
CREER DE LA VALEUR
A PARTIR DES
DONNEES
EN 2021
SÉBASTIEN QUINAULT / S1 2021/ M1 SARADS

Qui suis-je ?

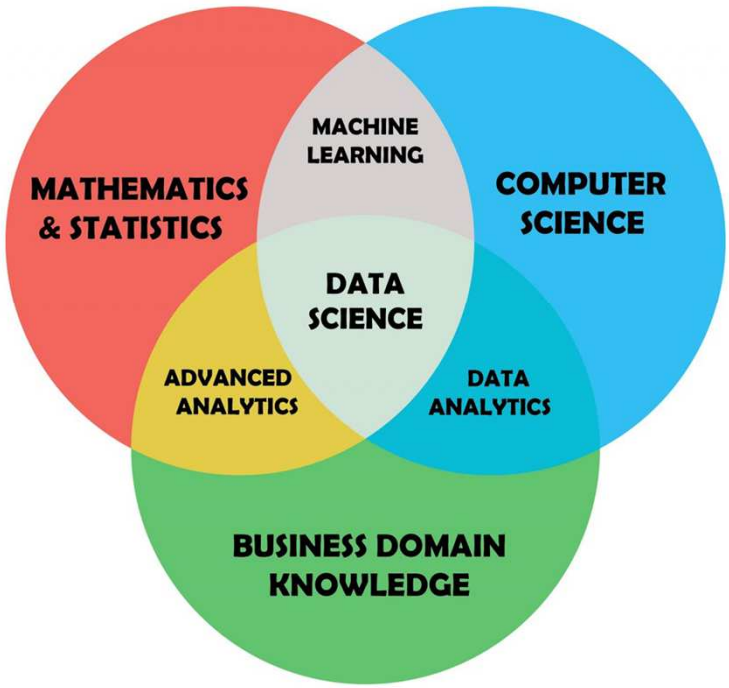
Sébastien QUINAULT
Data scientist – Groupe Covéa

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

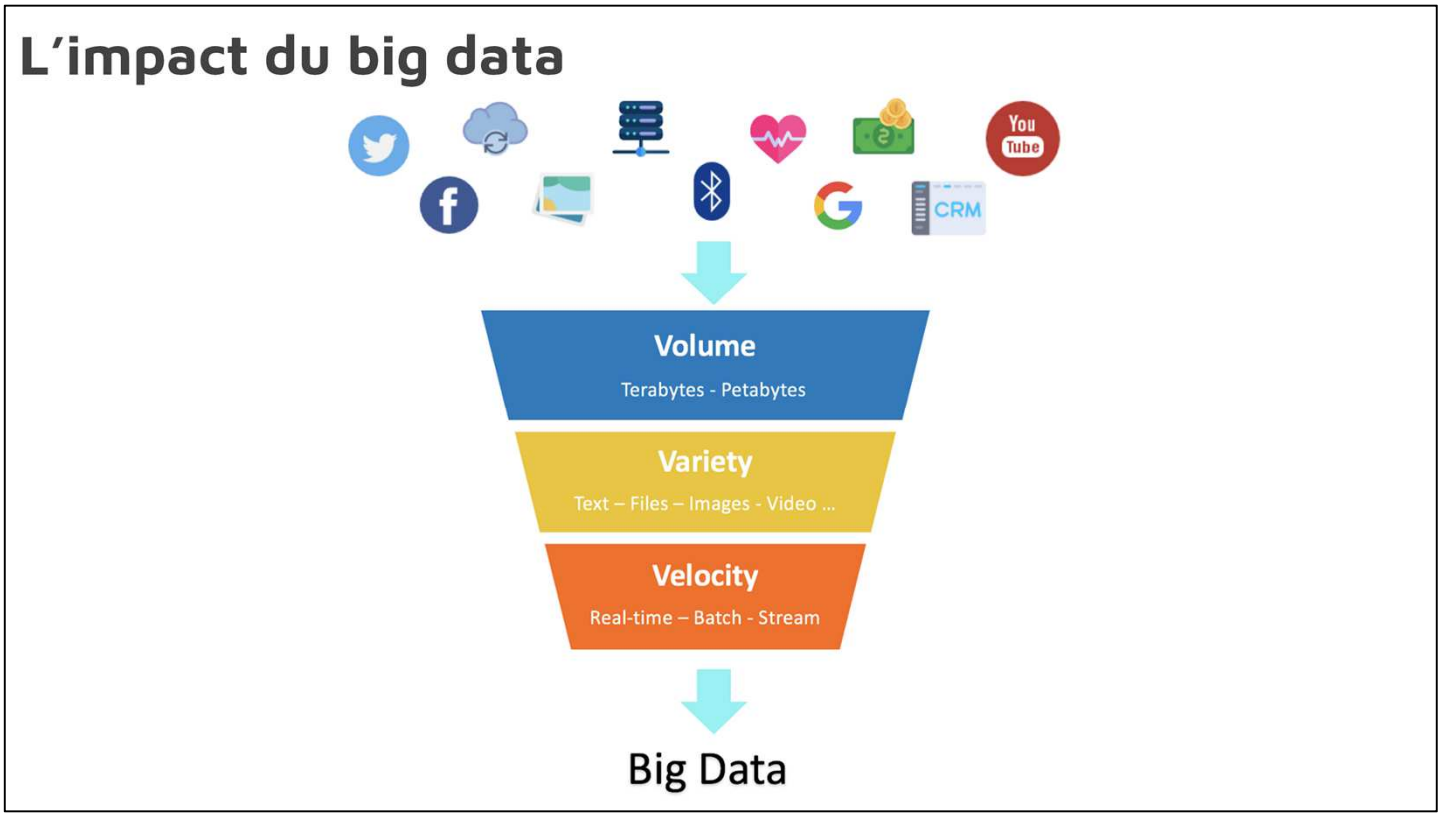
https://www.linkedin.com/in/sebastien-quinault

DATA SCIENTIST ?

DATA SCIENTIST




En 2021, l'entreprise « est » data centric
La valeur se créé a partir de la donnée



Pourquoi le domaine de la data est-il si tendance ?
Qu'est ce qui fait qu'on en parle plus ajd.

