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

CT ACADEMY | DATA ACTION LAB

5. DATA QUALITY

DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

DRIVERS FOR DATA QUALITY

GoC data governance policies/directives

Departmental data policies

Departmental data directives

Facilitates move towards GoC Open Data

But, it’s mostly just **good business practice**!

- Ability to utilize data more efficiently and effectively
- More timely access to data
- Increased confidence in decision making
- Preparation for advanced analytics such as AI and Machine Learning

“Improving data quality leads to improved decision-making throughout the enterprise. The more high-quality data you have, the more confidence you will have in your decision-making.”

WHAT IS DATA QUALITY?

Data quality (DQ) is the degree to which data **meets** or **exceeds** business requirements (i.e., the extent to which data is "fit for purpose").

It involves the **planning** and **implementation** of quality management techniques to **measure**, **assess**, and **improve** the fitness of data for use within an organization.

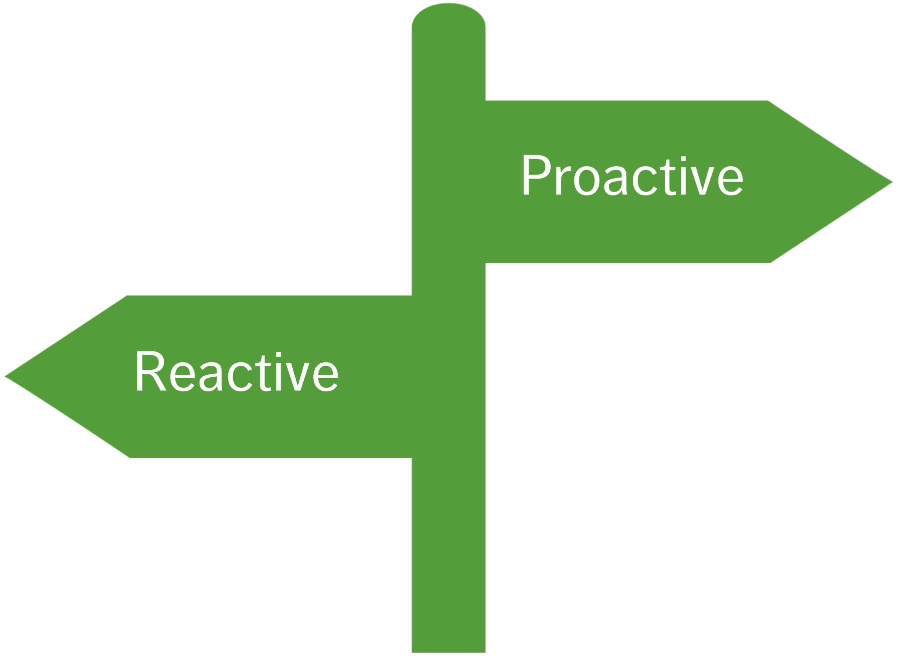
WHY IS DATA QUALITY IMPORTANT?

Data has **intrinsic value** due to its **information content**. The impacts of poor-quality data can include:

- Bad decisions, wasted resources, lowered performance chasing after issues.
- Escalating costs to remediate if issues are not caught early.
- Lack of trust in data for decision-making.
- Outdated and meaningless data bloat.
- Challenges with data sharing, integration and access to historical data.
- Poor-quality data is detrimental to analytics.

Data is the currency of **control:** we can’t control what we don’t know.

DATA QUALITY IS REACTIVE AND PROACTIVE



Quality Assurance

Preventing data quality issues ***before*** they impact operations and programs

Quality Control

Identifying and fixing data quality issues ***after*** they occur

A data quality program needs elements of both **control** and **assurance** to be fully effective.

DATA QUALITY DIMENSIONS

To implement data quality, we typically introduce categories called "**dimensions**".

These dimensions are **measurement attributes** of data, which can be individually **assessed**, **interpreted**, and **improved**.

We can use them to help build **data quality dashboards**.

They can also help us group **issues** and **risks**, helping us:

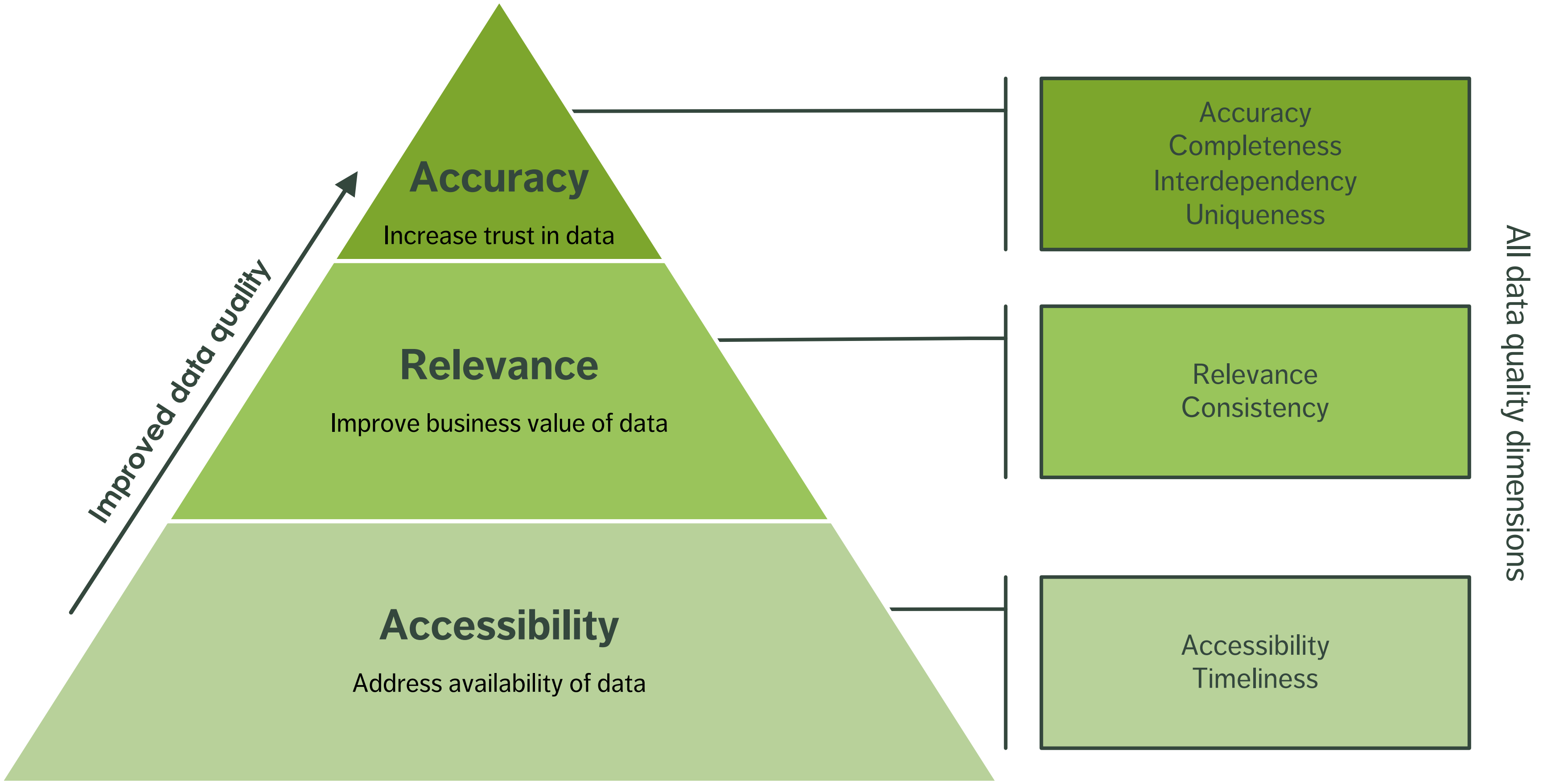
- perform **trend analyses**;
- identify underlying, **systemic issues**;
- set-up **data quality testing programs**, etc.

DATA QUALITY DIMENSIONS

Typical set of data quality dimensions

Accessibility	Business processes and consumers can access and use the data asset
Timeliness	Data values are sufficiently up-to-date for business processes and consumers needs
Relevance	Data asset is of value to and used by business processes and data consumers
Consistency	Data representations are the same within and across data assets and repositories
Accuracy	Data accurately represents a real-world entity, object, concept, etc.
Completeness	Data asset has no missing data values
Interdependency	Relationships between data elements are preserved within or across data assets
Uniqueness	Data representations are not duplicated within or across data assets

DATA QUALITY DIMENSIONS



data-action-lab.com

DATA QUALITY IS A PROCESS

There are three focus areas in addressing data quality:

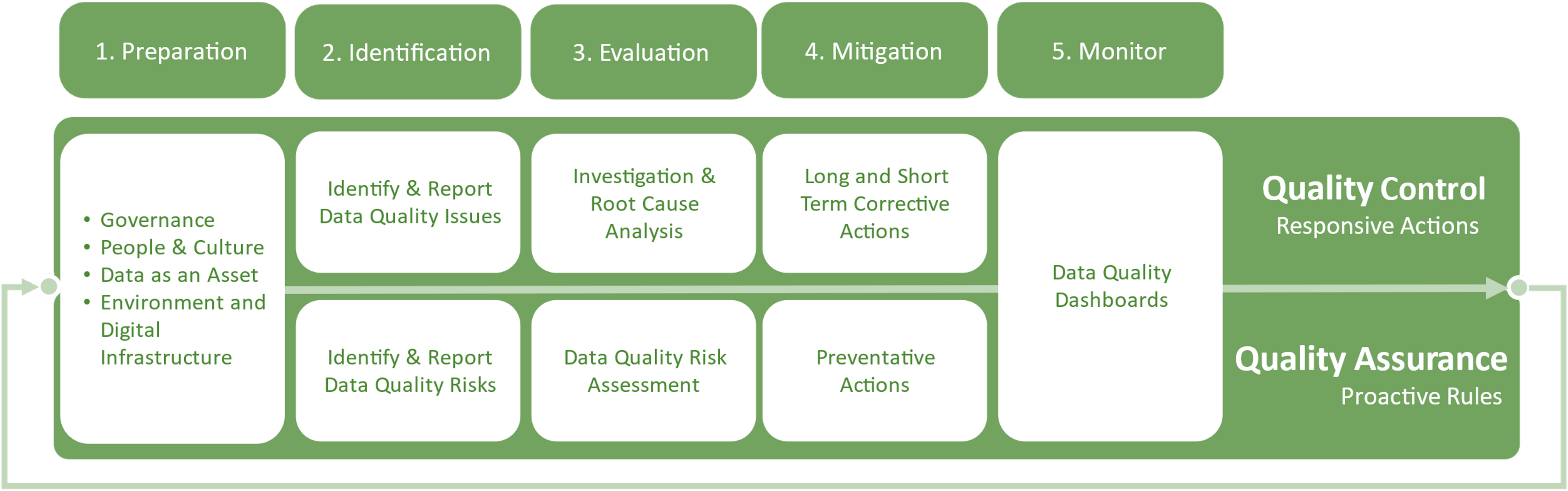
1. identify and mitigate **existing** data quality issues through quality control (e.g., testing a database to identify incorrect values then replacing them);
2. identify sources of high risk that could **create** quality issues and mitigate those risks through quality assurance (e.g., replacing a free form text field on a new software app with a dropdown list), and
3. track and **monitor** all known data quality issues and report them on a regular basis through quality monitoring (e.g., creating a list of all known issues and monitoring when they get fixed).

