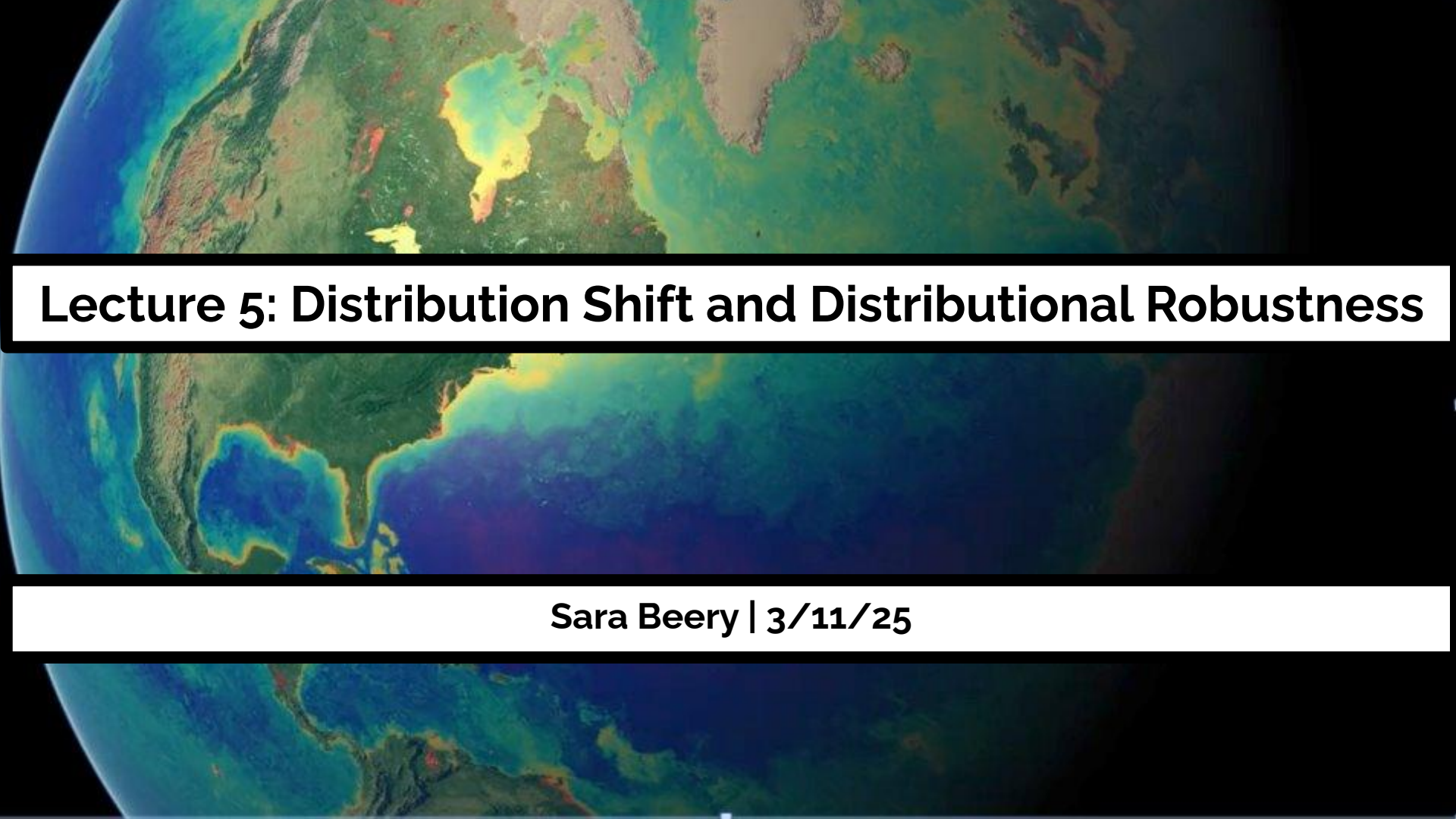
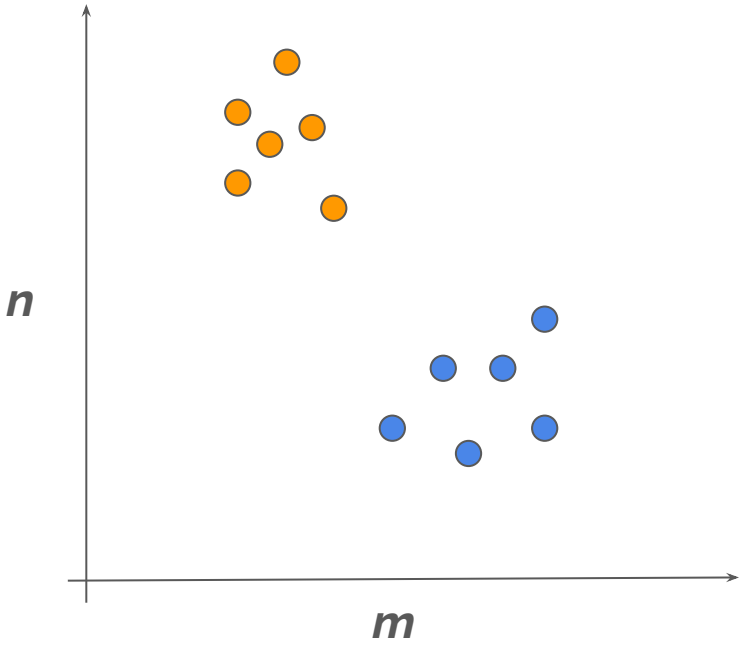


Lecture 5: Distribution Shift and Distributional Robustness

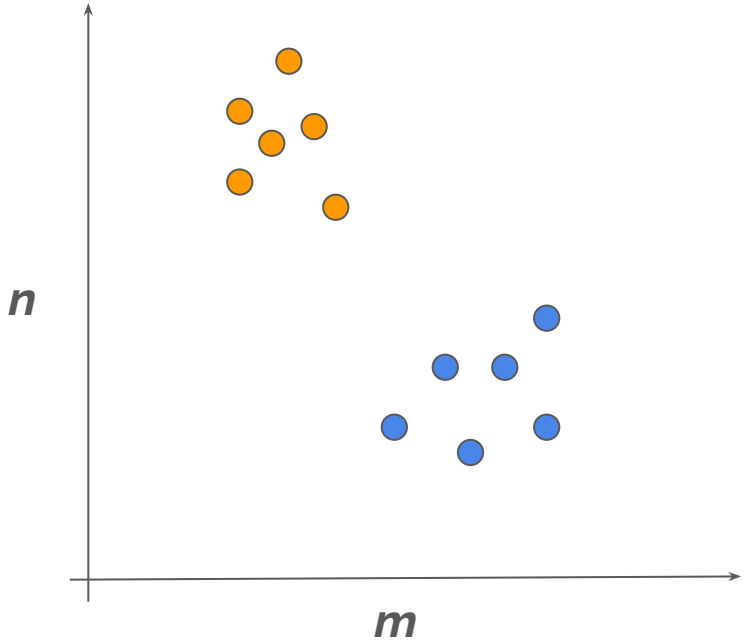
Sara Beery | 3/11/25

Learning models from data



Let's consider a simple case: learning to categorize points on a plane.

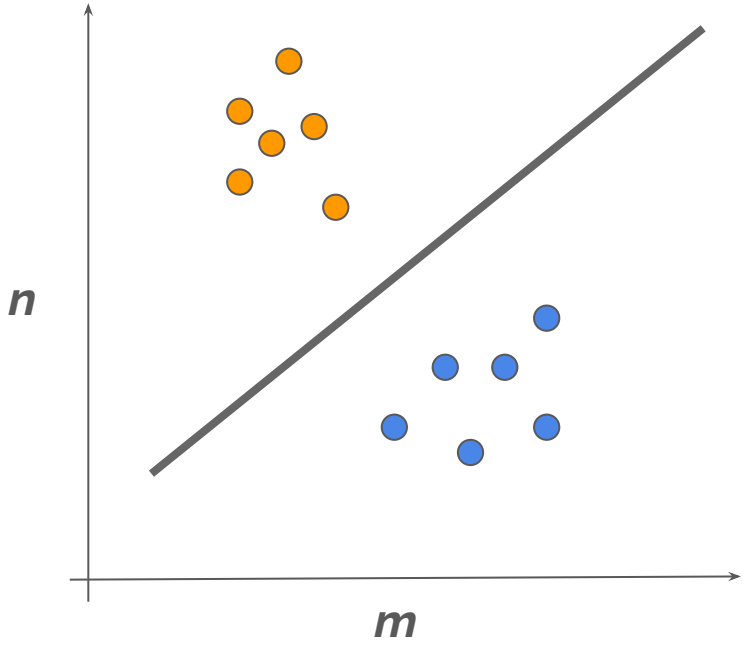
Learning models from data



Let's consider a simple case: learning to categorize point on a plane.

