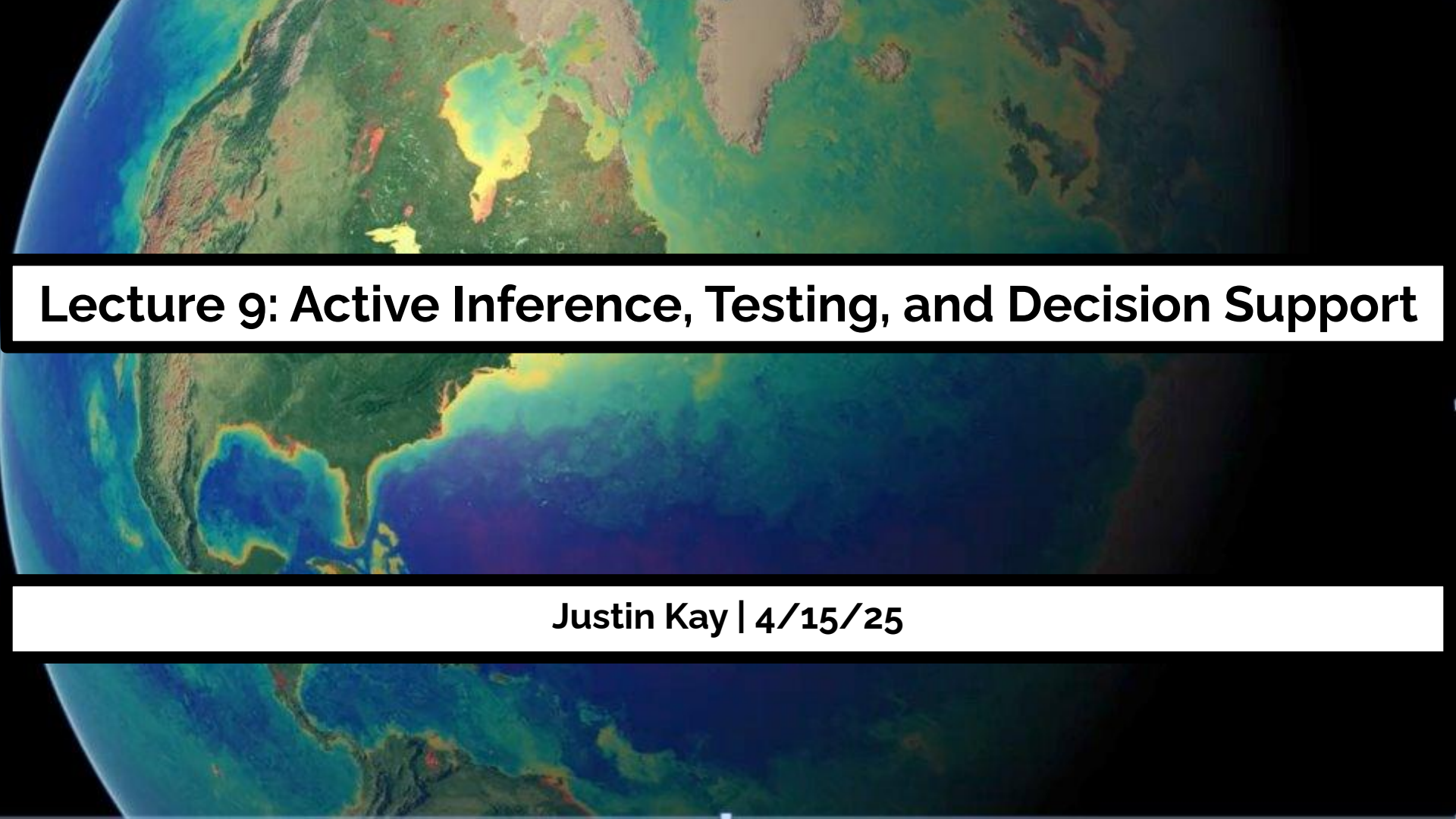
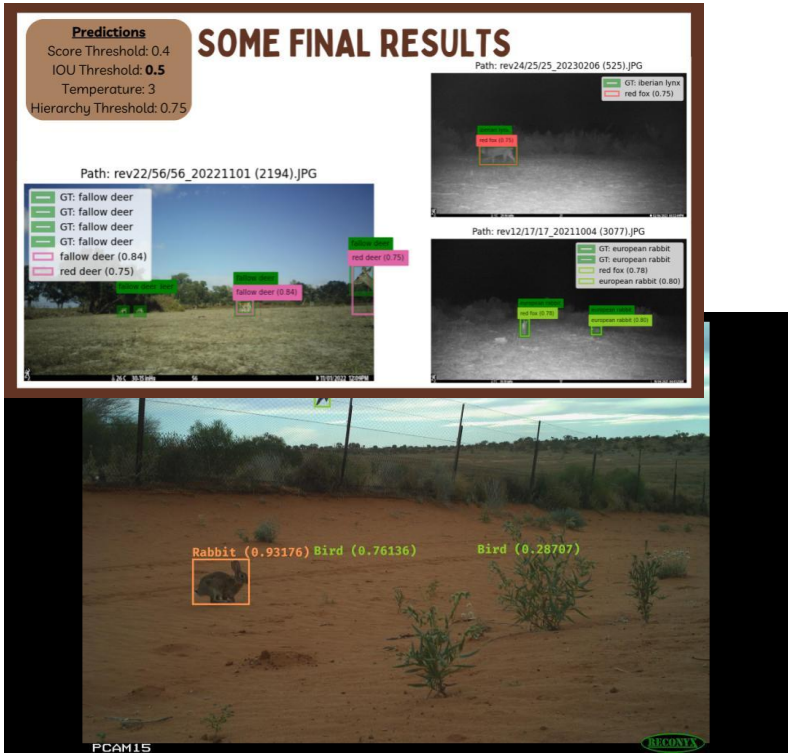


