L'intelligence d'affaires:
la statistique dans nos vies de consommateurs

Jean-François Plante, HEC Montréal
Marc Fredette, HEC Montréal

Congrès de l'ACFAS, Université Laval, 6 mai 2013

Intelligence d'affaires

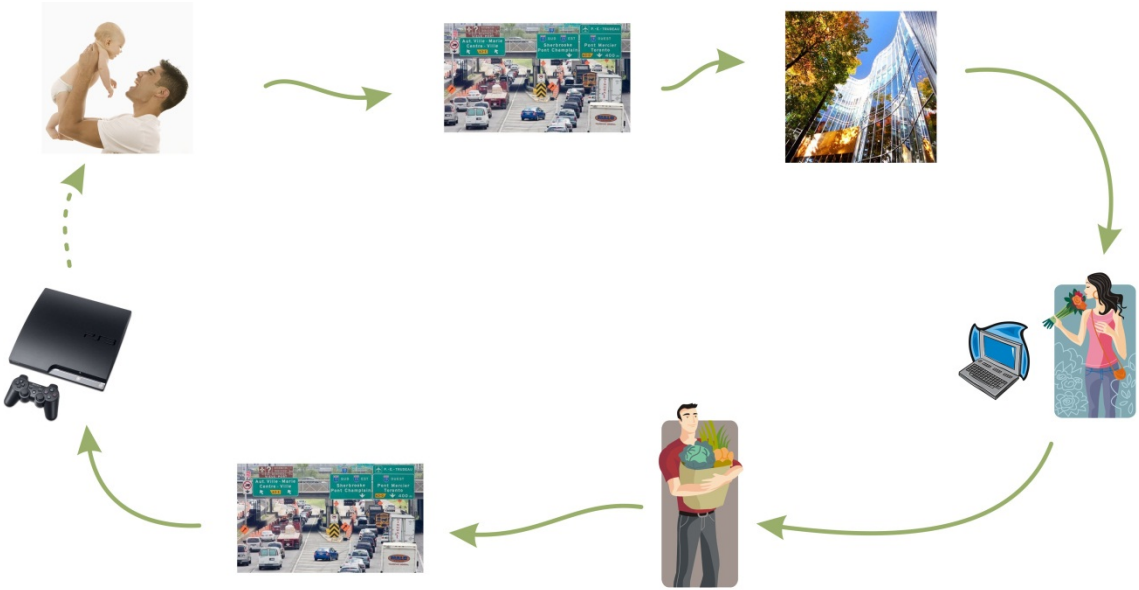
[Wikipedia] Business intelligence (BI) is a set of theories, methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information for business purposes.

jfplante.ca/Talks/ACFAS2013.pdf

Plan :

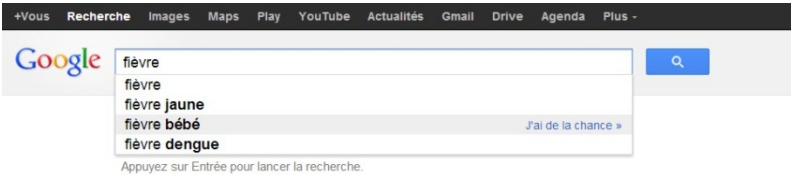
1. Une vie dans la journée d'un consommateur
2. Présence (cachée?) de la statistique dans cette journée
3. Les défis reliés à l'intelligence d'affaire
4. Conclusion : importance de l'intelligence d'affaires

