Apprentissage statistique

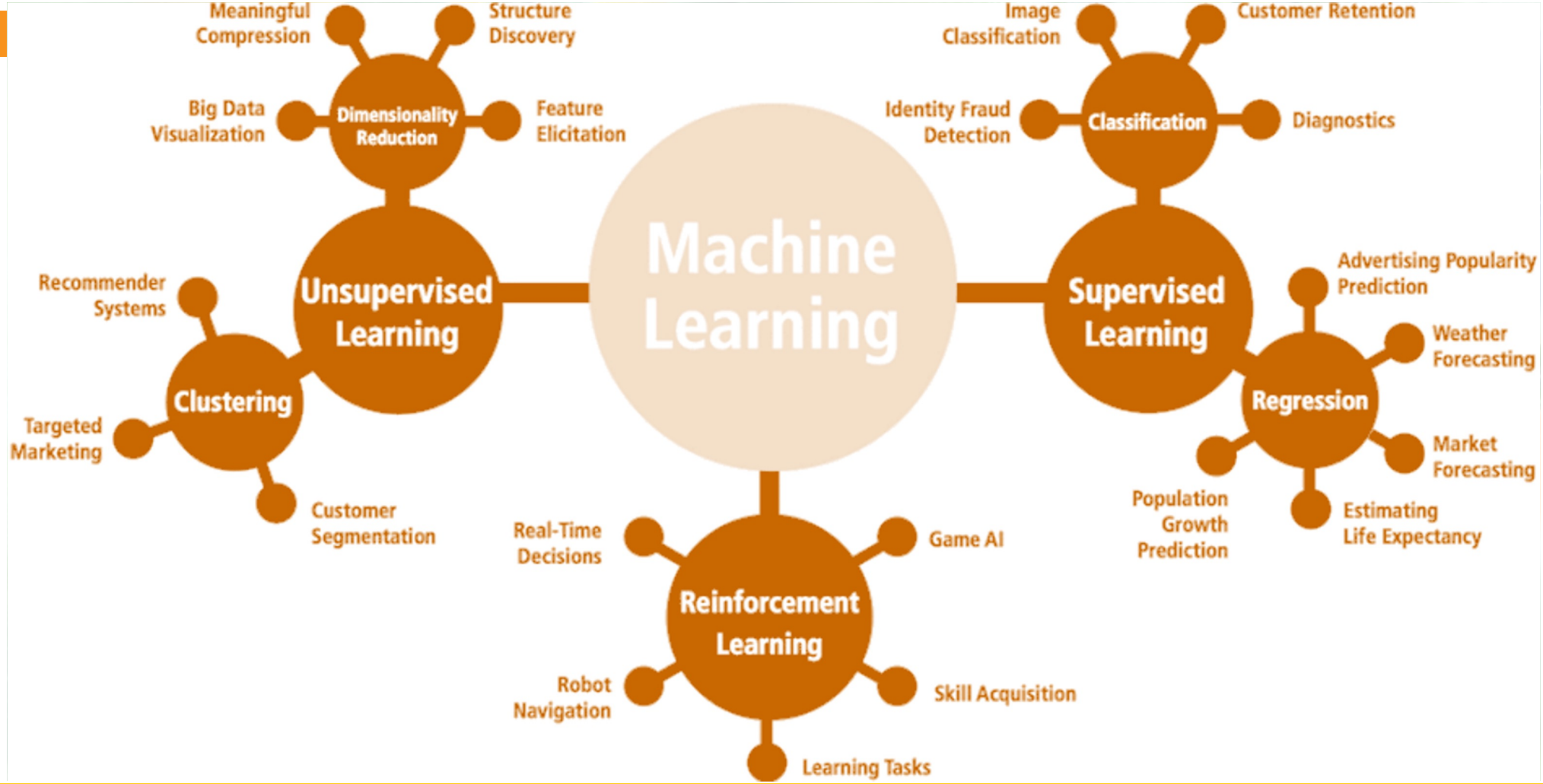
INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE

Nous apprenons de l'échec, et non du succès !

[B. Stoker, Dracula]

Les données ne sont pas des informations, les informations ne sont pas des connaissances, les connaissances ne sont pas de la compréhension, la compréhension n'est pas de la sagesse.

[C. Stoll (attribué), *Nothing to Hide : Privacy in the 21st Century*, 2006].



