Continual Machine Learning

Summer 2024

Teacher

Dr. Martin Mundt,

Research Group on Open World Lifelong Learning

Time

Every Friday 14:25 - 16:05 CEST

Course Homepage

http://owll-lab.com/teaching/cl_lecture_24

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

Week 1: Introduction and Motivation

Course requirements

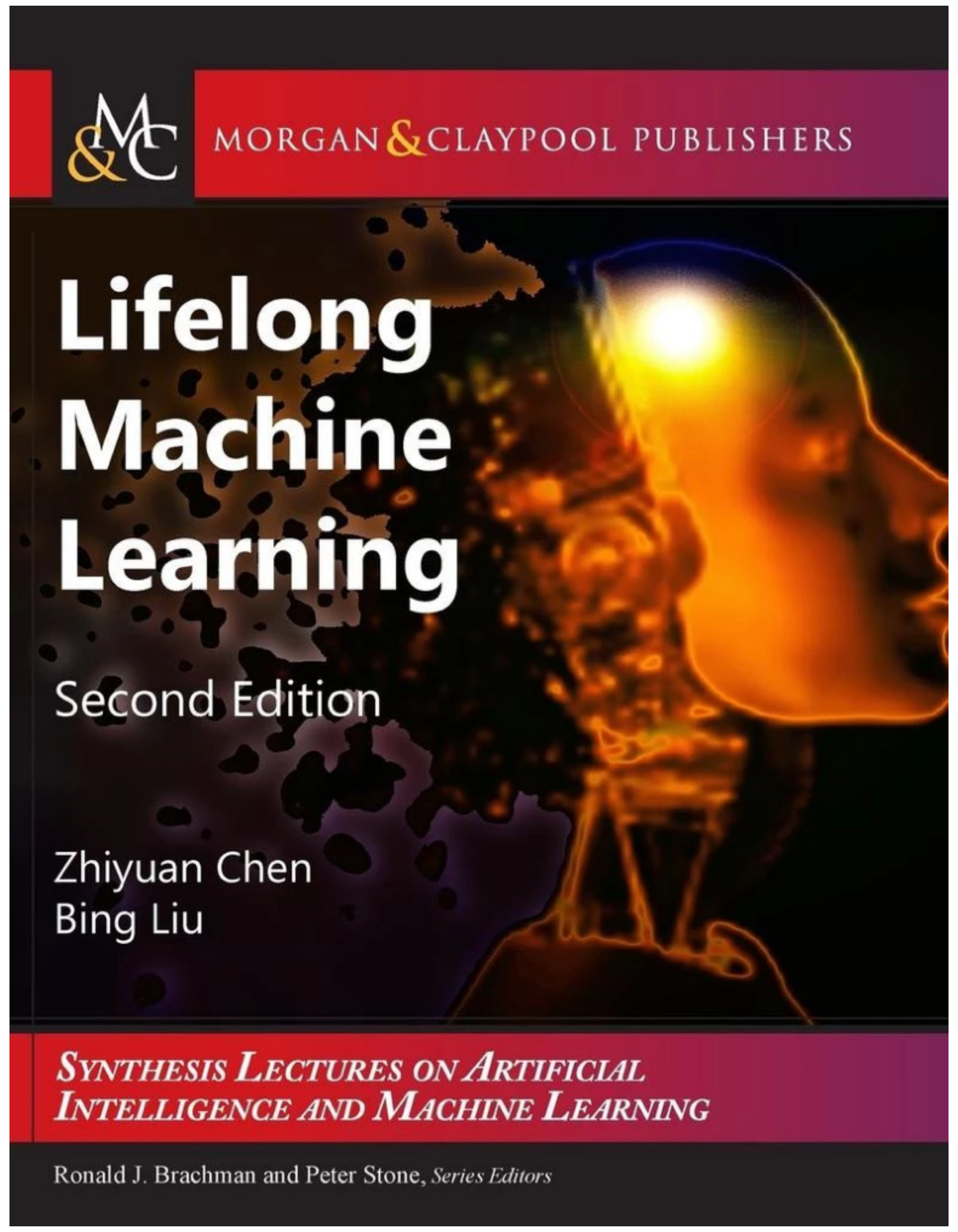
- Basic understanding of the ideas behind artificial intelligence, machine learning, deep learning

- We'll revisit some basics of models on a few slides, but just to the extent that we need to make out the difference/importance to/for continual learning

- In-depth knowledge of algorithms will be beneficial, but is not a requirement

Course materials

- Mainly the lectures, slides + linked materials
- Potentially helpful "Lifelong Machine Learning" by Chen & Liu
- Field is rapidly evolving & consolidation of works is largely still open



Motivation - what do you think: what is machine learning?

The static ML workflow

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”.

Machine Learning,
T. M. Mitchell, McGraw-Hill,1997

ML recap: train - test splits

*"The result of running the machine learning algorithm can be expressed as a **function**. The precise form of the function is determined during the **training** phase, also known as the **learning** phase, on the basis of the **training data**.*

*Once the model is trained it can then determine the identity of new images, which are said to comprise a **test set**. The ability to categorize correctly new examples that differ from those used for training us known as **generalization**".*

Pattern Recognition and Machine Learning,

C. M. Bishop, Springer 2006,

example on image classification: introduction page 2

ML recap: error/loss & learning

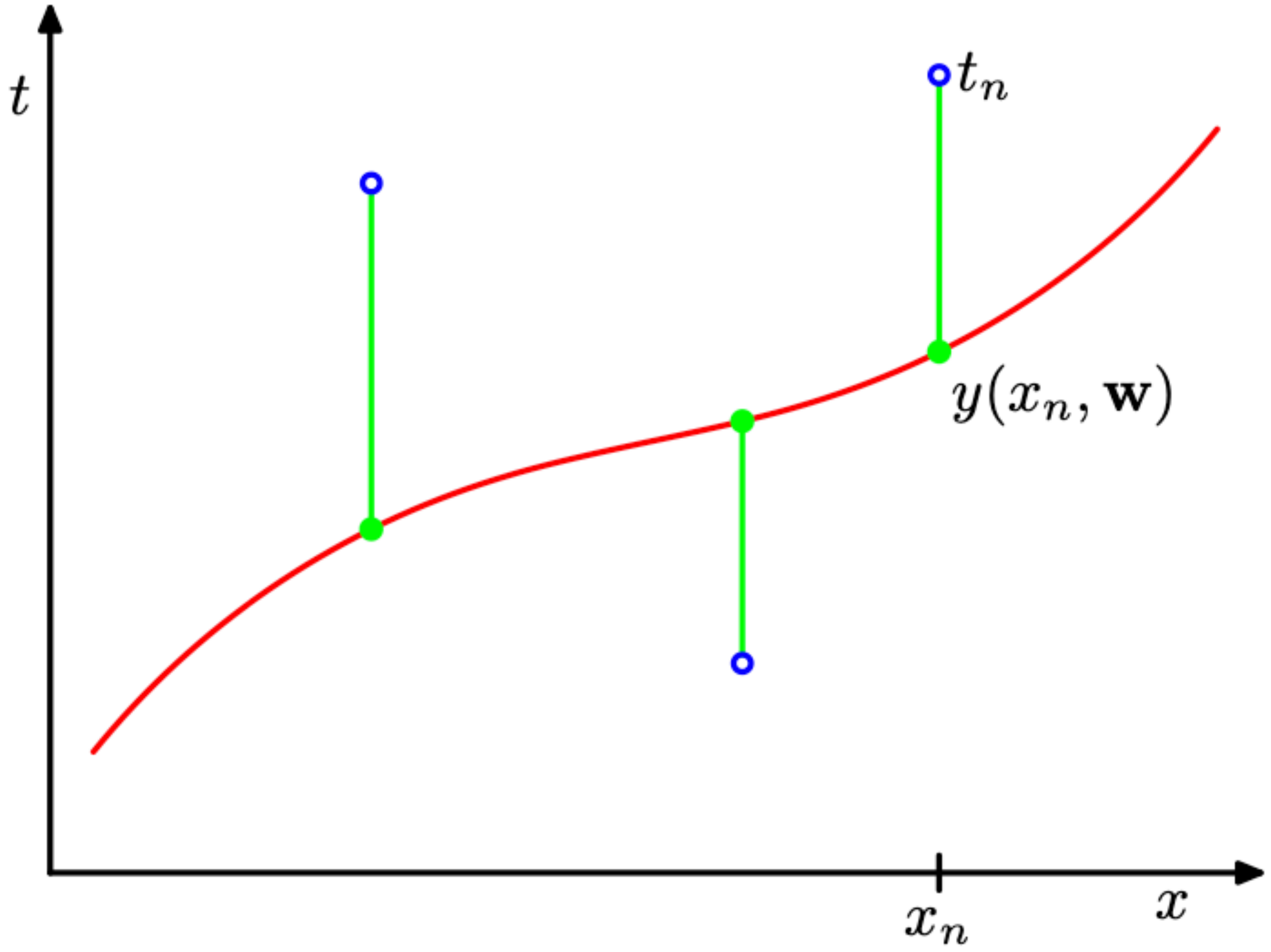


Figure 1.3 The error function (1.2) corresponds to (one half of) the sum of the squares of the displacements (shown by the vertical green bars) of each data point from the function $y(x, \mathbf{w})$.

Pattern Recognition and Machine Learning, C. M. Bishop, Springer 2006, example on polynomial curve fitting: intro page 6

ML recap: under & overfitting

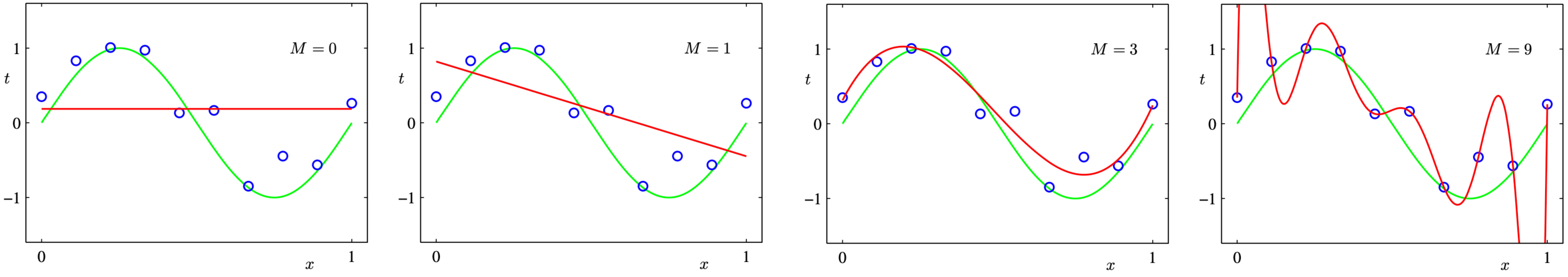


Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

Pattern Recognition and Machine Learning, C. M. Bishop, Springer 2006, example on polynomial curve fitting: introduction page 7

