



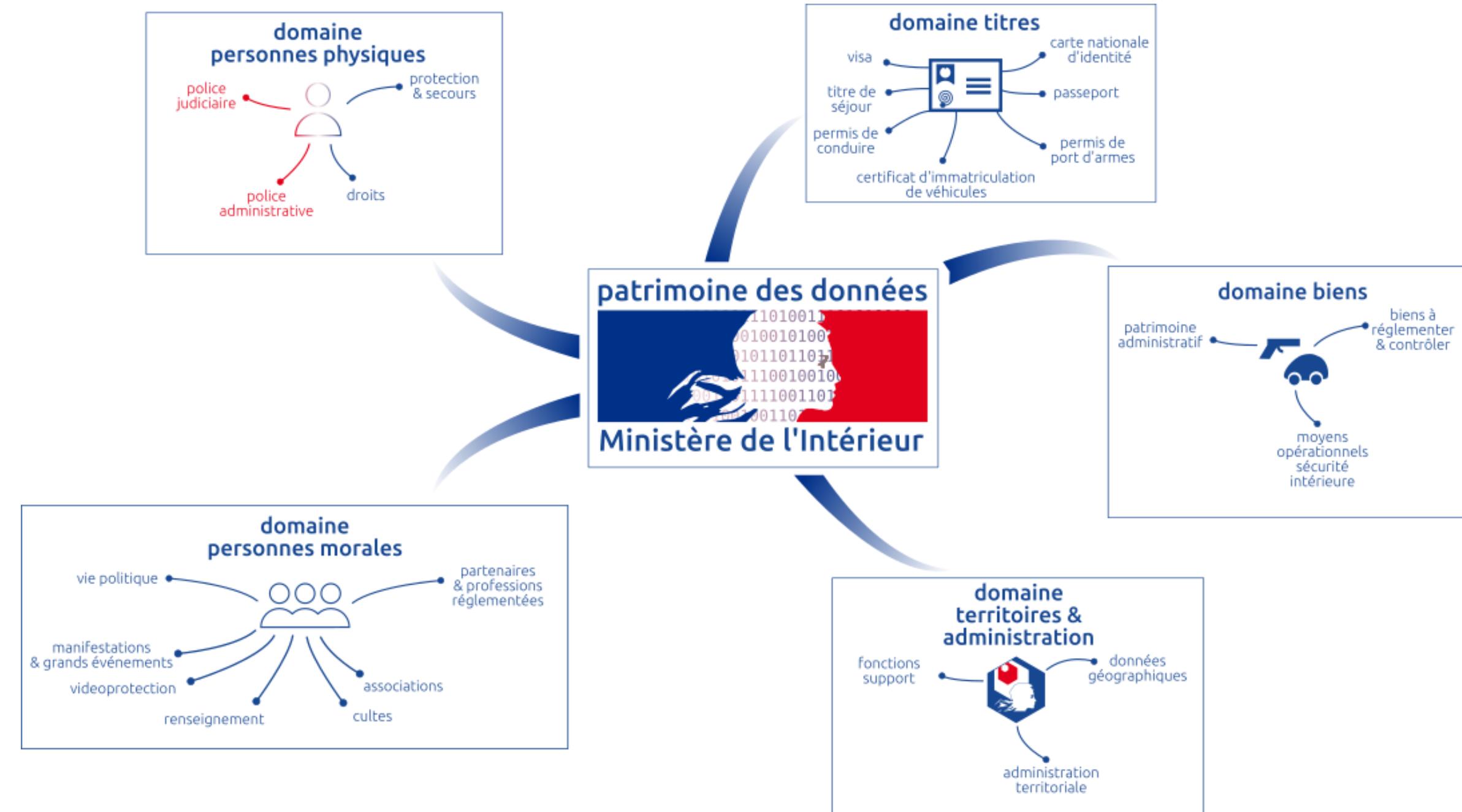
ACCÉLÉRER LA VALORISATION DES DONNÉES



Fabien Antoine
Ministère de l'Intérieur



LES DONNÉES





VARIÉTÉ

- structurées et référencées (passages au frontières)
- opendata (insee, IGN, météo, ...)
- libellés libres (villes, véhicules, noms, sociétés)
- géographiques (zones de compétences : sécurité intérieure, civile, communes, ...)
- textes (procédures police)
- images (radars, vidéosurveillance)



VOLUMES

- titres : > 100 millions
- infractions : > 1 Mds
- traffic: > 1000 Mds



VELOCITÉ

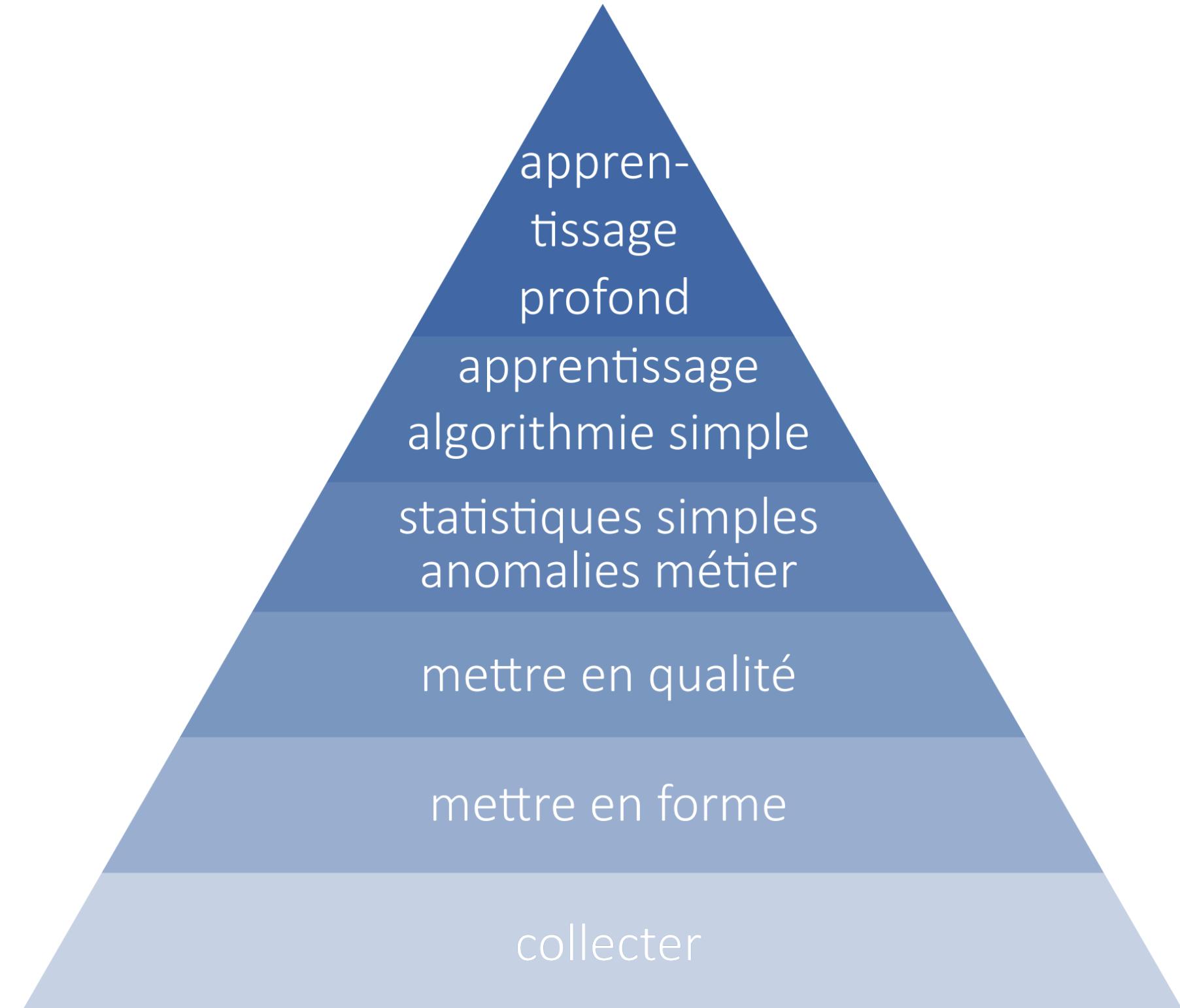
- temps statistiques : année / mois / semaine
(observatoires sécurité intérieure, routière ...)
- temps opérationnel :
 - jour / minute (supervision opérationnelle - sécurité civile, intérieure)
 - seconde (systèmes d'identification: contrôle frontière, ...)