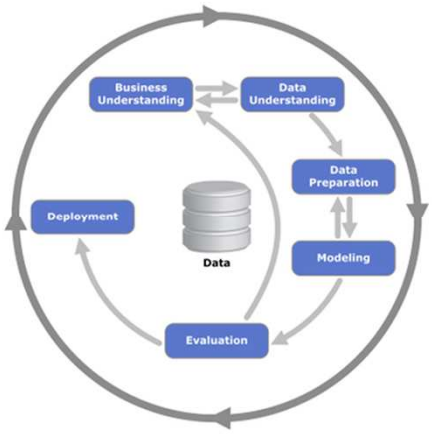
Tout d'abord le big data : les 3 V

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

En utilisant les bonnes méthodes et les bons outils

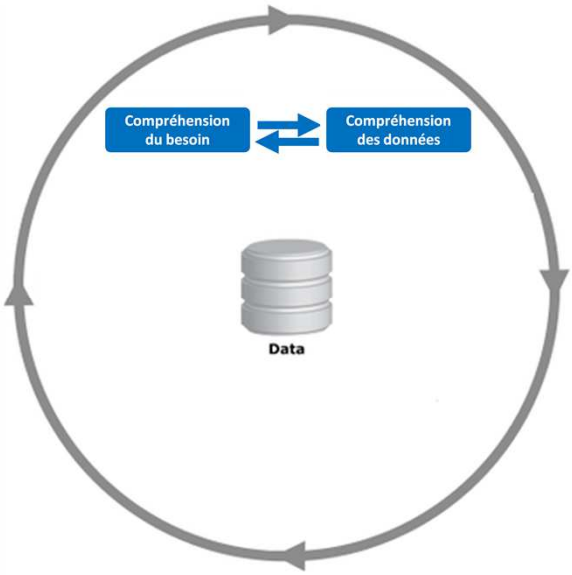
Une méthode qui fonctionne

CRISP-DM



CRISP-DM signifie **Cross Industry Standard Process for Data Mining**[1]. Il s'agit d'un Modèle de Processus de data mining qui décrit une approche communément utilisée par les experts en data mining pour résoudre les problèmes qui se posent à eux

Comprendre le besoin métier et les données disponibles



1ere étape : comprendre la question posée, échanger avec les équipes métier

2ème étape :
le référencement et la collecte des données. Les sources peuvent être :
- internes ou externes
- documentées ou non
- de format différent

Il est donc primordial de connaitre l'origine des données (producteur) ainsi que le sens des données (au travers d'un dictionnaire), leur date de fraicheur…

L'accès aux sources peut se faire au travers de nombreuses technologies : fichier plat, SGBD, API… L'idéal est de maitriser un langage de programmation (r, sas, python) car il permettra d'accéder à une très grande variété de données.

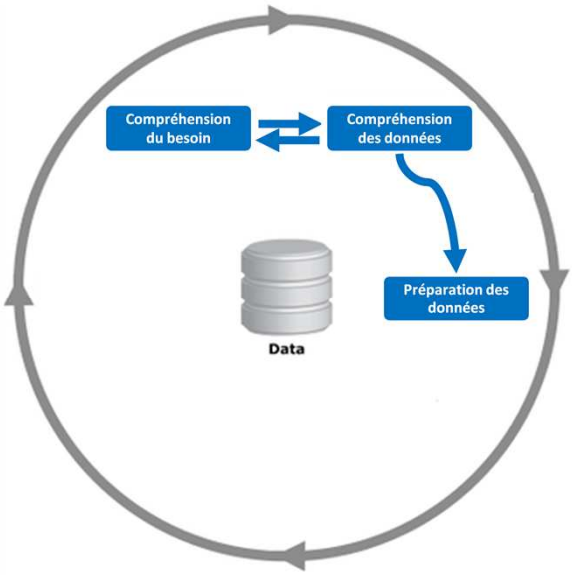
Des outils se développent aujourd'hui permettant d'accéder à diverses source en mode clique bouton

L'exploration des données est une étape à ne pas négliger. C'est grace à elle qu'on va se familiariser avec son jeu données :
- De combien de variables je dispose
- Quel est le type de ces variables : quali, quanti, logiques, géographiques….
- Quelles sont les modalités/valeurs de chacune des variables :
    - une variable avec 1000 modalités sera peut etre difficile à exploiter au sens statistique
    - des valeurs aberrantes apparaissent-elles ?

- quelle est la complétude des données : une variable avec 95% de manquants ne sera peut etre pas exploitable
- Calculer des stats univariées / multivariées
- Calculer les corrélations entre les variables (coef de corrélation, V de Kramer…)

Important : visualiser vos données !

Préparer les données



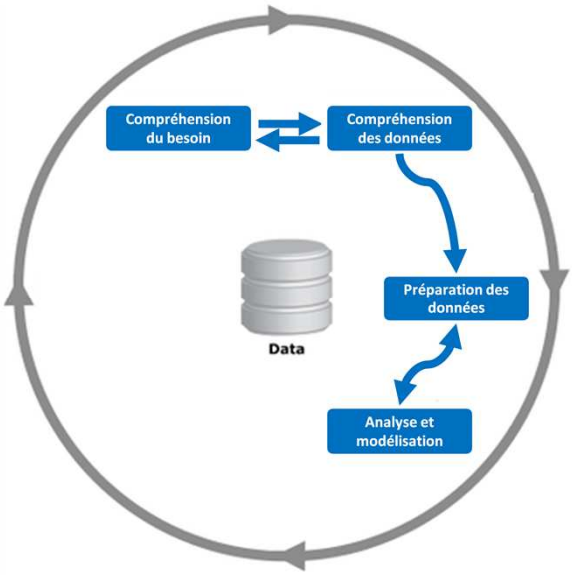
Le traitement et l'enrichissement des données est un élément différenciant entre les data scientists.
Aujourd'hui tout le monde à accès aux derniers algorithmes, à de la puissance de calcul (AWS).
Mais vous être l'intelligence derrière la machine, vous savez quelles données choisir, comment les transformer, comment les enrichir.
Exemples :

- Vous saurez comment regrouper des modalités pour en obtenir 25 au lieu de 1000
- Vous saurez définir la bonne stratégie de remplacement des valeurs manquantes (mode, moyenne, médiane, kmeans, glm …)
- Vous saurez comment traiter vos valeurs aberrantes : les supprimer, les plafonner, les catégoriser…
- Vous saurez construire de nouvelles données à partir des données existantes :
    - vous avez une adresse, transformez la en code iris puis enrichissez votre record avec les données Insee
    - vous avez une date : transformez la en jour ouvré, jour férié, jour de la semaine, trimestre, mois….

Essayez du mieux possible d'apporter un contexte à vos données. La clé du traitement de données reste l'intelligence humaine

Ces 3 phases concentreront 80% de votre temps de travail. A vous d'acquérir les bonnes méthodes, les bons outils pour optimiser ce temps

Analyser et modéliser



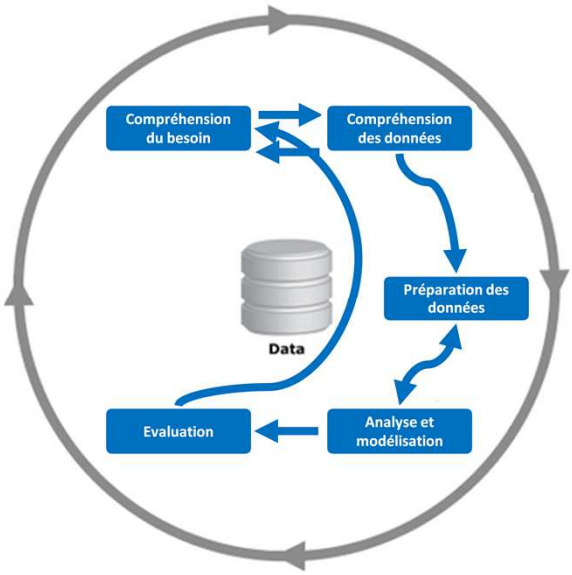
le travail d'analyse peut commencer.
Avant cela il est primordial de connaitre la question à laquelle on doit répondre :

- Analyser le cout de la réparation automobile -> BOF / Pourquoi le cout de la réparation automobile augmente ? OUI

Vous serez un métier support au service des « métiers », il sera important d'échanger avec eux sur les attendus de vos travaux. Une heure d'échange avec le métier, c'est plusieurs heures d'études statistiques de gagnées.

