Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction à l'analyse des données

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 ?

SOURCES DES DONNÉES

COLLECTE ET GESTION DES DONNÉES

QUESTIONS FONDAMENTALES

Pourquoi collectons-nous des données ?

Que peut-on faire avec les données ?

D'où viennent les données ?

À quoi ressemble une collection de données ?

Comment pourrait-on la décrire ?

Devons-nous faire la distinction entre données, informations et connaissances ?

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

CULTURES DE DONNÉES ET TERMES

Intelligence économique :

- entrepôt de données + data mart
- « dimension » des données (= ensemble de données)
- données hiérarchiques (tranches)
- élément de données
- table de dimensions + table de faits

Sciences/Statistiques :

- données expérimentales
- essais
- participants
- variables
- corrélation

Gestion des dossiers :

- architecture de l'information
- plan de classement
- ressource d'information
- champ
- forme et sujet

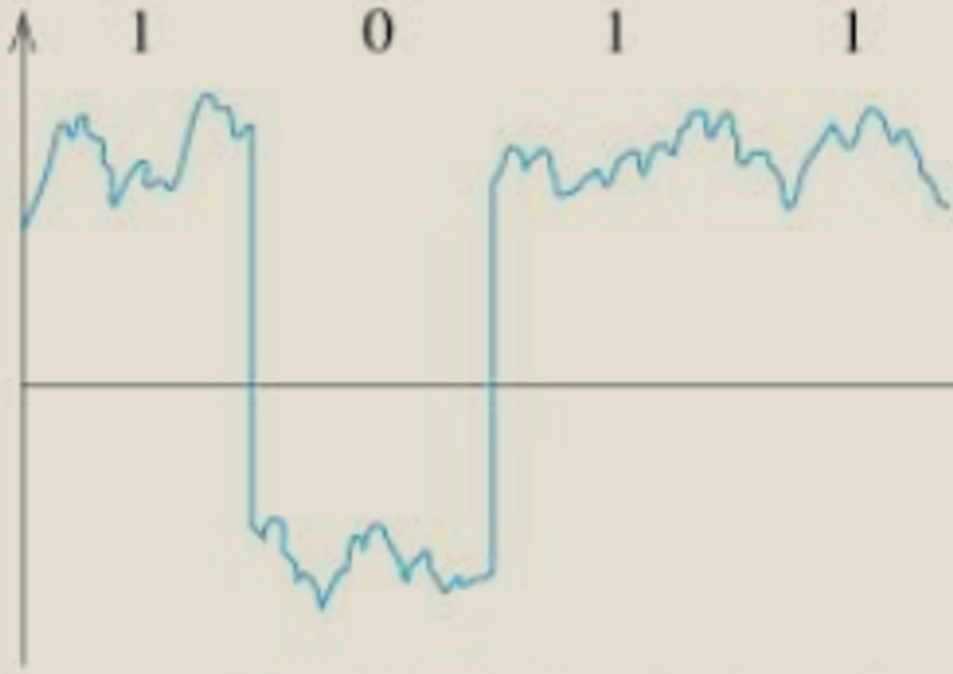
Données **Connaissances** **Information**

ORDINATEURS ET DONNÉES

L'informatique/science de l'information a son propre point de vue théorique et **fondamental** sur les données et l'information.

Les données deviennent numériques : les ordinateurs opèrent sur des données au sens fondamental du terme - des 1 et des 0 représentant des chiffres, des lettres, etc.

D'un point de vue pragmatique, les données sont enregistrées sur des ordinateurs et sont accessibles via des réseaux mondiaux.



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.

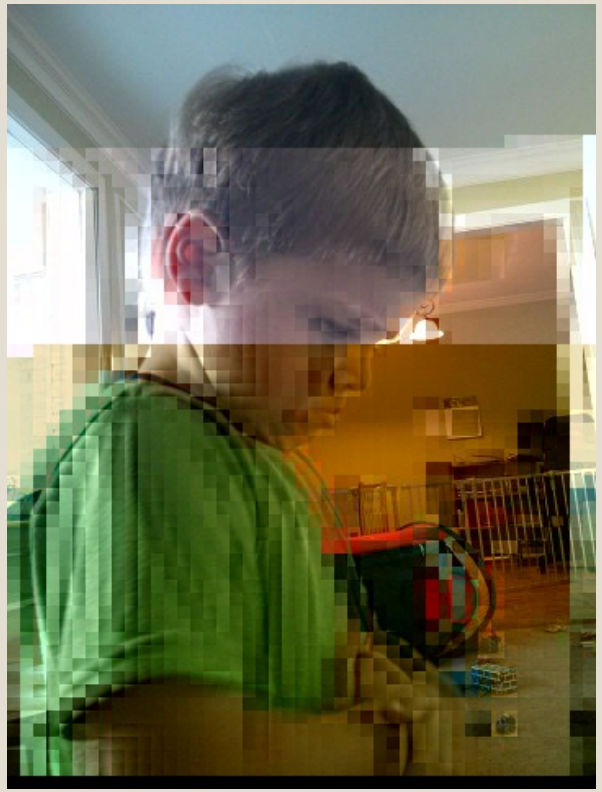
DÉGRADATION DES DONNÉES

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 !).



THÉORIE DE L'ÉCHANTILLONNAGE ET PLANIFICATION D'ÉTUDE

COLLECTE ET GESTION DES DONNÉES

« La dernière enquête montre que 3 personnes sur 4 représentent 75% de la population. »

D. Letterman

[G. Smith, The Exaggerated Promise of So-Called Unbiased Data Mining]

« À l'aide d'un appareil d'imagerie par résonance magnétique (IRM), un diplômé de Dartmouth a étudié l'activité cérébrale d'un saumon lorsqu'on lui montrait des photographies et qu'on lui posait des questions. L'aspect le plus intéressant de cette recherche, ce n'est pas qu'on ait étudié un saumon, mais que ce saumon était mort. Hé oui! On a acheté un saumon mort au marché local, on l'a placé dans un appareil d'IRM, et on a observé certains schémas. Il y avait inévitablement des schémas, mais ils étaient invariablement dénués de sens. »

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**).

ÉTUDES, ENQUÊTES ET MODÈLES D'ÉCHANTILLONNAGE

Une **enquête** est une activité qui consiste à recueillir de l'information sur des caractéristiques d'intérêt :

- de manière **organisée** et **méthodique**;
- sur une partie ou la totalité des **unités** d'une population;
- à l'aide de concepts, de méthodes et de procédures **bien définis**;
- grâce à la compilation de renseignements sous forme d'un résumé **significatif**.

Un **recensement** est une collecte de données sur toutes les unités d'une population; une **enquête sur échantillon** n'utilise qu'une fraction des unités.

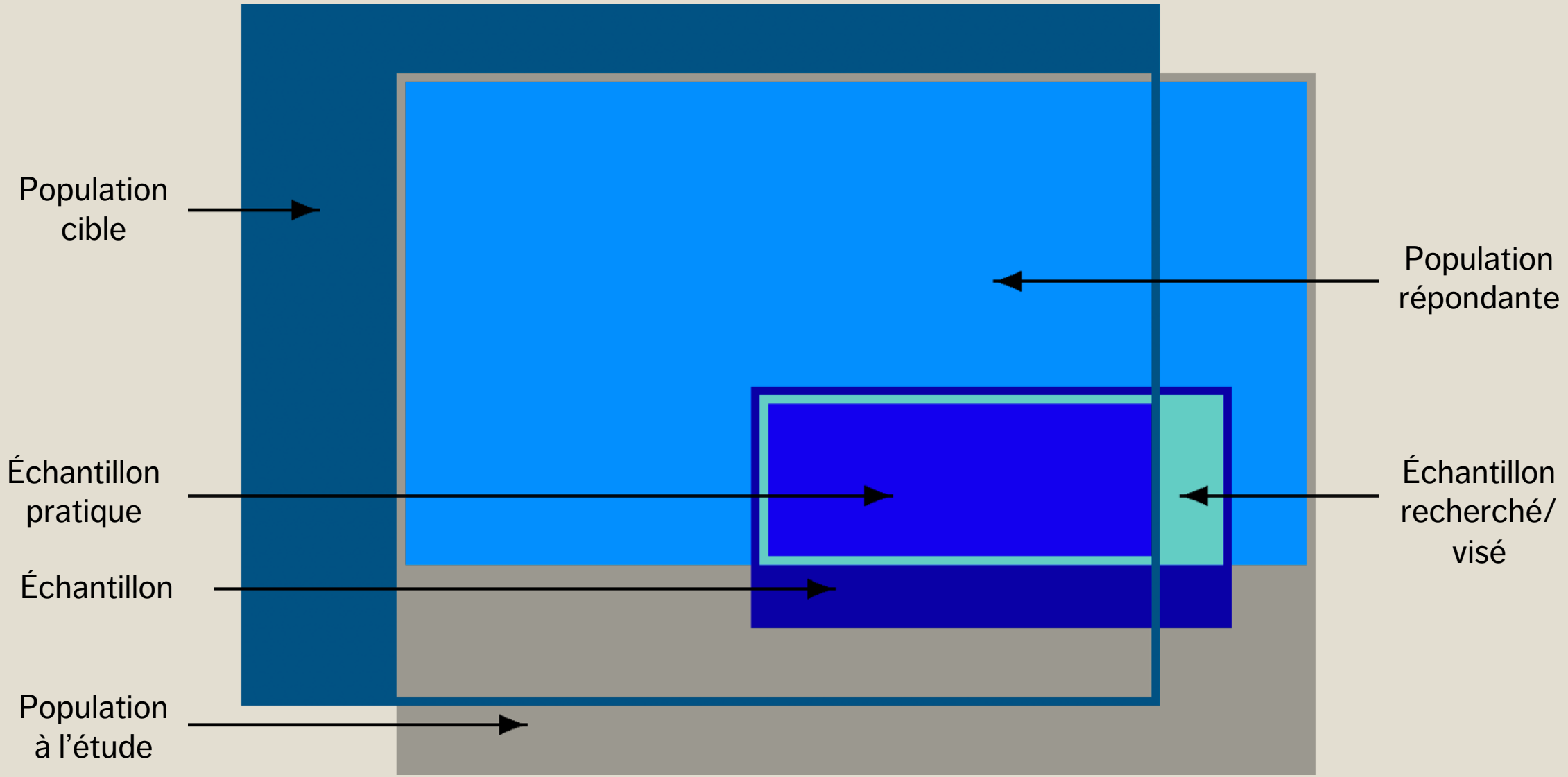
Lorsque l'échantillonnage est effectué correctement, il est possible de recourir à diverses **méthodes statistiques** pour faire des **inférences** sur la **population cible** en échantillonnant un (comparativement) petit nombre d'unités dans la **population étudiée**.