LA DATASCIENCE



"PYRAMIDE DE MASLOW" DE LA DATASCIENCE





BON À SAVOIR

- 60% du travail = préparation
- 20% = algorithmie/statistiques simples
- IA si cas mature ou ciblé
- complexité \perp bénéfice

ITÉRER AVEC LE MÉTIER POUR ÉVALUER LA VALEUR



RÔLES ET COMPÉTENCES VARIÉES

- **data architect** : architecture big data, sécurité, machine as a code
- **data engineer** : transformation, automatisation, optimisation, qualité
- **AMOA data**: modélisation et bénéfice métier, rapports visuels
- **data scientist** : algorithmie générale: graphes, fuzzy, apprentissage simple ...
- **data expert** : géographique, deep learning image ou texte
- **développeur** : valorisations de données interactives



L'ACCOMPAGNEMENT

DE LA MISSION DE COORDINATION ET D'APPUI À LA VALORISATION DES DONNÉES

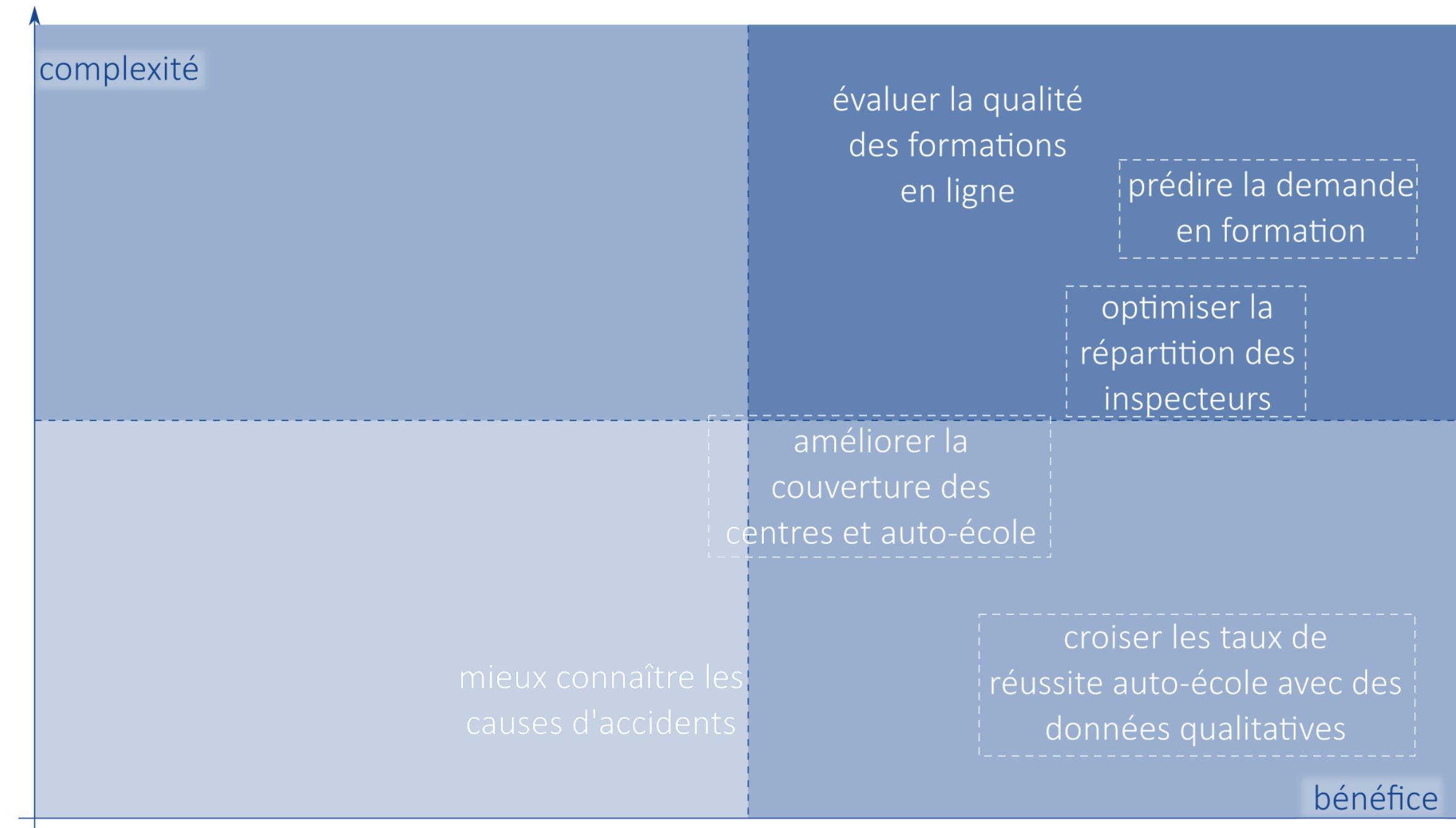


- recensement des cas d'usage data : méthode, capitalisation
- POC cas d'usage : développements, challenges EIG, pré-production



