CES
Data Scientist

Institut Mines Télécom

Comment a-t-on inventé le Data scientist ?

- Des données exploitables en quantité folles et dont le rythme de croissance est exponentiel

- Des besoins métiers pour proposer des nouvelles fonctionnalités, des nouveaux services, mieux comprendre les usagers

- Validation d'hypothèses et de modèles à portée de main

Le cas d'école : LinkedIn « People you may know »

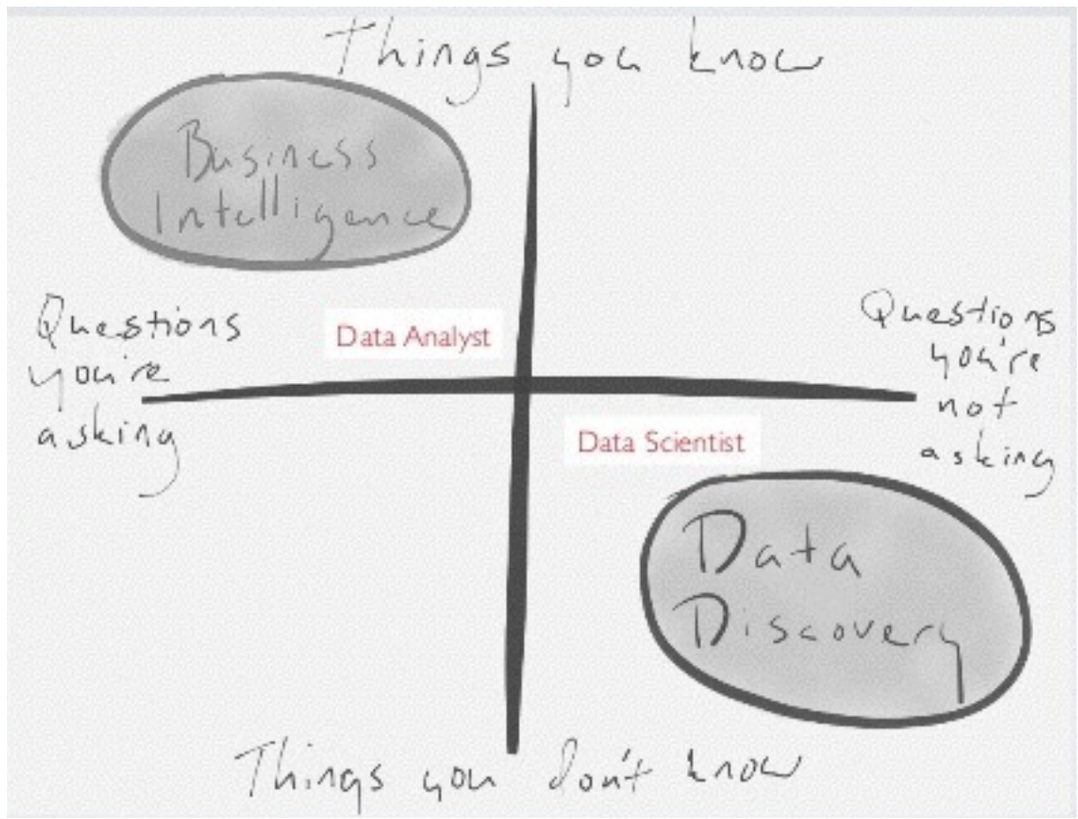
- Les gens du réseau n'ajoutaient pas spontanément beaucoup de personnes ce qui freinait la croissance

- Un doctorant a bati un modèle de recommandation basé sur les profils des gens pour encourager les gens à ajouter les personnes qu'elles pouvaient connaitre

