

Visualisation de données massives

Ahmed ADJIBADE

ahmed.adjibade@bivariant.com

Année universitaire : 2025-2026

L'explosion des données

- 2,5 quintillions d'octets créés chaque jour ($2,5 \times 10^{18}$)
- 90% des données mondiales ont été créées ces 2 dernières années
- Un individu génère en moyenne 1,7 Mo de données par seconde
- D'ici 2025 : 463 exaoctets créés quotidiennement

Ces données proviennent des réseaux sociaux, des objets connectés, des transactions, des logs, etc.



Source : [Decideo](#)

Plus de données \neq Plus de connaissances

Que représente cette feuille Excel ?

Calibri

11

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

A

Plus de données \neq Plus de connaissances

- Avoir des données ne suffit pas !
- Les données brutes sont inexploitable par l'humain
- Une feuille Excel peut contenir jusqu'à 1 048 576 lignes mais ne donne aucune information spontanée.

Mais alors,

- Comment gérer cette surcharge informationnelle ?
- Comment extraire du sens de ces masses de données ?
- Comment prendre des décisions éclairés ?

L'homme contre la machine

Où se trouve le panda ?



L'homme contre la machine

Où sont les maisons ?



Images pré-catastrophe annotée et non annotée dans le jeu de données xBD

L'homme contre la machine

Qu'est-ce que vous remarquez ?



*Images pré et post catastrophe annotées (hurricane-harvey) avec des dommages mineurs et des dommages majeurs
issu du jeu de données xBD*

L'homme contre la machine

L'ordinateur excelle quand ...

- La question est précise et bien définie
- Il faut traiter des volumes énormes
- Il faut se répéter sans erreur
- Il faut calculer rapidement

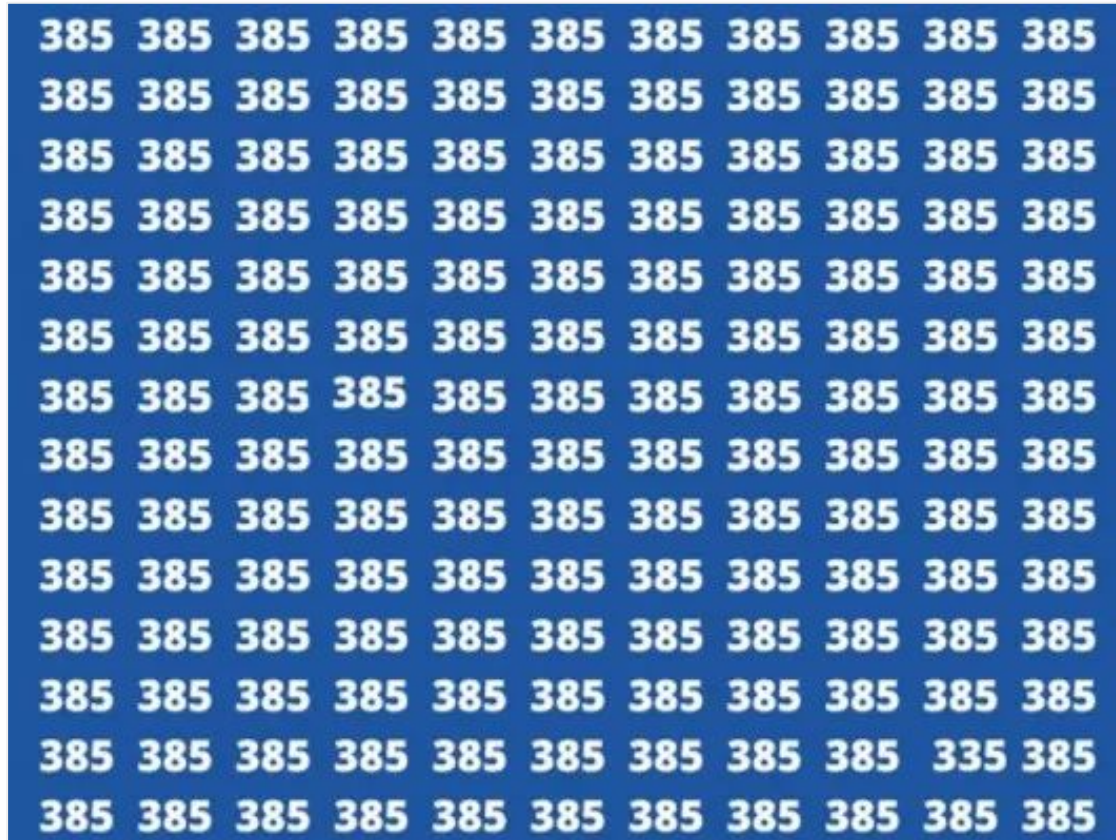
L'homme excelle quand ...

