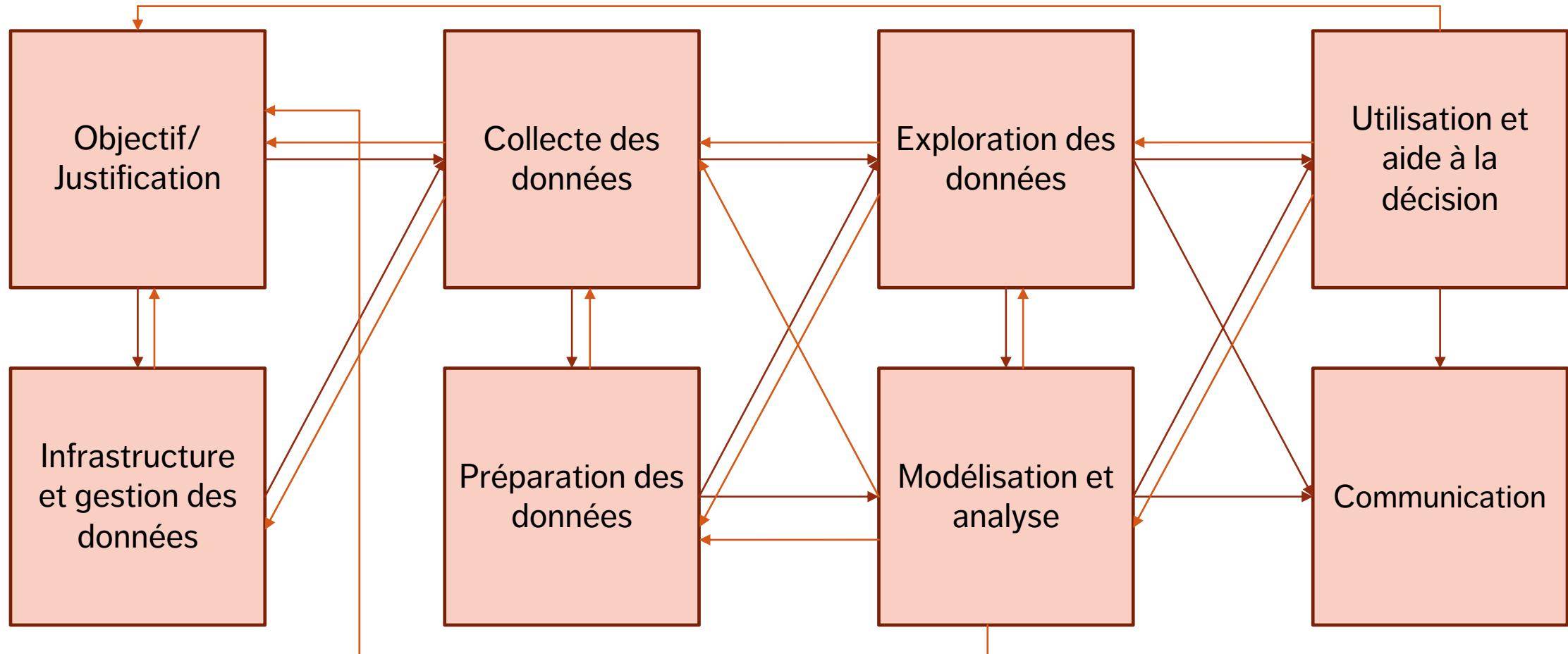

ARCHITECTURE DE DONNÉES ET DE L'INFORMATION

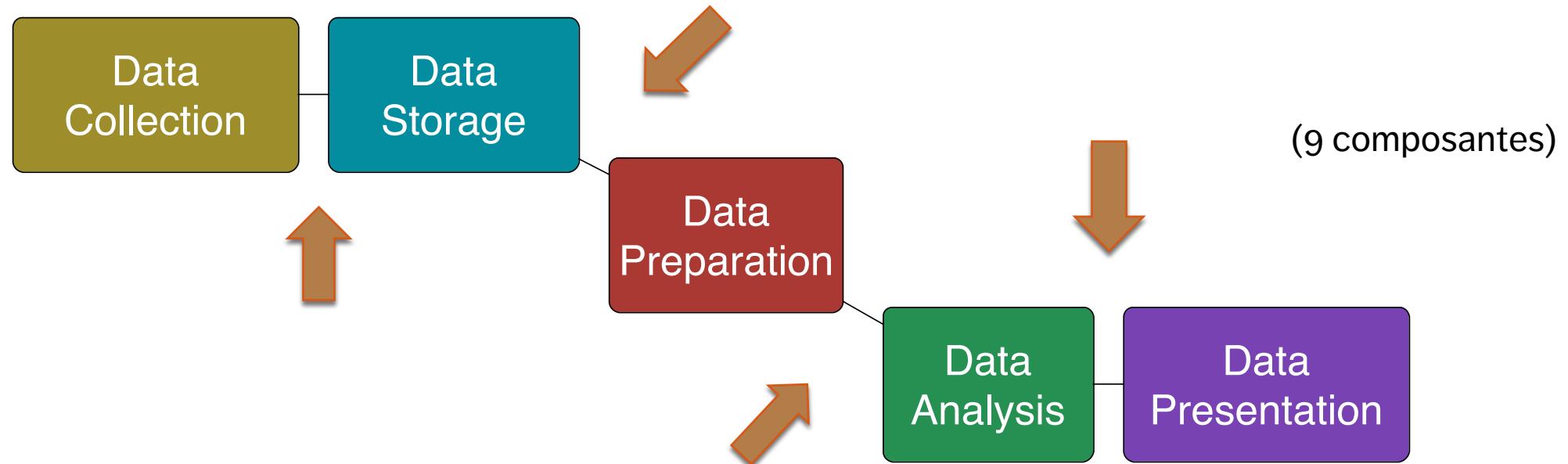
PRÉPARATION DU TERRAIN

LE PROCESSUS D'ANALYSE (DÉSORDONNÉ)



