# MODULE 2: DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

CT ACADEMY | DATA ACTION LAB

# 7. DATA ANALYTICS

## DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

# QUANTITATIVE SKILLS

Suggestions:

- **keep up with trends**
- become **conversant in your non-expertise areas**
- know **where to find information**

In many instances (70%?), only the basics (2nd–3rd year mandatory courses at uOttawa, say) are sufficient to meet government/industry needs.

**Focus:** make sure you really **understand** the basics, stepping stones.

In the rest of the cases, more sophisticated knowledge is required.

# QUANTITATIVE SKILLS

- survey sampling and data collection
- data processing and data cleaning
- data visualization
- mathematical modelling
- statistical methods
- regression analysis
- queueing models
- machine learning
- deep learning
- reinforcement learning
- stochastic modelling (MC simulations)
- optimization and operations research
- survival analysis
- Bayesian data analysis
- anomaly detection and outlier analysis
- feature selection/dimensions reduction
- trend extraction and forecasting
- cryptography and coding theory
- design of experiment
- graph and network theory
- text mining/natural language processing
- etc.

# SOFTWARE AND TOOLS

## Programming (and Related)

- Python, R, C/C++/C#, Perl, Julia, regexps (, Visual Basic?), Java, Ruby, etc.

## Database Management

- SQL and variants, ArangoDB, MongoDB, Redis, Amazon DynamoDB (, Access?), Big Query, Redshift, Synapse, etc.

## Data Visualization

- ggplot2, seaborn, plot.ly, Power BI, Tableau, D3.js, Google Data Studio, proprietary software, etc.

## Simulations, Statistical Analysis, Data Analysis, Machine Learning

- tidyverse, scikit-learn, numpy, pandas, scipy, MATLAB, Simulink, SAS, SPSS, STATA (, Excel?), Visio, TensorFlow, keras, Spark, Scala, etc.

## Typesetting and Reporting

- LaTeX, R Markdown, Adobe Illustrator, GIMP (, Word?, PowerPoint?), etc.

# EXERCISES

1. Which of the quantitative skills presented in this section do you possess? Which interest you? Which do you plan on learning about?

2. Which of the software skills presented in this section do you possess? Which interest you? Which do you plan on learning about?

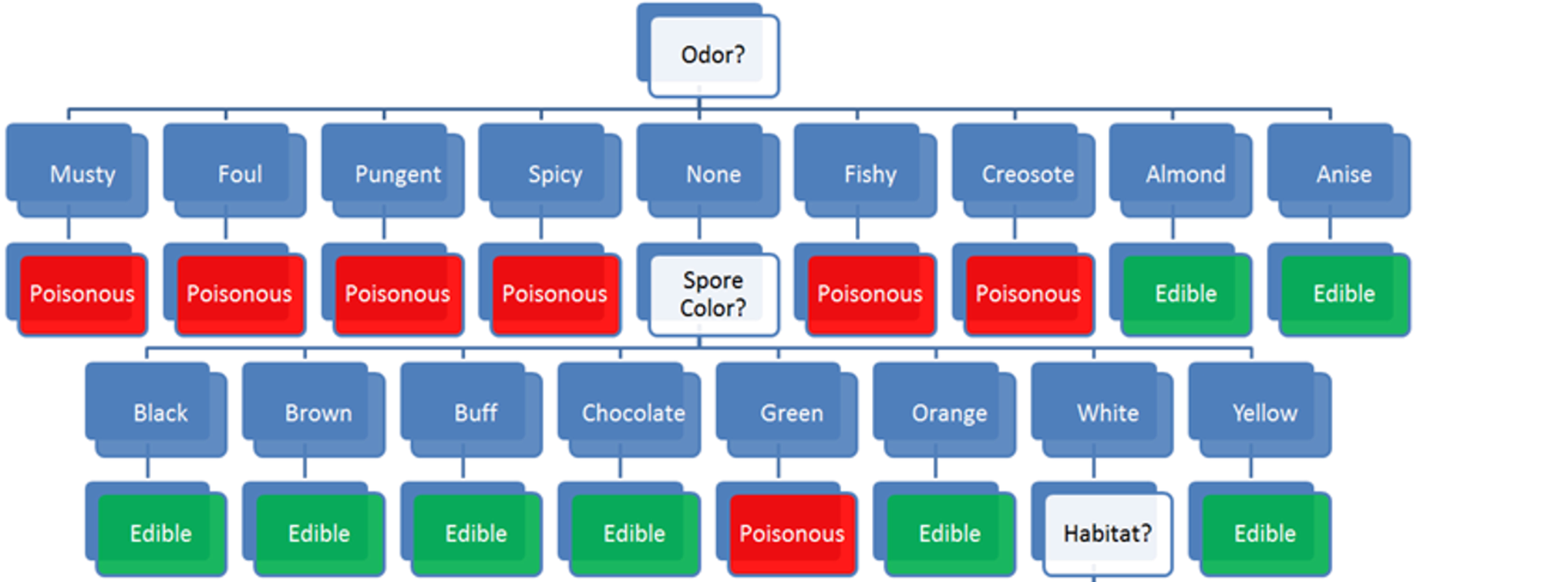
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

