Do You Follow What I'm Explaining?

A Practitioner's Guide to Opening the AI "Black Box" for Humans

Hello!

Kilian Kluge

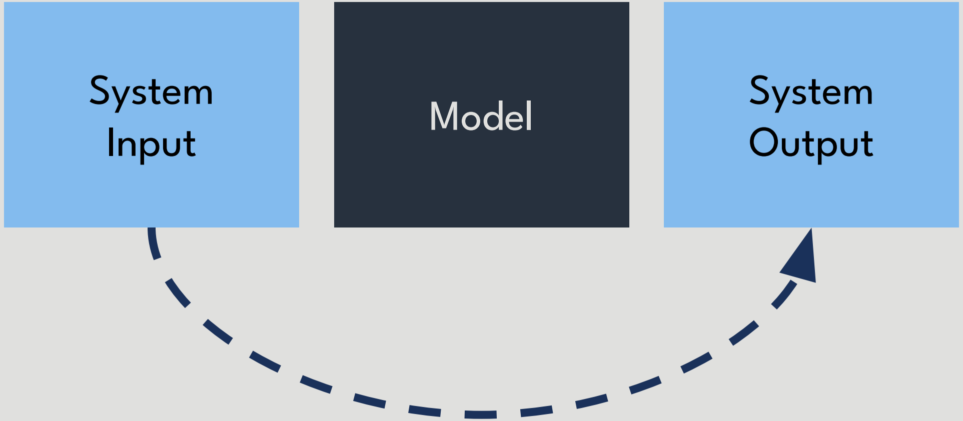
🎓

Design of Human-AI Interactions with Explainable AI

🏗️

inlinity
trusting artificial intelligence

https://github.com/ionicsolutions/do-you-follow-what-im-explaining

