Big Data in the French Public Health System

Emmanuel Bacry

Researcher at CNRS
Associate Professor
Head of the “Data Science Initiative”

Centre de Mathématiques Appliquées
Ecole Polytechnique

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

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

l’X
ÉCOLE
POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

Data Science and Public Health

- Data Science (Satistical Analysis) has always been at the heart of health-related problematics
- Strong health impact (HIV, cigarettes, Mediator, . . .)
- Strong economical impact (first state budget in France)
- Many “Big Data” sets in France : CNAMTS, AP-HP, . . .
- But “Big Data techniques” hardly used

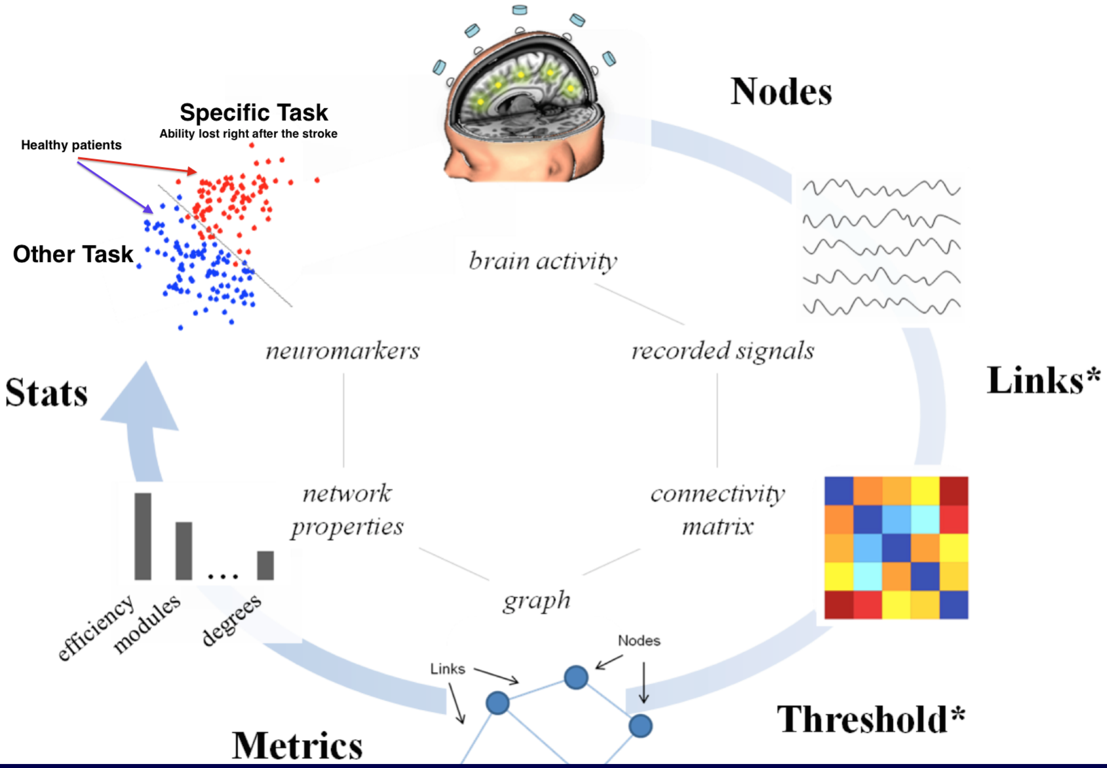
Big Data and Public Health at Ecole Polytechnique

- A team with various skills : signal/image processing, statistics, machine learning, computer science, . . .
- Both Maths Lab (CMAP) and Computer Lab (LIX) are involved
  - 10 Researchers : S.Allassonière, E.B., Y.Diao, S.Gaiffas, A.Guilloux (UPMC/X), J.Josse, M.Lavielle, E.Moulines, E.Scornet, M.Vazirgiannis
  - 11 Phd students or PostDocs
  - 5 "Big Data" engineers
  - Many internships
  - More to come !
- Many partners : AP-HP, CNAMTS, HEGP, Tenon, ICM, . . .