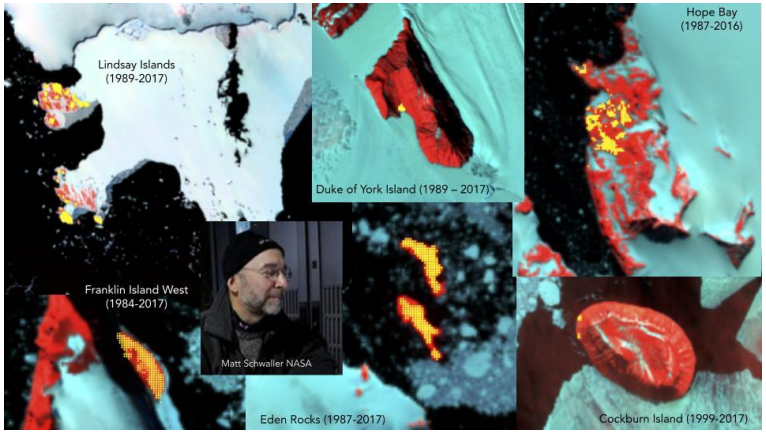
Lecture 9: Active Inference, Testing, and Decision Support

Justin Kay | 4/15/25

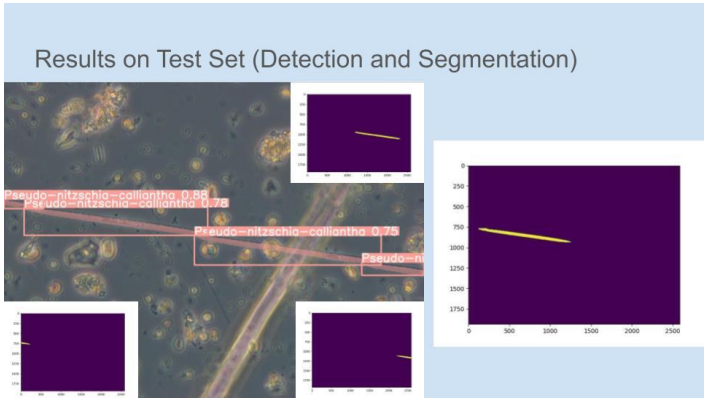
Typical model development / benchmarking



