Méthodologie des appariements de données individuelles

Lucas Malherbe, Insee

Insee Mesurer pour comprendre

DEUX OBJECTIFS

1) Définir les concepts et présenter les principales méthodes en suivant les étapes d'un processus d'appariement

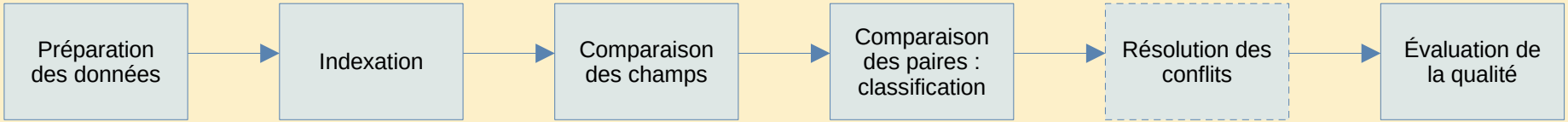
2) Proposer un ensemble de conseils pratiques issus de divers appariements

- Un appariement consiste à rapprocher deux bases de données d'origine distincte partageant des unités statistiques communes mais contenant des informations différentes.
- La tâche est aisée si les deux bases disposent d'un identifiant unique commun ou avec des données de parfaite qualité. Autrement, la tâche est complexe et l'intérêt méthodologique réel.
- Différents cas d'usage :
  - Identification à un fichier de référence exhaustif
  - Appariement de deux fichiers ne couvrant pas tout à fait la même population
  - Dédoublonnage (~ apparier un fichier avec lui-même)
- Différents objectifs : administratif or statistique ?

1 ÉTAPES D'UN APPARIEMENT

2 CONSEILS PRATIQUES

01 ÉTAPES D'UN APPARIEMENT



CONTEXTE

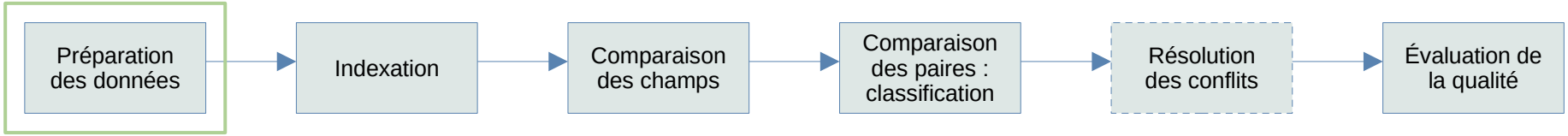
- Première étape indispensable, la préparation des données conditionne la réussite de l'appariement.

OBJECTIF

- Nettoyer, normaliser et préparer les données pour l'appariement

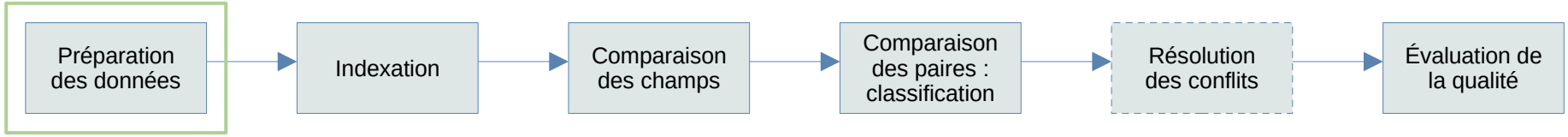
ÉCUEIL

- Une normalisation trop brutale conduit à une baisse significative de la variance : il y a une perte d'information.



EXEMPLES DE TRAITEMENTS

- **Capitalisation des lettres** (Dupont → DUPONT)
- **Suppression des caractères superflus** (JEAN-MICHEL L'HÉRITIER → JEANMICHEL LHERITIER)
- **Suppression des mots vides (*stop words*)** (le, la, de, du, ce)
- **Prise en compte des variations d'écriture communes** (av. → avenue)
- **Contrôle et correction des anomalies** (age > 120, ou négatif)
- **Segmentation de l'information** (date de naissance = 13/01/1970 → jour = 13, mois = 01 et année = 1970)



CONTEXTE

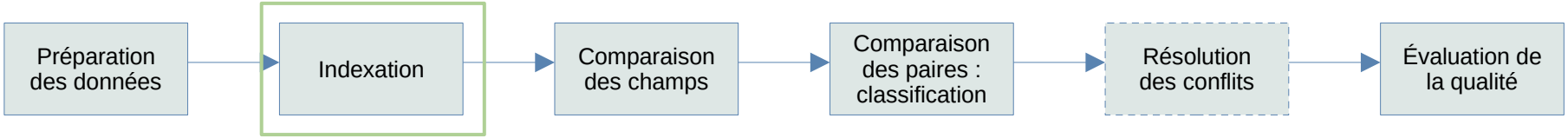
- Il est souvent impossible de comparer l'ensemble des paires possibles pour effectuer l'appariement (problème en $O(N^2)$).
- La plupart des paires sont très faciles à écarter.

