Base du traitement de l'information en santé

Coucou ! On se retrouve pour un cours pas si simple que ça mais qui est faisable. Je vais essayer de la rendre plus agréable pour vous à la lire. Préparez-vous une petite collation et un truc à boire puis c'est partiii !!!

I. Donnée, information, connaissance

A) Position du problème

Il n'est pas rare que **l'emploi de certains termes** se fasse au détriment de leur **sens originel.** Certains parleront d'évolution du langage mais il s'agit généralement d'une simple ignorance ou de motivation marketing, un terme fait plus sérieux, plus moderne, plus vendeur. Il en résulte une certaine **confusion**.

L'informatique est une des facettes des sciences de l'information.
➔ L'informaticien travaille sur l'information ou la donnée ?
➔ Est-ce qu'on lui fournit une information ou une connaissance ?

B) La donnée

Donnée = **description élémentaire** d'une réalité qui résulte d'une observation ou d'une mesure avec un instrument. **+++**
➔ C'est une notion **abstraite typée** : elle ne porte pas de sens en elle-même.

- Il y a des données numériques, symboliques, textuelles, logiques...

Exemple : On prend la fonction y=sin(x).
L'angle représenté par la valeur de x n'a pas d'importance, qu'il s'agisse d'un angle dans une pièce, la trajectoire d'un véhicule, la pente d'une courbe d'évolution d'un paramètre biologique.

Lorsqu'on range des données dans une base de données, leur « signification » importe peu.

La grande majorité des traitements réalisés par les informaticiens concernent des données dont le **sens** porté par leurs valeurs **n'est pas déterminant** au sein du traitement.

La **performance** de l'algorithme de stockage et de restitution est **uniquement** liée au type et au volume des données, à la fréquence et à la nature des accès à ces données.

C. L'information

L'information est ce qui donne une forme à l'esprit. Elle vient du verbe latin « *informare* », qui signifie « donner forme à » ou « se former une idée de ».

L'information est aussi une **notion abstraite**, mais d'un niveau d'abstraction supérieur à celui de la **donnée**. **+++**

➔ Pour simplifier : information = une donnée + un sens +++

Comparer deux informations s'avère bien plus complexe que comparer deux données.

➔ Comparer 2 données.
Exemple : On veut comparer 2 adresses. On peut alors faire appel à une « fonction de comparaison de chaînes de caractères »

➔ Comparer 2 informations.
Il faut traiter le « sens »
Exemple : Une température mesurée par un thermomètre est une donnée.
Son expression dans un référentiel d'unité (°C) est une première information. Si on ajoute l'heure et le lieu de la mesure, on enrichit l'information : température corporelle à 8h du matin avant toute activité : 37,2°C.

L'information est donc cet ensemble intelligible de **données**, qui prend un sens **+++**.

À ce sujet, il est possible de distinguer une définition objective et une définition subjective de l'information.

D. Les connaissances

Une fois les **données** décryptées et après leur avoir restitué le sens informatif, il reste à structurer ces **informations** en vue de leur conférer un sens plus large : **la connaissance**.

L'information en soi n'a donc qu'un intérêt très relatif. Elle ne vaudra que parce qu'elle sert de marchepied pour accéder à la **connaissance**. **L'information** n'en est seulement que le vecteur ; tout comme le document est celui de **l'information**.

Un **faisceau d'informations** permet de constituer, de reconstituer ou d'enrichir une **connaissance** sur un sujet.

Exemple : La comparaison d'une mesure de la température corporelle (effectuée dans des conditions spécifiques) à une valeur seuil va permettre de parler de **fébricule**, **d'hyperthermie**. Il s'agit alors d'une **connaissance élémentaire** (interprétation). Cette connaissance peut être enrichie d'une analyse d'un train de mesures pour qualifier **l'évolution** de la température : soudaine ou d'installation progressive.

La connaissance est une **notion abstraite**, d'un niveau d'abstraction supérieur à celui de **l'information** **+++**. La **connaissance**, à la différence de **l'information**, est partagée et s'appuie sur un **référentiel collectif.**

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).

