The Challenge of Computing
Responsible AI

Professor Thomas B. Moeslund

Aalborg University, Denmark

PIONEER CENTRE FOR
ARTIFICIAL INTELLIGENCE

AALBORG UNIVERSITY
DENMARK

VISUAL ANALYSIS &
PERCEPTION LAB

Agenda

- Who am I?
- Why are we talking about Responsible AI?
- How do we compute Responsible AI?
- The end-game of AI
- Q&A

Who am I?

- Head of Section for Media Technology (40 researchers)
- Head of AI for the People Center (150 researchers)
- Head of Visual Analysis and Perception lab (35 researchers)

- Pioneer Center for AI (co-lead): 50M EUR
- Center for AI in Society (co-lead): 7M EUR
- Responsible AI for Value Creation (lead): 3M EUR

AALBORG UNIVERSITY
DENMARK

Visual Analysis and Perception (VAP) Lab

Started in 2011 as ‘Visual Analysis of People’ Lab

Research field: Computer Vision & AI

Research interest:

Building intelligent systems that make sense out of (visual) data

AALBORG UNIVERSITY

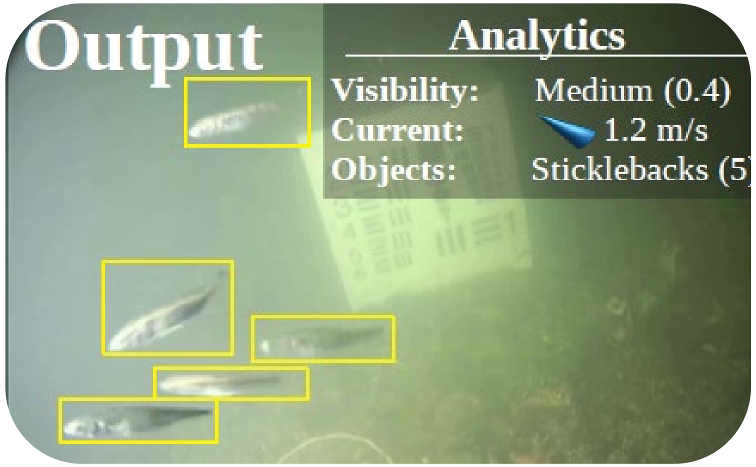
DENMARK

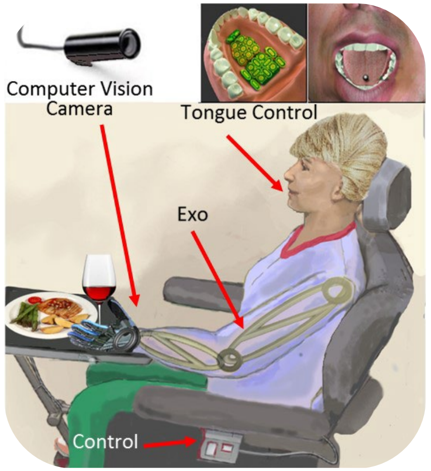
The people of VAP

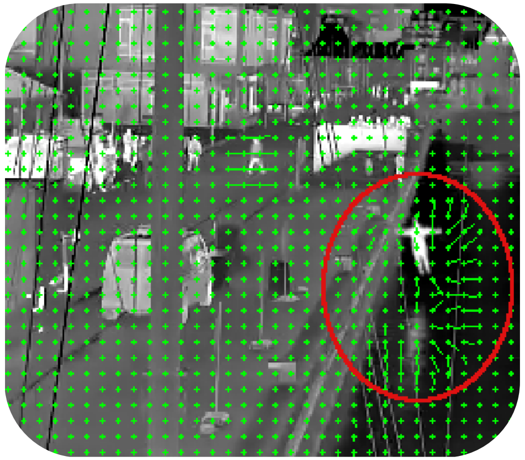
Professor	3
Associate Professor	3
Assistant Professor	4
Postdoc	8
PhD	14
Research assistants	3
Total	**35**

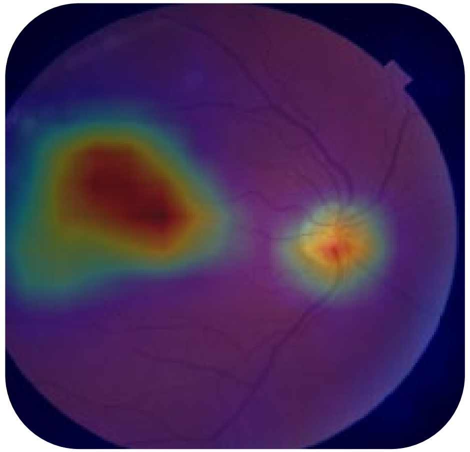
Research

- **Drivers:**
  - Curiosity
  - Real-world problems
  - Different sensors
- **Domains:**
  - Surveillance
  - Traffic
  - Robotics
  - Sports
  - Healthcare
  - Machine vision
  - Underwater
  - Responsible AI

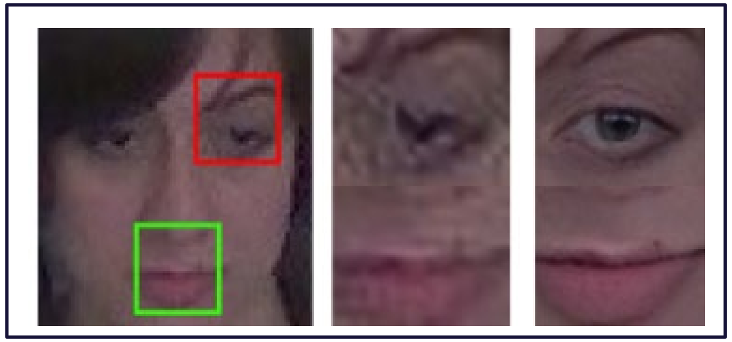




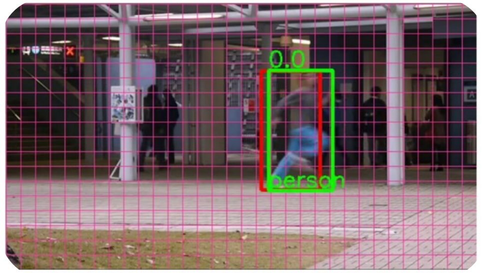




Fish & other animals

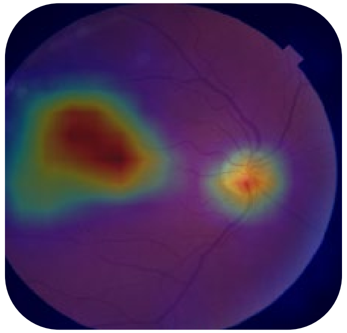


Surveillance & sports



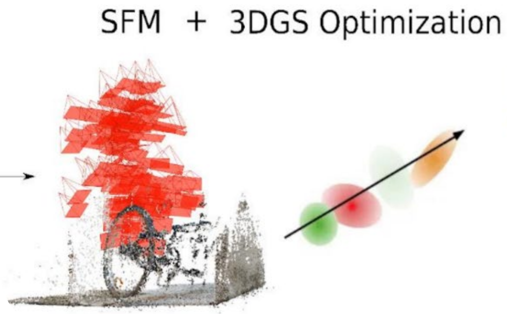
SoccerNet
2025 CHALLENGES
5TH EDITION
BALL ACTION SPOTTING
GAME STATE RECONSTRUCTION
MULTI-VIEW FOUL RECOGNITION
MONOCULAR DEPTH ESTIMATION
DEADLINE
24/04/2025
WWW.SOCCER-NET.ORG

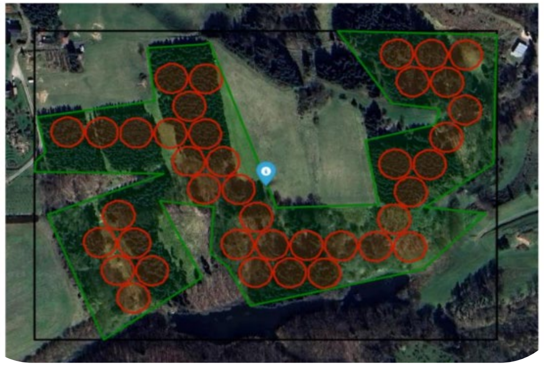
VAP LAB 2025



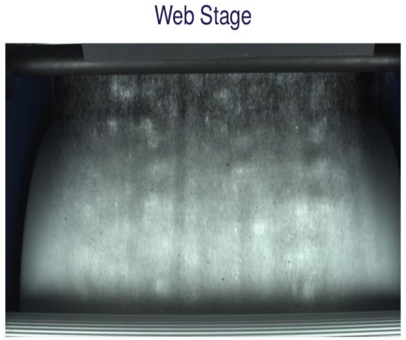
Responsible AI

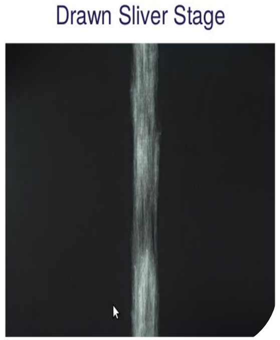
3D Vision





Quality inspection





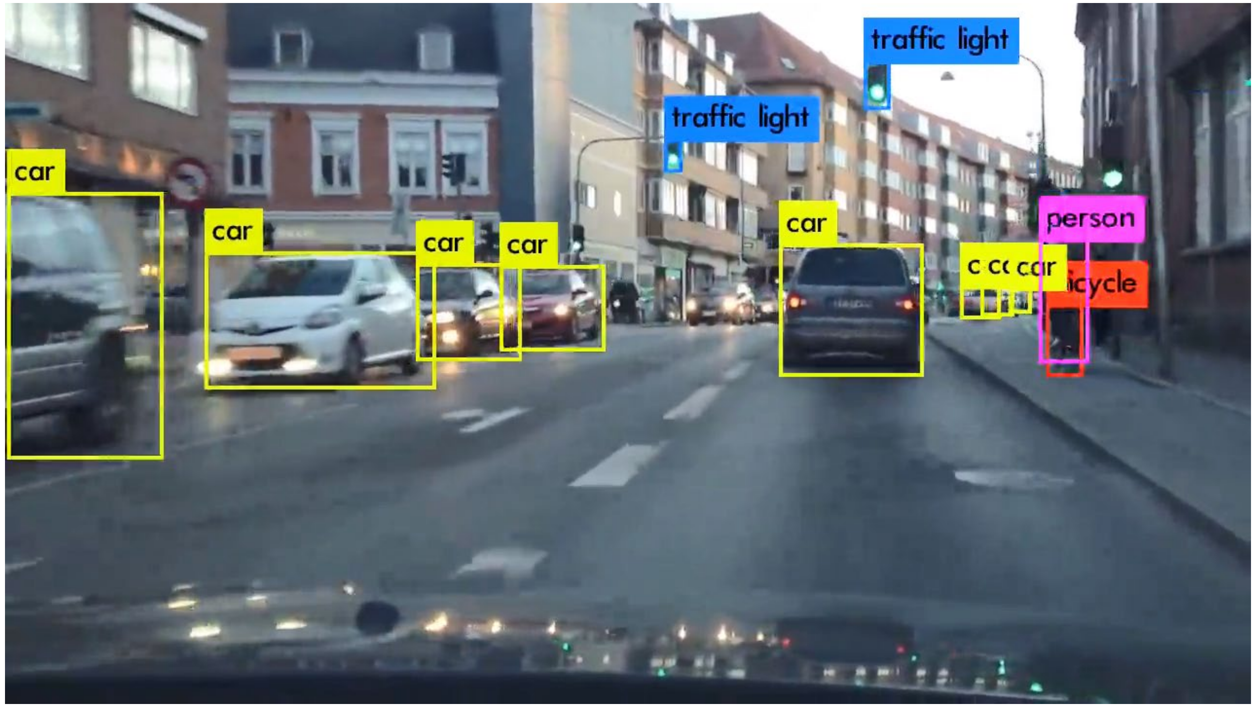
Agenda

- Who am I?
- Why are we talking about Responsible AI?
- How do we compute Responsible AI?
- The end-game of AI
- Q&A

