Continual Learning with Real-World Impact: Beyond Catastrophic Forgetting

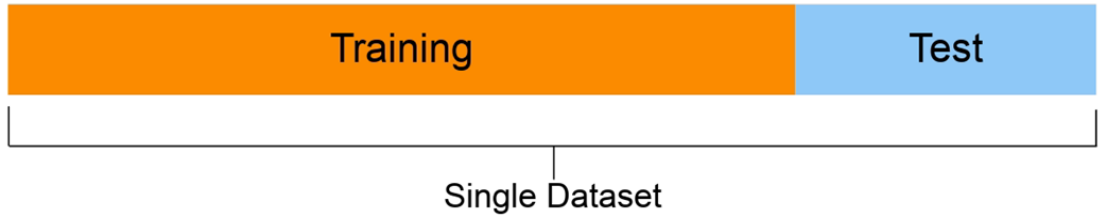
Christopher Kanan

Associate Professor, University of Rochester

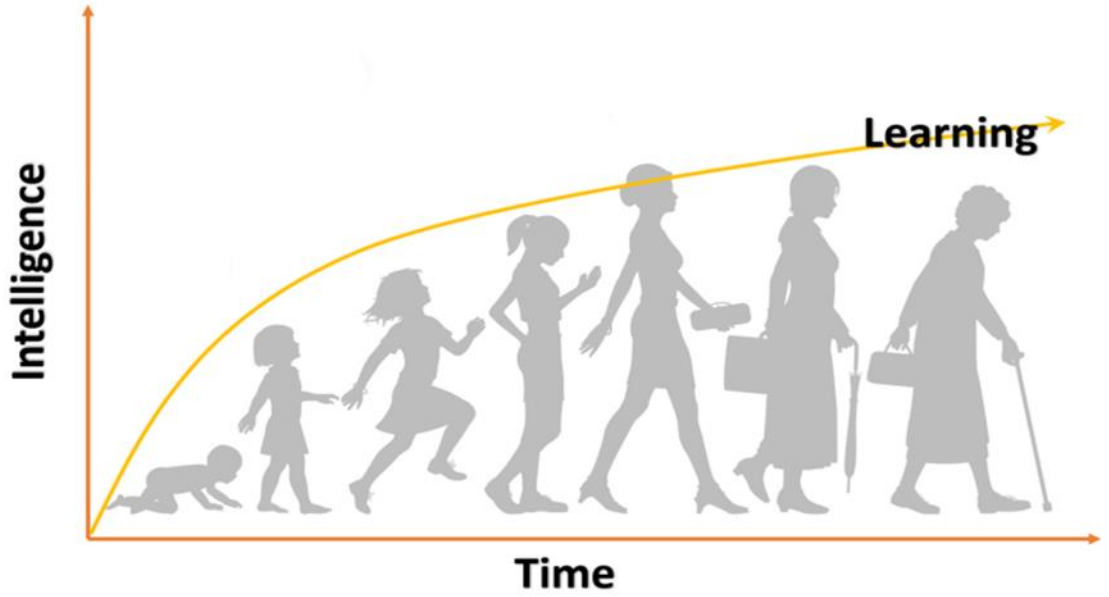
https://chriskanan.com

Why Do I Work on Continual Learning?

- Humans and animals do not have fixed train and test sets.
- We learn over the course of a lifetime.
  - We are continual learners.
- An AGI would need to learn more like us:
  - Incrementally
  - Efficiently



VS.

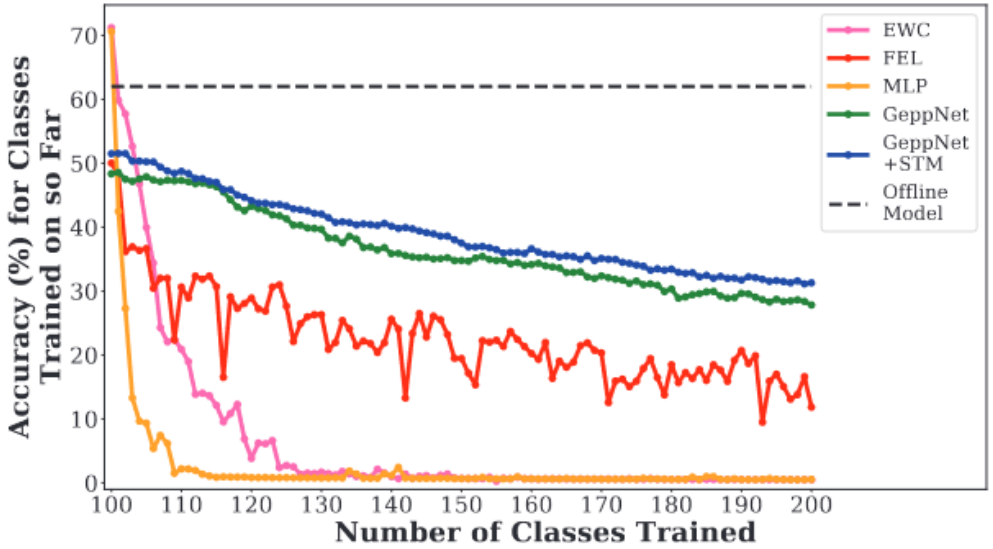


Continual Learning has Focused on Mitigating Catastrophic Forgetting

- In deep learning, we normally shuffle our training data to simulate making it independent and identically distributed (iid), which is the assumption that backpropagation makes.

- But in the real world, we often have temporally correlated inputs and outputs.
  - My experiences are not shuffled constantly. They are temporally contiguous.

- If we violate the iid assumption, we get catastrophic forgetting.
  - It is especially severe for class-incremental learning.



Kemker, McClure, Abitino, Hayes, & Kanan
AAAI-2018

Outline

1. What's continual learning good for?
2. What capabilities does a continual learning system need?
3. How might causal learning advance continual learning?

What’s Continual Learning Good For?

Efficiently Updating Very Large Production AI Systems

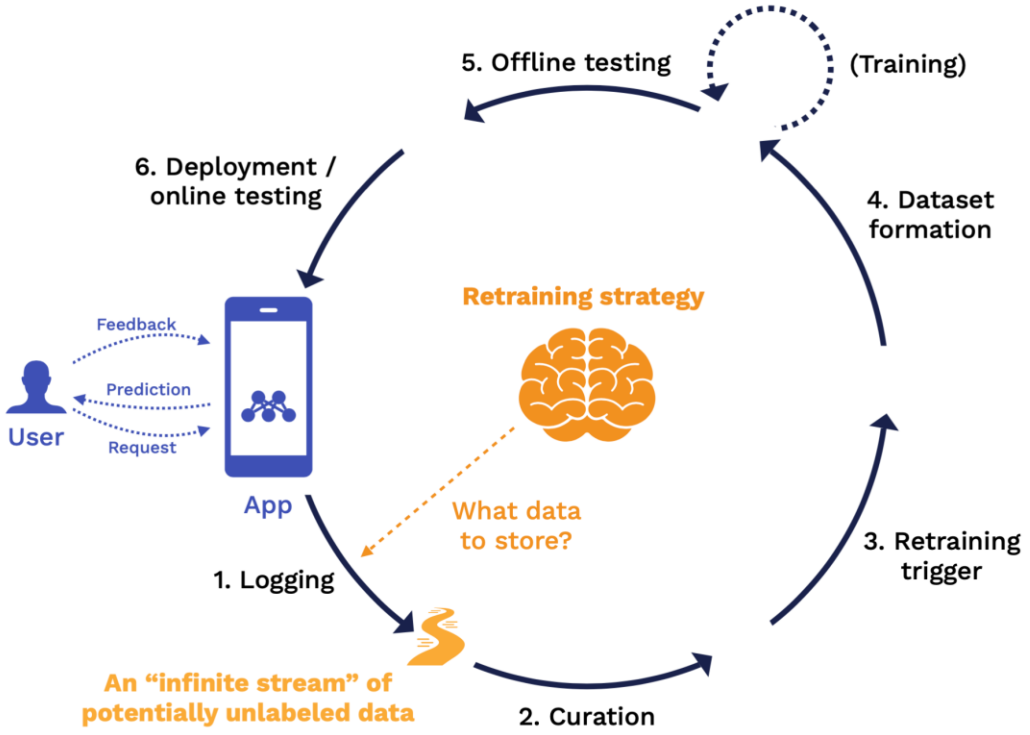
WILL KNIGHT BUSINESS 01.21.2020 07:00 AM