OBJECTIF

- Réduire la dimension en ne considérant que les paires qui ont une chance raisonnable d'être une paire d'individus identiques

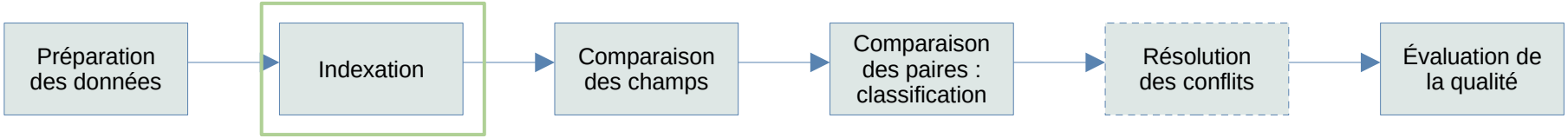
ÉCUEIL

- Une indexation trop drastique crée des faux négatifs : certaines paires d'individus identiques sont rejetées à cause de l'indexation. Il y a un compromis à trouver entre réduction de la dimension et qualité de l'appariement.



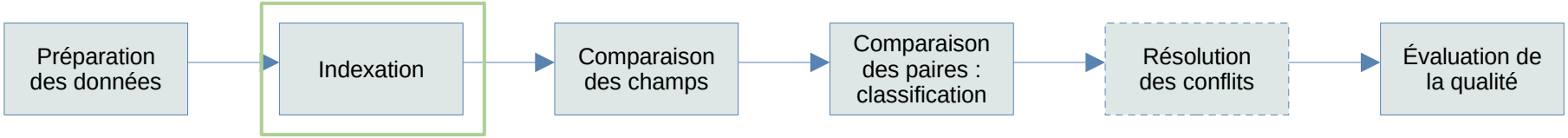
MÉTHODES

- Approche classique : **le blocage**
  - L'un des champs est choisi comme clé de blocage
  - Seules les paires d'individus possédant la même valeur sur la clé de blocage sont conservées comme paires potentielles.
  - Ex : si la clé de blocage est l'année de naissance, seuls les individus nés la même année seront comparés.
  - La clé de blocage doit être d'excellente qualité pour éviter de créer des faux négatifs.



MÉTHODES

- Variations autour du blocage
  - Combiner plusieurs champs pour construire la clé de blocage
  - Utiliser une distance à la place ou en complément d'une comparaison exacte
- Approche du voisinage trié
  - Les fichiers sont triés selon une clé de tri.
  - Une fenêtre glissante de taille fixe permet de sélectionner les paires avec une valeur proche sur la clé de tri.



CONTEXTE

- Après l'indexation, il faut examiner en détail les paires restantes. Pour la plupart des champs, une comparaison exacte ne suffit pas.

OBJECTIF

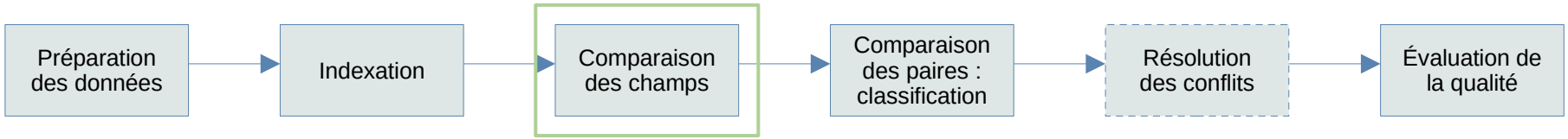
- Calculer des mesures de similarité pour chaque champ identifiant et chaque paire potentielle conservée à l'issue de l'indexation

ÉCUEIL

- Cette étape est intimement liée à la suivante. Certaines mesures de similarité sont plus adaptées à certains algorithmes de classification.

