NOTIONS UNIVERSELLES SUR L'ANALYSE DES DONNÉES

« 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

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

DONNÉES, APPRENTISSAGE AUTOMATIQUE ET INTELLIGENCE ARTIFICIELLE DANS L'ACTUALITÉ

NOTIONS UNIVERSELLES SUR L'ANALYSE DE DONNÉES

OBJECTIFS D'APPRENTISSAGE DU MODULE

Accroître la prise de conscience du rôle croissant de la science des données, de l'apprentissage automatique et de l'intelligence artificielle dans différents domaines de la société.

Accroître la sensibilisation au sujet des fonctionnalités/capacités possibles de ces technologies.

Accroître la sensibilisation concernant certains des enjeux sociaux découlant du rôle croissant de ces technologies.

News

Robots are better than doctors at diagnosing some cancers, major study finds

share

Save 7

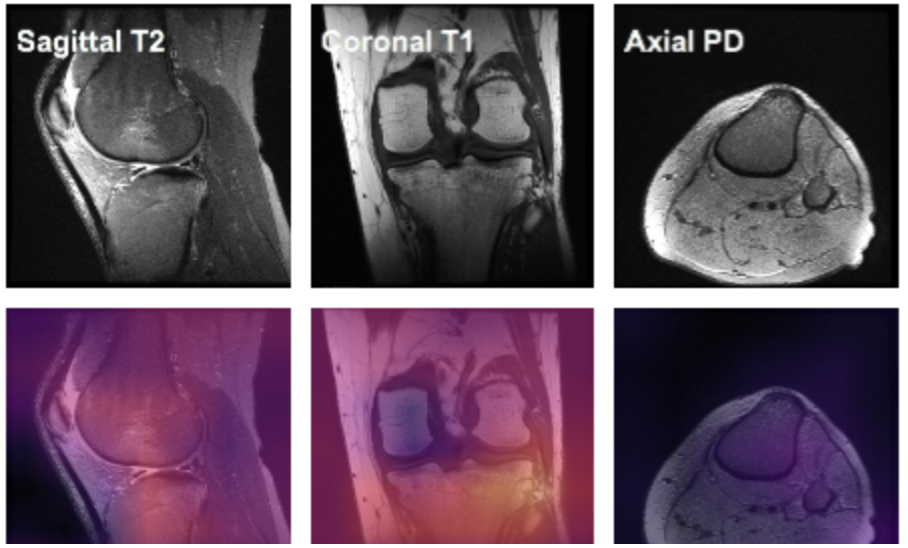
Stanford ML Group

MRNet: Deep-learning-assisted diagnosis for knee magnetic resonance imaging

Nicholas Bien *, Pranav Rajpurkar *, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng†, Matthew P. Lungren†

We developed an algorithm to predict abnormalities in knee MRI exams, and measured the clinical utility of providing the algorithm's predictions to radiologists and surgeons during interpretation.

Magnetic resonance (MR) imaging of the knee is the standard of care imaging modality to evaluate knee



Google AI Claims 99 Percent Accuracy In Metastatic Breast Cancer Detection

34

Posted by **BeauHD** on Friday October 12, 2018 @08:00PM from the promising-solutions dept.

Researchers at the Naval Medical Center San Diego and Google AI, a division within Google dedicated to artificial intelligence research, are using cancer-detecting algorithms to detect metastatic tumors by autonomously evaluating lymph node biopsies. VentureBeat reports:

> Their AI system -- dubbed Lymph Node Assistant, or LYNA -- is described in a paper titled "Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection," published in The American Journal of Surgical Pathology. In tests, it achieved an area under the receiver operating characteristic (AUC) -- a measure of detection accuracy -- of 99 percent. That's superior to human pathologists, who according to one recent assessment miss small metastases on individual slides as much as 62 percent of the time when under time constraints. LYNA is based on Inception-v3, an open source image recognition deep learning model that's been shown to achieve greater than 78.1 percent accuracy on Stanford's ImageNet dataset. As the researchers explained, it takes as input a 299-pixel image (Inception-v3's default input size), outlines tumors at the pixel level, and, in the course of training, extracts labels -- i.e., predictions -- of the tissue patch ("benign" or "tumor") and adjusts the model's algorithmic weights to reduce error.
>
> In tests, LYNA achieved 99.3 percent slide-level accuracy. When the model's sensitivity threshold was adjusted to detect all tumors on every slide, it exhibited 69 percent sensitivity, accurately identifying all 40 metastases in the evaluation dataset without any false positives. Moreover, it was unaffected by artifacts in the slides such as air bubbles, poor processing, hemorrhage, and overstaining. LYNA wasn't perfect -- it occasionally misidentified giant cells, germinal cancers, and bone marrow-derived white blood cells known as histiocytes -- but managed to perform better than a practicing pathologist tasked with evaluating the same slides. And in a second paper published by Google AI and Verily, Google parent company Alphabet's life sciences subsidiary, the model halved the amount of time it took for a six-person team of board-certified pathologists to detect metastases in lymph nodes.

Data scientists find connections between birth month and health

Date: June 8, 2015

Source: Columbia University Medical Center

Summary: Scientists have developed a computational method to investigate the relationship between birth month and disease risk. The researchers used this algorithm to examine New York City medical databases and found 55 diseases that correlated with the season of birth. Overall, the study indicated people born in May had the lowest disease risk, and those born in October the highest.

Share:

9
Oct
2018

Scientists Using GPS Tracking on Endangered Dhole Wild Dogs

Researchers Successfully Tag a Dhole. Wildlife scientists around the globe are ecstatic to hear that researchers were able to successfully place a GPS tracking device onto a dhole. This marks the first time in history that conservationists have been able to place a collar on one of these very rare Indian wild dogs. It's estimated that less than 2,500 of these creatures still exist globally.

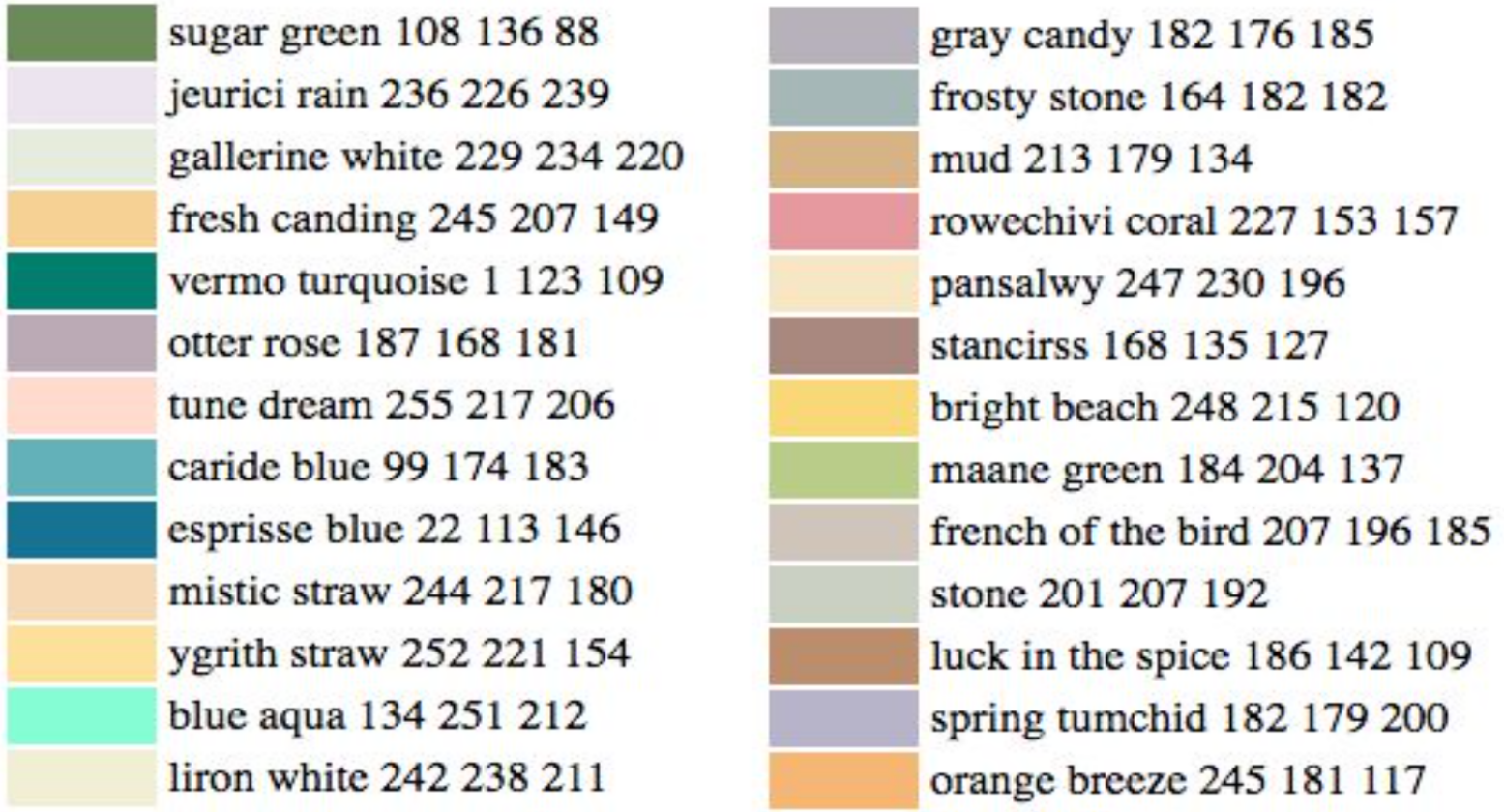
DESIGN

These AI-invented paint color names are so bad they're good

1

What's in a (paint) name?

By **Sam Reichman** | May 31, 2017, 5:12pm EDT

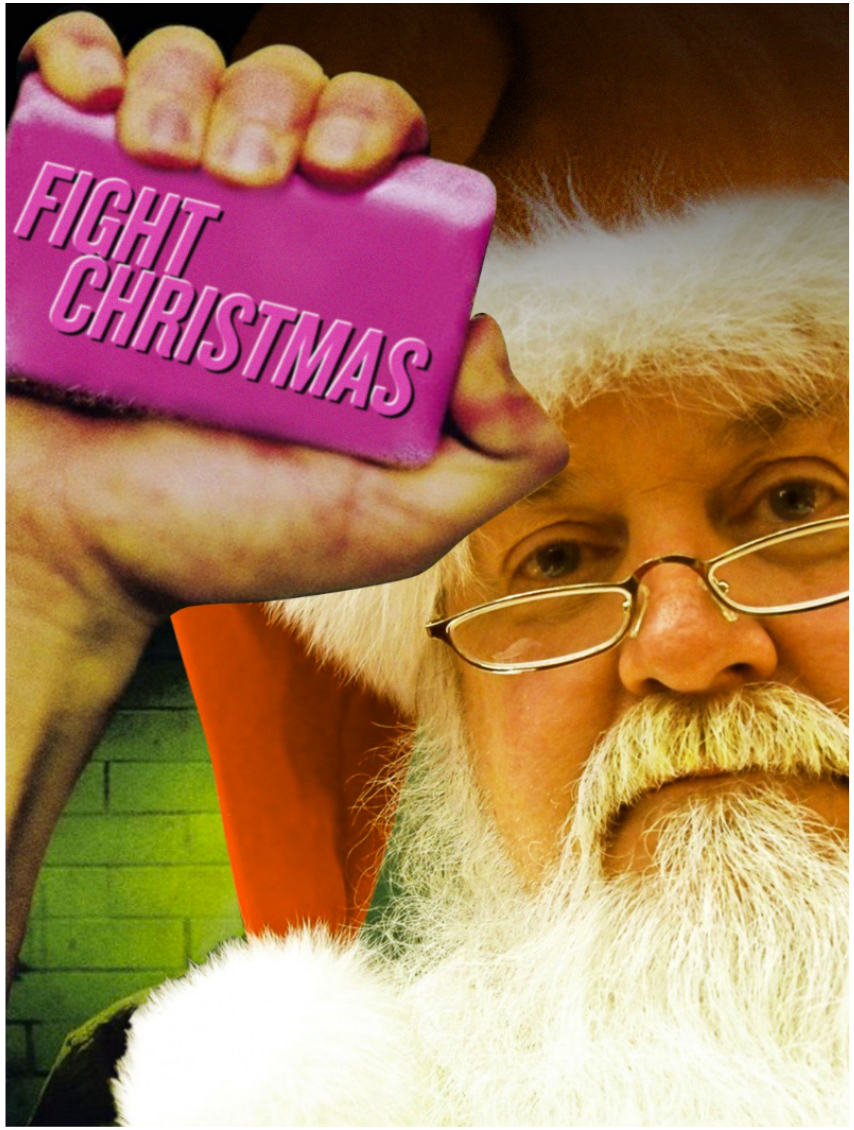


Intelligent Machines

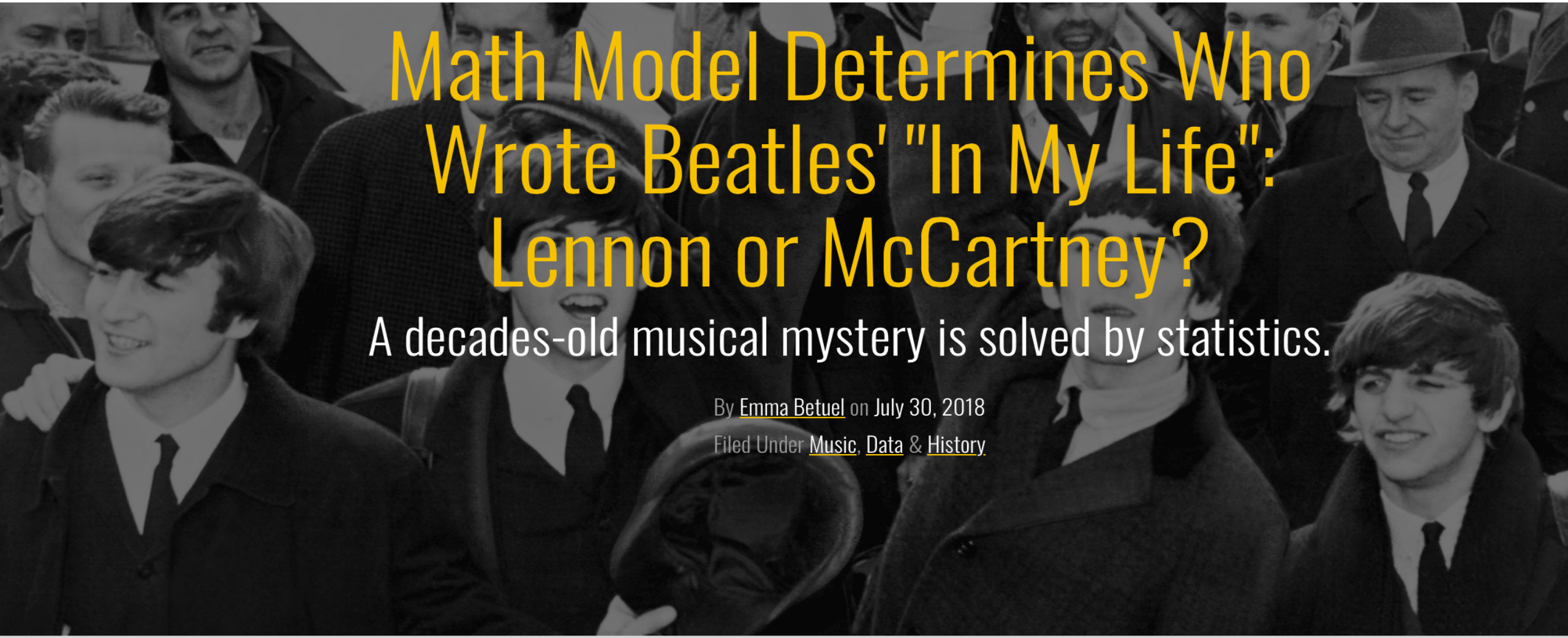
We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually.