Animal occupancy/abundance, Marquez-Rodriguez, Tamm

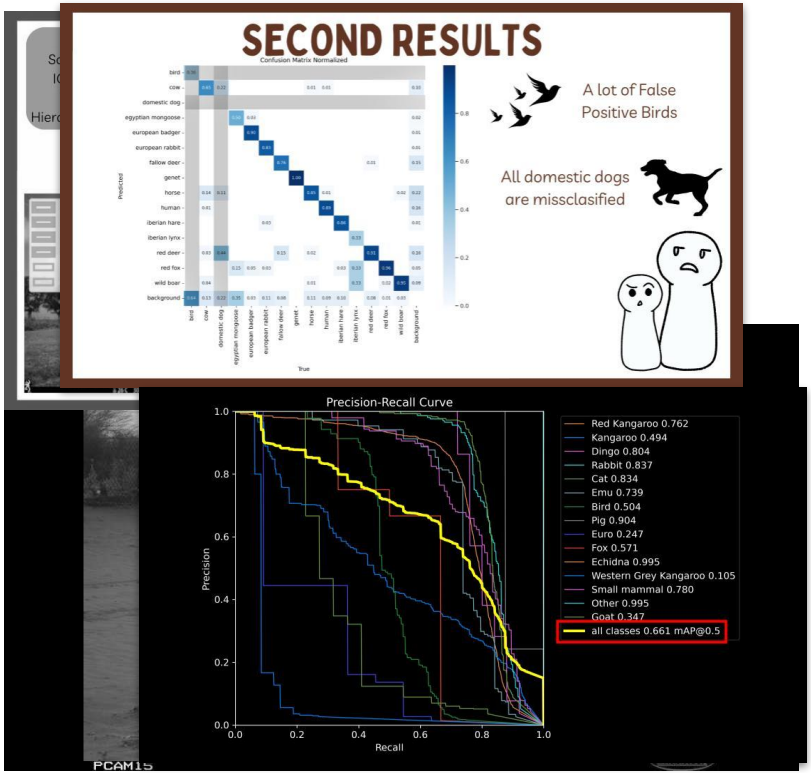


Guano surface area, Che-Castaldo

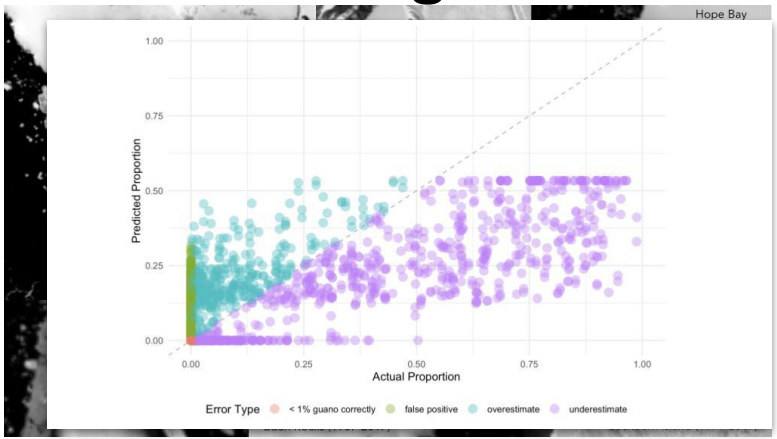


Phytoplankton biovolume, Marzidovšek

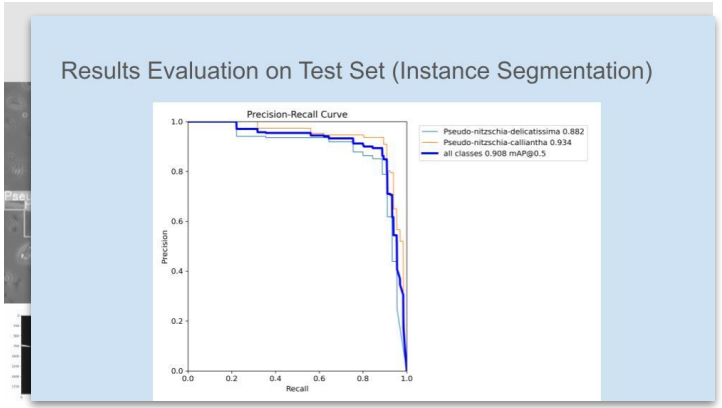
Typical model development / benchmarking



