
NOTIONS UNIVERSELLES SUR L'ANALYSE DES DONNÉES

PRÉPARATION DU TERRAIN

APERÇU

1. Données, AA et IA dans l'actualité
2. Données 101 – Notions de données de base
3. Quelques définitions pratiques
4. Flux de travail et sources – le processus de travail avec les données
5. Modèles et pensée systémique
6. Considérations éthiques et pratiques exemplaires

À LA UNE

“Robots are better than doctors at diagnosing some cancers, major study finds”
[The Telegraph, UK, 29-05-2018]

“MRNet: Deep-learning-assisted diagnosis for knee magnetic resonance imaging”
[Stanford ML Group]

“Data scientists find connections between birth month and health” [Columbia University Medical Centre]

“We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually.”
[MIT Technical Review, 21-12-2018]

OBJETS ET ATTRIBUTS



Objet : pomme

Forme : sphérique

Couleur : rouge

Fonction : alimentaire

Emplacement : réfrigérateur

Propriétaire : Jen

Rappelez-vous : une personne ou un objet n'est pas simplement la somme de ses attributs!

DES VARIABLES AUX ENSEMBLES DE DONNÉES

Les attributs sont les **champs** (ou les colonnes) d'une banque de données; les objets en sont les **instances** (ou les rangées).

On décrit un objet à l'aide de son **vecteur-signature**, l'ensemble des valeurs associées à ses attributs.

ID#	Shape	Colour	Function	Location	Owner
1	spherical	red	food	fridge	Jen
2	rectangle	brown	food	office	Pat
3	round	white	tell time	lounge	School
...

ENSEMBLE DE DONNÉES SUR LES CHAMPIGNONS VÉNÉNEUX



Amanita muscaria

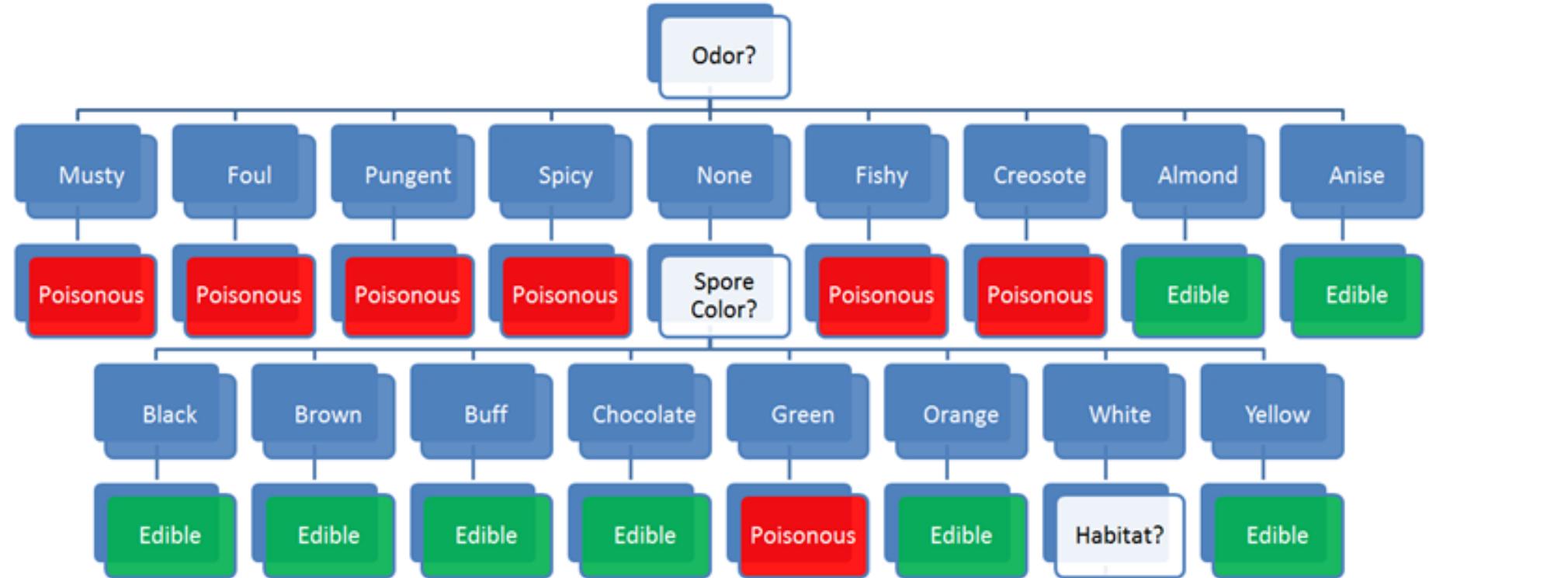
Habitat : bois

Taille du feuillet : étroit

Odeur : aucune

Spores : blancs

Problème de classification : L'*Amanita muscaria* est-il comestible ou vénéneux?



