Continually Learning Deep Machines that Understand What They Don't Know

Dr. Martin Mundt

TUDa & hessian.AI - Junior Research Group Leader

ContinualAI Board Member

http://owll-lab.com

Prof. Dr. Kristian Kersting

TECHNISCHE UNIVERSITÄT DARMSTADT

hessian.AI

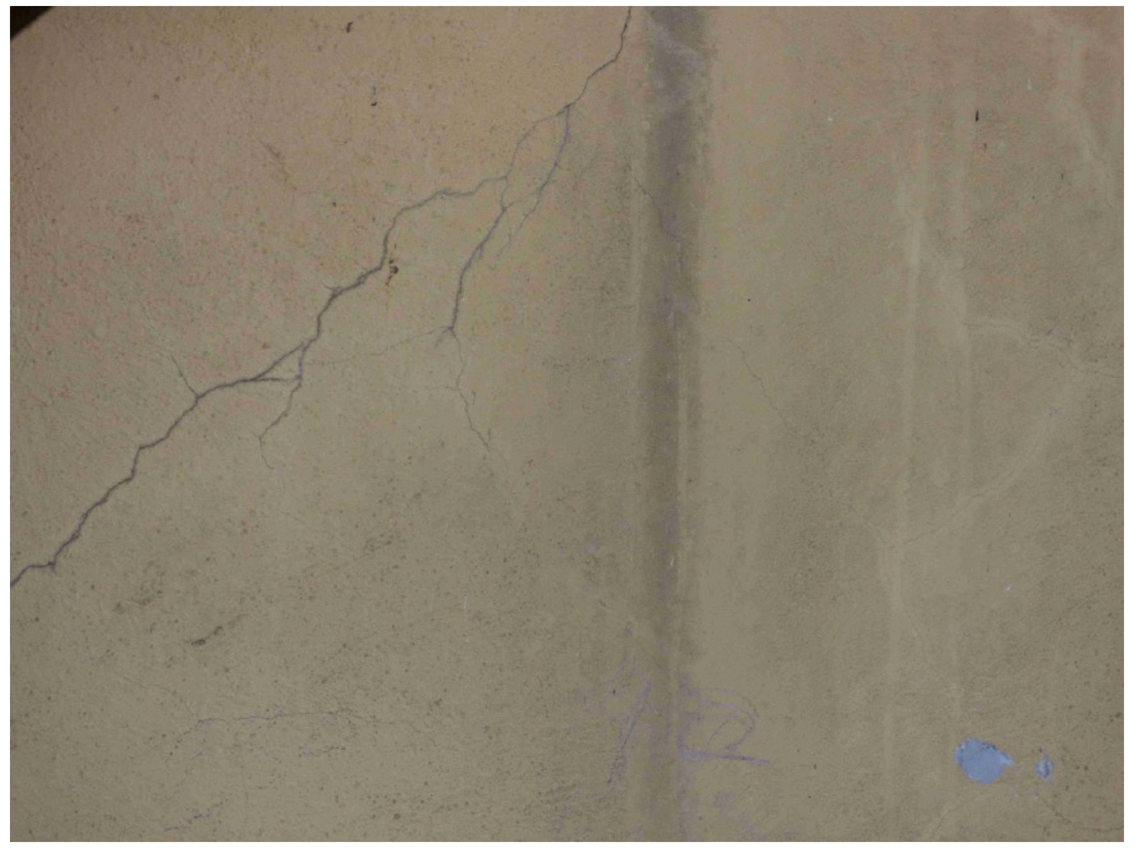
OWLL

ContinualAI

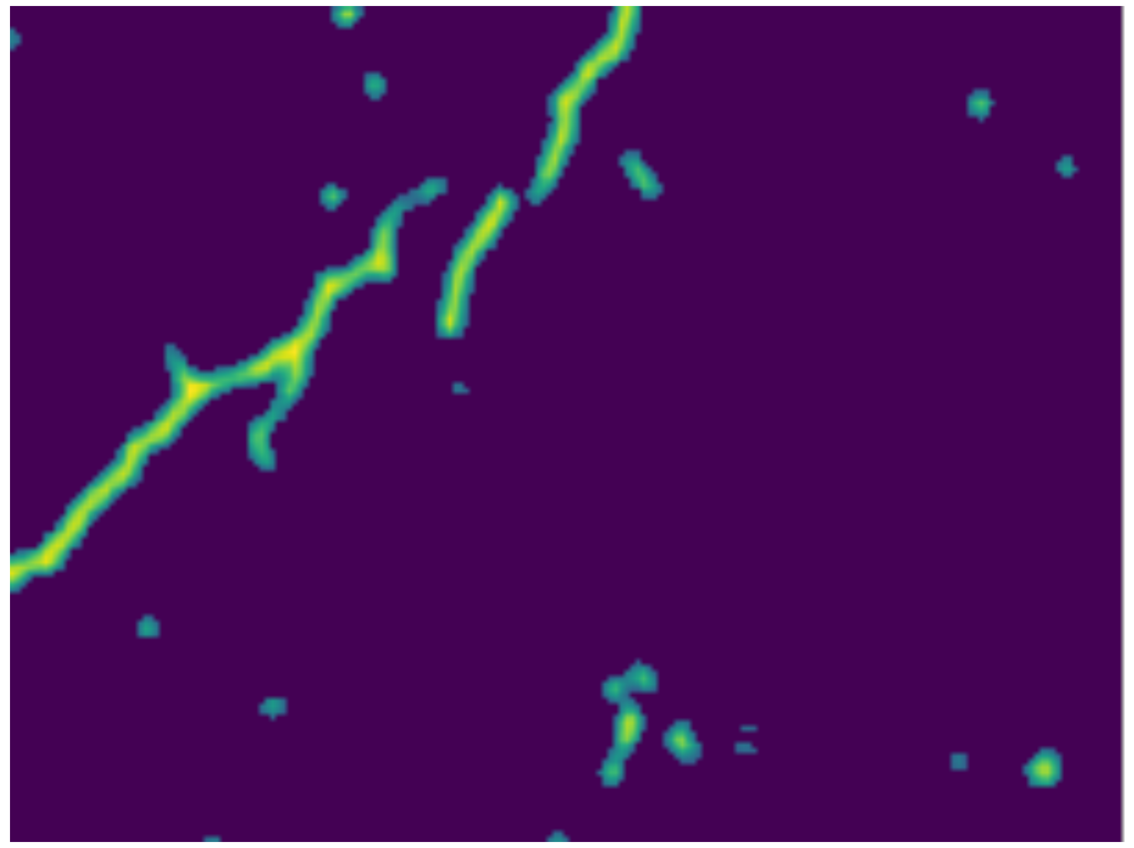
We could talk about AI applications…

Fly drone

Scan bridge



Inspect surface



Measure defects

Many factors: low amounts of data, experts are rare, annotation is cumbersome, predictions need to be robust (safety critical), tons of variation when system is deployed

Mundt, CVPR 2019

The standard machine learning workflow

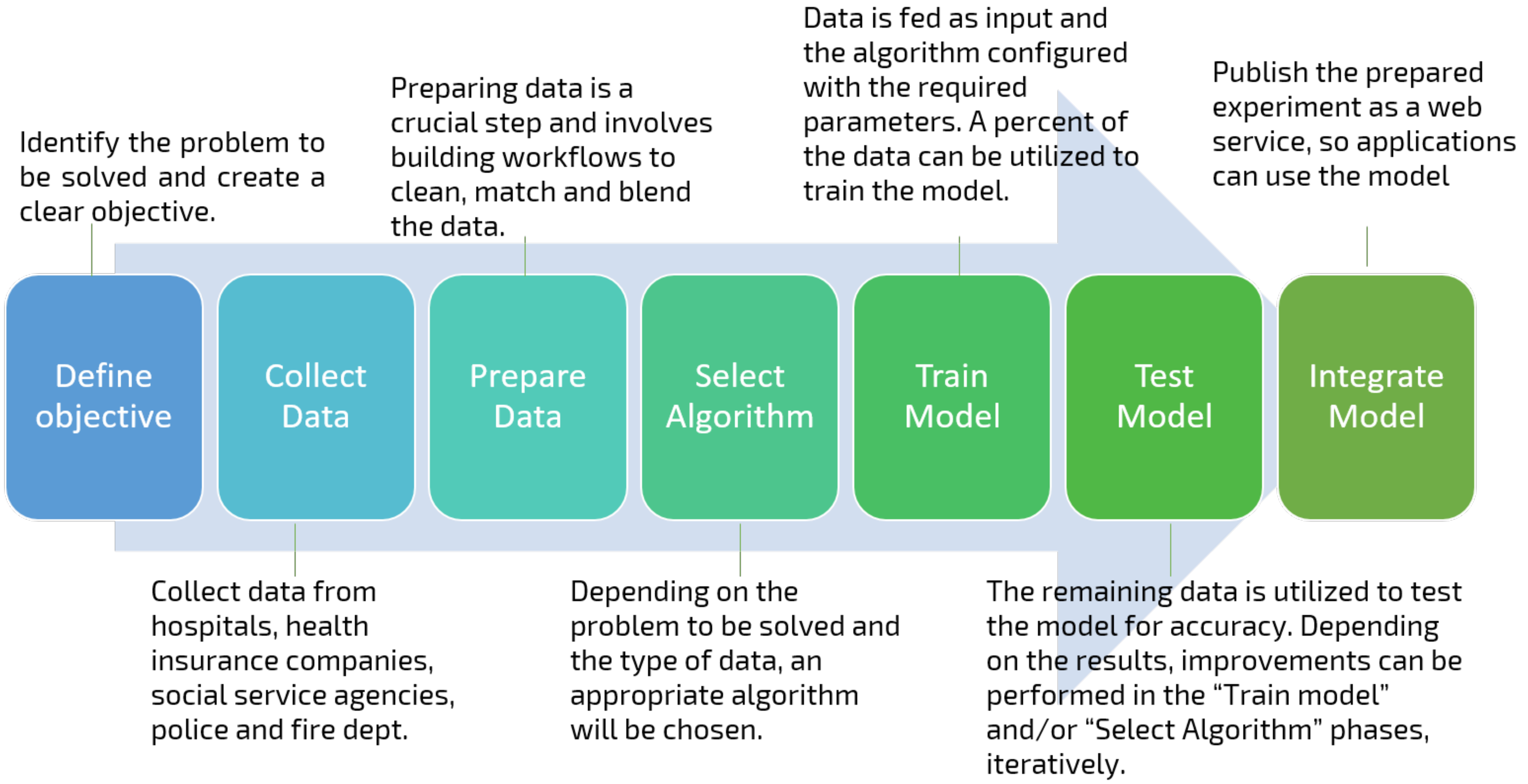
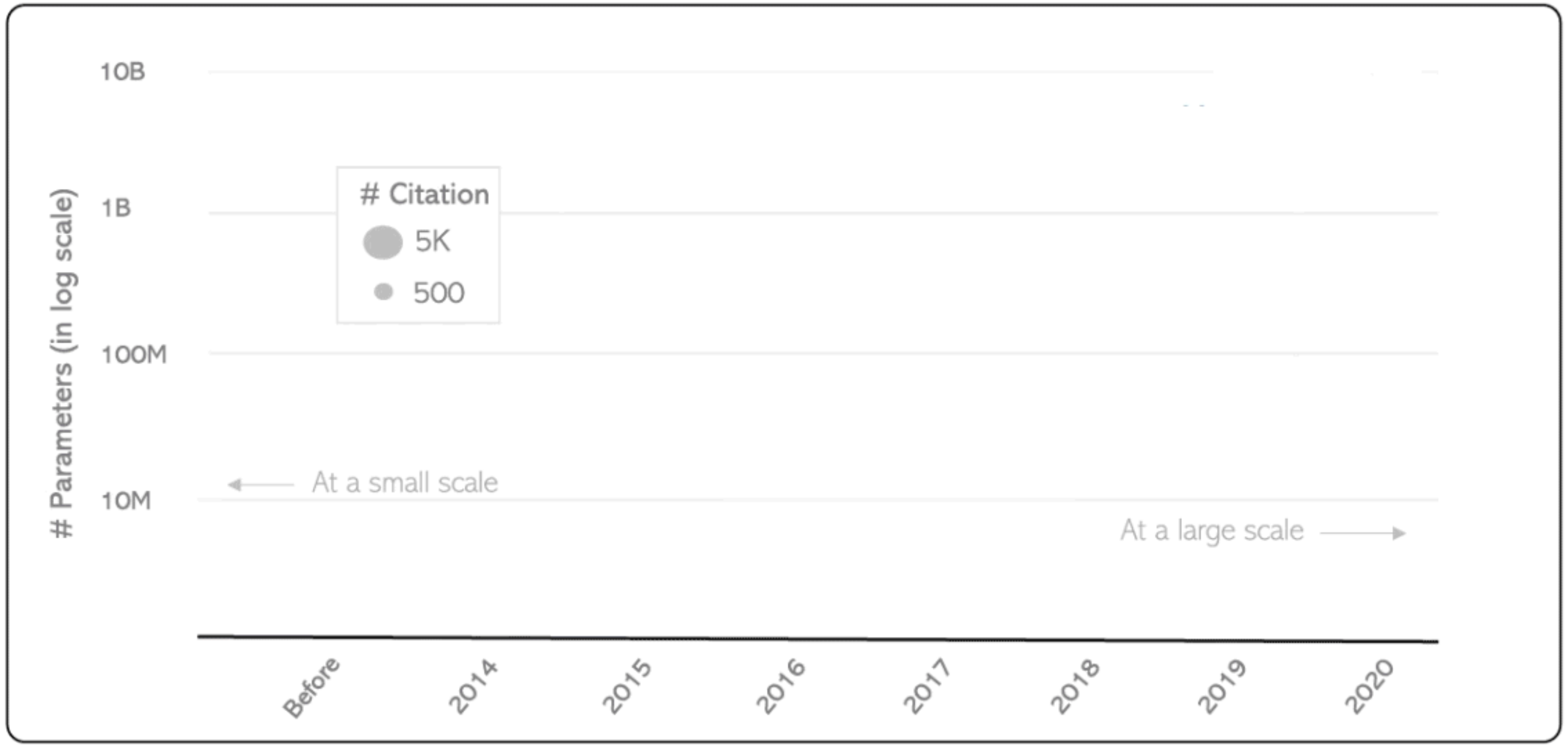


Figure from https://www.congrelate.com/get-workflow-machine-learning-images/

Is a static machine learning workflow + scale all we need?



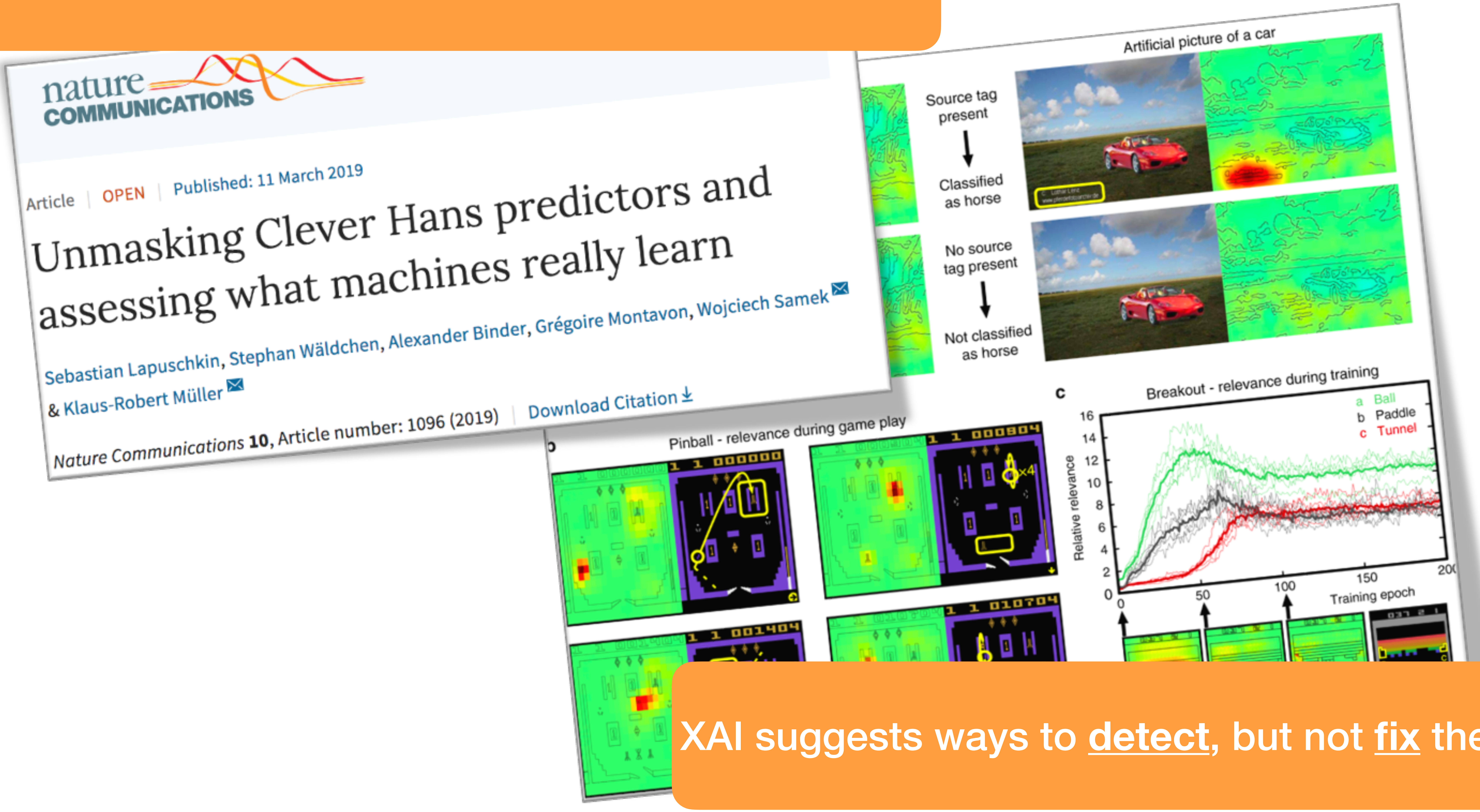
Li & Gao, "A deep generative model trifecta: three advances that work towards harnessing large-scale power, Microsoft Research Blog, 2020: https://www.microsoft.com/en-us/research/blog/a-deep-generative-model-trifecta-three-advances-that-work-towards-harnessing-large-scale-power/

