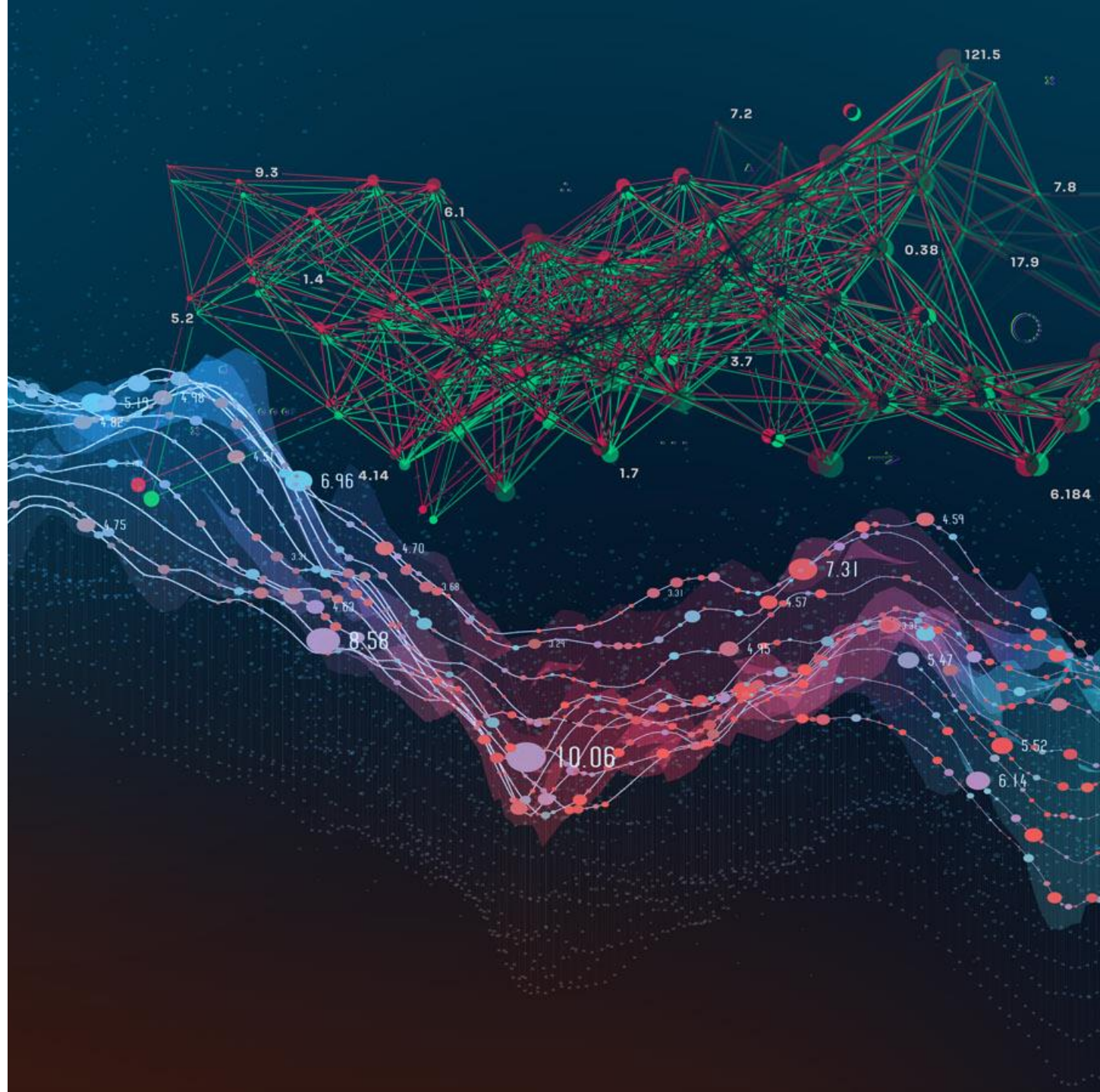


# Day 1 – Intro to Few-Shot Learning



# Welcome!

- Lauren Phillips
  - Data Scientist - PhD Psychology (UC Irvine)
    - Few-shot learning
    - NLP
- Zach New
  - Data Scientist - MS Mathematics (WWU)
    - Computer vision
    - Interpretability and Generalization
- Kayla Duskin
  - Data Scientist - BS Applied Math, Spanish (WWU)
    - Computer Vision
    - Natural Language Processing
    - Graphs



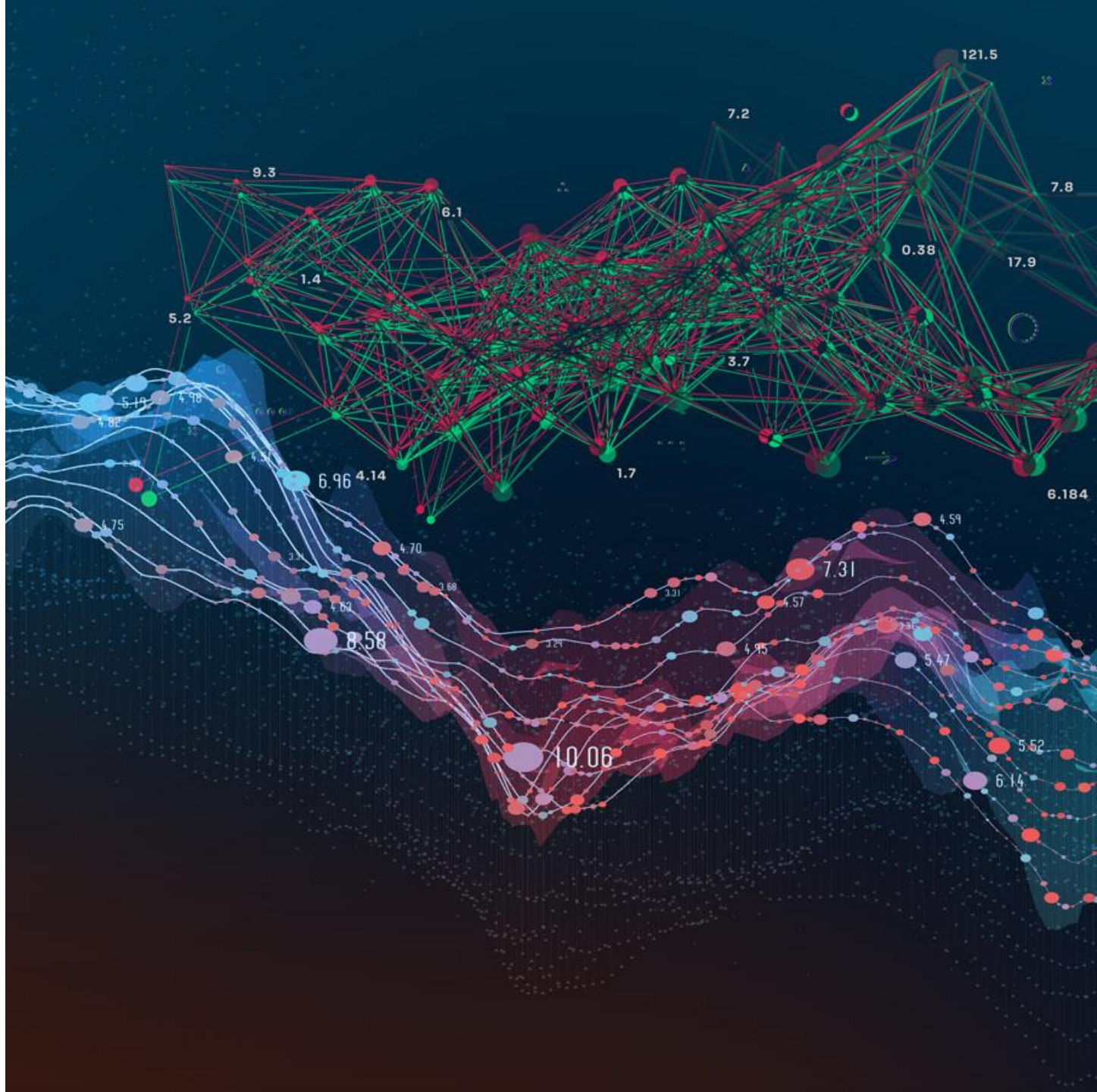


# Day 1 agenda

- Brief Recap on state of Deep Learning
- What is Few-Shot Learning?
  - What is it?
  - How does it work? (just the basics)
  - And why should I care?
- Applications
  - Sharkzor - Few-shot for image filtering
- Coding Examples
  - Applying few-shot to new data

