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

Dr. Martin Mundt TUDa & hessian.Al - Junior Research Group Leader **ContinualAl Board Member**







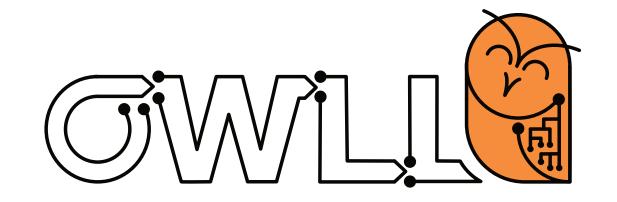




http://owll-lab.com



Prof. Dr. Kristian Kersting







We could talk about AI applications...



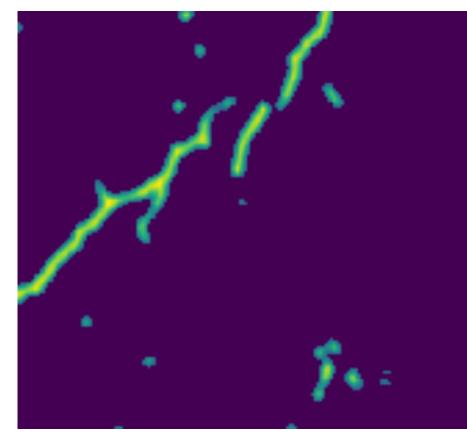


Fly drone

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



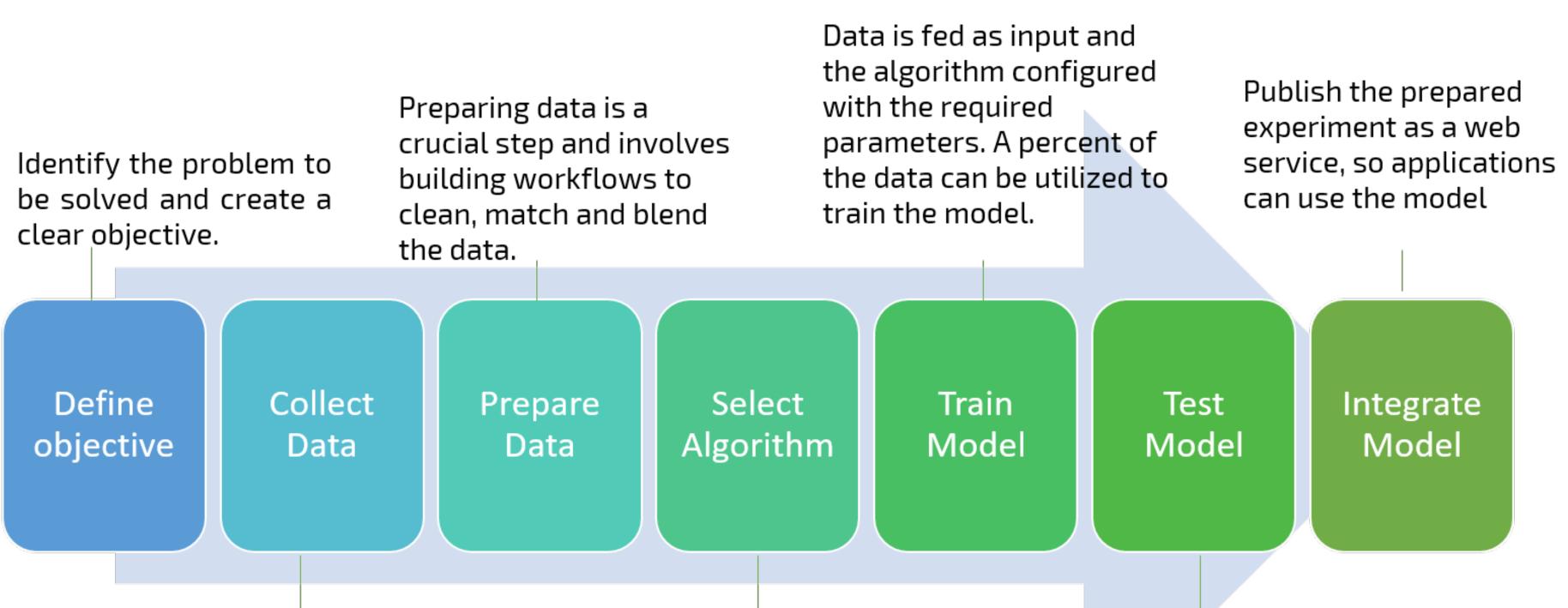








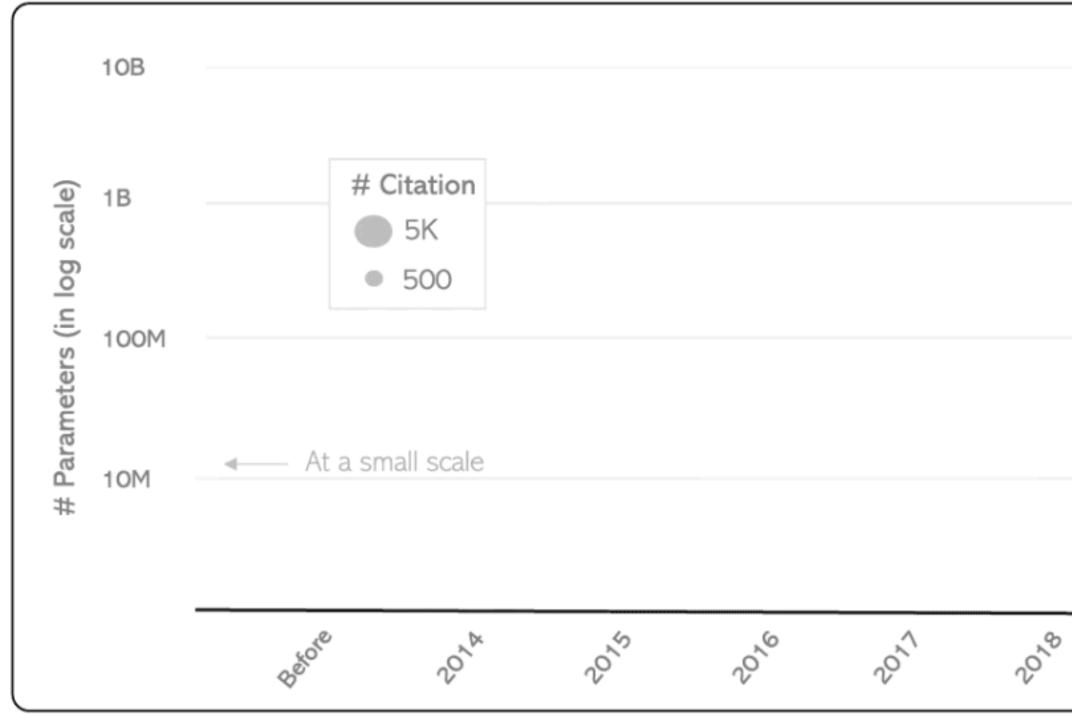
The standard machine learning workflow



Collect data from hospitals, health insurance companies, social service agencies, police and fire dept. Depending on the problem to be solved and the type of data, an appropriate algorithm will be chosen. The remaining data is utilized to test the model for accuracy. Depending on the results, improvements can be performed in the "Train model" and/or "Select Algorithm" phases, iteratively.



Is a static machine learning workflow + <u>scale</u> 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/

At a large scale ——	_
At a large scale	
.9 .0	-
2020 2020	

(AI) Research Director at Deepmind says all we need now is scaling

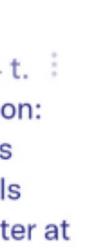


Nando de Freitas 📰 @Nando... · 4 t. Someone's opinion article. My opinion: It's all about scale now! The Game is Over! It's about making these models bigger, safer, compute efficient, faster at sampling, smarter memory, more modalities, INNOVATIVE DATA, on/ offline, ... 1/N

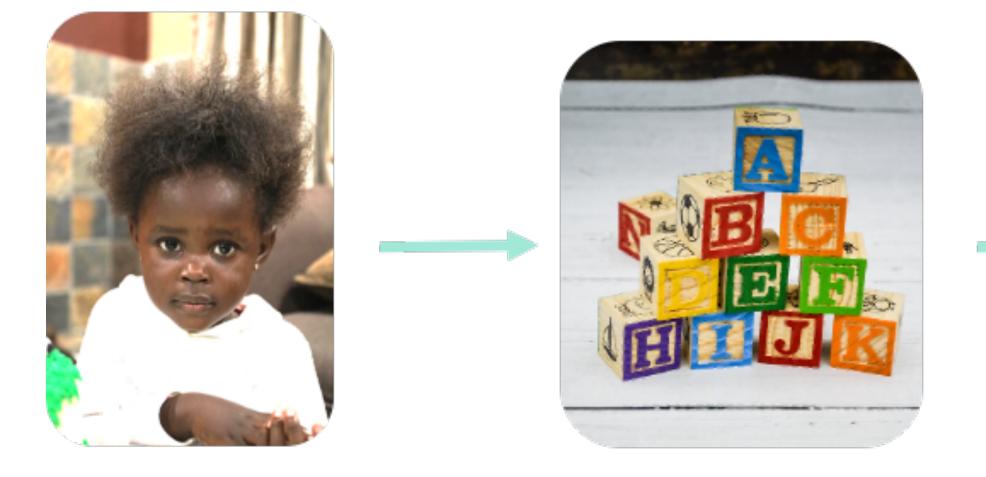


Q 10 1, 22 ♡ 78

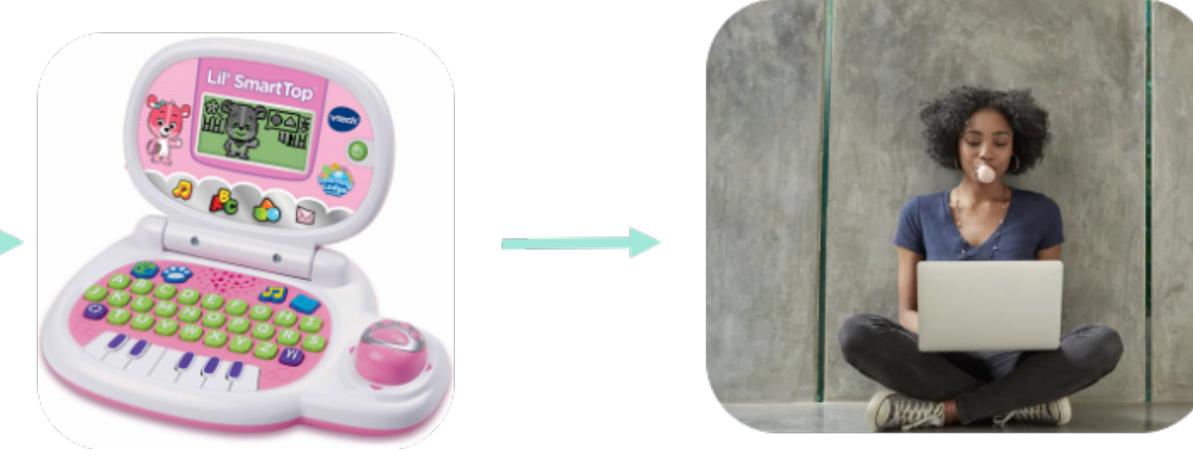




We have "foundation" models now, but humans learn & reason



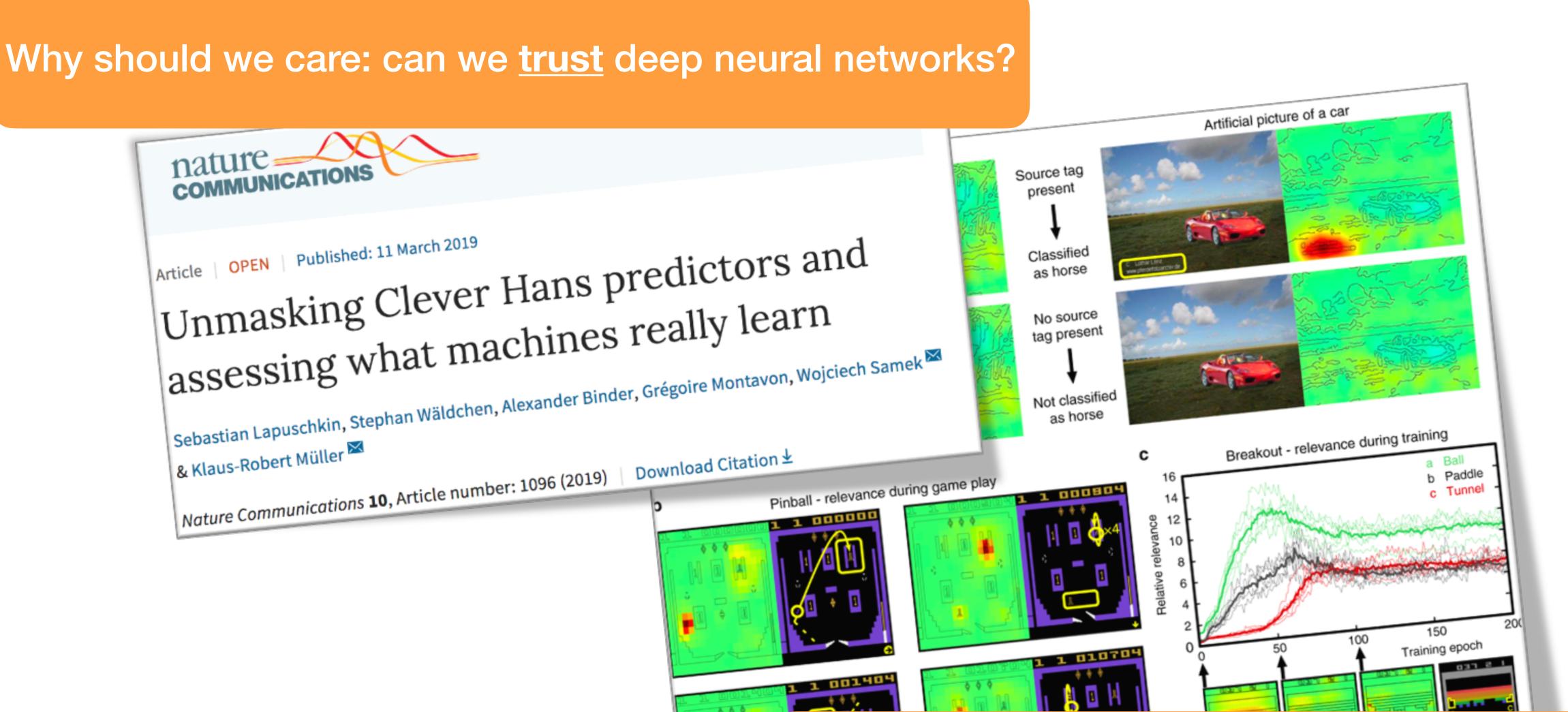
Importantly, humans revise their knowledge & continue adapting







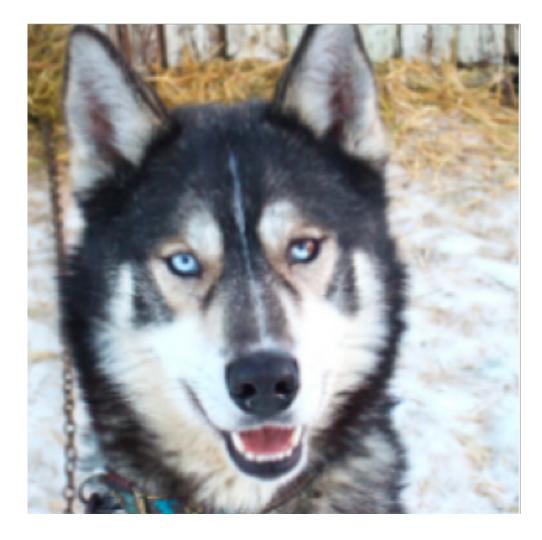




XAI suggests ways to <u>detect</u>, but not <u>fix</u> the issue!

