TECHNIQUES DE BASE D'ANALYSE DES DONNÉES

PRÉPARATION DU TERRAIN

APERÇU

1. Contexte et processus
2. Extraire de l'information de grands volumes de données : quelques concepts de base
3. Extraire de l'information de grands volumes de données : quelques techniques de base
4. La semaine prochaine : un contexte différent

CE QUE NOUS AVONS ABORDÉ JUSQU'À PRÉSENT…

- Modélisation des données et analyse conceptuelle
- Collecte de données
- Transformation des données
- Stockage des données
- Exploration des données
- Présentation des données

CE QUE NOUS AVONS ABORDÉ JUSQU'À PRÉSENT…

- Modélisation des données et analyse conceptuelle
- Collecte de données
- Transformation des données
- Stockage des données
- Exploration des données
- Présentation des données

Aujourd'hui
Intégration du tout dans le contexte de l'informatique décisionnelle et de l'analyse de données connexe.

VEILLE OPÉRATIONNELLE – ANALYSE DES ACTIVITÉS

Utiliser les données (et l'information) sur les opérations internes et l'état du marché pour appuyer la **prise de décisions éclairées** sur les **opérations** et la **stratégie commerciale**.

Il n'existe aucune définition fermement établie de ces termes – l'un est-il un sous-ensemble de l'autre?

Objectifs : meilleure connaissance de la situation + meilleure prévoyance

HISTOIRE DE L'INFORMATIQUE DÉCISIONNELLE

Fin des années 1800 : les gens ont commencé à reconnaître qu'ils pouvaient utiliser les données pour obtenir un avantage concurrentiel.

Années 1950 : avènement de la première base de données d'affaires aux fins d'aide à la décision

Années 1980-1990 : les ordinateurs et les données deviennent de plus en plus accessibles – entrepôts de données, exploration de données – la discipline est encore très technique et spécialisée

Années 2000 : tentative de faire passer l'analyse des activités des mains des explorateurs de données et d'autres spécialistes aux mains d'experts du domaine

Aujourd'hui : des données et des techniques spécialisées importantes sont arrivées sur le marché, ainsi que la visualisation des données, les tableaux de bord et les logiciels en tant que service.

1865

Années 1950

Années 1980-1990

Années 2000

2019

INFORMATIQUE DÉCISIONNELLE ET SCIENCE DES DONNÉES

Sur le plan historique, l'un des volets contribuant à la science des données moderne

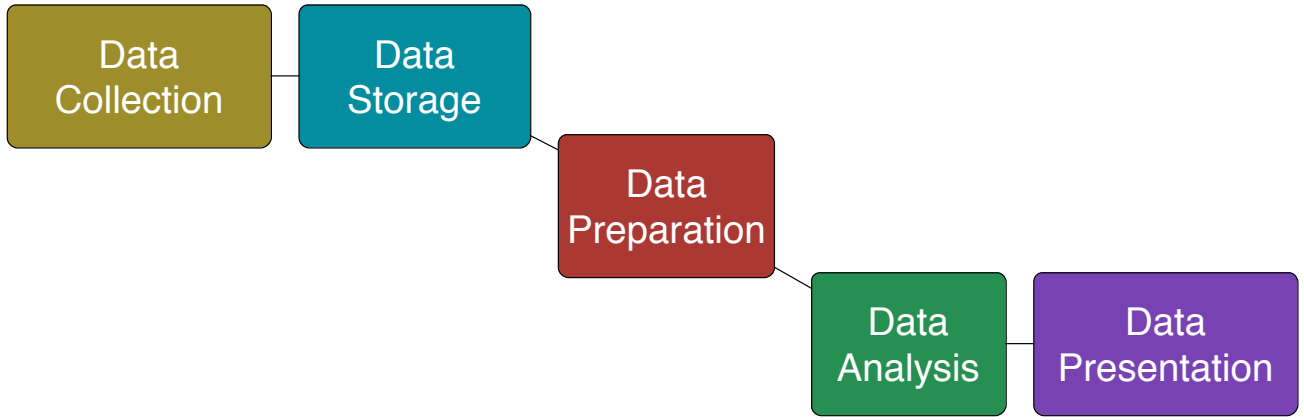
- **Système visé** : le domaine commercial – le marché auquel vous participez
- **Sources des données :** données de transaction, données financières, données de ventes, données organisationnelles.
- **Objectifs :** mieux comprendre les concurrents, les consommateurs et les activités internes et se service de ces connaissances pour appuyer la prise de décisions.
- **Culture et techniques préférées :** tableaux de données, indicateurs clés de performance, comportement des consommateurs, découpage en tranches et en dés, « observations » commerciales.

