Machine Learning for Geospatial Raster Data

Workshop #2 Please visit bit.ly/geoml-2

Presented by Juan Carrillo & Jaydeep Mistry

On behalf of the Geospatial Club - Winter 2019
In partnership with the University of Waterloo Geospatial Centre

Agenda

1. Why should we care about Machine Learning for Raster data?
2. Problem context: Understanding the Amazon from Space
3. Setup of the cloud-computing platform
4. Data exploration
5. Convolutional Neural Networks
6. Wrap-up

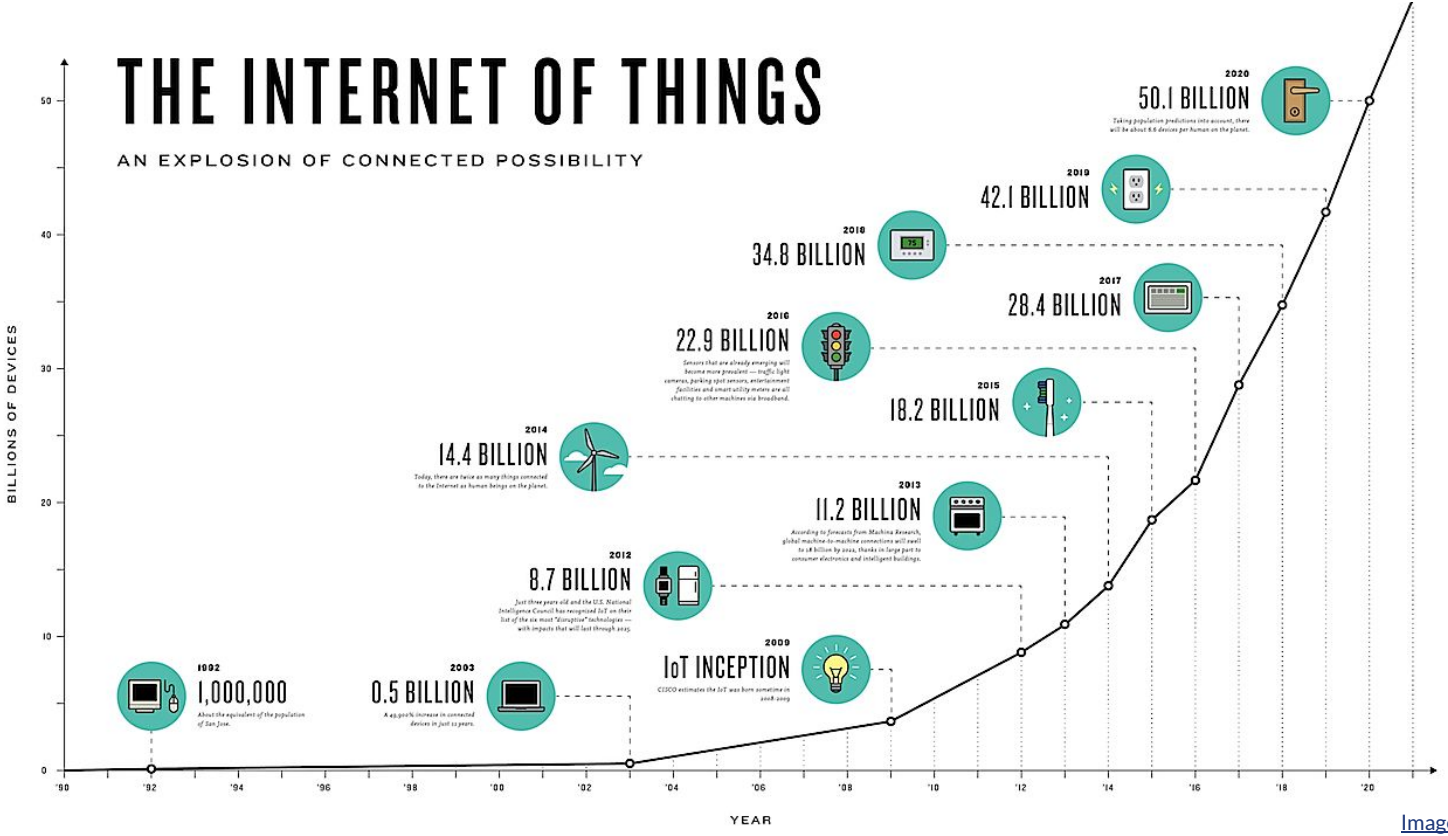


Image Source

Why does Machine Learning matter?

