Explaining AI systems

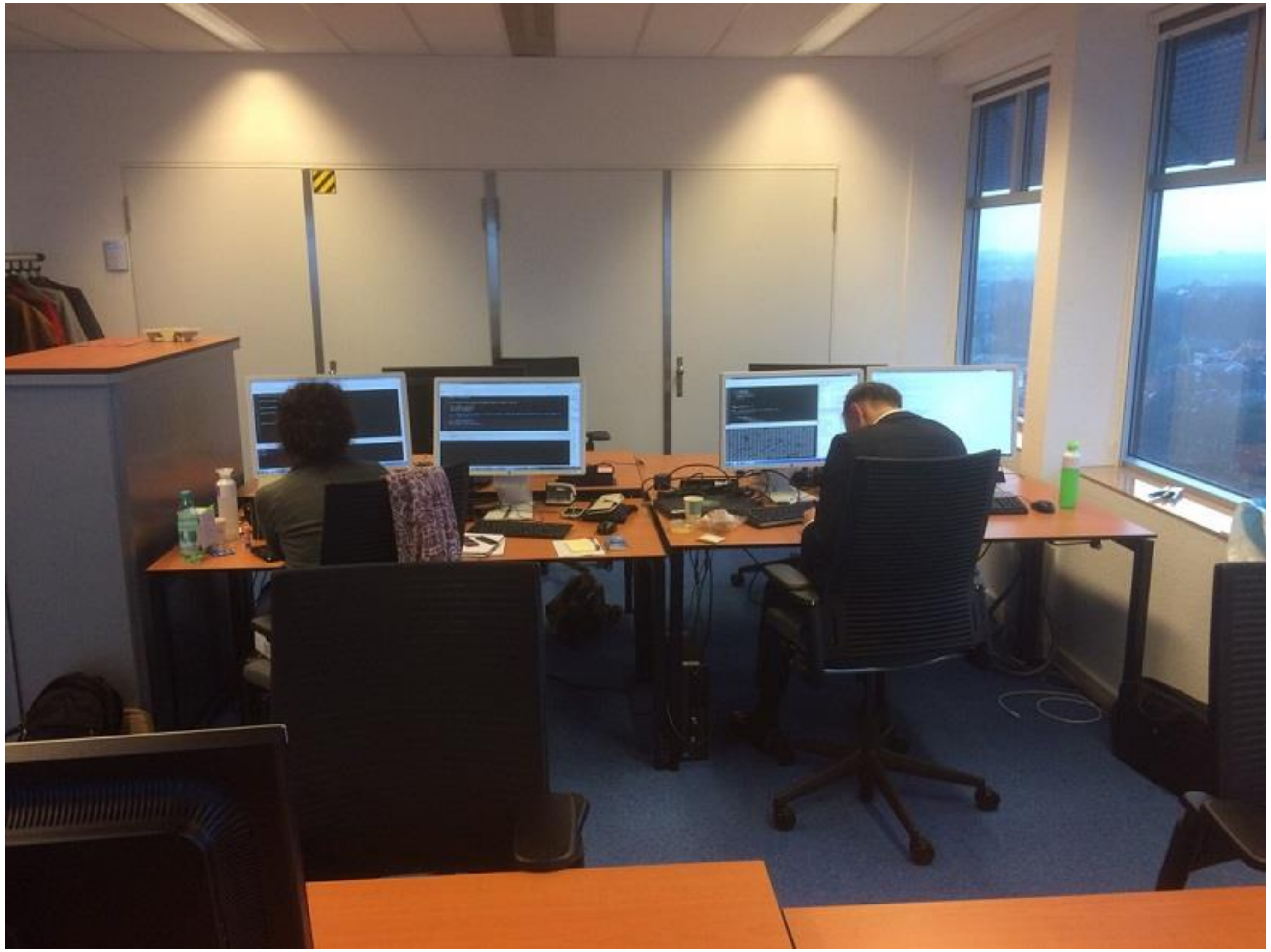
Hinda Haned
June 20^{th} 2019

Ahold Delhaize
