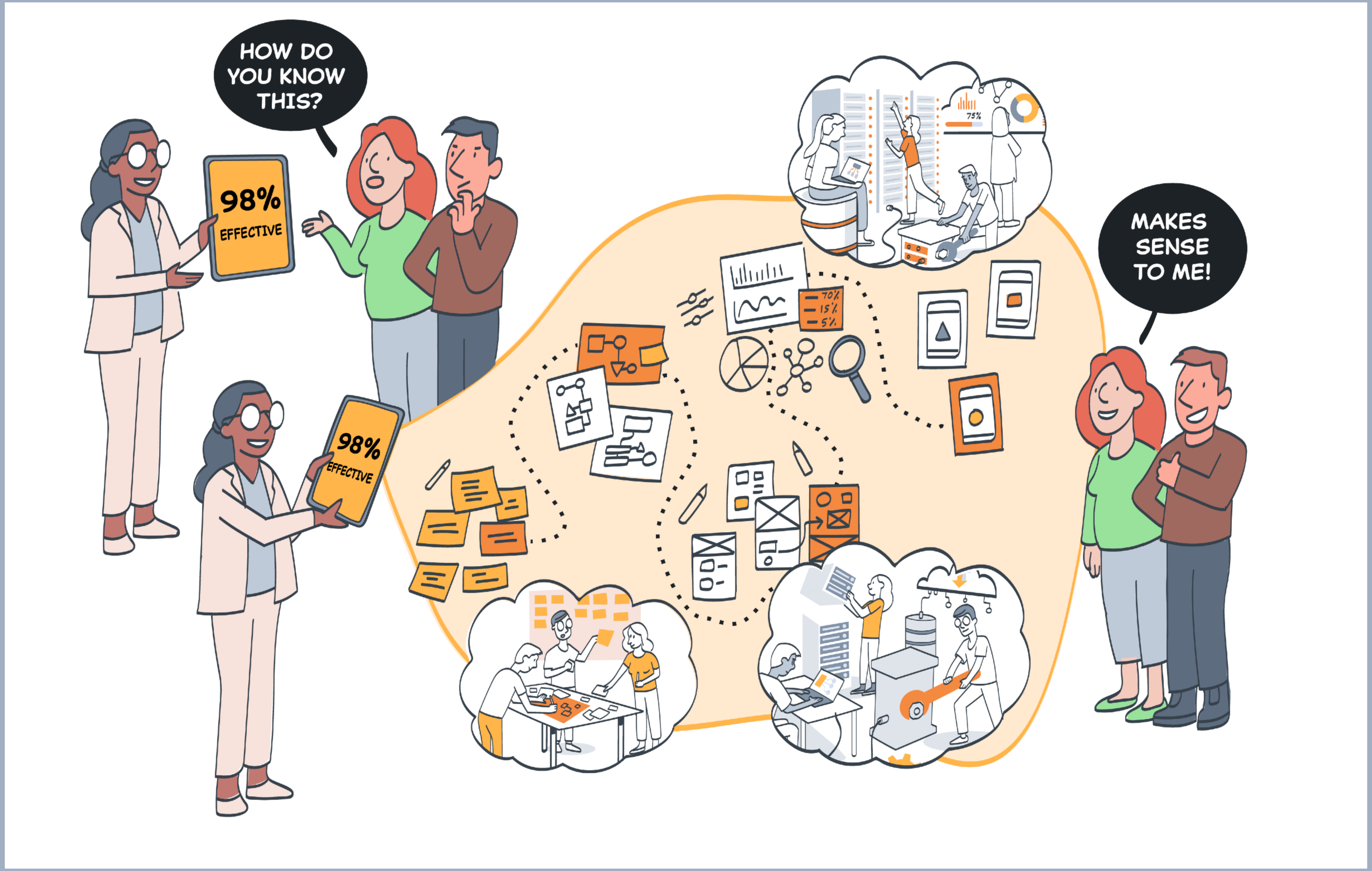
The Alan Turing Institute

Turing Commons

Explainability

Responsible Research and Innovation Skills Track



Dr Christopher Burr
Claudia Fischer

Skills Tracks

Core Modules

+

Optional Modules

+

Case Studies

+

Activities

Explainability Module

01

What is explainability?

02

Project Transparency

03

Model Interpretability

04

Situated Explanations

Feedback

Were any of the sections too difficult or too easy?

Did we give you enough time during discussion and activities?

Did we miss any important topics or concepts?

Any other feedback?

What does ‘explainability’ mean to you?

Go to https://menti.com and use the code
6465 7060

01

WHAT IS EXPLAINABILITY?

SECTION 1

Introduction

The scope of explainability

What is explainability?

Factors that support explanations

SECTION 1

Introduction

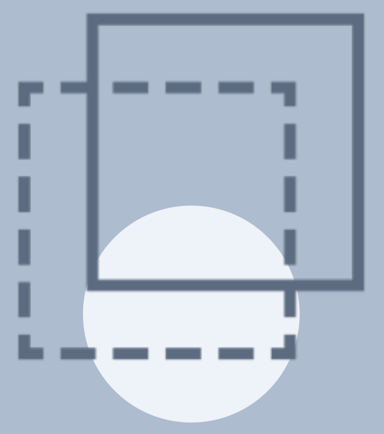
The scope of explainability

What is explainability?

Factors that support explanations

Use of generative AI, and even simpler algorithmic tools depend crucially on properties such as:

Use of generative AI, and even simpler algorithmic tools depend crucially on properties such as:



Transparency

Use of generative AI, and even simpler algorithmic tools depend crucially on properties such as:

Transparency

Interpretability

Use of generative AI, and even simpler algorithmic tools depend crucially on properties such as:

Transparency

Interpretability

Accessible explanations

SECTION 1

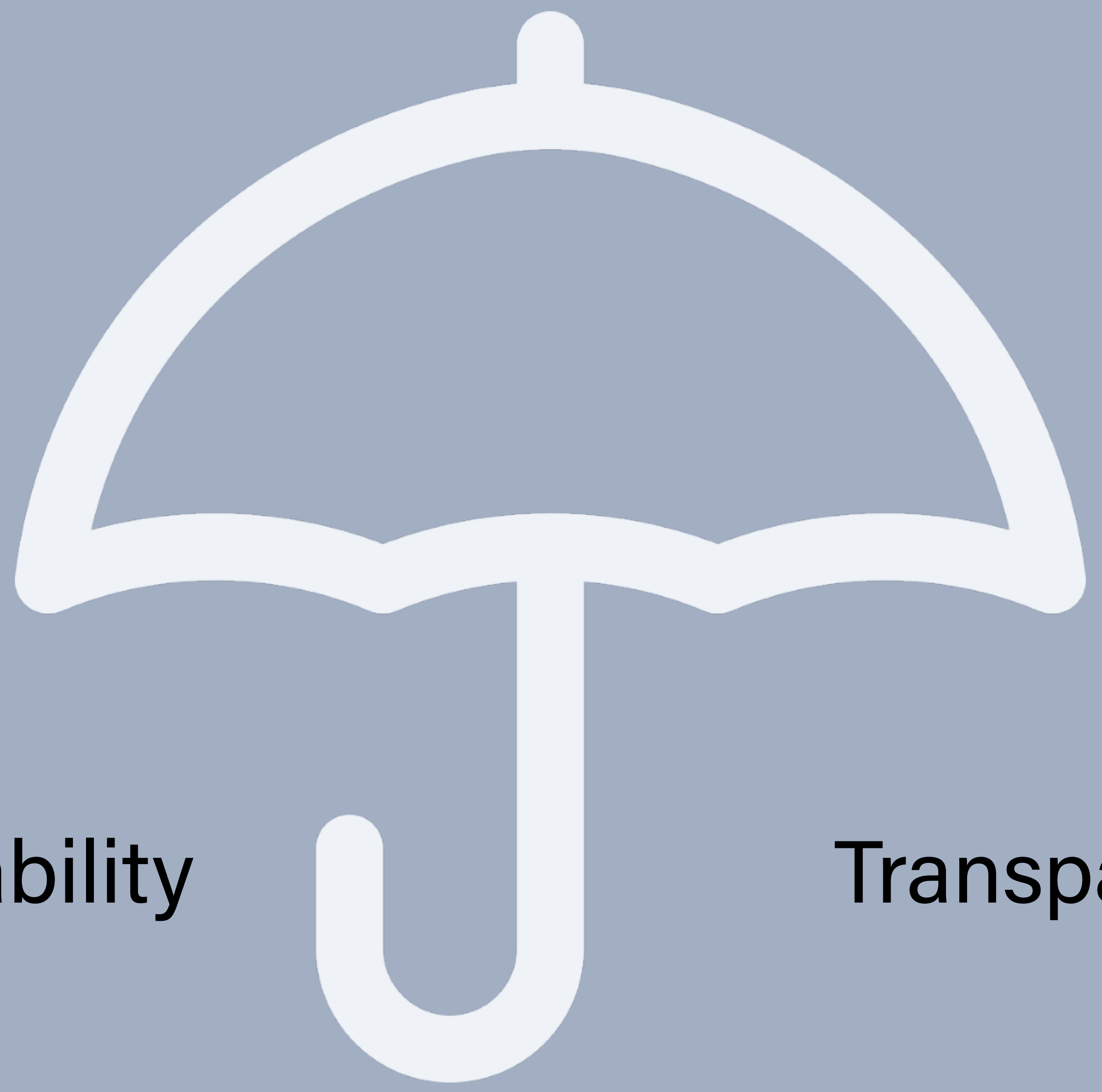
Introduction

The scope of explainability

What is explainability?

Factors that support explanations

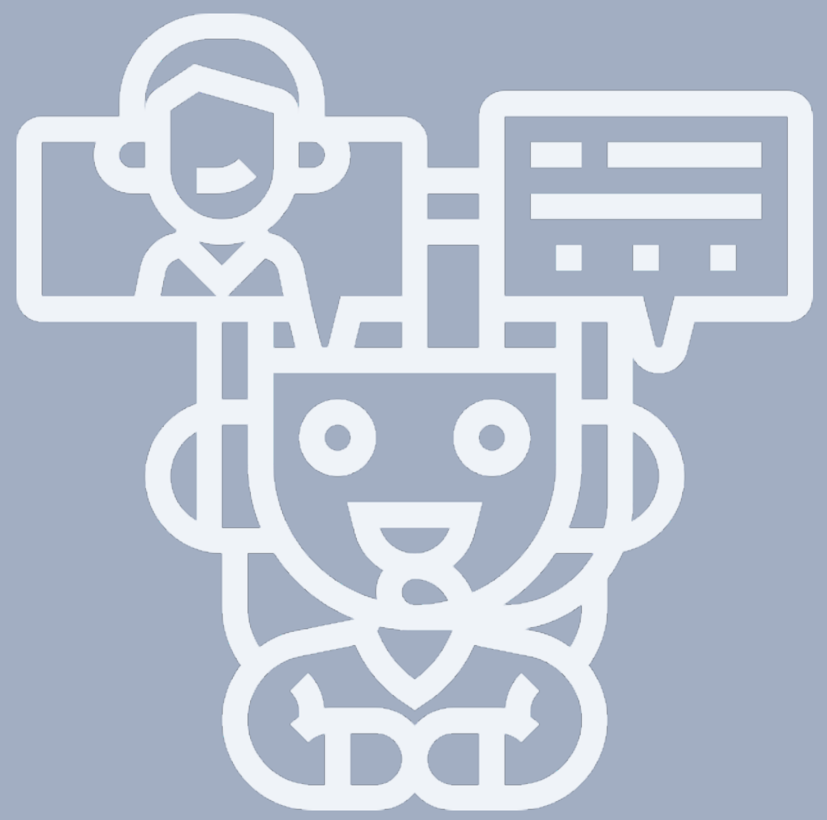
Explainability as a *catch-all*, umbrella term



Interpretability

Transparency

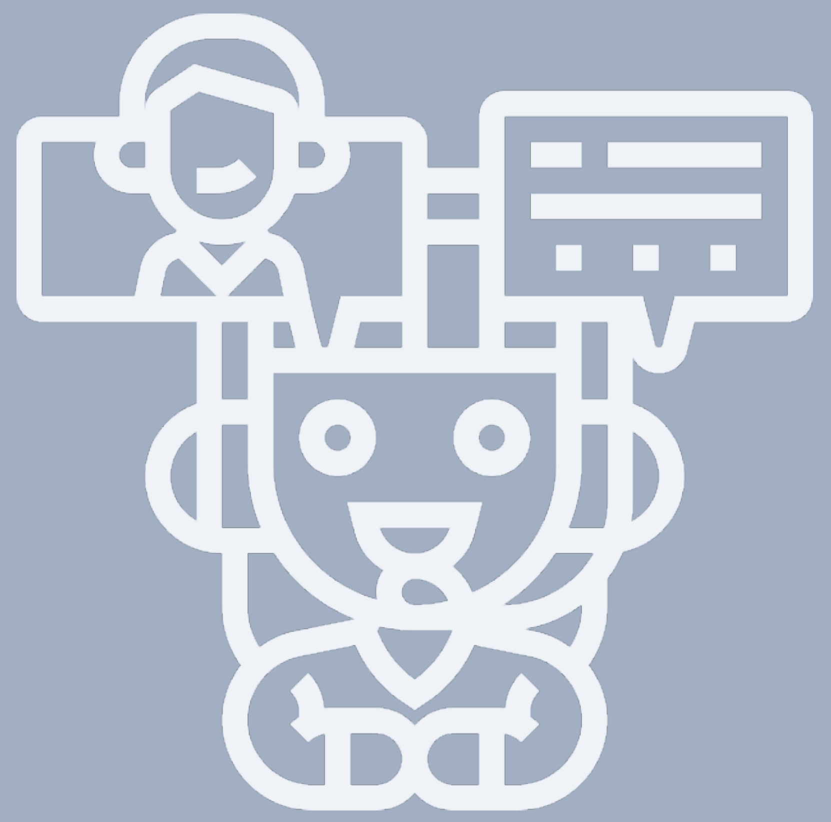
Situated Explanations



The focus of this module is on understanding why explainability matters for responsible research and innovation

Two relevant caveats:

The focus of this module is on understanding why explainability matters for responsible research and innovation



Two relevant caveats:

- This is not a module teaching how to use or implement existing methods or techniques.

The focus of this module is on understanding why explainability matters for responsible research and innovation

Two relevant caveats:

- This is not a module teaching how to use or implement existing methods or techniques.
- This module aims to be consistent with widely agreed uses of concepts and terminology, but also has its own unique perspective on the topic.