Une journée dans la vie d'un consommateur



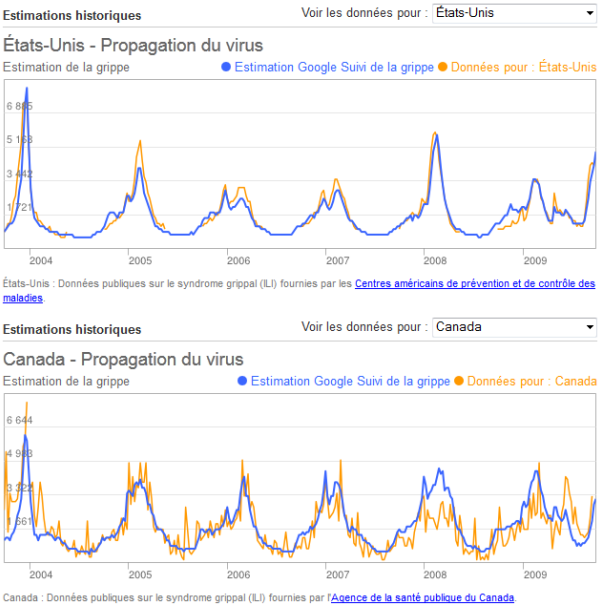
1. Lever du bébé – il a une fièvre légère et des petits boutons

Réflexe moderne :



Google TREND

Le comportement de la population contient de l'information! (Article dans Nature, limites)

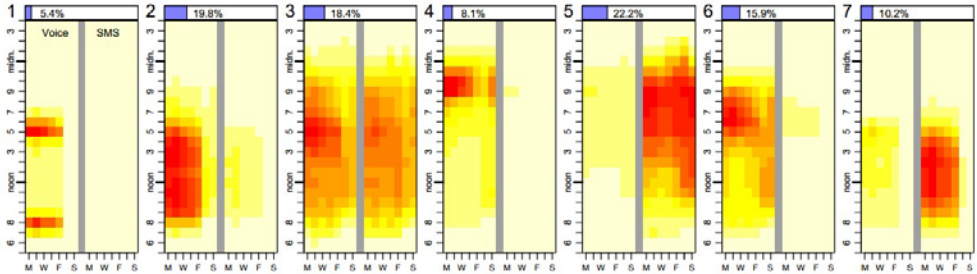


2. Se rendre au travail

- Même si on ne parle pas au téléphone en conduisant, on laisse souvent l'appareil sous tension.
- L'appareil sous tension se connectera à différentes tours de communications…
- AT&T research :
  - Estimation de la vitesse des véhicules (et donc du trafic).
  - Estimation de niveau d'utilisation des routes et de l'empreinte carbone de différentes communautés.

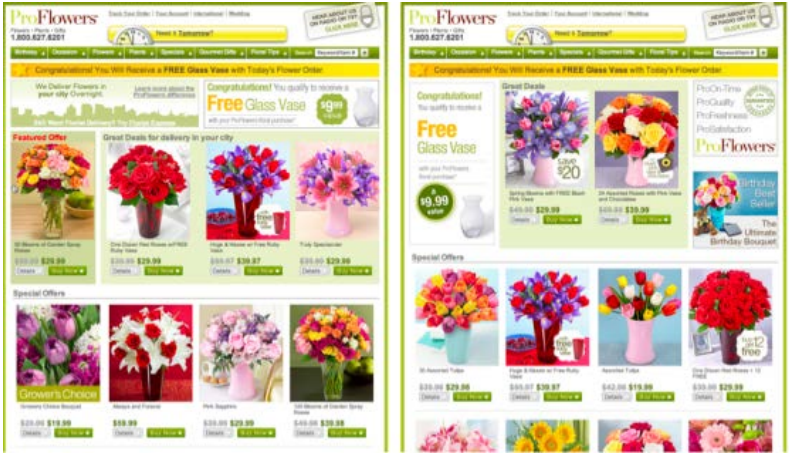
Parlant de cellulaire, ils ont l'avantage d'être associés au même individu…
Même avec des données anonymes, on peut donc :

- Détecter où vous demeurez et où vous travaillez (puisque vous téléphonez le plus souvent de la maison et du travail).
- Vous classer dans un segment de consommateurs selon votre usage du cellulaire.



3. C'est la fête de mon épouse, je lui achète des fleurs

Je visite le même site de fleuriste, mais il est différent?!?



Pourquoi?

Plusieurs sites font du « A/B-testing » (c'est le terme utilisé en Marketing).

Par exemple, en choisissant une femme plutôt qu'un homme, DHL augmente son taux de conversion de 15% (source non scientifique)...

