

# BASES DU TRAITEMENT DE L'INFORMATION EN SANTÉ

## I. DONNÉES, INFORMATION & CONNAISSANCE

<b>DONNÉE</b>	<ul style="list-style-type: none"> <li>• <b>Description élémentaire d'une réalité</b> qui résulte d'une <b>observation</b> ou d'une <b>mesure</b>.</li> <li>• Notion <b>abstraite, typée</b> (<u>ex</u> : données numériques, symboliques, textuelles, logiques,...).</li> <li>• Ne porte <b>pas de sens</b> en elle-même.</li> <li>• Lorsqu'on range des données dans une base de données, peu importe leur signification.</li> <li>• La majorité des traitements réalisés par les informaticiens concernent des données dont le sens porté par leurs valeurs n'est <b>pas déterminant</b> au sein du traitement.</li> </ul>
<b>INFORMATION</b>	<ul style="list-style-type: none"> <li>• Ce qui donne une forme à l'esprit.</li> <li>• Notion <b>abstraite</b>, mais d'un niveau d'abstraction supérieur à celui de la donnée.</li> <li>• <u>Information</u> = une <b>donnée</b> + un <b>sens</b> = ensemble intelligible de données qui prend un sens (on distingue une définition <u>objective</u> et une définition <u>subjective</u> de l'information).</li> <li>• Comparer 2 informations est <b>plus complexe</b> que comparer 2 données.</li> <li>• Une température mesurée par un thermomètre est une <b>donnée</b>. Son expression dans un référentiel d'unité (°C) est une <b>première information</b>. Si on ajoute l'heure et le lieu de la mesure, on <b>enrichit l'information</b>.</li> </ul>
<b>CONNAISSANCE</b>	<ul style="list-style-type: none"> <li>• Notion <b>abstraite</b>, d'un niveau d'abstraction supérieur à celui de l'information.</li> <li>• <b>Partagée</b>, à la différence de l'information, et s'appuie sur un <b>référentiel collectif</b>.</li> <li>• Un faisceau d'informations permet de <b>constituer</b>, de <b>reconstituer</b> ou <b>d'enrichir une connaissance</b> sur un sujet. L'<u>information</u> n'est seulement <b>que le vecteur</b> pour accéder à la connaissance (<i>tout comme le document est celui de l'information</i>).</li> <li>• Des informations peuvent être communiquées sans pour autant devenir des connaissances. Il faut alors les accompagner de leur référentiel puisque celui-ci ne sera pas partagé (non-implicite).</li> <li>• <b>CONNAISSANCE TACITE</b> = celle que possèdent les individus, <b>pas formalisée</b> et <b>difficilement transmissible</b> (<u>ex</u> : compétences, expériences, intuition,...).</li> <li>• <b>CONNAISSANCE EXPLICITE</b> = <b>formalisée</b> et <b>transmissible</b> sous forme de documents réutilisables (<u>ex</u> : informations concernant les processus, projets, clients,...).</li> </ul>

## II. TRAITEMENT DE L'INFORMATION

### 1. SCHÉMA GÉNÉRAL

Le traitement de l'information représente la **façon dont on perçoit et assimile une information**.

Traiter l'information c'est = la **capturer** (ou la **créer**), la **diffuser** (ou la communiquer en même temps), la **conserver**, la **restituer** (puis la communiquer).

### 2. TECHNOLOGIES TRADITIONNELLES

Durant l'Antiquité et le 18<sup>ème</sup> siècle :

- **Pas de compréhension ou d'analyse** de la nature des phénomènes.
- Expression d'une **perception** (non-reproductible).
- Outils de création **manufacturés** et supports de conservation **naturels**.

