

# Les principes fondamentaux de la science des données

---

P. BOILY

UNIVERSITÉ D'OTTAWA | FACULTÉ DES SCIENCES | DÉPARTEMENT DES MATHÉMATIQUES ET DES STATISTIQUES  
DATA ACTION LAB | IDLEWYLD ANALYTICS

# Instructeur – Patrick Boily

## Emploi

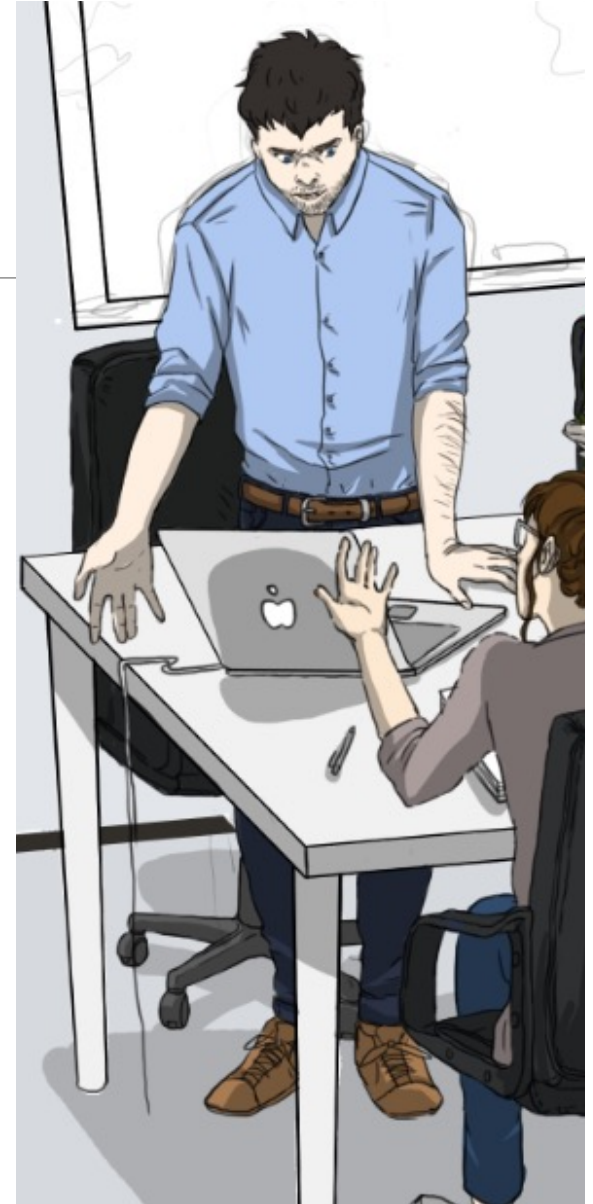
- Professeur Math/Stat [depuis '19, uOttawa]
- Président et conseiller principal [depuis '16, Idlewyld Analytics]
- Directeur et consultant principal ['12 - '19, CQADS, Carleton]
- Fonctionnaire GdC ['08 - '12, ASFC | StatCan | TC | TPSGC]
- 60+ cours universitaires ; 250+ jours d'atelier

## Projets

- GAC ; NWMO ; CATSA ; etc.
- 40+ projets

## Spécialisation

- Visualisation des données ; nettoyage des données (... malheureusement)
- Application d'un large éventail de techniques à tous les types de données
- Modélisation mathématique/statistique



# Matériel de cours

---

**Page Web du cours :**

<https://data-action-lab.com/101-dse>

**Notes de cours :**

<https://idlewyldanalytics.com>

**Contact :**

[pboily@uottawa.ca](mailto:pboily@uottawa.ca)

**Espace de travail Slack :**

<https://dspdi.slack.com>

# Description du cours

---

Ce cours offre aux participants l'occasion de maîtriser les connaissances et compétences fondamentales nécessaires à l'analyse des données.

