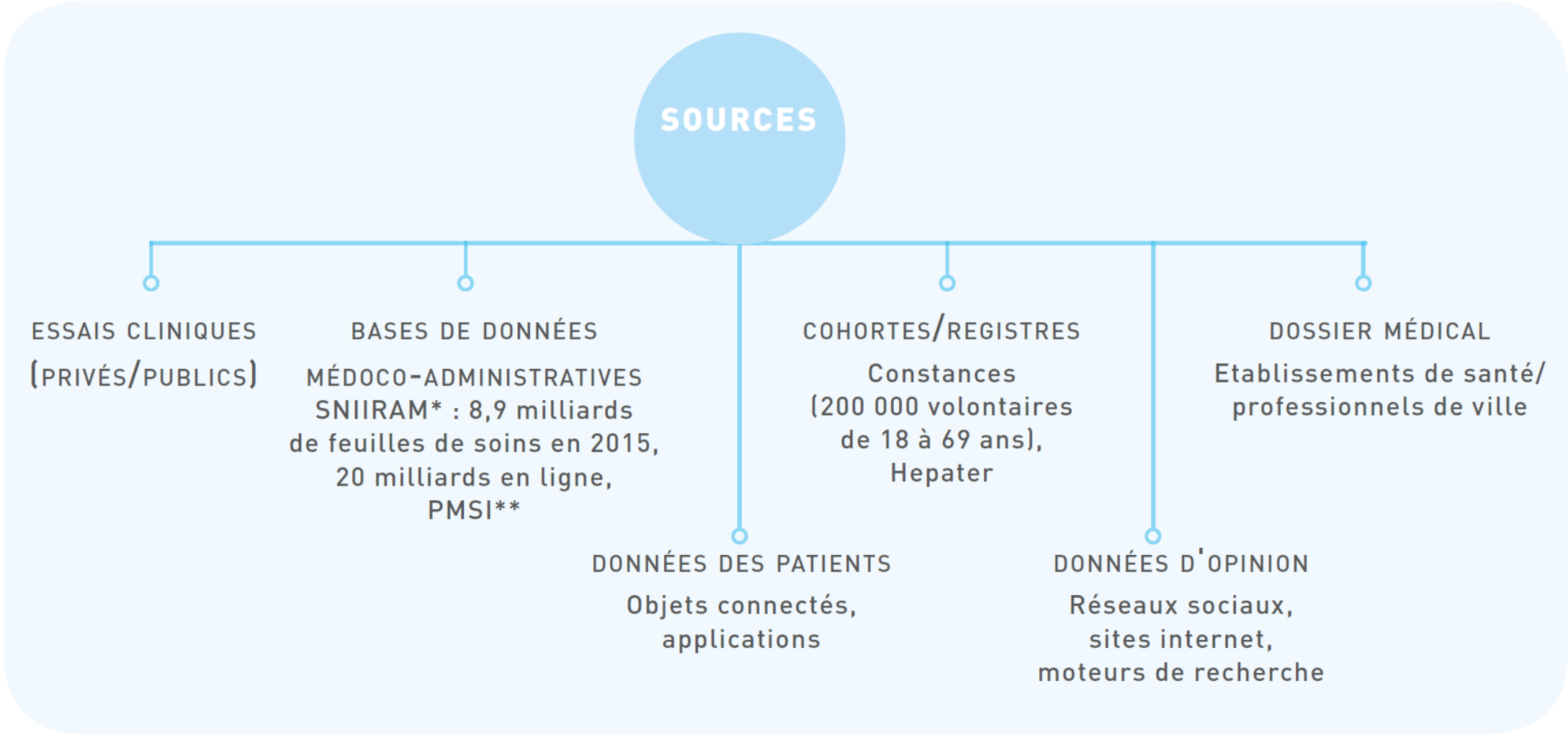
Crédits infographie : étude sur l'usage des données de santé (2018) lir.asso.fr

Hétérogénéité des données de santé



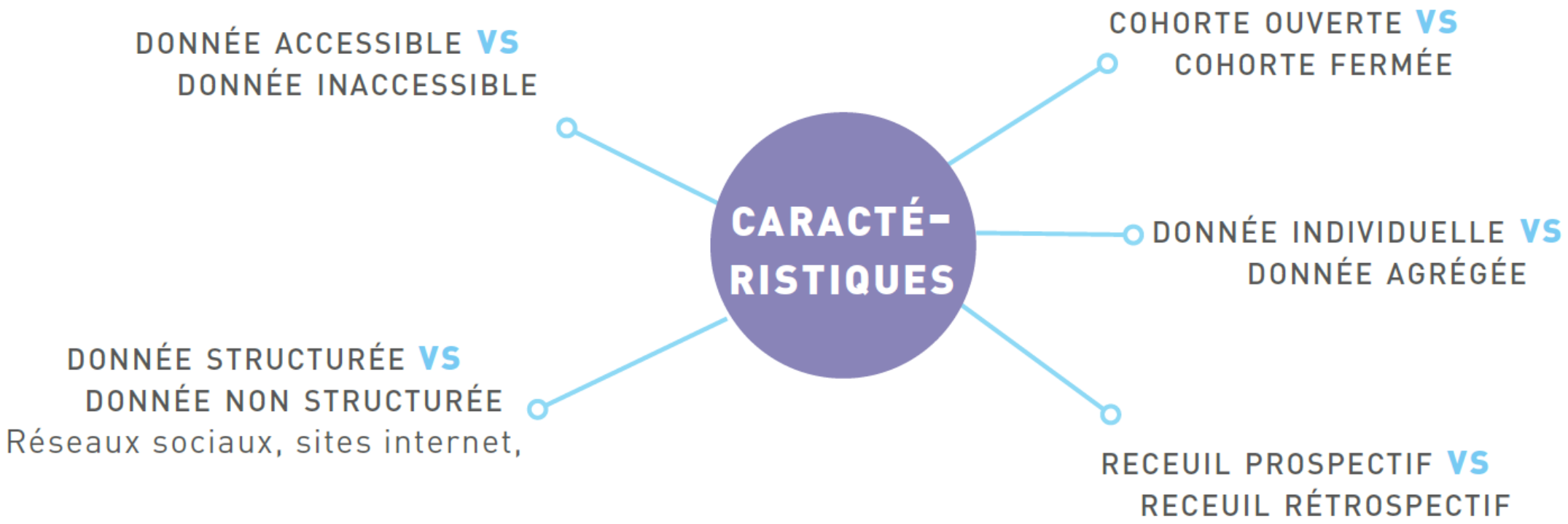
Crédits infographie : étude sur l'usage des données de santé (2018) lir.asso.fr

Hétérogénéité des données de santé



Crédits infographie : étude sur l'usage des données de santé (2018) lir.asso.fr

Hétérogénéité des données de santé



Crédits infographie : étude sur l'usage des données de santé (2018) lir.asso.fr

Evolution concernant les données et leur (ré)utilisation

Les données directement numériques sont des ressources (Big data, IA)

La recherches guidées par les données
en complément de la recherche classique guidée par les hypothèses

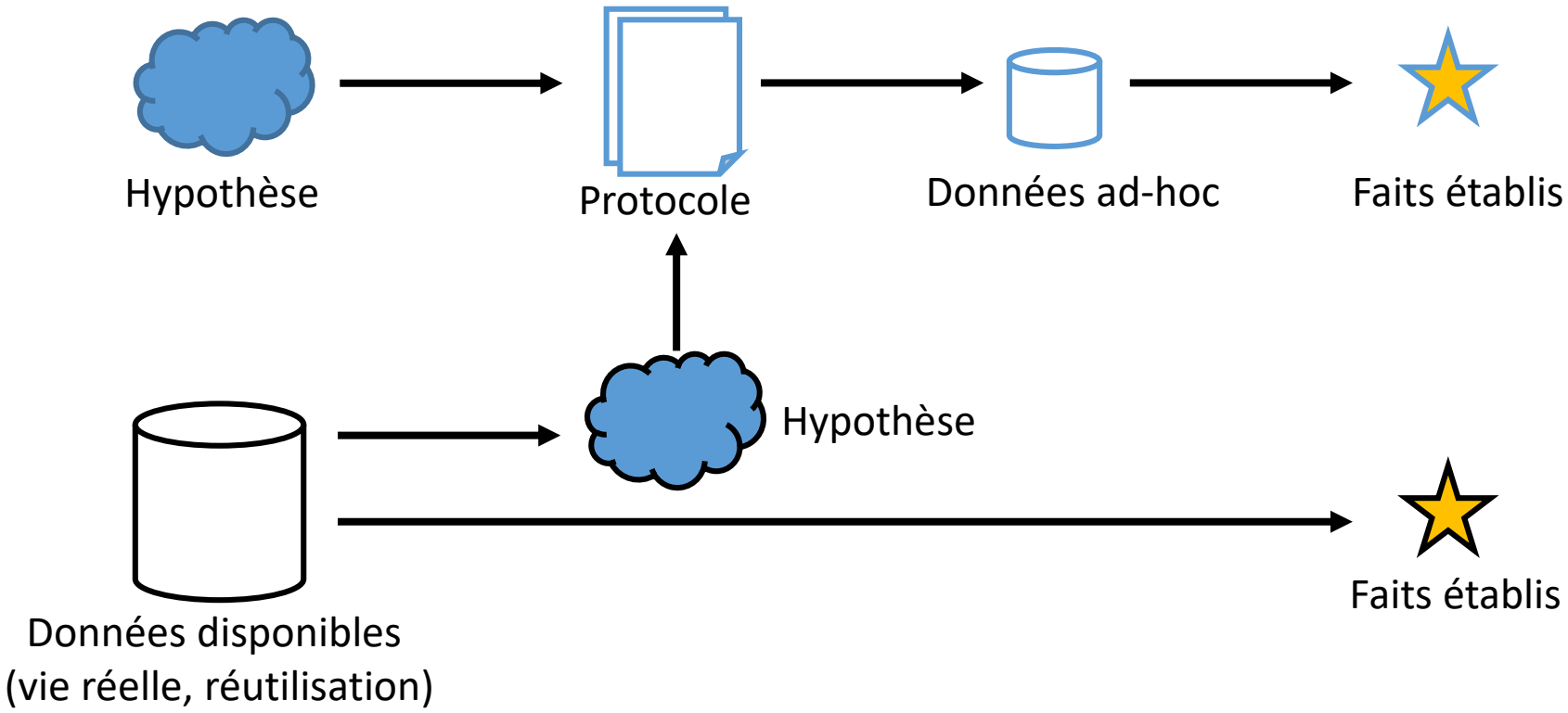


Schéma d'après M. Fieschi

Complémentarité avec les ECR

Données de « vie réelle »

