SINTEF

Building trust in AI systems

With missclassification detectors

Building Trust in AI systems with missclassification detectors

- Themis 5.0:
  - "Co-create an innovative AI-driven and human-centric trustworthiness AI ecosystem"
- How we contribute to the project?
  - "SafetyCage" - Set of probabilistic approaches to detecting miss-classification

Who am I?

Milan
SINTEF Digital

- Research scientist at SINTEF Digital, Norway
- My research interests
  - Applications of **Machine Learning** to facilitate decision making

Themis 5.0

I. Project, partners, use case
II. Our contribution

Themis 5.0

- Mission statement
  - Co-create an innovative AI-driven and human-centric trustworthiness AI ecosystem
  - Strengthen the dialogue between AI users and practitioners
  - Converge to a better understanding of model accuracy, robustness, trustworthiness and fairness.
- 16 Partners in Europe
- themis-trust.eu

Themis 5.0 - Use-cases

- Related to Critical Application and Industry Sectors
- Can we help partners to gain trust in their models?

USE CASE 1

Healthcare: AI Powered Personalised Risk Assessment

MEDICAL UNIVERSITY OF PLOVDIV

Human genetic data can provide insights into the risks of diseases. AI has appeared to be the most effective tool for analysing complex biological data sets. However, existing AI systems lack explainability and transparency on how their results are generated which raises questions over the trustworthiness of their assessments. Professionals having trust in AI is fundamental in healthcare systems as decisions can have significant impacts.

This use-case focuses on the enhancement of datasets that are used for training of AI

USE CASE 2

Managing Disruptions and Dynamic Situations

PORT OF VALENCIA

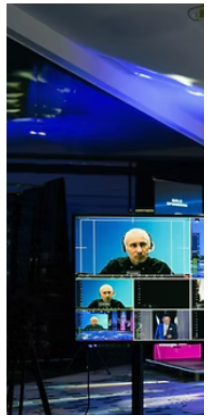
Ports have high import and export activity to address the demands of society and the economy. To ensure port infrastructures can support demand, timely maintenance of the docks is necessary to detect any risks and damages. If not monitored, serious risk can be posed to workers, stevedores and even visitors. AI can be used to predict disruptive events and damages that may pose a risk to those who use the port. However, many of these predictions are difficult to interpret.

