TD 3 – Notes

Intro à la DS : du DM au BD – Enjeux et opportunités

Warning : Ce document est à lire en complément des slides de Ricco Rakotomalala. C'est une version condensée-simplifiée pour les besoins d'une présentation en TD, étant donné que le temps y est limité.

En somme, on parle toujours de découverte de connaissances à partir des données, à l'aide d'outils divers et variés en vue d'apporter de l'aide à la décision. En ce sens, on mettra aussi l'accent sur la méthodologie et la connaissance métier. Ces démarches sont souvent un mix science-business[1], et il faudra parfois que le praticien mette lui-même les questions au jour. Ex de domaines :

- Statistique,
- Mathématiques (optimisation, etc.),
- Informatique (BDD, IA, etc.),
- Recherche Opérationnelle,
- Marketing & Sales,
- Etc.

En bref, tout domaine peut potentiellement être utile, et on les fait souvent travailler en synergie.

C'est important, aujourd'hui en particulier, car :

- On est à l'heure de la data : importance cruciale de la BDD, que cela soit la BDD relationnelle (Business Intelligence) ou la BDD non-relationnelle (Big Data). Pour les entreprises « traditionnelles » (banques, grande distribution), cela fait des années que l'on recourt à la data au service des métiers et de la prise de décision, avec diverses orientations : description, prédiction et même prescription.
- Les entreprises se sentent un peu plus concernées (d'où les « data awareness », « data centric », « data driven », etc.), et la législation a évolué dans le sens de cette prise de conscience (RGPD, etc.). Résultante : création de nouveaux postes comme le Data Protection Officer, qui fait également office, en général, de référent RGPD. Dans le même sens, on a davantage de demande sur des métiers comme le Data Engineer, le Data Architect, le Data Analyst, le Data Scientist, ou encore le Data Steward, sans parler de tous les métiers qui existaient déjà (analyste BI, statisticien, etc.).

[1] Lorsque je dis "business", cela s'appliquerait en fait à tout métier. Je prends l'exemple de l'entreprise car c'est de loin le cas le plus parlant.

La chaîne de traitement de l'information (très grossièrement) :

La collecte de données, par qualité+volumes décroissants :

- Expérimentations/enquêtes (ad hoc) → c'est le travail du Statisticien,
- sources de données métier (BDD relationnelles) comme le stocks, les ventes, etc. → C'est le travail de l'ingénieur Business Intelligence (BI),
- « tout venant » (Big Data) → c'est le travail du Data Engineer et parfois du Data Scientist.

Le but est d'obtenir des données de la meilleure qualité possible en vue de la prise de décision.

L'organisation : faire des données collectées des ensembles cohérents et spécialisés, propres à la description de phénomènes passés/présents. C'est plutôt le travail du praticien BI, mais les Data Engineers créent aussi des choses dans ce sens. Ce ne sont pas les mêmes techniques qui sont en jeu, et dans les faits, pour les « Big Data Analytics », on panache souvent données métier internes et données externes issues de l'Internet, par exemple.

L'analyse : décrire, prédire voire prescrire. C'est le travail du Data Analyst, Data Miner, du Data Scientist.

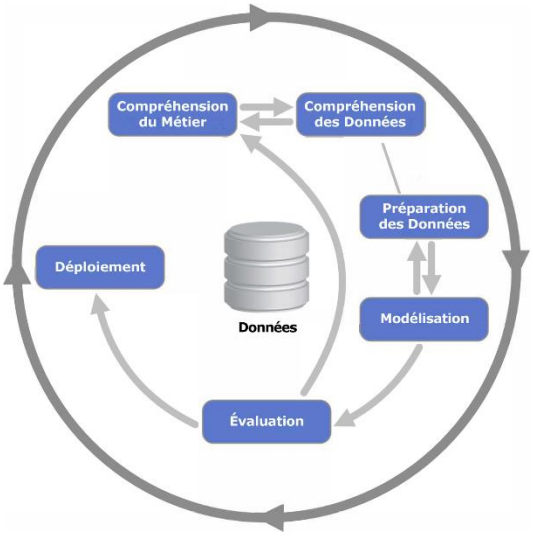
On voit bien le brassage des cultures à chaque étape : grossièrement, on a des informaticiens et des statisticiens à chaque étape, même si les informaticiens ont une part bien plus belle lorsqu'il s'agit de collecter et d'organiser les données. C'est logique en même temps, le fonds de l'affaire reste la base de données. Si l'on doit vraiment faire brosser un portrait minimaliste de qui fait quoi :

Rôles purement IT (Info)	**Rôles « gris »**	**Rôles purement Stat**
Data engineer	Data analyst	Statisticien
Ingénierie BI	Data scientist	
Data Architect	Data miner	

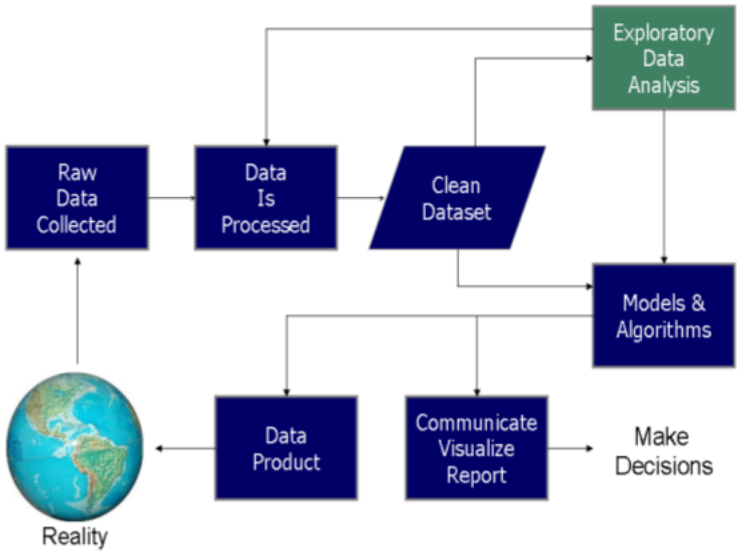
Même si on fait les choses autour de la donnée, il n'est jamais question de remplacer un expert métier, mais d'en prendre le relais et de bénéficier de l'expérience de l'entreprise. Il y a pratiquement toujours un expert pour chapeauter chaque étape ! Un ex. concret : certains établissements bancaires sont très anciens. Les experts qui y sont actuellement présents n'ont pas du tout le même âge que leur employeur, et pour autant il existe de l'expérience contenue dans la base de données ! Il est donc utile d'aller y jeter un œil et de tenter d'extraire quelque chose de la BDD, mais ce ne sont pas les données qui prendront les

décisions ; elles ne feront qu'y aider. Et d'ailleurs, les données, les modèles n'ont de valeur que parce que l'on peut en tirer quelque chose.

Un standard industriel pour le Data Mining : CRISP-DM, du fruit de la collaboration entre plusieurs géants :



On voit bien que les premières étapes sont fondamentalement en lien avec le business. C'est différent pour la Data Science :



Non seulement le métier ne figure plus sur le diagramme, mais on dirait qu'il a totalement disparu. En pratique, ce n'est pas toujours le cas, mais il est difficile de trouver des experts sur de la donnée « non-métier » (externe).

Le projet de Stat Desc 2 dans tout ça : c'est du Data Mining, on va faire de la classification sur la base de choses contenues dans une supposée BDD d'entreprise.