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

[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

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
…	…	…	…	…	…

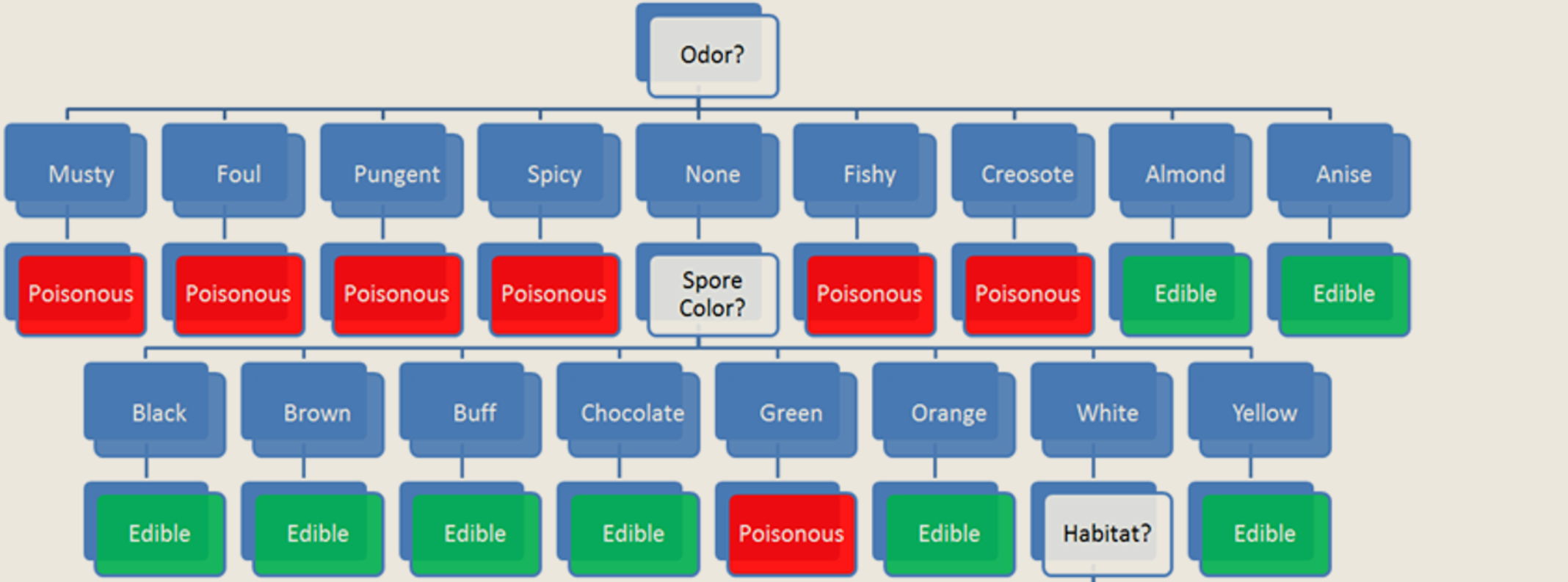
POISONOUS MUSHROOM DATASET

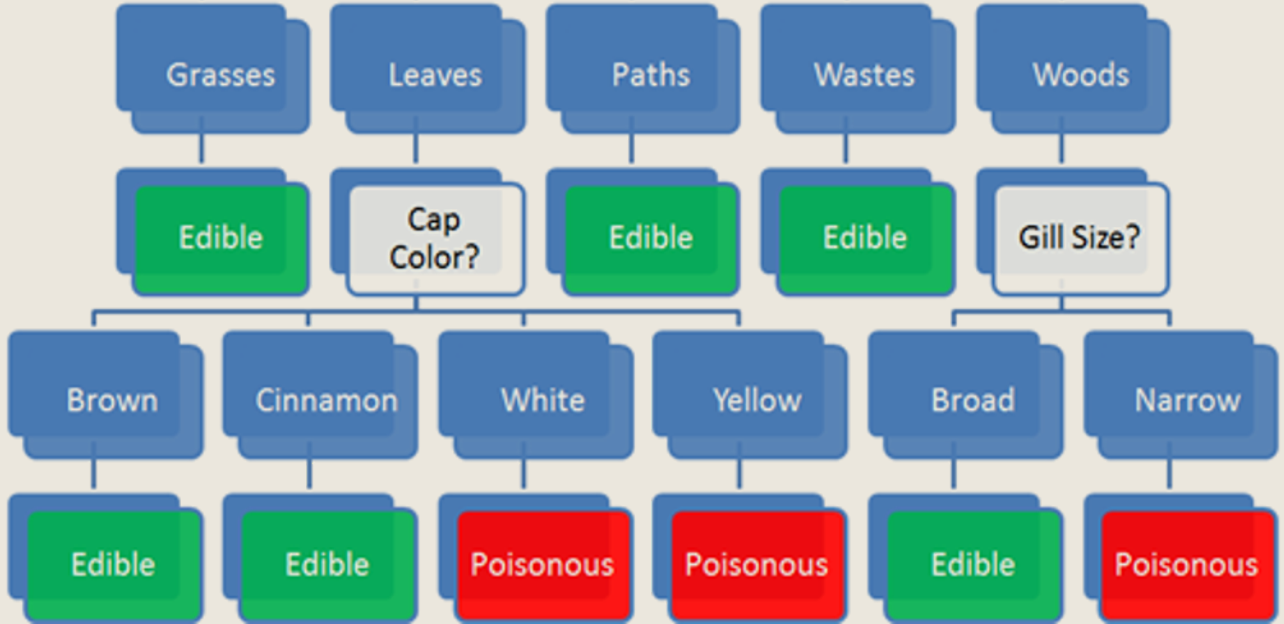
Amanita muscaria