Principales sources :

- Dossiers Patients Informatisés (DPI)
- Bases de données médico-administratives
- Registre et cohortes
- Web, réseaux sociaux
- Objets connectés

← complémentarité →

Données « expérimentales »

- Essais Cliniques Randomisé (ECR)

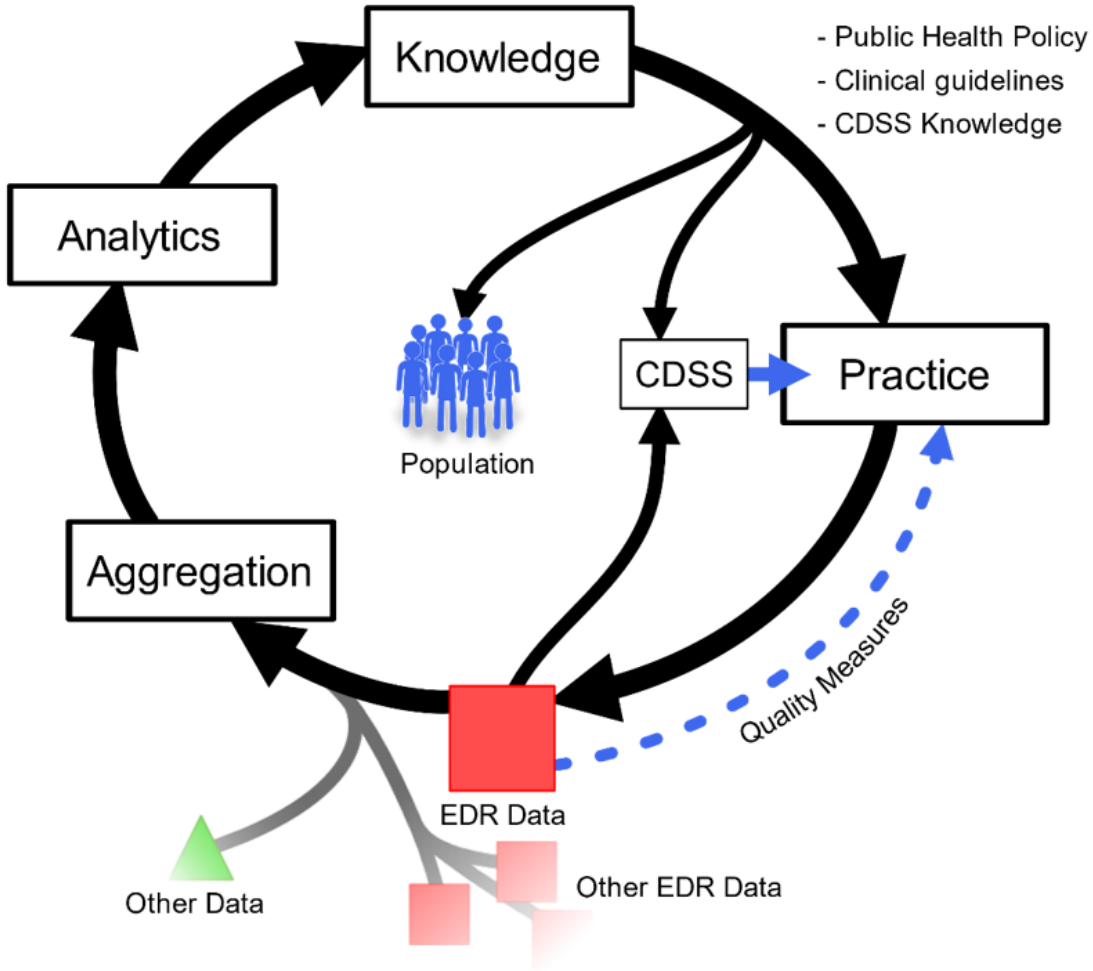
(petit effectifs, durée courte, patients plus jeune, moins de comorbidités,…)

- Etudes cas-témoins
- Série de cas
- …

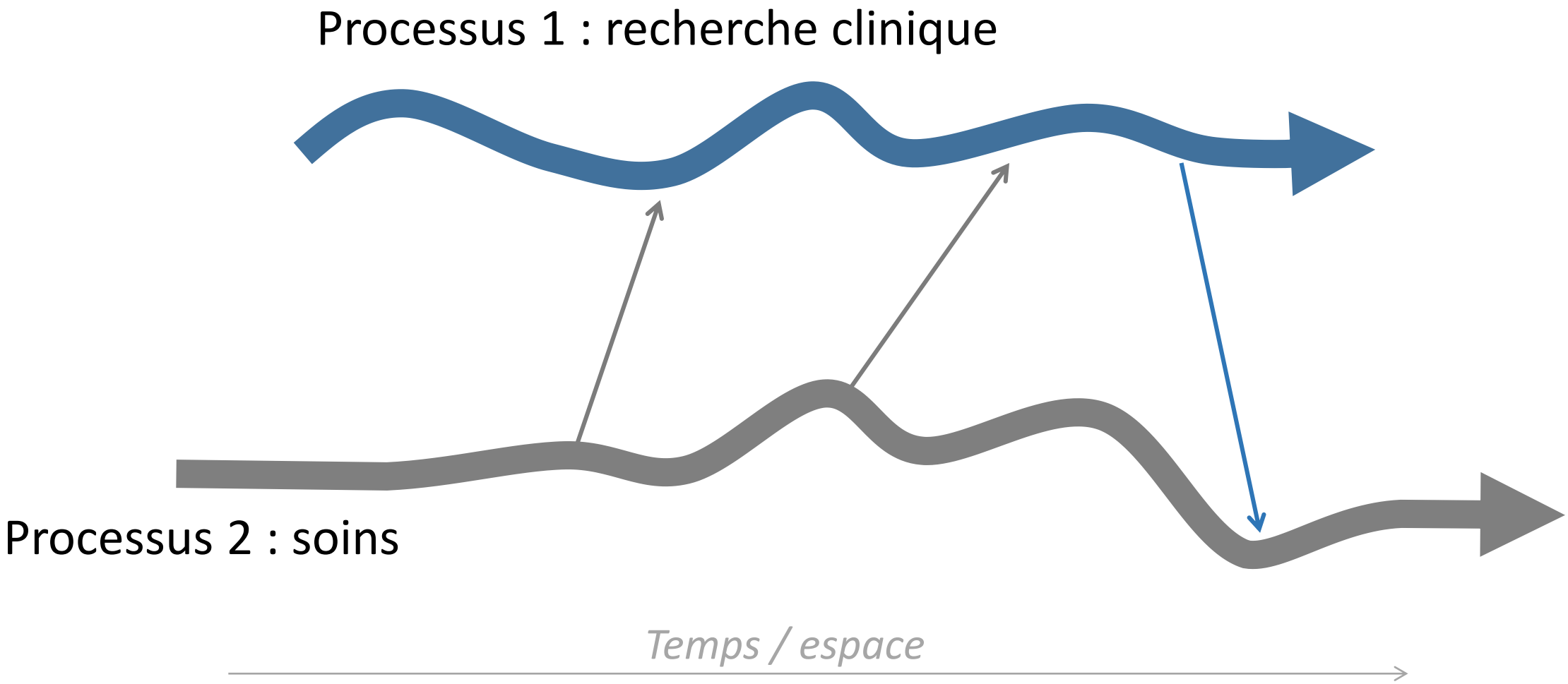
Paradigme des « Learning Healthcare Systems »

1) **médecine** fondée sur les **preuves** (EBM)
+
2) **preuves** fondées sur la **pratique**

