



INSTITUT DES
ACTUAIRES

Innovation Open Data

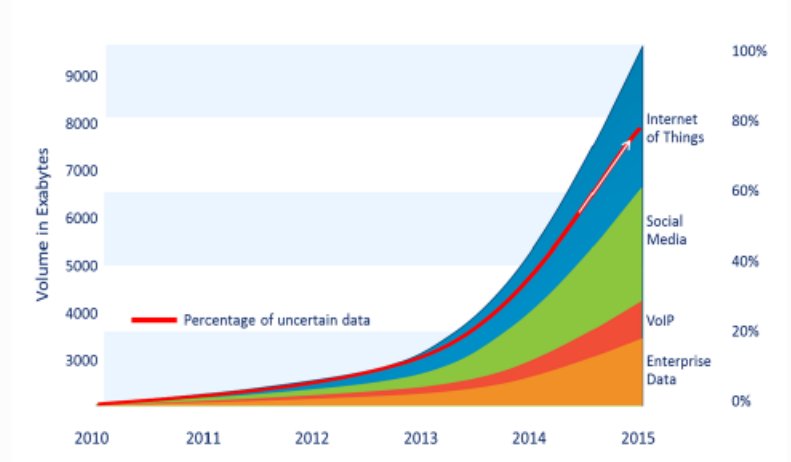
Comment les nouvelles bases de données santé en open data peuvent-elles être source d'inspiration pour l'assurance santé de demain.

Pascale Quennelle & Marc Raymond

Les nouvelles données

Les nouvelles données

D'où proviennent les données ?



data.gouv.fr

Jeux de données à la une Derniers jeux de données

	Base de données publique des médicaments (base officielle) 10 15		Cartographie des bases de données publiques en santé 1 10
	FINESX Extraction du Fichier des établissements 8 48		Fréquence des séjours selon codes diagnostics principaux ou actes classant 0 0
	Indicateur Avancé Sanitaire IAS® - SYNDROME GRIPPAL 7 59		Indicateurs du thème « Infections Associées aux Soins » 1 2
	Les indicateurs relatifs aux infections nosocomiales dans les établissements de santé - 2011 1 2		Open DAMIR : base complète sur les dépenses d'assurance maladie inter régimes 7 5

Les données santé

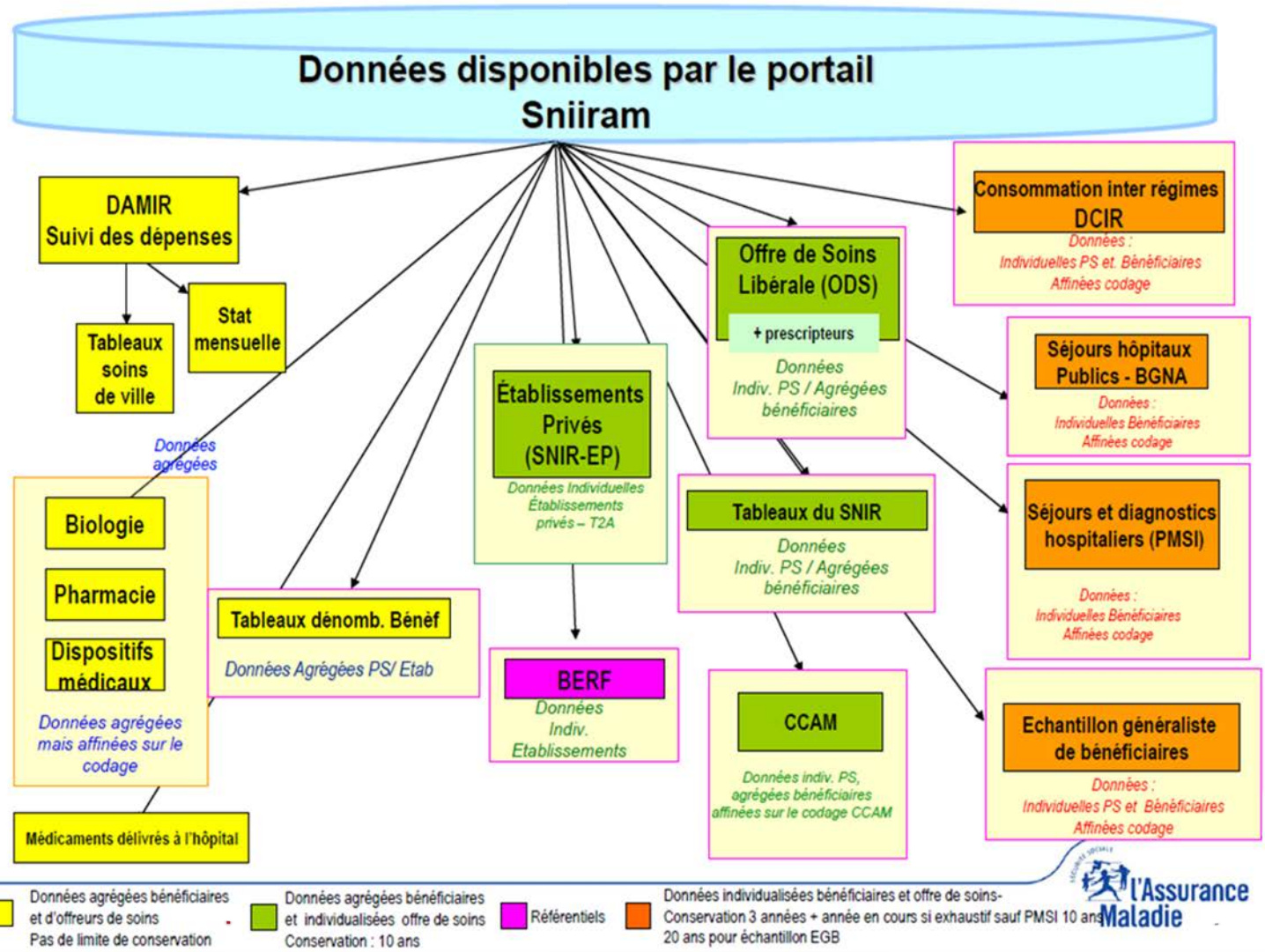
Les données Santé : à quoi servent-elles ?

« Les données santé sont destinées à la recherche, aux statistiques, à l'aide à la décision et à l'information du public. Elles permettent de mieux connaître le système de santé, pour mieux l'utiliser, pour en débattre démocratiquement et pour l'améliorer. »

Rapport sur la gouvernance et l'utilisation des données de santé 2013



Diverses bases de données disponibles



Article 47 de la loi Santé

Les finalités du Système National des Données de Santé (SNDS) :

- 1) L'information sur la santé, les soins et la prise en charge médico-sociale,
- 2) La définition, la mise en œuvre et l'évaluation des politiques de santé et de protection sociale,
- 3) La connaissance des dépenses de santé, des dépenses de l'assurance maladie et des dépenses médico-sociales,
- 4) L'information des professionnels, structures et établissements de santé ou médico sociaux sur leur activité,
- 5) La surveillance, la veille et la sécurité sanitaires,
- 6) La recherche, aux études et à l'innovation dans les domaines de la santé et de la prise en charge médico-sociale.

Il existe 3 sortes d'accès aux données de santé :

- En « routine » : Ministère de la santé, CNAMTS, DREES, ... (fixé par décret en conseil d'État),
- Sur demande pour des recherches, études ou évaluations,
- En open-data.

Article 47 de la loi Santé

Cas de l'open-data.

« Les données du système national des données de santé qui font l'objet d'une mise à la disposition du public sont traitées pour prendre la forme de statistiques agrégées ou de données individuelles constituées de telle sorte que l'identification directe ou indirecte des personnes concernées y est impossible. »

Exemples :

- SCORE-Santé
 - Éco-Santé
 - Le cube de la DREES
- DAMIR

Analyse des données DAMIR

Description de la base DAMIR

- Depuis le 26 Janvier 2015, l'assurance maladie met à disposition des bases de données sur les prestations des différents organismes d'assurance maladie, en particulier la base Open DAMIR qui est une base sur les **dépenses d'assurance maladie inter-régimes**.
- Ce jeu de données concerne l'ensemble des prestations prises en charge par l'Assurance Maladie obligatoire. Ainsi, l'ensemble des dépenses de remboursement, tous régimes confondus, de l'Assurance Maladie est couvert par cette base de données, **à l'exception d'une grande majorité des prestations hospitalières du secteur public**.
- Cette base a été établie tout en **préservant l'anonymat** des professionnels de santé et des bénéficiaires des soins.

Analyse de la base DAMIR

Base Open DAMIR

3 Acteurs : patient, professionnel de santé et OAM.

Actes associés à un remboursement décrits par :

- La nature de la prestation...
- La nature de l'assurance,
- Le complément d'acte.

Patient, bénéficiaire de l'acte connu par :

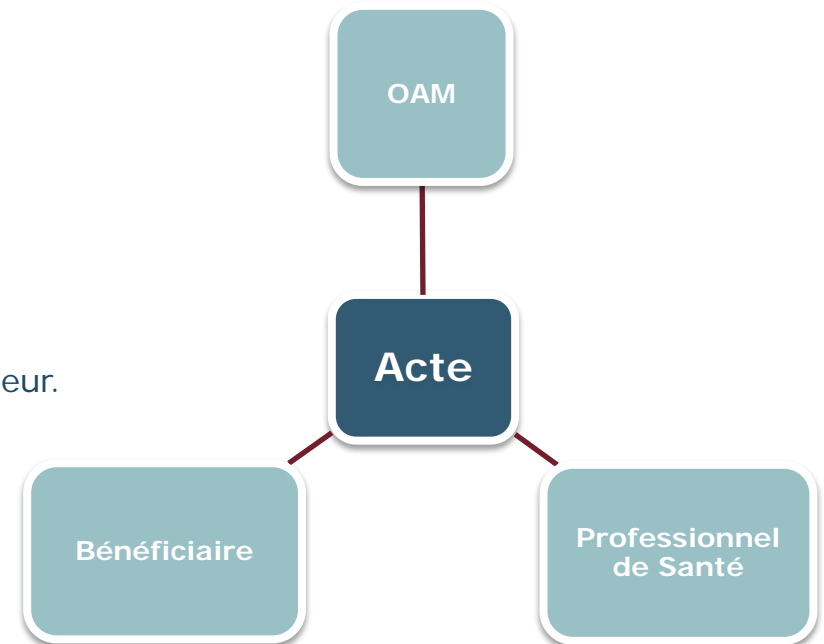
- Son âge,
- Son sexe,
- S'il est bénéficiaire de la CMU ou non,
- Et son lieu de résidence.

