Data Science Essentials

P. BOILY | UNIVERSITY OF OTTAWA | FACULTY OF SCIENCE | DEPARTMENT OF MATHEMATICS AND STATISTICS
DATA ACTION LAB | IDLEWYLD ANALYTICS

WITH CONTRIBUTIONS FROM **J. SCHELLINCK** | SYSABEE | DATA ACTION LAB

Instructor – Patrick Boily

Employment

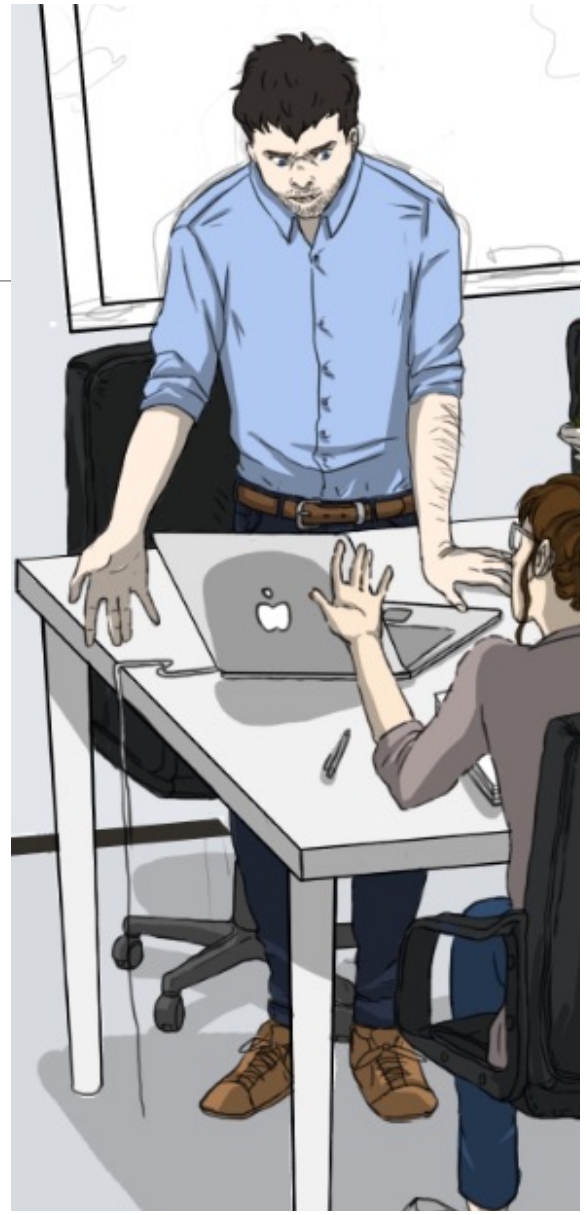
- Professor Math/Stat [‘19 – now, uOttawa]
- President [‘16 – now, Idlewyld Analytics]
- Manager and Senior Consultant [’12 – ‘19, CQADS, Carleton]
- Public Service [’08 – ’12, ASFC | StatCan | TC | TPSGC]
- 60+ uni course; 250+ workshop days

Projects

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

Specialization

- Data visualization; data cleaning (... unfortunately)
- Application of wide breadth of techniques to all kinds of data
- Mathematical/statistical modeling



Course Material

Course Webpage:

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

Course Notes:

https://idlewyldanalytics.com

Contact Info:

pboily@uottawa.ca

Slack Workspace:

https://dspdi.slack.com

Course Description

This course gives participants the opportunity to master foundational knowledge and skills needed for data analysis, along with a discussion of common challenges and pitfalls.

Participants will be introduced to various methods of data preparation, and to some intrinsic limitations of data and data analysis, and to easily avoidable pre-analysis mistakes.

Following the course, the participants have the option of working on a guided project, getting feedback from the instructor.

Additional Information

Exposure to programming frameworks would be beneficial but not necessary. Participants must be comfortable (not necessarily experts) with the concepts introduced in a Probability and Statistics university-level course.

Participants are required to bring a laptop/personal computer on which the current version of R/RStudio (Posit) are installed (for which they may require administrative authorisation to install packages).