MÉTHODES

- Le choix de la mesure dépend du type de champ (texte, nombre, date, adresse)
- Pour les champs textuels, le choix est large (Levenshtein, Jaro-Winkler, Editex, n-grams…)
- Après normalisation, chaque mesure de similarité est comprise entre 0 et 1 (1 pour des valeurs identiques et 0 pour des valeurs totalement différentes)



CONTEXTE

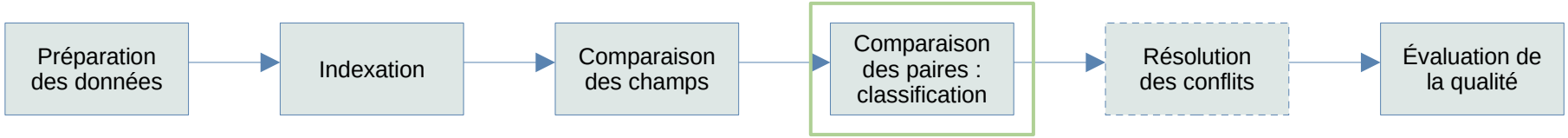
- Les mesures de similarité calculées à l'étape précédente sont mobilisées pour décider du statut de chaque paire.

OBJECTIF

- Classer les paires retenues après l'indexation en deux catégories : les paires liées et les paires non-liées

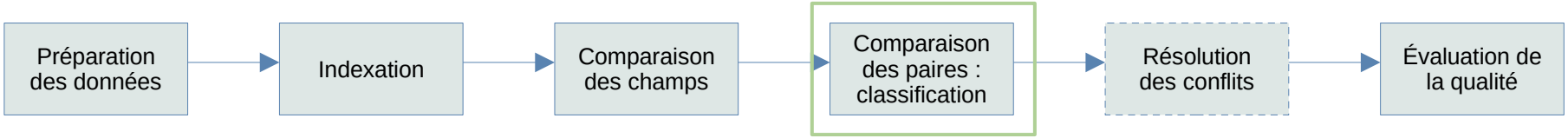
MÉTHODES

- Deux types d'approches : déterministe ou probabiliste



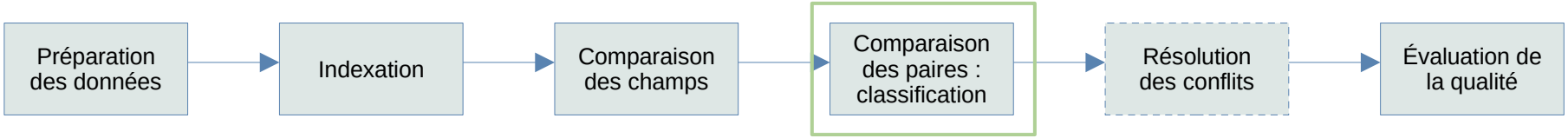
EXEMPLES DE MÉTHODES DÉTERMINISTES

- **Tours de clés successifs**
  - Succession d'étapes en commençant par des règles strictes et en relâchant progressivement les contraintes.
  - Les individus appariés à une étape ne sont plus considérés pour les étapes suivantes.
- **Méthode du plus proche écho**
  - Une moyenne pondérée des mesures de similarité donne un score pour chaque paire.
  - Si le score dépasse un seuil, la paire est liée.
- ***Machine learning* supervisé**
  - Cette méthode nécessite un échantillon de paires annotées.
  - Un algorithme de machine learning (ex : SVM ou arbre de décision) apprend à prédire le statut de nouvelles paires à partir de leurs mesures de similarité.



APPROCHE PROBABILISTE

- Les méthodes probabilistes dérivent toutes du cadre décrit par Fellegi et Sunter (1969).
- Elles se caractérisent par un processus d'inférence bayésienne qui conduit au calcul d'une probabilité pour chaque paire.
- Le modèle repose sur deux probabilités conditionnelles, m et u, calculées pour chaque variable identifiante :
  - m mesure la qualité des données ;
  - u représente la probabilité d'observer la même valeur par chance pour deux individus pris au hasard.
- Des poids sont ensuite calculés. Ils représentent le pouvoir prédictif de chaque champ pour déterminer le statut d'une paire.
- L'estimation des paramètres s'effectue de manière non supervisée via l'algorithme Espérance-Maximisation.