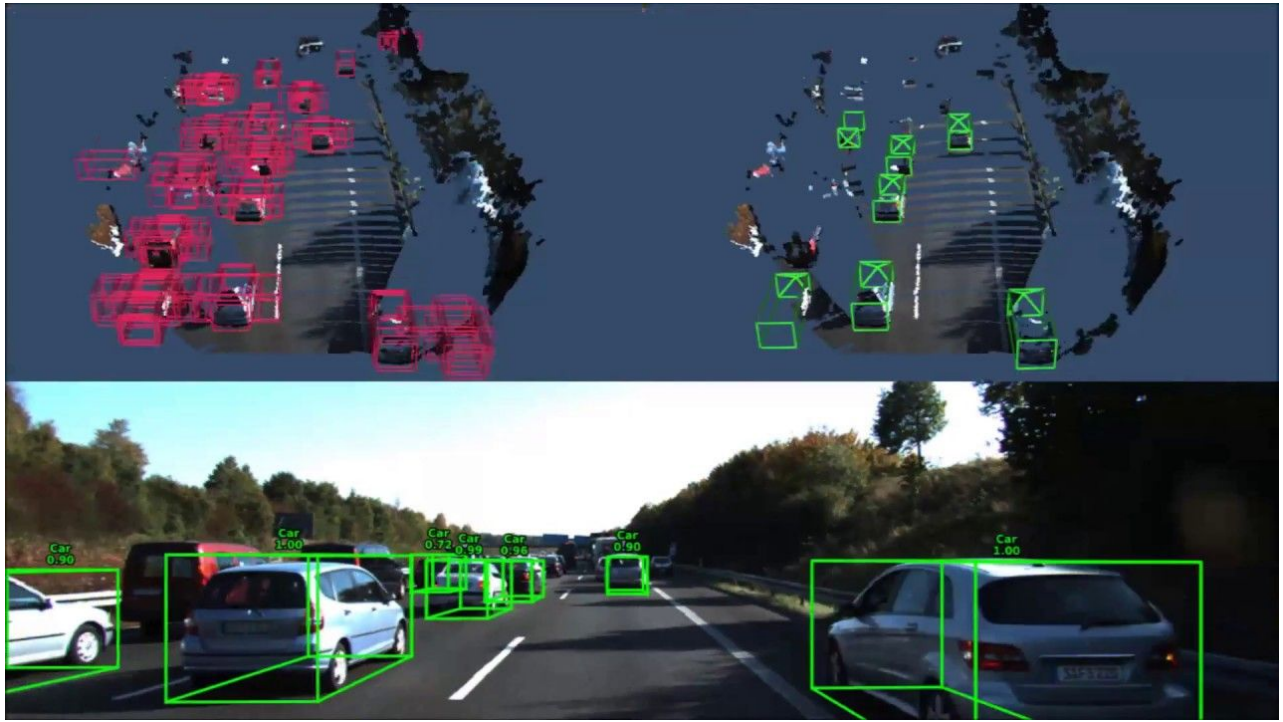


Image Source

Common Computer Vision tasks

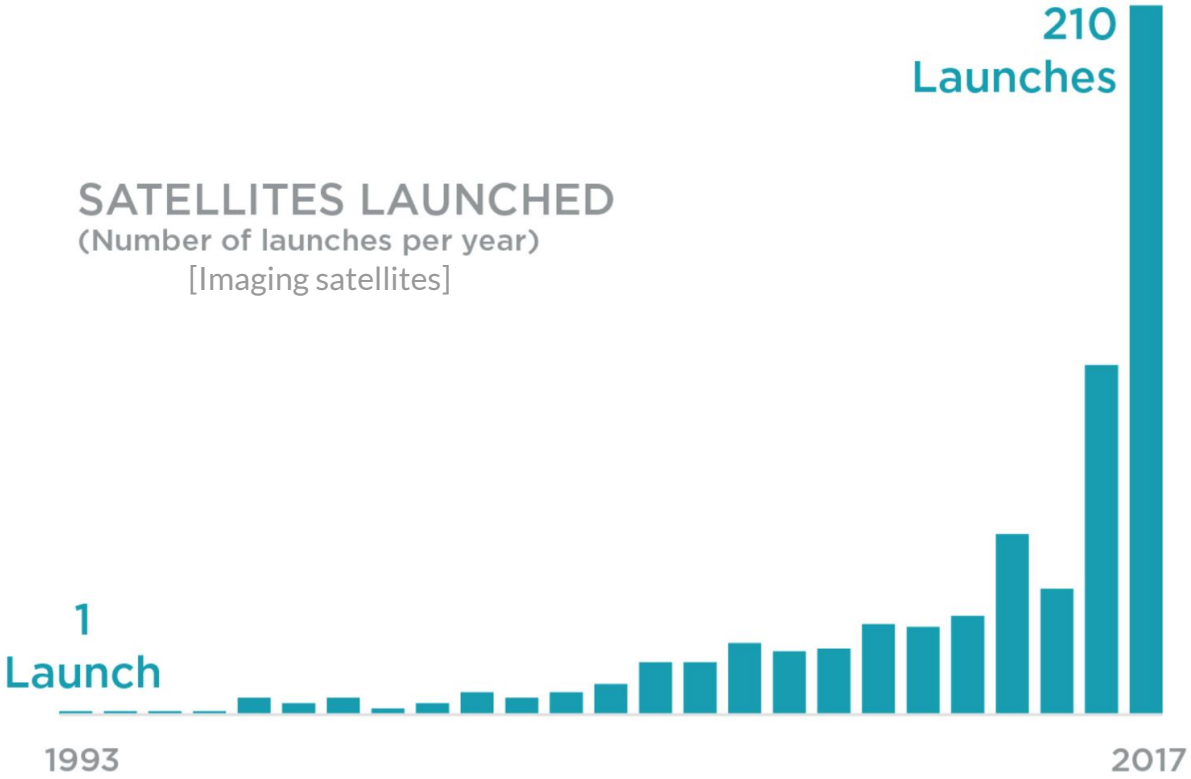


Image Source

Why does Raster data matter?