- La question est floue ou ouverte
- Il faut interpréter, contextualiser
- Il faut faire preuve de créativité
- Il faut détecter des anomalies visuelles

La visualisation de données permet à l'homme de tirer parti de la puissance de calcul de la machine tout en utilisant ses capacités cognitives.

Le cerveau humain est une machine visuelle

Trouvez 335 ...



Le cerveau humain est une machine visuelle

Eurêka !



Le cerveau humain est une machine visuelle

- 50% du cortex cérébral est dédié au traitement visuel
- Le cerveau traite une image en 13 millisecondes
- Nous pouvons distinguer jusqu'à 10 millions de couleurs

On détecte donc instantanément :

- Les motifs et les irrégularités
- Les anomalies
- Les tendances et les évolutions
- Les clusters

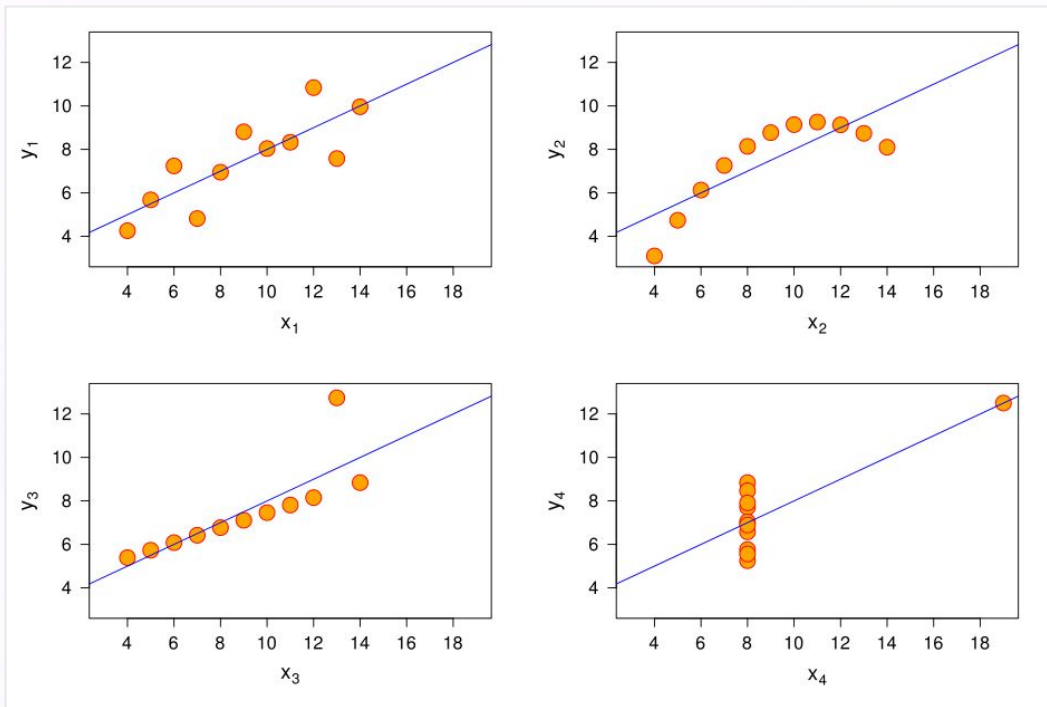
Même données, perception différente

Le Quartet d'Anscombe (1973)

4 jeux de données avec des statistiques identiques :

- Même moyenne de X et Y
- Même variance
- Même corrélation
- Même droite de régression

Mais les distributions sont complètement différentes.



Qu'est-ce que la visualisation de données ?

“ L'utilisation de représentations visuelles, interactives et informatiques de données abstraites pour amplifier la cognition.”