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

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

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





Habitat: woods

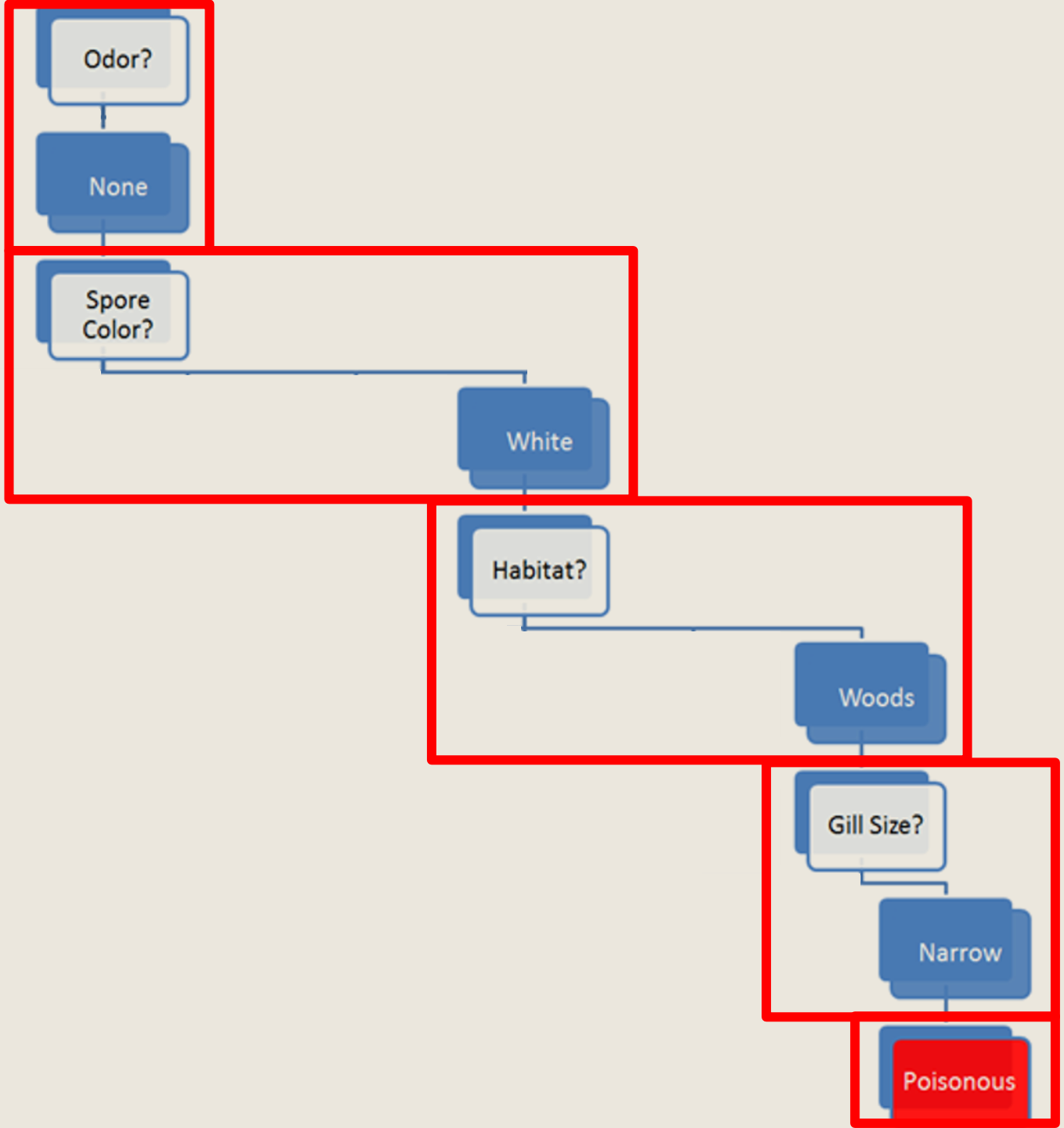
Gill Size: narrow

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**

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.