We have “foundation" models now, but humans learn & reason

Importantly, humans revise their knowledge & continue adapting

Slides adapted from our AAAI-23 Continual Causality Tutorial, Cooper & Mundt

Why should we care: can we trust deep neural networks?


nature COMMUNICATIONS

Article | OPEN | Published: 11 March 2019

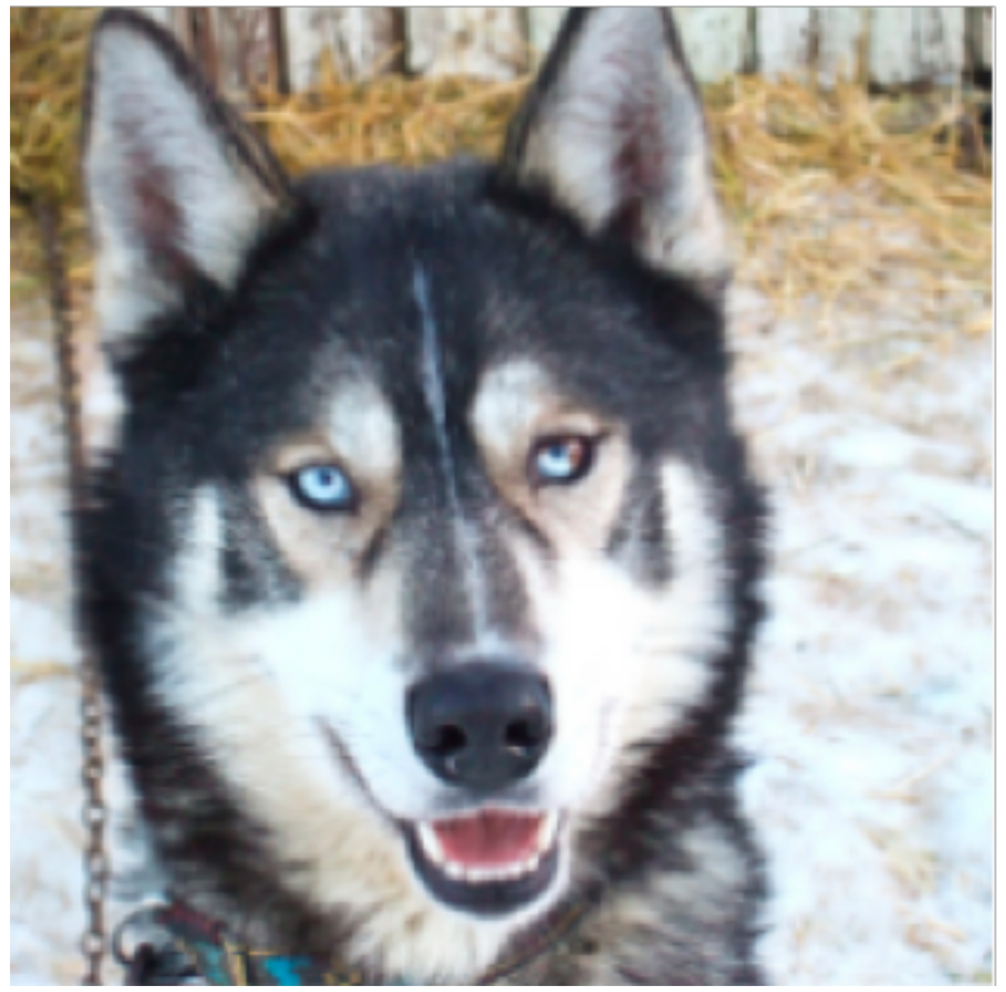
Unmasking Clever Hans predictors and assessing what machines really learn

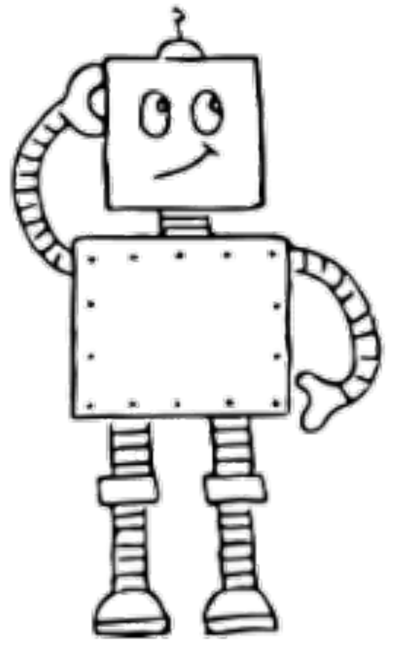
Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek & Klaus-Robert Müller

Nature Communications **10**, Article number: 1096 (2019) | Download Citation

XAI suggests ways to detect, but not fix the issue!

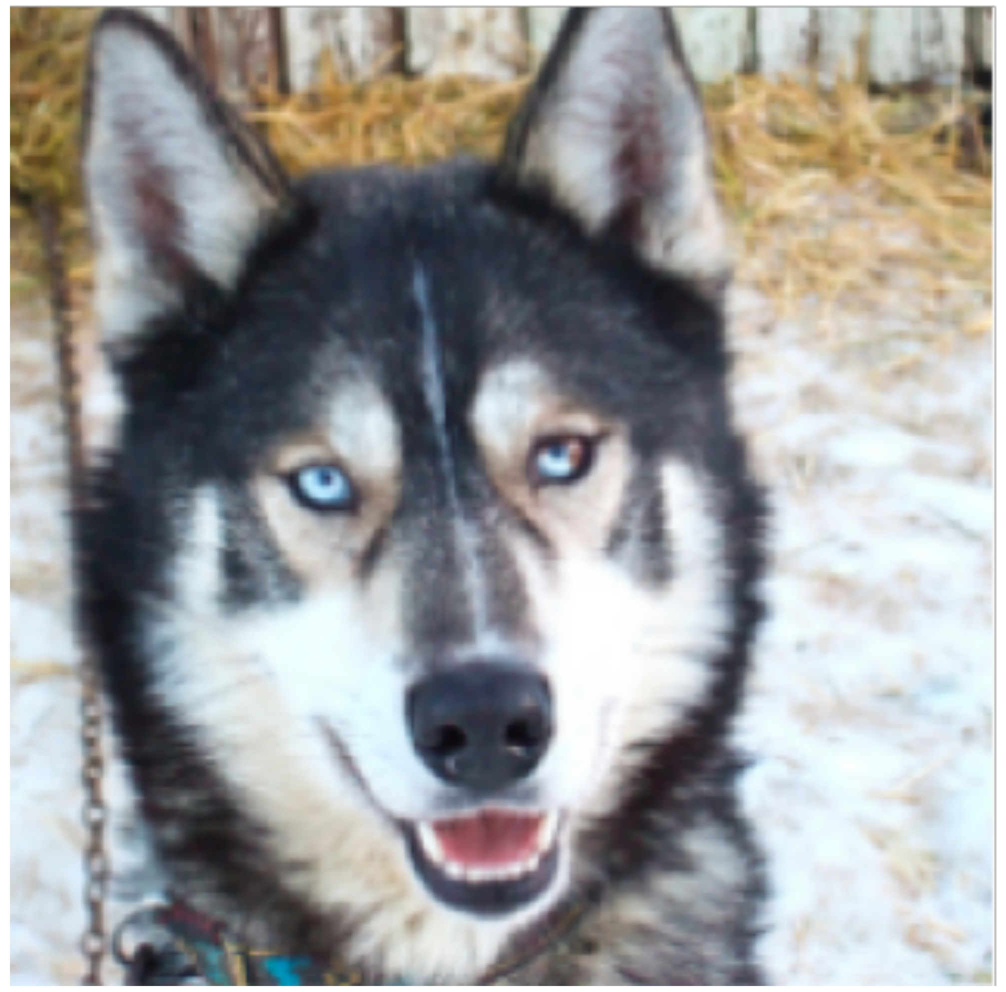
Example: Husky or Wolf?

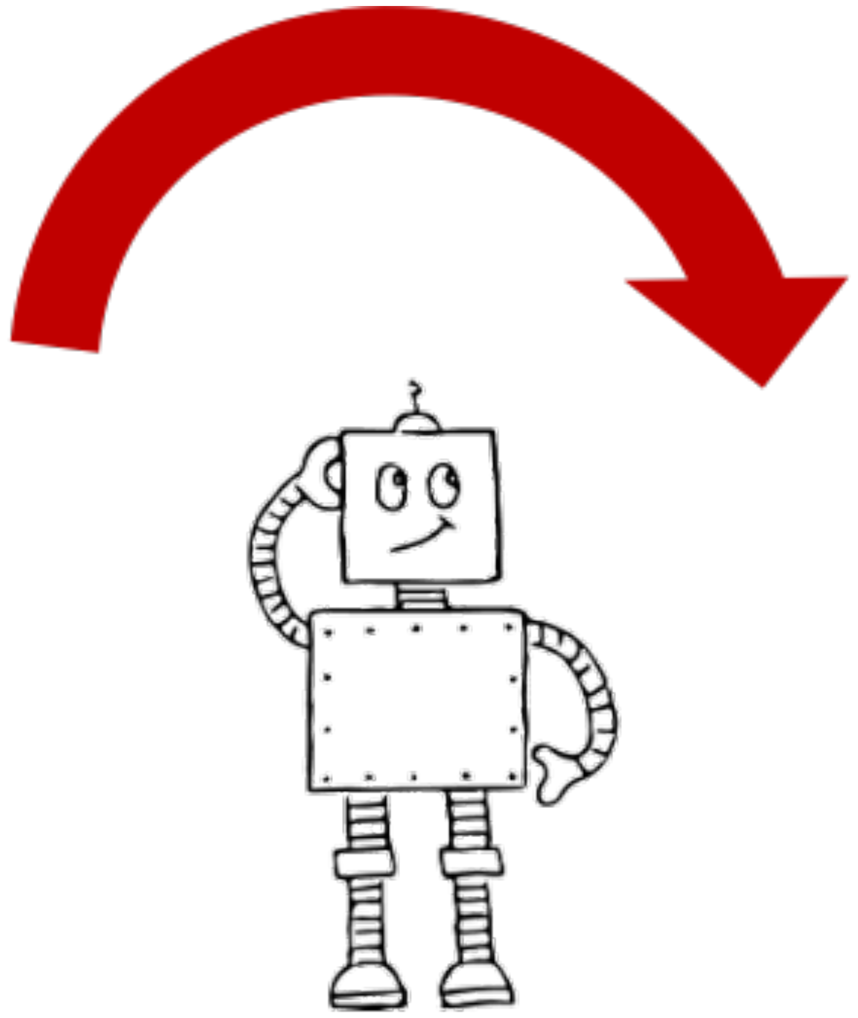




Consider an example image classification task about distinguishing between **husky dogs** and **wolves**

Example: Husky or Wolf? … and <u>why</u>?

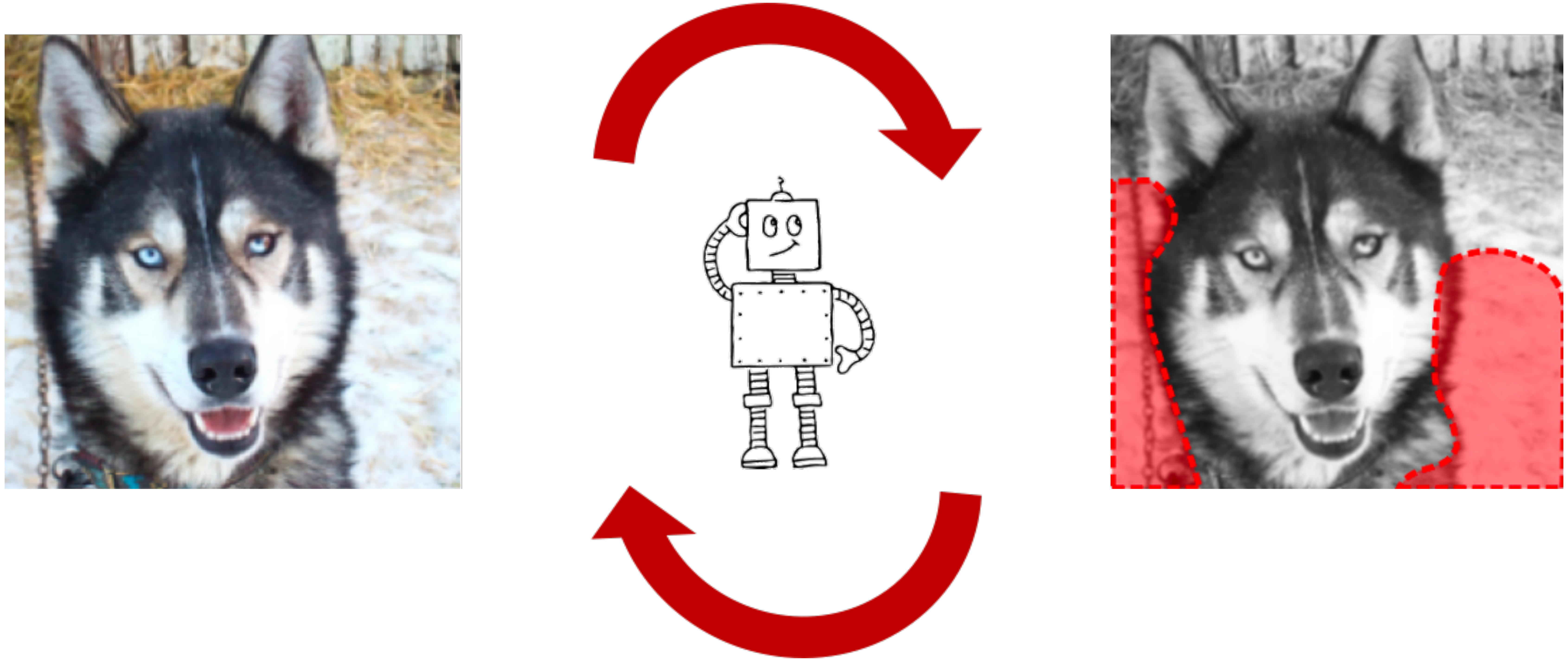




Consider an example image classification task about distinguishing between **husky dogs** and **wolves**

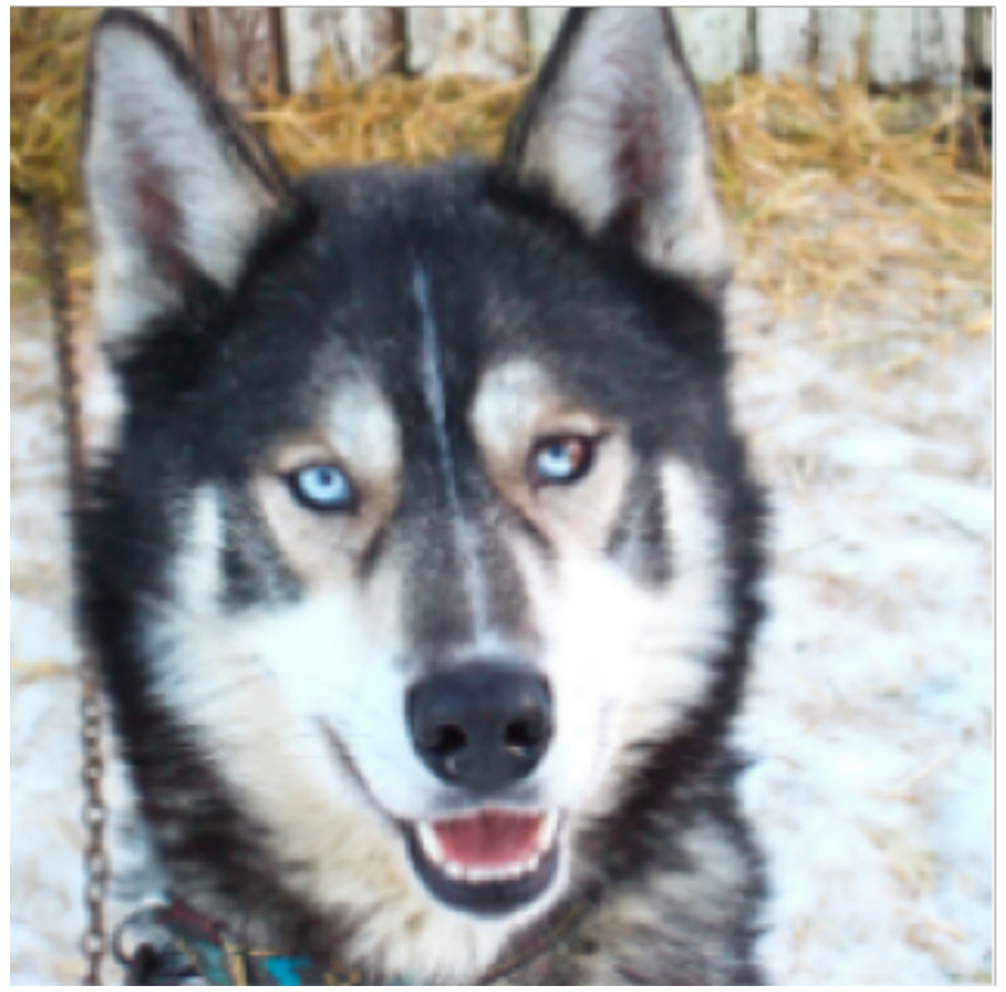
Local explanations allow to spot cases where the model is **right** for the **wrong reasons**