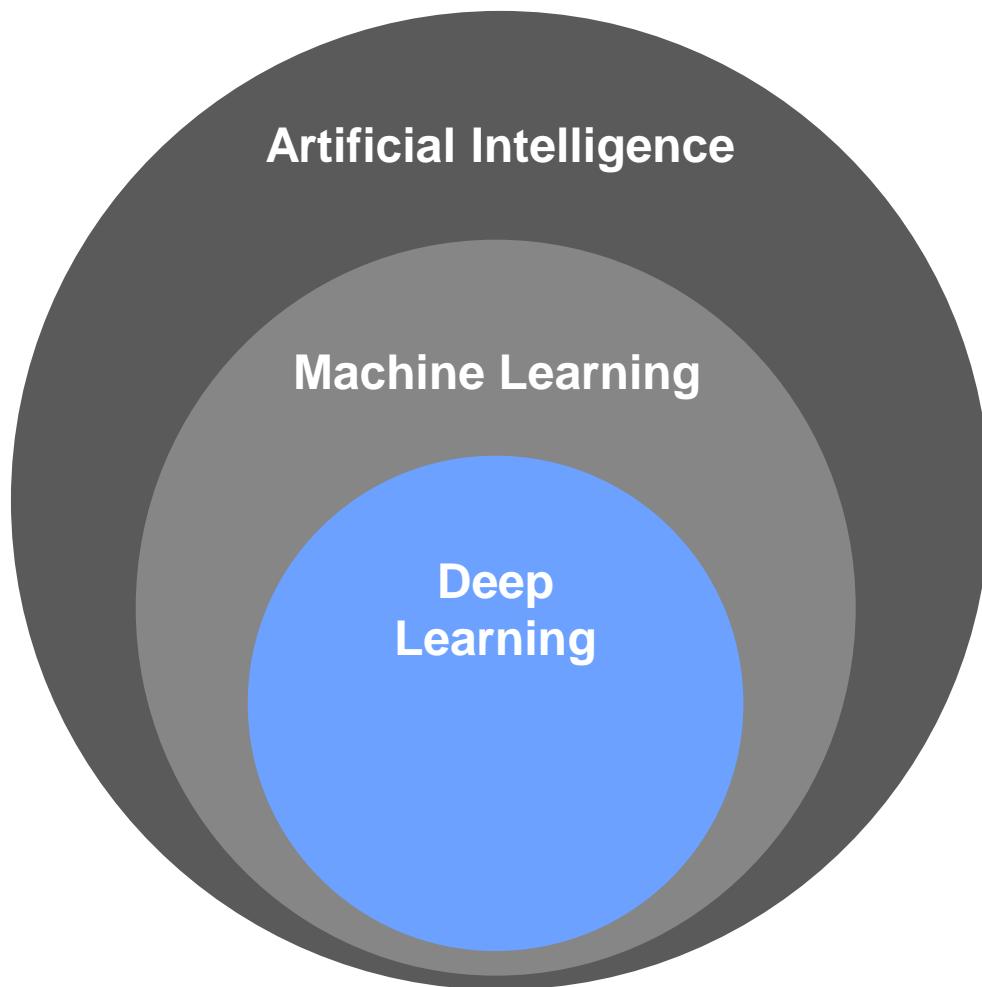
## A sneak peek at Days 2 – 4

- Day 2
  - Implementing few-shot learning
  - Models
  - Training
- Day 3
  - Beyond images - Text, Audio, Video
  - Visual attention / Localization
- Day 4
  - Evaluating few-shot learning
  - Alternatives to few-shot
  - Other topics



PNL is operated by Battelle for the U.S. Department of Energy

# Deep Learning vs. Machine Learning vs. Artificial Intelligence



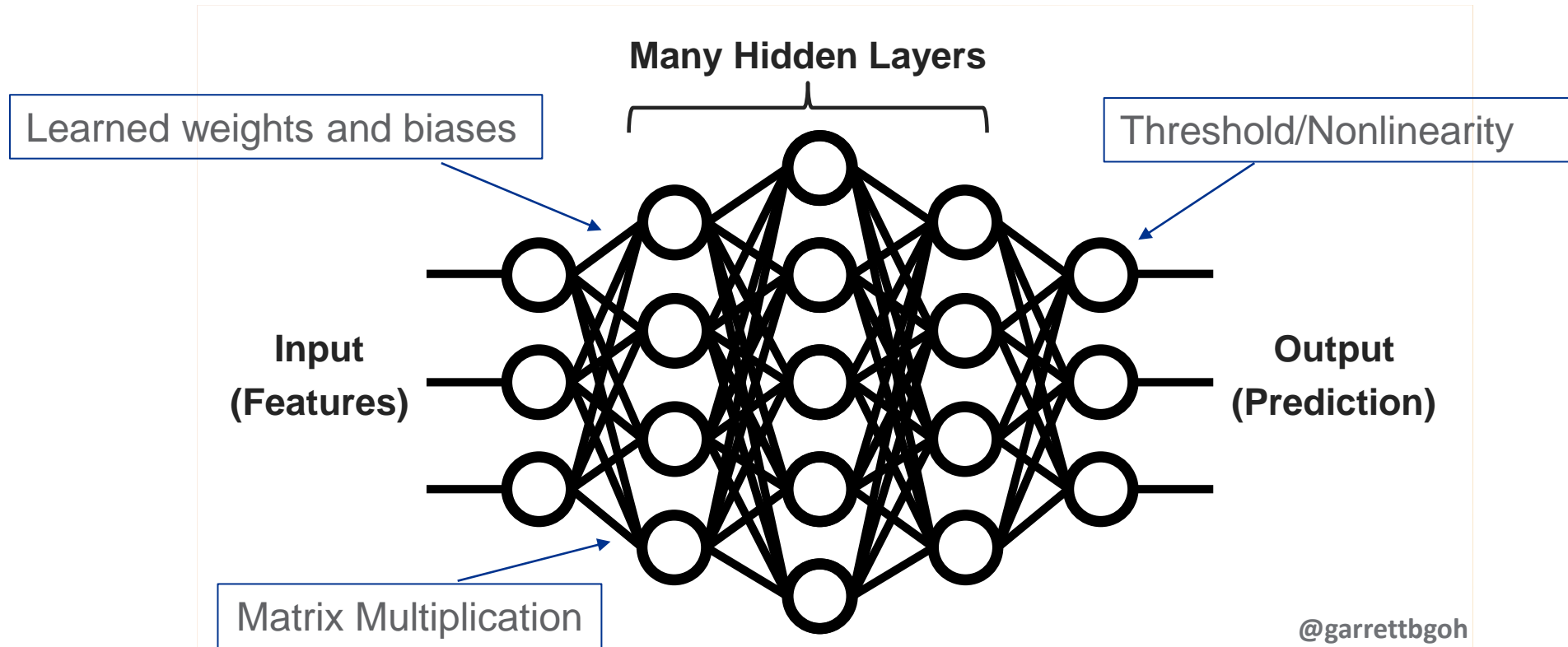
## Deep Learning

- Machine learning is a family of algorithms that improve by learning from data
- A neural network is one of those algorithms
- Neural networks are made up of successive *layers*. Deep learning models have many layers



# What is “Deep Learning”?

- A **Deep** neural network is a **multi-layer** neural network
- This broad definition allows for limitless architectures and applications



@garrettbgoth

<https://arxiv.org/pdf/1706.06689.pdf>

# Types of Neural Network Layers

Network architectures are designed for a particular task. A full model may mix and match different components/layers

- **Linear/Dense/Fully-connected**
  - Unordered data
- **Convolutional**
  - Image processing
  - Spatial data
  - Sequential data
- **Recurrent**
  - Sequential data
- **Attention**
  - Sequential data
  - Image data
  - Anything



# Learning Paradigms

## Supervised Learning

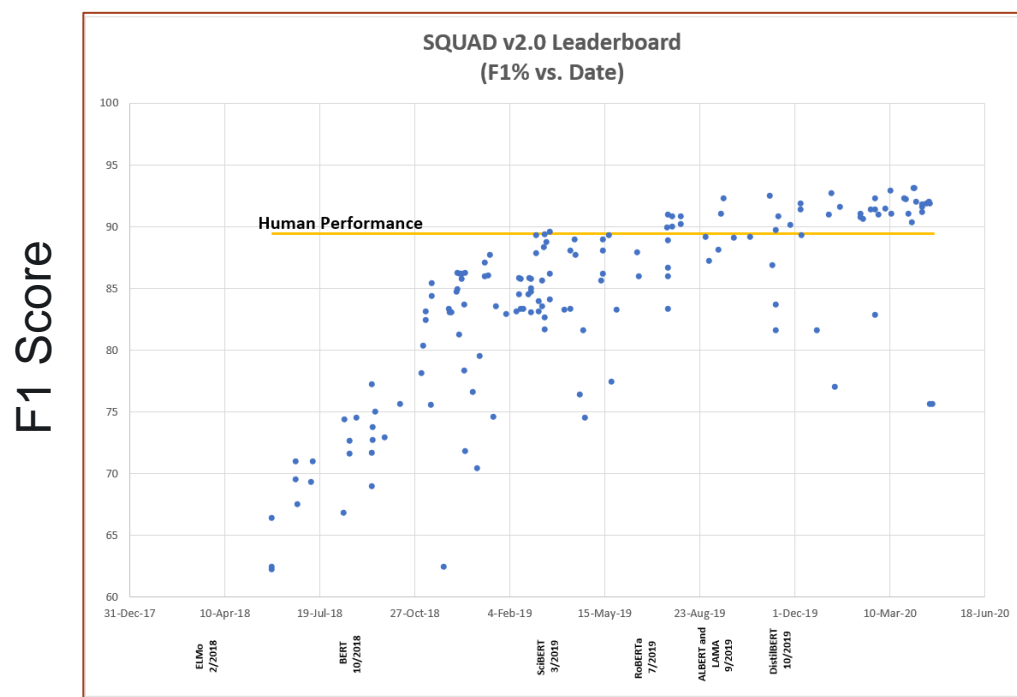
- Requires explicitly labeled data points
- Examples:
  - Image classification
  - Object Detection
  - Text sentiment classification
- Assumption that training data is *in-distribution*