Animal occupancy/abundance, Marquez-Rodriguez, Tamm



Guano surface area, Che-Castaldo



Phytoplankton biovolume, Marzidovšek

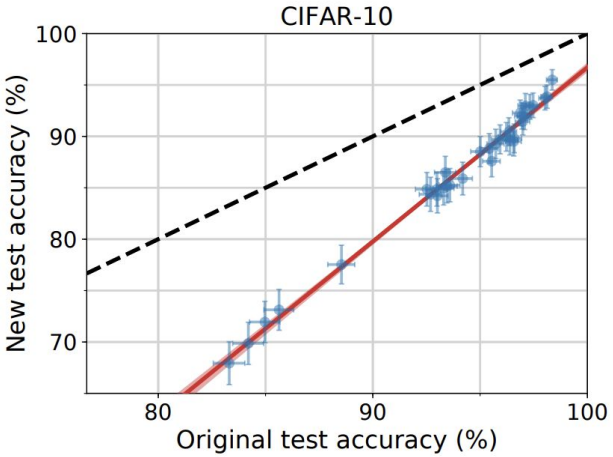
Performance on new data?

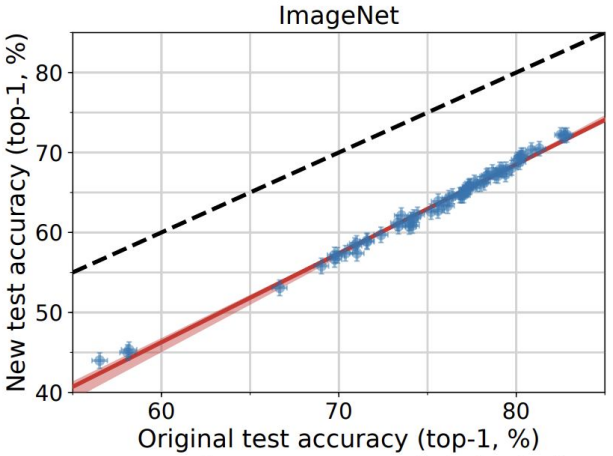
Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht* UC Berkeley
Rebecca Roelofs UC Berkeley
Ludwig Schmidt UC Berkeley
Vaishaal Shankar UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.