AI Can Do Great Things—if It Doesn't Burn the Planet

The computing power required for AI landmarks, such as recognizing images and defeating humans at Go, increased 300,000-fold from 2012 to 2018.

One algorithm that lets a robot manipulate a Rubik's Cube used as much energy as 3 nuclear plants produce in an hour. PHOTOGRAPH: GETTY IMAGES

Efficiently Updating Very Large Production AI Systems



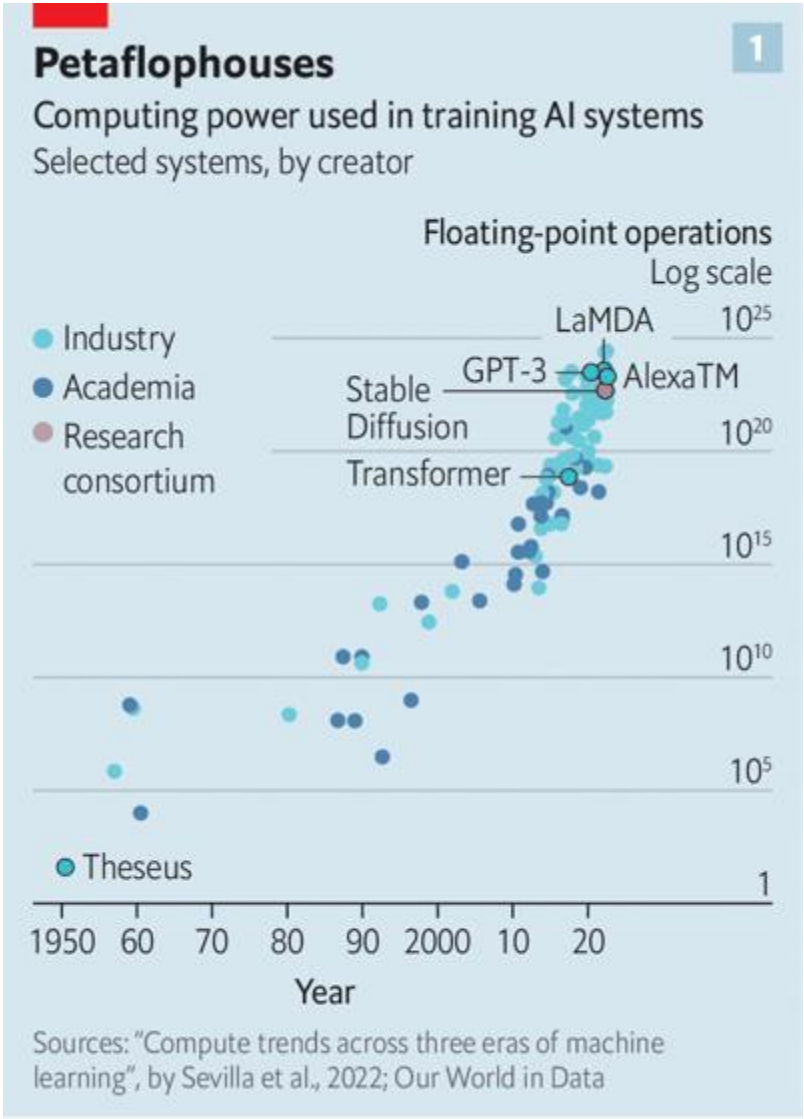
- In production AI, giant neural network systems are often periodically retrained from scratch.
- Continual learning could enable these systems to be continually updated with new data, rather than just updated periodic jobs 1-2 times per month.

Image Credit: Full Stack Deep Learning

Continual Learning and Production AI

- Can just re-train from scratch as a job or can easily employ cumulative replay for continual learning.
- Cumulative replay:
  - Get new examples, add them to the database.
  - Progressively fine-tune the network! Many ways work well.
    - Simplest approach is to mix new samples with some randomly selected examples from the database and do some iterations of backprop.
    - May have to deal with the warm-start problem with fine-tuning to get out of local optima.

We showed that on average over 99% of the offline learner's performance can be achieved using cumulative replay (Hayes et al., ICRA 2019)



Continual Learning Can Help Production AI But it (Probably) Isn't Essential

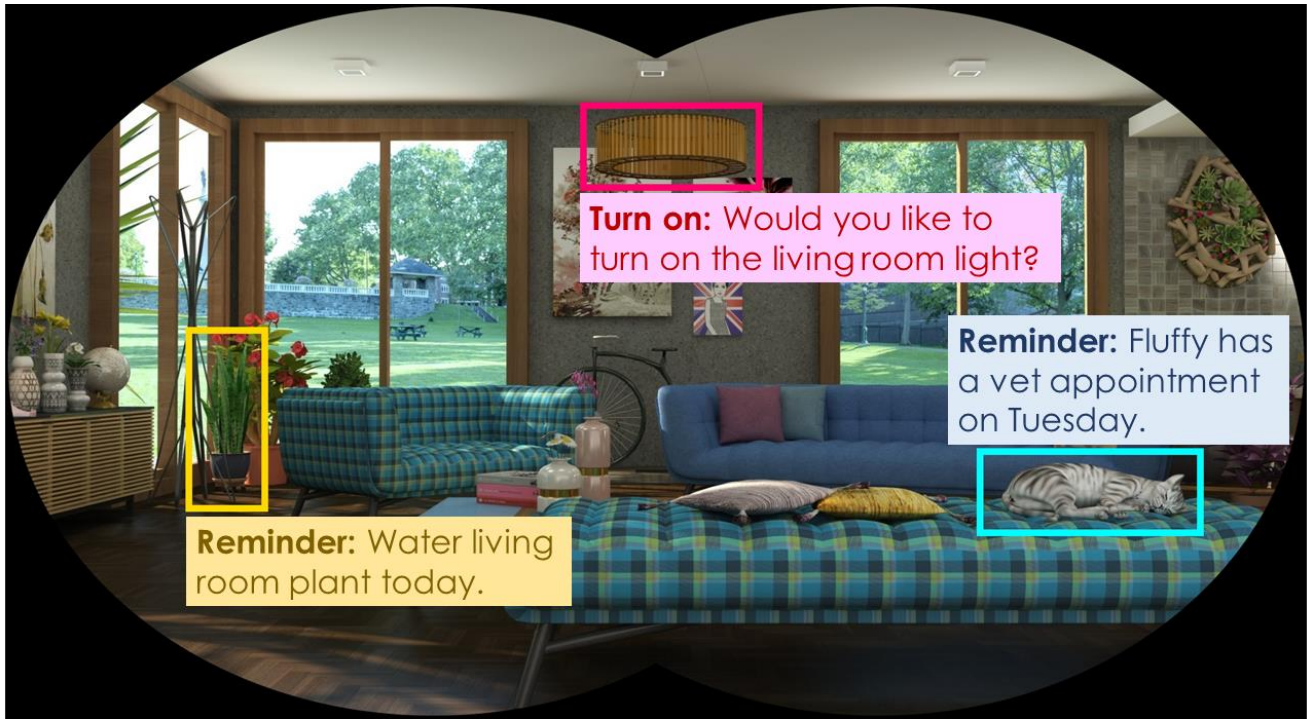
- Big companies do not need continual learning because they could just re-train from scratch or use cumulative replay.
- But continual learning could lead to huge savings in power and reduce costs!
  - Could we update LLMs with recent information?
- Questions for Continual Learning:
  - Are we focusing on efficient learning? This is rarely measured. **Minimize neural network updates.**
  - We cannot assume the data is non-iid, just that we have more data. Are our algorithms too rigid?