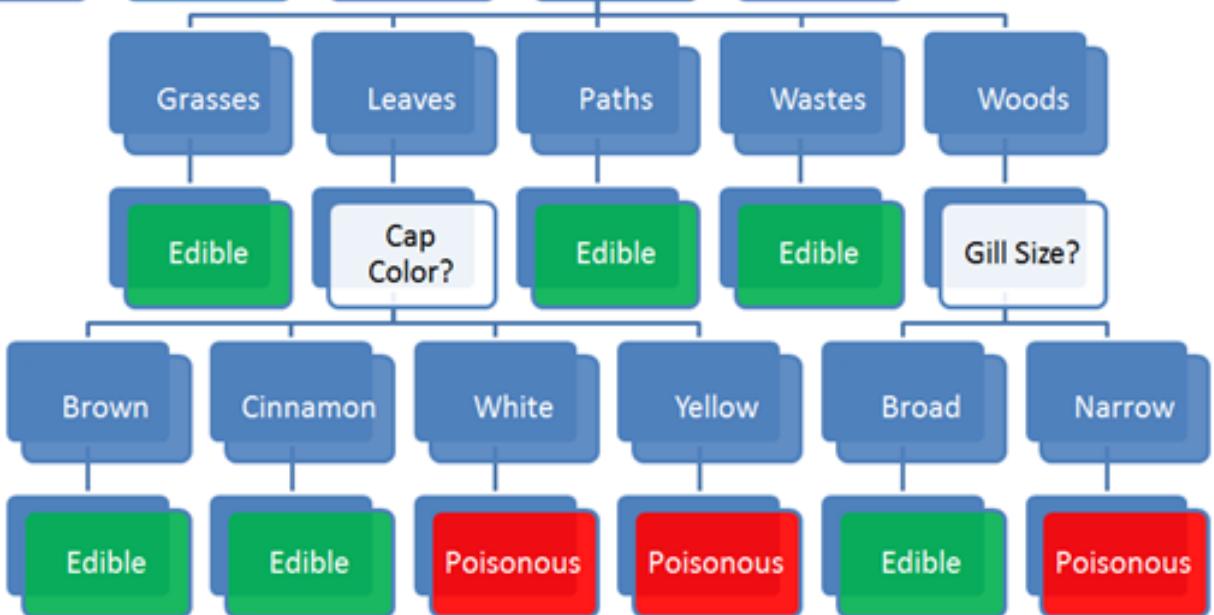
Habitat : bois

Taille du feuillet : étroit

Odeur : aucune

Spores : blancs

Problème de classification :
L'*Amanita muscaria* est-il comestible ou vénéneux?