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.

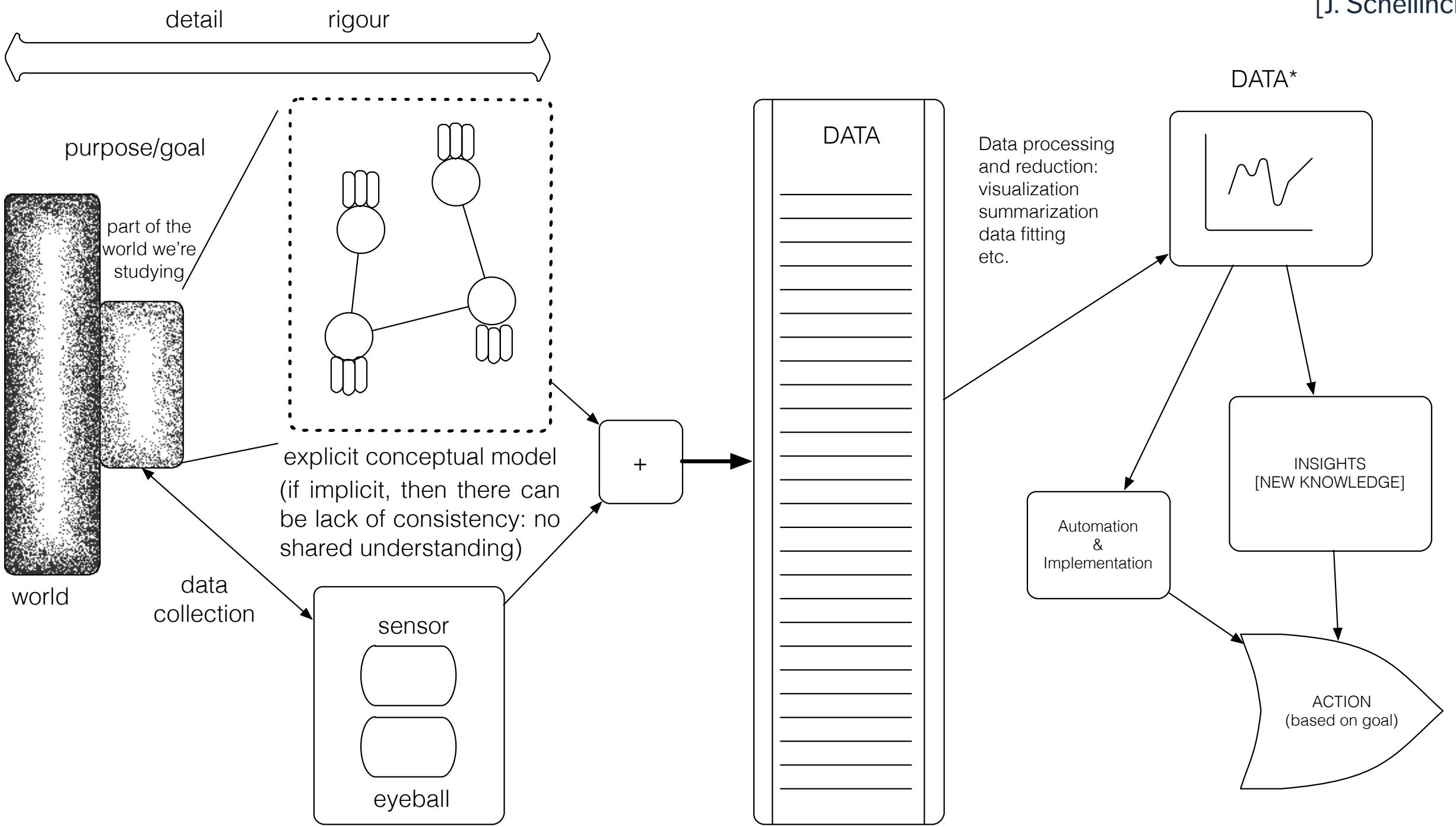
T. Kwartler (paraphrased)

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)

[StatCan, Steps of the Data Journey]

[J. Schellinck]



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

Real World

Model

Theory

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

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.

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.

[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