Using a neural network to create ridiculous plot lines takes a lot of work—and reveals the challenges of generating human language.

by Karen Hao December 21, 2018



MR. TECH



Math Model Determines Who Wrote Beatles' "In My Life": Lennon or McCartney?

A decades-old musical mystery is solved by statistics.

By Emma Betuel on July 30, 2018

Filed Under Music, Data & History

Scientists use Instagram data to forecast top models at New York Fashion Week

Method is 80 percent accurate in identifying most popular models for the following season

Date: September 3, 2015

Source: Indiana University

Summary: Researchers have predicted the popularity of new faces to the world of fashion modeling with over 80 percent accuracy using advanced computational methods and data from Instagram.

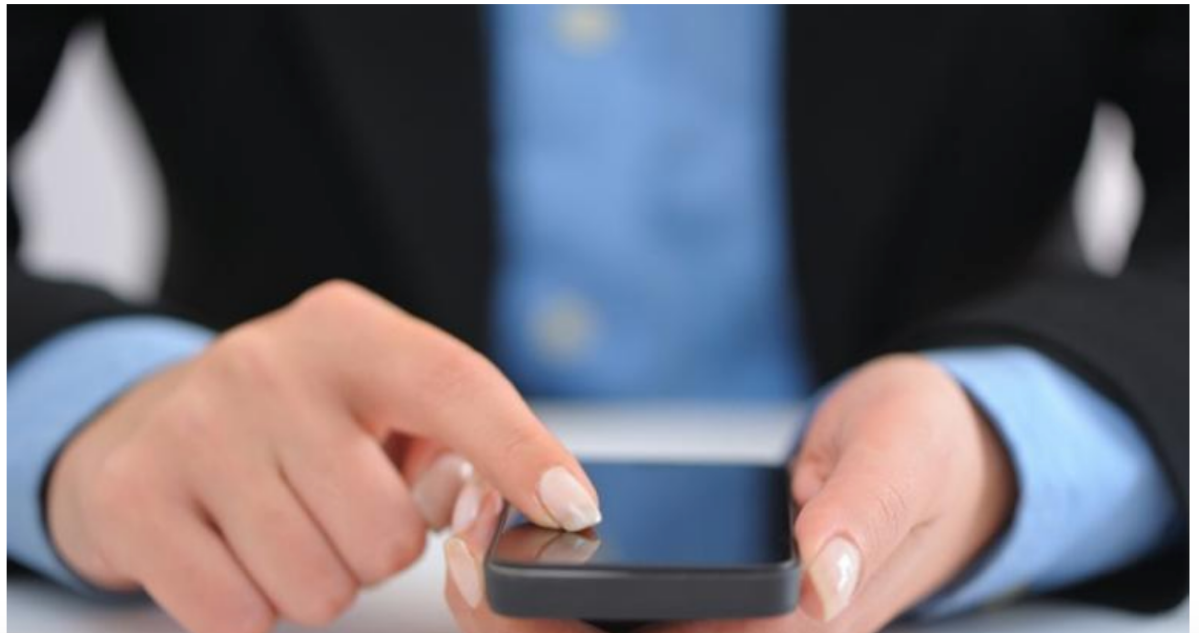
Share:

How big data will solve your email problem

That deluge in your inbox needs to be handled. A team of Israeli researchers thinks big data has some answers that can help.

By Jason Hiner | October 2, 2013 -- 16:05 GMT (09:05 PDT) | Topic: Going Deep on Big Data

6



NEWSLETTERS

ZDNet Big Data

Keep up with the latest developments in extracting maximum information value for today's business.

Your email address SUBSCRIBE

SEE ALL

MORE RESOURCES

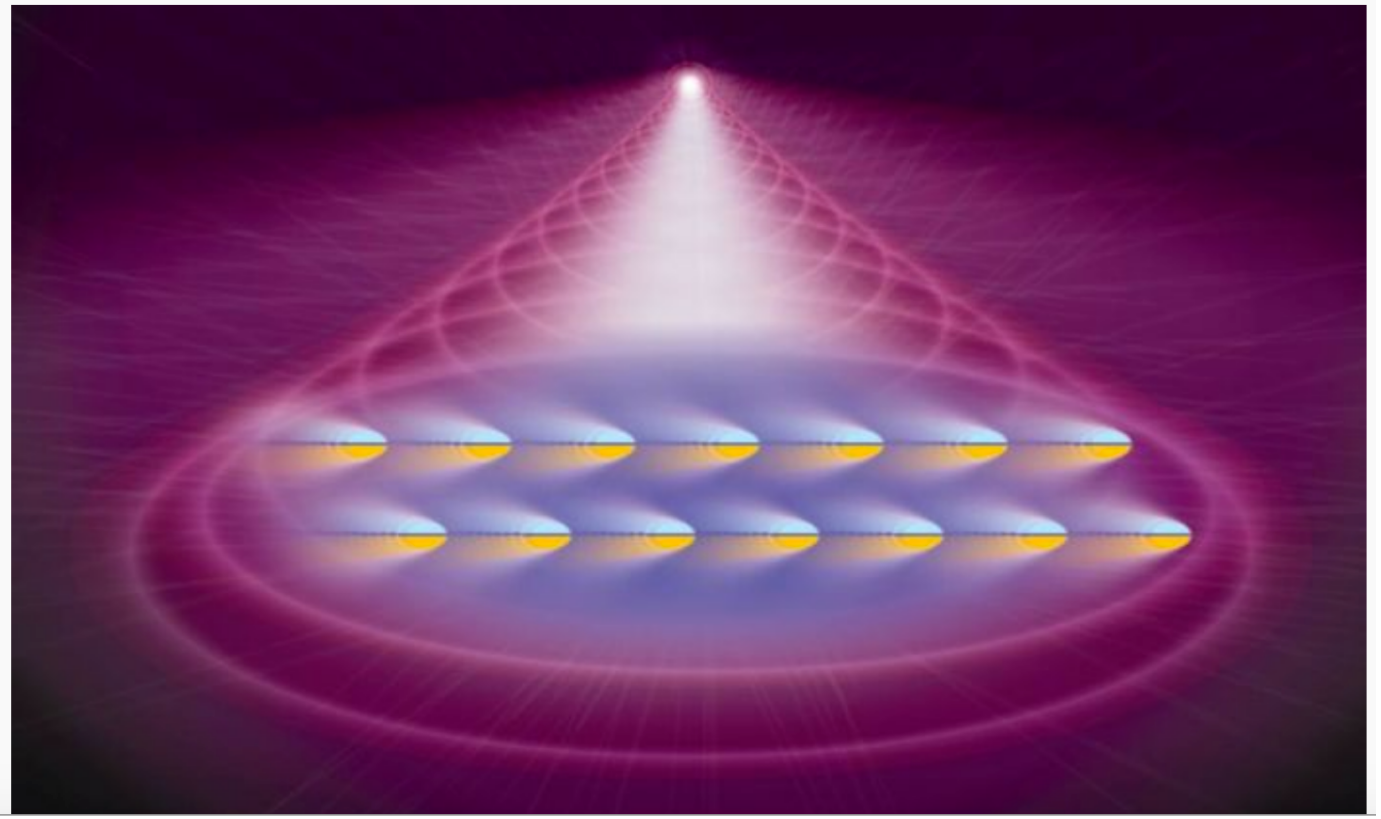
Special report: From cloud

Artificial intelligence better than physicists at designing quantum science experiments

Share on Facebook | Share on Twitter

ABC Science By science reporter Belinda Smith

Posted 19 October 2018 at 3:36 pm



Wonkblog • Analysis

This researcher studied 400,000 knitters and discovered what turns a hobby into a business

Most Read Business

1 **Perspective**
I ordered a box of crickets from the Internet and it went about as well as you'd expect

2 As a grocery chain is dismantled, investors recover their money. Worker pensions are short millions.

3 Markets poised to finish year with worst performance in a decade — and the volatility seems certain to continue

SCIENCE

Wait, Have We Really Wiped Out 60 Percent of Animals?

The findings of a major new report have been widely mischaracterized —although the actual news is still grim.

ED YONG OCT 31, 2018

MORE STORIES

Animals Are Riding an Escalator to Extinction
ED YONG

It Will Take Millions of Years for Mammals to Recover From Us
ED YONG

In a Few Centuries, Cows Could Be the Largest Land Animals Left
ED YONG

An Ancient Tradition

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 2 DAYS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin
8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Facebook documents seized by MPs investigating privacy breach

25 November 2018 | Business

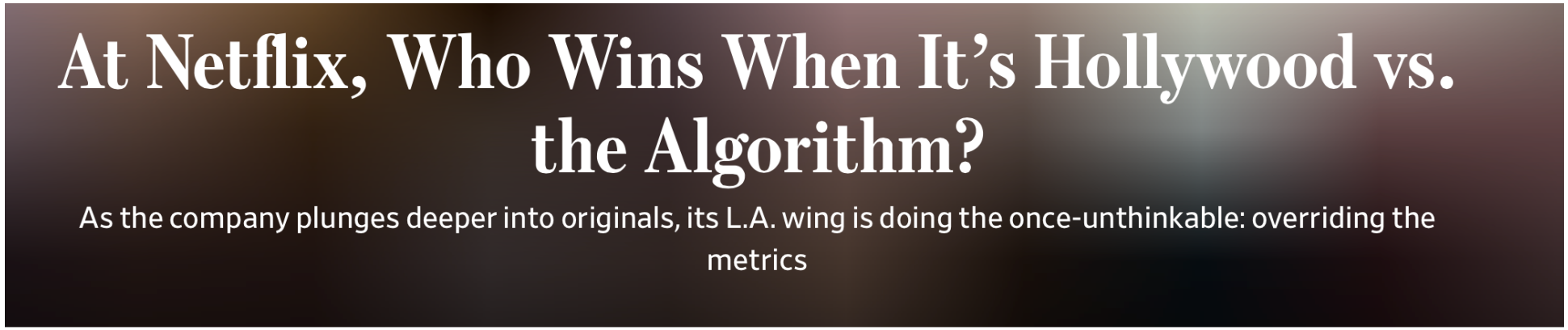
A cache of Facebook documents has been seized by MPs investigating the Cambridge Analytica data scandal.

The New York Times

Firm Led by Google Veterans Uses A.I. to 'Nudge' Workers Toward Happiness

At Netflix, Who Wins When It's Hollywood vs. the Algorithm?

As the company plunges deeper into originals, its L.A. wing is doing the once-unthinkable: overriding the metrics



The cast of Netflix original show 'GLOW.' NETFLIX

By *Shalini Ramachandran* and *Joe Flint*

100 COMMENTS

Nov. 10, 2018 12:00 a.m. ET

Netflix Inc.'s executives were torn. On the one hand they trusted the company's algorithm. On the other they were worried about ticking off Jane Fonda.

After the streaming-video giant released the second season of the comedy "Grace and Frankie" in 2016, its product team put up an image to promote the show to U.S. subscribers that only included Ms. Fonda's co-star, Lily Tomlin. Tests showed that more users clicked on the show when the photo didn't include Ms. Fonda.

DONNÉES 101 – NOTIONS DE DONNÉES DE BASE

NOTIONS UNIVERSELLES SUR L'ANALYSE DE DONNÉES

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

Attribué à Daniel Keys Moran

OBJECTIFS D'APPRENTISSAGE DU MODULE

Connaissance préliminaire des notions suivantes :

- Données, attribut (propriété, facteur, variable)
- Modèles prédictifs, modèles explicatifs
- Classification, estimation des probabilités de classe, regroupement, règles d'association, analyse des séries chronologiques, détection des anomalies, arbre décisionnel, apprentissage supervisé, apprentissage non supervisé

Comparer et mettre en contraste : la science des données par rapport à l'analyse (veille stratégique).

Connaissance des niveaux appropriés de confiance dans les modèles.

QU'EST-CE QU'UNE DONNÉE? D'OÙ PROVIENT-ELLE?