DATA QUALITY IS A PROCESS

We cannot implement a **data quality program** all at once, so we typically break it down to 5 stages:

1. preparation
2. issue and risk identification
3. issue and risk evaluation
4. issue and risk mitigation
5. ongoing monitoring

DATA QUALITY IS A PROCESS



1. Preparation
- Governance
- People & Culture
- Data as an Asset
- Environment and Digital Infrastructure

2. Identification

Identify & Report Data Quality Issues

Identify & Report Data Quality Risks

3. Evaluation

Investigation & Root Cause Analysis

Data Quality Risk Assessment

4. Mitigation

Long and Short Term Corrective Actions

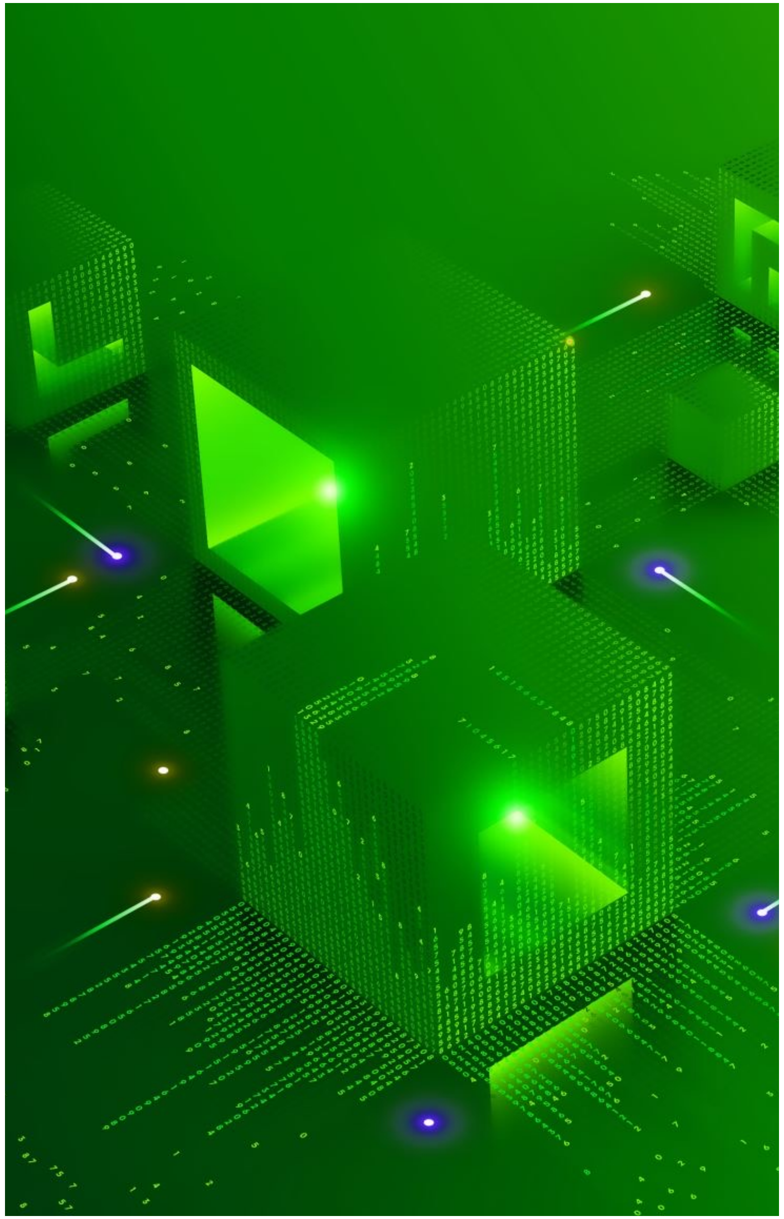
Preventative Actions

5. Monitor

Data Quality Dashboards

Quality Control
Responsive Actions

Quality Assurance
Proactive Rules



STAGE 1: PREPARATION

We can improve data quality by implementing programs such as:

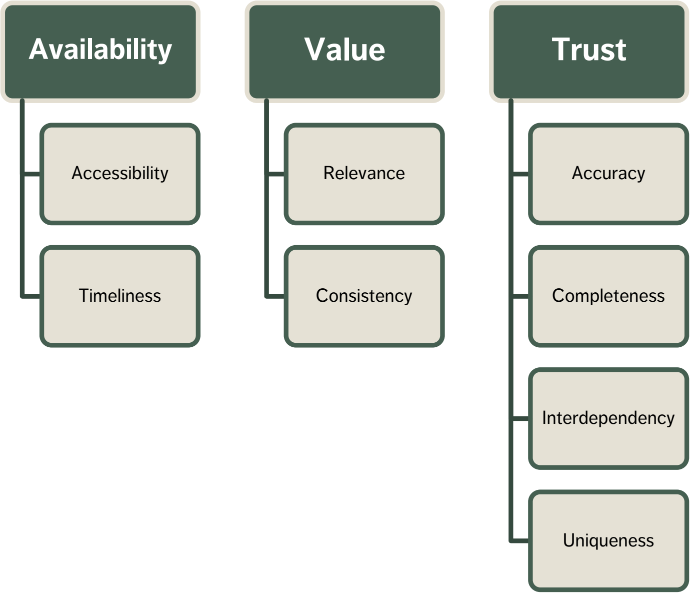
- **People & Culture** (data literacy, culture, and communication)
- **Environment & Digital Infrastructure** (tools, data asset catalogue)
- **Data Management** (metadata, reference/master data, dimensions & rules)
- **Governance** (roles & responsibilities, DQ planning, process definition)

Although it isn't critical to have all the above activities in place before starting on Data Quality, they do make a significant impact on the effectiveness of all DQ activities.

STEP 2: IDENTIFICATION

The second step in the process is to identify **data quality issues**. Currently-existing issues are called **data quality non-conformances**; issues that are yet to appear are known as **data quality risks**.

- DQ dimensions identify data attributes that can be used to measure data quality (often defined in an organization’s **Data Quality Framework**).
- Business rules define the **business requirements** for data and how **data quality tests** are performed.
- Data quality metrics track the results of these tests over time.



Availability
- Accessibility
- Timeliness

Value
- Relevance
- Consistency

Trust
- Accuracy
- Completeness
- Interdependency
- Uniqueness

STAGE 2: IDENTIFICATION METHODS

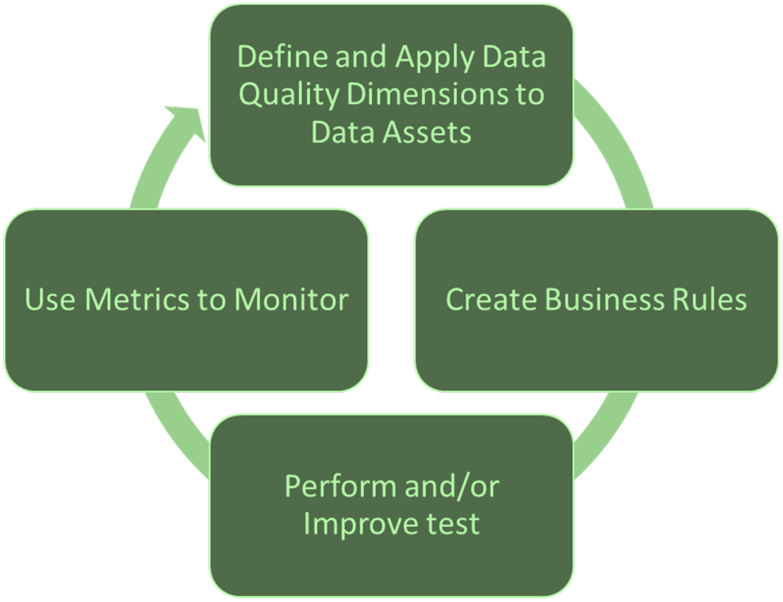
Methods for identifying **DQ non-conformances** and **DQ risks** include:

Quality Control

- DQ testing using software
- systems, process, and procedure auditing
- data consumer feedback

Quality Assurance

- creation of risk register



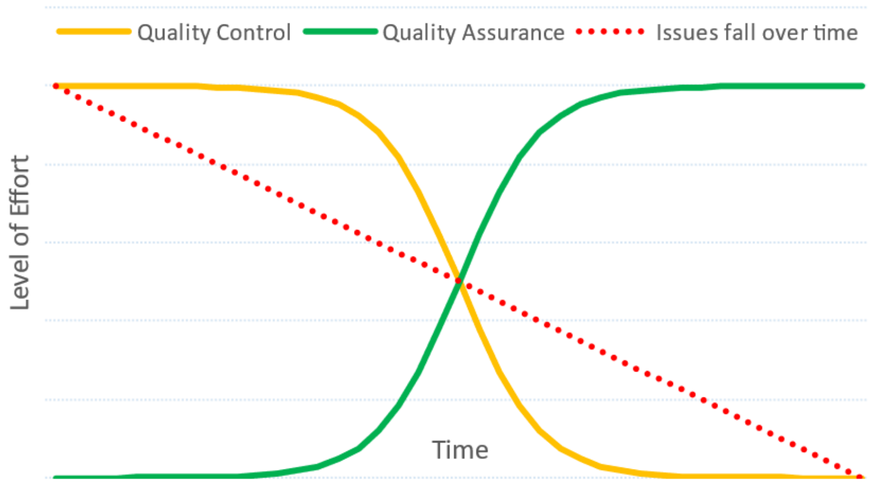
Define and Apply Data Quality Dimensions to Data Assets

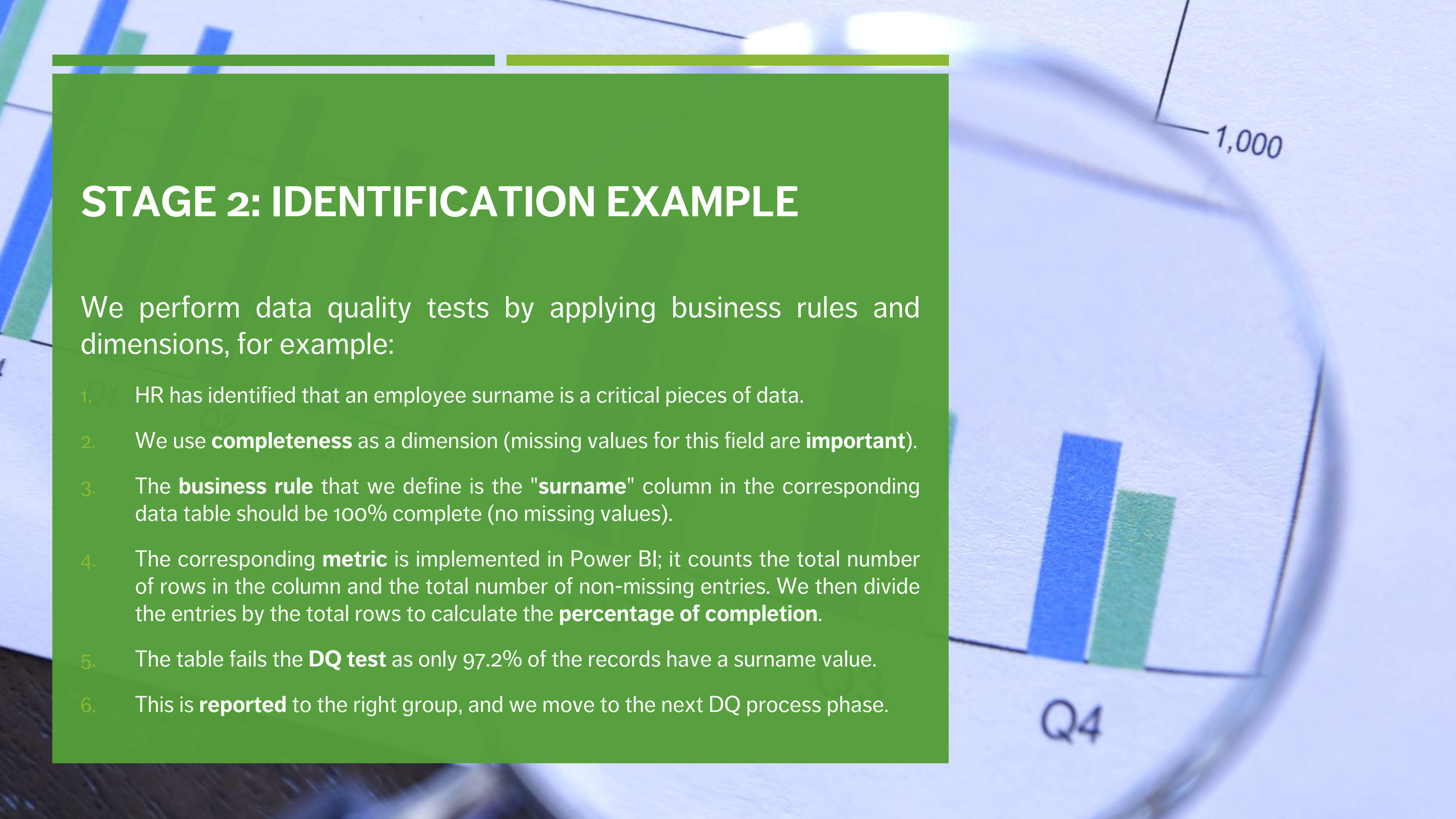
Create Business Rules

Perform and/or Improve test

Use Metrics to Monitor

Quality Assurance vs Quality Control over time



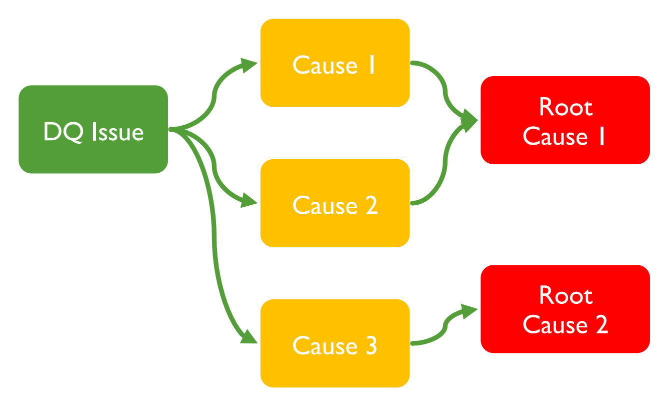


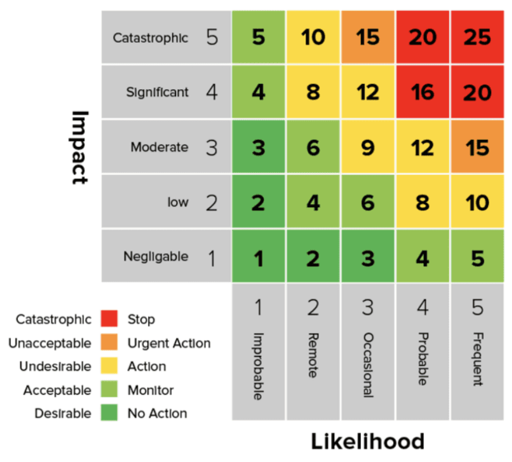
STAGE 2: IDENTIFICATION EXAMPLE

We perform data quality tests by applying business rules and dimensions, for example:

1. HR has identified that an employee surname is a critical pieces of data.
2. We use **completeness** as a dimension (missing values for this field are **important**).
3. The **business rule** that we define is the "**surname**" column in the corresponding data table should be 100% complete (no missing values).
4. The corresponding **metric** is implemented in Power BI; it counts the total number of rows in the column and the total number of non-missing entries. We then divide the entries by the total rows to calculate the **percentage of completion**.
5. The table fails the **DQ test** as only 97.2% of the records have a surname value.
6. This is **reported** to the right group, and we move to the next DQ process phase.