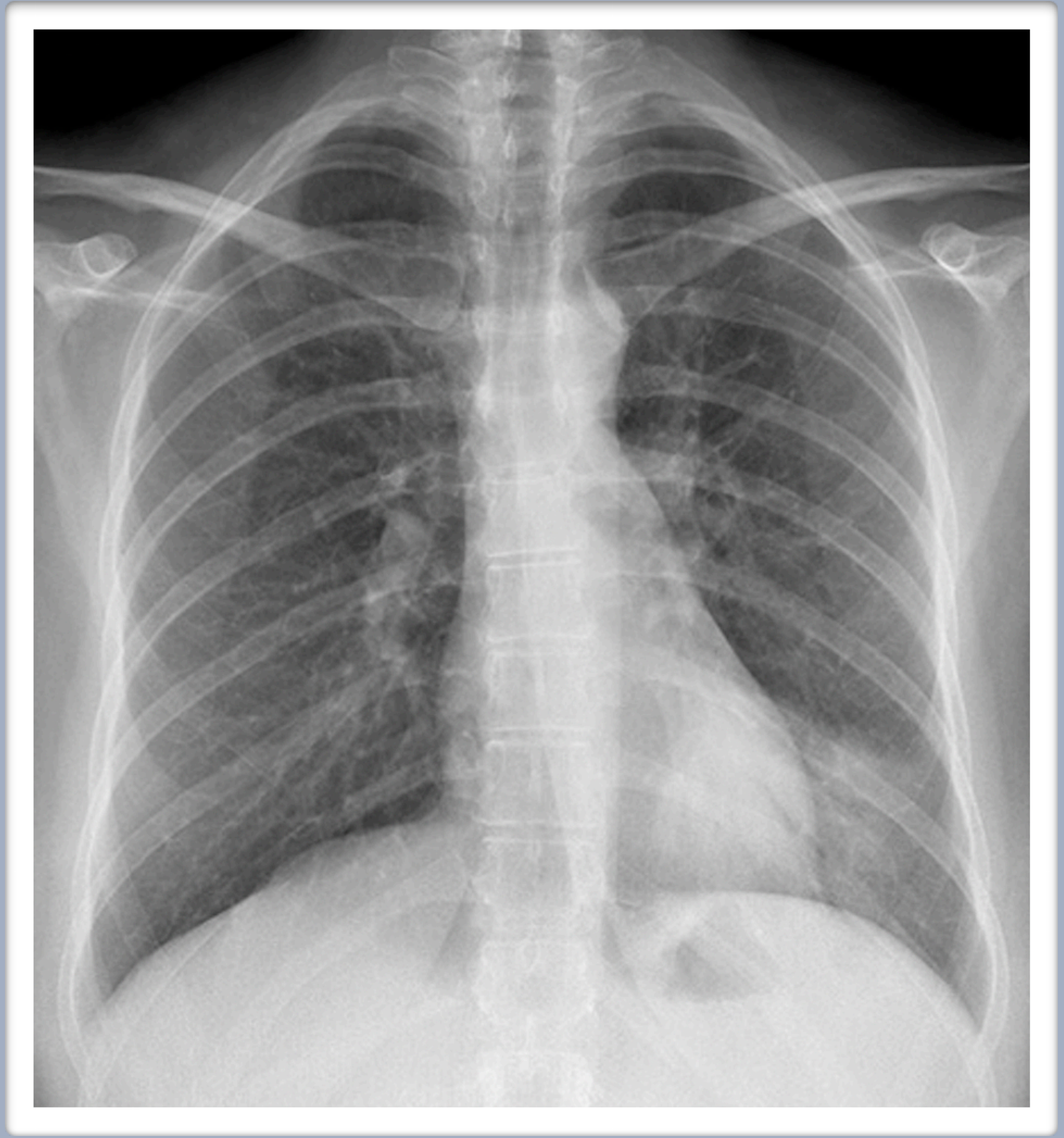
SECTION 1

What is explainability?

Interpretability is the degree to which a human can understand the cause of a decision.

— Miller (2019)

Interpreting images



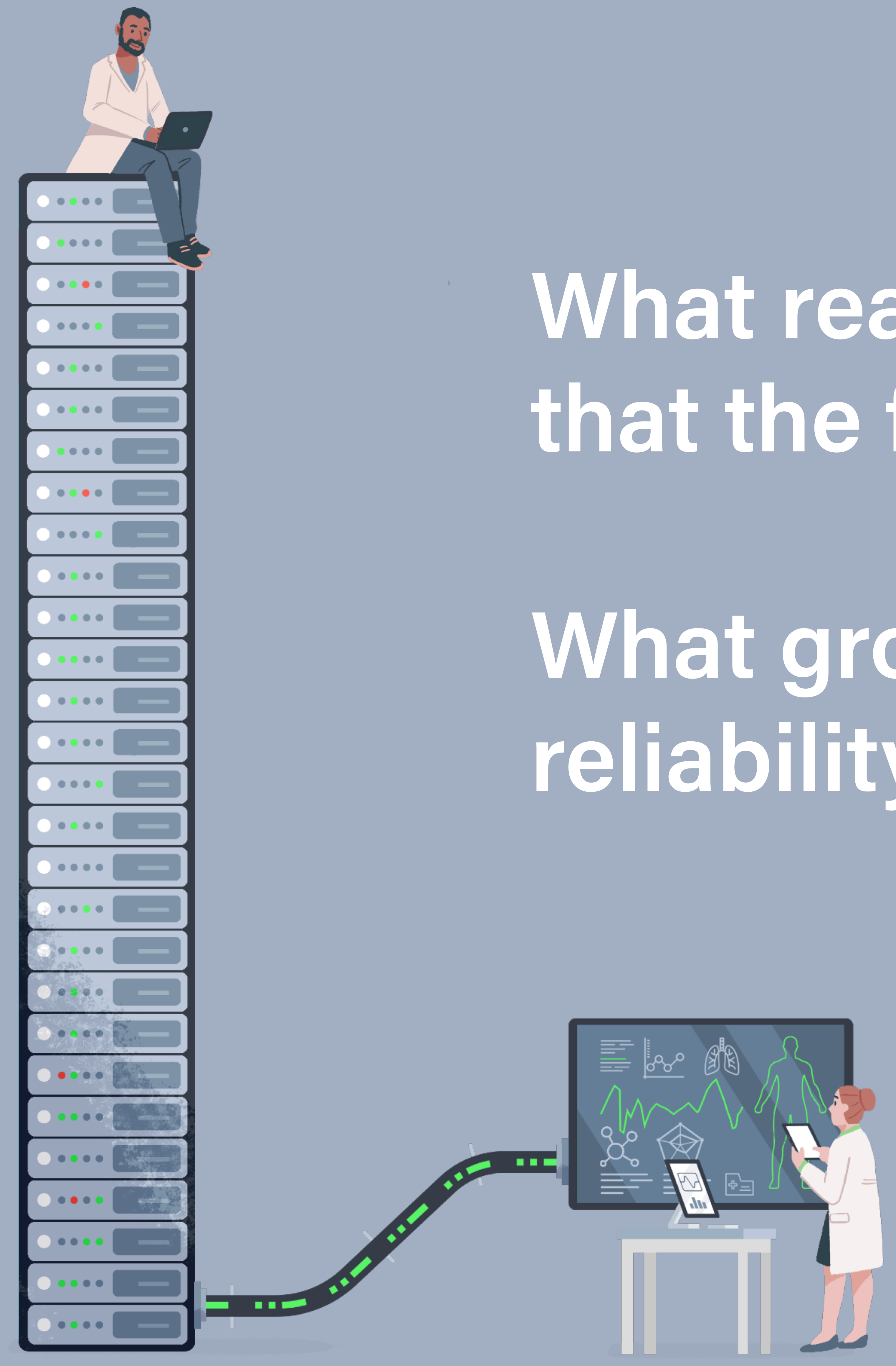
Requests for Explanations

Requests for Explanations


CYCLIST
ETA
3:59

The degree to which a system or set of tools support a person's ability to explain and communicate the behaviour of the system or the processes of designing, developing, and deploying the system within a particular context.

The Problem of Induction



What reasons do we have to believe, and justify, that the future will be like the past?

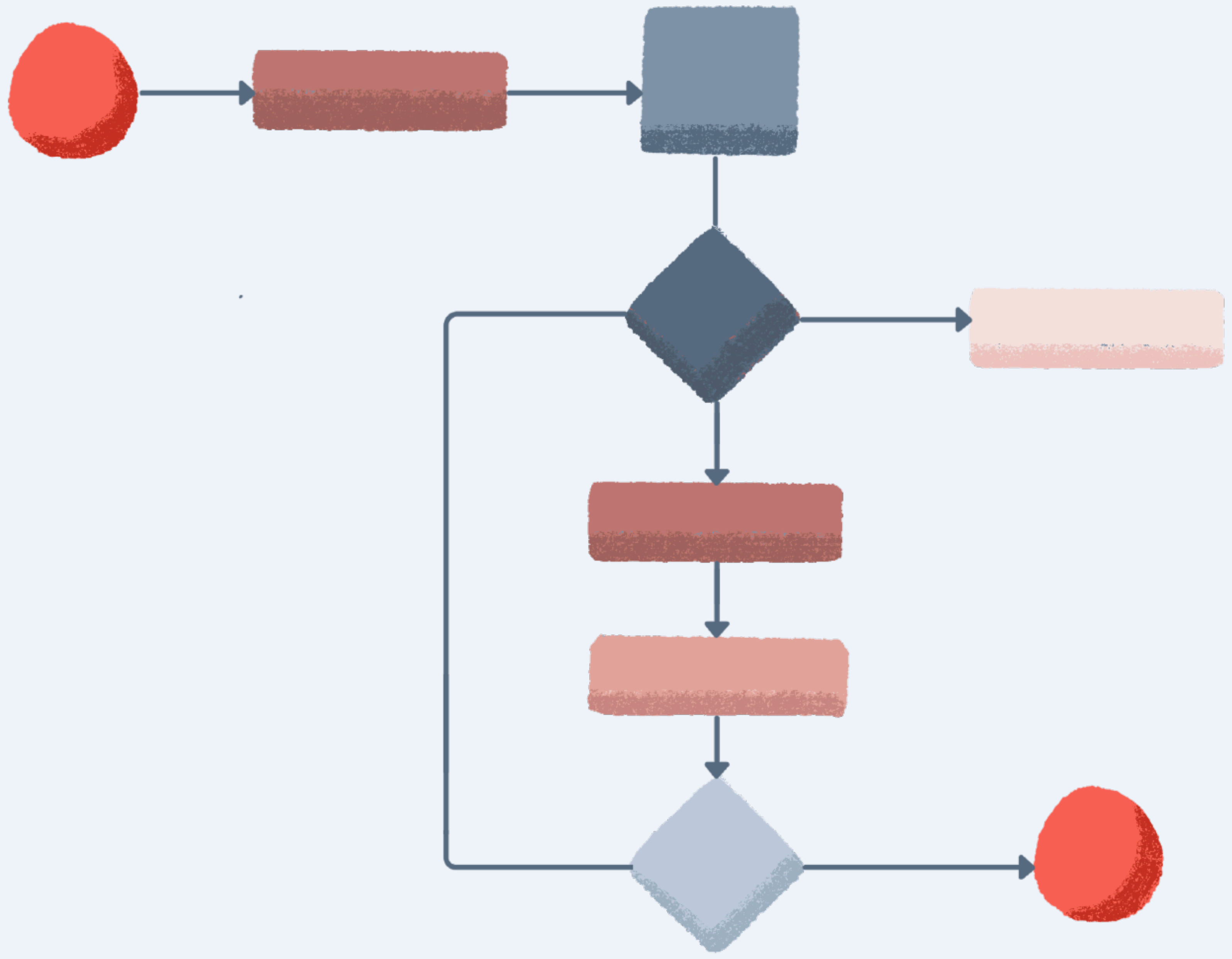
What grounds do we have for justifying the reliability of our predictions?

The Problem of Induction

The turkey's predictions were highly accurate (99.7% over the course of its life)…
but the one time it was wrong really mattered!

We want reliable and valid reasons for why we can trust the predictions made by our systems, especially those that are embedded within safety critical parts of our society and infrastructure.

Explainability as **justifiable reasons and evidence** for why the predictions and behaviour of a model are trustworthy and valid



SECTION 1

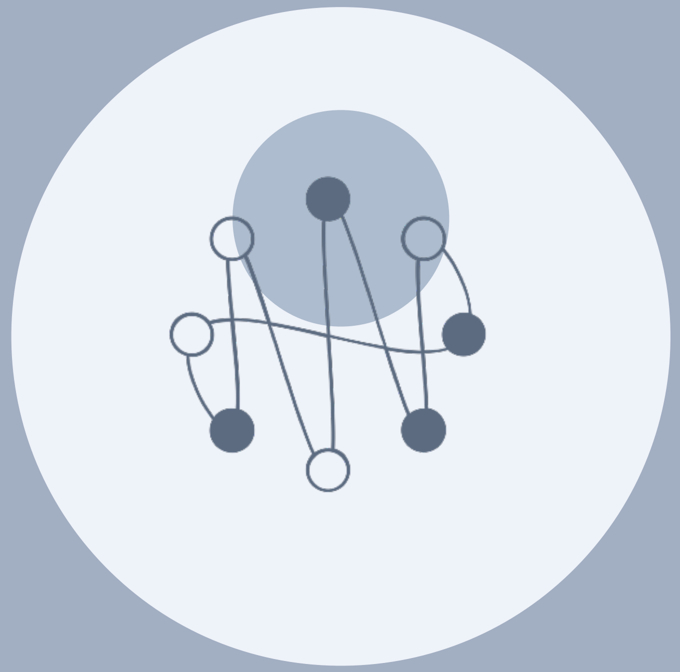
Introduction

The scope of explainability

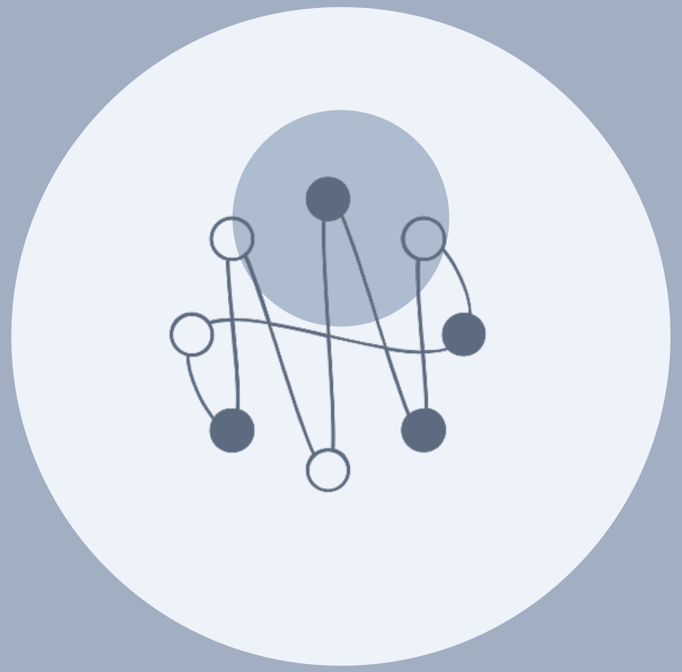
What is explainability?

