Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction to Data Analysis

DATA INSIGHT FUNDAMENTALS

Patrick Boily

Data Action Lab | uOttawa | Idlewyld Analytics

pboily@uottawa.ca

[with files from Jen Schellinck | Sysabee]

“Reports that say that something hasn't happened are always interesting to me, because as we know, there are **known knowns**; there are things we know that we know. There are **known unknowns**; that is to say, there are things that we now know we don't know. But there are also **unknown unknowns** – there are things we do not know we don't know.”

Donald Rumsfeld, US Department of Defense News Briefing, 2002

ANALYSIS PLANNING

DATA INSIGHT FUNDAMENTALS

"Plans are nothing. Planning is everything."

Dwight D. Eisenhower

[CDC, Creating an Analysis Plan]

ANALYSIS PLAN OVERVIEW

Formulate research questions/hypotheses

Identify necessary (and available) datasets

Establish inclusion/exclusion criteria for records/observations

Select variables for use in the analyses

Chose statistical methods and software

DATA 101: BASIC DATA CONCEPTS

DATA INSIGHT FUNDAMENTALS

"You can have data without information, but you cannot have information without data."

Daniel Keys Moran (attributed)

WHAT IS DATA?

4,529 ‘red’ 25.782 ‘Y’

OBJECTS AND ATTRIBUTES

Object: apple

Shape: spherical

Colour: red

Function: food

Location: fridge

Owner: Jen

Remember: a person or an object is not simply the sum of its attributes!

FROM ATTRIBUTES TO DATASETS

Attributes are **fields** (columns) in a database; objects are **instances** (rows).

Objects are described by their **feature vector**, the collection of attributes associated with value(s) of interest.

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
…	…	…	…	…	…

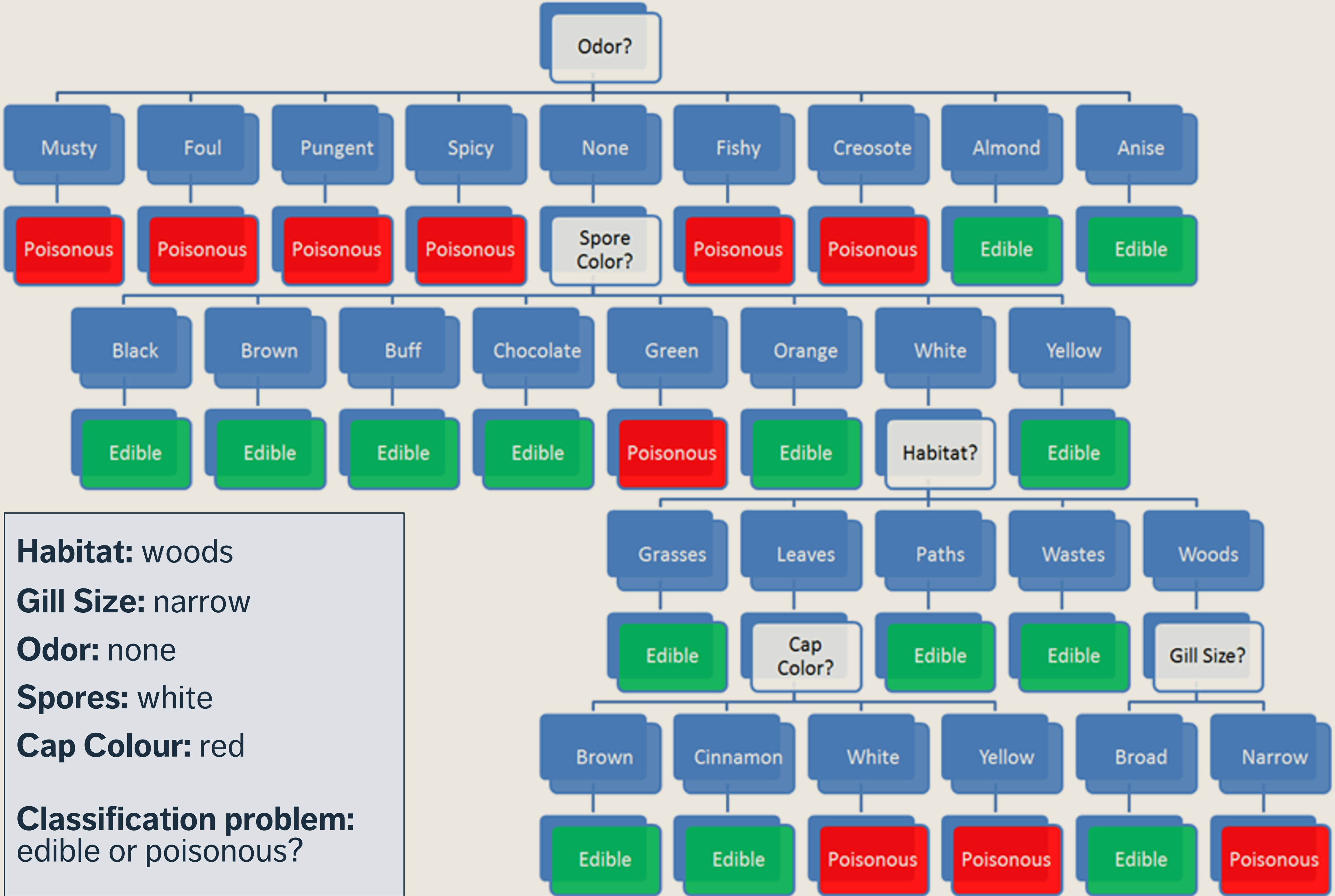
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

POISONOUS MUSHROOM DATASET

Amanita muscaria

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

Classification problem:
Is *Amanita muscaria* edible, or poisonous?



