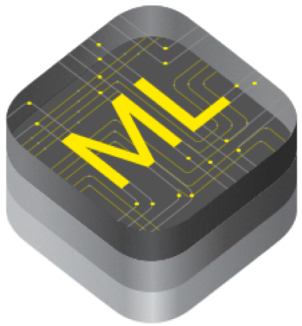
# AI On Your Phone

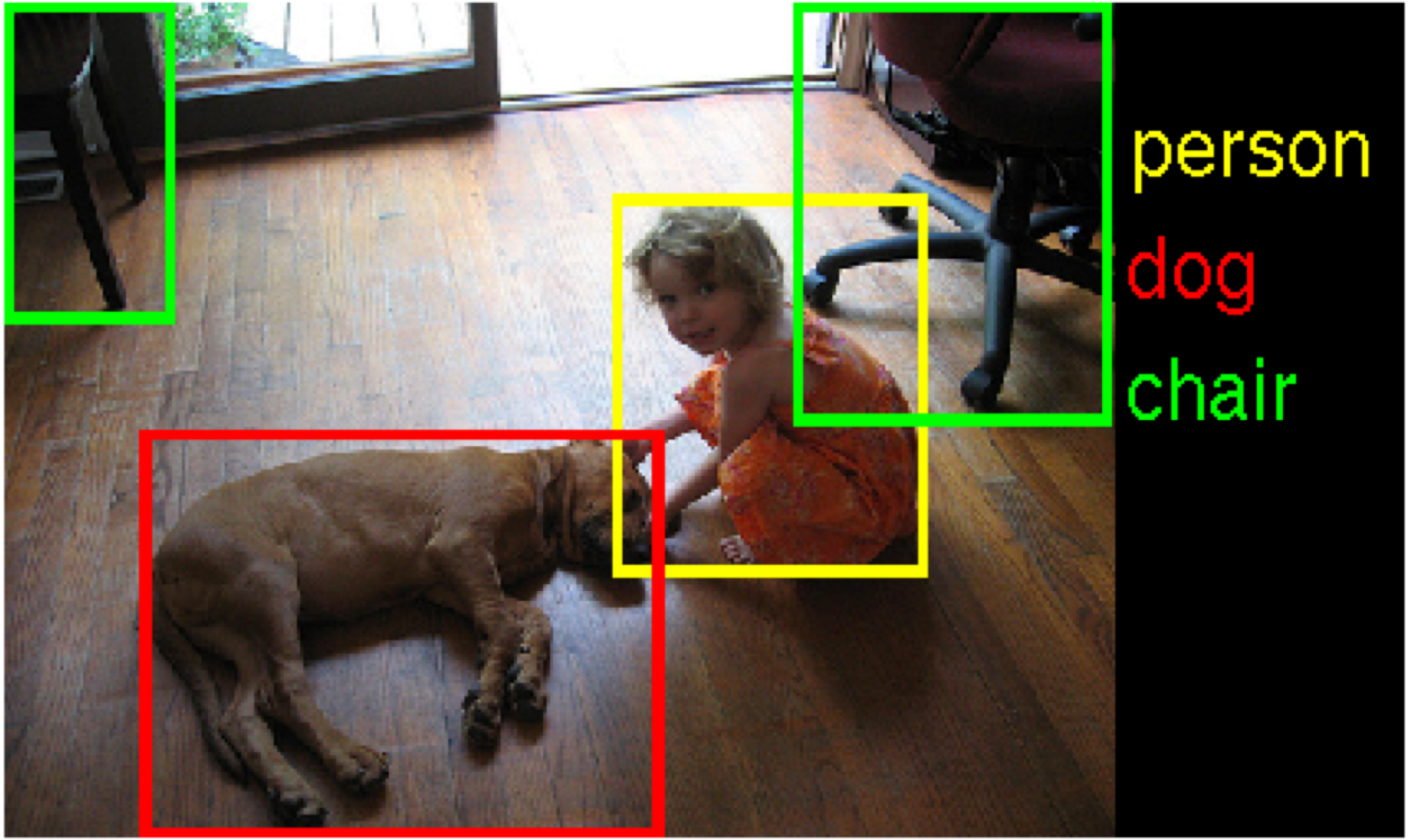
@ray_deck

github.com/rhdeck/boscc29



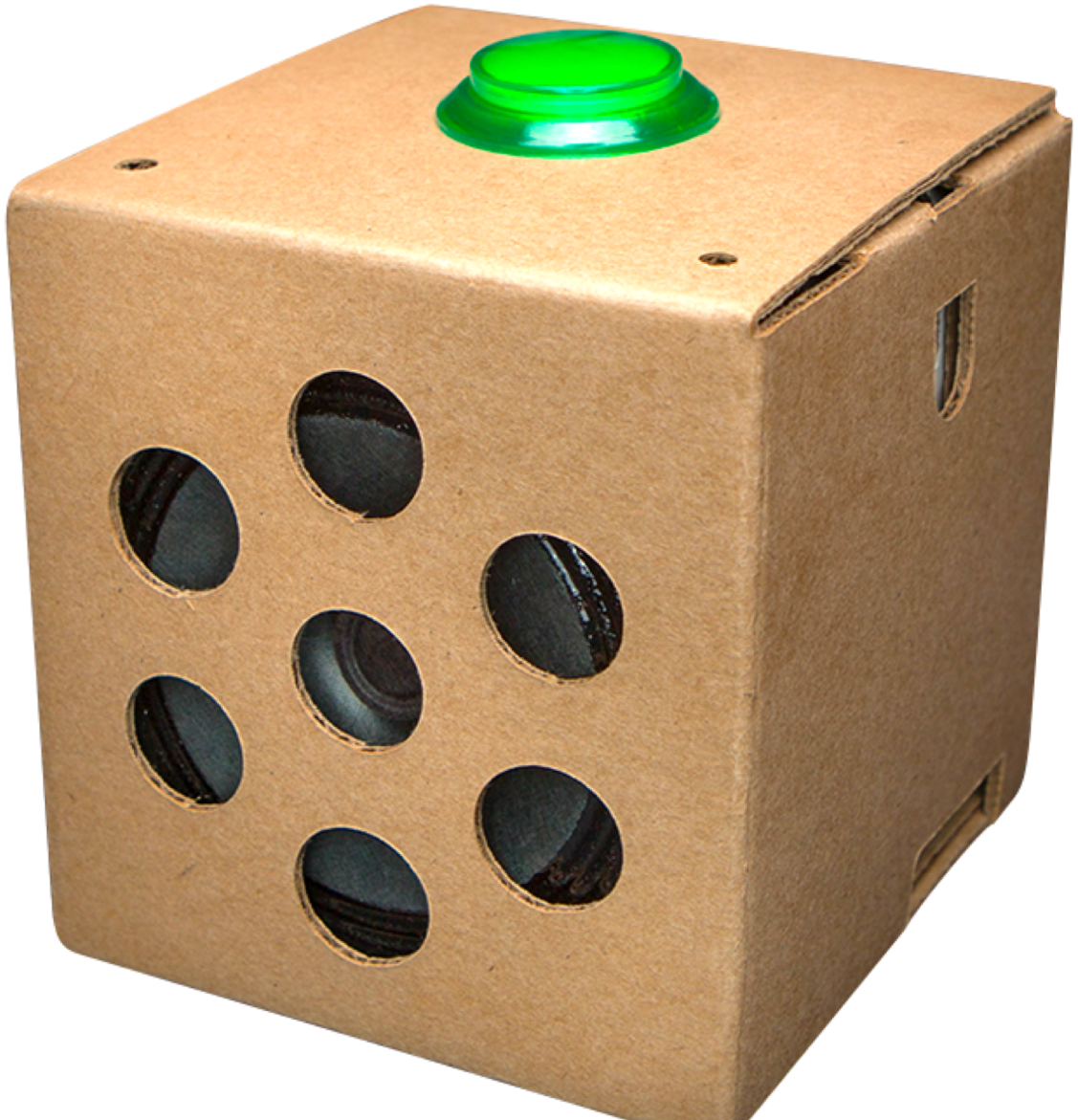

# What’s it Good For?

# Identification

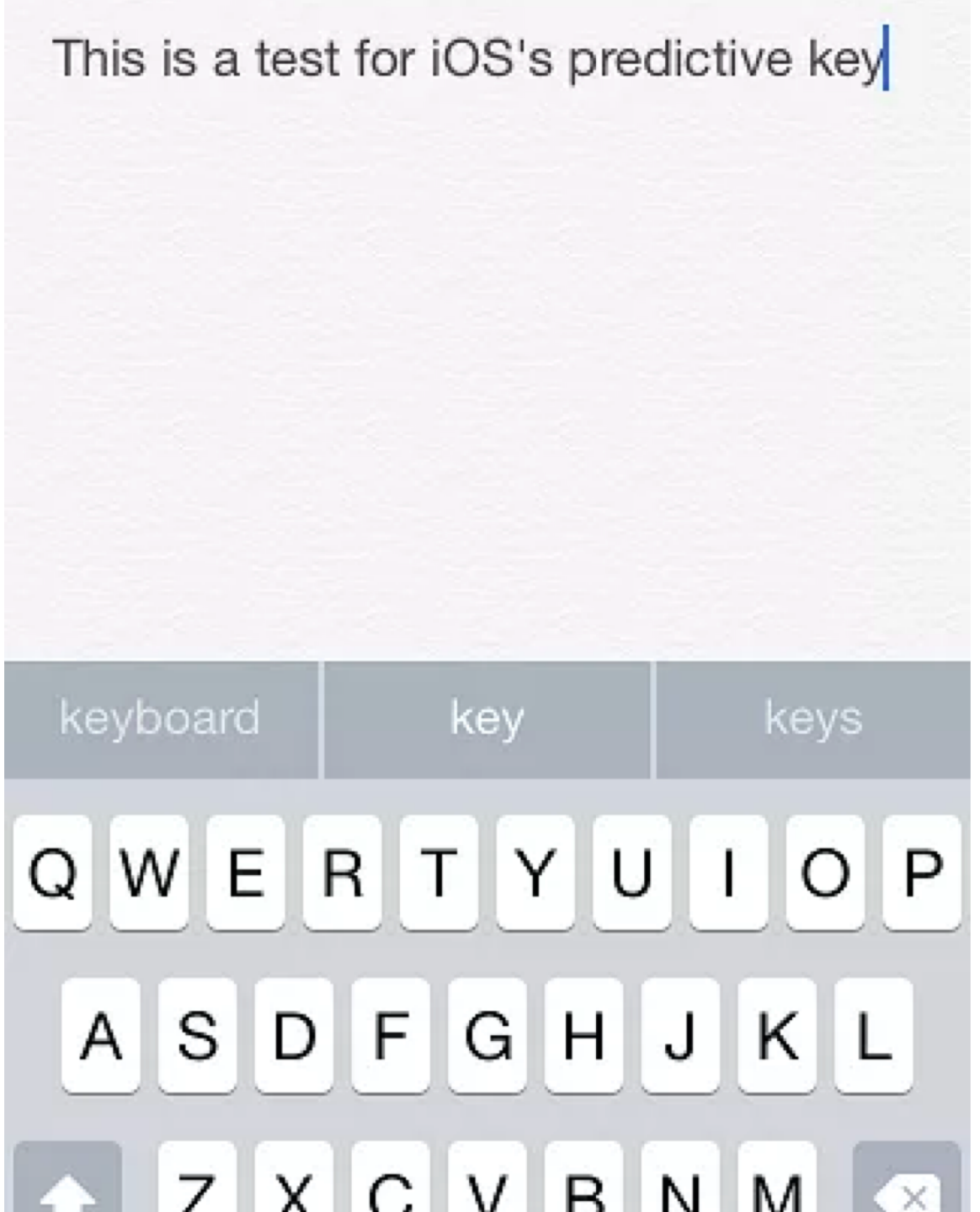


# Generation

# Voice



# Prediction

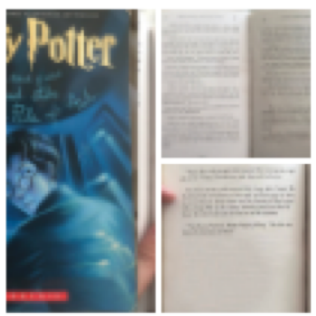

This is a test for iOS's predictive key
keyboard
key
keys
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M

# Prediction

**Al Dente's Inferno**
@dirtylonghair

Follow

"The Death Eaters were dead now, and Harry was hungrier than he had ever been."