Factors that support explanations

Factors that support explanations



Factors that support explanations

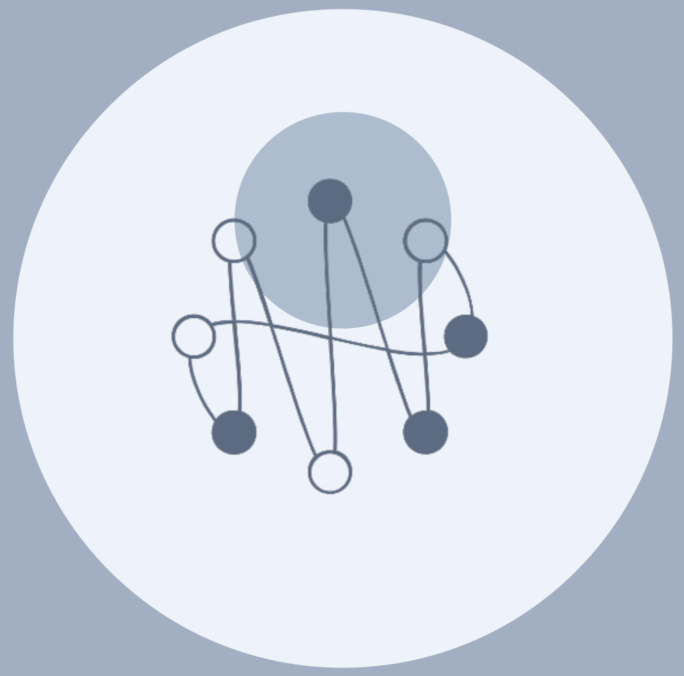


Transparent and accountable processes of project governance

Factors that support explanations

Transparent and accountable processes of project governance

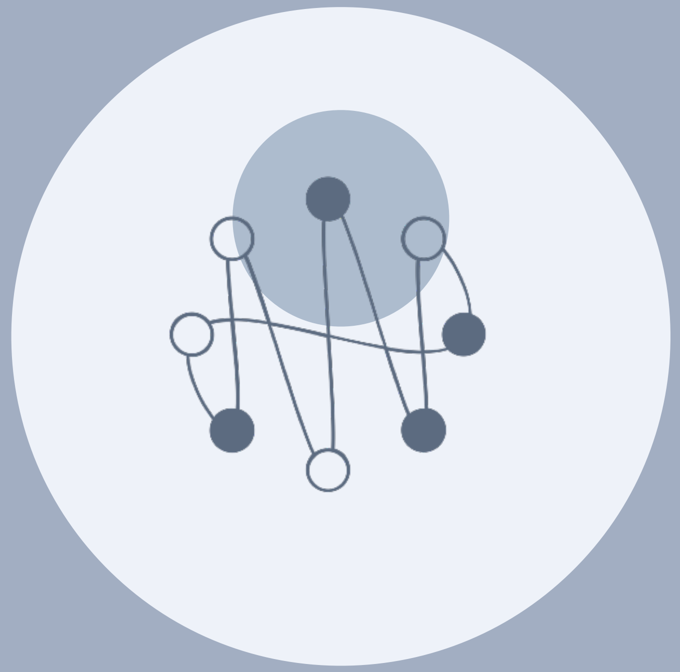
Interpretable models



Factors that support explanations

Transparent and accountable processes of project governance

Interpretable models



Awareness of the sociocultural context

Are there other factors you think we may have missed or under-emphasised?

Summary

- ‘Explainability’ is used as an umbrella term, capturing several important factors.
- Problem of induction prompts us to consider whether we have valid and reliable reasons (or justification) for our explanations.
- Requests for explanation are shaped by sociocultural expectations (e.g. folk psychological versus professional)
- Important factors include transparency, interpretability, and awareness of sociocultural context.

Q&A

02

PROJECT TRANSPARENCY

SECTION 2

01

Introduction

02

What are we trying to explain?

03

What does responsible project transparency look like?

04

Limits of transparency

SECTION 2

01
Introduction

02
What are we trying to explain?

03
What does responsible project transparency look like?

04
Limits of transparency

A team of lawyers are carrying out discovery.

A team of lawyers are carrying out discovery.

1. The other team have sent across mountains of documents and files

A team of lawyers are carrying out discovery.

1. The other team have sent across mountains of documents and files

2. Information about the structure of the algorithm is written in technical jargon that is hard for the lawyers to understand.

This hypothetical scenario highlight two relevant issues

This hypothetical scenario highlight two relevant issues

1. Transparency is not the same as *accessibility*

This hypothetical scenario highlight two relevant issues

1. Transparency is not the same as *accessibility*

2. Transparency is necessary for *explainability*

SECTION 2

Introduction

01

What are we trying to explain?

02

What does responsible project transparency look like?

03

Limits of transparency

04

Consider the following scenario:

> A team of data analysts who work for a travel booking website are asked to explain why a model has altered its predictions about customer purchasing behaviour.

But now let's assume that there is another change, which results in a significant drop in conversion rate.

The fault turns out to be related to a software dependency issue in their data pipeline (e.g. a broken plugin).

All customers are now being shown the same (expensive) holiday packages regardless of their location—used as a proxy for socioeconomic status.

The locus of our explanation will not always be the model.

The locus of our explanation will not always be the model.

Therefore, the sort of transparency that we are interested in is not merely the transparency of the model itself, but rather the **transparency of the project (and system) as a whole.**

What about the transparency of the learning algorithm?

(a) Husky classified as wolf (b) Explanation

Ribeiro et al. (2016)

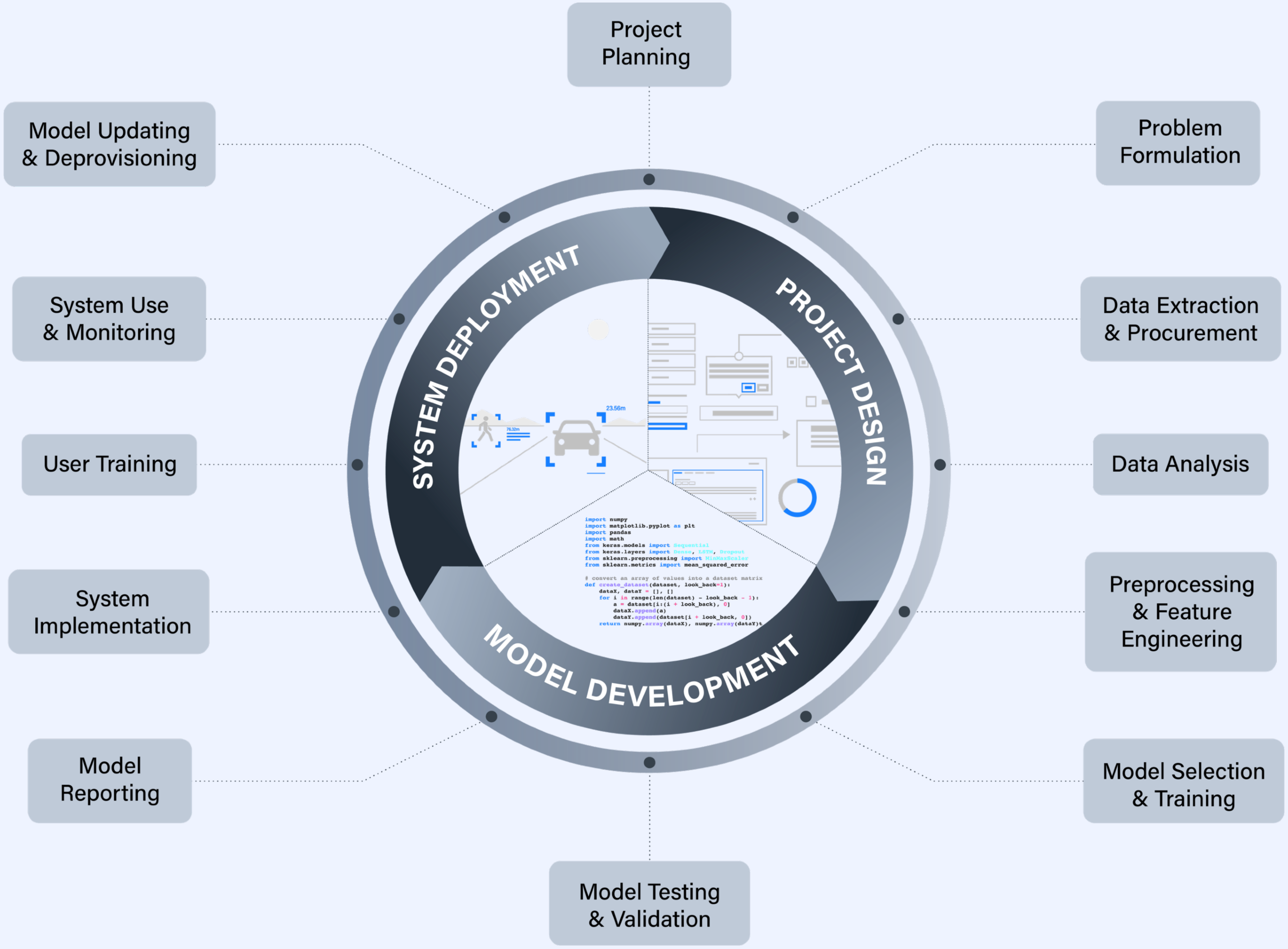
SECTION 2

01 Introduction

02 What are we trying to explain?

03 **What does responsible project transparency look like?**

04 Limits of transparency

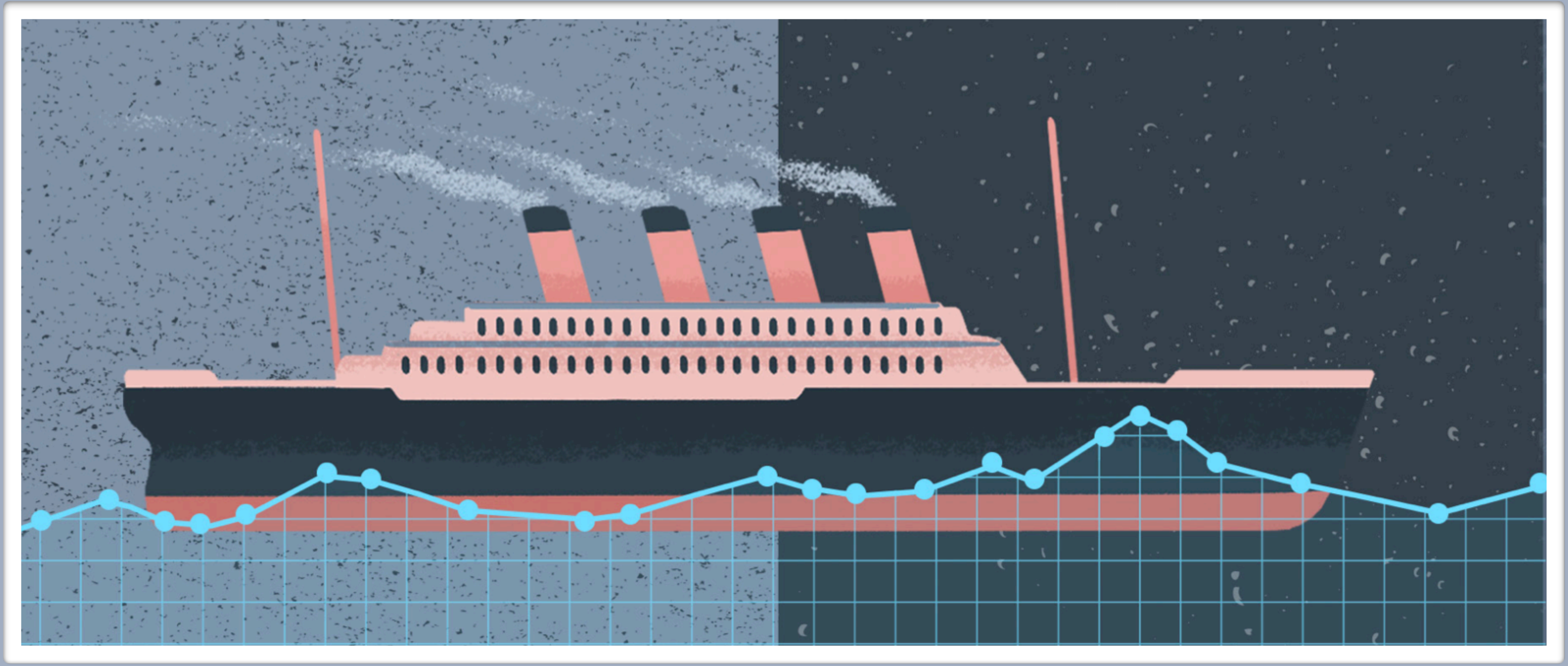

Project Planning
Problem Formulation
Data Extraction & Procurement
Data Analysis
Preprocessing & Feature Engineering
Model Selection & Training
Model Testing & Validation
Model Reporting
System Implementation
User Training
System Use & Monitoring
Model Updating & Deprovisioning
PROJECT DESIGN
MODEL DEVELOPMENT
SYSTEM DEPLOYMENT

Practical mechanisms and processes for project transparency

- ↗ Tasks that involve choices about how a project should be governed
- ↗ Tasks that involve what we can term 'data stewardship'
- ↗ Tasks that involve the engagement of stakeholders

Tasks that involve choices about how a project should be governed

Tasks that involve what we can term 'data stewardship'



Tasks that involve the engagement of stakeholders

What other tasks can you think of, which may occur during one of the project lifecycle stages, that would require transparency?

What other tasks can you think of, which may occur during one of the project lifecycle stages, that would require transparency?

How would this transparency be achieved and how would it contribute to explaining any decisions or actions taken?

SECTION 2

Introduction

01

What are we trying to explain?

02

What does responsible project transparency look like?

03

Limits of transparency

04

Limits of transparency

Limits of transparency

- Resource barrier—especially for smaller teams

Limits of transparency

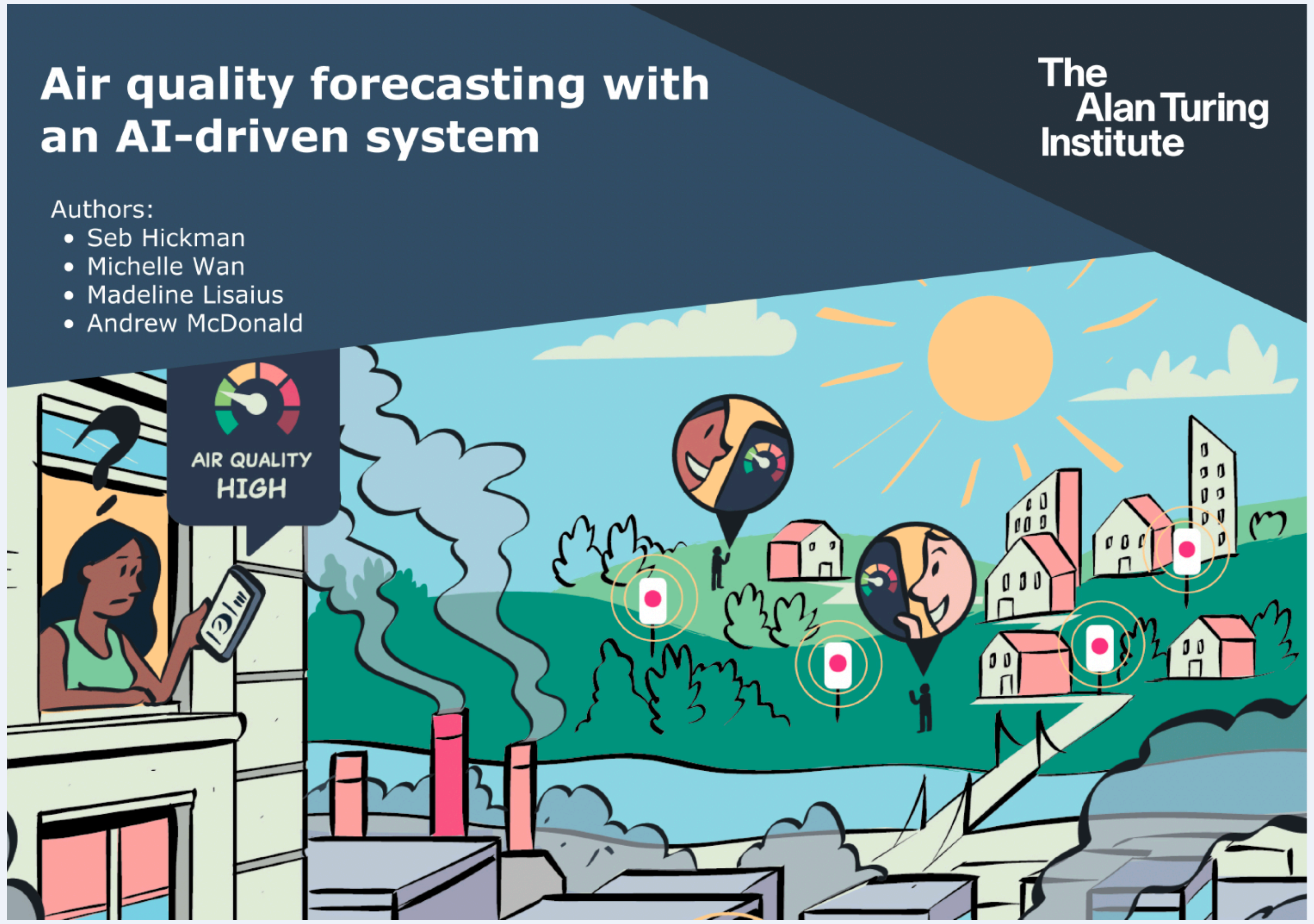
- Resource barrier—especially for smaller teams
- Intellectual property or other legal restrictions

Summary

- Transparency is necessary for explainability but is not the same as ‘accessibility’
- Locus of an explanation may go beyond the model that powers a system.
- Project lifecycle model provides a scaffold for identifying tasks that may be a source of information.
- There are limits to how much transparency can be obtained—transparency is not a universal good.

