LES NOUVEAUX OUTILS DE TRAITEMENT DE DONNEES

SÉBASTIEN QUINAULT / S1 2019 / M1 SARADS

OBJECTIF

Vous présenter une démarche et des outils ‘tendance’ utilisés en entreprise

Un exemple de démarche de travail : *expliquer la démarche d'analyse de données en entreprise pour ensuite introduire l'étendue des outils disponibles et la « nécessité » de cette diversité*

Un éco système technique en grande mutation : *montrer la multitude d'outils actuels en faisant un focus sur outils « data » et « stats », qq définitions de basd*

Quels outils pour quelle étape

Exemple d'outils de gestion de projet

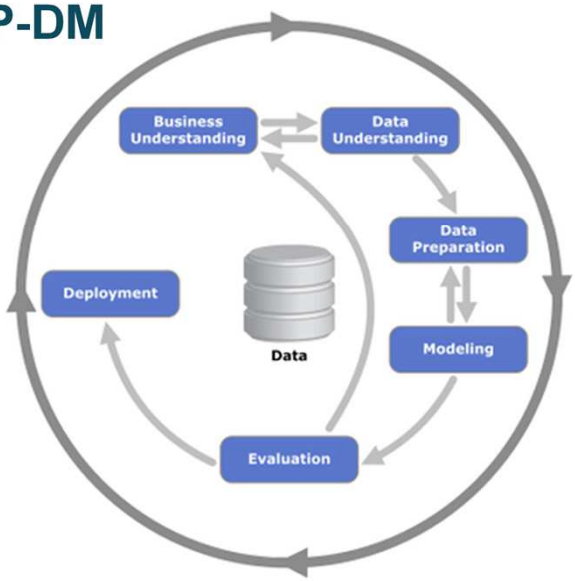
Se former

1

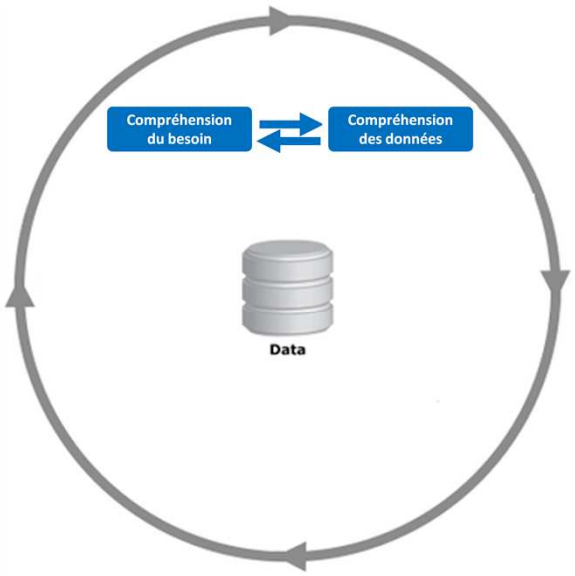
UNE DEMARCHE D'ANALYSE DE DONNEES EN ENTREPRISE

LA METHODE CRISP-DM

Méthode 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



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…

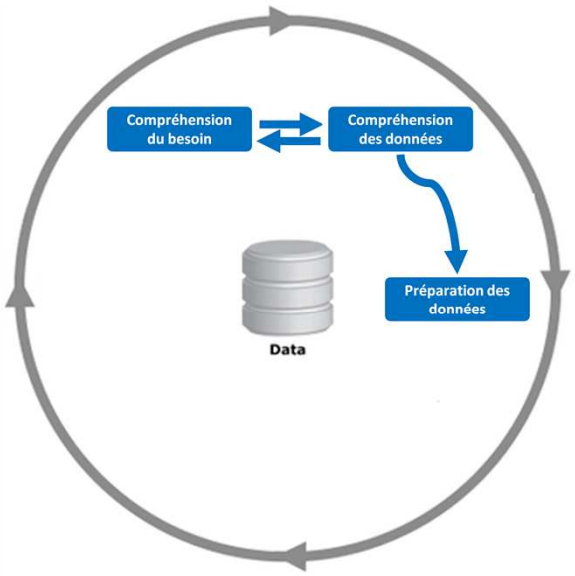
L'accès aux sources peut se faire au travers de nombreuses technologies : fichier plat, SGBD, API… L'idéal est de maitriser un langage de programmation (r, sas, python) car il permettra d'accéder à une très grande variété de données.