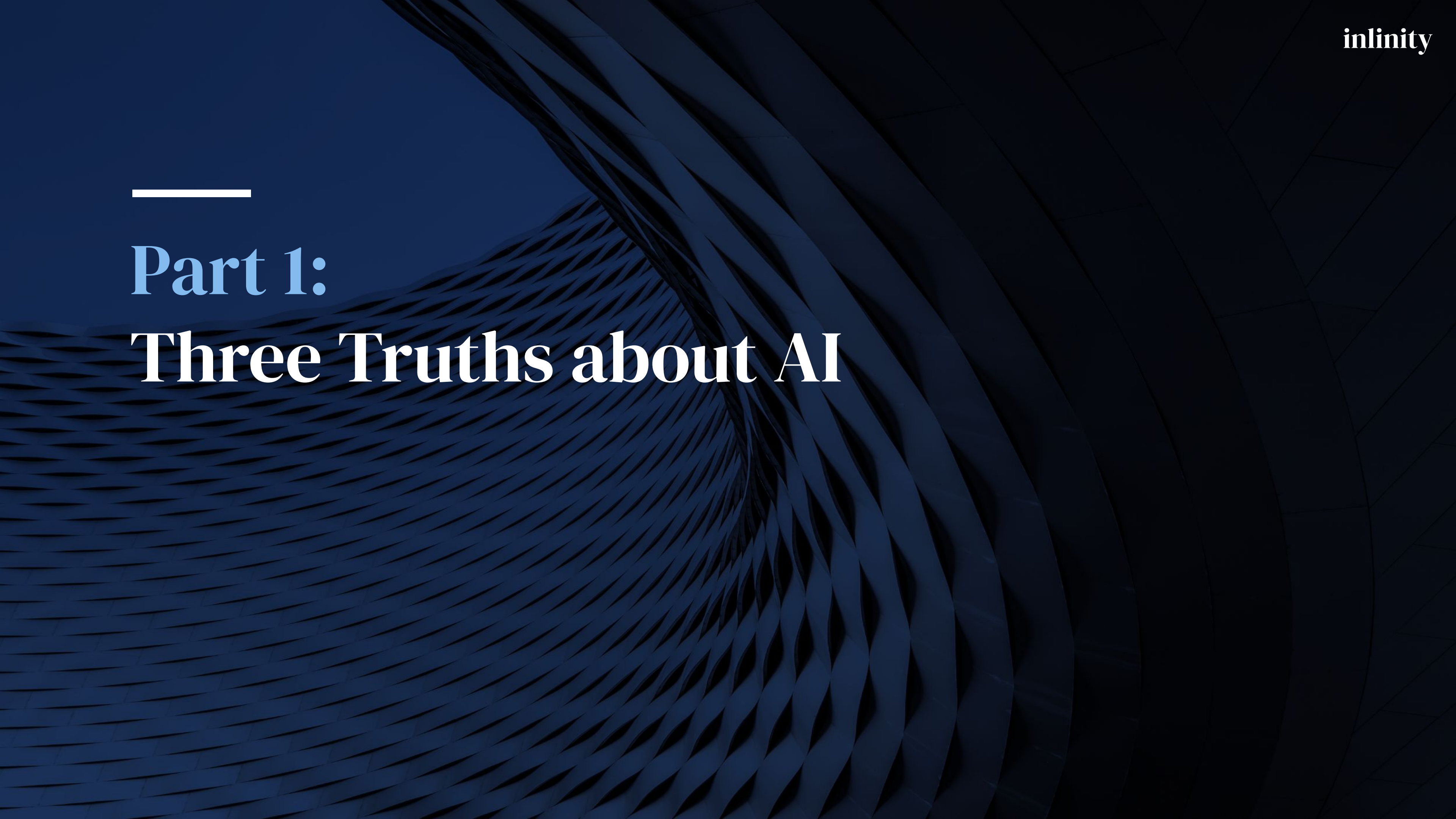


System Input

Model

System Output

🤔 😞



Part 1:
Three Truths about AI

There are no true “black boxes” in AI.

It’s all maths and bit-flipping.

Almost everything is a “black box”
to your users.

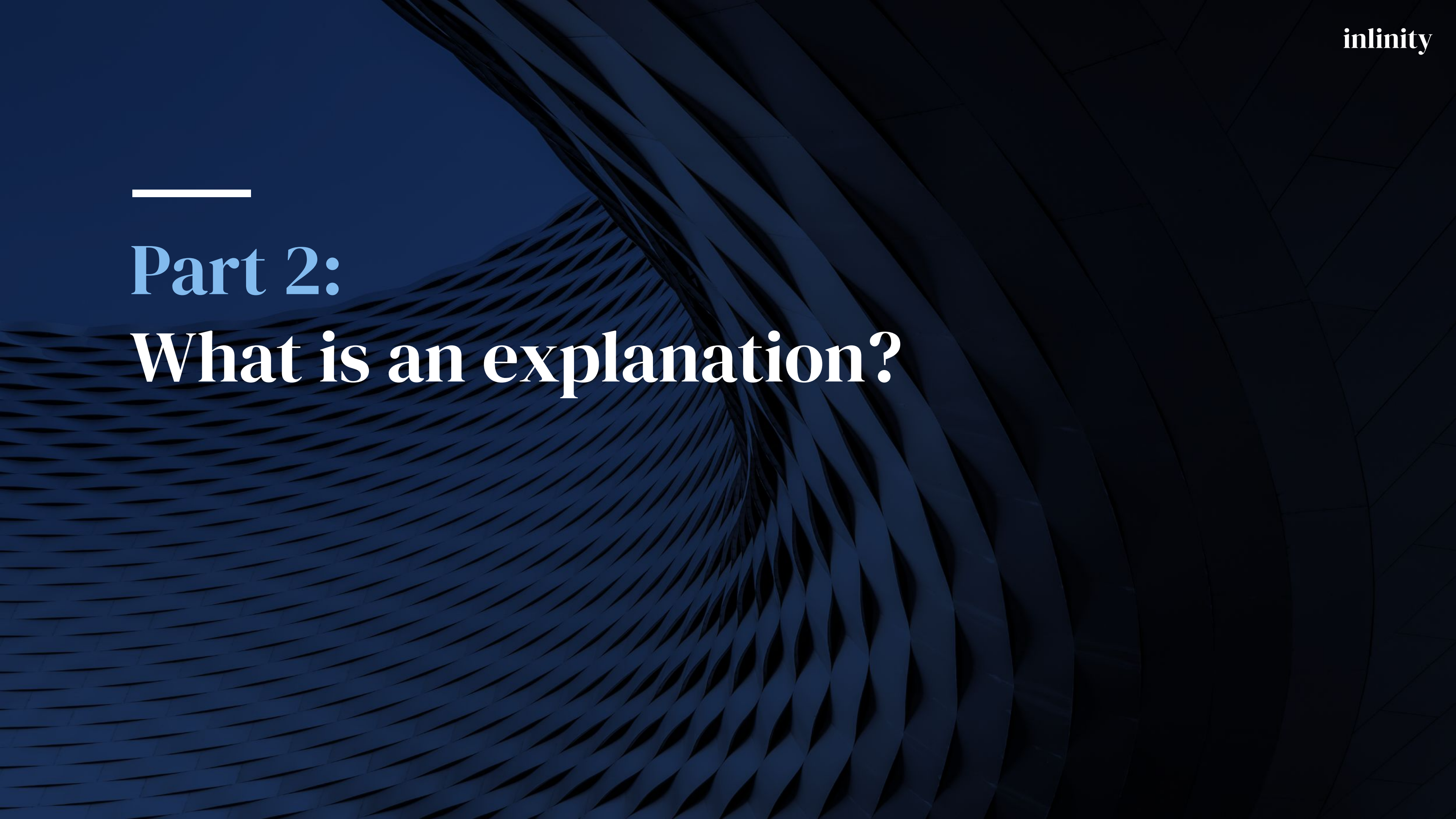
Even the most simple of models.