Meta's new AI Research SuperCluster has 760 NVIDIA DGX A100s (6080 GPUs).

Their existing AI cluster has 22,000 NVIDIA V100 GPUs.

On-Device Continual Learning

On-Device Continual Learning



- Less compute means can learn on-device.
- No cloud computing in space or in Internet deprived locations.
- Customized for each user without sending personal information through the web.
- On-device learning for AR/VR, Smart toys, robots, phones, and more.

MOTHERBOARD
TECH BY VICE

How This Internet of Things Stuffed Animal Can Be Remotely Turned Into a Spy Device

More bad news for toymaker Spiral Toys, which left customer data from its "CloudPets" brand exposed online.

By Lorenzo Franceschi-Bicchierai

February 28, 2017, 12:19pm Share Tweet Snap

PAUL STONE/YOUTUBE

IMAGE: CLOUDPETS

MOTHERBOARD
TECH BY VICE

Internet of Things Teddy Bear Leaked 2 Million Parent and Kids Message Recordings

A company that sells “smart” teddy bears leaked 800,000 user account credentials—and then hackers locked it and held it for ransom.

By Lorenzo Franceschi-Bicchierai

What Properties Do We Need for On-Device Continual Learning?

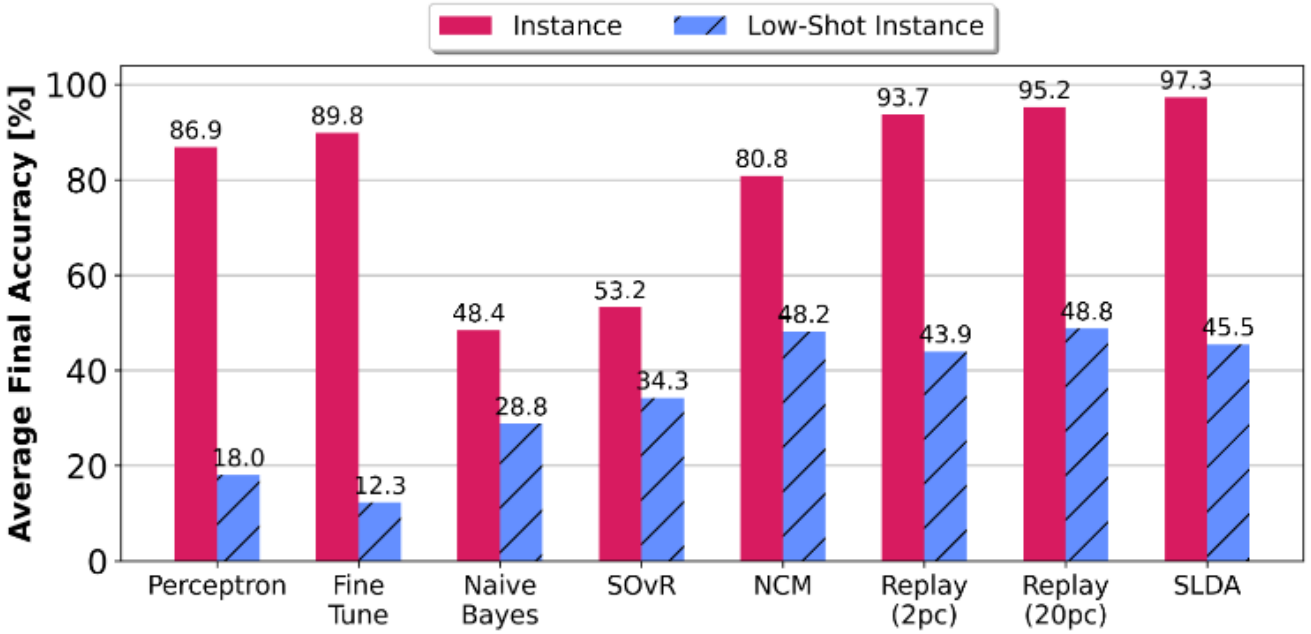
- Being able to update neural networks targeted for embedded devices.
  - Would a huge vision transformer be appropriate?
- Updating neural networks without much memory or storage.
  - Many existing continual learning systems require far more memory for replay than would be permissible.
- Effective generalization from only a few samples (low-shot learning).
  - Are we measuring this?

Online Continual Learning for Embedded Devices

Tyler L. Hayes[1], Christopher Kanan[1,2]

Published in CoLLAs (2022)

- Compared continual learning methods suitable for on-device learning across multiple data orderings: iid, class incremental, etc.
- SLDA (Hayes & Kanan, 2020) worked best, but low-shot learning has a long way to go.
- Demonstrated that mobile networks outperform popular architectures used in continual learning research, e.g., ResNet18 << MobileNetV3.



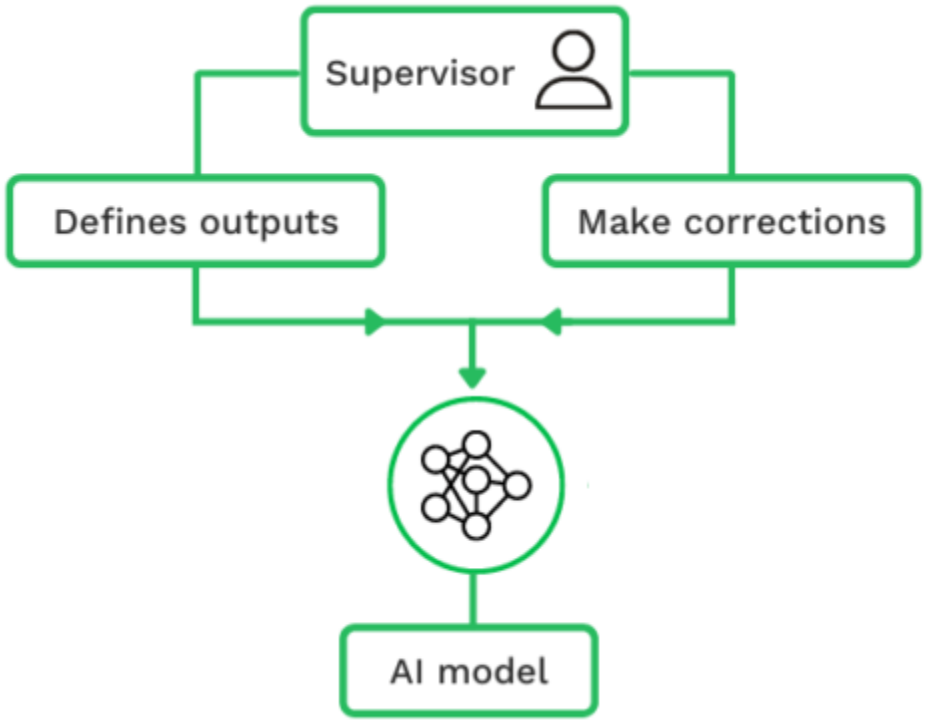
Low-shot learning has a long way to go.