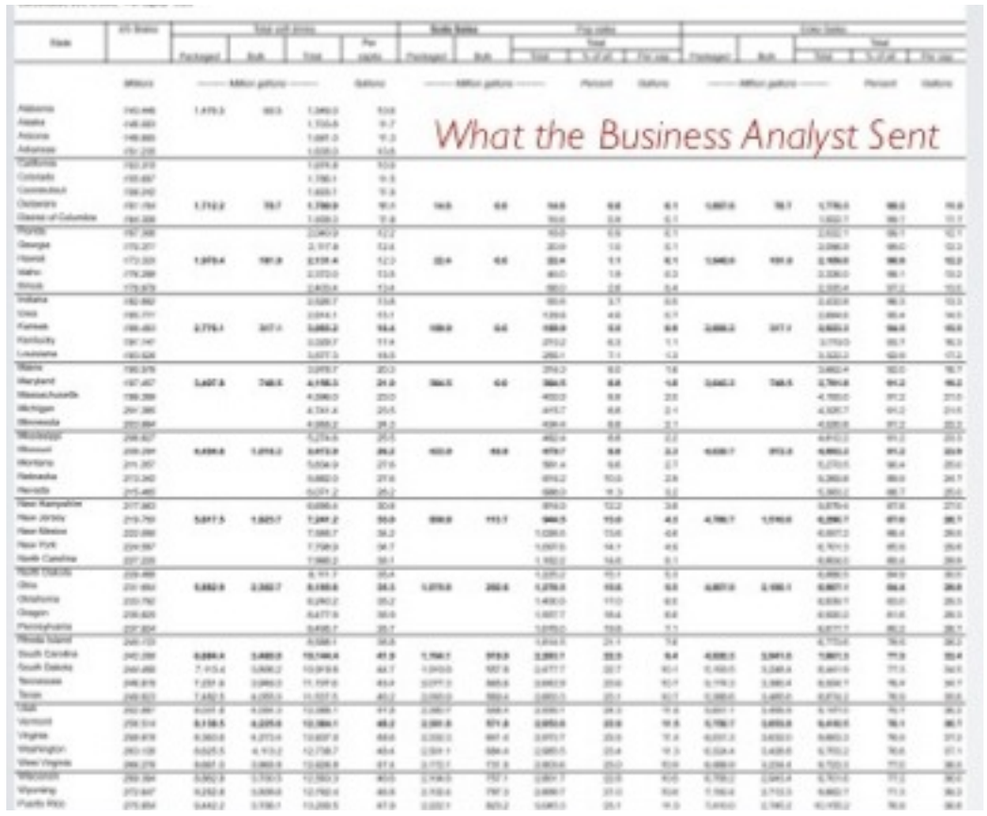
Quelles différences avec les métiers existants auparavant?

- Data scientist vs data miner
- Data scientist vs data analyst
- Data scientist vs developer