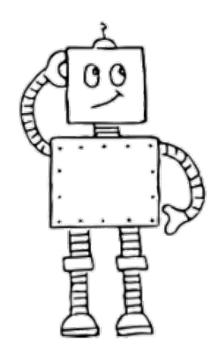






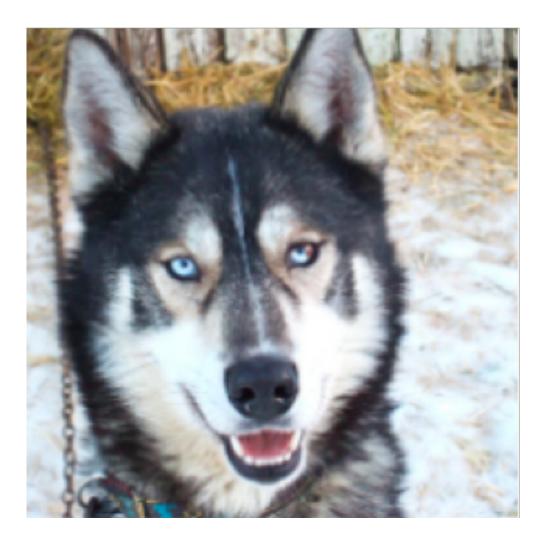
Consider an example image classification task about distinguishing between husky dogs and wolves

Example: Husky or Wolf?

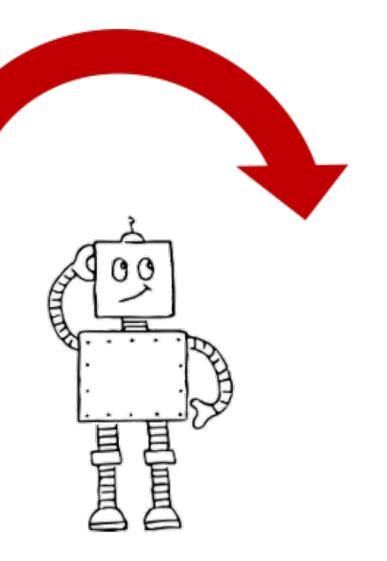




Example: Husky or Wolf? ... and why?



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

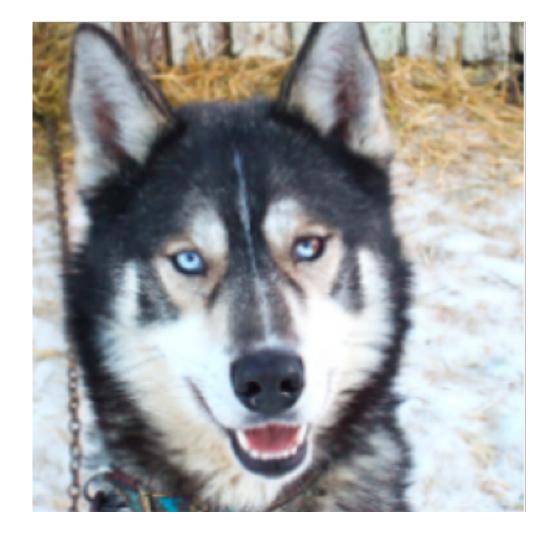


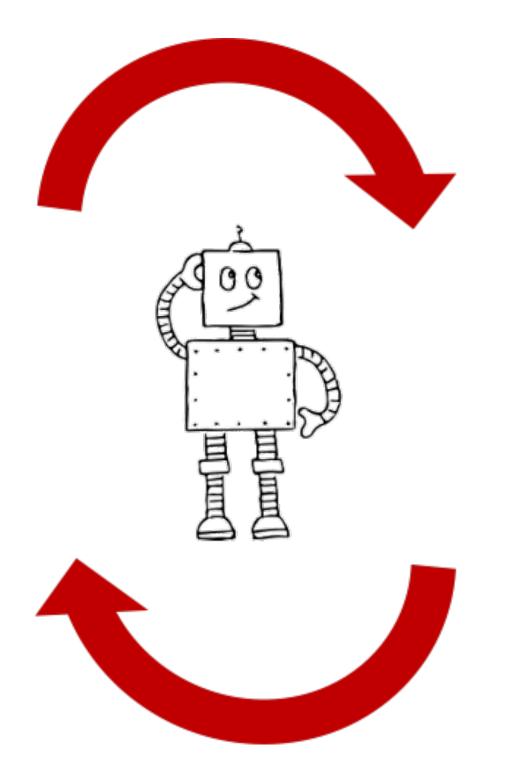


Local explanations allow to spot cases where the model is **right** for the **wrong reasons**



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



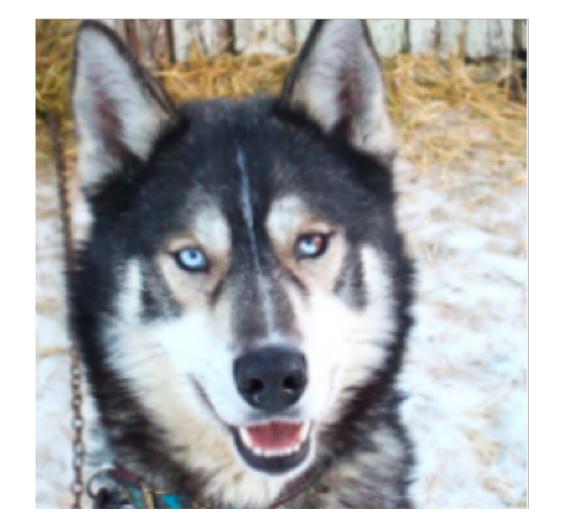


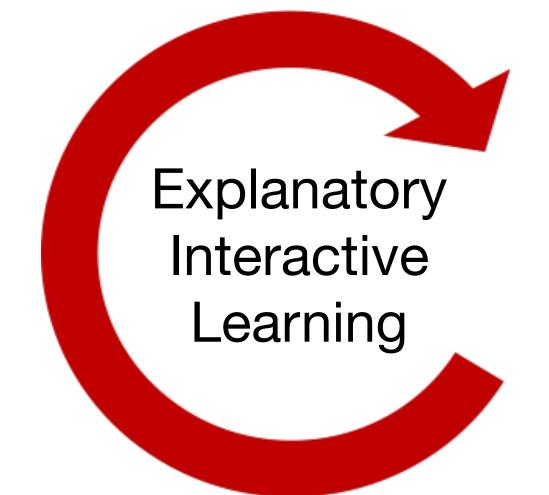
It is a husky, but not because of the highlighted pixels!





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







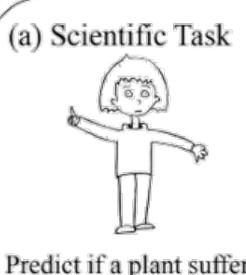
Explain predictions to users (competence, understandability) Allow user to correct explanations (directability)

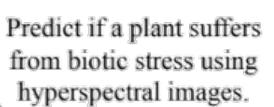


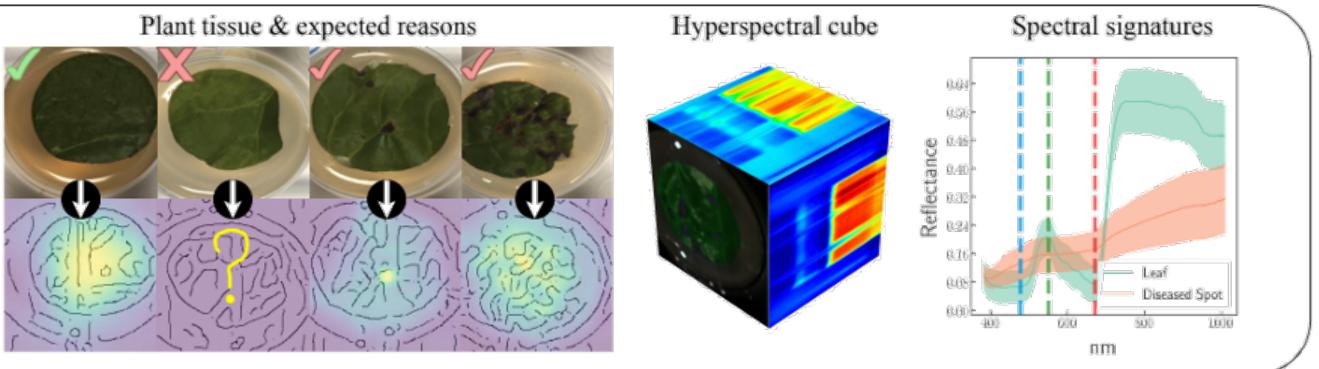


Of course, the pattern is not exclusive to images

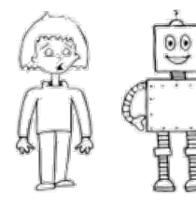
Example: plant phenotyping

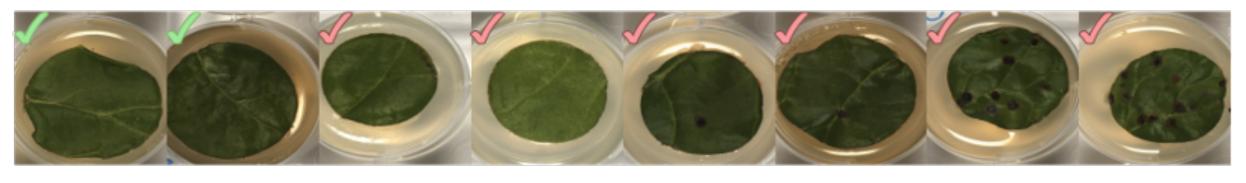






(b) Machine Learning: Often no interaction. Expert just provides the hyperspectral data and the labels.





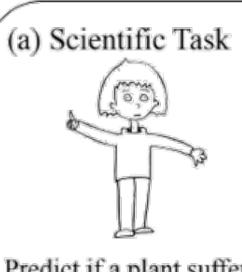
Schramowski et al. Nature Machine Intelligence 2020

