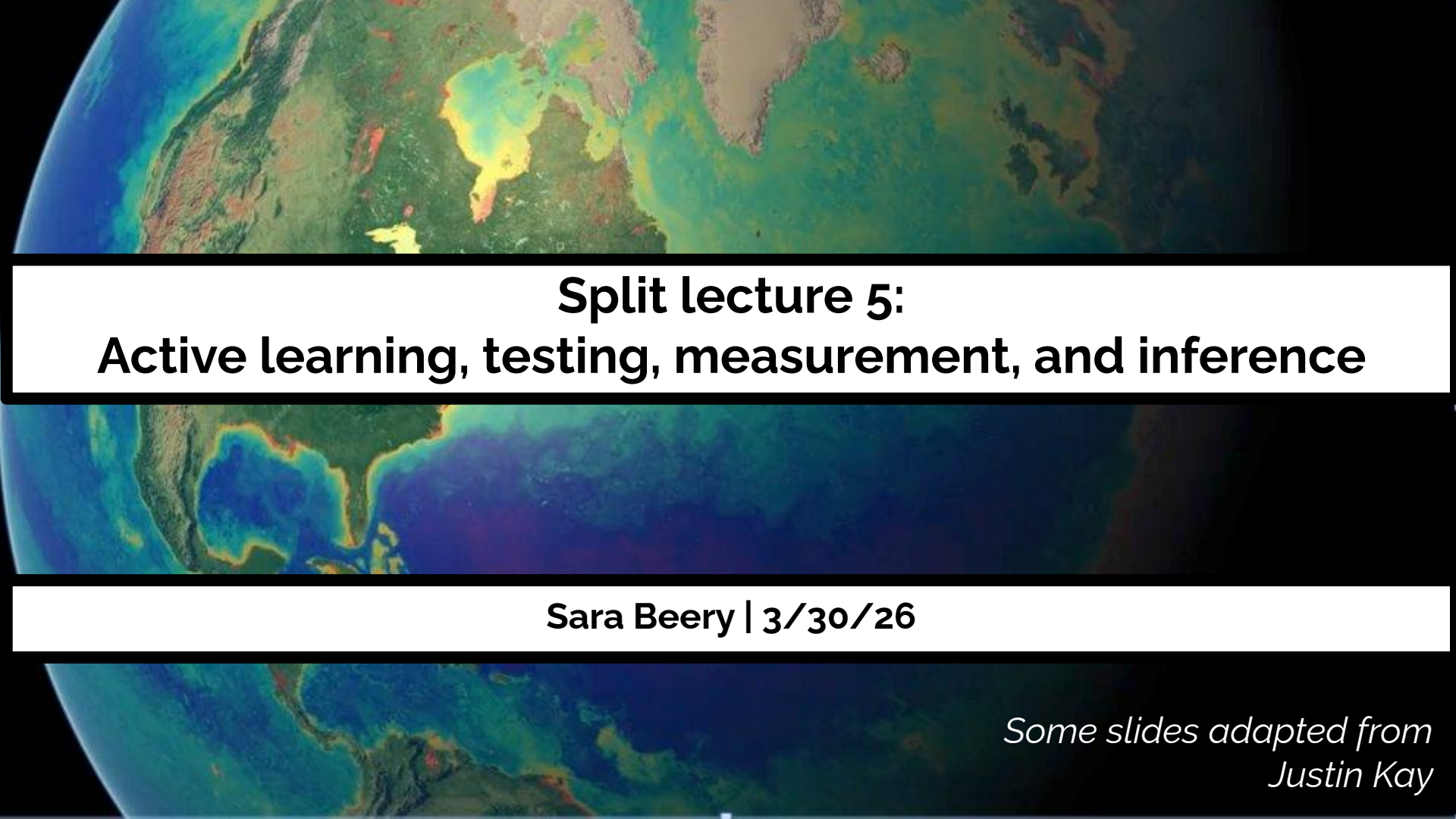
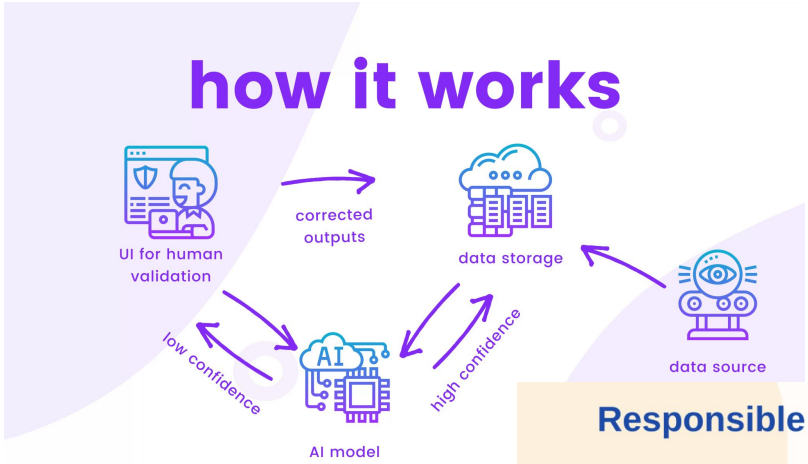


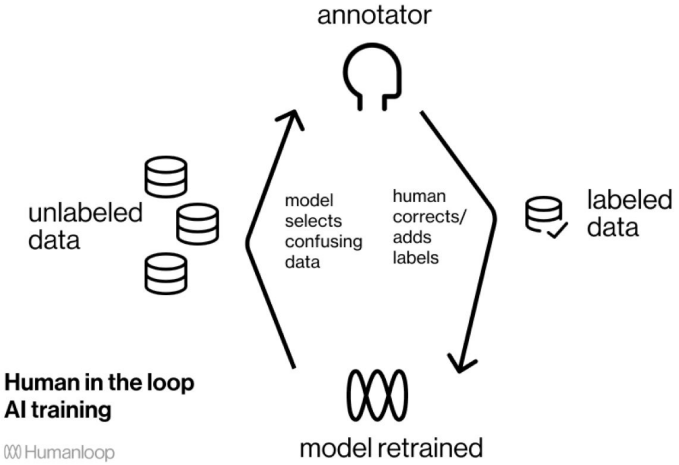
Split lecture 5:
Active learning, testing, measurement, and inference

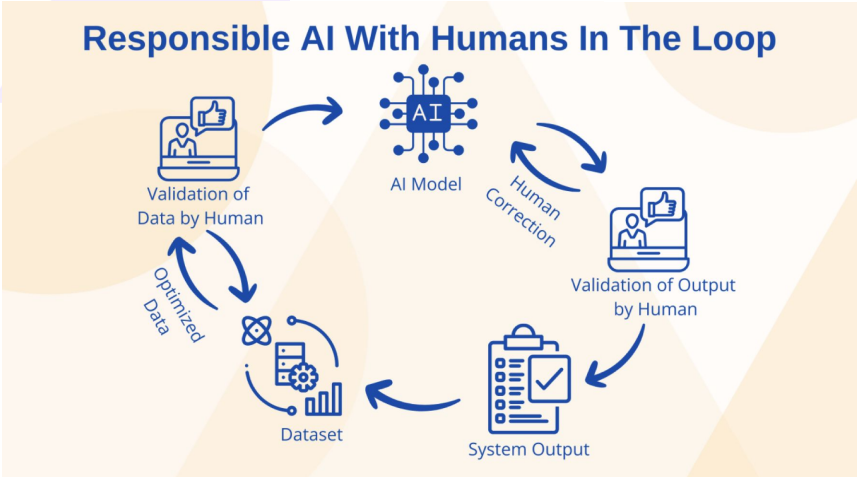
Sara Beery | 3/30/26

Some slides adapted from Justin Kay

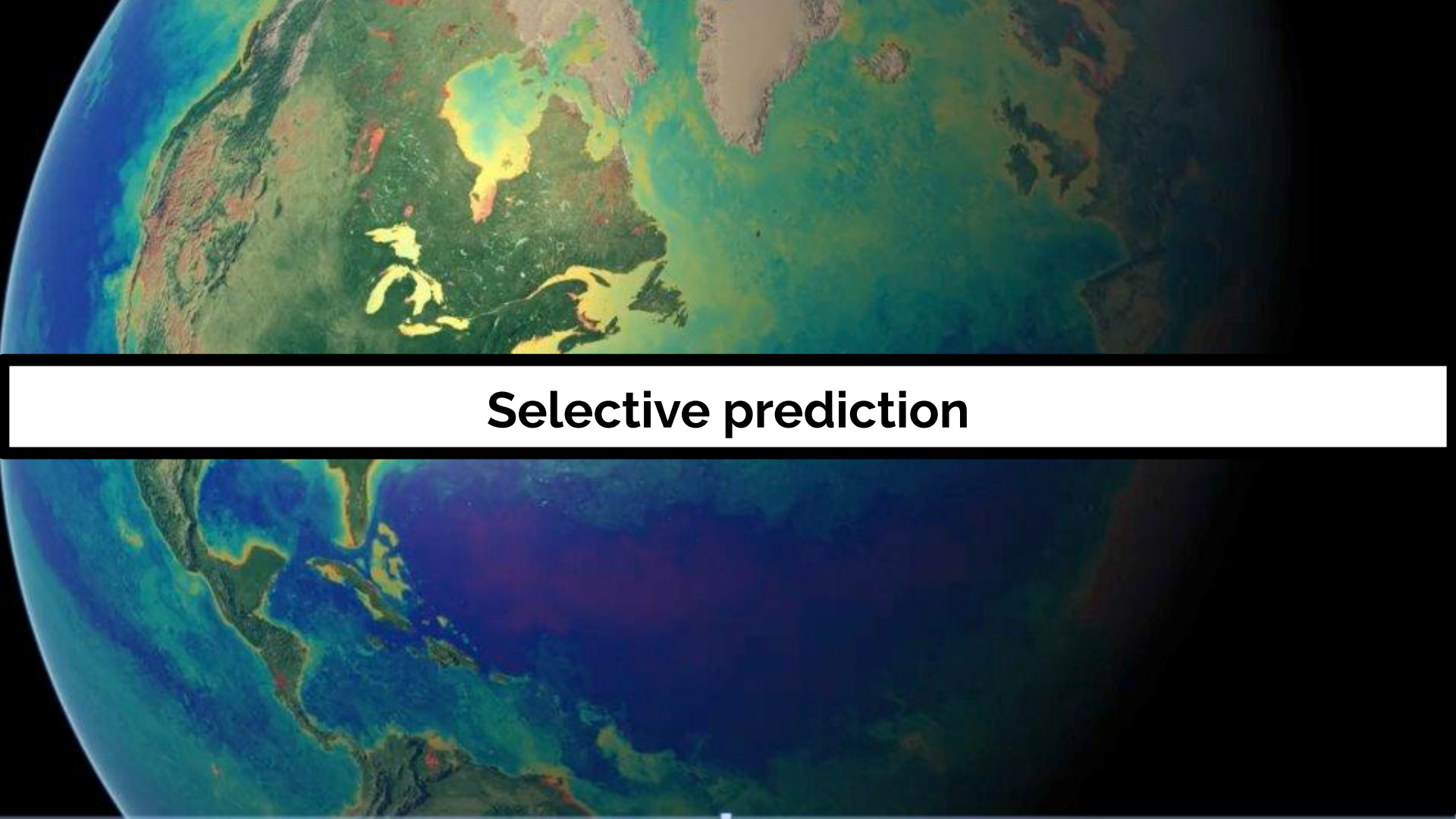
Bringing humans “in the loop”







Diagrams courtesy of industry PR:
SalesForce
HumanLoop
Humans in the Loop



Selective prediction

Prediction

$\hat{\mathbf{y}}$

$f_\theta : X \rightarrow \mathbb{R}^K$

$\mathbf{x}$

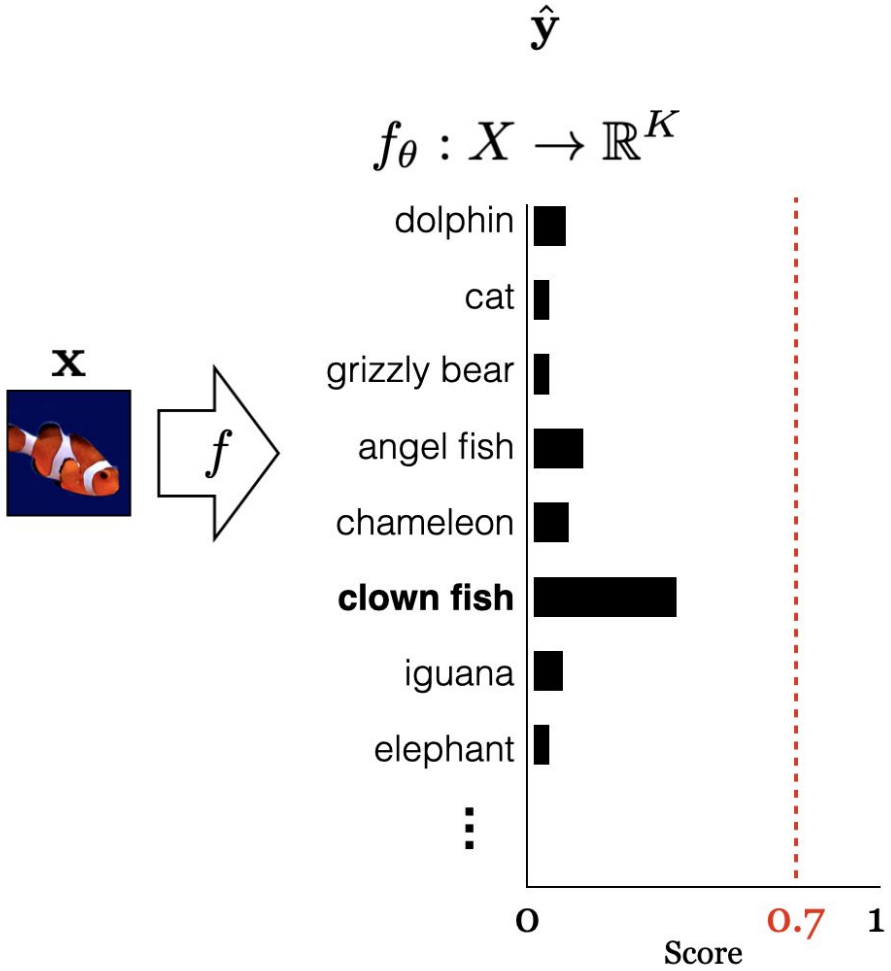
f

dolphin
cat
grizzly bear
angel fish
chameleon
clown fish
iguana
elephant
⋮

0 0.4 1
Score

"clown fish"

Prediction



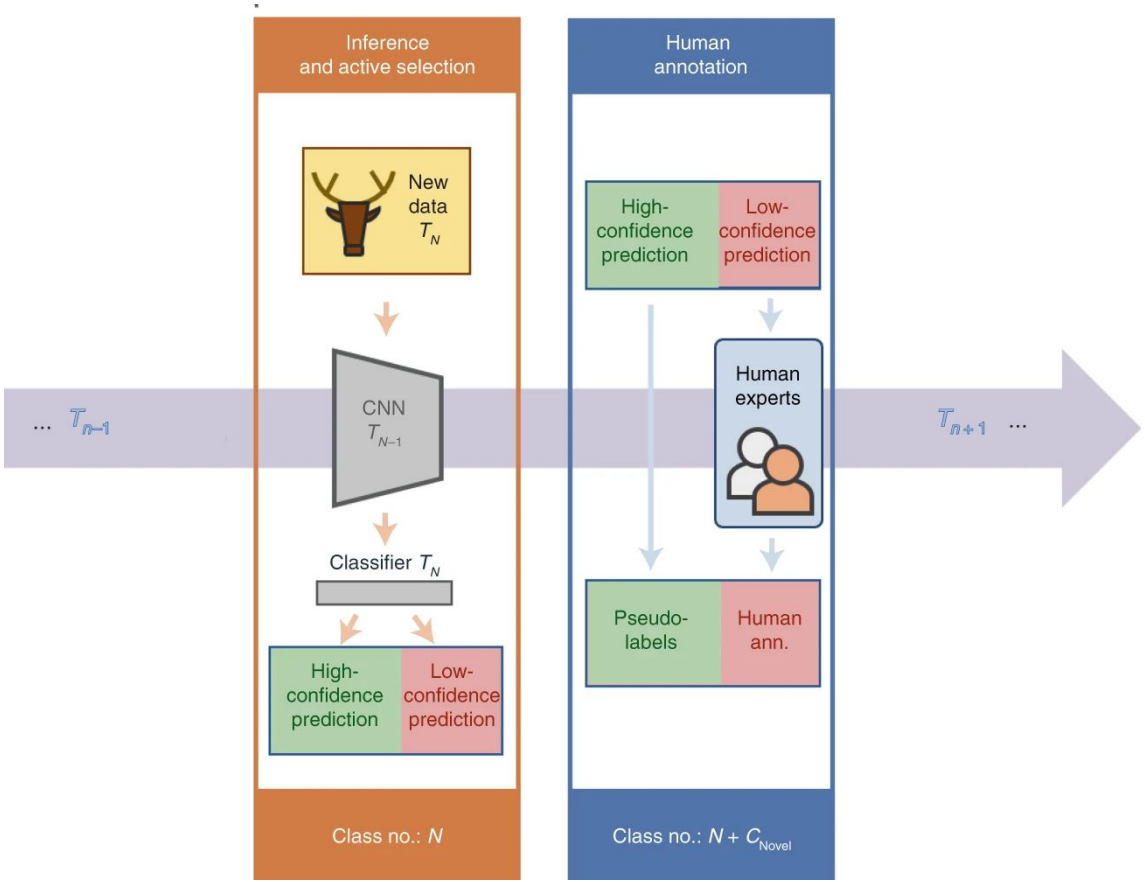
Selective prediction

"I don't know"