Card, Mackinlay & Shneiderman (1999)

De manière plus détaillée :

- **Représentations visuelles** : graphiques, cartes, diagrammes
- **Interactives** : l'utilisateur peut explorer, filtrer, zoomer
- **Informatiques** : générées par ordinateur, dynamiques
- **Données abstraites** : pas seulement des objets physiques
- **Amplifier la cognition** : augmenter nos capacités de compréhension

Critères d'une visualisation

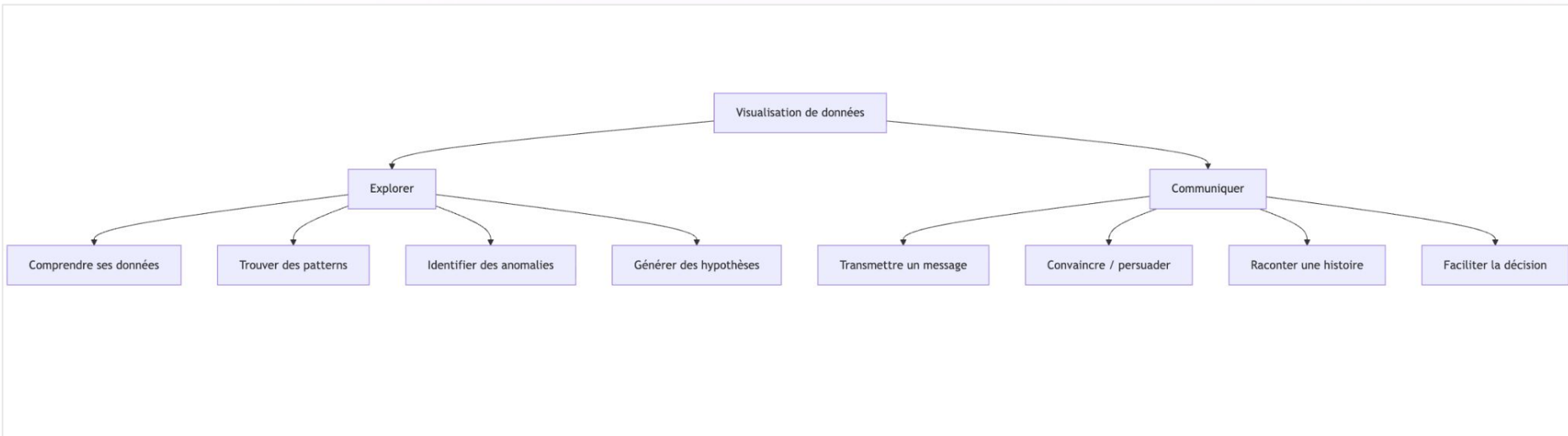
- Facile à **lire**
- Facile à **comprendre**
- **Difficile** à mal interpréter
- Conduit à se poser des questions
 - Plus de détails ?
 - Plus de contexte ?
 - Une autre façon de présenter les choses ?
- Une visualisation est juste un point de départ ...

Ce que la visualisation n'est PAS

La visualisation de données n'est pas :

- **De la décoration** : Chaque élément visuel doit porter de l'information utile.
- **De l'infographie** : La visualisation se concentre spécifiquement sur la représentation de données.
- **Un exercice artistique** : La clarté prime toujours sur la beauté.
- **Un processus automatique** : Un bon graphique nécessite des choix réfléchis et contextualisés.

Les deux grands objectifs : Explorer vs Communiquer



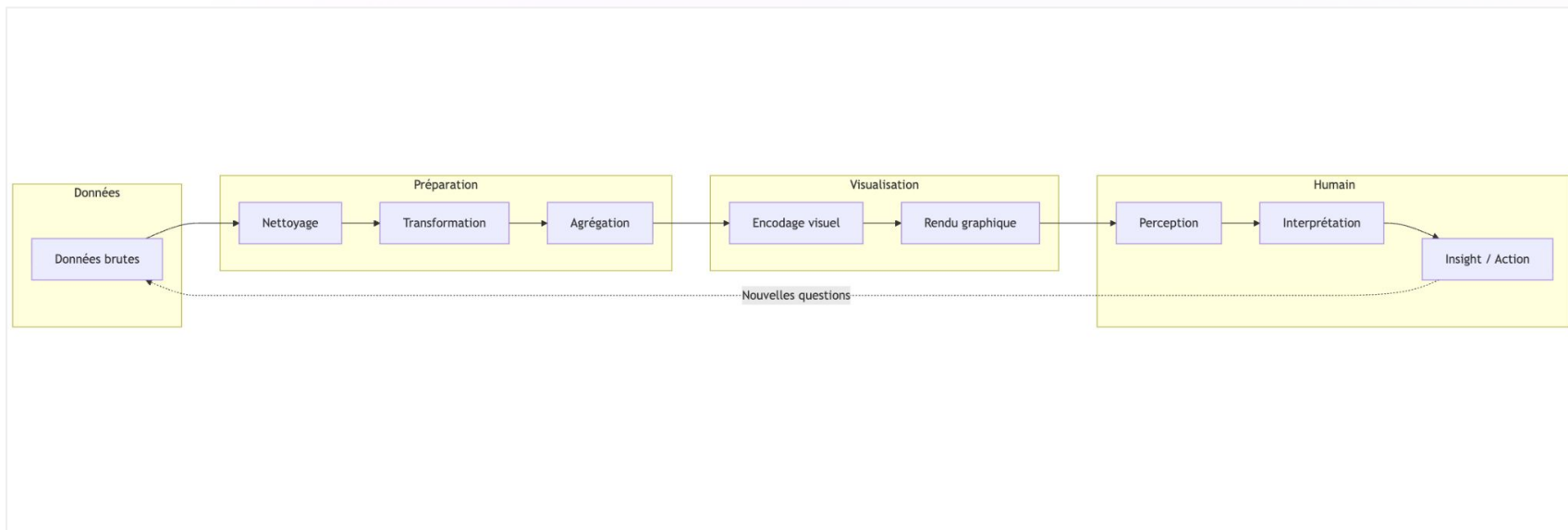
Les deux grands objectifs : Explorer vs Communiquer