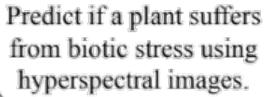
nature machine intell

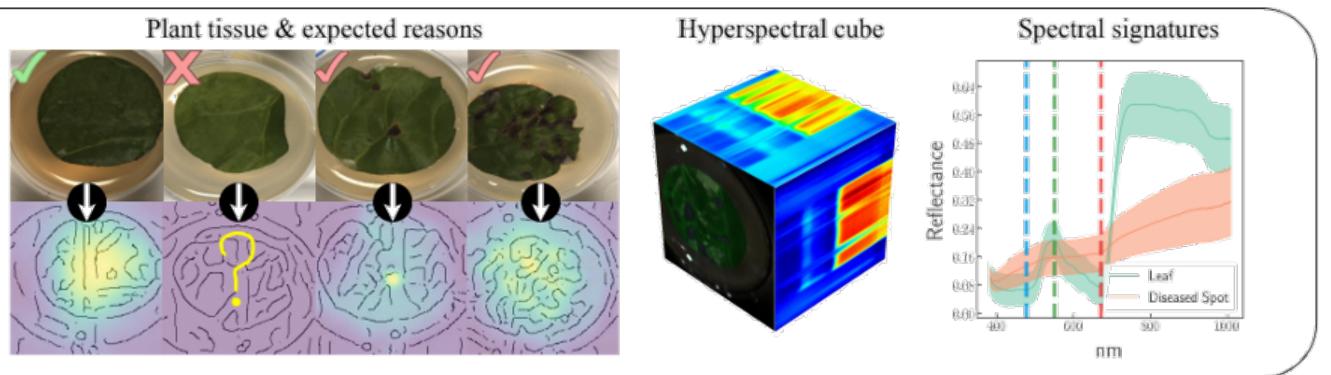


Of course, the pattern is not exclusive to images

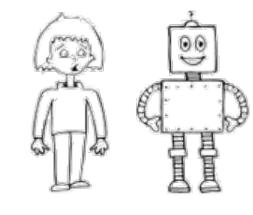
> Example: plant phenotypic

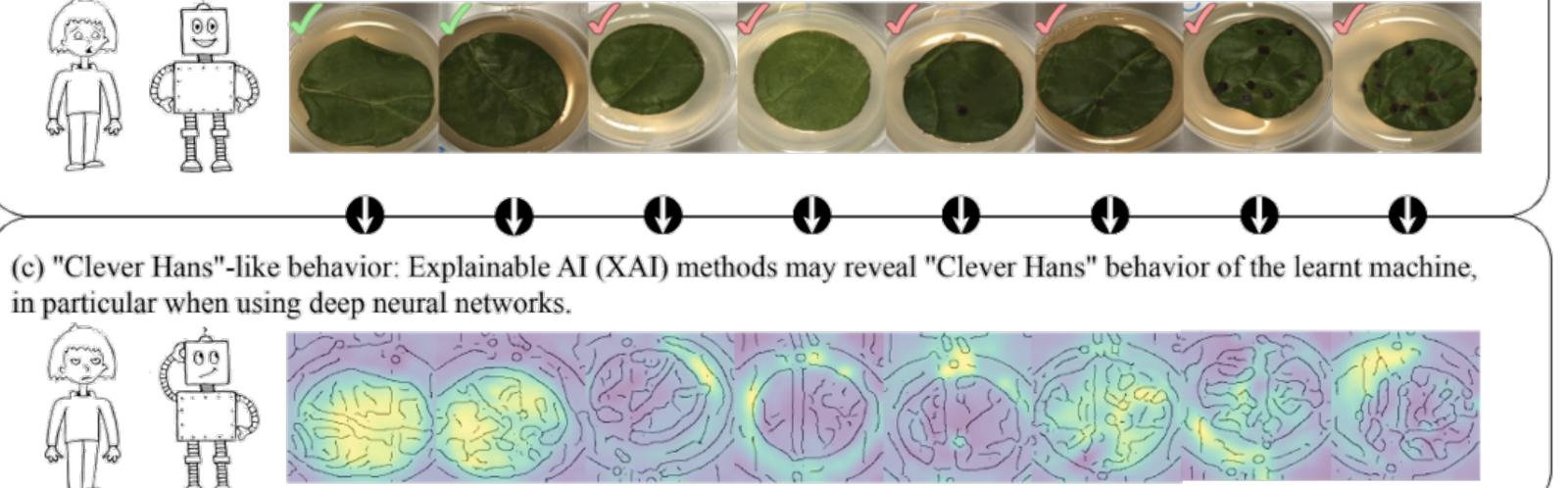




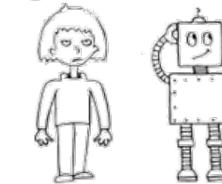


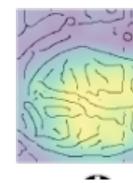
(b) Machine Learning: Often no interaction. Expert just provides the hyperspectral data and the labels.





in particular when using deep neural networks.





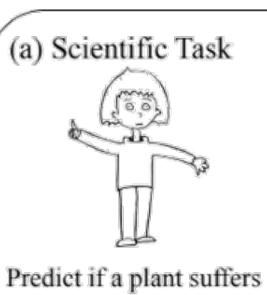
Schramowski et al. Nature Machine Intelligence 2020

nature machine intell

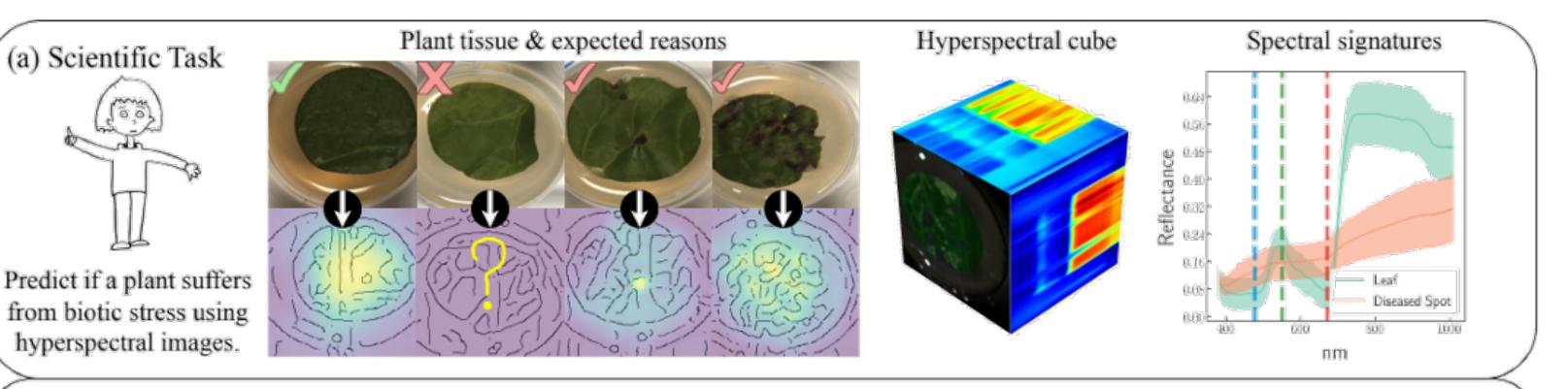


Of course, the pattern is not exclusive to images

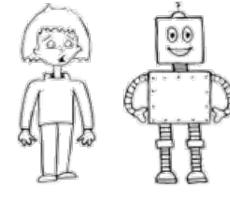
> Example: plant phenotypic

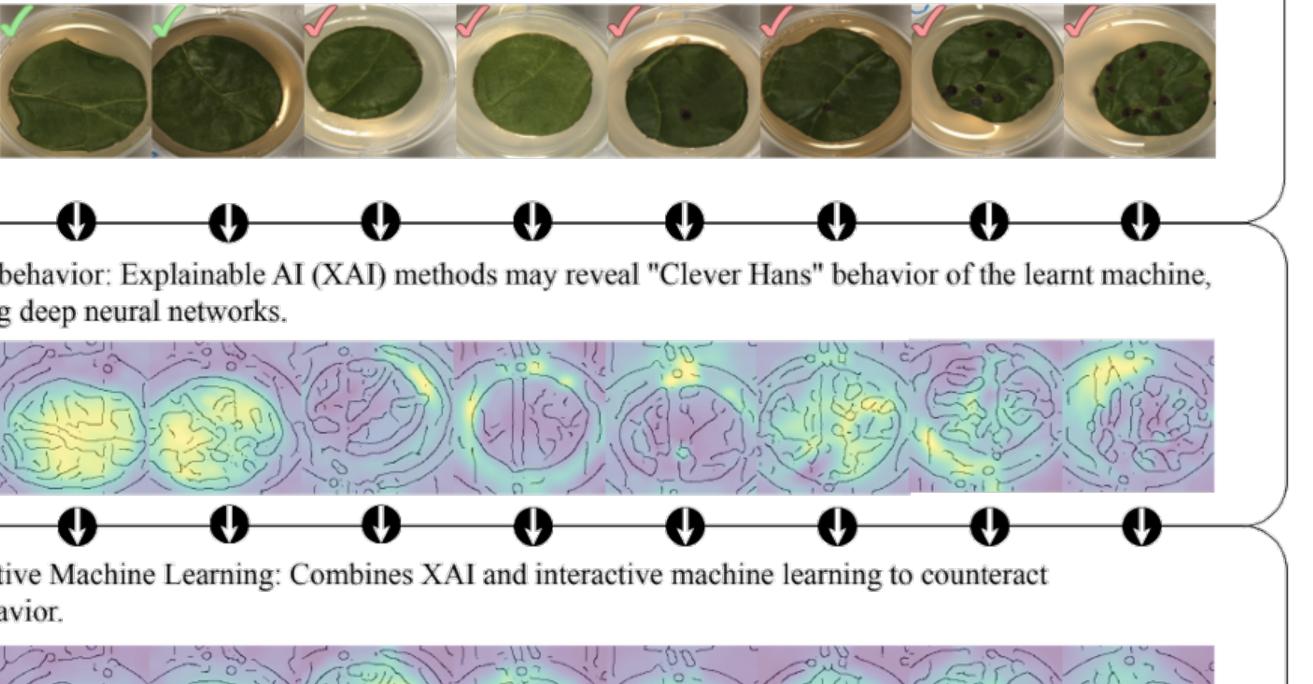


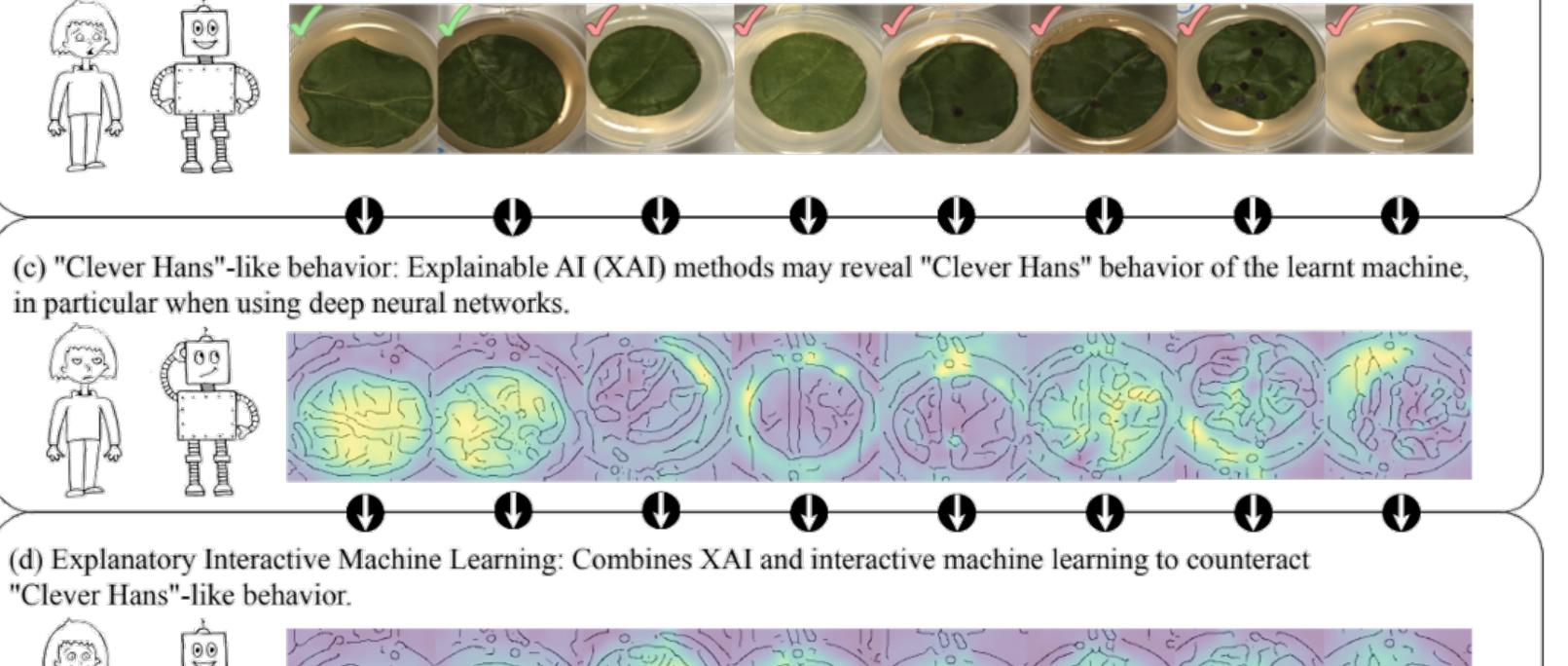
hyperspectral images.

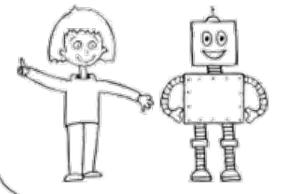


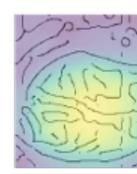
(b) Machine Learning: Often no interaction. Expert just provides the hyperspectral data and the labels.









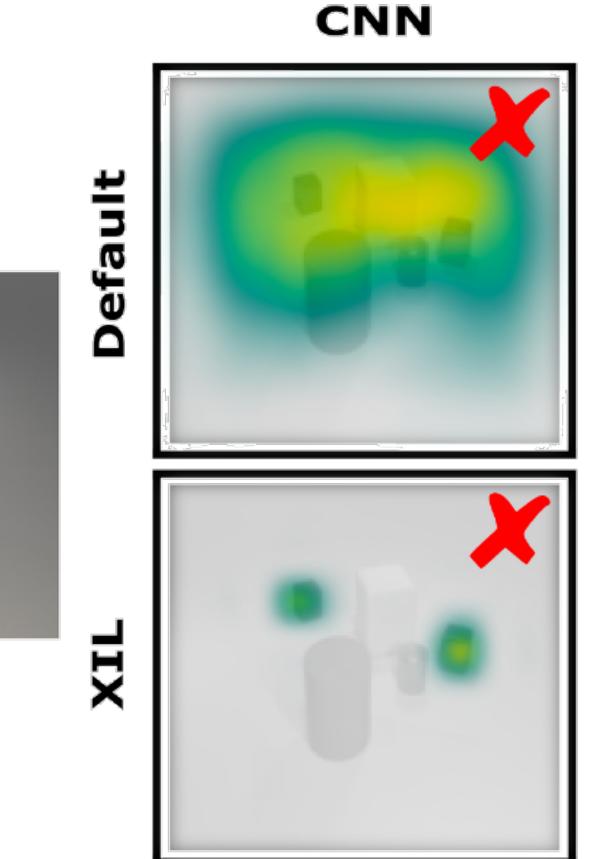


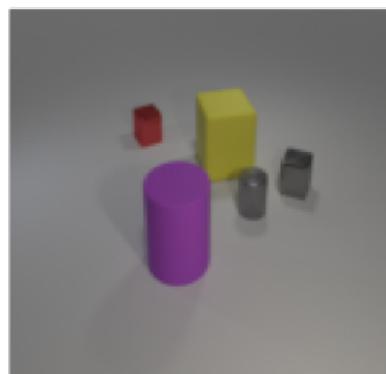
Schramowski et al. Nature Machine Intelligence 2020