Selective prediction gives an abstain option, it doesn't force a decision but instead takes model confidence into consideration

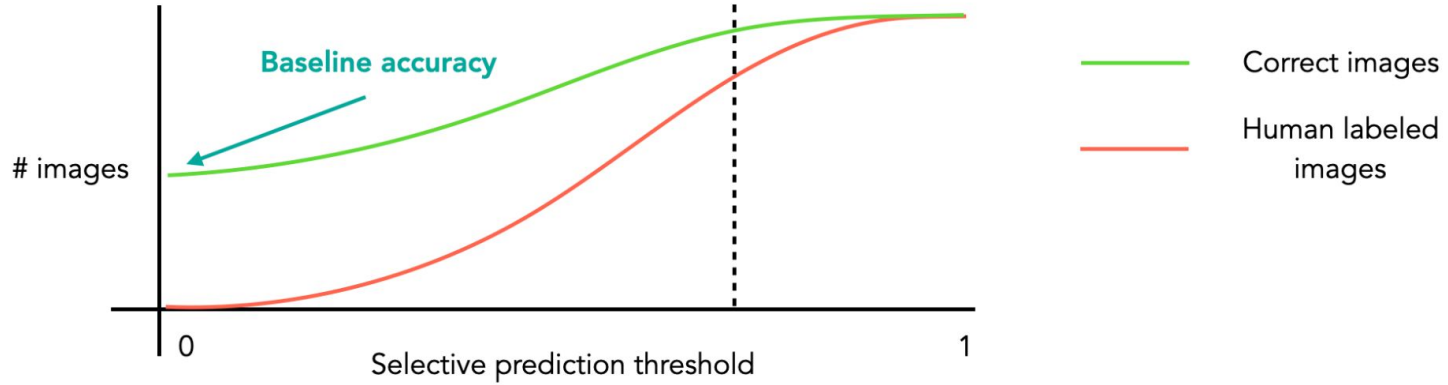
In practice, a human would then identify images that a model abstains

Selective prediction

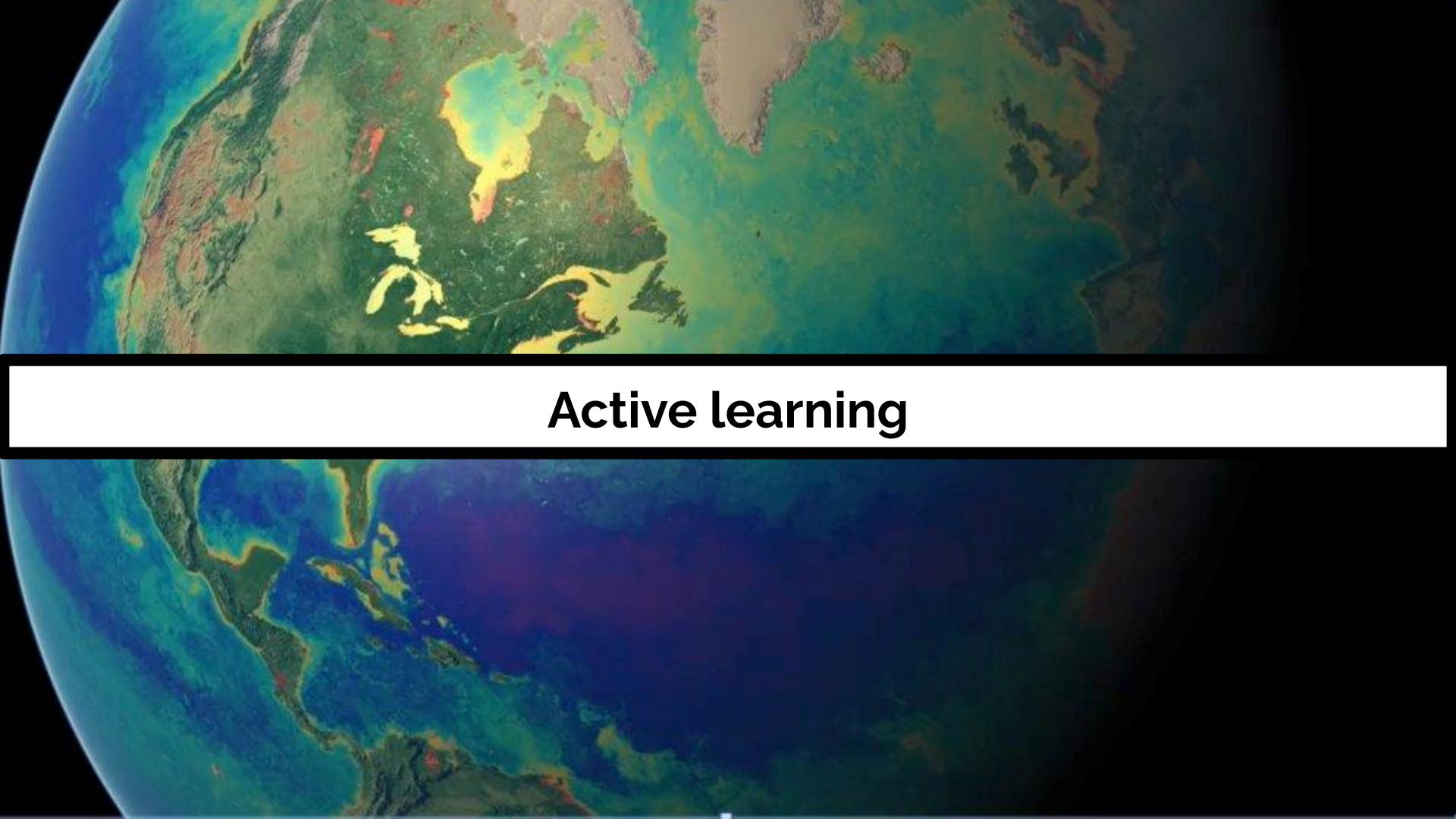


https://www.nature.com/articles/s42256-021-00393-0

Accuracy vs human effort in selective prediction



- Low thresholds mean the model is trusted more, thus less human effort needed to identify all the data but there is more possibility of error
- High thresholds mean the model is trusted less, thus humans ID more data but quality is easier to guarantee
- Threshold selection is an active area of research, calibrated models make this easier



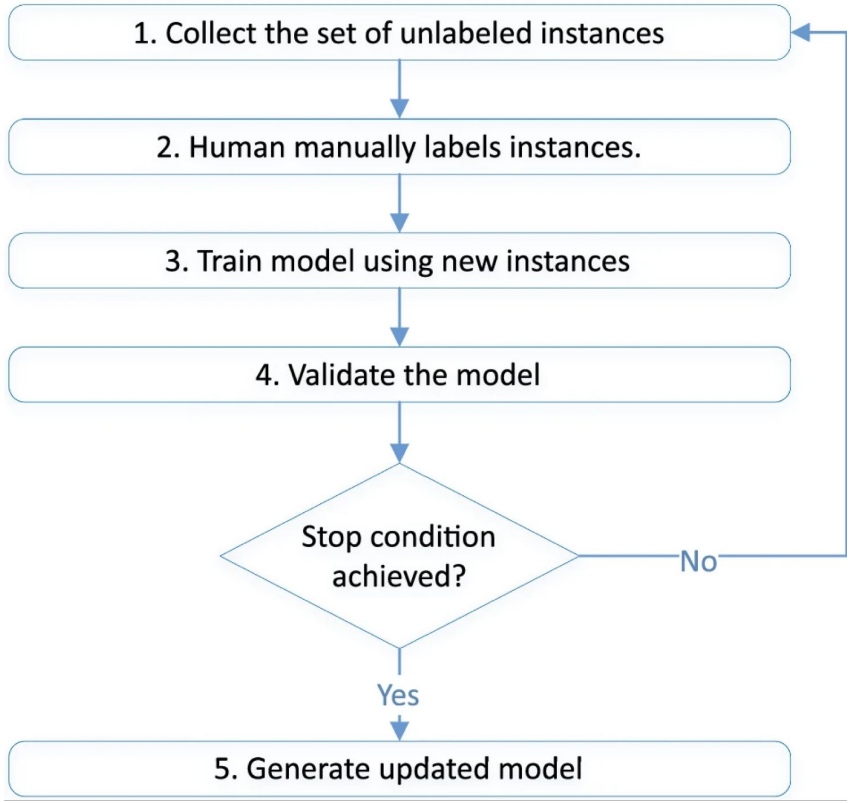
Active learning

Active learning

Learn to sample next data for human labeling automatically to optimize performance while minimizing human effort

Sampling criteria:

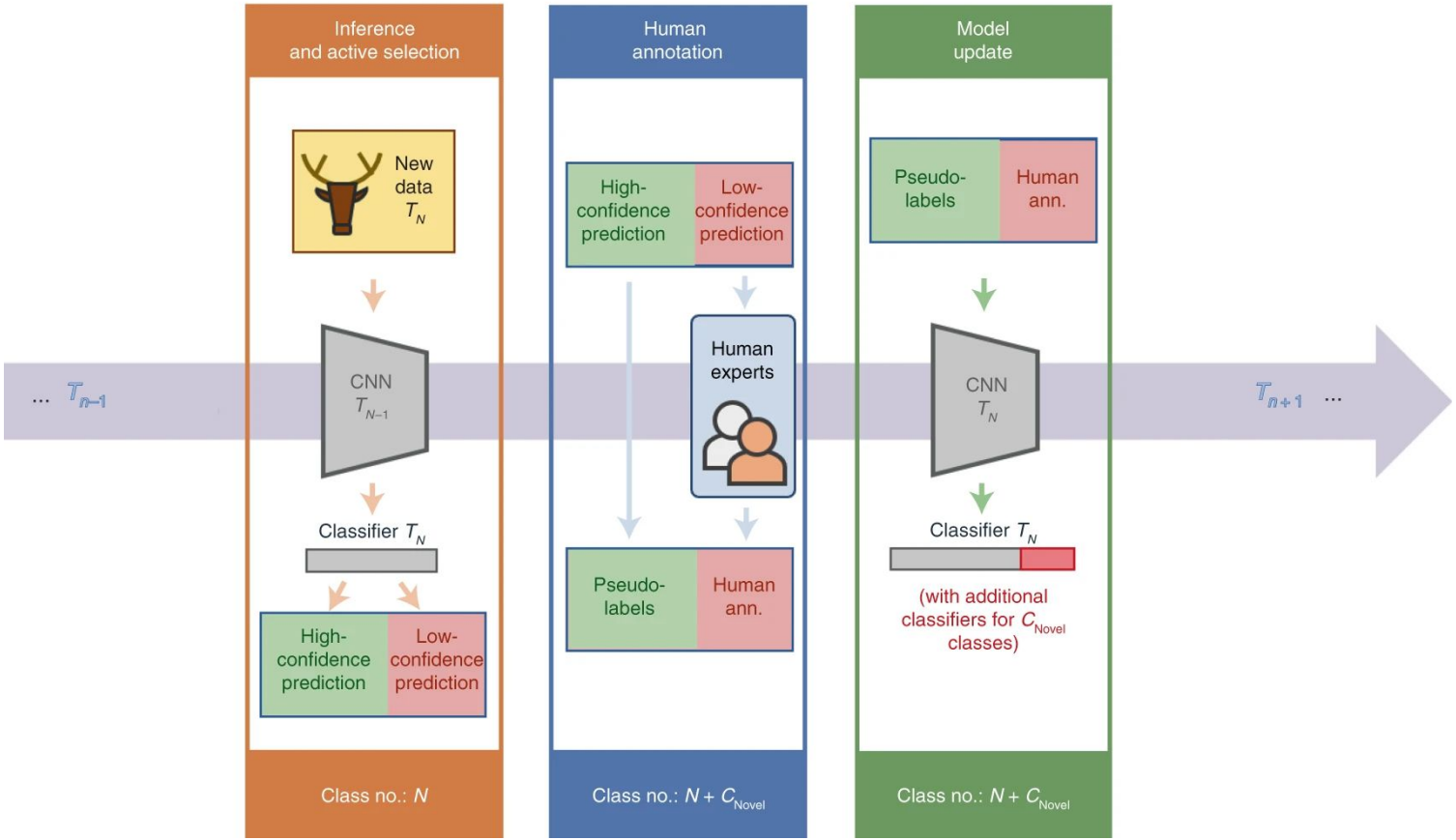
- Random
- Uncertainty (Exploit)
- Diversity (Explore)



Human-in-the-loop machine learning: a state of the art, Mosqueira-Rey et al., Artificial Intelligence Review 2022

Human-in-the-loop machine learning, Munro, Manning Publishing 2020

Active learning *via* selective prediction



https://www.nature.com/articles/s42256-021-00393-0

Active learning based on representations

One example:

- Use the MegaDetector to crop
- Cluster animals based on visual similarity in new cameras
- Humans ID examples from each cluster (active learning criteria)
- Gets same accuracy with **99.5% fewer labels**

A deep active learning system for species identification and counting in camera trap images, Norouzzadeh, Morris, Beery, et al., Methods in Ecology and Evolution 2021

Role of Human-AI Interaction in Selective Prediction

User would see one of the 4 conditions shown here:

Image 1

Image 40: AI model deferred.

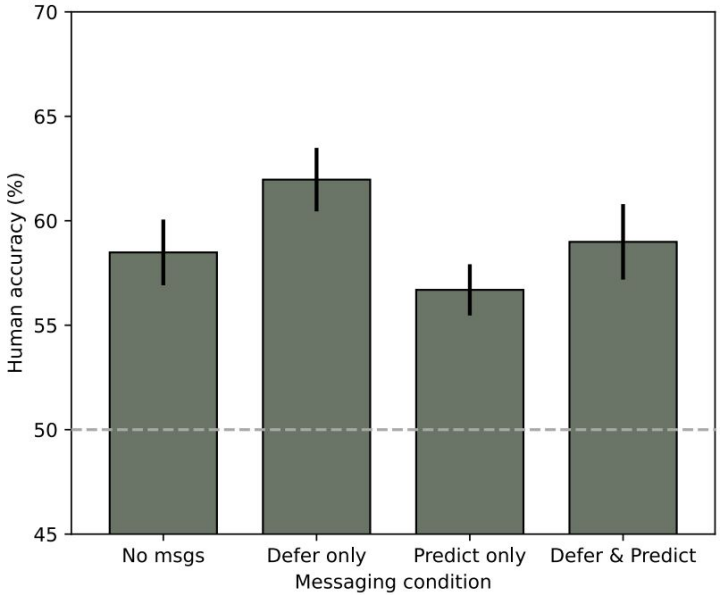
Image 6: AI model predicts no animal present.

Image 37: AI model deferred, but predicts no animal present.

1 2 3 4 5

Definitely no animal present ○ ○ ○ ○ ◉ Definitely animal present

Human accuracy decreases when model results are presented



https://ojs.aaai.org/index.php/AAAI/article/view/20465

Role of Human-AI Interaction in Selective Prediction

User would see one of the 4 conditions shown here:

Image 1

Image 40: AI model deferred.

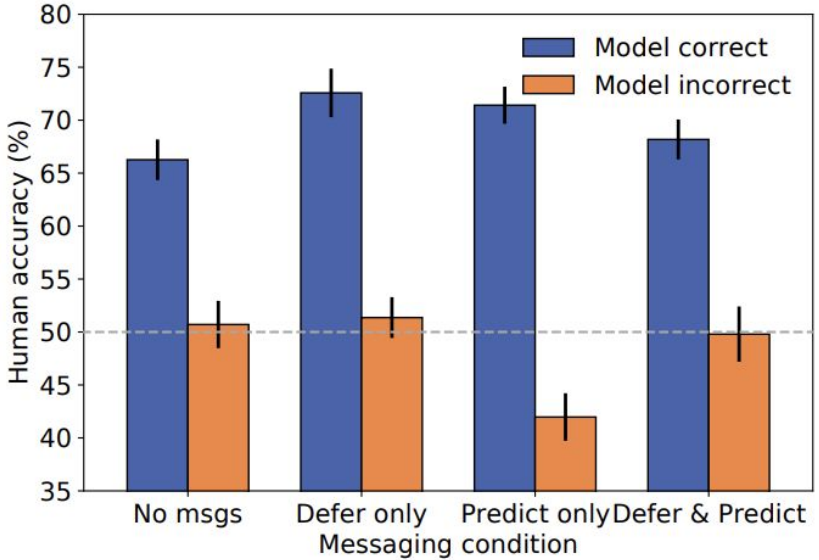
Image 6: AI model predicts no animal present.

Image 37: AI model deferred, but predicts no animal present.