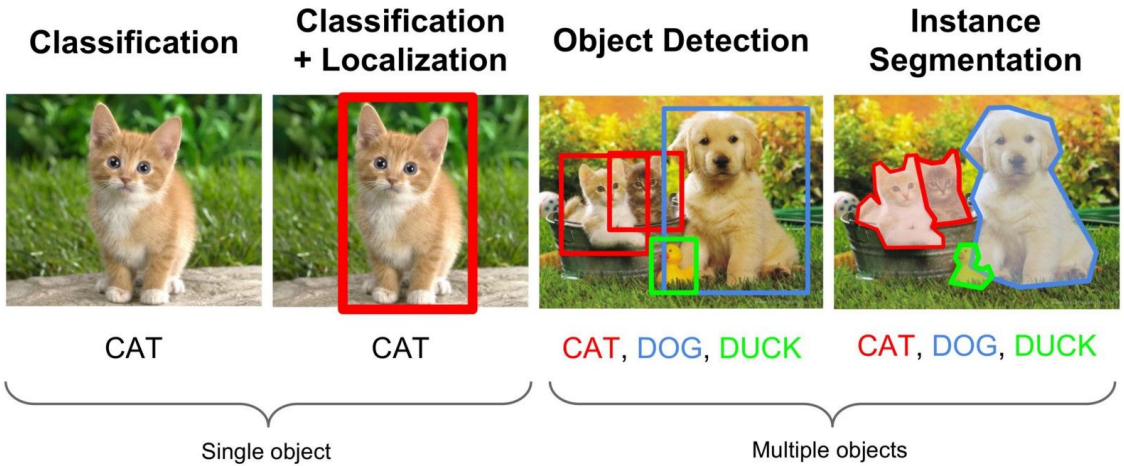


Image Source

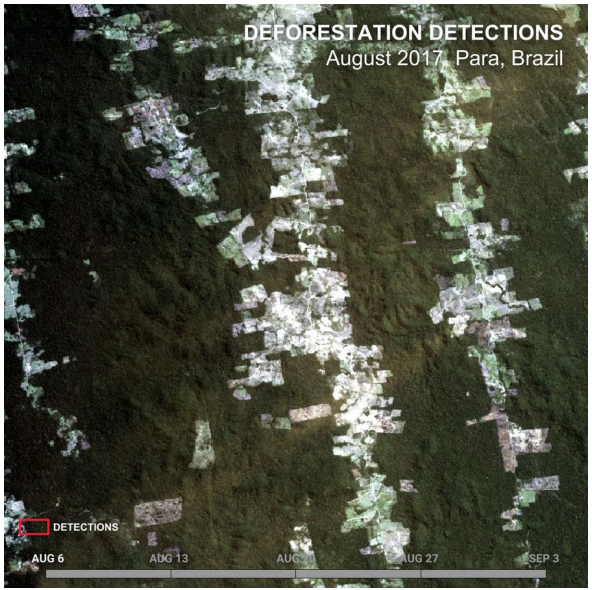


Image Source

Common Computer Vision tasks

The Problem:
Identify Human Activity in the Amazon Rainforest

1. Use satellite data
2. Track human footprint in the Amazon rainforest
3. Popular Kaggle Competition for $60,000

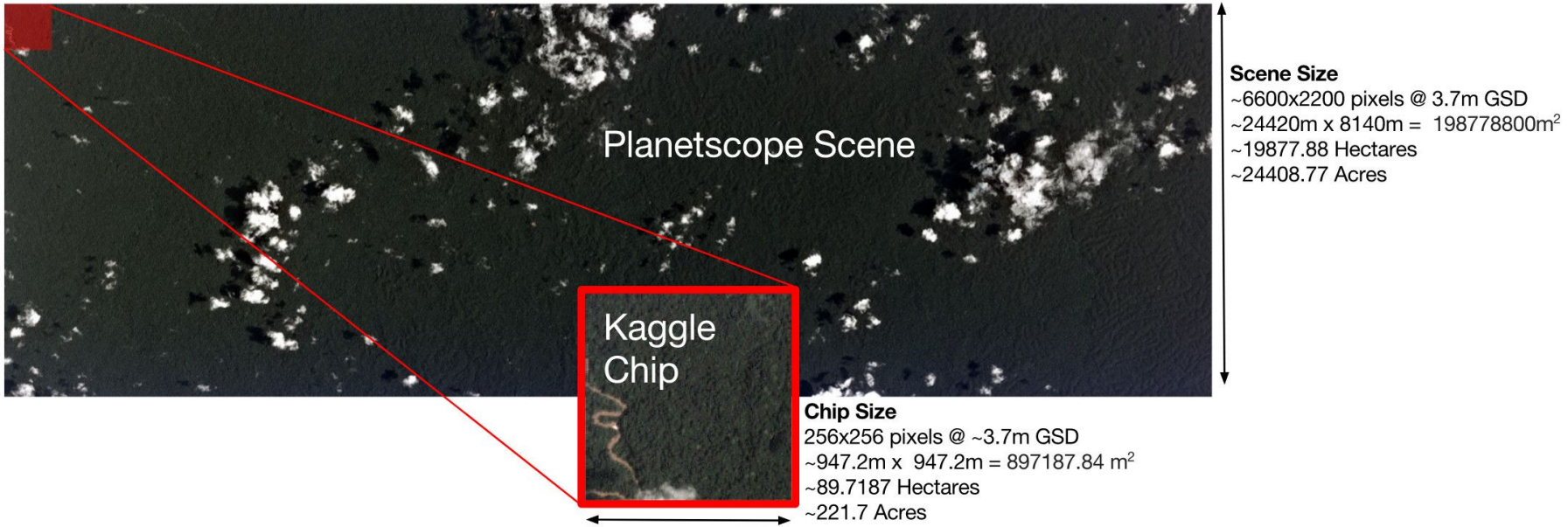


Image Source

Dataset: 40,000+ Satellite Images

How to perform Machine Learning online?