RECENSEMENT DES CAS D'USAGES

ÉDUCATION ROUTIÈRE





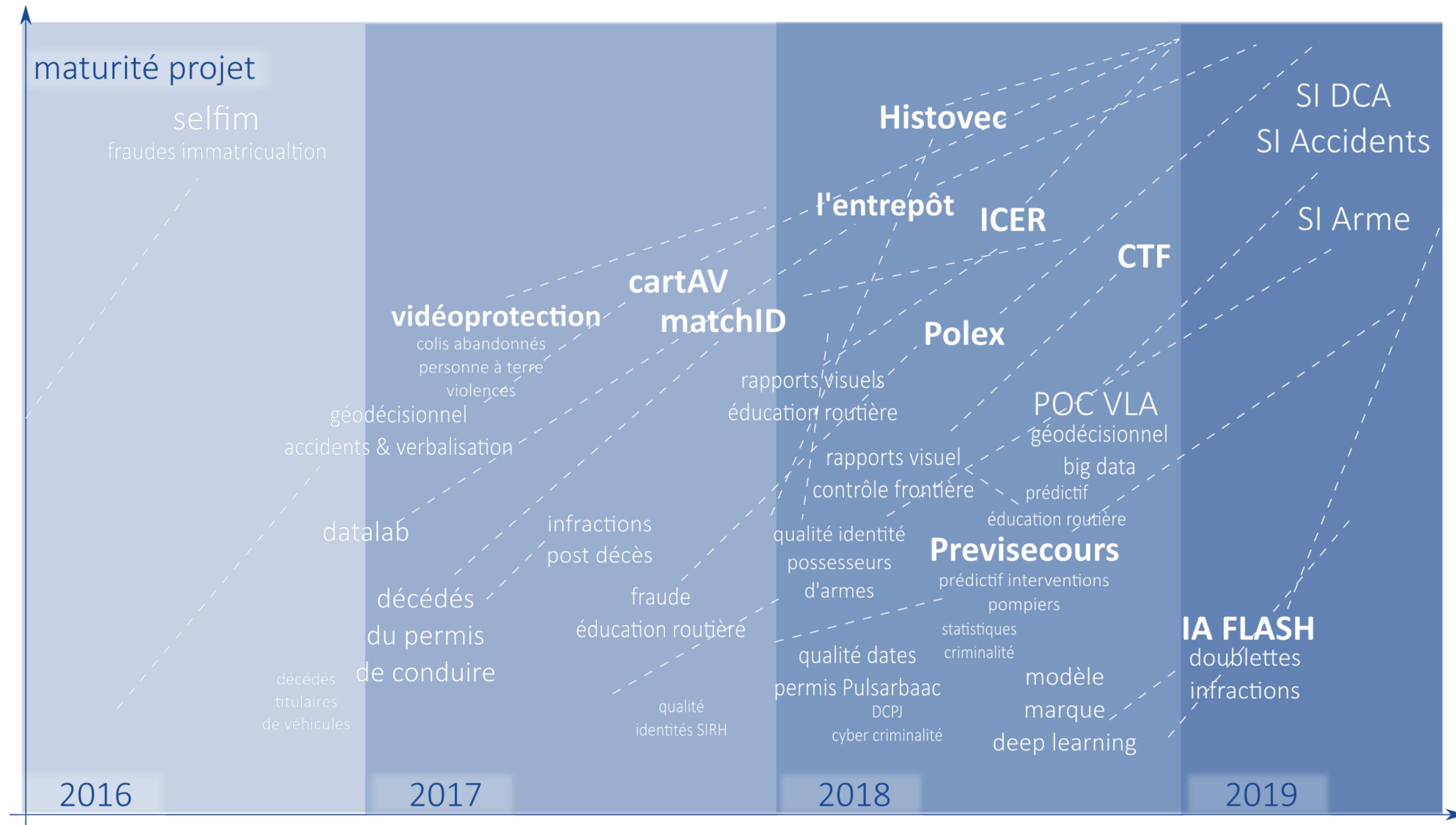
UN RECENSEMENT À GÉNÉRALISER

- favoriser l'apport de la datascience au métier
- analyser et suivre le bénéfice des actions data
- prioriser
- diffuser, communiquer



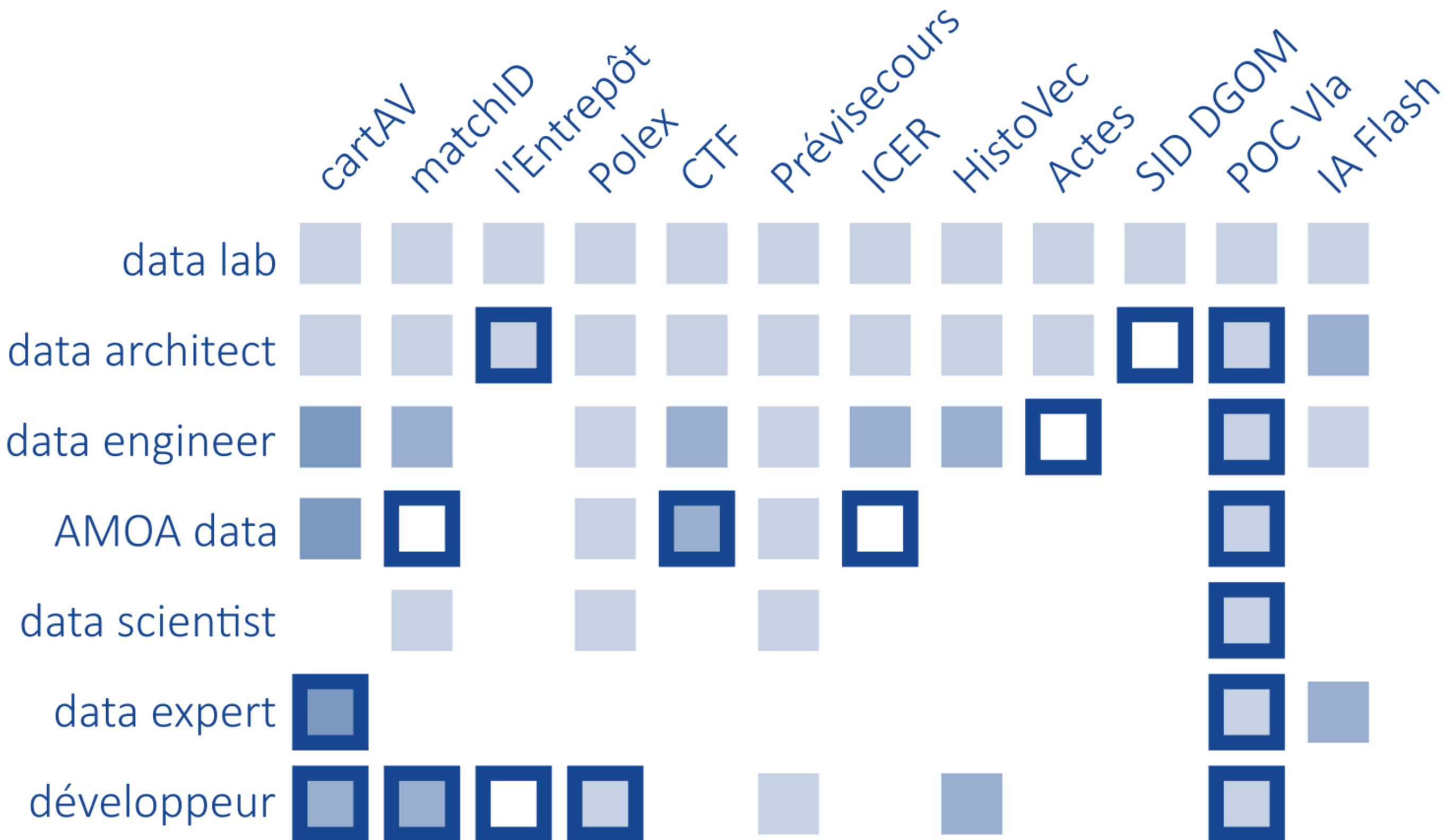
POCS CAS D'USAGE

- accompagner depuis l'idée initiale
- montage projet financement/sponsoring, FTAP, EIG, stages, interne ...
- maquetter/faire/encadrer
 - architectures, analyses, algorithmes, développements ...
- héberger / maintenir sur le datalab
 - jusqu'à industrialisation sur l'entrepôt





MOBILISATION DES COMPÉTENCES





FACTEURS DE SUCCÈS

- sponsor direction et implication du métier opérationnel
- cadre de protection des données
 - mandats, engagements personnels, sécurité
- disponibilité des données anticipation de 3 à 6 mois
- dispositif léger : reporting minimal et documentation a posteriori
- savoir réviser / décomposer ses objectifs
- anticiper la mise en production capitaliser sur l'entrepôt
- capitaliser y/c les échecs code source, rapports, communication



Données et cas d'usage sur <http://catalog.datalab.mi>



LES OUTILS



- Dataiku/DSS

- Dataiku/DSS
 - collecter, mettre en forme, en qualité, algorithmie, apprentissage simple et profond

- Tableau rapports visuels, statistiques évoluées
- bases analytiques

- bases analytiques
 - vertica (agrégation jusqu'à l'exabyte), elasticsearch (fuzzy), postgis (géographique)

- matériel (datalab: 400 vCpu, l'entrepôt: scalabilité du cloud MI)

