Introduction à la science des données

Instructeur: Patrick Boily

uOttawa

Institut de développement professionnel
Professional Development Institute

COLLECTE ET GESTION DES DONNÉES

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics
pboily@uottawa.ca

[en collaboration avec Jen Schellinck | Sysabee]

OBJECTIF

Nous recherchons des données qui peuvent :

- fournir un **aperçu légitime** de notre système d'intérêt ;
- fournir des réponses **correctes** et **précises** aux questions pertinentes ;
- **soutenir** l'élaboration de conclusions **valables**, avec la capacité de **qualifier/quantifier** ces conclusions en termes de portée et de précision.

Cela ne peut se faire sans la mise en place d'un **plan d'étude** : quelles données devons-nous collecter, et comment les collecter ?

MOTIVATIONS POUR LA COLLECTE DE DONNÉES

Trois fonctions, historiquement :

- la tenue de registres (gestion des personnes/de la société)
- science - nouvelles connaissances générales
- renseignement - affaires, militaire ? police ? social ? domestique ? personnel ?

Chacune de ces trois fonctions utilise des sources d'information différentes.

- ils ont collecté différents types de données
- ils ont également des cultures de données et des terminologies différentes

LES DONNÉES SONT RÉELLES

Les données sont une représentation, mais les données sont **physiques**.

Elles ont des propriétés physiques, elles nécessitent un espace physique et de l'énergie pour être utilisées.

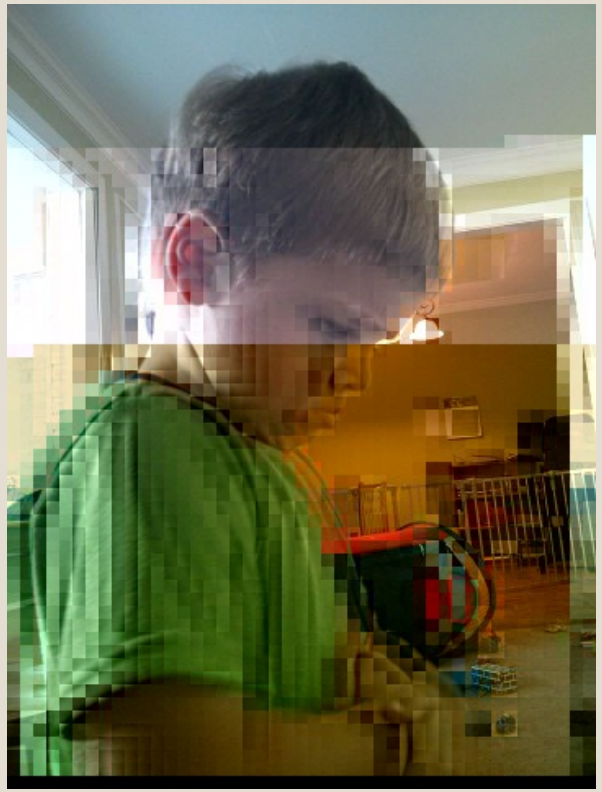
DÉGRADATION DES DONNÉES

Les données vieillissent; elles ont une **date d'expiration**.

« Données pourries » ou « données en décomposition » :

- **littéralement** – le support de stockage des données peut se détériorer
- **métaphoriquement**, lorsque les données **ne représentent plus** fidèlement les objets et les relations pertinents, voire lorsque ces objets n'existent plus de la même manière.

Les données doivent rester « fraîches » et « actuelles », et non « périmées » (selon le contexte et le modèle !).