FACTEURS DÉTERMINANTS

Parfois, des informations sur l'**ensemble** de la population sont nécessaires afin de répondre aux questions ; à d'autres moments, ce n'est pas nécessaire.

Le **type d'enquête** dépend de multiples facteurs :

- le type de question à laquelle il faut répondre ;
- la précision requise ;
- le coût du levé d'une unité ;
- le temps nécessaire à l'enquête sur une unité ;
- la taille de la population étudiée, et
- la prévalence des attributs d'intérêt.


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.**

CADRES D'ENQUÊTE

La **base de sondage** fournit les moyens **d'identifier** et de **contacter** les unités de la population étudiée. Elle est généralement coûteuse à créer et à maintenir.

Idéale, elle contient des données d'identification, de contact, de classification, de maintenance et de liaison. Elle doit minimiser le risque de **sur/sous-dénombrement**, le nombre de dédoublements et d'erreurs de classification .

Une approche d'échantillonnage statistique est contre-indiquée à moins que la base d'enquête choisie ne soit :

- **pertinente** (autrement dit qu'elle corresponde et permette l'accessibilité à la population cible);
- **exacte** (l'information qu'elle contient est valide);
- **opportune** (elle est à jour) et **offerte à un prix compétitif.**

ERREUR D'ENQUÊTE

Erreure totale =

Err. échantillonnage + Err. mesure + Err. non-réponse + Error couverture

- Err. échantillonnage: enquête, pas recensement
- Err. mesure: manque d'exactitude dans la mesure des observations
- Err. non-réponse: non-répondants présentant des différences d'observation systématiques
- Error couverture: dégradation ou corruption de la base

L'échantillonnage statistique permet de fournir des estimations, mais surtout de contrôler, dans une certaine mesure, l'**erreur totale** (ET) des estimations.

Idéalement, $\mathrm{ET} = 0$. Dans la pratique, les deux principaux éléments qui contribuent à l'ET sont: l'**erreur d'échantillonnage** (attribuable au choix du plan d'échantillonnage), et les **erreurs non attribuables à l'échantillonnage** (tout le reste).

ERREUR NON ATTRIBUABLE À L'ÉCHANTILLONNAGE

Dans une certaine mesure, il est possible de contrôler une erreur non attribuable à l'échantillonnage :

- l'**erreur de couverture** peut être réduite au minimum en choisissant des bases d'enquête à jour et de grande qualité;
- l'**erreur de non-réponse** peut être atténuée en choisissant soigneusement le mode de collecte des données et le plan du questionnaire, et au moyen de « rappels » et de « suivis »;
- l'**erreur de mesure** peut être grandement diminuée par une conception minutieuse du questionnaire, un essai préliminaire de l'appareil de mesure et une validation croisée des réponses.

En pratique, ces suggestions ne sont pas si utiles à l'époque moderne.

Cela explique, l'utilisation du **moissonnage du web/échantillonnage non-probabiliste**.

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.

MÉTHODES D'ÉCHANTILLONNAGE NON PROBABILISTE

Dans certains contextes, les méthodes NPS peuvent répondre aux besoins d'un client ou d'une organisation, mais il faut les informer des inconvénients et leur présenter des alternatives probabilistes.

- **Accidentel :** personne de la rue, dépend de la disponibilité des unités et du biais de l'enquêteur.
- **Volontaire :** biais d'auto-sélection
- **Jugement :** biais dû à des idées préconçues inexactes sur la population cible.
- **Quota :** sondage de sortie, ignore le biais de non-réponse.
- **Modifié :** commence par être probabiliste, passe au quota en réaction à des taux de non-réponse élevés.
- **Boule de neige :** système pyramidal

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)

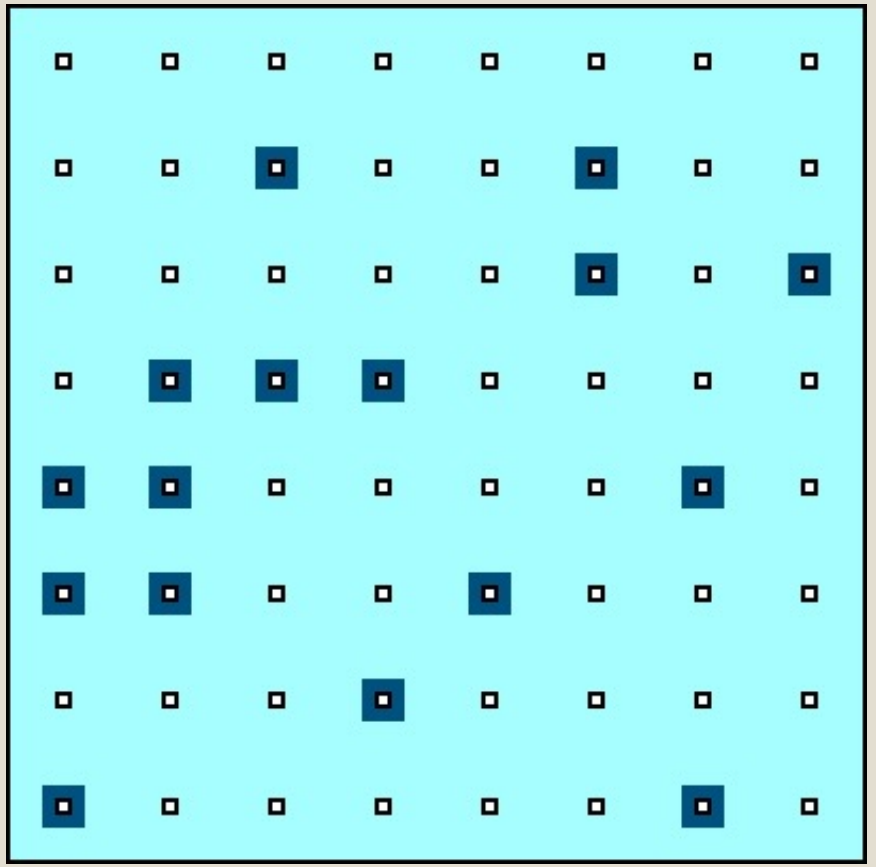
Échantillonnage répété (REP)

Échantillonnage à plusieurs degrés (EPD)

Échantillonnage à plusieurs phases (EPP)

ÉCHANTILLONNAGE ALÉATOIRE SIMPLE (EAS)

n unités sont choisies au hasard.



Avantages :

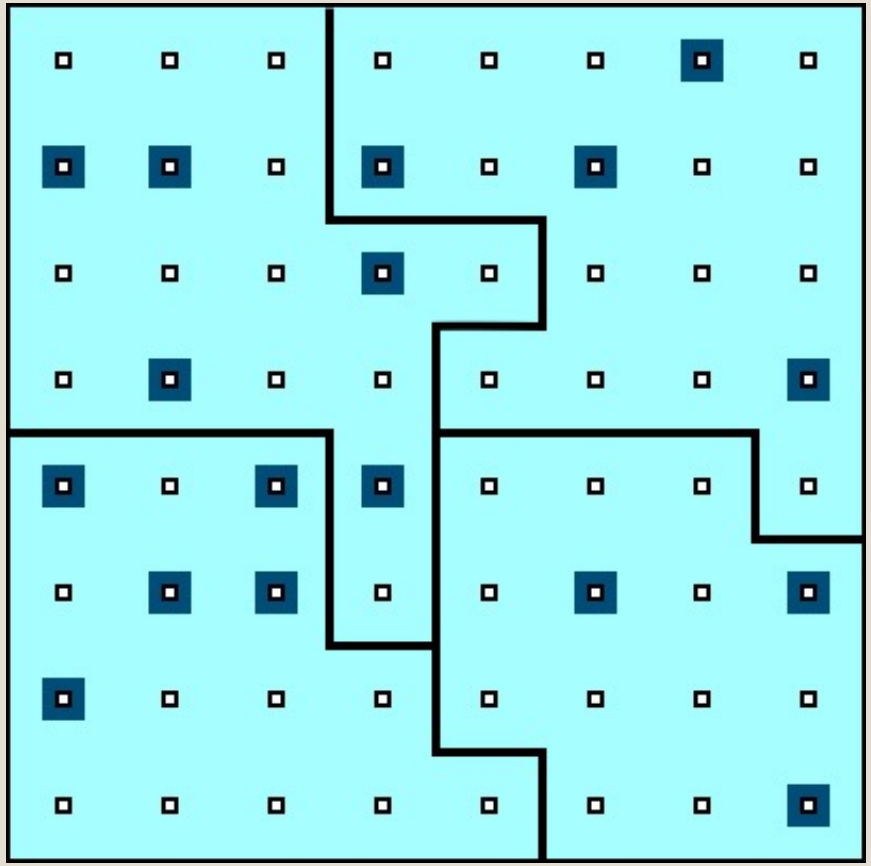
- plan d'échantillonage facile à mettre en oeuvre
- erreur d'estimation connue et facile à calculer
- ne requiert pas de données auxiliaires

Inconvénients :

- n'utilise pas l'information des données auxiliaires
- aucune garantie quant à la représentativité
- peut s'avérer coûteux si l'échantillon a une grande étendue géographique

ÉCHANTILLONNAGE STRATIFIÉ (STR)

$n = n_1 + \cdots + n_k$ unités sont choisies au hasard, à même k strates.