nature machine intell



Unfortunately, visual explanations alone are not all we need either



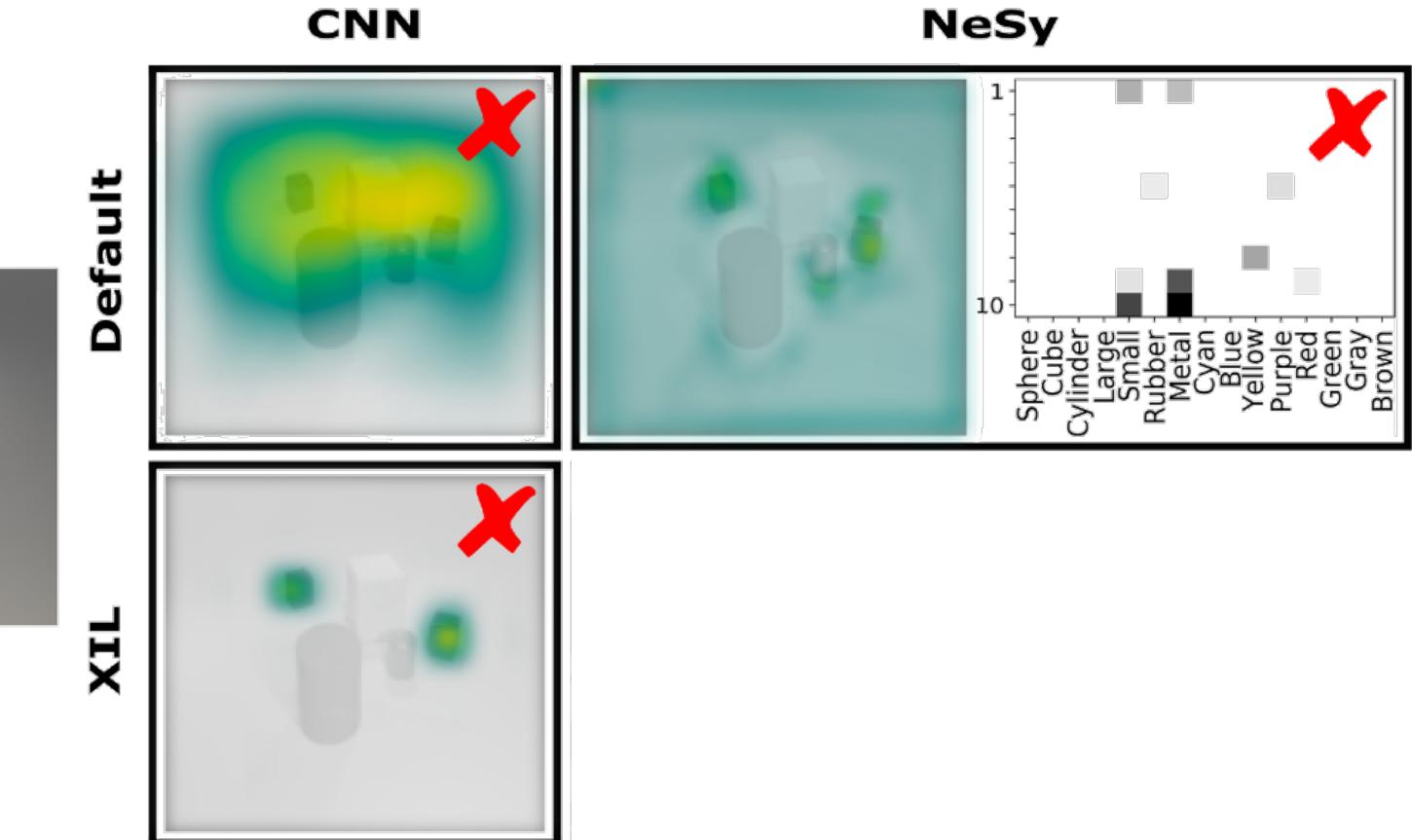


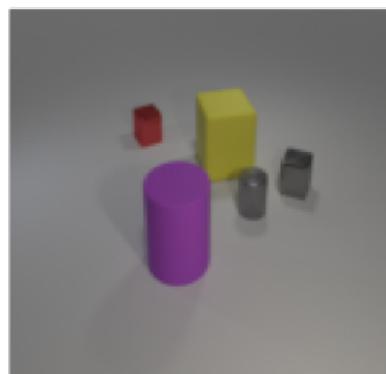
Underlying concept: the image contains a large cube & a large cylinder





Unfortunately, visual explanations alone are not all we need either

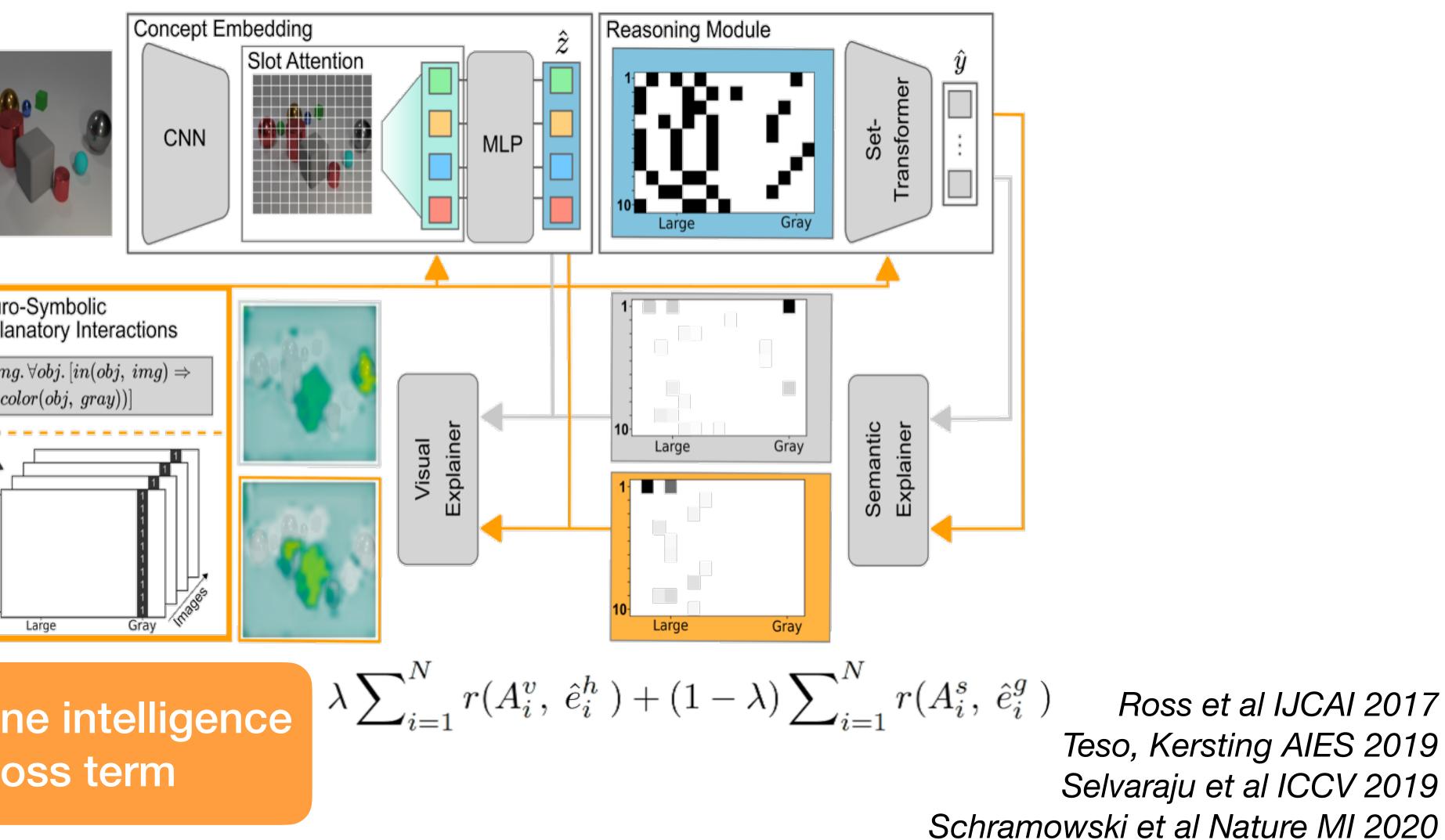


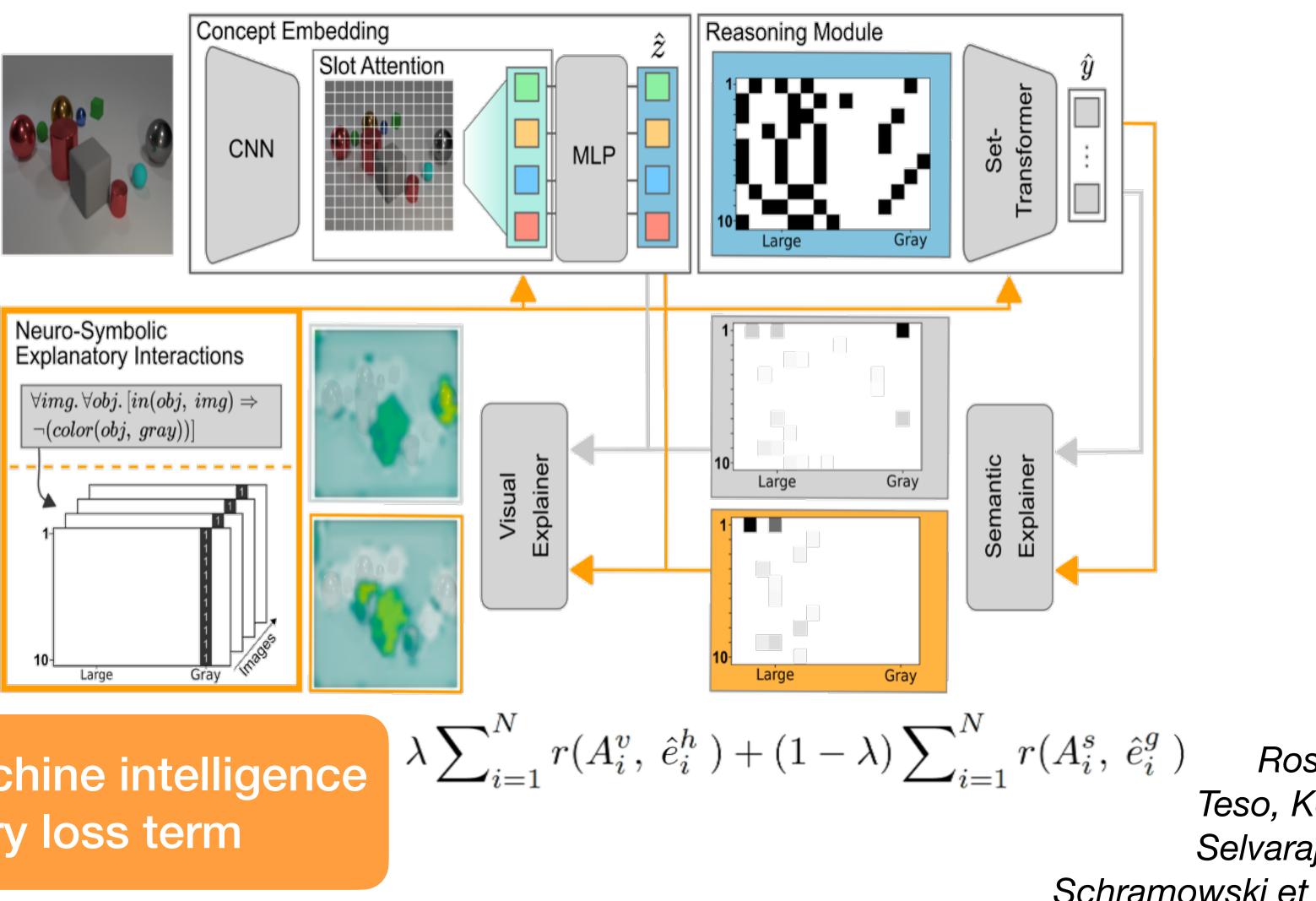


NeSy



Right for the right neuro-symbolic reasons!