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 !

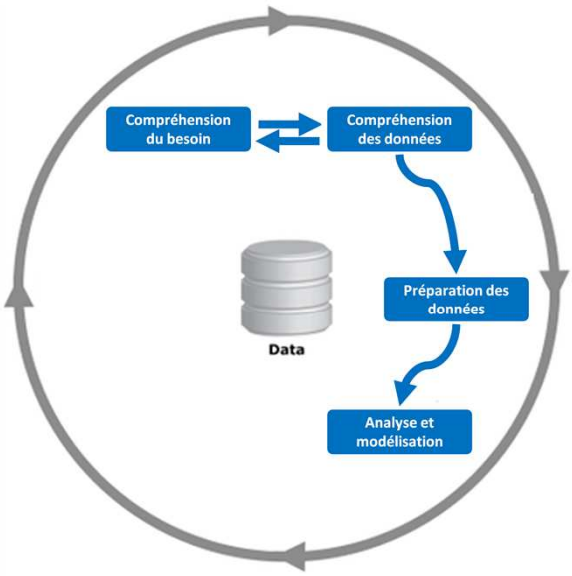


Le traitement et l'enrichissement des données est un élément différenciant entre les data scientists.
Aujourd'hui tout le monde à accès aux derniers algorithmes, à de la puissance de calcul (AWS).
Mais vous être l'intelligence derrière la machine, vous savez quelles données choisir, comment les transformer,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

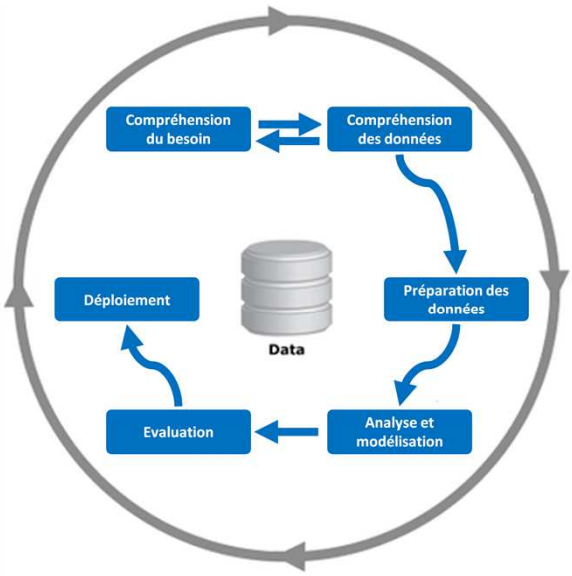


le travail d'analyse peut commencer.
Avant cela il est primordial de connaitre la question à laquelle on doit répondre :

- 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.



Le produit final peut prendre plusieurs formes :

