2023 AAAI Continual Causality Bridge Tutorial:

Putting the *Continual* in Continual Causality

Keiland Cooper

UCI University of California, Irvine

ContinualAI

NSF

Center for the Neurobiology of Learning and Memory

UCI Data Science Initiative

Martin Mundt

OWL

ContinualAI

TECHNISCHE UNIVERSITÄT DARMSTADT

hessian.AI

What is Continual Learning?

What is Continual Learning?

Continual Learning (CL) strives to capture the principals of how humans and animals learn adaptively and continuously about the world

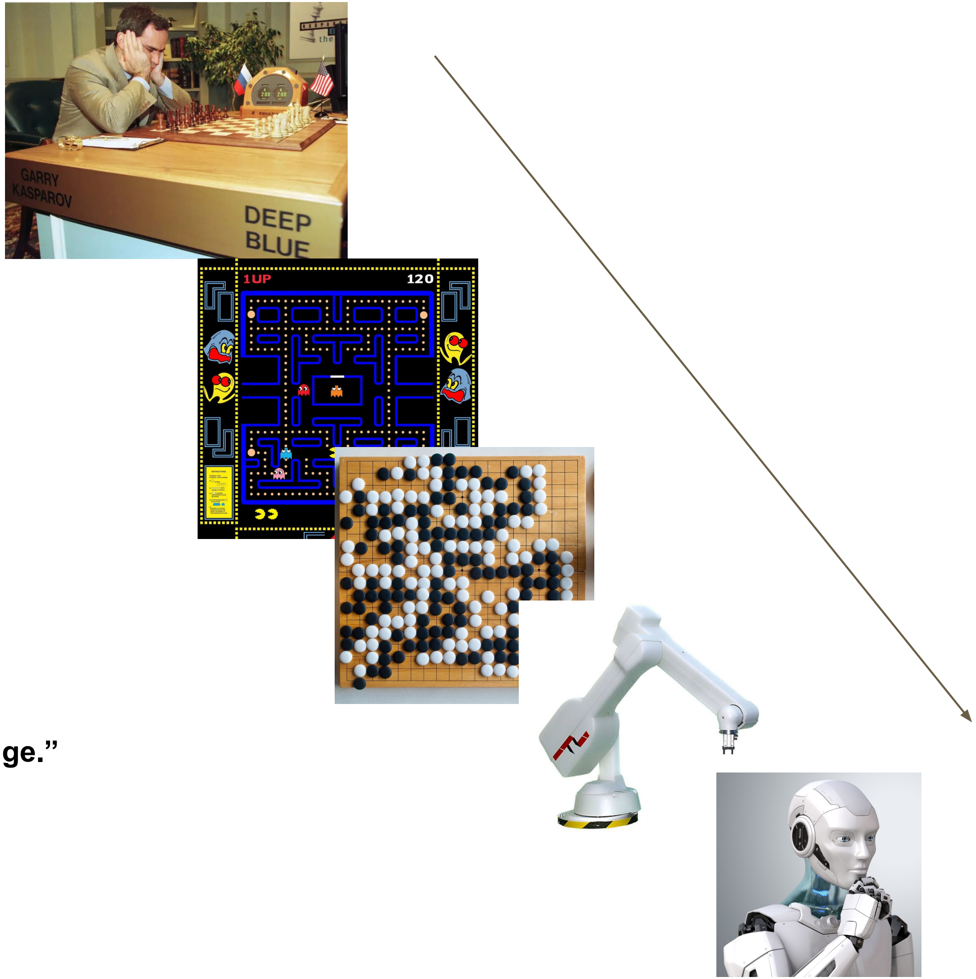
How does learning provide the autonomous, incremental, development of increasingly complex knowledge, skills, and behaviors?

The Promise

So if humans and animals learn continually, why shouldn’t our machines?

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

At the most, its one pathway towards strong, general artificial intelligence

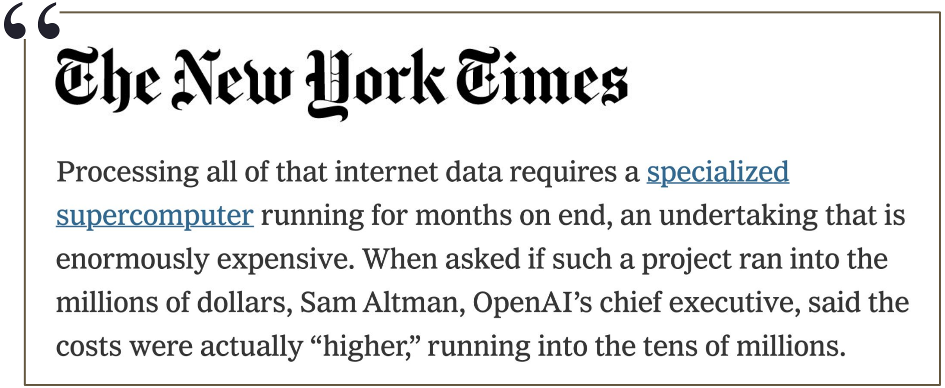


“Intelligence is the ability to adapt to change.”
- Stephen Hawking

The Practicalities

In the meantime, CL has direct benefits towards improving AI systems across research and 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.

The Problem

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

- Narrow, fixed models, lacking robustness
- Incomplete and growing datasets
- Catastrophic forgetting

Thus, CL research is growing more and more by the day to solve these problems

CHANDLER
DEVELOPING STORY
CRASH INVOLVING SELF-DRIVING WAYMO
15.COM 15 SUSPECTS ARE CONSIDERED DANGEROUS. CALL 911 IF YOU SEE THEM. 15 TUNE

Old ideas...

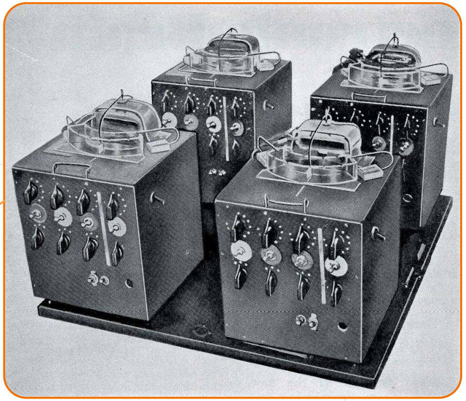
Mon July 23
1956

Ashby: Mon
July 23, 1956

Problem of homeostat

1) No memory of previous solns. i.e. learns to handle A, then learns B → then going back to A takes as much time as before

2) time to go to =lib. is too long.



Ray Solomonoff's notes on Ross Ashby's talk during the Dartmouth Summer Research Project in Artificial Intelligence, 1956

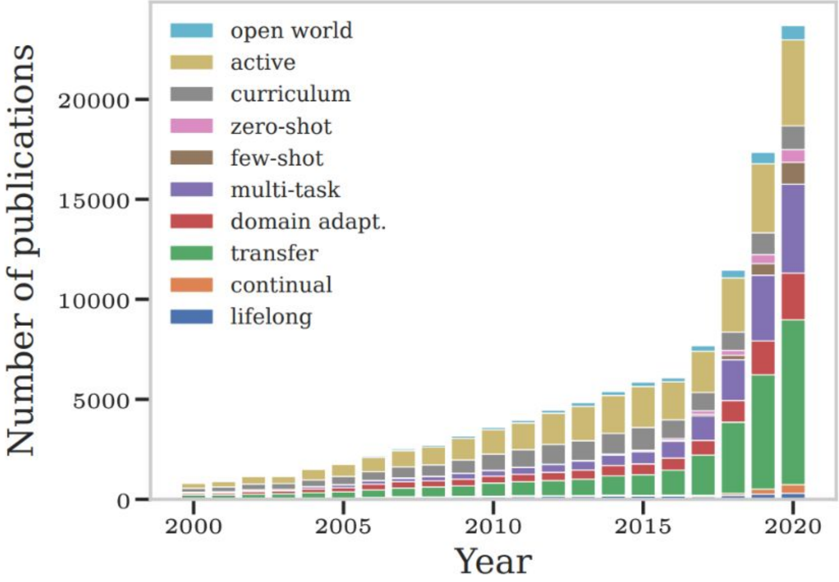
Old(er) Ideas…

"There exists in the mind of man a block of wax … harder, moister, and having more or less of purity in one than another… **the soft are good at learning, but apt to forget; and the hard are the reverse**"

– *Plato, Theaetetus, ~369 BCE*

(a) blossoming field(s)

Number of Continual Learning and related publications over time

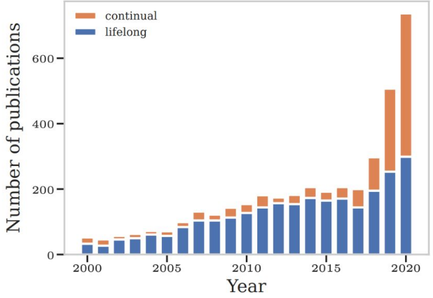


Adapted from Mundt et al. 2022, ICLR

A rose by any other name...

- **Continual Learning**
- Continuous Learning
- Lifelong Learning
- Sequence learning
- Online Learning
  - Streaming Learning
- Never-ending-learning
- Knowledge Aggregation
- The Stability / Plasticity dilemma
- And more!

- One shot / few shot learning
- Transfer learning
  - Domain Adaptation
- Curriculum Learning
- Meta-learning
  - Learning to learn
- Active Learning
- Multi-task learning
- Meta-learning