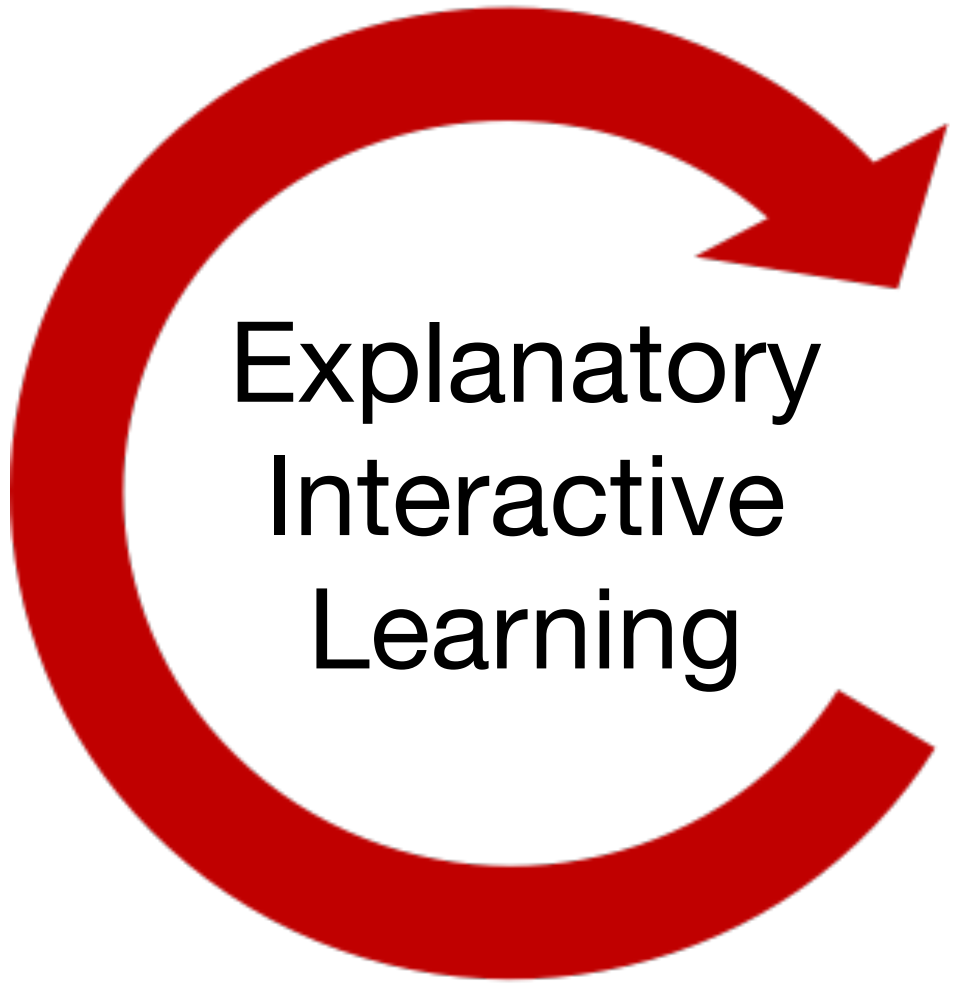
Example: Husky or Wolf? … and <u>why</u>? … and <u>feedback</u>!



It is a husky, but <u>not because</u> of the <u>highlighted pixels</u>!

Example: Husky or Wolf? … and why? … and feedback!

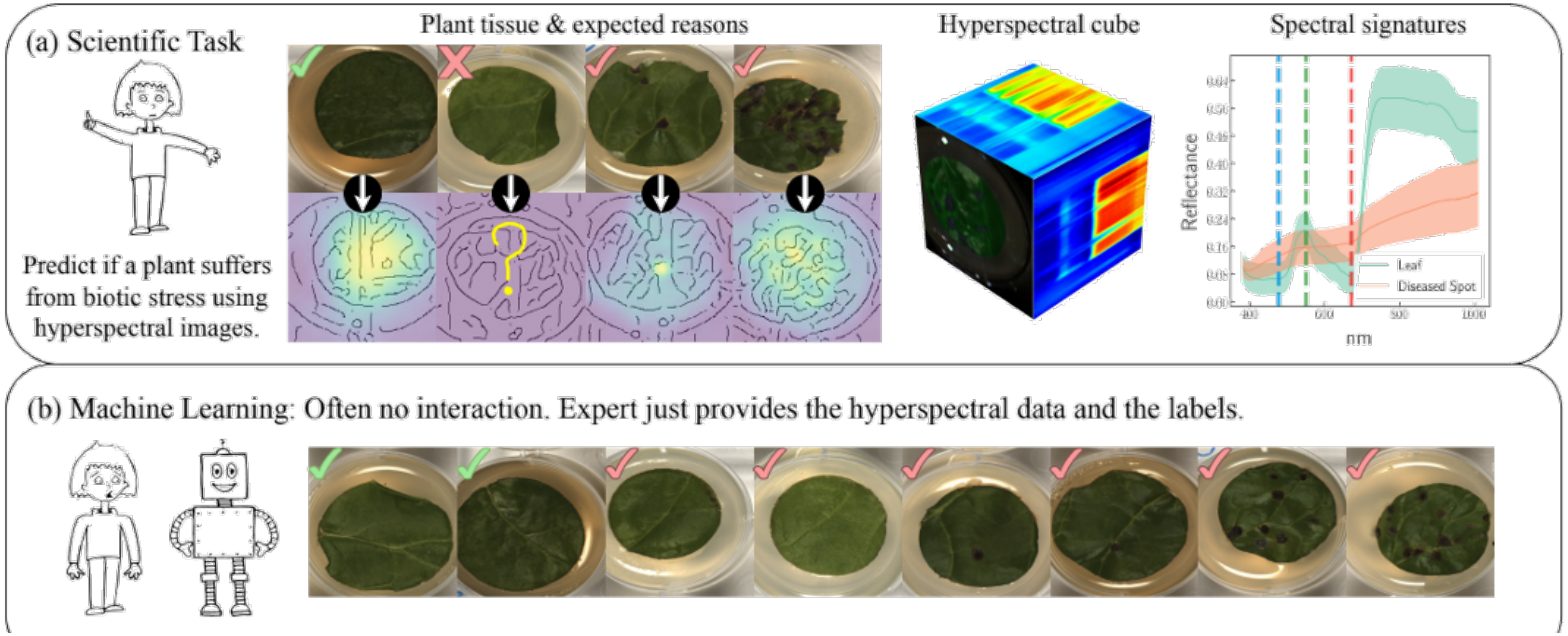




A. <u>Explain predictions</u> to users (*competence*, *understandability*)
B. Allow user to <u>correct explanations</u> (*directability*)