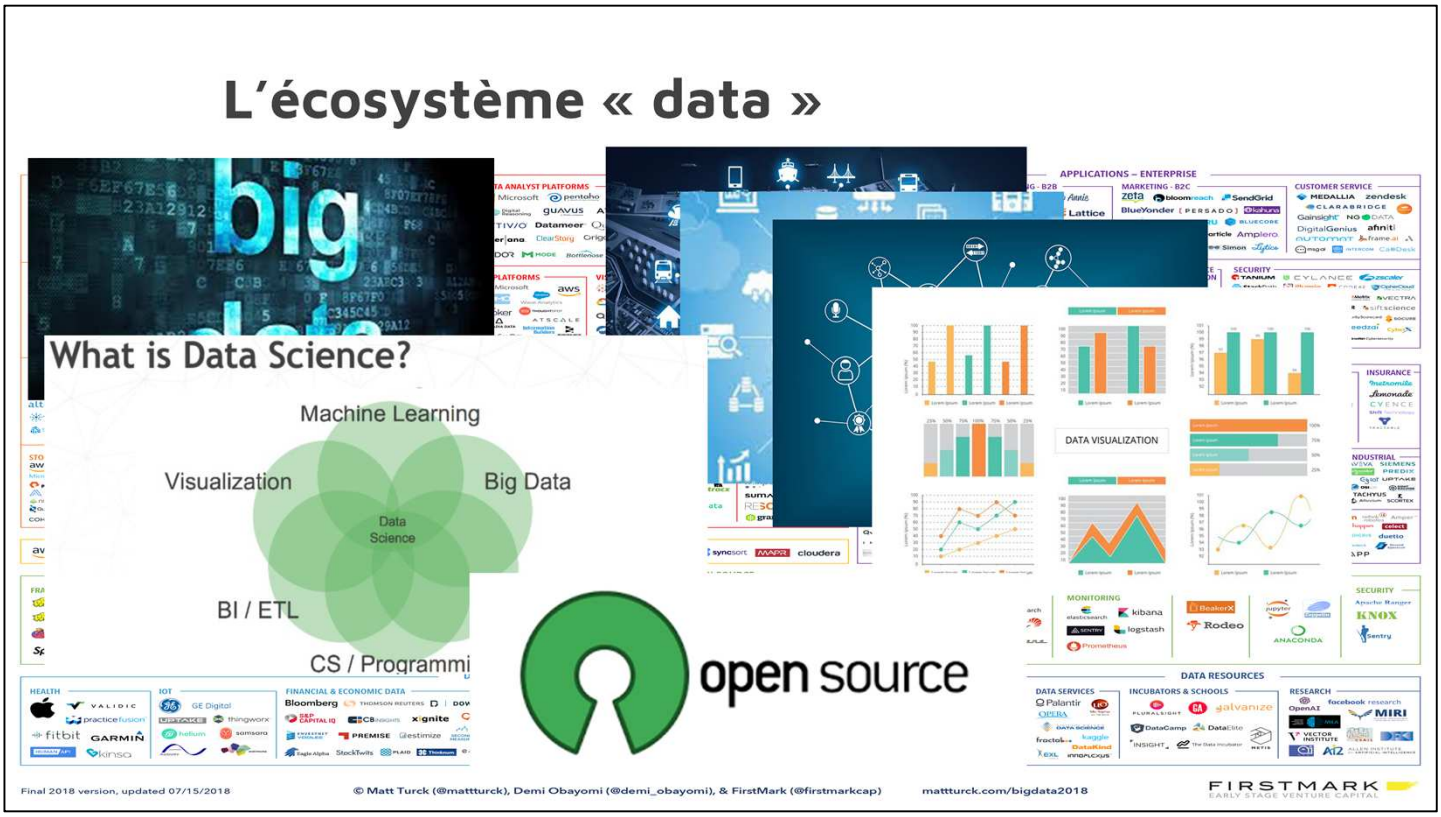
- un reporting récurrent
- un tableau de bord abouti avec des indicateurs clés
- une note économique : évitez les notes de 12 pages dont 10 de méthodo
- un applicatif complet permettant de simuler des impacts de décision tarifaire
- un algorithme de machine learning permettant de scorer en temps réél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

2

UN ECO SYSTÈME TECHNOLOGIQUE EN PLEINE EVOLUTION



Panorama 2019 assez complet des outils liés à la data
Quelques points de vocabulaires à éclaircir :

- Big data : données massives, 3V (vitesse, variété, volume)
- AI : intelligence articifielle, il s'agit d'un concept selon lequel la machine est apprenante. L'IA couvre plusieurs domaines : la robotique, le traitement du langage, le machine learning…
  https://fr.wikipedia.org/wiki/Intelligence_artificielle

- Data science : extraire des connaissances d'un ensemble de données

- Machine learning : champ d'application de l'IA / concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique

- Data viz : ensemble de techniques permettant la visualisation de données

- Open source : logiciels distribués librement (pas forcément gratuitement). Le code source de l'applicatif est libre et modifiable