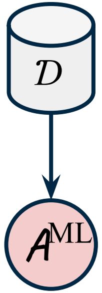


A Formalism

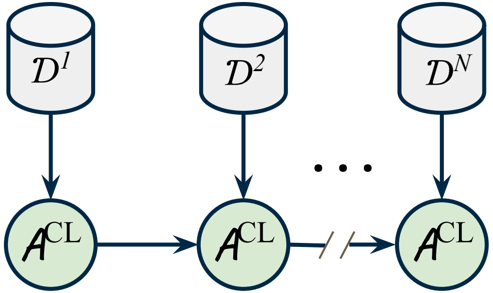
Traditional Machine Learning

Continual Learning

Data



□ =



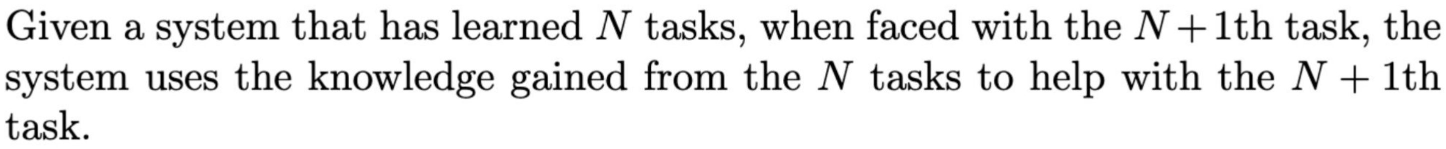
Model

Loss

Minimize loss over $\mathcal{D}$

Minimize loss over □

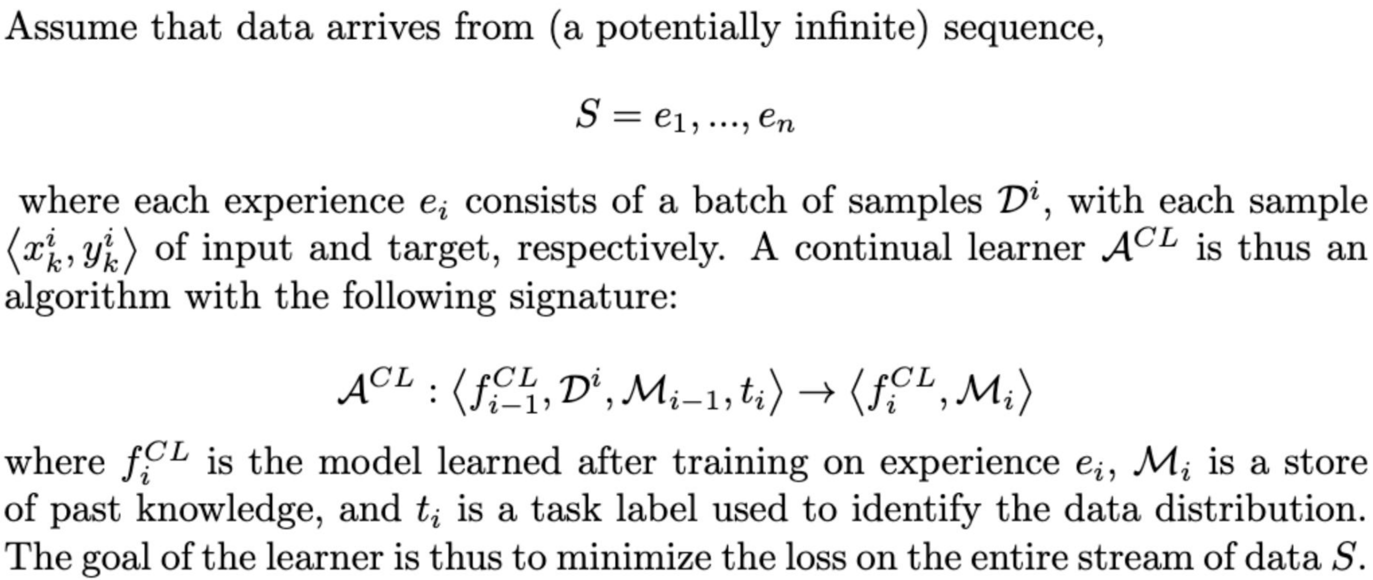
A formalism



Given a system that has learned N tasks, when faced with the $N+1$th task, the system uses the knowledge gained from the N tasks to help with the $N+1$th task.

Adapted from Thrun et al., 1996

A formalism



Assume that data arrives from (a potentially infinite) sequence,

$$S = e_1, ..., e_n$$

where each experience e_i consists of a batch of samples $\mathcal{D}^i$, with each sample $\langle x_k^i, y_k^i \rangle$ of input and target, respectively. A continual learner $\mathcal{A}^{CL}$ is thus an algorithm with the following signature:

$$\mathcal{A}^{CL} : \langle f_{i-1}^{CL}, \mathcal{D}^i, \mathcal{M}_{i-1}, t_i \rangle \rightarrow \langle f_i^{CL}, \mathcal{M}_i \rangle$$

where f_i^{CL} is the model learned after training on experience e_i, $\mathcal{M}_i$ is a store of past knowledge, and t_i is a task label used to identify the data distribution. The goal of the learner is thus to minimize the loss on the entire stream of data S.

Adapted from Lesort et al., 2020

One (major) hiccup

How to learn from every book?

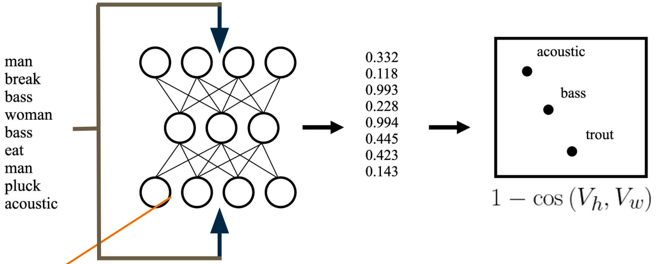
Step 1: Grab each book as a training corpus

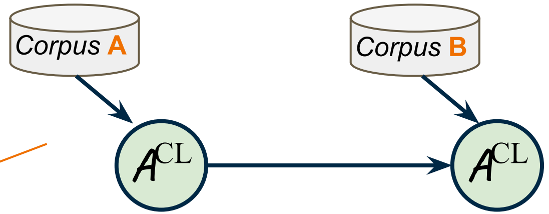
Step 2: Take an off-the-shelf word2vec model

Step 3: Train the model on each book *sequentially*

Step 4: Use the learned weights as a marker for learned semantics

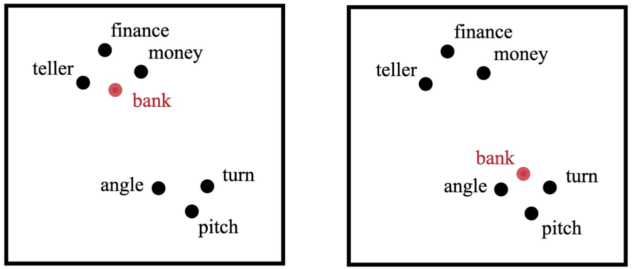
Step 5: Assert the semantics are meaningful…





Semantics after training on **Corpus A**

Semantics after training on **Corpus B**



Prior learning

Recent learning

Adapted from Cooper et al. 2018, MWCSC

The sequential learning problem

Connectionist architectures fail to learn sequentially

Michael McCloskey & Neal J.Cohen (1989)

Taught networks addition & multiplication problems, as well as a retroactive interference psychology problem

Noticed that learning new tasks disrupted the old task

Enter: catastrophic interference

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

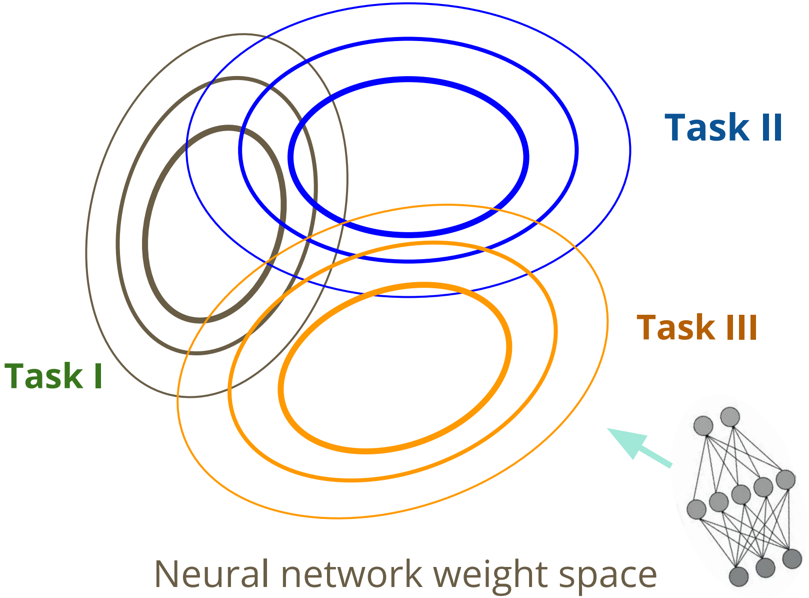
Why might NN’s be so forgetful?

Let’s say we want to teach a network 3 tasks.

We train on each task sequentially, with no direct overlap of task examples

We can think of the networks weights as occupying some possible space or landscape of configurations to solve a given task

The center of each distribution solves that task



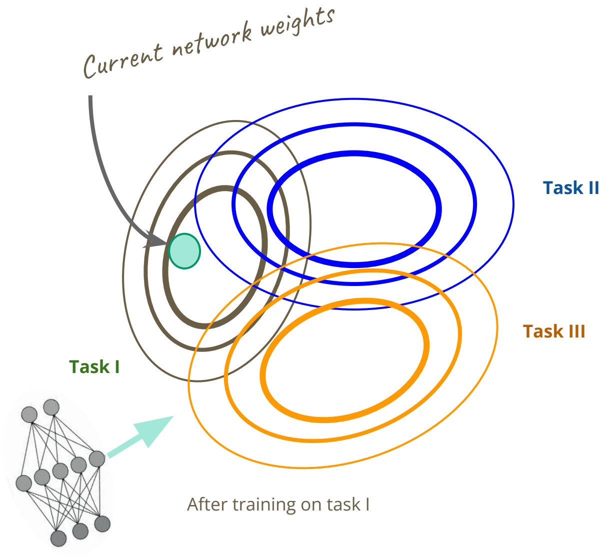
Task II

Task III

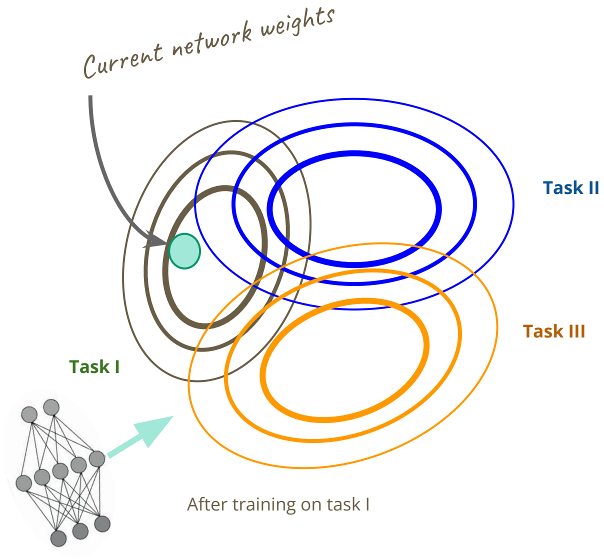
Task I

Neural network weight space

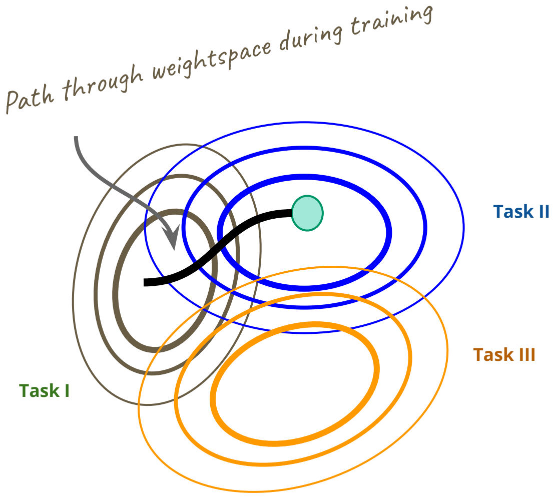
Why might NN's be so forgetful?



Why might NN's be so forgetful?