Les participants seront initiés à diverses méthodes de préparation des données, à certaines limites intrinsèques des données et de l'analyse des données, ainsi qu'à des erreurs de pré-analyse facilement évitables.

Après le cours, les participants ont la possibilité de travailler sur un projet guidé, en recevant du feedback de l'instructeur.

# Informations supplémentaires

---

Une exposition à la programmation et aux concepts introduits dans un premier cours universitaire de probabilités et de statistiques serait bénéfique (mais pas nécessaire).

Les participants sont encouragés à apporter un ordinateur portable/personnel sur lequel la version actuelle de R/Rstudio est installée (pour lequel ils peuvent avoir besoin d'une autorisation administrative pour installer des paquets).

Les participants au projet guidé doivent être familiers avec R et/ou Python.

# Objectifs d'apprentissage

---

À la fin de ce cours, les participants seront en mesure de :

- sélectionner des méthodes appropriées pour préparer leurs données à l'analyse
- anticiper les défis et les limites inhérents aux données et aux résultats d'analyse souhaités
- appliquer des stratégies de nettoyage des données à leurs données
- effectuer des analyses simples
- construire de simples pipelines de science des données

# Plan du cours

---

## Les aspects techniques et non techniques des données

1. Les compétences quantitatives  
Les logiciels et les outils  
L'approche des "I" multiples  
Les rôles et les responsabilités  
Aide-mémoire

## Les bases de la science des données

2. Les préliminaires
3. Les cadres conceptuels
4. L'éthique de la science des données
5. Le flux de travail analytique
6. Les données et les renseignements

Session 1

Session 2

Session 3

Session 4

# Plan du cours

---

## La préparation des données

7. La qualité et le traitement des données
8. Les valeurs manquantes
9. Les observations anormales
10. La dimensionnalité et les transformations de données