**Botnik Studios** @botnikstudios
We used predictive keyboards trained on all seven books to ghostwrite this spellbinding new Harry Potter chapter botnik.org/content/harry-…

Show this thread

3:59 PM - 12 Dec 2017

# What is a Machine Learning Model?

Artificial Intelligence

Deep Learning

**Machine Learning**

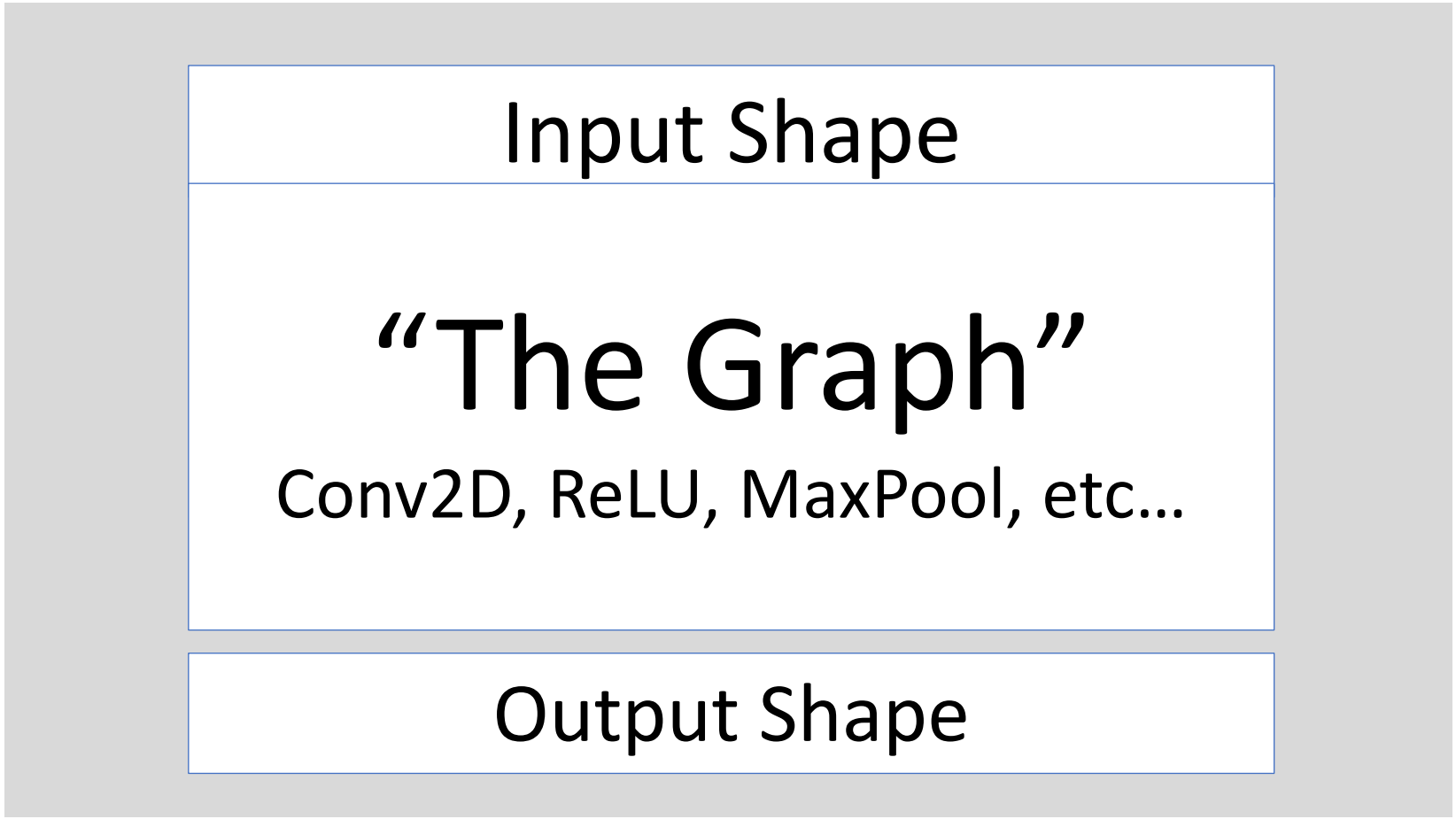
Statistical Computing

Supervised Learning

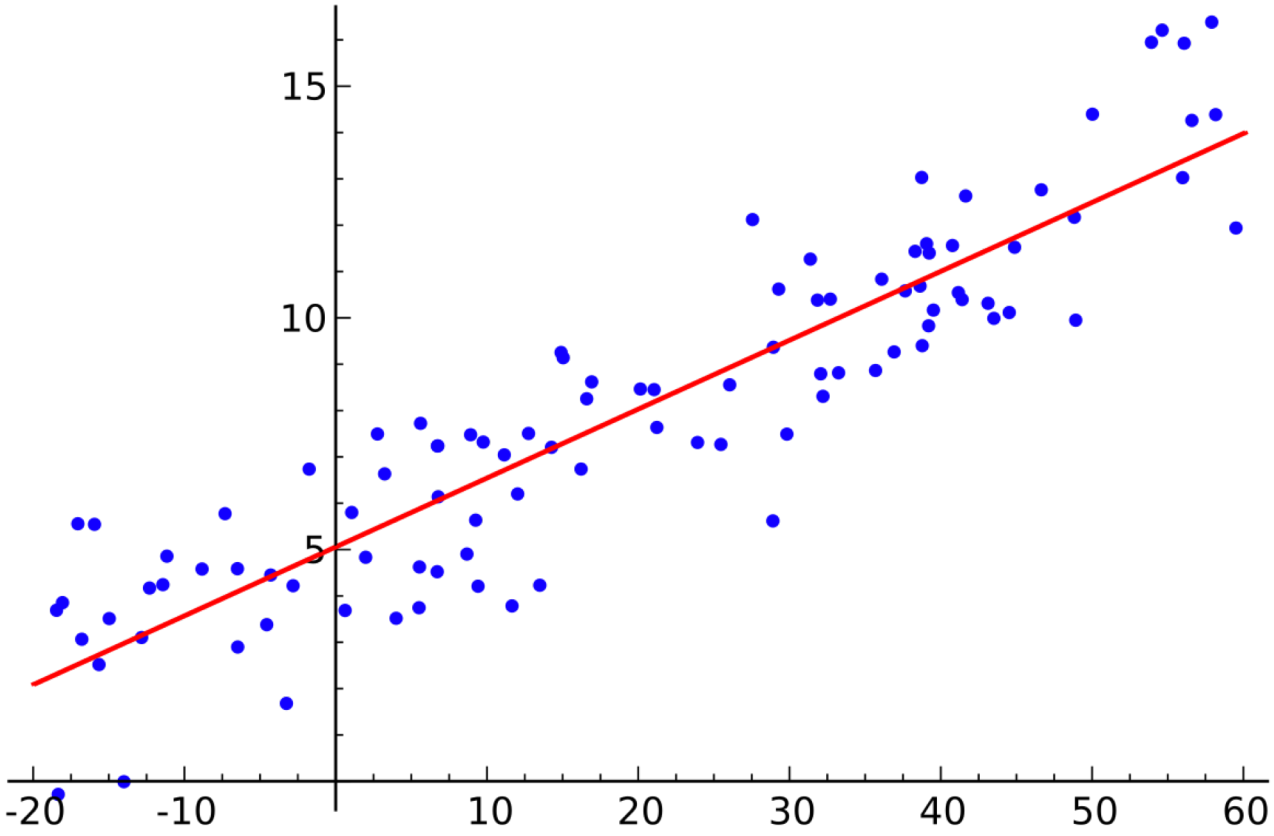
Unsupervised Learning

Reinforcement Learning

# Ops and Numbers

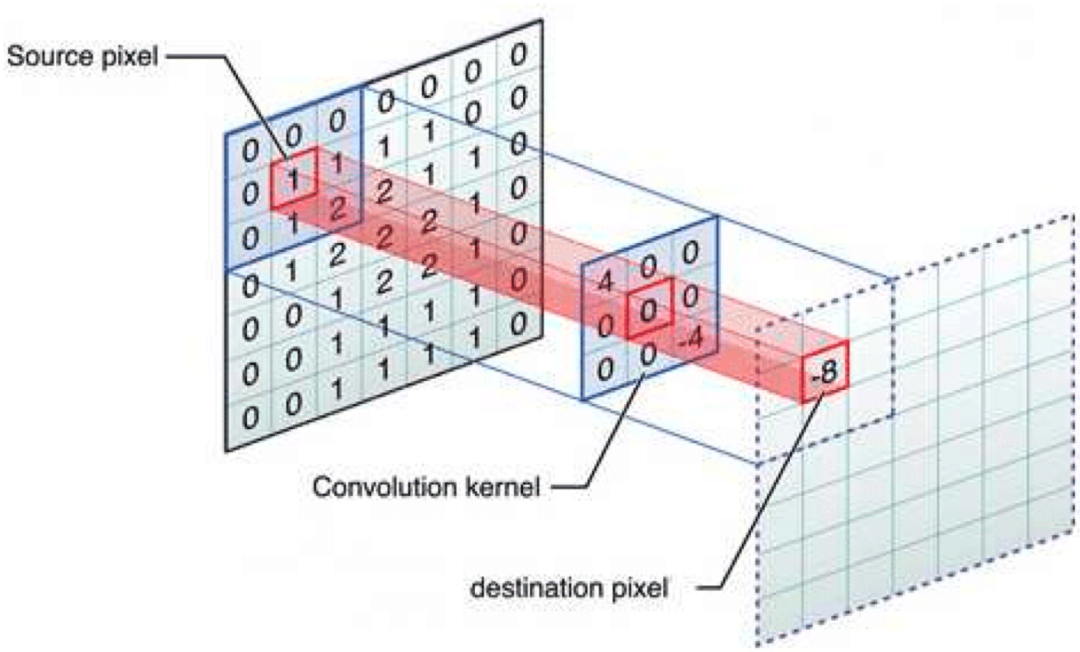