Q&A

Group Discussion and Case Studies


Air quality forecasting with an AI-driven system
The Alan Turing Institute
Authors:
• Seb Hickman
• Michelle Wan
• Madeline Lisaius
• Andrew McDonald
AIR QUALITY
HIGH


Remote Sensing in Context
Data and model exploitation for surveillance
Authors:
• Herbie Bradley
• Sofija Stepanović
• Orlando Timmerman
The Alan Turing Institute
GREEN MINING CORP
GMC

Has your concept of explainability changed since the initial word cloud activity?

If so, how and why?

Are there important factors related to explainability that you would like to add to your case studies?

Are there new deliberative prompts or key issues?

03

MODEL INTERPRETABILITY

SECTION 3

What is model interpretability?

01

Methods for interpreting models

02

Model interpretability and RRI

03

Building explanations

04

SECTION 3

01
What is model interpretability?

02
Methods for interpreting models

03
Model interpretability and RRI

04
Building explanations

Interpretability is the degree to which a human can understand the cause of a decision.

— Miller (2019)

Interpretability is the degree to which a human can understand the cause of a model's prediction or behaviour.

A model, trained on a dataset of cars, predicts a car's top speed based on the following three features:

A model, trained on a dataset of cars, predicts a car's top speed based on the following three features:

1. Colour

A model, trained on a dataset of cars, predicts a car's top speed based on the following three features:

1. Colour

2. Number of doors

A model, trained on a dataset of cars, predicts a car's top speed based on the following three features:

1. Colour
2. Number of doors
3. Convertible (binary value)

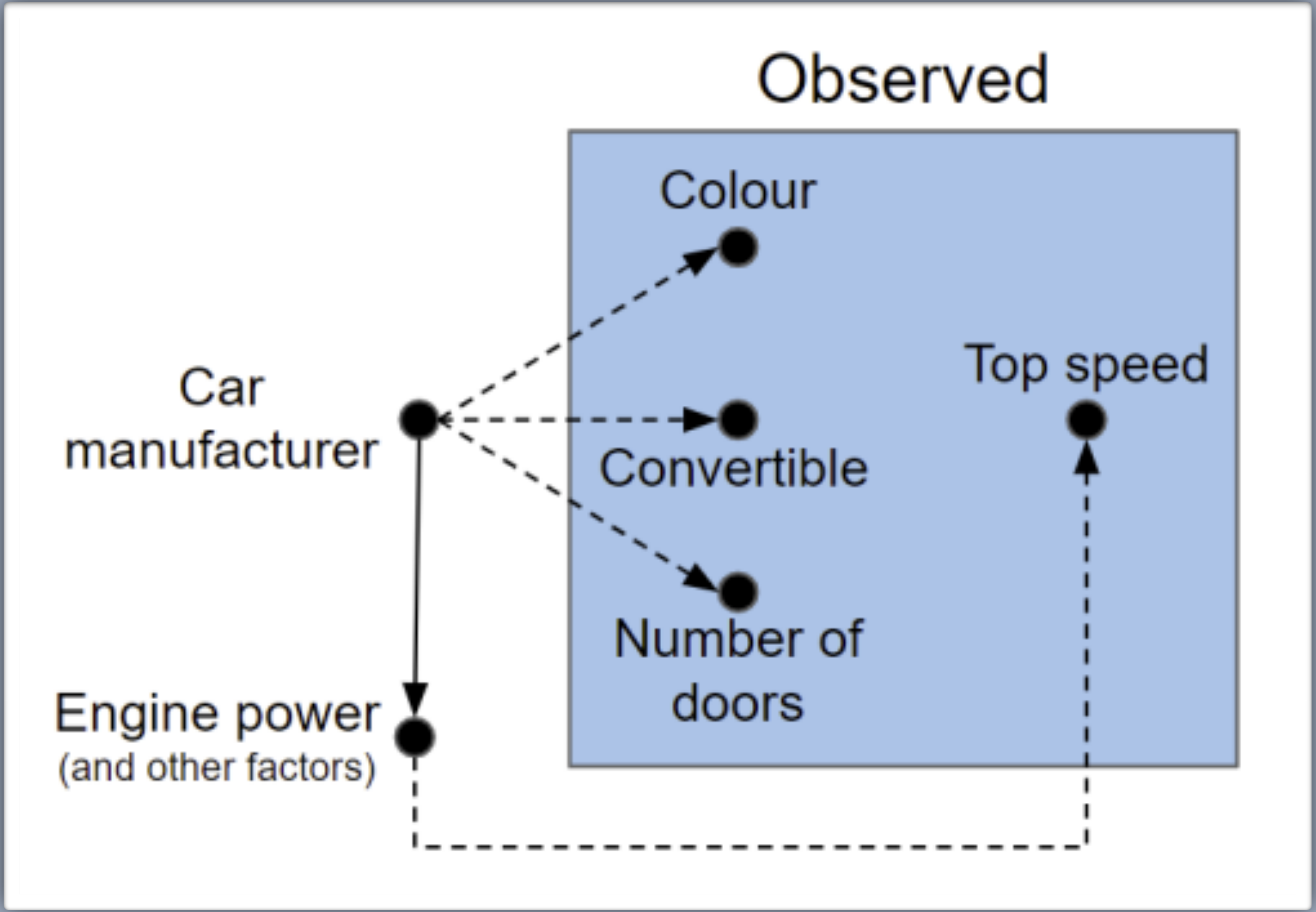
A model, trained on a dataset of cars, predicts a car's top speed based on the following three features:

1. Colour: **Green**

2. Number of doors: **Two doors**

3. Convertible (binary value): **Yes**


LAMBORGHINI
Lamborghini Warszawa



Reprinted from Lazaridis (2021)

SECTION 3

01
What is model interpretability?

02
Methods for interpreting models

03
Model interpretability and RRI

04
Building explanations

Methods for interpreting models:

1. Rule-based models
2. Linear models
3. Feature importance techniques
4. Prototypes and criticisms
5. Surrogate models
6. Visualisations
7. Concept activation vectors
8. Counterfactual explanations
9. Bayesian networks

Methods for interpreting models:

1. Rule-based models
2. Linear models
3. **Feature importance techniques**
4. Prototypes and criticisms
5. **Surrogate models**
6. **Visualisations**
7. Concept activation vectors
8. Counterfactual explanations
9. Bayesian networks

