Données massives en santé :

Changements techniques et culturels

Présentation Cnis - 2019.11.28
Matthieu Doutreligne - Drees

Contexte, les données

Des données de plus en plus massives à traiter:

- Enquête Santé Européenne, une année (EHIS) : 10141 personnes, 12 GB
- Système National de Données de Santé (SNDS), une année : 66.6 millions de personnes, 3.1 TB, ~ 5 milliards de lignes
- SNDS, 10 ans : 31 TB, ~ 50 milliards de lignes

Mais des volumes faibles comparés à l'industrie du numérique

- En santé, **IQVIA** possède une base de 500 millions de personnes
- Google, **Youtube analytics** : 100 milliards de lignes supplémentaires **par jour** à analyser en temps quasi réel (Chattopadhyay et al. 2019)

-> C'est réalisable, les solutions techniques existent et de nombreuses sont gratuites en open-source

Contexte, le SNDS

- Système National de Données de Santé (SNDS) géré par la Caisse Nationale d'Assurance Maladie (CNAM) contient actuellement :
  - les données de l'Assurance Maladie (consommations de soins en ville et en établissements remontées dans le SNIIRAM),
  - les données de facturation hospitalière du PMSI issues de l'ATIH,
  - les données sur les causes médicales de décès du CépiDC-Inserm.

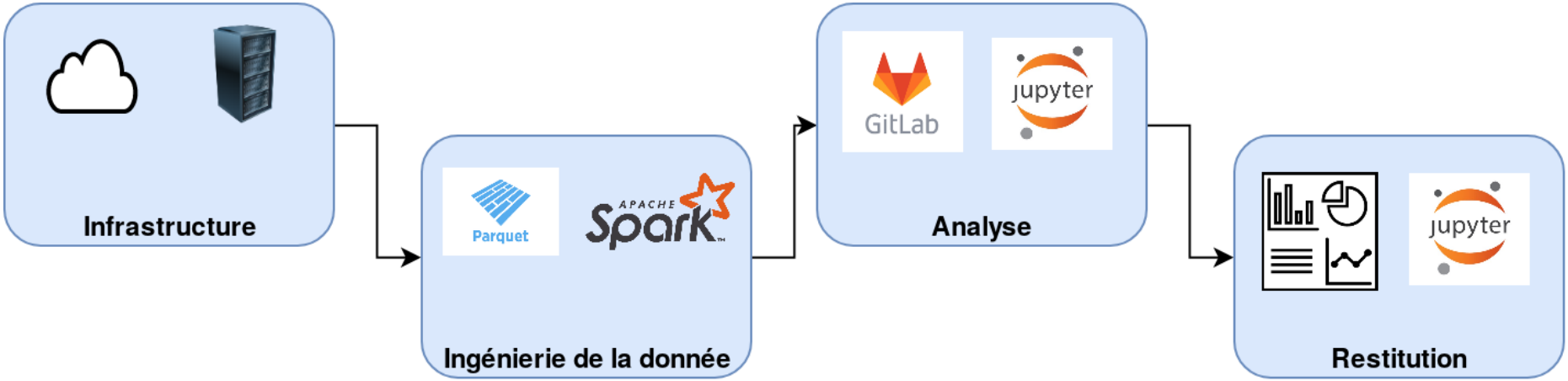
- La principale limitation de ces données est de ne rien contenir sur les résultats des différents examens.

- Appariement avec les données socio-démographiques de l'Échantillon Démographique Permanent de l'Insee (EDP).

Objectif et opportunités pour le SNDS

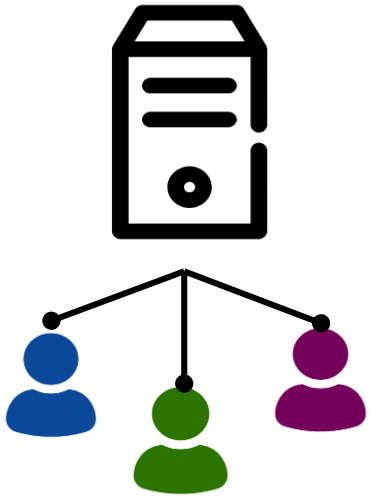
- **Documentation :** Lieu unique, ouvert, collaboratif, rassemblant les diverses connaissances existantes
- **Temporalité :** Analyser des données complexes comme le SNDS en quasi-temps réel (requête complexe de l'ordre la minute)
- **Accessibilité de la donnée :** Proposer des formats simplifiés voire standards avec des données de qualité
- Mettre en oeuvre des **méthodes d'analyse innovantes**