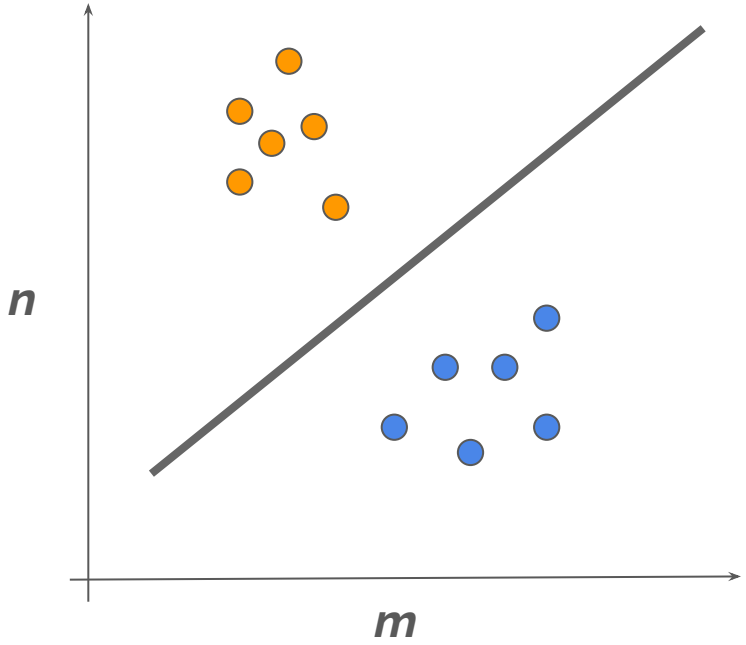
We assume we have **representative** data - aka **IID** - with labels

Learning models from data



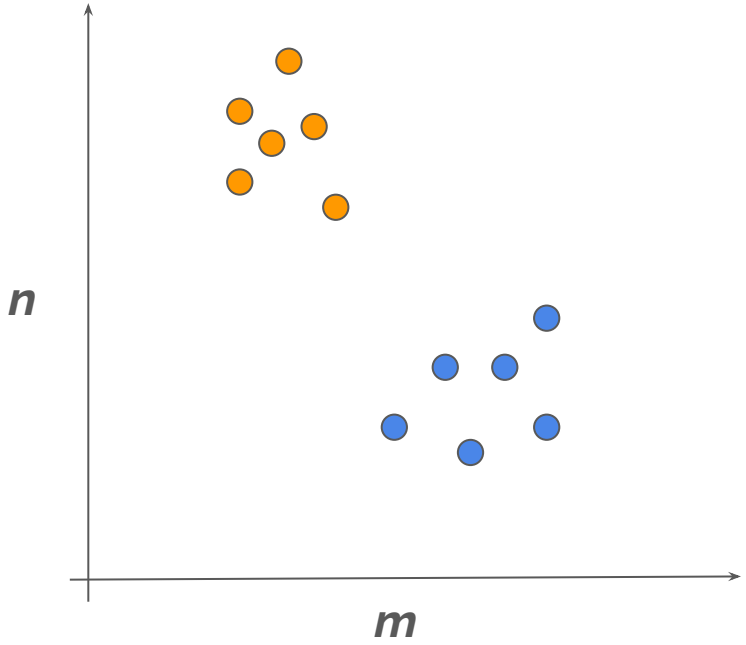
Using our labeled data, we learn this linear classifier

Learning models from data



How do we know if our model works?

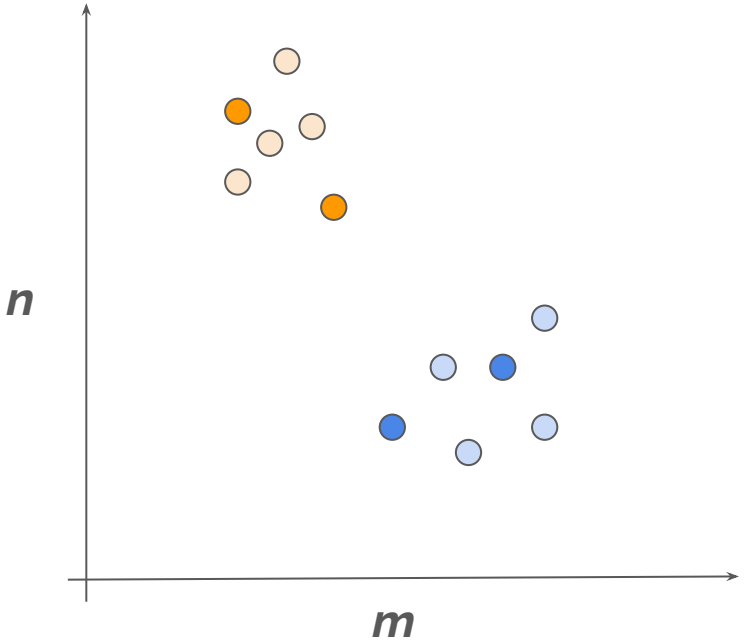
Learning models from data



How do we know if our model works?

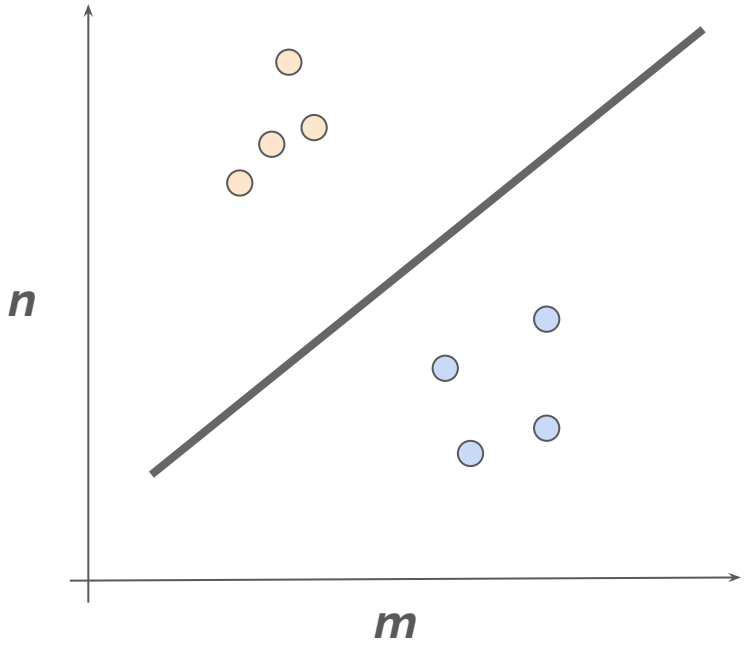
- We could get more data, label it, and test our model on the new data
- We can *pretend* we don't have some of the data during training and test the model on that data (most common)

Learning models from data



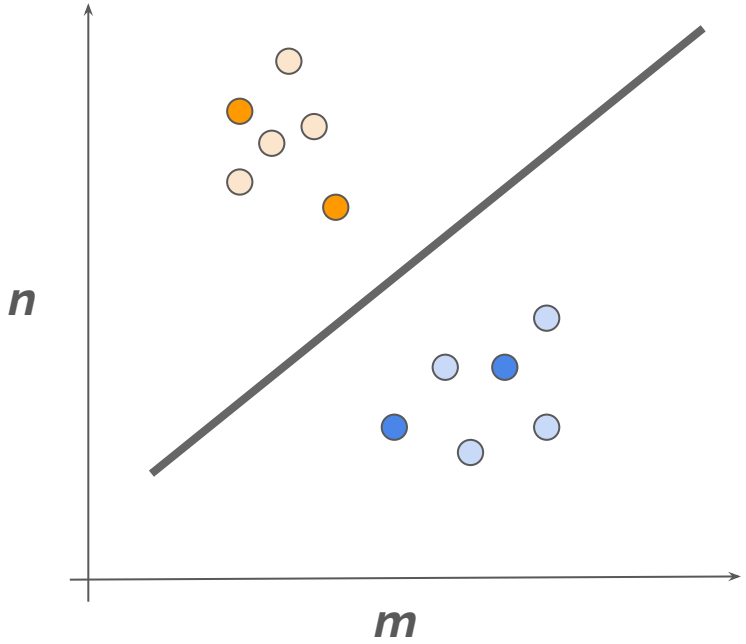
Since the data is assumed to be IID, we take a random subset to use as evaluation data

Learning models from data



Now we would learn our model using only the “training data”

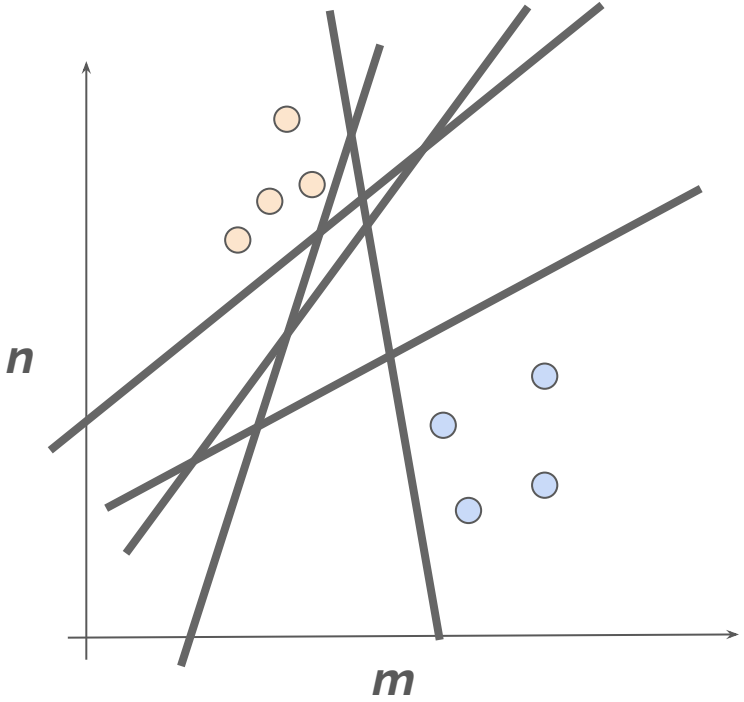
Learning models from data



Now we would learn our model using only the “training data”