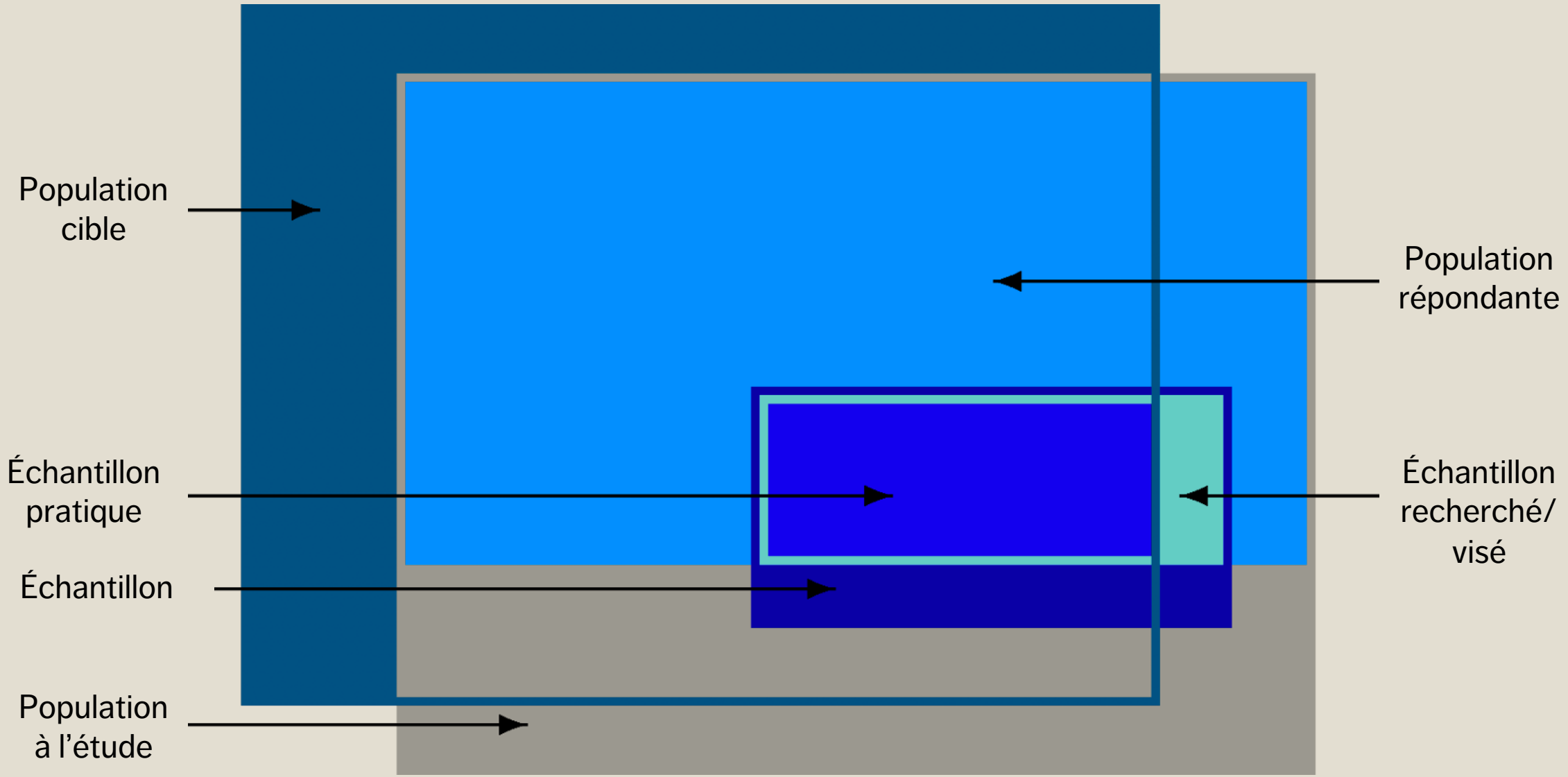


ÉCHANTILLONNAGE NON PROBABILISTE ET « PÊCHE » AUX TENDANCES

Deux situations distinctes peuvent s'associer pour causer des **problèmes** d'analyse des données :

- la formulation de conclusions (inférences) à partir d'un échantillon de population qui ne se justifient pas par la méthode de collecte de l'échantillon (symptomatique d'un échantillonnage non probabiliste)
- la recherche d'un quelconque schéma dans les données, suivie d'une formulation d'explications *a posteriori* concernant ces schémas

Seules ou combinées, ces deux situations conduisent à des conclusions médiocres (et **potentiellement nuisibles**).


