Machine Learning Beyond Static Datasets

ESSAI 2023

Dr. Martin Mundt,

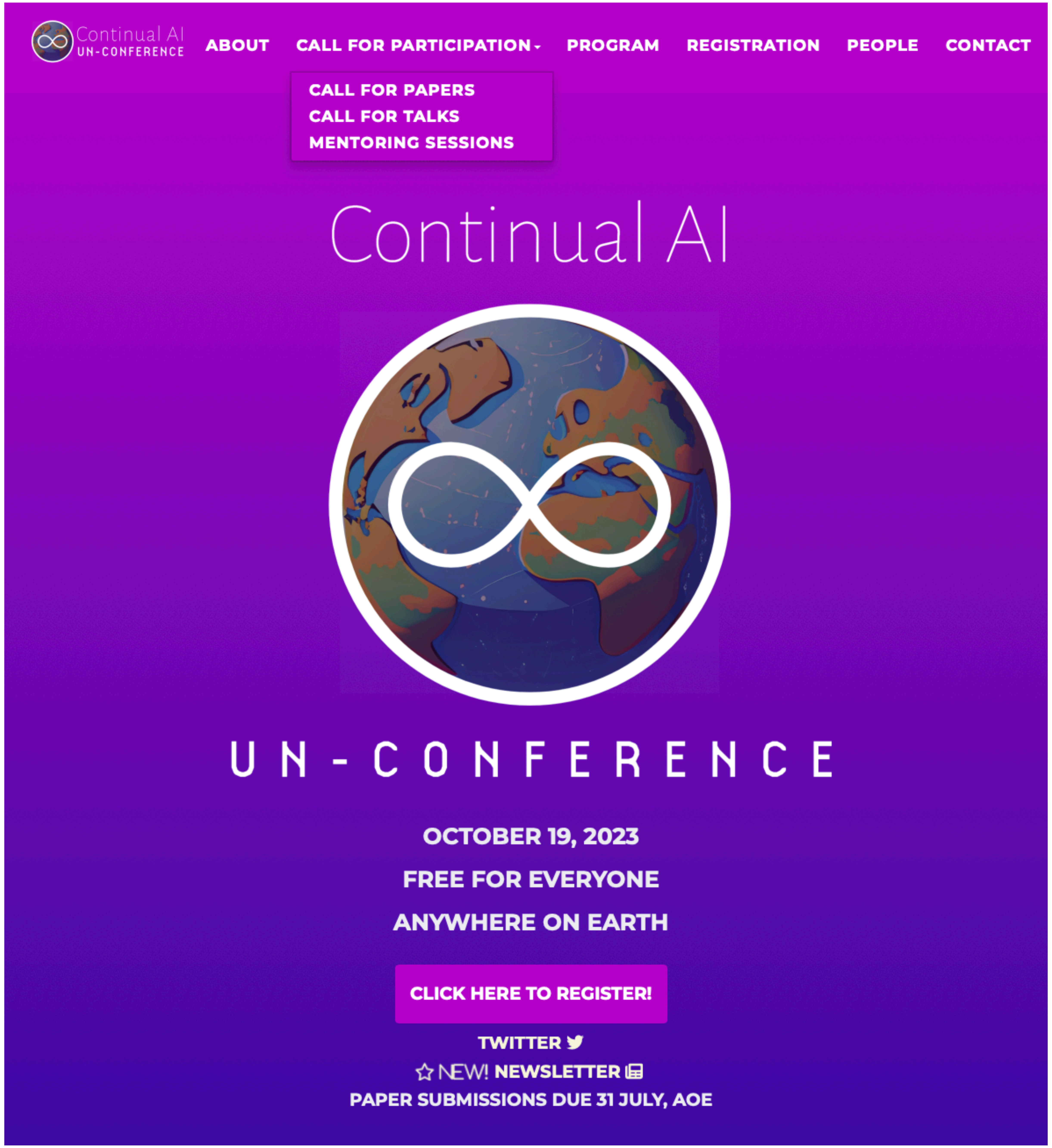
Research Group Leader, TU Darmstadt & hessian.AI

Board Member of Directors, ContinualAI

Course: http://owll-lab.com/teaching/essai-23

Day 5: The Unknown
Open World Learning & Evaluation

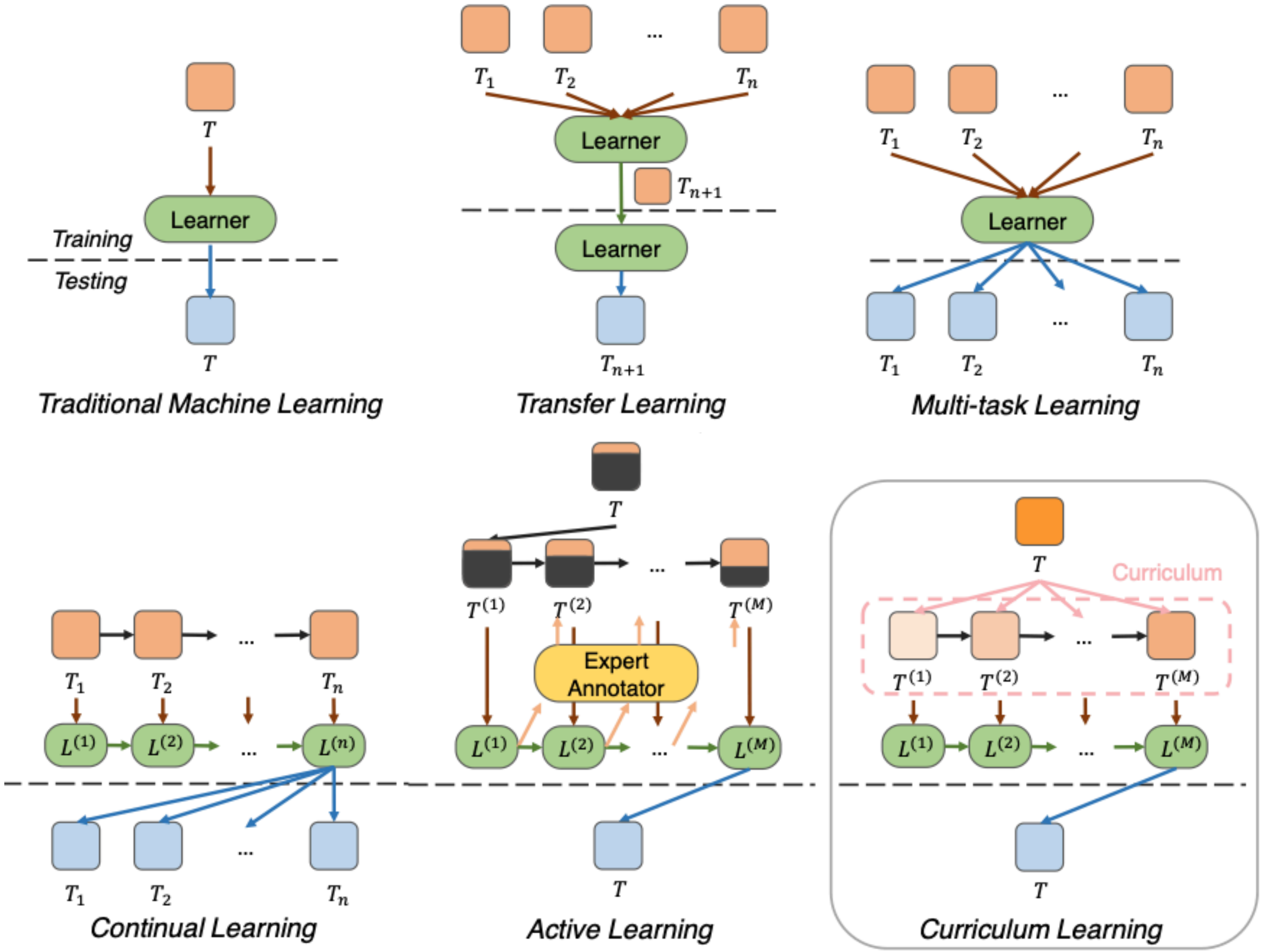
ContinualAI Un-Conference

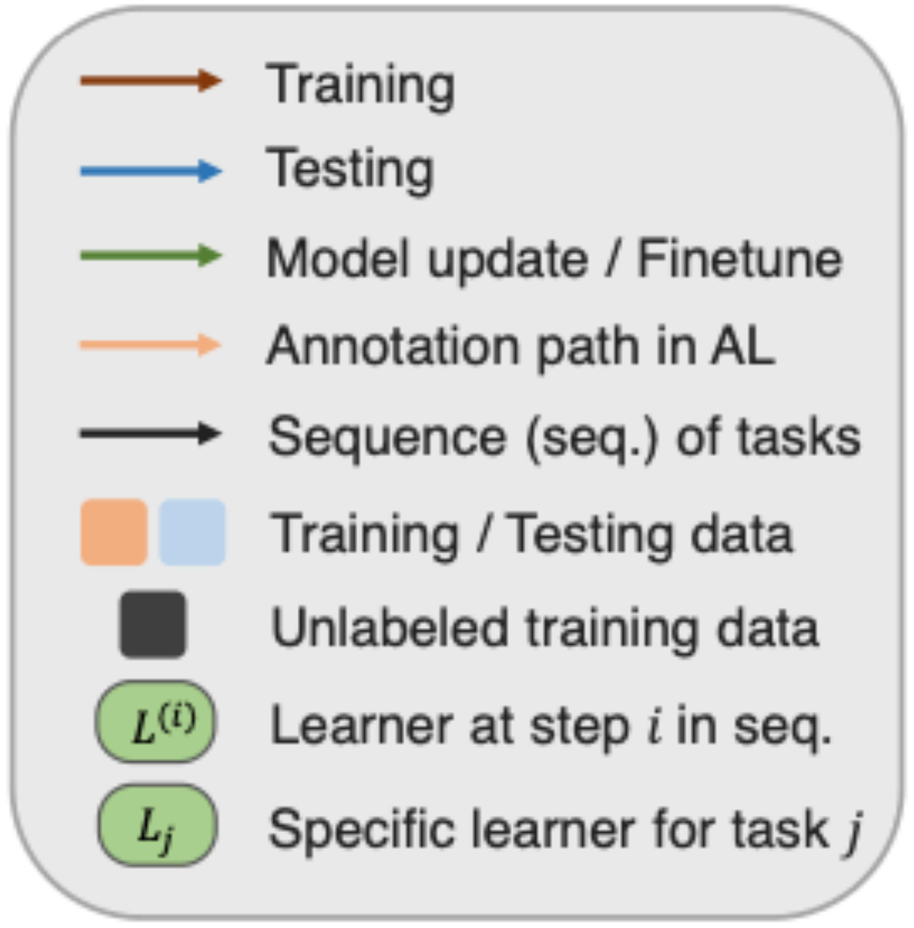


https://unconf.continualai.org/

- Free -> Get your registration :)
- Calls for pre-registered papers (already 31st), talks (mid of August), mentoring
- Multi-time zone & fully virtual

It’s about set-up & evaluation





Wang et al, “A Survey on Curriculum Learning”, TPAMI 2021

The types of tasks that are frequently considered

What if we don't know the boundary & aren't constrained to test examples?

What if future or unrelated data is in the test set?

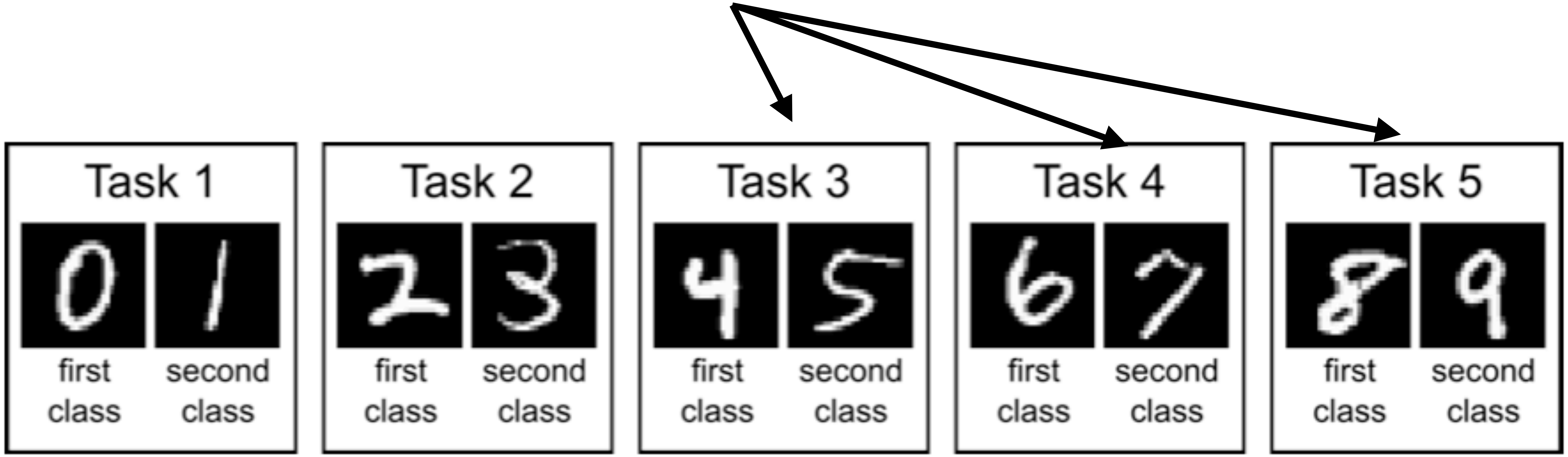


Figure 1: Schematic of split MNIST task protocol.

van de Ven et al, "Three types of incremental learning", Nature MI 2022

Challenge: the world is “open"

Challenge: the world is “open”

The threat of unknown unknowns

What do you think the prediction will be for a ML based classifier?

Challenge: the world is “open”

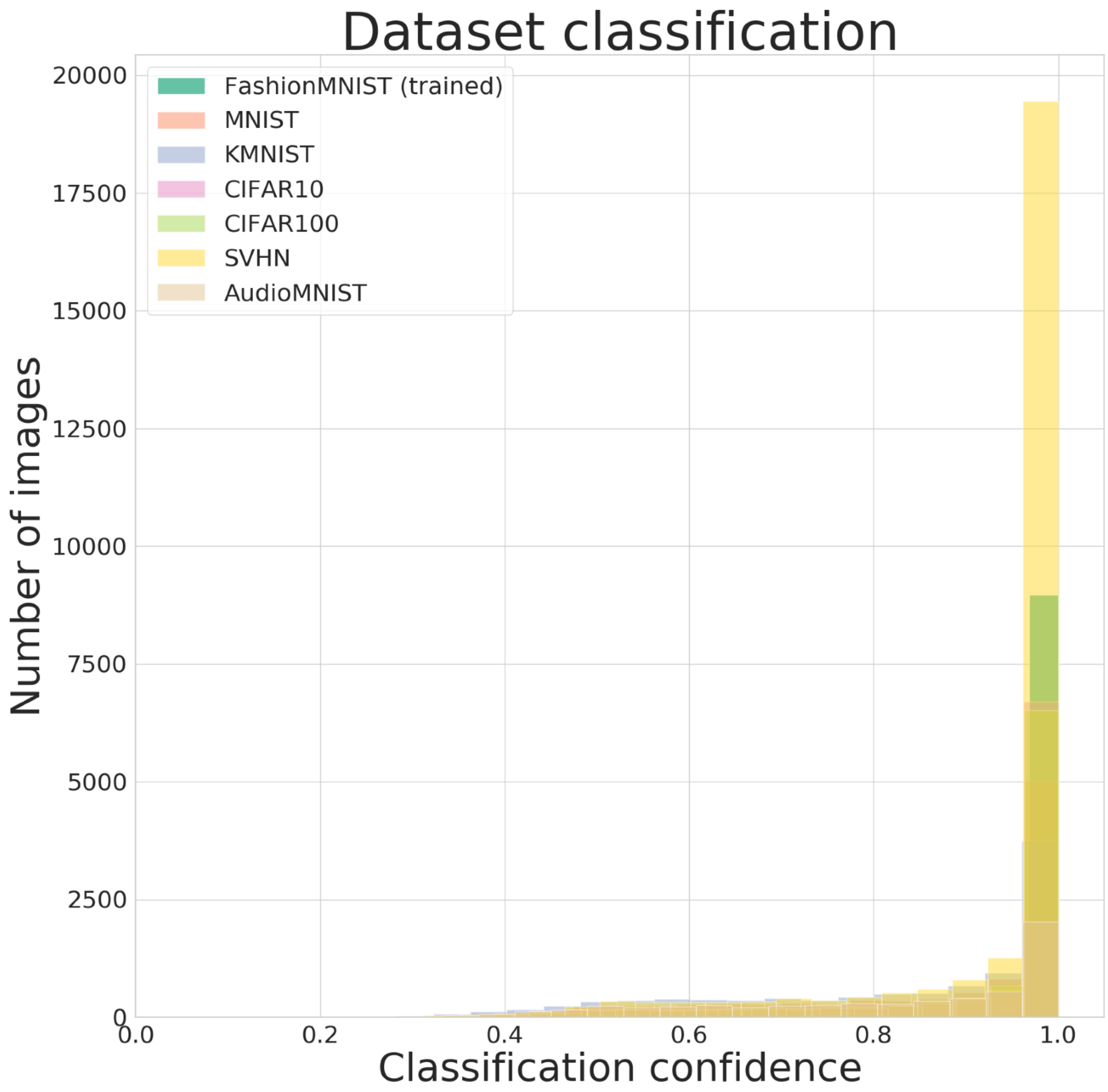
The threat of unknown unknowns

What do you think the prediction will be for a ML based classifier?

Most ML models are overconfident

"They don’t know when they don’t know”

Challenge: the world is "open"



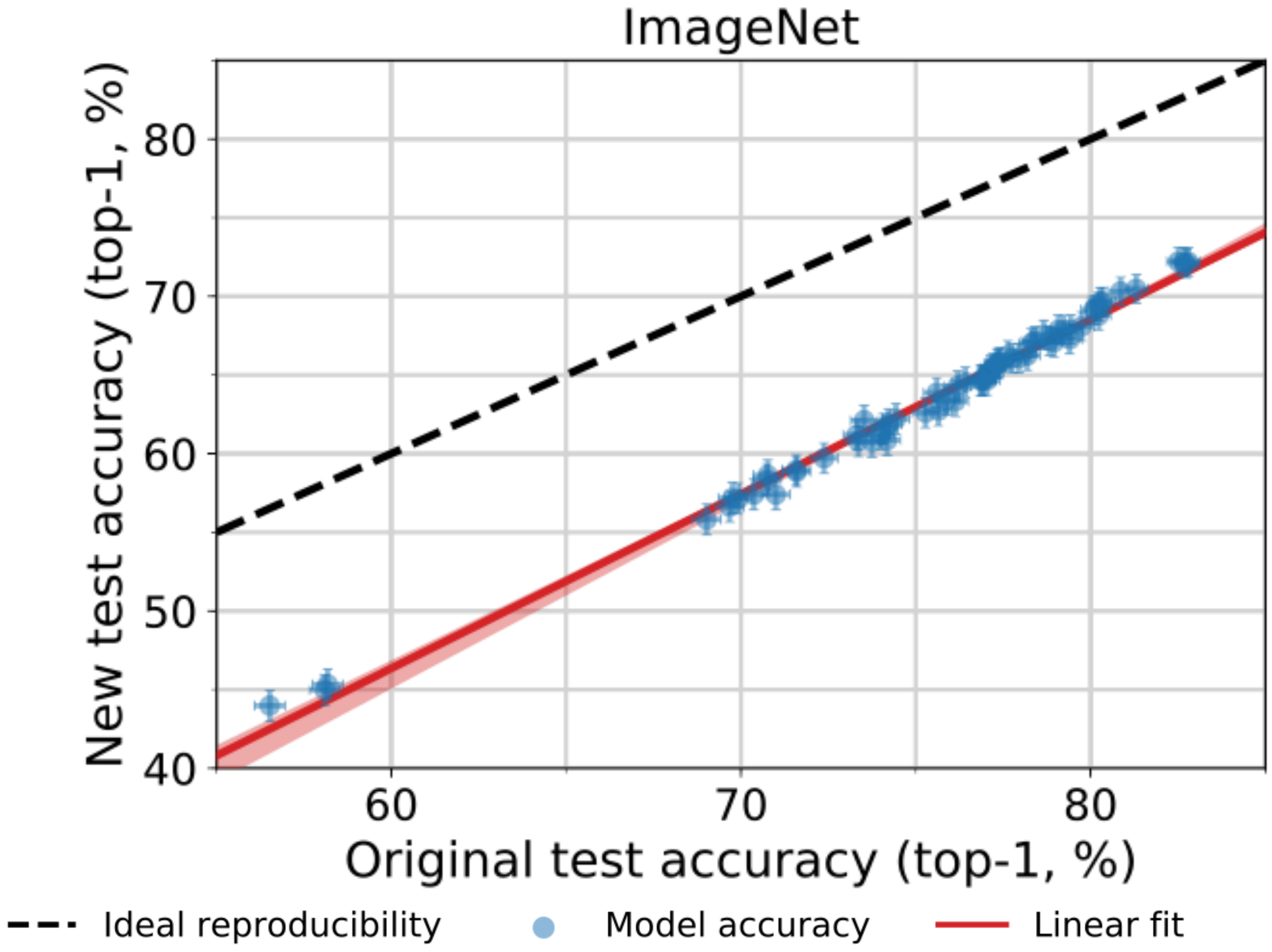
A quantitative example:

- Train a neural network classifier on a dataset (here fashion items)
- Log predictions for arbitrary other datasets
- Observe that majority of misclassifications happen with large output "probability"

Mundt et al "Open Set Recognition Through Deep Neural Network Uncertainty, Does Out-of-Distribution Detection Require Generative Classifiers?", ICCV Statistical Deep Learning Workshop 2019 (Based on a long-known problem, Matan1990)

“But this example is unrealistic in practice”!