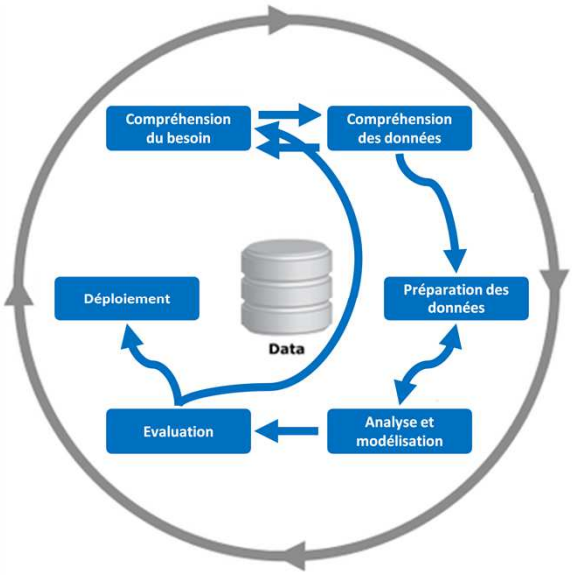
La suite, c'est votre métier. Utilisez la méthode stat la plus adaptée à votre sujet en prenant compte des délais demandés, du niveau de précision demandé…
Restez pragmatiques ! Votre client se souciera parfois peu des « super » méthodes que vous avez réussi à mettre en œuvre pour répondre à la question.
Vous voulez valoriser votre travail statistique : rédigez un article de synthèse que vous pourrez échanger avec vos collègues, avec une communauté sur le web… Vous favoriserez en plus la reproductibilité de vos travaux.

Evaluer vos résultats



Challenger les résultats de vos études : avec des études existantes, des chiffres de références, des métriques de performance

Délivrer un produit



Le produit final peut prendre plusieurs formes :
- un reporting récurrent
- un tableau de bord abouti avec des indicateurs clés
- une note économique : évitez les notes de 12 pages dont 10 de méthodo
- un applicatif complet permettant de simuler des impacts de décision tarifaire
- un algorithme de machine learning permettant de scorer en temps réél un risque quelconque

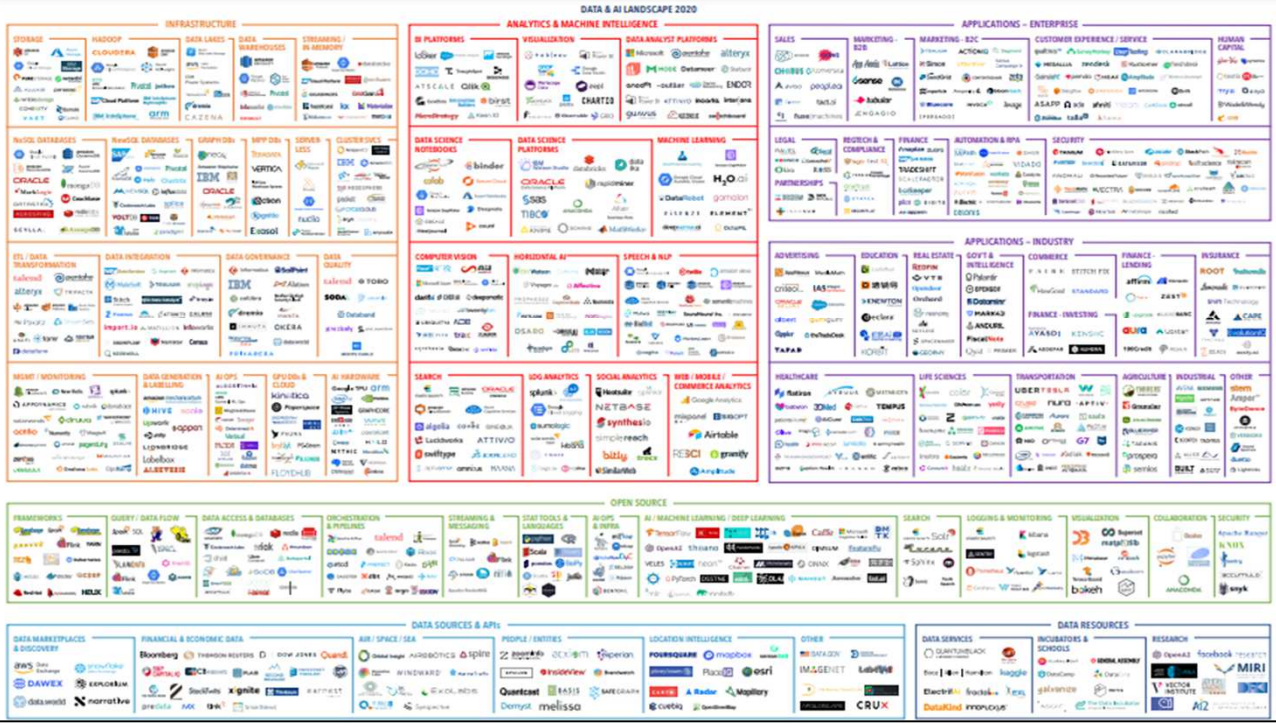
Adaptez votre produit à votre « client ». Vous vous adressez à un directeur, il aura le temps de lire un document d'une demi page maximum.
Vous vous adressez à un expert métier, il saura apprécier une note complète mettant en avant grace aux statistiques des éléments qu'il n'avait pas vu.
Ex : l'expert métier sait que le cout de la réparation automobile augmente mais vous lui avez montré que c'est à cause du prix des pièces de réparation sur les véhicules allemands…

Soignez votre produit final ! C'est ce produit qui sera jugé par votre client, pas les 10000 lignes de code que vous avez développées.
En entreprise, le statisticien doit apprendre à gérer une frustration : tout le monde (ou presque) se moque de votre démonstration. Le résultat est le plus important.
Vous voulez valoriser votre méthodo, communiquez avec vos pairs

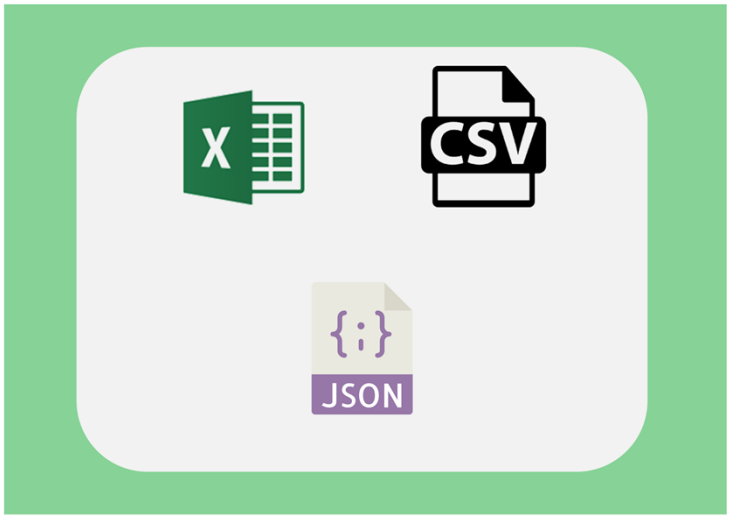
Déployer votre produit de manière industrielle

