Application Big Data sur le SNDS

Emmanuel Bacry

Directeur de recherche CNRS Université Paris Dauphine.

Directeur scientifique de l'INDS.

Responsable projets data/santé Ecole Polytechnique.

Youcef Sebiat

Chef de projet data/santé Ecole Polytechnique.

Contexte

- Partenariat CNAM-Polytechnique.

- Objectif : agrandir les champs d'application de traitement de la donné SNDS en utilisant des nouvelles méthodes issues du Big Data et du Machine Learning.

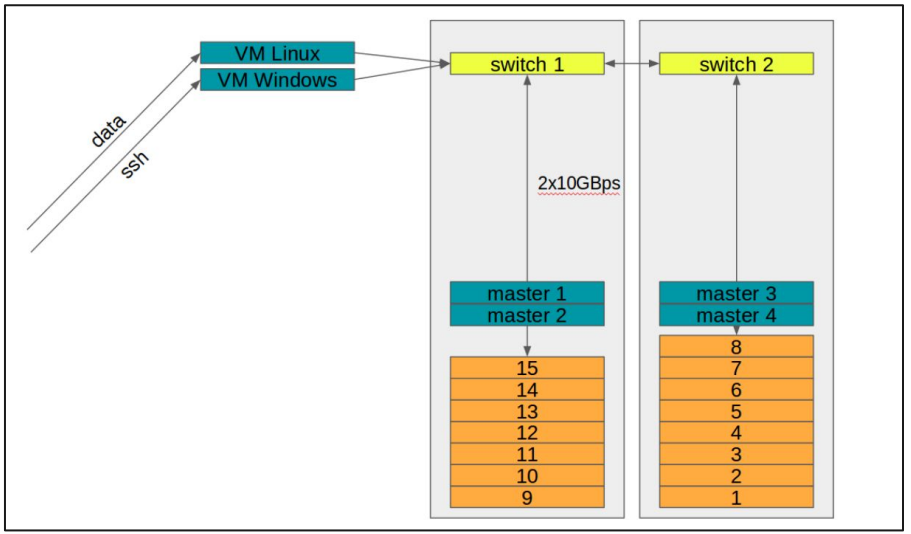
- Partenariat initial 2015-2017 reconduit pour 2018-2020.

Existant

- Portail SNIIRAM.
  - Basé sur Oracle Exadata.
  - Requêtage SQL.
  - Utilisation de SAS pour les analyses statistiques.
- Cadre d'utilisation :
  - Orienté transactionnel.
  - Qualité de données et fiabilité.
  - Excellent pour le suivi des remboursements.
- Analyses machine learning:
  - Besoin de redondance de données, impossible avec SQL.
  - Développement de nouvelles méthodes statistiques impossible avec SAS.
  - BDD existante orientée transaction et non patient.

Infrastructure

- Cluster de machines ordinaires:
  - Master node (1 à 4 dans la figure ci-dessous)
    - 2 processeurs de 4 cœurs cadencés à 3.0 GHz
    - 64 Go de RAM
    - 3 disques SAS 300 Go 15.000 tours/min en RAID 5
  - Slave nodes (1 à 15 dans la figure ci-dessous)
    - 2 processeurs de 8 cœurs cadencés à 2.4Ghz
    - 128 Go de RAM
    - 8 disques SATA 4To 7.2 tours/min en JBOD
    - 2 disques SAS SSD en RAID 0



Choix technologiques

- Stack Open Source
- Spark 2.x pour les calculs et workflows (Seul choix contraignant).
  - Framework logiciel de calcul distribué, basé sur MapReduce.
- Mesos pour la gestion des ressources.
  - Gestionnaire de ressources matérielles sur un cluster de machines.
- HDFS pour le stockage de données.
  - Système de stockage de fichier, distribué et scalable.

Dev

- 5 ans de développement, 2 ans pour une plateforme viable pour du ML.
- Jusqu'à 5 développeurs.
- Création d’une API/ETL de traitement de données.
- En Scala/Spark:
  - Transformer la donnée et la traiter.
- En Python/Spark:
  - Explorer la donnée et tourner les modèles de machine learning.
- Développement dans les bonnes pratiques
  - Couverture en tests unitaires 92%.
  - Jira gestion de tâche / git + GitHub pour le versionning du code.