Challenge: so many elements can shift



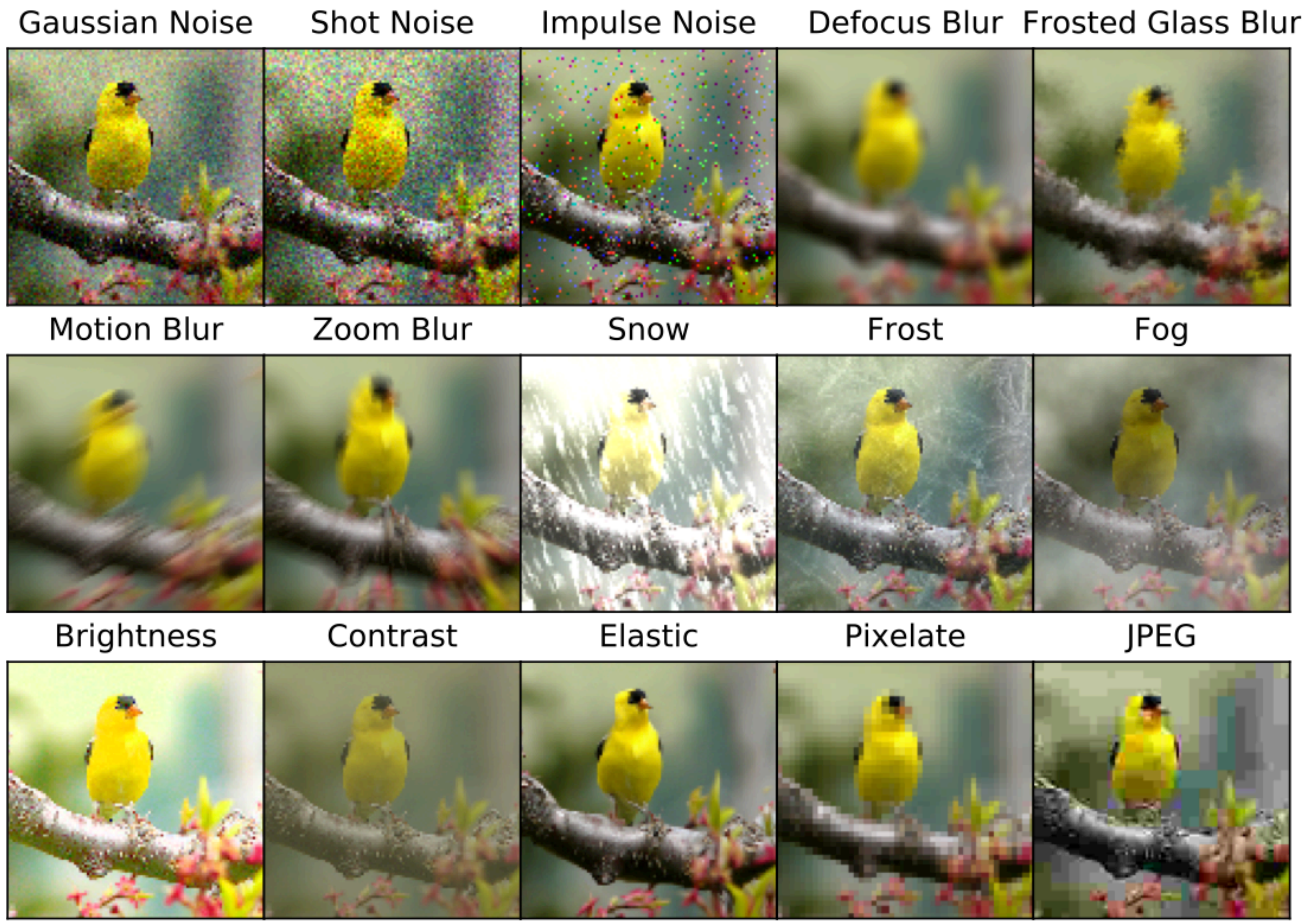
Recht et al, “Do ImageNet Classifiers Generalize to ImageNet?”, ICML 2019

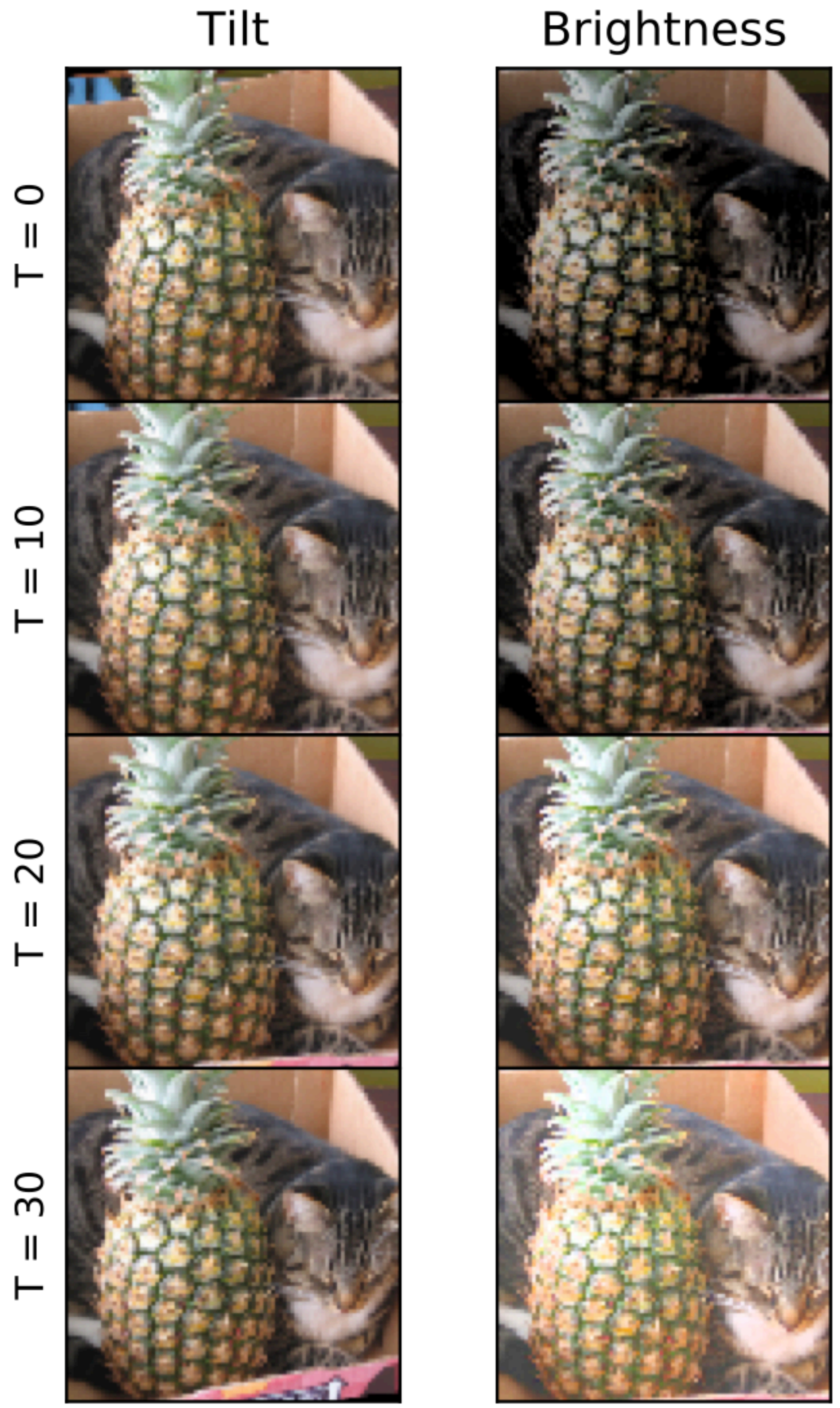
Performance loss even happens if we recollect another “test" set with the same instructions a 2nd time!

“Do ImageNet classifiers generalize to ImageNet?"

Challenge: so many elements can shift

Lots of natural perturbations & corruptions

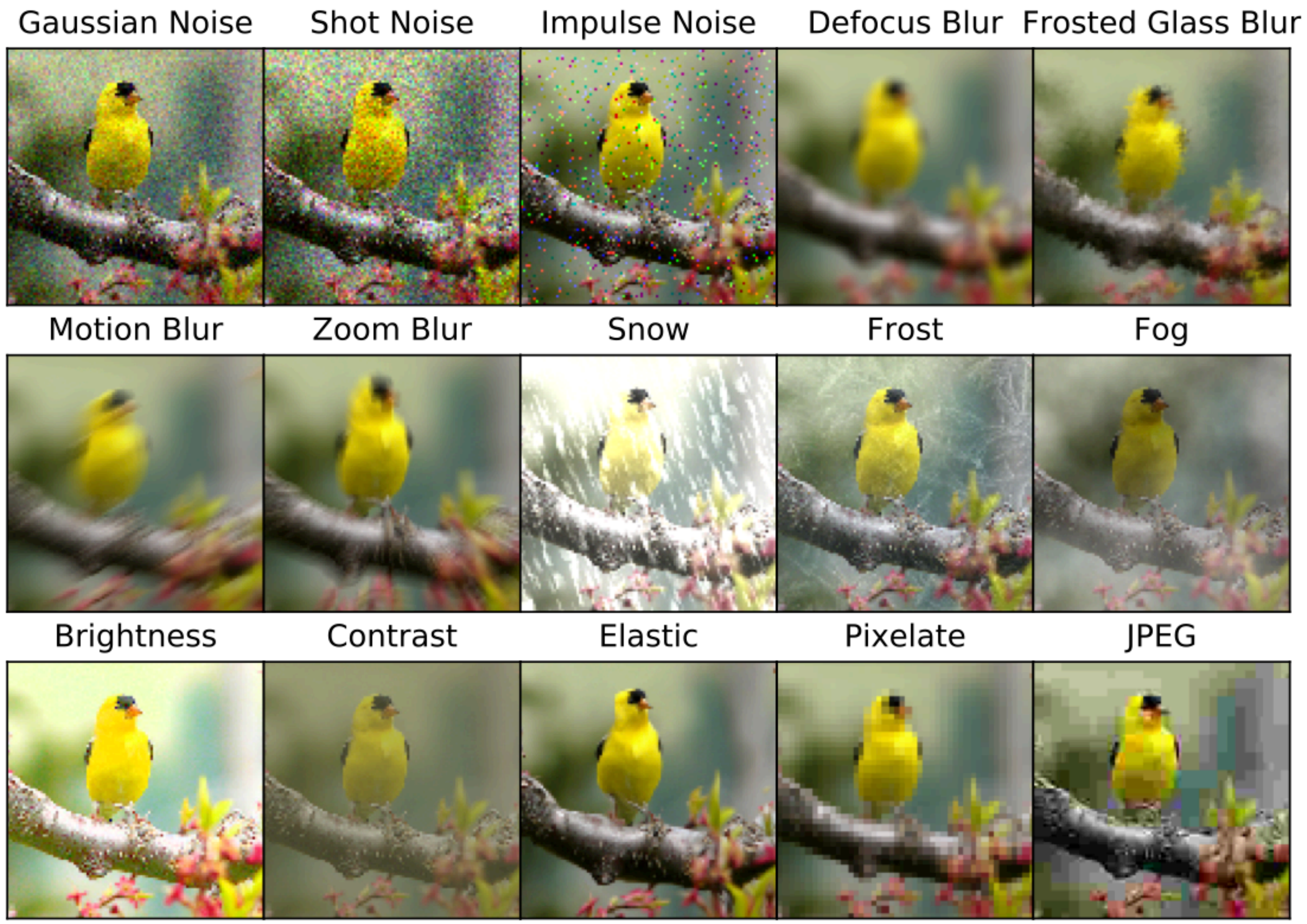


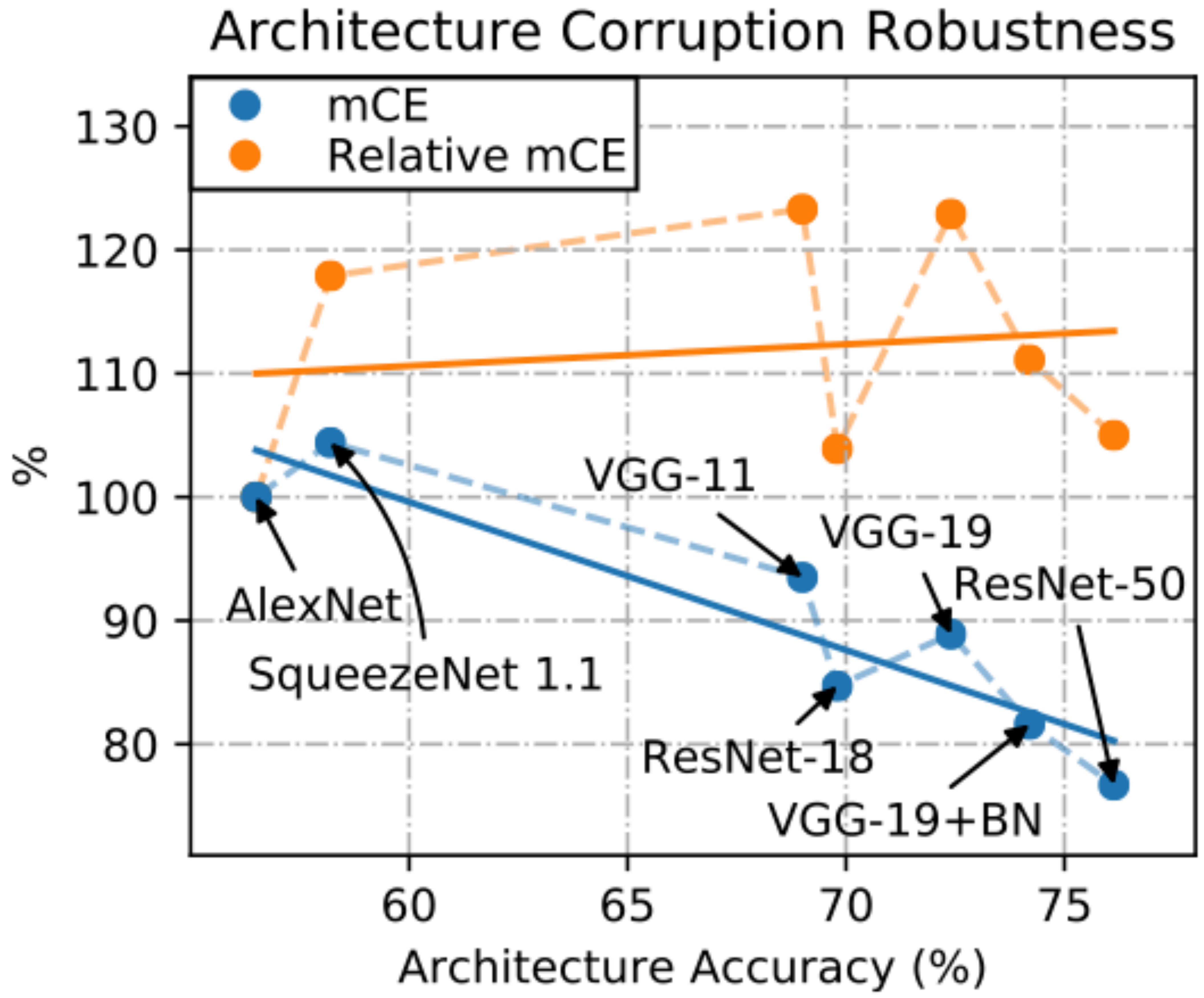


Hendricks & Dietterich, “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”, ICLR 2019

Accuracy in ImageNet has seemingly increased at the expense of robustness.

Lots of natural perturbations & corruptions

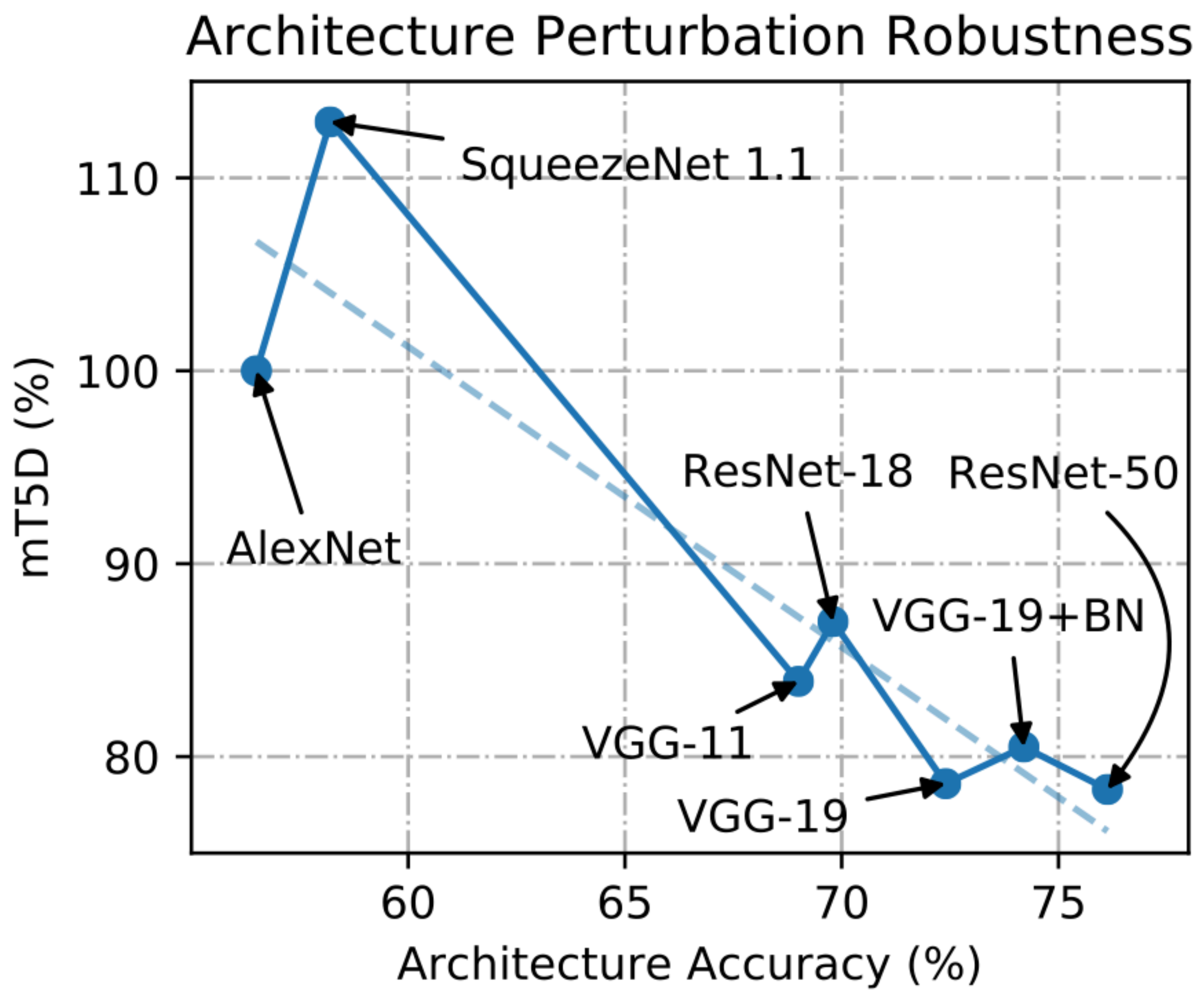


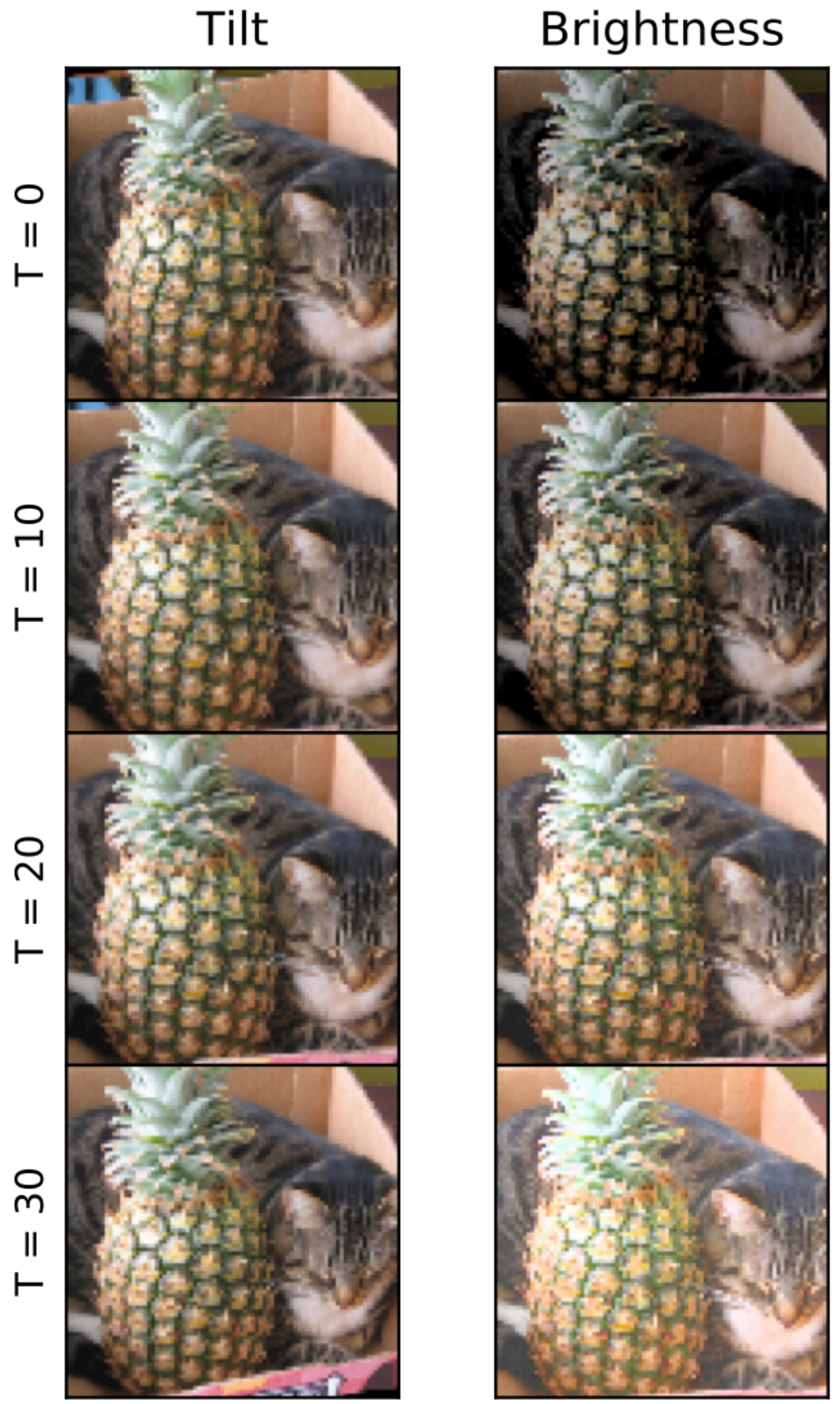


Hendricks & Dietterich, “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”, ICLR 2019

Accuracy in ImageNet has seemingly increased at the expense of robustness.

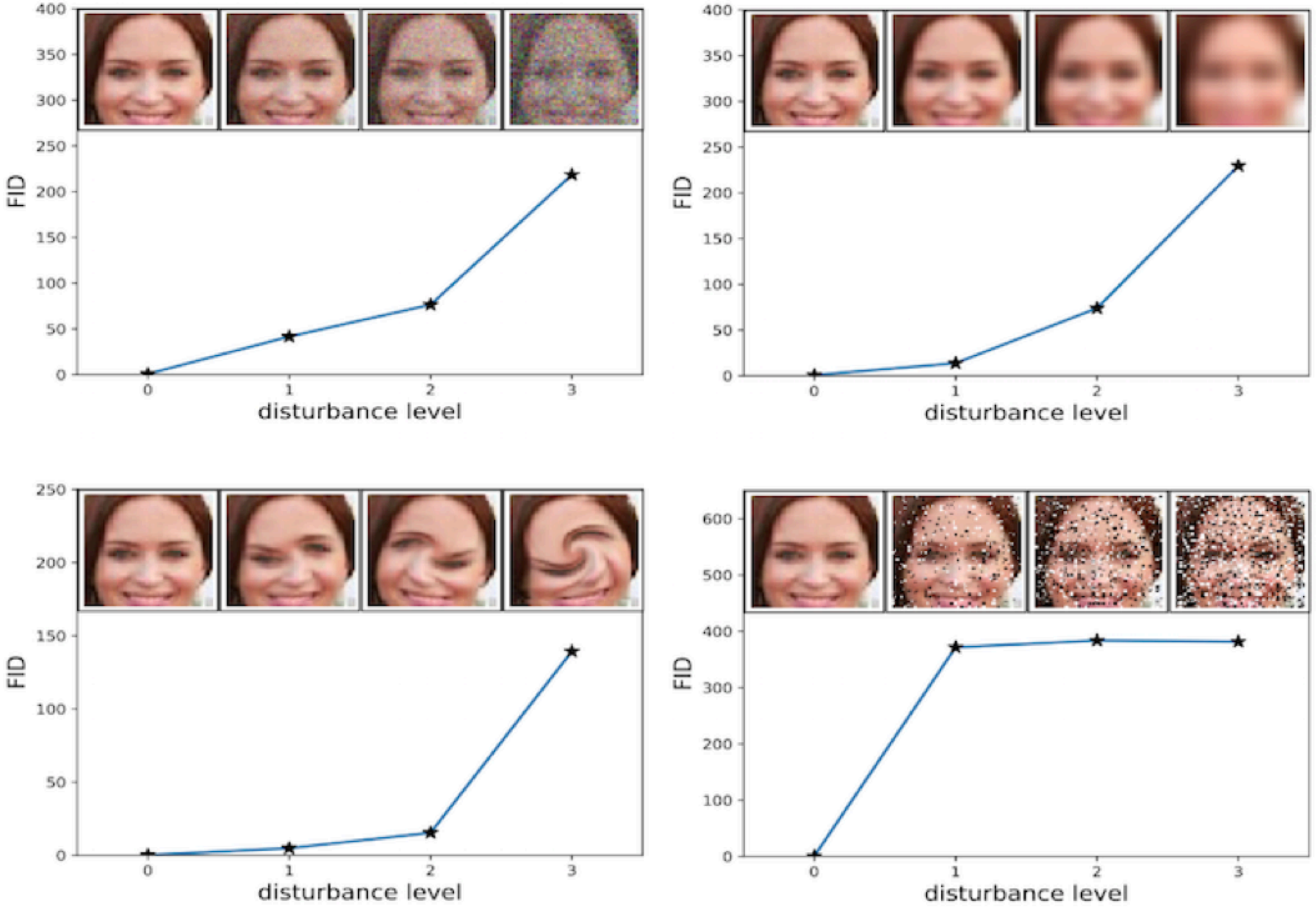
Lots of natural perturbations & corruptions





Hendricks & Dietterich, “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”, ICLR 2019

“Accuracy" in generation (FID) score, suffers from similar challenges with the way we typically measure



Recall: our losses & evaluation measures are often proxies for what we really want

Fréchet Inception Distance (FID) makes use of a pre-trained model to gauge generation "quality"

Ali Borji, "Pros and Cons of GAN Evaluation Measures”, 2018

Perspectives to address these challenges

More than known vs. unknown

1. **Known knowns:**

   From same distribution as train. Assumption: accurate & confident prediction.