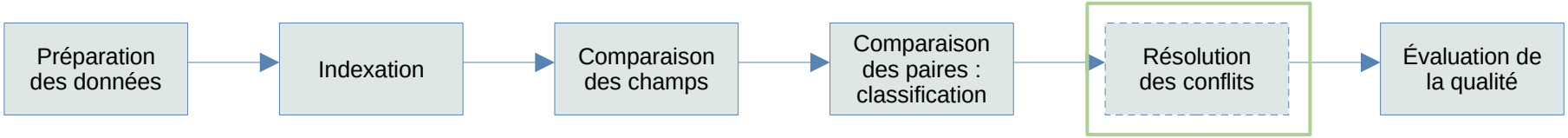


CONTEXTE

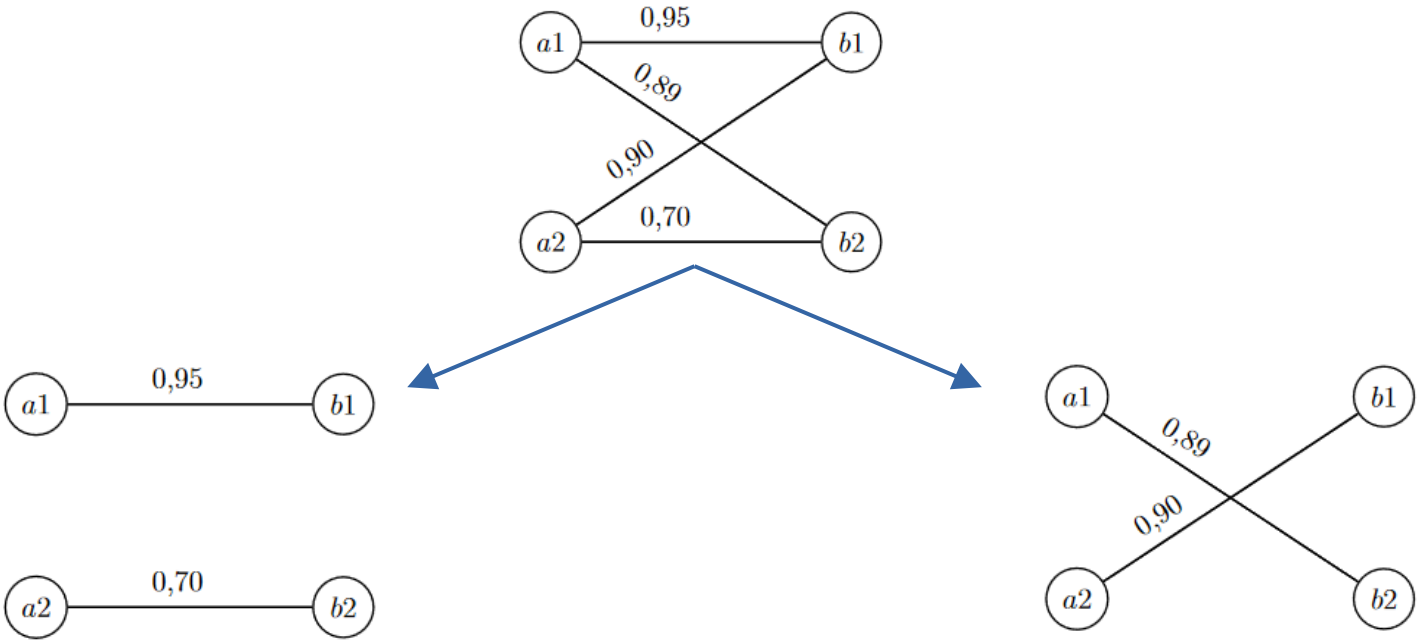
- Les algorithmes de classification traitent généralement les paires de façon indépendante, mais il existe souvent des restrictions sur les combinaisons de paires pouvant être liées.

MÉTHODES

- Algorithme glouton
- Recherche d'une solution optimale

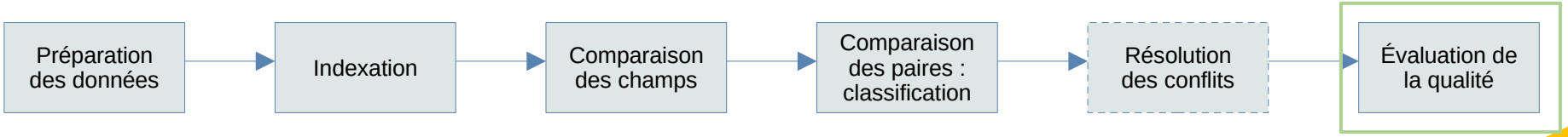


Exemple de conflit



Méthode gloutonne

Solution optimale



ÉVALUATION DE LA QUALITÉ (1/2)

CONTEXTE

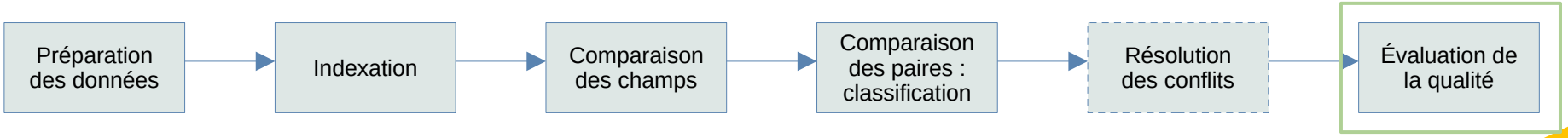
- L'évaluation de la qualité est une étape essentielle du processus, particulièrement lorsque des études reposent sur les résultats de l'appariement.