## Self-supervised Learning

- Usually, unlabeled data is leveraged to create a supervised task
- Examples:
  - Autoencoders
  - Clustering algorithms
  - Masked language modeling
  - Generative Adversarial Networks
- Still assumed that training data is representative of the full data distribution

# Why Deep Learning?

- Deep learning with large data has surpassed human performance on tasks in many domains



Text-based Question Answering

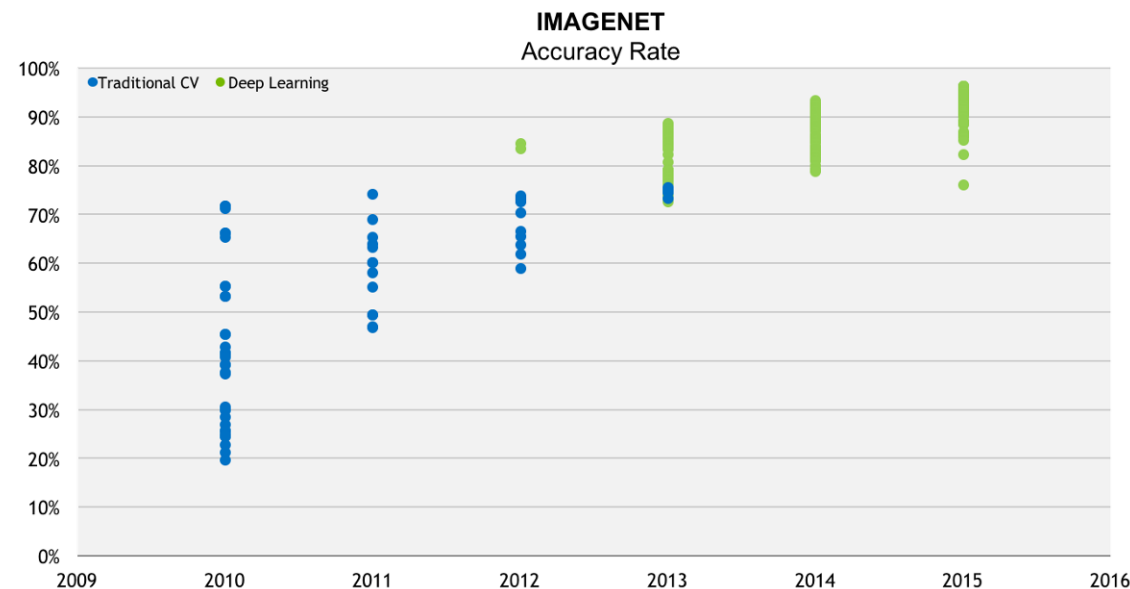


Image Classification

# Why Deep Learning?

Microsoft said Tuesday the court will find Microsoft's goal.

"We remain confident that once all the facts are presented in the forefront of helping developers have the option of taking advantage of Windows features when writing software Developers use the Java Compatibility Logo on its packaging and websites for Internet Explorer and Software using the best tools and provide them the tools they need to write cutting-edge applications. The company would comply with this decision, but we will immediately comply with this decision, but we will immediately comply with a preliminary ruling by Federal District Court Judge Ronald H. Whyte that Microsoft is no longer able to use the Java language," added Tod Nielsen, General Manager for Microsoft's goal.

2000  
Markov Model

<http://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>  
<https://www.cs.princeton.edu/courses/archive/spr05/cos126/assignments/markov.html>

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict. Copyright was the succession of independence in the slop of Syrian influence that was a famous German movement based on a more popular servicious, non-doctrinal and sexual power post. Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]]

2015  
Recurrent Neural Network

<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Deep learning has achieved impressive success in solving complex tasks, and in some cases its learned representations have been shown to match those in the brain [13, 20, 22, 29, 33]. However, there is much debate over how well the backpropagation algorithm commonly used in deep learning resembles biological learning algorithms. Recent studies using different training algorithms have shown the importance of various factors in neural learning, including network depth, choice of activation functions, and randomness in the training set [8, 19, 20, 22]. In this paper we focus on how feedback connections (i.e. feedback to previously visited layers) interact with the backpropagation learning algorithm. We have found that in most cases training without such connections fails to learn when using backpropagation. To illustrate this, we demonstrate that networks employing both feedforward and feedback connections but no learning can produce a surprisingly similar error curve when using gradient descent for learning, but fail to converge at the same point. This is not a general failure of gradient descent to produce the expected error curve, because both shallow and deep networks employing such connections have an error curve similar to that obtained by backpropagation...

2020  
GPT3

<https://arxiv.org/abs/2005.14165>  
<https://www.gwern.net/GPT-3#arxiv-paper>



# What can we do with Deep Learning?

- Unique architectures and training procedures have enabled progress on a huge variety of difficult tasks
  - Machine Translation
  - Speech Recognition (and understanding)
  - Image Classification
  - Question Answering systems
  - Autonomous Systems
  - Natural Language Processing
  - Game Playing
  - Generative Modeling



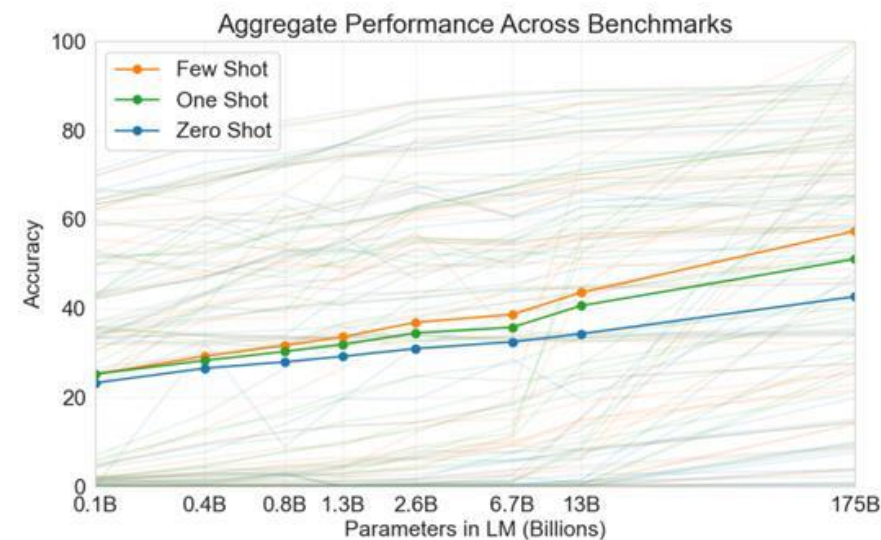
# Data Requirements

- Neural networks are notoriously data hungry
- The creation of large benchmark data sets like Imagenet helped prove the value of deep learning (1.2 million images, 1000 categories)
- Recent years have seen a push toward self-supervised models that leverage huge unlabeled datasets.
- Natural language processing is leading the charge in leveraging big data
  - GPT2 saw breakthrough performance with their novel data collection strategy that resulted in the WebText dataset (8 million documents, 40GB)
  - GPT3 produces even better results using an even larger dataset that includes a filtered version of Common Crawl (570GB)
- While large datasets deliver results, there are trade offs
  - Bias is rampant in “in-the-wild” data
  - Proprietary datasets inhibit collaboration

# Model Size

- Model sizes have increased in tandem with dataset sizes
- Total parameters in state-of-the-art models have crossed from millions into billions
- Language models dominate in model size, but image models also benefit from more parameters
- Model size can be prohibitive in deploying models and in some cases feels like an arms race for tech giants

