Observatoire participatif de la qualité de l'air

Méthodes et challenges pour l'analyse contextuelle de données de capteurs mobiles

Karine Zeitouni, Laboratoire DAVID – UVSQ

Journée thématique « Capteurs pour le suivi de la qualité de l'air »

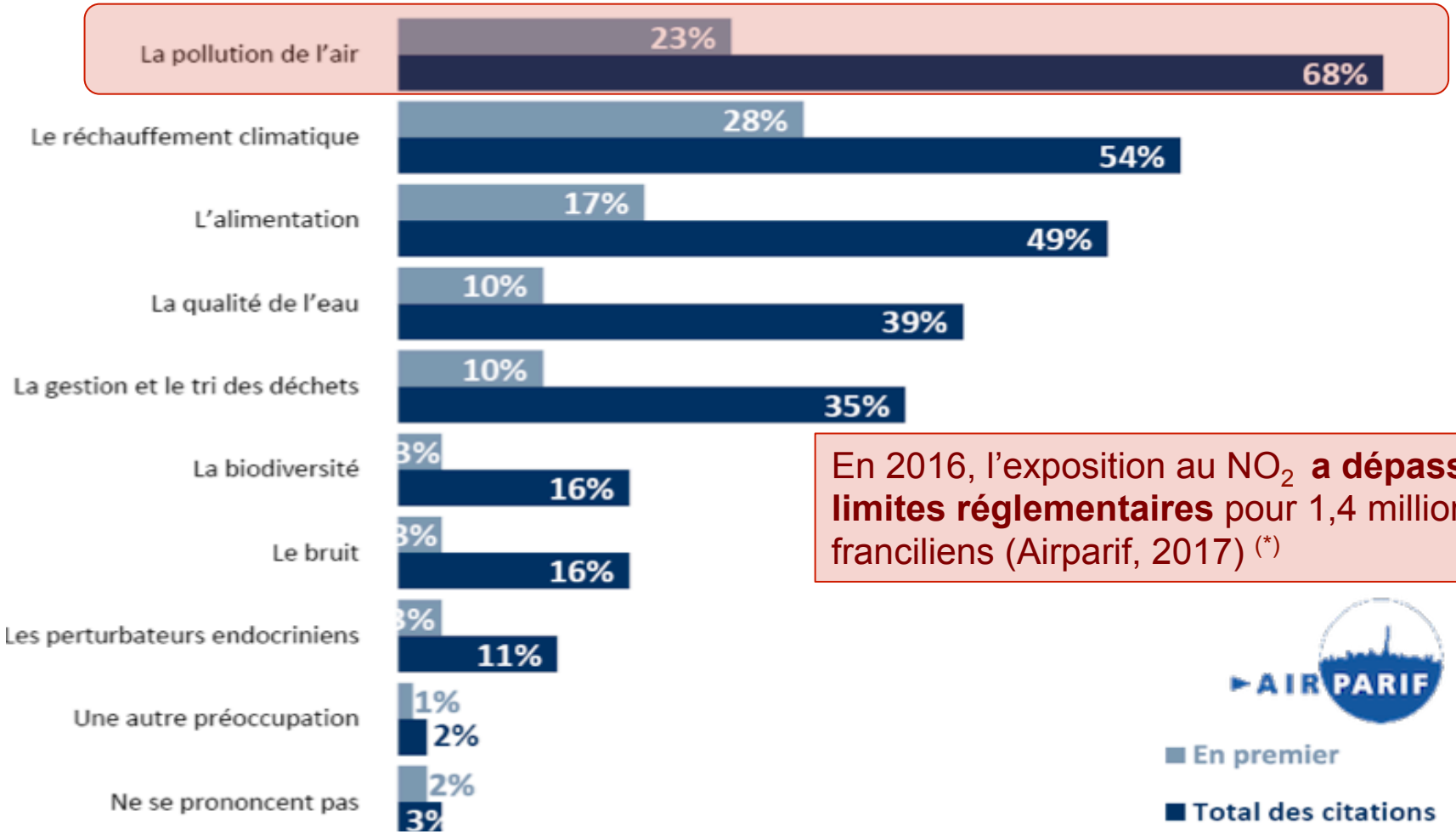
Paris, le 9 novembre 2017

david

UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

Qualité de l'air : 1ère préoccupation environnementale des Franciliens



En 2016, l'exposition au NO_2 **a dépassé les limites réglementaires** pour 1,4 million de franciliens (Airparif, 2017) (*)

Sondage Ifop pour AIRPARIF réalisé par questionnaire auto-administré en ligne du 9 au 14 octobre 2014 auprès d'un échantillon de 500 personnes, représentatif de la population francilienne âgée de 15 ans et plus.

(*) AIRPARIF 2017, Surveillance et information sur la qualité de l'air en Île-de-France Bilan année 2016

Limites de l'observation de la qualité de l'air vis-à-vis de l'exposition

- Les associations de surveillance de la qualité de l'air utilisent un réseau de stations de mesures fixes et de la modélisation

Capteurs de dioxyde d'azote (NO2)

Crédit Airparif

- **Mais on manque de données sur :**
  1. La quantification de **l'exposition individuelle réelle**
  2. L'analyse d'impacts de la pollution sur la **santé individuelle**
  3. La compréhension des **disparités de risque sanitaire** observées entre des groupes de population

Comment mesurer et analyser l'exposition individuelle ?

