IA et Système National des Données de Santé
Le Health Data Hub

Emmanuel Bacry

Directeur de Recherche au CNRS
Univ. Paris-Dauphine, PSL

Directeur Scientifique
du Health Data Hub

Directeur Projets Data/santé
Ecole Polytechnique

emmanuel.bacry@polytechnique.fr
http://www.cmap.polytechnique.fr/~bacry

Les bases de données santé en France

- Bases de données hospitalières (entrepôt AP-HP, Institut Curie, . . .),
- Bases de données d'entreprises privées (CEGEDIM, Sanofi, . . .)
- . . .
- **Une spécificité franccaise : le SNDS** (Système National des Données de Santé)

Le SNDS : une des plus grosses bases de données “santé” au monde

- base de données comptable (SNIIRAM/DCIR = Données de consommation inter-régimes) +
- PMSI : Programme de médicalisation des systèmes d'information (Parcours hospitalier) +
- CepiDC +
- ...

Quelques chiffres ...

- plus de 65 millions de personnes
- 1,2 Md de feuilles de soins par an
- 11 millions de séjours hospitaliers par an
- plus de 250To gérés par la CNAM

Un impact sociétal potentiel énorme sur la santé

$\longrightarrow$ **pharmaco-vigilance** : Identifier des médicaments actuellement sur le marché qui provoquent des effets secondaires néfastes

- 2013 : lien entre les pillules de 3ième génération et le risque d'embolie pumonaire
- 2015: le Mediator !

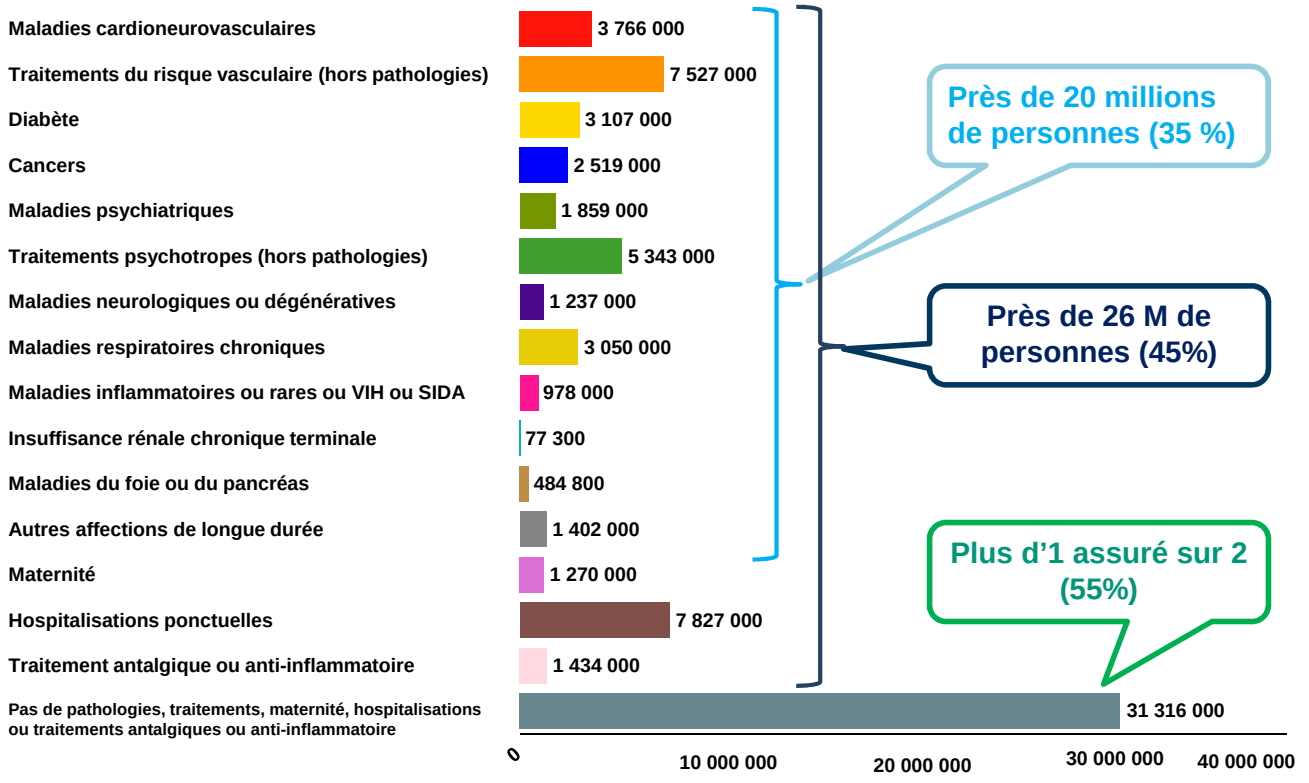
Mais aussi un impact sociétal potentiel énorme sur l'économie