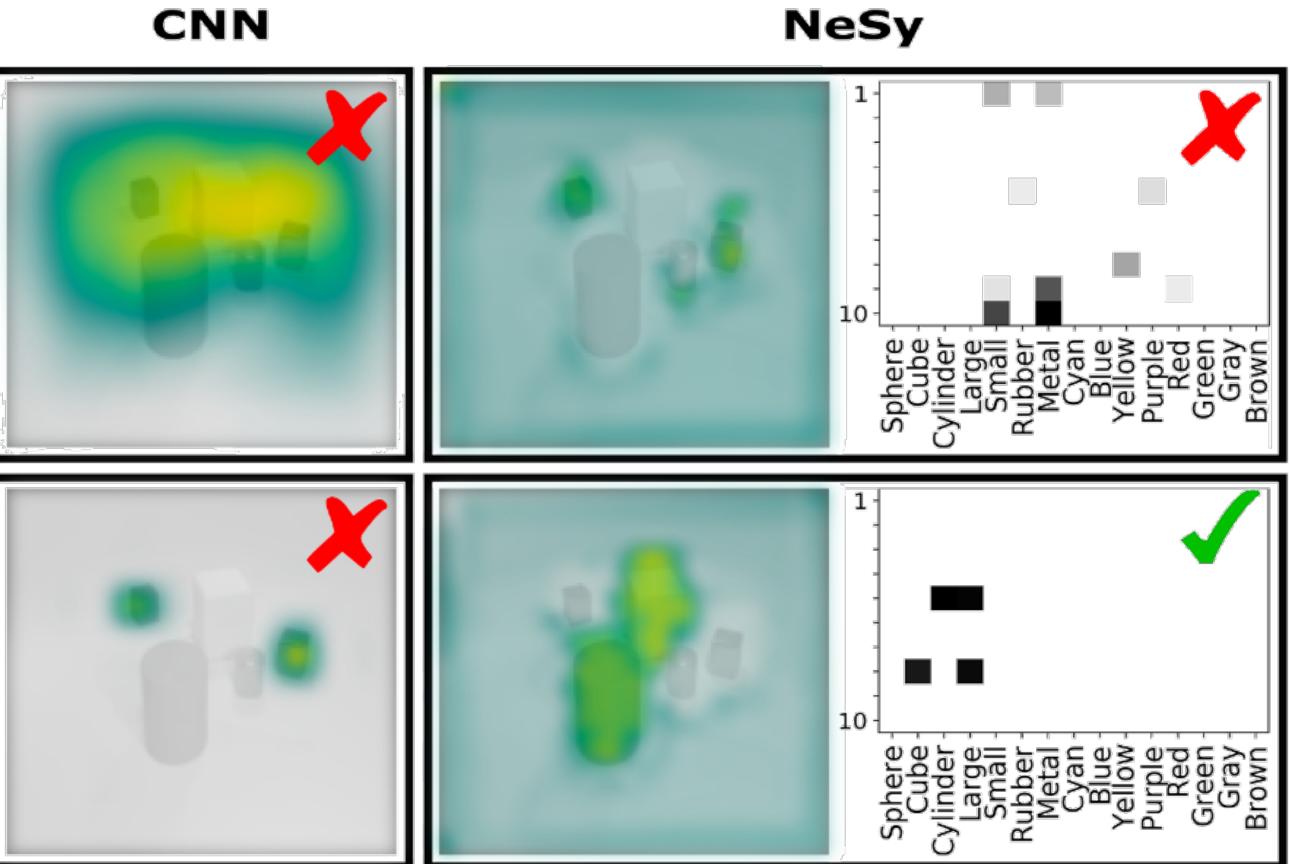


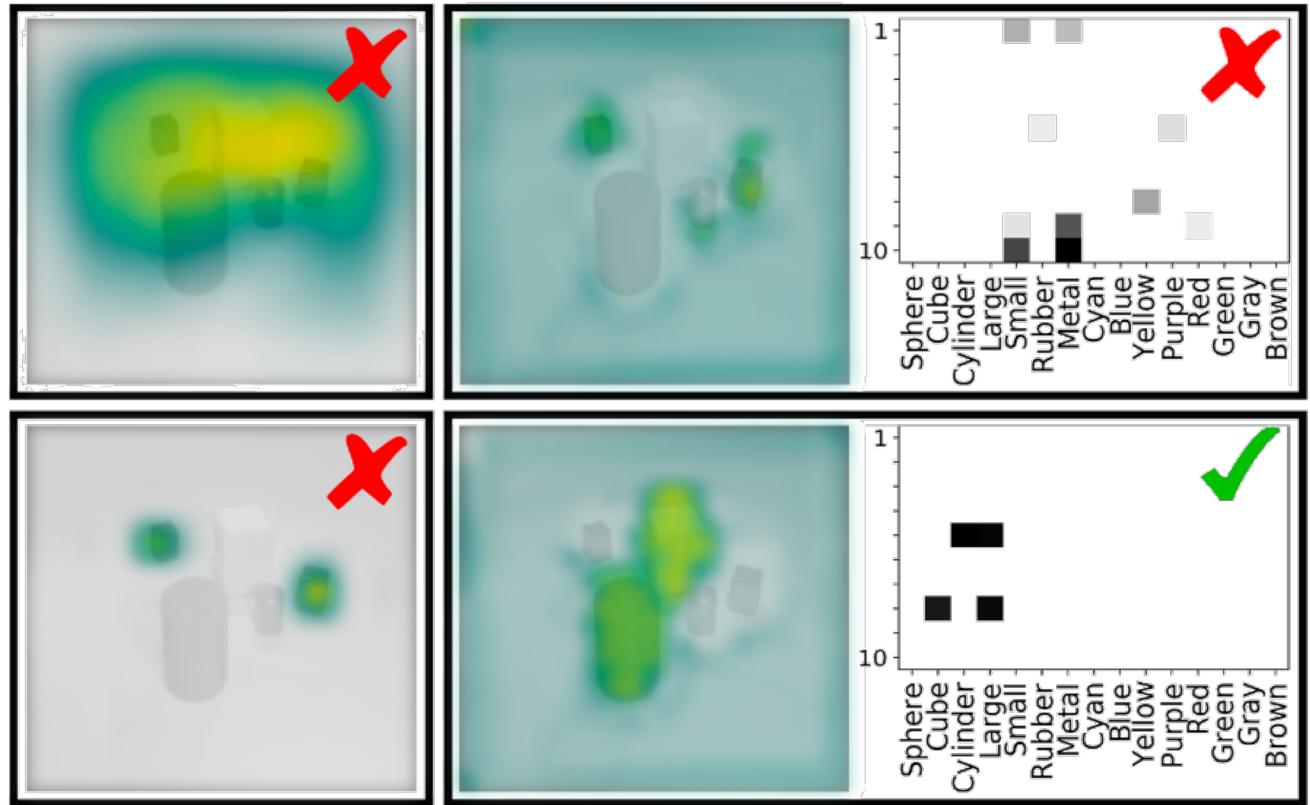
Combine human & machine intelligence via an explanatory loss term



Ross et al IJCAI 2017

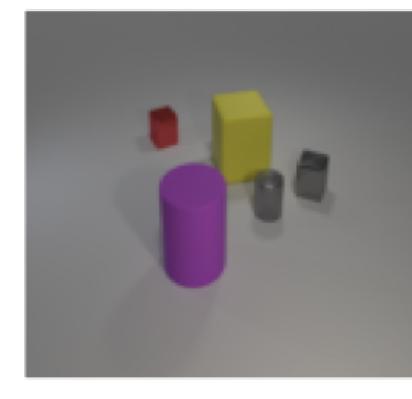
Symbolic feedback: "Never base your decision on gray cubes!"