Participants doing the guided project must be familiar with R and/or Python.

Learning Outcomes

At the end of this course, participants will be able to:

- select appropriate methods to prepare their data for analysis
- anticipate challenges and limitations inherent to data and desired analysis outcomes
- apply data cleaning strategies to their data
- conduct simple analyses
- build simple data science pipelines to provide actionable insights

Course Outline

Technical and Non-Technical Aspects of Data Work

1. Quantitative Skills
   - Software and Tools
   - Multiple I's Approach
   - Roles and Responsibilities
   - Analysis Cheat Sheet

Data Science Basics

2. Preliminaries
3. Conceptual Frameworks
4. Data Science Ethics
5. Analytics Workflows
6. Getting Insight From Data

Session 1
Session 2
Session 3
Session 4

Course Outline

Data Preparation

7. Data Quality and Data Wrangling
8. Missing Values
9. Anomalous Observations
10. Dimensionality & Data Transformations

Miscellanea

11. Data Engineering
12. Data Management

Session 1

Session 2

Session 3

Session 4

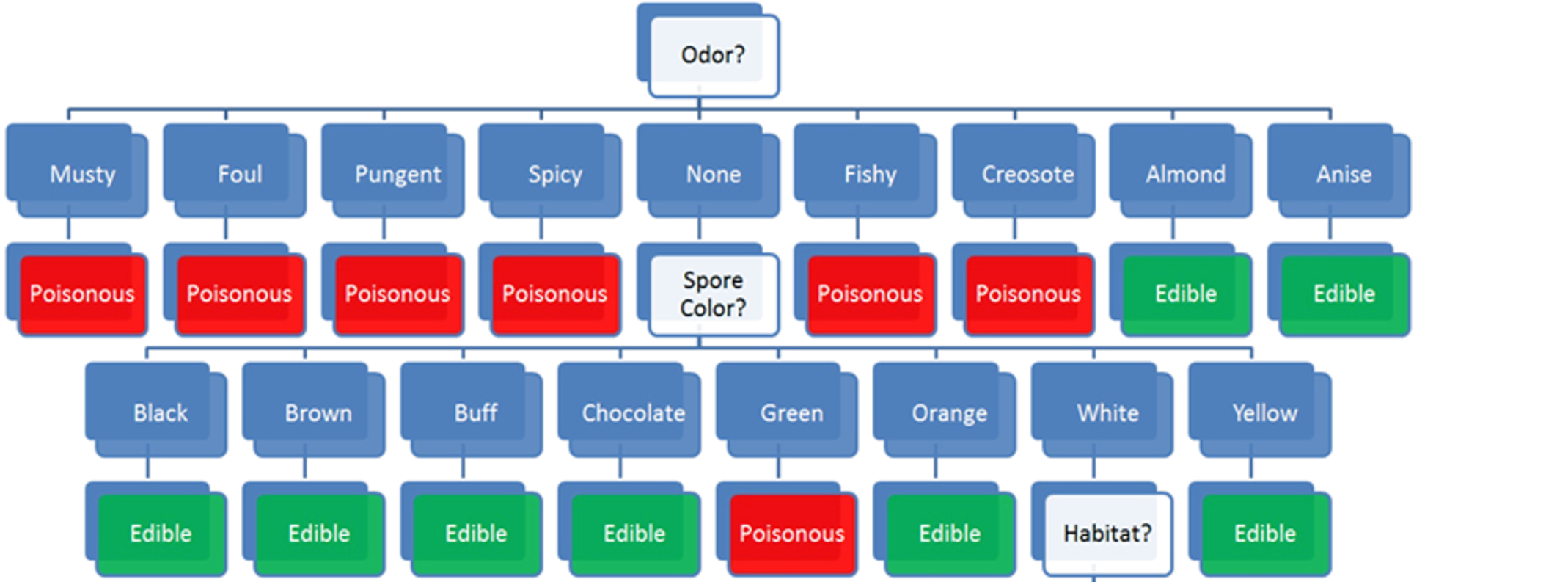
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

Poisonous Mushroom Problem

Amanita muscaria

Habitat: woods
Gill Size: narrow
Odor: none
Spores: white
Cap Colour: red

Classification problem:
Is Amanita muscaria edible, or poisonous?