Population cible
Population répondante
Échantillon pratique
Échantillon recherché/ visé
Échantillon
Population à l'étude

ÉTAPES DE L'ÉTUDE/DE L'ENQUÊTE

Les enquêtes suivent les mêmes étapes générales :

1. énoncé de l'objectif
2. sélection de la cadre d'enquête
3. plan d'échantillonnage
4. conception du questionnaire
5. collecte des données
6. saisie et codage des données
7. traitement des données et imputation
8. estimation
9. analyse des données
10. diffusion
11. documentation

Le processus n'est pas toujours linéaire, mais il y a un mouvement défini de **l'objectif à la diffusion.**

ÉCHANTILLONNAGE NON PROBABILISTE

Les méthodes d'**échantillonnage non probabiliste** (ENP) sélectionnent les unités d'échantillonnage de la population cible à l'aide d'approches subjectives et non aléatoires.

- Les ENP ont le mérite d'être rapide, relativement peu coûteux et pratique.
- Les ENP sont idéales pour l'analyse exploratoire et l'élaboration des enquêtes.

On a souvent recours aux ENP au lieu des échantillonnages probabilistes (**problématique**).

- Le biais de sélection associé rend les ENP peu sûres en matière d'inférences
- La collecte automatisée des données tombe souvent dans le champ des ENP – il est toujours possible d'analyser les données recueillies selon ces méthodes, mais pas nécessairement de généraliser les résultats à la population cible.

ÉCHANTILLONNAGE PROBABILISTE

Les plans d'échantillonnage probabiliste sont généralement plus **difficiles** et plus **coûteux** à mettre en place (car ils requièrent une base d'enquête de qualité), et ils prennent plus de temps à réaliser.

Ils fournissent des **estimations fiables** de la caractéristique d'intérêt et de **l'erreur d'échantillonnage**, ouvrant la voie à l'utilisation de petits échantillons pour tirer des inférences sur des populations cibles plus vastes (en théorie, du moins, les composantes de l'erreur non attribuable à l'échantillonnage peuvent tout de même jouer sur les résultats et la généralisation).

PLANS D’ÉCHANTILLONNAGE

Les différents **plans d'échantillonnage** présentent des avantages et des désavantages distincts.

Ils peuvent être utilisés pour calculer des estimations

- pour divers attributs de la population : moyenne, total, proportion, rapport, différence, etc.
- pour les intervalles de confiance à 95% correspondants.

Nous pourrions également vouloir calculer les tailles d'échantillon pour une **limite d'erreur** donnée (une limite supérieure du rayon de l'intervalle de confiance à 95% souhaité), et comment déterminer la **répartition de l'échantillon** (combien d'unités à échantillonner dans les différents groupes de sous-population).

PLANS D'ÉCHANTILLONNAGE PROBABILISTES

Échantillonnage aléatoire simple (EAS)

Échantillonnage aléatoire stratifié (STR)

Échantillonnage systématique (SYS)

Échantillonnage en grappes (EPG)

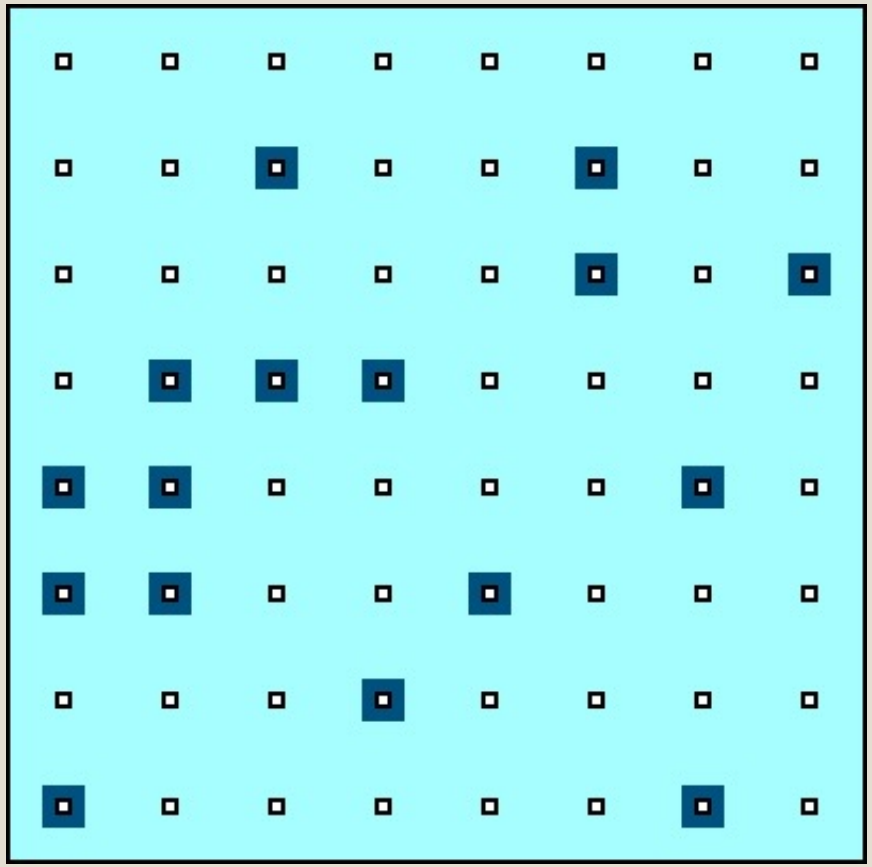
Échantillonnage avec probabilité proportionnelle à la taille (PPT)