3

DES OUTILS ADAPTES A CHAQUE ETAPE

L'ACCÈS AUX DONNÉES

Fichiers plats et SGBD

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

L'ACCÈS AUX DONNÉES

Les techno big data

- 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

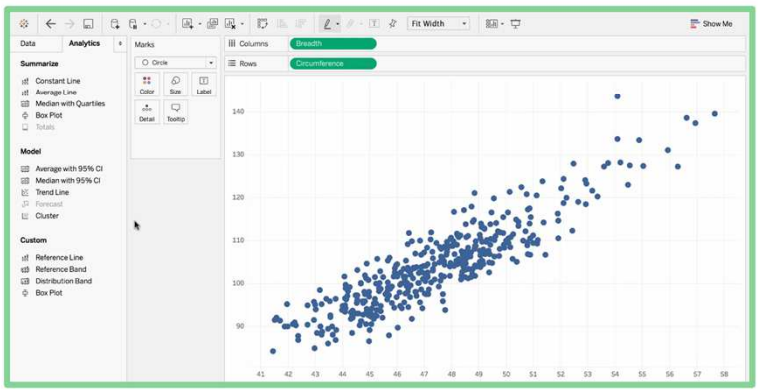
L'ACCÈS AUX DONNÉES

Un élément clé

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

L'EXPLORATION DES DONNEES

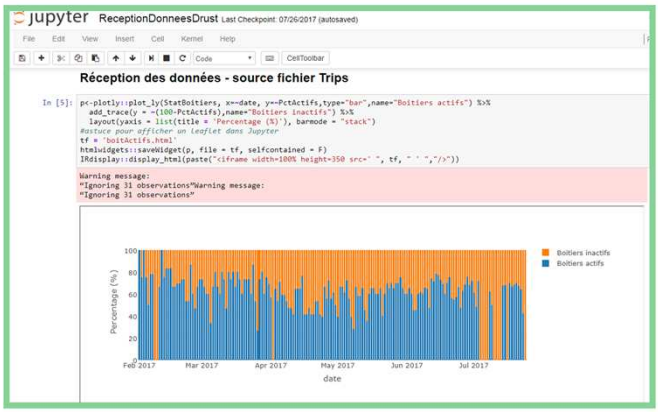
Les outils "clic-bouton"



- 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

L'EXPLORATION DES DONNEES

Les langages



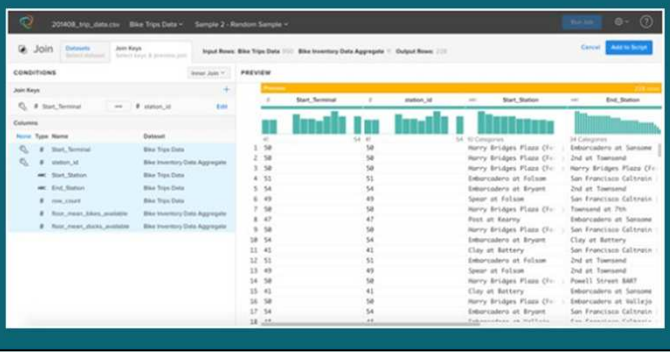
- Possibilité d'utiliser les langages au travers de notebook ou de IDE : RStudio pour R, Jupyter pour Python

Jupyter lab

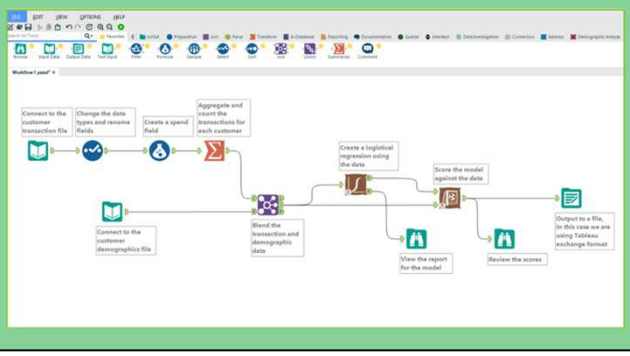
LA TRANSFORMATION DES DONNÉES

Les outils “clic bouton”

TRIFACTA



alteryx



- 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

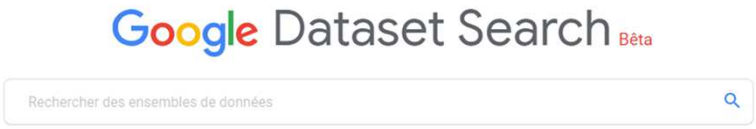
- Outils clic : alteryx / trifacta /exploratory mais souvent à la main de l’IT