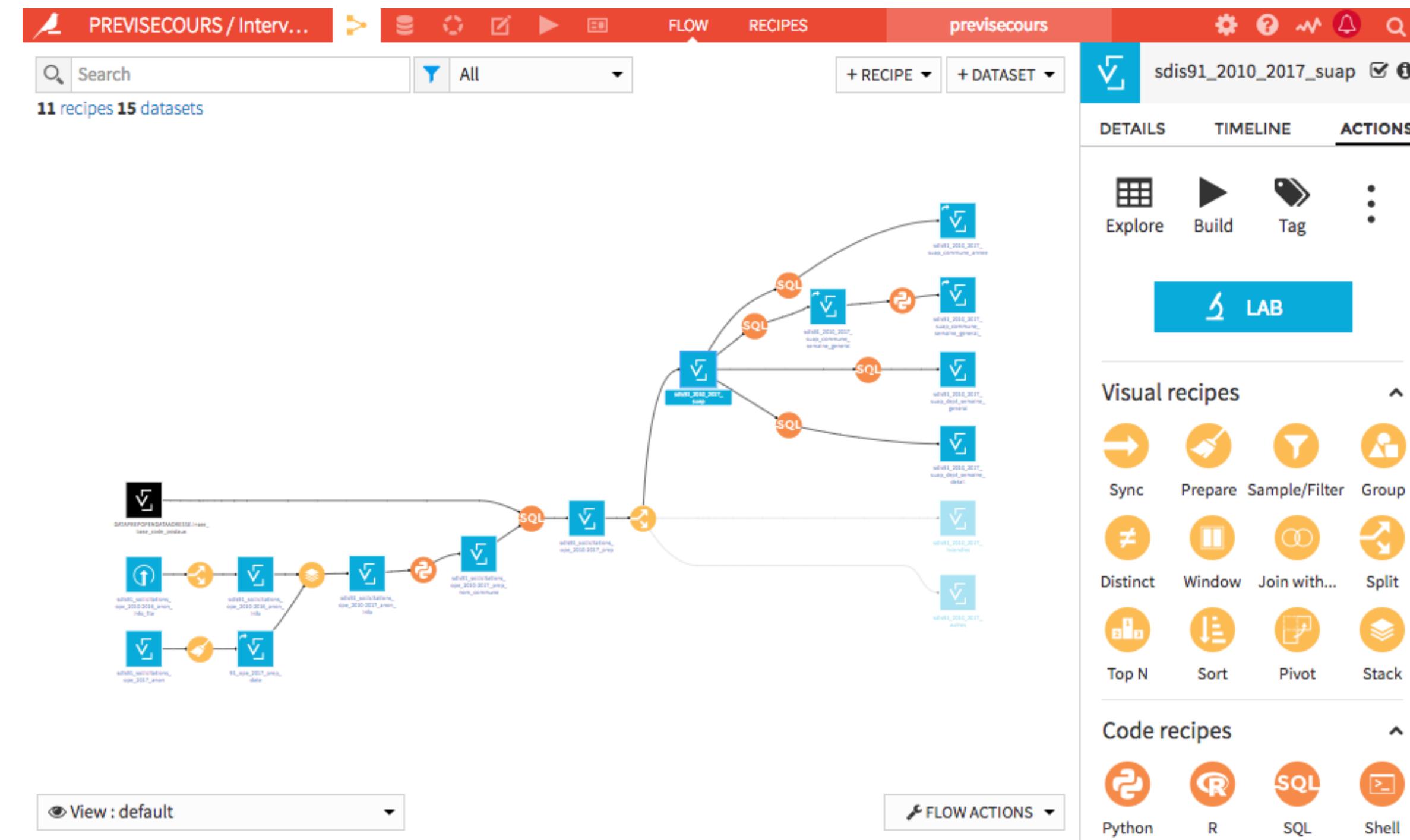


DATAIKU

- pour: data architect, data engineer, data scientist, data expert, développeur
- collaboratif
- technologies: SQL, python/scikit, R, scala, hadoop, tensorflow, ...
- collecte des données jusqu'à deep learning
- API de scoring, web apps

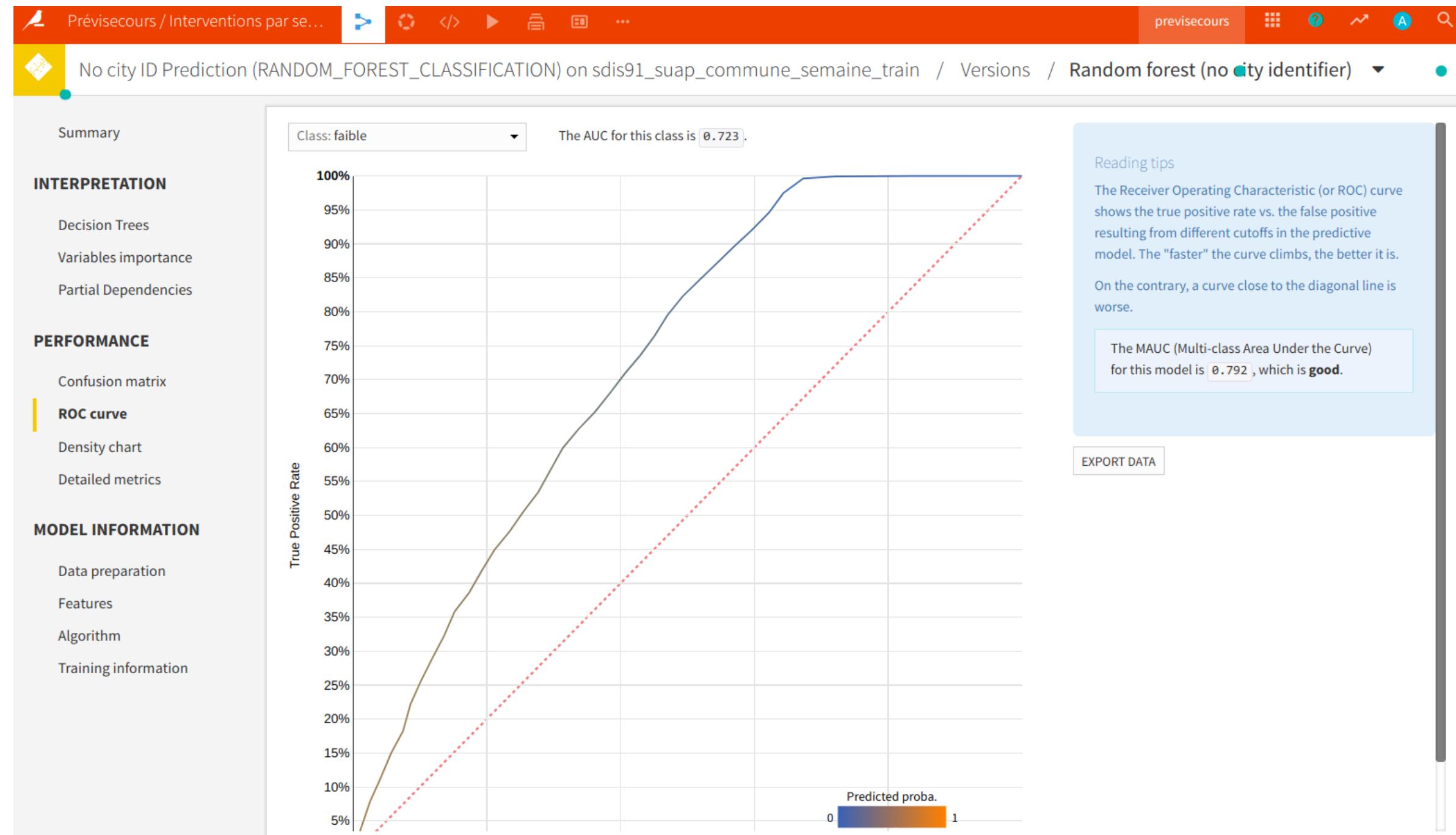


DATAIKU: L'ORCHESTRATION DES TRAITEMENTS





DATAIKU : LE LABORATOIRE PRÉDICTIF





TABLEAU

- pour: data scientist, statisticiens, contrôle de gestion
- calculs dynamiques jusqu'au milliard d'enregistrements
- visualisations avancée, jusqu'au géographique (simple)