Échantillonnage répété (REP)

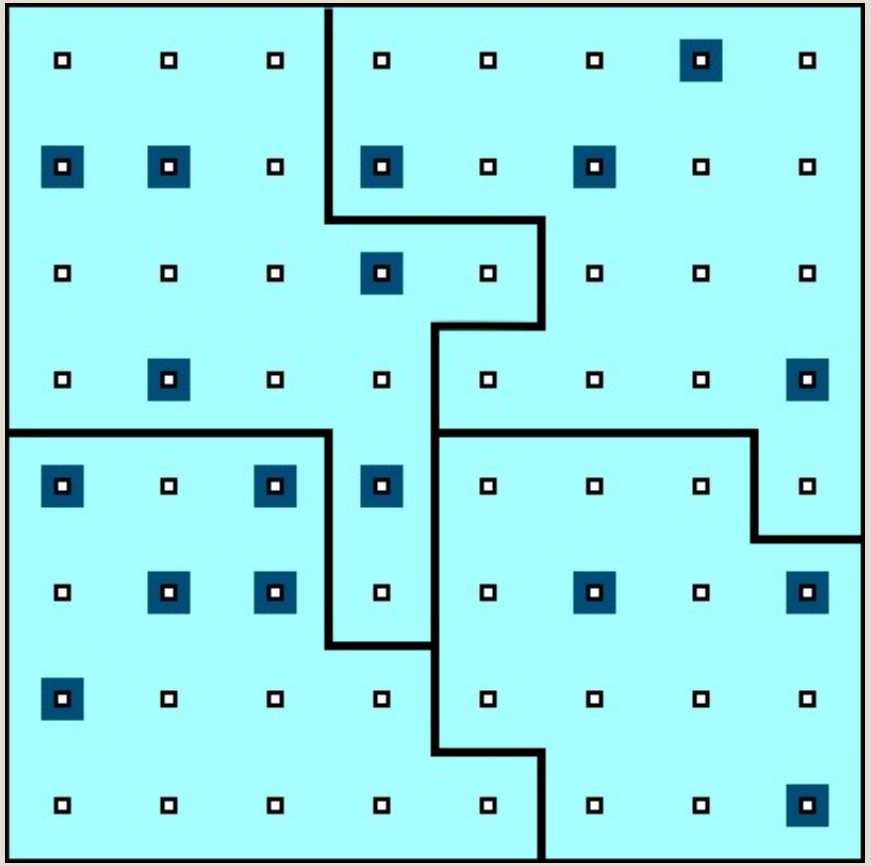
Échantillonnage à plusieurs degrés (EPD)

Échantillonnage à plusieurs phases (EPP)