### 3. TECHNOLOGIES ANALOGIQUES

Au 19<sup>ème</sup>-20<sup>ème</sup> siècle :

- **Compréhension** de la nature physique des phénomènes.
- L'acquisition des phénomènes, des appareils et des instruments de mesure est représentée par une information exprimant la variation d'une grandeur physique.
- Un **transducteur analogique** est un dispositif matériel permettant la conservation (par analogie) d'un phénomène physique en un autre, en vue de sa diffusion ou de son stockage.

### 4. TECHNOLOGIES NUMÉRIQUES

Depuis l'apparition de l'informatique, l'homme utilise des **codes informatiques** pour représenter l'information, stocker cette représentation et la traiter.

L'aspect du traitement de l'information devient un facteur très important dans le domaine de l'intelligence artificielle et pour les logiciels de modélisation qui essaient d'encourager certains types de raisonnement ou d'exploration.

## III. TRAITEMENT NUMERIQUE

### 1. INFORMATION NUMÉRIQUE

Sur une machine, **toute l'information se trouve sous forme numérique**, que ce soit dans la mémoire de masse (stockage), dans la mémoire vive, dans le microprocesseur et au niveau de tous les périphériques, et notamment ceux de communication.

Les informations rencontrées sur les machines sont de **natures différentes** : *principalement du texte, des images, du son, des vidéos, ... et des programmes.*

Chaque type d'information fait l'objet de **standards de codage**, selon sa **nature** ou sa **destination** (*stockage, utilisation, communication, ...*).

## 2. NUMÉRISATION DE L'INFORMATION

Certaines informations, portées par des grandeurs physiques, sont constituées de **signaux analogiques** : une grandeur analogique peut prendre dans un intervalle fini donné une infinité de valeurs (ce qui est aussi le cas du temps dans lequel ces grandeurs évoluent).

La **numérisation d'un signal analogique** peut se faire **par échantillonnage** :

- On découpe le temps en intervalles réguliers → **fréquence d'échantillonnage** (Hz)
- A chaque période d'échantillonnage, on mesure l'amplitude du signal et on la convertit en un nombre entier (on parle de quantification) → **résolution** (nombre de bits)
- Plus la **fréquence d'échantillonnage** et la **résolution** sont **élevées**, plus la **numérisation** est **fidèle**  
+++++

## 3. CODAGE DE L'INFORMATION

- **Quantum d'information** :
  - élément binaire (0 ou 1) → Bit (*Binary digit*)
  - 1 octet = 8 bits
  - Mot machine (8, 16, 32, 64 bits)
- **Numération binaire** (base 2) → calculs numériques
- **Logique** : Vrai=1 et Faux=0, raisonnement
- n bits permettent de coder  $2^n$  objets (ex : 8 bits = 256 objets)

**POINT TUT** : En informatique, un « mot » est une **unité élémentaire de mémoire**. Comme par exemple, un **octet** est un « mot » de **8 bits**. Donc, en programmation, on va représenter des nombres avec des « mots » de longueur variable, entre autres si on veut stocker de grand nombres ou pas. (Merci Chiron pour ces précisions, quel boss décidément !)

## 4. DONNÉES TEXTUELLES

- **Caractères** :
  - 26 lettres + blanc = 5 bits
  - 26 lettres + 10 chiffres = 6 bits
  - Majuscules + minuscules + chiffres = 7 bits
  - Caractères spéciaux = 8 bits = code ASCII
- **Unicode sur 16 bits** (63 536 objets, tous les alphabets + idéogrammes)

## 5. DONNÉES CHIFFRÉES

- **Nombres entiers** :
  - Naturels : 8 bits (de 0 à 255). D'une manière générale, un codage sur n bits permet de représenter des nombres entiers naturels compris entre 0 et  $2^n - 1$ .
  - Relatifs : 16 bits (-32 768 à +32 767). D'une manière générale, le plus grand entier relatif positif codé sur n bits sera  $2^{n-1} - 1$ .
- **Nombres réels** : Mantisse Exposant, 32 bits
  - Le signe est représenté par 1 seul bit : le bit de poids fort (celui le plus à gauche)
  - L'exposant est codé sur les 8 bits consécutifs au signe



- **Possibilité de compression** : de facteur 4 (sans perte) à 10 (avec perte d'information).

