Données, informations et connaissances

M162 — Analyser et modéliser des données

Jérôme Frossard

EPAI

19 août 2025

Introduction

Même si nous avons toutes et tous une compréhension intuitive des notions de données, d'information et de connaissance, il nous semble utile, pour introduire un module sur l'analyse et la modélisation des données, de clarifier ces notions et examiner la manière dont elles s'articulent entre elles, et avec un certain nombre de notions connexes.

Le but de cette présentation est de :

- Définir les notions de données, d'information et connaissance
- Examiner les liens entre ces trois notions
- Aborder brièvement les notions de codage, de message et de document.

Plan

1. Données, informations et connaissances
2. Traitement de données
3. Code, message et document
4. Codage analogique
5. Codage numérique
6. Données structurées et données non structurées

Plan

1. Données, informations et connaissances
2. Traitement de données
3. Code, message et document
4. Codage analogique
5. Codage numérique
6. Données structurées et données non structurées

Qu'est-ce qu'une donnée ?

De manière générale, on peut définir une **donnée** comme étant :

- Ce qui est connu et admis (un fait, un événement)
- Le point de départ d'un raisonnement.
- Ce à quoi l'on peut donner du sens pour produire une information.

Une **donnée brute** est une donnée qui n'a subi aucun traitement depuis son acquisition. Cela peut être :

- La mesure d'une grandeur physique (masse, tension électrique, luminosité, position d'une manette, etc.).
- Le rapport objectif d'un fait ou d'un événement (témoignage).
- Les réponses d'un questionnaire.

Qu’est-ce qu’une information ?

Pour interpréter des données, on doit en connaitre le contexte. Elles n’ont pas de sens propre. Par exemple :

- Si je vois les données « 7 » et « 30 » sur une feuille, cela ne m’apprend rien. Ce n’est pas une information.
- Si je vois « température : 7°C ; précipitation : 30 mm », j’en sais un peu plus, mais je ne sais pas quoi en faire. Ce n’est pas une information.
- Si je vois « Bulletin météo pour demain. température : 7°C ; précipitation : 30 mm », je sais qu’il fera frais et qu’il tombera des cordes. Cela me permet d’agir, par exemple, en annulant le pique-nique au bord du lac. C’est une information.

Une information est une donnée ou un ensemble de données contextualisées qui amène une connaissance nouvelle (elle répond à une question ou à un besoin spécifique) et peut permettre de décider et agir (elle peut avoir un effet).

Qu'est-ce qu'une connaissance ?

Pour faire court, une connaissance est une **croyance vraie justifiée**.

Voyons cela de plus près :

- **Croyance.** Une croyance est quelque chose que je tiens pour vrai. Je ne peux pas dire que je sais quelque chose si je ne crois pas que ce quelque chose est vrai.
- **Vraie.** Une croyance est une connaissance si elle se vérifie effectivement dans les faits (vérité correspondance) ou qu'elle me permet d'agir avec succès (pragmatisme).
- **Justifiée.** Je ne peux pas dire que je sais quelque chose, si je ne sais pas comment je le sais. La justification est souvent une expérience directe (évidentialisme) ou une source d'information fiable (fiabilisme).

Données, informations et connaissances

Données, connaissances, et information sont étroitement liées.

Nos connaissances (ainsi que d'autres croyances : présupposés métaphysiques, valeurs, croyances ordinaires, heuristiques, etc.) nous permettent :

- D'analyser et d'interpréter les données perçues par nos sens pour en extraire l'information et construire de nouvelles connaissances.
- De produire de l'information pour partager nos connaissances, par exemple, sous la forme de textes, de graphiques, de dessins, ou encore de formules mathématiques.

La manière spécifique dont nous analysons et interprétons les données à partir de nos connaissances et autres croyances pour produire de l'information (de nouvelles données) constitue un **traitement de données**.