Budget de la santé publique en France $\simeq$ 170 milliards

- 2014 : Cartographie de 54 pathologies (HIV $\simeq$ 50 criteres)

$\longrightarrow$ **Optimisation de parcours de soin** pour une pathologie donnée

Parmi les 57,1 millions de bénéficiaires du régime général, en 2015
Les effectifs



Champ : Régime général (y compris SLM)- France entière
Source : Cnamts (cartographie 2015- version G4)

L'Assurance Maladie Caisse Nationale 4

- Partenariat de recherche entre Polytechnique et CNAM 2015-2017, 2018-2020
- Equipe de plus de 11 personnes à temps plein (donc 5 développeurs)
- **But** : tester le potentiel des techniques/méthodologies "big data" sur les données du SNDS

$\Longrightarrow$ **Refonte de l'infrastructure de la CNAM !**

Architecture actuelle (entièrement propriétaire : Oracle + SAS)

- **Accès relativement lent**
  $\longrightarrow$ aucun besoin de streaming (batches délivrés au coup par coup par les CPAM)
- Structure : **bases de données relationnelles** (plus de 800 tables) **avec un schéma en étoile**
- Volume : **20 milliards d'événements, 450 To** (3 ans de données)
- Outil d'analyse statistique (SAS) limité

$\Longrightarrow$ **Freins à la recherche méthodologique**

- **Infrastructure des données**
  Base de données SQL $\longrightarrow$ grand "tableau" (parquet) centré sur l'individu
- **Infrastructure machine**
  Architecture "verticale" propriétaire (Exadata) $\longrightarrow$ Architecture "horizontale" (cluster Hadoop)
- **Infrastructure logiciel pour l'analyse statistique**
  Logiciels propriétaires (SAS) $\longrightarrow$ Scala, Spark, C++, Python,

Tout le framework de la pipeline dévelopée est open source.

SCALPEL3: a scalable open-source library forhealthcare claims databases, E.B., S.Gaiffas, F.Leroy, M.Morel, D.P.Nguyenc, Y.Sebiat, D.Sun, to come (very) soon

Le “projet pilote” en pharmaco-vigilance

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

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

Application

- Cohorte : diabétiques de type 2
- Médicament : ensemble des antidiabétiques
- Effet indésirable : Cancer de la vessie

$\longrightarrow$ dépistage du Pioglitazone (retiré du marché en 2011)

Setting

- We have individuals $i = 1, \ldots, n$
- Time $[0, T]$ is partitioned in intervals $I_1, \ldots, I_B$ (length=month, week or day)
- We observe the number of adverse events $y_{i,b} \in \mathbb{N}$
- We put $n_i = \sum_{b=1}^{B} y_{i,b}$ = total number of adverse events of individual i
- We observe longitudinal features $x_{i,b} = (x_{i,b}^1, \ldots, x_{i,b}^d) \in \mathbb{R}^d$ over time intervals $b = 1, \ldots, B$ (drugs exposures, etc.)
- We observe "static" features $z_i = (z_i^1, \ldots, z_i^p) \in \mathbb{R}^p$ (gender, age if B is small, etc.)

