Machine learning et big data en santé
Le partenariat X / CNAM
Data Drink – Etalab

Stéphane Gaïffas

5 avril 2018

LPSM, Univ. Paris Diderot and CMAP, Ecole polytechnique

1. Présentation du projet

L'équipe

Chercheurs confirmés

- E. Bacry (Univ. Paris Dauphine, Ecole polytechnique, CNRS)
- A. Guilloux (Univ. d'Evry)
- S. G. (Univ. Paris-Diderot, Ecole polytechnique)
- Nombreuses personnes impliquées à la CNAM

Temps complet

- Bio-statisticienne: F. Leroy (CNAMTS)
- Ingénieurs: F. Ben Sassi (CNAMTS), P. Burq (X), C. Giatsidis (X), S. Kumar (X), D. de Paula Silva (X), Y. Sebiat (X)

Le partenariat entre l'Ecole polytechnique et la CNAMTS

Partenariat de recherche sur 3 ans (2015-2017)

- Renouvelé cette année (2018–2020)

But : "big data" et machine-learning sur les données du SNIIRAM/PMSI

- Pharmacovigilance
- Efficacité/efficience des parcours de soins
- etc.

Le partenariat entre l'Ecole polytechnique et la CNAMTS

Alors que :

- SNIIRAM/PMSI n'est pas fait pour ça (à l'origine) !
- C'est la base qui rembourse les soins aux français
- Orientée "événements santé"

Mais :

- Extrêmement riche en informations
- Une des plus grande base de données "electronic healthcare records" du monde

2. Infrastructure

L'existant: infrastructure "verticale" de la CNAMTS

1. **Infrastructure machine** : architecture machine "verticale" propriétaire (IBM Exadata)
2. **Infrastructure base de données** : base relationnelle SQL (800 tables, plusieurs centaines de To) propriétaire (Oracle)
3. **Infrastructure logicielle** : logiciel propriétaire (SAS)

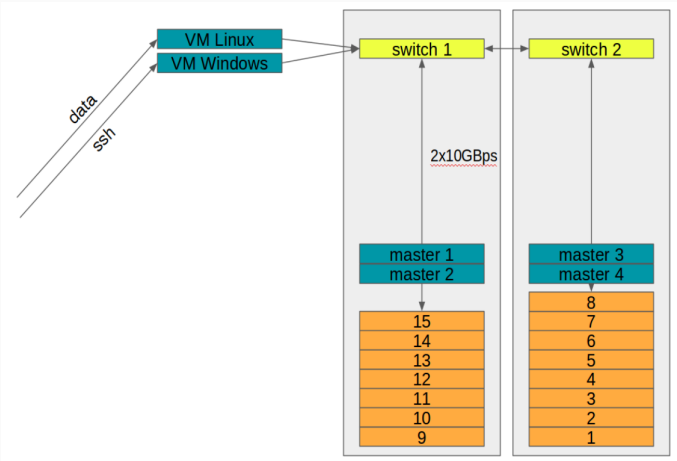
Inadéquation **pour de nouvelles approches type "big data"**

- Architecture très fermée limitant la recherche méthodologique
- Architecture peu adaptée au calcul distribué

Nouvelle infrastructure “horizontale”

Deux cluster big-data (à l’X et à la CNAM)

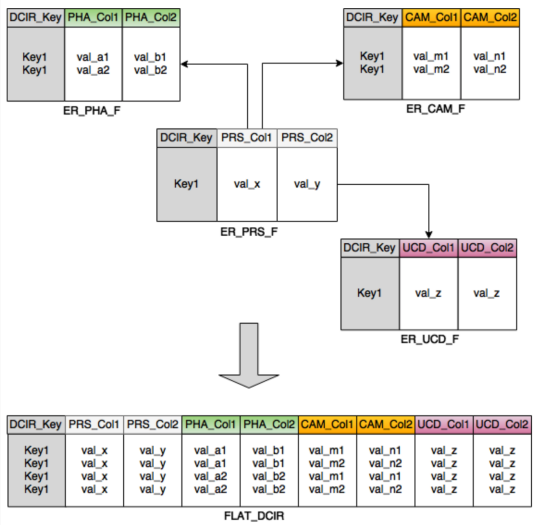
- Données et calculs distribués
- 4 masters, 15 slaves
- 240 cores
- 1.9To RAM, 480To (120 DD)
- Spark, Scala
- Uniquement technologies open-source