Depuis les années 70, **3 nouvelles techniques**, basées sur le traitement informatique, ont **bouleversé l'imagerie médicale** : la tomodensitométrie (scanner), l'angiographie numérisée, l'imagerie par résonance magnétique nucléaire (IRM).

La diffusion croissante des systèmes informatiques a bénéficié également à la scintigraphie, à l'échographie, à l'endoscopie vidéo et à la radiologie conventionnelle. Plus récemment est apparu le concept de **système informatique dédié à l'imagerie**.

**L'interprétation automatique des images**, comme aide au diagnostic, est **complexe** et reste du domaine de la **recherche**. Elle fait appel à de nombreuses techniques, notamment de reconnaissance des formes et d'intelligence artificielle, et combine des informations de natures diverses.

### Exemples d'application :

L'**évaluation de la contractilité myocardique** peut s'effectuer à partir de la numérisation d'images échocardiographiques.

L'**identification de chromosomes** pour la détermination du caryotype, repose sur des techniques de reconnaissance des formes. L'image de la préparation cytologique est numérisée. L'analyse des variations de densité permet de repérer les chromosomes.

La **transmission d'images par réseau** pour consultation par un expert permet des applications de **télédiagnostic** ou de **télésurveillance**, notamment en radiologie ou cytopathologie.

La **reconstruction en 3D d'images des organes** montre les rapports des structures entre elles (organes, tumeurs, structures vasculaires). Particulièrement employée dans le domaine de la **neurologie**.

La **chirurgie assistée par ordinateur** associe aux phases d'acquisition et d'interprétation d'images, 2 étapes de **raisonnement** et de **commande robotique**. L'objectif est de faciliter la réalisation de gestes médico-chirurgicaux complexes. A partir d'images reconstruites, souvent à partir de plusieurs sources, le raisonnement constitue un modèle du patient et permet de simuler l'intervention (geste virtuel). La commande robotique peut prendre la forme d'une aide **passive** (détection d'écarts au geste prévu), **semi-active** (système de contraintes) voire **active** (autonomie du robot).

---

## 8. DONNÉES SIGNAL : SIGNAUX PHYSIOLOGIQUES

**Signaux numériques** : la taille mémoire dépend de la fréquence d'échantillonnage.

Le **signal électrique analogique** produit par un capteur est un signal **continu**, variant en fonction du temps, à 2 dimensions, sa fréquence et son intensité.

Il doit être mis sous forme **binaire** pour être manipulable par un ordinateur, c'est l'opération de **conversion analogique - digitale** (ou **numérisation**) qui procède en trois étapes :

- le signal est d'abord découpé en segments de durées égales, c'est l'**échantillonnage**
- la hauteur de chaque segment est alors **quantifiée** (en prenant une valeur moyenne) ;
- cette valeur est ensuite **codée sous forme numérique** ; plus la longueur du mot binaire utilisé pour représenter la hauteur est grande, plus on peut définir de niveaux différents d'intensité du signal et donc plus la précision sera importante (1 bit ne permet de coder que deux niveaux et correspond à un signal en tout ou rien (*par exemple, froid ou chaud*), 2 bits autorisent 4 niveaux possibles alors qu'un octet (8 bits) correspond à 256 ( $2^8$ ) niveaux différents possibles).

La **séquence de traitement** comporte 4 phases :

- **acquisition du signal analogique** par un capteur et **numérisation** par un convertisseur analogique-digital ;
- **pré traitement simple** visant à l'amélioration de la qualité du signal (extraction du signal sur le bruit, amplification, filtration) ;
- **traitement analytique** permettant l'extraction de paramètres, *par exemple les complexes QRS d'un ECG*, le plus souvent par des méthodes mathématiques ;
- **interprétation** des résultats.