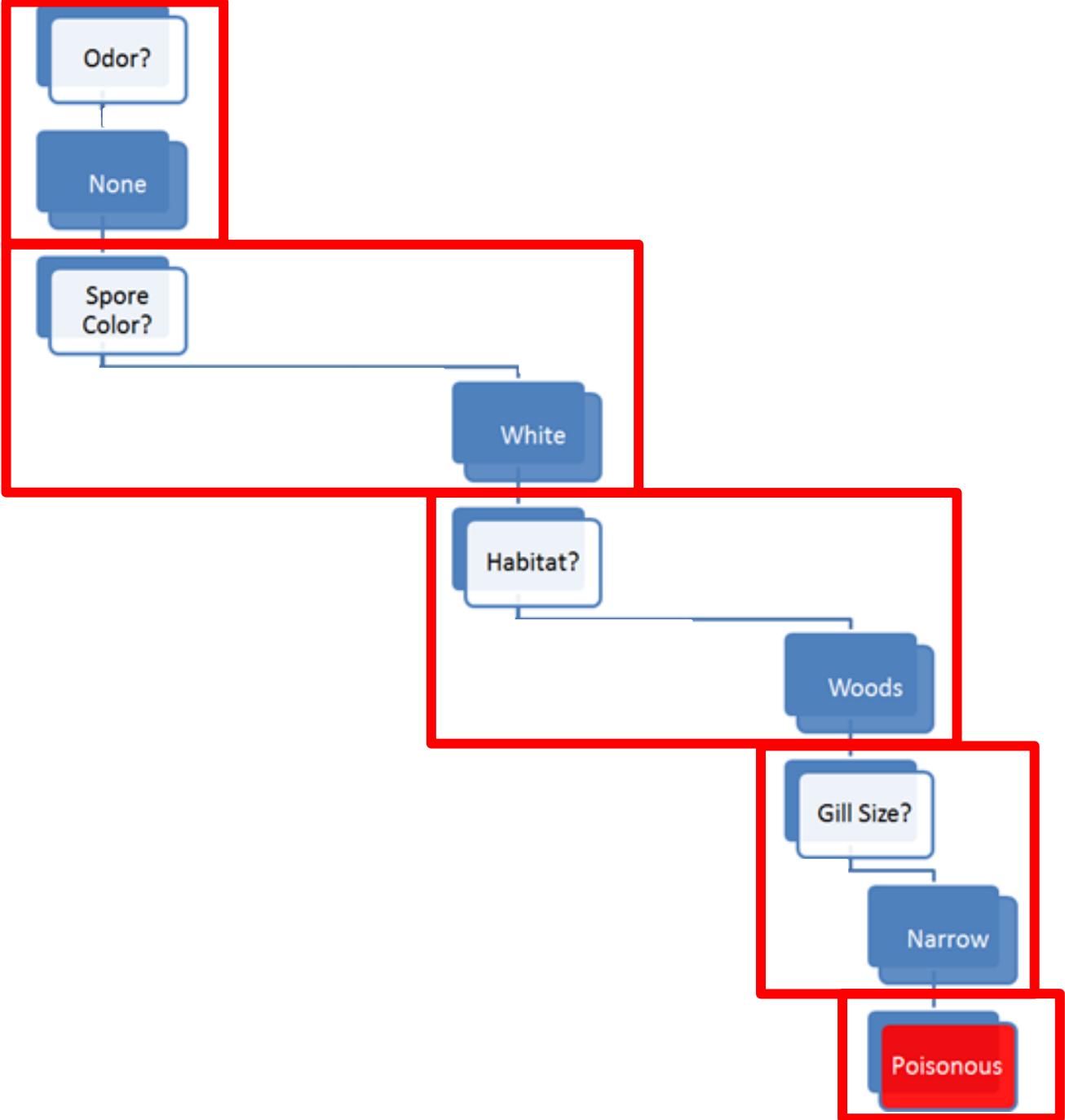
Habitat : bois

Taille du feuillet : étroit

Odeur : aucune

Spores : blancs

Problème de classification :
L'*Amanita muscaria* est-il comestible
ou **vénéneux**?



POSER LES BONNES QUESTIONS

La science des données consiste en réalité à poser des questions et à y répondre :

- **Analytique** : « Combien de fois a-t-on cliqué sur ce lien? »
- **Science des données** : « D'après l'historique des achats de cet utilisateur, puis-je prédire sur quels liens il cliquera la prochaine fois qu'il accèdera au site? »

Les modèles d'exploration/de science des données sont habituellement **prédictifs** (non **explicatifs**) : ils montrent les liens, mais ne révèlent pas pourquoi ils existent.

Attention : Toutes les situations n'exigent pas de faire appel à la science des données, à l'intelligence artificielle, à l'apprentissage automatique ou à l'analyse.

TÂCHES DE LA SCIENCE DES DONNÉES / L'APPRENTISSAGE AUTOMATIQUE / L'I.A.

Classification et **estimation de la probabilité de la classe** : quels clients sont susceptibles d'être des clients réguliers?

Regroupement : les clients forment-ils des groupes naturels?

Découverte de règles d'association : quels sont les livres couramment achetés ensemble?

Autres :

Profilage et description du comportement; prédition des liens; estimation de la valeur (combien un client est-il susceptible de dépenser dans un restaurant); **appariement des similitudes** (quels clients potentiels sont semblables aux meilleurs clients d'une entreprise?); **réduction des données; modélisation de l'influence et modélisation causale**, etc.

QU'EST-CE QUE L'ANALYSE DES DONNÉES?

Trouver **des tendances** dans les données

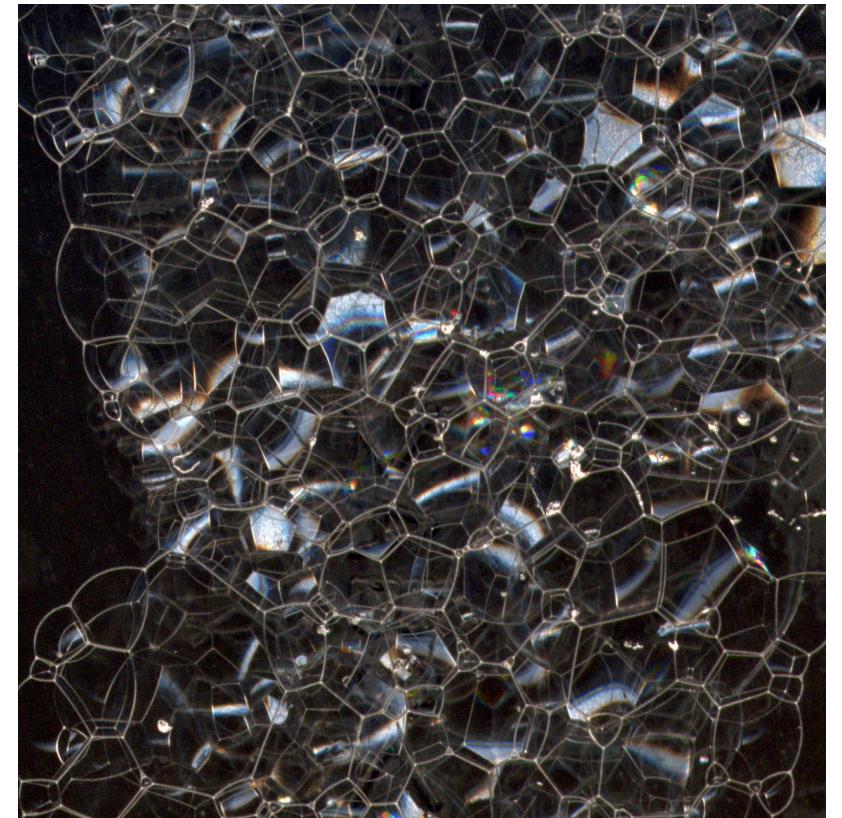
Utiliser les données pour faire quelque chose
(répondre à une question, aider à la prise de décision,
prédir l'avenir, tirer une conclusion)

Créer des modèles à partir de vos données

Décrire ou expliquer votre situation (votre **système**)

(Tester des hypothèses [scientifiques]?)