1. Types d'apprentissage et tâches

Un défi moderne

L'un des défis du travail dans les domaines de la **science des données** (DS), de l'**apprentissage automatique** (ML) et de l'**intelligence artificielle** (AI) est que la plupart des travaux quantitatifs peuvent être décrits comme DS/ML/AI (ce qui est souvent étiré jusqu'au ridicule).

La DS/ML/AI consiste en des **processus quantitatifs** (ce que Mason a appelé l'**intersection** des statistiques, de l'ingénierie, de l'informatique, de l'expertise du domaine et du "hacking") qui aident les utilisateurs à **obtenir des informations exploitables** sur leur situation sans pour autant renoncer complètement à leur responsabilité décisionnelle.

Un cadre hiérarchique

Robinson propose une "**structure hiérarchique inclusive**" :

1. dans un premier temps, le DS fournit des "**aperçus**" *via la* visualisation et l'analyse inférentielle (manuelle) ;
2. dans un deuxième temps, le ML produit des "**prédictions**" (ou "**conseils**"), tout en réduisant (sans l'éliminer) la charge de travail analytique, inférentielle et décisionnelle de l'opérateur ;
3. dans un dernier temps, l'AI **supprime le besoin de surveillance continue**, permettant à un système généralement sans surveillance de prendre des "mesures" automatiques (AI générale par opposition à intelligence augmentée).

Dans la pratique, les parties prenantes ne devraient probablement pas chercher à abdiquer **tout** leur pouvoir dans le processus de prise de décision.

Apprentissage

Les êtres humains apprennent (à toutes les étapes) en s'imprégnant d'abord de leur **environnement,** puis en:

- répondant à des questions à ce sujet ;
- testant des hypothèses ;
- créant de concepts ;
- faisant des prédictions ;
- créant des catégories, et
- en classant et en regroupant les différents objets et attributs.

Apprentissage statistique/automatique

L'objectif principal du DS/ML/AI est d'essayer d'**apprendre** aux machines à extraire des informations des données, de manière correcte et efficace, et sans préjugés ni idées préconçues – **pouvons-nous** (devrions-nous ?) **concevoir des algorithmes capables d'apprendre ?**

La méthode DS/ML/AI la plus simple consiste à **explorer des données représentatives** :

- fournir un résumé à l'aide de statistiques de base - moyenne, mode, histogrammes, etc ;
- rendre sa structure multidimensionnelle évidente grâce à la visualisation des données
- rechercher la cohérence, en tenant compte de ce qui est présent et de ce qui manque.

Apprentissage supervisé

L'apprentissage supervisé (SL) s'apparente à "l'**apprentissage avec un professeur**" : les étudiants donnent une réponse à chaque question d'examen sur la base de ce qu'ils ont appris à partir d'exemples résolus fournis par le professeur/le manuel ; le professeur fournit les réponses correctes et note les questions d'examen à l'aide d'une clé de réponse.

Les tâches typiques comprennent :

- la **classification**
- la régression
- les classements
- les recommandations

Apprentissage supervisé

Dans le cadre du SL, les algorithmes utilisent des **données d'apprentissage étiquetées** pour construire (former) un **modèle prédictif** ; la performance de chaque algorithme est évaluée à l'aide de **données de test** dont l'étiquette est connue, mais qui n'est pas utilisée dans la prédiction.

Il existe des **cibles** fixes pour entraîner le modèle (telles que les catégories d'âge ou les espèces végétales) – ces catégories/classes (et leur nombre) sont **connues avant l'analyse.**

Apprentissage non supervisé

L'apprentissage non supervisé (UL) s'apparente à "l'**auto-apprentissage par le regroupement d'exercices similaires en tant que guide d'étude**" : l'enseignante n'est pas impliquée dans le processus de découverte et les étudiantes peuvent aboutir à des regroupements différents.

Les tâches typiques comprennent :

- le **regroupement**
- la **découverte de règles d'association**
- le profilage des liens
- la détection des anomalies

Apprentissage non supervisé

Les algorithmes non supervisés utilisent des **données non étiquetées** pour trouver des **modèles naturels** dans les données ; l'inconvénient est que la précision **ne peut pas être évaluée** avec le même degré de satisfaction.