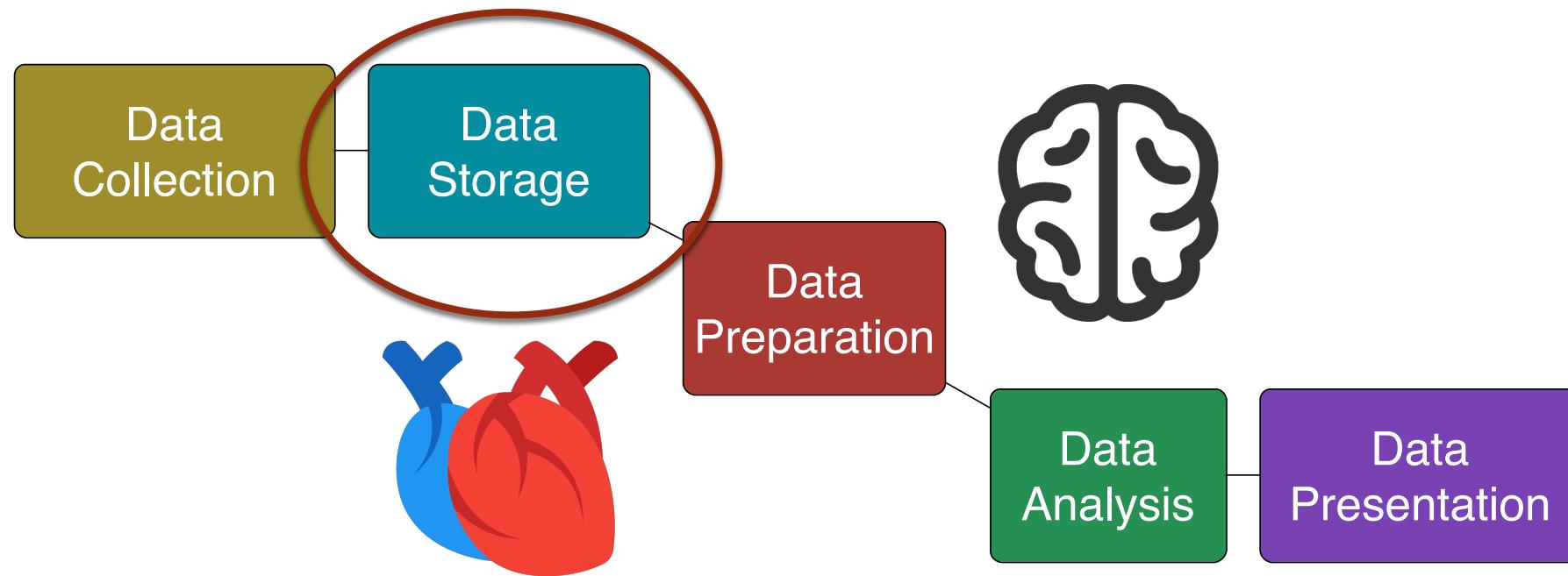
UN PIPELINE DE DONNÉES (AUTOMATISÉ) MIS EN ŒUVRE

Dans le **contexte de la prestation de services**, il se peut que vous souhaitiez avoir l'une de ces composantes...



(Comme toujours, attention au modèle « drift »)

UN PIPELINE DE DONNÉES (AUTOMATISÉ) MIS EN ŒUVRE



MOTIVATIONS POUR LA COLLECTION

Trois fonctions, historiquement :

- Tenue de dossiers (gestion des personnes et de la société!)
- Science - Nouvelles connaissances générales
- Renseignements - commerciaux, militaires, de police, sociaux? Au Canada? Personnel!



DIFFÉRENTES CULTURES DE DONNÉES, DIFFÉRENTS TERMES

Renseignements opérationnels :

- entrepôt de données + mini-entrepôt de données
- « dimension » de données (= ensemble de données)
- données hiérarchiques (tranches)
- élément de données
- tableau des dimensions + tableau des faits

Données

Sciences/statistiques :

- données expérimentales
- essais
- participants
- variables
- corrélation

Connaissances

Gestion des documents :

- architecture de l'information
- plan de classement
- ressource d'information
- champ
- formulaire
- objet