Dev

- Transformer la vision sur le SNDS.
  - Orienté patient.
  - Orienté analyses et statistiques.
  - Pour chaque patient construire sa trajectoire :
    - Diagnostiques
    - Actes
    - Achats de médicament
- Ouverture du code en Open Source.
- Code disponible dans le Health Data Hub prochainement.

Dev

Quelques chiffres:

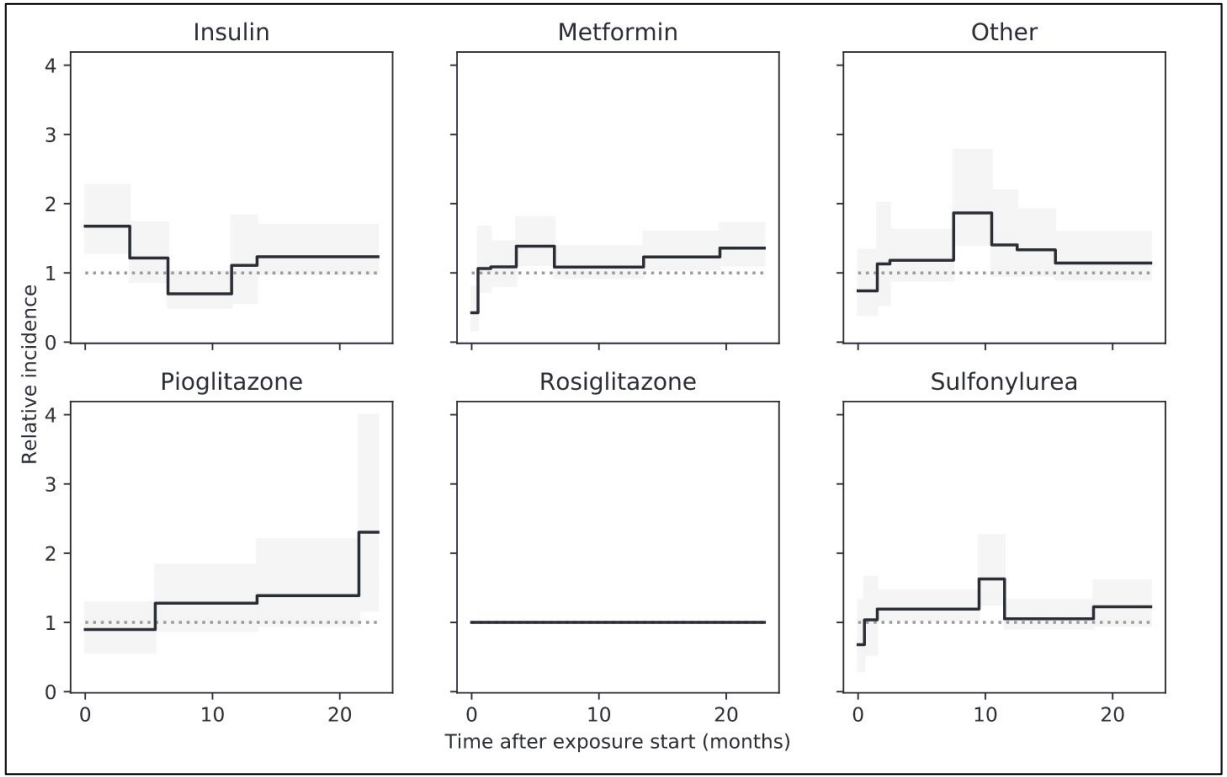
- 20 milliards d'évènements, 450 T (sous-échantillon de 3 ans de données)
- Aplatissement de données (jointure en entrée pour faciliter les analytics) : une année avec 11 milliards de lignes, réparties sur 7 tables avec 9 clé de jointure en 3H.
- Pour 13 millions de patients, extractions de concepts médicaux en 30 mn.