Your users don’t care about your model.

They care about the decision, recommendation, or prediction.

User-Centric Explainable AI

aims to enable humans to manage
the complexity of AI systems
by generating explanations.



Part 2:

What is an explanation?

Explaining is a social activity.

People don’t want to know the entire causality chain.

They want an explanation.

An explanation delivers or contains accompanying evidence or reason(s) for outputs and/or processes

Philipps et al.: Four Principles of Explainable Artificial Intelligence (NIST, 2021)

Part 3:

What is a good explanation?

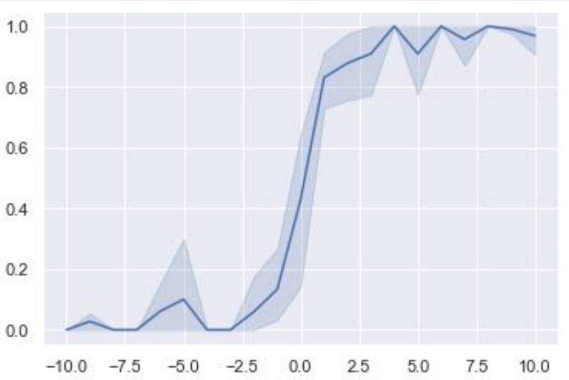
The quality and usefulness of explanations are highly subjective.

You cannot capture these concepts with simple metrics.

"Short explanations are better."

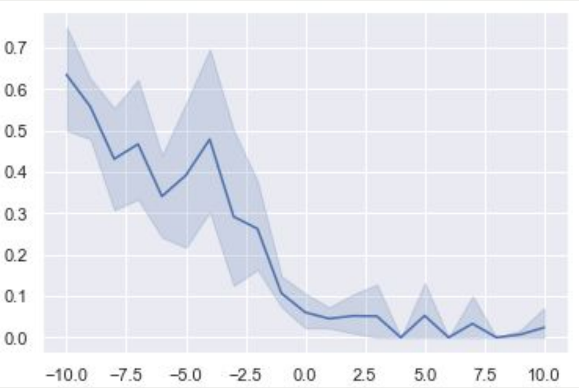
Obviously. No! Maybe?

“It’s longer”



“measured” length difference

“It’s shorter”