Enabling AI Systems that Require Continual Learning

Interactive learning

- AI systems interacting with humans often need to be updated:
  - Correct the AI's mistakes
  - Update the AI with new information

- For production AI systems, these updates do not need to be immediate, but for embedded devices and personal assistants, users would expect the model to immediately correct itself.
  - This requires continual learning.



Open World Continual Learning

Open-world Experience

- Continuously discover new object categories.
- Ability to label objects as known vs unknown.
- Ability to discover new sub-categories of known object types.
- Closely related area: Autonomous, Never-Ending Learning

Learning in Multimodal Open-world

How are new labels obtained?
- Active learning / human query
- Automatic object discovery

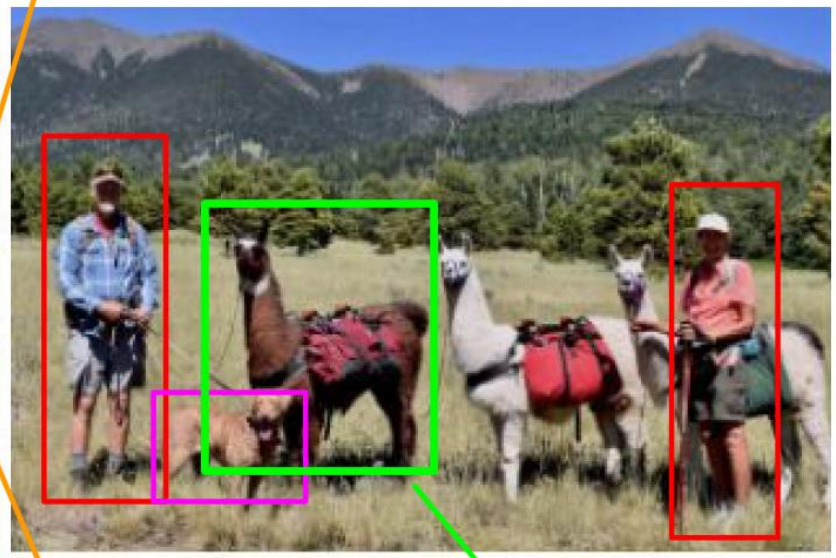
What are the people doing?

I think they are hiking with dogs and X.

Actually, X are Llamas?

Thinking ....

Show me the black Llamas.



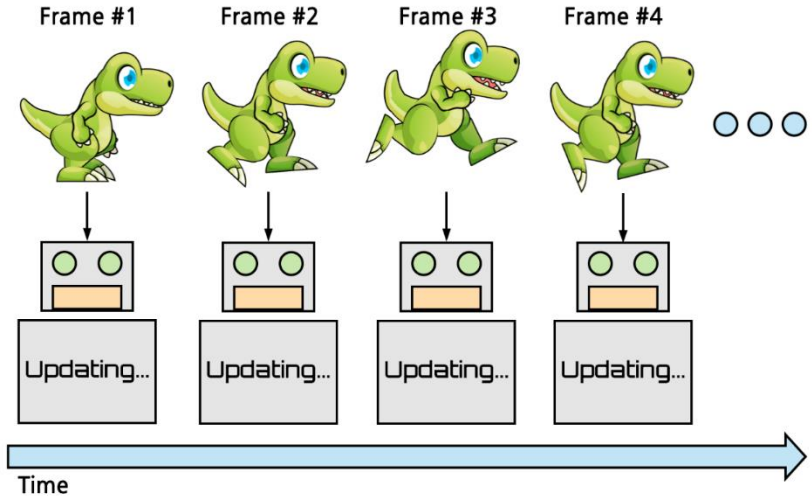
Quick learning allows to adapt to new information

Certificate in the form of bounding boxes

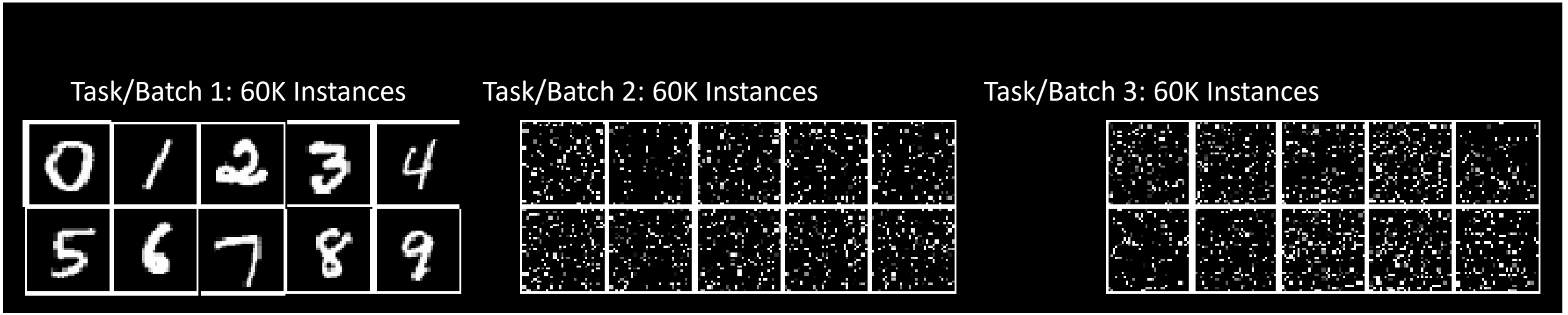
Most Continual Learning Work Has Limited Relationship to These Applications

What Would an Ideal Continual Learner Look Like?

- A continual learning algorithm is capable of incrementally learning from a data-stream without assuming the stream is sampled iid.
  - Should work effectively regardless of how the data is sampled, whether it is iid or extremely non-iid, e.g., class incremental learning.
- Ideally a continual learner:
  - Learns online without requiring large batches
  - Is sample, computationally, and memory efficient.
  - No task-labels or auxiliary information.
  - Should scale to real-world datasets
- Do most algorithms have these attributes?
  - Are they rigorously evaluated?



Incremental Learning of Permuted MNIST Batches



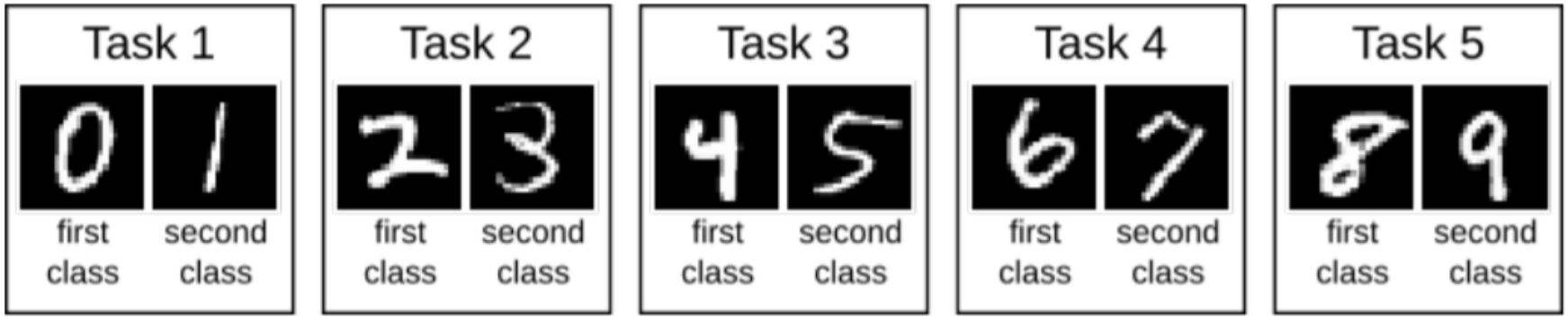
- Learn the original 60K MNIST images in a batch.
- Apply a random permutation matrix to the next batch of 60K to create the next task.
- Classifier typically has the "task" associated with each permutation at test time.
- Very popular paradigm! Hundreds of papers in the last year!!!