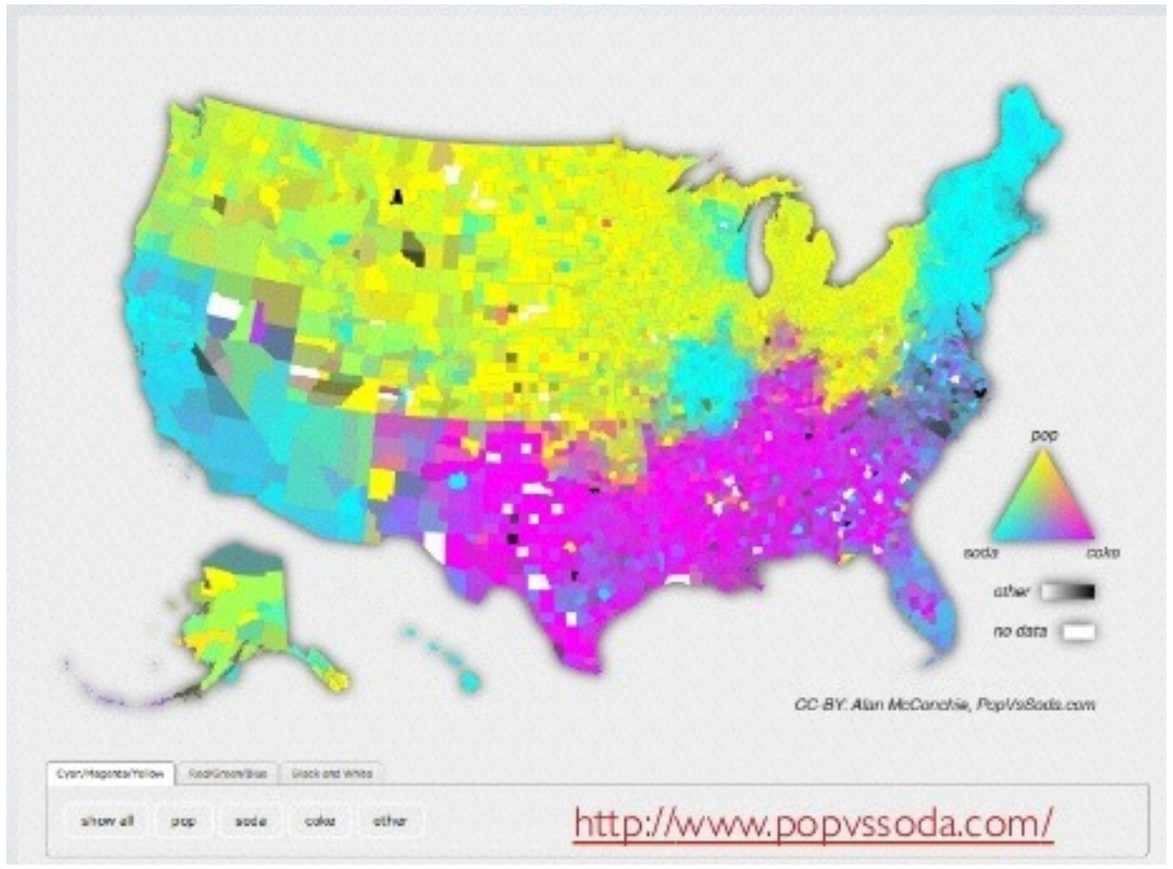
Quelles différences avec les métiers existants auparavant?



Quelles différences avec les métiers existants auparavant?



vs



business analyst

data scientist

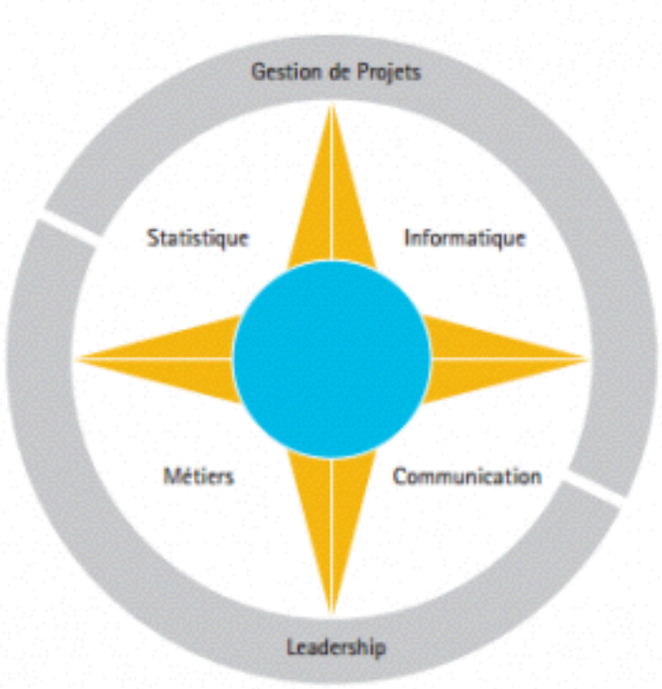
TOUCAN TOCO

Qu'attend-t-on d'un data-scientist?