- Free-to-use platform
- Designed for Data Science Competitions
- Full integration with **Python** and **R** for Cloud Computing

Optional: Google Colab, Earth Engine

Specs of a Kaggle Kernel virtual machine

16 Gigabytes of disk storage

4 CPU cores

17 Gigabytes of RAM

NVidia K80 GPU (12.5X speedup)

Source: Kaggle and BH seller

Let's get started!

- Dataset
- Notebook

Image source

Normalized Differential Vegetation Index (NDVI)

- Quantifies vegetation by using difference between Near-Infrared and Red light

- Near-Infrared reflected by vegetation
- Red light absorbed by vegetation
- Ranges between -1 to +1

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

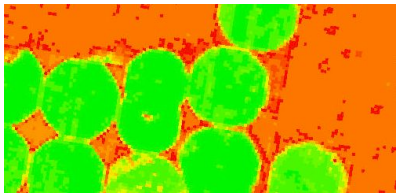


Image Source

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- For data dimensionality reduction
- Visualization of high dimensional datasets
- Uses t-Distribution to avoid having clusters group together

Creator; Visual Examples

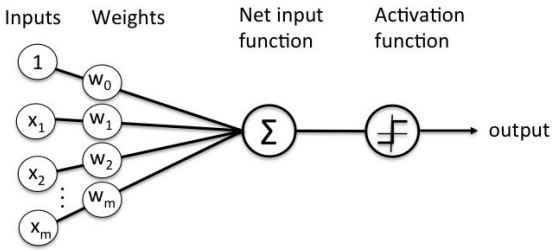
Why use this?

- Our Amazon rainforest dataset has
  - 40,000+ images
  - 256 x 256 pixels per image
  - Each image has 4 bands of data: Blue, Green, Red, Near-Infrared

Convolutional Neural Networks (CNN)

What are Neural Networks?

- System of connected ‘Neurons’ inspired by biological neural networks



Schematic of Rosenblatt’s perceptron.

What is CNN?

- Multi-layered system of Neurons, where the data gets Pooled to remove noise

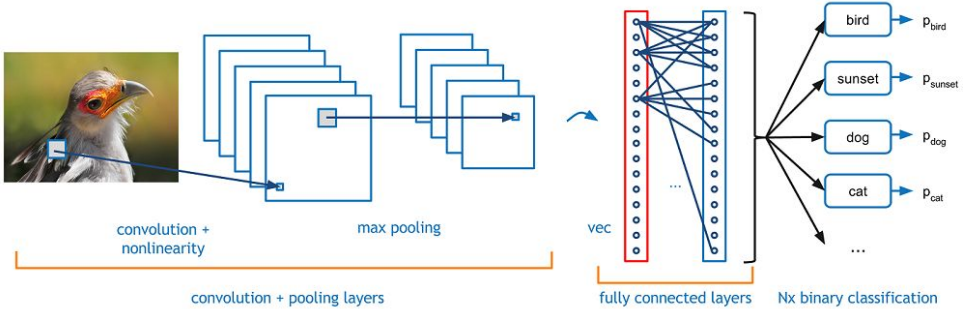


Image Source

Convolutional Neural Networks (CNN)