Renseignements

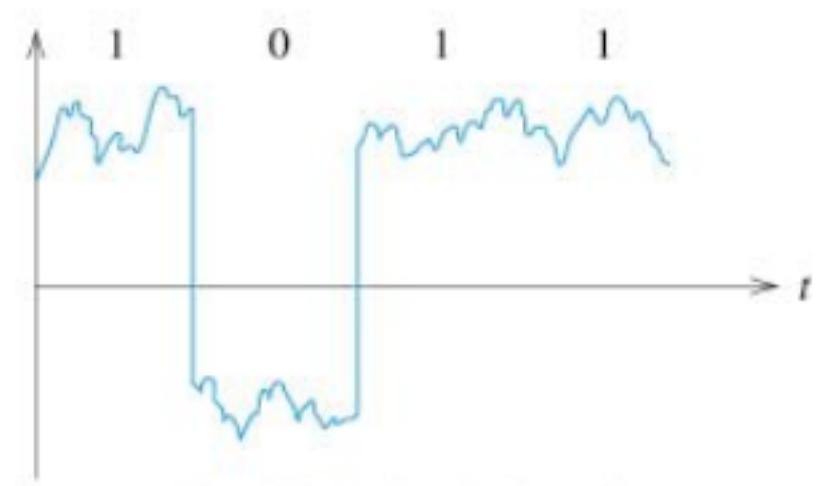
Science des données

ENTREZ DANS LE MONDE DE LA SCIENCE INFORMATIQUE! ET DE L'INFORMATIQUE!

L'informatique (et les sciences de l'information) a son propre point de vue théorique **fondamental** sur les données et l'information.

Essentiellement, les ordinateurs fonctionnent avec des données - des 1 et des 0 qui représentent des chiffres, des lettres, etc. - **les données deviennent numériques.**

De façon pragmatique, les données sont maintenant stockées sur des ordinateurs, et elles sont accessibles par l'intermédiaire de notre réseau informatique mondial.



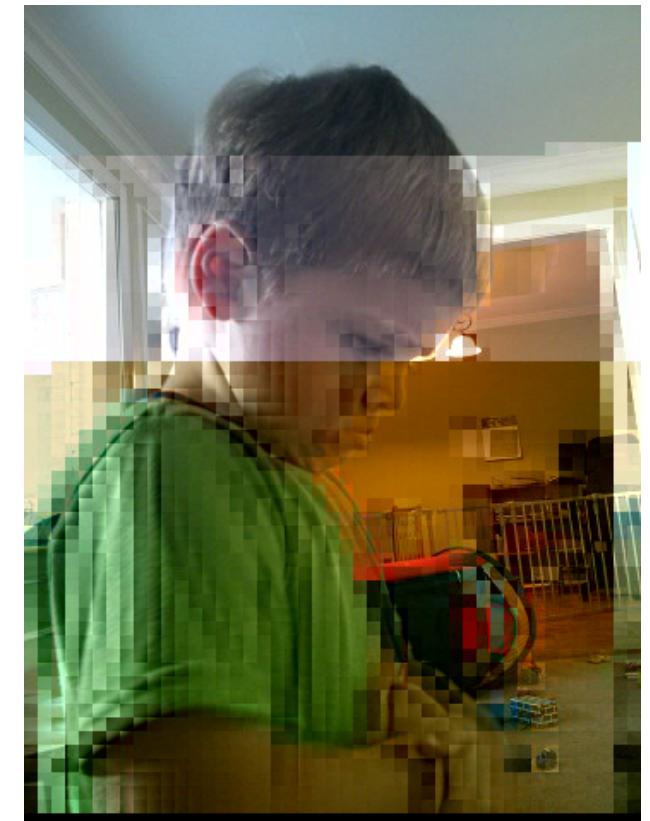
DÉGRADATION DES DONNÉES

Les données vieillissent avec le temps - elles ont une durée de vie.

Nous utilisons les expressions « données pourries », « données en dégradation » :

- **littéralement** - dans le sens où le support de stockage de données pourrait se dégrader;
- **métaphoriquement** - lorsque les données ne **représentent** plus exactement les relations et les objets pertinents ou même lorsque ces objets n'existent plus de la même manière.

Vos données doivent être « fraîches », « actuelles », et non « périmées » (en fonction du contexte et du modèle!)