University of Amsterdam

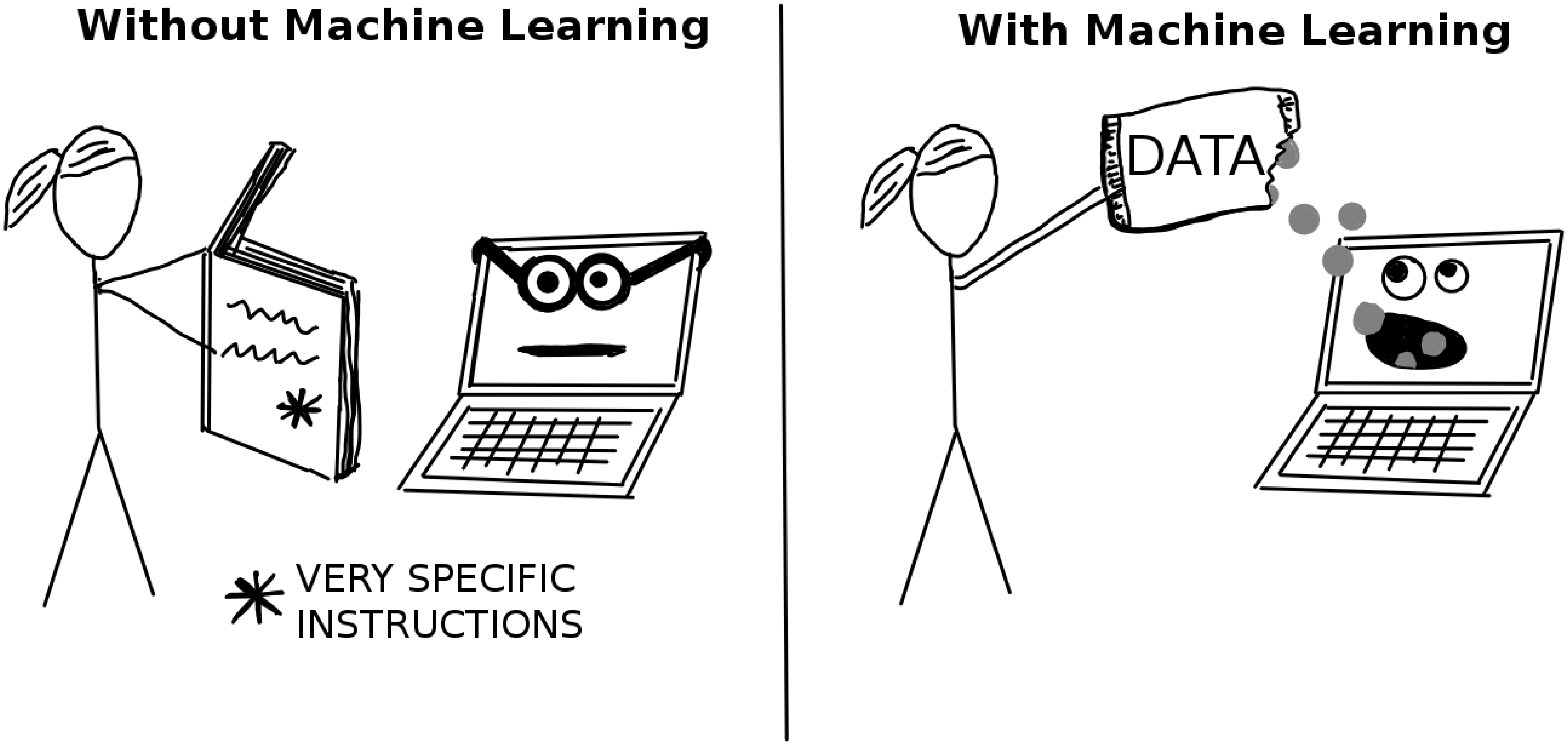
Artificial Intelligence (AI)

