It’s The Data, Stupid!

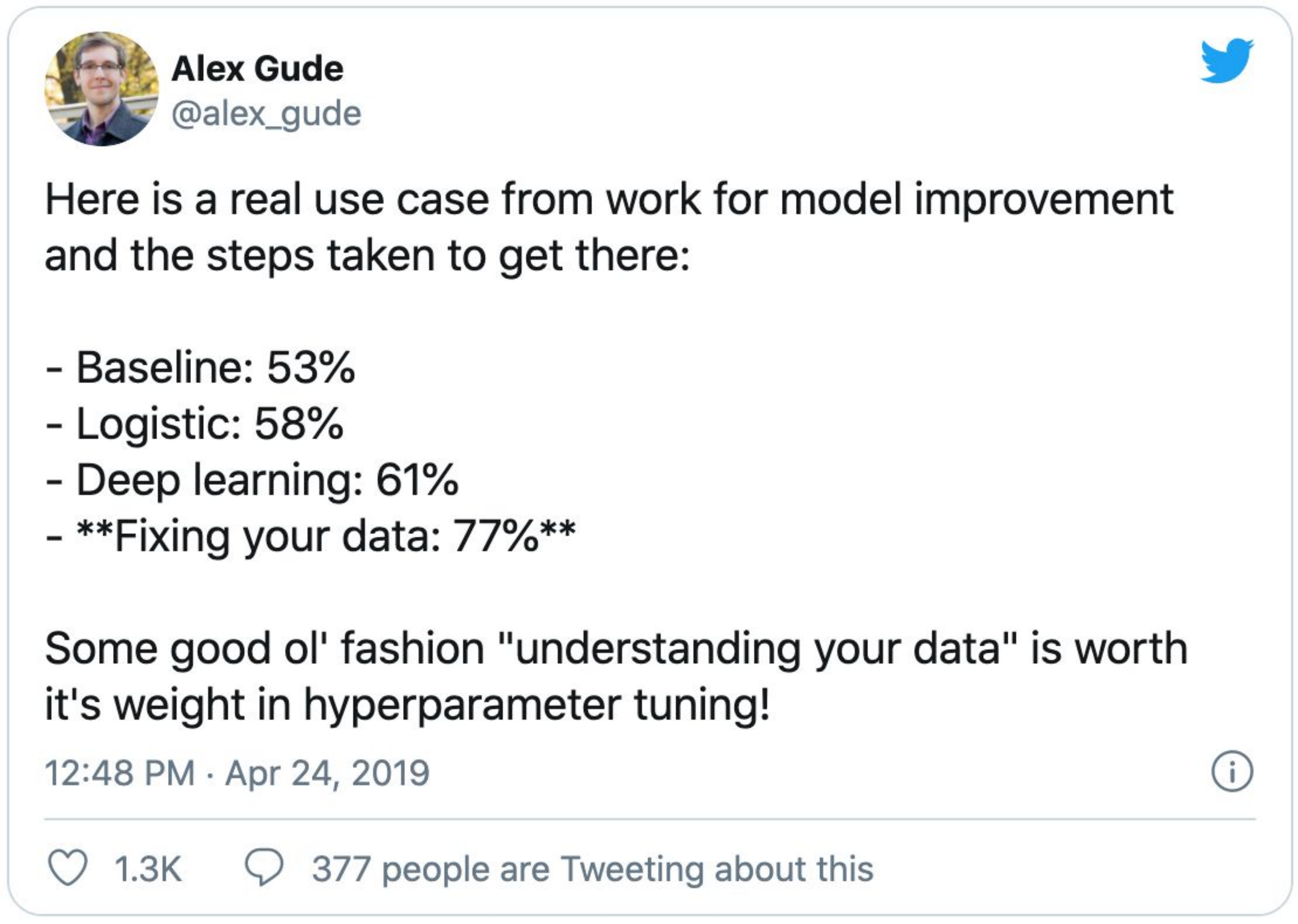
March 24, 2022
Peter Gao

Peter Gao

Cofounder and CEO at Aquarium

Early employee at Cruise, led CV team

Deep Learning research at Berkeley



Alex Gude
@alex_gude

Here is a real use case from work for model improvement and the steps taken to get there:

- Baseline: 53%
- Logistic: 58%
- Deep learning: 61%
- **Fixing your data: 77%**

Some good ol' fashion "understanding your data" is worth it's weight in hyperparameter tuning!