Default

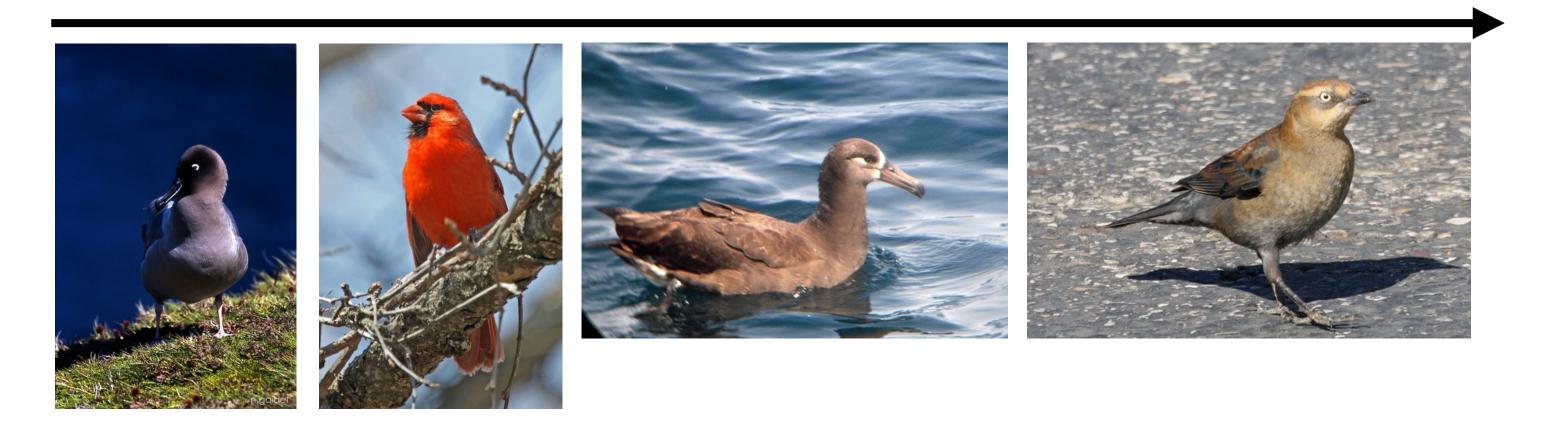


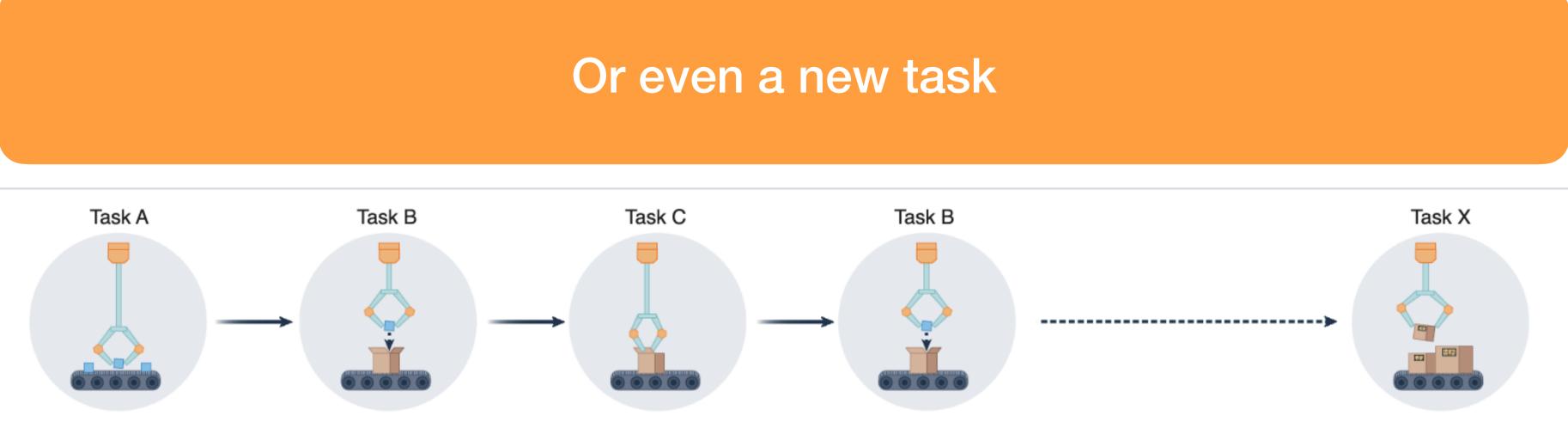


NeSy



What if there is a new concept





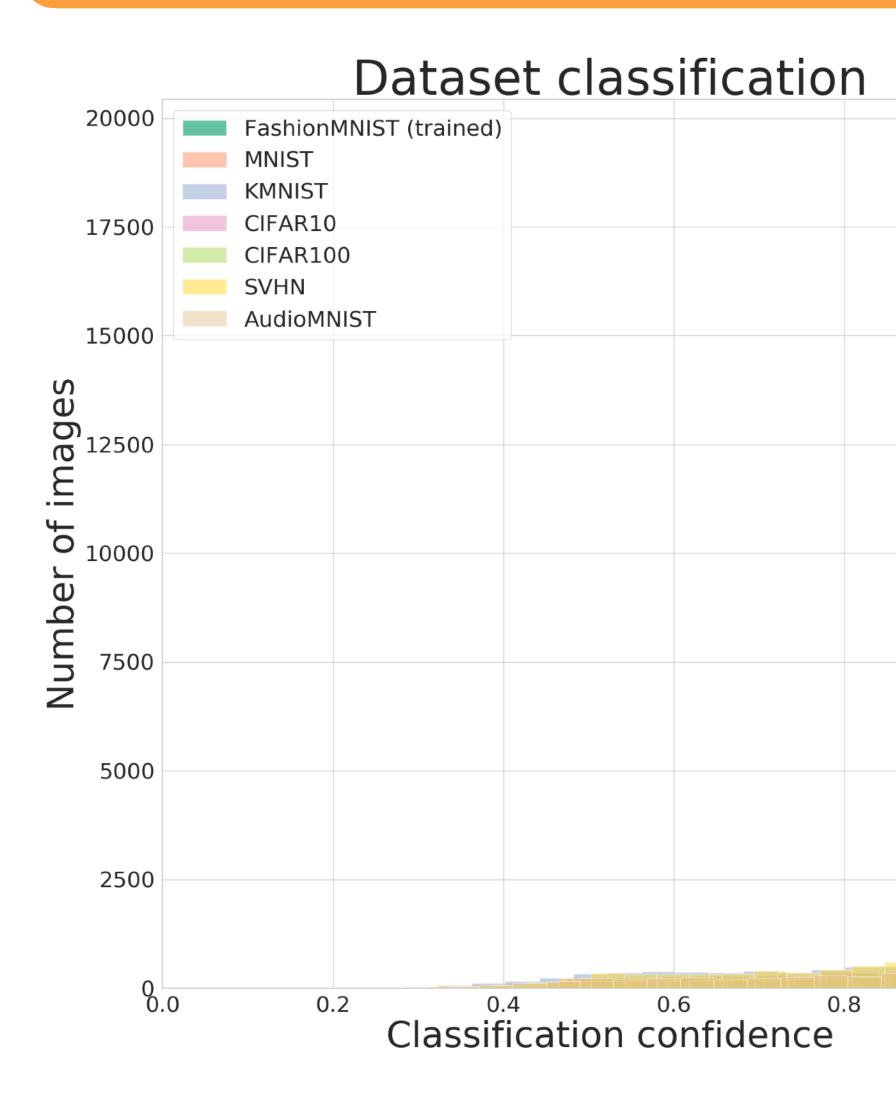
Welinder et al, Caltech-UCSD Birds 200

Kudithipudi et al, Nature MI 2022





Another challenge: deep models don't know when they don't know

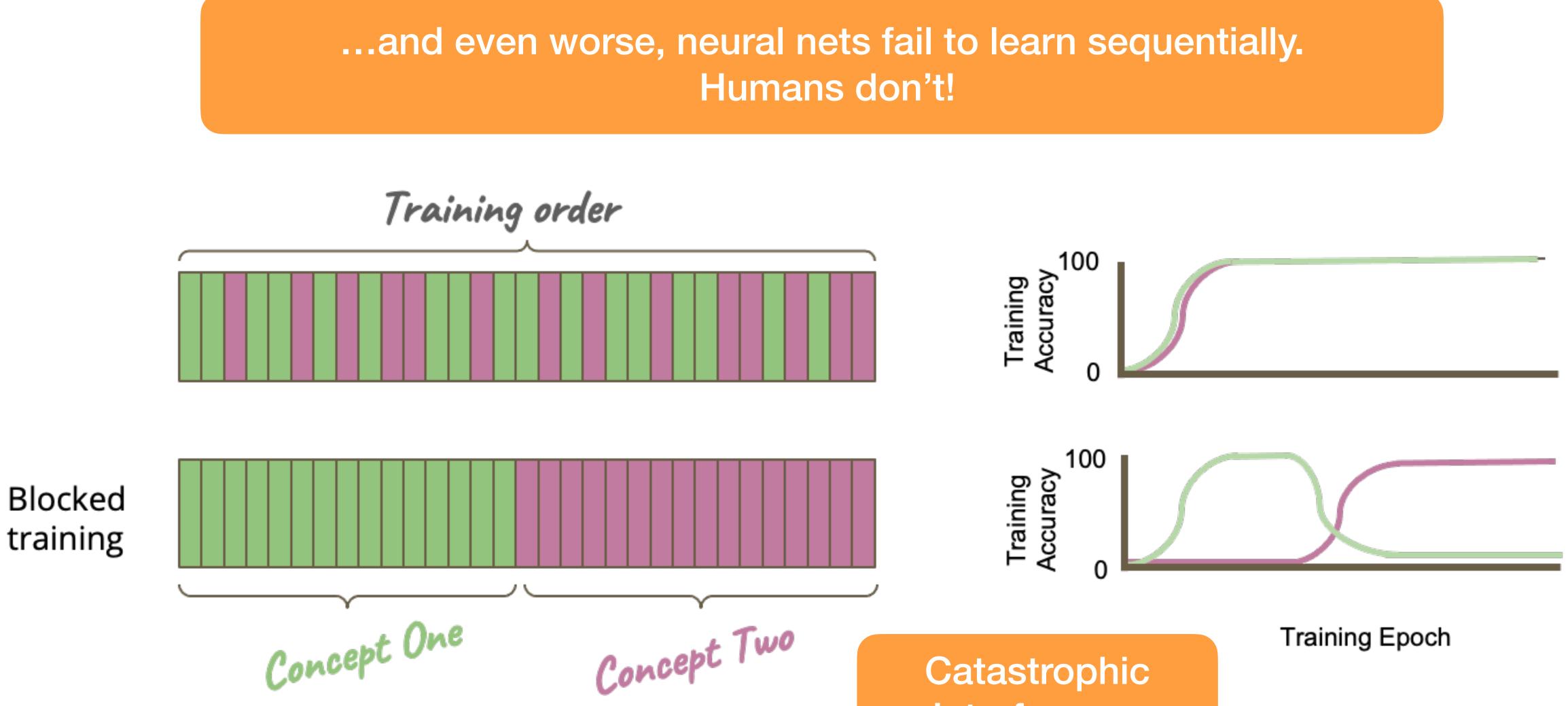


(Deep neural) models are overconfident on datasets they were never trained on

> Mundt et al. ICCV 2019 Mundt et al. Journal of Imaging 2022



Matan 1990

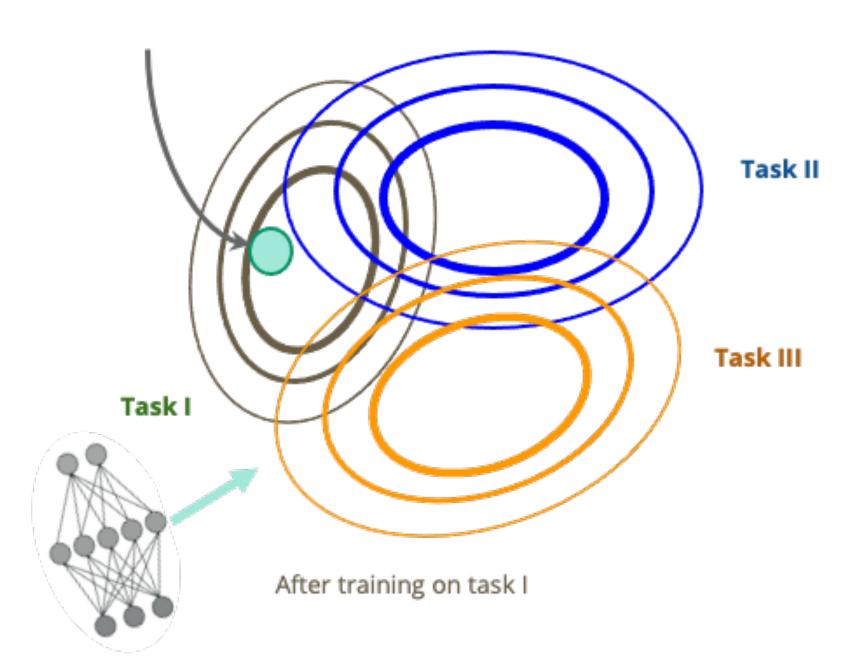


Adapted from Flesch et al 2022

Catastrophic Interference (McCloskey & Cohen 89)

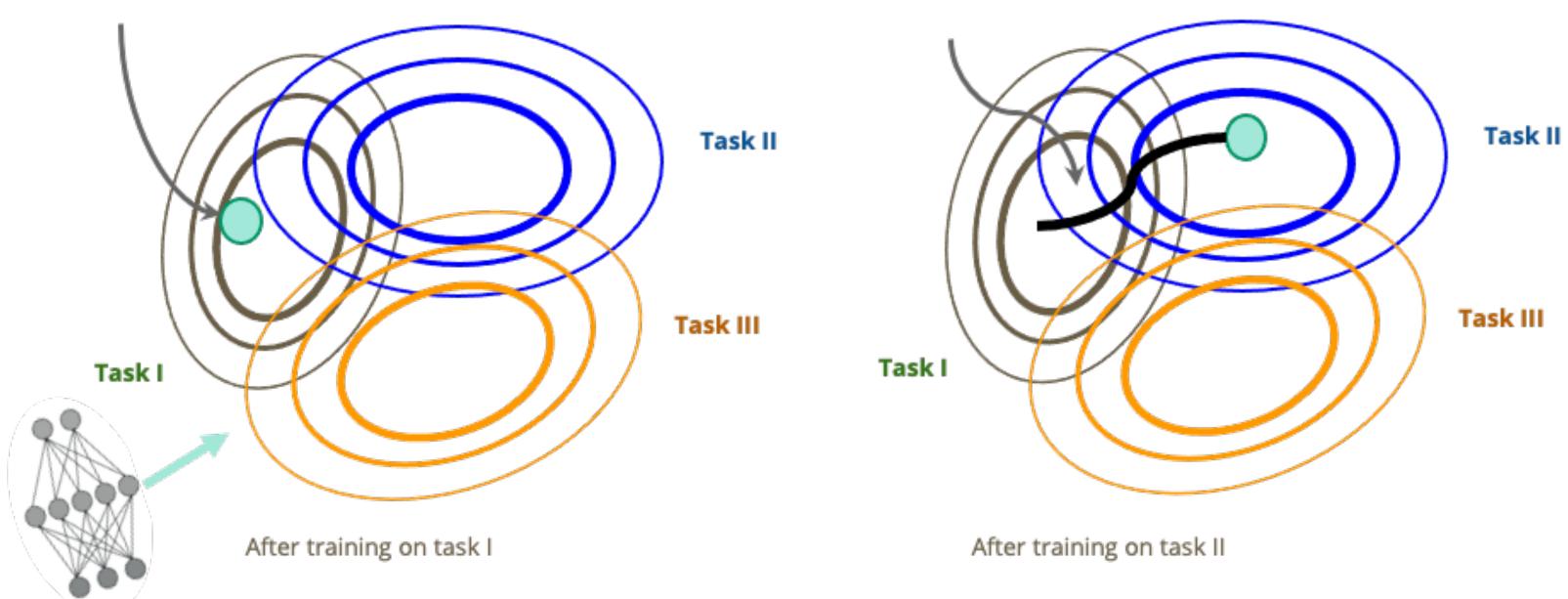


Why might neural networks be so forgetful? Is it surprising?



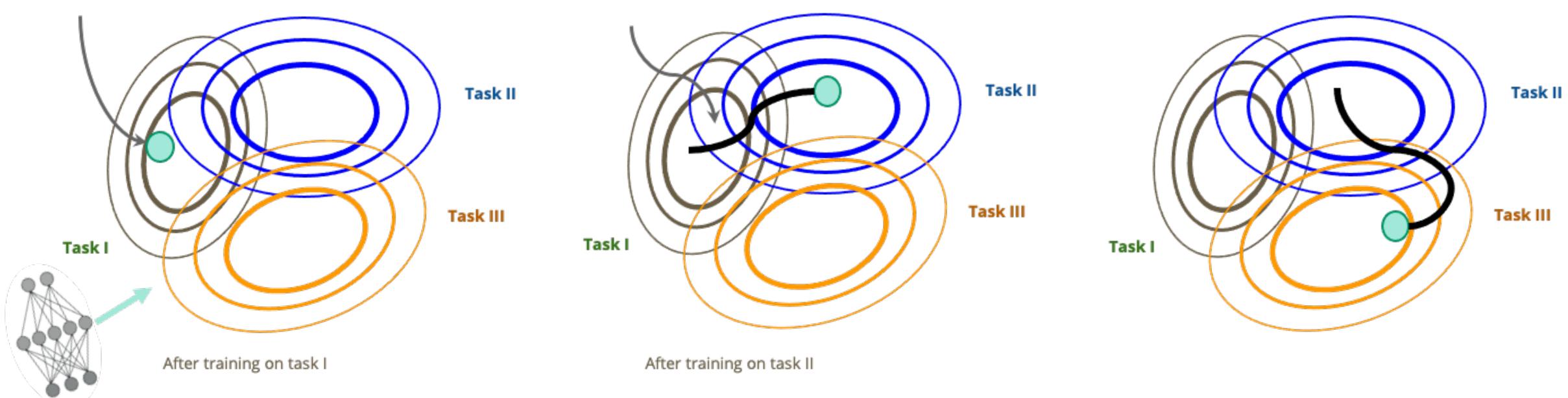


Why might neural networks be so forgetful? Is it surprising?