## GPT3

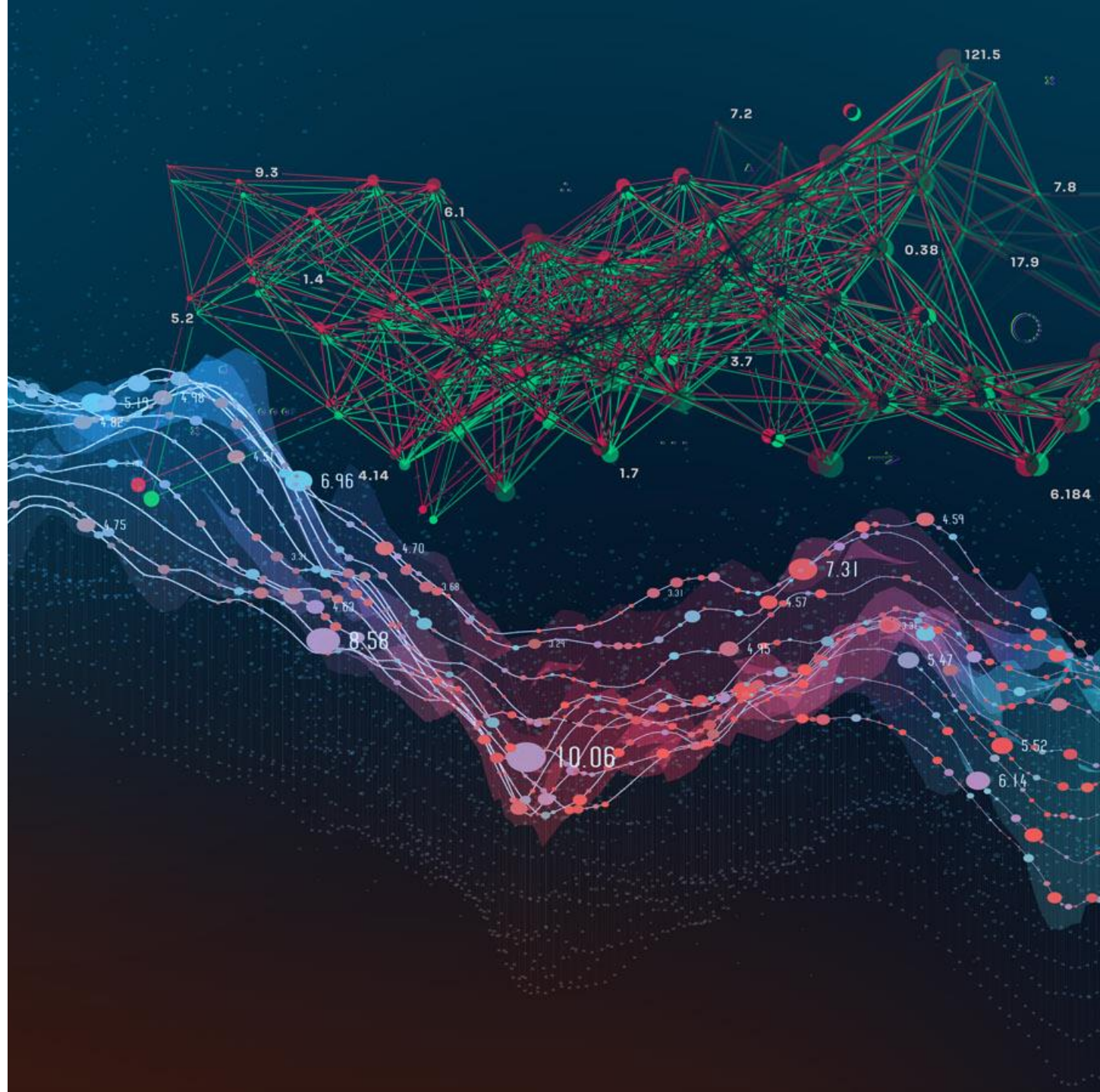


**Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks** While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

<https://arxiv.org/abs/2005.14165>



# What is Few-Shot Learning?



# Deep Learning for Small Data

- Deep learning efficiently leverages large quantities of data to learn representations of classes
- Deep learning *can* be applied to small data
  - ... but restrictions *will* apply ...
- But what if the class you care about has few examples?
  - Ignore differences and just apply existing model (e.g. information retrieval)
  - Adapt a pre-trained models (e.g. transfer learning)
  - Train model to generalize to new classes (e.g. ***few-shot learning***)

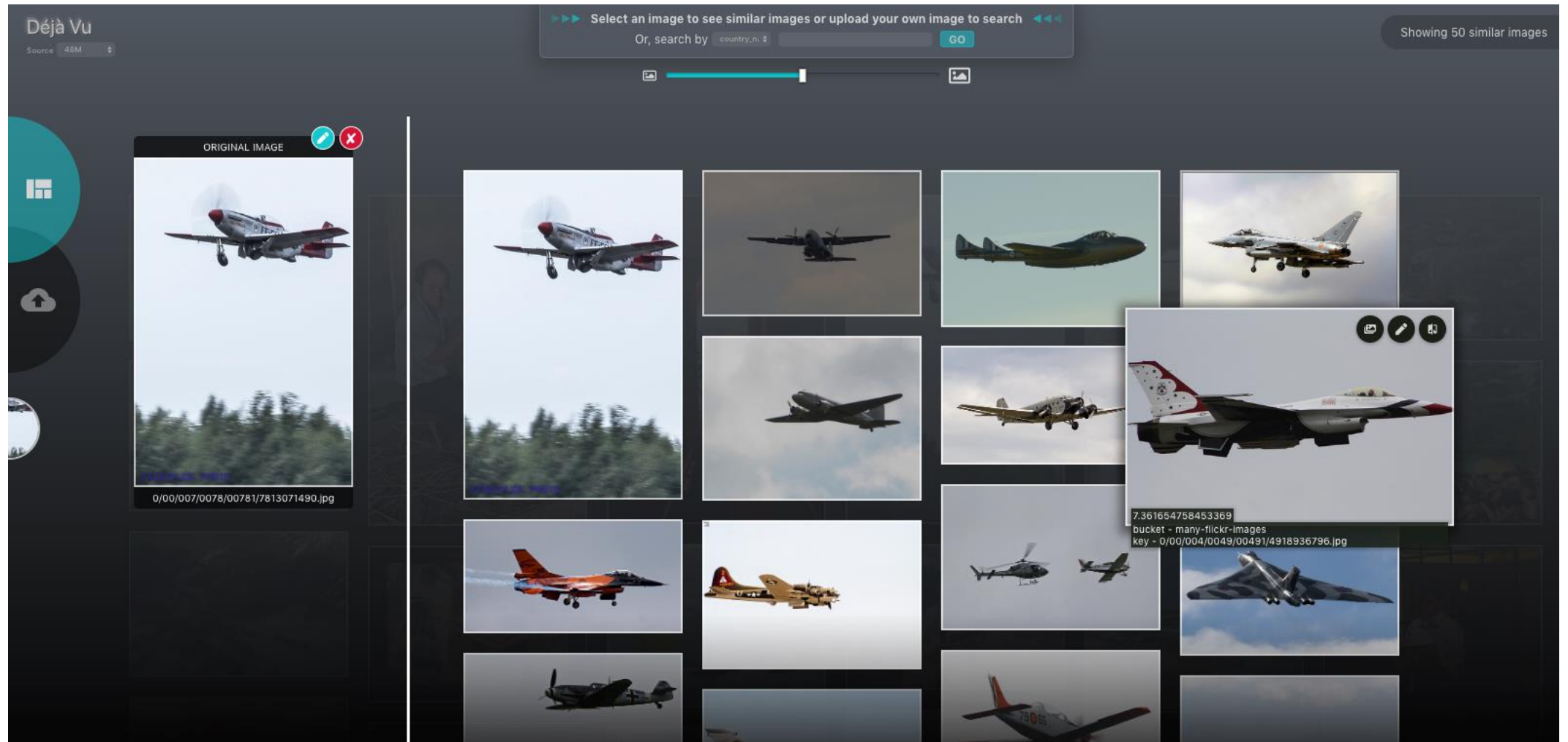
# What even is small data?

- Not all small data is created equal. What is the restriction?
  - How many examples (of the class of interest) exist in the data?
  - How many examples are labeled?
  - Do classes of interest change frequently?

Quantity	# Labels	# Data
Supervised	High	High
Semi-supervised	Low	High
Self-supervised	---	High
Transfer learning	Low/Medium	Low/Medium
Few-shot	Very low	Very low

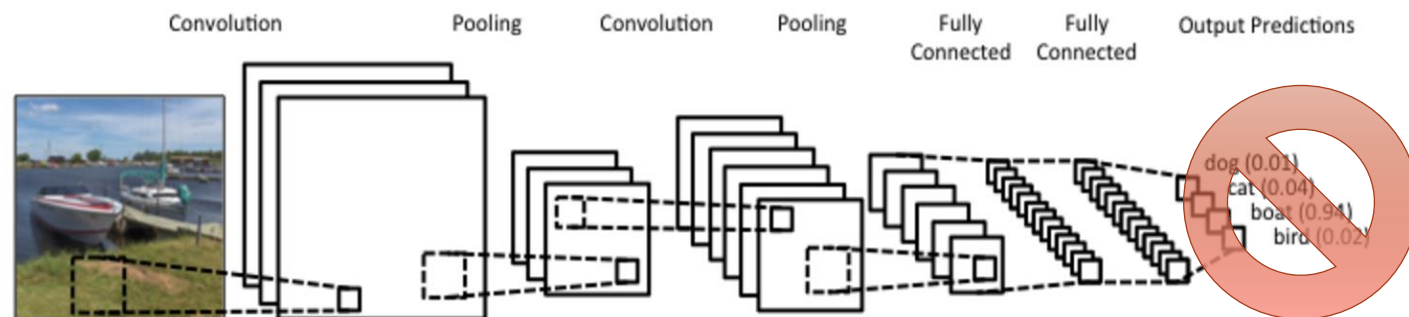


# Apply pre-trained model / Information Retrieval



# Transfer Learning

- For more difficult cases, fine-tuning can improve model performance
  - Supports prediction of new classes by replacing final classification layers
  - Or predict the same classes but on a different type of data
- Improvement typically scales with # of examples
  - Can still be usefully applied on images for 20 examples per class (Dhillon et al., 2020)
  - Retraining needs to occur whenever new classes appear



Get rid of this!