1 2 3 4 5

Definitely no animal present ○ ○ ○ ○ ◉ Definitely animal present

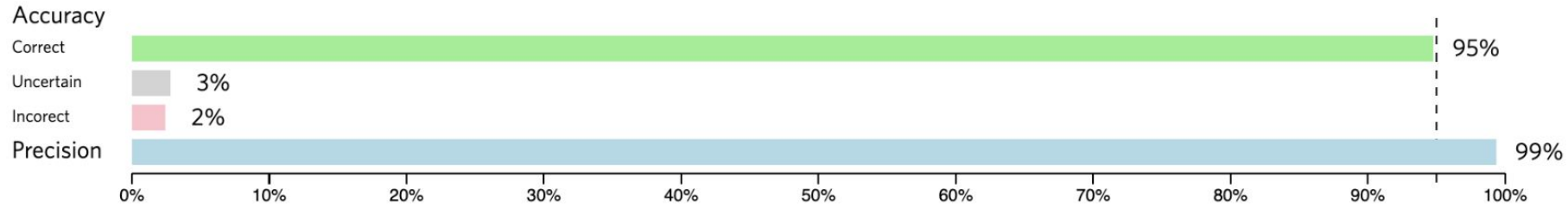
Human accuracy decreases when model results are presented



https://ojs.aaai.org/index.php/AAAI/article/view/20465

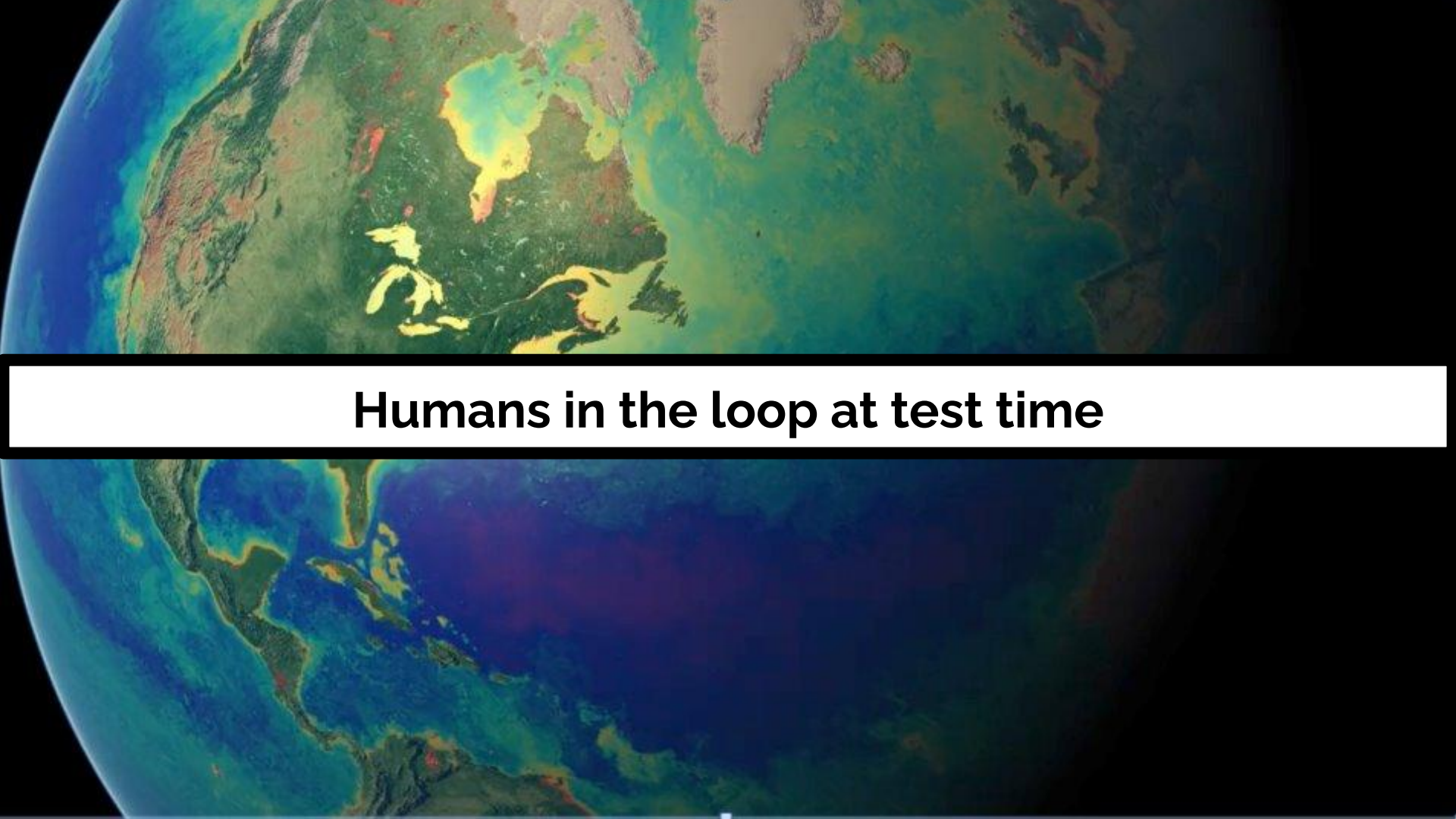
Confirmation bias

For the Research Grade subset, 95% were Correct, 3% were Uncertain and 2% were Incorrect. The average Precision was 99%.



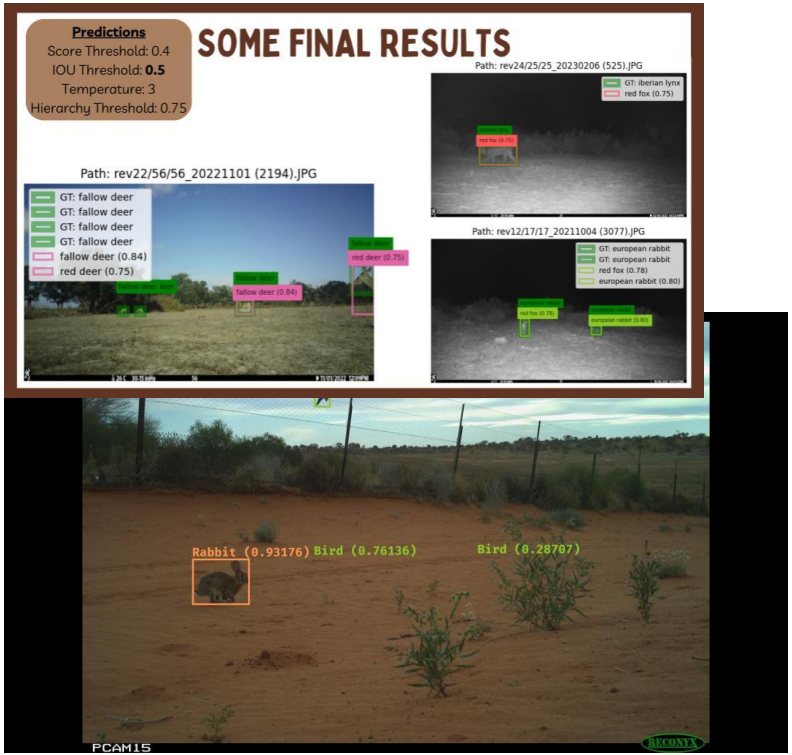
I had actually (not long ago) studied the question of subspecies of Apis mellifera in Africa and therefore knew, that bees from NE Namibia, SW Zambia and the Zambezi valley can't be identified to a subspecies, this area is a zone of introgression between A. m. scutellata and A. m. adansonii.
(My "wisdom" comes from a PHD thesis available for download here: Radloff, S. 1996. Multivariate analysis of selected honeybee populations in Africa https://commons.ru.ac.za/vital/access/manager/Repository/vital:5734/SOURCEPDF?site_name=Rhodes+University)
Obviously none of the other identifiers was aware of this. And this is when the confirmation bias sets in - you just agree without actually considering that you do not know how to identify this taxon.

Posted by traianbertau about 1 year ago

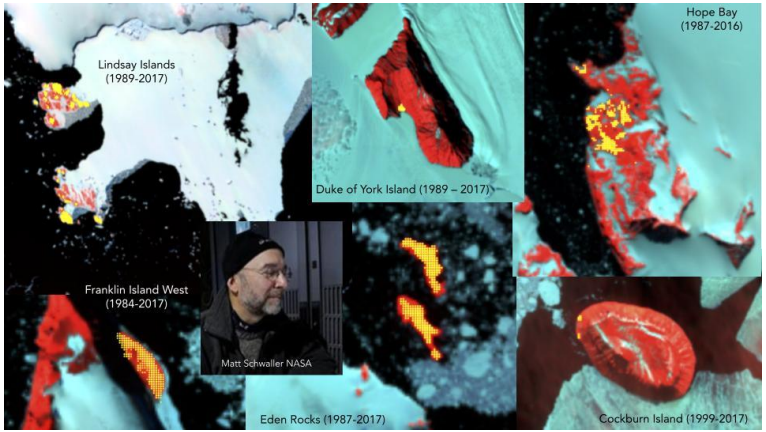


Humans in the loop at test time

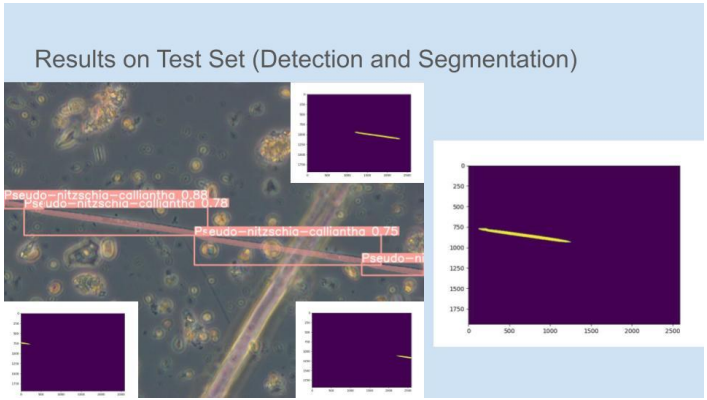
Typical model development / benchmarking



Animal occupancy/abundance, Marquez-Rodriguez, Tamm

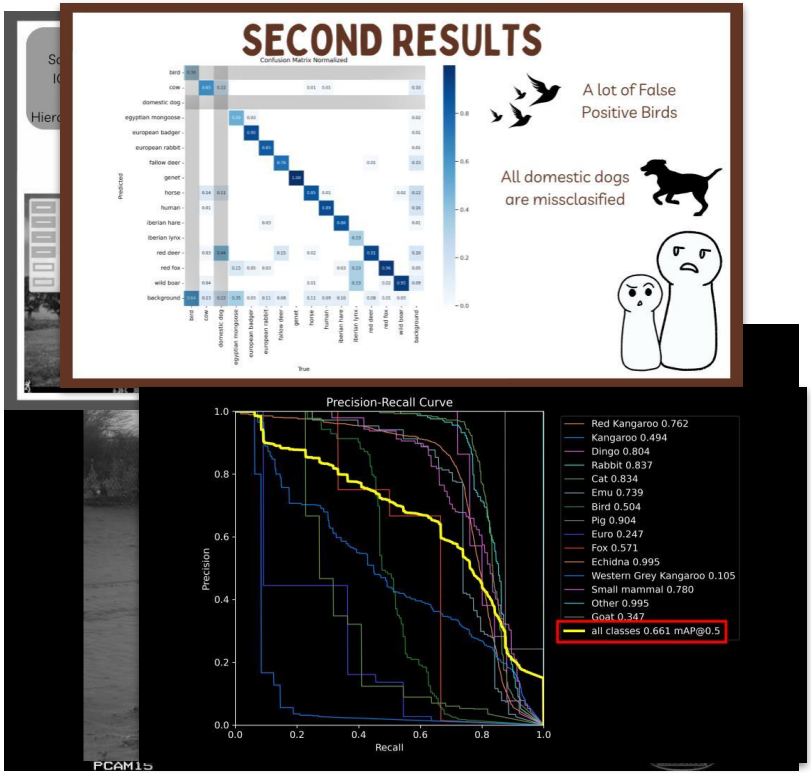


Guano surface area, Che-Castaldo

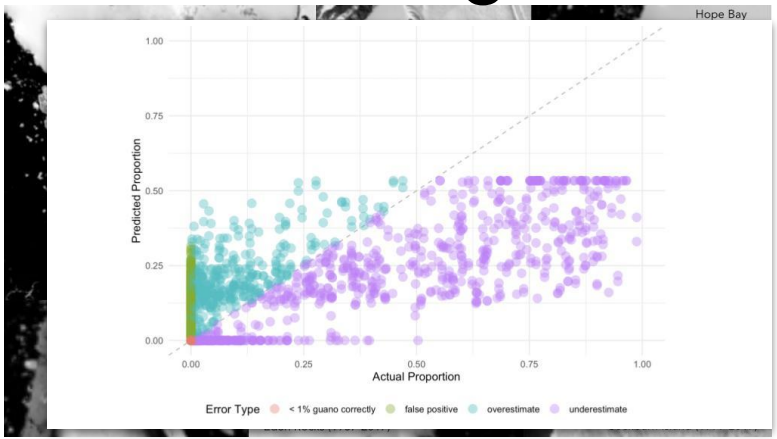


Phytoplankton biovolume, Marzidovšek

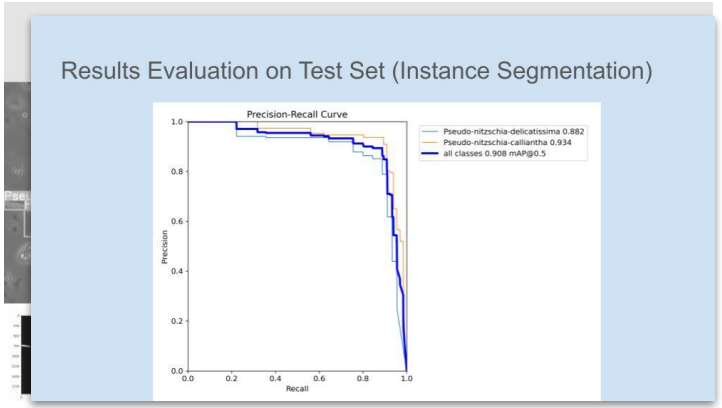
Typical model development / benchmarking



Animal occupancy/abundance, Marquez-Rodriguez, Tamm



Guano surface area, Che-Castaldo



Phytoplankton biovolume, Marzidovšek

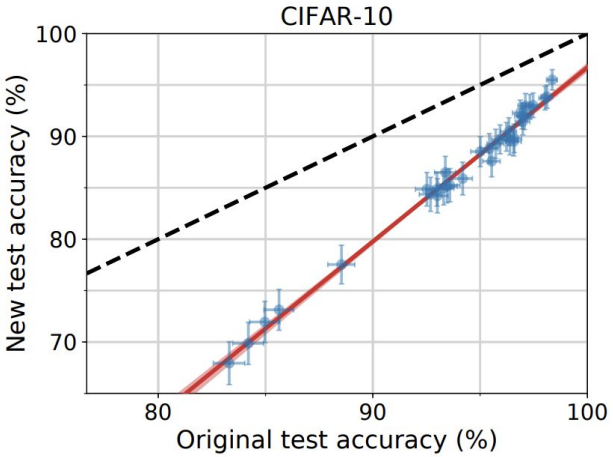
Performance on new data?

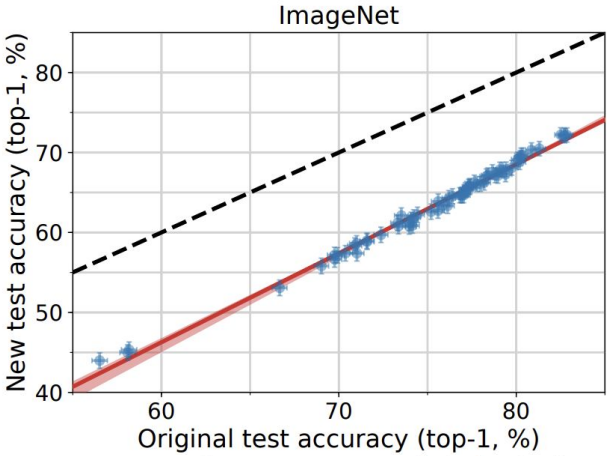
Do ImageNet Classifiers Generalize to ImageNet?

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.