Dans l'UL, nous ne savons pas quelle est la cible, ni même s'il y en a une – nous recherchons simplement des **groupes naturels** dans les données :

- les élèves du secondaire qui aiment la littérature, ont des cheveux longs et savent cuisiner **vs.**
- les étudiants qui font partie d'une équipe sportive et qui ont des frères et sœurs **vs.**
- les professionnels de la finance qui ont un penchant pour les films de super-héros, la bière artisanale et les sacs à dos Hello Kitty **vs.** ...

Autres cadres d'apprentissage

Certaines techniques de DS/ML/AI s'inscrivent dans les deux camps, mais il existe d'autres **approches conceptuelles** :

- **l'apprentissage semi-supervisé** dans lequel certains points de données sont étiquetés mais la plupart ne le sont pas, ce qui se produit souvent lorsque l'acquisition de données est coûteuse ("l'enseignant fournit des exemples travaillés et une liste de problèmes non résolus à essayer ; les étudiants essaient de trouver des groupes similaires de problèmes non résolus et les comparent aux problèmes résolus pour trouver des correspondances").
- **l'apprentissage par renforcement**, où un agent tente d'obtenir autant de récompenses (à court terme) que possible tout en minimisant les regrets (à long terme) ("se lancer dans un doctorat avec une superviseure; il y a des hauts et des bas, et le diplôme n'est pas certain").

Apprentissage statistique/automatique

L'expression "apprentissage statistique" n'est pas courante dans la pratique ; on a plutôt tendance à parler d'**apprentissage automatique** (ML).

Si une distinction doit être faite, on pourrait affirmer que :

- l'apprentissage statistique s'appuie sur des modèles de type statistique et l'accent est généralement mis sur l'**interprétabilité**, la **précision** et l'**incertitude**, alors que
- l'apprentissage automatique est issu des études sur l'intelligence artificielle et met l'accent sur les **applications à grande échelle** et la **précision des prédictions**.

La ligne de démarcation entre les deux termes est floue – le vocabulaire utilisé en pratique trahit surtout le niveau/champs d'éducation.

Questions sur la DS/ML

En pratique, les méthodes DS/ML/AI ne sont vraiment intéressantes que lorsqu'elles aident les utilisateurs à poser des questions utiles et à y répondre :

- **Analytique** – "Combien de clics ce lien a-t-il obtenu ?"
- **Science des données** – "Sur la base de l'historique précédent des clics sur les liens du site, puis-je prédire combien de Manitobains liront une page spécifique dans les trois prochaines heures ?" ou "Existe-t-il une relation entre l'historique des clics sur les liens et le nombre de personnes du Manitoba qui liront cette page spécifique ?"
- **Méthodes quantitatives** – "Nous n'avons pas de pages similaires dont l'historique pourrait être consulté pour faire une prédiction, mais nous avons des raisons de penser que le nombre de visites sera fortement corrélé avec la température à Winnipeg. En utilisant les prévisions météorologiques pour la semaine prochaine, pouvons-nous prédire combien de personnes accèderont à cette page spécifique au cours de cette période ?"

Questions sur la DS/ML

Les modèles DS/ML sont **prédictifs/descriptifs** (et non **explicatifs/ prescriptifs**) : ils montrent les connexions et exploitent les corrélations pour faire des prédictions, mais ils ne révèlent pas **pourquoi** ces connexions existent (réseaux bayésiens).

Les méthodes quantitatives, quant à elles, supposent généralement un certain niveau de **compréhension causale** basé sur divers **principes de base**.

Cette distinction n'est pas toujours bien comprise par les analystes et les parties prenantes.

Tâches du ML

Voici quelques tâches courantes de ML (avec questions représentatives) :

- **classification** – quels étudiants du premier cycle sont susceptibles de réussir au second cycle ?
- **estimation des probabilités** – quelle est la probabilité qu'un candidat donné remporte une élection ?
- **estimation des valeurs** – combien un client donné va-t-il dépenser dans un restaurant ?
- **appariement des similitudes** – quels sont les clients potentiels qui ressemblent le plus aux meilleurs clients existants ?
- **regroupement** – les signaux provenant d'un capteur forment-ils des groupes naturels ?
- **découverte de règles d'association** – quels sont les livres communément achetés ensemble en ligne ?
- **profilage** et **description du comportement** – quelle est l'utilisation typique d'un téléphone portable dans un segment de clientèle donné ?
- **link prediction** – J. et K. ont 20 amis en commun : pourraient-elles être amies ?