Des outils

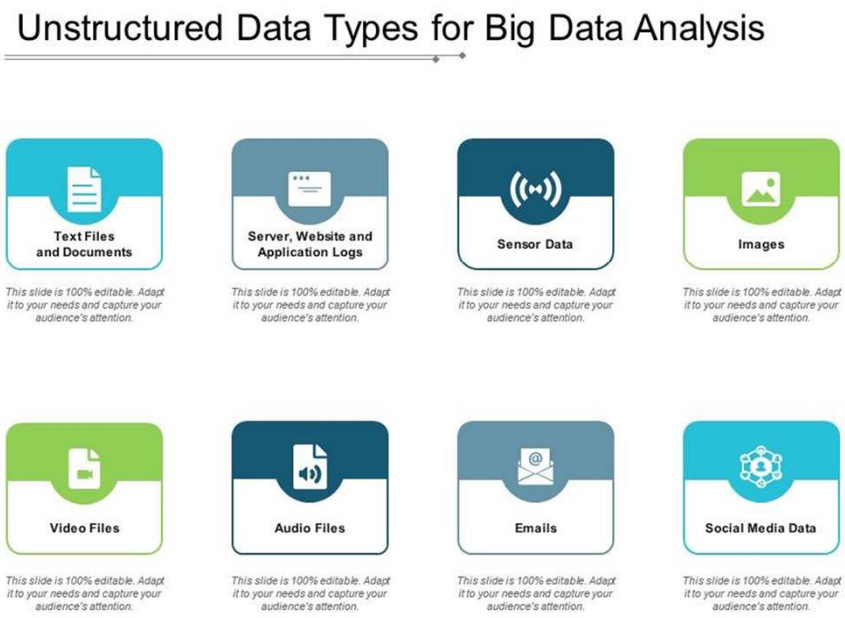
Mais lesquels ?



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

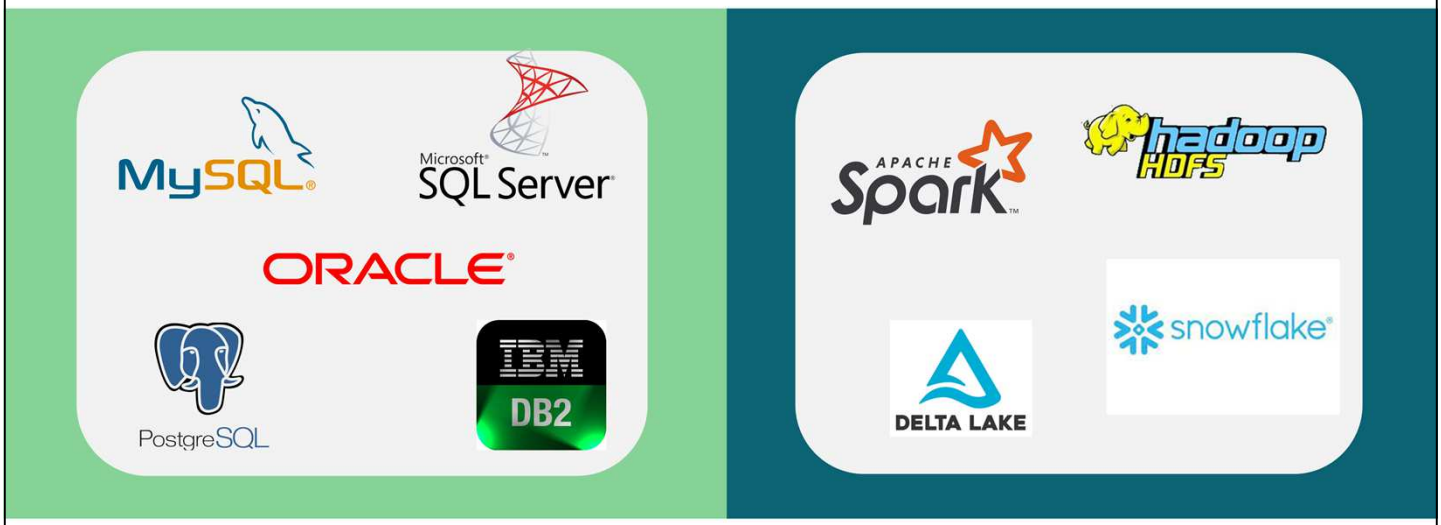


Comprendre l'écosystème – les données (2/2)



Accéder aux données

SGBD vs Big data



- SGDB relationels (Oracle, MySQL…) ; ensemble de données organisées suivant un modèle relationnel (entité<->relation)

- Hadoop : framework permettant de créer des applications distribuées (stockage et traitement) et scalables. Il est composé de plusieurs modules dont HDFS stockage de fichiers distribué

- Spark : Framework de calcul distribué (RAM + CPU) disposant d'API pour les langages Python, R

- .parquet : format de stockage orienté colonne

Accéder aux données – quel langage ?

La valeur sure !

- SQL : le langage des bases relationnelles / utilisable en environnement big data au travers des interfaces Hive et SparkSQL