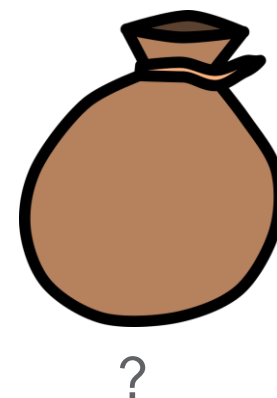
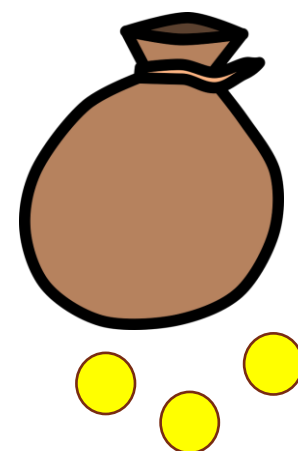
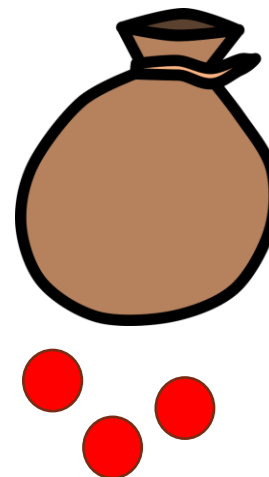
# Few-Shot Learning – Defining Goals

- From just a handful of examples (1 – 5), can we identify other examples that belong to the same class?
- To make this possible, we have to leverage similarities between classes
- Closely related to zero-shot learning, where each class is annotated with “attributes”
  - Model learns relationship between attributes and the class, allowing it to make predictions for previously unseen combinations of attributes



## An Interlude – Of Marbles

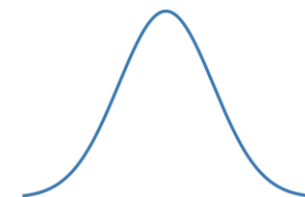
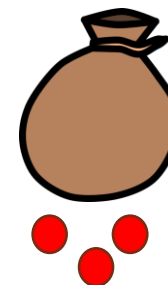
1. You find a bag of marbles, you can't see inside. If you grab a marble, what color will it be?
2. You grab three marbles one at a time, and each is red. What color will the next be?
3. You find another bag, pull out a marble and it's yellow. What color will the next be?
4. You find a third bag, what can you say about the marbles in it?



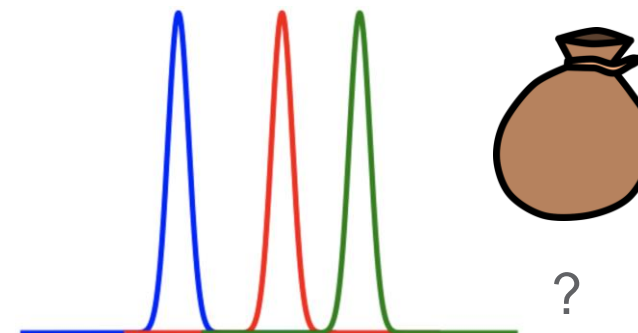
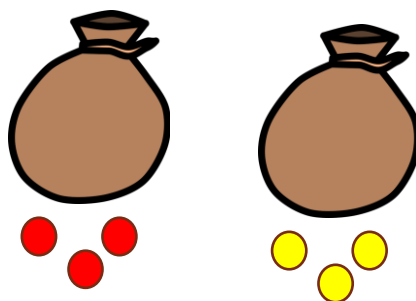
## An Interlude – Of Marbles

- What can we learn from bags of marbles?
  - For each bag, we form beliefs about the distribution of colors
  - But we also form beliefs about *all* of the bags together

- Learning from just one bag is like fitting a single Gaussian

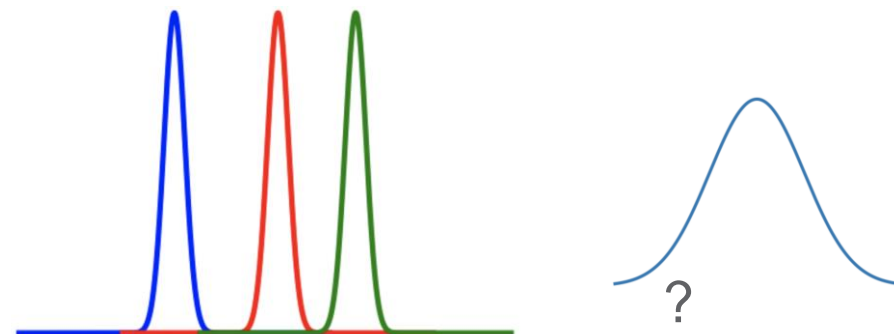


- But if we learn from all bags, we can better generalize to a new Gaussian with fewer examples



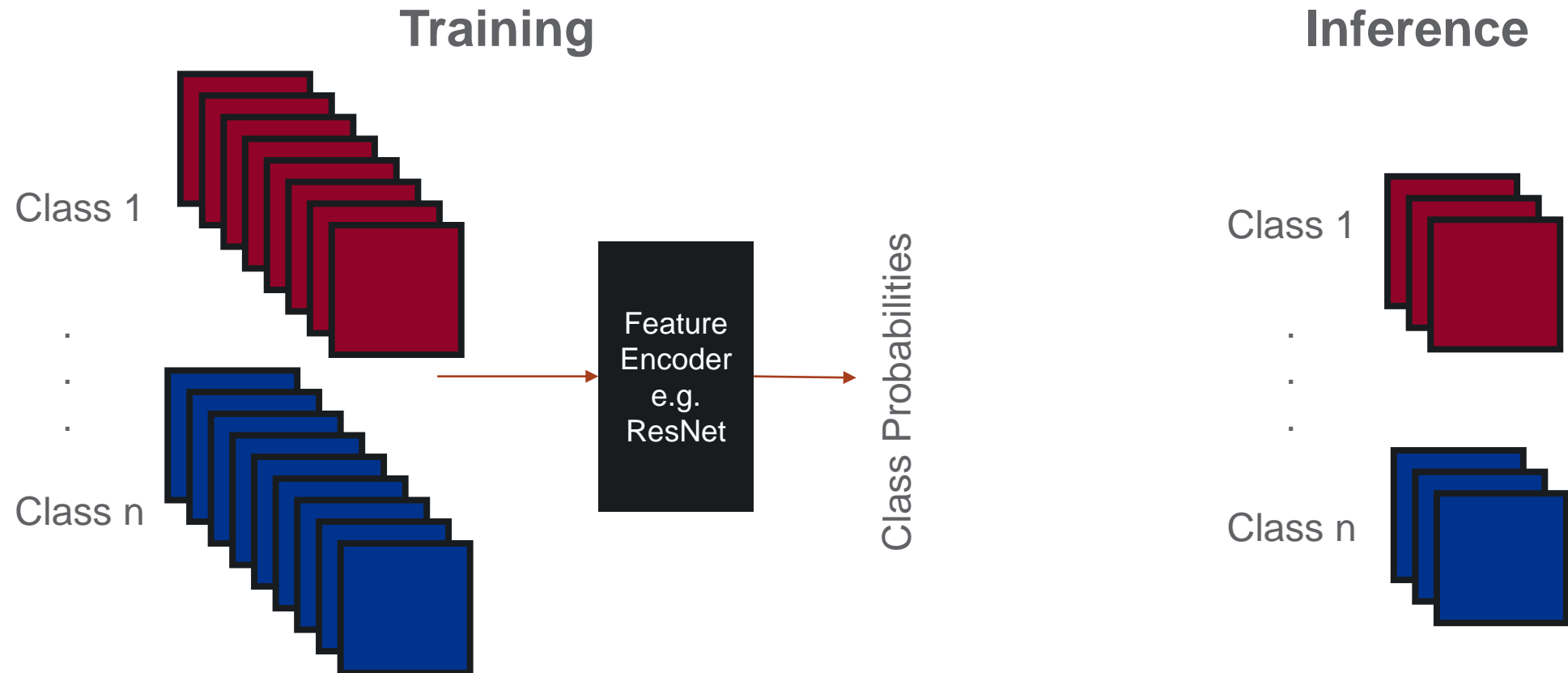
# Few-Shot Learning

- Few-shot learning *is NOT magic*
- Few-shot learning *is* an alternate way to train models which focuses on generalization
  - We're learning how to fit a distribution of distributions (hopefully!)
- Rather than fitting individual classes, we explicitly train the model to generalize to a new class
  - From the same distribution! (important caveat)



