Open World Lifelong Learning
A Continual Machine Learning Course

Teacher

Dr. Martin Mundt,

hessian.AI-DEPTH junior research group leader on Open World Lifelong Learning (OWLL)
& researcher in the Artificial Intelligence and Machine Learning (AIML) group at TU Darmstadt

Time

Every Tuesday 17:30 - 19:00 CEST

Course Homepage

http://owll-lab.com/teaching/cl_lecture

https://www.youtube.com/playlist?list=PLm6QXeaB-XkA5-lVBB-h7XeYzFzgSh6sk

TECHNISCHE UNIVERSITÄT DARMSTADT

Week 7: Evaluation

Evaluation

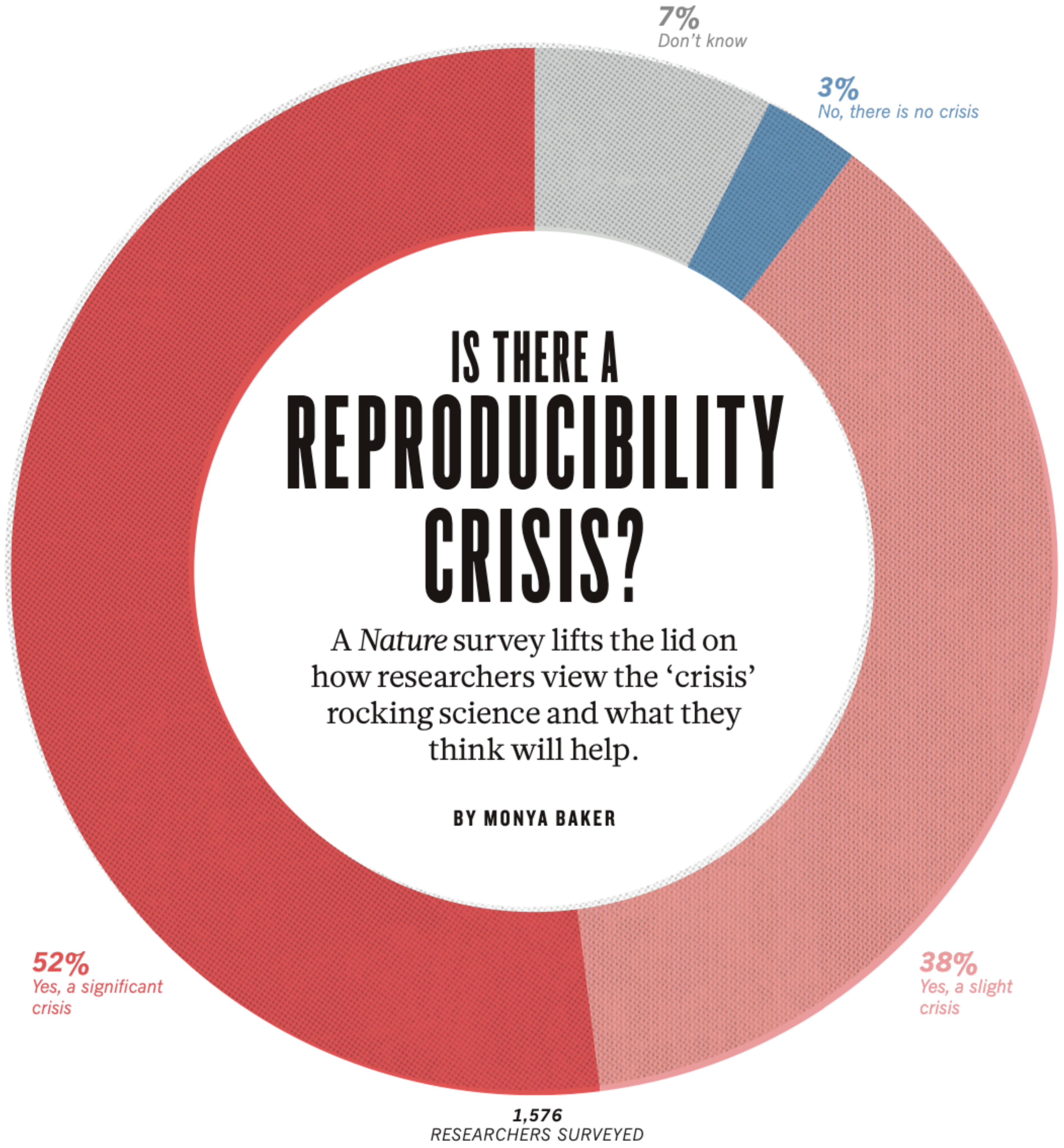
Why is evaluation challenging in machine learning?

Dimensions of evaluation in continual/lifelong learning

Why evaluation is even more challenging in continual/lifelong learning

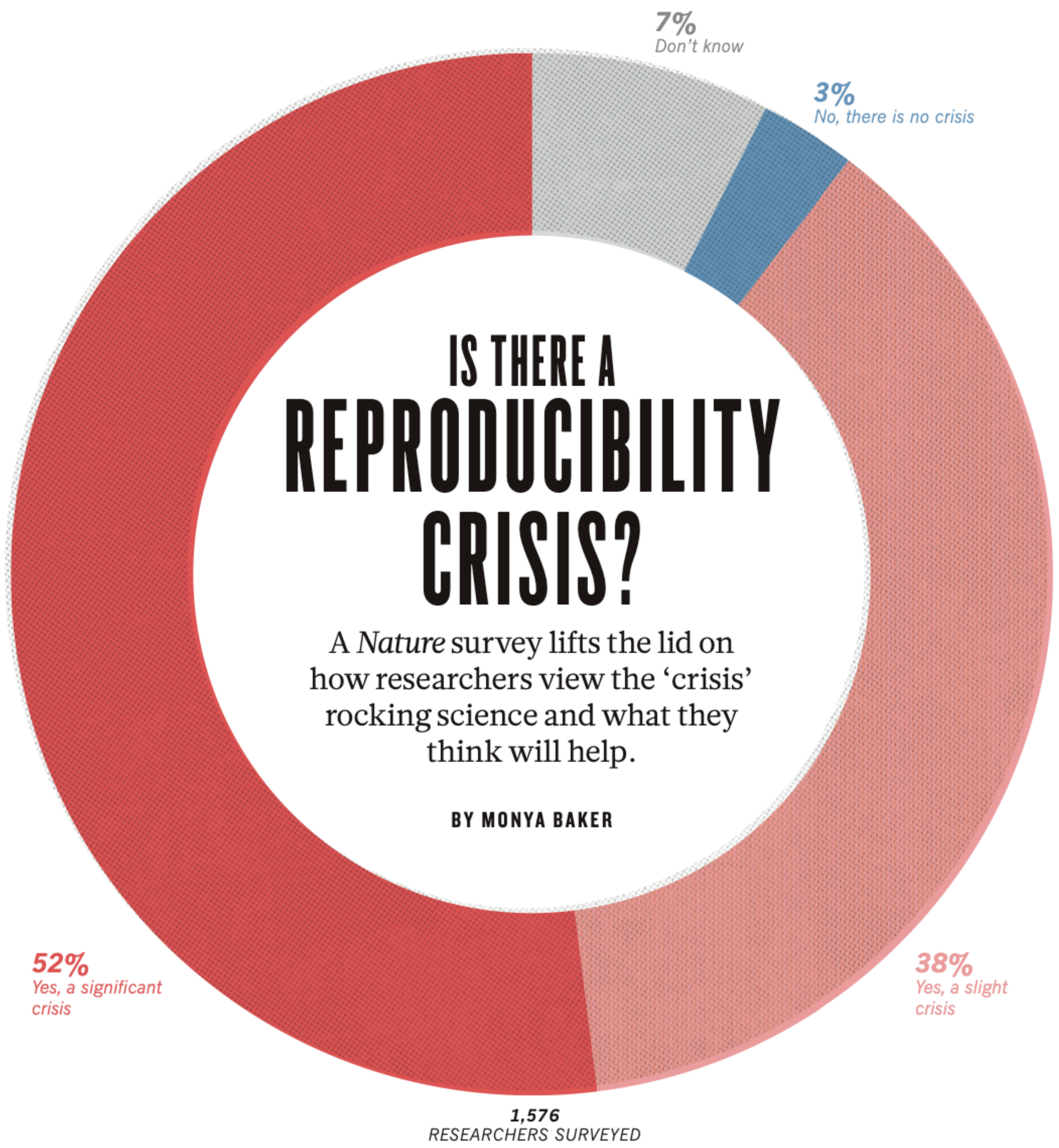
How can we move forward?

Is reproducibility in a crisis?



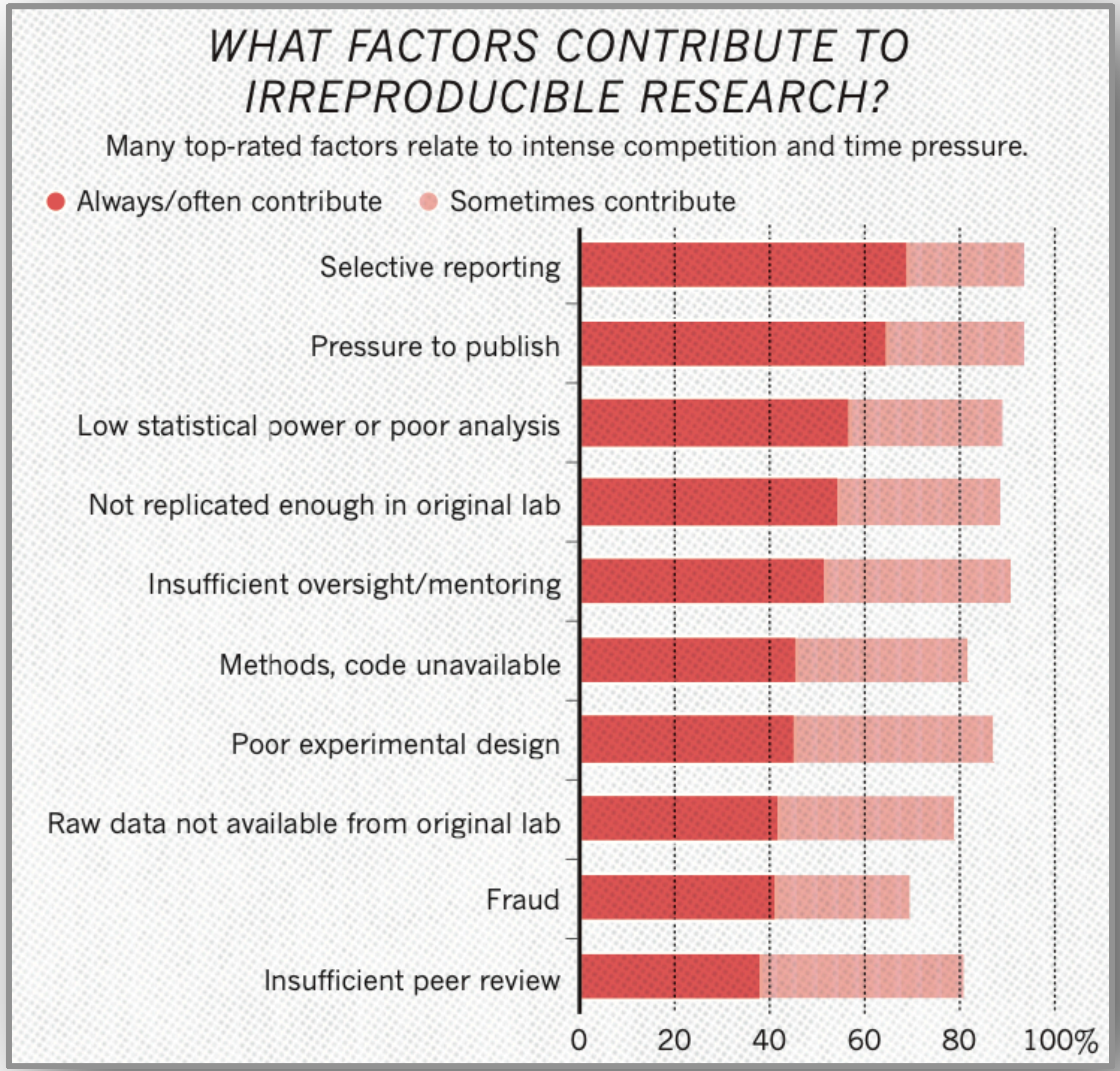
"1500 scientists lift the lid on reproducibility", Baker, Nature, issue 533, 2016

Is reproducibility in a crisis?



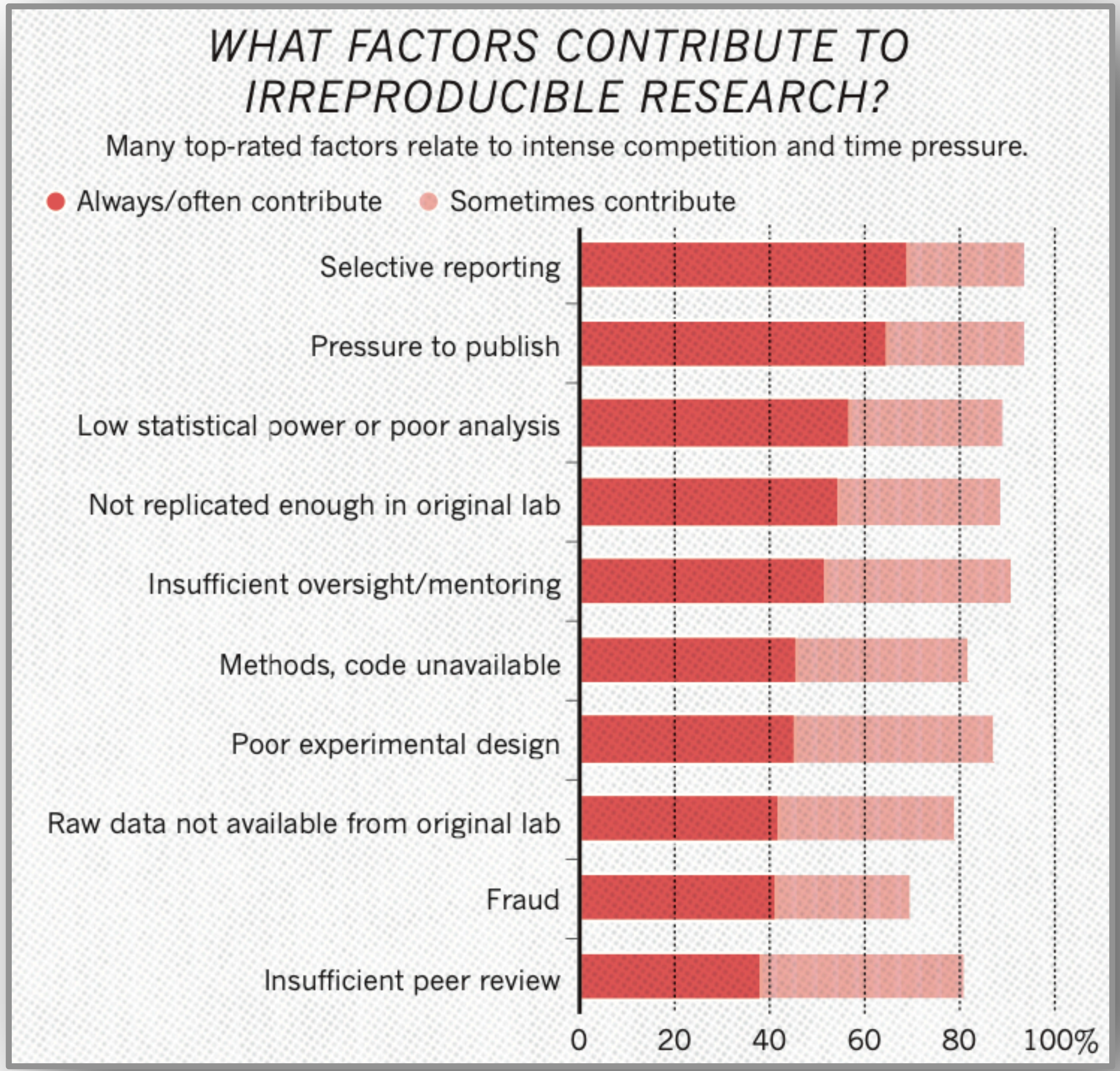
"1500 scientists lift the lid on reproducibility", Baker, Nature, issue 533, 2016