Professionnels de santé classés en exécutant et prescripteur.

Le professionnel de santé est décrit par :

- Sa catégorie,
- Sa spécialité,
- Son statut juridique et sa localisation.

L'organisme d'assurance maladie (OAM)



Base Open DAMIR

Axe spatial

Information sur plusieurs lieux tels que :

- La localisation de la résidence de l'assuré
- La localisation de la caisse locale d'affiliation mais aussi de celles des professionnels de santé,
- D'autres informations spatiales liées à un régime de remboursement différent.

Axe temporel

Information sur des axes temporels : La date de soin, la date de remboursement, le moment de l'acte.

Axe protocole médical

L'association prescripteur/exécutant donne une information sur la séquence.

L'information sur le parcours de soin peut aussi donner une indication.

Lien prescripteur et prestation

L'acte est prescrit par le prescripteur qui n'est pas forcément le professionnel de santé exécutant.

Analyse de la base DAMIR

Source : Base entière

55 variables, 800 types d'actes, 1 année = 220 millions de lignes de prestations

1/ Élimination des variables insuffisamment renseignées : perte d'information éventuelle sur l'impact d'une variable

Pre-processing : 55 variables, 800 types d'actes, 1 année = 220 millions de lignes de prestations

2/ Sélection des variables pertinentes : choix pour faciliter l'analyse

Connaissance à dire d'expert

3/ Retraitement des données : correction des données incomplètes

Traitement des modalités inconnues des variables sélectionnées. Deux cas :

- Répartition de ces valeurs sur les autres modalités (seuil de 5%).
- Sinon, conservation de la valeur « Inconnu ».

4/ Identification des valeurs aberrantes : complexe compte tenu des valeurs traitées

Base d'étude

12 variables restantes dans l'analyse, 11 groupements de frais de soins

Analyse de la base DAMIR

Traitement de la problématique liée aux 800 types d'actes et de celle liée aux 55 variables de la base

Méthode de Machine Learning inadaptée

Connaissances métier à dire d'expert pour choisir :

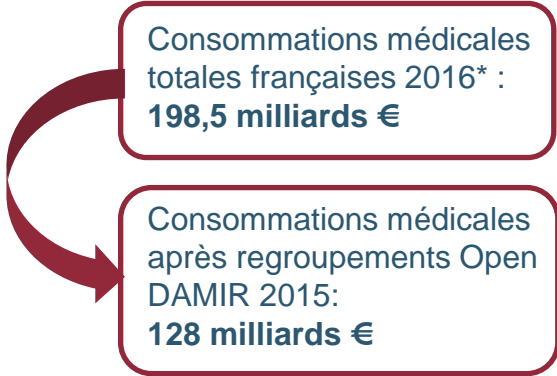
- Les regroupements de frais de soins qui ont du sens,
- Les variables à sélectionner pour notre analyse.

Chaque type d'actes appartient à l'un des regroupements de frais de soins suivants :

- Médecins généralistes
- Optique
- Audition
- Analyses
- Indemnités journalières
- Autres
- Médecins spécialistes
- Dentaire
- Médicaments
- Auxiliaires
- Transport

De plus, nous avons sélectionné 12 variables dans notre analyse :

- Année et mois de traitement
- Nature de prestation
- Libellé de prestation
- Catégorie d'activité de l'exécutant
- Catégorie d'activité du prescripteur
- Taux de remboursement
- Âge du bénéficiaire
- Sexe du bénéficiaire
- Zone géographique du bénéficiaire
- Montant de la dépense de la prestation
- Montant du dépassement de la prestation
- Montant du remboursement de la prestation



*Ministère des Affaires sociales et de la Santé

Utilisation d'une typologie en deux niveaux	
Typologie 1	Typologie 2
Santé publique et épidémiologie	<ul style="list-style-type: none"> • Etat de santé • Épidémiologie • Environnement • Habitudes de vie • Inégalités de santé • Autres
Offre de soins	<ul style="list-style-type: none"> • Infrastructures • Services proposés • Personnel • Tarifs, honoraires • Infrastructure de recherche • Informations médicales • Autres
Consommation de soins	<ul style="list-style-type: none"> • Activité des établissements de santé • Consultations • Biologie • Transports • Médicaments et dispositifs • Indemnités et accidents du travail • Activité des établissements médicaux-sociaux • Dépenses
Performance et opérations	<ul style="list-style-type: none"> • Performance financière • Performance opérationnelle • Qualité

Traitement de la problématique liée aux données manquantes

Analyse des 55 variables en distinguant :

- **les variables qualitatives,**
- **les variables quantitatives.**

→ La procédure sous SAS a révélé que sur les 55 variables et les 7 années, seules 3 variables contenaient des observations manquantes.

Ces 3 variables :

- Sont remplacées par d'autres variables (intérêt des variables préfiltrées);
 - Ou présentent peu d'intérêt
- **Ainsi la problématique des valeurs manquantes a été contournée.**

Traitement de la problématique liée aux valeurs inconnues

Répartition des observations sur l'ensemble des autres modalités de la variables. Traitement réalisé par une procédure sous SAS.

1ère étape : clé d'identification

Par la combinaison des modalités des variables qualitatives comme :

- résidence,
- sexe,
- année
- mois de traitement,
- nature de la prestation,
- taux de remboursement,
- catégorie de l'exécutant

2ème étape : coefficient de répartition des variables quantitatives

Le montant des variables quantitatives comme

- le montant de la dépense de la prestation,
- le montant du dépassement,
- le montant versé et remboursé

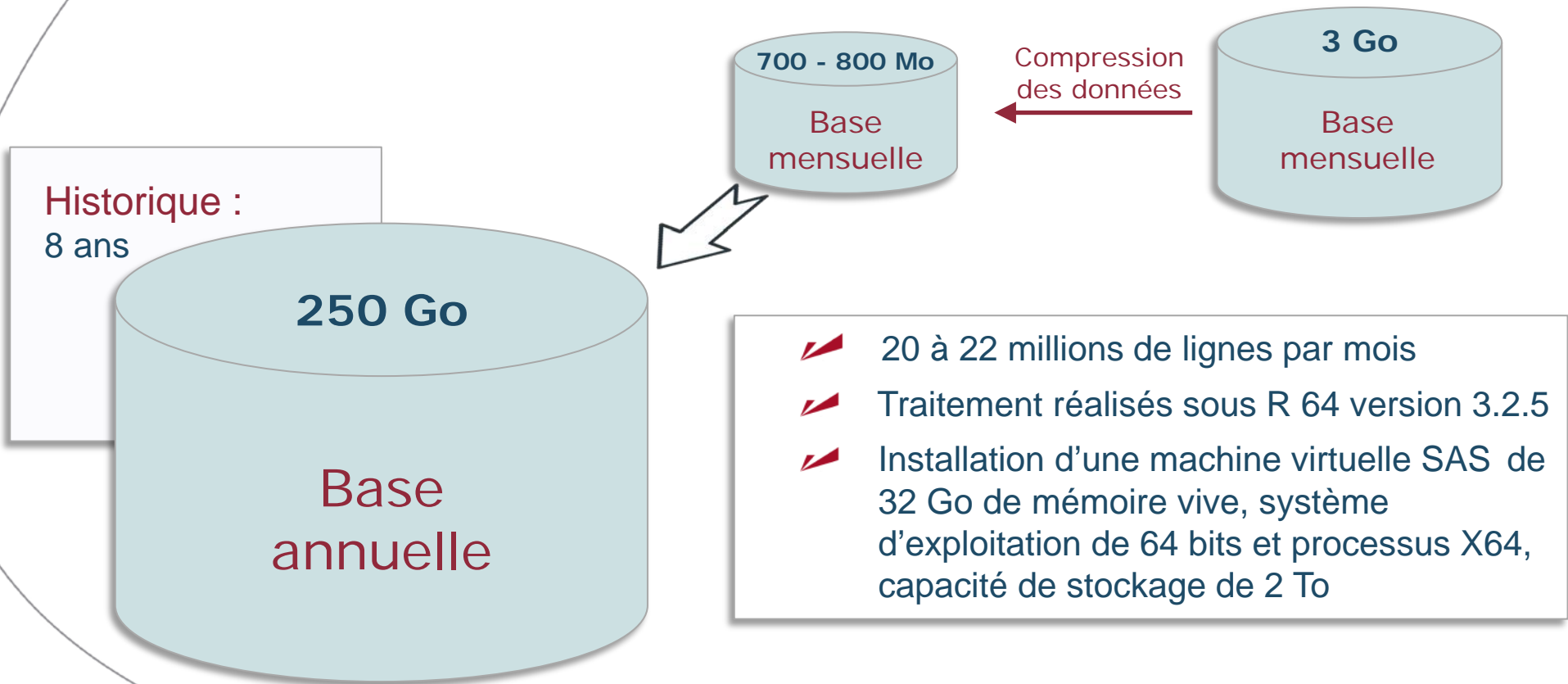
sont répartis sur les différentes modalités de la variable : « méthode imputation ».