1. Intrinsic vs post hoc

2. Model specific vs model agnostic

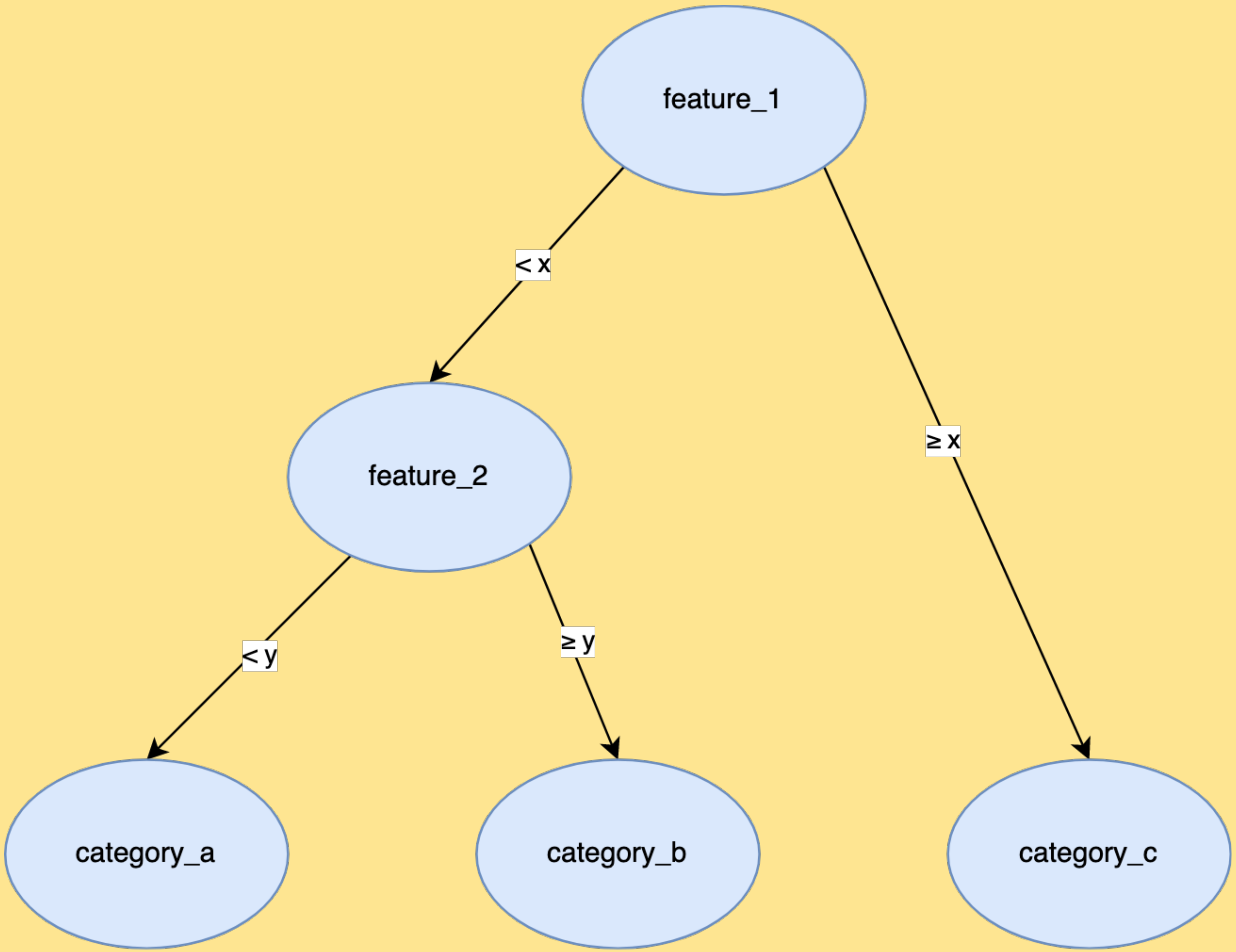
3. Global vs local

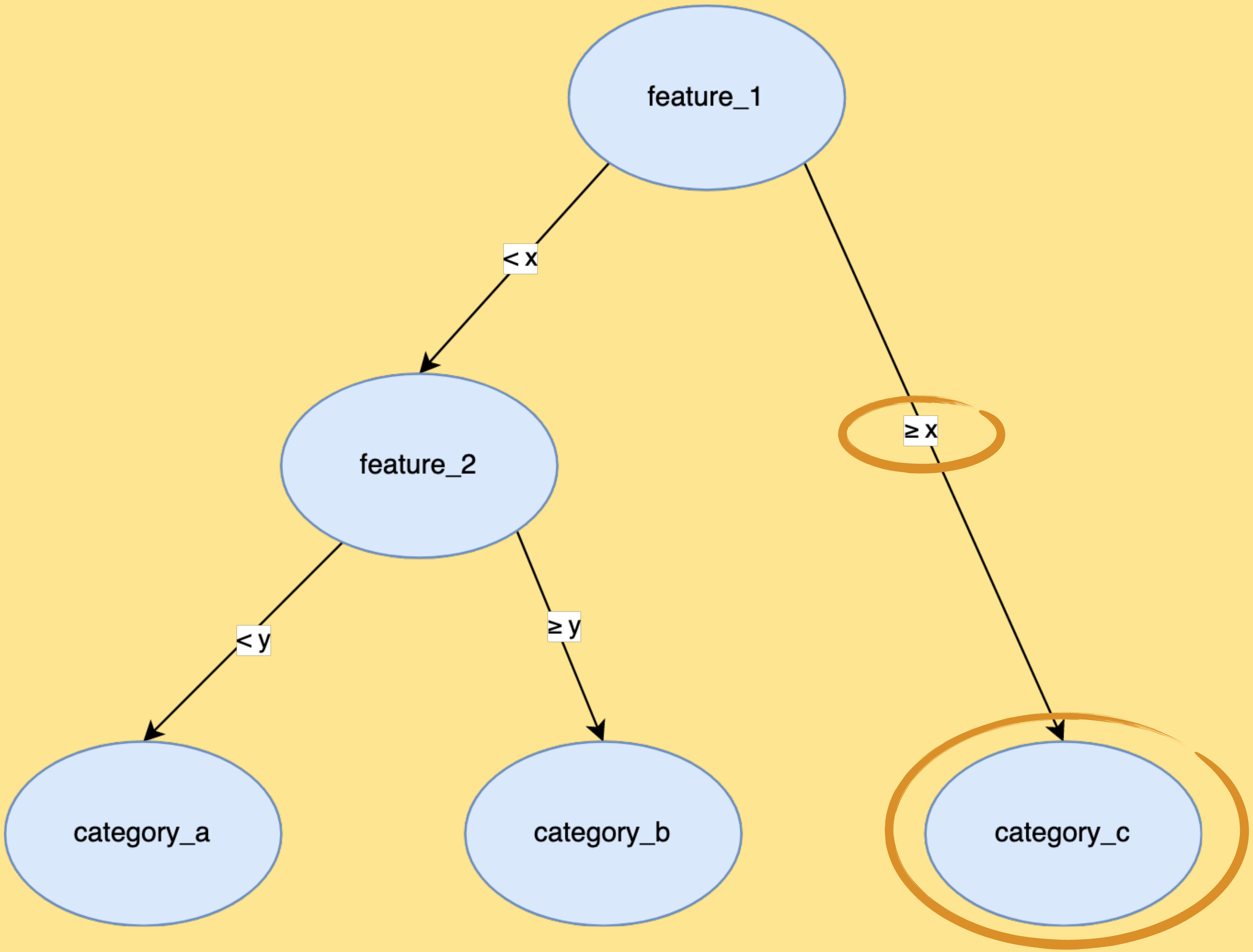
4. Results of interpretability methods

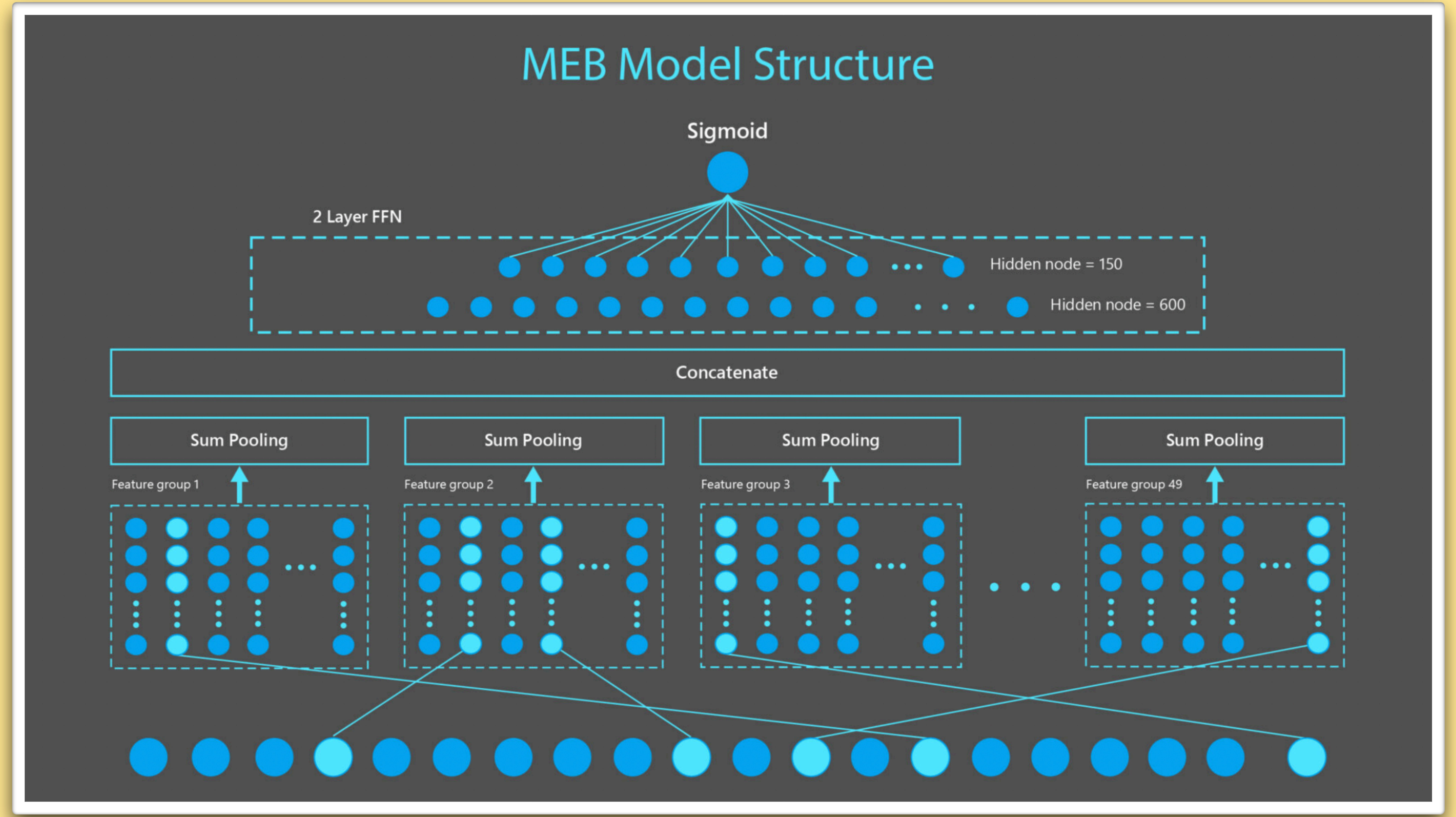
1. Intrinsic vs post hoc

Intrinsically interpretable

Low to no intrinsic interpretability

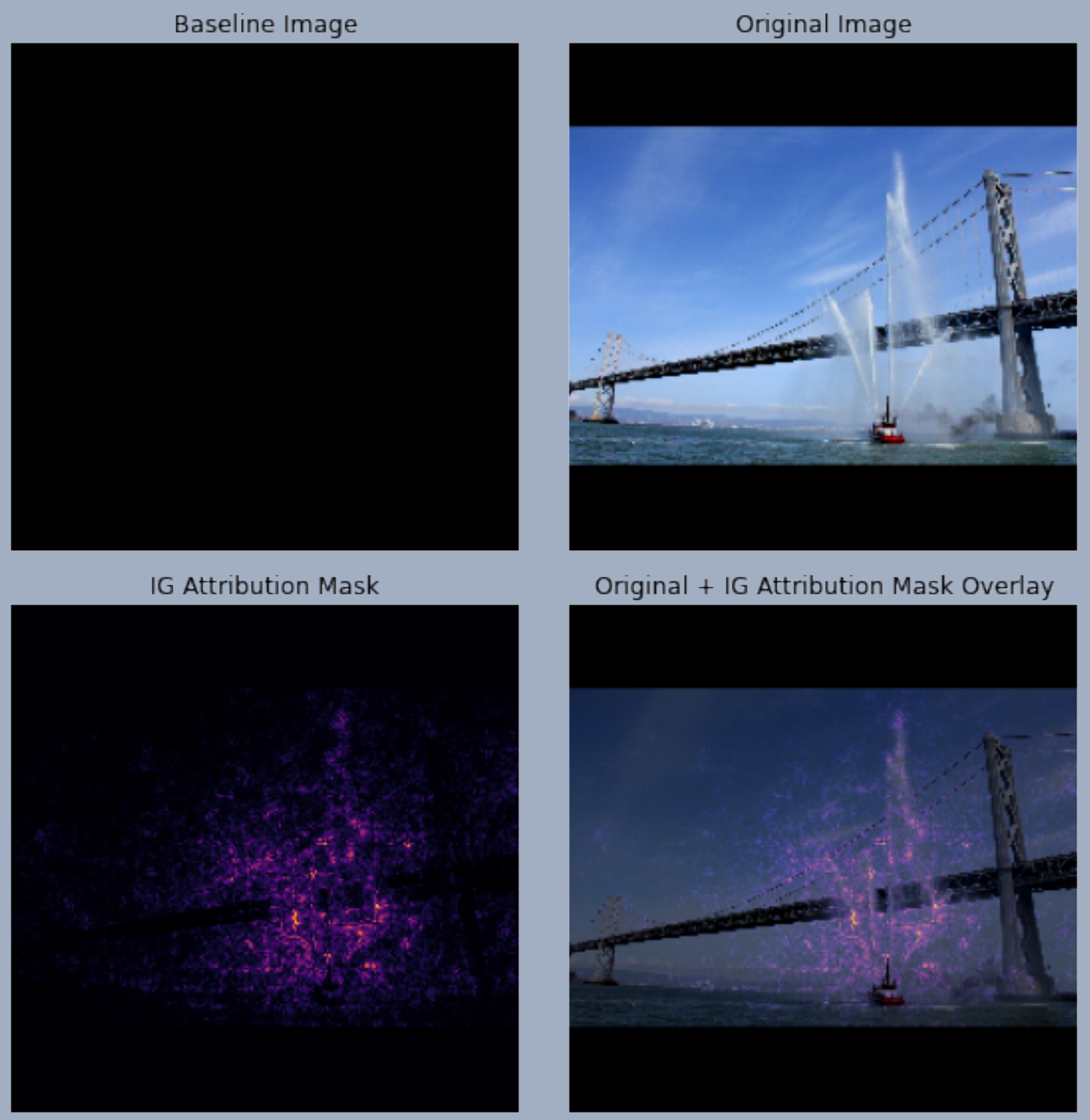

feature_1
< x
≥ x
feature_2
< y
≥ y
category_a
category_b
category_c


feature_1
< x
≥ x
feature_2
< y
≥ y
category_a
category_b
category_c


MEB Model Structure
Sigmoid
2 Layer FFN
Hidden node = 150
Hidden node = 600
Concatenate
Sum Pooling
Sum Pooling
Sum Pooling
Sum Pooling
Feature group 1
Feature group 2
Feature group 3
Feature group 49

2. Model-specific vs model agnostic

Integrated Gradients are specific to neural networks, and help visualise feature important

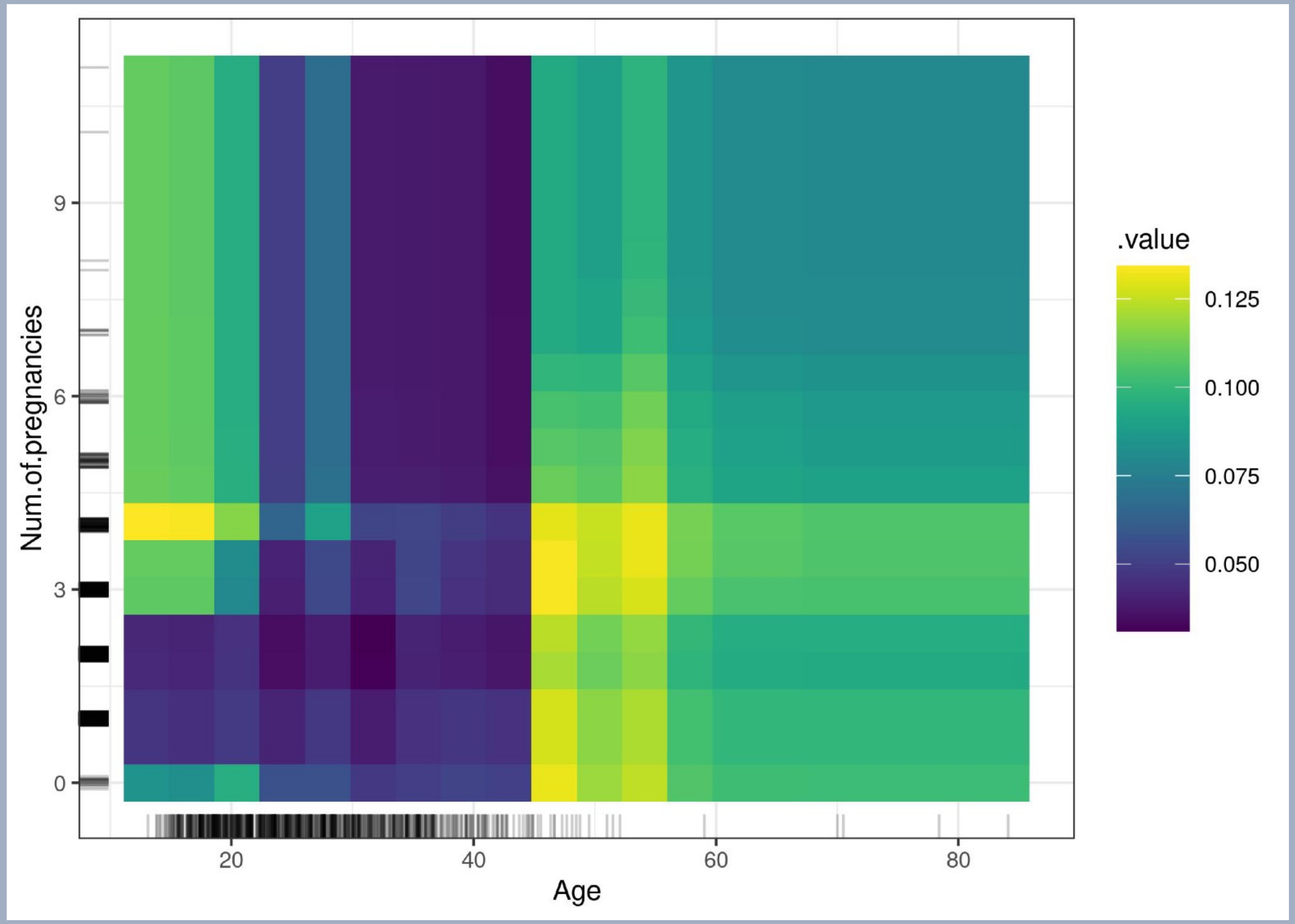


Reprinted from TensorFlow (2023)

2. Model-specific vs model agnostic

Partial dependency plots, however, are model agnostic.

They help visualise the relationship between a feature and the model's predictions.



Reprinted from Molnar (2021)

3. Global vs local

?

4. Results of interpretability methods

4. Results of interpretability methods

Feature summary statistics

4. Results of interpretability methods

Feature summary statistics

Feature summary visualisation

4. Results of interpretability methods

Feature summary statistics

Feature summary visualisation

Model internals

4. Results of interpretability methods

- Feature summary statistics
- Feature summary visualisation
- Model internals
- Data points

4. Results of interpretability methods

Feature summary statistics

Feature summary visualisation

Model internals

Data points

Intrinsically interpretable models

SECTION 3

01 What is model interpretability?

02 Methods for interpreting models

03 **Model interpretability and RRI**

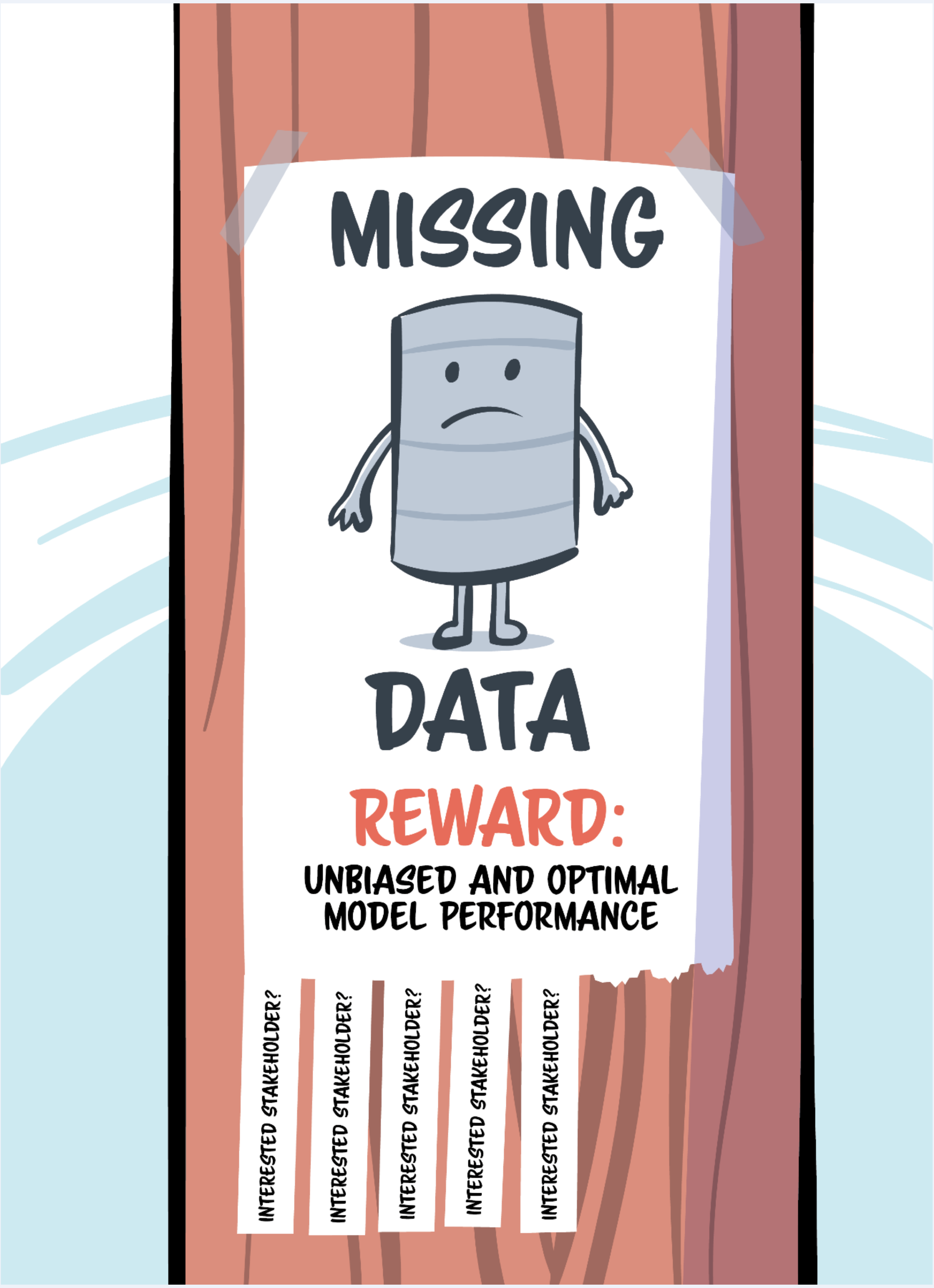
04 Building explanations

We have looked at some of the methods of model interpretability.