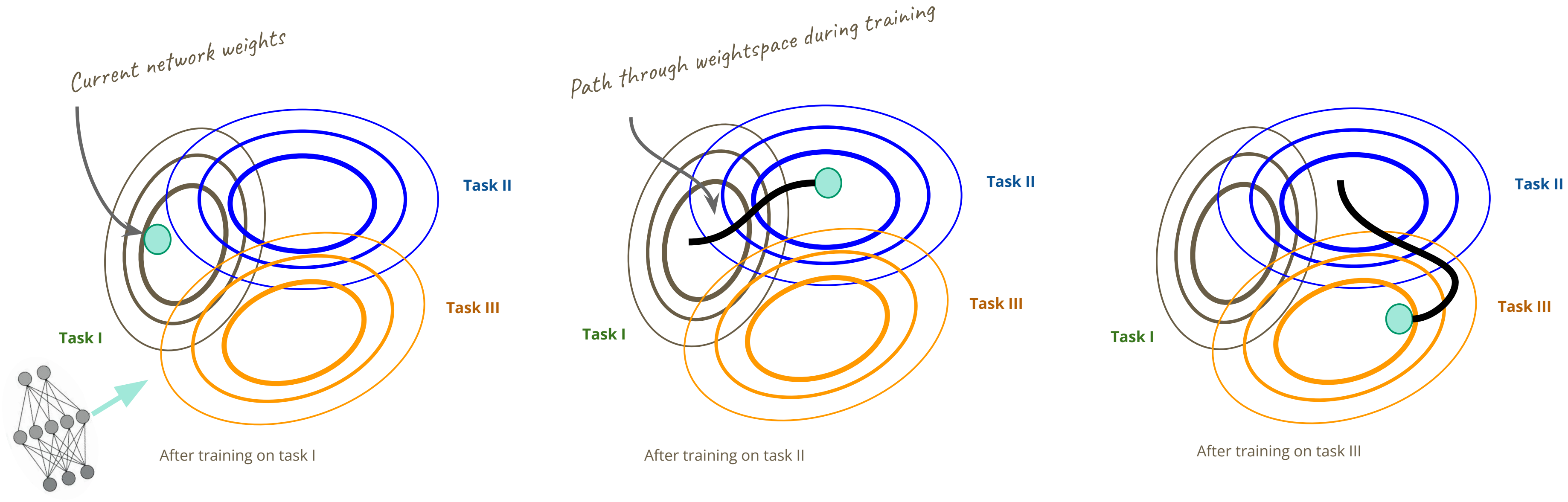


After training on task I



After training on task II

Why might NN’s be so forgetful?



Current network weights

Path through weightspace during training

After training on task I

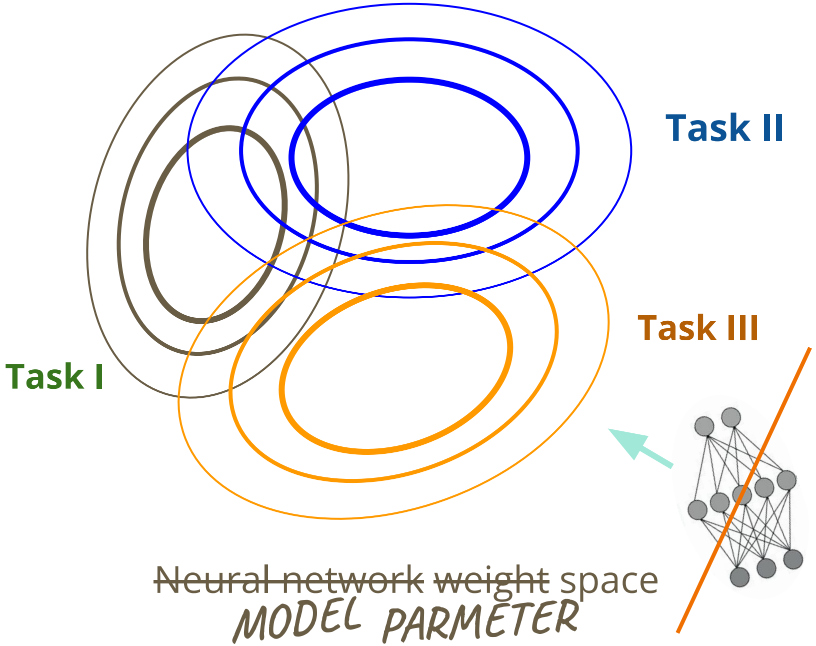
After training on task II

After training on task III

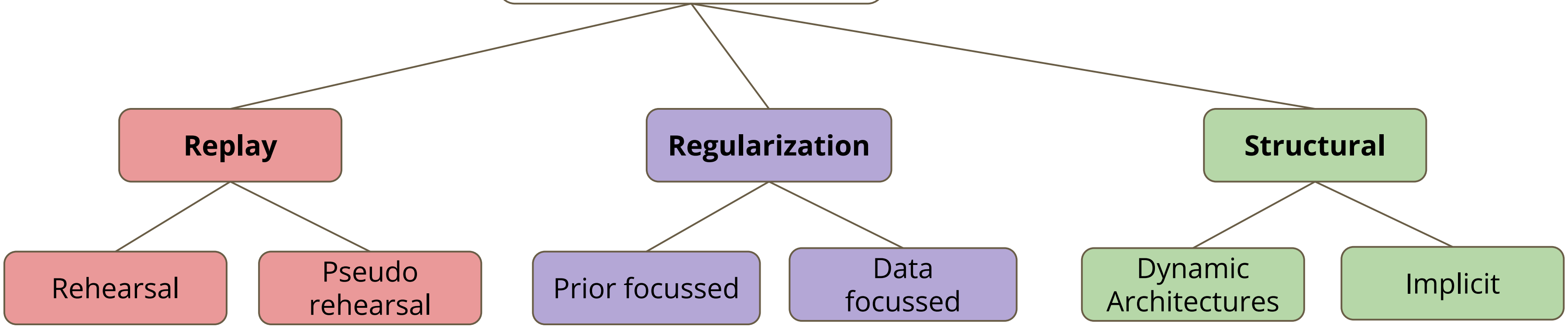
Not just neural networks

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 main CL Strategies



Replay

Rehearsal

Pseudo rehearsal

Regularization

Prior focussed

Data focussed

Structural

Dynamic Architectures

Implicit

Leverage past samples of previous task data

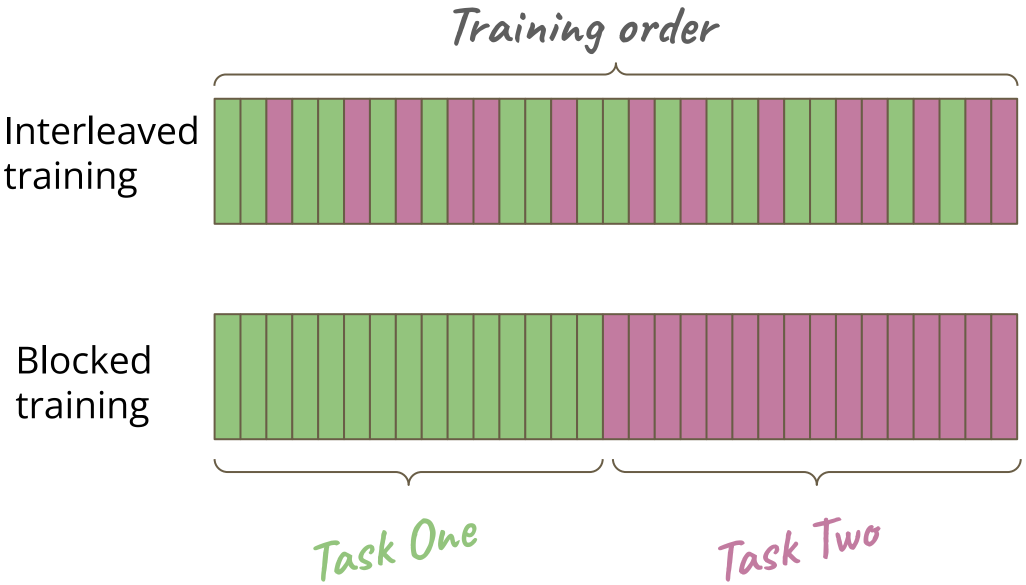
Alter the weight dynamics as a function of tasks

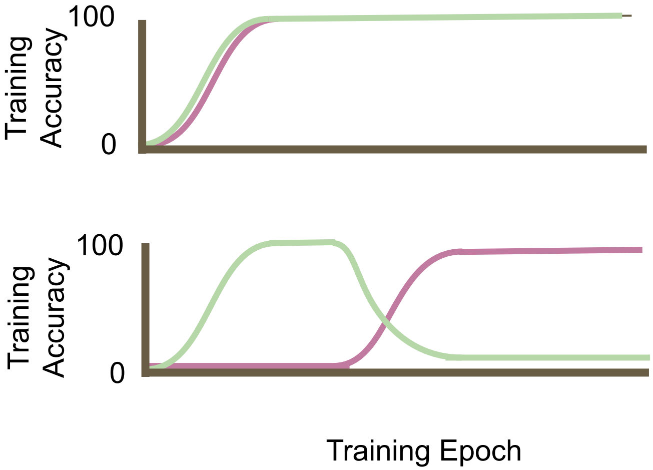
Change the macro or micro architecture of the network

Adapted from De Lange et al. 2021, TPAMI

“Replay” approaches 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 task data in memory buffer

Replace memory buffer with new examples

Disadvantages:

- Utilize separate memory
- Does not respect data privacy

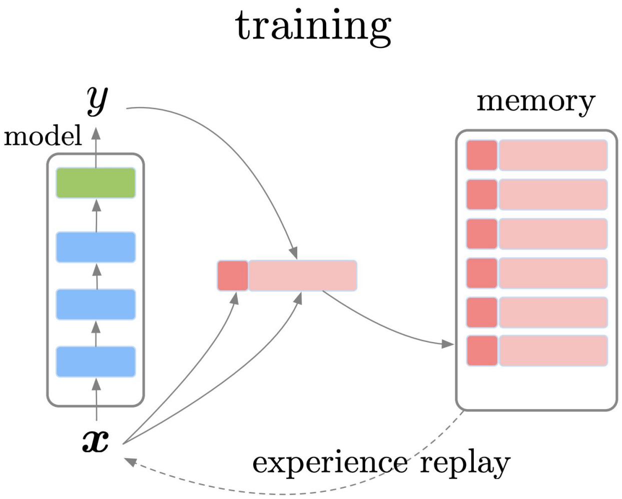


Figure from Masson d'Autume et al. 2019, ArXiv

A caveat...

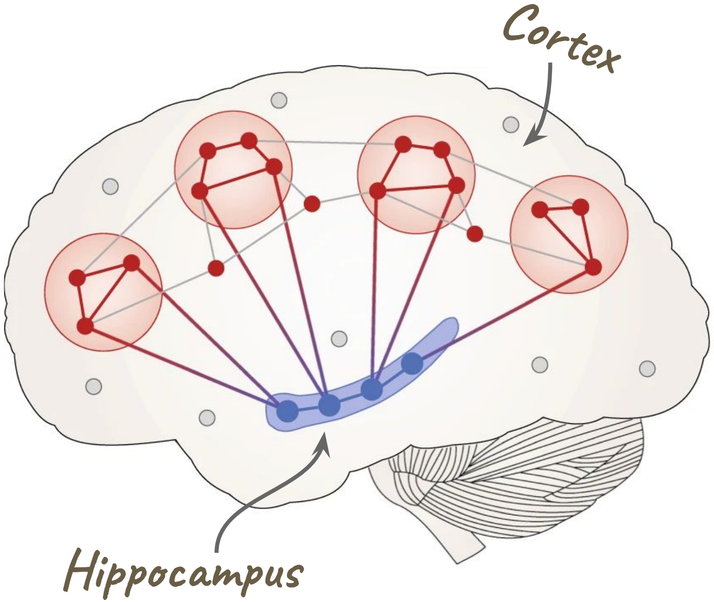
> “While it is an effective method in ANNs, rehearsal is unlikely to be a realistic model of biological learning mechanisms, as in this context the actual old information (accurate and complete representation of all items ever learned by the organism) is not available.”
>
> – *Robert French, 1997*

Replay IS biologically plausible

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

1. Hippocampus is a fast learning system
2. Cortex is a slow learning system
3. Hippocampus replays memories to cortex
4. Cortex generalizes memories
5. 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*

