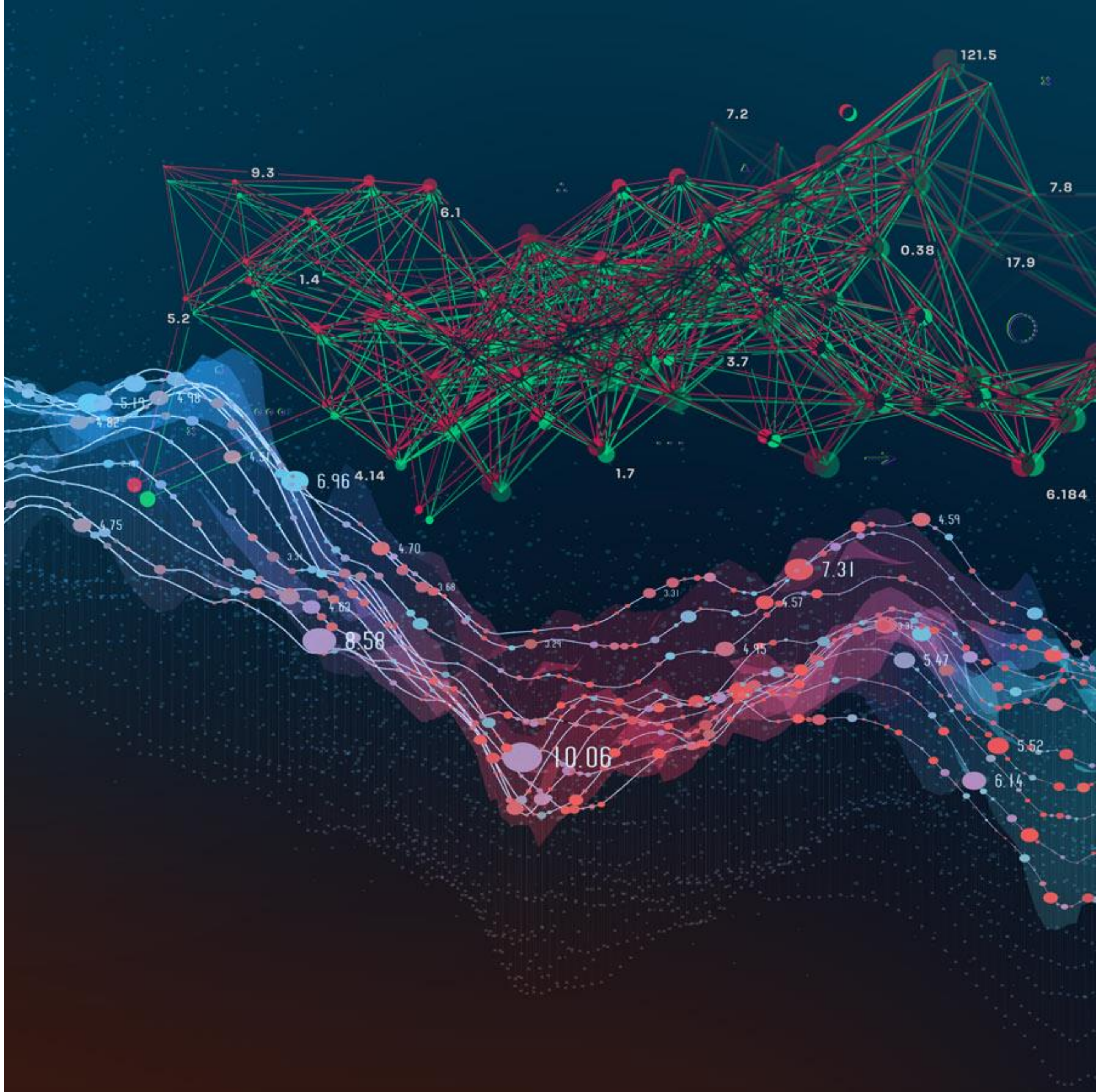
Pacific
Northwest
NATIONAL LABORATORY

Day 4 – Evaluating Few-shot learning

U.S. DEPARTMENT OF ENERGY BATTELLE

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



Day 3 recap

- Few-shot on non-image modalities
    - Modality-specific encoders

- Video
    - Activity recognition
- Audio
    - Audio event detection
- Text
    - Authorship attribution

Pacific Northwest
NATIONAL LABORATORY

Day 4 agenda

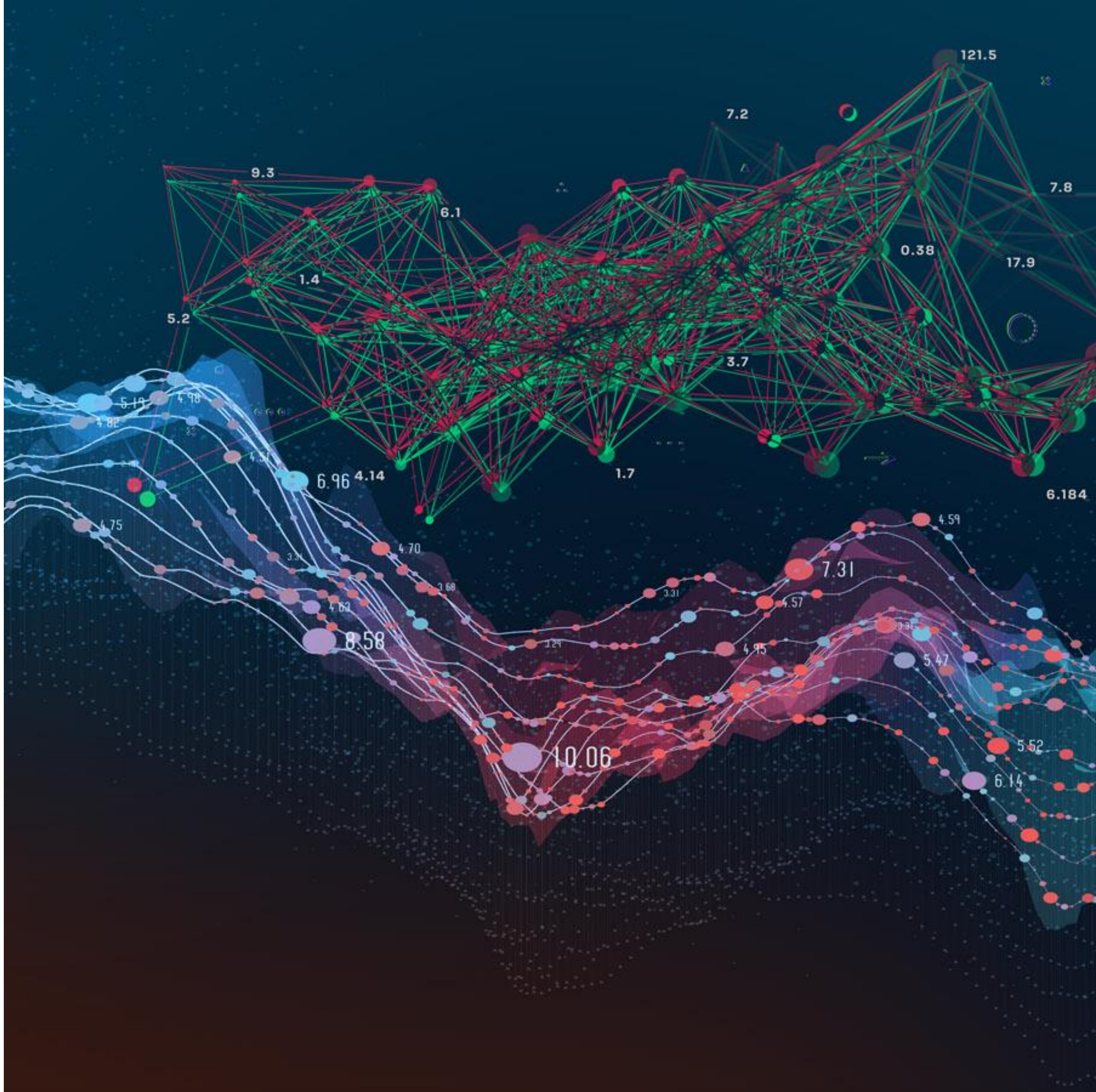
- Baselining few-shot performance
- Alternatives to few-shot learning
- Generalization and better evaluation
- Other considerations around bias
- Practical Examples
  - Model fine-tuning

