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

[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

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
…	…	…	…	…	…

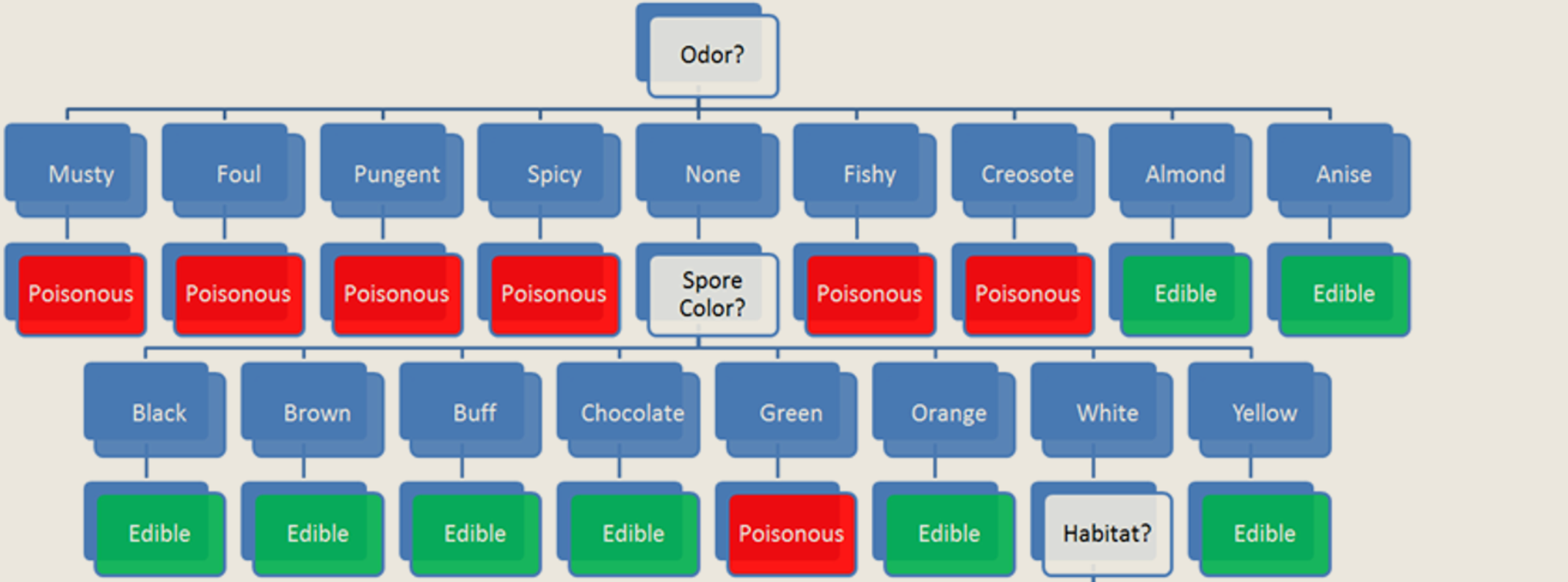
ENSEMBLE DE DONNÉES SUR LES CHAMPIGNONS VÉNÉNEUX