Habitat: woods

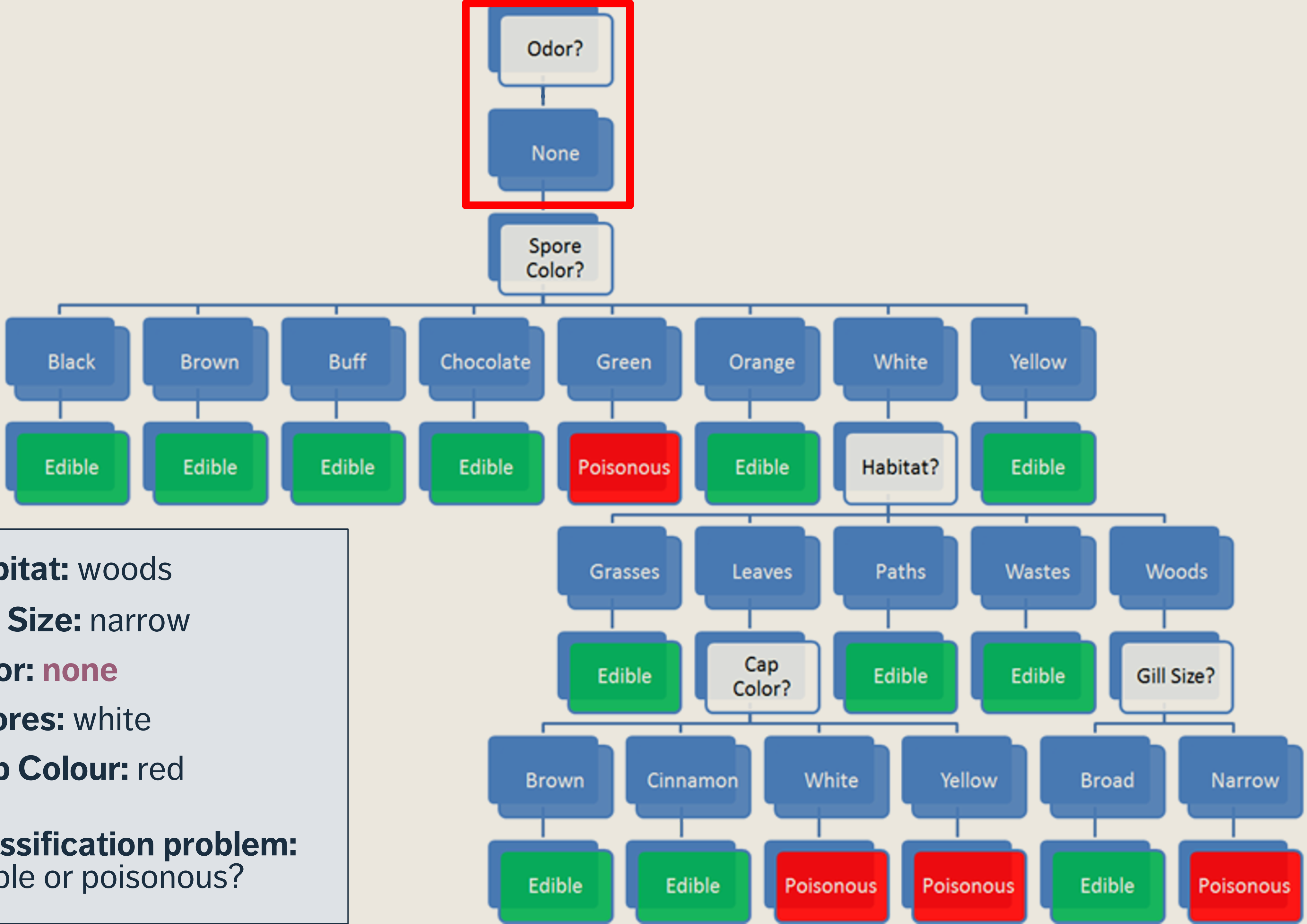
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous?



