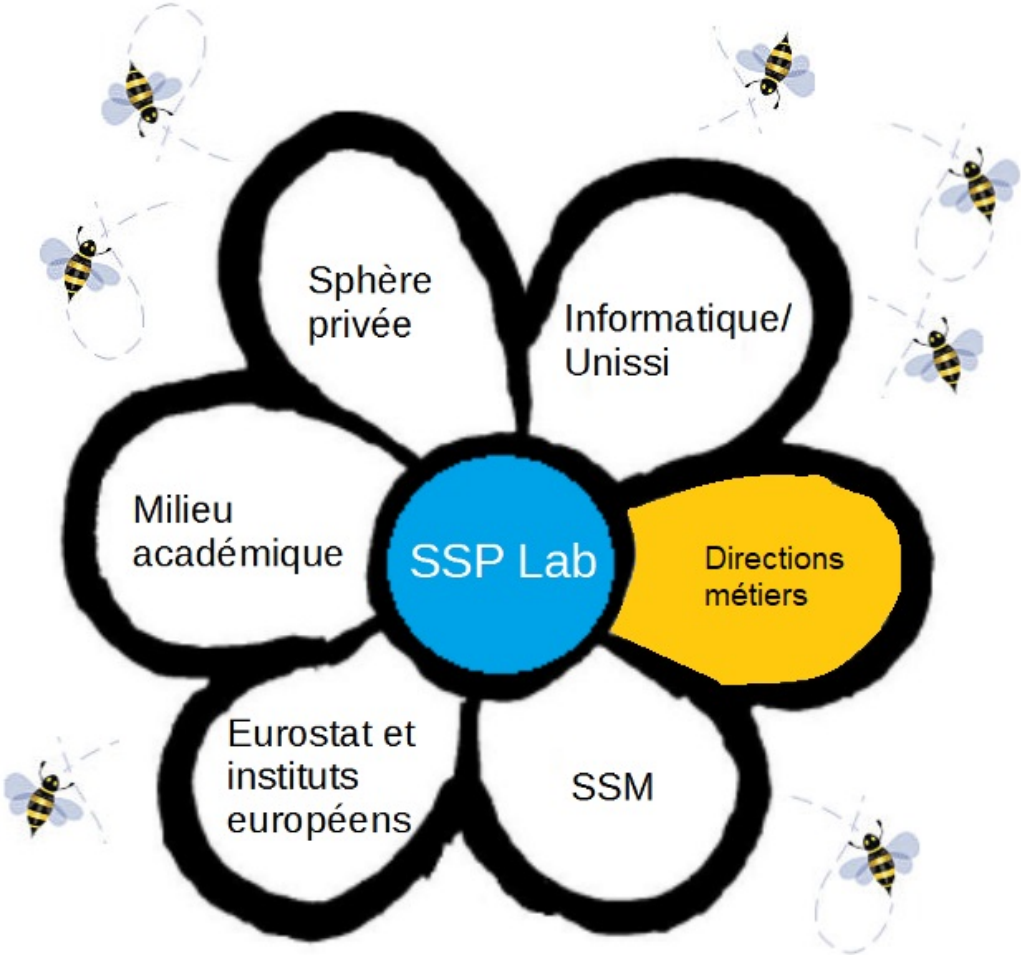
5324 euros de l'heure : outlier ou footballeur ?

Méthodes d'apprentissage non supervisé pour la détection d'anomalies : application au cas de la Déclaration Sociale Nominative

[Sources et Méthodes]

4 juin 2018

Des collaborations pour répondre à des problématiques métiers



La détection d'anomalies : d'un pré-traitement à la finalité

- Objectifs traditionnels de l'analyse statistique : spécification de la distribution statistique d'une variable ou sur l'exploration d'un nuage de points et détection de points anormaux longtemps été considérée comme annexe

- Détection d'*outliers* : attention particulière de la part de spécialistes de nombreux domaines, tels que le secteur bancaire, le secteur informatique et le secteur médical, dans lesquels la sécurité, la détection de fraudes, le décèlement d'une intrusion ou encore le repérage d'une cellule anormale

Un enjeu inhérent à la gestion de bases de données

- Sources multiples concernées : données administratives, données d'enquêtes ...