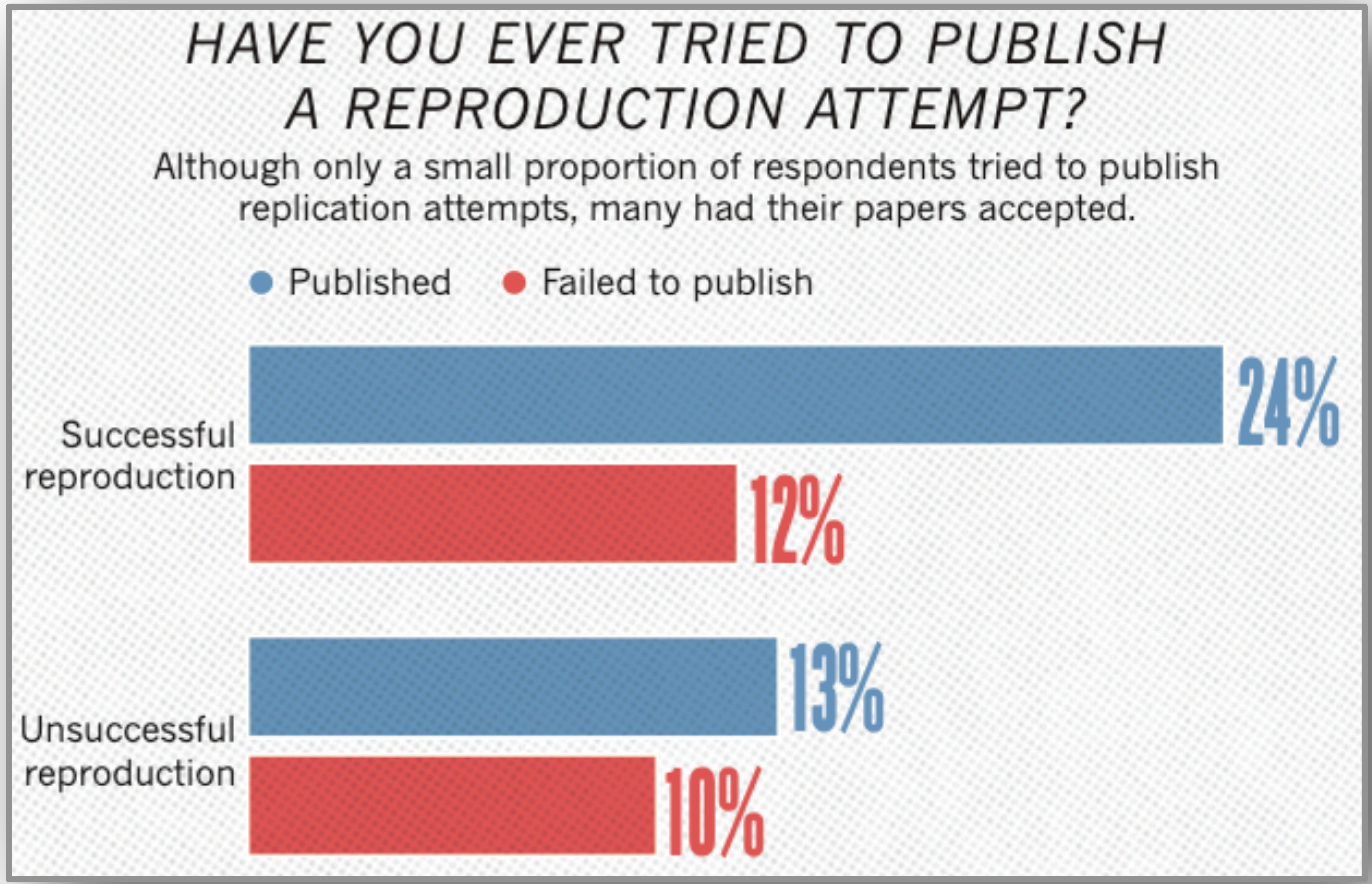
Is reproducibility in a crisis?



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

Is reproducibility in a crisis?





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

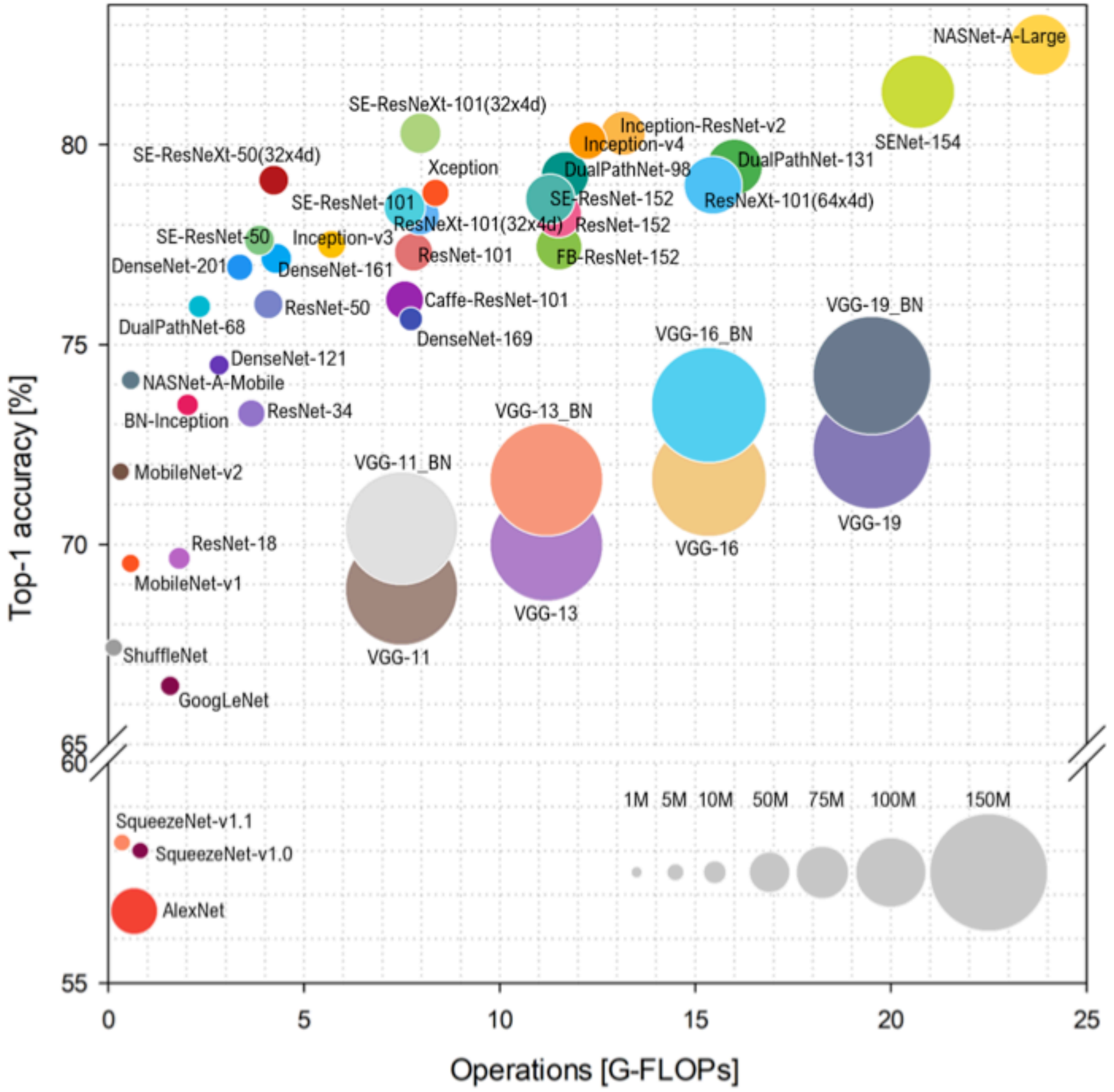
Are we in a crisis in ML too?

ML & the reproducibility 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

Recall: “static” models/data



Bianco et al, “Benchmark Analysis of Representative Deep Neural Network Architectures”, IEEE Access, 2018

Even in “static” scenarios:

- Many aspects of variation/interest!
- Fair comparisons, statistical significance, exhaustive & factual reporting

- (Misaligned?) research incentives
- Code, data, assets, accessibility…

Evaluation

Why is evaluation challenging in machine learning?

Dimensions of evaluation in continual/lifelong learning

Why evaluation is even more challenging in continual/lifelong learning

How can we move forward?

Recall: scenarios so far

What where some of the sequences of tasks we have seen so far?

- 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

Recall: scenarios so far

Benchmarks commonly based on popular vision datasets, language datasets, or reinforcement tasks (such as games)

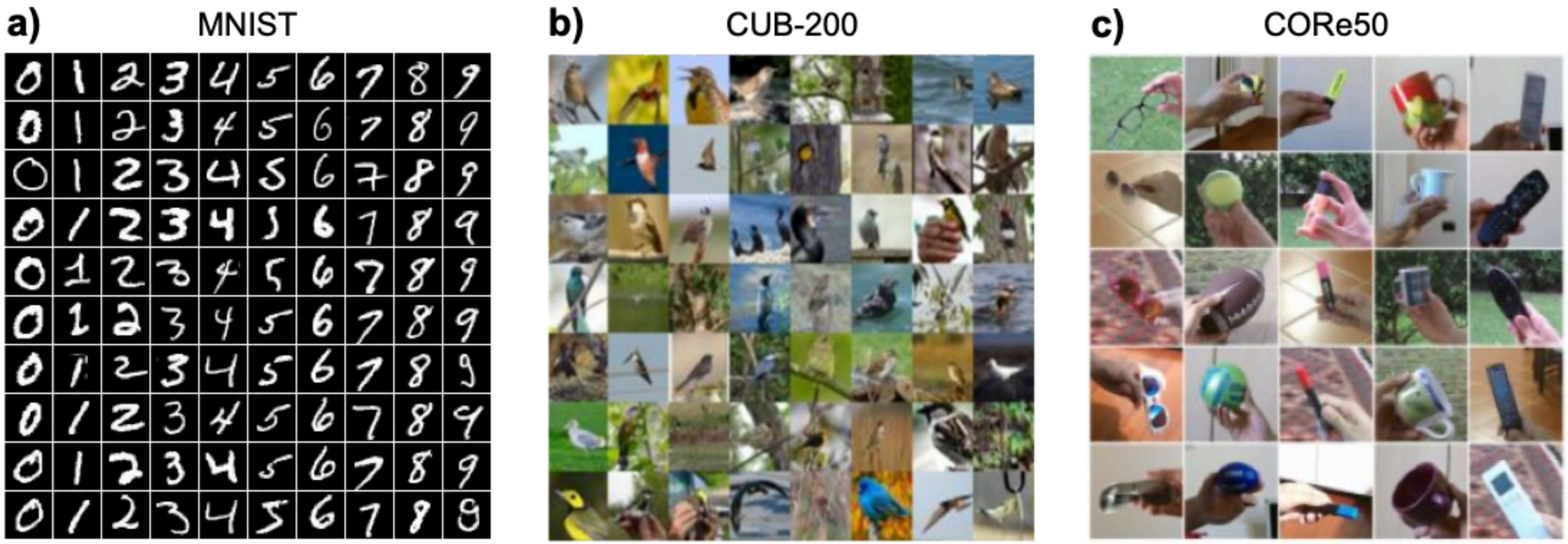


Figure 3: Example images from benchmark datasets used for the evaluation of lifelong learning

Parisi et al, "Continual Lifelong Learning with Neural Networks: A Review", Neural Networks 2019

Recall: scenarios so far

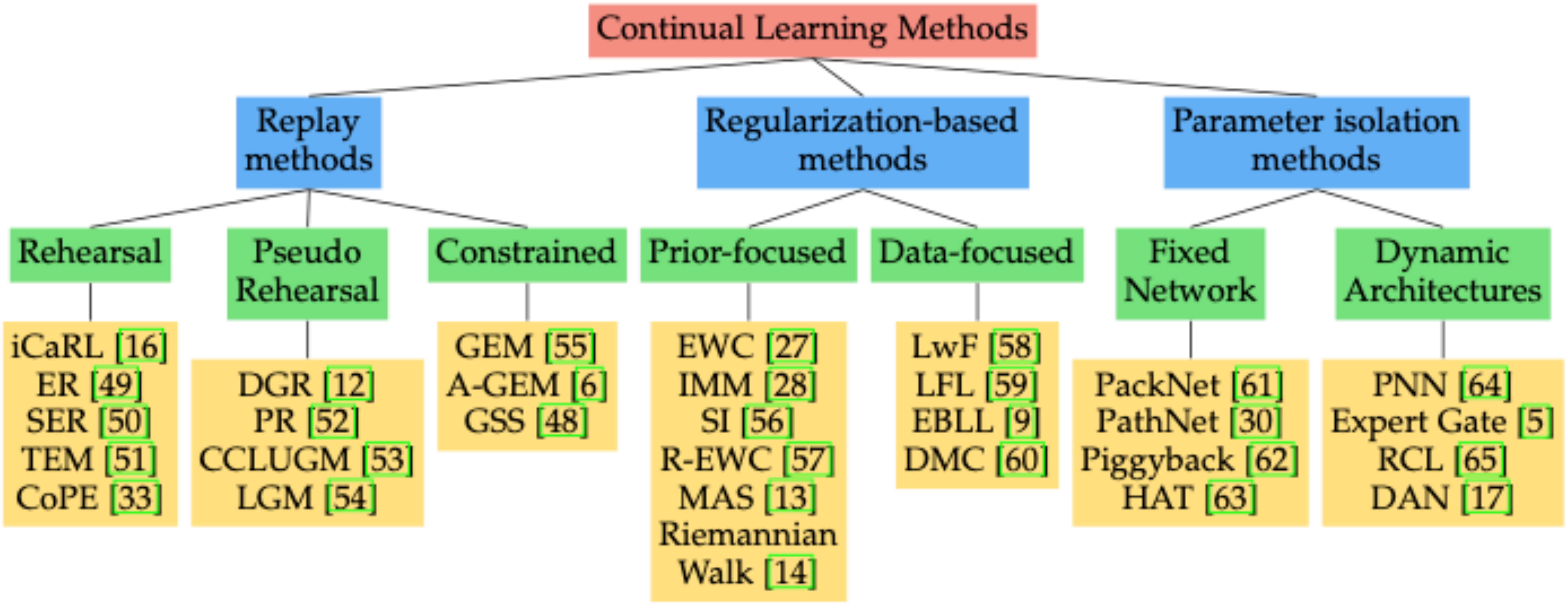
For now: let's assume that we know the sequence of tasks, i.e. a dedicated test set for each "experience/task" exists

