



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

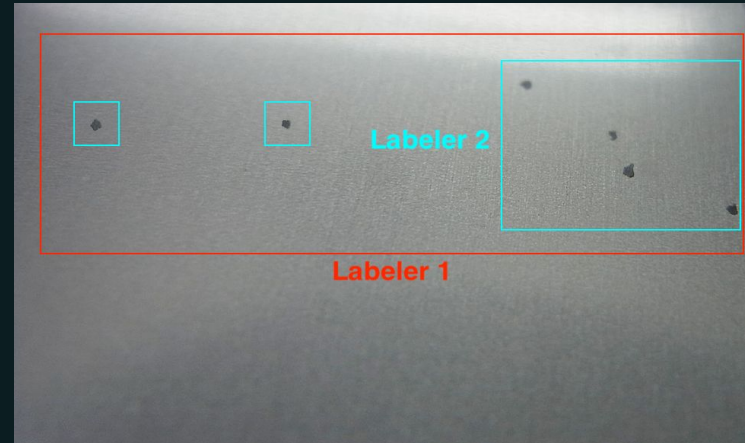
	Old School ML	Deep Learning
Example Tasks	Forecasting, recommendations	Object detection, seq2seq
Data types	Structured / tabular data	Unstructured data (imagery, audio, etc.)
Labeling	Labels come “for free”	Pay people to label data
Algorithms	Logistic regression, SVMs, random forests	Neural networks
Development Emphasis	<ul style="list-style-type: none">- Data pipelines + infrastructure- Feature engineering- Model experimentation (ex: sparsity)	<ul style="list-style-type: none">- Data pipelines + infrastructure- Fine tuning pretrained models- <u>Improving quality + variety of datasets</u>

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



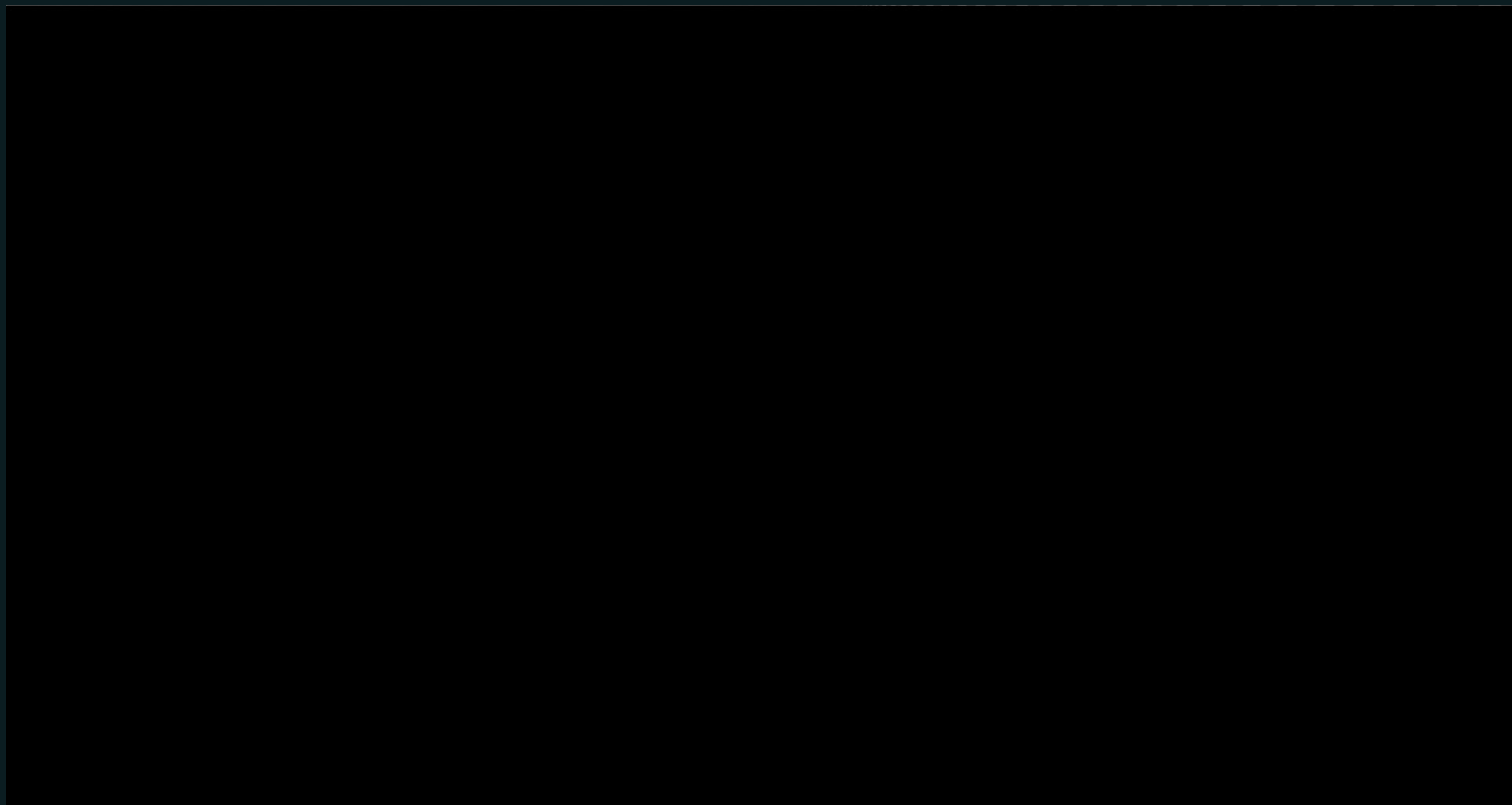
How Do You Improve Datasets Efficiently?

