Session 1

Statistical Learning

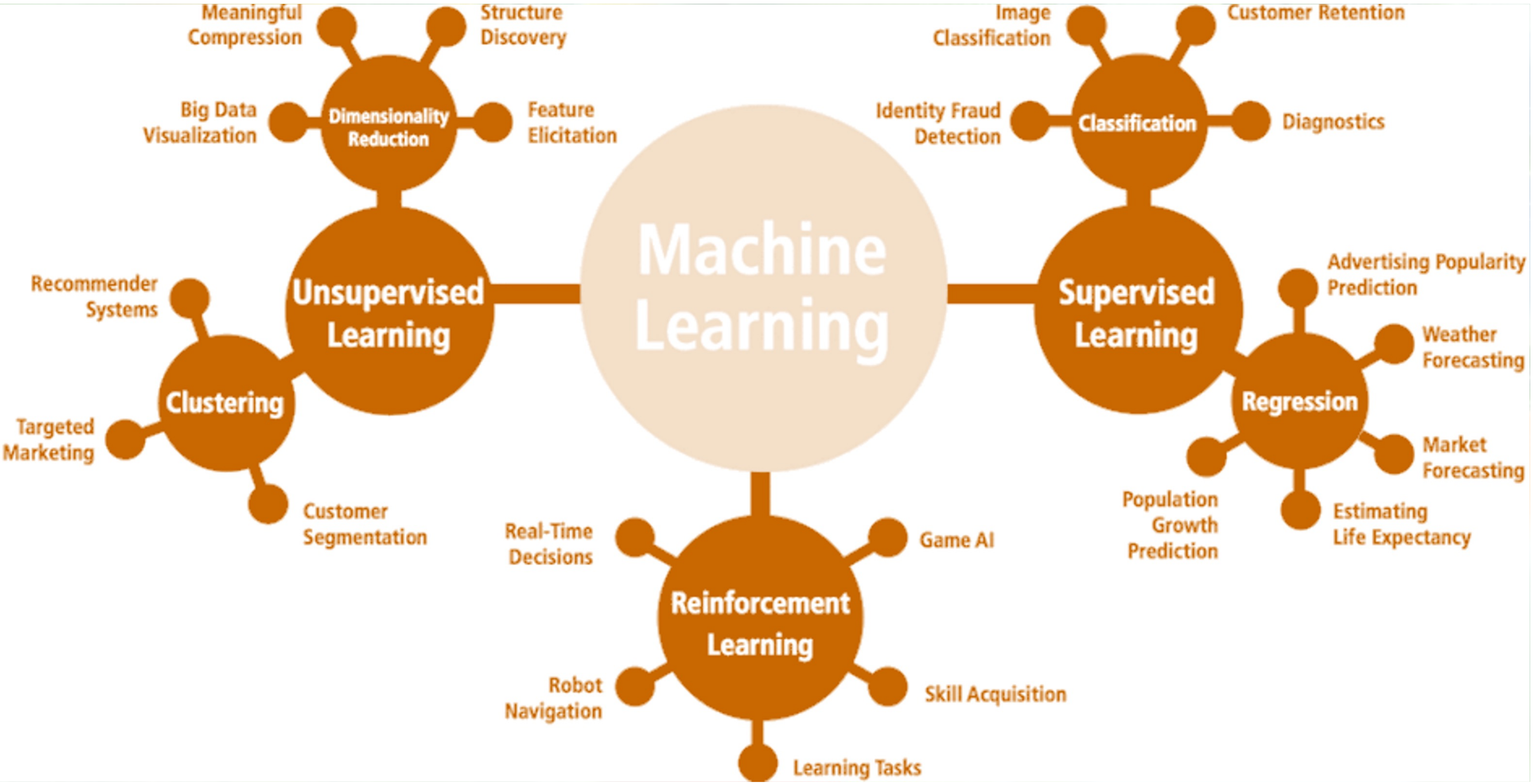
INTRODUCTION TO MACHINE LEARNING

We learn from failure, not from success!

[B. Stoker, Dracula]

Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom.

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



1. Types of Learning and Machine Learning Tasks

A Modern Challenge

One challenge of working in the **data science** (DS), **machine learning** (ML), and **artificial intelligence** (AI) fields : most quantitative work can be described as DS/ML/AI (this is often stretched to a ridiculous extent).