Autoregressive features

The intensity of occurrence of adverse events at time b depends on feature j :

$$\lambda_{i,b} = e^{\langle \boldsymbol{X}_{i,b'}, \boldsymbol{\theta} \rangle + \beta^\top z + c_b},$$

where:

- θ_j^k = effect of feature j when exposure occurred k time intervals before the current one
- $x_{i,b}^{j,k}$ = exposure of individual i to drug j that occurred k intervals before interval b

Leads to a **translation-invariant parametrization** of the model

- **no “time-realignment”** between individuals is required
- strong improvement compared to SCCS literature, where only one type of exposure, i.e. a single molecule, can be used !

More precisely, we have a Multinomial law (conditionnaly on n and x) :

$$\mathbb{P}(y_{i,1},\dots,y_{1,B}|n_i,x_i) = \frac{n_i!}{\prod_{b=1}^B y_{ib}!}\prod_{b=1}^B\left(\frac{e^{\langle \boldsymbol{x}_{i,b},\boldsymbol{\theta}\rangle}}{\sum_{b'=1}^B e^{\langle \boldsymbol{x}_{i,b'},\boldsymbol{\theta}\rangle}}\right)^{y_{i,b}}$$

Penalization

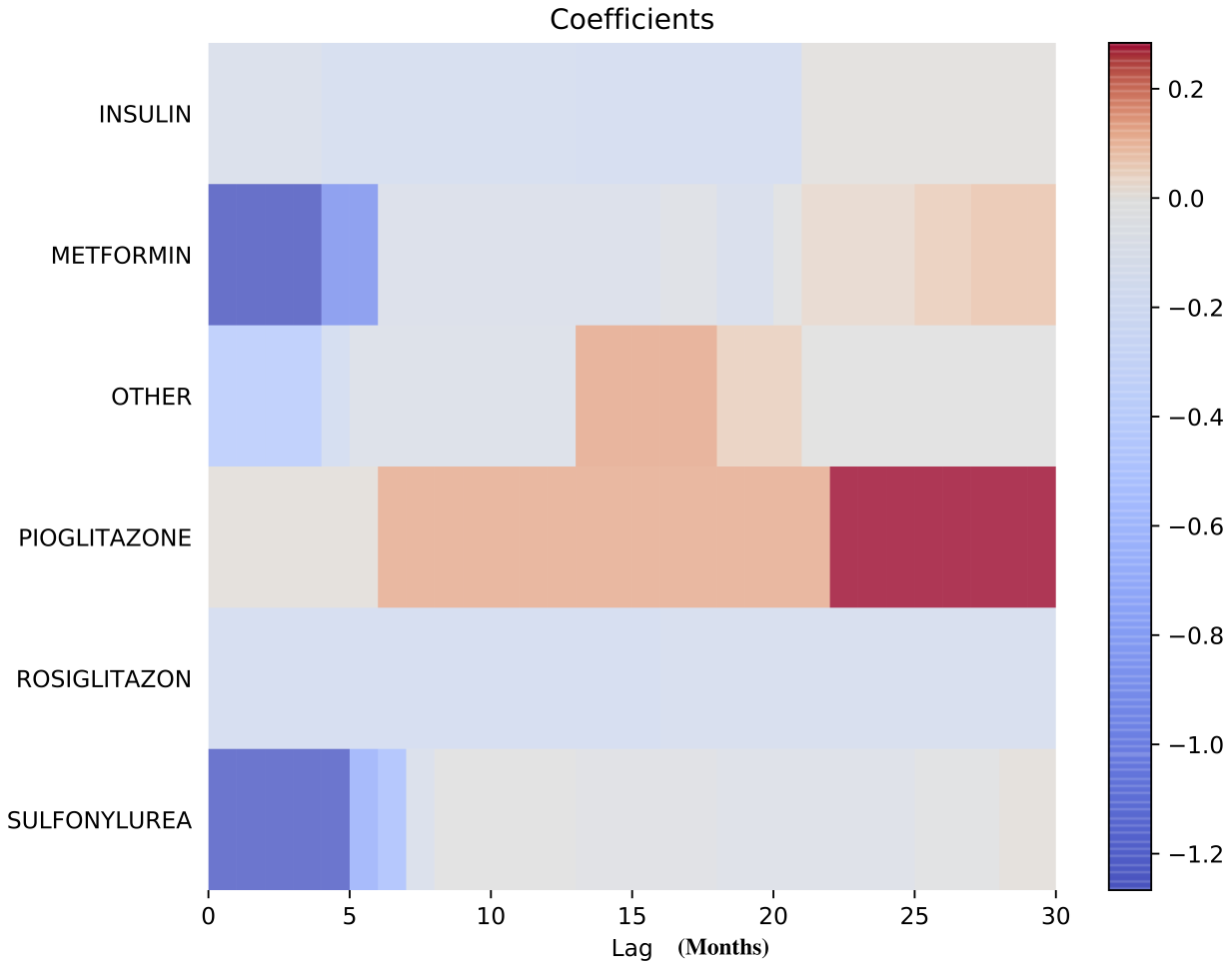
- We want to consider a large number of lags K, but we want to "smooth" time-adjacent coefficients $\theta^j_1,\dots,\theta^j_B$
- We use "group" total-variation penalization

