LA DATA SCIENCE COMME OUTIL DE VALORISATION DE LA DONNEE

SÉBASTIEN QUINAULT / S1 2021/ M1 SARADS

Qui suis-je ?

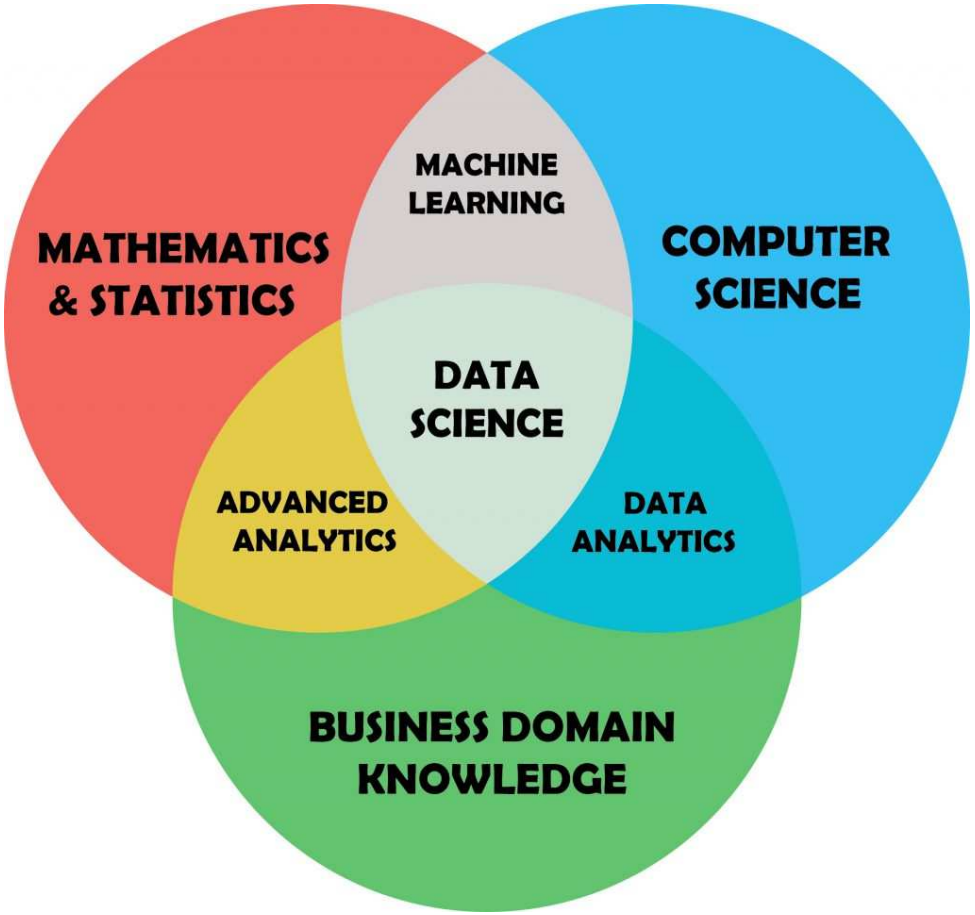
Sébastien QUINAULT
Data scientist – Groupe Covéa

Mon parcours :
DUT STID -> Maitrise GIS
Développeur BI
Chargé études statistiques
Data analyst
Data scientist

https://www.linkedin.com/in/sebastien-quinault

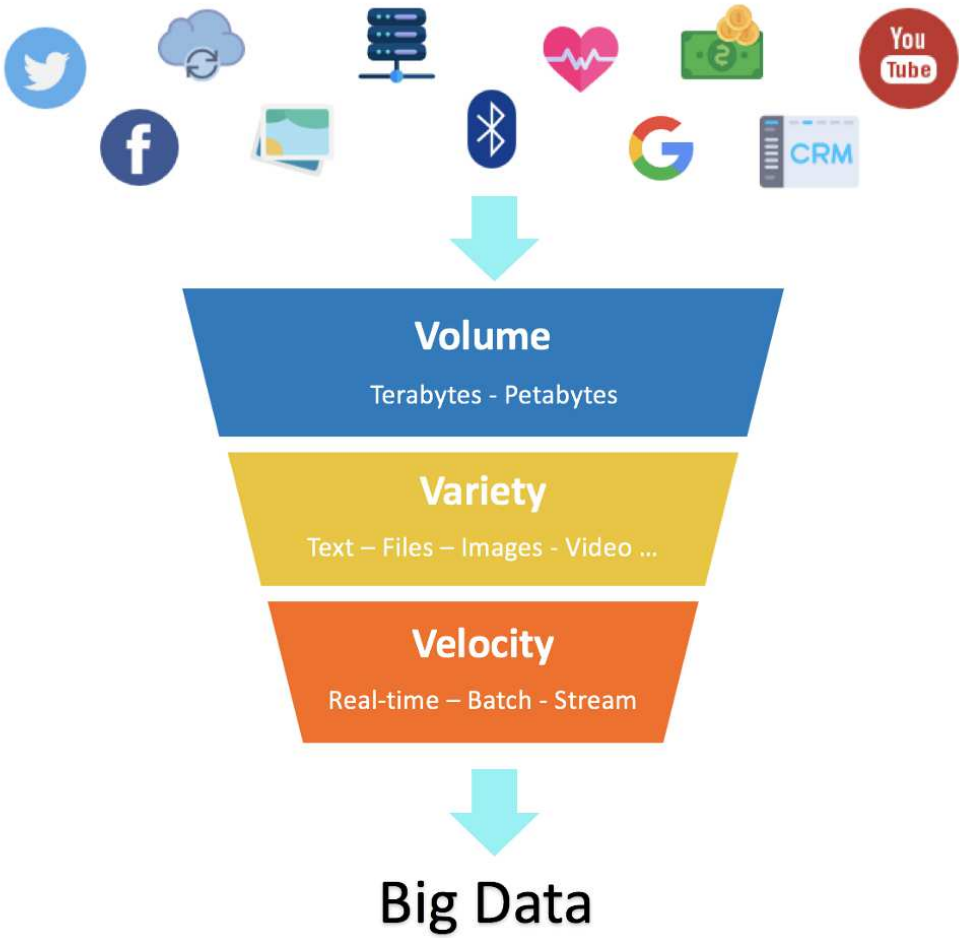
DATA SCIENTIST ?