2. **Known unknowns:**

3. **Unknown unknowns:**

4. **Unknown knowns:**

More than known vs. unknown

1. **Known knowns:**

   From same distribution as train. Assumption: accurate & confident prediction.

2. **Known unknowns:**

   Existing unknown “non-"examples or examples with high uncertainty.

3. **Unknown unknowns:**

4. **Unknown knowns:**

More than known vs. unknown

1. **Known knowns:**

   From same distribution as train. Assumption: accurate & confident prediction.

2. **Known unknowns:**

   Existing unknown “non-"examples or examples with high uncertainty.

3. **Unknown unknowns:**

   Unseen instances belonging to unexplored & unknown data distributions.

4. **Unknown knowns:**

More than known vs. unknown

1. **Known knowns (*or simply knowns*):**

   From same distribution as train. Assumption: accurate & confident prediction.

2. **Known unknowns:**

   Existing unknown “non-"examples or examples with high uncertainty.

3. **Unknown unknowns:**

   Unseen instances belonging to unexplored & unknown data distributions.

4. **Unknown knowns:**

   Usually not considered: known concept but choose to treat it as unknown (willful ignorance?) or our ML system cannot represent the concept + structure altogether

What do you think: how can we solve our challenge?

Three categories of approaches

Anomalies in predictions:

The unsuspecting angle, where out-of-distribution are hopefully separable through anomalous output values

Incorporating prior knowledge:

The intuitive idea to include “background” or “non-example” data population explicitly.

Open Set recognition:

The more formal approach ensures that we only rely on predictions from our “covered space”; we create bounds.

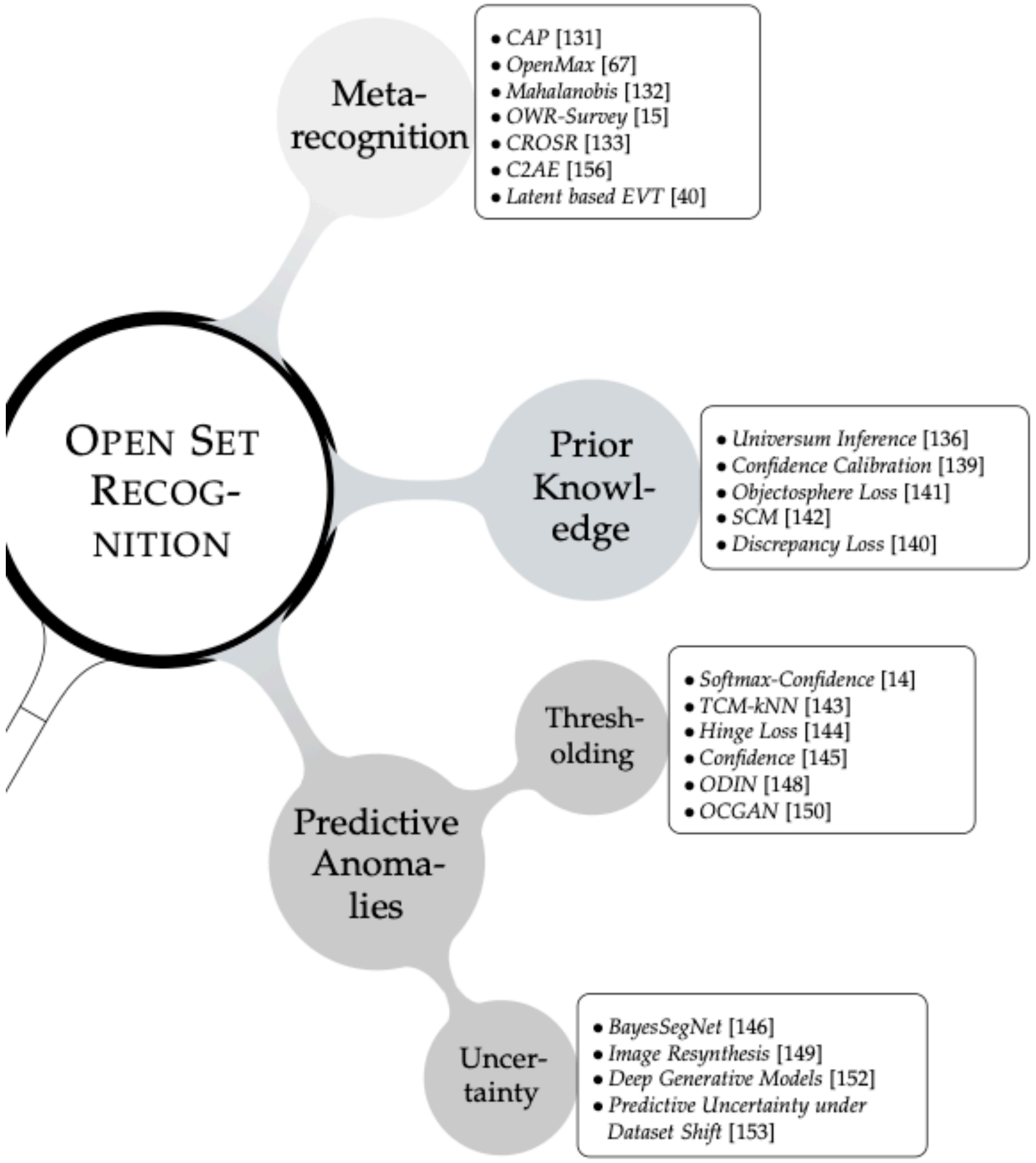


Figure from “A Wholistic View of Deep Neural Networks: Forgotten Lessons and the Bridge to Active and Open World Learning”, Mundt et al, Neural Networks, 2023

Predictive anomalies:
the unfortunate part of the story

Disclaimer: I'll use many figures from our papers for convenience, without trying to imply that we discovered these phenomena

Overconfidence & uncertainty

Unfortunately uncertainty is not a necessarily a “fix”

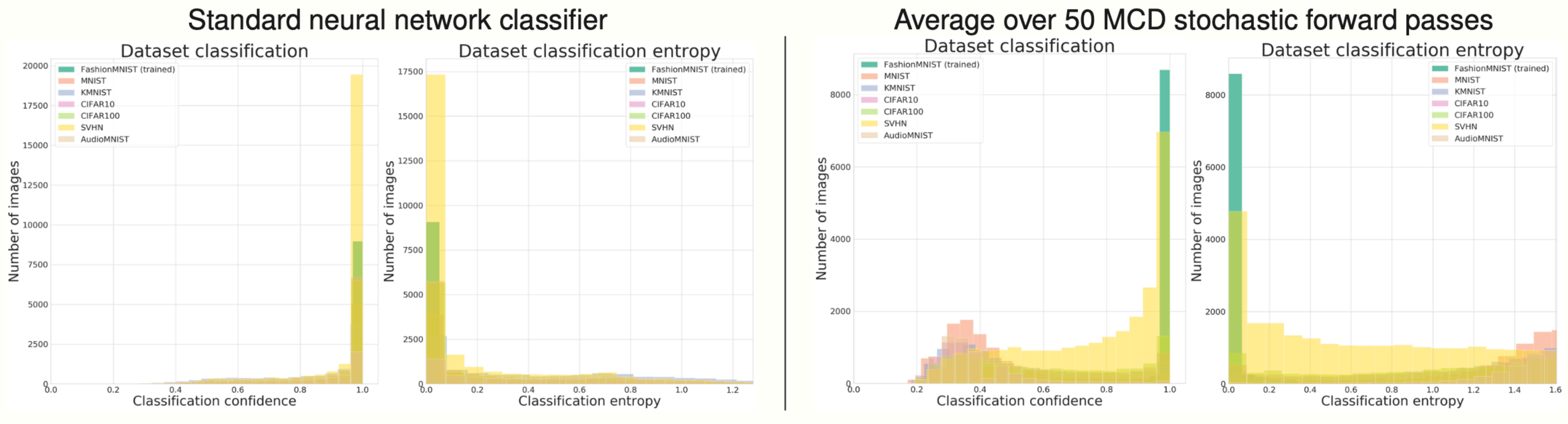


Figure from Mundt et al “Open Set Recognition Through Deep Neural Network Uncertainty, Does Out-of-Distribution Detection Require Generative Classifiers?”, ICCVW 2019

Overconfidence & gen. models

It get’s even harder when we try to select a threshold

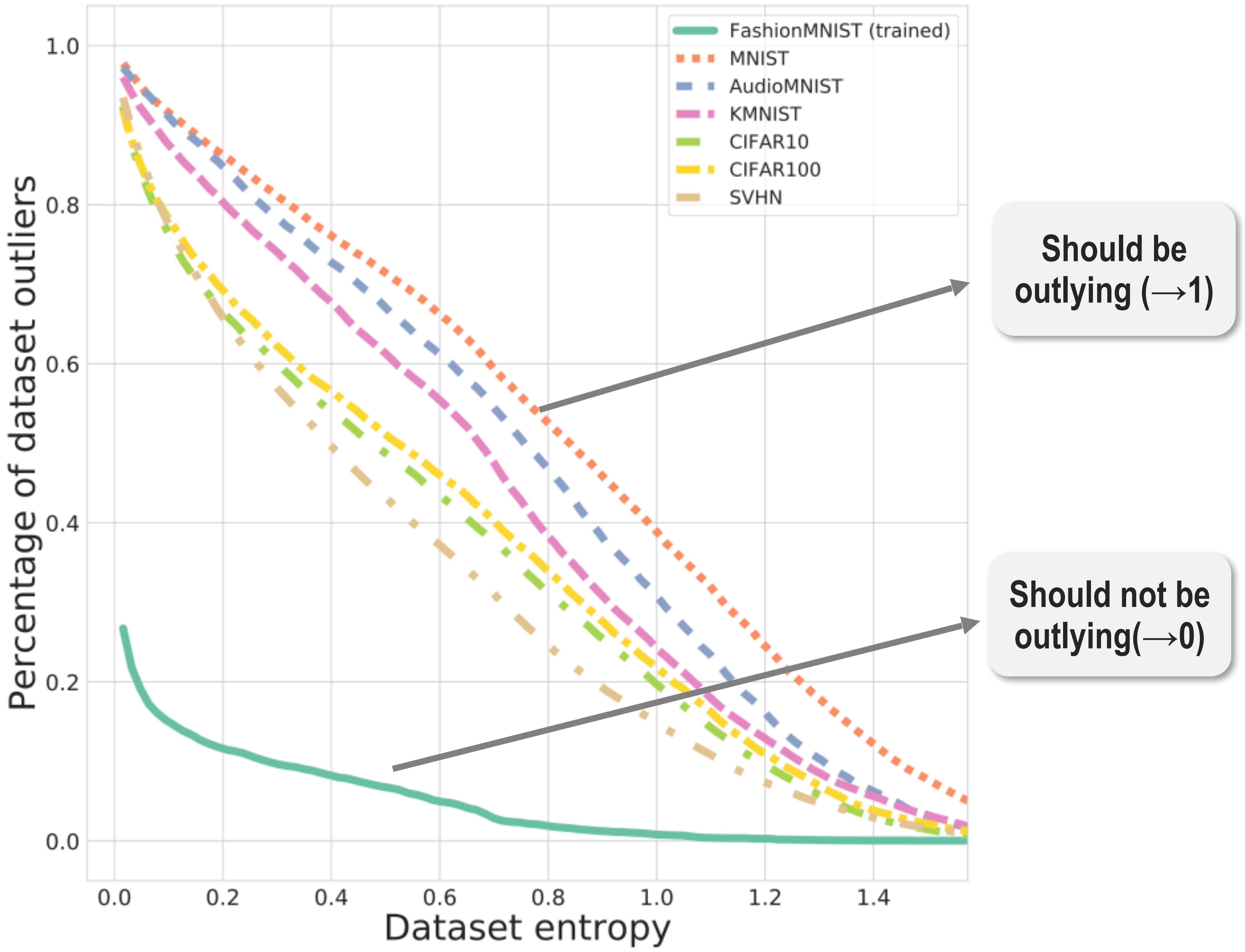


Figure from Mundt et al, “Unified Probabilistic Deep Continual Learning Through Open Set Recognition and Generative Replay”, Journal of Imaging, Volume 8, Issue 4, 2022

Overconfidence & uncertainty

It get’s even harder when we try to select a threshold

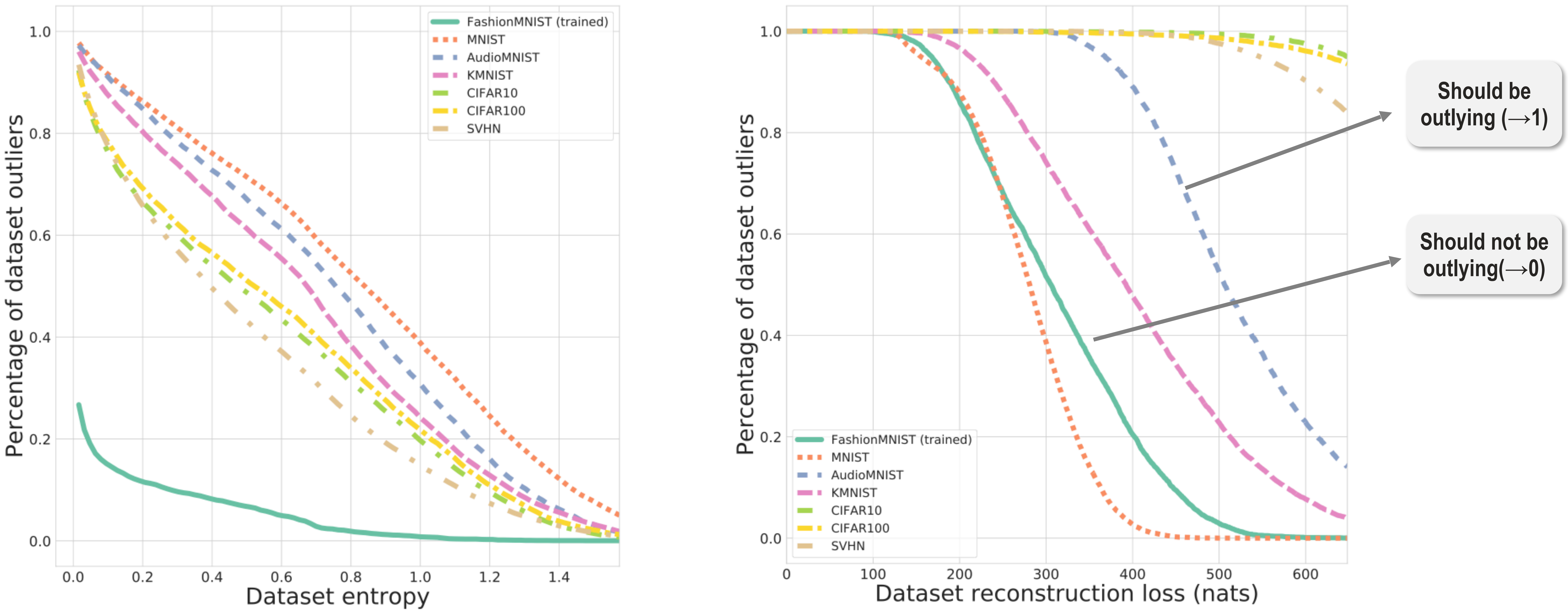
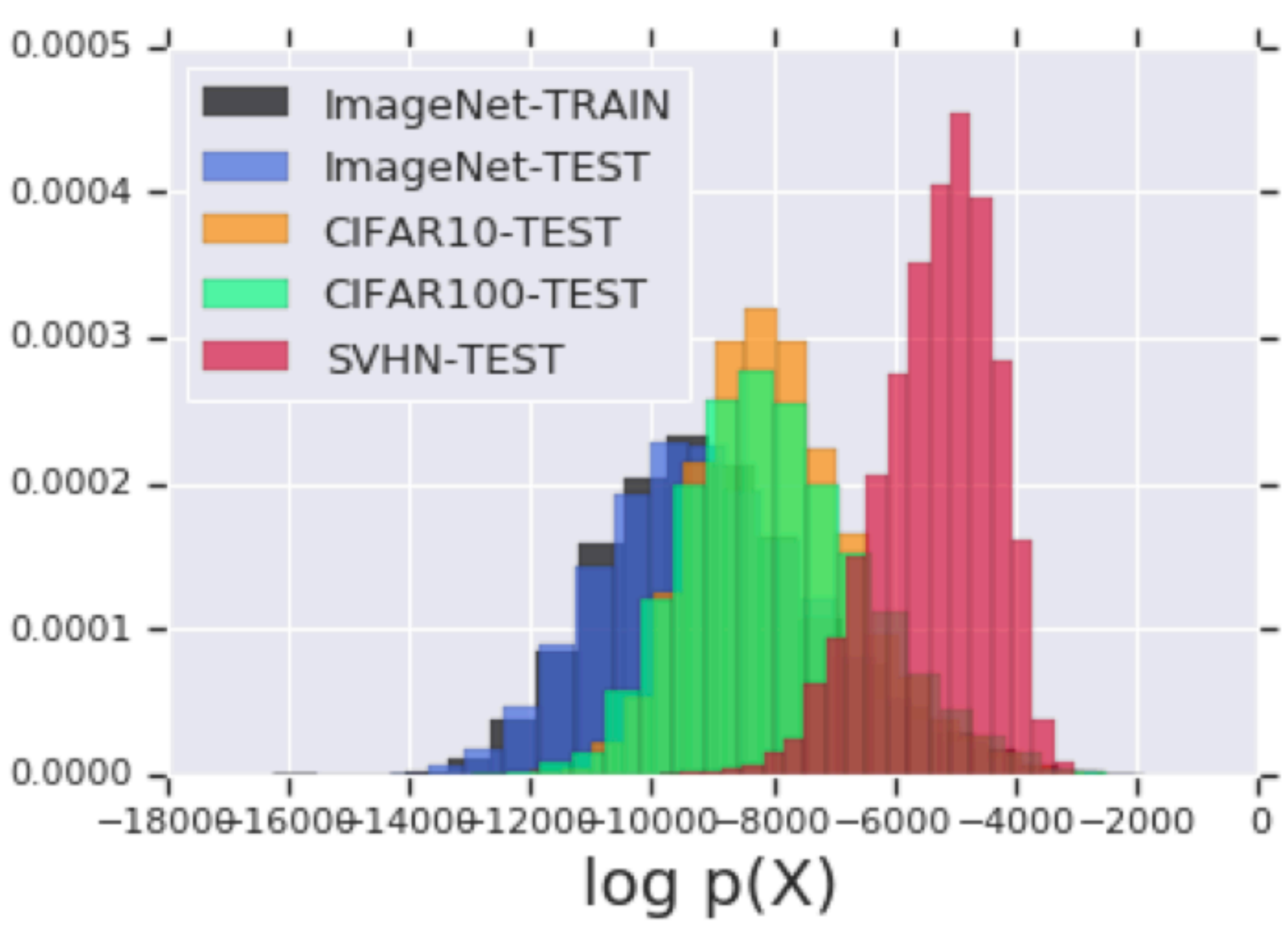


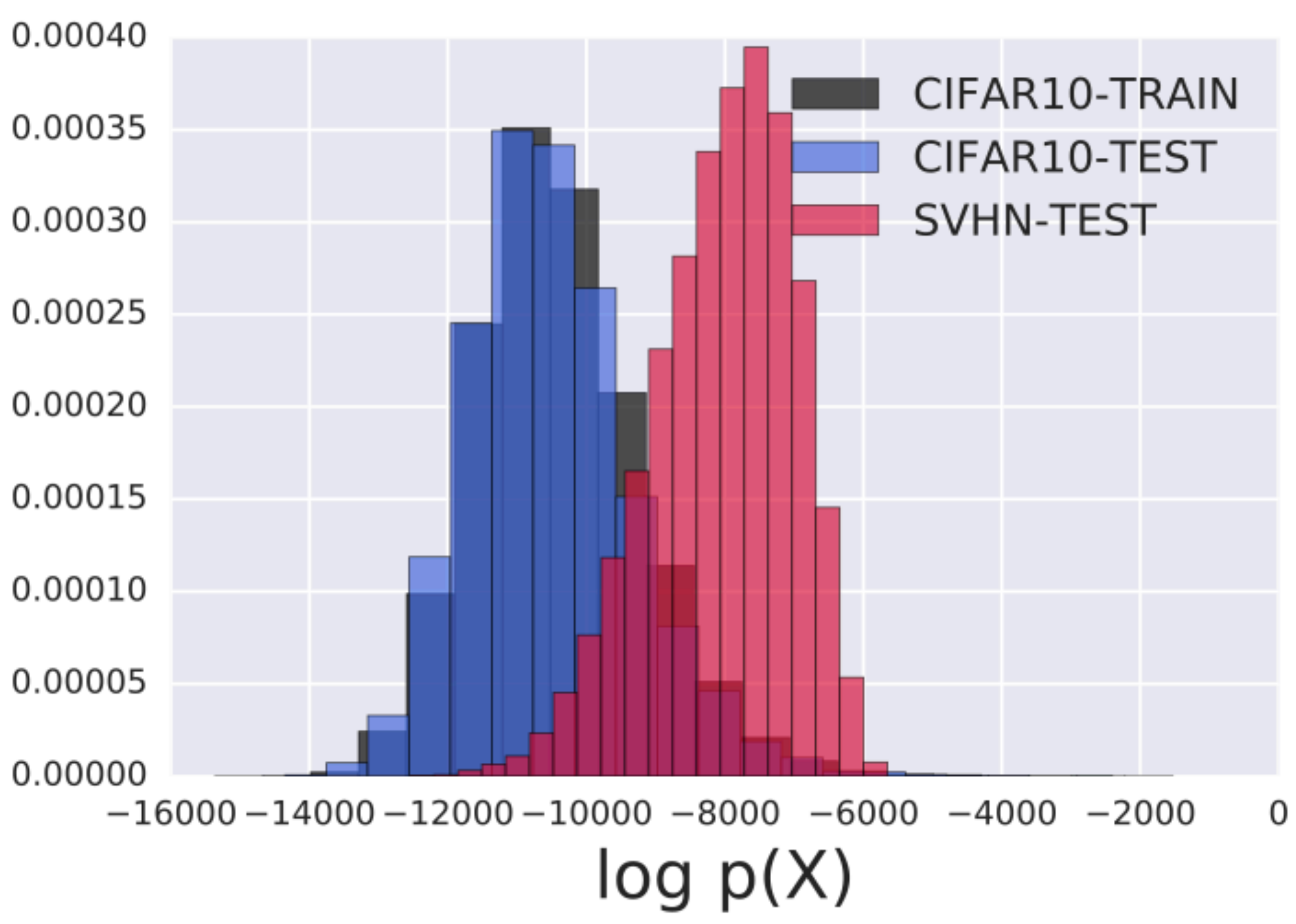
Figure from Mundt et al, “Unified Probabilistic Deep Continual Learning Through Open Set Recognition and Generative Replay”, Journal of Imaging, Volume 8, Issue 4, 2022

