Abyssal

Big (Unlabeled) Data:

A quest to create a fish detection dataset

Pedro Costa

Abyssal's *Head of Research* | INESC TEC Researcher

Agenda

- Abyssal
- How to annotate your dataset
- Creating a fish dataset


abyssal

abyssal

Founded in January 2012

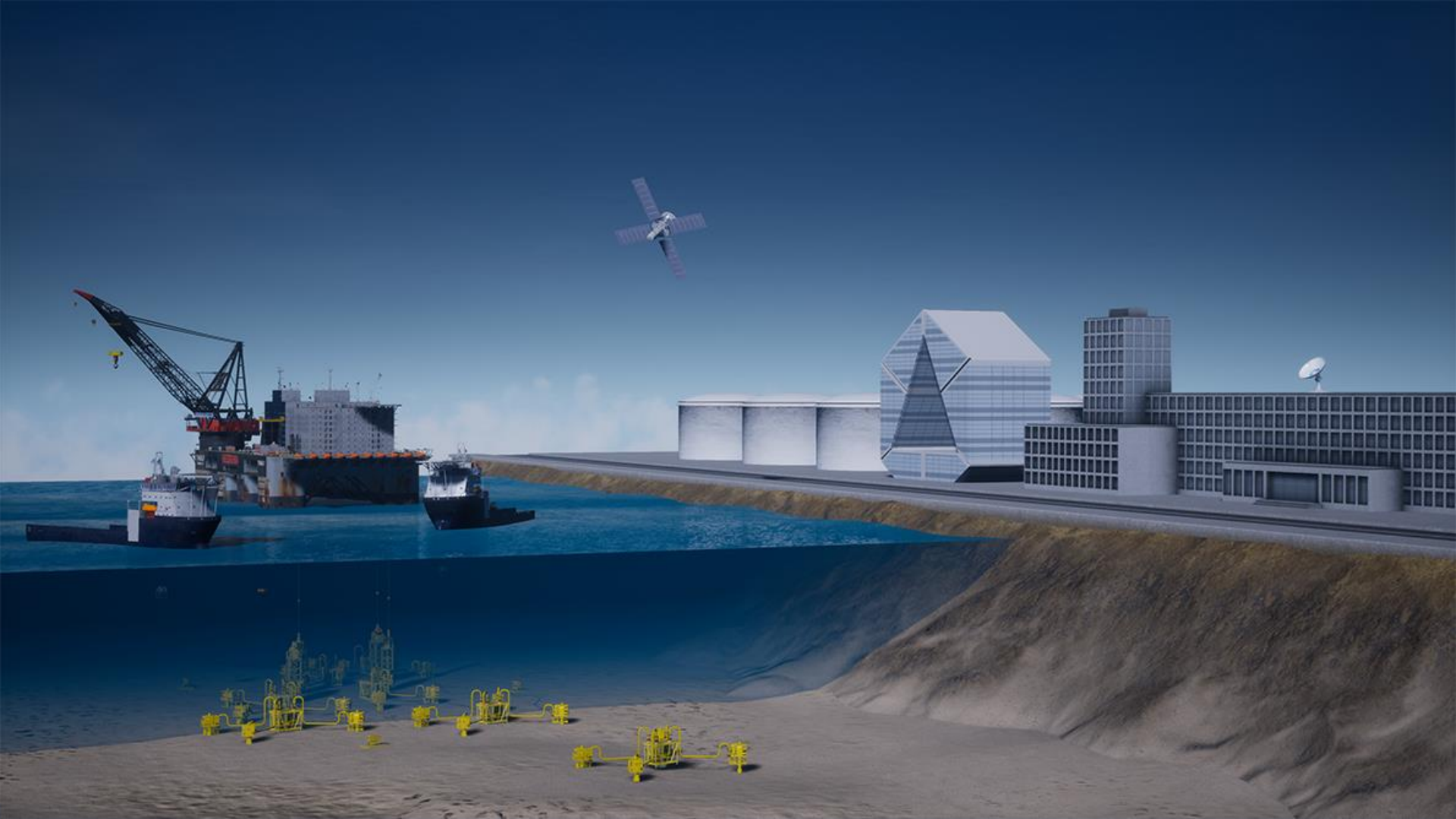
Team of 25+ Software Engineers, 3D Artists, ...

Augmented Reality for subsea Remotely Operated Vehicles (ROVs)

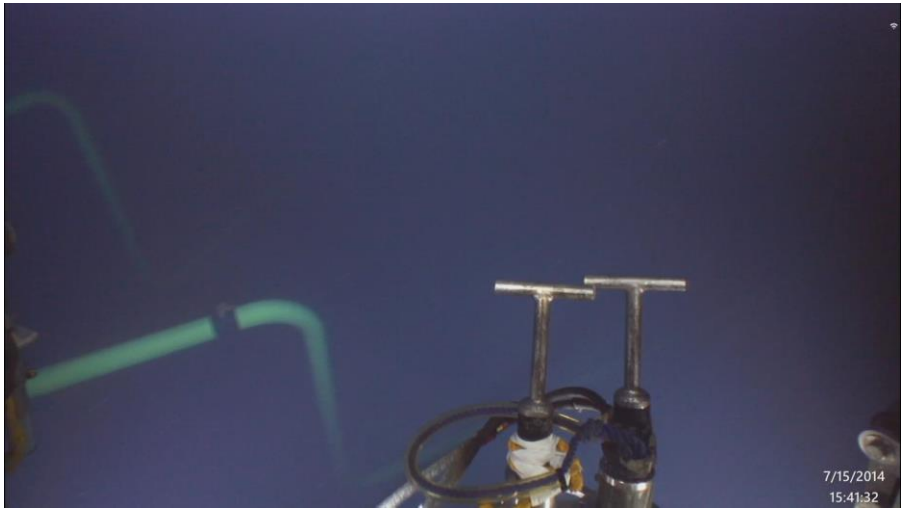
RESEARCH TEAM GOAL

Autonomous subsea operations!