PLANS D'ÉCHANTILLONNAGE

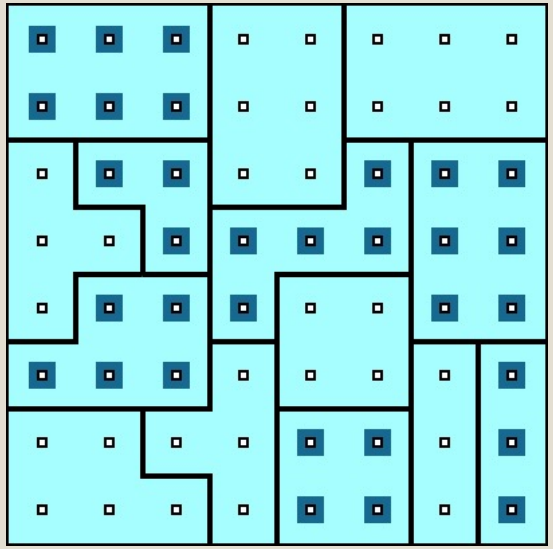


Échantillonnage
aléatoire simple

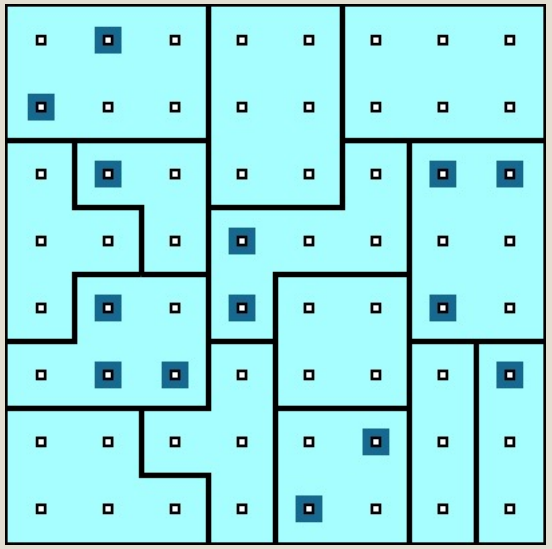


Échantillonnage
aléatoire stratifié

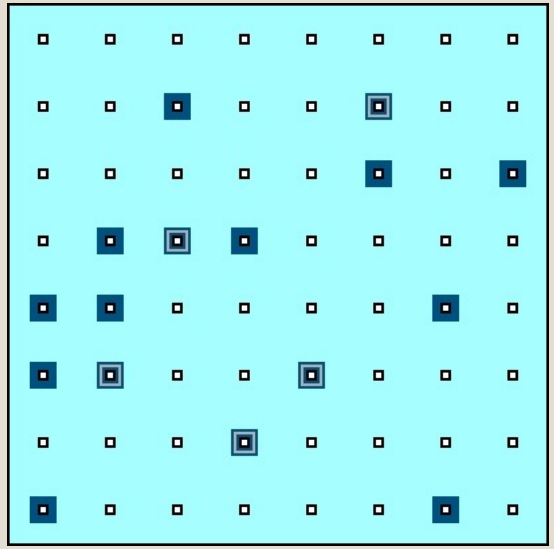
AUTRES PLANS D'ÉCHANTILLONNAGE



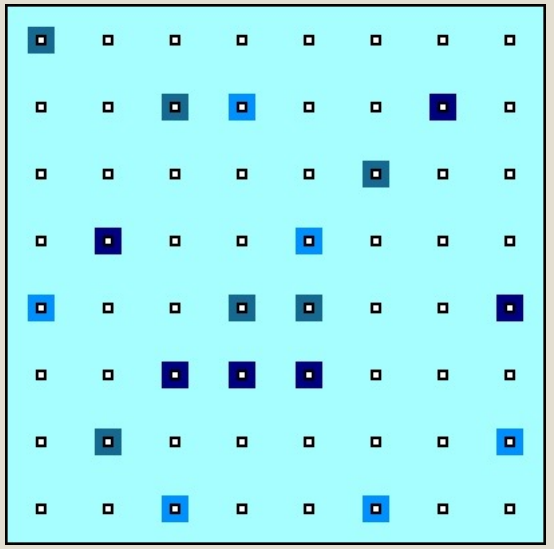
Échantillonnage en grappes



Échantillonnage à plusieurs degrés



Échantillonnage à plusieurs phases



Échantillonnage répété

LISTE DE VÉRIFICATION APPLICABLE À LA COLLECTE AUTOMATISÉE

Le **moissonnage du Web** ou est-il absolument nécessaire?

Critères :

- Prévoyez-vous répéter l'opération de temps à autre, p. ex. pour mettre à jour votre base de données?
- Désirez-vous que d'autres puissent reproduire votre processus de collecte des données?
- Traitez-vous fréquemment avec des sources de données en ligne?
- La tâche est-elle non négligeable en termes de portée et de complexité?
- Si la tâche peut être effectuée manuellement, manquez-vous de ressources pour laisser les autres faire le travail ?
- Êtes-vous prêt à automatiser le processus par le biais de la programmation ?