## IV. GESTION INFORMATIQUE DES DONNEES

### 1. GESTION INFORMATIQUE DES DONNÉES

Une **structure de données** correspond à une manière d'organiser et de représenter les données. Les deux types de renseignements contenus dans une structure de données sont les **données proprement dites** et les **liens qui peuvent exister entre elles**, formalisés par leur organisation. L'organisation de ces données en informatique est essentiellement celui de leur **stockage** et de leur **accès** sur une mémoire secondaire. Deux classes de systèmes peuvent être utilisées : les **fichiers** et les **bases de données**.

### 2. FICHIERS

**FICHIER** = ensemble de données organisées en vue d'une application déterminée. *Un fichier informatique peut contenir un programme, du texte libre ou des données.*

Les **fichiers de données** contiennent des informations de même nature (*un fichier est un ensemble de fiches de même type*) et surtout disposent d'une structure interne (*qui dit à quel endroit ce trouve tel type d'information*). Cette structure, ensemble de relations entre les différents éléments, permet l'exploitation des informations.

Les entités auxquelles on s'intéresse sont décrites par un certain nombre de **caractéristiques**, analogues pour tous les éléments d'un fichier, les entités se distinguant par les valeurs qui sont affectées à ces caractéristiques. *Par exemple, des malades seront tous décrits par leur nom, leur prénom, etc... et seules changent les valeurs de ces caractéristiques pour chaque individu.*

On appelle **enregistrement**, **article** ou **fiche**, l'ensemble des informations décrivant une entité. Les caractéristiques ou attributs sont appelés rubriques ou champs et peuvent recevoir des valeurs, appelées occurrences d'enregistrement ou réalisations.

Afin d'optimiser la gestion informatique des rubriques, les champs sont généralement définis par leur nom, le type de donnée qu'ils vont contenir (texte, nombre, date voire image) et leur taille maximale.

### 3. ACCÈS AUX DONNÉES

• **Accès SÉQUENTIEL** : Soit un fichier de malades enregistré sur une bande magnétique. Les informations (fiches et rubriques) sont écrites les unes à la suite des autres : nom1-prénom1-age1 — nom2-prénom2-age2 — nom3... La recherche d'un malade par son nom ne peut se faire qu'en **lisant séquentiellement tous les enregistrements le précédant**, ce qui peut être très long s'il y a beaucoup d'enregistrements.

• **Accès DIRECT** : Chaque enregistrement a une adresse formée d'un numéro de piste et d'un numéro de secteur. On peut positionner directement la tête de lecture sur la piste et **lire séquentiellement le secteur** sans être obligé de lire tous les enregistrements des pistes précédentes. On parle alors d'**organisation directe** et d'**accès direct**.

Pour **accéder à un enregistrement**, le problème est qu'on ne connaît pas toujours le numéro d'ordre de l'individu recherché, à moins d'avoir la liste complète et à jour des enregistrements.

Un **index** est une table de correspondance indiquant en face de la valeur du critère de recherche de chaque enregistrement (*par exemple, le nom*) le numéro d'ordre de cet enregistrement. L'index peut être **unique** ou **associé à d'autres index** (*on parle d'index primaire ou maître et d'index secondaires*), afin de permettre un **accès rapide** sur d'autres clés. La **gestion de l'index** est normalement assurée par le logiciel de gestion de données. La **clé d'index** permet **d'identifier de façon unique un enregistrement**. La clé d'index peut être **simple** ou **composée de plusieurs critères** (*par exemple, nom et prénom*) afin d'être **plus discriminante**.

• **Gestion informatique de fiches :**

- déclarer ou redéfinir la structure des enregistrements, c'est-à-dire le nom, le type et la taille des diverses rubriques ;
- saisir, modifier, ajouter des données ou les supprimer ;
- déclarer des clés d'index ou de trier le fichier ;
- retrouver des données répondant à des critères plus ou moins complexes ;
- éditer ou d'imprimer le fichier, en totalité ou partie, sous une présentation variable ;
- créer des masques facilitant la saisie à l'écran.