Le coefficient pour la modalité i de la variable ayant n modalités :

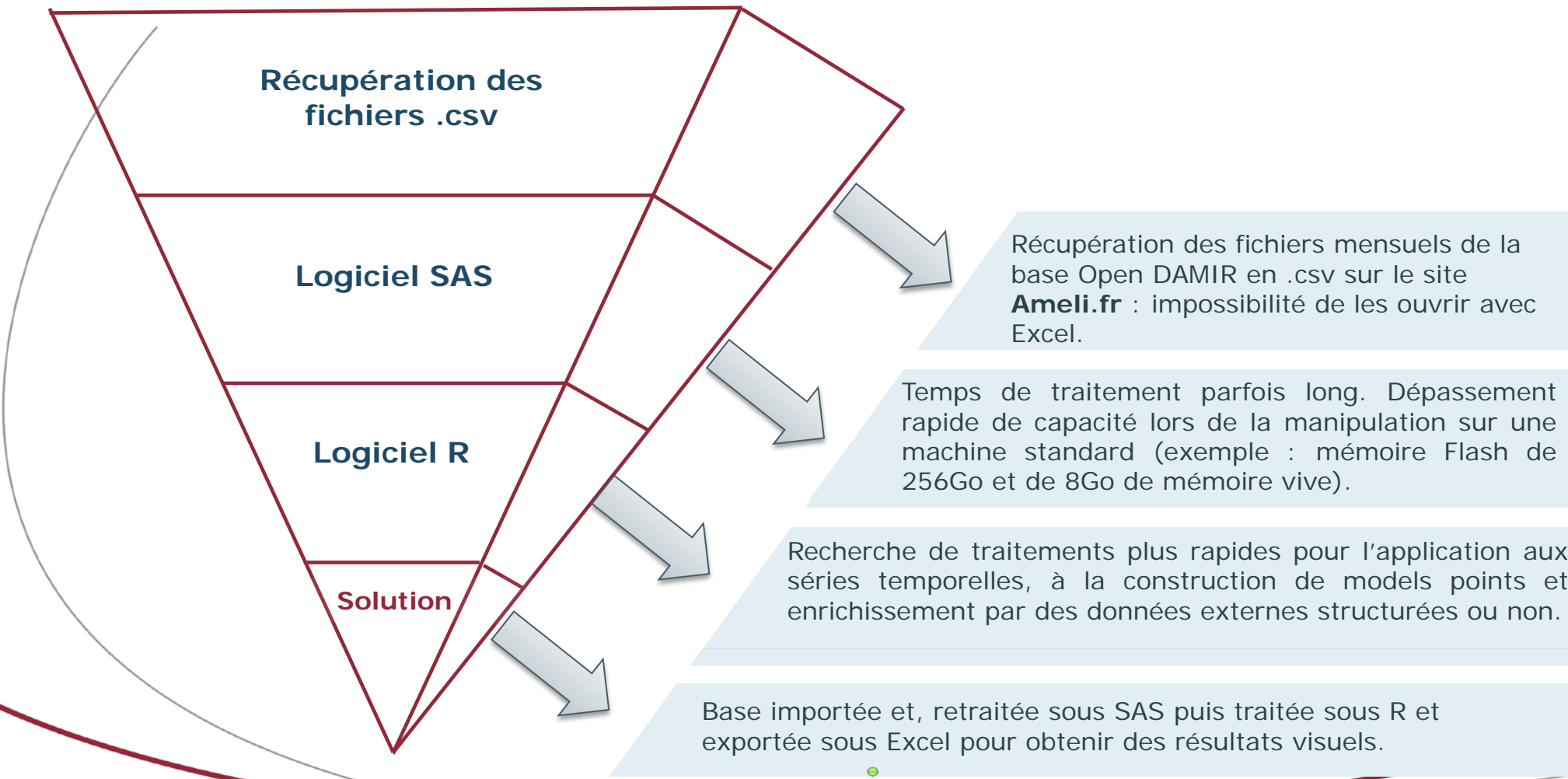
$$\text{Coef}(i, \text{dépense}) = \frac{M(i, \text{dépense})}{\sum_{k=1}^n M(k, \text{dépense})}$$

Note : $\sum \text{Coef} = 1$

Traitement de la problématique liée au volume de la base



Analyse de la base DAMIR



Analyse de la base DAMIR

Limites et réserves

Les données ont été agrégées afin qu'il ne soit plus possible d'identifier un individu : comprendre chaque ligne comme étant la somme des actes et des montants associés à toutes les variables catégorielles (lieu, âge, type d'acte, etc.).



Axe de recherche

Le nombre d'actes par prestation n'est pas la variable adéquate pour mesurer le poids de la ligne.
La nomenclature des actes médicaux CCAM évolue dans le temps



Axe de développement



Limites et réserves

Étudier l'homogénéité des données issues de différentes sources (différents systèmes d'information qui gèrent le RO)

Solution
complexe ?

**Axe de
recherche**

En l'absence de données exactes sur le nombres de personnes couvertes par le RO (tout régimes confondus), il conviendrait de déterminer le bon périmètre de l'extraction Open DAMIR.

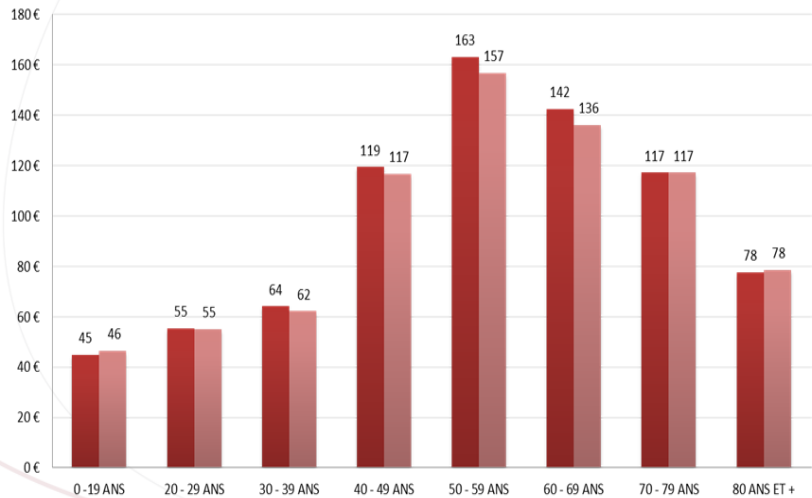
Améliorations
possibles

**Axe de
développement**

Perspectives de dépenses de santé

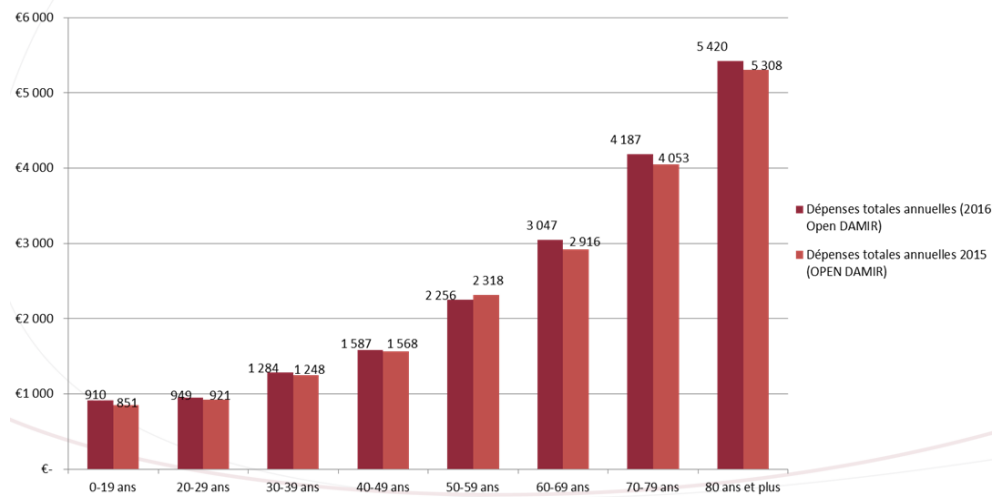
Evolution par âge optique

Dépenses annuelles moyennes



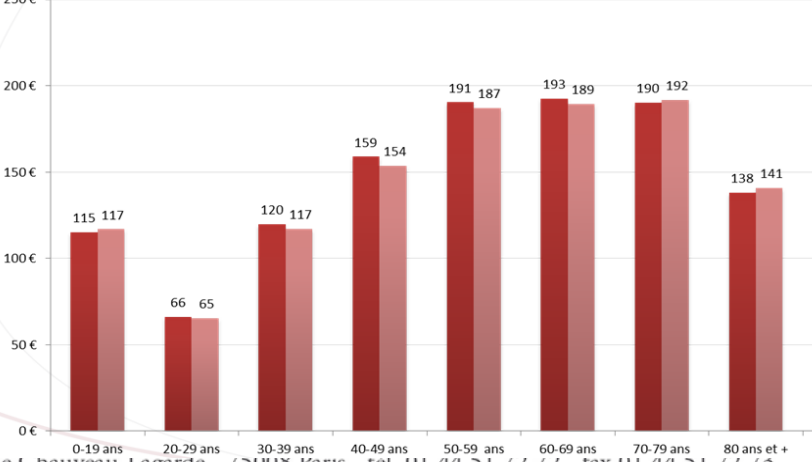
Evolution par âge toutes prestations confondues