(Effectuer des calculs à partir des données?)



Plus la tendance est compliquée, plus
l'analyse est compliquée.

QU'EST-CE QUE LA SCIENCE DES DONNÉES?

La science des données est l'ensemble des processus par lesquels nous extrayons
des informations utiles et exploitables des données.

T. Kwartler (paraphrasé)

La science des données est l'**intersection pratique** de la statistique, de l'ingénierie, de l'informatique, de l'expertise du domaine et du « piratage ». Elle s'articule autour de deux axes principaux : l'**analyse** (compter les choses) et l'**invention de nouvelles techniques** pour tirer des enseignements des données.

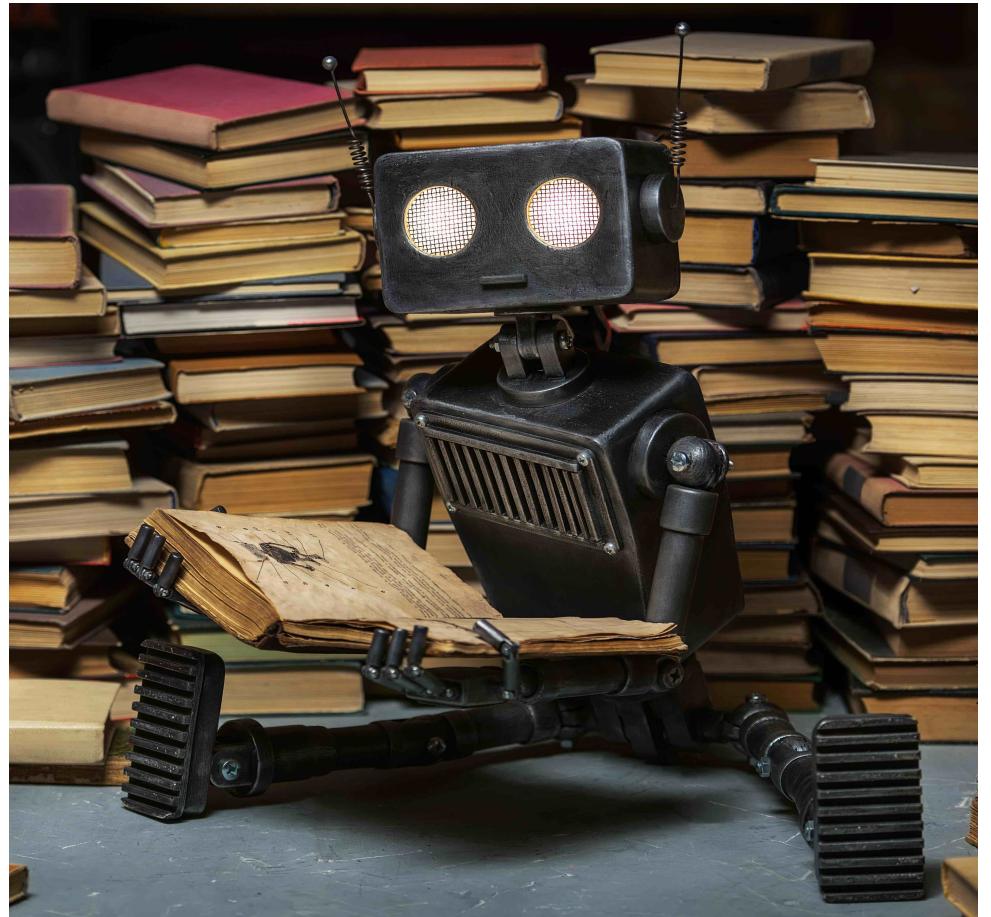
H. Mason (paraphrasé)

QU'EST-CE QUE L'APPRENTISSAGE AUTOMATIQUE?

À partir des années 1940, les chercheurs ont commencé à enseigner sérieusement aux machines comment apprendre.

Le but de l'**apprentissage automatique** était de créer des machines capables d'apprendre, de s'adapter et de répondre à des situations nouvelles.

De nombreuses techniques, accompagnées d'un grand nombre de fondements théoriques, ont été créées dans le but d'atteindre cet objectif.