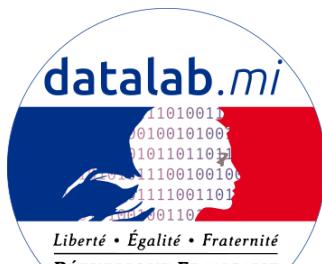
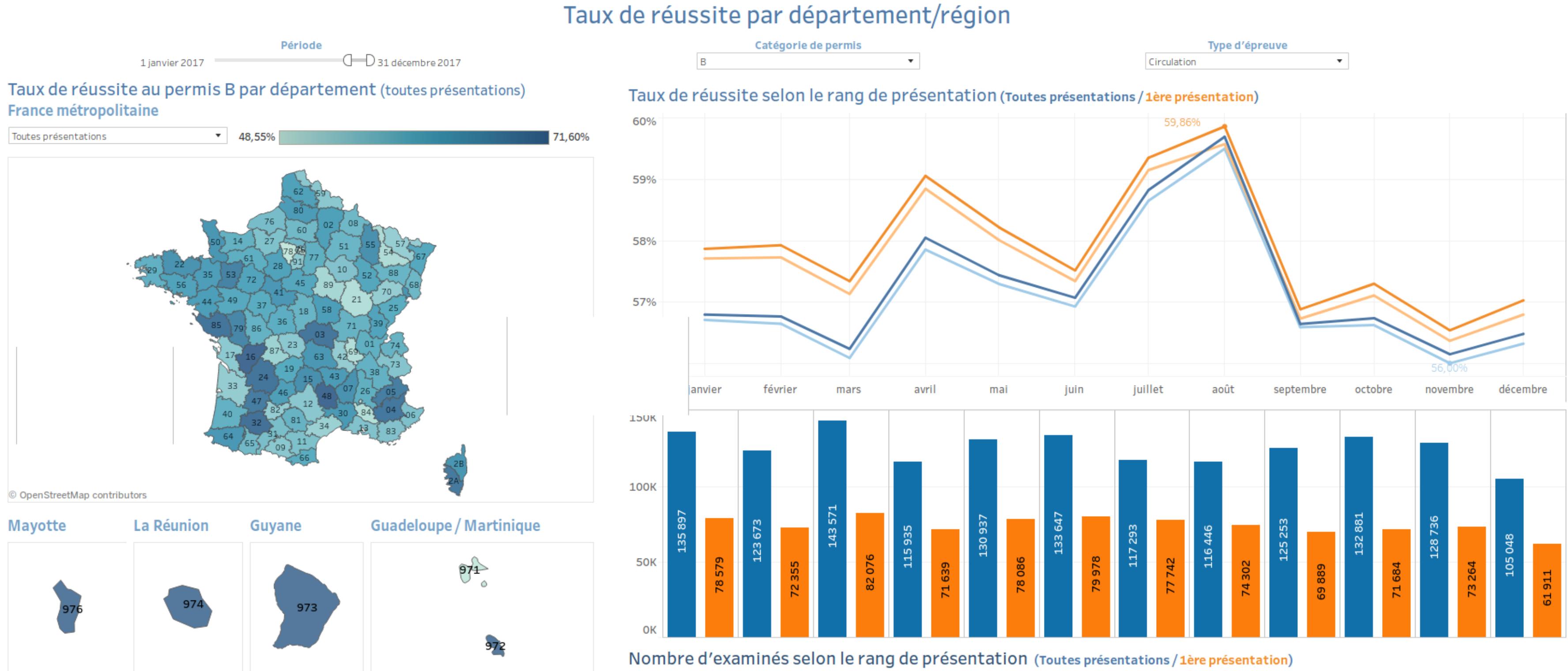


TABLEAU : LES TABLEAUX DE BORD INTERACTIFS



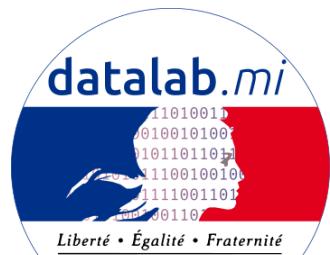


TABLEAU : FORMULES TYPE EXCEL SUR DU BIG DATA

Tableau de bord Dispositif

Taille Automatique

Feuilles Liste CE Liste CEs Carte CE Repartition / CE Titre Histo CE Total Ex TR EcartDep

Période 1 janvier 2017 31 décembre 2017

Département ARDECHE

Catégorie de permis B

Type d'épreuve Circulation

Rang présentation Toutes présentations

Analyse du taux de réussite des centres d'examen

Répartition des candidats TP examinés par centre d'examen

Centre	Nombre de candidats	Taux de réussite (%)
ANNONAY n°00780710	25,81%	64,53%
AUBENAS B n°00780700	23,25%	61,56% (-2,97)
TOURNON B n°00780751	21,76%	63,88% (-0,65)
LE TEIL B		68,95% (+4,42)
PRIVAS B n°00780720	13,80%	70,63% (+6,1)
BOURG ST ANDEOL B		63,90% (-0,63)
LE CHEYLARD		63,34% (-1,19)
		72,83% (+8,31)

Taux de réussite TP des centres d'examen du département

Liste des centres d'examen

N° CE	Nom
00780700	AUBENAS B
00780710	ANNONAY
00780720	PRIVAS B
00780731	BOURG ST ANDEOL B
00780741	LE TEIL
00780751	TOURNON
00780771	LE CHEYLARD

TR TP Dep

```
{FIXED [Cex Dep] :  
SUM(IIF([Ict Resultat Examen Label]=="réussite",1,0))  
/COUNT([Ict Resultat Examen Label])}
```

Le calcul est valide.

Feuilles affectées Appliquer OK

Total examinés TP % Réussite TP
6 769 64,53%

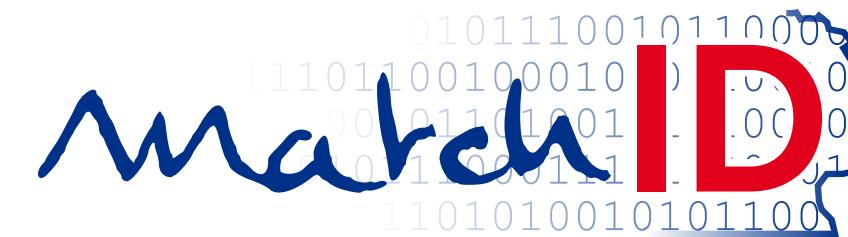
Ecart des TR CE et Dep TP (en pts)
2,96

Ecart entre le TR du centre et le TR du département (en pts)
-3,56 8,31

Carte des centres d'exams du département avec indications du nombre de candidats examinés TP



RÉSULTATS



fuzzy matching et apprentissage automatisé

- rapprochements d'états civils jusqu'à 100 millions d'identités
- des millions de morts détectés dans les bases du ministère
- une dizaine de cas traités (permis, immatriculations, armes, accidents,)
- gros volumes, haute précision, API de scoring, doublons, graphes
- score moyen de 95% (rappel=précision), AUC 99%
- publication d'un logiciel libre avec jeu opendata
- travaux avec d'autres administration (Finances, TRACFIN, Travail)



<https://matchid-project.github.io/>



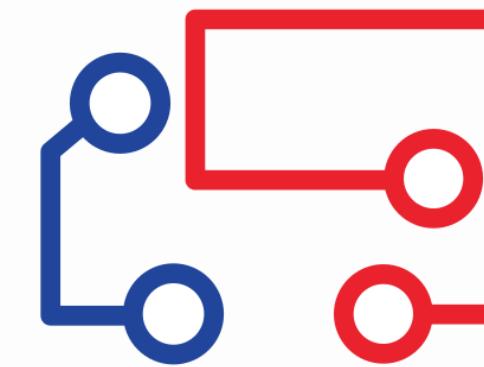
CARTAV ⚡

géocodage et géodécisionnel avancé

- croisement géographique accident et verbalisation
- intégration de **15 sources** de données dont géographiques (IGN, opendata accidents)
- 98,7% de couverture du géocodage (contre 55% pour le meilleur algorithme : addok/BAN d'étalab)
- 90% de géocodage de qualité suffisante pour une exploitation en dataviz
- 5 niveau géographiques interactifs filtrables
(région/départements/circonscriptions/communes/routes/points)
- utilisation expérimentée sur 100 circonscriptions de police
- déploiement DGPN 2019
- code opensource publié



<http://beta.datalab.mi/av>



opendata et modélisation prédictive

- aider les **sapeurs-pompiers** à intervenir plus efficacement en anticipant leur volume d'interventions
- intégration de **12 sources** de données opendata (météo, pollen, épidémies, ...) et 2 métiers (ssmsi, sdis 91)
- prévision semaines : score AUC 90% sur le secours à la personne, 78% incendies
- expérimentation SDIS 91 en cours
- hackathon en 2019 pour favoriser l'approfondissement et la généralisation
- partenariat Nexas envisagé