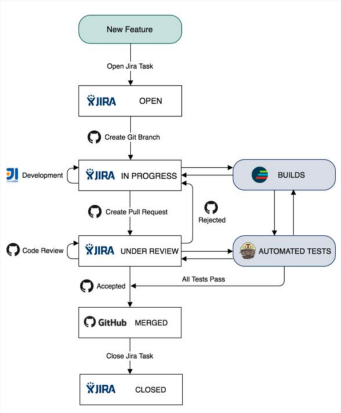
Preprocessing des données

Flattening

- Compréhension, Aplatissement, Organisation
- Stockage distribué
- Open source
- Production en mode “agile”
- Code organisé en ETL



Etape cruciale : passer d’un stockage idéal pour SQL (random access) à un stockage idéal pour traitements “batch”



Architecture logicielle

Featuring. Transformation la donnée brute en une matrice numérique utilisable par les algorithmes d'apprentissage.

- Codé en **scala**
- Utilise le framework **spark** pour distribuer les calculs
- Dépend du modèle et du cas concret à étudier

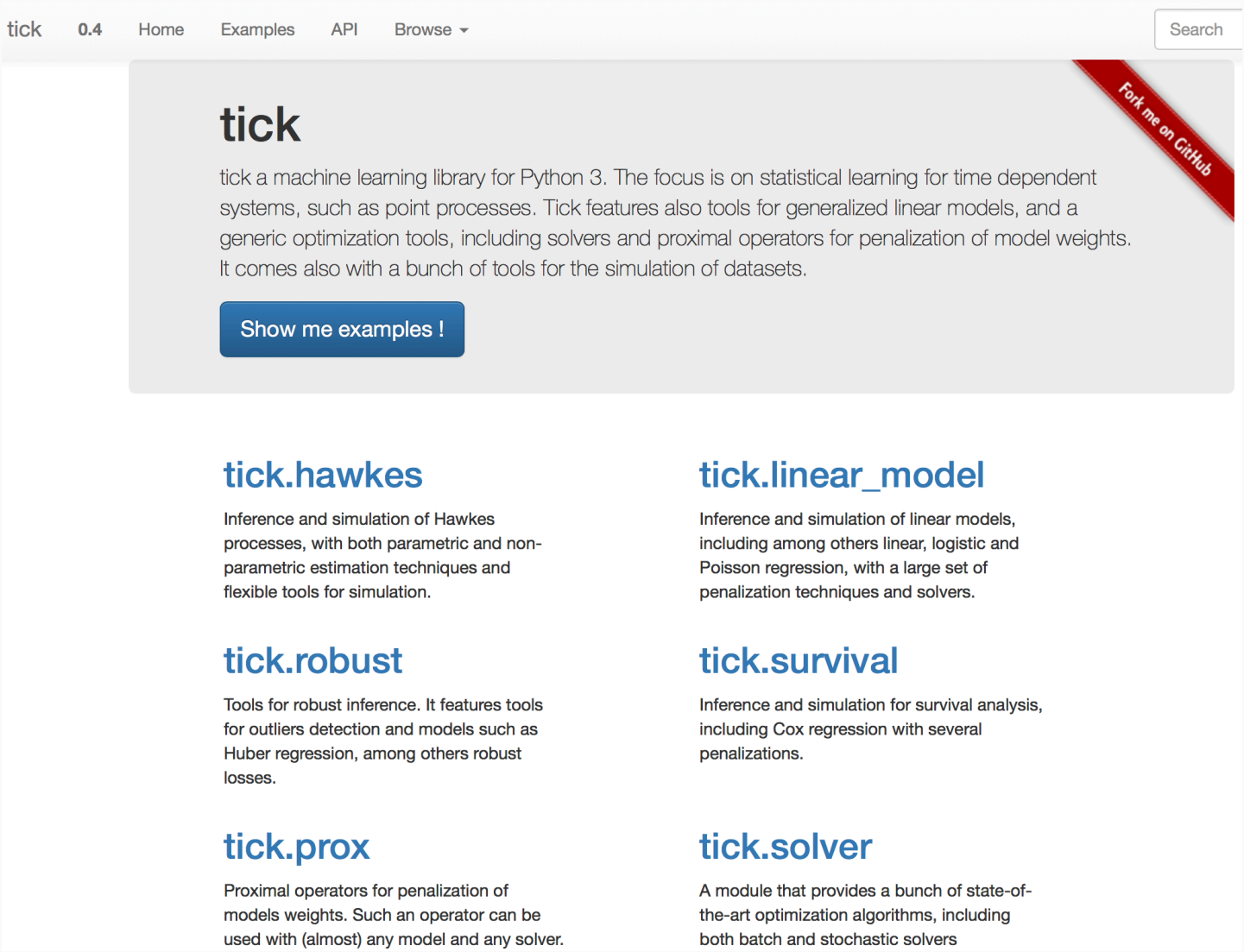
Machine Learning. Inférence statistique

- Utilisation de **R** dans le cas de modèles "classiques"
- Librairie `tick` (Python et C++) développée par l'équipe

Software: `tick` library

- Python 3 et C++11, Open-source (BSD-3 License)
- `pip install tick`
- `https://x-datainitiative.github.io/tick`
- Librairie d'apprentissage statistique pour modèles avec une composante temporelle (santé, finance, etc.)
- Partenariat avec Intel (use-case sur des prototypes, 180 cores)
- Contributeurs bienvenus!

Software: `tick` library



3. Le projet pilote

Cancer de la vessie pour une cohorte de diabétiques

Le “projet pilote” en pharmaco-vigilance

But : développer une **méthode de “screening”** permettant un premier balayage automatique sur plusieurs médicaments.

- $\neq$ validation d’hypothèse
- Etape simplifiée de préparation de cohorte

Application

- Cohorte : diabétiques de type 2
- Effet indésirable : Cancer de la vessie

Identification par screening du Pioglitazone (retiré du marché en 2011)

Cancer de la vessie pour une cohorte de diabétiques

Quelques chiffres

- 2.5 millions de personnes
- 4 ans d'historique
- 1.3 To ($\simeq$ 250Go par an)
- 2 milliards de "lignes" (500 millions de lignes par an)
- "Applatissement de la base" $\simeq$ 40 minutes (depuis spark 2.1)
- Featuring $\simeq$ 10mn