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









We can mitigate this problem with generative models



Continual Machine Learning That Can Identify What It **Doesn't Yet Know**

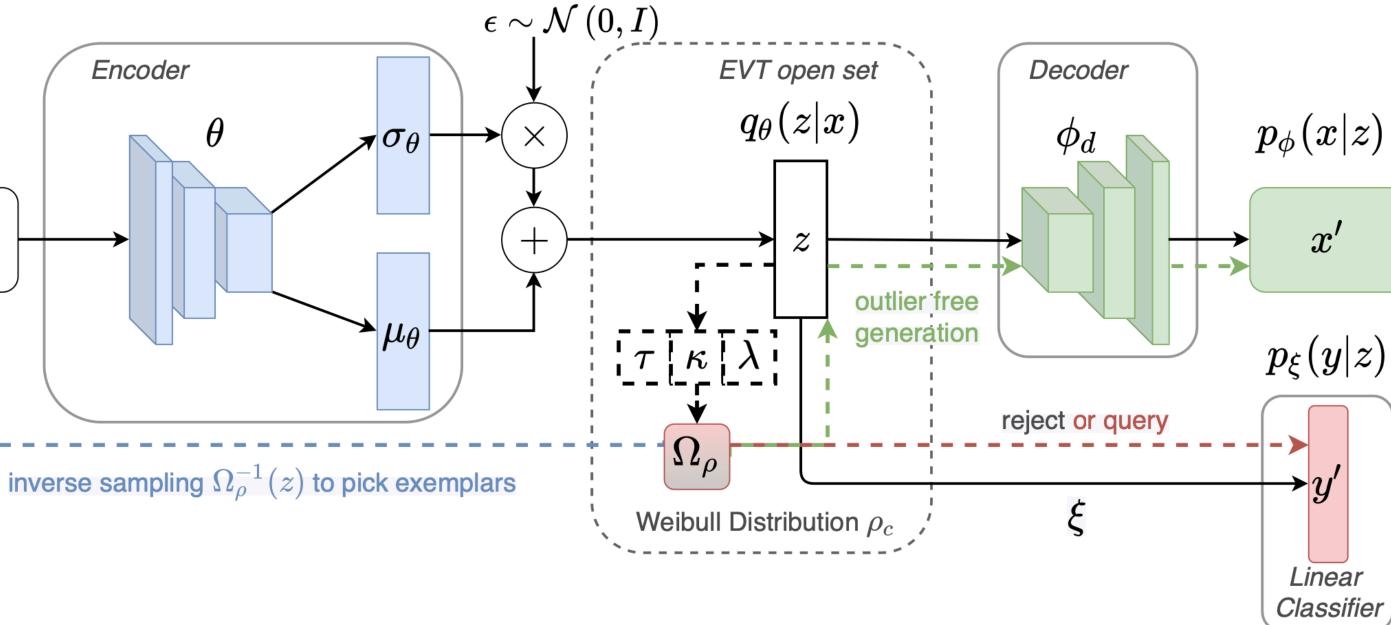
Volume 8 · Issue 4 | April 2022



Encoder x

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

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



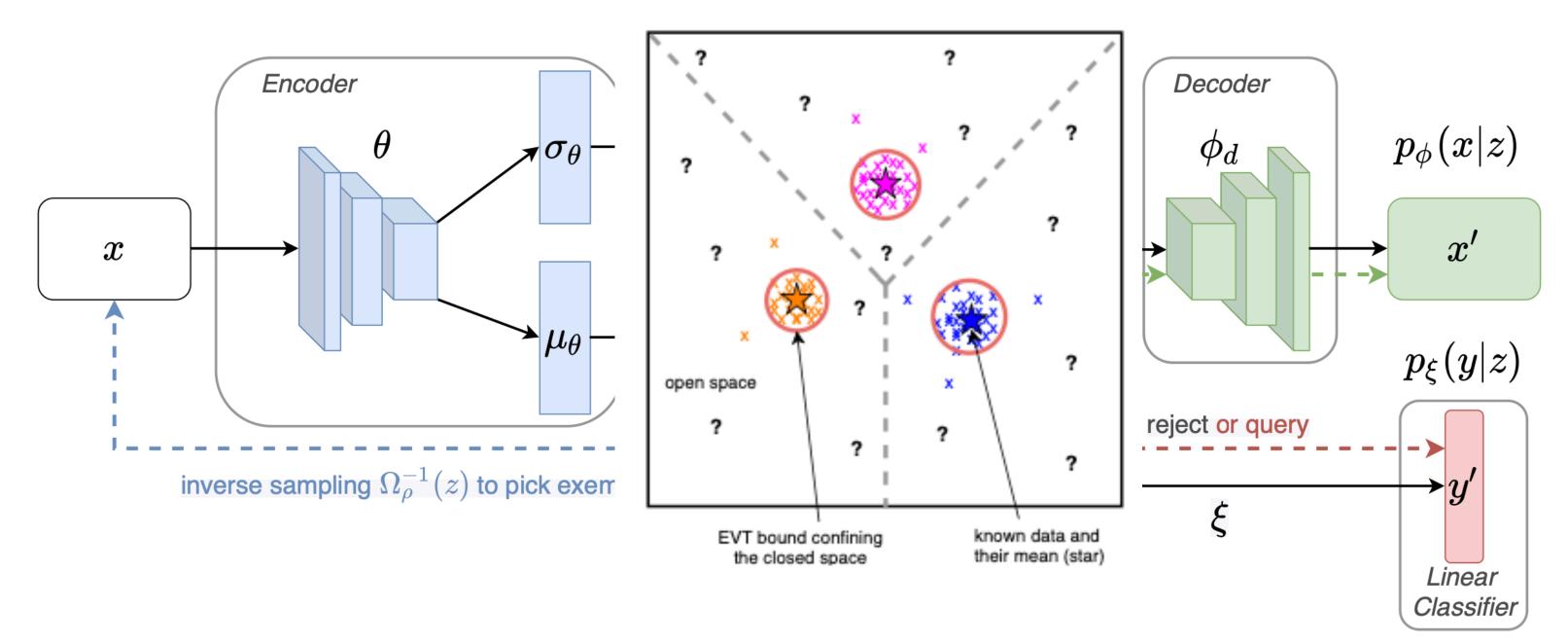






We can mitigate this problem with generative models





Continual Machine Learning That Can Identify What It **Doesn't Yet Know**

Volume 8 · Issue 4 | April 2022

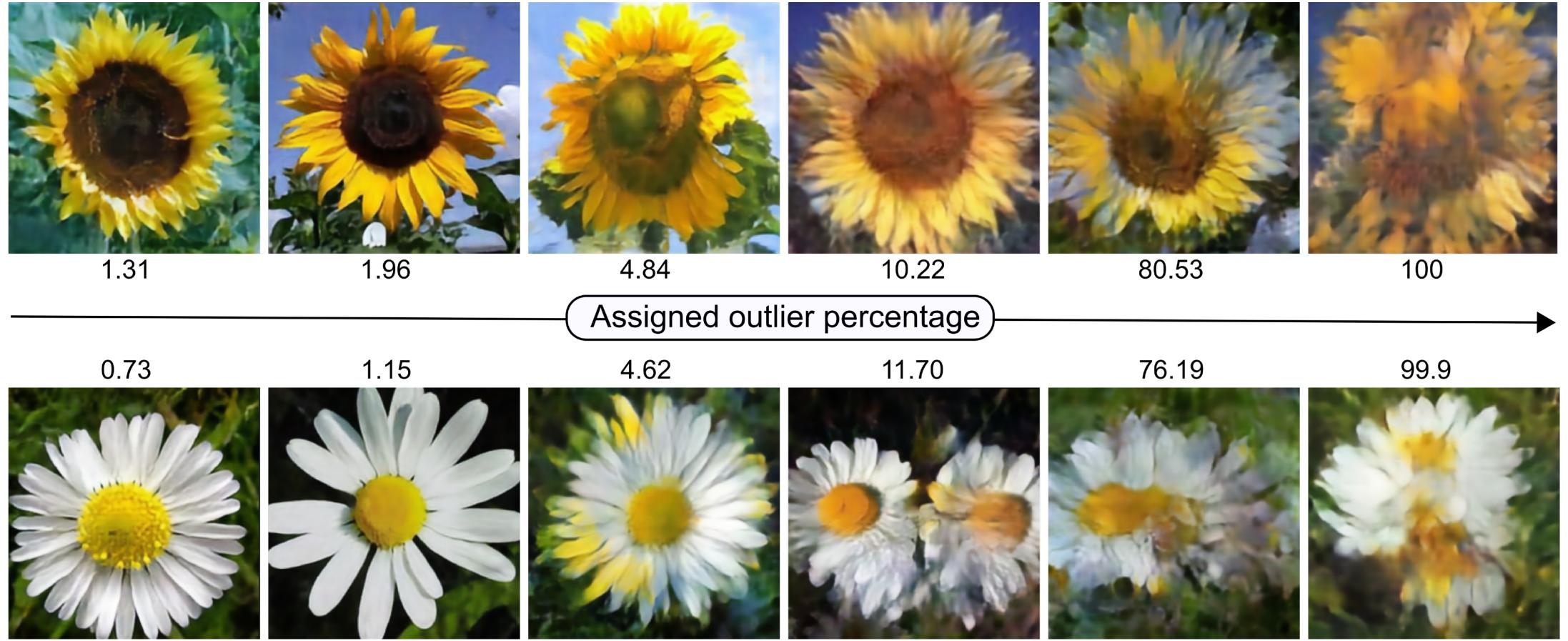


We can then measure similarity to factors we have already observed & replay knowledge from already seen ones





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



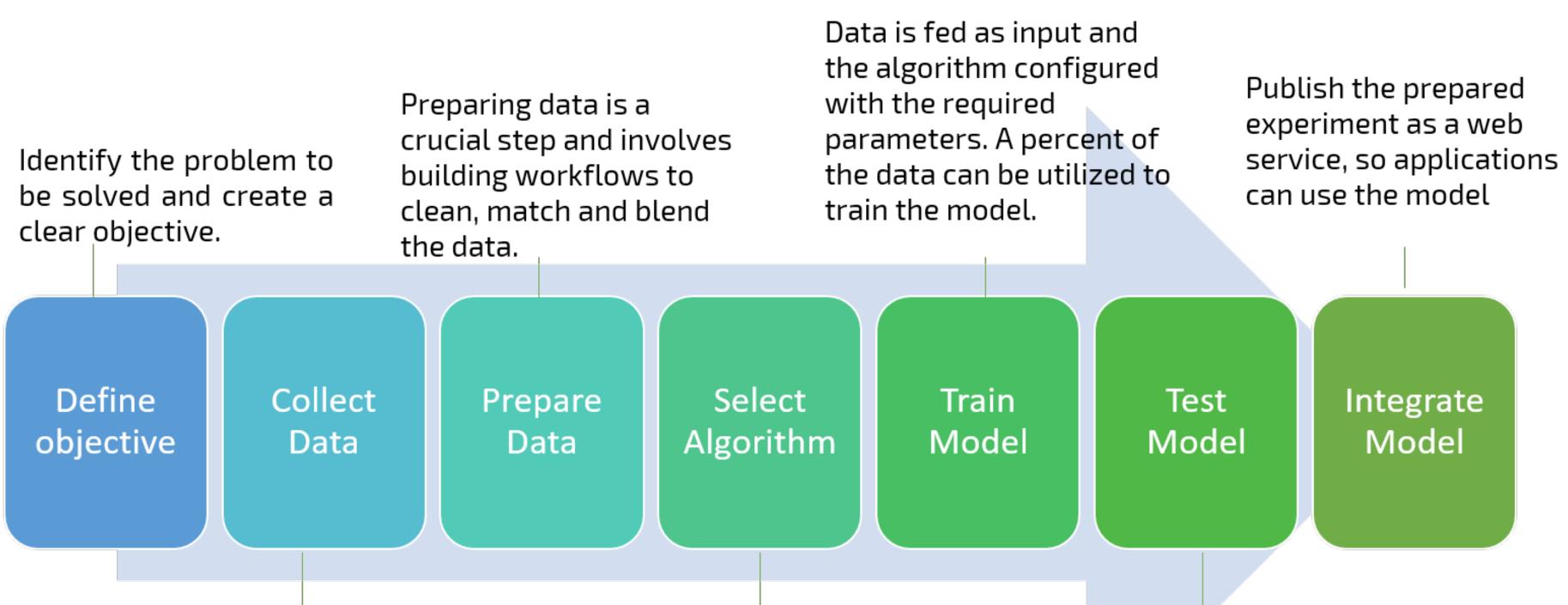
Mundt et al. Journal of Imaging 2022 Hong & Mundt Neural Networks 2022







So ultimately, the prevalent common ML pipeline is unrealistic

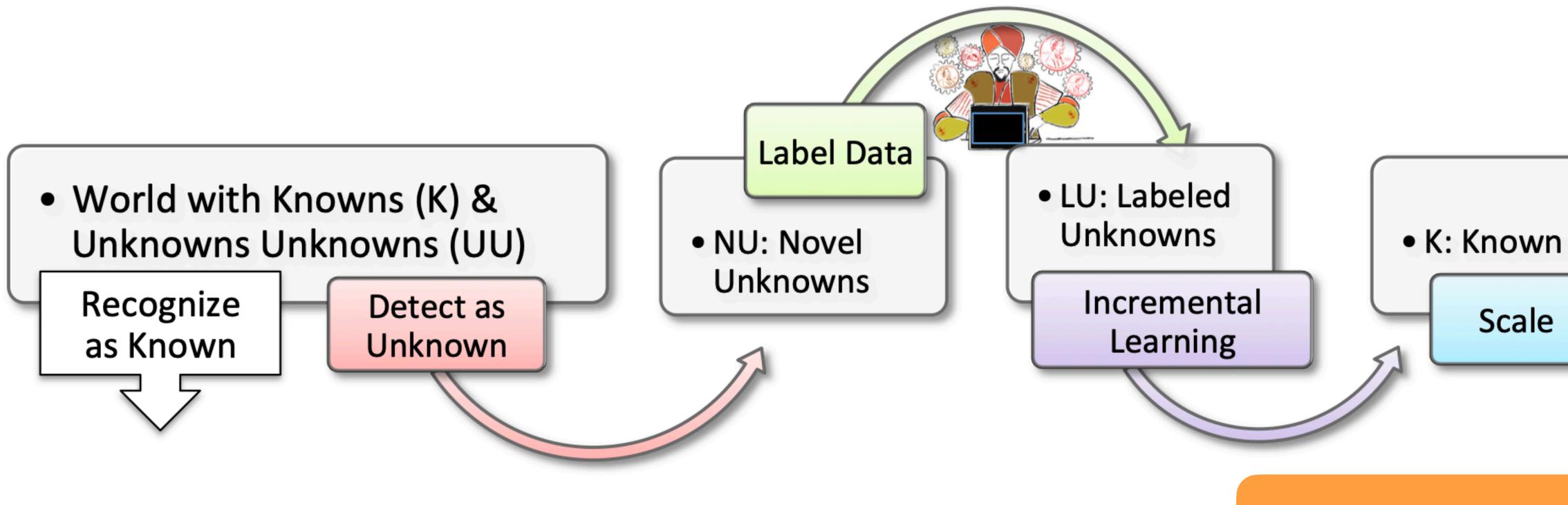


Collect data from hospitals, health insurance companies, social service agencies, police and fire dept. Depending on the problem to be solved and the type of data, an appropriate algorithm will be chosen.

The remaining data is utilized to test the model for accuracy. Depending on the results, improvements can be performed in the "Train model" and/or "Select Algorithm" phases, iteratively.



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



Bendale & Boult, CVPR 2015 Mundt et al, Neural Networks 2023 And perhaps even further







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

Versioning: stage versions according to prediction evaluation and deployment

discretized vs. continuous versions, backward compatibility

Prediction: test set evaluation, failure modes and robustness

evolving test set, inherent noise and perturbations, open world scenario

Monitor Predictions

Deploy Model

Deployment: model saving, platform compatibility, serving and cloud

> optimizer states and meta-data, distributing continuous updates, communication cost

Data: amount, redundancy vs. diversity,

cleaning, preprocessing

Code ML Mode/



data selection and ordering, task similarity, noisy streams, distribution shifts

(Continual) Machine Learning Workflow

ISDOM SUNT + MENT

Model: architecture, inductive bias, discriminative/generative, functions, parameters

model extensions, task-specific parameter identification

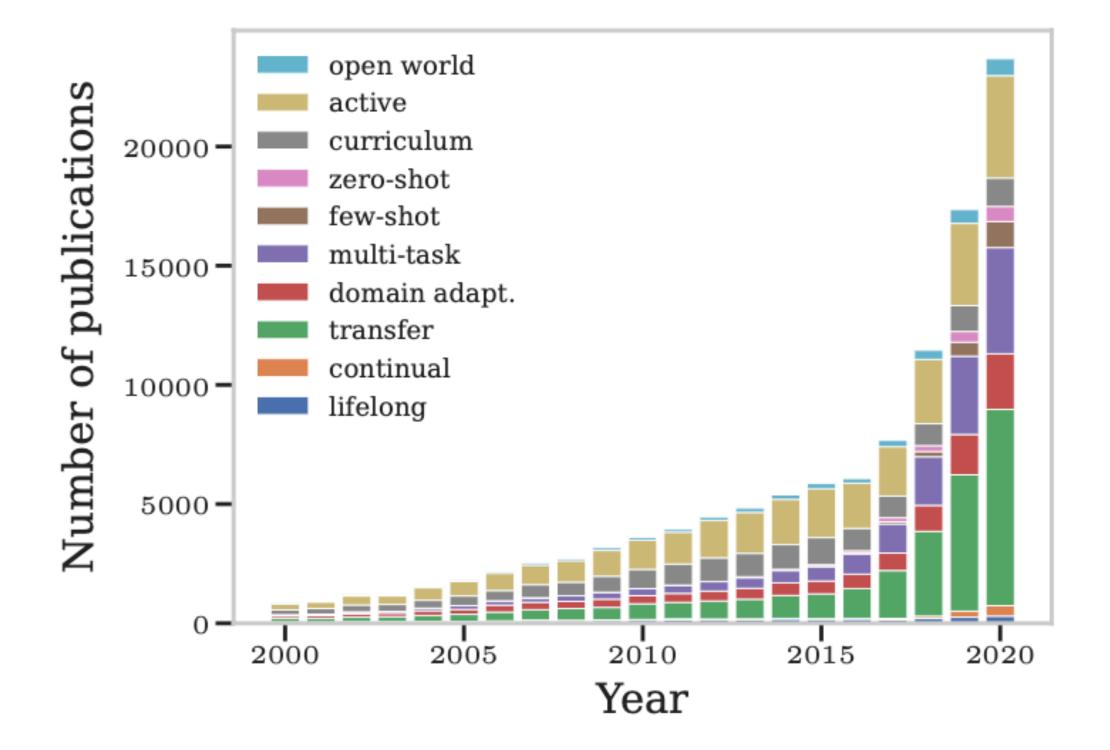
Training: loss function, optimizer, hyperparameters, convergence