Connaissance tacite	Connaissance explicite
= connaissance que possèdent les individus → Elle n'est **pas formalisée** et **difficilement transmissible**. Ce sont les compétences, les expériences, l'intuition, les secrets de métiers, les tours de main qu'un individu a acquis et échangés lors d'échanges internes et externes à l'entreprise. → Elle se transmet par **imitation** et **imprégnation**. On le sait sans le savoir. On met en œuvre des pratiques sans vraiment s'en rendre compte.	= connaissance formalisée et transmissible sous forme de documents réutilisables. → Ce sont les informations concernant les processus, les projets, les clients, les fournisseurs, etc… → Elle se transmet par des documents **formalisés et normalisés**

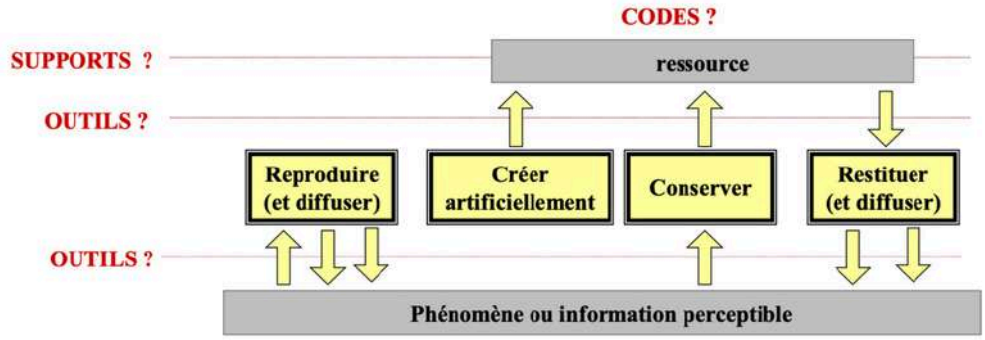
II. Traitement de l'information

Traitement de l'information = C'est la façon dont on **aperçoit** et assimile une information. Le cerveau humain lui **traite** de l'information.

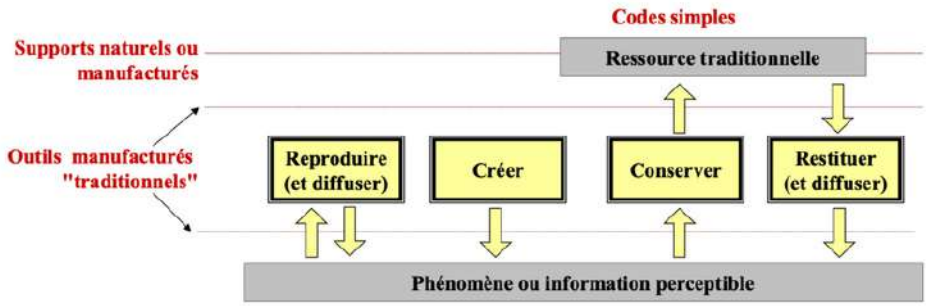
Sur ce traitement existent différents modèles dont un est celui du **double codage de l'information** et de la **formation d'un modèle mental** à partir des deux types de traitement (systèmes « verbal » et « figuratif »).

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**. S'il est confronté à de nouvelles informations, le cerveau va normalement essayer de les intégrer dans les conceptions préalables et ses modèles mentaux préexistants.

Nous allons voir maintenant les différents types de traitements de l'information.



A. Les technologies traditionnelles (Antiquité → 18ème siècle)



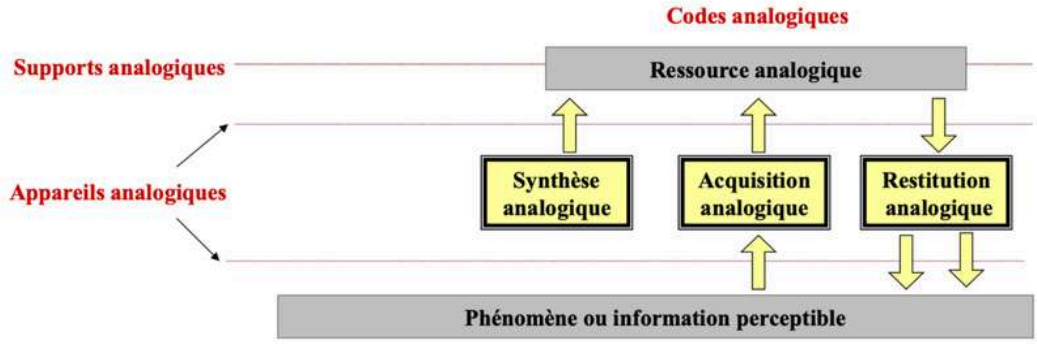
L'homme perçoit les phénomènes sans comprendre (ou analyser) leur nature. Il exprime sa perception (pas reproductible à l'identique).

