DATA INSIGHT FUNDAMENTALS

SETTING THE STAGE

OUTLINE

1. Data, M.L., and A.I. in the News
2. Data 101 – Basic Data Concepts
3. Some Practical Definitions
4. Workflows and Pipelines – the Process of Working with Data
5. Models and Systems Thinking
6. Ethical Considerations and Best Practices

HEADLINES

Robots are better than doctors at diagnosing some cancers, major study finds. [reported on by The Telegraph, UK, 29-05-2018]

MRNet: Deep-learning-assisted diagnosis for knee magnetic resonance imaging [Stanford ML Group]

Data scientists find connections between birth month and health [Columbia University Medical Centre]

We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually. [MIT Technical Review, 21-12-2018]

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** (or columns) in a database; objects are **instances** (or 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
…	…	…	…	…	…

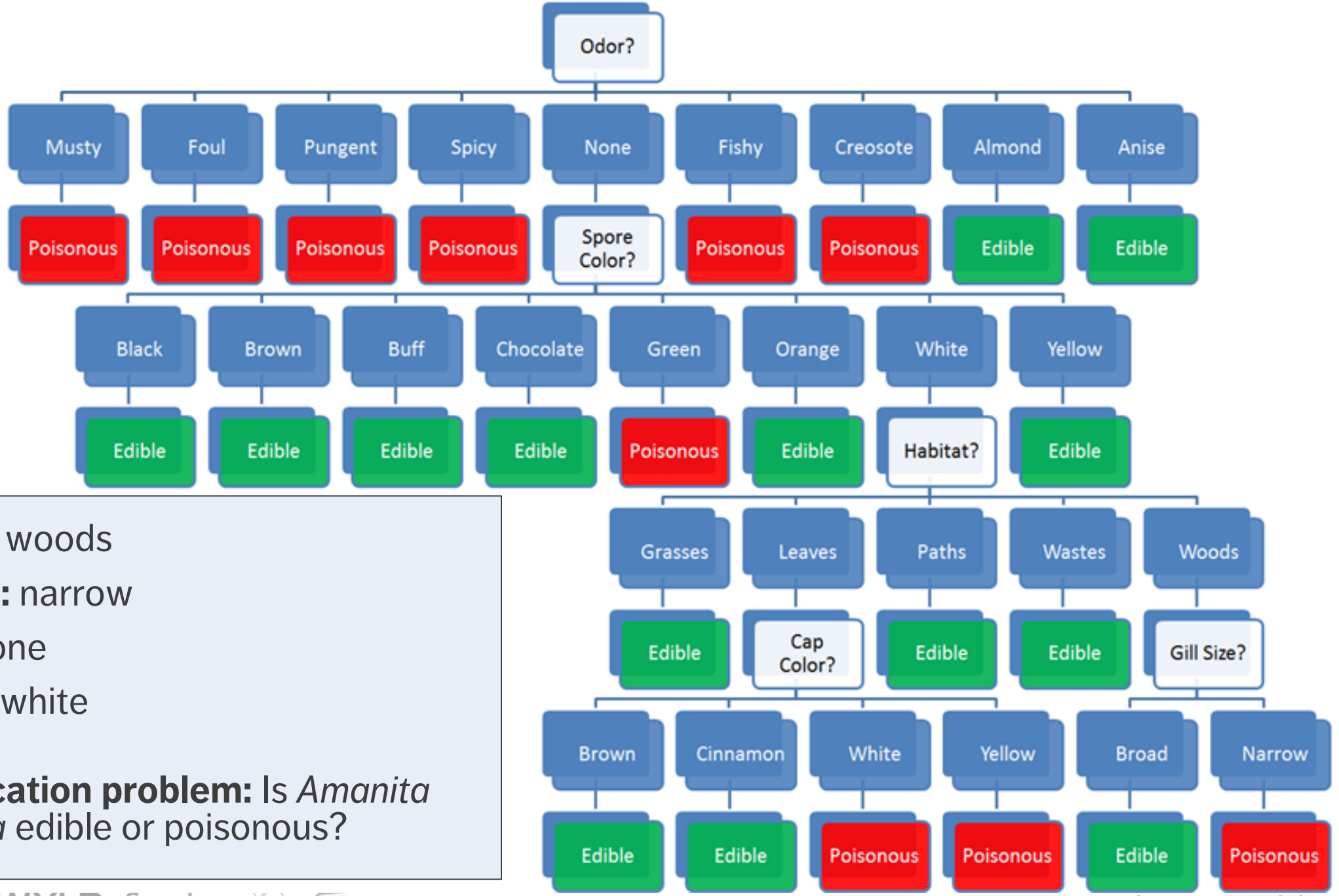
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

