Analyse des données basée sur l'ACP

INGMAN1 – Projet GM1

Année 2017–2018

1 Préambule

Dans le monde des entreprises, extraire de la connaissance en utilisant toutes les informations disponibles et le faire quasiment en temps réel est devenu un enjeu stratégique dans le cadre d'une mondialisation féroce. Depuis quelques années, pour répondre à cette problématique, est apparue une nouvelle discipline : la science des données ou *Data Science*. Autour de cette spécialité s'est créé un métier *Data Scientist* dont le marché est en train littéralement d'exploser. On identifie un certain nombre de fondamentaux dans cette discipline : Mathématique, Statistiques, Analyse des données, Algorithmique, bases données et visualisation des données. Les spécialistes de cette discipline ont donc un spectre très large de compétences. Cela explique entre autre qu'ils sont aujourd'hui très recherchés.

Nous avons extrait de wikipedia.org, les paragraphes suivants :

En termes généraux, la science de données est l'extraction de connaissance de données. Elle emploie des techniques et des théories dessinées (tirées) de beaucoup de champs dans les larges domaines des mathématiques, la statistique, la théorie de l'information et la technologie de l'information, y com-pris le traitement de signal, des modèles de probabilité, l'apprentissage automatique, l'apprentissage statistique, la programmation informatique, l'ingénierie de données, la reconnaissance de formes et l'apprentissage, la visualisation, l'analytique prophétique, la modélisation d'incertitude, le stockage de données, la compression de données et le calcul à haute performance. Les méthodes qui s'adaptent aux données de masse sont particulièrement intéressantes dans la science de données, bien que la discipline ne soit généralement pas considérée comme limitée à ces données.

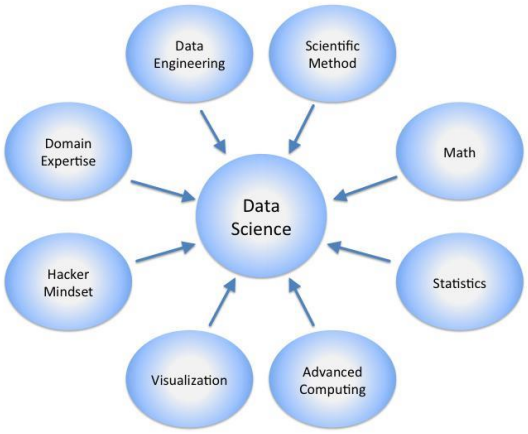


FIGURE 1 – Les différentes disciplines de la science des données.

La ***science des données*** (en anglais ***data science***) est une nouvelle discipline qui s'appuie sur des outils mathématiques, de statistiques, d'informatique (cette science est principalement une « *science des données numériques* ») et de visualisation des données. Elle est en plein développement, dans le monde universitaire ainsi que dans le secteur privé et le secteur public. Moore en 1991 a défini la statistique comme la science des données (définition reprise par d'autres dont par exemple James T. McClave et al. en 1997) et U. Beck en 2001 oppose la *science des données* à la ***science de l'expérience***, voyant une dissociation croissante entre ces deux types de science, que tendrait selon lui à encourager une société de la gestion du risque au sein d'une « *civilisation du danger* ».

2 Statistiques, Analyse des Données et Data Mining

Historiquement, les méthodes pour extraire des connaissances tournent autour de 3 spécialités : Statistique descriptive et inférentielle, Analyse des Données Multivariées et Data Mining. Historiquement, elles sont apparues dans cet ordre. Cette chronologie est étroitement liée au progrès de l'informatique. Il est d'ailleurs intéressant de noter que les deux premières sont l'apanage des mathématiciens tandis que la dernière a été créée par les informaticiens. Au-delà de ce petit débat sans grand intérêt, il est vraiment important de comprendre qu'il faut maîtriser ces trois spécialités pour être un bon Data Scientist.

On vous donne ci-dessous un extrait du cours Explore Data enseigné dans le parcours MI de l'EISTI. Il présente brièvement les trois thèmes :

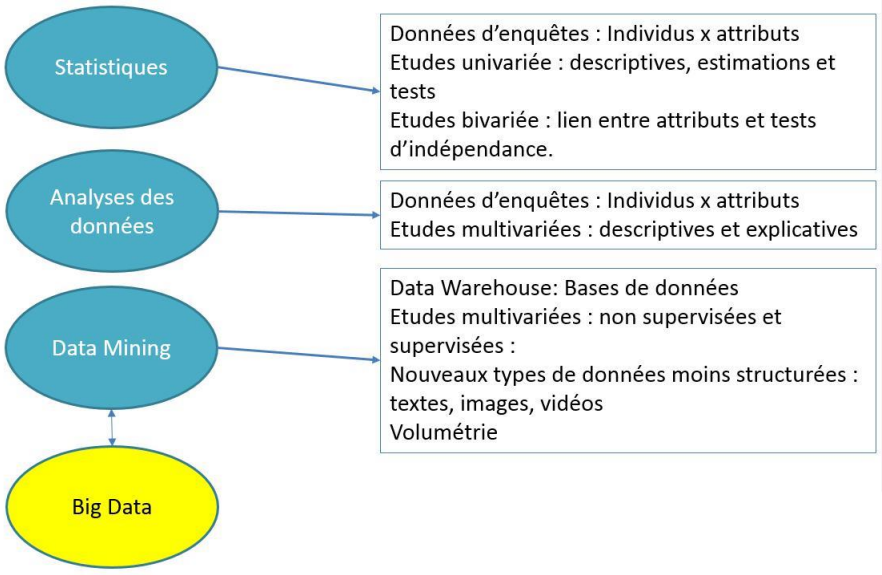


FIGURE 2 – Explore Data : Les différentes approches.