OBJECTIF

- Obtenir le plus d'informations possibles sur la qualité de l'appariement qui vient d'être effectué

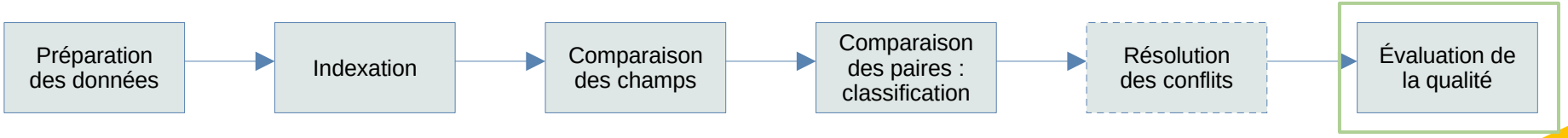
ÉCUEIL

- La plupart des indicateurs de qualité nécessitent un échantillon de paires annotées.
  - Étalon-or : échantillon représentatif de paires dont le statut réel est connu
  - Échantillon annoté manuellement



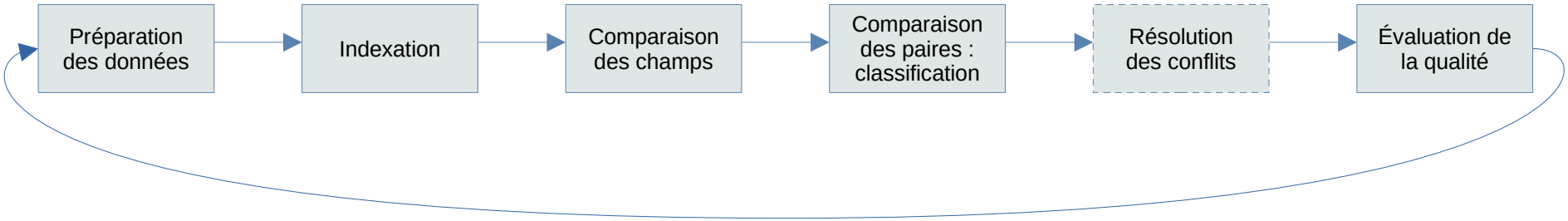
INDICATEURS

- Taux d'appariement
- Indicateurs des problèmes de classification binaire :
  - Vrais / faux positifs, vrais / faux négatifs
  - Précision, rappel et F-mesure
- Analyse de la distribution des paires liées, des paires non liées, ainsi que des paires mal classées.
- Évaluation de l'impact des erreurs sur les études subséquentes



02 CONSEILS PRATIQUES

1. Réussir un appariement prend du temps



- **Mettre au point un appariement est un processus itératif.**
- **Aucun algorithme n'est applicable de A à Z sur tous les fichiers : il faut prendre en compte leurs spécificités.**
- **Des indicateurs de qualité fiables participent à l'amélioration de l'appariement.**

2. Prendre au sérieux l'évaluation de la qualité

- Ne pas avoir une confiance aveugle en les résultats de l'appariement
- Le taux d'appariement n'est pas suffisant.
- Annoter un échantillon représentatif (via un tirage stratifié sur le score de chaque paire, par exemple)
- Fixer des contraintes de qualité à vérifier

3. Prendre en compte les principaux cas particuliers

- Le choix de la distance (Levenshtein, Jaro-Winkler...) a un impact modéré.
- Ce qui fait la différence, c'est d'adapter les mesures de similarité aux cas particuliers. Par exemple :
  - Interversion nom / prénom
  - Changement de nom suite à un mariage

4. Les appariements de fichiers volumineux

- Volume critique : 100 000 à 1M de lignes
- 2 limites informatiques :
  - La mémoire vive
  - Le temps de calcul
- Pistes de solutions
  - Procéder à une indexation très stricte, ou faire plusieurs tours d'appariement en relâchant progressivement l'indexation
  - Utiliser des outils spécifiques pour les gros volumes (ex : Spark)
  - Utiliser un moteur de recherche textuelle (ex : ElasticSearch)
  - Limiter les comparaisons floues

5. Choisir une méthode d'appariement

- Outil clés en main → probabiliste
- À développer soi-même → tours de clés successifs ou plus proche écho
- Il existe un échantillon annoté → *machine learning*
- Le plus performant → ...