Dépenses annuelles moyennes



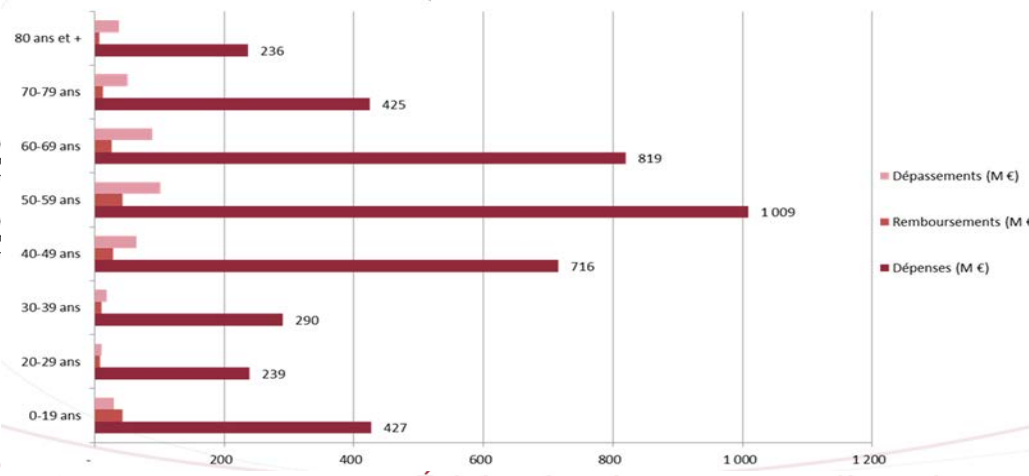
Evolution par âge traitements dentaire

Dépenses annuelles moyennes

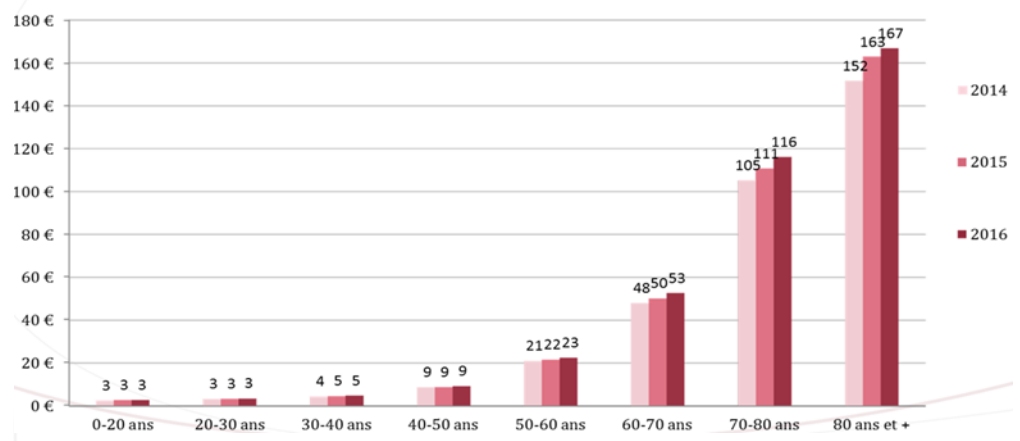


Evolution par âge verres en 2016

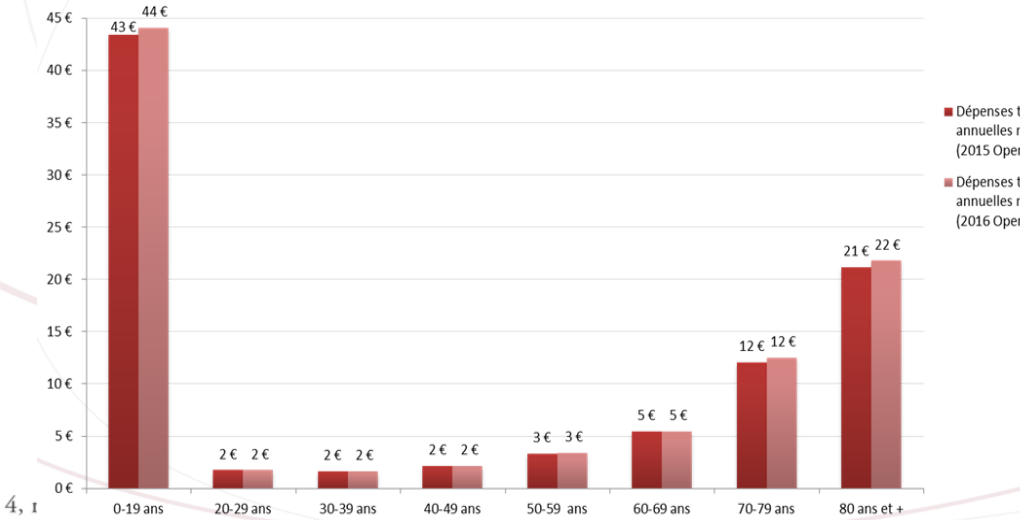
Dépenses annuelles



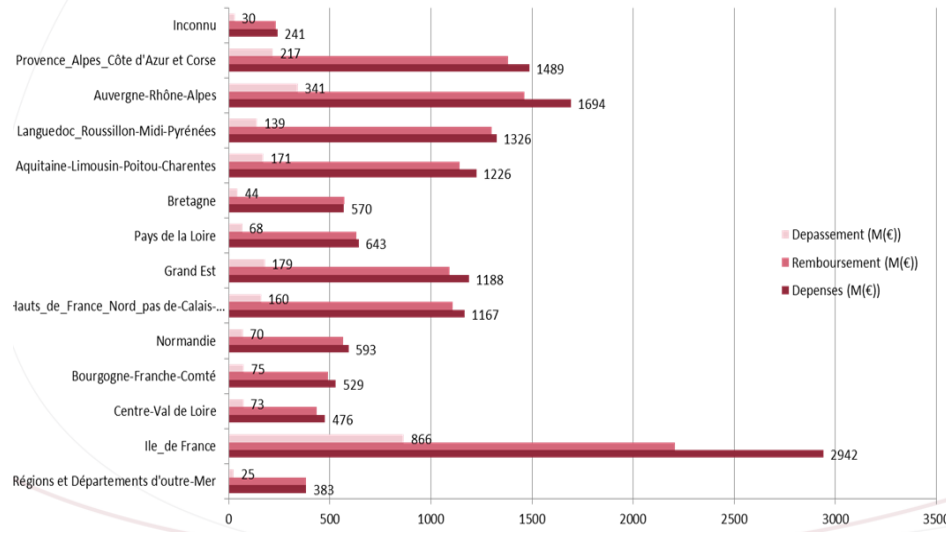
Evolution des dépenses annuelles moyennes en audition par segment d'âges de 2014 à 2016 en France



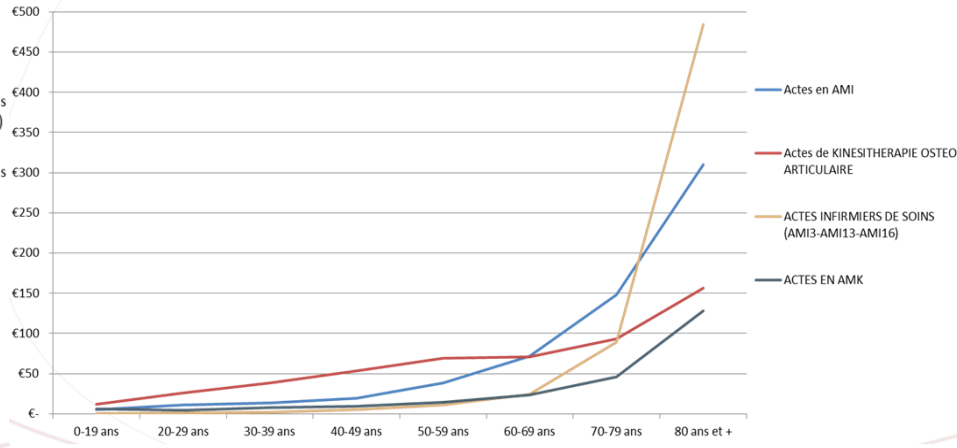
Evolution par âge actes des orthophonistes
Dépenses annuelles moyennes



Dépassements d'honoraires des médecins spécialistes par régions 2016



Evolution par âge actes auxiliaires médicaux
Dépenses annuelles moyennes

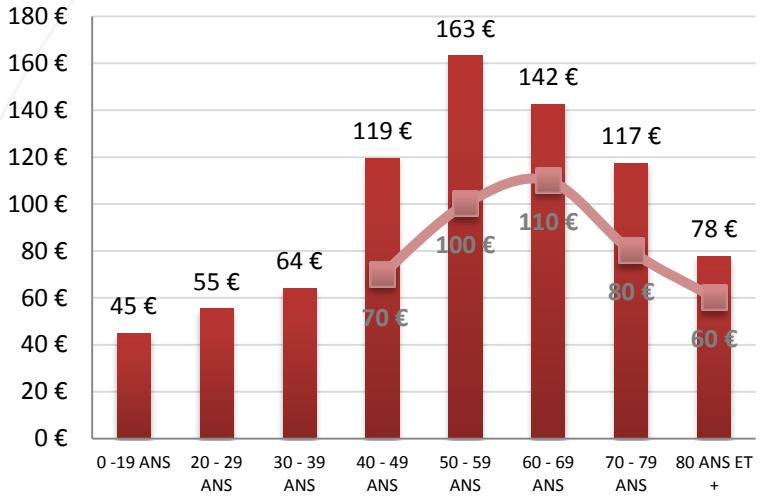


Éclairer les risques, tracer l'avenir

Construction de référentiel pour les dépenses de santé Selon différents champs de vision