USE CASE 3

Journalism: Preventing Disinformation

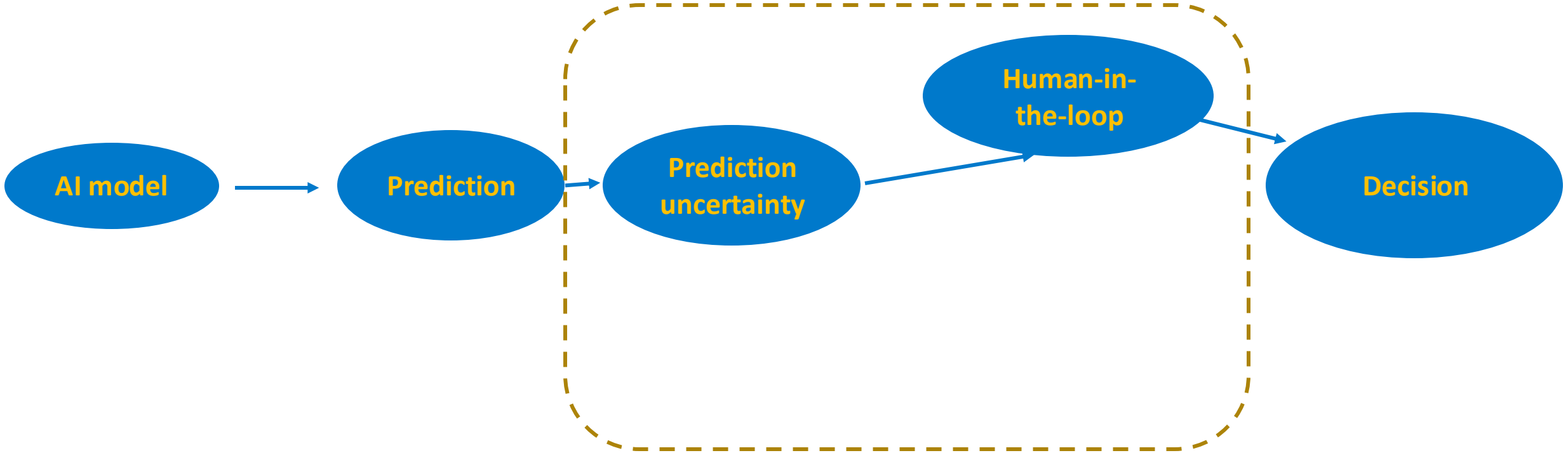
ATHENS-MACEDONIA NEWS AGENCY

Social media has cemented itself as a powerful new technology that makes it easy to manipulate and fabricate information, affecting democracy and trust in society. Social networks via computational propaganda techniques, such as 'trolling', can amplify fake news and disinformation. Journalists risk being manipulated by actors who frequently intimidate and discredit them, and media institutions. Increasingly, AI-based tools are being used to help avoid spreading disinformation and publishing of unchecked information.