DS/ML/AI consists of quantitative processes (what H. Mason has called “the **working intersection** of statistics, engineering, computer science, domain expertise, and”hacking”) that help users **learn actionable insights** about their situation without completely abdicating their decision-making responsibility.

A Hierarchical Framework

Robinson suggests an "**inclusive hierarchical structure**":

1. in a first stage, DS provides "**insights**" *via* visualization and (manual) inferential analysis;
2. in a second stage, ML yields "**predictions**" (or "**advice**"), while reducing (not eliminating) the operator's analytical, inferential and decisional workload;
3. in the final stage, AI **removes the need for oversight**, allowing for automatic "actions" to be taken by a mostly unattended system (general AI vs. augmented intelligence).

In practice, stakeholders should probably not seek to abdicate **all** of their agency in the decision-making process.

Learning

Humans learn (at all stages) by first **taking in their environment**, and then by:

- answering questions about it;
- testing hypotheses;
- creating concepts;
- making predictions;
- creating categories, and
- classifying and grouping its various objects and attributes.

Statistical/Machine Learning

The main goal of DS/ML/AI is to try to **teach** machines to extract insight from data, properly and efficiently, and free of biases and pre-conceived notions – in other words, **can (should?) we design algorithms that can learn?**

The simplest DS/ML/AI method is **exploring representative data** to:

- provide a summary through basic statistics – mean, mode, histograms, etc.;
- make its multi-dimensional structure evident through data data visualization; and
- look for consistency, considering what is in there and what is missing.

Supervised Learning

Supervised learning (SL) is akin to "**learning with a teacher**": students give an answer to each exam question based on what they learned from worked-out examples provided by the teacher/textbook; the teacher provides the correct answers and marks the exam questions using a key.

Typical tasks include:

- **classification**
- regression
- rankings
- recommendations

Supervised Learning

In SL, algorithms use **labeled training data** to build (or train) a **predictive model**; each algorithm's performance is evaluated using **test data** for which the label is known but not used in the prediction.

There are fixed **targets** against which to train the model (such as age categories, or plant species) – these categories/classes (and their number) are known **prior to the analysis**.

Unsupervised Learning

Unsupervised learning (UL) is akin to “**self-learning by grouping similar exercises together as a study guide**”: the teacher is not involved in the discovery process and students might end up with different groupings.

Typical tasks include:

- **clustering**
- **association rules discovery**
- link profiling
- anomaly detection

Unsupervised Learning

Unsupervised algorithms use **unlabeled data** to find **natural patterns** in the data; the drawback is that accuracy **cannot be evaluated** with the same degree of satisfaction.

In UL, we don't know what the target is, or even if there is one – we are simply looking for **natural groups/associations** in the data, such as:

- junior students who like literature, have longish hair, and know how to cook **vs.**
- students who are on a sports team and have siblings **vs.**
- financial professionals with a penchant for superhero movies, craft beer and Hello Kitty backpack **vs.** …

Other Learning Frameworks

Some DS/ML/AI techniques fit into both camps; others can be either SL or UL, but there are other **conceptual approaches** (usually for AI tasks):

- **semi-supervised learning** in which some data points have labels but most do not, which often occurs when acquiring data is costly ("the teacher provides worked-out examples and a list of unsolved problems to try out; the students try to find similar groups of unsolved problems and compare them with the solved problems to find close matches")
- **reinforcement learning**, where an agent attempts to collect as much (short-term) reward as possible while minimizing (long-term) regret ("embarking on a Ph.D. with an advisor... with all the highs and the lows and **maybe** a diploma at the end of the process?")

Statistical Learning/Machine Learning

The term “statistical learning” is not used frequently in practice (except by mathematicians and statisticians); the tendency is to speak instead of **machine learning**.

If a distinction must be made, it could be argued that:

- statistical learning arises from statistical-like models, and the emphasis is usually placed on **interpretability**, **precision**, and **uncertainty**, whereas
- machine learning arise from artificial intelligence studies, with emphasis on **large scale applications** and **prediction accuracy**.

The dividing line between the terms is blurry – the vocabulary used by practitioners mostly betrays their educational backgrounds.

DS/ML Questions

Outside of academia, DS/ML/AI methods are only really interesting when they help users ask and answer useful questions. Compare, for instance:

- **Analytics** – “How many clicks did this link get?”
- **Data Science** – “Based on the previous history of clicks on links of this publisher’s site, can I predict how many people from Manitoba will read this specific page in the next three hours?” or “Is there a relationship between the history of clicks on links and the number of people from Manitoba who will read this specific page?”
- **Quantitative Methods** – “We have no similar pages whose history could be consulted to make a prediction, but we have reasons to believe that the number of hits will be strongly correlated with the temperature in Winnipeg. Using the weather forecast over the next week, can we predict how many people will access the specific page during that period?”