STAGE 3: EVALUATION





Impact		5	10	15	20	25
Catastrophic	5	5	10	15	20	25
Significant	4	4	8	12	16	20
Moderate	3	3	6	9	12	15
low	2	2	4	6	8	10
Negligible	1	1	2	3	4	5
		1 Improbable	2 Remote	3 Occasional	4 Probable	5 Frequent

Likelihood

Catastrophic	🟥	Stop
Unacceptable	🟧	Urgent Action
Undesirable	🟨	Action
Acceptable	🟩	Monitor
Desirable	🟩	No Action

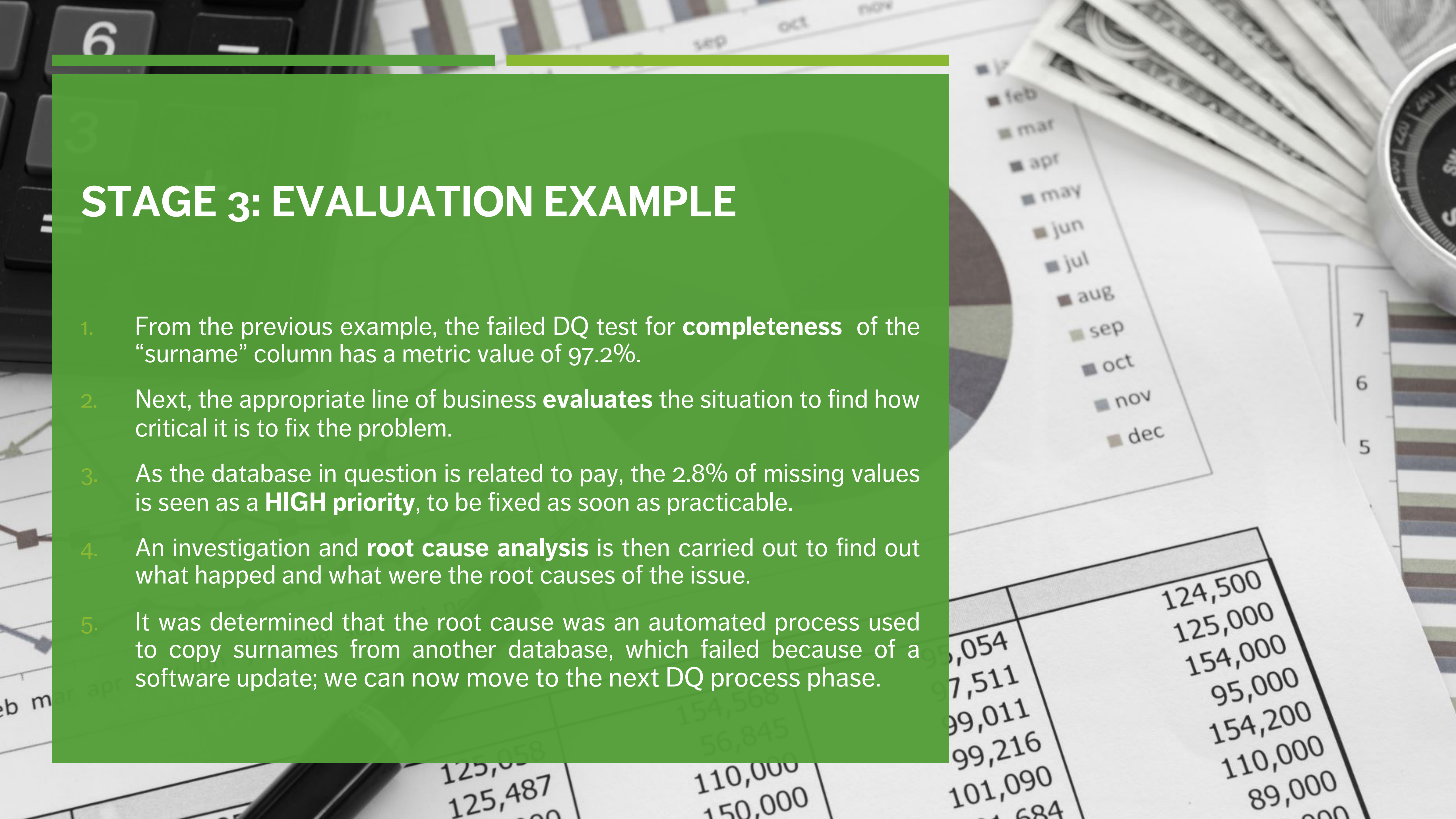
Once a DQ issue has been identified, we must **evaluate** it to prioritize **mitigation activities**. Evaluation methods typically include:

Quality Control

- formally investigate DQ Issues
- perform a **root cause analysis** to evaluate causes not symptoms

Quality Assurance

- evaluate and prioritize risks regarding impact



STAGE 3: EVALUATION EXAMPLE

1. From the previous example, the failed DQ test for **completeness** of the "surname" column has a metric value of 97.2%.
2. Next, the appropriate line of business **evaluates** the situation to find how critical it is to fix the problem.
3. As the database in question is related to pay, the 2.8% of missing values is seen as a **HIGH priority**, to be fixed as soon as practicable.
4. An investigation and **root cause analysis** is then carried out to find out what happed and what were the root causes of the issue.
5. It was determined that the root cause was an automated process used to copy surnames from another database, which failed because of a software update; we can now move to the next DQ process phase.

STAGE 4: MITIGATION

Once a DQ issue has been evaluated it may then require **mitigation** (fixing the issue). The actual fix will vary depending on the issue itself, but they fall into different categories.

Quality Control

- short term corrective actions to immediately fix the issue
- long term corrective actions ensure the issue does not recur

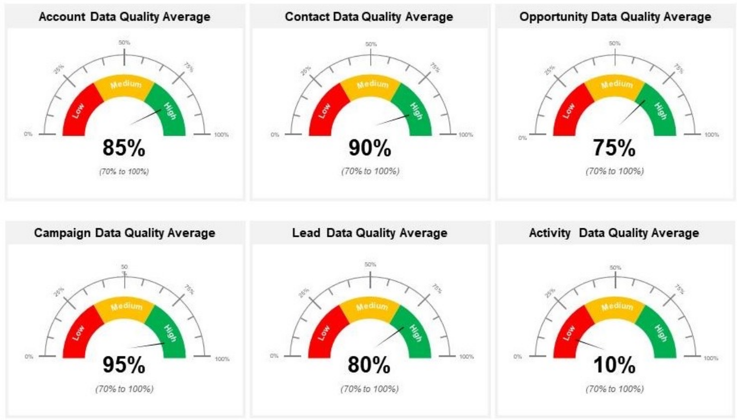
Quality Assurance

- preventative actions ensure that high risks do not turn into DQ issues

STAGE 4: MITIGATION EXAMPLE

From the previous example the following mitigation activities may be carried out.

1. A **short-term corrective action** is to re-run the data transfer process manually to ensure that the column is updated and 100% complete.
2. A **long-term corrective action** is to implement an automated DQ test once the transfer is complete to ensure that all records had been transferred between data assets.
3. We can now move to the next DQ process phase.


Account Data Quality Average
85%
(70% to 100%)
Contact Data Quality Average
90%
(70% to 100%)
Opportunity Data Quality Average
75%
(70% to 100%)
Campaign Data Quality Average
95%
(70% to 100%)
Lead Data Quality Average
80%
(70% to 100%)
Activity Data Quality Average
10%
(70% to 100%)
Low
Medium
High

STAGE 5: MONITOR

It is best practice to **monitor** DQ on an **ongoing** basis.

Quality Control & Quality Assurance

- status of DQ issues (non-conformances and risks)
- status of DQ investigations and root cause analysis
- status of internal quality audits
- status of corrective actions and preventative actions

STAGE 5: MONITOR EXAMPLE

As an issue had occurred in the “surname” column, the following items were recorded and included in the **ongoing monitoring** of DQ:

1. the date the issue was identified;
2. the criticality of the issue (high priority);
3. the type of issue (“completeness” non-conformance);
4. the actual test result (97.2%);
5. the date the short-term corrective action was completed;
6. the date the long-term corrective action was completed, and
7. a measure of the effectiveness of the long-term corrective action (“highly effective”).

As the line of business continues to monitor non-conformances and risks, a profile of the health of the data asset is created which shows **improvement over time** (?)

This is replicated for **all onboarded assets** for aggregated monitoring & reporting.

“Any substantial improvement must come from action on the system and is the responsibility of management. Wishing and pleading and begging the workers to do better is totally futile.”

(W. Edwards Deming, *Out of the Crises*)

DATA QUALITY IS ENACTED BY PEOPLE

Role	Description	Focus on Data Quality
Chief Data Officer	Responsible for all data related activities at department	Accountable for DQ program
Branch/Regional/ Program Heads	Responsible for Branch/Regional/Programs	Participation in the DQ program
Data Trustee	Strategically manages data assets and ensures compliance with data-related strategies, regulations, policies, directives	Creates a proactive, risk-based approach to the reduction of DQ issues
Data Steward	Advises, enacts, and enforces data policies and standards	Performs testing, risk assessment, investigations and auditing. Implements ongoing monitoring of DQ
Data Custodian	Ensures safe custody and integrity of hosted data	Provides technical support for corrective and preventative actions
Data Contributor	Ensures the data they provide aligns with technical and business policies, procedures, and standards	Ensures quality of data prior to inclusion in data asset
Data Consumer	Ensures usage of data supports all objectives and mandates	Reports on data issues found when using data

6. ASKING QUESTIONS & NON-TECHNICAL ASPECTS OF DATA WORK

DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

ASKING THE RIGHT QUESTIONS

Data science is 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 or useless.

ROADMAP TO FRAMING QUESTIONS

Understand the problem (opportunity vs. problem)

What initial assumptions do I have about the situation?

How will the results be used?

What are the risks and/or benefits of answering this question?

What stakeholder questions might arise based on the answer(s)?

Do I have access to the data necessary to answering this question?

How will I measure my ‘success’ criteria?

YES/NO TRAP

Examples of **bad** questions:

- Are our revenues **increasing** over time? **Has it** increased year-over-year?
- Are most of our customers from **this demographic**?
- **Does this project have** valuable ambitions to the broader department?
- **How great** is our hard-working customer success team?
- How often do you **triple check** your work?

Examples of **good** questions:

- What's the **distribution** of our revenues over the past three months?
- Where are our **top 5** high-spending cohorts from?
- What are the **different benefits** of pursuing this project?
- What are **three good** *and* **bad traits** of our customer success team?
- Do you **tend to** do quality assurance testing on your deliverables?

QUESTION AUDIT CHECKLIST

1. Did I avoid creating any **yes/no** questions?
2. Would **anyone** in my team/department understand the question irrespective of their backgrounds?
3. Does the question need more than one sentence to express?
4. Is the question '**balanced**' – scope is not too broad that the question will never truly be answered, or too small that the resulting impact is minimal?
5. Is the question being **skewed** to what may be easier to answer for my/my team's particular skillset(s)?

MULTIPLE I'S APPROACH TO DATA WORK

Technical and quantitative proficiency (or expertise) is **necessary** to do good quantitative work *in the real world*, but it is **not sufficient** – optimal real-world solutions may not always be the optimal academic or analytical solutions.

This might be the biggest surprise for those transitioning out of academia.

What works for one person, one job application, one project, one client, etc. may not work for another – **beware the tyranny of past success**!

The focus of quantitative work must include the delivery of **useful analyses/ products** (Multiple "I"s).

MULTIPLE I’S APPROACH TO DATA WORK

- **intuition**
  understanding context
- **initiative**
  establishing an analysis plan
- **innovation**
  new ways to obtain results, if required
- **insurance**
  trying multiple approaches
- **interpretability**
  providing explainable results
- **inquisitiveness**
  not only asking the same questions repeatedly
- **integrity**
  staying true to objectives and results
- **independence**
  self-learning and self-teaching
- **interactions**
  strong analyses through teamwork
- **interest**
  finding and reporting on interesting results
- **intangibles**
  thinking “outside the box”
- **insights**
  providing actionable results

MULTIPLE I’S APPROACH TO DATA WORK

Prospective employees/analysts are not solely gauged on technical know-how, but also on the ability to **contribute positively** to the workplace/project:

- communication
- team work and multi-disciplinary abilities
- social niceties and flexibility
- non-technical interests

Employers rarely chose robots when human beings are available; stakeholders are more likely to accept data recommendations from **well-rounded people**.

You should also evaluate eventual employers/clients on these axes.

ROLES AND RESPONSIBILITIES

A data analyst or a data scientist (in the **singular**) is unlikely to get meaningful results – there are simply too many moving parts to any data project.

Successful projects require **teams** of highly-skilled individuals who understand the **data**, the **context**, and the **challenges**.

Team *size* could vary from a few to several dozens; typically easier to manage small-ish teams (with 1-4 members, say, with **role overlaps**).