Des changements techniques et culturels



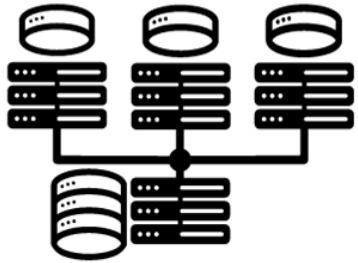
Différentes étapes de la donnée à la Drees

Infrastructures et ingénierie de la donnée

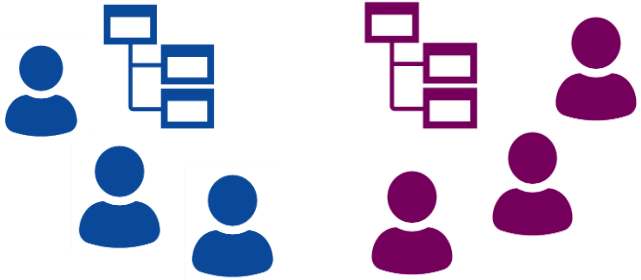


Infrastructure centralisée et formats disparates

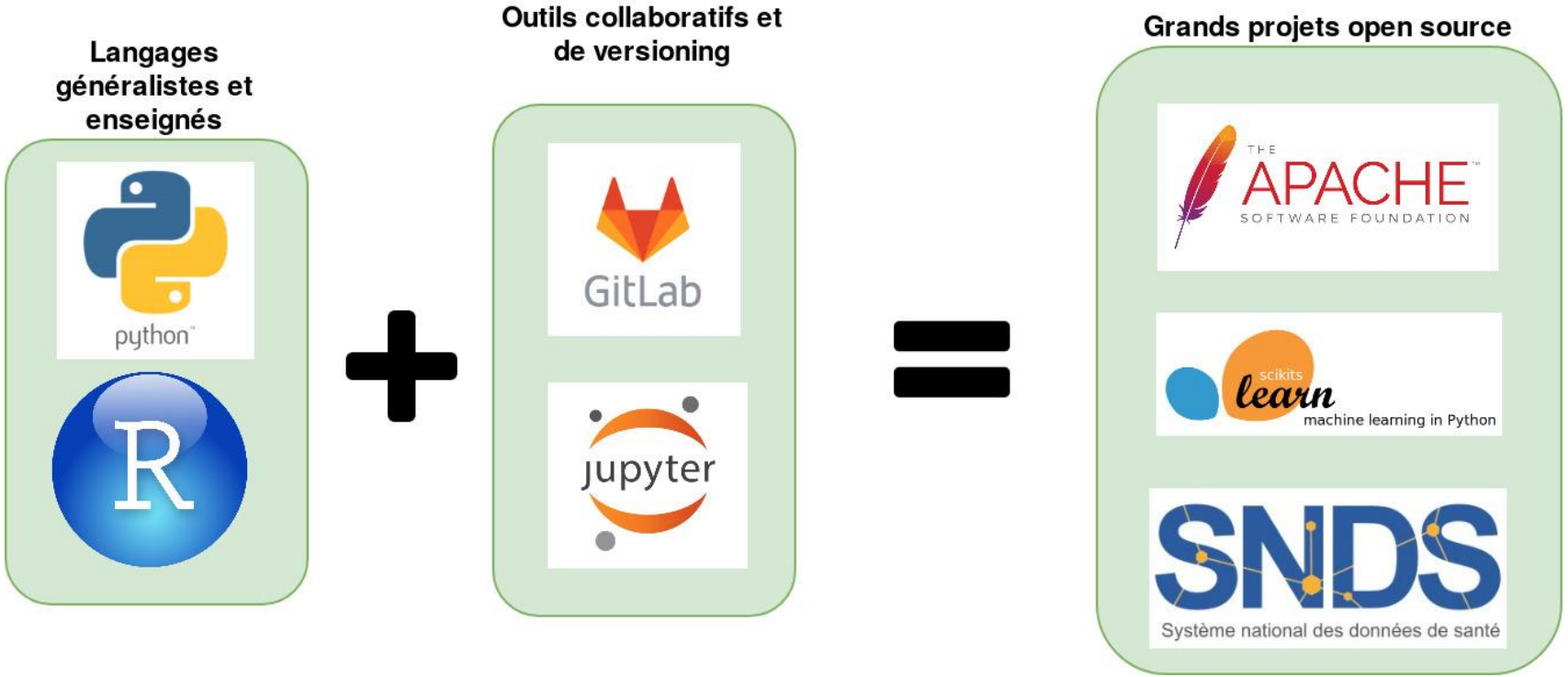
Infrastructure distribuée



Ingénierie de la donnée



Mutualisation des ressources pour l'analyse

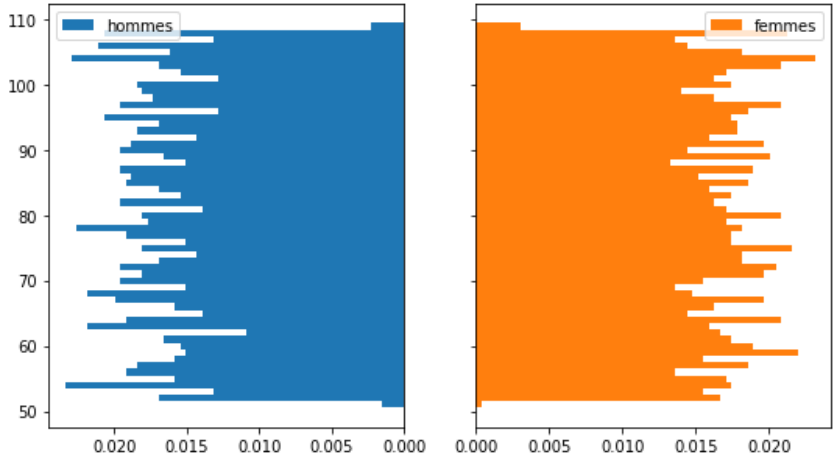