“measured” length difference

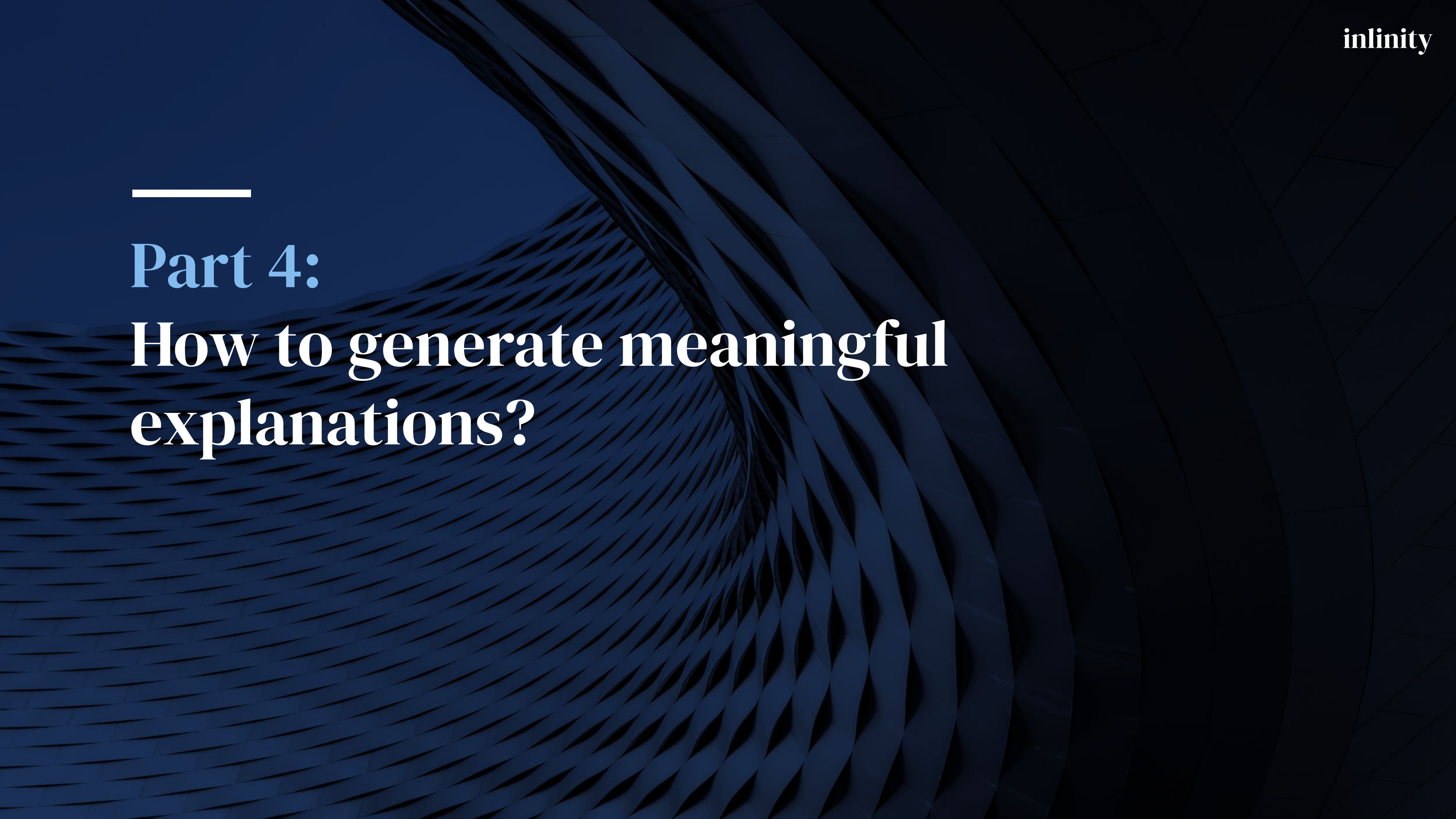
In 60% of cases, users preferred explanations because they were longer.

Accuracy: The explanation correctly reflects the reason for generating the AI system’s output and/or accurately reflects its internal processes

Philipps et al.: Four Principles of Explainable Artificial Intelligence (NIST, 2021)