Name	*Details*	*Related works*
XCOPA - Cross-lingual Choice of Plausible Alternatives	• a typologically diverse multilingual dataset for causal commonsense reasoning, which is the translation and reannotation • covers 11 languages from distinct families	(Edoardo M. Ponti and Korhonen, 2020)
WEBTEXT	• a dataset of millions of webpages suitable for learning language models without supervision • 45 million links scraped from Reddit, 40 GB dataset	(Radford et al., 2019)
C4 - Colossal Clean Crawled Corpus	• a dataset constructed from Common Crawl's web crawl corpus and serves as a source of unlabeled text data • 17 GB dataset	(Raffel et al., 2020)
LIFELONG FEWREL - Lifelong Few-Shot Relation Classification Dataset	• sentence-relation pairs derived from Wikipedia distributed over 10 disjoint clusters (representing different tasks)	(Wang et al., 2019b) (Obamuyide and Vlachos, 2019)
LIFELONG SIMPLE QUESTIONS	• single-relation questions divided into 20 disjoint clusters (i.e. resulting in 20 tasks)	(Wang et al., 2019b)

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

Recall: forgetting

Depending on choice of method, we will likely be interested in different measures



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

Aspects of the mechanisms

Rehearsal methods:

- What do you think should be here?

Regularization methods:

- …

Architecture/parameter methods:

- …

Aspects of the mechanisms

Rehearsal methods:

- Original data amount, generated data, (constant?) memory size, computational expense…

Regularization methods:

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

Architecture/parameter methods:

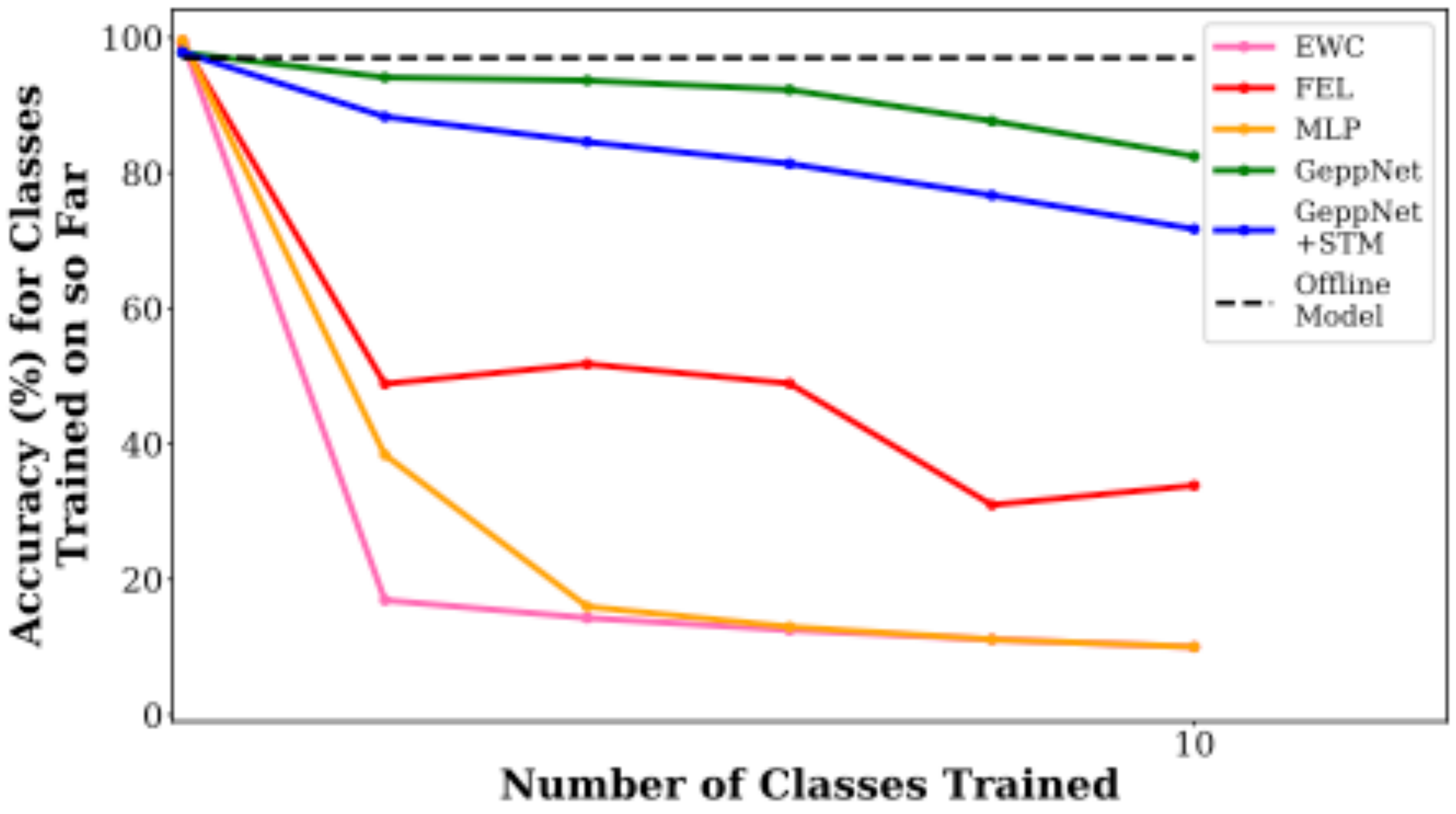
- Number of parameters, number of models, expert heads,
  memory expense, computational expense…

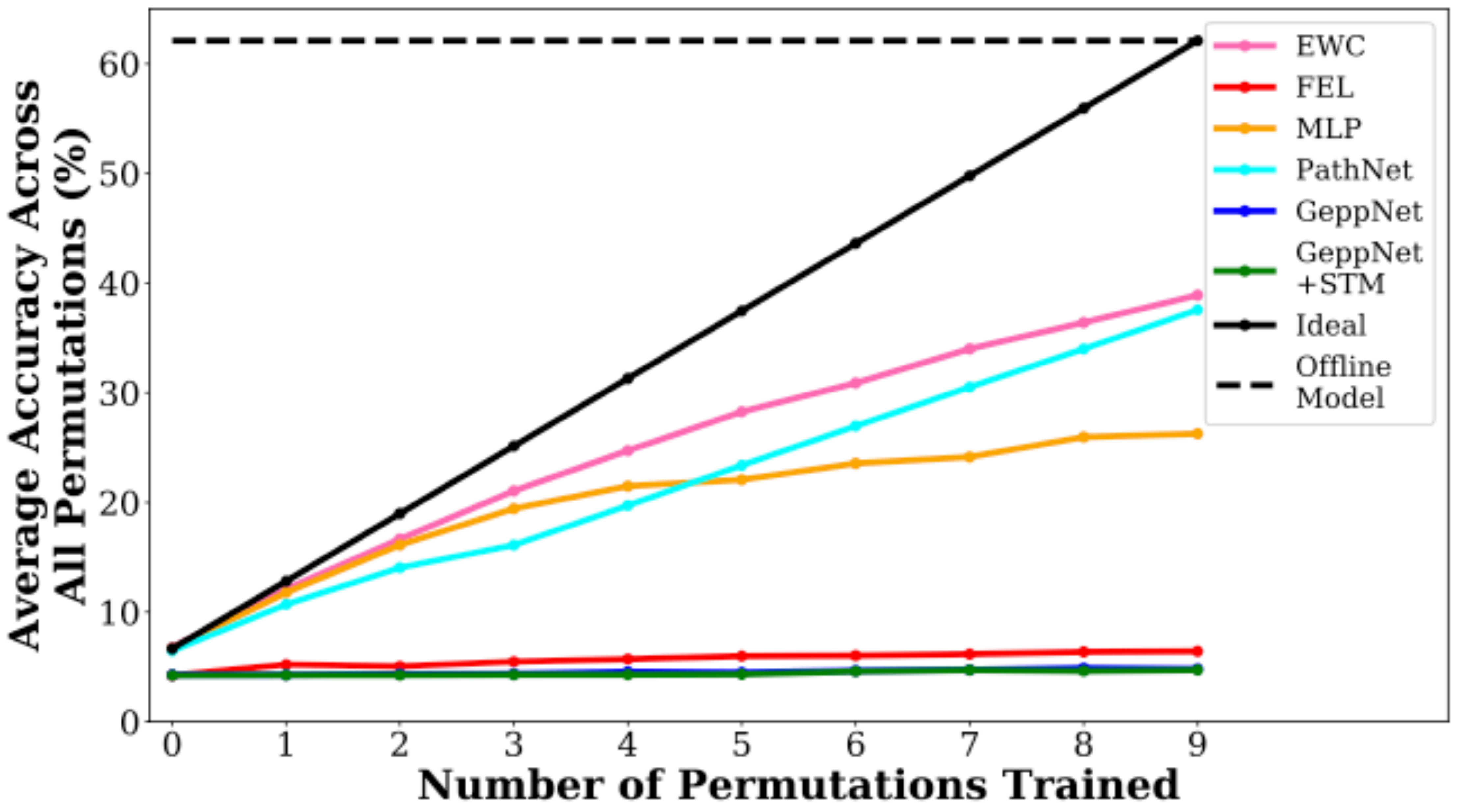
Final average losses seem insufficient

Let’s take a look at some further suggestions

(Some) ways to measure

Do we care about the overall performance?
Or the one up to the current point in time?





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

Per "task" measures

- "**Base**" loss: the initial (an old) task after i new experiences
- "**New**" loss: the newest task only
- "**All**" loss: average up to the present point in time
- "**Ideal**" loss: offline value trained at once

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

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

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

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

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

“Forgetting”

“We define forgetting for a particular task (or label) as the difference between the *maximum* knowledge gained about the task throughout the learning process in the past and the knowledge the model currently has about it.”

(Chaudhry et al, “Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence”, ECCV 2018)

For the j-th task after being trained up to task k > j:

“

$$f_j^k = \max_{l \in \{1, \cdots, k-1\}} a_{l,j} - a_{k,j} \; , \quad \forall j < k$$

“Intransigence”

“We define *intransigence* as the inability of a model to learn new tasks. Since we wish to quantify the *inability* to learn, we compare to the standard classification model which has access to all the datasets at all times”

(Chaudhry et al, “Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence”, ECCV 2018)

For a reference model for task k (denoted by *):

“

$$I_k = a_k^* - a_{k,k}$$

Forward & backward transfer

(Avg.) **Forward transfer** (with random baseline b): influence of a learning task on future tasks;

$$\mathrm{FWT}_{t,j} = a_{t-1,j} - \bar{b}_j \qquad \mathrm{FWT}_t = \frac{1}{t-1}\sum_{j=2}^{t-1} \mathrm{FWT}_{j-1,j}$$

(Avg.) **Backward transfer**: influence of a task on previous tasks; negative = forgetting, positive = retrospective improvement

$$\mathrm{BWT}_{t,j} = a_{t,j} - a_{j,j} \qquad \mathrm{BWT}_t = \frac{1}{t-1}\sum_{j=1}^{t-1} \mathrm{BWT}_{t,j}$$

R	Te_1	Te_2	Te_3
Tr_1	R^*	R_{ij}	R_{ij}
Tr_2	R_{ij}	R^*	R_{ij}
Tr_3	R_{ij}	R_{ij}	R^*

Lopez-Paz & Ranzato, “Gradient Episodic Memory for Continual Learning”, 2017,
See also: Díaz-Rodríguez & Lomonaco et al, "Don’t forget, there is more than forgetting: new metrics for Continual Learning", 2018

Forward & backward transfer

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

$$Z_b = \frac{1}{T} \sum_{k=1}^{T} a_{k,b,k}$$

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

Memory, size & compute

We can construct similar measures for memory, size & compute (Here tasks are called 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)

There are plenty of other interesting ideas of what to measure

Evaluation

Why is evaluation challenging in machine learning?

Dimensions of evaluation in continual/lifelong learning

Why evaluation is even more challenging in continual/lifelong learning

How can we move forward?

What should we report now?

The challenge of comparison

How do we compare & draw conclusions with various metrics + set-ups?

Model	Dataset	Data Permutation Ω_{base}	Data Permutation Ω_{new}	Data Permutation Ω_{all}	Incremental Class Ω_{base}	Incremental Class Ω_{new}	Incremental Class Ω_{all}	Multi-Modal Ω_{base}	Multi-Modal Ω_{new}	Multi-Modal Ω_{all}	Memory Constraints	Model Size (MB)
MLP	**MNIST**	0.434	0.996	0.702	0.060	1.000	0.181	N/A	N/A	N/A	Fixed-size	1.91
	CUB	0.488	0.917	0.635	0.020	1.000	0.031	0.327	0.412	0.610		4.24
	AS	0.186	0.957	0.446	0.016	1.000	0.044	0.197	0.609	0.589		2.85
EWC	**MNIST**	0.437	0.992	0.746	0.001	1.000	0.133	N/A	N/A	N/A	Fixed-size	3.83
	CUB	0.765	0.869	0.762	0.105	0.000	0.094	0.944	0.369	0.872		8.48
	AS	0.129	0.687	0.251	0.021	0.580	0.034	1.000	0.588	0.984		5.70
PathNet	**MNIST**	0.687	0.887	0.848	N/A	N/A	N/A	N/A	N/A	N/A	New output layer for each task	2.80
	CUB	0.538	0.701	0.655	N/A	N/A	N/A	0.908	0.376	0.862		7.46
	AS	0.414	0.750	0.615	N/A	N/A	N/A	0.069	0.540	0.469		4.68
GeppNet	**MNIST**	0.912	0.242	0.364	0.960	0.824	0.922	N/A	N/A	N/A	Stores all training data	190.08
	CUB	0.606	0.029	0.145	0.628	0.640	0.585	0.156	0.010	0.089		53.48
	AS	0.897	0.170	0.343	0.984	0.458	0.947	0.913	0.005	0.461		150.38
GeppNet+STM	**MNIST**	0.892	0.212	0.326	0.919	0.599	0.824	N/A	N/A	N/A	Stores all training data	191.02
	CUB	0.615	0.020	0.142	0.727	0.232	0.626	0.031	0.329	0.026		55.94
	AS	0.820	0.041	0.219	1.007	0.355	0.920	0.829	0.005	0.418		151.92
FEL	**MNIST**	0.117	0.990	0.279	0.451	1.000	0.439	N/A	N/A	N/A	Fixed-size	4.54
	CUB	0.043	0.764	0.184	0.316	1.000	0.361	0.110	0.329	0.412		6.16
	AS	0.081	0.848	0.239	0.283	1.000	0.240	0.473	0.320	0.494		6.06

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

The challenge of comparison

How do we compare & draw conclusions with various metrics + set-ups?