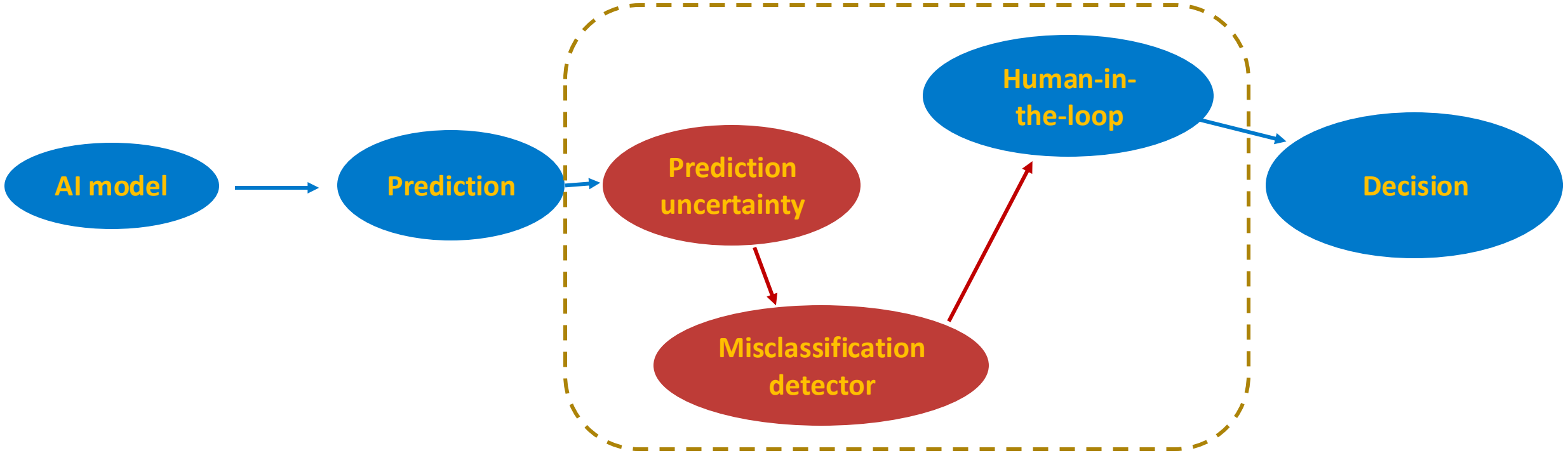
Themis 5.0 - Methodology

Themis 5.0 - Methodology



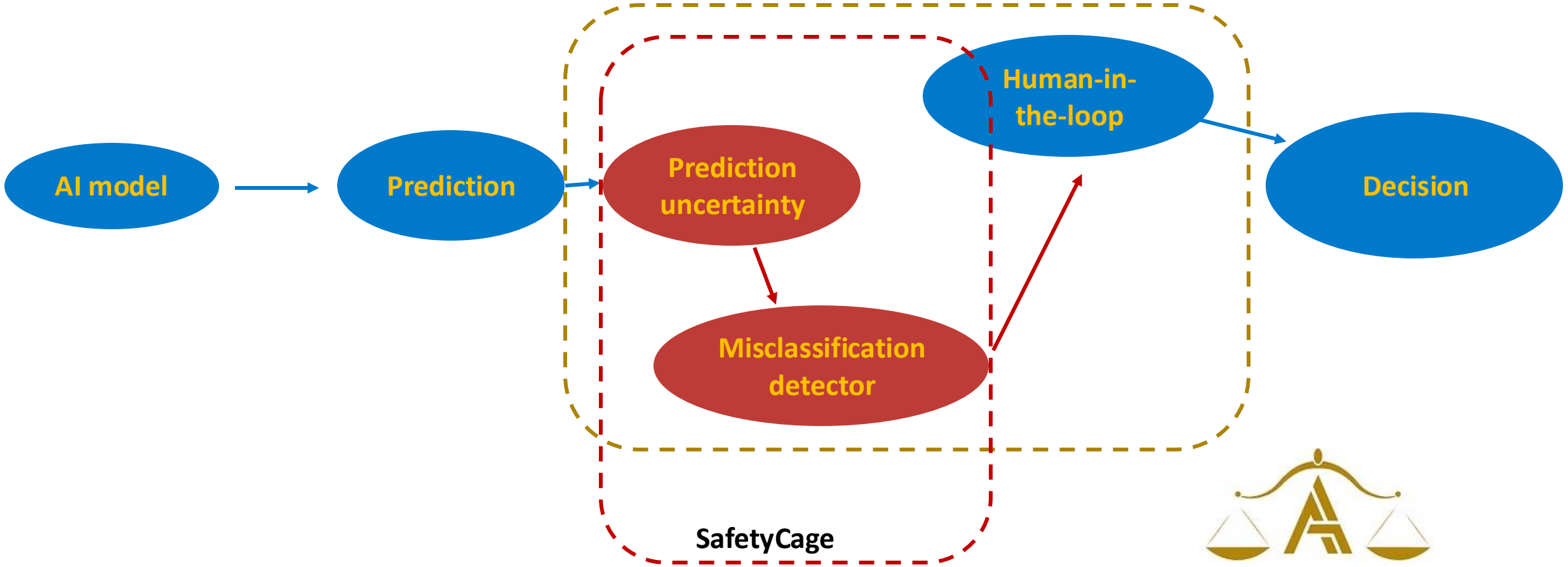
THEMIS 5.0

Themis 5.0 - Methodology



THEMIS 5.0

Themis 5.0 - SafetyCage



THEMIS 5.0

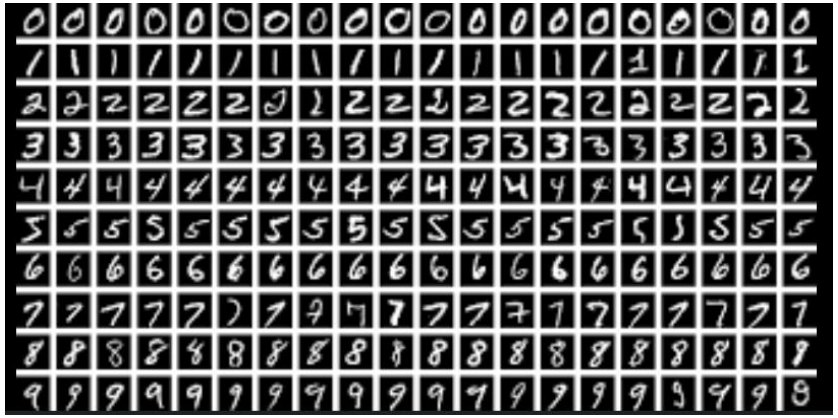