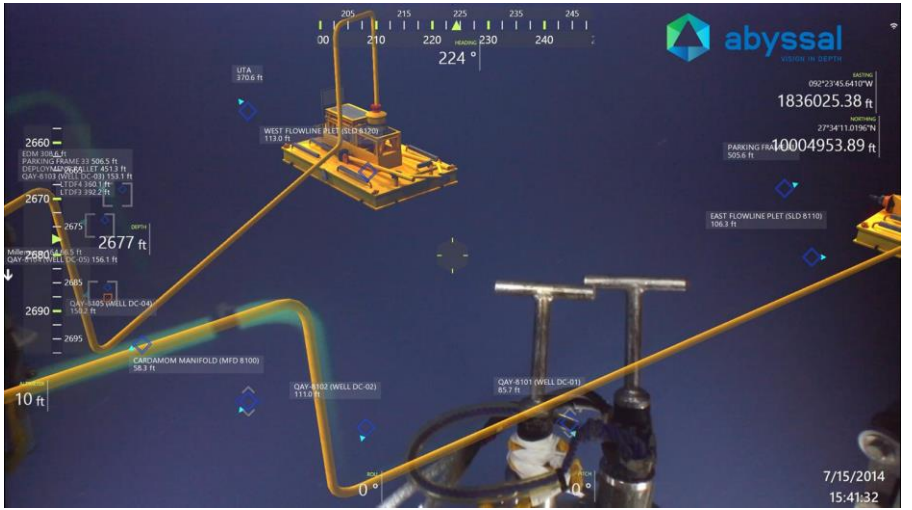
Planning and ROV control for inspection and maintenance.



Real



Real & Virtual



abyssal

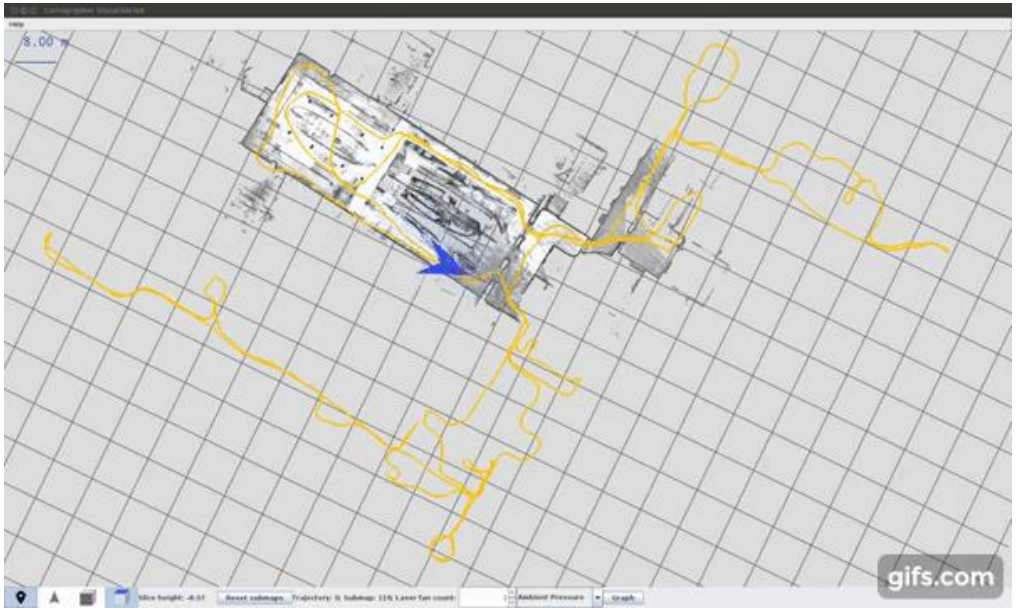
Simulator

- Train ROV pilots
- Planning new operations
- **Machine Learning!**

Research Projects

SLAM

SLAM - Simultaneous Localization and Mapping

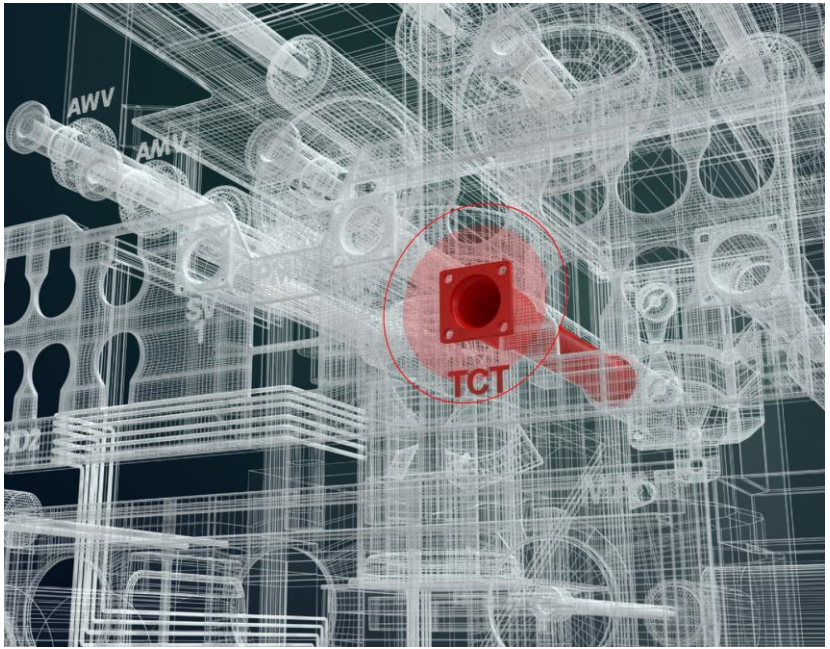


abyssal

Anomaly Detection

Automatically detect anomalies in marine renewables' subsea structures.

- Biofouling
- Sacrificial anodes
- Leaks
- ...



abyssal

Problem Definition

Find ~~anomalies~~ fish in ROV videos!

Predict if a **fish** is visible in a given **frame**.

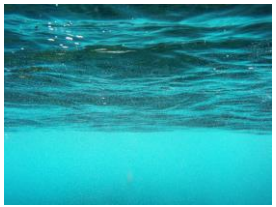
EMEPC gave us some of their videos:

- **3.9 TB!**
- More than 250 videos from 2008-2012.
- Some with less than 10 min other more than 1 hour.
- … Highly unbalanced!

Research Question

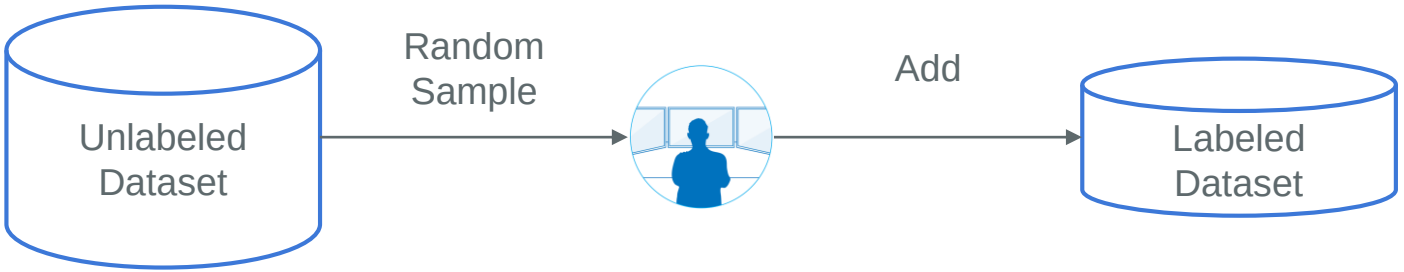
What data point (frame/video/segment) do we sample next to be labeled?