4 529 « rouge » 25 782 « Y »

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
…	…	…	…	…	…

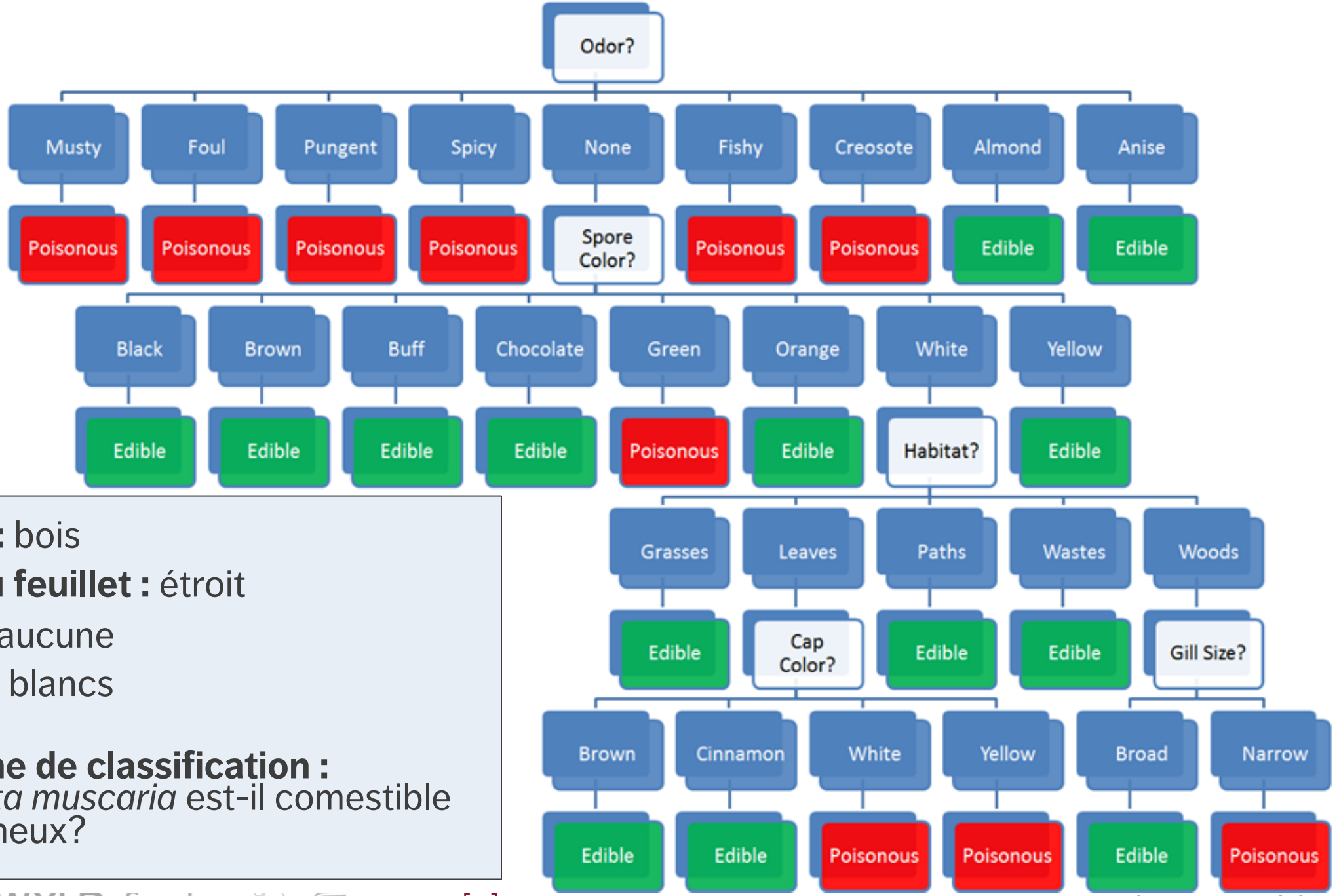
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

ENSEMBLE DE DONNÉES SUR LES CHAMPIGNONS VÉNÉNEUX

Amanita muscaria

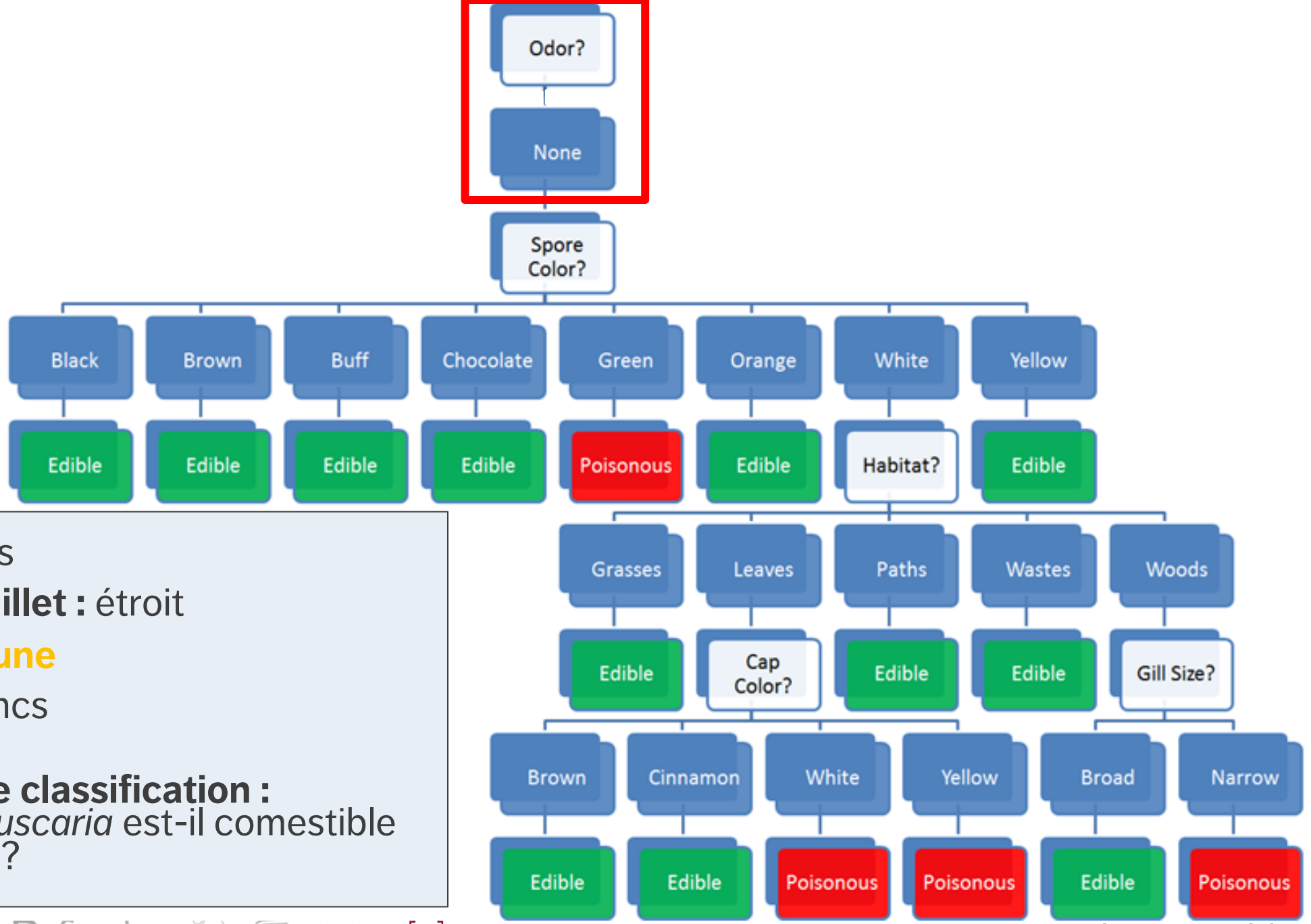
Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : blancs

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?