➔ Certes il y a une technique de peinture, mais il n'y a **pas de traitement de l'information pour reproduire/restituer.**
Exemple : L'écriture des livres au moyen âge constituait une œuvre unique.

➔ Cette œuvre pouvait être conservée mais restait difficilement diffusable.

➔ L'invention de l'imprimerie a permis la reproduction et la diffusion.

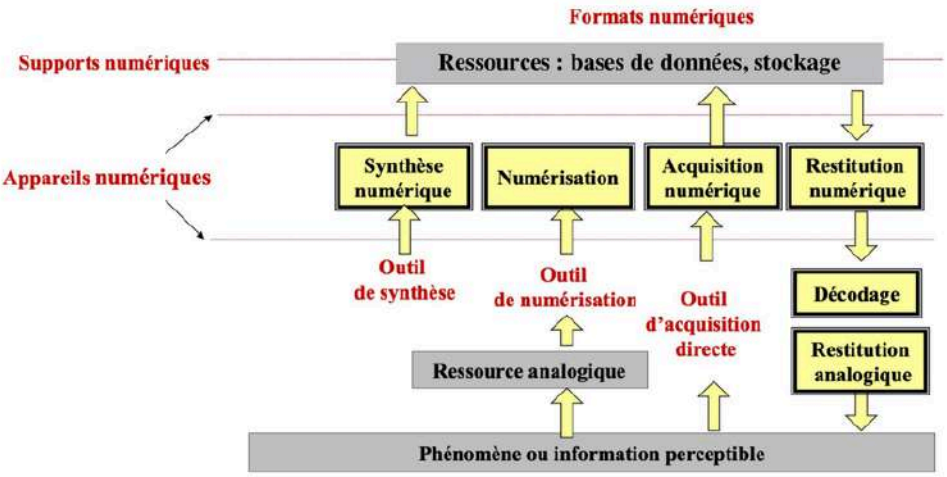
B. Les technologies analogiques (Antiquité → 18ème siècle)



Dans le monde analogique, l'acquisition des phénomènes, des appareils et des instruments de mesure est représentée par une **information exprimant la variation d'une grandeur physique** (une masse, une force...).
Ex : L'ampoule électrique convertit analogiquement l'énergie électrique en lumière.
Ex : La balance de pesage, qui mesure l'équilibre entre deux forces ou deux masses, est un des premiers exemples connus d'analogie mécanique.
Ex : La photographie argentique est une écriture analogique de la lumière.

Transducteur analogique : dispositif matériel permettant la conversion (par analogie) d'un phénomène physique en un autre phénomène physique en vue de sa diffusion ou de son stockage.
Ex : le microphone convertit les vibrations sonores en signaux électriques pour la diffusion

C. Les technologies numériques (Depuis l'antiquité de l'information)



L'homme utilise des **codes informatiques** pour représenter l'information, stocker cette représentation et la traiter.

➔ Ici, il va utiliser **différents outils** pour chaque étapes (synthèse, numérisation, acquisition).

III. Traitement numérique

A. Information numérique

Sur une machine (ordinateur, tablette, smartphone), 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 (affichage, réseau, ...).

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...).

B. Numérisation de l'information

Certaines informations, portées par des **grandeurs physiques** (tension électrique, intensité lumineuse) 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.

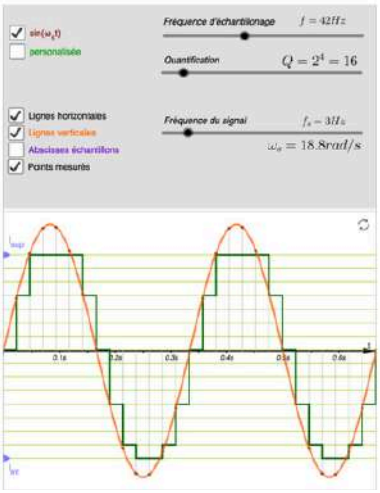
La numérisation d'un signal analogique peut se faire par **échantillonnage** :

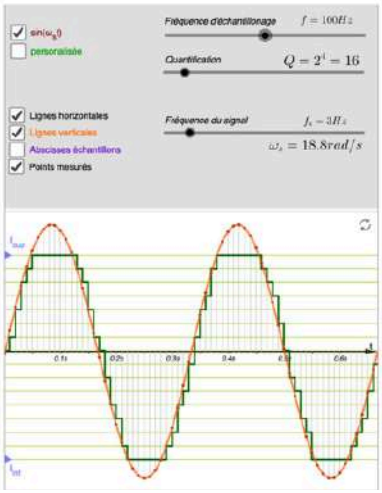
➔ On découpe le temps en intervalles réguliers → fréquence d'échantillonnage (Hz).