POISONOUS MUSHROOMS DATASET

Amanita muscaria

Habitat: woods
Gill Size: narrow
Odor: none
Spores: white

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



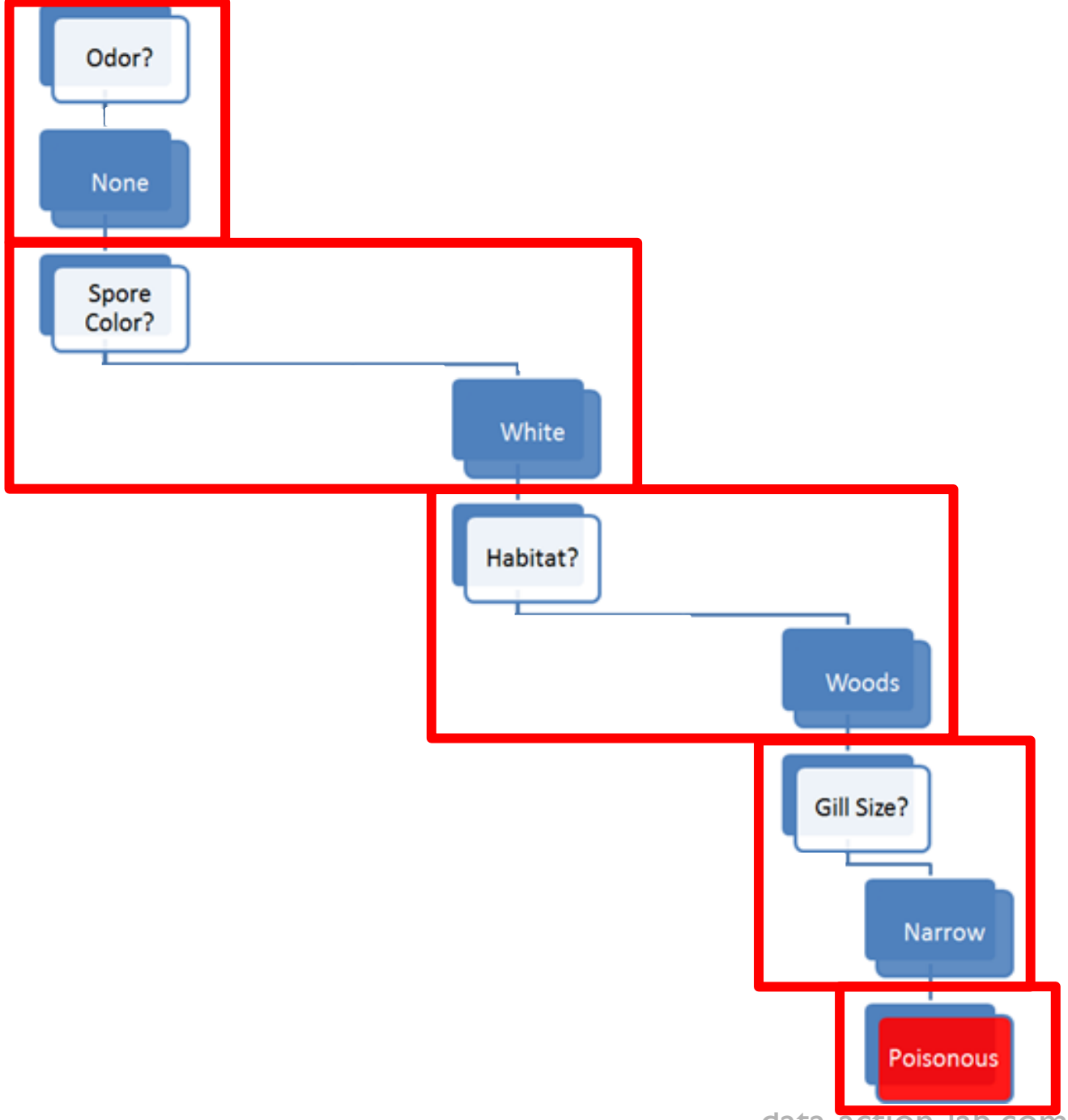
Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

