Innovation, marché et science des données (IMSD)
2020

DATAMINING

Vincent CHALMEL

Ingénieur ENSAI – Systèmes d'Information Statistique

EXPLORE
Solutions de veille et bases de données

Diplômé de l'ENSAI en 2011

- spécialisé en Systèmes d'Information Statistique
- première formation de dataminer / datascientist « moderne » en France
- (datamining, informatique et technologies de data management)

Expériences Professionnelles

- Industrie
  - Inergy Automotive – stabilité des processus de productions de systèmes de carburants
- Développement Logiciel
  - Coheris SPAD / Liberty – solutions de datamining et Business Intelligence
- Marketing et relation client
  - Epsilon International – email marketing
  - Coheris CRM – connaissance client
  - Bisnode / Brand & Consumer TECHNOLOGIES – nouveaux paradigmes de la relation client, « self care »
- Recherche & Développement
  - Explore : Traitement et qualification automatisé de volumes massifs textes

EXPLORE – Solutions de veille et bases de données

Des signaux générateurs de business

Permis de construire des particuliers + **120 000 permis / an**

Marchés publics et attributaires
+ 850 000 avis / an

Patrimoine immobilier
+ 18 millions de biens
+ 2,3 millions de propriétaires identifiés

Mouvements et projets immobiliers des entreprises
+ 25 000 projets / an

Annonces légales et juridiques
+ 2,7 millions d'avis / an

Actualités économiques et financières des entreprises
+ 60 000 évènements / an

Positionnement de l'enseignement

Prérequis

- Connaissance générale des concepts mathématiques usuels de la Statistique, des méthodes de statistique multivariée, d'analyse factorielle et de régression
- Appréciation des enjeux technologiques actuels

Articulation dans le cursus du master

- Structurer et valider les acquis en statistique et usage des logiciels d'analyse
- Positionner les techniques et technologies dans l'ordre d'un projet de datamining
- Fonder la réflexion sur les usages et l'interconnexion des apprentissages suivants

Plan du Cours

I. Introduction et Concepts
   1. Introduction au datamining
   2. La donnée, Définition
   3. Le projet de datamining

II. Méthodes statistiques du Datamining
   1. Statistique Descriptive, business Intelligence et organisation de la donnée
   2. Statistique Exploratoire et typologies
   3. Méthodes de modélisation

III. Ouverture Applicative
   1. Exemple de données non structurées, le text-mining
   2. Connaissance Client
   3. Moteurs de recommandation

INTRODUCTION ET CONCEPTS

1. Introduction au datamining
    1) Concepts Clefs et définitions
    2) Problématiques historiques et nouvelles
    3) Vers une définition précise du datamining
2. La donnée, Définition
    1) Données structurées et non structurées
    2) Types de données et usages
    3) Outils et processus de DataManagement
    4) Audit de données
3. Le projet de datamining
    1) Identifier les problématiques
    2) Identifier et assembler la donnée
    3) Cycle Analytique
    4) Technologies et production

INTRODUCTION AU DATAMINING

1) Concepts Clefs et définitions
   A. Donnée et Connaissance
   B. L'individu, Artefact Statistique ou Personne Humaine
   C. Discipline et méthodes
2) Problématiques historiques et nouvelles
   A. Des données en croissance
   B. Innovation technologiques
   C. Nouveaux usages
3) Vers une définition précise du datamining
   A. De la donnée (data) au savoir (knowledge)
   B. Fondement et pertinence de la démarche datamining
   C. Data Mining et Big Data

Selon le dictionnaire :

- Ensemble de **données** d'observation relatives à un groupe d'individus ou d'unités (souvent pluriel).
- Ensemble des **méthodes** qui ont pour objet la **collecte, le traitement et l'interprétation** de ces données.
- Ensemble des **données** **numériques** concernant un phénomène quelconque **et dont on tire certaines conclusions.**

Etymologie :

- 1875 Gottfried Achenwall introduit le néologisme *statistik* (all.) de *statista* (it.) « Homme d'État »
- Littéralement : Ensemble des **connaissances** nécessaires [à l'Homme d'Etat] pour **gouverner**.

Véritable paradoxe, entre **donnée**, **méthodes**, et **connaissances**…

Et pourtant véritable base historique évidente du datamining, de l'exploration des données…

DIVERS HÉRITAGES STATISTIQUES

EXPLORE

Depuis Les « inventaires royaux » (Sully, Colbert, Vauban) la statistique publique est l'un des outils majeurs de tous les régimes politiques Français

Insee

Depuis le XVIIe siècle la théorie des probabilités est immédiatement appliquée à l'économie

Dès le XIXe siècle la statistique anglo-saxonne se développe au service des banques et assurances

XXe La statistique Russe avant sa purge politique triomphe en socio-démographie, et planification

Cinquante dernières années, Intelligence artificielle et bases de données informatiques

Notion de KPI (*Key Performance Indicator* - Indicateur Clef d'Efficience)

Indicateur, pilotable, explicable, dont les **facteurs** sont identifiables, dont l'évolution est **modélisable** :

- N'est pas un montant (un revenu, un coût), un indicateur financier
- **Mesuré** fréquemment
- **Compréhensible** par la direction
- Simple et métier, ses tenants et aboutissants évidents à tous
- Orienté équipe, **responsabilités** identifiables
- Significatif en terme d'impact
- N'induit pas de comportements dysfonctionnels

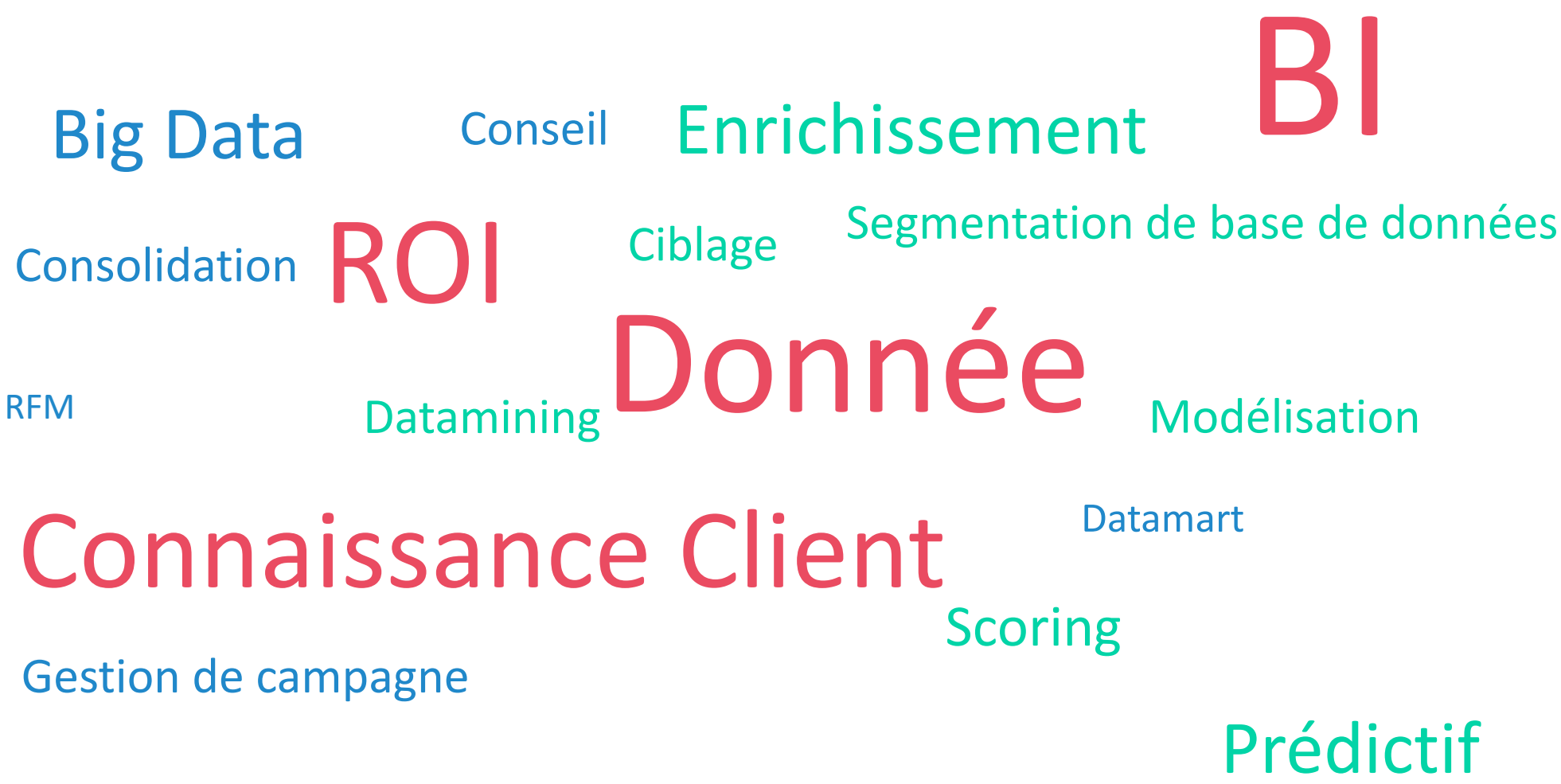
Le KPI répond à un besoin de **pilotage**, de **mesurer** et **améliorer** la performance, de prouver la stabilité de la qualité.

Problématiques de gestion de la relation client (CRM)

- Traitement des retours clients
- Etudes de satisfaction
- Phénomènes d'attrition (churn), etc.
- Notion de ROI

- Intelligence Artificielle
- Algorithmes d'apprentissage (supervisés ou non)
  - Réseaux de neurones
  - Machines à vecteurs de support
- Traitement de graphs, de signaux, d'images…
- Bases de données
- Processus de traitement et transformation de **grands volumes** de données

EXPLORE


BI
Big Data
Conseil
Enrichissement
Segmentation de base de données
Consolidation
ROI
Ciblage
Donnée
RFM
Datamining
Modélisation
Datamart
Connaissance Client
Scoring
Gestion de campagne
Prédictif

NOTION DE DATAMINING

EXPLORE

- La notion de datamining apparait véritablement en 1995
  - Accommodation de concepts et techniques issues du développement des sciences de l'information et de l'informatique, réconciliées avec la théorie statistique dont ils s'étaient éloignés

- Data-mining ou datascience ?
  - Part du data-management, des processus d'import et de qualité des données dans le projet
  - Importance conjointe du « machine learning » et du data-mining
  - Préoccupation pour l'analyse des Big Data
  - Maitrise des formalismes de stockage et de traitement distribués des données apportées par les technologies NoSQL, Hadoop, MapReduce…

Donnée et Connaissance

- Quelle définition de la donnée ?
- Comment évaluer si une donnée est exploitable ?
- Qu'est ce que l'information ? La Connaissance ?

EXPLORE

Connu ou admis comme tel

Renseignement fondamental

Représentation conventionnelle

Collection ou opportunité

La Donnée

Selon le Larousse

Point de départ pour une recherche

Information à traiter

Quantifiable

Soumis aux statistiques

CONCEPTS CLEFS – DONNÉE ET CONNAISSANCE

la Donnée

Information

Connaissance

- Information
  - Une information s'échange, se produit, se présente, elle est un outil de transmission, de communication des faits.
  - L'information est la mise en forme de la donnée, l'extraction des faits notables de la masse des messages possibles apportés par les données, à l'aide de conventions, pour être stocké, traité, ou communiqué.

- Connaissance
  - La connaissance est la capacité à comprendre, connaître les propriétés et caractéristiques, prendre conscience des traits spécifiques de l'objet exposé à travers les données (l'individu statistique)
  - La connaissance est un produit d'analyse, l'issue de l'exercice de l'apprentissage, servant à définir une réalité,

EXPLORE

L'individu, Artefact Statistique ou Personne Humaine

- Notion d'individu statistique et d'échantillon
- Notion d'événement, de variable « temporelle »
- Exemples : Le cycle de vie client, Donateur Caritatif, Cycle de production industriel, log serveur

L'individu statistique est un objet mathématique

Soit une Population Ω, on note (w_1,…, w_N) les individus de la population d'effectif N

Soit une Variable X suivant une loi de probabilité quelconque, les individus réalisent (x_1,…,x_N)

- L'individu est l'élément unitaire de la « population » d'un problème probabiliste
- L'individu est exprimé sur chaque variable étudiée
- La nature des individus statistiques, l'effectif de la population ou encore la capacité de l'échantillonner conditionnent la résolution d'un problème statistique

QUELLE NATURE RÉELLE DE L'INDIVIDU EN DATAMINING ?

- Une table de base ou un fichier de données ne porte pas de notion d'individu. (Lignes, Objets…)
- Individu choisi en fonction de la problématique
- Dicte les choix d'agrégats et mise à plat des données
- Oriente la définition des nouvelles variables et leur enrichissement
- Possibilité d'alterner des scopes successifs, plusieurs univers

Ex : Si je reçoit une base de tickets de caisse, quelle est sa granularité minimale ? Le ticket ? Un exemplaire du produit ?

Ma problématique porte t-elle sur un panier ? Une transaction ? Un client considéré sur l'ensemble de son cycle de vie ? Un produit ? Une gamme de produit ? De multiples « individus » peuvent faire l'objet de mes conclusions.

GARDER EN TÊTE LA VÉRITABLE IDENTITÉ DE L'INDIVIDU.

EXPLORE

Certaines variables modifient la nature même de l'objet étudié :

- Définissent des populations qu'on ne peut étudier en commun, n'obéissent pas aux mêmes lois et modèles

  Ex : nature d'une transaction, ancienneté d'un client

- Représentent un « temps »
  - Peuvent être une chronologie réelle (ancienneté, âge, année d'une enquête)
  - Ou en différer complètement (CSP, taille de ville, nombre d'occurrences d'un évènement dans un cycle de vie client)
  - Mais les individus doivent pouvoir être comparés lors de leur progression sur cette variable, que cette progression ne puisse être inversée
  - Certaines méthodes statistiques exploitent cette notion (ex : l'étude des trajectoires factorielles des modalités de variables)

Bien garder en tête les particularités, la réalité métier derrière le tableau de donnée, qui orienteront toute votre étude (ex : don à une ONG, « montants ronds » et discrétisation =/= montants d'achats)

Discipline et méthodes

- De la statistique publique au datamining moderne
- Multiples disciplines, synonymes ou faux-amis
  - Informatique décisionnelle
  - Analyse de données
  - Connaissance Client
- Une Palette de méthodes considérable
- Des démarches parfois complémentaires

DE LA STATISTIQUE PUBLIQUE AU DATAMINING MODERNE

EXPLORE

- On peut comprendre le datamining en « creux » par tout ce qu'il apporte à une démarche de statistique institutionnelle comme ce qu'il peut y puiser vis-à-vis des plus récentes nouveautés
- Réactivité, volumétrie et volatilité des données
- Exigences de ROI et d'estimation quantifiable de la pertinence des résultats
- Innovation vis à vies des technologies informatiques les plus modernes
- Ne pas sous-estimer malgré tout l'apport des méthodes factorielles, de classifications et clustering, dans les thématiques les plus récentes (réduction de dimensionnalité, apprentissage non supervisé sur de larges volumétries d'utilisateurs)

L’informatique décisionnelle (business intelligence)

Solution d’aide à la décision, informatisée

- Née en 1970 avec les premiers infocentres
- D’abord centrée sur les activités internes les plus génératrices de données (Comptabilité, planification budgétaire) puis étendu avec l’informatisation de la donnée

- L’informatique décisionnelle a pour but de consolider l’information des bases de données de l’entreprise
  - Consolider l’information depuis la donnée brute (Agrégations, Création de nouvelles variables calculées, Croisements, Graphiques)
  - Exploiter des données de sources multiples par des jointures parfois complexes
  - Permettre des visions sous des angles multiples, des clefs d’entrées diverses

L’informatique décisionnelle (business intelligence) (2)

L’informatique décisionnelle a pour but de rendre disponible cette information à grande échelle

- Extraction, Transfert et consolidation de données (ETL)
- Centralisation des données de l’entreprise dans un datawarehouse unifié
- Structuration, historisation, organisation en datamarts
- Analyse multidimensionnelle (OLAP ou relationnelle)
- Création de référentiels de données orientés métier (Univers BO, Dictionnaires Coheris Liberty…)
- Template de reportings
- Applications agile de requêtage, organisation de tableaux, de graphiques et création de tableaux de bord
- Travail collaboratif sur les requêtes
- Emission d’alertes, d’emailings de reporting
- Application de reporting intégrées à un intranet d’entreprise

Analyse et modélisation des données

Répondre à une problématique précise par l'exploitation de données diverses

- Déterminer une problématique, des besoins
- Choisir les données
- Auditer la qualité et l'exploitabilité des données
- Réaliser à l'aide de la BI ou d'outils ad-hoc une prise de connaissance et d'information des données
- Mesurer un phénomène
- Modéliser ce phénomène et ses facteurs
- Dégager des tendances

- Utiliser une palette complète de techniques statistiques et informatiques
- Etablir un rapport d'analyse, en collaboration avec le métier
- Mettre en production / Industrialiser les modèles (recalcul pour améliorer l'apprentissage et stockage en base des affectations de chaque individu

Analyse et modélisation des données (2)

L'analyse et la modélisation des données par le datamining ne se borne pas à l'exploitation des données stockées dans les bases métier

- Exemples de projets :
  - Détection de visage ou autres types d'images par les moteurs de recherches
  - Prévision de consommation électrique géolocalisée pour les smartgrids
  - Localisation et suivi de personne dans une pièce pour une application de communication ou de réalité virtuelle
  - Identification de la signature radar d'un certain modèle d'avion…

Ce type d'application R&D complexes, embarquées à des systèmes informatiques ou industriels ne nécessitent pas de méthodes statistiques plus exotiques, l'apprentissage statistique couvre tout types de champs d'application

Il faut par contre savoir prendre en main des données non structurées, en temps réel, sans latence dans l'application de modèle et s'adapter aux spécificités du projet et des données

- Ex : détection de visages dans banque d'image, modèle à couche (exclure images sans teintes chair, identifier la zone visage dans une image, reconstruire le visage par réseau de neurone, etc.) Chaque modèle a des performances et des risques complémentaires

La connaissance client, application concrète du datamining au quotidien du marché

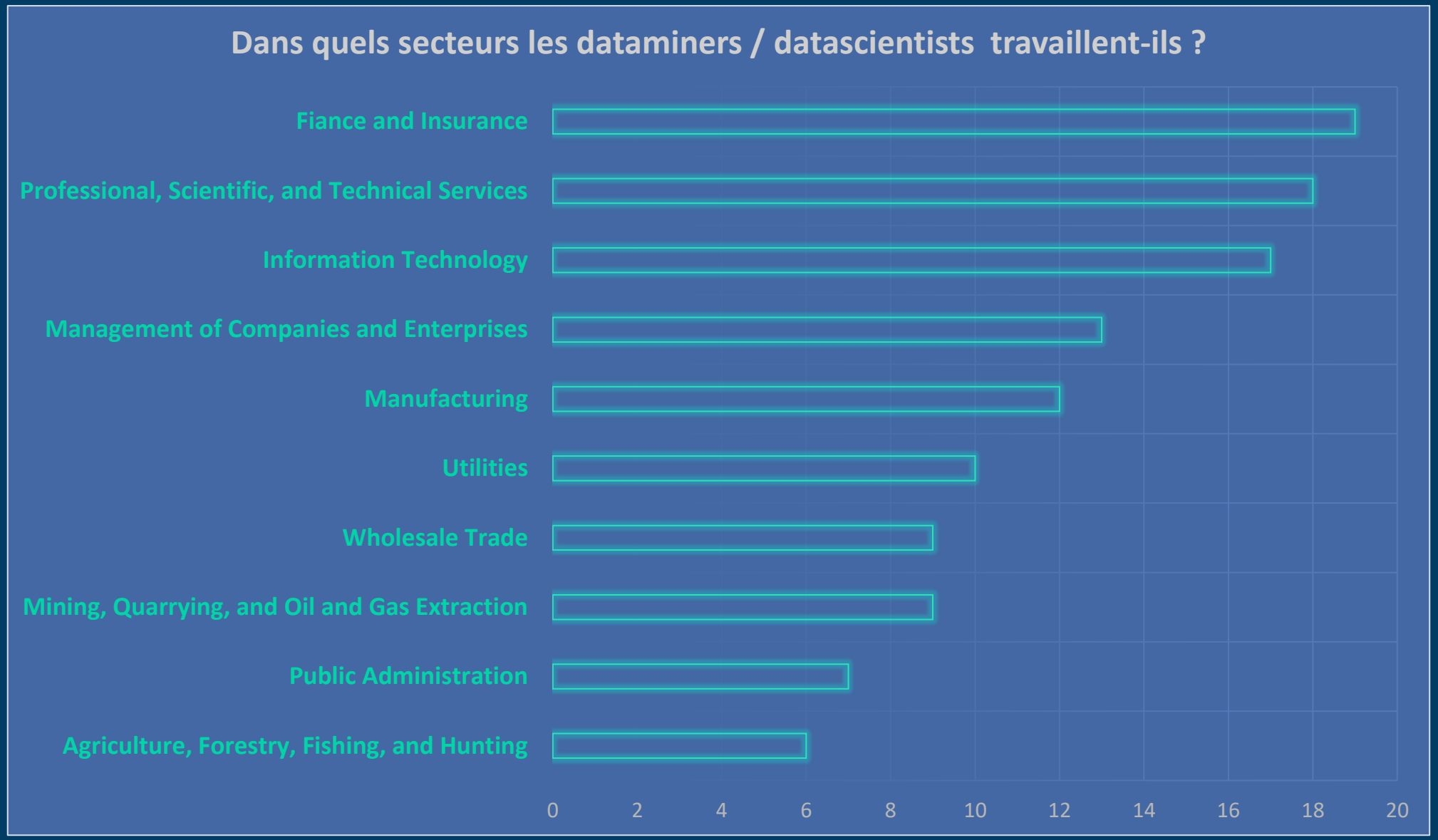
Utiliser l'information donnée par l'informatique et la connaissance produite par le datamining pour créer un véritable capital client

- Nécessité d'éléments quantitatif, de mesures précises d'activité et de valeur client, à laquelle répond la statistique
- Utiliser des ressources et technologies informatiques en constante réinvention pour exploiter au mieux cette connaissance

Compter sur une véritable expertise métier

- Détecter des tendances, des comportements, pressentis par l'expérience du marché
- Répondre à des demandes des clients, des phénomènes comportementaux
- Etablir des indicateurs de performance (valeur client, rétention, risques…)

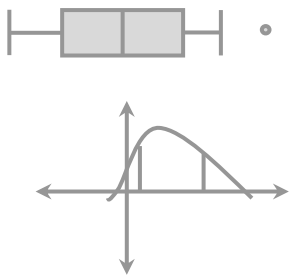
DISCIPLINES « UNIVERSELLES »



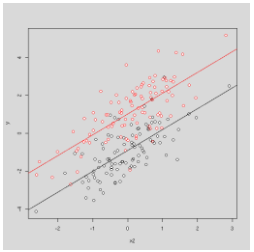
EXPLORE

Statistique descriptive et exploratoire

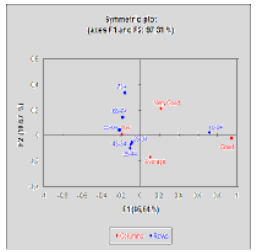
Établir les **caractères** principaux décrivant **une population**



Description



Tests,
Corrélations

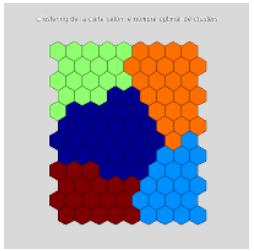


Analyses factorielles

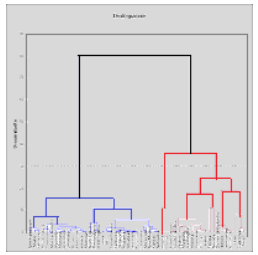
EXPLORE

Apprentissage statistique non supervisé

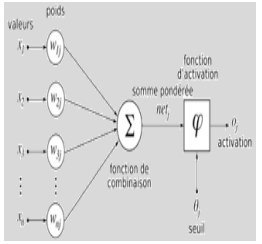
Établir les **structures** principales et sous-ensembles d'**une population**



K-Means,
Cartes de Kohonen



Classifications
ascendantes



réseaux neuronaux
non supervisés

Apprentissage statistique supervisé

Caractère
de **l'individu**

Analyse de variance
Régressions simples
Régressions multiples

Evènement
du **parcours** **client**

Régression Logistique
Analyse Discriminante
Arbres
Réseaux Bayésiens
SVM

Ex : **Prédire** l'achat, le risque de désabonnement d'un client
, le revenu potentiel attendu d'un **prospect** sur
la base de ses **caractères communs connus** avec les clients, les acheteurs, les churneurs.