Retrieved from Google image search - June 17th 2019

In real life

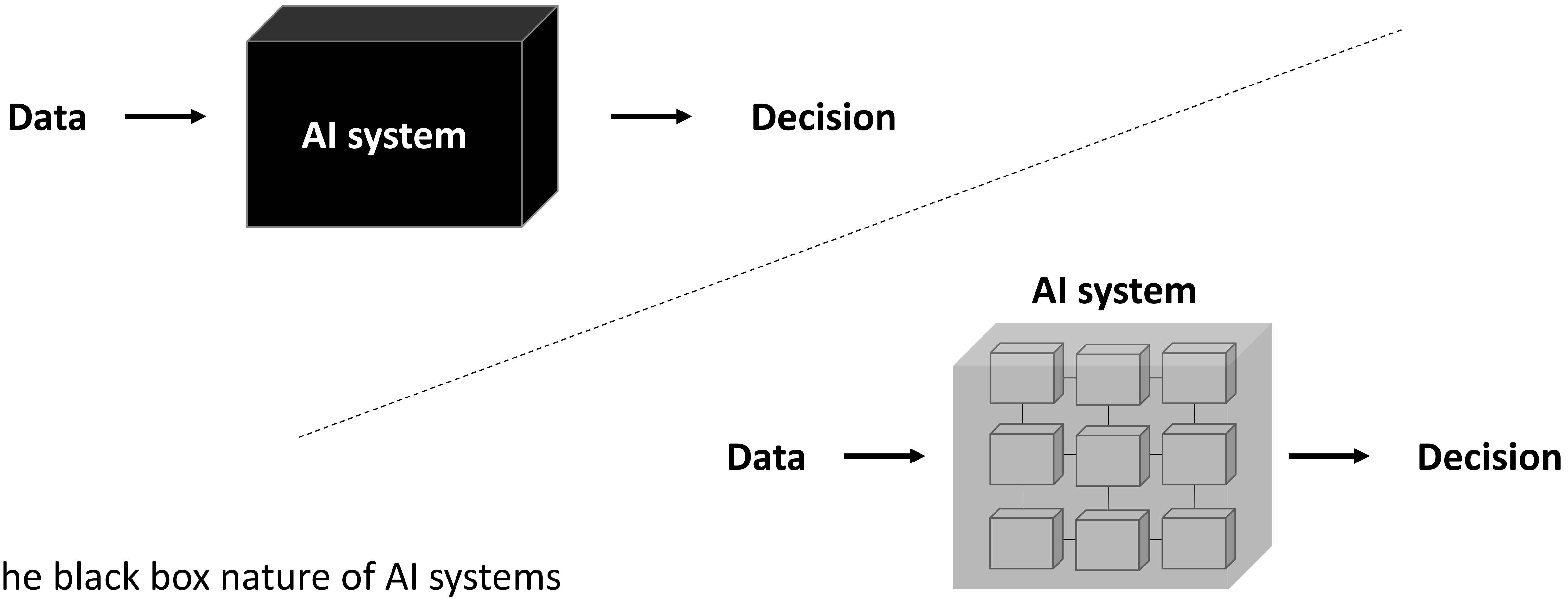


Machine learning



Source: https://christophm.github.io/interpretable-ml-book/

Black box metaphor



The black box nature of AI systems comes from the interaction of many simple components

AI & Explanations: debate

Against

- Not compatible with the purpose of AI systems to begin with
- Holding AI to above humans standards is unrealistic
- Explainability sacrifices performance for user-trust

In favor

- Transparency helps improve the AI-system
- Explanations help users trust AI-driven systems
- Compliance with regulations (Europe)

Explaining AI: why bother?

Compliance with General Data Protection Regulation (GDPR – 2018) – articles 13-15

“Data subjects have a right to “meaningful information about the logic involved” and to “the significance and the envisaged consequences” of automated decision-making”

An example: SPUI25 Forecaster

Prediction Task How many people will attend an event at SPUI 25?