ML recap: under & overfitting

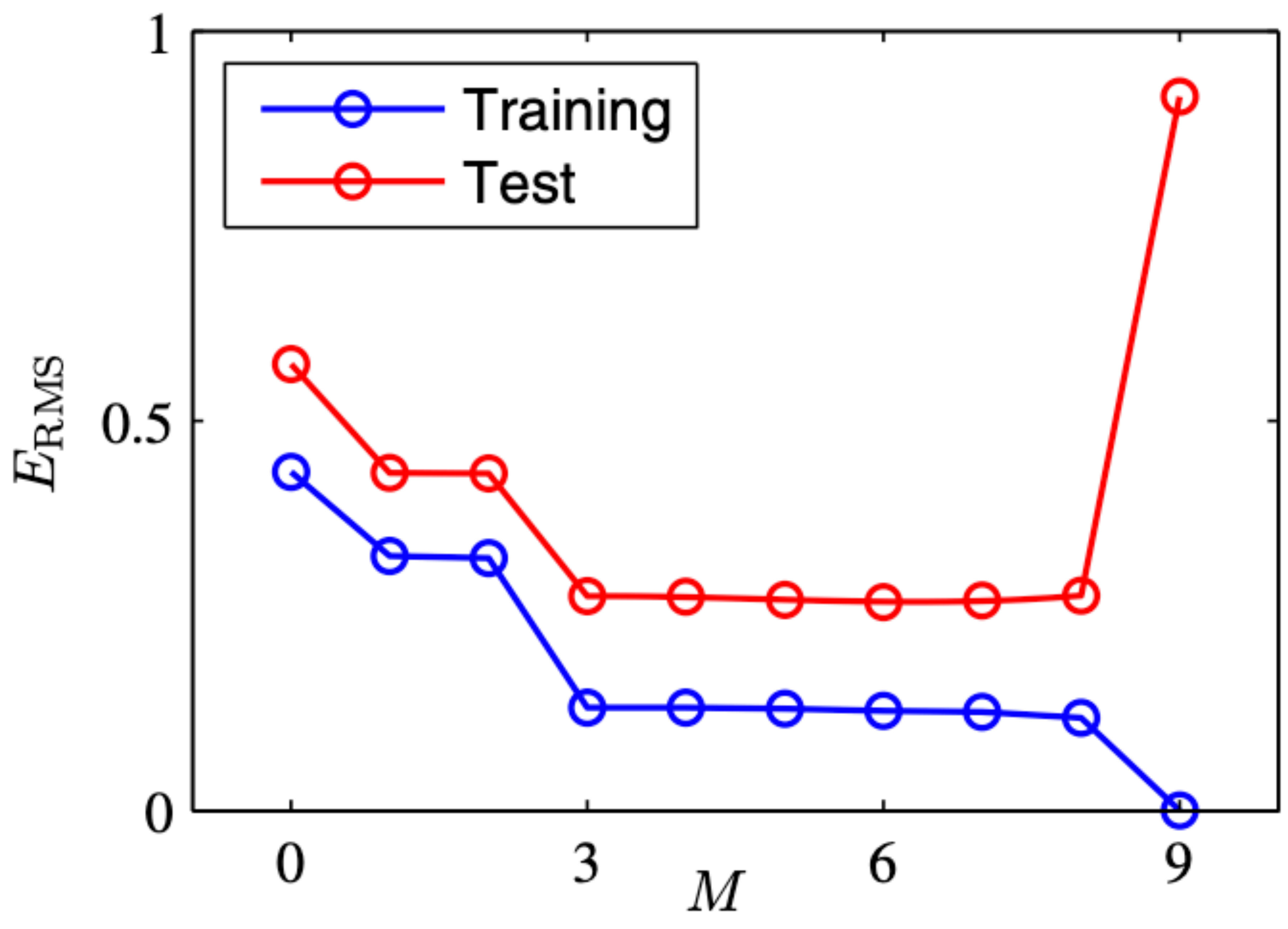


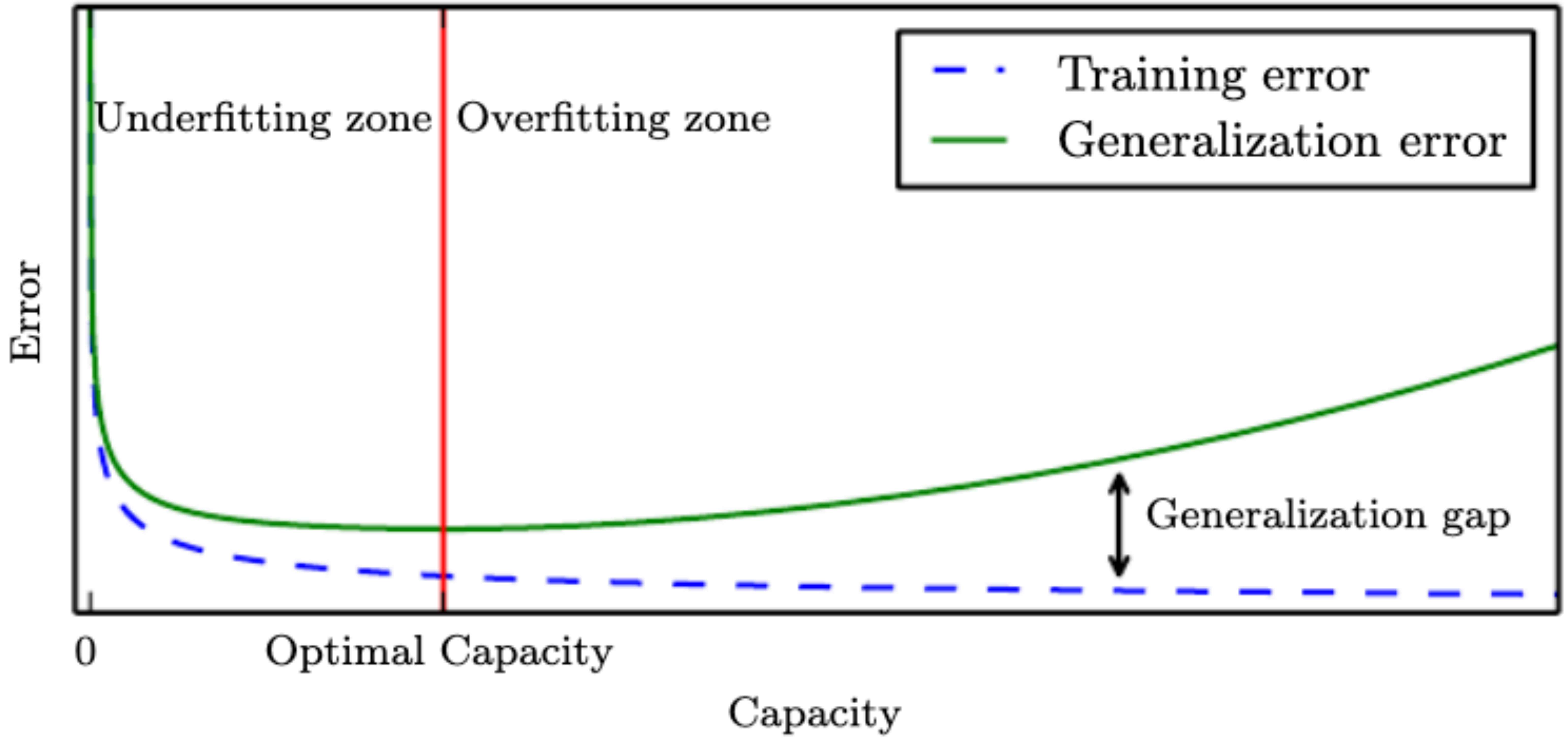
Figure 1.5 Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of M.

"Intuitively, what is happening is that the more flexible polynomials with larger values of M are becoming increasingly tuned to the random noise on the target values".

Pattern Recognition and Machine Learning, C. M. Bishop, Springer 2006, example on polynomial curve (over-)fitting in the introduction on page 8

ML recap: under & overfitting

This picture is still very much the same in the “deep learning era”



Deep Learning, Goodfellow, Bengio, Courville, MIT Press 2016,

Machine Learning Basics chapter, page 112.

What do you think are the goals of ML?

The static ML workflow: goals

"Of course, when we use a machine learning algorithm, we ***do not fix the parameters ahead of time****, then sample both datasets. We* ***sample the training set****,* ***then*** *use it to* ***choose the parameters*** *to reduce training set error,* ***then sample the test set****.*

The factors determining how well a machine learning algorithm will perform are its ability to:

1. Make the training error small.

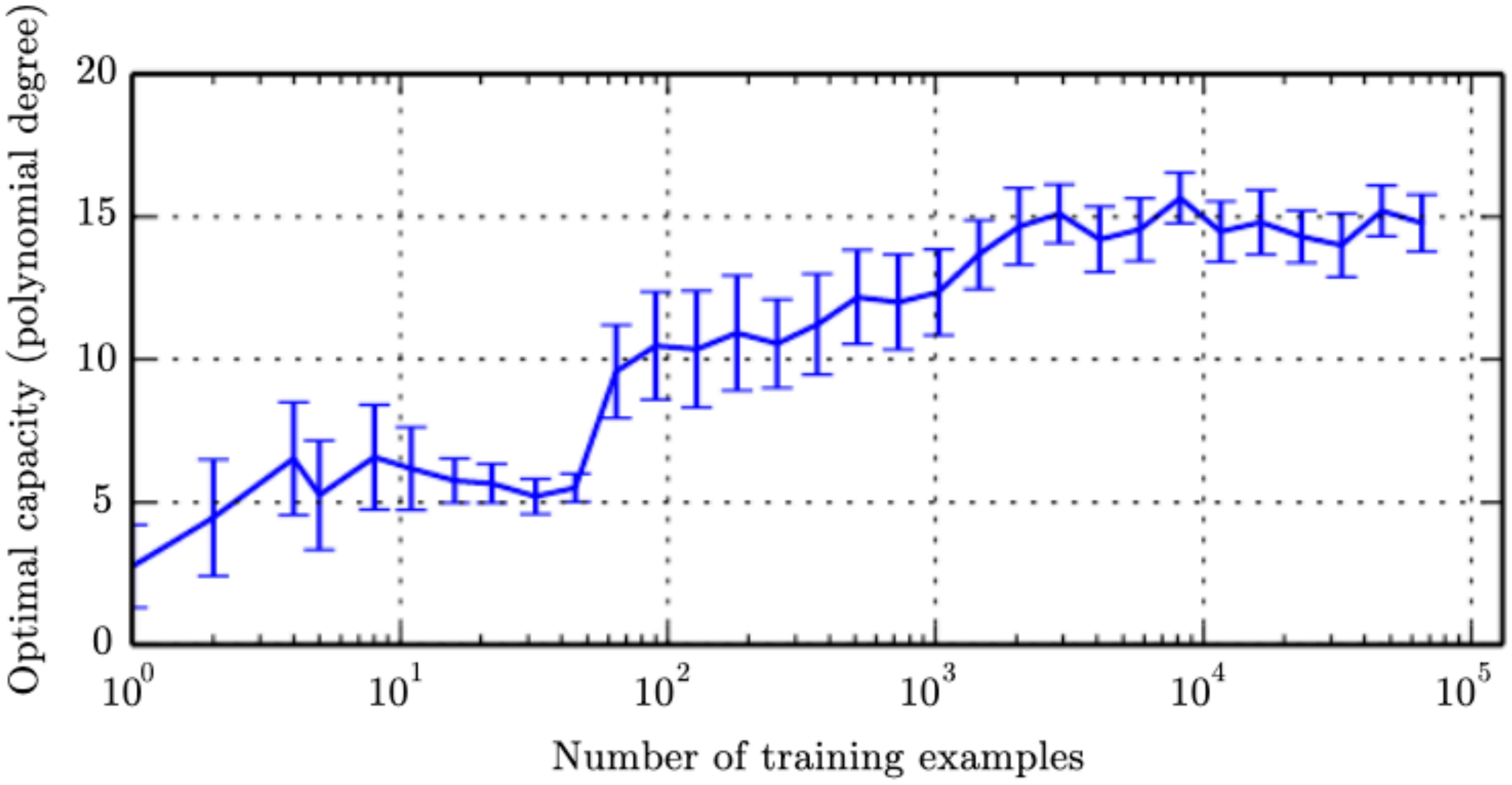
2. Make the gap between training and test error small".

Deep Learning, Goodfellow, Bengio, Courville, MIT Press 2016,

Machine Learning Basics chapter, page 108.

The static ML workflow: goals

So is ML all about finding a large dataset & a right capacity model?



Deep Learning, Goodfellow, Bengio, Courville, MIT Press 2016,

Machine Learning Basics chapter, page 114.

How do you think datasets should be acquired?

Static datasets: controlled

Small scale, but (some) controlled acquisition parameters

Image number	Object pose			Illumination direction		
	Frontal	22.5 ° right	22.5 ° left	Frontal	≈ 45 ° from top	≈ 45 ° from side
1	x			x		
2	x				x	
3	x					x
4		x		x		
5		x			x	
6		x				x
7			x	x		
8			x		x	
9			x			x

Table 3: The labeling of images within each scale in the KTH-TIPS database.

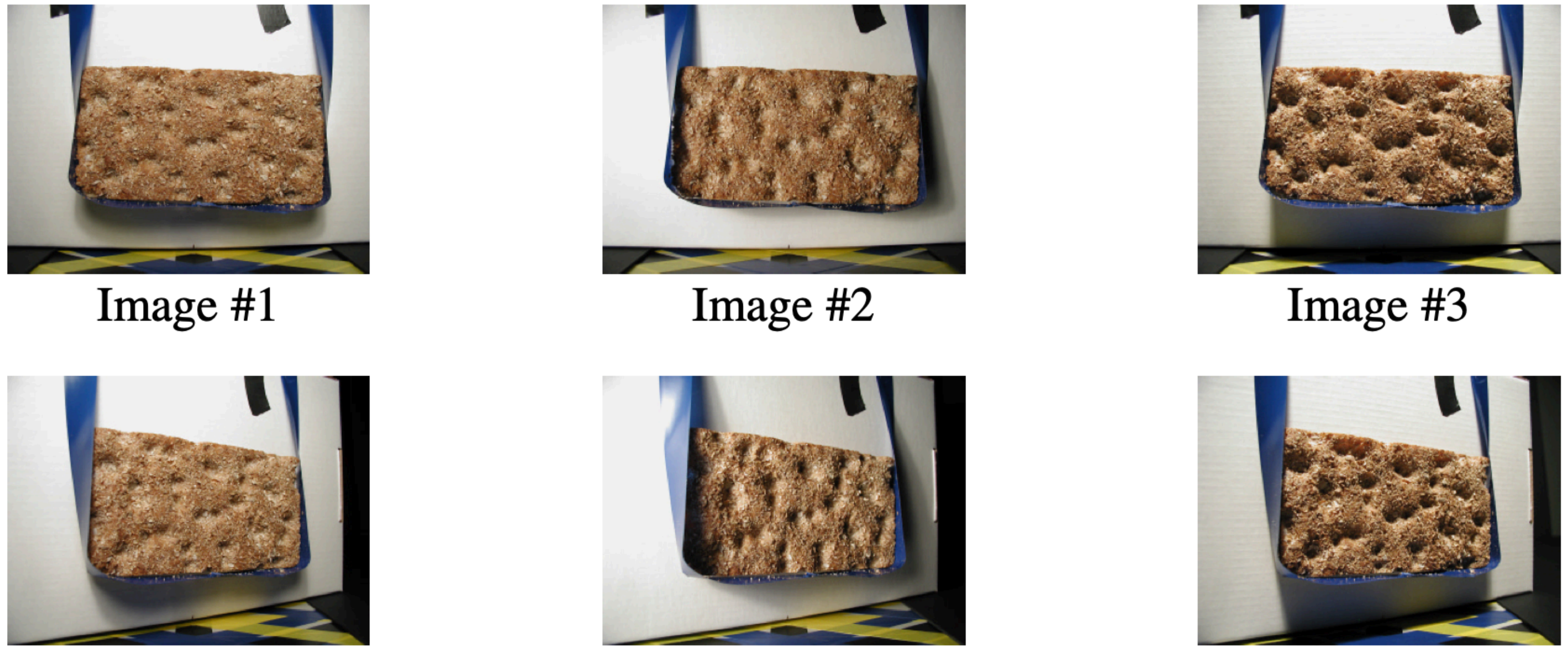


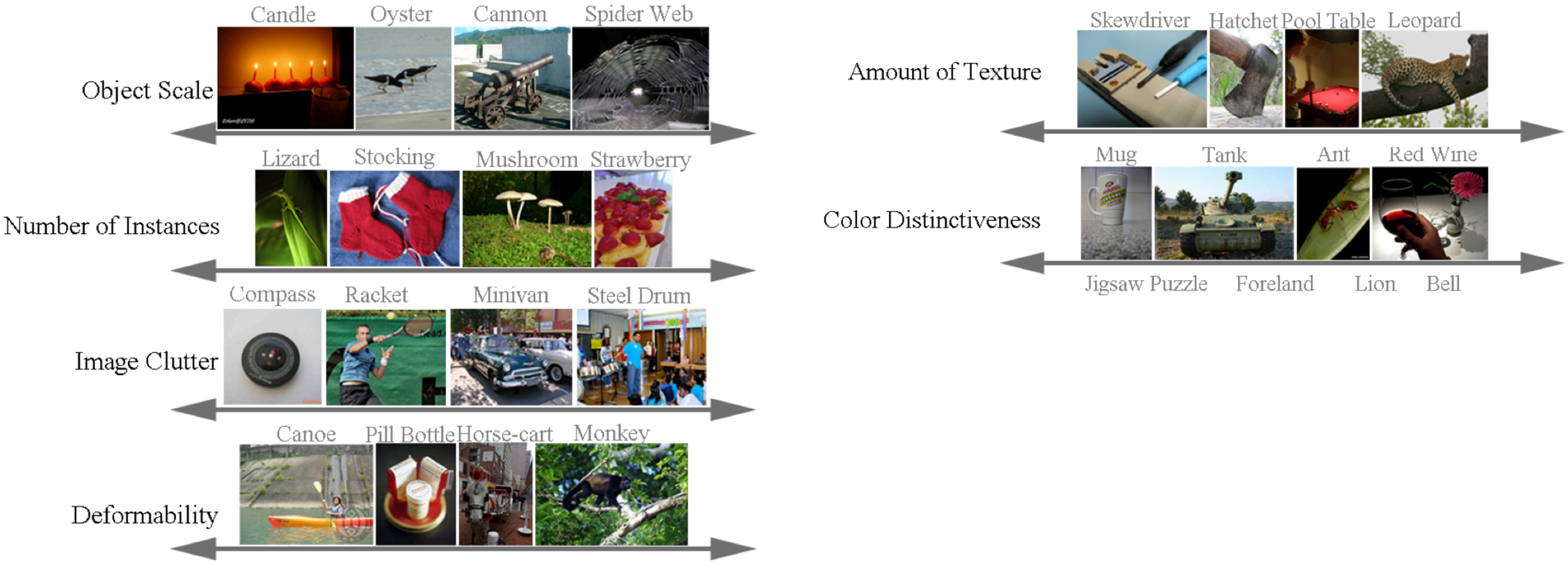
Image #1 Image #2 Image #3

Image #4 Image #5 Image #6

Hayman et al, “On the significance of real-world conditions for material classification”, ECCV 2004
& Fritz, Hayman et al, “The KTH-TIPS database”, technical report 2004