DATA SCIENTIST




L’entreprise « est » data centric
La valeur se créé a partir de la donnée

L’impact du big data



Comment produire cette valeur ? Grace a vous spécialistes de la donnée

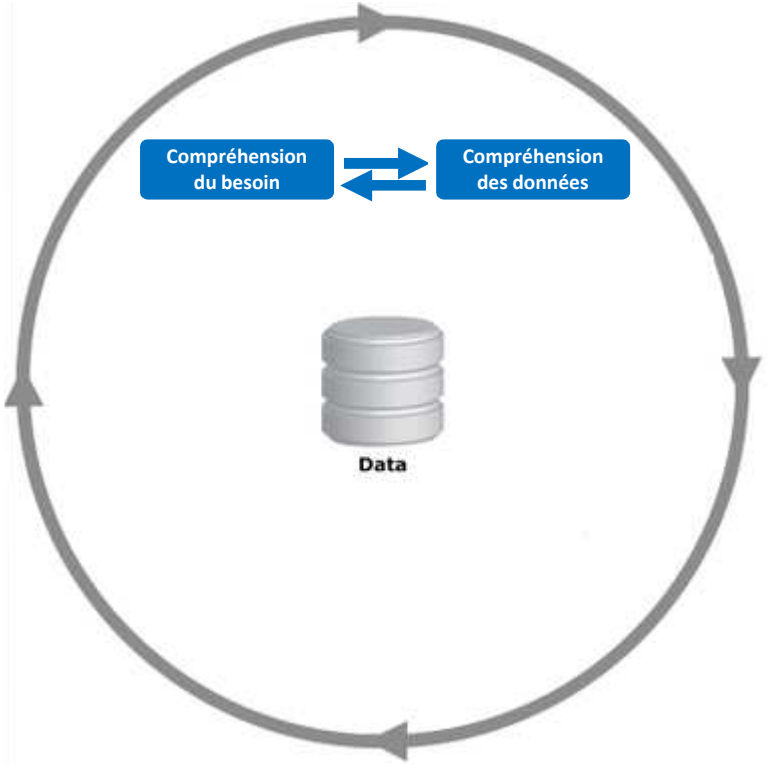
En utilisant les bonnes méthodes et les bons outils

Une approche standard du marché

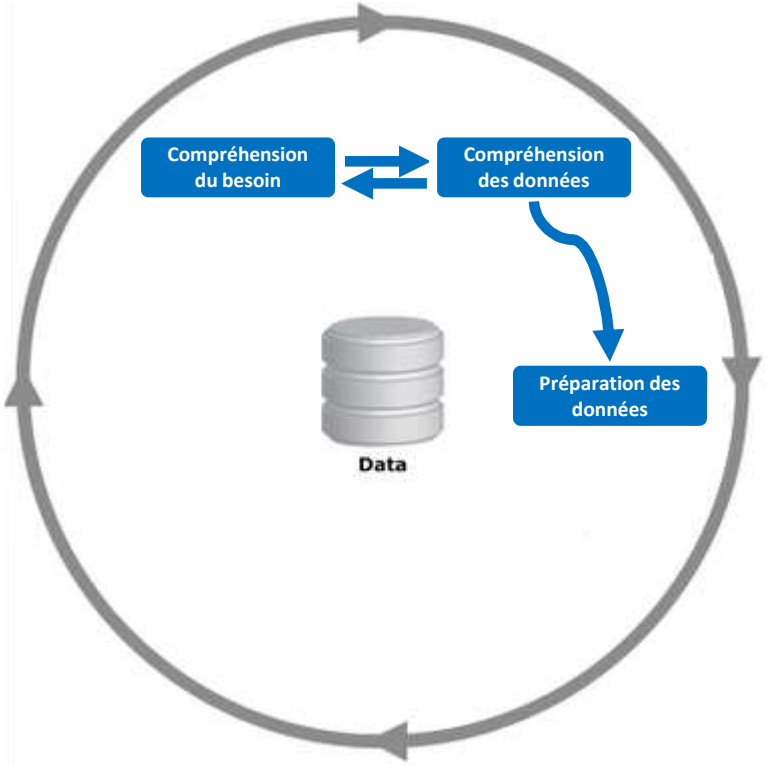
CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

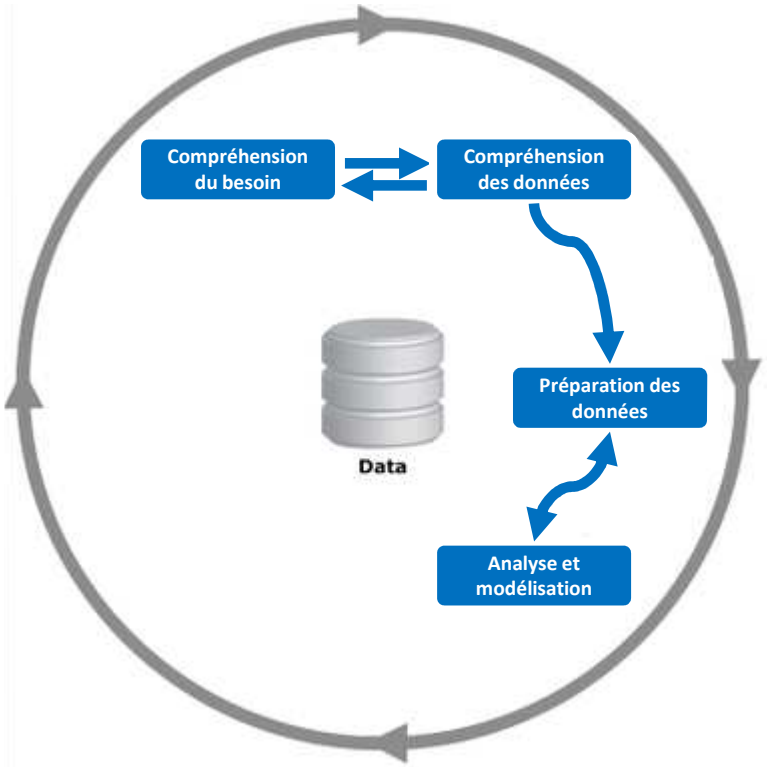
Comprendre le besoin métier et les données disponibles