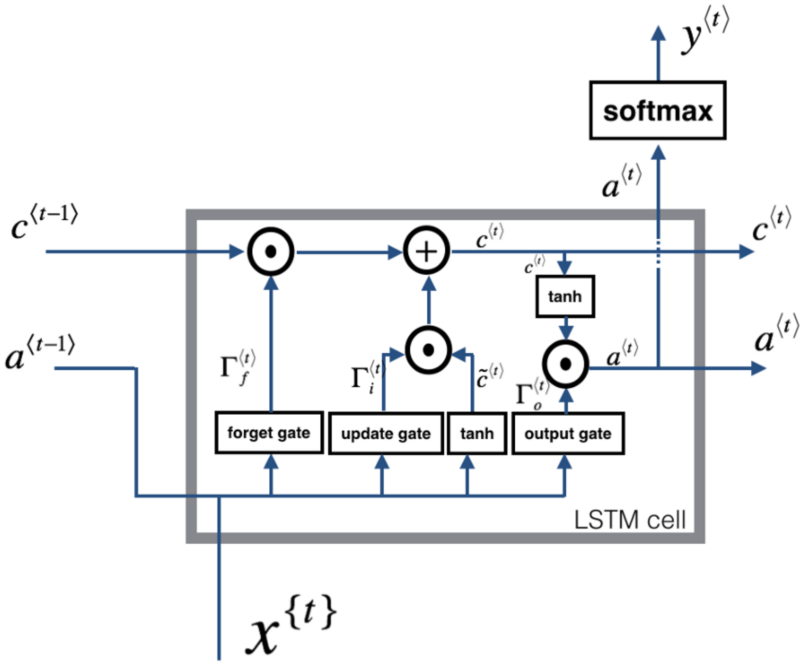
SPUI25 Forecaster

An example: SPUI25 Forecaster

Data

- Past experience at SPUI 25
- Weather on Thursday night?
- Is it raining?
- Free drinks afterwards?
- Is there football tonight?
- Subject popularity score
- Speakers popularity score
- ….

Model



. . .

$$\Gamma_f^{\langle t\rangle} = \sigma(W_f[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_f)$$
$$\Gamma_u^{\langle t\rangle} = \sigma(W_u[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_u)$$
$$\tilde{c}^{\langle t\rangle} = \tanh(W_C[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_C)$$
$$c^{\langle t\rangle} = \Gamma_f^{\langle t\rangle} \circ c^{\langle t-1\rangle} + \Gamma_u^{\langle t\rangle} \circ \tilde{c}^{\langle t\rangle}$$
$$\Gamma_o^{\langle t\rangle} = \sigma(W_o[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_o)$$
$$a^{\langle t\rangle} = \Gamma_o^{\langle t\rangle} \circ \tanh(c^{\langle t\rangle})$$