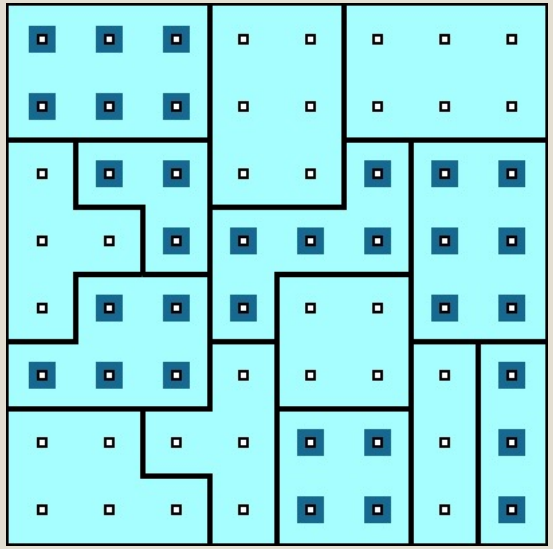
Avantages :

- peut produire une erreur d'estimation inférieure
- peut être moins dispendieux (strates adéquates)
- peut fournir des estimations pour sous-populations

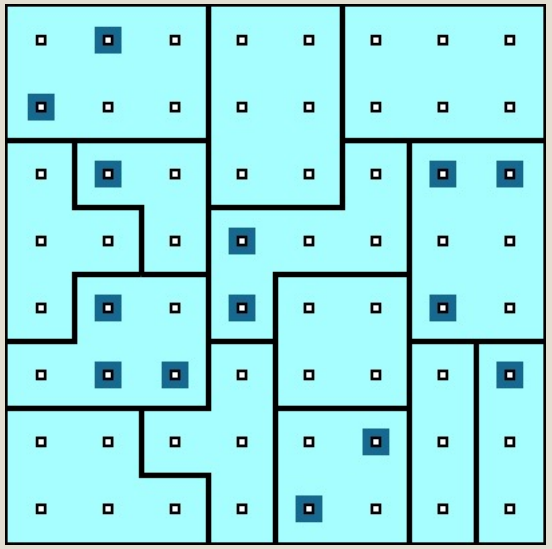
Inconvénients :

- aucun inconvénient majeur
- sans moyen naturel de stratifier la base d'enquête, à peu près équivalent à un EAS

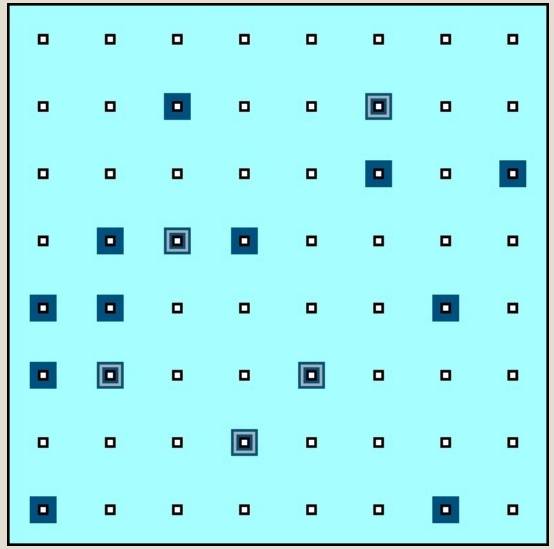
AUTRES PLANS D'ÉCHANTILLONNAGE



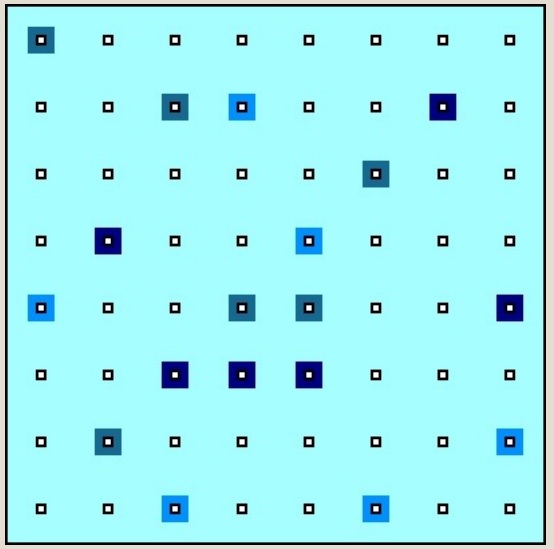
Échantillonnage en grappes



Échantillonnage à plusieurs degrés



Échantillonnage à plusieurs phases



Échantillonnage répété

MOISSONNAGE DU WEB ET ET COLLECTE AUTOMATISÉE DE DONNÉES

COLLECTE ET GESTION DES DONNÉES

« Les rues du Web sont pavées de données qui n'attendent que d'être recueillies. »

Munzart, Rubba, Meissner, Nyhuis, Automated Data Collection with R

WORLD WIDE WEB (« LA TOILE »)

La façon dont nous **partageons**, **collectons**, et **publions** les données a changé au cours des dernières années en raison de l'omniprésence du World Wide Web (WWW).

Les **entreprises privées**, les **gouvernements** et les **utilisateurs individuels** publient et partagent toutes sortes de données et d'informations.

À chaque instant, de nouveaux canaux génèrent de grandes quantités de données sur le comportement humain.

WORLD WIDE WEB

Il fut un temps, assez récent, où tant la rareté des données que leur inaccessibilité constituaient un problème pour les chercheurs et les décideurs. Tel n'est **manifestement** plus le cas désormais.

L'abondance des données présente son propre lot de problèmes particuliers :

- des masses de données enchevêtrées
- les méthodes classiques de collecte des données et les techniques usuelles d'analyse des données (en petites quantités) peuvent ne plus suffire aujourd'hui

EXEMPLE DE MOISSONNAGE DU WEB – NOUVEAU TÉLÉPHONE

Supposons que nous voulons savoir ce que la population pense d'un nouveau téléphone.

Approche standard : étude de marché (e.g. sondage téléphonique, système de récompenses, etc.).

Pièges :

- échantillon non représentatif : il se pourrait que l'échantillon sélectionné ne représente pas la population visée
- non-réponse systématique : les personnes qui n'aiment pas les sondages téléphoniques pourraient être moins (ou plus) portées à ne pas aimer le nouveau téléphone
- erreur de couverture : à titre d'exemple, il serait impossible de joindre les personnes qui ne disposeraient pas d'un téléphone filaire
- erreur de mesure : les questions du sondage fournissent-elles des renseignements convenant au problème posé?

EXEMPLE DE MOISSONNAGE DU WEB – NOUVEAU TÉLÉPHONE

Ces solutions peuvent être **onéreuses**, **chronophages**, **inefficaces**.

Variables de substitution – indicateurs qui sont étroitement reliés à la popularité du produit, sans mesurer directement celle-ci pour autant.

Si **popularité** = grands groupes de gens préfèrent un produit par rapport à un produit concurrent, les statistiques de vente que l'on retrouve sur un site Web commercial pourraient constituer un substitut de la popularité.

Les classements sur Amazon pourraient offrir une idée plus **complète** du marché par rapport à ce que permettrait d'obtenir un sondage classique.

EXEMPLE DE MOISSONNAGE DU WEB – NOUVEAU TÉLÉPHONE

Représentativité des **produits répertoriés**

- Tous les téléphones sont-ils répertoriés?
- Si tel n'est pas le cas, est-ce parce que le site Web ne les vend pas?
- Y a-t-il une autre raison?

Représentativité des **clients**

- Certains groupes spécifiques achètent-ils/n'achètent-ils pas de produits en ligne?
- Certains groupes spécifiques achètent-ils sur des sites spécifiques?
- Certains groupes spécifiques laissent-ils ou non des commentaires?

Honnêteté des clients et **crédibilité** des commentaires.

POURQUOI PROCÉDER À LA COLLECTE AUTOMATISÉE DES DONNÉES ?

En ce qui concerne les données scientifiques sociales :

- caractère limité des ressources financières
- peu de temps ou de désir de recueillir les données manuellement
- désir de travailler avec des sources riches en données à jour et de grande qualité
- documenter le processus du début (collecte des données) à la fin (publication) de sorte qu'il puisse être reproduit

Problèmes que pose la collecte manuelle :

- processus non reproductible
- présente des risques d'erreur en plus d'être lourd
- présente un risque plus élevé de « mourir d'ennui »

Advantages:

- fiabilité
- reproductibilité
- rapidité
- ensembles de données de meilleure qualité

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.

PROCESSUS DE COLLECTE DES DONNÉES

1. Savoir exactement de quel type d'information vous avez besoin

- Spécifique : ventes des dix principales marques de chaussures en 2017
- Vague : l'opinion des gens sur la marque de chaussures X

2. Déterminer s'il existe des sources de données sur le Web qui pourraient fournir de l'information directe ou indirecte sur votre problème

- Plus facile dans le cas de faits spécifiques : la page Web d'un magasin de chaussures fournira de l'information sur les chaussures qui sont actuellement prisées, etc.
- Les gazouillis (« tweets ») peuvent permettre de dégager des tendances en matière d'opinion sur tout et n'importe quoi
- Les plateformes commerciales peuvent fournir de l'information sur le niveau de satisfaction à l'égard d'un produit

PROCESSUS DE COLLECTE DES DONNÉES

3. Élaborer une théorie quant au processus de production des données lorsque l'on se penche sur des sources éventuelles

- Quand les données ont-elles été générées?
- Quand ont-elles été téléchargées sur le Web?
- Qui a téléchargé les données?
- Y a-t-il d'autres aspects qui pourraient ne pas avoir été couverts? Cohérence? Précision?
- À quelle fréquence les données sont-elles mises à jour?

PROCESSUS DE COLLECTE DES DONNÉES

4. Trouver un équilibre entre les avantages et les inconvénients des sources de données potentielles

- Valider la qualité des données utilisées
- Existe-t-il d'autres sources indépendantes qui fournissent de l'information similaire, et par rapport auxquelles il serait possible de procéder à une vérification croisée?
- Pouvez-vous identifier la source originale des données secondaires?

5. Prendre une decision

- Choisir la source de données qui semble la plus appropriée
- Documenter les raisons de cette décision
- Recueillir des données de plusieurs sources afin de valider les sources de données

QUALITÉ DES DONNÉES

Questions :

- Quel type de données est le plus approprié pour répondre à vos questions? is the quality of the data sufficiently high to answer the questions?
- L'information est-elle systématiquement déficiente?
- Les données sont-elles utilisées parce que "ce sont les meilleures données dont nous disposons" ?

La qualité des données dépend de **l'application**.

- Un échantillon de tweets collectés un jour au hasard pourrait être utilisé pour analyser l'utilisation d'un hashtag ou l'utilisation de mots spécifiques au genre.
- Moins utile s'il est collecté pendant le septième match de la finale de la Coupe Stanley (biais de collecte).

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**.

LÉGALITÉ DE LA MOISSONNAGE DU WEB

Qu'est-ce qu'une araignée?

- Il s'agit d'un programme qui parcourt ou arpente le Web pour en extraire de l'information rapidement
- L'araignée, ou programme collecteur, saute d'une page à l'autre, en en extrayant l'intégralité du contenu

Le **moissonnage** consiste à extraire de l'information spécifique de sites Web spécifiques : en quoi ces méthodes sont-elles **différentes**?

« Fondamentalement, le moissonnage consiste à **copier** de l'information : l'une des revendications les plus évidentes à l'encontre des dispositifs de récupération de données tient à la violation du droit d'auteur. »

LÉGALITÉ DE LA MOISSONNAGE DU WEB

L'exploration des informations d'une autre entreprise pour les traiter et les revendre est une plainte courante.

Directives éthiques :

- travailler de manière aussi transparente que possible
- documenter les sources de données à tout moment
- créditez ceux qui ont initialement collecté et publié les données
- si les données sont collectées par un autre organisme, demandez l'autorisation de les reproduire
- ne faites rien d'illégal !