Static datasets: large scale

A big focus of modern dataset has been on large scale & diversity



Russakovsky & Deng et al, “ImageNet Large Scale Visual Recognition Challenge, IJCV 2015, (challenges since 2010)

Static datasets: large scale

And trying to ensure reasonable train, validation, test splits through complex collection processes

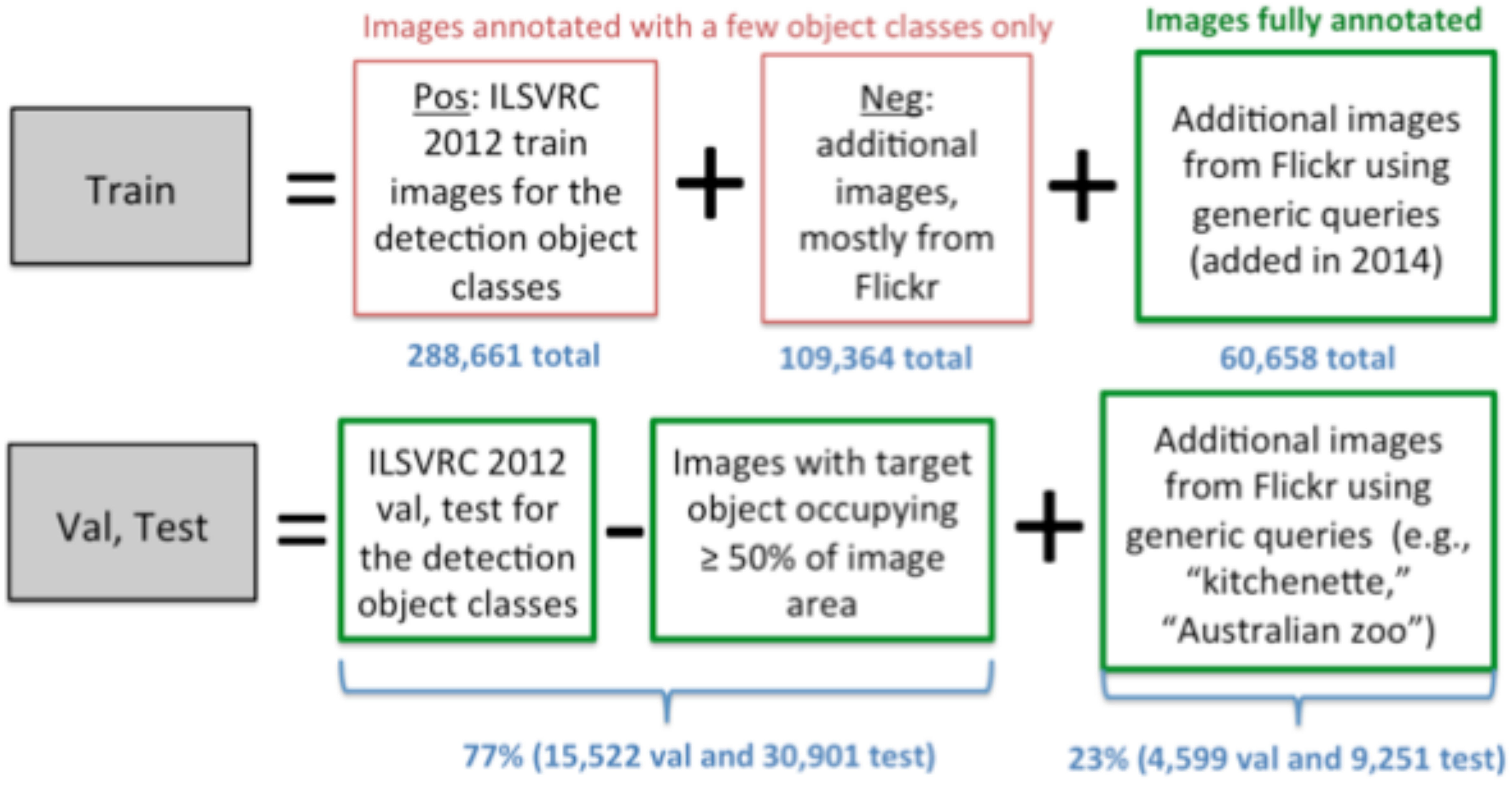


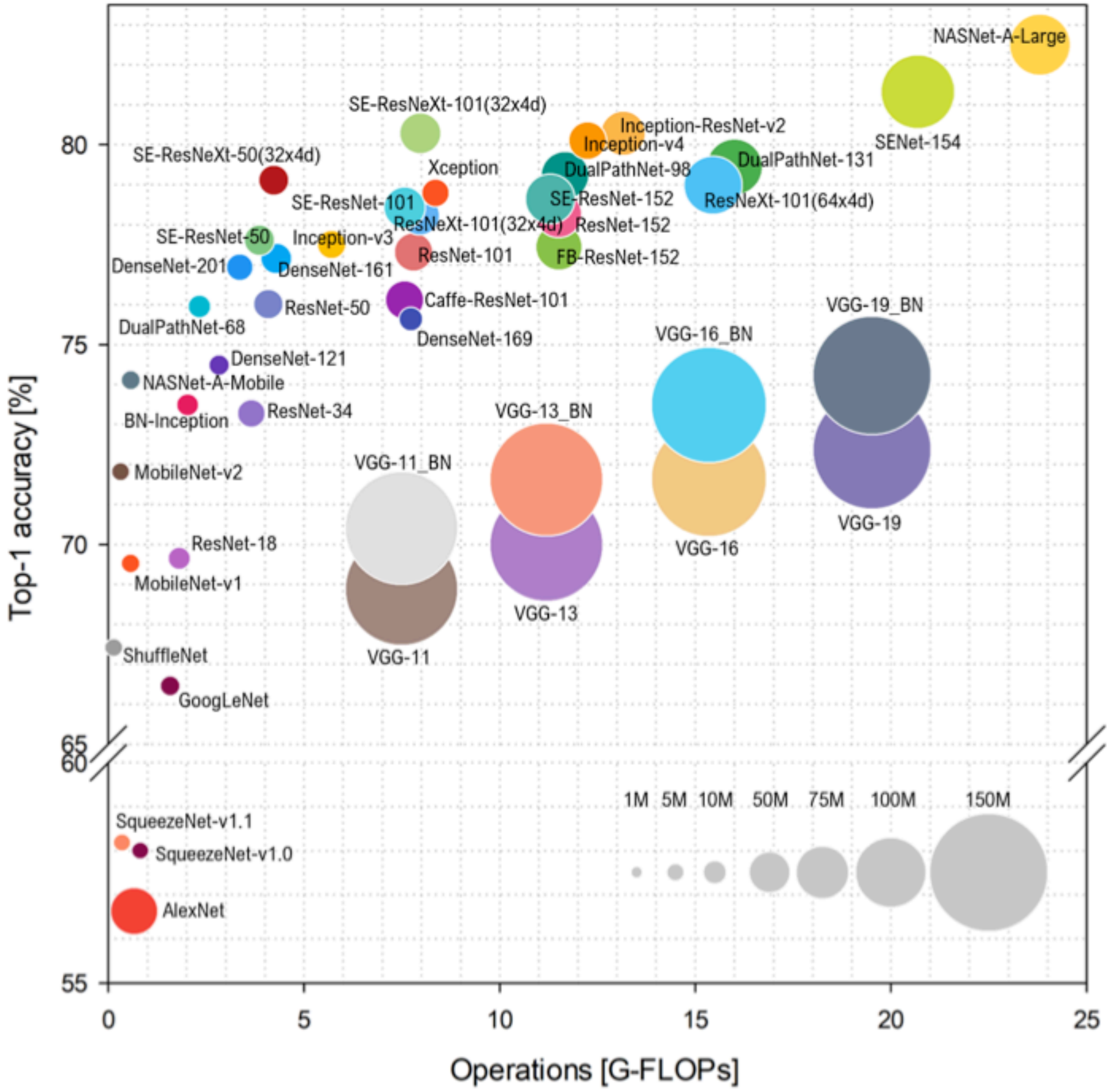
Image classification annotations (1000 object classes)

Year	Train images (per class)	Val images (per class)	Test images (per class)
ILSVRC2010	1,261,406 (668-3047)	50,000 (50)	150,000 (150)
ILSVRC2011	1,229,413 (384-1300)	50,000 (50)	100,000 (100)
ILSVRC2012-14	1,281,167 (732-1300)	50,000 (50)	100,000 (100)

Russakovsky & Deng et al, "ImageNet Large Scale Visual Recognition Challenge, IJCV 2015, (challenges since 2010)

What do you think:
should our primary goal be the solution to such benchmarks?

Static models

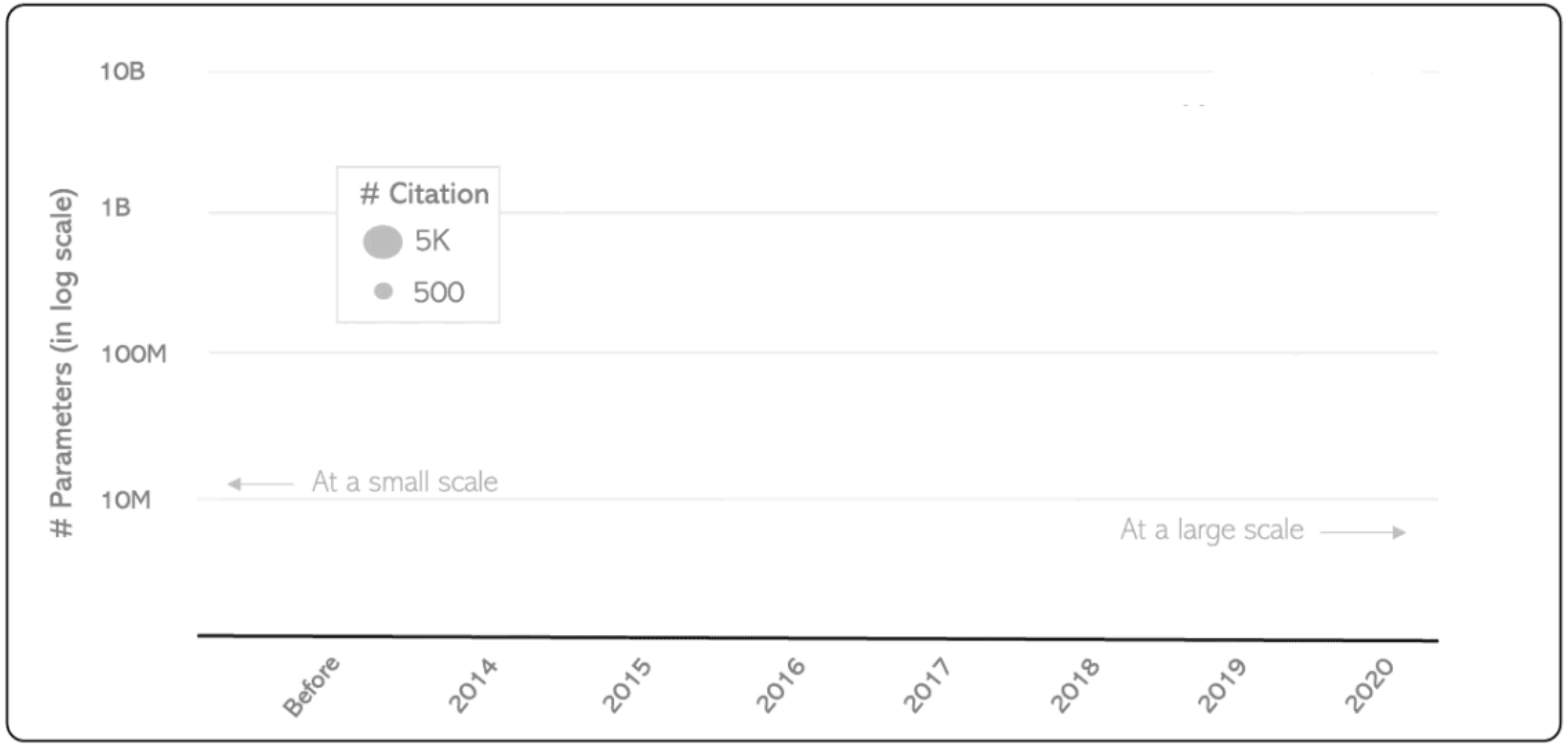


A very big emphasis has then been on "solving" such benchmarks

ImageNet is a prime example, where models & compute got bigger and more accurate over time

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

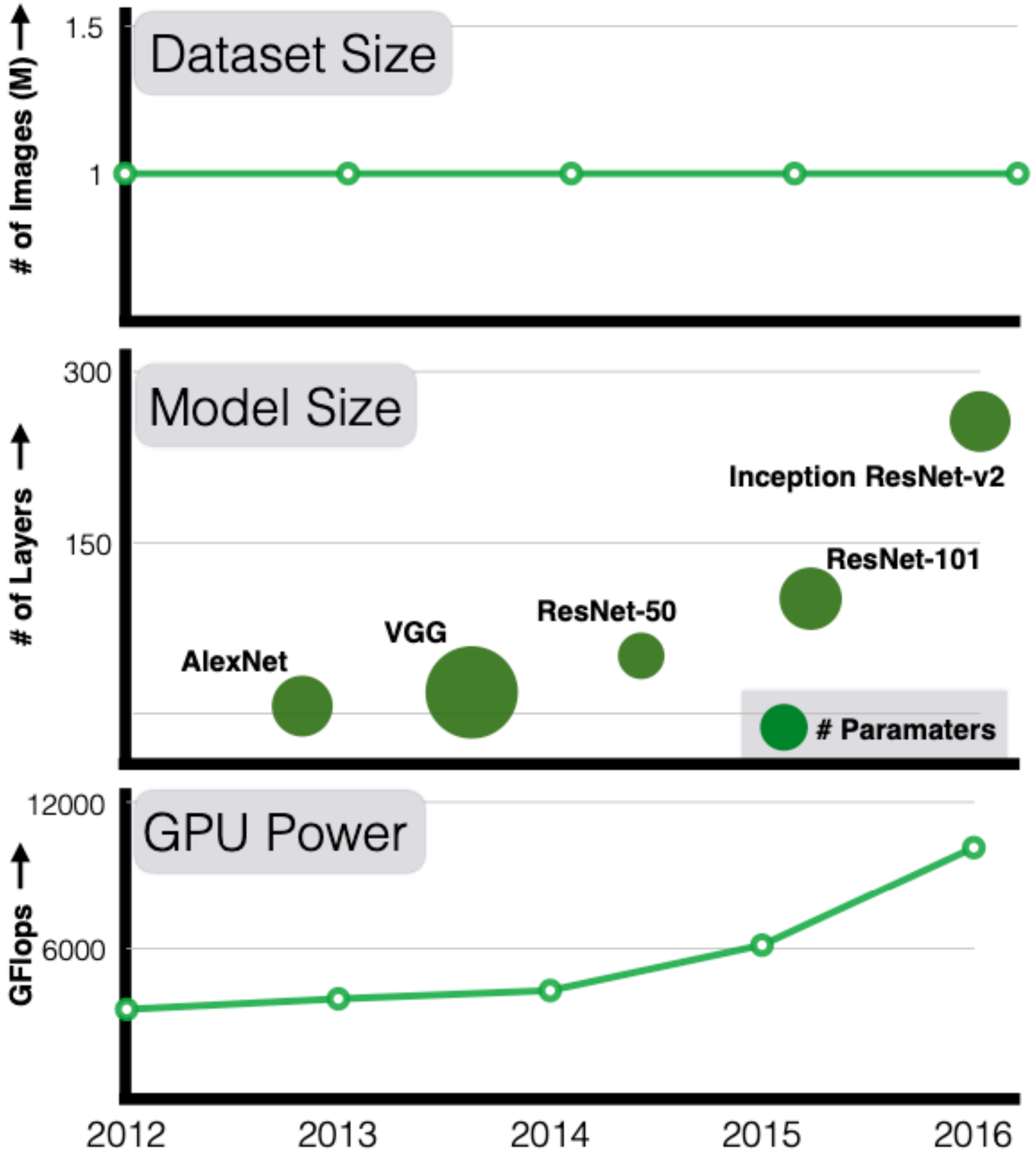
Static models



This trend continues even today

Li & Gao, “A deep generative model trifecta: three advances that work towards harnessing large-scale power, Microsoft Research Blog, 2020: https://www.microsoft.com/en-us/research/blog/a-deep-generative-model-trifecta-three-advances-that-work-towards-harnessing-large-scale-power/