Si la plupart des réponses sont "Oui", alors le recouvrement automatisé peut être le bon choix.

MOISSONNAGE DU WEB – QUALITÉ DES DONNÉES

Informations de première main : par exemple, un tweet ou un article de presse.

Données de seconde main : données qui ont été copiées à partir d'une source hors ligne ou récupérées ailleurs.

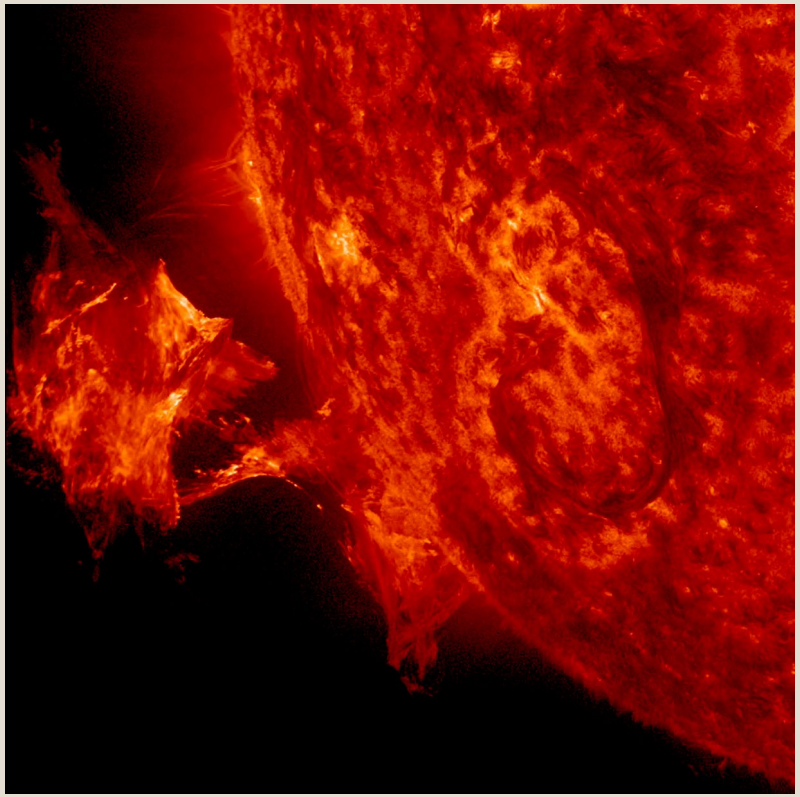
- Parfois, il est impossible de se rappeler ou de retracer la source de ces données.
- Est-il encore utile de les utiliser ? Cela dépend.

Toute utilisation de données secondaires nécessite un **recoupement** et une **validation**.

DONNÉES STRUCTURÉES PAR RAPPORT AUX DONNÉES NON STRUCTURÉES

La disponibilité croissante de données non structurées et de grands objets binaires « **blob** » est l'une des principales motivations de certains des nouveaux développements dans les types de bases de données et autres stratégies de stockage de données.

- **Données structurées** : étiquetées, organisées, discrètes, selon une structure limitée et prédéfinie
- **Données non structurées** : non organisées, pas de modèle de données structuré prédéfini précis – p. ex. texte dans un document
- **Données « blob »** : grand objet binaire – images, audio, multimédia