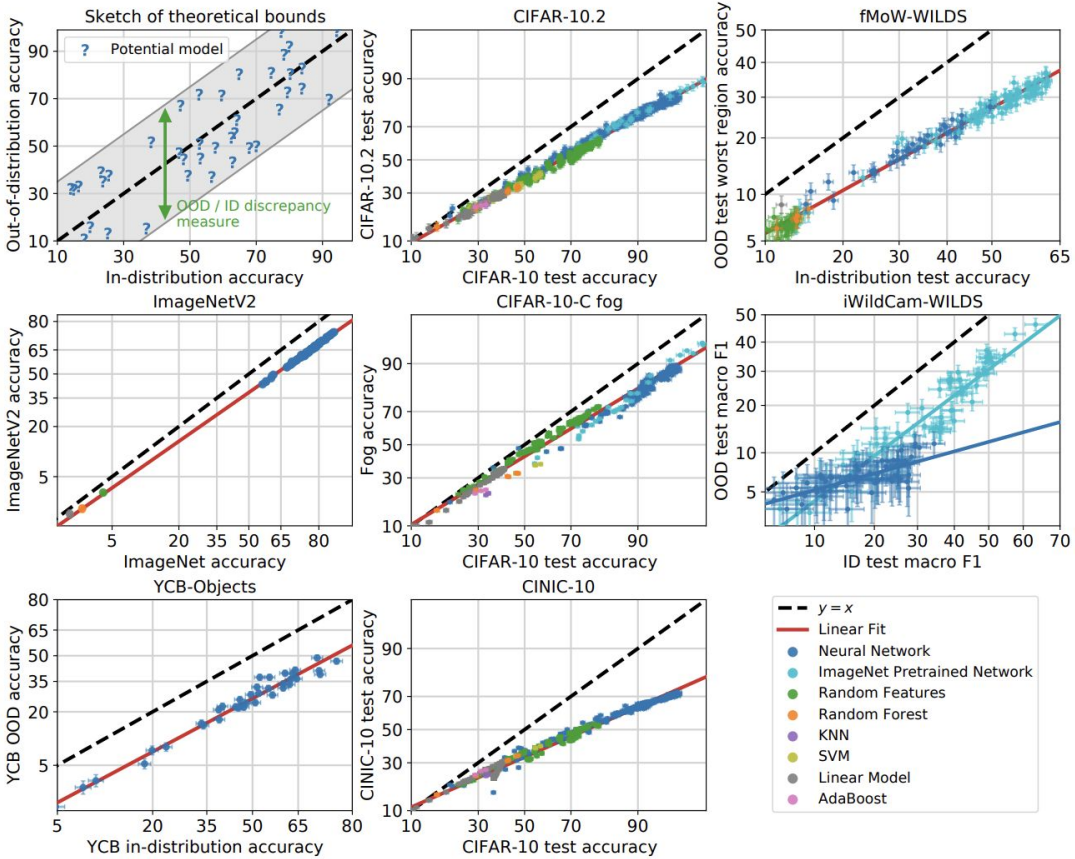


https://arxiv.org/pdf/1902.10811

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

We discussed **training techniques** to mitigate distribution shift

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

Test time

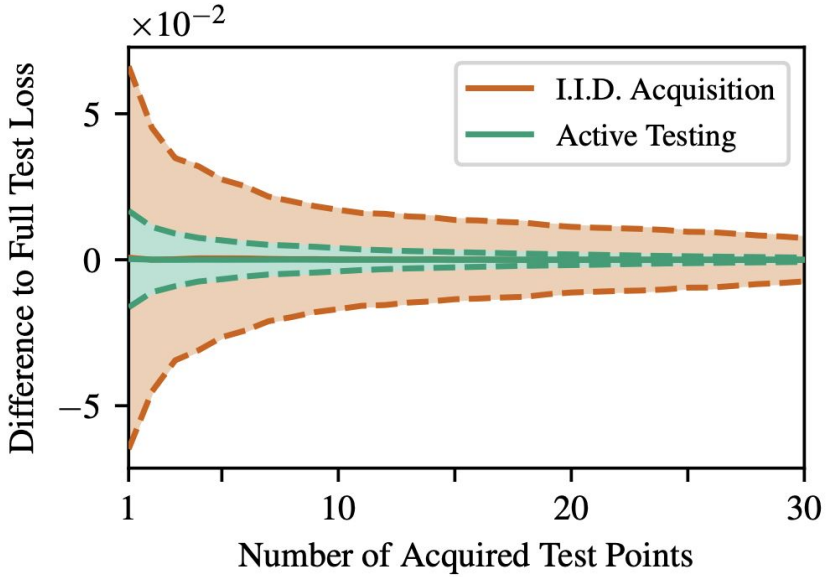
Keep the model fixed and focus on **post-training techniques** to interpret and utilize (imperfect) model predictions

- 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.



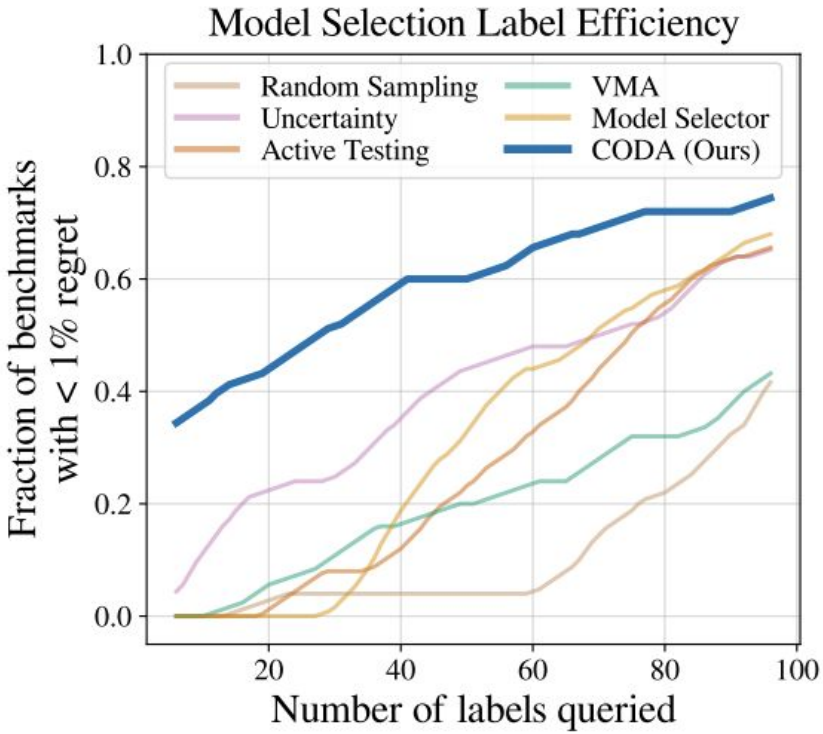
https://arxiv.org/pdf/2103.05331

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

Active Inference with ML+experts

What kinds of questions will you ask?

- Can we detect [very rare thing]?
- How many of [something interesting] are there?
- What are the spatial patterns of [species behavior]?
- How does [covariate x] affect [presence/abundance/behavior] of [species]?

Your use case will determine how you use the model.

Your model doesn’t have to be perfect.

1 min think-pair-share
(thinking optional)

What **ecological question** do you want to answer with the help of your model?

A useful paradigm - ML models as:

1. End-to-end flexible modeling approaches.
2. Tools for measurement/data collection.

	Cat
File 1	1
File 2	0
File 3	1

A useful paradigm - ML models as:

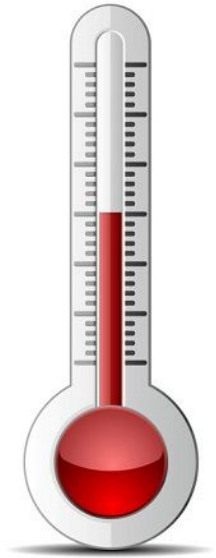
1. End-to-end flexible modeling approaches.

2. **Tools for measurement/data collection.**

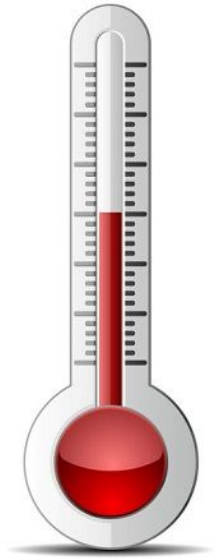
	Cat
File 1	1
File 2	0
File 3	1

I'm mostly going to talk about this

ML models as tools for measurement

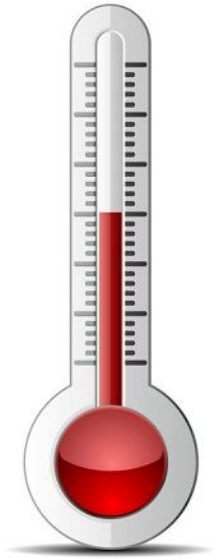


ML models as tools for measurement



	A thermometer
Noisy measurements	**±0.5 °C**
Biases	E.g. don’t put in direct sunlight
Operating range	**-20<T<140°C**

ML models as tools for measurement



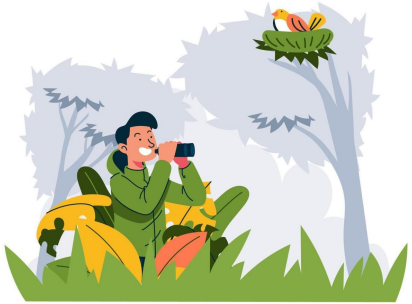
	A thermometer	**Your neural net**
Noisy measurements	**±0.5 °C**	Precision/Recall/False positive rate/ etc.
Biases	E.g. don’t put in direct sunlight	**Hard to foresee**
Operating range	**-20<T<140°C**	**In-distribution** (+ how much it generalizes)

Imperfect detection is not a new thing in Ecological data

Ecology, 83(8), 2002, pp. 2248–2255

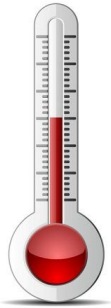
ESTIMATING SITE OCCUPANCY RATES WHEN DETECTION PROBABILITIES ARE LESS THAN ONE

DARRYL I. MACKENZIE,[1,5] JAMES D. NICHOLS,[2] GIDEON B. LACHMAN,[2,6] SAM DROEGE,[2] J. ANDREW ROYLE,[3] AND CATHERINE A. LANGTIMM[4]

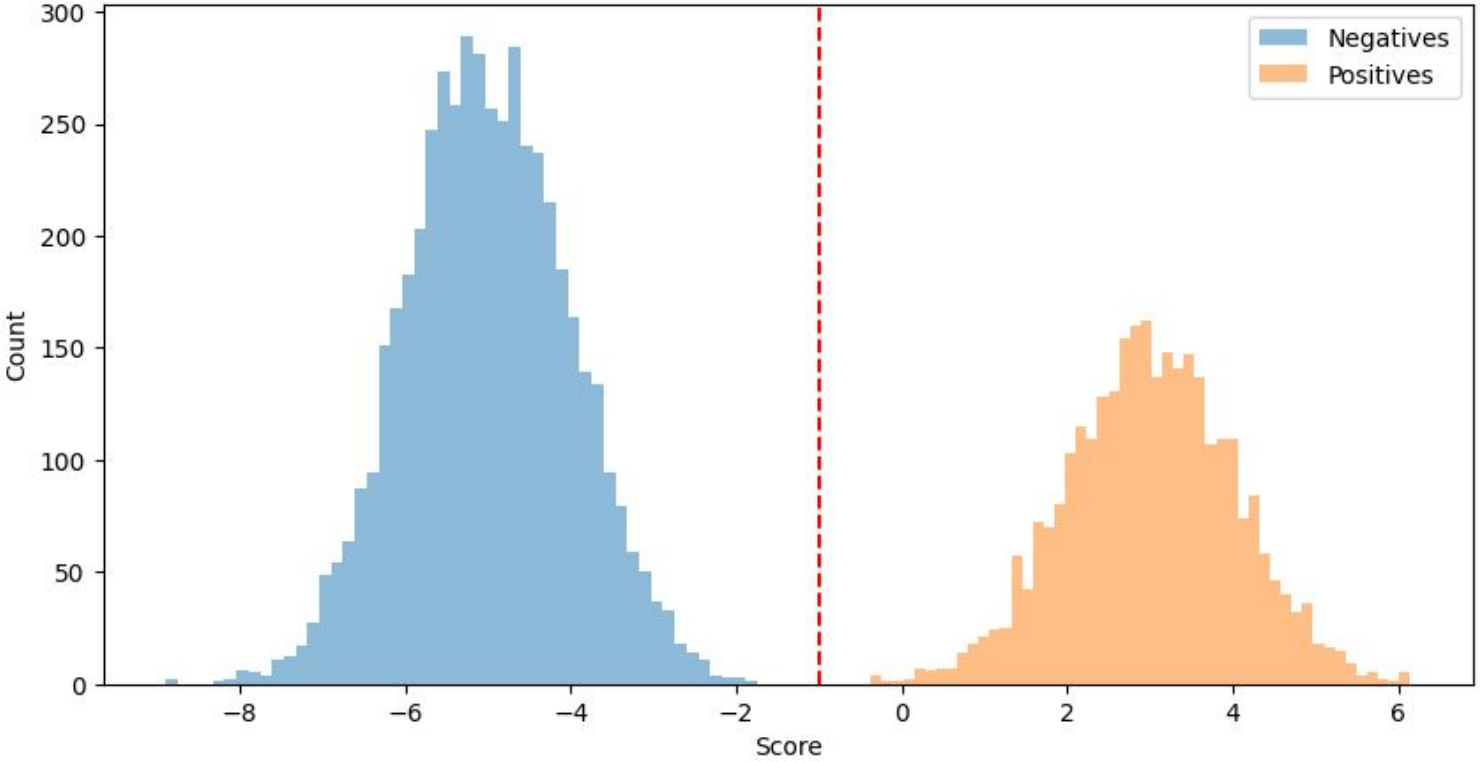


1 min think-pair-share

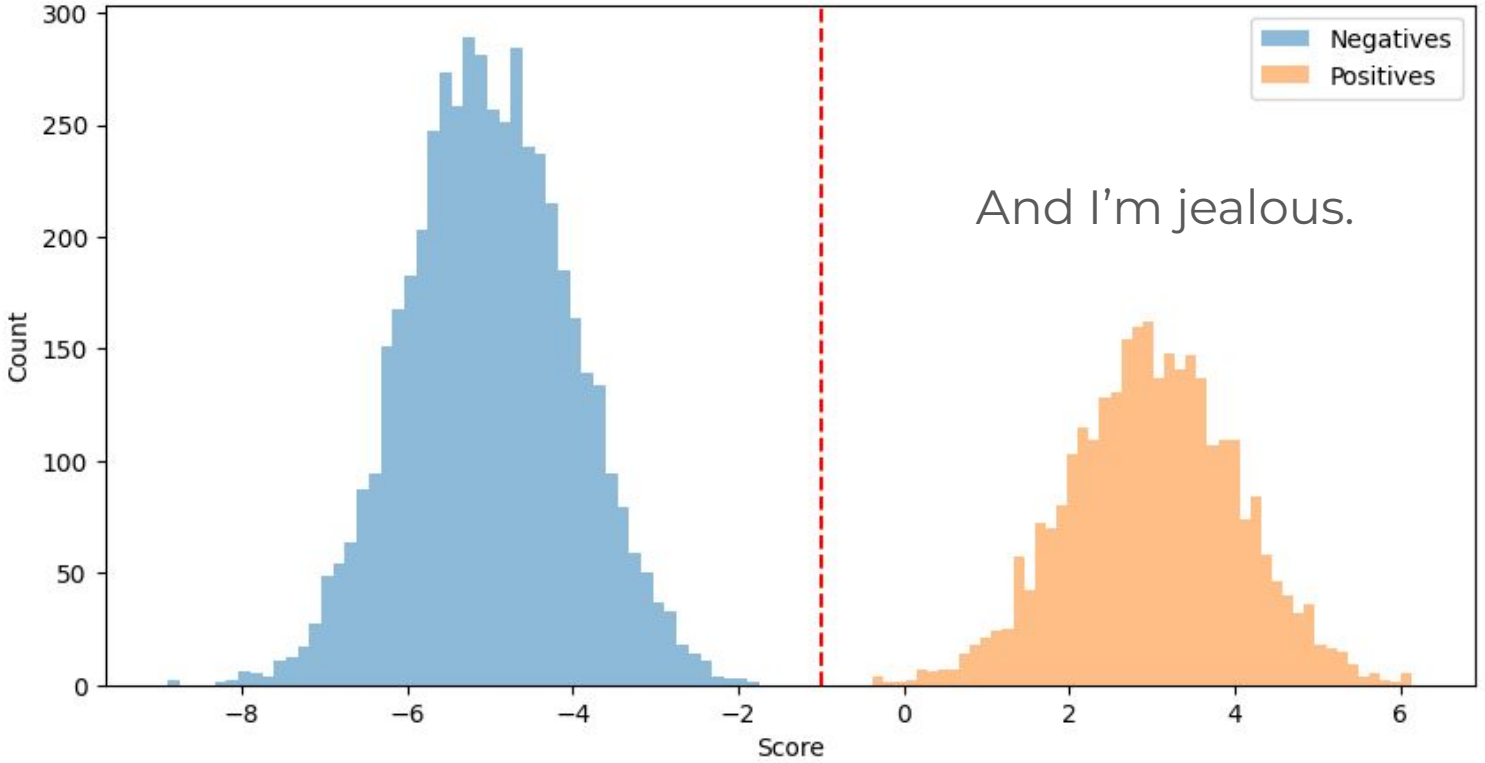
What is your model **measuring**, and can you **foresee any potential biases**?