- 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

jinja PROJECTS EXPLORE ISSUES

DOCS kitti lidar INTERNAL DEMO

Image Task 002362 [🔗](#)

Add Tags

DISPLAY SETTINGS

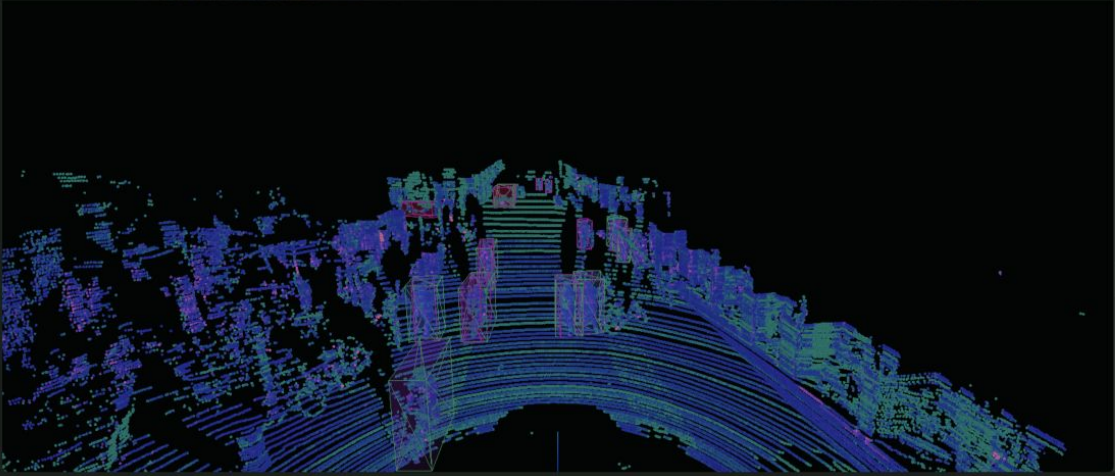



Image Details

Labels

- pedestrian f2baa0d-ba72-4954-bcd4-8e7bb0b545a8
- pedestrian be722cb4-56a6-4caa-98e4-a64f41cc5af3
- pedestrian 93a9c841-6743-4264-a3dc-347c21205e6f
- cyclist 3faef872-4026-4fea-ad35-49284eed565d
- cyclist 6f01f0c5-e604-4de5-a5b7-980f864db1b9
- cyclist c5a52db5-90fb-4cec-afa1-b5cbf1c487ad
- pedestrian d1cddda4-d6e2-414d-902f-9f62258c1c1
- pedestrian 17337367-5142-4db5-a397-0af054c4ef00
- car e8242e6f-6ecd-42c0-ae1f-104de3a74009
- dontcare ba7a15ec-f27f-442e-b018-8482840d15ad
- dontcare 706d6316-2845-4c1c-a483-fbad0d4e512
- dontcare 68a574dd-9307-4e77-9645-e1e07d5f5d2