Applications

- Pharmacovigilance avec le cas du Pioglitazone:
    - Antidiabétique retiré du marché en 2011, cause le cancer de la vessie.
- Reproduire le même résultat avec de nouvelles méthodes.
- M. Morel, E. Bacry, S. Gaïffas, A. Guilloux, F. Leroy, ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection, *Biostatistics*

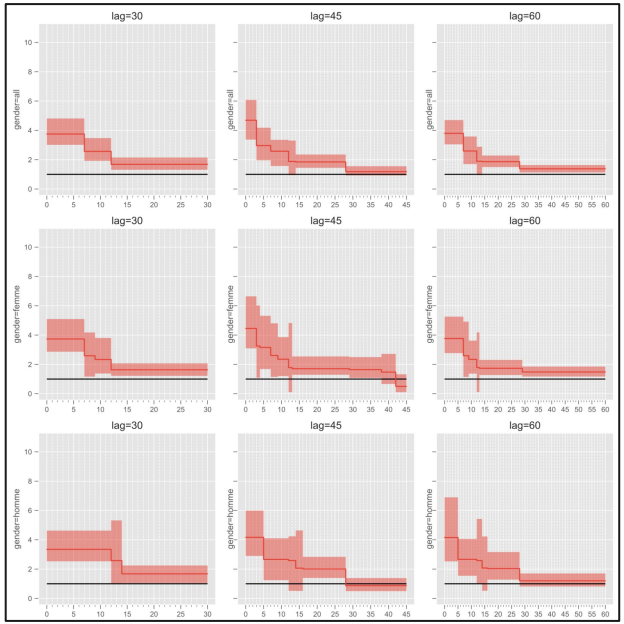
Applications

- 4 ans de données.
- 3.5 millions de patient en entrée.
- 30 mn de traitement pour créer les features.
- 10 mn pour transformer les features en matrices prêt pour modèle ML.



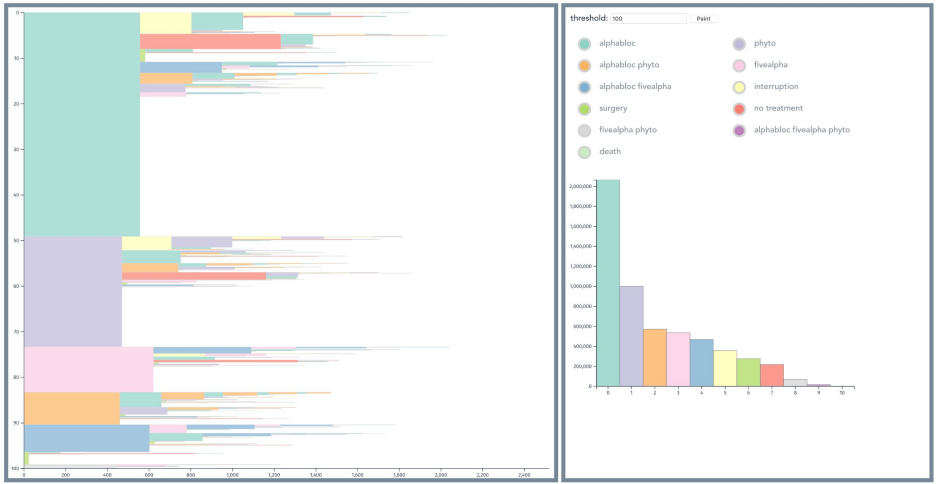
Applications : chute des personnes âgés

- Cas “réel”.
- Plus de 150 molécules étudiées.
- Plus de 13 millions de patients.
- 1H de traitement pour créer les features.
- 10 mn pour transformer les features en matrices prêt pour modèle ML.



Applications : Visualisation de parcours de soins

- Outil de visualisation de parcours patient.
- Scalabilité.
- Facilité d'utilisation.



Applications : Fraude.

- Repérer les trafics de médicaments ainsi que les pratiques dangereuses et fautives des professionnels de santé.

- Techniques machine learning avec contexte Big Data (utilise la même pipeline de données).

Applications : Patient2Vec

- Apprentissage de représentations des parcours patients en utilisant des algorithmes de deep learning.