Pacific Northwest
NATIONAL LABORATORY

Few-shot Baselines

U.S. DEPARTMENT OF ENERGY BATTELLE

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

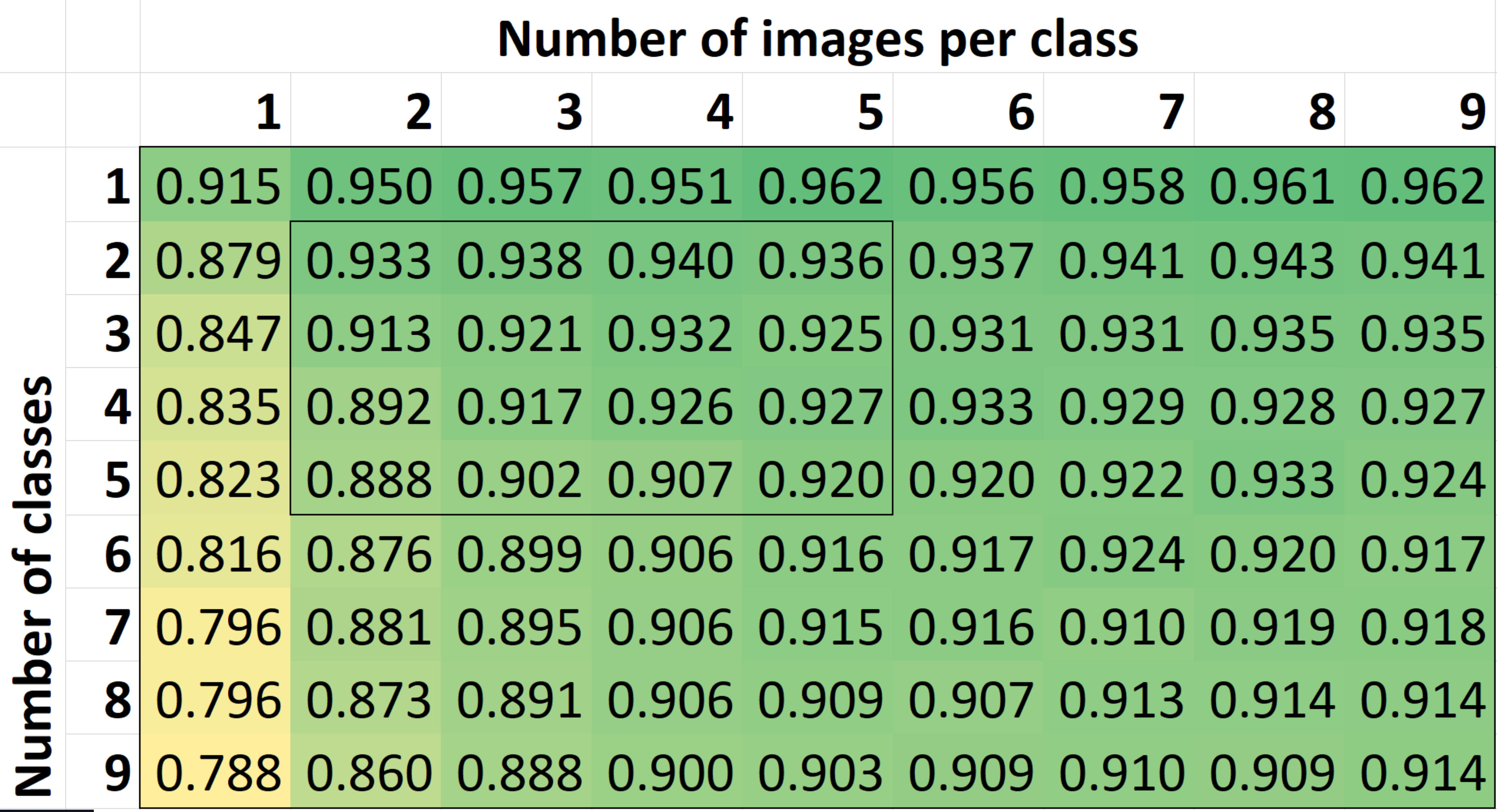


Model Evaluation

- With the way support sets are sampled it wouldn't be good to just pick one and run all you test data against it. With that in mind we average the performance over multiple episodes

- The shot and way define the problem and greatly effect model performance
  - **Shot**: The number of examples in you support set
  - **Way**: The number of classes you are trying to predict for

- A 5-shot 5-way is an easier problem than a 1-shot 10-way problem



		Number of images per class								
		1	2	3	4	5	6	7	8	9
Number of classes	1	0.915	0.950	0.957	0.951	0.962	0.956	0.958	0.961	0.962
	2	0.879	0.933	0.938	0.940	0.936	0.937	0.941	0.943	0.941
	3	0.847	0.913	0.921	0.932	0.925	0.931	0.931	0.935	0.935
	4	0.835	0.892	0.917	0.926	0.927	0.933	0.929	0.928	0.927
	5	0.823	0.888	0.902	0.907	0.920	0.920	0.922	0.933	0.924
	6	0.816	0.876	0.899	0.906	0.916	0.917	0.924	0.920	0.917
	7	0.796	0.881	0.895	0.906	0.915	0.916	0.910	0.919	0.918
	8	0.796	0.873	0.891	0.906	0.909	0.907	0.913	0.914	0.914
	9	0.788	0.860	0.888	0.900	0.903	0.909	0.910	0.909	0.914

Evaluating Few-shot Models

- Throughout the workshop, we've seen a variety of models and datasets. But is there a clear winner?

- Comparing results of different approaches can be difficult for a variety of reasons
  - Different implementations
  - Different datasets / data splits
  - Transductive vs Non-transductive models

- Combined, this makes it quite challenging to know what performance to expect on a new dataset, or which existing method will perform best

MiniImageNet

- As an example, 5-shot, 5-way performance on MiniImageNet is the most reported performance metric in the academic literature
  - There are two train/test splits in common usage…
- Is fine-tuning a good approach?
  - Maybe, but it also uses a bigger encoder…
- Regardless, all of these models are trained on 84x84 pixel images
  - Not useful for real-world tasks

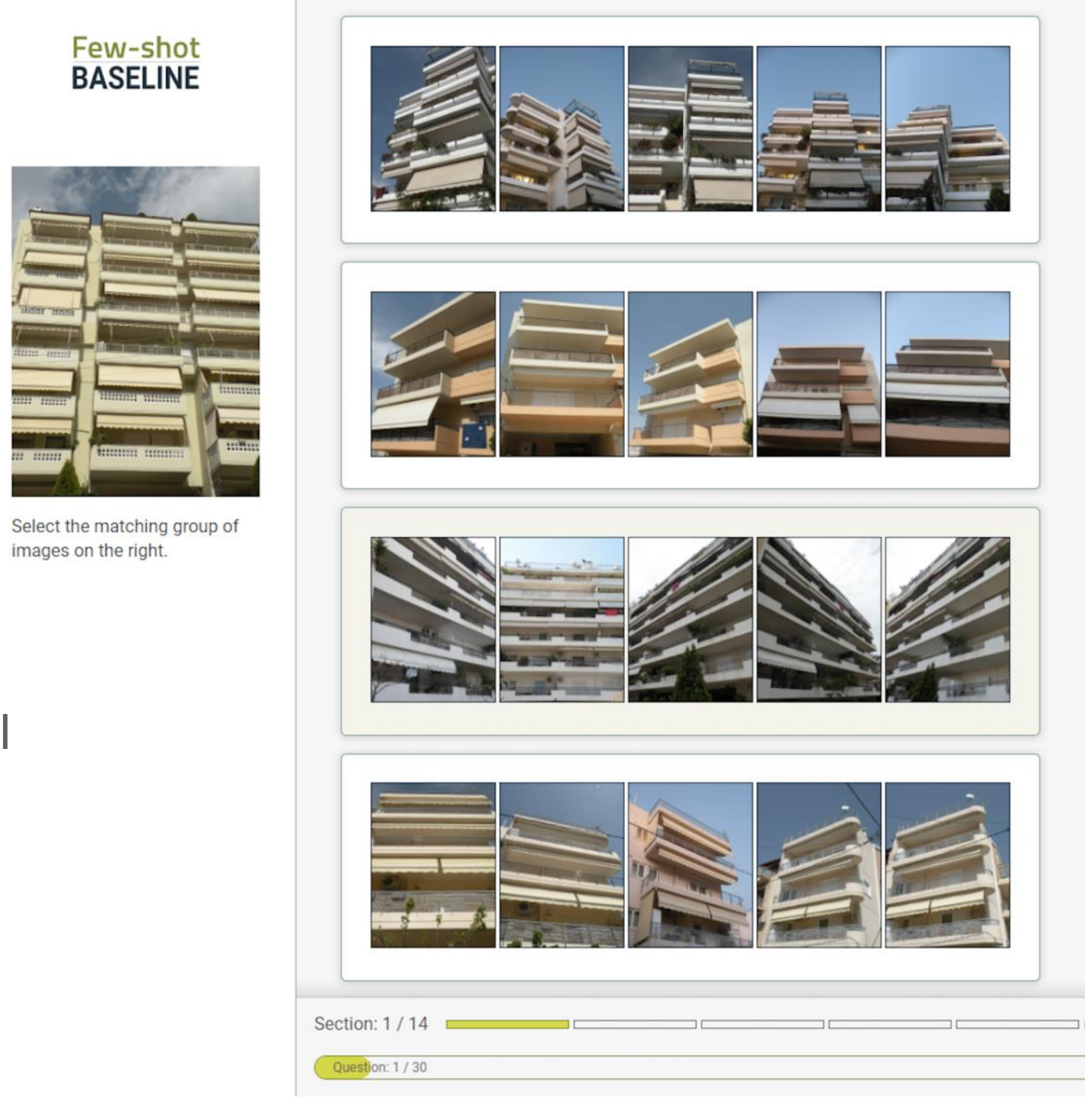
	5-shot, 5-way	Encoder	Transductive
MatchingNets	60.0	Conv (64)x4	No
ProtoNets	68.2	Conv (64)x4	No
MAML	63.1	Conv (32)x4	Yes* Not stated in paper
R2D2	68.4	Conv (96)x4	No
TADAM	76.7	Resnet-12	No
Fine-tuning	**78.2**	WRN-28-10	No
Transductive fine-tuning	**78.4**	WRN-28-10	Yes
LEO	77.6	WRN-28-10	Yes?

Results as reported in https://arxiv.org/pdf/1909.02729.pdf

Establishing a human baseline

- These problems remain for almost all few-shot baselines
  - E.g. Metadataset also limits models to 84x84 pixel resolution
  - E.g. Kinetics video dataset has Kinetics 400, two versions of Kinetics 600…
- Few papers consider the possibility that some images may belong to no class
- The end result of this:
  - Few-shot models trained for real-world use can’t be directly compared against existing published results
- Compare few-shot models against human baseline
  - How well could a human do the same task?

Baseline User Interface



Few-shot
BASELINE

Select the matching group of images on the right.