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



Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Classification problem: Is *Amanita muscaria* 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, or analytics.

DATA SCIENCE/MACHINE LEARNING/A.I. TASKS

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

Clustering: do customers form natural groups?

Association rule discovery: what books are commonly purchased together?

Others:

profiling and behaviour description; **link prediction**; **value estimation** (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**, etc.

WHAT IS DATA ANALYSIS?

Finding **patterns** in data

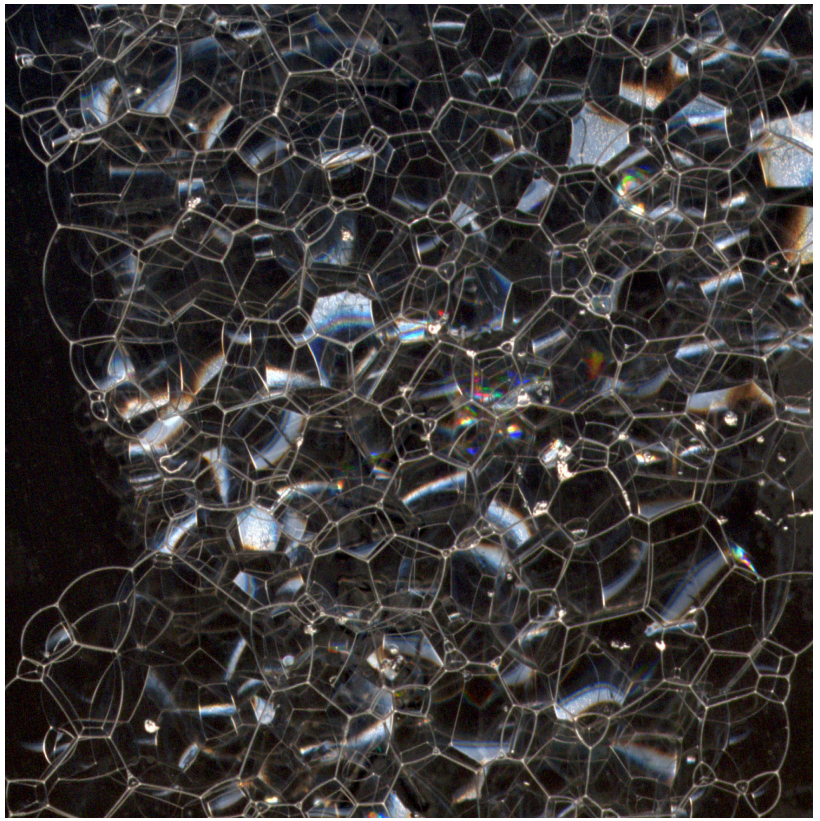
Using data to do something (answer a question, help decision-making, predict the future, draw a conclusion)

Creating models of your data

Describing or explaining your situation (your **system**)

(Testing (scientific) hypotheses?)

(Carrying out calculations on data?)



The more complicated the pattern, the more complicated the analysis

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.