Domain Experts / SMEs

- are authorities in a particular area or topic
- guide team through unexpected complications and knowledge gaps

ROLES AND RESPONSIBILITIES

Project Managers / Team Leads

- understand the process enough to recognize whether what is being done makes sense
- provide realistic estimates of the time and effort required to complete tasks
- act as intermediary between team and clients/stakeholders
- responsible for project deliverables.

Data Translators

- have a good grasp on the data and the data dictionary
- help SMEs transmit the underlying context to the data science team

Data Engineers / Database Specialists

- work with clients and stakeholders to acquire useable data sources
- may participate in the analyses, but are not necessarily specialists

ROLES AND RESPONSIBILITIES

Data Analysts

- clean and process data
- prepare initial visualizations
- have a decent understanding of quantitative methods (at most 1 area of expertise)
- conduct preliminary analyses

Data Scientists

- work with processed data to build sophisticated models
- focus on actionable insights
- have a sound understanding of algorithms/quantitative methods (2 or 3 areas of expertise)
- can apply them to a variety of data scenarios
- can be counted on to catch up on new material quickly

ROLES AND RESPONSIBILITIES

Computer Engineers

- design and build computer systems and pipelines
- are involved in software development and deployment of data science solutions

AI/ML Quality Assurance/Quality Control Specialists

- design testing plans for solutions that implement AI/ML models
- help the team determine whether the models are able to learn

Communication Specialists

- communicate actionable insights to managers, policy analysts, decision-makers, stake holders
- may participate in the analyses, but are not necessarily specialists (often data translators)
- keep abreast of popular accounts of quantitative results and developments

THE DIGITAL/ANALOG DATA DICHOTOMY

Humans have been collecting data for a long time; J.C. Scott argues that data collection was a major enabler of the modern nation-state.

For most of the history of data collection, we have lived in the **analogue world** (understanding grounded in continuous experience of **physical reality**).

Our data collection activities were the first steps towards a different strategy for understanding and interacting with the world.

Data leads us to conceptualize the world in a way that is **more discrete than continuous**..

THE DIGITAL/ANALOG DATA DICHOTOMY

Translating our experiences into numbers and categories, we create **sharper** and more definable boundaries than experience might suggest.

This discretization strategy leads to the **digital computer** (1;0), which is surprisingly successful at representing our physical world: the **digital world** takes on a reality as pervasive and important as the physical one.

This digital world is built on top of the physical world, but it **does not operate under the same set of rules:**

- in the physical world, the default is to **forget**; in the digital world, it is to **remember**;
- in the physical world, the default is **private**; in the digital world, the default is **public**;
- in the physical world, copying is **hard**; in the digital world, it is **easy**.

THE DIGITAL/ANALOG DATA DICHOTOMY

Digitization is making things that were **once hidden, visible; once veiled, transparent**.

Data scientists are scientists of the **digital world**. They seek to understand:

- the **fundamental principles of data**
- how these fundamental principles manifest themselves in different digital phenomena

Ultimately, data and the digital world are **tied to the physical world**. What is done with data has repercussions in the physical world; and it is crucial for data scientists to have a solid grasp of the fundamentals and context of data work before leaping into the tools and techniques that drive it forward.

ANALYSIS CHEAT SHEET

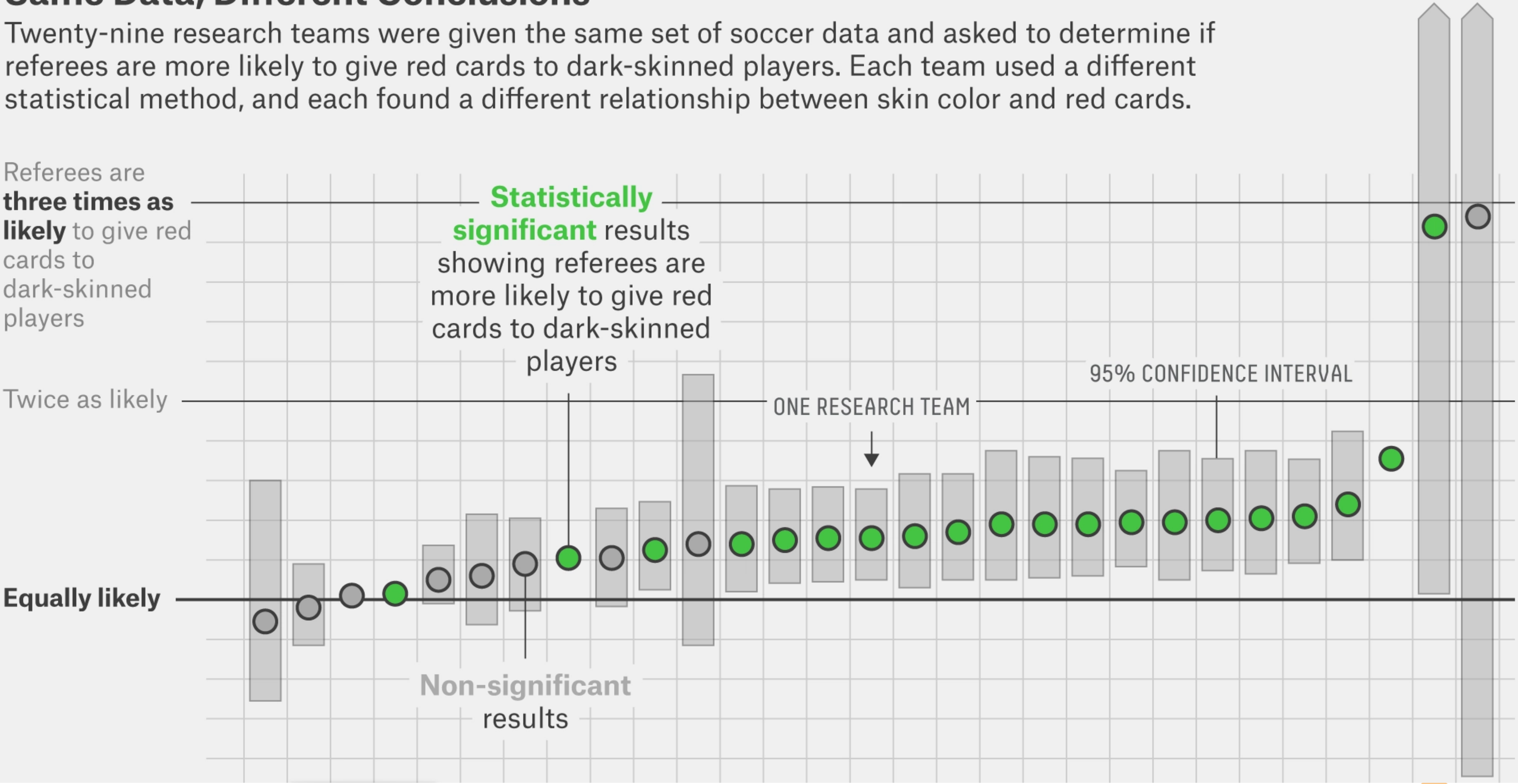
1. Business solutions are not always academic solutions.
2. Data and models don't always support the stakeholder's hopes, wants, needs.
3. Timely communication is key – externally and internally.
4. Data scientists need to be flexible (within reason), and willing and able to learn something new, quickly.
5. Not every problem calls for data science methods.
6. We should learn from both our good and our bad experiences.

ANALYSIS CHEAT SHEET

7. Manage projects and expectations.
8. Maintain a healthy work-life balance.
9. Respect the stakeholders, the project, the methods, and the team.
10. Data science is not about how smart we are; it is about how we can provide actionable insight.
11. When what the client wants can’t be done, offer alternatives.
12. “There ain’t no such thing as a free lunch.”

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



From: Science isn't broken - It's just a hell of a lot harder than we give it credit for. [Christie Aschwanden, 2015]

DATA IN THE NEWS

Here is a sample of headlines and article titles showcasing the growing role of **data science** (DS), **machine learning** (ML), and **artificial/augmented intelligence** (AI) in different domains of society.

While these demonstrate some of the functionality/capabilities of DS/ML/AI technologies, it is important to remain aware that new technologies are always accompanied by emerging social consequences (not always positive).

DATA IN THE NEWS

- “Robots are better than doctors at diagnosing some cancers, major study finds”
- “Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet”
- “Google AI claims 99% accuracy in metastatic breast cancer detection”
- “Data scientists find connections between birth month and health”
- “Scientists using GPS tracking on endangered Dhole wild dogs”
- “These AI-invented paint color names are so bad they’re good”
- “We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually.”
- “Math model determines who wrote Beatles’ "In My Life": Lennon or McCartney?”

DATA IN THE NEWS

- "Scientists use Instagram data to forecast top models at New York Fashion Week"
- "How big data will solve your email problem"
- "Artificial intelligence better than physicists at designing quantum science experiments"
- "This researcher studied 400,000 knitters and discovered what turns a hobby into a business"
- "Wait, have we really wiped out 60% of animals?"
- "Amazon scraps secret AI recruiting tool that showed bias against women"
- "Facebook documents seized by MPs investigating privacy breach"
- "Firm led by Google veterans uses A.I. to 'nudge' workers toward happiness"
- "At Netflix, who wins when it's Hollywood vs.the algorithm?"

DATA IN THE NEWS

- “AlphaGo vanquishes world’s top Go player, marking A.I.’s superiority over human mind”
- “An AI-written novella almost won a literary prize”
- “Elon Musk: Artificial intelligence may spark World War III”
- “A.I. hype has peaked so what’s next?”

Opinions on the topic are varied – to some, DS/ML/AI provide examples of **brilliant successes**, while to others it is the **dangerous failures** that are at the forefront.

What do you think?

Are you a glass half-full or glass half-empty sort of person when it comes to data and applications?

7. DATA ANALYTICS

DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

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.

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

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

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.

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**.

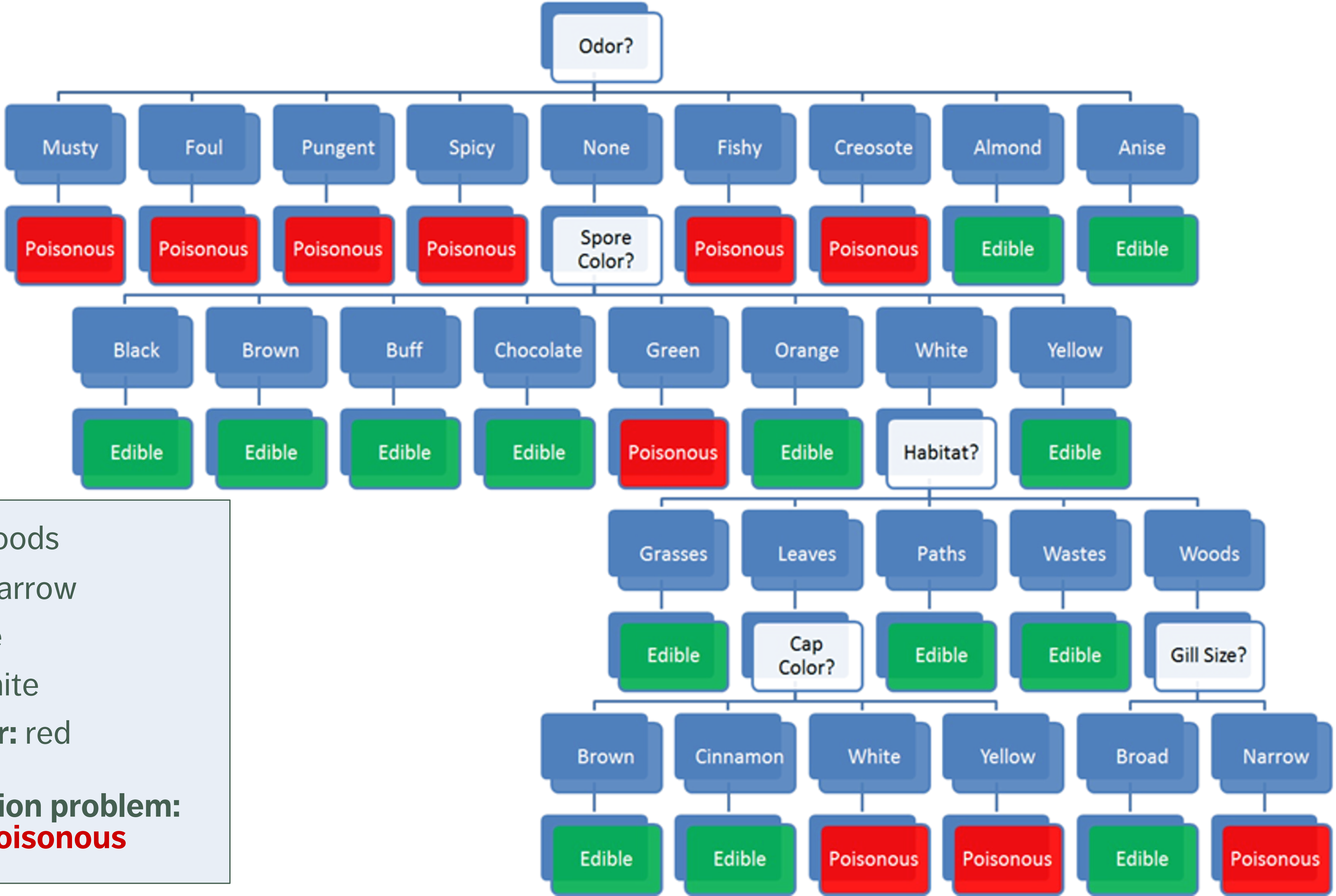
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