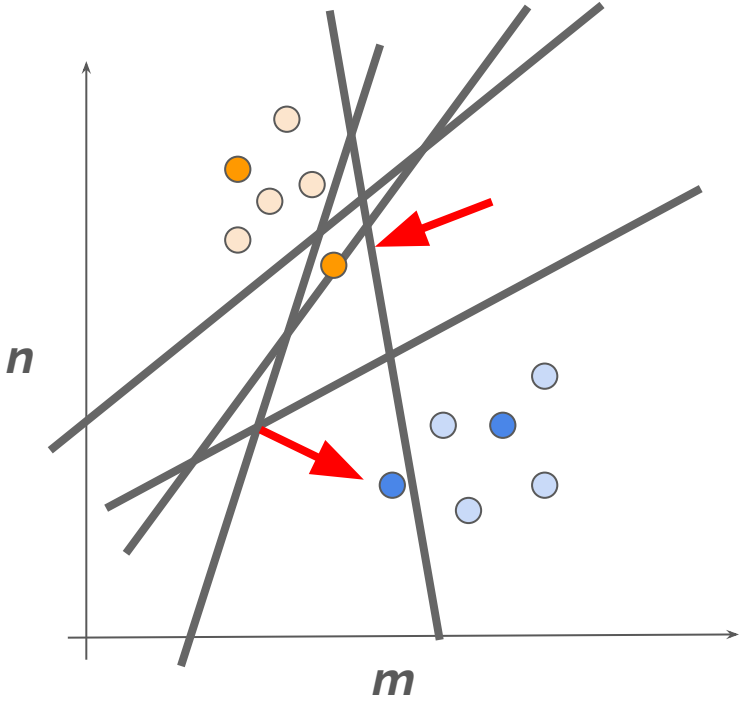
And then see if our model is a good classifier by checking our predictions on the “validation data” are correct.

Learning models from data



Note that there are lots of possible linear classifiers that would perfectly solve the training set

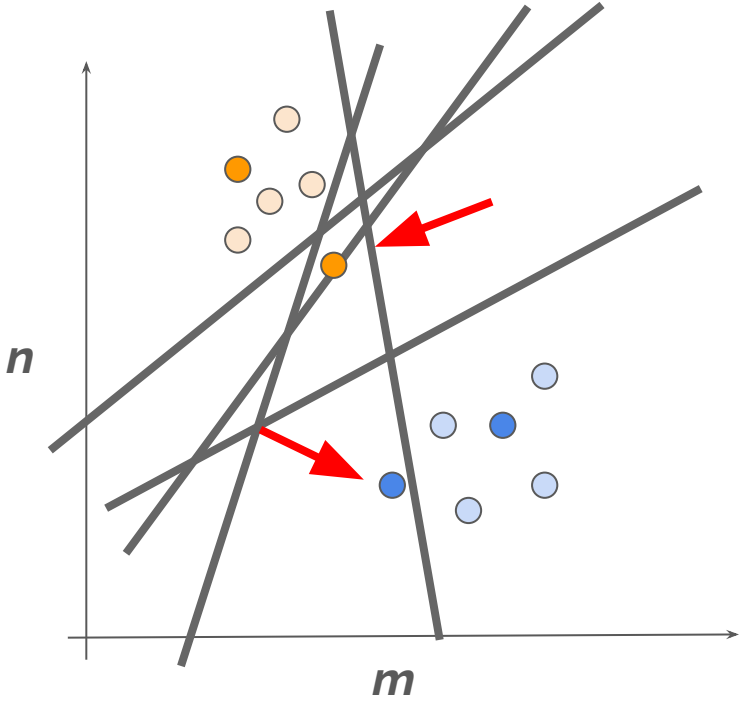
Learning models from data



Note that there are lots of possible linear classifiers that would perfectly solve the training set

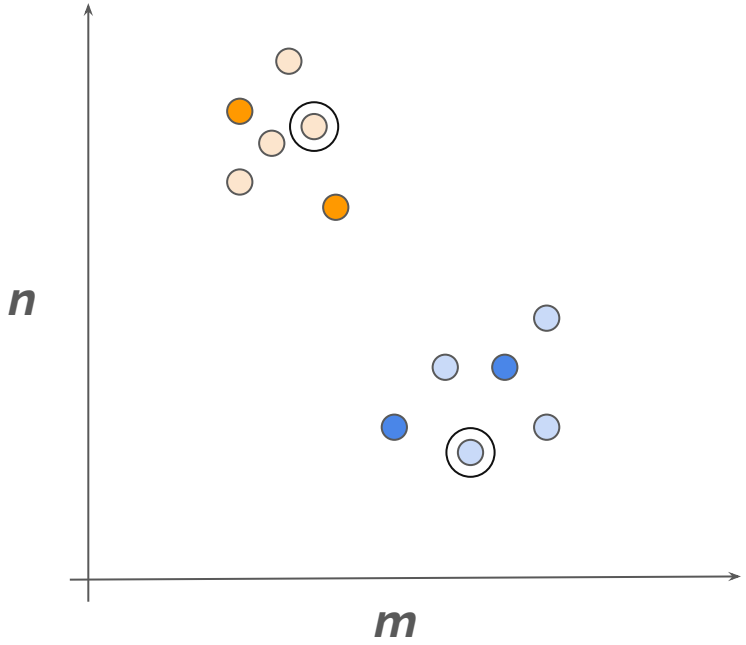
But not all of these would work well on the validation data

Learning models from data



If we decide which classifier to use based on validation data performance, then that means we're using some of our data for ***model selection***.

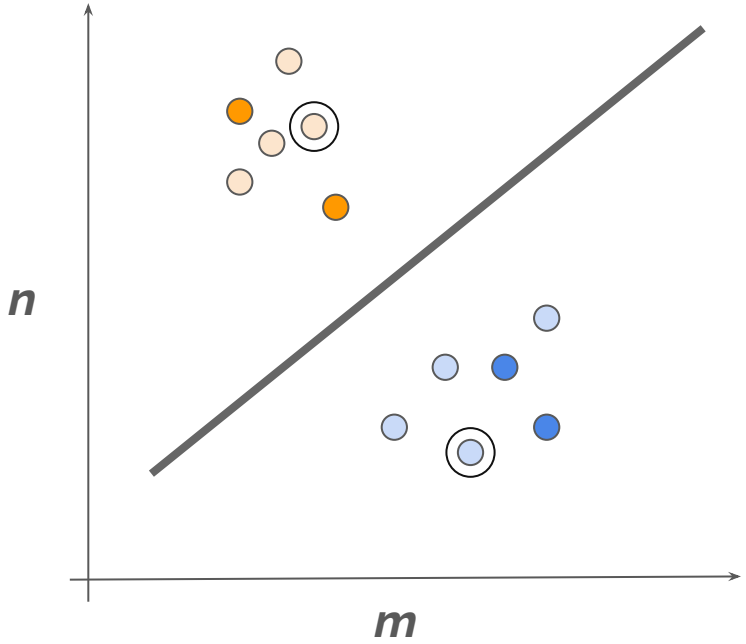
Learning models from data



This means that our validation dataset isn't truly unseen, so to maintain the integrity of our model evaluation, we further subsample the dataset to include a "test set", which we shouldn't ever look at until we've chosen our final model.

(note, ML researchers/subfields are **VERY BAD AT THIS**)

Learning models from data

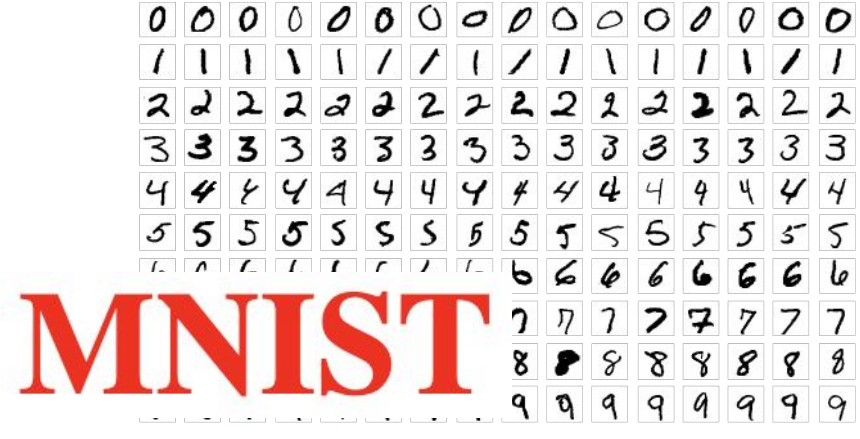


This combination of

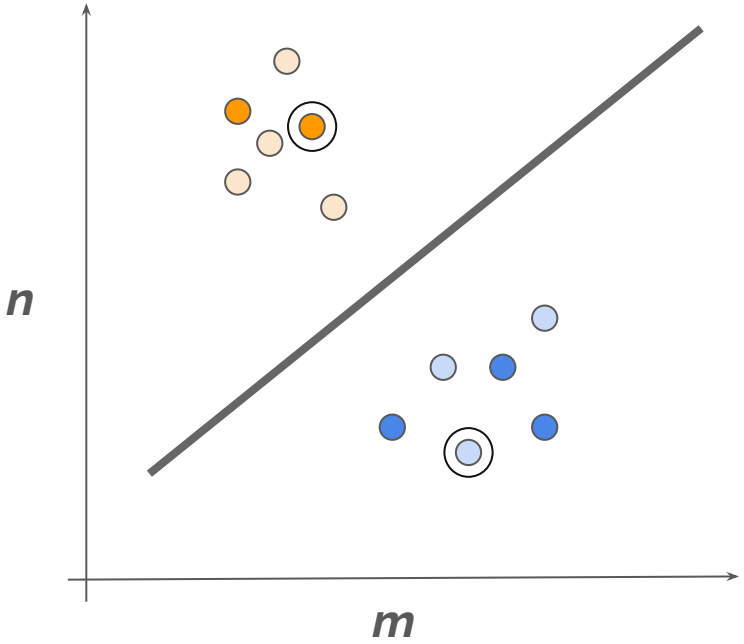
1. A dataset
2. A task (and how to evaluate it)
3. A train/val/test data split