➔ A chaque période d'échantillonnage, on mesure l'amplitude du signal et on la convertit en un nombre entier (on parle de « quantification ») → résolution (nb de bits).

➔ **Plus la fréquence d'échantillonnage et la résolution sont élevées, plus la numérisation est fidèle +++**

En **orange** : courbe du son originel
En **vert** : c'est le signal analogique
Rq : on voit bien que le son est découpé en intervalles réguliers (Hz)





C. Codage de l'information

★ Quantum d'information :

- élément binaire (0 ou 1) → Bit (Binary digit)
- Octet = 8 bits
- Mot machine (8, 16, 32, 64 bits)

★ Numération binaire (base 2)

- Calculs numériques

★ Logique

- Vrai = 1 Faux = 0
- Raisonnement

★ n bits permettent de coder

- 2^n objets
  - ➢ 8 bits (2^8 = 256 objets)
  - ➢ 16 bits (2^{16} = 65 536 objets)

★ Pour N Objets Combien de bits ?

- Entier supérieur à $\log_2(N)$
  - ➢ 1000 objets ($\log_2$ (1000) = 10 bits)
  - ➢ 128 objets ($\log_2$ (128) = 7 bits)

D. 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

- 65 536 objets
- Tous les alphabets + idéogrammes

E. Données chiffrées

★ Nombres entiers

- naturels
  - ➢ 8 bits (0 à 255)
  - ➢ d'une manière générale un codage sur n bits pourra permettre 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 un seul bit, le bit de poids fort (celui le plus à gauche)
➢ L'exposant est codé sur les 8 bits consécutifs au signe
➢ La mantisse (les bits situés après la virgule) sur les 23 bits restants
➢ Ainsi le codage se fait sous la forme suivante :
seeeeeeeemmmmmmmmmmmmmmmmmmmmm
➢ Le **s** représente le bit relatif au signe, les **e** représentent les bits relatifs à l'exposant, les **m** représentent les bits relatifs à la mantisse

Quand j'étais à votre place je n'avais jamais compris cette partie et c'est okay, en vrai retenez juste comment calculer un bit et N objet, qu'il faut 10 bits pour couvrir 1000 objets et 7 bits pour 128 objets, je pense que c'est le plus important !

Numérisation de données de biologie :

L'informatisation des laboratoires est un des domaines d'application les plus anciens et on en est désormais au stade de l'exploitation industrielle.

L'intérêt de ces systèmes est généralement dans le gain de productivité et de qualité du service effectué.

L'origine de la demande d'informatisation est l'augmentation du volume d'informations à gérer.
Exemple : inflation de la consommation d'examens, augmentation du nombre de résultats fournis par des systèmes d'analyse de plus en plus automatiques.

Quelle que soit la discipline (biochimie, hématologie ou bactériologie), ces systèmes permettent les mêmes fonctionnalités :

➔ **L'enregistrement** des demandes d'analyse
➔ Le **tri** des examens et ventilation par poste de travail
➔ **L'acquisition des résultats** :

➢ Acquisition automatique par interfaçage (connexion par interface de plusieurs systèmes) avec un analyseur automatique qui exige néanmoins une vérification et une double validation par le technicien de laboratoire et par le biologiste.
➢ Certains examens (hématologie, cytologie, anatomo-pathologie, bactériologie) sont largement **manuels** et **la saisie** l'est donc également.
➢ Pour les comptes-rendus, les contraintes liées à la formalisation du langage médical s'appliquent ici et ont débouché sur des solutions de type **questionnaire standardisé** et **langage de classification** (*exemple* : ADICAP *en anatomopathologie*).

➔ La consultation et l'édition des résultats analysés.
➔ La gestion du laboratoire.
➔ L'archivage des dossiers.

