Introduction à la science des données

Instructeur: Patrick Boily

uOttawa

Institut de développement professionnel
Professional Development Institute

Patrick Boily

Carrière :

Professeur [uOttawa] (~55 cours/ ~150 journées d'atelier)

Gérant ['12 – '19, CQADS/CAQAD, Carleton]

Fonctionnaire ['08 – '12, ASFC | StatCan | TC | TPSGC]

Clients :

AMC, SGDN, ACSTA, plusieurs autres (~40 projets)

Spécialités :

Visualisation des données, nettoyage des données, application d'un large éventail de méthodes quantitatives.

PRINCIPES FONDAMENTAUX DE L'ANALYSE DES DONNÉES

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

[En collaboration avec Jen Schellinck | Sysabee]

« Les rapports qui disent que quelque chose ne s'est pas passé sont toujours intéressants pour moi, parce que, comme nous le savons, il y a des **connus connus**; des choses connues comme étant connues. Nous savons aussi qu'il y a des **connus inconnus**, c'est-à-dire, qu'il y a des choses que nous savons que nous ne savons pas. Mais il y a aussi des **inconnus inconnus**, des choses que nous ne savons pas que nous ne savons pas.» [Traduction]

Donald Rumsfeld, point de presse du Département de la défense des États-Unis, 2002

PLANIFICATION DES ANALYSES

PRINCIPES FONDAMENTAUX DE L'ANALYSE DES DONNÉES

« Les plans ne valent rien; la planification compte pour tout. »

Dwight D. Eisenhower

[CDC, Creating an Analysis Plan]

APERÇU DU PLAN D'ANALYSE

Formuler des questions/hypothèses de recherche

Identifier les ensembles de données nécessaires (et disponibles)

Établir des critères d'inclusion/exclusion pour les données/observations.

Sélectionner les variables à utiliser dans les analyses

Choisir les méthodes et logiciels statistiques

DONNÉES 101 – NOTIONS DE DONNÉES DE BASE

PRINCIPES FONDAMENTAUX DE L'ANALYSE DES DONNÉES

« Vous pouvez avoir des données sans information, mais vous ne pouvez pas avoir d'information sans données. » [Traduction]

Attribué à Daniel Keys Moran

QU’EST-CE QU’UNE DONNÉE?

4,529 « rouge » « Y » 25.782

OBJETS ET ATTRIBUTS

Object : pomme

Forme : sphérique

Couleur : rouge

Fonction : alimentaire

Emplacement : réfrigérateur

Propriétaire : Jen

Rappel : une personne ou un objet n’est pas simplement la somme de ses attributs!

DES VARIABLES AUX 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
…	…	…	…	…	…