?

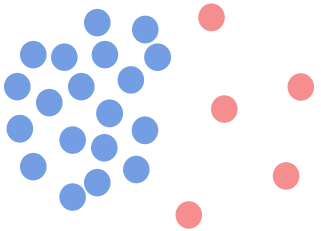
Typical approach



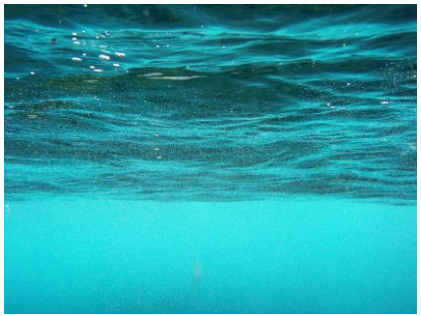
Is random sampling the best we can do?

Random Sampling problems

Unbalanced Datasets

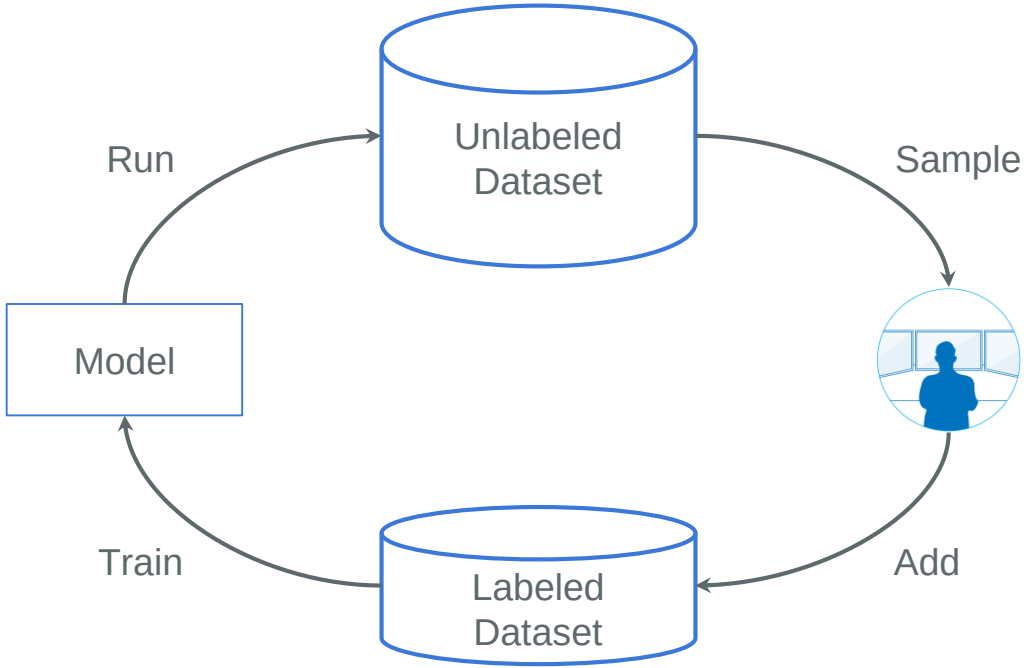


Repeated/Similar/Correlated examples

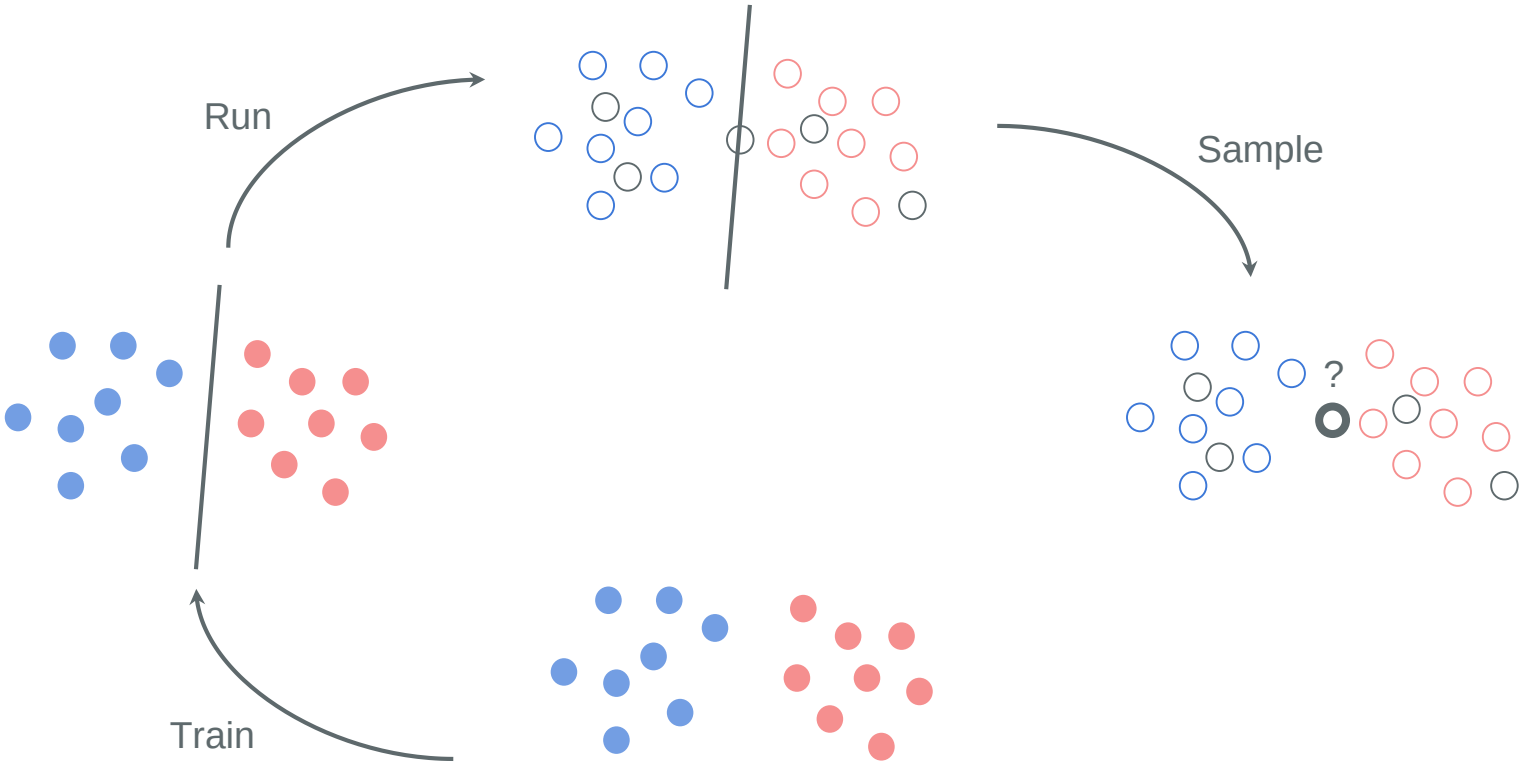


Active Learning

Use the model to sample new examples to be annotated.