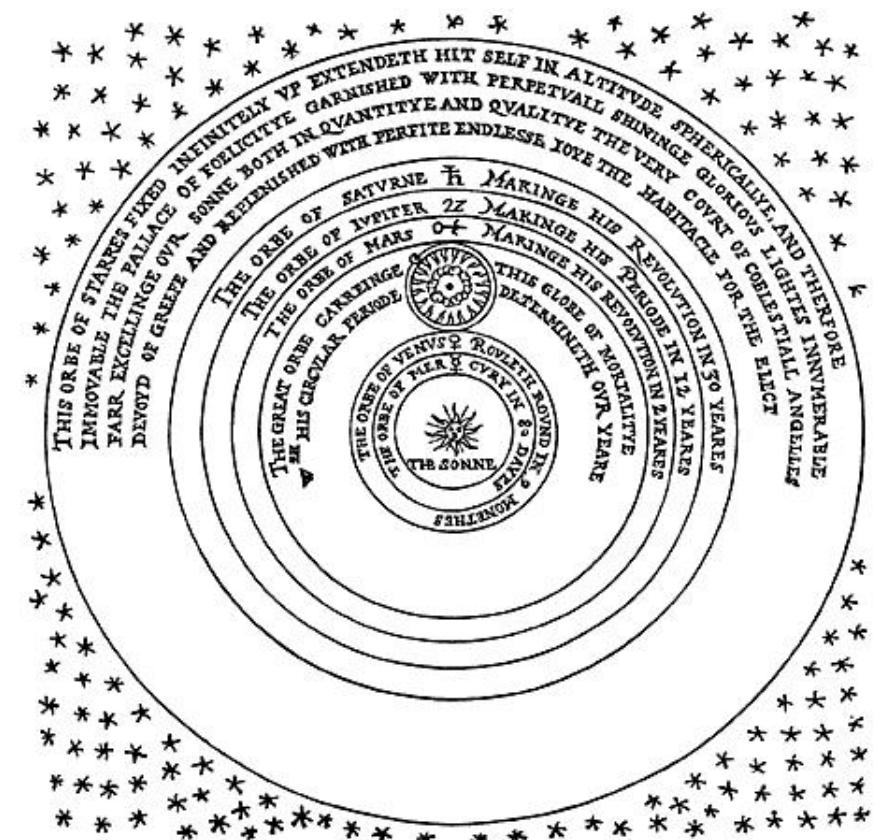


CONCEPTS DE BASE

Comment recouper les différentes disciplines qui utilisent des données?

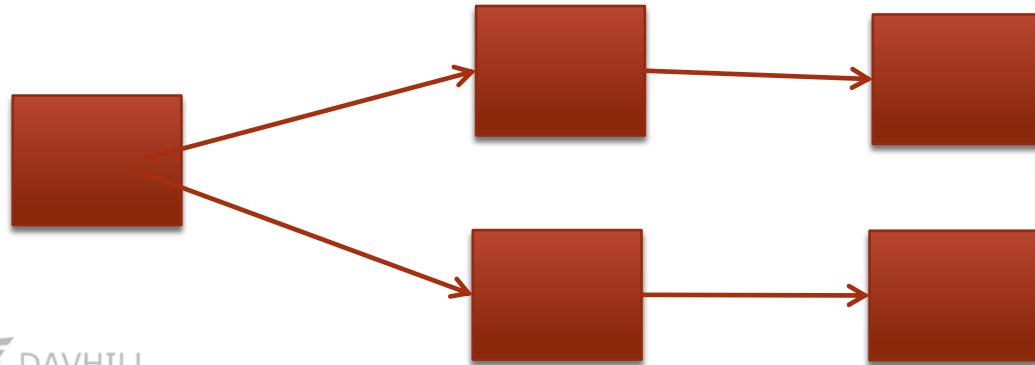
Concepts ou éléments de base (systèmes) :

- objet - attributs (concrets ou abstraits)
- objets multiples - **liens** entre ces objets/attributs
- façon dont ces éléments évoluent au fil du temps



LIENS AU SEIN DES SYSTÈMES

- quelques liens fondamentaux :
 - en partie
 - est une
 - est un type de
 - cardinalité (un à un, un à plusieurs, plusieurs à plusieurs)
- quelques autres liens liés à l'objet :
 - propriété,
 - liens sociaux
 - devient,
 - mène à



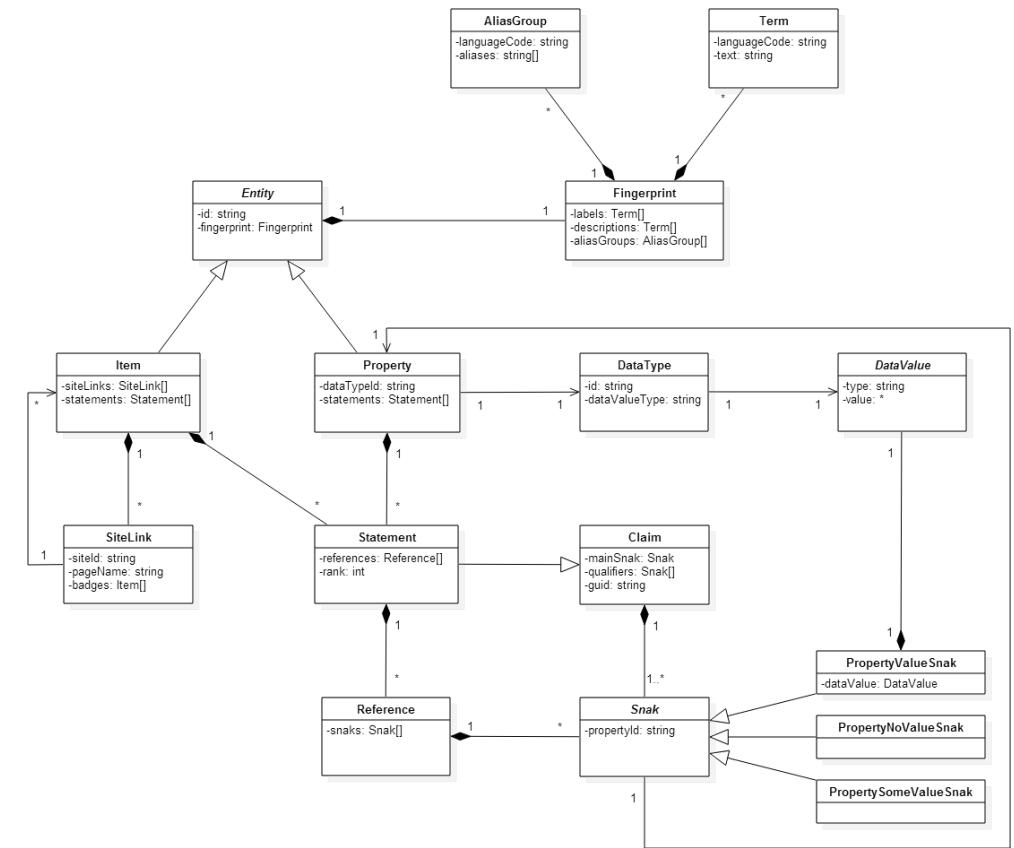
LANGAGES FORMELS DE MODÉLISATION CONCEPTUELLE

La modélisation conceptuelle n'est pas une science exacte – il s'agit plutôt de rendre vos modèles conceptuels internes **explicites** et **tangibles**.