Of course, the pattern is not exclusive to images

Example: plant phenotyping

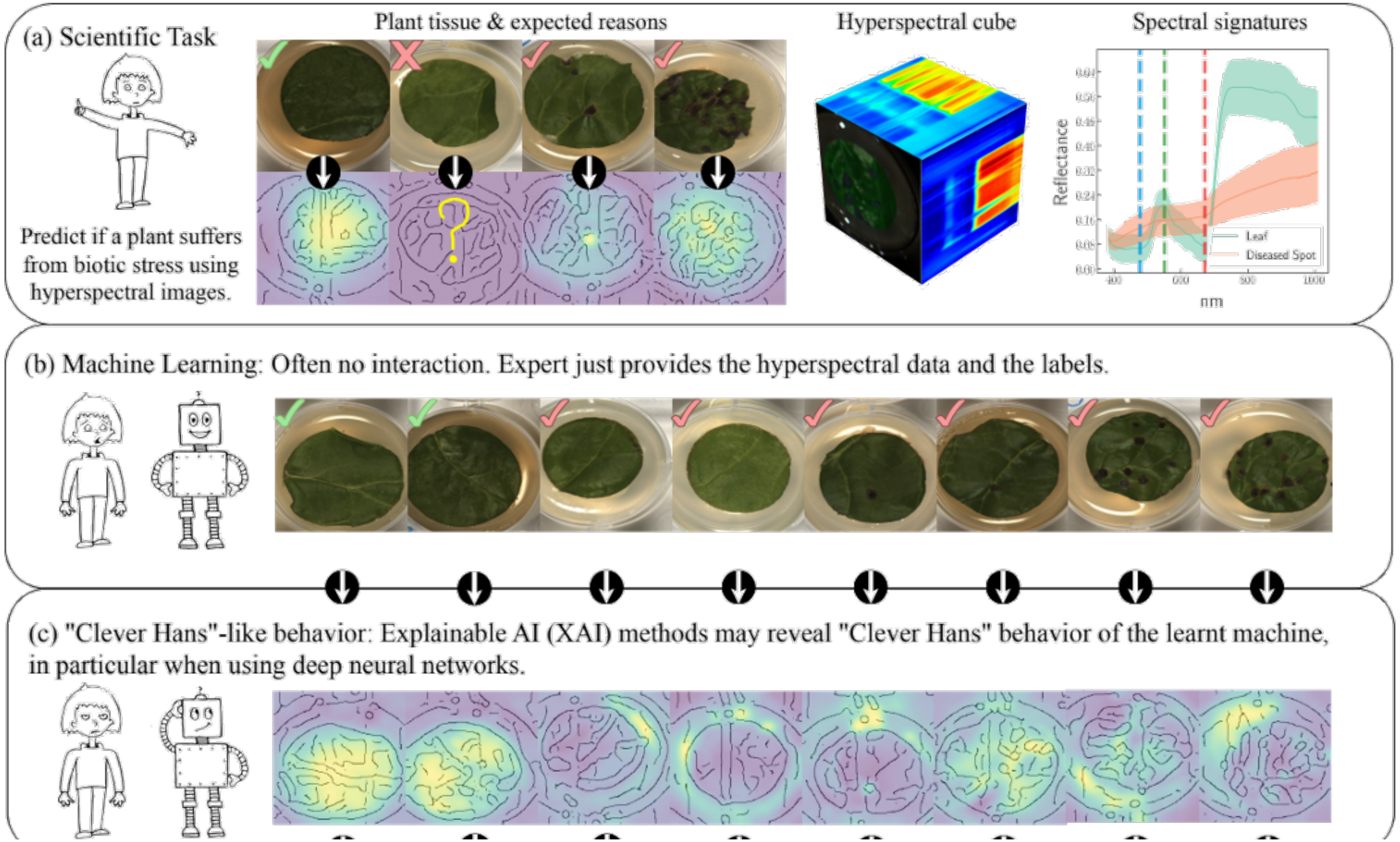


Schramowski et al. Nature Machine Intelligence 2020

nature machine intell

Of course, the pattern is not exclusive to images

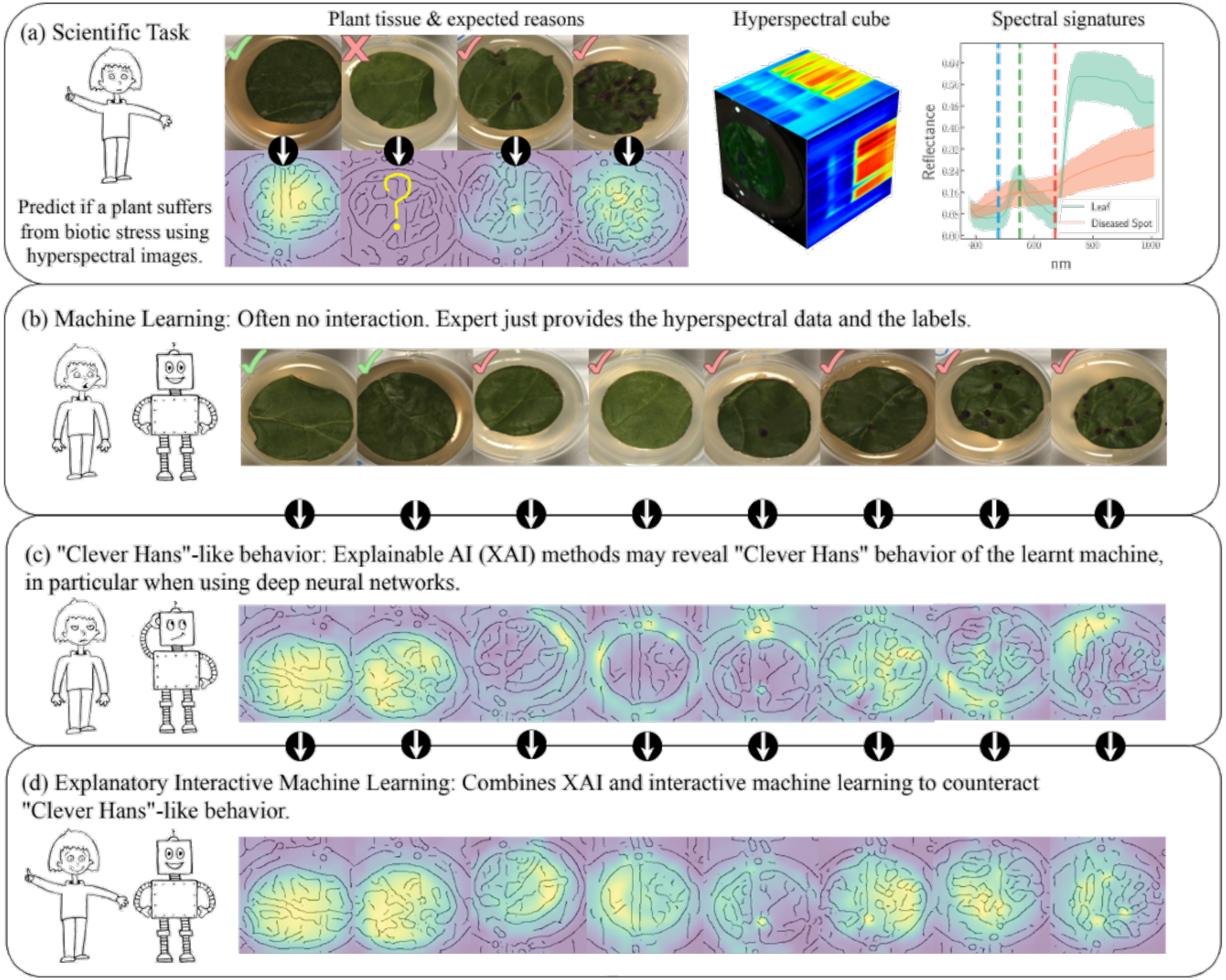
Example: plant phenotypic



Schramowski et al. Nature Machine Intelligence 2020

Of course, the pattern is not exclusive to images

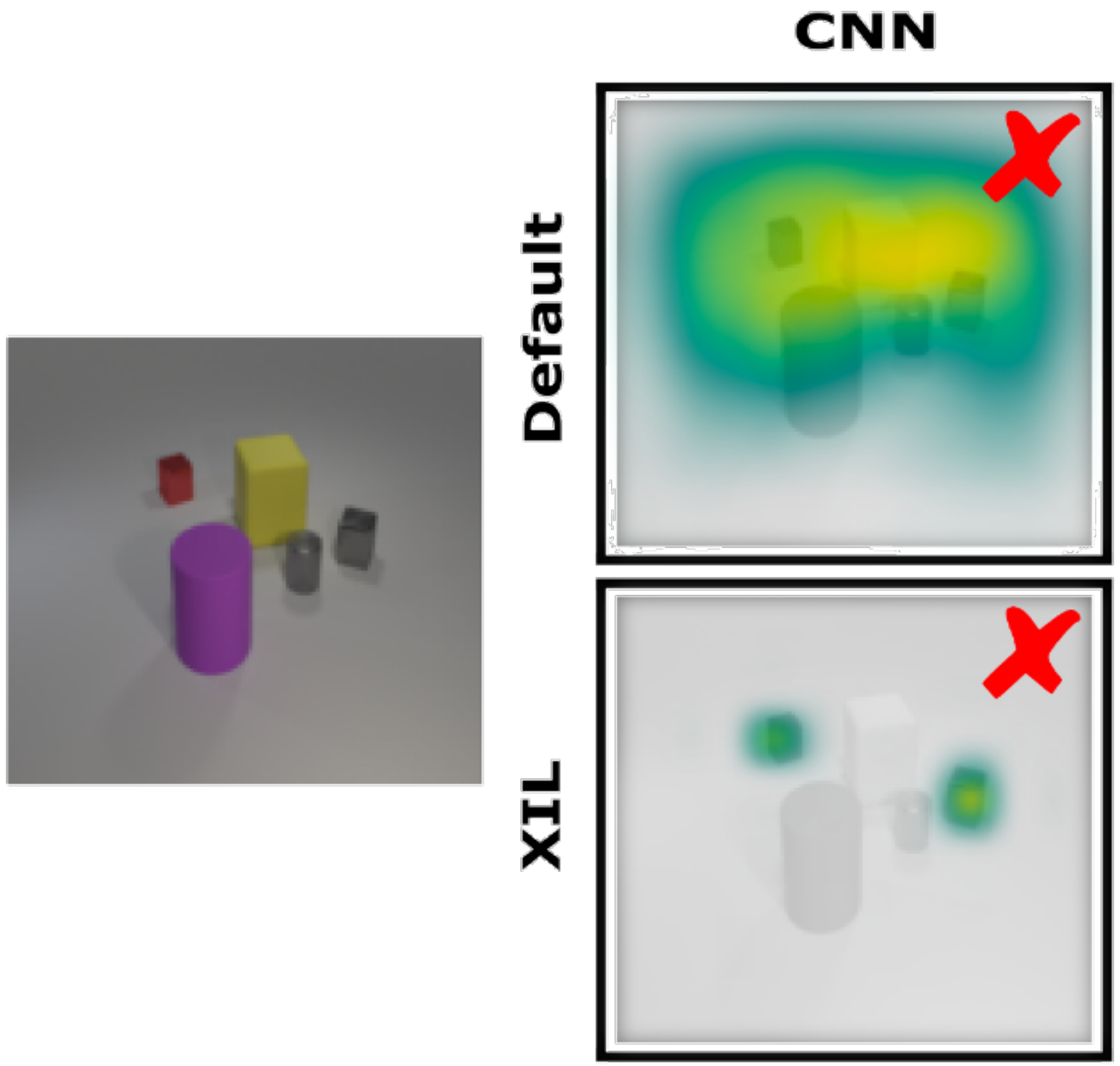
Example: plant phenotypic



Schramowski et al. Nature Machine Intelligence 2020

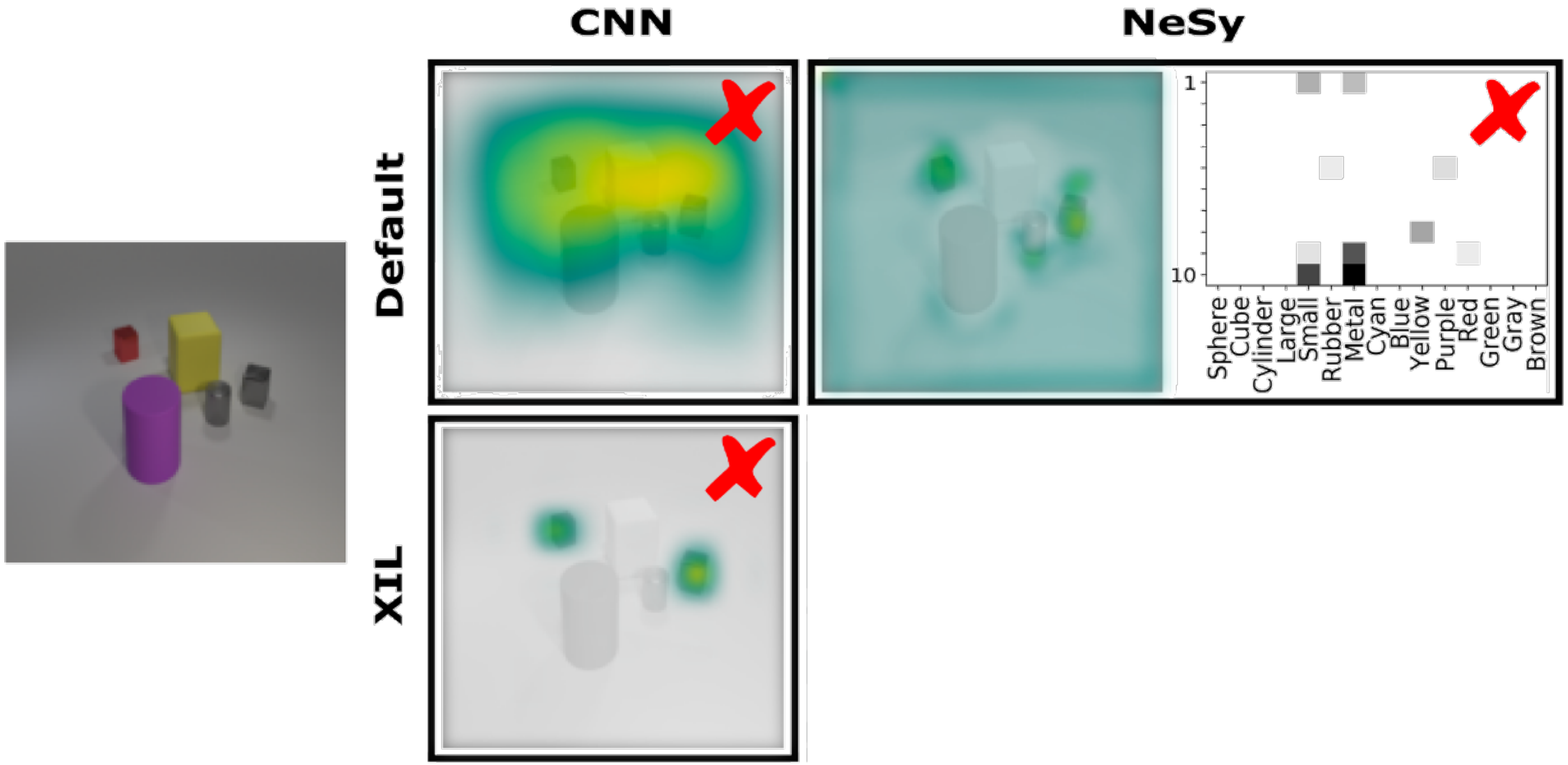
nature machine intell

Unfortunately, visual explanations alone are not all we need either

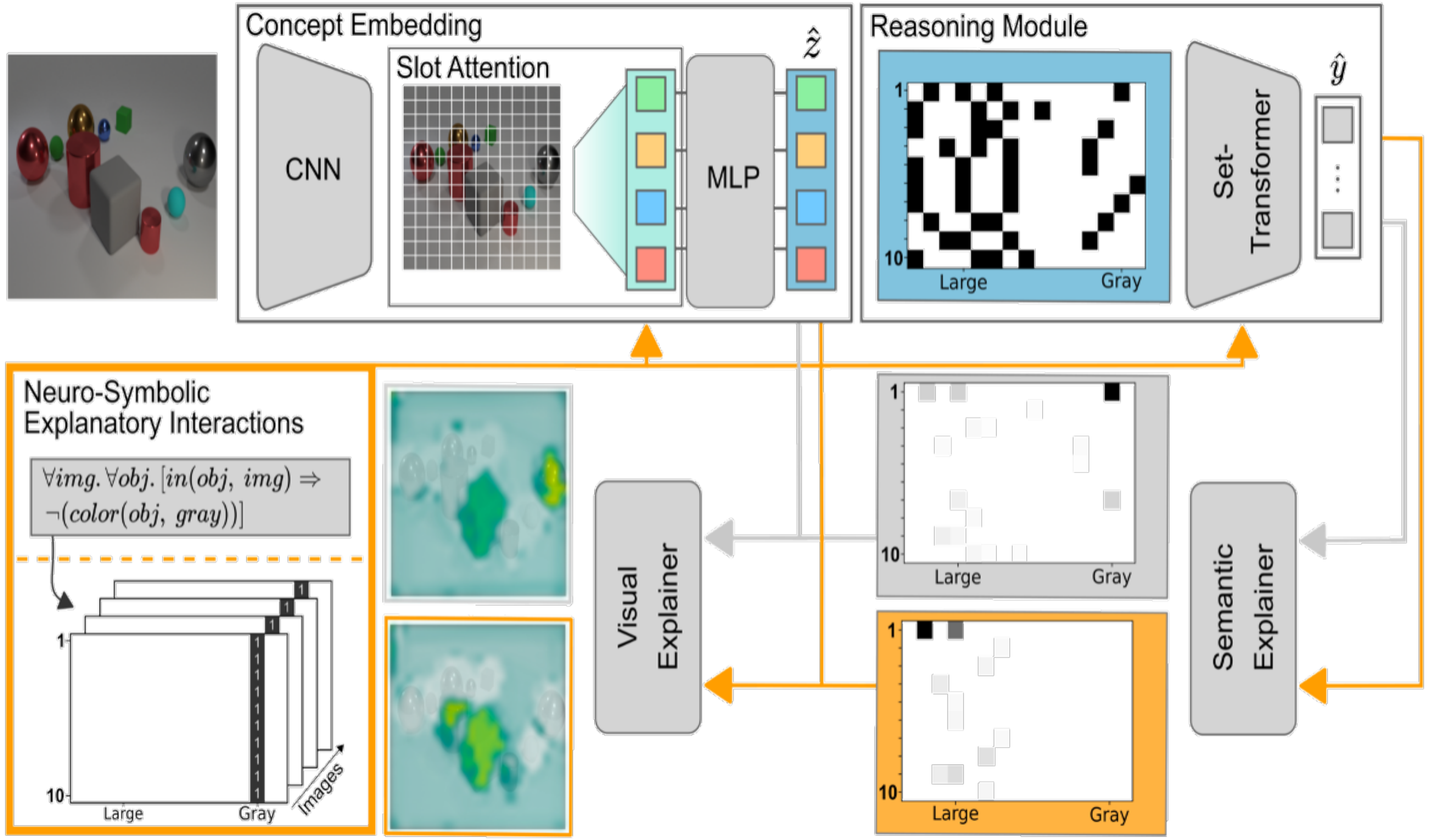


Underlying concept: the image contains a large cube & a large cylinder

Unfortunately, visual explanations alone are not all we need either



Right for the right neuro-symbolic reasons!

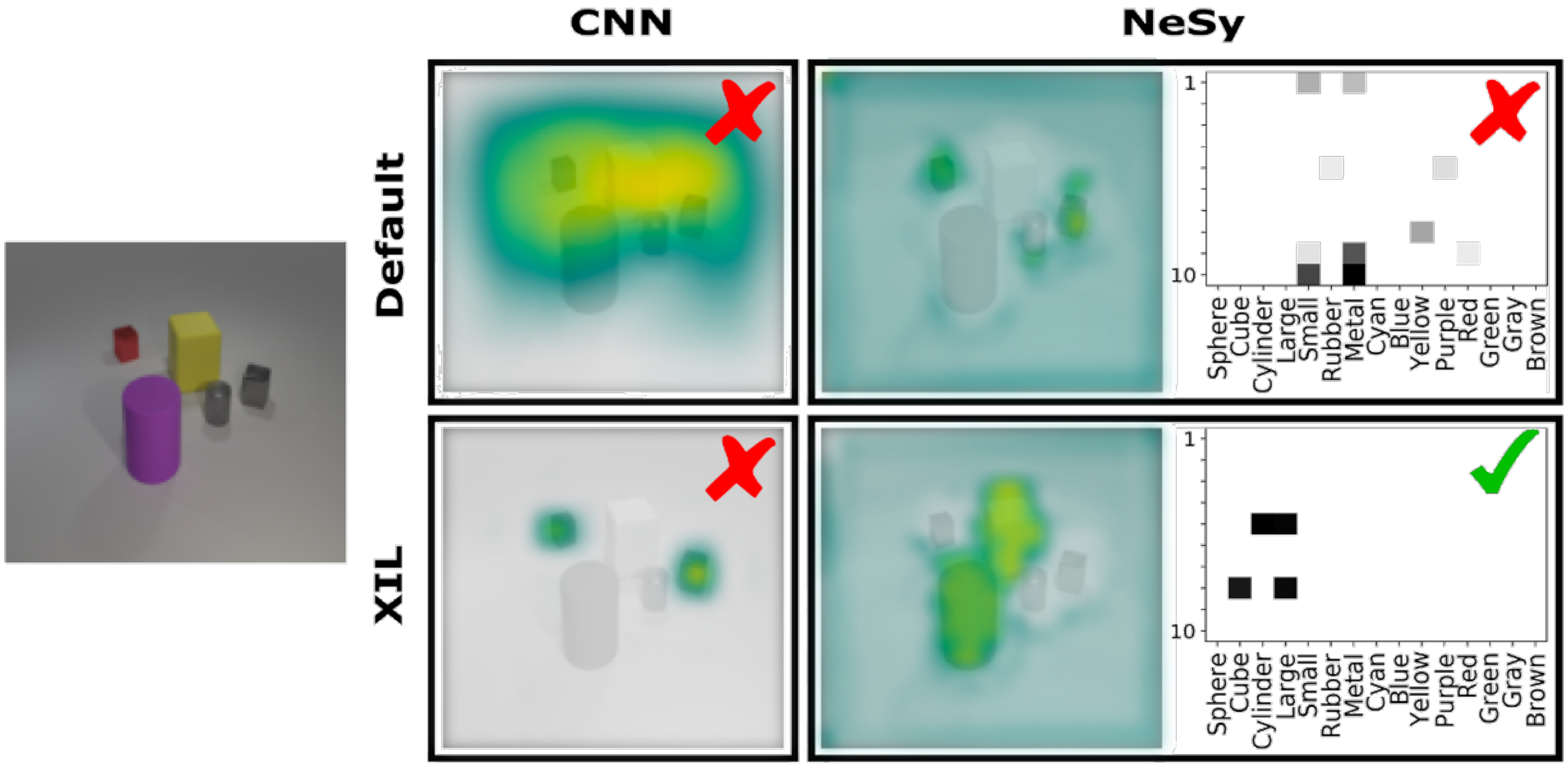


$$\lambda \sum_{i=1}^{N} r(A_i^v, \hat{e}_i^h) + (1-\lambda) \sum_{i=1}^{N} r(A_i^s, \hat{e}_i^g)$$

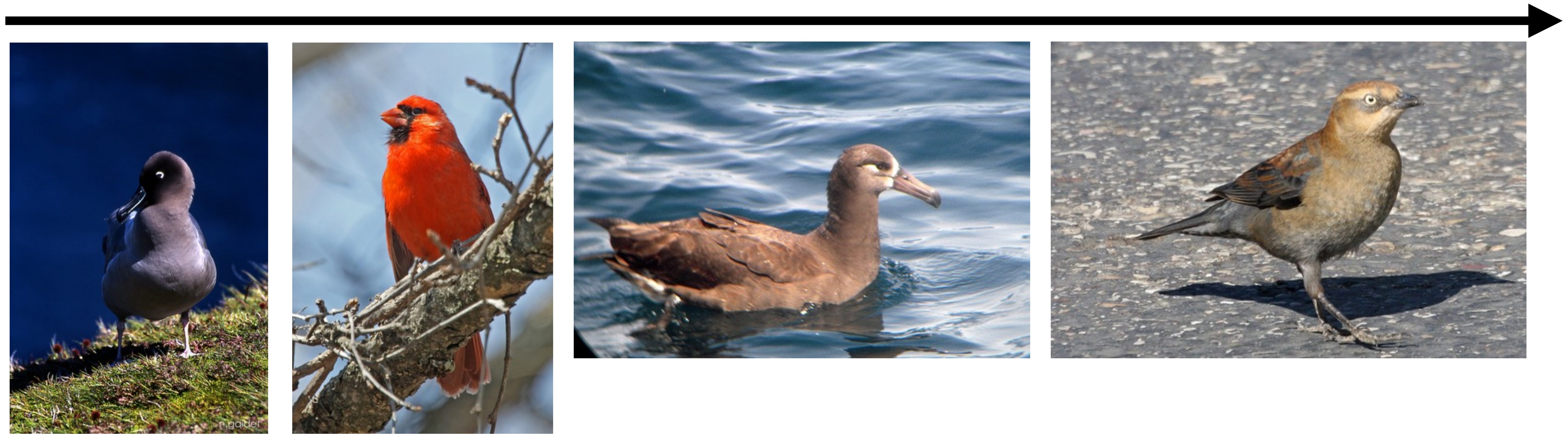
Combine human & machine intelligence via an explanatory loss term

Ross et al IJCAI 2017
Teso, Kersting AIES 2019
Selvaraju et al ICCV 2019
Schramowski et al Nature MI 2020

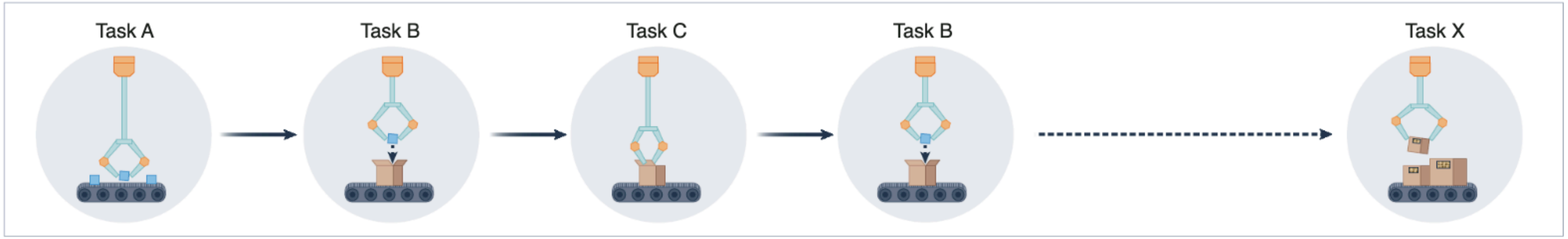
Symbolic feedback:
"Never base your decision on gray cubes!"