Algorithm

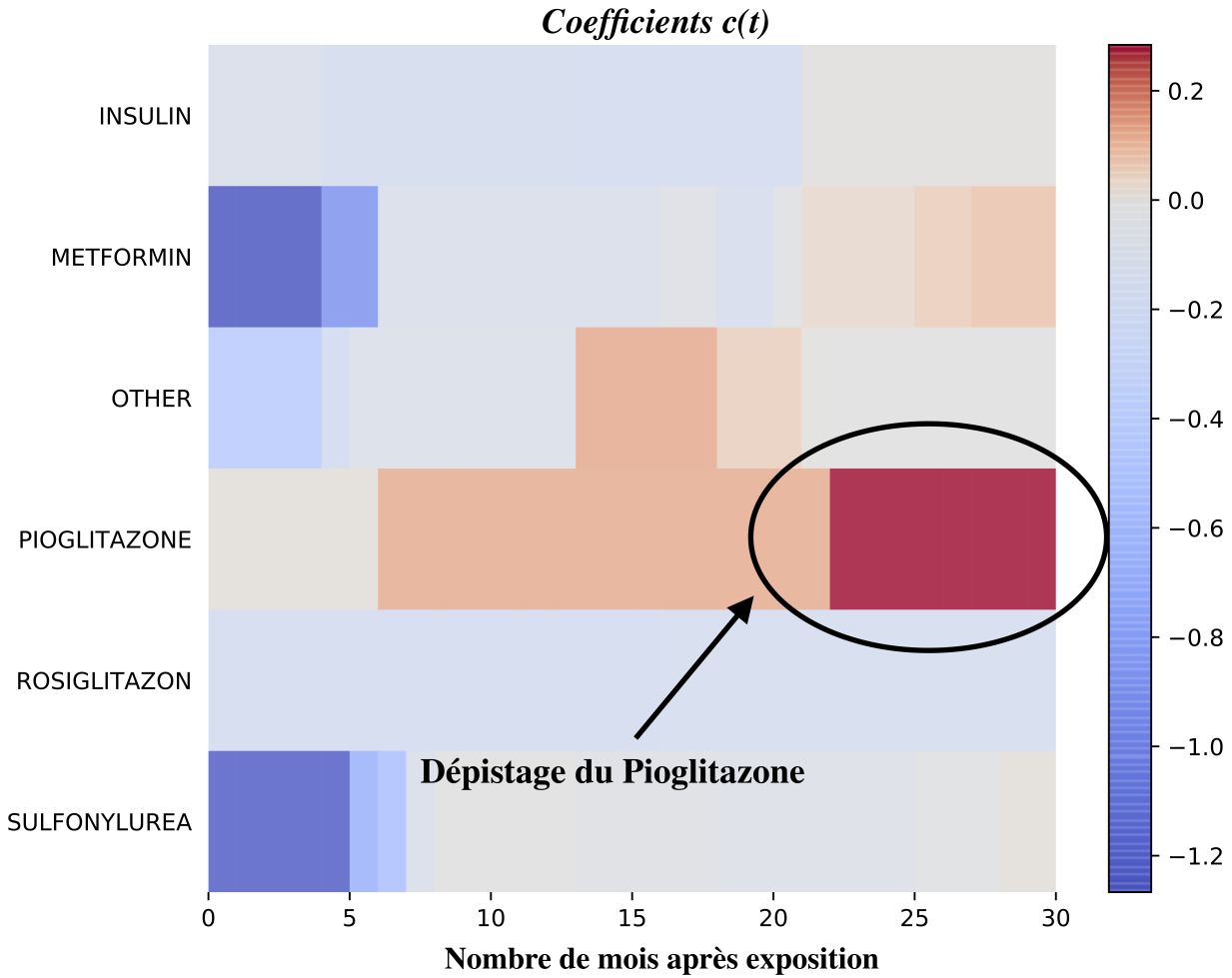
We minimize the following over $\boldsymbol{\theta}$