Data and model centrism



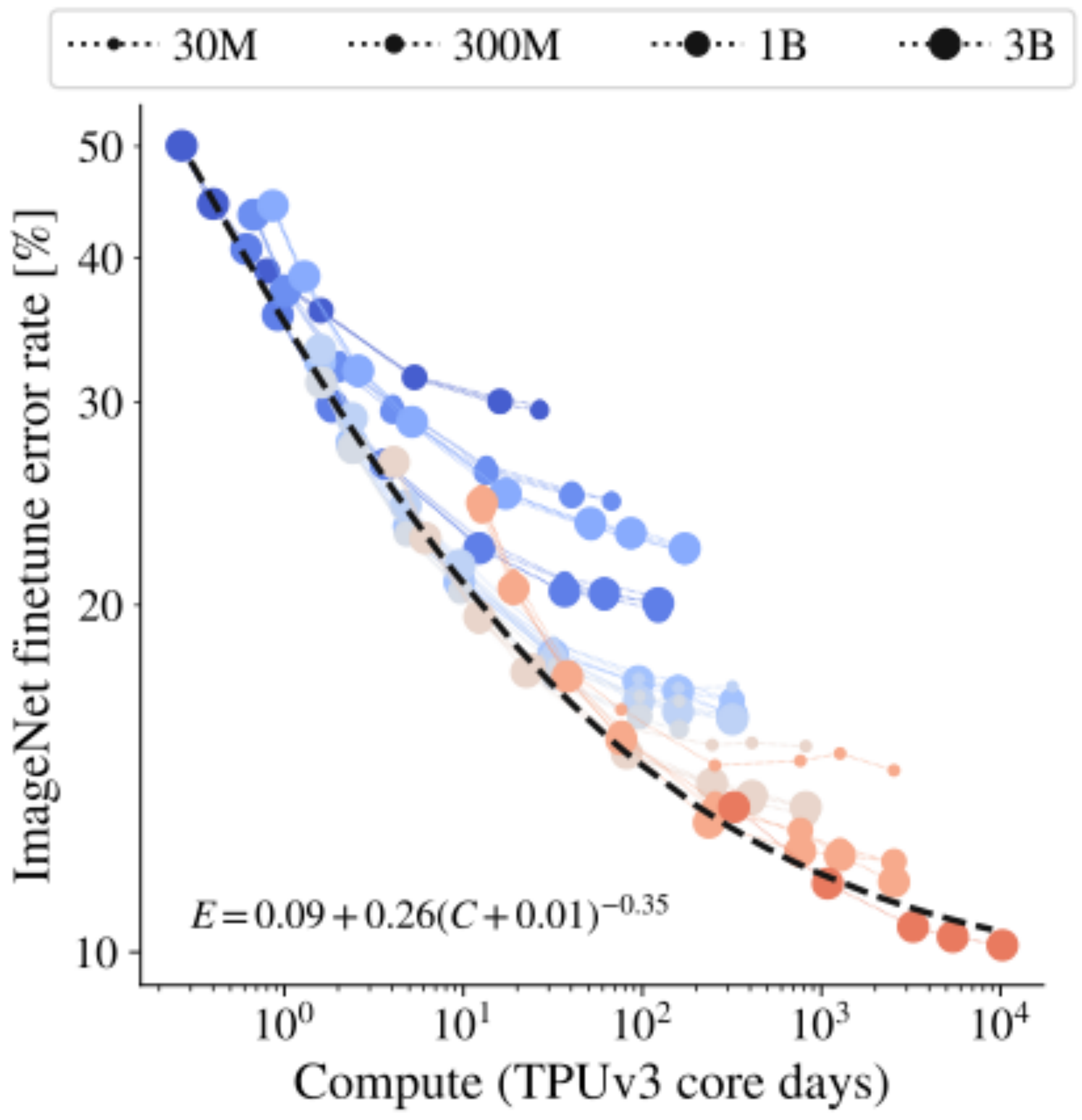
At the same time, it's often "either" models or data

For example, ImageNet has remained largely static* over time

* (excluding some concerns over fair representation)

Sun et al, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era", ICCV 2017

Data and model centrism



Zhao et al, "Scaling Vision Transformers", preprint 2021

Or conversely, a model is picked (here a transformer) and datasets are extended

Example from ImageNet to the (non-public) JFT 300M & JFT-3B

Summary: static ML workflow

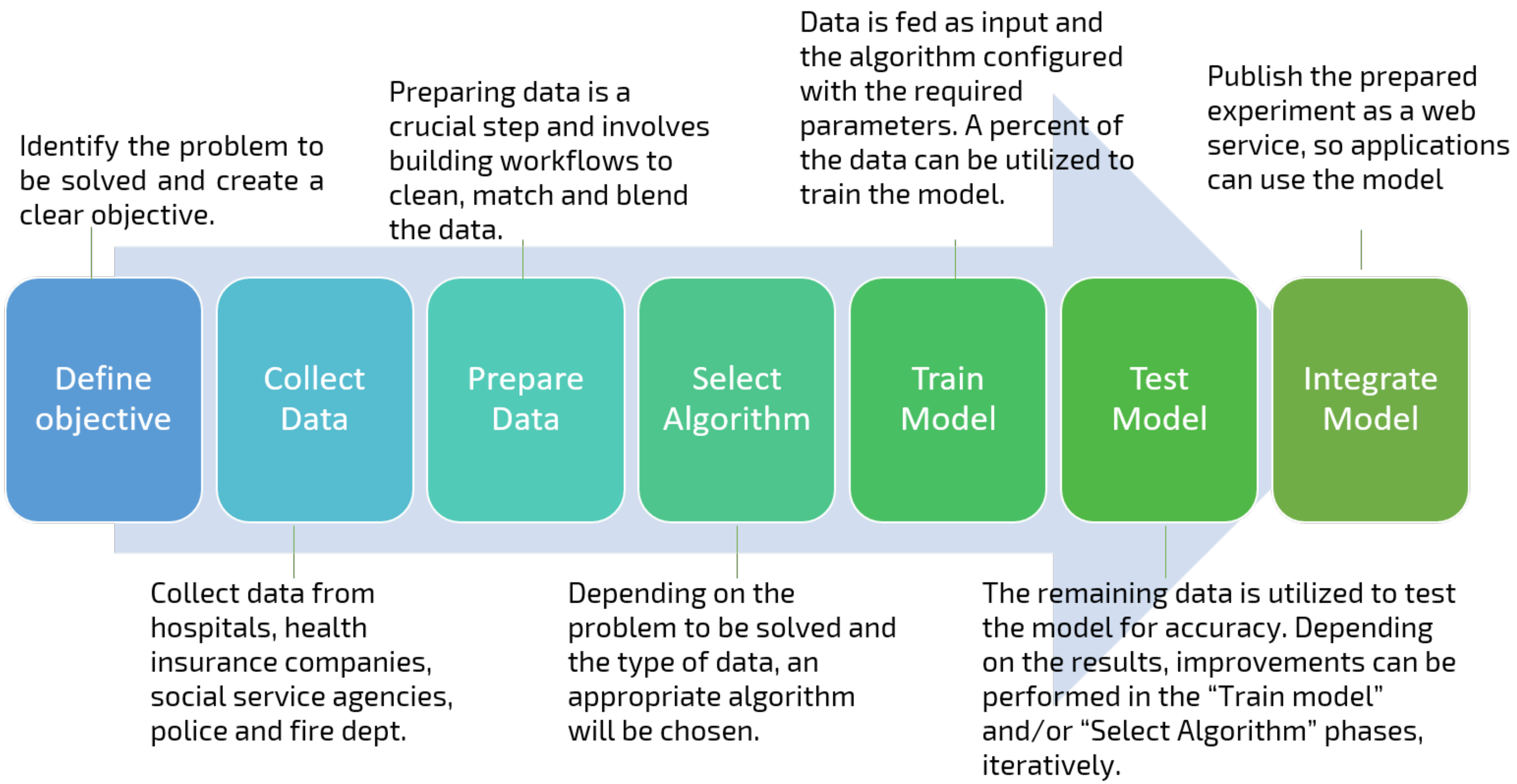
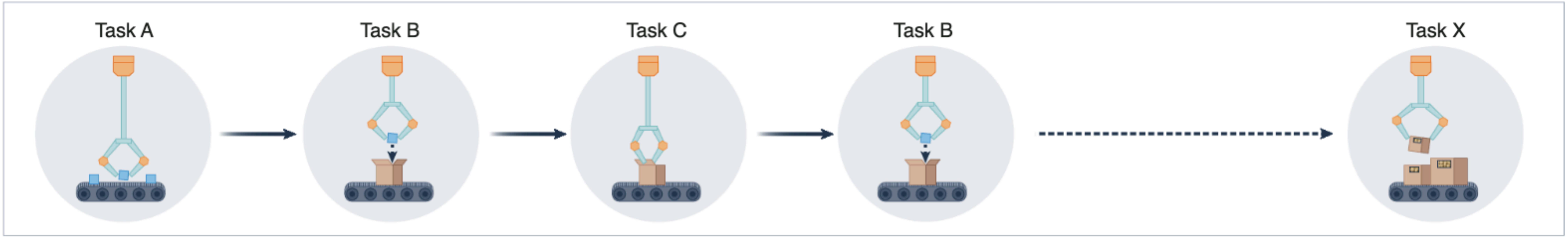


Figure from https://www.congrelate.com/get-workflow-machine-learning-images/

But what if we want to continue learning tasks? …

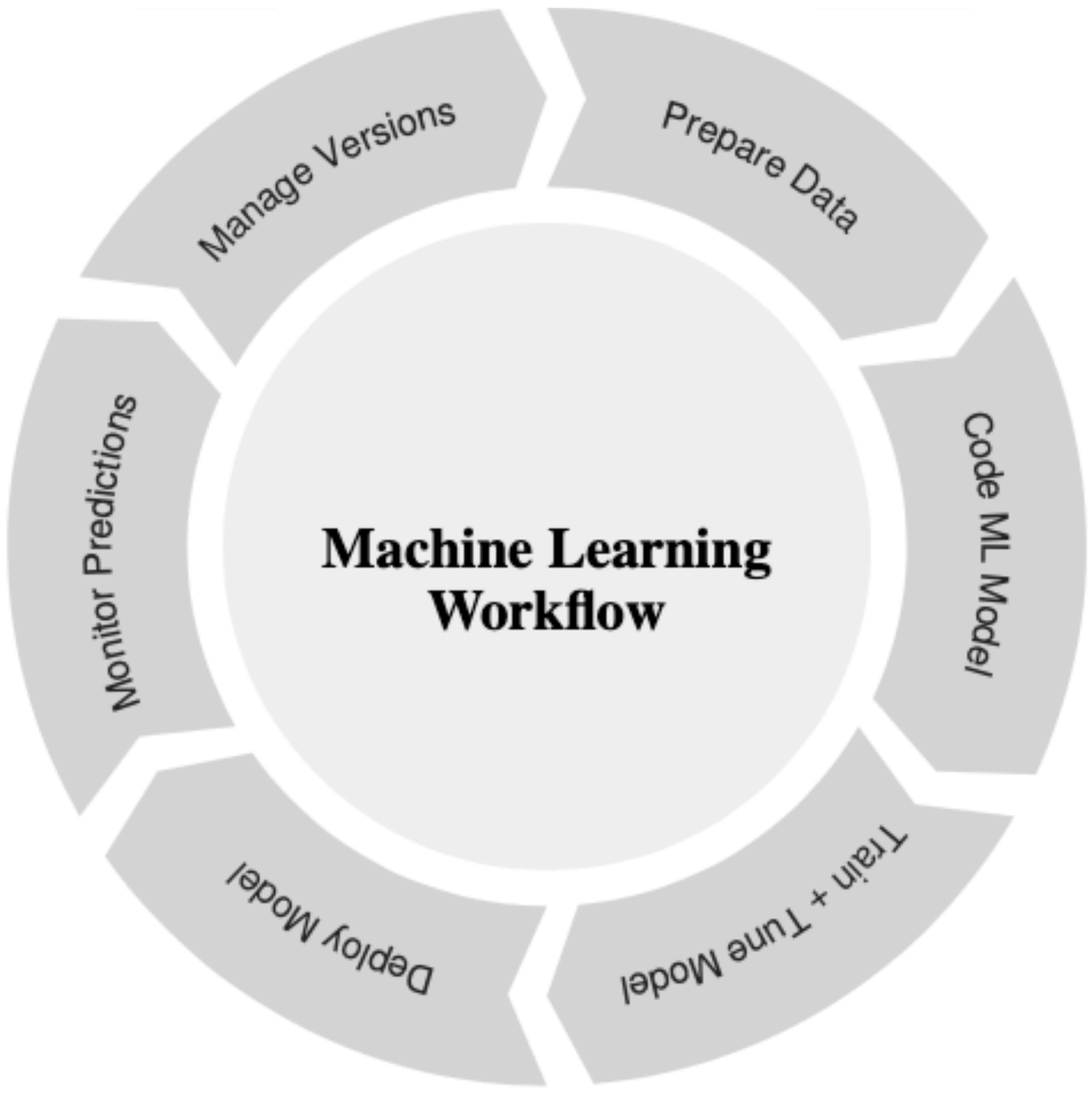


Kudithipudi et al, "Biological underpinnings for lifelong learning machines", Nature Machine Intelligence (4), 2022

Or add more categories?

Image examples from CUB200: “black footed albatross”, “rusty blackbird”, “sooty albatross”, and “cardinal”.
Welinder et al, Caltech-UCSD Birds 200, CNS-TR-2010-001, California Institute of Technology, 2010

Can we just iterate?