Elle vous donne l'occasion d'examiner et d'explorer vos idées et vos hypothèses.

Cela dit, divers efforts ont été faits pour formaliser la modélisation conceptuelle :

- UML - Langage de modélisation unifié
- Modèles de rapport entre entités - mais ils sont généralement connectés à des bases de données relationnelles



MODÈLES MÉTHODOLOGIQUES

Modèle conceptuel



Modèle mathématique

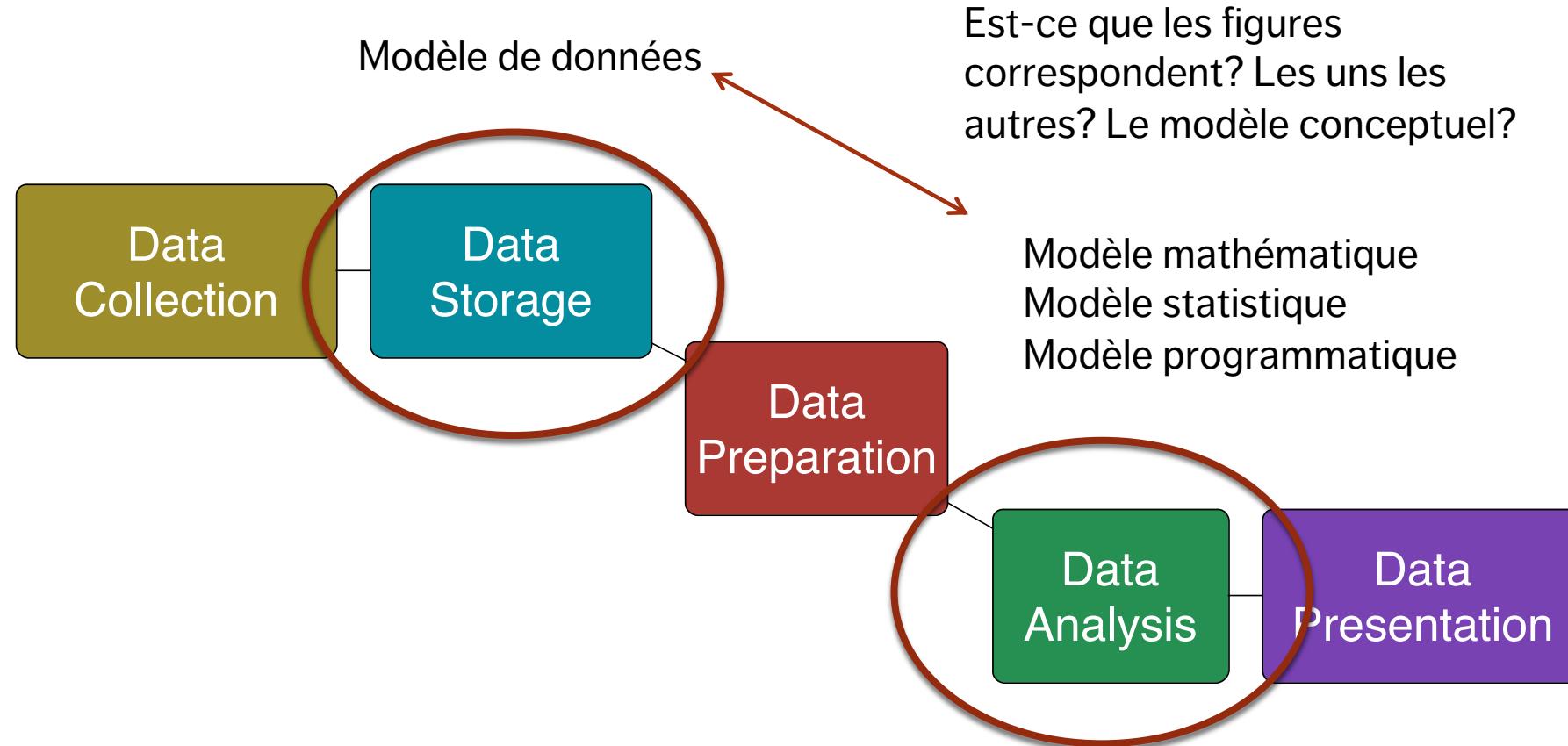
Modèle statistique

Modèle programmatique (informatique)

Ontologie (modèle de connaissances)
(Exercice de simulation)