Rassembler la **donnée**
Extraire **l'information**
Discriminer les **individus**
Prédire un **évènement** / une **valeur**

Un modèle de **datamining** industrialise **l'extraction** de la **connaissance** depuis la **donnée**

Des données en croissances

- La volumétrie des données explose
- Le net met à disposition des données de plus en plus diverses
  - En nature
  - En source
  - En stockage, encodage…
- Le paradoxe du digital, abondance et aridité

EXPLORE

Chaque minute :

500 000 tweets / 3 millions de posts facebook

3,7 millions de **recherches** google

4,2 millions de vidéos youtube vues

27k **avis** sur yelp

160 millions **d'emails** envoyés

Volume de données & croissance

…dont 75% de **spam**

Volume data **double** chaque année

25% sont **exploitables**

3% sont **exploitées**

Source : InternetLiveStats, WeAreSocial

EXPLORE

Diversité des données

Parcours **Client**

Réactivité **Email**

Navigation **Web**

Questionnaires de **Satisfaction**

Données **Sociodémographiques**

Données **Publiques**

Commentaires **Sociaux**

Images et **Vidéos**

LE PARADOXE DU MARKETING DIGITAL

EXPLORE

L'abondance est **partout**

Les clients sont **partout**

Les données sont **partout**

La connexion au net **partout disponible** à grande vitesse

Mais

Les vraies **ressources** sont **rares** et mal identifiées.

Quelles sont ces **ressources rares** ?

EXPLORE

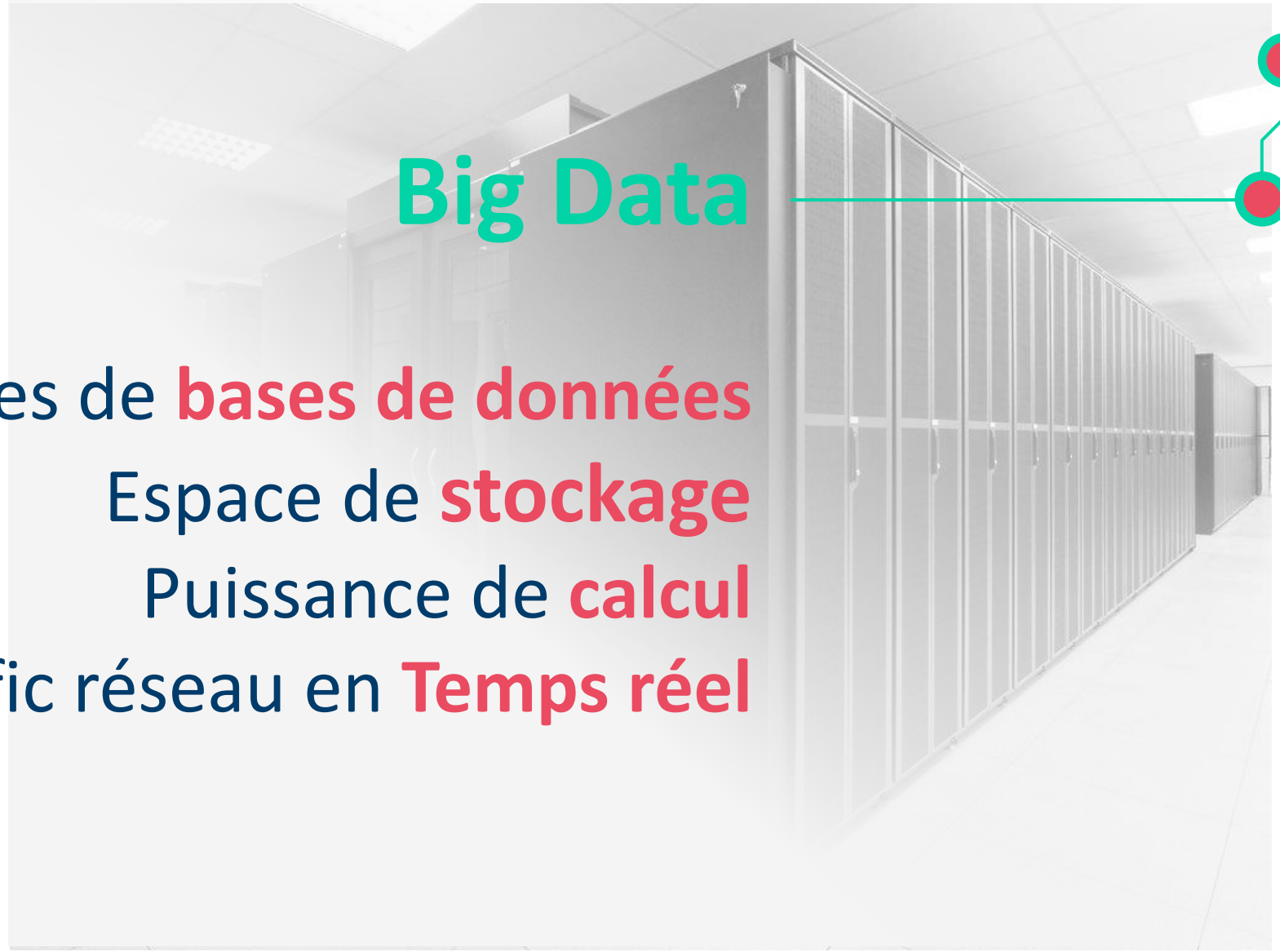
L'innovation technologique, le Big Data

> Conjugaison opportuniste d'innovations

> Hardware, Software et infrastructurelle

> Au service de nouveaux usages

EXPLORE



Big Data

Architectures de **bases de données**

Espace de **stockage**

Puissance de **calcul**

Trafic réseau en **Temps réel**

Bases de données NoSQL (MongoDB, Cassandra…)

- Clef/Valeur
- Document
- Colonne
- Graphe

Traitement massivement parallèle (Hadoop : HDFS + HBASE + MapReduce)

- Infrastructure de serveurs
- Traitements distribués sur quelques centaines ou milliers de serveurs

Exploitation de la mémoire, traitement en mémoire et non sur disque (Memtables)

- Accélère considérablement le temps de traitement des requêtes.

EVOLUTION DE L'ESPACE DE STOCKAGE ET DE LA PUISSANCE DE CALCUL

EXPLORE

Le disque dur reste l'élément le plus archaïque de l'informatique moderne

- HDD : Espace disque bon marché mais cycles de lecture/écriture restent lents
- Architectures RAID trop onéreuses pour utiliser des SSD NAND 3D sur des serveurs de calcul

Nouvelles technologies émergentes, ex : 3D Xpoint (Intel / Micron)

- 1000 fois plus rapide qu'un SSD
- Aussi robuste qu'un HDD
- Persistance (récupération d'incident exceptionnelle)
- Conception simple, potentiellement peu onéreuse (bémol du monopole)
- Spécifiquement pensée pour le big data

COUVERTURE WEB ET VITESSES DE TRANSFERTS

EXPLORE

D’un point de vue client

- Couverture ADSL presque totale du territoire Français, réduction des zones blanches
- Très haut débit terrestre et mobile en progression, près de 20% du territoire

=> Accès aux offres de e-commerce, au jeu, au Software as a Service, marchés en croissance énorme

D’un point de vue technologique

- Capacité de collecte, transmission, traitement, et retour de données en temps réel
- Développements d’API et de webservice de plus en plus efficaces en ressources et capables en complexité

Développement rapide et spécialisation du Cloud Computing, serveurs alloués à la volée pour une gestion fine de charge des services, capacités de calcul et de diversification d’activité en croissance.

L'Evolution des usages

- Nouvelles pratiques d'entreprise
  - E-marketing
  - Community Management
  - CRM
- Nouvelles pratiques individuelles
  - Méfiance vis-à-vis de l'email
  - Lassitude du display
  - Contact direct
- La relation client self-care
  - Vue unique du client
  - Personnalisation de la communication
  - Plateformes de « self-care »

Nouvelles pratiques d'entreprise

Développement du **CRM**

Relation en **Multicanal**

Connaissance Client

Traçabilité du **Parcours Client**

Collecte Navigation Web

Pages **Sociales**

Gouvernance de la donnée

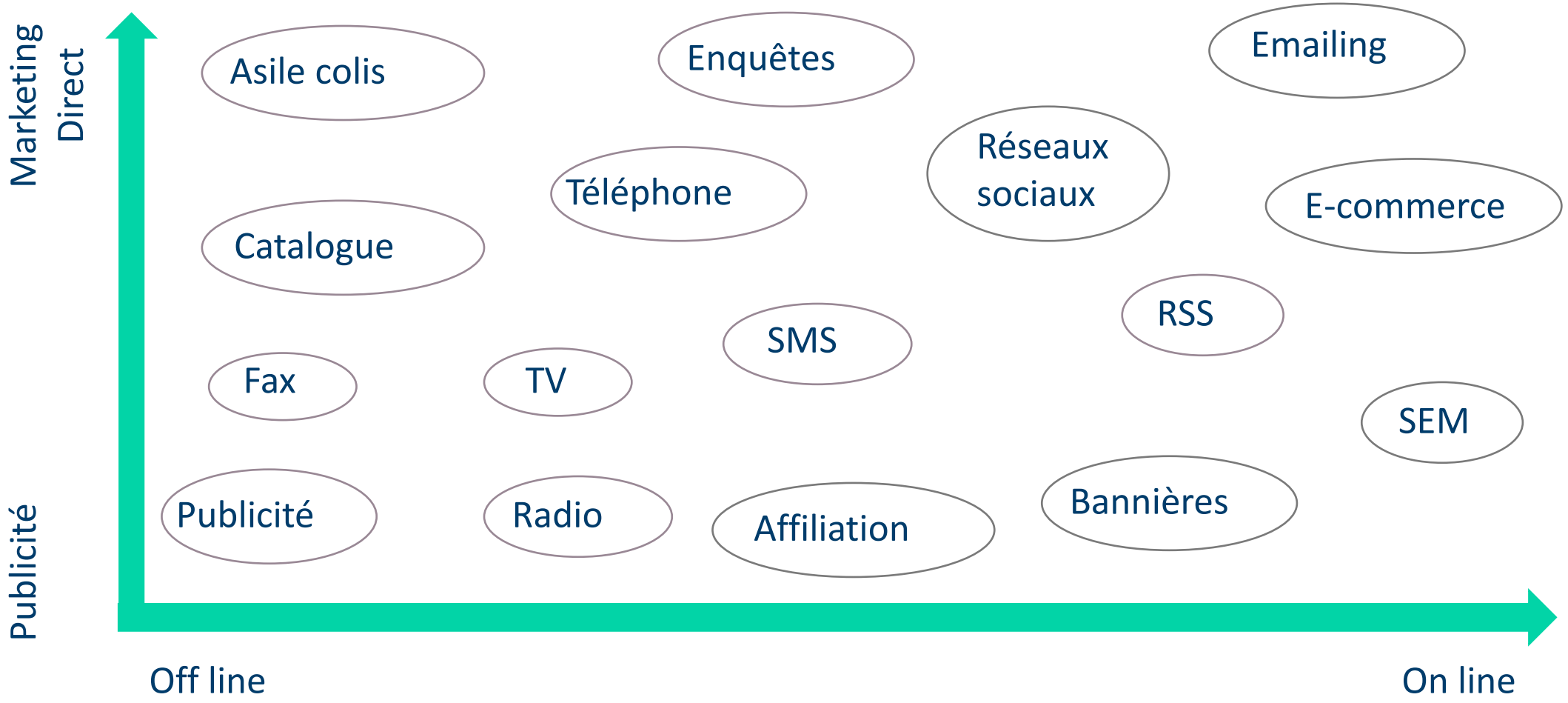
EXPLORE

Self Care
Réappropriation de la relation
Réseaux sociaux
Retours clients **publics** & **spontanés**

Relation
Client / Annonceur

Saturation Marketing
SPAM
Inquiétude vis-à-vis de la **vie privée**

MULTIPLICATION DES CANAUX MARKETING



Chaque source d'information a de la valeur

LES CANAUX MARKETING GÉNÈRENT DES DONNÉES …

Emailing

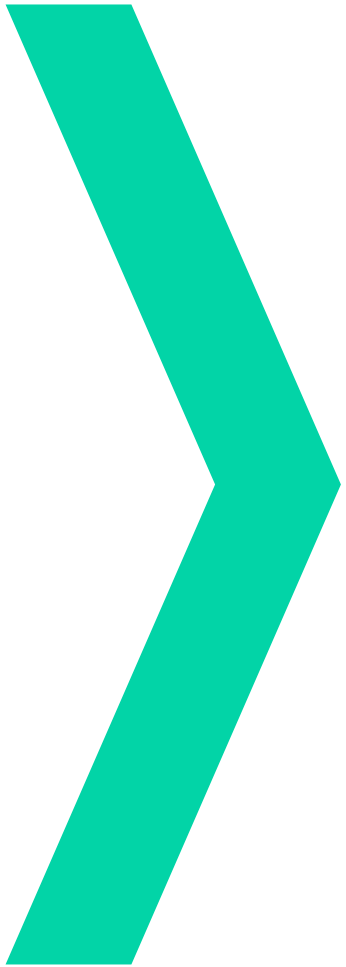
Réseaux sociaux

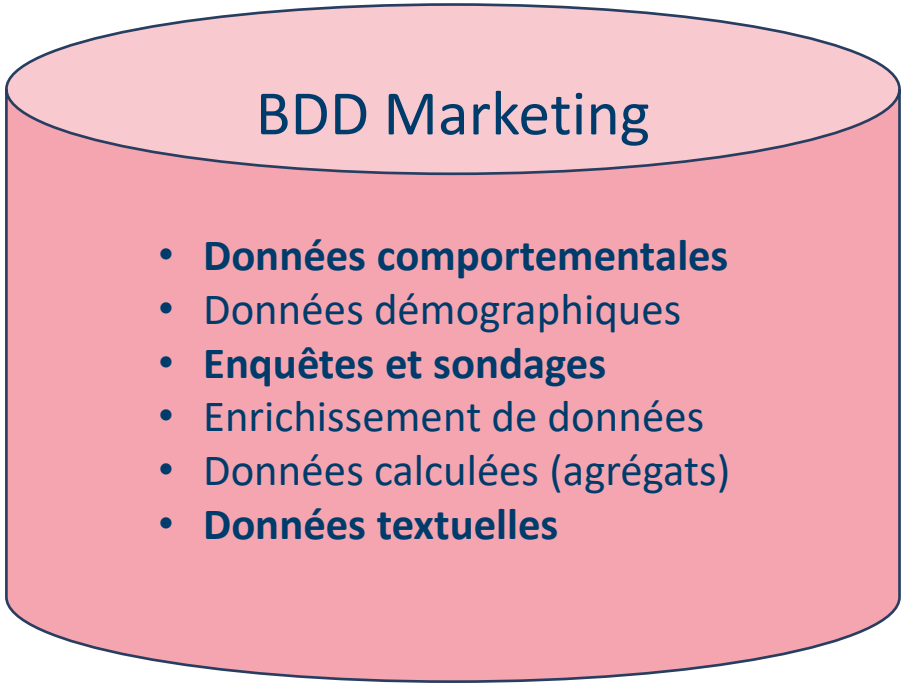
E-commerce

SEM

Call Centers

Téléphone





… et des signaux faibles qu'il faut traiter

Problématiques propres au **Marketing digital**

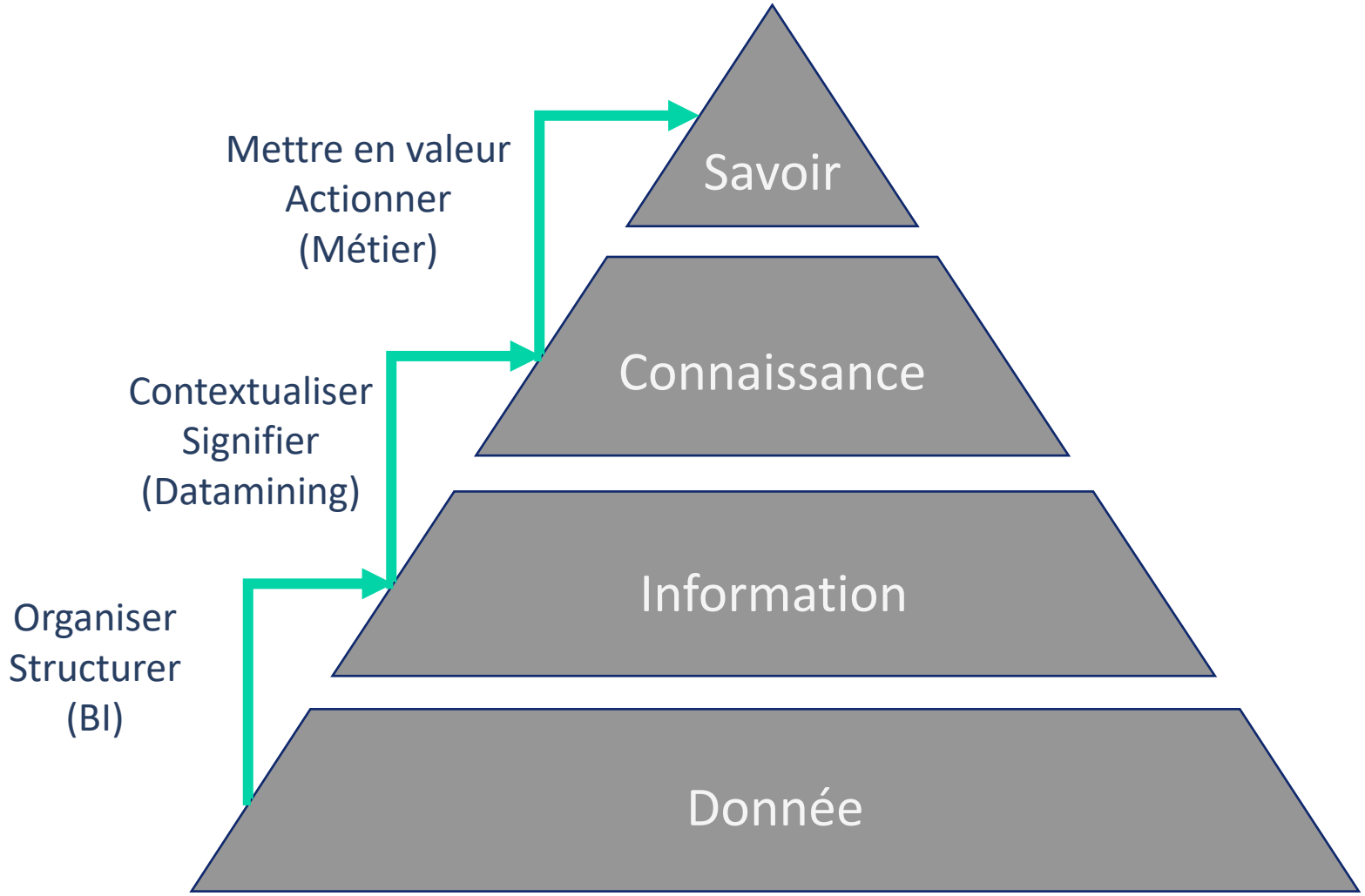
- Canal unique du **point de vue client**
  Faisceau de canaux pour l'annonceur (Email, site web, réseaux sociaux)
- **Désabonnement**, classement « **SPAM** »
- Relation **Bidirectionnelle**
- Exigence communication **personnalisée**

VERS UNE DÉFINITION PRÉCISE DU DATAMINING

De la donnée au savoir

- Démarches complémentaires,
  - de business intelligence,
  - de Datamining,
  - et de l'expertise métier (marketing par exemple)
- La donnée et sa valeur se consolident à chaque étape
- Des questions et problématiques précises orientent la pensée à chaque développement

DE LA DONNÉE AU SAVOIR



DE LA DONNÉE AU SAVOIR

EXPLORE

Données brutes  **information**  **Connaissance actionnable**

BI Datamining

1) Données brutes, multiples sources
2) Centralisation et transformation en information (traitement de la donnée), mise en place d'un CRM et/ou d'un projet de BI pour rendre l'information accessible
3) Analyse statistique (datamining) et transformation de la donnée en connaissance actionnable
4) Déclenchement des campagnes marketing sur la base de la connaissance acquise pour générer du profit

DE LA DONNÉE AU SAVOIR

EXPLORE

Fondement et pertinence

- Qu'invoquons nous pour justifier la pertinence des conclusions d'une étude datamining ?
- Des hypothèses mathématiques à la réalité du terrain
- Comment garantir les résultats dans un contexte réaliste ?