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

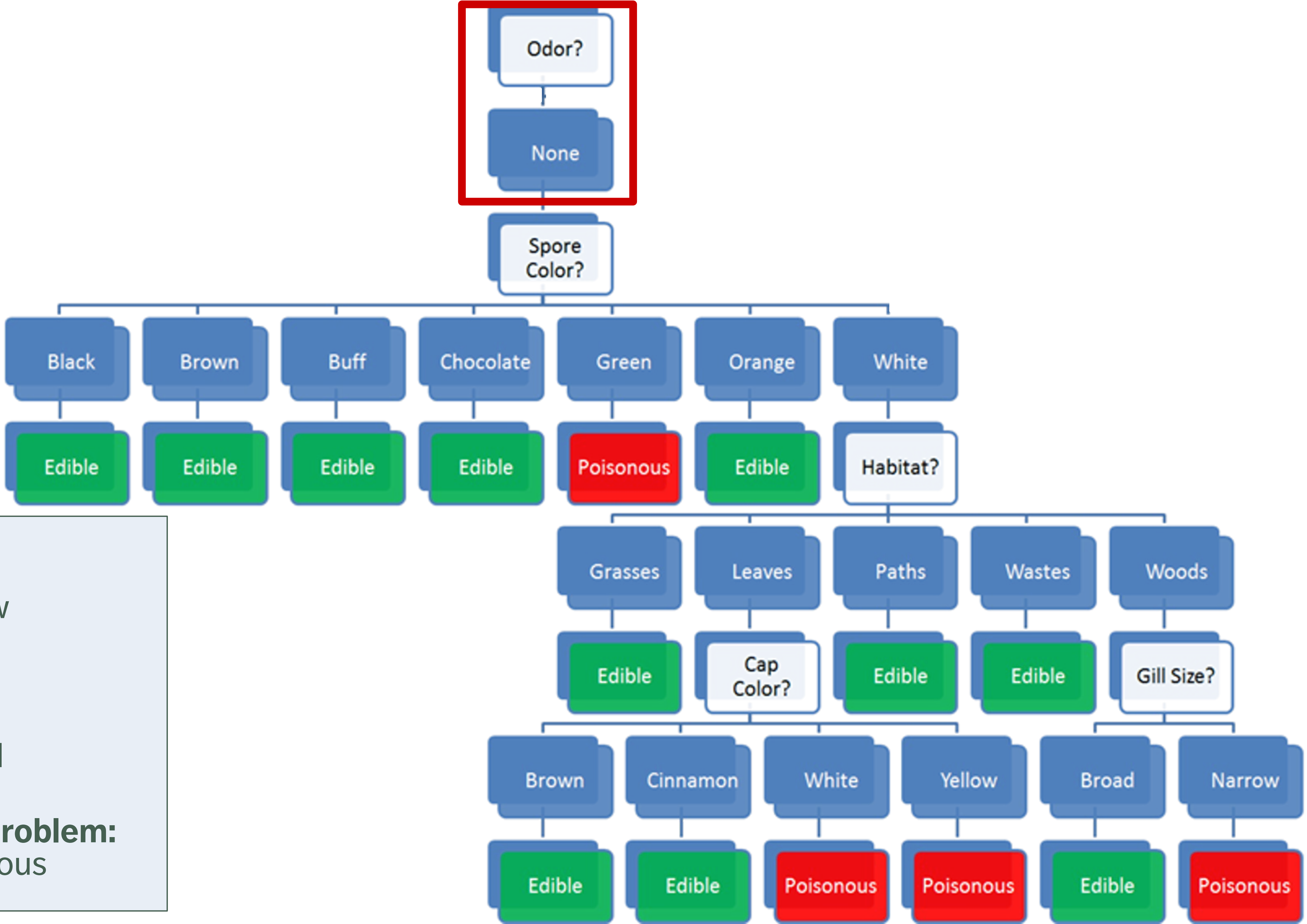
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or **poisonous**



Habitat: woods

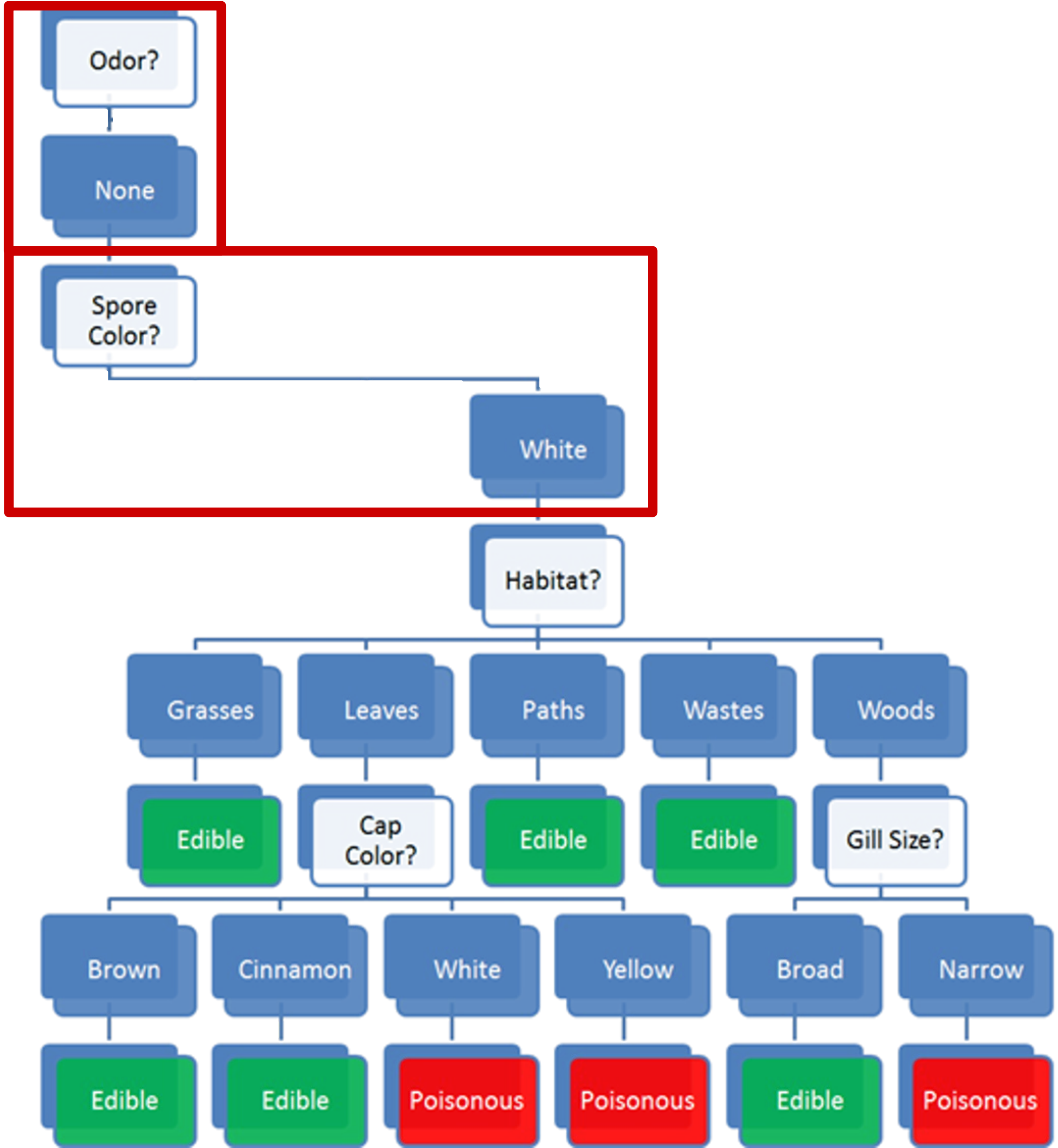
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous



Habitat: woods

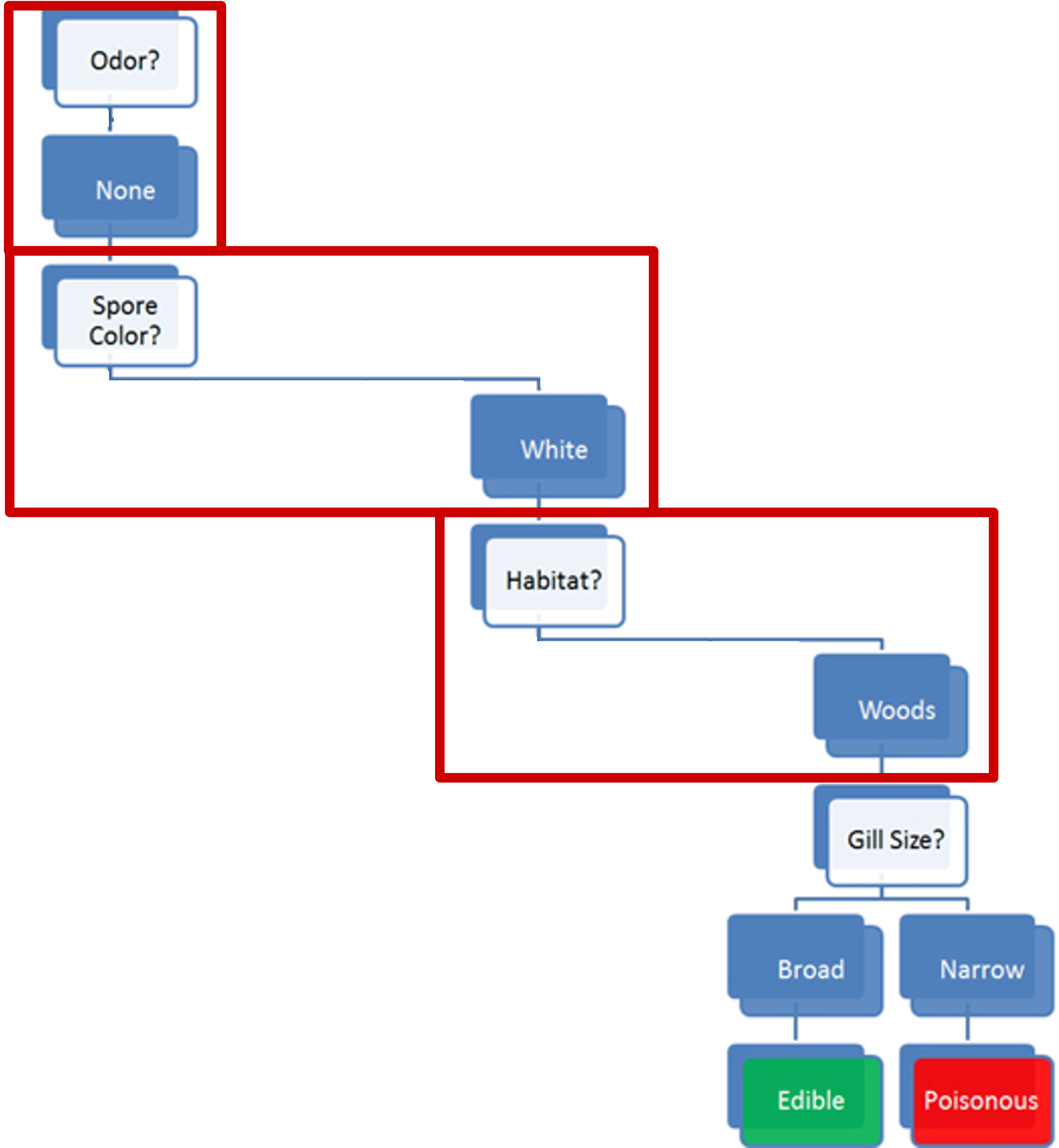
Gill Size: narrow

Odor: none

Spores: **white**

Cap Colour: red

Classification problem: edible or poisonous



Habitat: **woods**

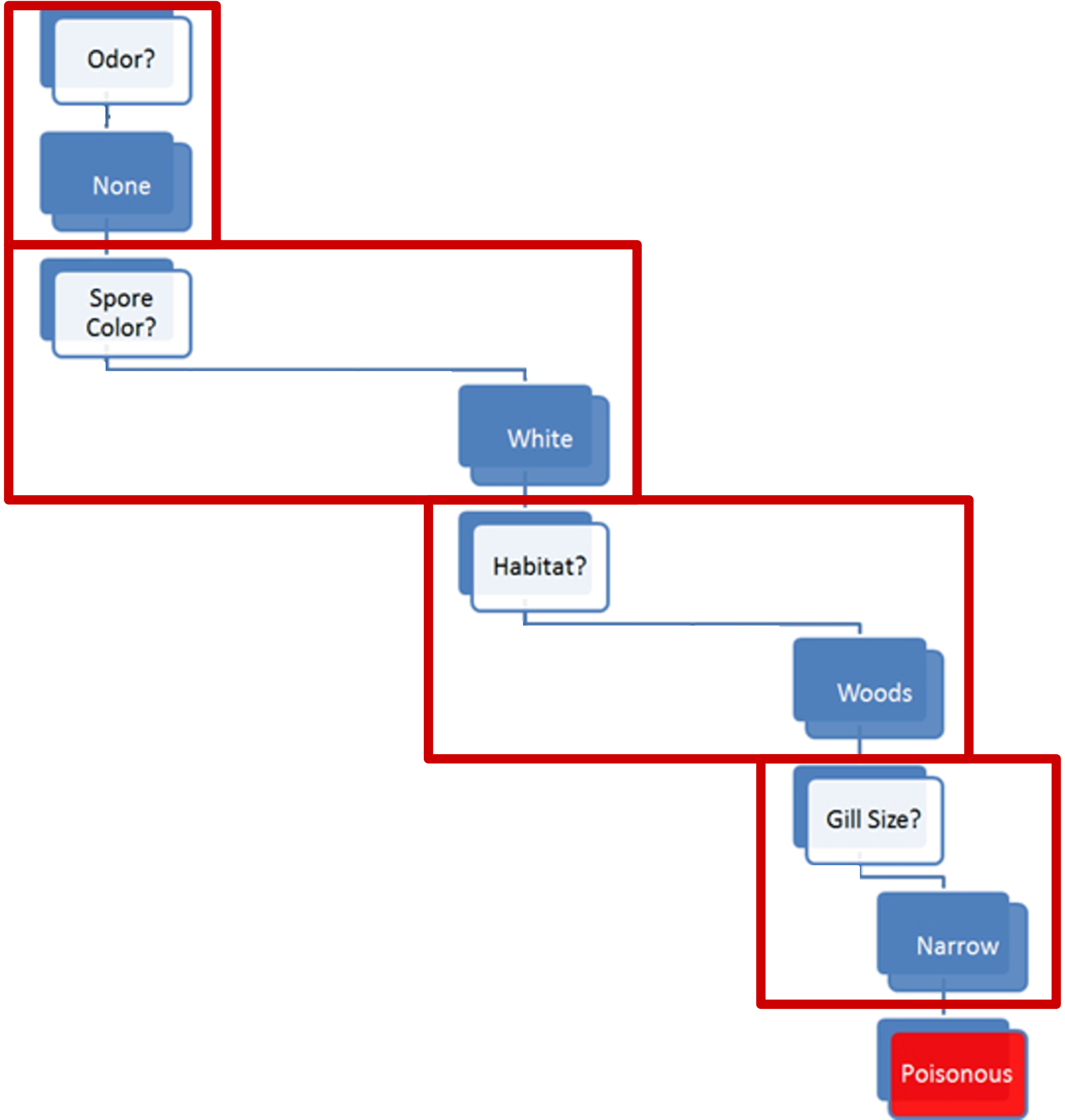
Gill Size: narrow

Odor: none

Spores: white

Cap Colour: red

Classification problem:
edible or poisonous



Habitat: woods

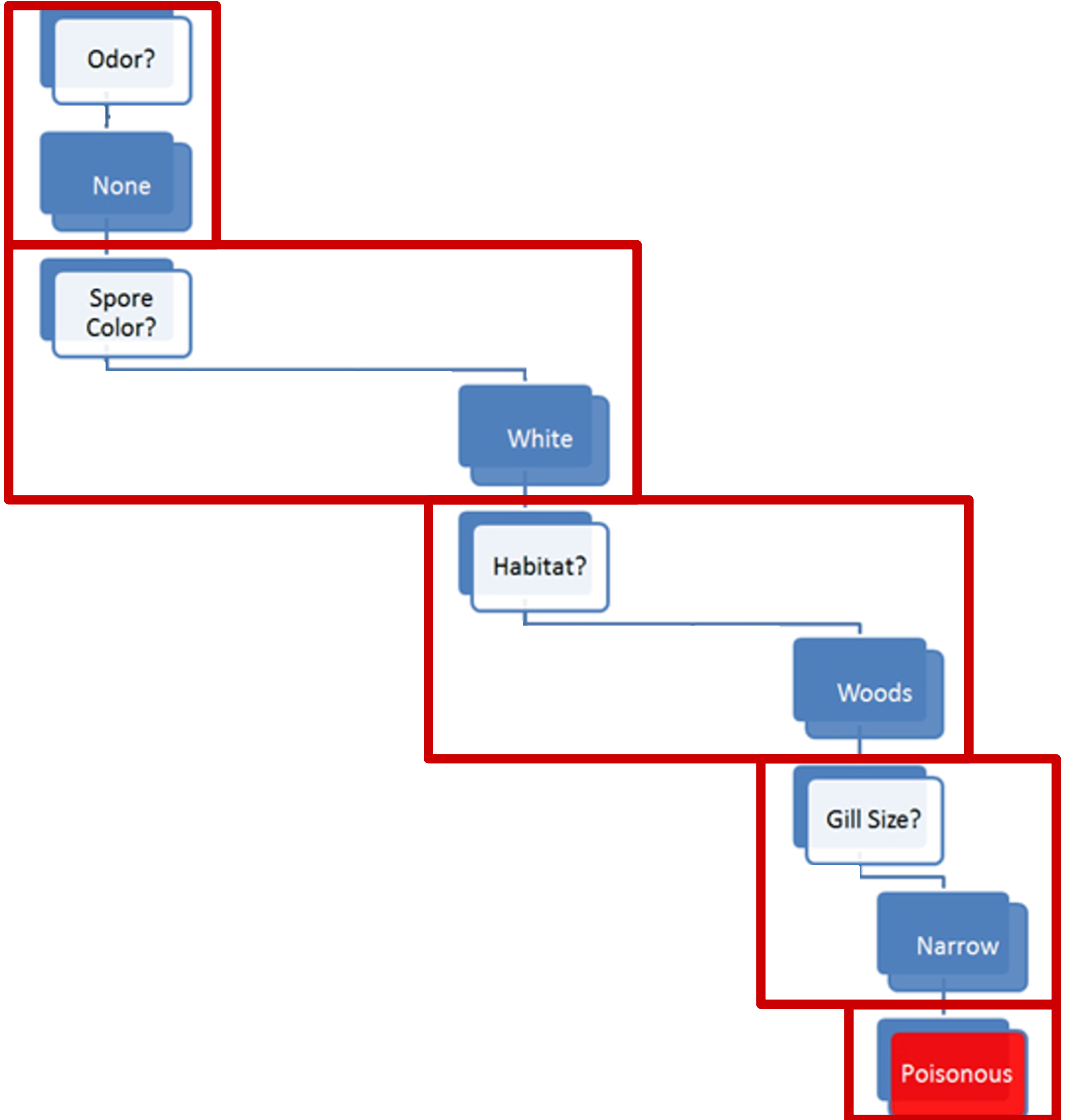
Gill Size: narrow

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 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

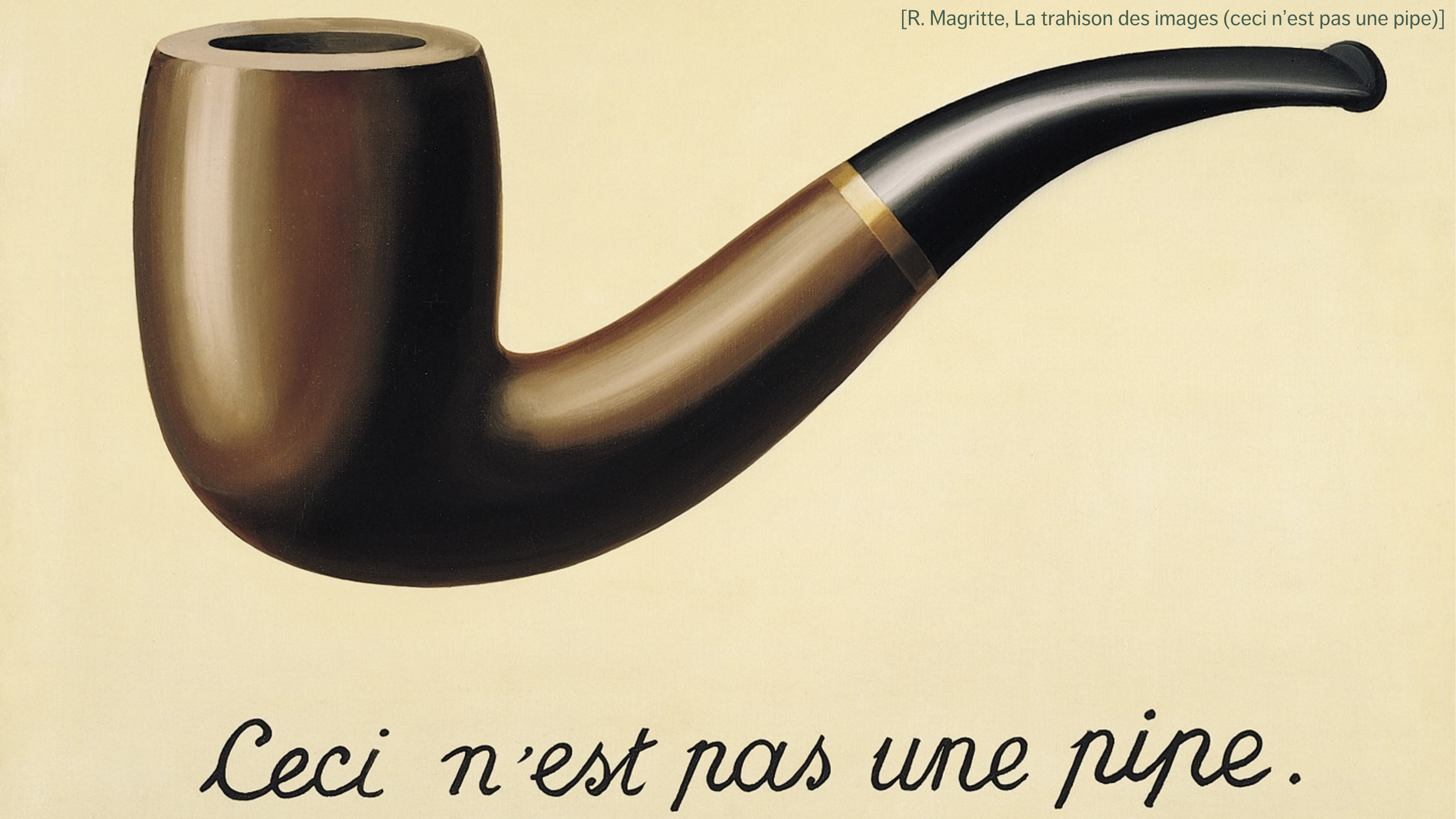
Remember: an object is not simply **the sum of its attributes**.

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)]

FROM OBJECTS AND ATTRIBUTES TO DATASETS

Raw data may exist in any format.

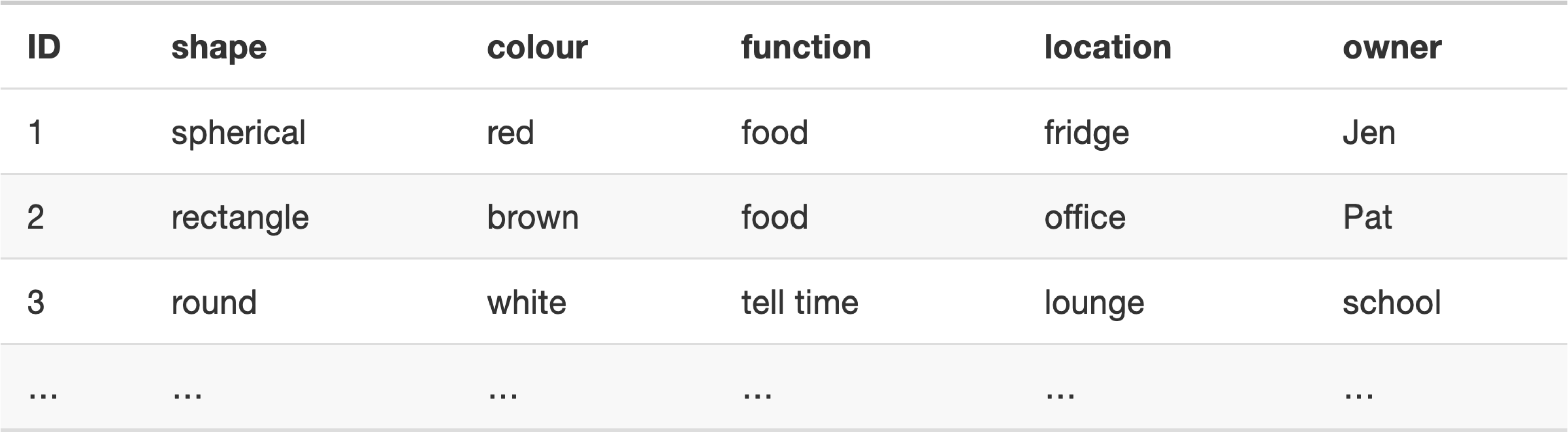
A **dataset** represents a collection of data that could conceivably be fed into algorithms for analytical purposes.