# EXAMPLE: 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

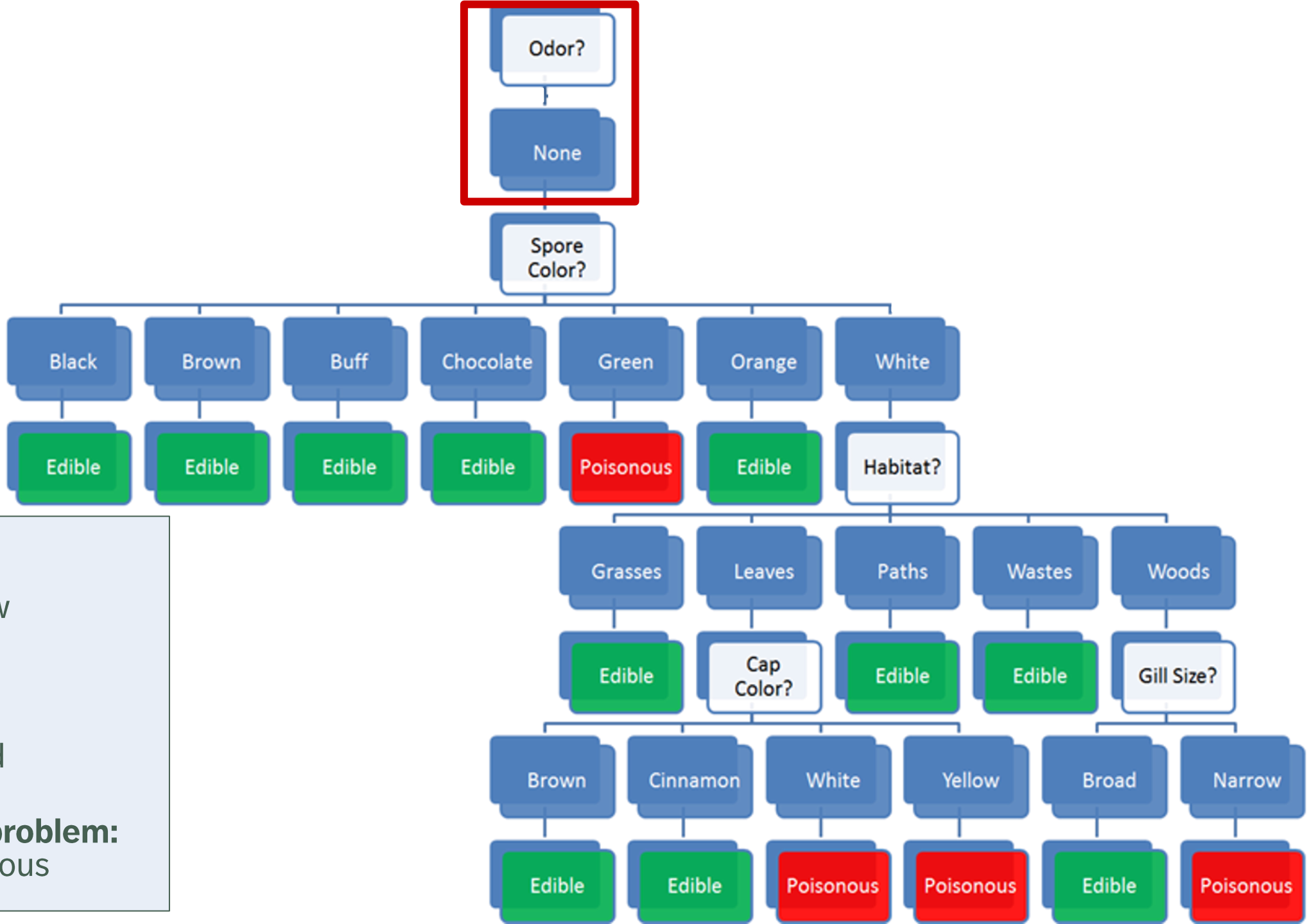
**Gill Size:** narrow

**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

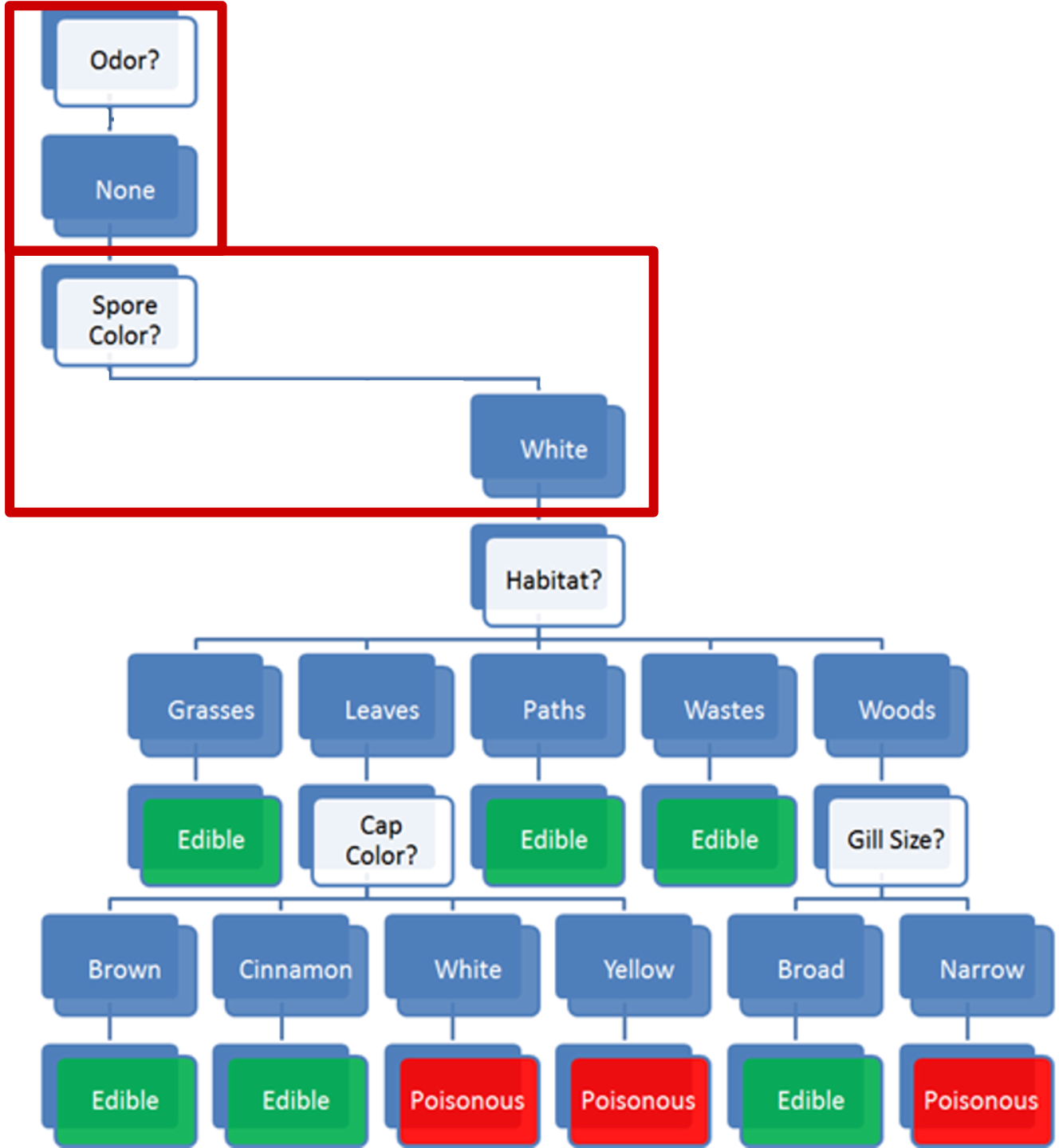
**Gill Size:** narrow

**Odor:** none

**Spores: white**

**Cap Colour:** red

**Classification problem:**
edible or poisonous



**Habitat: woods**

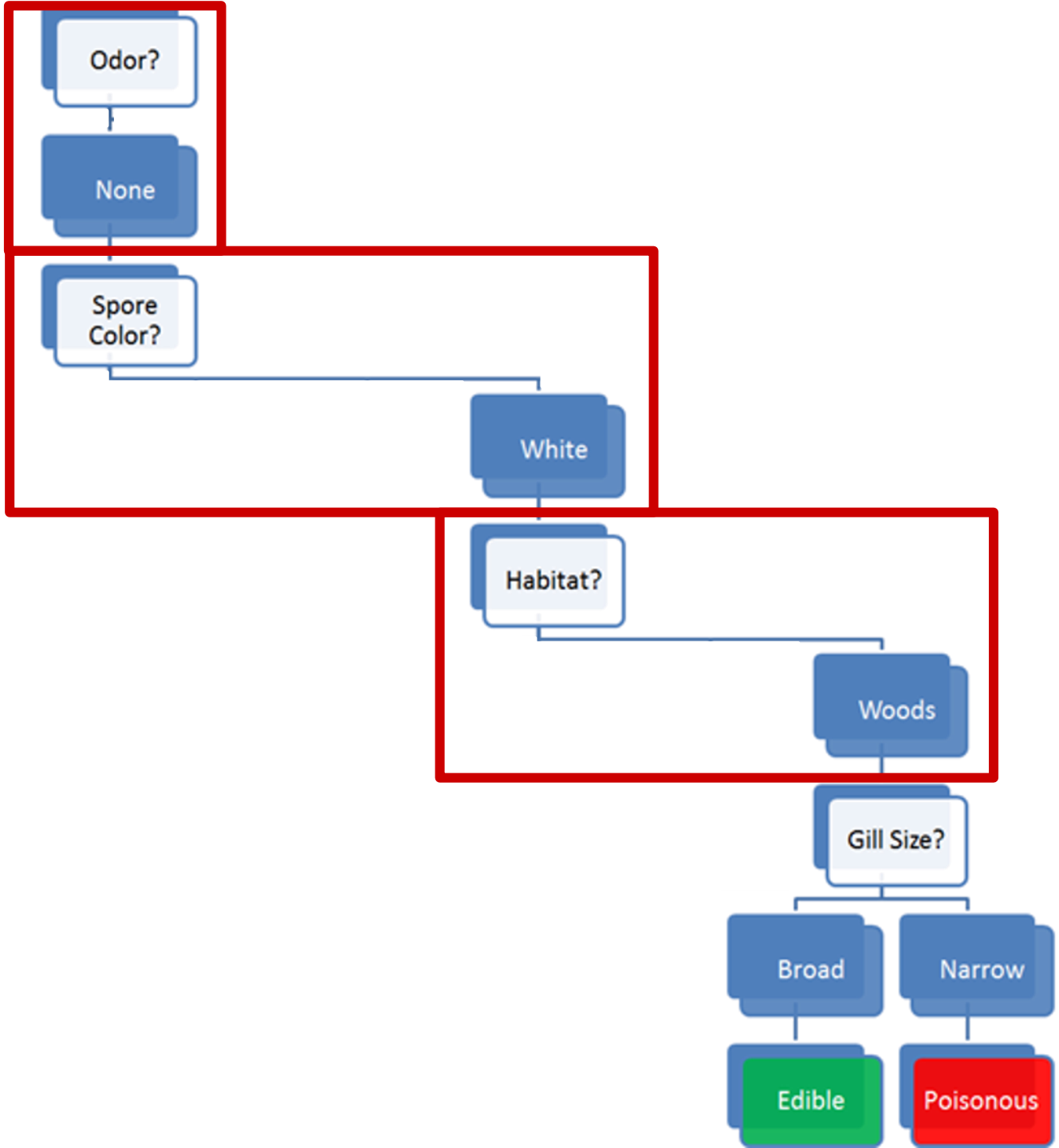
**Gill Size:** narrow

**Odor:** none

**Spores:** white

**Cap Colour:** red

**Classification problem:**
edible or poisonous



**Habitat:** woods

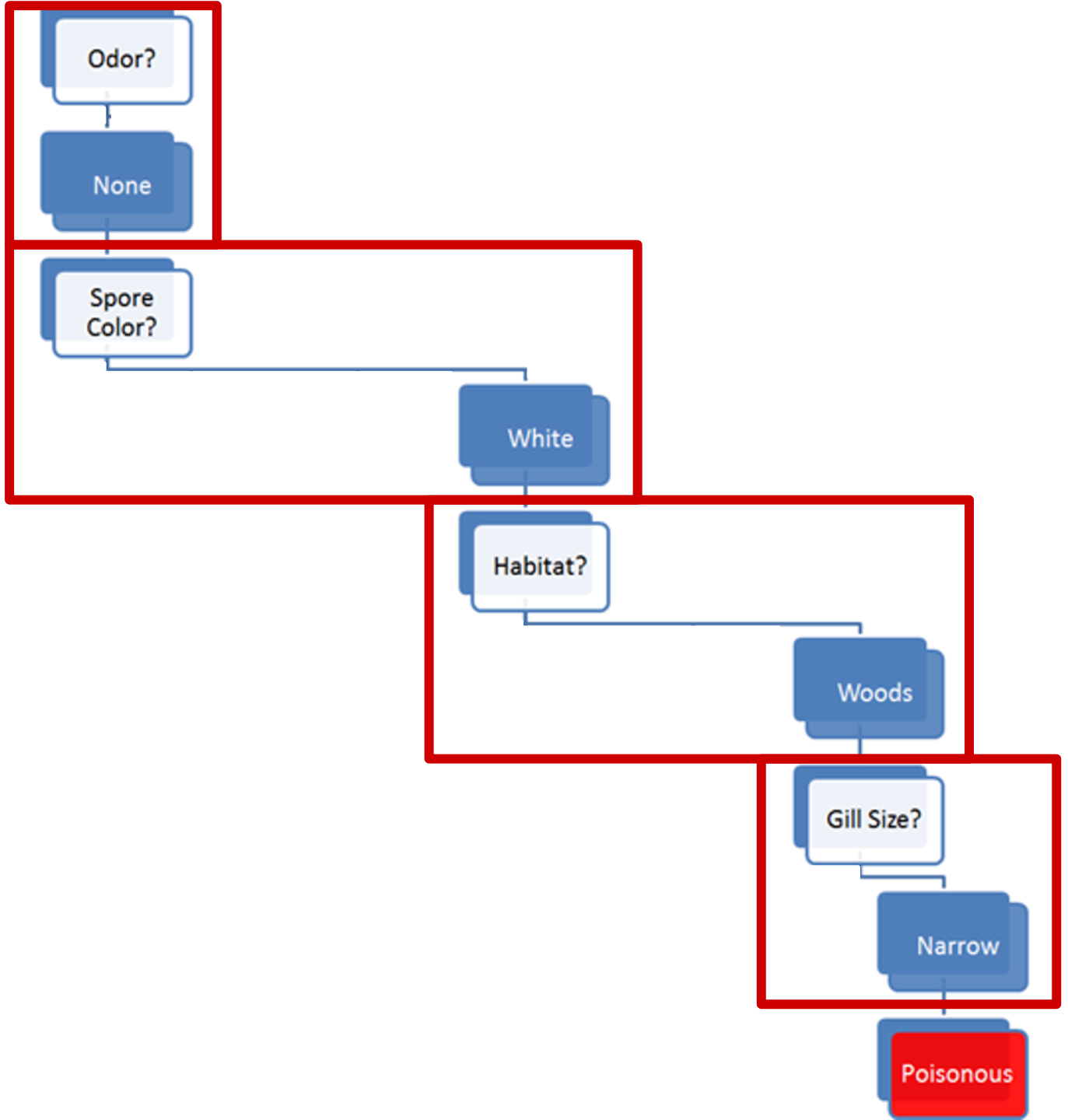
**Gill Size:** narrow

**Odor:** **none**

**Spores:** white

**Cap Colour:** red

**Classification problem:**
edible or poisonous



**Habitat:** woods

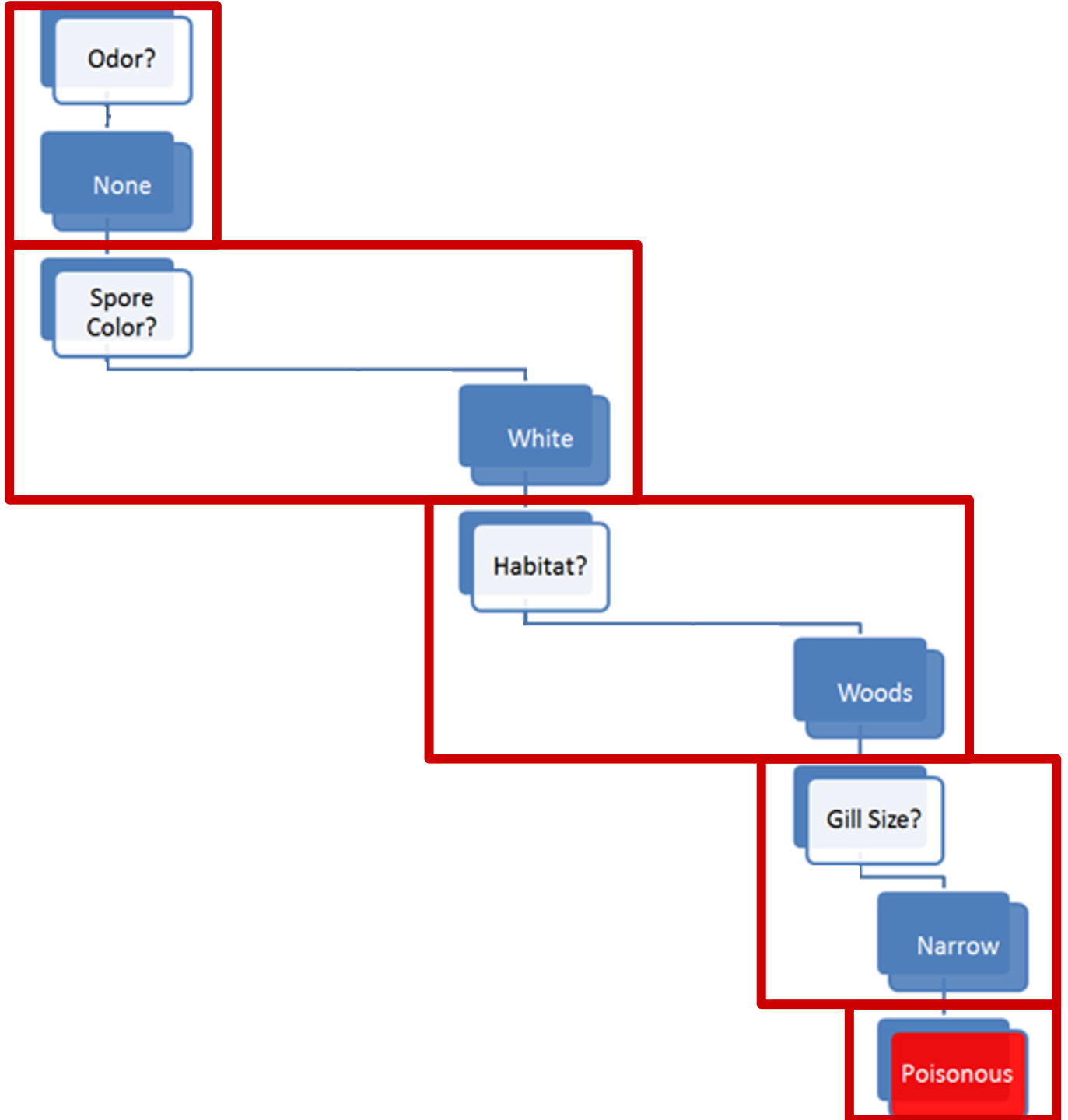
**Gill Size:** narrow

**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?

What is the financial equivalent of this example?

# WHAT IS DATA?

It is difficult to give a clear-cut definition of **data** (is it singular or plural?).