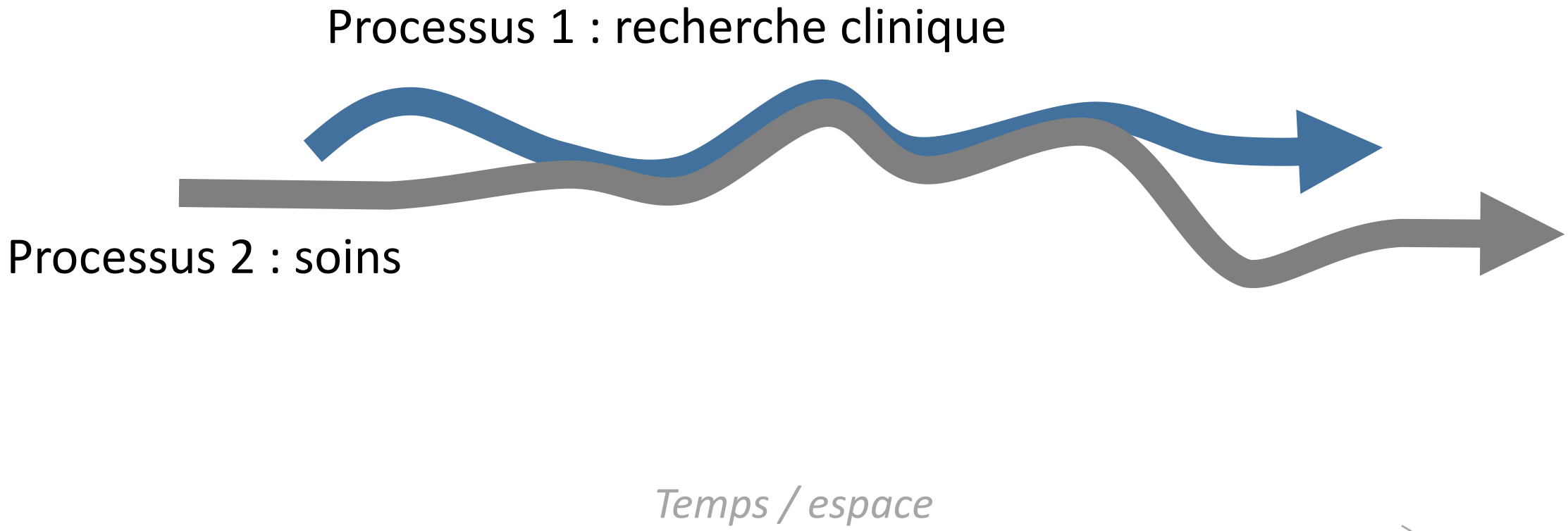
<u>Eléments clés</u> :
1) Organisation qui soutient la formation et la collaboration de patients, professionnels de la santé et chercheurs pour produire et utiliser des connaissances
2) Grands ensembles de données de santé (Big Data)
3) Chercher à améliorer la qualité directement sur les lieux de soin (point of care) grâce aux nouvelles connaissances générées par la recherche
4) Recherche effectuée dans contextes de soins courants



Réutilisation, Echange



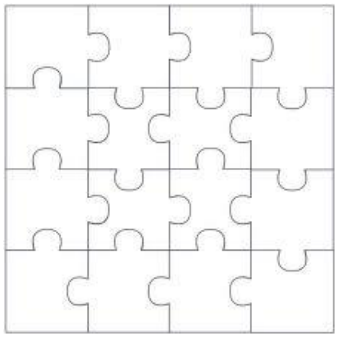
Partage => notion de mutualisation

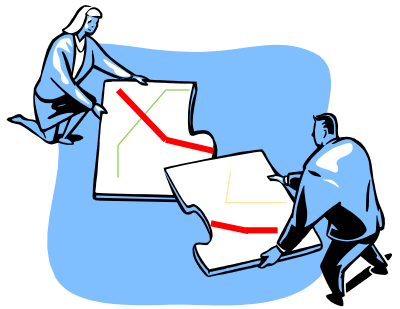


Notion d'interopérabilité

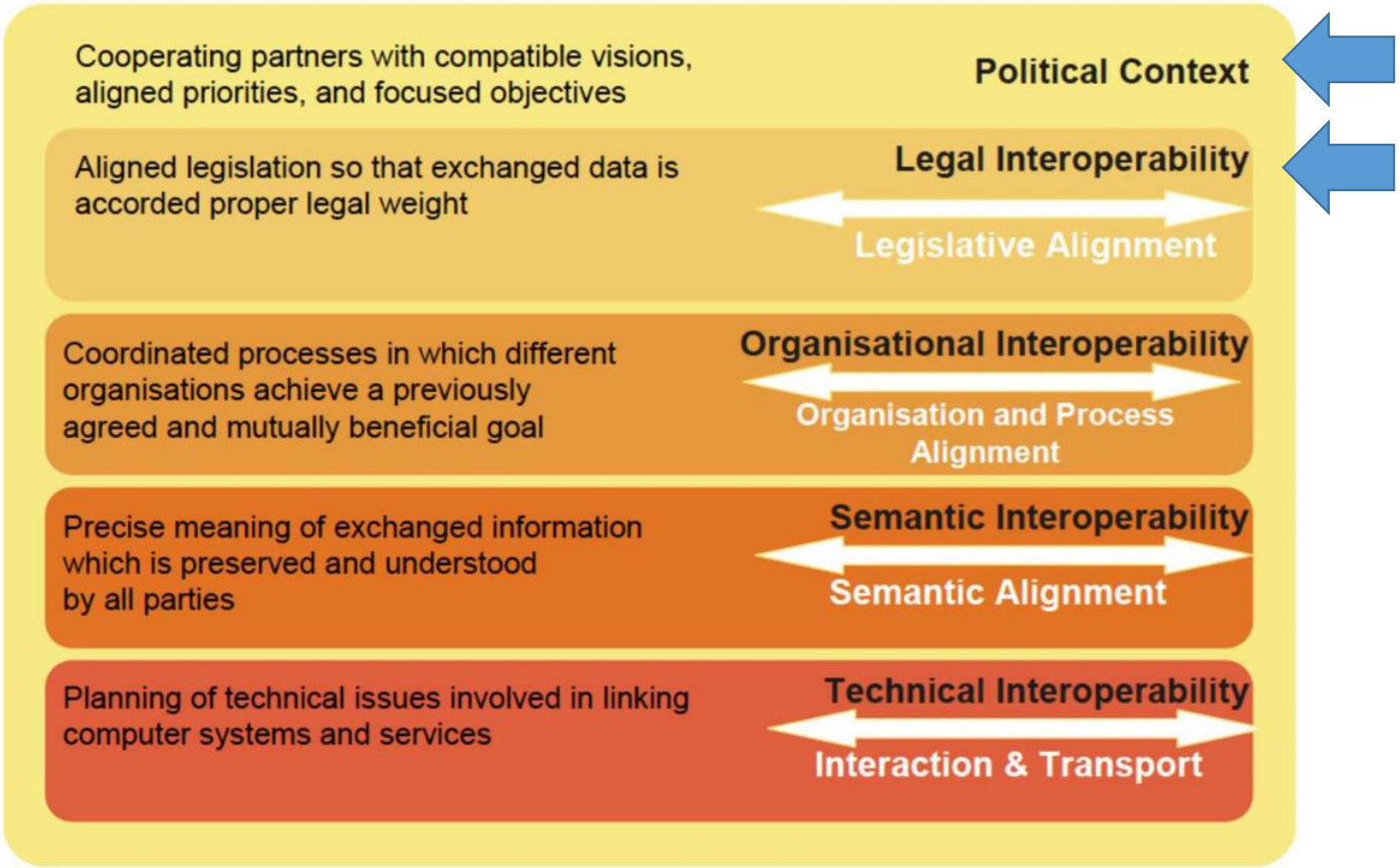
Interopérabilité des système d'information (SI) : différents niveaux

- Syntaxique → Technique
- Sémantique → Modèle et concepts (nomenclature/classification/référentiels de codages)
- Organisationnelle → Interopérabilité des processus





Niveaux d'interopérabilité plus généraux



D'après : Le référentiel general d'interopérabilité et les travaux de l'European Interoperability Framework (EIF)
Voir : http://references.modernisation.gouv.fr/interoperabilite et https://ec.europa.eu/isa2/eif
Vidéo illustrative : https://www.youtube.com/watch?v=g-CzHHJ0ZTM&

Grand domaines bénéficiant de la réutilisation des données de santé

- Innovation (production de connaissance pour développer de nouvelles modalités de prise en charge)
- Pilotage (établissements, régions, national)
- Régulation des soins
- Recherche

Clinical Care
- Direct Patient Care
- Quality and Safety
- Improving Efficiency
- Comparing Effectiveness
- Population Health

Governmental
- Biosurveillance
- Immunization Tracking
- Developing Public Policy

Business
- Fraud Detection
- Calculation of Insurance Premiums and Risk
- Marketing & Sales
- Drug Development
- Post-Marketing Surveillance

Research
- Design of Trials
- Recruitment
- Prediction
- Discovery

Safran C. Update on Data Reuse in Health Care. Yearb Med Inform. 2017 Aug;26(1):24-27.

Bégaud B, Polton D, von Lennep F. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé. Rapport 2017. https://solidarites-sante.gouv.fr/IMG/pdf/rapport_donnees_de_vie_reelle_medicaments_mai_2017vf.pdf

Quelques domaines plus spécifiques

- Recherche de patients similaires (aide à la décision)
- Pré-screening (études de faisabilité) / inclusion dans les essais cliniques
- Pharmacovigilance (suivi post AMM)
- Repositionnement / nouvelles indications de médicaments
- …