<https://viz.previseours.fr>



analyse d'anomalie statistiques

- lutte contre la fraude aux examens du permis
- 6 cas traités, un centre d'examen déjà fermé
- incubateur LAB-mi et label start'up d'état
- 1er déploiement octobre 2018



publication de données personnelles sécurisées

- permettre au vendeur de véhicule d'occasion de connaître l'historique de son véhicule et le communiquer à tout acheteur potentiel
- pseudonymisation et hash salé
- incubateur LAB-mi et label start'up d'état
- réalisation/mise en production homologuée en 2,5 mois
- évaluations avec Leboncoin, Armis auto



rendez-vous sur <https://histovet.interieur.gouv.fr>



CONTRÔLE FRONTIÈRES

statistiques opérationnelles

- statistiques générales du passage aux frontières (150M enregistrements)
- suivi opérationnel des déploiements (Parafe, Covadis)
- statistiques des dysfonctionnement techniques (ouverture Sas, certificats ...)
- réduction de la charge opérationnelle de suivi de l'équipe projet (~50k€/an)
- calculs avancés sur données de détail (100M d'enregistrement)
- opérationnel à la semaine voire / quart d'heure (Covadis/FPR2)
- 20 tableaux de bords, 144 vues



101001
01001000
01001000
0101101101
1100100101
111001101
0110
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE



DÉCISIONNEL ÉDUCATION ROUTIÈRE

statistiques nationales et départementales

- résultats des examens théorique et pratique du permis (20M d'enregistrements)
- activité des inspecteurs
- 30 tableaux de bords, 300 vues
- organisation dévops du dataprep au dataviz
- qualification en cours, production début 2019
- suites: cas datascience (prévisions volumes d'exams, positionnement autoécoles)

Projets 8

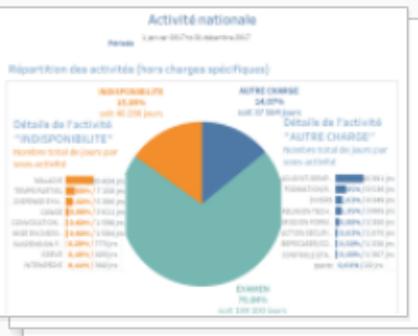
Classeurs 62

Vues 615

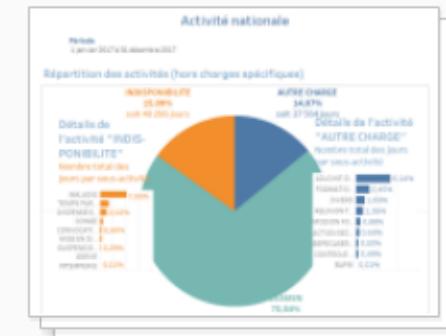
Sources de données 19

▼ 0 élément sélectionné

Trier par Nom (A-Z)



Activité nationale annuelle
15 vues ⭐ 0



Activité nationale annuelle
117 vues ⭐ 0



Activité permis sur plusieurs années
12 vues ⭐ 0



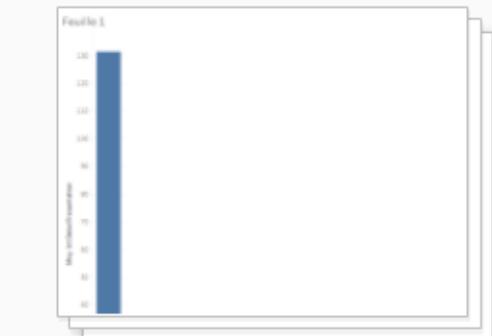
Activité permis sur plusieurs années
20 vues ⭐ 0



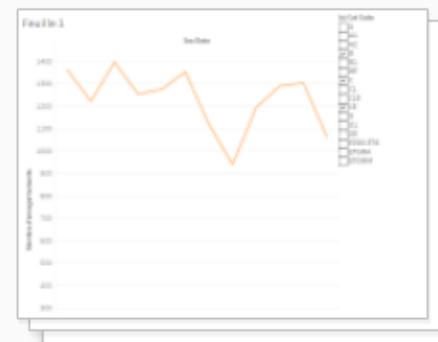
Analyse des CE
13 vues ⭐ 0



Analyse des CE
12 vues ⭐ 0



Feuille 3
24 vues ⭐ 0



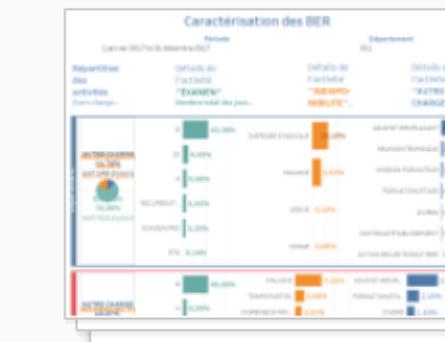
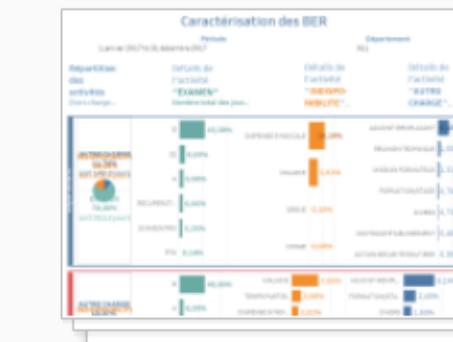
BacASableAdmin
55 vues ⭐ 0



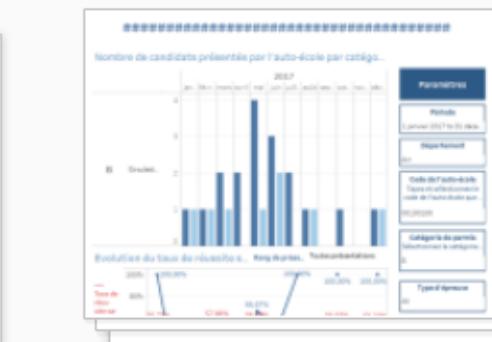
Candidats AE pour détection de fraude
30 vues ⭐ 0



Caractérisation BER
37 vues ⭐ 0



Caractérisation des auto-écoles
0 vue ⭐ 0



Caractérisation des auto-écoles
20 vues ⭐ 0



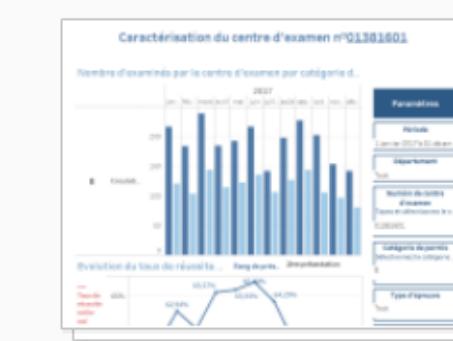
Caractérisation des candidats
0 vue ⚡ 0



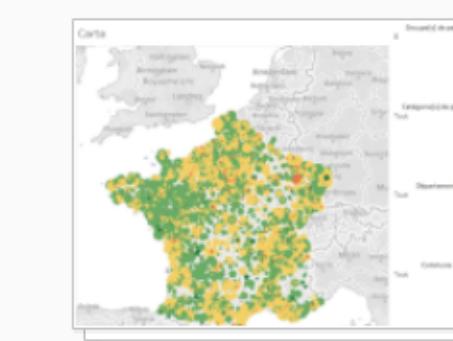
Caractérisation des candidats
10 vues ⚡ 0



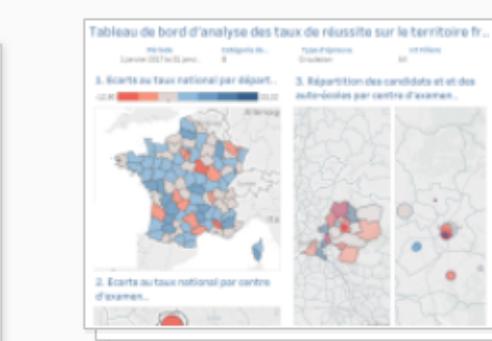
Caractérisation des centres d'examen
0 vue ⚡ 0