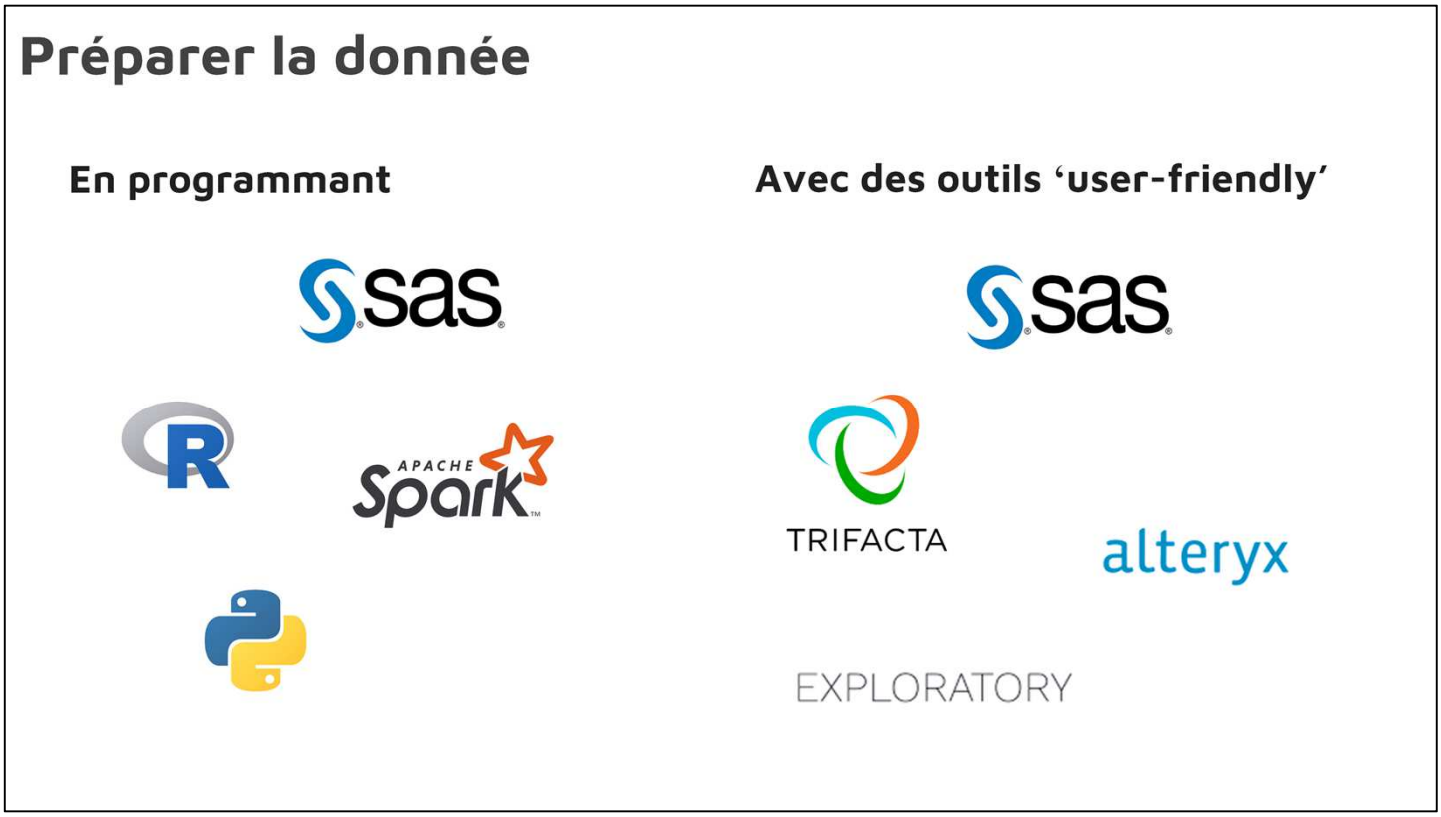
Comprendre la donnée

En programmant

Avec des outils ‘user-friendly’

- Sortir des statistiques simples : moyenne, ecart type, correlation, distribution
- Visualiser les données de manière basique : histo, boxplot, oucrbes, maps….
- Deux approches : les outils clés en main / les langages de programmation.
- Possibilité d’utiliser les langages au travers de notebook

Préparer la donnée



- Transformer la donnée :
  - la nettoyer : valeurs aberrantes, manquantes, filtres
  - la rendre compatible avec des traitements statistiques plus ou moins évolués

- Créer de nouvelles données : à partir des données existantes, à partir de données externes

- Outils clic : alteryx / trifacta /exploratory mais souvent à la main de l’IT

Analyser et modéliser

En programmant

Avec des outils ‘user-friendly’

DataRobot

- A part SAS, très présent dans les grands comptes, les autres sont assez peu répandu pour le moment
- Dataiku dispose d’une version gratuite limitée aux fichiers plats en source

Analyser et modéliser

Les frameworks de programmation

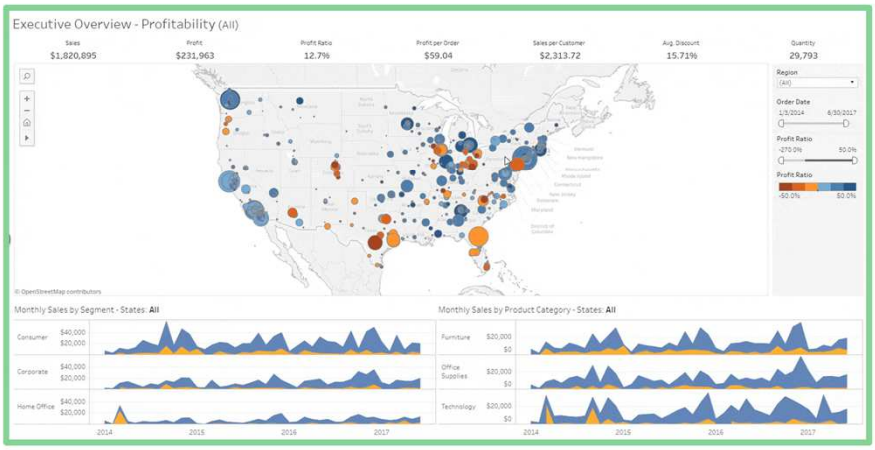
- Possibilité d'accéder à des frameworks de ML très puissantes au travers des packages
- Framework : accessibles au travers des langages de programmation (r/python/scala)…

Présenter les résultats

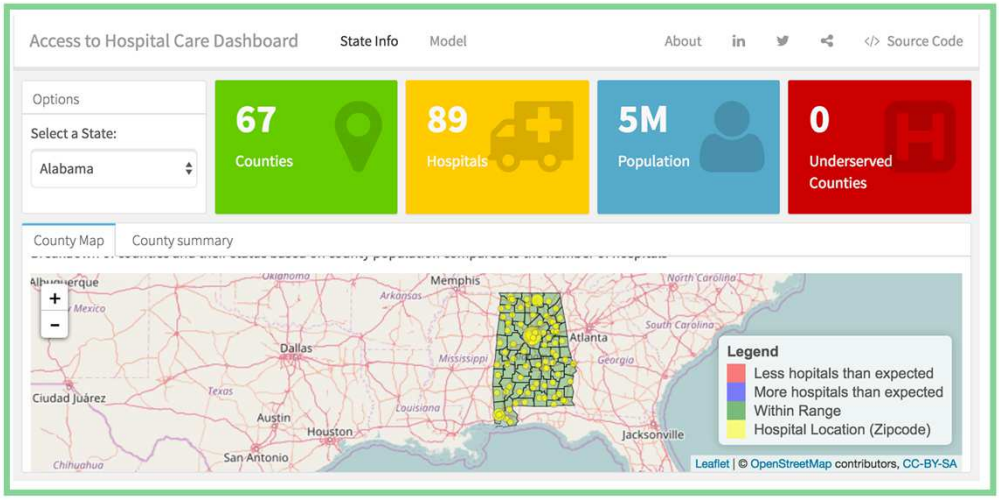
Qlik

+ a b | e a u

TOUCAN TOCO



Présenter les résultats



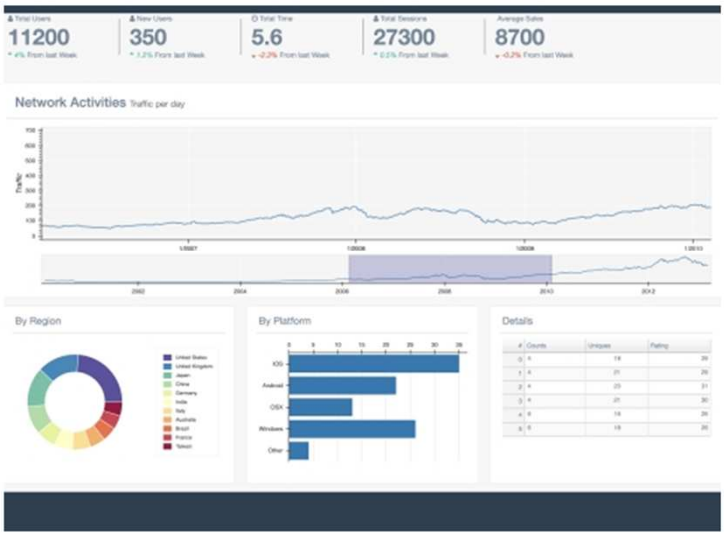
- R+Shiny / Markdown
- Pour les geeks, d3.js : programmation en javascript