Quelques exemples (1) …déjà anciens

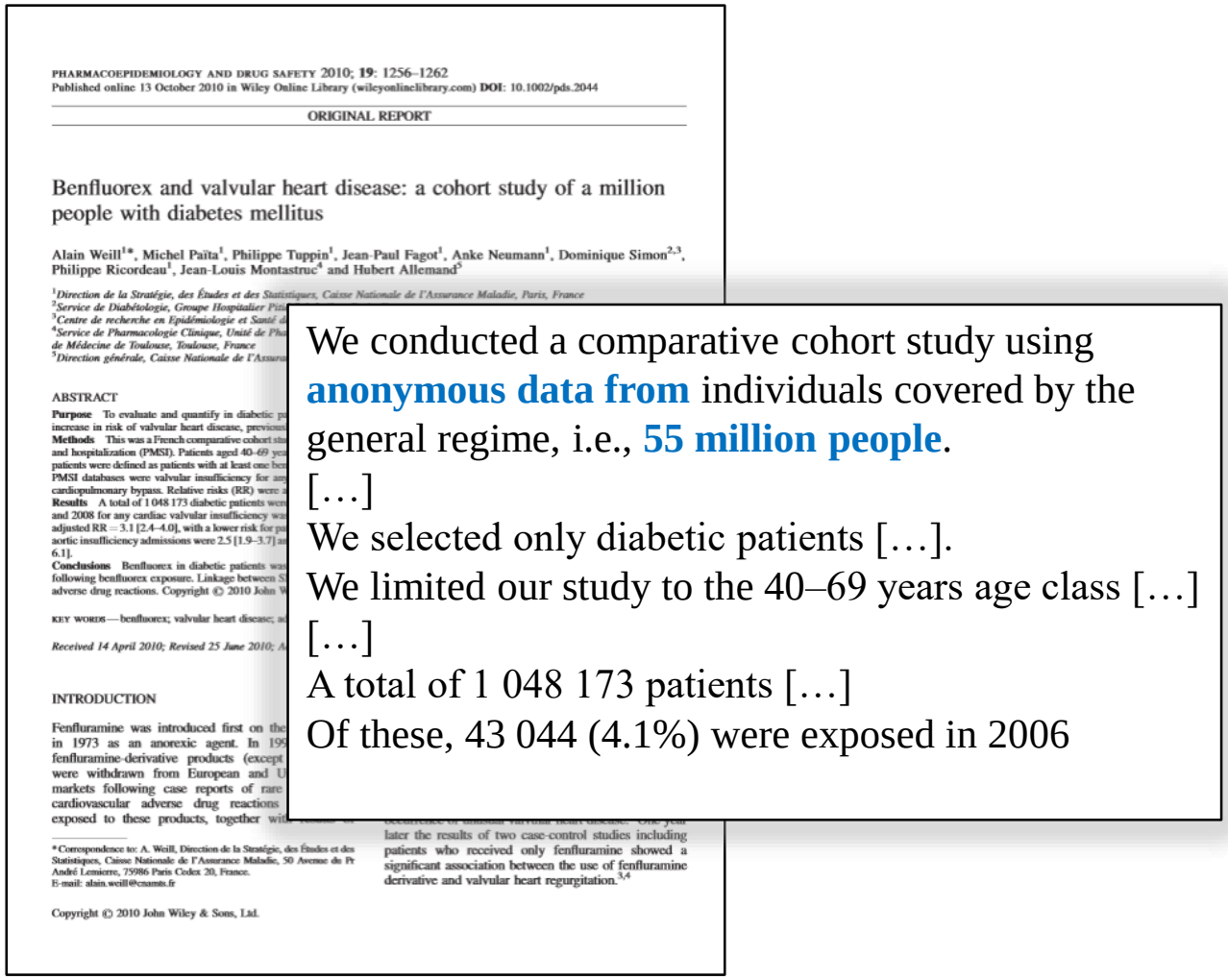
- Detecting influenza epidemics using search engine query data. Ginsberg J et al. Nature. 2009
- The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. Signorini A et al. PLoS One. 2011.
- The other Twitter revolution: how social media are helping to monitor the NHS reforms. McKee M et al. BMJ. 2011

Quelques exemples (2)

- Identification des patients à haut risque → baisse du taux de réadmission des patients cardiaques

Amarasingham R, Patel PC, Toto K, et al. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. BMJ Quality & Safety 2013;22:998-1005.

Quelques exemples (3)



PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2010; **19**: 1256–1262
Published online 13 October 2010 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.2044

ORIGINAL REPORT

Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus

Alain Weill[1]*, Michel Païta[1], Philippe Tuppin[1], Jean-Paul Fagot[1], Anke Neumann[1], Dominique Simon[2,3], Philippe Ricordeau[1], Jean-Louis Montastruc[4] and Hubert Allemand[5]

[1]*Direction de la Stratégie, des Études et des Statistiques, Caisse Nationale de l'Assurance Maladie, Paris, France*
[2]*Service de Diabétologie, Groupe Hospitalier Pitié*
[3]*Centre de recherche en Épidémiologie et Santé d*
[4]*Service de Pharmacologie Clinique, Unité de Pha*
de Médecine de Toulouse, Toulouse, France
[5]*Direction générale, Caisse Nationale de l'Assura*

ABSTRACT

Purpose To evaluate and quantify in diabetic pa
increase in risk of valvular heart disease, previous
Methods This was a French comparative cohort stu
and hospitalization (PMSI). Patients aged 40–69 yea
patients were defined as patients with at least one ben
PMSI databases were valvular insufficiency for any
cardiopulmonary bypass. Relative risks (RR) were a
Results A total of 1 048 173 diabetic patients were
and 2008 for any cardiac valvular insufficiency was
adjusted RR = 3.1 [2.4–4.0], with a lower risk for pa
aortic insufficiency admissions were 2.5 [1.9–3.7] an
6.1].
Conclusions Benfluorex in diabetic patients was
following benfluorex exposure. Linkage between S
adverse drug reactions. Copyright © 2010 John W

KEY WORDS — benfluorex; valvular heart disease; ad

Received 14 April 2010; Revised 25 June 2010; A

INTRODUCTION

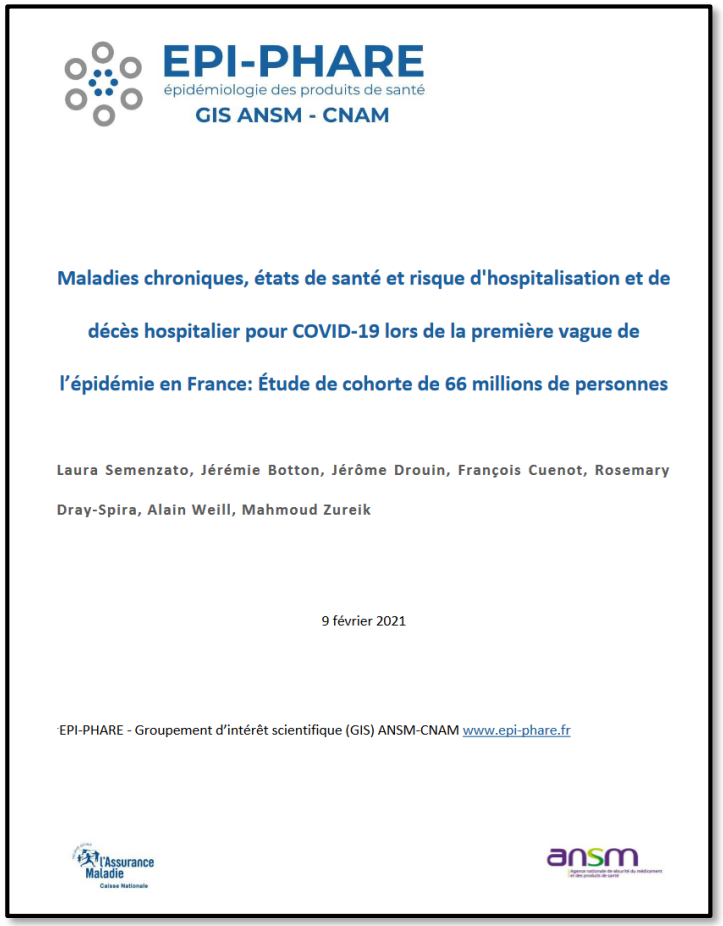
Fenfluramine was introduced first on the
in 1973 as an anorexic agent. In 199
fenfluramine-derivative products (except
were withdrawn from European and U
markets following case reports of rare
cardiovascular adverse drug reactions
exposed to these products, together with

later the results of two case-control studies including patients who received only fenfluramine showed a significant association between the use of fenfluramine derivative and valvular heart regurgitation.[3,4]

*Correspondence to: A. Weill, Direction de la Stratégie, des Études et des Statistiques, Caisse Nationale de l'Assurance Maladie, 50 Avenue du Pr André Lemierre, 75986 Paris Cedex 20, France.
E-mail: alain.weill@cnamts.fr

We conducted a comparative cohort study using **anonymous data from** individuals covered by the general regime, i.e., **55 million people**.
[…]
We selected only diabetic patients […].
We limited our study to the 40–69 years age class […]
[…]
A total of 1 048 173 patients […]
Of these, 43 044 (4.1%) were exposed in 2006

Quelques exemples (4)