Inferences

- cyclist 6b55fec9-bd58-4e4a-a22f-6656e024685
- cyclist 028f95ec-6bbe-4351-bd98-10d380ac9444
- pedestrian 409d7ace-78cd-47f0-9de7-1359845fcab9
- car 2ade8b52-6154-4a4e-9211-204927885705
- pedestrian c7137806-4564-4cbc-aa68-43a675f1e648

Example: KITTI

Image Task 002362 [↗](#)

Add Tags

DISPLAY SETTINGS

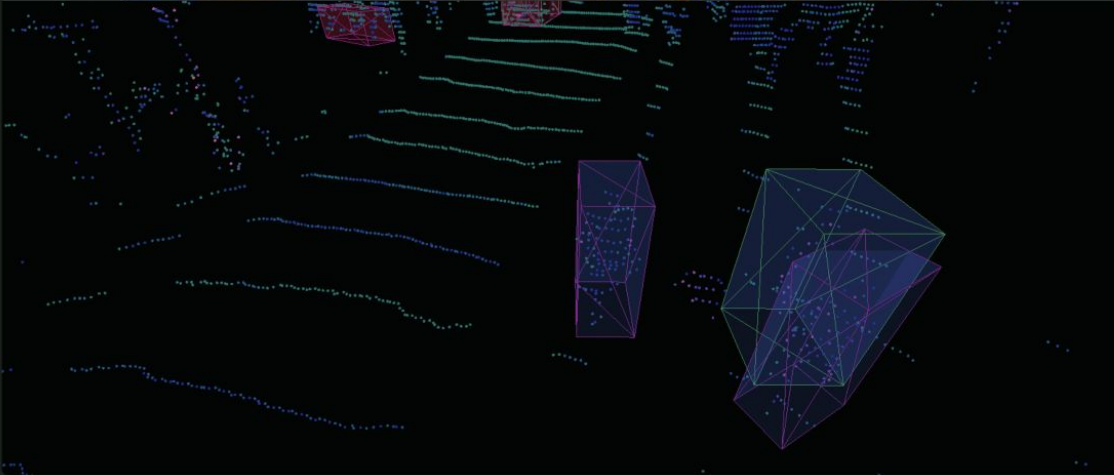



Image Details

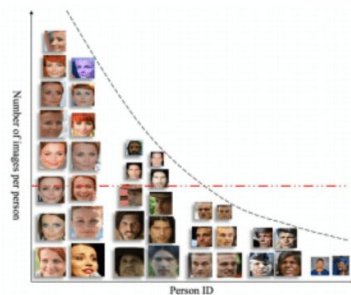
Labels

- pedestrian f2ba0a0d-baf2-4954-bcd4-8e7bb0b545a8
- pedestrian be722cb4-56af-4caa-98e4-a64f1cc5af3
- pedestrian 93a9c841-6743-4264-a3dc-347c721205eef
- cyclist 3faef872-f026-4fe0-ad35-69284eed565d
- cyclist 6f01f0c5-e606-4de6-a5b7-980f864db1b9
- cyclist c5a52db5-90fb-4cec-afa1-b5cbf1c487ad
- pedestrian d1cdd5a4-d6e2-414d-902f-9fd258bc1c1
- pedestrian 17337367-5142-4db5-a397-0af0504ef00
- car e8242eef-6ecd-42c0-ae1f-104de3a74009
- dontcare ba7a15ec-f27f-442e-b018-8482840d15ad
- dontcare 706d631e-2845-4c1c-a483-fbad0d04e512
- dontcare 68a57add-9307-4e77-9645-e1e077d5fd2

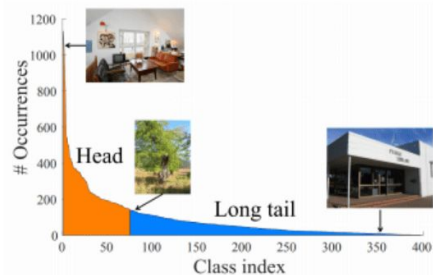
Inferences

- cyclist 6b55fec9-bd58-4e4a-a22f-6b55e024685
- cyclist 028fa5ec-6bbe-4351-b098-10d380ac9444
- pedestrian 409d7ace-78cd-47f0-9de7-1369845fca9
- car 2a8e8b52-6154-494e-9211-204927885705
- pedestrian c7137806-4564-4cbc-aa69-43ae75f1ea48