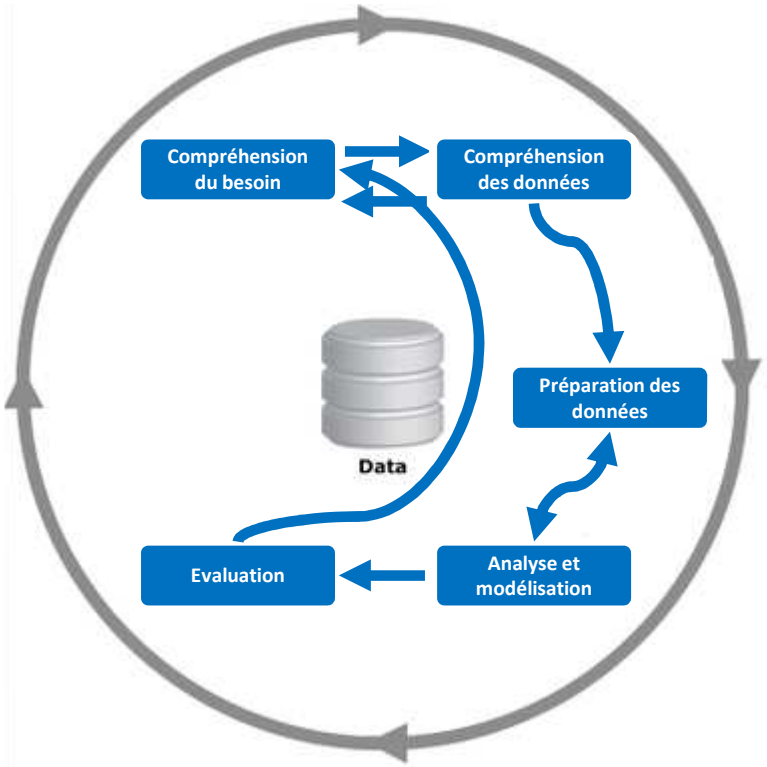
Préparer les données



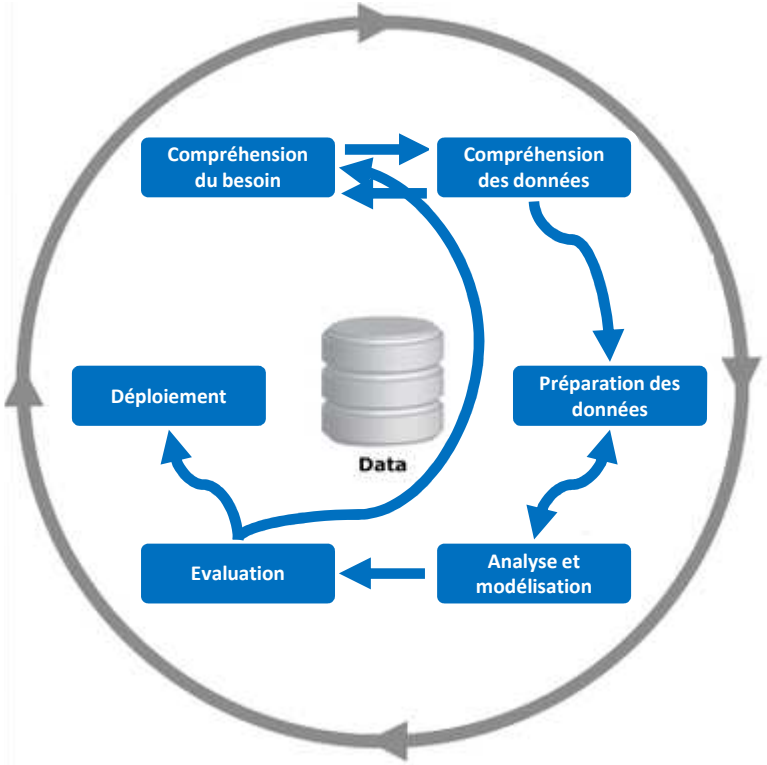
Analyser et modéliser



Evaluer vos résultats

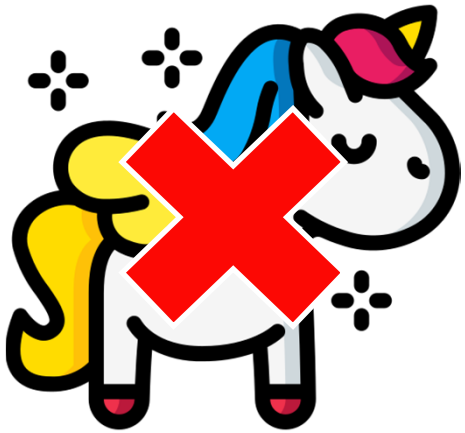


Délivrer un produit



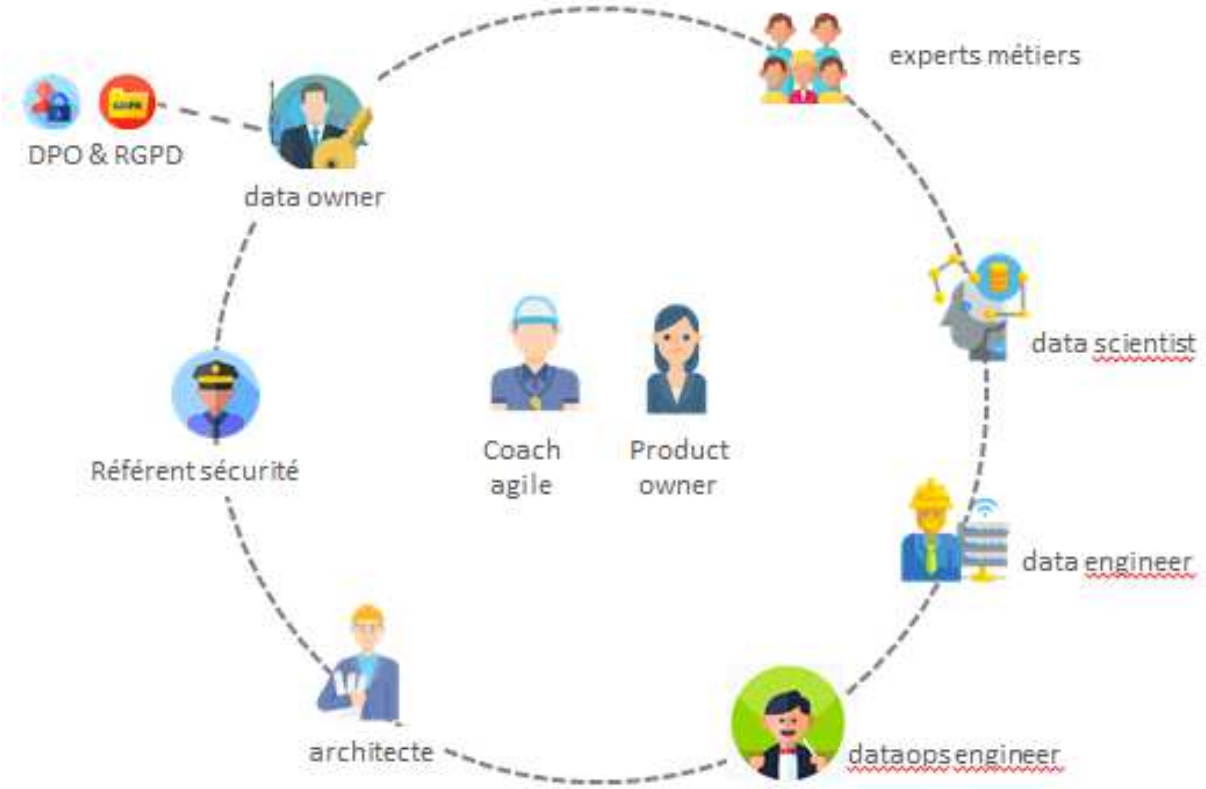

Une transformation dans
l'approche de la donnée

La data est un sport d’équipe



Beaucoup d’outils
Beaucoup de données
Beaucoup de méthodes

Une équipe composée de spécialistes



Détail des roles

Coach Agile
Le travail du coach agile consiste à épauler une équipe dans son parcours vers l'agilité. L'objectif de ce poste est d'avoir comme output des résultats améliorés et mesurables.
Il accompagne ses interlocuteurs sur les bonnes pratiques de définition des besoins. Il agit également sur la planification et les pratiques de développement. Globalement, il assure l'avancement des projets dans toutes les phases du cycle de développement.