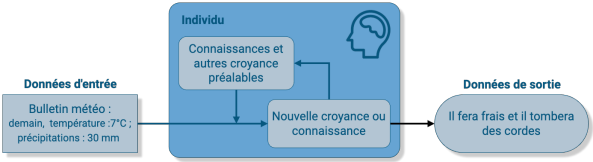
Plan

1. Données, informations et connaissances
2. Traitement de données
3. Code, message et document
4. Codage analogique
5. Codage numérique
6. Données structurées et données non structurées

Qu'est-ce qu'un traitement de données ?

Dans notre exemple, sur la base de mon expérience et de la confiance que j'accorde aux prévisions du bulletin météo (demain, 7°C et 30mm de pluie), je peux vous dire qu'il fera frais et qu'il tombera des cordes.

En faisant cela, j'effectue un **traitement de données** (*data processing*).



Un traitement de données est effectué par un **système de traitement de données** (*data processing system* ou DPS). Selon la cybernétique, les êtres vivants sont de tels systèmes.

Exemples de traitement de données

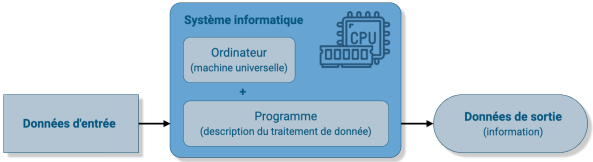
Un traitement de données n'est pas nécessairement compliqué. On peut construire des traitements très élaborés à partir de traitements simples.

Par exemple :

- Écrire (mémoriser) et lire (se rappeler)
- Comparer
- Trier et classer
- Filtrer et sélectionner
- Compter
- Mesurer la tendance centrale (mode, médiane, moyenne, etc.)
- Effectuer une opération arithmétique ou logique

Traitement de données et ordinateur

Le terme **ordinateur** a été choisi en 1955 par IBM France pour traduire EDPM (*electronic data processing machine*) : machine de traitement de données électronique.



Mais toutes les machines peuvent être considérées comme des systèmes de traitement de données. La particularité d'un ordinateur est d'être une **machine universelle** qui peut :

- simuler n'importe quelle machine (y compris elle-même),
- et donc, effectuer n'importe quel traitement de données (également appelé "calcul")

Pour cela, un ordinateur exécute un **programme**. Un programme est une **description d'une machine ou d'un traitement de données**, codée à l'aide d'un langage de programmation.

Plan

1. Données, informations et connaissances
2. Traitement de données
3. Code, message et document
4. Codage analogique
5. Codage numérique
6. Données structurées et données non structurées

Qu'est-ce que le codage ?

Pour être communiquée, une connaissance ou une idée doit être « matérialisée » d'une manière ou d'une autre, elle doit être **codée**.

Par exemple, pour partager une idée, je peux l'exprimer :

- De manière verbale, en produisant un texte oral ou écrit dans une certaine langue.
- De manière non verbale, en dessinant sur du papier un schéma, un diagramme, un graphique, etc.

L'expression de mon idée constitue des **données**. Le texte oral, le texte écrit, le schéma et le diagramme sont **différentes manières de coder** ces données.

Message et document

Les données codées prennent la forme d'un ensemble de signes sonores ou visuels qui constitue un **message**.

Dans le cas du texte écrit, du schéma et du diagramme, la feuille sur laquelle il se trouve est un **support transportable, reproductible et stable** qui constitue un **document**.

Un document a deux fonctions essentielles :

- **Transmission :** Un document a pour but de communiquer des données.
- **Mémorisation :** Un document constitue un enregistrement des données.

Dans le cas du texte oral, si celui-ci n'a pas été enregistré d'une manière ou d'une autre sur un support physique, il n'y a pas de document.

Code et convention

Si une personne a entendu et mémorisé le texte oral, ou lu le document qui contient le texte écrit, le schéma ou le diagramme, alors cette personne a reçu mon message.

Pour extraire l'information des données du message et connaître mon idée, cette personne doit être capable de les **interpréter**.

Pour cela, elle doit connaître le code que j'ai utilisé pour coder les données :

- La langue parlée ou écrite dans laquelle le texte est formulé.
- Le formalisme du schéma ou du diagramme.

L'échange d'information implique donc d'avoir des connaissances communes et de s'accorder sur la manière de les utiliser. En d'autres termes, l'échange d'information implique des **conventions**.