Now let's embed this understanding in the context of Responsible Research and Innovation.

Understanding your data

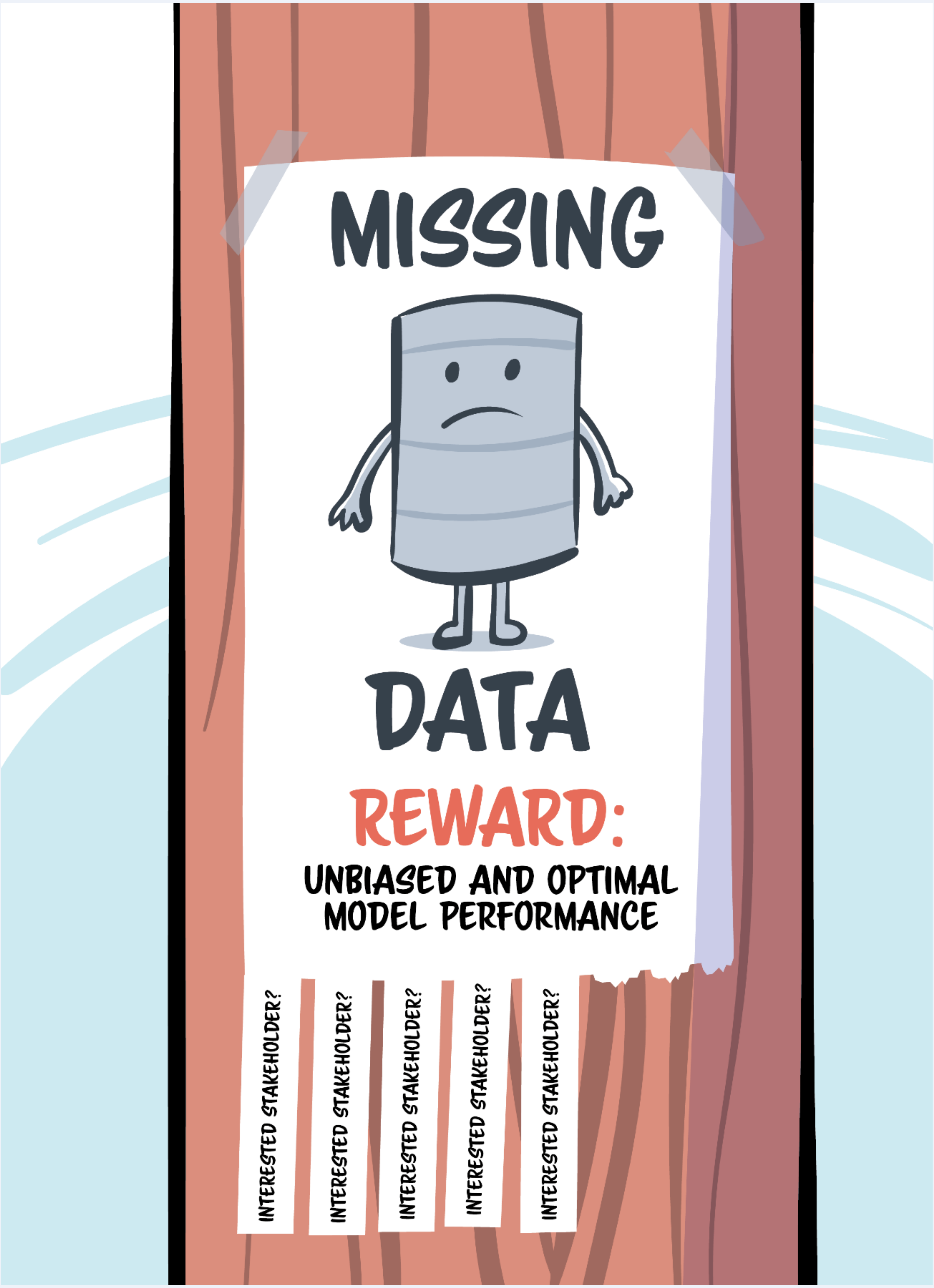
Interpretable ML methods can:



Understanding your data

Interpretable ML methods can:

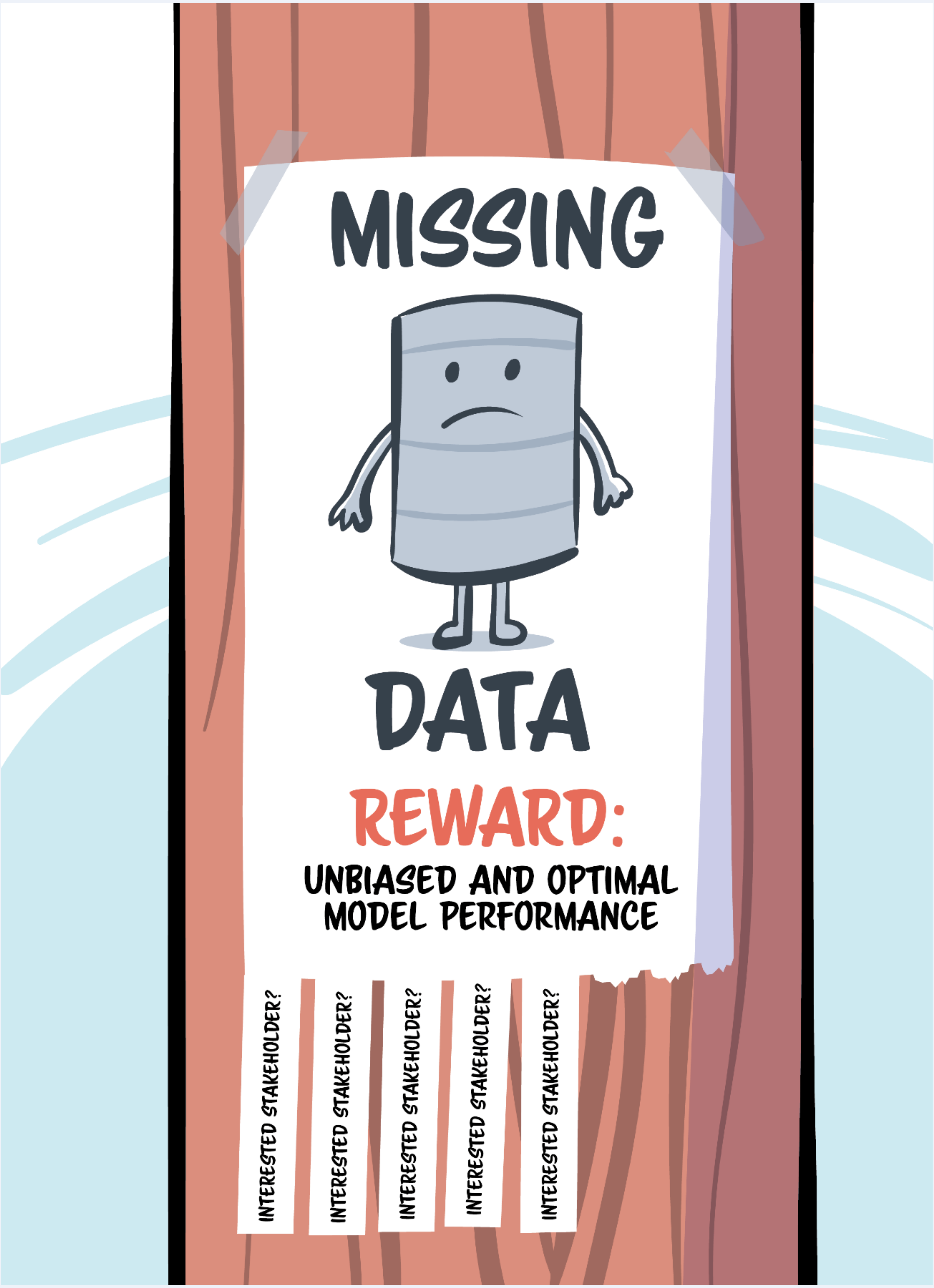
- Provide insights into input variables



Understanding your data

Interpretable ML methods can:

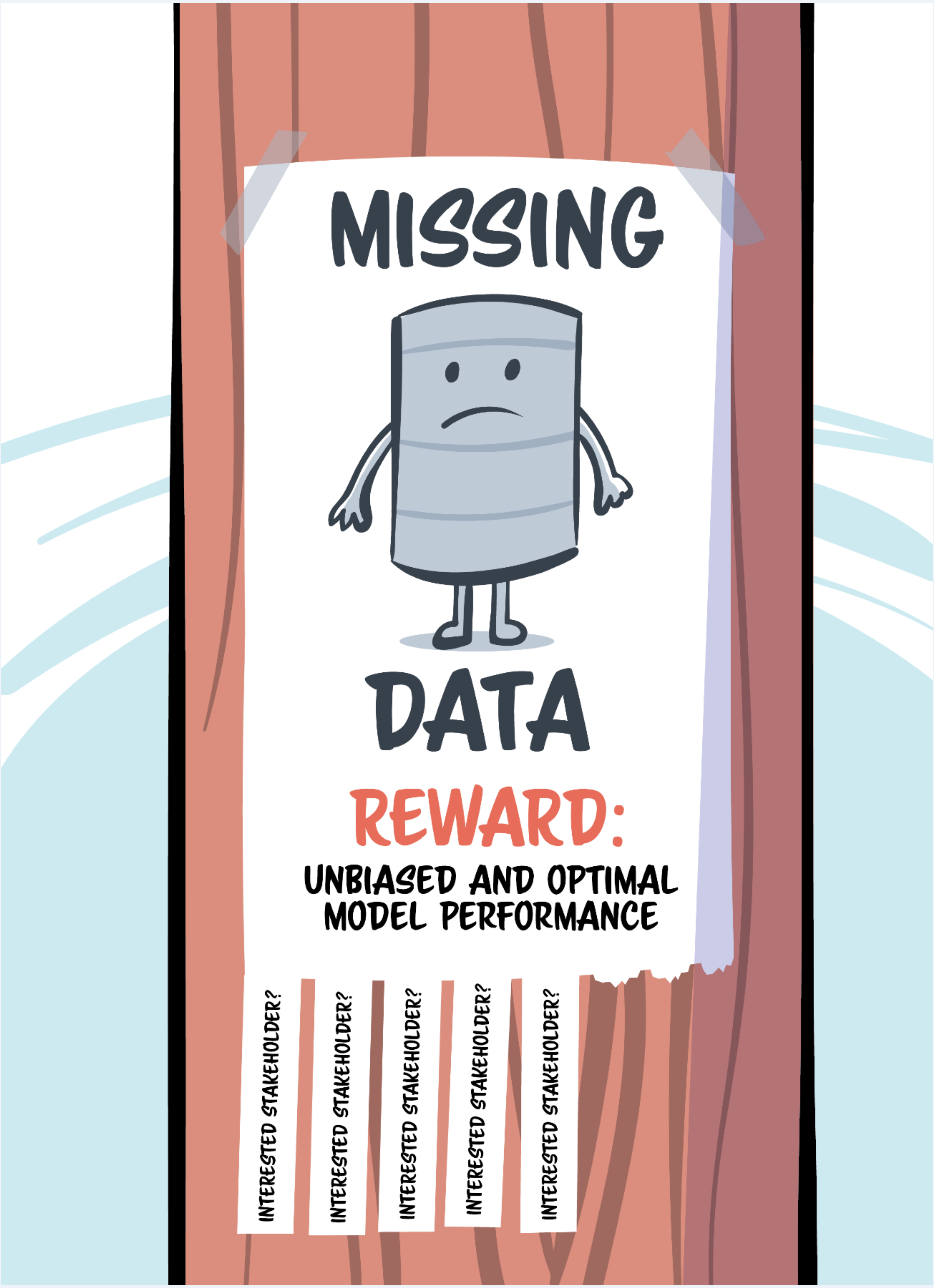
- Provide insights into input variables
- Help identify errors, biases, or gaps in dataset (e.g. missing data)



Understanding your data

Interpretable ML methods can:

- Provide insights into input variables
- Help identify errors, biases, or gaps in dataset (e.g. missing data)
- Identify patterns in large, complex, or high-dimensional datasets



However, selecting the *right interpretability method* is crucial

However, selecting the *right interpretability method* is crucial

Not all methods are created equally

Selecting the right tool for the job is an important maxim

The interpretation of a single weight always comes with the footnote that the other input features remain at the same value, which is not the case with many real applications...

A linear model that predicts the value of a house, that takes into account both the size of the house and the number of rooms, can have a negative weight for the room feature. It can happen because there is already the highly correlated house size feature. In a market where people prefer larger rooms, a house with fewer rooms could be worth more than a house with more rooms if both have the same size…

The weights only make sense in the context of the other features in the model.

— Molnar (2021)

However, selecting the *right interpretability method* is crucial

Not all methods are created equally

Selecting the right tool for the job is an important maxim

What the right tool is will be context-dependant

SECTION 3

What is model interpretability?