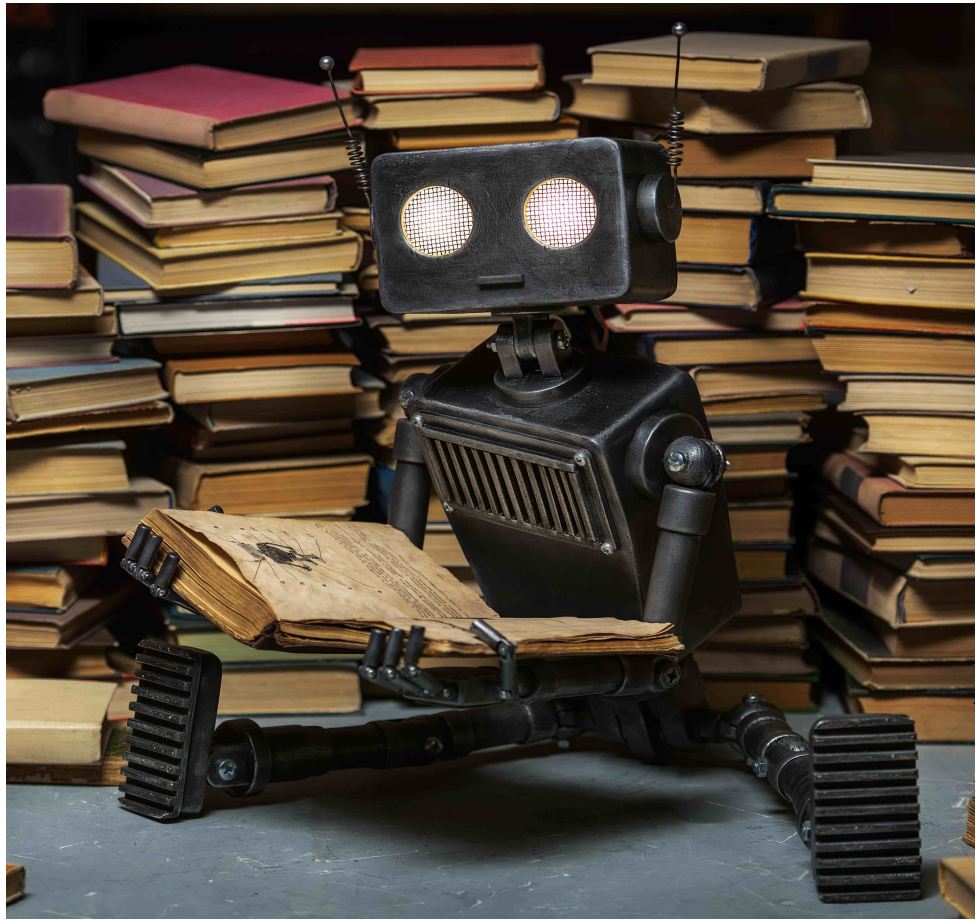
H. Mason (paraphrased)

WHAT IS MACHINE LEARNING?

Starting around the 1940s researchers began in earnest to teach machines how to learn

The goal of **machine learning** was to create machines that could learn and adapt and respond to novel situations

A wide variety of techniques, accompanied by a great deal of theoretical underpinning, was created in an effort to achieve this goal



WHAT IS ARTIFICIAL/AUGMENTED INTELLIGENCE?

Artificial Intelligence (A.I.) is non-human intelligence that has been engineered rather than one that has evolved naturally.

Artificial intelligence research is research carried out in pursuit of this goal.

Pragmatically speaking, A.I. is “computers carrying out tasks that only humans can usually do”.

Augmented Intelligence is human intelligence that is supported or enhanced by machine intelligence.

THE DATA SCIENCE "WORKFLOW"