## Miscellanea

11. L'ingénierie des données
12. La gestion des données

Session 1

Session 2

Session 3

Session 4



# Problème des champignons vénéneux

*Amanita muscaria*

**Habitat :** bois

**Taille des branchies :** étroites

**Odeur :** aucune

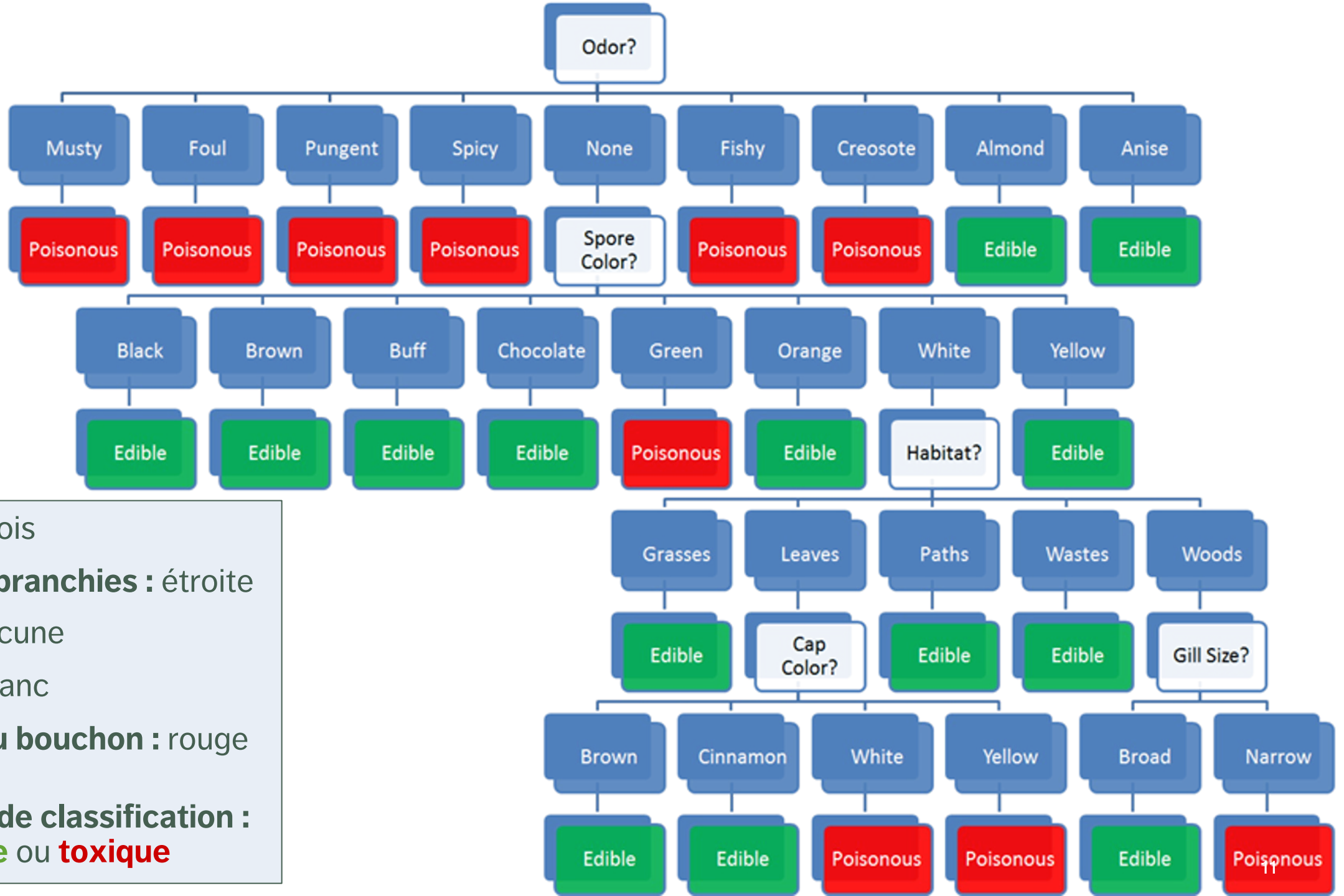
**Spores :** blanc

**Couleur du chapeau :** rouge

**Problème de classification :**

L'*Amanita muscaria* est-elle comestible ou toxique ?





