MICCAI 23 - DAICOW Tutorial

Pillars of Forgetting & Lifelong Evaluation

Martin Mundt

TU Darmstadt & hessian.AI - Independent Research Group Leader

ContinualAI - Board Member

http://owll-lab.com

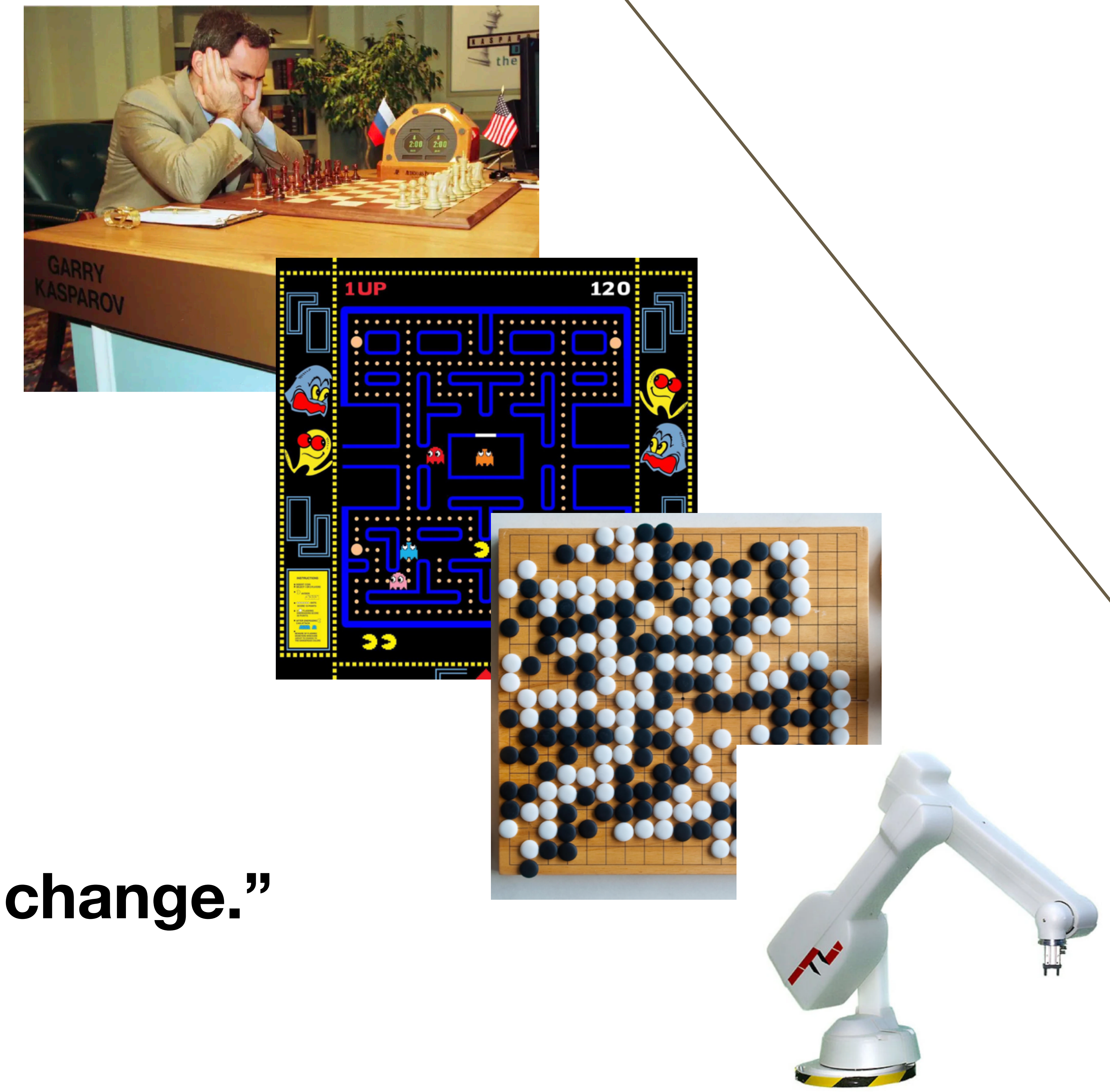
Parts of the upcoming tutorial are adapted from our previous "Continual Causality" tutorial at AAAI-23: Cooper & Mundt

Lifelong Learning: The Promise

If humans & animals learn continually, why shouldn’t our machines?

At the least, lifelong learning may be one pathway to more human-like intelligence

At the most, it's one pathway towards stronger artificial intelligence



“Intelligence is the ability to adapt to change.”
- Stephen Hawking

Lifelong Learning: The Practicalities

In the meantime, lifelong learning has direct benefits towards improving AI systems across research & real-world deployment

- Efficiency and Scalability
- Fairness, Privacy & Security
- Robustness and Accuracy

> **The New York Times**
>
> Processing all of that internet data requires a specialized supercomputer running for months on end, an undertaking that is enormously expensive. When asked if such a project ran into the millions of dollars, Sam Altman, OpenAI's chief executive, said the costs were actually "higher," running into the tens of millions.

Lifelong Learning: The Problem

Despite the achievements of many AI systems, few, if any, truly can learn continually over time:

- Narrow, fixed models, lacking robustness
- Incomplete and growing datasets
- Forgetting of prior knowledge

Connectionist models fail to learn sequentially

CATASTROPHIC INTERFERENCE IN CONNECTIONIST NETWORKS: THE SEQUENTIAL LEARNING PROBLEM

Michael McCloskey
Neal J. Cohen

I. Introduction

Connectionist networks in which information is stored in weights on connections between simple processing units have attracted considerable interest in cognitive science (e.g., Rumelhart, McClelland, & the PDP Research Group, 1986; McClelland, Rumelhart, & the PDP Research Group, 1986). Much of the interest centers around two characteristics of these networks. First, the weights on connections between units need not be prewired by the model builder but rather may be established through training in which items to be learned are presented repeatedly to the network and the connection weights are adjusted in small increments according to a learning algorithm (e.g., Ackley, Hinton, & Sejnowski, 1985; Rumelhart, Hinton, & Williams, 1986; Hinton & Sejnowski, 1986). Second, the networks may represent information in a distributed fashion. That is, the representation of an item may be spread, or distributed, across many different processing units and connections, and each unit and connection may be involved in representing many different items.

Distributed representations established through the application of learning algorithms have several properties that are claimed to be desirable from the standpoint of modeling human cognition (e.g., Hinton, McClelland, & Rumelhart, 1986; McClelland, Rumelhart, & Hinton, 1986; but see Prince & Pinker, 1988; Fodor & Pylyshyn, 1988; Lachter &

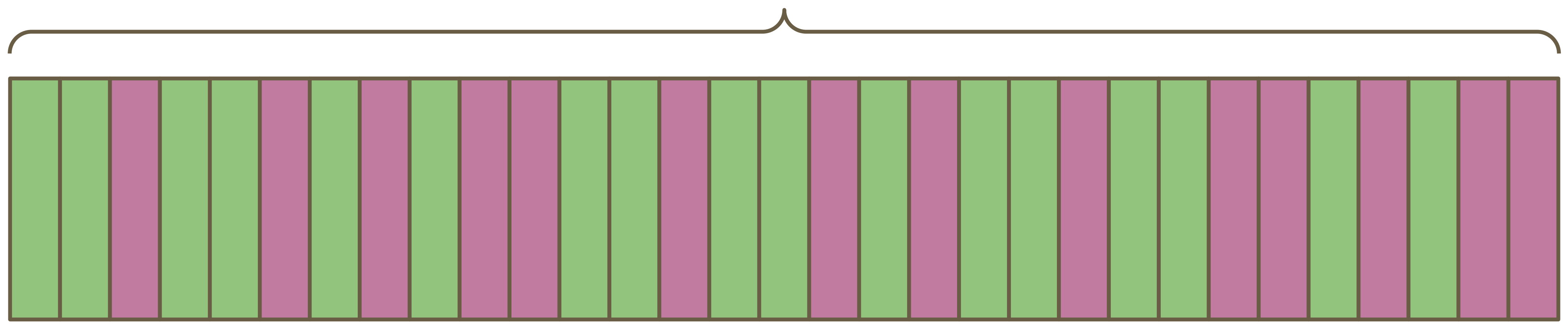
THE PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL. 24