Objective/ Rationale

Data Collection

Data Exploration

Utilization and Decision Support

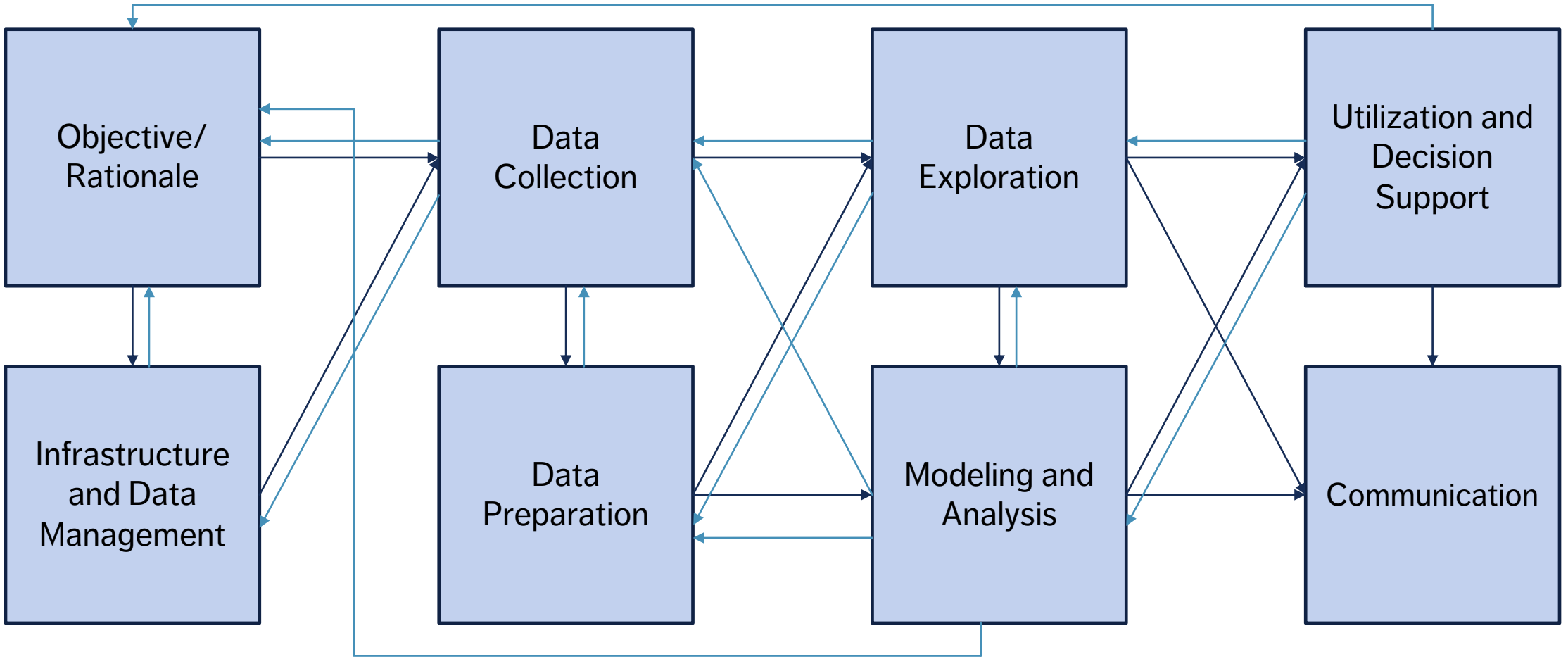
Infrastructure and Data Management

Data Preparation

Modeling and Analysis

Communication

THE DATA SCIENCE "WORKFLOW"