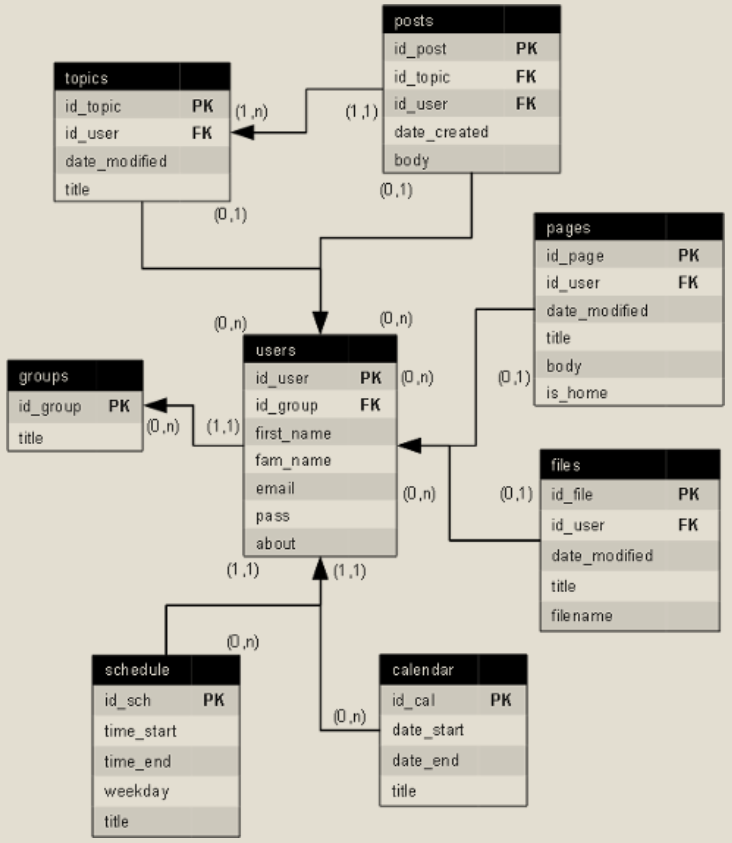
BASES DE DONNÉES RELATIONNELLES

Données stockées dans une série de **tableaux**.

En gros, chaque tableau représente un objet et des propriétés liées à cet objet.

Des colonnes spéciales dans les tables **relient** les instances d'objets entre les tables (ce qui permet les fusions).

L'approche traditionnelle du stockage des données.