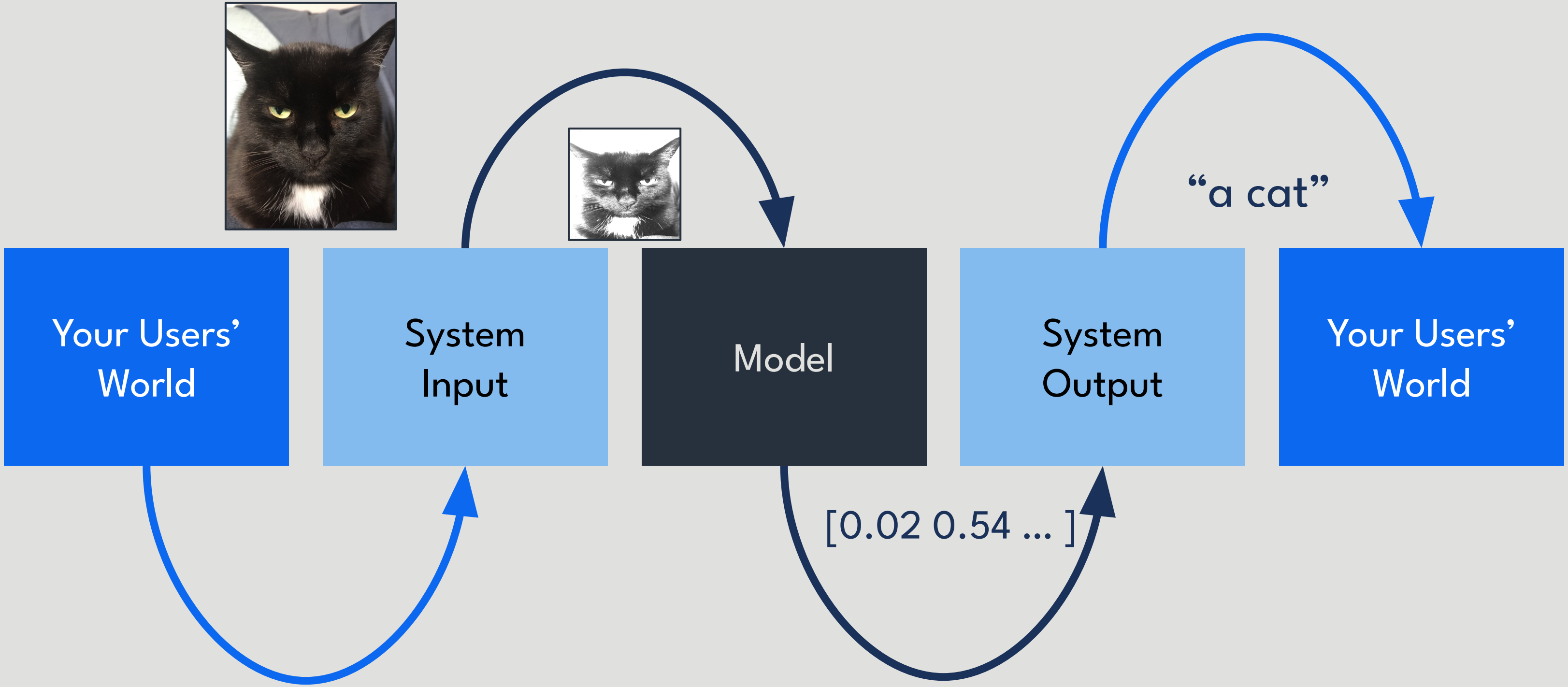
Meaningfulness: The provided explanations are understandable to the intended consumer(s)

Philipps et al.: Four Principles of Explainable Artificial Intelligence (NIST, 2021)



Part 4:
How to generate meaningful explanations?

From the users' point of view, there are two stages of pre- and postprocessing.



Your Users’ World

System Input

Model

[0.02 0.54 ...]

System Output

“a cat”

Your Users’ World

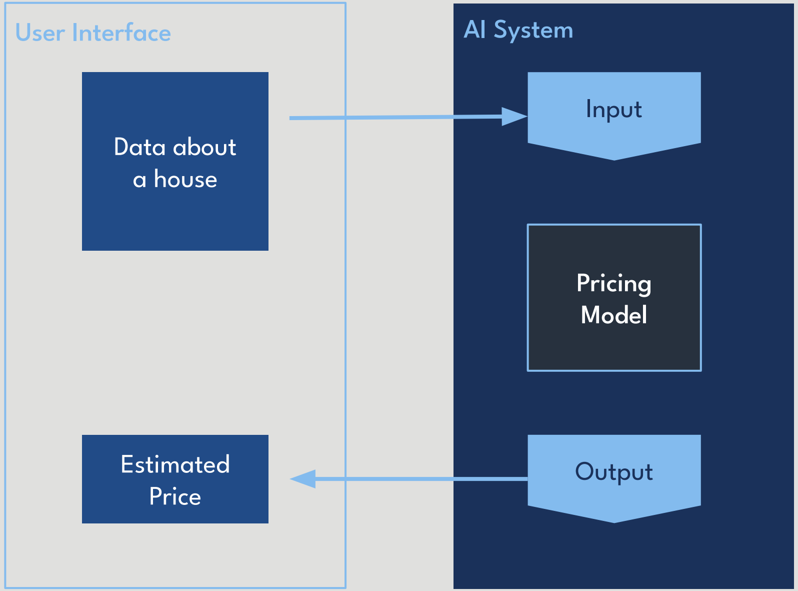
Only your users know what’s meaningful.