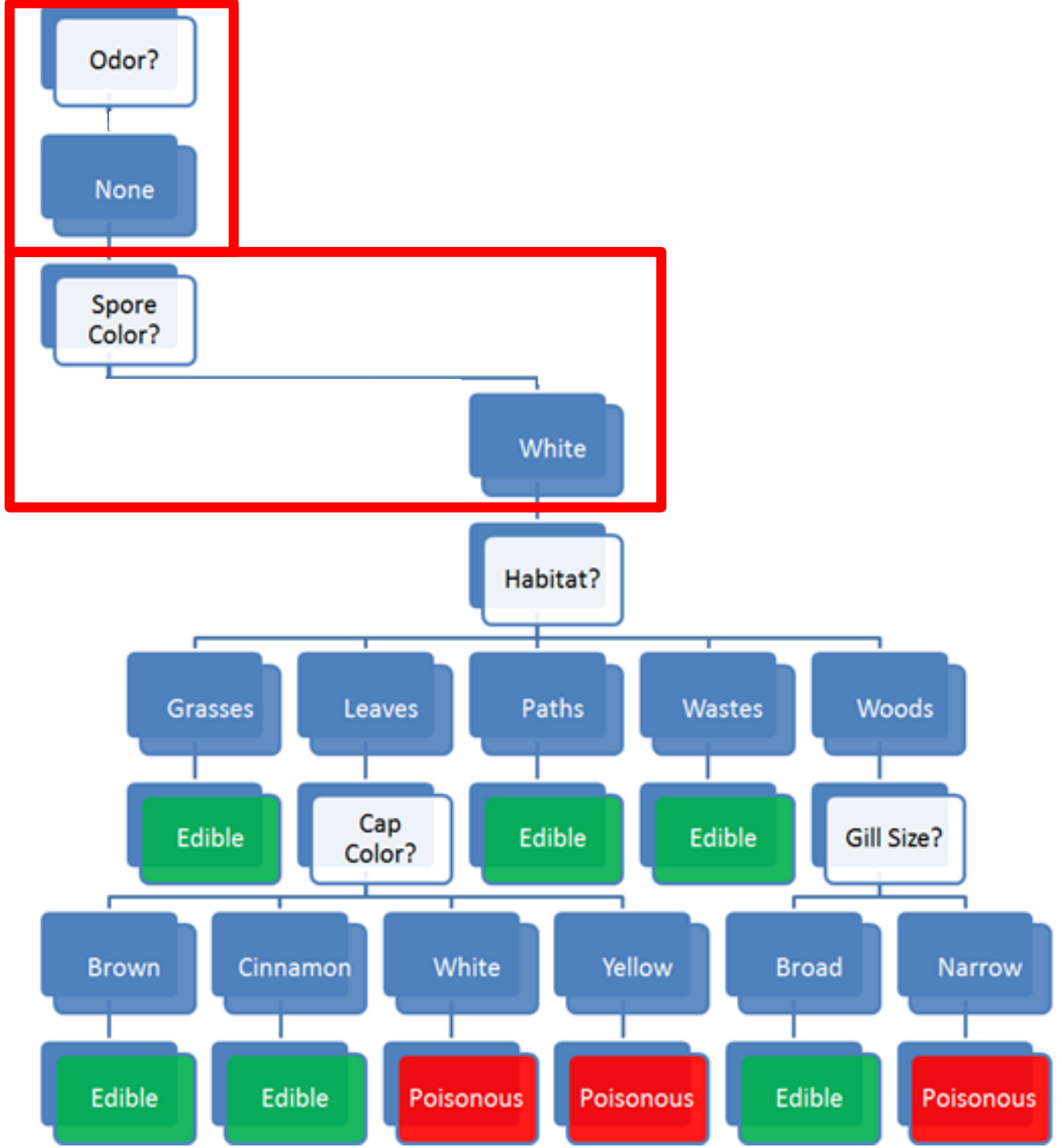
Habitat : bois

Taille du feuillet : étroit

Odeur : aucune

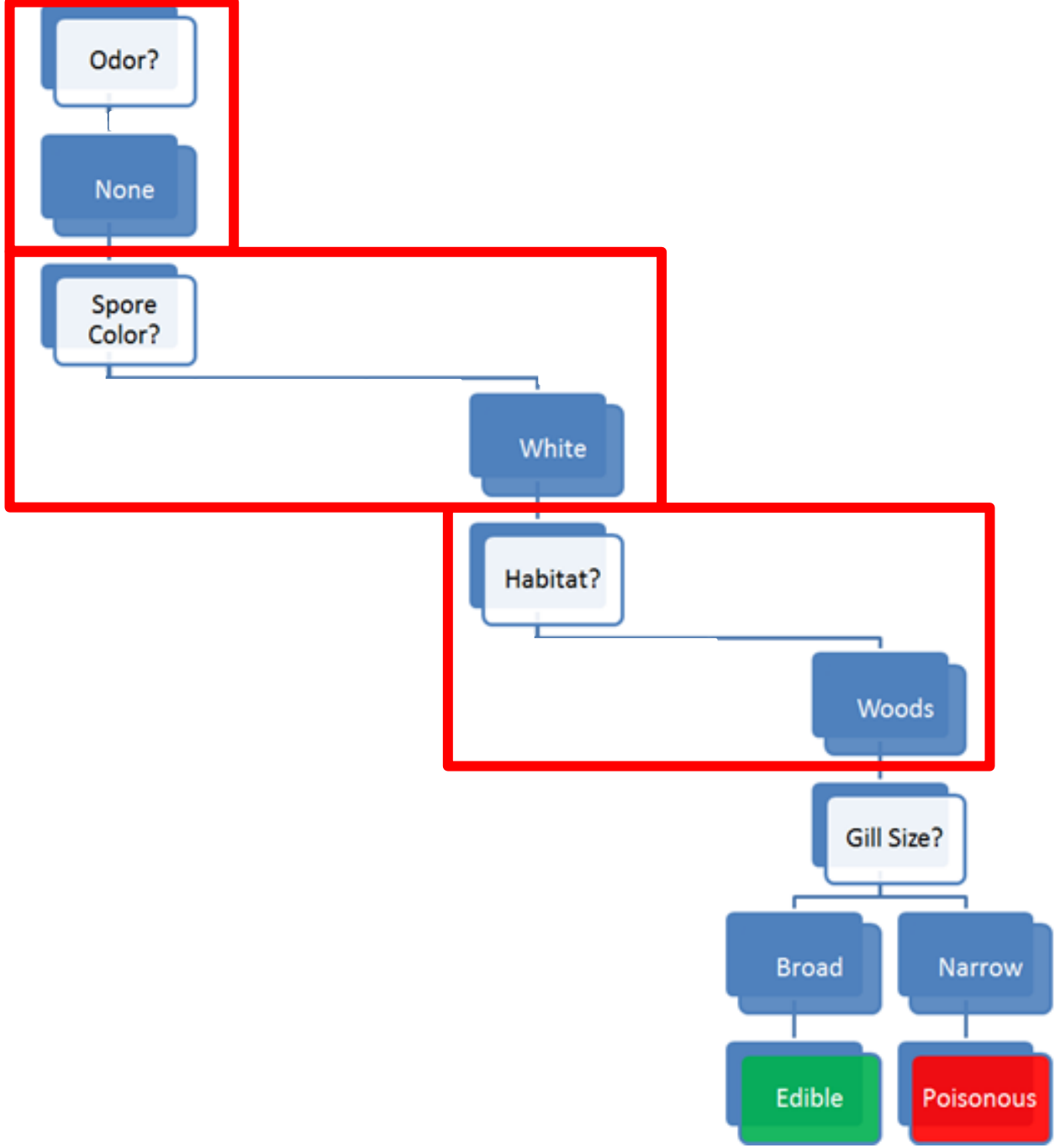
Spores : blancs

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



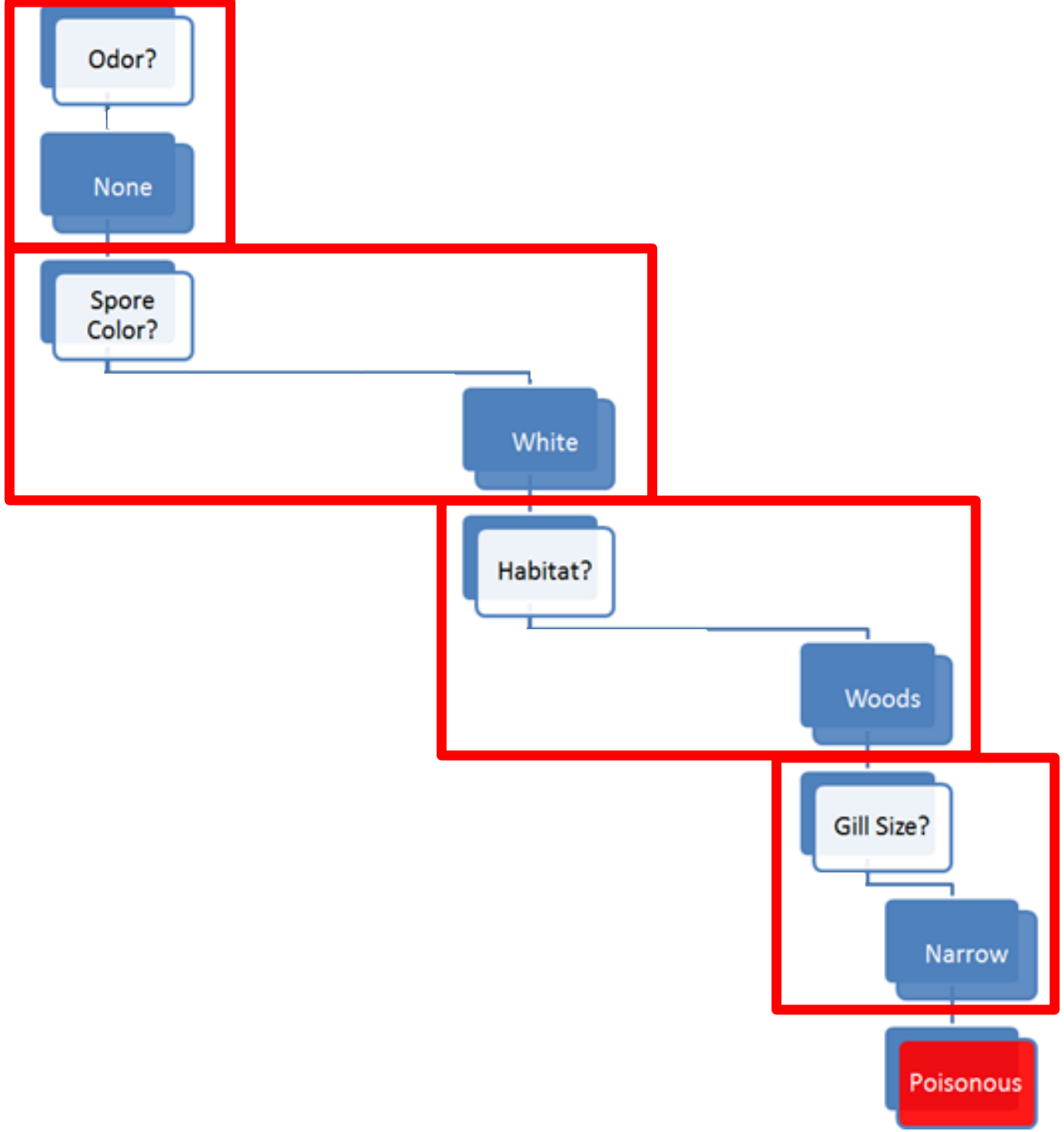
Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : **blancs**

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



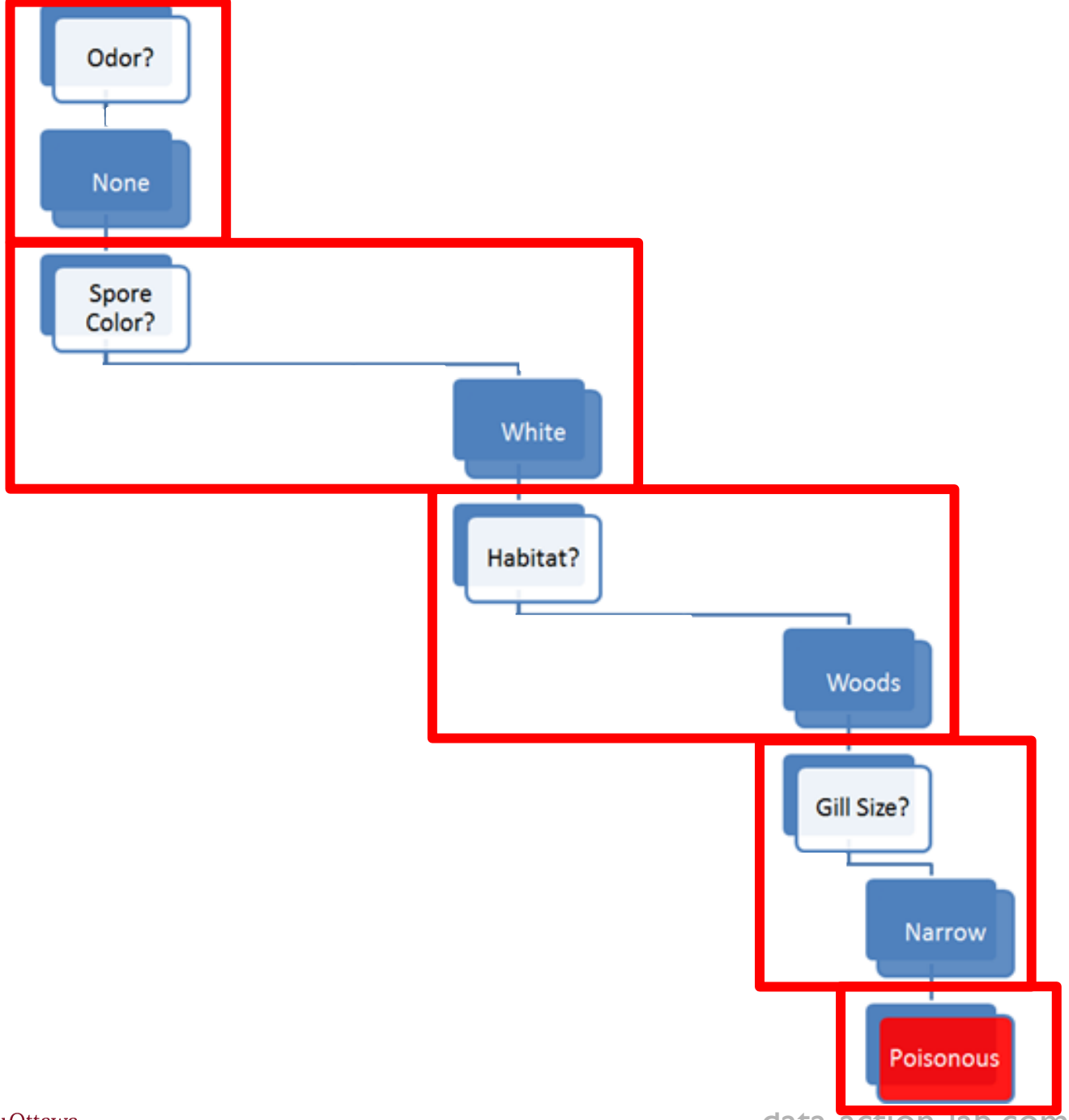
Habitat : bois
Taille du feuillet : étroit
Odeur : aucune
Spores : blancs

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?



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

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 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; **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.

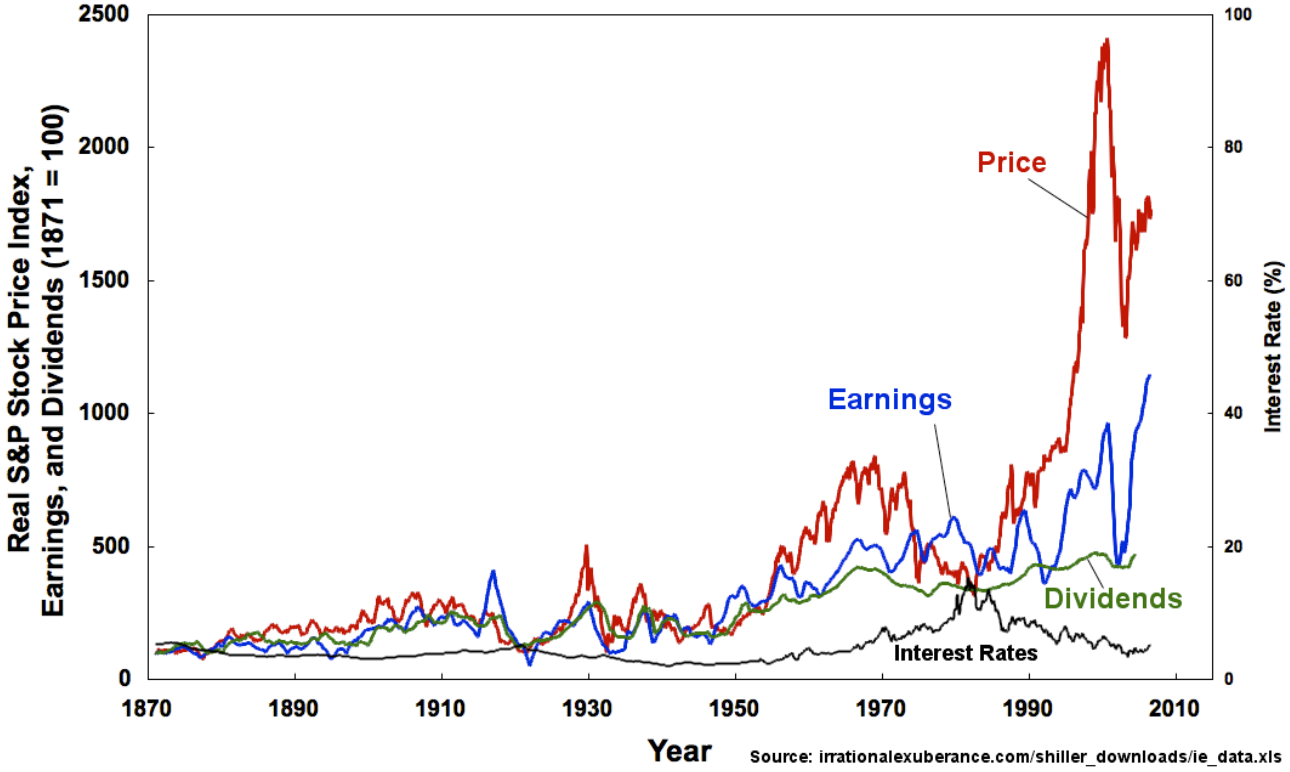
ANALYSE DES SÉRIES CHRONOLOGIQUES

Une série chronologique simple :

- Possède deux variables : le temps + une 2^{e} variable
- La deuxième variable est *séquentielle*

Quel est le comportement de cette deuxième variable au fil du temps? Par rapport à d'autres variables?

Pouvons-nous utiliser cette information pour prévoir le comportement de la variable à l'avenir?



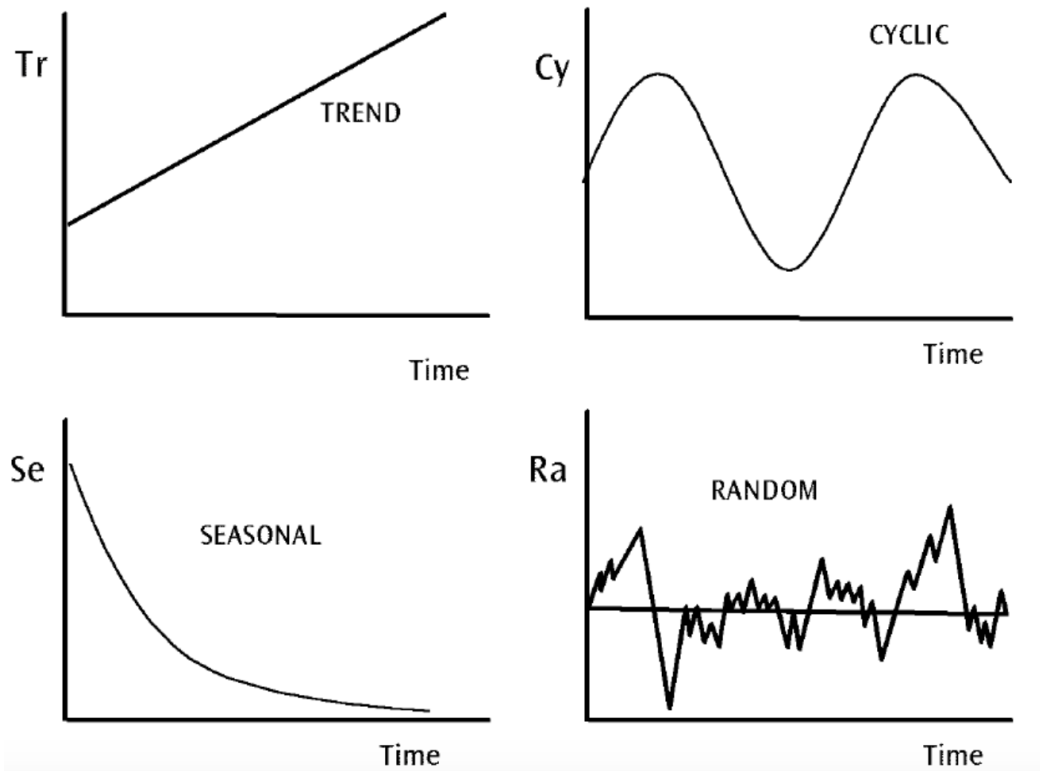
SCHÉMAS TEMPORELS

Il s'agit ici de nos objectifs d'analyse habituels :

- Trouver des tendances dans les données
- Créer un modèle (mathématique) qui saisit l'essence de ces tendances

Les tendances peuvent être assez complexes – une analyse poussée est généralement nécessaire!