Is called a “benchmark”

Benchmarks have played an integral role in CVML research

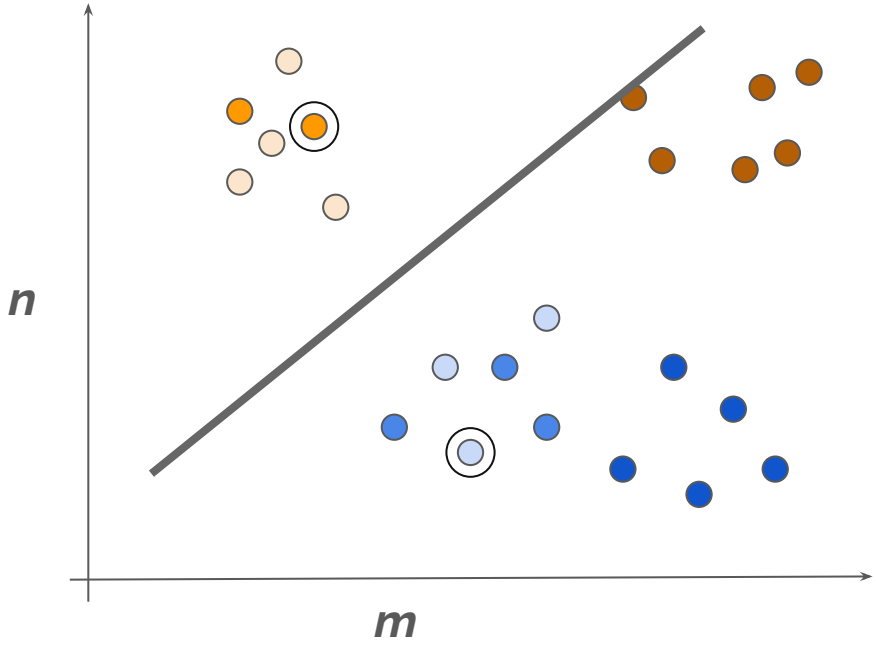


Learning models from data



What if our initial assumption was wrong and the labeled data we started with wasn't representative?

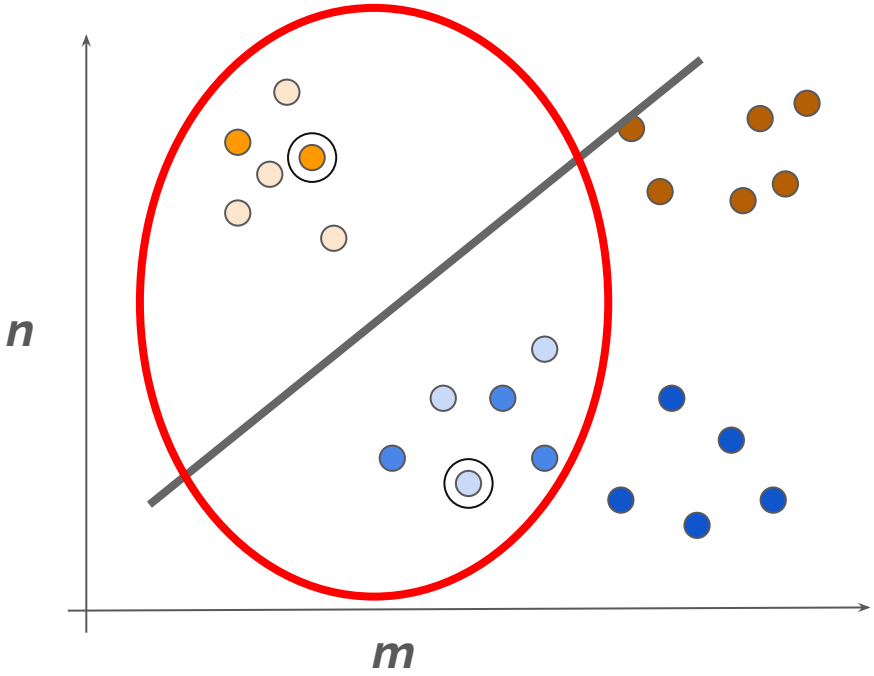
Learning models from data



Now say we try to use our model, and the data we use it on is sampled with larger values of *m*.

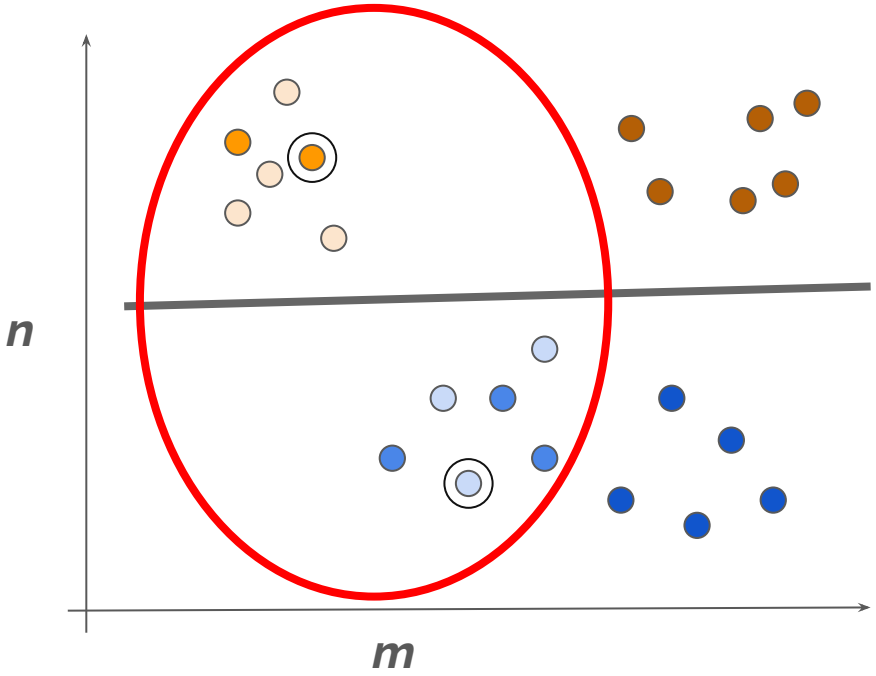
It turns out our model is **dead wrong**, it would classify all these points as blue.

Learning models from data



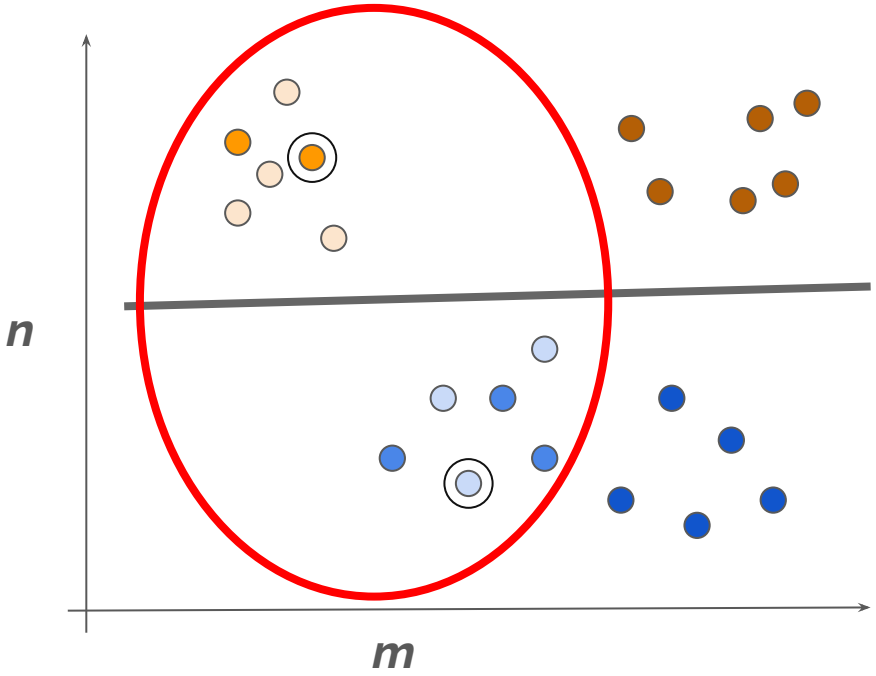
The “domain” of the original data, fell within the red circle, and the new data was *out of domain* (**OOD**)

Learning models from data



Ideally we would learn a classifier that would work on **all possible data**, but that's really hard if we've never seen data outside the red circle before