Data engineer
Prépare la donnée pour les data modeler, package le code informatique (data collecte, le modèle, le monitoring…) pour opérationnaliser le produit DS
Il a une vision du patrimoine de données en s'appuyant sur les data owner, acteur du patrimoine de données.

Datascientist : répond à des enjeux métiers à l'aide des méthodes de modélisation prédictive et analytique sur des donnée variées et/ou volumineuses et en s'appuyant sur une infrastructure dédiée

Data Ops engineer
Facilite/automatise la mise à disposition d'environnements (infrastructure, des données de test, des ressources machines (CPU, RAM…) , version du code informatique). Orchestre et automatise les pipelines (ex : data collect, preparation, train/test, prediction, monitoring) . Conduit le projet vers la mise en production dans le du cadre informatique est respecté.

Détail des roles

Product owner
Est le point d'entrée du projet côté métier. Il doit être appétent/acculturé au domaine de la data science.
Il est un membre de l'équipe à part entière. Son rôle est de comprendre les problèmes de fond du métier et de valoriser la résolution de ces problèmes. Il guide l'équipe dans la compréhension du problème. Il récupère les feedbacks des utilisateurs pour ajuster la trajectoire du produit.

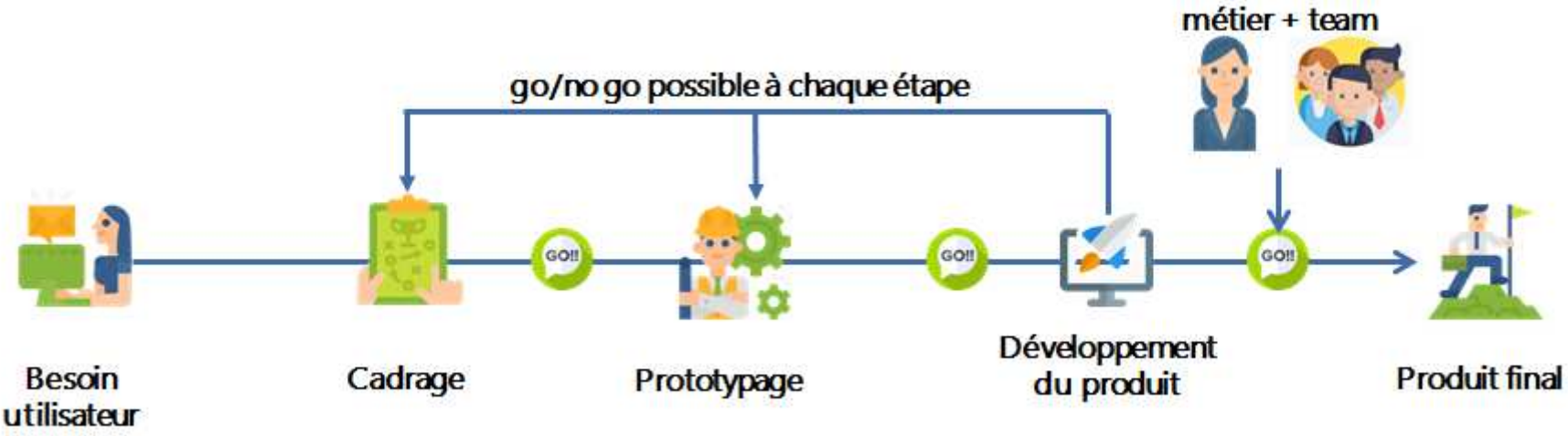
Experts métiers
Représente les personnes sur lesquelles la team peut s'appuyer pour des besoins d'expertise et de compréhension du business. Ils sont potentiellement les futurs utilisateurs du produit.

Data owner
Est garant de l'alignement stratégique et des politiques de gouvernance et de qualité des données sur un domaine de données dont il est propriétaire. Il accompagne les différents rôles du projet pour trouver des données, les comprendre et les utiliser correctement. Il est garant de la qualité des données rentrant dans son périmètre. (vision cible)

Architecte
Accompagne, supporte les projets dans la définition et la formalisation de l'architecture fonctionnelle, applicative, technique de leur projet.
S'assure de la cohérence du patrimoine applicatif créé / modifié à l'occasion du projet avec les principes et standards définis dans le cadre d'architecture et avec la cible du SI

En synthèse, tout le monde a sa place dans la data quelques soient vos points forts/points faibles

Une équipe organisée dans un cycle agile



Concrètement…

L'utilisateur final, direction de la conformité, souhaite automatiser la détection des cas frauduleux sur les sinistres matériels auto. Il estime que la détection de cas avérés lui permettrait de gagner 5% de la charge sinistre. La solution devra répondre au RGPD et aux normes de sécurité informatique.

L'équipe imagine une solution qui permettrait de faire un scoring à chaud des dossiers sinistres s'appuyant sur une double approche : apprentissage supervisé à partir de cas existants et une approche non supervisée avec la détection « d'anomalies ». La solution devra être rapide et s'interfacer avec le SI Sinistre. L'équipe s'appuiera sur les data owners pour s'assurer du cadre d'utilisation des données.