Section: 1 / 14

Question: 1 / 30

Image to label

Classes to choose from

Results – By Dataset

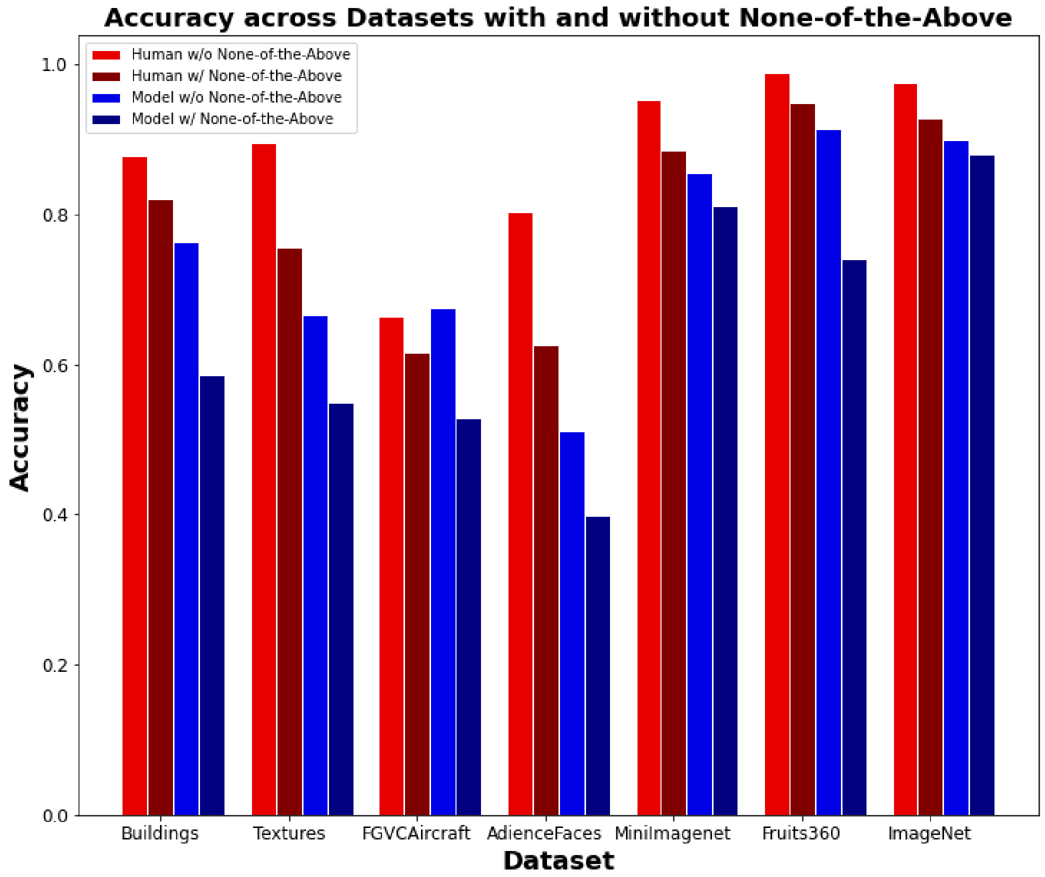
- Results for all participants (n=46)
  - On average, only 14.0%pt. gap between model and human accuracy, min. 3.8% gap
  - Accuracy depends heavily on dataset and is correlated with speed (r=-.113, p<<0.0001)
  - Model speed depends on compute available. On a single P100 GPU = 0.07 seconds per decision

Dataset	RCNet Accuracy (%)	Human Accuracy (%)	Human speed (seconds)
ImageNet	88.9	95.1	6.4
miniImageNet	83.3	92.0	7.6
FGVC Aircraft	60.3	64.1	17.3
Describable Textures	60.9	82.6	7.9
Fruits 360	82.8	96.8	5.8
Urban Buildings	67.5	85.0	15.0
Adience Faces	45.5	71.4	12.6

Pacific Northwest
NATIONAL LABORATORY

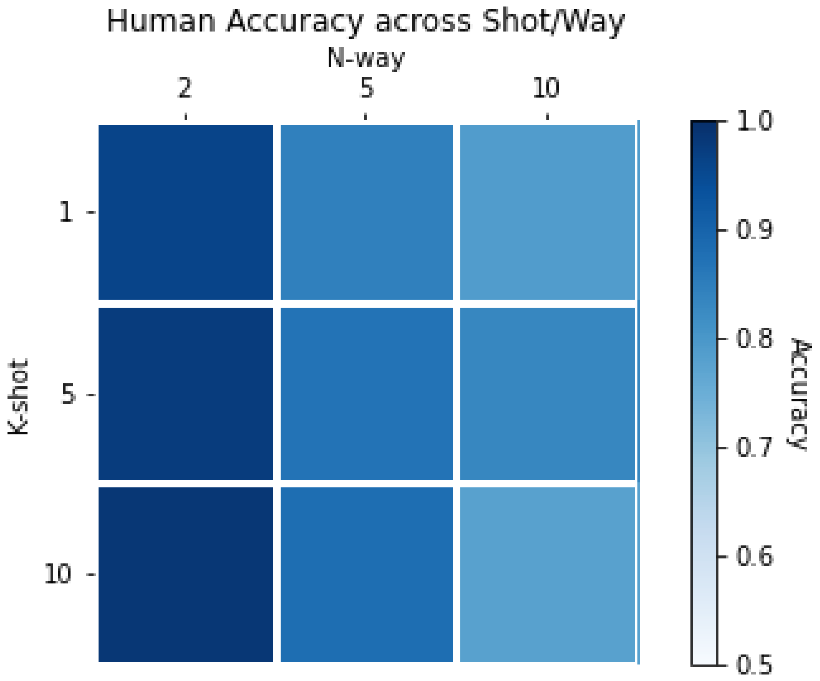
Results – None-of-the-above

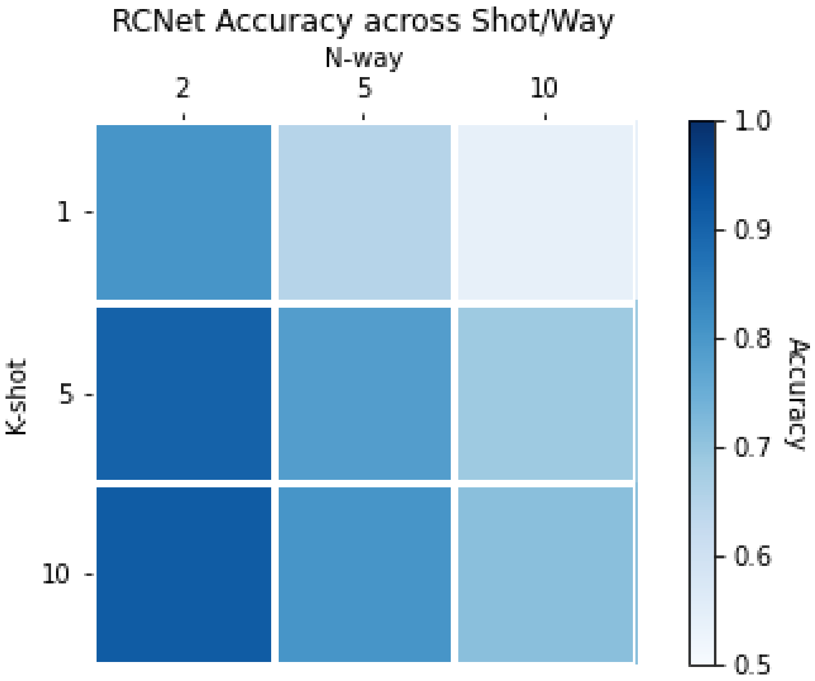
- Inclusion of none-of-the-above images decreases performance by 8.2%pts for participants and 11.3%pts for the model
- Pattern is consistent across datasets



Results – Shot/Way Combinations

- Humans perform worse as the number of classes increases (n-shot)
- Model performs worse with more classes and fewer examples