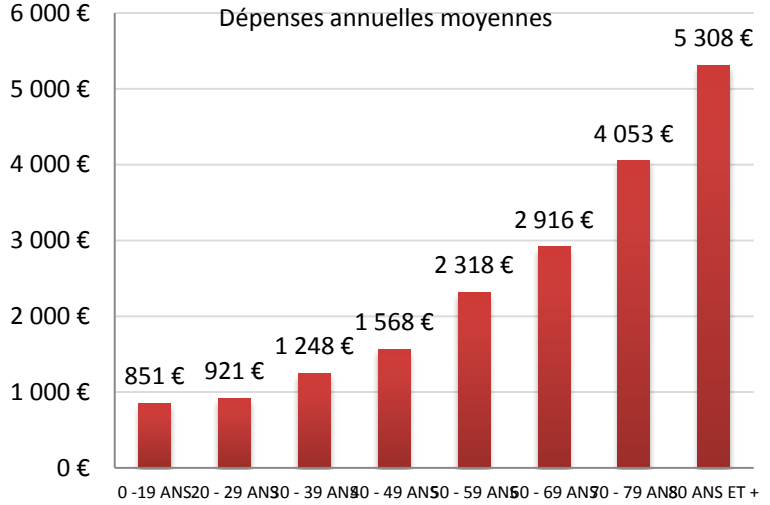
Evolution par âge optique

Dépenses annuelles moyennes



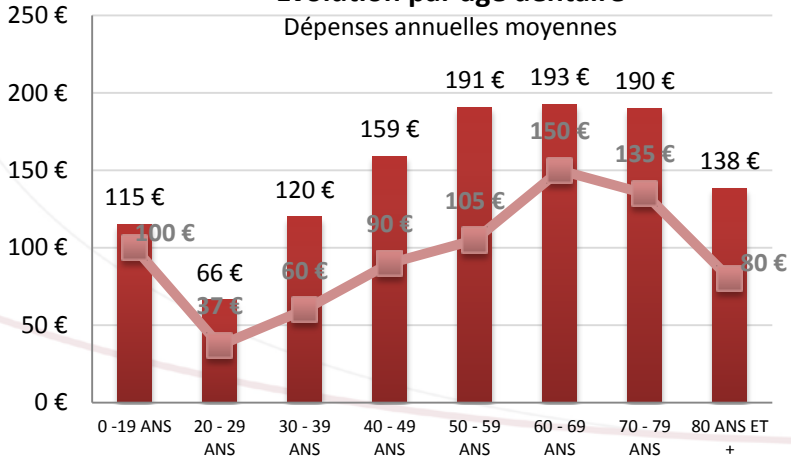
Evolution par âge toutes prestations confondues

Dépenses annuelles moyennes



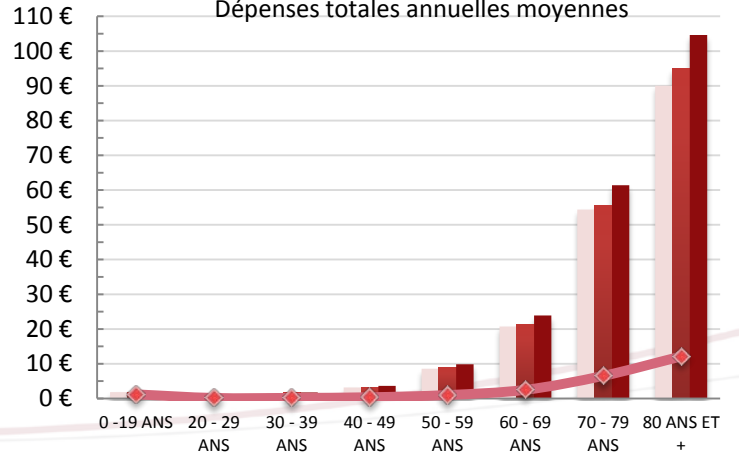
Evolution par âge dentaire

Dépenses annuelles moyennes



Evolution par âge appareils électroniques de surdité

Dépenses totales annuelles moyennes



Prédiction Méthode des Séries Temporelles

Prédiction selon la méthode des Séries Temporelles

L'objectif est l'appréhension des valeurs futures x_{T+1} , x_{T+2} à partir de l'observation des valeurs connues x_1, x_2, \dots, x_T



Modélisation et mise en place d'intervalle de prédiction valeur prévue & niveau de certitude.

Décomposition additive de la Serie :

$$X_t = M_t + S_t + Y_t$$

- M_t , tendance déterministe, comportement long terme
- S_t , tendance saisonnière déterministe
- Y_t , composante irrégulière et aléatoire

Prédiction selon la méthode des Séries Temporelles

Exemple d'application :

L'étude porte sur la prédiction du montant de la dépense de consommation en appareils auditifs pour les séniors.

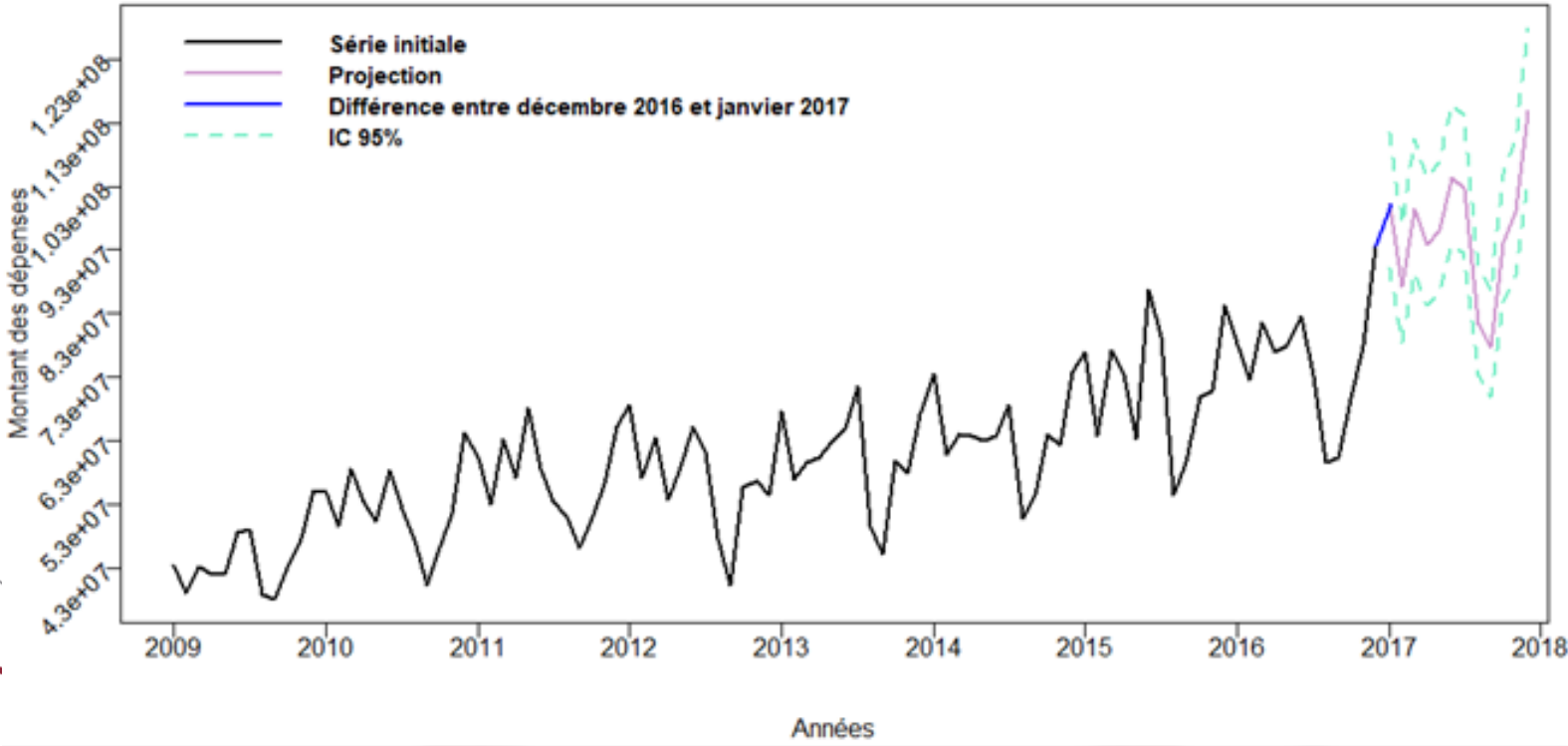


Prédiction selon la méthode des Séries Temporelles

Exemple d'application :

L'étude porte sur la prédiction du montant de la dépense de consommation en appareils auditifs pour les séniors.

Prévisions des montants de dépenses en appareils auditifs pour les personnes de 60 ans et plus



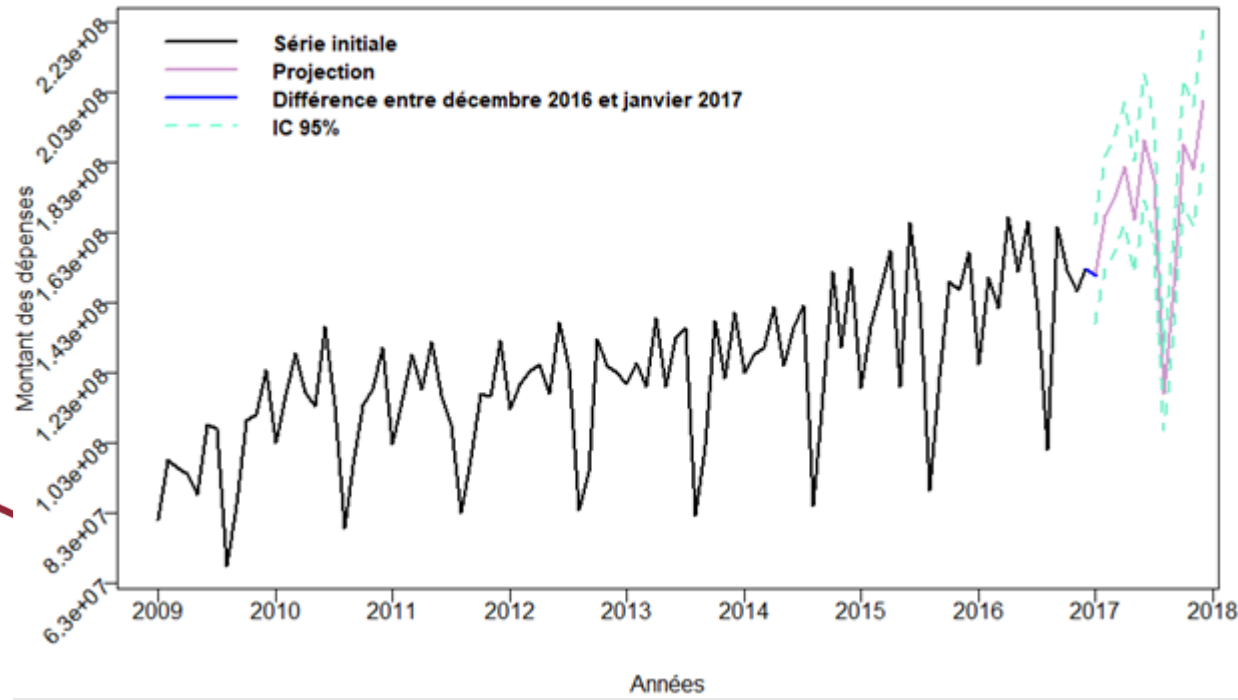
Prédiction selon la méthode des Séries Temporelles

Exemple d'application :

L'étude porte sur la prédiction du montant de la dépense de consommation en forfait journalier hospitalier en appareils auditifs pour les séniors.