En particulier, l'ensemble de la série peut souvent être décomposé en plusieurs **modèles de composantes**. Il existe des bibliothèques de logiciels qui peuvent vous aider!



ÉTUDES DE CAS DES SÉRIES CHRONOLOGIQUES

Communication Research
2018, Vol. 45(7) 1012–1030

DOI: 10.1177/0093650215581370
journals.sagepub.com/home/crx
SAGE

A Time-Series Analysis of International Public Relations Expenditure and Economic Outcome

Suman Lee[1] and Byungwook Kim[2]

Abstract

This study tested a causal relationship between international public relations (PR) expenditure and its economic outcome at the country level by using a time-series analysis. International PR expenditures of four client countries (Japan, Colombia, Belgium, and the Philippines) were collected from the semi-annual reports of the Foreign Agency Registration Act (FARA) from 1996 to 2009. Economic outcome was measured by U.S. imports from the client countries and U.S. foreign direct investment (FDI) toward them. This study found that the past PR expenditure holds power in forecasting future economic outcomes for Japan, Belgium, and the Philippines except Colombia.

Keywords

international public relations, PR return on investment, bottom-line effect, time-series analysis, Granger causality test

RESEARCH ARTICLE

Seiya MAKI, Shuichi ASHINA, Minoru FUJII, Tsuyoshi FUJITA, Norio YABE, Kenji UCHIDA, Gito GINTING, Rizaldi BOER, Remi CHANDRAN

Employing electricity-consumption monitoring systems and integrative time-series analysis models: A case study in Bogor, Indonesia

Abstract The Paris Agreement calls for maintaining a global temperature less than 2°C above the pre-industrial level and pursuing efforts to limit the temperature increase even further to 1.5°C. To realize this objective and promote a low-carbon society, and because energy production and use is the largest source of global greenhouse-gas (GHG) emissions, it is important to efficiently manage energy demand and supply systems. This, in turn, requires theoretical and practical research and innovation in smart energy monitoring technologies, the identification of appropriate methods for detailed time-series analysis, and the application of these technologies at urban and national scales. Further, because developing countries contribute increasing shares of domestic energy consumption, it is important to consider the application of such innovations in these areas. Motivated by the mandates set out in global agreements on climate change and low-carbon societies, this paper focuses on the development of a smart energy monitoring system (SEMS) and its deployment in households and public and commercial sectors in Bogor, Indonesia. An electricity demand prediction model is developed for each device using the Auto-Regression eXogenous model. The real-time SEMS data and time-series clustering to explore similarities in electricity consumption patterns between monitored units, such as residential, public, and commercial buildings, in Bogor is, then, used. These clusters are evaluated using peak demand and Ramadan term characteristics. The resulting energy-prediction models can be used for low-carbon planning.

Keywords electricity monitoring, electricity demand prediction, multiple-variable time-series modeling, time-series cluster analysis, Indonesia

Received Dec. 30, 2017; accepted Mar. 28, 2018; online May 30, 2018

1 Introduction

1.1 Background and objectives

To attain a low-carbon society, it is necessary to transform the centralized energy system into distributed systems at city and regional scales. Because energy demand patterns vary spatially, more detailed data on energy demand provided by innovative Information Communication Technologies (ICTs) is expected to enable local energy demand and supply system optimization in which distributed renewable energy resources can be integrated with large-scale grid energy supply systems.

Energy information and data at local scales, particularly in developing countries, is persistently unavailable. However, there is enormous potential to reduce energy use in various sectors through the use of rapidly developing ICT systems in energy management. The

MAKI ET COLL. : SYSTÈME D'ANALYSE DES DONNÉES

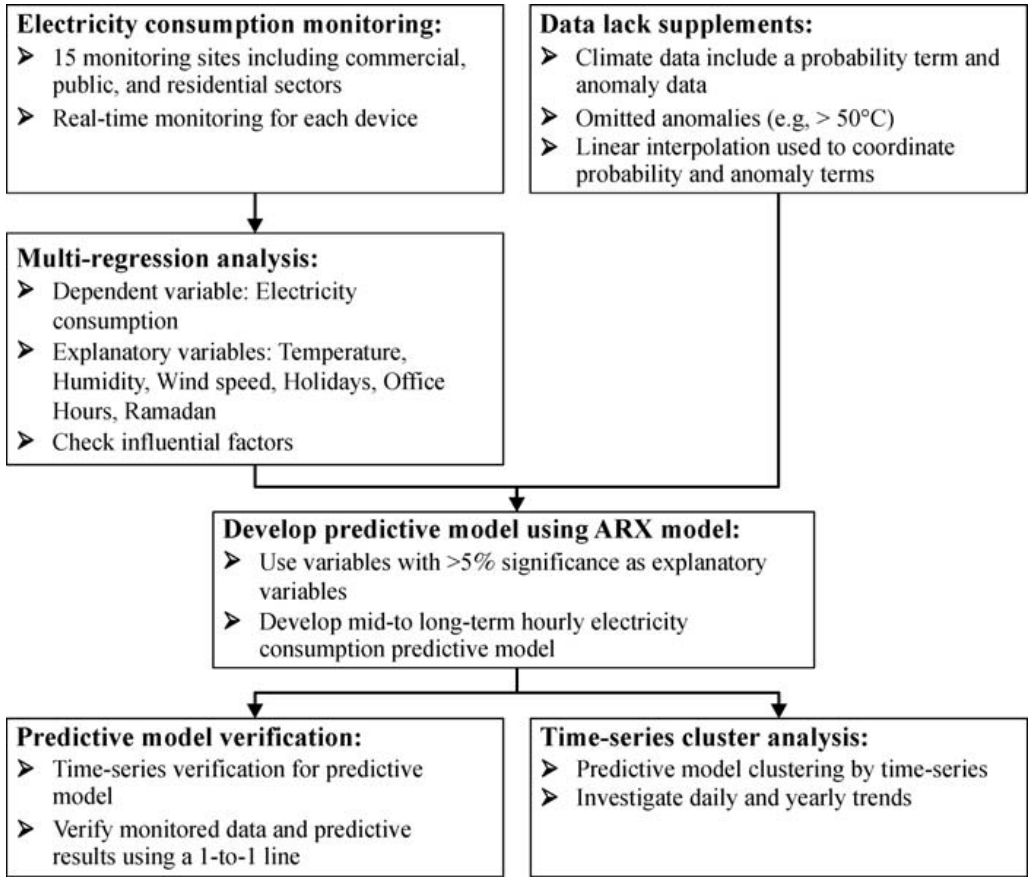


Fig. 1 Analytical procedure used in this paper

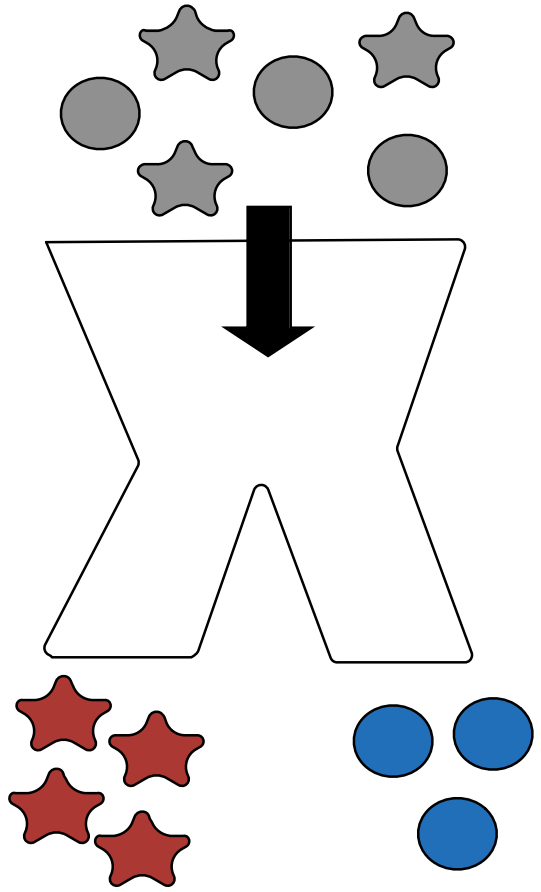
CLASSIFICATION

Classificateur : Si un objet m'est présenté, puis-je le classifier dans l'une des catégories prédéfinies?

Il existe beaucoup de techniques différentes pour réaliser cela, mais les étapes sont les mêmes :

- Utilisez une *trousse de formation* (« training set ») pour apprendre au classificateur à classifier
- Mettez à l'essai/validez le classificateur à l'aide de *nouvelles données*
- Utilisez le classificateur pour classifier les *nouvelles instances.*

Certains classificateurs (par exemple les réseaux neuronaux) sont très similaires à une « boîte noire ». Ils sont peut-être bons pour classer, mais vous ne savez pas pourquoi!



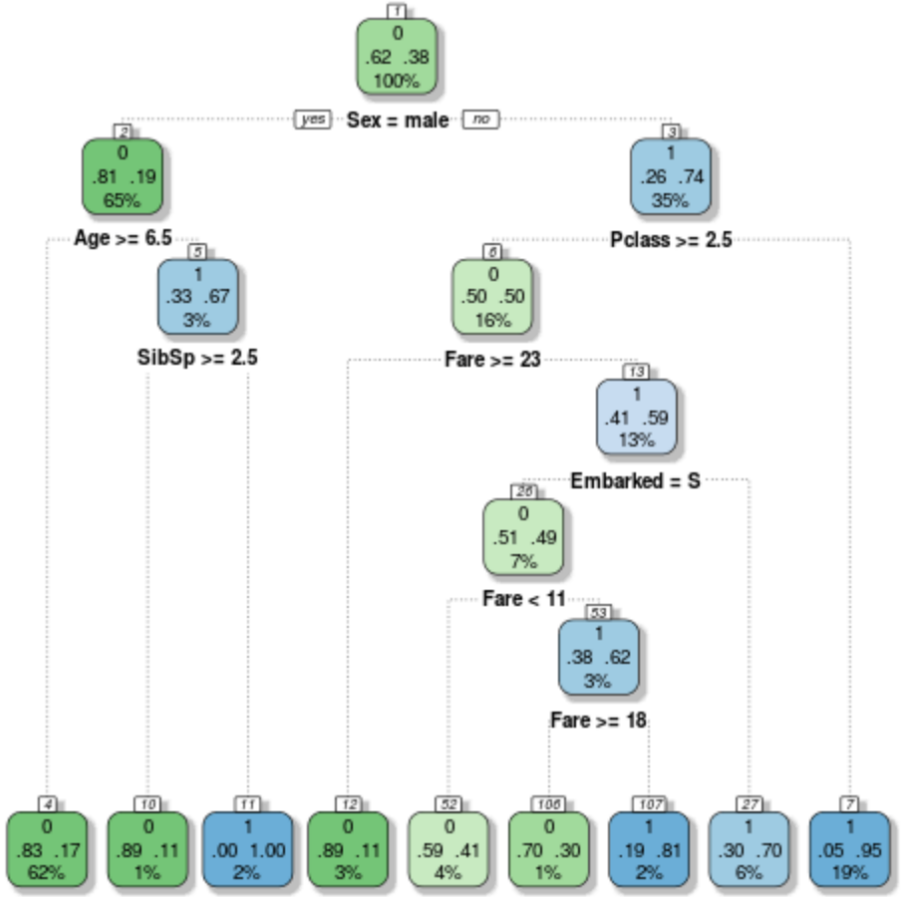
CLASSIFICATEURS D'ARBRES DE DÉCISION

Arbre de décision : Quelles sont vos propriétés? J'utiliserai (méthodiquement) cette information pour m'aider à vous classer.

Il existe des techniques que nous pouvons utiliser pour construire *automatiquement* ces arbres de décision.

Une fois l'arbre construit, nous pouvons voir comment la décision est prise.

Ils sont également utiles pour les systèmes experts.



ORIGINAL ARTICLE

Profiling Arthritis Pain with a Decision Tree

Man Hung, PhD; Jerry Bounsanga, BS; Fangzhou Liu, MS; Maren W. Voss, MS
Department of Orthopaedics, University of Utah, Salt Lake City, Utah, U.S.A.