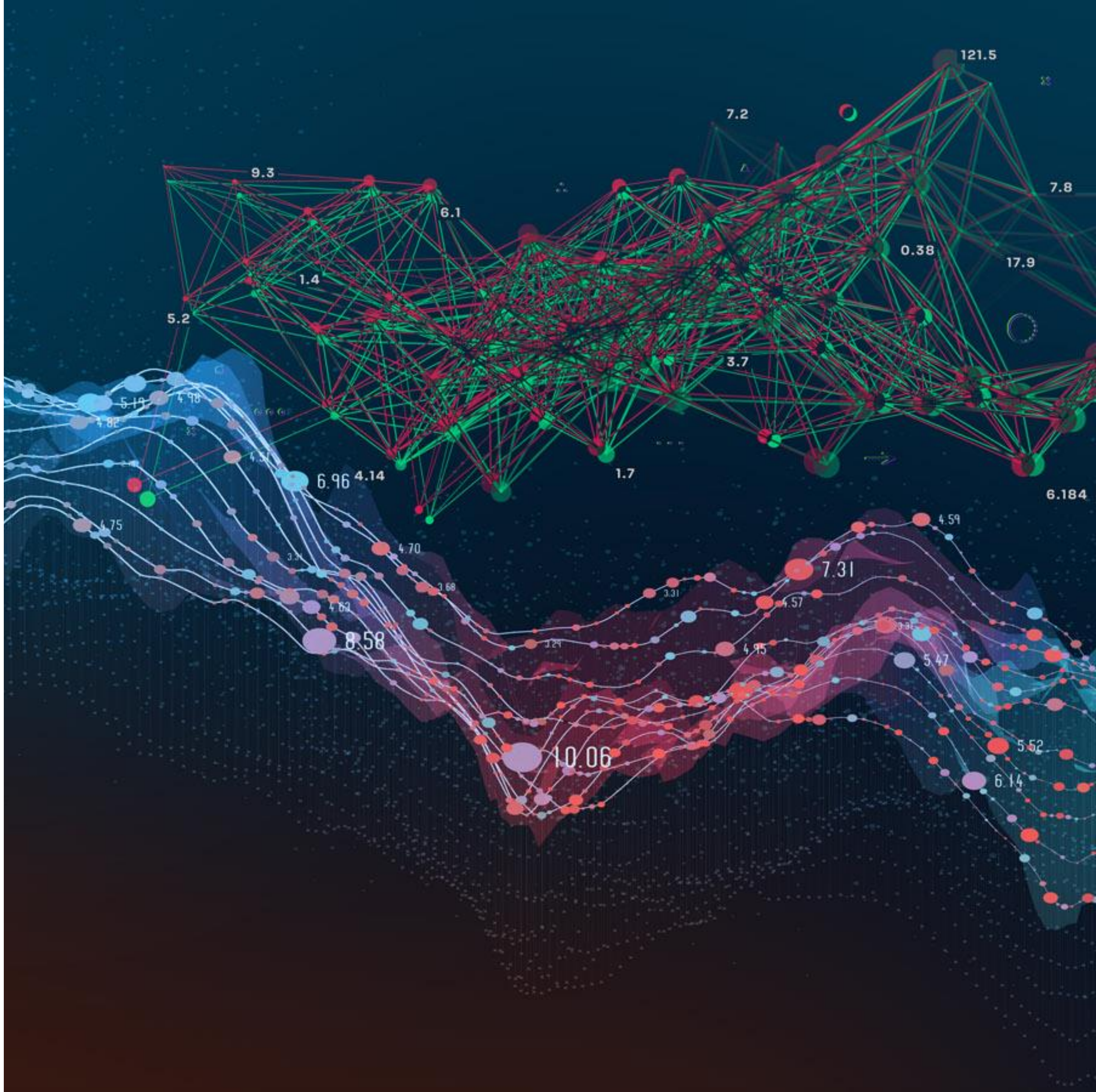


Pacific Northwest
NATIONAL LABORATORY

Alternatives to Few-Shot Learning

U.S. DEPARTMENT OF ENERGY BATTELLE

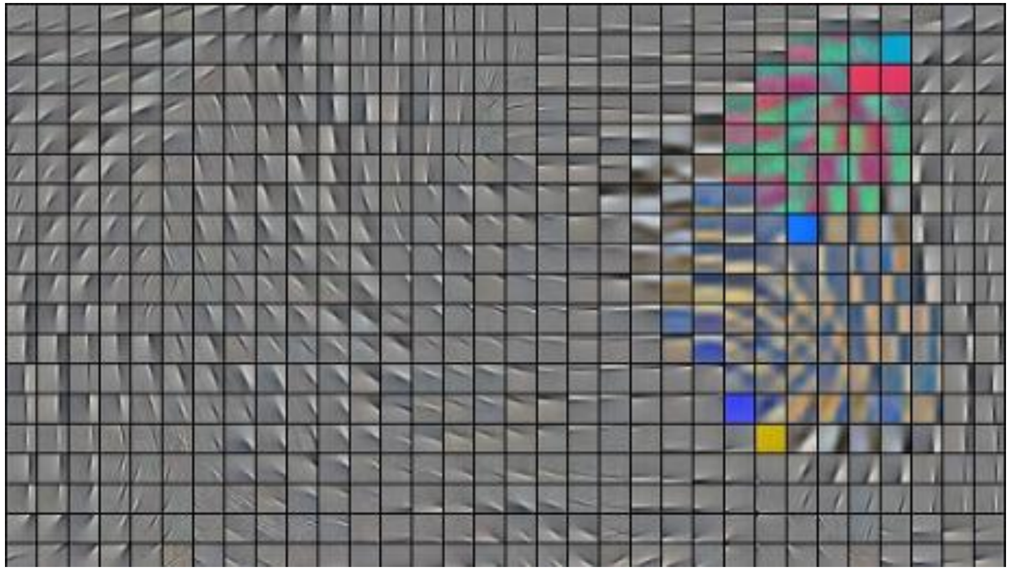
PNNL is operated by Battelle for the U.S. Department of Energy



What is Transfer Learning?

- Transfer Learning is the practice of leveraging a prior-trained network for a new task or domain

- Why does this work?
- Most CNNs show the same type of early features
- The layers near the output are tightly coupled with the task
- We want to exploit the generalizability of the early layers



Transfer Learning

- Benefits:
  - Spreads the benefit of state of the art research to the masses
  - Jumpstart training for a difficult problem
  - Reduce overfitting on a small data set

- Common scenarios for Transfer Learning:
  - Task Transfer: Have a model trained to identify dogs and cats, need a new one to detect horses and cows
  - Domain Transfer: Have an image classifier trained on professional photographs, need a model that runs on cell phone image data

Model Fine-tuning Considerations

- Things to keep in mind:
  - What type of transfer learning are you doing (domain or task)?
  - How much data do you have?
  - How similar is the target problem from the source problem?

- Typically, fine-tuning is much faster than training from scratch
  - ~1-5 epochs instead of 20+

- If your fine-tuning dataset is small, you *can* train too much -> overfitting
  - Overfitting is the bane of few-shot learning. If your model overfits to the point it does poorly on new images for old classes, how can it do well on entirely new classes?

Pacific Northwest NATIONAL LABORATORY

Transfer learning as few-shot learning

- A straight-forward approach might be to train a network on the few-shot training data as a traditional classifier, and then fine-tune on the support set.
  - Unfortunately, for small data this technique seems to perform poorly (Vinyals et al., 2016; Chen et al., 2018)

- But transfer learning can potentially become competitive when applied carefully (Chen et al., 2018; Dhillon et al., 2020)

Baseline++

- Baseline++ applies transfer learning to a few-shot scenario (Chen et al. 2018)

- Modifications to standard transfer learning:
  - Fix encoder weights and do not update
  - Rely on cosine distance instead of traditional linear classifier
  - Fine-tune the weight vector for each class
    - Similar to prototypical networks

- Matches similar level of performance to few-shot models on MiniImageNet

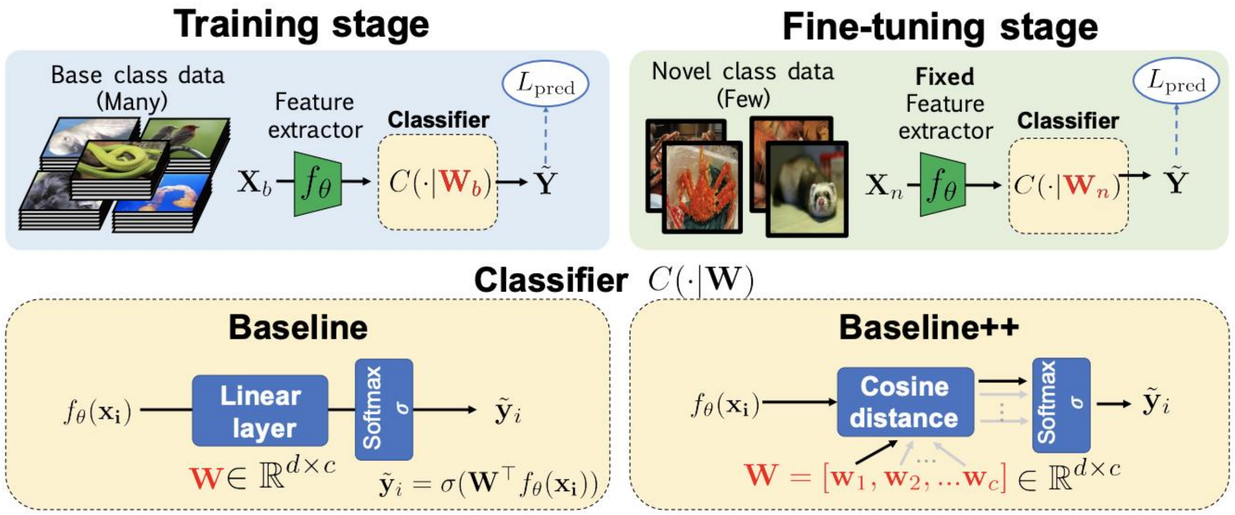


Figure 1 from https://arxiv.org/pdf/1904.04232.pdf

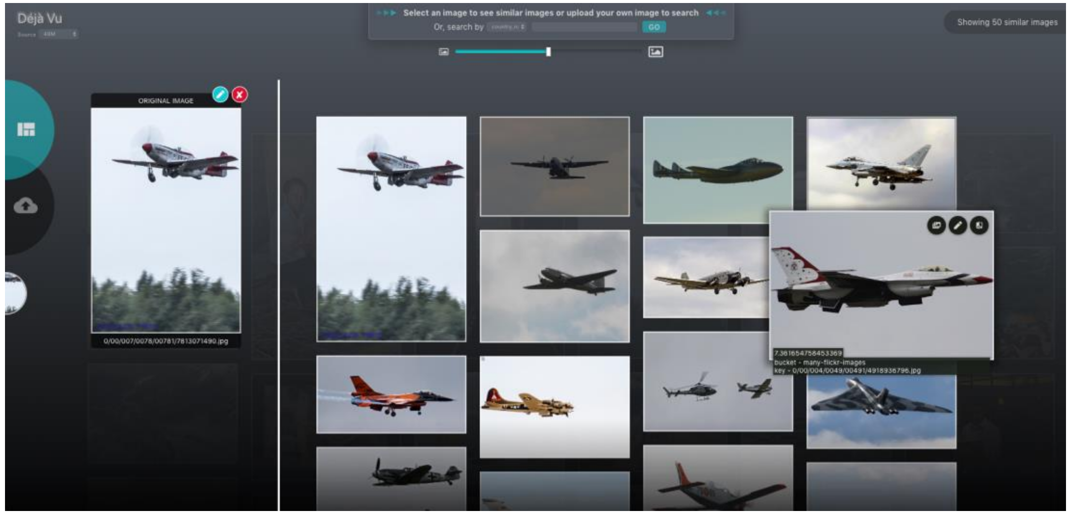
Transductive Fine-tuning (Dhillon et al., 2020)

- Transductive models use all datapoints in the test set to make predictions
  - E.g. Allows semi-supervised techniques to be applied at test time