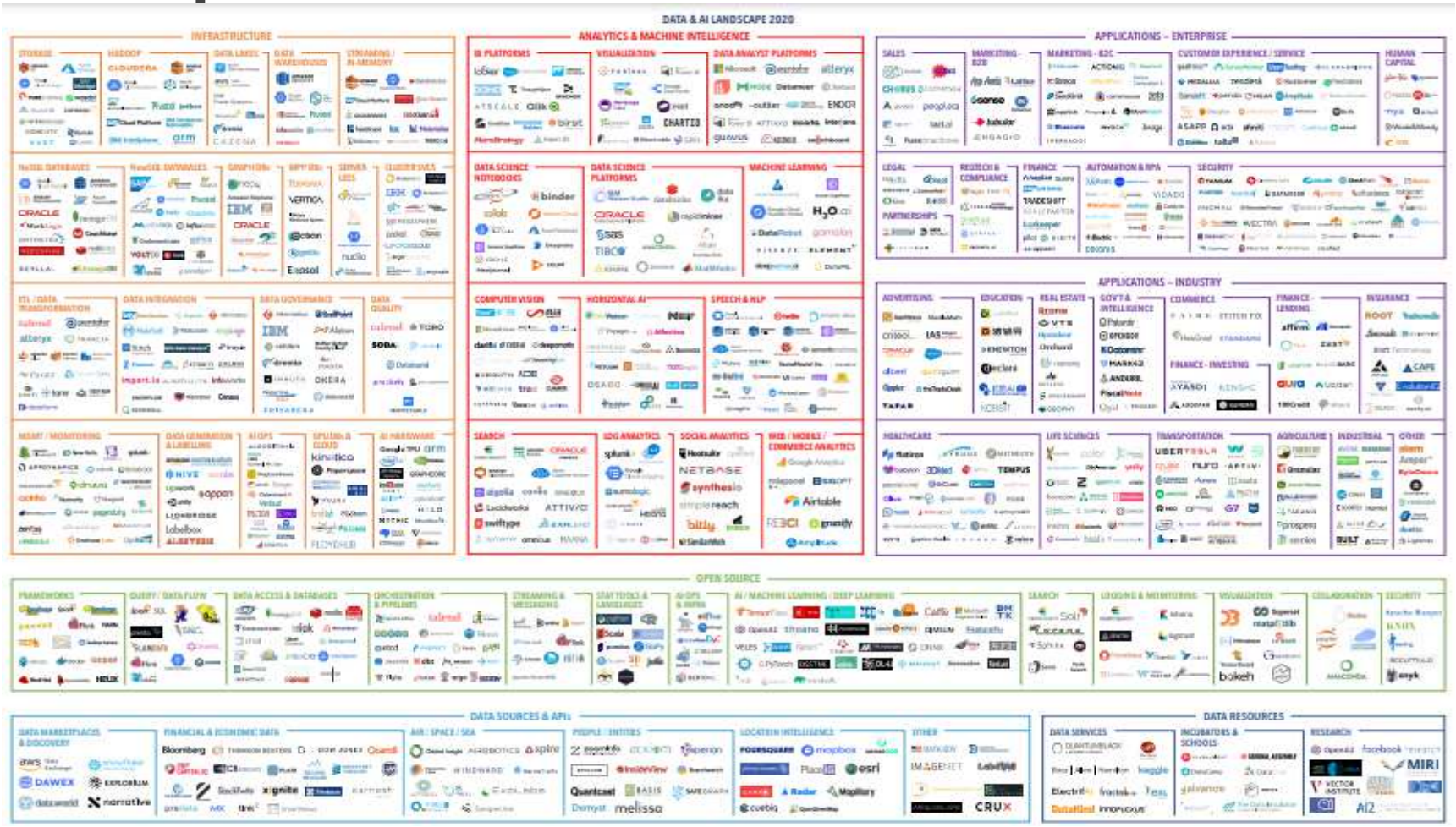
L'équipe a collecté un premier jeu de données, étudie la qualité des données et entraine les premiers modèles. Les cas potentiellement frauduleux détectés sont présentés au métier qui valide l'approche et la pertinence des résultats. Le produit peut passer en phase de développement

Le produit est développé sous la forme de sprints. Un sprint 0 a pu être réalisé pour mettre en place la team et s'approprier les technologies. Assez rapidement, une version simple de l'algorithme est déployé en production pour vérifier la capacité à s'intégrer dans le SI. Au fil du temps, l'algorithme gagne en précision et est redéployé de manière régulière. L'utilisateur dispose d'outils permettant de suivre les détections de fraude et les gains réalisés.

Le produit est maintenant pleinement opérationnel. Il fait partie du patrimoine applicatif et est suivi comme tout applicatif standard. Le métier souhaite désormais développer le même produit pour les déclarations de sinistre habitation.

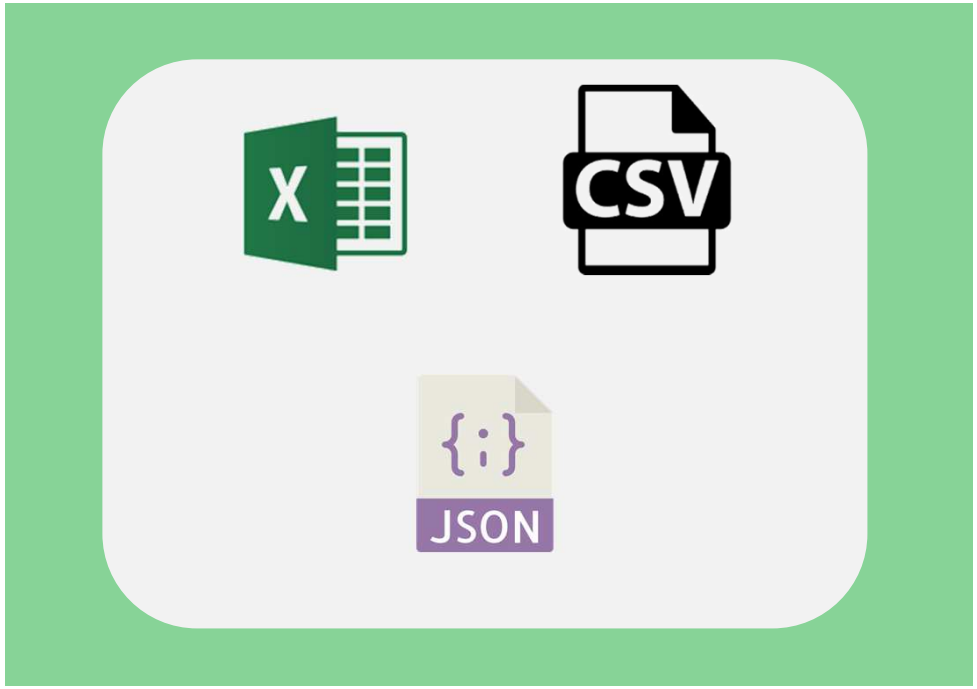
Des outils

Mais lesquels ?