Overconfidence & gen. models

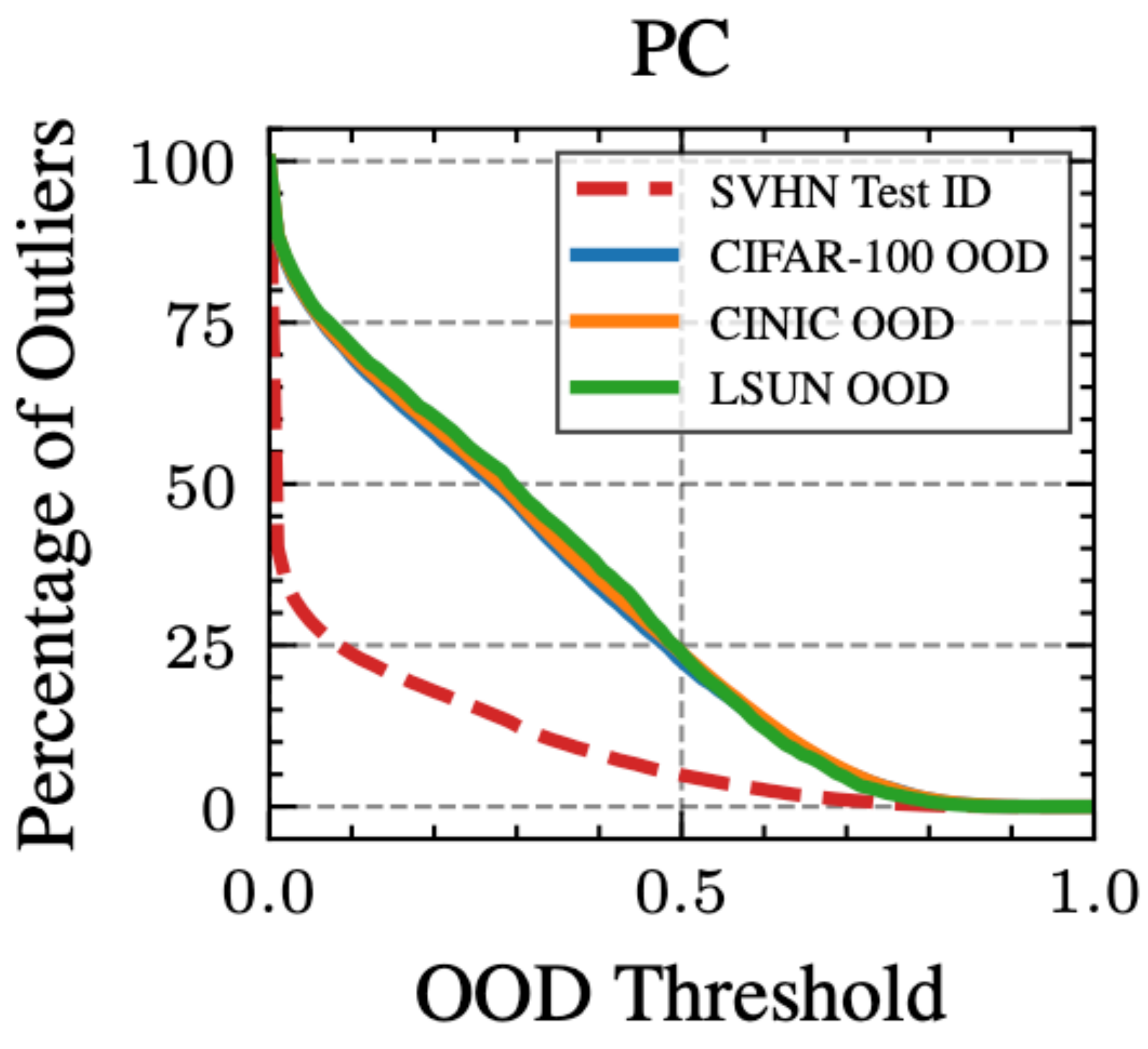
Overconfidence is not exclusive to discriminative models



Glow



PixelCNN



Probabilistic Circuit

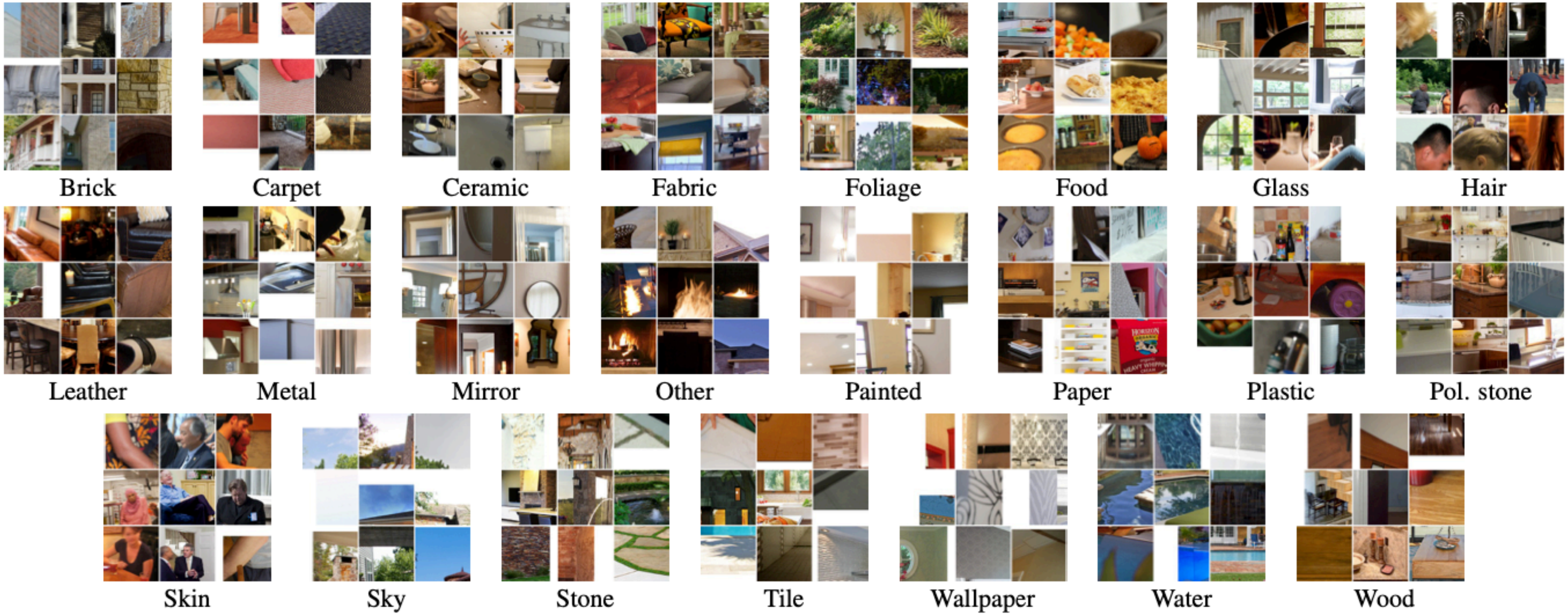
Nalisnick et al, “Do Deep Generative Models Know What They Don’t Know”, ICLR 2019

Ventola et al, UAI 2023. “Do Probabilistic Circuits Know What They Don’t Know”?

Including prior knowledge: an alternative?

The intuitive idea

Take a look at the Materials in Context (MINC) dataset: what do you notice?



Bell & Upchurch et al, “Material Recognition in the Wild with the Materials in Context Database”, CVPR 2015

The intuitive idea

Take a look at the Materials in Context (MINC) dataset: what do you notice?

Bell & Upchurch et al, "Material Recognition in the Wild with the Materials in Context Database", CVPR 2015

Inference with the universum

In essence: include **“non-examples”** that aren’t of interest but are available

(Some) key questions:

- How to implement the loss: many many conceivable conceivable
- “What part of the universum is useful” (“Inference with the universum”, Weston et al, ICML 2006)
- "What are we expected to see during prediction later”?

  (Noise? Other concepts? Etc.)

Calibration: some examples

1. We could let our predictions follow a uniform distribution for “out” data

(Kimin Lee et al, “Training confidence-calibrated classifiers for detecting out-of-distribution samples”, ICLR 2018)

$$\min_{\theta} \mathbb{E}_{P_{\mathrm{in}}(\widehat{\mathbf{x}},\widehat{y})}\left[-\log P_{\theta}\left(y=\widehat{y}|\widehat{\mathbf{x}}\right)\right]+\beta\mathbb{E}_{P_{\mathrm{out}}(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right)\parallel P_{\theta}\left(y|\mathbf{x}\right)\right)\right]$$

Calibration: some examples

1. We could let our predictions follow a uniform distribution for “out” data

(Kimin Lee et al, “Training confidence-calibrated classifiers for detecting out-of-distribution samples”, ICLR 2018)

$$\min_{\theta} \mathbb{E}_{P_{\text{in}}(\widehat{\mathbf{x}},\widehat{y})}\left[-\log P_\theta\left(y=\widehat{y}|\widehat{\mathbf{x}}\right)\right] + \beta\mathbb{E}_{P_{\text{out}}(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y|\mathbf{x}\right)\right)\right]$$

2. We could predict an “out” category or generally maximize uncertainty

Calibration: some examples

1. We could let our predictions follow a uniform distribution for "out" data

   (Kimin Lee et al, "Training confidence-calibrated classifiers for detecting out-of-distribution samples", ICLR 2018)

$$\min_{\theta} \mathbb{E}_{P_{\text{in}}(\widehat{\mathbf{x}},\widehat{y})}\left[-\log P_\theta\left(y=\widehat{y}|\widehat{\mathbf{x}}\right)\right] + \beta \mathbb{E}_{P_{\text{out}}(\mathbf{x})}\left[KL\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y|\mathbf{x}\right)\right)\right]$$

2. We could predict an "out" category or generally maximize uncertainty

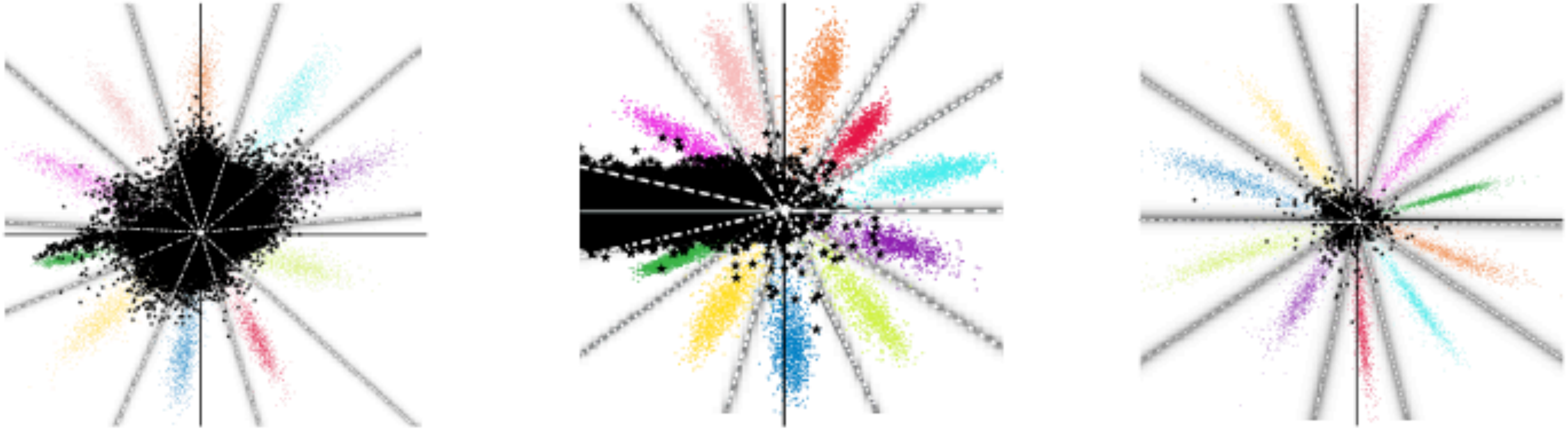
3. And many other versions to modify our loss to do something with "out",

   e.g. (Dhamija et al, "Reducing network agnostophobia", NeurIPS 2018)

$$J_E(x) = \begin{cases} -\log S_c(x) & \text{if } x \in \mathcal{D}'_c \text{ is from class } c \\ -\frac{1}{C}\sum_{c=1}^{C} \log S_c(x) & \text{if } x \in \mathcal{D}'_b \end{cases}$$

Background & Objectosphere

We could also construct variants for features/activations etc. to be zero



(a) Softmax (b) Background (c) Objectosphere

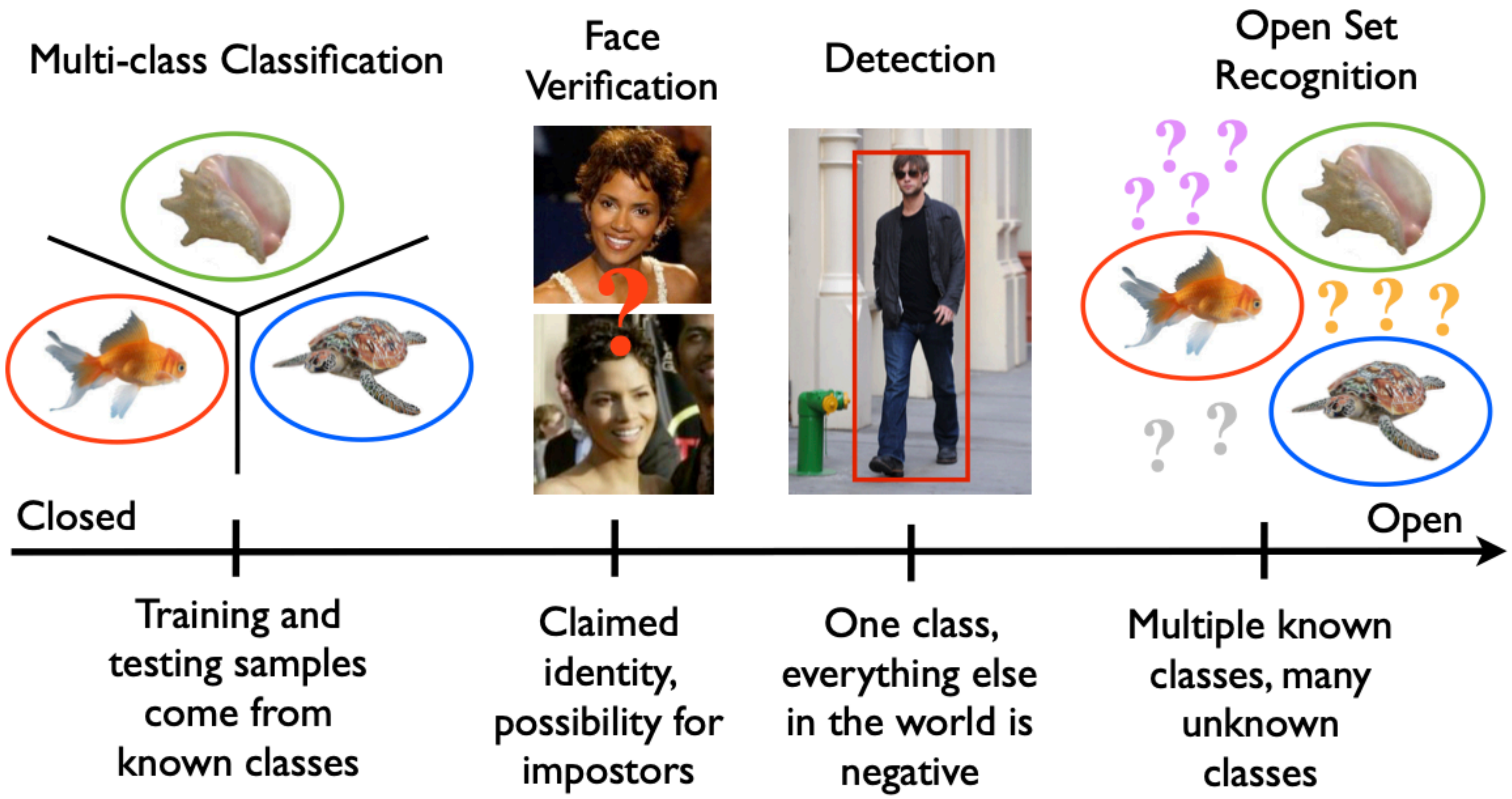
Figure 1: LENET++ RESPONSES TO KNOWNS AND UNKNOWNS. *The network in (a) was only trained to classify the 10 MNIST classes ($\mathcal{D}'_c$) using softmax, while the networks in (b) and (c) added NIST letters [15] as known unknowns ($\mathcal{D}'_b$) trained with softmax or our novel Objectosphere loss.*

Dhamija et al, “Reducing Network Agnostophobia”, NeurIPS 2018

What do you think are the up & downsides so far?

Closed to open world assumption

As the world grows more "open" we move from known unknowns to unknown unknowns. Our two perspectives only handle the former



Scheirer et al, "Towards Open Set Recognition", TPAMI 2012

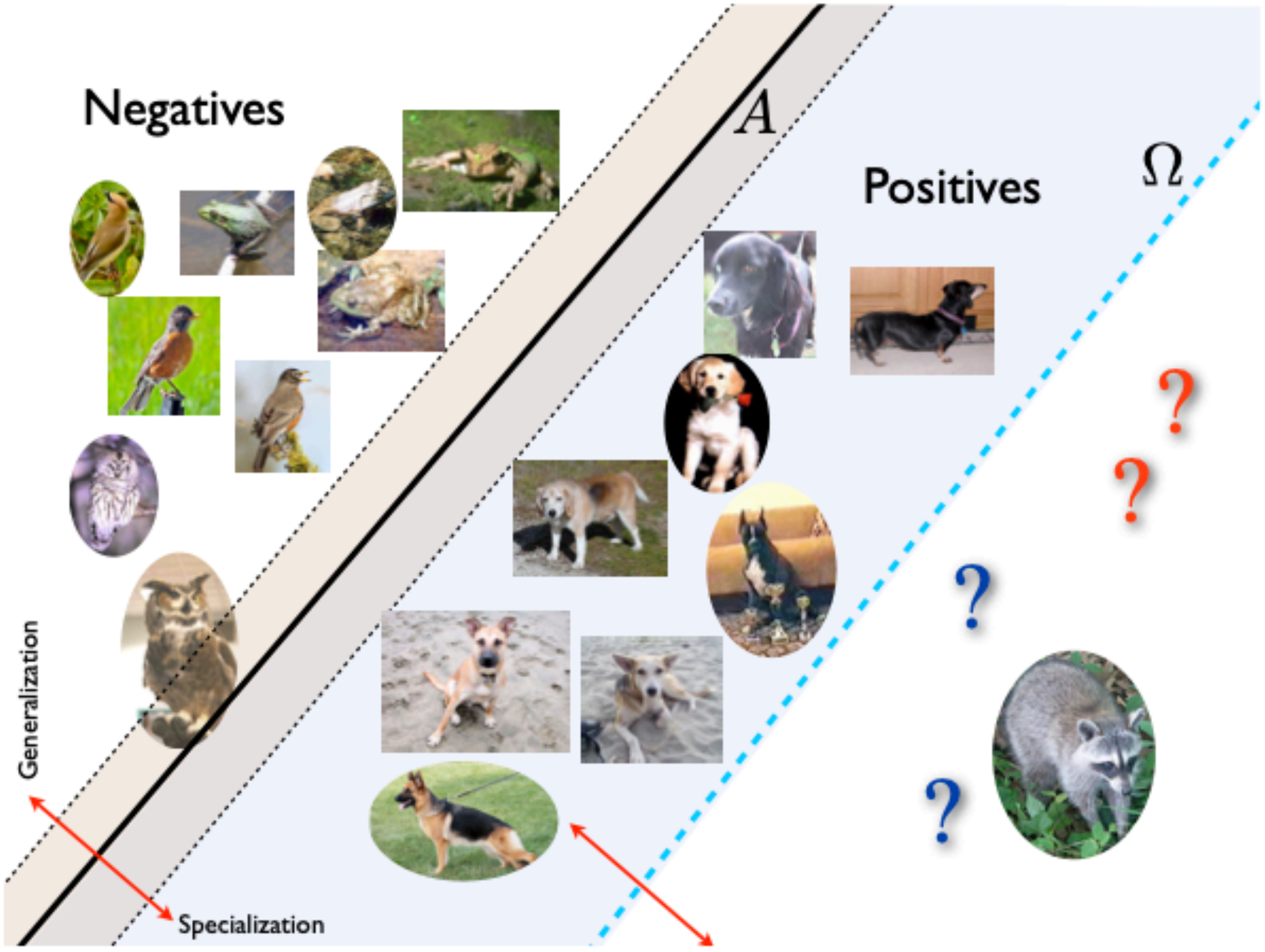
Open set recognition & explicit bounds

Intuition behind open space

Intuitively: take into account distances from known data points

SVM example: fit another parallel plane to reject, based on support set with large distances

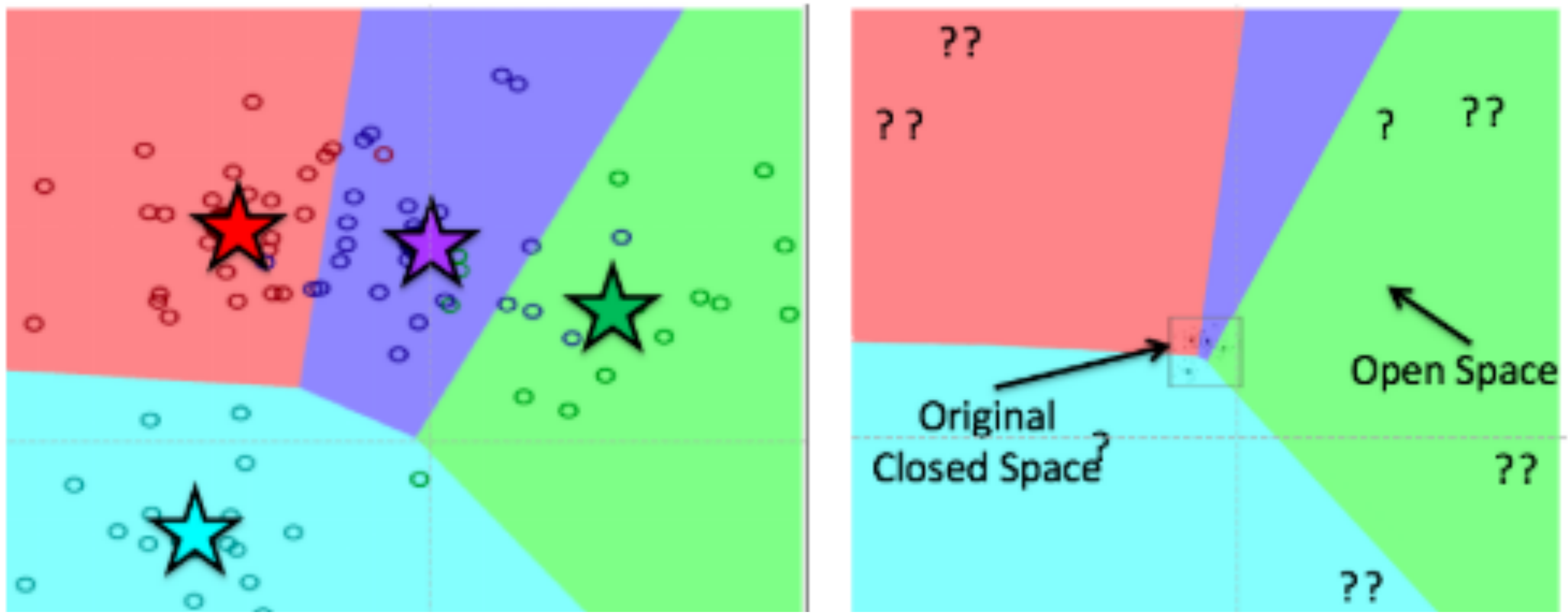
"Don't know & should not predict"



Scheirer et al, "Towards Open Set Recognition", TPAMI 2012

Formalizing open space/sets

Intuitively: open space is what is not covered with known data



"Learning and the Unknown", Boult et al, AAAI 2019

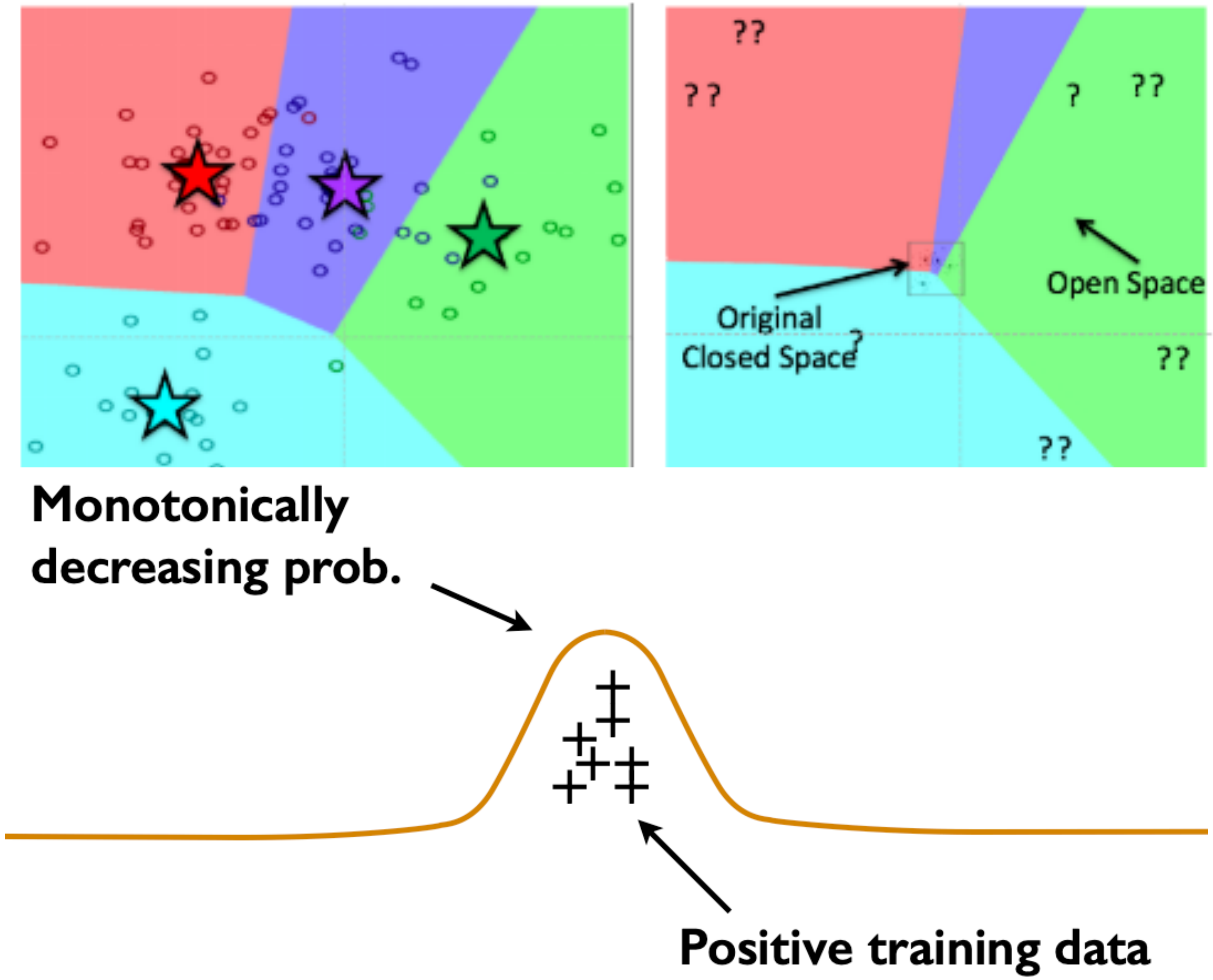
Scheirer et al, "Probability Models for Open Set Recognition", TPAMI 2014