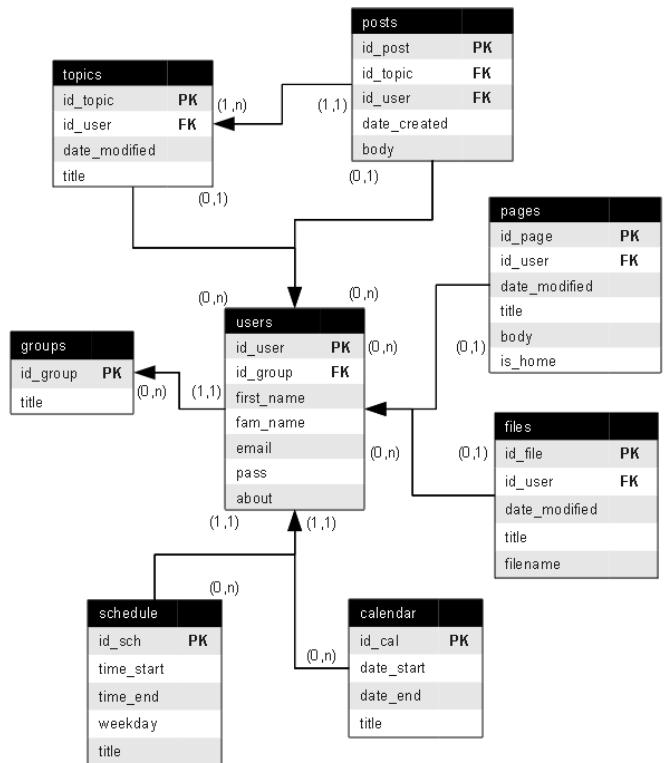
Modèle de données

UN PIPELINE DE DONNÉES (AUTOMATISÉ) MIS EN ŒUVRE



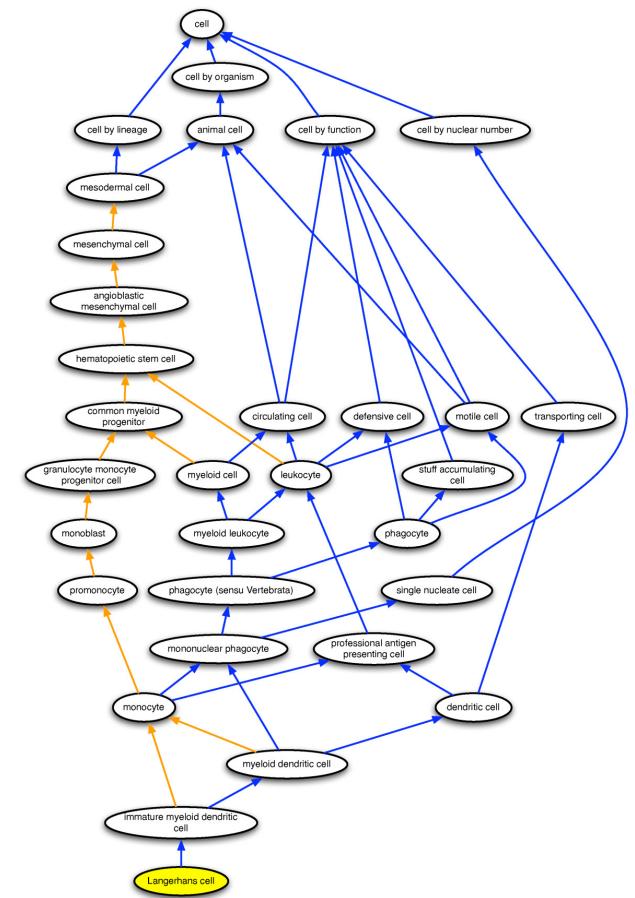
QU'EST-CE QUE LA MODÉLISATION DE DONNÉES?

- Un modèle de données est une description abstraite (logique) d'un système, construite en des termes qui peuvent ensuite être mis en œuvre comme la structure d'un type de logiciel de gestion de données.
- Vous pourriez soutenir que la modélisation de données est à mi-chemin entre un modèle conceptuel et une mise en œuvre de base de données.
- Les données elles-mêmes concernent les **instances** – le modèle concerne les **types d'objets**.
- Une autre option vaut la peine d'être considérée à ce sujet – **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!

