Machine Learning at Scale

Amy Unruh
Google
@amygdala

AI NEXTCon, 18th Jan 2018

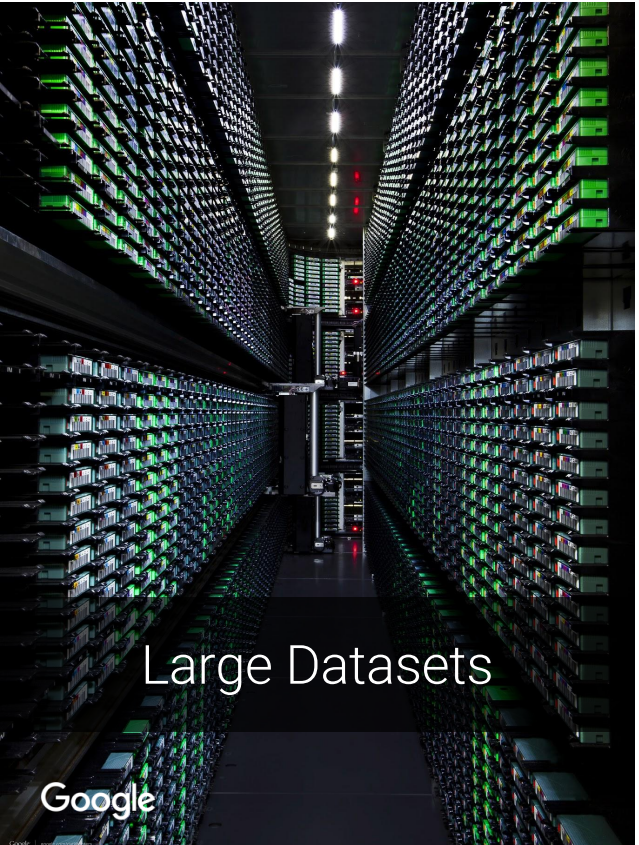
Google Cloud

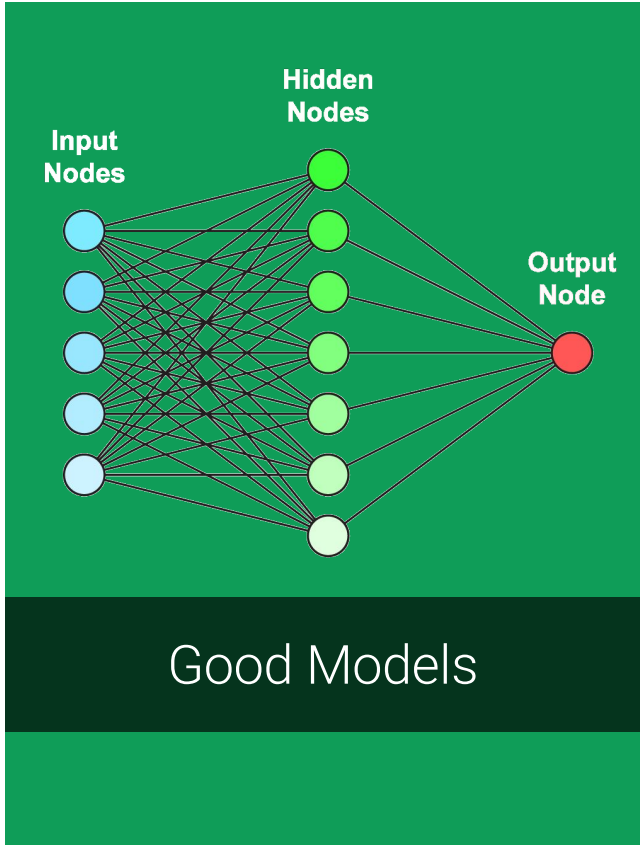
What this talk is about…

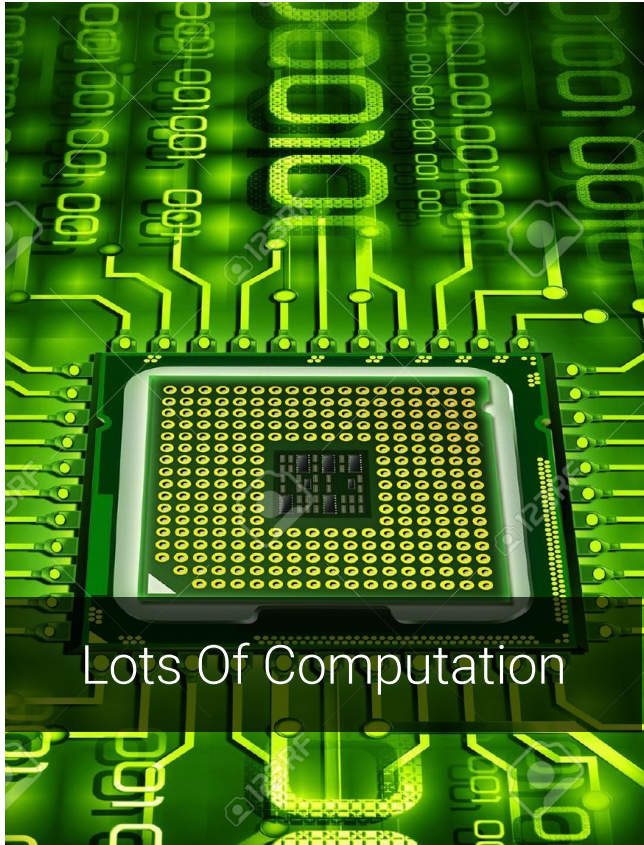
(Cloud AutoML Vision!)

ML on Google Cloud Platform (GCP)