Un code est une **manière conventionnelle de représenter des données.**

Code, conventions et normes

Les langues naturelles sont un peu particulières. Elles ont bien quelque chose de “conventionnel”, mais il s’agit de **systèmes conventionnels** complexes sur lesquels personne n’a véritablement de contrôle.

En informatique, en revanche, les conventions sont en général bien **documentées**, relativement **stables**, et dont l’évolution est souvent contrôlée par une organisation (entreprise, consortium, institut de normalisation national ou international, etc.).

Par exemple, dans ce module nous utiliserons deux conventions de codage :

- **Notation de Chen** : La notation de Chen, proposée par Peter Chen en 1976, est une manière de représenter un modèle conceptuel de données. Cette notation est largement utilisée et acceptée, mais elle n’est pas normalisée (**standard de fait**).
- **SQL** : Le SQL (Structured Query Language) est un langage standard pour la gestion des bases de données relationnelles, dont la spécification fait l’objet d’une **norme internationale** maintenue par l’ISO.

Code, conventions et normes — Autres exemples

- **Unicode** : L'Unicode est une norme pour le codage des caractères maintenue par le Consortium Unicode. Elle associe un numéro unique (point de code) à chaque caractère, et définit comment coder ces numéros en binaire (UTF-8, UTF-16)
- **HTML** : HTML (HyperText Markup Language) est le langage utilisé pour structurer et présenter le contenu des pages web. Ses spécifications sont développées et maintenues par le W3C (World Wide Web Consortium).
- **HTTP** : Le HTTP (HyperText Transfer Protocol) est le protocole utilisé pour la communication sur le web, notamment entre les navigateurs et les serveurs. Il est spécifié et maintenu par l'IETF (Internet Engineering Task Force).
- **Java** : Java est un langage de programmation et une plateforme maintenus par Oracle. Les spécifications sont développées via le Java Community Process (JCP), un système ouvert qui permet aux développeurs et aux entreprises de contribuer à son évolution.

Codage des données et ordinateur

Pour traiter des données avec un ordinateur, il doit être en mesure de les interpréter.

Il est facile de fabriquer des composants mécaniques, électromécaniques ou électroniques à deux états (allumé ou éteint). De plus, Claude Shannon a montré dès la fin des années 1930 que ces composants permettaient de réaliser non seulement des opérations logiques (ET, OU, NON), mais aussi des opérations d'arithmétique binaire.

Nos ordinateurs traitent donc les données en utilisant deux symboles qui correspondent à ces deux états, conventionnellement désignés par 0 et 1, les chiffres de la base 2 (*binary digit* ou bit). Puisque des groupes de bits permettent de représenter des nombres, on dit qu'un ordinateur est une machine numérique.

Pour traiter un texte, un schéma, ou un son avec un ordinateur, ce texte, ce schéma et ce son doivent être **codés avec des nombres**. Le traitement de données avec un ordinateur exige un **codage numérique**.

Chaine de codage

Une donnée codée est une nouvelle donnée qui peut être codée à son tour.

Par exemple :

- L'information correspondant à une idée peut être codée sous forme d'un **texte oral**.
- Ce texte peut ensuite être saisi sur un clavier et codé sous forme d'une **suite de nombres** binaires de 8 bits (octets).
- Cette suite de nombres binaires est stockée dans la mémoire de l'ordinateur sous forme de bits, représentés par des **charges électrostatiques**.
- Le contenu de la mémoire peut être enregistré dans un fichier sur un disque dur, où il est codé par l'orientation de **grains magnétiques**.
- Ce fichier peut être compressé (**codage de source**) et envoyé vers un serveur via un réseau Ethernet, où il est codé par la modulation d'une **onde électromagnétique**.
- Sur le serveur de fichiers, les données véhiculées par l'onde électromagnétique sont décodées, puis recodées sous forme de charges électrostatiques en mémoire, etc.