$$y = mx + b$$

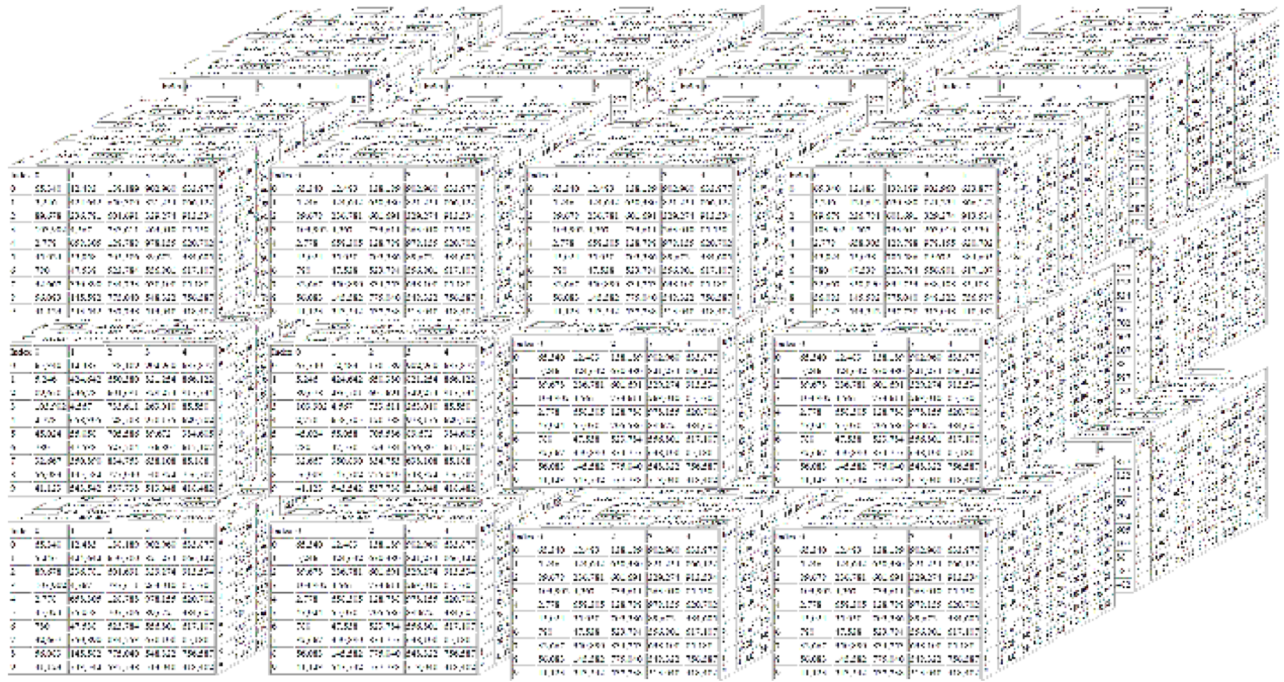


# Ops

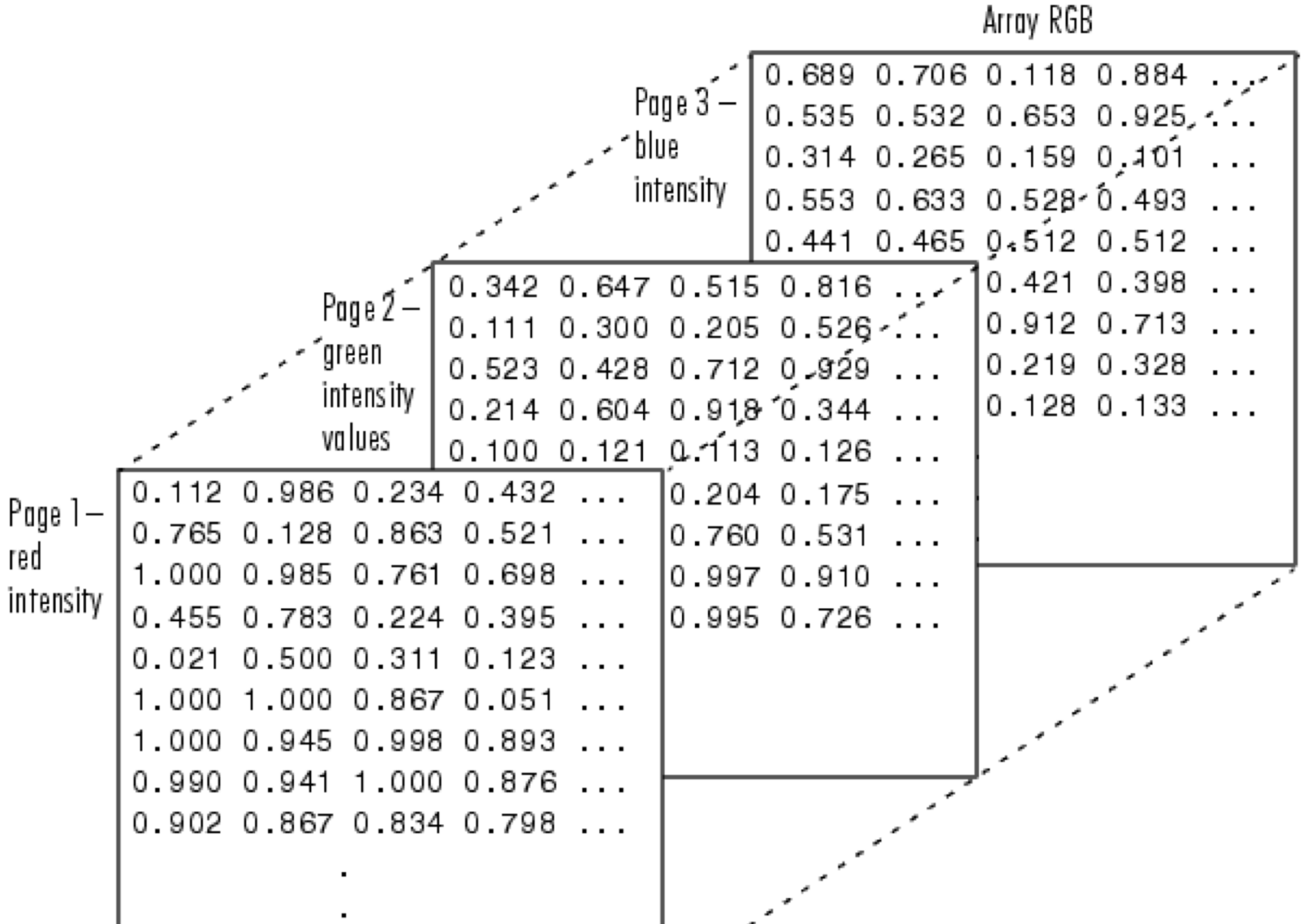
Convolutions
Pooling
ReLU
TanH
Sigmoid
LSTM
GRU
Dense/Fully-Connected
More…



# High-Dimensional Numbers



# Images are High-Dimensional Numbers!



# Old Stuff Works

https://arxiv.org/pdf/1803.01271.pdf

---

## An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling

---

**Shaojie Bai** [1] **J. Zico Kolter** [2] **Vladlen Koltun** [3]

**Abstract**

For most deep learning practitioners, sequence