01

Methods for interpreting models

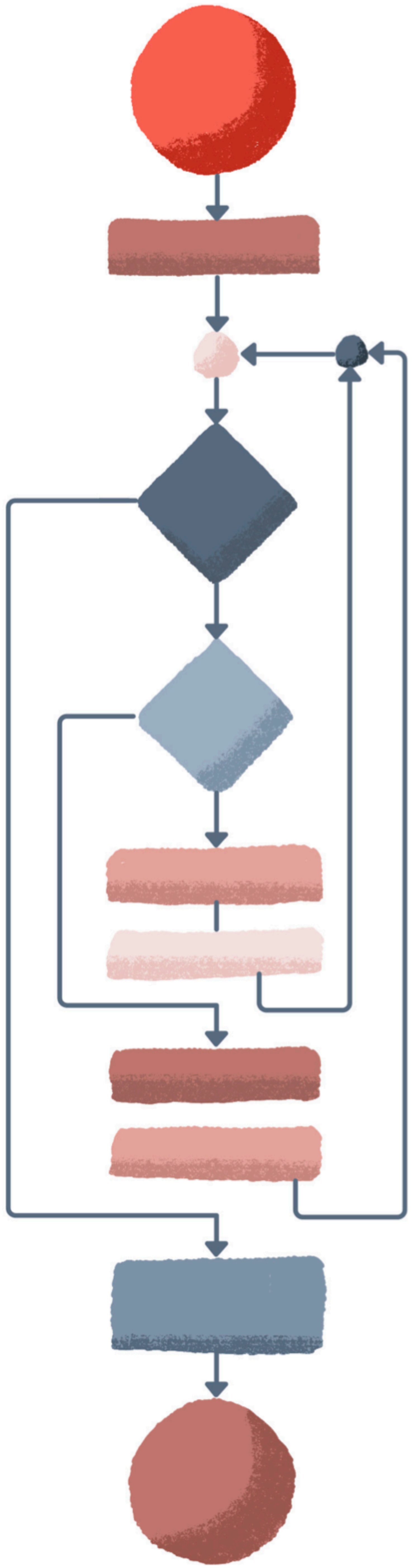
02

Model interpretability and RRI

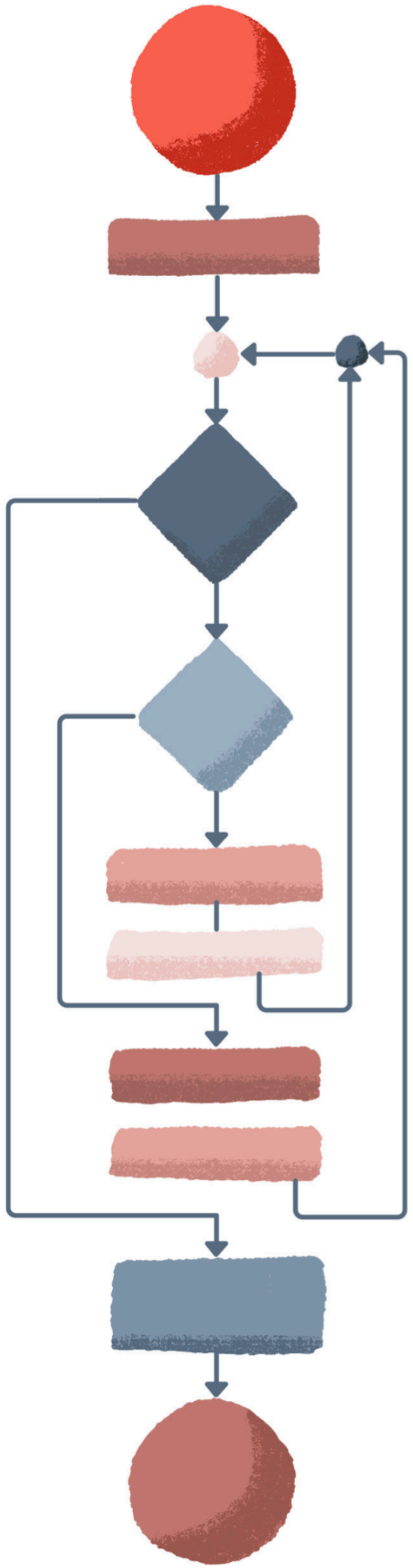
03

Building explanations

04

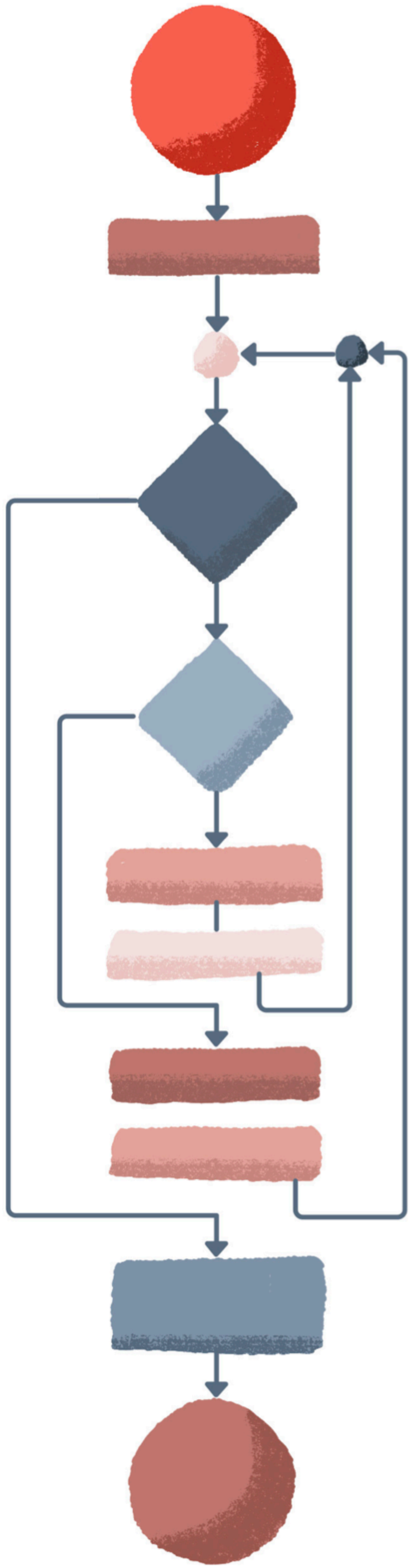


Multiple explanations for a system's behaviour may be required



Multiple explanations for a system's behaviour may be required

Therefore, even though you always need to choose the right tool for the job, you may sometimes need many tools



Multiple explanations for a system's behaviour may be required

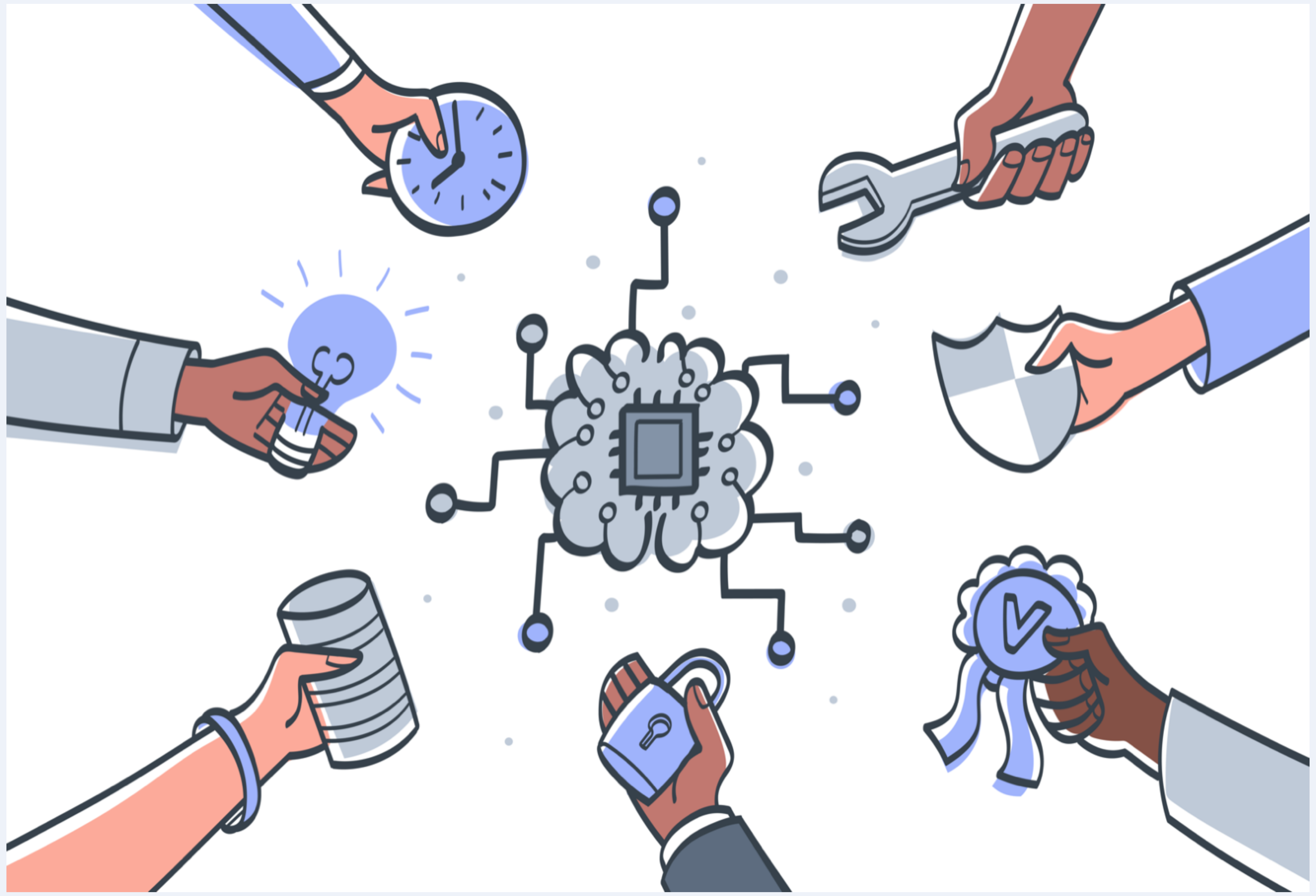
Therefore, even though you always need to choose the right tool for the job, you may sometimes need many tools

Here, again, we see another difference between interpretability and explainability—the tool you use is not the same as the objective.

Why would a stakeholder or user request an explanation for a model's behaviour or the outcomes of an AI system?

Accountability and Transparency

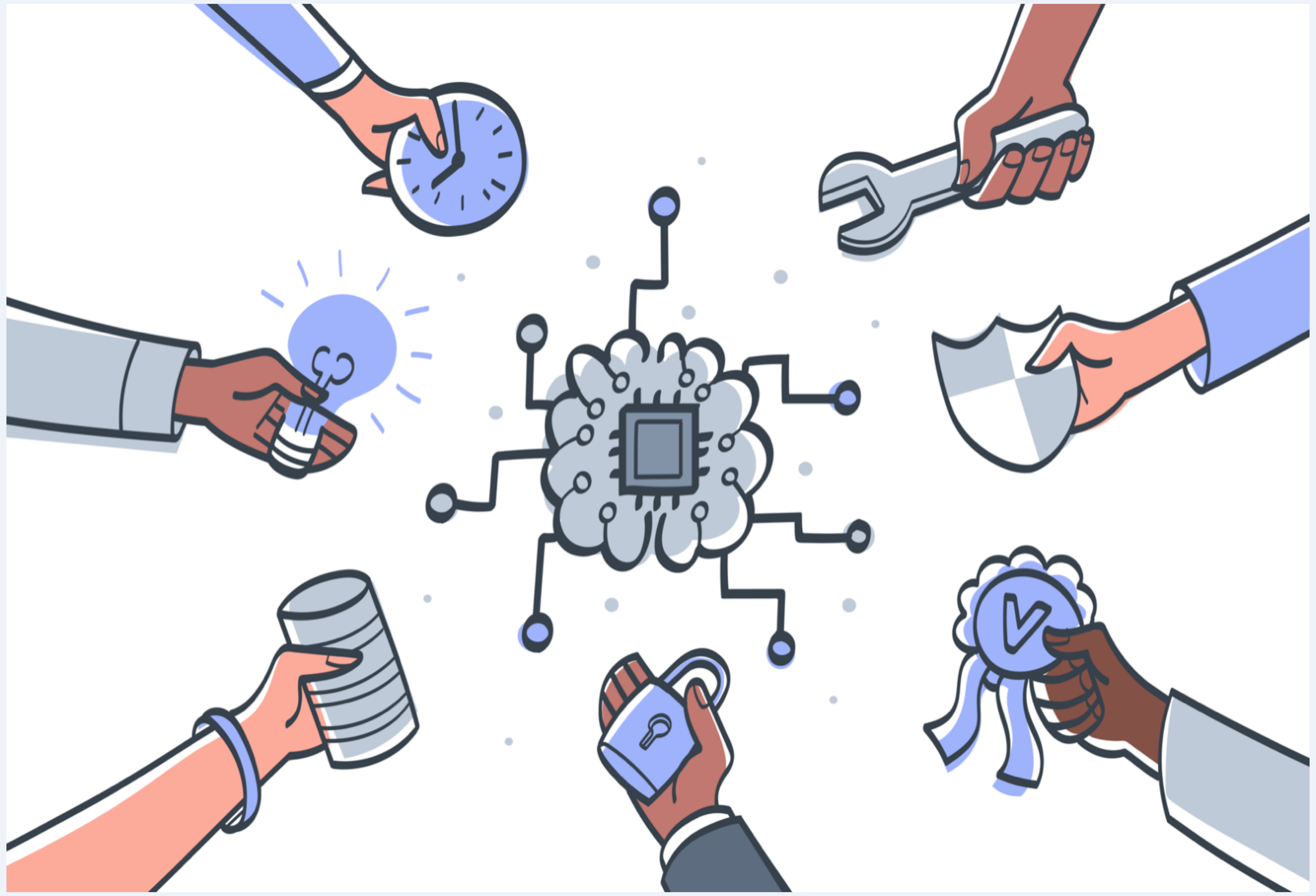
Compliance



Accountability and Transparency

Compliance

Bias detection

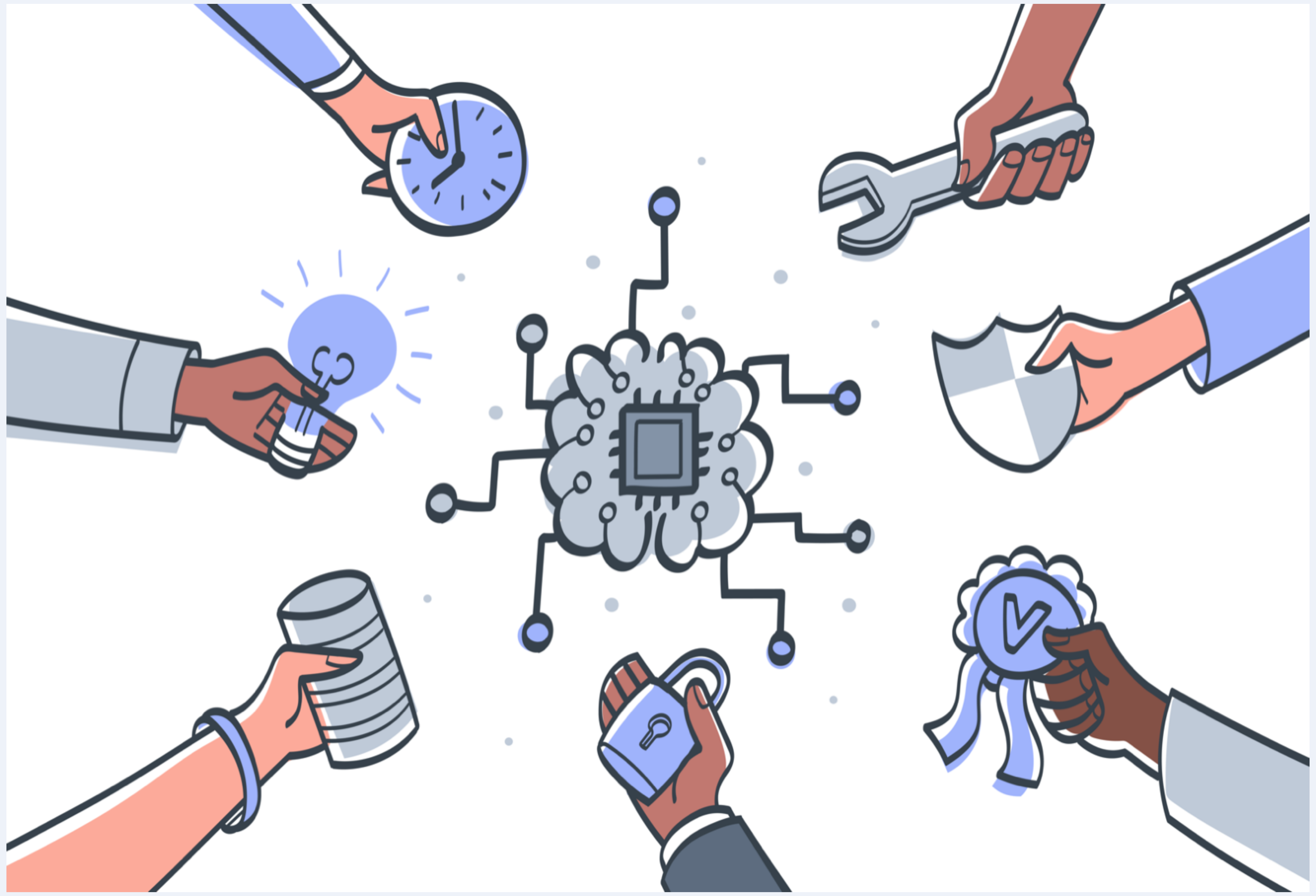


Accountability and Transparency

Compliance

Bias detection

Risk management and redress



Can you think of other reasons?