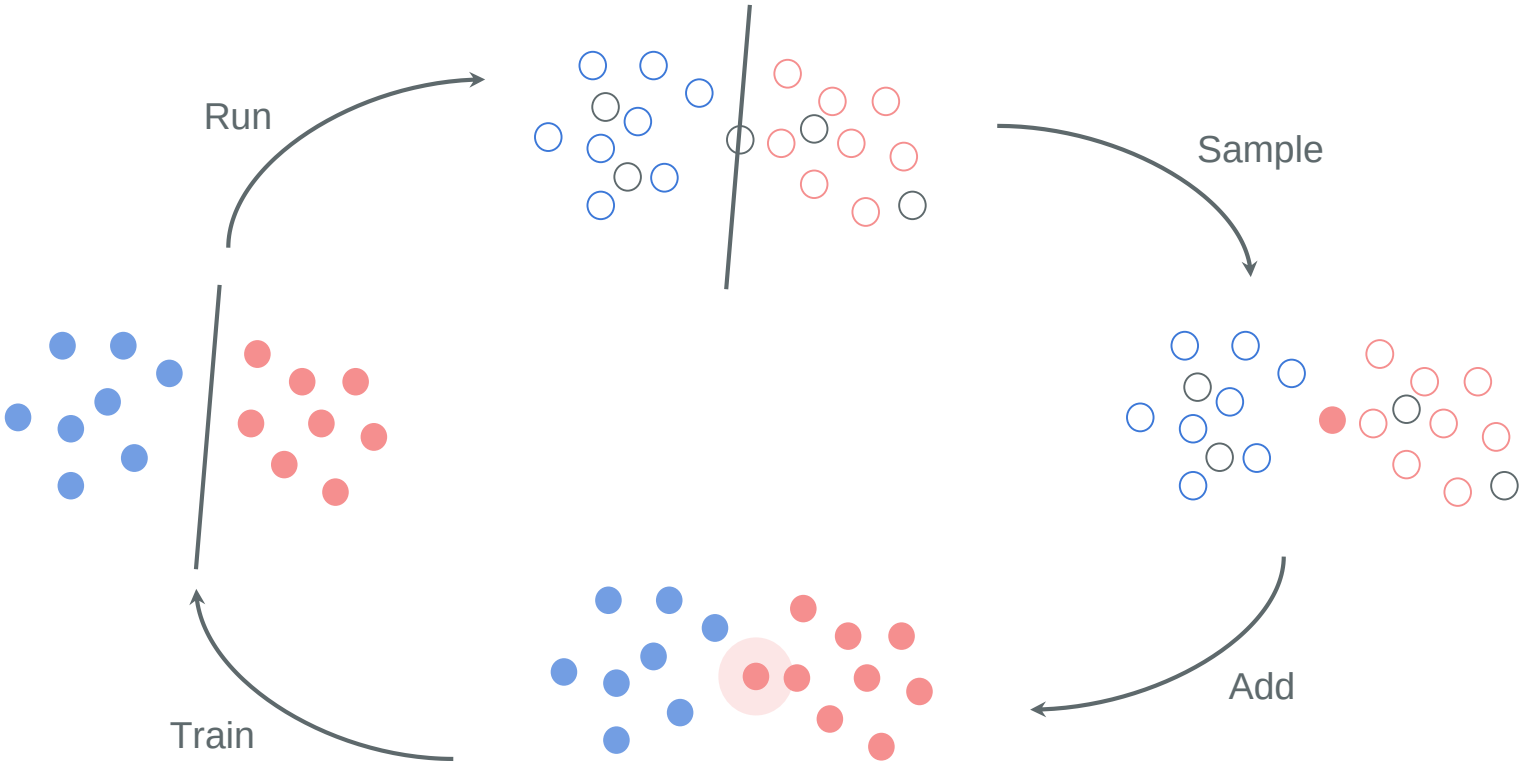


Intuition



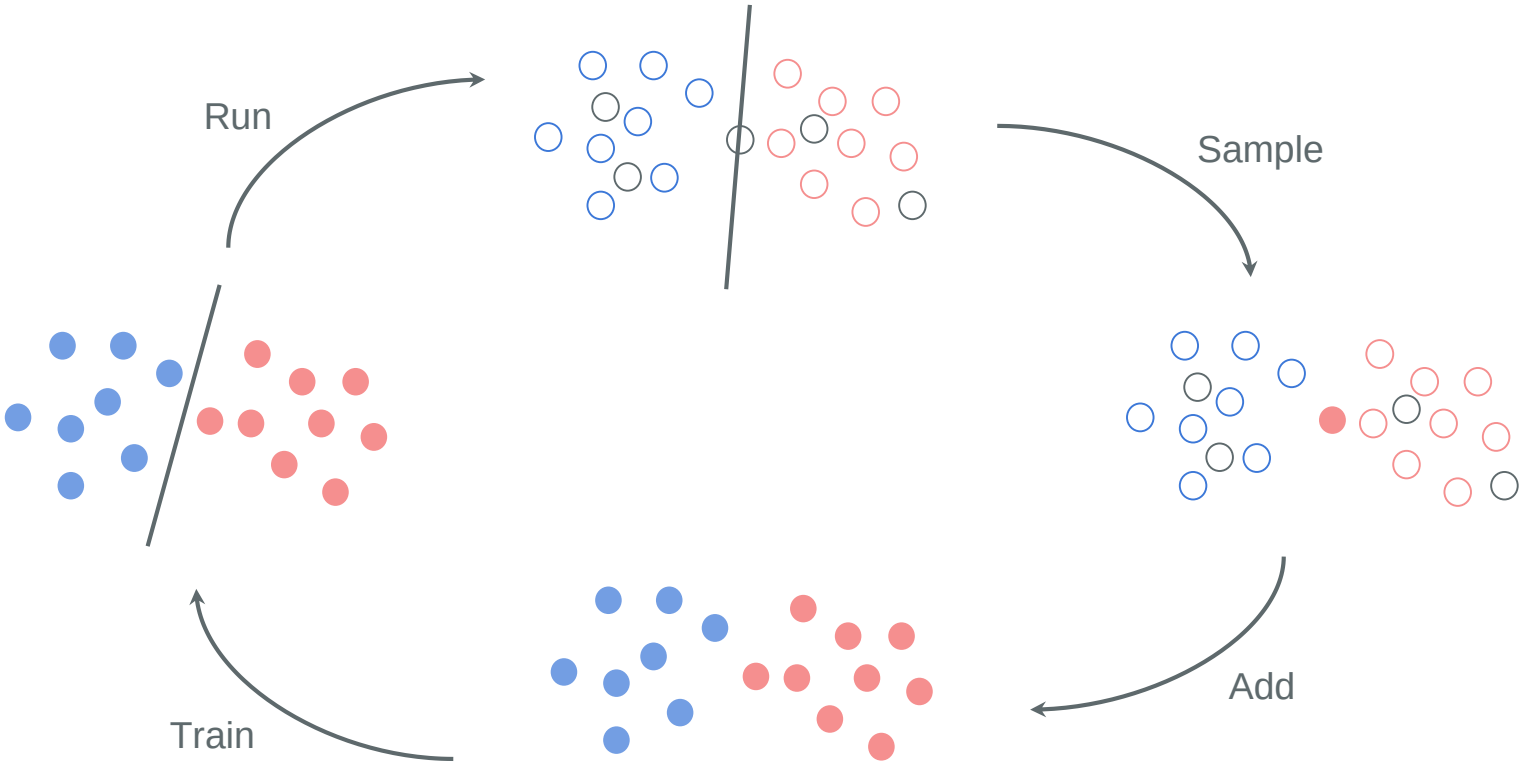
abyssal

Intuition



abyssal

Intuition



abyssal

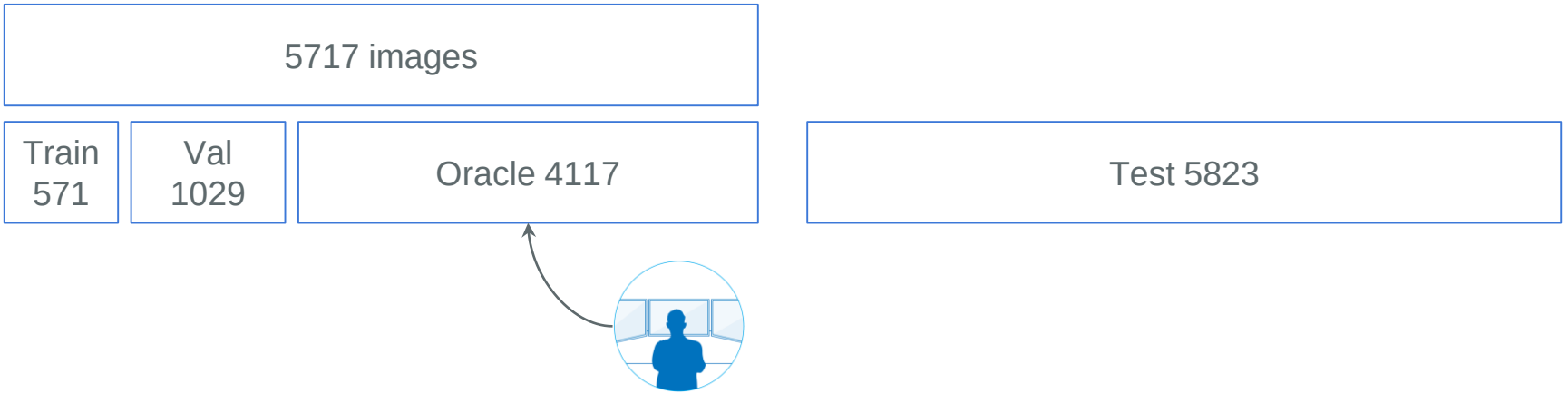
Active Learning

What do we need?

1. Dataset
2. Sampling Method
3. Model

Dataset: VOC 2012

Solve the **Cat vs No-Cat** task.



Unbalanced: Only 7.53% training images contain cats.