Learning models from data



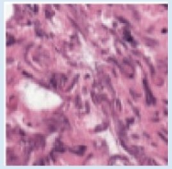
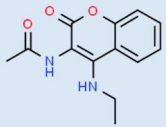
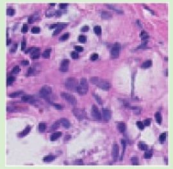
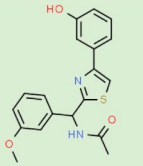
In ML, this challenge is often called "**distribution shift**" and the goal for models that handle it well is "**generalization**" or "**robustness**"

Domain shifts are ubiquitous in real-world scenarios

WILDS

https://wilds.stanford.edu/

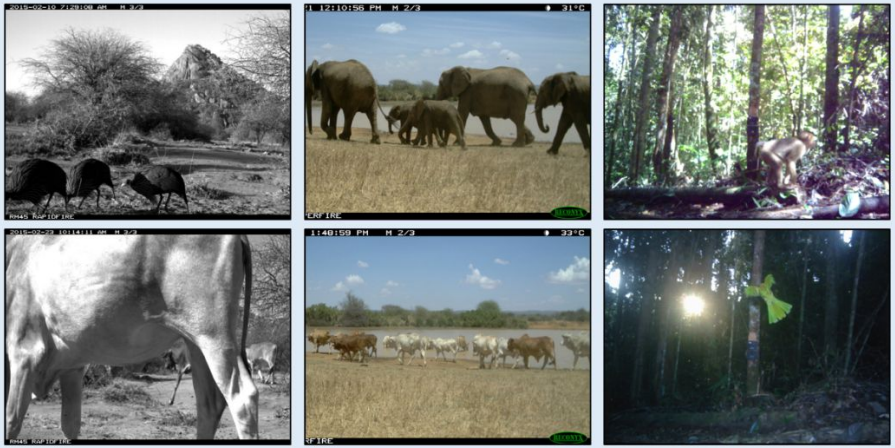
Pang Wei Koh*, Shiori Sagawa*, Henrik Marklund, Sang Michael Xie, Marvin Zhang,
Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery,
Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang

	Camelyon17	iWildCam	PovertyMap	FMoW	Amazon	CivilComments	OGB-MolPCBA
Shift	Hospitals	Locations	Countries	Time	Users	Demographics	Scaffold
Train					Overall a solid package that has a good quality of construction for the price.	What do Black and LGBT people have to do with bicycle licensing?	
Test					I *loved* my French press, it's so perfect and came with all this fun stuff!	As a Christian, I will not be patronizing any of those businesses.	
Adapted from	Bandi et al. 2018	Beery et al. 2020	Yeh et al. 2020	Christie et al. 2018	Ni et al. 2019	Borkan et al. 2019	Hu et al. 2020

We can build evaluation frameworks that measure ID vs OOD performance

Training data

Camera 1 | Camera 2 | … | Camera 245



Out-of-distribution (OOD) test data

Camera 246

…

Control: In-distribution (ID) test data

Camera 1 | Camera 2 | … | Camera 245

Macro F1

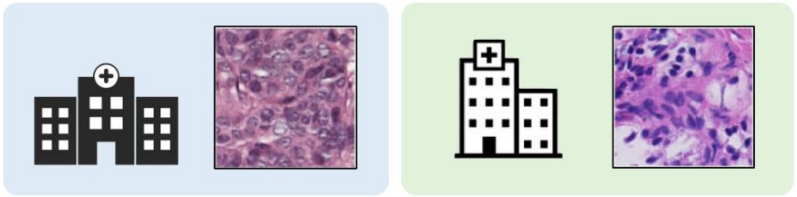
ID 47.0% → -16.0% → OOD 31.0%

[Beery et al., 2020; Koh et al., 2021]

Slide from Shiori Sagawa

Domain shifts are ubiquitous in real-world scenarios

shifts across hospitals in histopathology



ID accuracy 93.2% → -22.9% → OOD accuracy 70.3%

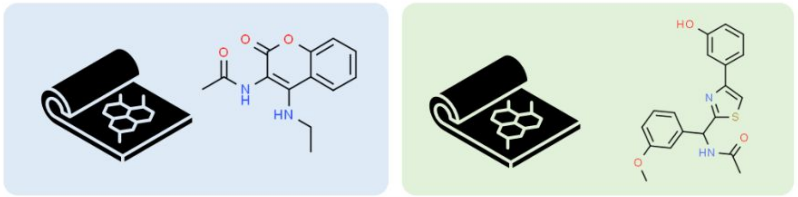
shifts across time in satellite imagery

ID accuracy 48.6% → -16.3% → OOD accuracy 32.3%

shifts across regions in wheat head detection

ID accuracy 63.3% → -13.7% → OOD accuracy 49.6%

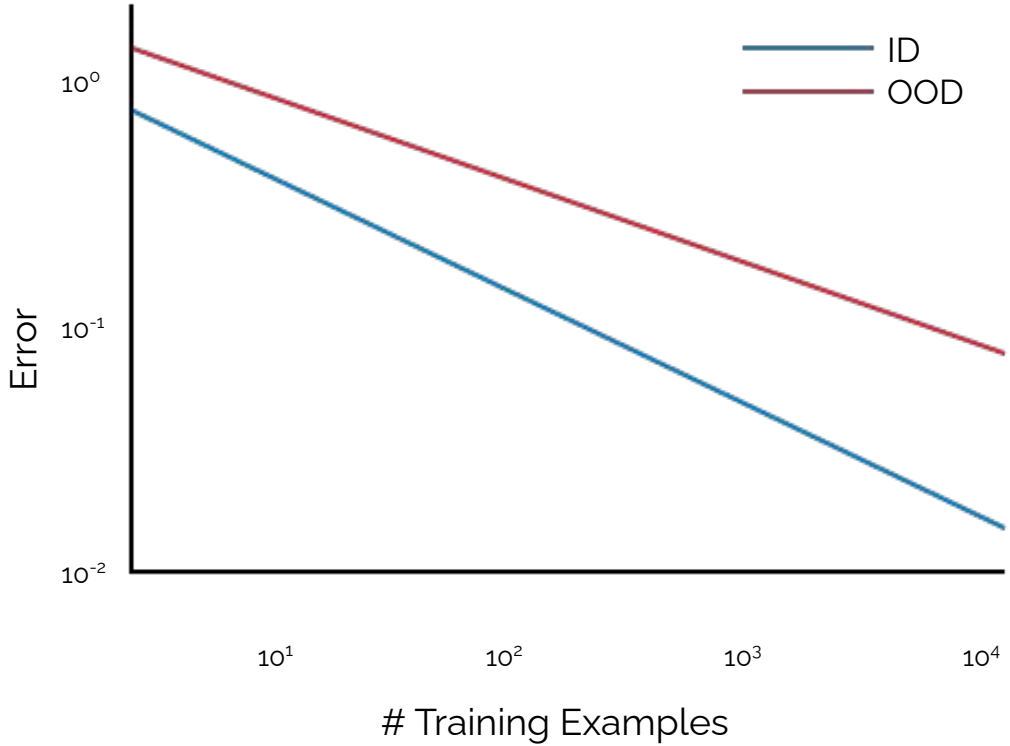
shifts across scaffold in bioassay prediction



ID AP 34.4% → -7.2% → OOD AP 27.2%

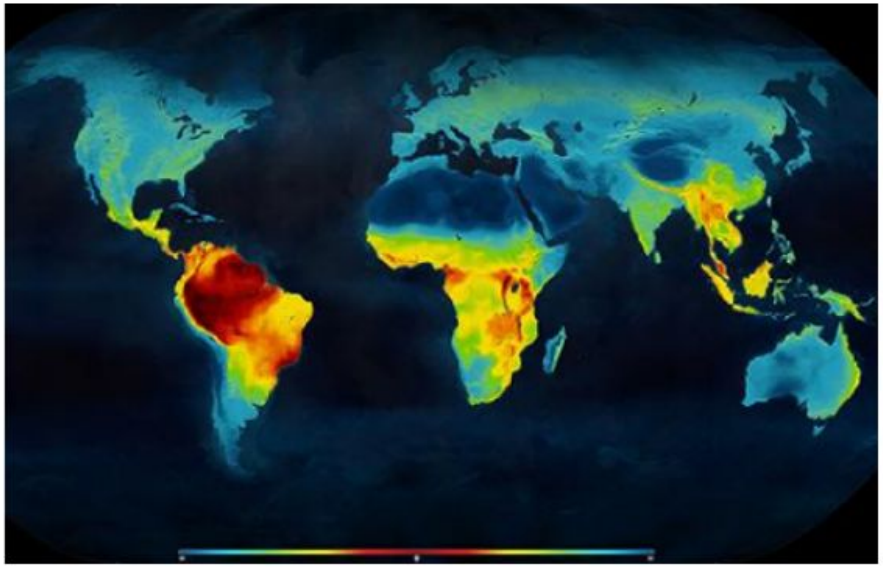
[Koh et al., 2021]