SINTEF
75 years

SafetyCage

I. A statistical framework for miss-classification detection

SafetyCage - Background

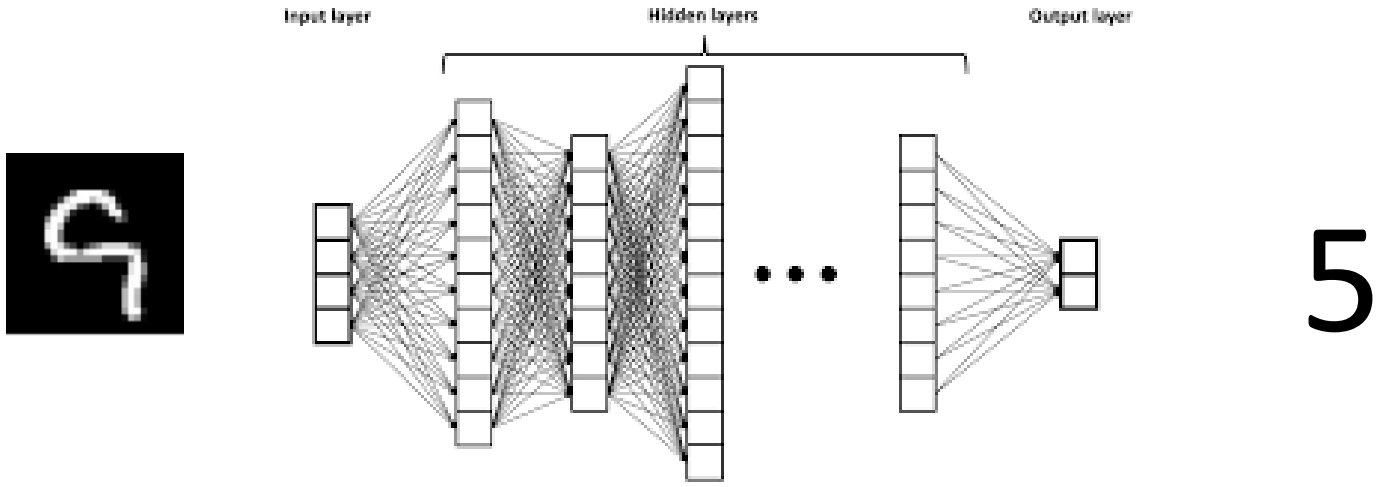
- Classification problems
  - Input – A set of features
  - Output – A class
- Examples:



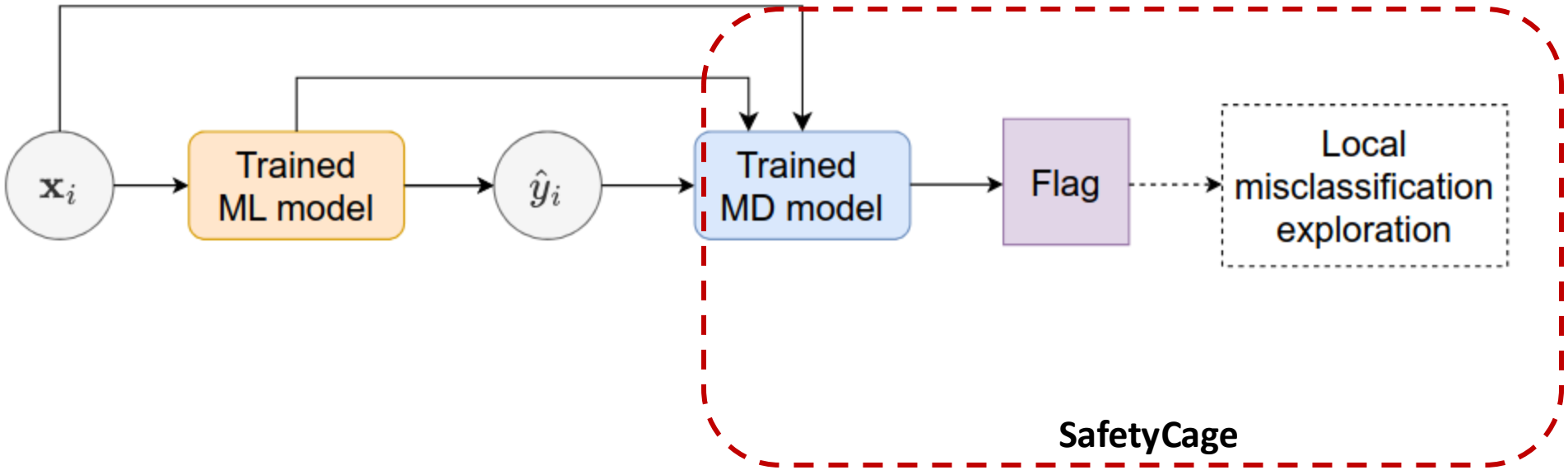
Data point	Class (label)
Image: hand drawn digits	Digit: 0, 1, 2, ...
Image	Chihuahua, Blueberry Muffin
Patient data	Cancer / No cancer
...	...

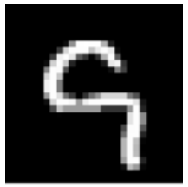
SafetyCage - Background

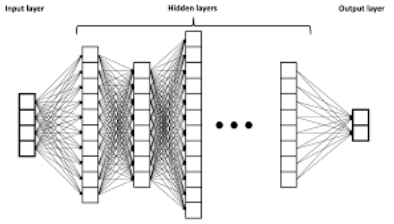
- Given a **trained neural** network predicting a class
- Can we train a system (Missclassification Detector - **MD**) capable of:
  - Predicting whether a prediction is incorrect?
  - Measuring a degree of confidence?



SafetyCage - Overview







5

Applying SafetyCage to a model predicting risk of pancreatic cancer

Patient health record (Age, gender, (co)morbidity, symptoms etc.