- Découvrir un sens caché derrière les données
- Obtenir des insights prédictifs et actionnables à partir de la donnée
- Créer des produits basés sur la donnée qui ont un impact direct sur le business
- Communiquer des histoires business pertinentes à partir de la donnée
- Apporter des informations fiables pour les prises de décision qui orientent le business

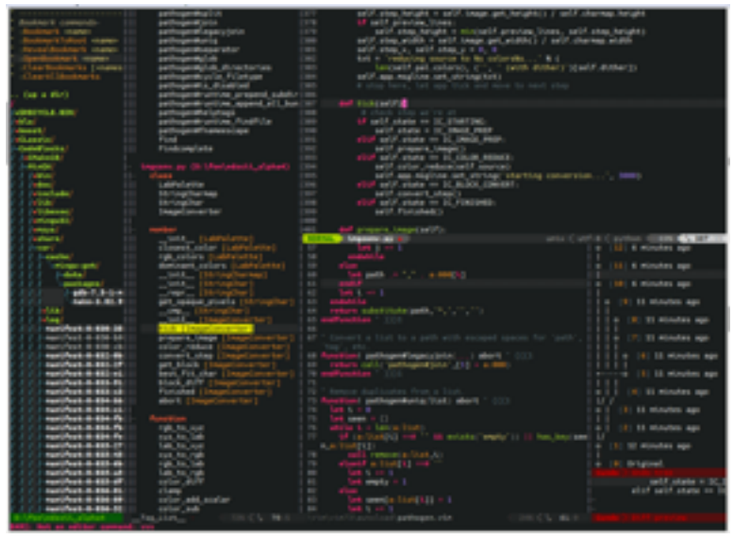
Les compétences du Data Scientist

- Expertise statistique et quantitative
- Sens business
- Capacités technique
- Communication



Qu'attend-t-on d'un data-scientist?

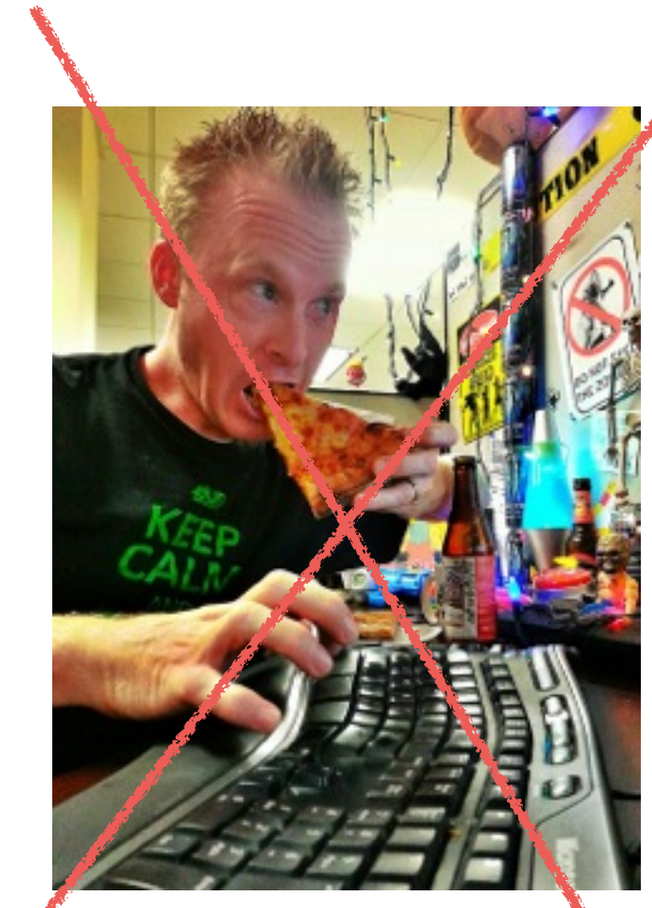
Capacités technique



- Savoir coder est indispensable
- Base : Excel, Python, R, Java
- Packages: Numpy, scikit-learn…
- Big Data : Hadoop, HDFS & MapReduce
- Visualisation: d3.js, Gephi
- Base de données: MySQL, MongoDB,…

Qu’attend-t-on d’un data-scientist?

Sens business



Qu’attend-t-on d’un data-scientist?