Sampling: Predictive Entropy

Also known as Uncertainty Sampling.

Select the example x^* that minimizes the predictive entropy:

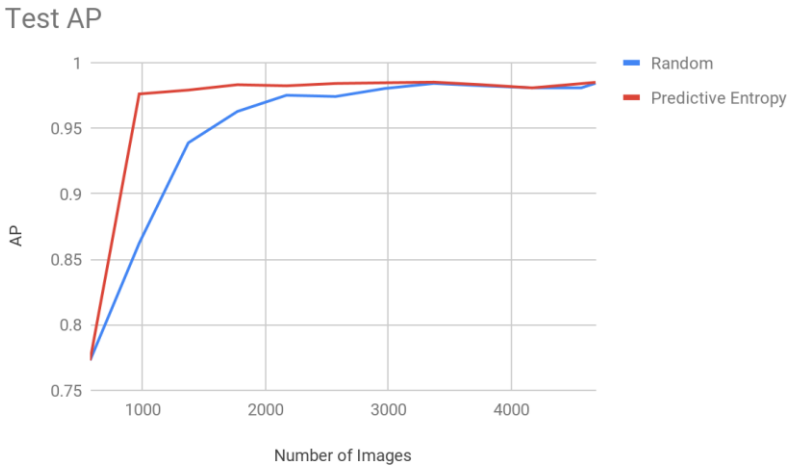
$$y = \text{model}(x)$$

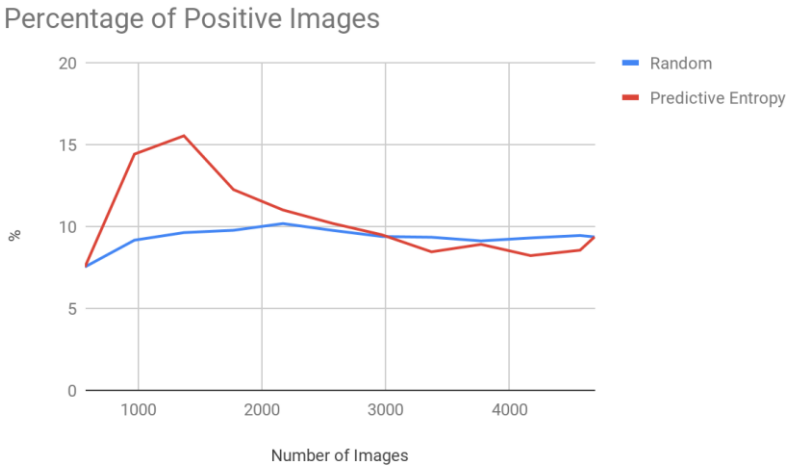
$$x^* = \arg\min y * \log(y) + (1 - y) * \log(1 - y)$$

Just select the example that is closer to the decision boundary!