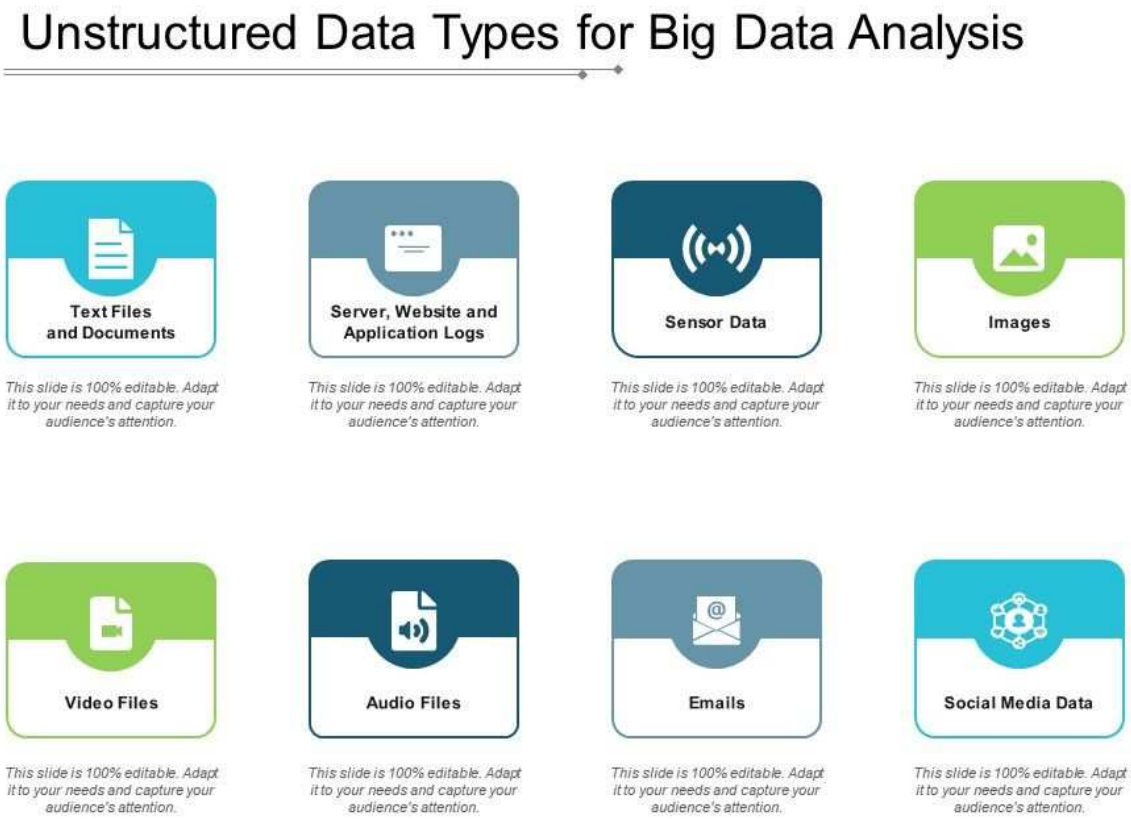


Mais lesquels ?

Comprendre l’écosystème – les données (1/2)

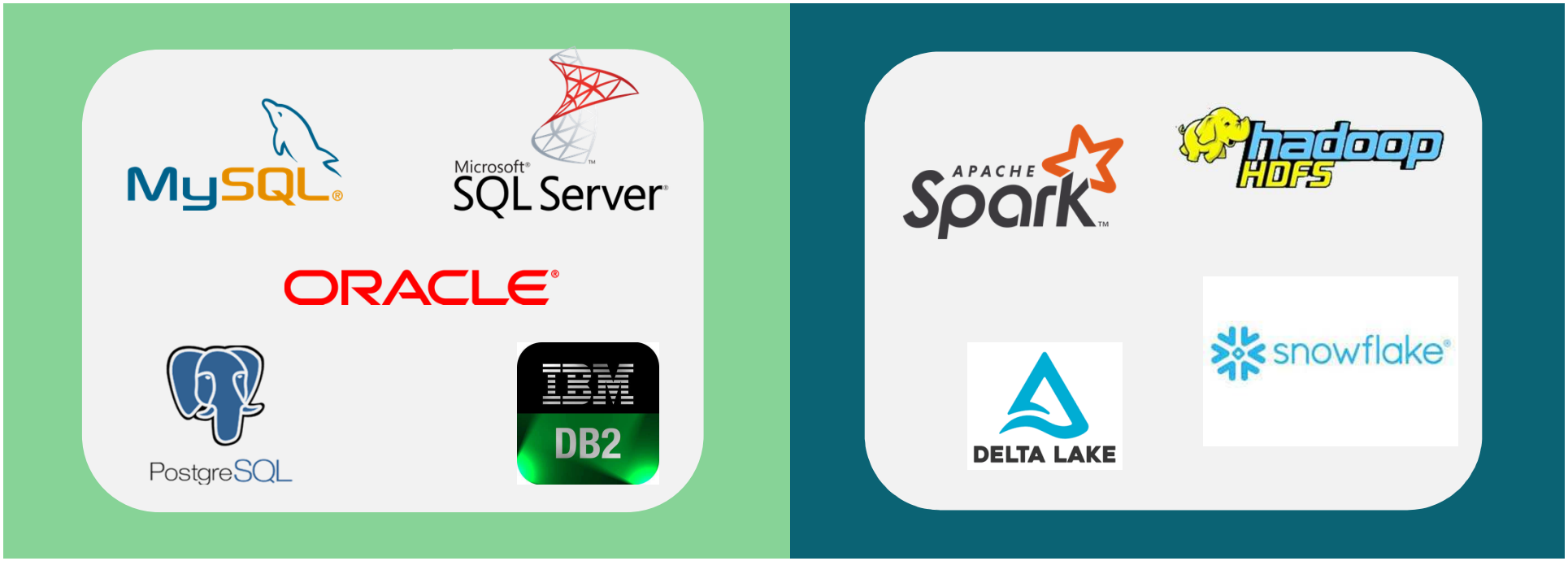


Comprendre l’écosystème – les données (2/2)



Accéder aux données

SGBD vs Big data



Accéder aux données – quel langage ?

La valeur sure !

Comprendre la donnée

En programmant

Avec des outils ‘user-friendly’

Préparer la donnée

En programmant

Avec des outils ‘user-friendly’

EXPLORATORY

alteryx

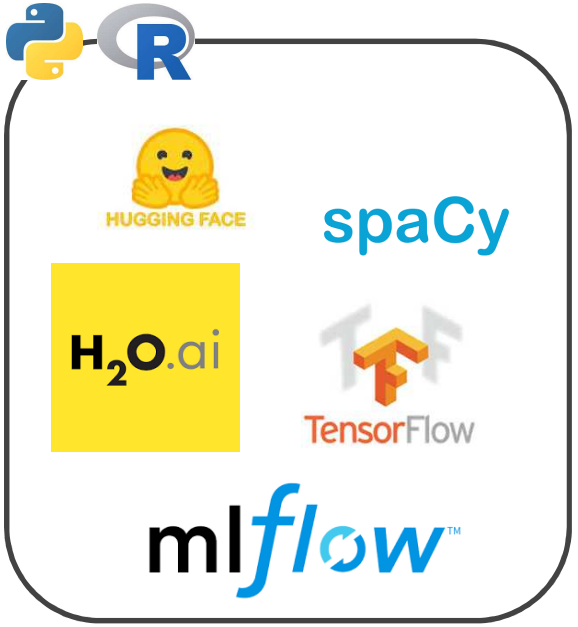
Analyser et modéliser

En programmant

Avec des outils ‘user-friendly’