You need to ask them!

Example:

Generating Meaningful Explanations for Real Estate Price Predictions

The Scenario



User Interface

Data about a house

Estimated Price

AI System

Input

Pricing Model

Output

The Question

Why is my house worth X?

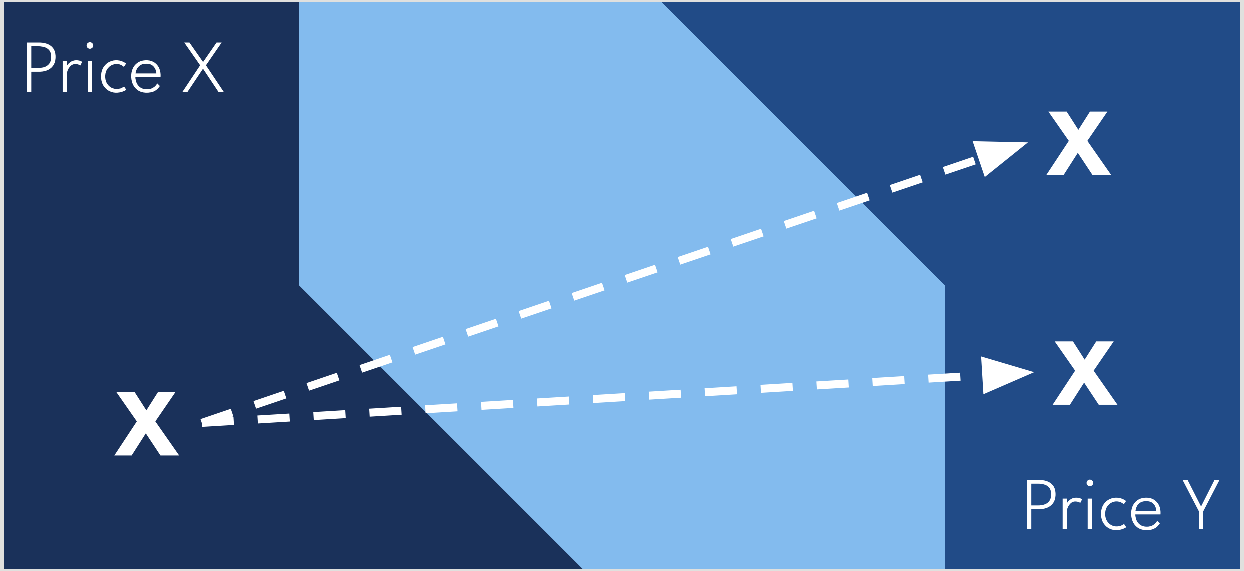
The Real Question

Why is my house worth X
instead of Y?