LA TRANSFORMATION DES DONNÉES

Les langages

L'ENRICHISSEMENT

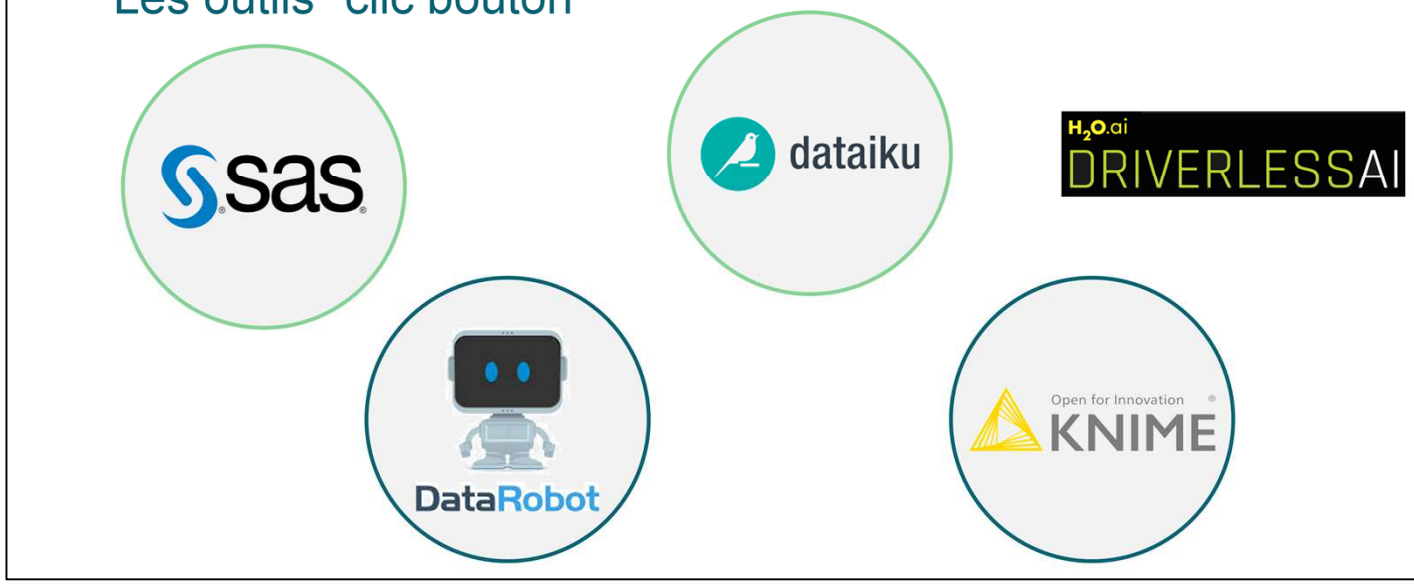
La data externe



Beaucoup de ressources disponibles en ligne
Privilégier les sources de données officielles, documentées, à jour

L'ANALYSE ET LA MODÉLISATION STATISTIQUE

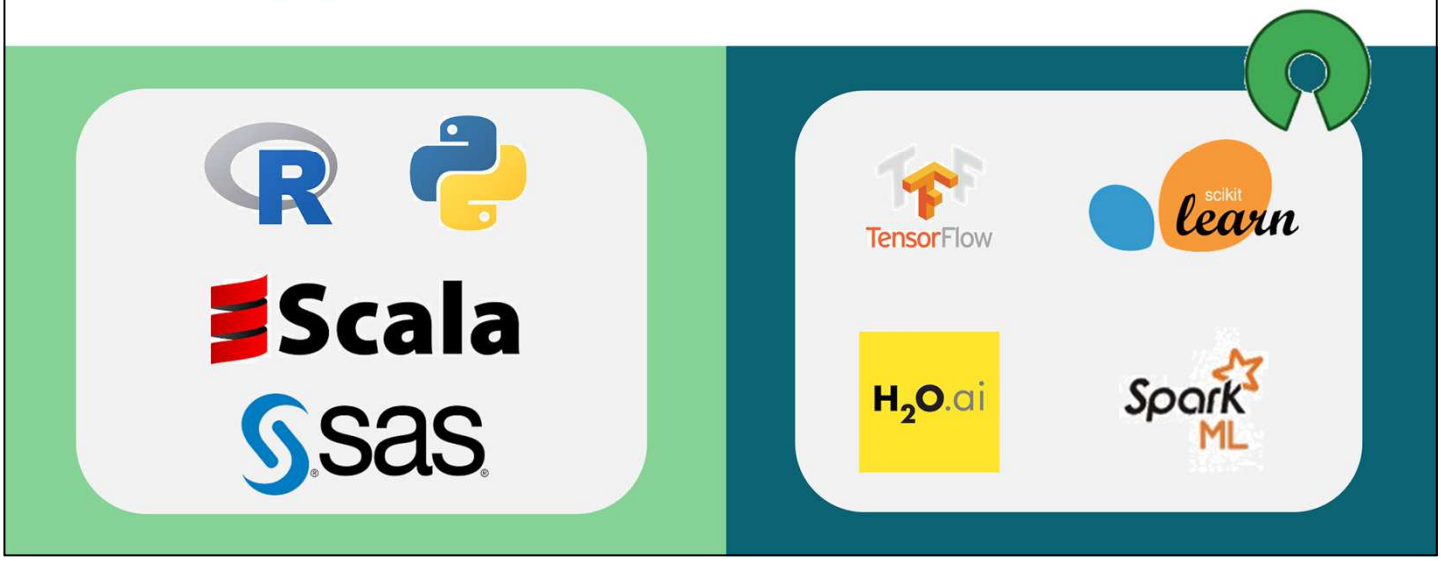
Les outils "clic bouton"