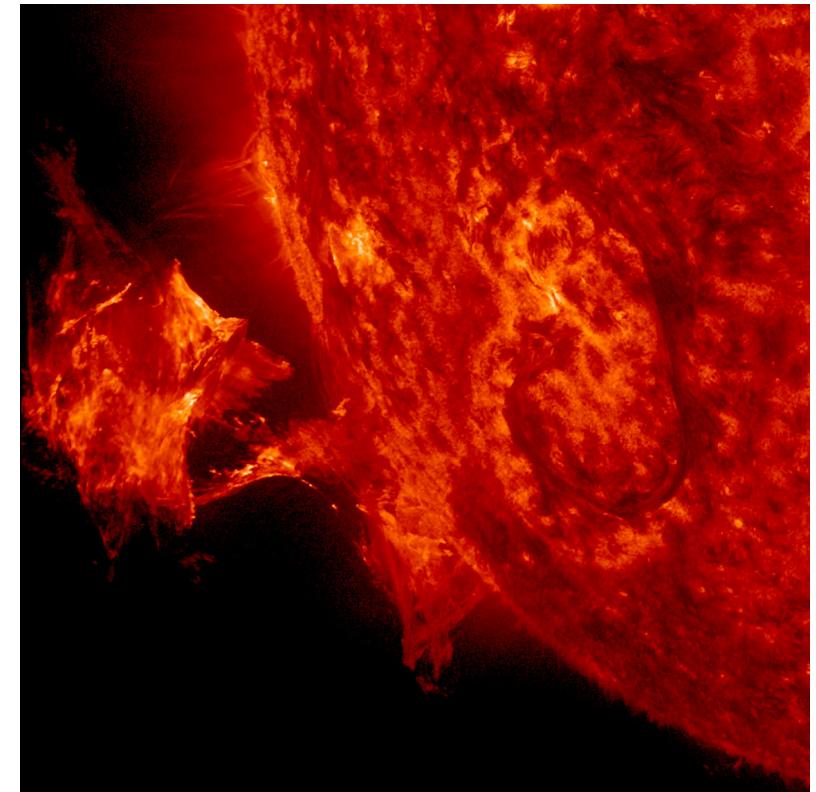
DONNÉES STRUCTURÉES PAR RAPPORT AUX DONNÉES NON STRUCTURÉES

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



QU'EST-CE QUE LA MODÉLISATION DE DONNÉES?

Nous allons examiner quatre options différentes qui sont actuellement populaires en matière de modélisation ou de stratégies de structuration **de données et de connaissances** fondamentales :

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

noSQL

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 et de valeurs, dans laquelle la clé est un nom ou une étiquette (possiblement d'un objet) et la valeur est une valeur qui y est associée.
- **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

mot - définition

URL - page Web

Nom de rapport - rapport (dossier de documentation)

personne - pointure de chaussures

Jen Schellinck - femmes, pointure 7

Colin Henein - hommes, pointure 10

Les triples ajoutent un verbe au mélange :

Personne - est - âge

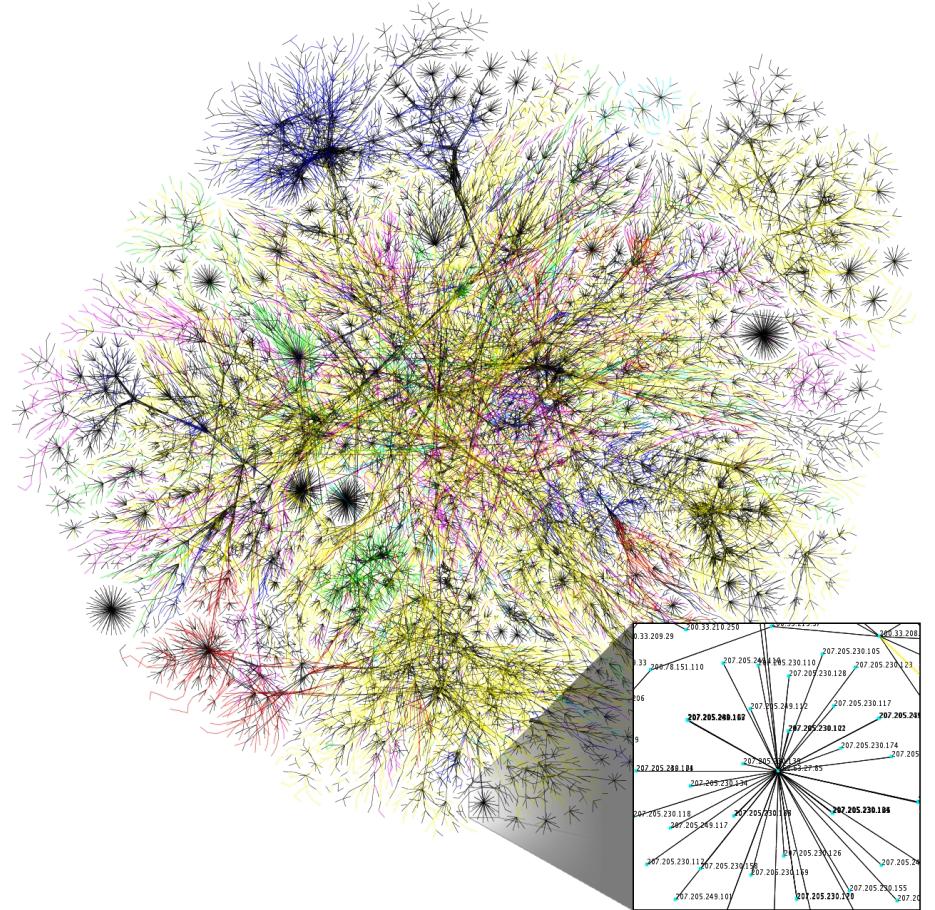
Objet - est de couleur - couleur

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.