$$-\frac{1}{n}\sum_{i=1}^n\sum_{b=1}^B \delta_i y_{i,b}\left(\langle \boldsymbol{x}_{i,b},\boldsymbol{\theta}\rangle - \log\Big(\sum_{b'=1}^B e^{\langle \boldsymbol{x}_{i,b'},\boldsymbol{\theta}\rangle}\Big)\right) \quad + \lambda\sum_{i=1}^d\sum_{k=1}^{B-1}|\theta^j_k-\theta^j_{k}$$

M.Morel, E.B., S.Gaïffas, A.Guilloux, F.Leroy, 2018



M.Morel, E.B., S.Gaïffas, A.Guilloux, F.Leroy, 2018



- Cohorte : **seniors**
- Médicaments : $\simeq 100$**) médicaments**
- Effet secondaire : **chute** (fractures)

Quelques chiffres (cluster HDFS 20 noeuds)

- 12 millions de personnes (versus 2.5 pour le projet pilote)
- 4 ans de suivi
- $\simeq$ 8To par an (versus 250Go pour le projet pilote) $\rightarrow$ 32To
- Applatissement de la base $\simeq$ 2-3 jours
- Préparation des données $\simeq$ 35 minutes
- Temps pour estimation $\simeq$ 2 minutes
- Cross validation $\simeq$ 45 minutes