Le but ultime est toujours le même : mieux comprendre le système visé.

INFORMATIQUE DÉCISIONNELLE ET PIPELINE DE DONNÉES

Notre modèle général de pipeline de données fonctionne également pour l'informatique décisionnelle.

Quels sont certains des aspects du pipeline d'informatique décisionnelle qui pourraient le distinguer d'un pipeline d'analyse plus générique?

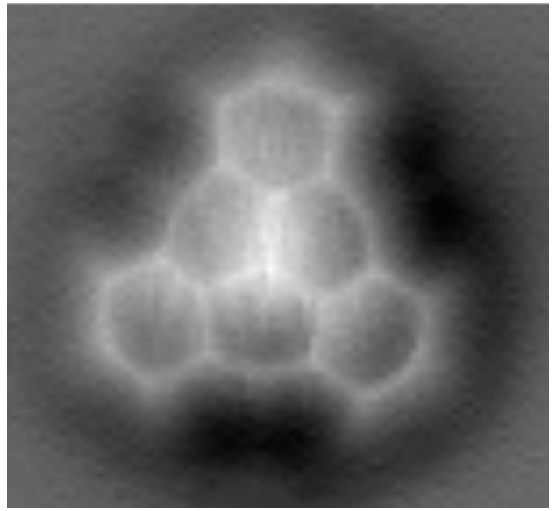


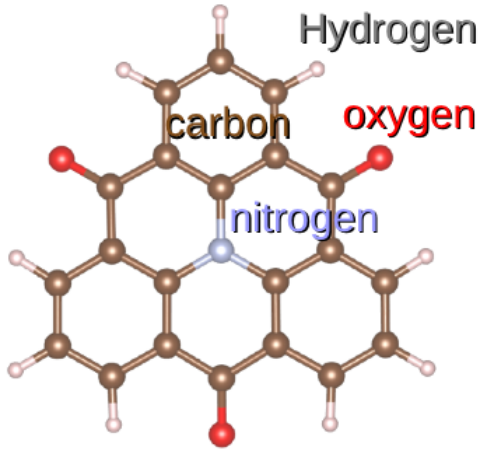
TROUVER DES TENDANCES, DES GÉNÉRALISATIONS ET UNE STRUCTURE

- **Tendance :** Régularité prévisible et répétitive
- **Structure :** Organisation d'éléments dans un système
- **Généralisation** : Création de concepts généraux ou abstraits à partir de concepts ou de cas spécifiques.

Objectif sous-jacent pendant l'analyse – trouver des tendances ou des structures dans nos données, tirer des conclusions à l'aide de ces tendances ou de ces structures.

Le fait de trouver des tendances et des structures n'a pas de valeur en soi, c'est la façon d'utiliser ces découvertes – les conclusions qui en sont tirées – qui est importante.





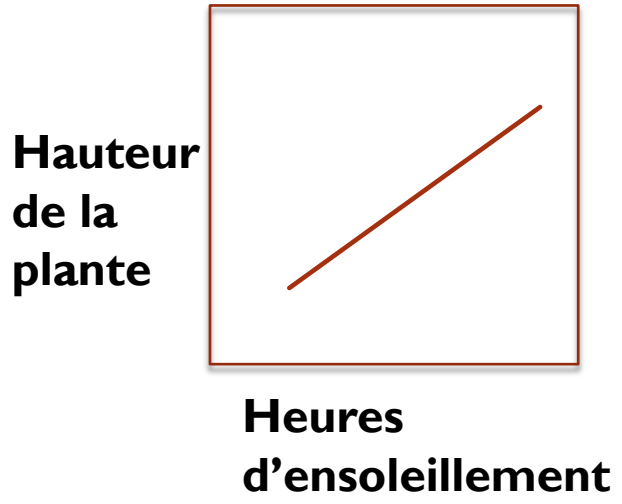
https://upload.wikimedia.org/wikipedia/commons/d/d0/TOAT_AFM.png