Performance degrades OOD even for common species

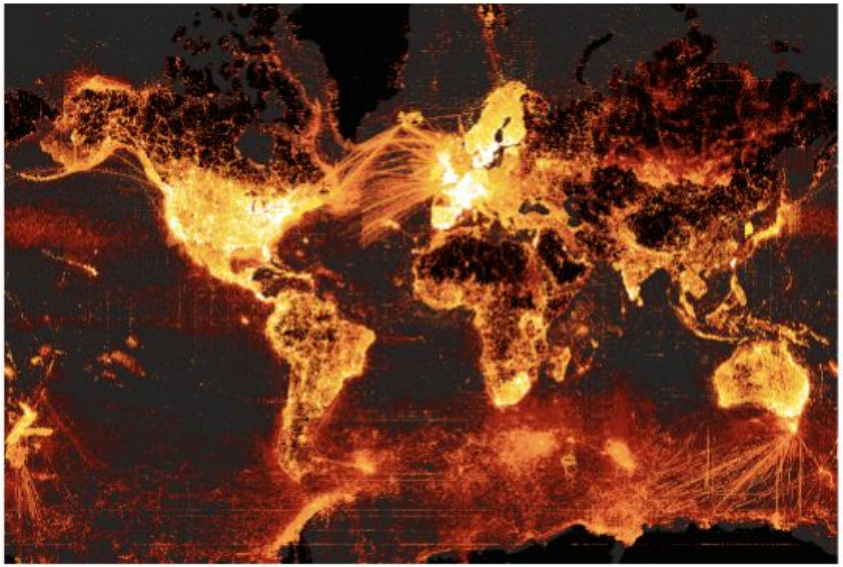


Recognition in Terra Incognita, Beery et al., ECCV 2018

Real-world data is never IID, we *always* have domain shift



Map of global biodiversity



Species occurrence data in GBIF

Scaling Biodiversity Monitoring for the Data Age - ACM XRDS 2021

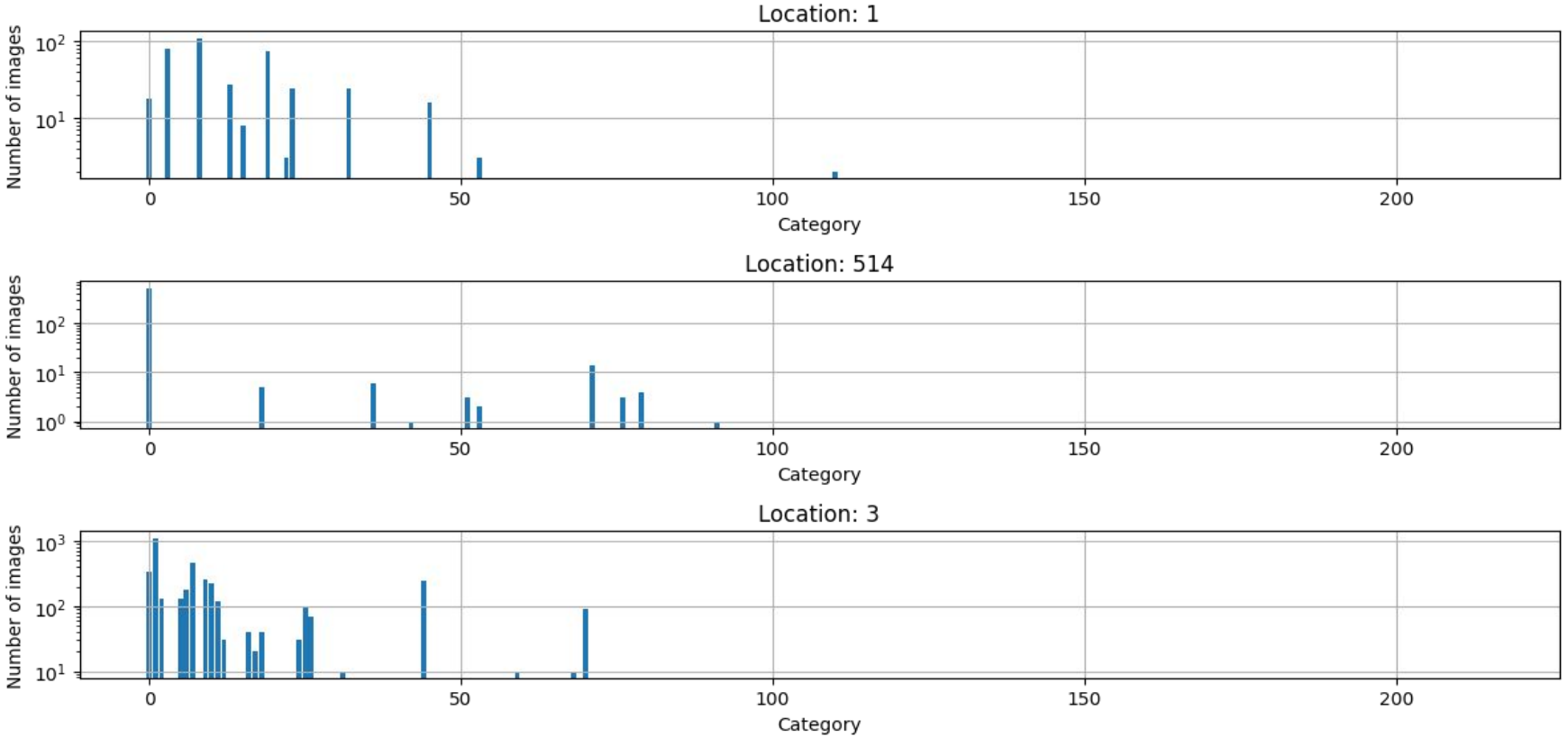
Why is this so hard?

Why is this so hard?

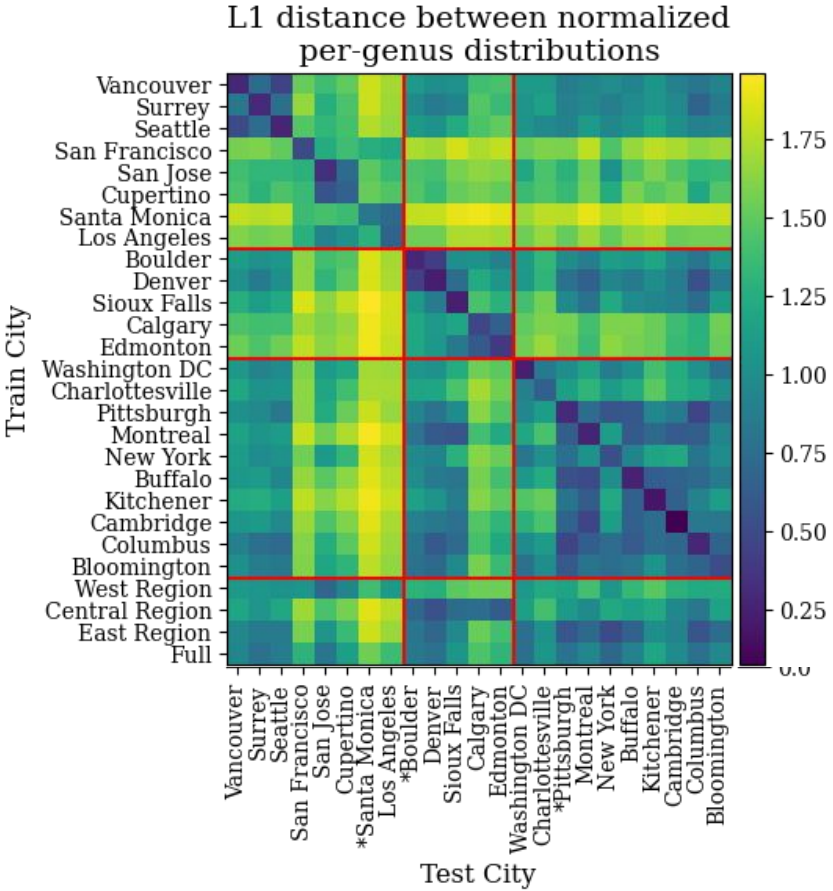
Subpopulation shift

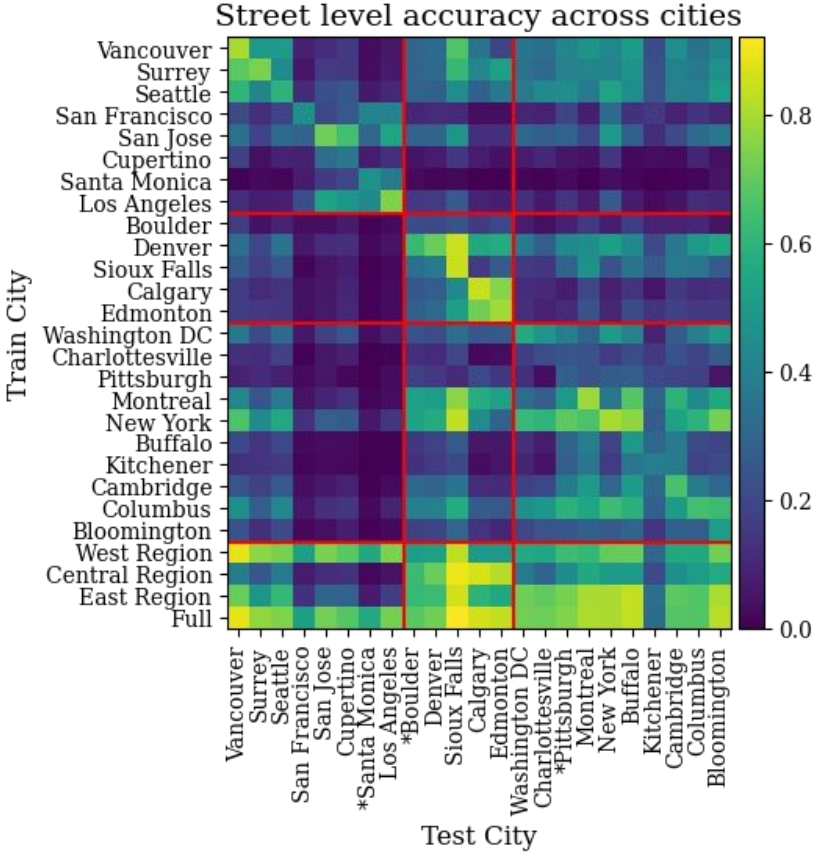
- Different class likelihoods within the same set of possible classes
- Different sets of possible classes (open set)

Class distribution is different for each static sensor location



Performance has strong correlation with distribution similarity in urban forests





Why is this so hard?