Deux approches, deux contextes

Aspect	Exploration	Communication
Public	Soi-même, équipe data	Décideurs, grand public
Objectif	Découvrir	Transmettre
Nombre de graphiques	Beaucoup, jetables	Peu, soignés
Interactivité	Essentielle	Optionnelle
Temps de création	Rapide, itératif	Long, réfléchi
Esthétique	Secondaire	Importante
Annotations	Minimales	Essentielles

Le processus de visualisation

De la donnée brute à la connaissance



Le processus de visualisation : Étape 1 - Préparation des données

- **Nettoyage**

- Gérer les valeurs manquantes
- Corriger les erreurs et incohérences (ex: str dans une colonne number)
- Standardiser les formats (dates, nombres)

- **Transformation**

- Calculer de nouvelles variables (ex: âge à partir de date de naissance)
- Normaliser ou discrétiser si nécessaire
- Pivoter ou dé-pivoter les tables

Le processus de visualisation : Étape 1 - Préparation des données

- **Agrégation**

- Résumer les données (somme, moyenne, comptage)
- Choisir le bon niveau de granularité

80% du temps d'un projet data est consacré à la préparation des données.

Le processus de visualisation : Étape 2 - Encodage visuel

L'encodage visuel consiste à faire correspondre une variable de données à une propriété visuelle pour la représenter graphiquement.

Les canaux visuels disponibles sont :

- **Position (x,y)** - le plus précis
- **Longueur** - très efficace pour comparer
- **Angle / Aire** - moins précis
- **Couleur** (teinte, saturation, etc)
- **Forme** - idéal pour les catégories
- **Taille** - idéal pour les quantités

Le défi est de pouvoir choisir le bon encodage pour chaque type de données.

Le processus de visualisation : Étape 3 - Perception et interprétation

- **Perception** : Le système visuel traite l'image de manière automatique, rapide et inconsciente.
- **Interprétation** : Le cerveau donne du sens aux informations perçues en les contextualisant avec ses connaissances.
- **Insight** : La découverte d'une compréhension nouvelle qui mène à une action ou une question.

Votre rôle c'est de faciliter le passage de la perception à l'insight.

Les types de données

Il est très important de comprendre ses données **AVANT** de les visualiser.

- Le type de donnée détermine les visualisations possibles
- Une erreur de classification conduit à une visualisation trompeuse
- Les outils (comme Tableau) catégorisent automatiquement mais ils peuvent se tromper

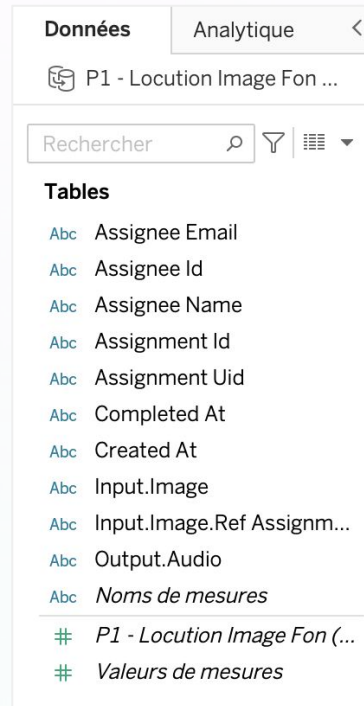
Les types de données

La grande distinction dans Tableau

Dimensions	Mesures
Variables qualitatives	Variables quantitatives
Catégorisent les données	Quantifient les données
Répondent à : Quoi ? Qui ? Où ? Quand ?	Répondent à : Combien ?
Bleues dans Tableau	Vertes dans Tableau