EXPLORE

Système de données contextualisées

Objectif préalablement établi

Adéquation au métier

Fondements & Pertinence

Données évaluées

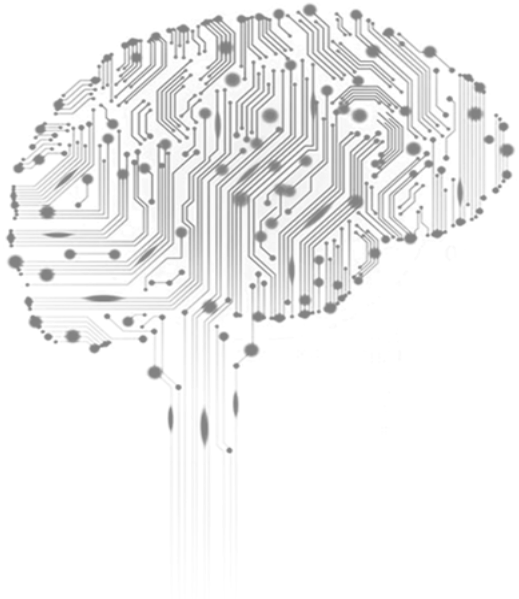
- En justesse
- En Certitude

« Garbage in - Garbage out »

EXPLORE

Datamining et Intelligence Artificielle

Machine **Learning**



Réseaux **Neuronaux**

Analyse prédictive issue de **l'informatique**

- Nécessite **moins de données et moins de temps** que la statistique
- « Boite noire » : **pas de modèle** explicatif au phénomène prédit

FONDEMENT ET PERTINENCE - BÉNÉFICES ATTENDUS

- Connaitre la validité et la qualité des données
- Connaitre l'adéquation d'une offre produit et d'une clientèle
- Comprendre les comportements et les attentes des clients
- Ne pas gaspiller l'effort marketing mais le concentrer sur les segments de clientèle les plus rentables
- Améliorer son ROI de communication et ses ventes
- Eviter la lassitude client, augmenter la valeur et la durée de vie des clients
- Cibler les clients à risques et les retenir
- …

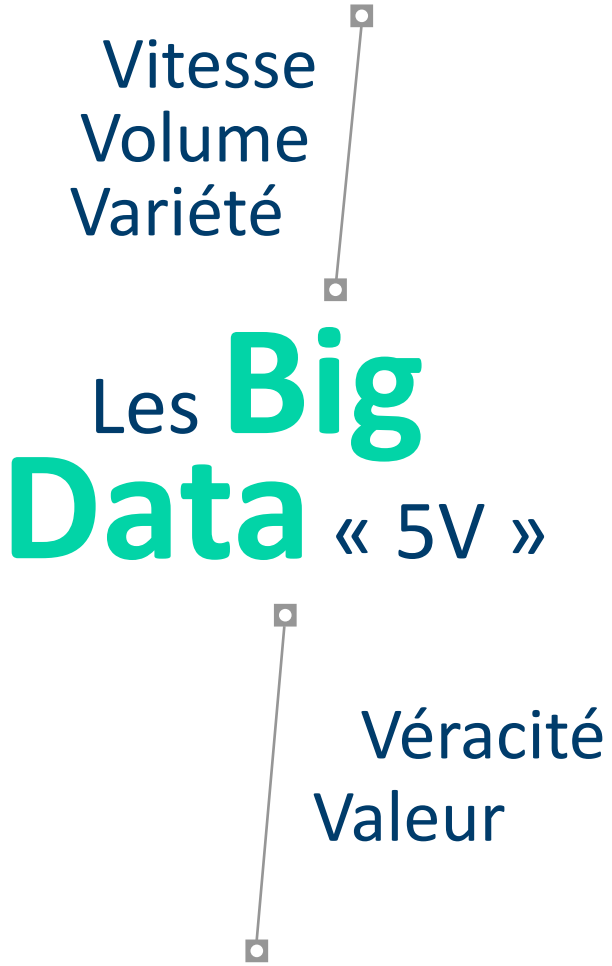
Définition 1

Le data-mining est une **démarche intellectuelle** d'appréhension des **règles** régissant une **population**, des **liens relationnels**, des **corrélations**, **dépendances** ou **causalités**, et donc d'**extraction de l'information** et de raffinage de la **connaissance** (prédictive par exemple), s'appuyant sur une grande quantité de **données**, et des méthodologies **statistiques**, **mathématiques**, algorithmes **informatiques**, et **reconnaissances de formes**

EXPLORE

Datamining et Big data

- Rappel : Définition actuelle du Big Data
- Comment le Datamining permet de réaliser ce qui reste souvent un fantasme ?
- Quelles sont les limites du datamining tel que nous l'entendons dans ce contexte ?


Vitesse
Volume
Variété
Les Big Data « 5V »
Véracité
Valeur

Une profonde **Adéquation**

Compréhension et **transformation** de la donnée

Sélection de variables et réduction de **dimensionnalité**

Intelligence des **modèles**, rapport au **métier**

Industrialisation aisée, **puissance** de calcul

Critères de **validité**, de **qualité**, **détection** d'erreurs

Certaines méthodes de datamining doivent **s'adapter**, ou **prennent de l'importance** pour le **Big Data**

- Exemples
  - **Text-Mining** (méthodes classiques de niche, énormément mises en avant par l'arrivée des réseaux sociaux, des modes de communications riches entre entreprises et individus)
  - **Hybridation de méthodes** (utilisation conjointe de diverses méthodes permettant de s'adapter aux problématiques big data par exemple Recommandation produit et filtrage collaboratif)
  - Réduction de dimensionnalité, usage abondant **d'Analyse Factorielle Multiple** pour écarter des variables en masse
  - **SVM** (Machines à vecteurs support) particulièrement adaptées à l'usage de Map/Reduce

A l'inverse les technologies spécifiques développées pour le big data sont indispensables à **l'analyse** en **temps réel** de **gros volumes de données**

- Le traitement des données a une complexité algorithmique très supérieure à o(n), couramment $o(n^2)$, doubler la puissance de calcul ne permet d'augmenter que de racine de 2 la capacité de traitement
  - ex dans le cas d'une SVM la seule vérification de l'existence d'une solution est déjà de complexité $o(n^2)$

LES FREINS AU BIG DATA

- Mettre ensemble un grand nombre de sources différentes
- Assurer la qualité des données malgré une grande variété de sources
- Convaincre son management et obtenir des ressources
- Interpréter et transmettre les résultats
- Trouver les compétences adéquates sur le marché
- Faire en sorte que les résultats soient disponibles au point d'interaction avec le client
- Etre capable de gérer le flux de données à la bonne vitesse
- Problèmes avec la protection juridique des données
- Ouverture d'esprit des parties prenantes aux résultats d'analyse

Source : Forrester Août 2012

EXPLORE

Quelques Exemples

- Prévention du risque d'incendie
- Recommandations multimédia
- Prévention du risque de rejet d'une greffe

QUELQUES EXEMPLES

Problématique : Prévention des incendies

Analyse exploratoire :

Typologie des quartiers, particularités des plus touchés : 60 **facteurs** identifiés

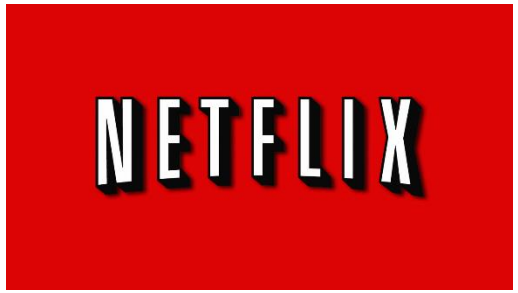
Conclusion :

Modèle **prédictif**

Programme de contrôle des immeubles à **risque**

Préconisations de politique d'urbanisme

QUELQUES EXEMPLES



Problématique : Recommander des programmes

Analyse exploratoire :

Profondeur du catalogue,
Faible nb d'avis/consommateur,
programmes de niches
Référencement sémantique **riche**

Conclusion :

Modèle mixte,
Conjugue deux types de similarités
- Intérêts communs des spectateurs
- Similarités Objectives (Thèmes, Genre, Acteurs)

Problématique : Dangerosité détection de risques de rejets de greffe par biopsie

Analyse exploratoire :

Collecte de sang systématique avant biopsie,
analyse des marqueurs sanguins en quête de **corrélations**

Conclusion :

Modèle **prédictif** basé sur un éventail de marqueurs sanguins, moitié moins de biopsies nécessaires.

Recherche sur ces marqueurs sanguins et leurs causes ?
Corrélation ou Causalité ?

Mieux **comprendre** les rejets.

LA DONNÉE

INTRODUCTION ET CONCEPTS - LA DONNÉE

1) Données structurées et non structurées
   A. Définition générale de la donnée
   B. Définition et exemple de données structurées
   C. Définition et exemple de données non structurées

2) Types de données et usages
   A. Données quantitatives
   B. Données Qualitatives
   C. Types de données et usages
      Métriques et distances

3) Outils et processus de DataManagement
   A. Les SGBD
   B. Les ETL
   C. La Collecte des données

4) Audit de données
   A. Donnée Aberrante
   B. Donnée Atypique
   C. Remplacement de données manquantes

Etymologie :

- Du latin « datum », ce qui est donné
- Apparait dans son acception actuelle dès 1500
- Usage moderne (via l'anglais data) date de 1940 ou 1950, avec l'émergence de l'informatique. L'une des façons de définir l'ordinateur est en effet une machine capable de « recevoir, traiter, et réémettre la donnée »

"*La donnée n'est pas l'information, l'information n'est pas la connaissance, la connaissance n'est pas la compréhension, la compréhension n'est pas la sagesse*"

Tim Berners-Lee

Les **informations** composant la donnée sont des **exemples** dotés d'attributs

- On dispose généralement d'un ensemble de N données, que l'on nomme **population**

Attributs

- Un **attribut** est un descripteur d'une **entité**, d'un **individu**. On l'appelle également **variable**, **champs** ou **caractéristiques**

Exemple

- Un **exemple** est une entité, une instance caractérisant un **objet** et est constitué **d'attributs**.
- synonymes : **point**, **vecteur** (souvent dans $\mathbb{R}^d$)

EXPLORE

Définition générale de la donnée

- La donnée comme matériau
- Définitions et exemples

DÉFINITION GÉNÉRALE DE LA DONNÉE

La donnée est notre instrument de travail, c'est une matière brute

La donnée porte l'information en puissance, la connaissance en est l'entéléchie

La donnée peut être :

- collectée spécifiquement en vue de répondre à une problématique
- stockée et générée automatiquement à l'usage d'un système et analysée ultérieurement

La donnée n'est plus limitée dans sa définition par les nécessités d'être informatisable, stockée par un SGBD et analysée par un système informatique

La donnée, organisée en séries se déploie sur autant de dimensions qu'il y a d'«objets» participants du système étudié

La donnée est l'ensemble des ressources disponibles et exploitables à l'analyse statistique ou informatique portant un lien de sens ou de nature avec l'objet de l'analyse, décliné ou déclinable sur une population d'étude, et pouvant contribuer à identifier les caractéristiques, les particularismes et états ou à mieux définir et comprendre cette population ou le caractère objet de l'analyse.

EXPLORE

Données structurées

- Définition
- Exemples

LES DONNÉES STRUCTURÉES - DÉFINITION

EXPLORE

- La donnée structurée est organisée en variables, en séries pour lesquelles on peut affecter une valeur à chaque individu

- Ces variables sont quantifiables, c'est-à-dire quantitatives (valeur numériques, discrète, i.e. entière ou bien continue) ou qualitatives (donnée non numérique recevant une valeur parmi un ensemble, dénombrable)

- Une variable qualitative peut être a priori sans ordre (ex : la marque d'une voiture, une région, …) ou hiérarchique (tranches d'âge, de revenu, CSP, …) on dira « ordinale », dans le cas contraire, elle est « nominale »

- Certains types de variables (binaires, dates, etc.) peuvent tout à la fois être considérées comme nominales ou continues

LES DONNÉES STRUCTURÉES - EXEMPLES

EXPLORE

- Données d'un logiciel de CRM
  - Comprendre et anticiper le comportement client
- Données commerciales
  - Soutenir la démarche marketing, up ou cross-sell
- Données de sondages, de questionnaire de satisfaction
  - Comprendre la structure d'une population, ses attentes, points forts, etc.
- Descriptifs de catalogue produit
  - Organiser un moteur de recherche, de recommandation, des gammes, des opérations de soldes

EXPLORE

Données non structurées

- Définition
- Exemples

LES DONNÉES NON STRUCTURÉES - DÉFINITION

EXPLORE

- Elément symbolique porteur de signification mais qui contrairement à la donnée structurée ne parle pas pour lui-même

- En effet avant toute chose, une donnée dont l'ensemble des valeurs possibles, n'est pas connu et incommensurable par avance !

- Inexploitable par des méthodes classiques en raison
  - De sa nature (Objet riche, réseau,)
  - De son format (image, texte long, logs)
  - D'un caractère non quantifiable (catalogue de produits, de films, de musiques)
  - Non organisé en variables

- Ne peut être soumis à un algorithme analytique, à un modèle statistique, non quantifiable en l'état

LES DONNÉES NON STRUCTURÉES - EXEMPLES

EXPLORE

- Images ou vidéos
  - Reconnaître des formes, des objets, suivre un mouvement
- Son
  - Isoler un discours dans un brouhaha, suivre une source sonore pour l'isoler du bruit, reconnaître une voix, une intonation
- Discours longs
  - Analyser le vocabulaire, en déduire une intention, l'adresser à la bonne personne
- Réseaux
  - Rapprocher des individus, des comportements de consommation, affiner des recommandations, répartir une ressource sur un réseau
- Logs, pages webs, favoris, verbatims « en vrac », métadonnées (e.g. coordonnées gps d'une photographie) …

EXPLOITER LES DONNÉES NON STRUCTURÉES

EXPLORE

- Comment transformer une image, un son, en donnée exploitable ?
- L'indexation : stocker, dater et rapprocher la donnée non structurée à l'individu
- Contextualiser et enrichir la donnée par les meta-données (contexte, descriptif qualitatif, titres, catégories, mots-clefs)
  - Démarche chronophage
  - Souvent laissée à l'utilisateur, source d'erreur
  - Le plus souvent incomplète

- Référencer l'information intrinsèque contenue dans la donnée : « image de visage » « mouvement » « discours thématique »

- Numériser. (Innombrables techniques apportées par l'informatique pour transformer une image, un texte long, des sons sur une bande, en multiples variables, exploitables conjointement aux métadonnées)
  - Nécessité de chercher la bonne méthode, spécifique à la nature et l'exploitation désirée de la donnée non structurée

EXPLORE

Données quantitatives

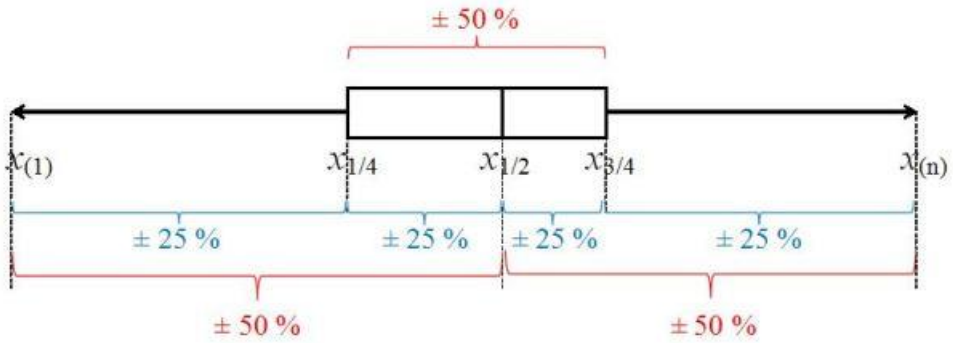
- Définition et exemples
- Représentation graphique
- Usage

DONNÉES QUANTITATIVES

EXPLORE

- Une donnée est dite quantitative si elle code une grandeur numérique
- Les variables quantitatives sont appelées continues lorsqu'elles prennent leurs valeurs dans $\mathbb{R}$ (ex : montants, durée, fréquence…), et discrètes lorsqu'elles prennent leurs valeurs dans $\mathbb{Z}$ (ex : âge, nombre de commandes…)
- Immédiatement exploitables par les méthodes statistiques les plus simples (corrélations, régressions, ACP, analyse discriminante)
- Peuvent suivre des lois de probabilités identifiables (algorithmes de mélange de lois…)
- Soumises à des contraintes fortes (Outliers, Corrélations multiples, etc.) et pourvue d'une **unité** (distance, durée, poids, montant…) ce qui la soumet à des phénomènes d'échelle
- Mais également aisément manipulables pour améliorer l'adéquation au modèle (centrer réduire, appliquer une fonction de transfert, etc.)

Boîte à moustache – Box Plot



Représente l'étendue d'une variable ainsi que ses tendances centrales (médiane et/ou moyenne)

La version de base présente les valeurs extrêmes (« moustaches »), ainsi que la boite de demi-étendue figurant l'écart inter-quartiles et séparée en son « centre » par la médiane. Les graduations ne correspondent pas aux valeurs théoriques mais bien aux valeurs prises par la série les plus proches.

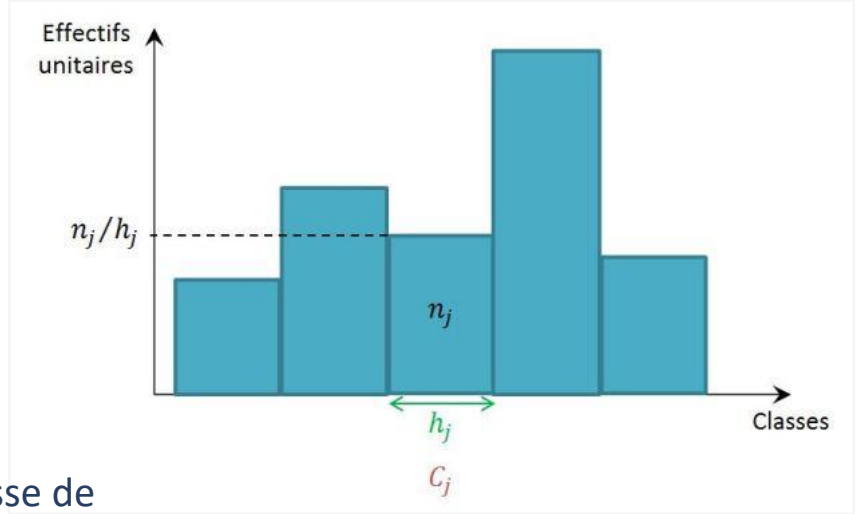
Dans les faits les valeurs extrêmes d'une série sont généralement atypiques, on modifie donc la structure du graphique afin de présenter les valeurs adjacentes des « pivots », souvent placés comme 1.5 fois l'écart interquartile (ou bien pour les séries les plus dispersées 2 fois l'écart interquartiles, les valeurs extérieures portent alors une très forte présomption d'aberration).

$$\begin{cases} P_d = x_{1/4} - 1.5\,(x_{3/4} - x_{1/4}) \\ P_g = x_{3/4} + 1.5\,(x_{3/4} - x_{1/4}) \end{cases}$$

Les valeurs adjacentes sont respectivement la plus petite valeur supérieure ou égale à Pd et la plus grande valeur inférieure ou égale à Pg

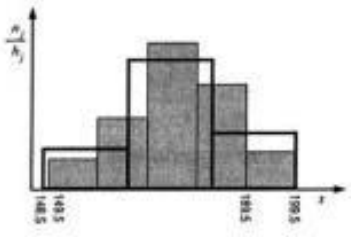
EXPLORE

Histogramme, Approximations de densité et de distribution



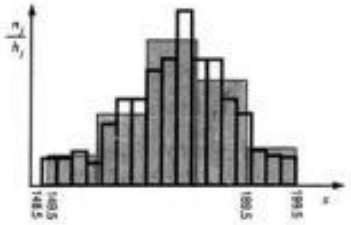
Pour une variable discrète, on peut utiliser un histogramme unitaire appelé également « diagramme en bâton » donnant une vision immédiate de la densité/distribution de population

La conception d'un histogramme est un exercice avancé de mise en classe de la variable, il est caractérisé par le choix du nombre de classes, leur largeur (égales ou non), et l'effectif unitaire ou fréquence de la classe



Groupement en 3 classes

Les classes peuvent être égales, ou s'élargir avec la raréfaction des données pour mieux représenter la distribution des données, et s'adapter aux particularismes des données (ex : montants de dons, variable « semi discrètes » - sur-représentation des valeurs rondes)



Groupement en 17 classes

Des « règles » existent pour déterminer un nombre de classes idéales

Ex : Règle de sturges, k classes telles que $k = 1 + \frac{10}{3}\log_{10} n$

Groupements en 5,3 classes et en 17 classes d'une distribution

Audit d'une donnée quantitative

Comprendre sa distribution

Homogène ou mélange de lois

Distribution approximable ou non par une loi de probabilité connue (normale…)

Mise en classe ou discrétisation évidente, bornes

Isoler et traiter les outliers (données aberrantes et/ou atypiques)

Valeurs extrêmes des distributions correspondant à des erreurs ou des individus trop particuliers pour participer à la définition d'un modèle

Non réponse codée en 0

Contextualiser les grandeurs, unités et plages de données rares

Certains modèles sont sensibles à des phénomènes d'échelle, deux longueurs, équivalentes, l'une codée en centimètres et l'autre en mètres, deux montants, en k€ et en € ne seraient pas traitées de la même façon même si leur étendue réelle est similaire.

Il est parfois nécessaire de « centrer-réduire » un certain nombre de variables quantitatives, ou de les mettre en classes et les traiter comme des variables qualitatives

Exploitation

Corrélation

L'indicateur le plus évident de pertinence d'une variable quantitative dans une analyse est évidemment la corrélation – évaluée selon sa pertinence et son estimation – entre variables.

Transformation

La corrélation de deux variables est souvent entendue au sens de colinéarité, de corrélation linéaire, c'est évidemment rarement le cas de données réelles. Mais elle constitue souvent une approximation tout à fait suffisante.

La représentation graphique en nuage de points des deux variables suggérera parfois une structure quadratique et logarithmique, et donc une fonction de transfert permettant de rapprocher la liaison d'une relation linéaire.

Multi-colinéarité

De nombreuses méthodes – les régressions en particulier – sont vulnérables aux multi-colinéarités, imposant une sélection drastique des variables quantitatives

Données qualitatives

- Définition et exemples
- Représentation graphique
- Usage

DONNÉES QUALITATIVES

Un caractère (une variable) donné est dit **qualitatif** s'il prend ses valeurs dans un ensemble E de **modalités** fini et labellisé

Ex : Couleur, marque de voiture, état matrimonial, sexe

Ses modalités doivent être **exhaustives** et **incompatibles** entre elles. Tout individu se voit donct attribué une et une seule modalité

(une question à choix multiple, par exemple, ne peut être traitée comme une seule variable)

Une variable qualitative est dite « ordinale » si ses modalités sont hiérarchisées (ex : tranches d'âge, de revenu…)

Une variable qualitative est « repérable » mais non « mesurable»

Caractérisation :

- nombre de modalités qu'elles admettent
- Distribution
  - mode (modalité majoritaire)
  - Effectifs de chaque modalité.

Un caractère quantitatif peut être considéré comme qualitatif s'il admet une valeur entière parmi un petit nombre

DONNÉES QUALITATIVES

EXPLORE

Une variable qualitative n'admet pas à proprement parler d'outliers, mais on considérera :

- Le regroupement de modalités rares (intuit métier, ou supervisé)
- Le choix entre ventilation des non-réponses, ou traitement de celle-ci comme une modalité

Des méthodes spécifiques remplacent la notion de corrélation afin d'étudier leurs liaisons :

- sur-représentations de modalités de l'une sur l'autre,
- variances intra et inter-classes

Les méthodes de datamining conçues pour recevoir des variables qualitatives ne nécessitent en général pas de sélection préalable des variables.

L'usage des variables qualitatives par des méthodes « fonctionnelles » (modèles) impose une transition (numérisation)

- binarisations (tableau disjonctif complet),
- tableaux de fréquences,
- distances/produits scalaires

Les méthodes descriptives ou exploratoires peuvent elle s'intéresser aux répartitions d'effectifs dans des ensembles et croisements de modalités, ou aux variances inter/intra classes de variables continues de référence

Tableau disjonctif complet

Soit un ensemble $C_1 \dots C_d$ un ensemble de caractères de dimensionnalité d

pour tout C_i de cet ensemble, on pose p_i la cardinalité de l'ensemble E_i de ses modalités, $E_i = \{M_{i,1} \dots M_{i,p_i}\}$.