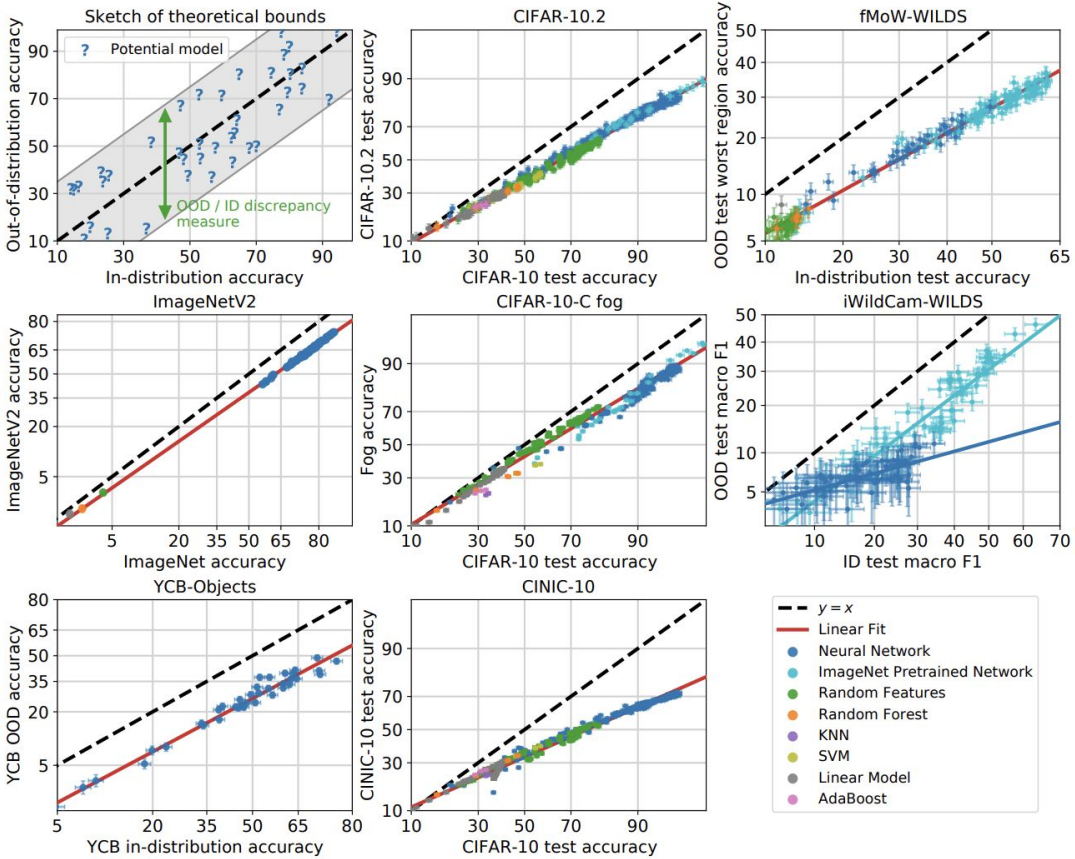


https://arxiv.org/pdf/1902.10811

Ideal reproducibility Model accuracy Linear fit

Performance on new data?

In-distribution and out-of-distribution accuracy often correlated, but *differently* correlated on different test data



https://proceedings.mlr.press/v139/miller21b/miller21b.pdf

Up until now

Training techniques to mitigate distribution shift

- Domain generalization and robustness
- Domain adaptation, specialization, transfer learning

This week: Test time

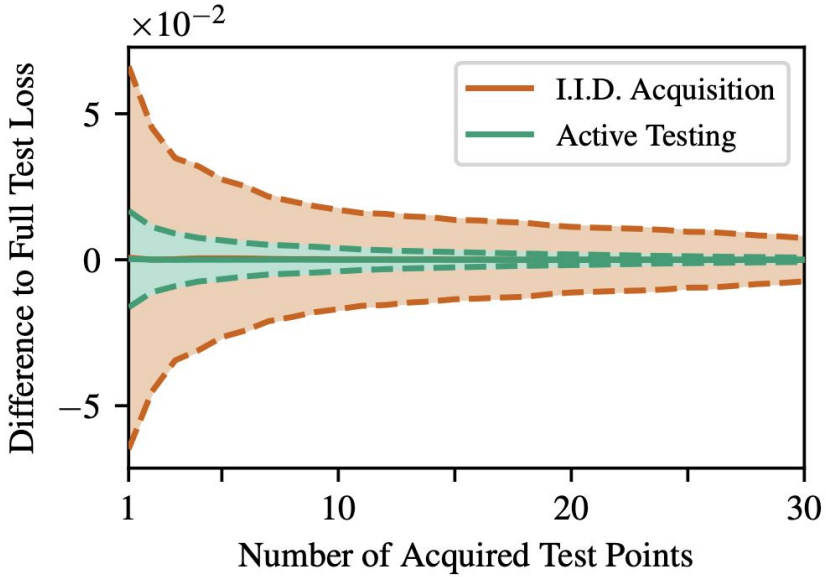
This week we **keep the model fixed** and focus on **post-training techniques** to interpret and utilize (imperfect) model predictions

- Active testing and model selection
- ML + statistical inference

Active testing

Understand how model will perform on some data with as few human labels as possible.