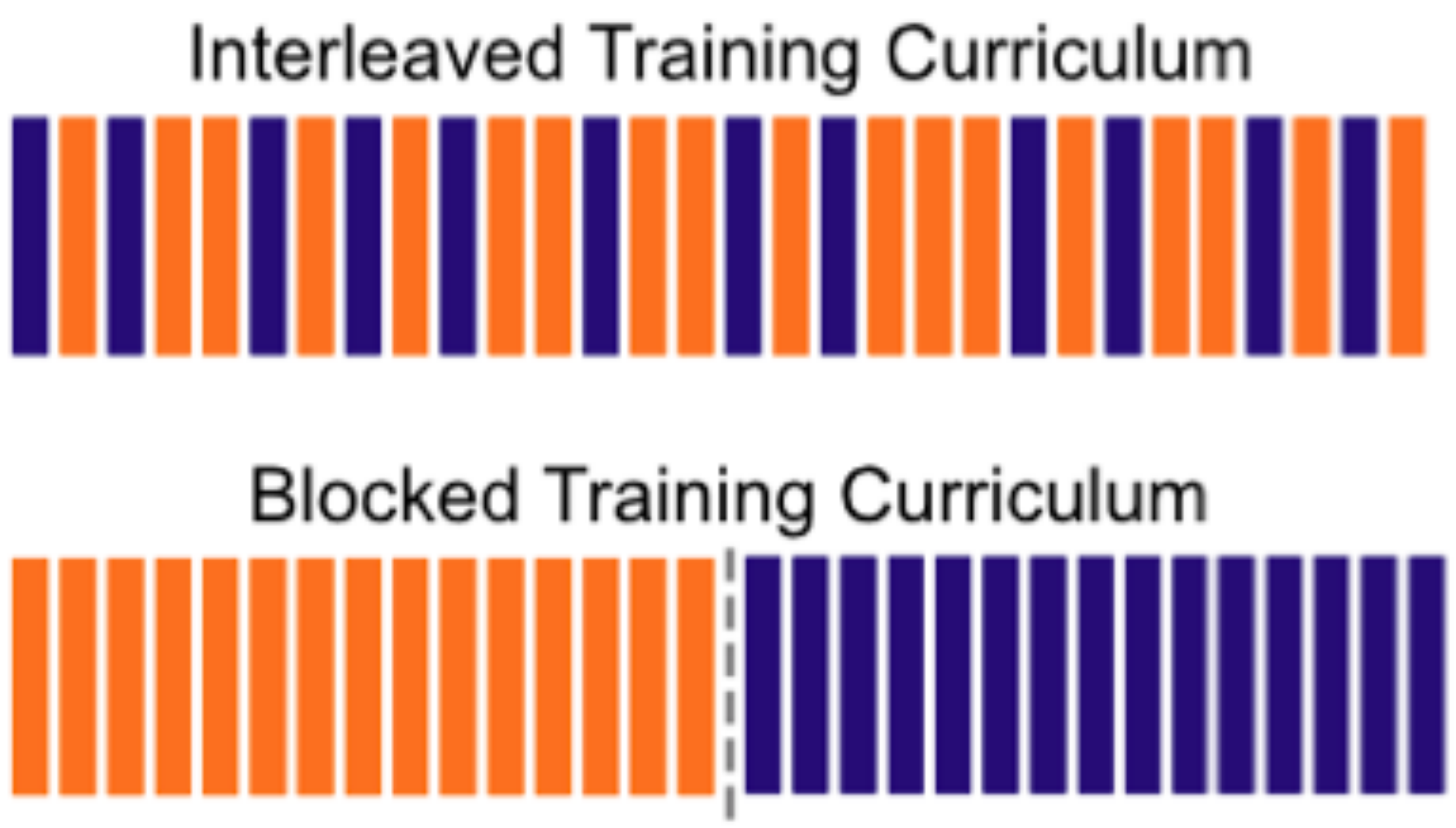


What do you think could happen?

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

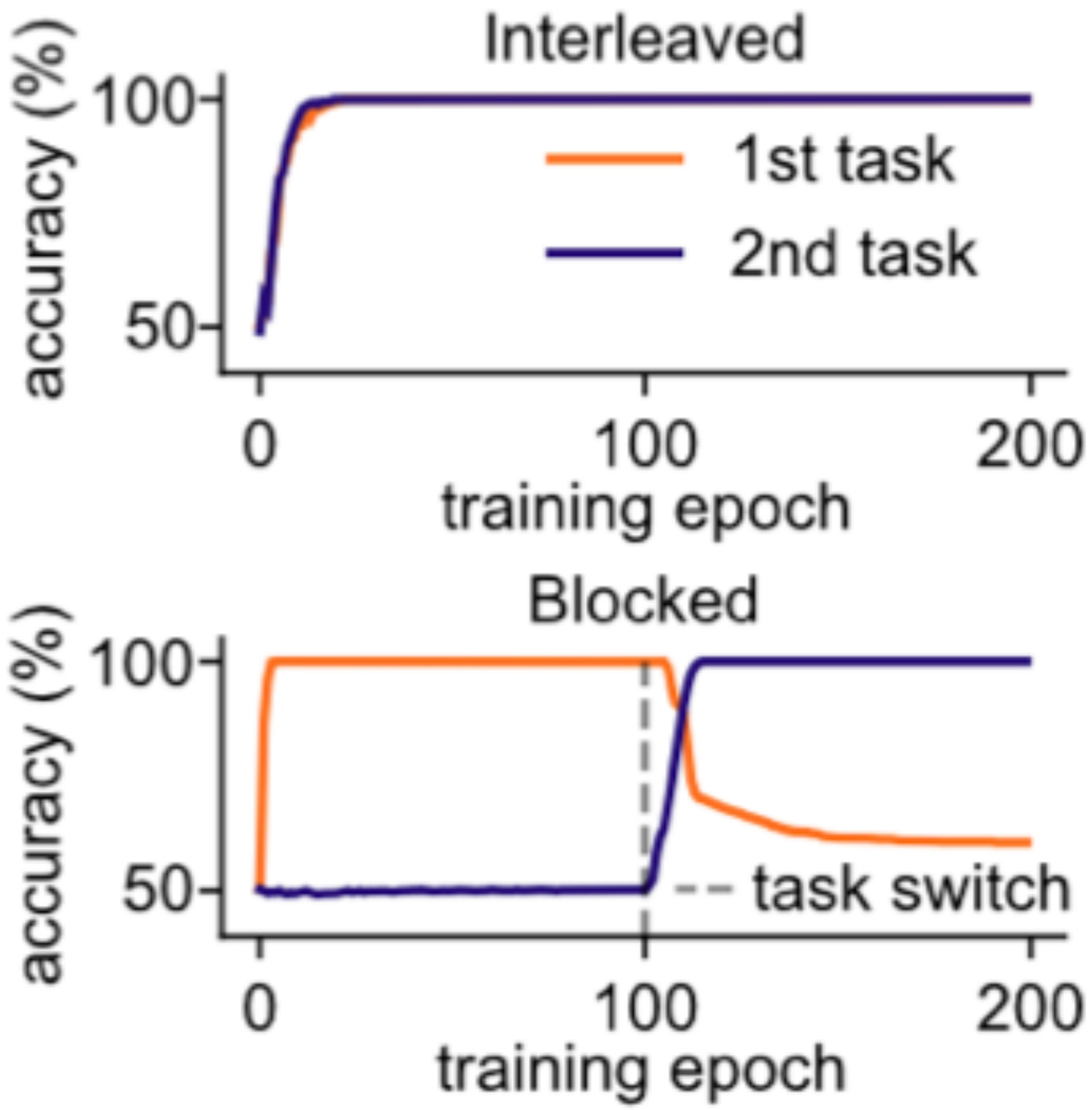
Continual learning

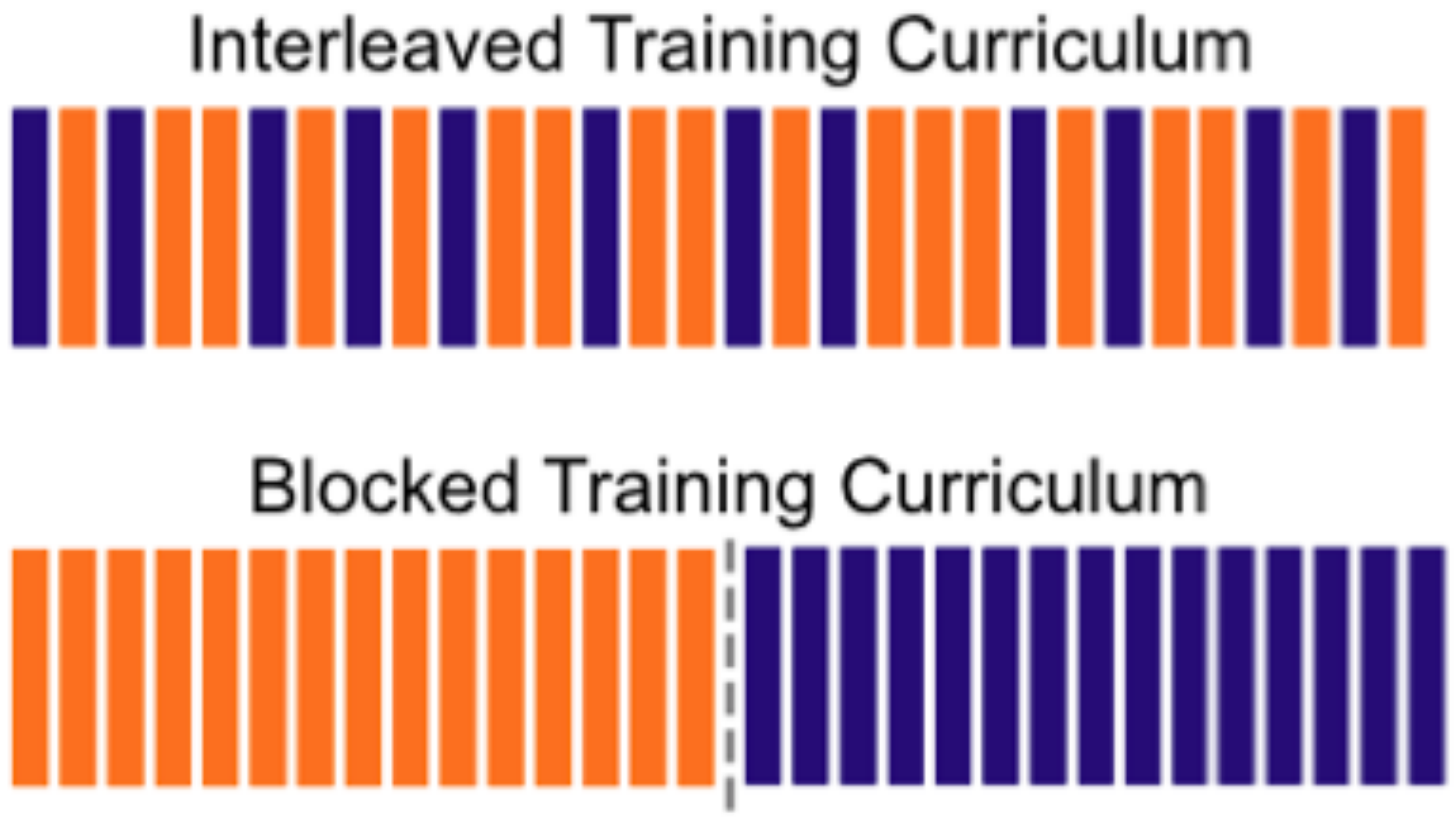
What do you think will happen if we present both of these to a machine learner?



Flesch et al, “Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals”, preprint, 2022

Continual learning





Machine learning typically shuffles data & performs poorly when data is ordered

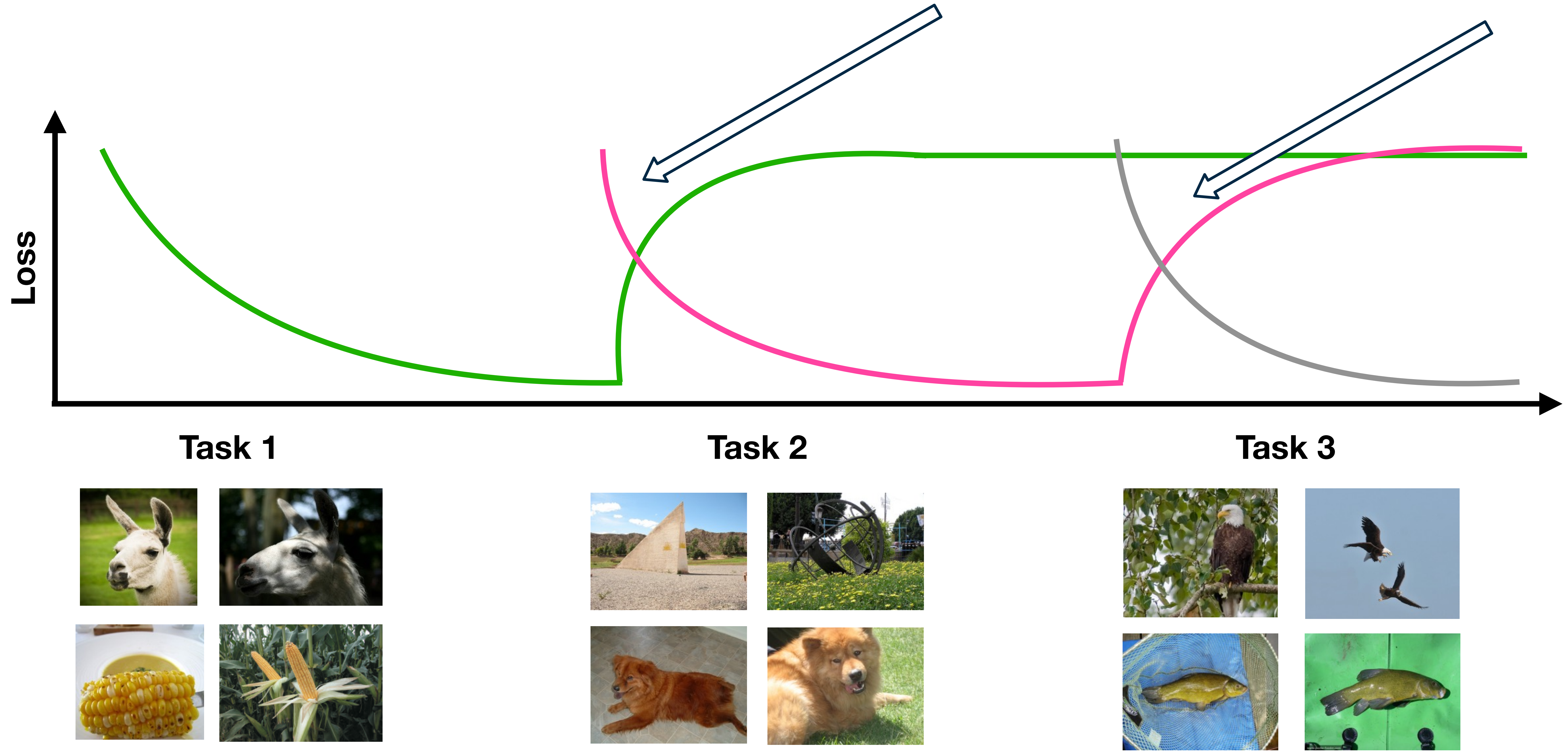
Flesch et al, “Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals”, preprint, 2022

Why do we need an entire lecture?

Challenge: forgetting

A popular example

Catastrophic forgetting



Key assumption: no access to/ revisiting of prior "task" data!