Subpopulation shift

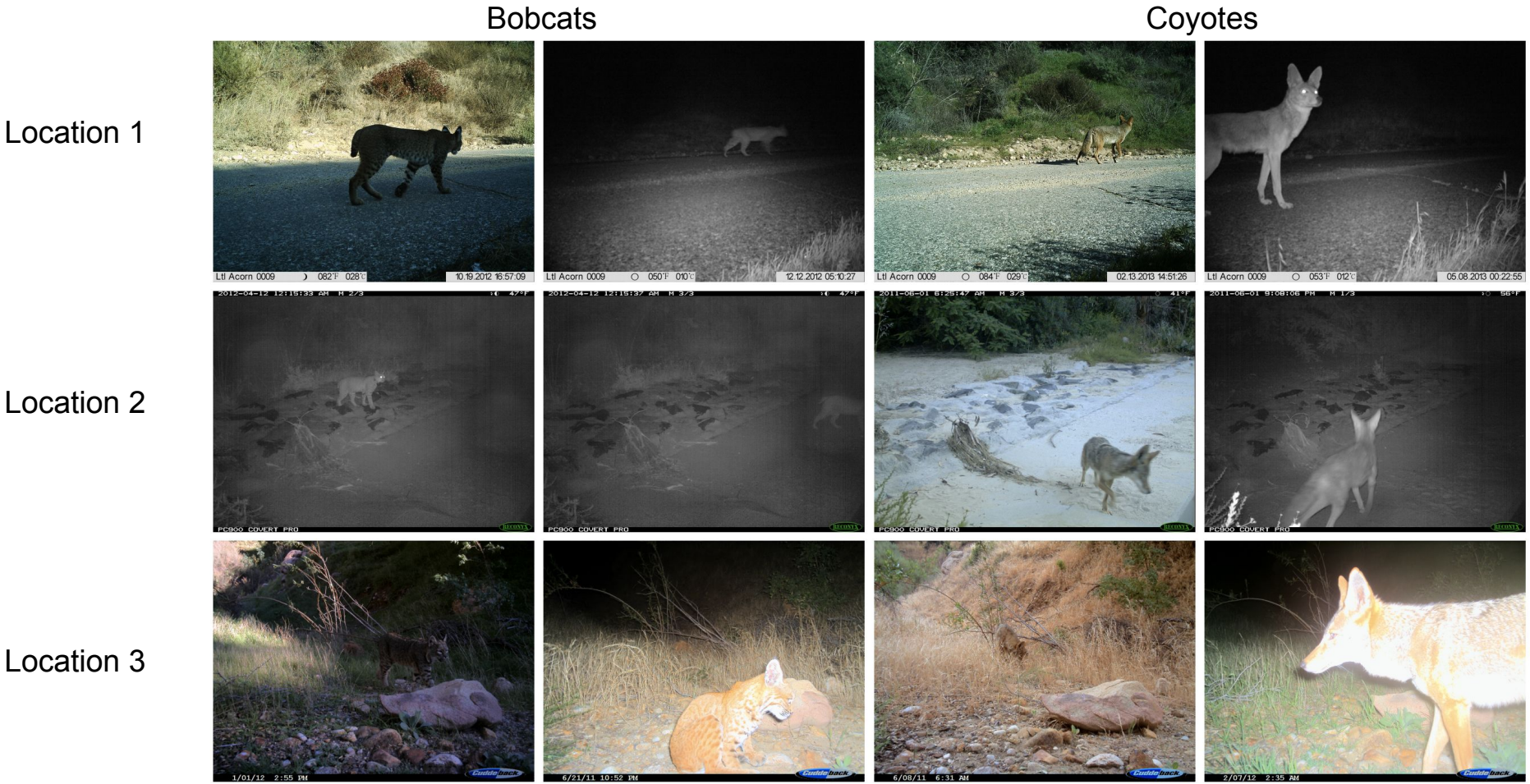
- Different species likelihoods within the same set of possible species
- Different sets of possible species (open set)

Visual shift

- Different camera angles
- Different sensor types
- Different backgrounds
- Different individuals
- Different ages
- Night/day
- Summer/winter
- Wet/dry
- …

High visual similarity in data from one static camera trap over long time periods