Restitution

- **Le notebook :**
  A la croisée de l'analyse et de la restitution

On peut regarder rapidement quelques variables démographiques de la poulation sélectionnée:

```
[9]: aged_buckets = age_pyramide_plot(population_patients, bucket_size=1)
     demographics = describe_demographics(population_patients)
```

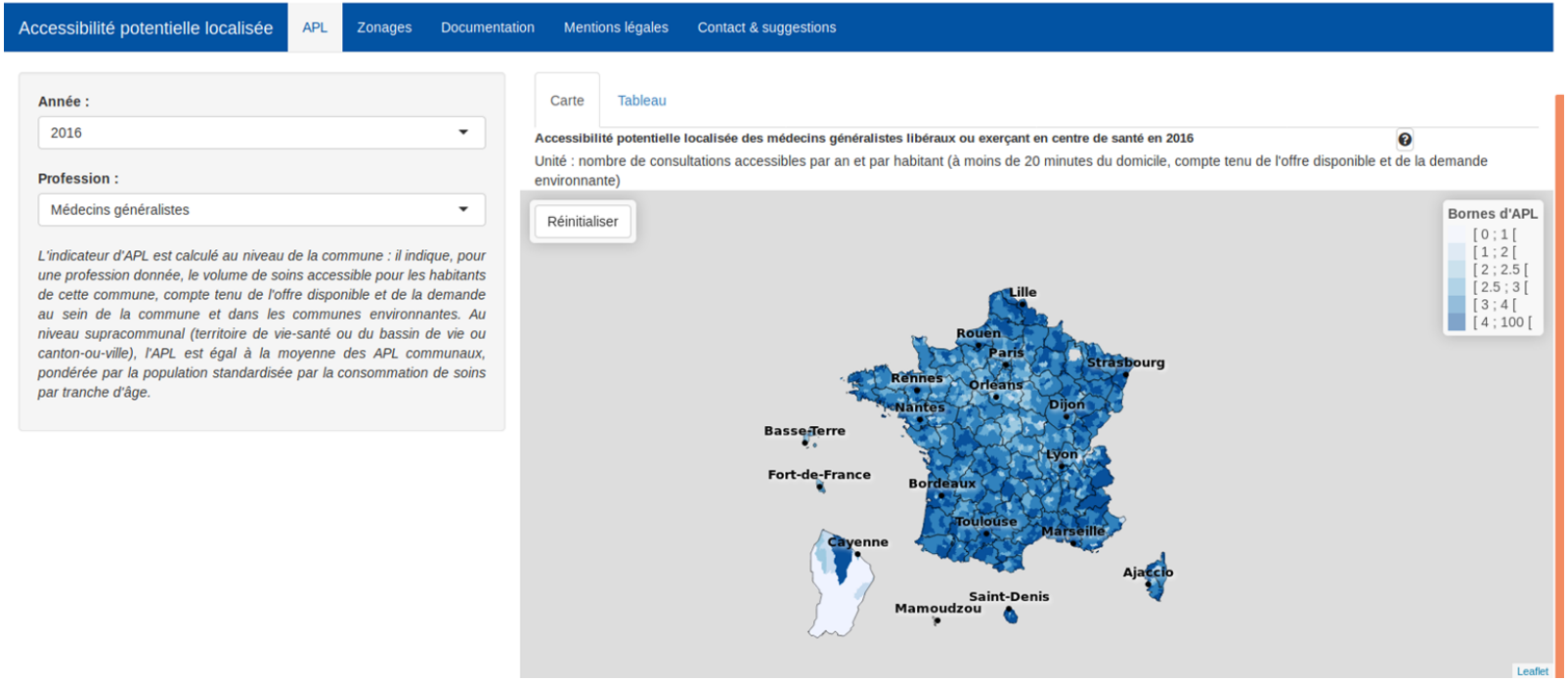



```
Taille population : 5299
Part femmes : 2658/5299 = 0.5016
Part DCD : 2768/5299 = 0.5224
```