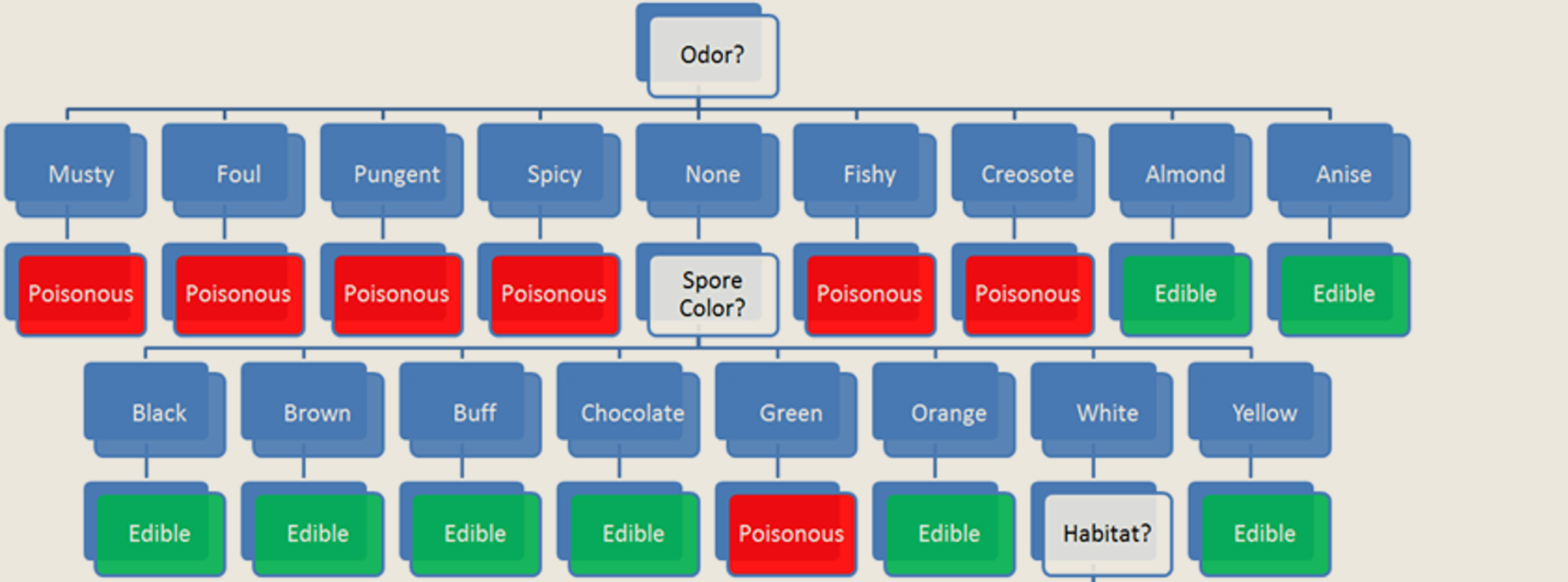
ENSEMBLE DE DONNÉES SUR LES CHAMPIGNONS VÉNÉNEUX

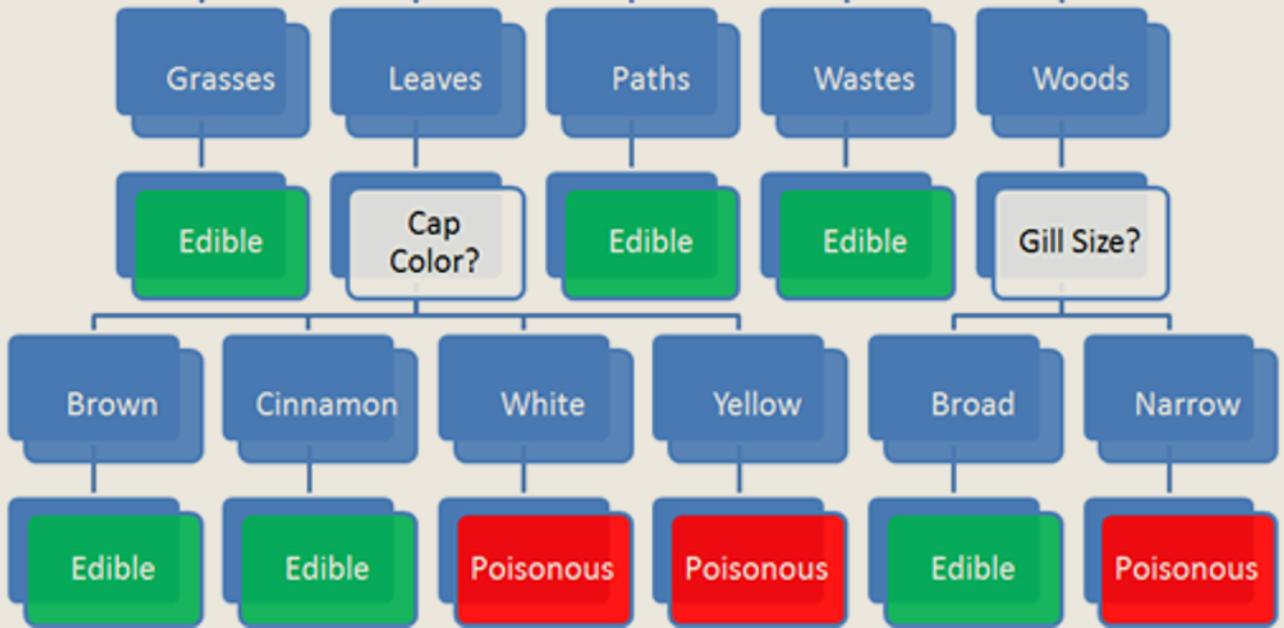
Amanita muscaria

Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : blancs