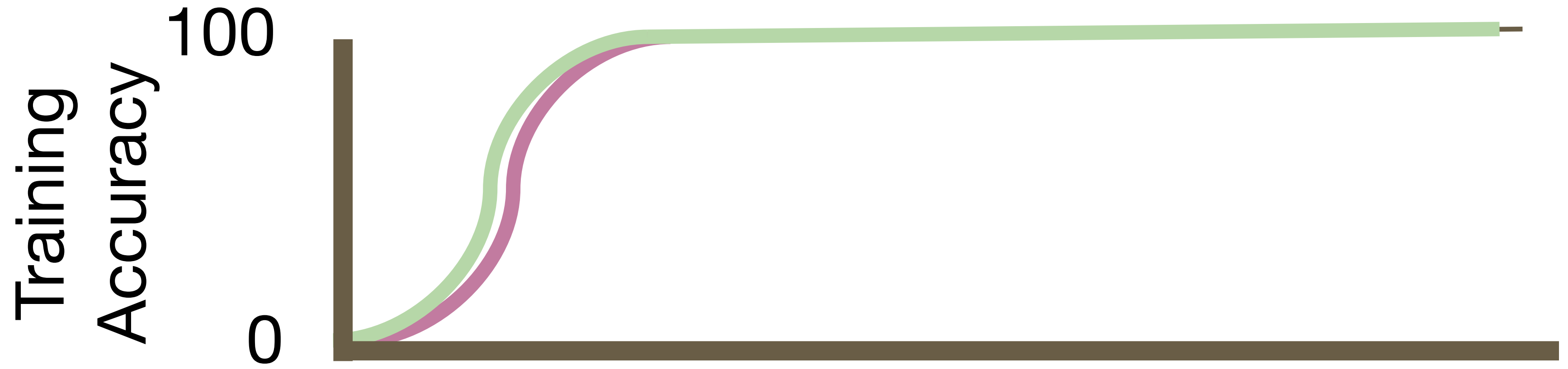
109

The sequential learning problem

Training Order

Interleaved training



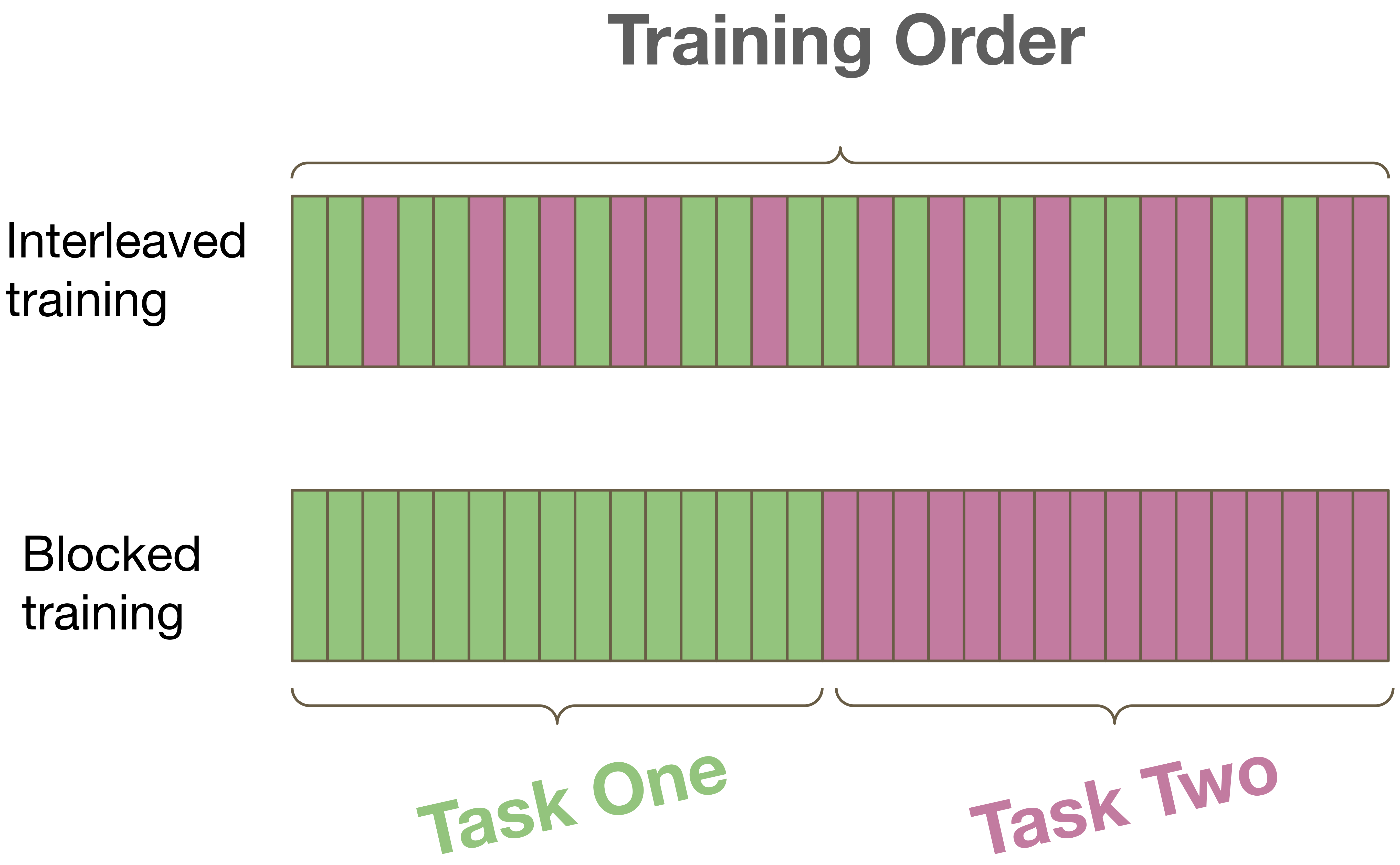


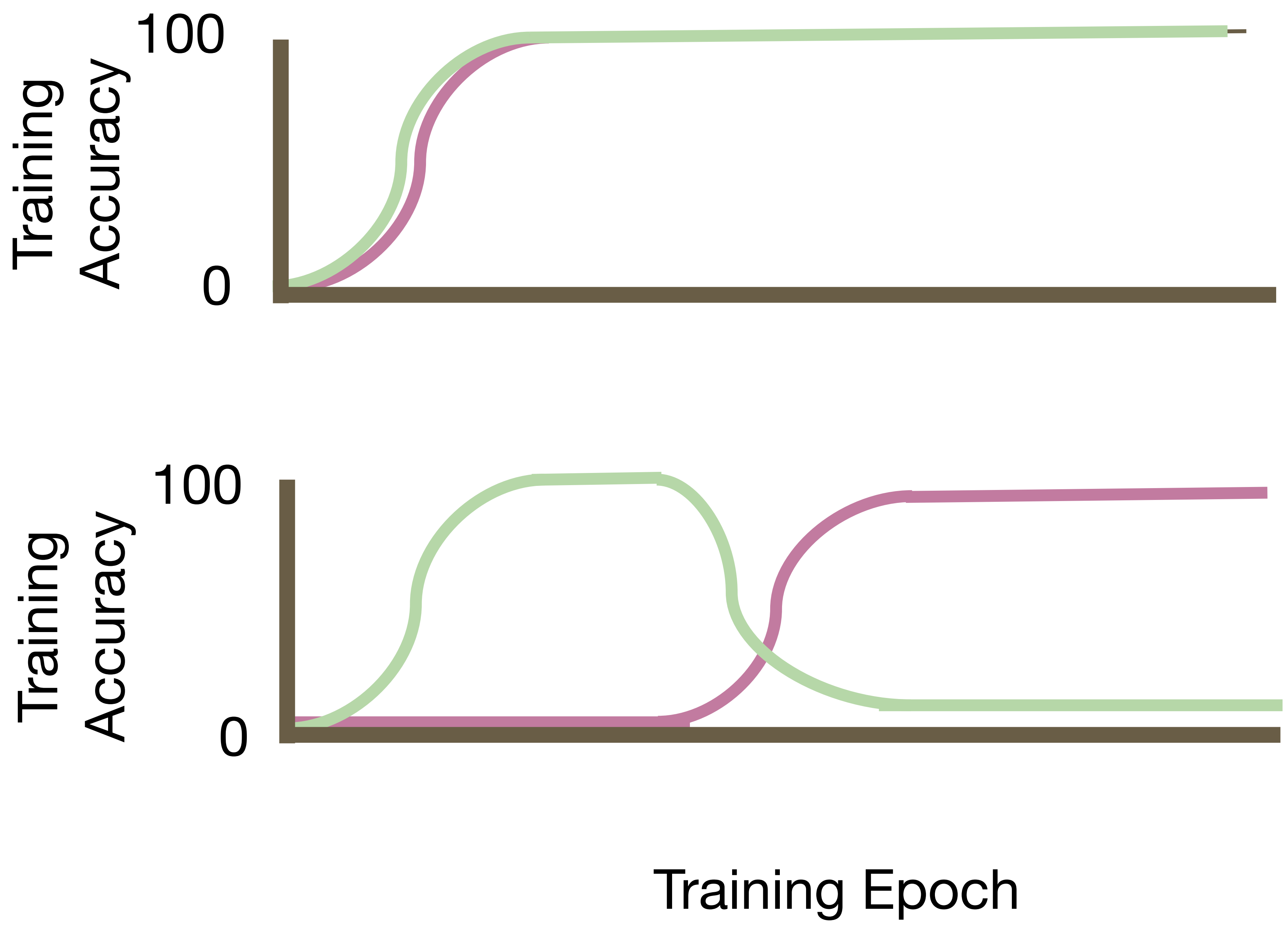
Task One

Task Two

Adapted from Flesch et al, 2022

The sequential learning problem





Adapted from Flesch et al, 2022

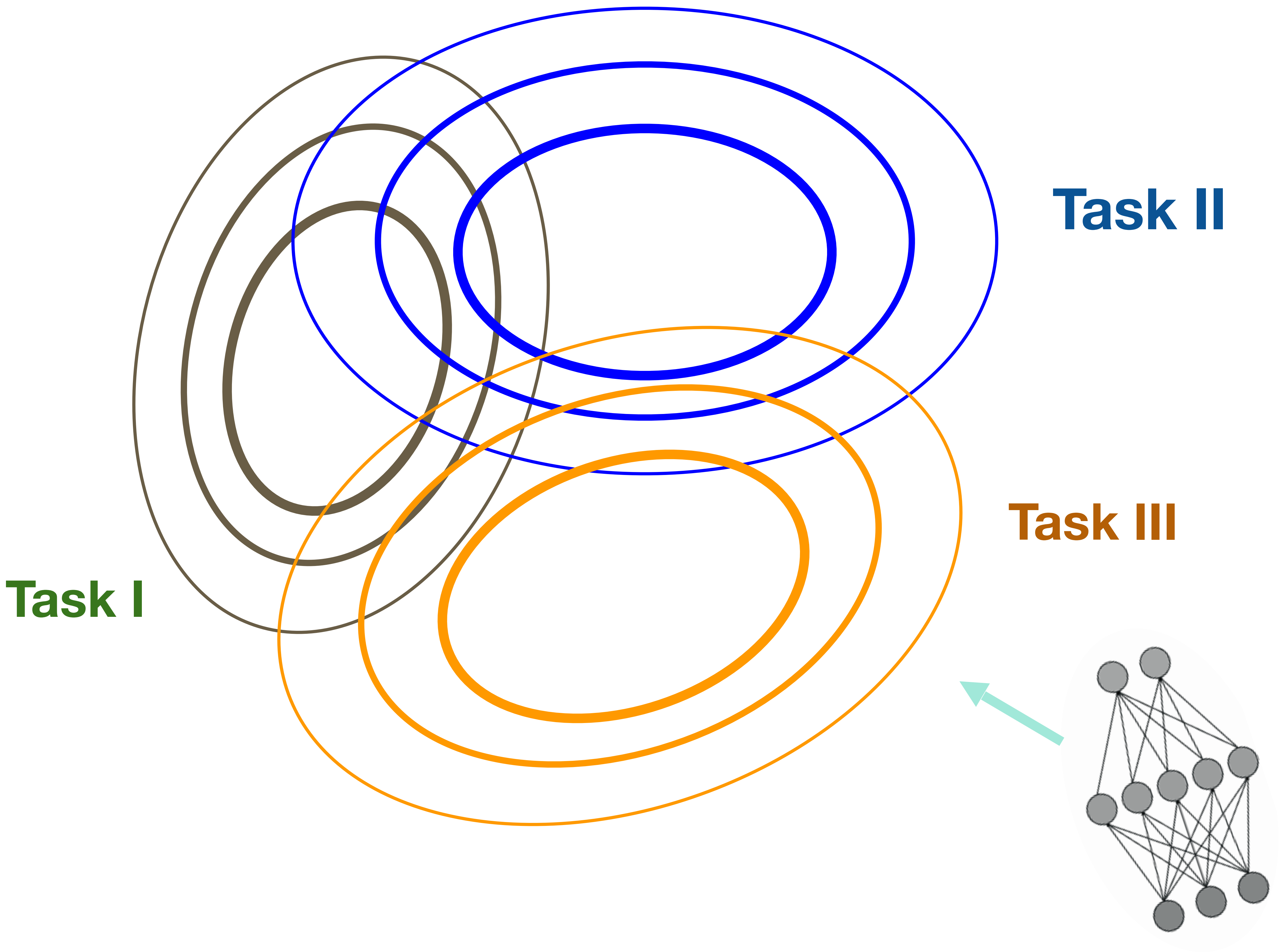
Is the forgetfulness of our ML models surprising?

Say we teach a network 3 tasks

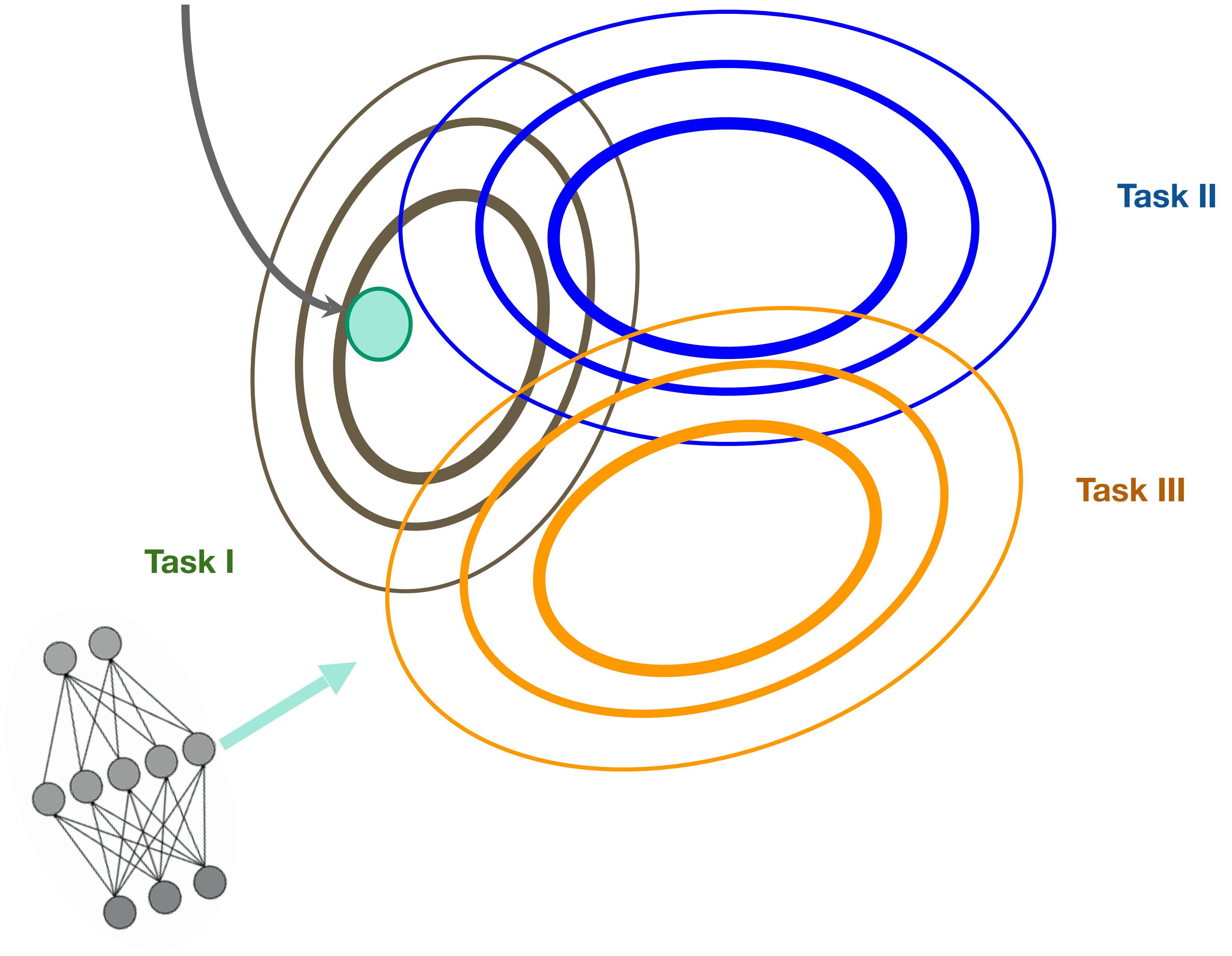
Train on each task sequentially, with no direct overlap of task examples

Think of the network's weights as occupying a landscape of configurations to solve a given task