Linguistically, a *datum* is “a piece of information”; **data** means “pieces of information,” or a **collection** of “pieces of information”.

*Data* represents the whole (greater than the sum of its parts?) or simply the idealized concept.

Is that clear?

# WHAT IS DATA?

Is the following data?

4,529 red 25.782 Y

Why? Why not? What, if anything is missing?

The Stewart approach: “we know it when we see it.”

Pragmatically, we think of data as a collection of facts about **objects** and their **attributes**.

# OBJECTS AND ATTRIBUTES

Object: *apple*

- **Shape:** spherical
- **Colour:** red
- **Function:** food
- **Location:** fridge
- **Owner:** Jen

Object: *sandwich*

- **Shape:** rectangle
- **Colour:** brown
- **Function:** food
- **Location:** office
- **Owner:** Pat

Remember: an object is not simply **the sum of its attributes**.

# OBJECTS AND ATTRIBUTES

Ambiguities when it comes to **measuring** (and **recording**) the attributes:

- apple picture is a 2-dimensional representation of a 3-dimensional object
- overall shape of the sandwich is vaguely rectangular, it is not exact (**measurement error**?)
- insignificant for most, but not necessarily all, analytical purposes
- apple's shape = volume, sandwich's shape = area (**incompatible measurements)**
- a number of potential attributes are not mentioned: size, weight, time, etc.
- are there other issues?

Measurement errors and incomplete lists are always part of the picture; is this collection of attributes providing a reasonable **description** of the objects?

[R. Magritte, La trahison des images (ceci n'est pas une pipe)]

Ceci n'est pas une pipe.

# FROM OBJECTS AND ATTRIBUTES TO DATASETS

**Raw data** may exist in any format.

A **dataset** represents a collection of data that could conceivably be fed into algorithms for analytical purposes.

Datasets appear in a **table** format, with rows and columns; attributes are the **fields** (or columns, variables); objects are **instances** (or cases, rows, records).

