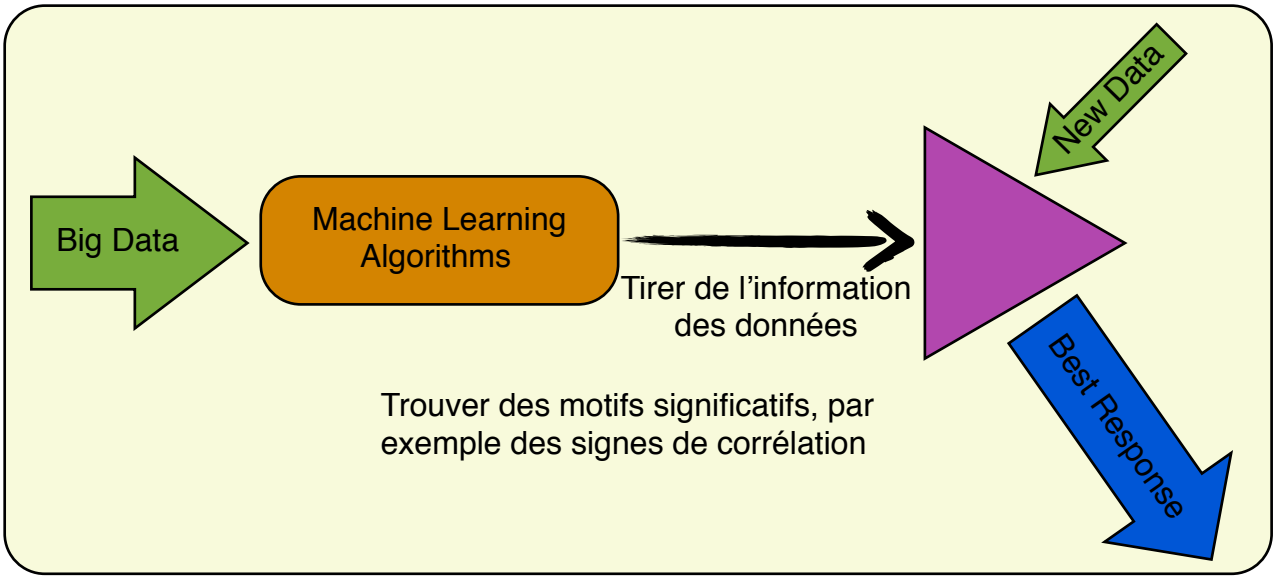
Conduite de la séance de discussion autour de «The Promise and Peril of Big Data»

I. Tour de table pour résumer brièvement le document

champ lexical : Big Data, machine-learning, patterns, information, correlation, models, best response, hypotheses, basis, questions, prediction, ontology, theory, statistics, bias, causality...

C'est quoi «Big Data» ? Exemples :

- exploitation de bases de données : web click data, dossiers médicaux, données biologiques ou géologiques (...), mesures de capteurs, réseaux sociaux...
- applications qu'on ne peut pas programmer «à la main» : hélicoptères autonomes, reconnaissance d'écriture manuscrite, traitement des langues naturelles, vision par ordinateur...
- programmes «auto-personnalisants» : Amazon/Allociné/Netflix, Google Ad...



Causalité vs. Corrélation : la corrélation est mesurable statistiquement à partir des données brutes, alors que la causalité nécessite des modèles sur les données. Si un évènement A apparaît en corrélation avec l'évènement B, cela n'implique pas que A cause B. Au moins 5 implications sont en fait possibles : 1. A cause B, 2. B cause A, 3. un facteur tiers C cause à la fois A et B, 4. combinaison des 3 possibilités précédentes, c'est-à-dire système auto-renforçant (*positive feedback*), 5. coïncidence.
Cf page Wikipedia
Exercice : quelle implication dans les cas suivants ?

1. S'endormir avec une chaussure au pied est fortement corrélé avec le fait de se réveiller avec un mal de crâne, donc s'endormir avec une chaussure au pied cause le mal de tête au réveil. (facteur tiers)

2. Plus il y a de pompiers autour d'un feu, plus on observe que le feu est important, donc les pompiers sont la cause du feu. (cause renversée)
3. Les jeunes enfants qui dorment avec une lumière allumée sont plus susceptibles de développer une myopie dans leur vie future, donc la lumière allumée la nuit cause la myopie. (facteur tiers : la myopie des parents, études publiées à partir de 1999)
4. Lors du vingtième siècle, le nombre de pirates a chuté en même temps que la température moyenne mondiale a augmenté, donc le réchauffement climatique est causé par le manque de pirates. (coïncidence utilisée de manière satyrique par le Pastafarianisme)
5. Les ventes de crèmes croissent proportionnellement avec le nombre de morts par noyade, donc la consommation de crèmes glacées cause la mort par noyade. (facteur tiers : été et chaleur)
6. Une augmentation de la pression est souvent associée avec une augmentation de la température, donc l'augmentation de la pression cause l'augmentation de la température. (système auto-renforçant suivant la loi des gaz parfaits : PV = nRT)
7. La quantité de bon cholestérol (HDL) st négativement corrélée avec l'incidence de crise cardiaque, ainsi prendre des médicaments pour augmenter le taux de bon cholestérol décroît les risques d'avoir une crise cardiaque. (facteur tiers ? remis en cause par des papiers récents)