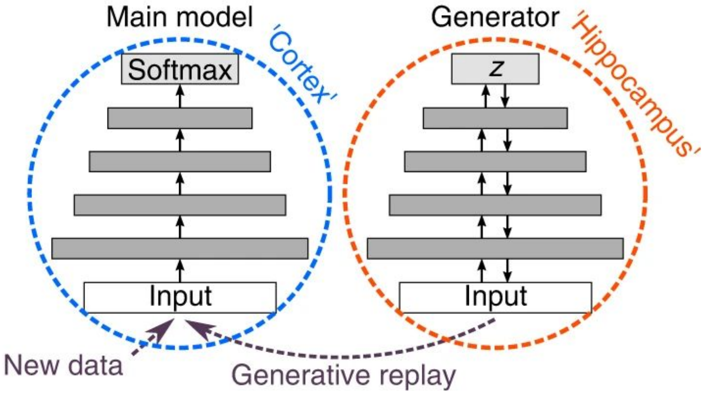
Pseudo-Replay is biologically plausible

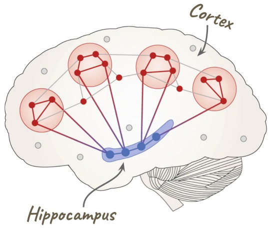
Generative replay:

- 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

“Regularization” approaches to alleviate forgetting

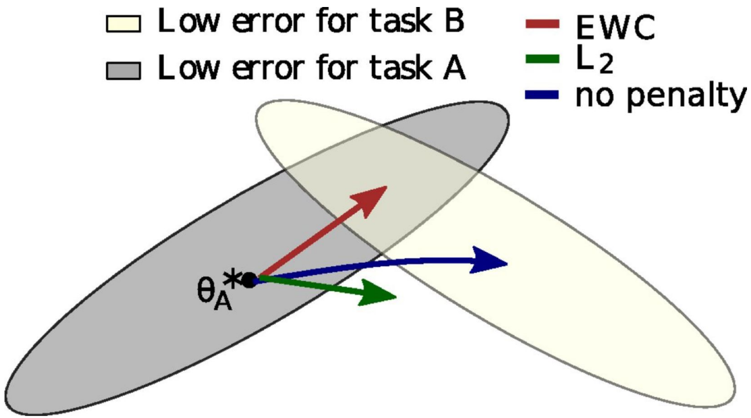
Elastic Weight Consolidation

Don’t jump directly to optimizing for a new task, preserve the old weights

Instead, penalize previously learned parameters

Step-1: Approximate the Fisher Information Matrix *F*

Step-2: Apply a squared regularisation loss to penalise any drastic 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

Learning and novel sensory experience promote rapid dendritic spine alterations

How to reconcile this with memory stability?

While novel experience promotes spine elimination,

A fraction of spines are maintained over long durations

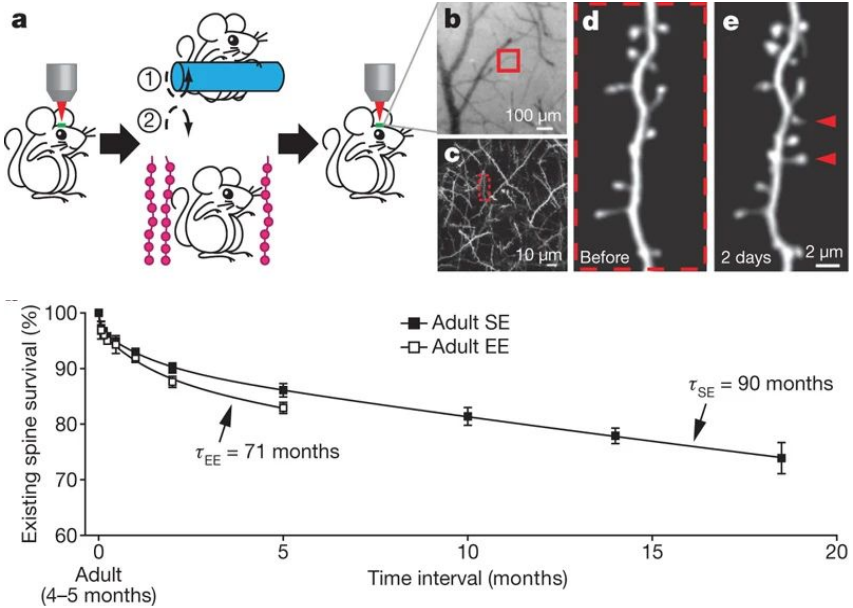


Figure adapted from Cichon & Gan 2015

Regularization and replay are complementary

> "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*

“Structural” approaches 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.”

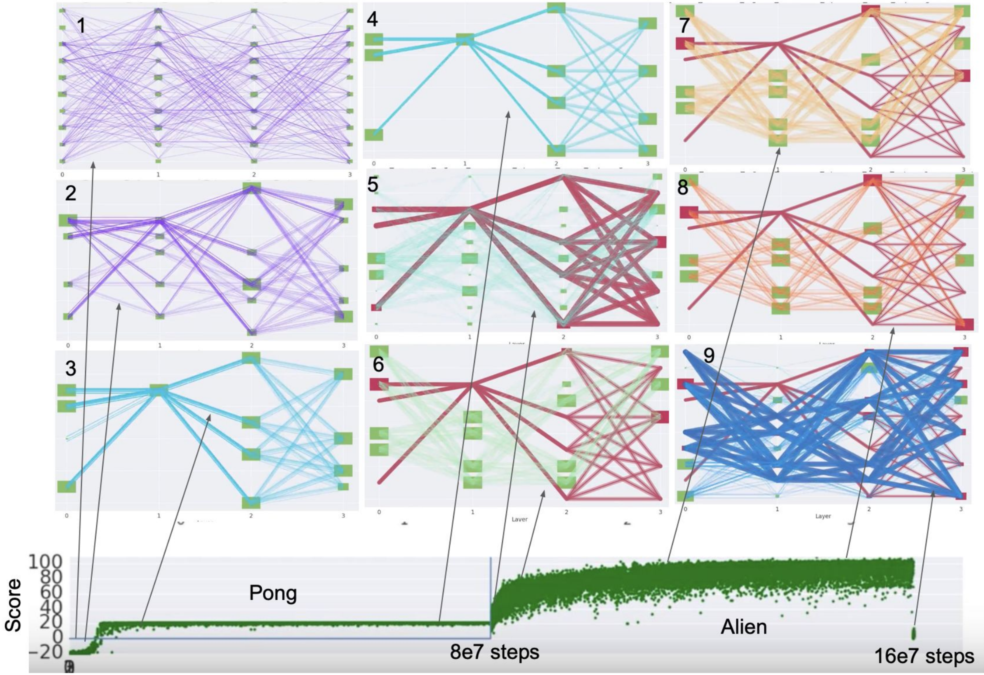
The implicit perspective

- Recall the regularization perspective: identify + constrain important params
- We could assume over-parametrization + try to “sparsify” our parameters
- We create “sub-models” that are primarily responsible for a specific task

The implicit perspective

Example: *Pathways/PathNets*

- Start with an over-parametrized model
- Constrain a task to use a subset of parameters
- 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

The explicit perspective

Inspiration from *neurogenesis*?

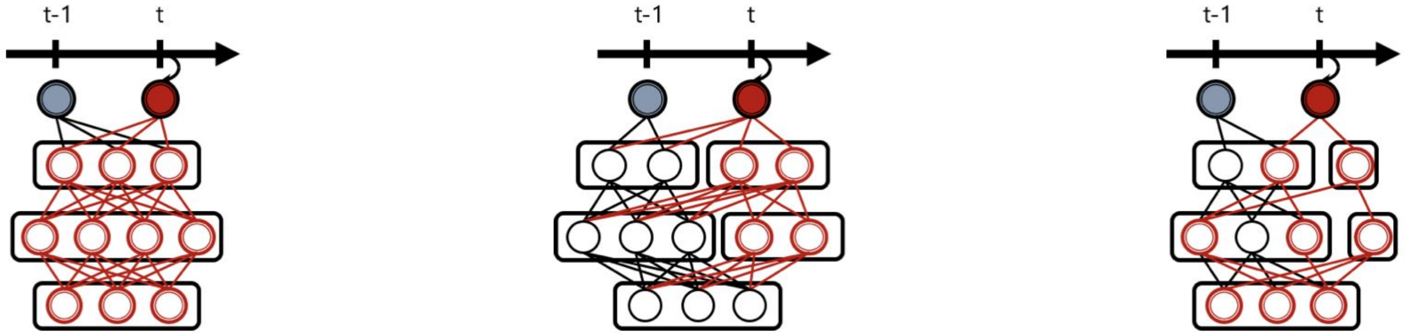
“After two decades of research, the neurosciences have come a long way from accepting that neural stem/progenitor cells generate new neurons in the adult mammalian hippocampus to unraveling the functional role of adult-born neurons in cognition and emotional control. The finding that new neurons are 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 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

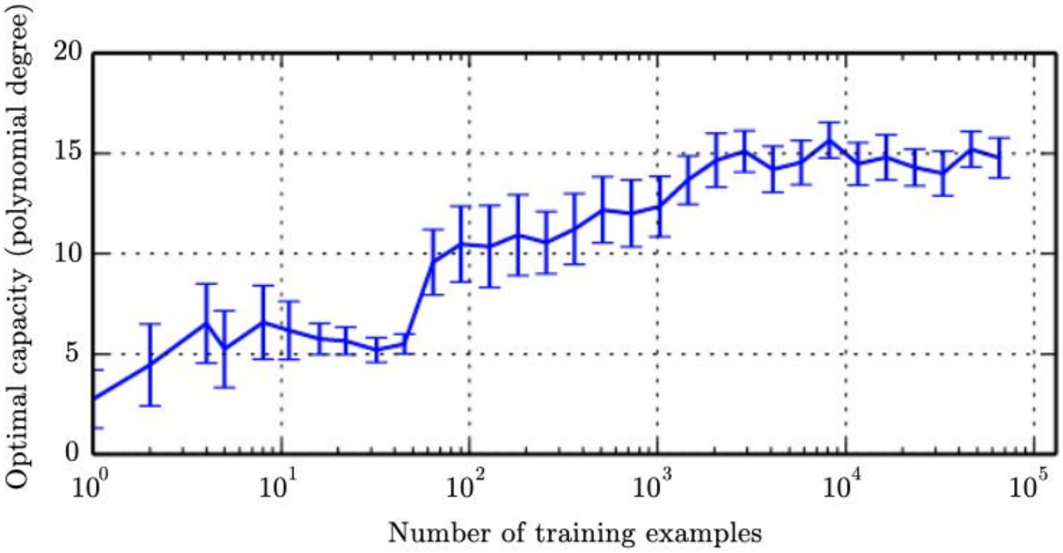
(b) No-retraining w/ expansion

(c) Partial retraining w/ expansion

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

Evaluation is challenging

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

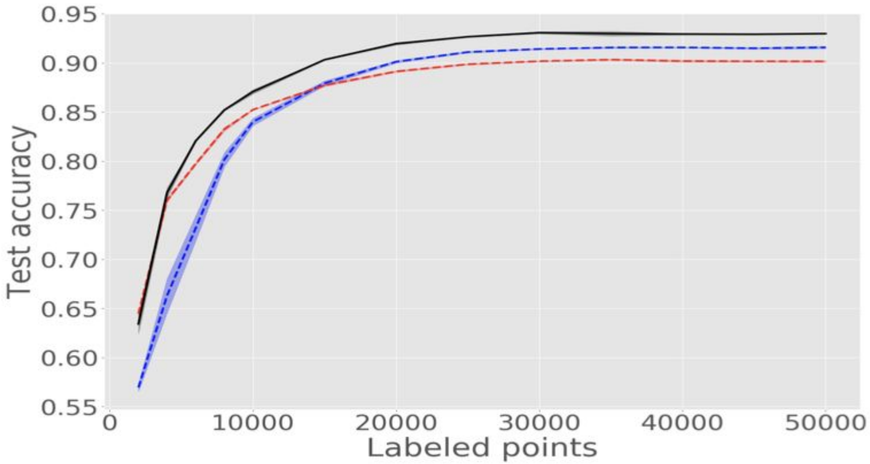


Deep Learning, Goodfellow, Bengio, Courville, MIT Press 2016, Machine Learning Basics chapter, page 114.