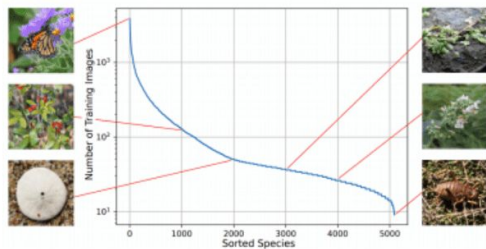
Long Tail Is Long



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



Example: Oxford IIT Pets

The screenshot displays the Oxford IIT Pets web application interface. On the left, a sidebar lists various dog breeds, each with a checkbox and a small circular icon. The 'shiba_inu' breed is selected, indicated by an orange checkbox and a cluster of orange circles. The main area shows a gallery of 188 selected Shiba Inu crops, arranged in a grid. Each crop is displayed with a small thumbnail, a label 'Label: shiba_inu', and a unique ID. The gallery is titled 'Viewing 188 Selected Crops' and includes a 'DISPLAY SETTINGS' button. A '6 - 10 / 188' indicator is visible at the bottom of the gallery.

Breed Selection List:

- 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

Gallery Crops (Examples):

- 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

The screenshot displays the Oxford IIT Pets web application. On the left is a vertical list of dog breeds, each with a checkbox and a small circular indicator. The 'shiba_inu' breed is selected, indicated by an orange circle. On the right, a window titled 'Viewing 5 Selected Crops' shows a grid of five images of Shiba Inu dogs. Each image has a label 'Label: shiba_inu' and a small '1' in a blue circle. The images show a white Shiba Inu lying on grass, a Shiba Inu in a brown coat, a white Shiba Inu standing on a wooden floor, and a close-up of a white Shiba Inu's face. A fifth image is partially visible at the bottom.

Viewing 5 Selected Crops DISPLAY SETTINGS

- 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

shiba_inu 01512e1b9cbcf0847dc2474b2b6... Label: shiba_inu

shiba_inu 475541e3820fe038d7d6fe86490... Label: shiba_inu

shiba_inu 7eeaf439b45ecf4eaf4ba6cc84e7... Label: shiba_inu

shiba_inu 7ff6c31a87e7a86d1f25216943d... Label: shiba_inu

shiba_inu 8dd0d11ec046afb65ade989c4fe... 1 / 5

Example: Oxford IIT Pets

Find Similar Dataset Elements

ADD SELECTED FRAMES TO ISSUE


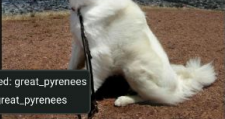







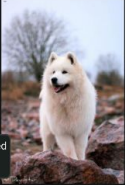
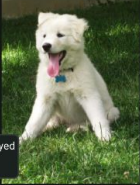

DISPLAY SETTINGS COLLECTION SETTINGS

Filters

Show Associated Labels Show Inference Labels

Sort by Similarity Score (Desc)

UNSORTED 149 ADDED TO ISSUE 0 DISCARDED 0

 <p>100 Predicted: shiba_inu Label: shiba_inu</p> <p>Similarity Score: 0.55</p>	 <p>100 Predicted: great_pyrenees Label: great_pyrenees</p> <p>Similarity Score: 0.54</p>	 <p>88 Predicted: staffordshire_bull_terrier Label: staffordshire_bull_terrier</p> <p>Similarity Score: 0.53</p>	 <p>100 Predicted: samoyed Label: samoyed</p> <p>Similarity Score: 0.53</p>
 <p>100 Predicted: samoyed Label: samoyed</p> <p>Similarity Score: 0.53</p>	 <p>100 Predicted: great_pyrenees Label: great_pyrenees</p> <p>Similarity Score: 0.53</p>	 <p>100 Predicted: pomeranian Label: shiba_inu</p> <p>Similarity Score: 0.53</p>	 <p>100 Predicted: american_bulldog Label: american_bulldog</p> <p>Similarity Score: 0.52</p>
 <p>99 Predicted: samoyed Label: samoyed</p>	 <p>100 Predicted: samoyed Label: samoyed</p>	 <p>97 Predicted: samoyed Label: samoyed</p>	 <p>100 Predicted: shiba_inu Label: shiba_inu</p>

5-16 / 149



Make it easier to build and improve production ML systems!

Q&A