modeling is synonymous with recurrent networks. Yet recent results indicate that convolutional ar-

chine translation (van den Oord et al., 2016; Kalchbrenner et al., 2016; Dauphin et al., 2017; Gehring et al., 2017a;b). This raises the question of whether these successes of convolutional sequence modeling are confined to specific ap-

018

# Packaging

# Models (Can Be) Files

Shapes of Input(s)

Operations

Hyperparameters

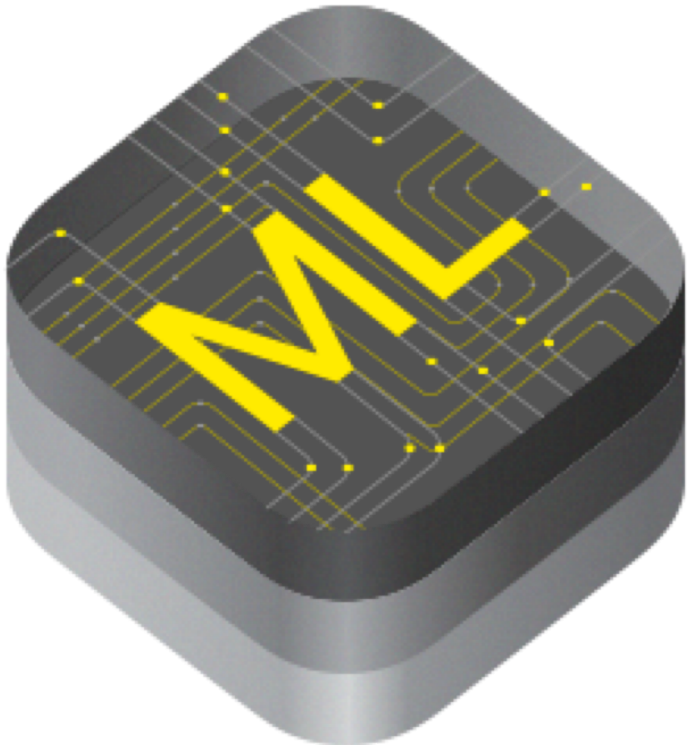
Weights

Sources & Drains

Shapes of Output(s)