AALBORG UNIVERSITY
DENMARK

Deep learning

It works ☺

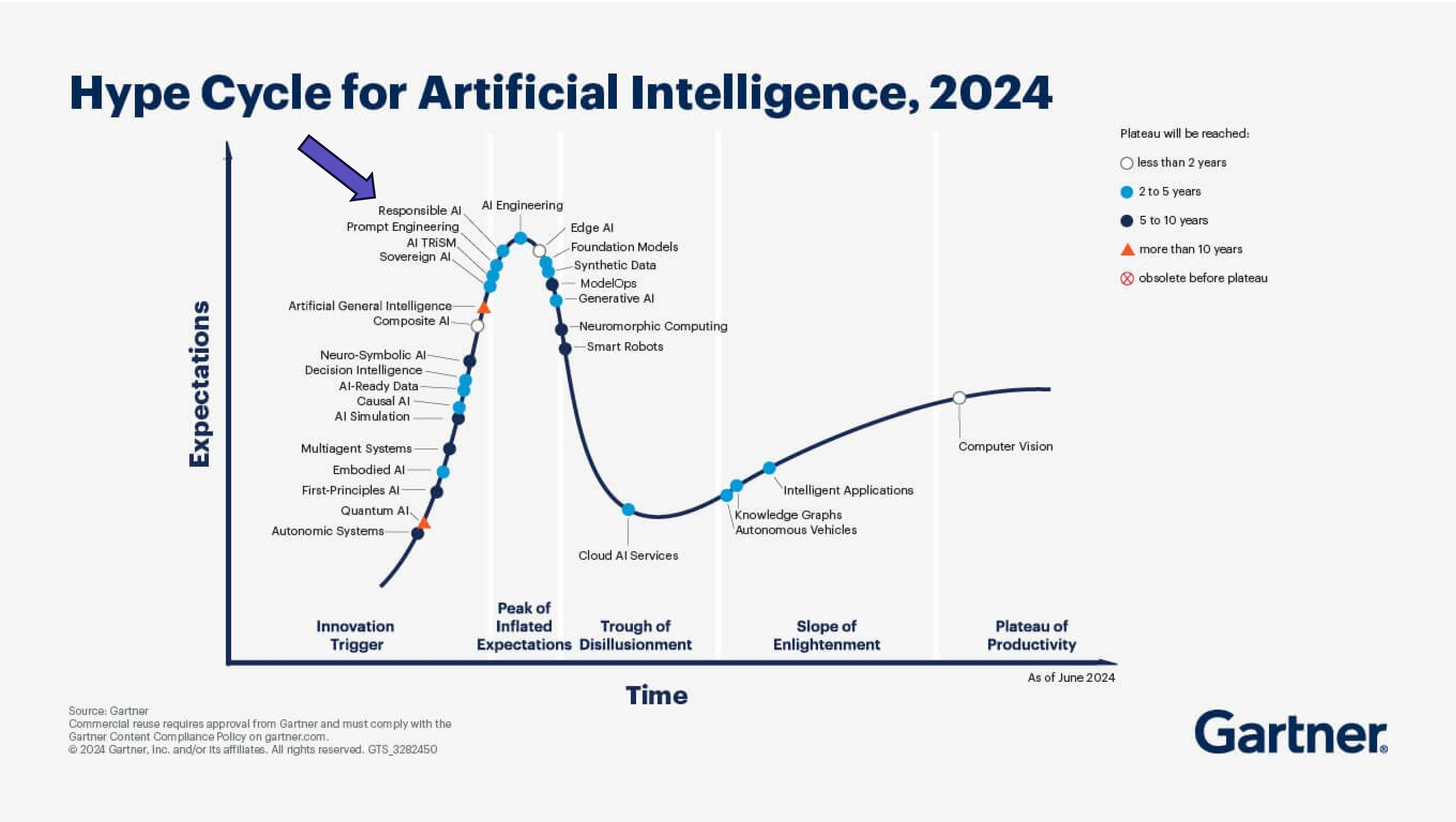


- The tech is working ☺
- But…

AALBORG UNIVERSITY
DENMARK

How do we deal with this?

Responsible AI



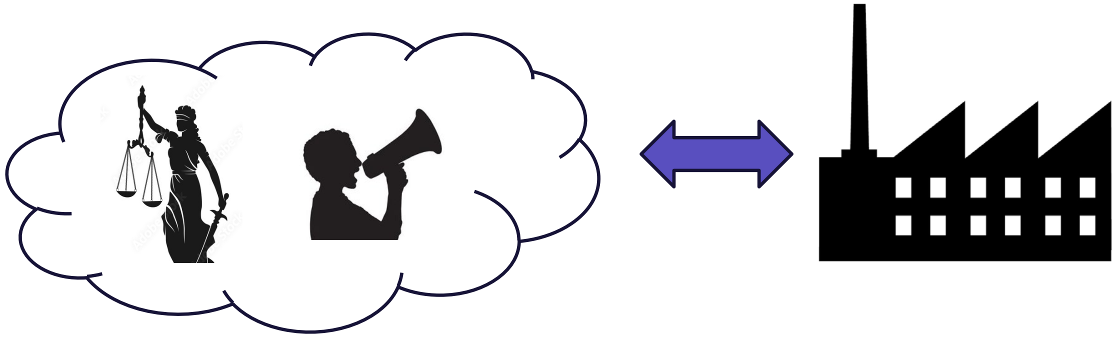
Hype Cycle for Artificial Intelligence, 2024

As of June 2024

Source: Gartner
Commercial reuse requires approval from Gartner and must comply with the Gartner Content Compliance Policy on gartner.com.
© 2024 Gartner, Inc. and/or its affiliates. All rights reserved. GTS_3282450

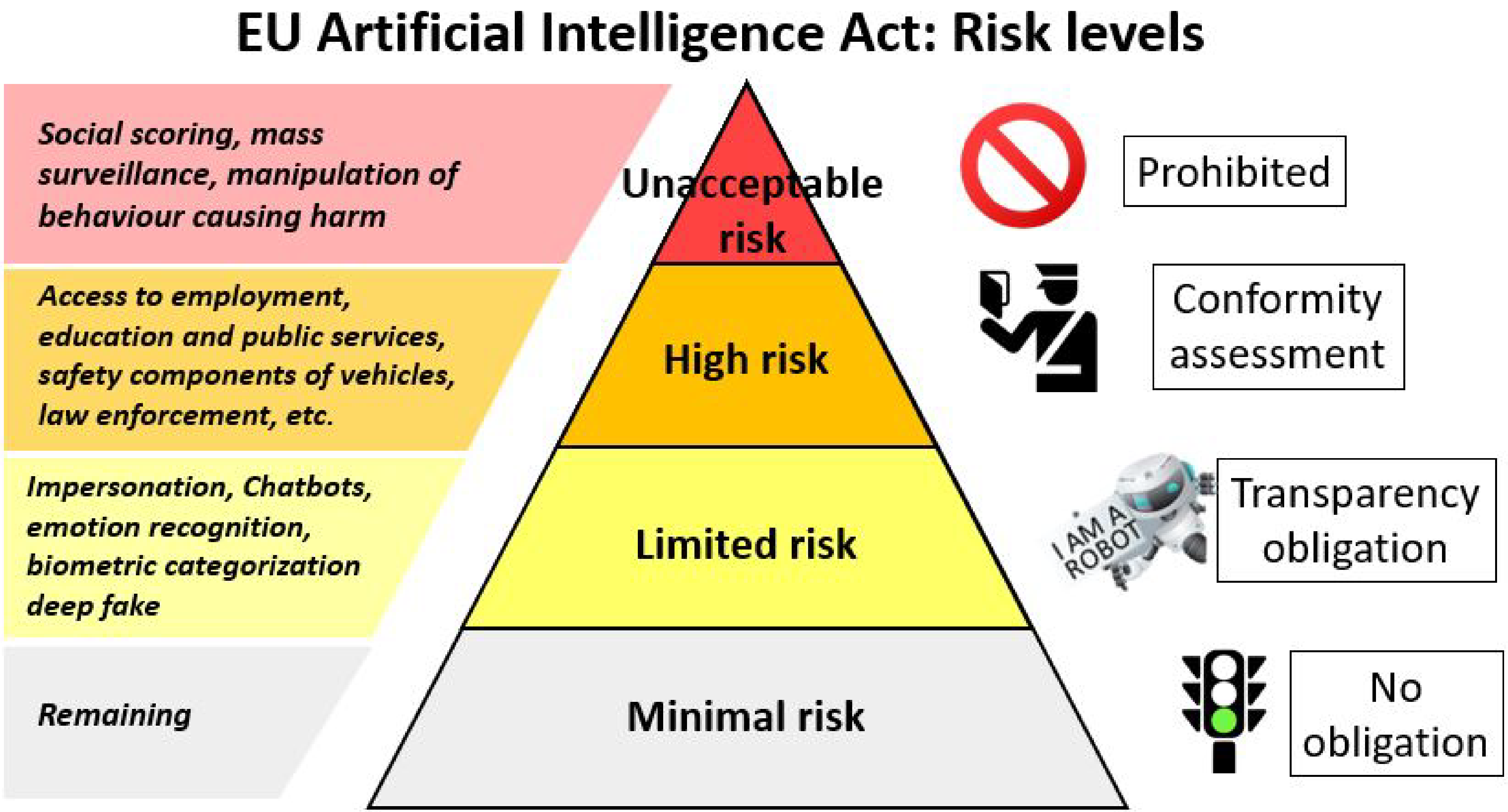
Different attitudes towards Responsible AI

- Innovation vs regulation
  - Fast vs slow
- Fear of missing out
- Start-up mindset: Move fast and break stuff until it works….
- Trust & democracies are eroded
- Hold back until we know what we are doing



EU's approach: Responsible AI via Regulation

- Responsible TECH: Nuclear weapon limited spread. Cloning



EU Artificial Intelligence Act: Risk levels

Examples	Risk level	Obligation
Social scoring, mass surveillance, manipulation of behaviour causing harm	Unacceptable risk	Prohibited
Access to employment, education and public services, safety components of vehicles, law enforcement, etc.	High risk	Conformity assessment
Impersonation, Chatbots, emotion recognition, biometric categorization deep fake	Limited risk	Transparency obligation
Remaining	Minimal risk	No obligation

Agenda

- Who am I?
- Why are we talking about Responsible AI?
- How do we compute Responsible AI?
- The end-game of AI
- Q&A

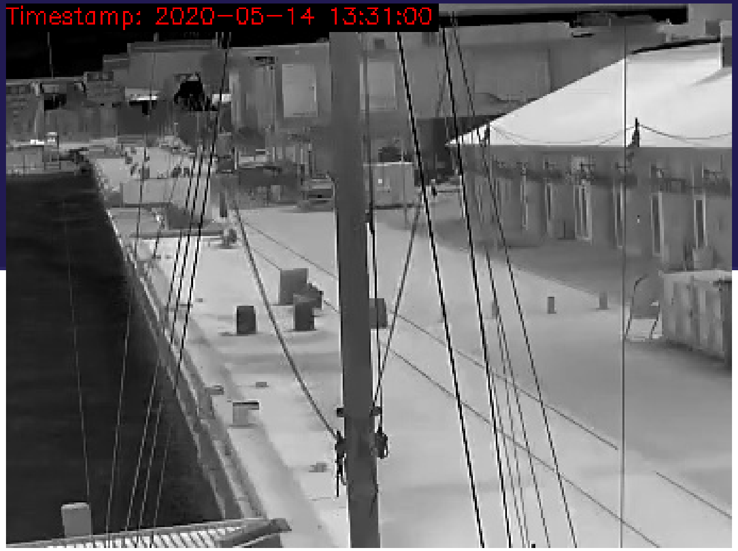
How to compute Responsible AI (RAI)

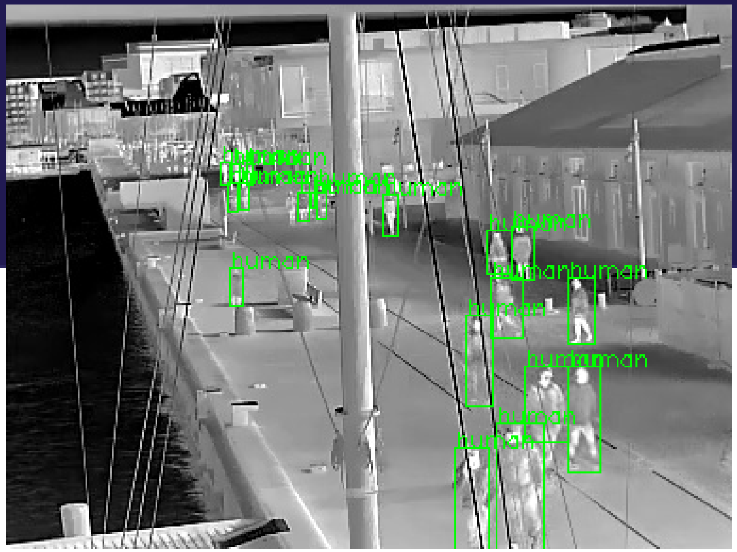
	Data	Model	Result
Terms:	‘The right to be forgotten’	‘Transparency’	‘Robustness’
Tech:	Machine Unlearning	XAI	Data drift Out of distribution