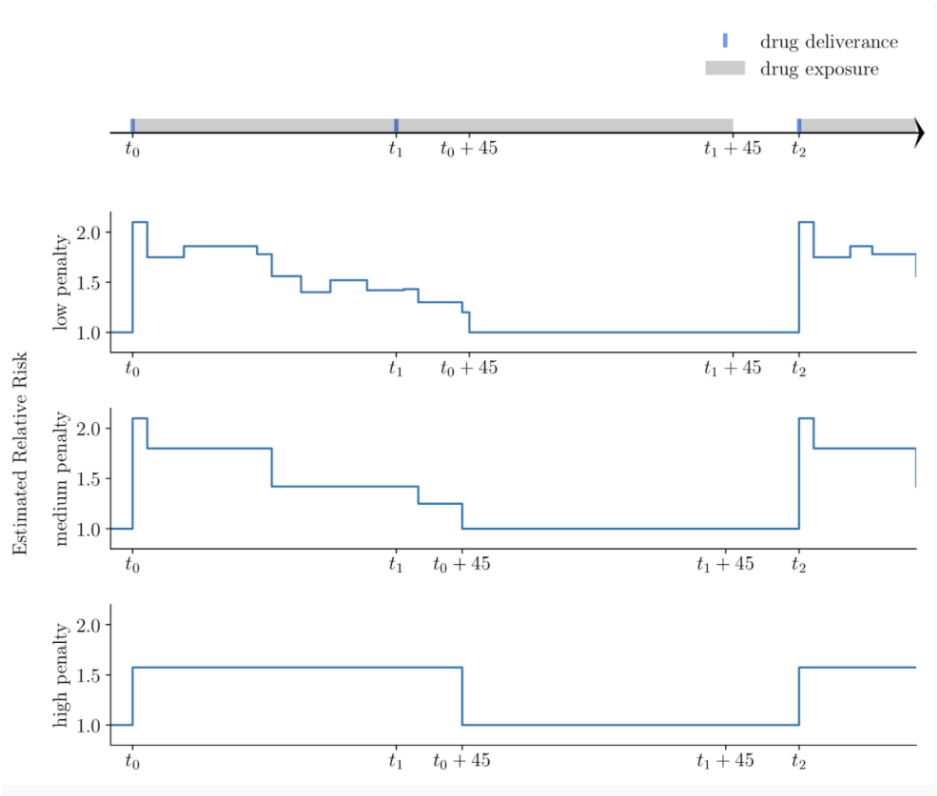
Population ciblée

- Personnes de plus de 65 ans
- Observées de 2014 à 2016
- Nouveaux utilisateurs d'Anxiolytiques, Hypnotiques, Antidépresseurs, Neuroleptiques (AHAN) en 2015
- Suivis de 01/2015 à 12/2016
- Ayant eu au moins une fracture

$$13.746.652 \text{ personnes} \longrightarrow 76.629 \text{ personnes}$$

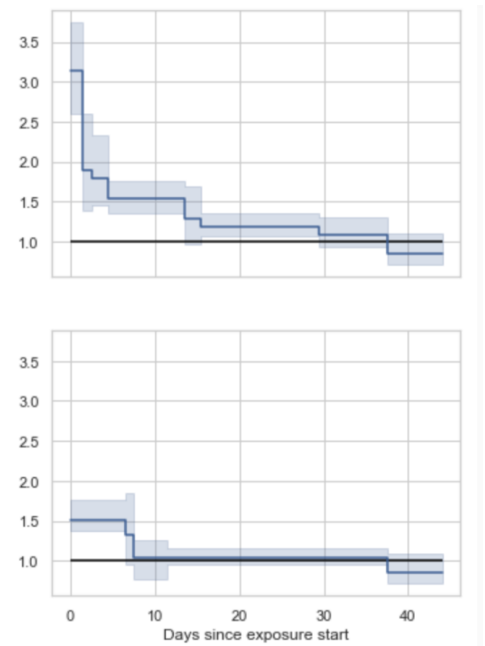
M.Morel, E.B., S.Gaïffas, A.Guilloux, F.Leroy, 2018

- Le patient est son propre controle
- Estimation longitudinale du risque
- Peu d'a priori sur les risques
- Robuste à l'omission de contrôles statiques