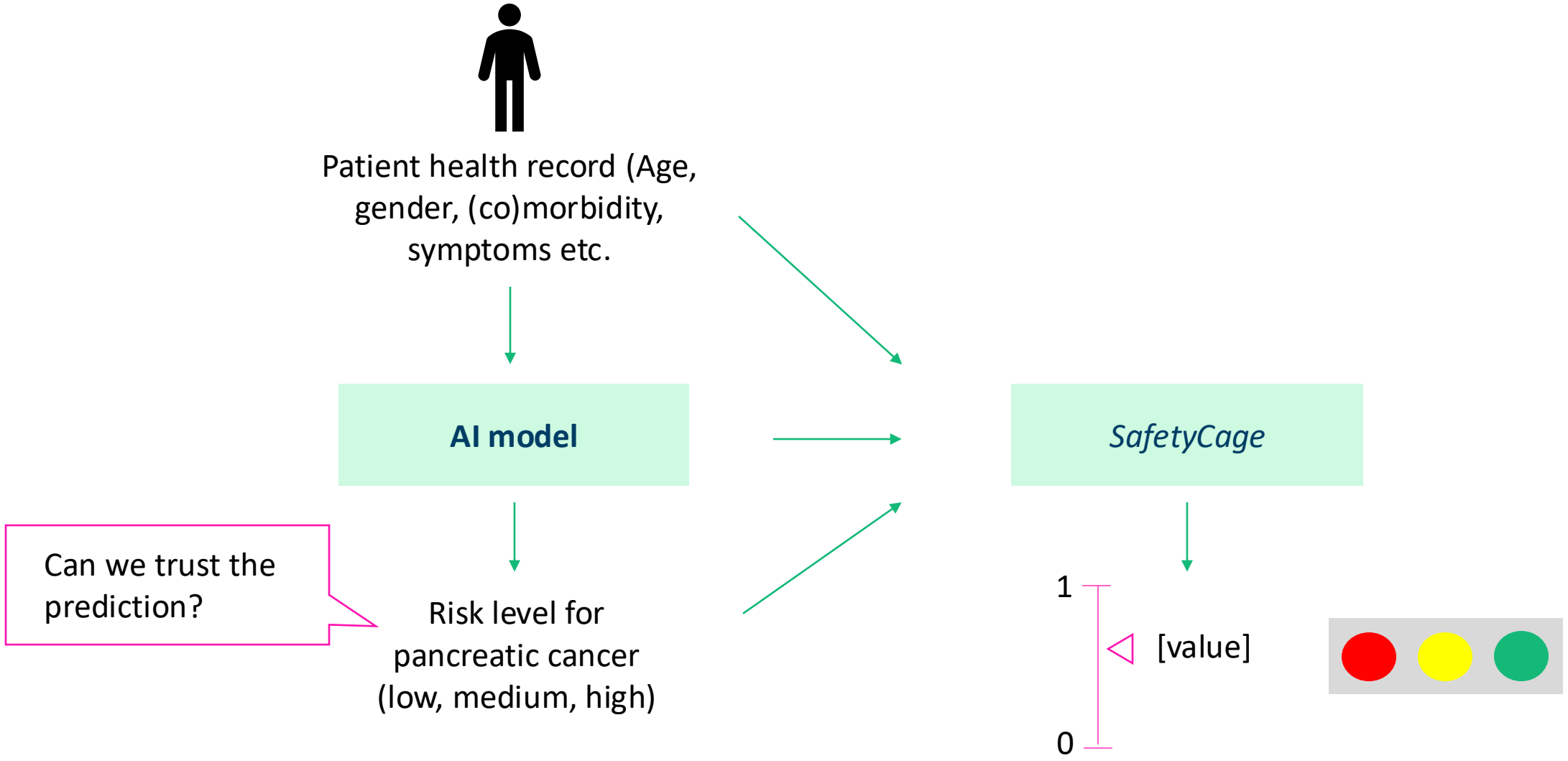
AI model

Risk level for pancreatic cancer (low, medium, high)

General performance on *historical data*
Classification accuracy = ***81%*** *of predictions correct*
Precision low risk factor (proportion of predictions for risk factor 0 that is correct) = ***90 %***
Precision medium risk factor (proportion of predictions for risk factor 1 that is correct) = ***73 %***
Precision high risk factor *(proportion of predictions for risk factor 2 that is correct) =* **78 %**

Note:
The table shows what we **can expect in general**, but says nothing about what we can **expect for one input sample**

Applying SafetyCage to a model predicting risk of pancreatic cancer



Applying SafetyCage to a model predicting risk of pancreatic cancer

Patient A health record (Age, gender, (co)morbidity, symptoms etc.

AI model

Risk level for pancreatic cancer: **High risk**

Can we trust the prediction?

SafetyCage

1

0

0,35

Prediction cannot be trusted

Applying SafetyCage to a model predicting risk of pancreatic cancer

Patient B health record (Age, gender, (co)morbidity, symptoms etc.

AI model

Can we trust the prediction?