LÉGALITÉ DE LA MOISSONNAGE DU WEB

eBay c. Bidder's Edge (BE)

- BE a eu recours à des programmes automatisés pour extraire de l'information de différents sites de vente aux enchères.
- Les utilisateurs pouvaient consulter les listes sur la page Web de BE, plutôt que d'avoir à se rendre sur les différents sites de vente aux enchères.
- En 1999, BE a accédé aux sites d'eBay environ 100 000 fois par jour (1,53 % du nombre de requêtes, 1,1 % de l'ensemble des données transférées par eBay).
- eBay a réclamé des dommages-intérêts allant de 45 000 $ et 62 000 $, sur une période de 10 mois.
- BE n'a volé aucune information qui n'était pas déjà publique, mais l'augmentation du trafic a imposé une charge additionnelle aux serveurs d'eBay.
- **Votre verdict?**

LÉGALITÉ DE LA MOISSONNAGE DU WEB

Facebook c. Pete Warden: Facebook a fait valoir que robots.txt n'avait aucune valeur juridique et qu'elle pouvait poursuivre quiconque accédait à son site, même si cette personne ou ce groupe se conformait aux instructions en matière de moissonnage.

Associated Press (AP) c. Meltwater: Meltwater propose un logiciel qui récupère des informations sur l'actualité à partir de mots-clés spécifiques ; le juge a donné raison à AP qui affirmait que son contenu avait été **volé par un concurrent**.

États-Unis c. Aaron Swartz: Swartz a été arrêté en 2011 pour avoir téléchargé illégalement des millions d'articles des archives de JSTOR.

LEÇONS APPRISES

On ne peut établir clairement quelles mesures de moissonnage sont illégales et lesquelles sont légales.

On estime que le fait de publier de nouveau du contenu à des fins commerciales est plus grave que ne l'est celui qui consiste à télécharger des pages à des fins de recherche ou d'analyse.

Robots.txt : le protocole d'exclusion des robots est un fichier qui indique au logiciel de récupération quelle information peut être recueillie sur le site.

Soyez gentil! Il n'est pas nécessaire de récupérer tout ce qui peut être récupéré. Les programmes de récupération devraient se comporter « gentiment », et ensuite d'être efficients, dans cet ordre de priorité.

COOPÉRATION AMICALE AVEC LES « API »

Il est toujours préférable d'opter pour la prudence : contactez les fournisseurs de données en cas de doute, surtout si de grands ensembles de données doivent être collectés.

« API » signifie **interface de programme d'application**, qui est un ensemble de routines, de protocoles et d'outils permettant de créer des applications logicielles.

De nombreuses API limitent l'utilisateur à un certain nombre d'appels d'API par jour (ou à d'autres limites). Ces limites doivent être respectées.

UTILISATION DES APIs

Une API est un moyen pour un site Web de donner aux programmes l'accès à ses données, sans avoir à recourir au moissonnage.

En d'autres termes, une API fournit un **accès structuré** à des **données structurées**.

Par exemple, un site financier peut offrir une API avec des données financières, ou le New York Times peut offrir une API pour les articles d'actualité.

Dans les deux cas, les données sont dans un format structuré et prédéfini (souvent JSON).

À FAIRE ET À NE PAS FAIRE EN MATIÈRE DE MOISSONNAGE

1. Demeurer identifiable

2. Réduire le traffic : Accepter les fichiers comprimés; en cas de moissonnage des mêmes ressources à plusieurs reprises, vérifiez tout d'abord si celles-ci ont changé avant d'y accéder de nouveau; ne récupérer que des parties de fichier.

3. Ne pas soumettre de demandes multiples au serveur : Le fait de soumettre de nombreuses demandes par seconde peut entraîner la mise hors service des serveurs peu puissants; les webmaîtres peuvent vous bloquer; une ou deux demandes par seconde est un rythme acceptable

4. Récupération de données modeste (efficient et poli) : Il est inutile de récupérer des pages quotidiennement ou de répéter la même tâche sans cesse; faites en sorte que votre programme de récupération de données soit aussi efficient que possible; Ne pas soumettre des pages à un trop grand nombre de demandes récupération; sélectionner les ressources que vous souhaitez utiliser et laisser le reste intact

[Collecte automatisée des données avec R]

Robert Gentleman

'What we have is nice, but we need something very different'

Source: Statistical Computing 2003, Reisensburg

Rolf Turner

'R is wonderful, but it cannot work magic'
answering a request for automatic generation of 'data from a known mean and 95% CI'