Couleur : rouge

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

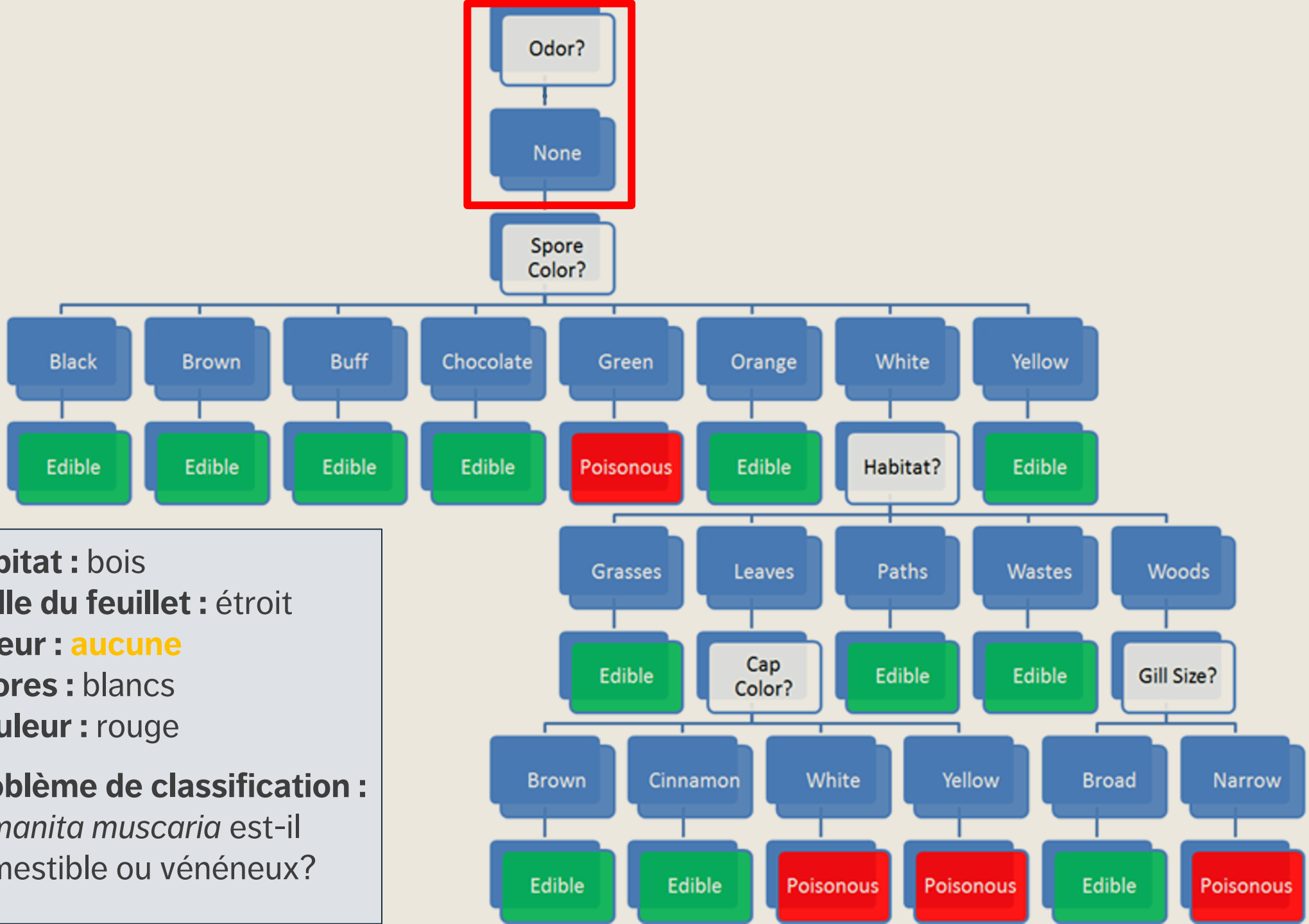
[https://archive.ics.uci.edu/ml/datasets/Mushroom]





Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : blancs
Couleur : rouge

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

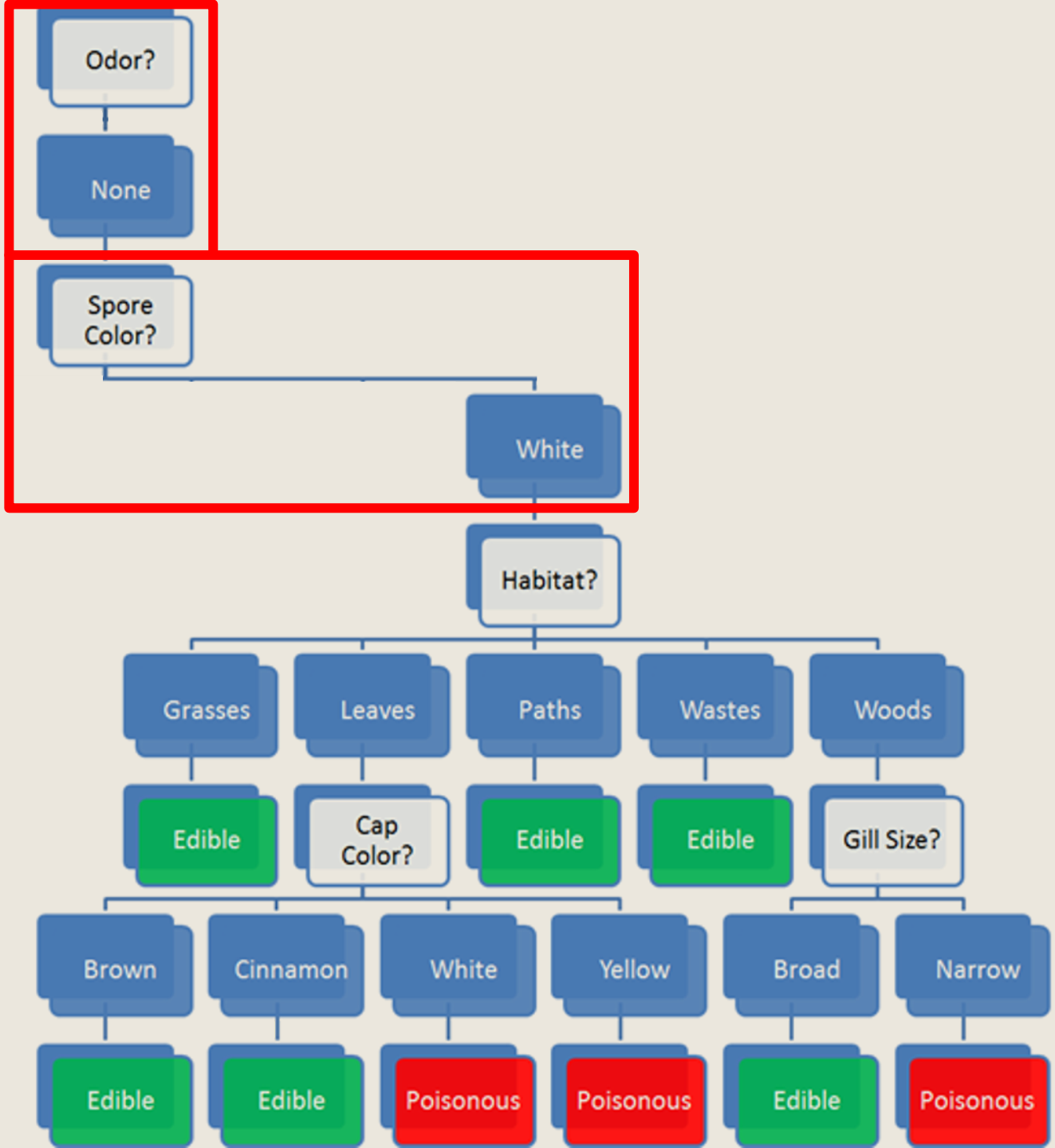


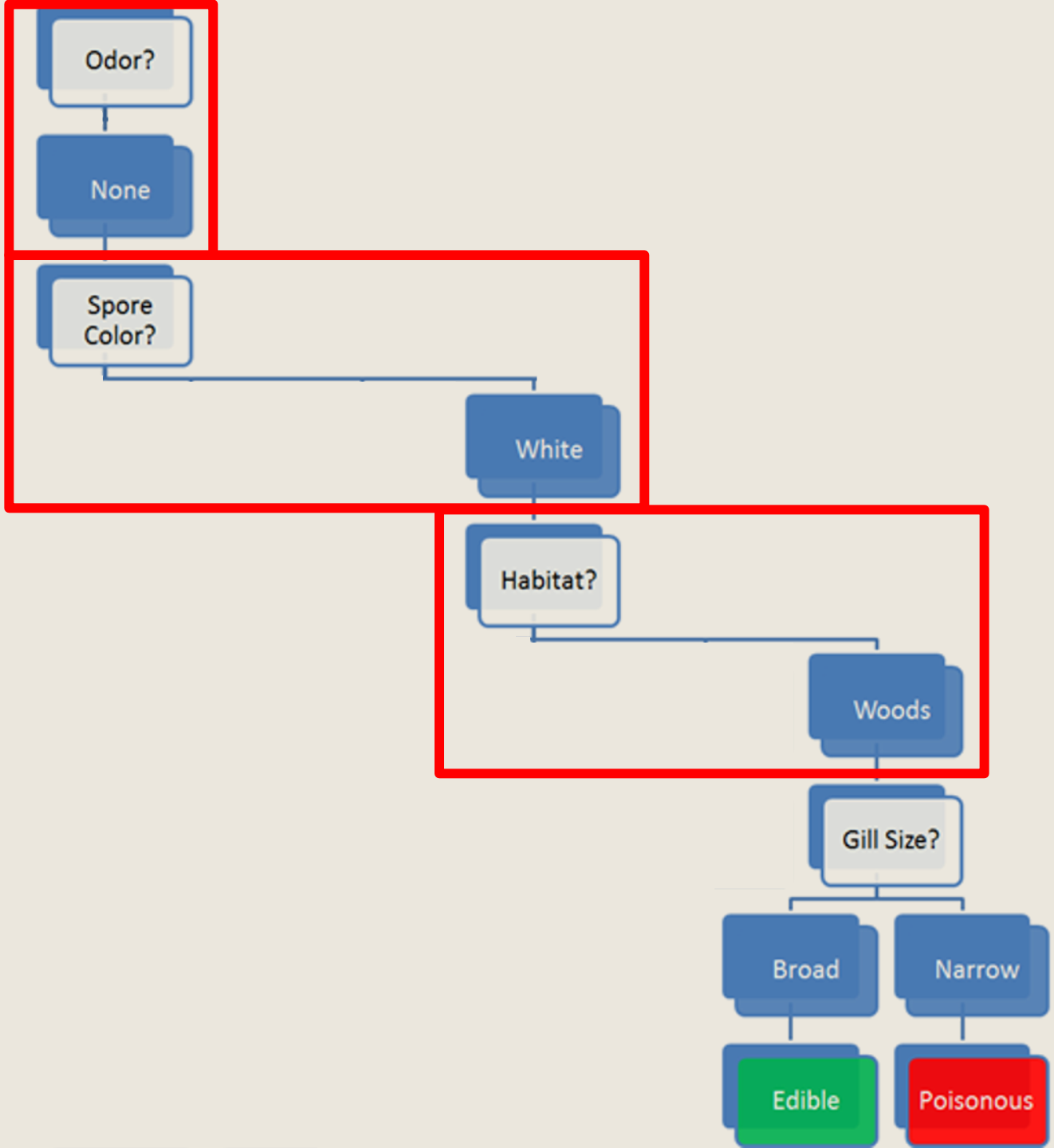
Habitat : bois
Taille du feuillet : étroit
Odeur : **aucune**
Spores : blancs
Couleur : rouge

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

Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : **blancs**
Couleur : rouge