Formalizing open space/sets

Intuitively: open space is what is not covered with known data

Formally: For a recognition function f over space $\mathscr{X}$ & a union of balls with radius r that includes all known training examples:

$$\mathcal{O} = \mathscr{X} - \cup_{i \in N} B_r(x_i)$$



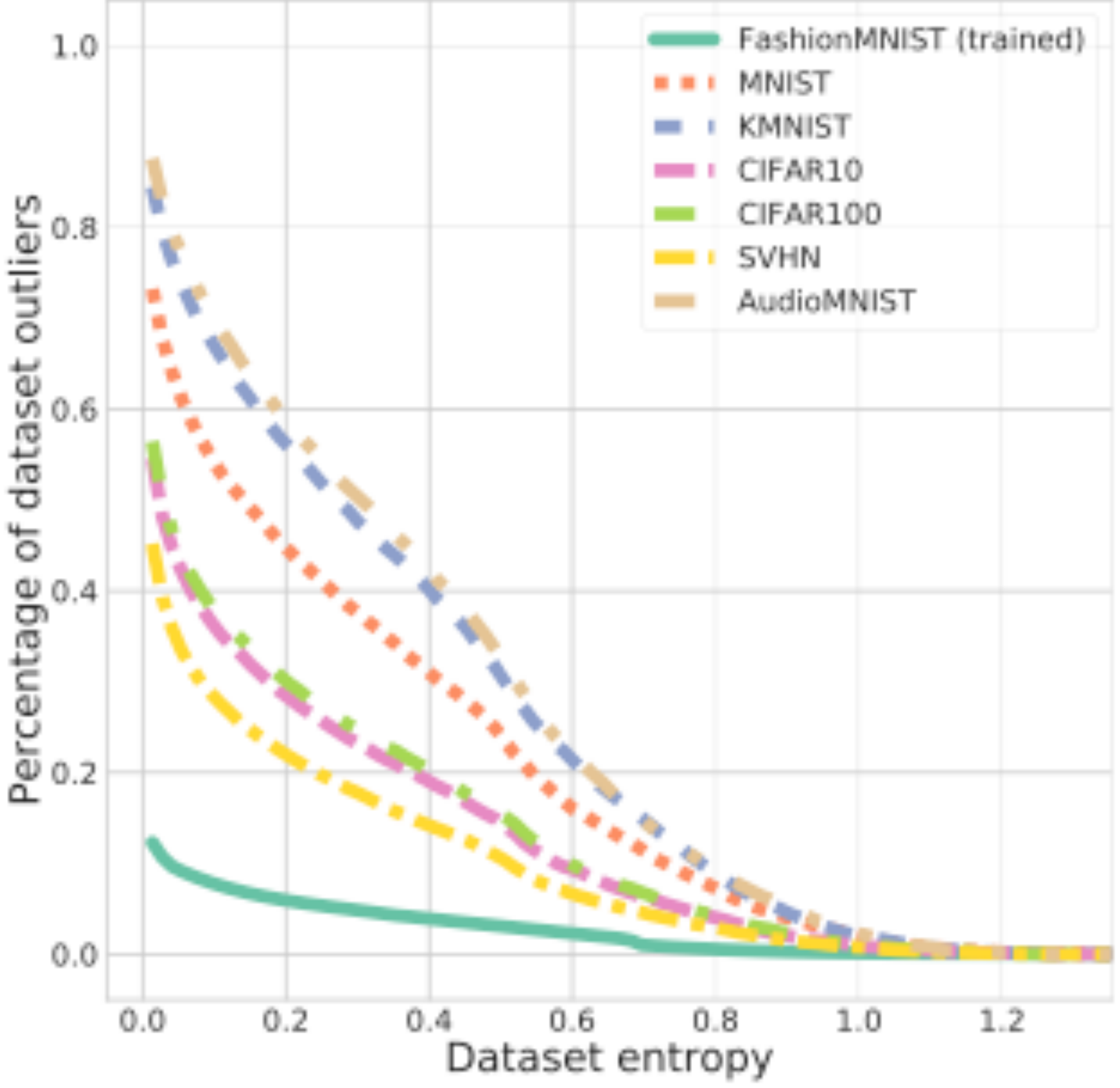
“Learning and the Unknown”, Boult et al, AAAI 2019

Scheirer et al, “Probability Models for Open Set Recognition”, TPAMI 2014

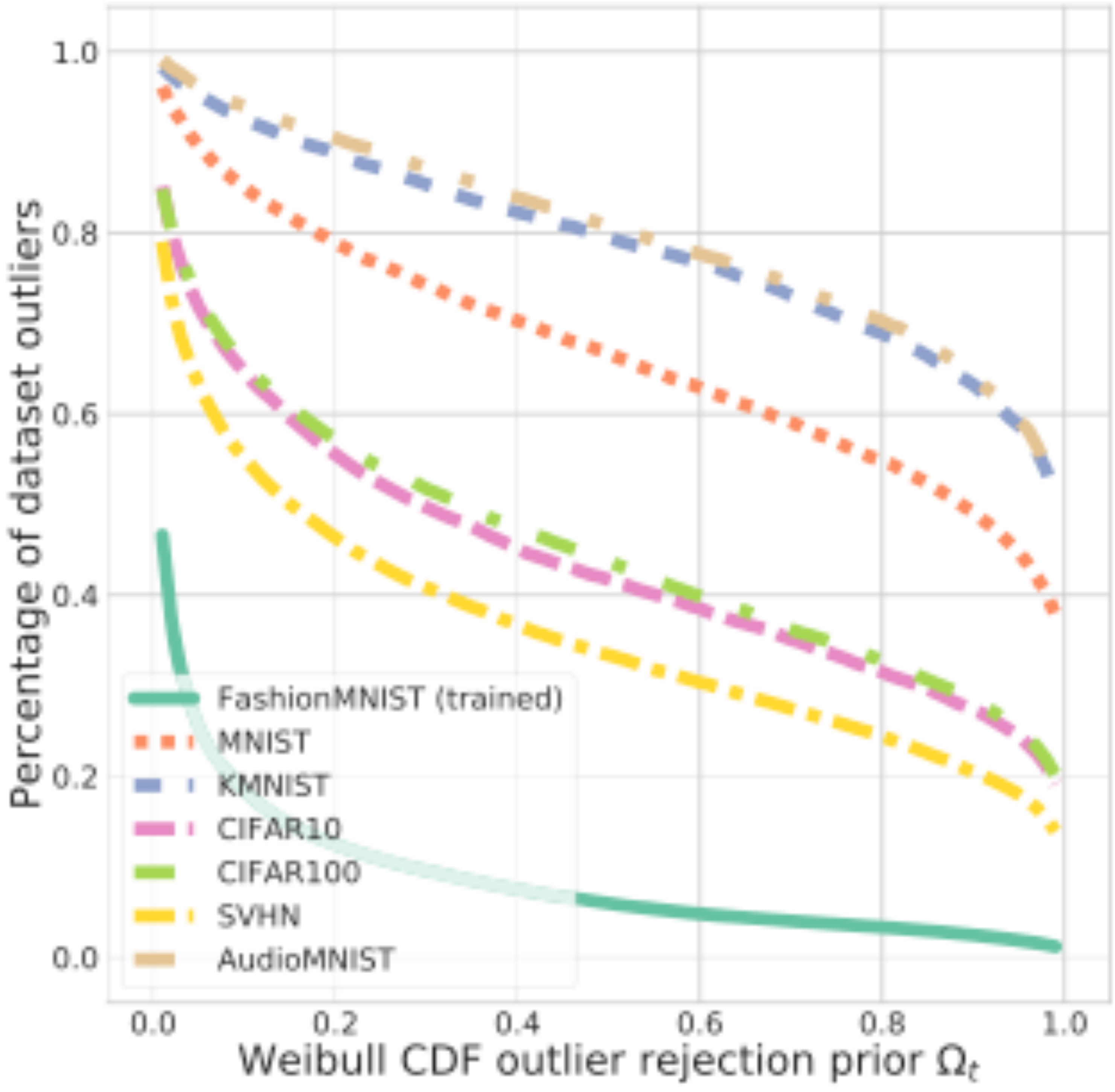
Some system examples that follow this intuition

There exist systems that use this idea, e.g. by extreme observed value fits

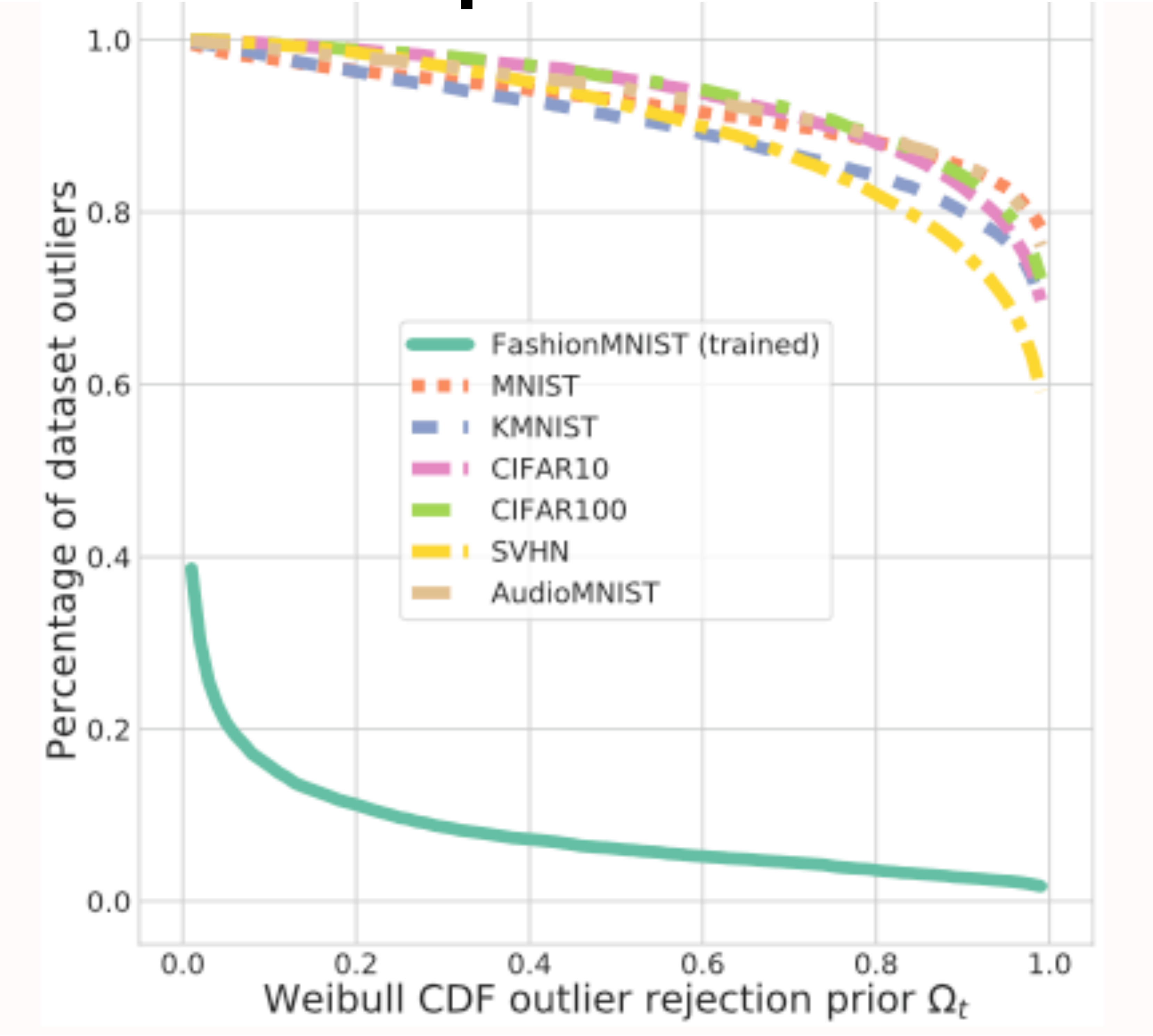
"Standard Model"



"OpenMax"



"OpenVAE"



Bendale & Boult et al, "Towards Open Set Deep Networks", CVPR 2016

Mundt et al, "Unified Probabilistic Deep Continual Learning Through Open Set Recognition and Generative Replay", Journal of Imaging 8:4, 2022

Open world learning: combining ideas

Open world learning: set-up & evaluation

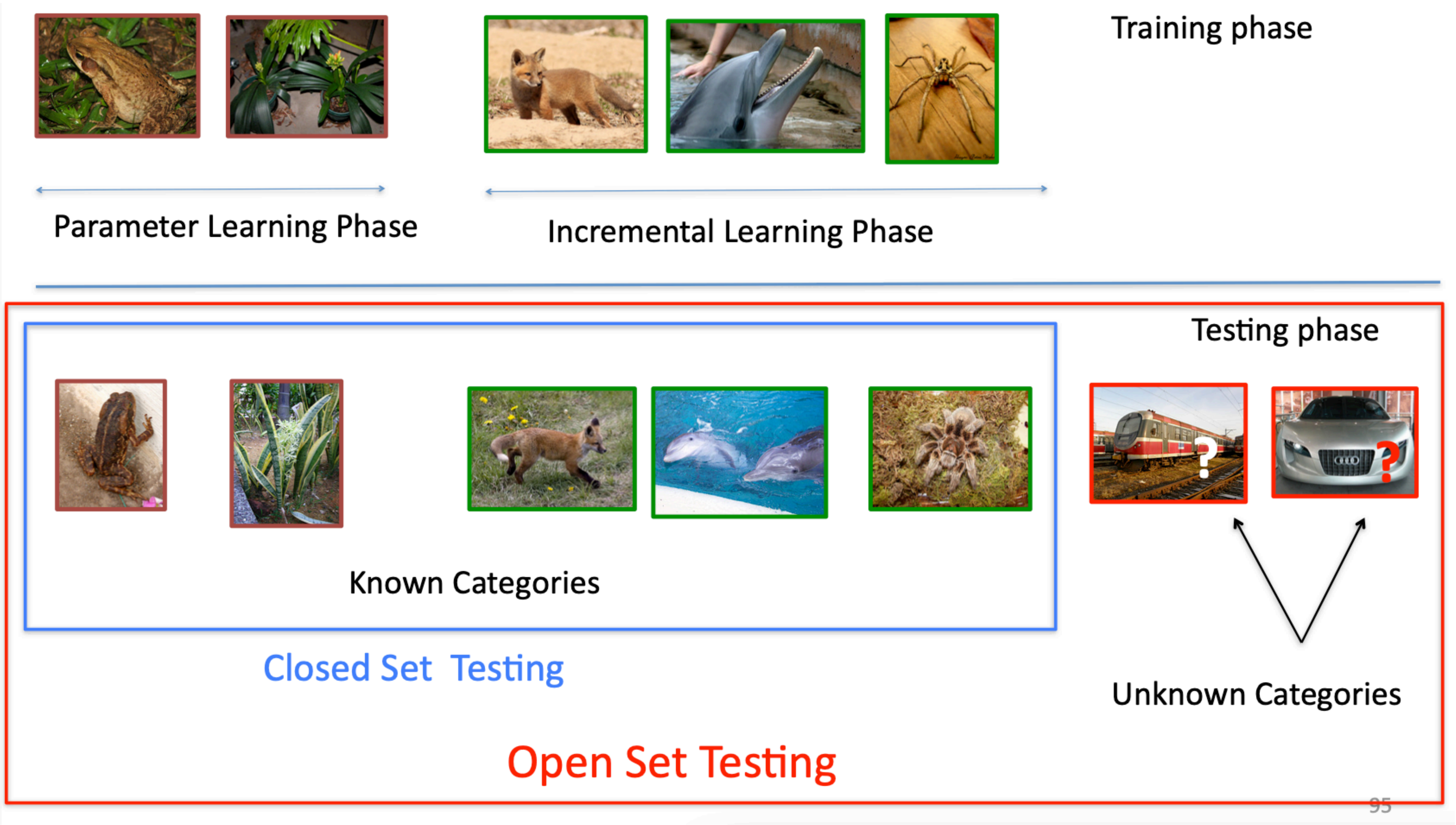
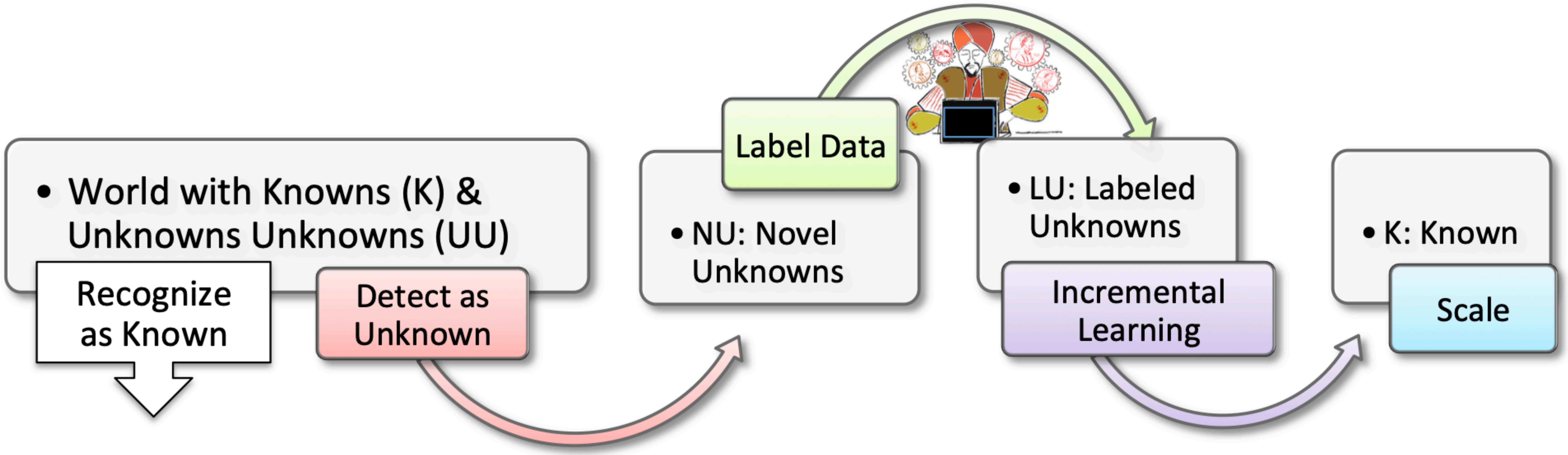


Figure from CVPR16 “Statistical Methods for Open Set Recognition” by Scheirer & Boult, https://www.wjscheirer.com/misc/openset/cvpr2016-open-set-part3.pdf

Open world learning: set-up & evaluation

Open world learning tries to "puzzle together" some (not all) of our seen pieces

"An effective open world recognition system must efficiently perform four tasks: detect unknown, choose which points to label for addition to the model, label the points, and update the model" (Boult et al, "Learning and the Unknown", AAAI 2019)



Bendale & Boult ,"Towards Open World Recognition", CVPR 2015

Finally: all together? An invitation to read two surveys!
1. A wholistic view of CL, Mundt et al, Neural Networks 2023

M. Mundt, Y. Hong, I. Pliushch et al. Neural Networks 160 (2023) 306–336

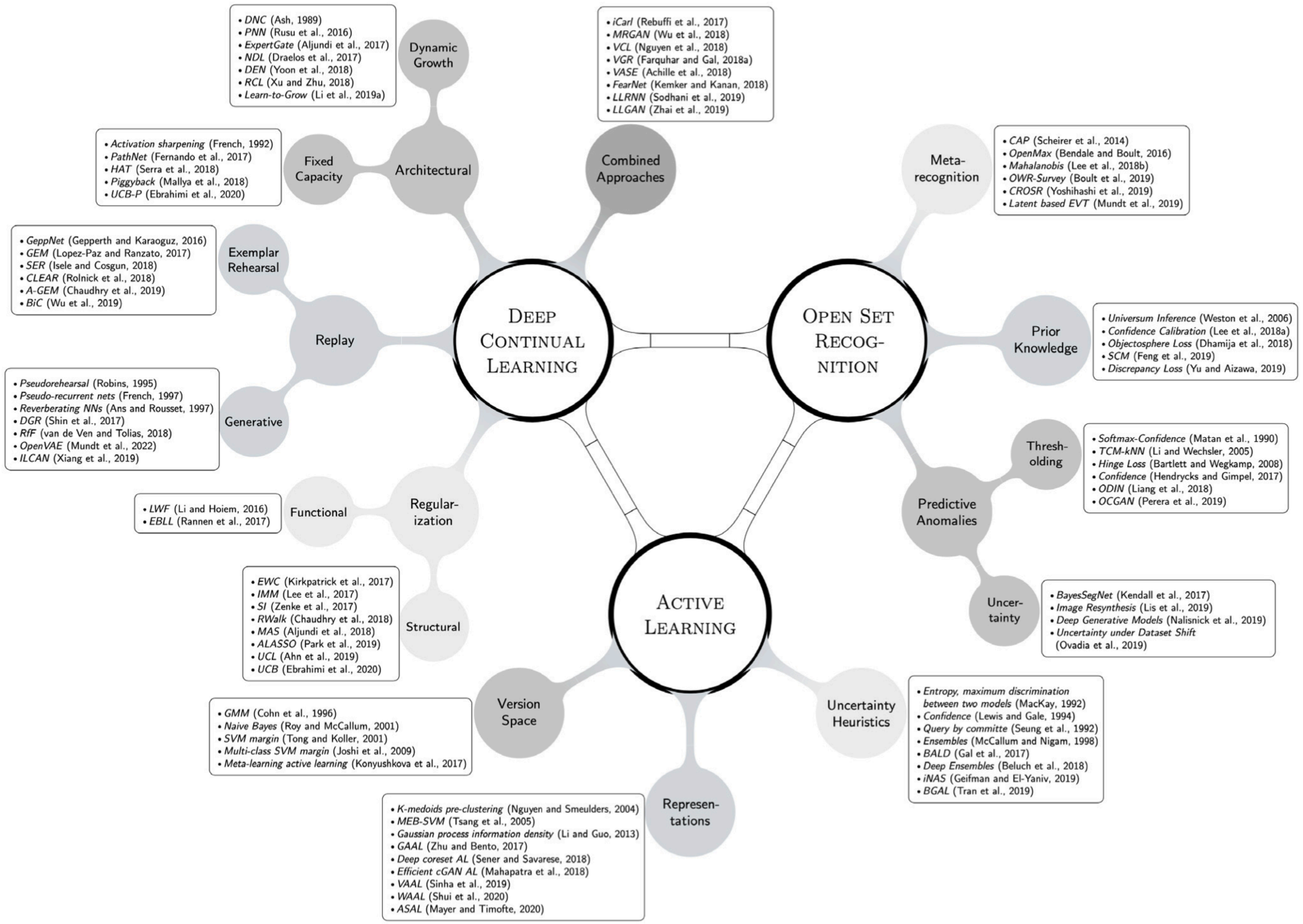
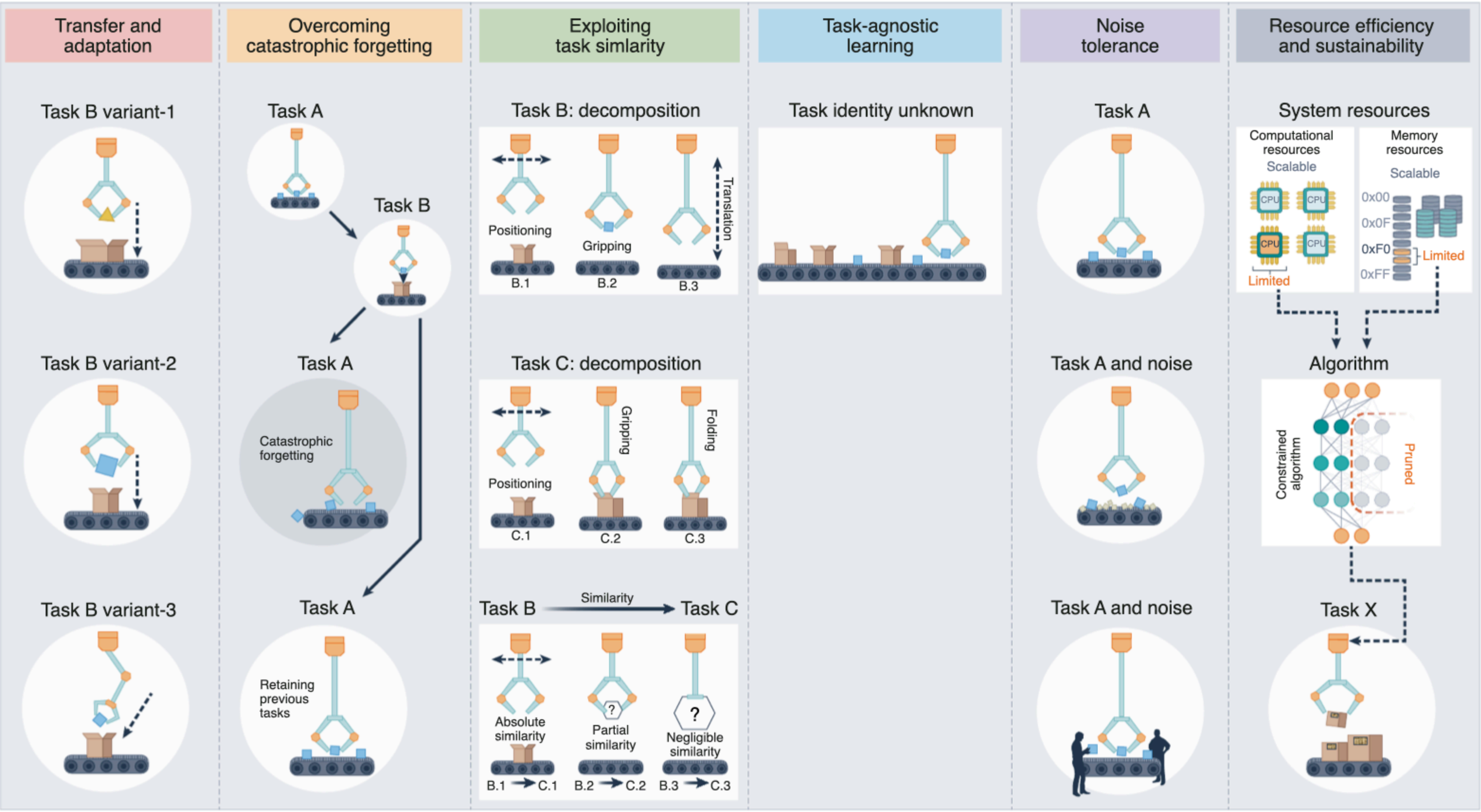