VARIABLES INDÉPENDANTES ET VARIABLES DÉPENDANTES

Dans un cadre expérimental

- **Variables contrôlées/parasites :** Nous faisons de notre mieux pour garder ces variables contrôlées et immuables pendant que d'autres variables sont modifiées.
- **Indépendante :** Nous contrôlons les valeurs de la variable. Nous soupçonnons qu'elles influencent les variables dépendantes.
- **Dépendante :** Nous ne contrôlons pas les valeurs – elles sont générées d'une manière ou d'une autre pendant l'expérience, et dépendent probablement de tout.

Comment ces concepts s'appliquent-ils à d'autres ensembles de données?



TYPES DE DONNÉES

Données numériques : nombres entiers ou continus

- 1, 7, 34,654, 0,000004

Données textuelles : chaînes de texte – peuvent être limitées à un certain nombre de caractères

- « Bienvenue au parc », « AAAAA », « 345 », « 45,678 »

Données nominales : un nombre fixe de valeurs, qui peuvent être numériques ou représentées par du texte. **Il n'y a pas d'ordre particulier ou inhérent** .

- (« rouge », « bleu », « vert"), (« 1 », « 2 », « 3 »)

Données ordinales : Données nominales avec un ordre inhérent. Contrairement aux données en entiers, l'espacement entre les valeurs **n'est pas** défini.

- (très froid, froid, tiède, chaud, très chaud)

TRANSFORMATION DES DONNÉES NOMINALES EN DONNÉES NUMÉRIQUES (DÉNOMBREMENT)

Nous pouvons transformer des données nominales en données numériques en dénombrant la fréquence des différentes valeurs de la variable catégorielle.

Ceci nous permet d'appliquer des techniques d'analyse numérique.

Couleur de la maison	Fréquence
rouge	40
bleu	13
vert	2

RÔLE PARTICULIER DES DONNÉES NOMINALES

Les données nominales jouent un rôle particulier :

- En *science des données*, on parle d'une variable nominale avec un ensemble prédéfini de valeurs.
- En *science expérimentale*, un facteur est une variable indépendante dont les niveaux sont définis – il peut aussi être considéré comme une catégorie de traitement.
- Dans l'*analyse des activités*, il est question de dimensions (qui ont des membres) par opposition aux mesures.

Quelle que soit la façon dont nous étiquetons ces types de variables, nous pouvons les utiliser pour **créer un sous-ensemble** de nos données ou pour **les compiler ou les résumer**.

DONNÉES HIÉRARCHIQUES/IMBRIQUÉES/MULTINIVEAU/MODÈLES

Si une variable nominale a plusieurs niveaux d'abstraction, nous pouvons créer des niveaux à partir de cette variable.

Nous pouvons, dans un sens, considérer ces niveaux comme de nouvelles variables nominales.

La « nouvelle » variable nominale a une relation prédéfinie avec le niveau plus détaillé.

Cette situation est courante avec les variables de temps et d'espace – nous pouvons faire un zoom avant ou arrière.

Cela nous permet de parler de la **granularité** des données – quel est le « zoom maximum »?

Année	Trimestre	Dénombrement
2012	1	34
2012	2	12
2012	3	52
2012	4	0
2013	1	21
2013	2	9
2013	3	112
2103	4	8

RÉCAPITULATION DES DONNÉES

Min : Plus petite valeur de la variable

Max : Plus grande valeur de la variable

Médiane : Valeur moyenne de la variable

Mode : Valeur la plus fréquente

Valeurs uniques : Liste des valeurs uniques

Évaluation	Type
4,31	Bleu
5,34	Orange
3,79	Bleu
5,19	Bleu
4,93	Vert
5,76	Orange
3,25	Orange
7,12	Orange
2,85	Bleu

COMPILER LES DONNÉES

Nous pouvons effectuer une opération sur un ensemble (ou sous-ensemble) de données – généralement sur une colonne de données.