Alors le tableau disjonctif complet de l'ensemble de la population $\Omega = (o_1, o_2, \dots, o_N)$ est la matrice comportant N lignes et $\sum_{i=1}^{d} p_i$ colonnes et telle que :

$$
\begin{array}{c}
 \\
o_1 \\
\vdots \\
o_k \\
\vdots \\
o_N
\end{array}
\begin{array}{c}
\begin{array}{cccccccccccc}
M_{1,1} & \dots & M_{1,p_1} & \dots & M_{i,1} & \dots & & M_{i,j} & & \dots & M_{1,p_i} & \dots & M_{d,1} & \dots & M_{d,p_d}
\end{array} \\
\begin{pmatrix}
1 & \dots & 0 & \dots & 0 & \dots & & 1 & & \dots & 0 & \dots & 0 & \dots & 1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & & \vdots & & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \dots & 1 & \dots & 1 & \dots & Si\ X_i(o_k) = M_{i,j} : 1, sinon\ 0 & & & \dots & 0 & \dots & 0 & \dots & 1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & & \vdots & & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \dots & 0 & \dots & 0 & \dots & & 0 & & \dots & 1 & \dots & 1 & \dots & 0
\end{pmatrix}
\end{array}
$$

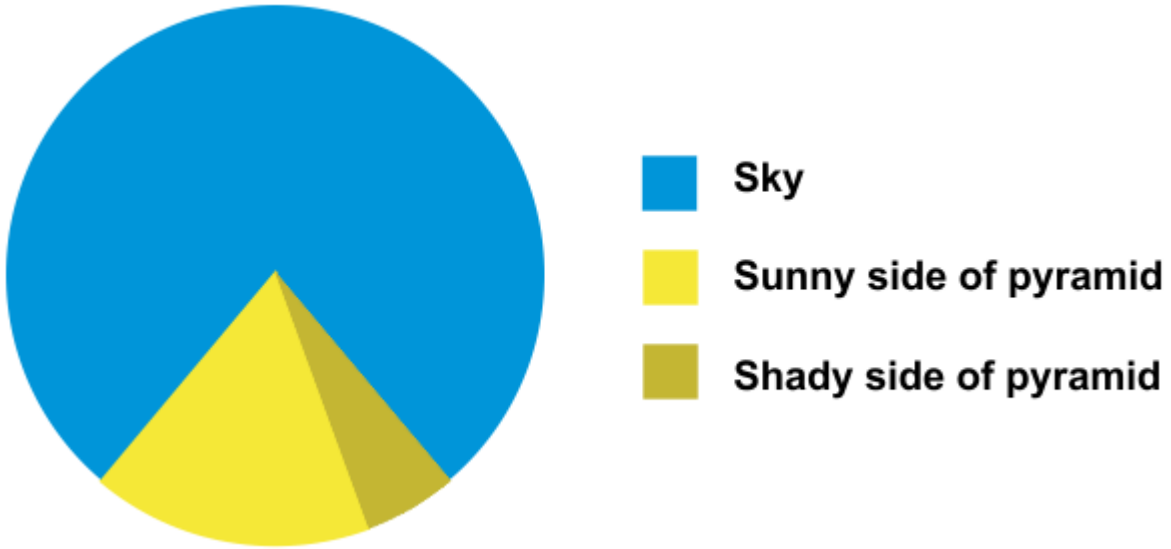
La somme d'une ligne du tableau disjonctif complet vaut d et la somme d'une colonne vaut N

Représentation graphique

Camemberts (Pie-Chart) et Histogrammes

- Le mode de représentation évident d'une variable qualitative est évidemment l'histogramme, ou le « camembert » permettant de visualiser côte-à-côte les effectifs de chaque classe.

- Le camembert à l'inconvénient de n'être lisible que pour un petit nombre de modalités (<5) et souvent détourné (représentation en 3D, ordre des classes)

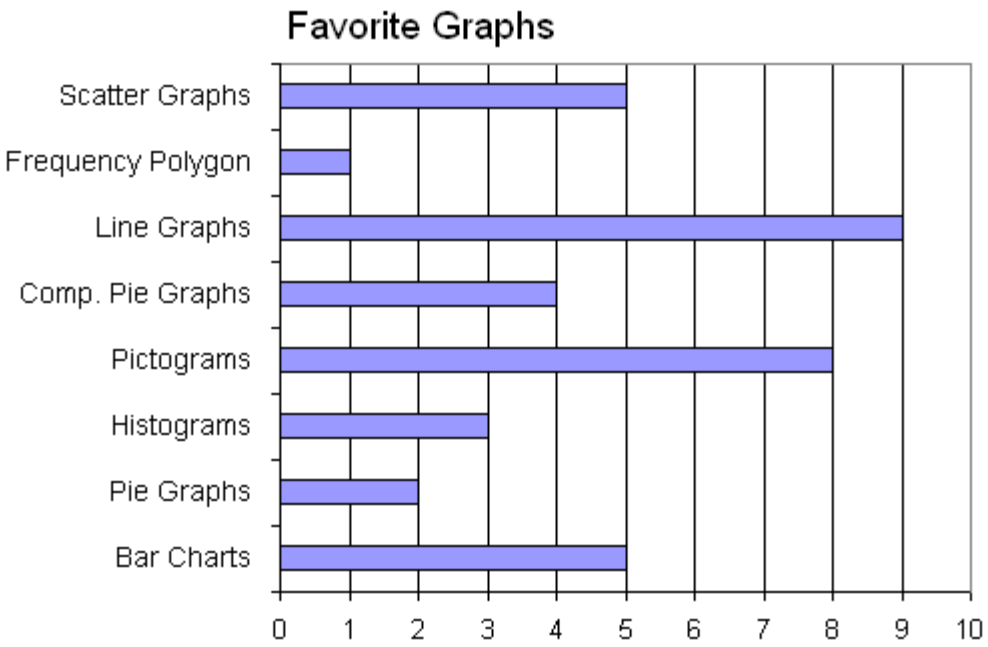


Représentation graphique

Camemberts (Pie-Chart) et Histogrammes

- L'Histogramme peut être enrichi plus aisément (superposition de moyennes d'une variable quantitative…) Il est plus adapté pour les modalités hiérarchisables (ordinales)

- Ces représentations sont impératives pour visualiser la prépondérance de certaines classes
  - Si une modalité regroupe une part anormalement élevée de l'effectif, ou si certaines modalités sont plus rares et doivent impérativement être regroupées.
  - Le regroupement de classe de modalités (recodage) demande les mêmes précautions (information perdue, sens des regroupements) que la mise en classe d'une variable continue



Cas particuliers

Variables Ordinales

Les variables ordinales sont très comparables aux variables quantitatives discrètes

Une variables quantitative mise en classe est toujours ordinale

Si les modalités d'une variable ordinales sont assez régulières (ex : satisfaction) et/ou trop nombreuses elles peuvent être notées et considérées comme continues

Des méthodes, (non supervisées en particulier) exploitent la particularité hiérarchique des variables ordinales pour rapprocher les individus de classes différentes

Variables Binaires

Peuvent également être considérées comme continues

Les plus aisées à prédire, les classifieurs linéaires, ou encore la régression logistique, sont conçues pour rapporter des grandeurs diverses à des variables binaires considérées comme numériques

Données et métriques

- Exemples d’autres types de données
- Notion de distance
- Notion de poids

Données composites

Certaines variables se comprennent nécessairement comme un tout, telles que les fréquences, ou les données transactionnelles

Agréger les données selon les modalités d'une variable qualitative cible afin d'obtenir, sur un ensemble de modalités d'autres variables quantitatives, les fréquences conjointes par sommation du tableau disjonctif complet donne lieux à un ensemble de méthodes spécifiques.
E.g. AFC : analyse factorielle des correspondances

Des données transactionnelles (ex : identifiant de client, identifiant de panier, identifiant de produit, quantité ou montant, ordres ou rang des transactions, etc.) bien que les identifiants, étant des métadonnées, ne soient pas porteurs de sens en eux-mêmes, représenteront une structure très aisément exploitables pour les algorithmes de datamining (ex : règles d'association produits, moteurs de recherche, analyse de navigation web…)

Distribution statistique

Soit une **population** d'observations $\Omega = (o_1, o_2, \dots, o_N)$ d'effectifs N individus

Soit un **caractère** C dont l'ensemble des p **modalités** possible est E $card(E) = p$

Considérons $A_i \subset \Omega, i \in [1, p]$, l'ensemble des individus possédant la modalité $M_i \in E$, n_i **l'effectif** de A_i

Par définition du découpage en modalités les A_i forment une **partition** de Ω, leur intersection 2 à 2 est nulle et leur union globale Ω

Une **variable statistique** est une application X: $\Omega \rightarrow E$ qui à chaque individus $o \in \Omega$ associe une modalité $M_i \in E$, on l'identifie à l'ensemble des triplets (A_i, M_i, n_i) Bien qu'on se préoccupe peu de l'ensemble A_i passé la phase d'audit. La **distribution statistique** de la variable X est $\{(M_i, n_i), i \in [1, p]\}$ (dans le cas où X ne réalise qu'une modalité de E, on parle de distribution statistique constante)

La notion d'effectifs d'une modalité étant absolue, ne permettant de comparaison on introduit celle de fréquence de la modalité $f(A_i) = \frac{n_i}{N}$

Distance

L'analyse des données repose TOUJOURS sur une notion de **distance** entre individus ou entre variables, une métrique commune entre des critères hétérogènes, l'originalité d'un algorithme de datamining tient souvent à la définition atypique de cette distance.

La distance entre deux individus est notre mesure de similarité entre deux individus x et x' ou sous ensemble de Ω, à l'aune des caractères $(X_1, \dots, X_d)$ connus (d est la dimensionnalité de l'espace d'analyse).

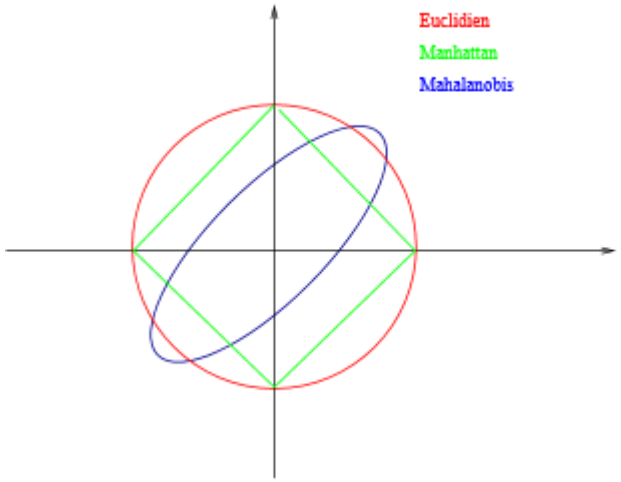
On redéfinit donc l'individu $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ comme un vecteur des modalités qu'il exprime des variables d'analyse.

Cet espace, dont chaque variable d'analyse est une dimension, est appelé espace des individus.

A l'inverse, l'espace des variables, de dimension N l'effectif de la population, est doté d'une norme permettant d'estimer la distance entre deux variables,

Distance

- La distance euclidienne « classique » est évidemment $\mathrm{d}(\mathrm{x},\mathrm{x}') = \sqrt{\sum_{i\in[1..d]}(x'_i - x_i)^2}$
  - C'est la distance habituellement retenue en classification hiérarchique et en typologie. L'inconvénient est le poids important des points se trouvant à une grande distance de l'origine de la mesure. Ceci ne les isole pas davantage, au contraire.
  - en statistique on raisonne souvent en carré des distances pour en faciliter la décomposition.
- Autres exemples de distances pour des variables continues, distance de manhattan, de mahalanobis
  - Respectivement $\sum_{i=1}^{d}|(x_i - x'_i)|$ et $\sqrt{(x-x')^t\sum^{-1}(x-x')}$
- Il existe bien des manières de mesurer des distances entre individus, c'est encore plus vrai lorsqu'on introduit des caractères qualitatifs (cas de l'ACM ou d'une régression avec Analyse de Variance – ANOVA – par exemple)
- Si on considère des variables quantitatives et qualitatives conjointement la pratique la plus courante consiste à binariser chaque modalité (construire le tableau disjonctif complet) et considérer ces nouvelles variables binaires comme continues.



Pondération

- La définition de distance doit aussi intégrer celle de **pondération**.

- Si les performances informatiques actuelles permettent d'évaluer un modèle sur la globalité des données à disposition de l'analyste, on peut désirer effectuer un **sondage** parmi les informations disponibles, ou simplifier les calculs en éliminant la **redondance** (ou quasi-redondance) parmi les données et la remplacer par un indicateur de pondération.

- On introduit alors la notion **d'inertie** à partir d'une distance d
  $I = \sum_{i=1}^{n} p_i d_i^2$

Corrélation

Les méthodes de datamining estiment souvent également une distance entre variables (liaison linéaire de régression, covariance, angle entre vecteurs…), le plus souvent assimilée à une **corrélation entre les phénomènes exprimés**

Soient deux variables X et Y, on note $x_i \; et \; y_i$ les valeurs prises par X et Y pour $o_i \in \Omega$

- On estime la moyenne de X et Y respectivement par $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ et $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$
- On estime les variances de X et Y respectivement par $\widehat{\sigma_x} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ et $\widehat{\sigma_y} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$
- Et la covariance de X et Y par $\widehat{\sigma_{xy}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})$
- Le coefficient de corrélation linéaire de X et Y est alors estimé par $\widehat{r_p} = \frac{\widehat{\sigma_{xy}}}{\widehat{\sigma_x} \cdot \widehat{\sigma_y}}$

Corrélation

Incidemment, avec la définition vectorielle des variables $X = (x_1, \dots, x_N)$ et $\mathrm{Y} = (y_1, \dots, y_N)$ l'angle formé par X et Y a pour cosinus selon la distance euclidienne :

$$\cos \alpha = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Cette définition alternative de la corrélation est non seulement un outil excellent pour la compréhension de l'ACP et des méthodes de sélection de variable en régression, mais confirme également pourquoi le coefficient de corrélation est toujours compris entre -1 et 1

La corrélation linéaire est cependant incapable d'exprimer des liaisons réelles mais non linéaires, pour lesquelles il est indispensable d'utiliser une transformation de variable (fonction de transfert)

Liaison de variables nominales

Le tableau disjonctif complet ne permet que de mesurer la liaison des modalités des variables nominales

On souhaite mesurer la liaison entre deux variables X et Y nominales admettant les modalités $(M_1 \dots M_p)\ et\ (N_1 \dots N_q)$ via les effectifs (ou fréquences) conjointes des modalités, n étant l'effectif total

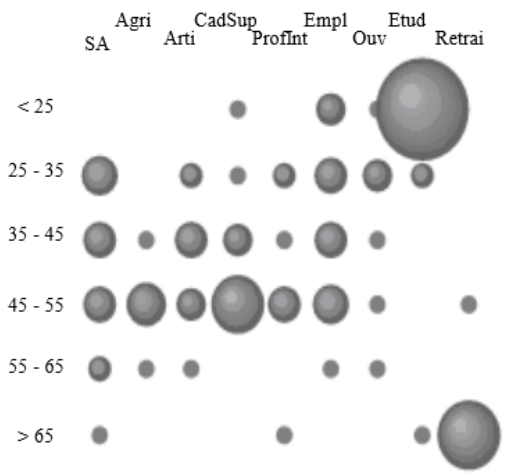
On dresse le tableau dit de « contingence »

Modalités Effectifs	M_1		M_p
N_1	n_{N_1,M_1}		n_{N_1,M_p}
		n_{N_i,M_j}	
N_q	n_{N_q,M_1}		n_{N_q,M_p}

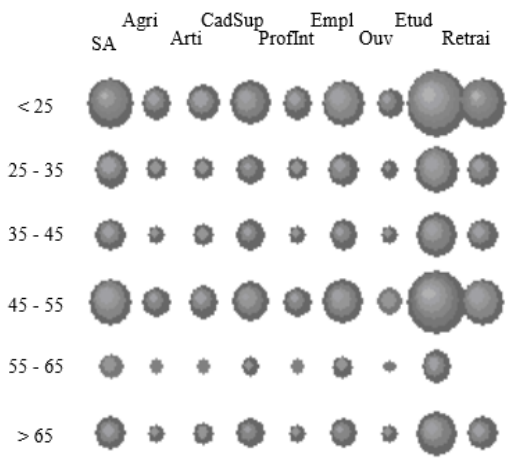
Liaison de variables nominales

La distance du χ^2 (Chi-deux) est construite en supposant l'indépendance des deux variables

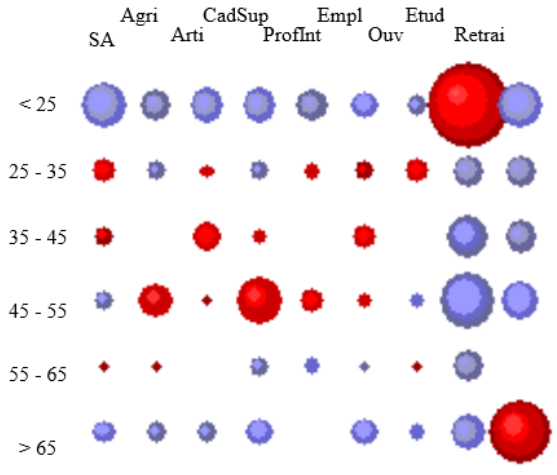
i.e. que pour tout (i,j) : $\frac{n_{N_i,M_j}}{n_{M_j}} = \frac{n_{N_i}}{n}$ d'où l'effectif théorique $n'_{N_i,M_j} = \frac{n_{M_j} * n_{N_i}}{n}$



Cartogramme des effectifs observés



Cartogramme des effectifs théoriques



Cartogramme des différences (bleu, négatives)

Le taux d'écart à cette estimation nous donne la distance (nulle si indépendance)

$$0 \leq \mathrm{d}^2 = \sum_{i,j} \frac{(n_{N_i,M_j} - \frac{n_{M_j} * n_{N_i}}{n})^2}{\frac{n_{M_j} * n_{N_i}}{n}} = n\left(\sum_{i,j} \frac{n_{N_i,M_j}^2}{n_{M_j} * n_{N_i}} - 1\right) \leq n(\min(p,q) - 1)$$

Liaison de variables nominales

Le χ^2 ne mesure pas l'intensité de la liaison, pour cela d'autres statistiques ont été imaginées en voici deux exemples

Le V de Cramer $v = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n*[\min(l,c)-1]}}$: est la racine du χ^2 rapporté à son maximum, il a l'avantage de prendre ses valeurs entre -1 et 1

Le Coefficient de Contingence $cc = \sqrt{\frac{\chi^2}{\chi^2+n}}$: analogue au V de Cramer dans sa construction et son interprétation, mais toujours inférieur à lui et par construction conçu pour évaluer des tableaux d'effectifs identiques

Le Φ de Pearson $\frac{\chi^2}{n}$: métrique sur laquelle est basée l'AFC

Liaison d'une continue et d'une nominale

Deux grandeurs sont définies pour mesurer la liaison entre une variable continue X et une nominale Y (dont les modalité sont (A_1 … A_p)

La variance inter-classe, définie pour chaque modalité j, n_j étant l'effectif pour cette modalité, est la variance au sein de cette modalité

$$\sum_{i,Y_i=A_j} \frac{(x_i-\overline{x_j})^2}{n_j}$$

La variance intra-classe est la variance résiduelle $\frac{1}{p}\sum_j(\overline{x_j}-\bar{x})^2$

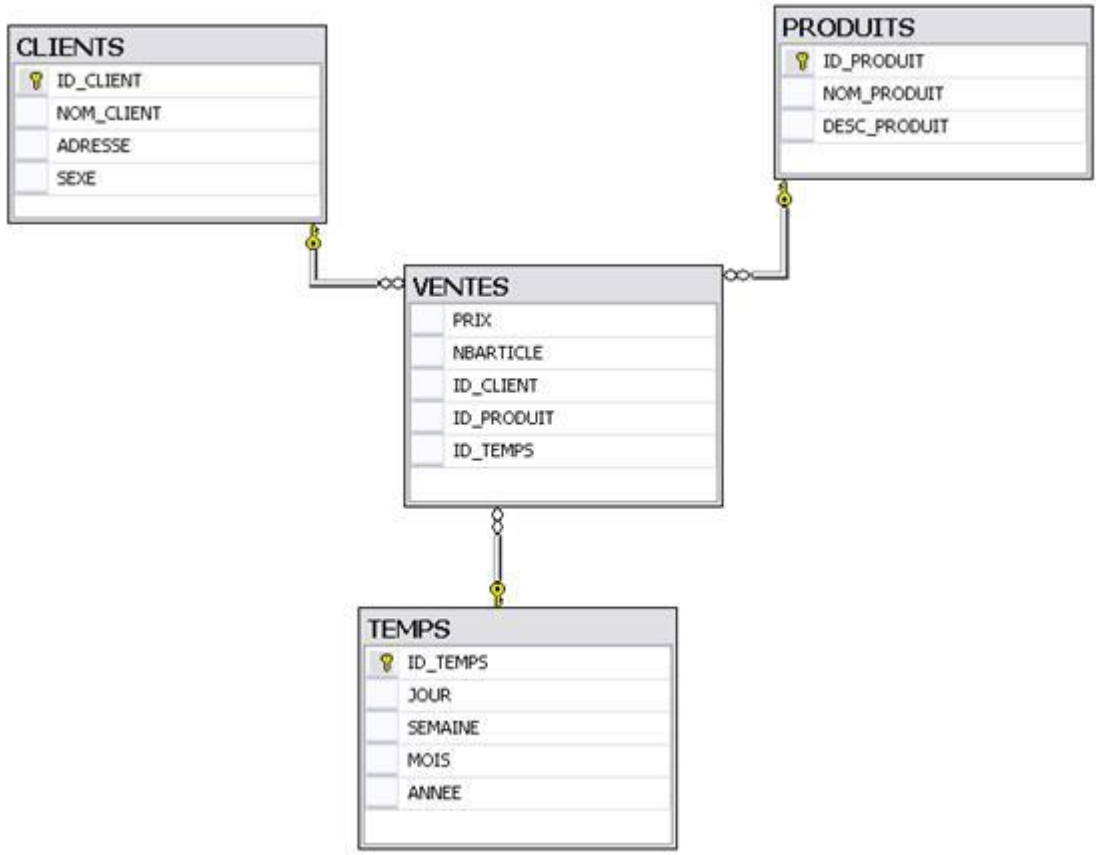
La minimisation des variances inter-classes et la maximisation de la variance intra-classe traduit la liaison de X et Y

EXPLORE

LES SGBD

- L'organisation du SI analytique
- Bases de données relationnelles (SQL)
- Bases de données NoSQL (e.g. MongoDB)
- Architectures BigData

BASES DE DONNÉES RELATIONNELLES



Données **structurées**, sur différents niveaux d'agrégation, interrogeable grâce à une syntaxe (**SQL**) standardisée

Structure classique des ressources CRM, structure et logiques apparus en 1980

Agrégats minimaux, concept de « **normalisation** » limitant la **redondance** d'information pour maximiser l'efficacité des accès aux données et faciliter le contrôle de la cohérence

Les données doivent être mises à plat par des **jointures** et des sélections fines facilités par la logique du langage

La **structure relationnelle** rend aisé la reproduction d'un modèle objet, métier ou logique

Les données sont **contraintes** par des ensembles de **clefs** (**primaires** au sein d'une table, **étrangères** pour les jointures…) et règles

SGBD tabulaires

Modèle de données : table, orienté ligne/colonne, opérateurs ensemblistes.

Architecture : partitionnement et réplication de tables, stockage en fichiers.

Exemples : Google Bigtable sur GFS, Hadoop Hbase sur HDSF, Apache Accumulo.

SGBD orientés-documents

Modèles XML et JSON, langages de requêtes: XQuery, SQL/XML, JSONiq, Json Query…

SGBD XML : BerkeleyDB XML, DB2 pureXML, EMC X-Hive/DB, Apache Xindice, etc.

SGBD JSON : MongoDB, CouchBase, LinkedIn Espresso, etc.

L'usage de XML/JSON s'étend aux SGBD relationnels : DB2, Oracle, SQLServer, MySQL, PostgreSQL. etc

SGBD orientés-graphes (relations objets, arbres et réseaux)

Modèle de données : graphe, RDF, opérateurs de parcours de graphes, langages de requêtes.

Architecture : partitionnement et réplication de graphes, stockage en fichiers, index.

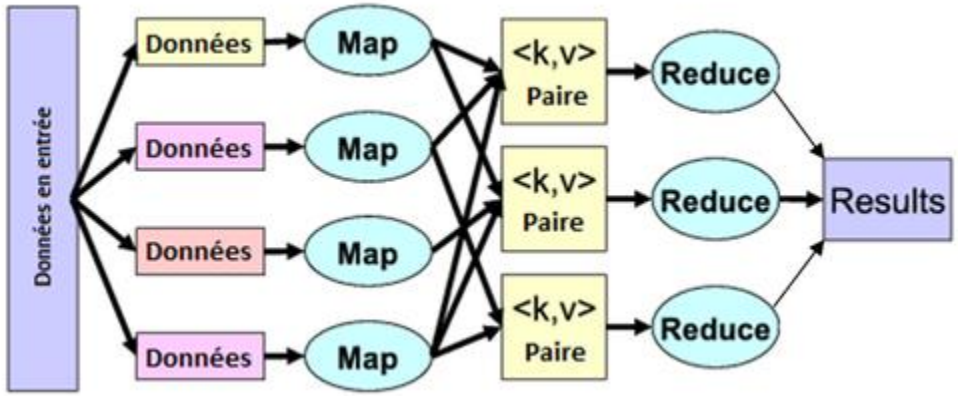
Exemples : Neo4J, DEX/Sparksee, AllegroGraph, InfiniteGraph, IBM DB2 Sparql.