Results: Pre-Trained

- ResNet50 pre-trained on ImageNet;
- Sampling 400 images at each iteration.





abyssal

Examples

“Difficult” images are added first!

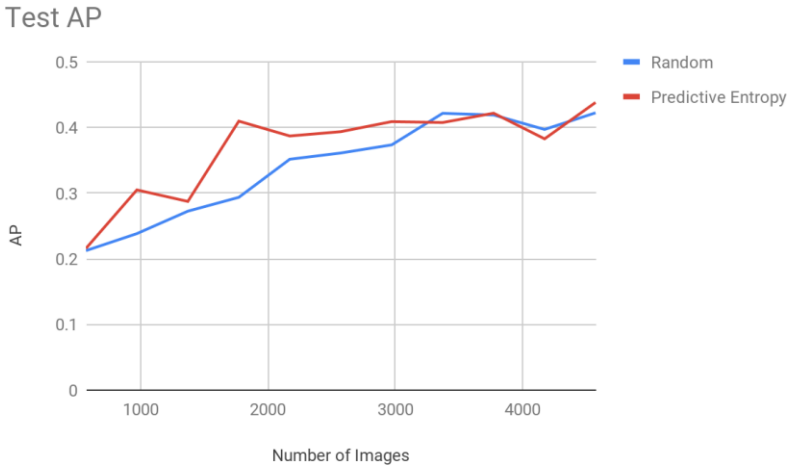
First Iteration

Last Iteration

abyssal

Results: Scratch

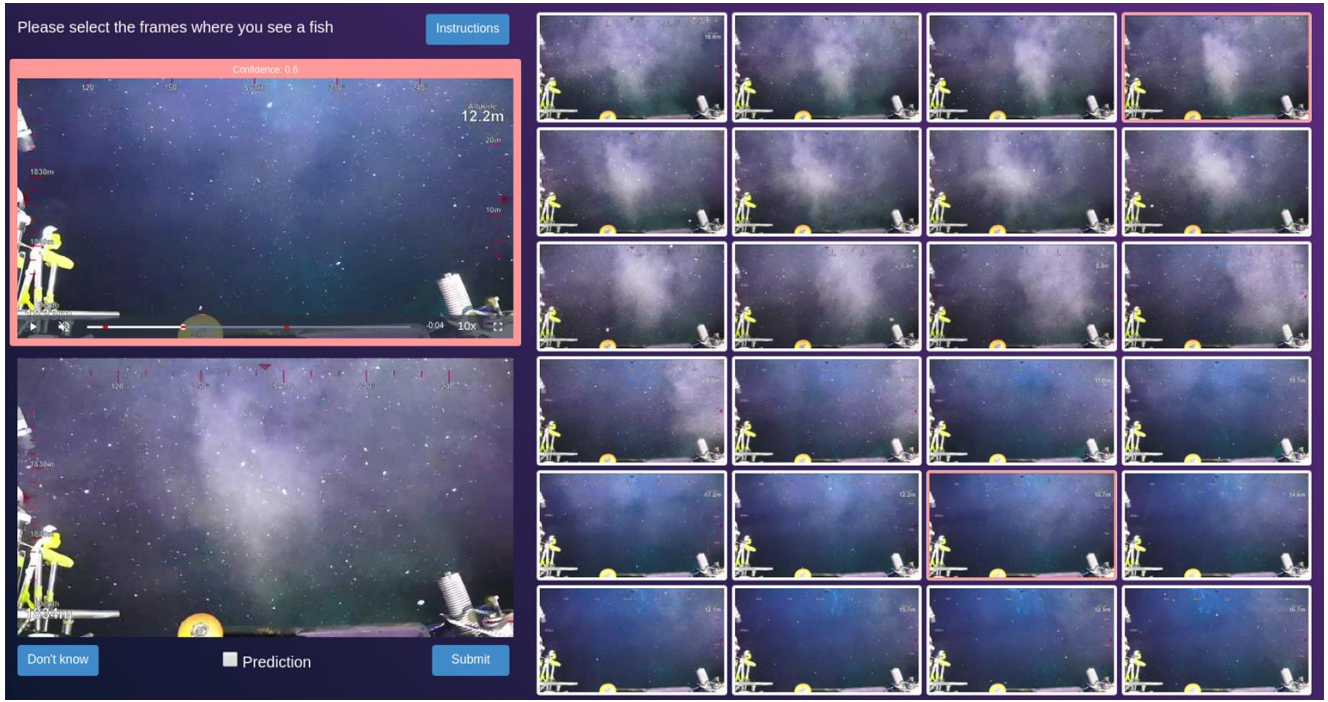
- SqueezeNet;
- Sampling 400 images at each iteration.