Model	Incremental Class	Similar Data	Dissimilar Data	Memory Efficient	Trains Quickly
MLP	✗	✗	✗	✓	✓
EWC	✗	✗	✓	✓	✓
PathNet	✗	✓	✗	✗	✗
GeppNet	✓	✗	✗	✗	✗
GeppNet+STM	✓	✗	✗	✗	✗
FEL	✗	✗	✗	✗	✓

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

The challenge of comparison

How do we compare & draw conclusions with various metrics + set-ups?

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

Unfortunately, it's not just about what to measure!

It's about assumptions, trade-offs, benchmarks,…

Should we strive for specific benchmarks & overall consensus or transparency?

Crisis worse in lifelong ML?

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

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

1. We propose fundamental desiderata for future evaluations, which can be applied regardless of dataset.
2. We analyse the shortcomings of existing widely used evaluations in continual learning.

"Towards Robust Evaluations of Continual Learning", Farquhar & Gal, Lifelong Learning workshop at ICML 2018

The challenge of defining a “task”

Challenge of defining a “task”

It’s not just challenging to compare across multiple metrics, it’s also challenging to agree on what “tasks” should be

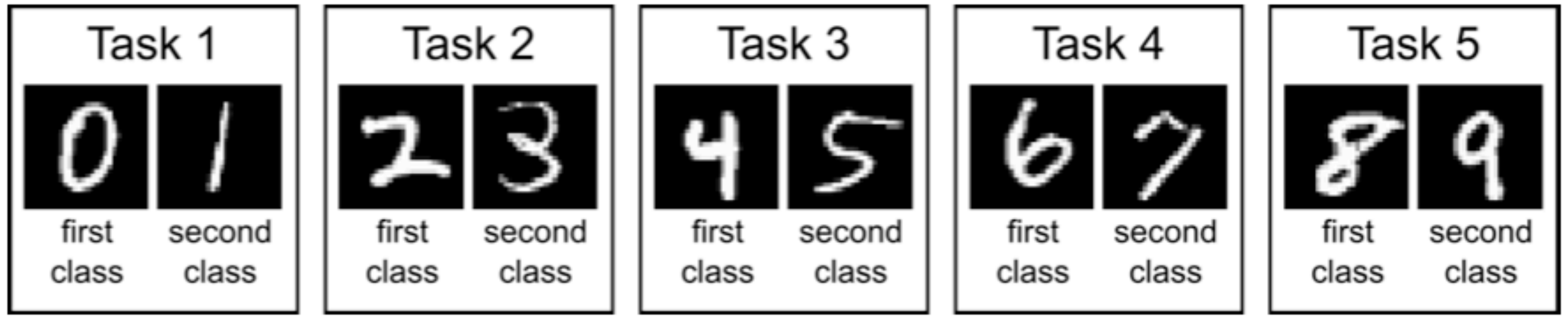


Figure 1: Schematic of split MNIST task protocol.

Table 2: Split MNIST according to each scenario.

Task-IL	With task given, is it the 1st or 2nd class? (e.g., 0 or 1)
Domain-IL	With task unknown, is it a 1st or 2nd class? (e.g., in $[0, 2, 4, 6, 8]$ or in $[1, 3, 5, 7, 9]$)
Class-IL	With task unknown, which digit is it? (i.e., choice from 0 to 9)

Table 1: Overview of the three continual learning scenarios.

Scenario	***Required at test time***
Task-IL	Solve tasks so far, task-ID provided
Domain-IL	Solve tasks so far, task-ID not provided
Class-IL	Solve tasks so far *and* infer task-ID

van de Ven & Tolias, “Three scenarios for continual learning”, arXiv:1904.07734, 2019

Challenge of defining a "task"

It's not just challenging to compare across multiple metrics, it's also challenging to agree on what "tasks" should be

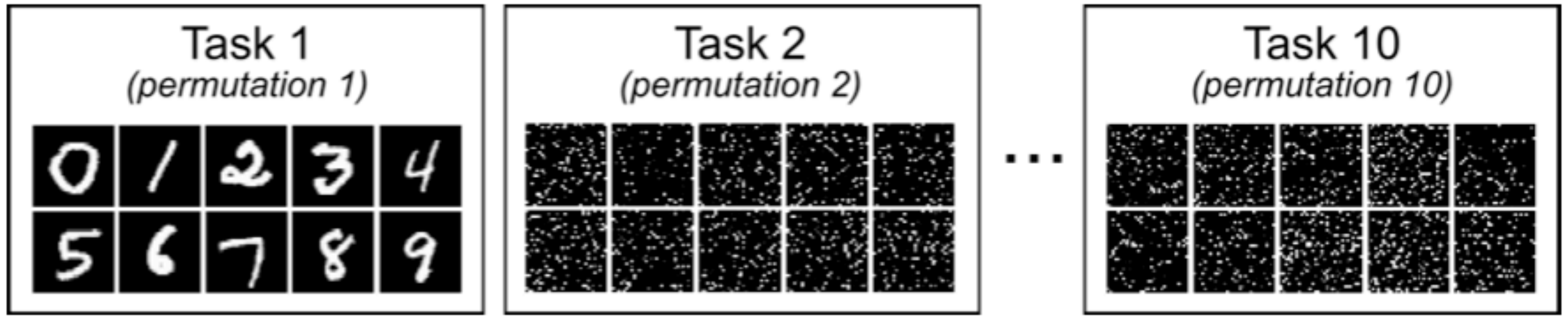


Figure 2: Schematic of permuted MNIST task protocol.

Table 3: Permuted MNIST according to each scenario.

Task-IL	Given permutation X, which digit?
Domain-IL	With permutation unknown, which digit?
Class-IL	Which digit *and* which permutation?

Table 1: Overview of the three continual learning scenarios.

Scenario	***Required at test time***
Task-IL	Solve tasks so far, task-ID provided
Domain-IL	Solve tasks so far, task-ID not provided
Class-IL	Solve tasks so far *and* infer task-ID

van de Ven & Tolias, "Three scenarios for continual learning", arXiv:1904.07734, 2019

Recall: expert heads

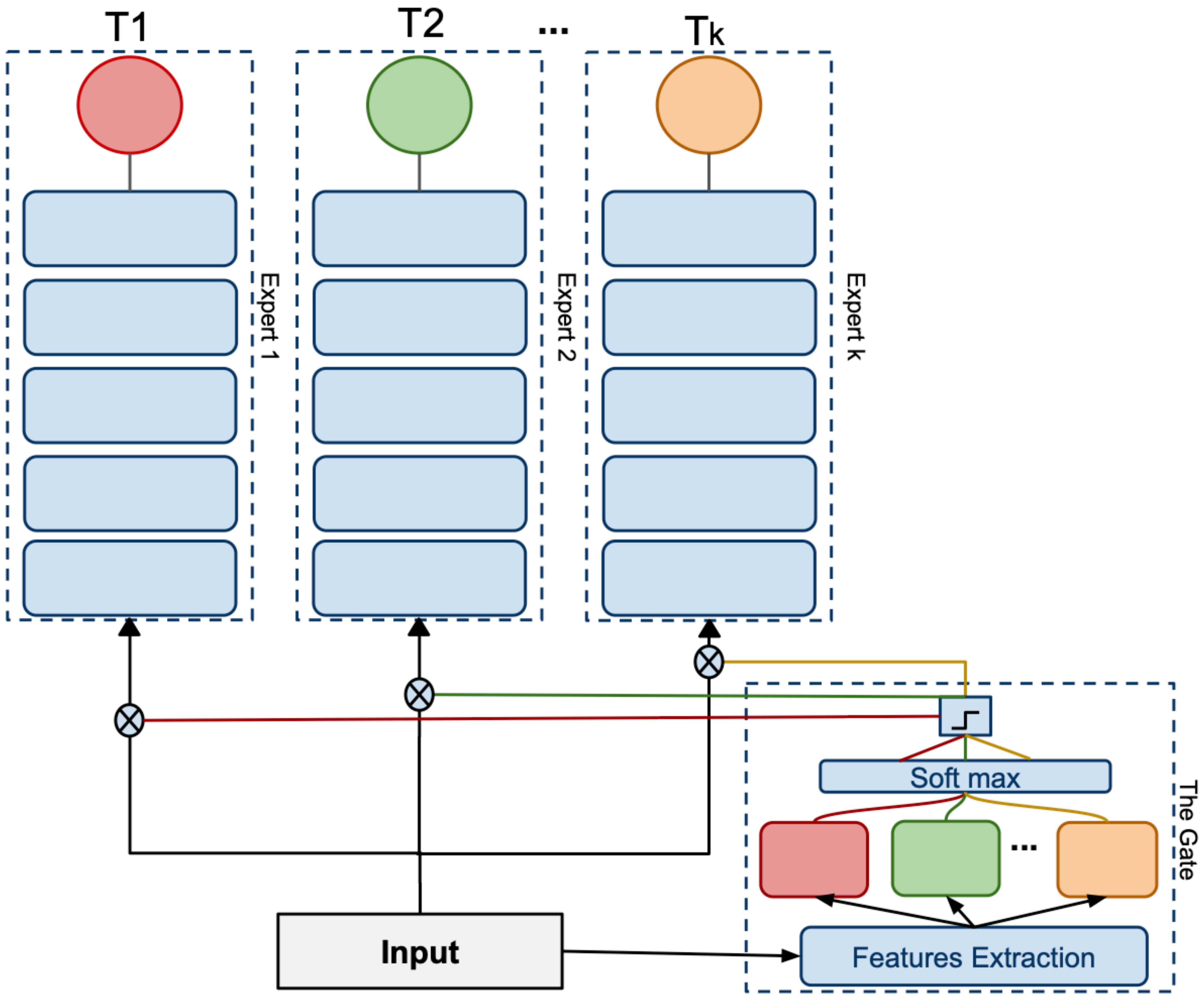


Figure 1. The architecture of our Expert Gate system.

Aljundi et al, "Expert Gate: Lifelong Learning with a Network of Experts", CVPR 2017

Why does such a scenario/"task" distinction even matter?

Recall the "**experts**" approach:

- We could share parts + add individual experts on top

The challenge of expert heads

Expert heads often evaluated from a “forgetting only" perspective.
Not only test set for each “experience/task, but also the task id is provided!