ICM (Institut du cerveau et de la moëlle épinière) - La Pitié Salpétrière Hospital

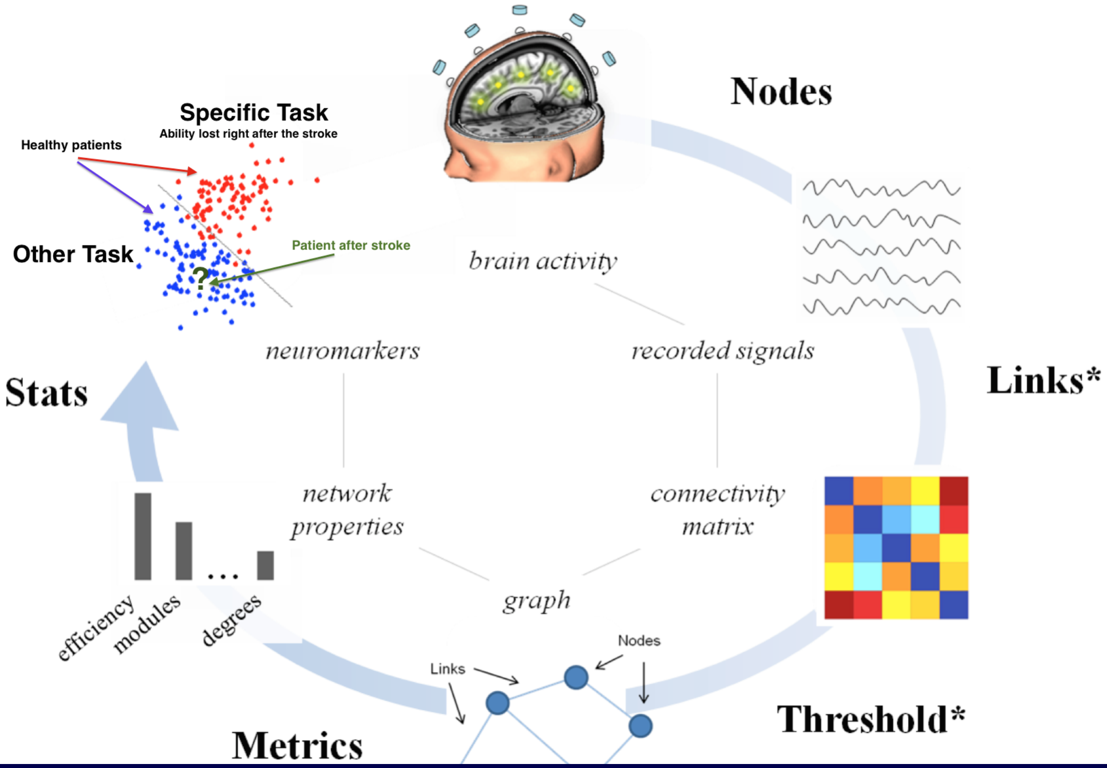
Modelling brain connectivity using sparse graphical model
Application to stroke evolution analysis

S.Allassonière, F.Deloche (Polytechnique), F. de Vico Fallani (ICM)



Modelling brain connectivity using sparse graphical model
Application to stroke evolution analysis

S.Allassonière, F.Deloche (Polytechnique), F. de Vico Fallani (ICM)



Diagnostic Aid

S.Gaïffas, S.Bussy (Polytechnique), A.Guilloux (UPMC,Polytechnique) A-S.Jannot (HEGP)

The Database at HEGP : one of the largest Hospital data center in France

- 1.4 million patients
- 15 year historic
- All the data of the patient's hospital stay (X-rays, biological data, prescriptions, . . .)
- Specialized in *complex* pathologies

One particular pathology : Vaso-Occlusive Crisis (in drepanocytosis) :