Challenge: the world is “open”

The threat of unknown unknowns

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

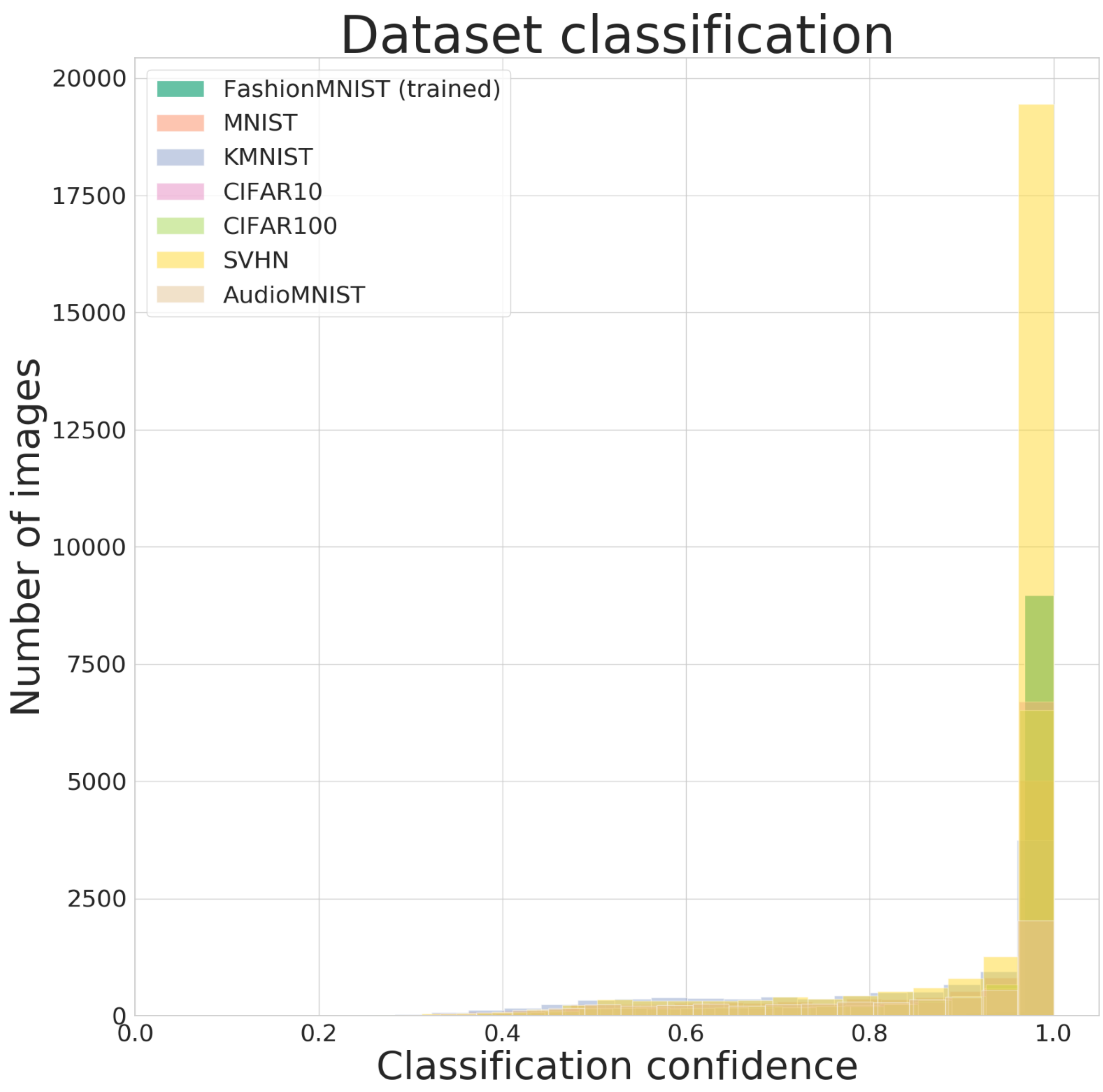
Challenge: the world is “open”

The threat of unknown unknowns

Most ML models are overconfident

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

Challenge: the world is "open"



A quantitative example:

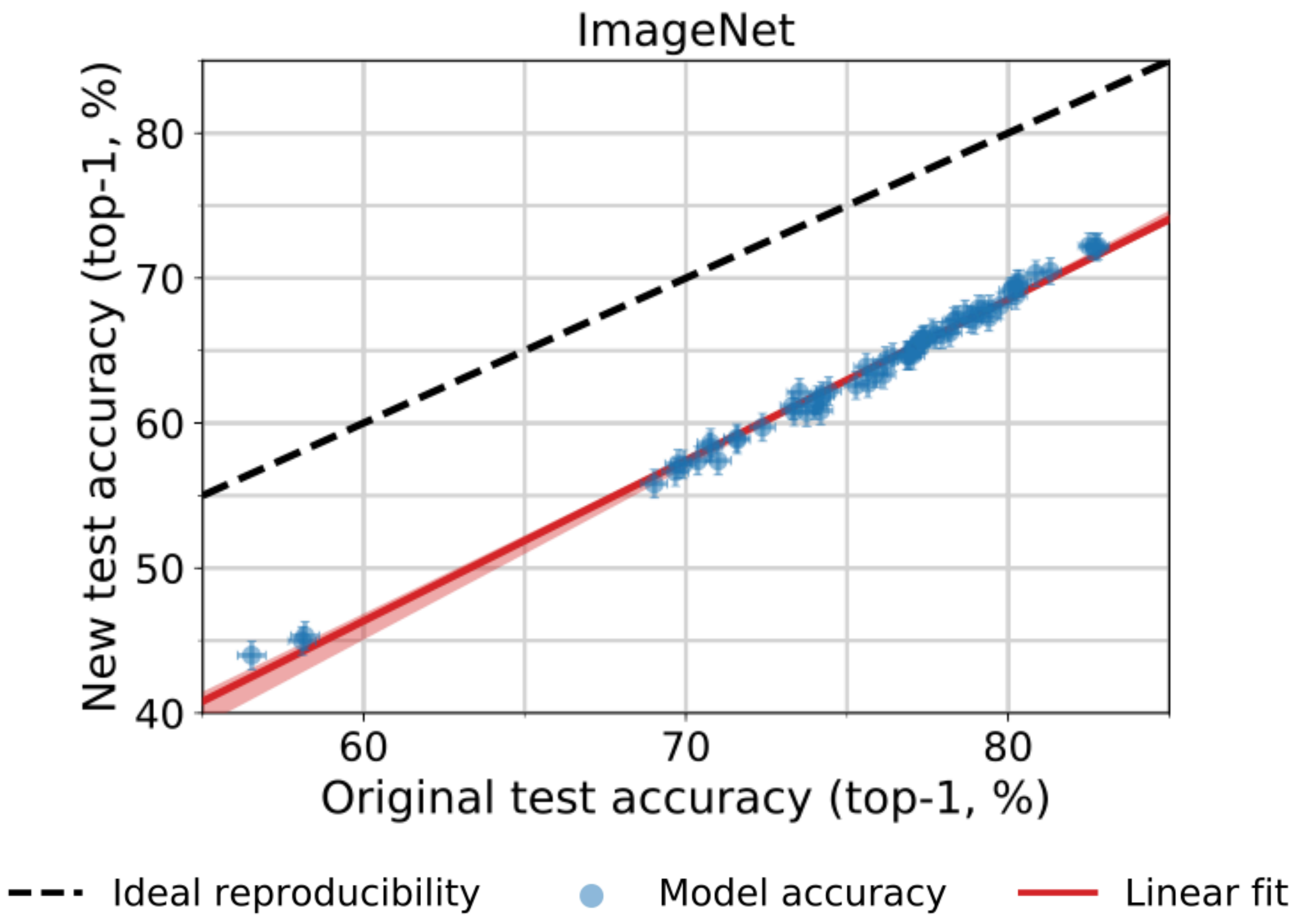
1. Train a neural network classifier on a dataset (here Fashion items)
2. Log predictions for arbitrary other datasets
3. Observe that majority of misclassifications happen with large output "probability"

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

“But this example is unrealistic”!

What do you think will happen if we collect a second test set (following the same procedure) & evaluate?

Challenge: distribution shifts

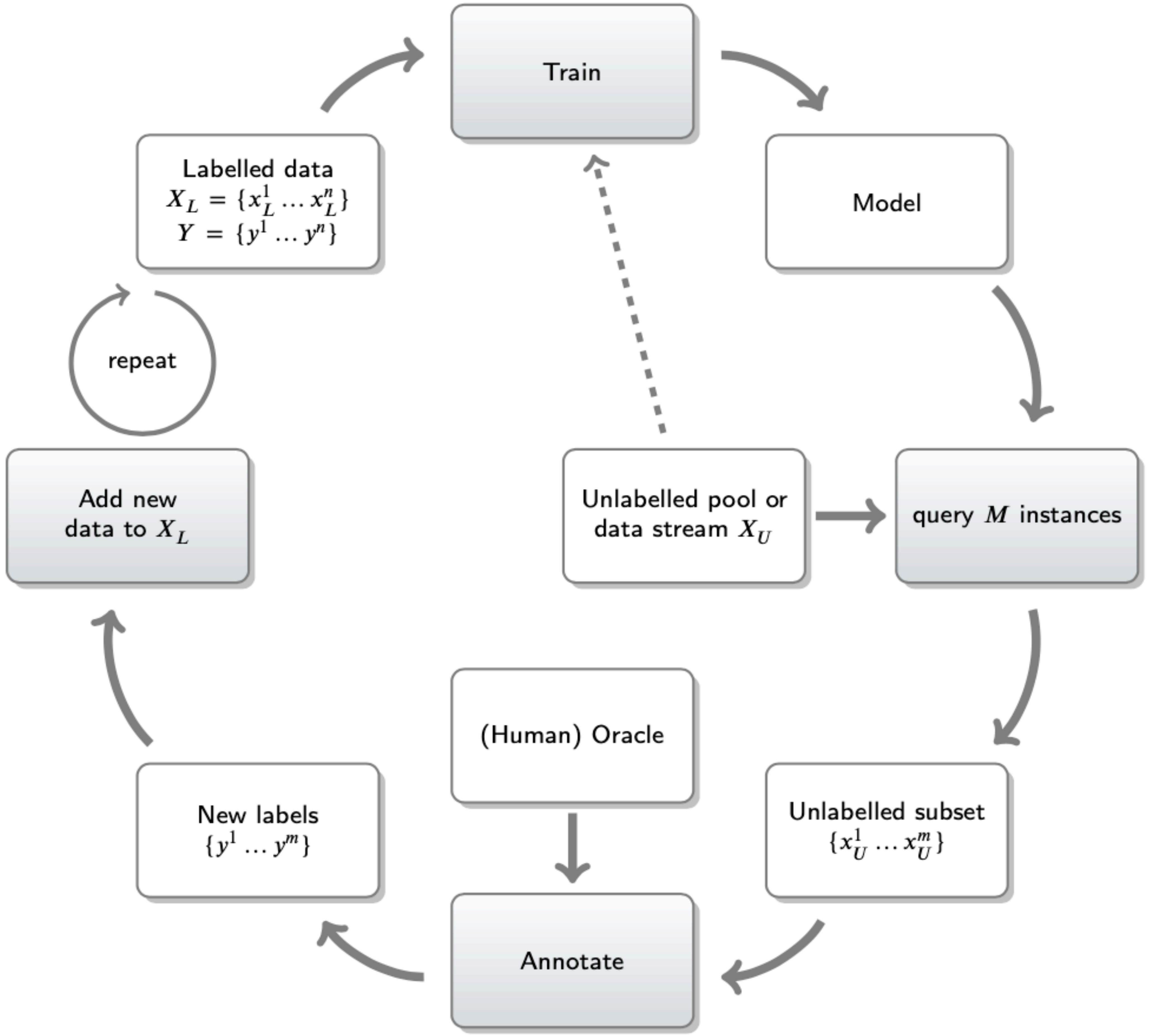


Natural data distributions are complex & can easily shift!

Performance loss even happens if we recollect another “test set” with the same instructions a second time!

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

Challenge: select & add data



What if we want to add data over time?

- How to pick data?
- Does the data belong to the task?
- How similar is the data?
- How optimize accumulated error (is this even what we want?)

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

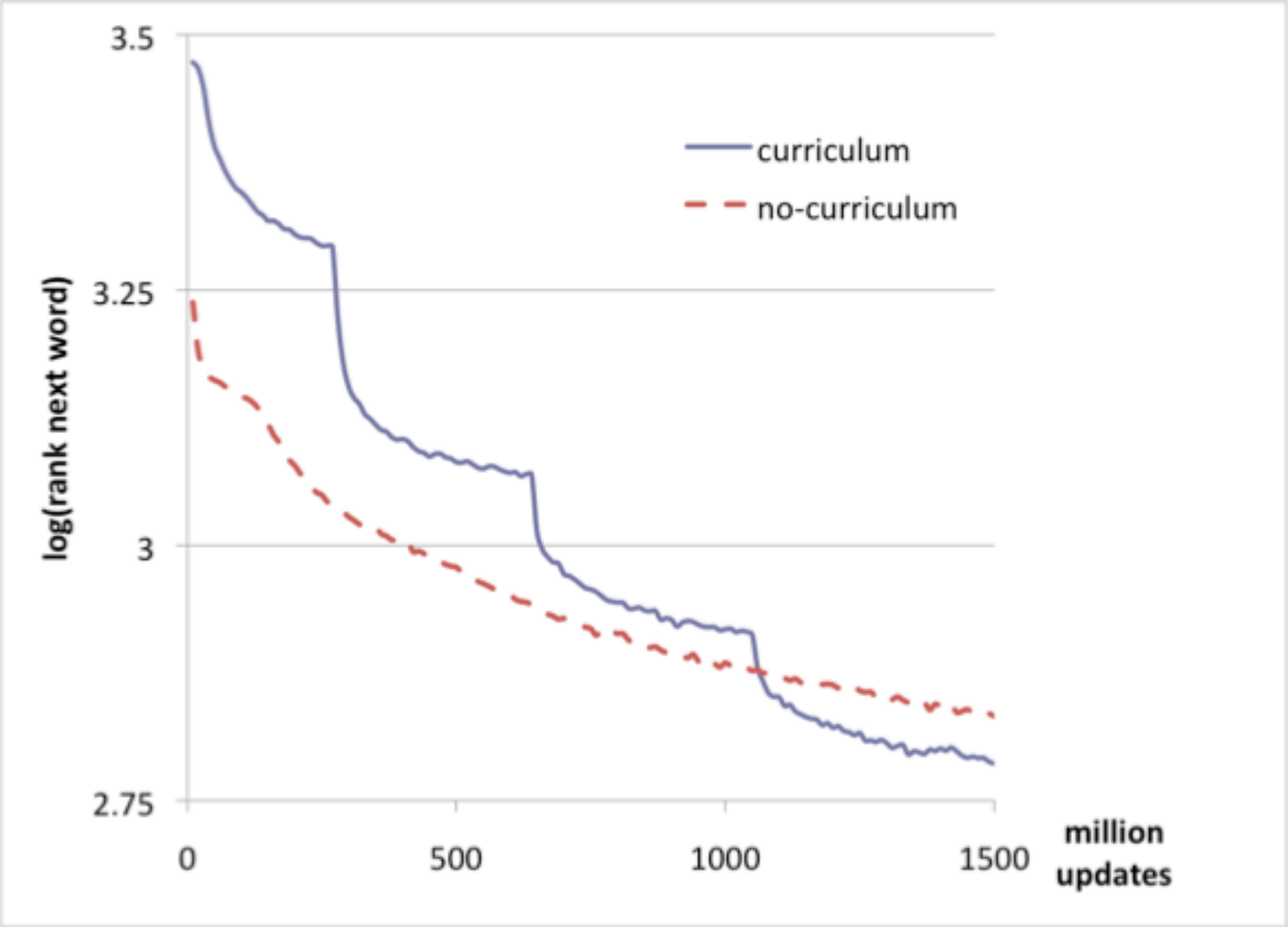
What kind of data would you intuitively pick?