Sens business

- Un data scientist n’est pas un geek développeur qui vit enfermé avec des bières et de la pizza

- Un data scientist connait les différents métiers de son entreprise

- Il va partir d’enjeux et d’objectifs pour tenter d’y apporter une solution via son expertise

Qu'attend-t-on d'un data-scientist?

Expertise statistique et probabilistique

- Savoir extraire les informations de base d'un jeu de données
- Pouvoir croiser différentes sources
- Construire des modèles prédictifs

Qu’attend-t-on d’un data-scientist?

Communication

- Produire sans communiquer est inutile comme souvent

- Un data scientist doit pouvoir capter l’attention des décideurs et leur apporter l’information digérée et bien présentée

Les besoins essentiels du Data Scientist

- Etre au contact du reste du business, notamment avec les personnes opérationnelles responsables des produits et services

- Etre au contact de ses pairs pour échanger sur les méthodes / modèles / outils

- Avoir accès facilement au base de données internes

Data Scientist et Big Data

- Le Data scientist n'est pas forcément un spécialiste de la big data ! Il peut travailler de pair avec un tel spécialiste

- Certains enjeux business peuvent être résolus par un data scientist sans utiliser des données massives mais avec une bonne intuition, les bons jeux de données…

La chaine de traitements de la donnée : Acquisition

- Où est la donnée ?
- Comment je peux la récupérer ?
- Scraping, API, RSS, Capteurs…

La chaine de traitements de la donnée : Normalisation / Cleaning

- Plusieurs sources de données différentes
- Des données incomplètes
- Comment faire le pont et remplir les trous ?

La chaine de traitements de la donnée : Stockage

- Différentes méthodes de stockage pour différents usages

- Fichiers, base de données, système de fichiers distribués…

La chaine de traitements de la donnée : Analyses

- Quelles informations statistiques je peux retirer ?

- Analyse par groupe, moyennes, variance, corrélations…

La chaine de traitements de la donnée: Modélisation et Apprentissage

- Est ce que je peux donner un sens / une cohérence aux données via un modèle ?
- Est ce que je peux prédire certains outcomes ?
- Régression, classification, graphs bayésiens…

La chaine de traitements de la donnée : Visualisation

- Le dernier kilomètre de la donnée

- Présenter ses résultats à une audience business qui s'intéressera à l'impact direct sur son activité

- La visualisation doit être simple, actionnable et apporter des insights

La production d'un data scientist : un data product

- bati à partir de la donnée
- résultat d'explorations et itérations successives
- un algorithme qui apprend de la donnée
- une réponse à des inconnues
- une valeur immédiate pour le business
- une capacité à prévoir certains évènements dans le futur

Exemples de projets de data science

Se former en ligne

- La révolution des MOOCS
- Cours gratuits d'établissements prestigieux dans la majorité des disciplines académiques
- Importance donnée aux activités, aux évaluations, et aux interactions entre participants.
- Reposent également sur des exercices et des examens

Se former en ligne

- Introduction to Data-Science – University of Washington – Coursera

  - Manipulation de données volumineuses, introduction à Hadoop, MapReduce et NoSQL. Modélisation statistique, Machine learning, Graph Analytics, Text-Mining, Filtres Collaboratifs. Communication des résultats, DataViz. Les cours s'accompagnent de 8 devoirs à rendre, dont 4 en langage de programmation, un concours Kaggle et une visualisation en utilisant le logiciel Tableau.

- Pré-requis : programmation basique (R, SQL, Python) et familiarité avec la manipulation de bases de données

Se former en ligne

- Autres formations en ligne:
- BigDataUniversity.com
- MIT Online X Program
- …

Tester ses compétences

- Portails en ligne de challenges
- Problèmes concrets proposés par les entreprises
- Communauté de data scientists
- Possibilité de se comparer à d'autres participants
- Kaggle, datascience.net …

Questions

Liens

- http://toucantoco.com
- http://dashboardeco.fr
- http://communes.lesechos.fr
- http://toucantoco.com/innovators-thomson-reuters/