12:48 PM · Apr 24, 2019

1.3K   377 people are Tweeting about this

How To Improve Your ML System

- Improve your model code
- Improve your training dataset
- Do these faster and more frequently

“Old School” ML Vs Deep Learning

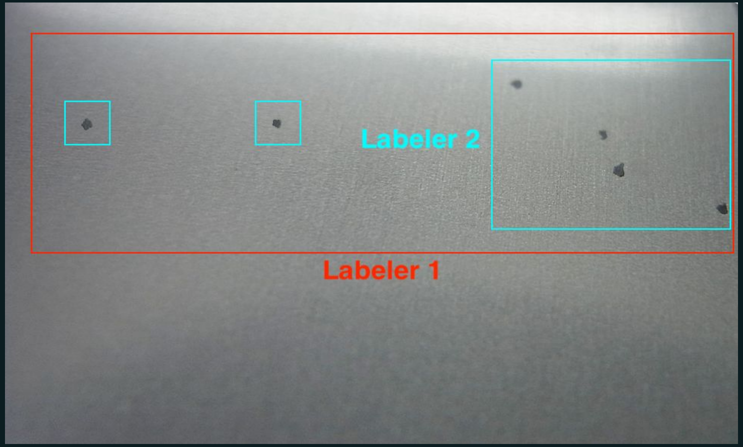
	Old School ML	Deep Learning
Example Tasks	Forecasting, recommendations	Object detection, seq2seq
Data types	Structured / tabular data	Unstructured data (imagery, audio, etc.)
Labeling	Labels come “for free”	Pay people to label data
Algorithms	Logistic regression, SVMs, random forests	Neural networks
Development Emphasis	- Data pipelines + infrastructure - Feature engineering - Model experimentation (ex: sparsity)	- Data pipelines + infrastructure - Fine tuning pretrained models - Improving quality + variety of datasets

How Do You Improve Your Data?

- Find problems in the data / model performance
- Figure out why the problems are happening
- Modify your dataset to fix the problems
- Make sure the problems are fixed as you retrain your model on the new dataset
- Deploy new model + repeat

Types Of Data Problems

- Invalid data
- Labeling errors, ambiguities
- Difficult edge cases
- Out of sample data



How Do You Improve Datasets Efficiently?

- Problem: Identifying failure cases in the data / model performance
  - Lots of labeled data, only a few examples are problematic. Labor intensive to dig through haystack looking for the needle
  - Example: Triple QA finds many issues but can 3x your labeling cost
  - Example: Hard to understand model failure modes without metadata to slice on

How Do You Improve Datasets Efficiently?

- Solution: Get feedback signal that tells you where to look
    - Feedback from double-checking
        - Human-check prod model outputs (example: customer feedback)
        - Check disagreement between automated systems
    - Feedback from model
        - High loss disagreements with labels (tend to be labeling errors)
        - Error patterns vs labels (in metadata + raw data)
        - Distributional shifts between training + prod environments