Le lissage par moyenne mobile est celui qui minimise la SSE.

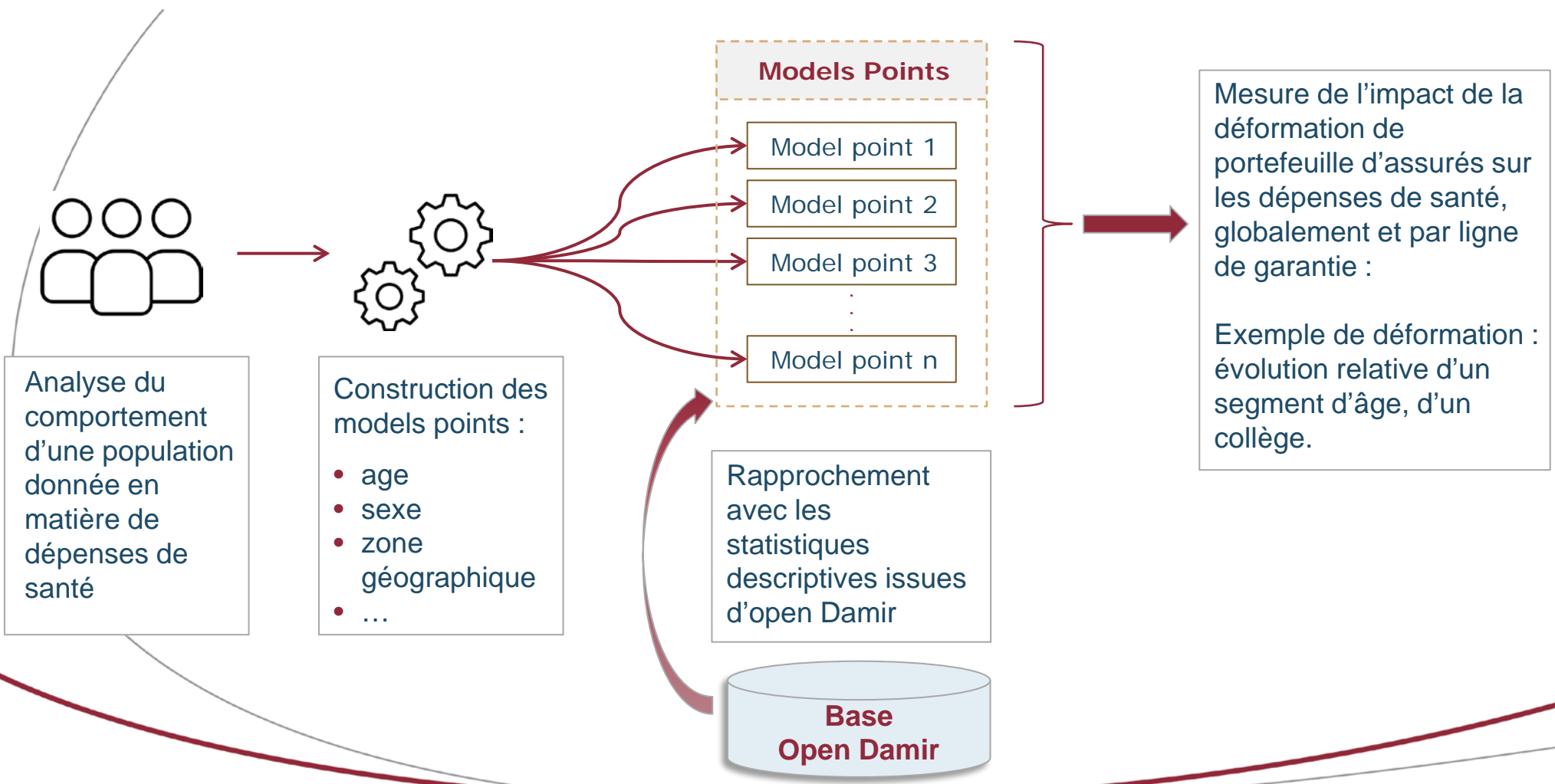
Prévisions des montants de dépenses en Forfait Journalier Hospitalier



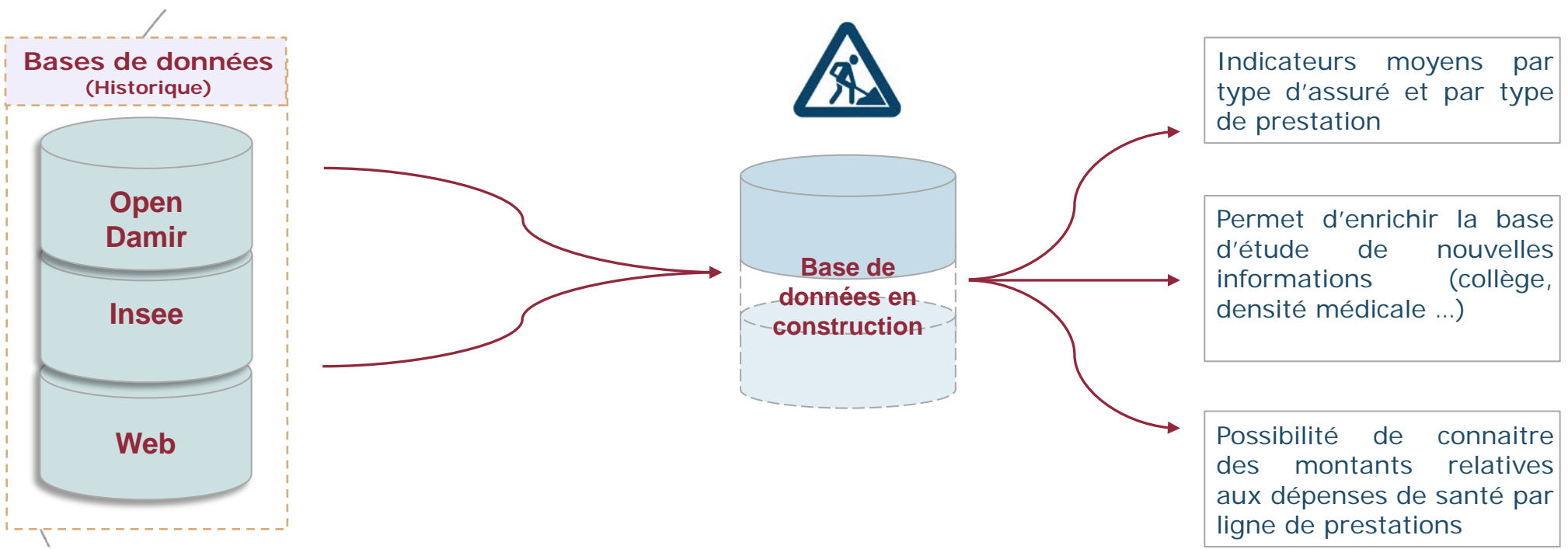
Méthode de projection	SSE
Moyennes Mobiles	0, 081823
LES	0, 117557
Holt	0, 116034
HWSA	0, 098358

Construction de benchmark et Enrichissement de la base

Construction de benchmark



Enrichissement par rapprochement avec d'autres bases

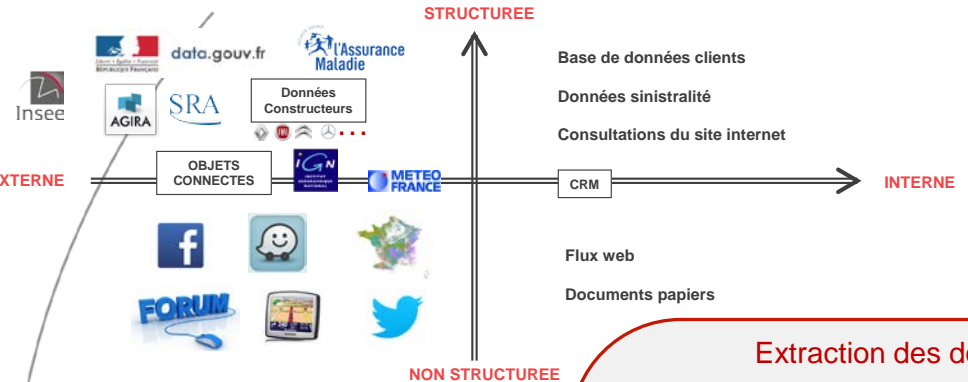


Hypothèse : les écarts entre la population Insee et la population Damir sont suffisamment faibles pour permettre le rapprochement des données issues des deux bases.