Visual differences across different sensors



Temporal/seasonal changes

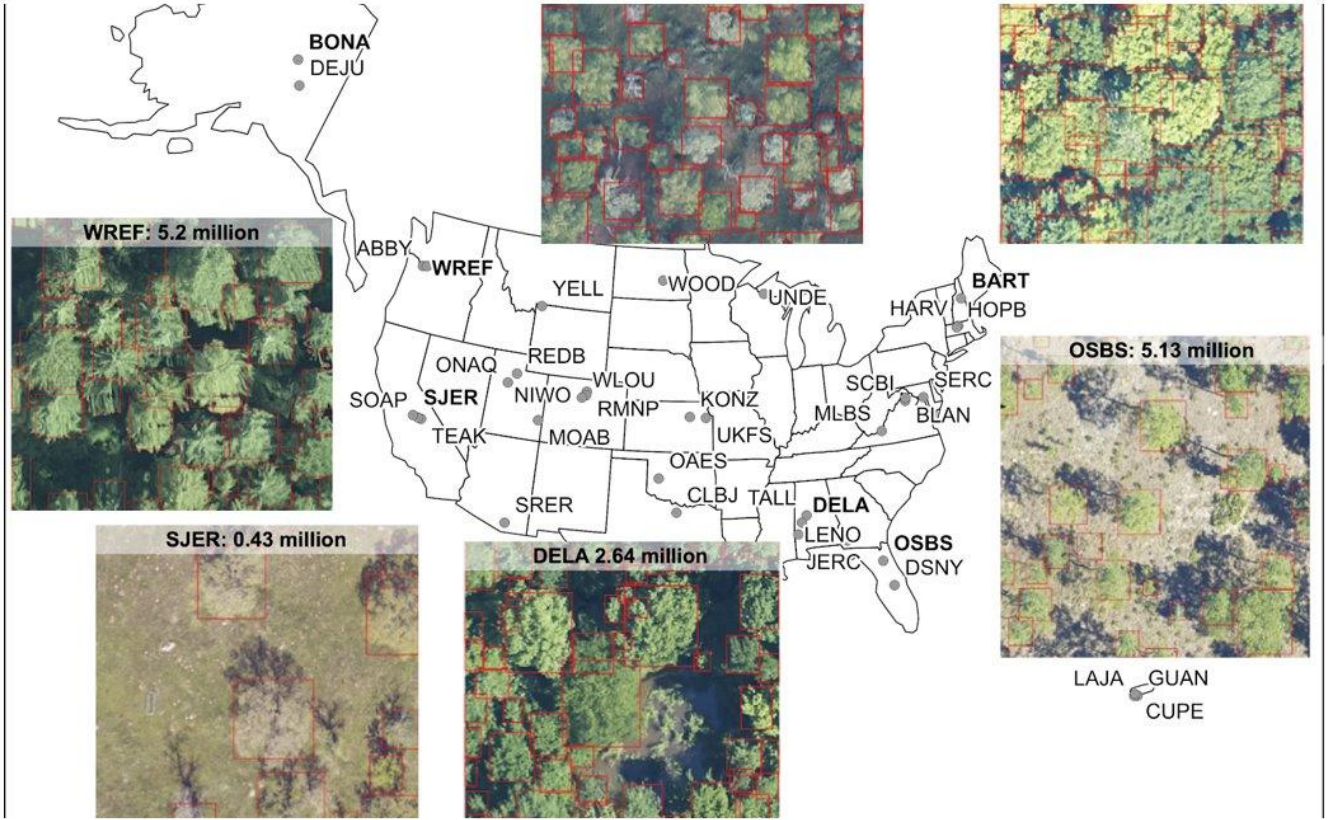

10/22/07 4:50 AM
Cuddeback


10/22/07 9:00 AM
Cuddeback


10/22/07 9:09 AM
Cuddeback


10/25/07 7:20 PM
Cuddeback

Different ecosystems have both subpopulation and visual shift



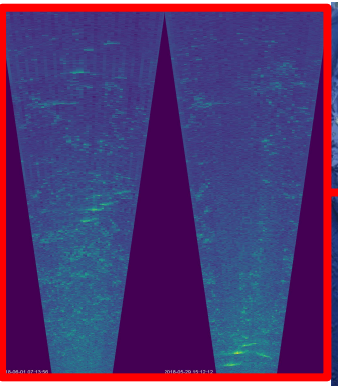
NEONCROWNS Dataset

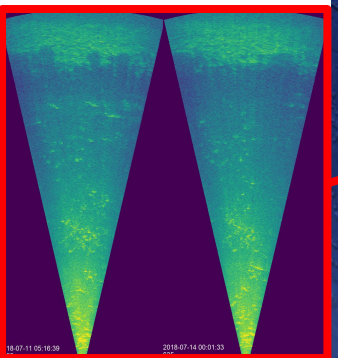
104,675,304 trees

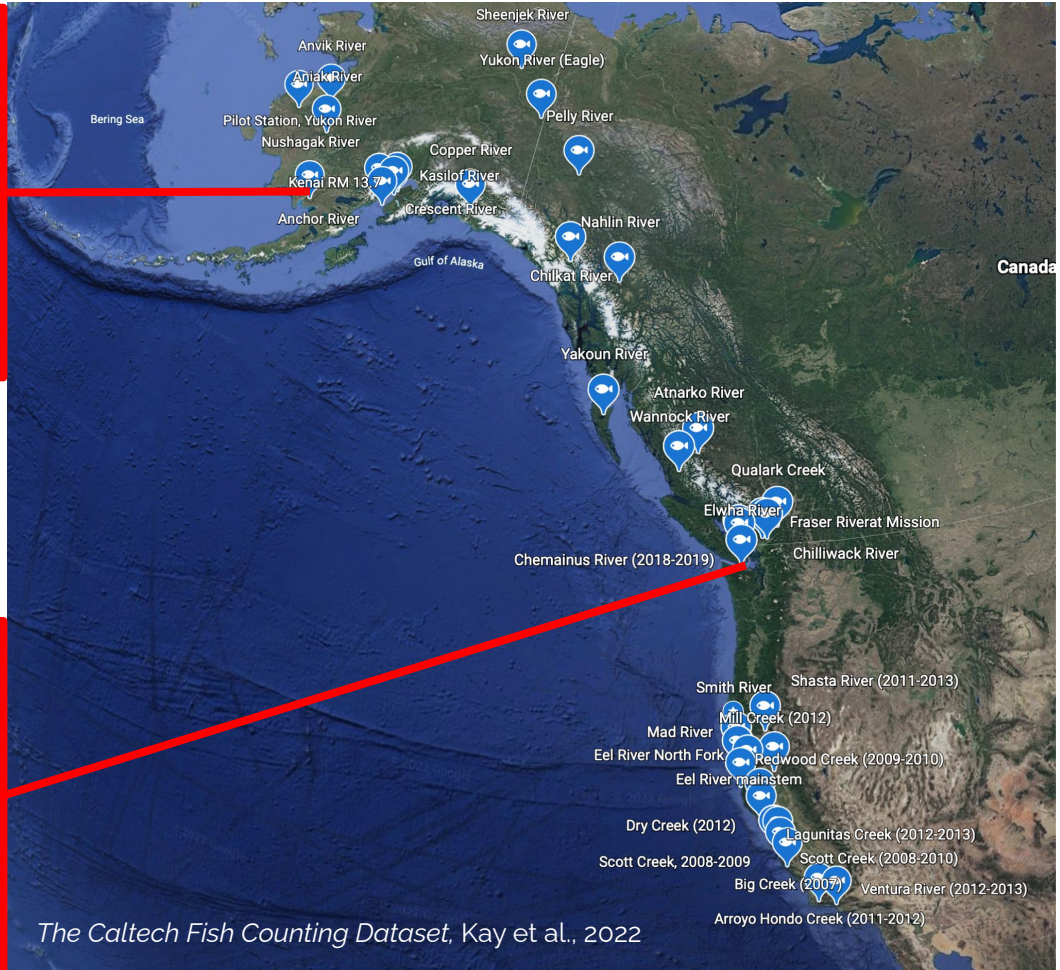
http://visualize.idtrees.org/

Weinstein et al., 2020

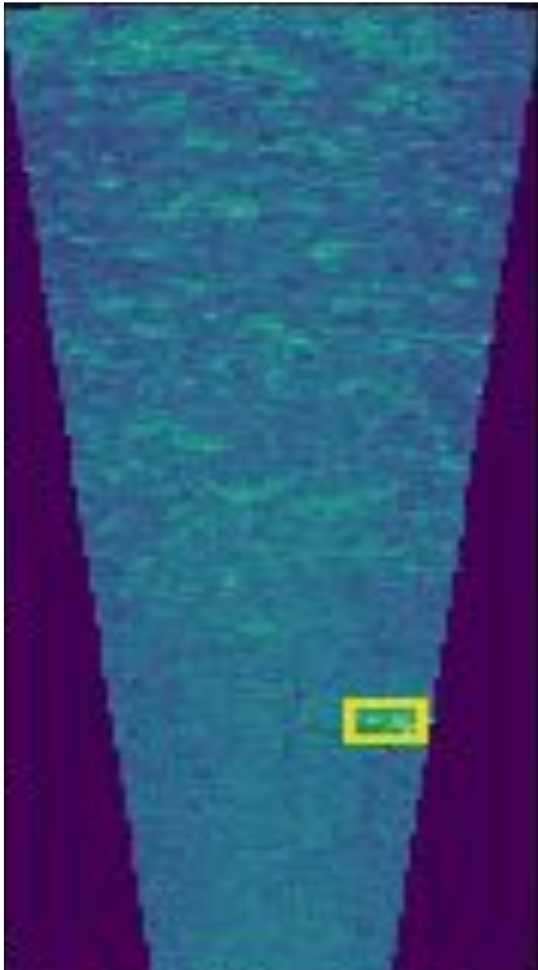
Detecting and counting salmon in static sonar



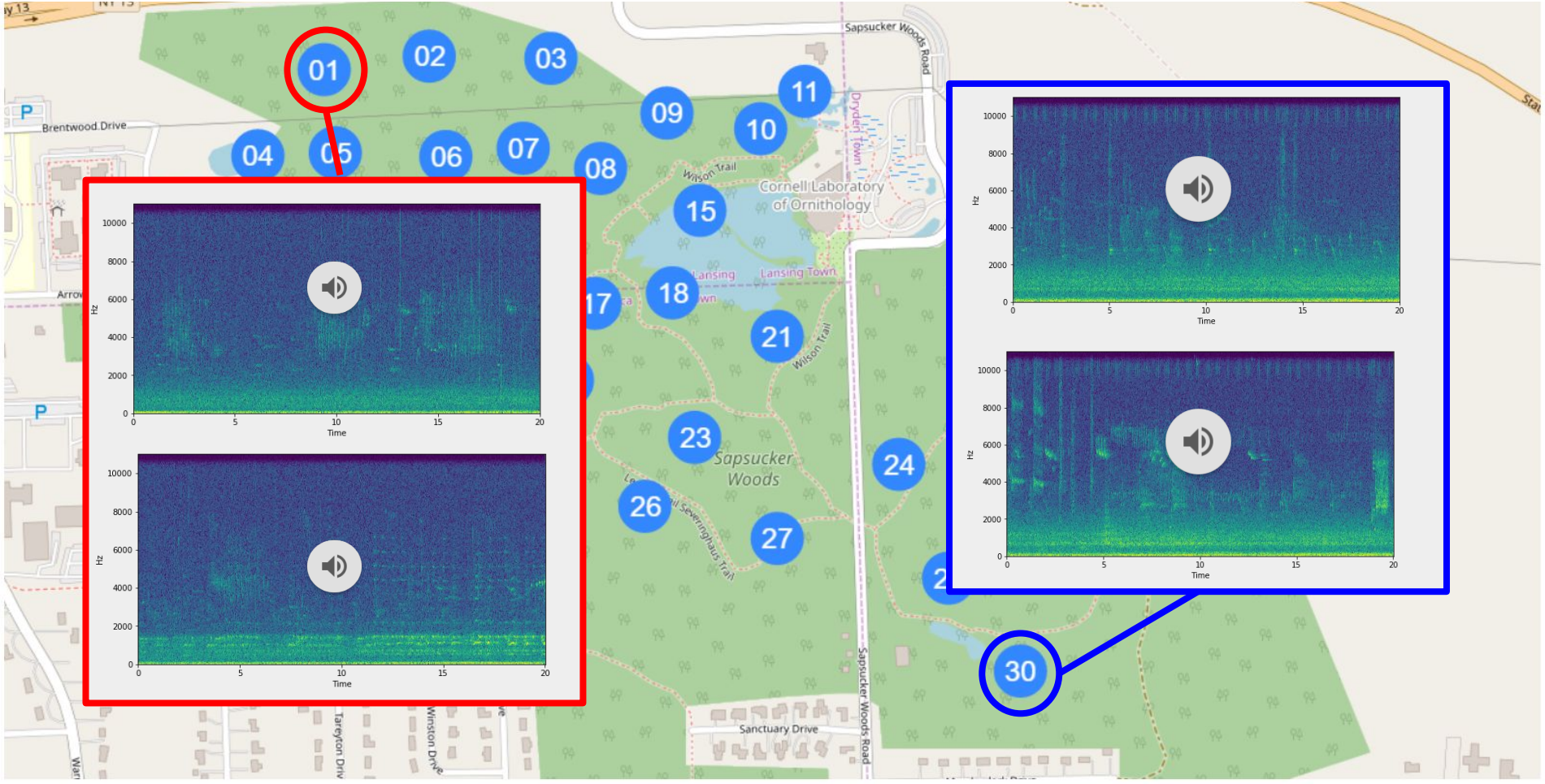




The Caltech Fish Counting Dataset, Kay et al., 2022



Detecting and categorizing birdsong in static bioacoustic sensors



What can we do about it?

- Data
  - Collect in-domain data, maybe with a human in the loop
  - Mimic in-domain data (augmentation, synthesis)
- Models
  - Disentangled representations (via localization (MegaDetector) or directly in the representation space a la Clarify: https://arxiv.org/abs/2402.03715)
  - Robust optimization (DRO, ERM)
- Priors
  - Geographic (SINR, TIML)
  - Or temporal, species, modality-specific, etc.