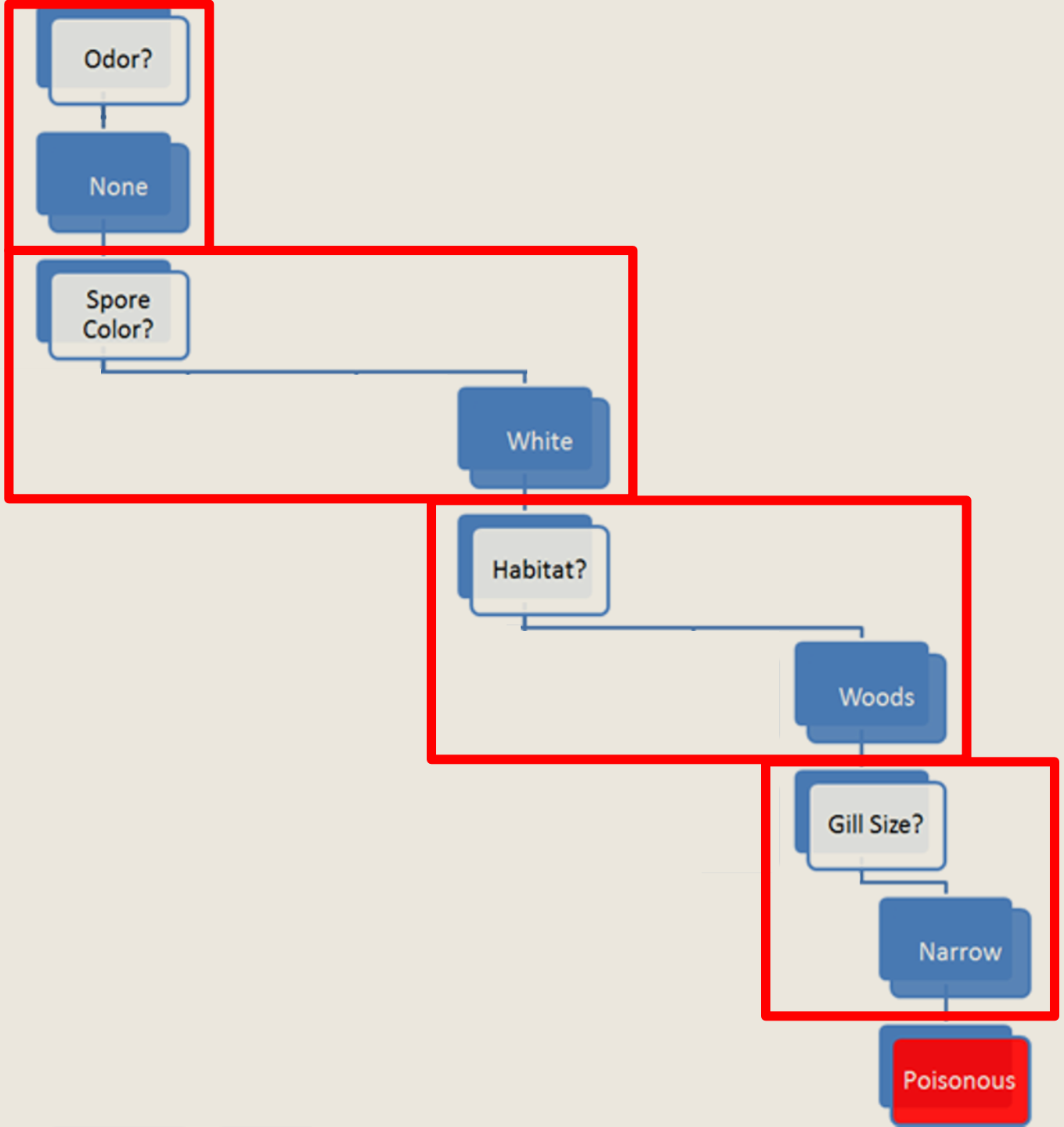
Problème de classification : l'*Amanita muscaria* est-il comestible ou vénéneux?





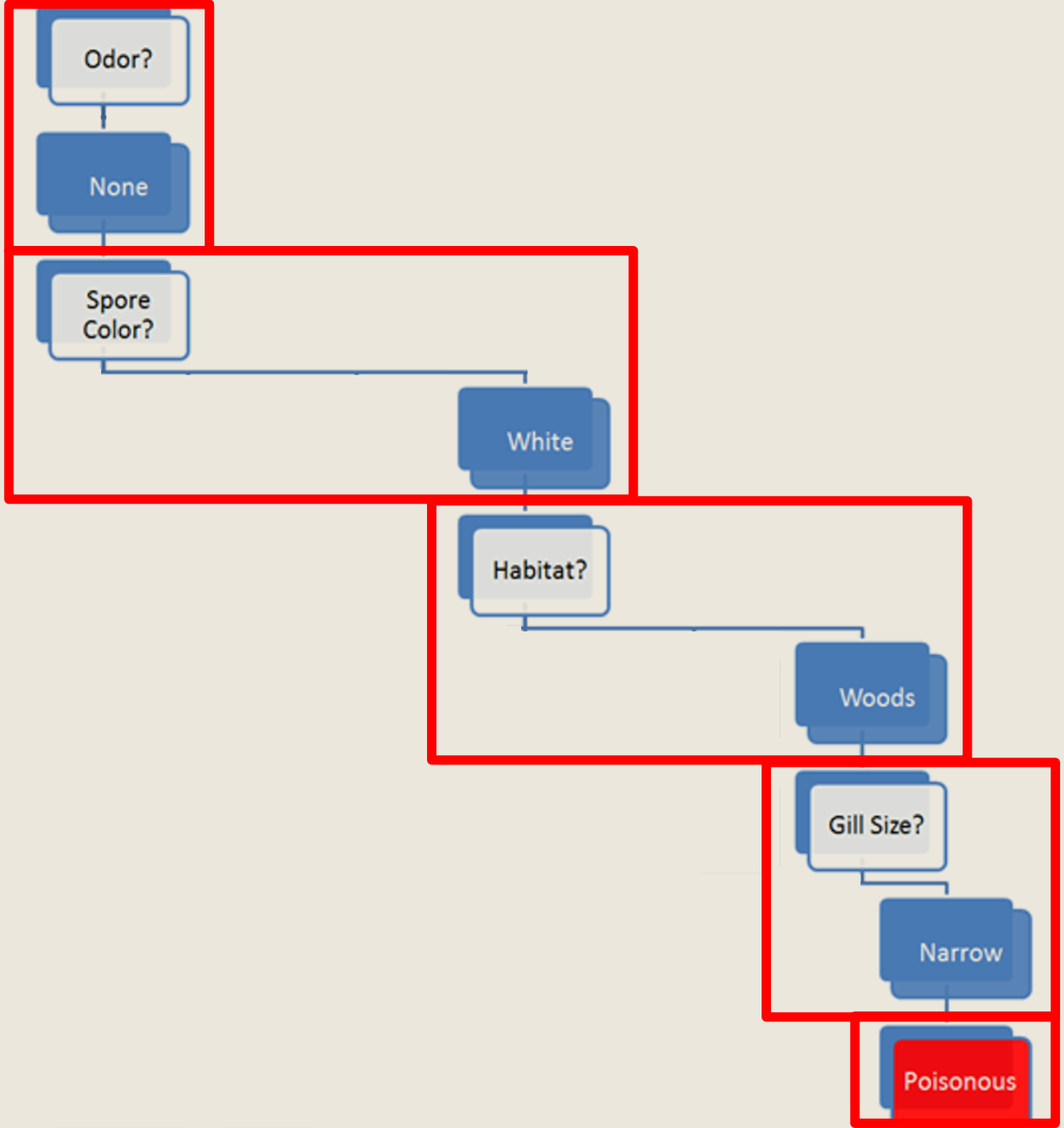
Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : blancs
Couleur : rouge

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



Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : blancs
Couleur : rouge

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



Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : blancs
Couleur : rouge

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

DISCUSSION

Feriez-vous confiance à une prédiction disant que l'*Amanita muscaria* est « **comestible** »?