DS/ML Questions

DS/ML models are **predictive/descriptive** (not **explanatory/prescriptive**): they show connections, and exploit correlations to make predictions, but they don’t reveal **why** such connections exist (Bayesian networks).

Quantitative methods, on the other hand, usually assume a certain level of **causal understanding** based on various **first principles**. That distinction is not always understood properly by analysts and stakeholders.

ML Tasks

Common ML tasks (with representative questions) include:

- **classification** – which undergraduates are likely to succeed at the graduate level?
- **probability estimation** – how likely is it that a given candidate will win an election?
- **value estimation** – how much is a given client going to spend at a restaurant?
- **similarity matching** – which prospective clients are most similar to established best clients?
- **clustering** – do signals from a sensor form natural groups?
- **association rules discovery** – what books are commonly purchased together online?
- **profiling** and **behaviour description** – what is the typical cell phone usage of a certain customer's segment?
- **link prediction** – J. and K. have 20 friends in common: perhaps they'd be great friends?

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

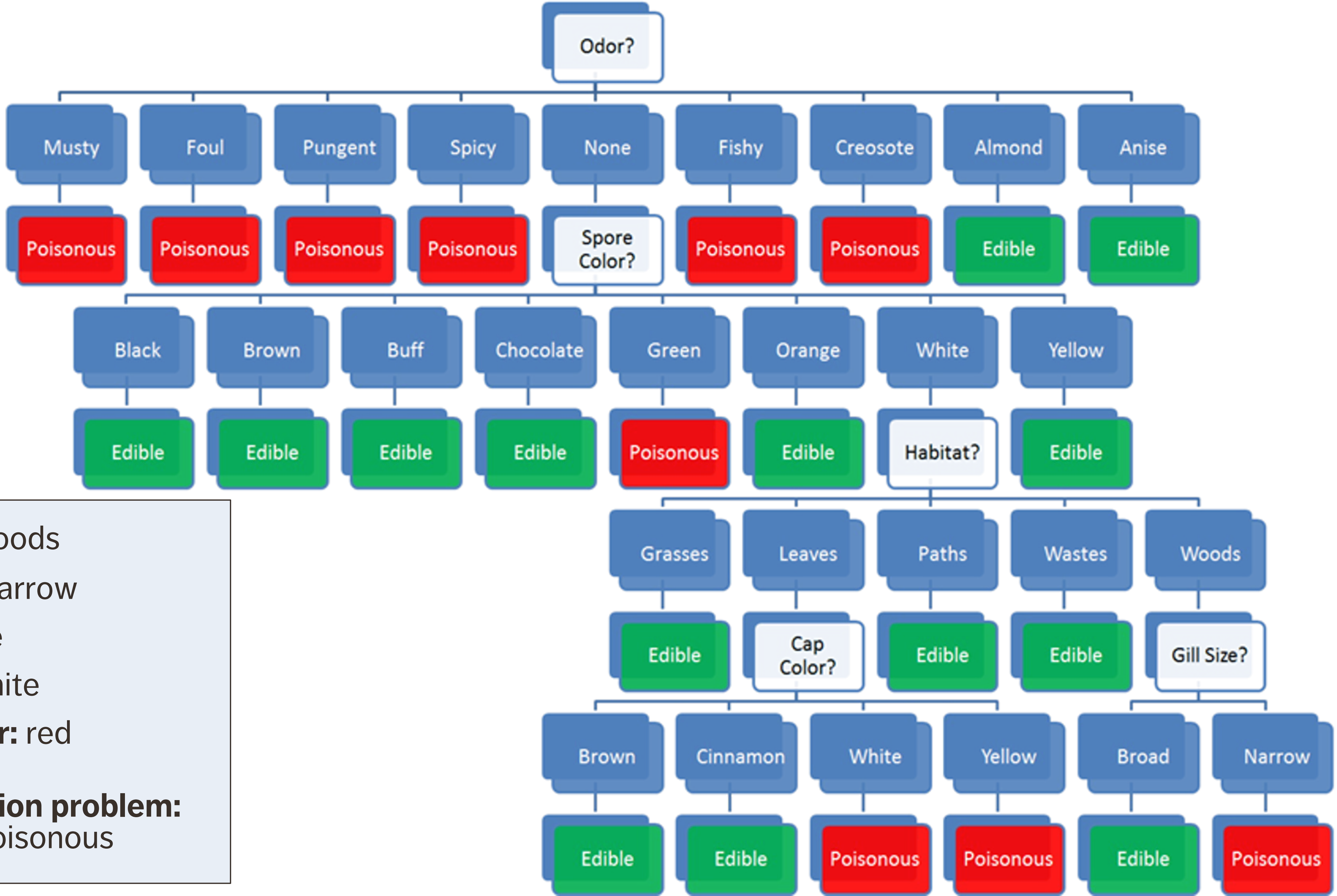
Mushroom Classification Problem

Amanita muscaria

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

Classification problem:
Is Amanita muscaria edible or poisonous?

Session 1



Odor?

Musty
Foul
Pungent
Spicy
None
Fishy
Creosote
Almond
Anise

Poisonous
Poisonous
Poisonous
Poisonous
Spore Color?
Poisonous
Poisonous
Edible
Edible

Black
Brown
Buff
Chocolate
Green
Orange
White
Yellow

Edible
Edible
Edible
Edible
Poisonous
Edible
Habitat?
Edible

Grasses
Leaves
Paths
Wastes
Woods

Edible
Cap Color?
Edible
Edible
Gill Size?

Brown
Cinnamon
White
Yellow
Broad
Narrow

Edible
Edible
Poisonous
Poisonous
Edible
Poisonous

Habitat: woods

Gill Size: narrow

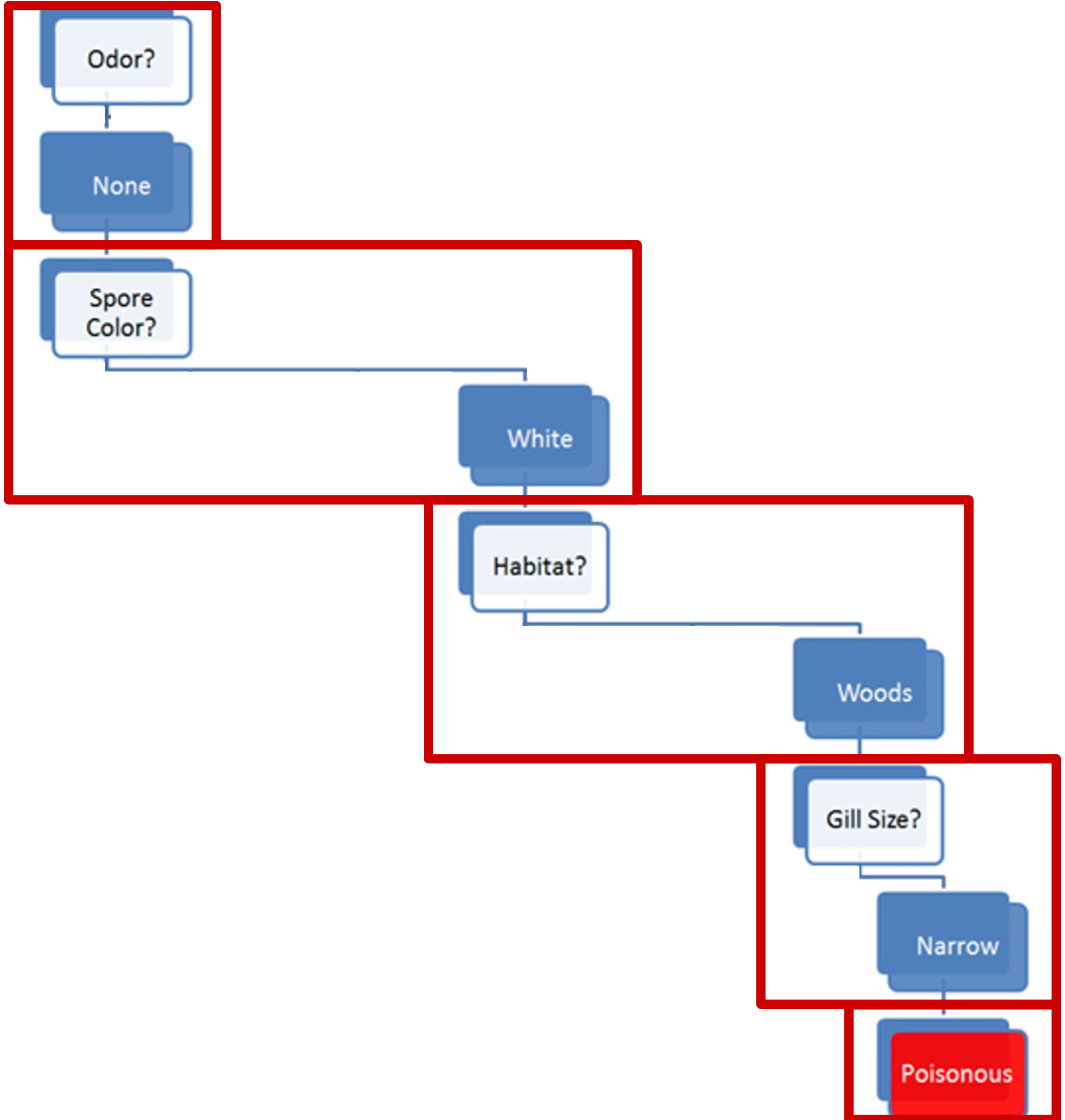
Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous

Session 1



Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem: edible or **poisonous**

Mushroom Classification Problem

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?

Suggested Reading

Types of Learning and Machine Learning Tasks

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

Exercises

Types of Learning and Machine Learning Tasks

1. Of what types of machine learning tasks are the following problems representative?
   - Identifying risk factors associated to breast/prostate cancer.
   - Predicting whether a patient will have a second, fatal heart attack within 30 days of the first on the basis of demographics, diet, clinical measurements, etc.
   - Establishing the relationship between salary and demographic information in population survey data.
   - Predicting the yearly inflation rate using various indicators.

2. What are some examples of supervised, unsupervised, semi-supervised, reinforcement machine learning tasks in the business world? In a public policy/government setting? In a scientific setting?

Exercises

Types of Learning and Machine Learning Tasks

3. Assuming that DS/ML/AI techniques are used in the following cases, identify whether the required task falls under SL or UL.
    a. Estimating the repair time required for an aircraft based on a trouble ticket.
    b. Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).
    c. In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying pattern in prior transactions.
    d. Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets with a known threat status.
    e. Identifying segments of similar customers.

Exercises

Types of Learning and Machine Learning Tasks

3. Assuming that DS/ML/AI techniques are used in the following cases, identify whether the required task falls under SL or UL.

   f. Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and non-bankrupt firms.

   g. Automated sorting of mail by zip code scanning.

   h. It is more difficult and expensive to win new customers than it is to retain existing customers. Scoring each customer on their likelihood to quit can help an organization design effective interventions, such as discounts or free services, to retain profitable customers in a cost-effective manner.

   i. Some medical practitioners conduct unnecessary tests and/or over-bill their government or insurance companies. Using audit data, it may be possible to identify such providers and take appropriate action.

Exercises

Types of Learning and Machine Learning Tasks

3. Assuming that DS/ML/AI techniques are used in the following cases, identify whether the required task falls under SL or UL.

   j. A market basket analysis can help develop predictive models to determine which products often sell together. This knowledge of affinities between products can help retailers create promotional bundles to push non-selling items along a set of products that sell well.

   k. Diagnosing the cause of a medical condition is the crucial first step in medical engagement. In addition to the current condition, other factors can be considered, including the patient’s health history, medication history, family’s history, and other environmental factors. A predictive model can absorb all of the information available to date (for this patient and others) and make probabilistic diagnoses, in the form of a decision tree, taking away most of the guess work involved.

   l. Schools can develop models to identify students who are at risk of not returning to school. Such students can be flagged to be on the receiving end of potential corrective measures.

Exercises

Types of Learning and Machine Learning Tasks

3. Assuming that DS/ML/AI techniques are used in the following cases, identify whether the required task falls under SL or UL.

   m. In addition to customer data, telecom companies also store call detail records (CDR), which precisely describe the calling behaviour of each customer. The unique data can be used to profile customers, who may be marketed to based on the similarity of their CDR to other customers'.

   n. Statistically, all equipment is likely to break down at some point in time. Predicting which machine is likely to shut down is a complex process. Decision models to forecast machinery failure could be constructed using past data, which can lead to savings provided by preventative maintenance.

   o. Identifying which tweets contain disinformation and which tweets are legitimate.