# What Do we Need?

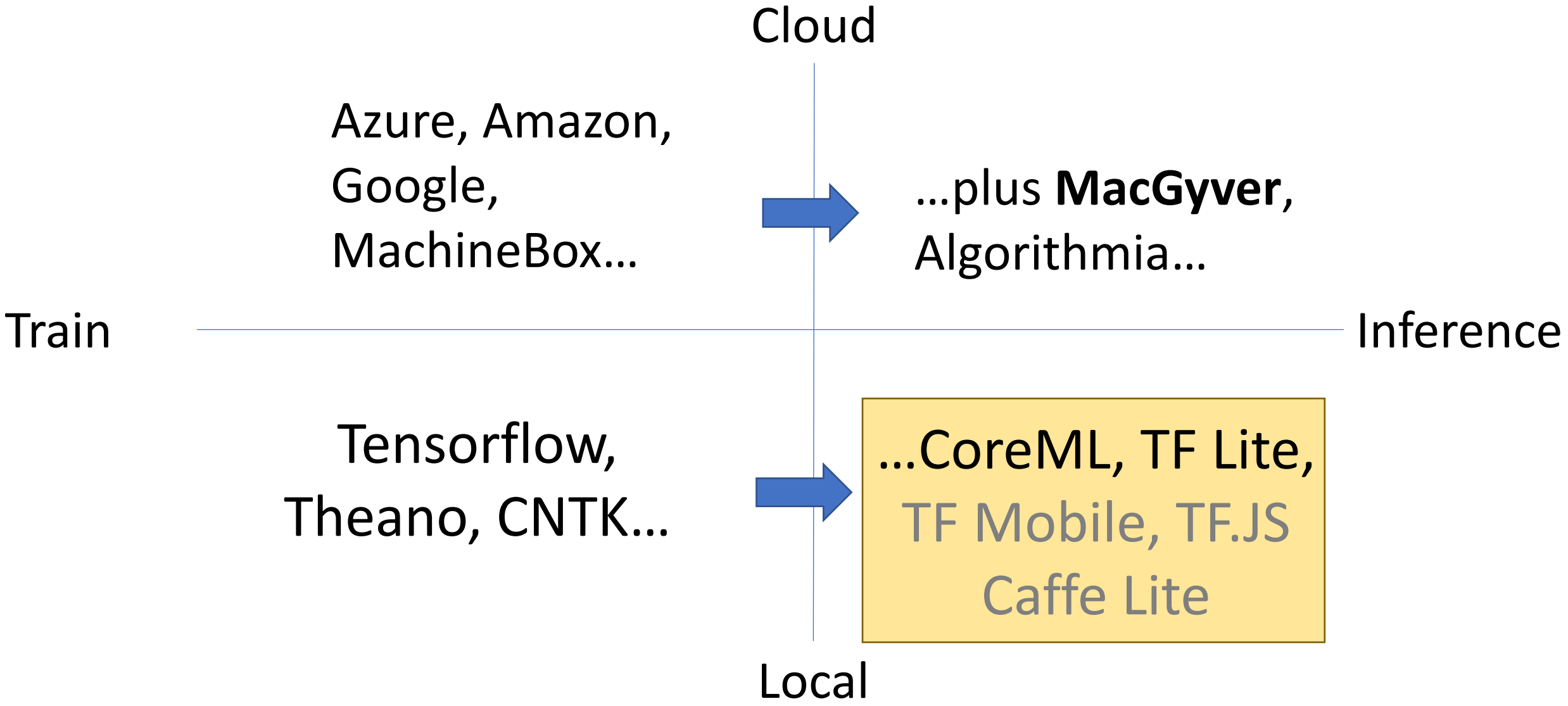
1. **Convert** Image/Frame to Normalized Number
2. Run the model efficiently on local resources
3. **Convert** output number to what we seek:
   - Dictionaries of labels
   - Generated Image
   - Prediction
   - etc

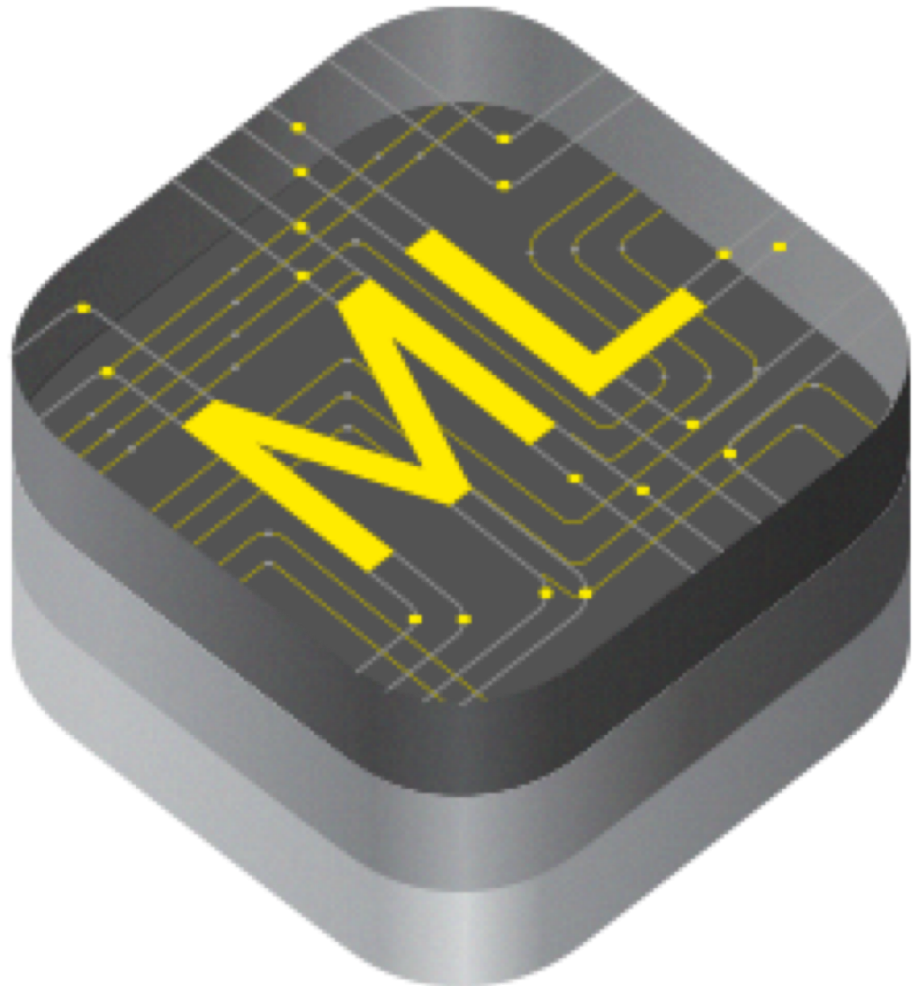


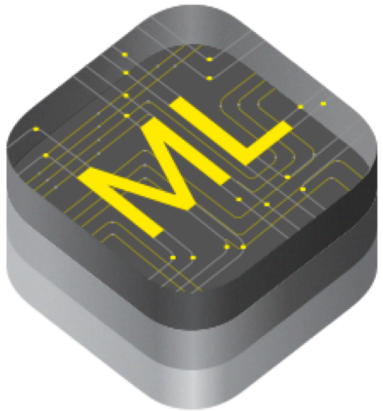

TensorFlow

Lite

# ML Universe


Cloud
Azure, Amazon,
Google,
MachineBox…
…plus **MacGyver**,
Algorithmia…
Train
Inference
Tensorflow,
Theano, CNTK…
…CoreML, TF Lite,
TF Mobile, TF.JS
Caffe Lite
Local