The Challenges of ML







The Datacenter as a Computer

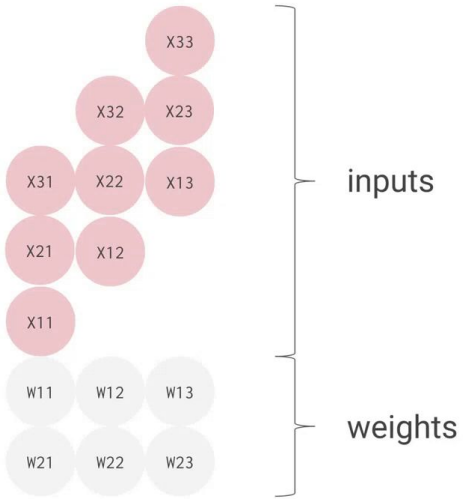
Google

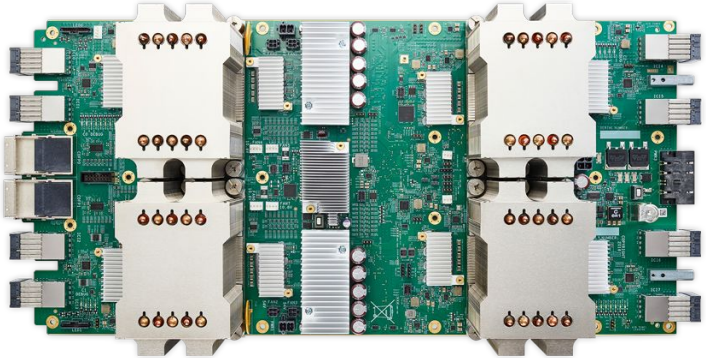
Jupiter network

10 GbE x 100 K = 1 Pbps

Connect servers with

microsec latency





2nd generation

Tensor Processing Unit

ASIC for TensorFlow

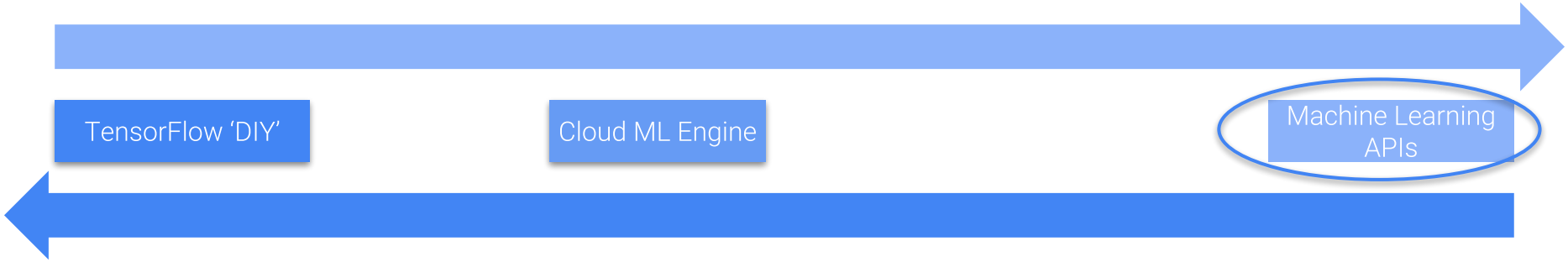
Designed by Google

180 Tflops / chip

11.5 Pflops / pod

The GCP ML Spectrum :: Technologies

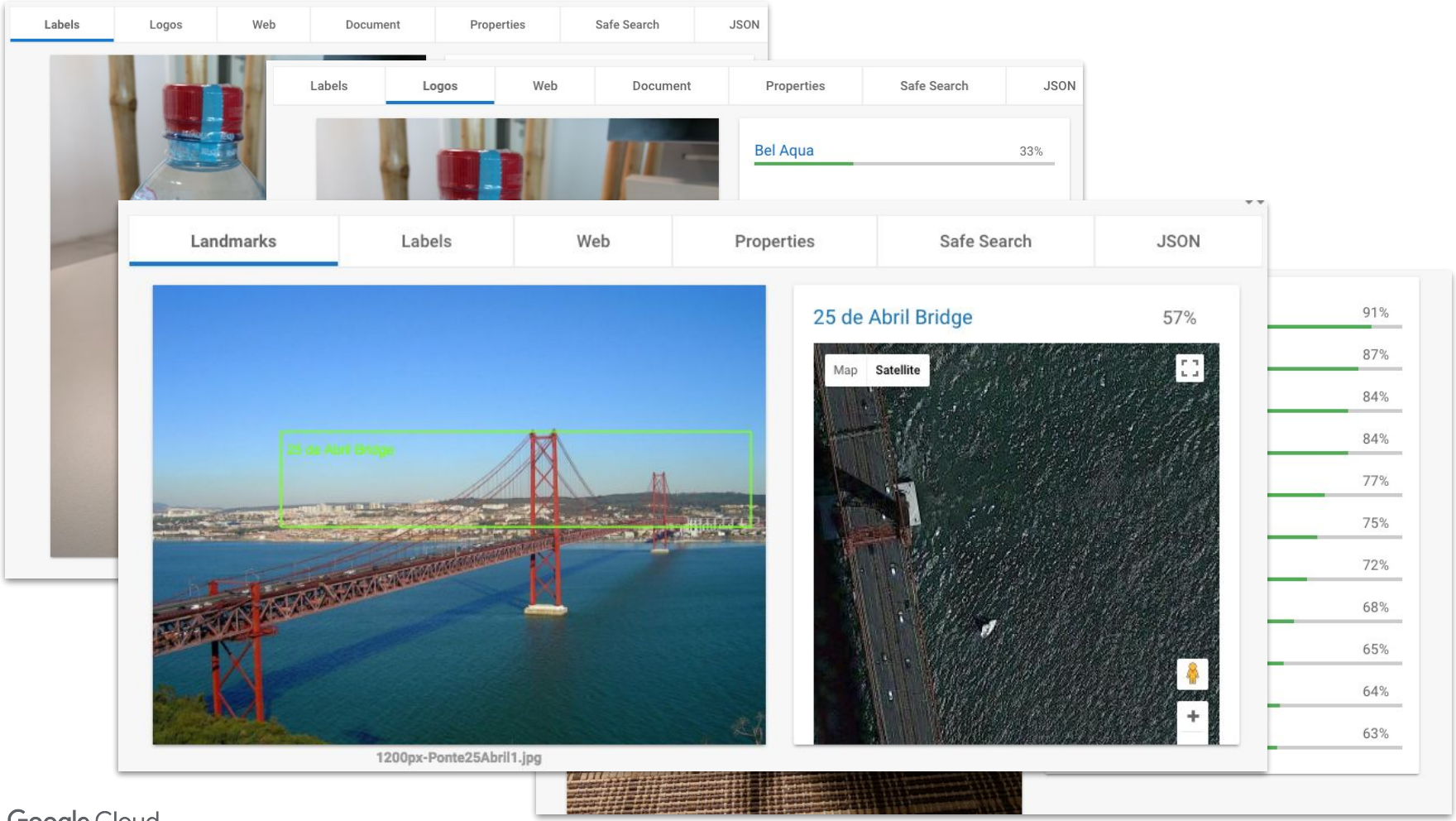
As it's looked until recently...



Cloud ML APIs

https://cloud.google.com/products/machine-learning/

(Visit their landing pages to try them out)



Google Cloud

mini-demo (Vision API)

Aerial Photography	96%
Water Resources	90%
Water	89%
Photography	83%
Bird's Eye View	77%
River Island	76%
Reservoir	74%
Estuary	72%
River	67%

scarps and steep slopes

Aerial Photography	94%
Water Resources	93%
Photography	78%
Atmosphere Of Earth	78%
Bird's Eye View	75%
Estuary	74%
Reservoir	64%
Inlet	62%
Floodplain	62%

exposed riprap* structures

Water Resources	93%
Floodplain	87%
Waterway	84%
Estuary	82%
Aerial Photography	81%
Land Lot	78%
River Island	77%
Bird's Eye View	71%
Ecoregion	69%

freshwater swamps and woody vegetation

*loose stone used to form a foundation for a breakwater or other structure.

Dataset courtesy of Texas A&M University. See https://storage.googleapis.com/tamucc_coastline/GooglePermissionForImages_20170119.pdf for details.

Google Cloud

hmmm… maybe I need a model that’s customized for my problem domain..

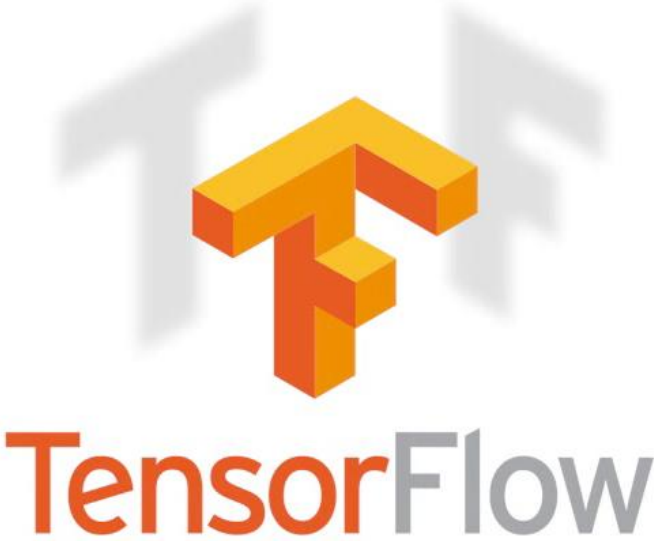
One route — build and train a TensorFlow model

Powered by Open Source

-Created by Google Brain team
-**Fast, flexible and production-ready framework** that scales from research to production.

Multiple deployment options:

- Mobile, Desktop, Server, Cloud
- CPU, GPU, Raspberry Pi!



https://github.com/tensorflow

TensorFlow and the **Open-Source Community**

Positive Reviews

85,000+
GitHub Stars

25,000+
GitHub repositories with 'TensorFlow' in the title

Rapid Development

1,000+
Contributors

27,000+
Commits

Direct Engagement

11,000+
Stack Overflow questions answered

100+
Community-submitted GitHub issues responded to weekly

The Google Brain Team — Looking Back on 2017 (Part 1 of 2)

The Google Brain Team — Looking Back on 2017 (Part 2 of 2)

Eager Execution: An imperative, define-by-run interface to TensorFlow

https://opensource.com/article/17/11/intro-tensorflow

https://github.com/tensorflow/models

+ model weights for **transfer learning**

The Google Brain Team — Looking Back on 2017 (Part 1 of 2)

The Google Brain Team — Looking Back on 2017 (Part 2 of 2)

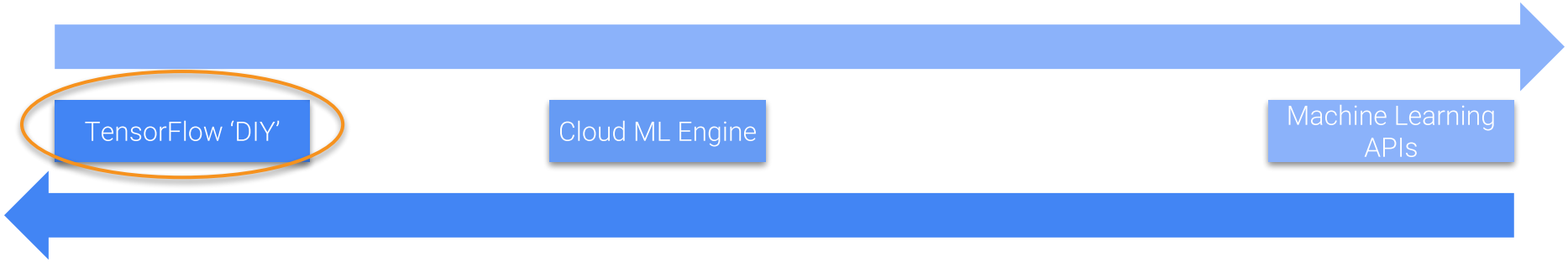
Eager Execution: An imperative, define-by-run interface to TensorFlow

https://opensource.com/article/17/11/intro-tensorflow

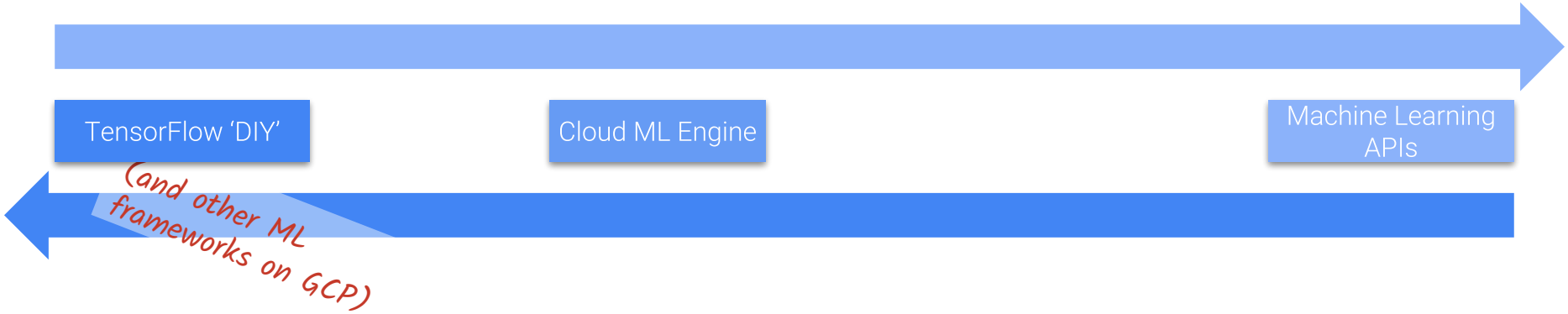
So… how to scale out TensorFlow?

