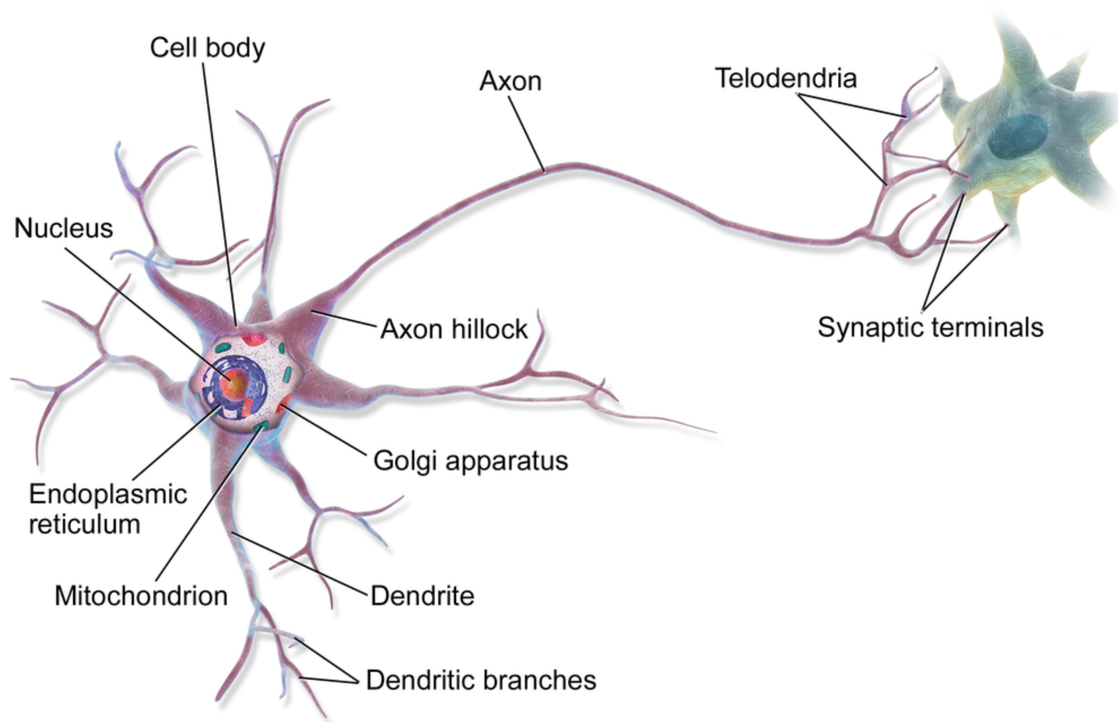
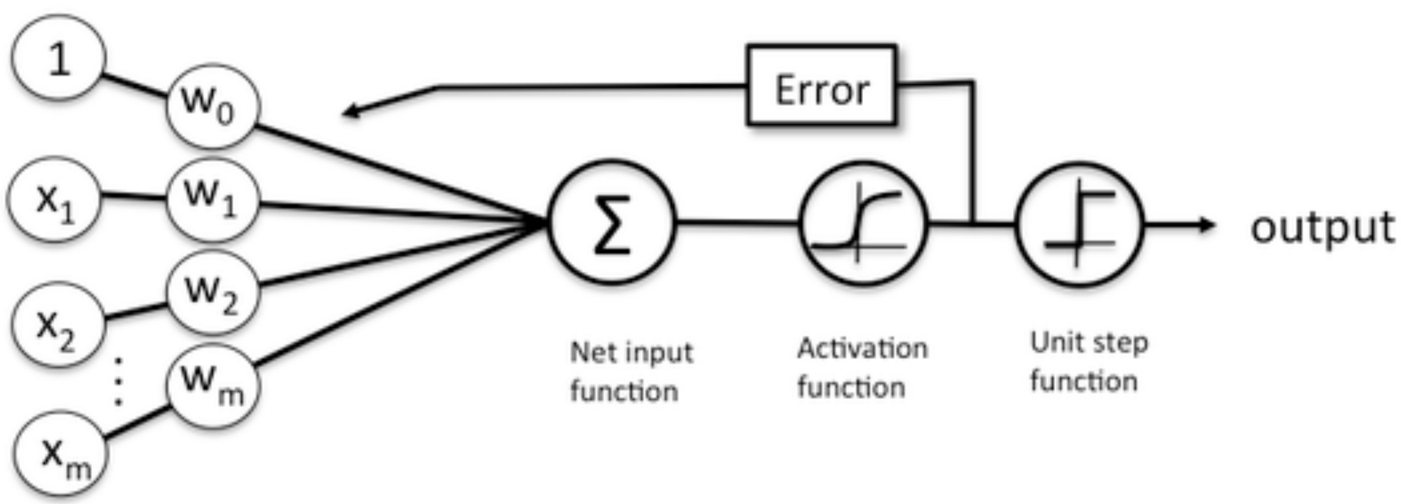
Deep Learning