# Traditional Classification

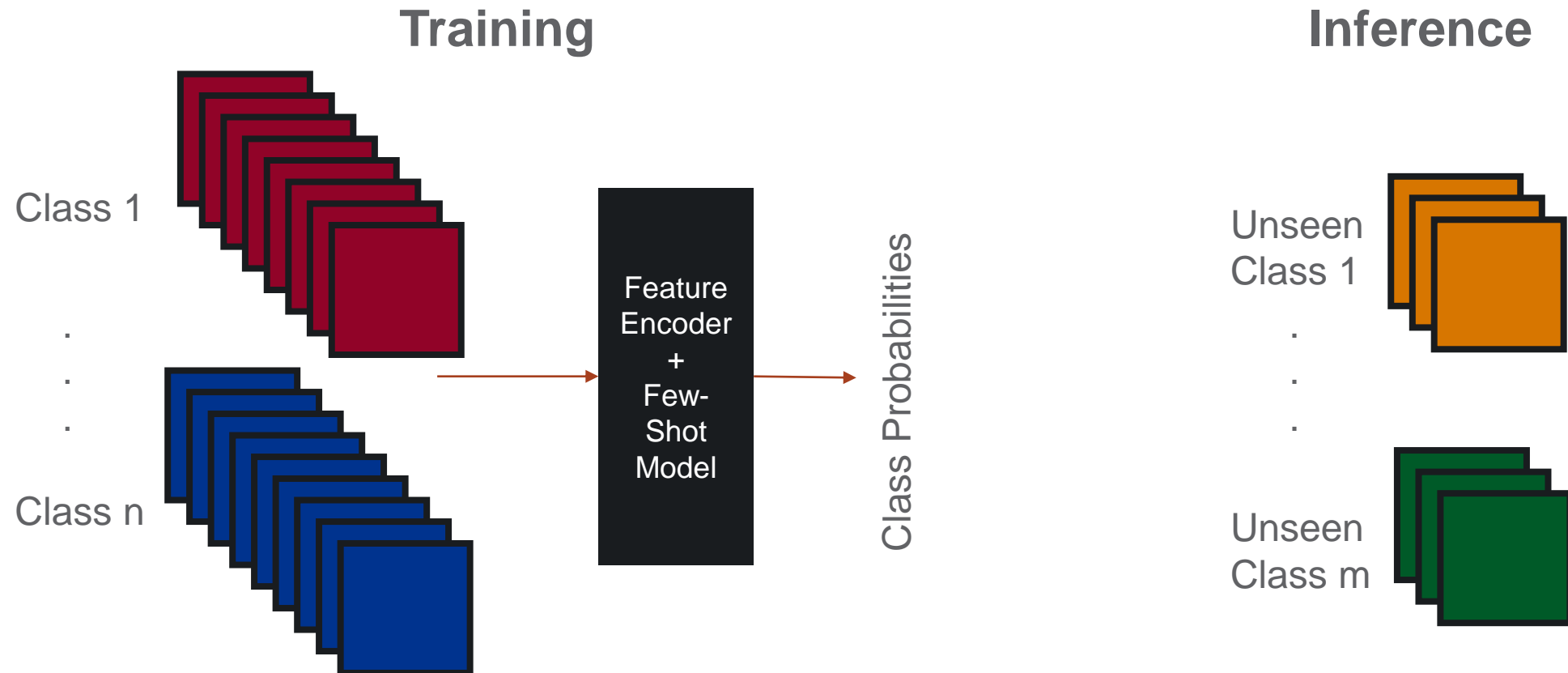
Focus on classification of specific classes with large datasets.



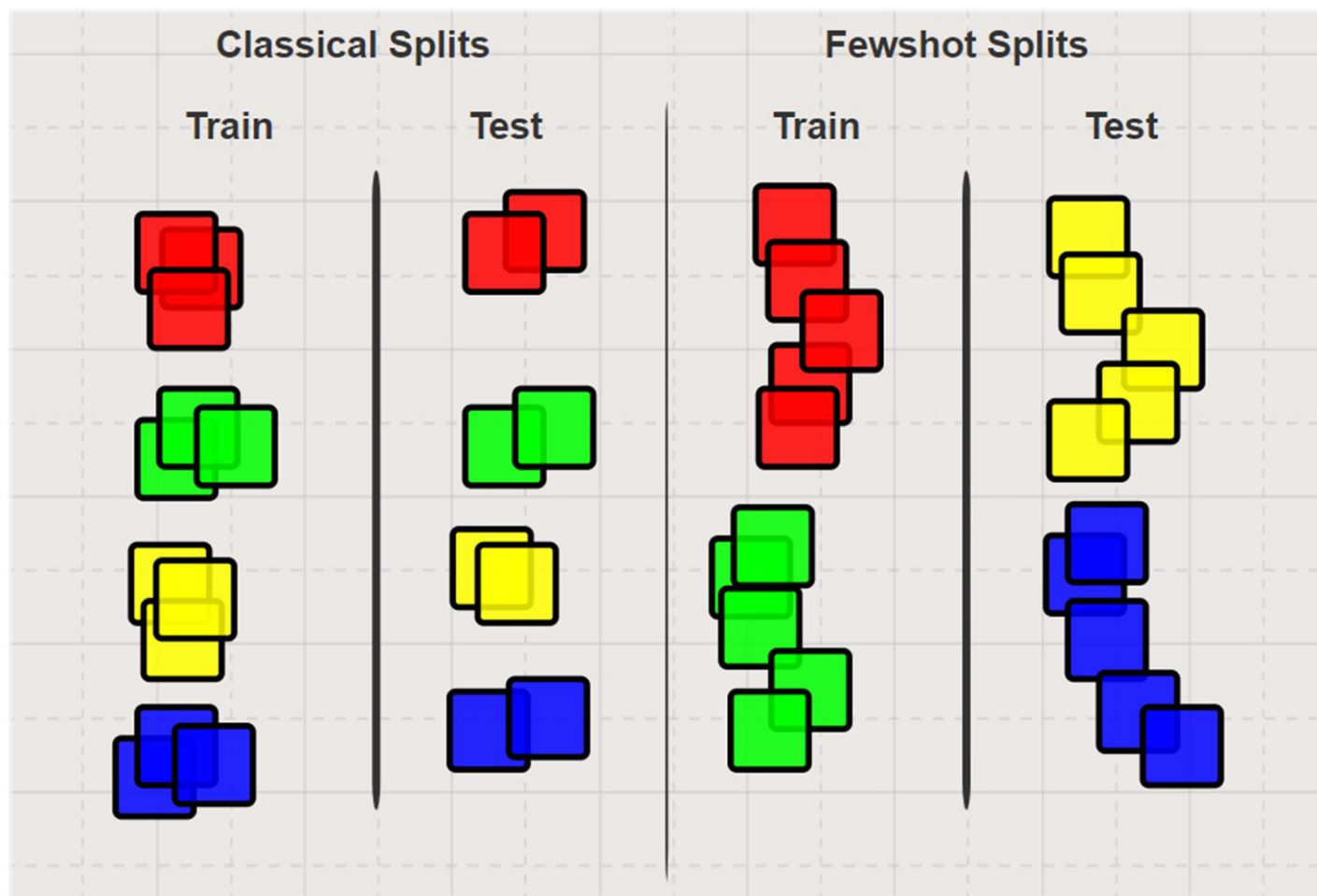


# Few-Shot Classification

Focus on comparison of arbitrary classes with small datasets.