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

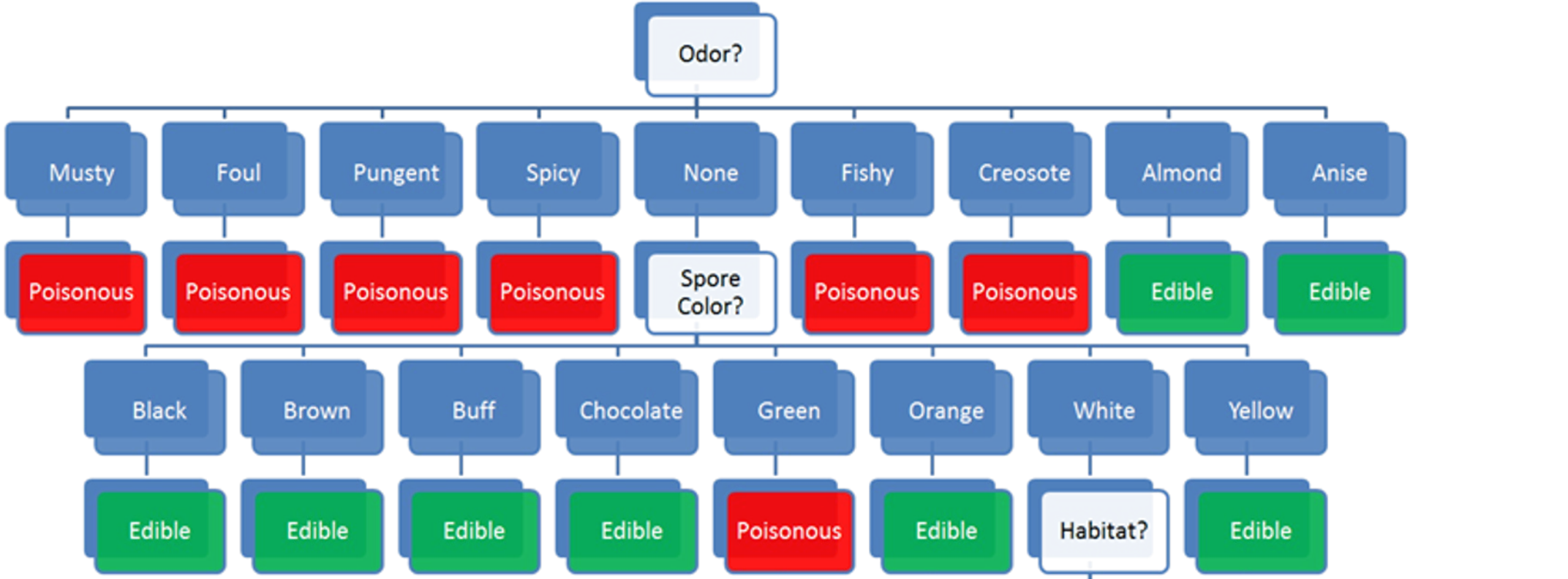
Problème de classification des champignons

Amanita muscaria

Habitat : bois
Taille des branchies : étroites
Odeur : aucune
Spores : blanches
Couleur du capuchon : rouge

Problème de classification :
Amanita muscaria est-il comestible ou toxique ?