Ashish Mahabal
<ashish@caltech.edu>
Lead Computational and Data Scientist
Astronomy/Center for Data Driven Discovery, Caltech
UC Riverside, Fields Program, 2020-02-27

Outline

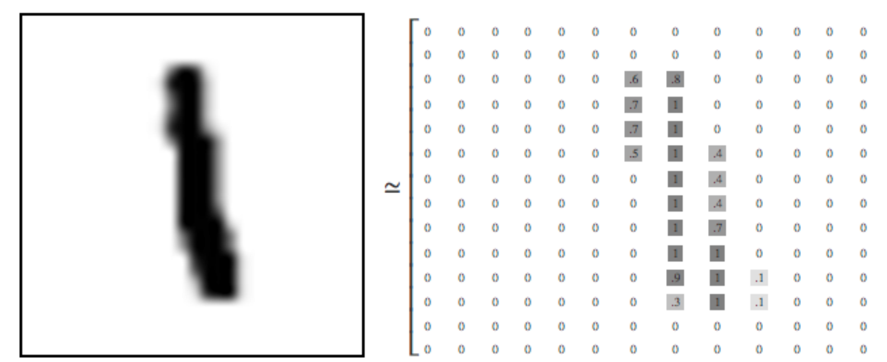
- Quick intro to neural networks
- Layers of a convolutional neural network
- Adversarial examples
- Generative and deconvolutional networks
- Example architectures
- Domain examples
- Non-image complex networks





Schematic of a logistic regression classifier.

By BruceBlaus - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=28761830
Quora: Sebastian Raschka



MNIST digits
0100000000

Real and artificial neurons

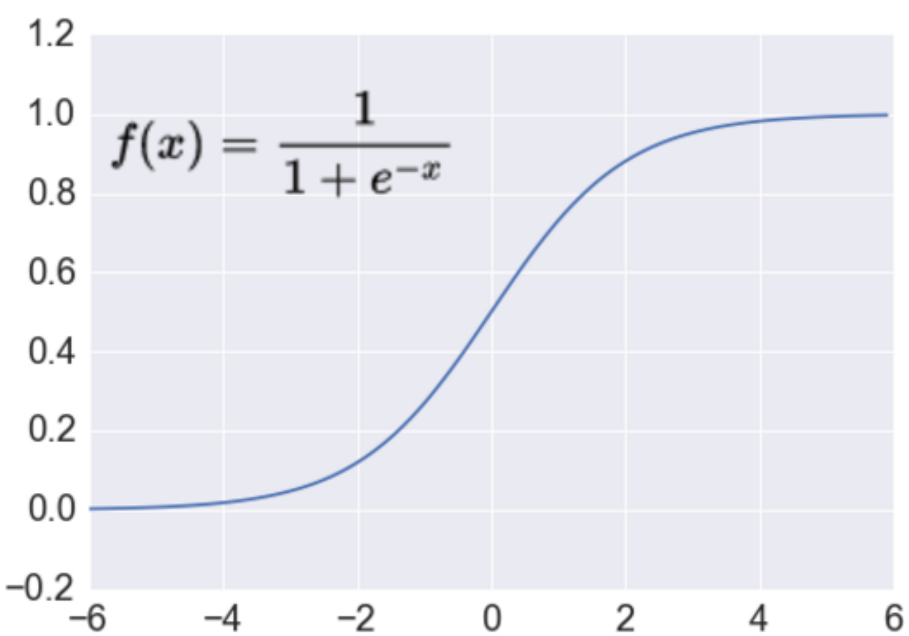
Softmax Regression

- This equation gives the probability of each classification
- Softmax regression first add up evidence of input being in a certain class and convert evidence into probabilities.
- There are 3 parts to this equation:
    1. Add weights (Wx)
    2. Add biases
    3. Softmax