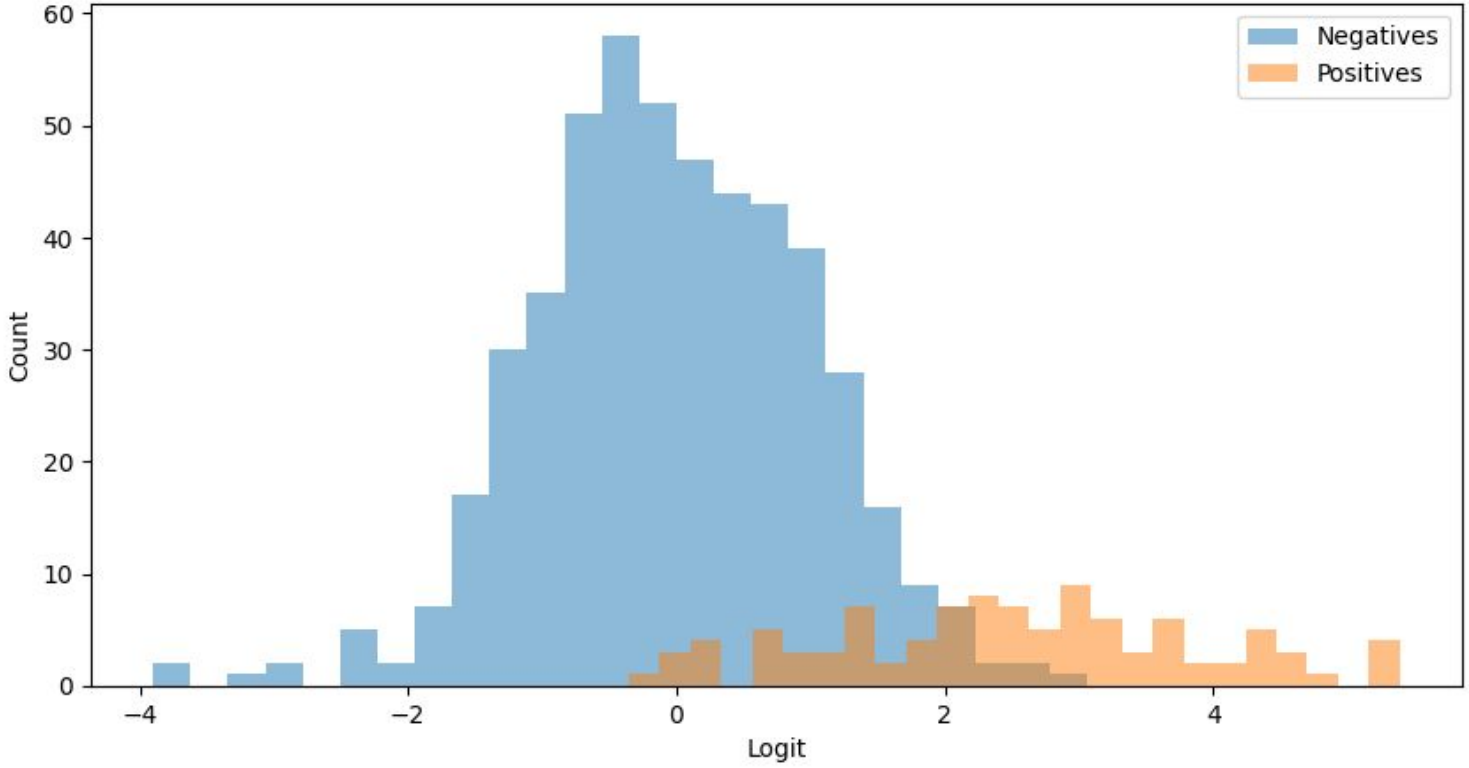
A binary classifier example


Negatives
Positives
Count
Score
300
250
200
150
100
50
0
−8
−6
−4
−2
0
2
4
6

For those with perfect classifiers… You don't need to think hard.

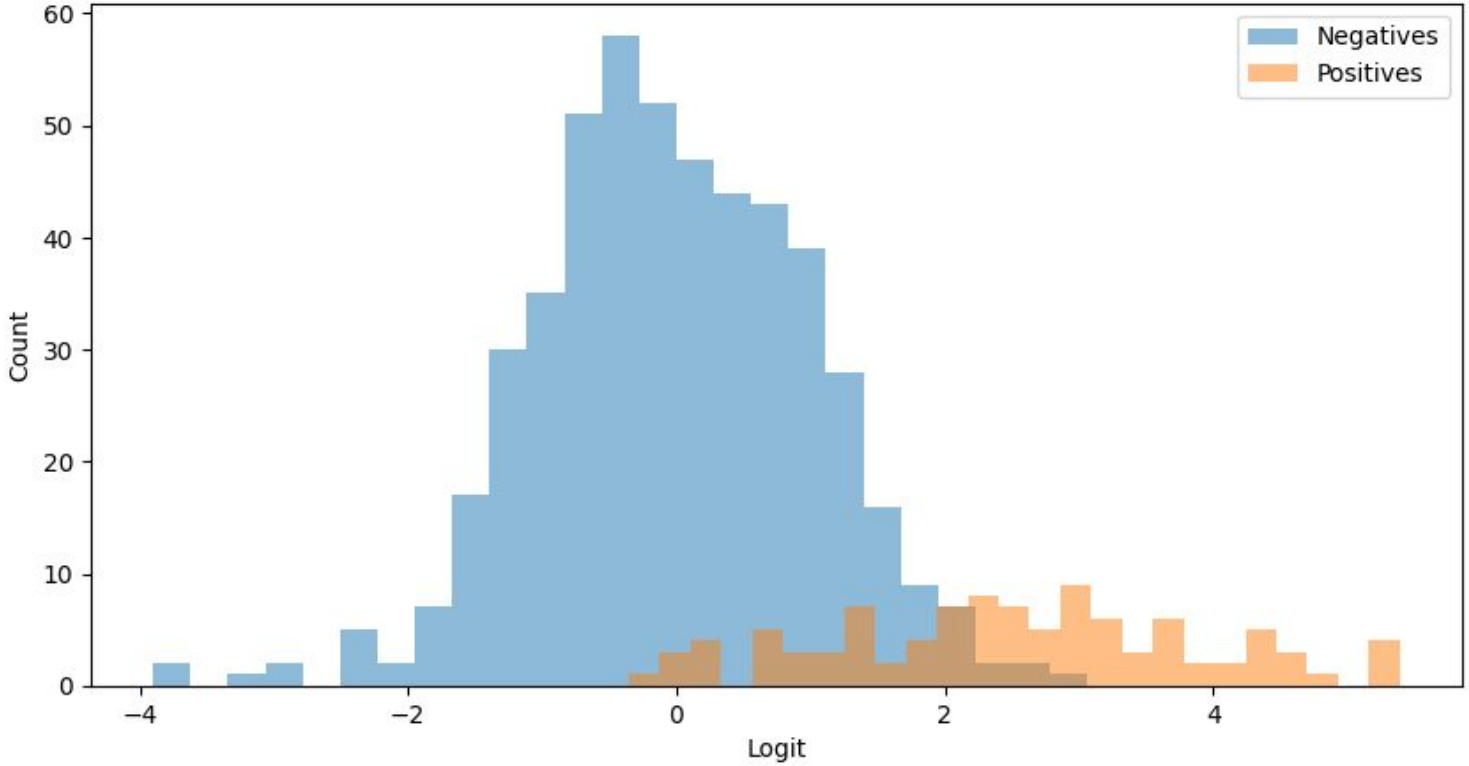


Typically, a classifier is imperfect.



Typically, a classifier is imperfect.

These predictions contain information - the challenge is how to use that information

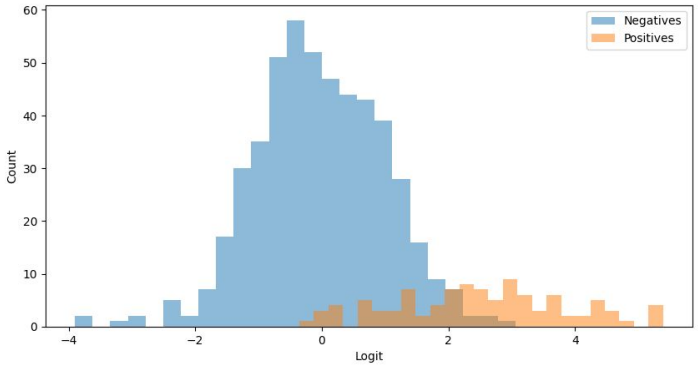


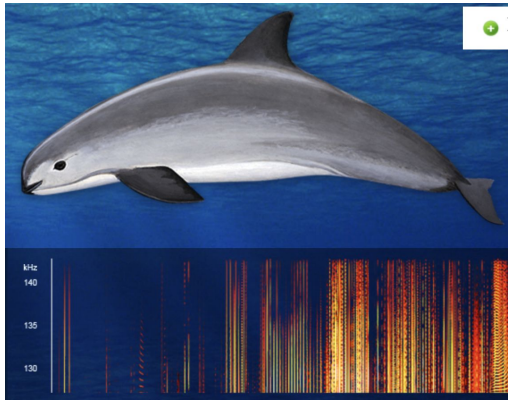
Use case 1: Detecting an endangered species

1,000,000 hours of audio recordings.

Limited human-expert listening effort available.

Presence-only records may be enough.

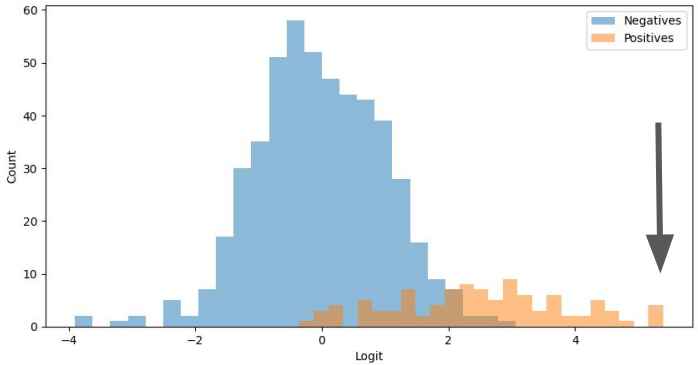




Use case 1: Detecting an endangered species

Presence-only records may be enough.

→ 'Top-down labeling': rank clips using the classifier and have a human verify them.

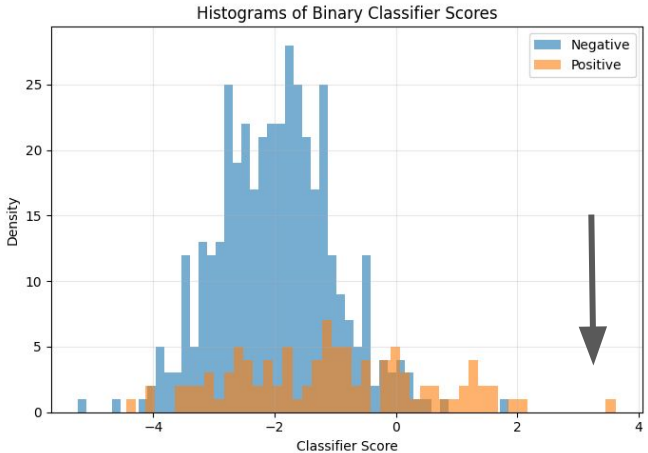


Use case 1: Detecting an endangered species

Presence-only records may be enough.

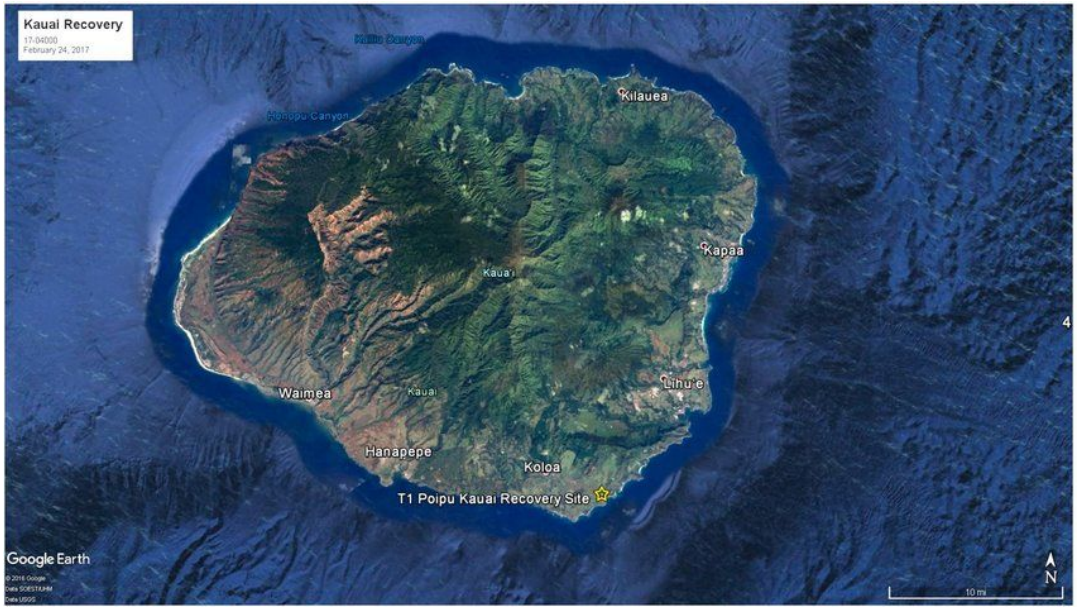
→ 'Top-down labeling': rank clips using the classifier and have a human verify them.

Even a 'bad' classifier can be useful.

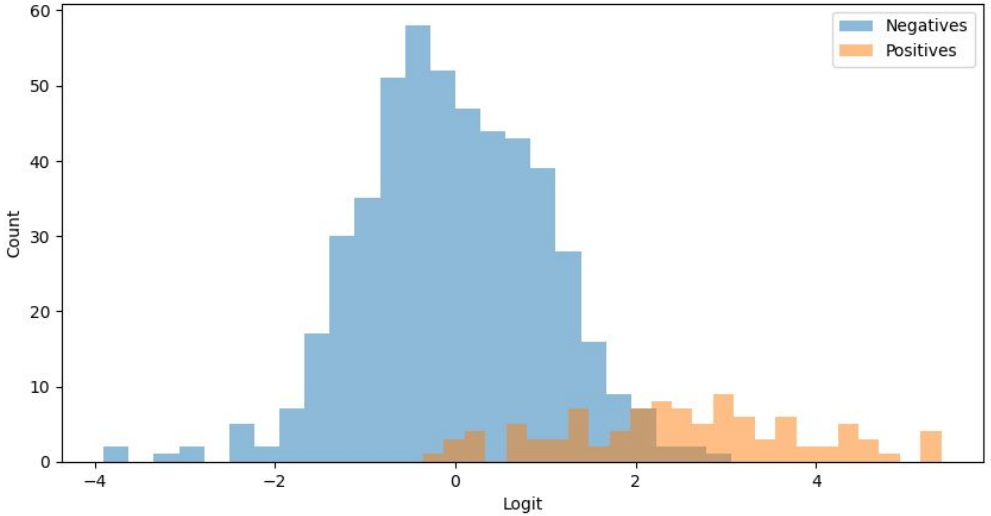


Use case 2: Counting detections at multiple sites

Kauai amakihi

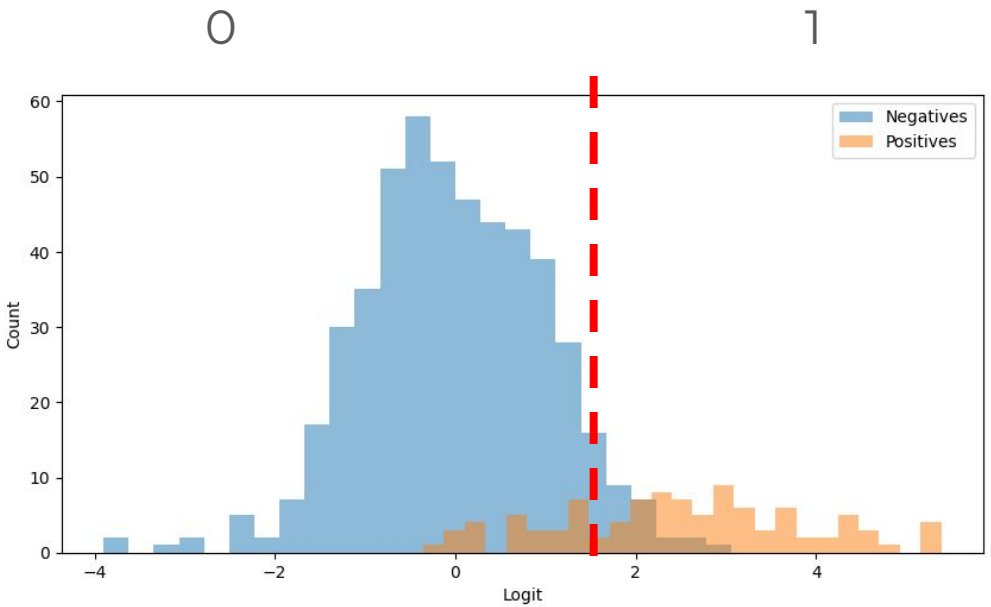


Use case 2: Counting detections at multiple sites



Use case 2: Counting detections at multiple sites

Threshold and count.

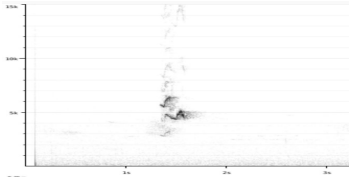


Use case 2: Counting detections at multiple sites

On your test set you achieved 95% precision, 60% recall at this threshold.

	Number of detections (out of 10,000 clips)
Site 1	500
Site 2	200
Site 3	50
Site 4	2

Kauai amakihi



Conclusion:

All sites **occupied**.

Site 1 **>** Site 2 **>** Site 3 **>** Site 4

	Number of detections (out of 10,000 clips)
Site 1	500
Site 2	200
Site 3	50
Site 4	2

Kauai amakihi

Conclusion:

All sites **occupied**.