Habitat: woods

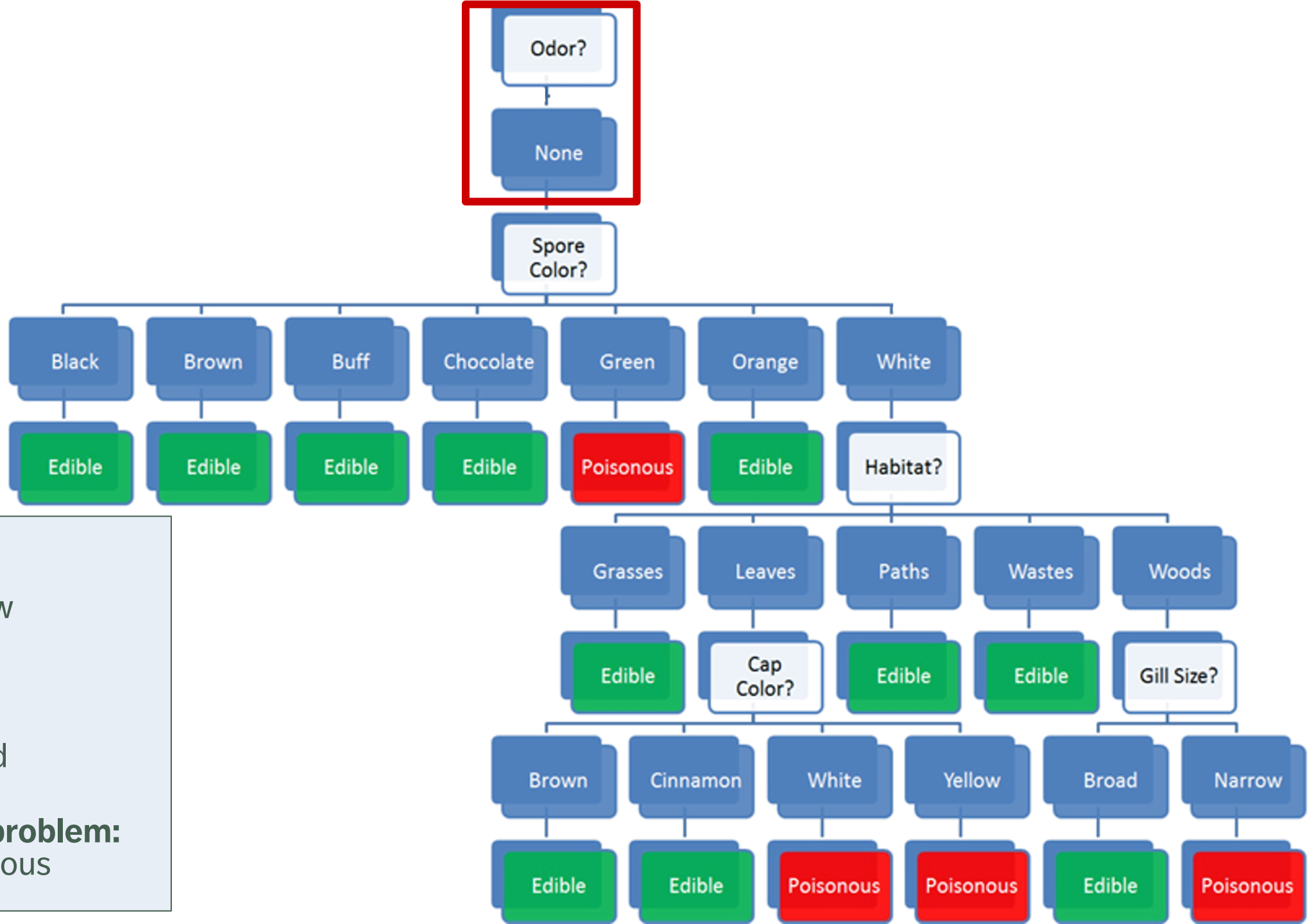
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or **poisonous**



Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem: edible or poisonous

Habitat: woods

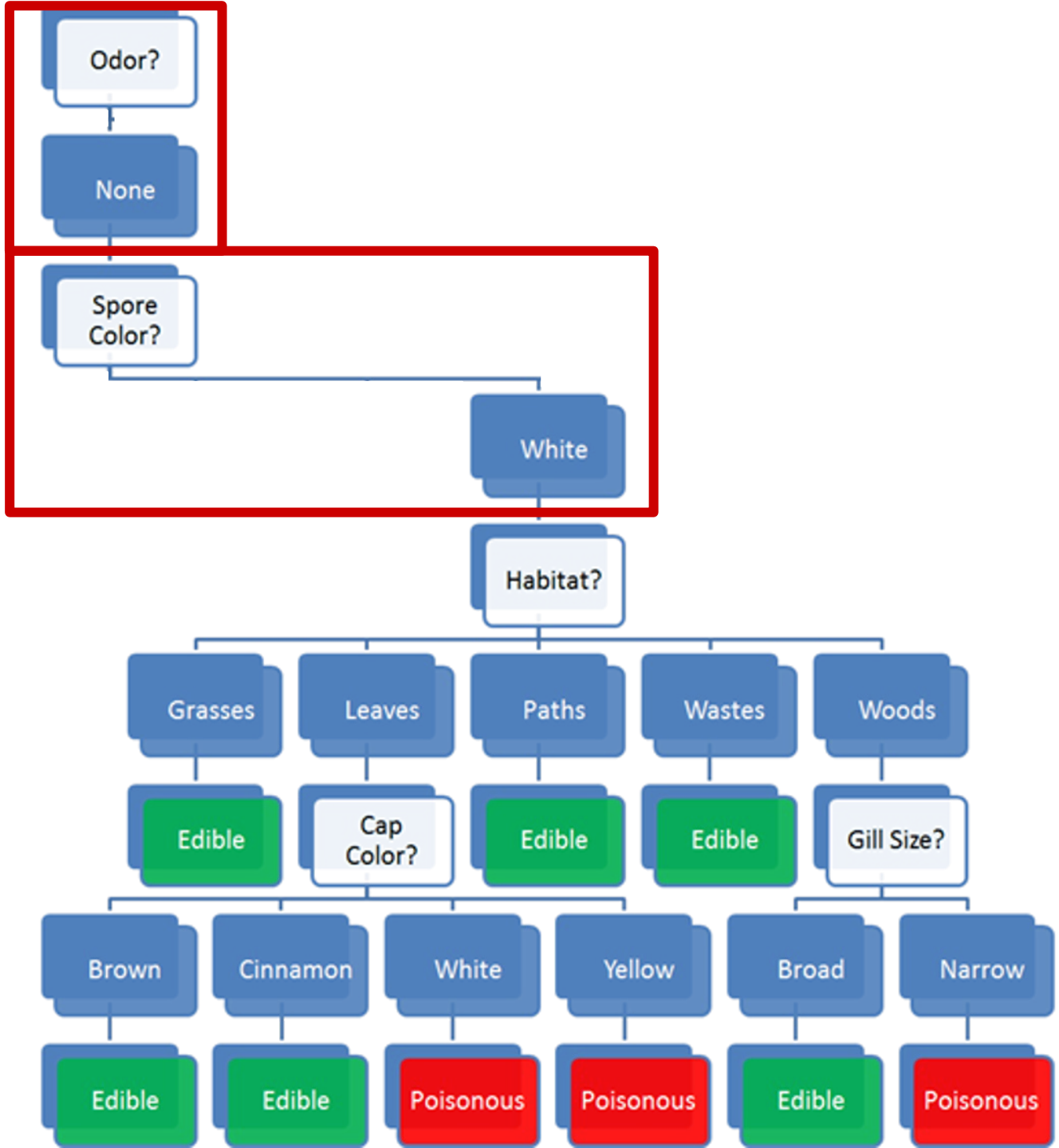
Gill Size: narrow

Odor: none

Spores: **white**