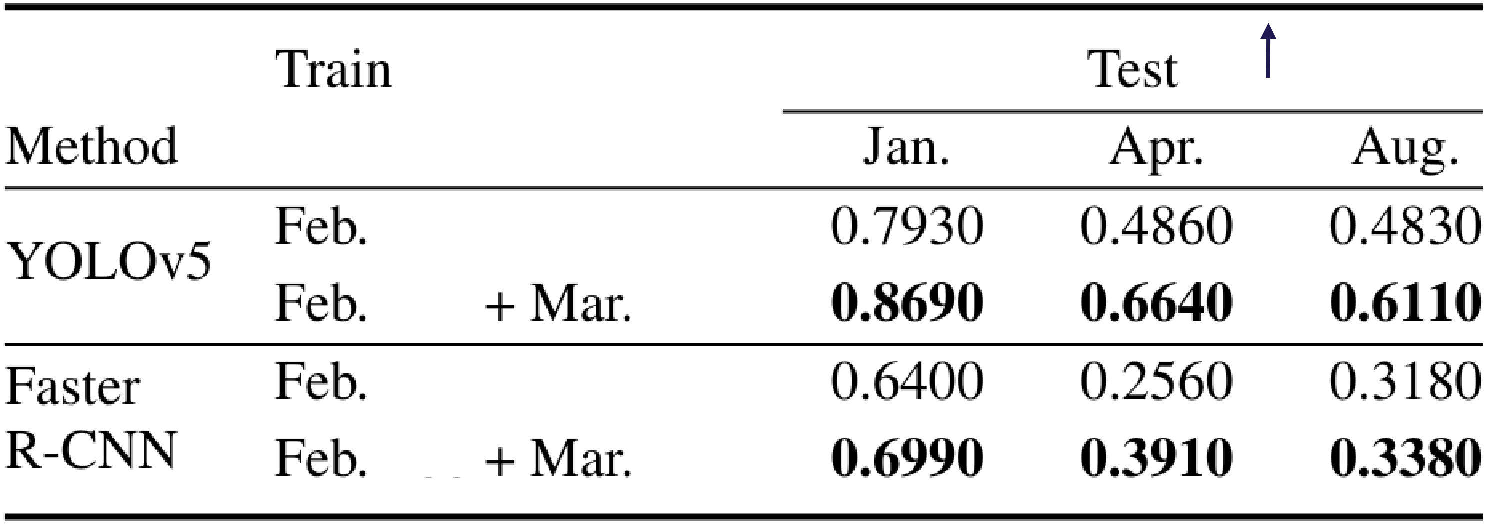
How to compute RAI

Robustness - Data drift

- 8 months
- Four classes
- 6.8M annotations







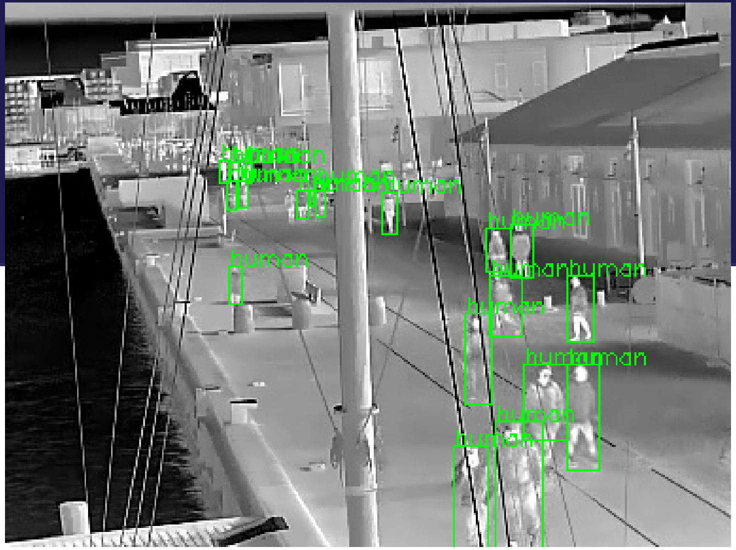
	Train		Test		
Method			Jan.	Apr.	Aug.
YOLOv5	Feb.		0.7930	0.4860	0.4830
	Feb.	+ Mar.	**0.8690**	**0.6640**	**0.6110**
Faster R-CNN	Feb.		0.6400	0.2560	0.3180
	Feb.	+ Mar.	**0.6990**	**0.3910**	**0.3380**

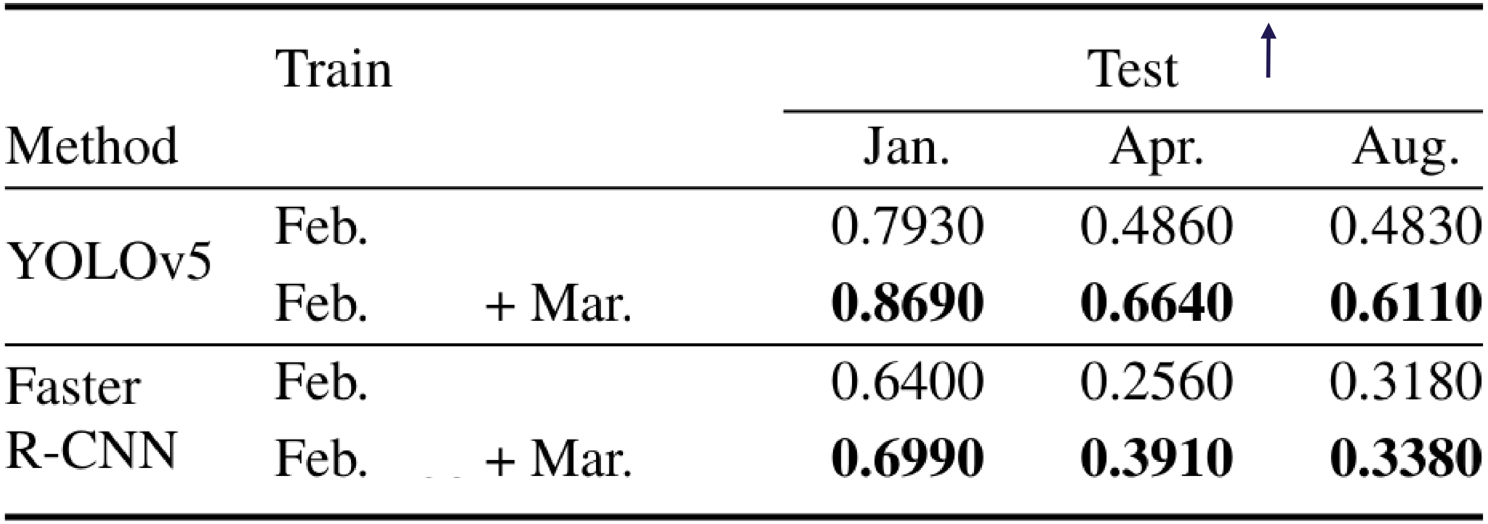
[Ivan Nikolov et al. Seasons in Drift. NeurIPS’21]

How to compute RAI

Robustness - Data drift

- 8 months
- Four classes
- 6.8M annotations



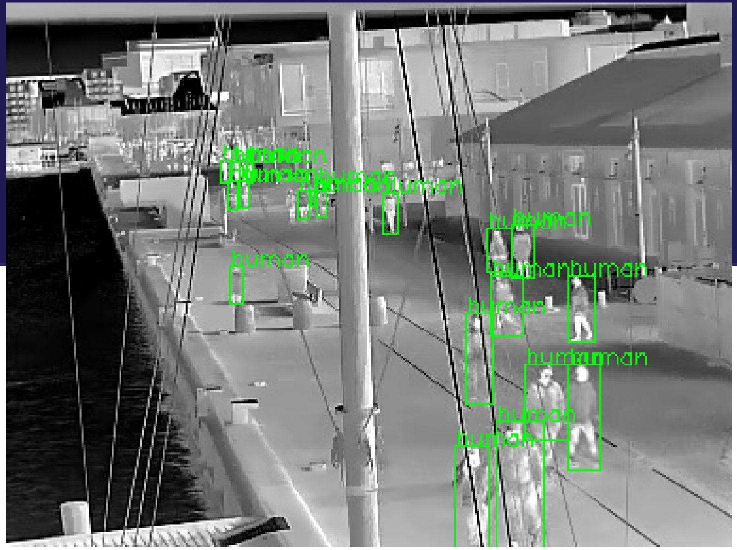


	Train		Test		
Method			Jan.	Apr.	Aug.
YOLOv5	Feb.		0.7930	0.4860	0.4830
	Feb.	+ Mar.	**0.8690**	**0.6640**	**0.6110**
Faster R-CNN	Feb.		0.6400	0.2560	0.3180
	Feb.	+ Mar.	**0.6990**	**0.3910**	**0.3380**

[Ivan Nikolov et al. Seasons in Drift. NeurIPS’21]

How to compute RAI

Robustness - Data drift



Takeaways

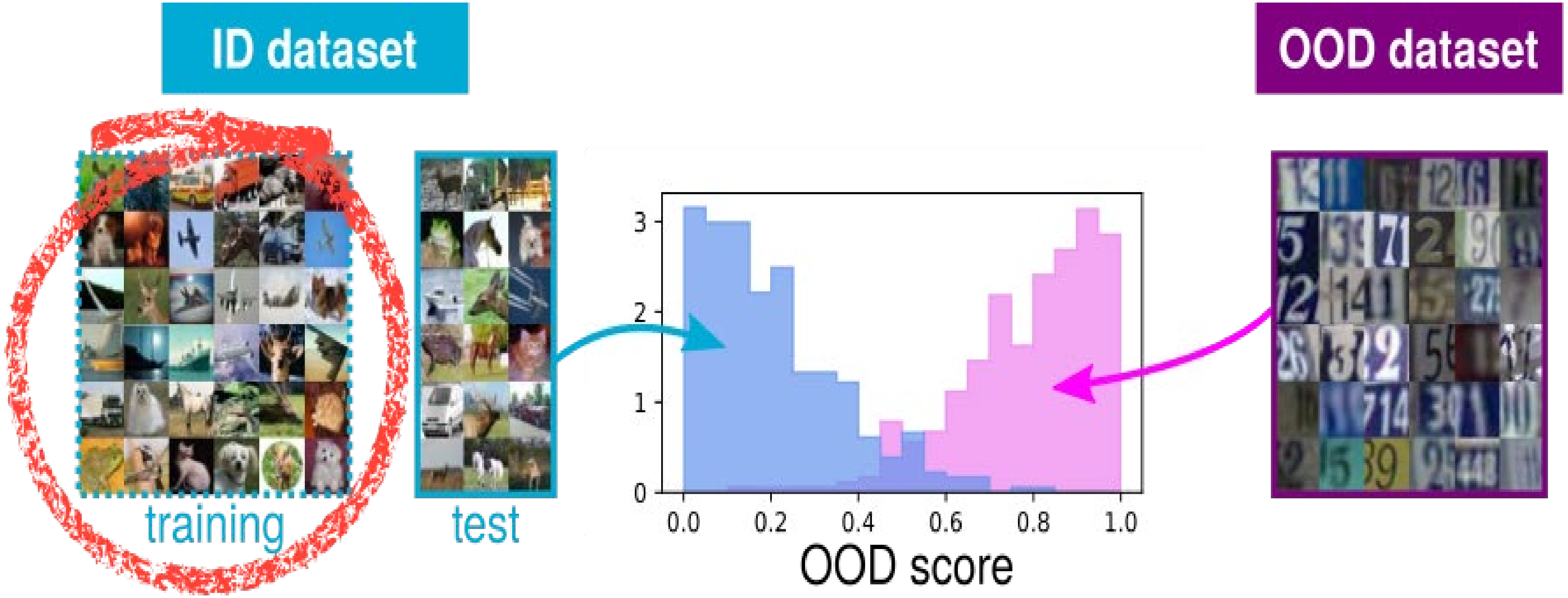
- We don’t have a good method for detecting drift automatically
- Not clear how to mitigate
- Drift metrics?
- Additional research needed

How to compute Responsible AI (RAI)

	Data	Model	Result
Terms:	‘The right to be forgotten’	‘Transparency’	‘Robustness’
Tech:	Machine Unlearning	XAI	Data drift Out of distribution