- Emergence de **mini-capteurs à bas coût** (*,**):
  - Une palette de **capteurs nomades**
  - Couplés avec la **géolocalisation** par GPS
  - Offrant un capacité de **stockage, de communication, faciles d'utilisation**
- Technologie prometteuse pour **mesurer en continu et partout l'exposition individuelle** et révéler les changements rapides et les pics d'exposition
- Permettent de **densifier le réseau de mesures** et de **couvrir tous les milieux** (les porteurs jouent le rôle de **sondes mobiles**)

(*) Air sensor guidebook. EPA/600/R-14/159, rapport de l'agence de protection de l'environnement des Etats Unis, Juin 2014

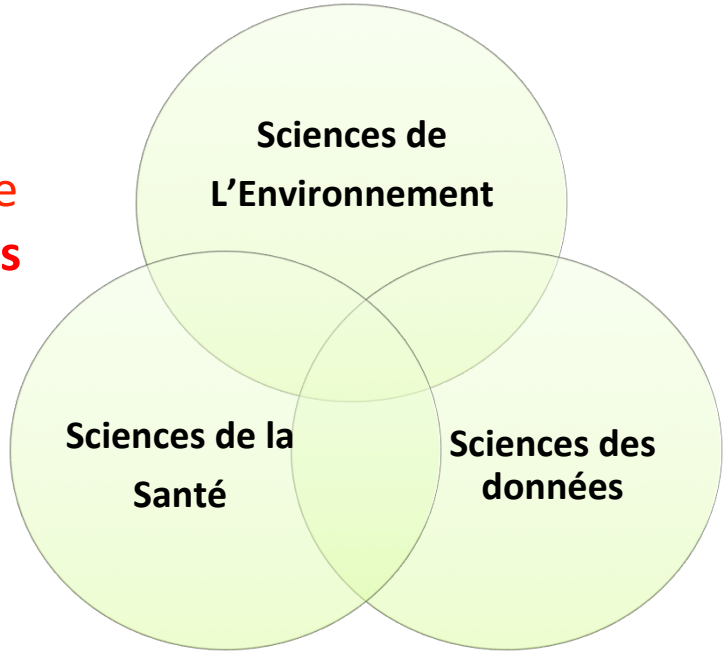
(**) Snyder, E. G et al. (2013). The changing paradigm of air pollution monitoring. Environmental science & technology, 47(20), 11369-11377.

Plan de la présentation

- ❑ Contexte
- ☑ **Présentation du projet POLLUSCOPE**
- ❑ Focus sur l'aspect Données

Objectif général du projet POLLUSCOPE

- Le projet Polluscope(*) adresse les questions méthodologiques et techniques visant à la fois **l'évaluation** des capteurs nomades et l'**analyse** de **l'exposition individuelle** à la pollution de l'air et de **ses effets sanitaires** sur la population à risque.

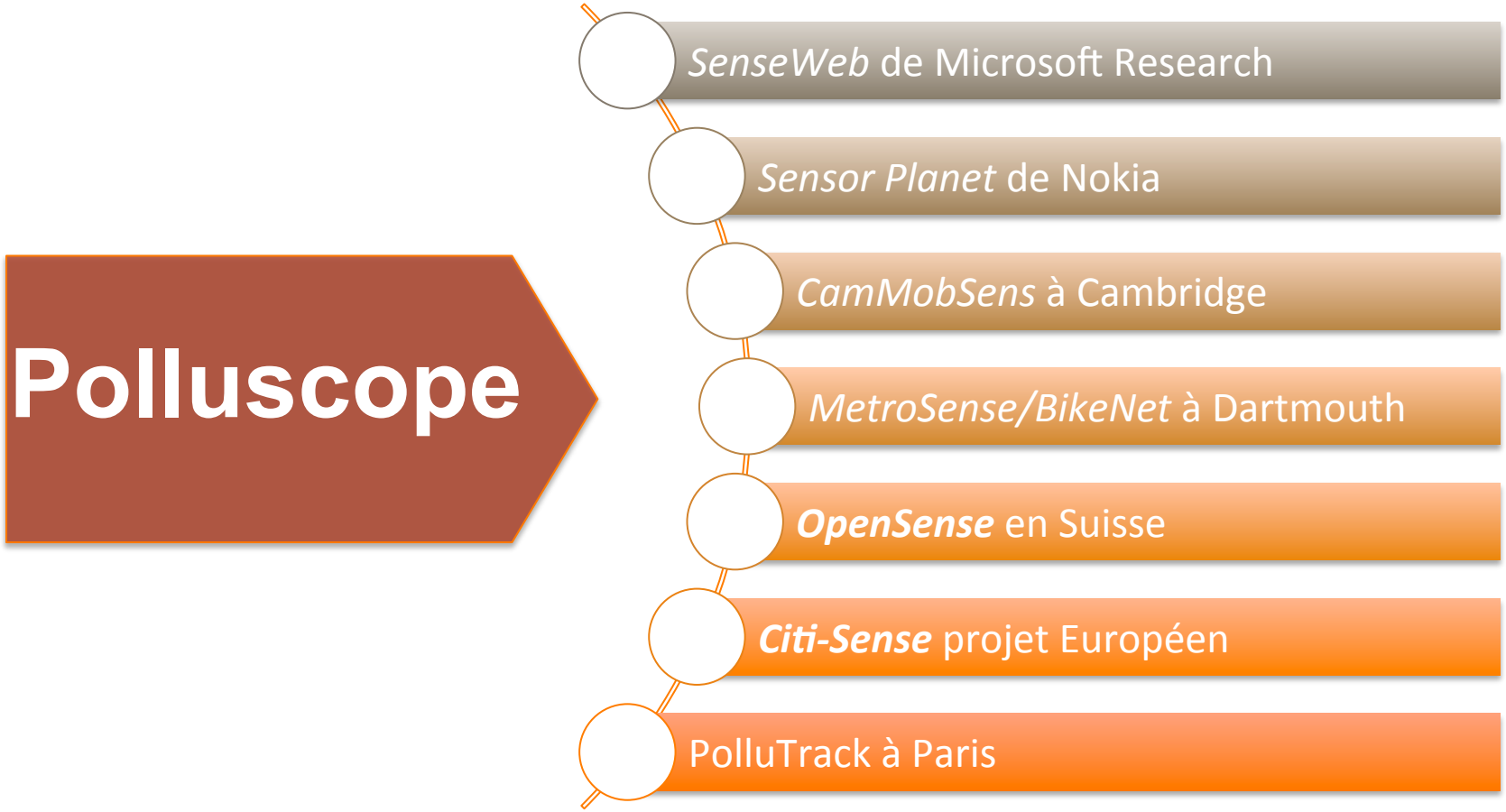


- On propose pour cela le développement d'une **plateforme** de **collecte**, de **gestion** et d'**analyse de données** issues de **capteurs environnementaux**, **d'activité** et **de santé**.