- Exigences diverses pouvant motiver la détection automatique des anomalies :
  - Données massives
  - Fréquence de données
  - Détection sans *a priori*, sans hypothèses sur la distribution des données

- Difficultés fréquentes face à la détection d'anomalies :
  - Absence d'échantillon d'apprentissage $\Rightarrow$ Se focaliser sur des méthodes d'apprentissage non supervisé
  - Absence d'échantillon d'évaluation $\Rightarrow$ Comment mesurer la performance des algorithmes appliqués ?
  - Nature diverse des anomalies $\Rightarrow$ Combiner plusieurs algorithmes

Un exemple de collaboration autour de cet enjeu

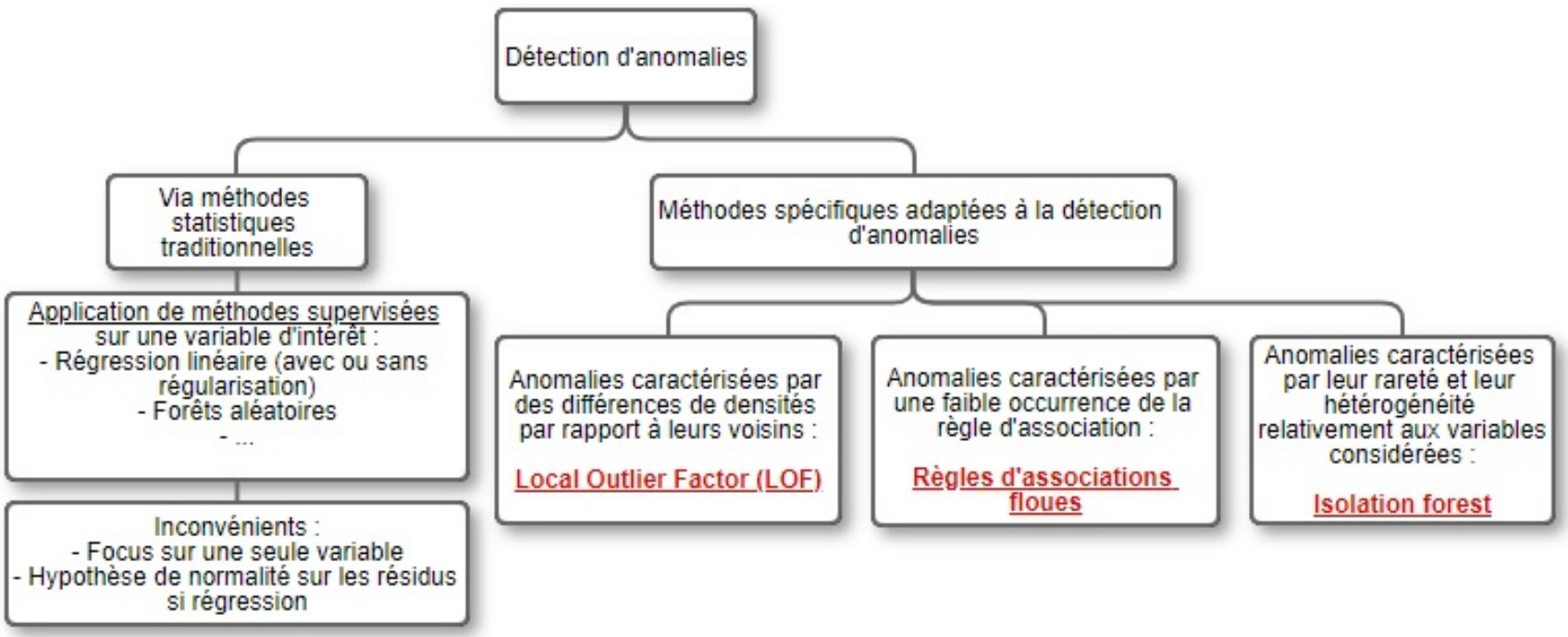
- Le passage à la **Déclaration Sociale Nominative (DSN)** :
  - Compilation de données sociales remplaçant notamment (à terme) les Déclarations Annuelles de Données Sociales (DADS) et des bulletins récapitulatifs de cotisations
  - Gestion de données mensuelles massives

  $\Rightarrow$ test des méthodes de Machine Learning pour **détecter automatiquement** les anomalies

- Une collaboration sous forme de *sprints* / ateliers successifs :
  - Des demi-journées ou journées de travail pour avancer à 2 ou 3 sur la compréhension d'une méthode et sa mise en œuvre
  - Une équipe mi-métier, mi-méthodo : 2 personnes du Dera (Département de l'Emploi et des Revenus d'activité), 1 du DMS et 2 du SSP Lab