L'aspect Big Data ne sera pas abordé dans ce projet. Cependant, il faut savoir que la volumétrie

des données augmente à un tel point que les technologies informatiques de stockage et de récupération en un temps acceptable des informations sont en train d'être totalement repensées. Cette problématique et ses solutions s'appellent le *Big Data*. Nous ne pouvons pas, dans ce projet, étudier toutes les méthodes d'extraction de connaissances. Notre périmètre sera l'**Analyse en Composantes Principales** (ACP). C'est une méthode au croisement de l'analyse des données en tant que méthode descriptive et le Data Mining en tant que méthode non supervisée.

3 Analyse en composantes principales

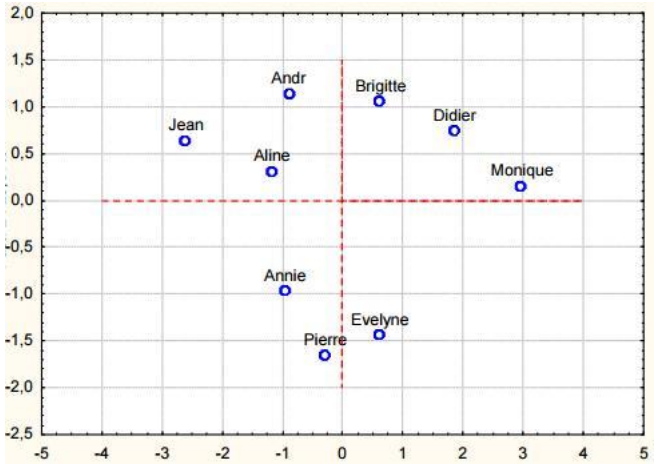
Pour résumer à l'extrême, cette méthode à pour objectif de représenter un ensemble d'observations décrites à l'aide de *p* critères avec *p* assez grand dans un espace de très faible dimension en perdant le moins de connaissances sur les différences entre les observations. La réduction de la dimension rend possible l'interprétation des données. Nous illustrons cette méthode avec un exemple publiée par Marie Chavent de l'université de Bordeaux :

3.1 Les données de l'exemple

Ci-dessous, un tableau de notes attribuées à 9 sujets dans 5 matières :

Sujet	Math	Sciences	Français	Latin	Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

3.2 La visualisation de l'exemple dans un espace à deux dimensions



Cette méthode illustre parfaitement la science des données car elle allie les points suivants :

— Les mathématiques dans le domaine des espaces vectoriels normés.

— L'informatique dans le domaine de la programmation.

— La visualisation.

4 Objectifs

Le but de ce projet est d'analyser un ensemble de données en utilisant l'ACP. Le sujet à développer est le suivant : à partir d'un ensemble des données fourni par vous-mêmes, vous devez algorithmer et programmer l'ACP en C/C++, pour analyser ces données. Votre jeu de donnés doit être assez riche pour utiliser au mieux le code réalisé.

Votre travail devra traiter les points suivants :

— Réaliser une étude mathématique de l'ACP avec la démonstration des différents théorèmes utilisés.

— Reformuler les besoins et planifier le projet en utilisant la méthode SIXO.

— Écrire des différents algorithmes pour obtenir une représentation en faible dimension en utili-sant le pseudo langage utilisé dans le cours d'algorithmique procédurale.

— Programmer en C/C++ des différents algorithmes et représentation graphique.

— Rendre une interprétation des résultats de votre jeu de données que vous avez utilisé.

5 Modalités

Ce projet est à effectuer par groupe de 2 (voir 3), la composition des groupes est libre. Les choix algorith-miques et d'implémentation sont libres. Le code doit être rédigé en C/C++. Le dépôt des différents livrables se fera simultanément sur GitLab et AREL. Les fonctionnalités de dépôt sur AREL permettent de faire simultanément les rendus sur les deux plateformes.

6 Déroulement des séances

À chaque séance, vous devrez rendre compte de votre travail à votre encadrant en déposant une fiche d'activités (modèle à récupérer sur la plateforme AREL). Le rendu final sera le dernier rendu publié sur GitLab et AREL, au plus tard à 23h59 le **samedi 02/06/2018**. Il sera également publié sur AREL. Le scénario associé à cette session 2016/2017 indique le déroulement du projet avec les étapes et les dates.

8 Évaluation

Seront notées l'étude mathématique de l'ACP et la vraisemblance de l'implémentation des algorithmes, la cohérence entre la réalisation et le cahier des charges, les techniques et méthodes d'implémentation, la gestion du suivi de projet et, bien entendu, la réponse apportée à l'objectif attendu.

Un rapport devra être fourni avec le rendu et être suffisamment pertinent, contenant notamment votre avis sur l'opportunité des algorithmes employées. Il ne sera pas noté mais pourra pénaliser lourdement la note donnée à la réalisation s'il se trouvait ne pas être satisfaisant.

En revanche, les points suivants seront sévèrement sanctionnés: archive invalide, incomplète ou mal nommée, travail rendu en retard, code pas ou peu commentés.