(*) Projet ANR débuté en septembre 2016 − Durée 4 ans

Positionnement versus l'état de l'art

- *Mobile Crowd Sensing mais pas axés sur l'exposition et ses risques*



Positionnement versus l'état de l'art

- Ce qui caractérise Polluscope :
  - **Focalisé du l'exposition individuelle** à la pollution atmosphérique et ses **effets sanitaires** sur les fonctions respiratoires et cardiaques
    - Capteurs portés 24h/24 7j/7 par des individus (pas que statiques ou fixés sur un véhicule)
    - Couvre tous les milieux intérieur et extérieur
  - **Polluants visés plus représentatifs** de la pollution de l'air :
    - Polluants gazeux (ozone, NO_2), particulaires (*black carbon*, PM) et COV
  - **Détection du contexte** par reconnaissance d'activité du participant et de son micro-environnement (domicile, travail, métro, route, activité physique…)
  - **Plateforme intégrant le processus complet** de pré-traitement, de requêtes et de fouille de données

Mode opératoire & Choix des capteurs

Polluants ciblés	NO_2, O_3, COV, Formaldéhyde Particules fines, Black Carbon, LDSA (lung-deposited surface area)
Plage de mesure	O_3 = 0 à 250 ppb NO_2 = 0 à 500 ppb Black Carbon = 0 à 50 µg/m3 Particules fines (PM10) = 0 à 1000 µg/m3 PM2.5 = 0 à 1000 µg/m3
Seuil de sensibilité	Le fabricant devra le spécifier (exemple : (+) 5ppb – Particules : 1 µg/m3, 0.1 µg/m3 sur BC)
Fréquence d'acquisition	Idéalement inférieure à 5 minutes (le fabricant devra la spécifier)
Dérive à long terme	À préciser pour 24 heures, une semaine et un mois
Dimensionnement de l'expérimentation	- 2x80 participants, 10 à 20 en // - 2 campagnes annuelles de 6 semaines
Calibration	- préciser les procédures de calibration - fréquence des calibrations - standard à utiliser

Cahier des charges :

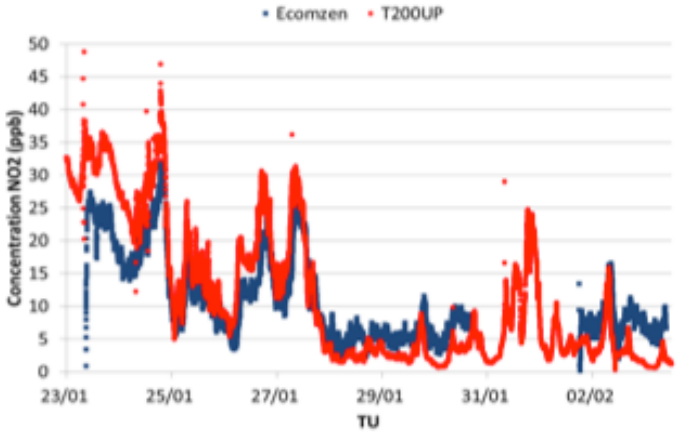
- ✓ Polluants règlementés /
- ✓ Air intérieur & extérieur
- ✓ Fiabilité / reproductibilité
- ✓ Performances
- ✓ Connectivité
- ✓ Ergonomie
- ✓ Coût / budget prévu

Evaluation en 4 phases:

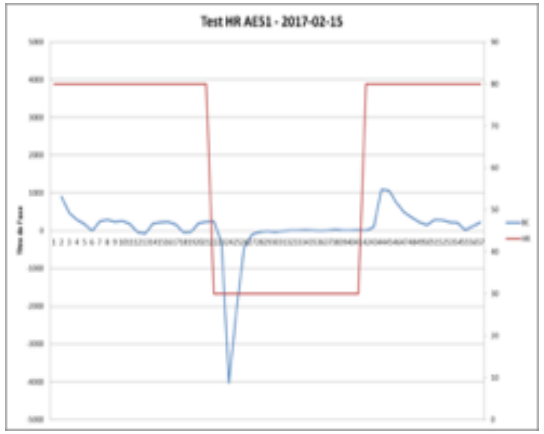
1. Etude bibliographique
2. Test / stations de référence
3. Test en conditions contrôlées
4. Test terrain en mobilité

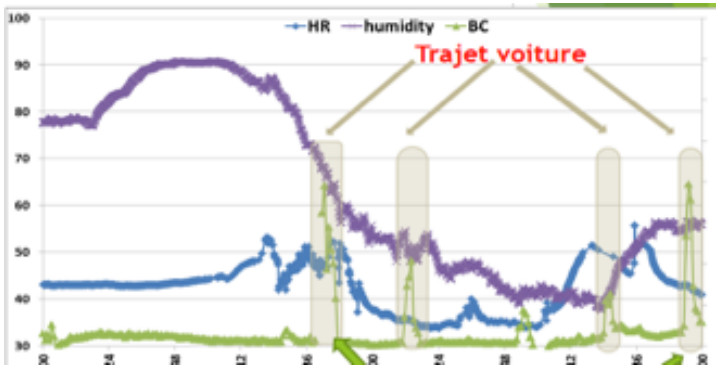
Mode opératoire & Choix des capteurs

- ✓ Test en extérieur près des instruments de référence de la station SIRTA @LSCE
- ✓ Tests en chambre avec contrôle de conditions climatiques @AirParif
- ✓ Tests en mobilité (Airparif, Cerama, LSCE impliqués)