Site 1 **>** Site 2 **>** Site 3 **>** Site 4

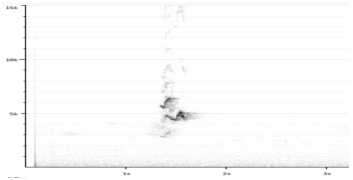
	Number of detections (out of 10,000 clips)
Kauai site 1	500
Costa Rica 2	200
Maui site 3	50
Kauai site 4	2

Kauai amakihi

Distribution shift means you can't trust the performance as measured on the test set

	Number of detections (out of 10,000 clips)
Kauai site 1	500
Costa Rica 2	200
Maui site 3	50
Kauai site 4	2

Kauai amakihi

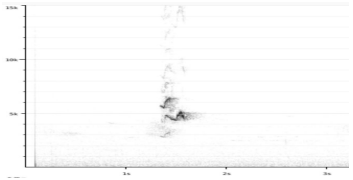


Changes between sites also constitute small **distribution shifts**

(even if the only thing that changes is the number of positives)

	Number of detections (out of 10,000 clips)
Kauai site 1	500
Costa Rica 2	200
Maui site 3	50
Kauai site 4	2

Kauai amakihi



To get trustworthy and interpretable numbers, you need to calibrate.

Uncertainty quantification and calibration

{ 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

Calibration of thresholded detections

1. For each subgroup of the data (e.g. site)label some random data.

	Predicted 1	Predicted 0
True 1	16	4
True 0	6	24

Calibration of thresholded detections

1. For each subgroup of the data (e.g. site) label some random data.

	Predicted 1	Predicted 0
True 1	16	4
True 0	6	24

2. Estimate:
   p(label=1 | prediction=1) *precision*
   p(label=1 | prediction=0) *false negative rate*

Calibration of thresholded detections

1. For each subgroup of the data (e.g. site) label some random data.

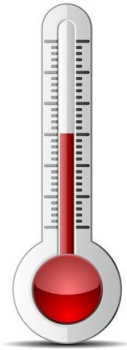
	Predicted 1	Predicted 0
True 1	16	4
True 0	6	24

2. Estimate:
   p(label=1 | prediction=1) *precision*
   p(label=1 | prediction=0) *false negative rate*
3. Use these to correct your estimates.

Calibration will vary by site

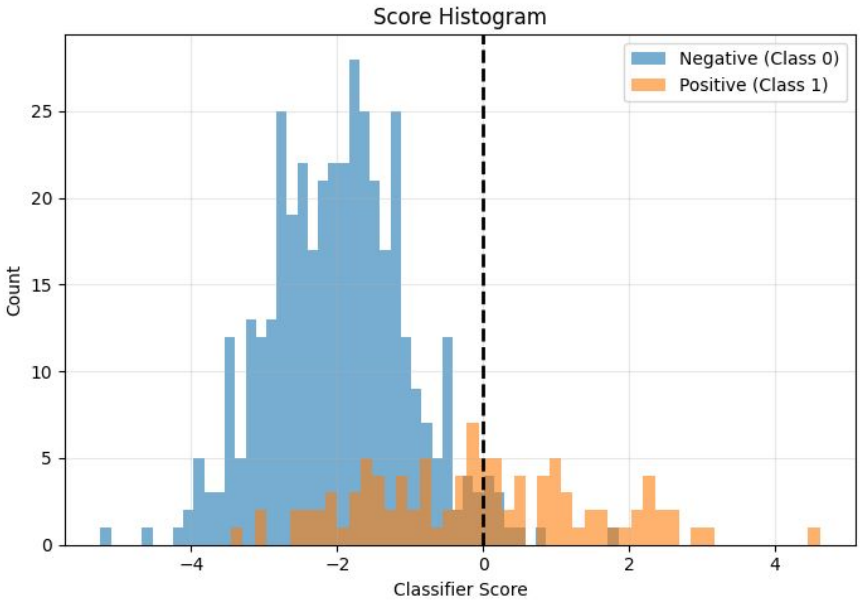
To get **statistically valid site-level estimates**, you need a **calibration set from each site**.

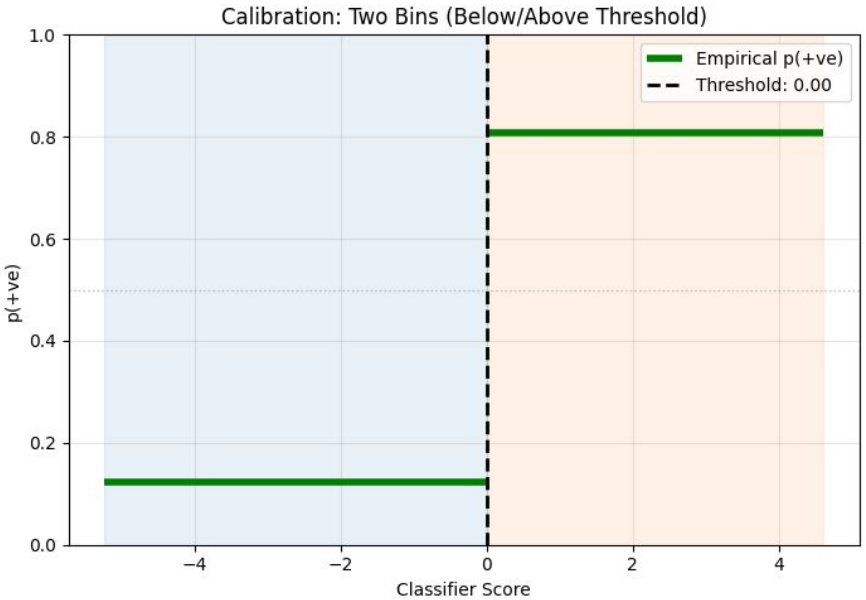
	Raw Detections	95%CI
Kauai site 1	500	400-550
Costa Rica	200	0-10
Maui site 1	50	0-10
Kauai site 2	2	5-25



Calibrating

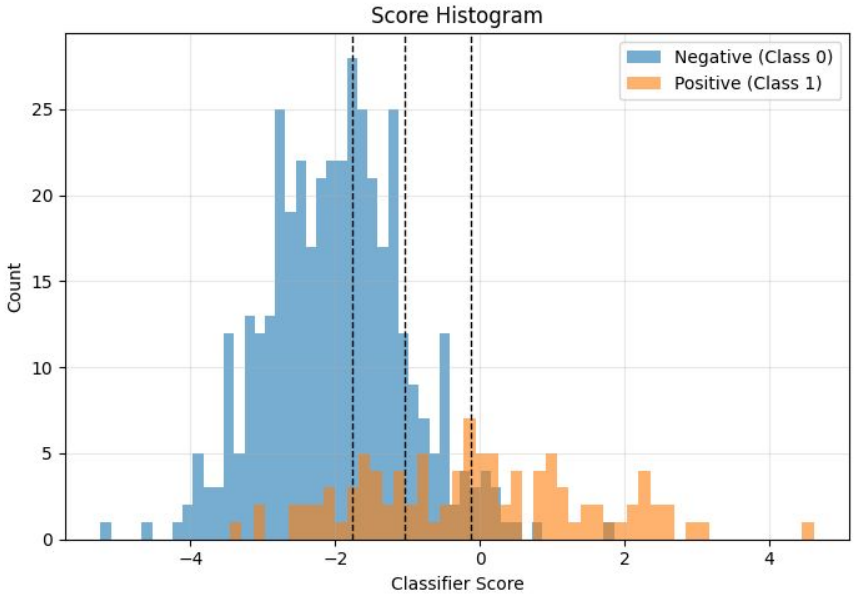
Isn’t thresholding a little arbitrary?

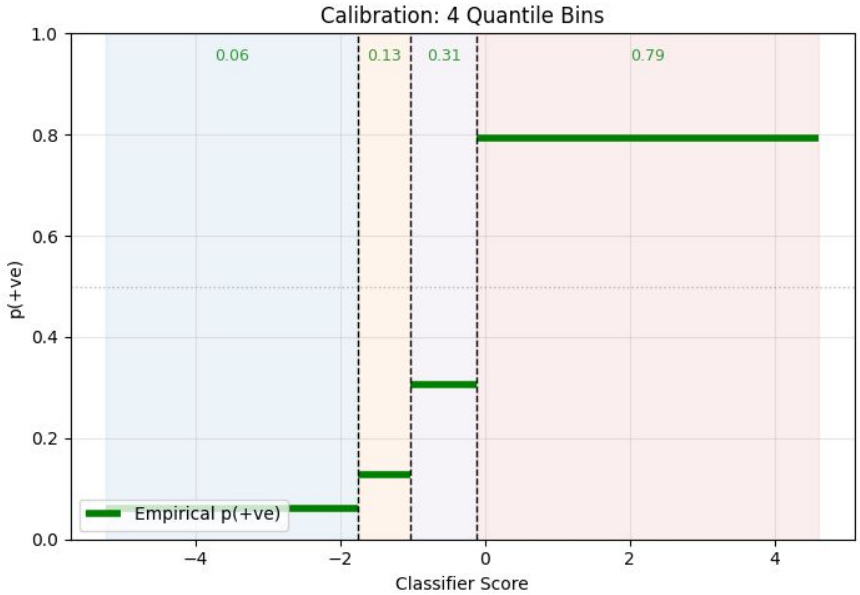




Calibration by histogram binning

Divide scores into multiple bins, label data from each bin.
Can be particularly helpful if the positive class is rare.

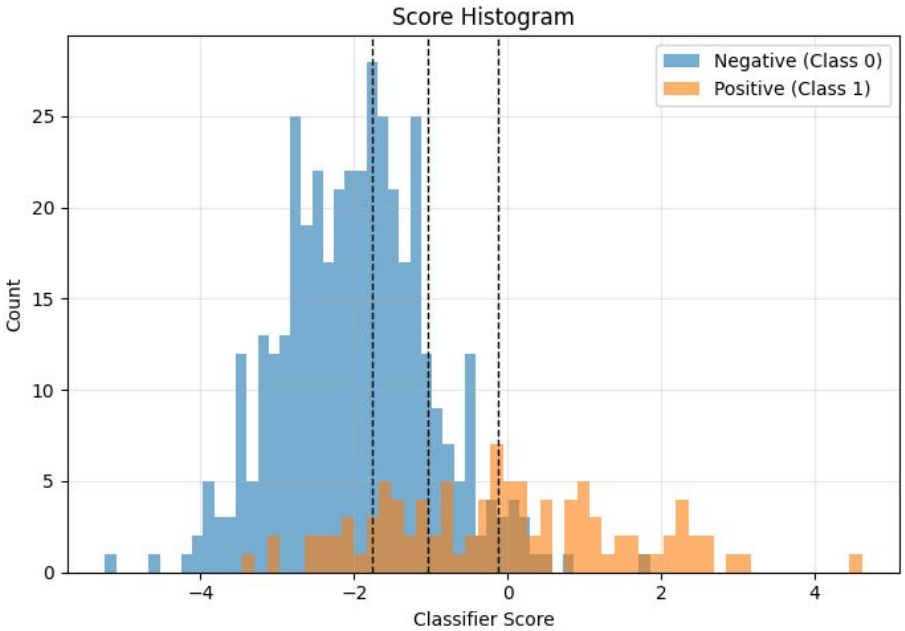


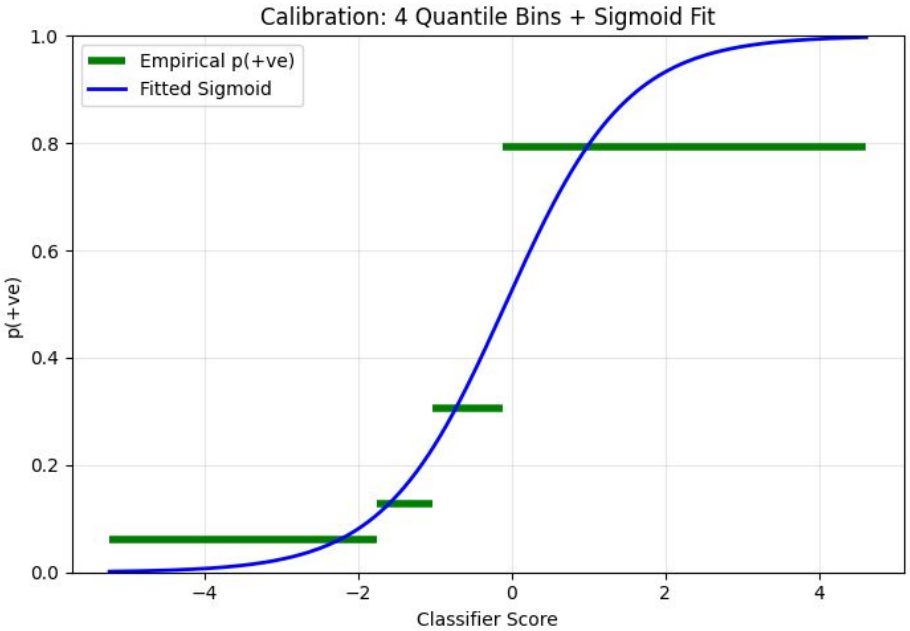


Navine, Denton, Weldy, Hart 2024

Calibration by Platt scaling

Platt scaling fits a logistic regression to learn a mapping from classifier score to p(+ve)

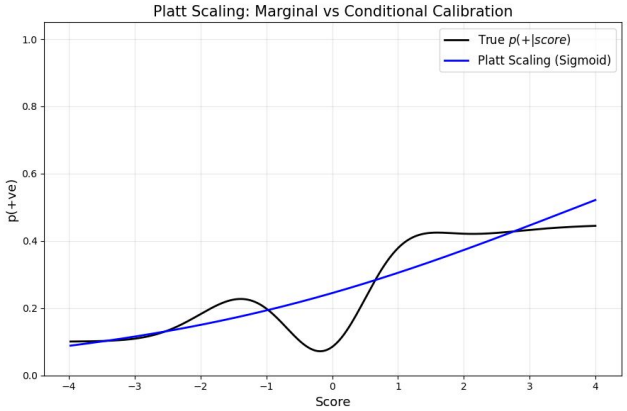




Calibration has a lot of subtlety

Platt scaling gives **marginal calibration**. (over ALL data points, on average the probability is right)

It doesn't give **conditional guarantees**. e.g. it does not guarantee that of all the data points with p(+ve) = 0.5, 50% are positives.



Classifier Score	Calibrated P(+ve)
1	0.80
-2	0.01
-1.5	0.10
…	…

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

Common goal: Statistical Inference

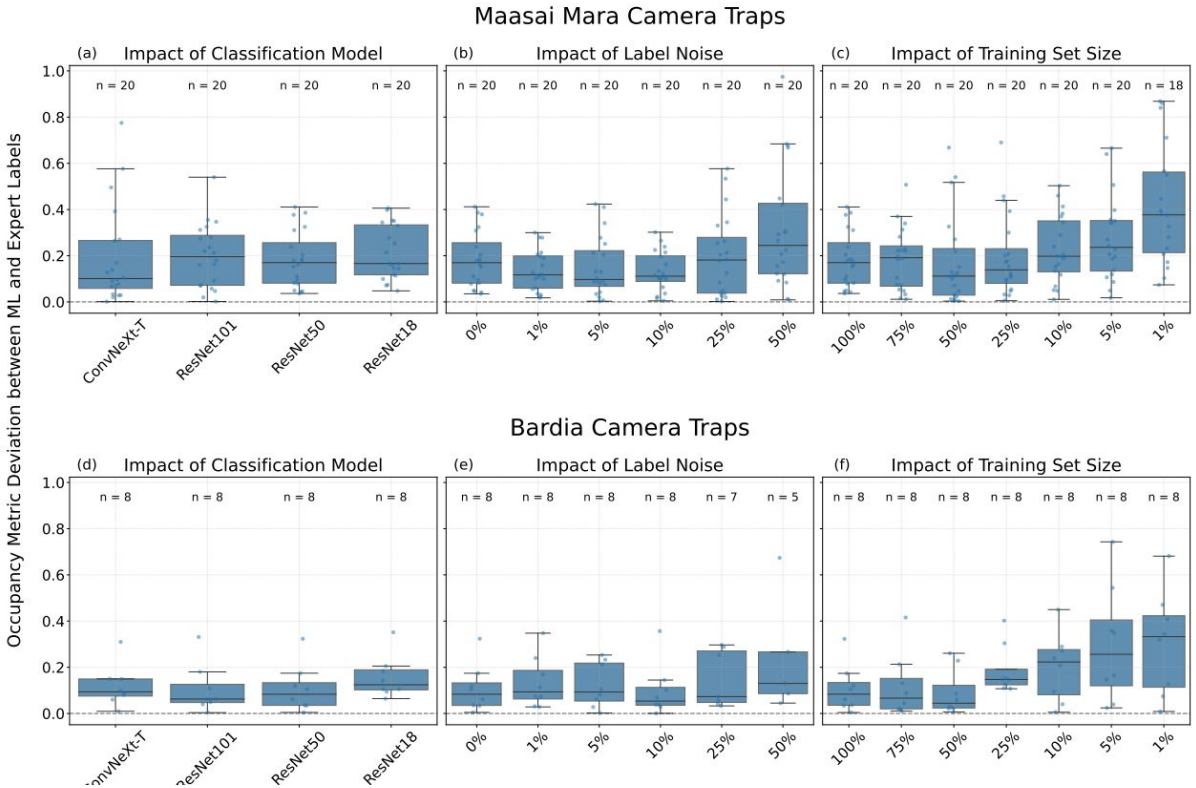
Use some data to infer some characteristics of the larger population:

- 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

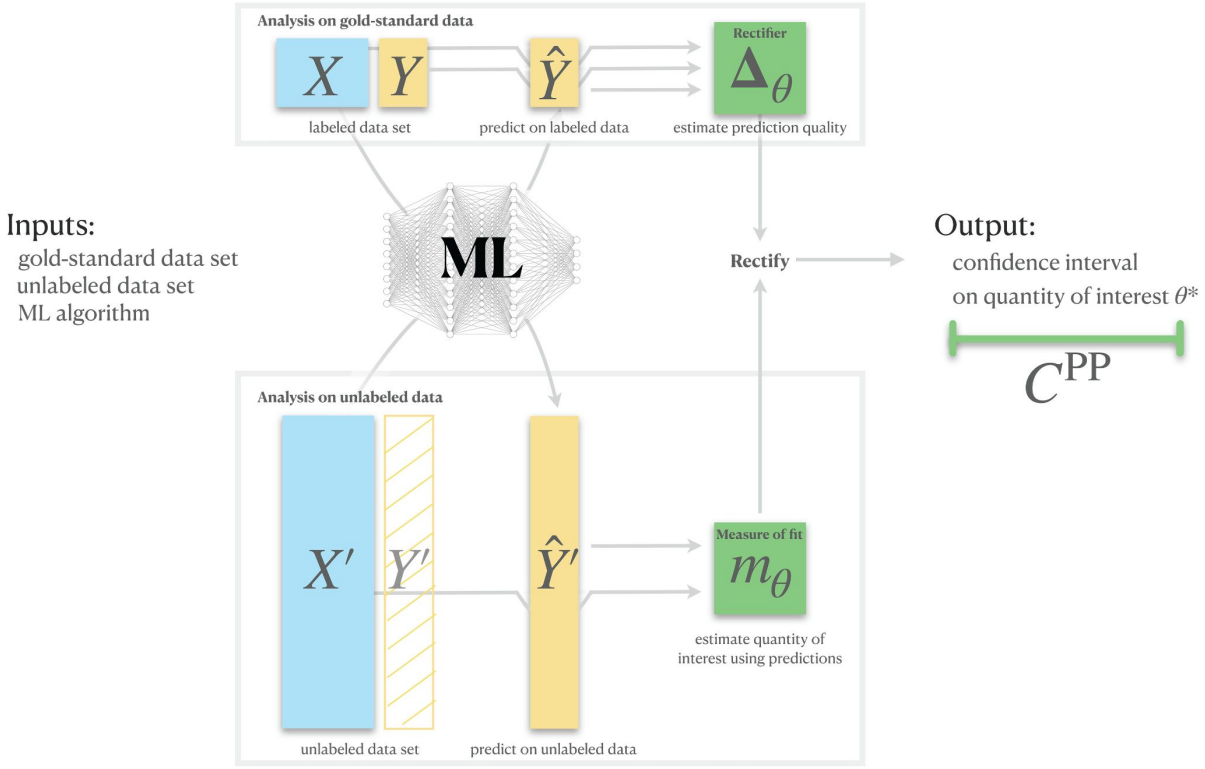


https://arxiv.org/pdf/2408.14348

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

Use case 3:
Doing Ecological Inference

Ecological process

Detection process

$$X \rightarrow Y$$

$$\theta \rightarrow Y \rightarrow \hat{Y}$$

Simple ecological inference: Turtle sex ratios

- At a healthy beach, 50% of turtles are born male.
- If the sand is too hot, 100% of turtles are born male.
- You have a camera trap and a baby turtle sex ID model.
- You want to know if your beach is healthy.

Inference - ignoring measurement error

H_0  $p = 0.5$

H_1  $p = 1$

Inference - ignoring measurement error

H_0  p = 0.5

H_1  p = 1

Data:

Boy, Boy, Boy, Boy, Boy, Boy

Inference - ignoring measurement error

H_0  p = 0.5

H_1  p = 1

Data:

Boy, Boy, Boy, Boy, Boy, Boy

6 bits of evidence in favor of H_1

$$P(H_1 \mid \text{data}) \approx 98.5\%$$

Inference - accounting for measurement error

H_0  p = 0.5

H_1  p = 1

Data:

Boy, Boy, Boy, Boy, Boy, Boy

Your model:
80% accurate

Inference - accounting for measurement error

H_0  p = 0.5

H_1  p = 1

Your model:
80% accurate

Data:

Boy, Boy, Boy, Boy, Boy, Boy

~4 bits of evidence in favor of H_1

$$P(H_1 \mid \text{Data}) \approx 94.4\%$$

Significant?

B, B, B, B, B, B

B, B, B, B, B, B

Inference - you can be infinitely wrong if you don't account for measurement error

H_0  p = 0.5

H_1  p = 1

Data:

Boy, Boy, Boy, **Girl**, Boy, Boy

Treating data as gold-standard

$P(H_1 \mid Data) = 0\ \%$ (∞ bits of evidence against H_1)

Accounting for your imperfect model

$P(H_1 \mid Data) = 81\ \%$ (2.07 bits of evidence in support of H_1)

Treating
ML predictions
as gold
standard data

Accounting
for uncertainty
in model
predictions.

Using ML predictions gets you way more power

x bits per observation

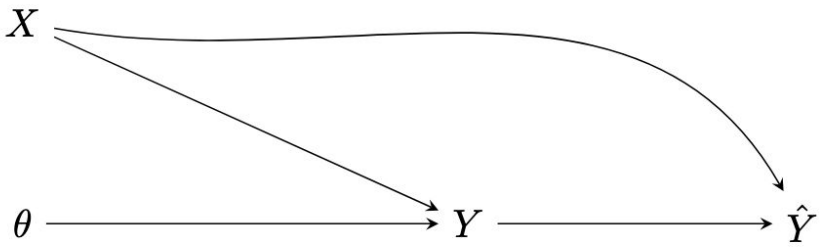
n observations

<x bits per observation

>>>>>>n observations

Nightmare fuel

When covariates of interest are confounded with model performance.



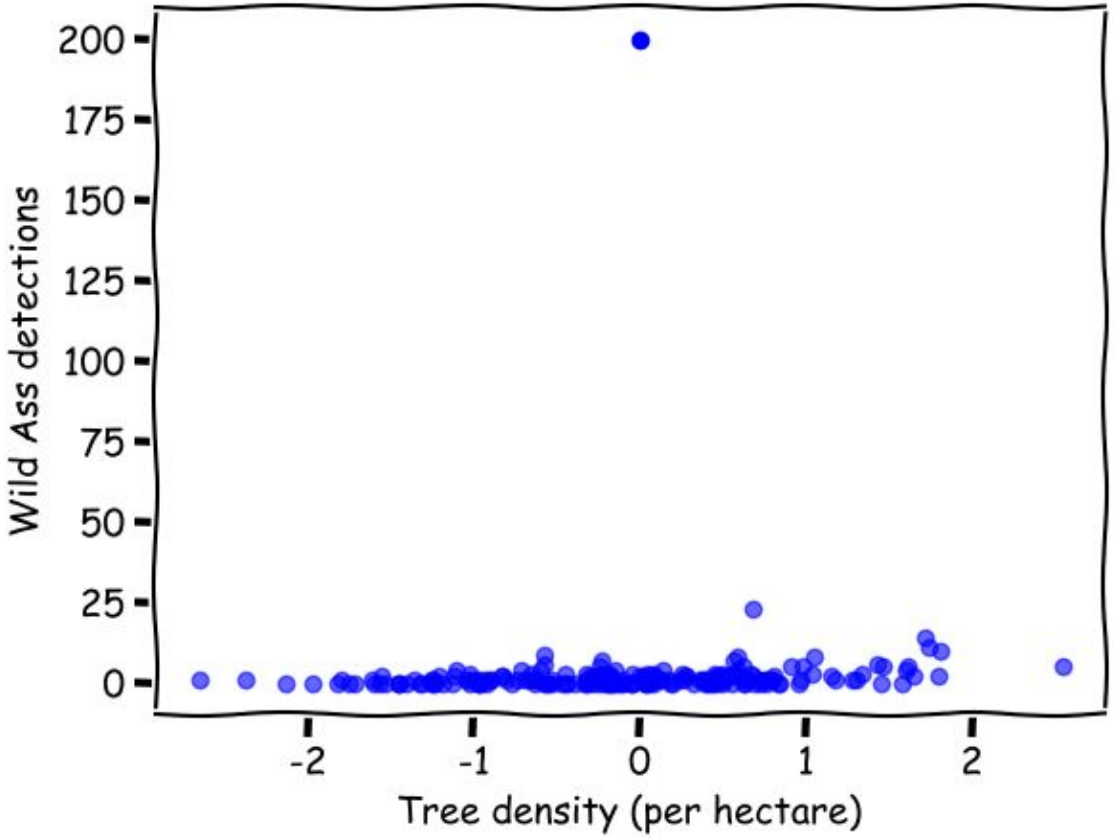
Inference problem:

Kazakhstan is facing serious desertification.

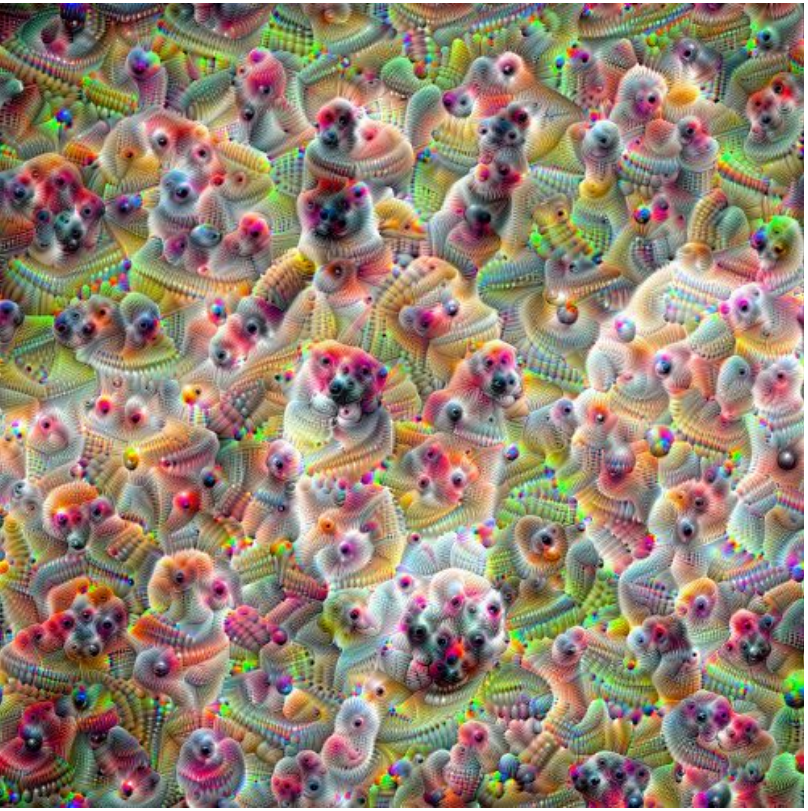
Will this affect the Asiatic Wild Ass population?

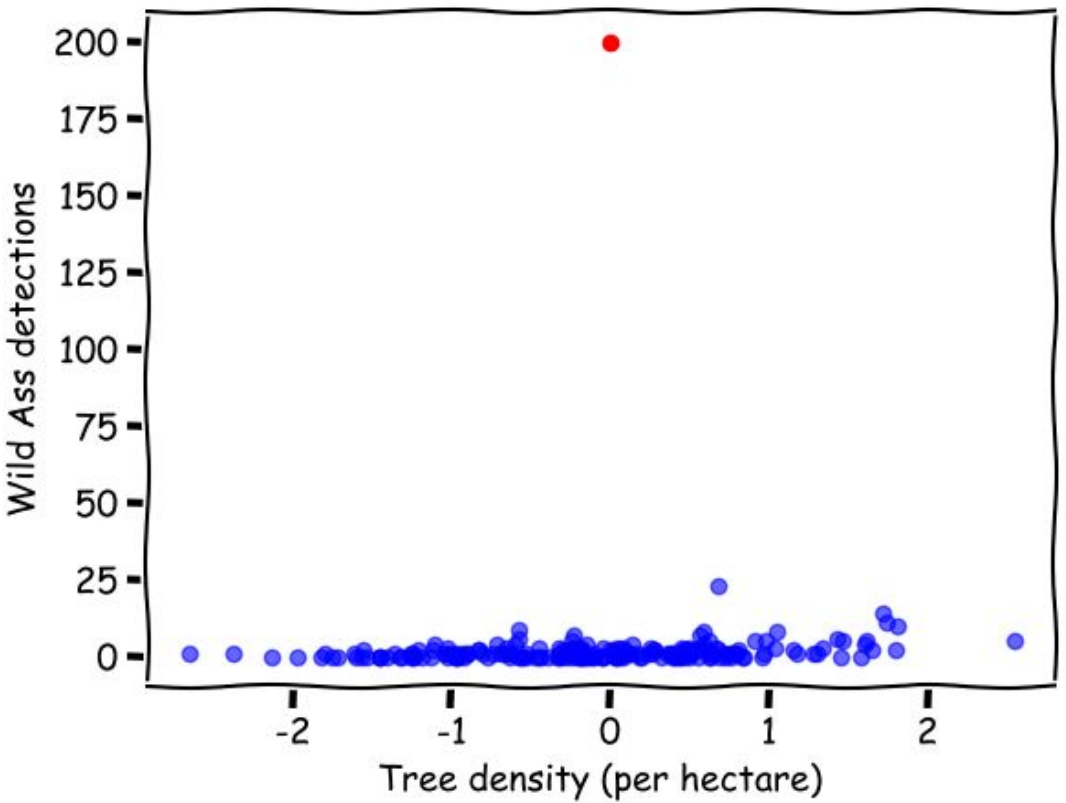
Does tree density affect Wild ass populations?

Does tree density affect Wild Ass populations?



Does tree density affect Wild Ass populations?

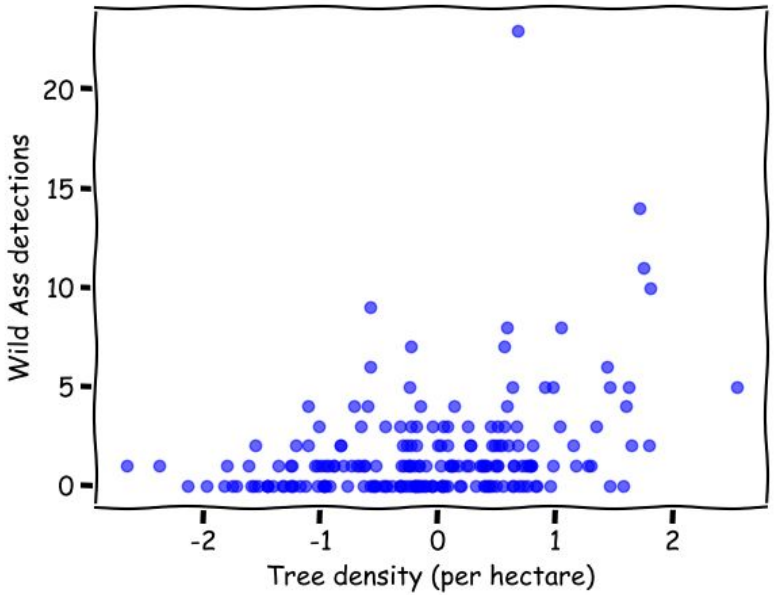




Does tree density affect Wild Ass populations?

1. Check outliers.
2. Calibrate by labeling some of the images.

	Predicted	True count
Image 1	5	4
Image 2	7	3
Image 3	0	0
Image 4	2	2



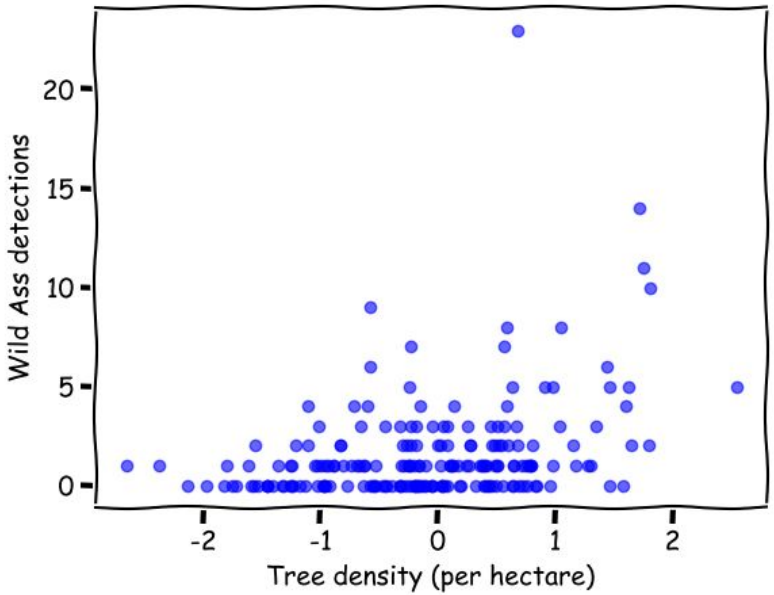
(I'm not plotting error bars because its ugly, but calibration gives you them)

Does tree density affect Wild Ass populations?

