Base du traitement de l'information en santé

Coucou mes guerriers! J'espère que vous êtes prêt pour ce cours!! La fiche est un peu plus longue que d'habitude mais j'ai mis pas mal d'exemple et d'illustration, j'ai rajouté certaines chose par rapport à l'année dernière! N'hésitez pas à le couper en deux si besoin. J'espère que ce cours va autant vous plaire qu'à moi.
Toujours petit rappel, en **vert** ce son mes exemples à ne pas apprendre!! Sur ce bonne lecture.
J'ai pas eu la place pour les dédis alors j'en fait une ici à Ellicase pour m'avoir donnez accès aux diapos!

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.
Mais l'informaticien travaille-t-il sur l'information ou la donnée ?
Est-ce qu'on lui fournit une information ou une connaissance ?

B) La donnée

Une donnée est une **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…
Si 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, peu importe leur signification. 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.
On peut dire que 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.

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.
On peut dire pour simplifier que l'information est **une donnée + un sens**.
Comparer deux informations s'avère bien plus complexe que comparer deux données.
Si on veut comparer deux adresses, on peut faire appel à une fonction de comparaison de chaînes de caractères. Si on veut comparer deux informations il faut traiter le « **sens** »
Par 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.

Ainsi 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).

- **La connaissance tacite** : c'est la 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.
La connaissance tacite se transmet par **imitation** et **imprégnation**. On le sait sans le savoir. On met en œuvre des pratiques sans vraiment s'en rendre compte.

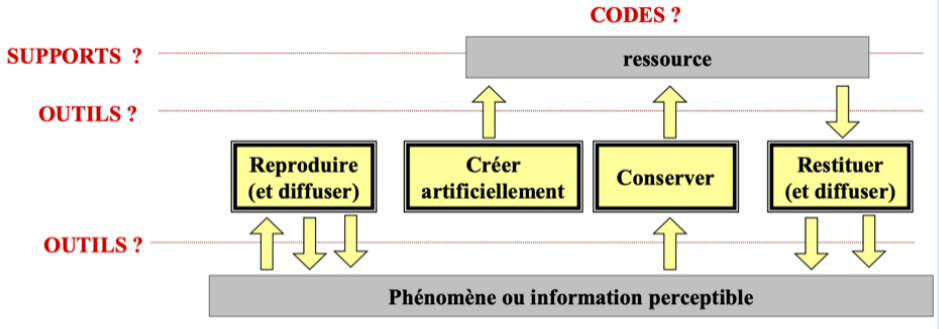
- **La connaissance explicite** : C'est la connaissance formalisée et transmissible sous forme de documents réutilisables. Ce sont les informations concernant les processus, les projets, les clients, les fournisseurs, etc.

La connaissance explicite se transmet par des **documents formalisés et normalisés.**

II- Traitement de l'information

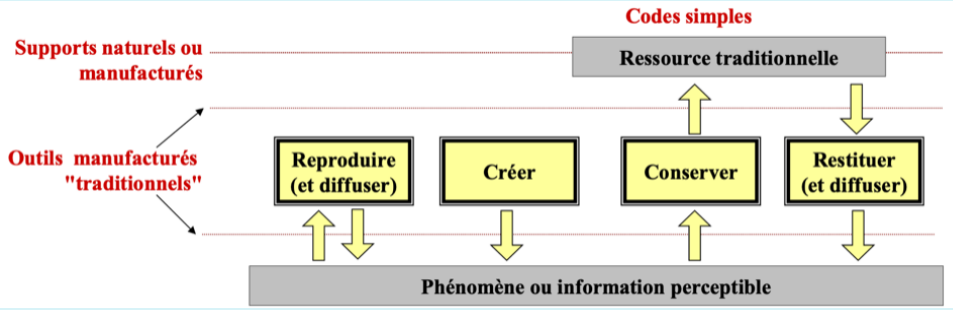
C'est la façon dont on aperçoit et assimile une information. Le cerveau humain lui traite de l'information.
Sur ce traitement existent différents modèles dont, un est celui du **double codage** de **l'information** et de la **formation** d'un modèle mental à partir des deux types de traitement (systèmes « 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.



