

Données — Structuration et traitement

Ce qui est noté au contrôle, ce n'est pas le résultat seul : c'est le raisonnement posé, puis le calcul mené jusqu'au bout avec l'unité.

1 Vocabulaire. Répondre en une phrase précise.

- ① a. Une **donnée** organisée selon un **schéma connu** — le plus souvent une table : une ligne par enregistrement, une colonne par descripteur.
- ② b. Le **CSV** écrit une table **plate** en texte brut ; le **JSON** organise des couples nom-valeur et accepte des structures **imbriquées**.
- ③ c. Une information **sur** la donnée : par exemple l'**unité** de mesure et la **date** de collecte.
- ④ d. Toute information se rapportant à une personne **identifiée ou identifiable**, y compris **par recoupement** — une adresse **IP** en est une.
- ⑤ e. La **finalité**, la **minimisation**, la **durée limitée** de conservation — ou encore le **consentement** et les droits d'accès et d'effacement.

2 Une table recense les emprunts d'une bibliothèque : identifiant, titre, catégorie, date d'emprunt, nombre de jours de retard.

- ① a. Un **filtrage** sur la condition « retard > 0 ».
- ② b. **Filtrer** les retards, **trier** par ordre décroissant, puis conserver les **cinq premiers**.
- ③ c. **Non**, sauf si les **métadonnées** l'indiquent. Une case vide peut signifier « aucun retard » ou « information non saisie » : les traiter comme des zéros fausse la moyenne dans le second cas.
- ④ d. **Regrouper** les enregistrements par catégorie, puis **calculer** la moyenne de la colonne retard dans chaque groupe — après avoir décidé du sort des cases vides.

3 Un établissement publie un fichier « anonymisé » : le nom est supprimé, mais il reste la date de naissance, le sexe, le code postal et la filière.

- ① a. **Non**. Ces quatre descripteurs combinés désignent souvent une **seule personne** dans une commune donnée : elle reste **identifiable**, ce qui suffit à faire de ces **données** des données personnelles.
- ② b. En **croisant** avec une autre source — un trombinoscope, une liste d'admis, un profil public — qui partage une ou deux de ces colonnes. Chaque recoupement réduit le nombre de candidats jusqu'à un seul.
- ③ c. Ne garder que l'**année** de naissance au lieu de la date complète, remplacer le code postal par le **département**, et **regrouper** les filières. On dégrade volontairement la précision pour casser l'unicité des combinaisons.
- ④ d. La **minimisation** : ne collecter que les descripteurs **nécessaires** à l'usage annoncé. Ce qui n'a pas été collecté ne peut pas fuir.
- ⑤ e. Parce que le risque ne tient pas à l'intention mais à la **structure des données**. Un fichier publié échappe à son auteur, et le croisement peut être fait par n'importe qui, plus tard, avec des sources qui n'existaient pas au moment de la publication.

4 Big data et interprétation. Analyser.

- ① a. Un ensemble de **données** dont le **volume**, la **vitesse** d'arrivée et la **variété** dépassent ce qu'un traitement classique peut absorber.
- ② b. L'étude classique part d'une **hypothèse** et la teste. Le **big data** cherche des **corrélations** dans les données **sans hypothèse préalable**, et formule l'explication après coup.
- ③ c. **Non**. Deux grandeurs peuvent varier ensemble parce qu'une **troisième** les influence toutes deux, ou par pur hasard. La corrélation constate, elle n'explique pas.
- ④ d. Parce qu'en testant un très grand nombre de paires, certaines coïncident **par hasard**. Plus on cherche, plus on trouve — et ce qu'on trouve ainsi ne se reproduit pas sur d'autres données. C'est pourquoi une corrélation exhumée d'un grand jeu de données doit être **vérifiée ailleurs** avant d'être publiée.

Corrige-toi toi-même. Compte tes points exercice par exercice : c'est ainsi qu'on voit où l'on perd, et ce n'est presque jamais là où on le croit.

Ce qui est évalué	Points	Ce qui fait perdre le point
Décrire une table et ses descripteurs	3 pts	Mélanger deux natures d'information dans une colonne

Distinguer CSV et JSON	3 pts	Écrire un CSV pour des données imbriquées
Expliquer le rôle des métadonnées	4 pts	Interpréter un nombre sans unité
Montrer les limites d'une anonymisation par suppression du nom	5 pts	Déclarer anonyme un fichier réidentifiable
Nommer des principes du RGPD	3 pts	Réduire le RGPD au consentement aux cookies
Distinguer corrélation et causalité	2 pts	Conclure à une cause à partir d'une corrélation