#### 4. BASES DE DONNÉES

La solution générale consiste à **organiser les fichiers en bases de données** qui regroupent de **grands ensembles de données interdépendantes**, selon les critères suivants :

- support informatique ;
- absence de répétition inutile ;
- partage et utilisation des données par des applications ou des utilisateurs distincts ;
- évolution indépendante des données et des applications ;
- protection et contrôle de l'accès aux données.

L'organisation et la gestion de ces bases de données, complexes, sont **assurées par un ensemble de programmes** rassemblés sous le terme de **SGBD** (Système de Gestion de Base de Données, Data Base Management System ou DBMS en anglais).

Il est fréquent que les mêmes données soient **dupliquées** en totalité ou en partie **dans plusieurs fichiers indépendants**.

Il en résulte une **perte de place sur les supports physiques** et des **difficultés de mise à jour** : certaines fiches sont mises à jour plus souvent que d'autres et des données deviennent périmées ou incohérentes.

D'autre part, l'enregistrement des données sous forme de fichiers simples **ne permet pas de prendre en compte efficacement certaines relations entre les informations** (*lien par exemple entre un patient et la liste de toutes ses venues à l'hôpital...*).

## V. BIG DATA ET SANTÉ

Le **Big Data** est la solution permettant à tout le monde **d'accéder en temps réel à des bases de données immenses** et propose ainsi une alternative aux solutions classiques devenues obsolètes face à autant de données.

Le Big Data aide à obtenir une **meilleure représentation de l'interaction** avec les clients, permet de **mieux comprendre leurs besoins** et garantit **la pertinence de l'information délivrée** améliorant ainsi la qualité des services.

Pour mieux comprendre ce qu'est le Big Data on a coutume de citer les 10 V qui le définissent : Volume, Vitesse, Variété, Variabilité, Véracité, Validité, Vulnérabilité, Volatilité, Visualisation, Valeur.