Challenge: concept difficulty

Example: Ranking language model trained with vs without curriculum on Wikipedia

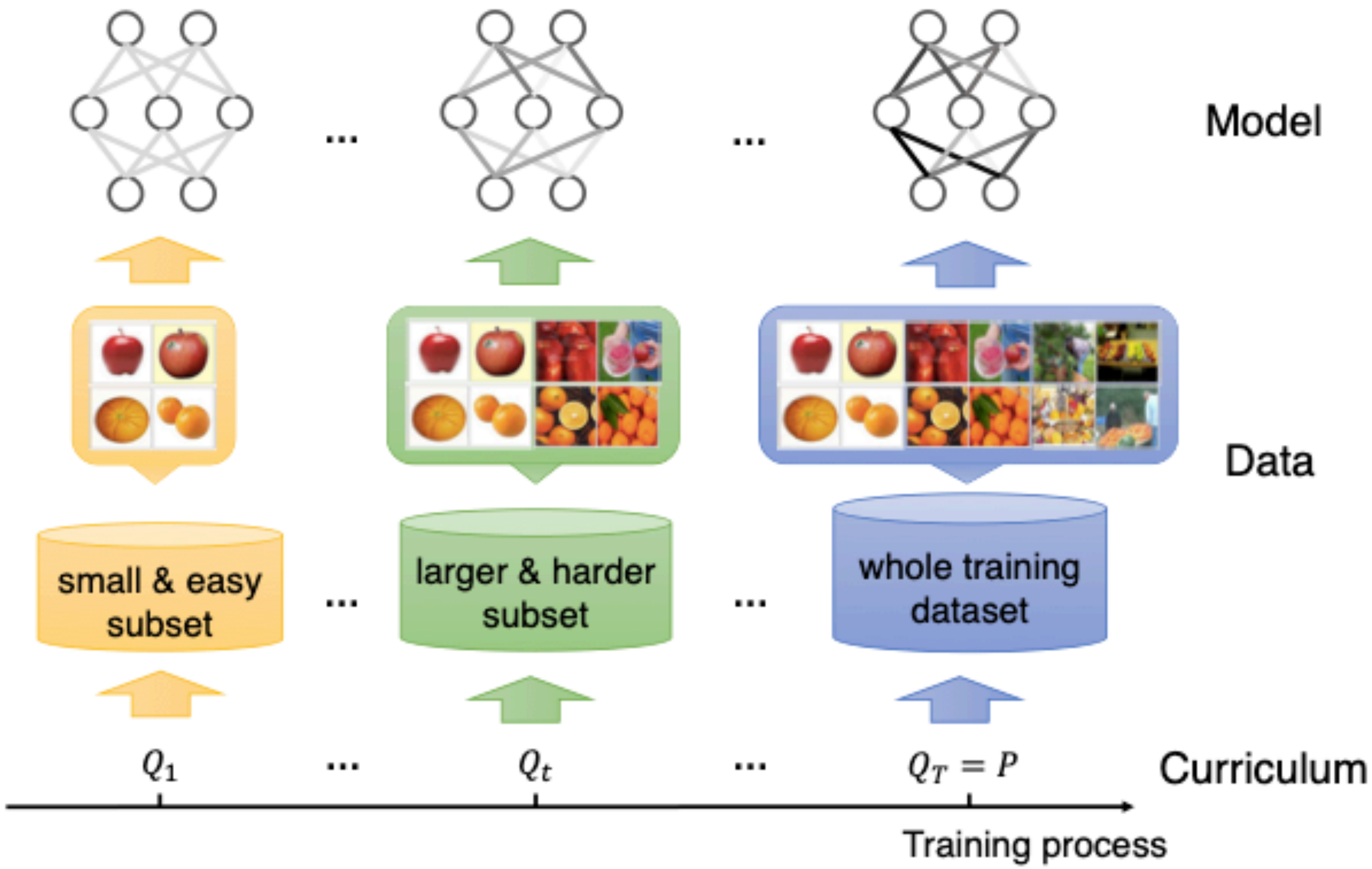
"Error" is log of the rank of the next word (within 20k-word vocabulary).

1. The curriculum-trained model skips examples with words outside of 5k most frequent words
2. Then skips examples outside 10k most frequent words and so on



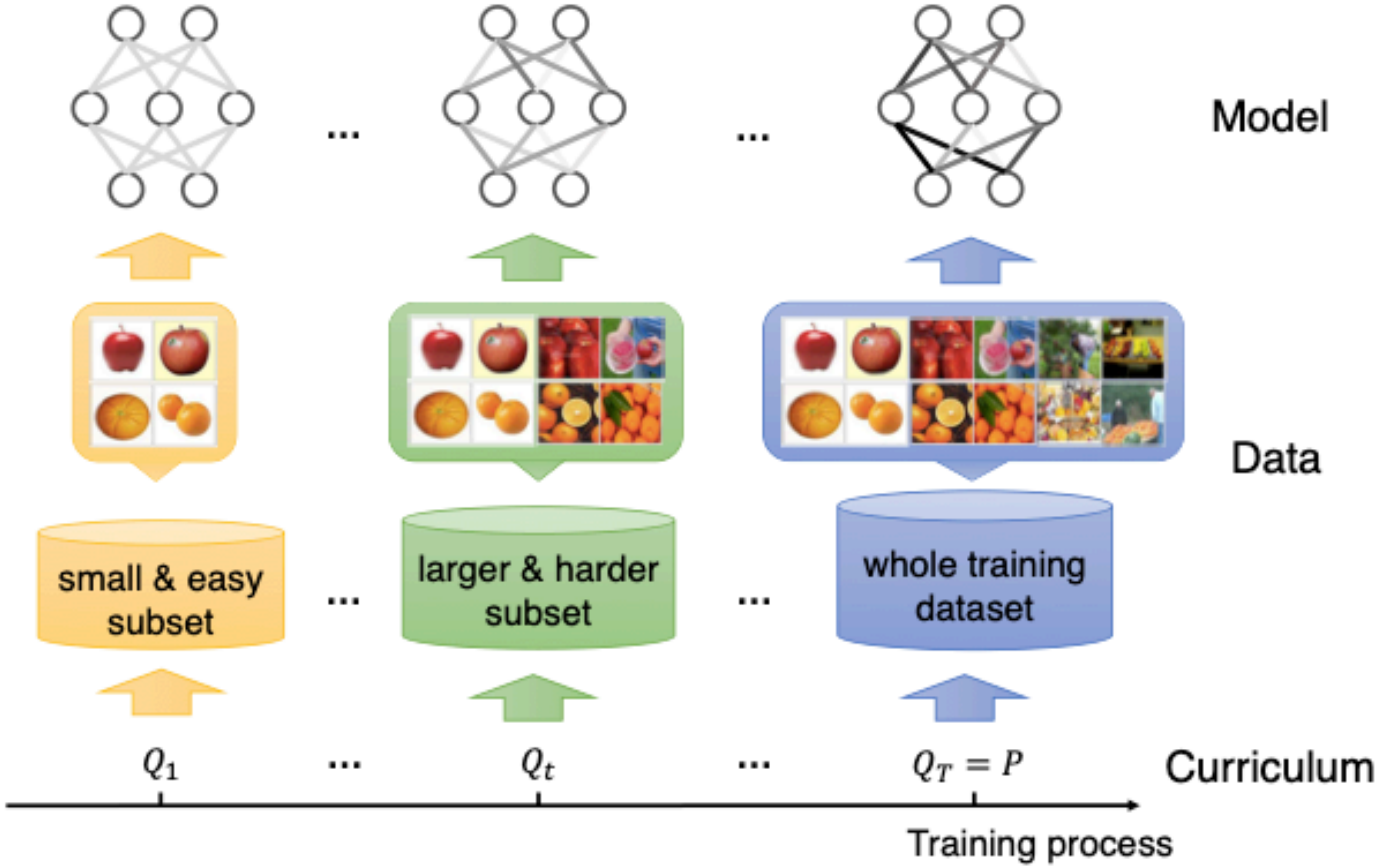
Bengio et al, "Curriculum Learning", ICML 2009

Challenge: concept difficulty



Wang et al, "A Survey on Curriculum Learning", TPAMI 2021

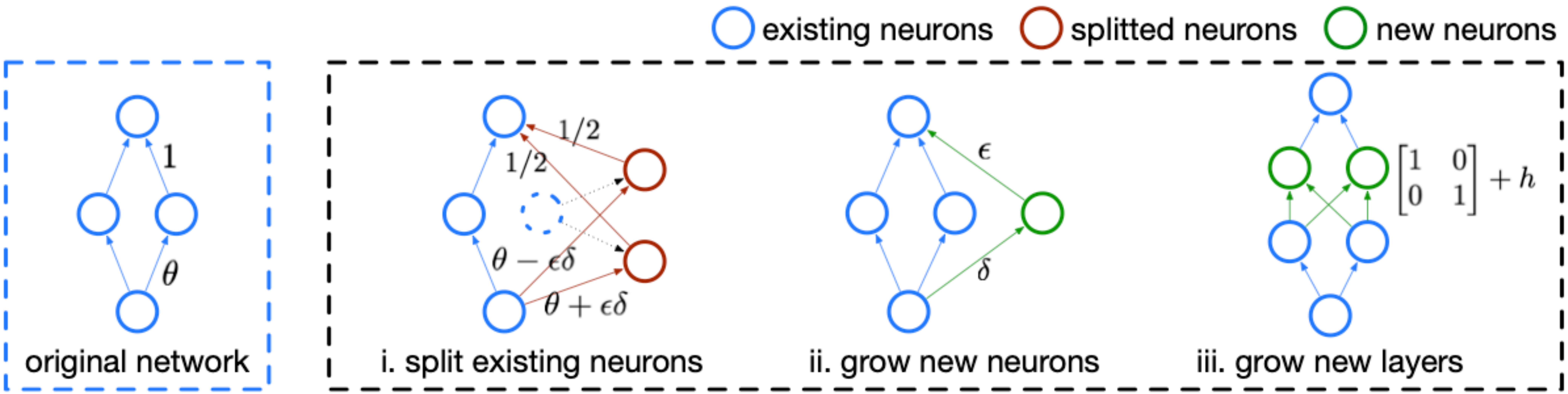
Challenge: concept difficulty



The model choice in this picture remains the same, do you think this is sufficient?

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

Challenge: adapting models



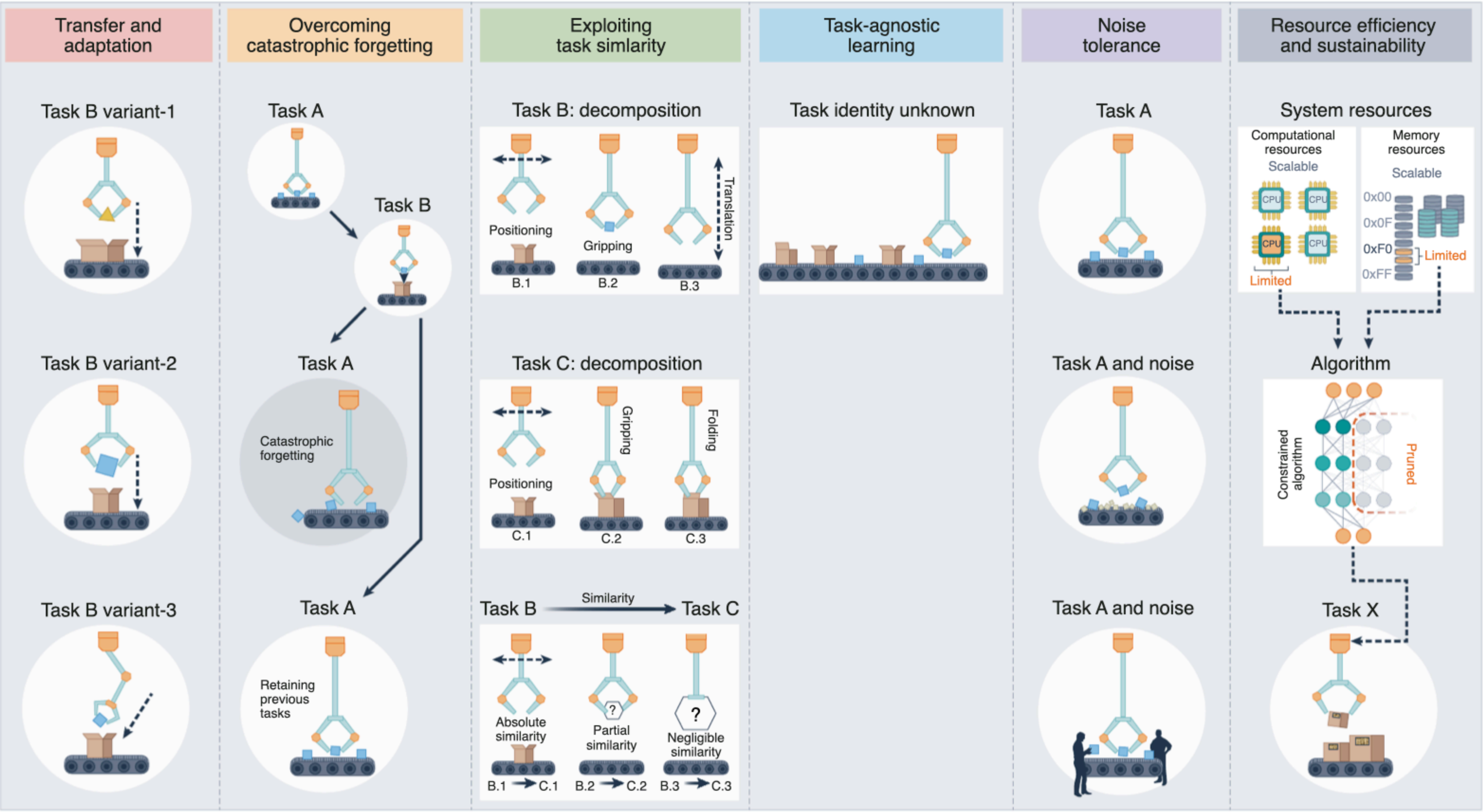
But is our initial model choice and its practical realization still good enough?

What if complexity changes?

Or even the inductive bias should be altered?