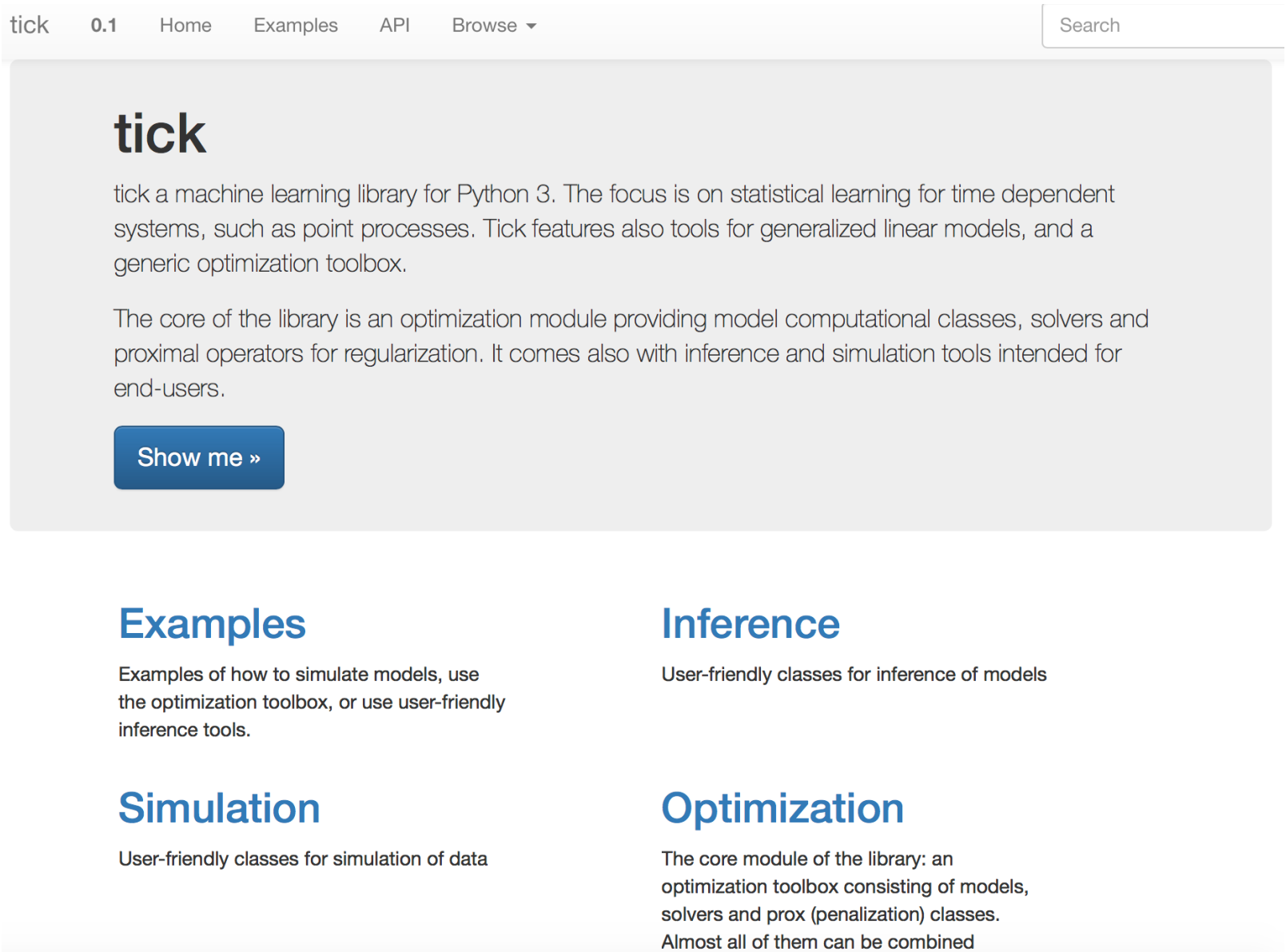
M.Morel, E.B., S.Gaiffas, A.Guilloux, M.Laanani, F.Leroy, D.P.Nguyen, Y.Sebiat, to come (very) soon

- Risques relatifs pour chacune des molécules étudiées
- Etudes de nombreuses molécules en une fois
- Résultats cohérents avec des études plus spécifiques
- Quelques résultats originaux . . .



E.B., M.Bompaire, S.Gaiffas, S.Poulsen, 2018, + P.Deegan, ...