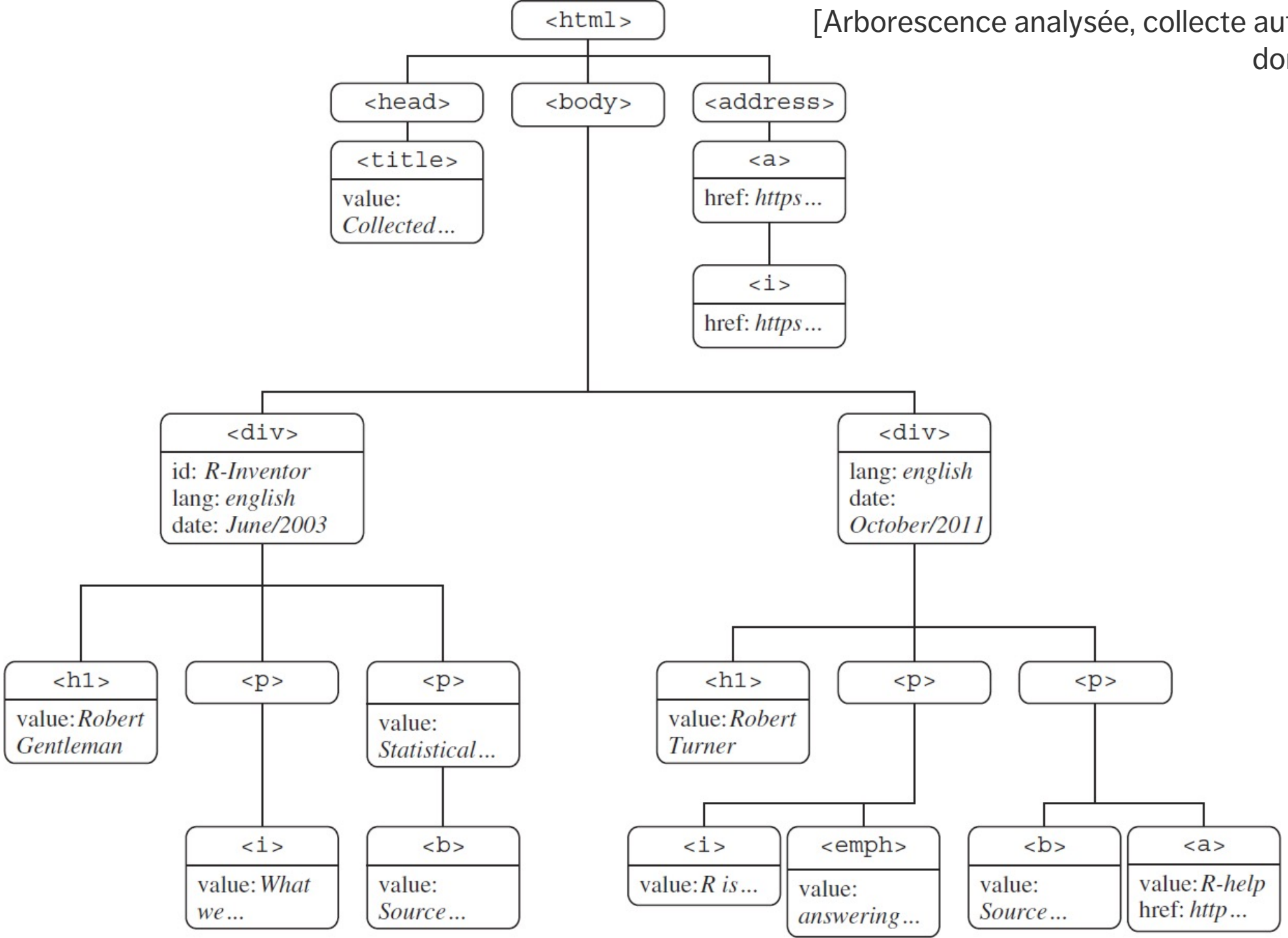
Source: R-help

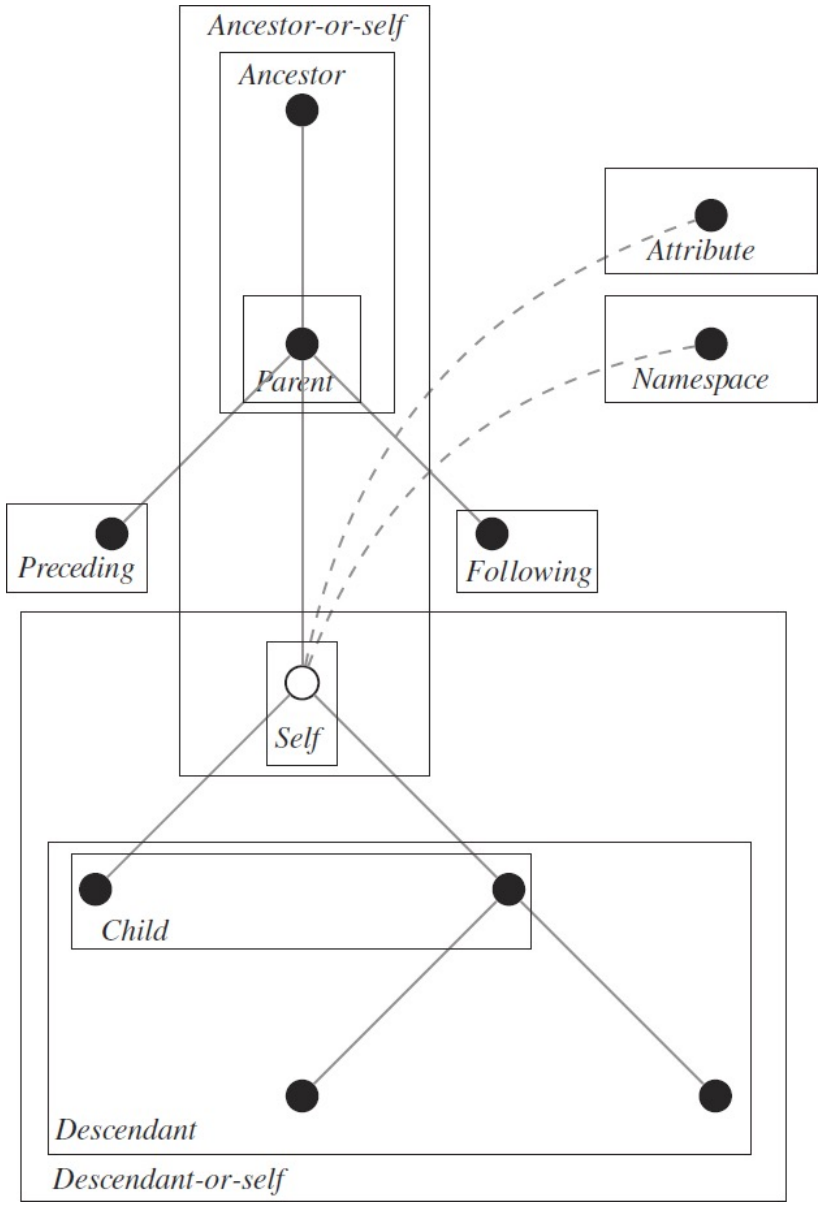
The book homepage

[Fichier HTML analysé, collecte automatisée des données avec R]

```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
<html>
<head><title>Collected R wisdoms</title></head>
<body>
<div id="R Inventor" lang="english" date="June/2003">
  <h1>Robert Gentleman</h1>
  <p><i>'What we have is nice, but we need something very different'</i></p>
  <p><b>Source: </b>Statistical Computing 2003, Reisensburg</p>
</div>

<div lang="english" date="October/2011">
  <h1>Rolf Turner</h1>
  <p><i>'R is wonderful, but it cannot work magic'</i> <br><emph>answering a request
for automatic generation of 'data from a known mean and 95% CI'</emph></p>
  <p><b>Source: </b><a href="https://stat.ethz.ch/mailman/listinfo/r-help">R-help</a>
</p>
</div>

<address>
<a href="http://www.r-datacollectionbook.com"><i>The book homepage</i></a><a></a>
</address>

</body>
</html>
```

[Arborescence analysée, collecte automatisée des données avec R]





Axis name	Result
ancestor	Selects all ancestors (parent, grandparent, etc.) of the current node
ancestor-or-self	Selects all ancestors (parent, grandparent, etc.) of the current node and the current node itself
attribute	Selects all attributes of the current node
child	Selects all children of the current node
descendant	Selects all descendants (children, grandchildren, etc.) of the current node
descendant-or-self	Selects all descendants (children, grandchildren, etc.) of the current node and the current node itself
following	Selects everything in the document after the closing tag of the current node
following-sibling	Selects all siblings after the current node
namespace	Selects all namespace nodes of the current node
parent	Selects the parent of the current node
preceding	Selects all nodes that appear before the current node in the document except ancestors, attribute nodes, and namespace nodes
preceding-sibling	Selects all siblings before the current node
self	Selects the current node

OUTILS DE MOISSONNAGE DU WEB

XPath

`rvest`

Beautiful Soup

Selenium

etc.

MODÉLISATION DES DONNÉES ET DES CONNAISSANCES

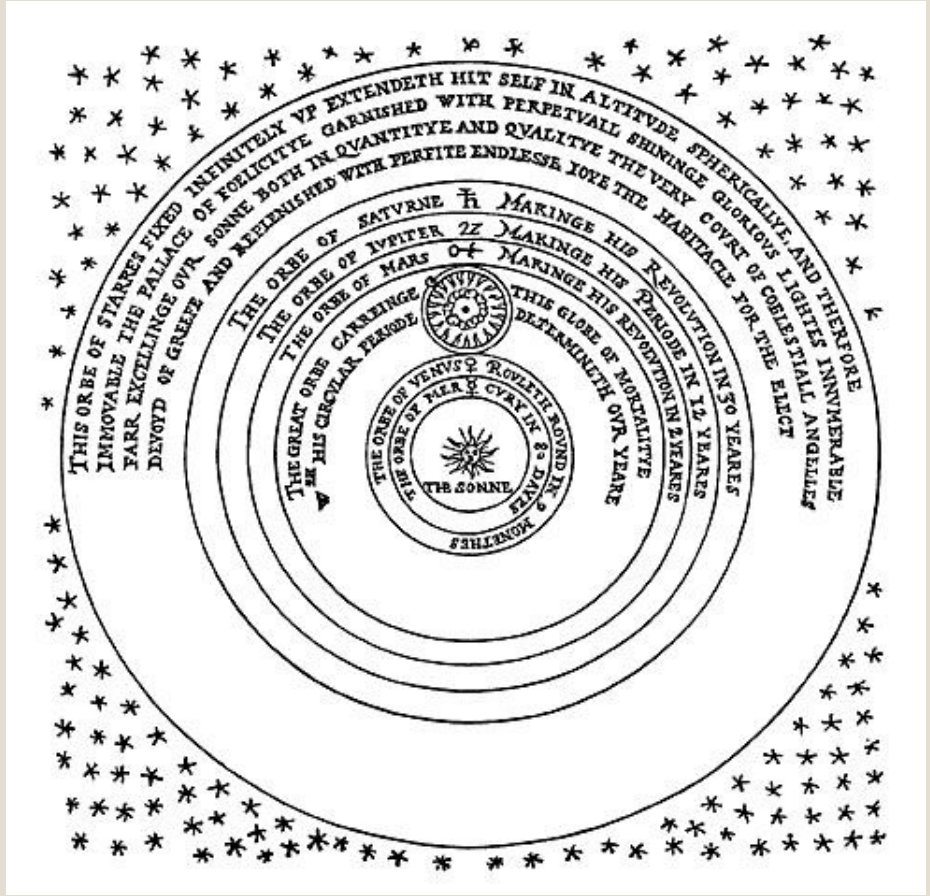
COLLECTE ET GESTION DES DONNÉES

CONCEPTS DE BASE

Comment faire le lien entre les différentes disciplines qui utilisent des données ?

Concepts ou éléments fondamentaux (systèmes) :

- objet - attributs (concrets ou abstraits)
- objets multiples - **relations** entre ces objets/attributs
- comment ces éléments évoluent dans le temps



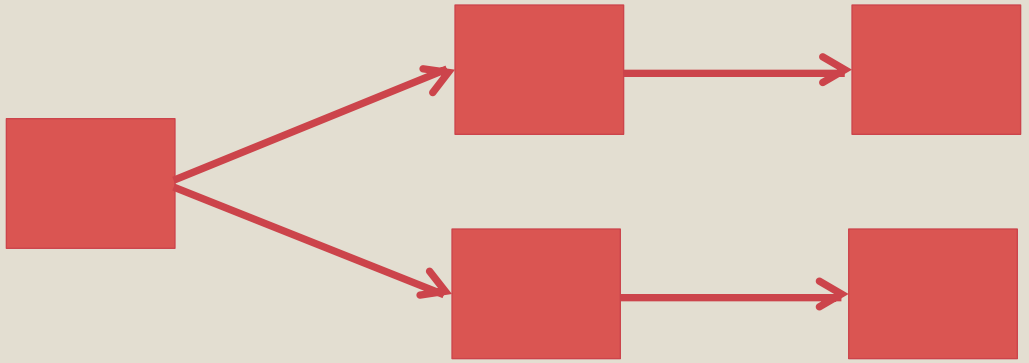
RELATIONS DE SYSTÈME

Quelques relations fondamentales :

- partie-entière
- est-un
- est-un-type-de
- cardinalité (un à un, un à plusieurs, plusieurs à plusieurs)

Quelques relations spécifiques à un objet :

- la propriété
- relations sociales
- devient
- mène à



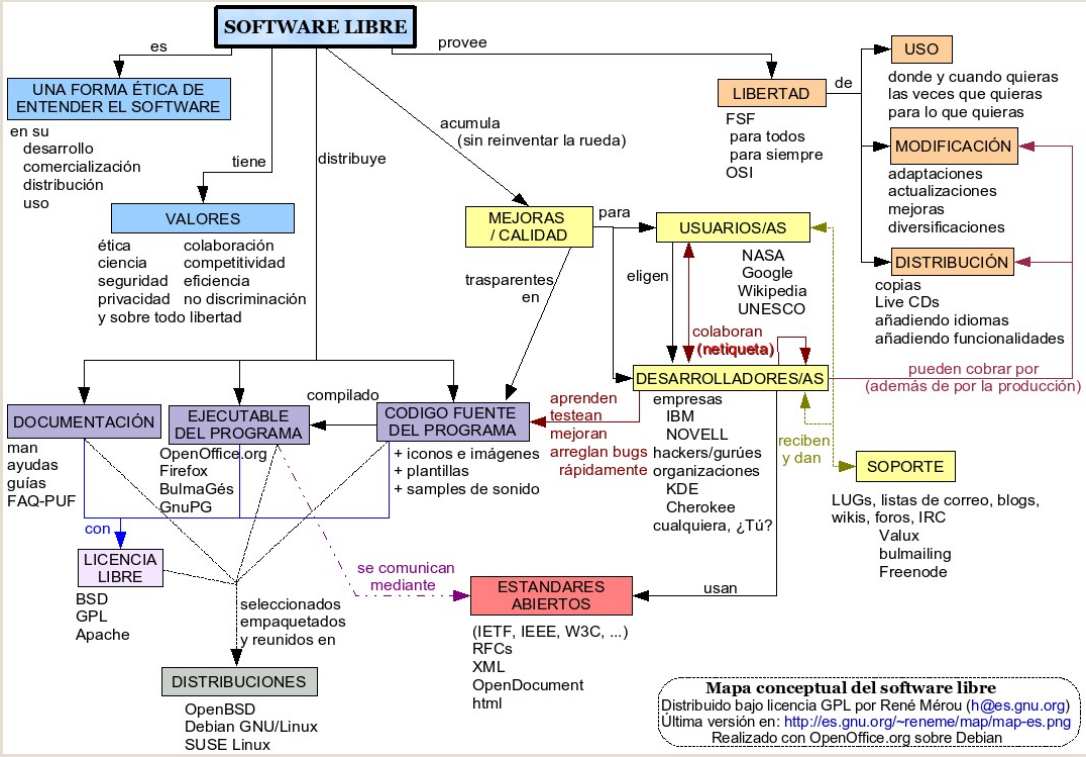
MODÈLE CONCEPTUEL

Un **modèle conceptuel** est, grosso modo :

- un modèle qui n'est pas mis en œuvre, qui n'existe que sur le plan conceptuel
- un diagramme ou une description verbale d'un système (par exemple, boîtes et flèches, cartes mentales, listes, définitions)

L'accent est mis sur :

- les **états possibles**, et non sur la capture de comportements spécifiques
- les types d'objets, et non sur des instances spécifiques ; le but est l'**abstraction**.