Wu & Liu et al, “Firefly Neural Architecture Descent: A General Approach for Growing Neural Networks”, NeurIPS 2020

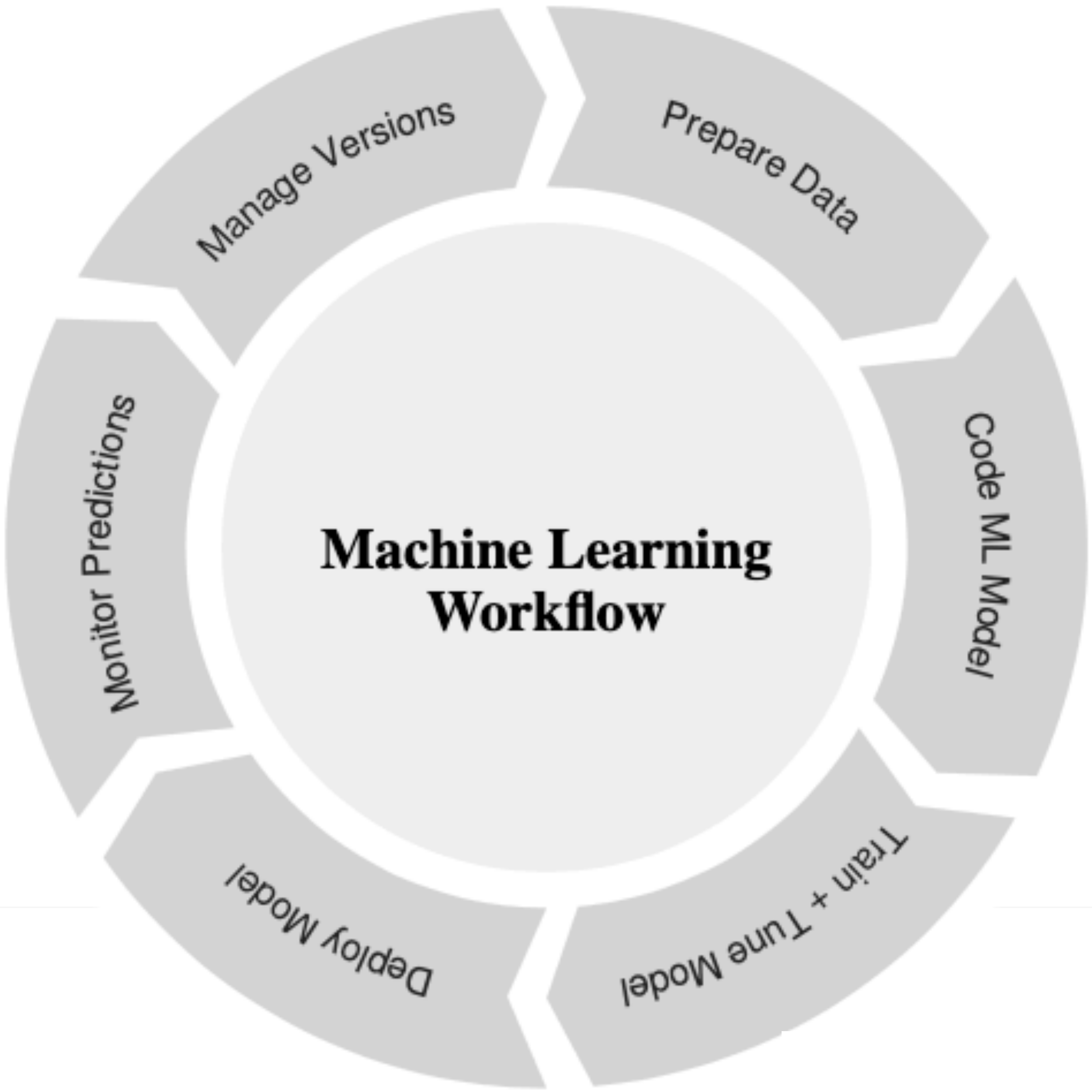
Challenges: all together?



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

Summary of course objectives & content

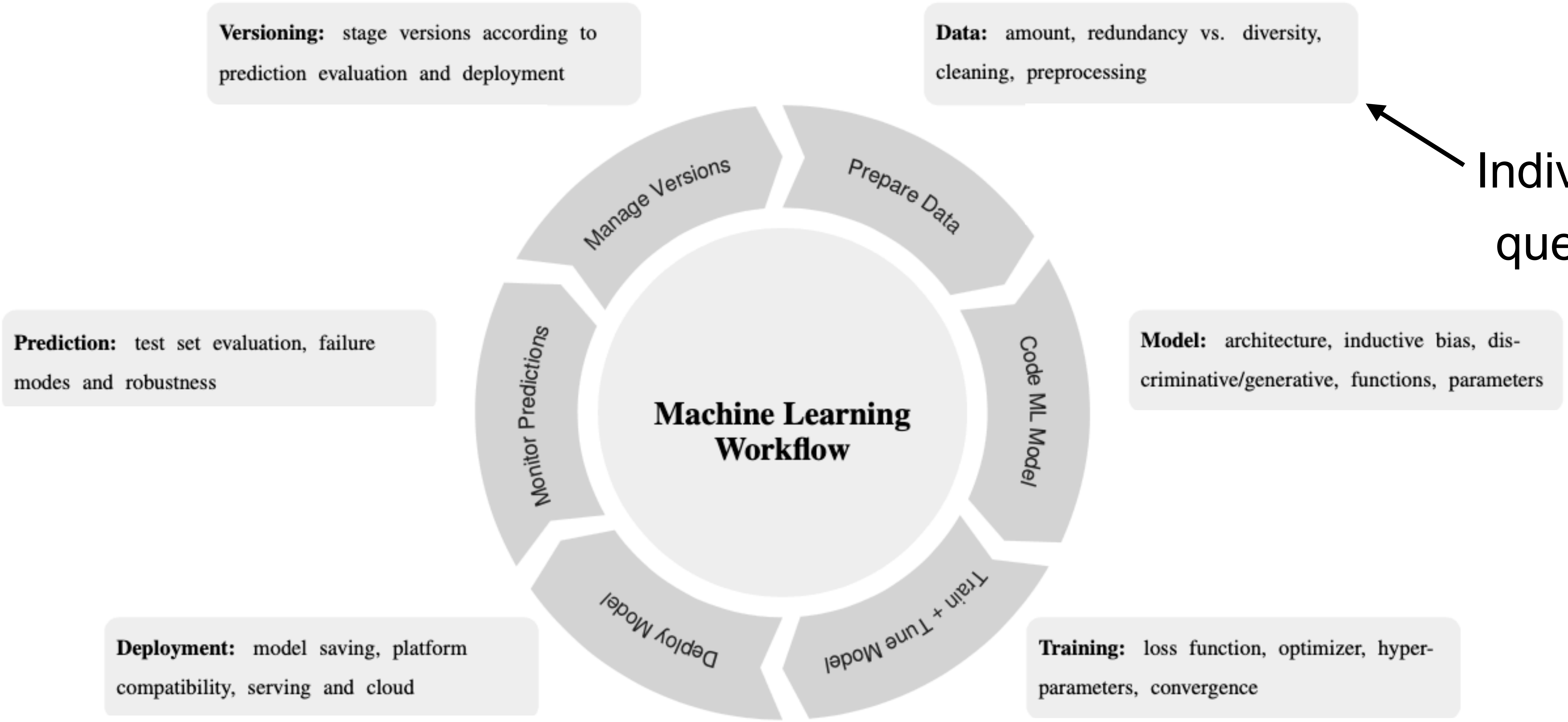
Can we just iterate?



Turns out that this is harder than expected!

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

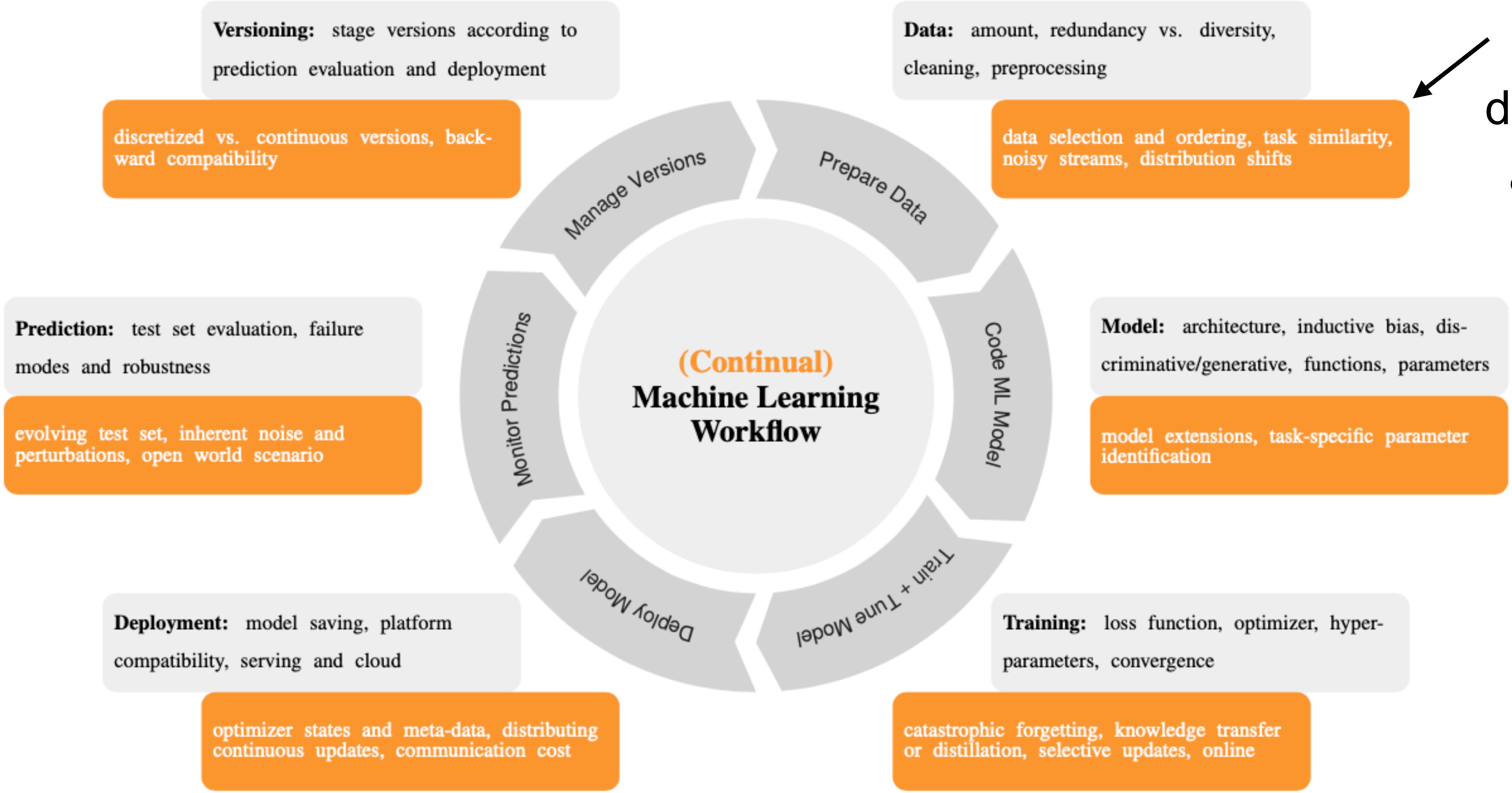
From static ML workflow …



Individual questions

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

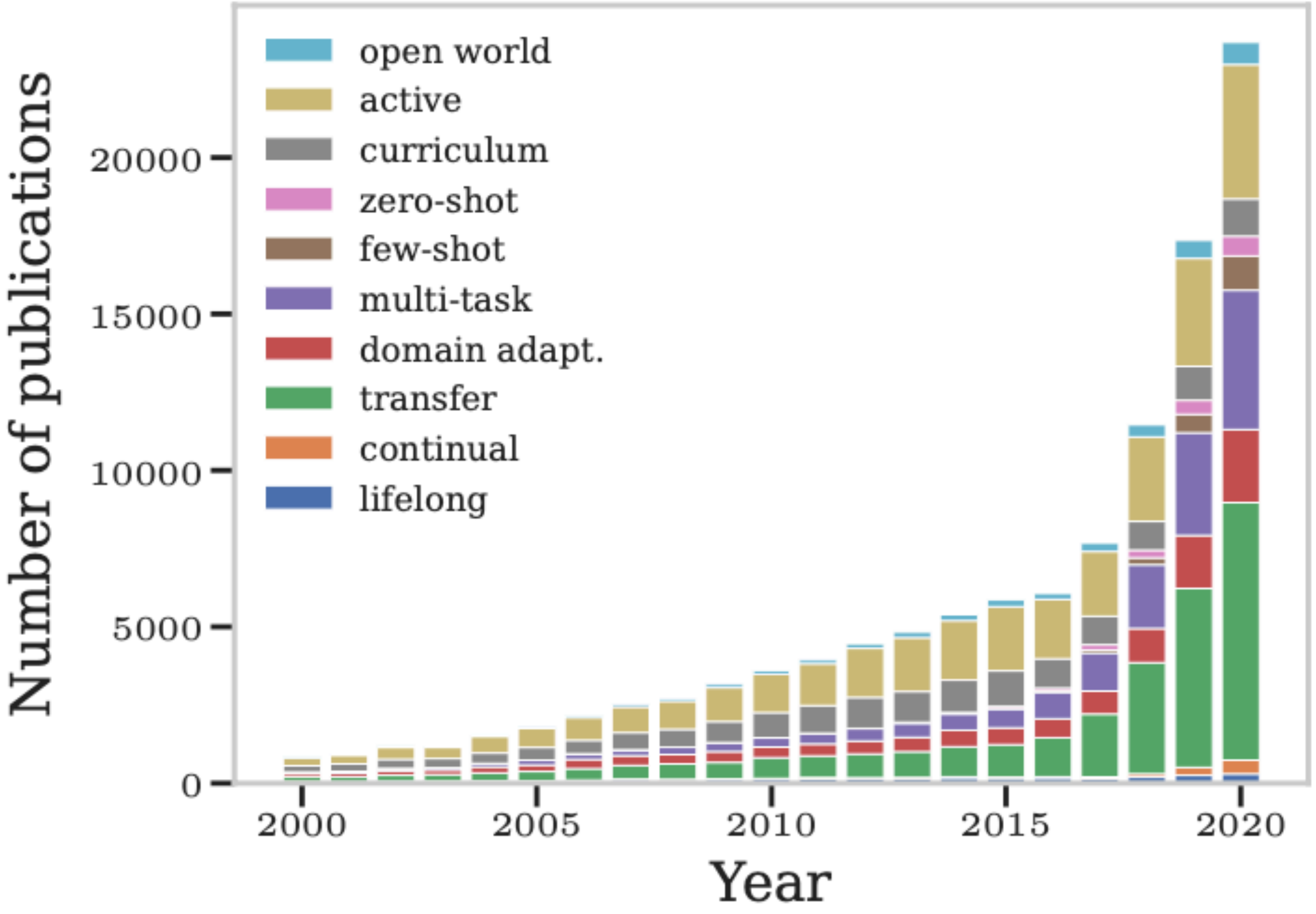
… to continual ML …



Continual dependencies & synergies

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

to dependencies & synergies



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

We try to gain understanding in this course