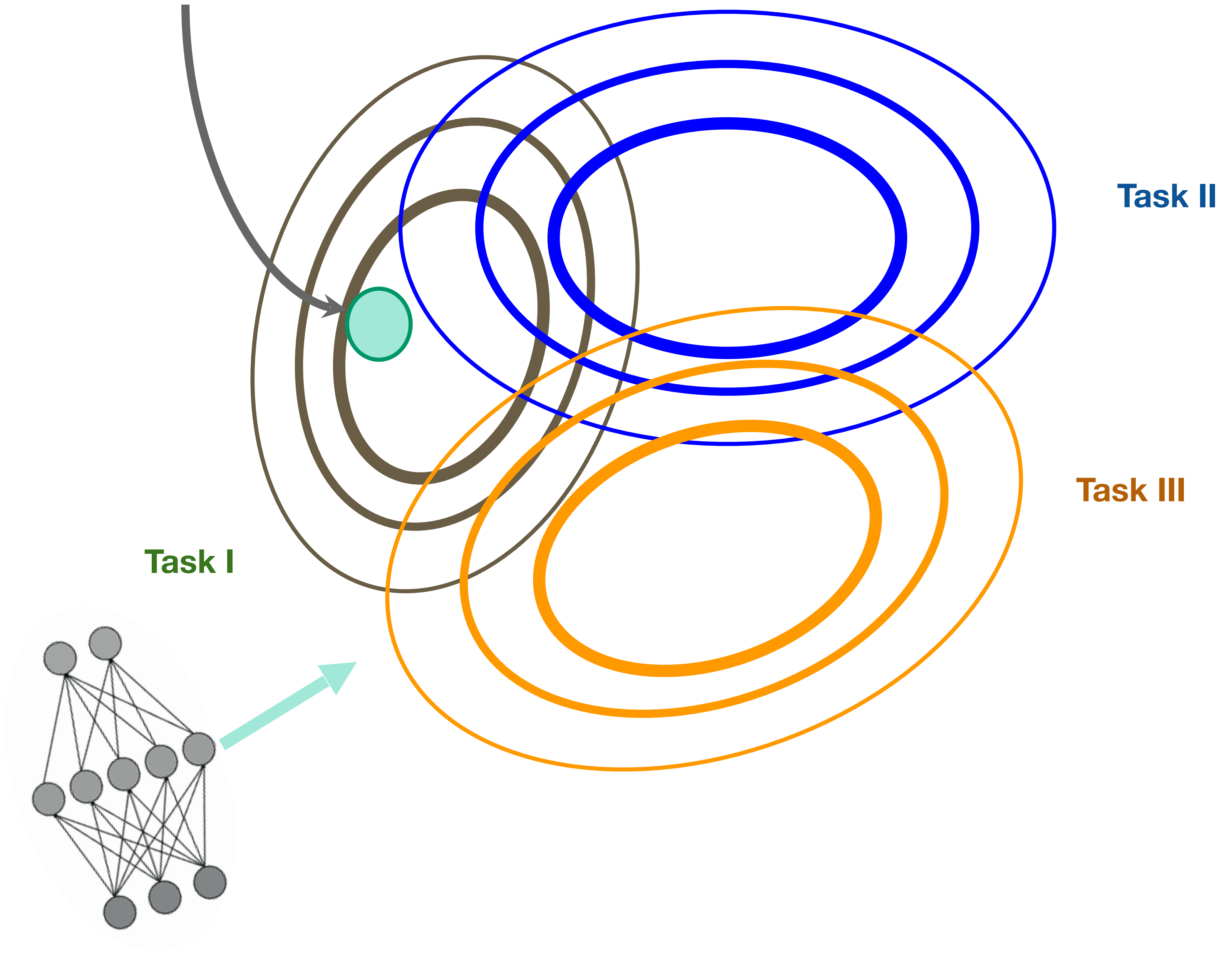
The center of each distribution on the right solves that task

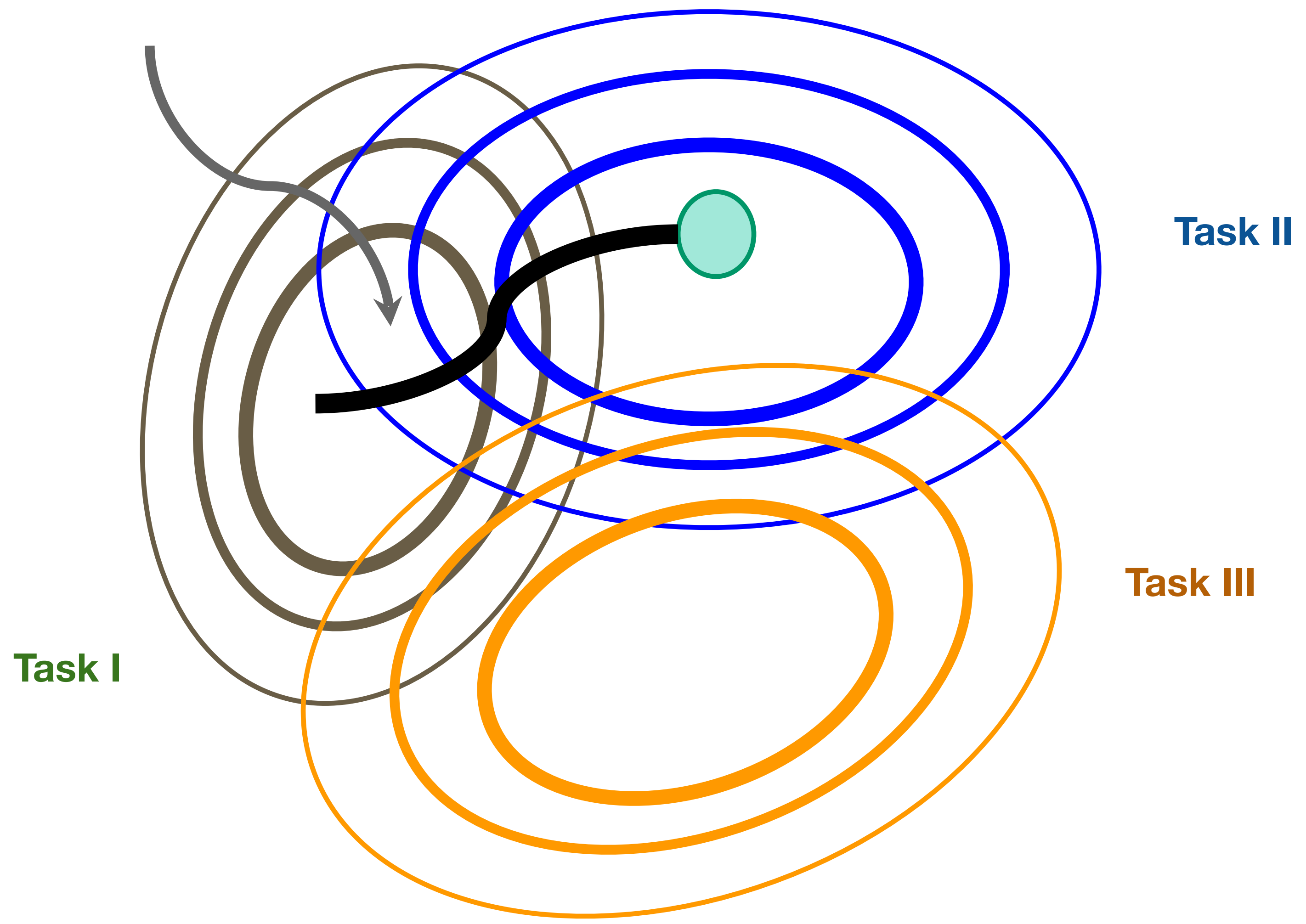


Is the forgetfulness of our ML models surprising?

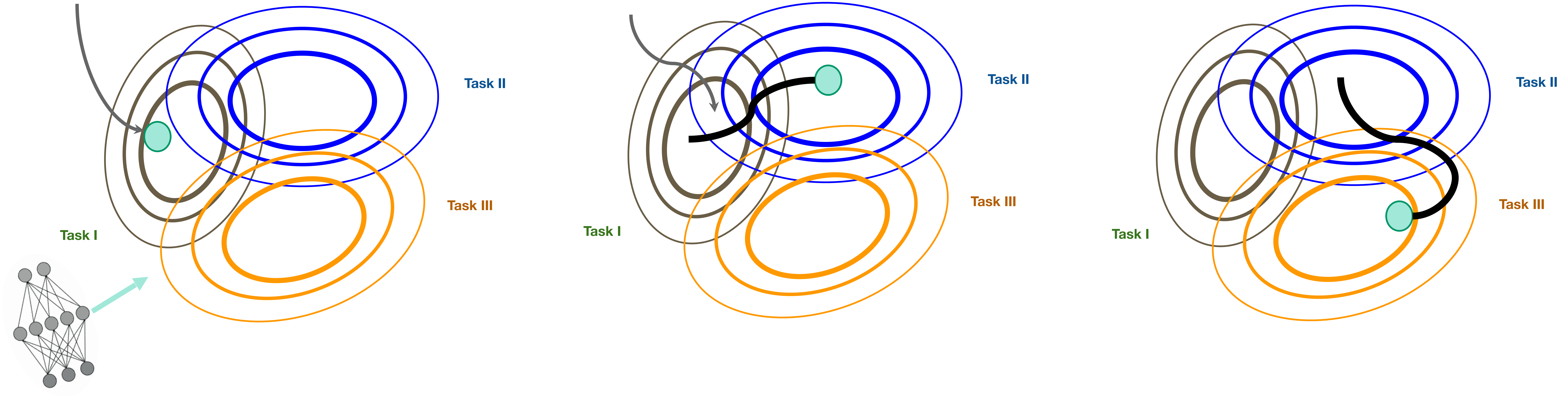


Is the forgetfulness of our ML models surprising?





Is the forgetfulness of our ML models surprising?

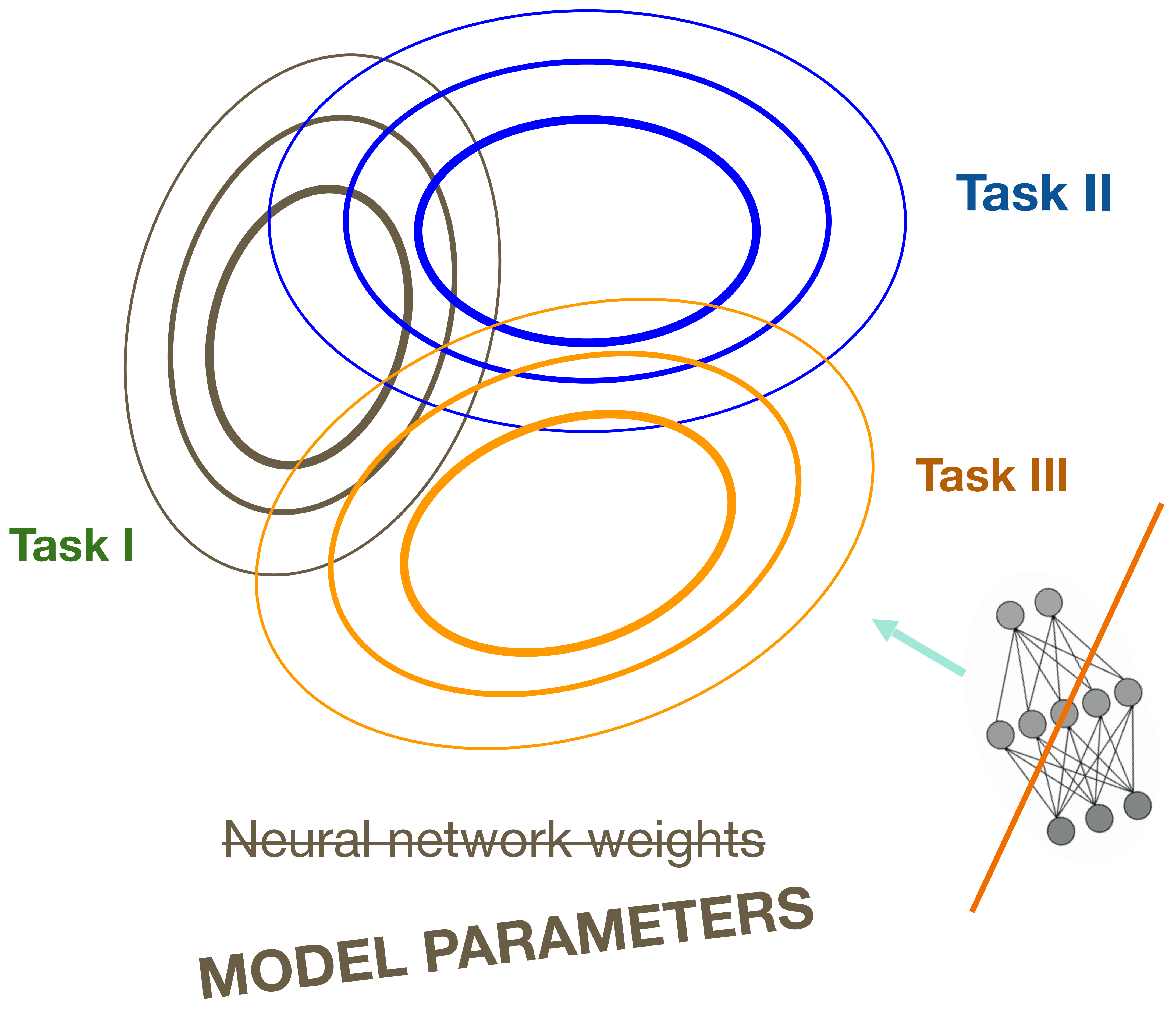


… not really

Not just neural networks

While most commonly associated with deep learning, catastrophic interference applies to a much broader class of algorithms

- Neural networks (McCloskey & Cohen 1989)
- Linear regression (Everon et al., 2022)
- SVM (Ayad 2014)
- Self organizing maps (Richardson & Thomas 2018)
- And more…