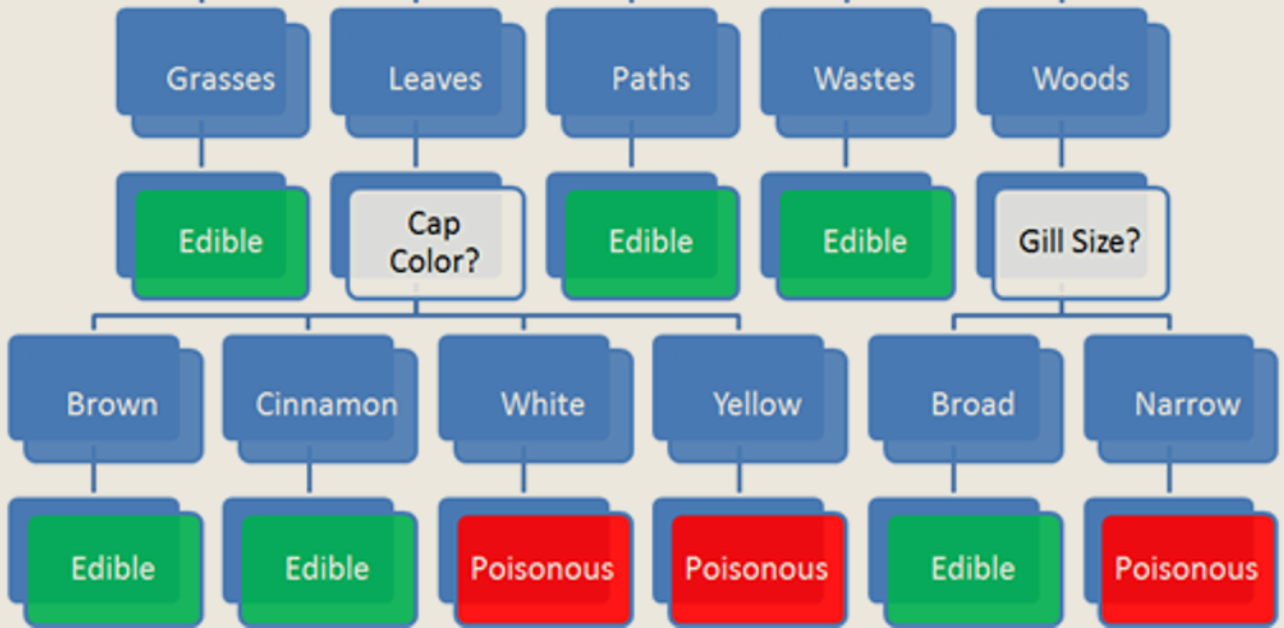
Amanita muscaria

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

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

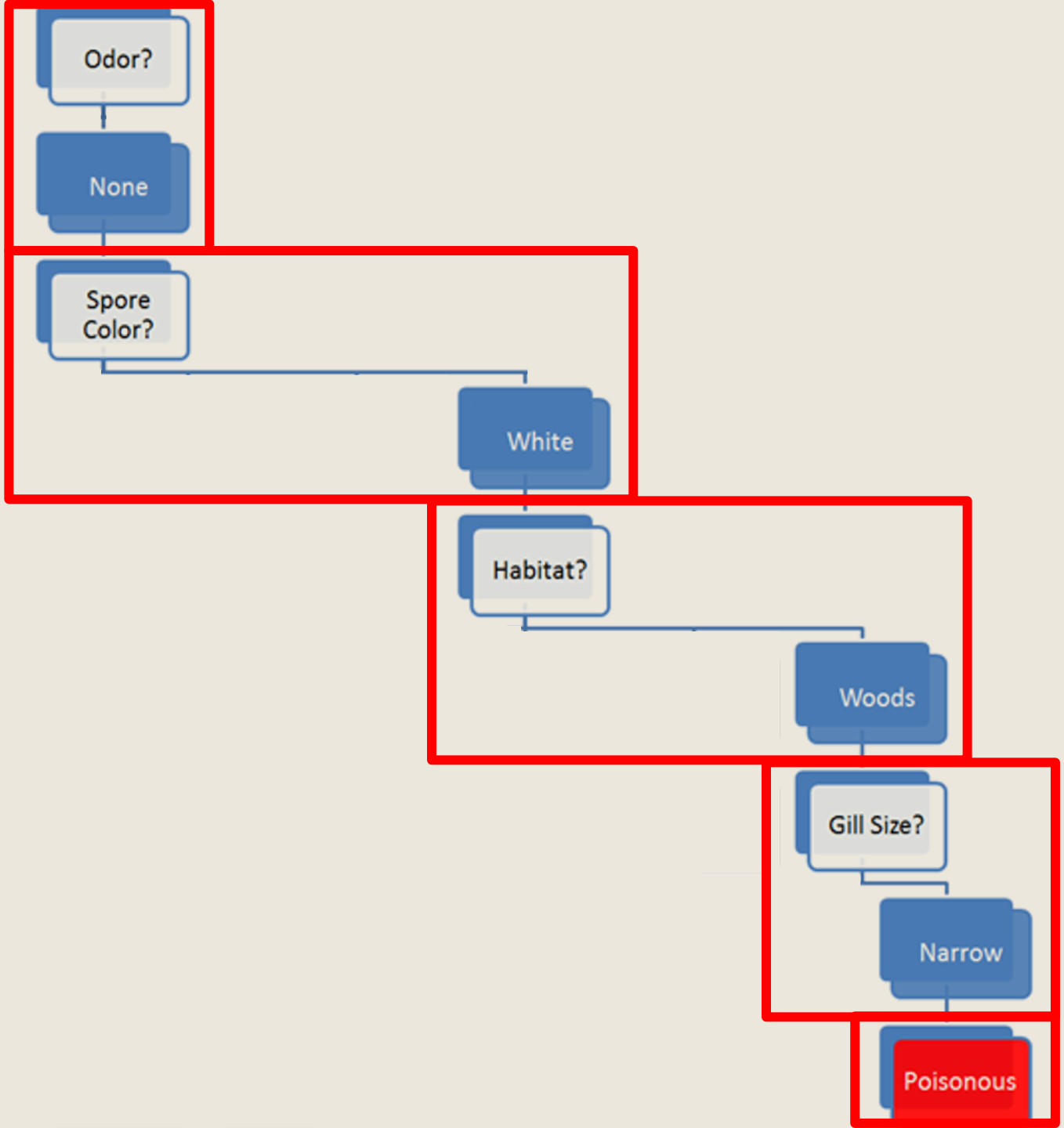
[https://archive.ics.uci.edu/ml/datasets/Mushroom]





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**?