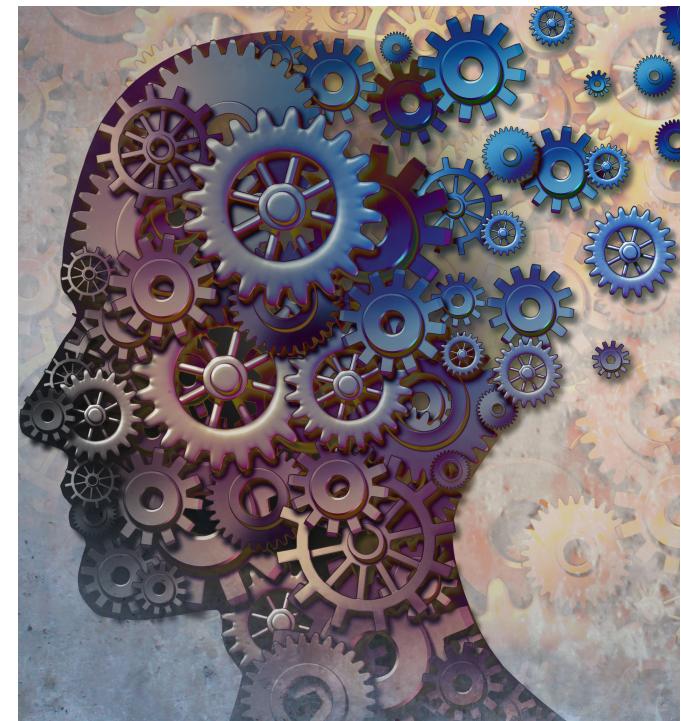
QU'EST-CE QUE L'INTELLIGENCE ARTIFICIELLE/AUGMENTÉE?

L'intelligence artificielle (I.A.) est une intelligence non humaine qui a été conçue par l'ingénierie plutôt qu'une intelligence qui a évolué naturellement.

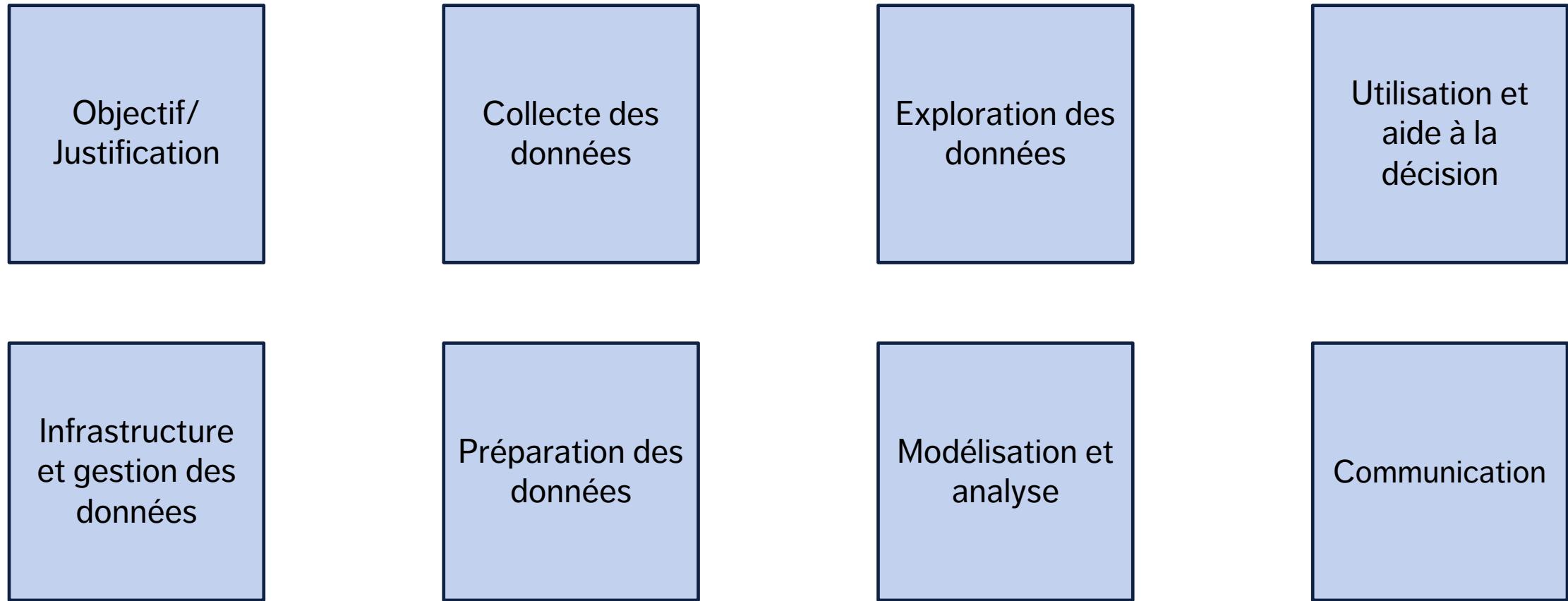
La recherche en intelligence artificielle est une recherche menée dans ce but.