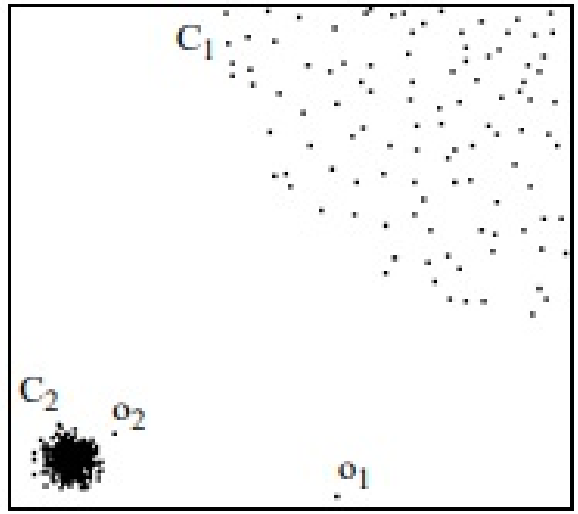
En pratique : comment détecter des anomalies ?

- **Triplet d'intérêt** sur lequel détecter les anomalies :
  (salaire brut horaire, salaire net horaire, nombre d'heures)
- De multiples variables de contrôles :
  Secteur d'activité, PCS, sexe, âge ...
- **Méthodes diverses pour détecter les anomalies** :



1) Le *Local Outlier Factor*

- Algorithme reposant sur le concept d'*outliers* locaux, par opposition aux *outliers* globaux généralement détectés avec une définition courante du terme d'*ouliers*



- Notions proches de celles caractéristiques à l'algorithme DBSCAN mais pas de constitution de *clusters* et pas une approche de densité globale

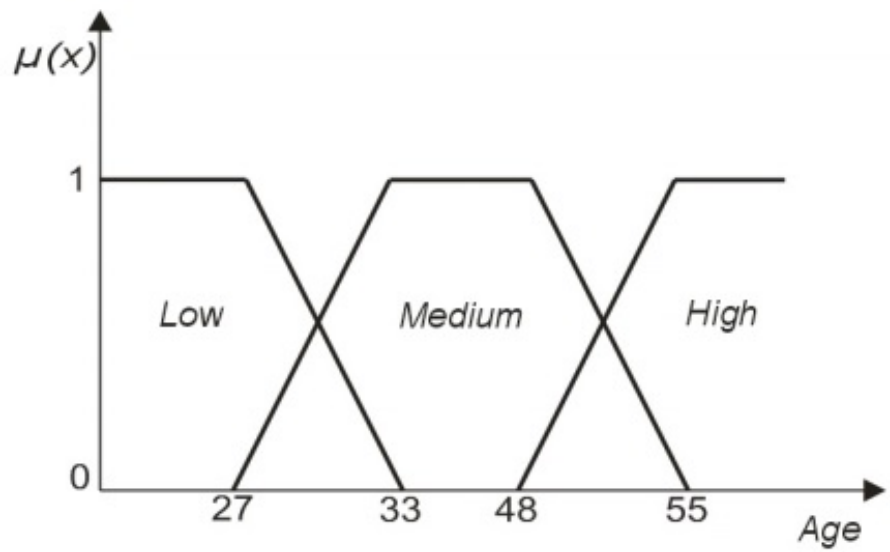
2) Les règles d'association floues

Format d'une règle d'association :
Catégorie sociale, sexe, tranche d'âge $\Rightarrow$ (salaire brut, salaire net, nombre d'heures)

Principe des règles d'association classiques :

- Recherche des *itemsets* fréquents
- Construction de règles à partir des *itemsets* fréquents

Règles d'association floues : Issues de la fuzzification des variables quantitatives



3) Les *isolation forests*

- Agrégation de nombreux arbres, dits *isolation trees*. Chaque arbre utilise un échantillon aléatoire d'observations et, à chaque scission, tire aléatoirement la variable une variable puis la valeur de découpage

- Hypothèse relative aux *outliers* : Les anomalies sont **peu nombreuses** et **différentes**

- Anomalies : observations très vite isolées lors de la construction des arbres alors que les observations normales sont isolées bien plus profondément (scissions de l'arbre très loin de la racine)