Habitat: woods

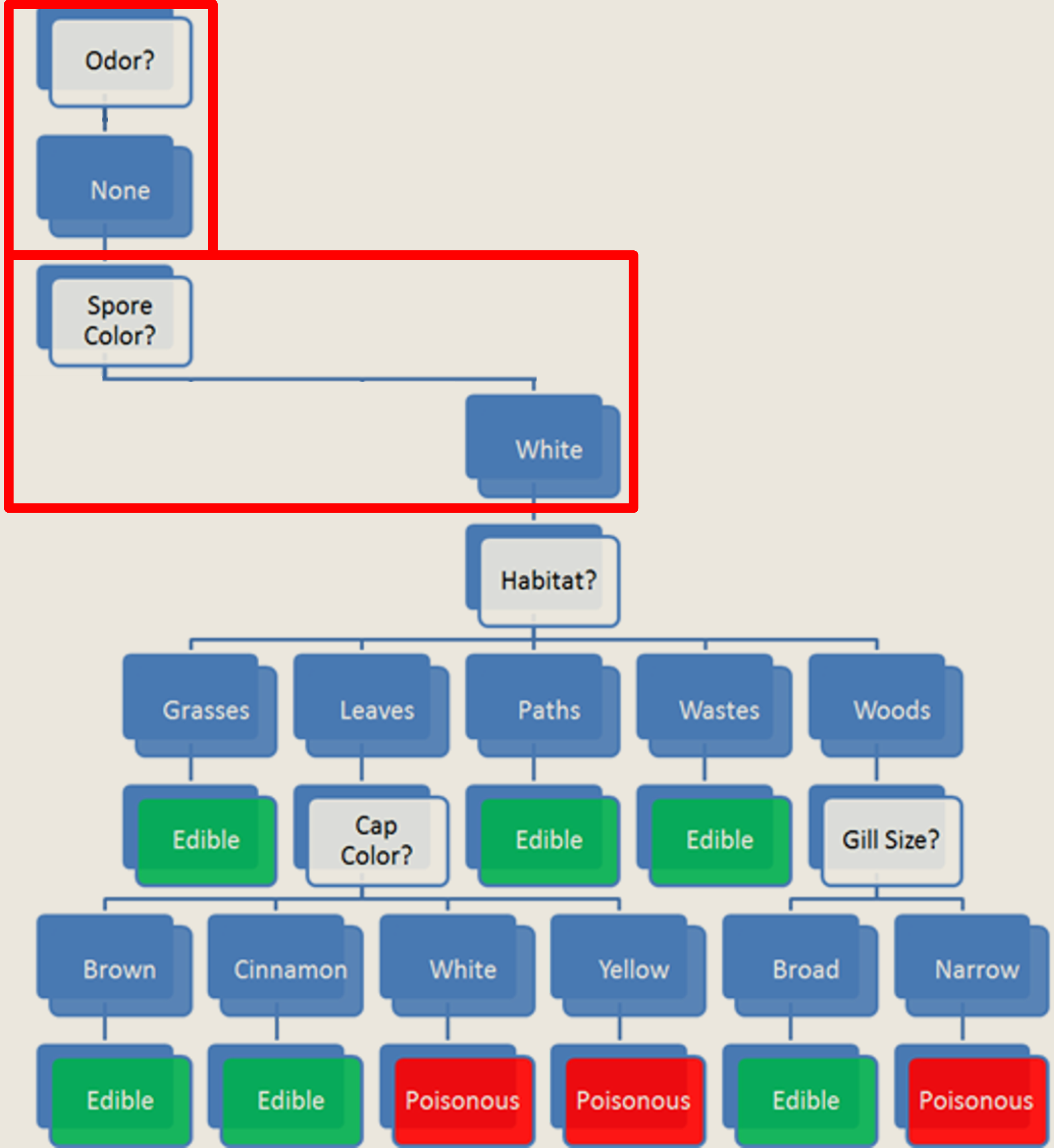
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous?



Habitat: woods

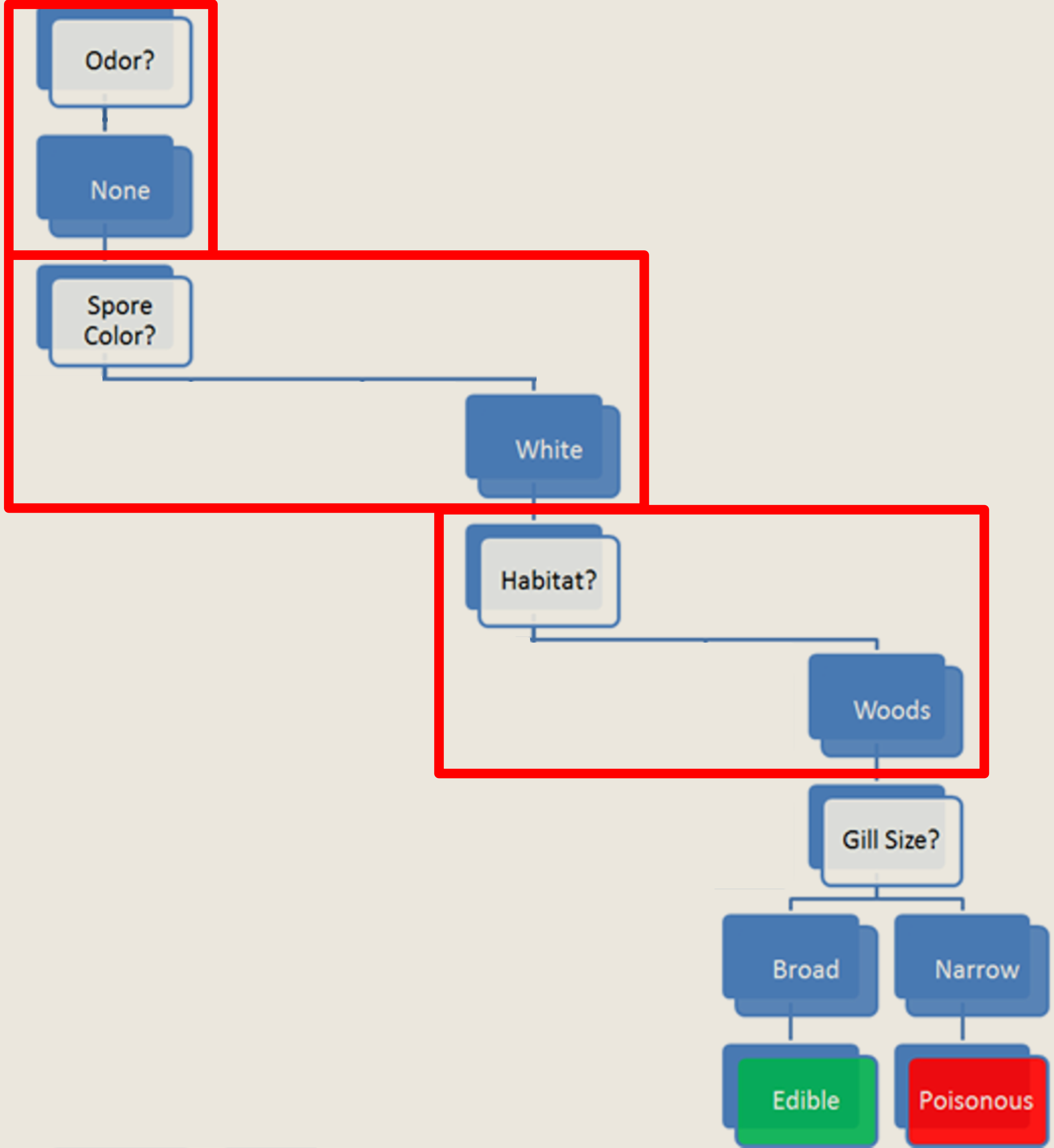
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous?