Example: KITTI

Example: KITTI

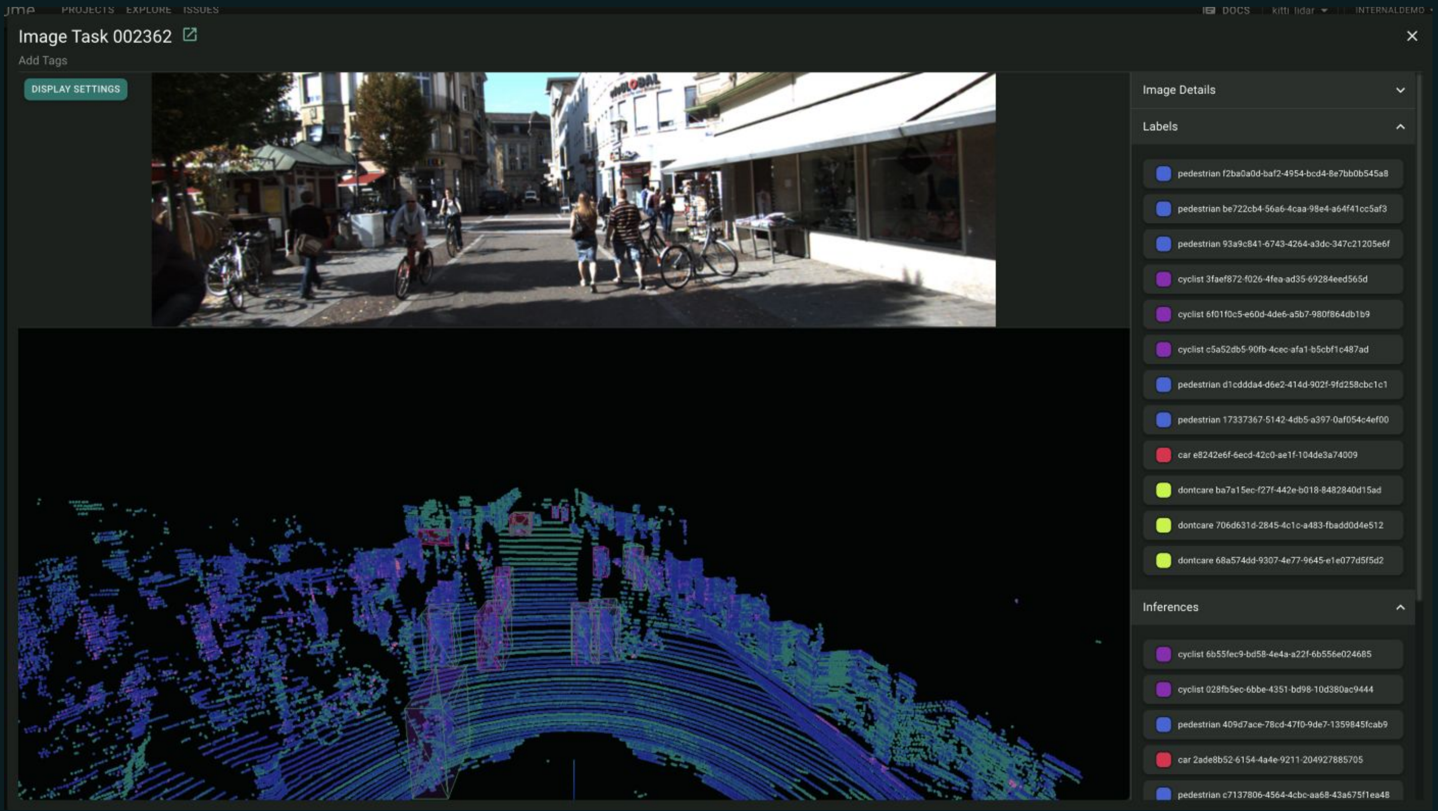


Image Task 002362

Add Tags

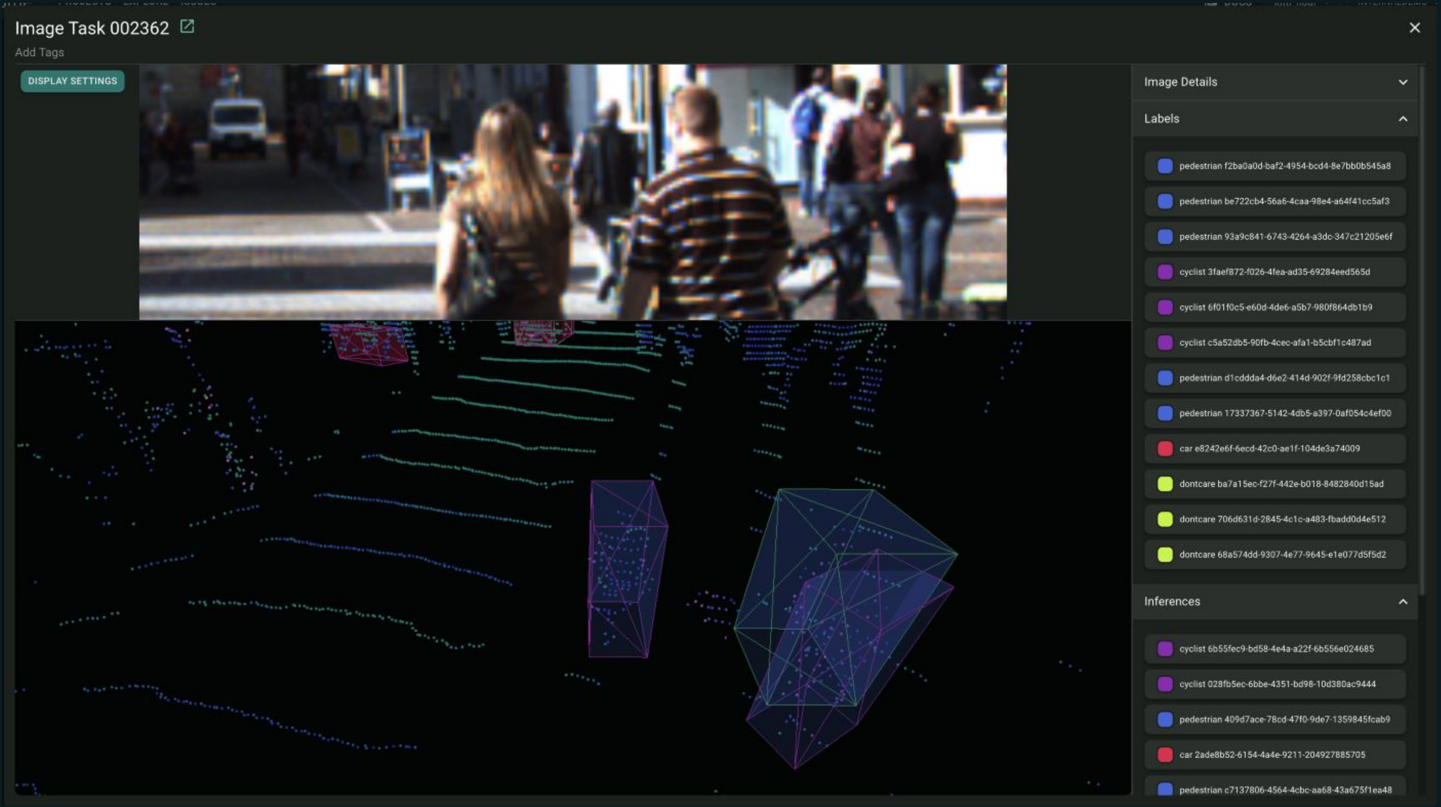
Labels

- 🟦 pedestrian f2ba0a0d-baf2-4954-bcd4-8e7bb0b545a8
- 🟦 pedestrian be722cb4-56a6-4caa-98e4-a64f41cc5af3
- 🟦 pedestrian 93a9c841-6743-4264-a3dc-347c21205e6f
- 🟪 cyclist 3faef872-f026-4fea-ad35-69284eed565d
- 🟪 cyclist 6f01f0c5-e60d-4de6-a5b7-980f864db1b9
- 🟪 cyclist c5a52db5-90fb-4cec-afa1-b5cbf1c487ad
- 🟦 pedestrian d1cddda4-d6e2-414d-902f-9fd258cbc1c1
- 🟦 pedestrian 17337367-5142-4db5-a397-0af054c4ef00
- 🟥 car e8242e6f-6ecd-42c0-ae1f-104de3a74009
- 🟨 dontcare ba7a15ec-f27f-442e-b018-8482840d15ad
- 🟨 dontcare 706d631d-2845-4c1c-a483-fbadd0d4e512
- 🟨 dontcare 68a574dd-9307-4e77-9645-e1e077d5f5d2

Inferences

- 🟪 cyclist 6b55fec9-bd58-4e4a-a22f-6b556e024685
- 🟪 cyclist 028fb5ec-6bbe-4351-bd98-10d380ac9444
- 🟦 pedestrian 409d7ace-78cd-47f0-9de7-1359845fcab9
- 🟥 car 2ade8b52-6154-4a4e-9211-204927885705
- 🟦 pedestrian c7137806-4564-4cbc-aa68-43a675f1ea48