The GCP ML Spectrum :: Technologies

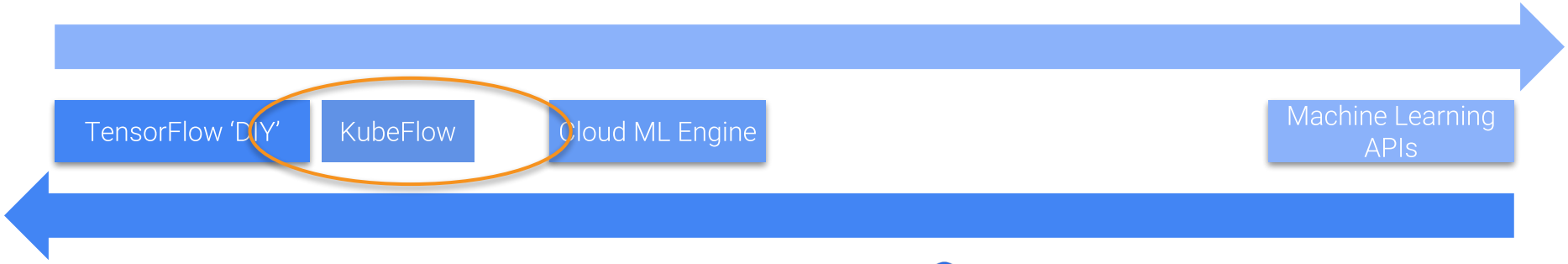
As it's looked until recently...



The GCP ML Spectrum :: Technologies

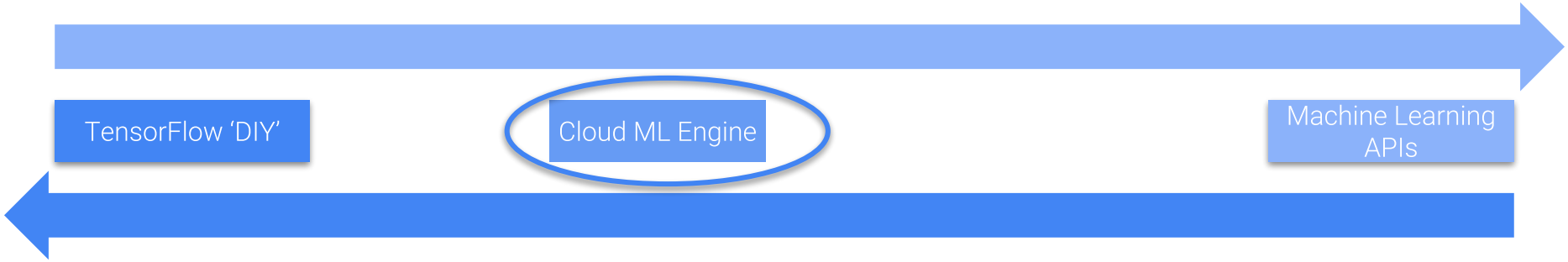


The GCP ML Spectrum :: Technologies



[Introducing Kubeflow - A Composable, Portable, Scalable ML Stack Built for Kubernetes]

The ML Spectrum :: Technologies



ML at Google scale:

Large scale training and prediction with **Cloud ML Engine**

Cloud Machine Learning Engine (**ML Engine**)

Fully managed platform for TensorFlow

Distributed Training with GPUs (and TPUs)

Fast and **scalable** online/batch **prediction**

Hyperparameter tuning using **HyperTune**

+ Online prediction alpha for scikit-learn + XGBoost: https://goo.gl/HabX5z

cloud.google.com/ml-engine

What resources do you need to solve an ML problem?