Session 1



Habitat : bois

Taille des branchies : étroites

Odeur : aucune

Spores : blanches

Couleur du capuchon : rouge

Problème de classification :
comestible ou toxique

Session 1

Habitat : bois

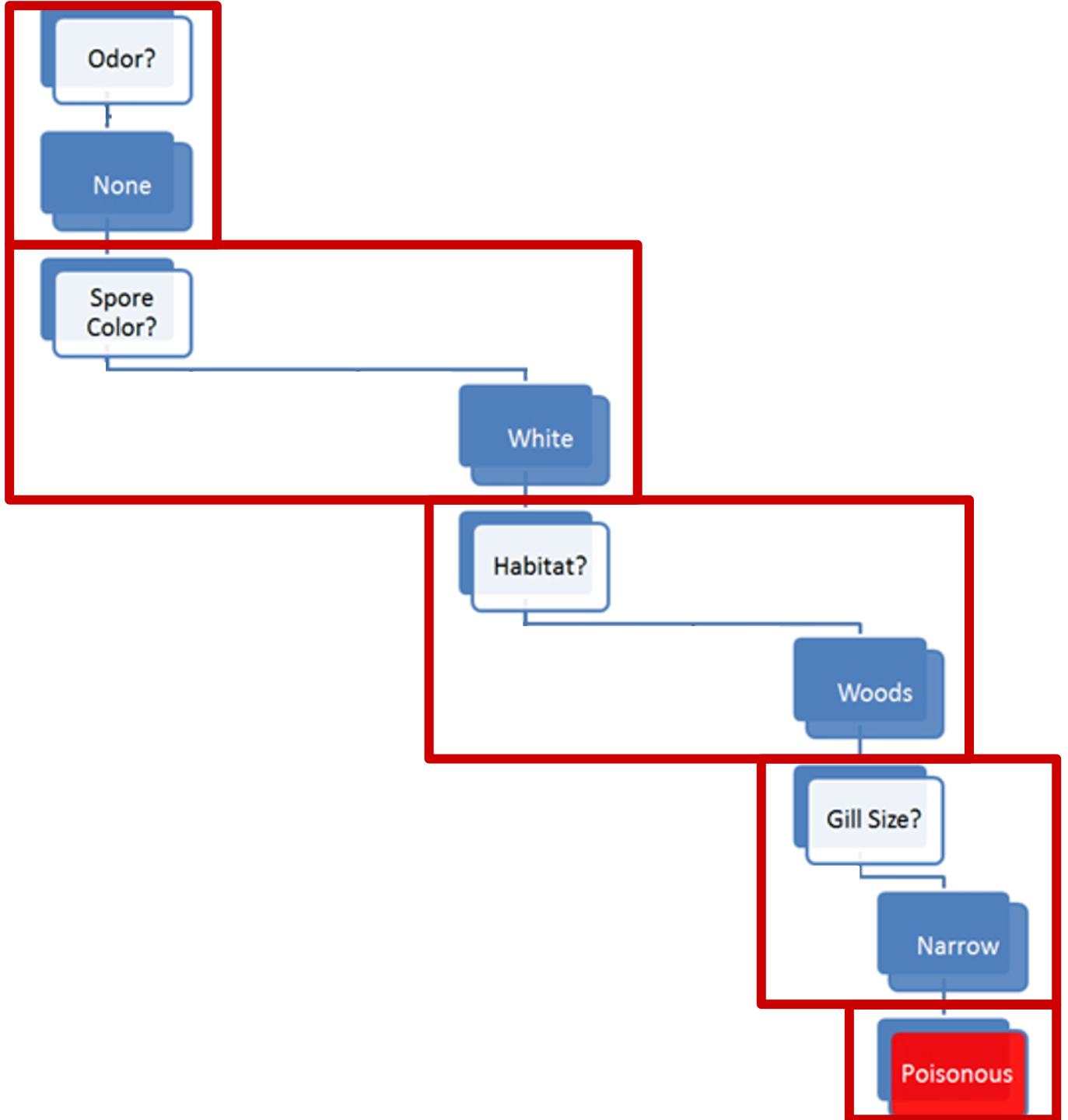
Taille des branchies : étroites

Odeur : aucune

Spores : blanches

Couleur du capuchon : rouge

Problème de classification :
comestible ou **toxique**



Classification des champignons

Auriez-vous fait confiance à une prédiction qui dit "**comestible**" ?

D'où vient le modèle ?

Que devez-vous savoir pour faire confiance au modèle ?

Quel est le coût d'une erreur de classification ?

Lectures suggérées

Types d'apprentissage et
tâches d'apprentissage automatique

D. Robinson, "What's the difference between data science, machine learning, and artificial intelligence?" *Variance Explained*, Jan. 2018.

D. Woods, "Bitly's Hilary Mason on "what is a data scientist?"," *Forbes*, Mar. 2012.

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    - 19.1 Preliminaries
    - 19.2 Statistical Learning

Exercices

Types d'apprentissage et tâches d'apprentissage automatique

1. De quels types de tâches d'apprentissage automatique les problèmes suivants sont-ils représentatifs ?
   - Identification des facteurs de risque associés au cancer du sein et de la prostate.
   - Prédire si un patient aura une deuxième crise cardiaque mortelle dans les 30 jours suivant la première sur la base des données démographiques, du régime alimentaire, des mesures cliniques, etc.
   - Établir la relation entre le salaire et les informations démographiques dans les données d'enquête sur la population.
   - Prévoir le taux d'inflation annuel à l'aide de différents indicateurs.