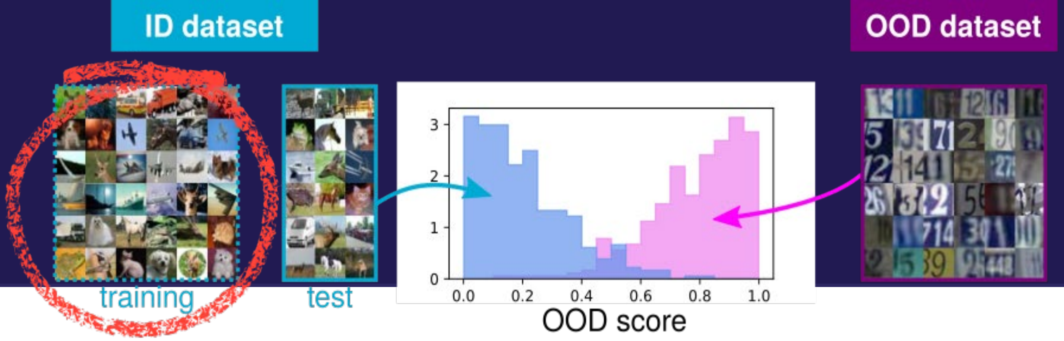
How to compute RAI

Robustness - Out of distribution

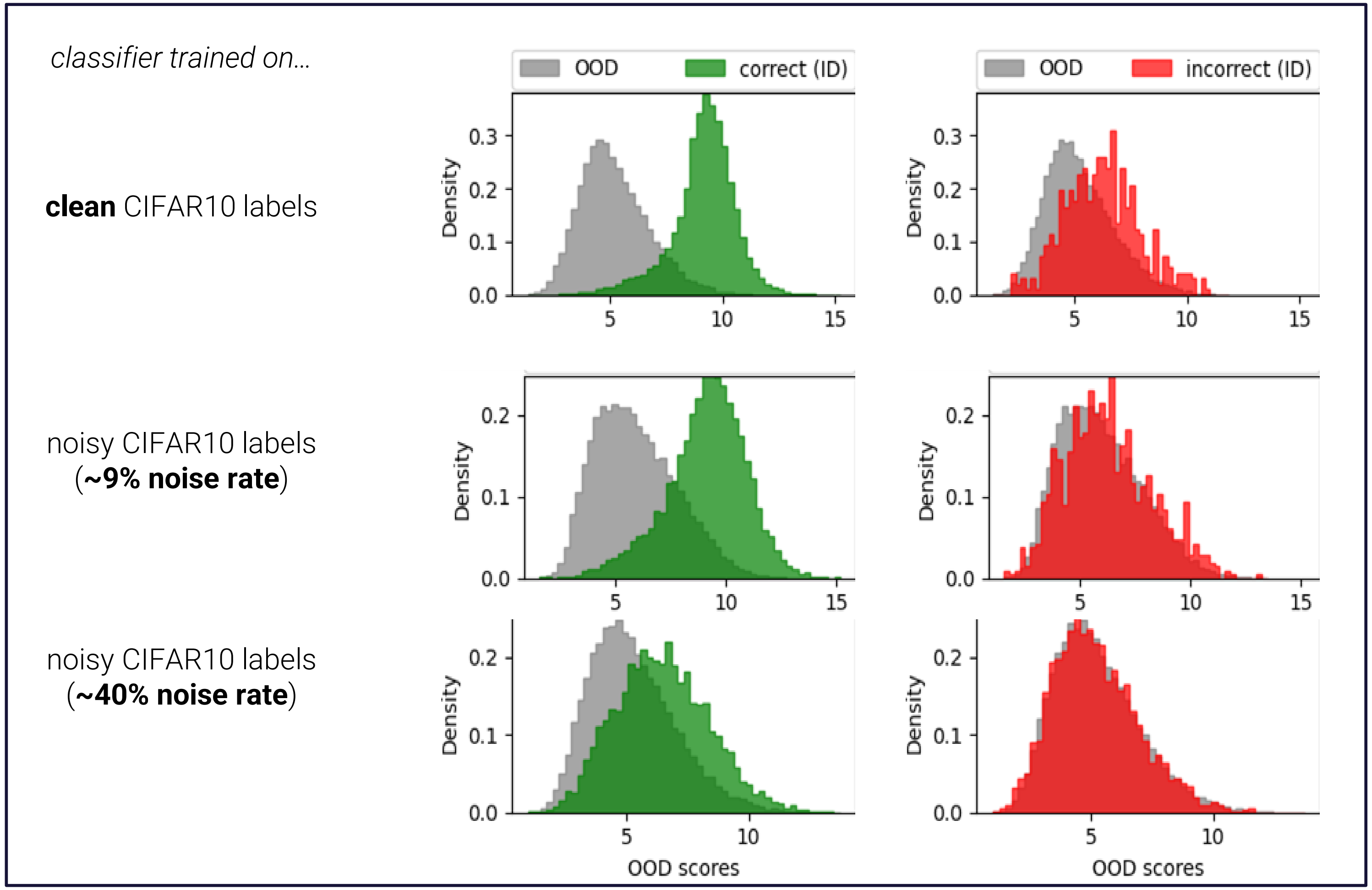


How to compute RAI

Robustness - Out of distribution



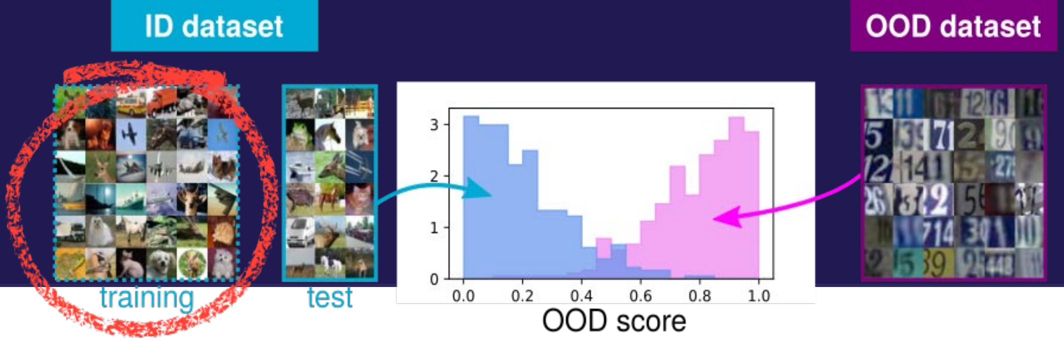
- 20 post-hoc OOD detectors
- 396 trained classifiers
- 7 OOD datasets



[Galadrielle Humblot-Renaux et al. A noisy elephant in the room. CVPR'24]

How to compute RAI

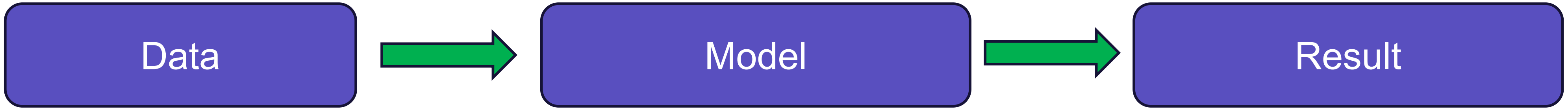
Robustness - Out of distribution



Takeaways

- We don’t have a good method for detecting OoD (in the face of label noise)
- Label noise is an underrated problem
- Metrics?
- Additional research needed

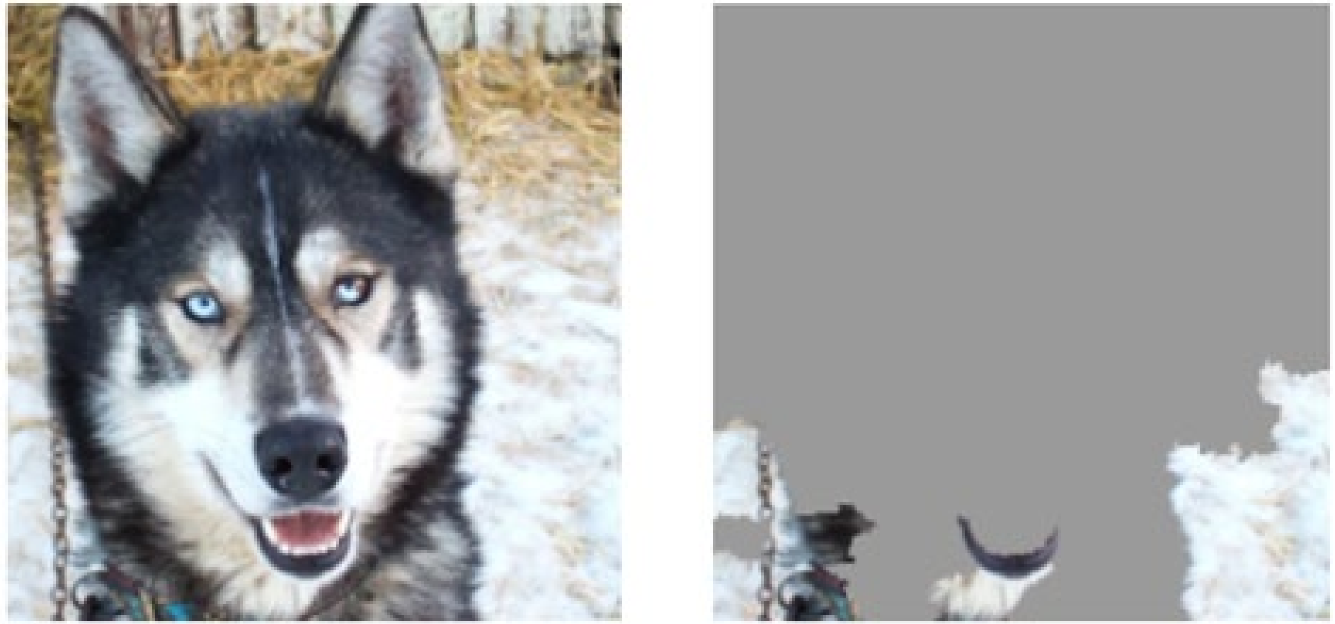
How to compute RAI



	Data	Model	Result
Terms:	‘The right to be forgotten’	‘Transparency’	‘Robustness’
Tech:	Machine Unlearning	XAI	Data drift Out of distribution

How to compute RAI

Transparency - XAI

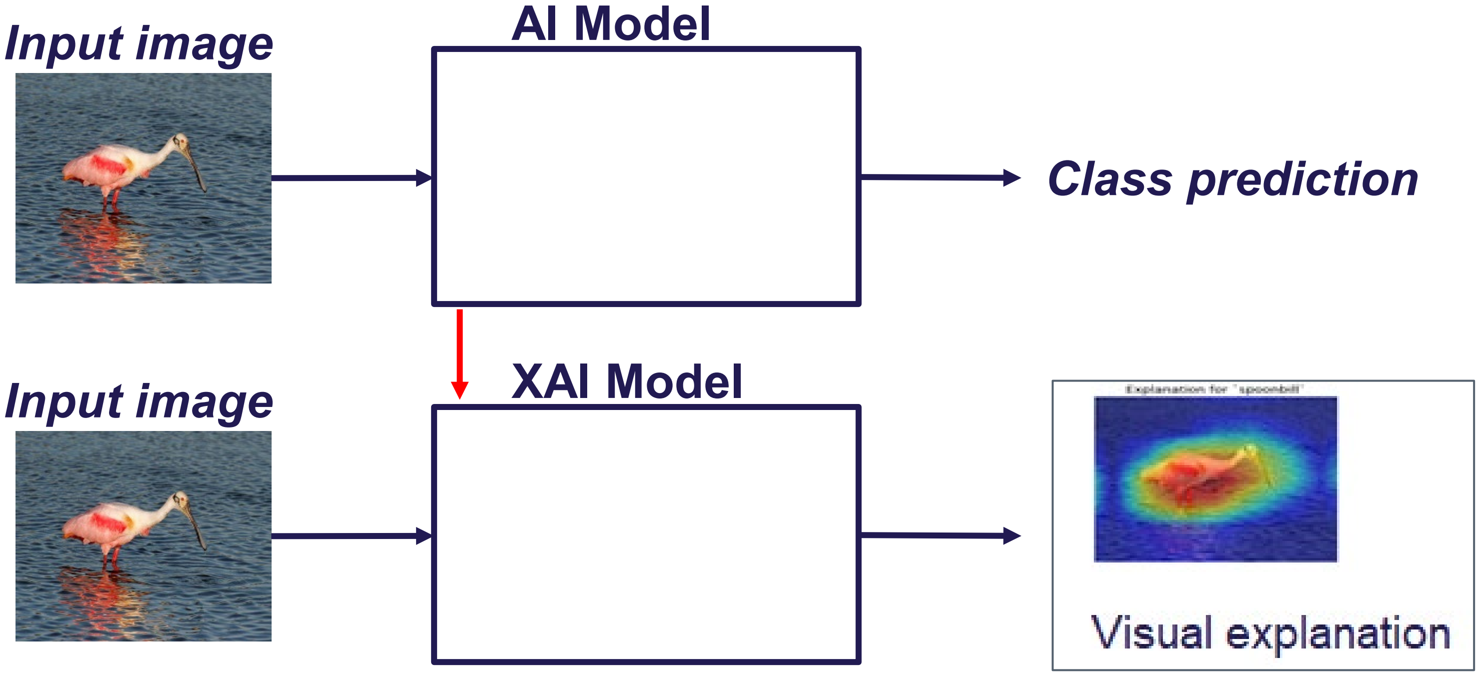


(a) Husky classified as wolf

(b) Explanation

How to compute RAI

Transparency - XAI



Input image

AI Model

Class prediction

Input image

XAI Model

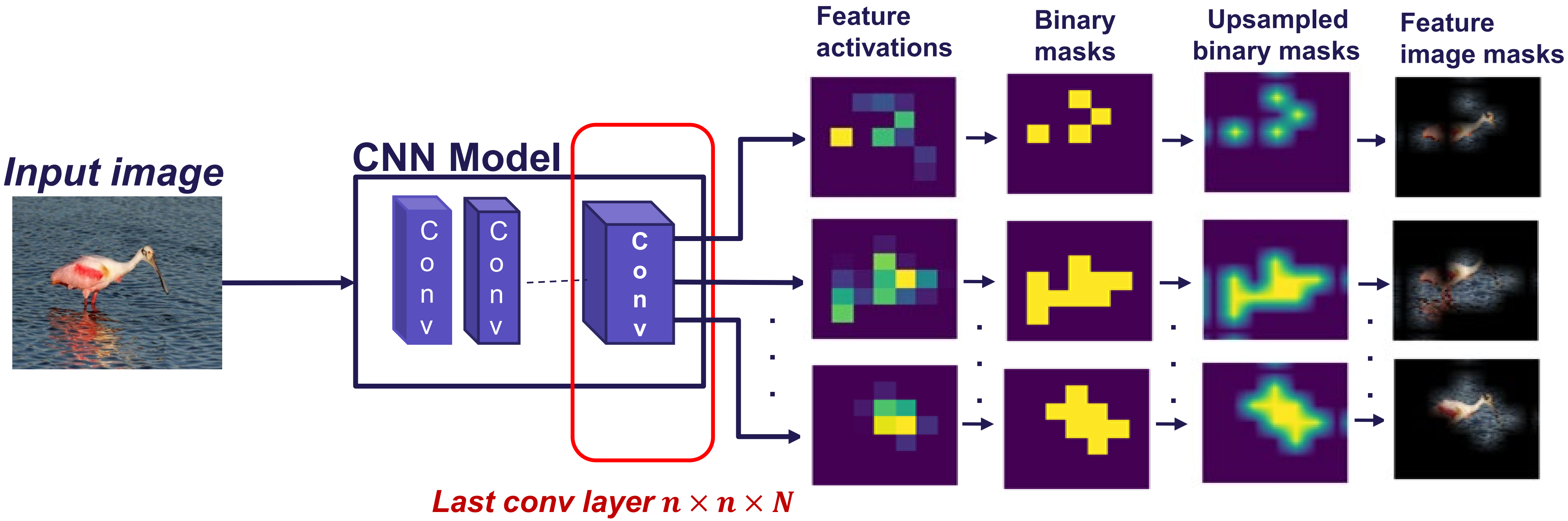
Visual explanation

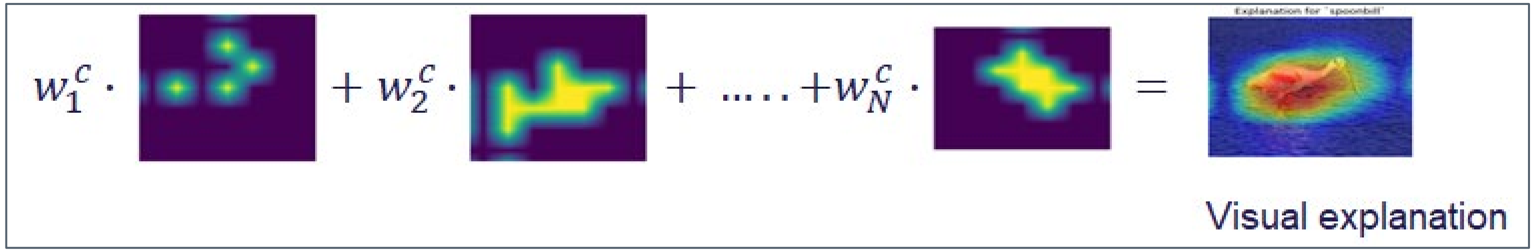
Method: Similarity Difference and Uniqueness (SIDU)