Récapi'Tut de Nono :

L'informatisation des laboratoires sert à gérer le grand volume d'examens et de résultats, en améliorant la productivité et la qualité.
Ces systèmes permettent d'enregistrer les demandes, d'organiser les analyses, d'acquérir et valider les résultats (automatiques ou manuels), de produire les comptes-rendus, puis d'assurer la consultation, la gestion et l'archivage.

F. Données images : imagerie médicale

(Le paragraphe en gris qui suit n'est pas à apprendre...)

★ *Image Statique :*
- *Image Bitmap : 1 bit par pixel (Noir/Blanc) 1024 X 1024 points : 1 M de Bits :128 Ko*
- *Image 512 × 512 × 8 bits (256 Niveaux de gris) 256 Ko*
- *Image 1024 × 1024 × 8 bits (256 Couleurs) 1 Mo*
- *Image 1024 × 1024 × 16 bits (65000 Couleurs) 2 Mo*

★ *Séquence d'images :*
- *Endoscopie*
- *Angiographie + coronarographie échographie*
- *Images 512 × 512 à 8 niveaux de gris (256 K octets) 4.5 Mo 5 secondes de film : 22 Mo*
- *Vidéo 24 images par seconde ; 6 Mo /seconde ; 1 minute = 36 Mo*

★ *Possibilité de compression :*
- *De Facteur 4 (sans perte) à 10 (avec perte d'information)*

Depuis les **années 70**, trois nouvelles techniques, basées sur le traitement informatique, ont bouleversé l'imagerie médicale :
★ 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 qui sont devenues peu ou pro-numériques par conversion d'images sources.

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 : le problème consiste à identifier les paramètres et les structures signifiants puis à les comparer à des structures connues ou à les confronter à des connaissances théoriques ou expérimentales.

On peut alors voir des exemples d'application :
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 3D d'images d'organes** montre les rapports des structures entre elles (organes, tumeurs, structures vasculaires). Particulièrement employée dans le domaine de la neurologie, elle peut déboucher sur la création d'un espace en réalité virtuelle où le médecin peut se déplacer ou sur la production automatique de moules en 3D, afin que le chirurgien puisse **repérer les voies** d'abord ou **répéter l'intervention.**

La **chirurgie assistée par ordinateur** associe aux phases d'acquisition et d'interprétation d'images, deux 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 dernière étape peut prendre la forme d'une **aide passive** (détection d'écarts au geste prévu), **semi-active** (système de contraintes) voire **active** (autonomie du robot).

Récapi'tut de Nono

🔹 Nouvelles techniques depuis les années 70 :
- Scanner (TDM)
- Angiographie numérisée
- IRM

🔹 Autres techniques devenues numériques : scintigraphie, échographie, endoscopie vidéo, radiologie.

🔹 Applications actuelles :
- Téléimagerie : transmission d'images pour télédiagnostic / télésurveillance.
- Reconstruction 3D : visualiser organes/tumeurs, préparer des chirurgies, réalité virtuelle.
- Chirurgie assistée par ordinateur : modéliser le patient, simuler l'opération, aide au geste (passive → semi-active → robot autonome).

👉 À retenir : l'informatique a transformé l'imagerie médicale (nouvelles techniques, numérisation des anciennes, nouvelles applications comme téléimagerie, 3D et chirurgie assistée).

G. Données signal : signaux physiologiques

★ Signaux numériques : la taille mémoire dépend de la fréquence d'échantillonnage
Exemple : ECG (électrocardiogramme), EEG (électroencéphalogramme), EMG (électromyogramme)

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 **3 étapes** :

1) Le signal est d'abord découpé en segments de durées égales, c'est **l'échantillonnage**.
2) La hauteur de chaque segment est alors **quantifiée** (en prenant une valeur moyenne).
3) 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 2 niveaux et correspond à un signal en « tout ou rien » (ex : froid ou chaud) ; 2 bits autorisent 4 niveaux possibles ; alors qu'1 octet (8 bits) correspond à 256 (28) niveaux différents possibles.)

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

- 1) **Acquisition du signal analogique** par un capteur et **numérisation** par un convertisseur analogique-digital.
- 2) **Pré-traitement** simple visant à l'amélioration de la qualité du signal (extraction du signal sur le bruit, amplification, filtration).
- 3) **Traitement analytique** permettant l'extraction de paramètres, (ex : les complexes QRS d'un ECG) le plus souvent par des méthodes mathématiques.
- 4) **Interprétation** des résultats.