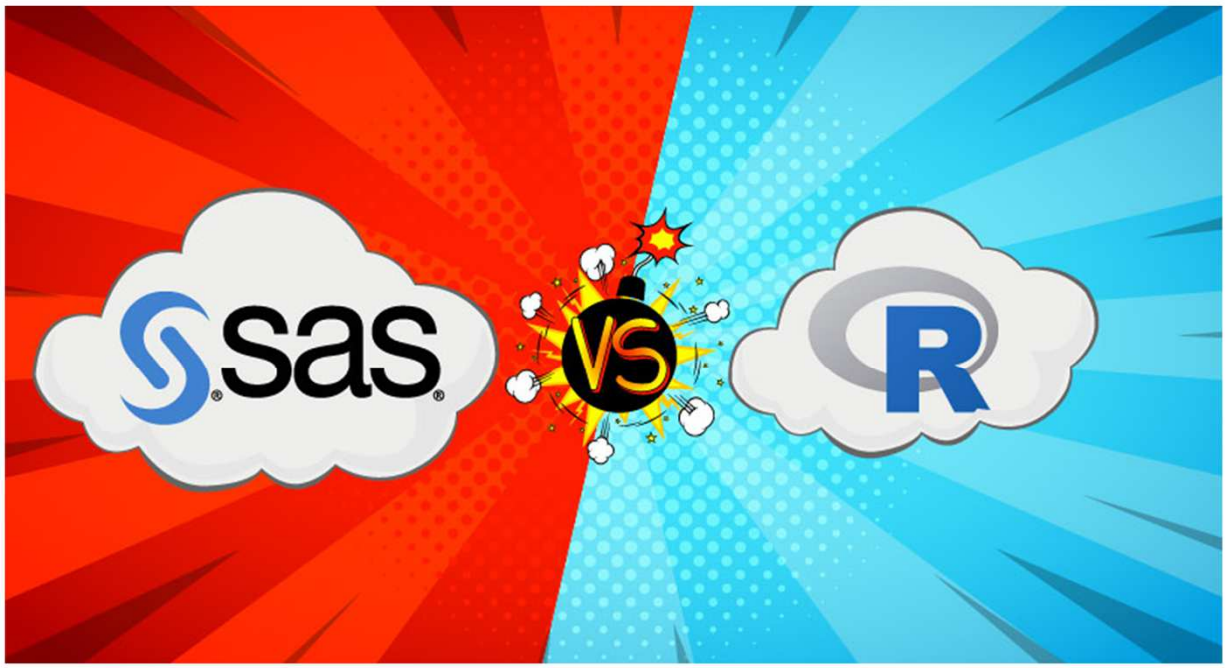
- A part SAS, très présent dans les grands comptes, les autres sont assez peu répandu pour le moment
- Dataiku dispose d'une version gratuite limitée aux fichiers plats en source