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

TREAMING

Enrichir vos analyses

- Prenez des sites de confiance
- Assurez vous de l'origine de la donnée, de sa documentation, de sa fréquence de mise à jour

Se former


DataCamp


DataScientest • com

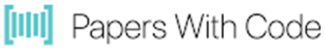

bloggers


KDnuggets


FUN
FRANCE
UNIVERSITÉ
NUMÉRIQUE
Se former en liberté

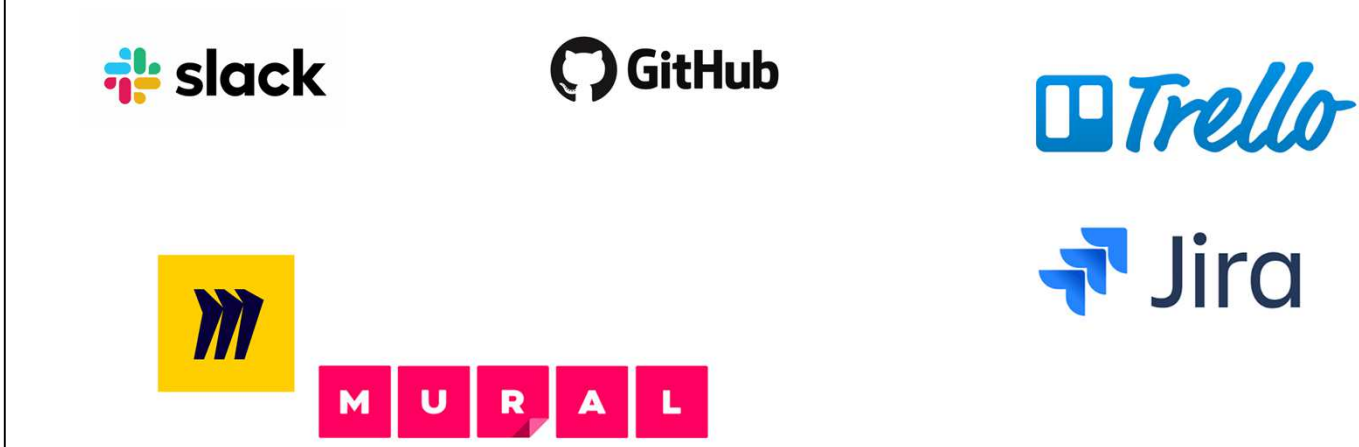

coursera


stackoverflow


Papers With Code


kaggle

Gérer vos projets

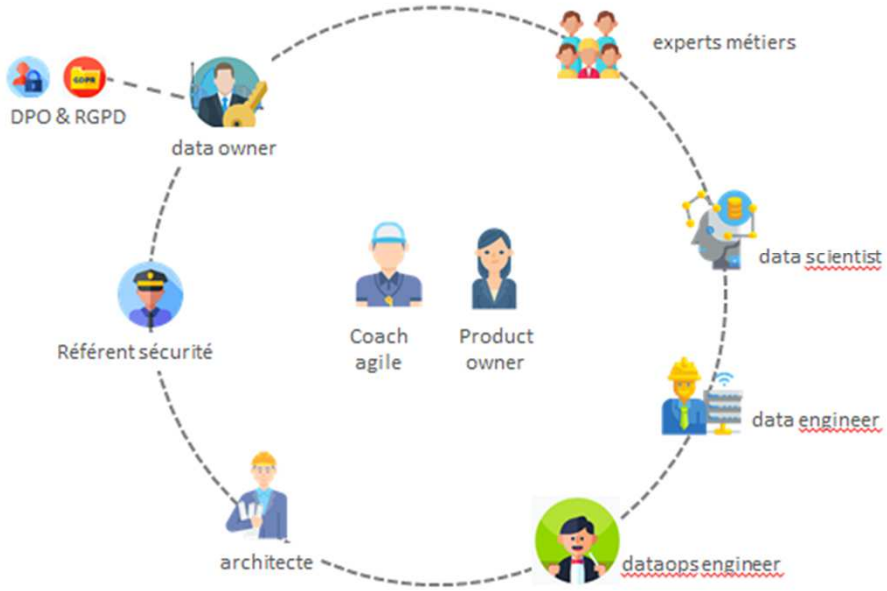


- Slack pour les canals de com
- Git pour le partage de code
- JIRA/Trello pour le backlog
- Miro/mural pour l’idéation


Une transformation dans
l'approche de la donnée


La data est un sport d'équipe
Beaucoup d'out
Beaucoup de donné
Beaucoup de méthod
20

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

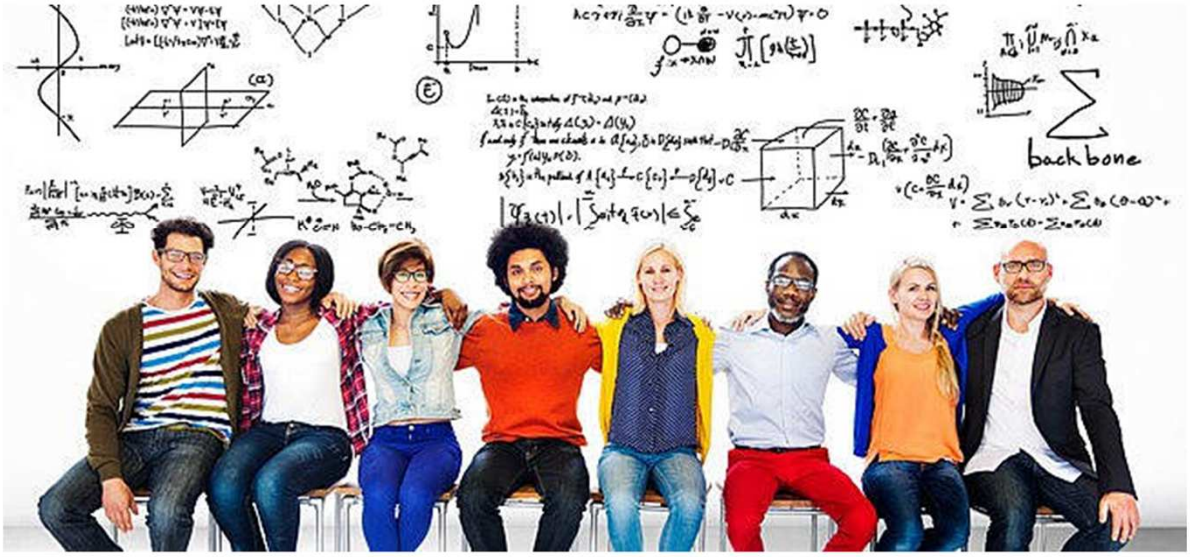
Product owner
Est le point d'entrée du projet côté métier. Il doit être appétent/acculturé au domaine de la data science.
Il est un membre de l'équipe à part entière. Son rôle est de comprendre les problèmes de fond du métier et de valoriser la résolution de ces problèmes. Il guide l'équipe dans la compréhension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