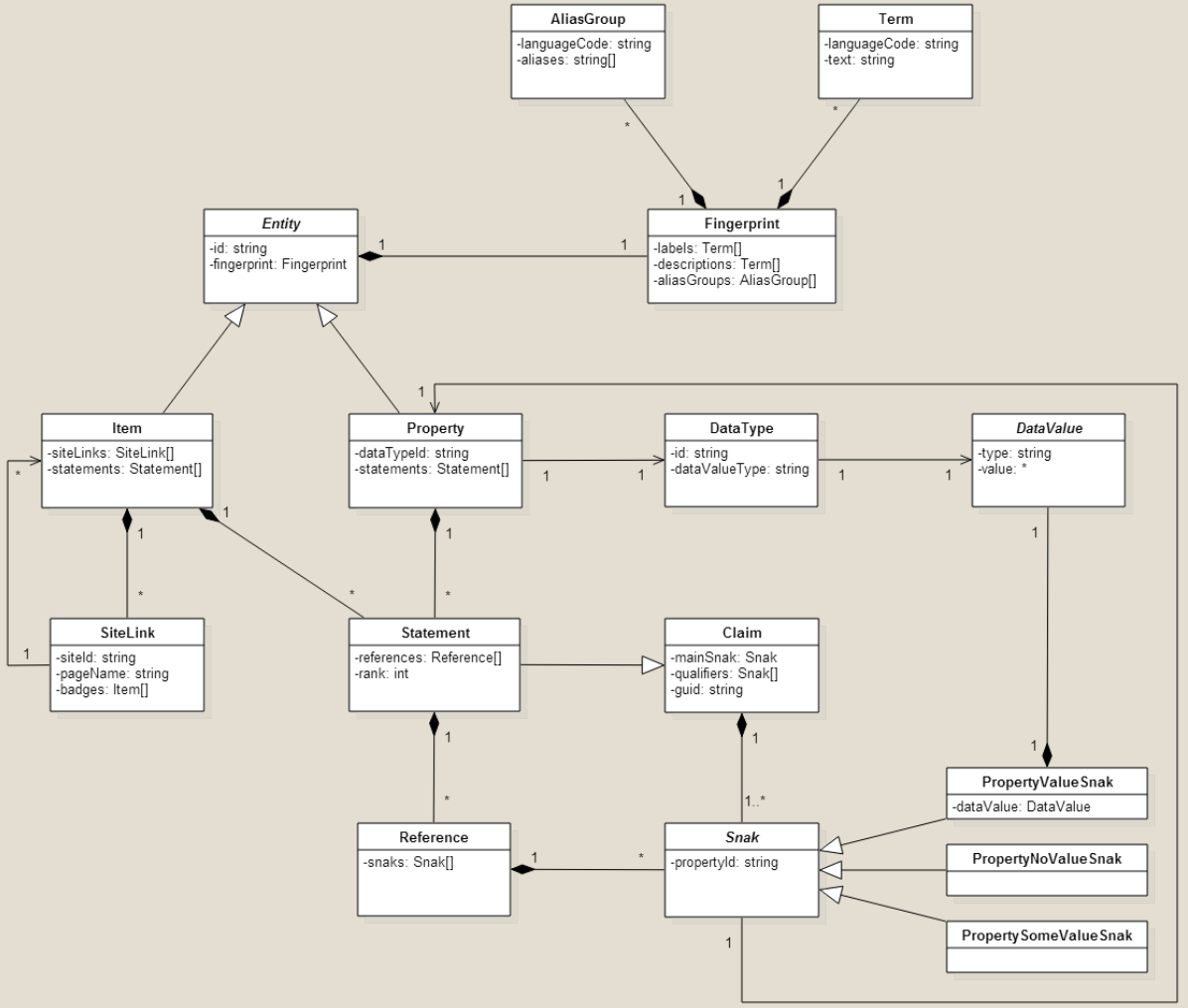
LA MODÉLISATION CONCEPTUELLE FORMELLE

La modélisation conceptuelle permet de transformer les modèles conceptuels internes en modèles **explicites** et **tangibles**.

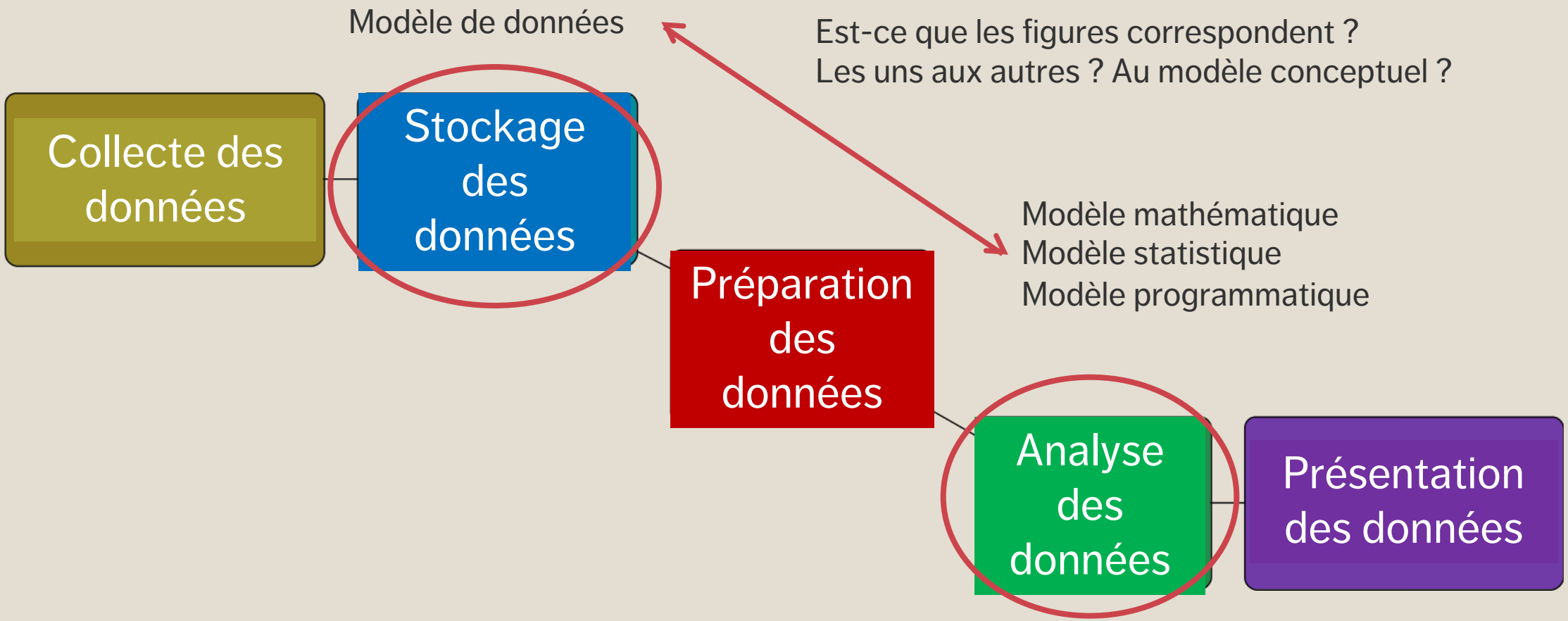
Elle offre la possibilité d'examiner et d'explorer des idées et des hypothèses.

Divers efforts ont été faits pour formaliser la modélisation conceptuelle :

- UML (langage universel de modélisation)
- Modèles de relations entre entités (ER)



PIPELINE DE DONNÉES AUTOMATISÉ



CONCEPTS FONDAMENTAUX

Il est important de structurer les **données** et les **connaissances** afin qu'elles puissent être :

- stockées et accessibles
- modifiables et ajoutables
- extraites utilement et efficacement (extraction - transformation - chargement)
- gérés par des humains et des ordinateurs (programmes, robots, IA)

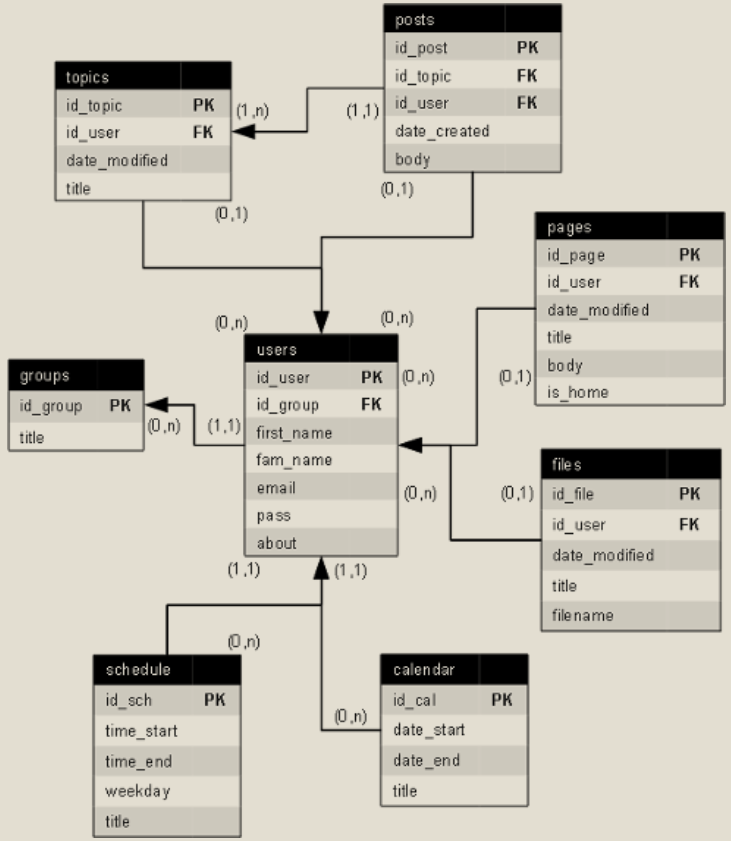
MODÉLISATION DES DONNÉES

Les **modèles de données** sont des descriptions **abstraites/logiques** d'un système, construites en termes qui peuvent ensuite être mis en œuvre comme structure d'un type de logiciel de gestion des données.

C'est à mi-chemin entre un modèle conceptuel et une mise en œuvre de la base de données.

Les données elles-mêmes concernent les **instances** - le modèle concerne les **types d'objets**.

Autre option à envisager : les **ontologies**.