Habitat: woods

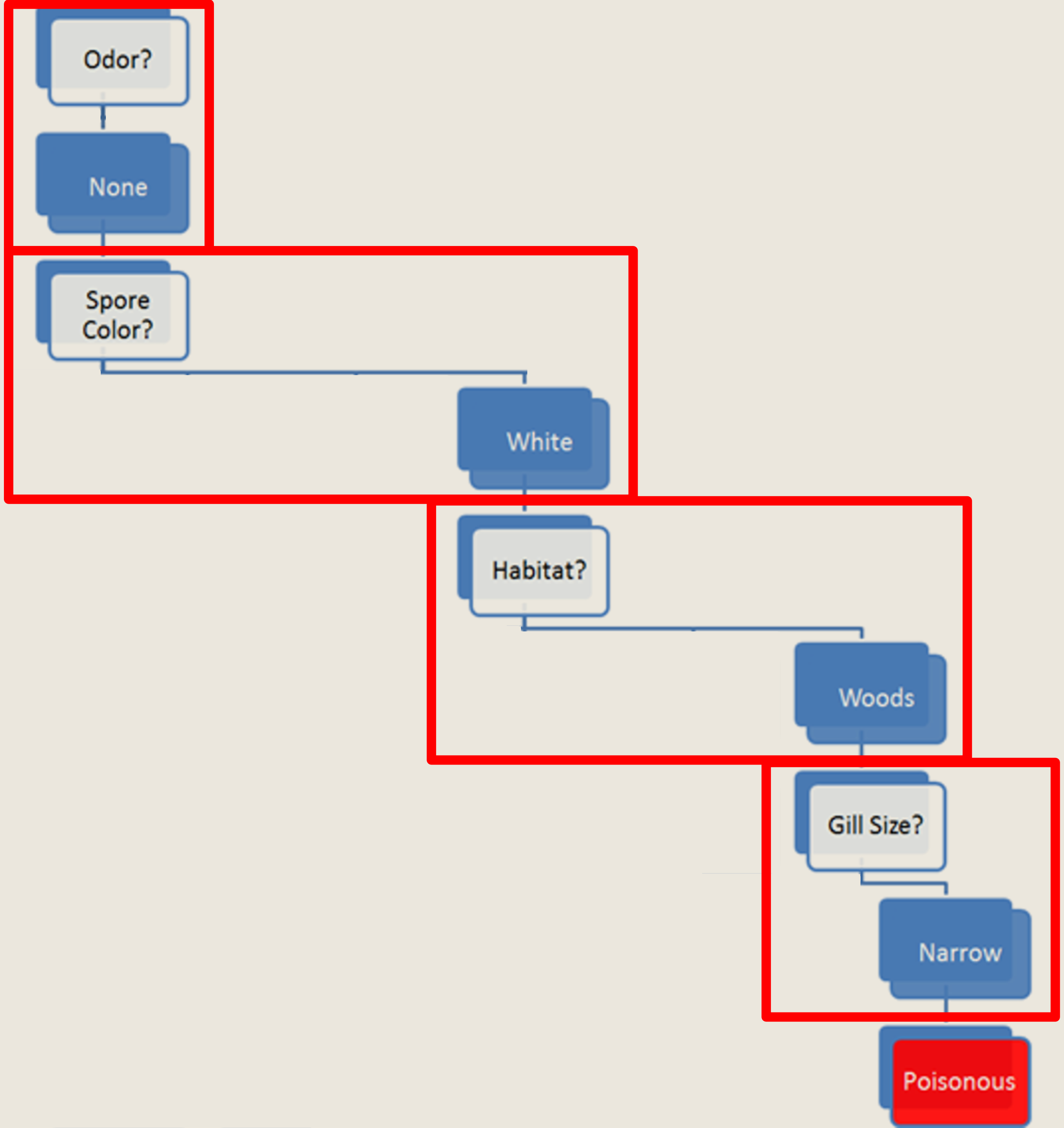
Gill Size: narrow

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

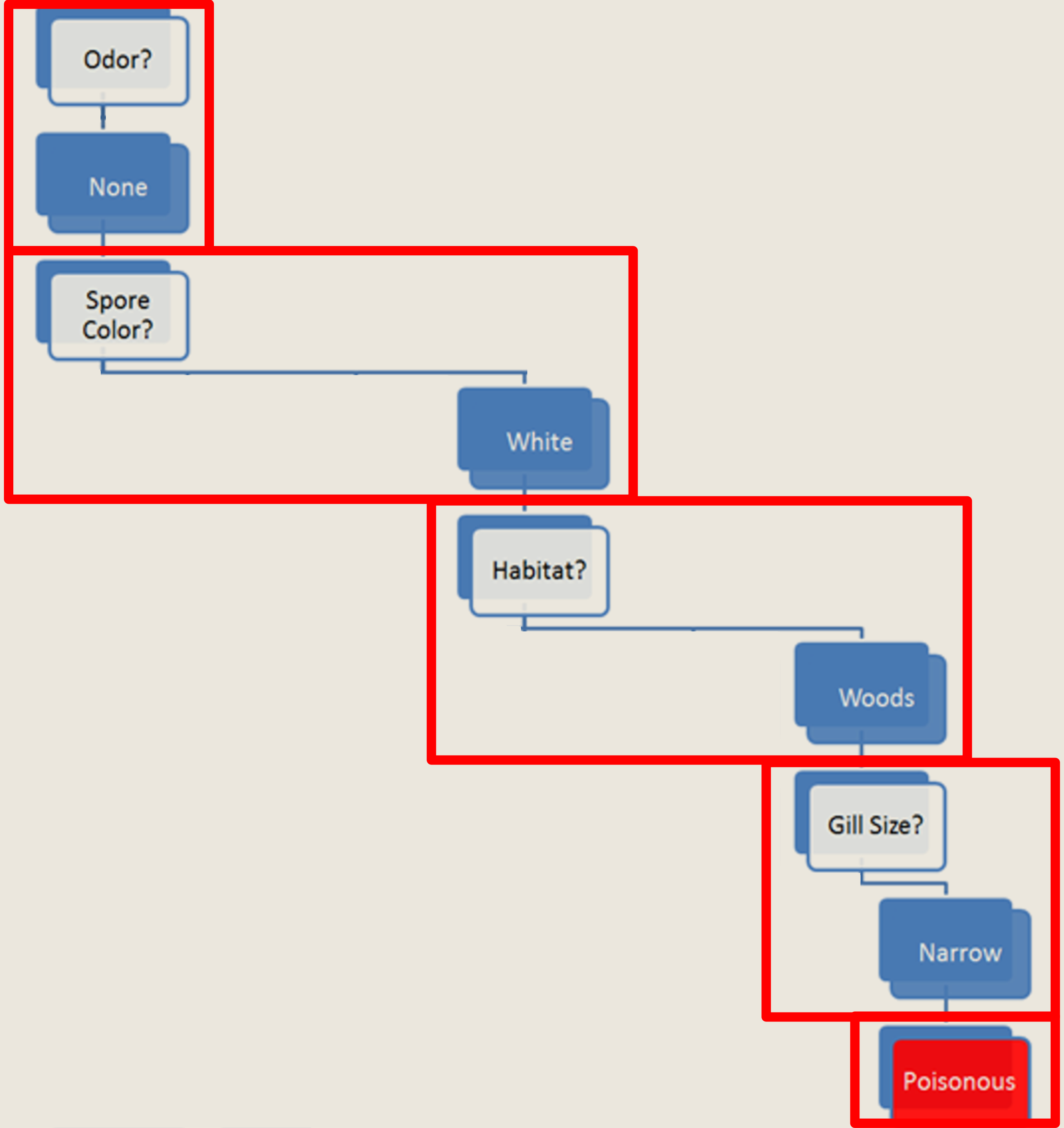
Gill Size: narrow

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?

ASKING THE RIGHT QUESTIONS

Data science is really about asking and answering questions:

- **Analytics:** “How many clicks did this link get?”
- **Data Science:** “Based on this user’s previous purchasing history, can I predict what links they will click on the next time they access the site?”

Data mining/science models are usually **predictive** (not **explanatory**): they show connections, but don't reveal why these exist.

Warning: not every situation calls for data science, artificial intelligence, machine learning, statistics, or analytics.

THE WRONG QUESTIONS

Too often, analysts are asking the **wrong questions**:

- questions that are **too broad** or **too narrow**
- questions that **no amount of data could ever answer**
- questions for which **data cannot reasonably be obtained**

The **best-case scenario** is that stakeholders will recognize the answers as irrelevant.

The **worst-case scenario** is that they will erroneously implement policies or make decisions based on answers that have not been identified as misleading and/or useless.

DATA SCIENCE/MACHINE LEARNING AUGMENTED INTELLIGENCE TASKS

Classification and **class probability estimation:** which clients are likely to be repeat customers?