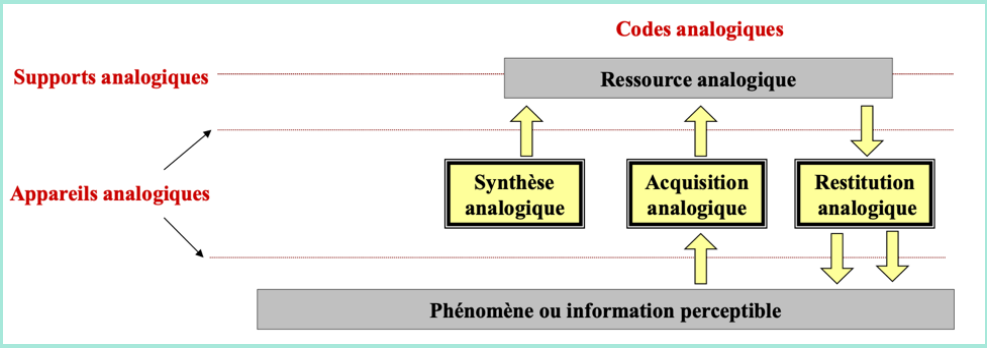
1°) On créer artificiellement que ce soit un phénomène ou bien une information [ex : éditer une fiche de cours]
2°) On conserve [ex : sur mon disque dure]
=> devient une ressource utilisable
3°) Restituer ou diffuser [ex : je publie ma fiche sur le forum]

A) Technologies traditionnelles (Antiquité au 18e siècle)



L'homme perçoit les **phénomènes** sans comprendre (ou analyser) leur **nature**. Il exprime sa **perception** (pas reproductible à l'identique). Les outils de **création** et les **supports** de conservation (manufacturés ou supports naturels [ex : du peintre et d'un tableau]) . Certes il y a une technique de peinture, mais il n'y a pas de traitement de l'information pour reproduire ou restituer. 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.

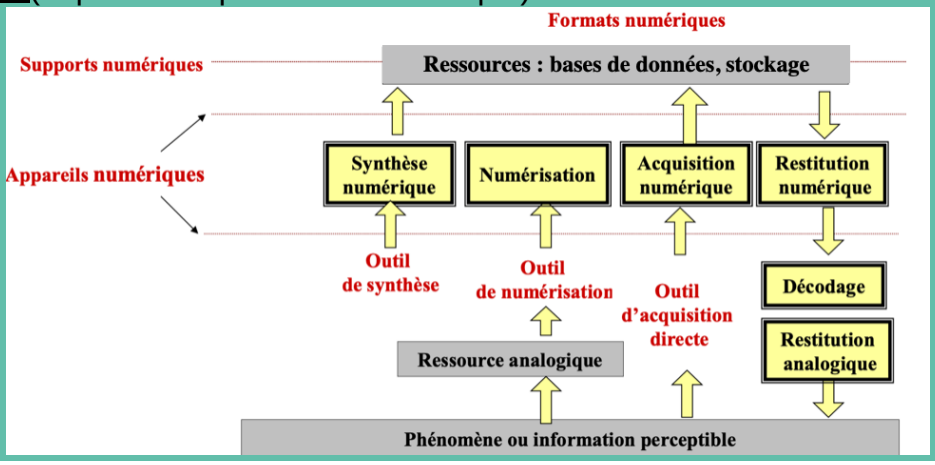
B) Technologies analogique (19e au 20e siècle)



L'homme comprend la **nature physique** des phénomènes. 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…). 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. 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) Technologies numériques (depuis l'antiquité de l'informatique)



L'homme utilise des codes informatiques pour représenter l'information, stocker cette réprésentation et la traiter. Ici il va utiliser **différents outils** pour chaque étapes (synthèse, numérisation, acquisition)

III- Traitement numérique

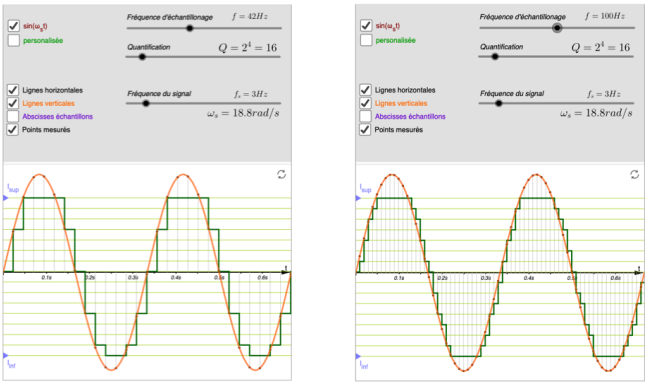
Sur une machine (ordinateur, tablette, smartphone), toute l'information se trouve sous forme numérique, que ce soit dans la mémoire de masse (stockage), dans la mémoire vive, dans le microprocesseur et au niveau de tous les périphériques, et notamment ceux de communication (affichage, réseau, …)
Les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

A) 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** (nombre de bits)

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



En jaune c'est la courbe du son originel
En vers c'est le signal analogique. On voit bien que le son est découpé en intervalles régulier (Hz)

B) Codage de l'information

• Quantum d'information :
- élément binaire (0 ou 1) → Bit (Binary digit)
- 1octet=8bits
- 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 2n objets (ex : 8 bits = 256 objets)

TUT'AIDE : 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.