Summary

- Many model interpretability methods are unable to generate causal explanations.
- Taxonomy of interpretability methods to help provide structure.
- Interpretability methods have many uses, but it is important to “choose the right tool”.
- The right tool is dependant on the objective (e.g. compliance versus fairness statement).

Q&A

04

SITUATED EXPLANATIONS

SECTION 4

Motivation

01

Developing situated explanations

02

Proportionality and the demands of explainability

03

SECTION 4

Motivation

01

Developing situated explanations

02

Proportionality and the demands of explainability

03

sentiments.ts write_sql.go parse_expenses.py addresses.rb

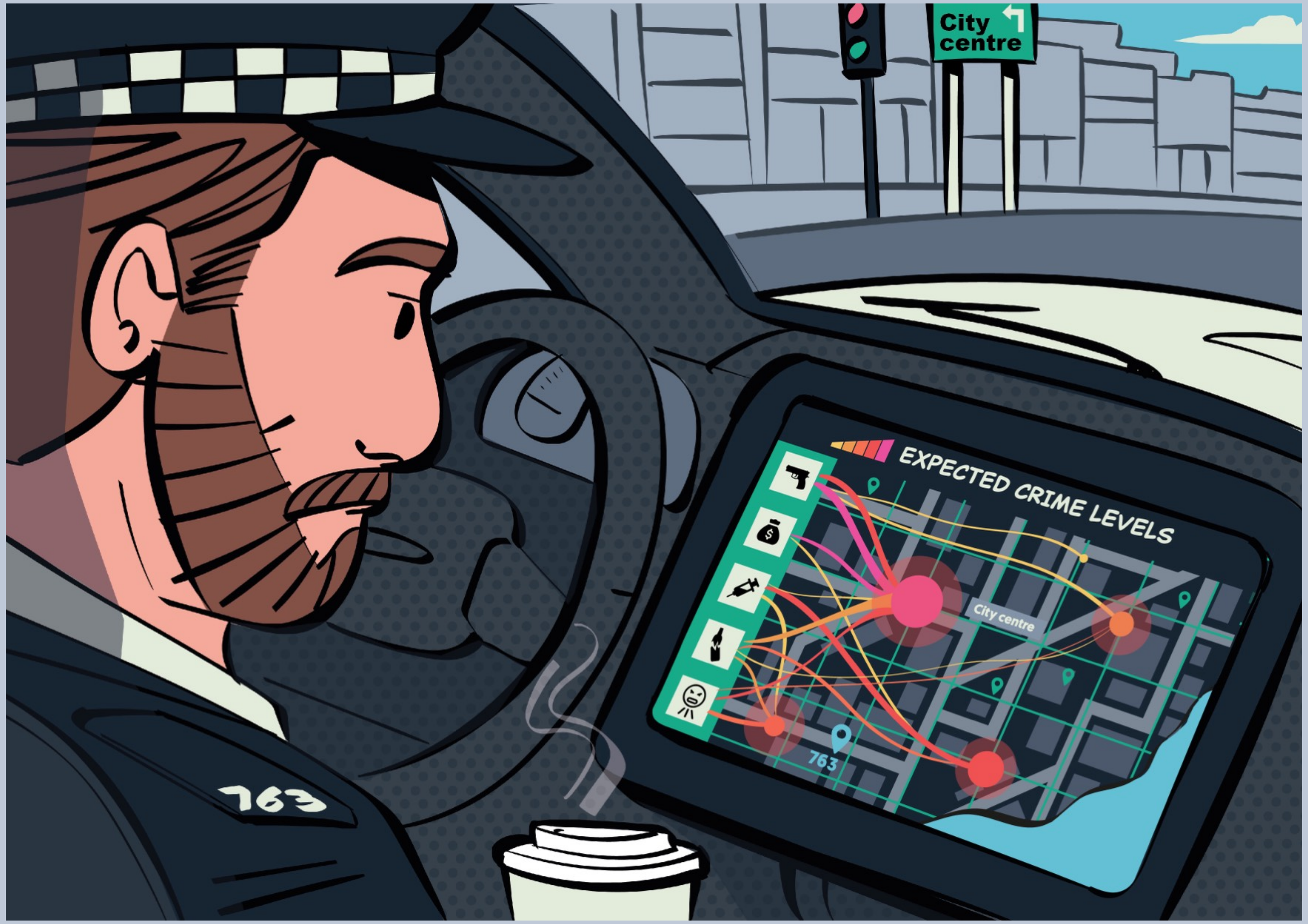
```typescript
#!/usr/bin/env ts-node

import { fetch } from "fetch-h2";

// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch(`http://text-processing.com/api/sentiment/`, {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
}
```

Copilot

The sociocultural context in which the systems are deployed matters for accessible explanations



SECTION 4

Motivation

01

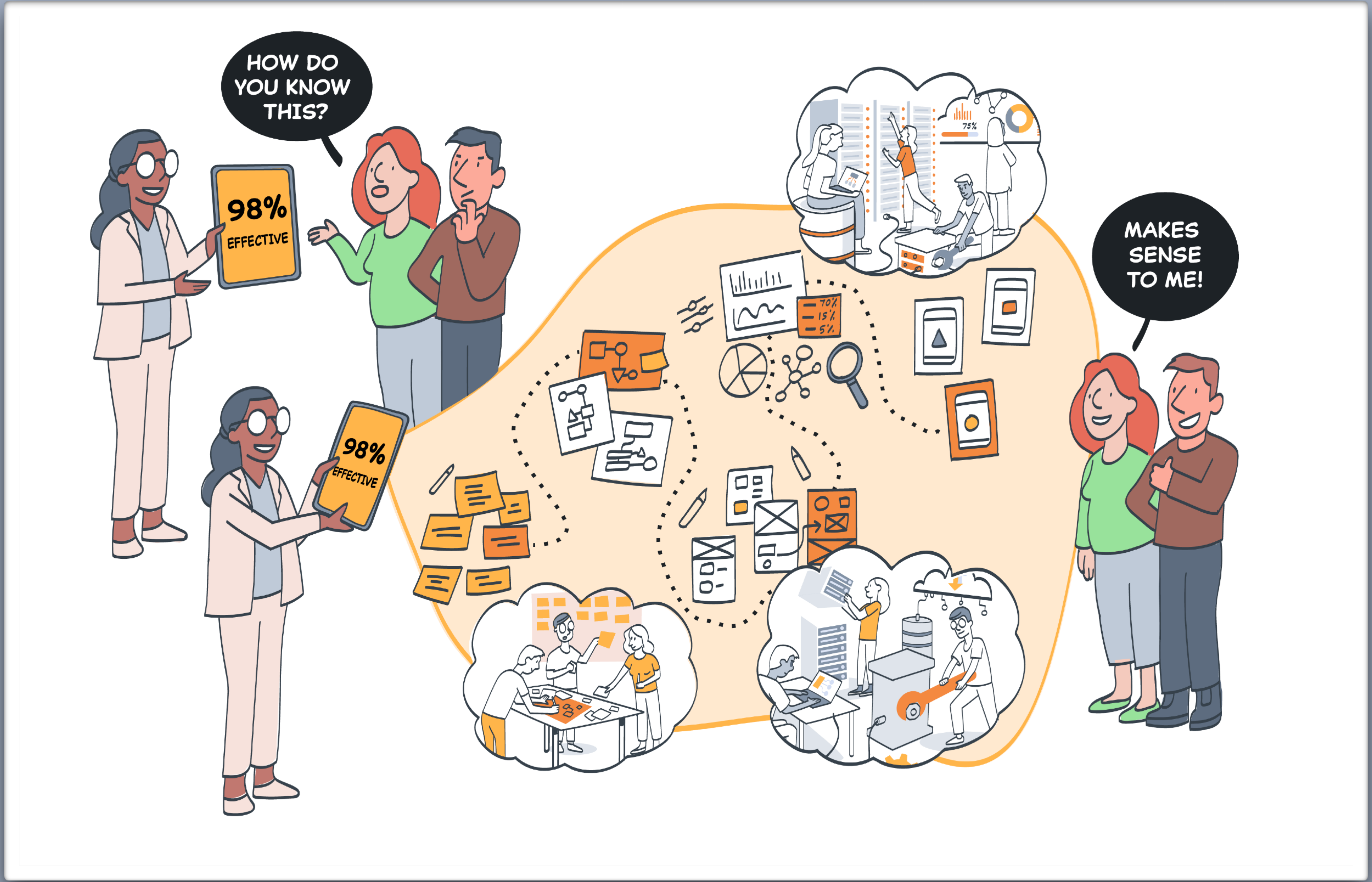
Developing situated explanations

02

Proportionality and the demands of explainability

03

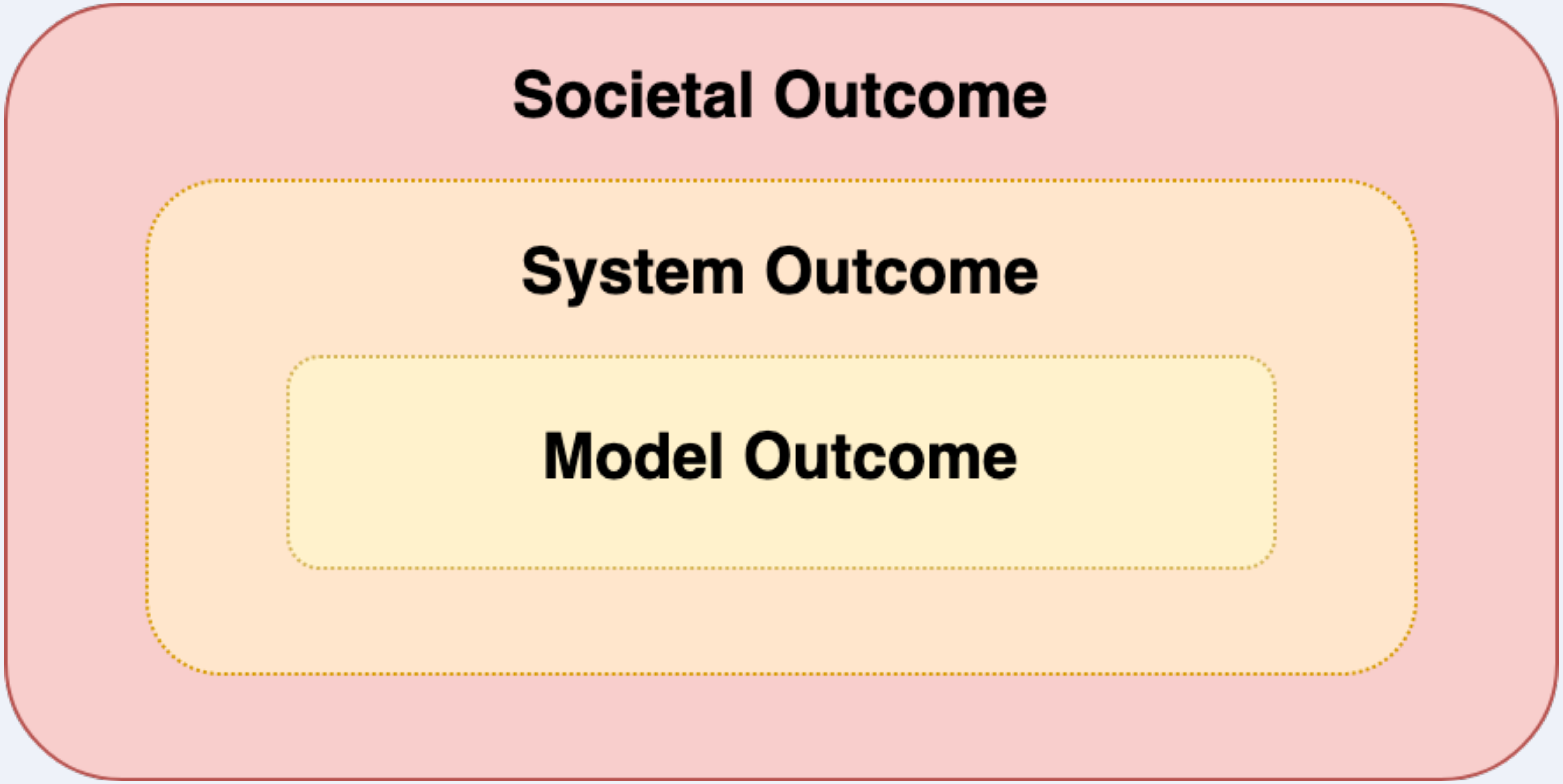
Know Your Stakeholder(s)



Two possible targets for situated explanations:

Process-based explanations

Outcome-based explanations


Societal Outcome
System Outcome
Model Outcome

Potential explanations needed:

↗ Which experts were involved in the design of the system, and how did their involvement lead to the choice of decision threshold for the classifier?

Potential explanations needed:

- Which experts were involved in the design of the system, and how did their involvement lead to the choice of decision threshold for the classifier?

- Which evaluation metrics were selected to assess the performance of the model, and why were these metrics chosen? How were biases assessed and mitigated?

Potential explanations needed:

- Which experts were involved in the design of the system, and how did their involvement lead to the choice of decision threshold for the classifier?
- Which evaluation metrics were selected to assess the performance of the model, and why were these metrics chosen? How were biases assessed and mitigated?
- Were any steps taken during the model's implementation to accommodate variations in the quality of input data (e.g. low resolution images)?


75%
?

SECTION 4

Motivation

01

Developing situated explanations

02

Proportionality and the demands of explainability

03

How much resource should a project team invest in improving the explainability of their model or system?

Improving the explainability of a system is not a trivial feat

Improving the explainability of a system is not a trivial feat

Requires technical and domain-specific expertise

Improving the explainability of a system is not a trivial feat

Requires technical and domain-specific expertise

Requires resources for clear and accesible documentation

Improving the explainability of a system is not a trivial feat

Requires technical and domain-specific expertise

Requires resources for clear and accesible documentation

Requires resources for meaningful engagement with stakeholders

Improving the explainability of a system is not a trivial feat

Requires technical and domain-specific expertise

Requires resources for clear and accesible documentation

Requires resources for meaningful engagement with stakeholders

Therefore proportionality will be required

The greater the impact and *scope* of a system, the greater the need for explainability.

Generative AI for Elevator Jingles

State Surveillance System

Summary

- Process-based explanations versus outcome-based explanations.
- Hierarchy of outcomes that may need explaining: model, system, societal.
- Knowing your stakeholders or users is essential to building situated explanations that address their requirements.
- An over-arching principle of proportionality should always be kept in mind.

Group Activity

Workshop Feedback

Thank You!

The Alan Turing Institute

Turing Commons

Responsible Research and Innovation - Explainability

Dr Chris Burr and Claudia Fischer

To find out more about Turing Commons guidebooks please visit
https://alan-turing-institute.github.io/turing-commons/