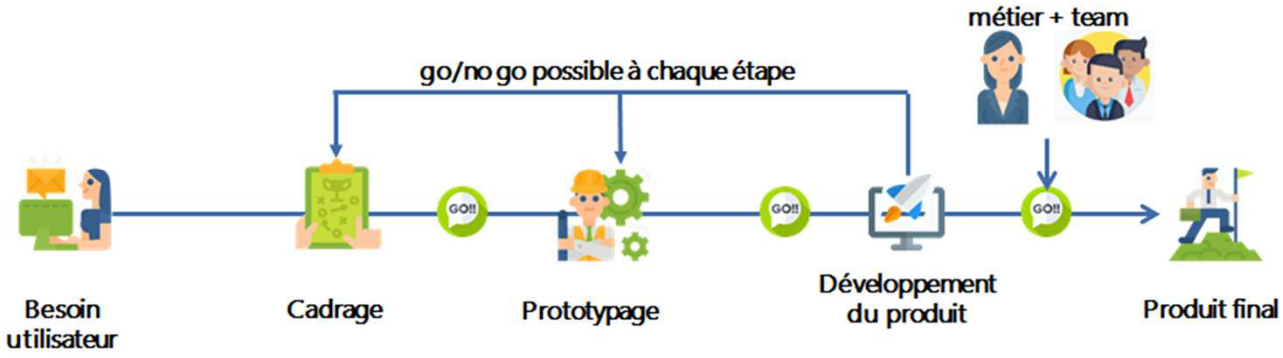
Data owner
Est garant de l'alignement stratégique et des politiques de gouvernance et de qualité des données sur un domaine de données dont il est propriétaire. Il accompagne les différents rôles du projet pour trouver des données, les comprendre et les utiliser correctement. Il est garant de la qualité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

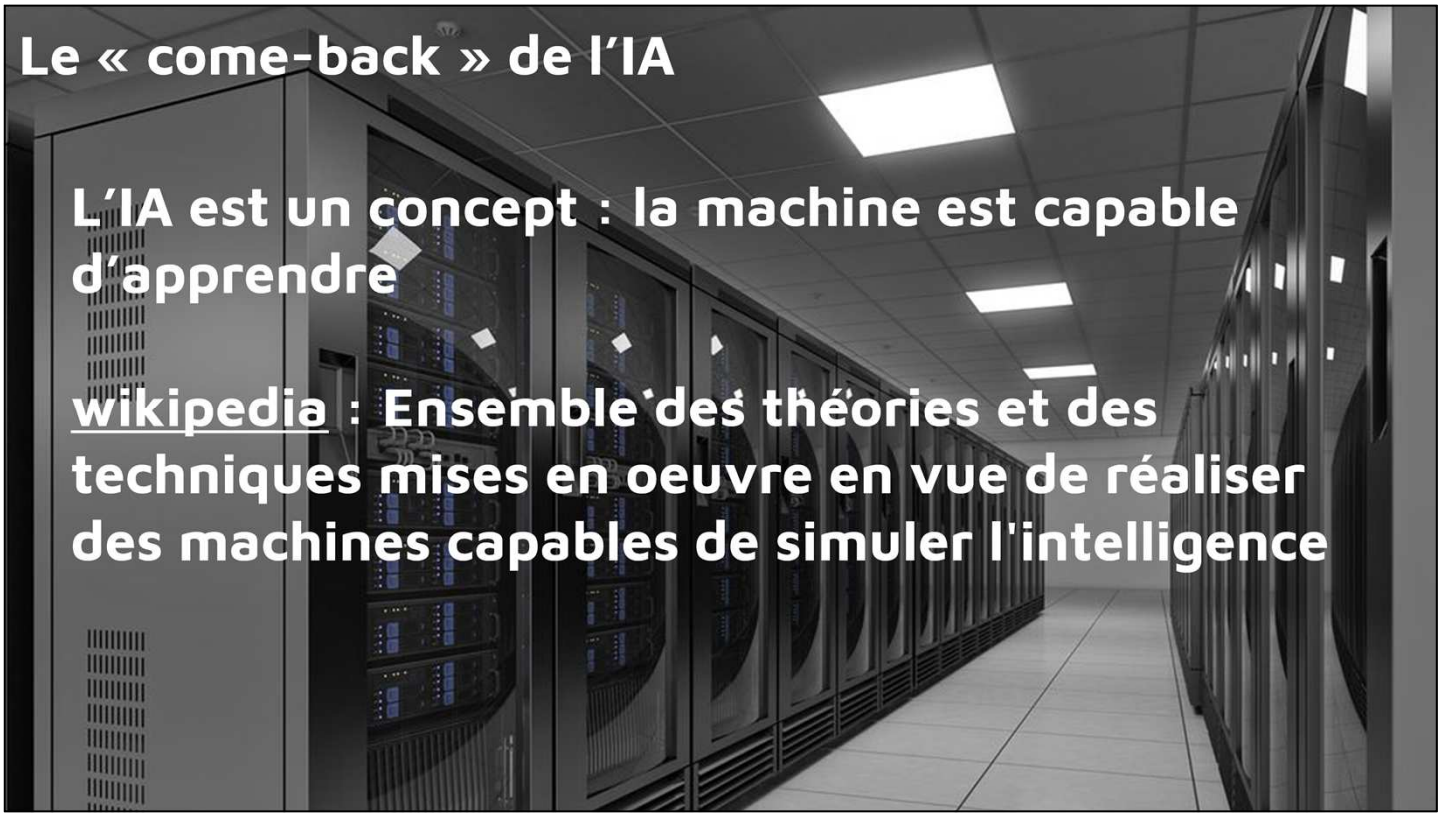
L'équipe imagine une solution qui permettrait de faire un scoring à chaud des dossiers sinistres s'appuyant sur une double approche : apprentissage supervisé à partir de cas existants et une approche non supervisée avec la détection « d'anomalies ». La solution devra être rapide et s'interfacer avec 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.