D'où vient le modèle?

Que devez-vous savoir pour faire confiance au modèle?

Quel est le coût d'une erreur de classification, dans ce cas-ci?

POSER LES BONNES QUESTIONS

La science des données consiste 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.

LES MAUVAISES QUESTIONS

Trop souvent, les analystes posent les **mauvaises questions** :

- des questions **trop vagues** ou **trop restraintes**
- des questions auxquelles **aucune quantité de données ne pourrait répondre**
- des questions pour lesquelles il est **impossible d'obtenir des données**

Dans le **meilleur des cas**, les parties prenantes reconnaîtront que les réponses ne sont pas pertinentes.

Dans le **pire des cas**, elles mettront en œuvre des politiques ou prendront des décisions erronées sur la base de réponses qui n'auront pas été identifiées comme trompeuses et/ou inutiles.

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?

Règles d'association : quels sont les livres couramment achetés ensemble?

Autres :

Profilage et description du comportement; **prédiction 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.

QUELQUES DÉFINITIONS PRATIQUES

PRINCIPES FONDAMENTAUX DE L'ANALYSE DES DONNÉES

«Que révèle un nom? Ce que nous appelons une rose
Par n'importe quel autre nom sentirait aussi bon. »

W. Shakespeare, Roméo et Juliette, acte II, scène 2

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, prédire 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?)

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)

FLUX DE TRAVAIL ET SOURCES

PRINCIPES FONDAMENTAUX DE L'ANALYSE DES DONNÉES

« Tous les modèles sont faux.
Certains modèles sont utiles. »

George Box

[Statistique Canada, *Étapes du cheminement des données*]

LE « FLUX DE TRAVAIL » DE LA SCIENCE DES DONNÉES

Objectif/ Justification

Collecte des données

Exploration des données

Utilisation et aide à la décision

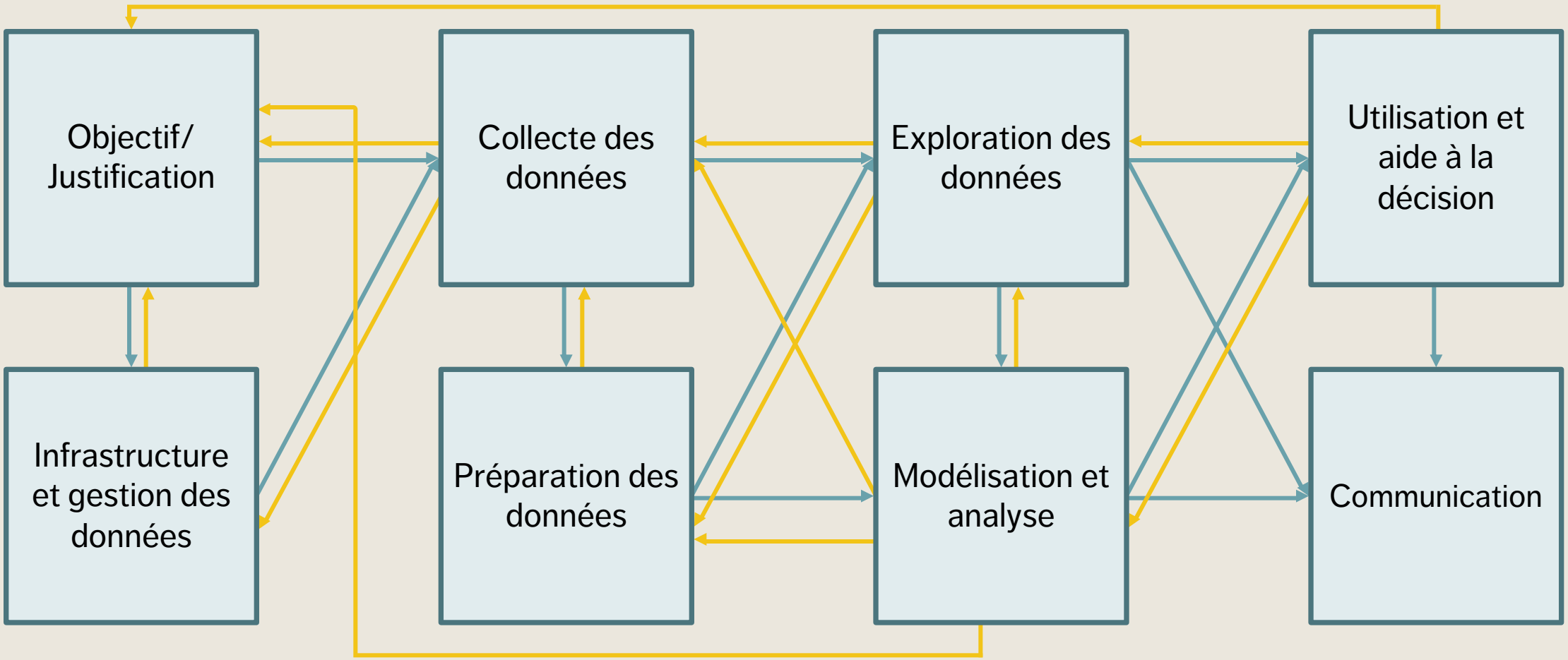
Infrastructure et gestion des données

Préparation des données

Modélisation et analyse

Communication

LE « FLUX DE TRAVAIL » DE LA SCIENCE DES DONNÉES



LE PROCESSUS D'ANALYSE DES DONNÉES

Un **grand nombre de modèles analytiques** doivent être générés avant qu'une sélection finale puisse être faite.