Example: KITTI



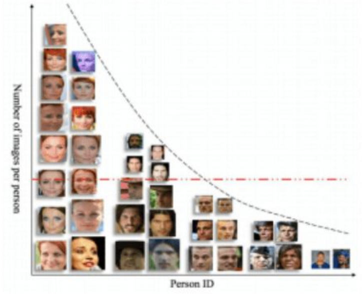
Labels

- 🟦 pedestrian f2ba0a0d-baf2-4954-bcd4-8e7bb0b545a8
- 🟦 pedestrian be722cb4-56a6-4caa-98e4-a64f41cc5af3
- 🟦 pedestrian 93a9c841-6743-4264-a3dc-347c21205e6f
- 🟪 cyclist 3faef872-f026-4fea-ad35-69284eed565d
- 🟪 cyclist 6f01f0c5-e60d-4de6-a5b7-980f864db1b9
- 🟪 cyclist c5a52db5-90fb-4cec-afa1-b5cbf1c487ad
- 🟦 pedestrian d1cddda4-d6e2-414d-902f-9fd258cbc1c1
- 🟦 pedestrian 17337367-5142-4db5-a397-0af054c4ef00
- 🟥 car e8242e6f-6ecd-42c0-ae1f-104de3a74009
- 🟨 dontcare ba7a15ec-f27f-442e-b018-8482840d15ad
- 🟨 dontcare 706d631d-2845-4c1c-a483-fbadd0d4e512
- 🟨 dontcare 68a574dd-9307-4e77-9645-e1e077d5f5d2

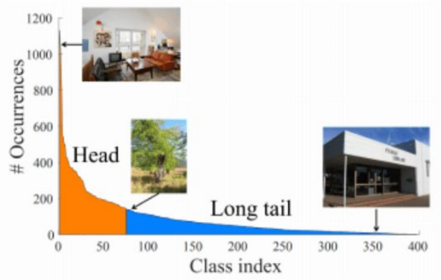
Inferences

- 🟪 cyclist 6b55fec9-bd58-4e4a-a22f-6b556e024685
- 🟪 cyclist 028fb5ec-6bbe-4351-bd98-10d380ac9444
- 🟦 pedestrian 409d7ace-78cd-47f0-9de7-1359845fcab9
- 🟥 car 2ade8b52-6154-4a4e-9211-204927885705
- 🟦 pedestrian c7137806-4564-4cbc-aa68-43a675f1ea48