1. **VOLUME** = Le **volume de données à traiter est considérable**. La quantité de données astronomiques générées par les entreprises et les personnes est en constante augmentation. Seul le Big Data est capable de traiter un nombre aussi conséquent de données et d'informations.
2. **VITESSE DE CRÉATION** = La vitesse est la **rapidité à laquelle les données affluent**. C'est-à-dire la **fréquence à laquelle elles sont générées, capturées et partagées**. Avec les nouvelles technologies les données sont générées toujours plus rapidement et dans des temps beaucoup plus courts. Les entreprises sont obligées de les collecter et de les partager en temps réel mais le cycle de génération de nouvelles données se renouvelle très vite, rendant rapidement les informations **obsolètes**.
3. **VARIÉTÉ D'INFORMATIONS** = Les **types de données** et leurs **sources** sont de plus en plus **diversifiés** supprimant ainsi les structures nettes et faciles à consommer des données classiques. Ces nouveaux types de données incluent un grand nombre de contenus très diversifié : *géolocalisation, connexion, mesures, processus, flux, réseaux sociaux, texte, web, images, vidéos, mails, livres, tweets, enregistrements audio...* De par cette diversité qui supprime la structure, **l'intégration des données** à des feuilles de calcul ou application de base de données est de plus en plus **complexe voire impossible**.
4. **VARIABILITÉ** = *A quelle vitesse la structure des données change-t-elle ? A quelle fréquence la forme des données change-t-elle ?* L'important est d'établir si la **structure contextuelle du flux de données est régulière et fiable** même dans des conditions d'imprévisibilité extrême. La variabilité définit la nécessité d'obtenir des **données significatives** en tenant compte de toutes les circonstances possibles.
5. **VÉRACITÉ** = La véracité, l'exactitude des données demeurent aujourd'hui le principal défi du Big Data. À l'heure actuelle, ces données ne sont pas encore suffisamment maîtrisées et la précision des analyses s'en trouve affectée.
6. **VULNÉRABILITÉ** = Le Big Data apporte de nouveaux **problèmes de sécurité**. Il y a quotidiennement des violations de données massives.
7. **VOLATILITÉ** = *A quel âge les données sont considérées comme non pertinentes, historiques ou obsolètes ? Combien de temps faut-il conserver les données ?* Avant l'ère big data, en général on stockait les données **indéfiniment**. Quelques téraoctets de données ne pouvaient pas engendrer de dépenses de stockage élevées. En raison de la vitesse et du volume de ces données massives, leur volatilité doit être soigneusement prise en compte. Il faut établir des **règles pour la disponibilité** et à la **mise à jour des données** afin de garantir une **récupération rapide des informations** en cas de besoin.
8. **VISUALISATION** = Une autre caractéristique du Big Data est la **difficulté à visualiser les données**. Les logiciels de visualisation de données volumineuses actuels sont confrontés à des problèmes techniques en raison des limitations de la technologie en mémoire, de leur faible évolutivité, de leur fonctionnalité et de leur temps de réponse. Il est donc nécessaire d'avoir différentes manières de représenter des données. Si on associe cela avec la multitude de composantes résultant de la variété et de la vélocité des données massives et des relations complexes qui les lient, il est possible de voir qu'il n'est **pas si simple de créer une visualisation qui a du sens**.
9. **VALEUR** = La valeur fait référence à l'objectif, au scénario ou au résultat commercial que la solution analytique doit prendre en compte. *Les données ont-elles une valeur, sinon valent-elles la peine d'être stockées ou collectées ?* L'**analyse** doit être effectuée pour répondre aux **considérations éthiques**.
10. **VALIDITÉ**

Dans le **domaine de la santé**, le **big data** (ou **données massives**) correspond à l'**ensemble des données socio-démographiques et de santé**, disponibles auprès de différentes sources qui les collectent pour diverses raisons.

L'**exploitation** de ces données présente de **nombreux intérêts** : *identification de facteurs de risque de maladie, aide au diagnostic, au choix et au suivi de l'efficacité des traitements, pharmacovigilance, épidémiologie...*

Elle n'en soulève pas moins de **nombreux défis techniques et humains**, et pose autant de **questions éthiques**.

**FIN.**

---

Temps total d'édition = 152 minutes.

Voilà pour cette fiche que je dédicace à tous les PASS qui bossent la Santé Num (et qui vont donc tout défoncer à l'examen). Oubliez pas que c'est des QRUs, donc chiiiiill, je compte sur vous pour me perfect ça, grâce à votre nouvelle superbe fiche faite par votre tutrice préférée (feat. Chiron, le sang).

Mes instructions pour ce cours : y a rien de trop complexe à comprendre, la partie informatique pure a été un peu dégrossie, et pour le reste, je vous conseille de LIRE simplement ça, tranquillement. C'est juste 8 pages à lire, et vous verrez que là vous ferez marcher votre mémoire à court terme, et si y a un QRU (pas trop bâtard, on l'espère) sur ce cours, vous penserez à moi et à cette fiche et à ce que vous avez lu, et vous l'aurez. C'est juste des concepts assez globaux à retenir, vous prenez pas la tête hein.

Une mention toute particulière à ces P1 qui m'ont demandé cette fiche : vous êtes juste adorables, et j'ai trop envie que vous réussissiez, donc voilà c'est pour ça que je vous l'ai faite. Des énormes bisous à Ampz80, Anis, Nova, Paulinepome et même à toi Oscar (j'espère que j'ai oublié personne par contre).

Je veux que vous arriviez jeudi avec un mental de guerrier qui va soulever des grosses daronnes. Le jour-J c'est votre état d'esprit qui fera la différence. On se retrouve en P2 les gars.