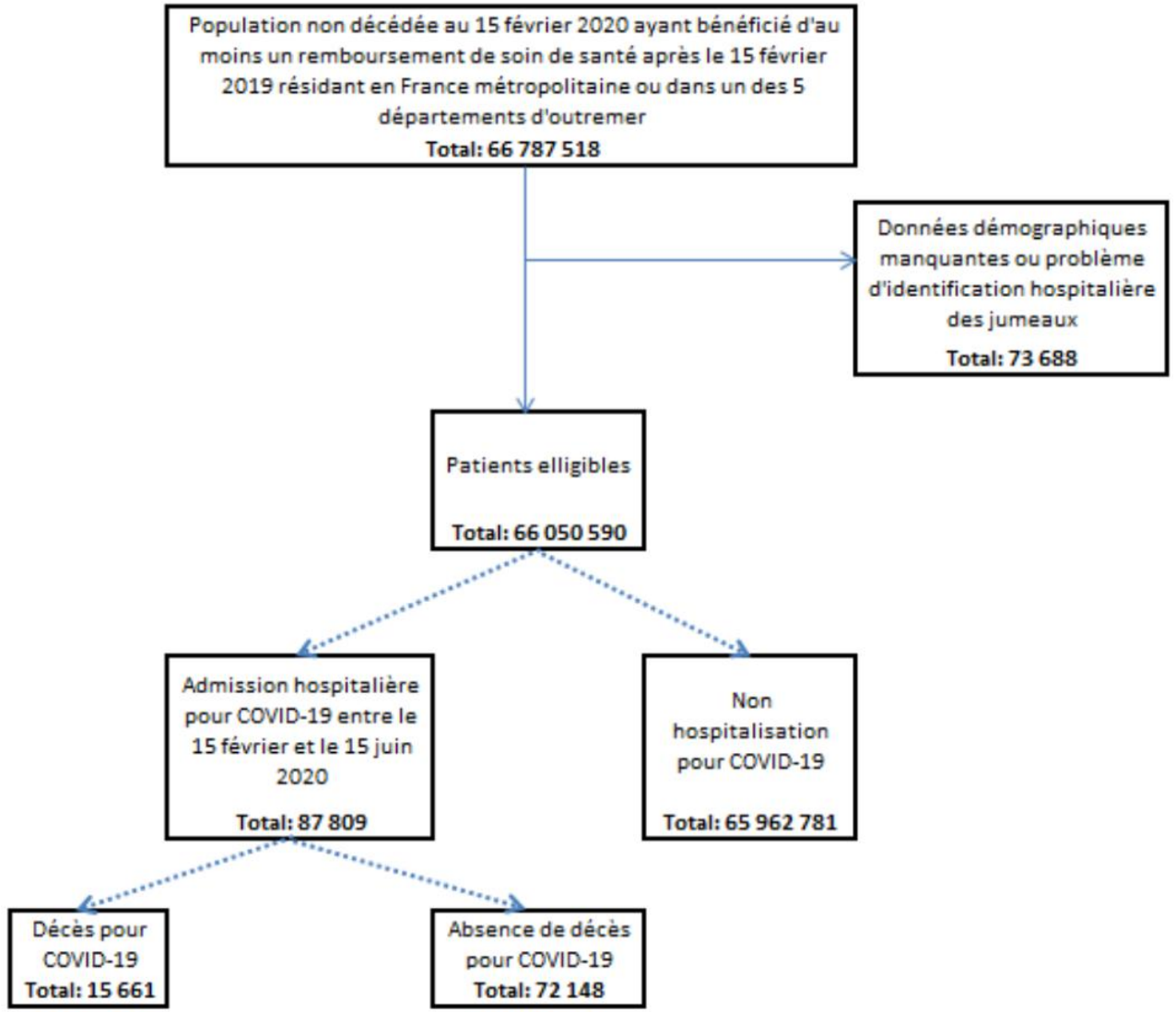
EPI-PHARE
épidémiologie des produits de santé
GIS ANSM - CNAM

Maladies chroniques, états de santé et risque d'hospitalisation et de décès hospitalier pour COVID-19 lors de la première vague de l'épidémie en France: Étude de cohorte de 66 millions de personnes

Laura Semenzato, Jérémie Botton, Jérôme Drouin, François Cuenot, Rosemary Dray-Spira, Alain Weill, Mahmoud Zureik

9 février 2021

EPI-PHARE - Groupement d'intérêt scientifique (GIS) ANSM-CNAM www.epi-phare.fr



SNIIR-AM (Système national d'information inter-régimes de l'Assurance maladie)

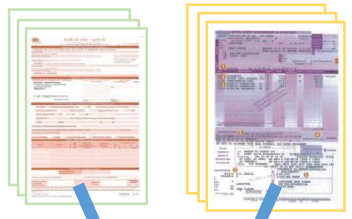
Contenu :

- Données socio-démographiques (sexe, âge, caisse d'affiliation, commune de résidence, affiliation CMUC)
- Numéro ALD
- Codes CIM 10 des diagnostics (PMSI)
- Informations sur les professionnels de santé et établissement qui ont prise en charge
- Codes CIP des médicaments remboursés
- Codes des actes de biologie et codes CCAM de actes techniques

**antériorité passe à 13 ans depuis 2012*

ALD = Allocation Longue Durée
CMUC = Couverture Maladie Universelle Complémentaire

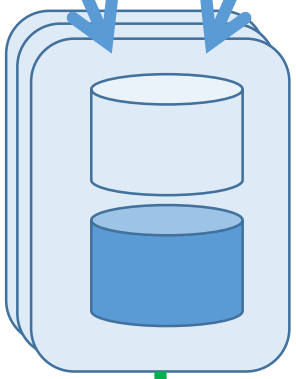
SNIIR-AM



Factures hôpitaux
Feuilles de soins
Factures actes biologie, pharmacies

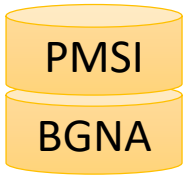
Niveau régional
CPAM

Contrôle assurées, prescripteurs,..
Données nominatives



Données de productions (consommation de soins, ALD, Accidents Travail, Maladies Pro)
Référentiels (annuaires des assurés, des établissements, des professionnels, …)

Anonymisation



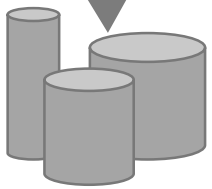
Données Hospitalières

Niveau national
CNAM

SNIIR-AM

Données individuelles anonymes

Extraction



« DataMarts »

SNIIR-AM

Quelques aspects techniques :

- Plus de 1 milliard de feuille de soins / an
- Plus de 25 teraoctets de données
- 785 000 objets (tables, index,…)
- Environ 20 milliards de lignes :
  - 9,0 milliards lignes de ventilation comptable
  - 8.4 milliards de lignes de nature de prestation
  - 3,4 milliards de lignes de pharmacie « affinée » (CIP)
  - 1,2 milliard de lignes de biologie « affinée » (codebiologie)
  - 270 millions ligne d'actes techniques médicaux (CCAM)
  - 57 millions ligne d'actes de transport médicaux

Données de santé – Aspects législatifs

Données à caractère personnel concernant la santé :

« **données** relatives à la santé physique ou mentale, **passée**, **présente** ou **future**, d'une **personne physique** qui **révèlent des informations sur l'état de santé** de cette personne. »

Règlement Général sur la Protection des Données (RGPD) Article 4, raison 35

Trois catégories de données de santé

Données de santé par nature

Croisement de données devenant des données de santé

Données de santé en raison de leur utilisation

Principes généraux de la mise en œuvre de traitement de données à caractère personnel – Aspects législatifs et éthique

PRINCIPE 1
LA FINALITÉ

DÉFINIR LES OBJECTIFS DU FICHIER
Avant toute collecte et utilisation de données personnelles, le responsable de traitement doit précisément annoncer aux personnes concernées ce à quoi elles vont lui servir. Ces objectifs, appelés "*finalités*", doivent respecter les droits et libertés des individus. Ils limitent la manière dont le responsable pourra utiliser ou réutiliser ces données dans le futur
> En savoir plus

PRINCIPE 2
LA PERTINENCE

VÉRIFIER LA PERTINENCE DES DONNÉES
Seules les données strictement nécessaires à la réalisation de l'objectif peuvent être collectées : c'est le principe de minimisation de la collecte. Le responsable de traitement ne doit donc pas collecter plus de données que ce dont il a vraiment besoin. Il doit également faire attention au caractère sensible de certaines données.
> En savoir plus

PRINCIPE 3
LA CONSERVATION

LIMITER LA CONSERVATION DES DONNÉES
Une fois que l'objectif poursuivi par la collecte des données est atteint, il n'y a plus lieu de les conserver et elles doivent être supprimées. Cette durée de conservation doit être définie au préalable par responsable du traitement, en tenant compte des éventuelles obligations à conserver certaines données.
> En savoir plus

PRINCIPE 4
LES DROITS

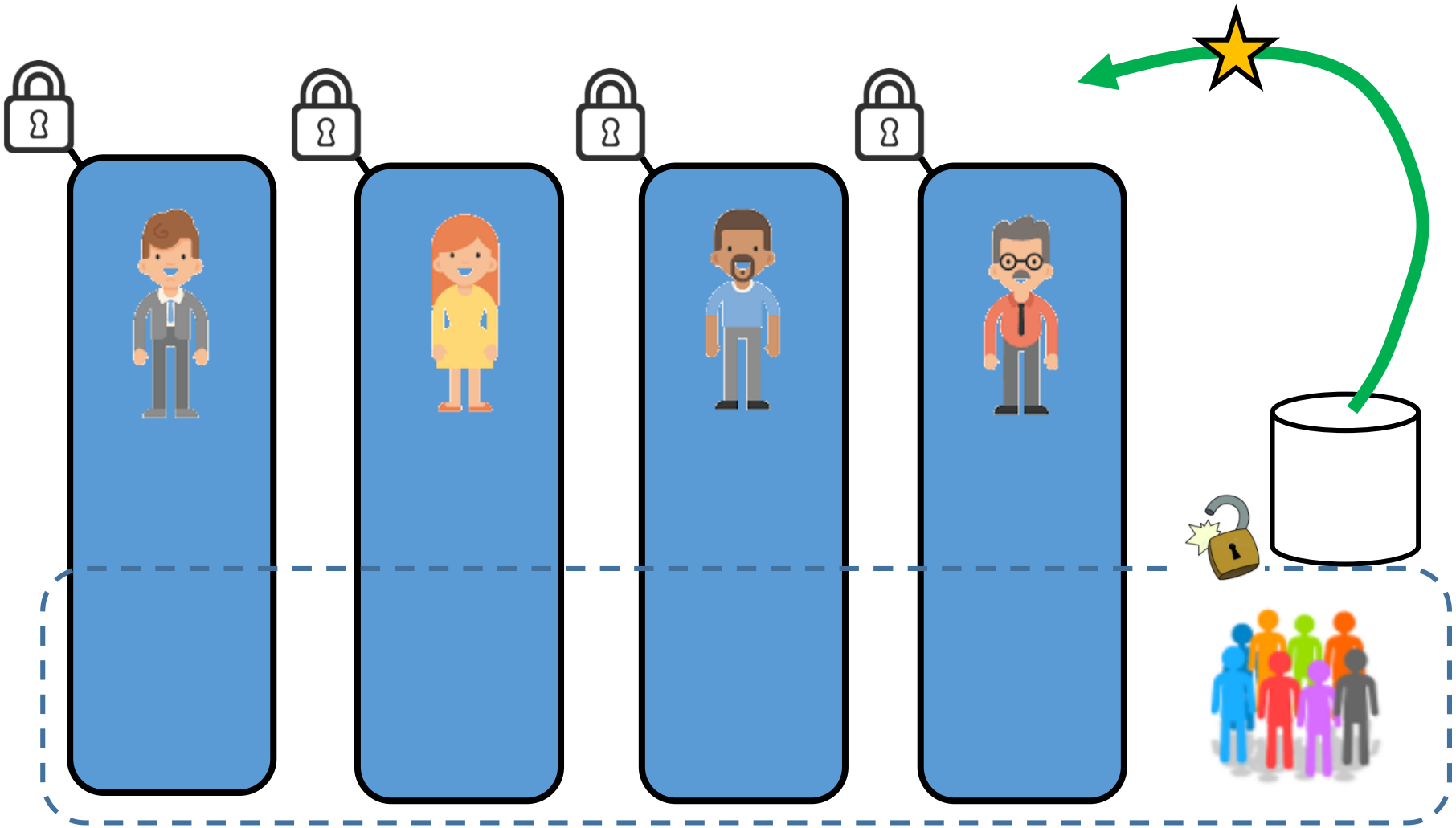
RESPECTER LES DROITS DES PERSONNES
Des données concernant des personnes peuvent être collectées à la condition essentielle qu'elles aient été informées de cette opération. Ces personnes disposent également de certains droits qu'elles peuvent exercer auprès de l'organisme qui détient ces données le concernant : un droit d'accéder à ces données, un droit de les rectifier et enfin un droit de s'opposer à leur utilisation.
> En savoir plus