Analyser et modéliser

Les frameworks de programmation

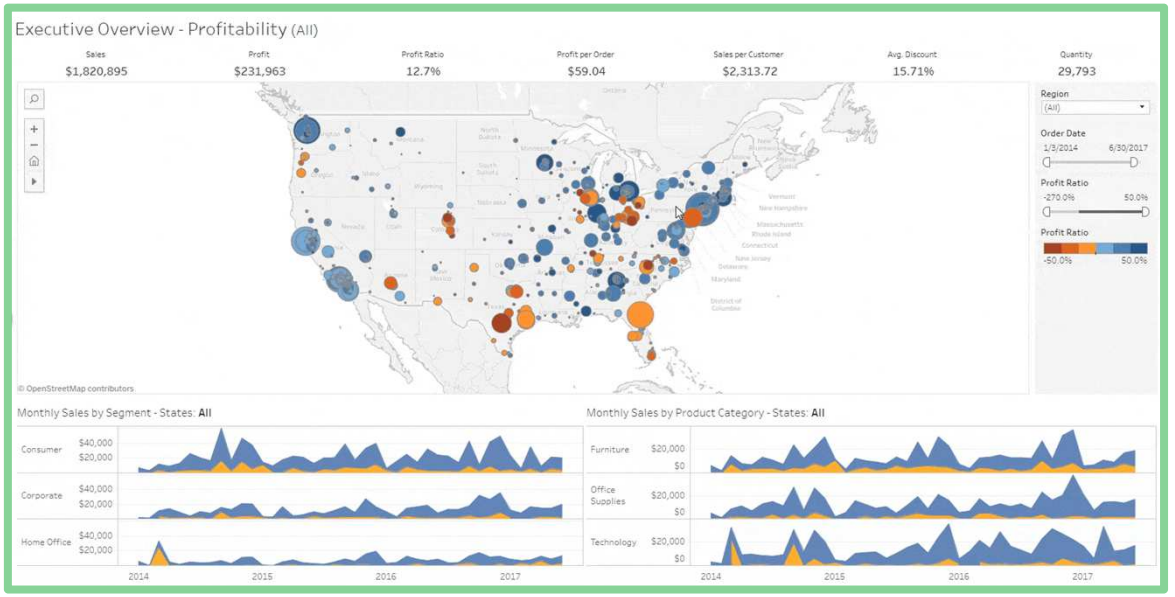


Présenter les résultats

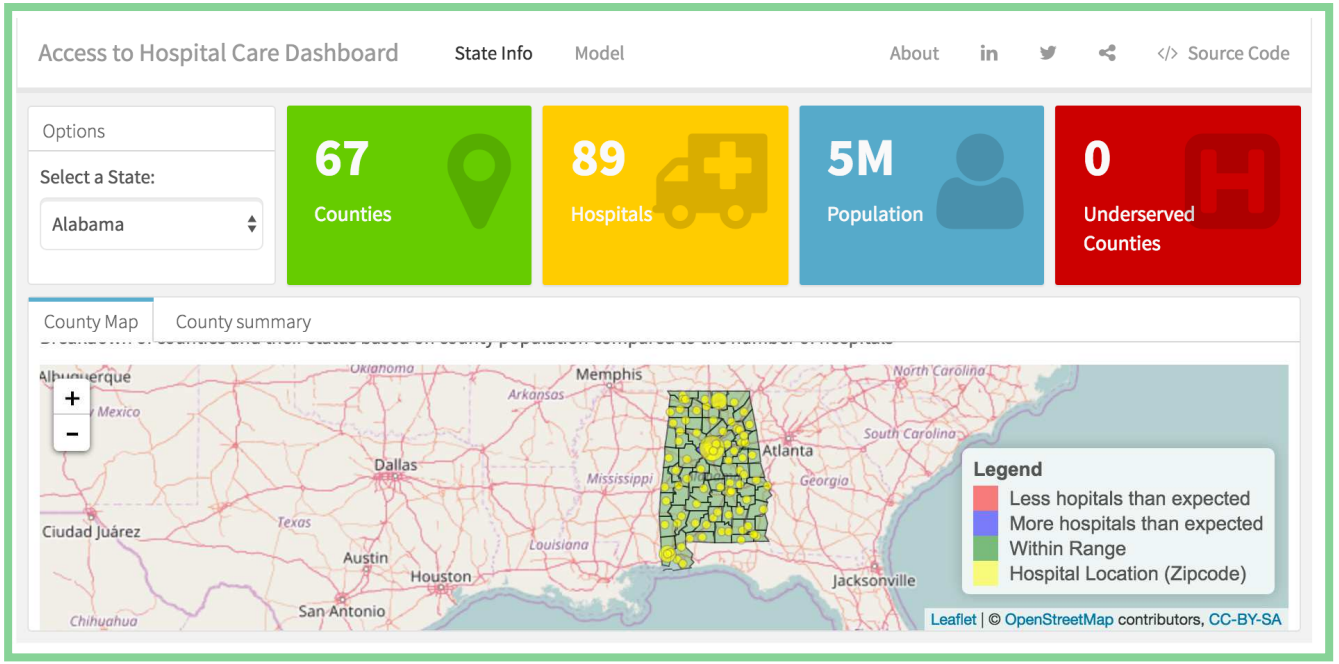
Qlik

+ a b | e a u

TOUCAN TOCO



Présenter les résultats

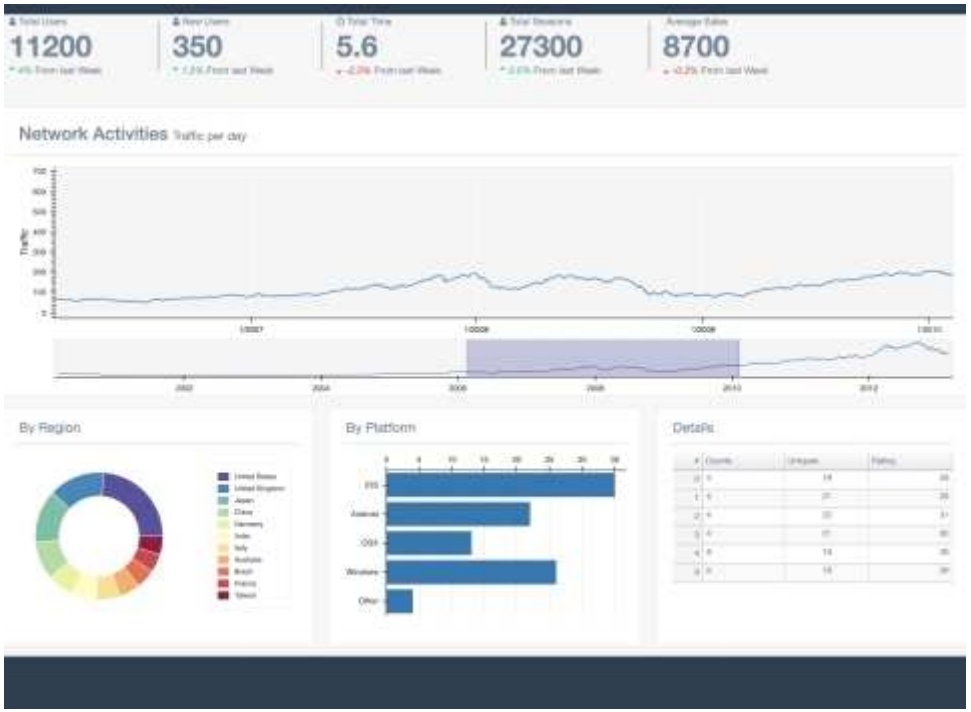


Présenter les résultats

Dash by plotly

bokeh

Streamlit



Déployer vos résultats

C’est la partie délicate

Pas de « silver bullet »