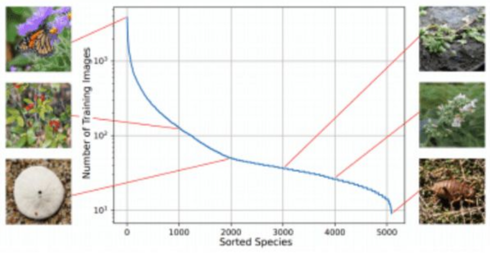
Long Tail Is Long



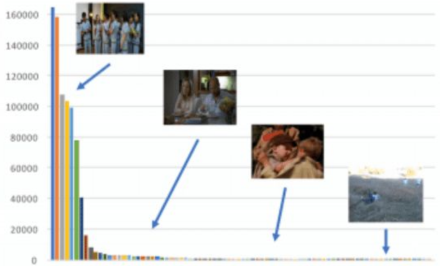
Faces [Zhang et al. 2017]



Places [Wang et al. 2017]



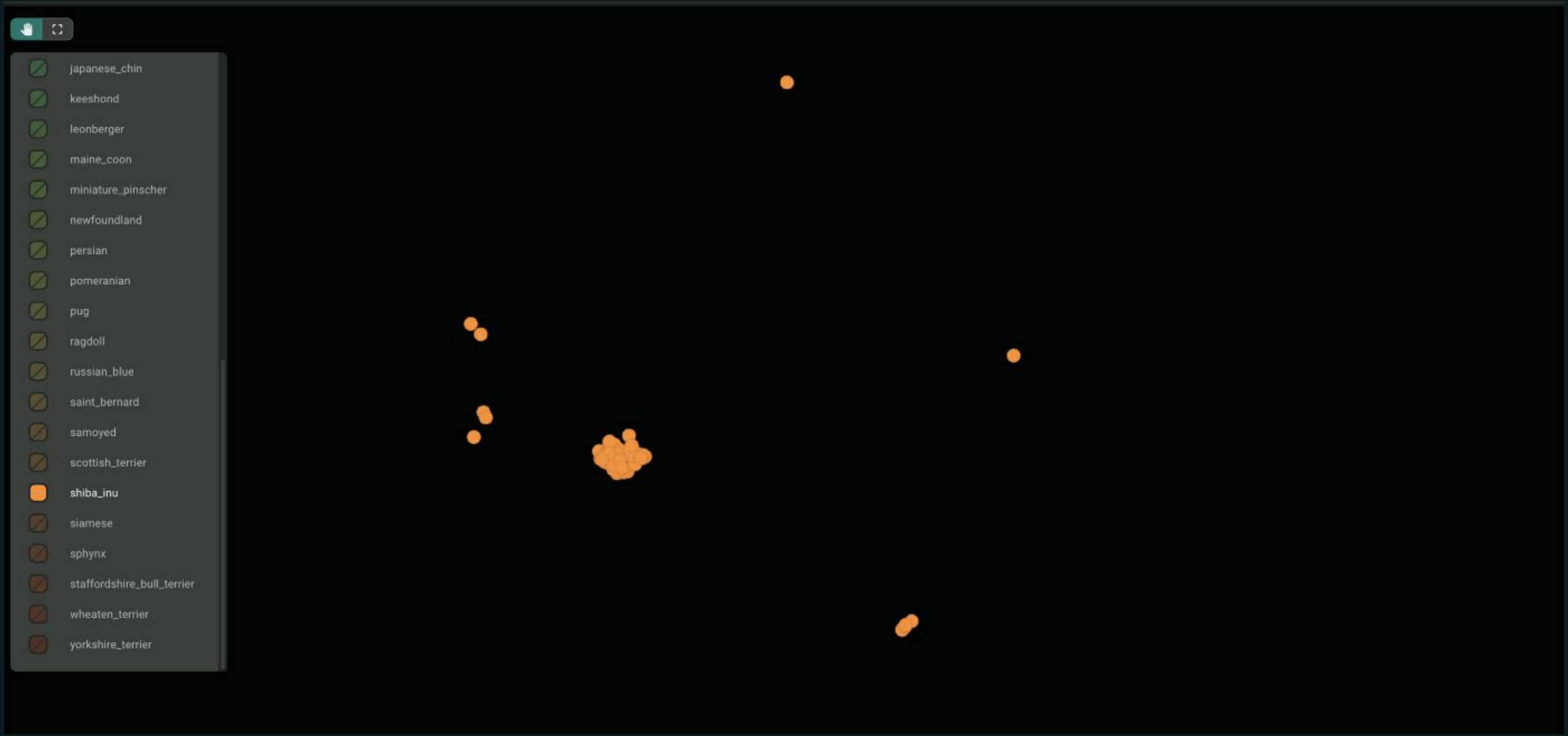
Species [Van Horn et al. 2019]