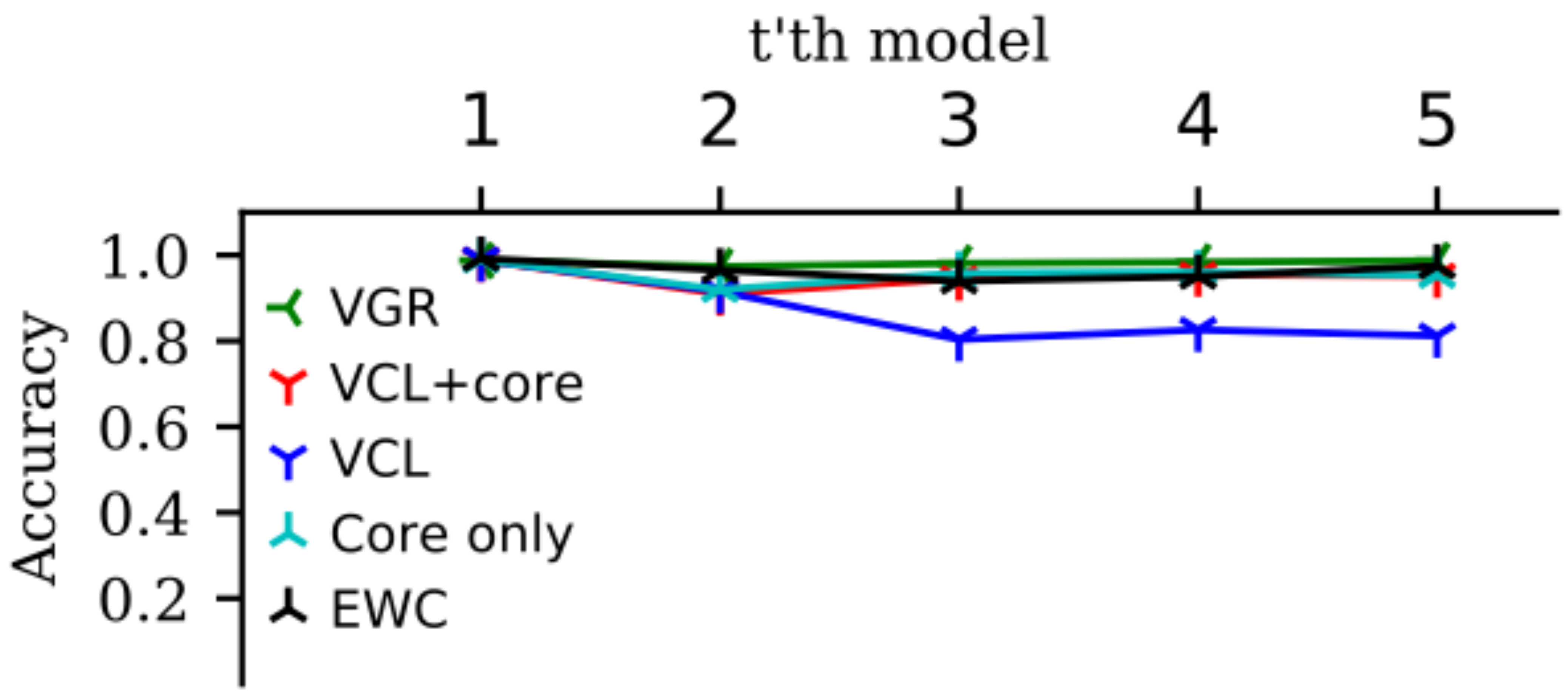


Figure 5. **Multi-headed Split FashionMNIST.**

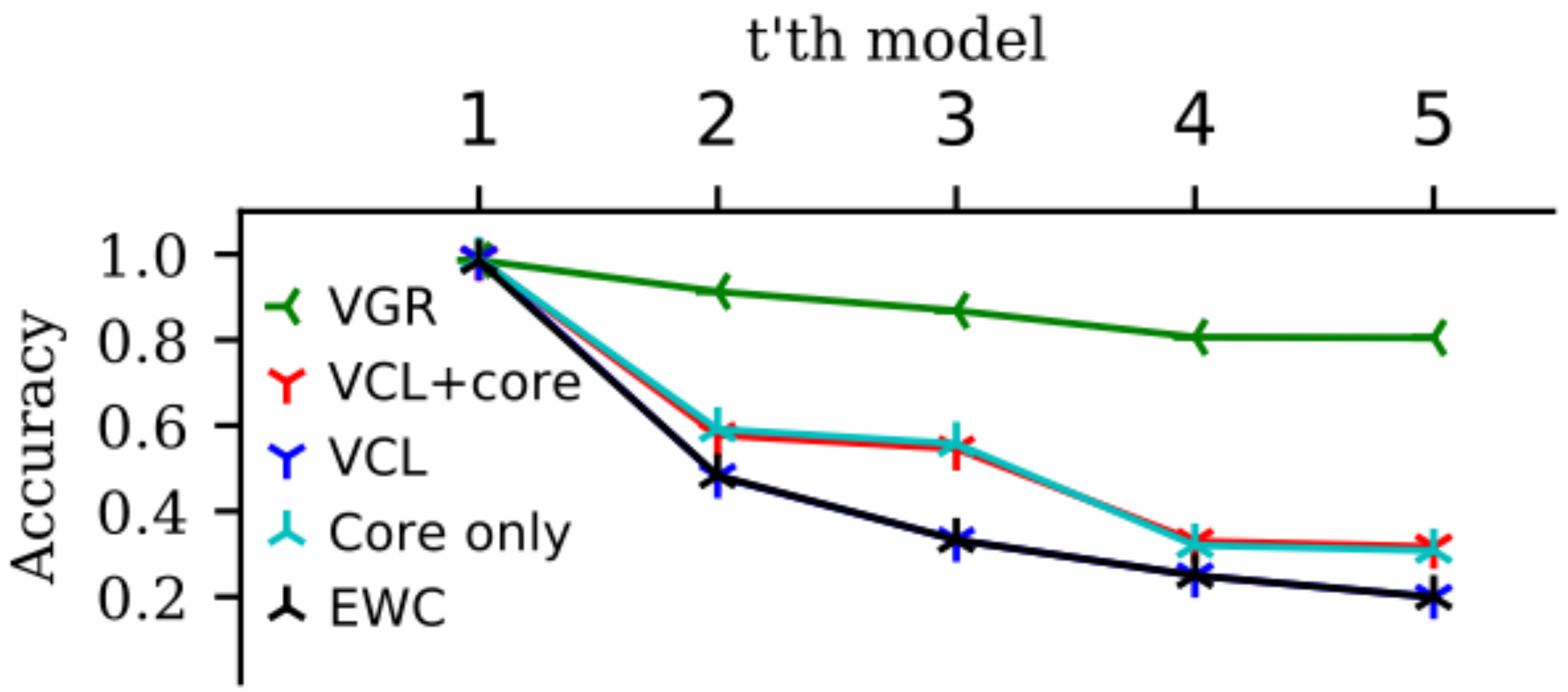


Figure 3. **Single-headed Split Fashion MNIST.**

Farquhar & Gal, “Towards Robust Evaluations of Continual Learning”, Lifelong Learning workshop at ICML 2018

The challenge of expert heads

Expert heads often evaluated from a “forgetting only" perspective.
Not only test set for each “experience/task, but also the task id is provided!

Approach	Method	Task-IL	Domain-IL	Class-IL
Baselines	*None – lower bound*	*87.19 (± 0.94)*	*59.21 (± 2.04)*	*19.90 (± 0.02)*
	Offline – upper bound	*99.66 (± 0.02)*	*98.42 (± 0.06)*	*97.94 (± 0.03)*
Task-specific	XdG	99.10 (± 0.08)	-	-
Regularization	EWC	98.64 (± 0.22)	63.95 (± 1.90)	20.01 (± 0.06)
	Online EWC	99.12 (± 0.11)	64.32 (± 1.90)	19.96 (± 0.07)
	SI	99.09 (± 0.15)	65.36 (± 1.57)	19.99 (± 0.06)
Replay	LwF	99.57 (± 0.02)	71.50 (± 1.63)	23.85 (± 0.44)
	DGR	99.50 (± 0.03)	95.72 (± 0.25)	90.79 (± 0.41)
	DGR+distill	99.61 (± 0.02)	96.83 (± 0.20)	91.79 (± 0.32)
Replay + Exemplars	iCaRL (budget = 2000)	-	-	94.57 (± 0.11)

van de Ven & Tolias, “Three scenarios for continual learning”, arXiv:1904.07734, 2019

The challenge of hyper-parameters in continual learning

The challenge of hyper-params

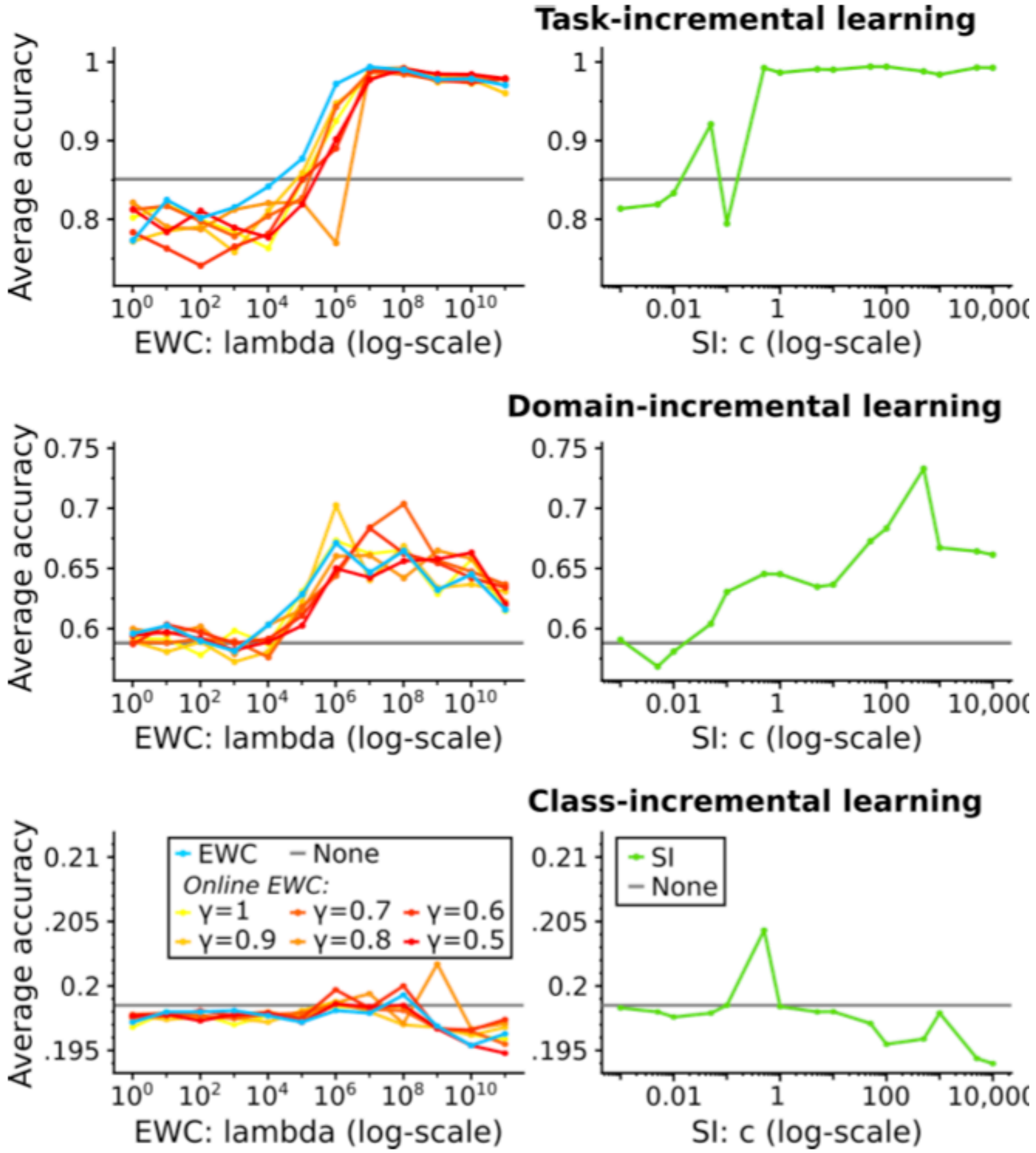
Algorithm 1 Learning and Evaluation Protocols

for h **in** hyper-parameter list **do** $\triangleright$ Cross-validation loop, executing multiple passes over $\mathcal{D}^{CV}$
    for $k = 1$ **to** T^{CV} **do** $\triangleright$ Learn over data stream $\mathcal{D}^{CV}$ using h
        for $i = 1$ **to** n_k **do** $\triangleright$ Single pass over $\mathcal{D}_k$
            Update f_θ using $(\mathbf{x}_i^k, t_i^k, y_i^k)$ and hyper-parameter h
            Update metrics on test set of $\mathcal{D}^{CV}$
        end for
    end for
end for
Select best hyper-parameter setting, h^*, based on average accuracy of test set of $\mathcal{D}^{CV}$, see Eq. $\boxed{1}$.
Reset f_θ.
Reset all metrics.
for $k = T^{CV} + 1$ **to** T **do** $\triangleright$ Actual learning over datastream $\mathcal{D}^{EV}$
    for $i = 1$ **to** n_k **do** $\triangleright$ Single pass over $\mathcal{D}_k$
        Update f_θ using $(\mathbf{x}_i^k, t_i^k, y_i^k)$ and hyper-parameter h^*
        Update metrics on test set of $\mathcal{D}^{EV}$
    end for
end for
Report metrics on test set of $\mathcal{D}^{EV}$.

There are more set-up assumptions: how do we select the continual hyper-parameters?

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

The challenge of hyper-params



There are more set-up assumptions: how do we select the continual hyper-parameters?

Recall: plasticity - sensitivity trade-off (algorithms such as EWC, SI, etc.)

$$L(\theta) = L_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

van de Ven & Tolias, “Three scenarios for continual learning”, arXiv:1904.07734, 2019

The challenge of formulating desiderata: consensus

Continual learning desiderata?

The challenge of consensus. Is it possible to postulate general desiderata?

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, strict privacy guarantees…

Continual learning desiderata?

The challenge of consensus. Is it possible to postulate general desiderata?

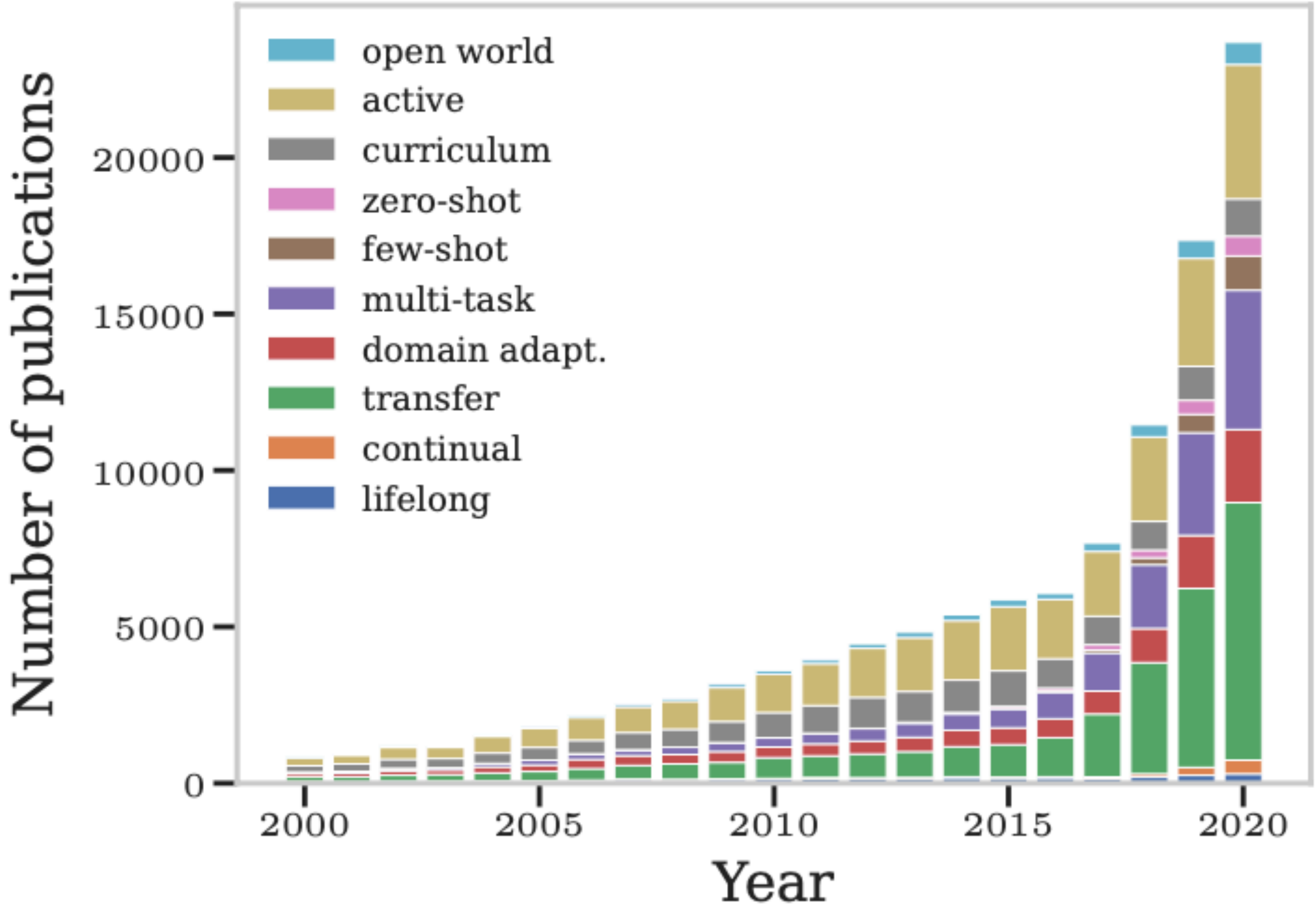
Property	*Definition*
Knowledge retention	The model is not prone to catastrophic forgetting.
Forward transfer	The model learns a new task while reusing knowledge acquired from previous tasks.
Backward transfer	The model achieves improved performance on previous tasks after learning a new task.
On-line 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

Assumptions, assumptions, assumptions…

Recall Lecture 1: continual ML



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

Why are there so many possible assumptions & things to measure?!