Clustering: do diplomatic missions form natural groups?

Association rule discovery: what books are commonly purchased together?

Others:

- **profiling and behaviour description**; **link prediction**; **value estimation/regression analysis** (how much is a client likely to spend in a restaurant?); **similarity matching** (which prospective clients are similar to a company's best clients?); **data reduction**; **influence/causal modeling**, **anomaly detection**, **time series analysis**, etc.

SOME PRACTICAL DEFINITIONS

DATA INSIGHT FUNDAMENTALS

“What’s in a name? That which we call a rose
By any other name would smell as sweet.”

W. Shakespeare, Romeo and Juliet, Act II, Scene 2

WHAT IS DATA ANALYSIS?

Finding **patterns** in data

Using data to do something (answer a question, assist in decision-making, predict a future occurrence, draw a conclusion)

Creating models of the data

Describing or explaining a situation (the **system**)

(Testing (scientific) hypotheses?)

(Carrying out calculations on data?)

WHAT IS DATA SCIENCE?

Data science is the collection of processes by which we extract useful and **actionable insights** from data.

T. Kwartler (paraphrased)

Data science is the **working intersection** of statistics, engineering, computer science, domain expertise, and “hacking.” It involves two main thrusts: **analytics** (counting things) and **inventing new techniques** to draw insights from data.

