C23_TD et cours **Collecter les données – l'échelle du big datas**

Activité 1 : Pourquoi trier les données collectées en informatique ?

Les sciences et technologies du numérique connaissent actuellement un changement d'échelle majeur par la taille et la complexité des données qu'elles manipulent. L'information est devenue un objet monnayable, négociable, et une cible d'investissements.

Ces données sont collectés à partir d'objets connectés, à partir de l'activité des internautes sur les sites de e-commerce (voir le document suivant sur le marketing digital), des statistiques d'utilisation de produits, de véhicules, de bâtiments, voire des données collectées suite à des évènements naturels, biologiques, etc.

Ces données peuvent servir les domaines de la santé (suivi de la propagation d'épidémies, aide au diagnostic,… des transports (analyse de flux,), de l'environnement (prévisions météorologiques, contrôle de la pollution), mais aussi dans l'analyse de la clientèle dans l'industrie et le commerce.

En se basant sur des informations passées, les techniciens spécialisés dans l'observation des grosses données (big datas) peuvent ainsi faire des prévisions dans chacun de ces domaines.

Jusqu'à maintenant, on ne savait pas bien techniquement gérer les gigantesques flux de données qui sont générés par ces exemples. Il était trop couteux de stocker, analyser et recouper entre elles via des algorithmes statistiques. Tout ceci doit se faire de manière rapide, aussi rapidement que les données affluent. Il faut aussi savoir sélectionner les bonnes des mauvaises données, sous peine d'affecter les conclusions générales. Mais maintenant, on est en capacité d'exploiter ces grosses quantités de données pour les transformer en connaissance et, ou, en profit.

Dans ce flot d'information, il est alors capital de savoir faire le tri pour faire l'analyse et ainsi obtenir un axe de travail pertinent.

Questions

1. Parmi les exemples de données collectées : lesquelles sont sous licence libres/propriétaire. Citer deux exemples de chaque.
2. Pourquoi est-il si important de faire le tri des données avant de les stocker et les analyser ?

Activité 2 : Marketing digital et big datas

A quel moment parle-t-on de données massives (Big datas) ? Faut-il collecter des données à une échelle mondiale, ou bien, le simple suivi de client dans un petit commerce de quartier, suffit-il à franchir ce seuil de « données massives » ?

Prenons l'exemple d'un petit supermarché de ville, avec 300 références produits sur ses rayons. La quantité de données générée ne ferait normalement pas partie de Big Data, qui s'entend souvent dans la zone du tera voire du pétaoctet.
On peut avoir intérêt à y faire une analyse de vente croisée, dite de « règles d'association » sous la forme : *« Si le client achète le produit X alors il (elle) achète le produit Y »*. Alors, le nombre de **combinaisons possibles , si l'on prend en compte toutes les possibilités** par client (par ligne dans un tableau) vaut $2^{300} \approx 10^{90}$. Ce qui est énorme…

Cette échelle est bien sûr gigantesque et ne peut être directement traitée par l'humain ni par nos ordinateurs de manière directe. Et il ne s'agit là que de 300 produits : imaginons l'échelle *« hypermarché»* ou *« géant de l'e-commerce »* ! Pour ces grands magasins en ligne, la collecte de donnée doit permettre de construire un portrait assez fiable du client, et être capable de le servir beaucoup mieux. Cette collecte peut consister alors à enregistrer les clics et le mouvement de la souris, l'achat, les requêtes de recherche, ce qui va aussi constituer un nombre important de données.

C'est donc aussi dans ces cas que nous pouvons parler de Big Data et que nous avons besoin d'algorithmes.

Il faudra diminuer le nombre de combinaisons, en réalisant une sélection et un tri des données à collecter selon des questions pertinentes.

Les mathématiques nous offrent de nouvelles façons de représenter ce cas du *panier de la ménagère*, et d'en sélectionner les bonnes informations. Ces techniques relèvent de l'optimisation algorithmique et de la théorie des graphes.

Questions :

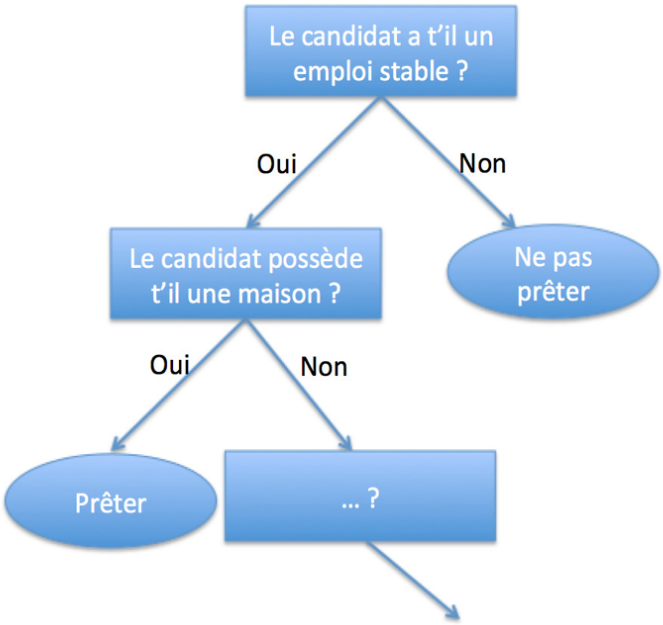
1. On cherche à numériser l'information suivante : pour un client donné, l'article X est acheté ou n'est pas acheté. Combien de combinaisons sont possibles ? Combien de bit(s) sont nécessaire(s) ?