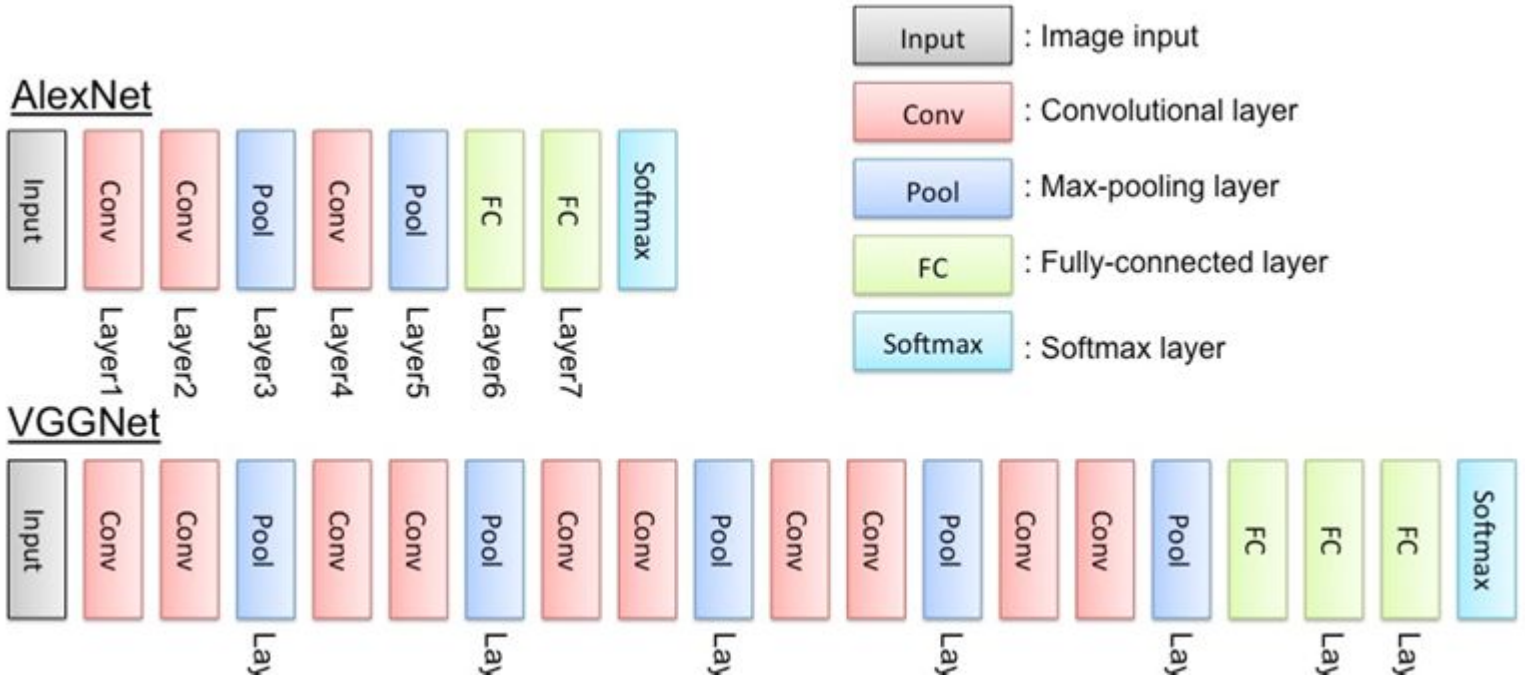


Image Source

Deep Convolutional Neural Networks (CNN)