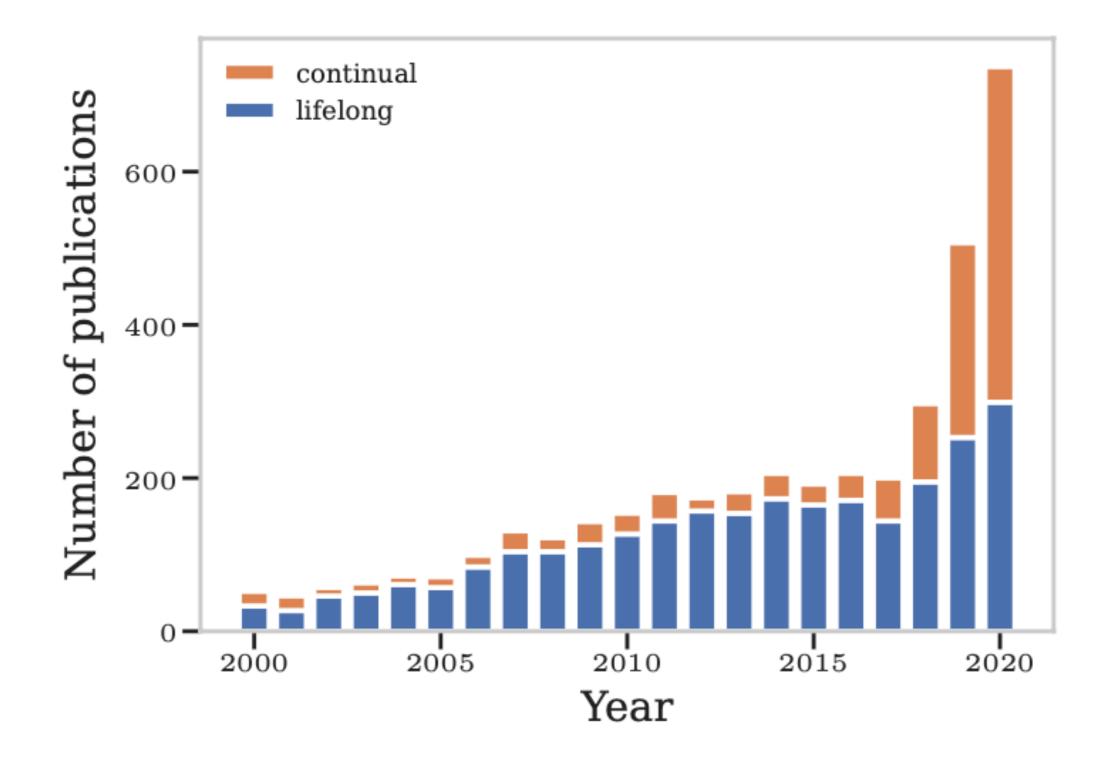
catastrophic forgetting, knowledge transfer or distillation, selective updates, online

Mundt et al, ICLR 2022



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

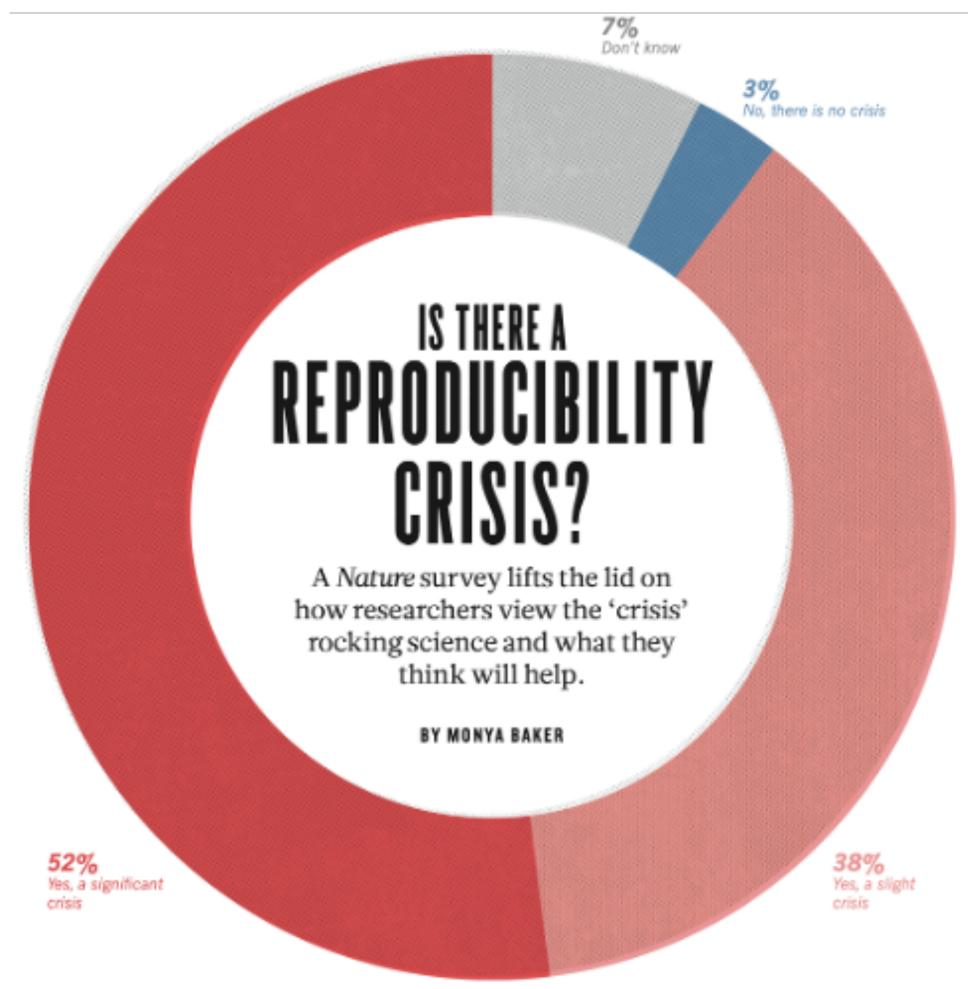




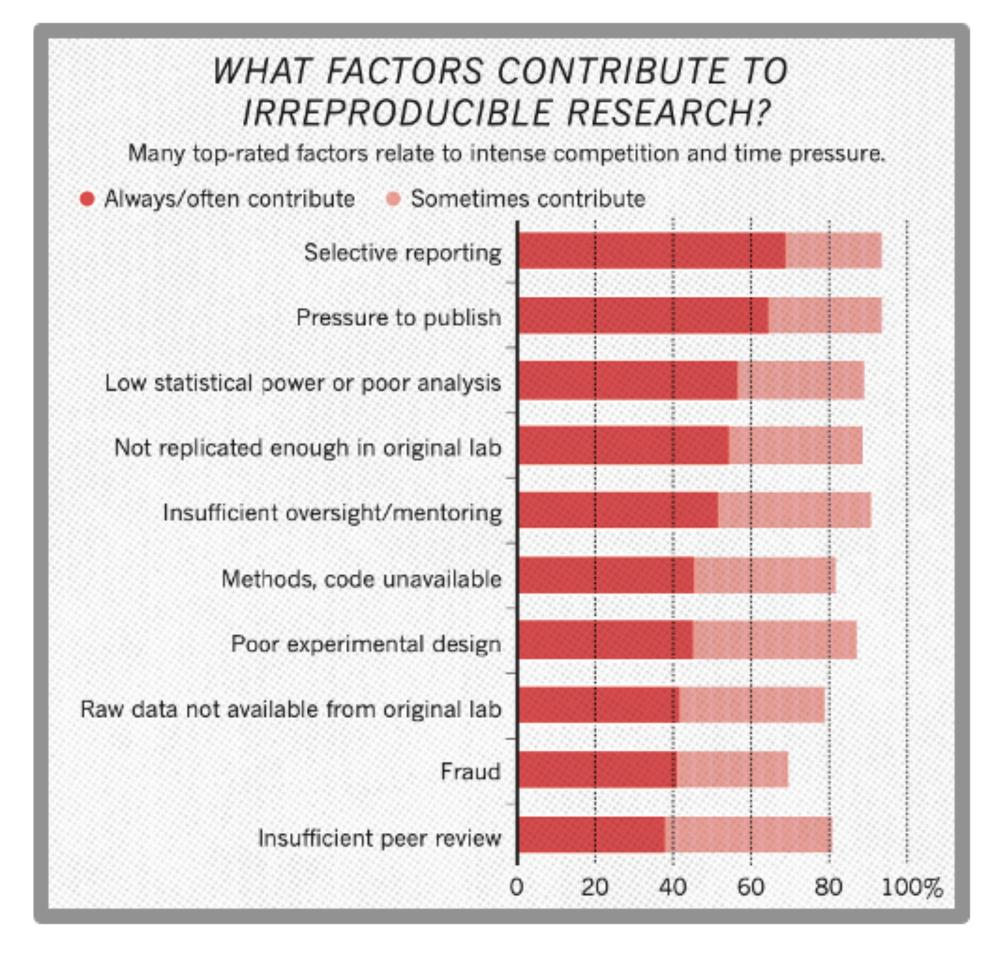
Mundt et al, ICLR 2022



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



1,576 RESEARCHERS SURVEYED

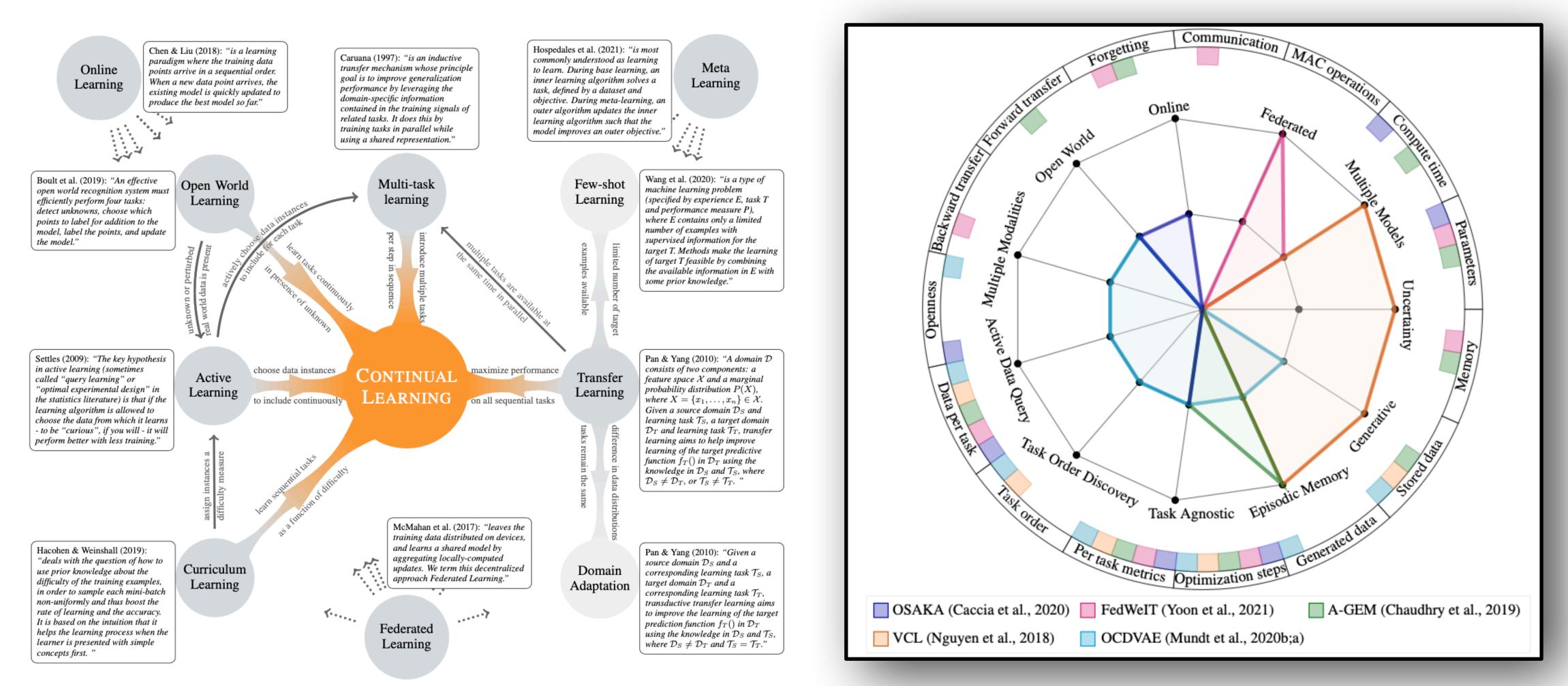


Baker, Nature 2016





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

- Standard deep neural networks are not right for the right reasons 1.
- 2. Standard deep neural networks don't know what they don't know

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

3. Standard deep neural networks are **bad at learning sequentially/continually**

But very powerful -> generative + symbolic + human





Thanks to ... and many more!



Visvanathan Ramesh Uni Frankfurt, hessian.Al



Kristian Kersting TU Darmstadt, hessian.Al



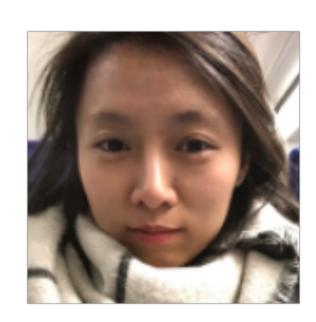
Iuliia Pliushch Uni Frankfurt



Patrick Schramowski TU Darmstadt, DFKI



Wolfgang Stammer TU Darmstadt



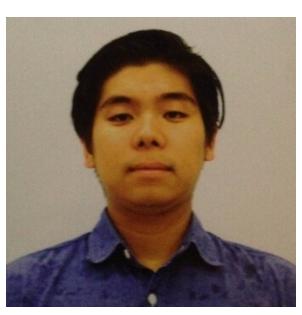
Xiaoting Shao TU Darmstadt, Evonik



TECHNISCHE UNIVERSITÄT DARMSTADT







Yongwon Hong, Yonsei University



Steven Braun TU Darmstadt

25



Vincenzo Lomonaco Uni Pisa, ContinualAl



Tyler Hayes NAVER, ContinualAI