Fig. 4. Visual taxonomy of neural network based methods for continual learning, active learning and open set recognition.

How forgetting, active data queries & order are connected to open set recognition & generative models

“A Wholistic View of Deep Neural Networks: Forgotten Lessons and the Bridge to Active and Open World Learning”, Mundt et al, Neural Networks, 2023

Finally: all together? An invitation to read two surveys!
2. Biological underpinnings of LML, Kudithipudi et al, Nature MI 2022



Ideally, we may also want all together, as hypothesized or even known for biological systems!

Kudithipudi et al, “Biological underpinnings for lifelong learning machines”, Nature Machine Intelligence (4), 2022

So what are the implications for evaluation measures?

It depends on the choices for our mechanisms.
Example: (catastrophic) forgetting

Generally: Average loss, final loss, learning speed, data dependency, transferability, forgetting (backward transfer), “openness”, robustness?

Rehearsal methods: (constant?) memory size, generated data amount, extra computational expense…?

Regularization methods: Regularization strength (hyper-parameters), memory expense, computational expense…?

Architecture/parameter methods: Number of parameters, number of models, expert heads, memory expense, computational expense…?

First good idea: per “task” measures

- “**Base**” loss: the initial (an old) task after i new experiences

$$\Omega_{base} = \frac{1}{T-1}\sum_{i=2}^{T}\frac{\alpha_{base,i}}{\alpha_{ideal}}$$

- “**New**” loss: the newest task only

$$\Omega_{new} = \frac{1}{T-1}\sum_{i=2}^{T}\alpha_{new,i}$$

- “**All**” loss: average up to the present point in time

$$\Omega_{all} = \frac{1}{T-1}\sum_{i=2}^{T}\frac{\alpha_{all,i}}{\alpha_{ideal}}$$

- “**Ideal**” loss: offline value trained at once

Kemker et al, “Measuring Catastrophic Forgetting in Neural Networks”, AAAI 2018

First good idea: per “task” measures

- “**Base**” loss: the initial (an old) task after i new experiences

  -> Measure retention

- “**New**” loss: the newest task only

  -> Measure ability to encode new tasks

- “**All**” loss: average up to the present point in time

  -> Measure present overall performance

- “**Ideal**” loss: offline value trained at once

  -> Measure achievable “baseline”

$$\Omega_{base} = \frac{1}{T-1}\sum_{i=2}^{T}\frac{\alpha_{base,i}}{\alpha_{ideal}}$$

$$\Omega_{new} = \frac{1}{T-1}\sum_{i=2}^{T}\alpha_{new,i}$$

$$\Omega_{all} = \frac{1}{T-1}\sum_{i=2}^{T}\frac{\alpha_{all,i}}{\alpha_{ideal}}$$

Kemker et al, “Measuring Catastrophic Forgetting in Neural Networks”, AAAI 2018

Second good idea: learning speed & data dependency

(Avg.) **b-shot performance** (b = mini-batch number) after the model has been trained on all tasks T

Chaudhry et al, “Efficient Lifelong Learning with A-GEM”, ICLR 2019

Second good idea: learning speed & data dependency

(Avg.) **b-shot performance** (b = mini-batch number) after the model has been trained on all tasks T

Learning Curve Area (LCA) at beta is the area of the convergence curve Z as a function of b in [0, beta]:

$$\mathrm{LCA}_\beta = \frac{1}{\beta+1}\int_0^\beta Z_b db = \frac{1}{\beta+1}\sum_{b=0}^{\beta} Z_b$$

Beta = 0 is zero-shot performance == Forward transfer

Chaudhry et al, “Efficient Lifelong Learning with A-GEM”, ICLR 2019

Third good idea: memory, size & compute

Similar measures for memory, size & compute (here tasks=N) (Díaz-Rodríguez & Lomonaco et al, "Don't forget, there is more than forgetting: new metrics for Continual Learning", 2018)

$$CE = min(1, \frac{\sum_{i=1}^{N} \frac{Ops\uparrow\downarrow(Tr_i)\cdot\varepsilon}{Ops(Tr_i)}}{N})$$

Computational Efficiency

Quantifies add/multiply ops (inference & updates)

$$MS = min(1, \frac{\sum_{i=1}^{N} \frac{Mem(\theta_1)}{Mem(\theta_i)}}{N})$$

Model Size Efficiency

Quantifies parameter growth

$$SSS = 1 - min(1, \frac{\sum_{i=1}^{N} \frac{Mem(M_i)}{Mem(D)}}{N})$$

Sample Storage Size Efficiency

Quantifies stored amount of data (for rehearsal)

We don’t yet have consensus, but we at least agree it’s more than “best in bold” of some average value

The challenge of definitions & formulating desiderata: consensus

Some suggestions (Farquhar & Gal, "Towards Robust Evaluations in Continual Learning"):

A. Cross-task resemblance
B. Shared output head
C. No test time task labels
D. No unconstrained re-training on old tasks
E. More than two tasks

And also questions: unclear task boundaries, continuous tasks, overlapping vs. disjoint tasks, long task sequences, time/compute/memory constraints, privacy guarantees…

The challenge of definitions & formulating desiderata: consensus

Is it at all possible to postulate general desiderata?

Property	*Definition*
Knowledge retention	The model is not prone to catastrophic forgetting.
Forward transfer	The model learns a new task while reusing knowledge acquired from previous tasks.
Backward transfer	The model achieves improved performance on previous tasks after learning a new task.
On-line learning	The model learns from a continuous data stream.
No task boundaries	The model learns without requiring neither clear task nor data boundaries.
Fixed model capacity	Memory size is constant regardless of the number of tasks and the length of a data stream.

Table 1: Desiderata of continual learning.

Biesialska et al, “Continual Learning in Natural Language Processing: A Survey”, COLING 2020

We seem to lack benchmarks that allow us to do principled investigation + non-static datasets at large-scale

Importantly: a lot of existing work (if not the most) “emulates" by re-purposing existing datasets

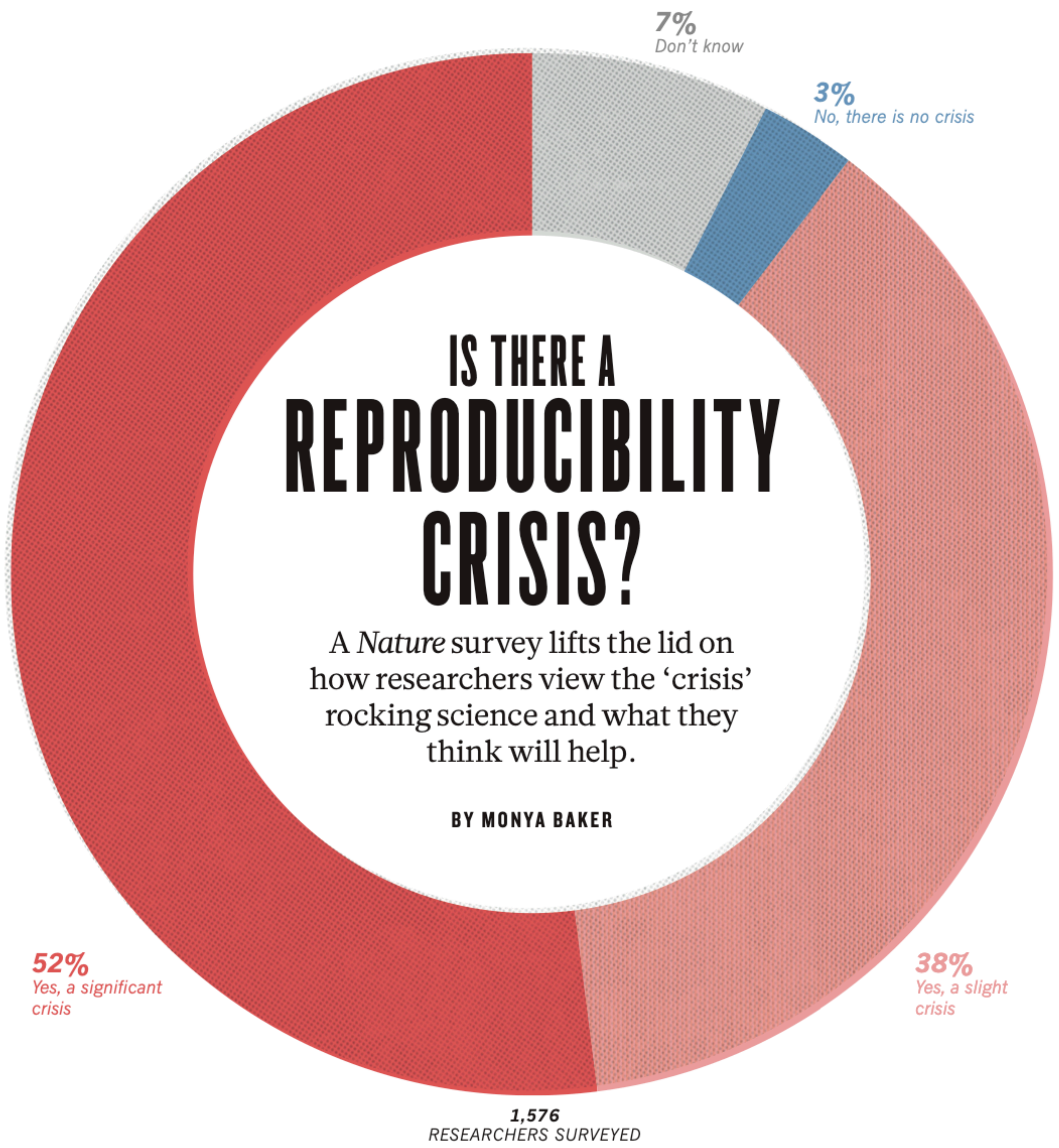
- A sequence of datasets
- Sequences of classes (from known datasets)
- Sequentially querying the instances of datasets
- Sequences of games (in RL), or languages etc.
- Sequences of the same task with shifting distribution

So what are good benchmarks & how do we evaluate?

So what are good benchmarks & how do we evaluate?
I don’t have full answers, but it is extremely important!

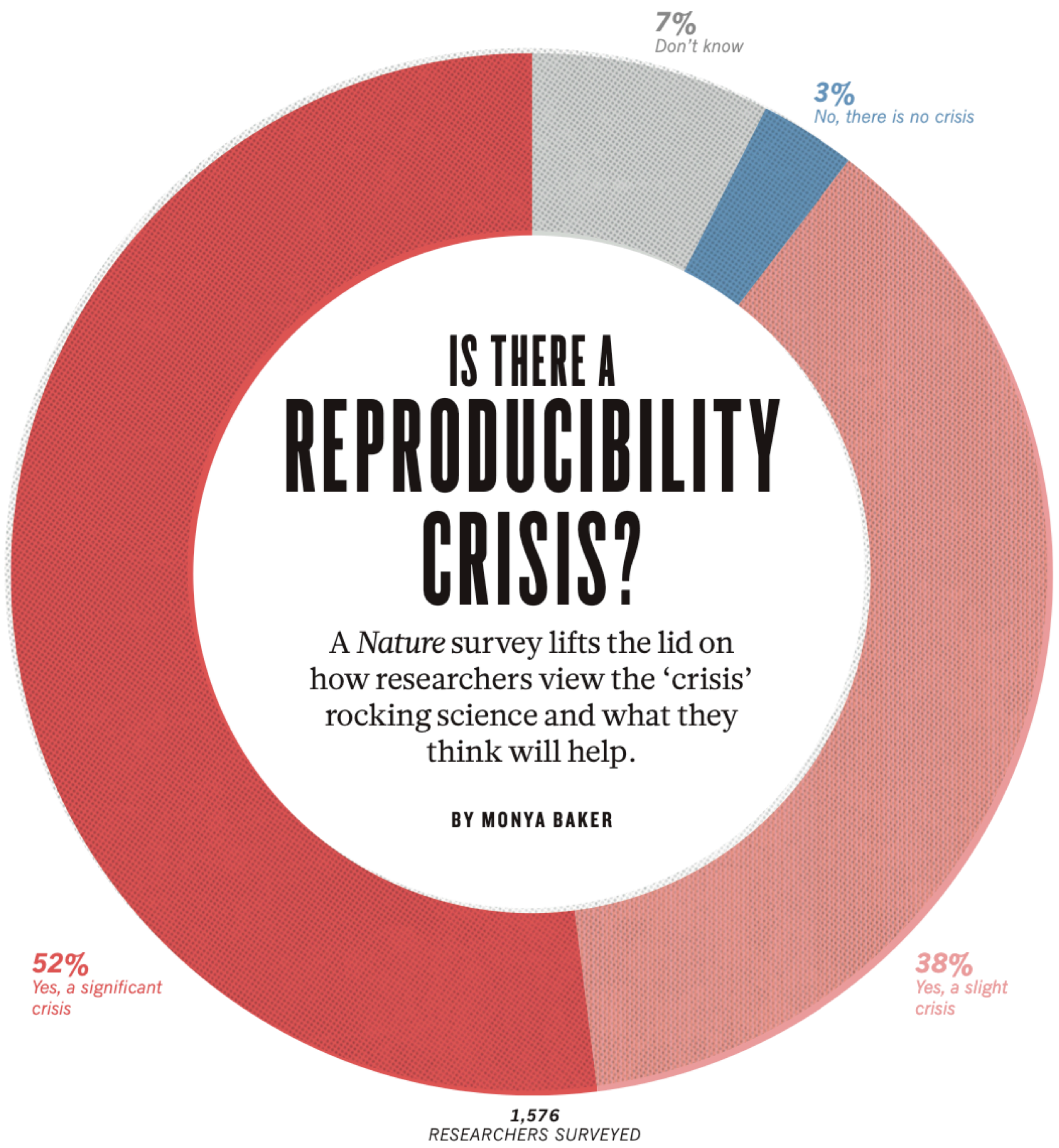
Why? Answer A:
Reproducibility Crisis

Why? Answer A) is reproducibility in a crisis?



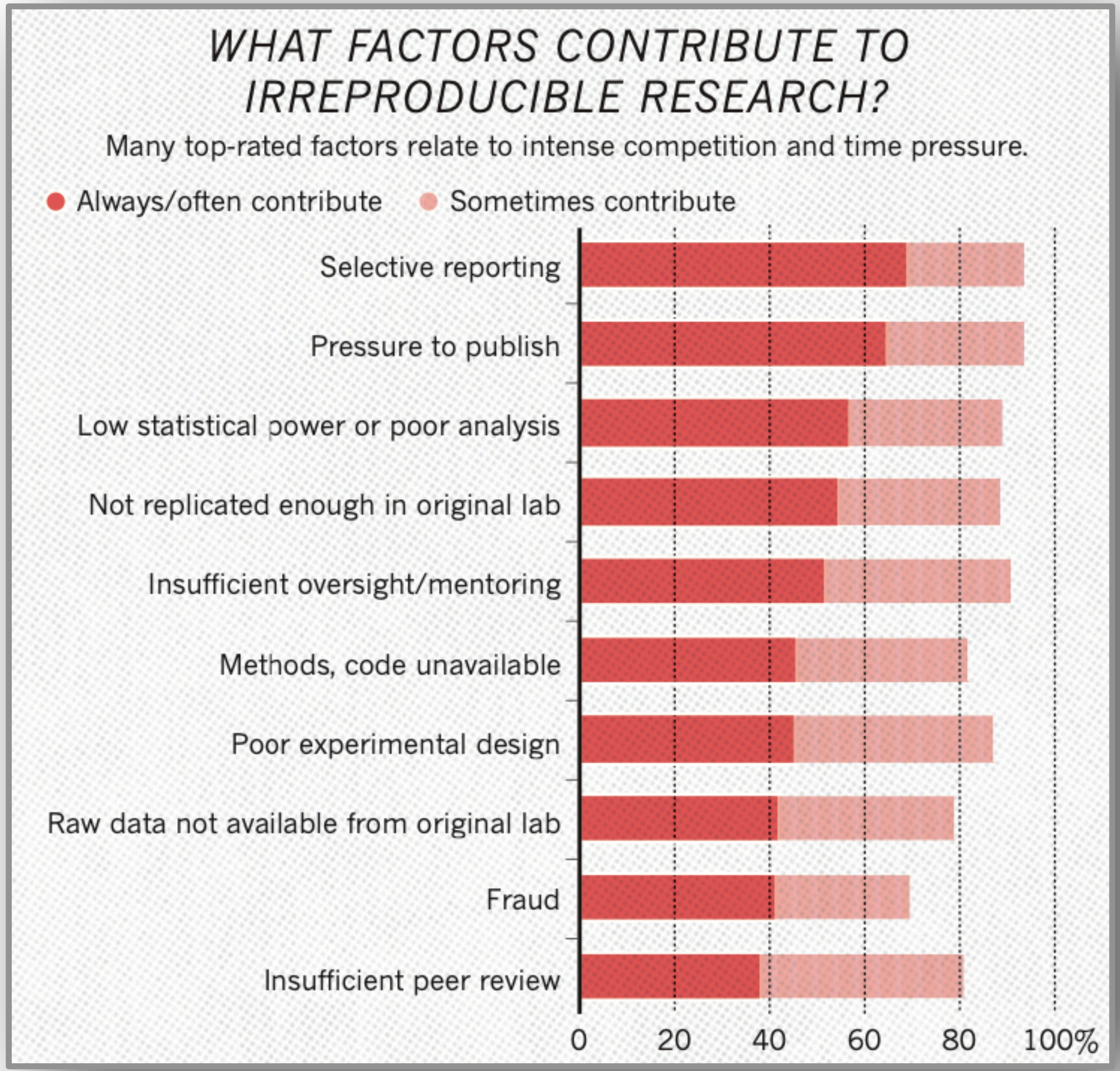
"1500 scientists lift the lid on reproducibility", Baker, Nature, issue 533, 2016

Why? Answer A) is reproducibility in a crisis?