- This crisis calls for hospitalization (morphine for a few days)
- When does the crisis stop ? When to stop morphine ?
- Hospitalization monitoring
- Minimize the rate of re-hospitalization

Diagnostic Aid

S.Gaïffas, S.Bussy (Polytechnique), A.Guilloux (UPMC,Polytechnique) A-S.Jannot (HEGP)

The Database at HEGP : one of the largest Hospital data center in France

- 1.4 million patients
- 15 year historic
- All the data of the patient's hospital stay (X-rays, biological data, prescriptions, . . .)
- Specialized in *complex* pathologies

One particular pathology : Vaso-Occlusive Crisis (in drepanocytosis) :

- This crisis calls for hospitalization (morphine for a few days)
- When does the crisis stop ? When to stop morphine ?
- Hospitalization monitoring
- Minimize the rate of re-hospitalization

Prediction of arrival flows in emergency services

P.Aegerter (AP-HP), E.B., S.Gaïffas, M.Wargon (AP-HP)

The Database :

- Arrival flows over 5 years
- $>$ 80 Emergency services in Ile-de-France
- Specific data on arrivals

Forecast :

- Forecast at various time-horizons and different scales
- Influence of various environmental factors
- Characterization of the emergency services network
- Typology of the different services

Prediction of arrival flows in emergency services

P.Aegerter (AP-HP), E.B., S.Gaïffas, M.Wargon (AP-HP)

The Database :

- Arrival flows over 5 years
- $>$ 80 Emergency services in Ile-de-France
- Specific data on arrivals

Forecast :

- Forecast at various time-horizons and different scales
- Influence of various environmental factors
- Characterization of the emergency services network
- Typology of the different services

The data from the Caisse Nationale d'Assurance Maladie (CNAMTS)

- **The SNIIRAM database** :
  - Accounting (main) database +
  - PMSI database (hospital data) +
  - Hypocrate database (physician data) +
  - . . .

The data from the Caisse Nationale d'Assurance Maladie (CNAMTS)

- **The SNIIRAM database** :
  - Accounting (main) database +
  - PMSI database (hospital data) +
  - Hypocrate database (physician data) +
  - ...

- **SNIIRAM** $\simeq$ **largest health database in the world**
  - 65 million people
  - $\simeq$ 500 Terabytes !

Statistical analysis of SNIIRAM database

Very strong potential impact

- **Health impact**
  → 2013 : used to show that 3rd generation contraception pill increases pulmonary embolism risk
  → 2014 : Cartography of 54 important pathologies (HIV $\simeq$ 50 criteria)
- **Economical impact** (budget $>$ 170 billion euros/year)
  → medico-economical chart to quantify the cost of the different pathology

The partnership Polytechnique-CNAMTS

- **3-year partnership** (2015-2017) Polytechnique-CNAMTS
- **CNAMTS opens all SNIIRAM to Polytechnique research team**
- **Many themes of research**
  - Identifying useful factors in medico-economic path-ways of given pathologies
  - Weak signal detection or anomaly detection in pharmacoepidemiology
  - Fraud detection
  - . . .