**Habitat :** bois

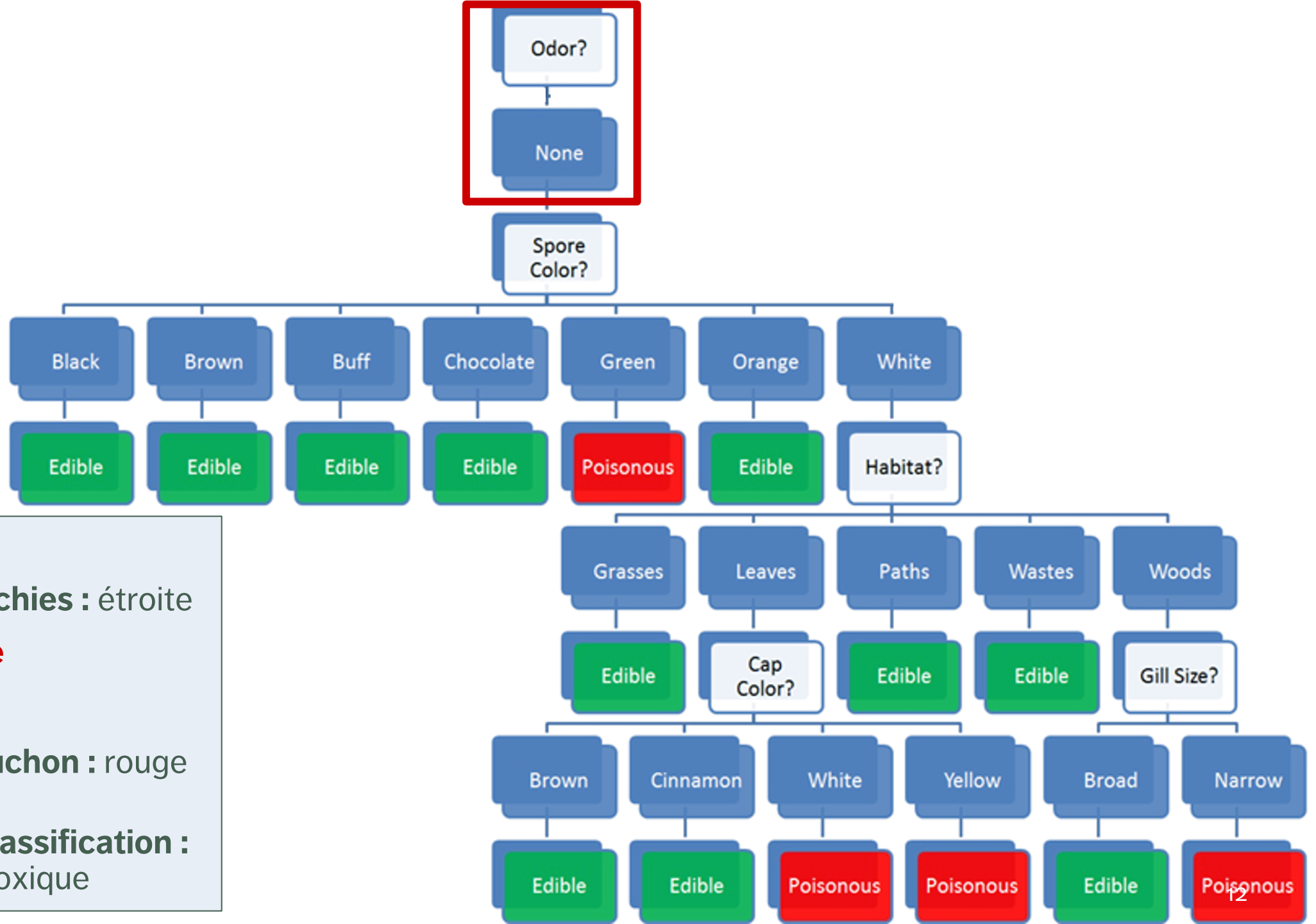
**Taille des branchies :** étroite

**Odeur :** aucune

**Spores :** blanc

**Couleur du bouchon :** rouge

**Problème de classification :**  
**comestible** ou **toxique**



**Habitat :** bois

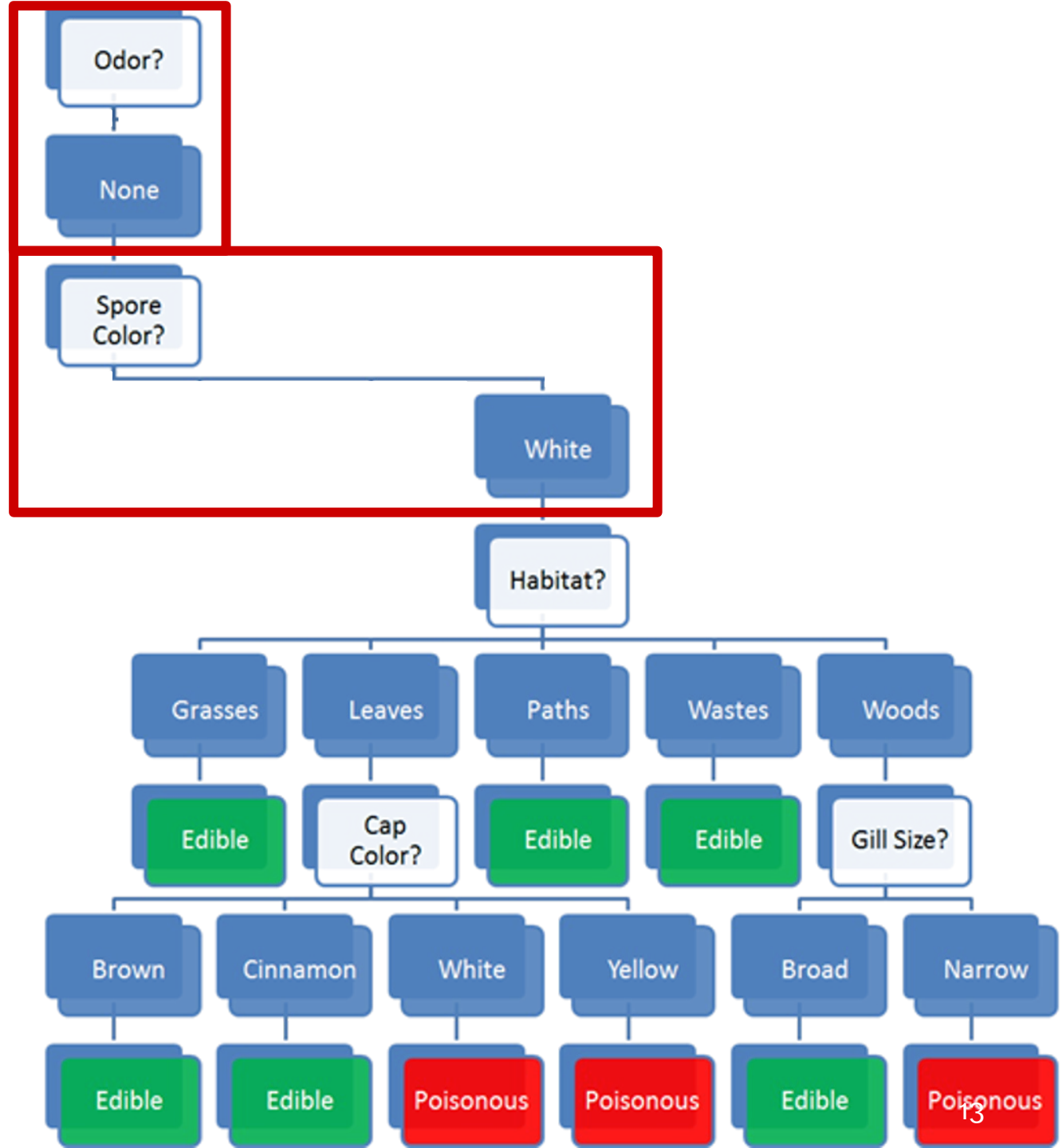
**Taille des branchies :** étroite

**Odeur :** aucune

**Spores :** blanc

**Couleur