Concaténation des différents types d'événements et écriture de la population de base

On joint et on écrit la `population_cohort` qui définit le périmètre général de notre étude. Cette étape en joignant les tables et en écrivant les événments est coûteuse en calcul et prendra quelques minutes (voi dizaines de minutes si on est gourmand sur le champ de l'étude).

Restitution



L'accessibilité potentielle localisée (APL), Bureau des Professions de Santé (B. Legendre) : http://dataviz.drees.solidarites-sante.gouv.fr/carto-apl/

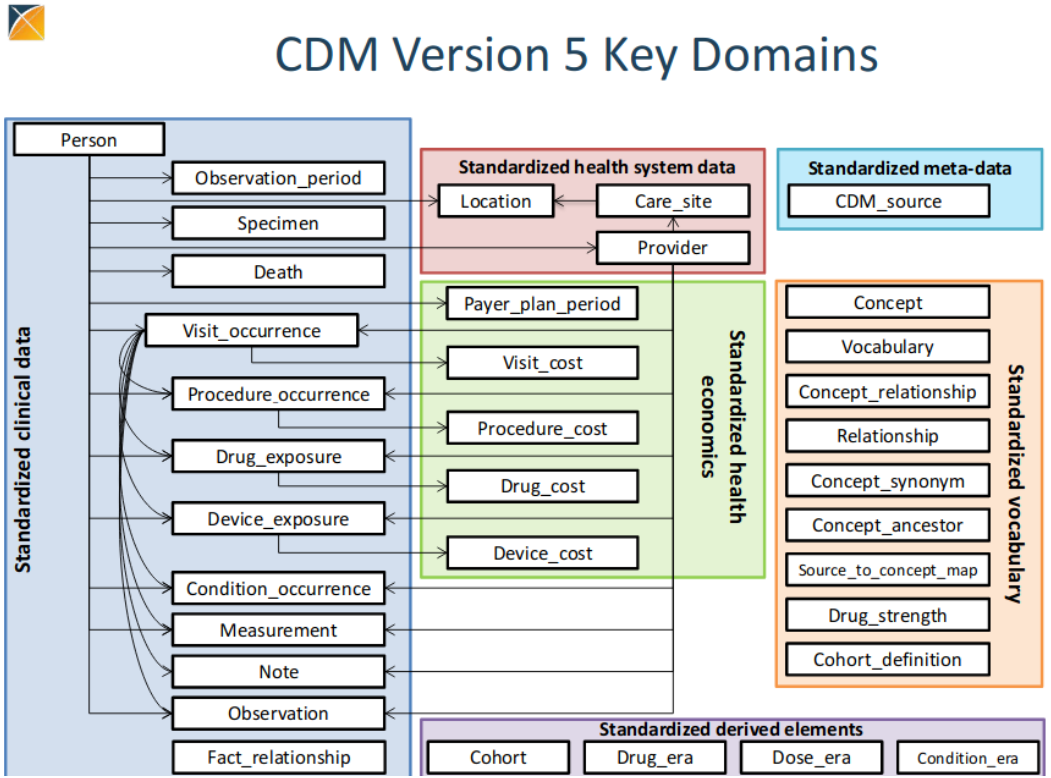
Calcul d'indicateurs de morbidités européens sur le modèle de données de santé **Omop**