Training data

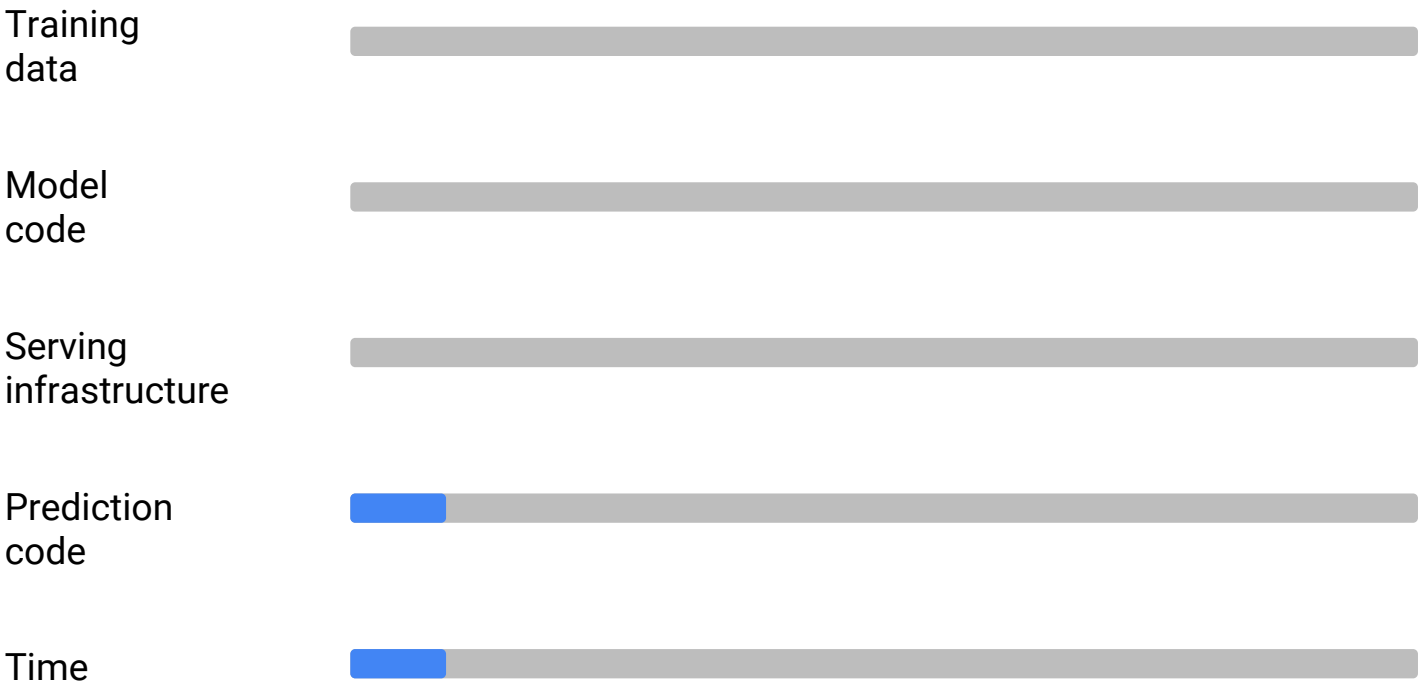
Model code

Serving infrastructure

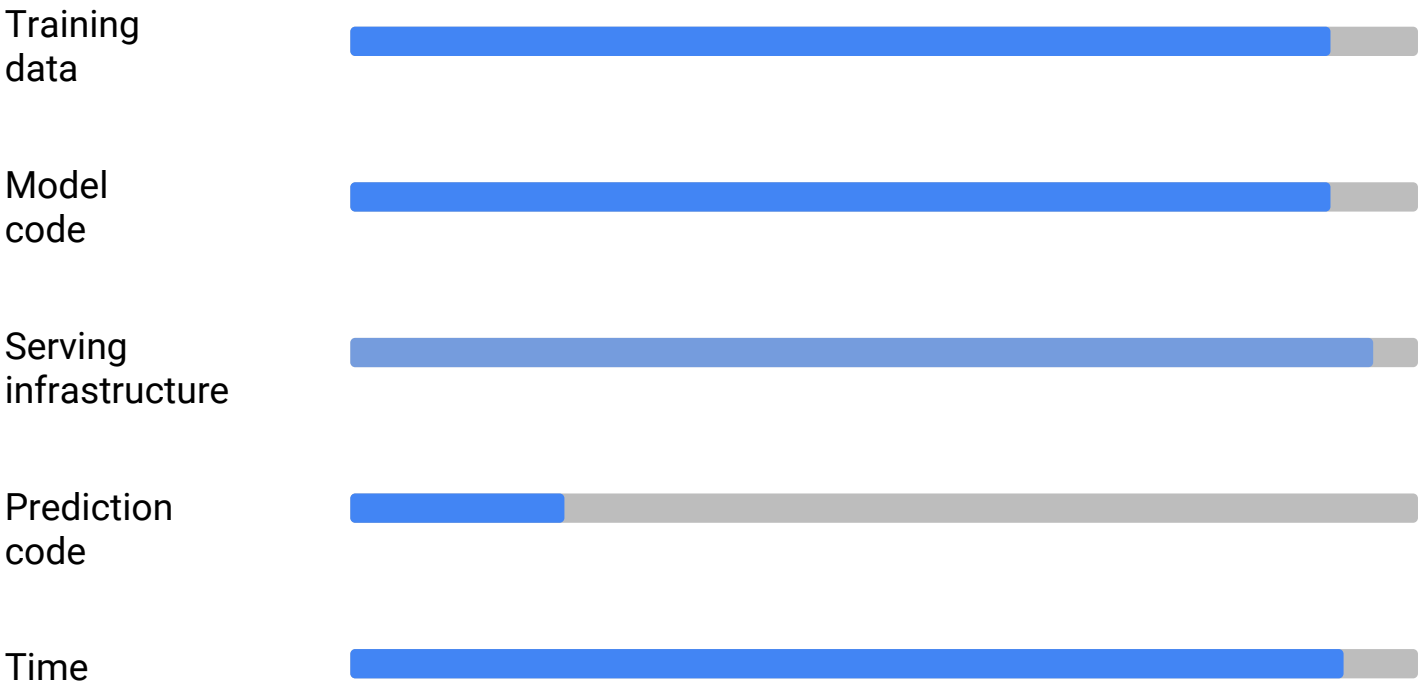
Prediction code

Time

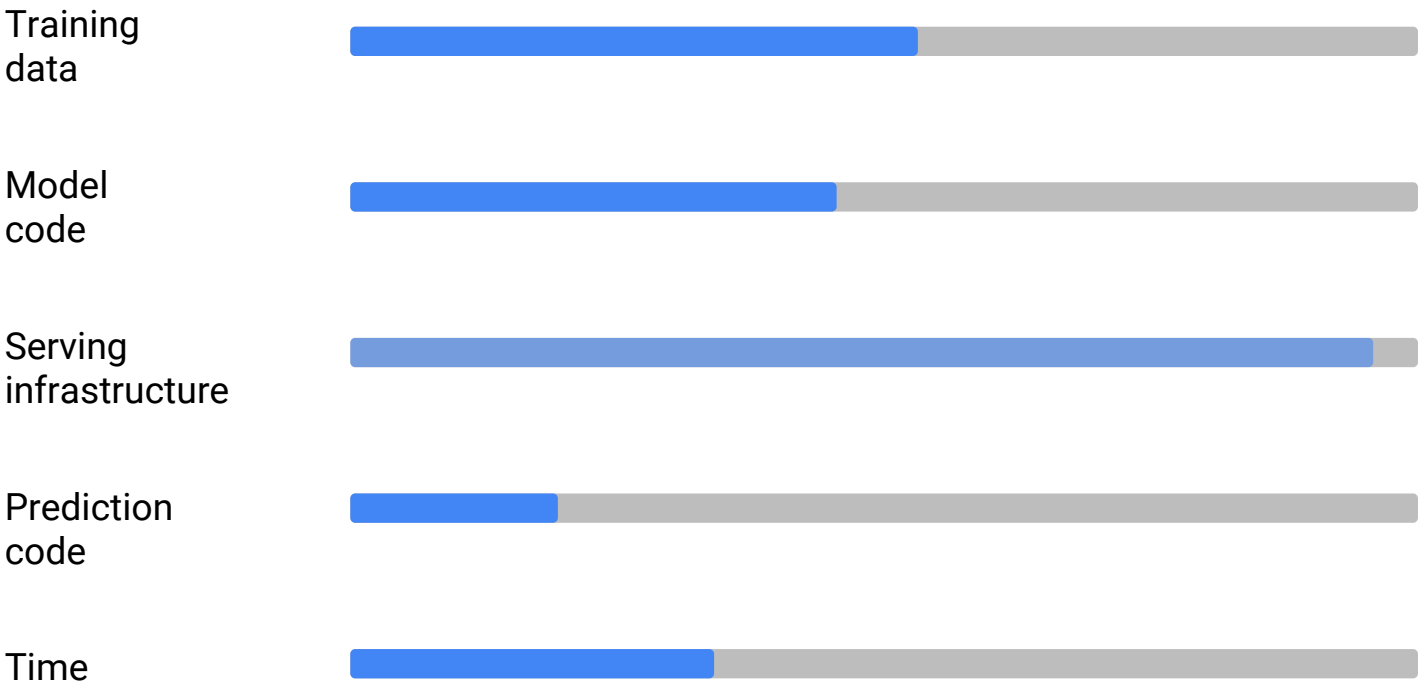
Machine Learning as an API



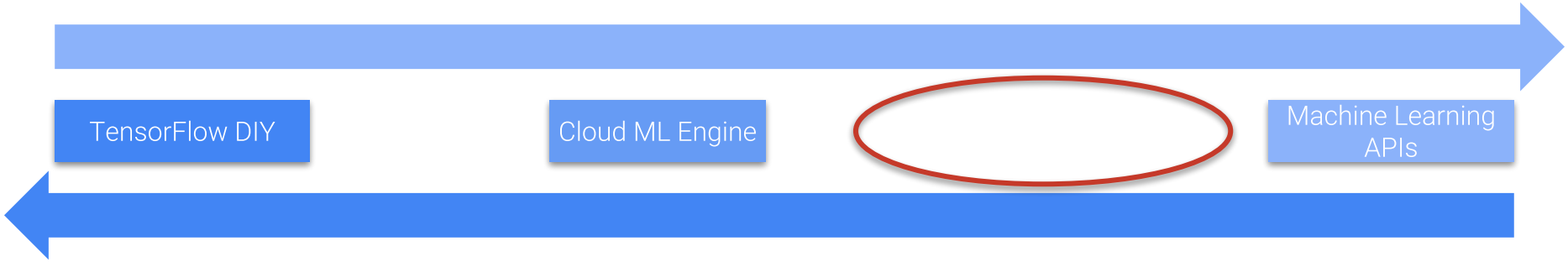
Custom model: build from scratch



Custom model: transfer learning

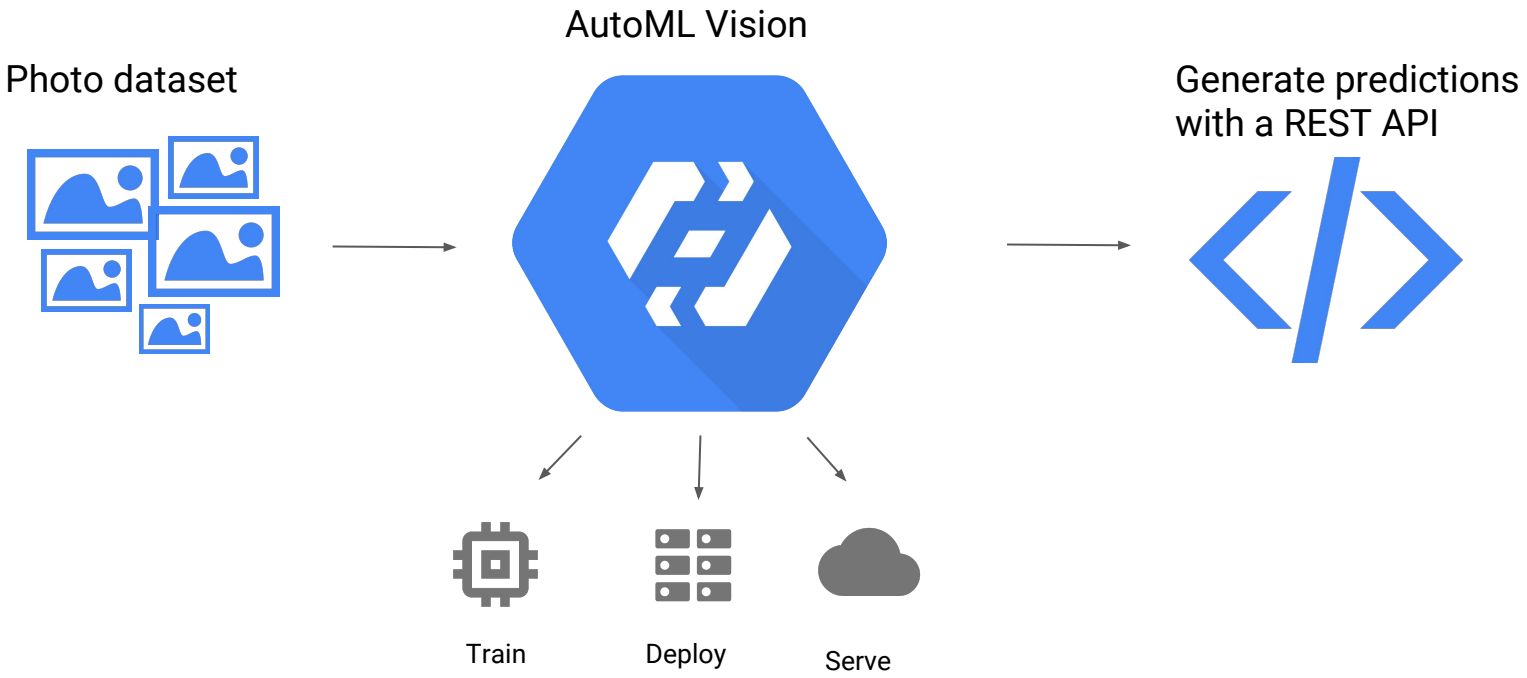


The GCP ML Spectrum :: Technologies