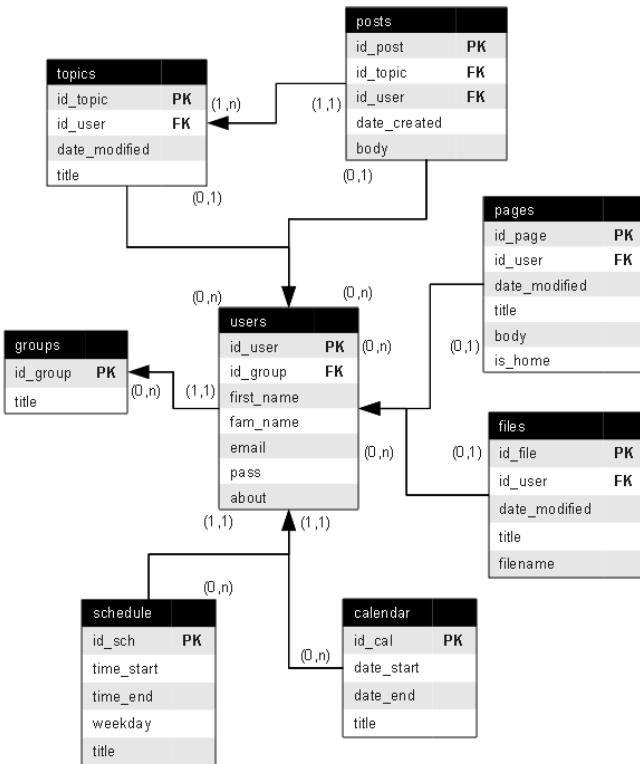
Le modèle de données :

- 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 tableaux relient les instances d'objets dans le tableau.



AVANTAGES, INCONVÉNIENTS ET CAS D'UTILISATION DE CHACUN

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 (malgré son nom).

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.

REMARQUE SUR LES 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 » (par opposition à un mini-entrepôt de données) pour décrire une situation avec plusieurs feuilles de calcul (ponctuelles) comme 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										
10 September 2017	42	42										
1 September 2017	38	43										
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 August 2017	40						40.3	43.4	6.7	3.1	3.5	1.6
11 August 2017	42						40.3	43.6	6.4	3.1	3.4	1.5
1 August 2017	41						40.3	43.8	6.4	3.1	3.4	1.6
19 July 2017	41						40.0	43.8	6.4	3.1	3.4	1.6
18 July 2017	41						40.5	43.8	6.4	3.1	3.0	1.7
16 July 2017	42						40.5	43.8	6.4	3.1	3.0	1.7
15 July 2017	39						40.4	43.9	6.5	3.1	2.8	1.6
14 July 2017	41						40.4	43.8	6.5	3.0	2.9	1.7
11 July 2017	40						40.8	43.5	6.4	2.9	2.7	1.8
6 July 2017	38						40.8	43.4	6.5	2.9	2.7	1.8
3 July 2017	41						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
3 July 2017	41						40.7	44.2	6.3	3.0	2.8	1.7
30 June 2017	41						40.7	44.2	6.3	3.0	2.8	1.7
29 June 2017	39						40.7	44.2	6.3	3.0	2.8	1.7

REMARQUE SUR LES 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 (p. ex. études scientifiques!); comme certains types d'analyse exigent que vous ayez toutes les données en un seul endroit, vous allez devoir générer un fichier non hiérarchique de toute façon. Facile à lire dans le logiciel d'analyse (p. ex. R) et il est facile d'effectuer des opérations quant à l'ensemble de données
- **Inconvénients** - il est très difficile de gérer l'intégrité des données à long terme si vous recueillez et travaillez continuellement avec les données. Ne fonctionne pas bien si vous traitez des données sur un système comprenant de nombreux types d'objets différents et leurs liens. Peut être très difficile d'exécuter des opérations d'interrogation de données

QUELQUES OUTILS ET MOTS À LA MODE

- MongoDB, ArangoDB
- Entrepôt de documents
- JSON, YAML
- API, GraphQL
- Données interreliées
- Web sémantique
- Langage d'ontologie Web (OWL)
- Protected

MISE EN ŒUVRE DE VOTRE MODÈLE

- Pour mettre en œuvre votre modèle de données/connaissances, vous devez avoir accès à un **logiciel de stockage et de gestion de données**.
- L'accès pourrait représenter un défi pour vous en tant que personne, parce que traditionnellement ces logiciels fonctionnent sur des serveurs.
- Les serveurs sont bons parce qu'ils permettent à plusieurs personnes d'accéder à une seule base de données en même temps, à partir de différents programmes clients. Il est toutefois difficile de « jouer » avec une base de données.
- SQLite à la rescousse!

RÔLE DU LOGICIEL DE GESTION DES DONNÉES

Le logiciel de gestion des données offre aux humains un moyen facile d'interagir avec leurs données.

Il s'agit essentiellement d'une interface – de données – humaines.

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.



UN QUOI DE DONNÉES???

Auparavant :

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

Maintenant :

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

De plus en plus de distinction entre la mémoire de données et le logiciel de gestion de données

DU MODÈLE DE DONNÉES À LA MISE EN ŒUVRE

Que faisons-nous une fois que nous avons terminé un modèle de données (logique)?

- instancier le modèle dans le logiciel de votre choix (p. ex : créer des tableaux dans MySQL);
- charger les données.
- **demandez** les données :
 - Les bases de données relationnelles utilisent le langage d'interrogation structuré (SQL)
 - D'autres types de bases de données utilisent des langages de requête totalement différents (AQL, moteurs sémantiques, etc.) ou s'appuient sur des programmes informatiques sur mesure (p. ex. écrits en R, Python).

GESTION DE LA COLLECTION DE DONNÉES

Une fois que vous avez créé une collection de données, vous devez également la gérer! Qu'est-ce que cela signifie? Essentiellement, cela signifie qu'il faut maintenir les données dans la base de données :

- exactes,
- précises,
- uniformes,
- complètes.

Ne laissez pas votre lac de données se transformer en marais de données...