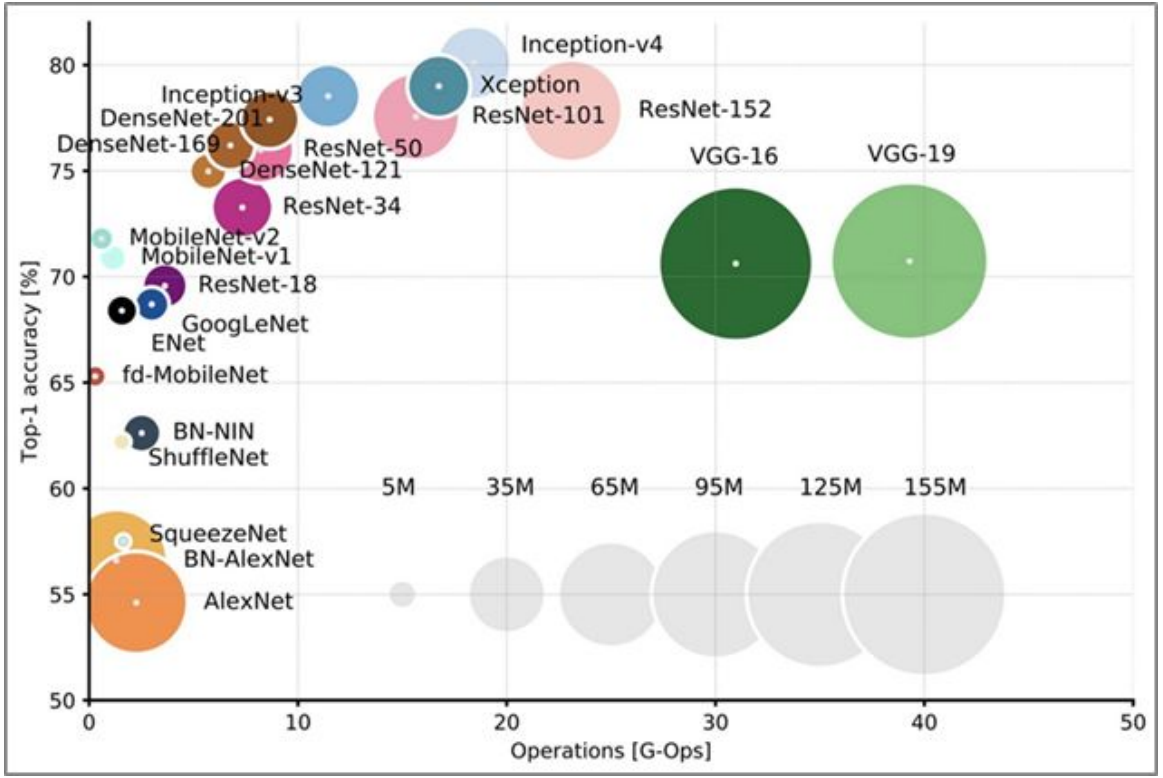


Image Source

Deep Convolutional Neural Networks (CNN)