A small active learning detour

Let's take a small detour on an active learning experiment:
data from a pool is queried to be added to the dataset over time (x-axis)

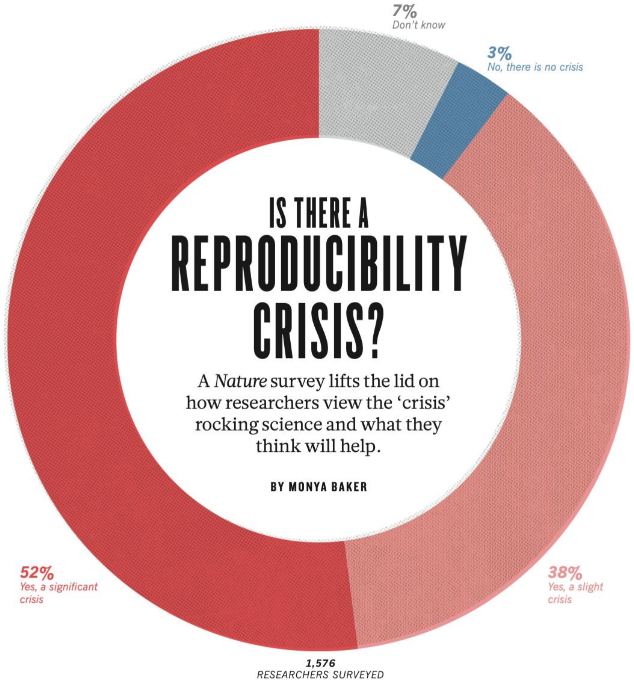
1. *Black line*: incremental architecture
2. *Blue line*: fixed Resnet (large)
3. *Red line*: fixed small architecture (start of the incremental one)



Geifman & El-Yaniv, "Deep Active Learning with a Neural Architecture Search", NeurIPS 2019

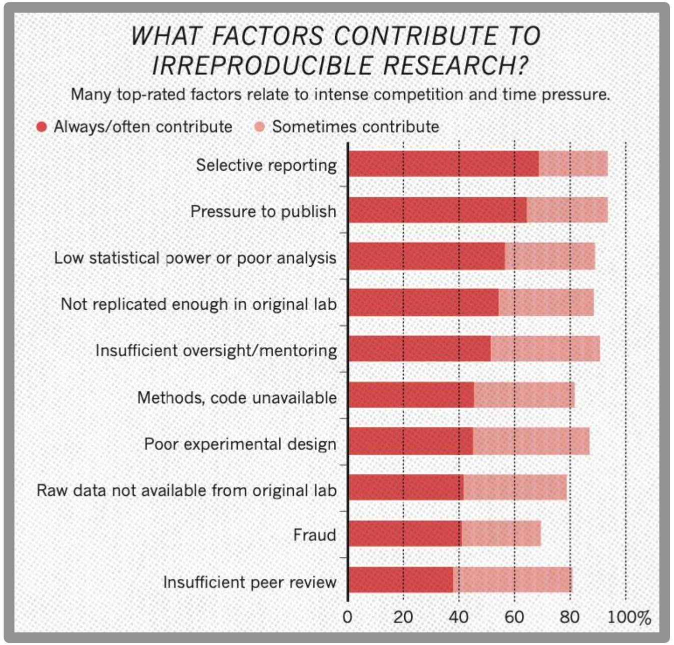
So how do we evaluate &
what do we care about in CL?

Reproducibility crisis?


IS THERE A REPRODUCIBILITY CRISIS?

A *Nature* survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

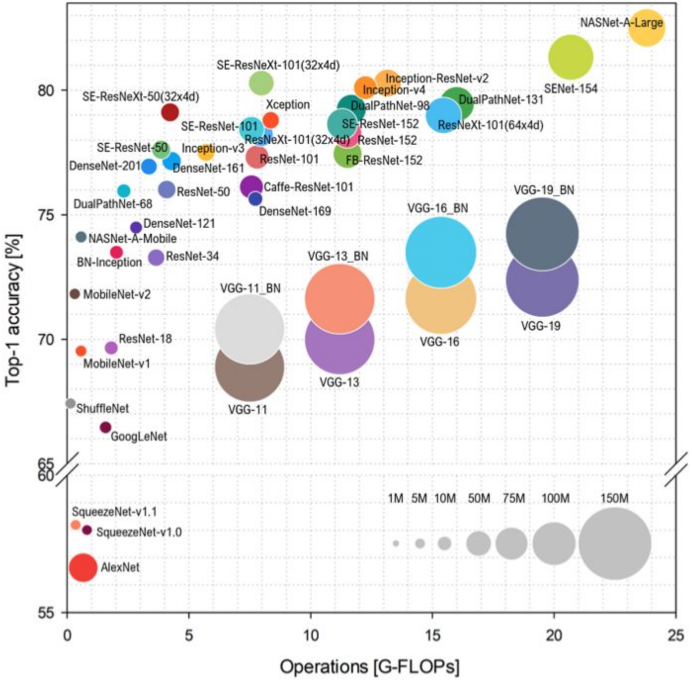
BY MONYA BAKER


WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

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

Reproducibility in (static) ML



Even in static scenarios:

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

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

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

Popular scenarios in continual learning

What are some of the sequences of tasks that are popular in continual learning?

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

Challenge of defining a continual learning "task"

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

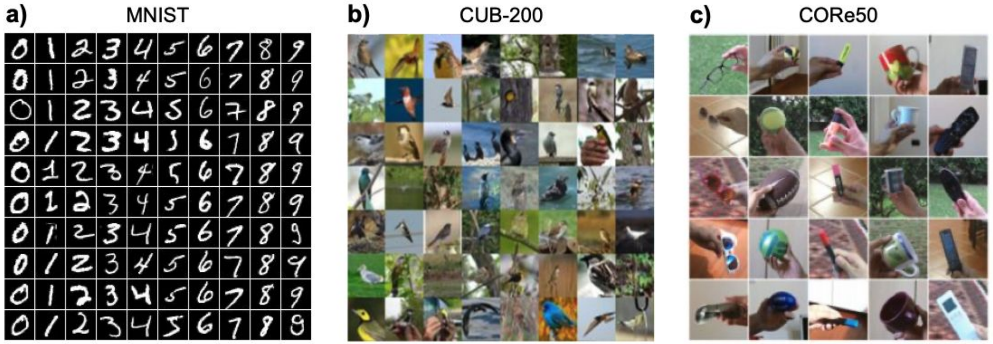
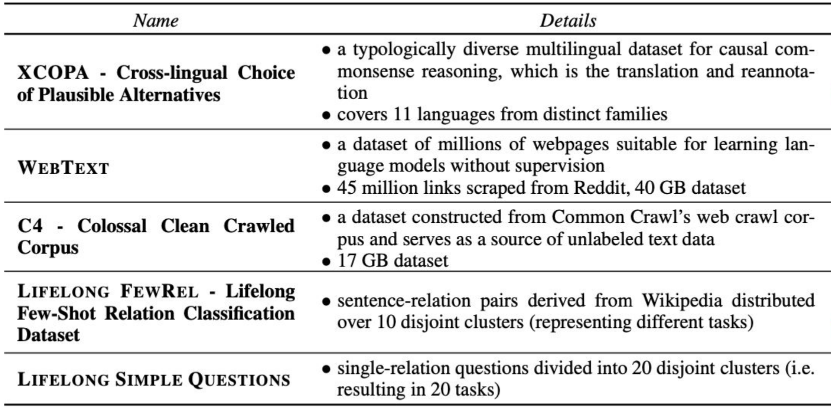


Figure 3: Example images from benchmark datasets used for the evaluation of lifelong learning



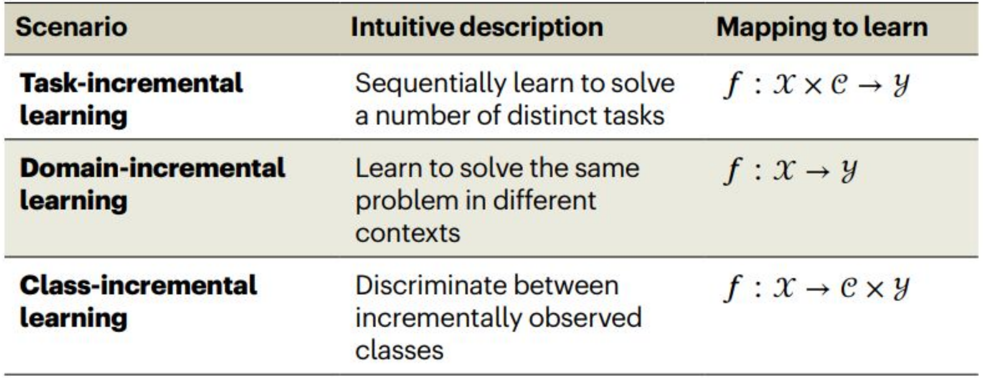
Name	*Details*
XCOPA - Cross-lingual Choice of Plausible Alternatives	• a typologically diverse multilingual dataset for causal commonsense reasoning, which is the translation and reannotation • covers 11 languages from distinct families
WEBTEXT	• a dataset of millions of webpages suitable for learning language models without supervision • 45 million links scraped from Reddit, 40 GB dataset
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
LIFELONG FEWREL - Lifelong Few-Shot Relation Classification Dataset	• sentence-relation pairs derived from Wikipedia distributed over 10 disjoint clusters (representing different tasks)
LIFELONG SIMPLE QUESTIONS	• single-relation questions divided into 20 disjoint clusters (i.e. resulting in 20 tasks)

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

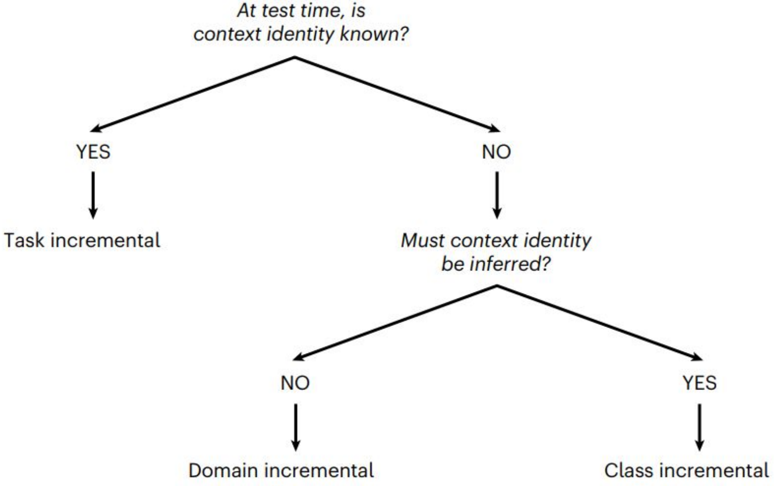
Biesialska et al, "Continual Learning in Natural Language Processing: A Survey", COLING 2020

Challenge of defining a continual learning “task”

It’s challenging to agree on *tasks* (and maybe we don’t need to agree)



Scenario	Intuitive description	Mapping to learn
Task-incremental learning	Sequentially learn to solve a number of distinct tasks	$f : \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{Y}$
Domain-incremental learning	Learn to solve the same problem in different contexts	$f : \mathcal{X} \rightarrow \mathcal{Y}$
Class-incremental learning	Discriminate between incrementally observed classes	$f : \mathcal{X} \rightarrow \mathcal{C} \times \mathcal{Y}$



At test time, is context identity known?