Processus itératif : la sélection des caractéristiques et la réduction des données peuvent nécessiter de nombreuses visites chez des experts du domaine avant que les modèles commencent à donner des résultats prometteurs.

Les connaissances spécifiques à un domaine doivent être intégrées dans les modèles afin d'éliminer les classificateurs aléatoires et les schémas de regroupement, **en moyenne**.

LA VIE 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, 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!**

*ÉVALUATION ET VALIDITÉ DU MODÈLE

Les modèles doivent être **actuels**, **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 est **dépassé** ou qu'il n'**est plus utile** ?

Les succès passés peuvent entraîner une réticence à la ré-évaluation.

MODÈLES ET PENSÉE SYSTÉMIQUE

PRINCIPES FONDAMENTAUX DE L'ANALYSE DES DONNÉES

« Et si le seul modèle valide de l'univers était l'univers lui-même? »

Inconnu

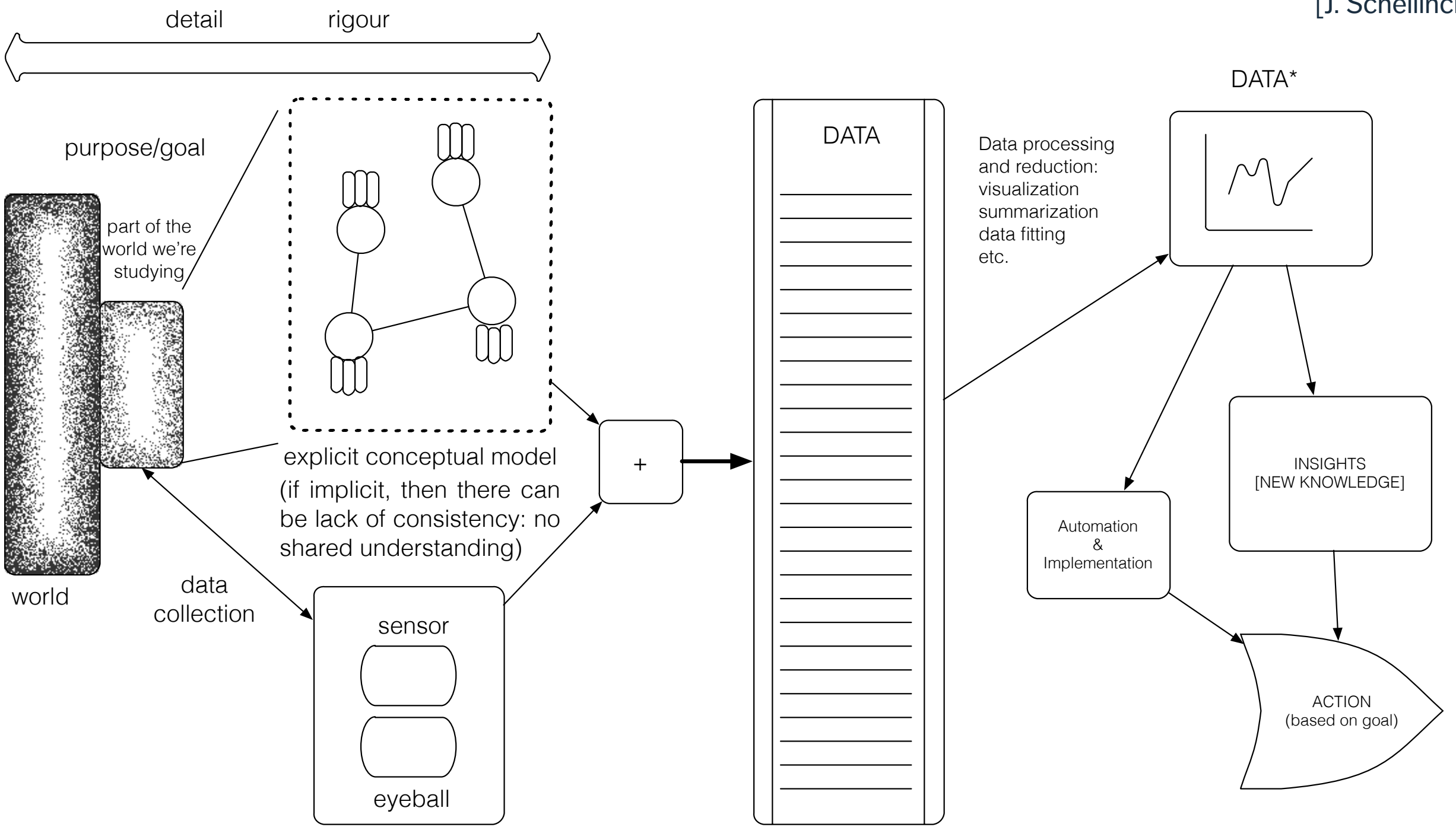
REPRÉSENTATION

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.