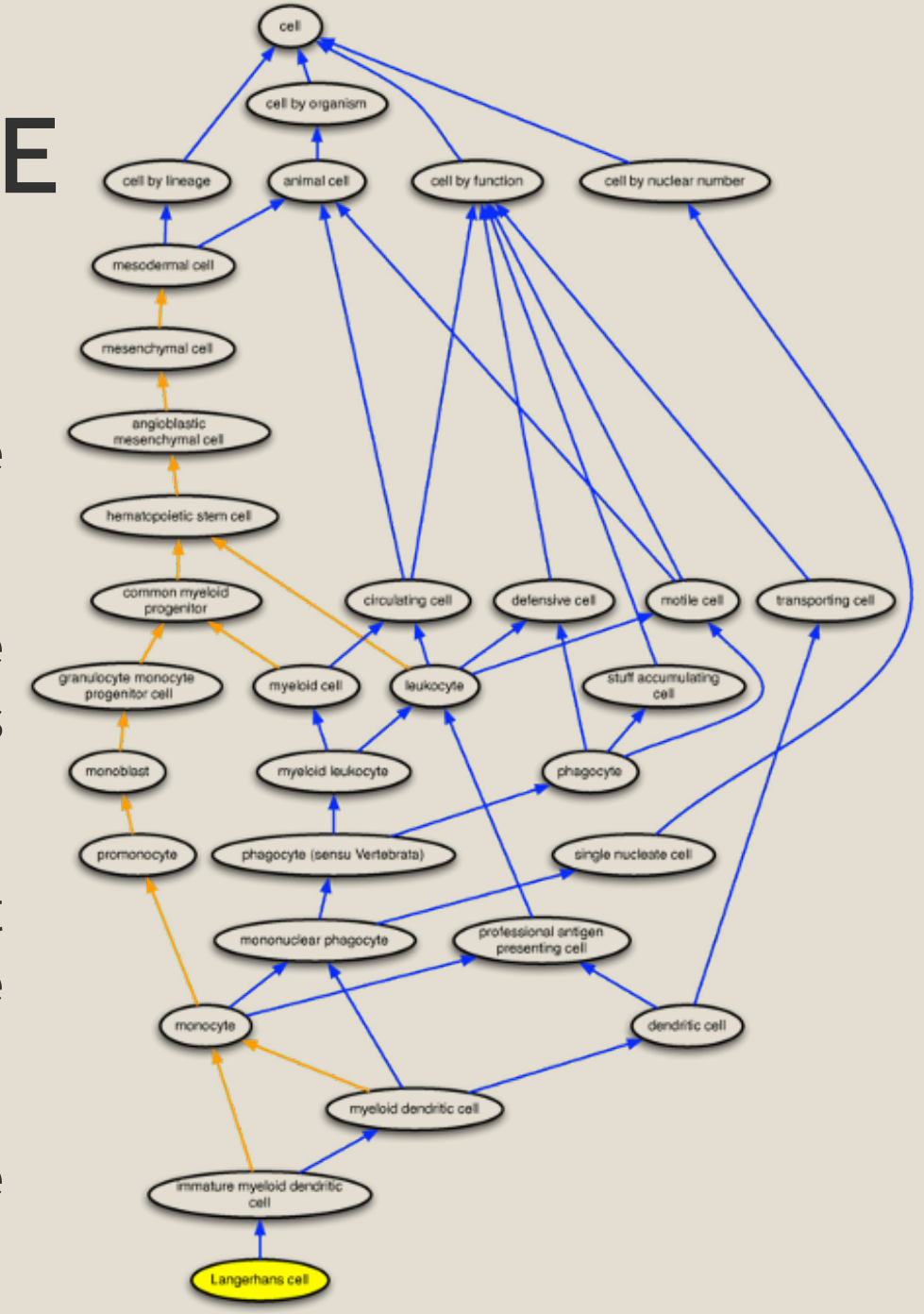
ONTOLOGIE – MODÈLE DE CONNAISSANCES

Une collection structurée et lisible par machine de **faits** sur un domaine.

Motivée par le désir de créer des données de plus en plus lisibles par machine, mais toujours **complexes sur le plan conceptuel**.

Vous pourriez décrire l'ontologie, en plaisantant un peu, comme « un modèle de données gonflé aux stéroïdes ».

Une tentative de se rapprocher du niveau de détail d'un **modèle conceptuel complet**.



MÉTADONNÉES POUR FOURNIR UN CONTEXTE

Nous perdons quelque chose lorsque nous passons de notre modèle conceptuel à un modèle de type spécifique – p. ex. le modèle de données ou de connaissances.

Une façon de conserver le contexte est de fournir des **métadonnées** (riches, si possible) – des données **sur** nos données!

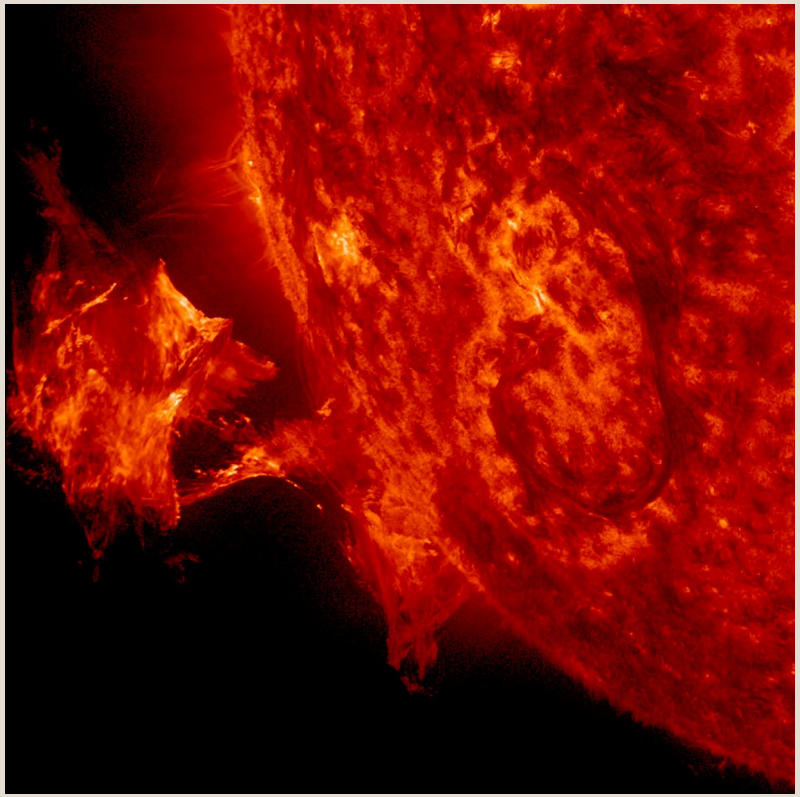
Les métadonnées sont essentielles lorsqu'il s'agit de mettre en œuvre des stratégies pour travailler d'un ensemble de données à l'autre.

L'ontologie peut aussi jouer un rôle ici!

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



MODÉLISATION DES DONNÉES (REPRISE)

Différentes options sont actuellement populaires en termes de **données fondamentales** et de stratégies de modélisation ou de structuration des **connaissances :**

- paires de valeurs clés (p. ex. JSON)
- triples (p. ex. modèle RDF)
- bases de données graphiques
- bases de données relationnelles

MÉMOIRES DE VALEURS CLÉS ET MÉMOIRES TRIPLES

Voici des moyens **relativement peu structurés** de stocker les données :

- **Valeur clé :** toutes les données sont simplement stockées sous la forme d'une liste géante de clés (noms ou étiquettes) et de valeurs (associées à la clé).
- **Triple** : les données sont stockées en tant que sujet – prédicat – objet

Exemples :

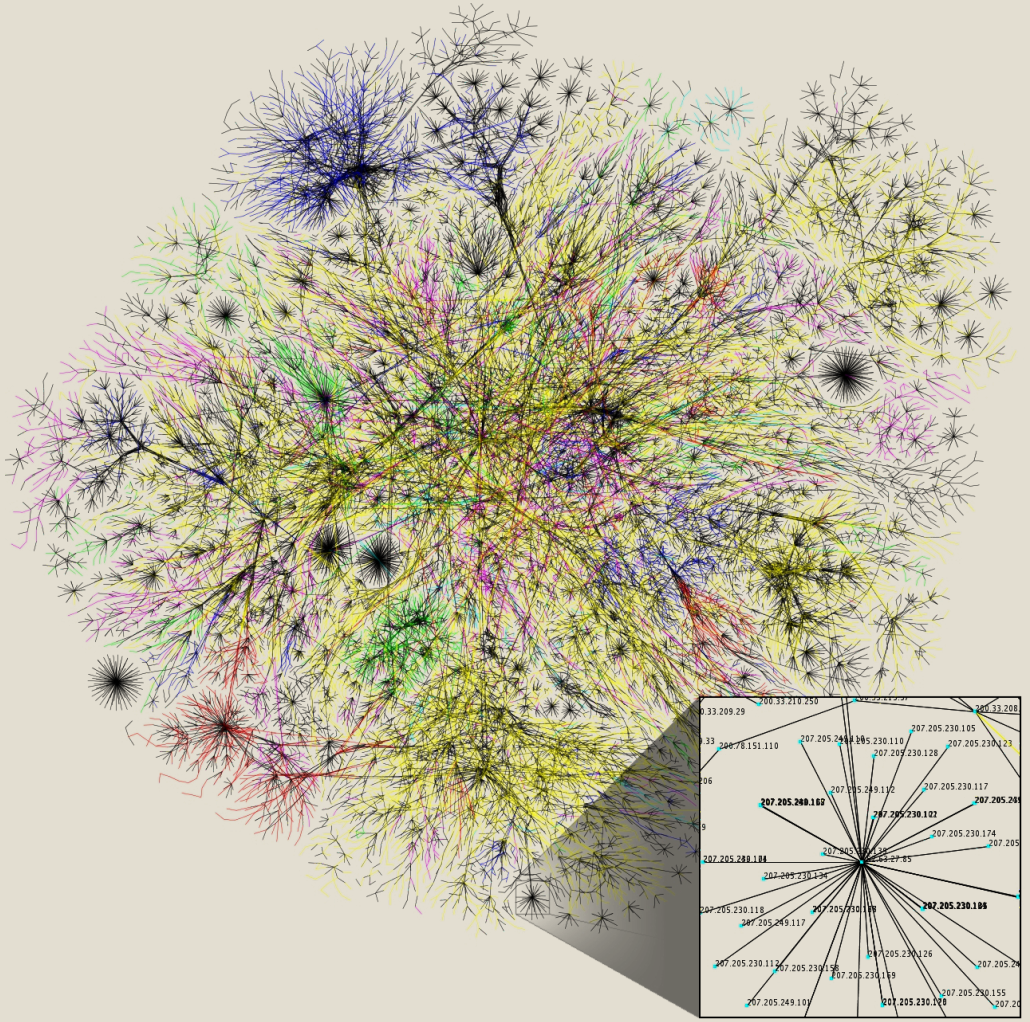
- **type de pomme - couleur de la pomme :** (Granny Smith – verte), (Red Delicious – rouge)
- **personne - pointure de chaussures :** (Gwynneth Rayfield – 2), (Llewellyn Rayfield – 6)
- **mot – définition:** (URL – page web), (nom de rapport – rapport [dossier de documentation])
- Les **triples** ajoutent un verbe au mélange : Personne - est - âge, Objet - est de couleur - couleur

BASES DE DONNÉES GRAPHIQUES

Accent mis sur les **liens** entre les différents types d'objets, plutôt que sur les liens entre un objet et les propriétés de cet objet.