2. Quels sont des exemples de tâches d'apprentissage automatique supervisé, non supervisé, semi-supervisé, par renforcement dans le monde de l'entreprise ? Dans un contexte de politique publique/de gouvernement ? Dans un contexte scientifique ?

Exercices

Types d'apprentissage et
tâches d'apprentissage automatique

3. En supposant que les techniques DS/ML/AI soient utilisées dans les cas suivants, déterminez si la tâche requise relève de SL ou UL.
   a. L'estimation du temps de réparation nécessaire pour un aéronef sur la base d'une fiche de panne.
   b. Décider d'accorder ou non un prêt à un demandeur sur la base de données démographiques et financières (en se référant à une base de données contenant des données similaires sur des clients antérieurs).
   c. Dans une librairie en ligne, le fait de recommander aux clients d'acheter d'autres articles en fonction de leurs habitudes d'achat lors de transactions antérieures.
   d. Identification d'un paquet de données réseau comme dangereux (virus, attaque de pirates informatiques) sur la base d'une comparaison avec d'autres paquets dont l'état de menace est connu.
   e. Identifier des segments de clients similaires.

Exercices

Types d'apprentissage et
tâches d'apprentissage automatique

3. En supposant que les techniques DS/ML/AI soient utilisées dans les cas suivants, déterminez si la tâche requise relève de SL ou UL.
   f. Prévoir si une entreprise fera faillite en comparant ses données financières à celles d'entreprises similaires en faillite ou non.
   g. Tri automatisé du courrier par balayage du code postal.
   h. Il est plus difficile et plus coûteux de gagner de nouveaux clients que de fidéliser les clients existants. L'évaluation de chaque client en fonction de sa probabilité de quitter l'entreprise peut aider une organisation à concevoir des interventions efficaces, telles que des remises ou des services gratuits, pour fidéliser les clients rentables de manière rentable.
   i. Certains praticiens médicaux effectuent des tests inutiles et/ou surfacturent leur gouvernement ou leur compagnie d'assurance. Grâce aux données d'audit, il peut être possible d'identifier ces prestataires et de prendre les mesures qui s'imposent.

Exercices

Types d'apprentissage et tâches d'apprentissage automatique

3. En supposant que les techniques DS/ML/AI soient utilisées dans les cas suivants, déterminez si la tâche requise relève de SL ou UL.
   j. Une analyse du panier de la ménagère peut aider à développer des modèles prédictifs pour déterminer quels sont les produits qui se vendent souvent ensemble. Cette connaissance des affinités entre les produits peut aider les détaillants à créer des offres promotionnelles groupées pour pousser les articles qui ne se vendent pas vers un ensemble de produits se vendant bien.
   k. Le diagnostique de la cause d'un état pathologique est la première étape cruciale de l'engagement médical. Outre l'état actuel, d'autres facteurs peuvent être pris en compte, tels les antécédents de santé du patient, ses antécédents médicaments, les antécédents de sa famille et d'autres facteurs environnementaux. Un modèle prédictif peut absorber toutes les informations disponibles à ce jour (pour ce patient et d'autres) et établir des diagnostiques probabilistes, sous la forme d'un arbre de décision, en supprimant la plupart des conjectures.
   l. Les écoles peuvent développer des modèles pour identifier les élèves qui risquent de ne pas retourner à l'école. Ces élèves peuvent être signalés pour faire l'objet de mesures correctives.

Exercices

Types d'apprentissage et tâches d'apprentissage automatique

3. En supposant que les techniques DS/ML/AI soient utilisées dans les cas suivants, déterminez si la tâche requise relève de SL ou UL.
    m. Outre les données relatives aux clients, les entreprises de télécommunications conservent également des enregistrements détaillés des appels (CDR), qui décrivent avec précision le comportement de chaque client en matière d'appels. Ces données uniques peuvent être utilisées pour établir le profil des clients, qui peuvent être commercialisés en fonction de la similitude de leur CDR avec celui d'autres clients.
    n. Statistiquement, tous les équipements sont susceptibles de tomber en panne à un moment ou à un autre. Prévoir quelle machine est susceptible de s'arrêter est un processus complexe. Des modèles de décision permettant de prévoir les défaillances des machines peuvent être élaborés à partir de données antérieures, ce qui peut permettre de réaliser des économies grâce à la maintenance préventive.
    o. Identifier les tweets qui contiennent de la désinformation et ceux qui sont légitimes.