H. Mason (paraphrased)

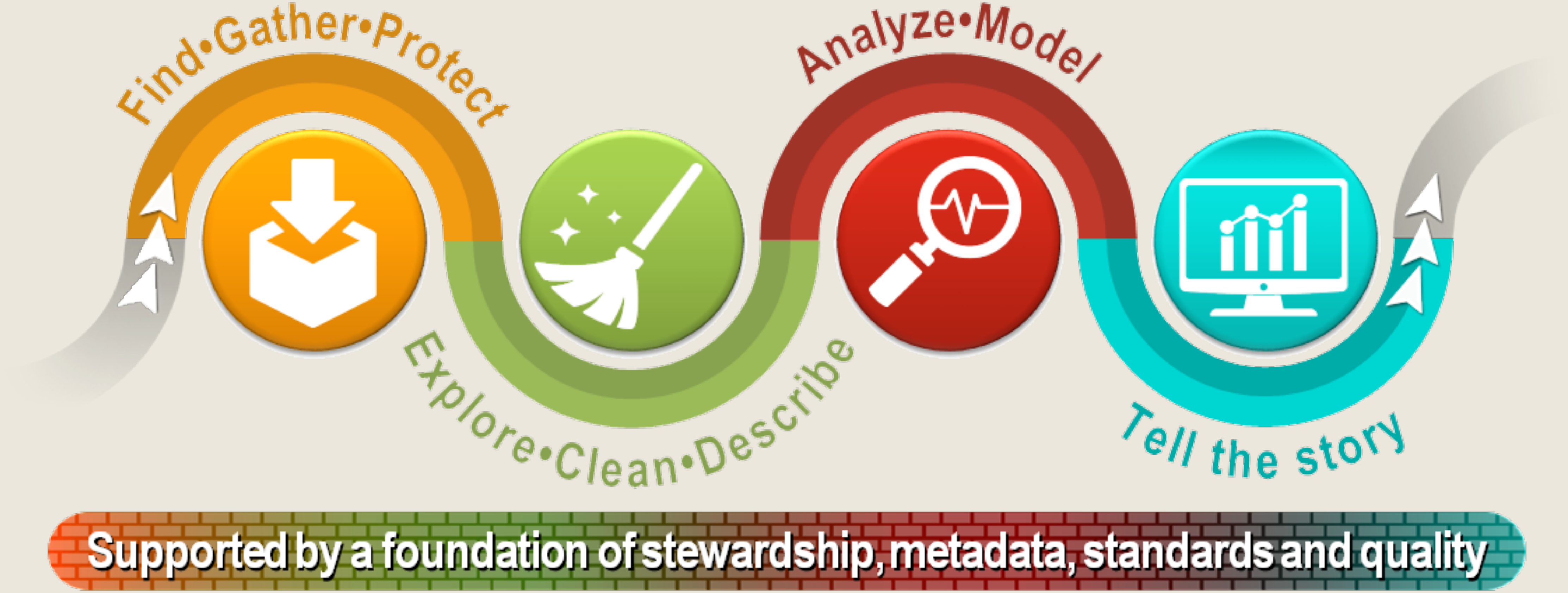
WORKFLOWS AND PIPELINES

DATA INSIGHT FUNDAMENTALS

“All models are wrong.
Some models are useful.”

George Box

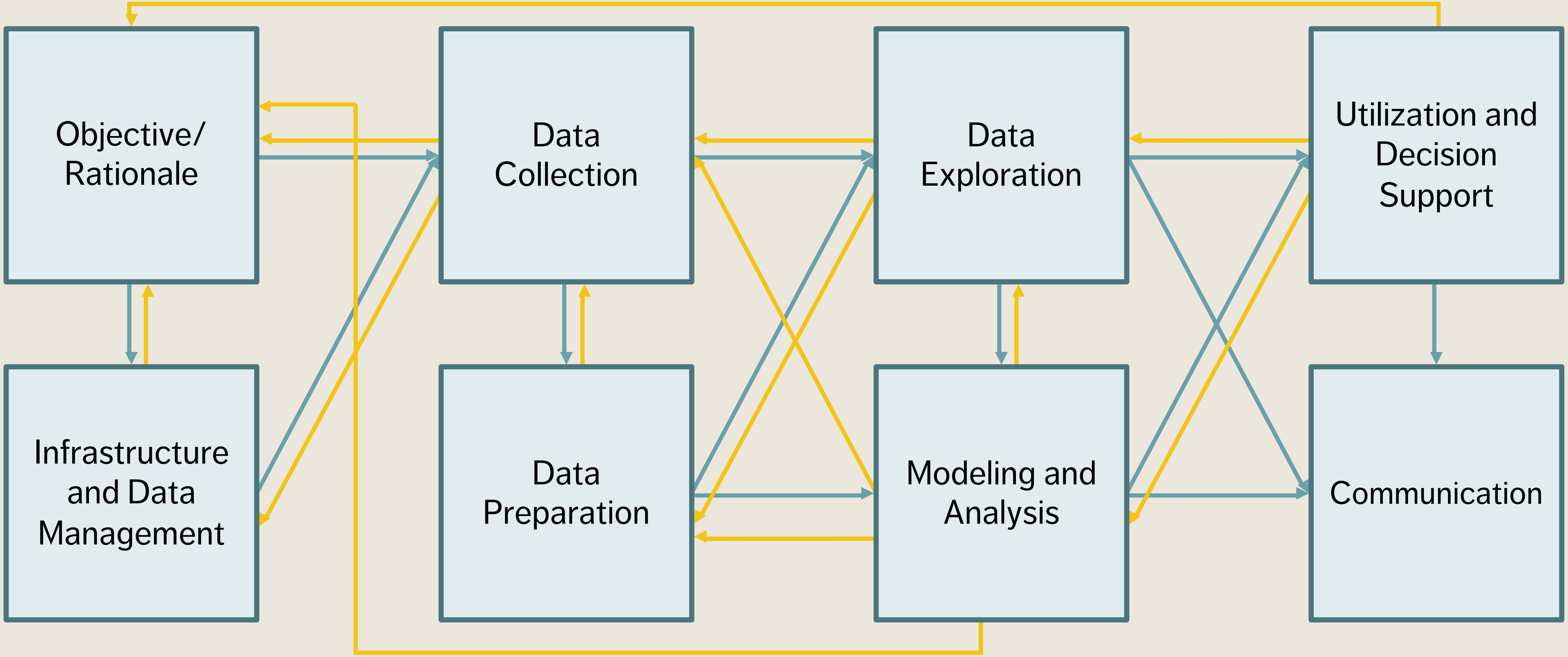
[StatCan, Steps of the Data Journey]



THE DATA SCIENCE WORKFLOW

- Objective/ Rationale
- Data Collection
- Data Exploration
- Utilization and Decision Support
- Infrastructure and Data Management
- Data Preparation
- Modeling and Analysis
- Communication

THE DATA SCIENCE WORKFLOW



THE DATA ANALYSIS PROCESS

A **large number of analytical models** have to be generated before a final selection can be made.