abyssal

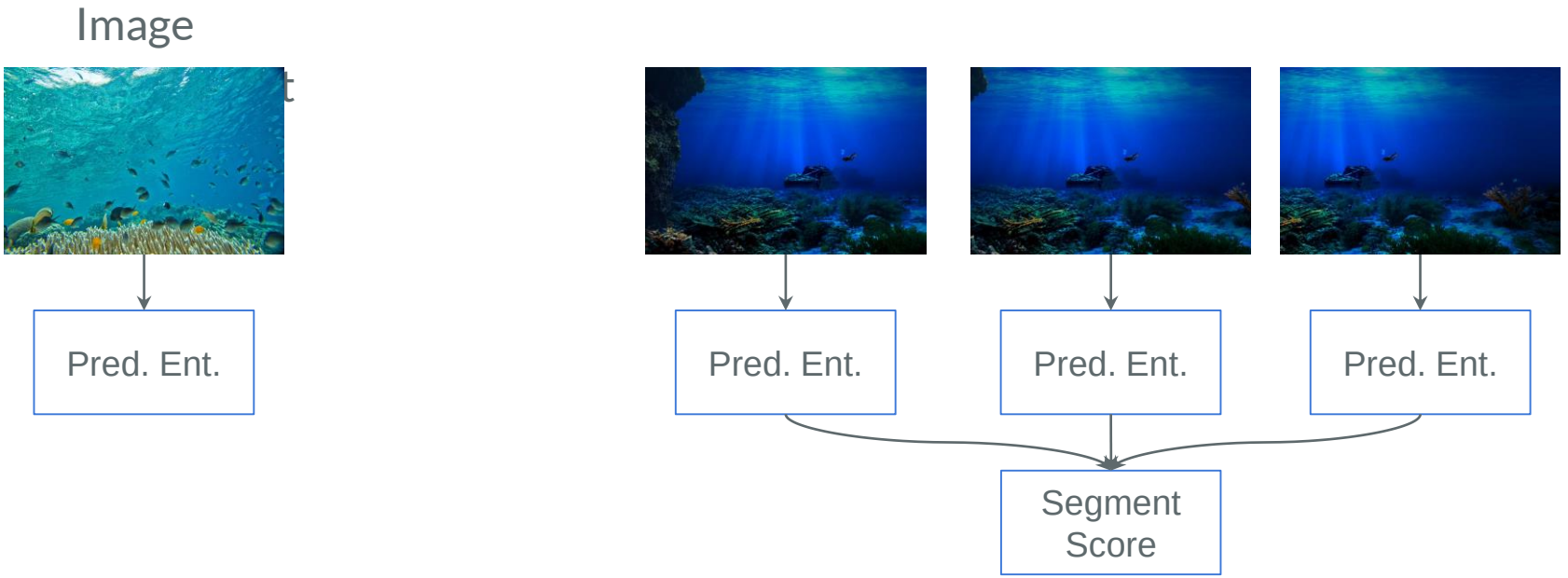
Now to the fish dataset!

We created a tool to label one minute video segments at a time.



abyssal

From images to videos



Combine frame scores using **mean/min** to get the segment score.

abyssal

Dataset

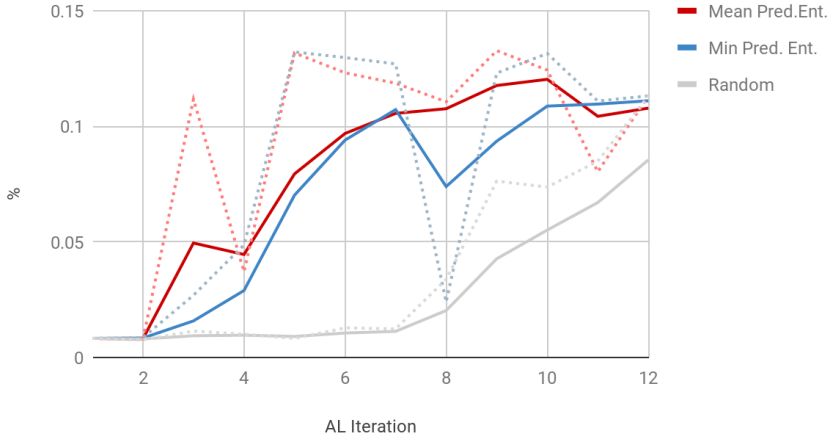
- Assign different videos from 2012 to train, val and oracle.
- Assign videos from 2008 to test.
- Randomly sample video segments.
- Unbalanced. Only 2 video segments in the test and val sets depict fish.
- Large intra-class variability.

661 segments 39087 images			
Train: 50 segments 2903 imgs. (4.2%)	Val: 61 segments 3611 imgs. (0.6%)	Oracle: 500 segments 29594 imgs. (0.8%)	Test: 50 segments 2979 imgs. (1.3%)

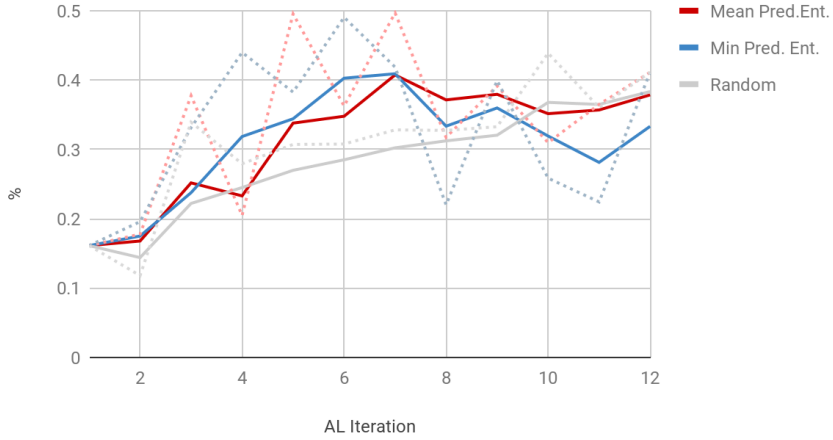
Results

- ResNet50 pre-trained on ImageNet;
- Sampling 50 segments at each iteration.