[Satya M. Muddamsettya et al. Visual Explanation of Black-Box Model. Pattern Recognition, 2022]

How to compute RAI

Transparency - XAI

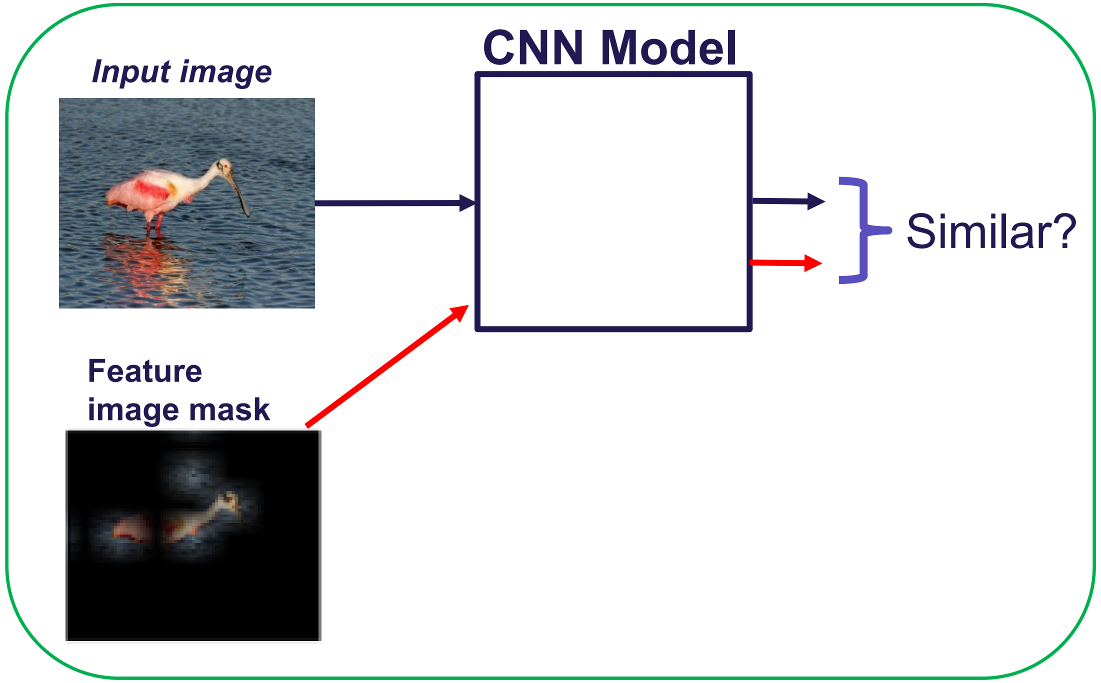


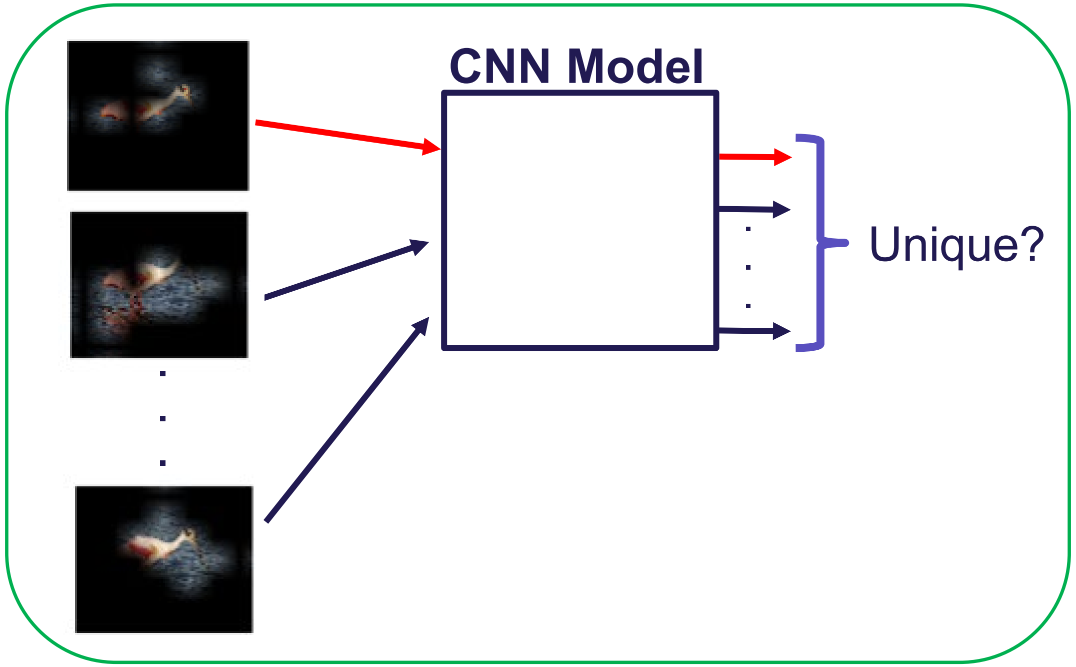


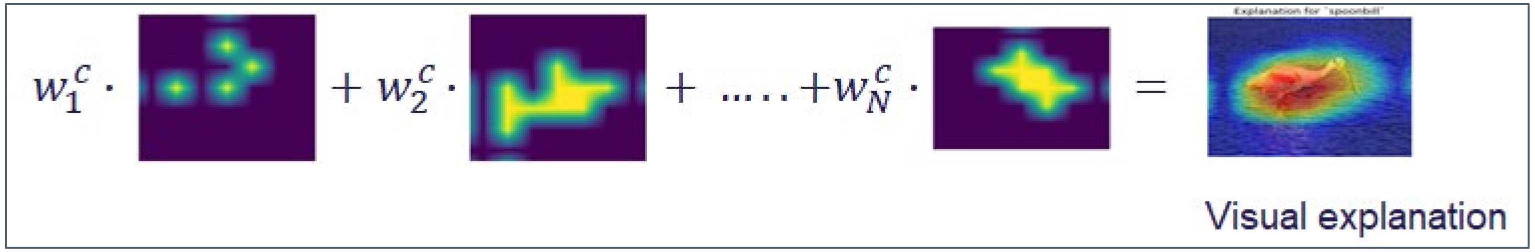
[Satya M. Muddamsettya et al. Visual Explanation of Black-Box Model. Pattern Recognition, 2022]

How to compute RAI

Transparency - XAI



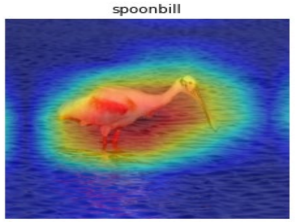




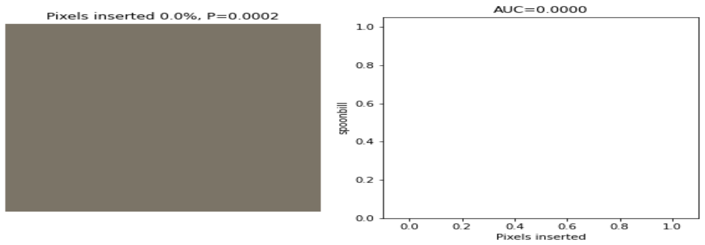
[Satya M. Muddamsettya et al. Visual Explanation of Black-Box Model. Pattern Recognition, 2022]

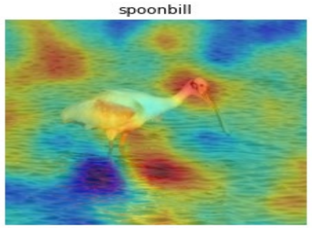
How to compute RAI

Transparency - XAI

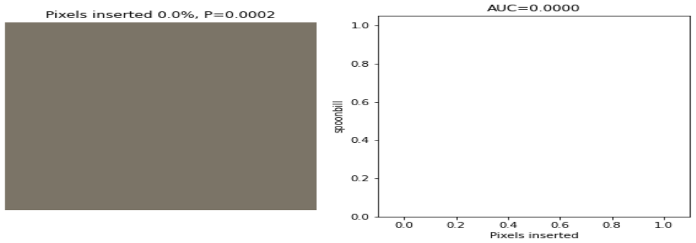


SIDU



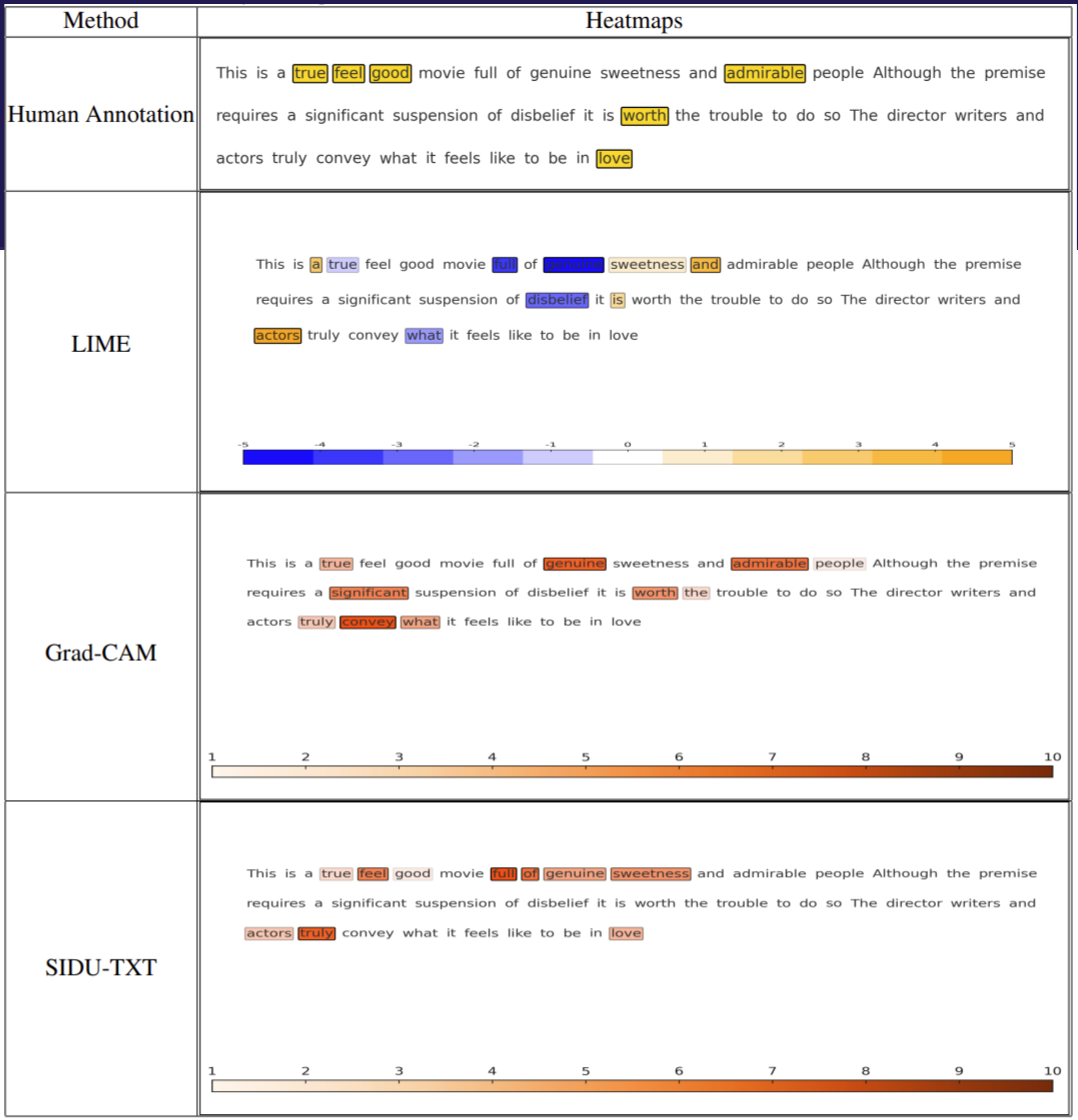


RISE



Insertion

METHODS	Insertion (Higher the better) ↑	Deletion (Lower the better) ↓
RISE	0.63571	0.13505
GRAD-CAM	0.6286	0.1539
SIDU	**0.65801**	**0.13424**