YES

Task incremental

NO

Must context identity be inferred?

NO

Domain incremental

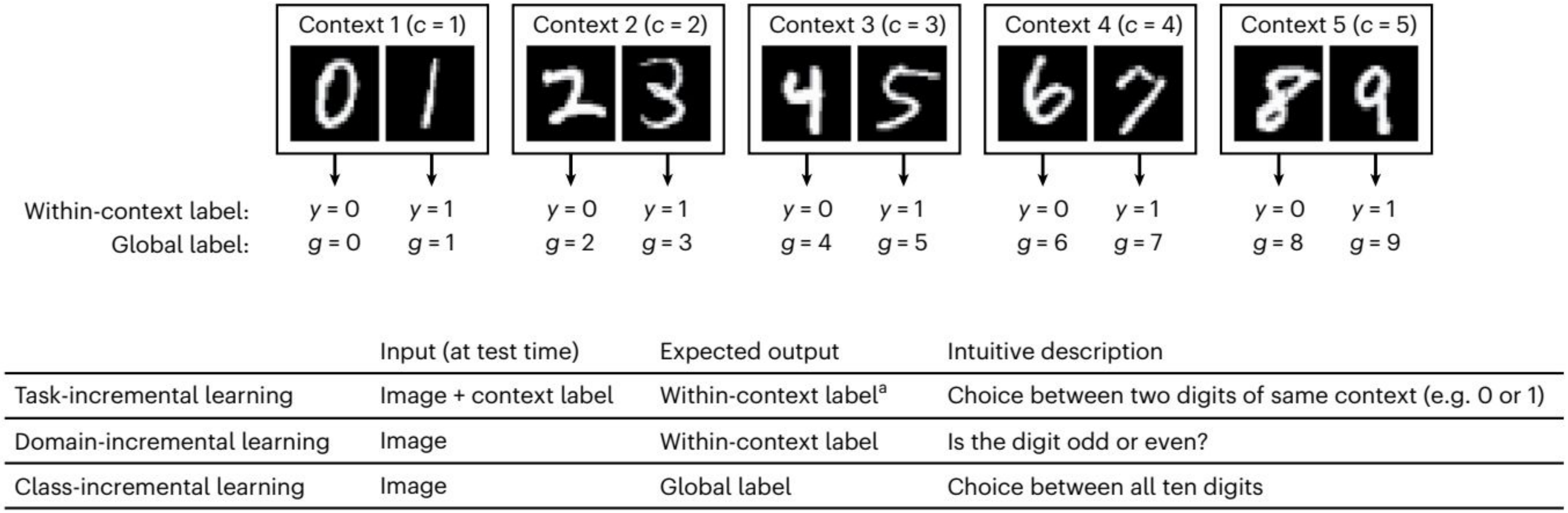
YES

Class incremental

van de Ven et al, “Three types of incremental learning”, Nature Machine Intelligence 4, 2022

Challenge of defining a continual learning "task"

It's challenging to agree on *tasks* (and maybe we don't need to agree)



	Input (at test time)	Expected output	Intuitive description
Task-incremental learning	Image + context label	Within-context label[a]	Choice between two digits of same context (e.g. 0 or 1)
Domain-incremental learning	Image	Within-context label	Is the digit odd or even?
Class-incremental learning	Image	Global label	Choice between all ten digits

van de Ven et al, "Three types of incremental learning", Nature Machine Intelligence 4, 2022

There are plenty of ideas of what & how to measure

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

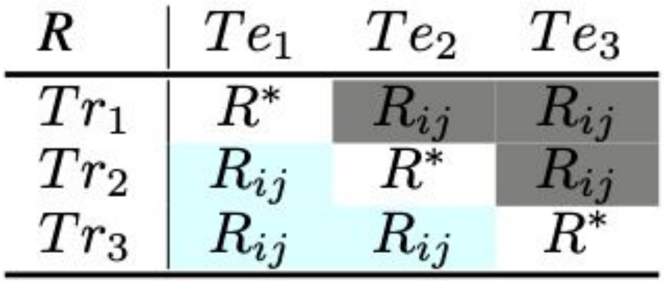
Forward and backward transfer

(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 et al, "Don’t forget, there is more than forgetting: new metrics for Continual Learning”, 2018

Memory, size and compute

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

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

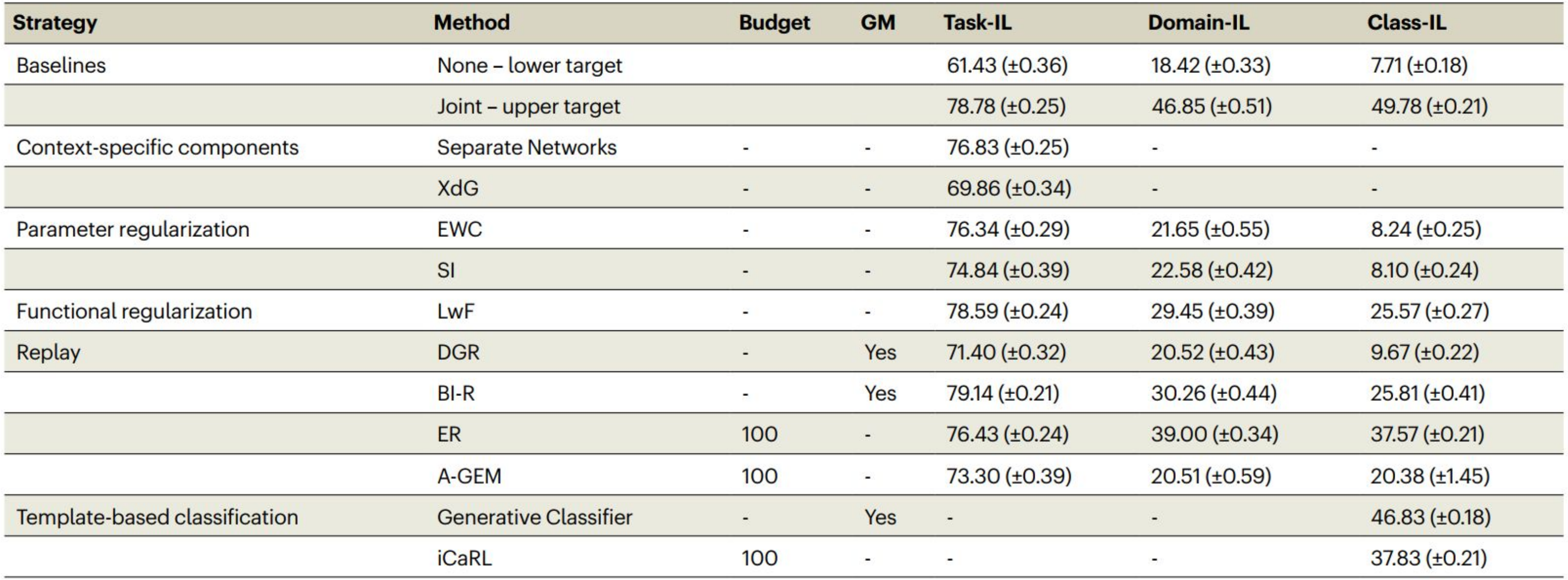
Quantifies stored amount of data (for rehearsal)

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

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

The challenge of comparison

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

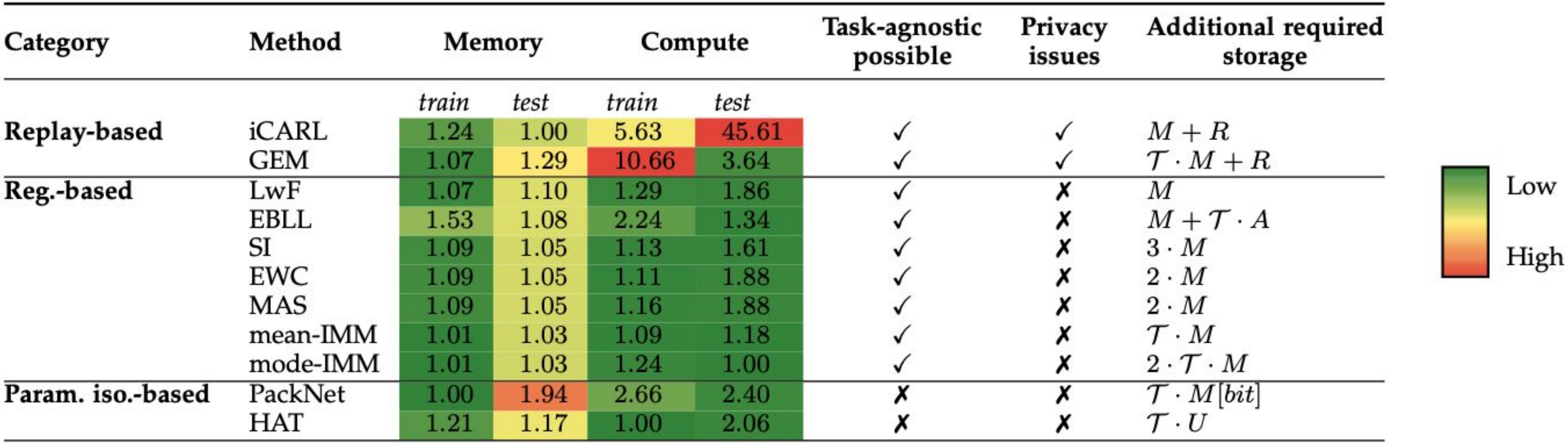


Strategy	Method	Budget	GM	Task-IL	Domain-IL	Class-IL
Baselines	None – lower target			61.43 (±0.36)	18.42 (±0.33)	7.71 (±0.18)
	Joint – upper target			78.78 (±0.25)	46.85 (±0.51)	49.78 (±0.21)
Context-specific components	Separate Networks	-	-	76.83 (±0.25)	-	-
	XdG	-	-	69.86 (±0.34)	-	-
Parameter regularization	EWC	-	-	76.34 (±0.29)	21.65 (±0.55)	8.24 (±0.25)
	SI	-	-	74.84 (±0.39)	22.58 (±0.42)	8.10 (±0.24)
Functional regularization	LwF	-	-	78.59 (±0.24)	29.45 (±0.39)	25.57 (±0.27)
Replay	DGR	-	Yes	71.40 (±0.32)	20.52 (±0.43)	9.67 (±0.22)
	BI-R	-	Yes	79.14 (±0.21)	30.26 (±0.44)	25.81 (±0.41)
	ER	100	-	76.43 (±0.24)	39.00 (±0.34)	37.57 (±0.21)
	A-GEM	100	-	73.30 (±0.39)	20.51 (±0.59)	20.38 (±1.45)
Template-based classification	Generative Classifier	-	Yes	-	-	46.83 (±0.18)
	iCaRL	100	-	-	-	37.83 (±0.21)

van de Ven et al, “Three types of incremental learning”, Nature Machine Intelligence 4, 2022