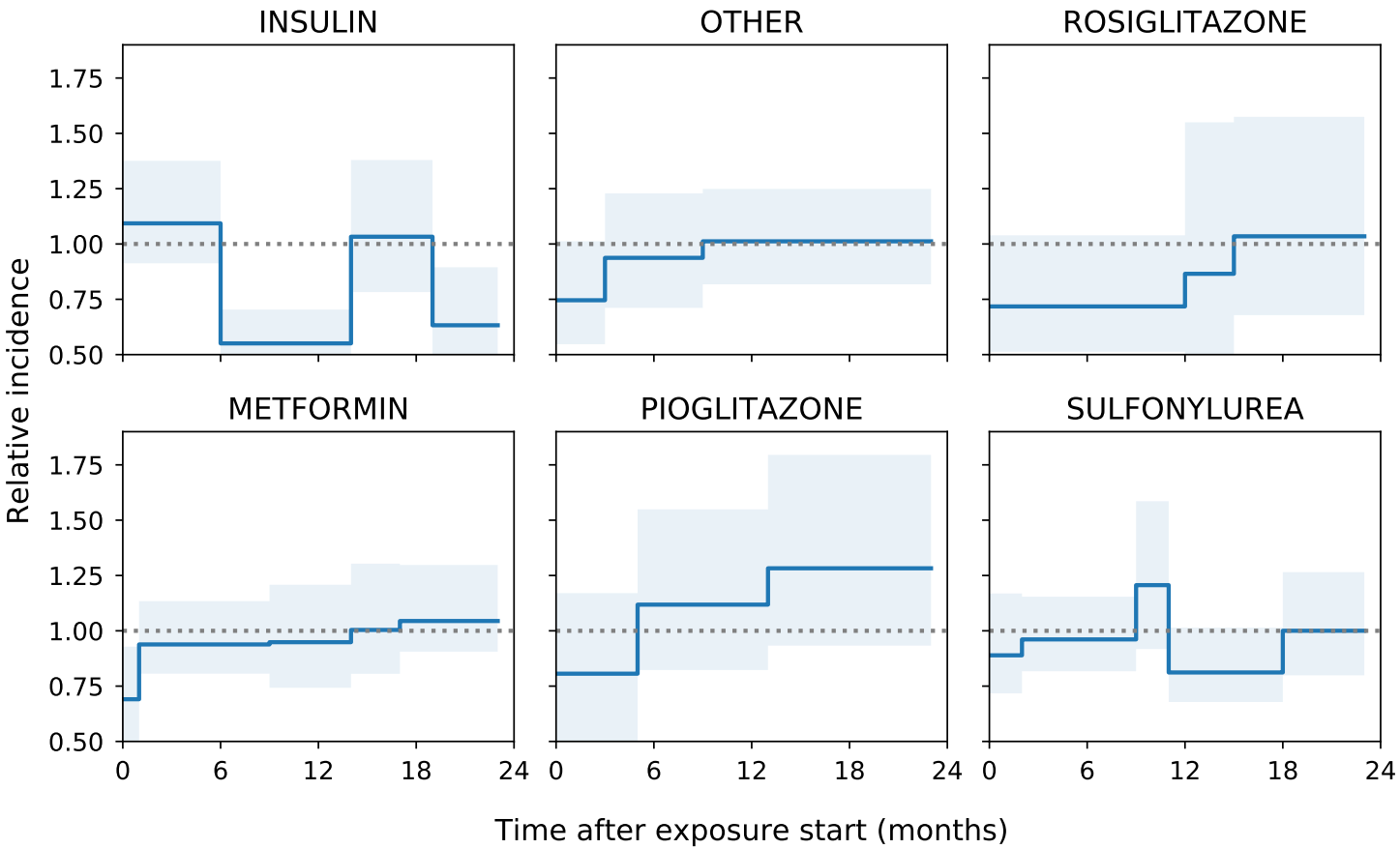
4. Résultats

Un nouveau modèle de screening

Un **modèle SCCS** :

- Modèle “self-controlled case-series” (SCCS) : on ne garde **que les assurés ayant eu un cancer**
- Patients forment leur propre contrôle
- **Travail de préparation de cohorte très simplifiée** : définition de l’exposition
- Modèle longitudinal
- Estimation de l’impact de l’exposition dans le temps à un médicament sur la probabilité de développer un cancer

Résultats obtenus par le nouveau modèle



Méthodologie

Création d'un nouveau modèle

- Approche SCCS: Self-Control Case Series

Revue et corrigée

- Structure auto-régressive des features
- Techniques de pénalisation et cross-validation (signal faible)
- Algorithmes d'optimisation extrêmement rapides

Méthodologie

Algorithme. Minimization par rapport à θ de

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log\left(\frac{\lambda_{ik}(\phi,\theta)}{\sum_{k'=1}^{K}\lambda_{ik}(\phi,\theta)}\right)+\gamma\sum_{j=1}^{J}\sum_{k=1}^{p-1}|\theta^j_{k+1}-\theta^j_k|$$

où