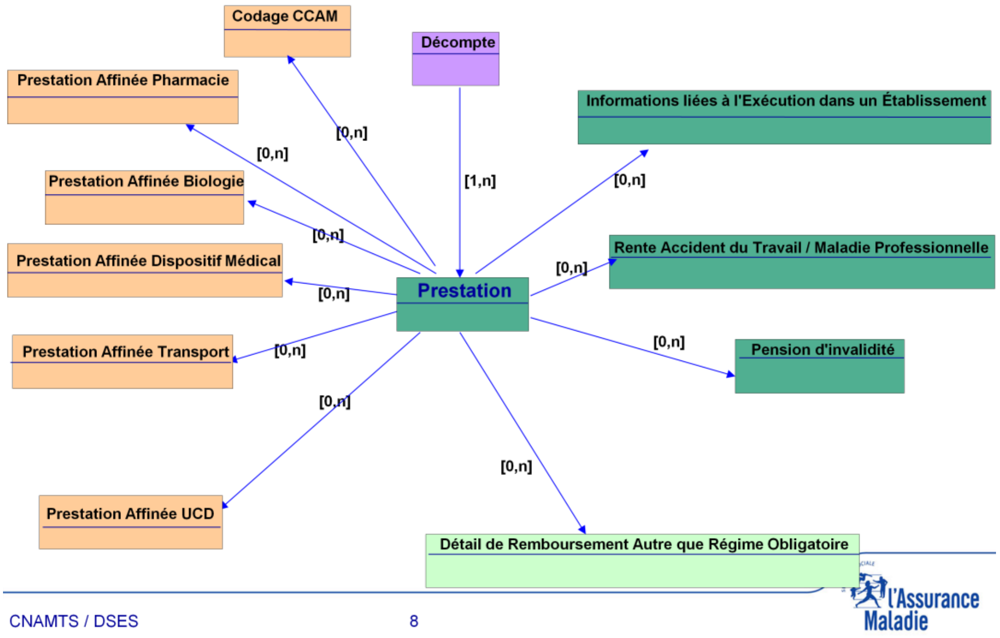
Main challenges

- **Design of scalable machine learning algorithms**

Main challenges

- **Design of scalable machine learning algorithms**

- **HOWEVER : Implementation requires . . .**
  $\Longrightarrow$ Pre Processing of the database
  - SNIIRAM : Oracle relational database with .... $\simeq$ 1000 tables !
  - Allowance oriented


Codage CCAM
Décompte
Prestation Affinée Pharmacie
Informations liées à l'Exécution dans un Établissement
[0,n]
[0,n]
[1,n]
[0,n]
Prestation Affinée Biologie
[0,n]
Rente Accident du Travail / Maladie Professionnelle
Prestation Affinée Dispositif Médical
[0,n]
Prestation
[0,n]
Prestation Affinée Transport
[0,n]
[0,n]
Pension d'invalidité
[0,n]
[0,n]
Prestation Affinée UCD
Détail de Remboursement Autre que Régime Obligatoire
CNAMTS / DSES
8
l'Assurance Maladie

Pre processing of the database

- SNIIRAM needs to be "flattened" (Parquet)
  - Patient oriented
  - Doctor/Institution oriented
  - . . .

⟹

- Very few constraints
- Efficient request
- BUT :
  - Significant increase in storage size (redundancy)
  - flattening process is a very heavy process

Pre processing of the database

Event representation

- **“Low-Level” structure** $\simeq$ SNIIRAM original structure
- **“High-level” structure** (done with a medical expert)
  - Pathology definitions ?
  - Structuring health path-ways (periodic/continuous treatment,. . .) ?
  - Medication structuring (same molecules, . . .) ?
  - . . .

A “smaller” project with Tenon Hospital : around the Prostate Cancer

- **Observapur database** : 10-year old subset of SNIIRAM database restricted on identified prostate cancer patients.
  - Pr. B.Lukacs, Tenon Hospital and Pr. E.Vicault, URC Saint-Louis Lariboisière Fernand-Widal
  - Since 2004 : $\simeq$ 2.4 million patients total
- **Information** (lightly) **structured** (thanks to expert)
- **Research themes**
  - Automatic structuring of implications of Prostate Cancer
    → Unsupervised learning for identifying “latent implications”
  - Specificity of Type II Diabetes in prostate cancer pathways
    → Design of new scalable algorithm in survival analysis
  - . . .

Big Data in French Public Health System at Ecole Polytechnique

- **Big** Data
- **Big** Projects
- **Big** scientific challenges (Maths + Computer science)
- (potentially) **Big** impacts
- not **Big** enough Team ☹ : **WE ARE HIRING !**

BIG Adventure!