[J. Schellinck]



PENSER EN TERMES SYSTÉMIQUES

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**.

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 entre pour la 1ière 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 réparer un fusible.
- (Comment le feriez-vous si la connaissance était malvoyante ?)

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?

[http://dailycatdrawings.tumblr.com; #309]

Le monde réel

Théorie

Modèle

Identification de détails pertinents pour la **description** et la **traduction** d'objets du monde réel en variables de modèle.

À RETENIR

Les systèmes peuvent se rapprocher de certains aspects de l'Univers.

Les modèles de systèmes fournissent la base sur laquelle les données sont identifiées et collectées, mais les données elles-mêmes sont approximatives et sélectives.

Des lacunes dans les connaissances peuvent survenir - soyez prêt à revoir régulièrement votre configuration.

La modélisation conceptuelle implicite peut conduire à des situations problématiques.

Si les données, le système et le monde ne sont pas alignés, les résultats de l'analyse des données peuvent s'avérer inutiles.

CONSIDÉRATIONS ÉTHIQUES ET MEILLEURES PRATIQUES

PRINCIPES FONDAMENTAUX DE L'ANALYSE DES DONNÉES

"Nous avons volé dans les airs comme des oiseaux et nagé dans la mer comme des poissons, mais nous devons encore apprendre le simple geste de marcher sur la Terre comme des frères."

Martin Luther King, Jr.

Quels dommages peuvent être causés par les données ?

LE BESOIN D'ÉTHIQUE

Anciennement : Une mentalité de « **Far West** » pour 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 élaborés (ils décrivent les manières responsables de pratiquer la science des données).

Il s'agit d'une responsabilité **supplémentaire** pour les spécialistes des données, mais aussi d'une **protection** contre les analyses douteuses.

Votre organisation dispose-t-elle d'un code d’éthique pour ses spécialistes des données ? Pour ses employés ?

QU'EST-CE QUE L'ÉTHIQUE ?

De manière générale, l'éthique fait référence à l'étude et à la définition des comportements corrects et incorrects :

- « pas [...] les conventions sociales, les croyances religieuses ou les lois. » (R.W. Paul, L. Elder)

Théories éthiques influentes :

- La **règle d'or** de Kant (faites aux autres...), le **conséquentialisme** (la fin justifie les moyens), **l'utilitarisme** (agir de manière à maximiser l'effet positif), etc.
- **Confucianisme, taoïsme, bouddhisme (?), etc.**
- **Ubuntu**, **Maori**, etc.

[PCAP® est une marque déposée du *Centre de gouvernance de l'information des Premières Nations* (CGIPN)]

QU'EST-CE QUE L'ÉTHIQUE ?

Les principes de PCAP® des Premières Nations

- **Propriété**
  les connaissances, les données et les informations culturelles sont la propriété des communautés
- **Contrôle**
  les communautés ont le droit de contrôler tous les aspects de la recherche et de la gestion de l'information qui les concernent
- **Accès**
  les communautés doivent avoir accès aux informations et aux données les concernant, quel que soit le lieu où elles sont conservées
- **Possession**
  les communautés doivent avoir le contrôle physique des données pertinentes

[University of Virginia, Centre for Big Data Ethics, Law, and Policy]

L'ÉTHIQUE DANS LE CONTEXTE DES DONNÉES

Questions relatives à l'éthique des données :

- **Qui**, le cas échéant, est propriétaire des données ?
- Y a-t-il des **limites** à la façon dont les données peuvent être utilisées ?
- Certaines analyses comportent-elles des **biais de valeur** ?
- Y a-t-il des catégories qui **ne devraient pas** être utilisées dans l'analyse des données personnelles ?
- Certaines données devraient-elles être **accessibles à tous** les chercheurs ?

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

BONNES PRATIQUES

Le principe de l'innocuité : les données recueillies auprès d'un individu ne doivent pas être utilisées pour lui nuire.

Consentement éclairé :

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

Respecter la « vie privée » : excessivement difficile à maintenir à l'ère du chalutage constant de l'internet pour les données personnelles.

BONNES PRATIQUES

Garder les données publiques : les données doivent être gardées publiques (toutes ? la majorité ?).

Choisir de participer ou de se retirer : Le consentement éclairé exige la possibilité de se retirer.

Anonymiser les données : suppression des champs d'identification des données avant l'analyse.

« Laisser parler les données » :

- pas de « picorage » (cherry picking)
- importance de la validation (nous y reviendrons plus tard)
- corrélation et causalité (nous y reviendrons plus tard)
- répétabilité

[https://www.statcan.gc.ca/fra/themes-debut/genre_diversite_et_inclusion]

ACS+

L'**analyse comparative entre les sexes +** est une manière d'évaluer l'impact de politiques, de programmes, d'initiatives sur les personnes de diverses identités de genre.

Exemple: Les arrêts de travail et la vulnérabilité financière, D. Messacar, R. Morrissette

- Si les données n'avaient pas été collectées et/ou analysées dans le cadre de l'ACS+, il serait plus difficile de voir comment la vulnérabilité financière affecte les différents groupes (si l'analyse avait porté uniquement sur les groupes d'âge et le sexe, par exemple, au lieu d'inclure également la composition de la famille).

Les politiques et les événements ont **un impact réel sur des personnes réelles**, et pas toujours de la même manière. Les méthodes d'analyse des données sont généralement utilisées pour prédire et/ou décrire les résultats **moyens** (ou centraux), mais ce sont souvent ceux qui sont loin du centre qui sont les plus touchés.