Pragmatiquement parlant, l'I.A. est « un ordinateur qui exécute des tâches que seuls les humains peuvent habituellement accomplir. »

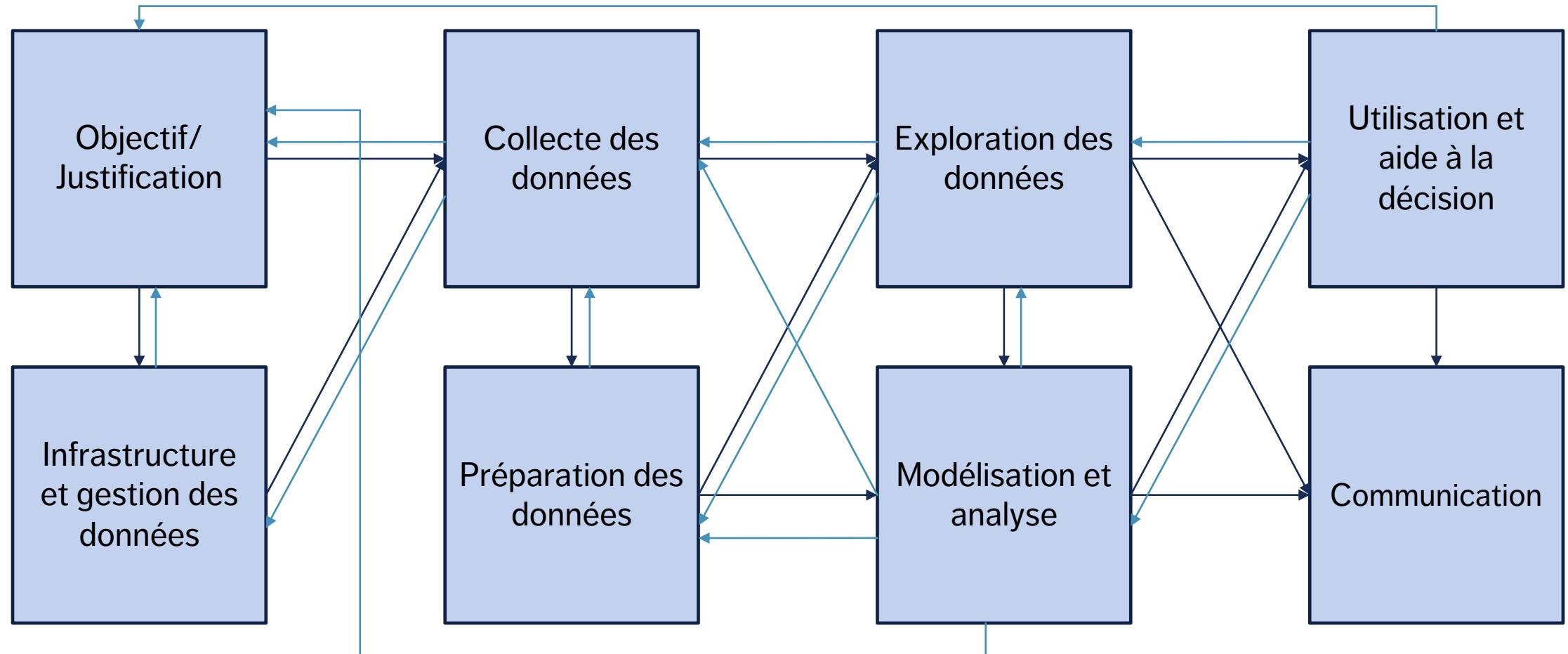
L'intelligence augmentée est l'intelligence humaine qui est soutenue ou améliorée par l'intelligence artificielle.



LE « FLUX DE TRAVAIL » DES DONNÉES



LE « FLUX DE TRAVAIL » DE LA SCIENCE DES DONNÉES



LA SUITE DES CHOSES APRÈS L'ANALYSE

Lorsqu'une analyse ou un modèle est « lâché dans la nature », il peut avoir une vie propre.

Les analystes pourraient éventuellement devoir abandonner le contrôle de la diffusion. Les résultats pourraient être détournés, mal compris ou mis au rancart. Que peut faire l'analyste pour éviter cela?

Enfin, en raison de la **décomposition analytique**, il est important de ne PAS considérer la dernière étape analytique comme une impasse statique, mais plutôt comme une invitation à revenir au début du processus.

ÉCOSYSTÈME DE LA SCIENCE DES DONNÉES

L'analyse des données est un **sport d'équipe**, les membres de l'équipe ayant besoin d'une bonne compréhension des **données** et du **contexte**.

- Gestion des données
- Préparation des données
- Analyse
- Communications

Même de légères améliorations par rapport à l'approche actuelle peuvent trouver une place utile dans une organisation – **la science des données ne concerne pas seulement les mégadonnées et les perturbations!**