Lorsque nous le faisons, nous pouvons considérer cela comme la compression ou la « compilation » des nombreuses valeurs de données en une seule valeur représentative.

Les fonctions typiques de compilation sont « moyenne », « somme » et « dénombrement ».

Si nous appliquons la même fonction de compilation à de nombreuses colonnes différentes, nous pouvons considérer cela comme une **mise en correspondance** (d'une liste) de colonnes par rapport aux fonctions.

Évaluation	Type
4,31	Bleu
5,34	Orange
3,79	Bleu
5,19	Bleu
4,93	Vert
5,76	Orange
3,25	Orange
7,12	Orange
2,85	Bleu

TABLEAUX DE CONTINGENCE/TABLEAUX CROISÉS DYNAMIQUES

Tableau de contingence : Tableau utilisé pour examiner la relation entre deux variables nominales – en particulier la fréquence d'une variable par rapport à une deuxième variable (tableau croisé).

Tableau croisé dynamique : Tableau généré dans une application logicielle par l'application d'opérations (p. ex. somme, dénombrement, moyenne) à des variables, éventuellement basées sur une autre variable (nominale). Peut être utilisé pour créer un tableau de contingence.

	Grand	Milieu	Petit
Bleu	1	32	31
Orange	14	11	0
Vert	5	5	5

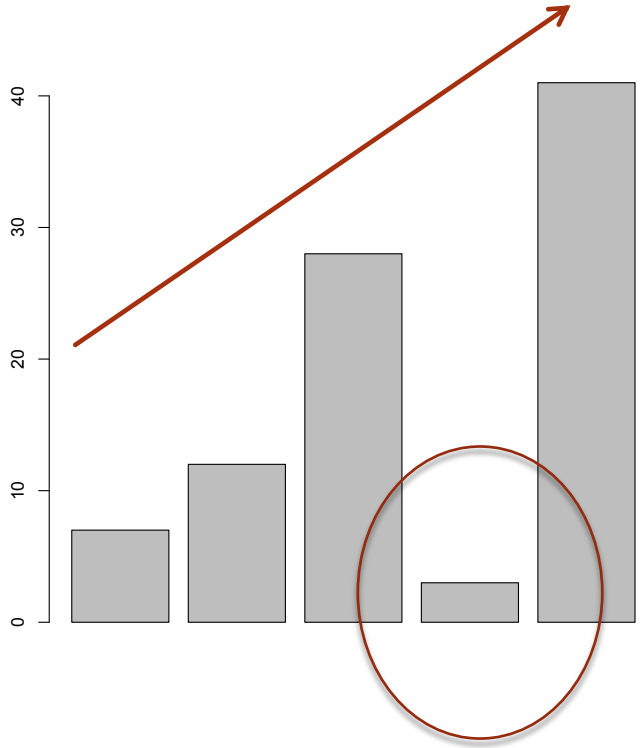
ANALYSE AU MOYEN DE LA VISUALISATION

Analyse au sens large :

- identification de modèles ou de structures
- ajout de sens à ces modèles ou à ces structures en les **interprétant** dans le contexte de votre système.

Option 1 : utiliser des techniques d'analyse pour y parvenir.

Option 2 : visualiser les données et utiliser la capacité analytique de notre cerveau (perception) pour arriver à des conclusions significatives sur ces modèles.



QUELQUES VISUALISATIONS SIMPLES POUR RÉVÉLER DES TENDANCES

Diagramme de dispersion : bien adapté à deux variables numériques

Diagramme linéaire : variable numérique et variable nominale

Diagramme à barres : bien adapté à une donnée nominale et une donnée numérique – ou plusieurs données nominales ou des données nominales imbriquées