Les grandes tendances 2021



Émergence de l'IA dans les années 50 / Ian Turing

Développement dans les années 80

Remise à jour au début des années 2000/2010 (puissance de calcul, données…)

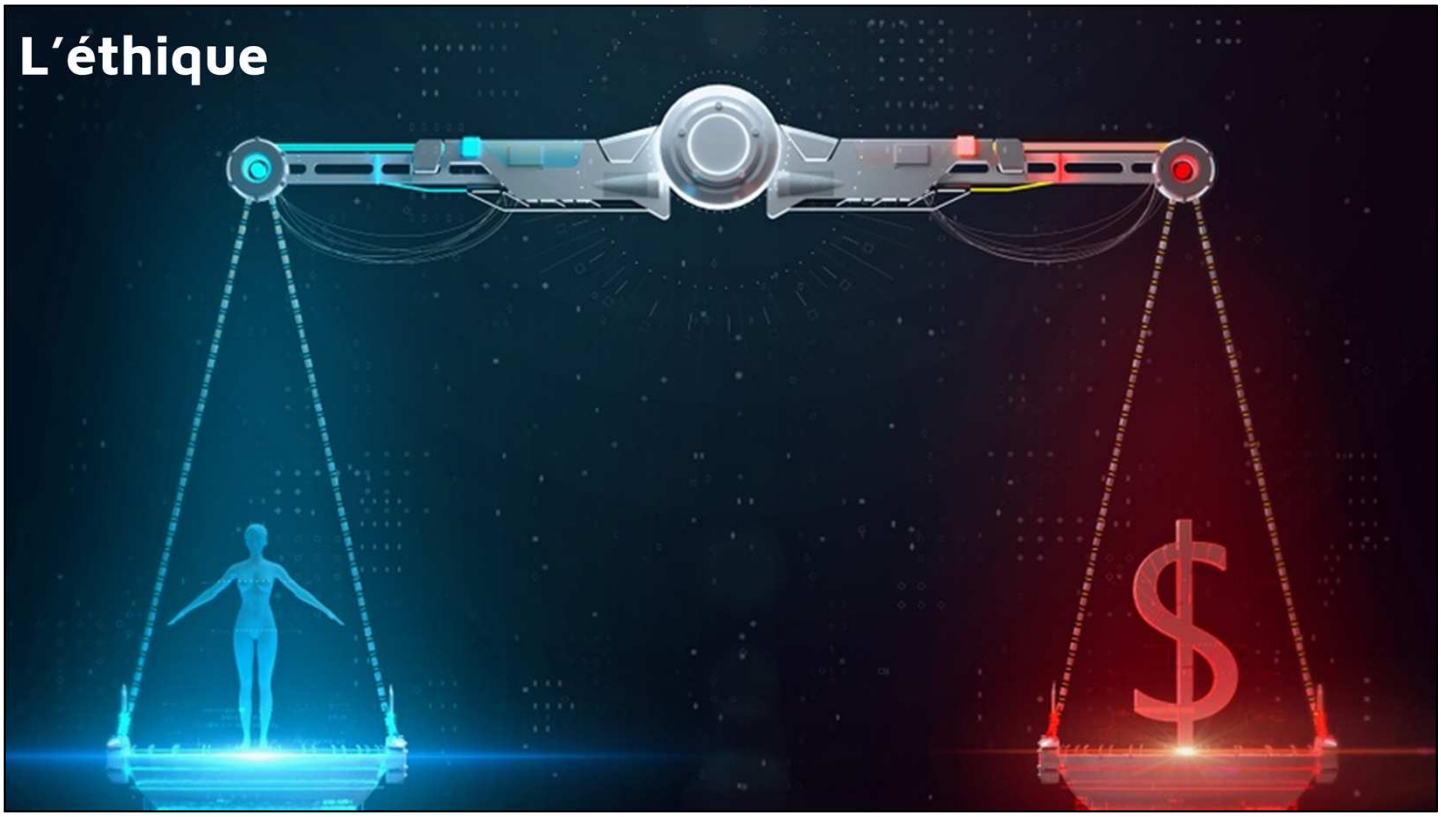
Domaines couverts : robotique, traitement du langage, machine learning

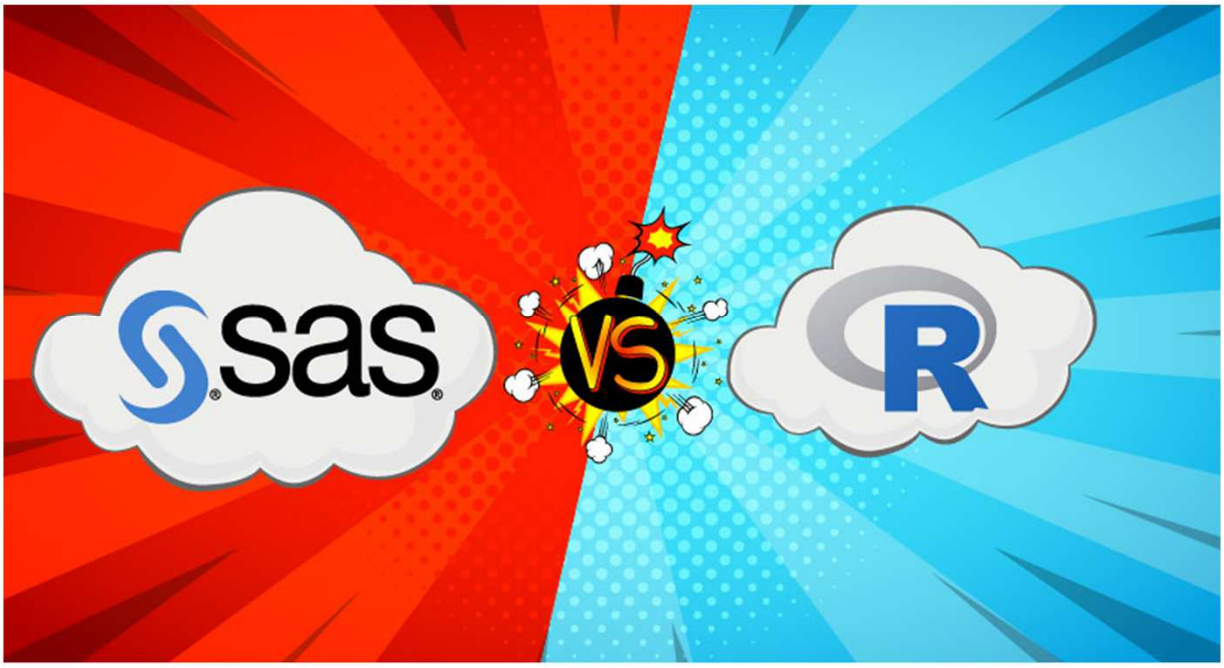
A lire : quand la machine apprend Yann Le Cun

La data science devient mature

L’enjeu est désormais de mettre en production des travaux surtout axés R&D a ce jour

Travailler sur la data ne peut plus se faire sans prise du compte du RGPD et du consentement de l’utilisateur


L’éthique



- Prenez des sites de confiance
- Assurez vous de l’origine de la donnée, de sa documentation, de sa fréquence de mise à jour

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
R

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 ?

R : langage de data analysis de stats
Python : langage adapté plus pour le déploiement et la mise en production, c'est un langage de programmeurs

Code Python est plus propre que le code R, et donc plus maintenable.
R est puissant sur la partie Dataviz
API ML souvent plus avancées en Python

Comment choisir alors :
si vous souhaitez développer des algos et les mettre en production->Python
Si vous souhaitez faire de l'analyse statistique et produire des résultats -> R