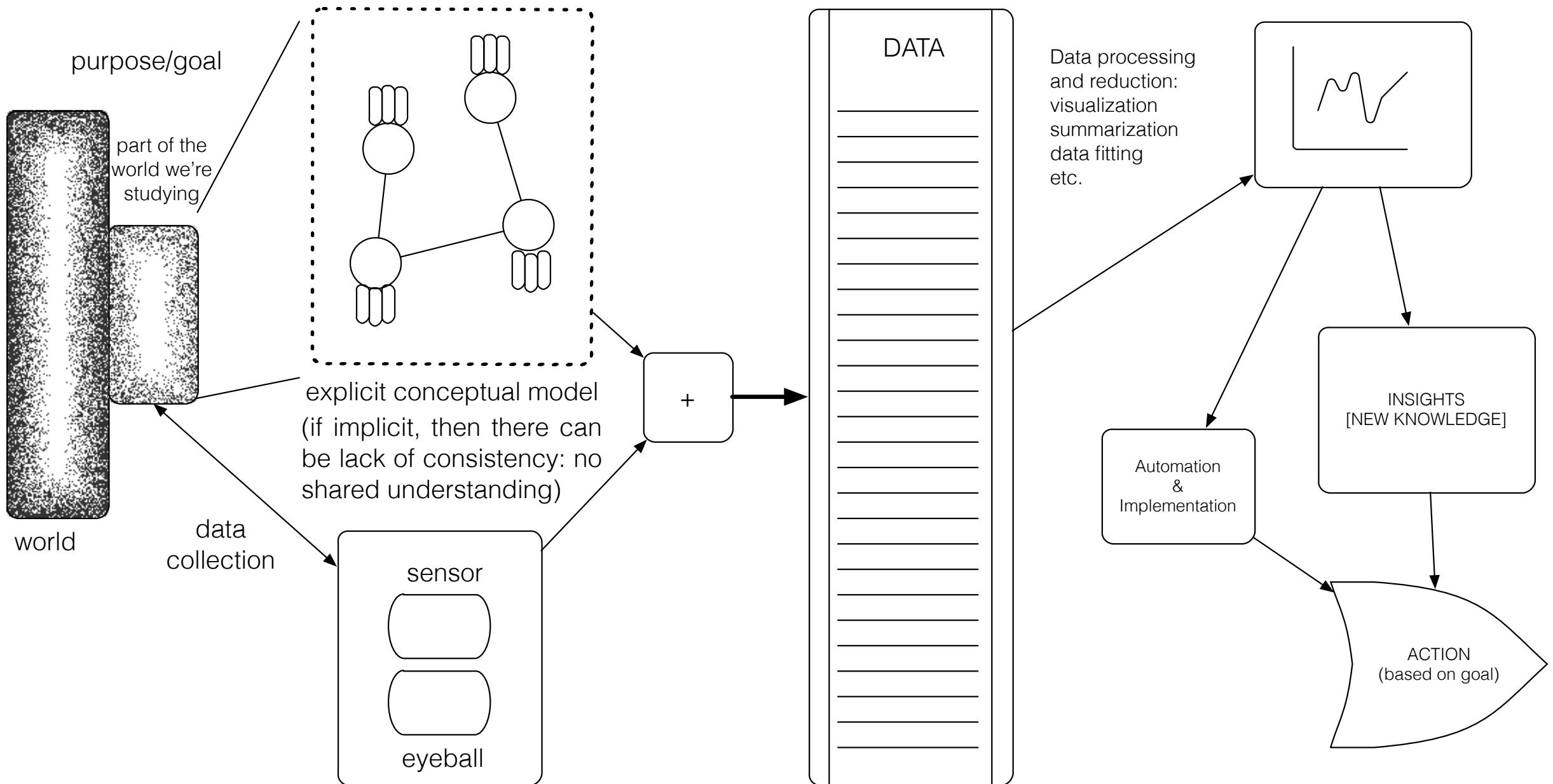
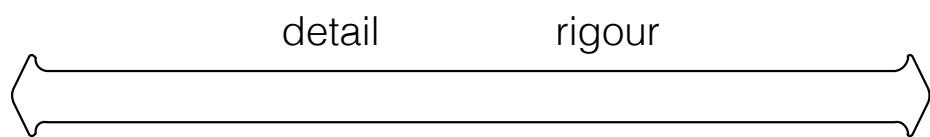
REPRÉSENTATION

Une représentation est un objet qui remplace un autre objet.

Une représentation peut ou non ressembler physiquement à l'objet qu'elle représente.

Les représentations du monde nous aident à comprendre, à naviguer et à manipuler le monde.





PENSER EN TERMES SYSTÉMIQUES

Afin de comprendre comment les divers aspects du monde interagissent les uns avec les autres, nous devons **découper des morceaux** correspondant à ces aspects et définir leurs **limites**.

Travailler avec d'autres intelligences exige une **compréhension commune** de ce qui est étudié.

Un **système** est composé d'**objets** dont les **propriétés** peuvent changer avec le temps. Au sein du système, nous percevons **des actions** et des **propriétés évolutives** qui nous amènent à penser en termes de **processus**.

PENSER EN TERMES SYSTÉMIQUES

Les **objets** eux-mêmes ont diverses propriétés. Les processus naturels génèrent (ou détruisent) des objets et peuvent modifier leurs propriétés avec le temps.

Nous **observons**, **quantifions** et **enregistrons** des valeurs particulières de ces propriétés à des moments précis.

Cela génère des points de données, saisissant la **réalité sous-jacente** avec un certain degré d'**exactitude** et d'**erreur** (biaisée ou non).

DÉTERMINER LES LACUNES DANS LES CONNAISSANCES

Une **lacune dans les connaissances** est déterminée lorsque nous nous rendons compte que ce que nous pensions savoir sur un système s'avère incomplet (ou faux).

Cela peut se répéter à n'importe quel moment du processus :

- Nettoyage des données
- Consolidation des données
- Analyse des données

La solution doit être flexible. Face à une telle lacune, **revenez en arrière, posez des questions et modifiez la représentation du système.**

MODÈLES CONCEPTUELS

Exercice :

- Imaginez qu'une connaissance vient d'entrer pour la première fois dans votre espace de vie.
- Vous êtes au téléphone avec elle, mais vous n'êtes pas à la maison en ce moment.
- Expliquez-lui comment préparer une tasse de sucre.