Iterative process: feature selection and data reduction may require numerous visits to domain experts before models start yielding promising results.

Domain-specific knowledge has to be integrated in the models in order to beat random classifiers and clustering schemes, **on average**.

LIFE AFTER ANALYSIS

When an analysis or model is ‘released into the wild’, it can take on a life of its own.

Analysts may eventually have to relinquish control over dissemination. Results may be misappropriated, misunderstood, or shelved. What can the analyst do to prevent this?

Finally, because of **analytic decay**, it’s important to view the last analytical step NOT as a static dead end, but rather as an invitation to return to the beginning of the process.

DATA SCIENCE ECOSYSTEM

Data analysis is a **team sport**, with team members needing a good understanding of both **data** and **context**

- data management
- data preparation
- analysis
- communications

Even slight improvements over a current approach can find a useful place in an organization – **data science is not solely about Big Data and disruption!**

MODEL ASSESSMENT AND VALIDITY

Models should be **current**, **useful**, and **valid**.

Data can be used in conjunction with existing models to come to some conclusions or can be used to update the model itself.

At what point does one determine that the current data model is **out-of-date** or is **not useful anymore**?

Past successes can lead to **reluctance** to re-assess/re-evaluate a model.

MODELS AND SYSTEMS THINKING

DATA INSIGHT FUNDAMENTALS

"What if the only valid model of the Universe is the Universe itself?"

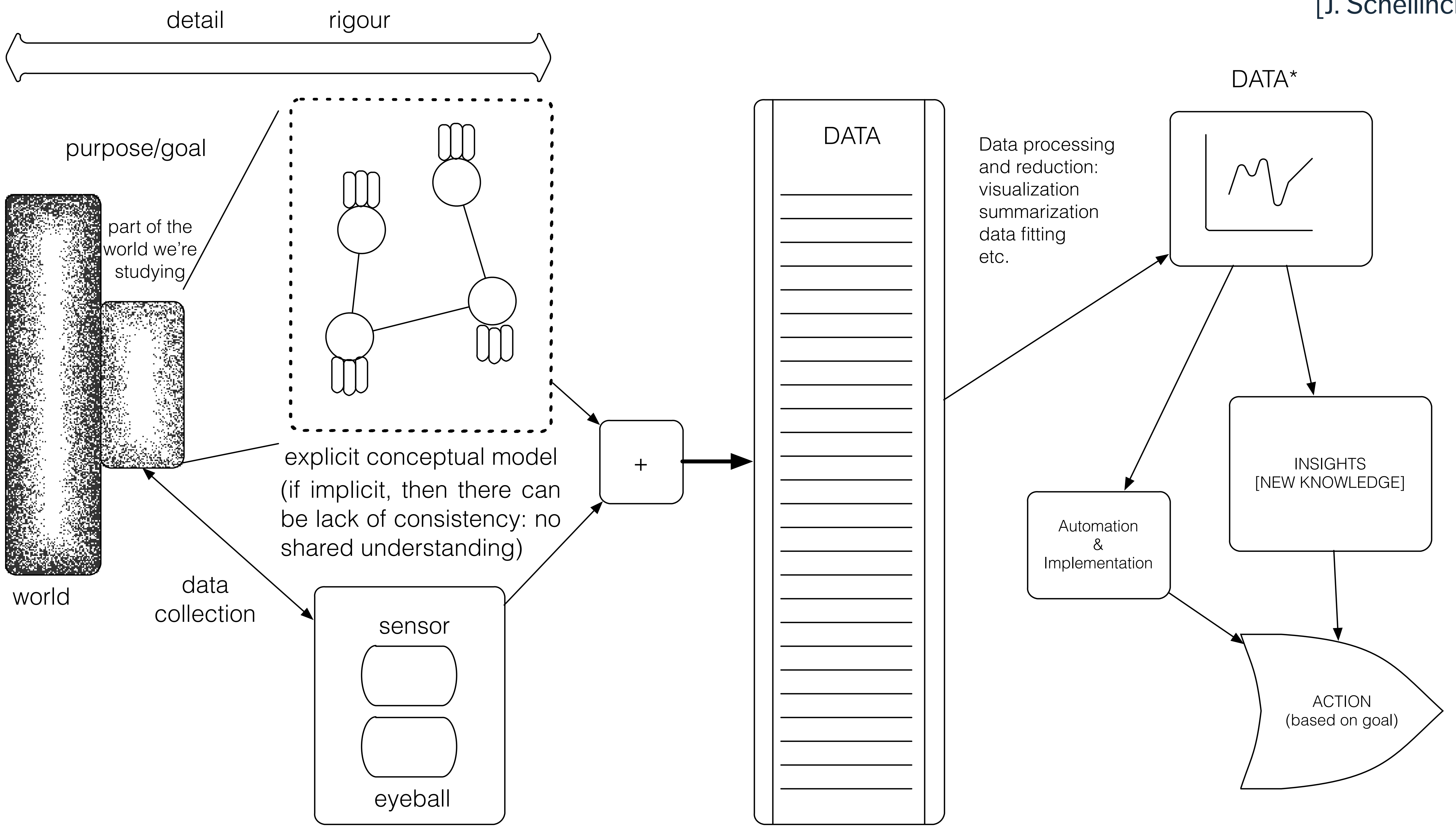
Unknown

REPRESENTATIONS

A **representation** is an object that stands in for another object.

A representation may or may not physically resemble the object it represents.

Representations of the world help us to **understand**, **navigate**, and **manipulate** the world.