- ✓ Utilisation de l'**infrastructure** drees pour tirer profit du calcul distribué
- ✓ Collecte des **informations d'alignement** sur un projet ouvert
- **Ingénierie de la donnée :** Implémentation du nettoyage de la donnée et de la transformation dans du code testé et versionné
- **Calcul des indicateurs :** taux de prévalences et incidences de maladies
- **Application shiny** pour la restitution

le Modèle Commun de données de santé **Omop**

Indicateur I50, infarctus du myocarde

- Code ICD10 I50 (éventuel élargissement)
- Période de 3 ans
- Sources:
  - **Condition_occurrence** : inpatients
  - **Death** : cause de décès
  - **Drugs** : médicaments spécifiques
  - **Visit_occurrence** : outpatients



Conclusion

- **Avantages de ces nouvelles sources de données**
  - Champs d'application et vitesses de calcul
  - Transparence, reproductibilité des traitements
  - Standardisation et mutualisation
- **Limitations**
  - Technicité exigeant des nouvelles compétences et de la formation
  - Centralisation de nombreuses données (risque de mésusage)
  - Impact environnemental

FIN

Slides complémentaires (informations techniques, chiffres, liens)

Infrastructures, rapides comparaisons

Infrastructure	Spécifications techniques	Coût	Usage
CASD, serveur VM 5	16 coeurs, 384 Go RAM, 11 To stockage	tarif PMSI : 1.6K€/mois	Projet
Serveur de calcul Drees (Bigoudi)	80 coeurs, 2 To RAM, +100 To stockage	75K€ achat	Organisme, recherche
Cluster de calcul de Polytechnique	240 coeurs, 1.8 To RAM, 480 To stockage (15 workers et un noeud maître)	150K €	labo recherche
Cluster Insee			Organisme, recherche

Sur les perspectives et la nature du cloud computing, je recommande les deux excellents papiers :

- *Above the clouds: A Berkeley View of Cloud Computing, M. Armbrust et al., 2009*
- *Cloud Programming Simplified : A Berkeley View on Serverless Computing, E. Jonas et al., 2019*

Ingénierie de la donnée

- **Changement de format** pour des formats standards orientés colonnes (csv vers parquet)
- Passage au **calcul distribué** tirant parti de l'infrastructure (logiciel spark)
- Transposition et enrichissement de la **chaîne de traitement** développée pendant 5 ans au cours du partenariat Polytechnique/Cnam : SCALPEL3: a scalable open-source library for healthcare claims databases, Bacry et al., 2019
- Code testé, documenté, versionné et ouvert
- La donnée a un cycle de vie (pipeline) qu'on doit décliner selon les finalités, et auditer régulièrement

Plus grande technicité en amont de l'analyse :
Nécessite des compétences en interne **et** une documentation exemplaire et ouverte sur les transformations effectuées

Analyse, mutualisation des ressources

- Exemples d'open source :
  - Apache Software foundation : nombreux projets fondamentaux en analyse de données (hadoop, spark, …), +7000 contributeurs, +350 projets
  - Scikit-learn : bibliothèque d'apprentissage statistique (origine Inria/Télécom), 1500 contributeurs, +75000 utilisateurs
  - Tensorflow : bibliothèque d'apprentissage profond de Google, +2200 contributeurs, +50000 utilisateurs
  - Documentation textuelle du SNDS : 30 contributeurs, utilisés par +120 personnes
  - Scalpel, nettoyage et extraction d'événements dans le SNDS : 12 contributeurs, utilisé par la CNAM, la Drees, Polytechnique