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.

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.

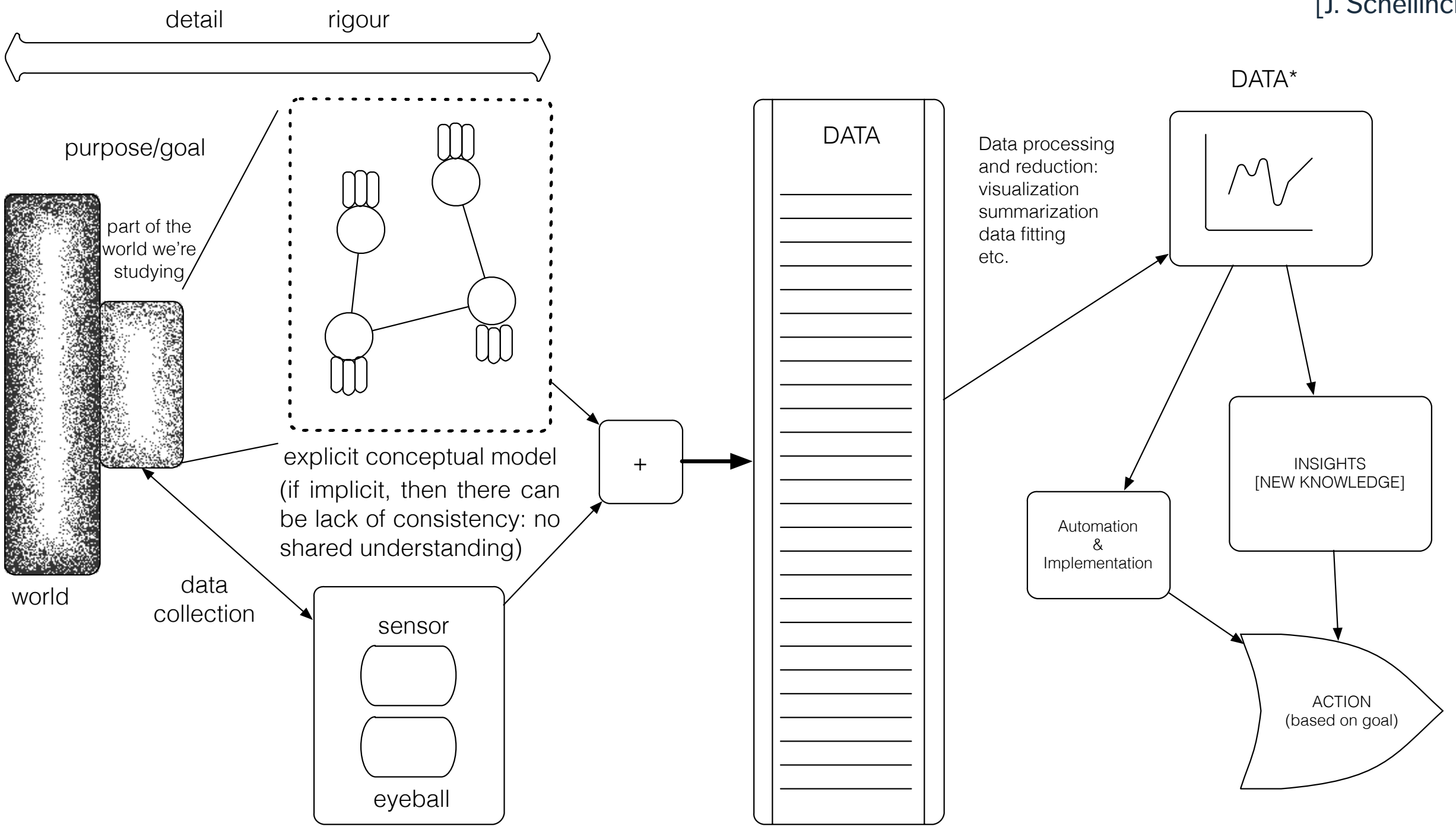
T. Kwartler (paraphrasé)

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)

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

[J. Schellinck]



[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.

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.

[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