ML

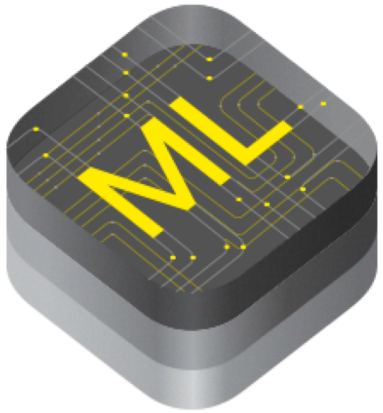



iOS Runtime and Bindings

File Format

Conversion Tooling

# Runtime and Bindings



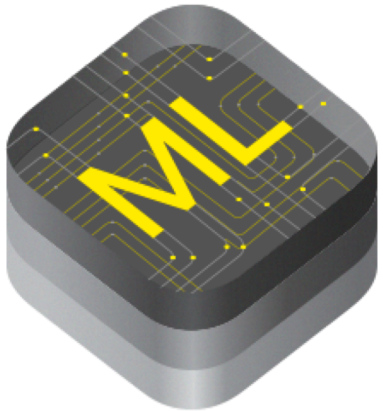

CVPixelBuffer in, Dictionary Out

Support for OTA/Runtime Download and Compilation

Uses Metal Shaders for processing

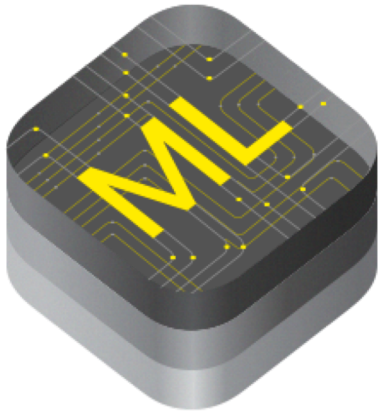
Limited Operation Support

# .mlmodel




1. Protobuf
2. Feature Normalization
   - Value Range of Numbers (0-255 -> 0-1)
   - Mean Value
3. Channel Ordering
   - BGR vs RGB
4. Output Labeling/Denormalization

# Conversion



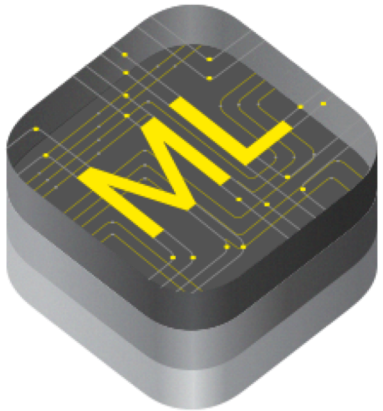

**Keras** – coremlconverter

**Tensorflow** – tf-coreml

**CNTK** - mmdnn

**MXNet** – mxnet-to-coreml

# Advantages




## Ease of Application

## Speed

Multiple analyses per second

## Separation of Concerns

Compilation vs Inference

## Source Agnostic


Lite

Multiple Runtimes

File Format

Converter

# Cross-Platform


Lite

iOS, Android, Raspberry Pi

feed() and run()