Long Short-Term Memory (LSTM) cell structure

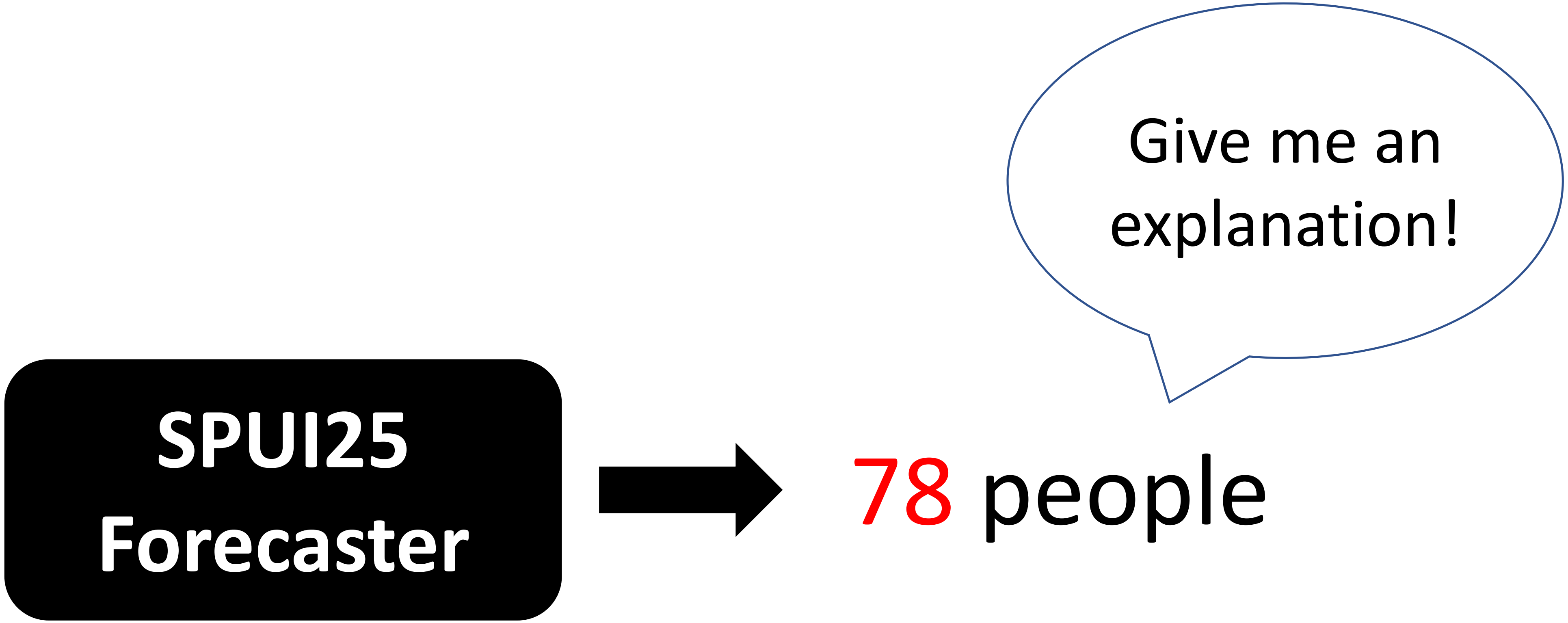
Source: https://iitmcvg.github.io/

An example: the SPUI25 Forecaster

Prediction How many people will attend tonight's event at SPUI 25?

SPUI25 Forecaster → 78 people

An example: the SPUI25 Forecaster



What is an explanation?

An explanation is the answer to a why-question (Miller 2017)

- Why this book recommendation?
- Why does this shopping app think I am pregnant?
- Why did I not get the loan?
- Why did the car crash?

What is a good explanation?

Research shows that people do not explain the causes for an event, but explain the cause of an event relative to some other event that did not occur; that is, an explanation is always of the form:

“Why A and not B”

This is called a contrastive explanation

T. Miller (2017)

Generating contrastive explanations

Counterfactual examples

describes the smallest change to the feature values that changes the prediction to a predefined output

how would the prediction have been if input X had been different?

Surrogate models

simplified local version of the black box model

features that influenced the predictions vs. those that were absent

Contrastive explanations

SPUI25
Forecaster

Which feature values **must be changed** to increase the number of people to 100?

- If the temperature was increased by 10 degrees, the prediction would be 100 people
- If the speaker popularity score was up by 7 points, the prediction would be 100 people

Which features **were not important** for the prediction?