Récapi'tut de Nono :

À retenir : un signal doit être numérisé (3 étapes) pour être manipulé, puis passe par 4 phases de traitement (acquisition → pré-traitement → analyse → interprétation).

IV. Gestion informatique des données

A. 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 celle 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.

B. 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 se 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.

Exemple : des malades seront tous décrits par leur nom, leur prénom... Seules changent les valeurs de ces caractéristiques pour chaque individu.

★ **Enregistrement** (article ou fichier) = 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.**

C. 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 précédent, ce qui peut être très long s'il y a beaucoup d'enregistrements.

★ Accès direct :

Sur les disques et les disquettes, les informations sont enregistrées sur des pistes concentriques, partagées en secteurs. Chaque enregistrement a une adresse formée d'un numéro de piste et d'un numéro de secteur. On peut donc positionner directement la tête de lecture sur la piste puis 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, de la même façon que l'index d'un livre indique à quelle page apparaît tel mot.

➔ La **clé d'index** permet d'identifier de façon **unique** un enregistrement.

➔ La gestion de l'index est normalement assurée par le logiciel de gestion de données.

L'utilisateur ne voit que le fichier principal et la clé, il demande "lire enregistrement de clé "Martin", le système récupère alors le numéro d'enregistrement dans la table d'index pour accéder directement à cet enregistrement.

La **clé d'index** peut être **simple** ou **composée** de plusieurs critères (par exemple, nom et prénom) afin d'être plus discriminante.

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 *(par exemple l'adresse ou le diagnostic).*

D. Gestion informatique des fichiers

➔ 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.

E. Base 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.** *(Exemple : lien par exemple entre un patient et la liste de toutes ses venues à l'hôpital...)*

V. Big Data et santé

A. Big Data

★ 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.

B. Définition du Big Data

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 +++.**

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.
Vitesse	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**.

Variété	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 diversifiés *(ex : 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**.
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. C'est particulièrement le cas lorsque la collecte de données repose sur le traitement de la langue. → En effet, les mots n'ont pas de définitions statiques et leur signification peut varier énormément selon le contexte.
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.
Validité	(Pas d'explication du prof sur ce point.)
Vulnérabilité	Le Big Data apporte de **nouveaux problèmes de sécurité.** Il y a quotidiennement des violations de données massives. *Exemple rapporté par CRN : en mai 2016, « un pirate informatique a posté des données sur le dark web pour les vendre, qui concernaient des informations sur 167 millions de comptes LinkedIn et … 360 millions d'e-mails et de mots de passe des utilisateurs de MySpace ».*

Volatilité	A quel âge les données sont-elles 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.
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 impossible de se fier aux graphiques traditionnels lorsqu'on essaie de tracer un milliard de points de données. Il est donc nécessaire d'avoir différentes manières de représenter des données telles que la mise en cluster de données ou l'utilisation de cartes d'arbres, de sunbursts, de coordonnées parallèles, de diagrammes de réseau circulaires ou de cônes. 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.**
Valeur	Ce V décrit la valeur qu'il est possible d'obtenir à partir des données et comment les mégadonnées obtiennent de meilleurs résultats à partir de données stockées. 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**.

C. Le Big Data en santé

Conclusion :

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.**

C'est la fin de ce long cours : félicitations à toi d'avoir survécu ! Faites à fond les QCMs des annatuts, des annales pour cibler les points importants du cours.

Maintenant place aux dédissss (je sais que vous attendez que ça) :

- Dédi à ma mère qui m'en a fait voir de toutes les couleurs, qui m'a checké quand je lui ai dit que je suis passée en P2, qui m'a supporté toute l'année au téléphone quand j'en avais marre de réviser, qui me motivait quand j'avais un coup de mou, qui m'a démotivé le soir avant les oraux, qui m'a motivé à nouveau juste avant les oraux (je sais d'où vient le fait que je sois indécise), qui m'a amené Nala l'amour de ma vie quand j'étais au plus bas de ma vie, qui a voulu vendre Nala 1 an après sans mon consentement libre, éclairé et conscient, bref qui est ma meilleure amie et ma pire ennemie en même temps
- Dédi à Lina, mon binôme de p1, LAS1 et LAS2, à nos cours de spé physique qu'on a enduré en terminale avec la prof qui a dit qu'on ne réussirait jamais la LAS qu'avec le tutorat #rageuse, à nos fous rires à la bu de Valrose, à nos chats qui portent le même nom (Nala #AucuneOriginalité), à nos TPs de SV (mention spéciale au TP de biochimie où c'était une catastrophe et que Lina s'est embrouillé avec le prof et l'a remis à sa place)
- Dédi à Anto : tous les NÂGAs mangés, à notre marathon seigneur des anneaux, notre marathon stranger things, nos sessions BU, nos cours d'espagnol à jouer au président en terminale, nos cours de philo avec Lina et notre prof très perchée (je ne sais pas si elle mérite sa dédi…), à nos cours de spé physique survécu, aux coups de pression que tu me mettais quand j'osais regarder mon téléphone une demi-seconde au lieu de réviser