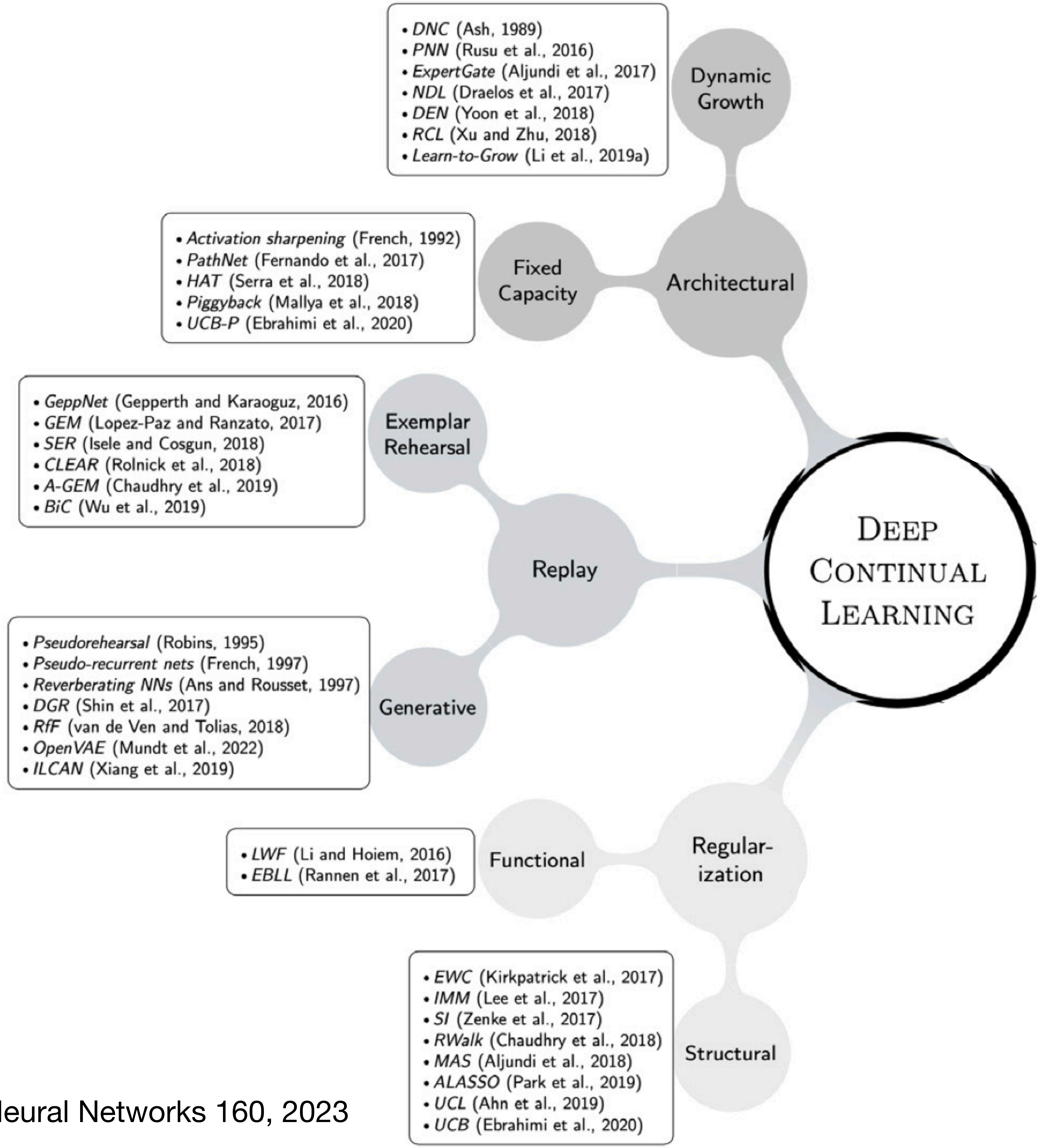


Overview of Strategies

Regularization: Alter the weight dynamics as a function of tasks

Replay: Leverage past samples of previous task data

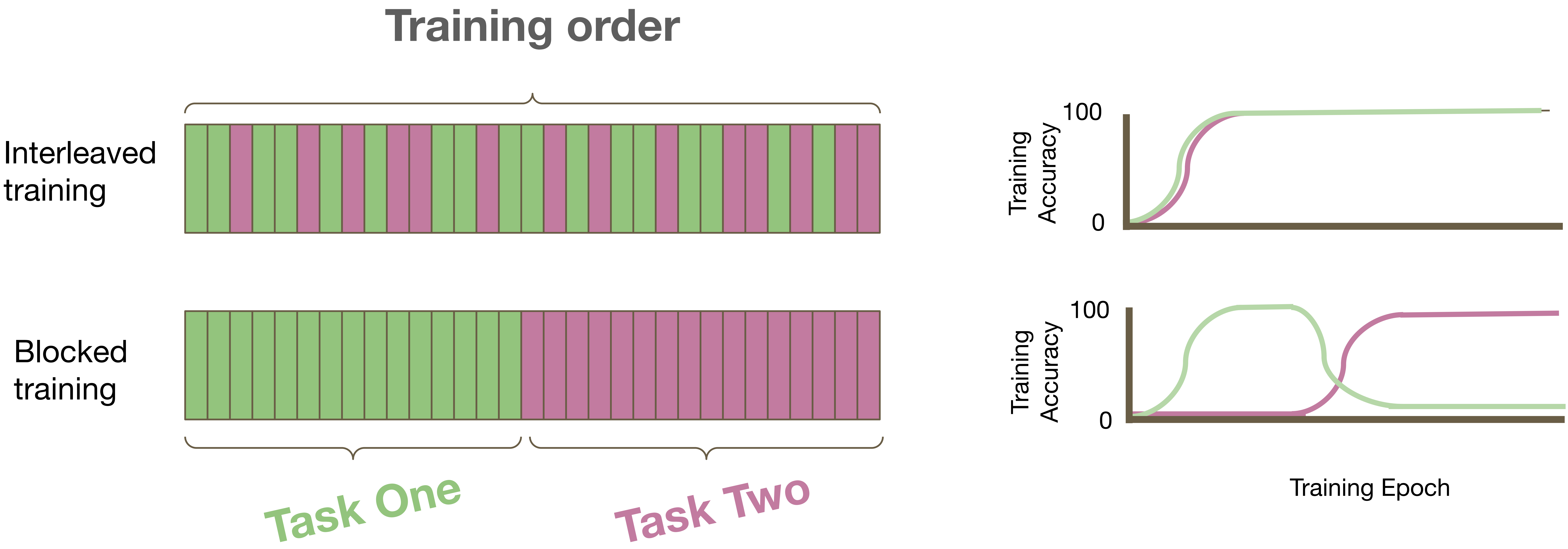
Architectural: Change the macro or micro architecture of the network



Mundt et al, “Wholistic View of Continual Learning with Deep Neural Networks”, Neural Networks 160, 2023

Pillar 1: “Replay” to alleviate forgetting

If interleaving samples rescues forgetting…



Adapted from Flesch et al, 2022

…then storing samples for later may be useful

Saves samples of each tasks' data in (external) memory buffer

Progressively replace parts of memory buffer with new examples

Disadvantages:

- Utilizes separate memory
- Violates data privacy
  (a key motivation for lifelong learning)

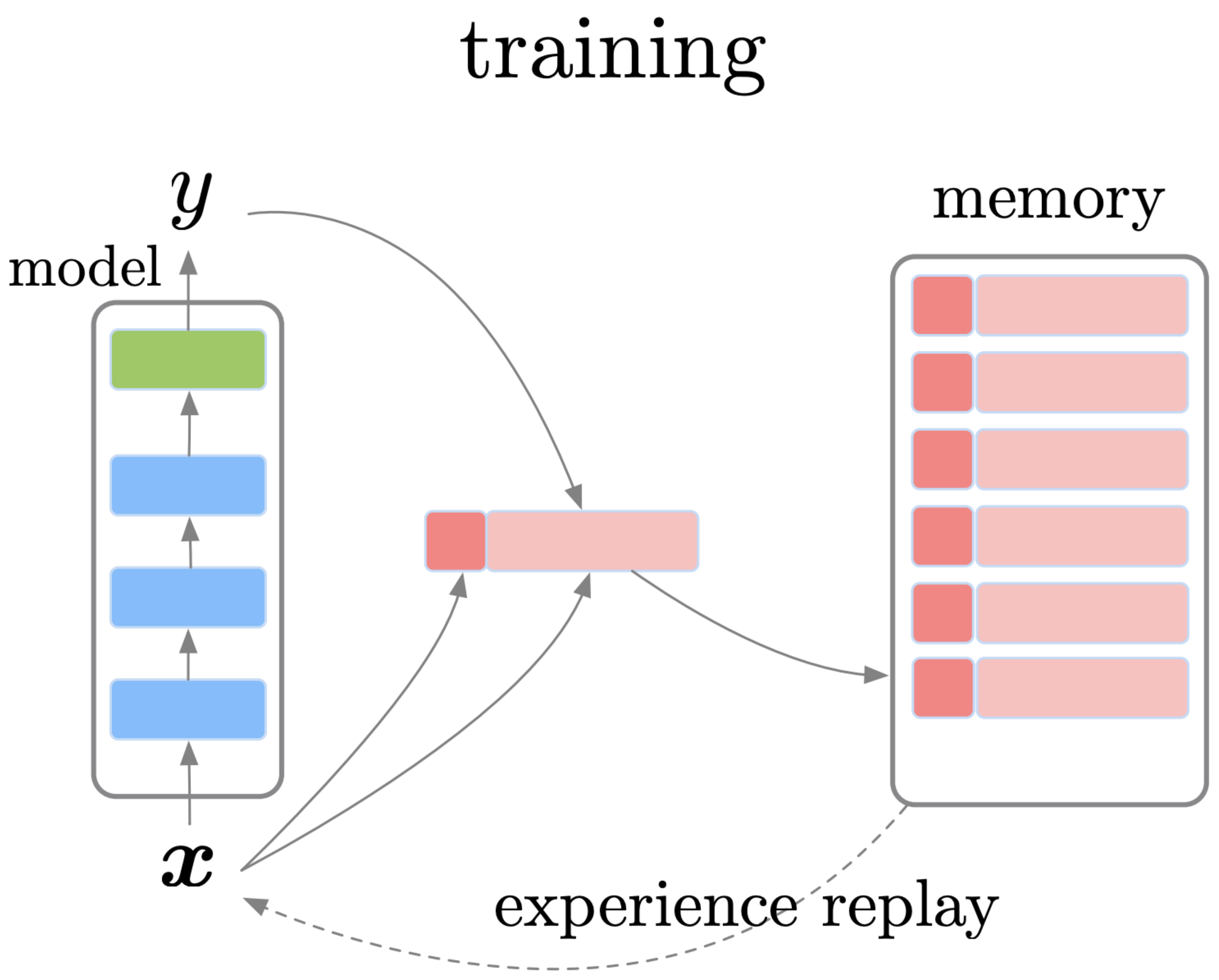


Figure from Masson d'Autume et al. 2019, ArXiv

A caveat…

“While it is an effective method in ANNs, rehearsal is unlikely to be a realistic model of biological learning mechanisms, as in this context the actual old information (accurate and complete representation of all items ever learned by the organism) is not available.

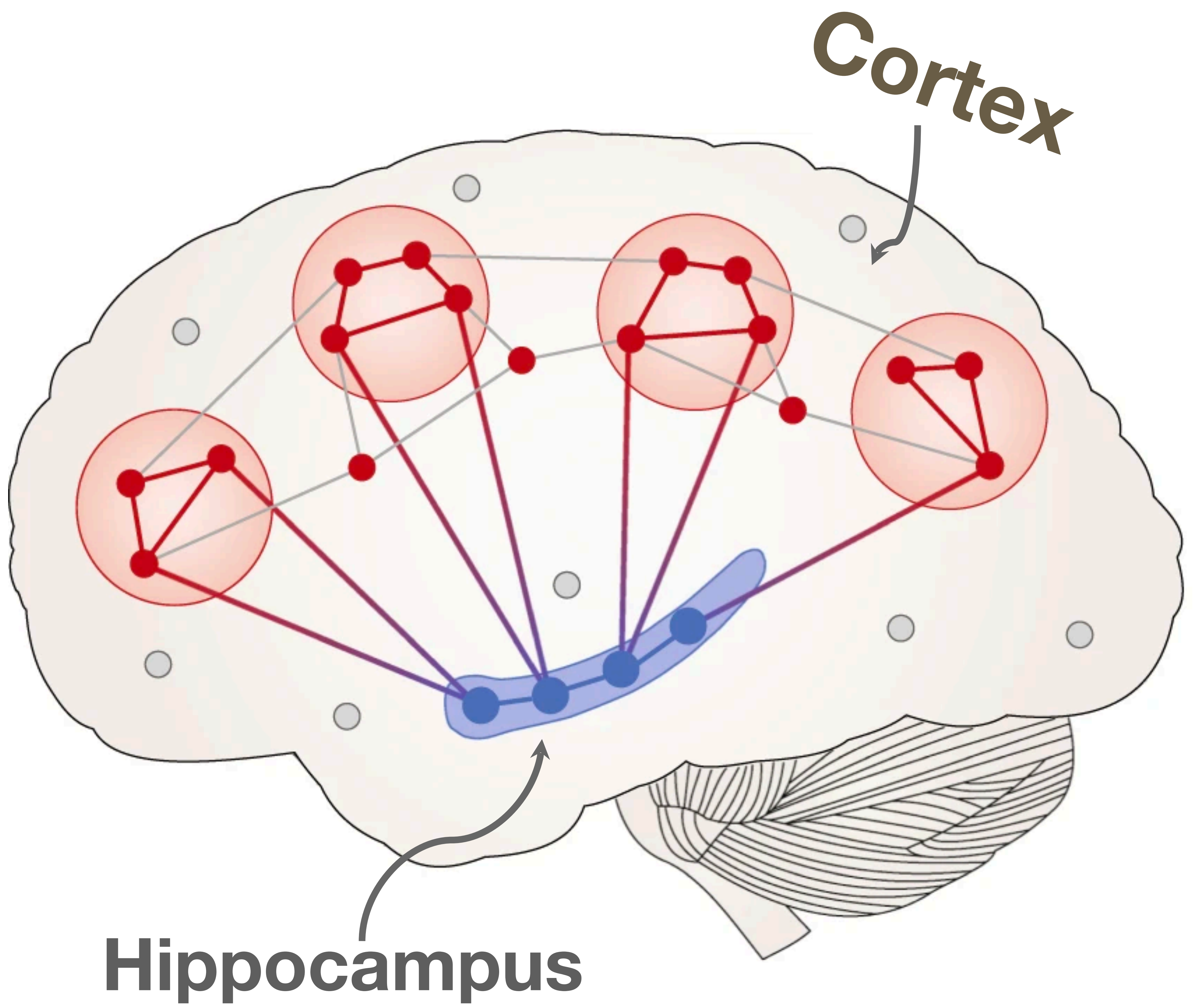
– Robert French, 1997

Replay IS biologically plausible

Complementary learning systems theory
(McClelland et al., 1995; Marr et al., 1971)

- Hippocampus is a fast learning system
- Cortex is a slow learning system
- Hippocampus replays memories to cortex
- Cortex generalizes memories
- Hippocampus becomes less necessary for recall

The “central dogma” of memory consolidation



Hayes et al., 2021 Neural Computation; Figure adapted from Klinzing et al., 2019

A caveat… solved?