Similar motivations and methodologies to active learning.



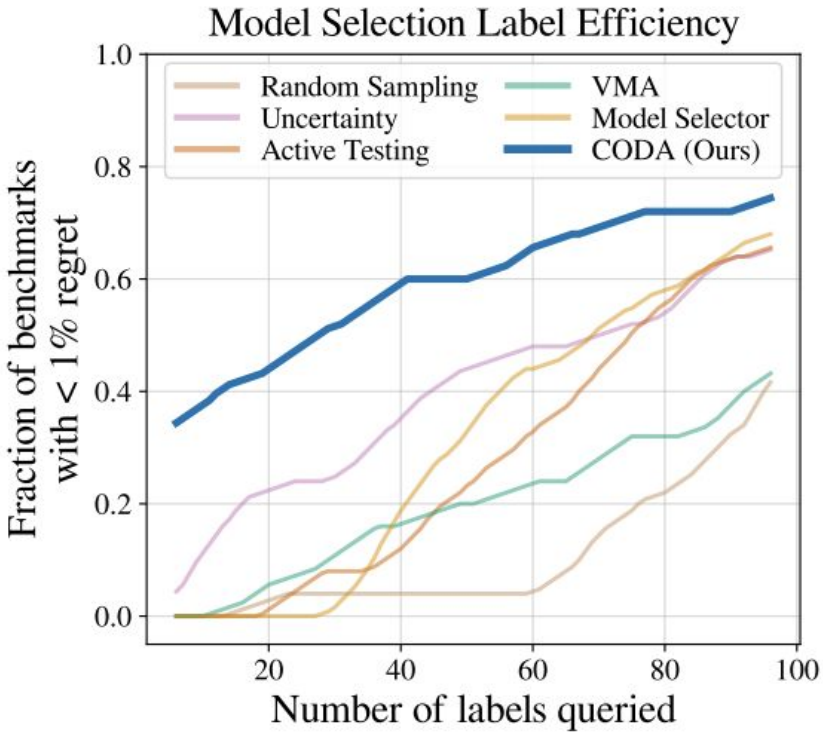
https://arxiv.org/pdf/2103.05331

Active model selection

You need to choose a model to use on your data, but don't have labels

- Model zoos
- Across checkpoint runs
- Domain adaptation

How to figure out what model you should use?



Consensus-Driven Active Model Selection, Kay et al. 2025

Uncertainty quantification

{ fox squirrel 0.99 }

{ fox squirrel 0.82, gray fox, 0.03, bucket, 0.02, rain barrel 0.02 }

{ marmot 0.30, fox squirrel 0.22, mink 0.18, weasel 0.16, beaver 0.03, polecat 0.01 }