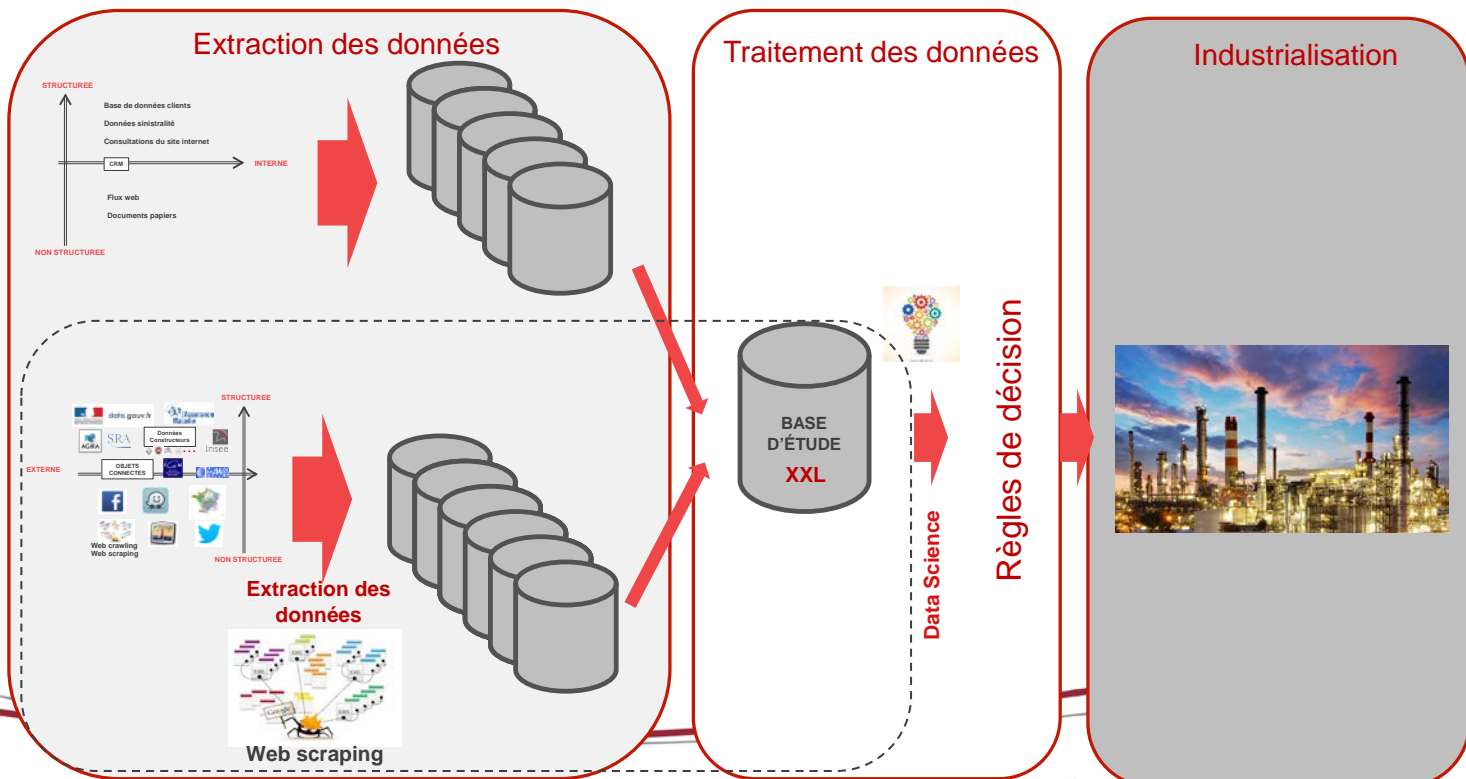
L'anonymisation de la base ne permet pas de mesurer le nombre d'actes de prestations ou le nombre de bénéficiaires. En revanche un rapprochement est possible avec le nombre de personnes couvertes par la sécurité sociale (population Open Damir) en utilisant les données Insee sur l'ensemble de l'historique.

Démarche générique

1. Les sources de données

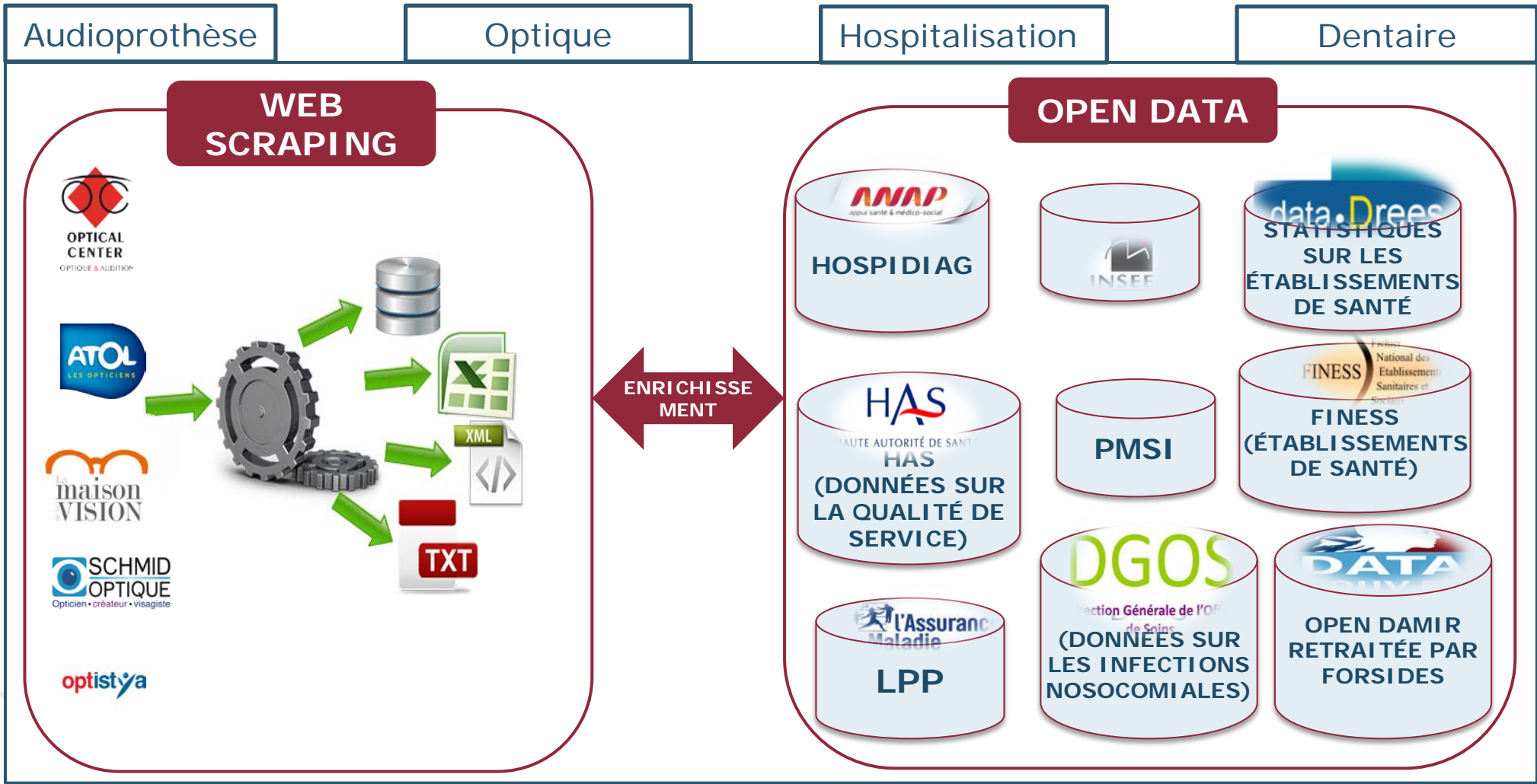


2. La démarche




Eclairer les risques, tracer l'avenir

Enrichissement par des sources de données externes



Le web scrapping

Les librairies



- Jsoup
- Selenium
- System.net



- HTMLAgilityPack
- WebBrowser
- System.Net







beautiful soup

Les outils



L'approche

- PHASE 1 Identification de la source de données
- PHASE 2 Chargement de la page Web
- PHASE 3 Interaction Web (clic, sélection, exécution de scripts ...)
- PHASE 4 Identification de la donnée à récupérer
- PHASE 5 Extraction de la donnée