Risk level for pancreatic cancer: **Low risk**

SafetyCage

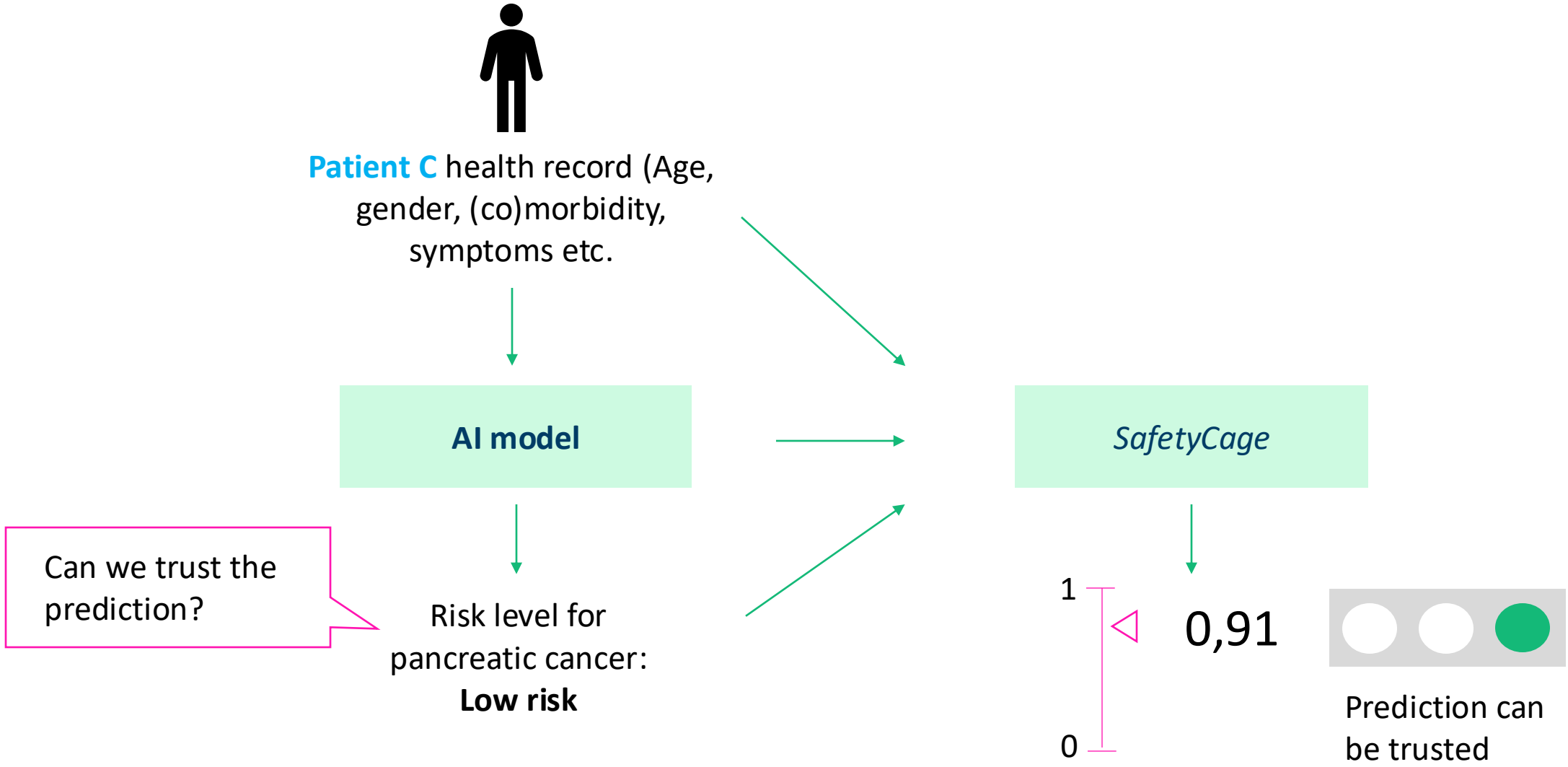
1

0

0,67

Prediction should be interpreted with care

Applying SafetyCage to a model predicting risk of pancreatic cancer



Applying SafetyCage to a model predicting risk of pancreatic cancer

Patient D health record (Age, gender, (co)morbidity, symptoms etc.

AI model

Risk level for pancreatic cancer: **High risk**

Can we trust the prediction?

SafetyCage

1

Threshold: 0.62

0

0,81

Prediction can be trusted

Detailed explanation

According to the misclassification detection method *MSP*, **the class prediction is correct.**

This method looks at the maximum softmax probability value, which is available in the output from the ML model, as a measure of the uncertainty in a particular class prediction.