- Modifications to standard transfer learning:
  - Cross-entropy loss is used on support examples, Shannon Entropy regularization applied to query examples
    - Transductive because it uses all queries to label a single query example
  - Initialize classifier weights and normalize model outputs to unit L2 norm
    - Maximizes cosine similarity between weights and encoder outputs
  - Use model logits as the input to final classification layer instead of penultimate layer's features
    - Logits show better clustering behavior

Information Retrieval vs Few-shot

- Many information retrieval systems rely on calculating distances based on features extracted from a pretrained model



Select an image to see similar images or upload your own image to search

Or, search by

0/00/007/0078/00781/7813071490.jpg

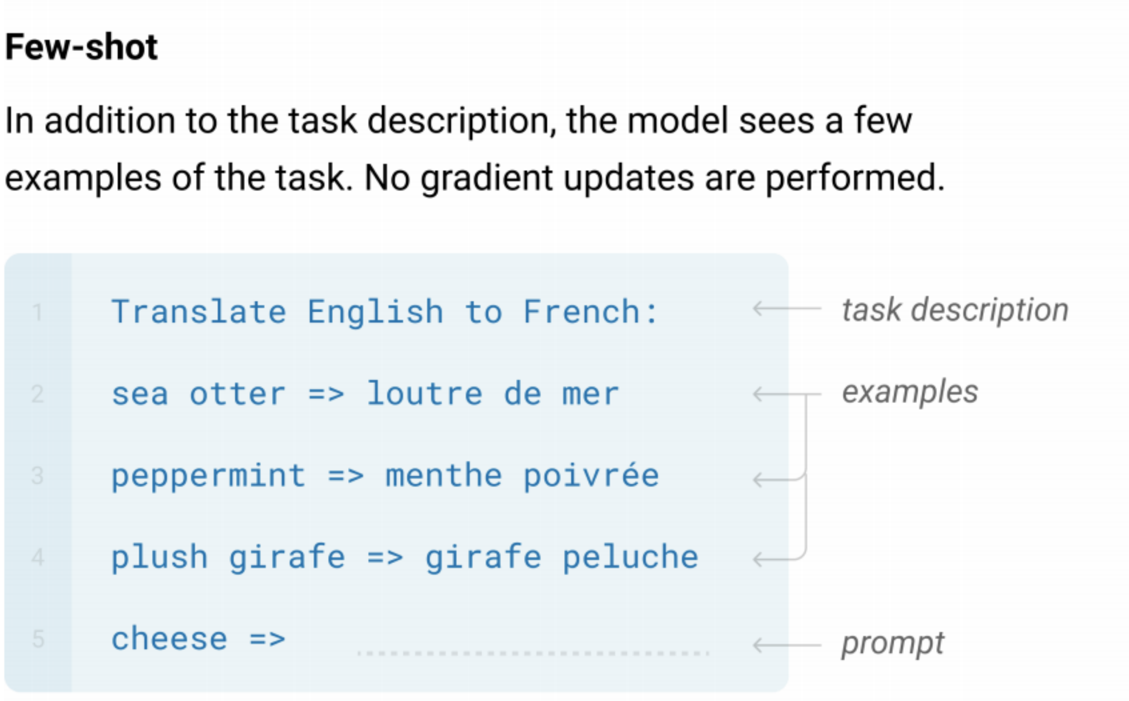
7.361654758453369
bucket - many-flickr-images
key - 0/00/004/0049/00491/4918936796.jpg

- When to choose information retrieval
  - Only have a single example
  - It’s a quick way to look through a larger dataset

- When to choose few-shot learning
  - Have multiple images you’d like to leverage, or multiple classes to choose from
  - Need to generalize beyond training dataset

GPT-3 - Language Models are Few-shot Learners

- Alternatively, in some cases pre-trained models may work for few-shot without modification…
- For large language models, support examples can be provided as context
- 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.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

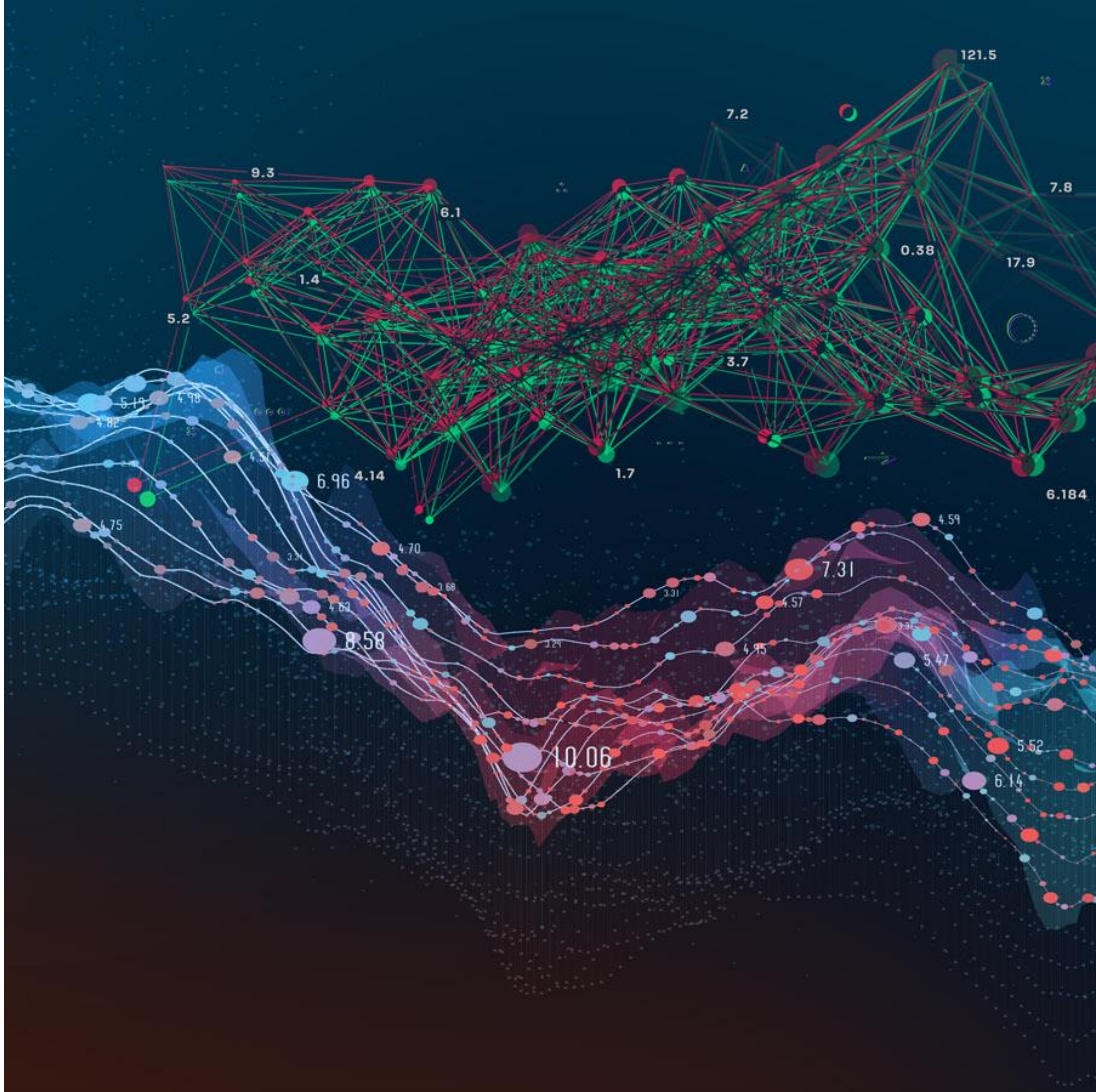
https://arxiv.org/pdf/2005.14165.pdf

Pacific Northwest
NATIONAL LABORATORY

More on Evaluation & Generalization

U.S. DEPARTMENT OF ENERGY BATTELLE

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



Evaluating few-shot generalization

- As performance on individual image datasets has begun to plateau, the field has moved towards better assessments of generalization

- How can we understand when few-shot models generalize to a new class (or don’t?)

- What if a specific type of class is absent from training?
  - Challenge test sets

- What if the type of class is very different from the training data?
  - Meta-dataset

Distributions of classes

- To understand where a few-shot model will perform well, we have to understand what defines a ***class*** during training
  - Performance will be best when test classes mirror training classes

- Simply put, classes are defined by our labels:
  - ImageNet: class = **object identity** (e.g. dog, cat, plane, train, etc.)
  - VisualGenome: class = **object identity** (for specific region of image)
  - Describable Textures: class = **texture of object**
  - Kinetics 600: class = **human activity in video**
  - Reddit comments: class = **authorship**

- If during inference, we change the definition of what makes up a class, performance will degrade

An example – Sorting cars by color

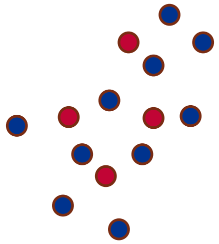
- Model was trained on Stanford Cars dataset
  - Class determined by Make/Model/Year

- Color represents an “out-of-distribution task” in that it does not follow the same pattern used to create classes for training
  - In fact, color may be largely ignored by the model since any make/model/year may have any number of different colors.

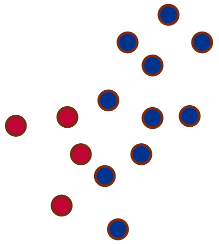
Challenge Test Sets

- What happens when few-shot models generalize to a category that doesn’t exist in the training data?