Intégration SQL/NoSQL

Le relâchement de la cohérence : problèmes pour les développeurs et les utilisateurs.

NoSQL versus relationnel. L'intégration SQL/NoSQL avec Google F1 ou Apache Drill (Syntaxe SQL adaptée au NoSQL).

ARCHITECTURES BIG DATA



Les architectures Big Data modernes se basent surtout sur des bases tabulaires ou documentaires (Ex : HBase / MongoDB)

L'archétype de l'architecture de données orienté Big Data est le framework Hadoop, qui ne se limite pas à des structures de données mais un environnement de développement Hard/Software

- Système de fichier HDFS (Hadoop Distributed File System) conçu pour les systèmes distribués
  - Distinction des métadonnées, des espaces de nommage et des données pour un accès instantané à des structures de données fractionnées et éclatées sur plusieurs serveurs (cluster HDFS)
  - Rend superflu l'usage du RAID pour la récupération d'incident par des réplications croisées de données
- Map/Reduce (Google) ensemble d'algorithmes d'agrégats et calcul parallélisés pour les Gros Volumes et le Temps Réel

  permet des manipuler des fractions de la donnée en batch et les réinsérer dans le processus afin de traiter avec des usages mémoires alternés les gros volumes de données
- Hbase ou Cassandra (Apache) : Bases de données tabulaires orientées colonnes
- Mais aussi ZooKeeper : gestion de configuration, Drill(Apache), Hive (Facebook) ou Pig (Yahoo!) pour l'analyse de données

Implémentations grand public : Cloudera / MapR

Décriées pour des manques de consistance, ou le non respect des normes POSIX rendant difficile la certification des données

LES ETL

> Définition et périmètre de l'ETL

> Différence entre ETL et ELT

Extract, Transform, Load

Les processus (ou logiciels) d'ETL regroupent les outils et solutions permettant de regrouper des données de sources diverses (bases de données de tous types, fichiers, flux) de les transformer et confronter, afin de les charger dans un unique dépôt de données

Les opérations classiques prises en charges par l'ETL regroupent

- Le dédoublonnage (rattacher une information à un objet existant –par exemple identifier un client par ses nom et adresse – et le confronter, le mettre à jour, l'enrichir, ou créer un nouvel objet)
- Le redressement (mise sous un format unique de dates, d'adresses, etc. disponibles sous divers formats)
- La détection et la correction d'anomalies
- Les aggrégats de base (casser une trop grande normalisation des données pour préparer un datamart analytique)
- La gestion des flux temps réel

ETL OU ELT

Selon les solutions technologiques de bases de données utilisées, la puissance disponible et la complexité des données et traitements on choisis, pour chaque projet une stratégie Extract, Transform, Load ou Extract, Load, Transform

La première solution (la plus classique) consiste à confier à notre solution (talend par exemple) toute les transformations, historisations, confrontations de données et ne charger dans la base de données que les données dans leur état terminal.

- C'est la solution la plus légère, idéale pour initier un projet, elle minimise les flux de transferts de données, permet l'exploitation en mémoire

La seconde solution est plus adaptée en cas de limitations de puissance ou de grandes volumétries ou de mise à jour différentielle ou temps réelle des données, les données sont alors centralisées, nettoyées a minima et chargées en base. On exploite ensuite la puissance des processus natifs de notre solution de base de données (ex : Map/Reduce) pour les aggrégats et traitements.

La collecte des données

- Sources de données, le SI analytique
- Hiérarchiser, qualifier, valoriser la donnée
- Choix de données applicatives
- Concevoir un questionnaire d’enquête

LIEN ENTRE STOCKAGE ET ANALYSE

EXPLORE

Deux principes de base :

Ce qui se dégage de notre observation du terrain est que moins les données sont structurées, plus elles vont nécessiter un traitement important afin de les transformer en connaissance actionnable via un processus de reformatage, qui en outre, se doit d'être intelligent.

Il faut noter par ailleurs, que la quantité de données disponibles ne donne pas de réelle indication de la complexité du traitement nécessaire ; à l'inverse, leur richesse, leur degré de fiabilité et leur structure (ou l'absence de structure) vont être des facteurs beaucoup plus importants à prendre en compte.

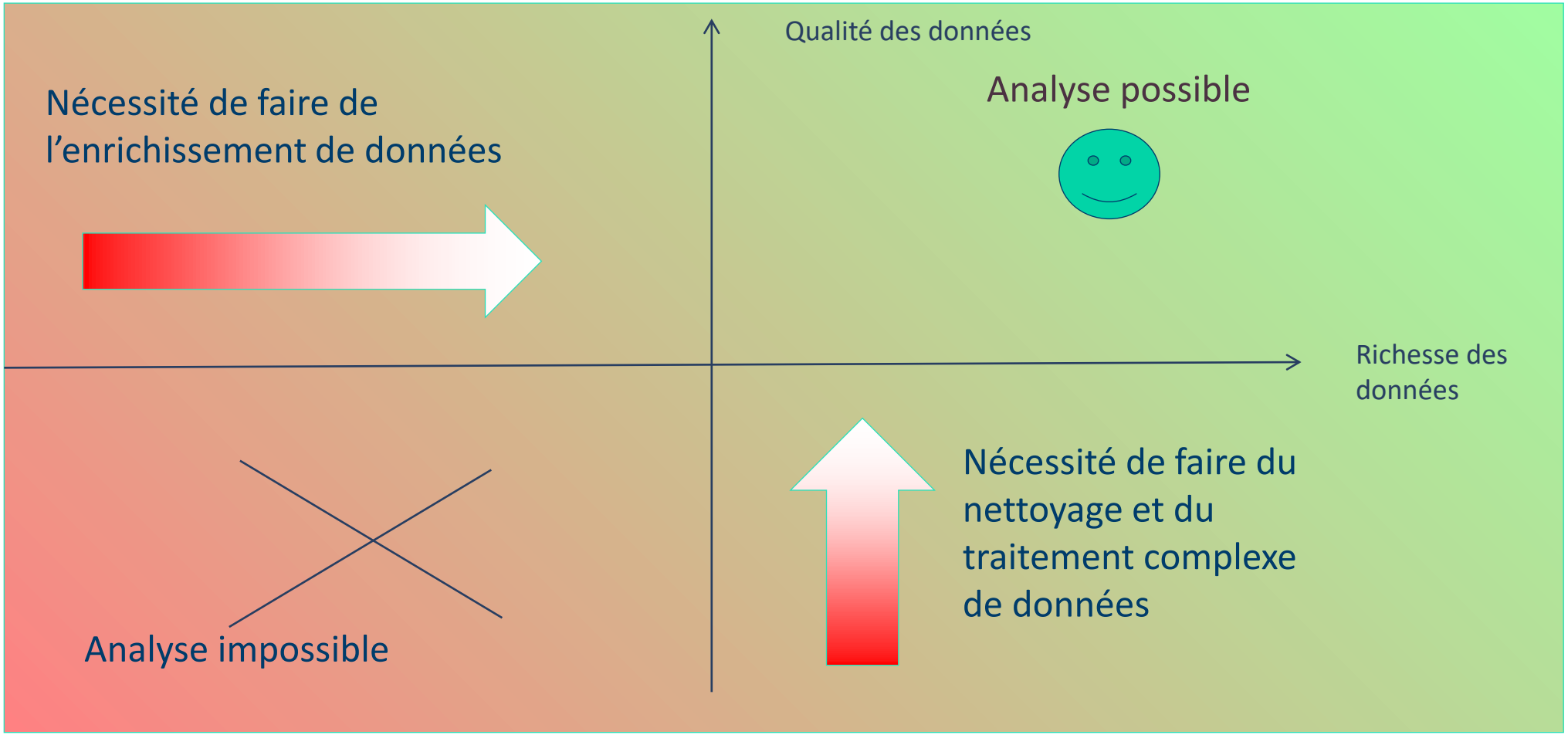
Une contrainte importante :

Tous les formats de stockage ne sont pas compatibles avec tous les types d'analyse

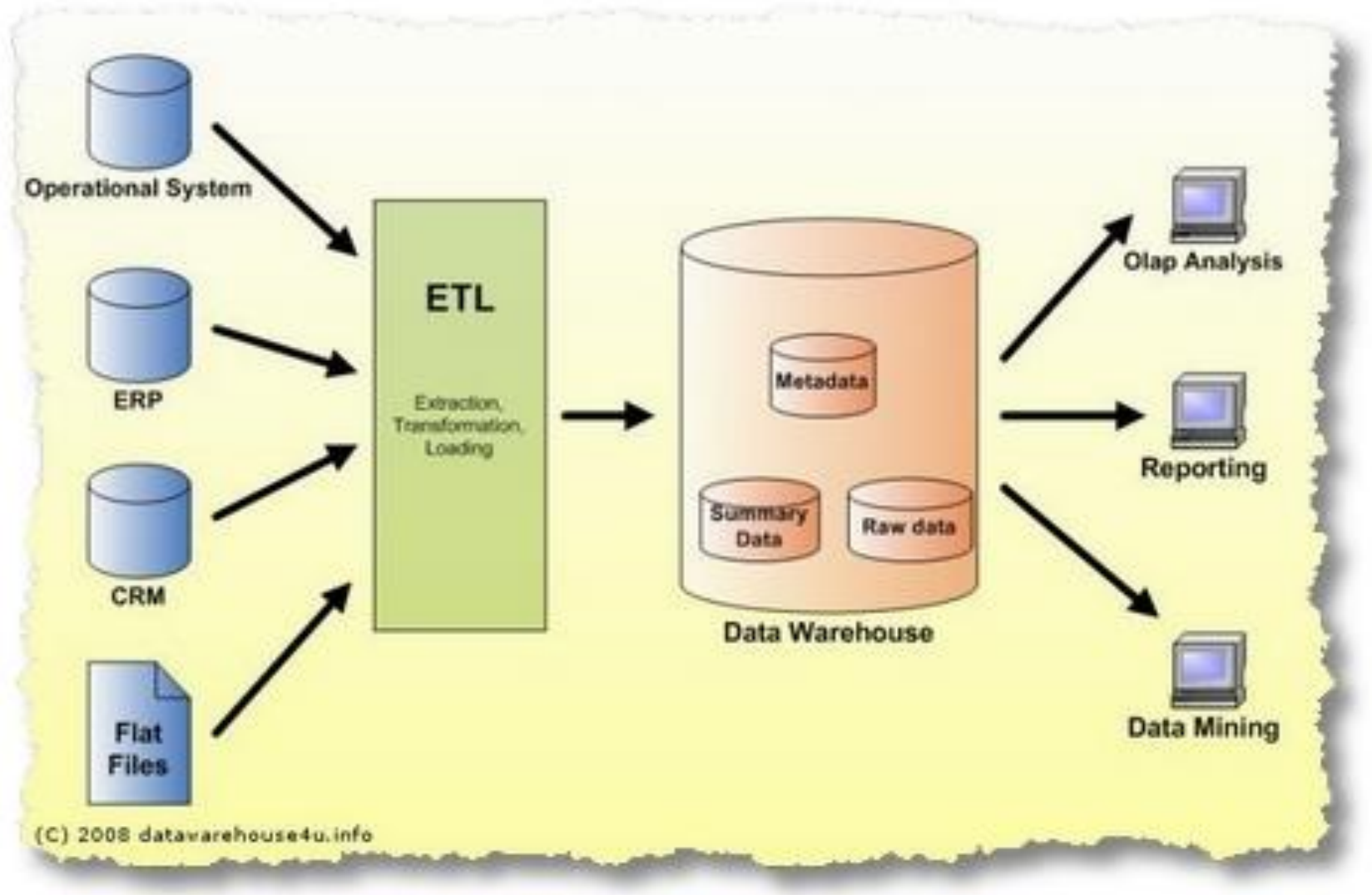
Il faut donc stocker les données avec un objectif en tête

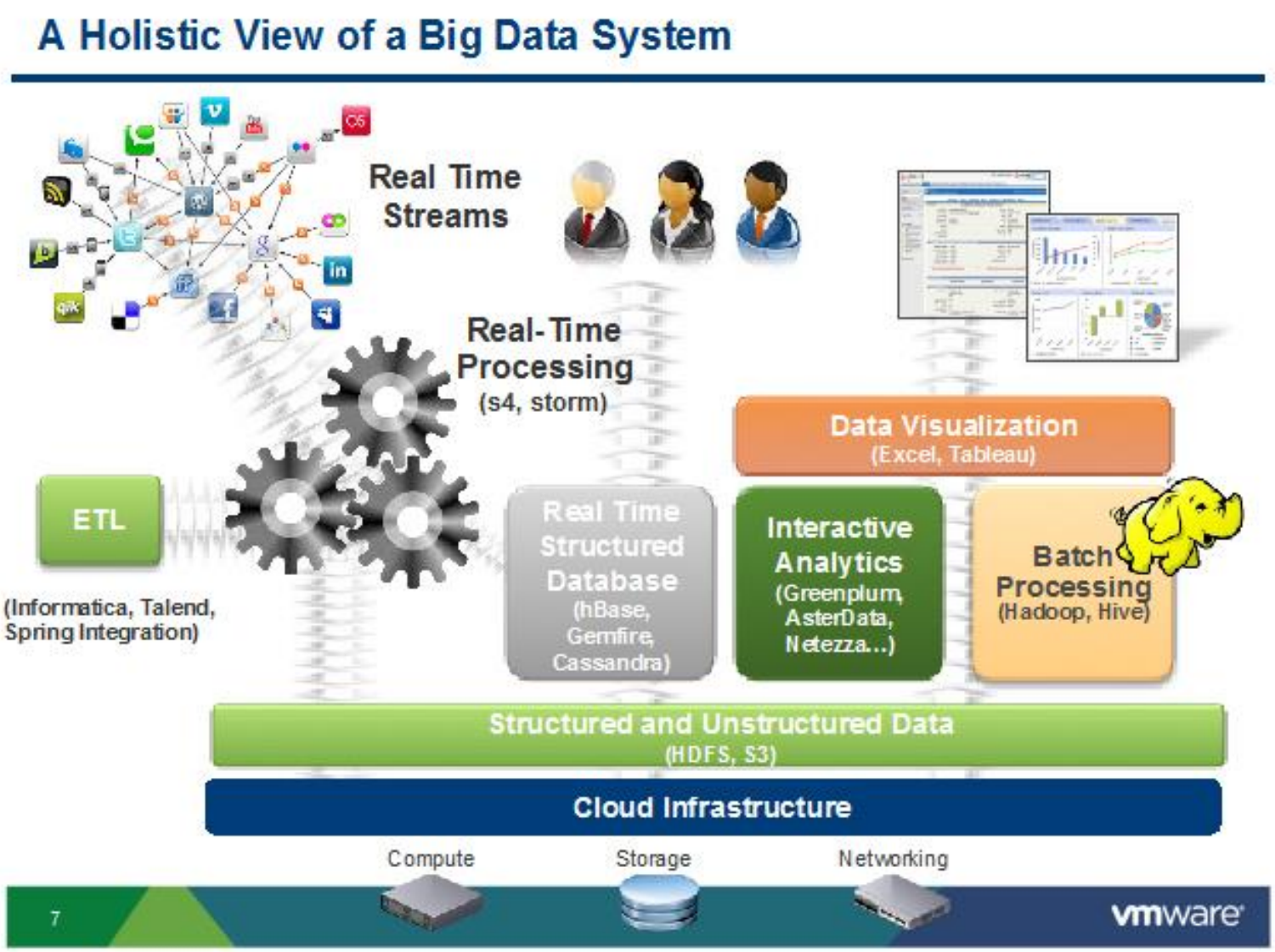
Sinon, le risque est de ne pas pouvoir analyser les données plus tard !

QUAND PEUT-ON FAIRE DU DATAMINING ?



ORGANISATION DU SI ANALYTIQUE




A Holistic View of a Big Data System
Real Time Streams
Real-Time Processing
(s4, storm)
Data Visualization
(Excel, Tableau)
ETL
(Informatica, Talend, Spring Integration)
Real Time Structured Database
(hBase, Gemfire, Cassandra)
Interactive Analytics
(Greenplum, AsterData, Netezza...)
Batch Processing
(Hadoop, Hive)
Structured and Unstructured Data
(HDFS, S3)
Cloud Infrastructure
Compute
Storage
Networking
7
vmware

LA « STACK » BIG DATA

EXPLORE

Dataviz

Analytics

Machine Learning

Nouvelles sources de données temps réel

IoT

Scrapping / Crawling Web

Web Analytics

- Google 360, etc.

Logging / Recherche / Analyse

Batchs

Files de messages

Streams

Bases NoSQL

Qualité des données & ETL

Multiples moyens de transformer la donnée

Système de fichier distribué

VALORISER LA DONNÉE

La démarche la plus évidente est d'insérer entre le SI et les applications analytiques (BI, Datamining) des datamarts spécifiques

Les architectures big data actuelles laissent souvent cohabiter SGBD tabulaires et documents, les uns spécialisés pour ces applications

Valoriser la donnée, c'est d'abord savoir la mobiliser, et mesurer sa valeur le dataminer doit bien connaître les sources de données, maîtriser les différentes couches du SI, travailler avec les administrateurs et architectes des structures de données, au besoin le dataminer est force de proposition sur les enjeux du client et ses besoins d'analyse

La conception d'un datamart analytique consiste à générer, à partir de la donnée brute, des agrégats, des variables calculées, des regroupements et pré-analyses pertinentes, en vue de la problématique qui nous est confiée

Dans le cas des données non structurées, la valorisation consiste également en premier lieu à leur donner une structure en vue de l'analyse

Inventorier

Une grande partie des DSI ignorent le contenu de leurs bases de données. (trop métier, trop diverses, accolées à une application, ad-hoc) Quand au métier, ils sont trop souvent séparés de la données par des interfaces et clients leur faisant perdre sa vision globale

- Le métier (récent) du « Chief Data Officer » i.e. avec une traduction française véritablement pertinente « Directeur d'exploitation de la donnée » est de prendre le relais, d'instaurer une politique de valorisation de la donnée, centralisation des SGBD, architecture globale de datawarehouses et datamarts, règles de constitutions systématiques de dictionnaires de données précis

- L'action conjointe du dataminer, de l'expert métier, et du SI doit permettre d'explorer les ressources de données disponibles afin de chercher filons et pépites

- Les deux premiers axes d'étude sont : La nature de la donnée (sa source, la qualité de sa collecte, l'aisance avec laquelle on peut la relier aux autres ressources) et sa valeur (monétaire, cette fois)

Classer

Inventorier les données c'est les qualifier selon leur rareté, et leur valeur à l'exploitation

- Données concurrencées :

  Données non-exclusives (Broker ou open-data) n'apportant de valeur ajoutée qu'en croisement avec les données propriétaires

  La qualité et pertinence sur le long terme dépendent de la collecte et la mise à jour, par un acteur tiers

  Les données d'un partenaire sont évaluées dans leur degré d'exclusivité selon la politique d'ouverture du partenaire

- Données Junk :

  Données sans rapport au métier ou impossibles à croiser pour des raisons de qualité

- Données Stratégiques : À l'inverse, générées en interne, sécurisées, spécialisées, sont parfois trop rares (**stratégiques**).

  Les données sont elles exploitables aisément et utilement ?

  (ex. données de visites d'un site, de réseaux sociaux, les clefs de valorisation sont parfois peu claires, le lien avec les clients ténu, une veille et une temporisation indispensables) (**dilemme**)

		Valeur pour l'entreprise	
		Faible	Forte
Nature de la donnée	Rare	Donnée dilemne	Donnée stratégique
	Publique	Donnée junk	Donnée concurrencée

©Marketing Analytics BtoB

Monétisation des données ?

N'apporte rien à l'entreprise (sinon du cash)
Peut même être nuisible si des partenaires peu scrupuleux engagent des démarches marketing lourdes

Source de valeur interne ?

Exploitation marketing, (acquisition de clients, conversion de prospects fidélisation)
ré-usage R&D, réduction couts industriels, image de marque, mieux connaitre sa gamme, ses clients etc.

Qualifier

Dans l'exemple de données prospects/clients l'adresse postale reste le point de départ qui sert le plus souvent à réunir les sources de données et les enrichir

- Après une frénésie dans les années 2000 on a constaté la contre-productivité d'accumuler les données sans vision et plan d'ensemble, parfois même sans référentiel unique
- Les données exogènes (socio-demo, géomarketing) doivent avoir un but précis palier à un manque de connaissance, de précision des informations endogènes

Encourager une double qualification des données

- Au-delà de l'acquisition de données et leur objectif de valorisation, encourager l'usage régulier de questionnaires, d'enquêtes, permettant de choisir la donnée que l'on souhaite obtenir de notre population

Mesurer la qualité des données, le taux de complétion des indicateurs clefs, des principaux agrégats et dimensions d'analyse, définir des seuils de tolérance en dessous duquel la donnée n'est pas réellement exploitable.

Historiser la donnée, mesurer sa rapidité d'évolution, la durée pendant laquelle elle reste pertinente

EXPLOITER LES METADATA

L’exploitation des méta-données est le meilleur exemple à donner de la viabilisation et la valorisation des données.

Cette méthodologie amène à chercher de la valeur, de l’information, dans des « artefacts » techniques (date d’enregistrement, géolocalisation des utilisateurs d’un site, volume de données collectées par période, pics de fréquentation, actions répétées etc.)

- Assurer la pertinence de ces informations, leur exploitabilité, les opportunités offertes
- Définir des processus d’analyse « quick win » permettant de vérifier des hypothèses métier sur le métadonnées avant d’en généraliser l’extraction
- Ex : est-ce que des pics de mention de mon produit sur les réseaux sociaux, les éléments de langage nouveaux sur les blogs spécialisés ou dans la presse, sont liés à des évolutions des ventes, des phénomènes de churn ?

Bâtir des modèles prédictifs d’abord très spécialisés, limités à quelques phénomènes, avant d’intégrer ceux-ci à mes analyses grand-angle

Notion de qualité des données

- Définition
- Sources d’anomalies
- Quel comportement, Prévenir ou guérir

Définition

La donnée est une ressource, l'évaluation des ressources et leur confrontation aux problématiques auxquelles on souhaite répondre est la base de toute démarche projet.

Le projet de datamining, et plus généralement l'appréhension des données d'une entreprise, d'une activité, par un dataminer ou en informatique décisionnelle doit commencer par un audit des données

Les buts de l'audit des données

- Le dataminer est amener à collaborer avec le métier et les administrateurs de bases de données.
- Poser la fondation, pour un ou plusieurs projets, non orienté par l'exposé d'une problématique.

Axes d'évaluation de qualité des données

Disponibilité

La disponibilité des données est mesurée par la capacité à centraliser et mettre en relation les données de sources diverses, elle rejoint les impératifs des projets de DMP, de vue unique du client.

Actualité et pertinence

L'actualité et la pertinence des données tient tant à leur mode d'alimentation, leur mise à jour régulière, à la détermination de la profondeur d'historique disponible et à la mesure de la dégradation de qualité des données et de maintient de leur cohérence avec l'ancienneté.

Diversité

La diversité des données assure l'exhaustivité des facteurs d'analyse, la pertinence et la valeur ajoutée des modèles.