https://arxiv.org/pdf/2009.14193

Per-data point predictions often aren't the end goal

Common goal: Statistical Inference

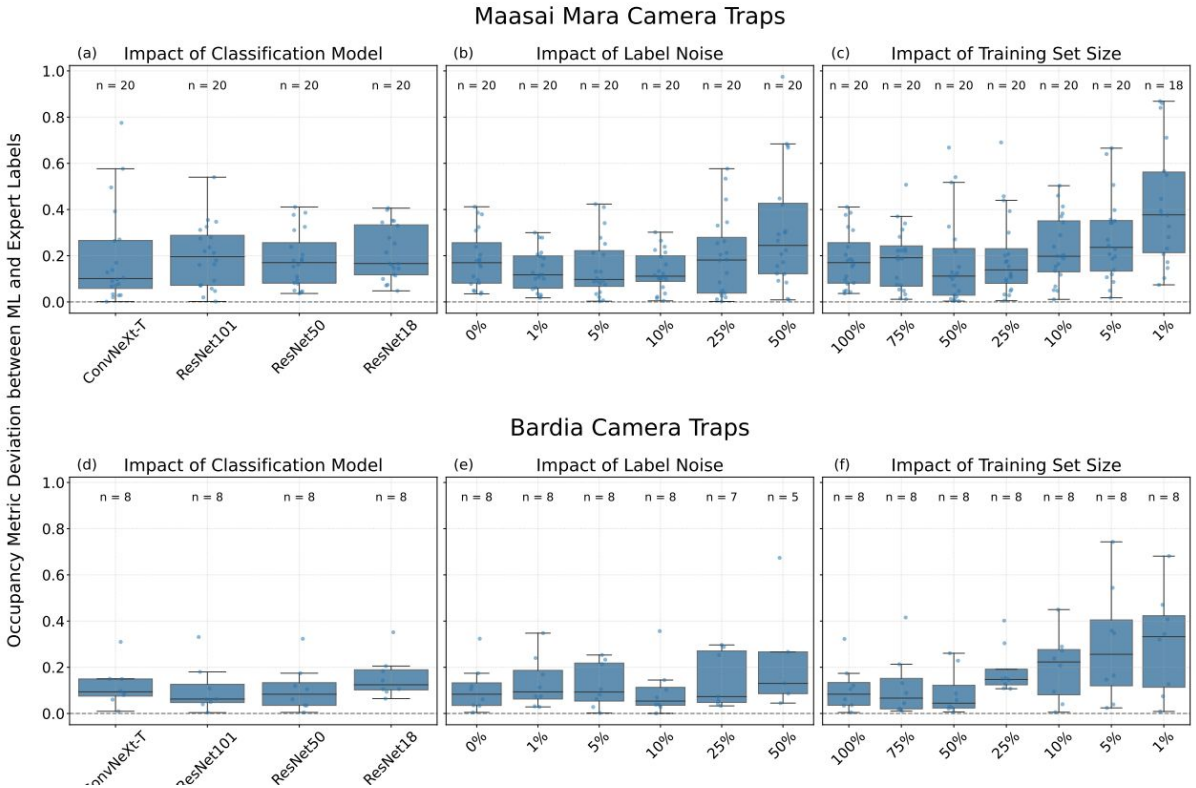
Use some data to infer some characteristics of the larger population:

- How many birds live here?
- What fraction of galaxies have spiral arms?
- What is the rate of deforestation of the Amazon?