- Python 3 et C++11
- Open-source (BSD-3 License)
- `pip install tick` (on MacOS and Linux...)
- `https://x-datainitiative.github.io/tick`
- Statistical learning for time-dependent models
- Point processes (Poisson, Hawkes), Survival analysis, GLMs (parallelized, sparse, etc.)
- A strong simulation and optimization toolbox
- Partnership with Intel (use-case for new processors with 256 cores)
- BNP-P, CNAM, Euronext, Médéric Malakoff, ...
- More contributors welcome!


tick
0.1
Home
Examples
API
Browse
Search
tick
tick a machine learning library for Python 3. The focus is on statistical learning for time dependent systems, such as point processes. Tick features also tools for generalized linear models, and a generic optimization toolbox.
The core of the library is an optimization module providing model computational classes, solvers and proximal operators for regularization. It comes also with inference and simulation tools intended for end-users.
Show me »
Examples
Examples of how to simulate models, use the optimization toolbox, or use user-friendly inference tools.
Inference
User-friendly classes for inference of models
Simulation
User-friendly classes for simulation of data
Optimization
The core module of the library: an optimization toolbox consisting of models, solvers and prox (penalization) classes. Almost all of them can be combined

- Pharmaco-vigilance
- Visualisation d'un très grand nombre de parcours de soin sous une pathologie donnée
- Détection de fraudes
- ...

Le Health Data Hub : Un projet national

Décrit dans la loi "Ma santé 2022" votée en Juillet 2019

- Le hub sera le portail privilégié du SNDS pour des projets de recherche d'intérêt public (opérés par des institutions publiques ou entreprises privées)
- Le SNDS a été étendu à toutes les données issues de soins remboursés partiellement par l'assurance maladie ($\simeq$ toutes les données de santé franccaises !)

Principal ambition : *Mettre au servicce du plus grand nombre toutes ces données dans le respect de l'éthique et du droit des citoyens*

Quelques chiffres

- Travail a débuté en Janvier 2019 (INDS + DREES + prestataires)
- Budget : 80m € pour les 4 premières années
- Loi votée en Juillet 2019
- La strcuture Health Data Hub verra le jour courant novembre 2019
- La platforme v0.1 sera prête courant novembre 2019

Quelques exemples de défis

1. Consolider des données de santé massives et hétérogènes sur une infrastructure moderne.
   - Données Massives : SNDS, Données hospitalières, Données du privé , ...
   - Données Hétérogènes : EHR, images, omiques, analyses biologiques , ...
2. Partage des données suivant une gouvernance unifiée
   - Décret définissant cette gouvernance

Quelques exemples de défis (suite)

3. Création d'un tiers de confiance pour accès sécurisé paprtagé (+ ressources partagées) pour la valorisation des données
   - Connections fortes avec les acteurs de cet écosystème (public/privé)
   - Devenir l'un des acteurs principaux du décelopement de cet écosystème
4. Identificcation de projets prometteurs en terme de santé publique et promotion du secteur industriel
   - Devenir l'une des meilleures plateforme de données de snaté au monde
5. Devenir une plateforme de référence de mise en relation compétences/intérêts

et ... un contre-pouvoir (essentiel) aux GAFA/BATX ?

Les premiers projets sélectionnés (10 sur . . . 189)

- **DeepSark** : Base de données **quasi exhaustive dur le sarcome** en France
- **NS-Park** : Cohorte de **20k patients Parkinson** (données cliniques, omiques, ...)
- **Rexetris** : Données de transplantés rénaux
- **Deep.Piste** : **250k mamographies** annotées
- **Hydro** : Données temps réel de plus de **8000 Pacemakers**

Les premiers projets sélectionnés (10 sur . . . 189)

- **Oscour** : Base de données exhaustives de **tous les services d'urgence** en France
- **Pimpon** : Base de données de Vidal
- **Ordei** : Base de données de pharmacovigilane de l'ANSM
- **Parcours IDM en IDF** : Base de données SAMU78
- **ARAC** : Données de remboursements de Malakoff Mederic (> 1 million de personnes)