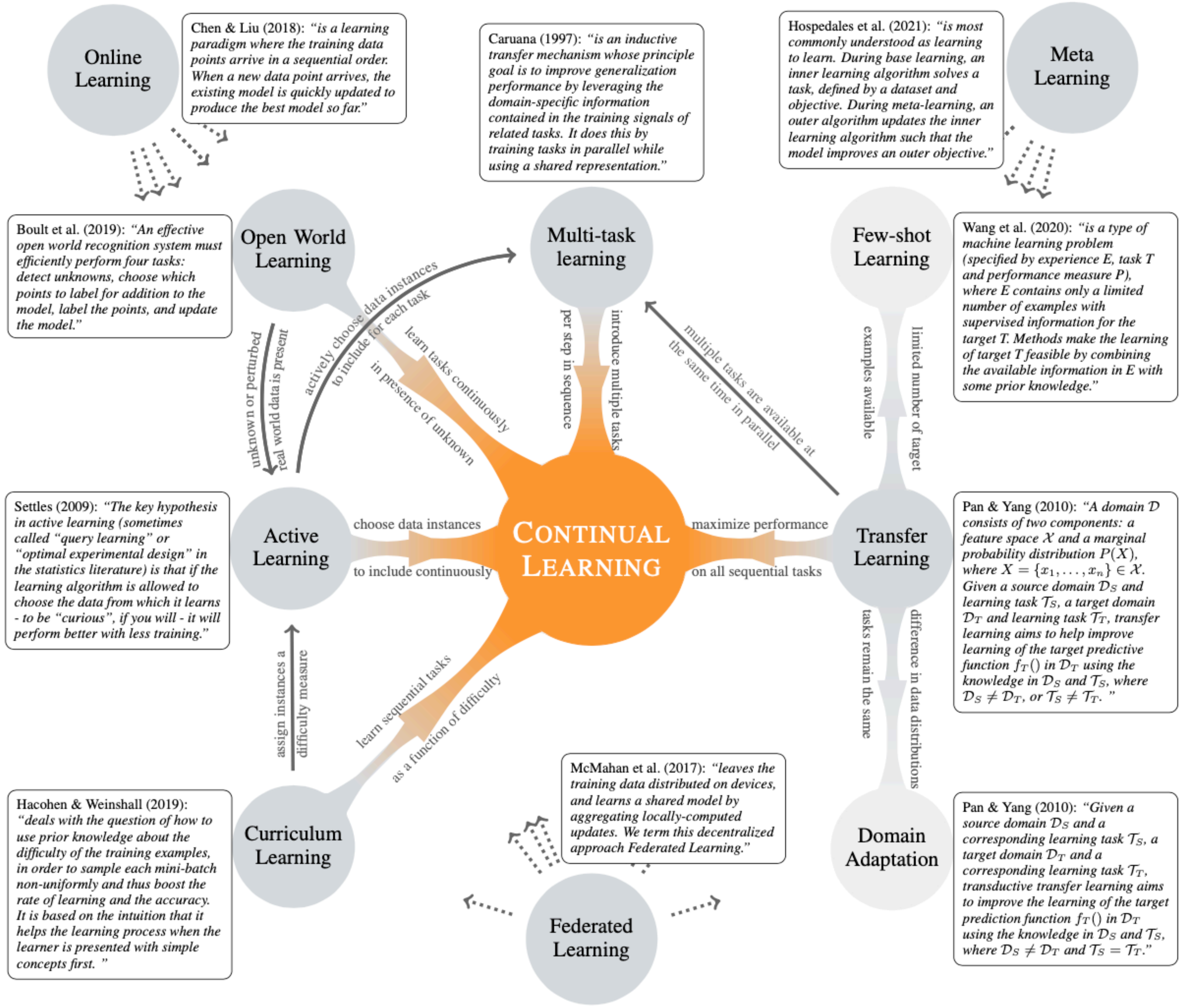
Let’s remind ourselves where they come from & the reason why we have waited to discuss evaluation for 7 weeks

Evaluation & related paradigms

The differences between machine learning paradigms with continuous components can be nuances

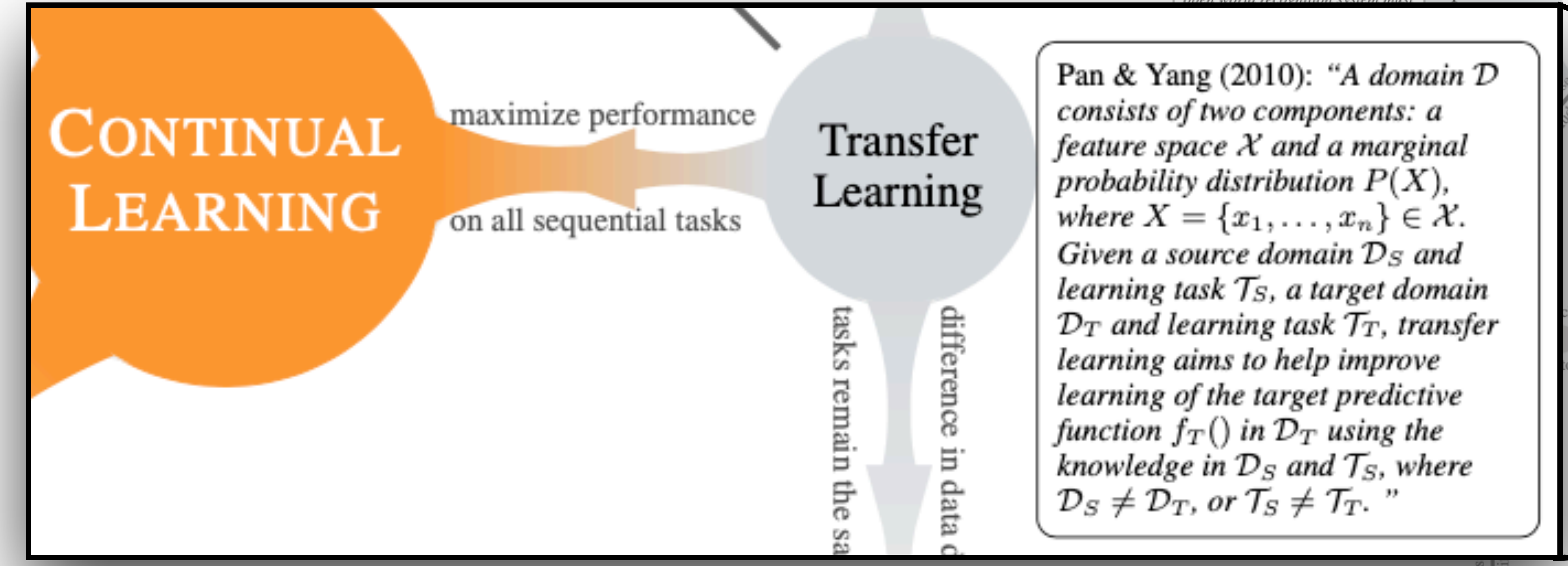
Key aspects often reside in how we evaluate

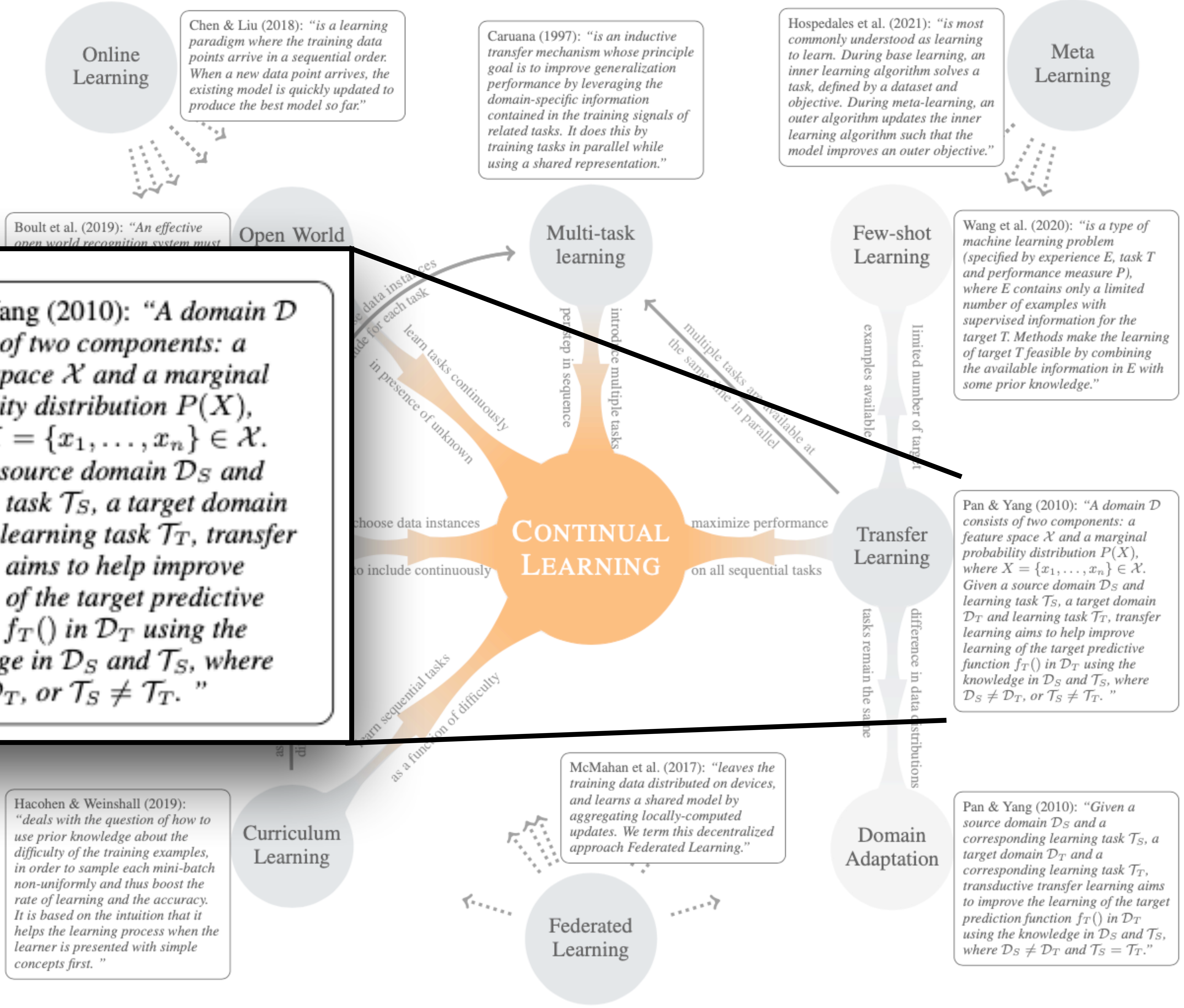
Each paradigm seems to have a particular preference (potentially neglecting other important factors)



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

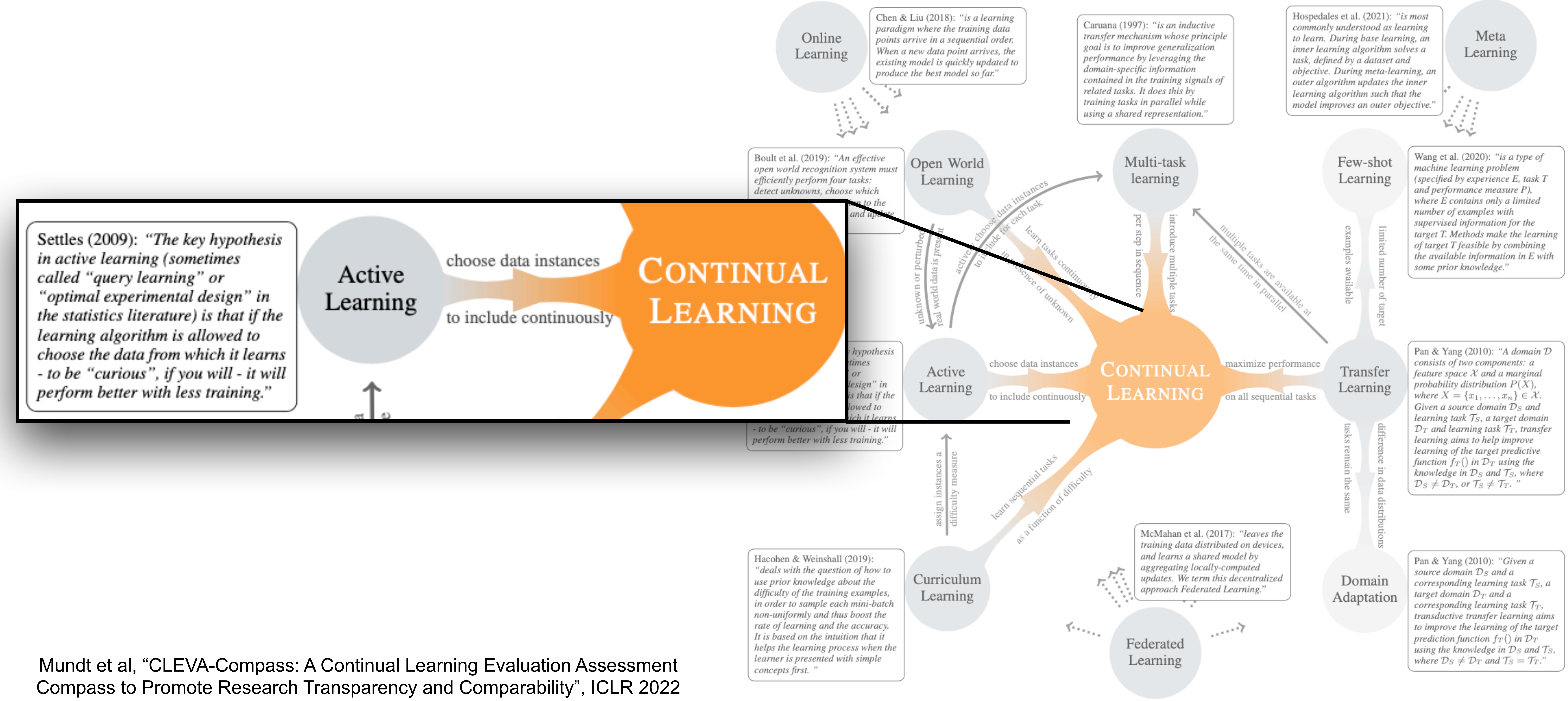
Evaluation & related paradigms





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

Evaluation & related paradigms



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

In all honesty, it is presently challenging to assess continual/lifelong learning systems

Evaluation

Science and evaluation: are we in a crisis? (Have we always been?)

Why is evaluation challenging in machine learning?

Different/additional dimensions of evaluation in continual/lifelong learning

Why evaluation is even more challenging in continual/lifelong learning

How can we move forward?