Meets none of our goals:

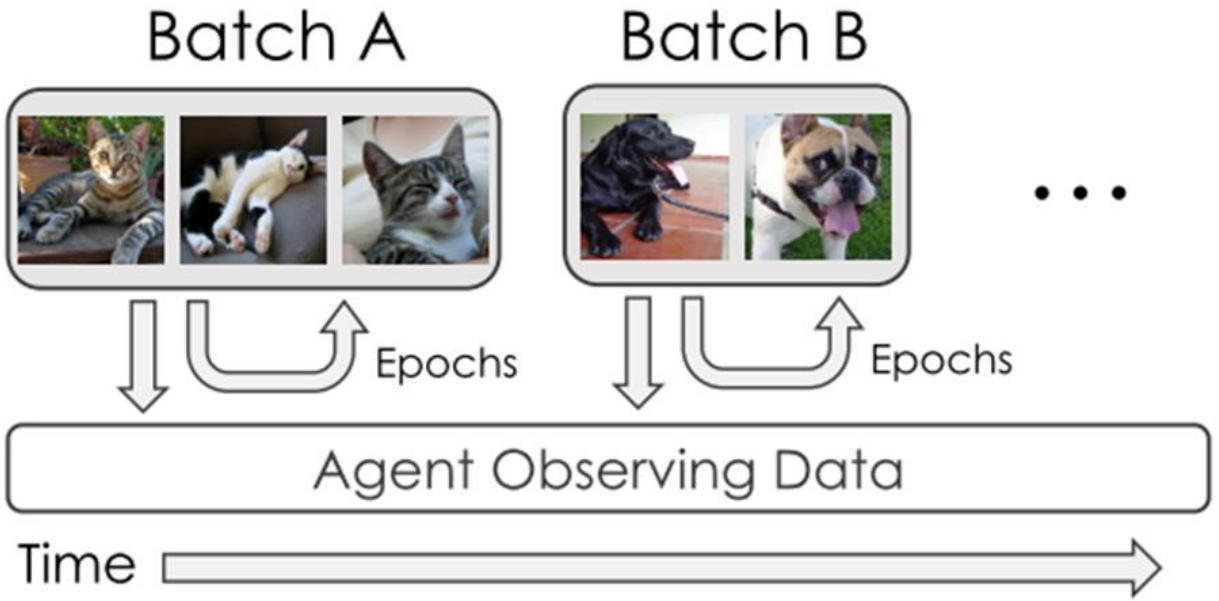
- Highly unnatural. Nothing like animal learning. Not applicable to our applications.
- Algorithms tested only on these problems often do not scale.
- Systems that work well for this fail in other more useful paradigms (Kemker et al., AAAI 2018)

Incremental Batch Learning with Task Labels Without Permutations

- Incremental batch learning with task labels at test time. Often studied with MNIST.



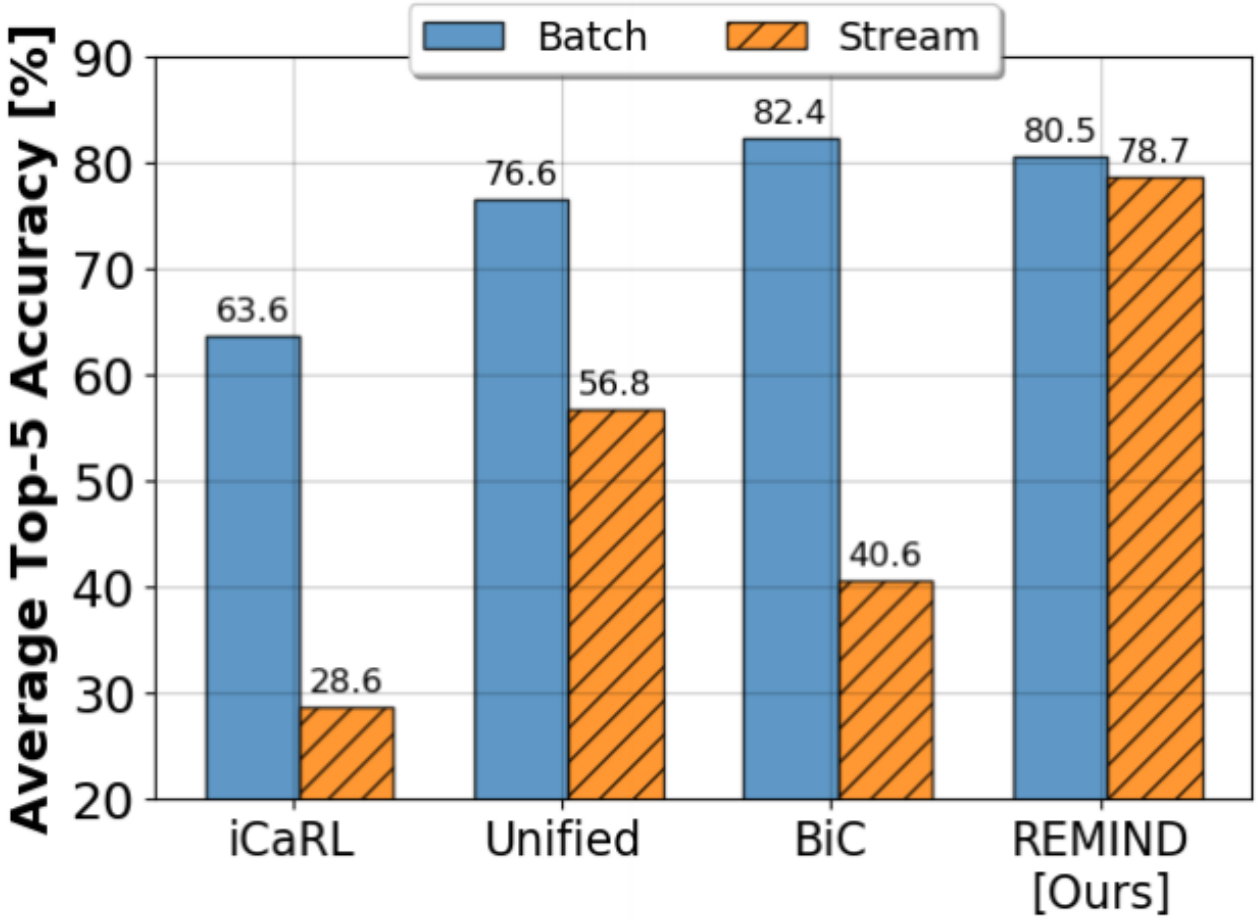
- **Problems:**
- Must know the task label during deployment. Often not available for real-world applications.
- Often tasks are binary.
- Little applicability to real-world applications.
- Heavily studied using MNIST, CIFAR-100, TinyImageNet, etc..
- Algorithms that work well on MNIST is not strong evidence that they will scale up to large datasets with natural images.