Crédit Polluscope/Airparif, Cerema et LSCE

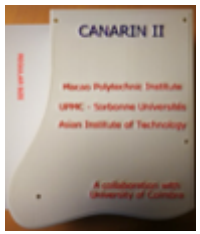




Evaluation des capteurs - Bilan

➢ Capteurs retenus :

AE51 (Black Carbon)



Canarin (PM)

Cairsens (NOx)

Crédit fabricants

➢ Des avancées mais des limites

- **Résultats contrastés** entre les différents capteurs testés : données parfois aléatoires et peu reproductibles entre appareils différents
- **Pertes de données** pour plusieurs capteurs
- Seul le **capteur de Black Carbon est très fiable** (**retenu**)
- Seuls **deux capteurs (PM et NO2) mesurent correctement la tendance** (**retenus**)
- Solutions **diverses** en termes d'ergonomie, de connectivité,… et **peu intégrées**
- Pour couvrir les polluant visés dans Polluscope, il a fallu combiner plusieurs capteurs dont les mesures sont exploitables et offrant un compromis entre les autres critères.

➢ Le boitier multi-capteurs idéal n'existe pas encore !

Plan de la présentation

- ❑ Contexte
- ❑ Présentation du projet POLLUSCOPE
- ☑ **Focus sur l'aspect Données**

Objectifs du traitement des données

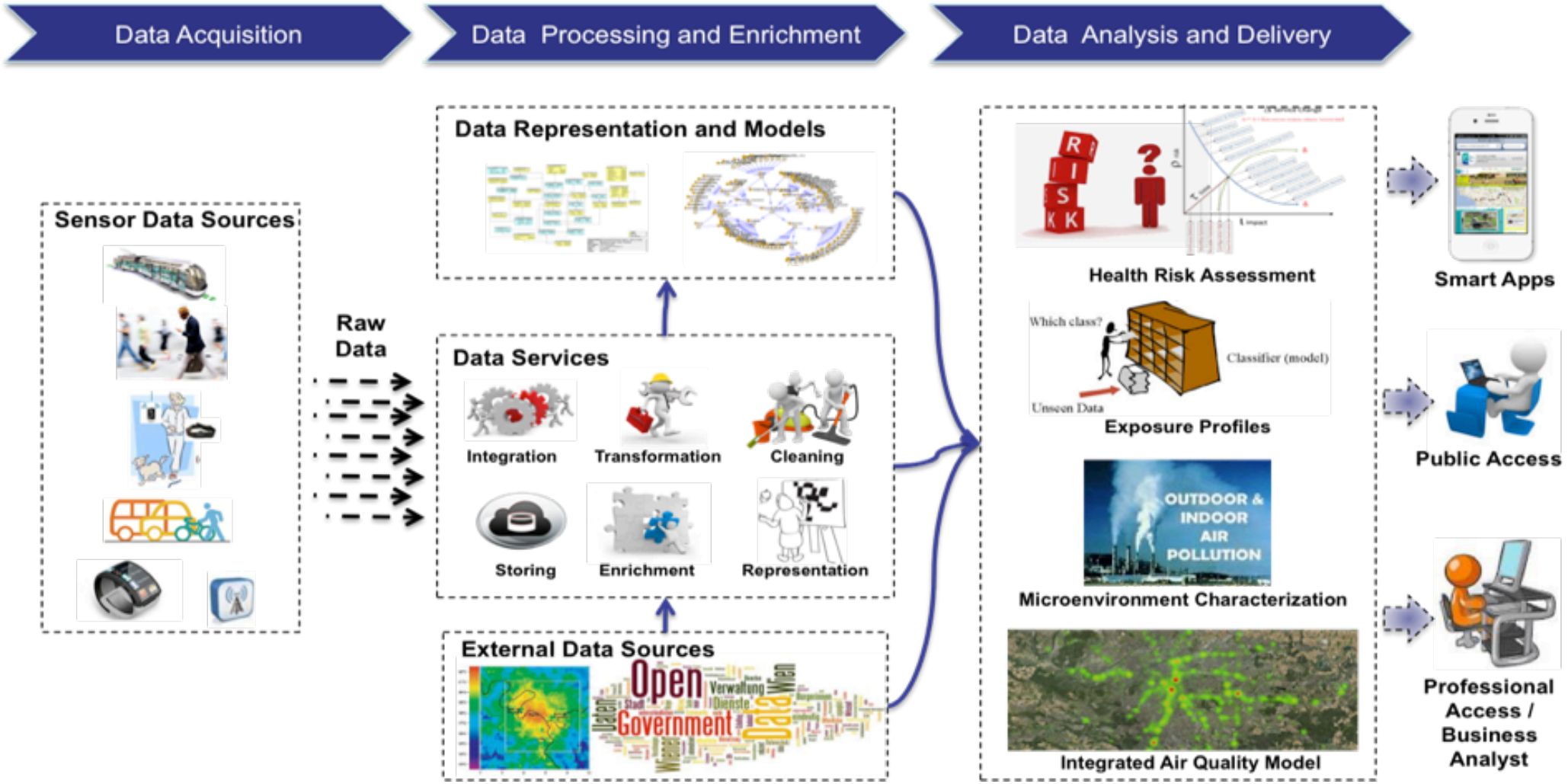
- **Exploiter** les données collectées **pour mieux cerner le phénomène d'exposition** à une mauvaise qualité de l'air : **qui, où, quand, combien et comment ?**
- **Etudier finement les liens** entre **l'exposition, l'activité et l'état de santé**
- **Evaluer l'apport des données participatives** à la **massification des mesures** de la qualité de l'air