NeurIPS checklist

Whether a crisis or not, there is much room for general improvement!
… on the incentives & presentation part …

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]**
   (b) Did you describe the limitations of your work? **[TODO]**
   (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]**
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
   (b) Did you include complete proofs of all theoretical results? **[TODO]**

Checklist blog: https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b , checklist taken from formatting instructions

NeurIPS checklist

Whether a crisis or not, there is much room for general improvement!
… on the empirical experimentation parts …

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]**

Checklist blog: https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b , checklist taken from formatting instructions

NeurIPS checklist

Whether a crisis or not, there is much room for general improvement!
… and on many other fronts: assets, data, ethics etc.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**
   (b) Did you mention the license of the assets? **[TODO]**
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**

Checklist blog: https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b , checklist taken from formatting instructions

Reproduction & replication

ML Reproducibility Challenge 2021

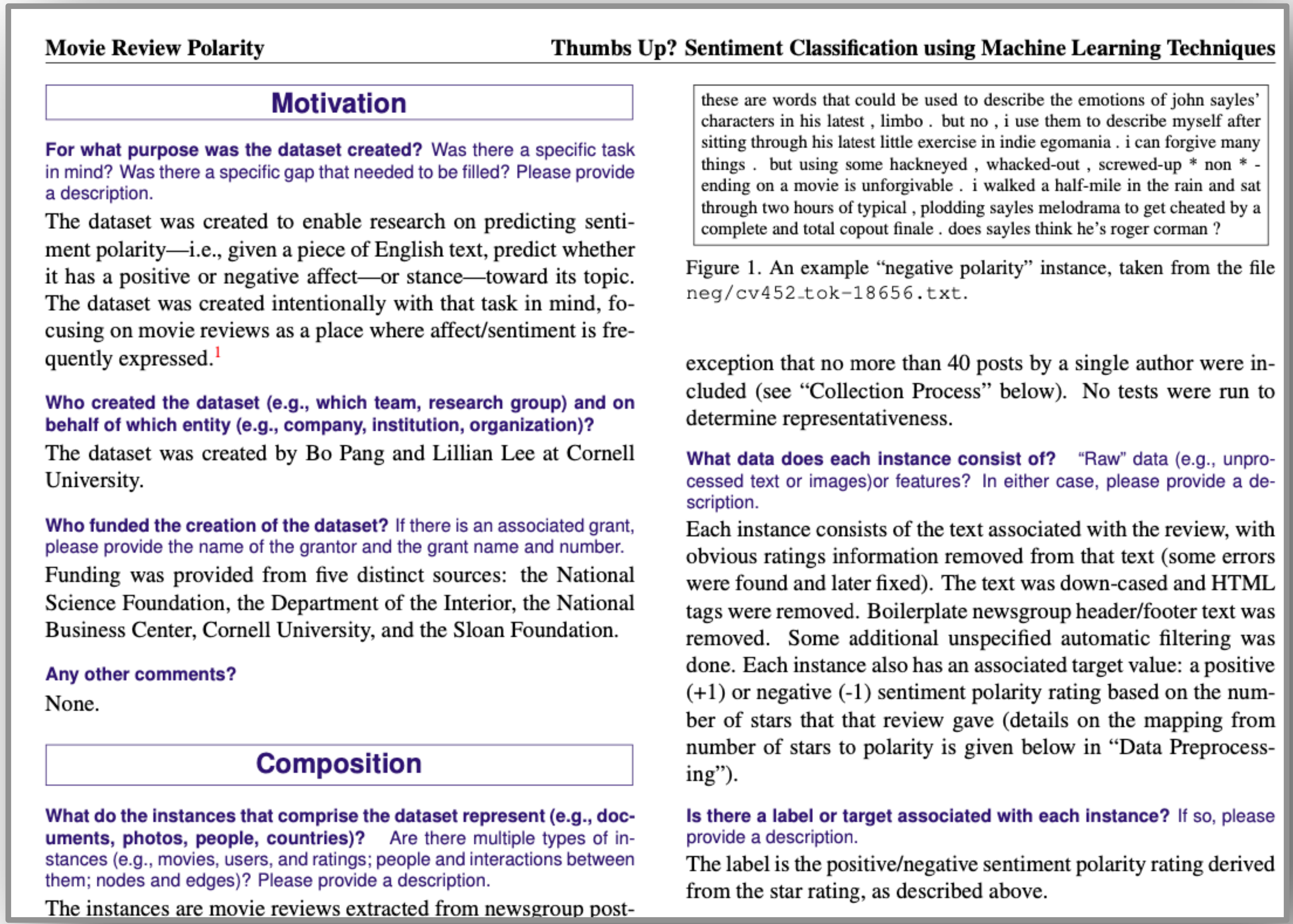
Welcome to the ML Reproducibility Challenge 2021 Fall Edition! This is the fifth edition of this event, and a successor of the ML Reproducibility Challenge 2020 (and previous editions V1, V2, V3), and we are excited this year to broaden our coverage of conferences and papers to cover **nine** top venues of 2021, including: NeurIPS, ICML, ICLR, ACL-IJCNLP, EMNLP, CVPR, ICCV, AAAI and IJCAI.

The primary goal of this event is to encourage the publishing and sharing of scientific results that are reliable and reproducible. In support of this, the objective of this challenge is to investigate reproducibility of papers accepted for publication at top conferences by inviting members of the community at large to select a paper, and verify the empirical results and claims in the paper by reproducing the computational experiments, either via a new implementation or using code/data or other information provided by the authors.

https://paperswithcode.com/rc2021

Dataset sheets & model cards

Dataset sheets



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 removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

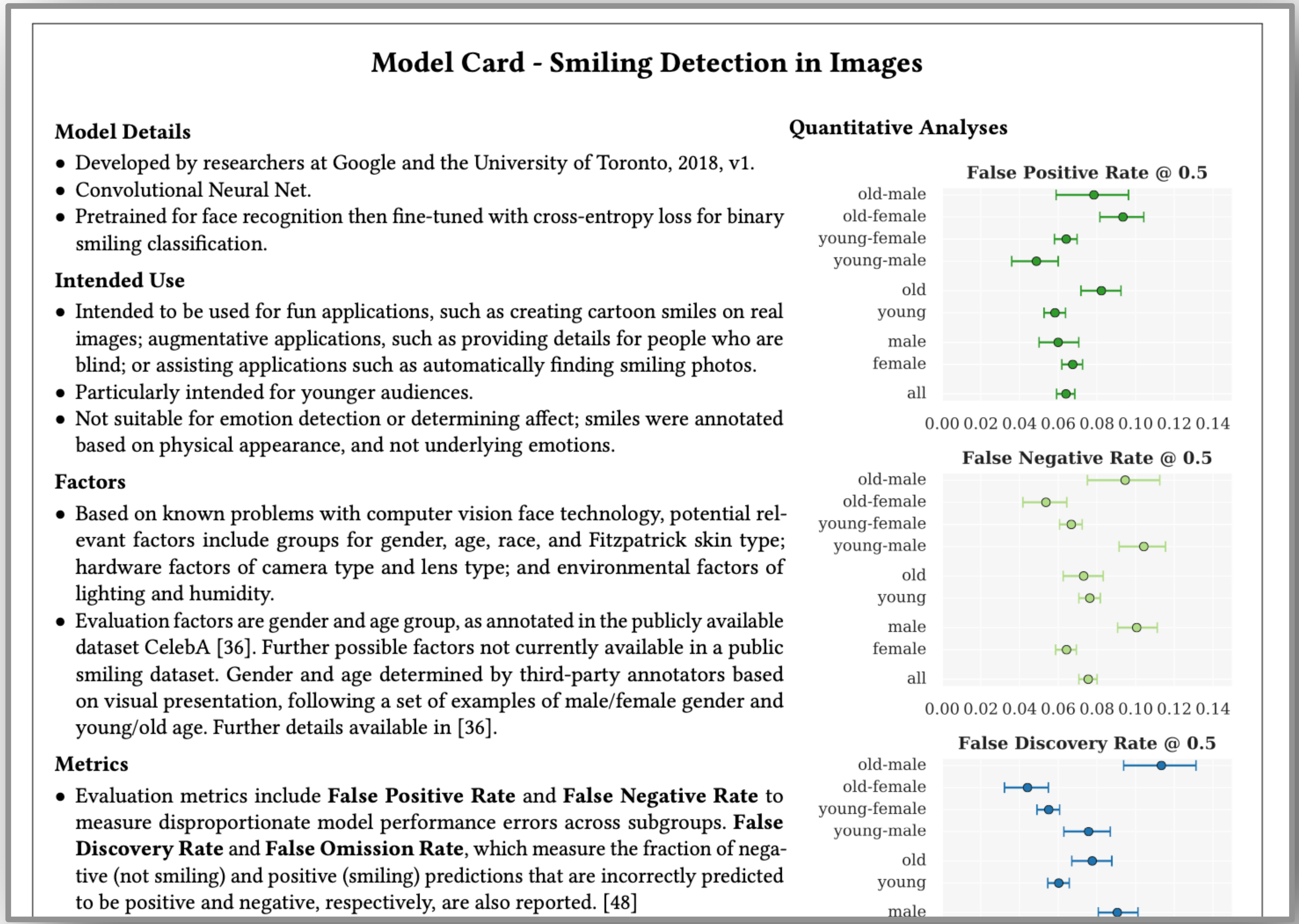
The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

"Datasheets for Datasets", Gebru et al, CACM 2021

Specify motivation, composition, collection process, pre-processing, cleaning, labeling, distribution, maintenance, ethical considerations etc.

Dataset sheets & model cards

Model cards


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.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]

Quantitative Analyses

"Model Cards for Model Reporting", Mitchell et al, FAccT 2019

Specify model details, intended use, human-centric application intent, organization developing the model; considerations on deployment, limits, and ethics; descriptions of metrics, model version, license etc.

Reporting limitations

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.

Limitations

A sign of bad research or an exercise of self-reflection?

Smith et al, “REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research”, FAccT 2022

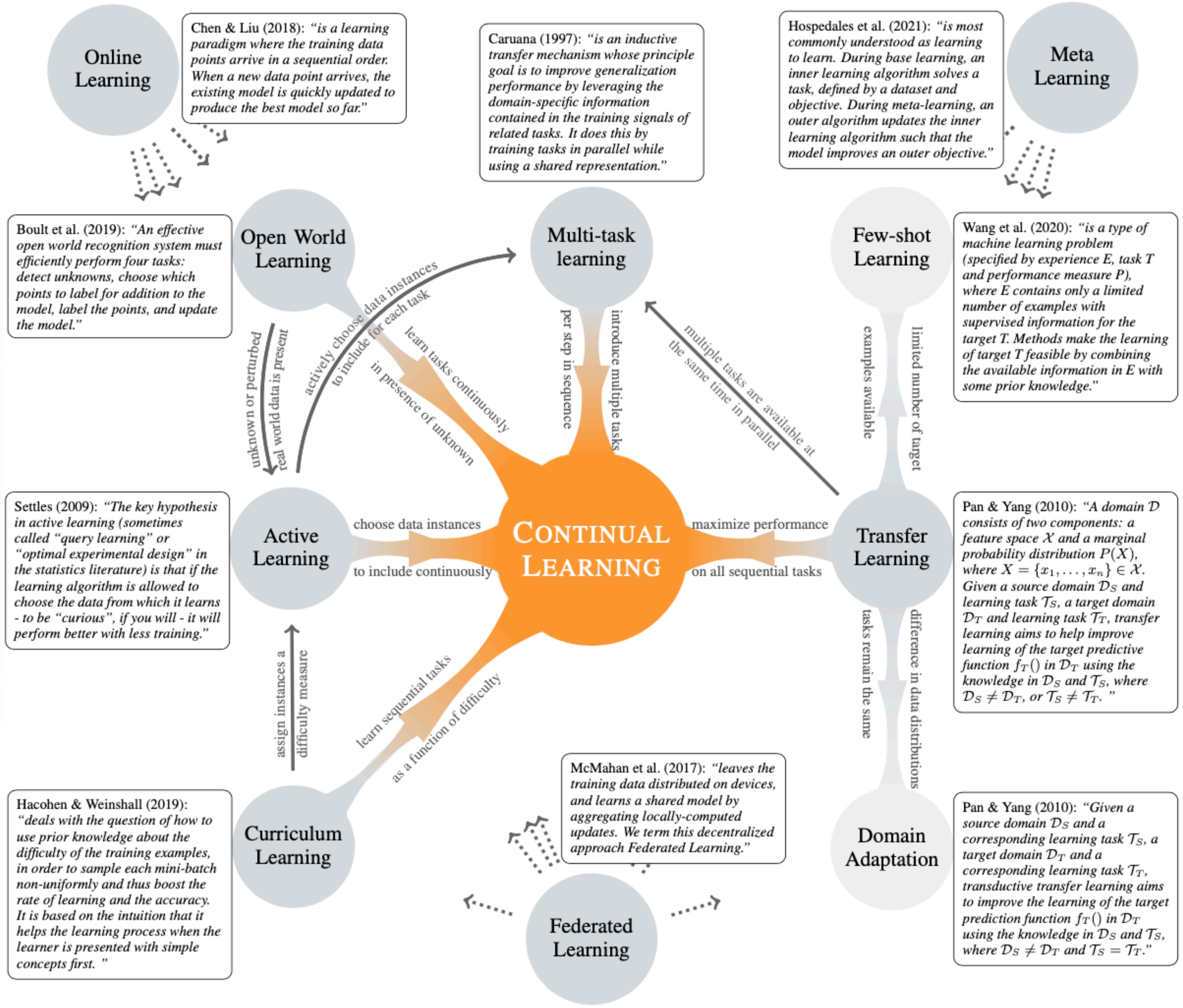
Important note: previous efforts are largely yet to develop for continual/lifelong learning

Evaluation & related paradigms

ContinualAI hessian.AI TECHNISCHE UNIVERSITÄT DARMSTADT

Do distinct applications warrant the existence of numerous scenarios?