What if there is a new concept

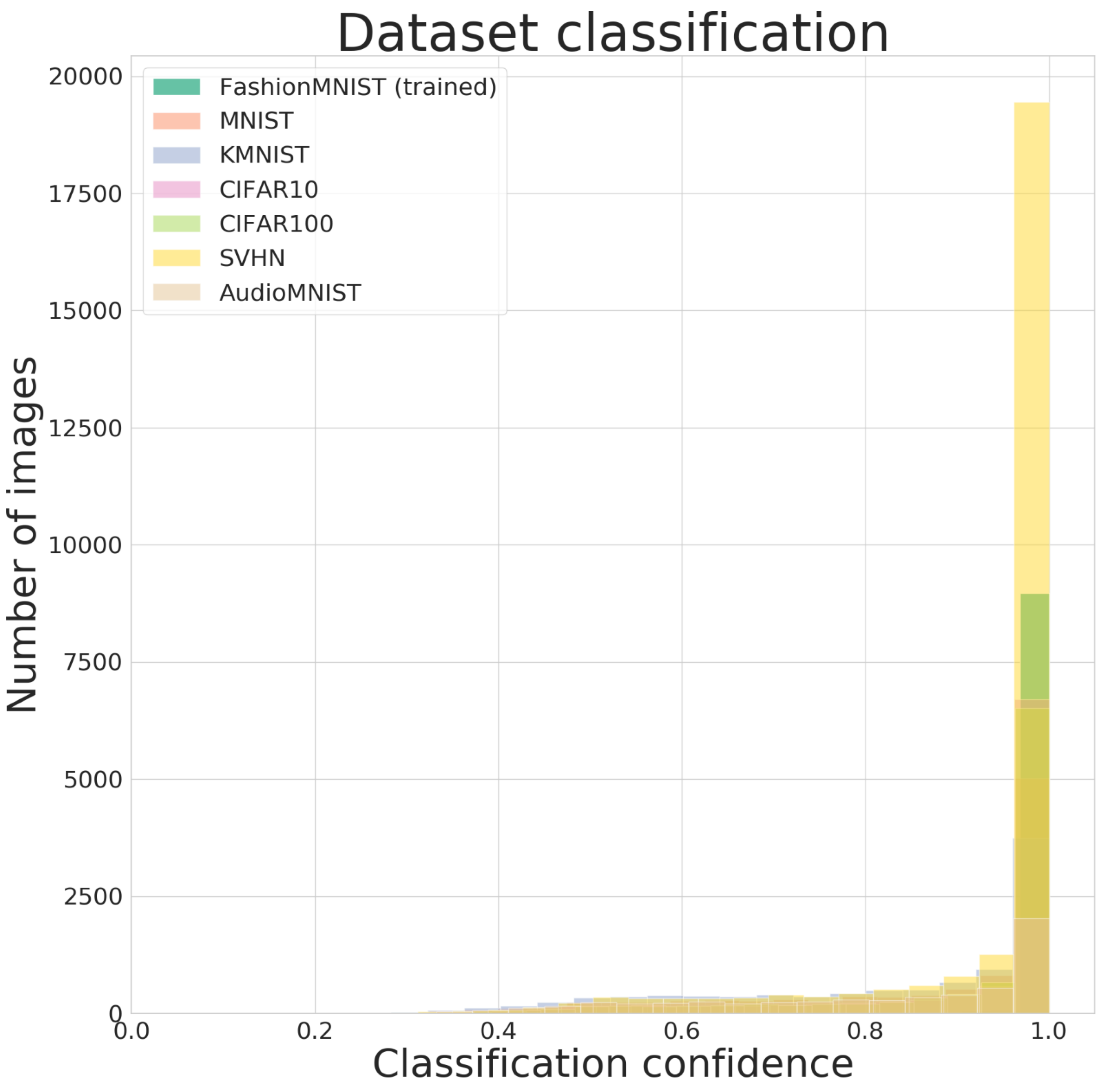


Or even a new task



Welinder et al, Caltech-UCSD Birds 200 *Kudithipudi et al, Nature MI 2022*

Another challenge: deep models **don’t know when they don’t know**

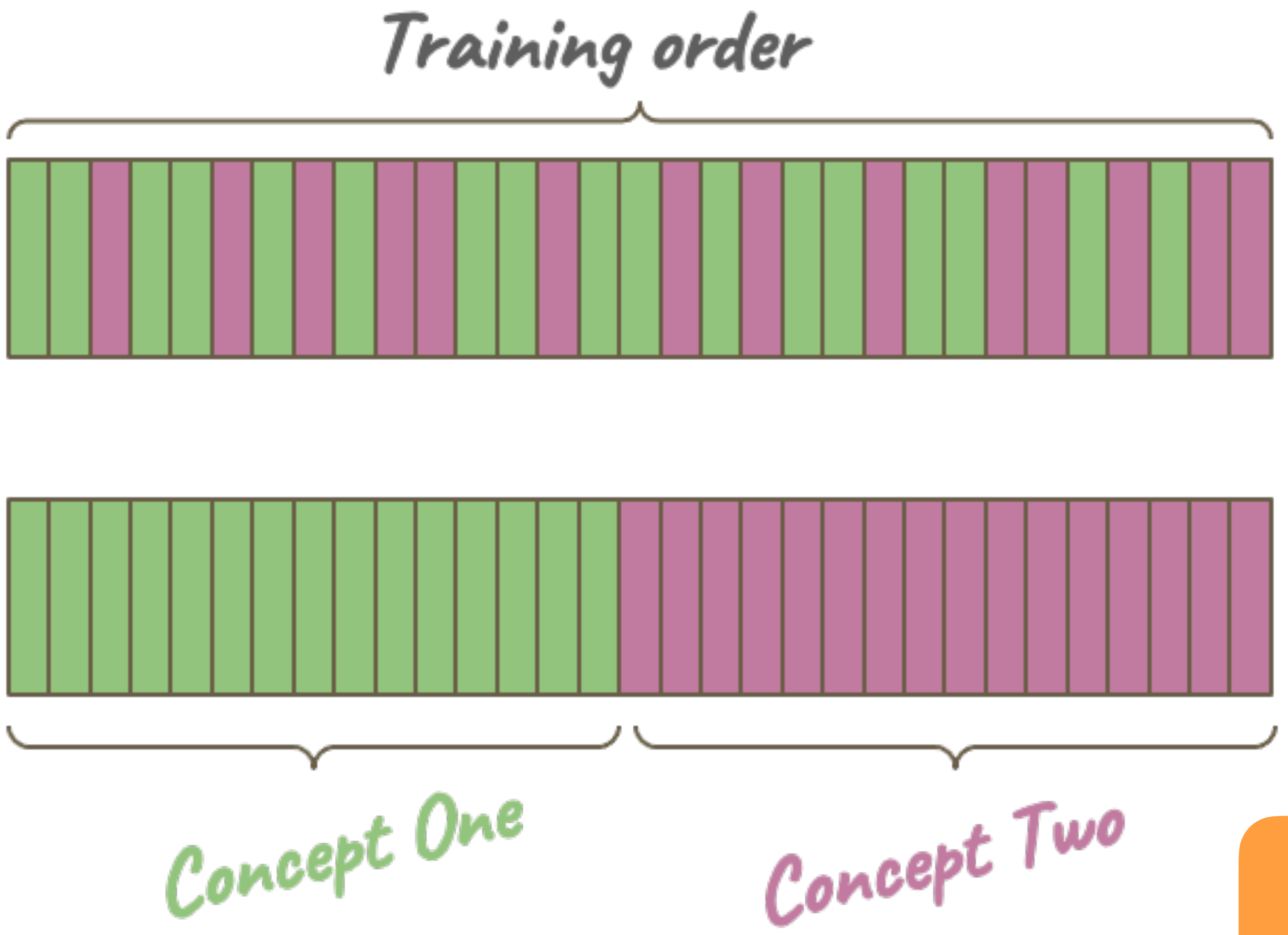


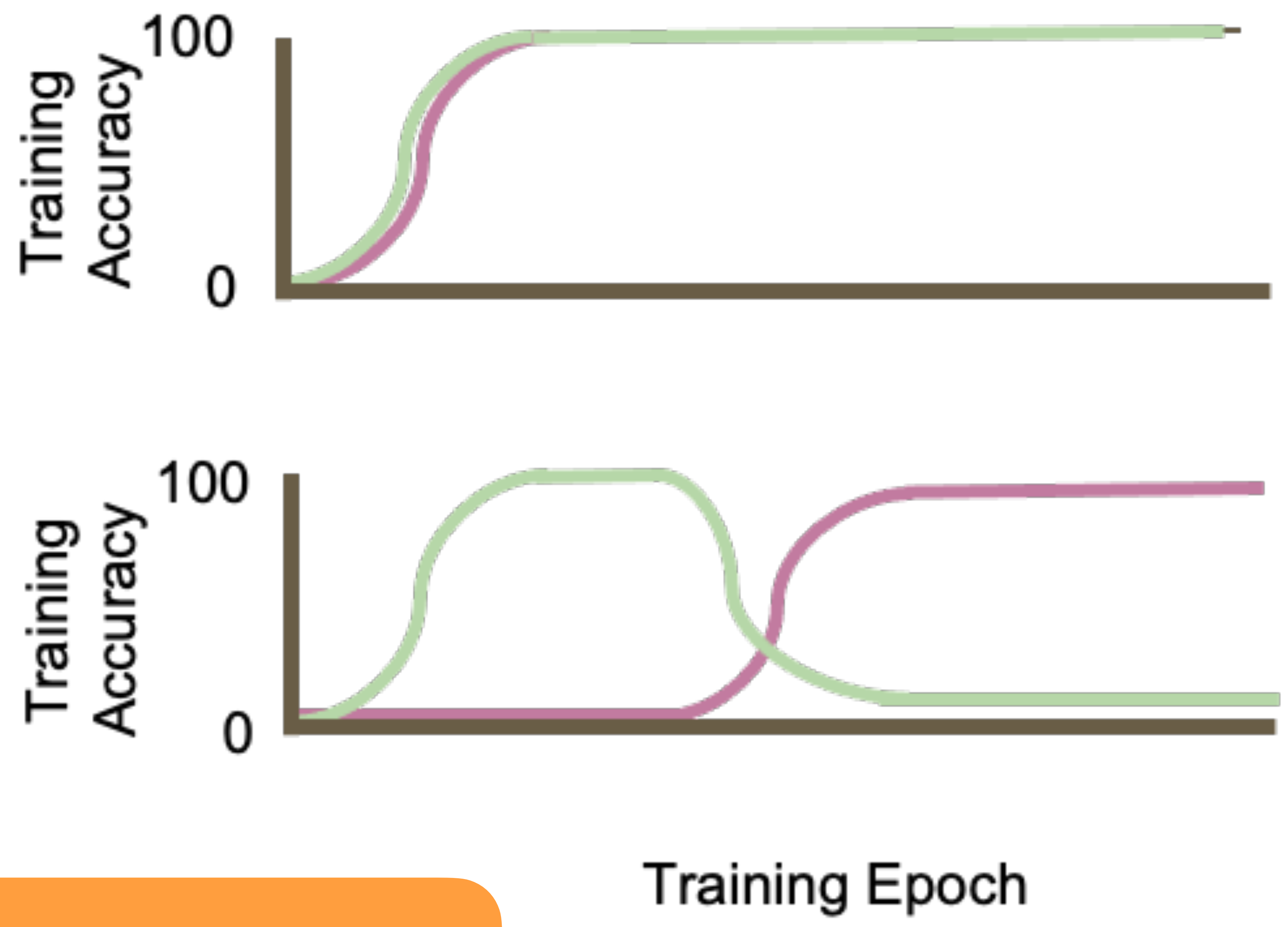
(Deep neural) models are **overconfident** on datasets they were never **trained on**

Matan 1990
Mundt et al. ICCV 2019
Mundt et al. Journal of Imaging 2022

...and even worse, neural nets fail to learn sequentially. Humans don't!

Blocked training

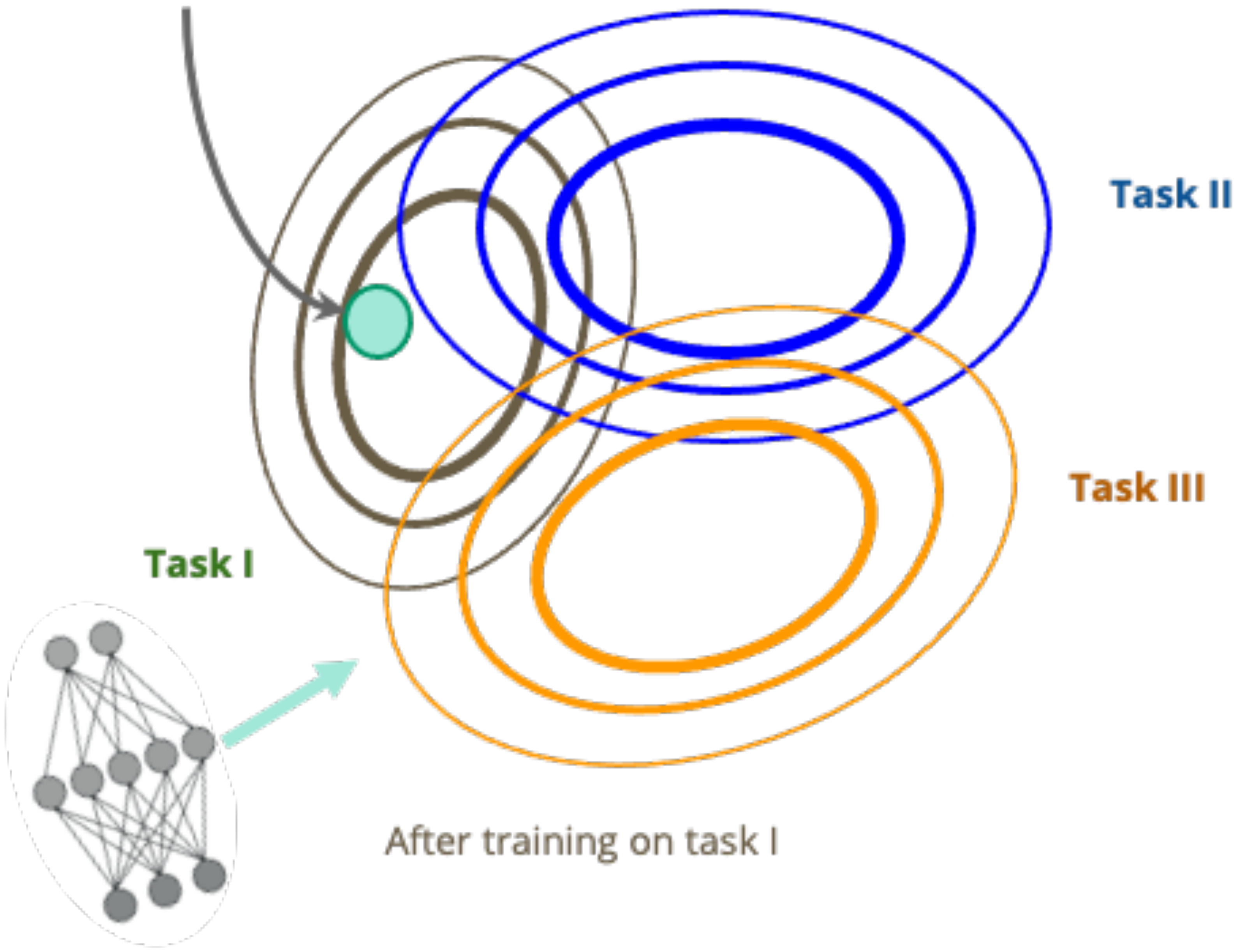




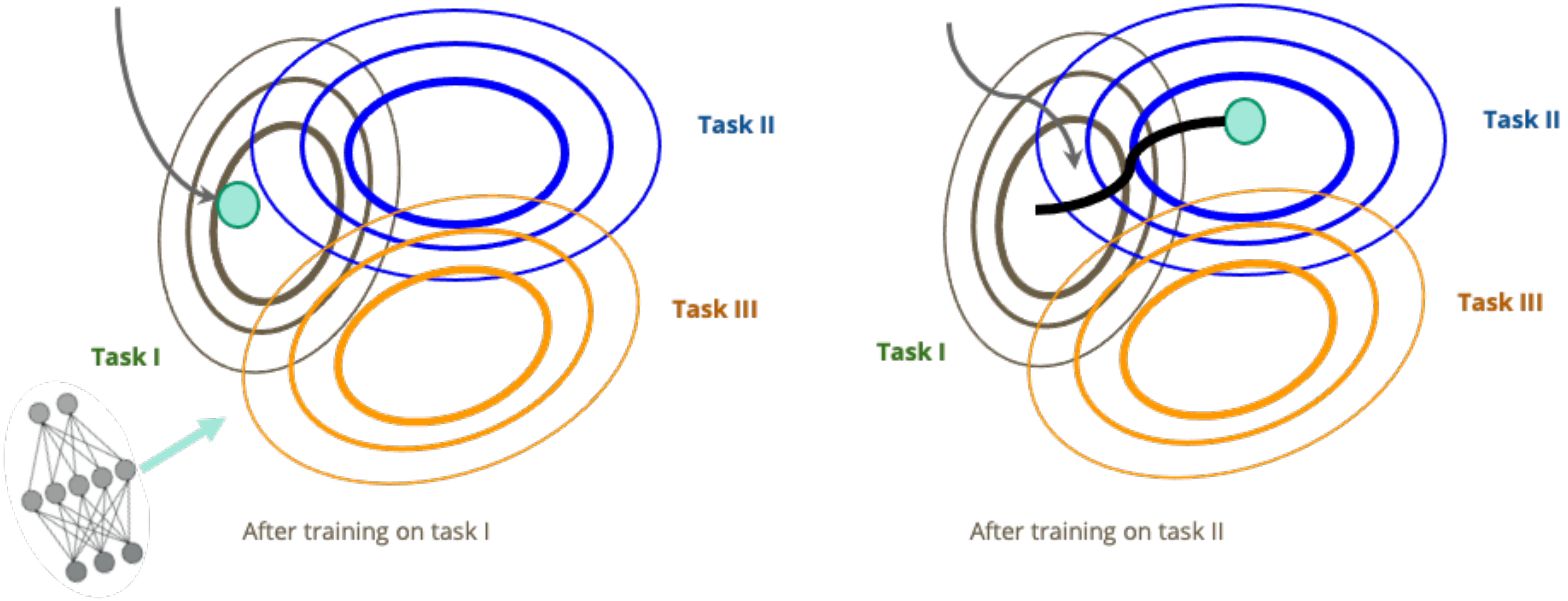
Catastrophic Interference (McCloskey & Cohen 89)

Adapted from Flesch et al 2022

Why might neural networks be so forgetful? Is it surprising?

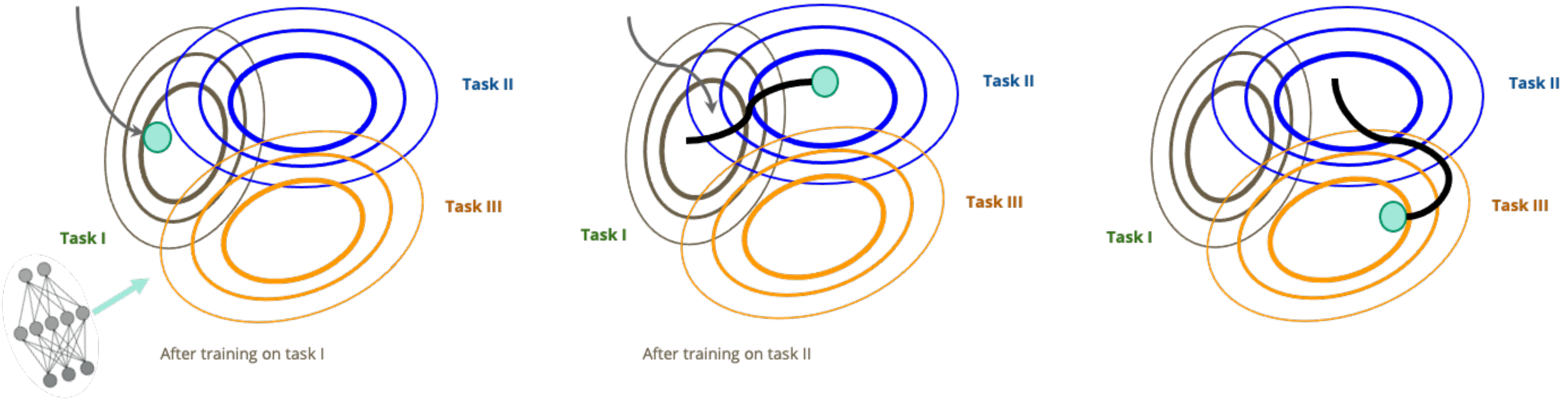


Why might neural networks be so forgetful? Is it surprising?