➢ **Tirer le meilleur profit de cette nouvelle source de données.**

Proposition - Plateforme de données Polluscope

- **Architecture** ouverte et évolutive, offrant des **services de données**
- **Modèle de données** capturant la **variabilité spatiale et temporelle** des données
- **Requêtage efficace** de données **potentiellement volumineuses**
- **Pré-traitement** pour **pallier l'imperfection et l'hétérogénéité** des données
- **Enrichissement par le contexte** **géographique** et le **type d'activité**
- **Analyse exploratoire** multi-dimensionnelle
- **Fouille de données** pour détecter des profiles d'exposition, pour caractériser des micro-environnement, pour expliquer / prédire le niveau d'exposition selon la catégorie de population, ses activités, la période et/ou le lieu, …
- **Confidentialité** des données des participants

Plateforme de données Polluscope – Architecture Fonctionnelle



Challenge Modèle de données

➢ Spécificité des données de mesures :

- Séries temporelles **multi-variées** et **géo-localisées** (t, localisation, mesures)

➢ Problèmes :

- ✓ Issues de capteurs divers, souvent **asynchrones**
- ✓ Échelles **hétérogènes**
- ✓ Comporte des **données manquantes** et du bruit
- ✓ Fortement dépendantes de facteurs externes (ou du **contexte**)

➢ Le modèle de données doit permettre :

1. une **vue unifiée** des données synchrones et comparables à différentes échelles
2. Le **nettoyage** des données : débruitage et completion des données
3. **L'analyse contextuelle** spatiale et temporelle

Proposition - Modèle de données basé fonctions

- Données en entrée comprenant typiquement :

  (t, mesures) et (t, longitude, latitude) parfois asynchrones, bruitées et incomplètes

Idée : *Représenter **l'interpolation** au lieu les données brutes !*

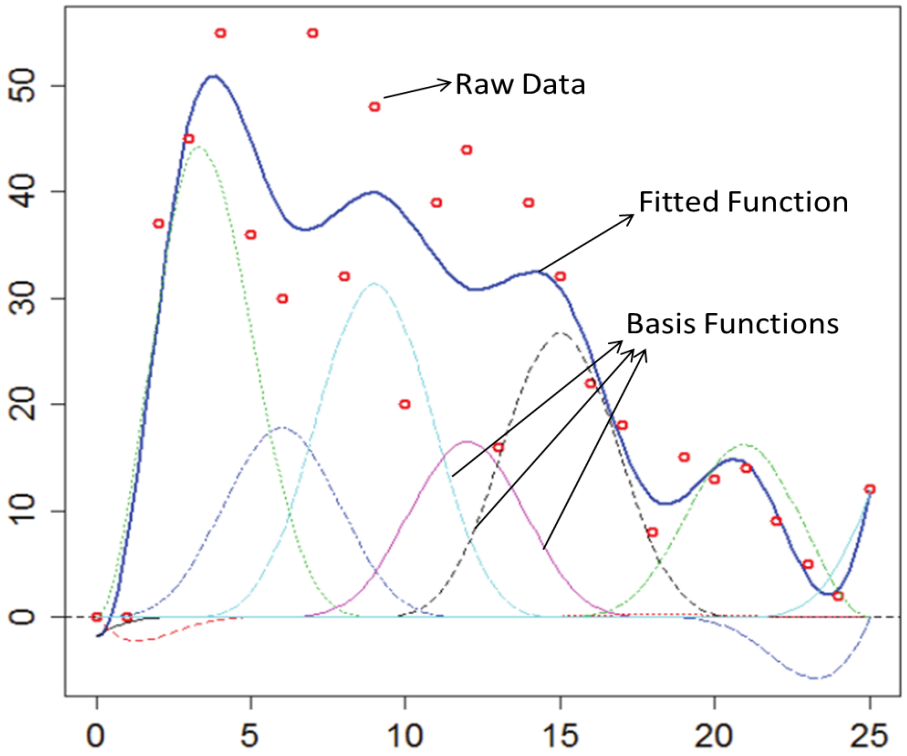
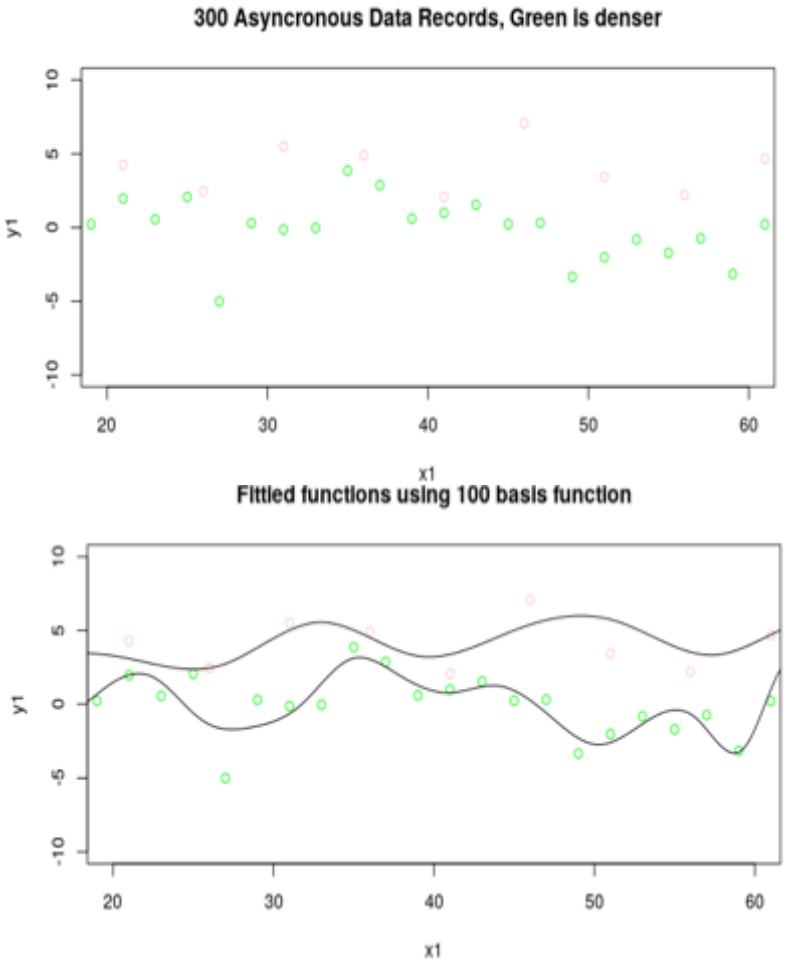