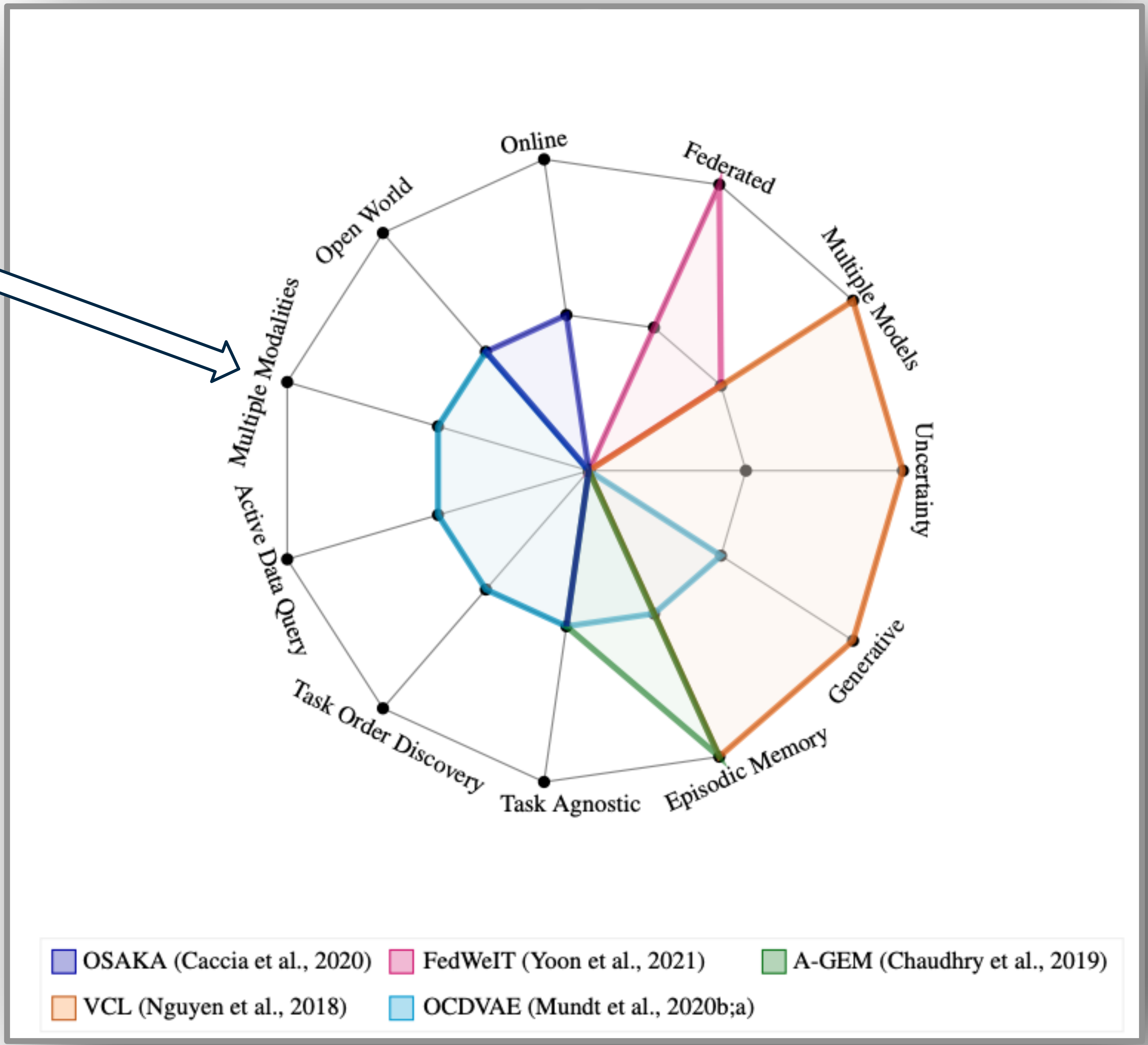
—> Make inspiration in set-up transparent and promote comparability!



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

CLEVA-Compass

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

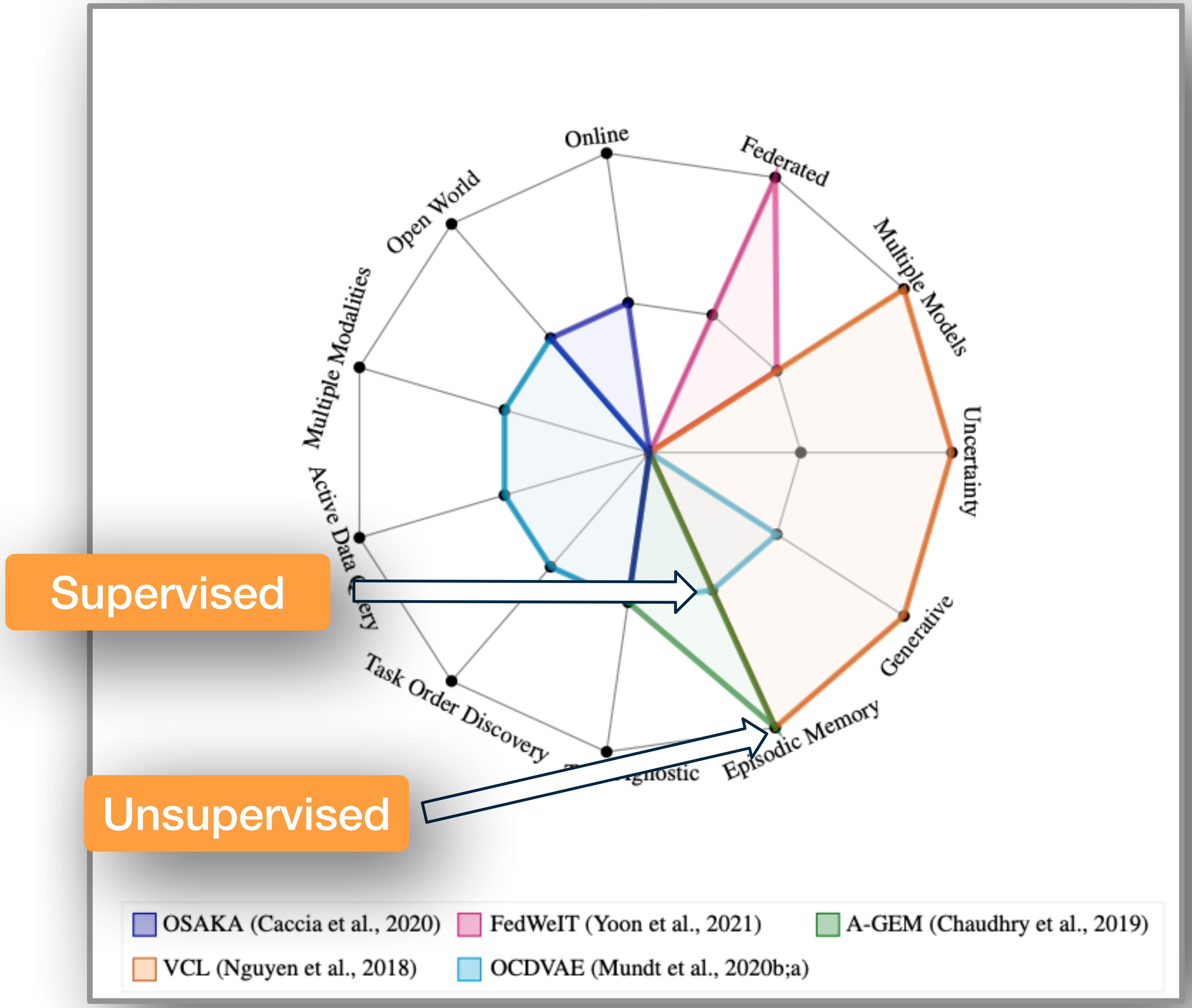


Mundt et al, “CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability”, ICLR 2022

CLEVA-Compass

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

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



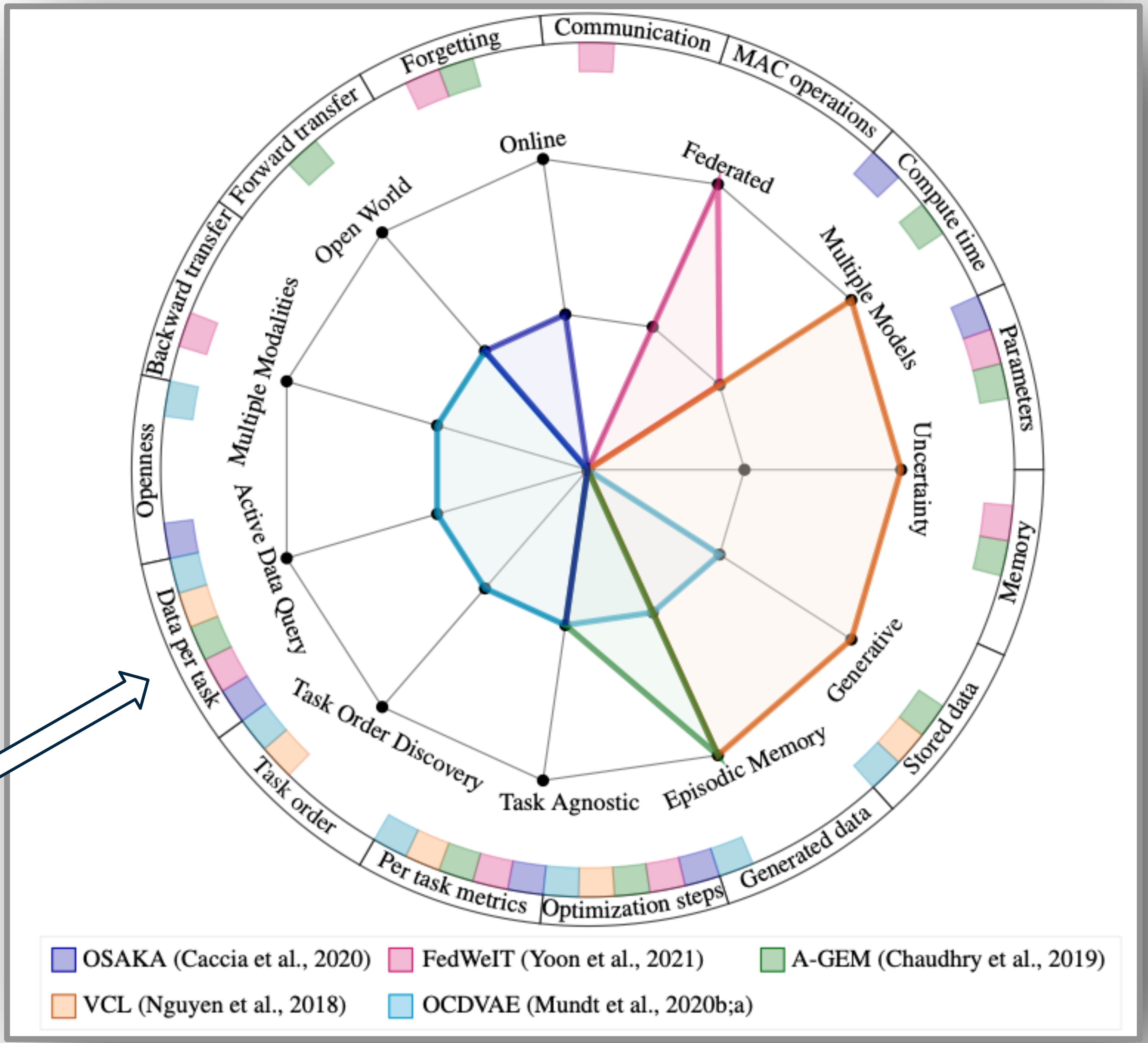
Mundt et al, “CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability”, ICLR 2022

CLEVA-Compass

Inner compass level (star plot):
indicates related paradigm inspiration & continual 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

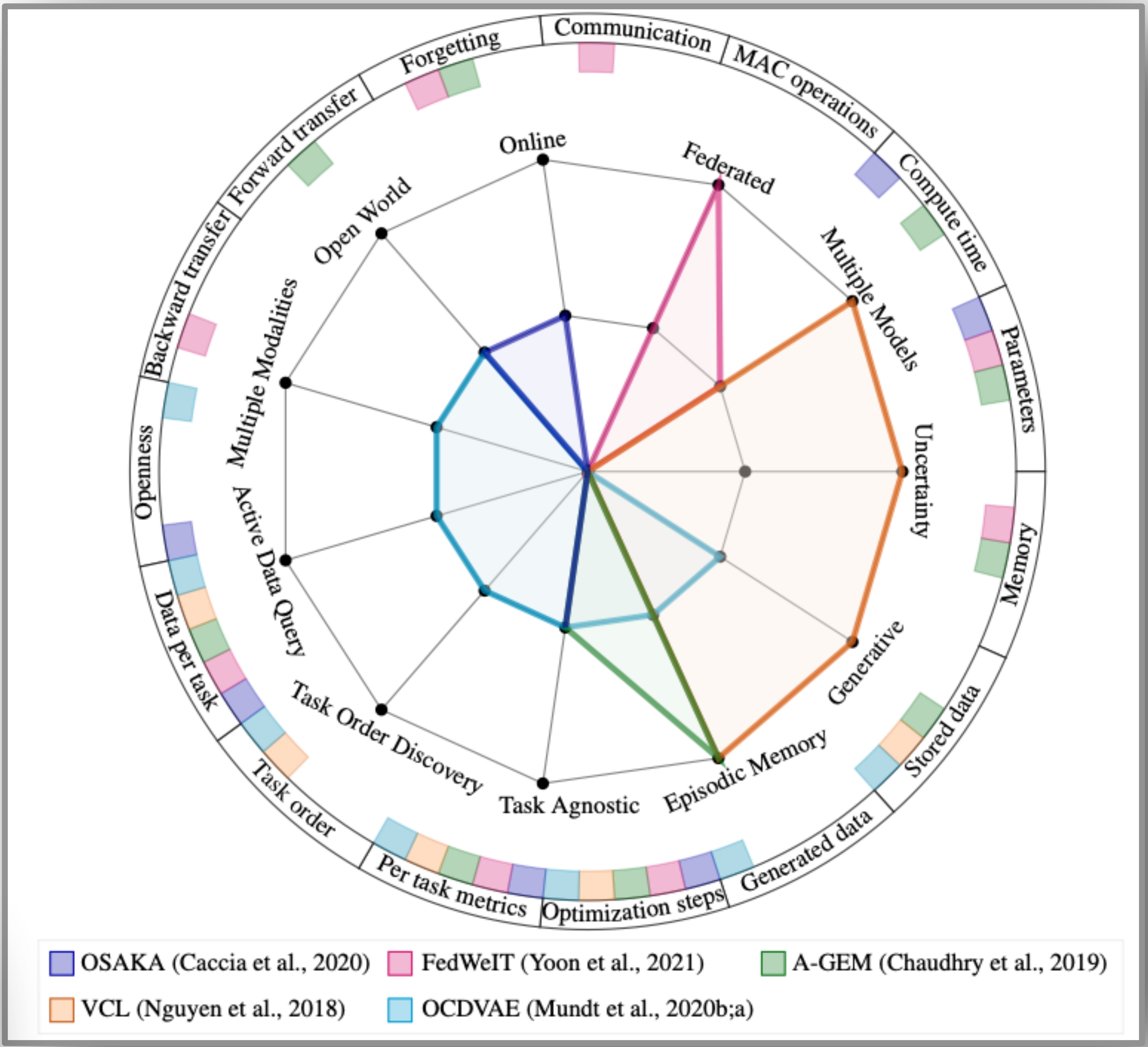
CLEVA-Compass

Inner compass level (star plot):
indicates related paradigm inspiration & continual 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

—> Encourages transparency, summarizes incentives, & promotes comparability in a compact visual form



We'll continue to talk about scenarios + assumptions next week, when we transition to the "open world"

Primarily: what if we don't know what to test on?