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

# FROM OBJECTS AND ATTRIBUTES TO DATASETS

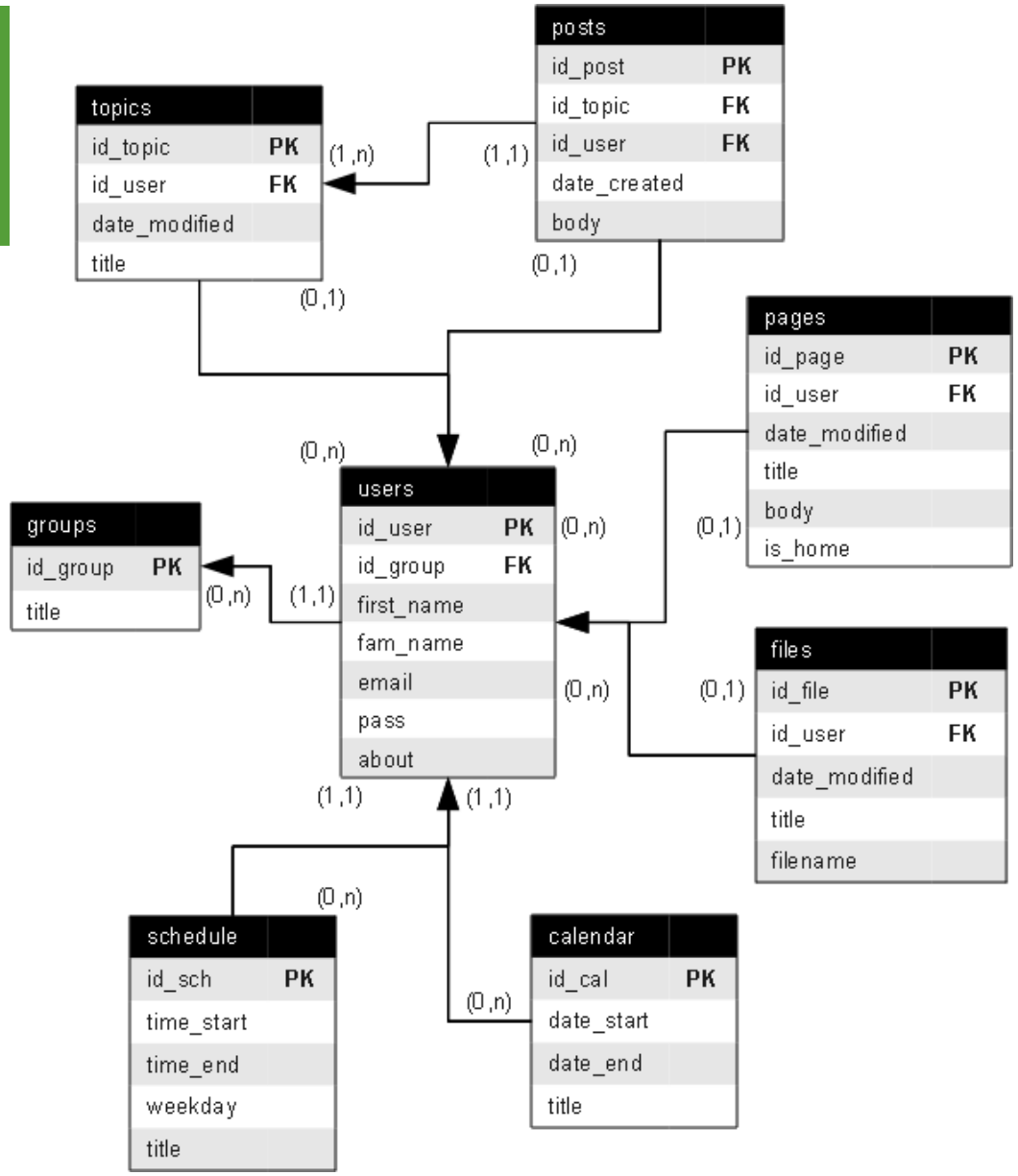
The dataset of physical objects could start with:

| 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 |
| ... | ... | ... | ... | ... | ... |

[Wooptoo, "Entity - relationship model." Wikimedia]

# FROM OBJECTS AND ATTRIBUTES TO DATA

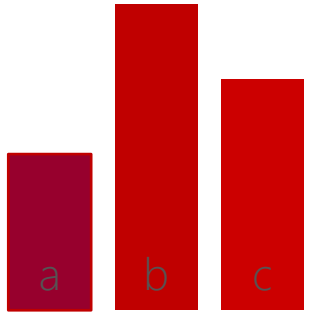
In practice, more complex **databases** are used, for a variety of reasons that we briefly discuss at a later stage.



# ANALYTICS MODES

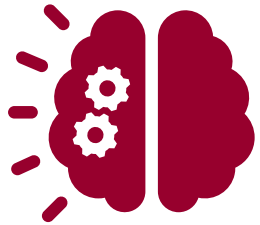
Looking Backwards

## Descriptive




Show **what** happened

## Diagnostic



Explain **why** something happened

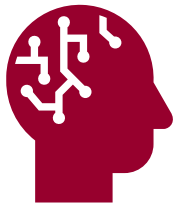
Easy

Difficult

## Predictive

Guess **what will** happen

## Prescriptive



Suggest **what should** happen

Looking Ahead

# THE “ANALYTICAL” METHOD

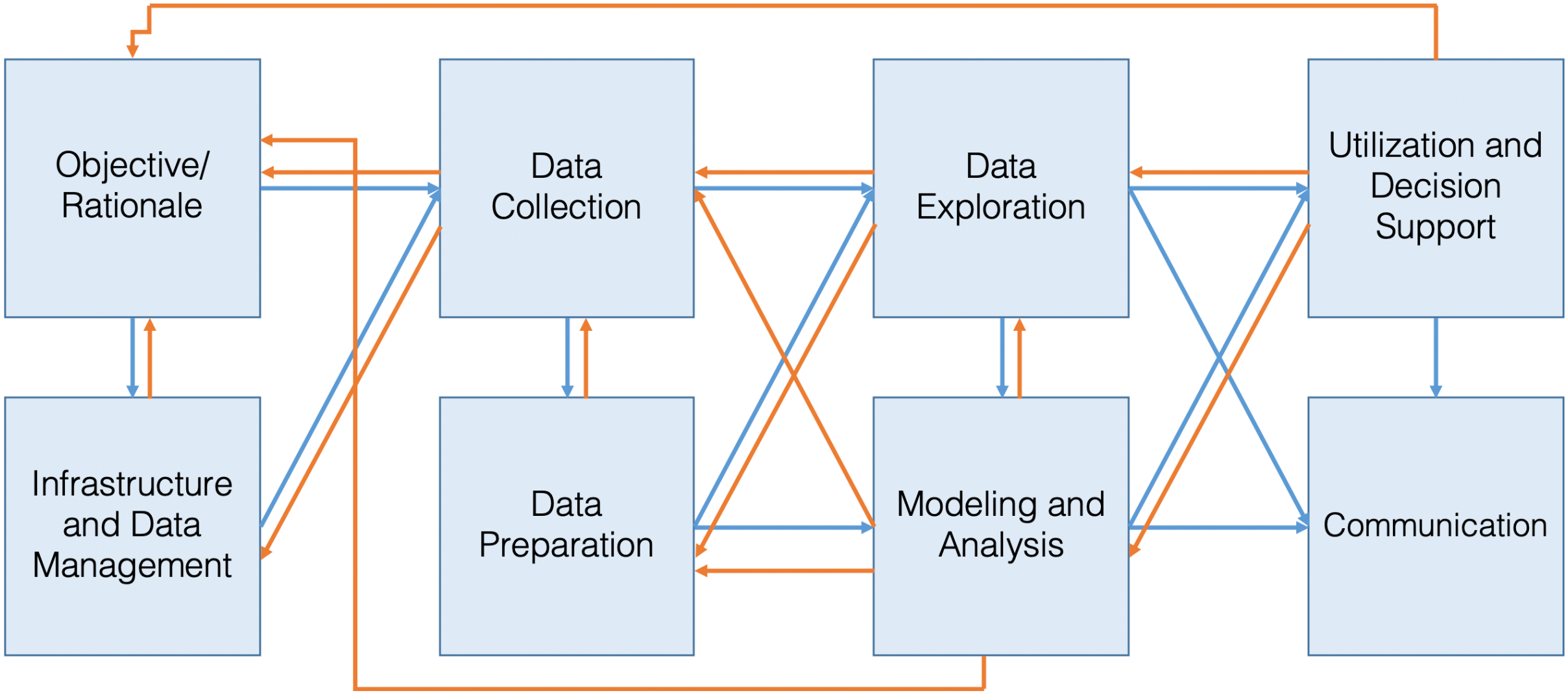
As with the **scientific method**, there is a “step-by-step” guide to data analysis:

- statement of objective
- data collection
- data clean-up
- data analysis/analytics
- dissemination
- documentation

Notice that **data analysis** only makes up a small segment of the entire flow.

In practice, the process is quite often **messy**, with steps added in and taken out of the sequence, repetitions, re-takes, etc.

Surprisingly, it tends to work… when **conducted correctly**.


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

# THE "ANALYTICAL" METHODS

In practice, data analysis is often corrupted by:

- lack of clarity
- mindless rework
- blind hand-off to IT
- failure to iterate

All approaches have a common core

- data science projects are **iterative**
- (often) **non-sequential**.

Helping stakeholders recognize this **central truth** makes it easier for data scientists to:

- plan the **data science process**
- obtain **actionable insights**

**Take-away:** there is a lot to consider in advance of modeling and analysis

- **data analysis is not just about data analysis**.

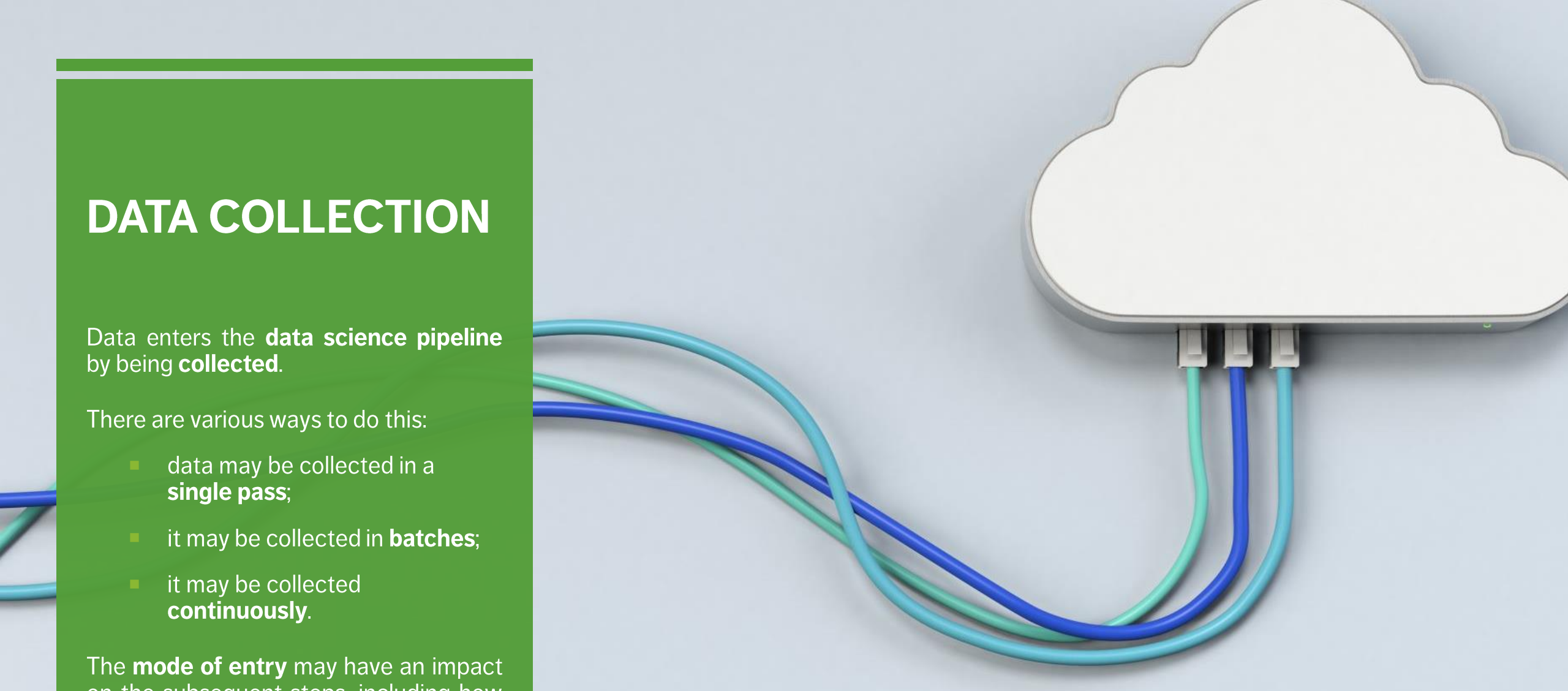
# DATA COLLECTION

Data enters the **data science pipeline** by being **collected**.

There are various ways to do this:

- data may be collected in a **single pass**;
- it may be collected in **batches**;
- it may be collected **continuously**.

The **mode of entry** may have an impact on the subsequent steps, including how frequently models, metrics, and other outputs are **updated**.