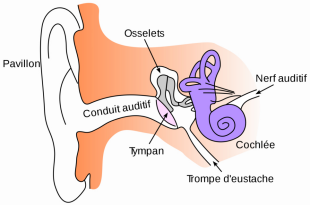
Plan

1. Données, informations et connaissances
2. Traitement de données
3. Code, message et document
4. Codage analogique
5. Codage numérique
6. Données structurées et données non structurées

Codage analogique

Le mot **analogique** vient du mot « analogue », qui signifie « semblable ».

Lorsque je parle, le codage du texte est analogique. Mes cordes vocales produisent une onde sonore **continue dans le temps et en amplitude**, analogue au sens du texte qui encode mon idée.



Quand l'onde sonore qui voyage dans l'air parvient à votre oreille, le tympan et les osselets oscillent de manière analogue à cette onde sonore, et produisent avec la cochlée un influx nerveux analogue à ces oscillations.

Exemple — Disque à microsillon (vinyle)

Un autre exemple de codage analogique est le disque vinyle. Les sillons gravés dans le disque ont une forme analogue à celle d'une onde sonore.

À la lecture, l'aiguille oscille de manière analogue à la forme des sillons. Comme cette forme est analogue à l'onde sonore enregistrée, l'aiguille oscille elle aussi de manière analogue à cette onde sonore.

Le mouvement de l'aiguille est converti en un signal électrique avec un capteur piézoélectrique, puis ce signal est amplifié et converti en une onde sonore par un haut-parleur.

Exemple — Montre analogique

Dans une montre analogique, la position angulaire de la petite aiguille indique l'heure qu'il est sur un cycle de douze heures. La position angulaire est analogue à l'heure.

La montre analogique met en évidence un problème important du codage analogique : la précision de lecture.

La rotation de la petite aiguille est continue (ou peut être considérée comme telle), sa position angulaire à chaque instant correspond à une heure précise dans le cycle de 12h. Mais il est difficile de la lire dans cette position avec précision sur le cadran, une minute correspondant à une variation d'angle de seulement un demi-degré.

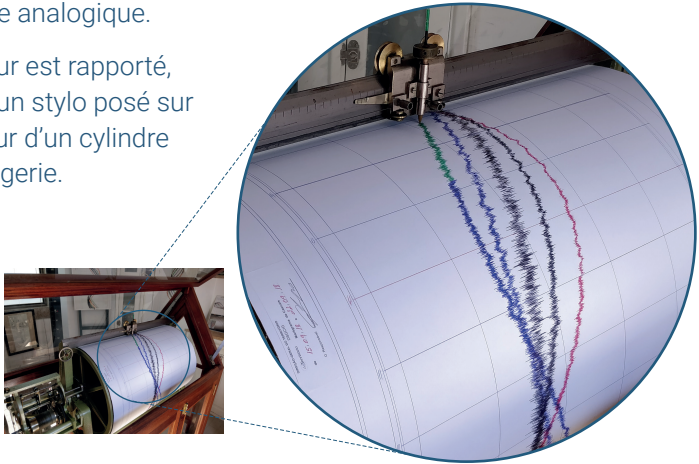
Pour faciliter la lecture, une grande aiguille indique la minute dans l'heure, et une trotteuse, la seconde dans la minute.

Exemple — Enregistreur de marées

Enfin, un enregistreur de marées analogique comme celui de Cascais au Portugal est encore un autre exemple de codage analogique.

Le déplacement vertical d'un flotteur est rapporté, par l'intermédiaire d'une poulies, à un stylo posé sur une feuille de papier enroulée autour d'un cylindre entraîné par un mécanisme d'horlogerie.

La courbe tracée sur le papier est analogue à l'élévation de l'eau.



Exemple — Avantages et inconvénients

Jusqu'aux années 1980, les systèmes analogiques avaient l'avantage de permettre un traitement en temps réel, ce que les ordinateurs de l'époque ne pouvaient pas toujours assurer.

Jusqu'à cette époque, pour des applications telles que l'enregistrement ou la reproduction audio et vidéo, la transmission d'informations, la simulation de vol, la conduite de tire, ou encore la surveillance de centrale nucléaire, on utilise donc des systèmes analogiques.