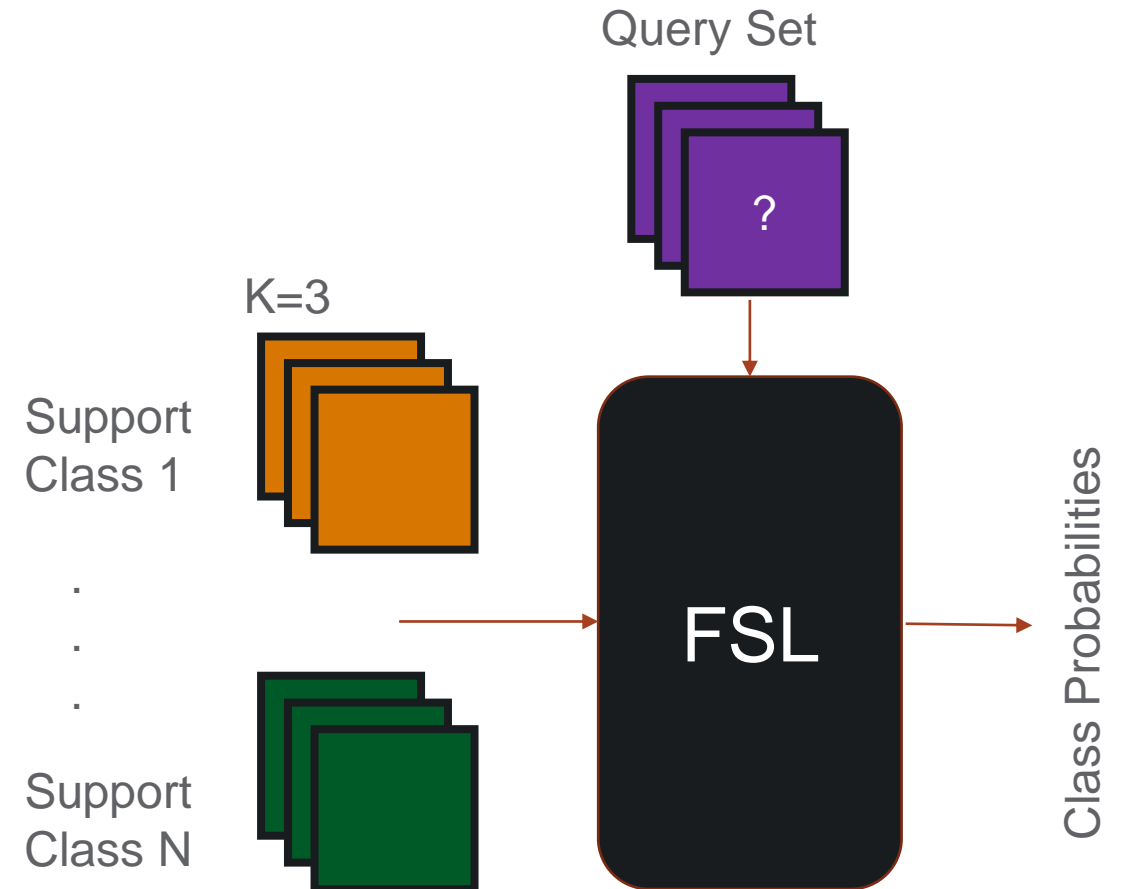


# Few-Shot Splits



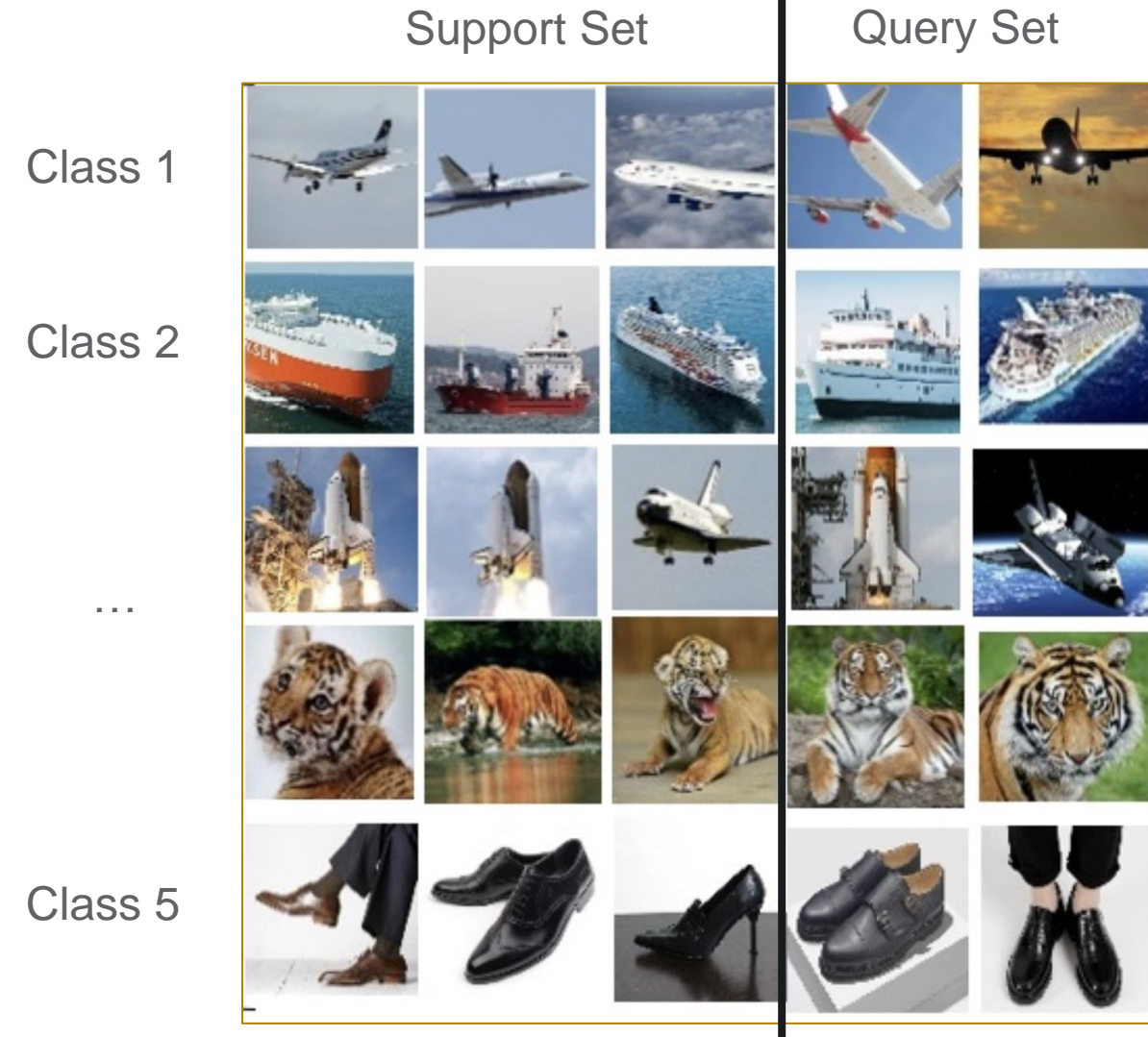
# Few-Shot Episodes

- “Support set” defines new classes
  - $N$ -way = number of new classes
  - $K$ -shot = number of examples per class
- “Query set” are unlabeled data
- Each combination of support and query makes up an “episode”



# Few-shot Episodes

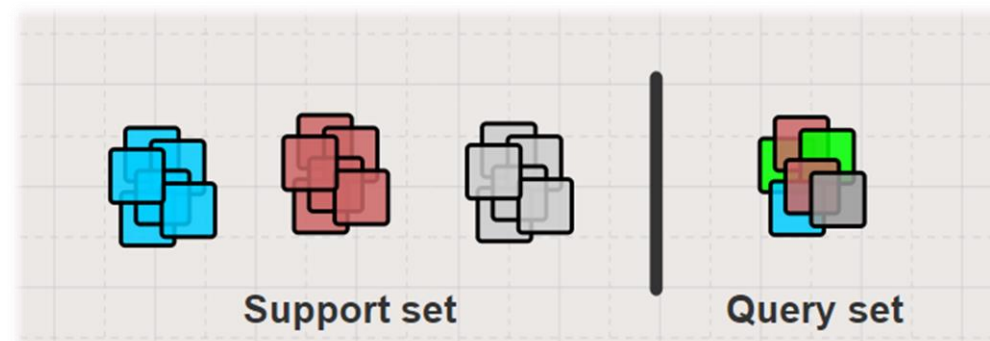
- To train, we sample episodes with  $K$  classes and  $N$  support examples per class
- $N$ -shot,  $K$ -way
  - e.g. in this example, the episode is 3-shot, 5-way
- As opposed to standard labels which are fixed, class #s here are arbitrary and change from episode to episode





# Few-shot Training

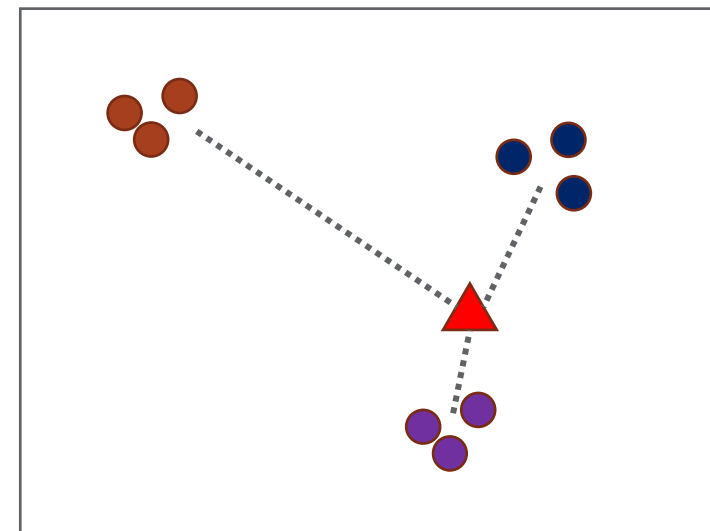
- “Classical” Training
  - Train encoder on large data task, e.g. ImageNet
  - Slow, requires hundreds of examples per class
- Few-Shot Training
  - Randomly sample episodes instead of batches
  - Slow, requires hundreds of examples per class *during training*
- Specifically training in a way that matches the task at inference



## An example model – Prototypical Networks

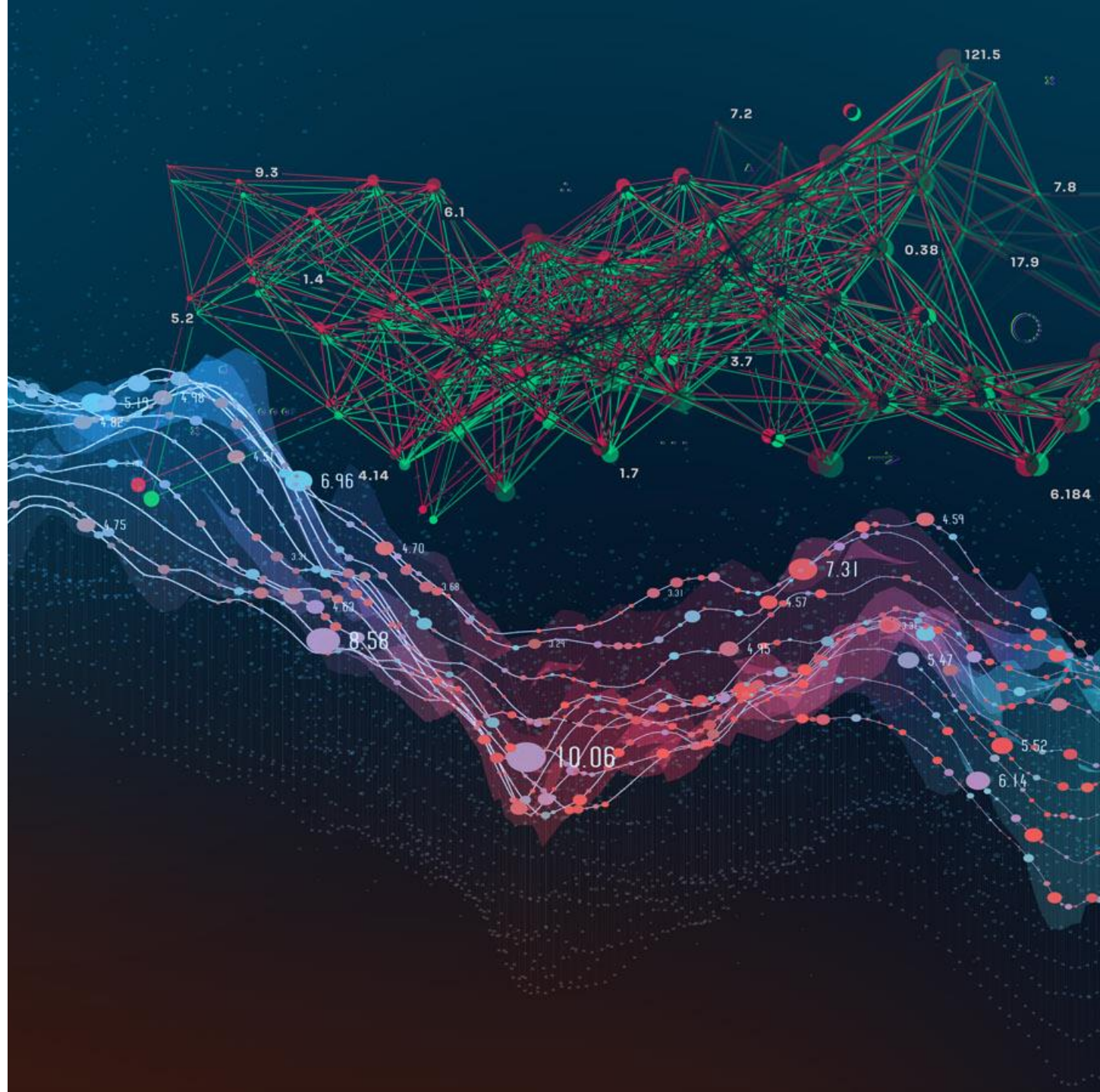
1. Place datapoints into an embedded vector space
2. Each class is represented by a “prototype”, the average of its members
3. Unlabeled examples are placed into the nearest group based on distance to prototypes