"While it is an effective method in ANNs, rehearsal is unlikely to be a realistic model of biological learning mechanisms, as in this context the actual old information (accurate and complete representation of all items ever learned by the organism) is not available. Pseudo-rehearsal is significantly more likely to be a mechanism which could actually be employed by organisms as it does not require access to this old information, it just requires a way of approximating it."

– Robert French, 1997

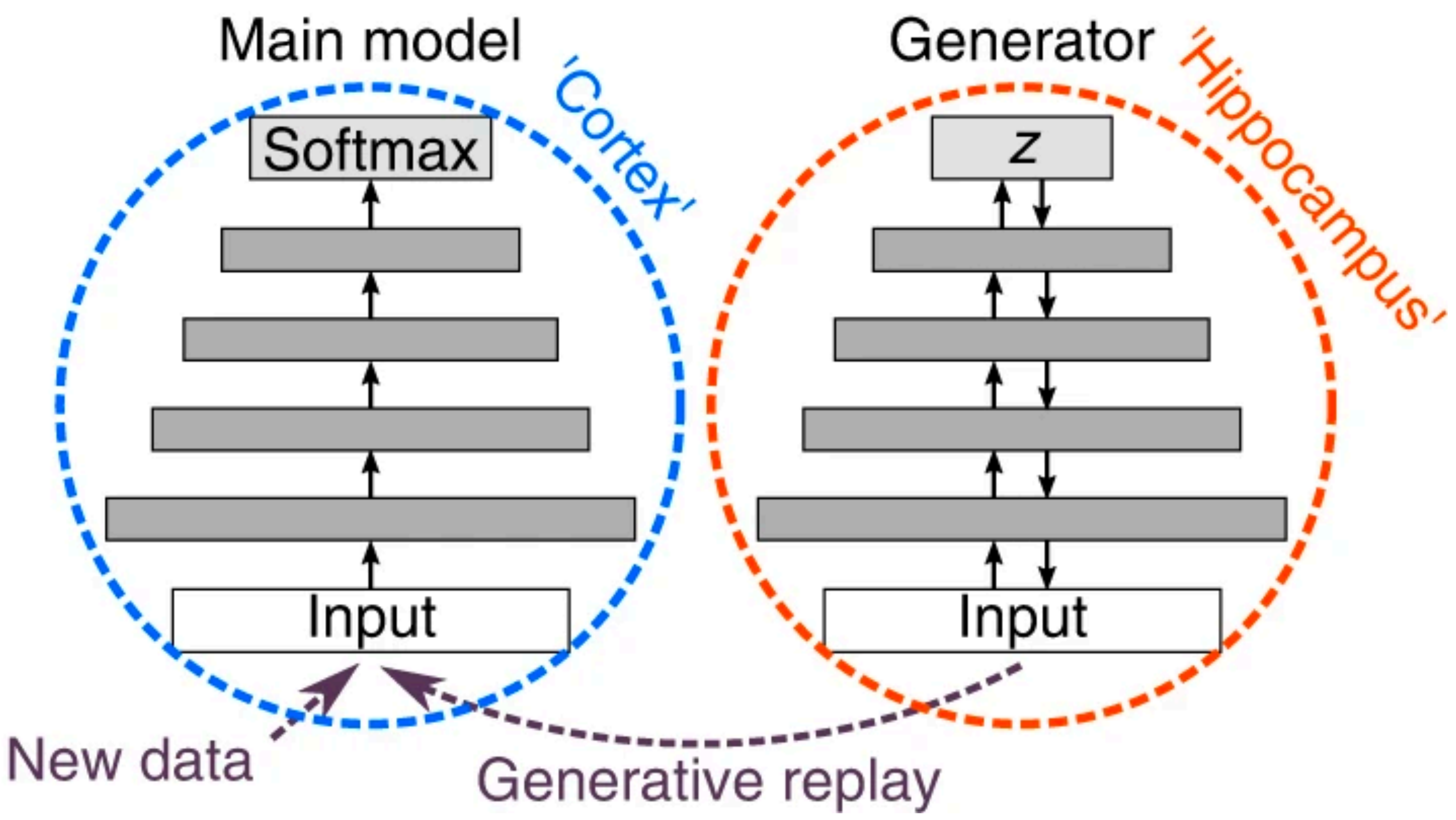
Pseudo-Replay is biologically plausible

Generative replay:

- Don't memorize samples directly
- Instead, memorize their exemplars
- Replay generated samples instead

Increasing evidence biological replay is not a simple function of experience:

- Replay is weighted by novelty
- Replay samples all routes in an environment



Lesort et al., 2019; Figure adapted from van de Ven et al., 2020; Klinzing et al., 2019

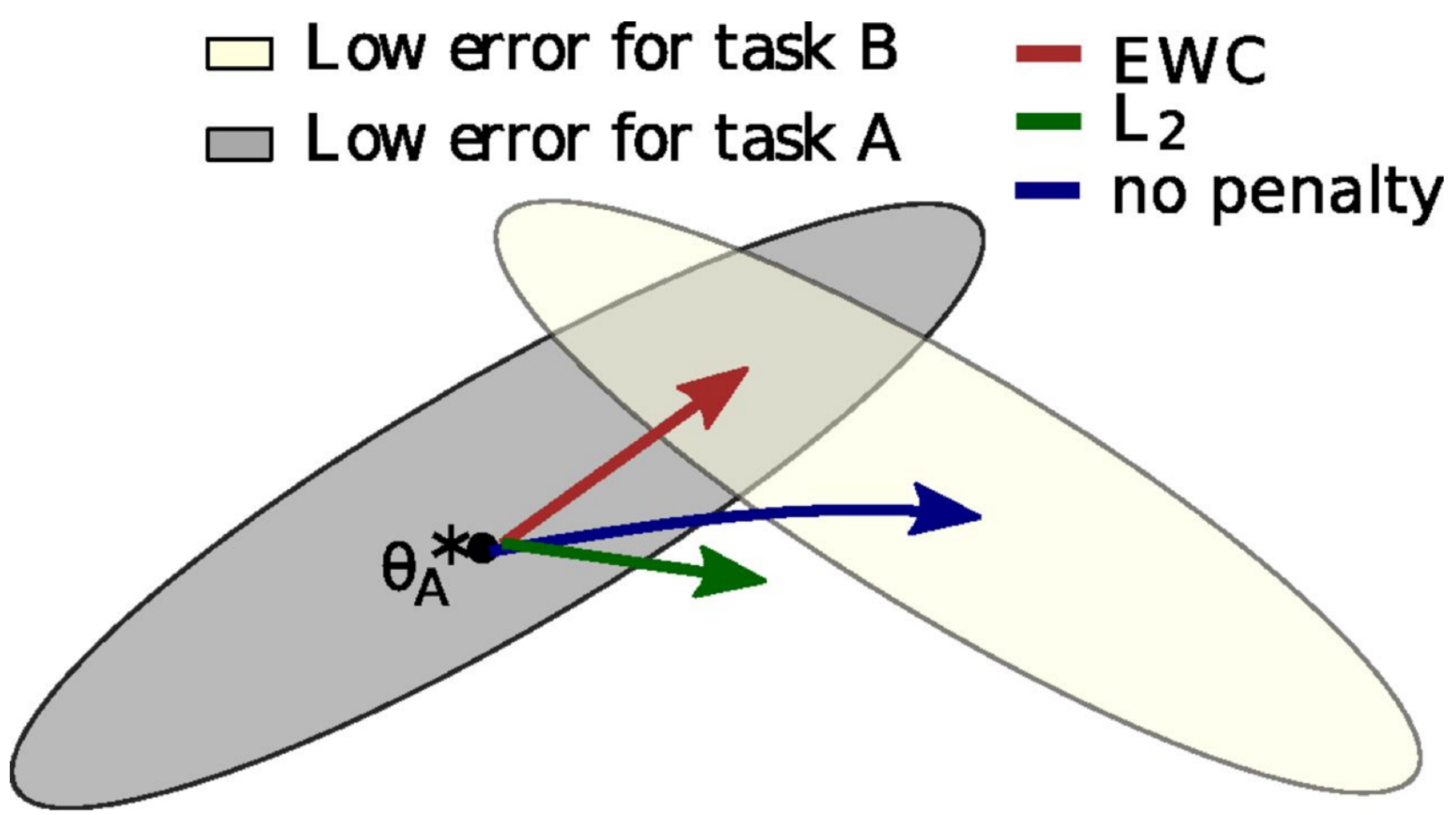
Pillar 2: “Regularization” to alleviate forgetting

Elastic Weight Consolidation

Don't greedily optimize for a new task, preserve the old weights by penalizing updates to already learned parameters

Step-1: Approximate Fisher Information *(parameter importance)*

Step-2: Apply a squared regularization loss to penalize shift in important weights from the previous task



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

Regularization IS biologically plausible & complementary to replay

“Instead of viewing cellular and systems consolidation as separate entities, we need to focus more on their interactive dynamics. …After more than a century of research, one thing has become abundantly clear: consolidation is not a simple process.”

– Lisa Genzel and John Wixted, 2017

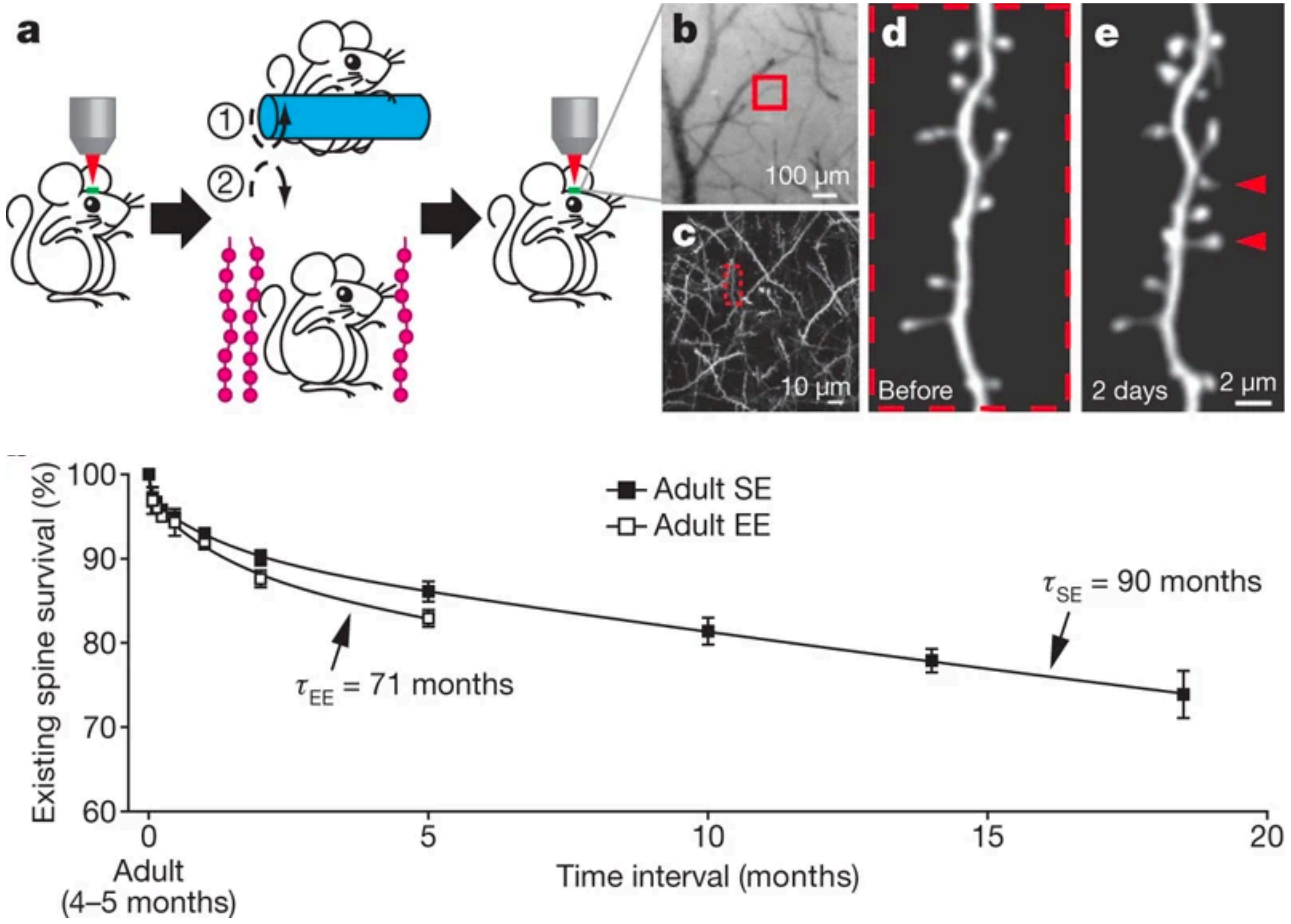


Figure adapted from Cichon & Gan 2015

Pillar 3: “Structure” to alleviate forgetting

Why dynamic architectures?