Mais le codage analogique, a deux gros inconvénients :

- Une précision de lecture limitée et dépendante de l'habileté de l'opérateur
- Une grande **susceptibilité au bruit**.

Codage analogique — Sensibilité au bruit

Le bruit est un signal qui ne contient pas d'information utile et qui se superpose au signal des données. Dans l'enregistreur de marées, le clapotement des vagues, les jeux dans le mécanisme, les variations de température, etc., constituent autant de sources de bruit.

Un signal analogique, continu en amplitude est particulièrement sensible au bruit, car même un bruit relativement faible réduit la précision de lecture. Comme une certaine quantité de bruit est inévitable, cette précision est limitée de façon irréductible.

Le bruit est particulièrement gênant dans la transmission de données. Alors qu'un chuchotement suffit à se faire entendre dans un lieu silencieux, il faudra sensiblement élever la voix dans le brouhaha d'une foule. Si l'amplitude du bruit augmente, il faut augmenter l'amplitude du signal pour conserver un bon rapport signal sur bruit.

Plan

1. Données, informations et connaissances
2. Traitement de données
3. Code, message et document
4. Codage analogique
5. Codage numérique
6. Données structurées et données non structurées

Numérisation

Le codage numérique consiste à représenter un signal par **une suite de nombres**. Alors que le codage analogique est généralement continu, le codage numérique est **discret**. L'information est représentée par une séquence de symboles séparé les uns des autres.

Le processus qui consiste à coder un signal analogique sous forme numérique est appelé **numérisation** (*digitization* en anglais).

Par exemple, plutôt que de dessiner en continu la variation de hauteur de la surface de l'eau, on mesure la hauteur à intervalle régulier (**échantillonnage**) à l'aide d'une échelle graduée (**quantification**), et on enregistre la valeur de la mesure, par exemple dans un carnet.

Tant la fréquence d'échantillonnage est au moins le double de la fréquence maximale du signal à mesurer, on ne perd pas d'information. Dans le cas des marées, une mesure toutes les heures devrait permettre de reconstituer le signal avec une bonne fidélité.

Transmission de données numériques

Une fois numérisé, le signal est beaucoup plus résistant au bruit. Tant que les chiffres (bits) peuvent être distingués sans ambiguïté, on conserve toute l'information.

C'est particulièrement utile pour la transmission de données. Le signal peut être transmis sans aucune altération tant que le bruit reste en dessous d'un certain seuil. De plus, la **redondance** offerte par des **codes de détection et de correction d'erreur** permet d'accroitre encore cette résistance.

Mais cela n'est pas gratuit. Un traitement est nécessaire pour reconstituer le signal. Si celui-ci a été échantillonné à une cadence de plusieurs milliers de fois par seconde, cela demande une puissance de traitement considérable.

En outre, dans le cas d'un signal analogique, la dégradation est généralement progressive, tandis que dans le cas d'un signal numérique, le signal est parfaitement restitué, jusqu'au moment où il ne peut plus du tout l'être (*cliff effect*).

Traitement de données numérique

Contrairement à un signal analogique dont le traitement nécessite un dispositif spécifique pour chaque application, un signal codé numériquement peut être traité à l'aide de programmes exécutés par un ordinateur (une machine universelle).

Cela permet d'enchaîner les traitements :

- suppression du bruit, linéarisation, etc.
- chiffrement, compression, code de correction d'erreur, etc.

Les capteurs et actuateurs restent évidemment analogiques. Mais la miniaturisation et l'augmentation de la puissance des systèmes informatiques ont permis de rapprocher toujours plus les convertisseurs analogique-numérique (ADC) et numérique-analogique (DAC) des capteurs et des actuateurs.

Ce rapprochement, associé à la simplification de l'électronique de conditionnement du signal analogique, réduit considérablement le risque de perturbation avant la numérisation.