Cap Colour: red

Classification problem:
edible or poisonous



Habitat: woods

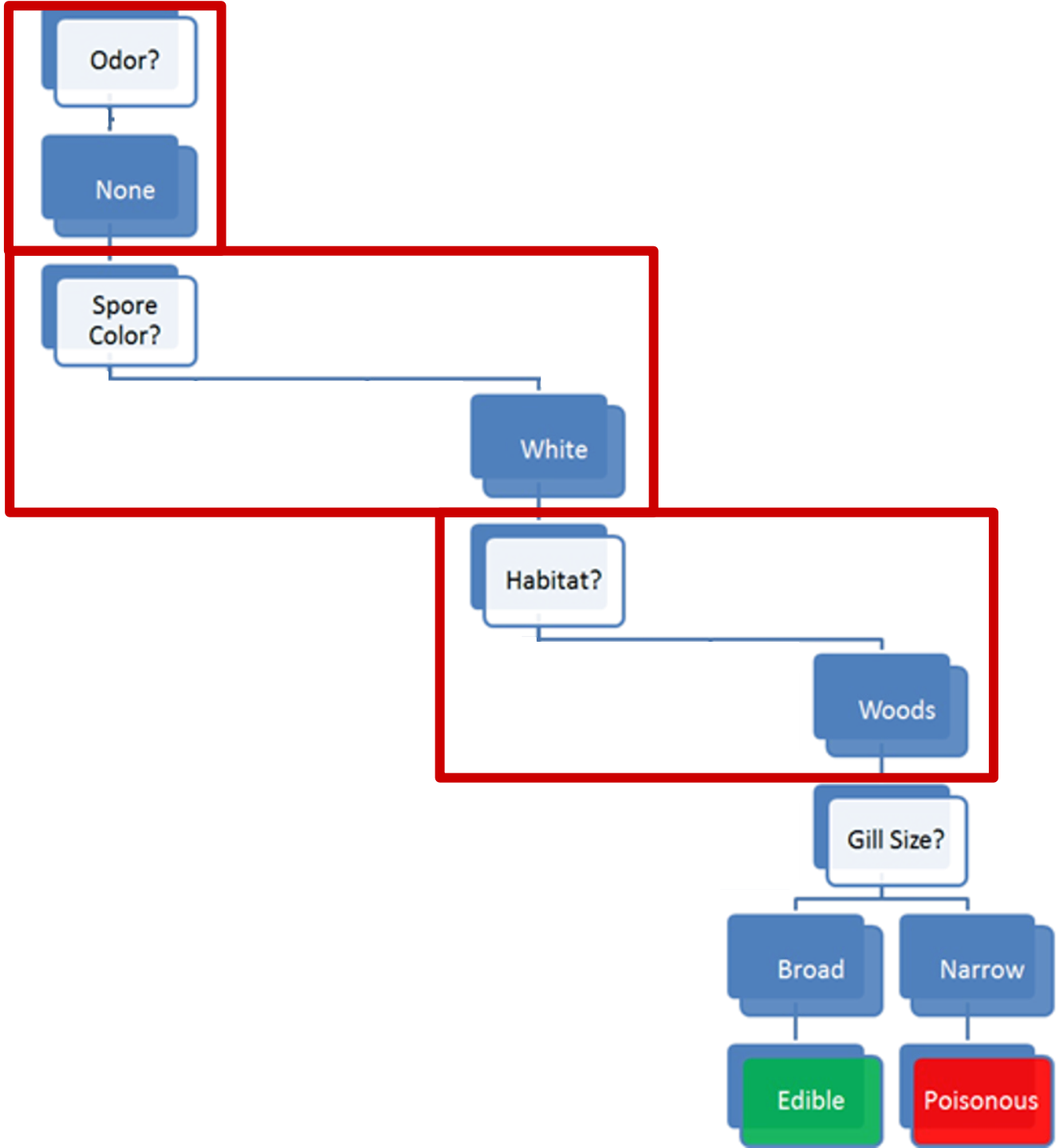
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous



Habitat: woods

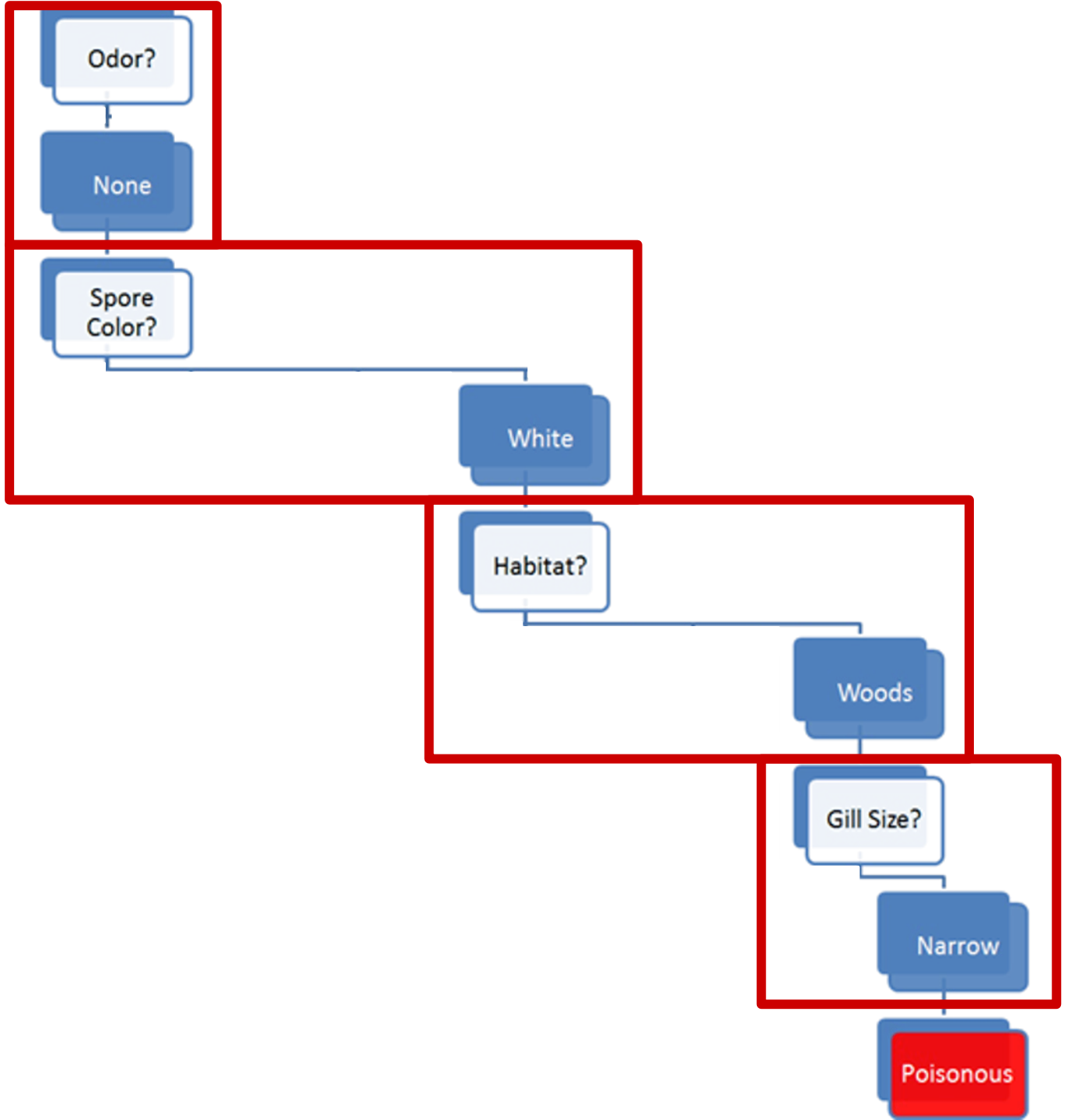
Gill Size: narrow

Odor: **none**

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous



Habitat: woods

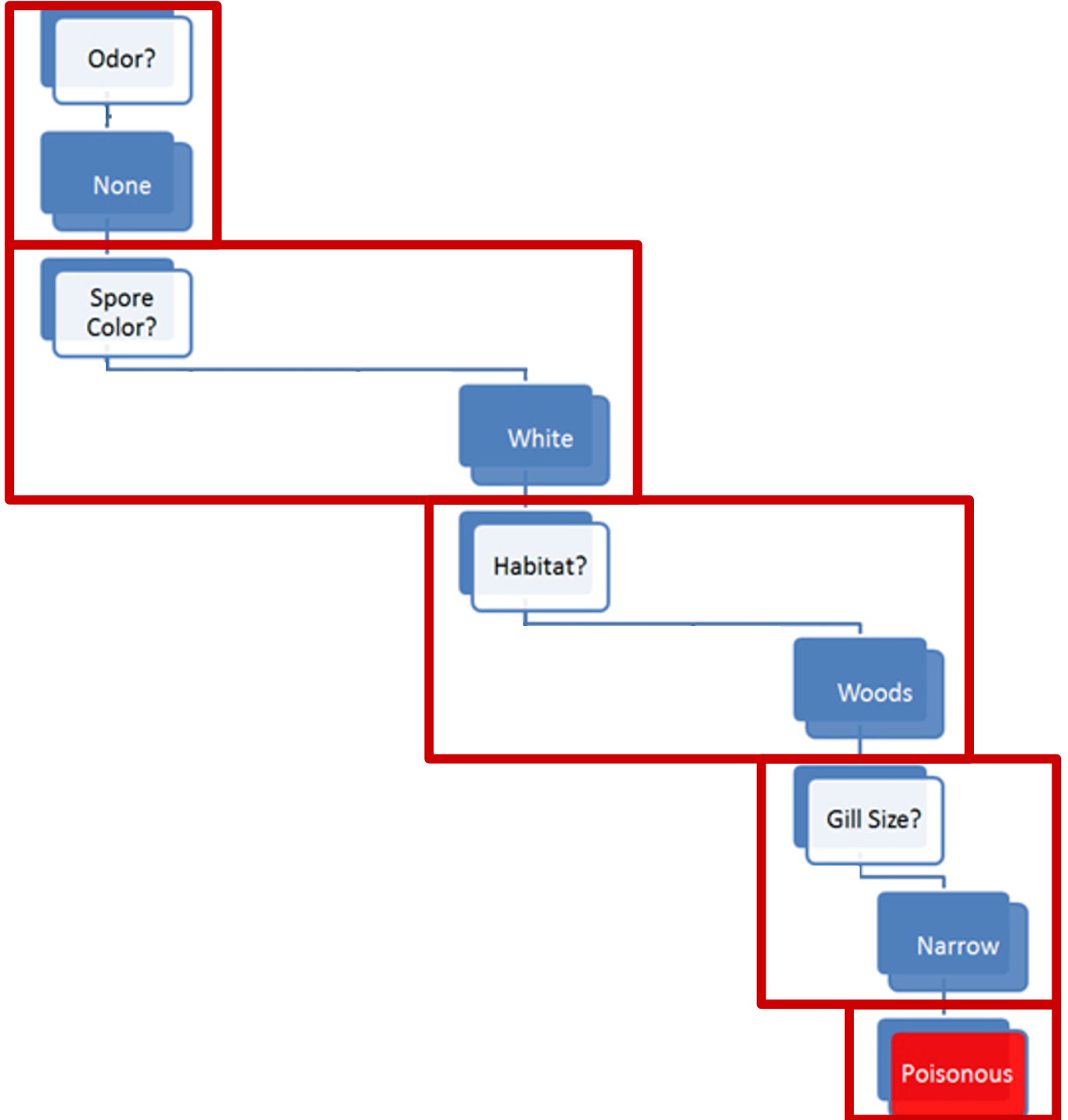
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or **poisonous**



Discussion

Would you have trusted an “**edible**” prediction?

Where is the model coming from?

What would you need to know to trust the model?

What’s the cost of making a classification mistake, in this case?