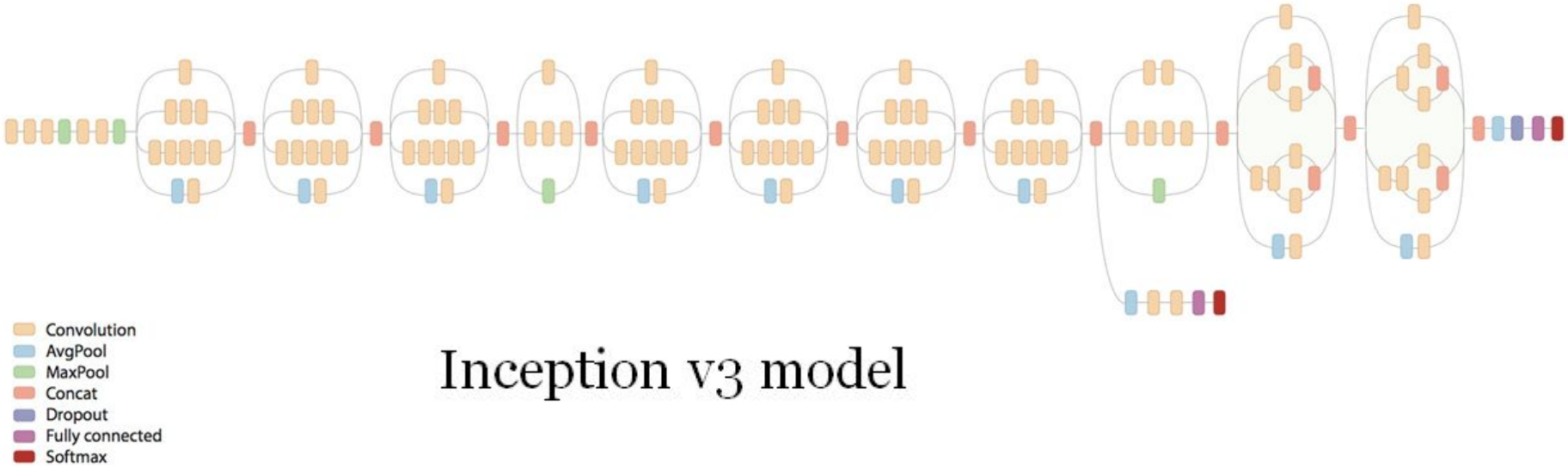
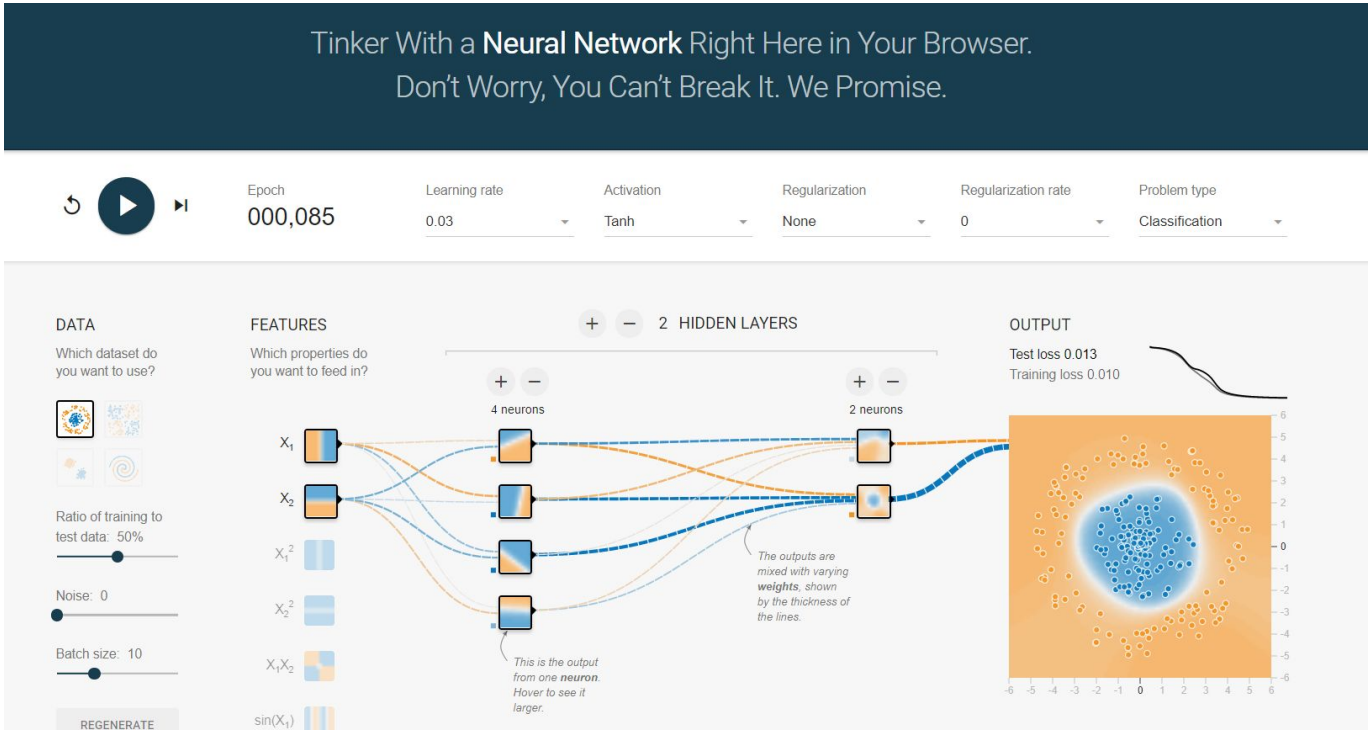