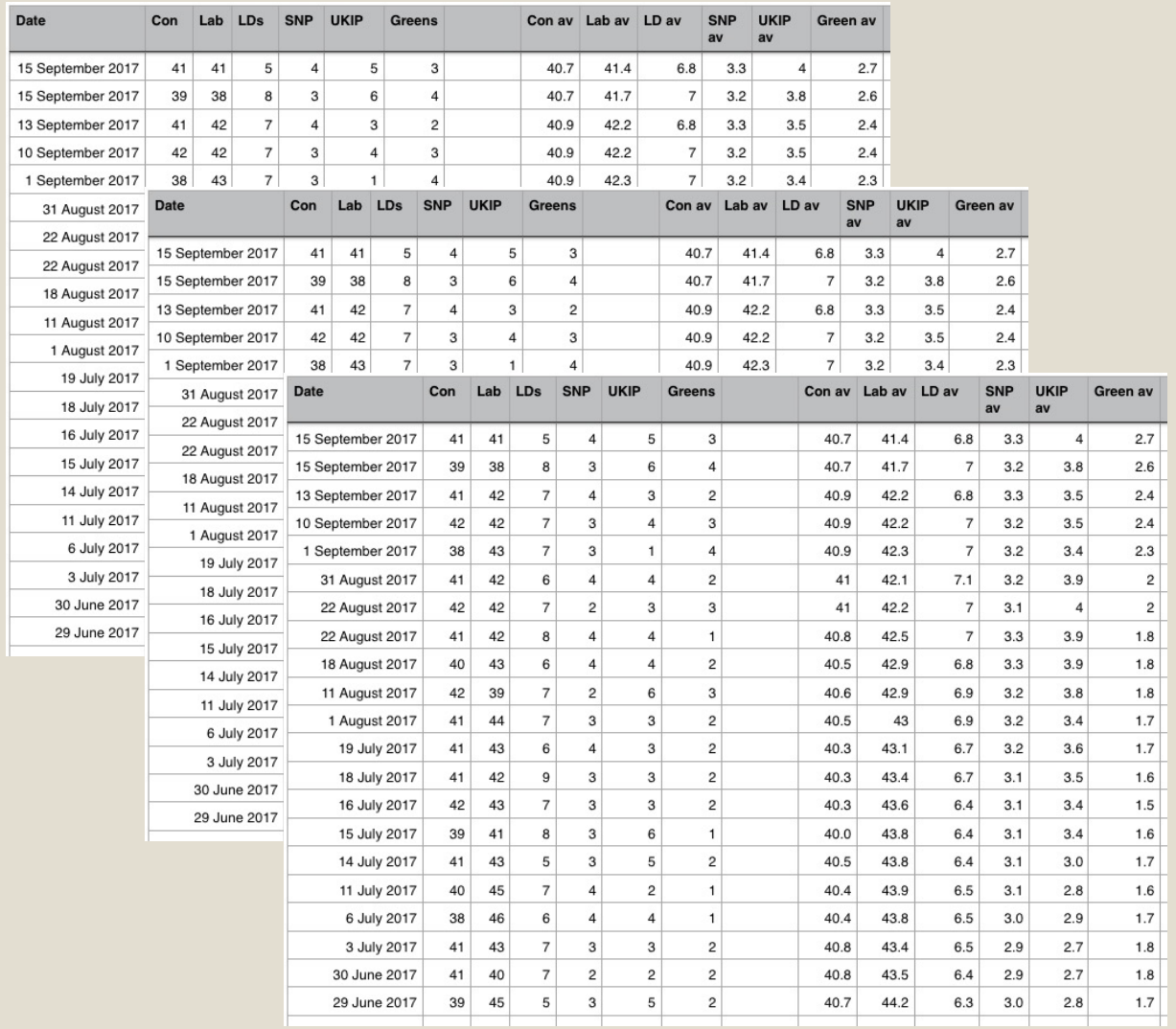
FICHIERS NON HIÉRARCHIQUES ET LES FEUILLES DE CALCUL

Qu'en est-il de la conservation de vos données dans un seul tableau géant (feuille de calcul)?

Ou plusieurs feuilles de calcul?

Ça ne peut pas être si terrible que ça!

Wayne Eckerson a inventé le terme « spreadmart » pour décrire une situation où de nombreuses feuilles de calcul (*ad hoc*) constituent une stratégie de données.



Date	Con	Lab	LDs	SNP	UKIP	Greens		Con av	Lab av	LD av	SNP av	UKIP av	Green av
15 September 2017	41	41	5	4	5	3		40.7	41.4	6.8	3.3	4	2.7
15 September 2017	39	38	8	3	6	4		40.7	41.7	7	3.2	3.8	2.6
13 September 2017	41	42	7	4	3	2		40.9	42.2	6.8	3.3	3.5	2.4
10 September 2017	42	42	7	3	4	3		40.9	42.2	7	3.2	3.5	2.4
1 September 2017	38	43	7	3	1	4		40.9	42.3	7	3.2	3.4	2.3
31 August 2017	41	42	6	4	4	2		41	42.1	7.1	3.2	3.9	2
22 August 2017	42	42	7	2	3	3		41	42.2	7	3.1	4	2
22 August 2017	41	42	8	4	4	1		40.8	42.5	7	3.3	3.9	1.8
18 August 2017	40	43	6	4	4	2		40.5	42.9	6.8	3.3	3.9	1.8
11 August 2017	42	39	7	2	6	3		40.6	42.9	6.9	3.2	3.8	1.8
1 August 2017	41	44	7	3	3	2		40.5	43	6.9	3.2	3.4	1.7
19 July 2017	41	43	6	4	3	2		40.3	43.1	6.7	3.2	3.6	1.7
18 July 2017	41	42	9	3	3	2		40.3	43.4	6.7	3.1	3.5	1.6
16 July 2017	42	43	7	3	3	2		40.3	43.6	6.4	3.1	3.4	1.5
15 July 2017	39	41	8	3	6	1		40.0	43.8	6.4	3.1	3.4	1.6
14 July 2017	41	43	5	3	5	2		40.5	43.8	6.4	3.1	3.0	1.7
11 July 2017	40	45	7	4	2	1		40.4	43.9	6.5	3.1	2.8	1.6
6 July 2017	38	46	6	4	4	1		40.4	43.8	6.5	3.0	2.9	1.7
3 July 2017	41	43	7	3	3	2		40.8	43.4	6.5	2.9	2.7	1.8
30 June 2017	41	40	7	2	2	2		40.8	43.5	6.4	2.9	2.7	1.8
29 June 2017	39	45	5	3	5	2		40.7	44.2	6.3	3.0	2.8	1.7