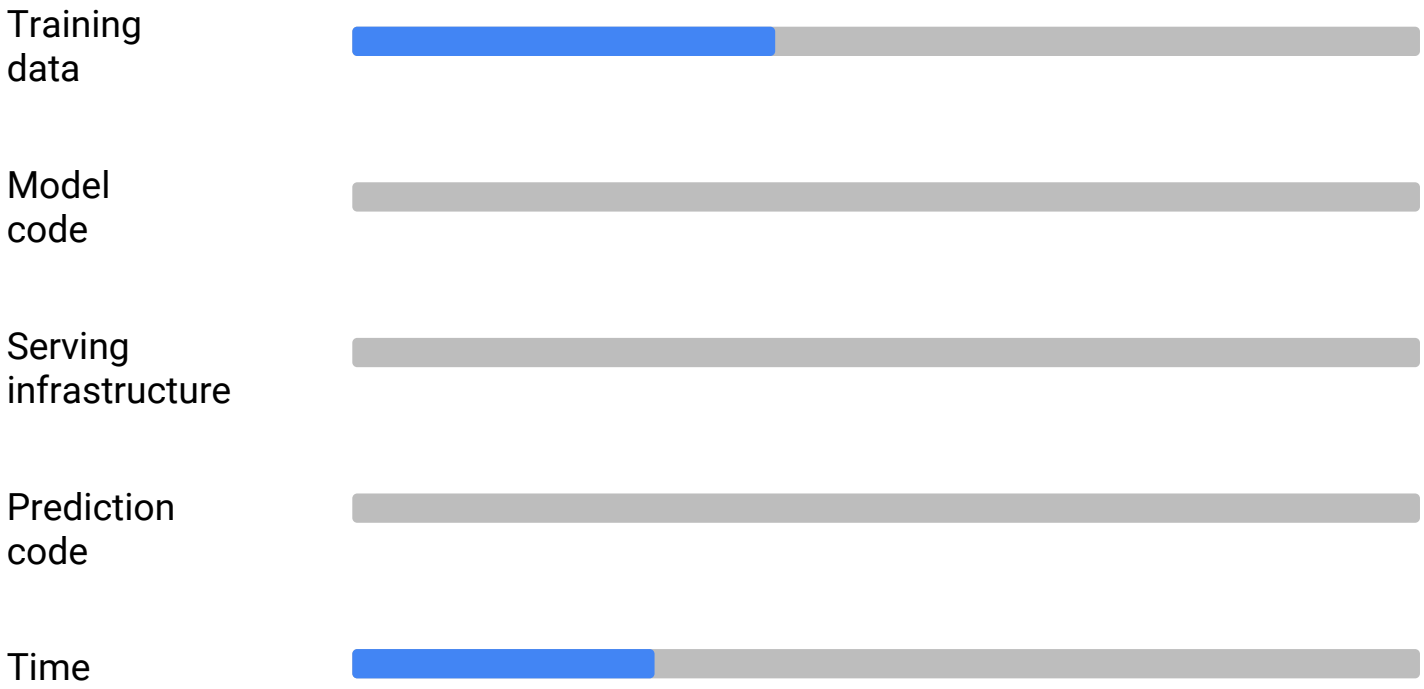


Introducing Cloud AutoML Vision

AutoML to the rescue



Cloud AutoML Vision

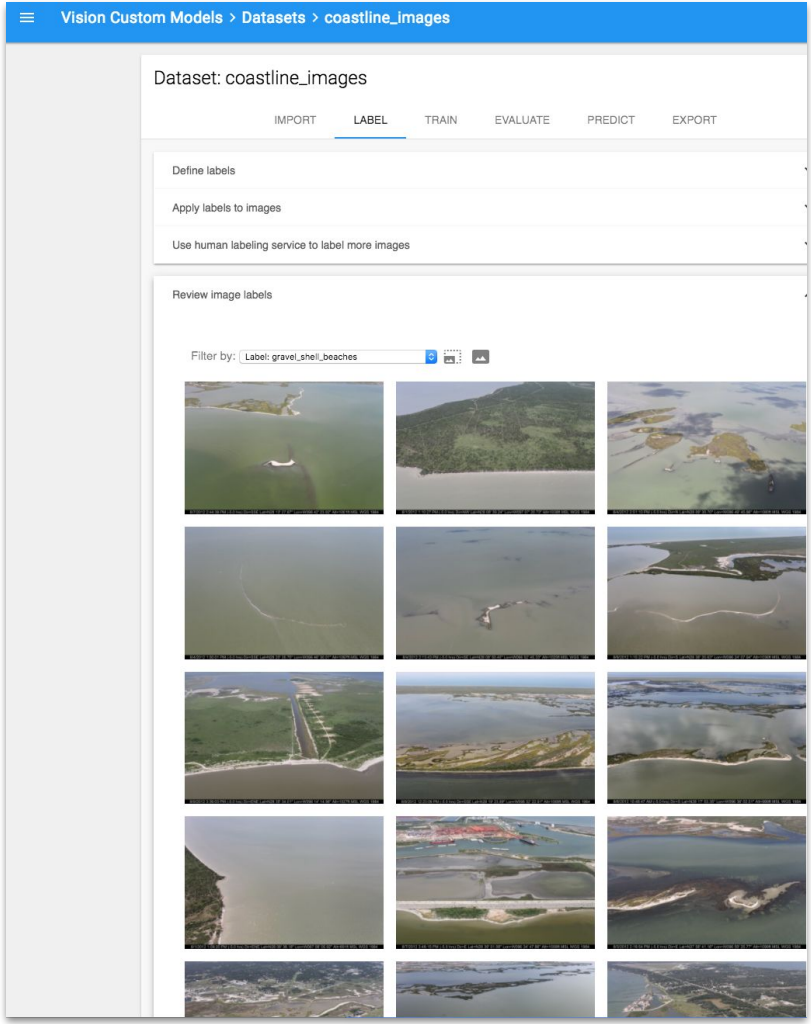


The ML Spectrum :: Technologies

Demo!

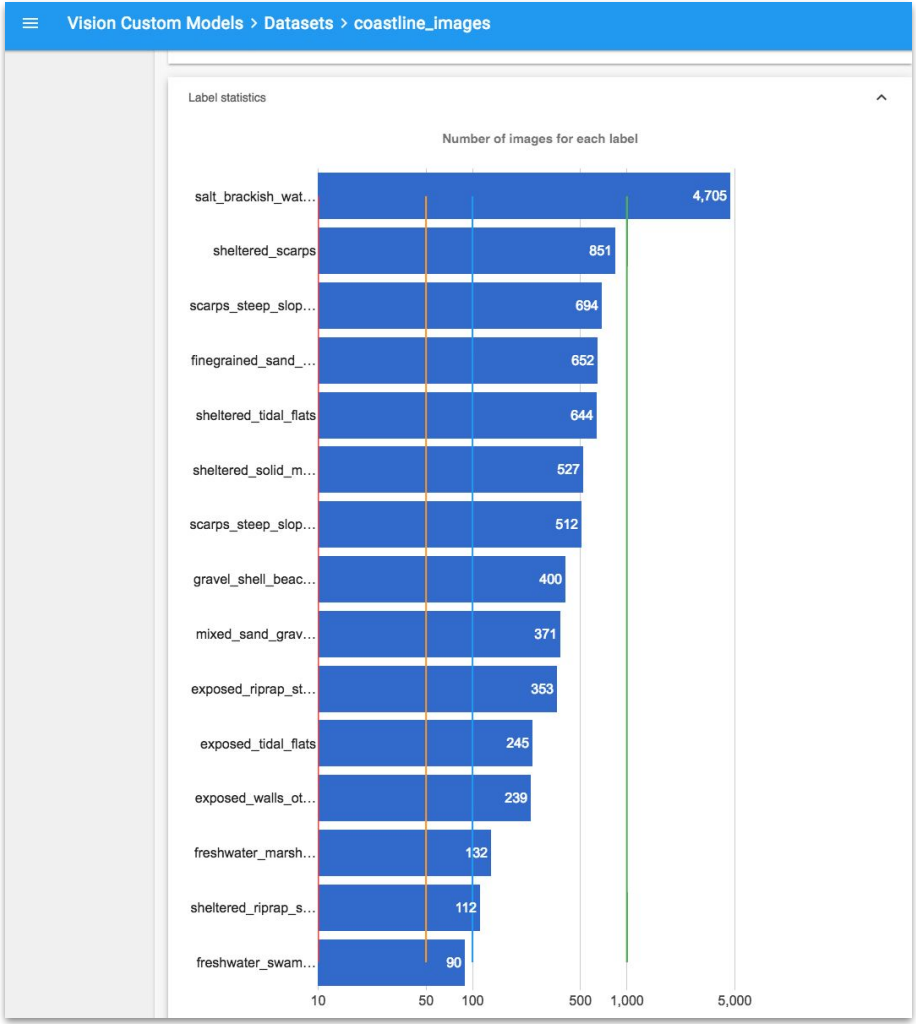
(The dataset used is courtesy of Texas A&M University. See https://storage.googleapis.com/tamucc_coastline/GooglePermissionForImages_20170119.pdf for details).