https://arxiv.org/pdf/2408.14348

Per-data point predictions often aren't the end goal

Better models don't always translate to better inference

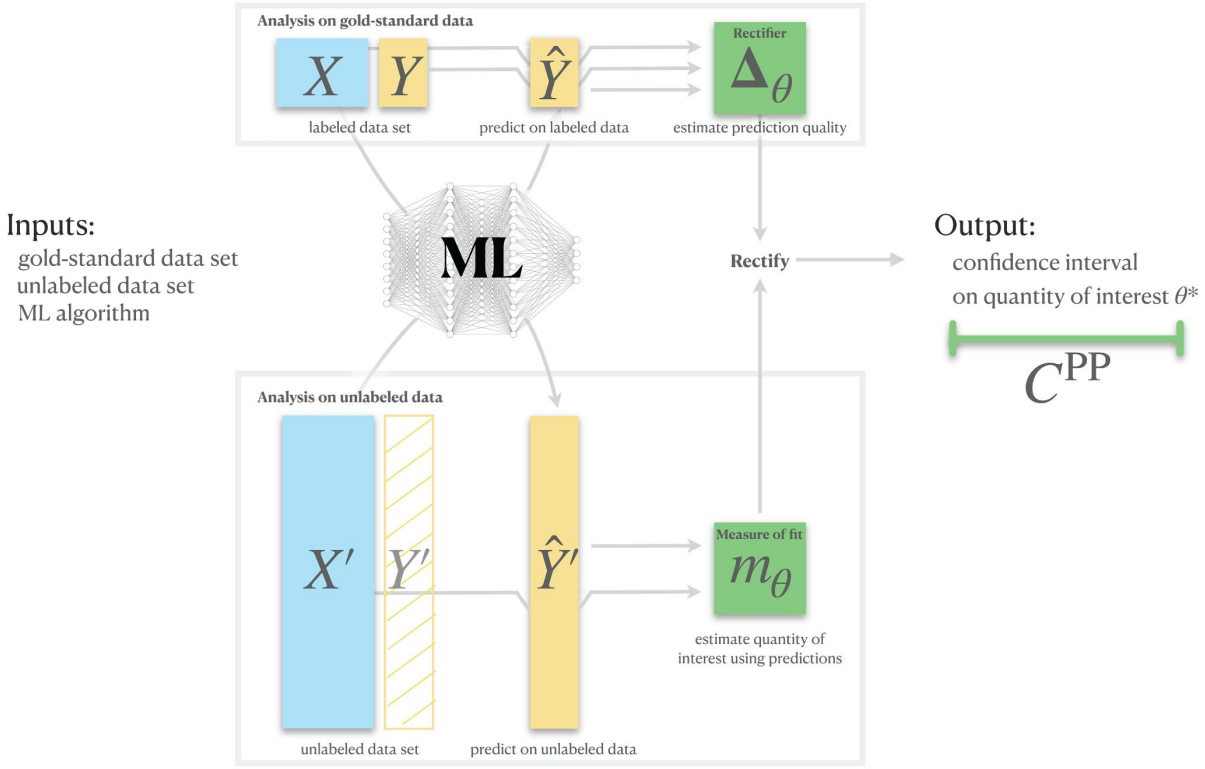


https://arxiv.org/pdf/2408.14348

Prediction-Powered & Active Inference

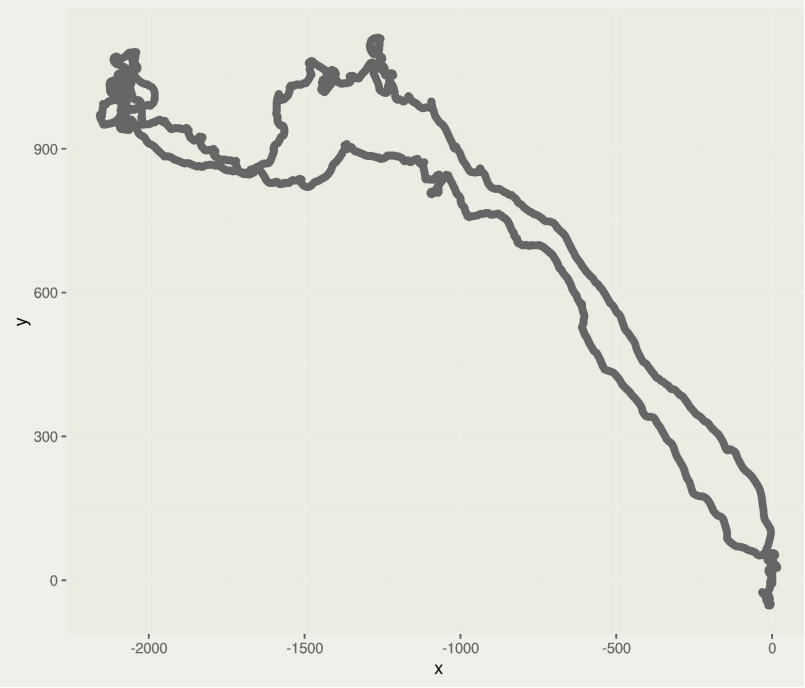
Use human labels to re-calibrate prediction-based inference rather than retrain model.

Can use active sampling to choose which data points to label.



https://www.science.org/doi/10.1126/science.adi6000

Sometimes we *can't* get ground truth



What is this animal doing?

Always – Class 1: Possible to distinguish in all photos		
Usually – Class 2: Possible to distinguish with most pictures		
Rarely – Class 3: Only possible to identify with ideal pictures, which are rare		
Never – Class 4. Impossible to distinguish with typical camera trap or citizen science pictures.		

https://marieetienne.github.io/Talks/_presentation/#1

https://academic.oup.com/jmammal/article/103/4/767/6564439