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

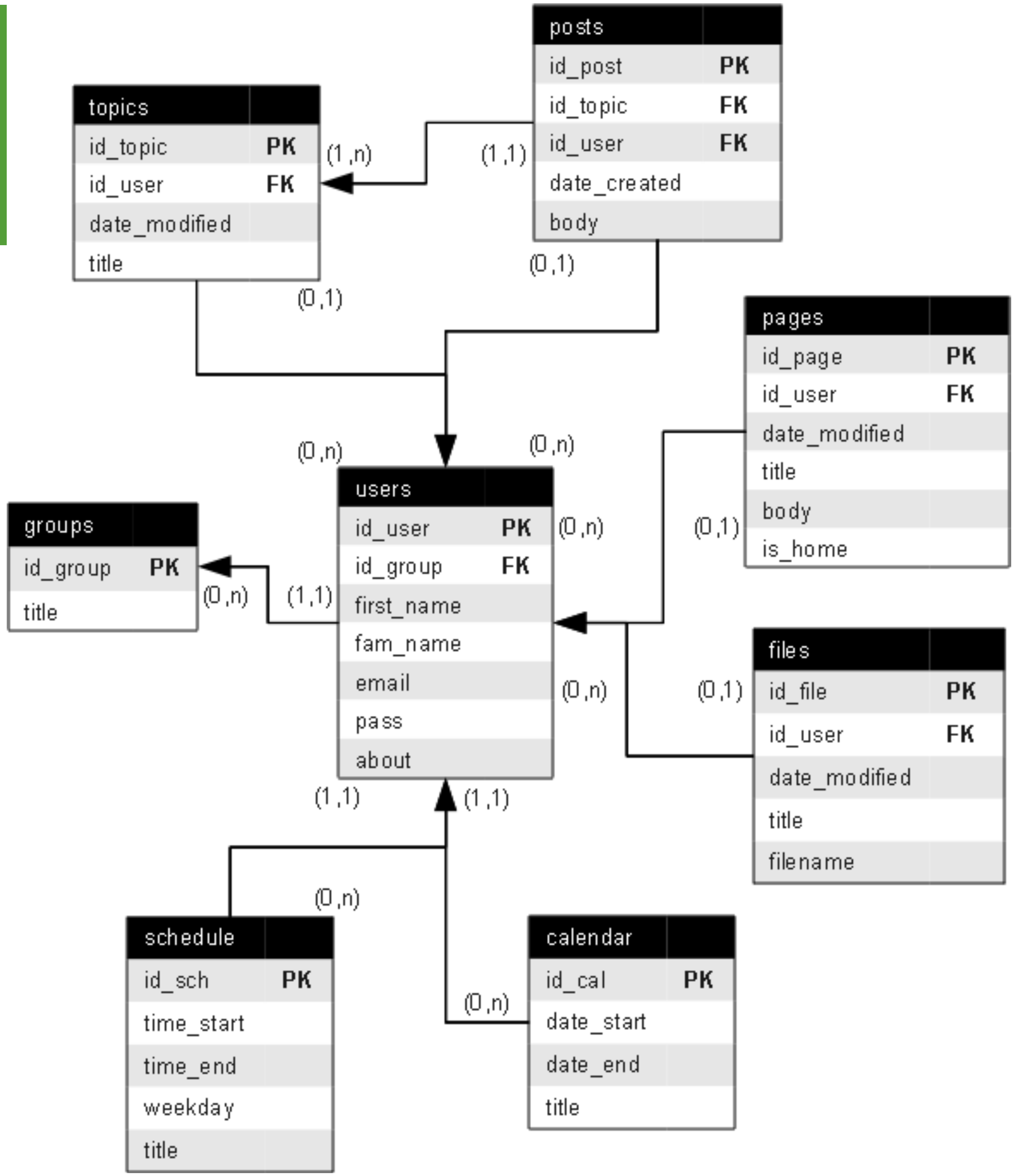
The dataset of physical objects could start with:



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
...	...	...	...	...	...

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.



topics
id_topic PK
id_user FK
date_modified
title

posts
id_post PK
id_topic FK
id_user FK
date_created
body

pages
id_page PK
id_user FK
date_modified
title
body
is_home

groups
id_group PK
title

users
id_user PK
id_group FK
first_name
fam_name
email
pass
about

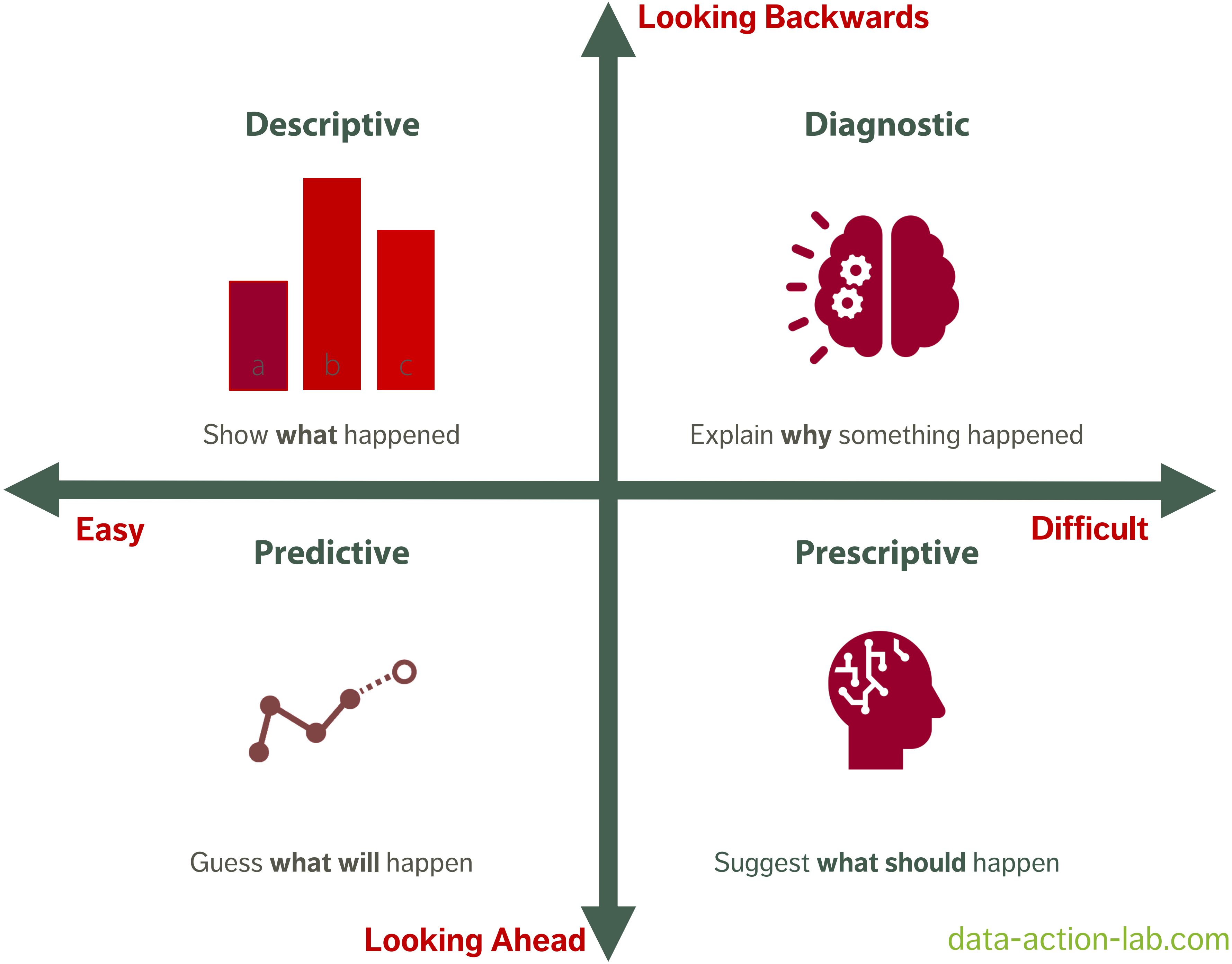
files
id_file PK
id_user FK
date_modified
title
filename

schedule
id_sch PK
time_start
time_end
weekday
title

calendar
id_cal PK
date_start
date_end
title

[Wooptoo, “Entity - relationship model.” Wikimedia]

ANALYTICS MODES



Looking Backwards

Easy

Difficult

Looking Ahead

Descriptive

Show **what** happened

Diagnostic

Explain **why** something happened

Predictive

Guess **what will** happen

Prescriptive

Suggest **what should** happen

data-action-lab.com

[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.

ANALYTICS WORKFLOWS

You are probably sick of **discussions about context** and would rather move to data analysis proper.

Very soon. One last thing, then: the **project context**.

Data science is more than just the analysis of data; this is apparent when we look at the typical steps involved in a **data science project.**

Data analysis pieces take place within this larger project context, as well as in the context of a larger **technical infrastructure** or **pre-existing system**.

THE “ANALYTICAL” METHOD

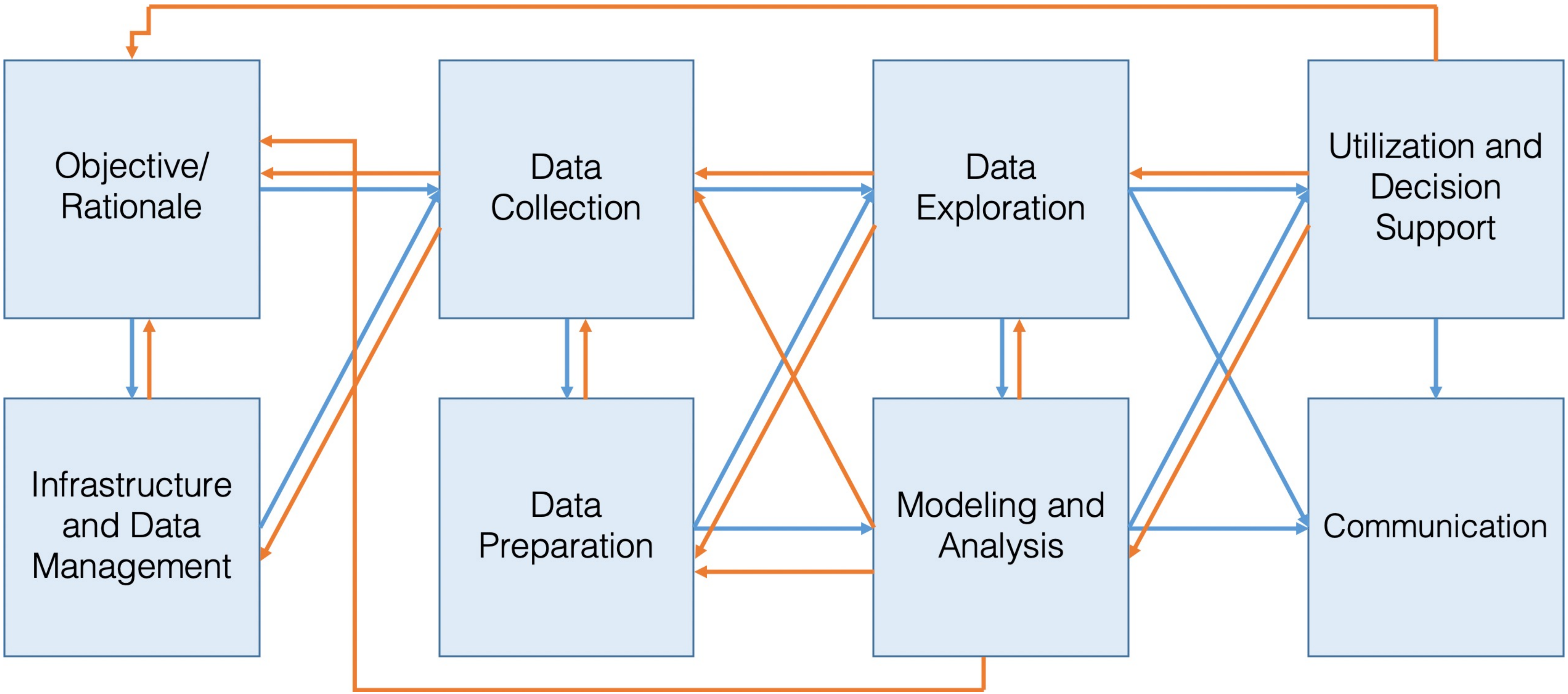
As with the **scientific method**, there is a “step-by-step” guide to data analysis:

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

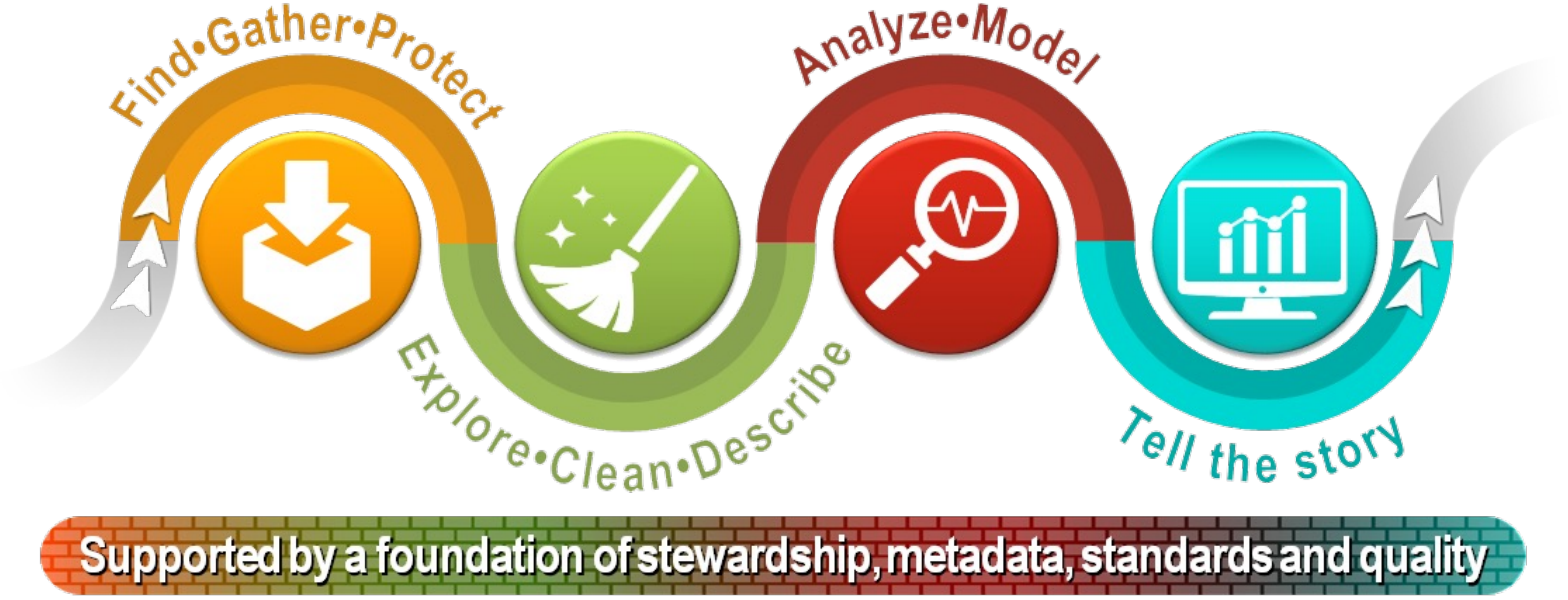
Notice that **data analysis** only makes up a small segment of the entire flow.