du bouchon :** rouge

**Problème de classification :**  
comestible ou toxique



**Habitat :** bois

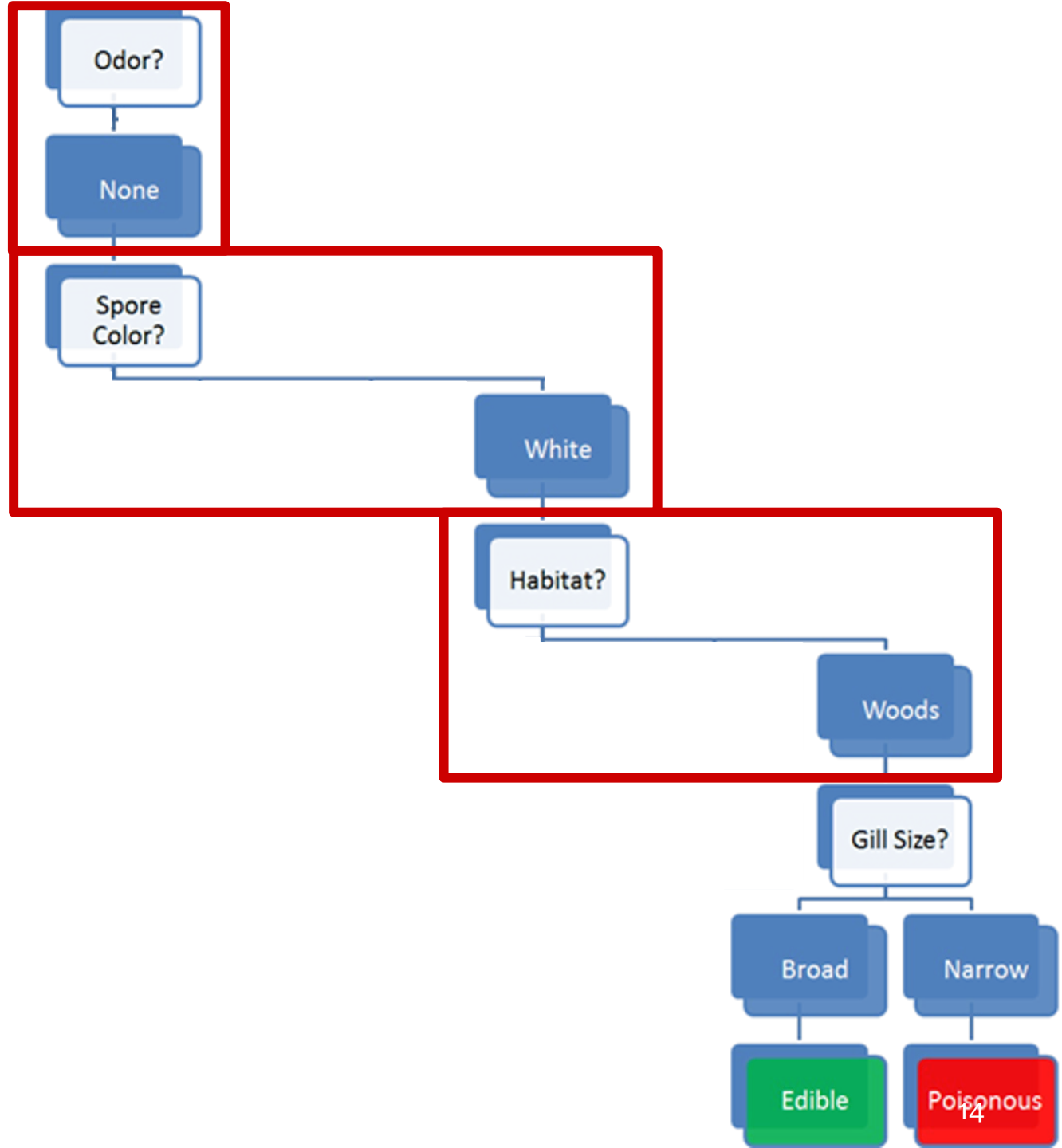
**Taille des branchies :** étroite

**Odeur :** aucune

**Spores :** **blanc**

**Couleur du bouchon :** rouge

**Problème de classification :** comestible ou toxique



**Habitat : bois**

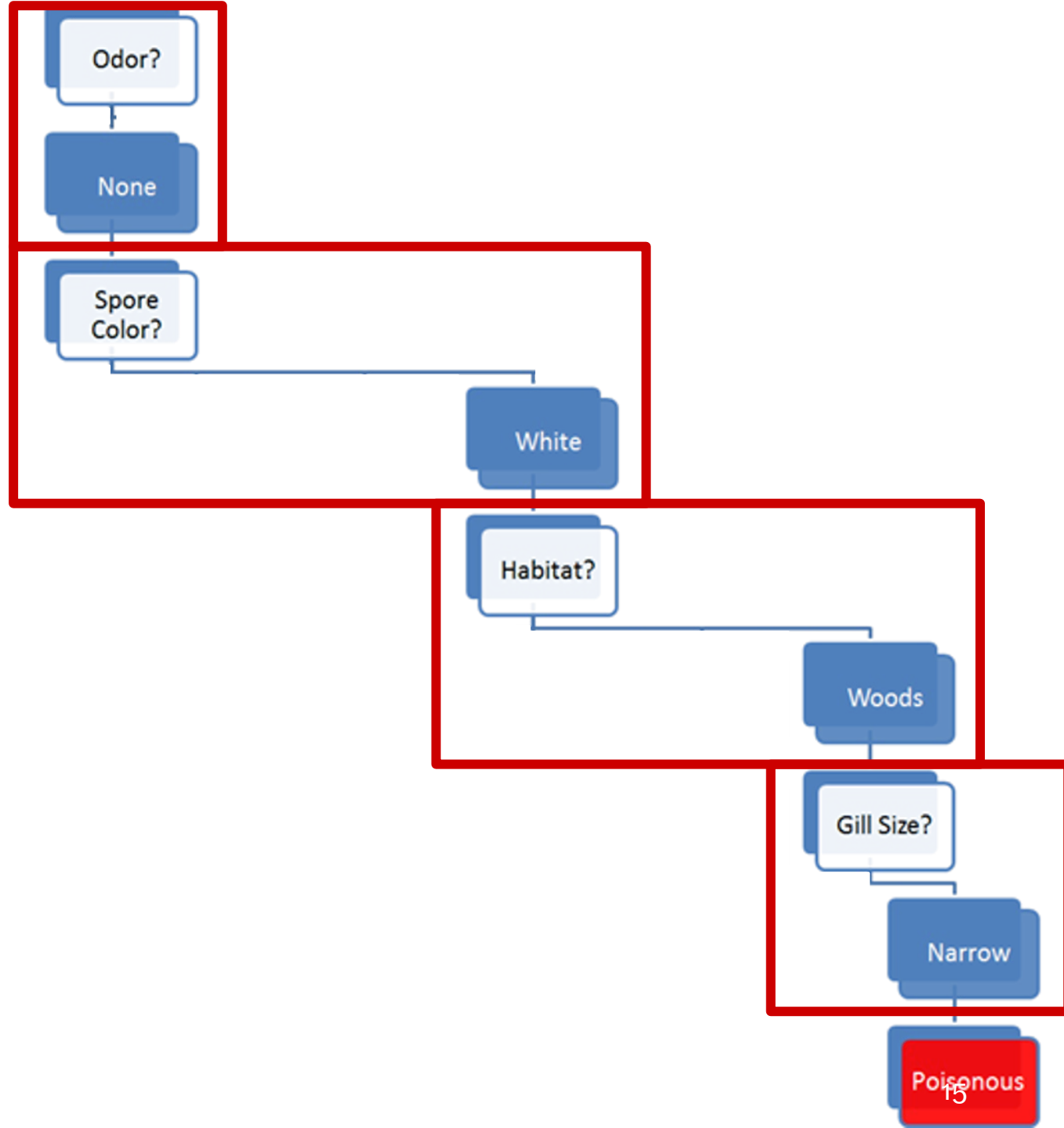