(a) Incremental Batch Learning

Incremental Batch Learning – Learn incrementally from batches of *N* examples, where each batch is only seen once, without catastrophic forgetting.

- For supervised learning, most use "continual learning" to mean this.
- Common scenarios: Learn ImageNet in batches of *N*=100000 where each batch has 100 classes not seen later. Unbounded resources during a batch.
- Popular setup in computer vision with ImageNet-1K.
  - Much closer real-world applications than MNIST.



REMIND Your Neural Network to Prevent Catastrophic Forgetting

Tyler L. Hayes[1,*][0000−0002−0875−7994], Kushal Kafle[2,*][0000−0002−0847−7861], Robik Shrestha[1,*][0000−0002−0945−3458], Manoj Acharya[1][0000−0003−0223−3556], and Christopher Kanan[1,3,4][0000−0002−6412−995X]

- Most systems fail when using batches of 50 samples rather than 100K samples during incremental learning of ImageNet.
- Most systems cannot revisit. They only work in the incremental class learning edge case!
- Small batches or learning online is critical for many applications.

	ImageNet	CORe50			
Model	CLS IID	IID	CLS IID	INST	CLS INST
Fine-Tune (θ_F)	0.288	0.961	0.334	0.851	0.334
ExStream	0.569	0.953	0.873	0.933	0.854
SLDA	0.752	0.976	0.958	0.963	0.959
iCaRL	0.306	-	0.690	-	0.644
Unified	0.614	-	0.510	-	0.527
BiC	0.440	-	0.410	-	0.415
REMIND	**0.855**	**0.985**	**0.978**	**0.980**	**0.979**
Offline (θ_F)	0.929	0.989	0.984	0.985	0.985
Offline	1.000	1.000	1.000	1.000	1.000

Class instance – See all instances of an object in a temporal sequence, and each class is seen in order.

- REMIND works well despite the order of the data.
- Performance is virtually identical if iid, sorted by class, or ordered by instances in a video stream.

REMIND Compared to Methods that Use Task Labels

- We compared REMIND to regularization methods that use task labels on CORe50.
- REMIND achieves the best results regardless of whether task labels are allowed.
- By using task-labels, it means that these methods know which "task" the current input belongs to.
  - Task 1: "The example is either class 1 or 2."
  - Task 2: "The example is either class 3 or 4"
  - Etc.

	CLS IID		CLS INST	
MODEL	TL		TL	
SI	0.895		0.905	
EWC	0.893		0.903	
MAS	0.897		0.905	
RWALK	0.903		0.912	
A-GEM	0.925		0.916	
REMIND	**0.995**		**0.995**	
Offline	1.000		1.000	

Continual Learning & Goodhart's Law

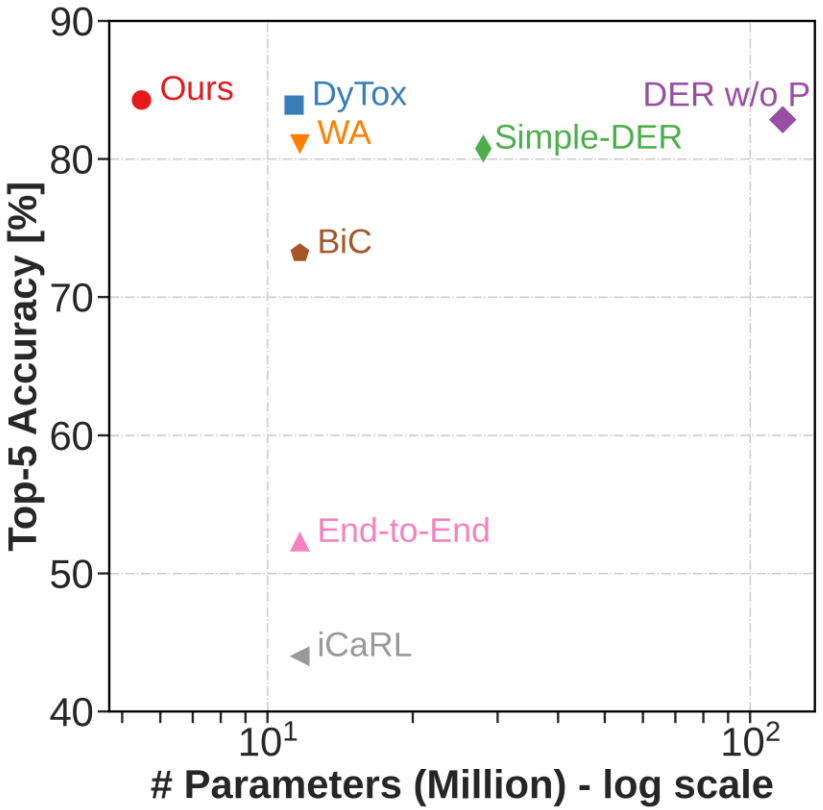
Goodhart's Law: When a measure becomes a target it ceases to be a good measure.

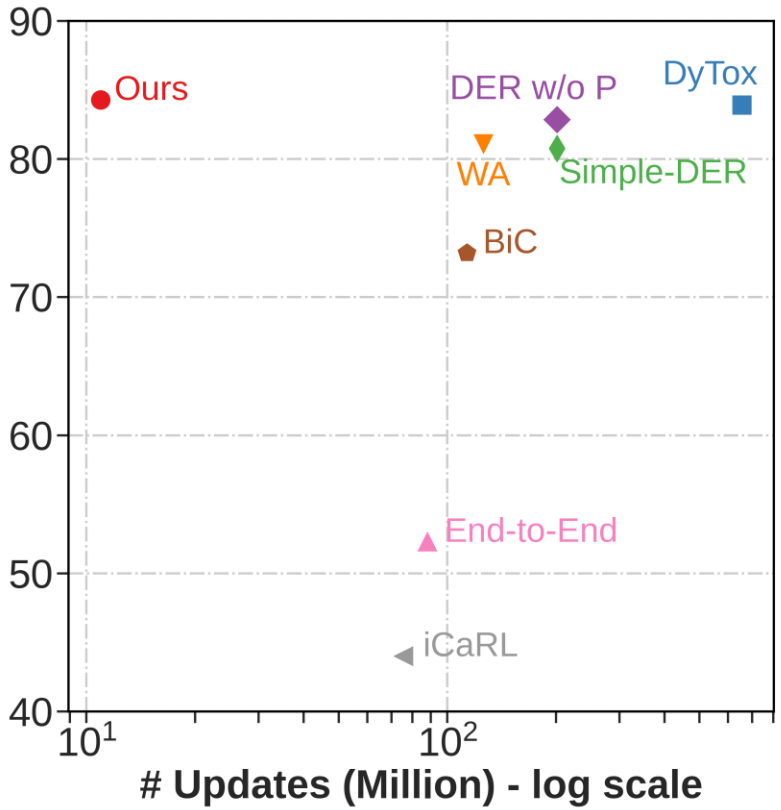
- Designed for the test, rather than designed to solve the continual learning problem.
- Systems aren't designed for real-world applications.
- Most systems are only evaluated or only function on edge cases: incremental class learning.
- Most systems have constraints that make them useless for these real-world applications, e.g., very large batches, huge replay buffers, ensembling many large models, using enormous amounts of memory.