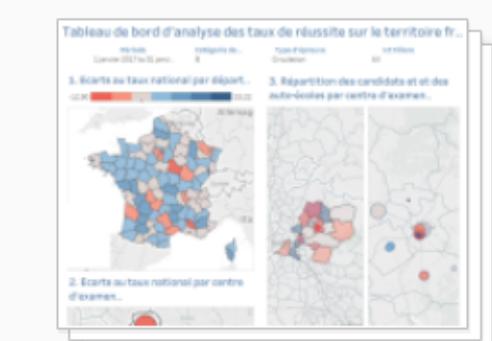
Caractérisation des centres d'examen
0 vue ⚡ 0



Carto_auto_ecoles
30 vues ⚡ 0



Cartographie des taux de réussite
11 vues ⚡ 0

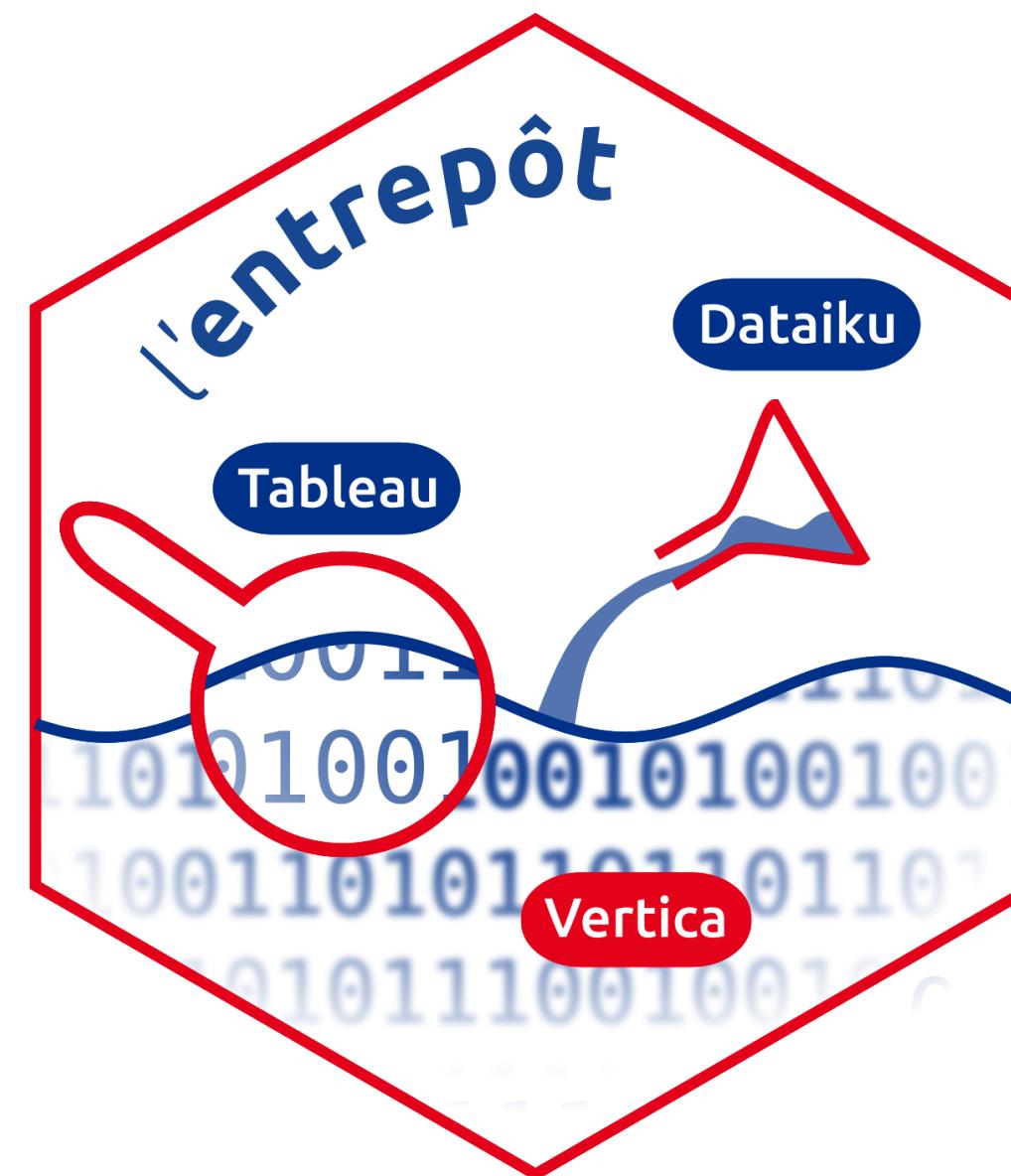


Cartographie des taux de réussite
0 vue ⚡ 0



INDUSTRIALISATION

DU DATALAB À L'ENTREPÔT





DATALAB

- de l'avant-projet jusqu'à qualification
- évolutions de l'architecture gros volume, nouvelle technologie