Actions [Zhang et al. 2019]

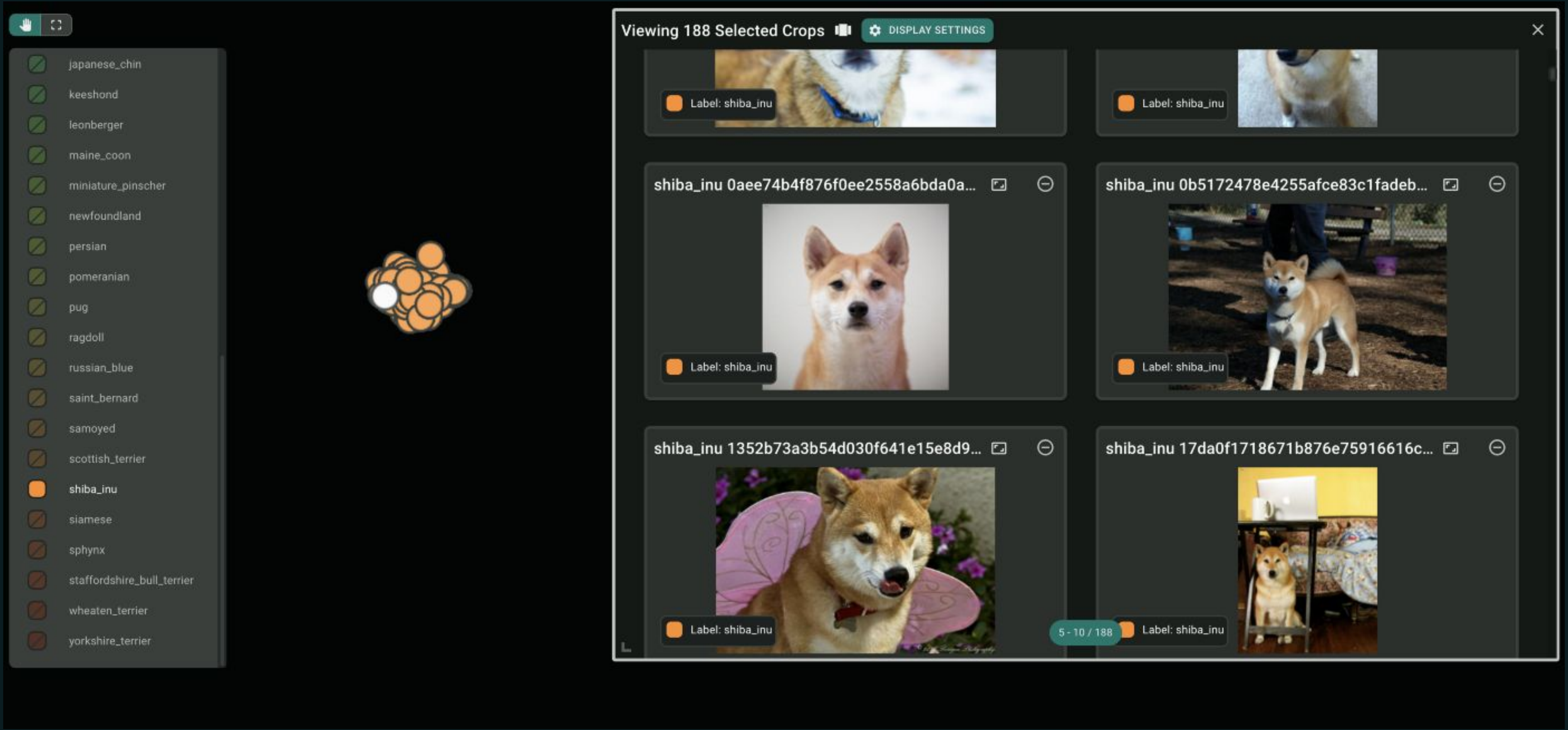
Source: Z. Liu, Z. Miao, et al

Example: Oxford IIT Pets



- ▨ japanese_chin
- ▨ keeshond
- ▨ leonberger
- ▨ maine_coon
- ▨ miniature_pinscher
- ▨ newfoundland
- ▨ persian
- ▨ pomeranian
- ▨ pug
- ▨ ragdoll
- ▨ russian_blue
- ▨ saint_bernard
- ▨ samoyed
- ▨ scottish_terrier
- 🟧 shiba_inu
- ▨ siamese
- ▨ sphynx
- ▨ staffordshire_bull_terrier
- ▨ wheaten_terrier
- ▨ yorkshire_terrier

Example: Oxford IIT Pets



Viewing 188 Selected Crops

Label: shiba_inu

Label: shiba_inu

shiba_inu 0aee74b4f876f0ee2558a6bda0a...