Preview image labels

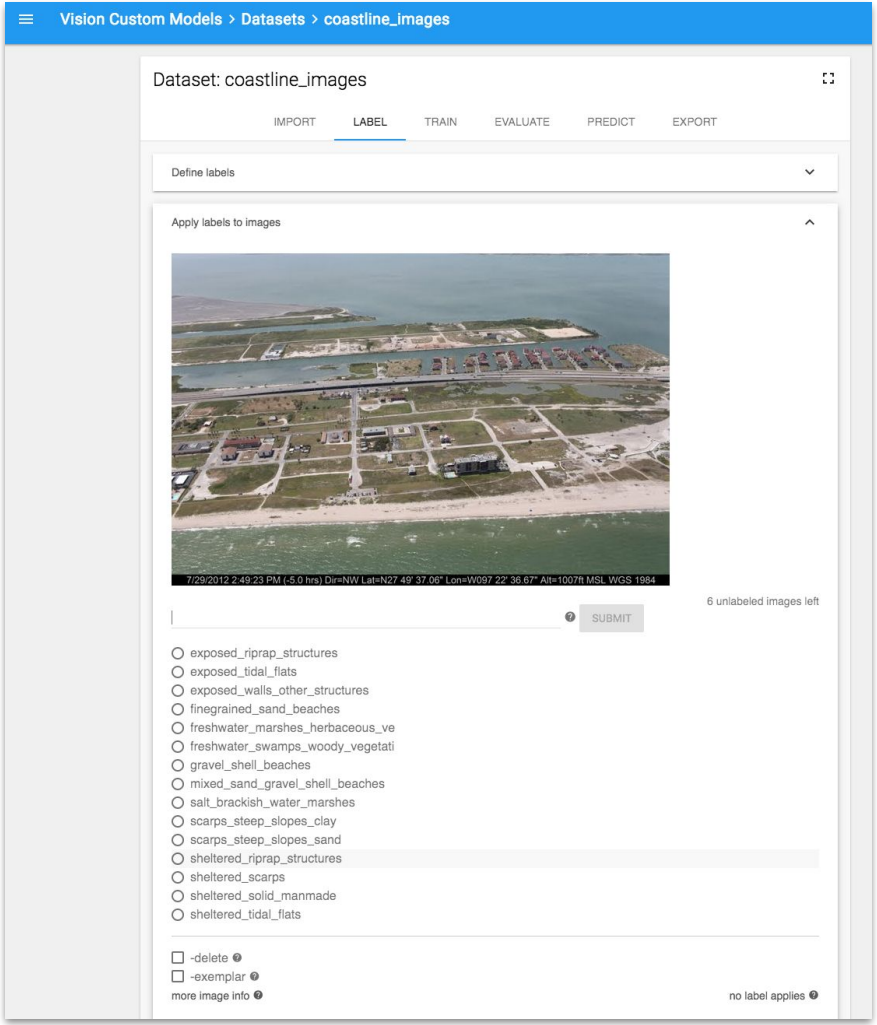


Google Cloud

Label statistics

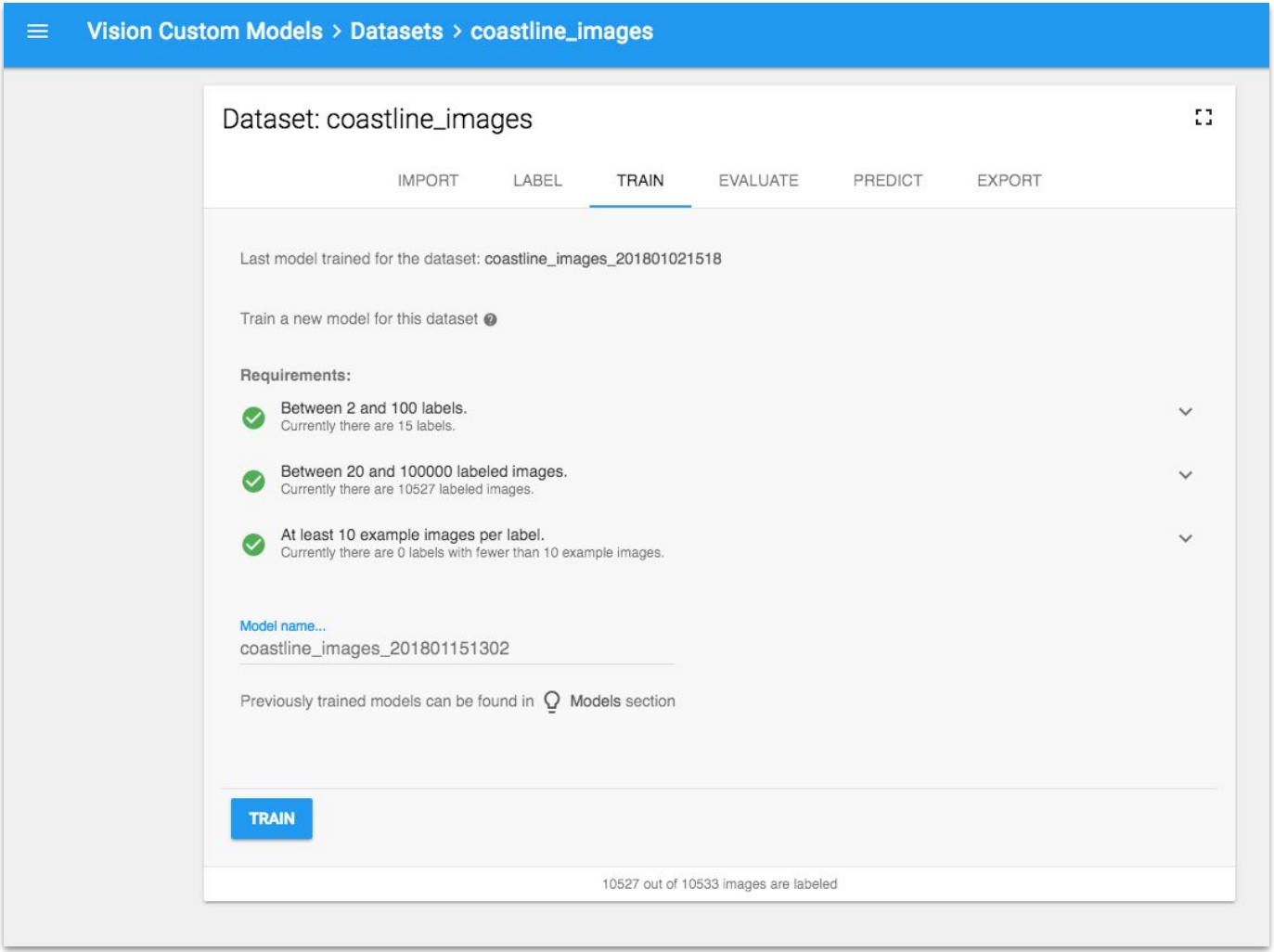


Apply labels to images



Google Cloud

Train a model for a dataset



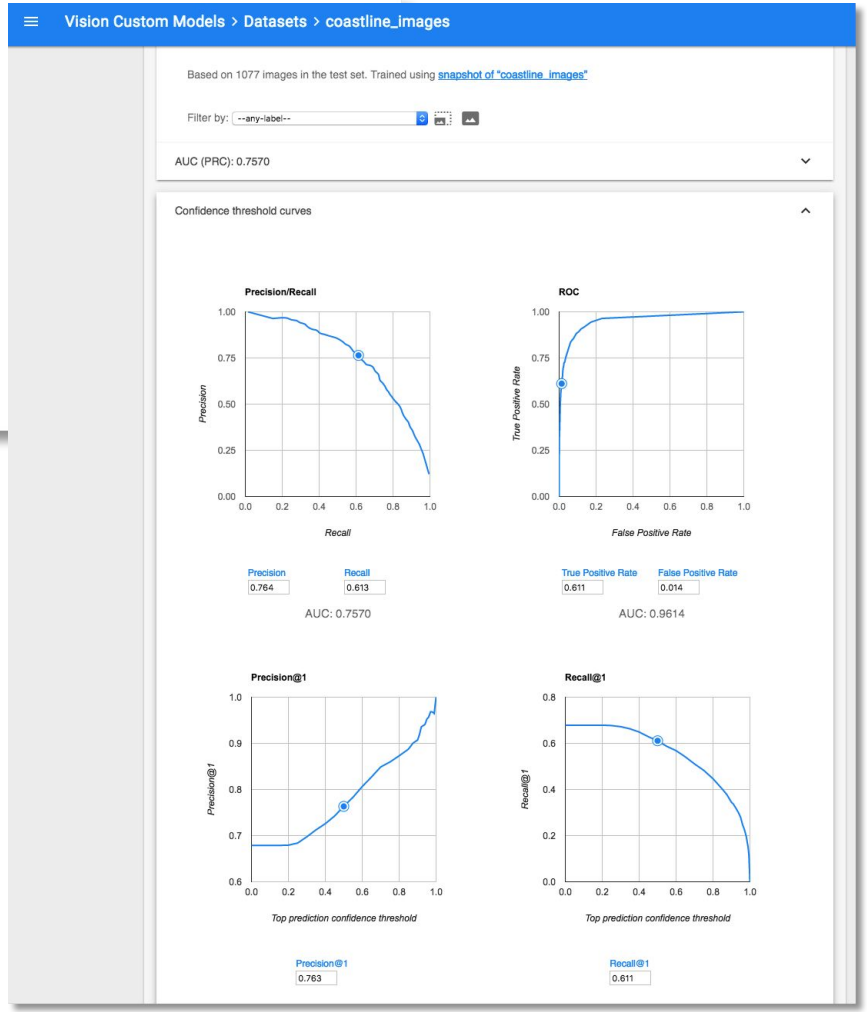
Google Cloud

AUC (PRC): 0.7570

Area Under the Precision/Recall curve.
Higher value is better. 1.0 is perfect, 0.5 is effectively random.
Current value: 0.7570