"Catastrophic forgetting is a direct consequence of the overlap of distributed representations and can be reduced by reducing this overlap."

Robert French, "Using Semi-Distributed Representations to Overcome Catastrophic Forgetting in Connectionist Networks", AAAI 1993

Why dynamic architectures?

"Catastrophic forgetting is a direct consequence of the overlap of distributed representations and can be reduced by reducing this overlap."

Robert French, "Using Semi-Distributed Representations to Overcome Catastrophic Forgetting in Connectionist Networks", AAAI 1993

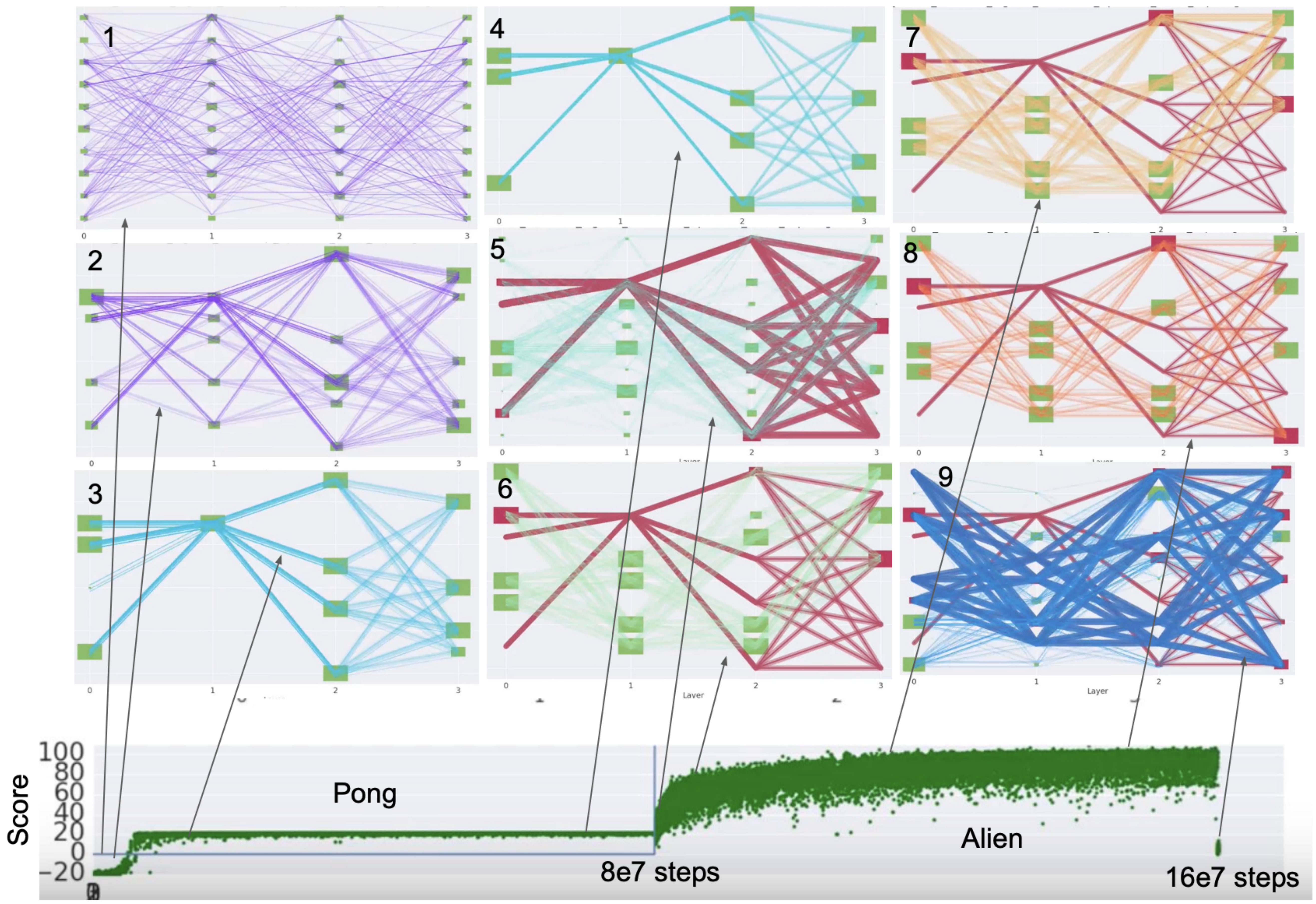
"Very local representations will not exhibit catastrophic forgetting because there is little interaction among representations. However, a look-up table lacks the all-important ability to generalize. The moral of the story is that you can't have it both ways."

The practical implicit perspective

- Start over-parametrized
- Constrain a task to use a subset of parameters, create “sub-models”

Example: *Pathways/PathNets*

Enforce a small/fixed number of active modules/“paths”



Fernando et al, “PathNet: Evolution Channels Gradient Descent in Super Neural Networks”, arXiv:1701.08734, 2017

Dynamic structure is biologically plausible

Inspiration from *neurogenesis*?

"After two decades of research, the neurosciences have come a long way from accepting that neural stem/progenitor cells generate new neurons in the adult mammalian hippocampus to unraveling the functional role of adult-born neurons in cognition and emotional control. The finding that new neurons are born and become integrated into a mature circuitry throughout life has challenged and subsequently reshaped our understanding of neural plasticity in the adult mammalian brain.*"*

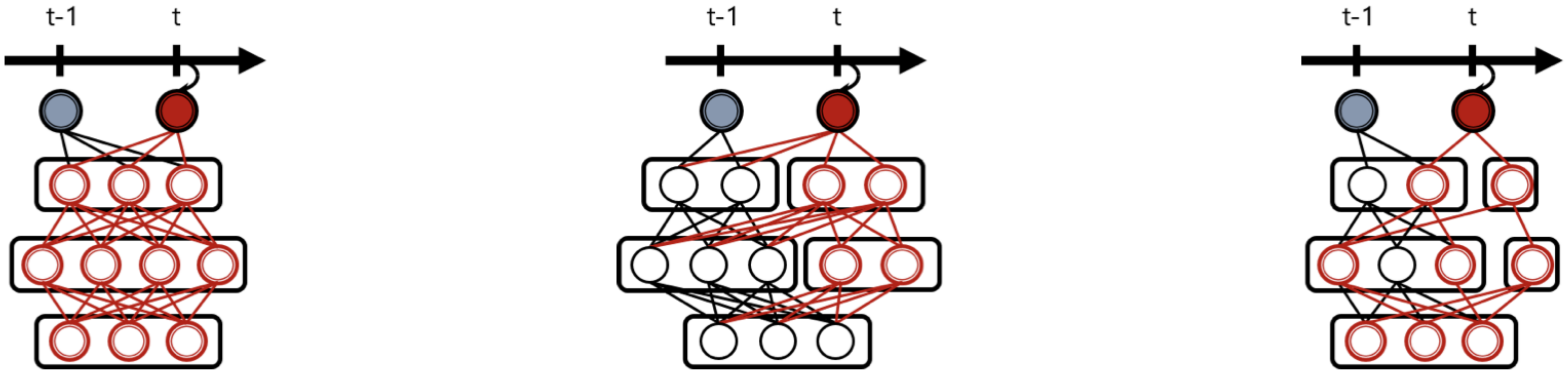
(Quote: Vadodaria & Jessberger, "Functional neurogenesis in the adult hippocampus: then and now", frontiers in neuroscience 8, 2014, see also C. Gross, "Neurogenesis in the adult brain: death of a dogma", Nature Reviews Neuroscience, 2000)

The practical explicit perspective

Various combinations with partial re-training with expansion - three questions:

1. *When* should we add?   2. What/how do we add?   3. When do we stop?

EWC   Progressive Nets   DEN



(a) Retraining w/o expansion   (b) No-retraining w/ expansion   (c) Partial retraining w/ expansion

Yoon et al, "Lifelong Learning with Dynamically Expandable Networks", ICLR 2018

So how do we evaluate lifelong learning?

First good idea: per task measures

- "**Base**" loss: the initial 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: forward and backward transfer

(Avg.) ***Forward transfer*** (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.

Third 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]:

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

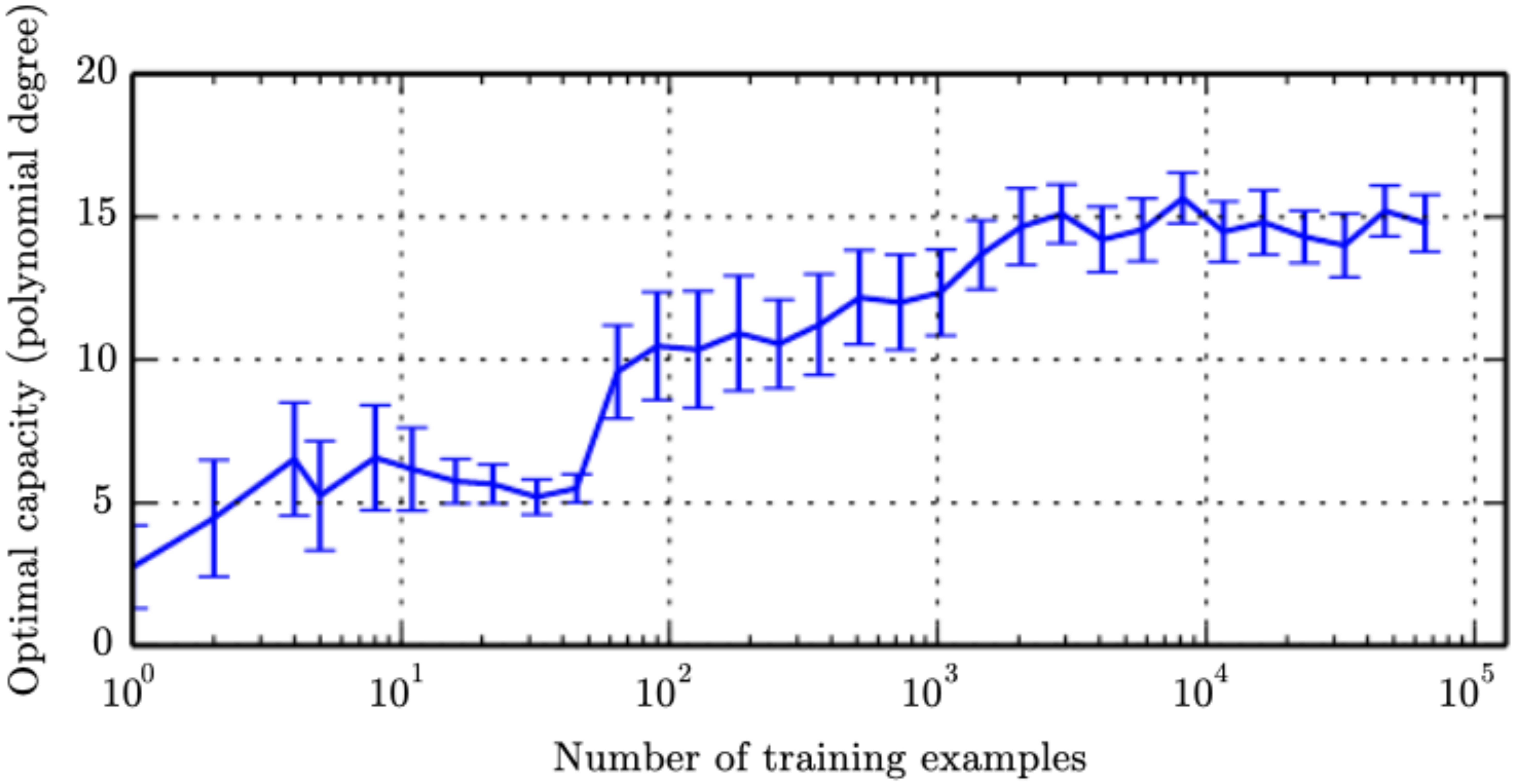
Beta = 0 is zero-shot performance == Forward transfer

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

Are we done? Can we solve lifelong learning?
No, forgetting is only one of many challenges

Challenge 1: evaluation axes are intertwined

Unfortunately it's not only about catastrophic forgetting, it's also about *capacity…*



Deep Learning, Goodfellow, Bengio, Courville, MIT Press 2016, Machine Learning Basics chapter, page 114.