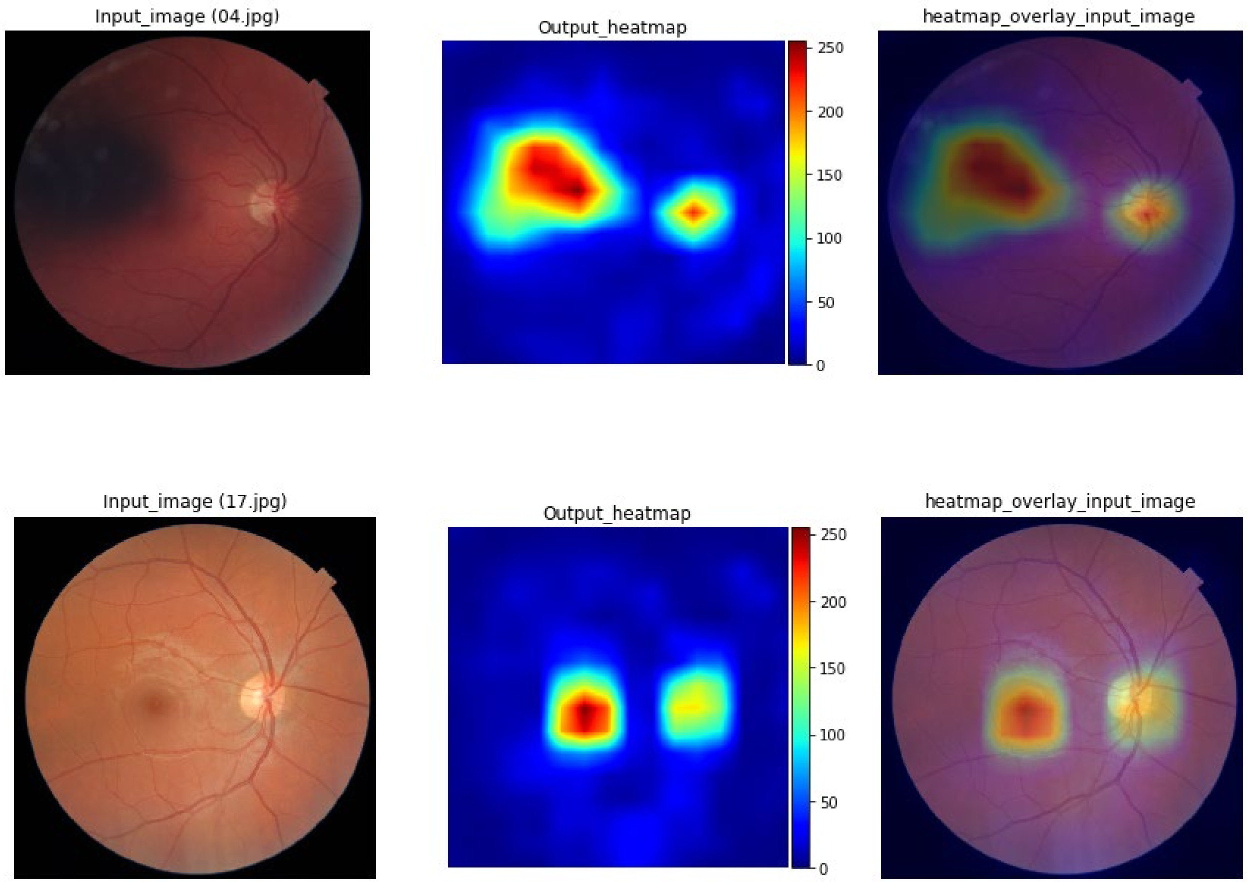


Method	Heatmaps
Human Annotation	This is a true feel good movie full of genuine sweetness and admirable people Although the premise requires a significant suspension of disbelief it is worth the trouble to do so The director writers and actors truly convey what it feels like to be in love
LIME	This is a true feel good movie full of genuine sweetness and admirable people Although the premise requires a significant suspension of disbelief it is worth the trouble to do so The director writers and actors truly convey what it feels like to be in love
Grad-CAM	This is a true feel good movie full of genuine sweetness and admirable people Although the premise requires a significant suspension of disbelief it is worth the trouble to do so The director writers and actors truly convey what it feels like to be in love
SIDU-TXT	This is a true feel good movie full of genuine sweetness and admirable people Although the premise requires a significant suspension of disbelief it is worth the trouble to do so The director writers and actors truly convey what it feels like to be in love

[Satya M. Muddamsettya et al. Visual Explanation of Black-Box Model. Pattern Recognition, 2022]

How to compute RAI

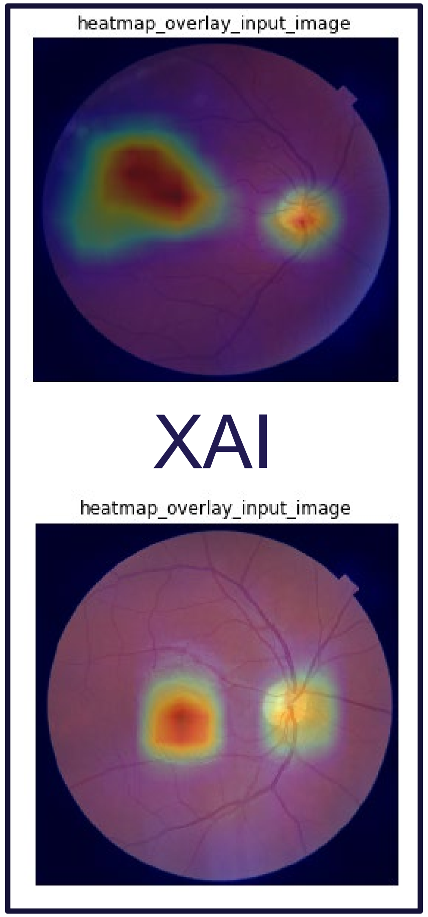
Transparency - XAI

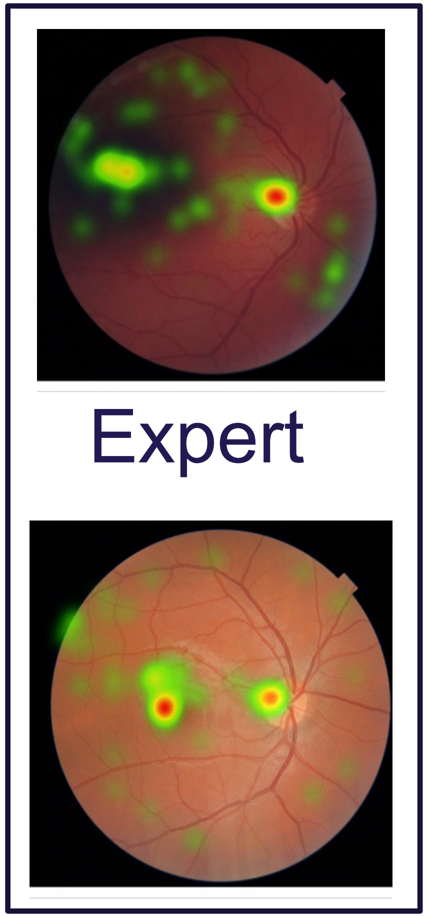


[Satya M. Muddamsettya et al. Visual Explanation of Black-Box Model. Pattern Recognition, 2022]

How to compute RAI

Transparency - XAI



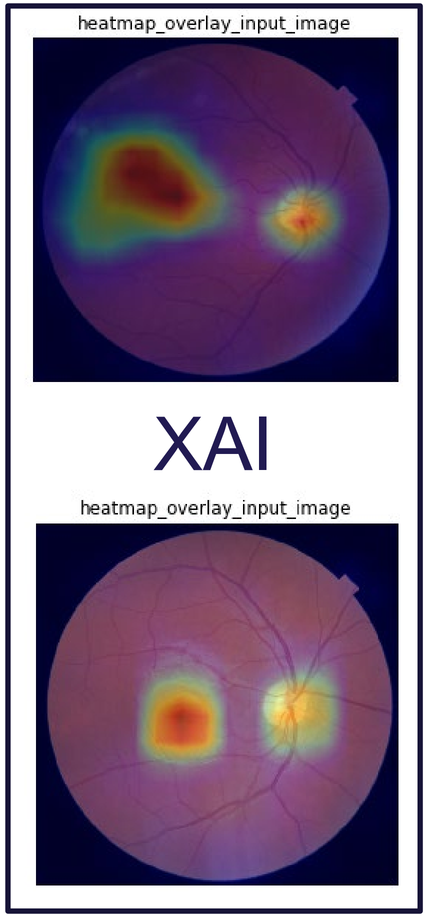


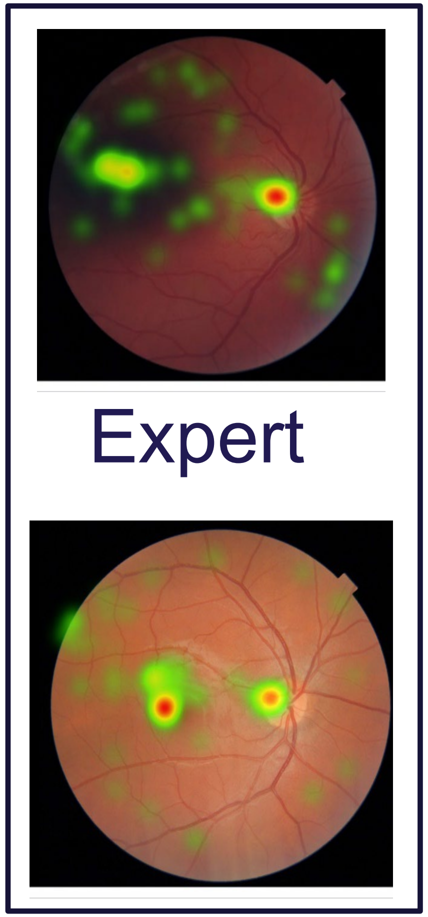
[Satya M. Muddamsettya et al. Visual Explanation of Black-Box Model. Pattern Recognition, 2022]

How to compute RAI

Transparency - XAI

- How do we quantify XAI?
- For whom is this explanation relevant?
  - Debugging tool

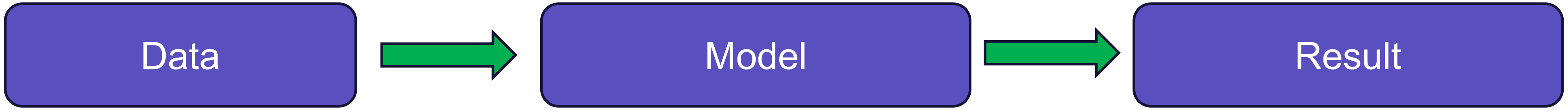




Takeaways

- We don't have a good metric for XAI performance
- Additional research needed
  - We need to involve end-users
  - UX
  - XAI interface
  - Human-XAI Interaction

How to compute RAI

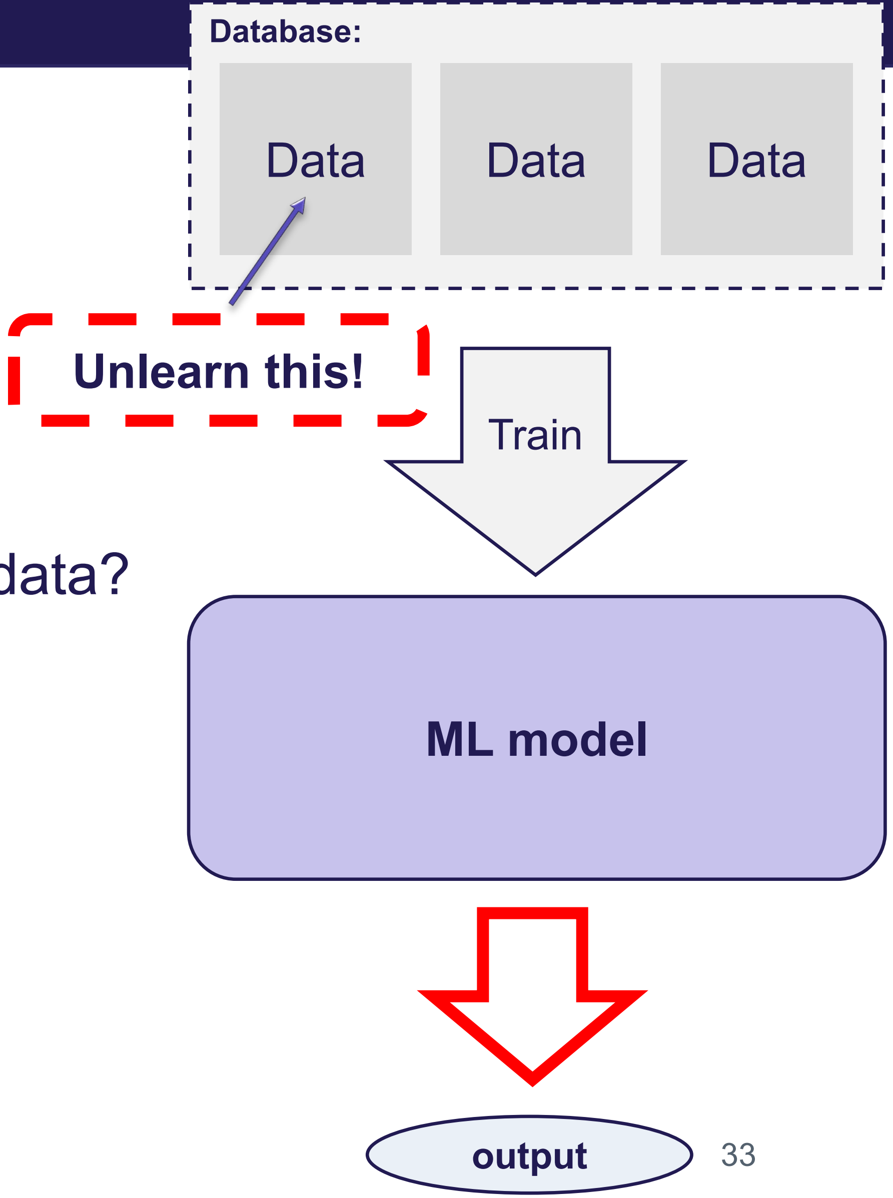


	Data	Model	Result
Terms:	‘The right to be forgotten’	‘Transparency’	‘Robustness’
Tech:	Machine Unlearning	XAI	Data drift Out of distribution