Axes d'évaluation de qualité des données

Complétion

La complétion des données tient à l'alimentation régulière et conjointe des sources de données diverses, à une profondeur d'historique analogue, à ce que certaines variables ne fassent pas défaut pour une part considérable des observations de la population.

Structures

Analyser et comprendre les structures de données par une analyse descriptive, soumise au métier, valide les variables les plus pertinentes pour tout analyse, les degrés d'agrégations divers à considérer, les distinctions à introduire et nouvelles variables à calculer.

Justesse

Le processus de données exempt la base d'analyse de données aberrantes, mal formatées, approximatives, conjugue à la fois des éléments déclaratifs et des éléments objectifs

Rôle du dataminer dans la qualité des données

- S'impliquer dans les processus de constitution de datamarts, pour assurer la disponibilité, le maintient de la cohérence et la mise à jour des données qui seront disponibles pour l'analyse.
- Faire remonter toute anomalie afin de permettre au S.I. les investigations nécessaires et la mise en place de contrôles de qualité en amont.
- Évaluer les volumétries réelles, c'est-à-dire les données mobilisables pour une analyse, les variables véritablement exploitables, apportant réellement une information nouvelle.
- Maintenir une veille sur les sources de données, l'opportunité de les enrichir de données publiques (opendata, sociodémo…)

LE PROJET DE DATAMINING

Le projet de datamining

1) Identifier les problématiques
   A. Le Dataminer et le Métier
   B. Recueil des besoins
   C. Hiérarchisation des données, agrégats métiers, etc.
2) Identifier et assembler la donnée
   A. Identifier les sources de données
   B. Estimer l'adéquation et l'exploitabilité des bases
   C. Consolider et réconcilier les données diverses
3) Cycle Analytique
   A. Audit des données
   B. Exploration des données
   C. Gestion des variables
   D. Modélisation
   E. Cycle
4) Technologies et production
   A. Mise en production et vie d'un modèle
   B. Logiciels d'analyse, temps différé
   C. Appliquer un modèle en temps réel

Indexation d'images de visages

Problématique :

Parmi un grand volume et une grande variété d'images, contenant ou non un ou plusieurs visages , reconnaître et indexer des visages

Enjeux :

Part de « bons » (I.e. de visages) parmi les images extrêmement faible

Nécessité de taux de fausses détections très faible, on privilégie le risque de manquer un visage plutôt que d'indexer d'autres images, mais il faut un taux de détection élevé.

Nécessité de détection très rapide, presque temps réel

Application possible :

Afficher le visage d'une personne dans les résultats de recherches la concernant à côté de la page contenant cette image (résumé de page enrichi, multimédia)

Source : R. Féraud, O. Bernier, J.E. Viallet and M. Collobert.
« A fast and Accurate Face Detector based on Neural Networks »
& « Neural Network Based Face Detection »

EXPLORE

Un modèle composite hiérarchique

92%

Parcours rapide

8%

60%

Filtre de couleurs

3.2%

93%

Réseau de neurone Rapide

0.2%

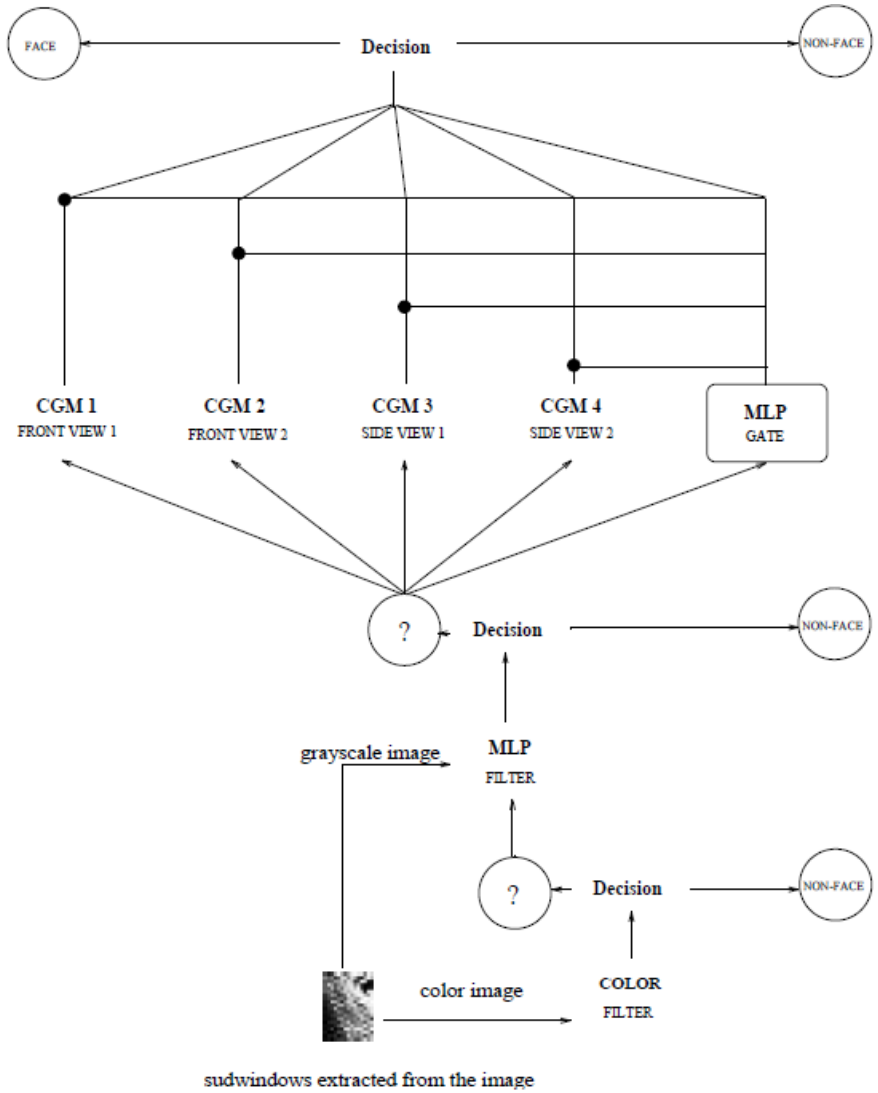
99.8%

Réseau de neurone modulaire

1.27 secondes

Problème, la quasi-totalité des images du web ne représentent pas des êtres humains, un modèle discriminant classique basé sur la distribution des visages, et des non-visages, est inapplicable car un « non-visage » n'est pas définissable, la distribution statistique des « non-visages » ne peut donc pas être discriminante

LE MODÈLE DÉCISIONNEL



sudwindows extracted from the image

The face detector is composed of three stages. The last stage is the only one which is able to decide if the analyzed subwindows is a face.

CGM : Constraint Generative Model

MLP : Multi-Layer Perceptron

CAS D'EXEMPLE : INDEXATION D'IMAGES DE VISAGES

EXPLORE

Quelles données sont à notre disposition :

Images (décomposition en pixels ou en zone, couleurs, etc.) de tailles différentes

Métadonnées des images

Une partie des images sont qualifiées par les métadonnées, ou connues (sélectionnées) on sait qu'elles comportent ou non un visage, ce sera notre échantillon d'apprentissage

Quelle est la cible ?

Quelle est la différence ente les images avec/sans visage ?

Comment sélectionner les images de visages ?

Comment isoler la portion d'image contenant le visage ?

La problématique spécifique ?

Comment ne pas avoir à faire de compromis entre robustesse du modèle et rapidité d'analyse ?

Réseau de neurone modulaire (Modèle génératif contraint)

Un réseau modulaire est constitué de plusieurs réseaux de neurones apprenant une portion du problème (visages de face, visages de profil) et d'un « portail » (ici une ACP) affectant à chaque image le modèle le plus adapté

Un modèle génératif est un modèle entrainé par une probabilité conditionnelle (ici un modèle linéaire discriminant) et non une correction d'erreur (on parle alors de modèles discriminatifs : régressions logistiques, perceptrons, SVM, etc.)

Le réseau génératif contraint consiste (avec des couches neurales de précisions variables) à reconstruire sur chaque output chaque élément estimé « non-visage » en le projetant sur l'espace des visages comme la moyenne des plus proches visages, couche après couche le réseau apprend donc à projeter les non-visages et donc, estimer la distance d'une image à l'espace des visages comme la distance entre l'image et sa projection

Un modèle génératif contraint utilise un certain nombre de « contre-exemples » (ici des images ne représentant pas de visages, mais de plus en plus ressemblants à des images) pour forcer la réduction de l'espace de détection.

CAS D'EXEMPLE : INDEXATION D'IMAGES DE VISAGES

EXPLORE

Filtre basé sur un perceptron multicouches (MLP) :

Le réseau est compose de 300 inputs (1 par pixel d'une fenêtre de 15x20 extraite de l'image initiale) , 20 neurones cachés et un output (visage / non-visage) soit 6041 pondérations. Le réseau de neurones est entrainé par une rétro-propagation standard. Ce réseau est rapide et léger, mais s'il reconnaît 99% des visages mais a toujours 1% de faux-positifs (images considérées à tord comme des visages) et ne peut être utilisé seul.

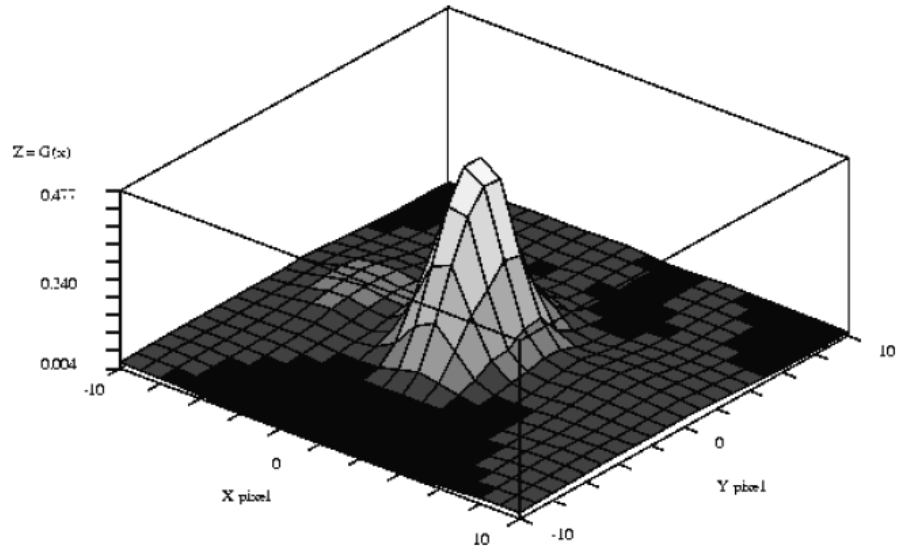
Filtre de couleur :

Avant de transformer chaque image en niveaux de gris, lissés, pour permettre une meilleure efficacité du MLP, le filtre exclue les portions d'image ne contenant pas assez de teintes de chair. L'image étant décomposée en zones de 15x20 pixels, les zones comportant une faible densité (<40%) de teintes de chair sont considérées comme « background » les autres sont soumises au MLP après égalisation d'histogramme (amélioration de contraste), lissage, et soustraction du « visage-moyen »

Algorithme de parcours rapide : excluant les images les plus improbables

N'importe quel modèle rapide avec un taux de faux négatifs quasi-nul sur les visages mais mais un taux de faux positifs potentiellement important, permettant d'exclure 92% des images du web

Usage du modèle dans un moteur de recherche



Pour une image, le résultat de l'algorithme décroit rapidement et de façon monotone autour d'un point central de visage

L'algorithme de recherche consiste donc à tester un certain nombre de points d'une grille (a), puis faire une recherche exhaustive autour du maximum (b)

Une vignette est ensuite extraite de l'image de départ, correspondant à une zone de bonne détection, la plus petite possible, à afficher à côté des résultats de recherche

(a) The grid

(b) The detected face

(c) The extracted cropped frame

PRÉREQUIS D'UN PROJET DE DATAMINING

EXPLORE

Mon entreprise a-t-elle besoin de lancer un projet de datamining ?

- Des données à valoriser
- Les données sont au cœur de mon activité
- Ma pratique doit évoluer

Poser une problématique

- Pourquoi ais-je besoin de ces analyses ?
- Que m'apporteront-elles ?
- Comment mesurerais-je l'apport de la démarche datamining ?
  - Quel retour sur investissement en attendrais je ?

Mobiliser la donnée

Se doter d'une infrastructure IT adaptée

EXPLORE

Quand et comment lancer un projet de datamining ?

Dois-je m’adresser à un conseil extérieur ou internaliser le datamining ?

Quels choix technologiques ais-je à faire ?

Qualifier mon projet : Connaissance Client ? Big Data ? Algorithme et processus ou étude ponctuelle ? Un support pour mon métier ou un outil au cœur de mon activité ?

Points de vigilance projet

L'erreur la plus fréquente est de stocker les données Big Data dans des entrepôts de données en se disant qu'on les analysera plus tard.

- On a vu que le format de stockage des données doit être adapté aux types d'analyses qu'on souhaite réaliser.
- Mais on doit aussi choisir ce que l'on stocke, à quelle fréquence, dans quel but, pour faire quels types d'analyses.
- Le format de stockage doit aussi être choisi en fonction du type de base que l'on va être amené à manipuler en terme de taille, de type de données, de variété des données, et du type d'utilisation que l'on va en avoir

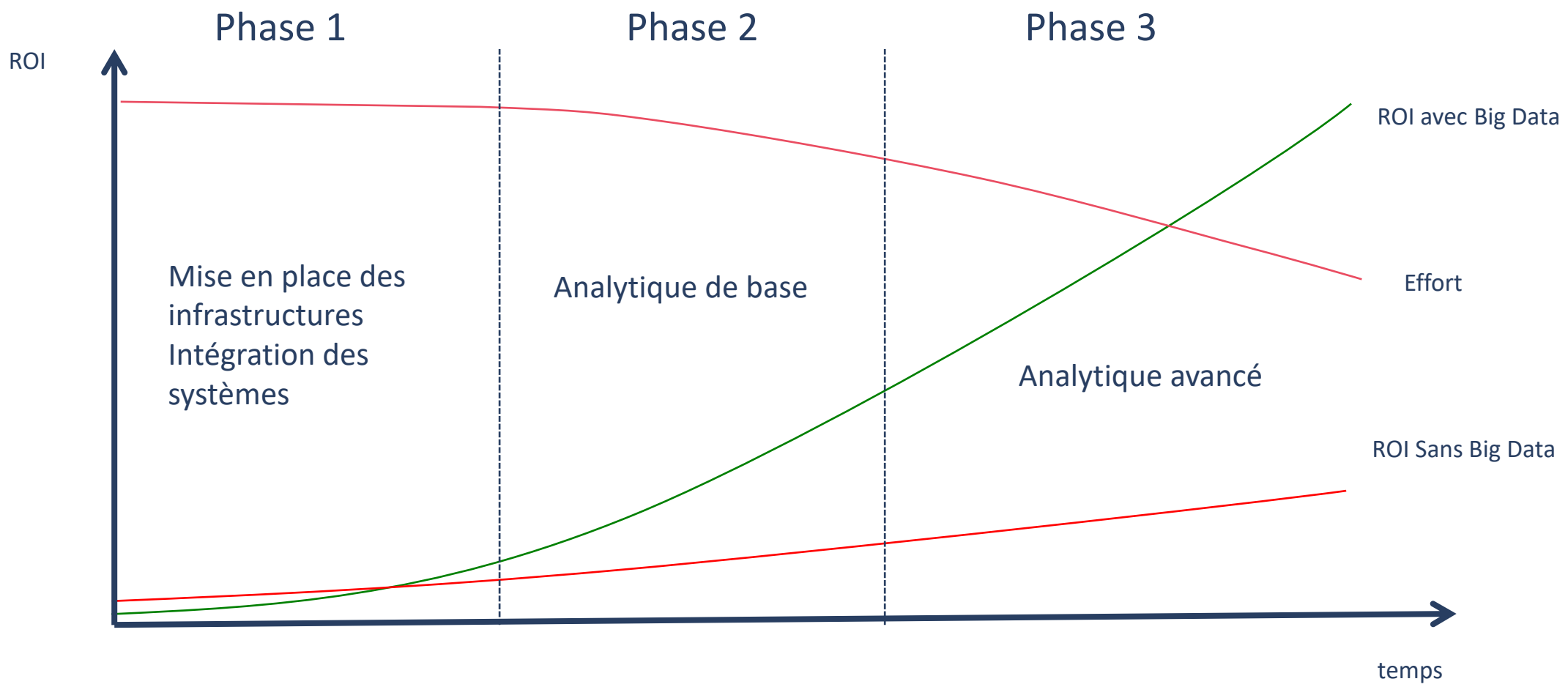
Il faut être vigilant dans la phase exploratoire, notamment sur les échanges nécessaire entre les acteurs métier et les analystes

- Il est vital de s'assurer que les analystes maitrisent bien les aspects et les enjeux « métier »
- La phase exploratoire dépend des données, et il faut donc accorder une certaine souplesse sur la durée de cette phase

Points de vigilance projet

- Privilégier un périmètre réduit mais maitrisé, plutôt que d'attendre pour démarrer un projet gigantesque potentiellement incontrôlable.
- Rigueur de la gestion de projet, le fait qu'il s'agisse d'un projet de type Big Data, nécessitant une certaine flexibilité dans la phase exploratoire ne doit pas faire oublier les bases.
- Veiller à la qualité et à la pertinence des données utilisées
- Traiter les projets Big Data comme des projets transverses et ne pas les cloisonner à un secteur, un département ou une division.
- Vérifier les aspects juridiques, notamment sur la protection des données privées par rapport aux données que vous compter utiliser et l'objectif poursuivi.

TYPE D'APPROCHE, IMPACTS ET BÉNÉFICES DANS LE TEMPS



COMPÉTENCES ET MÉTIERS IMPLIQUÉS

EXPLORE

- Administrateurs de Bases de données
- Architectes data
- L’Informatique décisionnelle
- L’analyste statisticien, dataminer, ou datascientist
- Les développeurs
- L’expert métier

Le CDO « Chief Data Officer » :

- Encore peu répandu en France,
- Niveau des décideurs de l'entreprise, il participe au comité de direction.
- Il est en charge de la gouvernance des données de l'entreprise
- Rôle stratégique et multiple

Le Data Scientist

- Le Data Scientist est celui qui produit la valeur de la donnée.
- Multiple compétence : métier, statistique et IT
- Capable d'analyser les données, et d'interpréter les résultats d'analyse
- Capable de faire des recommandations tactiques et stratégiques

Le Data Stewart

- Responsable de la mise en œuvre de la stratégie et de l'application de la gouvernance décidée par le CDO sur le terrain
- Responsable de la qualité des données

IDENTIFIER LES PROBLÉMATIQUES

Le dataminer et le métier

- Etre à l'écoute
- Etre force de proposition
- Garder la pertinence et la scientificité de la démarche

On peut établir des règles pour une relation saine et fructueuses entre les personnes expertes dans le métier cadre d'une analyse d'une part et l'analyste d'autre part.

1. Etre à l'écoute

2. Etre force de proposition

3. Faire preuve de pédagogie

4. Conserver la rigueur et scientificité de la discipline

Etre à l'écoute et force de proposition

- Dans une démarche visant à faire le pont entre des données brutes – obtenues dans le cadre de l'exercice du métier – l'information dont le métier à besoin, et la connaissance de ce qu'il ignore encore, la collaboration est indispensable.

- Cette collaboration doit commencer par l'écoute, sans imaginer maitriser les tenants et aboutissants de l'activité de tous les commanditaires, clients, entreprises avec lesquelles on est amené à travailler, on ne doit jamais se comporter en « boite noire » appliquant des règles strictes et des modèles incompréhensibles qu'elle qu'en soit la qualité.

- A toute étape du projet le métier peut, et doit, être sollicité, percevoir et orienter les premiers résultats de l'audit (choix de seuils de données atypiques, valeurs considérées aberrantes et éliminées, regroupement de modalités et leur sens, etc.), mais également de l'analyse exploratoire (liaison entre variables, profils de la population), et des modèles soumis au métier

- A l'interprétation du rôle des variables dans le modèle à la fois enfoncer des portes ouvertes (ce qui confirme son adéquation) confirmer des pressentis et intuitions métiers encore non prouvées, et lever des phénomènes insoupçonnés ou contre-intuitifs, que l'on validera séparément

Etre à l'écoute et force de proposition

- Le second point important est que l'expert métier (marketeur, responsable e-commerce, industriel, concepteur d'application etc.) n'a pas forcément une vision claire, ni du contenu de ses données, ni de ce qu'elles peuvent déceler, et du rôle que peut avoir l'analyse pour son business.
  - « Je veux un score » ? « Un score de quoi ? » « bah un score quoi… » /*en fait c'était une segmentation*/
- Une interview métier générale aura pour objectif d'exposer au dataminer les problématiques en cours, les enjeux de l'interlocuteur, le sentiment de la personne sur sa population (de clients, d'utilisateurs, de produits) les questions qu'il se pose… et l'amener à réfléchir sur les données et ce qu'elles pourront éventuellement permettre d'expliquer, de qualifier.
- Le but du dataminer sera d'orienter, guider la réflexion, et jauger de la faisabilité de l'analyse au regard des données disponibles
  - Ex: cibler une campagne marketing de prospection et évaluer sa pénétration est impossible si on ne dispose pas de données de réactivité passée d'autres campagnes mais uniqement de données d'activité client, on ne pourra qu'évaluer la valeur potentielle des futurs clients mais rien ne permet d'affirmer qu'ils soient plus facile à transformer que les prospects les moins rentables.

L'impératif de pédagogie

- Les principes et règles du datamining sont souvent acceptés mais rarement réellement comprises.
  - Concepts de modélisation et prédiction, de segmentation de population, de corrélation, variable à prédire vs variables explicatives…
  - Au mieux boîtes noires, au pire presque « magiques » capables de tout prédire dans n'importe quelle condition

- La pédagogie est en conséquence une qualité essentielle du dataminer,
  - la présentation de choix et décisions de datamanagement est expliquée, justifiée
  - La présentation des modèles essayés, de ceux rejetés, demande parfois l'explication de la méthodologie, voire, et particulièrement dans le cas où l'interprétation d'un modèle paramétrique suscite des recommandations métiers, des modes de calculs des modèles

- Par ailleurs, il est souvent indispensable pour une entreprise de faire collaborer profondément les analystes à la force de vente en raison de la difficulté de synthétiser un discours commercial sur le datamining et détecter les opportunités

La caution scientifique

- Le flou dans lequel peuvent être beaucoup d'interlocuteurs nous oblige à beaucoup de rigueur.

- Savoir dire NON
  - Expliquer l'insuffisance des données, en volume, en qualité, recommander une nouvelle problématique ou comment améliorer et rendre exploitable les données

- Expliquer et relativiser
  - Savoir expliquer les contraintes d'un modèle (stabilité d'une segmentation par exemple), sa qualité (interprétation d'une courbe de lift), ses insuffisances (seuils d'imprécision, vraisemblance, pertinence, risques de confusions, présentation des individus limites),
  - Savoir exposer la signification des indicateurs produits par l'outil statistique (odd-ratios, coefficients et paramètres et leur interprétation métier possible, veiller à relativiser grâce aux tests statistiques

- Accompagner
  - Sélection d'une cible de campagne, transcrire en recommandations métiers les conclusions d'étude

IDENTIFIER LES PROBLÉMATIQUES

EXPLORE

Recueil des besoins

- Problématiques Métier
- Données cibles et explicatives potentielles
- Degré de préparation, avancement et maturité dans la démarche de datamining

Etablir les objectifs du projet

Le budget disponible pour le projet doit être connu

- A minima son ordre de grandeur
- Est-on sur un projet one shot ou devant comporter une phase d'industrialisation, automatisé et devant donc compter un investissement IT ?