…and memory, size, compute & plenty of other things

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

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

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

Computational Efficiency

Quantifies add/multiply ops (inference & updates)

Model Size Efficiency

Quantifies parameter growth

Sample Storage Size Efficiency

Quantifies stored amount of data (for rehearsal)

(Díaz-Rodríguez & Lomonaco et al, "Don't forget, there is more than forgetting: new metrics for Continual Learning", 2018)

Challenge 2: what matters in 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	✗	✗	$\mathcal{T} \cdot U$

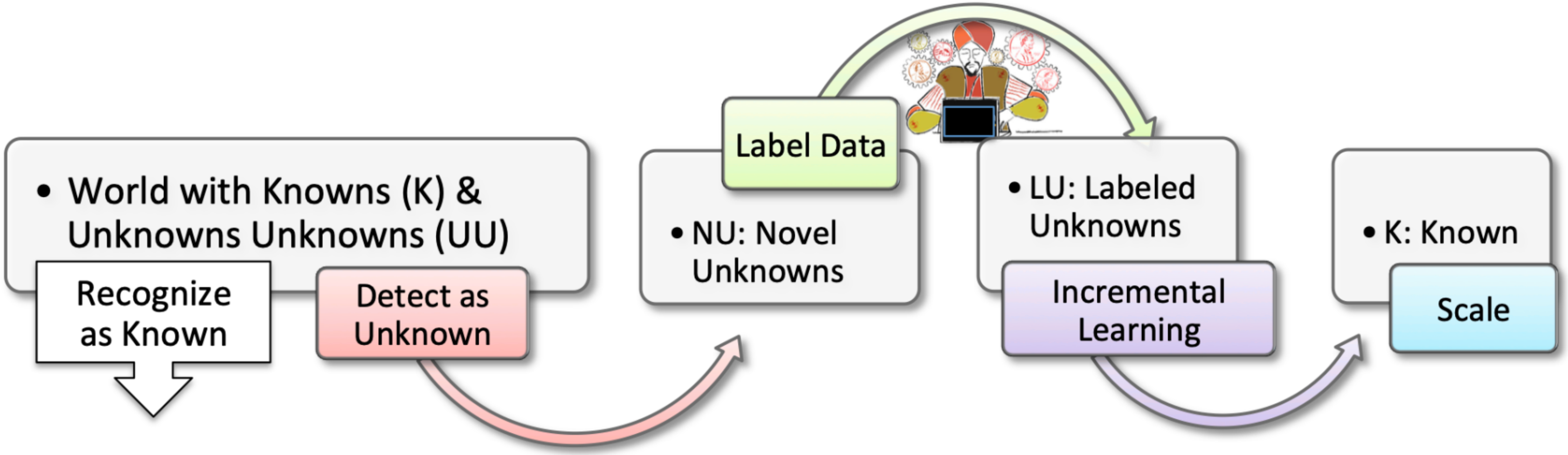
Low

High

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

Challenge 3: the world is “open” & not a benchmark

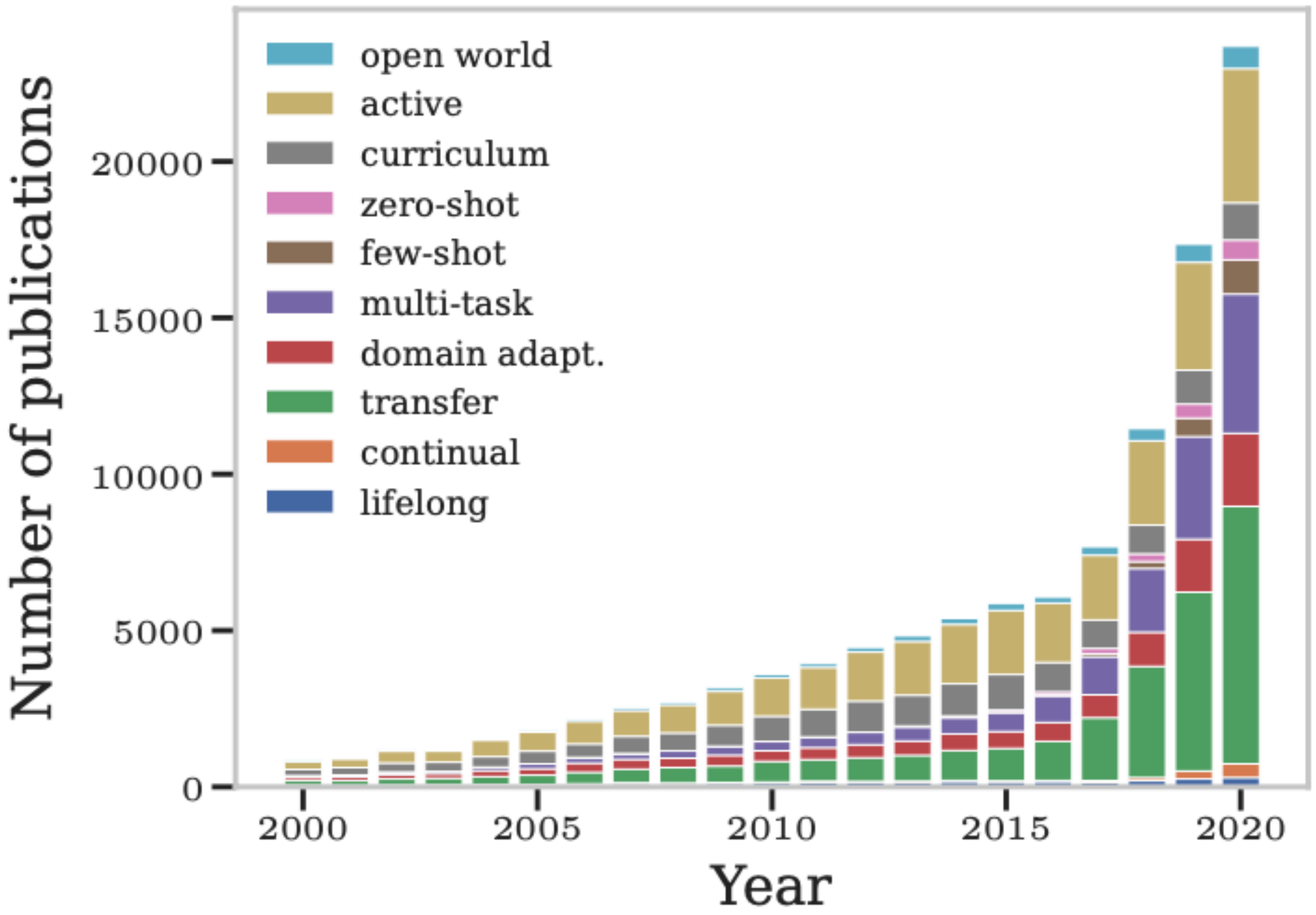
The challenge of consensus. Is it possible to postulate general *desiderata*?



Bendale & Boult, “Towards Open World Recognition”, CVPR 2015. Also see Mundt et al “A Wholistic View of Continual Learning with Deep Neural Networks: Forgotten Lessons and the Bridge to Active and Open World Learning”, Neural Networks 160, 2023

It’s unclear if there is a single set of desiderata…
but can we at least compare fairly?

Many more challenges - the way forward?



Where do we go from here?

Why are there so many possible *assumptions* and ways to *measure*?!

Let's wrap up by reminding ourselves about their *origin*!

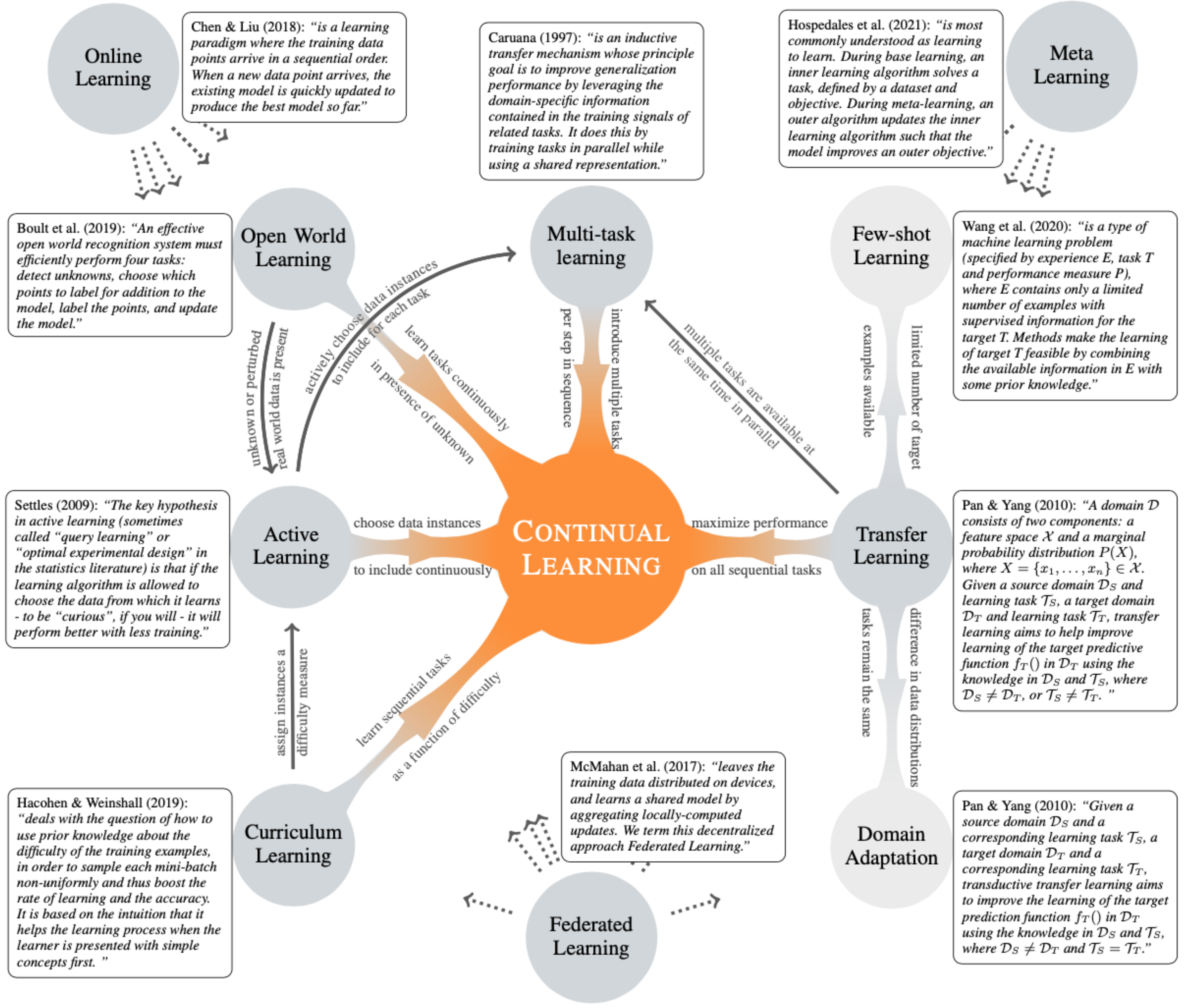
Mundt et al, "CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability", ICLR 2022

Evaluation & related paradigms

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

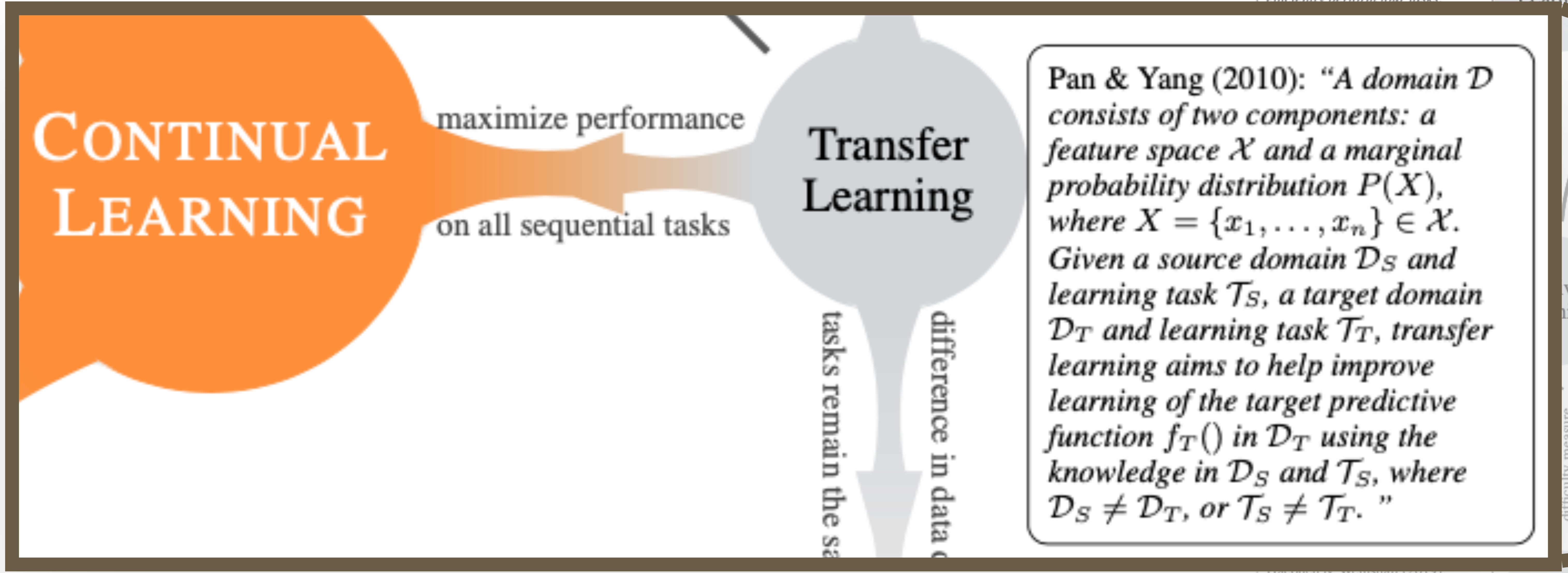
Key aspects often reside in how we *evaluate*