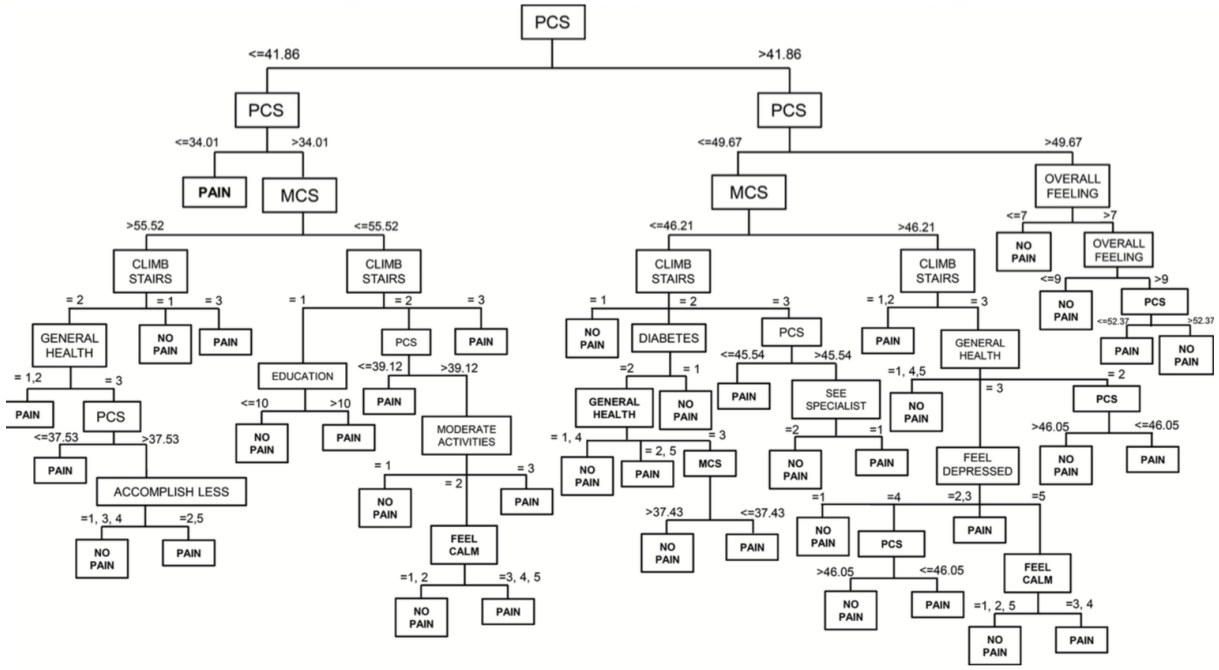
■ **Abstract**

Background: Arthritis is the leading cause of work disability and contributes to lost productivity. Previous studies showed that various factors predict pain, but they were limited in sample size and scope from a data analytics perspective.
Objectives: The current study applied machine learning algorithms to identify predictors of pain associated with arthritis in a large national sample.
Methods: Using data from the 2011 to 2012 Medical Expenditure Panel Survey, data mining was performed to develop algorithms to identify factors and patterns that contribute to risk of pain. The model incorporated over 200 variables within the algorithm development, including demographic data, medical claims, laboratory tests, patient-reported outcomes, and sociobehavioral characteristics.
Results: The developed algorithms to predict pain utilize variables readily available in patient medical records. Using the machine learning classification algorithm J48 with 50-fold cross-validations, we found that the model can significantly distinguish those with and without pain (*c*-statistics = 0.9108). The *F* measure was 0.856, accuracy rate was 85.68%, sensitivity was 0.862, specificity was 0.852, and precision was 0.849.
Conclusion: Physical and mental function scores, the ability to climb stairs, and overall assessment of feeling were the most discriminative predictors from the 12 identified variables, predicting pain with 86% accuracy for individuals with arthritis. In this era of rapid expansion of big data application, the nature of healthcare research is moving from hypothesis-driven to data-driven solutions. The algorithms generated in this study offer new insights on individualized pain prediction, allowing the development of cost-effective care management programs for those experiencing arthritis pain. ■

Key Words: arthritis, pain, big data analytics, data mining, predictive analytics

INTRODUCTION

Loss of productivity and permanent work disability can be caused by physical limitations that result from pain. The cost of pain in both increased healthcare costs and lowered work productivity has been estimated in a 2008 U.S. sample to range from $560 to $635 billion.[1] Prior research has linked associations among pain, arthritis, and productivity[2,3] and the Centers for Disease Control and Prevention reports that 80% of those with arthritis will have pain-related limitations in movement, with 14% requiring routine needs assistance.[4,5] Varying levels of pain are present in many different types of orthopedic conditions, such as arthritis, back pain, and other musculoskeletal problems.[2,3] Economically, the United States spends close to $80 billion on arthritic conditions in addition to $47 billion lost in consumer earnings.[6] Increased mortality rates, myocardial infarction, work disability,[7,8] fatigue,[9] and poor mental health[10–15] make arthritis and the pain it creates an important public health concern.



igure 3. Predictors of pain tree diagram. PCS, Physical Component Summary; MCS, Mental Component Summary.

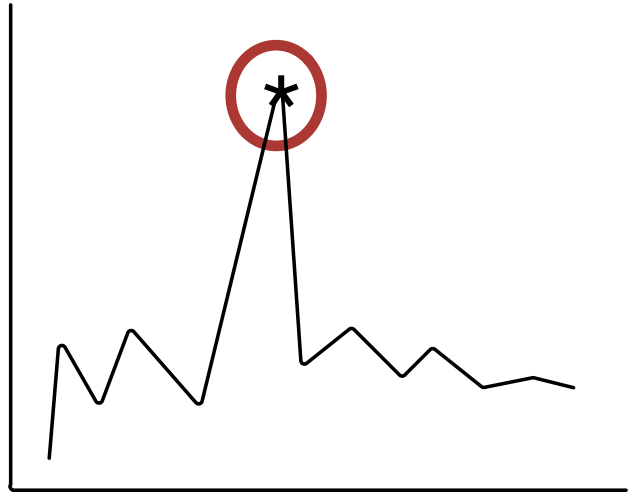
DÉTECTION DES ANOMALIES

Anomalie : événement inattendu, inhabituel, atypique ou statistiquement improbable.

Ne serait-il pas agréable d'avoir un pipeline d'analyse de données qui vous alerte lorsque les choses sortent de l'ordinaire?

De nombreuses approches analytiques différentes à adopter!

- Regroupement
- Méthode naïve de Bayes
- Dérogation aux règles d'association
- Techniques d'ensemble



ÉTUDE DE CAS DE DÉTECTION D'ANOMALIES

Energy 157 (2018) 336–352

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings

Alfonso Capozzoli*, Marco Savino Piscitelli, Silvio Brandi, Daniele Grassi, Gianfranco Chicco

Dipartimento Energia "Galileo Ferraris", Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

ARTICLE INFO

Article history:
Received 5 February 2018
Accepted 19 May 2018
Available online 21 May 2018

Keywords:
Energy consumption
Building energy management
Adaptive symbolic aggregate approximation
Anomaly detection
Data mining
Smart buildings

ABSTRACT

The energy management of buildings currently offers a powerful opportunity to enhance energy efficiency and reduce the mismatch between the actual and expected energy demand, which is often due to an anomalous operation of the equipment and control systems. In this context, the characterisation of energy consumption patterns over time is of fundamental importance. This paper proposes a novel methodology for the characterisation of energy time series in buildings and the identification of infrequent and unexpected energy patterns. The process is based on an enhanced Symbolic Aggregate approXimation (SAX) process, and it includes an optimised tuning of the time window width and of the symbol intervals according to the building energy behaviour. The methodology has been tested on the whole electrical load of buildings for two case studies, and its flexibility and robustness have been confirmed. In order to demonstrate the implications for a preliminary diagnosis, some unexpected trends of the total electrical load have also been discussed in a post-mining phase, using additional datasets related to heating and cooling electrical energy needs.

The process can be used to support stakeholders in characterising building behaviour, to define appropriate energy management strategies, and to send timely alerts based on anomaly detection outcomes.

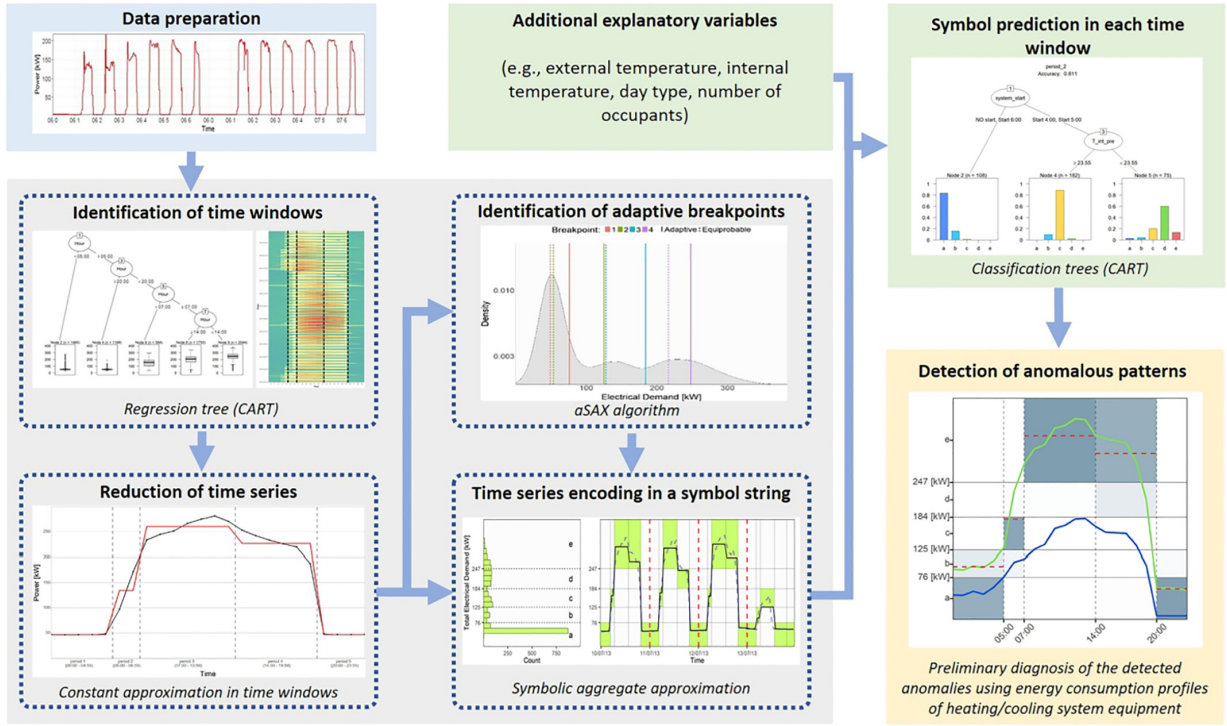


Fig. 2. – Framework for advanced energy consumption characterisation in buildings and anomalous pattern detection.

TECHNIQUES D'APPRENTISSAGE NON SUPERVISÉES

Comportements automatisés par rapport à comportements intelligents

Supervisé : nous vous donnons quelques exemples, vous apprenez d'eux

Non supervisé : vous apprenez par vous-même, en fonction de votre expérience

Techniques non supervisées :

- Règles d'association
- Moteurs de recommandation
- Nouvelles catégories (regroupement)

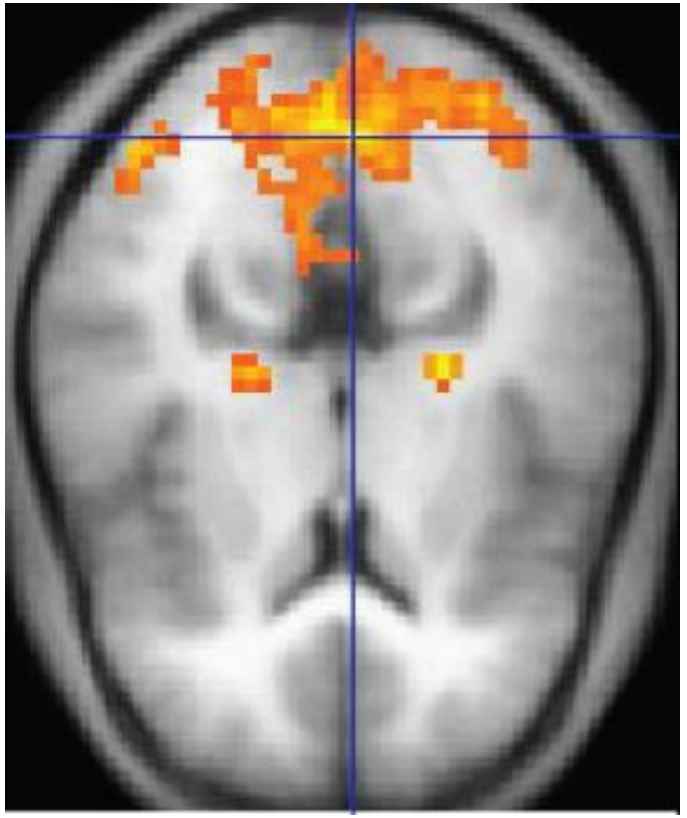
ÉTUDE DE CAS : GROUPEMENT

Les troubles cognitifs légers (TCL) sont un facteur de risque connu du développement de la maladie d'Alzheimer.

Les TCL s'accompagnent de changements dans la structure du cerveau.

Mais quels sont les changements indiquant qu'une personne développera la maladie?

Voici quelques techniques de la science des données qui s'appliquent aux données d'IRM : machines à vecteurs de support, statistiques bayésiennes, intervalles de fonctions de vote, extraction des caractéristiques et (la dernière, mais non la moindre) DBSCAN.



IRMf mettant en évidence certaines régions du cortex préfrontal

QUELQUES DÉFINITIONS PRATIQUES

NOTIONS UNIVERSELLES DE L'ANALYSE DE 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

OBJECTIFS D'APPRENTISSAGE DU MODULE

Connaissance préliminaire des notions suivantes :

- Analyse des données
- Science des données
- Apprentissage automatique
- Schémas/tendances
- Système
- Intelligence artificielle
- Intelligence augmentée

QU'EST-CE QUE L'ANALYSE DES DONNÉES?

Trouver **des tendances** dans les données

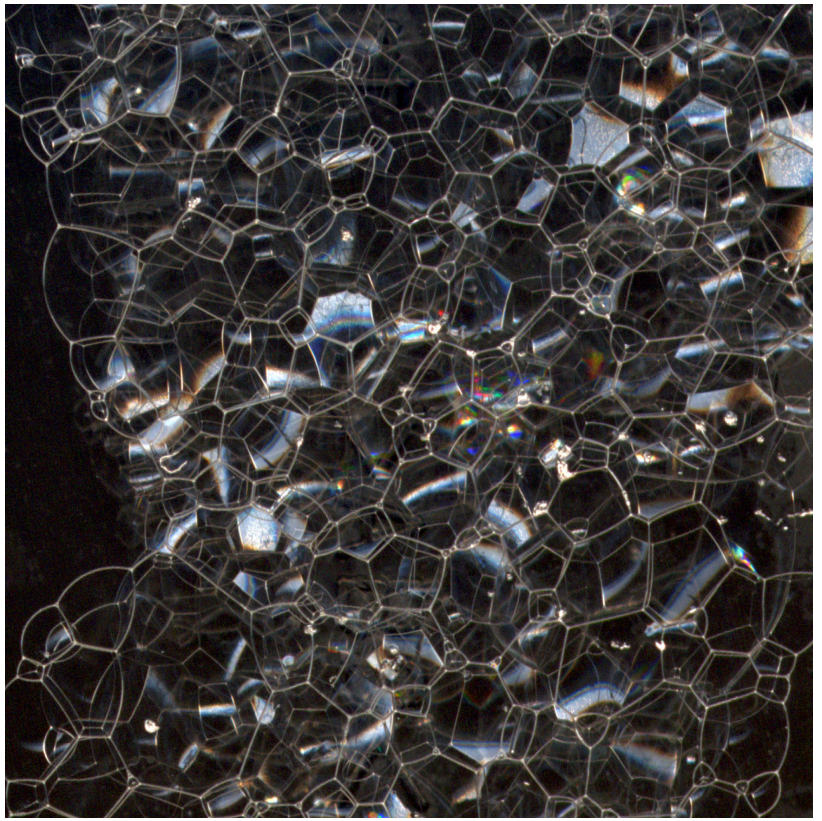
Utiliser les données pour faire quelque chose (répondre à une question, aider à la prise de décision, prédire l'avenir, tirer une conclusion)

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?)