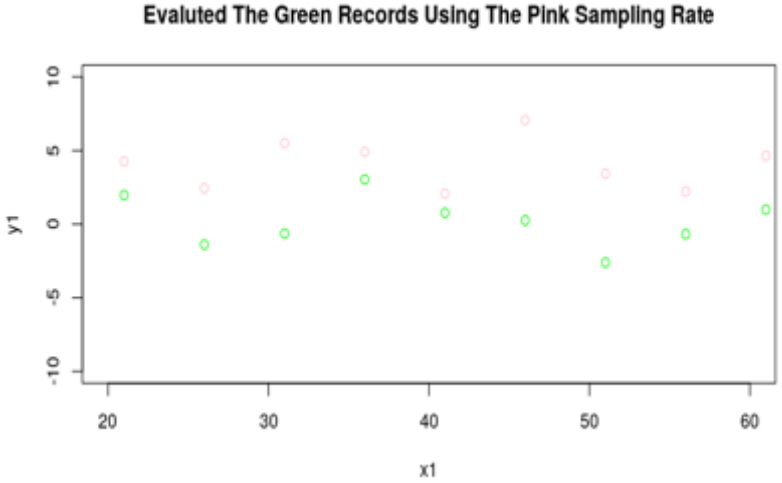
Illustration FDA
(Ramsay & Silverman, 2005)

Ramsay, J. O. (James O.) and B.W. Silverman (2005). Functional data analysis. Springer

Proposition - Modèle de données basé fonctions

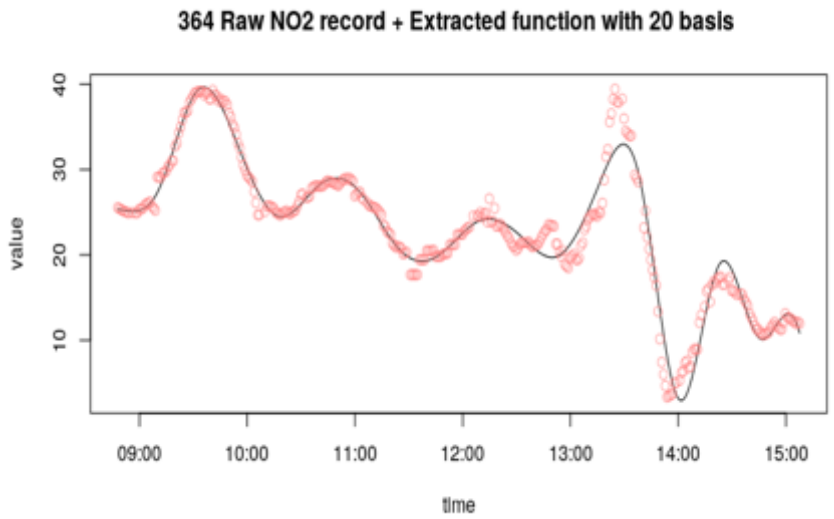
Intérêt 1 : *Permet* ***d'intégrer des données asynchrones***

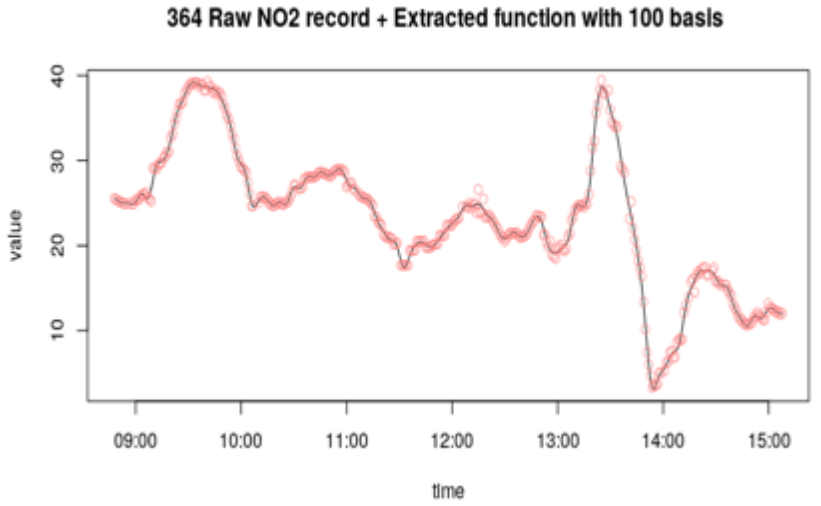




Proposition - Modèle de données basé fonctions

Intérêt 2 : *S'ajuste à l'échantillon de mesures tout en permettant de compléter les* ***données manquantes*** *et* ***d'éliminer le bruit***