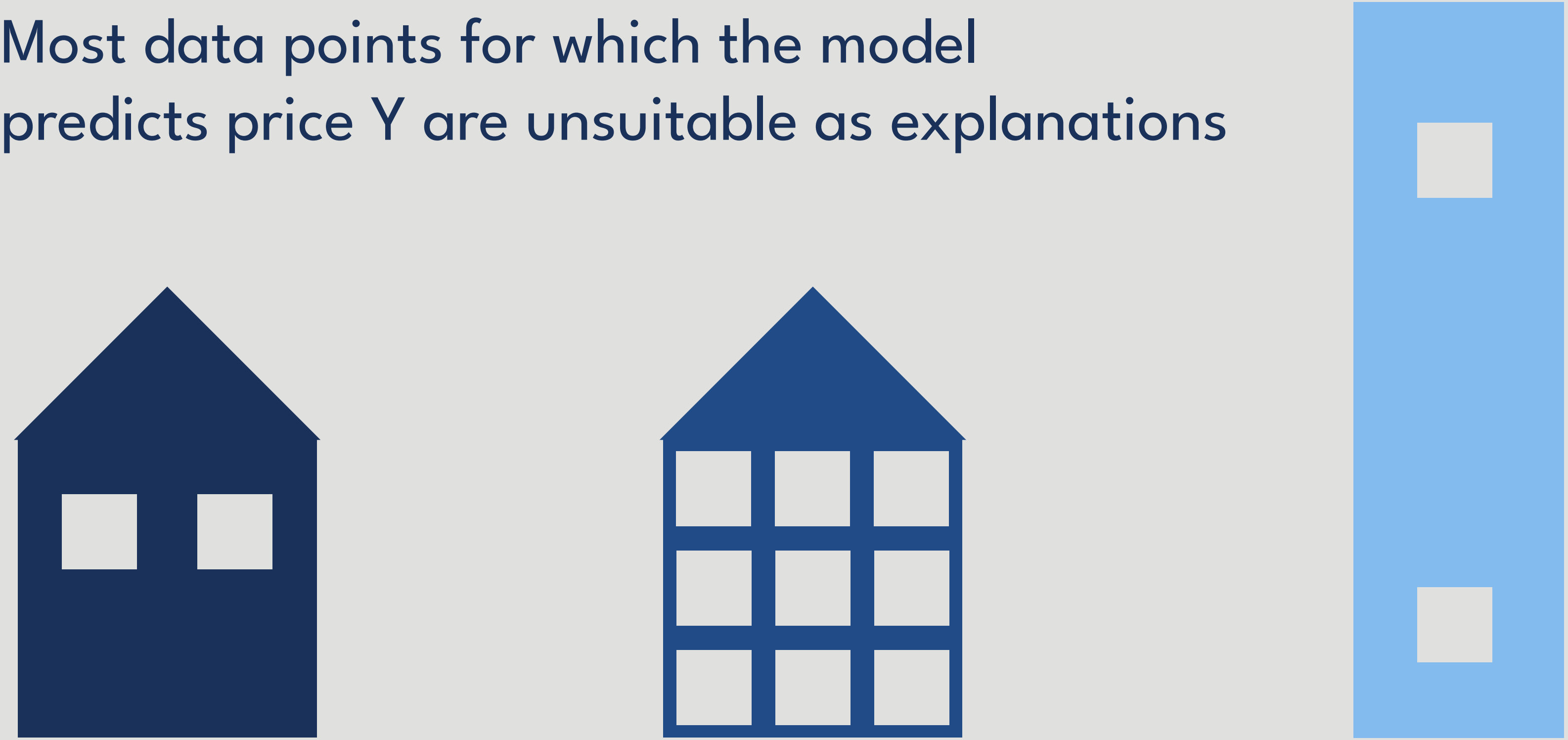
Find a data point for which the pricing model predicts a higher price.

inlinity

Counterfactual Explanations



Most data points for which the model predicts price Y are unsuitable as explanations



Find a data point for which the pricing model predicts a higher price that is perceived as realistic and typical.

Do users think the generated data points are realistic and typical?

Let's ask them!

... using established constructs

(in our case, from communications & media research)

Everything else being equal,
the house would be worth 600 k€ instead of 350 k€
if it was built in 2005 instead of 1999,
had a living space of 233 m^2 instead of 190 m^2,
and a lot size of 714 m^2 instead of 490 m^2.

Summary

Explanation

Delivers or contains accompanying evidence or reason(s) for outputs and/or processes

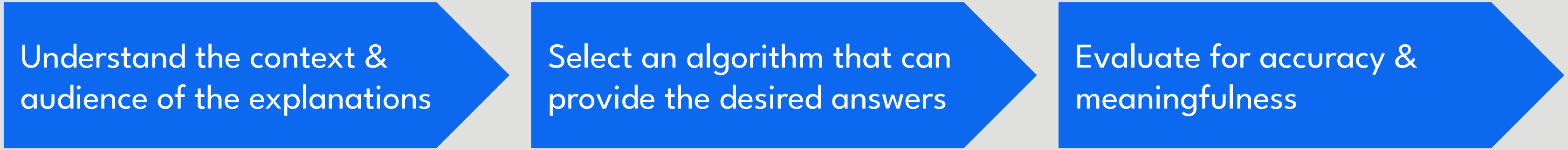
Meaningfulness

The provided explanations are understandable to the intended consumer(s)

Accuracy

The explanation correctly reflects the reason for generating the output and/or accurately reflects the AI system's internal processes

Philipps et al.: Four Principles of Explainable Artificial Intelligence (NIST, 2021)



Understand the context & audience of the explanations

Select an algorithm that can provide the desired answers

Evaluate for accuracy & meaningfulness

Thank You!

Explainable AI Slack Community

Twitter: @XAI_Research
explainableaiworld@gmail.com

kilian.kluge@inlinity.ai

https://github.com/ionicsolutions/do-you-follow-what-im-explaining