**Taille des branchies : étroite**

**Odeur : aucune**

**Spores : blanc**

**Couleur du bouchon : rouge**

**Problème de classification :  
comestible ou toxique**



**Habitat :** bois

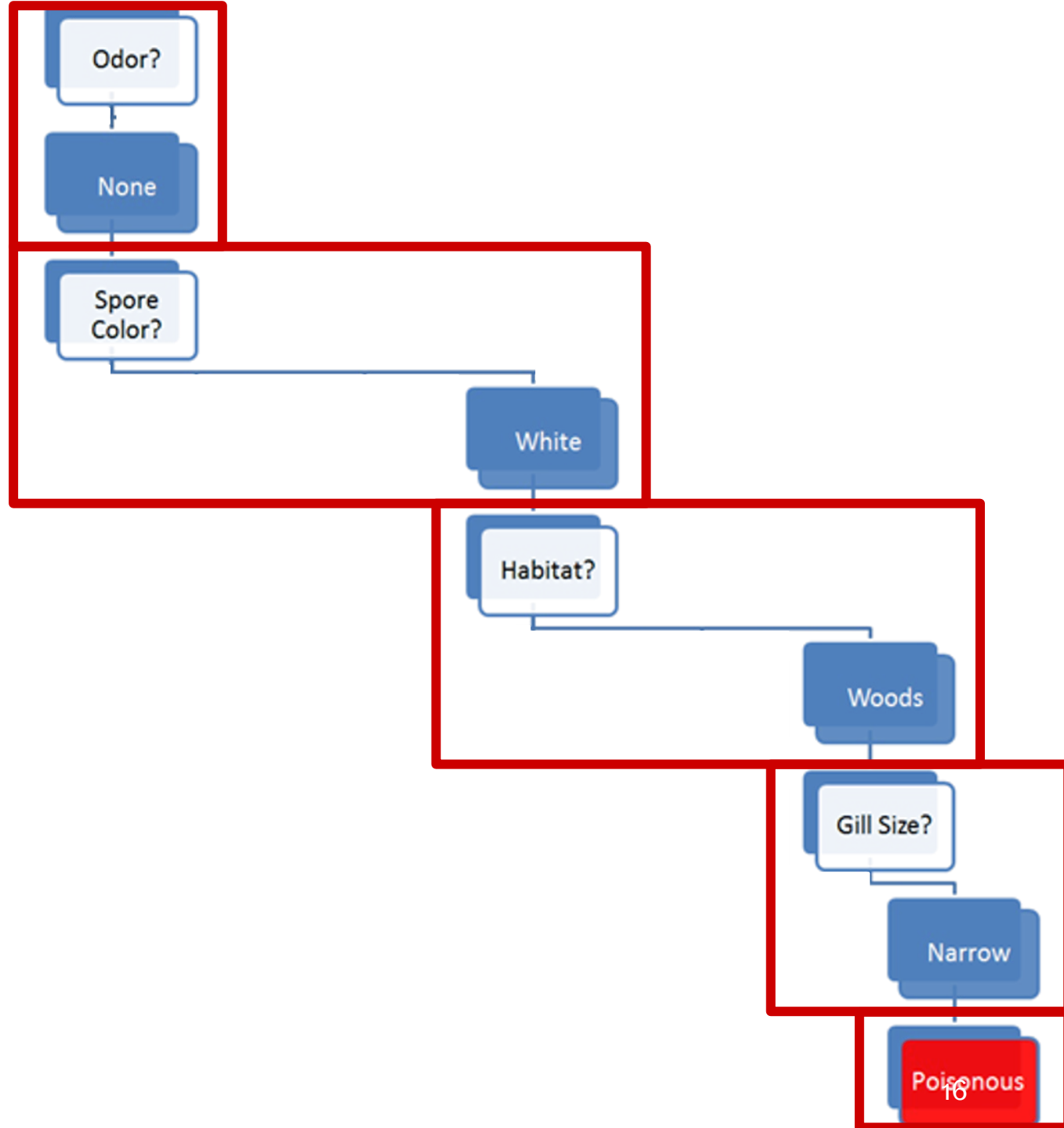
**Taille des branchies :** étroite

**Odeur :** aucune

**Spores :** blanc

**Couleur du bouchon :** rouge

**Problème de classification :**  
comestible ou toxique



**Habitat :** bois

**Taille des branchies :** étroite

**Odeur :** aucune

**Spores :** blanc

**Couleur du bouchon :** rouge

**Problème de classification :**  
comestible ou **toxique**

# Discussion

---

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

D'où vient le modèle ?

Qu'auriez-vous besoin de savoir pour faire confiance au modèle ?

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



# Cours sœurs

---

## LA VISUALISATION DES DONNÉES ET LES TABLEAUX DE BORD

1. Concepts fondamentaux
2. Les tableaux de bord
3. La mise en récit de données
4. Les visualisations avec ggplot2

## INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE

1. Apprentissage statistique
2. Classification
3. Regroupement
4. Problèmes et défis