How to compute RAI

The right to be forgotten - Machine Unlearning

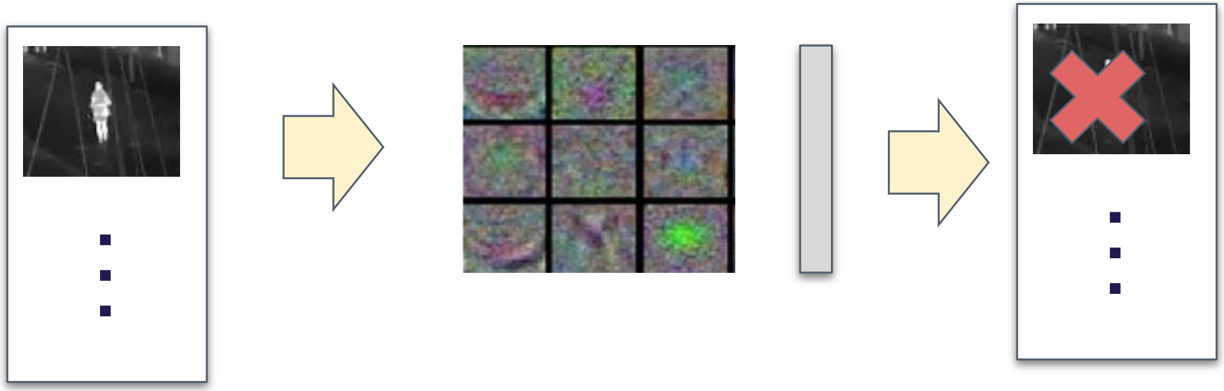
- Remove data-point & retrain (‘gold standard’)
  - Not always possible
  - Expensive
- Remove data-point & keep the model
  - Is it ok (legal & fair) that outputs are based on deleted data?
  - Deleted data can be recreated
- Other motivations for Machine Unlearning
  - Remove feature (gender, age, etc.)
  - Remove noise
  - Remove malicious data

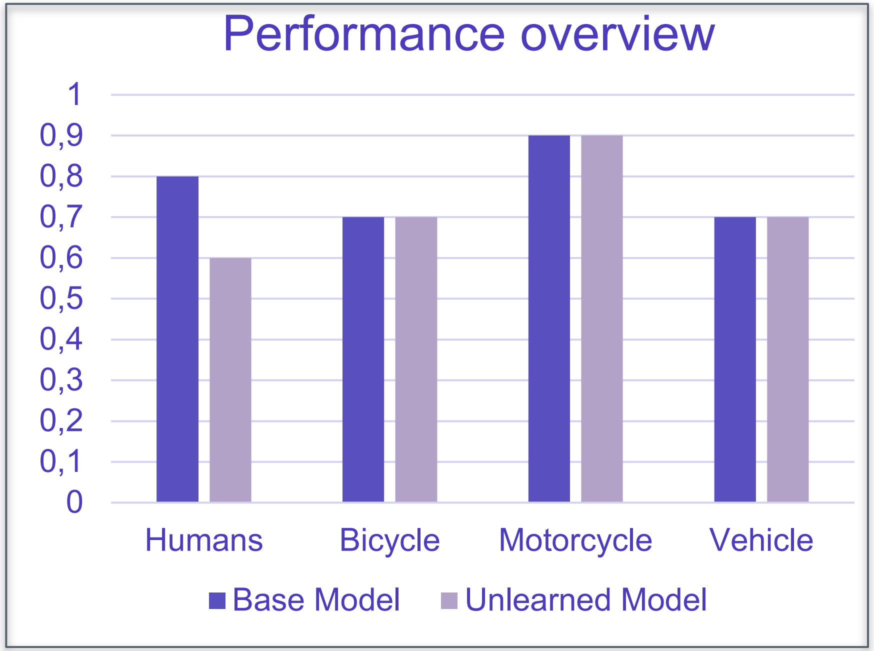


AALBORG UNIVERSITY
DENMARK

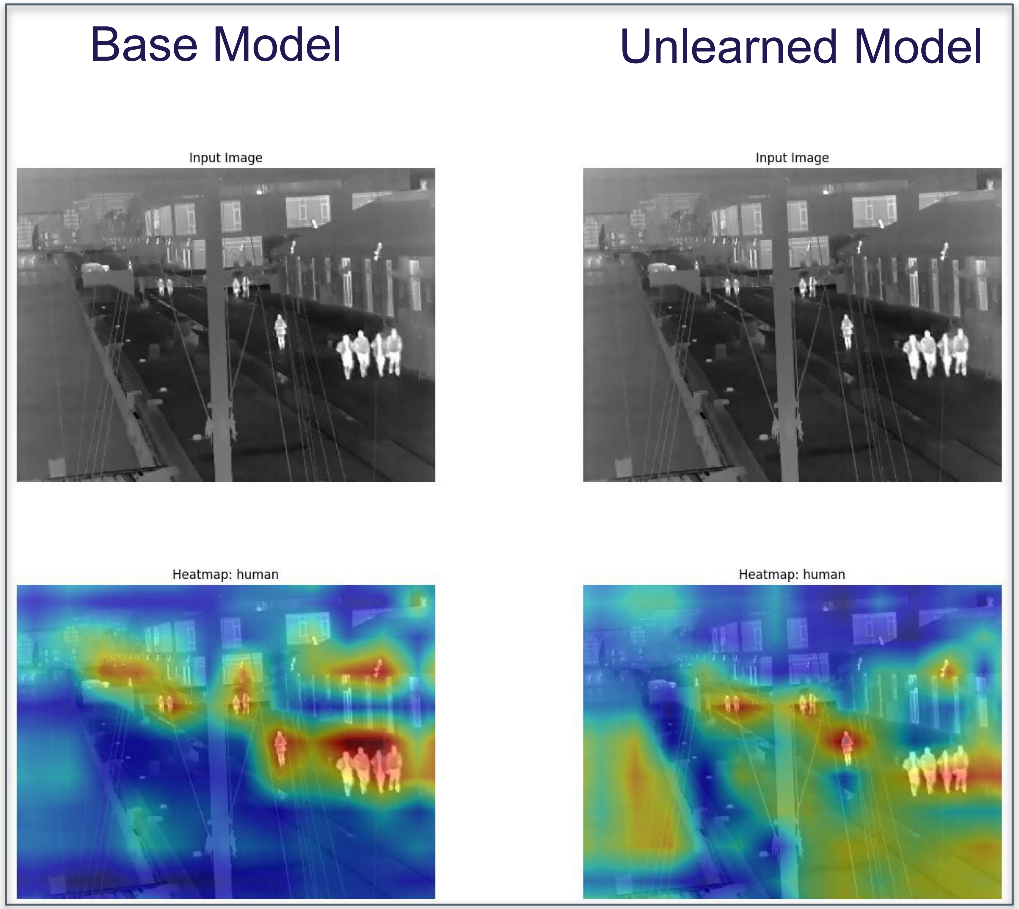
How to compute RAI

The right to be forgotten - Machine Unlearning





Performance overview



[Alex P. Vidal et al. Verifying Machine Unlearning with Explainable AI. ICPRW), 2024]

How to compute RAI

The right to be forgotten - Machine Unlearning

Takeaways

- Machine Unlearning is a very new topic
  - We don’t have good methods for unlearning
- Additional research needed
  - Degree of unlearning vs performance

AALBORG UNIVERSITY
DENMARK

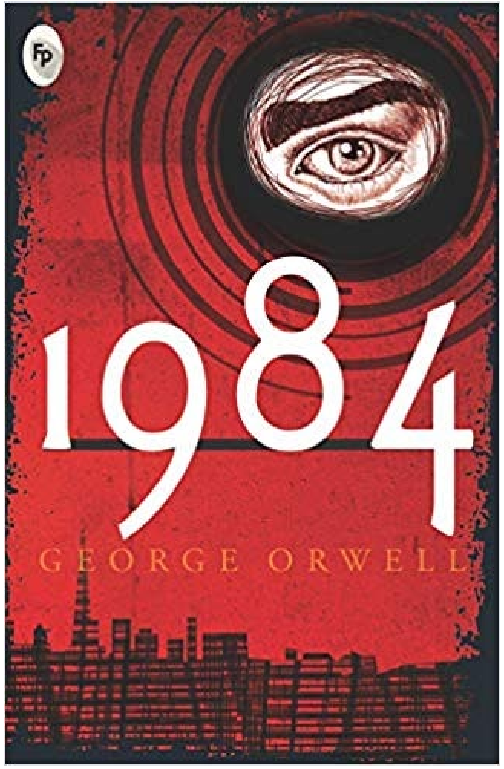
Agenda

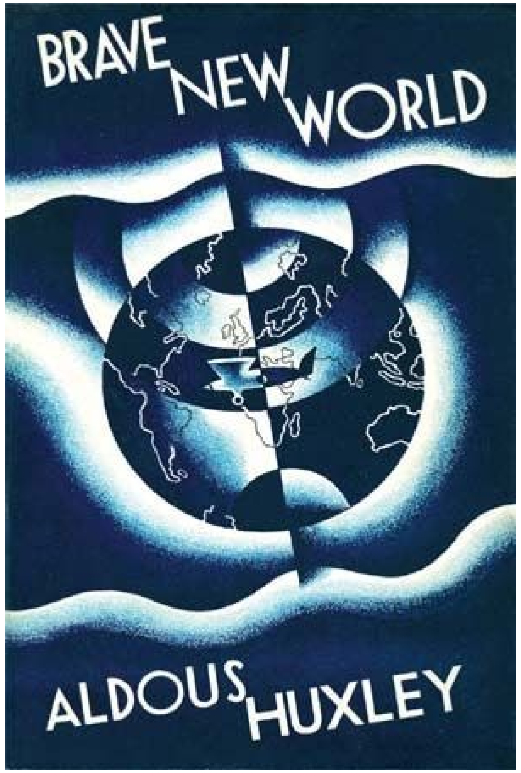
- Who am I?
- Why are we talking about Responsible AI?
- How do we compute Responsible AI?
- The end-game of AI
- Q&A