PRINCIPE 5
LA SÉCURITÉ

SÉCURISER LES DONNÉES
Le responsable de traitement doit prendre toutes les mesures nécessaires pour garantir la sécurité des données qu'il a collectées mais aussi leur confidentialité, c'est-à-dire s'assurer que seules les personnes autorisées y accèdent. Ces mesures pourront être déterminées en fonction des risques pesant sur ce fichier (sensibilité des données, objectif du traitement...)
> En savoir plus

https://www.cnil.fr/fr/comprendre-vos-obligations/les-principes-cles

Deux logiques à concilier (1)



Deux logiques à concilier (2)

- Continuité des soins :
  - Identifier la personne physique
  - Confidentialité +++
  - Partage/mutualisation
- Recherche :
  - Identifiant d'un individu (sans désigner une personne physique)
  - Confidentialité +/-
  - Réutilisation

Freins à la réutilisation

1. L'information n'est pas présente (temps/lieu)
2. L'information est présente mais son utilisation n'est pas autorisée
3. L'information est présente mais ne peut pas être utilisé telle quelle :
   - pour des raisons **organisationnelles et/ou techniques**
   - pour des raisons **politiques**
   - pour des raisons de **qualité intrinsèque** ou **extrinsèque** (fiabilités, complétude, pertinence)

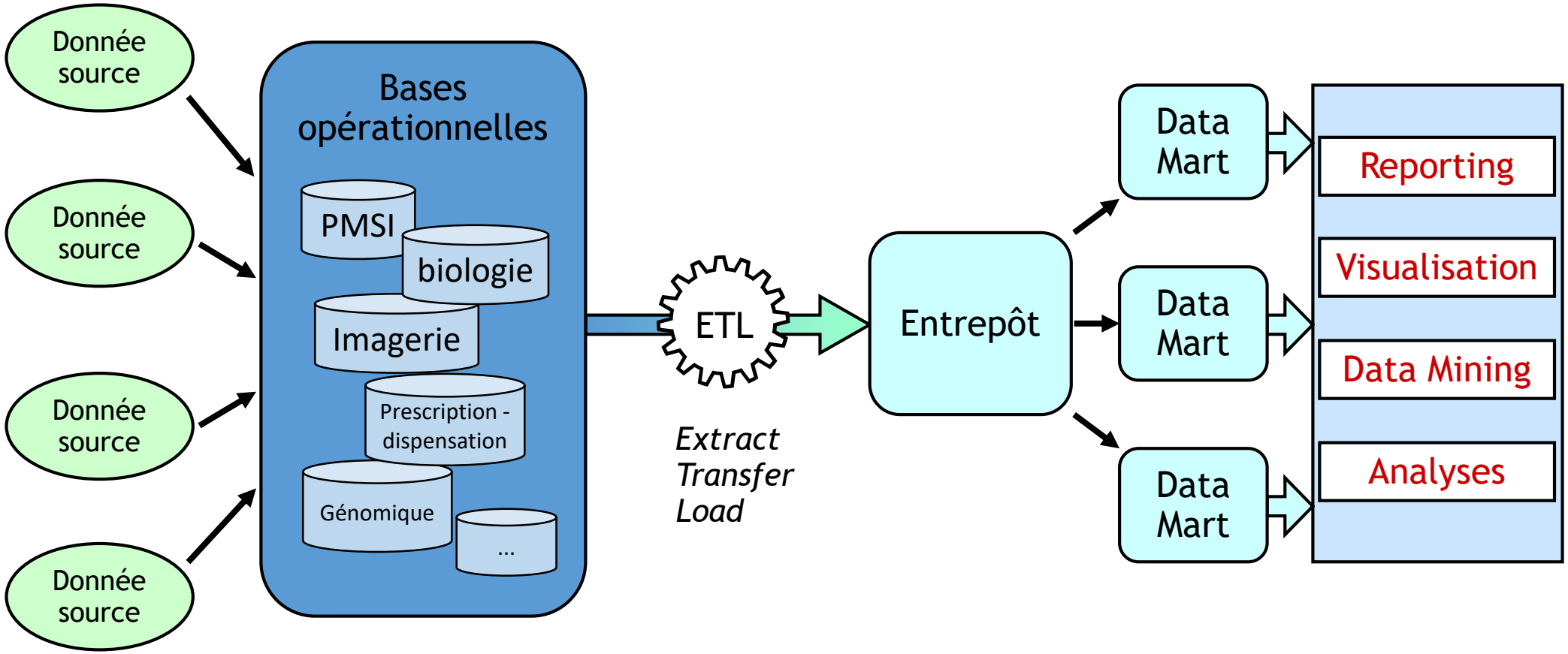
adapté de Galster G. 2012

Entrepôt de données

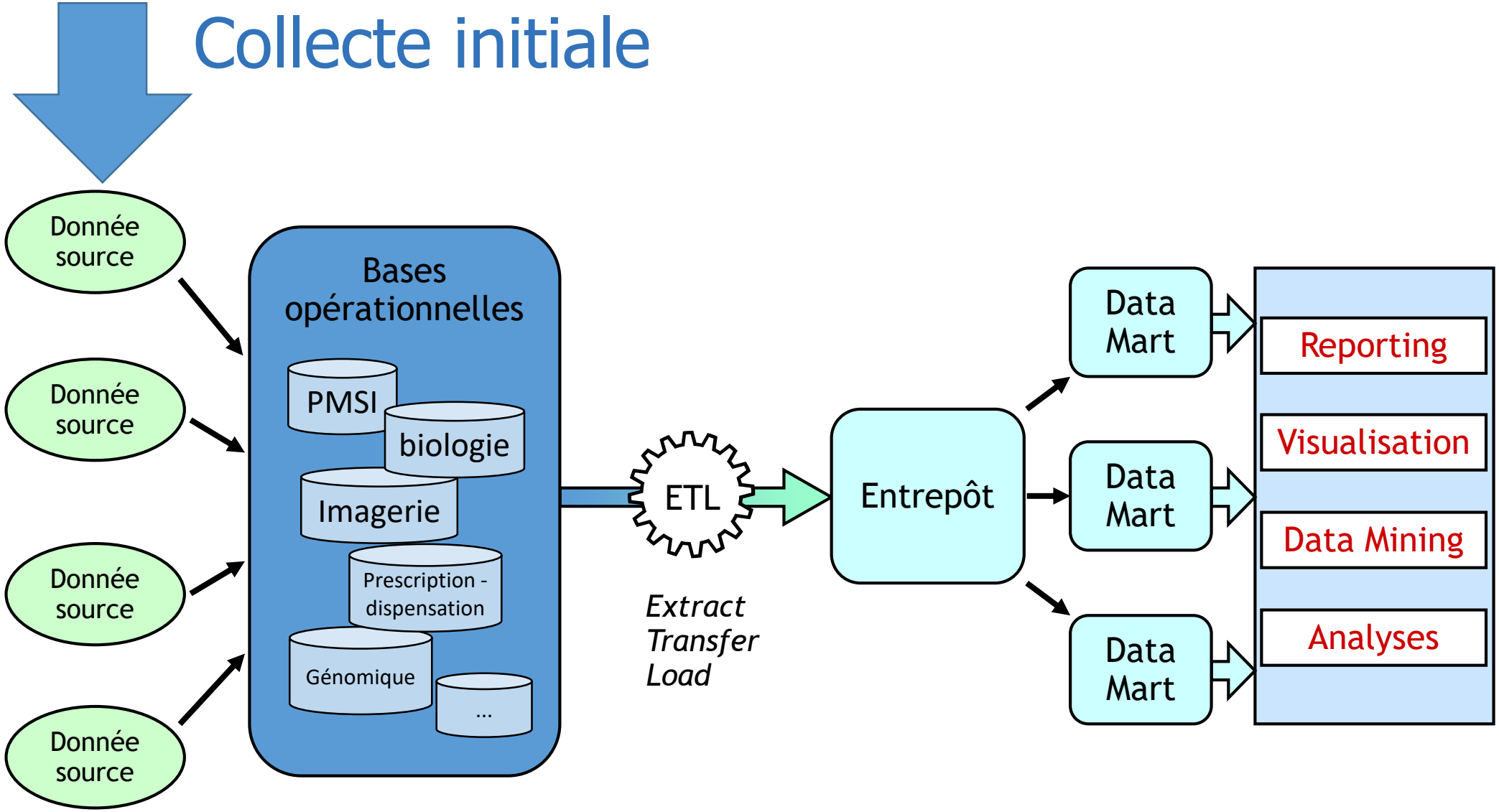
« Un entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision. » - William Inmon, 1994.