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

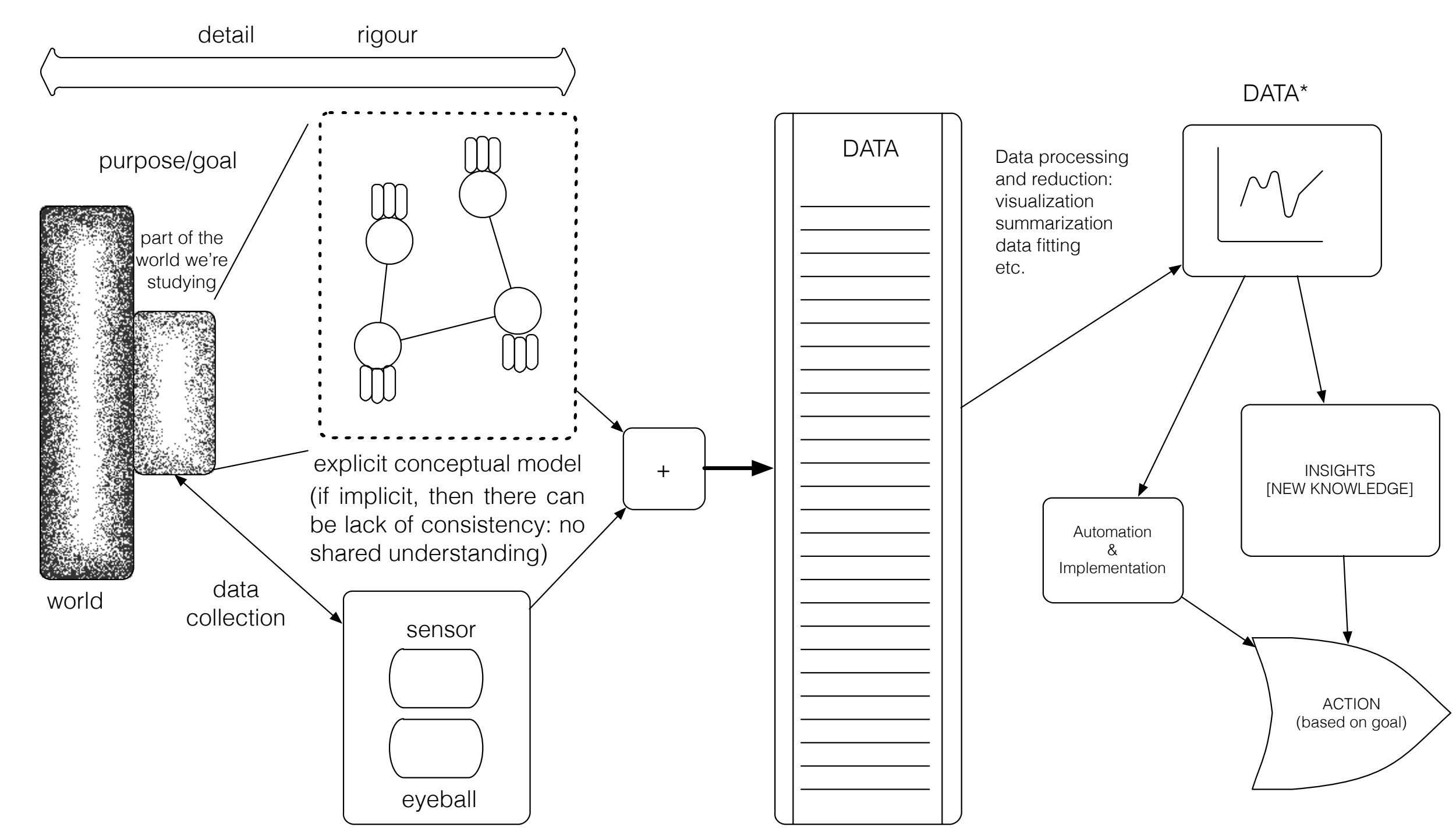
Even slight improvements over a current approach can find a useful place in an organization – **data science is not solely about Big Data and disruption!**

REPRESENTATION

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
explicit conceptual model (if implicit, then there can be lack of consistency: no shared understanding)
data collection
sensor
eyeball
+
DATA
DATA*
Data processing and reduction: visualization summarization data fitting etc.
Automation & Implementation
INSIGHTS [NEW KNOWLEDGE]
ACTION (based on goal)

THINKING IN SYSTEMS TERMS

In order to understand how various aspects of the World interact with one another, we need to **carve out chunks** corresponding to the aspects and define their **boundaries**.

Working with other intelligences requires **shared understanding** of what is being studied.

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**.

THINKING IN SYSTEMS TERMS

Objects themselves have various properties. Natural processes generate (or destroy) objects, and may change the properties of these objects over time.

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

This generates data points, capturing the **underlying reality** to some degree of **accuracy** and **error** (biased or unbiased).

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 MODELS

Exercise:

- assume that 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 preparing a cup of sugar.

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?

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.

Does your organization have a code of ethics for its data scientists? For its employees?

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 *Western* ethical theories:

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

Influential *Eastern* ethical theories:

- **Confucianism**, **Taoism**, **Buddhism** (?), etc.

WHAT ARE ETHICS?

First Nations Principles of **OCAP®**:

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

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.

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

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 and re-evaluate a model.