Les **modèles conceptuels** sont construits à l'aide d'outils d'analyse méthodique.

- Schémas
- Entrevues structurées
- Descriptions structurées
- Autres

RELATION ENTRE LES DONNÉES ET LE SYSTÈME

Les données recueillies et analysées seront-elles utiles pour comprendre le système?

On ne peut répondre à cette question que si nous comprenons :

- La façon dont les données sont **recueillies**
- La **nature approximative** des données et du système
- Ce que les données **représentent** (observations et caractéristiques)

La combinaison du système et des données **est-elle suffisante** pour comprendre les aspects du monde à l'étude?

LE BESOIN D'ÉTHIQUE

Autrefois : mentalité « **Far West** » dans la collecte (et l'utilisation) des données. Tout ce qui n'était pas technologiquement interdit était autorisé.

Aujourd'hui : des codes de conduite professionnels sont en cours d'élaboration pour les scientifiques des données (définir des façons responsables de pratiquer la science des données).

Responsabilité **supplémentaire** pour les scientifiques des données; mais aussi **protection** contre l'embauche en vue d'effectuer des analyses douteuses.

Votre organisation dispose-t-elle d'un code de déontologie pour ses scientifiques des données? Pour ses employés?

QU'EST-CE QUE L'ÉTHIQUE?

En termes généraux, l'éthique fait référence à l'**étude** et à la **définition** des **bonnes et des mauvaises conduites** :

- « ce n'est pas [...] des conventions sociales, des croyances religieuses ou des lois ». (R.W. Paul, L. Elder) [Traduction]

Théories éthiques *occidentales* influentes :

- La **règle d'or** de Kant (traite les autres comme...), le **conséquentialisme** (la fin justifie les moyens), l'**utilitarisme** (agir pour maximiser l'effet positif), etc.

Théories éthiques *orientales* influentes :

- **Confucianisme, taoïsme, bouddhisme (?)**, etc.

QU'EST-CE QUE L'ÉTHIQUE?

Principes de PCAP® des Premières Nations :

- **Propriété**

Les communautés des Premières Nations sont propriétaires de leur savoir culturel, de leurs données et des renseignements les concernant.

- **Contrôle**

Les communautés des Premières Nations ont le droit de contrôler l'intégralité de la recherche et de la gestion de l'information les concernant.

- **Accès**

Les communautés des Premières Nations doivent avoir accès aux renseignements et aux données les concernant, peu importe où ils se trouvent.

- **Possession**

Les communautés des Premières Nations doivent avoir le contrôle matériel des données pertinentes.

L'ÉTHIQUE DANS LE CONTEXTE DES DONNÉES

Questions relatives à l'éthique des données :

- **Qui**, le cas échéant, possède les données?
- Y a-t-il des **limites** à l'utilisation des données?
- Certaines analyses comportent-elles des **biais fondés sur les valeurs?**
- Y a-t-il des catégories qui ne devraient **pas** être utilisées dans l'analyse des données personnelles?
- Certaines données devraient-elles être **divulguées à tous** les chercheurs?

D'un point de vue analytique, on préfère le **général à l'empirique** – les décisions prises sur la base de l'apprentissage automatique et de l'I.A. (sécurité, finances, marketing, etc.) peuvent toucher des personnes réelles de **manière imprévisible**.

PRATIQUES EXEMPLAIRES

« Ne pas nuire » : Les données recueillies auprès d'une personne ne **doivent pas être utilisées pour lui nuire.**

Consentement éclairé :

- Les personnes doivent **consentir à la collecte et à l'utilisation** de leurs données.
- Les personnes doivent avoir une **compréhension réelle de ce à quoi ils consentent** et des **conséquences possibles** pour elles et pour les autres.

Respect de la « vie privée » : excessivement difficile à maintenir à l'ère du ratissage constant d'Internet à la recherche de données personnelles.

PRATIQUES EXEMPLAIRES

Garder les données publiques : les données devraient être gardées **publiques** (Toutes? La plupart? N'importe lesquelles?).

Inclusion/exclusion : le consentement éclairé exige la possibilité de **se retirer**.

Données anonymisées : suppression des champs d'identification des données avant l'analyse.

« Laissons parler les données » :

- Pas de picorage
- Importance de la validation (nous y reviendrons plus tard)
- Corrélation et causalité (nous y reviendrons plus tard également)
- Répétabilité

ÉVALUATION ET VALIDITÉ DU MODÈLE

Les modèles doivent être à **jour**, **utiles** et **valides**.

Les données peuvent être utilisées en conjonction avec les modèles existants pour arriver à certaines conclusions, ou peuvent être utilisées pour mettre à jour le modèle lui-même.

À quel moment détermine-t-on que le modèle de données actuel n'est plus à **jour** ou qu'il **n'est plus utile**?

Les succès passés peuvent entraîner une **réticence** à repenser et à réévaluer un modèle.