L'ENTREPÔT

production et évolutions itératives, sur le cloud MI



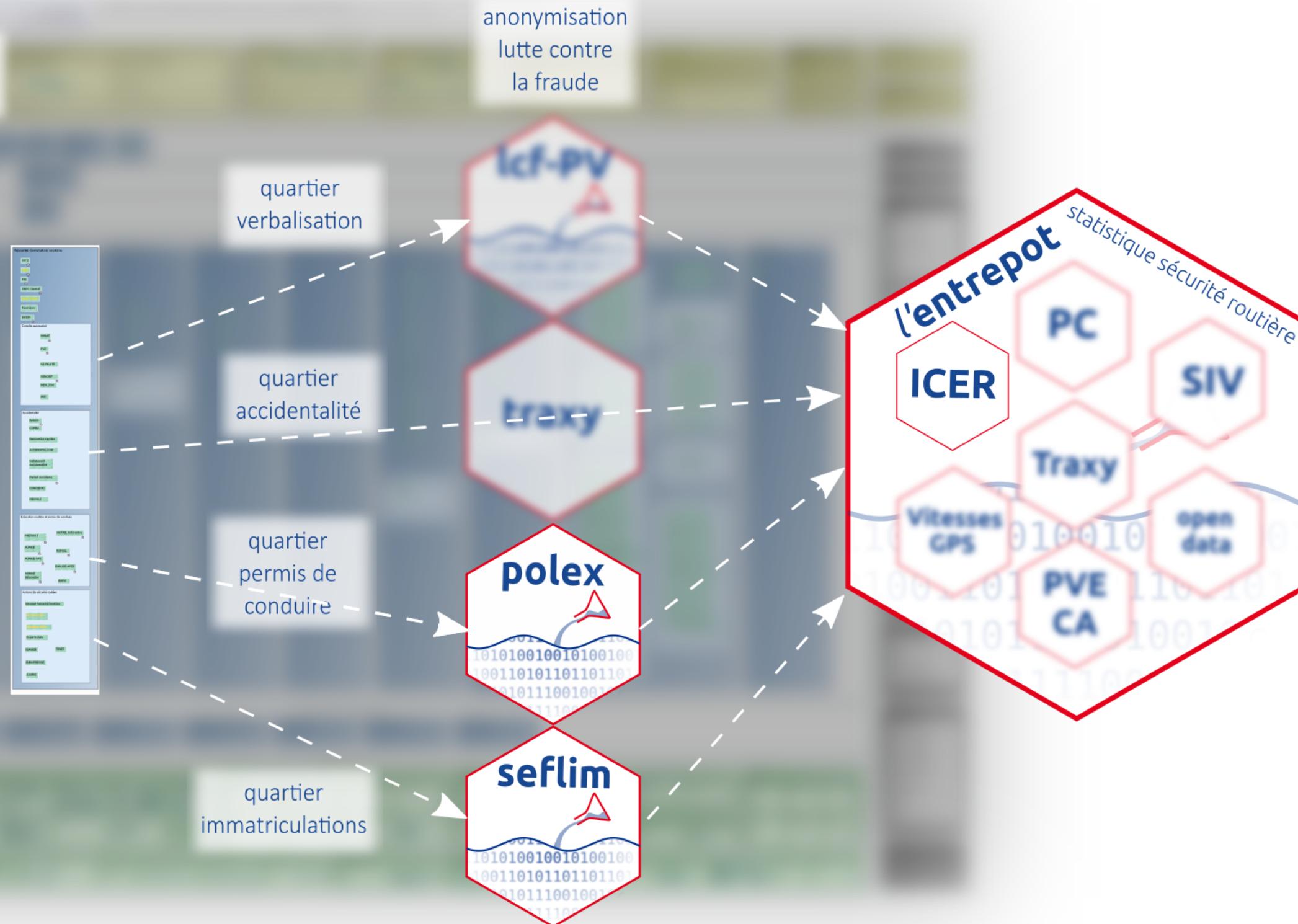
DÉPLOIEMENT & SÉCURITÉ DES DONNÉES

- un entrepôt statistique général
 - par zone fonctionnelle ou direction métier (e.g 360 conducteur et véhicules)
 - données anonymes ou pseudonymisées, croisements statistiques autorisés
- un entrepôt par donnée protégée
 - par zone fonctionnelle ou sous-direction : e.g. éducation routière (Polex) vs immatriculations (Selfim)
 - finalités de **lutte contre la fraude**, qualité et **anonymisation** des données pour chaque quartier fonctionnel, un entrepôt (e.g éducation routière, permis, immatriculations)
anonymisation



POS MI

zone fonctionnelle sécurité routière



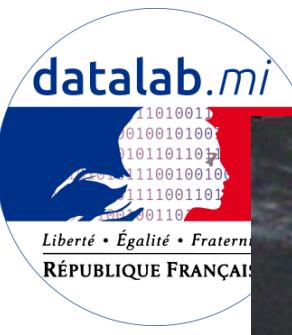


PERSPECTIVES



deep learning images

- détecter automatiquement marque et modèle de véhicules sur les images radars
- bloquer l'envoi de PV aux victimes de fausses doublettes
- architecture deep learning (GPU/TPU ? cluster Tensorflow ?)
- enjeux d'anonymisation (archivage, opendata/hackathons)



DT : [REDACTED]-2016 CSA : 00447 PK/PR : [REDACTED] IA : CNT CSA NR : RD1082 NPV : 4 NV : 1 VOIE INF : VOIEX

VM : 65 km/h VLA : 50 km/h VILLE : MONTROND LES BAINS CP : 42210 COND : Standard SENS : RAPP

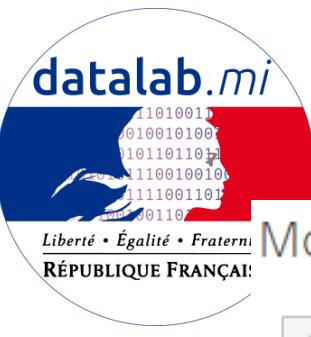
SC : [REDACTED] : -



POC VITESSES

1000 milliards d'enregistrements

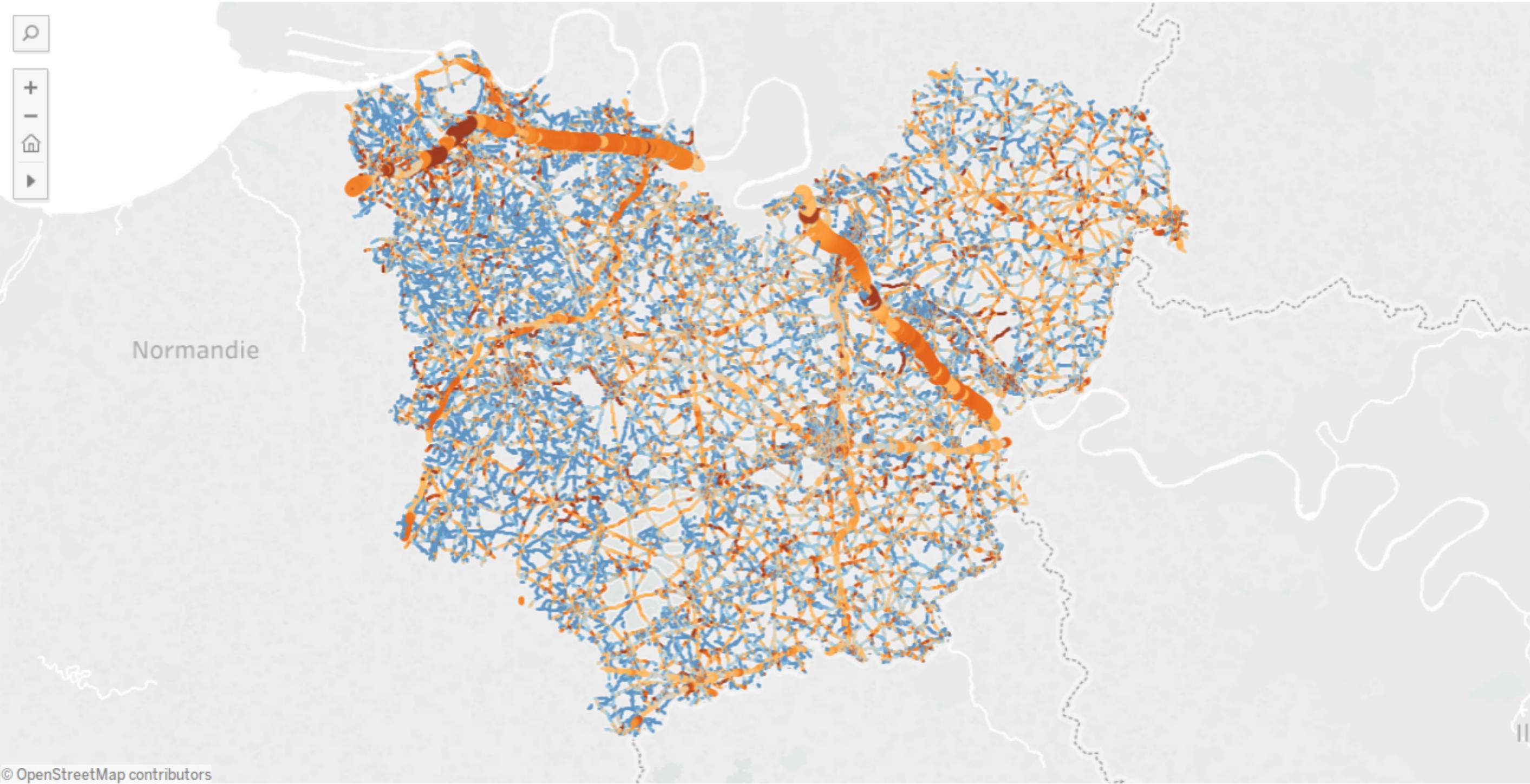
- définir les métriques pertinentes (dépassements de vitesse, ...)
- évaluer l'impact des politiques de sécurité routière sur la pratique
- enjeux d'architecture (vertica, Hadoop)
- décisionnel géographique gros volume (galligeo, custom, ... ?)



Liberté • Égalité • Fraternité

RÉPUBLIQUE FRANÇAISE

Moyenne des vitesses pour les véhicules dépassant la vitesse limite autorisée (Eure)



© OpenStreetMap contributors



DÉPLOIEMENT DE L'ENTREPÔT

- finalisation : éducation routière, contrôle frontière
- en cours: Actes, SID DGOM (+geo business)
- opportunités:
 - sécurité routière :accidents / verbalisations / immatriculations / permis
 - sécurité civile : infosdis



QUESTIONS ?

[télécharger la présentation en version pdf](#)