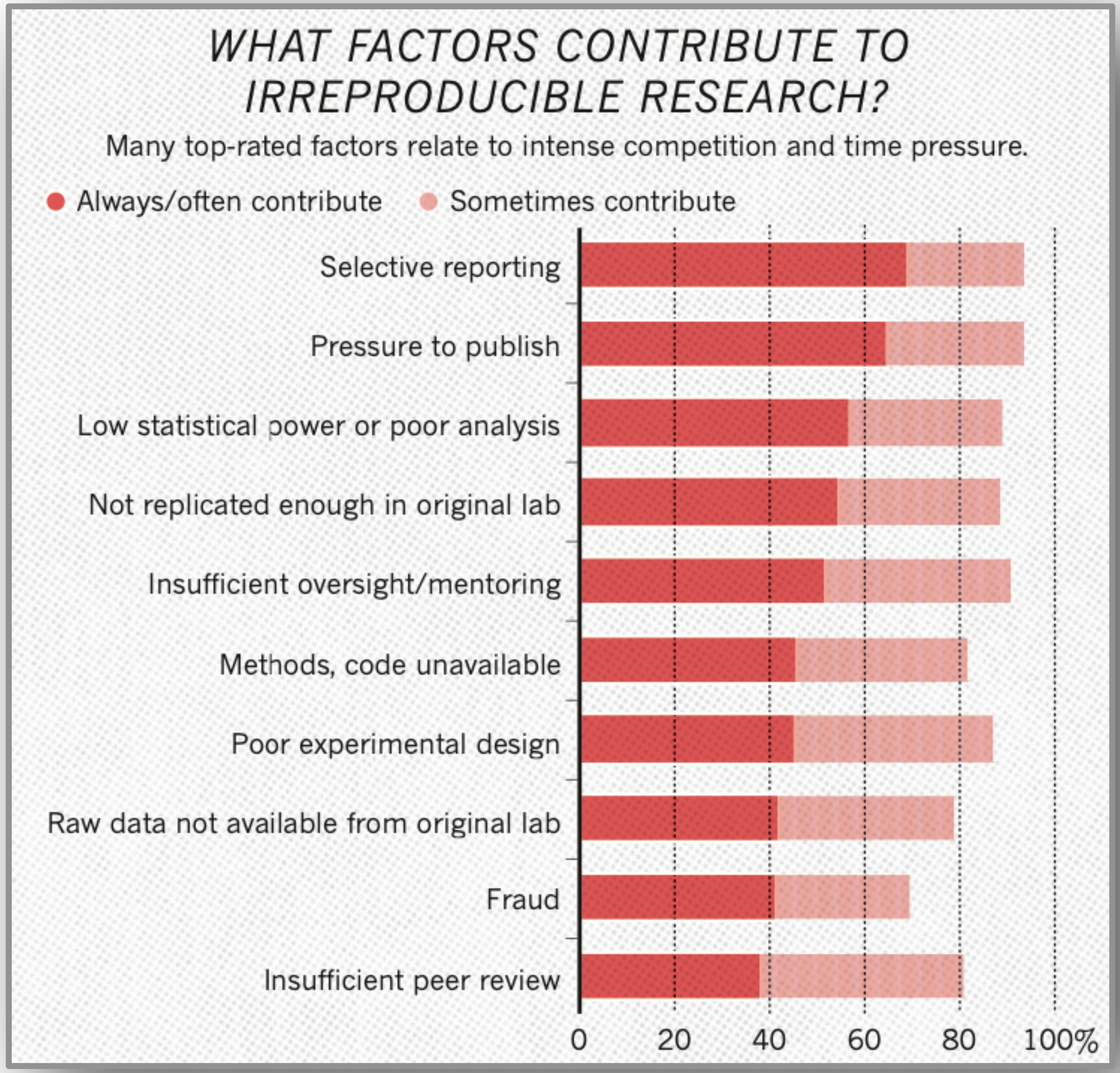
"1500 scientists lift the lid on reproducibility", Baker, Nature, issue 533, 2016

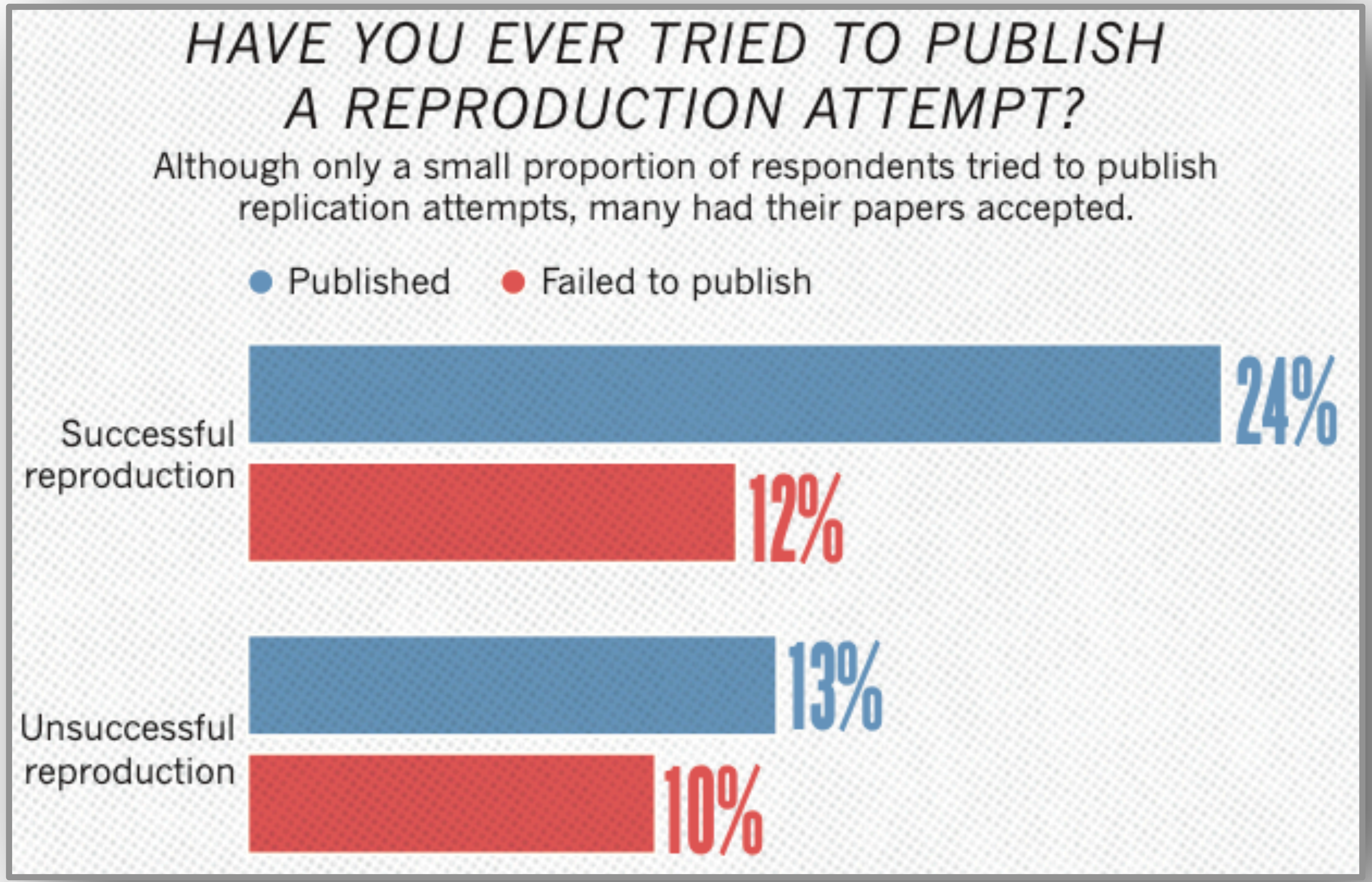
Why? Answer A) is reproducibility in a crisis?



"1500 scientists lift the lid on reproducibility", Baker, Nature, issue 533, 2016

Why? Answer A) is reproducibility in a crisis?





"1500 scientists lift the lid on reproducibility", Baker, Nature, issue 533, 2016

Why? Answer A) is ML reproducibility in a crisis?

Through experimental methods focusing on PG methods for continuous control, we investigate problems with reproducibility in deep RL. We find that both intrinsic (e.g. random seeds, environment properties) and extrinsic sources (e.g. hyperparameters, codebases) of non-determinism can contribute to difficulties in reproducing baseline algorithms.

"Deep Reinforcement Learning that Matters", Henderson et al, AAAI 2018

Why? Answer A) is LML reproducibility in a crisis?

The lack of consensus in evaluating continual learning algorithms and the almost exclusive focus on forgetting motivate us to propose a more comprehensive set of implementation independent metrics accounting for several factors we believe have practical implications worth considering in the deployment of real AI systems that learn continually: accuracy or performance over time, backward and forward knowledge transfer, memory overhead as well as computational efficiency.

"Don't forget, there is more than forgetting: new metrics for Continual Learning",
Díaz-Rodríguez et al, Continual Learning Workshop at NeurIPS 2018

Why? Answer A) is LML reproducibility in a crisis?

The lack of consensus in evaluating continual learning algorithms and the almost exclusive focus on forgetting motivate us to propose a more comprehensive set of implementation independent metrics accounting for several factors we believe have practical implications worth considering in the deployment of real AI systems that learn continually: accuracy or performance over time, backward and forward knowledge transfer, memory overhead as well as computational efficiency.

"Don't forget, there is more than forgetting: new metrics for Continual Learning",
Díaz-Rodríguez et al, Continual Learning Workshop at NeurIPS 2018

we evaluate CF behavior on the hitherto largest number of visual classification datasets, from each of which we construct a representative number of Sequential Learning Tasks (SLTs) in close alignment to previous works on CF. Our results clearly indicate that there is no model that avoids CF for all investigated datasets and SLTs under application conditions.

"A comprehensive, application-oriented study of catastrophic forgetting in DNNs",
Pfuelb & Gepperth, ICLR 2019

Why? Answer B:
Awareness of application relevant trade-offs

Why? Answer B) every application has different requirements, but we need to be aware of trade-offs

Category	Method	Memory		Compute		Task-agnostic possible	Privacy issues	Additional required storage
		train	*test*	*train*	*test*			
Replay-based	iCARL	1.24	1.00	5.63	45.61	✓	✓	$M + R$
	GEM	1.07	1.29	10.66	3.64	✓	✓	$\mathcal{T} \cdot M + R$
Reg.-based	LwF	1.07	1.10	1.29	1.86	✓	✗	M
	EBLL	1.53	1.08	2.24	1.34	✓	✗	$M + \mathcal{T} \cdot A$
	SI	1.09	1.05	1.13	1.61	✓	✗	$3 \cdot M$
	EWC	1.09	1.05	1.11	1.88	✓	✗	$2 \cdot M$
	MAS	1.09	1.05	1.16	1.88	✓	✗	$2 \cdot M$
	mean-IMM	1.01	1.03	1.09	1.18	✓	✗	$\mathcal{T} \cdot M$
	mode-IMM	1.01	1.03	1.24	1.00	✓	✗	$2 \cdot \mathcal{T} \cdot M$
Param. iso.-based	PackNet	1.00	1.94	2.66	2.40	✗	✗	$\mathcal{T} \cdot M[bit]$
	HAT	1.21	1.17	1.00	2.06	✗	✗	$\mathcal{T} \cdot U$

Low

High

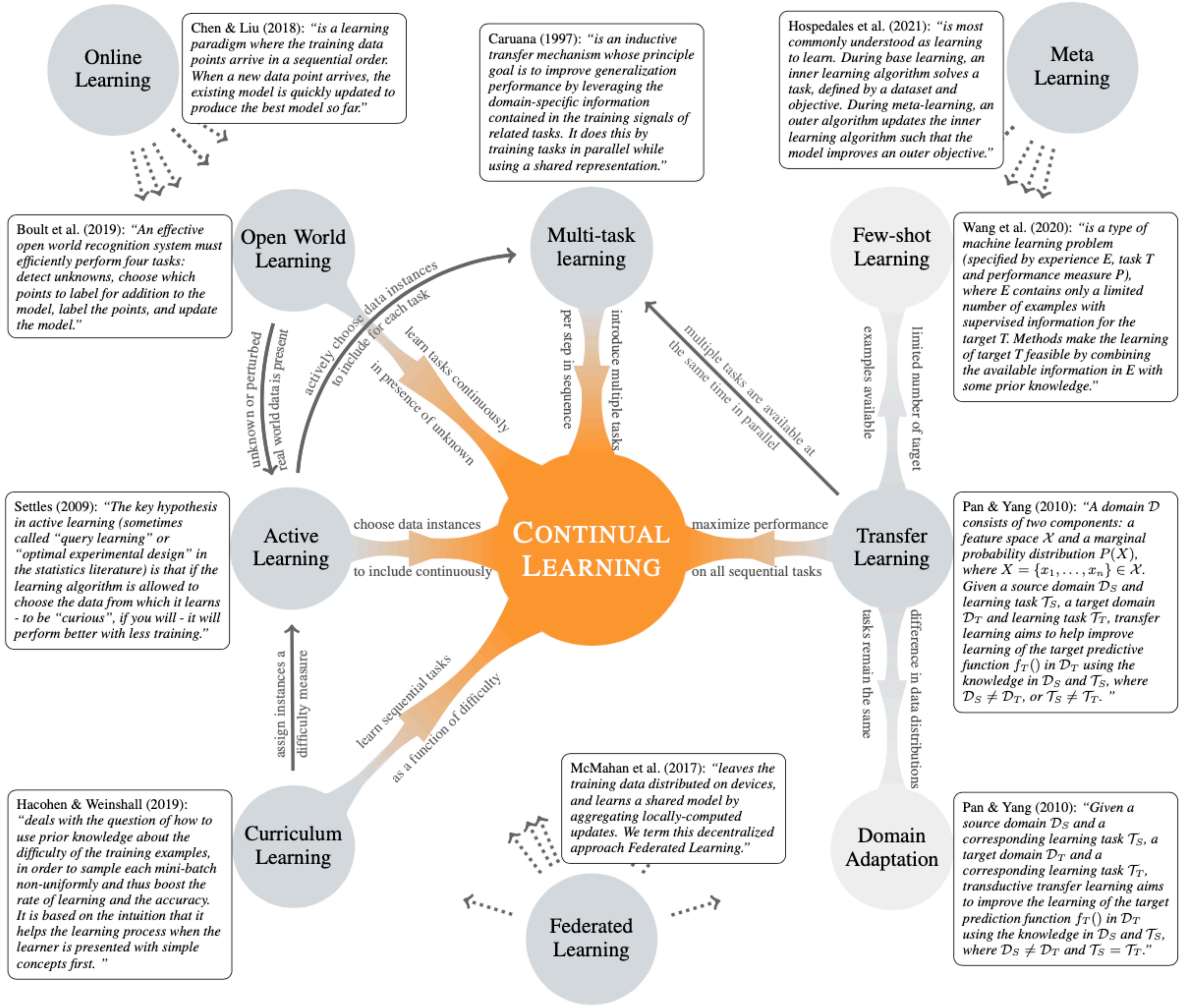
De Lange et al, "A continual learning survey: Defying forgetting in classification tasks", TPAMI 2021

Why? Answer B) every application has different requirements, but we need to be aware of trade-offs

The differences between ML paradigms with continuous components can be nuances

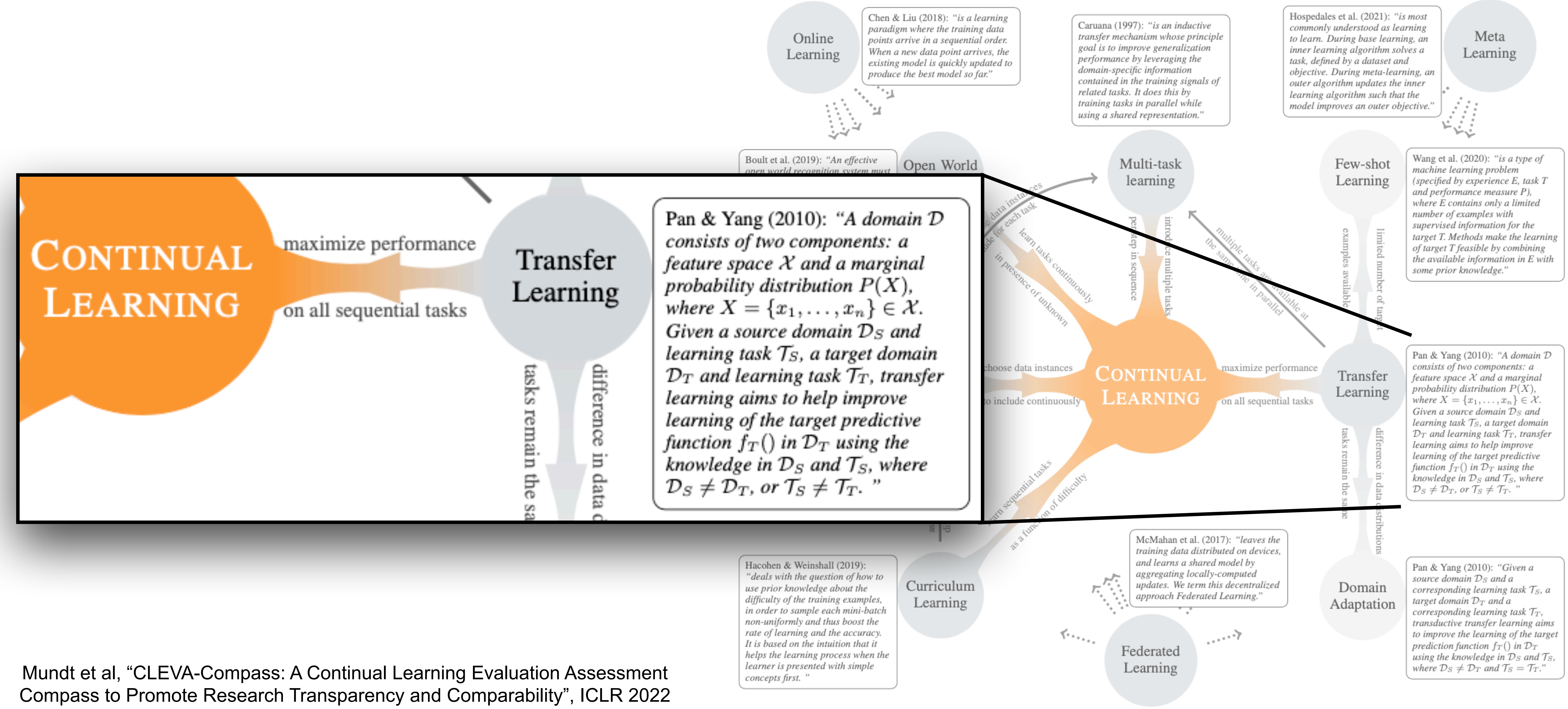
Key aspects often reside in how we evaluate

Each paradigm seems to have a particular preference (potentially neglecting other important factors)



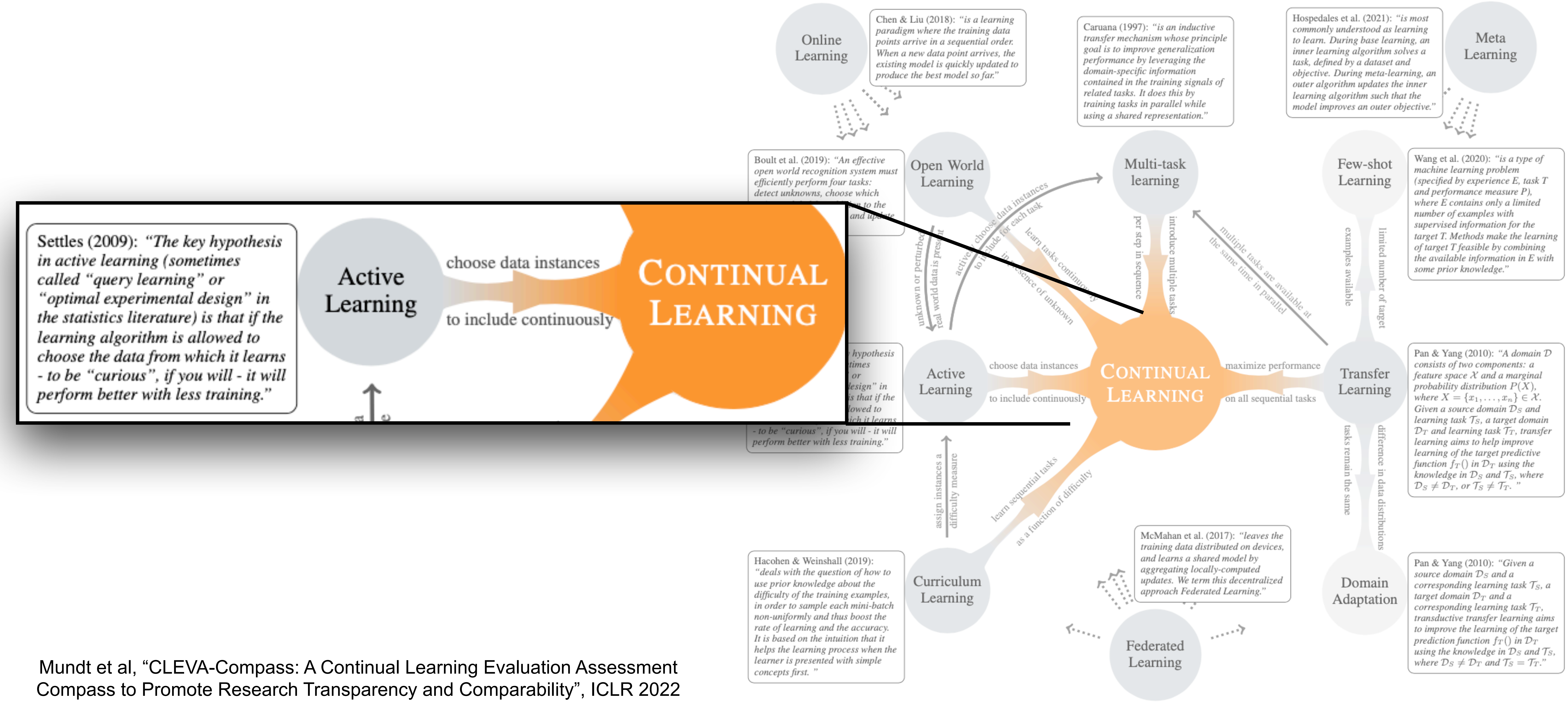
Mundt et al, "CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability", ICLR 2022

Why? Answer B) every application has different requirements, but we need to be aware of trade-offs



Mundt et al, "CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability", ICLR 2022

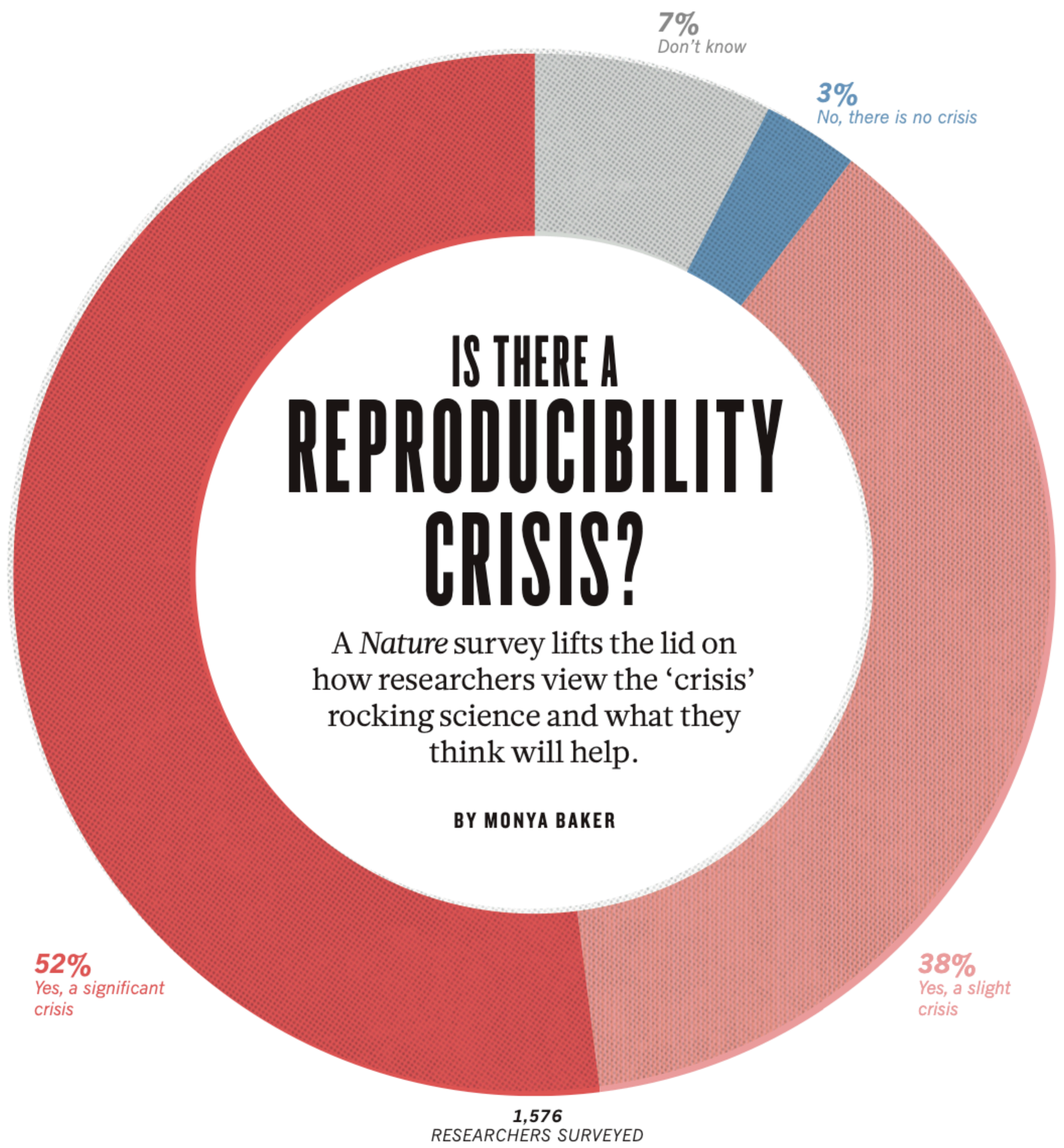
Why? Answer B) every application has different requirements, but we need to be aware of trade-offs

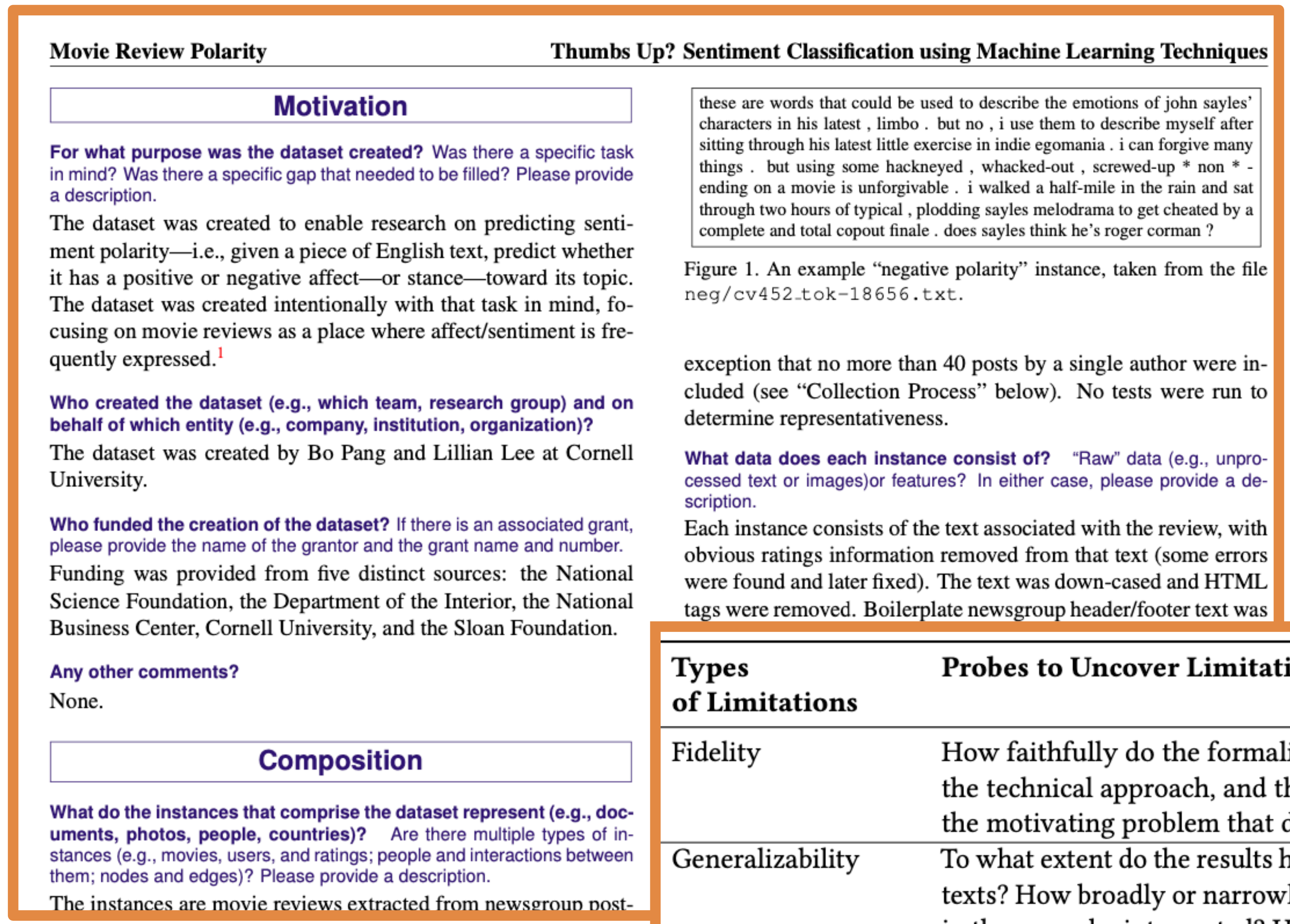


Mundt et al, "CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability", ICLR 2022

Apart from continuing research, what can we do now?