The challenge of comparison

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

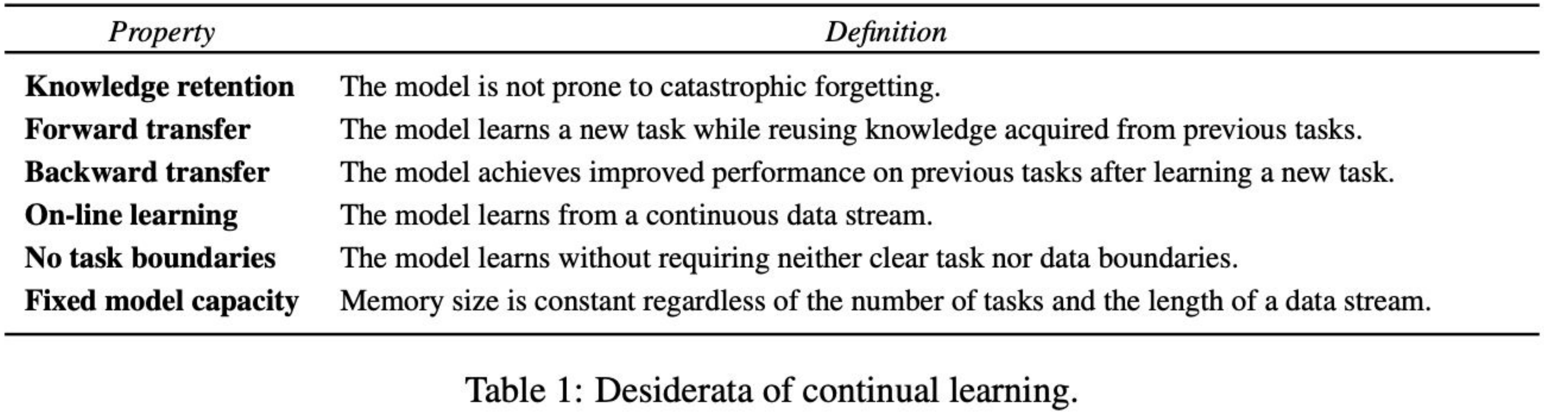


Category	Method	Memory		Compute		Task-agnostic possible	Privacy issues	Additional required storage
		train	*test*	*train*	*test*			
Replay-based	iCARL	1.24	1.00	5.63	45.61	✓	✓	$M + R$
	GEM	1.07	1.29	10.66	3.64	✓	✓	$\mathcal{T} \cdot M + R$
Reg.-based	LwF	1.07	1.10	1.29	1.86	✓	✗	M
	EBLL	1.53	1.08	2.24	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	✗	✗	$\mathcal{T} \cdot U$

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

Continual learning desiderata

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



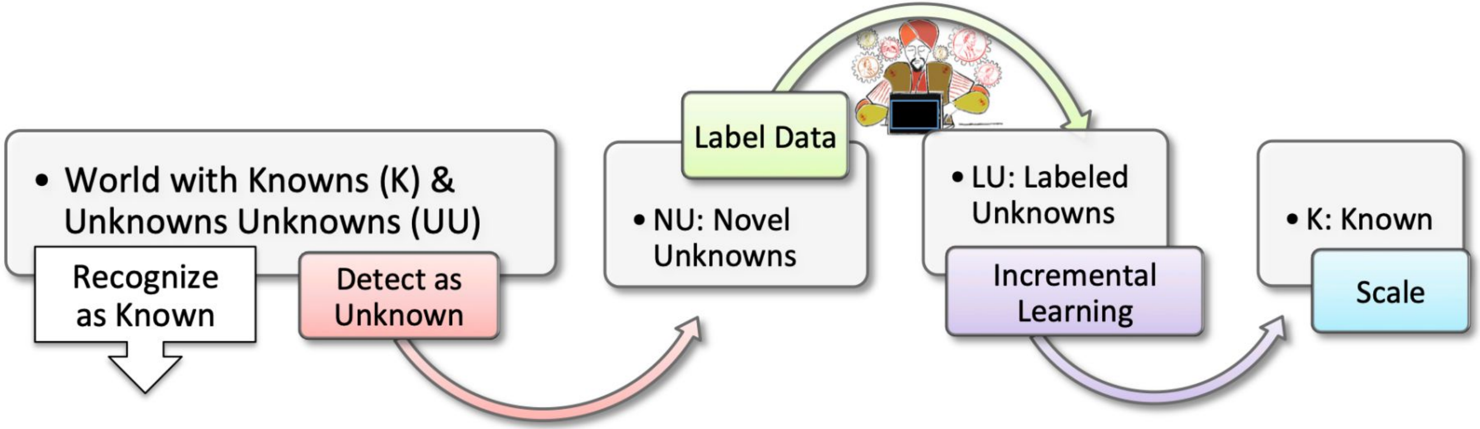
Property	*Definition*
Knowledge retention	The model is not prone to catastrophic forgetting.
Forward transfer	The model learns a new task while reusing knowledge acquired from previous tasks.
Backward transfer	The model achieves improved performance on previous tasks after learning a new task.
On-line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

Continual learning desiderata

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



- World with Knowns (K) & Unknowns Unknowns (UU)

Recognize as Known

Detect as Unknown

- NU: Novel Unknowns

Label Data

- LU: Labeled Unknowns

Incremental Learning

- K: Known

Scale

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?

Can we compare fairly?

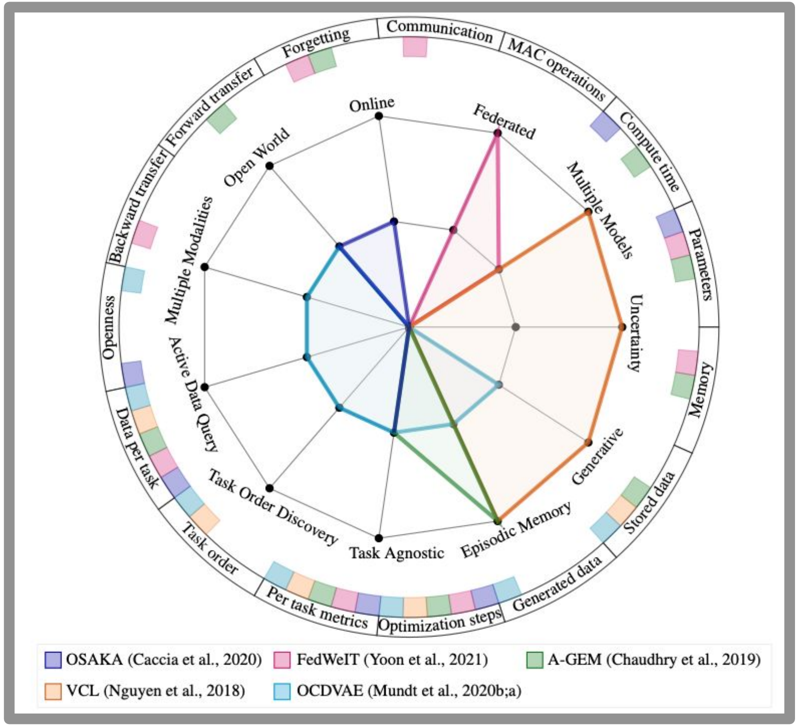
The Continual Learning EValuation Assessment (CLEVA-) Compass

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

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

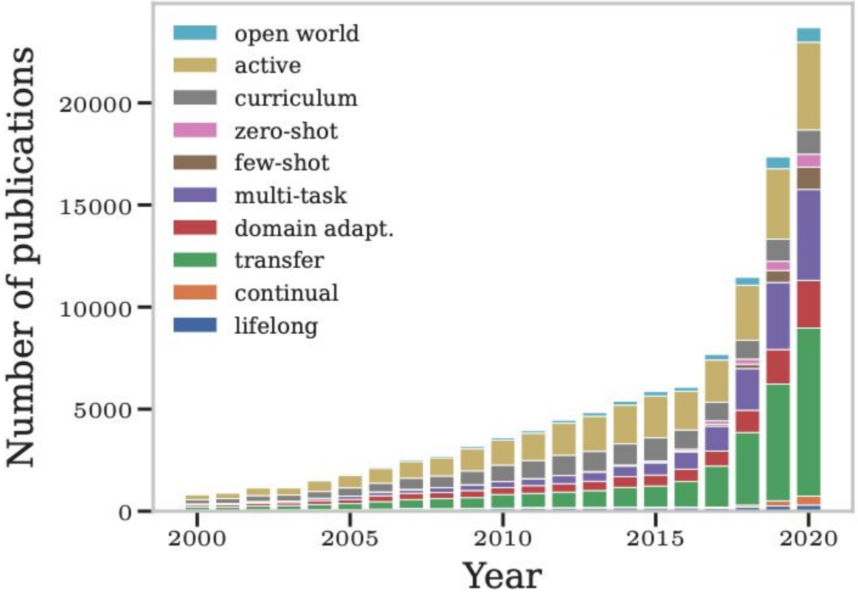
Outer compass level:
Comprehensive set of practical measures

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

The challenge of comparison & the way forward?



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

Where do we go from here?

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

Let’s think about their *origin*!

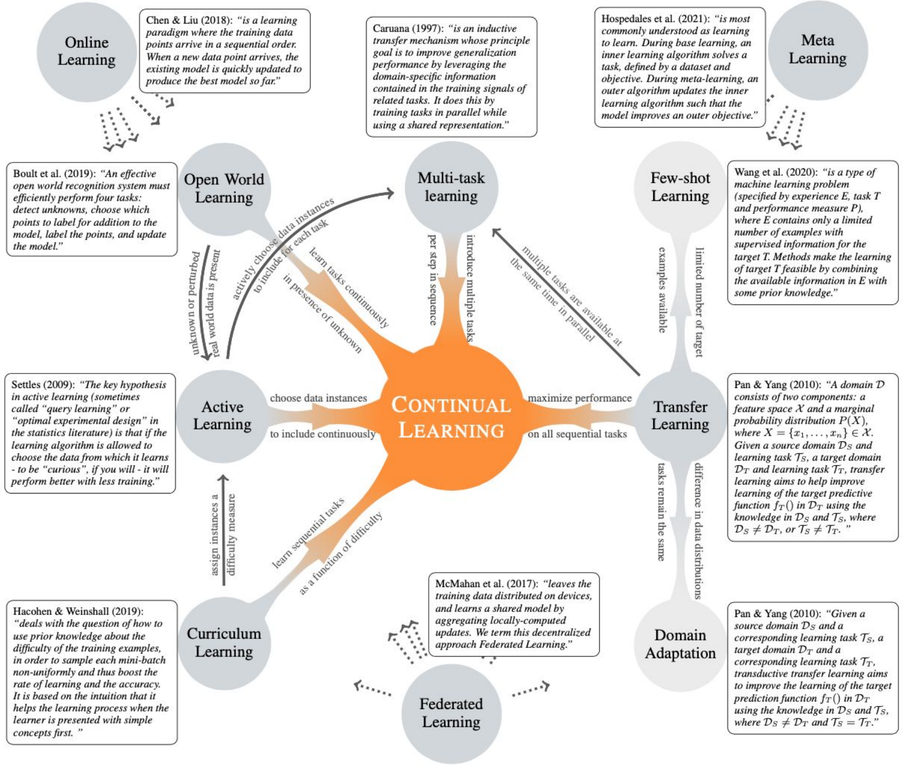
Evaluation & related paradigms

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



Chen & Liu (2018): *"is a learning paradigm where the training data points arrive in a sequential order. When a new data point arrives, the existing model is quickly updated to produce the best model so far."*

Caruana (1997): *"is an inductive transfer mechanism whose principle goal is to improve generalization performance by leveraging the domain-specific information contained in the training signals of related tasks. It does this by training tasks in parallel while using a shared representation."*

Hospedales et al. (2021): *"is most commonly understood as learning to learn. During base learning, an inner learning algorithm solves a task, defined by a dataset and objective. During meta-learning, an outer algorithm updates the inner learning algorithm such that the model improves an outer objective."*