- Dédi à Gaby : mon petit rayon de soleil sur pattes, l'année prochaine sera la tienne, l'administration de SV pue le caca qu'on se le dise et elle ne te mérite pas, la p2 t'attend avec impatience
- Dédi à Andreea : cette année c'est la bonne, j'espère que tu réussiras et je tiens ta place en p2 bien au chaud, crois en toi comme moi je crois en toi
- Dédi à Nala, mon chat d'amour, mon soutien émotionnel, qui se prend parfois pour un chien (oui elle va chercher la balle, si un jour elle aboie ça ne m'étonnerait même pas), qui me suit de PARTOUT #PotDeColle (à l'heure où j'écris cette dédi, elle est sur moi), qui martyrise toutes les personnes qui viennent chez moi (promis elle est gentille, juste un peu bipolaire sur les bords, Jannastomose et Meleviscère peuvent en témoigner), sans qui je m'ennuierais au quotidien (je pense que mes voisins me prennent pour une folle avec tous les fous rires que j'ai avec mon chat)
- À ma titine qui tient la route après 2 ans de possession, qui supporte tous les car'arokés de chaque personne qui monte dans ma voiture et de moi-même, qui me permet de faire le taxi de tout mon entourage
- Anti-dedi à la personne qui m'a rentré dedans au feu rouge 30 min avant la TTR
- Dédi à la team NYC qui me supporte depuis 10 ans (ça fait beaucoup là non?) : ma Carla, ma partenaire de concert, de voyage, de Kpop, de Stranger Things, de Disney, de StarAc, qui me suit dans toutes les bêtises que je fais, carrément sa maison c'est ma maison (je ne rigole même pas) ; ma Yaya avec qui je partage la passion de la danse (#HipHop), celle qui me fait mourir de rire depuis toujours, ses tantes me prévenaient du repas de famille avant elle carrément (ma deuxième famille), on se le fera ce voyage à NYC <3
- Dédi à Lindsey, ma première amie quand j'ai déménagé à Montauroux (y'a 10 ans aussi palalala #vieille), qui m'a fait rencontrer Yaya et Carla, mon panda, qui m'a soutenu dans les moments très difficiles de ma vie, j'en suis très reconnaissante <3
- Dédi à Elea et nos partiels en commun quand j'étais en LAS2 SV alors qu'on était pas du tout dans la même licence (j'ai jamais compris ce qu'elle faisait, je sais juste que ça tourne autour de l'écologie, écosystème etc.. et qu'elle a cours de PARTOUT)
- Dédi à David, le père de Carla, qui a mieux réagi que ma mère quand il a su que je passais en p2 (il m'a invité au resto et l'a dit à tout mon village, carrément chaque personne venait me voir à mon travail pour me féliciter), qui m'a invité à Zanzibar avec eux mais que j'ai du refusé pour les cours et le tut #TutriceInvestie
- Dédi à mes grands parents : ma mamie qui me carry les courses tous les dimanches, qui ramène le marché chez elle carrément (elle achète les cagettes de fruits et légumes quand elle sait que je viens), mon papi qui remplit mon portefeuille
- Dédi à mes oncles qui croyaient en moi depuis petite et qui espéraient que je suis la relève de la famille, qui ont cotisé pour m'offrir ma voiture bancale
- Dédi à ma tante qui m'a offert l'occasion de voyager avec elle, qui m'a fait vivre un rêve éveillée (Ibiza, Paris, Londres ma ville de coeur, la Bretagne, la Normandie), c'est la femme la plus forte que je connaisse <3
- Le meilleur pour la fin : dédi à mon papa que j'aime fort, celui qui m'a poussé depuis petite à faire les études de mes rêves <3

FIN