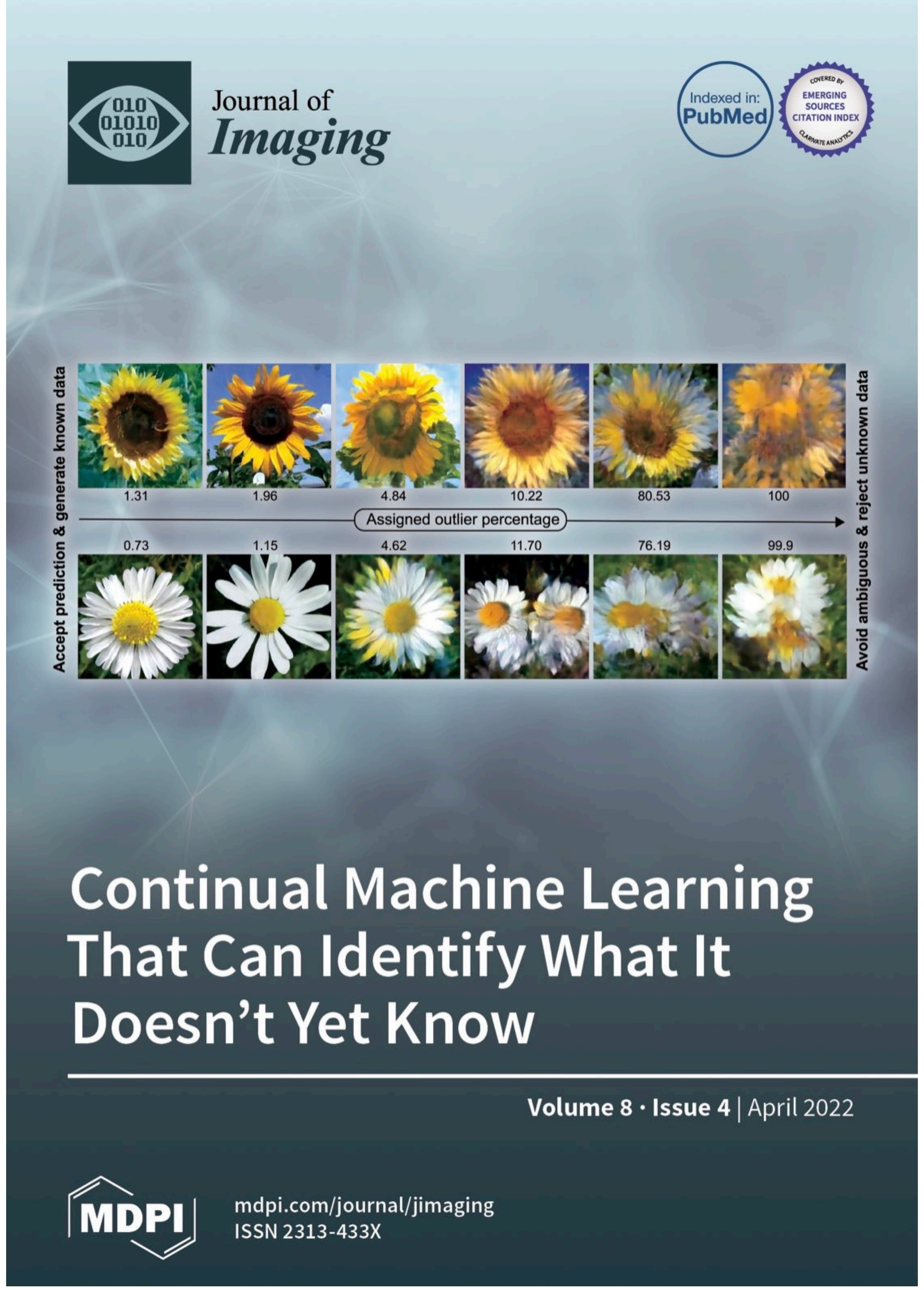
Slides adapted from our AAAI-23 Continual Causality Tutorial, Cooper & Mundt

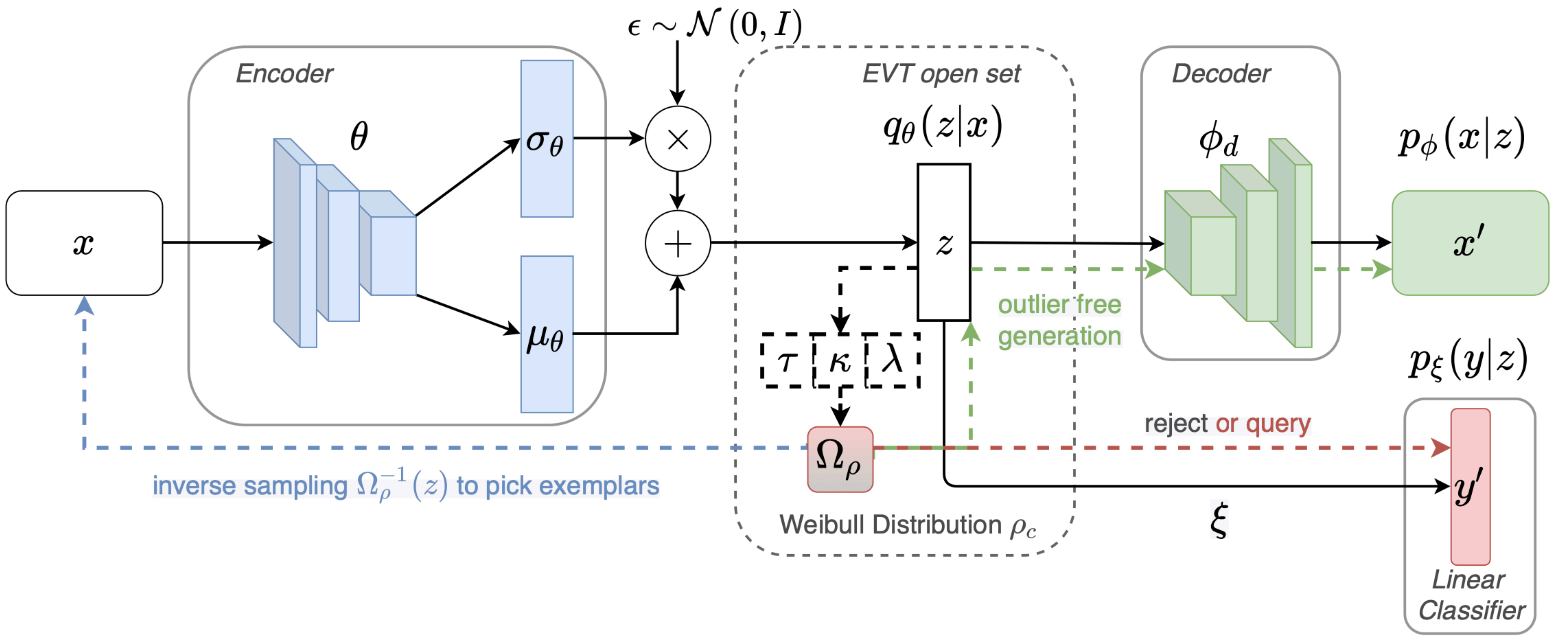
Why might neural networks be so forgetful? Is it surprising?



Slides adapted from our AAAI-23 Continual Causality Tutorial, Cooper & Mundt

We can mitigate this problem with generative models

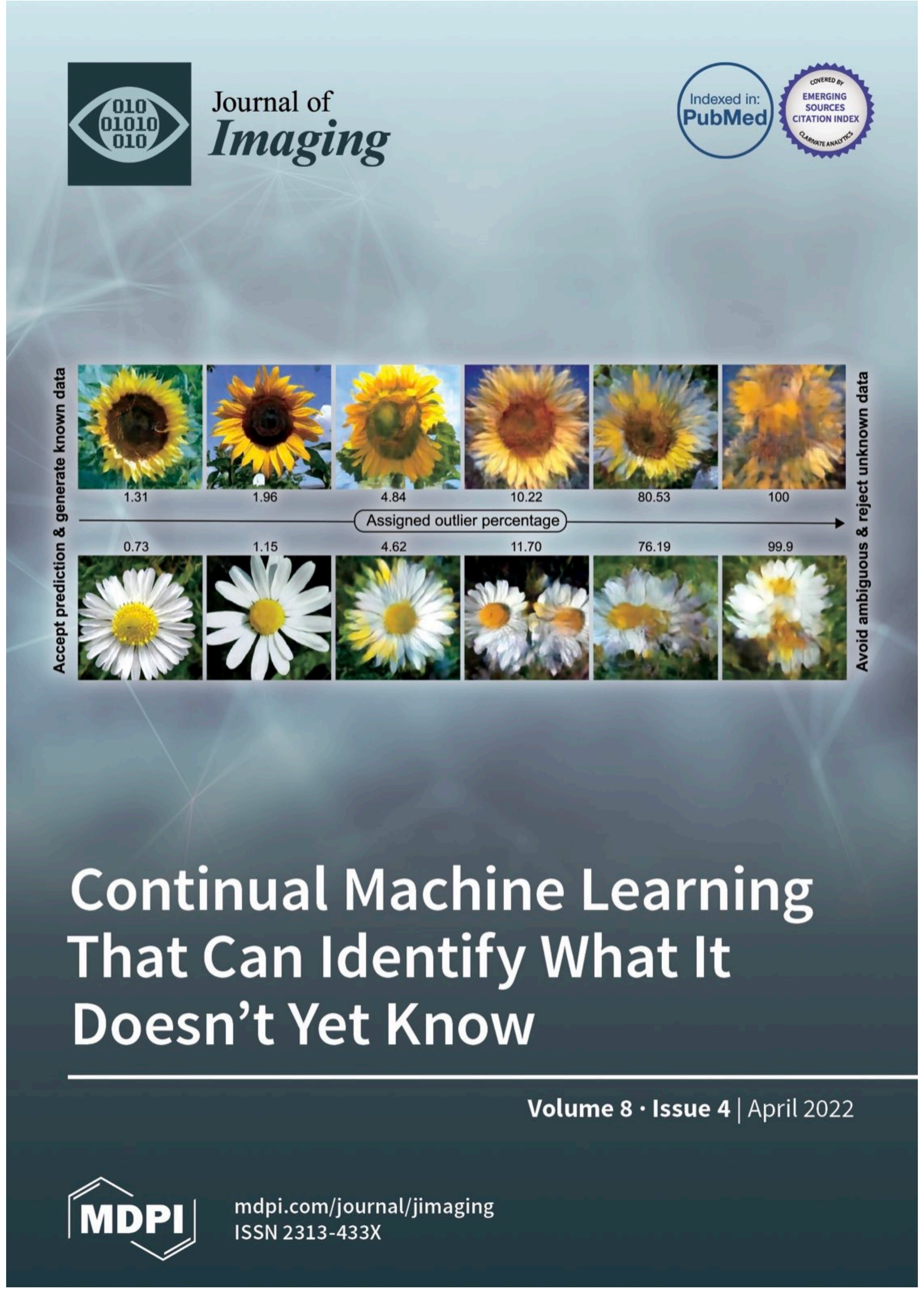


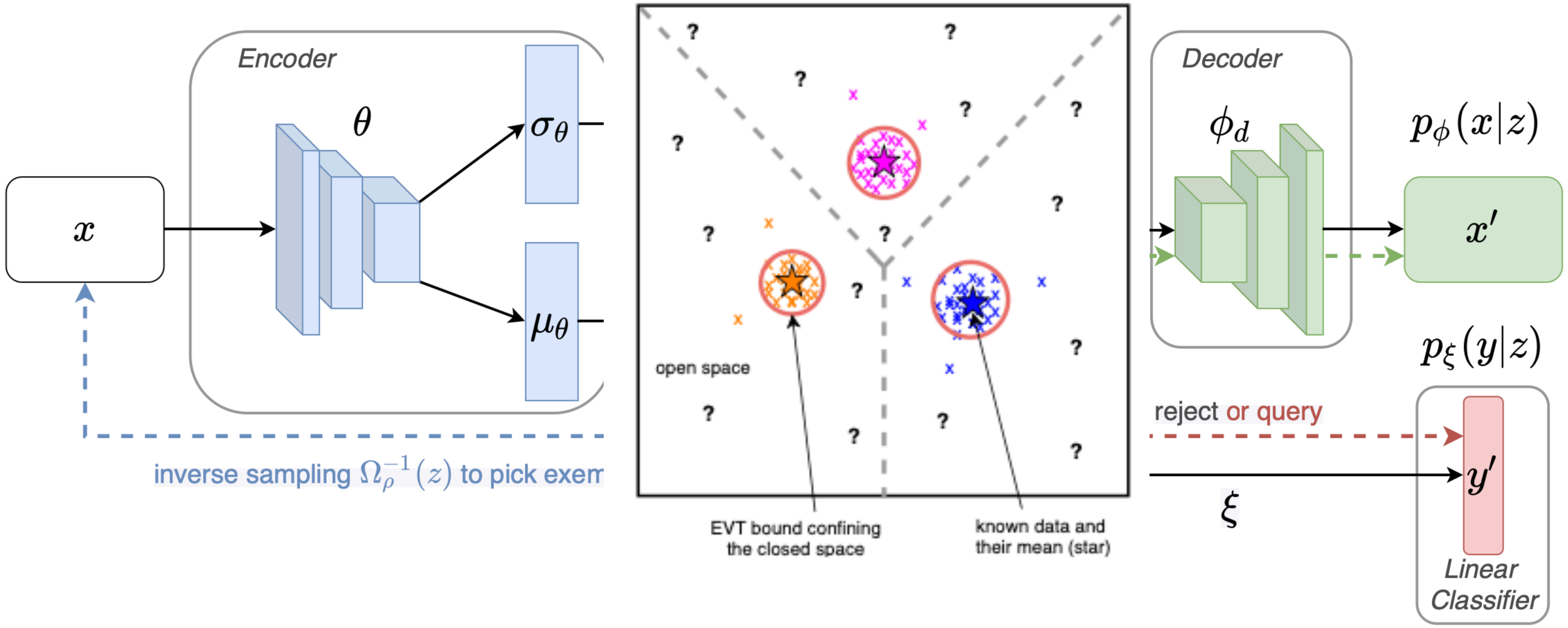


$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\xi}) = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{z}) + \log p_{\boldsymbol{\xi}}(\boldsymbol{y}|\boldsymbol{z})\right] - \beta KL(q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}))]$$

-> we learn how to encode data into generative factors & in turn how to decode (generate) these into data

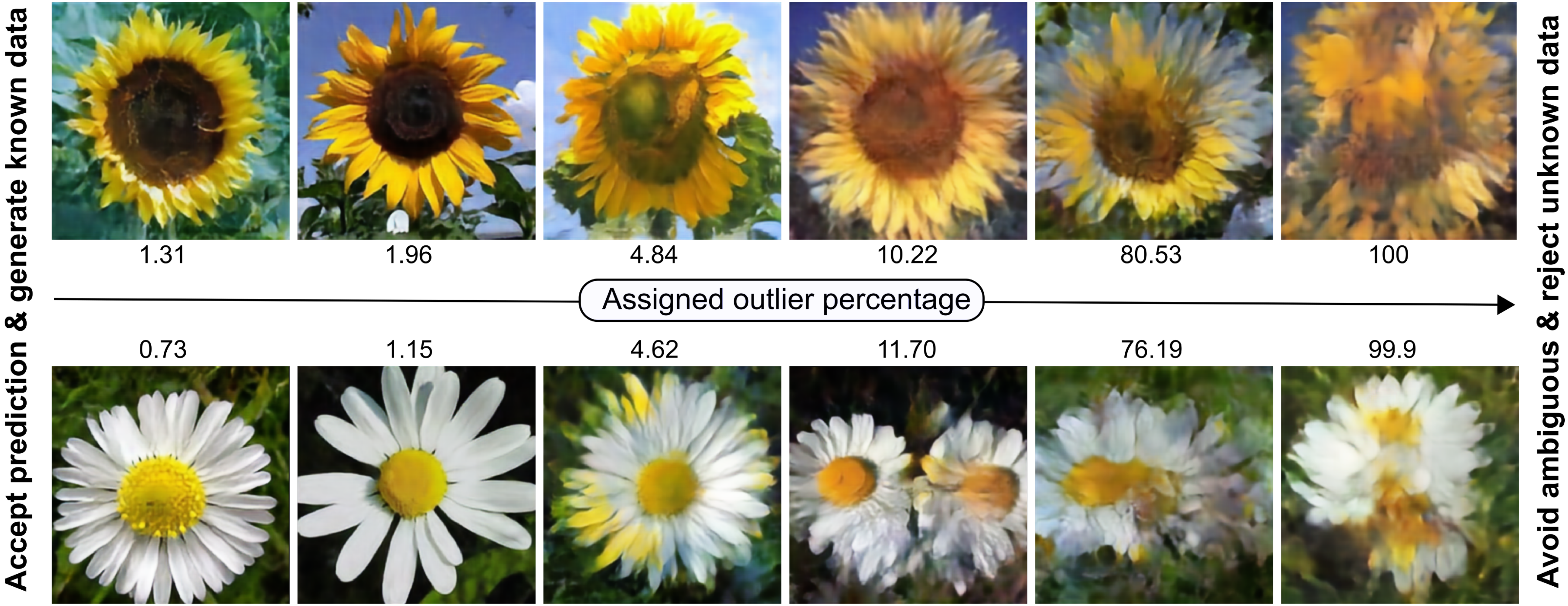
We can mitigate this problem with generative models