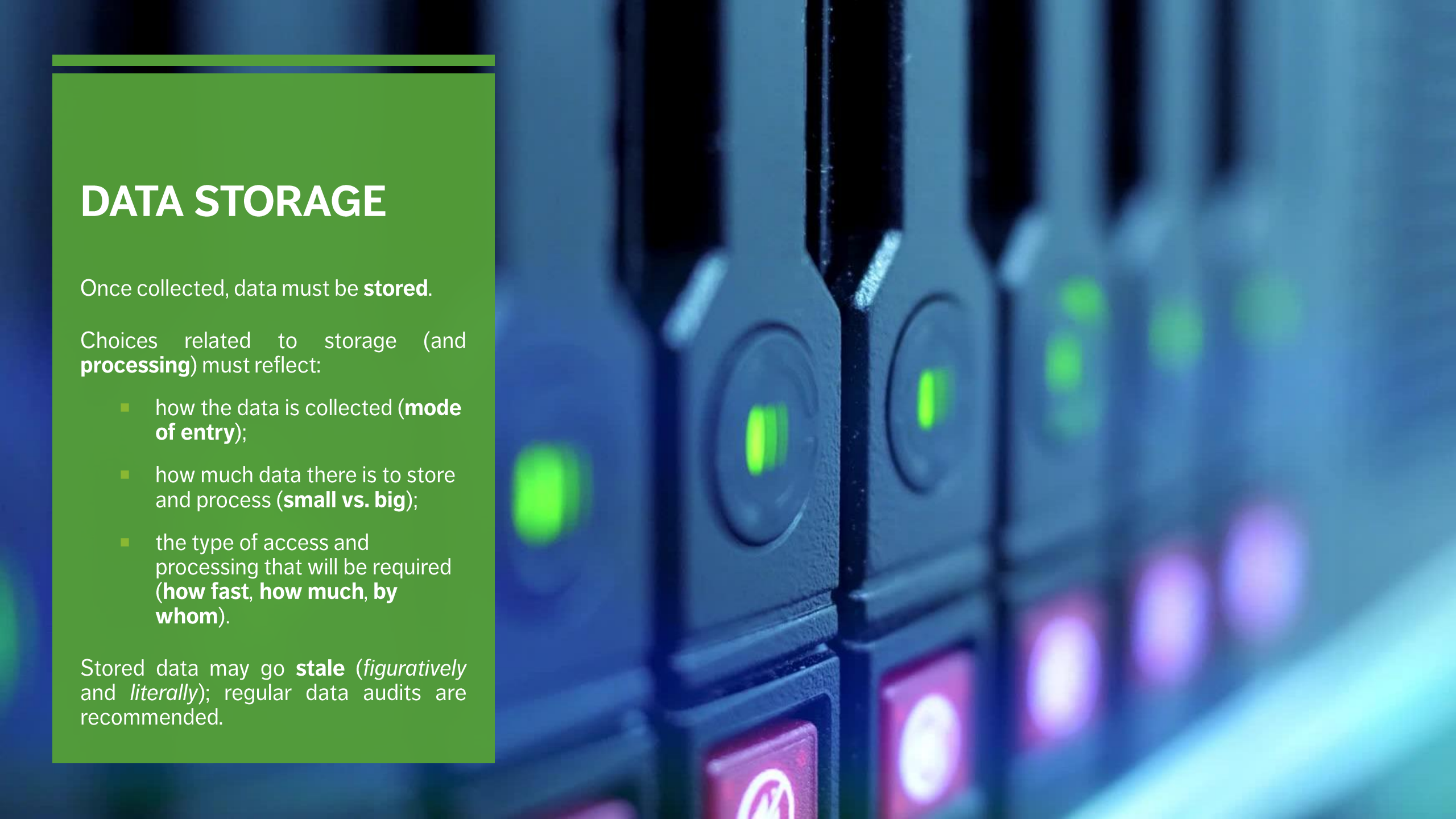
# DATA STORAGE

Once collected, data must be **stored**.

Choices related to storage (and **processing**) must reflect:

- how the data is collected (**mode of entry**);
- how much data there is to store and process (**small vs. big**);
- the type of access and processing that will be required (**how fast**, **how much**, **by whom**).

Stored data may go **stale** (*figuratively* and *literally*); regular data audits are recommended.



# DATA PROCESSING

The data must be **processed** before it can be analyzed.

The key point is that **raw data** has to be converted into a format that is **amenable to analysis**, by:

- identifying **invalid**, **unsound**, and **anomalous** entries
- dealing with **missing values**
- **transforming** the variables so that they meet the requirements of the selected algorithms

The **analysis** itself is almost anti-climactic: run the selected methods or algorithms on the processed data.

# MODELING

Data science teams should know:

- data cleaning
- descriptive statistics and correlation
- probability and inferential statistics
- regression analysis
- classification and supervised learning
- clustering and unsupervised learning
- anomaly detection and outlier analysis
- big data/high-dimensional data analysis
- stochastic modeling, etc.

These only represent a **small slice** of the analysis pie (see earlier slide).

No one analyst/data scientist could master all (or even a majority of them) at any moment, but that is one of the reasons why data science is a **team activity**.

# ASSESSMENT AND LIFE POST ANALYSIS

Before applying findings, we must first confirm that the model is reaching **valid conclusions** about the system.

Analytical processes are **reductive:** raw data is transformed into a small(er) **numerical summaries**, which we hope is **related** to the system of interest.

Data science methodologies include an **assessment phase**, an analytical sanity check: is anything **out of alignment?**

Beware the **tyranny of past success:** even if the analytical approach has been vetted and has given useful answers in the past, it may not always do so.

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

## Real World

## Model

**Theory**

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

# MODEL ASSESSMENT AND LIFE AFTER ANALYSIS

When an analysis or model is 'released into the wild', it often takes on a life of its own. When it inevitably ceases to be **current**, there may be little that data scientists can do to remedy the situation.

How do we determine if the current data model is:

- **out-of-date**?
- no longer **useful**?
- how long does it take a model to react to a **conceptual shift**?

Regular **audits** can be used to answer these questions.

# MODEL ASSESSMENT AND LIFE AFTER ANALYSIS

Data scientists rarely have full control over **model dissemination**.

- results may be misappropriated, misunderstood, shelved, or failed to be updated
- can conscientious analysts do anything to prevent this?

There is no easy answer: analysts should not only focus on the analysis, but also recognize opportunities that arises to **educate stakeholders** on the importance of these auxiliary concepts.

Due to **analytic decay**, the last step in the analytical process is not a **static dead end**, but an invitation to re-iterate to the beginning of the process.

# DATA PIPELINES (FIRST PASS)

In the **service delivery context**, the data analysis process is implemented as an **automated data pipeline** to enable automatic runs.