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

Wang et al. (2020): *"is a type of machine learning problem (specified by experience E, task T and performance measure P), where E contains only a limited number of examples with supervised information for the target T. Methods make the learning of target T feasible by combining the available information in E with some prior knowledge."*

Settles (2009): *"The key hypothesis in active learning (sometimes called "query learning" or "optimal experimental design" in the statistics literature) is that if the learning algorithm is allowed to choose the data from which it learns - to be "curious", if you will - it will perform better with less training."*

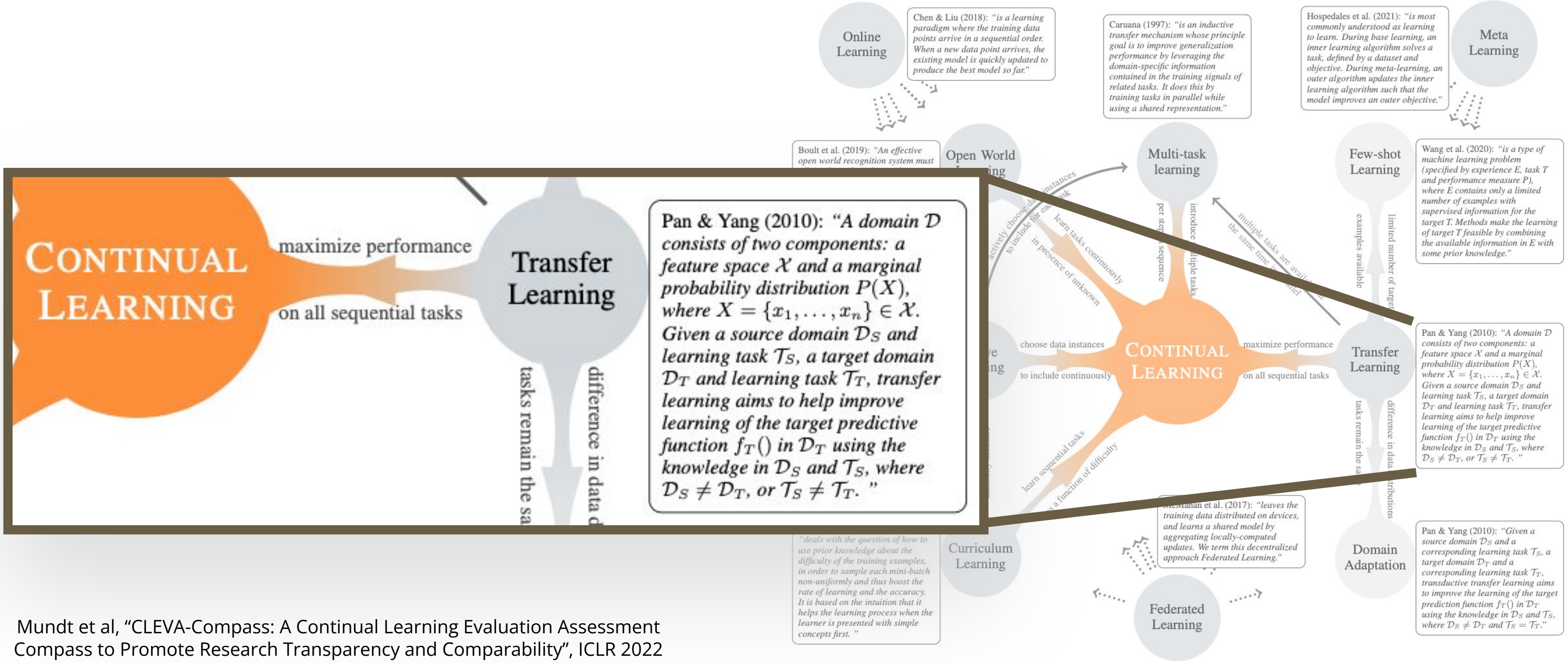
Pan & Yang (2010): *"A domain $\mathcal{D}$ consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to help improve learning of the target predictive function $f_T()$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$. "*

McMahan et al. (2017): *"leaves the training data distributed on devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning."*

Hacohen & Weinshall (2019): *"deals with the question of how to use prior knowledge about the difficulty of the training examples, in order to sample each mini-batch non-uniformly and thus boost the rate of learning and the accuracy. It is based on the intuition that it helps the learning process when the learner is presented with simple concepts first. "*

Pan & Yang (2010): *"Given a source domain $\mathcal{D}_S$ and a corresponding learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and a corresponding learning task $\mathcal{T}_T$, transductive transfer learning aims to improve the learning of the target prediction function $f_T()$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ and $\mathcal{T}_S = \mathcal{T}_T$."*

Evaluation & related paradigms



Pan & Yang (2010): *"A domain $\mathcal{D}$ consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to help improve learning of the target predictive function $f_T()$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$. "*

Chen & Liu (2018): *"is a learning paradigm where the training data points arrive in a sequential order. When a new data point arrives, the existing model is quickly updated to produce the best model so far."*

Caruana (1997): *"is an inductive transfer mechanism whose principle goal is to improve generalization performance by leveraging the domain-specific information contained in the training signals of related tasks. It does this by training tasks in parallel while using a shared representation."*

Hospedales et al. (2021): *"is most commonly understood as learning to learn. During base learning, an inner learning algorithm solves a task, defined by a dataset and objective. During meta-learning, an outer algorithm updates the inner learning algorithm such that the model improves an outer objective."*

Boult et al. (2019): *"An effective open world recognition system must*

Wang et al. (2020): *"is a type of machine learning problem (specified by experience E, task T and performance measure P), where E contains only a limited number of examples with supervised information for the target T. Methods make the learning of target T feasible by combining the available information in E with some prior knowledge."*

Pan & Yang (2010): *"A domain $\mathcal{D}$ consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to help improve learning of the target predictive function $f_T()$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$. "*

McMahan et al. (2017): *"leaves the training data distributed on devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning."*

"deals with the question of how to use prior knowledge about the difficulty of the training examples, in order to sample each mini-batch non-uniformly and thus boost the rate of learning and the accuracy. It is based on the intuition that it helps the learning process when the learner is presented with simple concepts first. "

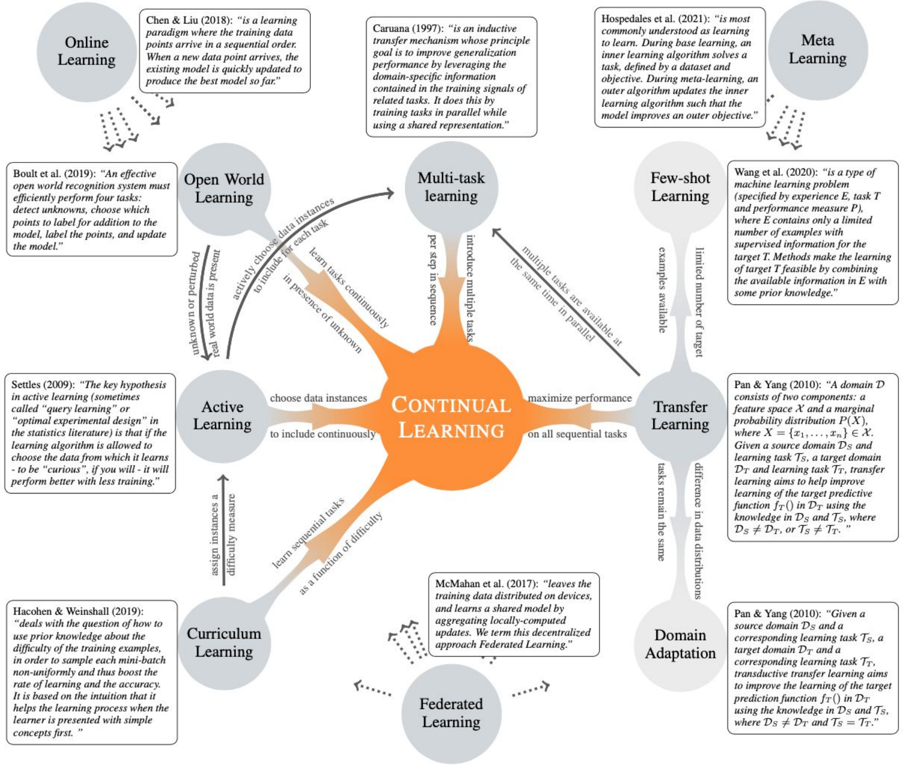
Pan & Yang (2010): *"Given a source domain $\mathcal{D}_S$ and a corresponding learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and a corresponding learning task $\mathcal{T}_T$, transductive transfer learning aims to improve the learning of the target prediction function $f_T()$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ and $\mathcal{T}_S = \mathcal{T}_T$."*

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!



Chen & Liu (2018): *“is a learning paradigm where the training data points arrive in a sequential order. When a new data point arrives, the existing model is quickly updated to produce the best model so far.”*

Caruana (1997): *“is an inductive transfer mechanism whose principle goal is to improve generalization performance by leveraging the domain-specific information contained in the training signals of related tasks. It does this by training tasks in parallel while using a shared representation.”*

Hospedales et al. (2021): *“is most commonly understood as learning to learn. During base learning, an inner learning algorithm solves a task, defined by a dataset and objective. During meta-learning, an outer algorithm updates the inner learning algorithm such that the model improves an outer objective.”*

Boult et al. (2019): *“An effective open world recognition system must efficiently perform four tasks: detect unknowns, choose which points to label for addition to the model, label the points, and update the model.”*

Wang et al. (2020): *“is a type of machine learning problem (specified by experience E, task T and performance measure P), where E contains only a limited number of examples with supervised information for the target T. Methods make the learning of target T feasible by combining the available information in E with some prior knowledge.”*

Settles (2009): *“The key hypothesis in active learning (sometimes called “query learning” or “optimal experimental design” in the statistics literature) is that if the learning algorithm is allowed to choose the data from which it learns - to be “curious”, if you will - it will perform better with less training.”*

Pan & Yang (2010): *“A domain $\mathcal{D}$ consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning aims to help improve learning of the target predictive function $f_T()$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$, or $\mathcal{T}_S \neq \mathcal{T}_T$. ”*

McMahan et al. (2017): *“leaves the training data distributed on devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning.”*

Hacohen & Weinshall (2019): *“deals with the question of how to use prior knowledge about the difficulty of the training examples, in order to sample each mini-batch non-uniformly and thus boost the rate of learning and the accuracy. It is based on the intuition that it helps the learning process when the learner is presented with simple concepts first. ”*

Pan & Yang (2010): *“Given a source domain $\mathcal{D}_S$ and a corresponding learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and a corresponding learning task $\mathcal{T}_T$, transductive transfer learning aims to improve the learning of the target prediction function $f_T()$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ and $\mathcal{T}_S = \mathcal{T}_T$.”*

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

Evaluation & related paradigms

But perhaps, we can also gain a more *thorough understanding*?

- Define & react to catastrophic forgetting & knowledge transfer in learning causal models?
- Understanding effective ways for causal tools to help interpret continual learning systems & distribution shifts?
- Develop next generation benchmarks beyond repurposing sequences of existing dataset?

Continual Causality

2023 AAAI Continual Causality Bridge Tutorial:

Thank You!

Keiland Cooper

UCI University of California, Irvine

ContinualAI

NSF

Center for the Neurobiology of Learning and Memory

UCI Data Science Initiative

Martin Mundt

OWLL

ContinualAI

TECHNISCHE UNIVERSITÄT DARMSTADT

hessian.AI