We can then measure similarity to factors we have already observed & replay knowledge from already seen ones

We can then measure similarity to factors we have already observed & replay knowledge from already seen ones



Mundt et al. Journal of Imaging 2022
Hong & Mundt Neural Networks 2022

So ultimately, the prevalent common ML pipeline is unrealistic

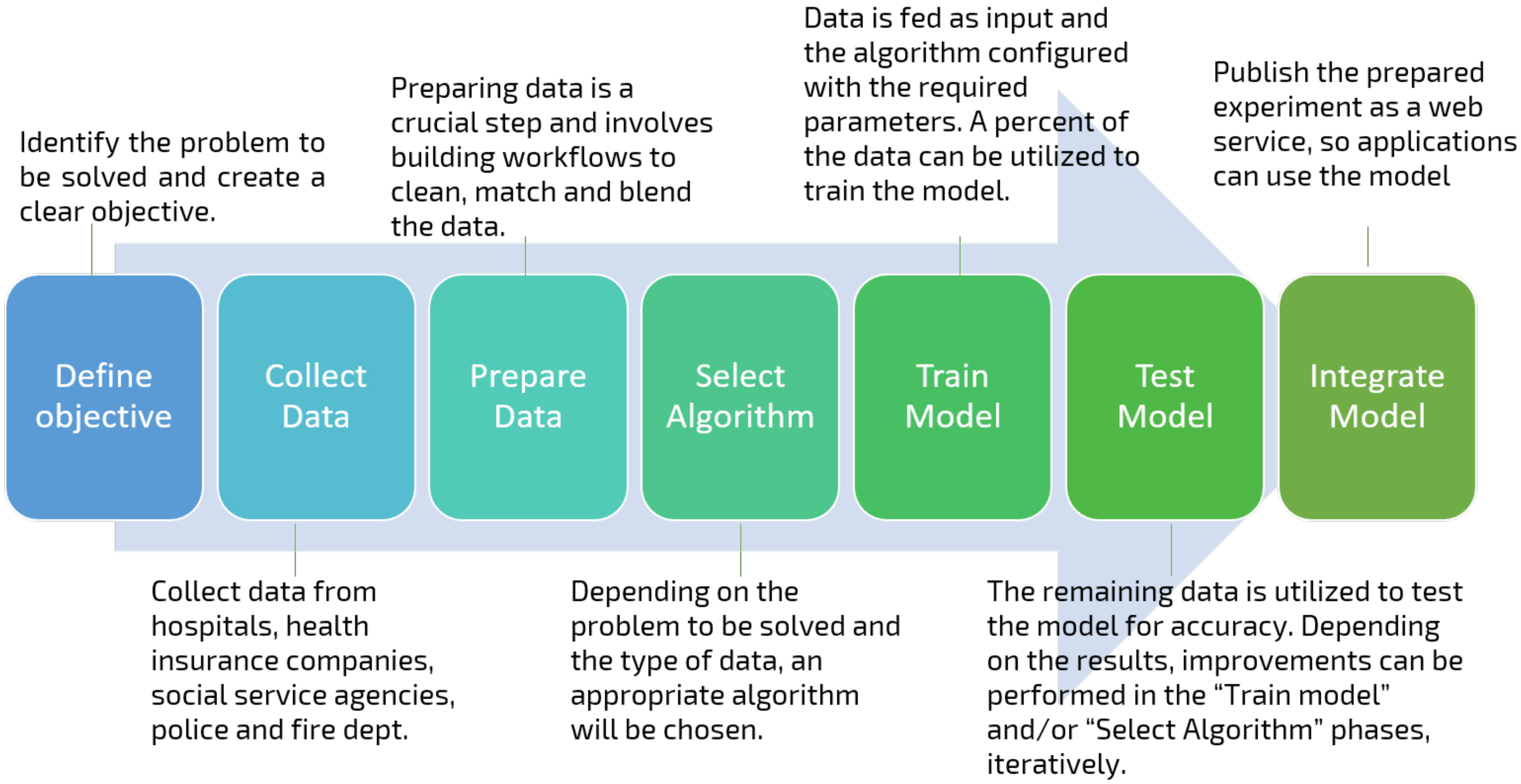
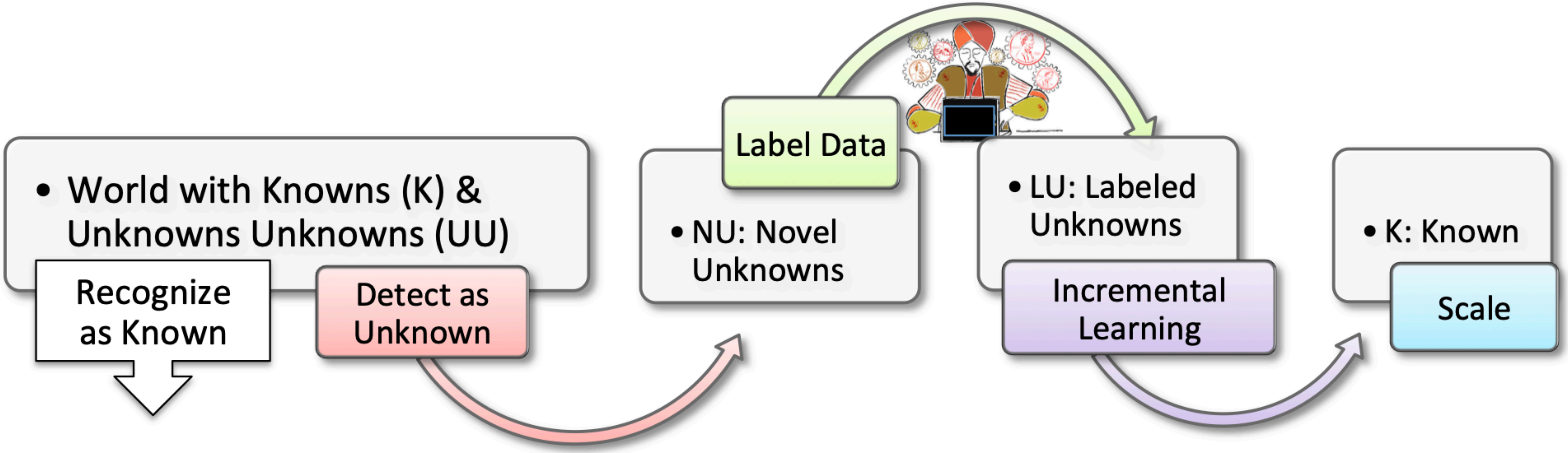


Figure from https://www.congrelate.com/get-workflow-machine-learning-images/

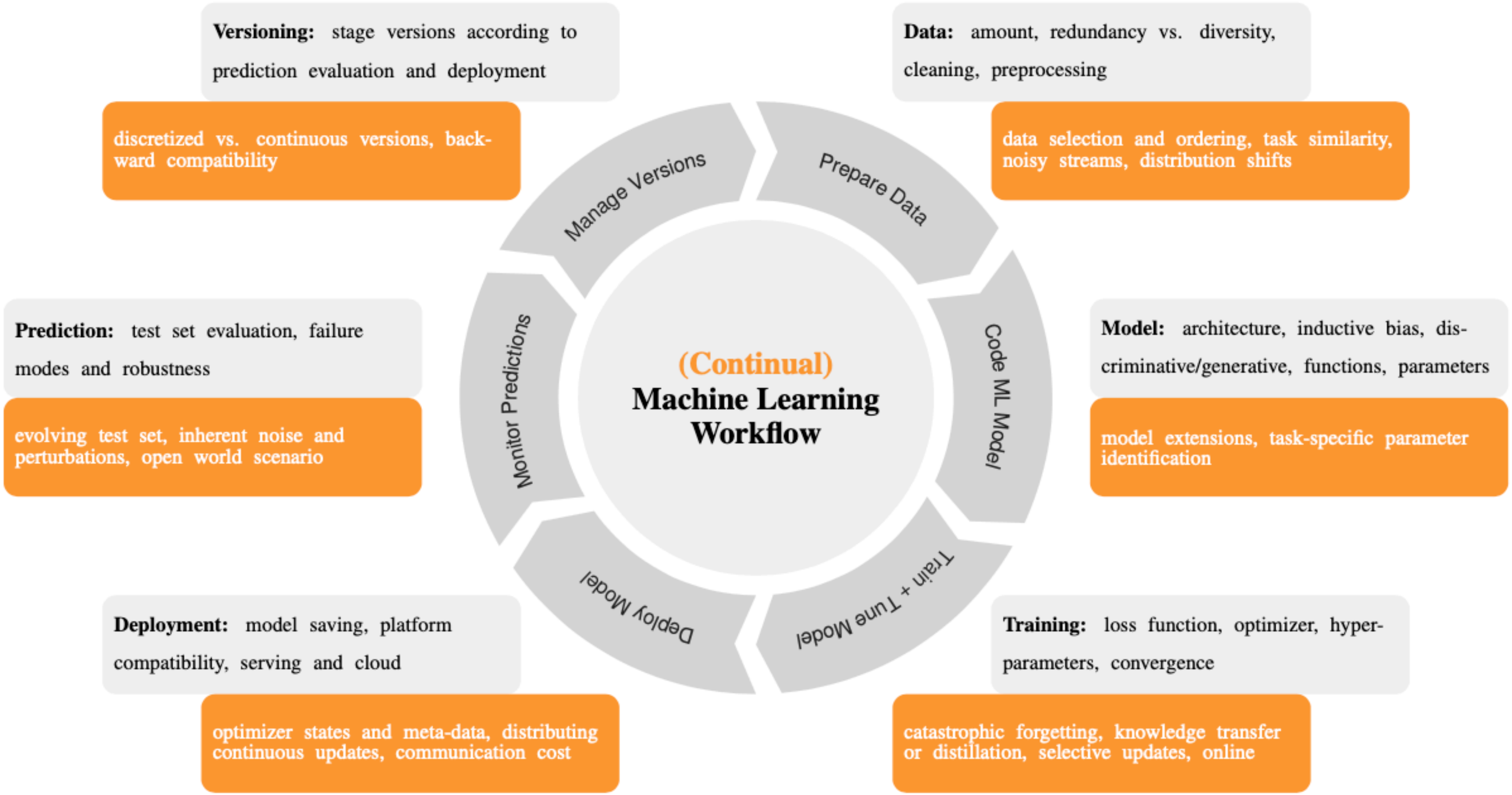
in reality it may look much more like this with sensor drifts, novel concepts in data, focus on wrong reasons, or even new tasks



And perhaps even further

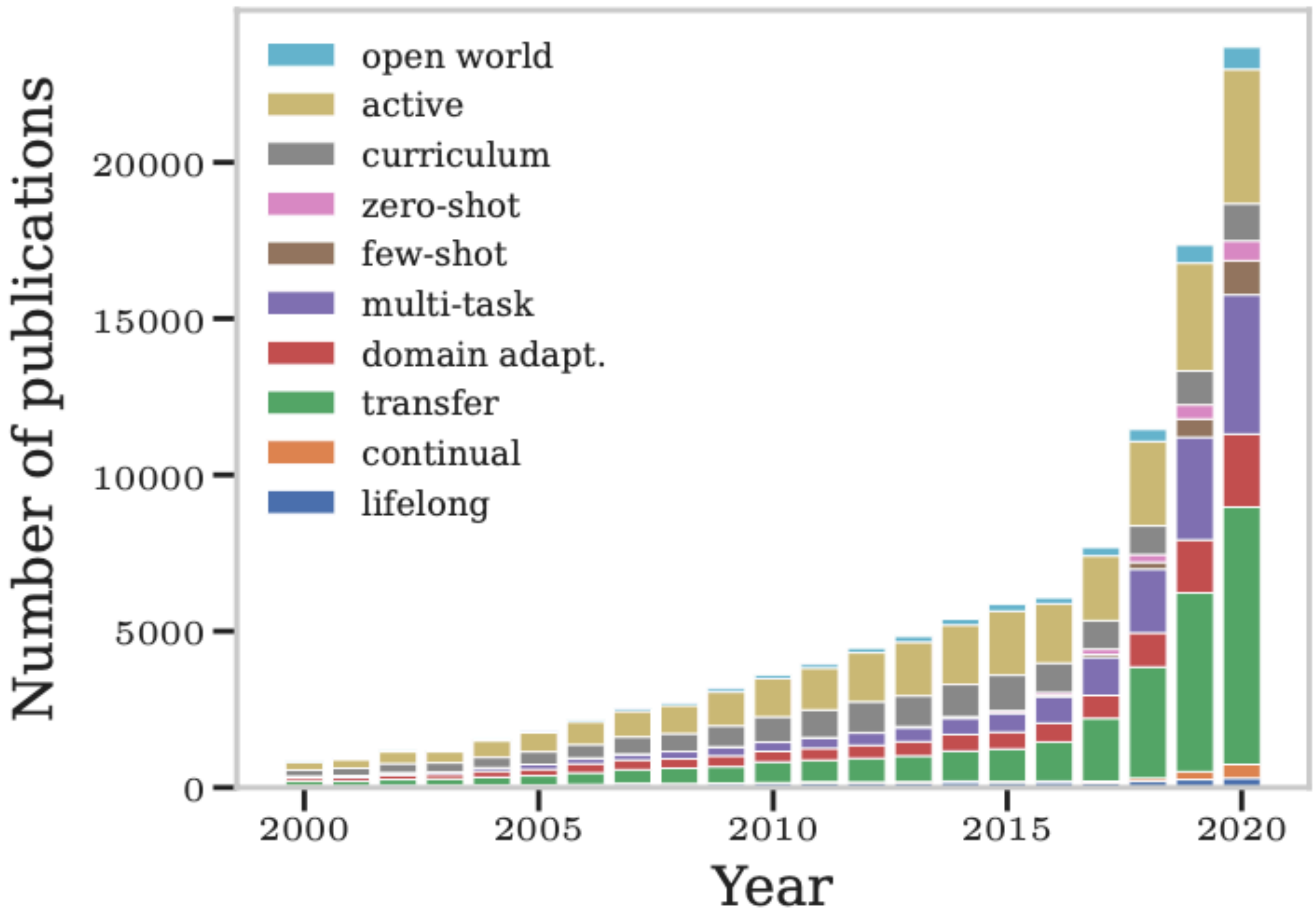
Bendale & Boult, CVPR 2015
Mundt et al, Neural Networks 2023

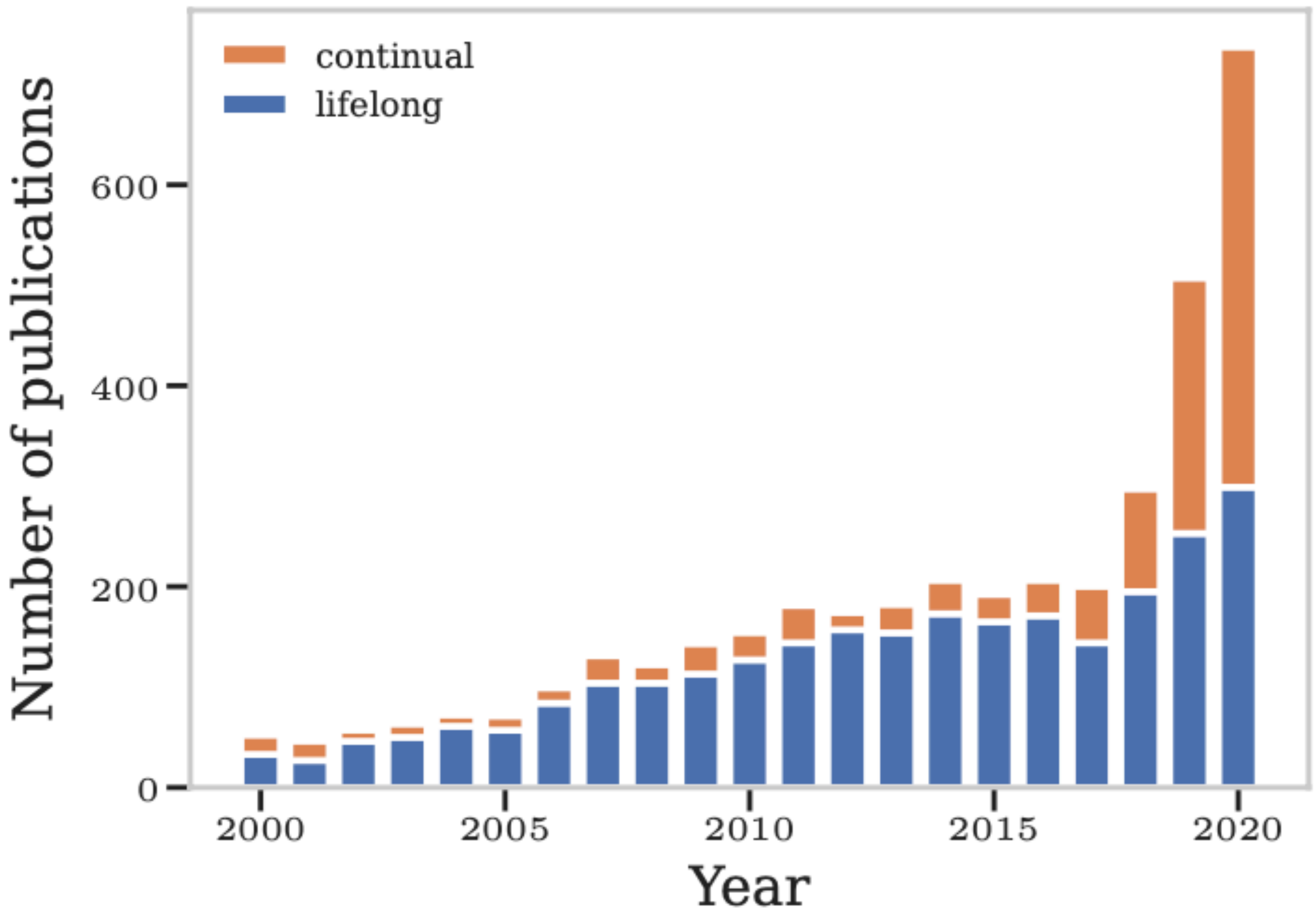
And in reality, likely much much more complex than six simple steps