Label: shiba_inu

shiba_inu 0b5172478e4255afce83c1fadeb...

Label: shiba_inu

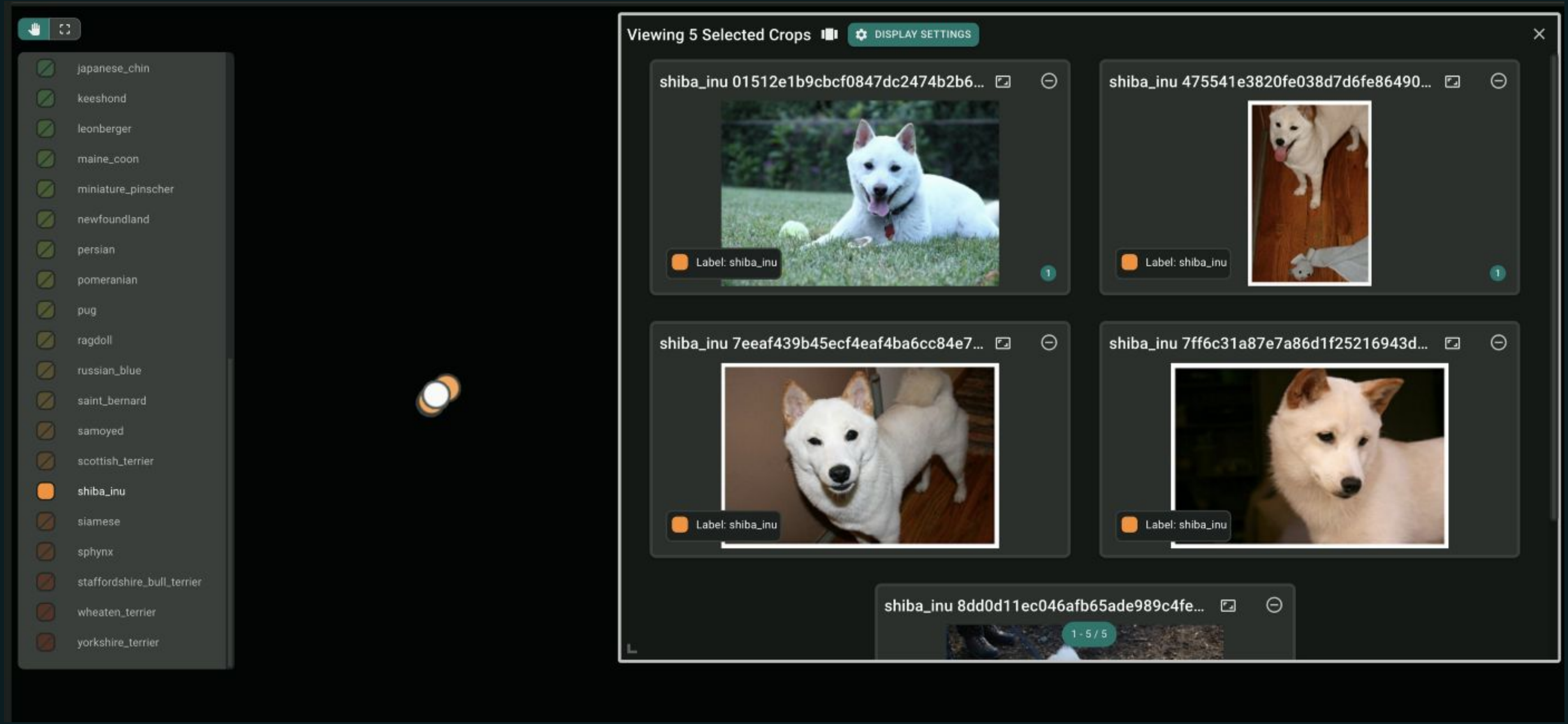
shiba_inu 1352b73a3b54d030f641e15e8d9...