- Number of free drinks after the talk

Limitations

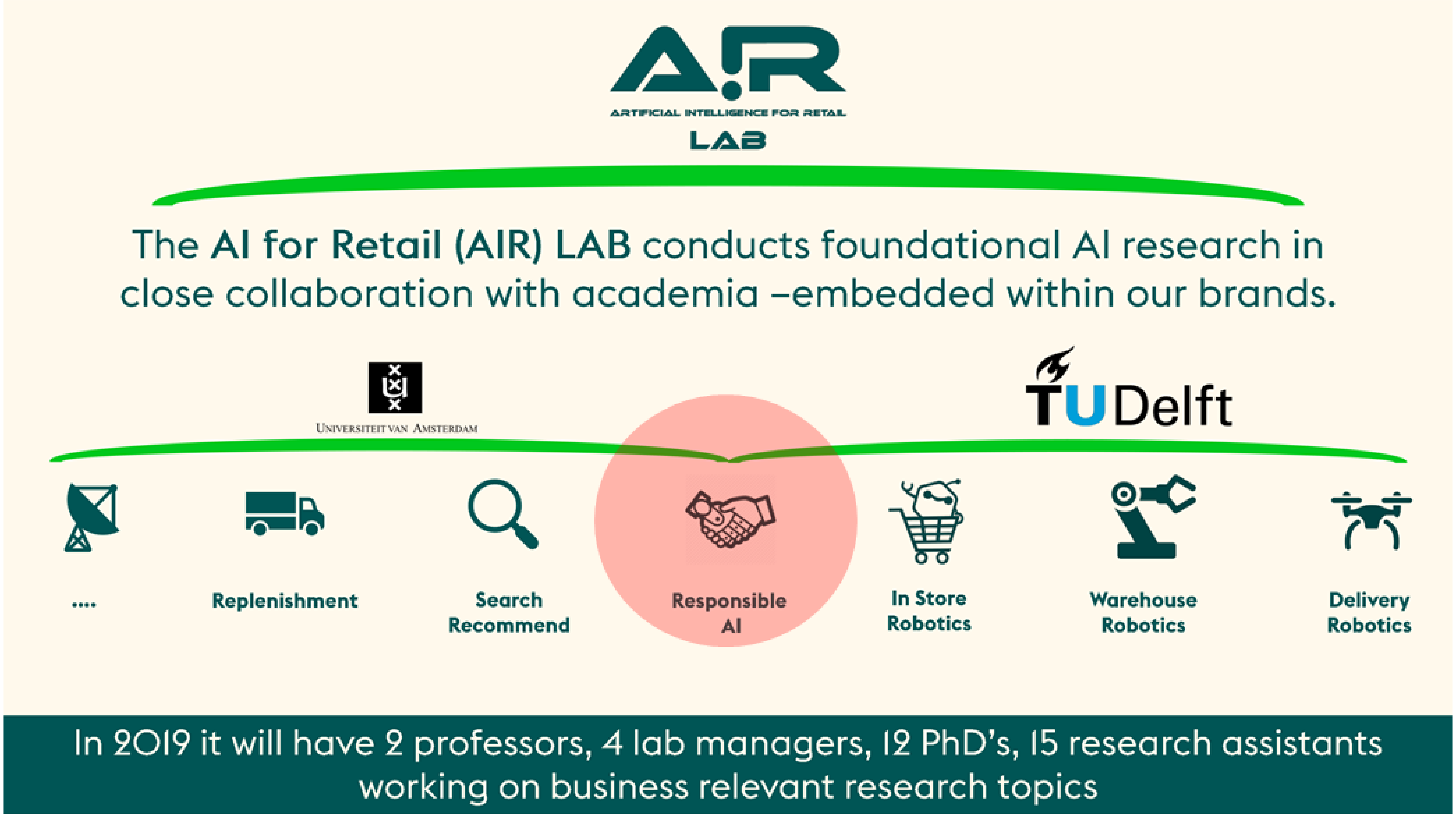
- We can find multiple contrastive explanations for the same prediction: how do we choose?
- It might not always be possible to find contrastive explanations
- Explanations can be very instable
- The explanations might not be actionable

Algorithmic aversion is a thing!

"We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake"

Dietvorst et al (2015)

Ongoing research



Ahold Delhaize

Hinda Haned
h.haned@uva.nl

What WE want

- Have some idea of what the system does
  - How do the personalized recommendations work?
- Able to challenge the system
  - What can I do to change my application's outcome?
- Able to understand when the system make
  - How do I trust the self-driving car?