Les attentes du projet doivent être connues

- Cherche t'on à comprendre et anticiper un phénomène (ex : comportement clients, incendie, reconnaissance de forme, recommandation de média…) ?
- Cherche t'on à gagner de l'argent (quick Wins) ou à investir, capitaliser sur des outils potentiels ?
- Cherche t'on à convaincre rapidement une direction, des investisseurs, un client sceptique ? A démontrer un concept (P.O.C.)
- Ces objectifs peuvent être complémentaires mais parfois contradictoires

Etablir le cadre du projet

Etablir les livrables du projet

- Les livrables doivent être mesurés à l'aide de KPIs.
- Les conditions permettant d'établir un succès, ou à tout le moins un jalon pouvant susciter un nouvel investissement être établis
- Au contraire les prérequis indispensables listés, ainsi que les conditions dans lesquelles, au regard des premiers résultats, le projet doit être abandonné, repoussé, ou développé et enrichi (nouvelle collecte de données par exemple) pour aboutir

Etablir des attentes réalistes

- En terme d'ordre de grandeur (R.O.I., volumétrie de données)
- En terme de temps (Départ, Jalons, Délai de mise en production)
- Le bénéfice le plus important n'est pas nécessairement le gain financier

Maturité

Un projet de datamining comprend nécessairement une reprise d'historique

- Cas 1 : Première expérimentation du datamining
  - Opportunité de quick wins dans le cadre du projet
  - Segmentation « métier » intuitives à reprendre et intégrer
  - Budget suffisant pour poser des « fondations » (segmentations globales, base client, base produit, ou dans un cadre hors connaissance client investigation sur les caractérisation statistiques d'une image de visage par exemple)

- Cas 2 : Développement
  - Reprise, audit et mise à jour si nécessaire des algorithmes de datamining déjà en place
  - Quel mode d'industrialisation ?
  - Meilleure exploitation et présentation des résultats d'analyse
  - Assumer un ROI en baisse, la courbe du gain apporté par chaque approfondissement du travail de datamining décroit rapidement

Hiérarchisation des données

- Lier données et problématique
- Niveaux d'agrégats
- Données explicatives, données de support

Ressources disponibles

Quelles sont les données les plus directement liées à la problématique ?

- Données suffisantes explicatives (i.e. hors « informations » telles que identification des individu / descriptif des produits) sur la population
- Données illustratives, ne pouvant contribuer à expliquer le phénomène ou trop anecdotiques ou rares pour être exploitées par une segmentation, mais pouvant être utilisées comme variables illustratives pour contribuer à mieux comprendre les modèles
- Si étude d'un phénomène précis (score, prédiction, décision, etc.) existence d'une « variable cible » ou nécessité de la générer ?

Quels sont les niveaux d'agrégats différents, (dans le cas d'un SGBD relationnel, d'un modèle objet, dé-normalisé : un par table) et leurs dimensionnalités ?

- Lesquels sont les plus susceptibles d'expliquer le phénomène ou caractériser la population
- Lesquels peuvent supporter et affiner nos conclusions
- Lesquels doivent être écartées

Ressources disponibles

Comment exploiter des agrégats de données a priori éloignées de la problématique afin d'enrichir notre analyse

- Données pouvant être agrégées immédiatement, détermination des agrégats pertinents (ex : scope temporel, activité moyenne mensuelle, trimestrielle ou annuelle ?)
- Lesquels peuvent faire l'objet d'une étude à part dont les conclusions peuvent enrichir l'analyse

Déterminer la nécessité d'ajouter d'éventuelles sources de données annexes

- Ajout d'Open data, ex : estimation de CSP sur les iris

Opportunités de raffiner la donnée,

- Ex 1 : investir dans une typologie de la population pour construire des modèles prédictifs différents sur différentes classes, déterminer des gammes de produit suscitant des comportements différents, etc.
- Ex 2 : remplacer un lot de données d'un thème commun assez éloigné du sujet en classes et utiliser cette classe comme unique variable pour ce thème dans l'analyse

EXPLORE

Identification des sources de données

- Interroger l'I.T. (D.S.I. ou C.D.O)
- Obtenir un dictionnaire des données pour chaque source
- Identifier les variables intéressantes pour l'analyse
- Identifier les éléments de jointure

Le rôle de l'I.T.

Dans une structure d'entreprise traditionnelle, le management des données (infrastructures, process de transfert, gestion des bases de données) est dévolue à la Direction des Services Informatiques.

- Les D.S.I. sans sensibilité particulière à la donnée, n'ont parfois pas une vision métier du contenu des bases
- Lesconventions de nommage des serveurs/instances excluent en général l'identification de leur contenu (par sécurité)
- il conviendra soit de maintenir un annuaire décrivant le contenu de chaque base, ou de s'adresser au métier (ou dans le pire des cas d'auditer les processus susceptibles de collecter une donnée pertinente) pour localiser les données que l'on souhaite mobiliser.

Une entreprise peut se doter d'un C.D.O. (Chief Data Officer) qui centralise et soulage la D.S.I de toute la gestion de la donnée et y adjoint le pilotage des fonctions analytiques et de l'informatique décisionnelle.

- Normalement cette structure permet de connaître beaucoup plus précisémment avec un interlocuteur (ou une équipe) unique la localisation, la qualité des données, les technologies utilisées, et les normes en vigueur pour accéder et centraliser dans un datamart les données mobilisées pour un projet

Le dictionnaire des données

- Il est rare d'avoir à tout niveau (nom de base, de tables ou de collections, de colonnes ou de propriétés…) d'un SGBD des noms et labels porteurs de sens.

- L'un des prérequis essentiels d'un projet est donc d'obtenir des personnes responsables de chaque source un « dictionnaire » complet du contenu des bases
  - Noms des tables
  - Jonctions entre les tables (systèmes de clefs primaires/étrangères ou correspondances de propriétés)
  - Nom de chaque colonne ou propriété
  - Formats de stockage, transformations à prévoir
  - Liste des modalités possible des variables non numériques (luxe)
  - Fréquence de mise à jour, fiabilité, taux de valeurs non-renseignées / vides … (rare)

Variables clefs

Construire la variable cible

N.B. Un problème de datamining n'admet pas toujours une variable cible en entrée. Les problèmes « non supervisés » (Classification, Typologies, Analyses factorielles…) visent au contraire à les créer

Qu'est ce qu'un « bon » ou un « mauvais » individu parmi notre population ?

- Individus clients : Client ayant ou susceptible d'avoir tel caractère, connu tel évènement (ex : achat de l'un ou l'autre produit, produire un CA élevé, etc.)
- Images : Correspondre à la cible à reconnaître, subir telle modification
- Logs : Régularité d'usage, non défaillance

Affiner la définition par l'avis de plusieurs experts métiers quelques exemples :

- Transformation d'une sollicitation marketing : Achat dans les x mois ? Augmentant son activité ? Achat d'un produit précis ou non ?
- Web : visite d'une page : Rester durée x sur la page ? Visualiser x% du contenu ? Action (clic) sur la page ? Poursuite navigation depuis cette page v.s. « retour » ?

Variables clefs

- Avant une sélection statistique des variables éligibles à l'analyse et une modélisation il convient d'identifier les éléments les plus pertinents pour le métier :
  - Variables à Prédire
  - Indicateurs clefs de performance
  - Dimensions et ventilations évidentes (origine, canaux, gammes, etc.)
  - Segments identifiés intuitivements
- Ces éléments sont identifiés par le métier comme des aides essentielles pour comprendre les phénomènes, les avoir dans le modèle permettra de mesurer la pertinence de la lecture intuitive de ce qui se produit, ces variables seront le plus souvent en bonne place dans la composition des modèles et permettront de donner une signification plus claire aux phénomènes observés, leurs liaisons pourront également être intrinsèquement riches en enseignement.

Estimer l'adéquation et l'exploitabilité des bases

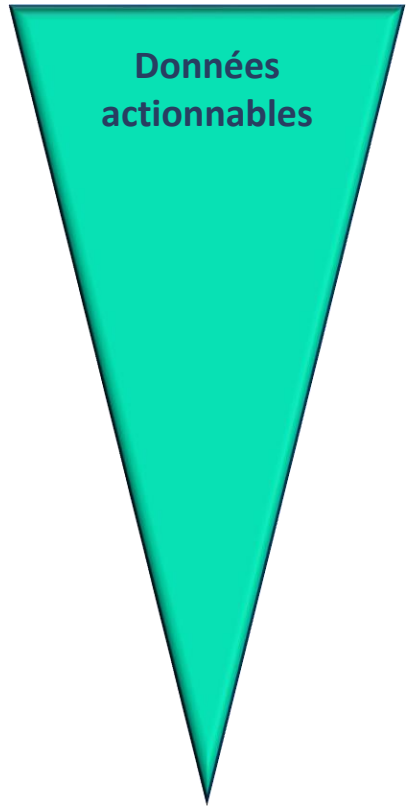
- Un compromis doit être établi entre finesse d'analyse et moyens (budget)
- Les données ont des rapports immédiats ou plus éloignés à la problématique
- Mais sont également plus ou moins facilement exploitables
- Et sont également plus ou moins fiables

IDENTIFICATION DES SOURCES DE DONNÉES

EXPLORE

Classer Les données par source et type

- Données transactionnelles et e-commerce
- Données CRM et ERP
- Données venant des interactions avec les services (email, call centers, …)
- Résultats d'enquête
- Données de campagnes (emails)
- Données Web Analytics
- Bases de externes données achetées
- Données externes libres (Open Data)
- Données venant des réseaux sociaux



Les données ont une relation étroite avec le savoir et la connaissance

Les données sont la base de cette connaissance :

Données justes + Raisonnement juste = Connaissance

Données fausses + Raisonnement juste = *Ignorance*

Données justes + Raisonnement faux = *Ignorance*

Données fausses + Raisonnement faux = *Ignorance*

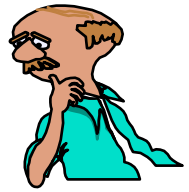
QUELQUES ERREURS COURANTES

- Couverture fausse
- Pas de réponse
- Erreur d'échantillon
- Information demandée

Population exclue

Suivi par téléphone des « non réponses »

Échantillon non représentatif



Mauvaise question

IDENTIFIER ET ASSEMBLER LA DONNÉE

Consolider et réconcilier les données diverses

- Poser les bonnes questions
- Méthodologie de conception d'un datamart
- Etablir les règles de jointure

QUESTIONS À SE POSER SUR LES DONNÉES

- Les données sont-elles représentatives de la réalité ?
- Les données sont-elles recueillies pour définir la réalité?
- Les données sont-elles bien renseignées ?
- Les calculs effectués à partir des données sont-ils justes?
- Le raisonnement effectué à partir des données calculées est-il juste ?
- Les données nécessaires ont-elles été relevées?
- Les données sont-elles séquencées correctement?

CONSTRUIRE LA VISION 360°

- La vision 360° passe nécessairement par un datamart orienté métier
- Quelles sont les informations dont on a besoin à court, moyen, long terme
- A-t'on accès à toutes les données nécessaires ?
- A-t'on la possibilité de construire les flux de données ?
- Comment les flux de données vont-ils alimenter le datamart ?
  - Quels ETL ?
  - Quelle fréquence de mise à jour ?
  - Batch ou temps réel ?
  - Quelles types de transformations et de recodages ?
  - Quelles sécurité en cas de défaillance ?
- Analyse du risque en cas de sinistre, de piratage, de vol de données
- Qui a accès ? Qui est responsable ?

CONSOLIDER ET RÉCONCILIER LES DONNÉES DIVERSES

- La démarche de conception d'un datamart analytique, d'un jeu de données d'étude doit être itérative, progressive

- A chaque adjonction de sources de données on évalue :
  - Les agrégats nécessaires
  - Les nouvelles variables à créer, les variables à enrichir/compléter d'une source sur l'autre (règles de rejet ou remplacement en cas de conflit)
  - La complétion des jointures (quelle part de la population est commune aux deux sources de données, pour que les nouvelles variables soient pertinentes pour l'analyse)
  - Les règles de gestion des deux sources (périodicité de mise à jour, quelle source est référence en cas d'incohérence)
  - Quelle profondeur d'historique est conservée
  - Les règles permettant de ne pas générer de doublons

CONSOLIDER ET RÉCONCILIER LES DONNÉES DIVERSES

EXPLORE

- La jointure de sources de données distinctes n'est pas toujours établie sur un seul identifiant
  - on peut concéder une perte de précision d'une part pour enrichir le sens métier
  - ex : adjoindre des données transactionnelles à l'étude d'un catalogue produit, comment considérer le remplacement naturel d'un produit par un autre lors du renouvellement de catalogue. Les transactions sur l'un et l'autre peuvent être confondues

- Chaque source de données a un rôle métier propre
  - Ex : groupe commercial : base des tickets magasins et base e-commerce, (données transactionnelles) , base du programme de fidélité, base des stocks, base marketing, CRM (retours clients/SAV) etc
  - Chacune peut obéir à des règles distinctes établies en cloisonnement et parfois contradictoires (ex: un client est-il un individu ou un foyer)
  - Pas forcément d'id commun (utiliser patronyme et bloc d'adresse ou email ? Comment détecter doublons ?)
  - La jonction de deux sources peut susciter des interogations spécifiques ex : comment la base marketing et les bases de ventes permettent de détecter une transformation ? (achat dans un délai x suivant une sollicitation, d'un produit plus ou moins, ou non, lié au thème de la campagne)

Audit des données

- Points de vigilance
- Données manquantes
- Données aberrantes
- Données atypiques

AUDIT DES DONNÉES EXISTANTES

- Utilisation des méthodes d'analyse descriptives multivariées
- Un point de vigilance particulier est mis sur :
  - Avoir la liste de ***toutes*** les sources de données
  - Les sources de données dupliquées
  - Les données contradictoires ou incohérentes
  - Les différences de codage des données
  - Les données manquantes
  - Les données aberrantes (outliers)
  - Les risque de perte d'intégrité des données
  - Les différents états d'une même données dans le temps
  - Les données mal (ou pas) définies
  - Les processus de collecte mal maitrisés
  - Les tables de références mal définies (ex : nomenclature produits)
- Traiter, Nettoyer les données
- Quelles sont les référentiels déjà construits, sont ils exploitables ?

Données aberrantes

Définition

- On parle de données aberrantes quand un ou plusieurs individus portent, pour un caractère (continu) une grandeur inconciliable avec le reste de la distribution
- Ex : OdM, don moyen à 50~100 euros, 95% de la distribution sous les 500€, quelques dons à >10k voire 100k€
- La donnée aberrante peut être issue d'une erreur de saisie, de calcul, ou un réel comportement anormal d'un individu de la population, l'avis du métier est indispensable

Risque induit par les données aberrantes

- La plupart des méthodes d'analyse conçues pour les variables continues (ACP, Régression) sont extrêmement sensibles aux données aberrantes, en ACP ils risquent de diriger un axe à eux seuls, en régression de détourner la pente…
- Pour les modèles prédictifs on parle de « sur-apprentissage » quand le modèle se cale sur quelques individus anormaux de l'échantillon d'apprentissage et perd donc en qualité

Traitement des données aberrantes

- On peut exclure totalement l'individu (solution à privilégier si la variable concernée est capitale dans le modèle) ou si l'individu en lui-même est réellement anormal.

- On peut considérer la donnée comme manquante

- On peut créer une nouvelle variable comptant le nombre d'épiphénomènes s'il est nécessaire d'agréger la donnée
  - Ex : variable montant présentant des valeurs aberrantes rares au-delà d'un seuil
  - Montant_max, montant_moy, montant_annuel ignoreront les valeurs au-delà de ce seuil
  - Nb_versements_exceptionnels comptera les évènements aberrants

Données atypiques

Définition

- Une donnée est atypique si elle est relativement rare, correspond aux extrémités de la distribution de la variable, mais ne peut être considérée comme aberrante d'un point de vue métier
- Les données atypiques peuvent correspondre à une « niche » et donc d'un point de vue statistique à un mélange de lois,

Risque induit par les données atypiques

- Les risques sont sensiblement les mêmes que pour les données aberrantes, dans des proportions moindres, un modèle peut être biaisé par les données atypiques surtout si la distribution est asymétriques

Traitement des données atypiques

- Il est hors de question de supprimer les individus concernés par les données atypiques, ils peuvent au contraire correspondre à des segments particulièrement intéressants de la population, mais on peut les séparer et évaluer deux modèles

- Si on considère que la qualité du modèle peut en souffrir, on préfèrera recoder la variable

- On peut également évaluer un mélange de loi et créer deux variables, ou exploiter des modèles de type ANOVA pour chercher à les séparer au regard des variables qualitatives de l'analyse

Modalités rares

Les modalités rares sont le pendant qualitatif des données aberrantes ou atypiques.

La caractérisation d'une modalité comme rare dépend évidemment :

- De la régularité de la distribution
- De l'effectif de la population
- Du nombre de modalités pour cette variable ET les autres variables de l'analyse

On traitera les données rares par des regroupements de modalités jusqu'à avoir un jeu de variables de cardinalité assez similaire pour ne pas impacter la qualité des modèles

Traitement des données manquantes

Une donnée est manquante quand l'information de réalisation d'une variable pour un individu est indisponible

Pour les variables qualitatives, et en particulier dans le cas d'un questionnaire, -on parle de non réponse- les modèles proposeront de considérer les manquantes comme une modalité, la non réponse peut porter une information

La non réponse peut également dépendre d'une autre variable (ex : nb d'achats, montant, ou questions subsidiaires « Si oui pourquoi » ?)

Points de vigilance

- Outils de manipulation de données qui vont introduire des sigles (NA, NULL, '') pour les données manquantes, ou pire, leur attribuer une valeur (un 0 par exemple)
- Une erreur lors d'un import ou dans les traitements d'un ETL et une chaine de caractère vide peut être introduite qui ne sera plus considérée comme une valeur manquante « null » par les SGBD mais (sauf cas exceptionnel de contexte) devra l'être
- Dans un modèle JSON, que signifie l'absence d'une propriété d'un document, v.s. sa présence vide ?

Traitement des données manquantes

Pour traiter les données manquantes, bien comprendre le rôle de chaque variable, l'avis du métier est indispensable, la donnée manquante peut en fait avoir du sens, pour les continues on pourra la remplacer par une valeur (0 par exemple pour un montant d'achat dans une catégorie si aucun achat n'a été effectué)

La plupart des modèles EXIGENT une valeur pour chaque entrée lors de la phase d'apprentissage mais peuvent les admettre en projection lors de l'application , il conviendra alors :

- D'estimer l'exploitabilité des individus, si trop de valeurs manquantes il faudra les supprimer, si un nombre suffisant d'individus n'ont aucune manquantes ils serviront aux échantillons d'apprentissage
- Dans le cas contraire, les outils statistiques proposeront de remplacer par diverses indicateurs de tendance centrale (moyenne, robuste ou non, mode d'une qualitative, demi-étendue, médiane..) ou une valeur aléatoire avec ventilation des manquantes selon la distribution.
- Si cette variable a un rôle éminent dans l'étude et que l'investissement en temps est possible, on pourra éventuellement la prédire grâce à un modèle

EXPLORE

Exploration des données

- Démarche d’analyse exploratoire
- Méthodologie
- Le rapport d’audit
- Le métier et l’exploration des données

Démarche d'analyse exploratoire

Souvent menée en partie de front ou de façon alternée avec l'audit de qualité des données (si on a regardé les distributions, les outliers…) c'est notre premier contact avec les données.

La démarche d'analyse exploratoire consiste, comme son nom l'indique, à découvrir notre environnement, à reconnaître le champ de bataille, à défricher les données.

C'est une démarche éminemment inductive, allant du plus particulier au plus général, demandant de la méthode, qui se déroule en trois temps.

- Examiner chaque variable
- Examiner les variables deux à deux, en mettant en avant les variables pressenties – ou indiquées par le métier – comme plus importantes
- Etablir les liaisons globales multivariées, la forme de notre univers dont chaque variable est une dimension

Tenir un rapport des observations effectuées, s'obliger à l'exhaustivité, à creuser les phénomènes, à accorder à chaque résultat le temps nécessaire est impératif ! Il est extrêmement facile de rater un élément important

Ne pas hésiter à interroger le métier sur l'allure de tel résultat ou telle distribution

Méthodologie

A chaque étape correspond sa méthodologie propre et, avec l'accroissement de la complexité et de la dimensionnalité, des méthodes statistiques de plus en plus évoluées.

Analyse univariée :

- Vérifier la distribution des variables, établir les contraintes de normalité éventuelles des modèles, évaluer les éventuelles mises en classes, regroupement de modalités nécessaires

Analyse bivariée :

- Visualiser graphiquement les corrélations, valider par les tests statistiques les liaisons perçues
- Visualiser les transformations fonctionnelles de variables nécessaires si relations non linéaires

Analyse multivariée :

- Recherche des facteurs principaux i.e. cartographie massive des liaisons, corrélations linéaires (ACP) ou multiples (ACM), réduction de dimensionnalité

REMARQUES GÉNÉRALES SUR L'ANALYSE EXPLORATOIRE

- L'analyse exploratoire porte du fruit sur les agrégats intermédiaires, pour préparer les dernières étapes du datamanagement,

- L'analyse exploratoire est planifiée, on gagne à classer les variables par thème, mesurer les multi-corrélations, voire aller jusqu'à amener sur les thèmes les moins qualifiants une analyse factorielle partielle pour réduire la dimensionnalité

- L'analyse exploratoire s'oriente déjà en fonction de l'objectif, on ne fera pas les mêmes tests pour
  - Préparer l'évaluation d'un modèle avec des contraintes fortes (type régression)
  - Préparer une segmentation – on cherchera alors des classes de population à séparer, des variables à devoir garder comme illustratives –
  - Chercher un algorithme global pour une base big data
    - Evaluer la raréfaction (sparcity) des données, quelles variables sont trop clairsemées pour être utilisées telles quelles
    - Comment traiter des phénomènes datés, quelle est l'importance du temps etc.)

EXPLORE

Gestion des variables, et individus

- Pourquoi transformer les données
- Comment transformer les données

- Une fois bien connues les données dont nous disposons il importe, afin d'assurer une qualité optimale et une bonne compréhension des modèles établis ultérieurement, d'améliorer la formes de nos données

- Le data management est affaire de compromis.
  - Chaque agrégation des données, chaque mise en classe, chaque variable mise à l'écart, nous prive d'une partie de l'information, il est donc nécessaire d'en évaluer le gain attendu.
  - Les choix établis à cette étape ne peuvent être arbitraire, la consultation du métier est encore capitale

GESTION DES INDIVIDUS

- Comprendre le datamining comme apprentissage statistique impose une vigilance certaine vis-à-vis des individus composant la population d'étude.