L'ANALYSE ET LA MODÉLISATION STATISTIQUE

Les langages et framework



- Beaucoup de packages et librairies stats dans R et Python
- 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)…


sas
VS
R

QUELQUES DIFFÉRENCES

- Modèle de tarification : licensing vs open source
- Courbe d'apprentissage + IDE
- Disponibilité des nouveautés
- Restitution des résultats
- Support éditeur / communauté / documentation
- Patrimoine applicatif existant

SAS + cher
R plus compliqué / pas de Guide comme dans SAS / mais R studio une bonne IDE

R évolue plus facilement : open source, bcp de contributeurs, bcp de packages
SAS assez faible en dataviz…on passe souvent par excel. R au top avec des packages comme ggplot2 / plotlty et la possibilité de faire du markdown et du shiny
Sas a une documentation de bonne qualité / R c'est aléatoire suivant les packages
Un des principaux « freins » à la croissance d'un langage comme R est l'existence d'un patrimoine applicatif important dans les grands comptes


VS
R

LES DIFFÉRENCES

- Deux langages open source avec 2 objectifs différent
- 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

LA RESTITUTION DES RÉSULTATS

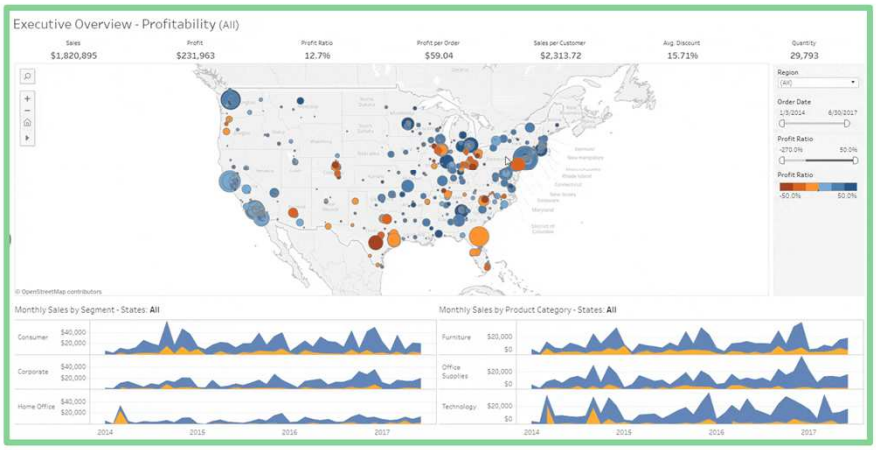
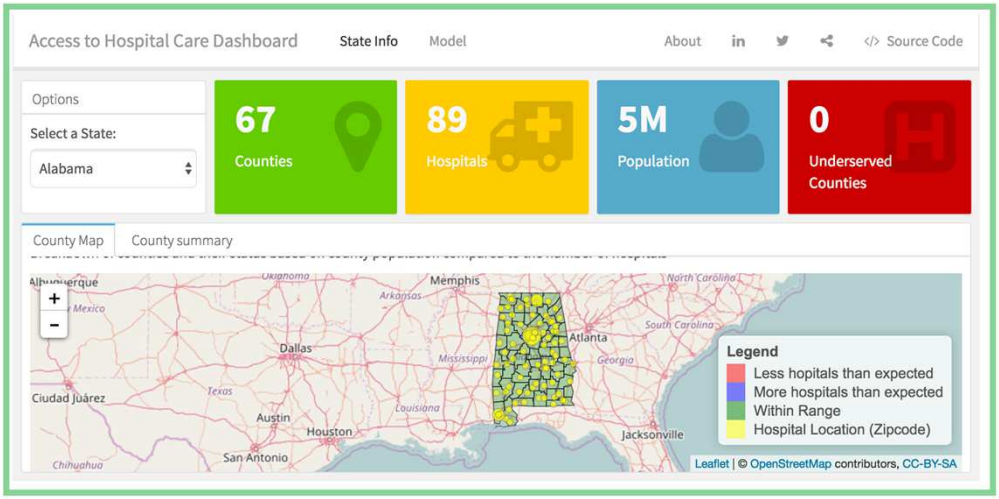


Tableau / Qlik / SAP BO

LA RESTITUTION DES RÉSULTATS



markdown exemple

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

4

LA GESTION DE PROJETS

LA GESTION DE PROJETS

5

SE FORMER

QUELQUES RESSOURCES

POUR ME CONTACTER

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