Ici, il y a une vraie randomisation. On peut élaborer un plan d'expérience à plusieurs facteurs et appliquer les tests d'hypothèses bien connus.

4. Après le lunch : pause YouTube

Les vidéos sont rarement des publicités déguisées…

Pour chaque nouveau vidéo YouTube (1h de nouveau vidéo à chaque minute) :

- La trame sonore de la vidéo est transcrite automatiquement.
- Des algorithmes de « Text Mining » transforment le texte en valeurs numériques.
- Une analyse factorielle permet de réduire le nombre de variable.
- Une régression logistique est utilisée pour évaluer la probabilité que la vidéo contienne des publicités cachées.
- La validation croisée permet d'évaluer la performance du modèle.

Beaucoup de puissance de calcul est requise, mais ce n'est pas un problème chez Google, le 4ème plus important constructeur de serveurs au monde! Ils utilisent des bibliothèques R conçues à l'interne pour le calcul parallèle.

5. Il faut passer à l'épicerie

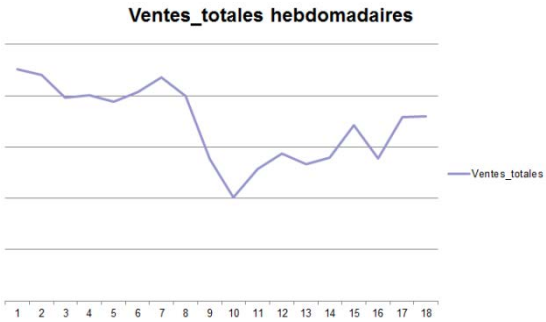
Grâce à ma carte de fidélisation, j'obtiens des rabais et des primes!

Qu'est-ce que le marchant peut bien en retirer?

a) Évaluer les effets de l'arrivée d'un compétiteur (projet supervisé à HEC) :

- Ouverture d'une nouvelle épicerie à proximité.

- Quel est l'effet sur la clientèle? On peut savoir quel type de clients on perd, et donc mieux cibler notre stratégie pour retenir les clients qui sont à risque de quitter.



- Avec PROC MIXED :
  - Si le concurrent est SuperC : les clients achetant du congelé sont plus à risque de quitter.
  - Distance entre le domicile et le compétiteur : plus importante en milieu urbain qu'en milieu rural.

b) Prédire les achats d'un client (Market Basket Analysis)

Même sans programme de fidélisation, on peut analyser les articles que les clients ont tendance à acheter simultanément. Des algorithmes ont été développés pour détecter les règles d'association (probabilités conditionnelles) les plus fortes. Évidemment, avec le programme de fidélisation on peut vous faire parvenir directement les offres ciblées.

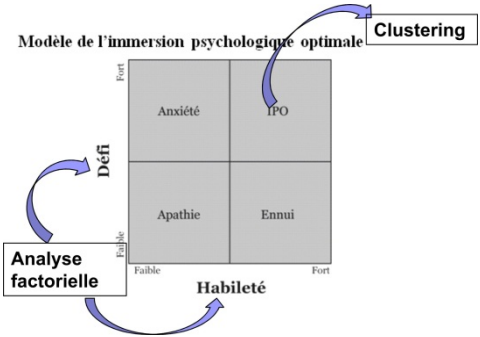
Exemples :

- Target envoie des promotions sur les articles de bébé à une adolescente aux États-Unis. Son père outré ne savait pas qu'elle était enceinte.
- Distributeur de pièces d'auto au Québec *(projet supervisé à HEC)*
    - En plus du MBA, régression binomiale négative pour évaluer l'effet des spéciaux et des promotions.

6. En soirée, je peux enfin me reposer un peu

Suis-je à l'abri de l'intelligence d'affaires?

Étude de l'efficacité du placement de produits dans les jeux vidéo (Mémoire à HEC).



Jeux en ligne : analyse du comportement des joueurs pour prédire le taux d'attrition (churn) et maximiser la rétention.