Area Under the Receiver Operating Characteristic curve.
Higher value is better. 1.0 is perfect, 0.5 is effectively random.
Current value: 0.9614

Precision@1. Rate of top scored label being correct
Higher value is better. 1.0 is perfect.
Current value: 0.6787



Confidence threshold curves and AUC stats

Google Cloud

Confusion matrix

Predicted label

True Label	sheltered_scarps	salt_brackish_water_marshes	scarps_steep_slopes_clay	sheltered_solid_manmade	sheltered_riprap_structures	scarps_steep_slopes_sand	sheltered_tidal_flats	mixed_sand_gravel_shell_beaches	gravel_shell_beaches	freshwater_swamps_woody_vegetati
sheltered_scarps	40%	33%	13%	4%	2%	0%	0%	0%	0%	0%
salt_brackish_water_marshes	1%	88%	2%	0%	0%	0%	3%	0%	1%	0%
scarps_steep_slopes_clay	6%	24%	63%	0%	0%	0%	4%	0%	0%	1%
sheltered_solid_manmade	7%	4%	2%	80%	4%	0%	0%	0%	0%	0%
sheltered_riprap_structures	0%	62%	0%	25%	12%	0%	0%	0%	0%	0%
scarps_steep_slopes_sand	7%	26%	2%	0%	0%	59%	2%	2%	0%	0%
sheltered_tidal_flats	0%	29%	2%	0%	0%	2%	65%	0%	0%	0%
mixed_sand_gravel_shell_beaches	2%	22%	7%	2%	0%	5%	0%	57%	2%	0%
gravel_shell_beaches	0%	42%	2%	2%	0%	0%	0%	2%	50%	0%
freshwater_swamps_woody_vegetati	16%	16%	0%	0%	0%	0%	0%	0%	0%	66%

Vision Custom Models > Datasets > coastline_images

Dataset: coastline_images

IMPORT LABEL TRAIN EVALUATE PREDICT EXPORT

Use model: coastline_images_201801021518

Query online

Select up to 10 images to make predictions on.

Choose Files salt_brackish...marshes1.jpg

PREDICT

Query Cloud ML Engine from the command line

Query Cloud ML Engine from Python

Query Vision API

10527 out of 10533 images are labeled

Vision Custom Models > Datasets > coastline_images

Dataset: coastline_images

IMPORT LABEL TRAIN EVALUATE PREDICT EXPORT

Query online

0.862
salt_brackish_water_marshes

Select up to 10 images to make predictions on.

Choose Files No file chosen

PREDICT

Query Cloud ML Engine from the command line

Query Cloud ML Engine from Python

Query Vision API

10527 out of 10533 images are labeled

Google Cloud

Top scored images

Display “gravel_shell_beaches”: Top scored

Google Cloud

salt_brackish_water_marshes

SUBMIT

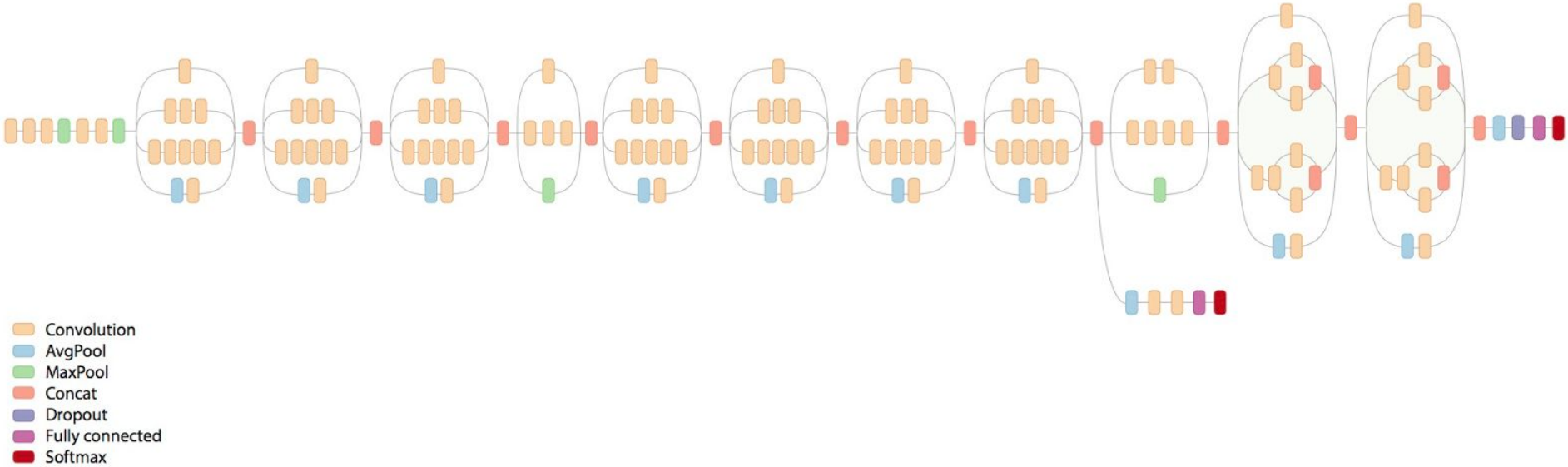
- [ ] exposed_riprap_structures
- [ ] exposed_tidal_flats
- [ ] exposed_walls_other_structures
- [ ] finegrained_sand_beaches
- [ ] freshwater_marshes_herbaceous_ve
- [ ] freshwater_swamps_woody_vegetati
- [ ] gravel_shell_beaches (0.88)
- [ ] mixed_sand_gravel_shell_beaches
- [x] salt_brackish_water_marshes (0.29)

Google Cloud

(Hmm… should this be a multi-labeled dataset?..)

Under the hood: Transfer learning and Neural Architecture Search

Transfer learning:
a powerful technique to build highly accurate models even with limited data



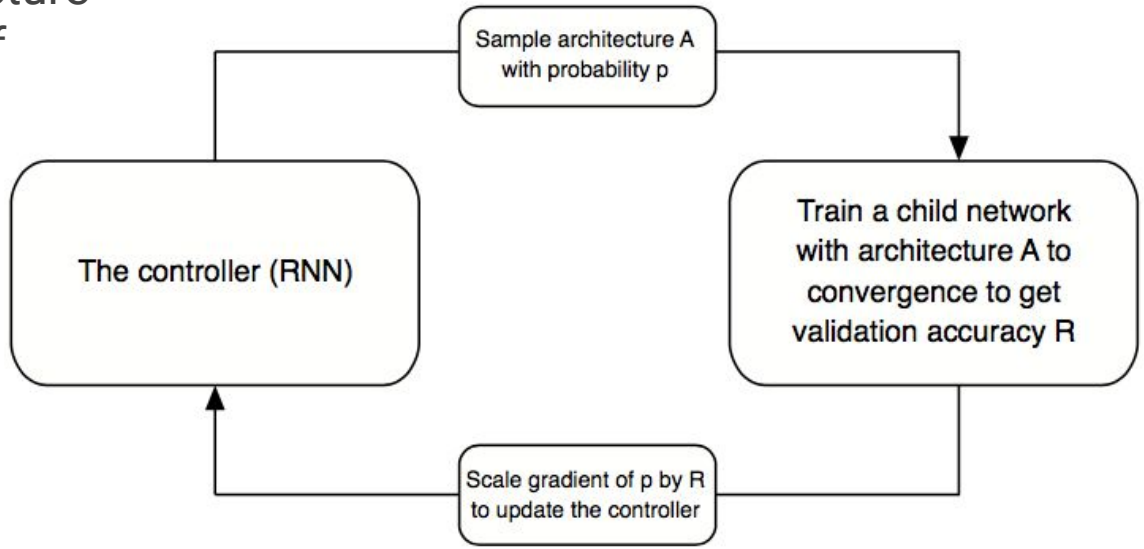
From: http://googleresearch.blogspot.com/2016_03_01_archive.html

AutoML and Neural Architecture Search

Google Cloud

AutoML and Neural Architecture Search

Learn the model architecture directly on the dataset of interest.



A controller RNN predicts architecture *A* from a search space with probability *p*. A child network with architecture *A* is trained to convergence achieving accuracy *R*. Scale the gradients of *p* by *R* to update the RNN controller.

https://research.google.com/pubs/pub45826.html , "Neural Architecture Search with Reinforcement Learning", Barret Zoph, Quoc V. Le.

AutoML and Neural Architecture Search

- Want to design a search space so that the **complexity of the architecture is independent of the depth of the network and the size of input images**.
- One inspiration for this search space is the recognition that CNN-based networks often have **repeated motifs** (such as combinations of convolutional filter banks).

From: https://arxiv.org/abs/1707.07012 , "Learning Transferable Architectures for Scalable Image Recognition" , Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V. Le.