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 processing and reduction:
visualization
summarization
data fitting
etc.
DATA*
Automation & Implementation
INSIGHTS [NEW KNOWLEDGE]
ACTION (based on goal)

THINKING IN SYSTEMS TERMS

A **system** is made up of **objects** with **properties** that potentially change over time. Within the system we perceive **actions** and **evolving** properties leading us to think in terms of **processes**.

We **observe**, **quantify**, and **record** particular values of these properties at particular moments in time.

This generates data points, capturing the **underlying reality** that the system approximates to some degree of **accuracy** and **error**.

IDENTIFYING GAPS IN KNOWLEDGE

A **gap in knowledge** is identified when we realize that what we thought we knew about a system proves incomplete (or false).

This might happen repeatedly, at any moment in the process:

- data cleaning
- data consolidation
- data analysis

The solution is to be flexible. When faced with such a gap, **go back**, **ask questions**, and **modify the system representation**.

CONCEPTUAL MODELING

Exercise:

- an acquaintance has just set foot in your living space for the first time
- you are on the phone with them but not currently at home
- explain to them how to go about changing a breaker
- (how would you do so if the acquaintance was visually impaired?)

Conceptual models are built using methodical investigation tools

- diagrams
- structured interviews
- structured descriptions
- etc.

RELATING THE DATA TO THE SYSTEM

Is the data which has been collected and analyzed going to be of any use when it comes to understanding the system?

This question can only be answered if we understand:

- how the data is **collected**
- the **approximate nature** of both data and system
- what the data **represents** (observations and features)

Is the combination of system and data **sufficient** to understand the aspects of the world under consideration?

[http://dailycatdrawings.tumblr.com; #309]

Real World

Theory

Identification of details relevant to **description** and **translation** of real-world objects into model variables

Model

TAKE-AWAYS

Systems can approximate certain aspects of the Universe.

System models provide the basis under which data is identified and collected, but data itself is approximate and selective.

Knowledge gaps happen – be ready to re-visit your set-up regularly.

Implicit conceptual modeling can lead to problematic situations.

If the data, the system, and the world are out of alignment, data analysis insights might ultimately prove useless.

ETHICAL CONSIDERATIONS & BEST PRACTICES

DATA INSIGHT FUNDAMENTALS

"We have flown the air like birds and swum the sea like fish but have yet to learn the simple act of walking the Earth like brothers."

Martin Luther King, Jr.

What harm can come from data?

THE NEED FOR ETHICS

Formerly: “**Wild West**” mentality to data collection (and use). Whatever wasn’t technologically forbidden was allowed.

Now: professional codes of conduct are being devised for data scientists (outline responsible ways to practice data science).

Additional responsibility for data scientists; but also, **protection** against being hired to carry out questionable analyses.

WHAT ARE ETHICS?

Broadly speaking, ethics refers to the **study** and **definition** of **right and wrong conducts:**

- "not [...] social convention, religious beliefs, or laws". (R.W. Paul, L. Elder)

Influential ethical theories:

- Kant's **golden rule** (do onto others...), **consequentialism** (the ends justify the means), **utilitarianism** (act in order to maximize positive effect), etc.
- **Confucianism**, **Taoism**, **Buddhism** (?), etc.
- **Ubuntu**, **Maori**, etc.

[OCAP® is a registered trademark of the *First Nations Information Governance Centre* (FNIGC)]

WHAT ARE ETHICS?

First Nations Principles of OCAP®:

- **Ownership**
  cultural knowledge, data, and information is owned by communities
- **Control**
  communities have the right to control all aspects of research and information management that impact them
- **Access**
  communities must have access to information and data about themselves no matter where it is held
- **Possession**
  communities must have physical control of relevant data

[University of Virginia, Centre for Big Data Ethics, Law, and Policy]

ETHICS IN THE DATA CONTEXT

Data ethics questions:

- **Who**, if anyone, owns data?
- Are there **limits** to how data can be used?
- Are there **value-biases** built into certain analytics?
- Are there categories that should **not** be used in analyzing personal data?
- Should some data be **publicly available** to **all** researchers?

Analytically, the **general** is preferred to the **anecdotal** – decisions made on the basis of machine learning and A.I. (security, financial, marketing, etc.) may affect real beings in **unpredictable ways**.

BEST PRACTICES

"Do No Harm": data collected from an individual **should not be used to harm** the individual.

Informed Consent:

- Individuals must **agree to the collection and use** of their data
- Individuals must have a **real understanding of what they are consenting to**, and of **possible consequences** for them and others

Respect "Privacy": excessively hard to maintain in the age of constant trawling of the Internet for personal data.

[More information on how to consider data privacy and human rights is found in the Intranet Data Wiki.]

BEST PRACTICES

Keep Data Public: data should be kept **public** (all? most? any?).

Opt-In/Opt-Out: Informed consent requires the ability to **opt out**.

Anonymize Data: removal of id fields from data prior to analysis.

“Let the Data Speak”:

- no cherry picking
- importance of validation (more on this later)
- correlation and causation (more on this later, too)
- repeatability

[https://www.statcan.gc.ca/eng/topics-start/gender_diversity_and_inclusion]

GBA+

Gender-Based Analysis Plus is an analytical process used to assess how different gendered people may experience policies, programs and initiatives.

Example: Work interruptions and financial vulnerability, D. Messacar, R. Morrissette

- If the data had not been collected and/or analyzed in a GBA+ manner, it would be harder to see how financial vulnerability affects different groups (if the analysis had looked only at age groups and gender, for example, instead of also including family composition).

Policies and events **impact real people in real way**, and not always in the same manner. Data analysis methods are typically used to predict and/or describe **average** (or central) outcomes, but it is often those who are far from the centre who are most affected.

Let’s be smart about this.