II. Usage de modèles le long du processus

Qui a besoin de modèles, et pourquoi ? Celui qui prédit, pour orienter sa prédiction, ou bien celui qui cherche à acquérir de nouvelles données, quitte à biaiser sa recherche ?

1. Utilisation de modèles sur les données : catégories sur les données (cancer/pas cancer) ou valeurs associées aux données (prix de vente des maisons) connues a priori sans erreur ce qui entraîne un mécanisme d'apprentissage «Top-Down Fashion»
   + très bien si on veut extrapoler la catégorie/valeur d'exemples nouveaux (c'est l'apprentissage supervisé)
   — suppose qu'on connaît parfaitement les catégories/valeurs (c'est-à-dire possibilité de biais)

Exemples de biais (démarche qui engendre des erreurs dans le résultat d'une étude) dans la conduite d'une expérience, par exemple un sondage : 3 niveaux d'échantillonnage (*sampling*, échantillon de population dans l'univers, échantillon des questions posées, échantillon des réponses fournies qui dépendent du lieu, du temps où la question est posée et de la personne qui la pose...), échantillon stratifié (exemple : stratifié selon l'âge)

2. Absence absolue de modèles : pure corrélation, recherche de motifs sans a priori (exemple : séquençage génétique)
   + pas de biais (vraiment ? oui, si on ne cherche pas à extrapoler !)
   — difficile d'extraire des modèles, de poser des questions sur les données : il faut une théorie pour extrapoler en dehors du domaine de données observé
   — pour les systèmes de recommandations, cela entraîne que dès qu'on sait que ça marche sans modèles, on peut intentionnellement piéger la machine (Google bombing)

3. Corrélation utilisée pour tester et raffiner des modèles sur les données et ainsi trier «le bon grain de l'ivraie» : détection/suppression de données aberrantes (*outliers*)
   + permet de créer automatiquement des ontologies (classifications) sur les données observées, en repérant les générateurs, les vecteurs de base : «Bottom-Up Fashion»
   + la machine apprend à poser les bonnes questions sur les données
   + permet d'éliminer les corrélation fallacieuse
   — enlève un peu de variabilité du modèle

4. Utiliser la variabilité présente dans les données pour une interprétation sans biais
   + pas de nettoyage des données basées sur des modèles biaisés
   + permet de mettre à jour les modèles a priori sur les données

Exploration de données (*data mining*) vs. Statistique : pour la communauté de fouille de données, les données sont l'univers dans lequel il s'agit de repérer des motifs sans a priori sur la façon dont ont été collectées les données ou leur raison d'être. Pour la communauté statisticienne, les données sont un échantillon (nécessairement bruité) de l'univers qu'il faut donc considérer comme un reflet déformé de la réalité.

III. Qu'est-ce qui change quand on possède «beaucoup» de données ?

Qu'est-ce qui change en termes de résolution de problèmes lorsque l'on dispose de «suffisamment» de données ?

Si on dispose d'un algorithme d'apprentissage qui fournit des résultats médiocres, à quelles conditions peut-on raisonnablement penser qu'il peut être meilleur si on ajoute des données pour apprendre plus ? Dans quel cas au contraire, est-on sûr que cela n'y changera rien et qu'il est plus profitable de chercher à améliorer notre algorithme ?

Imaginons qu'on possède un algorithme d'apprentissage : comment valider son efficacité ? On considère que notre ensemble de données est partitionner en un ensemble de données sur lequel on fait apprendre notre algorithme et un ensemble de données sur lequel on teste notre algorithme : ensemble d'apprentissage et ensemble de validation. Alors qu'on fait croître la taille de l'ensemble de données de départ et qu'on partitionne l'ensemble avec une proportion fixée, l'erreur sur l'ensemble d'apprentissage croît et l'erreur sur l'ensemble de validation décroît. Si les erreurs continuellement croissent/ décroissent, cela tend à faire penser que les données ont une forte <u>variance</u> et qu'il est donc intéressant de collecter des données supplémentaires pour faire converger les deux erreurs vers une valeur commune «faible». Au contraire, si les erreurs convergent rapidement vers une valeur commune «forte», cela tend à faire penser que l'algorithme possède un fort <u>biais</u> et qu'il est inutile de faire grossir la taille des données.