AutoML and Neural Architecture Search

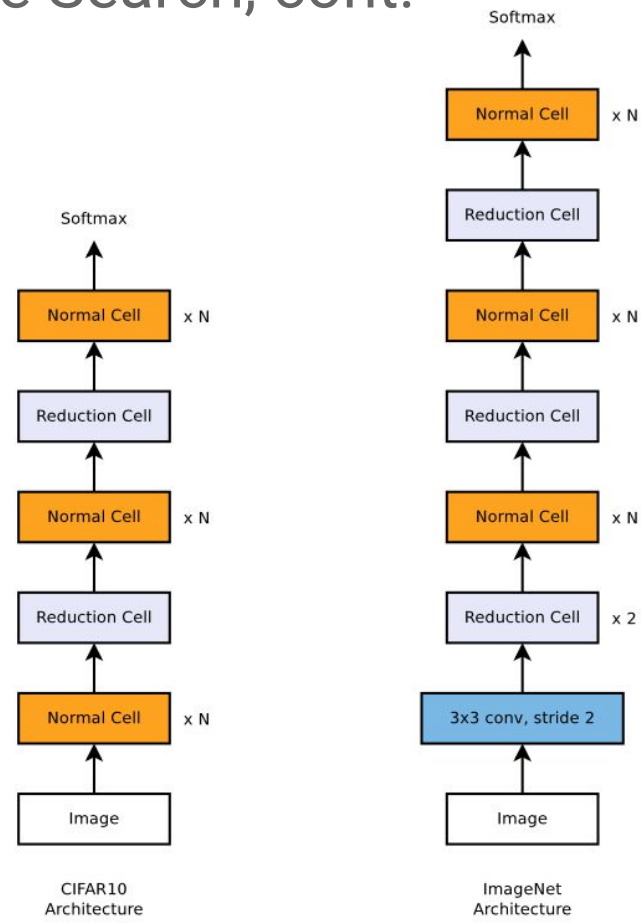
- This suggests that the controller RNN may be able to **predict a generic convolutional cell expressed in terms of these motifs**, that can be stacked.
- This approach can be expensive with large datasets, so what if you could search for an architectural building block on a small dataset and then transfer the block to a larger dataset?
- A key contribution of paper— the design of such a search space, that enables **transferability**. Search for building blocks using CIFAR-10, get good results when apply them to ImageNet.

From: https://arxiv.org/abs/1707.07012 , "Learning Transferable Architectures for Scalable Image Recognition"

AutoML and Neural Architecture Search, cont.

Overall architectures of the NNs are manually determined, using 2 types of cells: **'Normal'** and **'Reduction'**
+ and things like number of motif repetitions as free params.

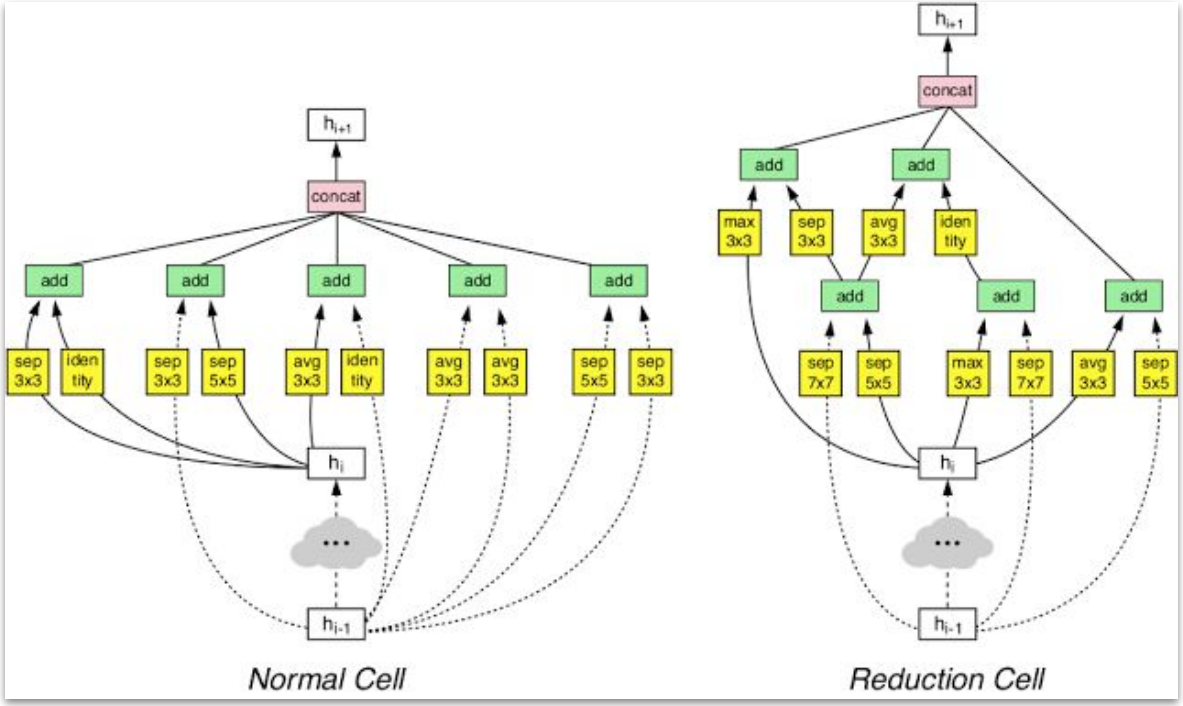
What varies are the structures of the Normal and Reduction cells. These are searched by the controller RNN.



Scalable architectures for image classification consist of two repeated motifs termed *Normal Cell* and *Reduction Cell*. This diagram highlights the model architecture for CIFAR-10 and ImageNet.

From: https://arxiv.org/abs/1707.07012 , "Learning Transferable Architectures for Scalable Image Recognition"

AutoML and Neural Architecture Search, cont.

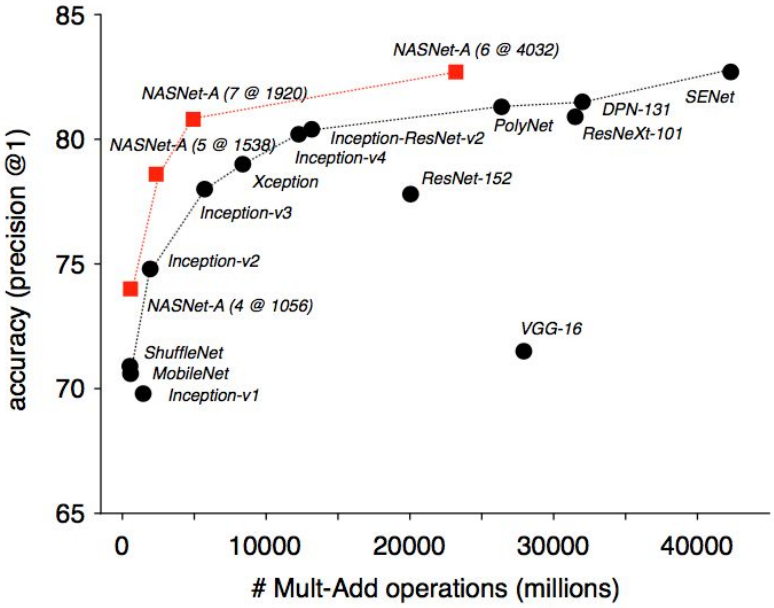


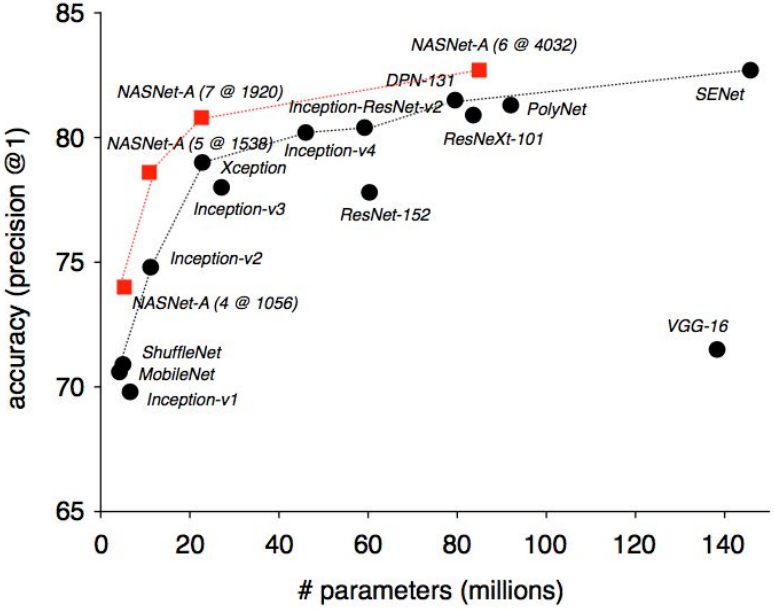
Architecture of the best convolutional cells (NASNet-A)

The best architecture found using CIFAR-10 achieves **state-of-the-art accuracy when transferred** to ImageNet classification

(From: https://arxiv.org/abs/1707.07012 , https://research.googleblog.com/2017/11/automl-for-large-scale-image.html)

AutoML and Neural Architecture Search





(From: https://arxiv.org/abs/1707.07012 , https://research.googleblog.com/2017/11/automl-for-large-scale-image.html)

Why Cloud AutoML?

Limited ML expertise needed

Unlike all other ML services, you need limited ML expertize to to train a machine learning model.

Your own custom models

Easily create and customize your own state-of-the-art ML models for your unique use case.

Simple

Simple graphical user interface (GUI) where you just have to specify content, and Cloud AutoML turn content into a high quality model from your dataset.

High Quality

State-of-the-art performance on public benchmark CIFAR and ImageNet

For more:
https://cloud.google.com/automl/
To sign up:
Cloud AutoML Alpha Program (https://goo.gl/NjQ8Ss)

Read more:
https://www.blog.google/topics/google-cloud/cloud-automl-making-ai-accessible-every-business/
https://www.nytimes.com/2018/01/17/technology/google-sells-ai.html

These slides: https://goo.gl/S6qt3k

Thank you!