AALBORG UNIVERSITY
DENMARK

Responsible AI - The end-game

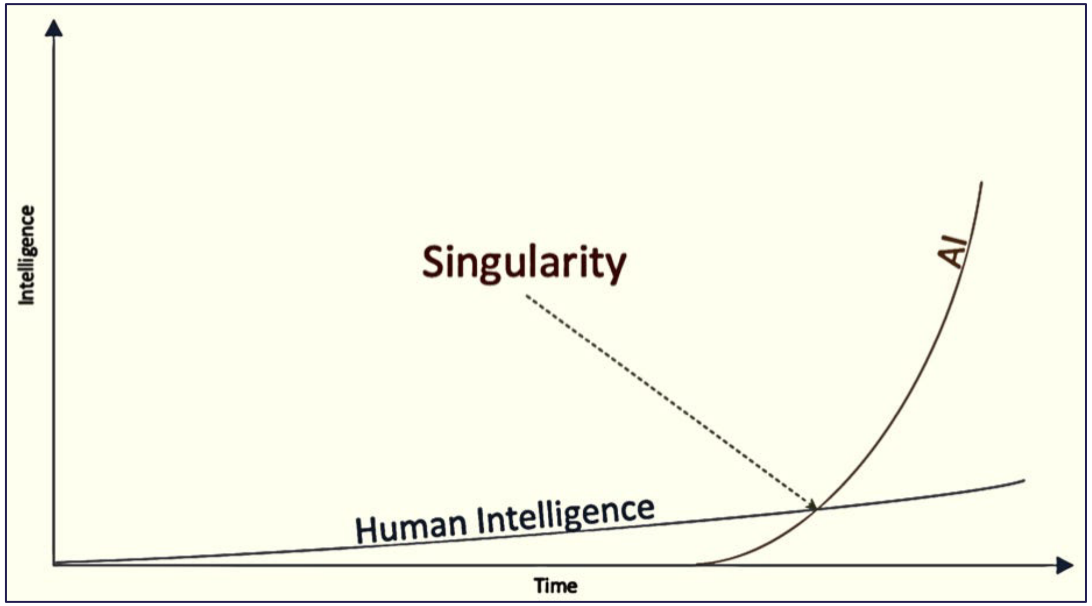
- Dystopia vs utopia

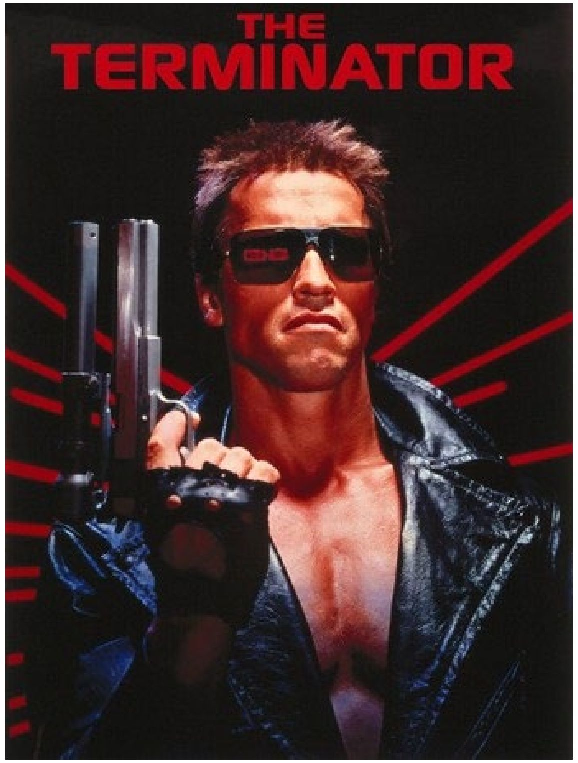


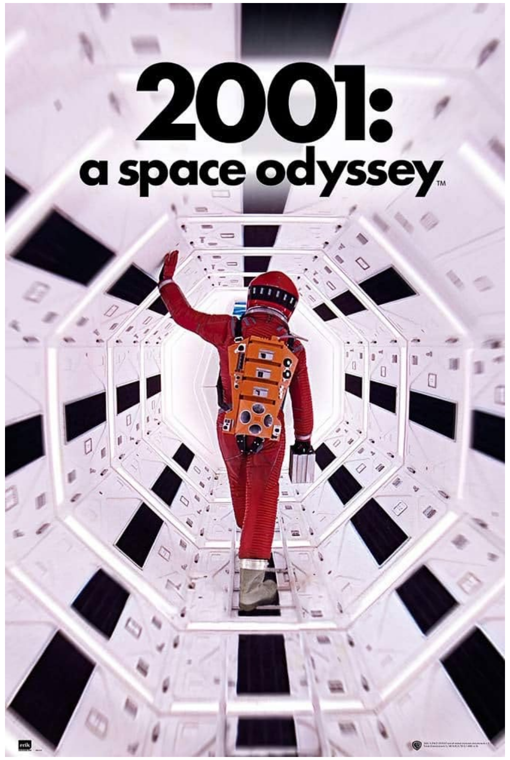


The end-game

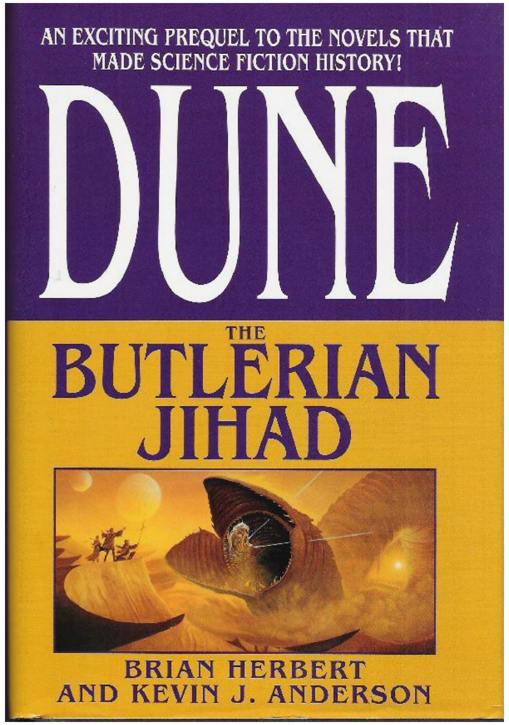
- Man vs tech







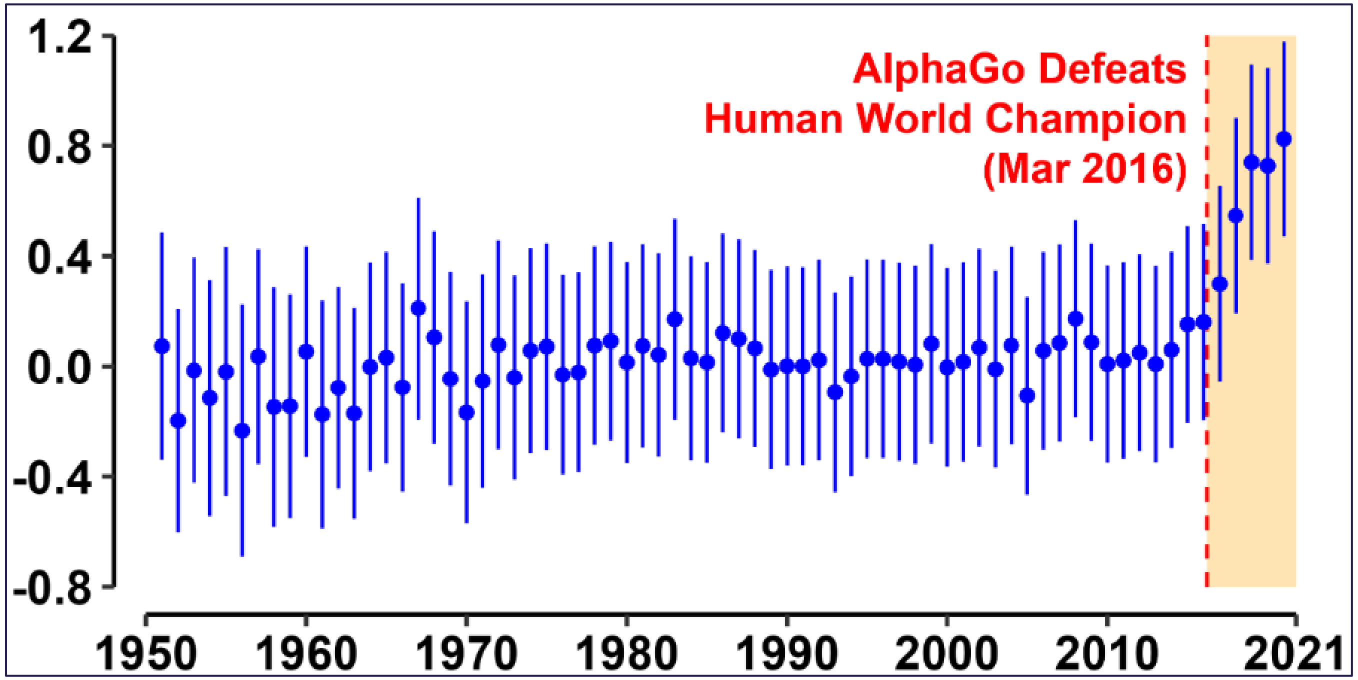
KEANU REEVES LAURENCE FISHBURNE
THE MATRIX
www.whatisthematrix.com
ON MARCH 31ST THE FIGHT FOR THE FUTURE BEGINS.

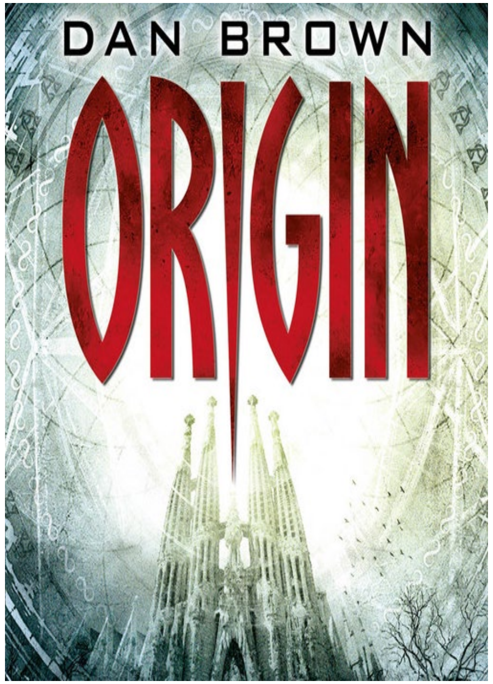


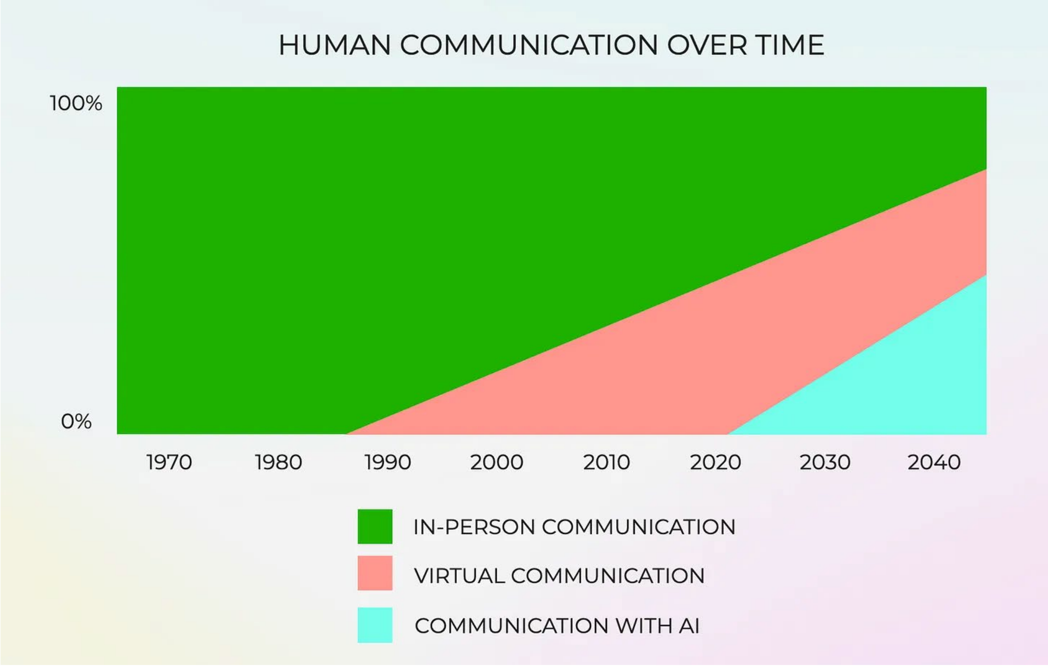
AN EXCITING PREQUEL TO THE NOVELS THAT MADE SCIENCE FICTION HISTORY!
DUNE
THE BUTLERIAN JIHAD
BRIAN HERBERT AND KEVIN J. ANDERSON

The end-game

- Man & tech

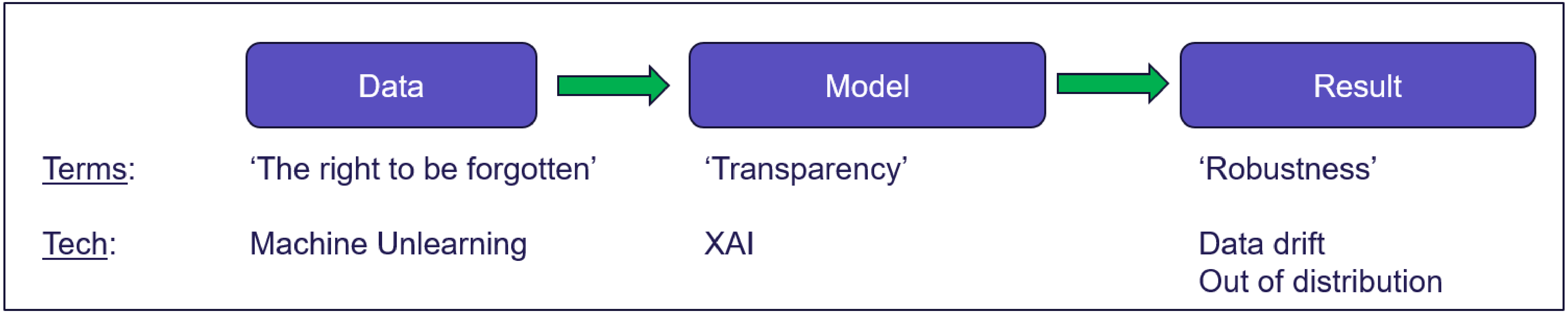






Conclusion

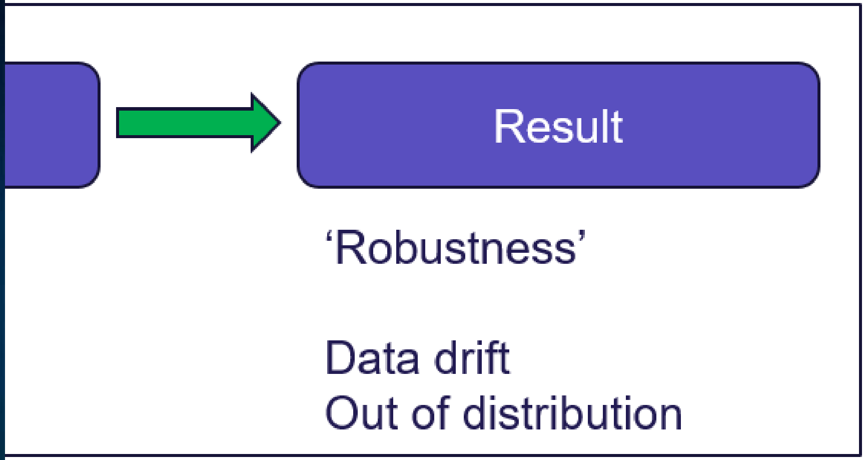
- Go home and think about it: where will it all end?
- AI will be regulated
- How to translation the general terms into computational methods/metrics
  - Still open research questions => But will be defined now…
  - EU: CEN-CENELEC



	Data	Model	Result
Terms:	‘The right to be forgotten’	‘Transparency’	‘Robustness’
Tech:	Machine Unlearning	XAI	Data drift Out of distribution

Conclusion

- Go home and think about it: where will it all end?
- AI will be regulated
- How to translation the general terms into computational methods/metrics
    - Still open research questions => But will be defined now…



Result

'Robustness'

Data drift
Out of distribution