Exemple : téléphone portable (I/II)

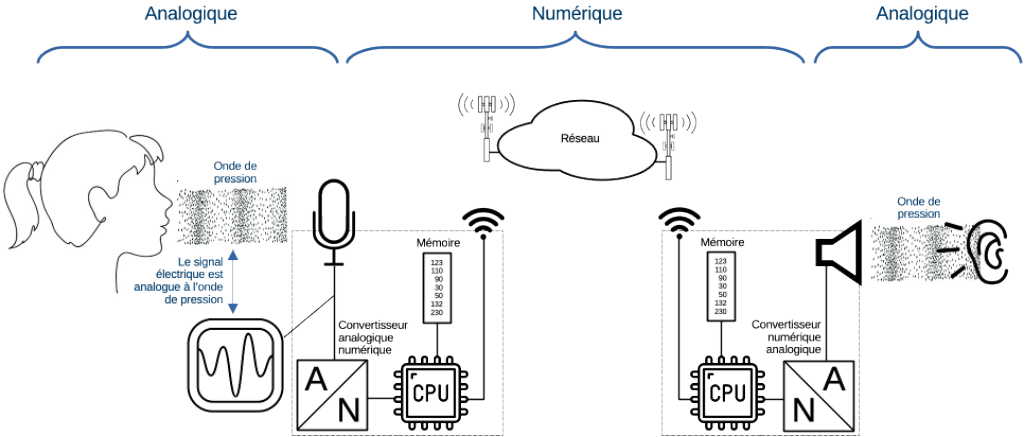
Lorsque je parle au téléphone, le microphone produit un signal électrique analogue à l'onde sonore de mes paroles. Ce signal n'est pas transmis tel quel sur le réseau téléphonique. Il est d'abord **numérisé**.

Son amplitude est mesurée plusieurs milliers de fois chaque seconde avec un ADC. Le résultat est une suite de nombres. Il s'agit d'un **codage numérique**.

Ces nombres sont brièvement stockés dans une mémoire pour être traités (compressées, chiffrées, etc.), puis le résultat est codé par la modulation d'une onde électromagnétique.

Chez l'interlocuteur, le signal électromagnétique est démodulé, puis les nombres sont décodés et convertis les uns après les autres en une tension électrique par un DAC. Si la cadence suit celle de la numérisation, le signal électrique produit est analogue au son de mes paroles. Enfin, ce signal est converti en une onde sonore par un haut-parleur.

Exemple : téléphone portable (II/II)



Plan

1. Données, informations et connaissances
2. Traitement de données
3. Code, message et document
4. Codage analogique
5. Codage numérique
6. Données structurées et données non structurées

Données structurées et données non structurées

Nous l'avons vu, un ordinateur peut traiter des données, si elles sont codées sous une forme numérique et stockées dans sa mémoire. Toutefois, certaines données sont plus faciles à traiter que d'autres. Pour un ordinateur :

Il est difficile, voire très difficile, de traiter des **données non structurées** :

- Les textes dans une langue naturelle (lettres, offres, rapport, courriel, etc.)
- Le son, l'image, la vidéo

Il est en revanche très facile de traiter des **données structurées** :

- Du texte dans un langage de programmation
- Des balises HTML
- Des feuilles Excel ou des fichiers CSV
- Des fichiers XML

Données non structurées et systèmes d'IA générative

Aujourd'hui, on compte beaucoup sur les systèmes dits d'intelligence artificielle générative pour le traitement de données non structurées.

Bien que ces systèmes soient impressionnants à première vue, ils sont loin d'être parfaits.

- Même si le taux de succès peut être relativement élevé, à grande échelle le nombre d'erreurs est significatif et les erreurs sont difficiles à détecter.
- Beaucoup de ces systèmes sont non déterministes, les mêmes données ne produisent pas toujours le même résultat.
- Ils sont plus coûteux en temps et en énergie qu'une solution classique ; si le traitement peut être réalisé par un algorithme ou une heuristique, il doit l'être.
- Enfin, ces systèmes reposent sur l'exploitation massive de ressources naturelles dont l'impact pèse surtout sur des populations qui ne bénéficient pas de ces technologies, et sur le travail mal rémunéré et souvent traumatisant d'un très grand nombre de personnes chargées d'annoter les données, principalement dans les pays du Sud.