C) 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)

D) 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 2n-1.
- Relatifs : 16 bits (-32 768 à +32 767). D'une manière générale, le plus grand entier relatif positif codé sur n bits sera 2n-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
- La mantisse (les bits situés après la virgule) est codée sur les 23 bits restants
- Ainsi, le codage se fait sous la forme suivante :

seeeeeeeemmmmmmmmmmmmmmmmmmmmm

Avec le s représentant le bit relatif au signe, les e représentant les b**it**s relatifs à l'exposant, et les m représentant les bits relatifs à la mantisse.

TUT'AIDE : Le système Mantisse Exposant permet la conversion d'un nombre décimal en nombre entier, un ordinateur ne pouvant stocker que ces derniers. Donc pour représenter des nombres décimaux, on est obligés de le faire de manière « détournée ».
Par exemple, ici on code un nombre réel en 32 bits par ce système de Mantisse Exposant. On va donc avoir un « mot » de 32 bits divisés en plusieurs sous-parties.
On a donc un mot de 1+8+23 = 32 bits (valeurs 0 ou 1), qui représente un nombre décimal où le premier bit donne le signe (+ ou -), et les 8 suivants donnent l'exposant en notation scientifique, et enfin le reste de bits permet de donner les chiffres après la virgule.

Numérisation de données de biologie :

L'informatisation des laboratoires est un des domaines d'application les plus anciens et 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 [ex : 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 : automatique par interfaçage 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 (par exemple ADICAP en anatomopathologie).

- **La consultation** et **l'édition** des résultats analysés
- **La gestion** du laboratoire
- **L'archivage** des dossiers

E) Données image (lisez juste pour vous informez)

- 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 :

- 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 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.
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 **trois dimensions** d'images des 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 trois dimensions, 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).

Données signales :signaux physiologiques

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

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 (28) niveaux différents possibles).

La séquence de 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, par exemple les complexes QRS d'un ECG, le plus souvent par des méthodes mathématiques ;
4. Interprétation des résultats.

IV- 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.

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.

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** :

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 [ex : 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 [c'est un peu son identité perso].
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 "nom", 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 [ex : nom et prénom] afin d'être plus discriminante (c'est pour éviter les homonymes et les doublons). 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).

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.

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…)

IV- Gestion informatique des données

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.

Volume	Volume de données considérables à traiter. La quantité 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	Rapidité à laquelle les données affluent. C'est à dire la fréquence à laquelles 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 d données classiques. Ce nouveau type de donnée incluent un très grand nombre de contenus très diversifié : (géolocalisation, connexion, mesures, processus, flux, réseaux sociaux, texte, web, images).
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 le flux de données est régulière et fiable même dans des condition d'imprévisibilité extrême. Nécessité d'obtenir des données significatives en tenant compte de toutes circonstances possibles. Ex : lorsque la collecte de données repose sur le traitement de la langue. Les mots n'ont pas de définitions statiques et leur significati peut varier énormément selon le contexte.
Véracité	La véracité, l'exactitude des données demeurent aujourd'hui le principal défi du Big Data. A l'heure actuelle, ces données ne sont pas encore suffisament maîtrisées et la précision des analyses est affectée.
Vulnérabilité	Le Big Data approte de nouveaux problèmes de sécurté. Il y a quotidiennement des violations de données massives
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, on stockait les données **indéfiniment**. En raison de la vitesse et du volume de ces données massives, leur volatilité doit être soigneusement prise en compte. Des **règles pour la disponibilité** et à la **mise à jour des données** pour garant une **récupération rapide des informations** en cas de besoin.
Visualisation	**Difficulté à visualiser les données**. Problèmes techniques en raison des limitations de la technologie en mémoire, faible évolutivité, fonctionnalité et temps de réponse. Nécessité d'avoir différentes manières de représenter des données. Associé 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	Objectif, scénario ou 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**.
Validité	

j'ai fais un tableau pour tous condenser en une fois <3

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, aux choix et aux suivsi 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!! Vous êtes trop fort! allez boire un coup <3)