Méthodes statistiques utilisées dans les exemples

- Analyse factorielle
- Régression (linéaire, logistique, binomiale négative, etc.)
- Modèles linéaires mixtes et/ou généralisés
- Analyse de survie (régression de Cox)
- Analyse de regroupement (k-means, k-nn, méthodes hiérarchiques, etc.)
- Autres modèles d'apprentissage (svm, arbres de décision, etc.)
- Bagging, boosting, forêts aléatoires
- Validation croisée
- Plans d'expérience
- Analyse de correspondance
- Courbes ROC, lift charts
- etc.

Défis de l'intelligence d'affaires

Au niveau modélisation :

- Données massives, peu ou mal structurées.
- Données d'observations (peu de données randomisées).
- Nettoyage et préparation des données nécessaire.
- Hypothèses des modèles difficiles à respecter (est-ce si grave)?
- Changements dans les données (Google Flu).
- Comparabilité des données (plates-formes mobiles).

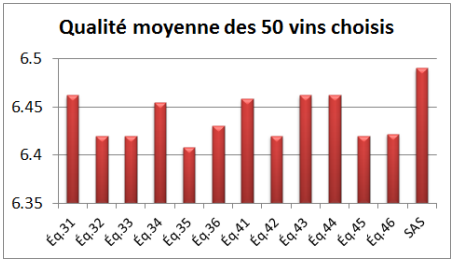
Au niveau communication :

- Il faut être capable de communiquer le résultat des modèles et leur qualité a des gens qui n'ont pas, ou peu, de formation quantitative.
- Dans la plupart des contextes, une régression linéaire multiple est déjà un outil sophistiqué. Imaginez une régression de Cox!

Mauvais modèle vs modèle optimal

Nous sommes entraînés à toujours chercher le meilleur modèle. En pratique, il y a souvent peu de différence entre une solution naïve et une solution optimale.

Exemple : Devoir au MBA, choisir 50 vins dans une liste de 500 en se basant sur une régression linéaire ajustée sur un autre échantillon de 1000 vins.



- La qualité moyenne du vin est de 5,44.
- Les équipes obtiennent entre 6,44 et 6,46.
- *SAS rapid predictive modeller* atteint 6,49 en utilisant des modèles beaucoup plus complexes (et du boosting).
- Avec un choix au hasard, on aurait en moyenne 5 bons vins dans le top 50. Les équipes en ont toutes 18 ou 19.
- On aurait aussi 5 mauvais vins parmi les 50 pires. On en obtient 0 ou 1.

Concours Netflix :

Base de données de 100 480 507 de notes données par 480 189 usagers à 17 770 films. Il faut prévoir la note que les usagers donneront à un nouveau film.

- Prévision = cote moyenne du film : $RMSE = 1.0540$
- *Cinematch* (modèle linéaire naïf après beaucoup de nettoyage et de préparation des données) : $RMSE = 0.9525$

Il y a 1 000 000 $ en jeu pour la première équipe à améliorer le RMSE de *Cinematch* d'un 10 % additionnel (i.e. atteindre 0.8572 sur l'échantillon test).

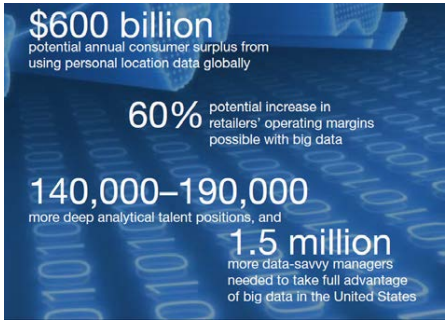
Lancement du concours : 2 octobre 2006.

Participants : 41 305 équipes de 186 pays.

Fin du concours : 21 septembre 2009.

Importance de l'intelligence d'affaires

« Big data revolution »



- Rapport de McKinsey International :
- La Maison Blanche investit 200 millions $ en recherche.
- Hal Varian, économiste en chef chez Google :
  "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding.",
  — NY Times, août 2009
- Programme d'intelligence d'affaires à HEC Montréal :