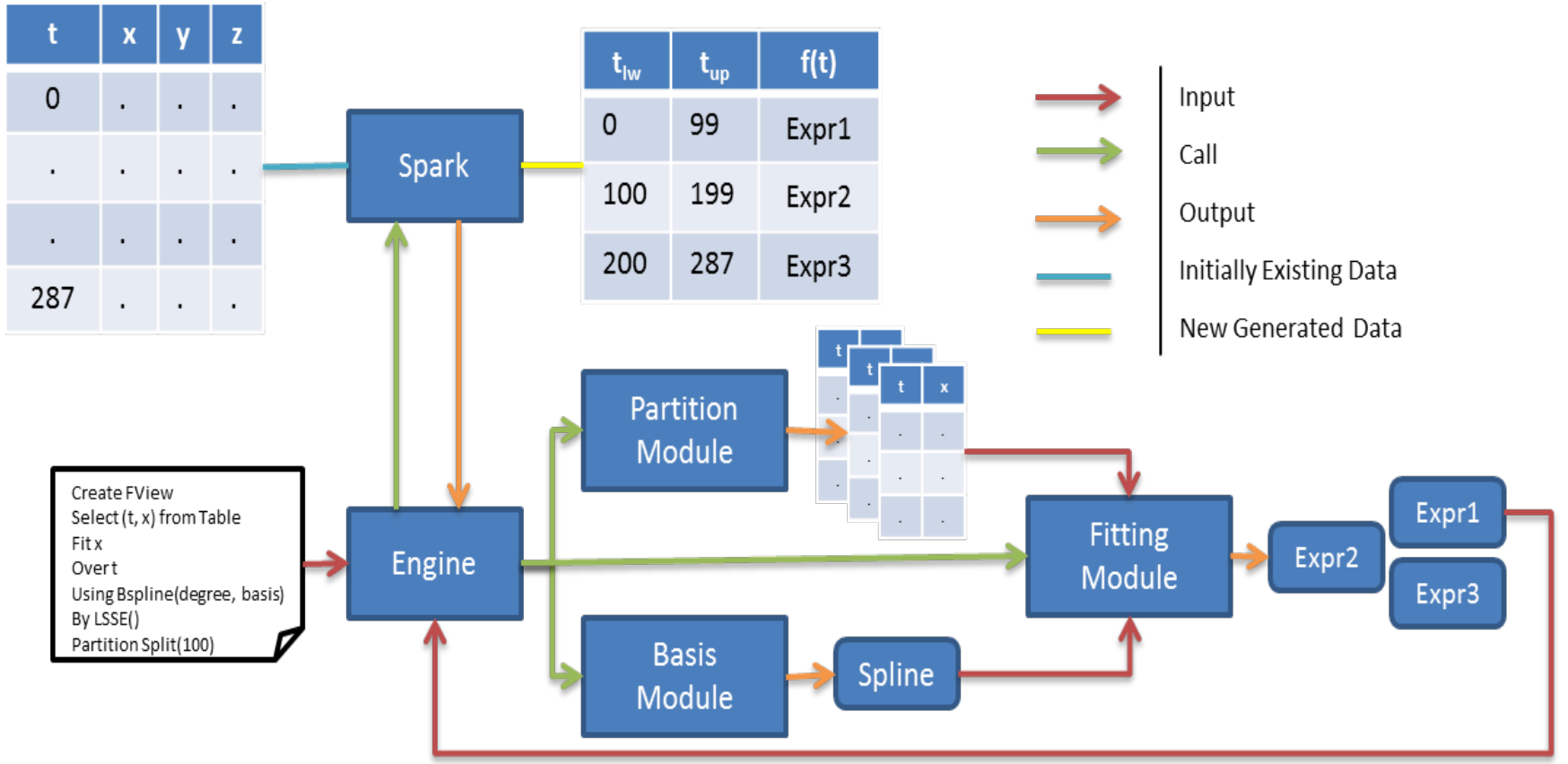
Intérêt 3 : ***plus compact*** *et* ***plus sémantique*** *que les données de départ*