Le modèle de données :

- objets représentés par des **nœuds**
- liens entre ces objets représentés par des **bords**
- les objets peuvent avoir un lien avec d'autres objets du même type - la personne est le frère ou la sœur de l'autre personne



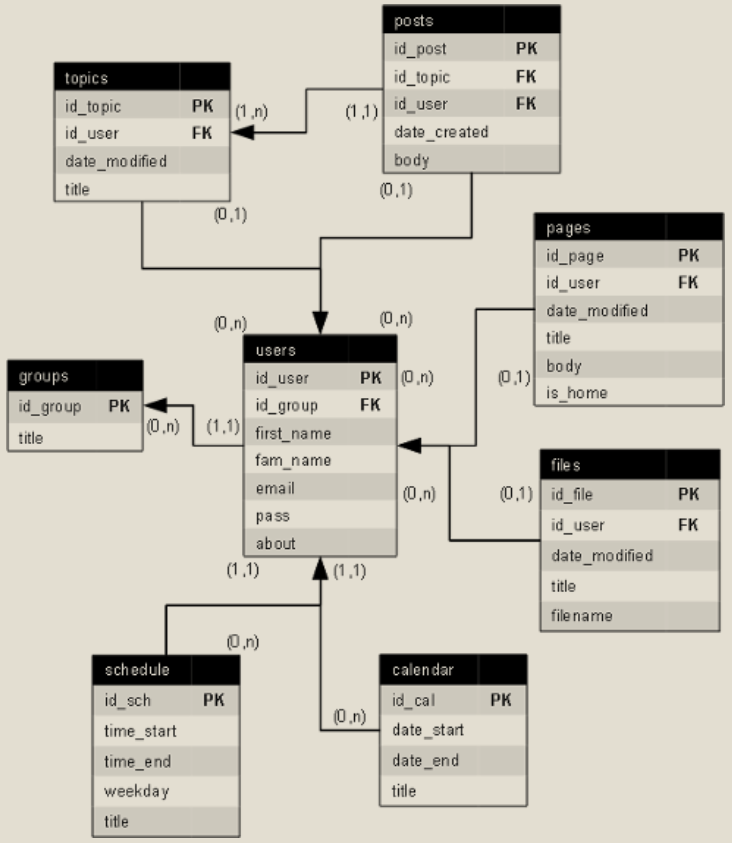
BASES DE DONNÉES RELATIONNELLES

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.



MÉMOIRES ET BASE DE DONNÉES

Base de données relationnelle:

- largement soutenue, bien comprise, fonctionne bien pour de nombreux types de systèmes et de cas d'utilisation. Base toutefois difficile à changer une fois mise en œuvre; ne gère pas bien les liens.

Mémoires de valeurs clés:

- peuvent prendre n'importe quel type de données; nul besoin de beaucoup de renseignements sur l'avancement de sa structure. Si vous avez beaucoup de valeurs manquantes, ces mémoires ne prendront pas de place. Peuvent toutefois être désordonnées et mystérieuses; difficile d'y trouver des données.

Bases de données graphiques:

- rapides et intuitives si vous utilisez des données fortement axées sur les liens; pourraient être la seule option si vos données sont ainsi parce que les bases de données traditionnelles peuvent ralentir énormément. Sont toutefois probablement trop spécialisées si vos données ne sont pas ainsi, pas encore supportées à grande échelle.

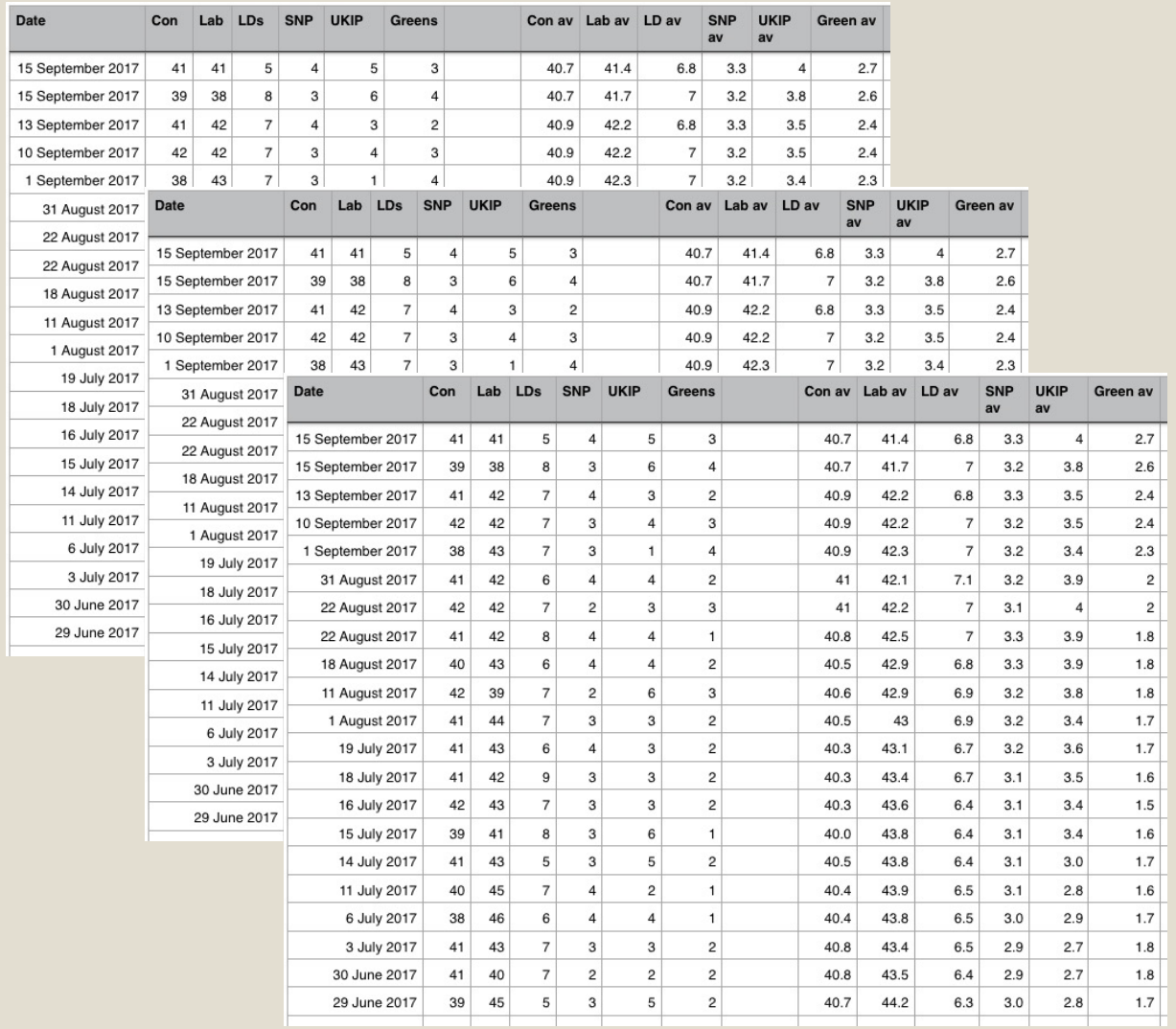
FICHIERS NON HIÉRARCHIQUES ET LES FEUILLES DE CALCUL

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

FICHIERS NON HIÉRARCHIQUES ET LES FEUILLES DE CALCUL

Avantages :

- très efficace si vous recueillez des données une seule fois, sur un type particulier d'objet
- certains types d'analyse exigent que vous ayez toutes les données en un seul endroit
- facile à lire dans un logiciel d'analyse et à effectuer des opérations sur l'ensemble des données

Inconvénients :

- très difficile de gérer l'intégrité des données à long terme si vous recueillez et travaillez continuellement avec les données
- n'est pas idéal pour les données de systèmes impliquant de multiples types d'objets et de relations
- il peut être très difficile d'effectuer des opérations d'interrogation de données

OUTILS ET MOTS À LA MODE

MongoDB, ArangoDB

Entrepôt de documents

JSON, YAML

API, GraphQL

Données interreliées

Web sémantique

Langage d’ontologie Web (OWL)

Protégé

MISE EN ŒUVRE DU MODÈLE

Pour mettre en œuvre votre modèle de données/connaissances, vous devez avoir accès à un **logiciel de stockage et de gestion des données**.

Cela peut constituer un défi pour les individus, car ces logiciels fonctionnent généralement sur des **serveurs**.

Les serveurs sont utiles car ils permettent à plusieurs utilisateurs d'accéder simultanément à une même base de données, à partir de différents programmes clients, mais il est difficile de « jouer » avec les données.

C'est là que **SQLite** entre en jeu.

RÔLE DU LOGICIEL DE GESTION DES DONNÉES

Les logiciels de gestion des données offrent aux utilisateurs un moyen facile d'interagir avec leurs données.

Il s'agit essentiellement d'une interface entre les **personnes** et les **données**.

Grâce à cette interface, vous pouvez:

- ajouter des données à votre collection de données
- extraire des sous-ensembles de données de votre collection en fonction de certains critères
- supprimer des données de votre collection ou en modifier

NOMS / TERMINOLOGIE

Auparavant :

- Base de données
- Entrepôt de données
- Mini-entrepôts de données
- Système de gestion de bases de données
- (SQL)

Maintenant :

- Lac de données
- Bassin de données
- Marais de données?
- Cimetière de données?
- (noSQL)

De plus en plus : une distinction entre l'**entrepot de données** et le **logiciel de gestion des données**.

DU MODÈLE DE DONNÉES À LA MISE EN ŒUVRE

Une fois que le modèle de données (logique) est achevé :

- **instancier** le modèle dans le logiciel de votre choix (p. ex : créer des tableaux dans MySQL);
- **télécharger/charger** les données.
- **interroger** les données :
  - Les bases de données relationnelles traditionnelles utilisent le **langage de requête structuré** (SQL : Structured Query Language)
  - d'autres types de bases de données utilisent des langages d'interrogation totalement différents (AQL, moteurs sémantiques, etc.) ou reposent sur des programmes informatiques sur mesure (par exemple, écrits en R, Python)

GESTION DE LA COLLECTION DE DONNÉES

Une fois les données collectées, il faut aussi les **gérer**.

Fondamentalement, cela signifie que la base de données doit être maintenue, afin que les données soient

- exactes
- précises
- uniformes
- complètes

Ne laissez pas votre lac de données se transformer en marécage!