Une seule solution: impliquer l’IT dès le départ

A STREAMING

Enrichir vos analyses

data.gouv.fr

géoservices

Gérer vos projets


Les grandes tendances

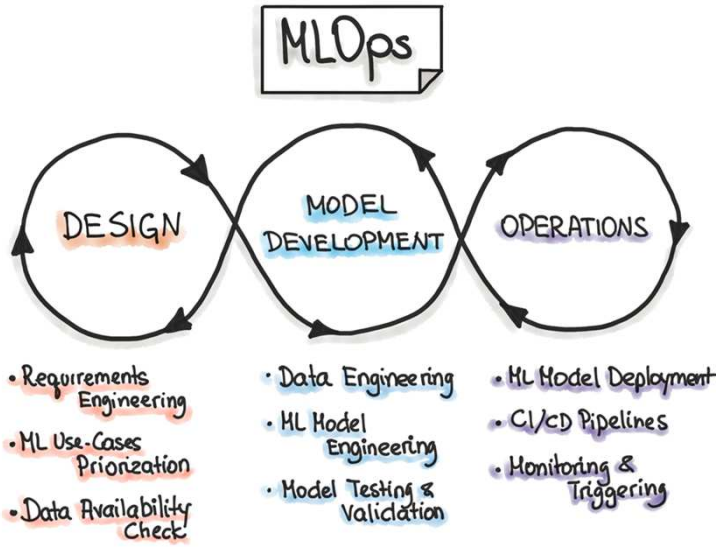
L’IA

L’IA est un concept : la machine est capable d’apprendre

wikipedia : Ensemble des théories et des techniques mises en oeuvre en vue de réaliser des machines capables de simuler l'intelligence

MLOPS

Processus transversal, collaboratif et itératif qui opérationnalise la Data Science

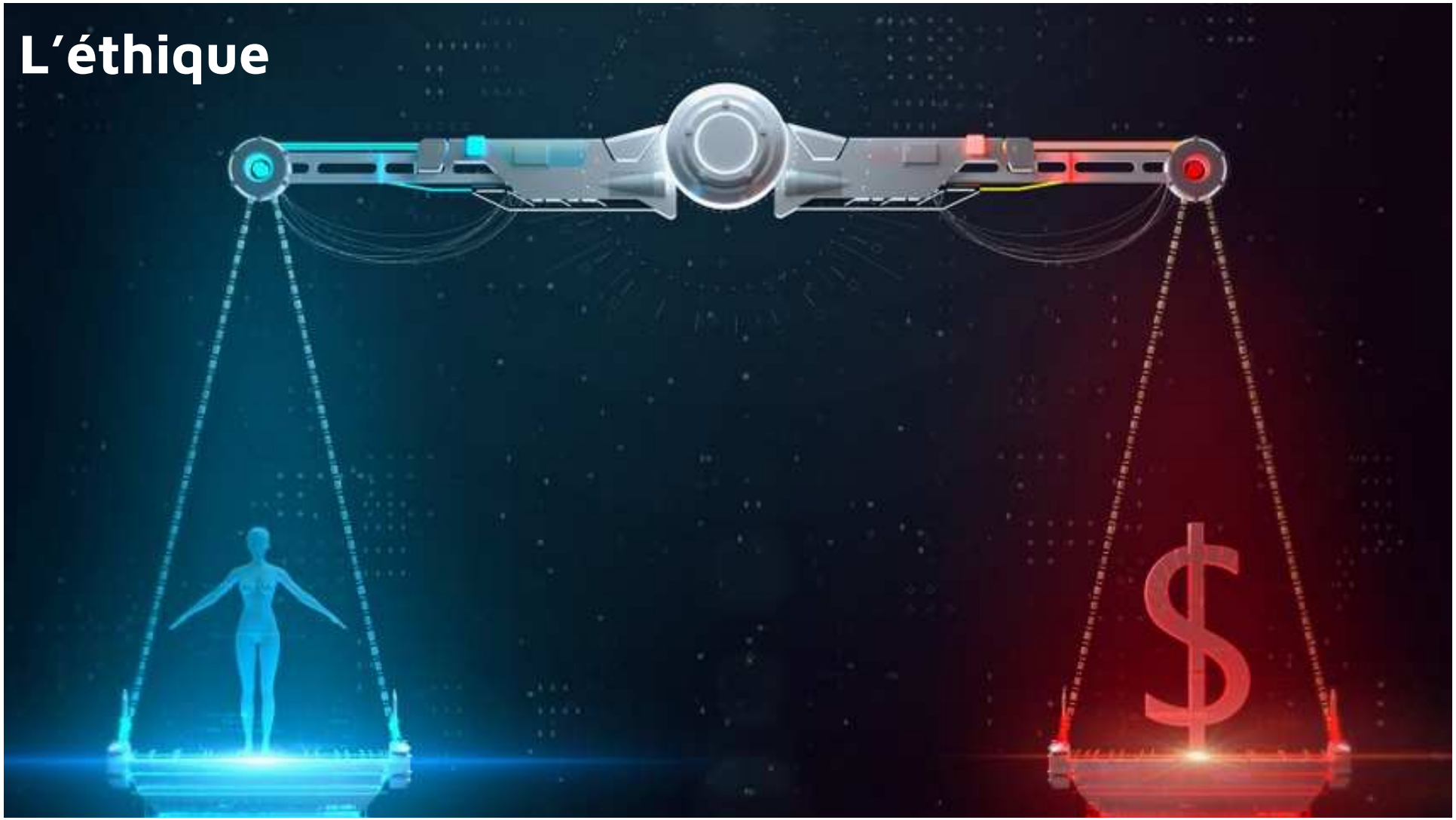


Cloud

Accès a des services informatiques (serveurs, stockage, mise en réseau, logiciels) via Internet (le « *cloud* » ou « nuage ») a partir d'un fournisseur

Le contexte réglementaire

L’éthique



Se former


DataCamp


DataScientest • com


bloggers


KDnuggets


FUN
FRANCE
UNIVERSITÉ
NUMÉRIQUE
Se former en liberté


coursera


stackoverflow


Papers With Code


kaggle


sas
VS
R

Quelques différences

Modèle économique : licensing vs open source

Courbe d’apprentissage + IDE

Disponibilité des nouveautés

Restitution des résultats

Support éditeur / communauté / documentation

Patrimoine applicatif existant


VS

Quelques différences

Deux langages open source avec 2 objectifs différents

Python + maintenable, + facile à apprendre

Présentation des résultats

API de machine learning

Les packages R

Comment choisir ?