We can develop & use transparent documentation





Movie Review Polarity — Thumbs Up? Sentiment Classification using Machine Learning Techniques

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup post-

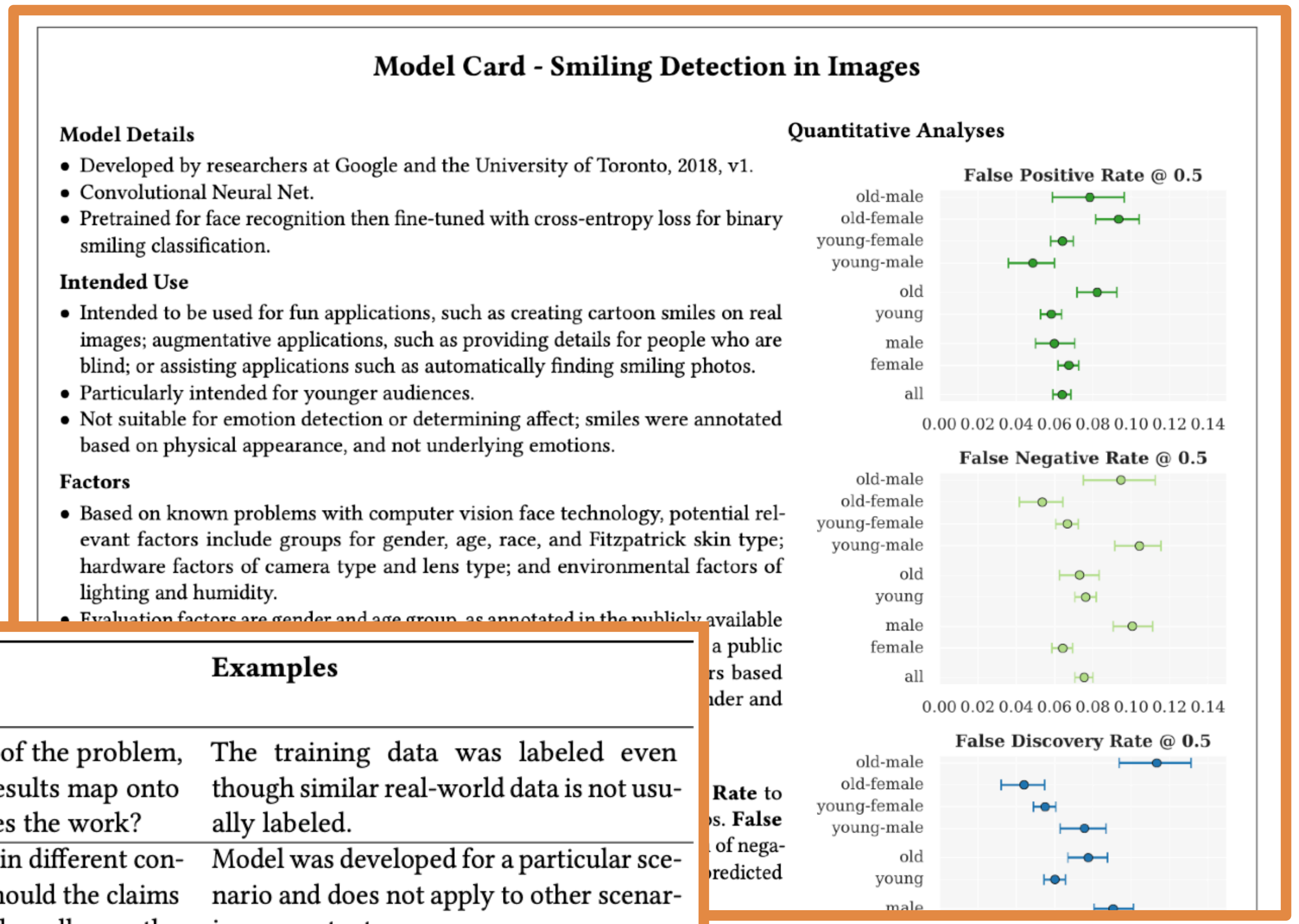
these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file neg/cv452_tok-18656.txt.

exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

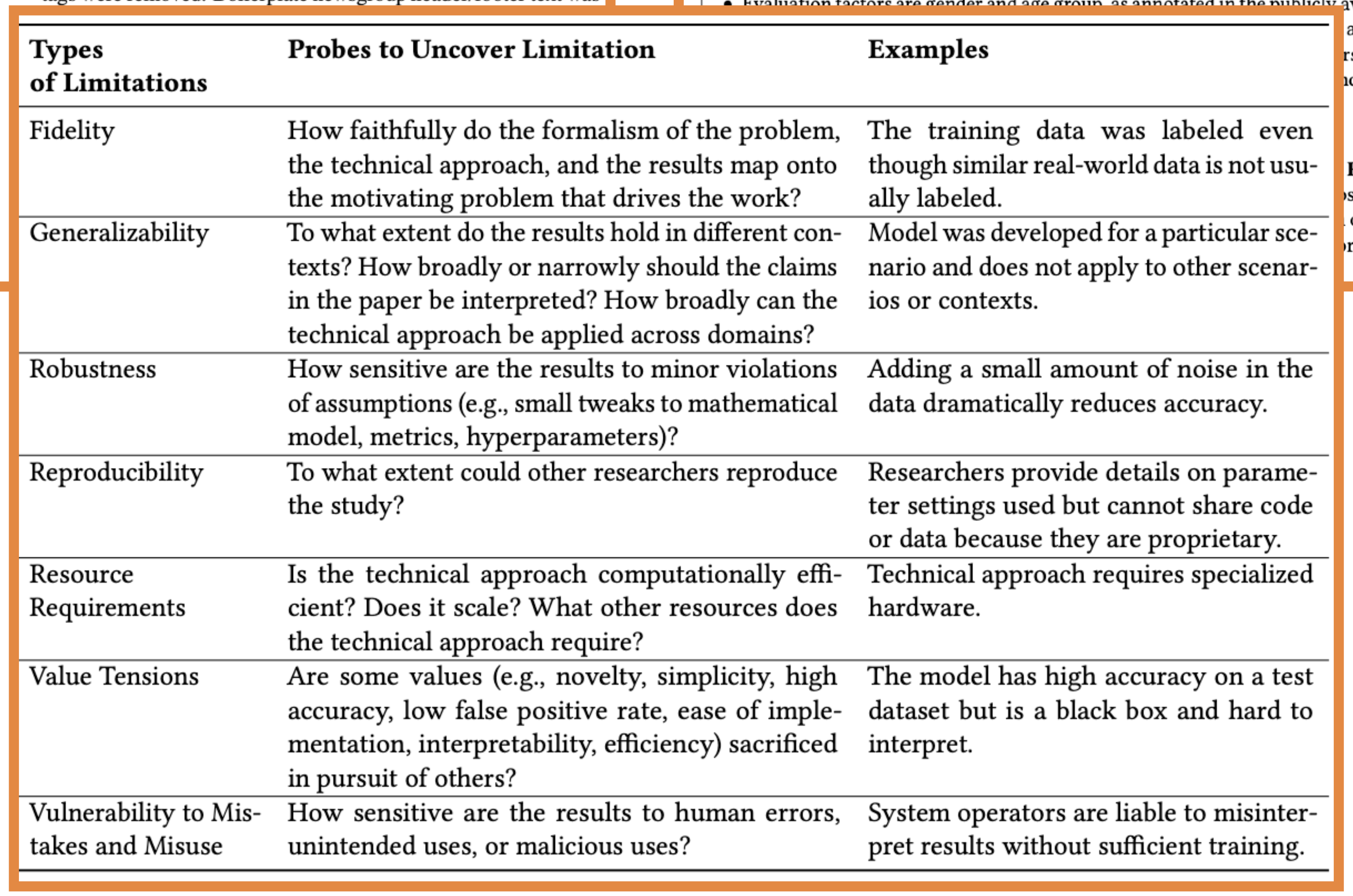
Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.

Quantitative Analyses

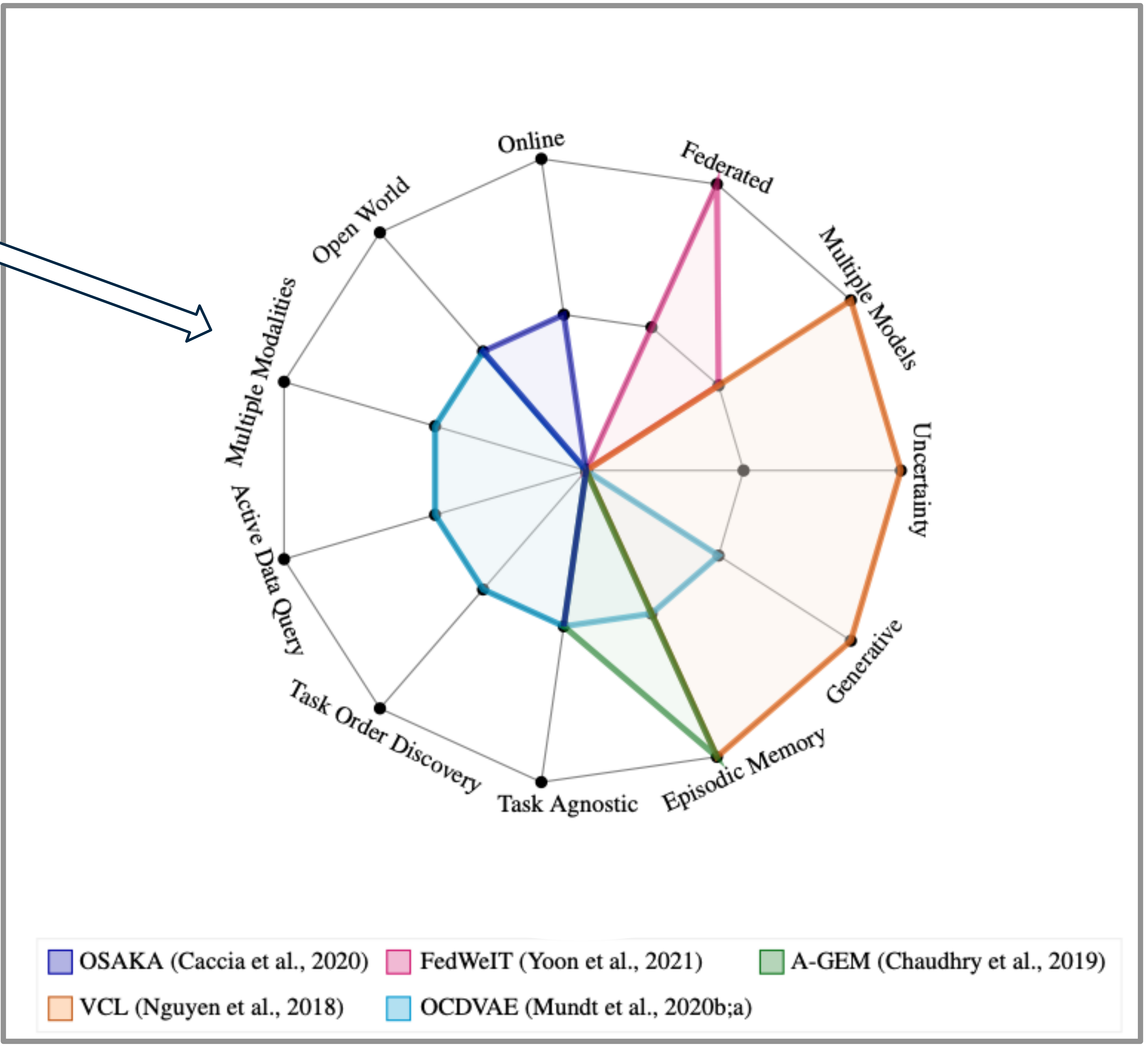


Types of Limitations	Probes to Uncover Limitation	Examples
Fidelity	How faithfully do the formalism of the problem, the technical approach, and the results map onto the motivating problem that drives the work?	The training data was labeled even though similar real-world data is not usually labeled.
Generalizability	To what extent do the results hold in different contexts? How broadly or narrowly should the claims in the paper be interpreted? How broadly can the technical approach be applied across domains?	Model was developed for a particular scenario and does not apply to other scenarios or contexts.
Robustness	How sensitive are the results to minor violations of assumptions (e.g., small tweaks to mathematical model, metrics, hyperparameters)?	Adding a small amount of noise in the data dramatically reduces accuracy.
Reproducibility	To what extent could other researchers reproduce the study?	Researchers provide details on parameter settings used but cannot share code or data because they are proprietary.
Resource Requirements	Is the technical approach computationally efficient? Does it scale? What other resources does the technical approach require?	Technical approach requires specialized hardware.
Value Tensions	Are some values (e.g., novelty, simplicity, high accuracy, low false positive rate, ease of implementation, interpretability, efficiency) sacrificed in pursuit of others?	The model has high accuracy on a test dataset but is a black box and hard to interpret.
Vulnerability to Mistakes and Misuse	How sensitive are the results to human errors, unintended uses, or malicious uses?	System operators are liable to misinterpret results without sufficient training.

- Reproducibility Crisis, Baker, Nature 2016
- Model Cards, Mitchell et al, FAccT 2019
- Data Sheets, Gebru et al, CACM 2021
- REAL ML: Smith et al, FAccT 2022

Continual Learning EValuation Assessment: CLEVA-Compass

Inner compass level (star plot):
indicates related paradigm inspiration & setting configuration (assumptions)

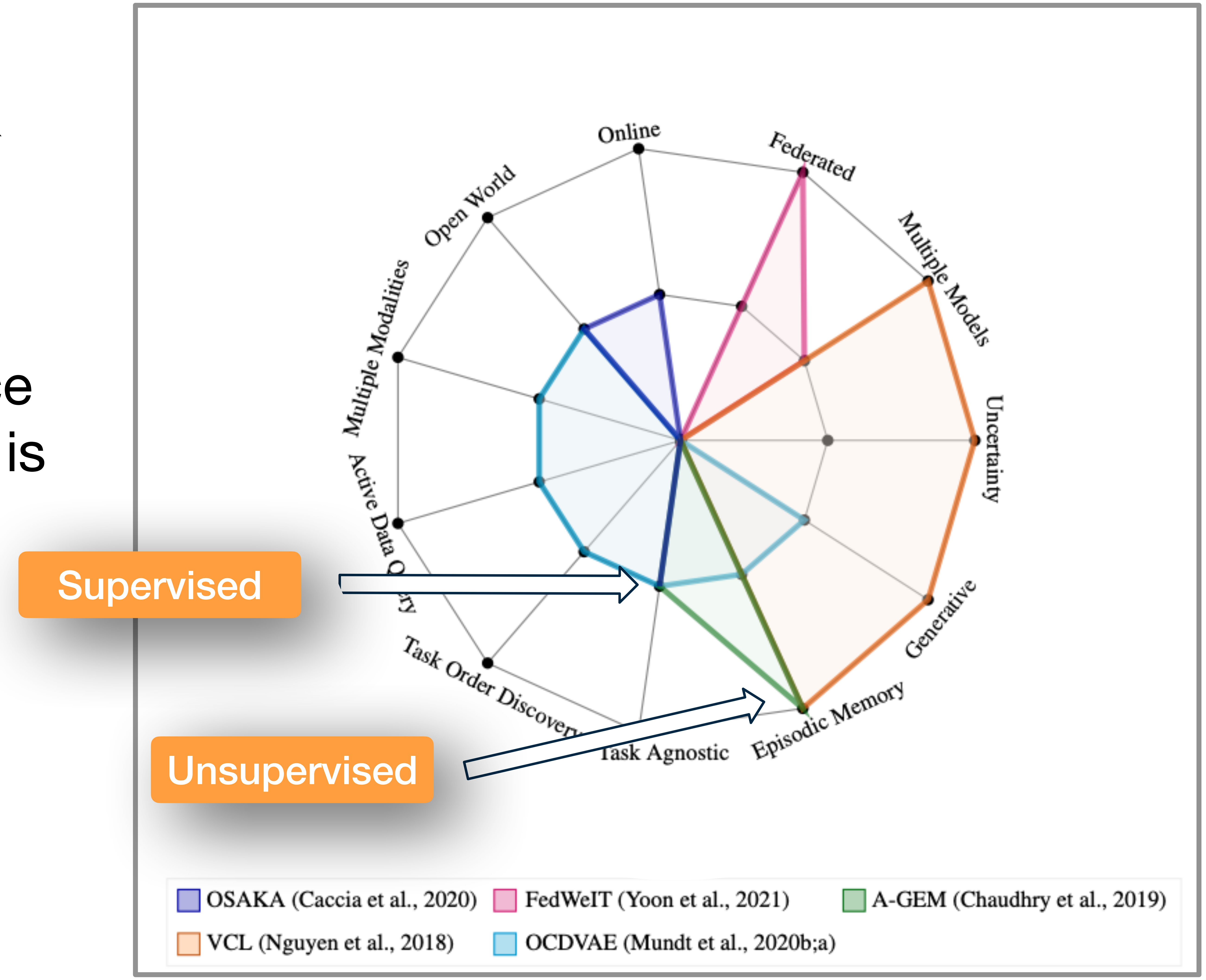


Mundt et al, "CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability", ICLR 2022

Continual Learning EValuation Assessment: CLEVA-Compass

Inner compass level (star plot):
indicates related paradigm inspiration & setting configuration (assumptions)

Inner compass level of supervision:
“rings” on the star plot indicate presence of supervision. Importantly: supervision is individual to each dimension!



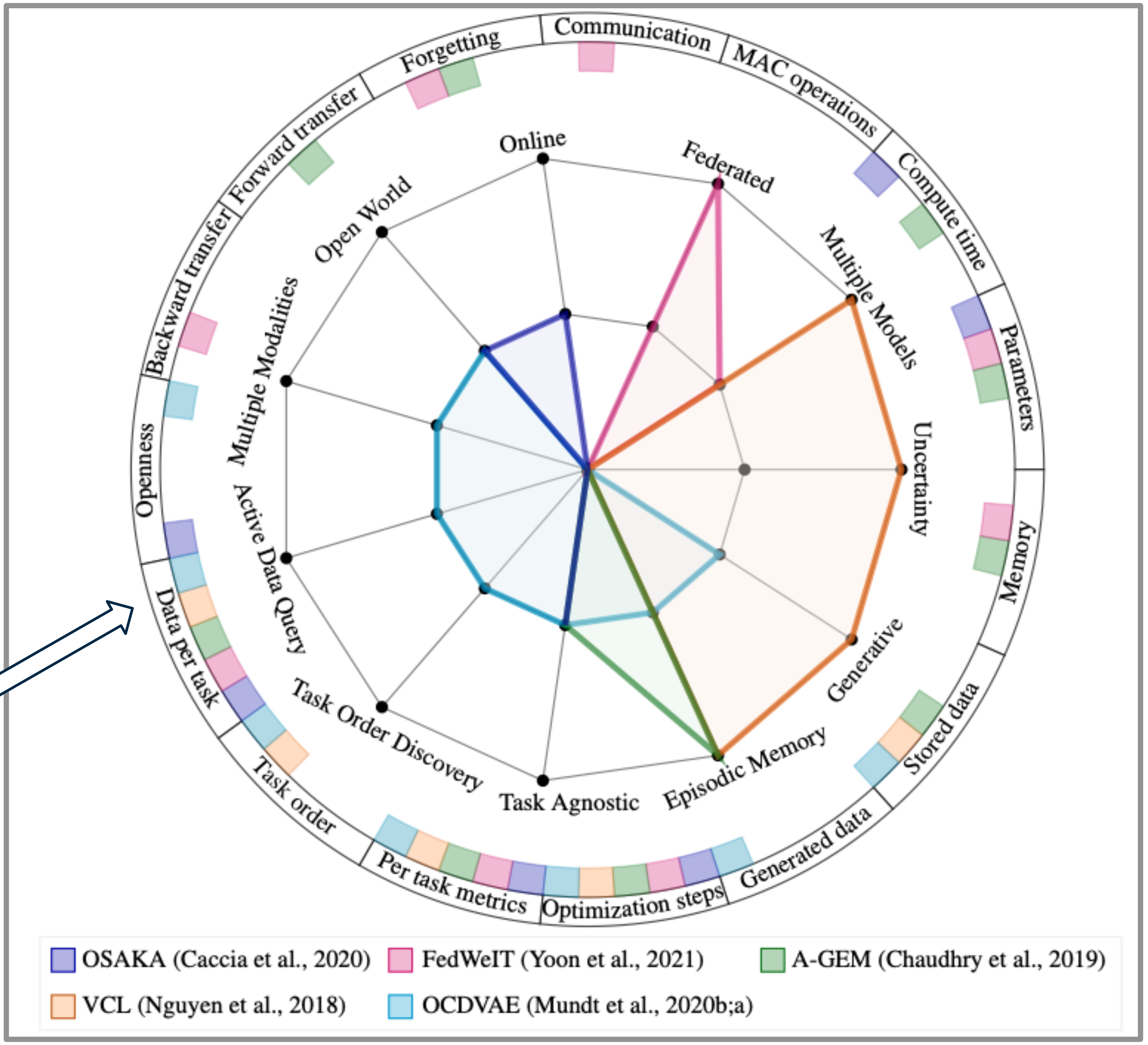
Mundt et al, “CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability”, ICLR 2022

Continual Learning EValuation Assessment: CLEVA-Compass

Inner compass level (star plot):
indicates related paradigm inspiration & setting configuration (assumptions)

Inner compass level of supervision:
"rings" on the star plot indicate presence of supervision. Importantly: supervision is individual to each dimension!

Outer compass level:
Contains a comprehensive set of practically reported measures



Mundt et al, "CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability", ICLR 2022

With gained understanding over the years & hopefully this course, let's acknowledge the opportunity!

With gained understanding over the years & hopefully this course, let's acknowledge the opportunity!

An opportunity to improve understanding, promote transparency & create lifelong learning systems!

With gained understanding over the years & hopefully this course, let's acknowledge the opportunity!

An opportunity to improve understanding, promote transparency & create lifelong learning systems!

Reach out: martin.mundt@tu-darmstadt.de, ContinualAI or QueerInAI Slacks, @mundt_martin on Twitter