Test AP



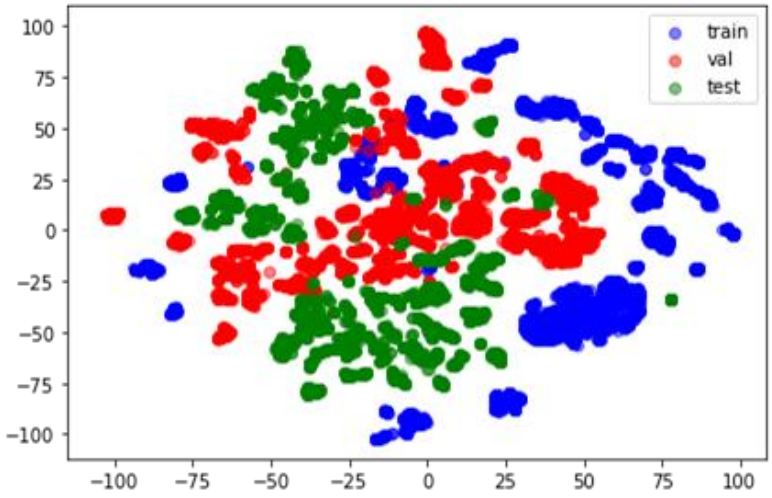
Val AP



abyssal

t-SNE on the features

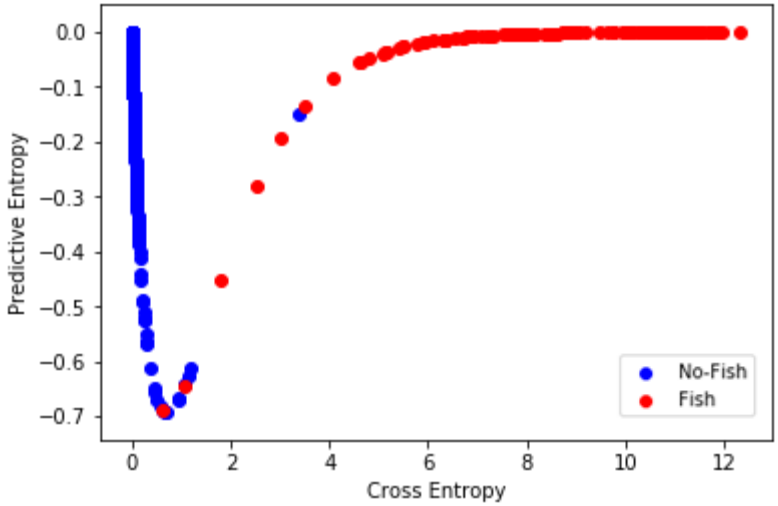
train/val/test features are very different!
This coupled with high intra-class variability may explain low performance.



More data and semi-supervised methods may be able to help...

Caution!

Fish images are being wrongly classified with absolute certainty!



abyssal

Deep Magic

Deep Learning methods can fit a dataset with random labels (even with regularization)!!!

Maybe Predictive Entropy is not the optimal sampling strategy.

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

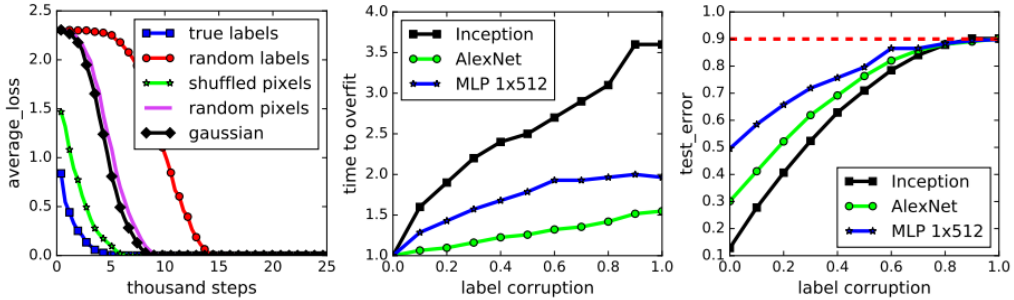
Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com



(a) learning curves (b) convergence slowdown (c) generalization error growth

AL from the start

53236 labeled frames.

Sampled segments are divided into train/val, making sure that there are no segments from the same video on different sets.

Potential Problem: val set is biased.

	New Dataset	Old Dataset
Pos. Imgs. (%)	**6.07%**	1.0%
Val AP	**70.6%**	49.5%
Val AUC	90.1%	**97.3%**

Future Work

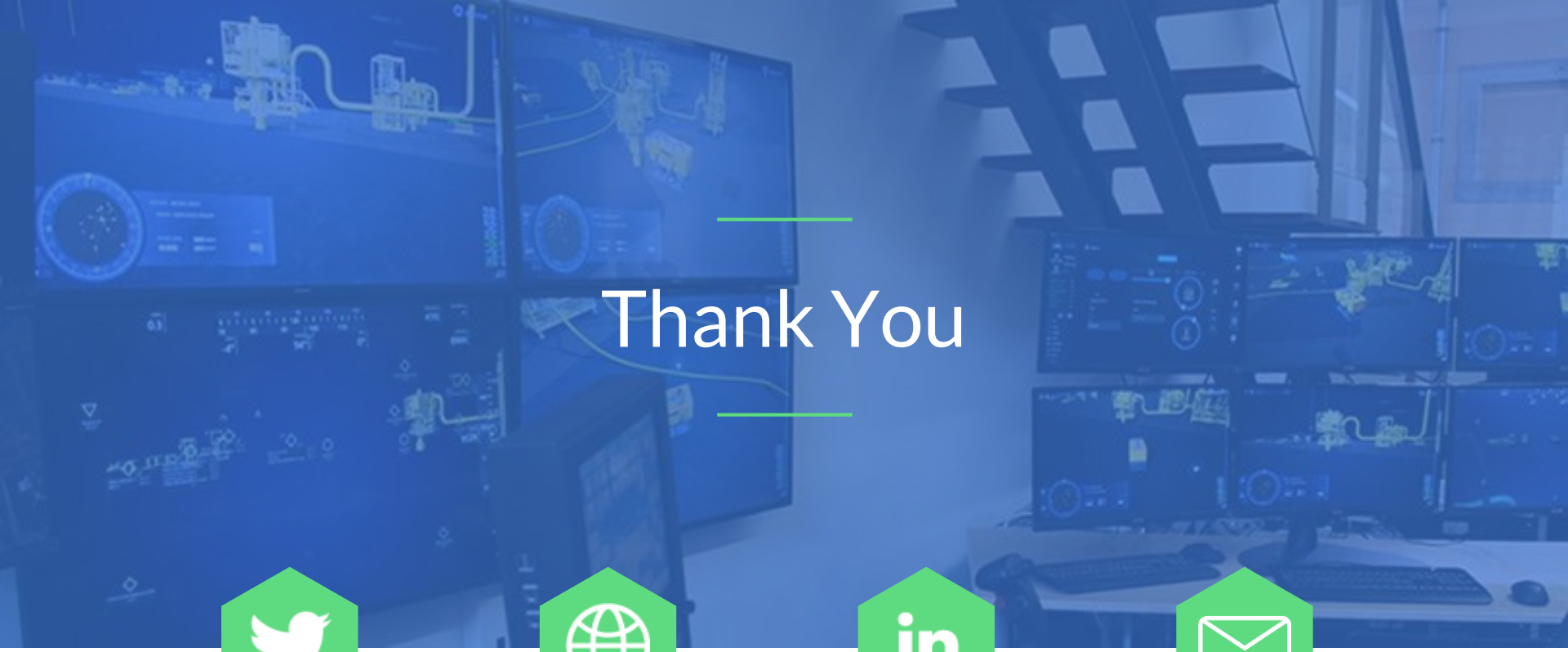
Move from heuristic-based sampling to learned sampling functions.

- Generative model on the feature space;
- Learn what regions of the feature space are more prone to error;
- Predict model improvements after adding the example to the training set.

Label more data.

abyssal

We are hiring!



Thank You

@abyssal_sa

abyssal.eu

Abyssal S.A.

Pedro Costa
pcosta@abyssal.eu