Entrepôt de données ≠ simple déversoir

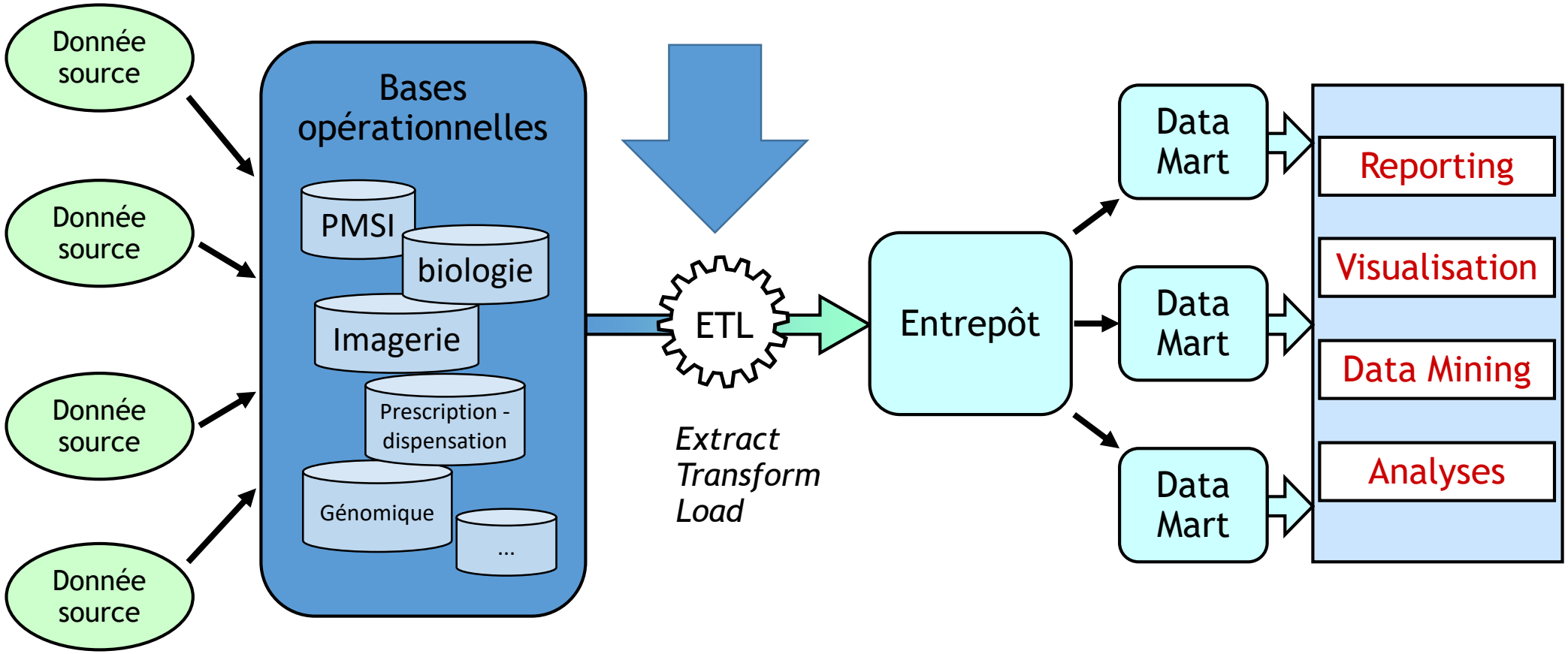
Principe de constitution d'un entrepôt de données hospitalier



Collecte initiale



Sélection, data cleaning puis alimentation de l'entrepôt



Entrepôts de données hospitaliers

Exemple : cartographie des données de l'entrepôt de données de l'AP-HP (date : février 2021)

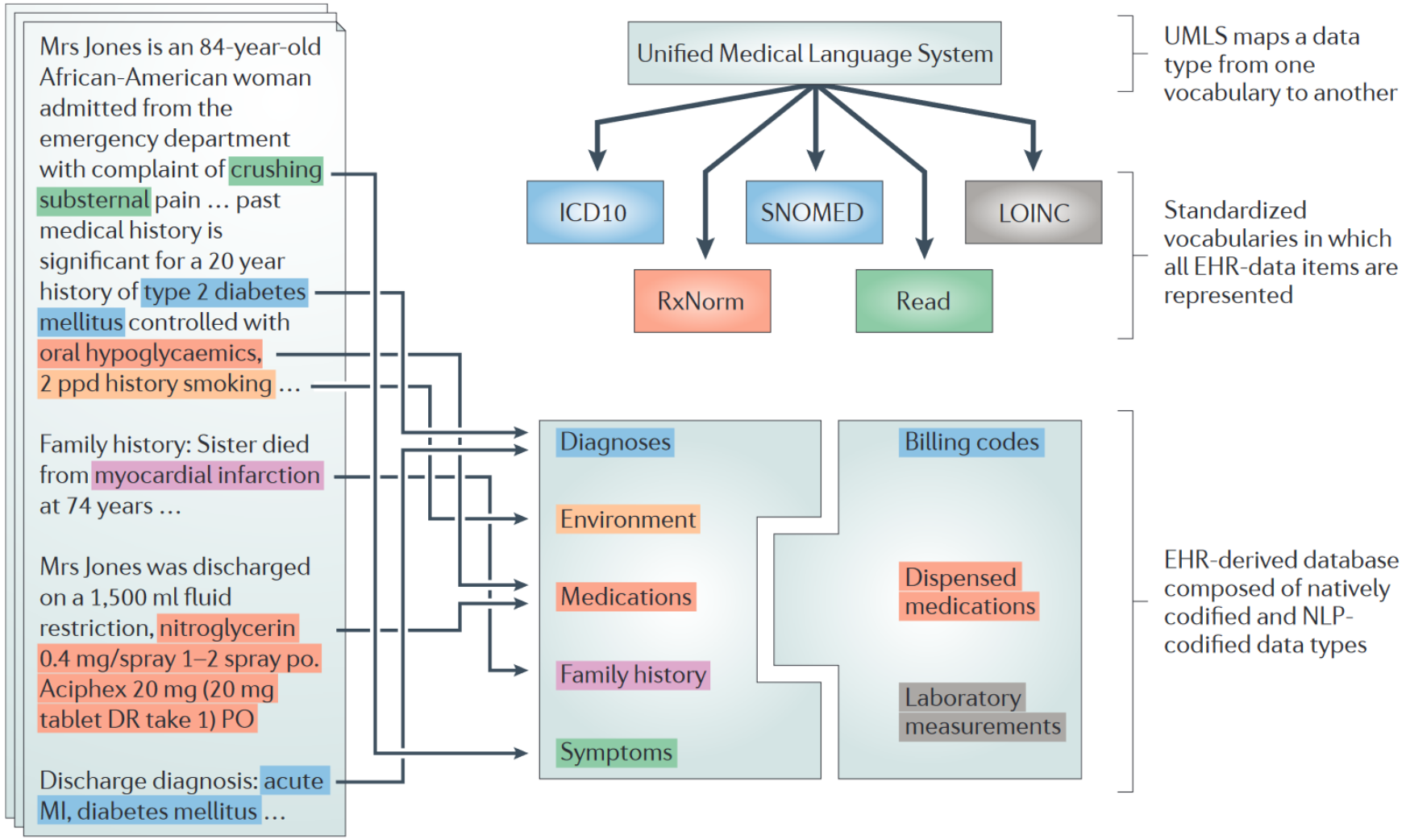
Transformation des données et extraction de caractéristiques

- Gestion des doublons
- Reformatage
  - Format physique de données
  - Transformation de type, de nom, de code
- Consolidation
- Uniformatisation d'échelle
- Discrétisation de variables
- Utilisation de méthodes avancées : NLP (Natural Langage Processing)
- …

A visionner !

Réutilisation de données hospitalières et intelligence artificielle : des données à l'intervention de santé, un chemin cahoteux. Emmanuel Chazard. https://sesstim.univ-amu.fr/video-box/webinar-quantim-emmanuel-chazard

Extraction de concepts par NLP



Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011 Jun;12(6):417-28.

Figure 1 | **From clinical notes to structured phenotypes.** Natural language processing (NLP) identifies various concept types in the textual records that are associated with each patient for each medical record.

Problèmes de la qualité des données

D'après Kahn MG, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244.

1. **Conformité** (le format de la donnée est-il bien conforme à ce qui est attendu ?)
   - Format de la valeur elle-même
   - Conformité du code ou libellé identifiant la donnée
2. **Complétude** (la valeur de la donnée est-elle bien présente ?)
   - Omission fortuite (**MCAR** *Missing Completely at Random*)
   - Omission dépendante du contexte (**MAR** *Missing At Random,* **MNAR** *Missing Not At Random*)
3. **Plausibilité** (la valeur de la donnée est-elle crédible ?)
   - Intemporelle
   - Temporelle

A visionner !