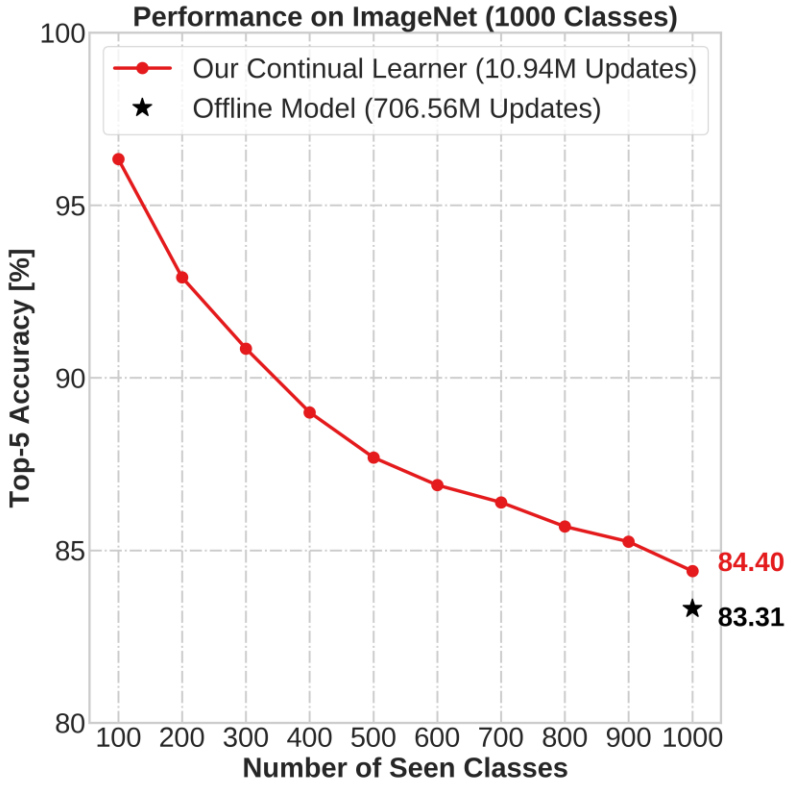
Continual Learning Needs More Rigor

- Systems need to be targeted at one or more real-world applications.
- Systems need to be tested on large-scale datasets and shown to scale regardless of “batch” size.
- Systems should be capable of performing well regardless of data ordering.
  - A system only capable of class incremental learning without revisiting has no real-world utility.
- We need established gauntlets and suites of tests for evaluating continual learning performance.

Catastrophic Forgetting is Largely Solved







- Continually learn 900 classes from ImageNet. First 100 classes used for initialization.
- **Learns 900 classes in only 4 hours on one GPU.** Focus is on efficient learning!
  - Comparison continual learning methods take 1+ days to train on same hardware!
- Works equally well regardless of data ordering. **Same performance as offline system!**
- Low parameter count. Can target on-device applications.
- Paper in preparation (results subject to improve). Probably on arXiv in late March.

Problems Beyond Catastrophic Forgetting & How Causal Machine Learning Can Help

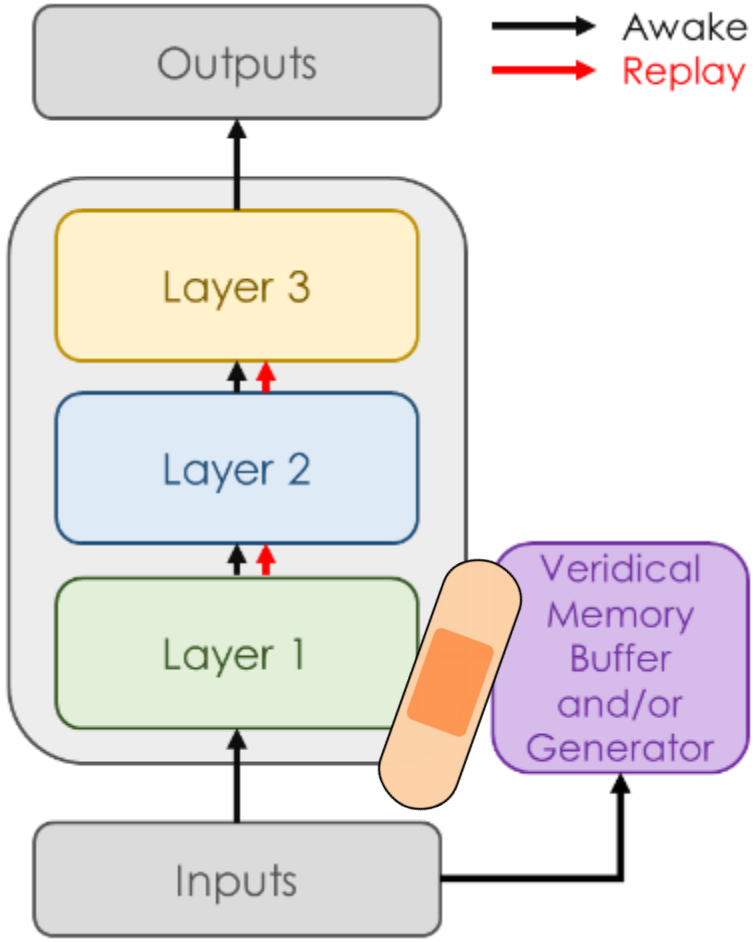
Problems Beyond Catastrophic Forgetting

1. Efficient learning
   - Minimizing neural network updates
   - Learning the next task more efficiently than the first task
2. Out-of-domain generalization and distribution shifts
3. Learning to overcome dataset bias

- **To tackle these problems we need algorithms that do not make the iid assumption.**
  - **Backprop alone is unlikely to suffice for solving these problems.**

We Won't Be Able to Tackle These Problems with Band Aids

- Catastrophic forgetting happens due to the data being iid and learning with algorithms that make this assumption.
- Continual learning methods are largely band aids:
  - How do we tweak learning so that the non-iid data will work?
  - Replay makes the data approximately iid.
- **Best case scenario**: We match an offline learner where the data is iid. It would work well given lots of data.
  - System won't forget, but we can't tackle problems beyond forgetting with band aids.