Mundt et al, ICLR 2022

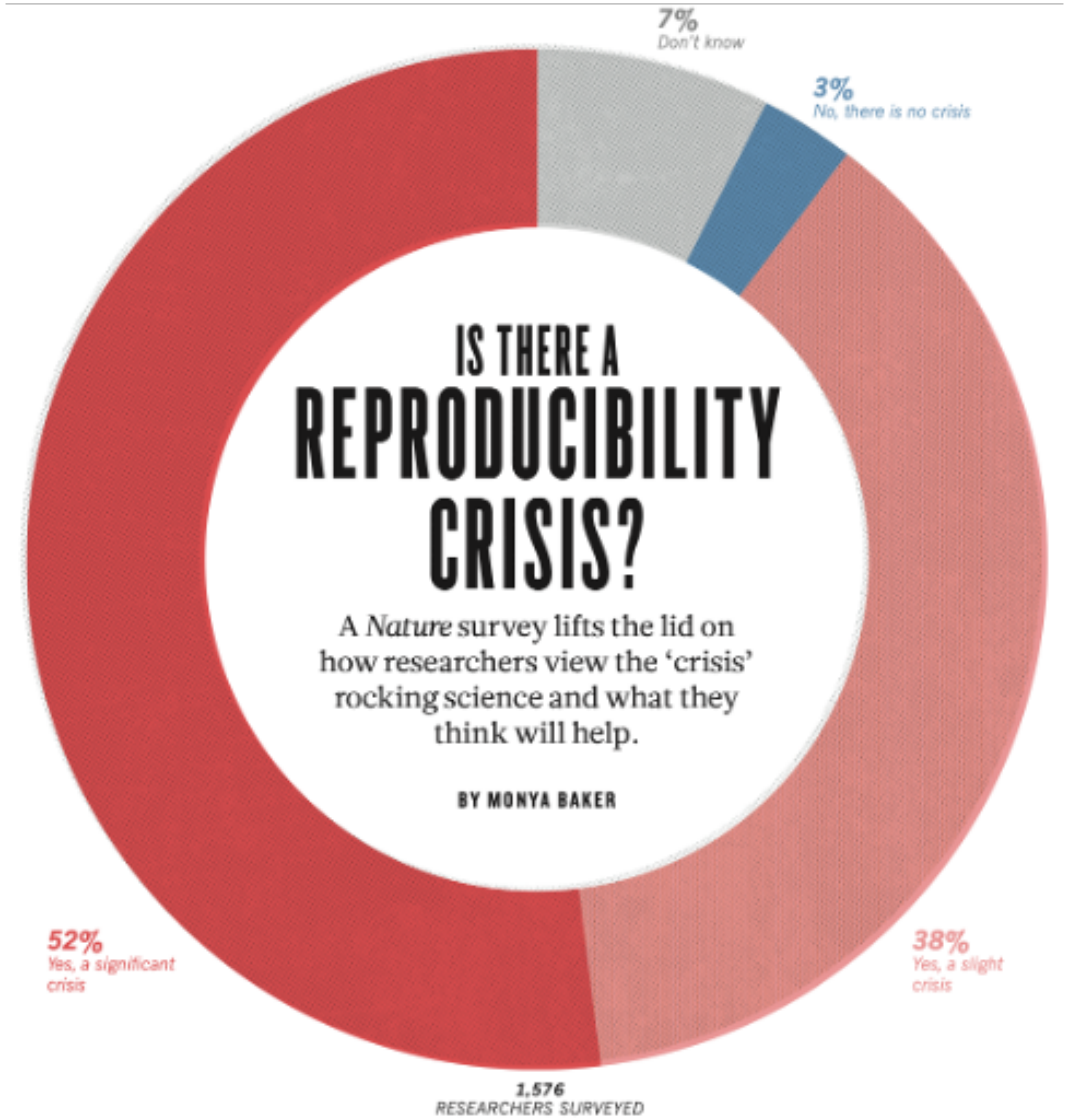
Not yet convinced? What if our application doesn’t require *all* of these factors?
Pragmatically: Why should we still care?

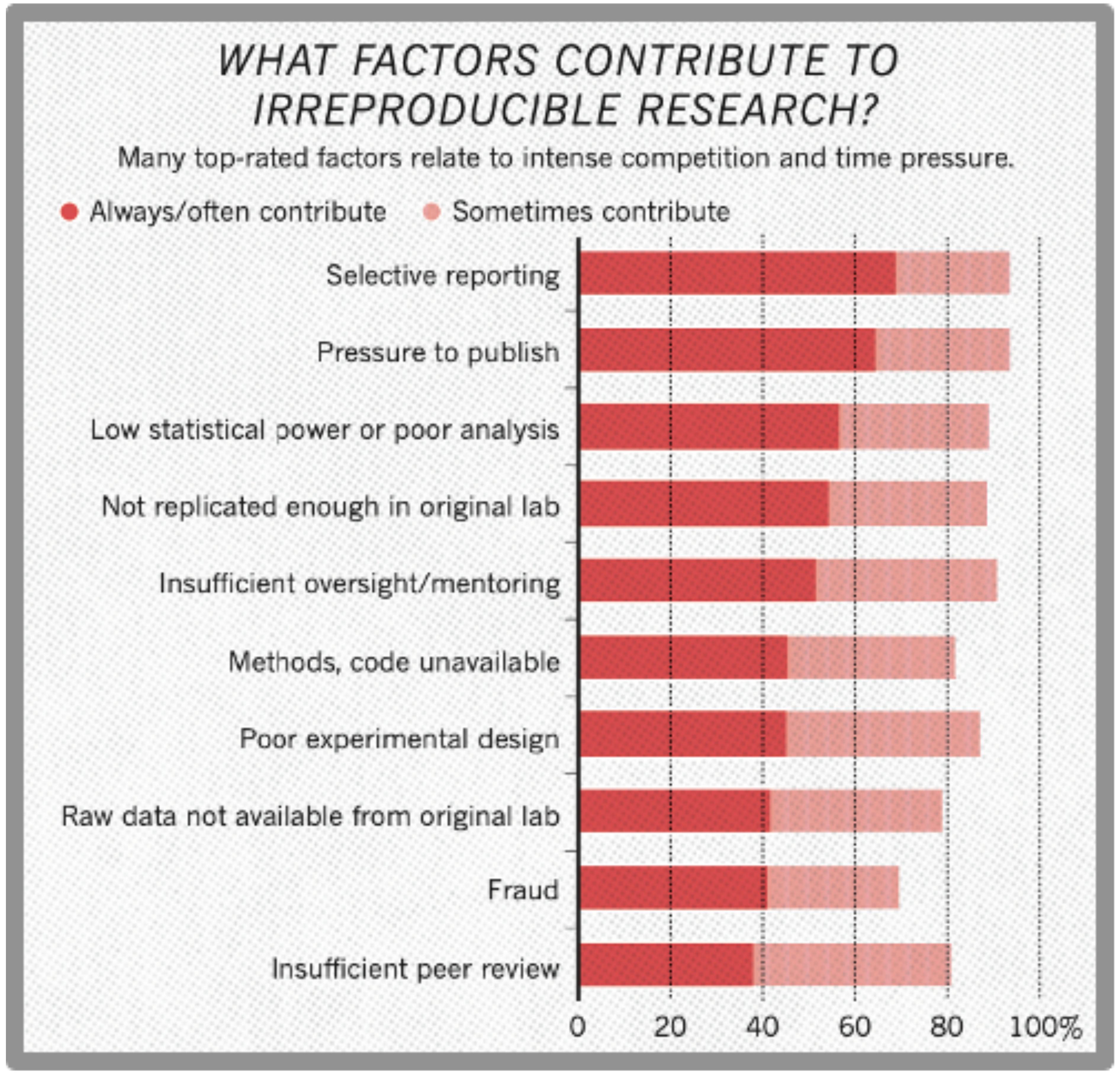




Mundt et al, ICLR 2022

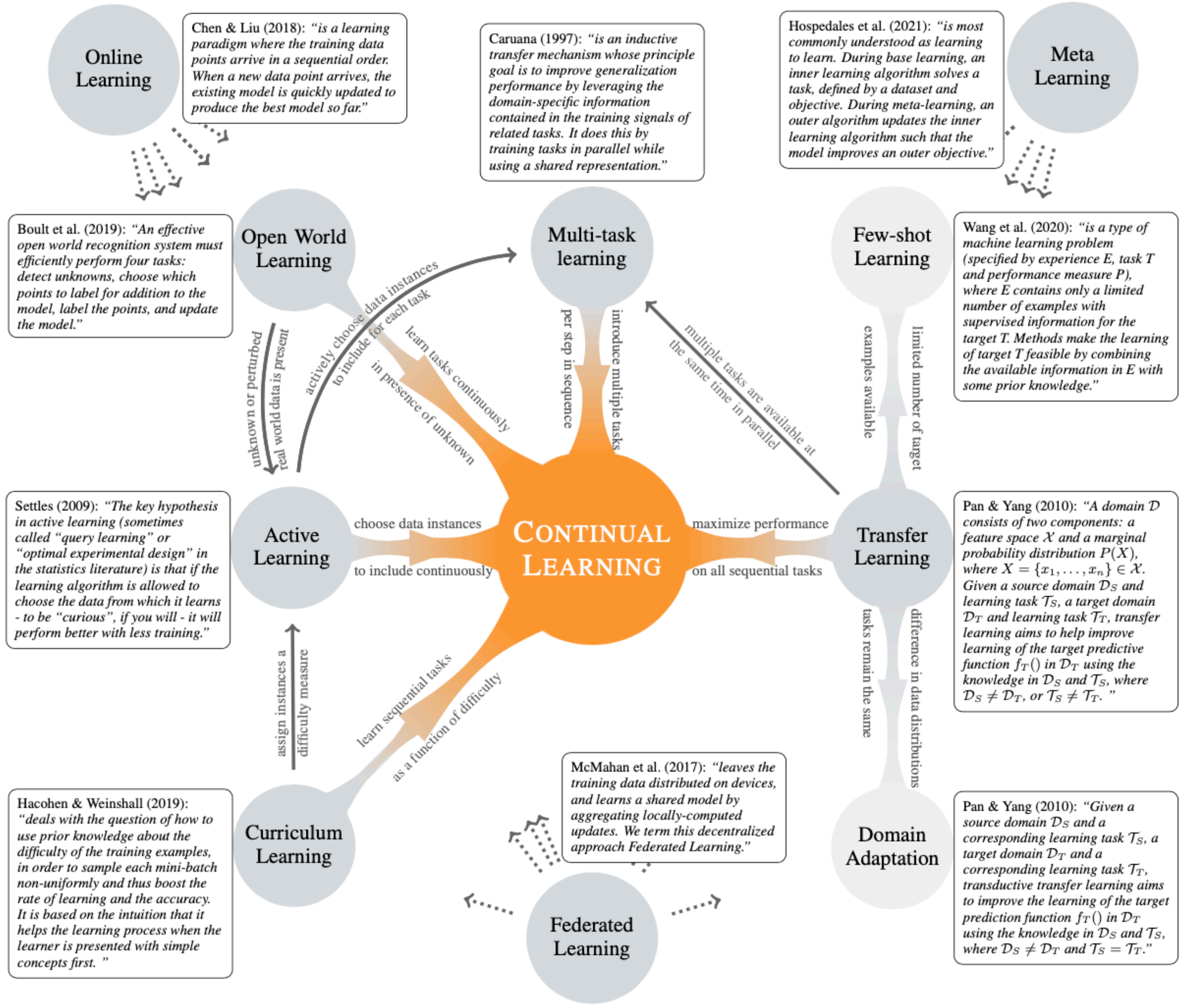
Not yet convinced? What if our application doesn't require *all* of these factors? Pragmatically: Why should we still care?

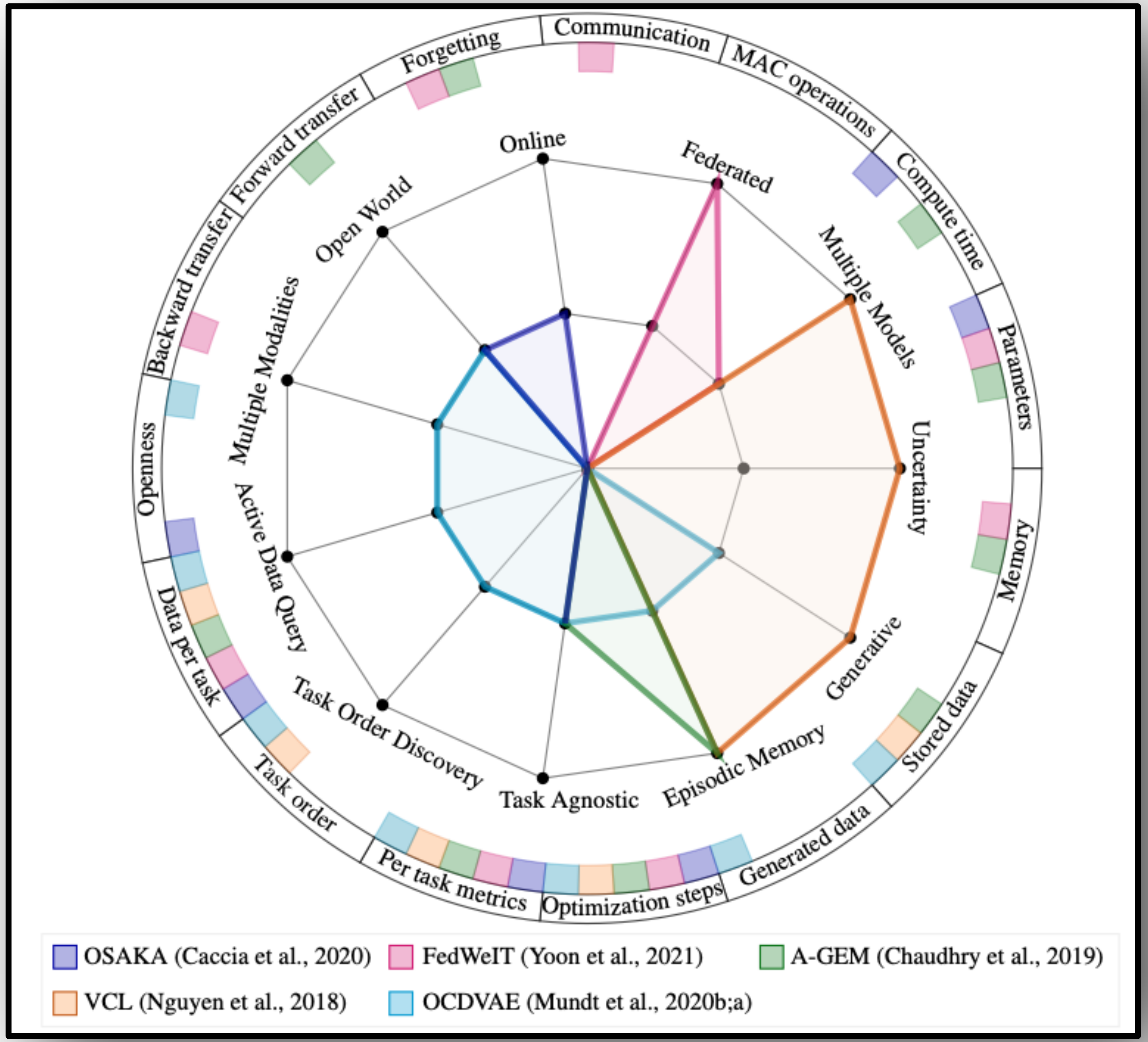




Baker, Nature 2016

Systems are complex! Assumptions & evaluation setups often collapse aspects into scalar measured quantities. Let’s acknowledge this fact & make it transparent





Mundt et al, ICLR 2022

Summary & take-aways

1. Standard deep neural networks are not **right for the right reasons**
2. Standard deep neural networks **don't know what they don't know**
3. Standard deep neural networks are **bad at learning sequentially/continually**

But very powerful -> **generative + symbolic + human**

With this we can enable explanatory interactive and continual learning & in the process make the machine learning workflow reflect our real-world desiderata more accurately than current static benchmarking

Thanks to … and many more!

Visvanathan Ramesh
Uni Frankfurt, hessian.AI

Kristian Kersting
TU Darmstadt, hessian.AI

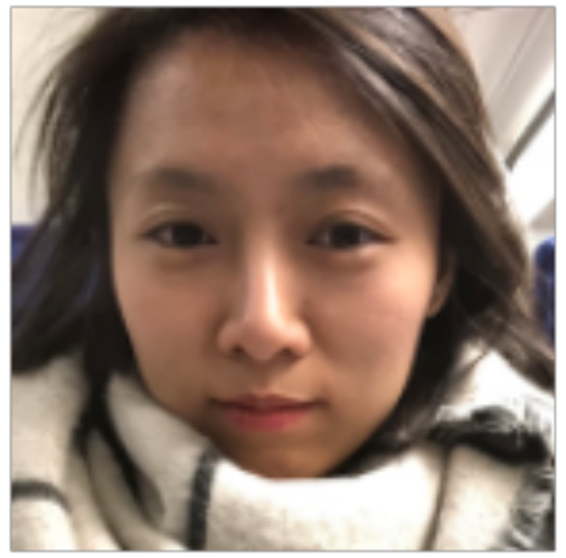
Iuliia Pliushch
Uni Frankfurt

Yongwon Hong,
Yonsei University

Vincenzo Lomonaco
Uni Pisa, ContinualAI

Patrick Schramowski
TU Darmstadt, DFKI

Wolfgang Stammer
TU Darmstadt



Xiaoting Shao
TU Darmstadt, Evonik

Steven Braun
TU Darmstadt

Tyler Hayes
NAVER, ContinualAI

TECHNISCHE UNIVERSITÄT DARMSTADT

hessian.AI

OWLL

ContinualAI