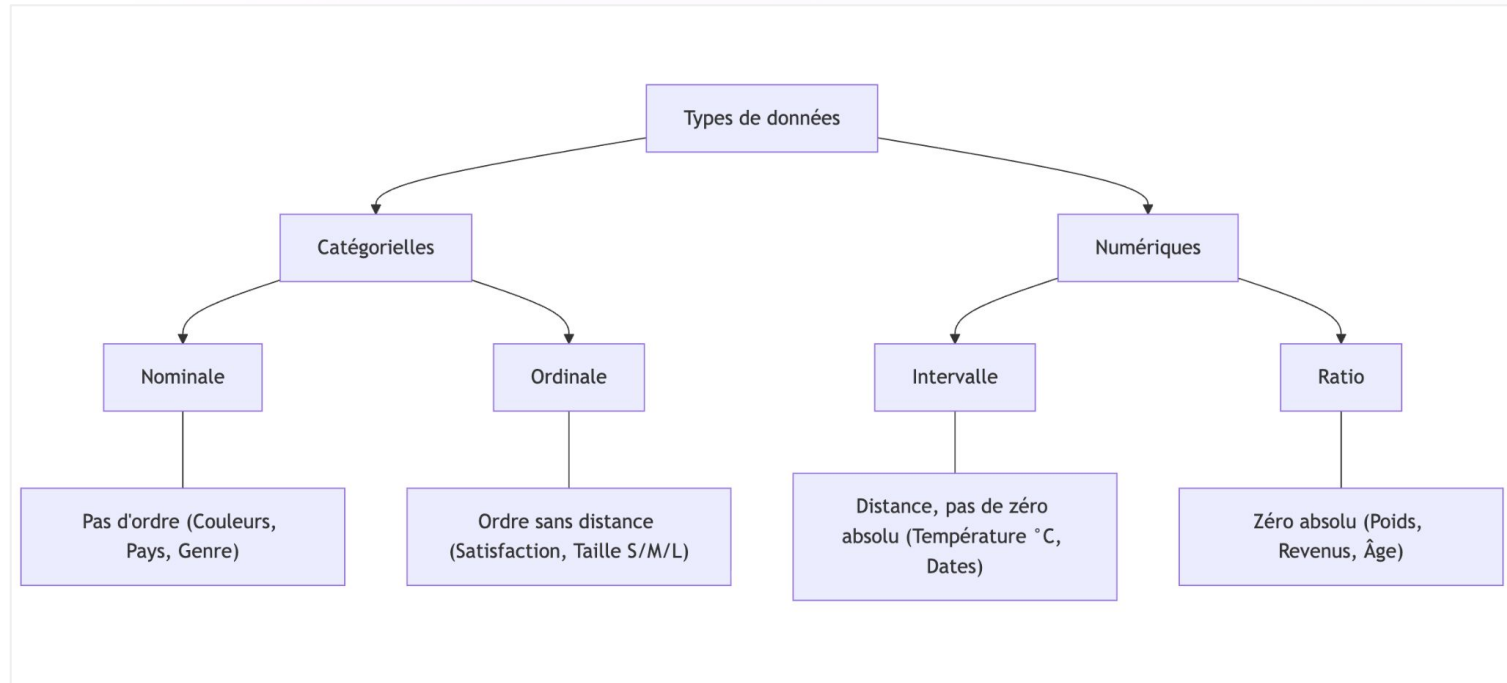
Exemples :

- Dimensions : Pays, Catégorie, Nom client, Date
- Mesures : Ventes, Quantité, Température



Les échelles de mesure

Quatre niveaux de mesures (Stevens, 1946)



Les échelles de mesure

Le type de donnée guide les choix visuels

Type	Encodages recommandés	À éviter
Nominal	Couleur (teinte), Forme, Position spatiale séparée	Taille, Dégradé de couleur
Ordinal	Couleur ordonnée (dégradé), Position sur un axe	Formes distinctes sans logique
Intervalle / Ratio	Position, Longueur, Taille, Saturation	Forme



Qu'est-ce que Tableau ?

- L'un des leaders de la visualisation de données dans le monde
- Interface drag-and-drop, pas de code
- Visualisations interactives et dashboards

Pour ce cours, nous utiliserons [Tableau Public](#) qui est une version libre et gratuite.

TP1

Objectifs du TP:

- Installer et configurer Tableau Public
- Découvrir l'interface par la pratique
- Créer vos premières visualisations

Vous aurez besoin d'un ordinateur avec Tableau Public installé et le fichier de données fourni.