No Compilation Step

Limited Operations

Number-to-Number

# .tflite


Lite

FlatBuf

Quicker Memory Access

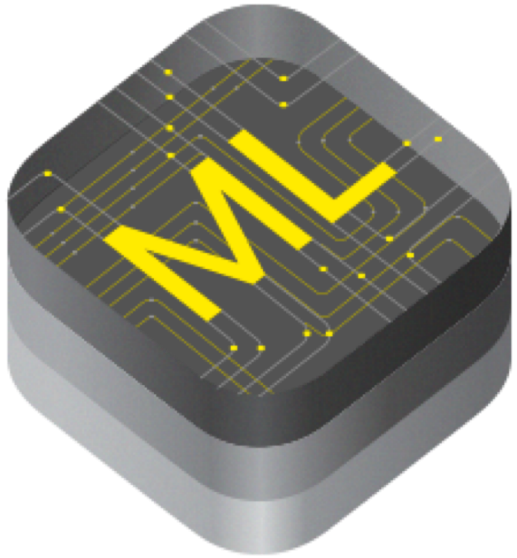
Alternative Precision: Bfloat16

# Conversion


Lite

toco

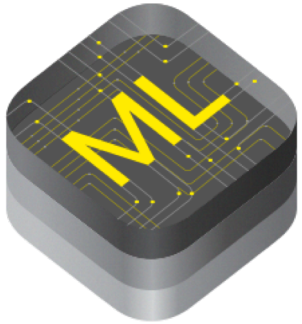
Tensorflow Optimizing Converter



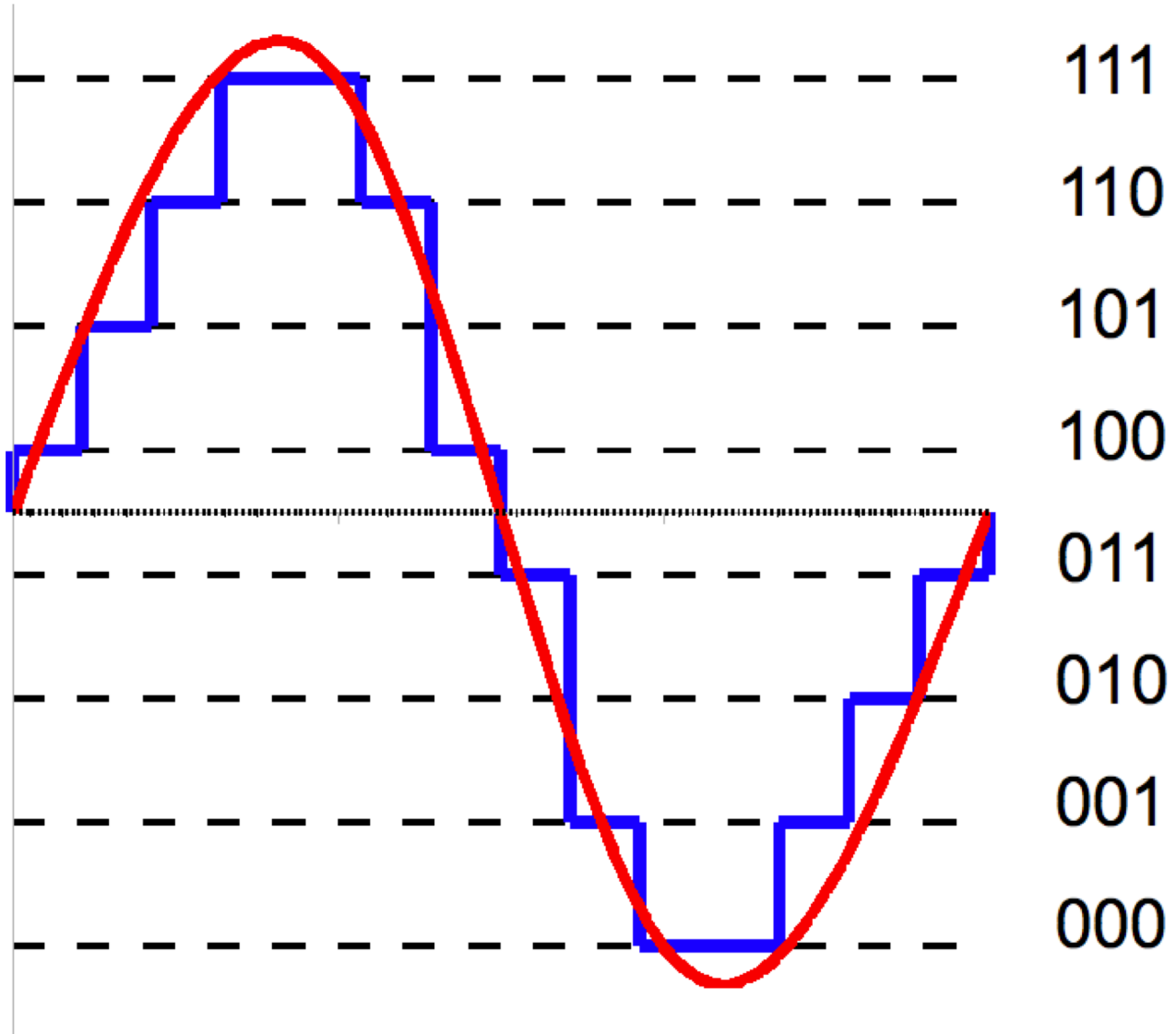


| | |
|---|---|
| iOS Only | Multiplatform |
| Input/Output Affordance | Just the Numbers |
| Production | Beta |
| Chunky | Small |

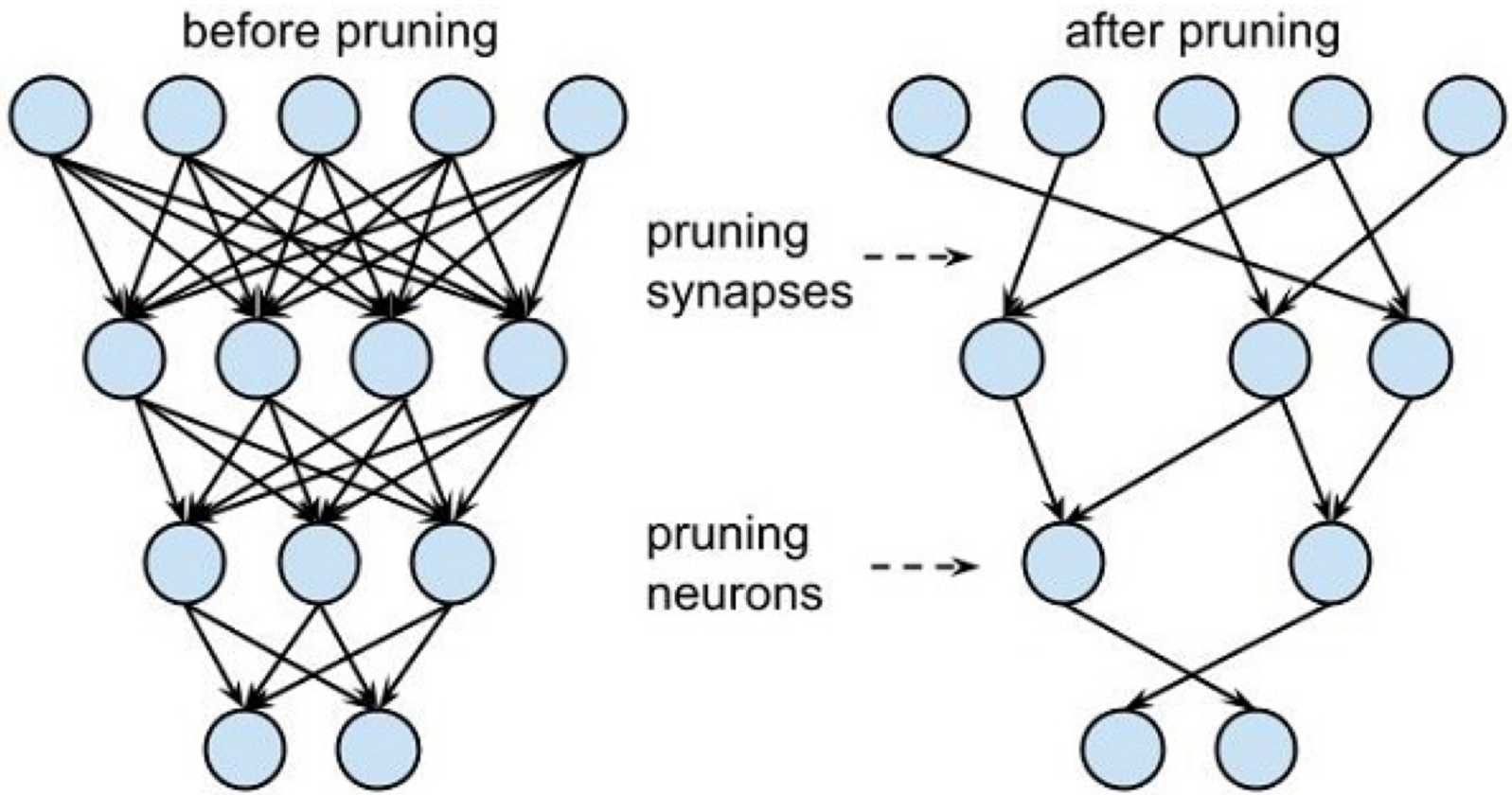
# Cross-Platform Optimizations




# Quantization



# Pruning


before pruning
after pruning
pruning
synapses
pruning
neurons

# Planning

| | |
|---|---|
| VGG16 | **550**MB |
| InceptionV3 | **94**MB |
| MobileNet | **17**MB |

# One More Thing


TensorFlow.js


WebGL

Functional Programming

Cross-Platform Nirvana?


@ray_deck

github.com/rhdeck/boscc29