- Notre objectif ici est de construire une partition de notre population composée d'une ou plusieurs bases d'apprentissages et de contrôle

- Nous avons vu comment apprécier la cohérence de notre population, détecter des individus atypiques ou aberrant, la problématique du projet et/ou des impératifs métiers sont également à prendre en ligne de compte
  - A l'issue d'une typologie, nous pouvons percevoir des groupes homogènes mais entre eux très différents parmi notre population, et souhaiterons les traiter différemment (ex : images de visages de face, de profil…)
  - Un impératif métier ou une caractéristique du problème peut contraindre à réduire notre base d'étude :
    - Ex : personnes n'ayant qu'un seul don à une ONG / clients n'ayant acheté qu'une fois / opportunistes ayant uniquement profité d'une promotion, clients d'un opérateur passés à la concurrence lors d'un déménagement dans une zone non couverte
    - La définition même d'un individu peut être revue : sur une étude ayant une forte dimension temporelle, ex : fidélité client v.s. ancienneté
  - Sélectionner des contre-exemples représentatifs et non trop nombreux permettant de modéliser correctement les distributions de « bons » et de « mauvais »

GESTION DES VARIABLES

- Bien évidemment un certain nombre de variables peuvent toujours être générées, au cours des agrégats éventuels (ex : tickets de caisse : montants totaux, ou périodiques, paniers moyens, fréquence, ancienneté, récence des achats, volumes, ventilation par gamme…)

- Nos variables d'études peuvent, et doivent, être optimisées pour :
  - améliorer les résultats obtenus
    - en terme de qualité et précision des modèles
    - mais également pour faciliter leur interprétation
  - optimiser les temps de calcul en ne multipliant pas les données qui n'apportent pas d'information supplémentaire

- Une sélection supervisée des variables permettra d'éliminer les multi-corrélations pouvant fausser de nombreux modèles (régression logistique en particulier) ainsi que les variables insuffisamment corrélés à notre objectif dans le cas d'un apprentissage supervisé

EXPLORE

Modélisation

- Apprentissage supervisé ou non supervisé
- Modèles prédictifs génératifs et discriminatifs
- Arbres

Définition :

- On appelle modèle statistique l'approximation du processus d'évolution d'une variable aléatoire, que l'on tente d'inférer à l'aide d'une fonction composite des distributions d'un ensemble de variables explicatives, dotée d'hypothèses sur ces variables.
- La variable à expliquer est souvent appelée également cible ou variable réponse

L'inférence d'un modèle statistique consiste à résoudre un problème d'optimisation consistant soit à maximiser une vraisemblance, soit à minimiser une erreur de prédiction

La qualité (opérationnelle) d'une modélisation tient

- Au respect de ses hypothèses
- A sa complexité, en temps de calcul et petit nombre de variables nécessaires
- A la qualité de la prédiction (dans un cadre métier, le point de vue bayésien sera renforcé par le R.O.I, un bon modèle est un modèle rentable)
    - La précision (ou sensibilité) du modèle est sa capacité à bien prédire / identifier un phénomène
    - La spécificité du modèle est sa capacité à ne pas exagérer sa prédiction et savoir émettre à raison un jugement négatif

Points de vigilance

- Le modèle est un raisonnement abstrait, une vue idéalisée d'un phénomène apprise des données d'apprentissage
- Le modèle statistique se base sur des hypothèses qui peuvent être erronées
- La qualité du modèle repose tout autant sur la qualité de la collecte des données et leur richesse que sur la qualité de la modélisation : Un raisonnement juste sur des données d'apprentissage fausses ou faussées n'aboutira qu'à des prédictions erronnées :
  - Ce principe est souvent formulé « Garbage In, Garbage Out. »

- Pour chaque problématique de datamining un grand nombre de modèles, (i.e. d'algorithmes – méthodologies statistiques ou de « Machine Learning » pouvant proposer un modèle) peuvent être mis en concurrence. Il convient de choisir le meilleur, suffisamment simple, mais adapté aux cas particuliers de la base d'études

La problématique particulière d'un problème non supervisé est l'absence d'une variable « cible » ce qui en fait un problème algorithmique autant que statistique

- La « cible » (une partition de l'univers) n'est connue ni par le nombre de ses classes ni par leurs caractéristiques.
- Autre façon de voir : une ou plusieurs (mises en concurrence) variable cible sont construites au fur et à mesure de l'apprentissage et un critère permet de sélectionner la plus riche en information

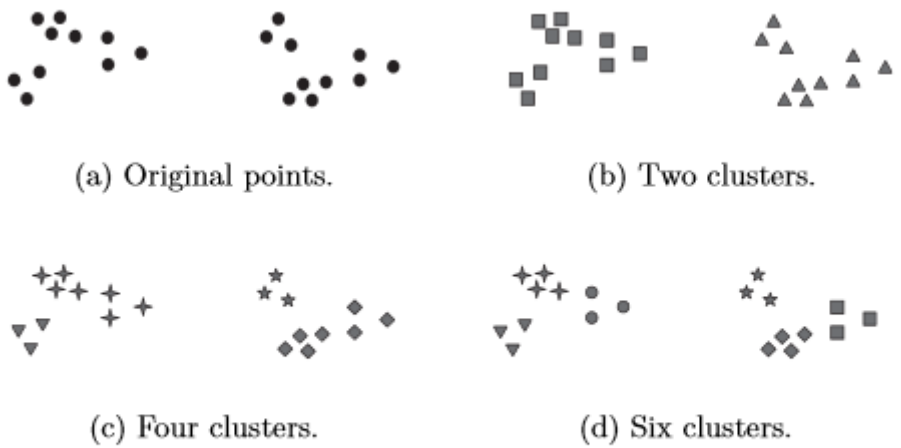


Figure 8.1. Different ways of clustering the same set of points.

Domaines d'application

Domaine	Forme des données	Clusters
Text mining	Textes Mails	Textes proches Dossiers automatiques
Web mining	Textes et images	Pages web proches
BioInformatique	Gènes	Gènes ressemblants
Marketing	Infos clients, produits achetés	Segmentation de la clientèle
Segmentation d'images	Images	Zones homogènes dans l'image
Web log analysis	Clickstream	Profils utilisateurs

Les regroupements établis peuvent être choisis selon :

- Leur sens (dans ce cas on choisir des méthodes, telles les analyses factorielles basées sur les structures de liaisons des données)
- Leur utilité, comme un « résumé » des données, non forcément interprétable, pour remplacer :
  - Un grand nombre de variables à des fins de réduction de dimensionnalité par la seule variable d'affectation
  - Un grand nombre d'individus par les seuls prototypes de classe (i.e. individu le plus proche du centre de classe) pour accélérer un modèle de décision (c'est alors une démarche semblable à un échantillonnage)

La typologie (en : classification) est l'exercice consistant à affecter au mieux un nouvel individu à la segmentation établie

Une segmentation est dite

- Hiérarchique, lorsqu'elle établit des partitions successives de chaque segments et impose de choisir un niveau de découpage des segments pour établir la partition
- Par partitionnement quand elle établit les frontières des segments
- Par modélisation quand elle établit des distributions de probabilités différentes par segment

Exemples : K-means, Classification Ascendante hiérarchique, DBSCAN (basé sur la densité des données), modèles de mélanges de lois normales

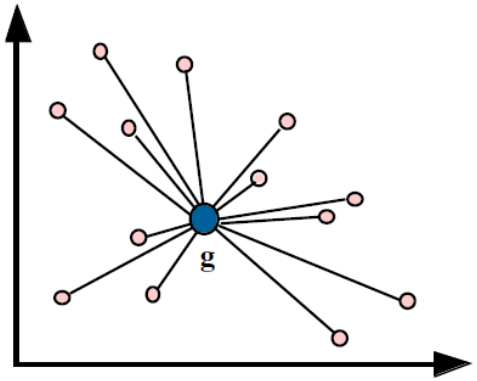
Comment choisir, estimer l'efficacité d'un algorithme de segmentation

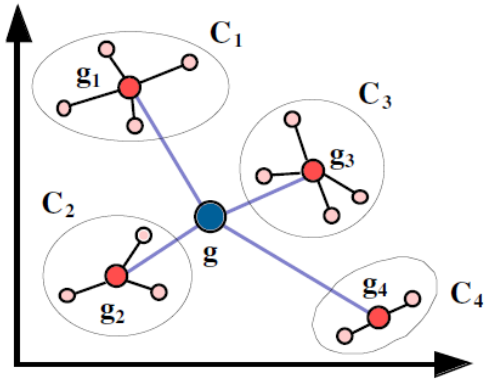
- Quelles sont les données que j'essaie de rassembler et/ou séparer ?
- Comment puis-je définir la similarité ou dissimilarité de deux individus ?
- Comment la méthode construit ses partitions et segments, comment puis-je évaluer leur pertinence ?
- Combien de segments ? (peu si je suis en quête de sens, modérés pour agréger l'information des variables, beaucoup pour une dimension de dimensionnalité sur les individus)
- Comment comparer les résultats de plusieurs algorithmes ?

Notions clefs :

- Un algorithme de segmentation travaille sur une mesure de distance entre les observations à minimiser, ou, au contraire de similarité à maximiser)
- On a déjà défini différentes notions de distance, tant pour des variables continues qu'à valeurs discrètes
- Les segments C_k d'une partition, d'effectif (ou cardinal) N_k sont définis par :
  - Un centre de gravité $\mu_k = \frac{1}{N_k}\sum_{i \in C_k} x_i$ ($\mu = \frac{1}{N}\sum_i x_i$ *étant celui de l'univers)*
  - Une Inertie, qui est une mesure de cohérence du segment, de concentration des points autour du centre de gravité, $\sum_{i \in C_k} D^2\,(x_i, \mu_k)$ D étant la distance choisie pour le problème, *l'inertie inter-cluser déterminant l'éloignement des centres des classes c'est-à-dire la qualité de la séparation* $\sum_k N_k D^2\,(\mu_k, \mu)$
  - Une matrice de Variance/Covariance $\sum_{i \in C_k}(x_i - \mu_k)(x_i - \mu_k)^{\mathrm{T}}$ (Variance/Covariance inter-cluster : $\sum_k(\mu_k - \mu)(\mu_k - \mu)^{\mathrm{T}}$)
  - On définit également une inertie intra-cluster $\sum_k \sum_{i \in C_k} D^2\,(x_i, \mu_k)$

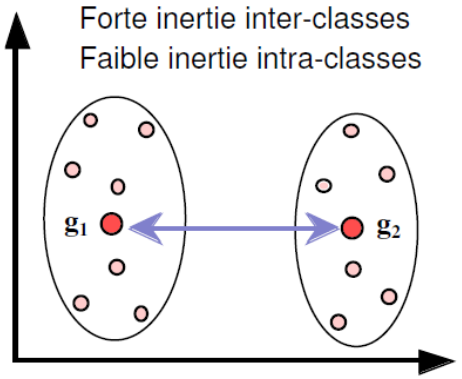
Evaluation de la qualité d'une partition

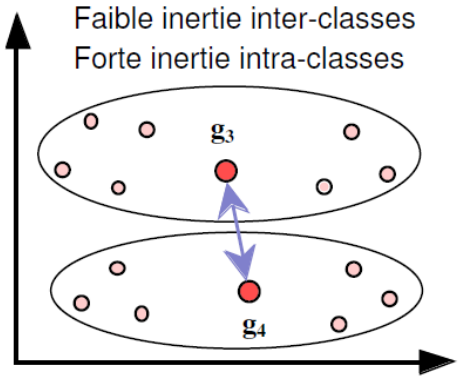




L'inertie totale des points est la somme de l'inertie intra-cluster et de l'inertie inter-cluster

Il faut minimiser l'inertie intra-cluster (classes plus homogènes) et maximiser l'inertie inter-cluster (classes mieux séparées)





Prédiction ou apprentissage ?

- Les modèles dit prédictifs visent à classer des individus, leur attribuer un score, ou une valeur d'un caractère, **non connu**, qui en ce sens peut souvent être attendu dans le **futur**. Par extension, il peut s'agir simplement d'une information existante, que l'on ignore et souhaite inférer.

- La population est séparée en deux sous-populations :
  - Un échantillon d'apprentissage, sur lequel on connait un ensemble de données $x_i \in X$ l'ensemble des variables dites « explicatives » ou espace des entrées, et également une ou plusieurs sorties $y_i \in Y$ ensemble de variables à prédire, à expliquer, ou espace des sorties
  - Une population sur laquelle on ne connait que l'espace des entrées et l'on souhaite inférer Y

- Ne connaissant pas le processus à l'origine des (x_i, y_i) on établit une loi de probabilité conjointe P(X,Y) ou P(Y|X)

- On parle d'apprentissage supervisé car la connaissance des sorties sur l'échantillon d'apprentissage permet aux algorithmes et méthodes que l'on présentera ici de se corriger selon une notion **d'erreur**

Objectif :

- Concevoir une fonction, à partir des n individus de notre échantillon d'apprentissage $d_1^n = ((x_1, y_1); \ldots; (x_n, y_n))$

$$f : X \rightarrow Y$$

$$x_i \longmapsto \hat{y}_i$$

- Résoudre le problème d'optimisation sur les paramètres de f minimisant l'erreur commise sur les valeurs inconnues, les exemples futurs des x_i

On identifie deux grandes catégories de problèmes prédictifs :

- Les régressions, visant à établir la valeur d'une propriété continue ($Y \equiv \mathbb{R}^k$) avec k le nombre de variables à prédire
- La classification visant à affecter les individus aux modalités d'une variable qualitative, $y_i \in \{C_1 ; \dots ; C_k\}$ une partition de l'univers

Un score peut être l'une ou l'autre

- Ex : Régression logistique, dont la variable continue cible est une transformation par la fonction logit de la probabilité d'être « bon » ou « mauvais »
- Ex : Analyse discriminante établissant une frontière entre deux populations

On définit à partir de la notion de distance celle de l'erreur commise, pour chaque individu i, entre l'estimateur $\widehat{y_i}$ et la valeur réalisée y_i.

A partir de l'erreur de la prédiction ou posera aussi une fonction de coût, dite également fonction perte, que l'on souhaitera optimiser, permettant de pénaliser l'erreur

- fonction de coût $l : Y \times Y \rightarrow \mathbb{R}^+$, $(y, y') \mapsto l(y, y')$ ex : perte quadratique $l\big(y, f(x)\big) = (\mathrm{y} - \mathrm{f(x)})^2$

  Dans le cadre d'une classification on parle également de matrice de coût

- Risque moyen ou erreur de généralisation de la règle f : $R_P(f) = \mathbb{E}_{(X,Y)\sim P}[l\big(Y, f(X)\big)]$

La fonction de coût est un élément extrêmement important de la prédiction dans une logique de marché, car elle peut intégrer des éléments factuels de coûts de campagnes marketing, permettant d'évaluer, par l'estimation du coût de la prédiction un R.O.I.

- Exemple : Lors d'une campagne de prospection, le coût d'un faux positif est celui de l'envoi de la campagne, le coût d'un faux négatif est tout le potentiel de revenu du client qui ne sera pas contacté.

La matrice de coût sera a priori asymétrique car toutes les erreurs ne sont pas aussi pénalisantes

Risque théorique

$$R_P(f) = \mathbb{E}_{(X,Y)\sim P}\big[l\big(Y, f(X)\big)\big] = \iint_{X,Y} l\big(y, f(x)\big)P(x,y)dxdy$$

Ou dans le cas d'une classification $\int_X \sum_{y\in Y} l\big(y, f(x)\big)P(x,y)dx$

Exemples :

- Classifieur optimal de Bayes. Par Bayes $P(x,y) = P_r(y|x)P_X(x)$

$\Rightarrow$f minimise R_P : $x \rightarrow argmin_{y'\in Y} \sum_{y\in Y} l(y,y')\, P_r(y|x)$ (on affecte x à la classe minimisant son risque conditionnel à y' plutôt que y)

$\Rightarrow$ avec un coût d'erreur unitaire on parle du classifieur optimal de Bayes $f(x) = argmin_{y'\in Y} P_r(y'|x)$

- Cas de la régression réelle, fonctions de régressions Espérance et Médiane

Avec l(y,y')=(y-y')² (dist.euclidienne / moindres carrés) $\inf_{f\in\mathcal{F}} R_P(f) = R_P(\mathbb{E}[Y|X=x])$

Avec l(y,y')=|y-y'|(erreur absolue) $\inf_{f\in\mathcal{F}} R_P(f) = R_P(mediane[Y|X=x])$

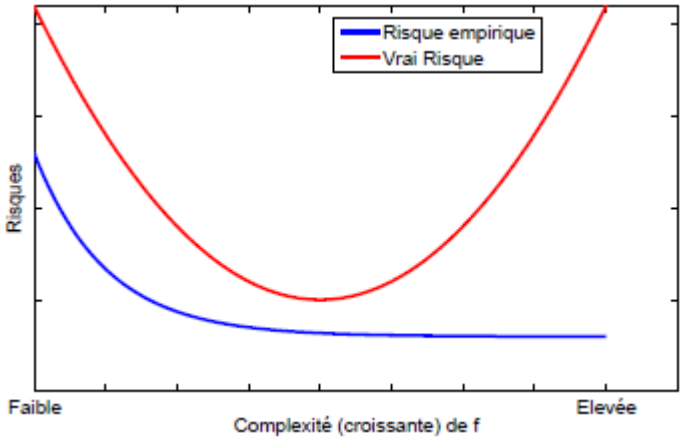
Risque Empirique

Nous ne disposons que de notre échantillon d'apprentissage $d_1^n = \{(x_i, y_i), 1 \leq i \leq n\}$

- Le risque empirique, à supposer que toutes nos observations sont également probables (i.e. $P(x_i, y_i) = \frac{1}{n}$) est évidemment $\frac{1}{n}\sum_{i=1}^{n} l(Y_i, f(X_i))$

- Avec l'augmentation de la dimensionnalité, de la complexité du modèle (nombre de variables croissants, prises en compte des croisements de variables de plus en plus complètes etc.) les modèles non paramétriques basés sur ce risque empirique montrent un risque de surapprentissage

- Un compromis sera fait entre simplicité du modèle et qualité de la prédiction sur l'échantillon d'apprentissage (capacité de généralisation du modèle)
- Pour éviter ce risque ou mesurera le vrai risque, par exemple par validation sur un second échantillon non utilisé dans la minimisation



On dispose de règles de prédiction optimales que l'on peut formuler à condition de connaître la loi P qui en pratique, est le plus souvent inconnu. Seules des observations de la loi sont connues.

La démarche prédictive consiste à déterminer un algorithme prédictif qui ne dépende pas de P mais de l'ensemble d'apprentissage

Définition

Un Algorithme de prédiction est représenté par une application (mesurable)

$\widehat{f}: (X, Y)^n \rightarrow F$ qui à un ensemble d'apprentissage $d_1^n = \{(x_i, y_i), 1 \leq i \leq n\}$ associe une règle de prédiction $\hat{f}(d_1^n)$ ou par une suite $\widehat{f_n}; n \geq 1$ d'applications mesurables telles que pour $n \geq 1, \widehat{f_n} : (X, Y)^n \rightarrow F$

Cycle et raffinement

> Méthodologie pratique

EXPLORE

CRoss Industry Standard Process for Data Mining

Compréhension de la problématique opérationnelle :

Définir les objectifs et les prérequis en termes business et les traduire en problématique technique Datamining.

Compréhension des données

Collecter des données, audit de qualité

Préparation des données

Construction des tables d'analyse

Data Management

Modélisation

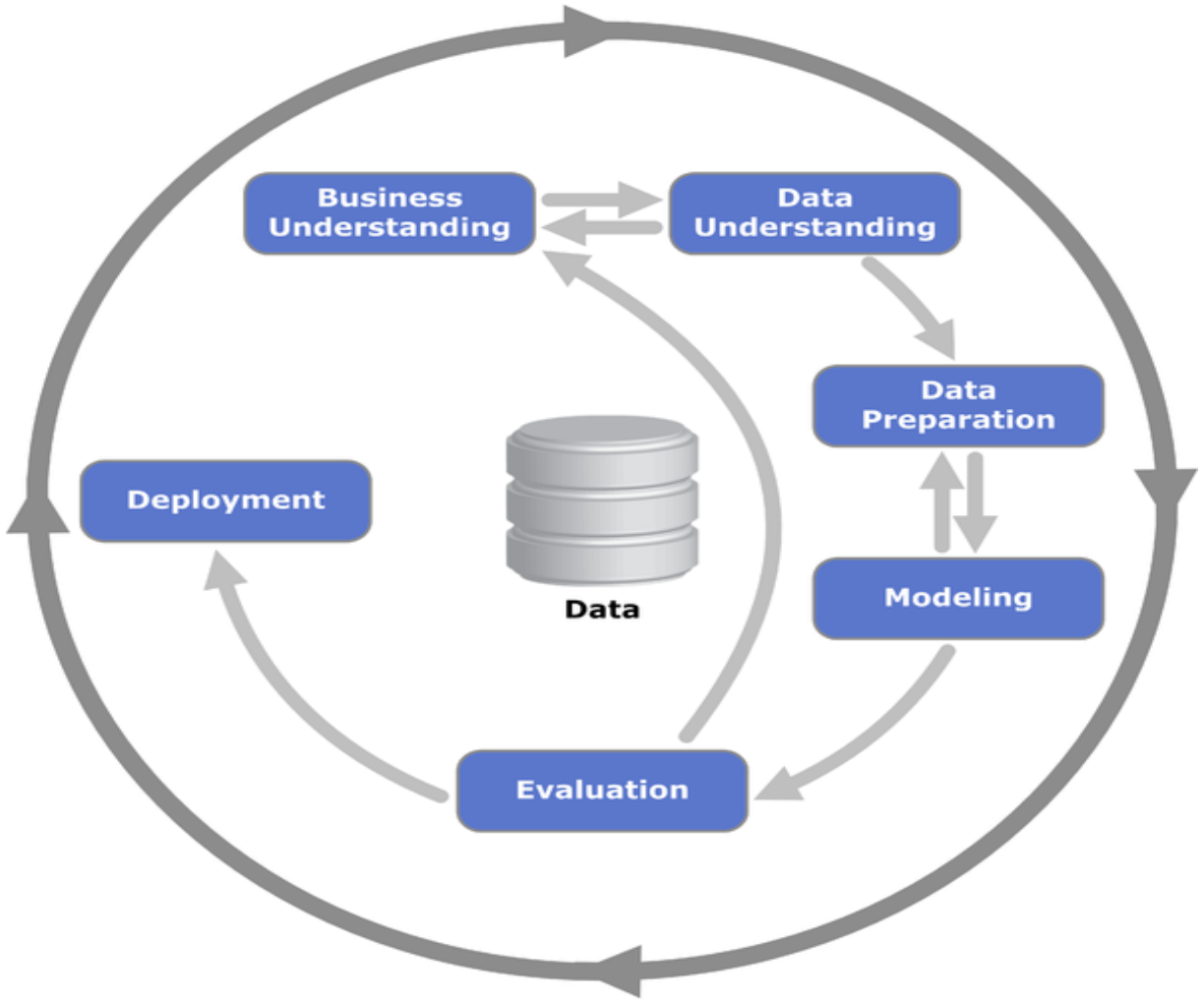
Sélection du meilleur modèle sur des critères statistiques

Evaluation

Evaluation de la pertinence du modèle par rapports aux objectifs business initiaux

Déploiement

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Mise en production et vie d'un modèle

- Logiciels d'analyse classiques
- Application d'un modèle en temps réel

APPLIQUER UN MODÈLE EN TEMPS RÉEL

Les outils de datamining tels que SPAD, SAS, ou SPSS permettent tous d'archiver un modèle et de le réappliquer à de nouveaux individus sur lesquels X est connu et Y non.

Ces outils nécessitent une intervention humaine : charger les nouvelles données, appliquer le modèle, exporter le résultat :

- Tout à fait adapté à des mises à jours massives de scores dans une base de données
- ils ne peuvent être intégrés dans un processus automatisé (ex : recommandation en temps réel d'un produit par un moteur de recherche, par transmission de la navigation du client sur un site de e-commerce, par web-service, en temps réel pour mettre à jour des scores de préférences)

Petit à petit les éditeurs développent des serveurs de déploiements, interrogables via une API

L'alternative d'un outil de script (R, Python) est généralement encore privilégiée avec le bémol de la performance

Les framework big data tels que hadoop doivent permettre d'intégrer des API d'analyse temps réel des big data.

Il demeure que la mise à jour d'un modèle d'apprentissage demande parfois des calculs matriciels massifs peu adaptés au temps réel => bien distinguer apprentissage et application du modèle dans un processus et les choix d'architecture.

MÉTHODES D'EXPLOITATION ET MISE EN PRODUCTION

EXPLORE

1) Mise en production d'un modèle

2) Logiciels d'analyse / temps différé

3) Appliquer un modèle en temps réel

4) Containers