Whenever the softmax probability is less than **alpha=0.62**, the safetycage method considers the class prediction as wrong.

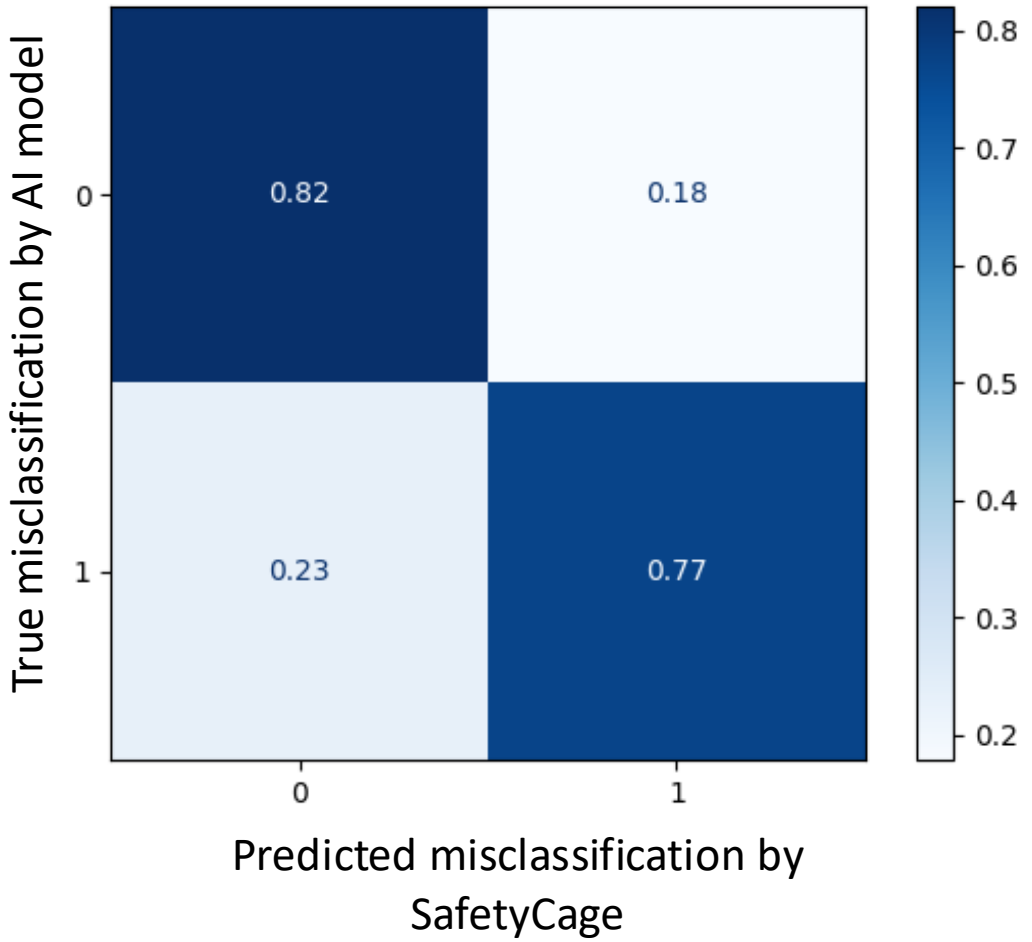
The maximum softmax probability value is in this case equal to **0.81.**

Can SafetyCage be trusted?

- SafetyCage is built from statistical arguments
  - Not always right!
  - 77% of the predicted miss classifications were indeed missclassified

- Perfect missclassification detector:

1	0
0	1



Further work

- Improving the misclassification detectors
  - Developing other/stronger statistical arguments to assess model confidence
  - Diversifying SafetyCage portfolio
- Demonstrating and deploying SafetyCage in pilot seetings

SINTEF

75 years

1950 – 2025
Technology for a better society

sintef.no/75