Easy Test set



Hard Test set



- Traditionally, deep learning has performed poorly in these scenarios

Challenge Test Sets - SCAN

- SCAN is a non-fewshot dataset which demands out-of-distribution generalization (Lake & Baroni, 2018)
- Sequence-to-sequence task: Text command -> ACTION
  - “Jump left” -> LTURN JUMP
  - ”Jump opposite left and walk thrice” -> LTURN LTURN JUMP WALK WALK WALK
- Easy test set: A random set of commands are withheld
  - 100% accuracy even when trained on only 8% of all combinations of commands
- Hard test set: Sees all primitives, but never the combination “turn left”
  - LSTM-based seq2seq = 90.3% accuracy

Challenge Test Sets - SCAN

- But “turn left” is very similar to “turn right”. “Jump” follows the same pattern as “walk” and “run”, but shows different generalization performance

- 2nd Hard test set: Sees all primitives, but never “jump” in combination
  - LSTM-based seq2seq = 1.2% accuracy (Lake & Baroni, 2018)
  - LSTM + Syntactic Attention = 78.4% (Russin et al., 2019)
  - Seq2seq + meta-training = 99.95% (Lake, 2019)

- What is meta-training?
  - Same support/query structure as used in few-shot learning

Pacific Northwest NATIONAL LABORATORY

Challenge Test sets

- Matching Networks (Vinyals et al., 2016)
    - Early few-shot learning model, introduced MiniImageNet

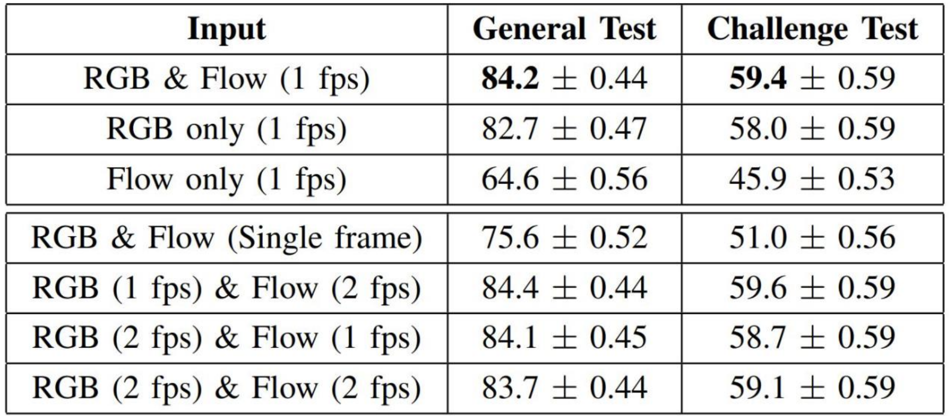
- Trained two few-shot models on ImageNet.
    - The first had a random train/test split
    - The second was trained on all *non-dog* categories and evaluated on all *dog* categories

- Even for few-shot models, this can be a challenging task

	Random split	Dogs Split
Train	97.0%	96.4%
Test	93.2%	**58.8%**

Challenge Test Sets

- Kinetics 600
  - Challenge test set is made up of “violent” actions
  - See similar performance drop



Input	General Test	Challenge Test
RGB & Flow (1 fps)	**84.2** ± 0.44	**59.4** ± 0.59
RGB only (1 fps)	82.7 ± 0.47	58.0 ± 0.59
Flow only (1 fps)	64.6 ± 0.56	45.9 ± 0.53
RGB & Flow (Single frame)	75.6 ± 0.52	51.0 ± 0.56
RGB (1 fps) & Flow (2 fps)	84.4 ± 0.44	59.6 ± 0.59
RGB (2 fps) & Flow (1 fps)	84.1 ± 0.45	58.7 ± 0.59
RGB (2 fps) & Flow (2 fps)	83.7 ± 0.44	59.1 ± 0.59

- Reddit authorship
  - “Hard” test set is made up of users who post on a very diverse set of subreddits
  - Intended to capture whether the model “solves” authorship by recognizing most authors tend to write on a small set of topics

	Easy	Hard
DistilBERT	83.2%	61.3%
All Stylometrics	67.3%	52.0%
All + DistilBERT	83.8%	61.8%

Meta-Dataset (& other meta-datasets)

- Introduced in Triantafillou et al. (2020), Meta-dataset is a few-shot dataset made up of 10 different few-shot datasets
  - Train on 8 datasets, test on 2 held-out datsets

- Speaking of distributions…
  - Traditional deep learning generalizes to new examples of *known* classes
  - Few-shot generalizes to new examples of *unknown* classes
  - Meta-dataset generalizes to new examples of *unknown* classes from *unknown* datasets!

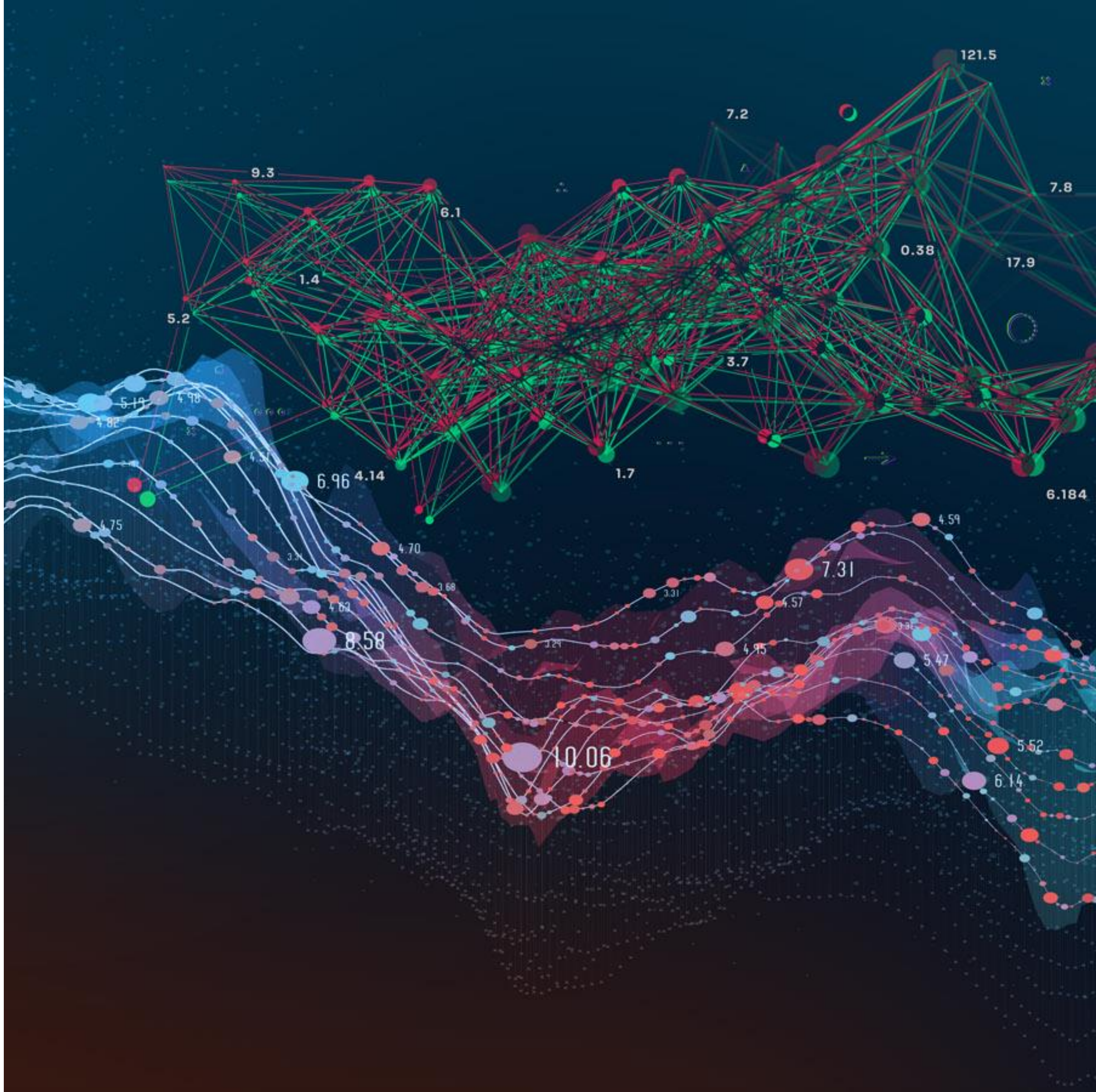
Training datasets		**Validation datasets**
ImageNet	Quick Draw	Traffic Signs
Omniglot	Fungi	MSCOCO
FGVC Aircraft	VGG Flower	
CUB Birds	Describable Textures	

Pacific Northwest NATIONAL LABORATORY

Bias and Ethics Considerations

U.S. DEPARTMENT OF ENERGY BATTELLE

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



Large Dataset challenges

- What are potential harms and injustices?
  - People’s images are constantly mined without any real consent.
    - ImageNet contains plenty of voyeuristic, non-consensual images.
    - Reverse image searching identities is not that difficult!

- Many categories are themselves unethical and offensive.
  - 80 Million Tiny Images has 10,000+ images of people categorized as slurs.

- Are there counterbalances against these harms?
  - Reactive
    - Public outcry, legal action...
  - Proactive
    - Ethics boards, regulation...

Large Dataset challenges - Opaque Datasets

- JFT-300M is a huge, effectively private image dataset.
  - Used in many celebrated papers (GANs, separable convolution, etc).
  - 300M+ images across 18,000 categories.
  - Hard to find any details...

- Open Images is an open subset of JFT-300M.
  - ...and contains non-consensual images of children in swimwear.

- Prabhu & Birhane (2020) provide an overview of large image datasets
  - https://arxiv.org/abs/2006.16923

- What incentives encourage these types of datasets to exist?
- What incentives could *discourage* their prevalence?

Pacific Northwest NATIONAL LABORATORY

Large Dataset challenges – Introducing Bias

- Training on large datasets often requires training on data from the internet

- Major source of bias in text models
  - GPT-2 and GPT-3 both have known issues of bias based on gender, race, and other factors.
  - E.g. GPT-3 has difficulty writing about Muslims without referencing violence
    - Content warning for islamophobia: https://twitter.com/abidlabs/status/1291165311329341440

Solutions: Power Distribution

- Who *actually* has a say in…
  - What categories and ontologies we construct?
  - How models are trained and evaluated?
  - How we hold people and groups accountable?