- *Metric-based* approach to few-shot (Snell et al., 2017)
  - Model is trained to learn features where class comparison can be done
  - Similar to nearest neighbors, but with improved performance on new classes



- Distance function forces strong inductive bias that helps model generalize

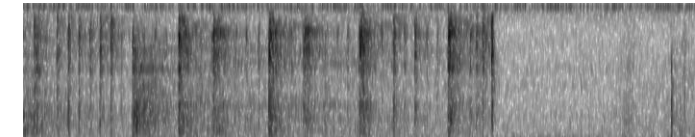
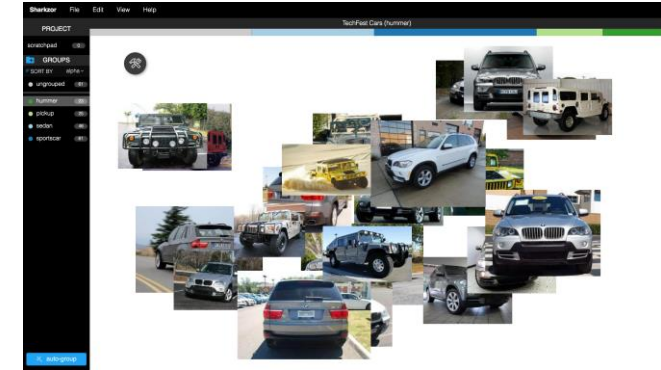
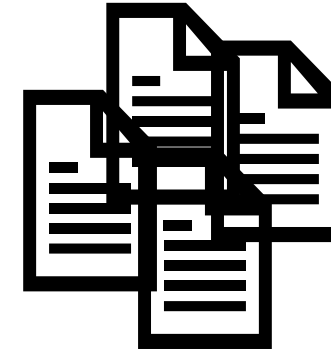
# Few-Shot in action






# Few-shot learning beyond images

- While image datasets are the most common benchmarks, few-shot has been successfully applied to many other data modalities
- Images
  - Image filtering
  - GeoINT
- Text
  - Authorship attribution
- Audio
  - Event detection, speaker identification
- Video
  - Activity recognition



ground truth			
predictions			



# GPT-3 - Language Models are Few-shot Learners

- For large language models, support examples can be provided as context (Brown et al., 2020)
- This trend to ever larger models has been stronger in NLP than other domains
  - Large training datasets have their own concerns, especially around bias

## Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

1	Translate English to French:	← task description
2	sea otter => loutre de mer	← examples
3	peppermint => menthe poivrée	
4	plush girafe => girafe peluche	
5	cheese => .....	← prompt

<https://arxiv.org/pdf/2005.14165.pdf>

# Sharkzor Demo

**SHARKZOR** File Edit View Help

Uranium Forensics (round) [838 images]

Columns: 4 Filter: All Show Masks: 0 IMAGES SELECTED

**GROUPS**

Sort by: Alpha

Ungrouped	77
rods	19
rough	735
round	7

Auto Group

New Group



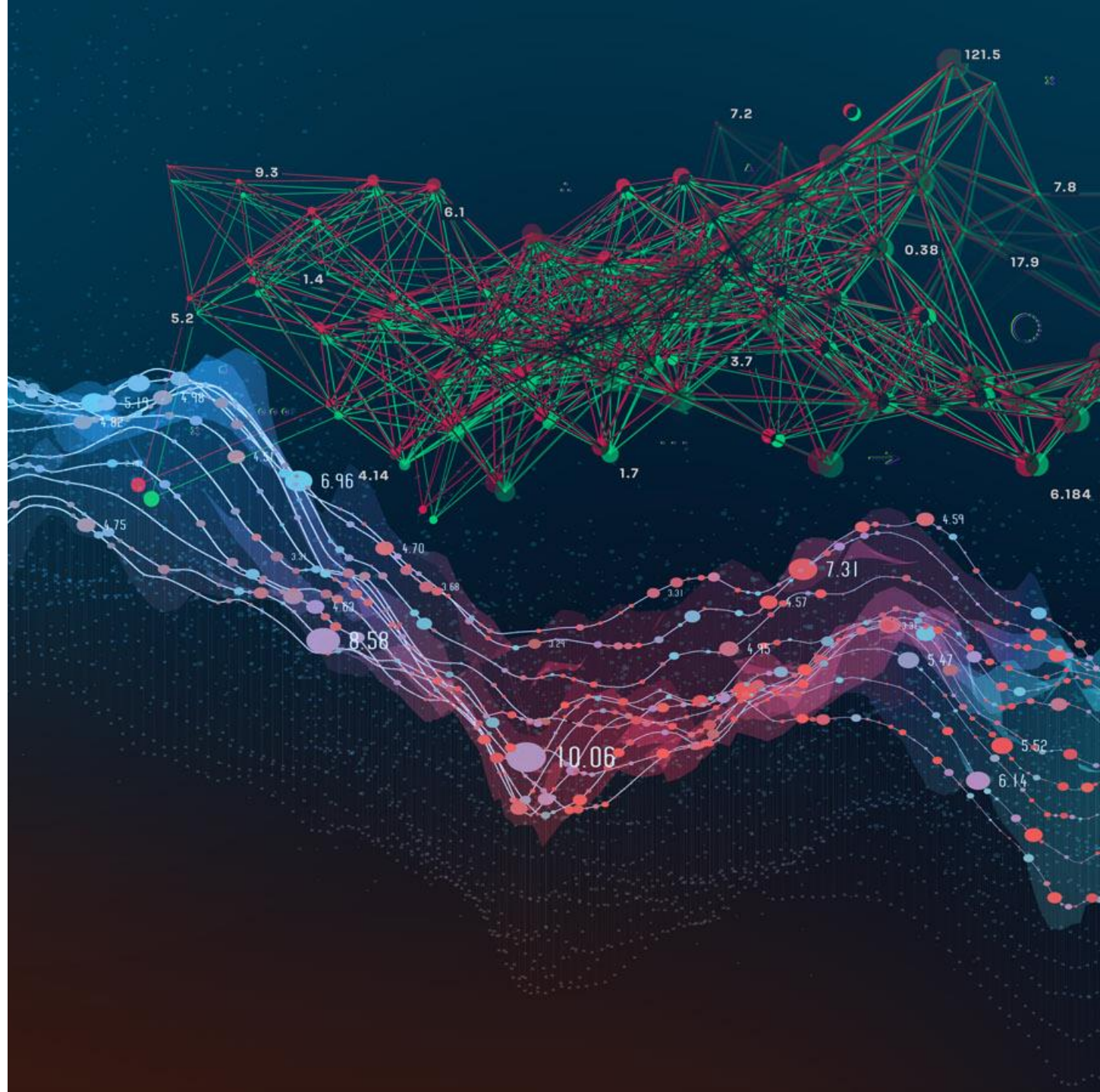




# State of the field



PNNL is operated by Battelle for the U.S. Department of Energy



## A brief history

- Current conception introduced in Lake et al., 2015
- Many algorithms proposed, two major contenders
  - ProtoNets (Snell et al., 2017)
  - MAML (Finn et al., 2017)
- Current trends:
  - Scaling up. Baselines are still focused on small image datasets
  - More work into understanding when few-shot succeeds and fails
  - Extending to real-world problems
    - By 2017 PNNL had developed methods for handling “none-of-the-above” images. This topic is only recently in other academic literature (Gao et al., 2019)
- Workshop on Meta-Learning (at NeurIPS) now in 4<sup>th</sup> year



Where is another?

𐀀	𐀁	𐀂	𐀃	𐀄
𐀅	𐀆	𐀇	𐀈	𐀉
𐀊	𐀋	𐀌	𐀍	𐀎
𐀏	𐀐	𐀑	𐀒	𐀓

Omniglot dataset, from Fig 1.  
Lake et al (2019)