2. On revient maintenant sur l'exemple proposé du commerce contenant 300 articles.
Chaque panier de client passant en caisse est décrit par les 300 bits correspondants aux 300 articles du magasin.
   a. Combien d'octets sont nécessaires pour enregistrer les achats réalisés sur une seule journée ? (hypothèse d'un client par minute).
   b. Vérifier que le volume de données générées n'est pas assez important pour atteindre l'échelle du *big datas*.
   c. Expliquer pourquoi ce volume est trop important pour être traité *directement par un humain, ou par un ordinateur de manière directe (sans l'utilisation d'un algorithme optimisé, issu du monde de l'intelligence artificielle)*.

Activité 2 : des exemples de tri de données utilisant un graphe

Partie 1 : *Accorder ou non un prêt bancaire*

Un banquier peut utiliser un arbre de décision pour déterminer s'il fera ou non le prêt au candidat :

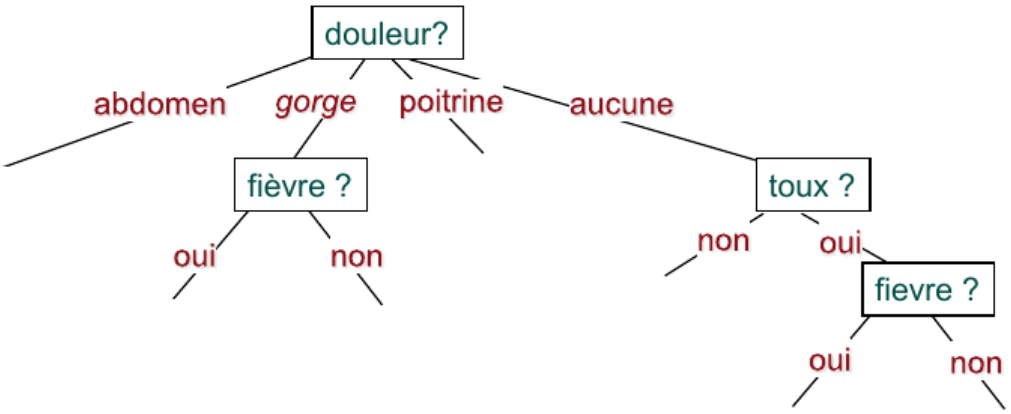


Accorder ou non un prêt bancaire.
Chaque individu est évalué sur un ensemble de variables testées dans les nœuds internes.
Les décisions sont prises dans les feuilles.

Travail : prolonger la branche au niveau du dernier noeud .. ? avec une nouvelle condition.

Lorsque les données sont catégorielle, leur traitement consiste à les classer. On utilise alors un algorithme de classement, qui peut être représenté par un arbre, comme celui vu dans l'exemple précédent.

Partie 2 : *algorithme du médecin*
Pour l'arbre suivant, les valeurs cibles (celles des feuilles) ne sont pas indiquées.



Travail :

1. Compléter cet arbre avec ces valeurs cibles, déduites de l'algorithme proposé :

```
si douleur == 'abdomen' alors :
      maladie = 'appendicite'
sinon si douleur == 'gorge' alors :
      si fievre alors :
            maladie = 'rhume'
      sinon maladie = 'mal de gorge'
sinon si douleur == 'poitrine' alors :
      maladie = 'infarctus'
sinon :
      si !toux :
            maladie = 'rien'
      sinon :
            si fievre :
                  maladie = 'rhume'
            sinon :
                  maladie = 'refroidissement'
```

2. On donne ici les donnée en tableau de quelques patients, décrits par plusieurs variables catégorielles (Toux, Fievre, Poids, Douleur).

	Toux	Fièvre	Poids	Douleur
Marie	non	oui	normal	gorge
Fred	non	oui	normal	abdomen
Julie	oui	oui	maigre	aucun
Elvis	oui	non	obèse	poitrine

   a. Utiliser l'arbre de décision pour déterminer leur maladie.
   b. Quels avantages y-a-t-il à utiliser un arbre de décision, plutôt qu'une liste présentant toutes les combinaisons possibles ?
   Exemple de liste (et donc de combinaison possible) : [toux, fievre, poids, douleur, maladie]

Remarque : On verra par la suite que ce type de représentation peut être très efficace pour classer des grandes collections d'objets, et établir des correspondances entre ces objets. C'est ce qui est réalisé par les algorithmes en *intelligence artificielle*.