Label: shiba_inu

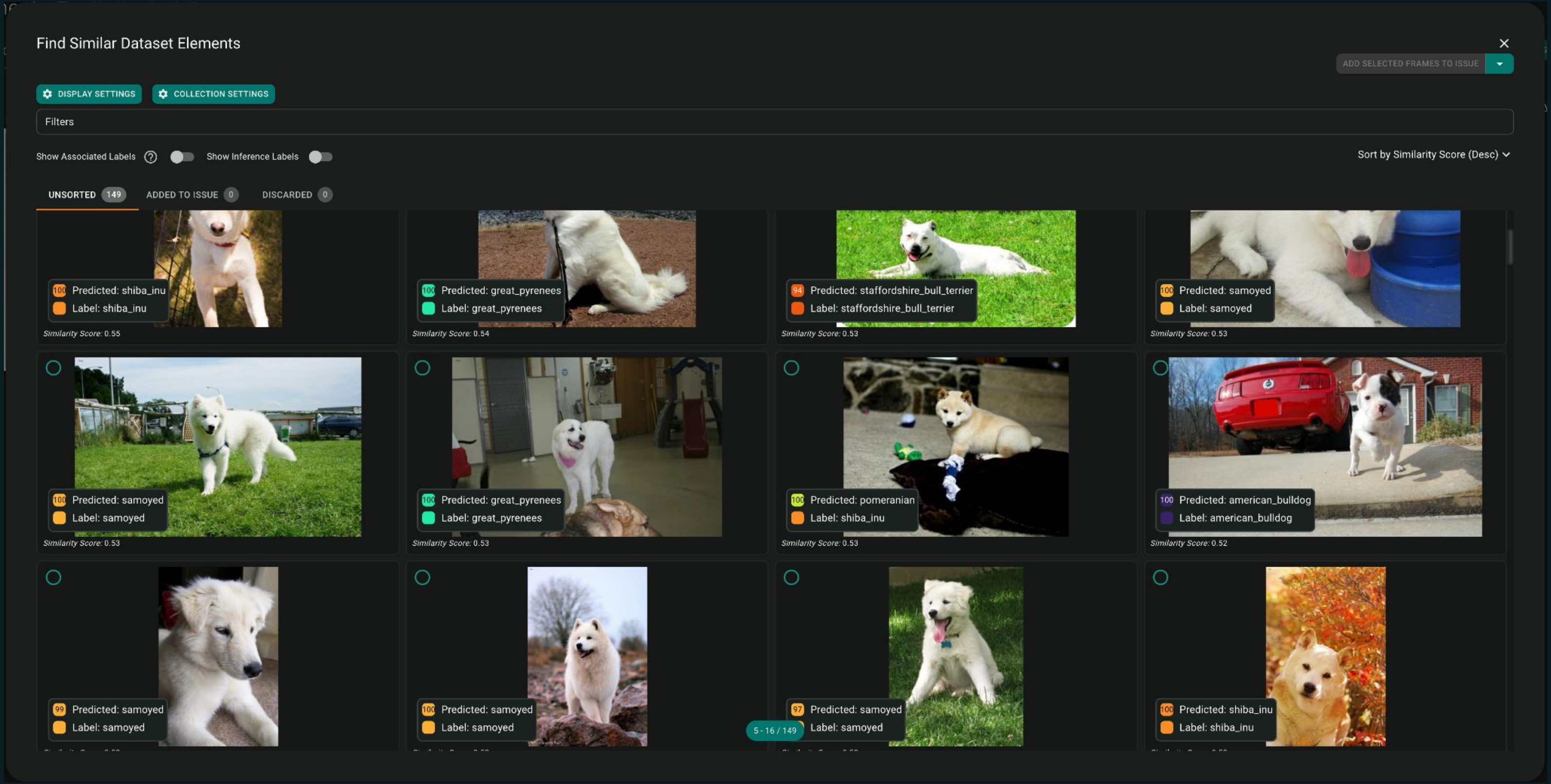
shiba_inu 17da0f1718671b876e75916616c...

Label: shiba_inu

Example: Oxford IIT Pets



Example: Oxford IIT Pets



Find Similar Dataset Elements

ADD SELECTED FRAMES TO ISSUE

DISPLAY SETTINGS

COLLECTION SETTINGS

Filters

Show Associated Labels

Show Inference Labels

Sort by Similarity Score (Desc)

UNSORTED 149

ADDED TO ISSUE 0

DISCARDED 0

100 Predicted: shiba_inu
Label: shiba_inu
Similarity Score: 0.55

100 Predicted: great_pyrenees
Label: great_pyrenees
Similarity Score: 0.54

94 Predicted: staffordshire_bull_terrier
Label: staffordshire_bull_terrier
Similarity Score: 0.53

100 Predicted: samoyed
Label: samoyed
Similarity Score: 0.53

○ 100 Predicted: samoyed
Label: samoyed
Similarity Score: 0.53

○ 100 Predicted: great_pyrenees
Label: great_pyrenees
Similarity Score: 0.53

○ 100 Predicted: pomeranian
Label: shiba_inu
Similarity Score: 0.53

○ 100 Predicted: american_bulldog
Label: american_bulldog
Similarity Score: 0.52

○ 99 Predicted: samoyed
Label: samoyed

○ 100 Predicted: samoyed
Label: samoyed

○ 97 Predicted: samoyed
Label: samoyed

○ 100 Predicted: shiba_inu
Label: shiba_inu

5 - 16 / 149

Aquarium

Make it easier to build and improve production ML systems!

Q&A