Image Source

More on Neural Networks - online demo



[Try it here!]

Time to Train a basic CNN

Computing on CPU - 5-6 mins Avg

```
Train on 35000 samples, validate on 5479 samples
Epoch 1/4
35000/35000 [==============================] - 91s 3ms/step
Epoch 2/4
35000/35000 [==============================] - 90s 3ms/step
Epoch 3/4
35000/35000 [==============================] - 90s 3ms/step
Epoch 4/4
35000/35000 [==============================] - 90s 3ms/step
```

Computing on GPU - 30s Avg **(10x Faster)**

```
Train on 35000 samples, validate on 5479 samples
Epoch 1/4
35000/35000 [==============================] - 11s 309us/step
Epoch 2/4
35000/35000 [==============================] - 6s 172us/step -
Epoch 3/4
35000/35000 [==============================] - 6s 172us/step -
Epoch 4/4
35000/35000 [==============================] - 6s 173us/step -
```

Wrap-Up

Thank You to the Geospatial Centre for this collaboration!

Slide of this workshop: bit.ly/geoml-2

Please upvote Kernel and dataset :)

Contact us at:

https://www.facebook.com/uwgeospatial/

Jaydeep Mistry
jaydeep.mistry@uwaterloo.ca

Juan Carrillo
jmcarril@uwaterloo.ca