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.

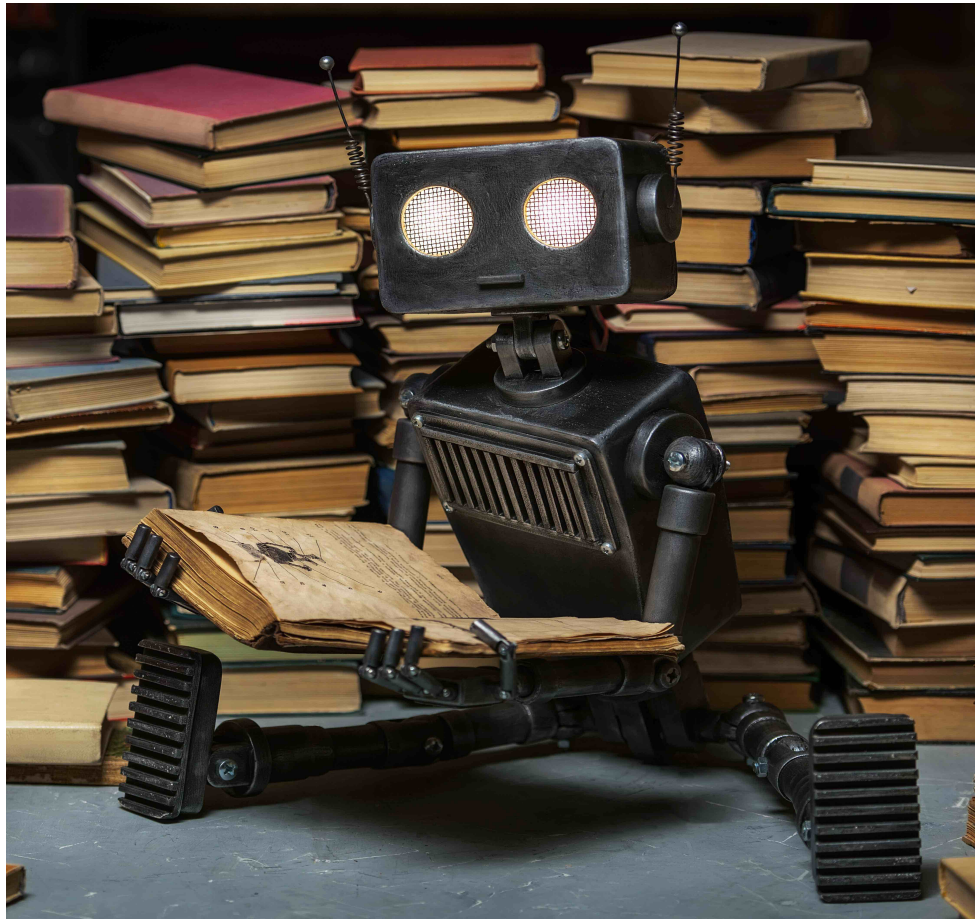
H. Mason (paraphrasé)

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.

FLUX DE TRAVAIL ET SOURCES

NOTIONS UNIVERSELLES DE L'ANALYSE DE DONNÉES

« Tous les modèles sont erronés, mais certains modèles sont utiles. »

George Box [Traduction]

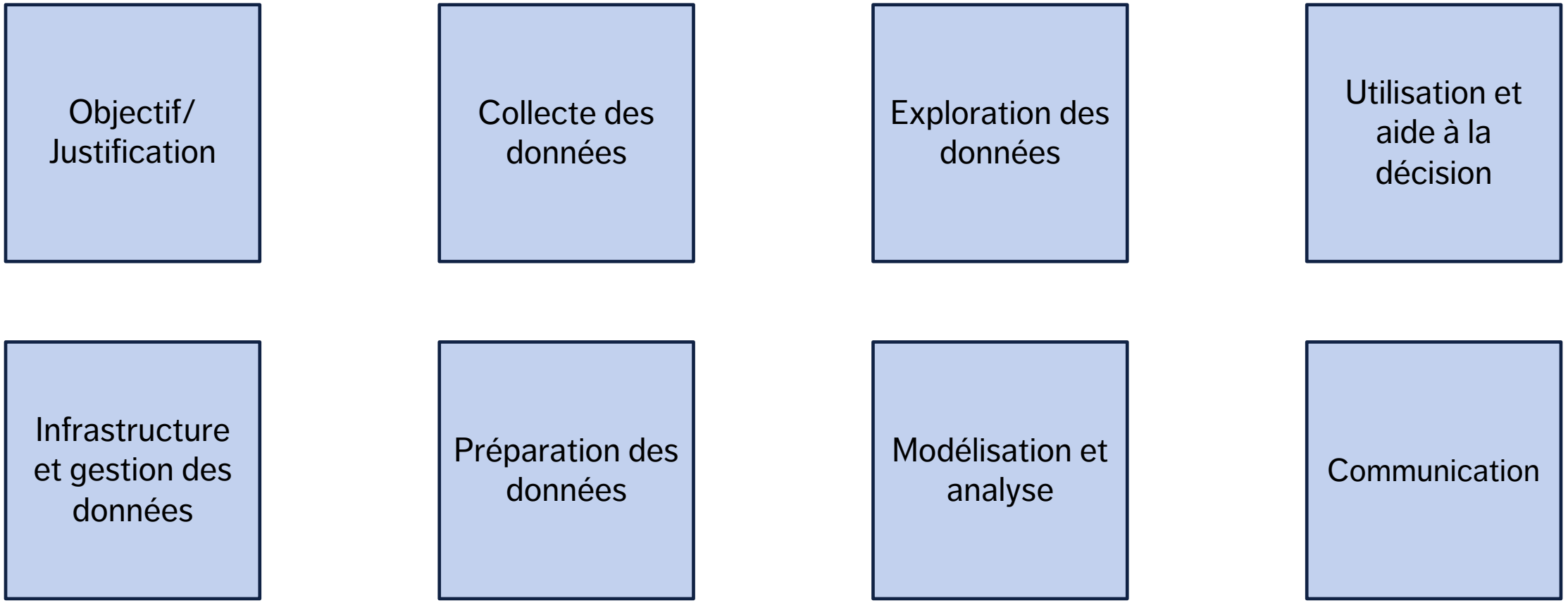
OBJECTIFS D'APPRENTISSAGE DU MODULE

Connaissance préliminaire des notions suivantes :

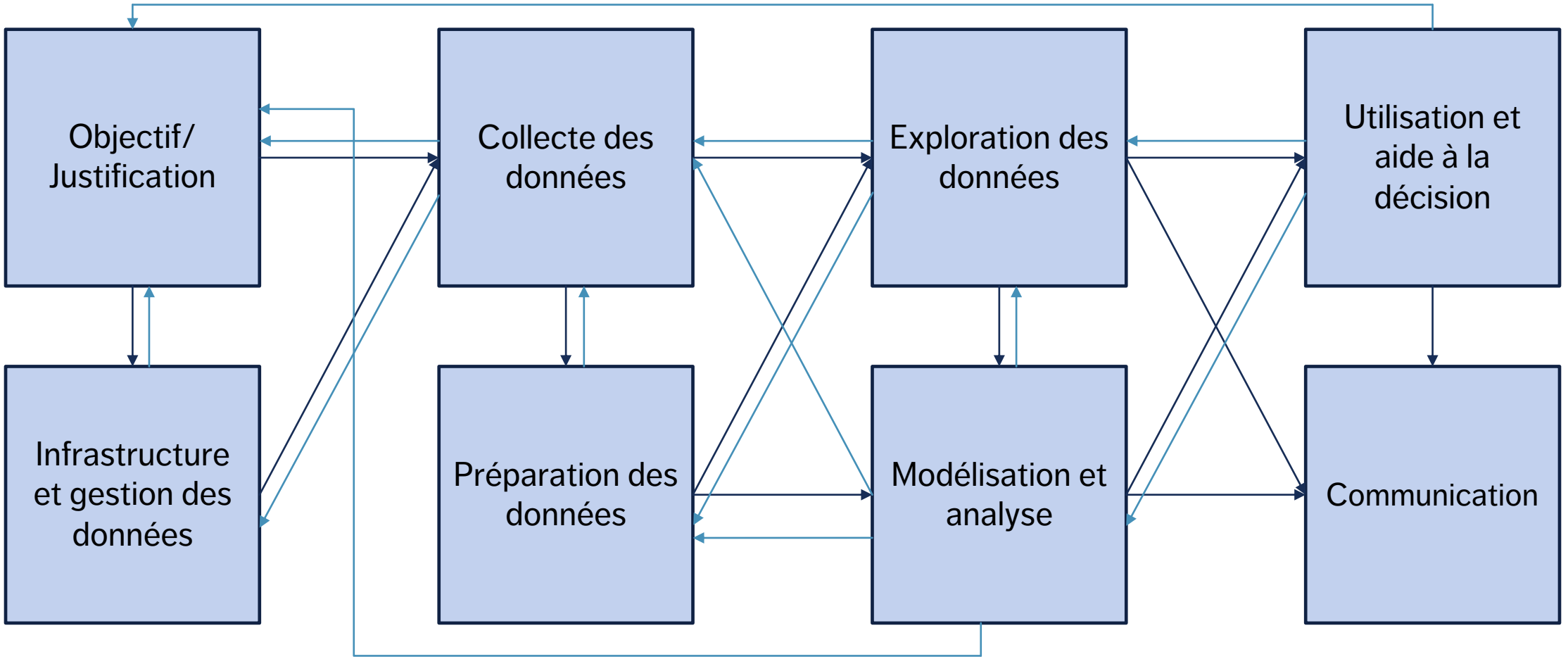
- Le flux de travail et ses composantes (collecte de données, exploration des données, etc.)
- Le modèle analytique
- L'exploration de données
- La décomposition analytique
- L'écosystème de la science des données
- Les équipes de science des données

Prise de conscience de la non-linéarité du processus d'analyse des données.

LE « FLUX DE TRAVAIL » DES DONNÉES



LE « FLUX DE TRAVAIL » DE LA SCIENCE DES DONNÉES



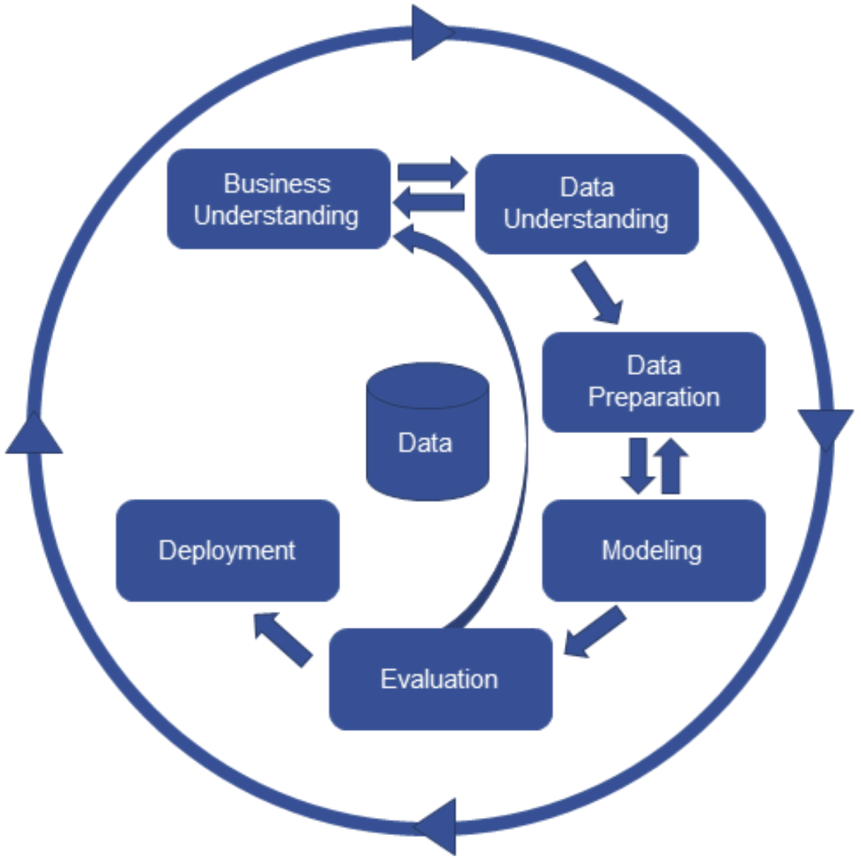
LE PROCESSUS D'ANALYSE DES DONNÉES

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

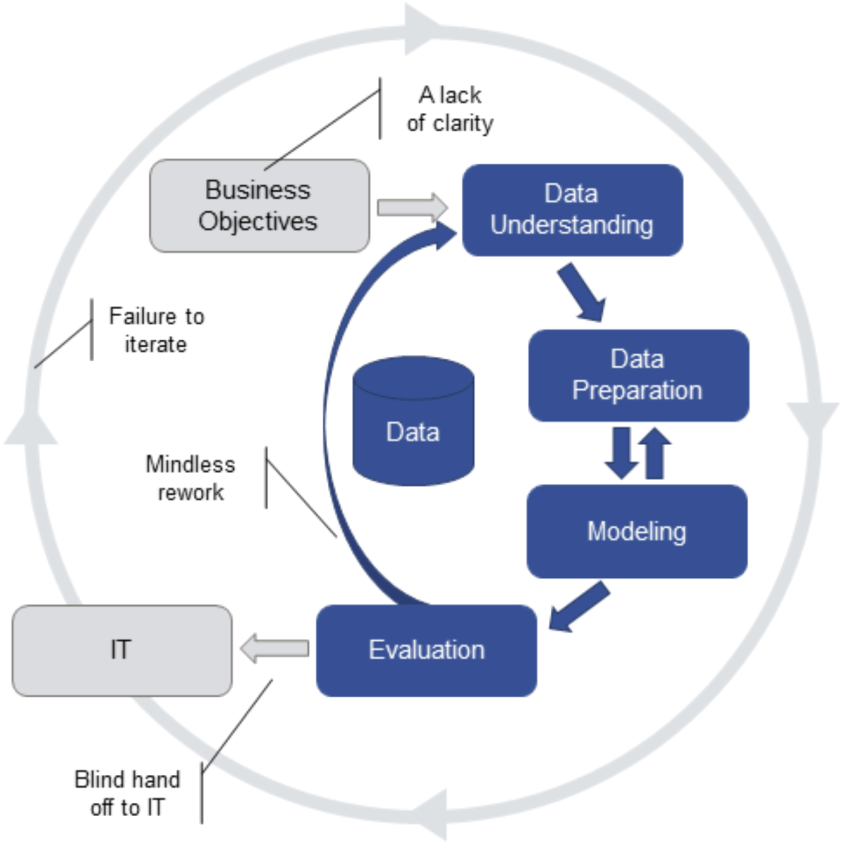
Processus itératif : la sélection des caractéristiques et la réduction des données peuvent nécessiter de nombreuses visites chez des experts du domaine avant que les modèles commencent à donner des résultats prometteurs.