- The answer is rarely "those most impacted by them".

- Avenues for shifting power...
  - Increased transparency and dataset auditing.
    - Facilitates outside review.
  - Ethics committees.
    - Reduces *some* burden on practitioners.
  - Others?

Small Dataset challenges

- Ever-larger datasets have many known issues…

- But small data can also pose challenges:
  - Instead of large-scale societal biases, small datasets may have biases introduced through data collection

  - If a model is asked to solve a problem based on very little information, it has to rely on other sources of information

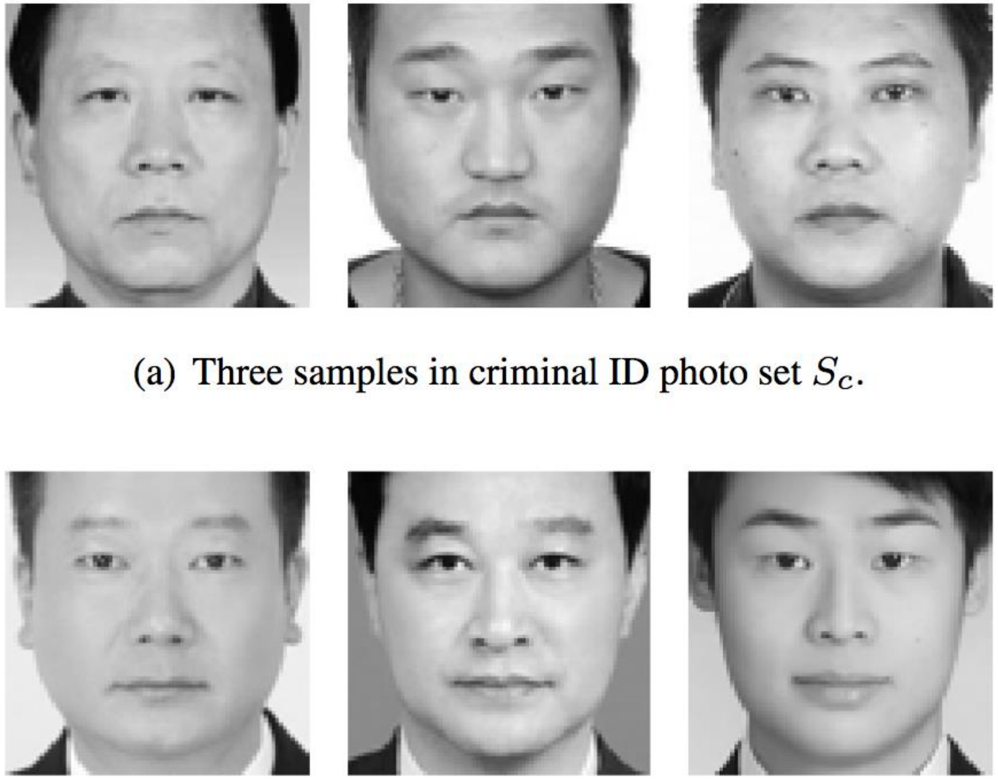
Small scale bias

- When we ask models to do impossible things, we set ourselves up for our models to learn from biased datasets

- Even small curated datasets can introduce bias

- “Automated Inference on criminality using Face Images”
- **Wu & Zhang (2016)**
  - Posted to ArXiv
  - 89.51% Accuracy
  - <u>Totally bogus results due to dataset</u>

https://www.semanticscholar.org/paper/Automated-Inference-on-Criminality-using-Face-Wu-Zhang/1cd357b675a659413e8abf2eafad2a463272a85f



(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Low Information Tasks

- PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models
  - Original paper, https://arxiv.org/pdf/2003.03808.pdf
  - Code, https://github.com/adamian98/pulse

- Leverages pre-trained model based on Flickr Face HQ Dataset
  - https://github.com/NVlabs/ffhq-dataset

- Trained on CelebA HQ dataset
  - https://www.tensorflow.org/datasets/catalog/celeb_a_hq

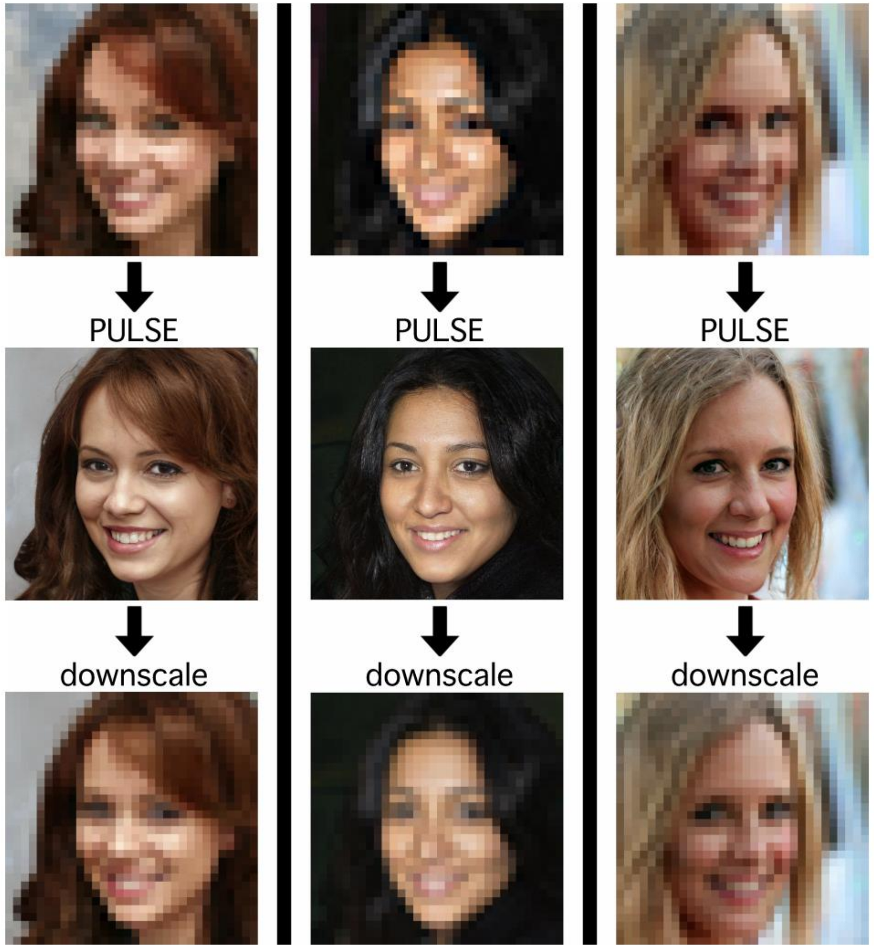
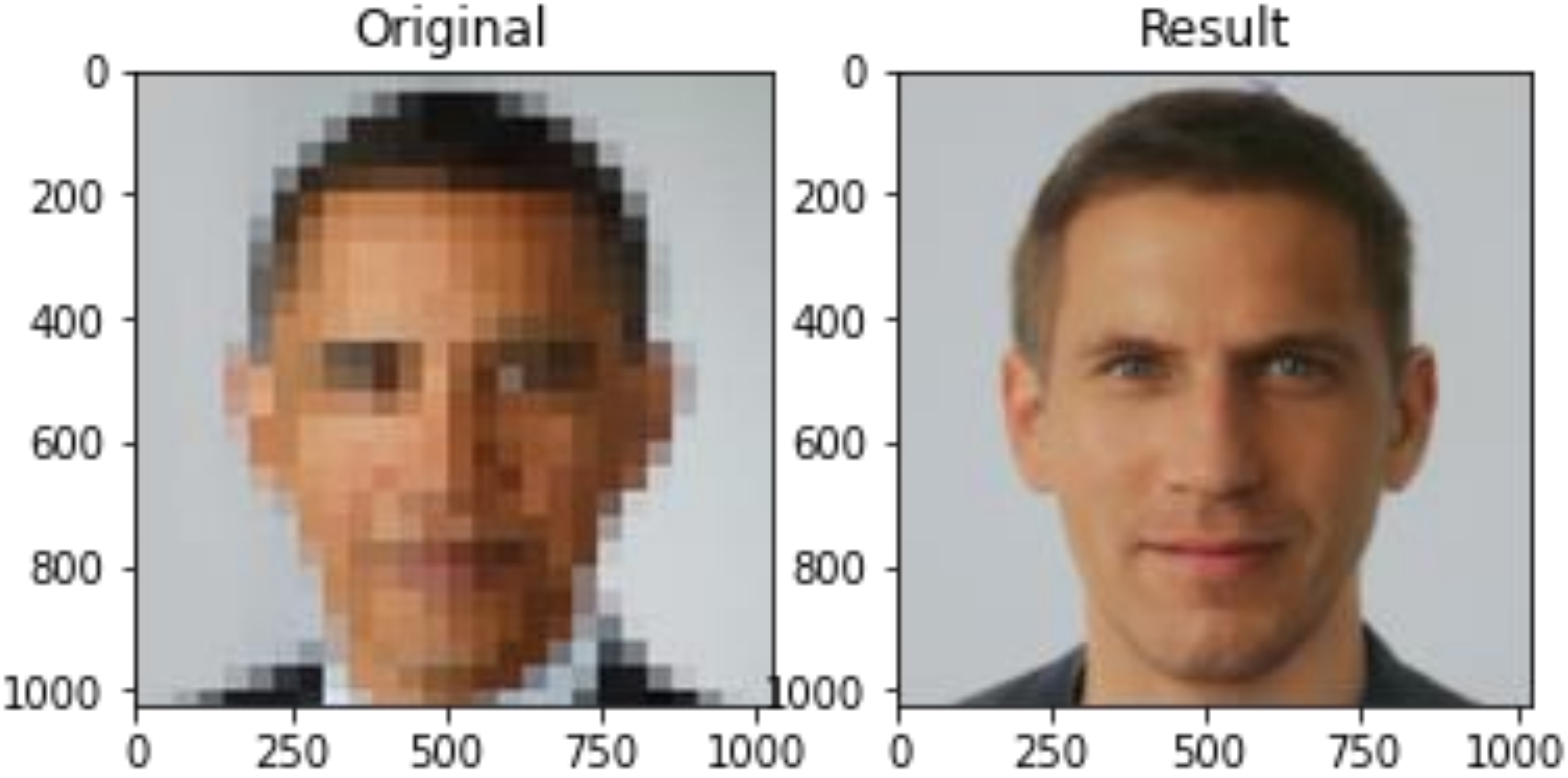