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**.

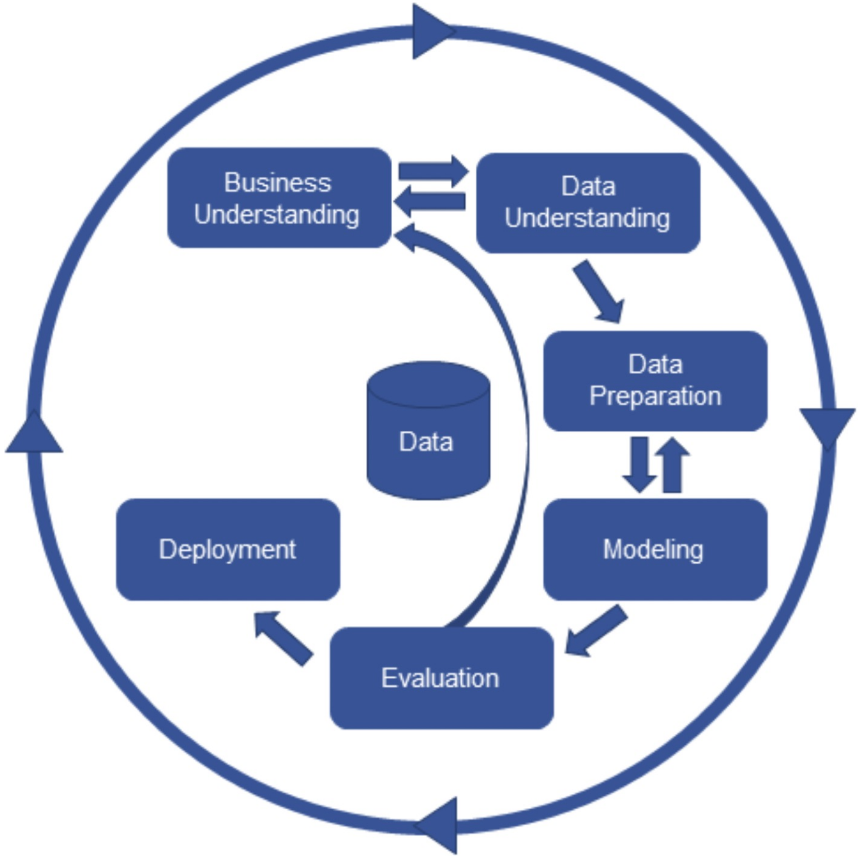

Objective/ Rationale
Data Collection
Data Exploration
Utilization and Decision Support
Infrastructure and Data Management
Data Preparation
Modeling and Analysis
Communication

[StatCan, *Steps of the Data Journey*]


Find•Gather•Protect
Explore•Clean•Describe
Analyze•Model
Tell the story
Supported by a foundation of stewardship, metadata, standards and quality

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

THE "ANALYTICAL" METHODS



Business Understanding

Data Understanding

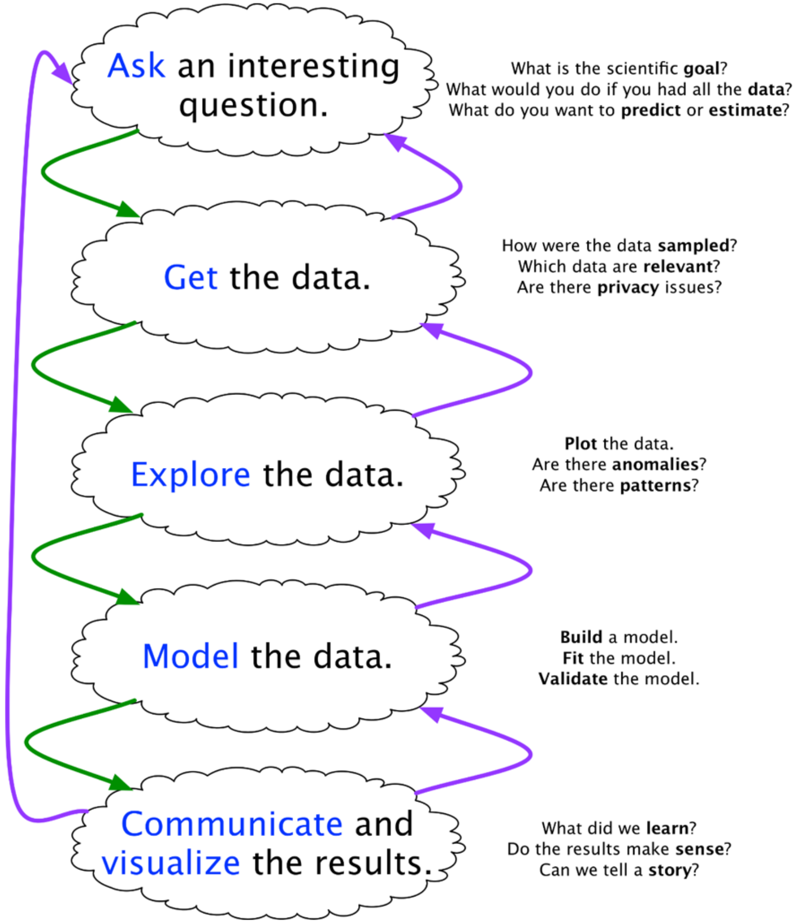
Data Preparation

Data

Modeling

Deployment

Evaluation



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

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**.

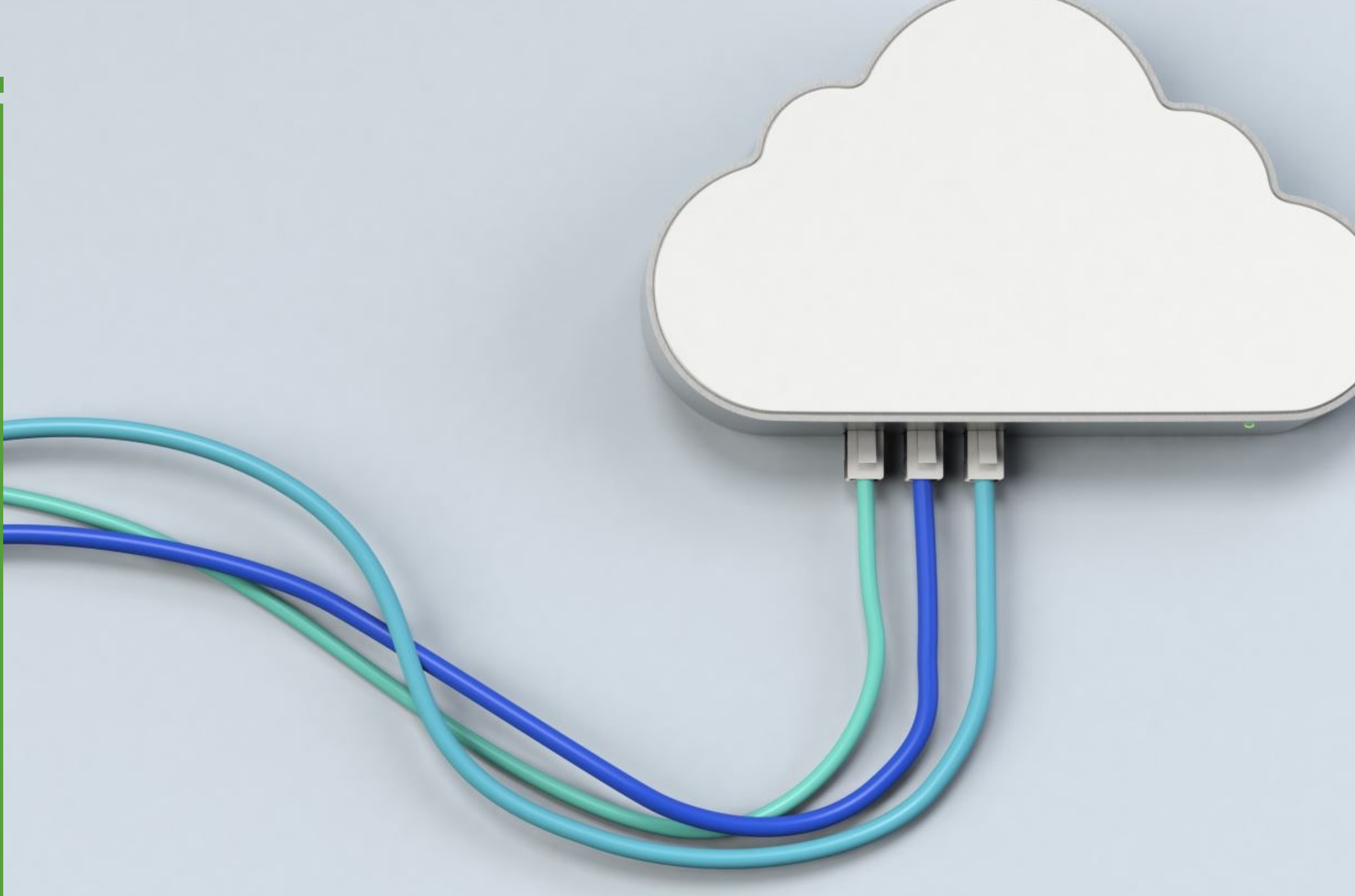
DATA COLLECTION

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.

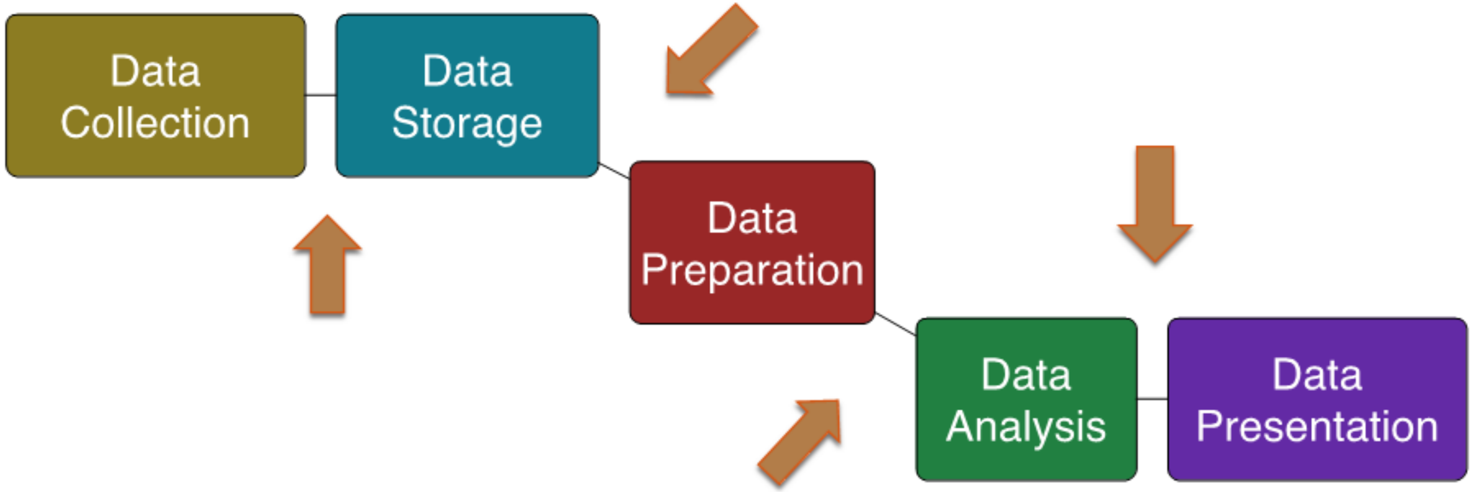
Data pipelines usually consist of 9 components (5 **stages** and 4 **transitions**):

- 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.


Data Collection
Data Storage
Data Preparation
Data Analysis
Data Presentation