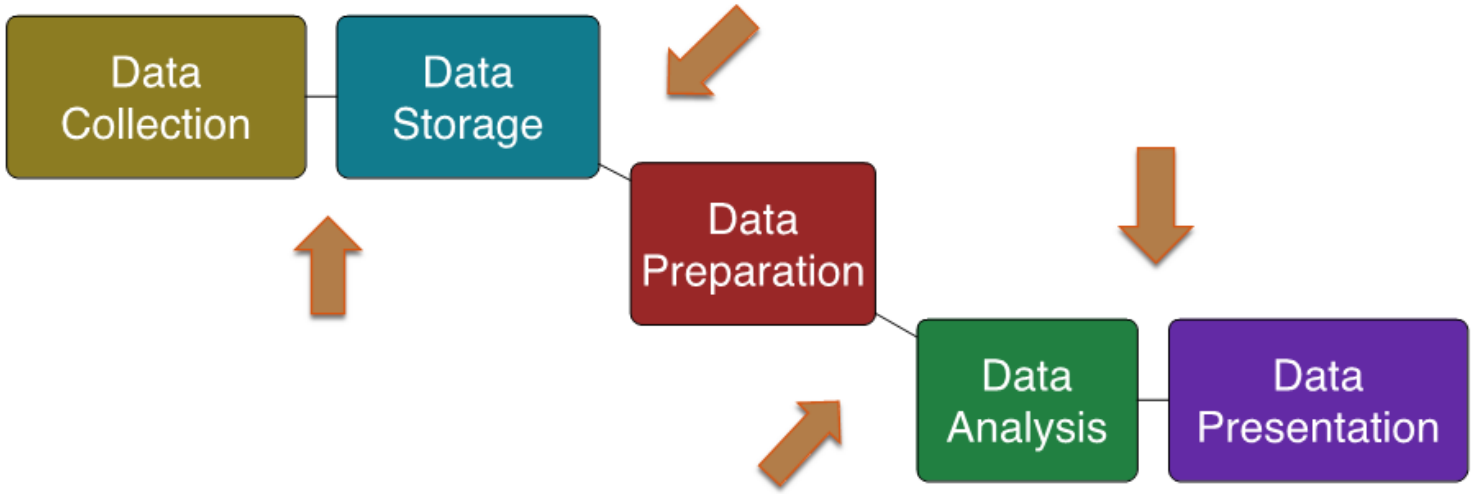
Data pipelines usually consist of 9 components (5 **stages** and 4 **transitions**):

- data collection
- data storage
- data preparation
- data analysis
- data presentation

# DATA PIPELINES (FIRST PASS)

Each components must be **designed** and then **implemented**.

Typically, at least one data analysis pass process must be done **manually** before the implementation is complete.



# EXERCISE

What does your financial data pipeline look like?

# SUPPLEMENTAL MATERIAL

## 7. DATA ANALYTICS

# QUANTITATIVE SKILLS

## Out-of-academia context:

- apply **quantitative methods** to (business) problems in order to obtain **actionable insight**
- difficult for any given individual to have expertise in **every** field of mathematics, statistics, computer science, data science, data engineering, etc.

With a graduate degree in math/stats, for instance:

- **expertise** in 2-3 areas
- **decent understanding** of related disciplines
- **passing knowledge** in various domains

Flexibility is an ally, perfectionism... only up to a point.

# SOFTWARE AND TOOLS

Modern quantitative work typically involves **programming** (or the use of point-and-click software, at the very least).

But programming languages **go in and out of style.**

It is important not just to understand the syntax of a particular language, but also how computer languages and computing infrastructure work in general.

**ALSO:** avoid getting caught up in programming/tool wars … they're more or less all functionally equivalent!

# SOFTWARE AND TOOLS

**Q:** At StatCan, R or SAS?

**A:** Not easy to answer as StatCan is in a slow transition period. The Agency is better equipped for SAS (with “Big Data” options, such as SAS Grid).

R is [...] not as ideal for large files (e.g., Census data), so it is not an option in such cases because it is still too slow (unless you have very powerful servers). But we would prefer to use the R packages, so it’s a dilemma.

**TL;DR:** R is our future, but SAS is still very much our present. In times of transition, **analysts/employees who know both are better positioned**.

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

# GBA+

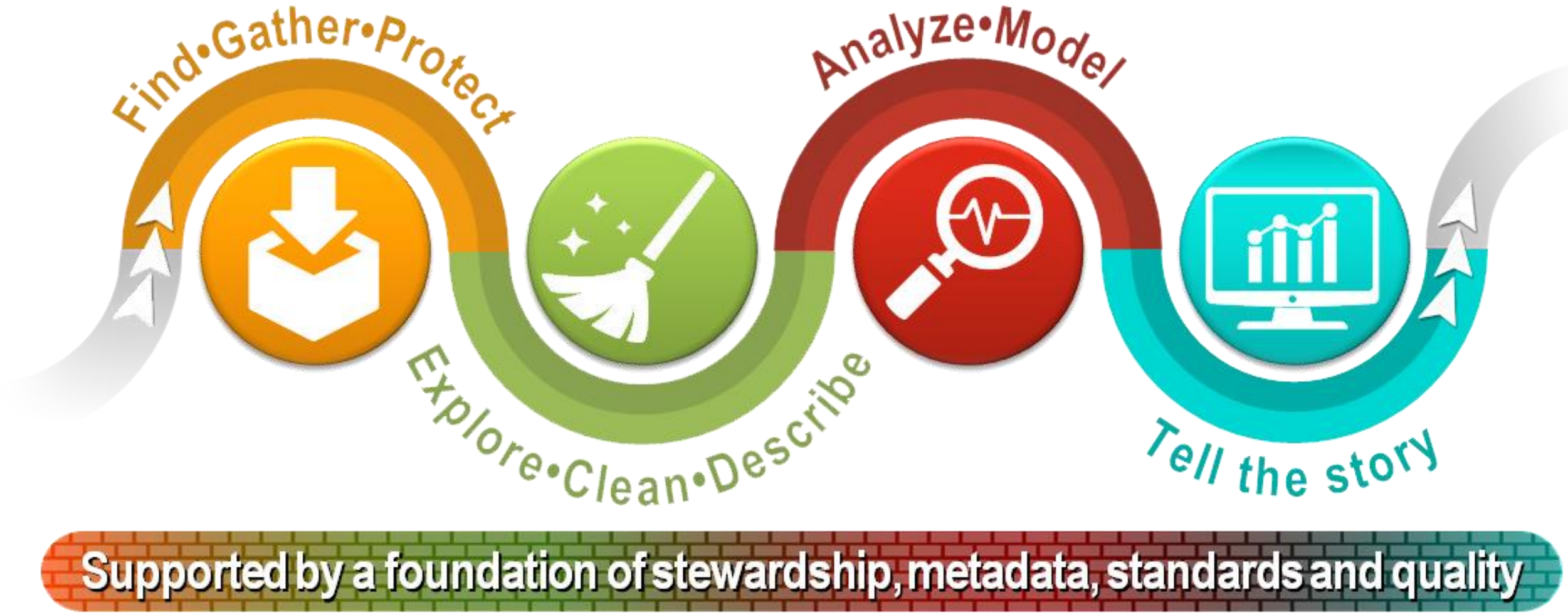
**Gender-Based Analysis Plus** is an analytical process used to assess how different gendered people may experience policies, programs and initiatives.

**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.

[StatCan, *Steps of the Data Journey*]



[CRISP-DM framework (left); Blitzstein and Pfister's data science workflow (right)]

# THE "ANALYTICAL" METHODS