Qualité temporelle dans les entrepôts de données cliniques. Bastien Rance. https://sesstim.univ-amu.fr/video-box/webinar-quantim-bastien-rance

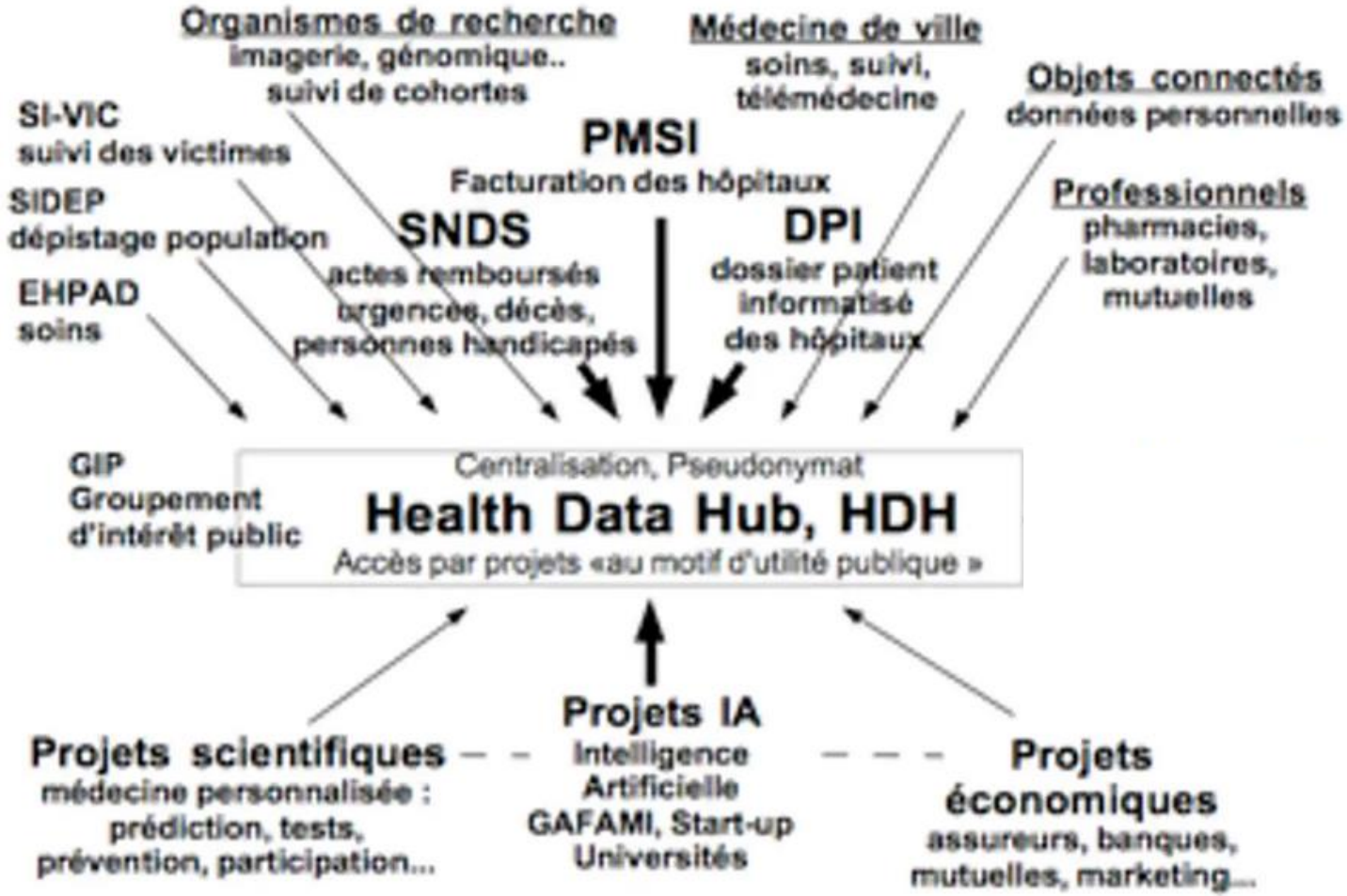
Encore plus de données !

Manque de données massives largement accessibles pour développer l'IA en santé (cf. Rapport C. Villani, mars 2018)

Health Data Hub

HEALTH DATA HUB

Plateforme des données de santé française (Health Data Hub élargie, HDH)



Emprunté à https://theconversation.com/donnees-de-sante-larbre-stopcovid-qui-cache-la-foret-health-data-hub-138852

Réseaux de données pour la recherche (Data Networks)

- Organisation et gouvernance
- Communication avec la clinique et ses données (et avec les patients !)
- Services communs (outils d'analyse, anonymisation de données, etc.)
- Plateforme facilitant l'accès et de traitement des données

Quelques exemples : PCORnet (Patient-Centered Clinical Research Network)

OHDSI, OMOP (Obervational Medical Outcomes Partnership)

TEHDAS (Towards the European Health DAta Space)

A visionner !

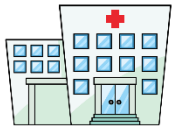
Vers une approche systémique du traitement de l'information en recherche en santé. Marius Fieschi. https://sesstim.univ-amu.fr/video-box/webinar-quantim-marius-fieschi

Big data en santé : Réutilisation des données massives en santé, définition, exemples de cas d'usage, infrastructures et méthodes. Marc Cuggia. https://sesstim.univ-amu.fr/video-box/webinar-quantim-marc-cuggia

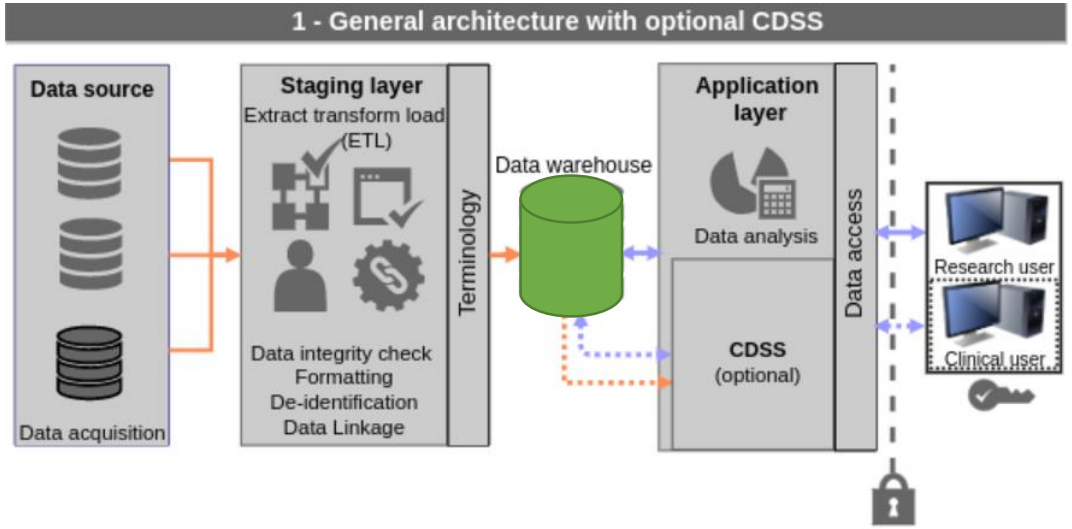
Fédération d'entrepôts et plateforme de données pour la recherche

- Utilisation de standards (nomenclatures, modèles de données, procédures)

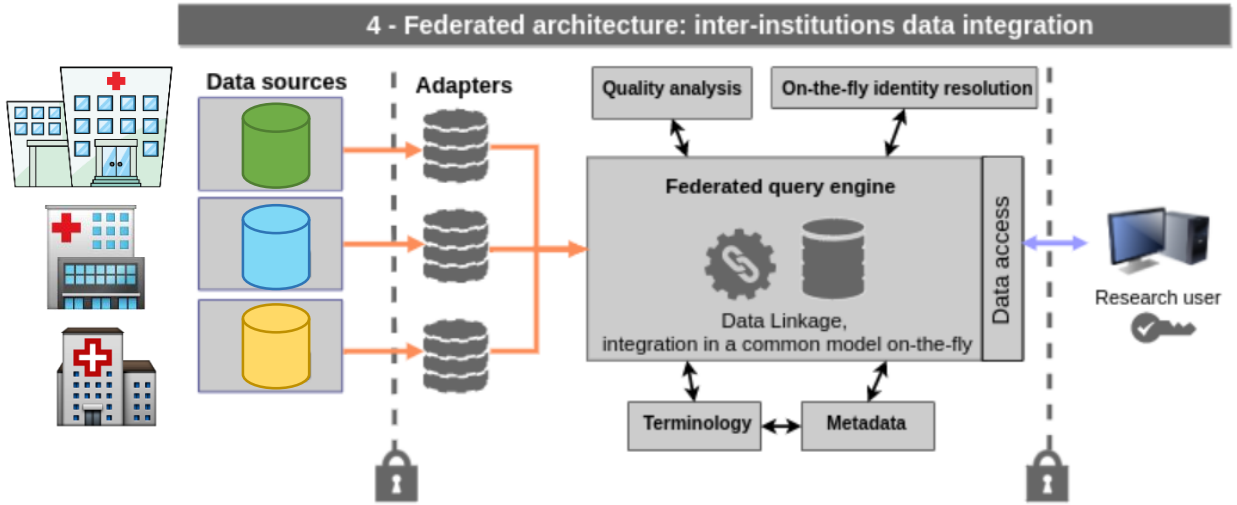
→ Interopérabilité → partage, réutilisation



Entrepôt « mono institution »



Plateforme « inter-institution »



D'après : Gagalova K et al. What You Need to Know Before Implementing a Clinical Research Data Warehouse: Comparative Review of Integrated Data Repositories in Health Care Institutions JMIR Form Res 2020;4(8):e17687

Evolution concernant les données

Les données directement numériques
et au plus proche du patient (mHealth)

Hier : Problème de la collecte de données

Aujourd'hui : Qualité et usage des données
et des algorithmes résultants

Demain : Multiplication des capteurs et IoT, données temps réel,
intégration de données hétérogènes (plateformes de données)

Merci pour votre attention !

Dr Jean-Charles DUFOUR
jean-charles.dufour@univ-amu.fr