Processus de construction du modèle de fonctions

Entrée : *données originales + commande SQL étendu (défini)*

Sortie : *Modèle décrit par des fonctions sous SPARK*

- ***Segmentation et recherche automatique de coefficients***

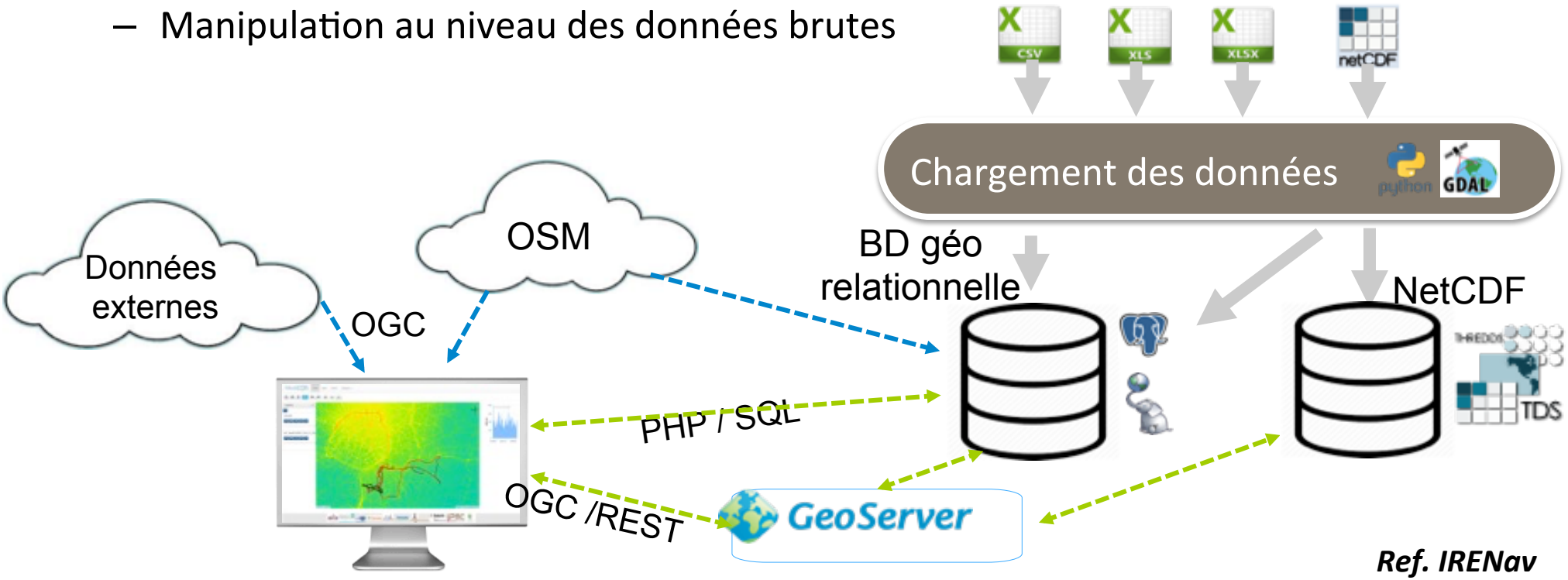


Challenge Variété et Volume de données

- Même limité en territoire et en nombre de participants, le format est variable et le volume parfois conséquent :
  - Données de mesures de capteur individuel (XLSX - ~ 5Mo/jour)
  - Données de suivi GPS (CSV - ~ 100Mo/jour)
  - Données d'annotation de déplacements (CSV - ~ 5Mo/jour)
  - Sorties de modèle de pollution (Airparif – NetCDF) - ~ 18Go/jour
  - Données de références (Shapefile – chargées une seule fois)
- ***Besoin d'une plateforme intégrant les techniques Big Data***

Architecture technique initiale

- **Solution basée SIG** (ici **Postgres-PotGIS/GeoServer**)
  - Similaire à l'état de l'art
  - Basé sur les standards, donc interopérable
  - Manipulation au niveau des données brutes

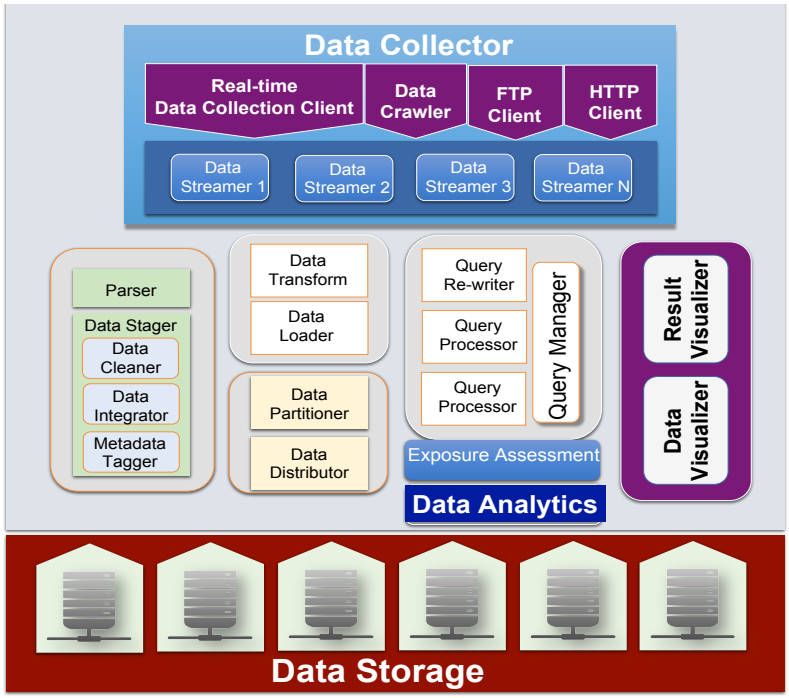


Ref. IRENav

Architecture technique cible

➢ Solution basée sur les technologies Big Data

- Intègre la modélisation par les fonctions dans un framework distribué (Spark)
- Intègrera **tous les services de données** allant de l'acquisition à l'analyse
- Correspond plus aux objectifs fixés dont le **passage à l'échelle**

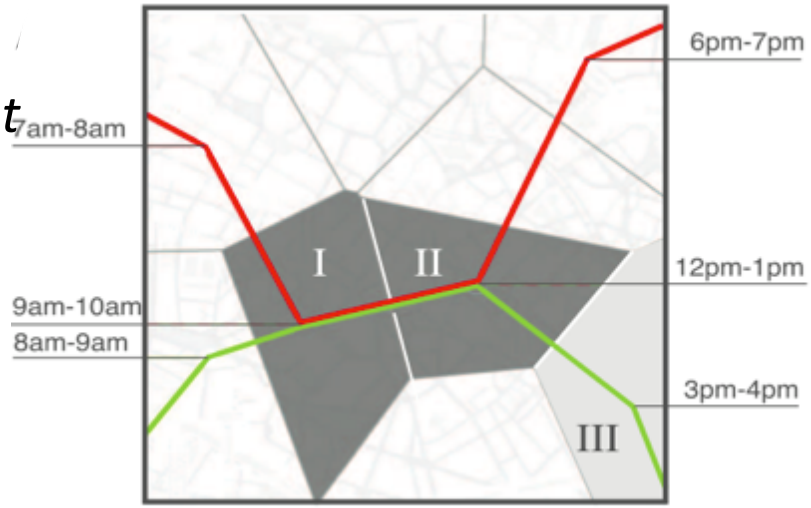


Confidentialité

- Les participants sont réticents à partager leurs données de localisation
  - L'anonymisation simple ne suffit pas

  Ex. en localisant le domicile et le travail, on peut souvent identifier une personne

- Solution envisagée :
  - Masquage de la localisation des lieux susceptibles de ré-identifier le participant
  - Seules les données agrégées sont publiées
  - Compromis utilité − anonymité



Ref. Montjoye Y. A. et al. (2013). *Unique in the crowd: The privacy bounds of human mobility. Scientific reports, vol. 3.*

L'Equipe POLLUSCOPE

Informatique

Géomatique

Urbanisme

Santé

Environne-ment

Métrologie

Site web : polluscope.uvsq.fr