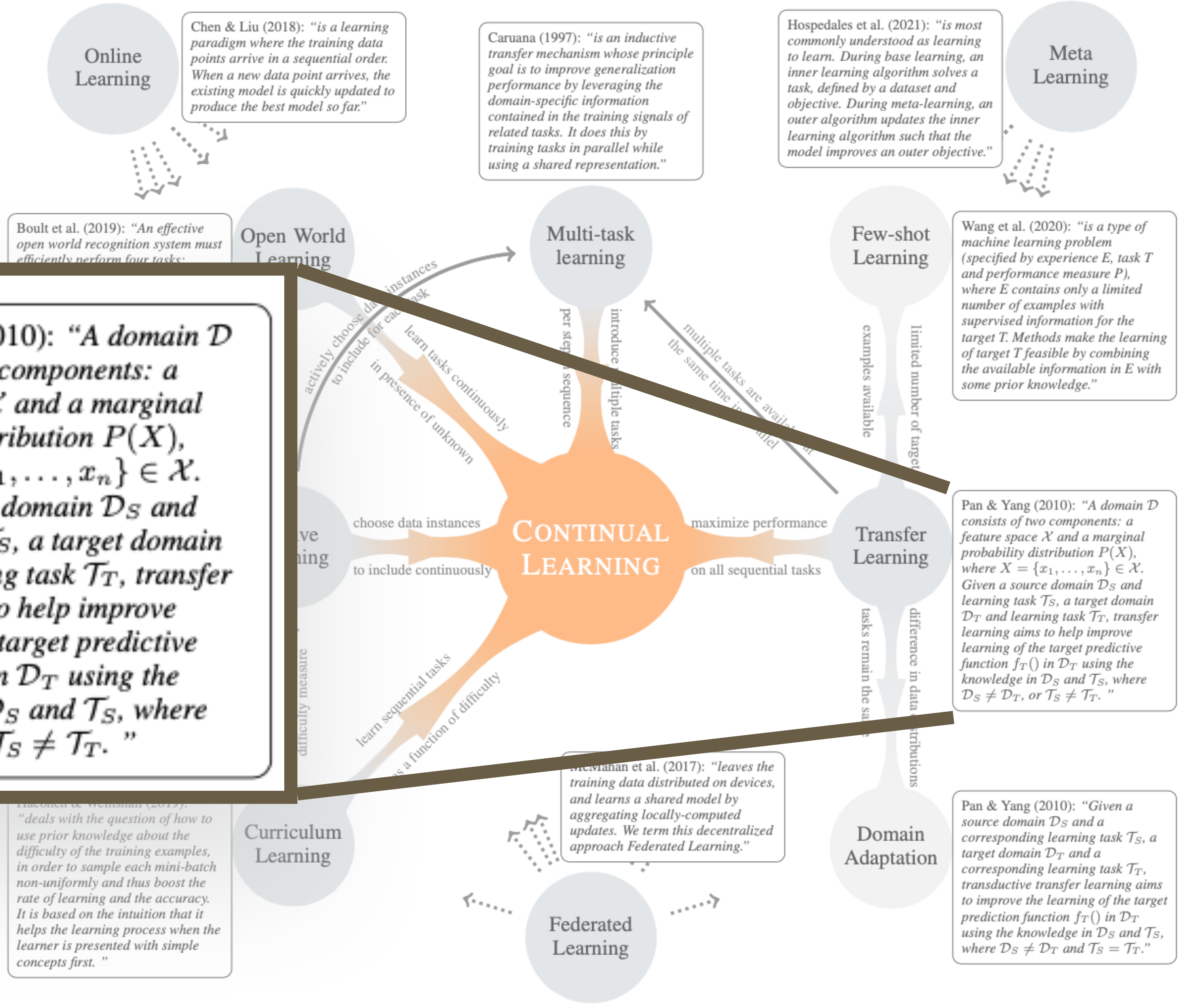
Each *paradigm* seems to have a particular *preference* (potentially neglecting other important factors)



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

Evaluation & related paradigms



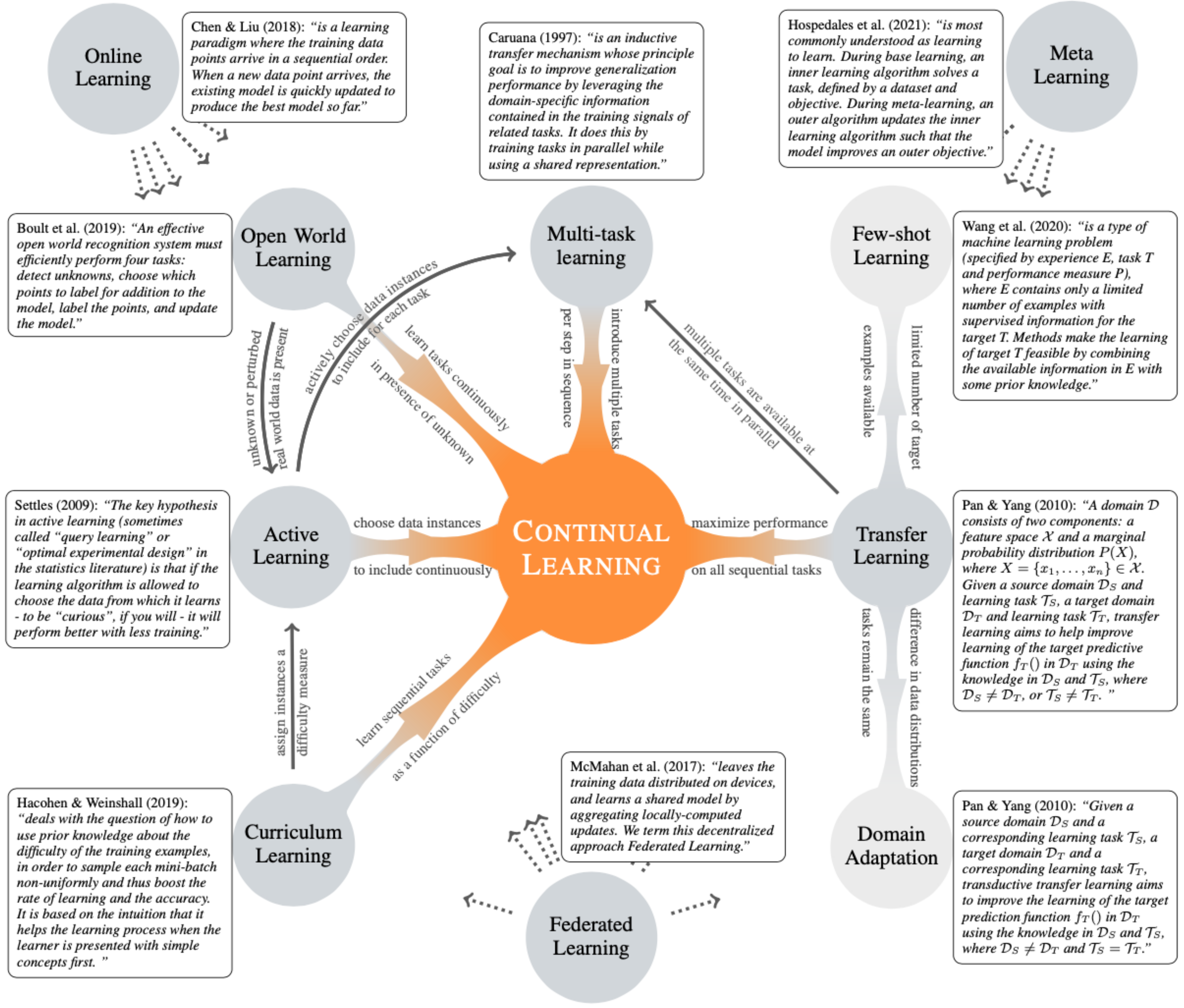


Mundt et al, "CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability", ICLR 2022

Evaluation & related paradigms

Do distinct applications warrant the existence of numerous scenarios?

Yes!
—> but make inspiration in set-up transparent & promote comparability!



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

Can we compare fairly?

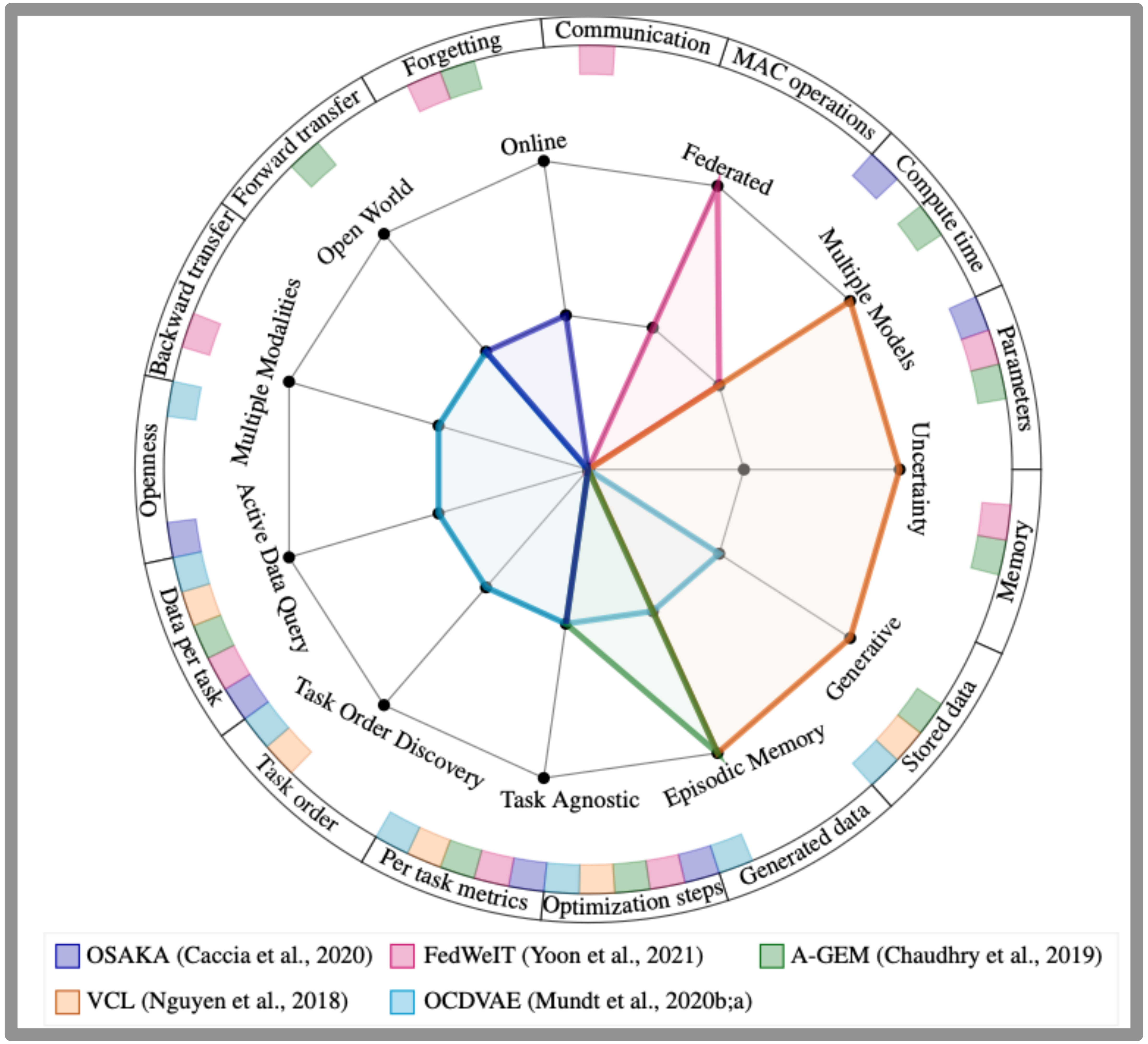
The Continual Learning EValuation Assessment (CLEVA-) Compass

Inner compass level (star plot):
paradigm inspiration + setting (assumptions)

Inner compass level of supervision:
“Rings” indicate level of supervision.

Outer compass level:
Comprehensive set of practical measures

Encourages transparency, summarizes incentives, and promotes comparability in a compact visual form



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

Thank You!

Martin Mundt

Email: martin.mundt@tu-darmstadt.de

Web: https://owll-lab.com/

Twitter/X: @mundt_martin