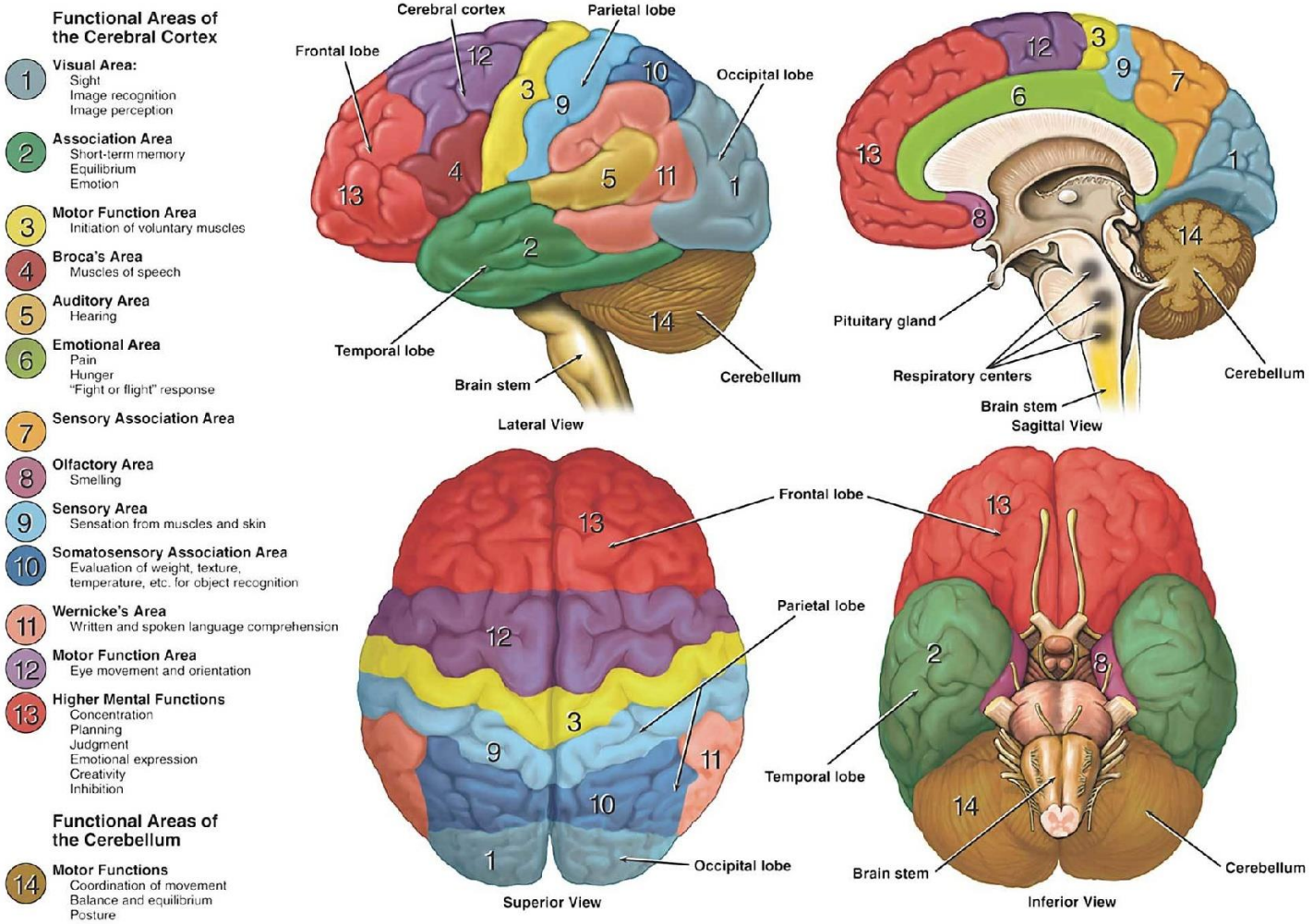
Continual Learning & Transfer Learning

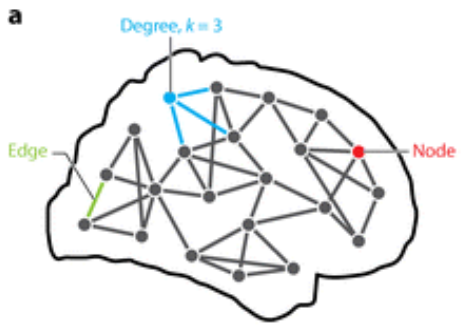
- Continual learning can be thought of as transferring from the past to the future, and vice versa.
- Existing continual learning systems are essentially progressive fine-tuners.
- The mechanism of transfer is crude:
  - Fine-tuning dense weights with various tricks to deal with non-iid.
- **Hypothesis**: Causal representation learning of factored representations could lead to significant benefits to increase learning efficiency.

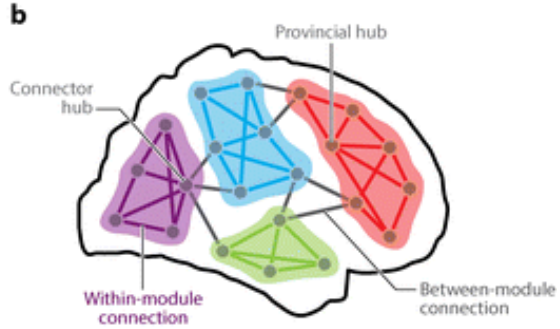
Causal Representation Learning

- Causal representation learning: neural networks that map low-level features to some high-level factorized variables (representations) supporting causal statements relevant to downstream tasks.
- For continual learning, we can think of these causal representations as being learning inductive biases over time that help with:
  - Robustness to bias
  - Efficient learning
  - Out-of-domain generalization

The Brain Has Functional Segregation and Modularity







Sporns O, Betzel RF. 2016. Annu. Rev. Psychol. 67:613–40

- The brain has hierarchical modules.
- Regions functionally close share information with the same module.

Learning Refined Abstractions Over Time

- How do we measure concept acquisition over time?
- How do we measure the ability to dynamically recombine learned concepts?
- We can at least measure whether learning is becoming more efficient over time.

Continual Learning Needs Better Evaluation Paradigms to Showcase the Benefits of Causal Learning

- Existing evaluation paradigms themselves need a massive upgrade to measure the things that matter in continual learning.
  - Incremental learning on ImageNet and other datasets won't suffice.

- We need to design experimental setups where the benefits of causal learning could be evaluated.

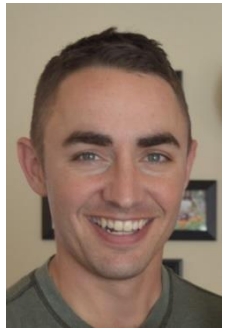
Acknowledgements: Past & Present Students and Sponsors


Dr. Kushal Kafle

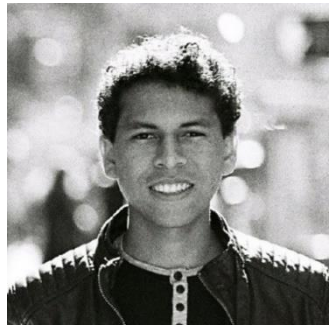

Robik Shrestha


Dr. Ron Kemker

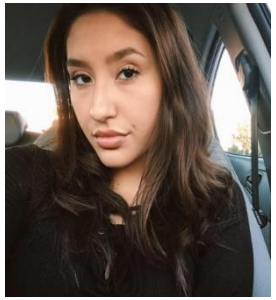

Dr. Tyler Hayes


Dr. Ryne Roady


Dr. Manoj Acharya

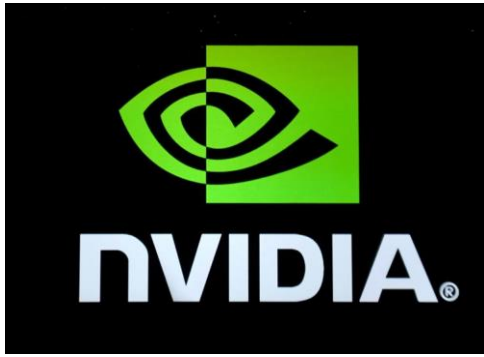

Gianmarco Callalli


Angelina Abitino


Ayesha Gonzalez


Dr. Usman Mahmood


Yousuf Harun



Thank You!

- Continual learning has many valuable real-world applications.
  - We should design algorithms for real-world applications
  - Their performance characteristics measured more rigorously.

- Toy problems are an okay place to start, but continual learning needs to grow up.

- There are much more interesting questions in continual learning than mitigating catastrophic forgetting.
  - Causal machine learning is an exciting area to mine for ideas to incorporate these abilities into continual learners.
  - We will need new evaluation setups and metrics to properly assess these capabilities.

- http://chriskanan.com