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**.

PROCESSUS INTERSECTORIEL STANDARD, EXPLORATION DE DONNÉES



PROCESSUS INTERSECTORIEL STANDARD, EXPLORATION DE 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!**

MODÈLES ET PENSÉE SYSTÉMIQUE

NOTIONS UNIVERSELLES DE L'ANALYSE DE DONNÉES

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

Inconnu

OBJECTIFS D'APPRENTISSAGE DU MODULE

Connaissance préliminaire des notions suivantes :

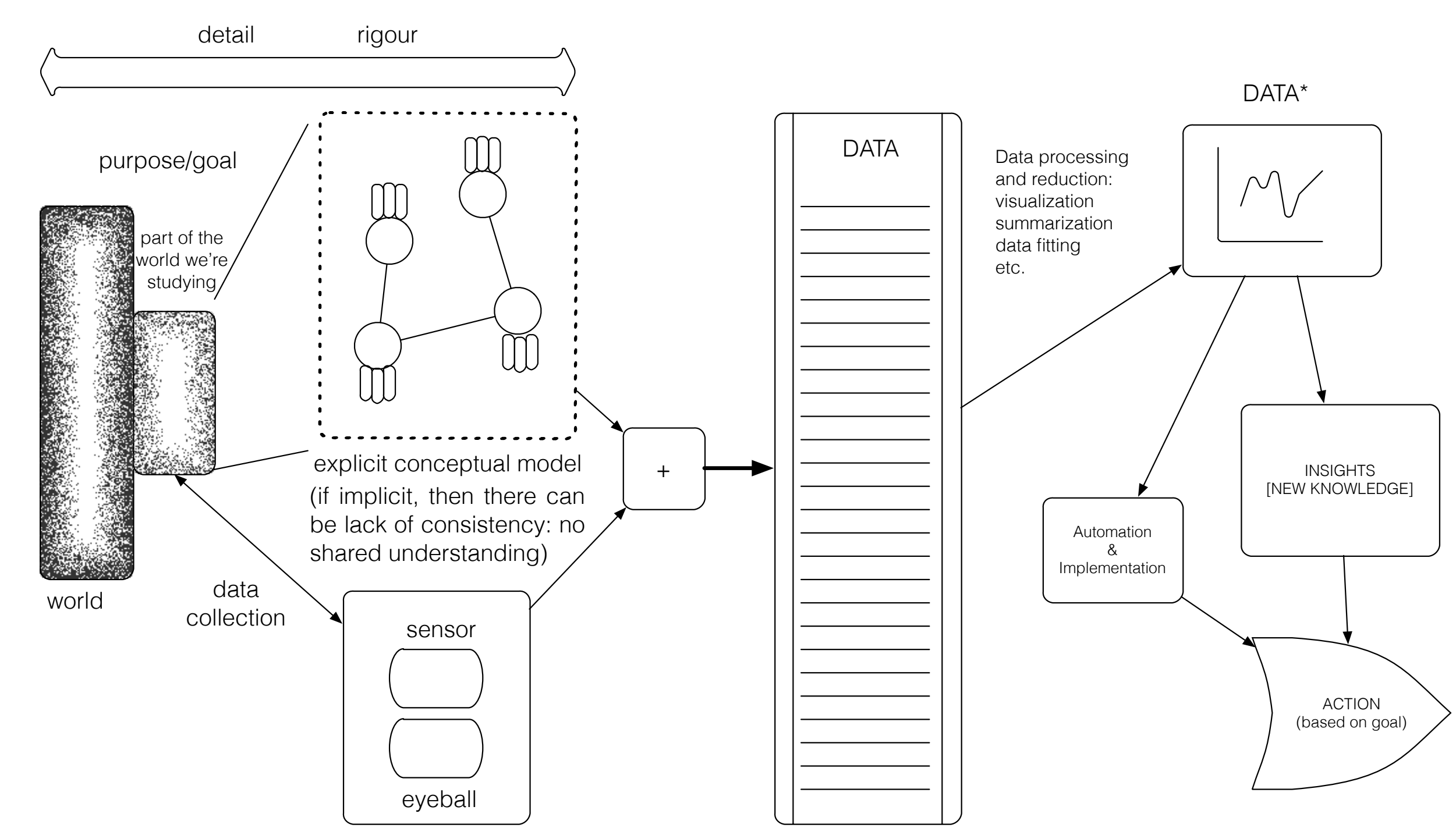
- Représentation
- Systèmes
- Modèles
- Propriétés
- Lacunes dans les connaissances
- Modèle conceptuel

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.


detail
rigour
purpose/goal
part of the world we're studying
world
explicit conceptual model (if implicit, then there can be lack of consistency: no shared understanding)
data collection
sensor
eyeball
+
DATA
DATA*
Data processing and reduction: visualization summarization data fitting etc.
Automation & Implementation
INSIGHTS [NEW KNOWLEDGE]
ACTION (based on goal)

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?

À RETENIR

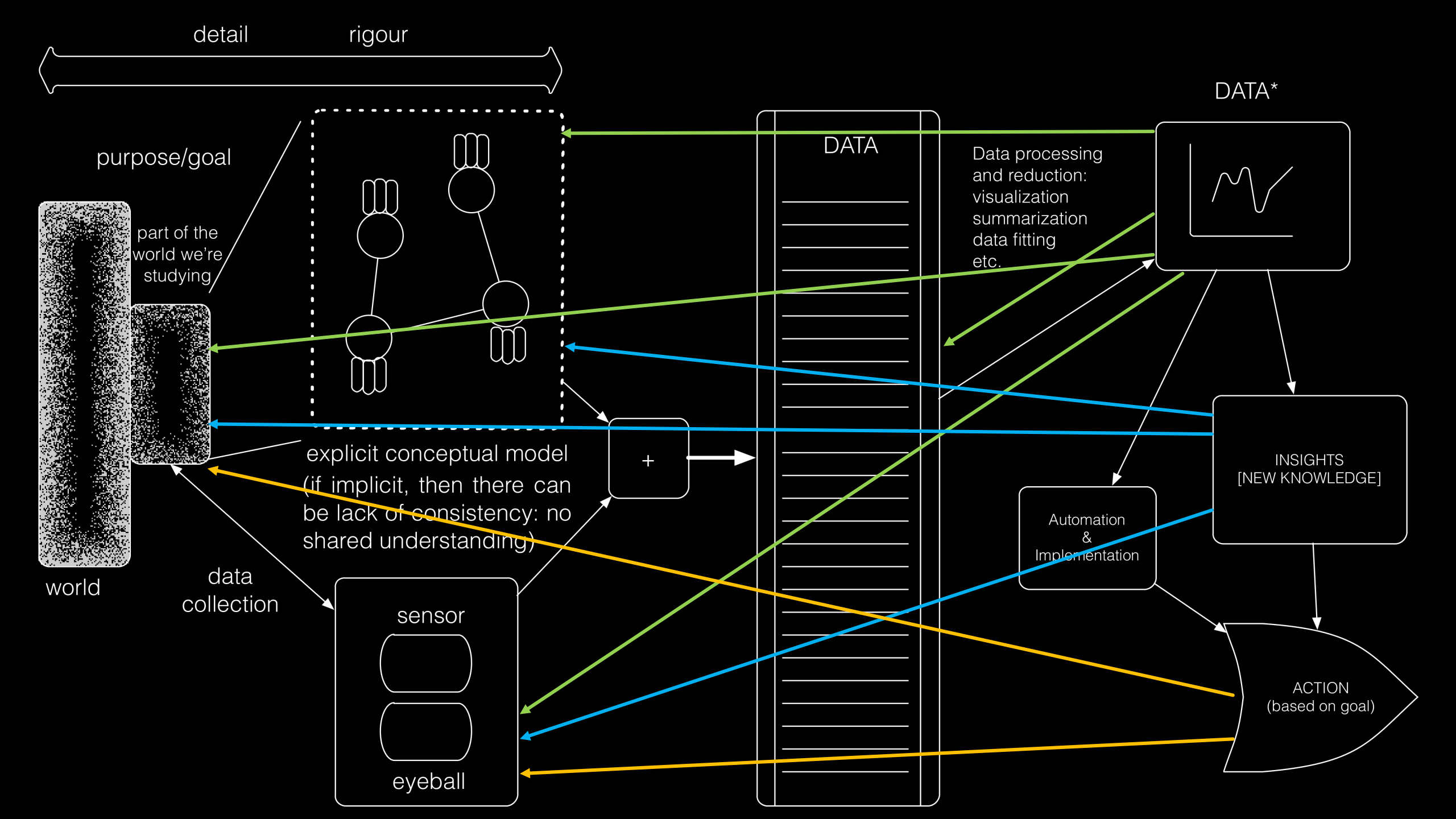
Certains aspects de l'univers peuvent être estimés approximativement à l'aide de systèmes.

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

Il y a des lacunes dans les connaissances. Soyez prêt à revoir votre configuration régulièrement.

Nous ne nous appuyons souvent que sur une modélisation conceptuelle implicite, mais cette méthode comporte des risques.

Si les données, le système et le monde ne sont pas harmonisés, les informations pourraient être inutiles.


detail
rigour
purpose/goal
part of the world we're studying
world
data collection
explicit conceptual model (if implicit, then there can be lack of consistency: no shared understanding)
+
sensor
eyeball
DATA
DATA*
Data processing and reduction: visualization summarization data fitting etc.
INSIGHTS [NEW KNOWLEDGE]
Automation & Implementation
ACTION (based on goal)

CONSIDÉRATIONS ÉTHIQUES ET PRATIQUES EXEMPLAIRES

NOTIONS UNIVERSELLES DE L'ANALYSE DE DONNÉES

« Nous avons volé dans les airs comme des oiseaux et nagé dans la mer comme des poissons, mais nous n'avons pas encore appris à marcher sur Terre comme des frères. »

Martin Luther King fils [Traduction]

OBJECTIFS D'APPRENTISSAGE DU MODULE

Connaissance préliminaire des concepts suivants :

- Éthique et pratiques exemplaires
- Principes des Premières Nations (PCAP)
- Ne pas nuire
- Consentement éclairé
- Vie privée
- Validité du modèle

DISCUSSION

Quels préjudices peuvent être causés par les données?

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.

[Université de Virginie, 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, 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.

LECTURES ET RÉFÉRENCES

NOTIONS UNIVERSELLES DE L'ANALYSE DE DONNÉES

RÉFÉRENCES

Premières Nations – PCAP

Article de Wikipédia sur l'apprentissage semi-supervisé

Article de Wikipédia sur l'apprentissage supervisé

Article de Wikipédia sur l'apprentissage par renforcement

Article de Wikipédia sur l'apprentissage non supervisé

J. Blitzen [2017], *What is it like to design a data science class?*, réponse sur le site *Quora*

J. Taylor [2017], *4 Problems with CRISP-DM*, KDNuggets.

Brin, D. [1998], *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?*, Perseus.

RÉFÉRENCES

Mayer-Schönberger, V. et Cukier, K. [2013], *Big Data : la révolution des données est en marche*, Robert Laffont.

Mayer-Schönberger, V. [2009], *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press.

Data Science Association, *Data Science Code of Professional Conduct*.

Chen, M. [2013], *Is 'Big Data' Actually Reinforcing Social Inequalities?*, The Nation.

Shin, L. [2013], *How the New Field of Data Science is Grappling With Ethics*, SmartPlanet.

Schutt, R. and O'Neill, C. [2013], *Doing Data Science: Straight Talk From the Front Line*, O'Reilly.

O'Neill, C. [2016], *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown.

RÉFÉRENCES

Chang, R.M., Kauffman, R.J., Kwon, Y. [2014], *Understanding the paradigm shift to computational social science in the presence of big data*, Decision Support Systems, 63:67–80, Elsevier.

Hurlburt, G.F., Voas, J. [2014], *Big Data, Networked Worlds*, IEEE Computer Society.

Introna, L.D. [2007], *Maintaining the reversibility of foldings: Making the ethics (politics) of information technology visible*, Ethics and Information Technology, 9:11–25, Springer.

Floridi, L. [2011], *The philosophy of information*, Oxford University Press.

Floridi, L. (ed) [2006], *The Cambridge handbook of information and computer ethics*, Cambridge University Press, 2006.

Big Data & Ethics

Mason, H. [2012], What is a Data Scientist?, Forbes.

RÉFÉRENCES

Schlimmer, J.S. [1987], *Concept Acquisition Through Representational Adjustment (Technical Report 87-19)*. Department of Information and Computer Science, UCalifornia, Irvine.

Iba, W., Wogulis, J., Langley, P. [1988], *Trading off Simplicity and Coverage in Incremental Concept Learning*, in Proceedings of the 5th International Conference on Machine Learning, 73-79. Ann Arbor, Michigan: Morgan Kaufmann.

Gorelik, B. [2017], Don't study data science as a career move; you'll waste your time!, gorelik.net

J. Leskovec, A. Rajaraman, J. Ullman [2015] *Mining of Massive Datasets*, Cambridge University Press.

Hastie, T., Tibshirani, R., and J. Friedman [2008], *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer.

Provost, F., Fawcett, T. [2013], *Data Science for Business*, O'Reilly.