Illustration : l'extraction de commentaires sur le Web

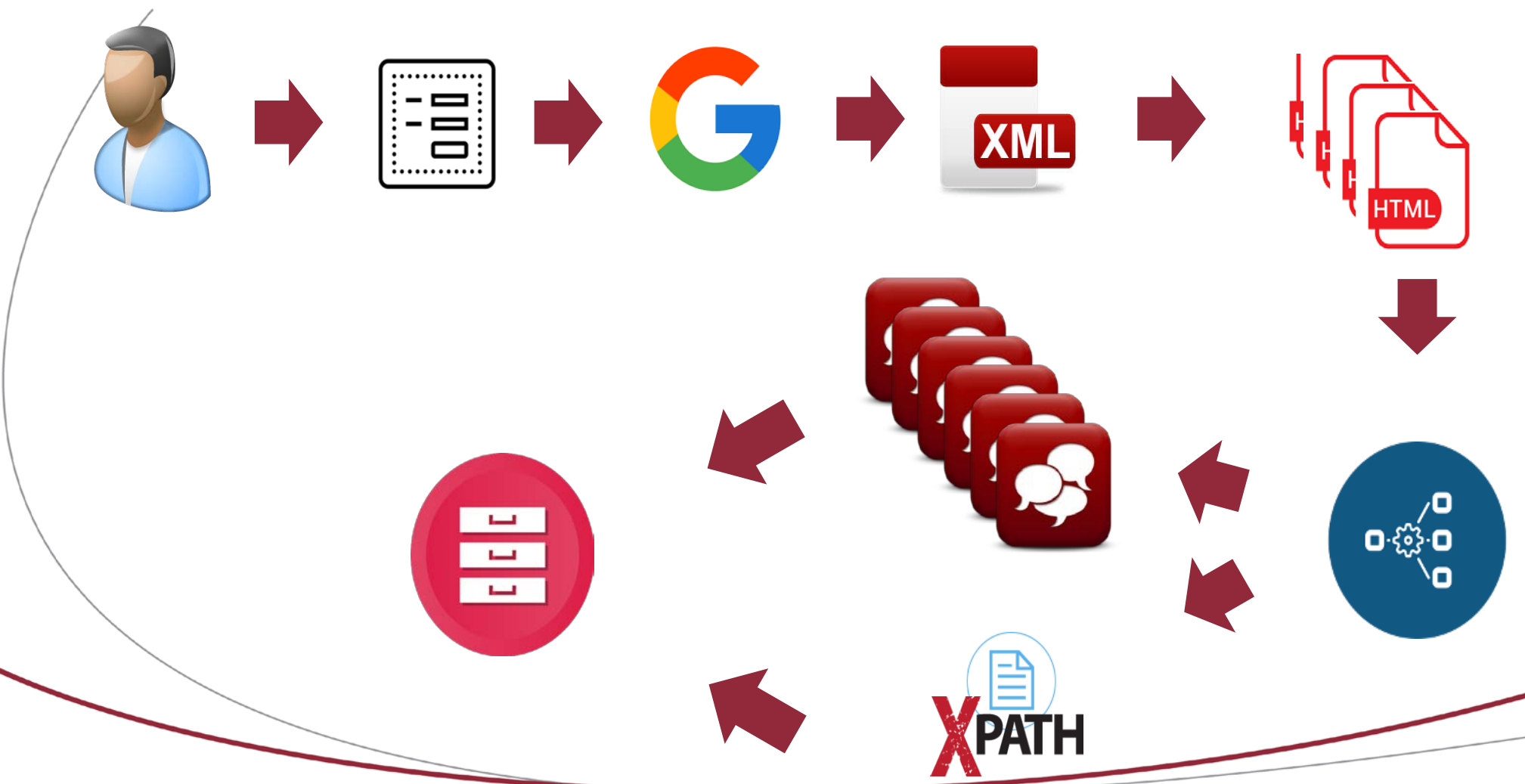


Illustration : l'extraction de données PDF

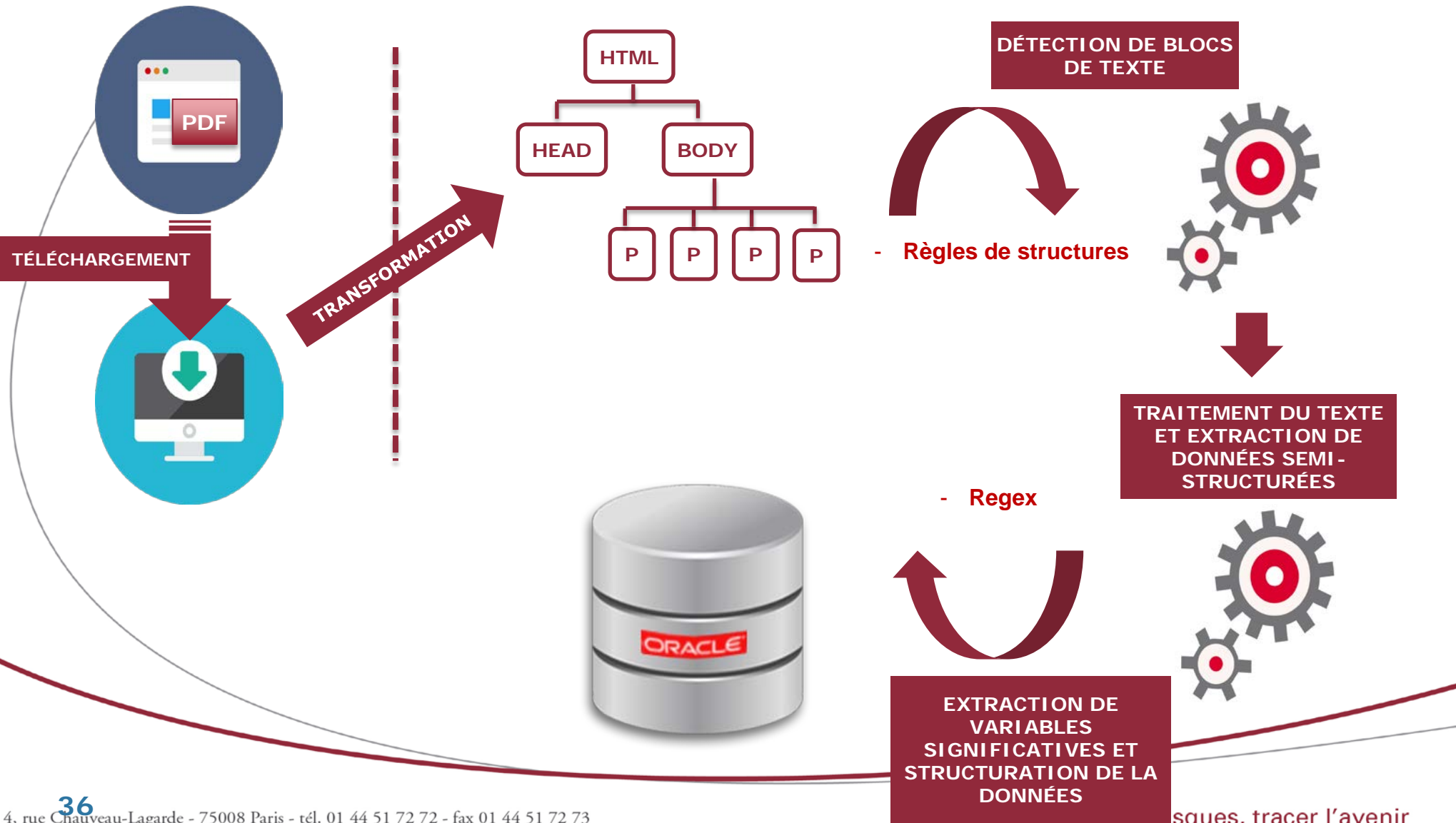
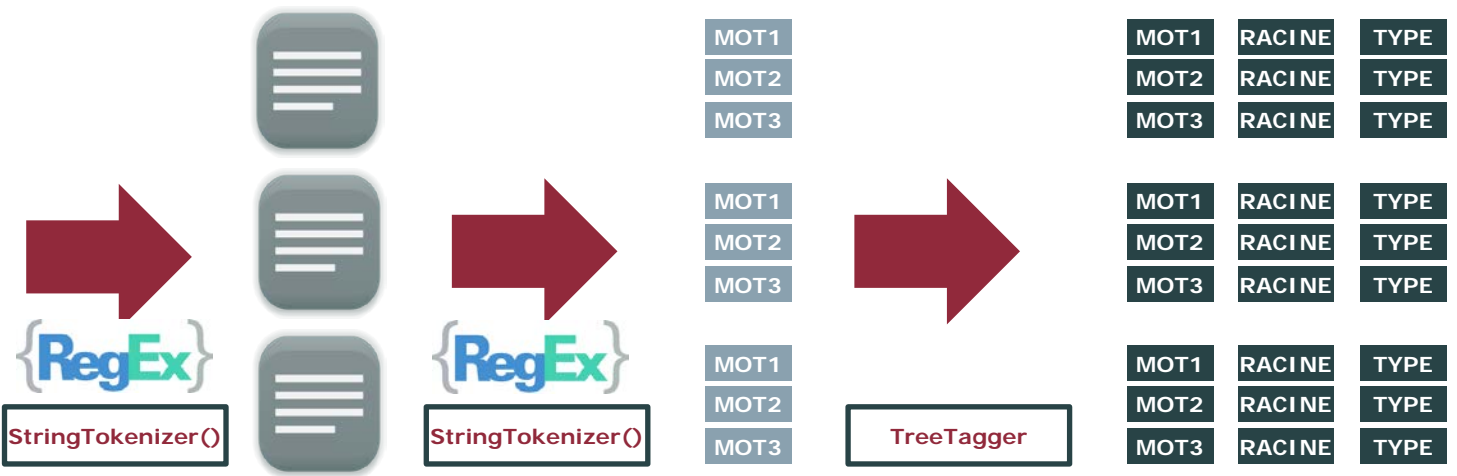


Illustration : le traitement de données textuelles



MOT1	RACINE	TYPE
MOT3	RACINE	TYPE
MOT1	MOT3	

MOT1	RACINE	TYPE
MOT3	RACINE	TYPE
MOT1	MOT3	

MOT1	RACINE	TYPE
MOT3	RACINE	TYPE
MOT1	MOT3	

MOT1	RACINE	TYPE
MOT3	RACINE	TYPE

MOT1	RACINE	TYPE
MOT3	RACINE	TYPE

MOT1	RACINE	TYPE
MOT3	RACINE	TYPE



Demain

Open data une opportunité pour l'assurance santé

Les données au service de l'innovation



Des évolutions produits en cours

De nouvelles données

Intégration des données des bracelets connectés au sein d'applications qui rassemblent des données de bien-être issues d'accessoires connectés et d'applications, permettant aux consommateurs de suivre et de partager :

- les distances parcourues,
- les calories brûlées,
- la fréquence cardiaque
- ...

Pour de nouveaux services

En échange de leurs données témoignant d'un mode de vie sain, les clients reçoivent des avantages financiers

Mais aussi :

- incitation à mener une vie plus saine,
- prévention,
- conseils

Des perspectives aussi en

Détection de la fraude

Détermination du tarif

Anticipation des évolutions

Gestion de la relation client