$$\lambda_{ik}(\phi,\theta)=\exp\left(\phi_k+\sum_{j=1}^{d}\sum_{l=1}^{L_i^j}\theta^j_{k-c_{il}^j}\mathbf{1}_{[0,p]}(k-c_{il}^j)\right),$$

avec :

- $\phi_k =$ effet age
- $\theta_t^j =$ effet de l'exposition à la molécule j après t mois d'exposition

5. Conclusion

Conséquences

- **Rapidité :** quelques secondes pour entrainer le modèle du projet pilote (après flattening et featuring)
- **Scalabilité :** sur une machine à de grosses cohortes ($\approx$ 10m individus, $\approx 1K$ features)
- **Plus besoin de réalignement** des expositions des individus (**invariance par translation**).
  Solution à un problème majeur des modèles SCCS, limité à l'exposition **à une seule** molécule dans la litérature !

En cours: étude sur les chutes de personnes agées

- Cohorte : personnes agées
- Médicaments : large classe (plusieurs dizaines)
- Effet indésirable : chute (fracture, fractures par sites)

Quelques chiffres

- 12 millions de personnes (contre 2.5 pour projet pilote)
- $\simeq$ 1.6To par an (contre 250Go pour projet pilote)
- 2 milliards de ligne par an (contre 400 millions pour projet pilote)

Merci !