1. Check outliers.
2. Calibrate by labeling some of the images.
3. Fit statistical model.

$$Y_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$



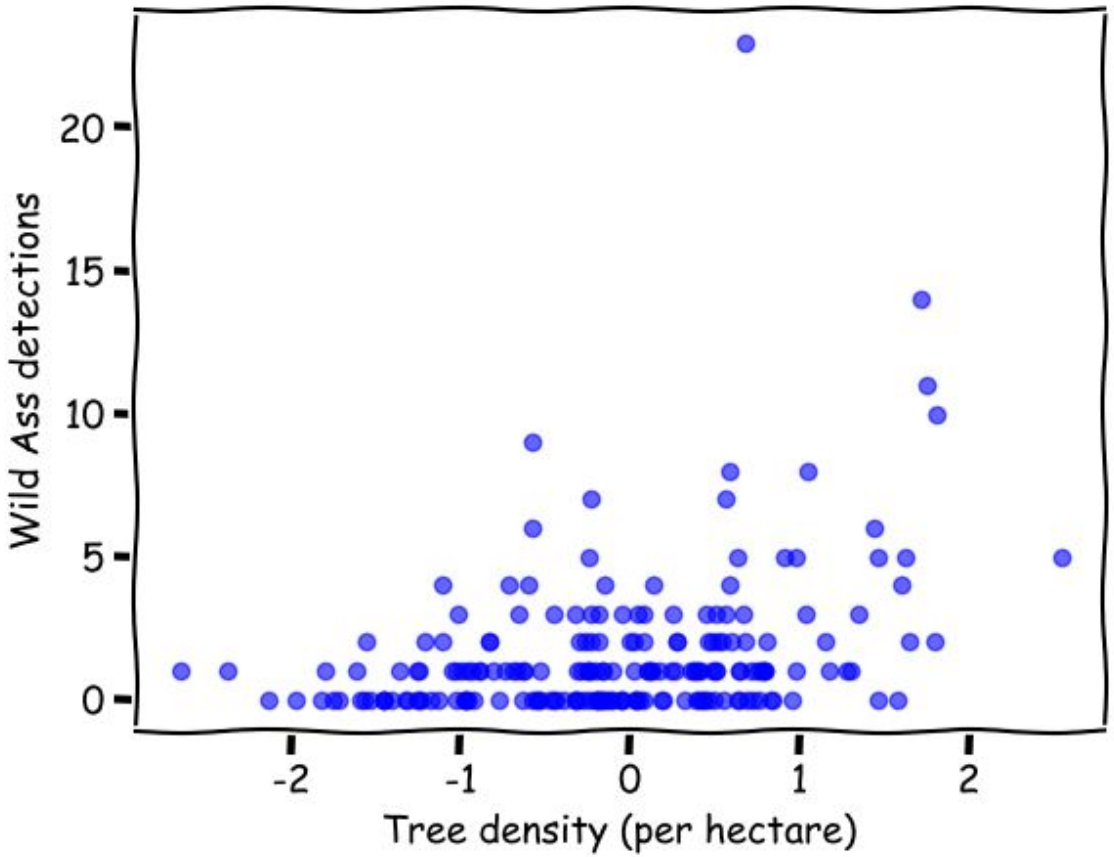
(I'm not plotting error bars because its ugly, but calibration gives you them)

Discuss

Have we done Ecological inference?

$$Y_i \sim \text{Poisson}(\lambda_i)$$

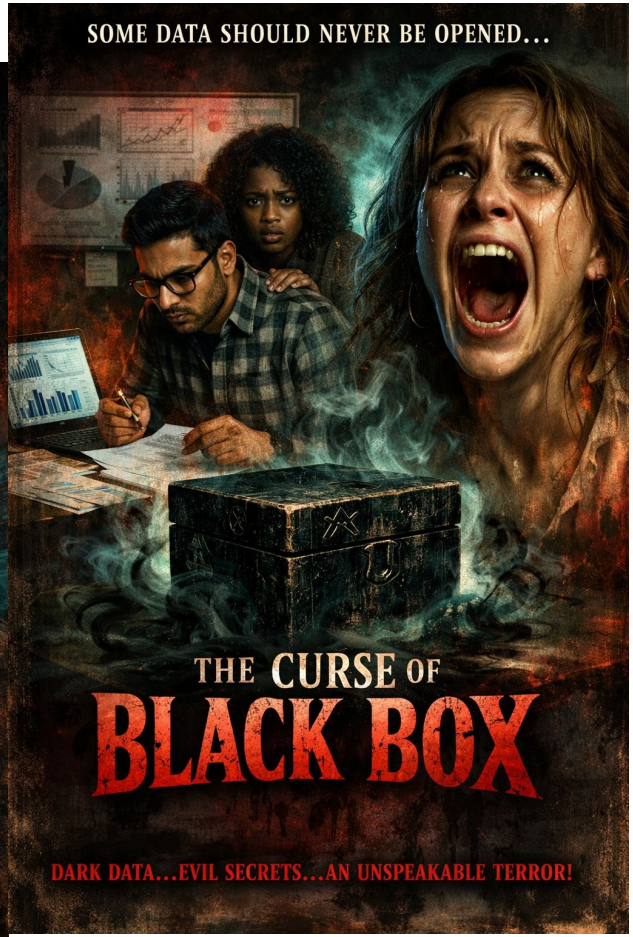
$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$



$\beta_1 = 1 \pm 0.3$


JUST BECAUSE YOU CAN
DETECT IT
DOESN'T MEAN IT'S REAL
with great statistical power comes great responsibility

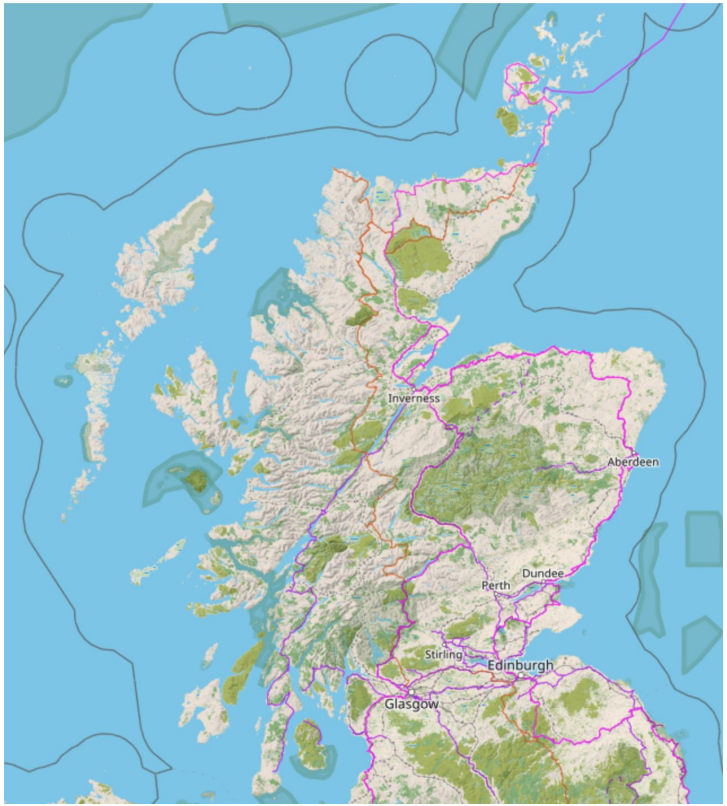

THE
UNIVERSAL FUNCTION
APPROXIMATOR
We proved it could represent the truth.
We never proved it would.


SOME DATA SHOULD NEVER BE OPENED...
THE CURSE OF
BLACK BOX
DARK DATA...EVIL SECRETS...AN UNSPEAKABLE TERROR!

A template:

<u>“Does proximity to a road affect Loon abundance?”</u>

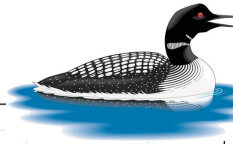
1. Collect audio from multiple sites.
2. Use classifier to estimate site-level Loon vocal density.
3. Do statistics!



Label and calibrate

Label random clips.

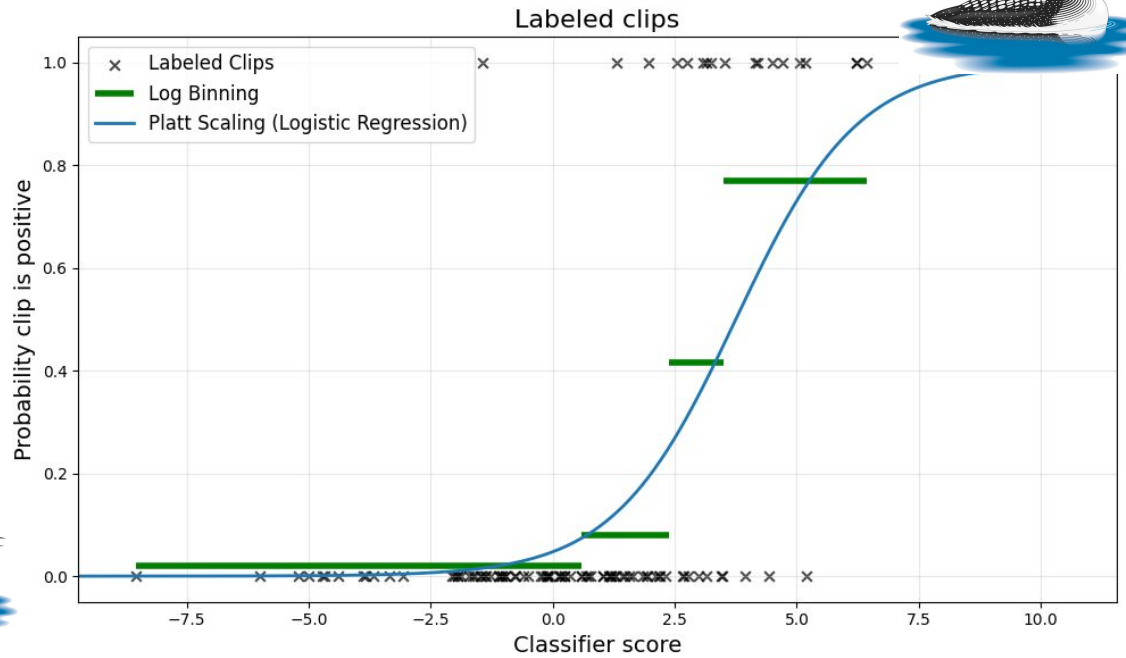
Recommendation:
Use log binning to stratify clips for labeling if your species is rare.



Label and calibrate

Calibrate

(Turn scores into probabilities)



Listen and calibrate

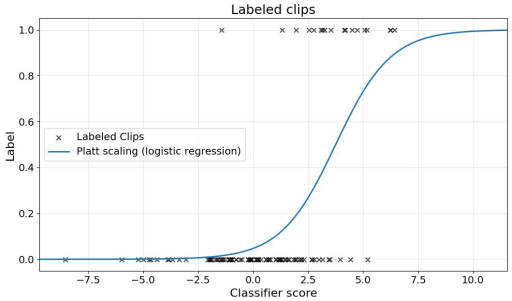
Site 1

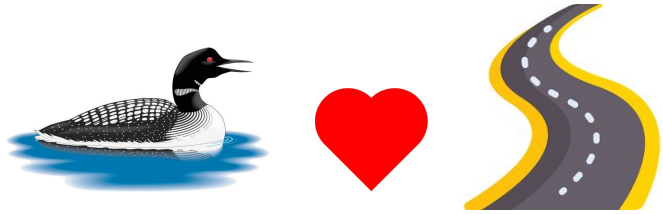
Score	Probability clip is a positive
1	0.30
-2	0.01
-1.5	0.05
…	…

Site 2

Score	Probability clip is a positive
-1	0.10
-2	0.01
3	0.50
…	…

. . .





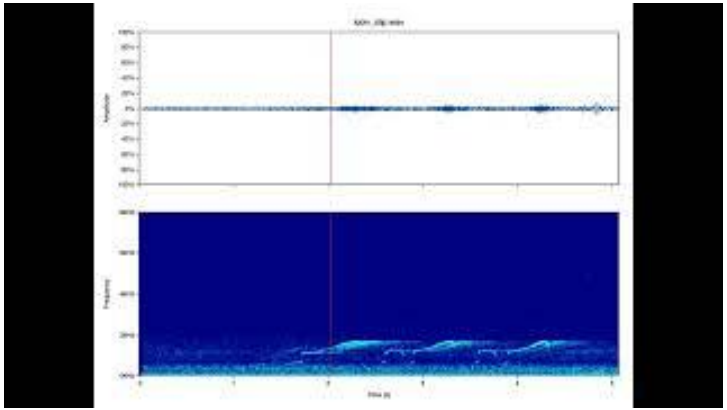
Loon vocal density
against proximity to road

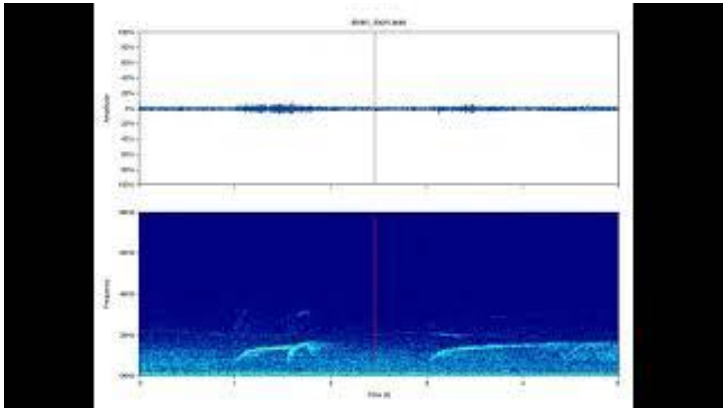
Vocal Density

0.30
0.25
0.20
0.15
0.10

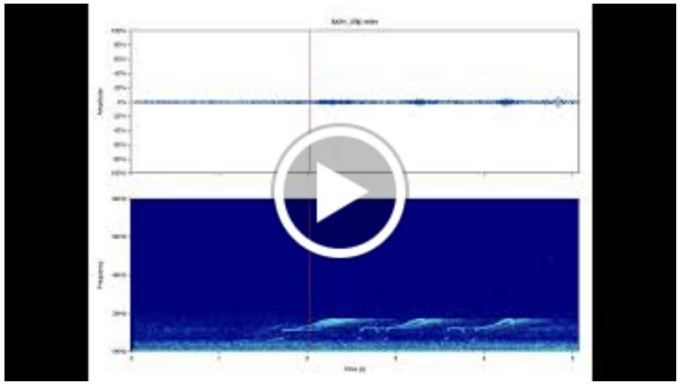
−2
−1
0
1

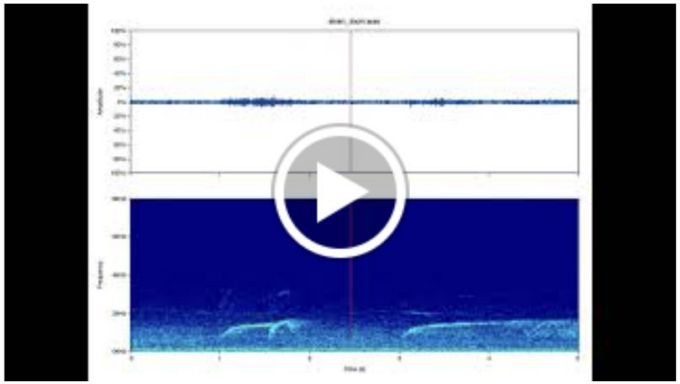
Proximity to road



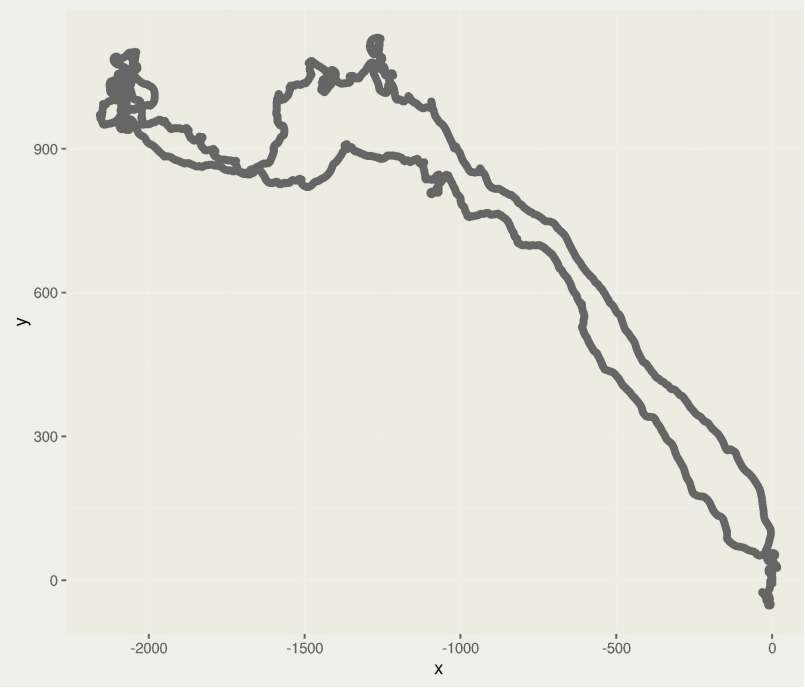



POLICE





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