Figure 1. (x32) The input (top) gets upsampled to the SR image (middle) which downscales (bottom) to the original image.

https://arxiv.org/pdf/2003.03808.pdf

PULSE - Testing it out



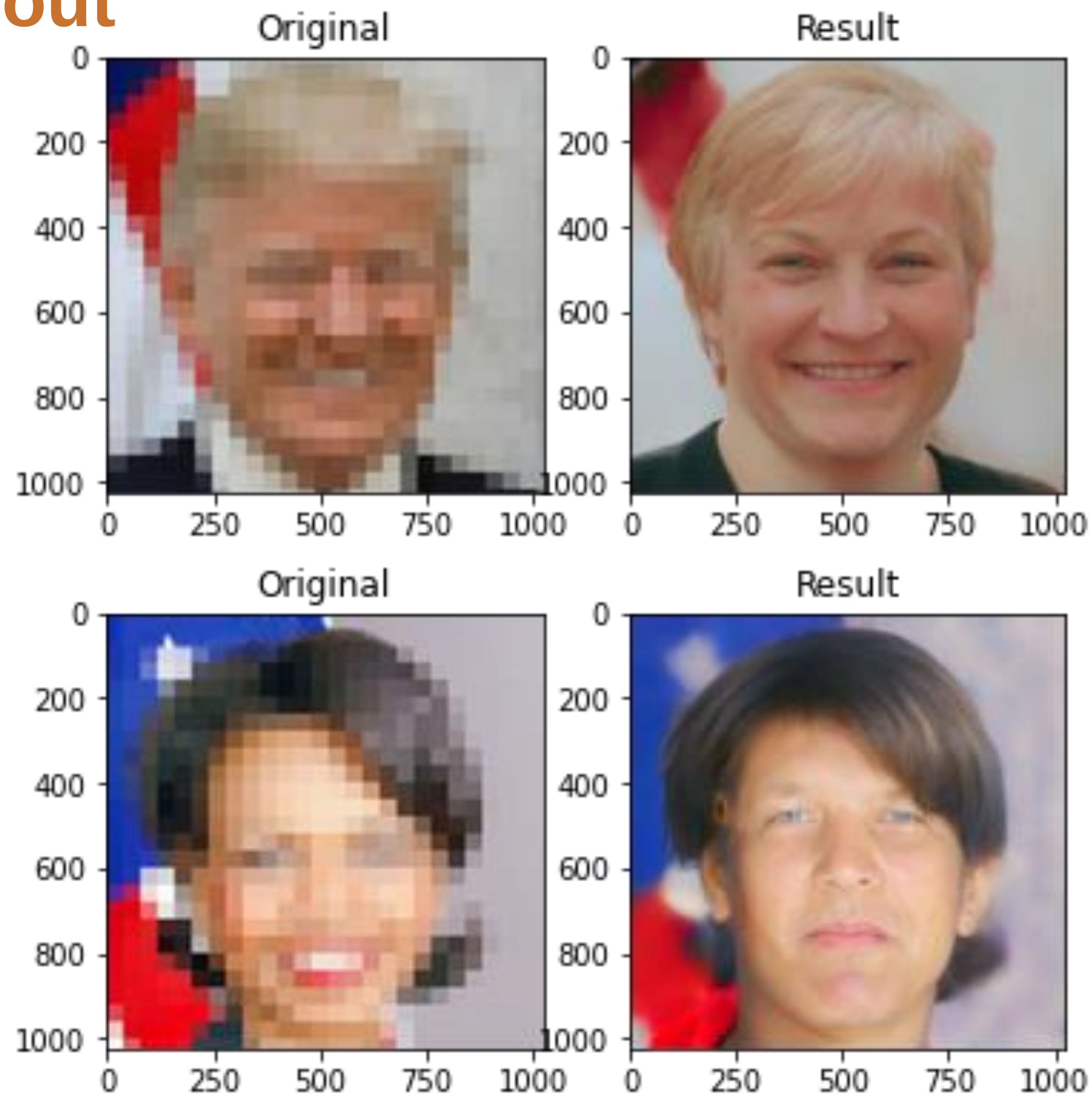
Original Obama photo: https://www.biography.com/us-president/barack-obama Photo Credit: Pete Souza

Image went around on AI/ML Twitter

PULSE - Testing it out

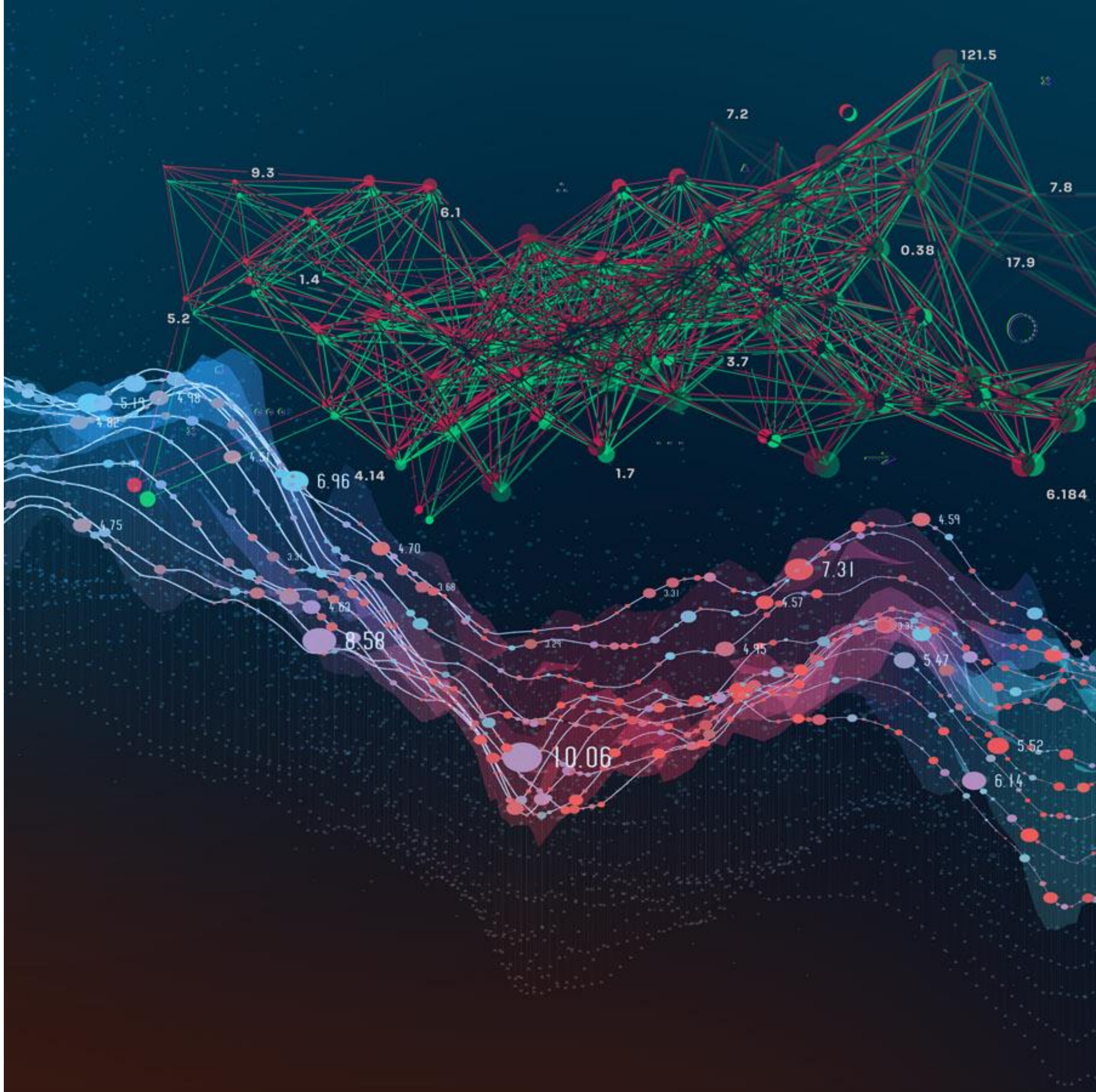
- Where did things go wrong?
  - Dataset biases
  - Pre-processing considerations
    - ✓ E.g. % of image taken up by face

- Upscaling requires additional information not in input
  - In these cases, models are forced to rely on general patterns learned from data
  - “Good” vs “bad” bias difficult to separate



Pacific Northwest
NATIONAL LABORATORY

Wrapping things up



U.S. DEPARTMENT OF ENERGY BATTELLE

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

Pacific Northwest
NATIONAL LABORATORY

Few-Shot Learning in Overview

- Increasingly powerful approach to handling small data problems
  - The field is still young and no clear winner has emerged in terms of approach
  - Expanding how we think of and measure generalization (e.g. meta-dataset)
- Can be applied to problems in a variety of data types
  - Image filtering, Authorship attribution, Activity recognition, Audio event detection
- Influencing other fields
  - Transfer learning is improving based on ideas drawn from few-shot
  - Reinforcement learning has incorporated MAML-style meta-learning to reduce training times and improve generalization