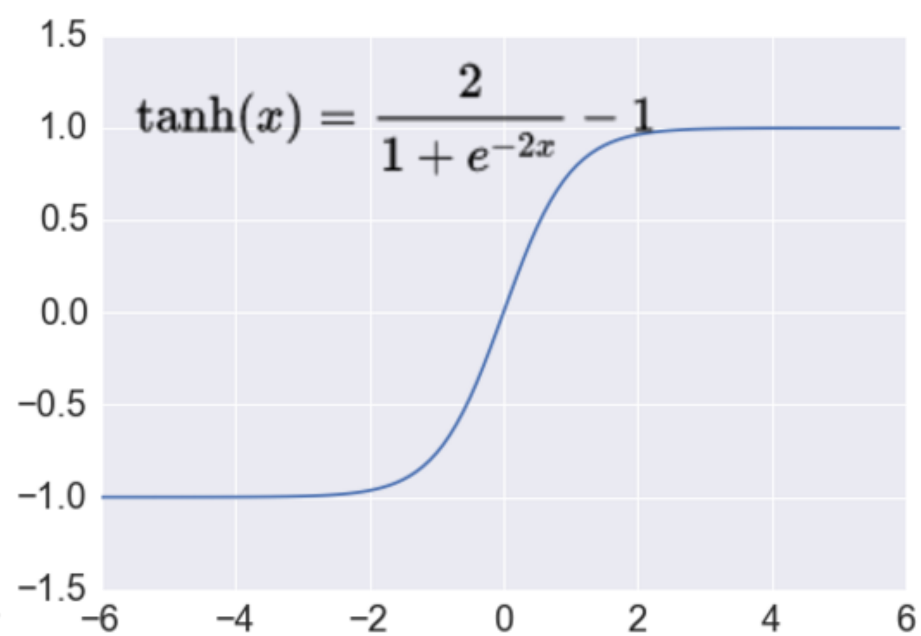
$$y = \mathrm{softmax}(Wx + b)$$

Rectified Linear Unit

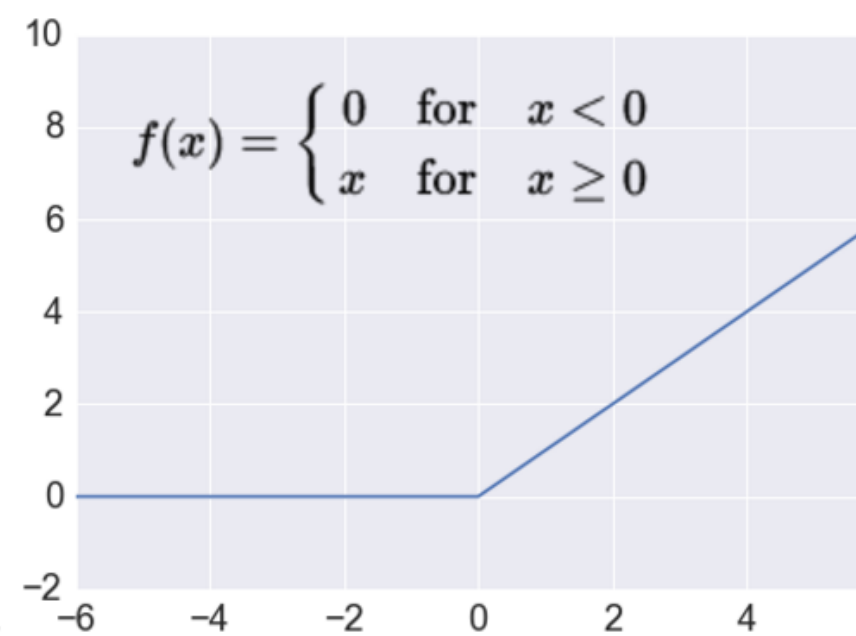
Sigmoid



TanH

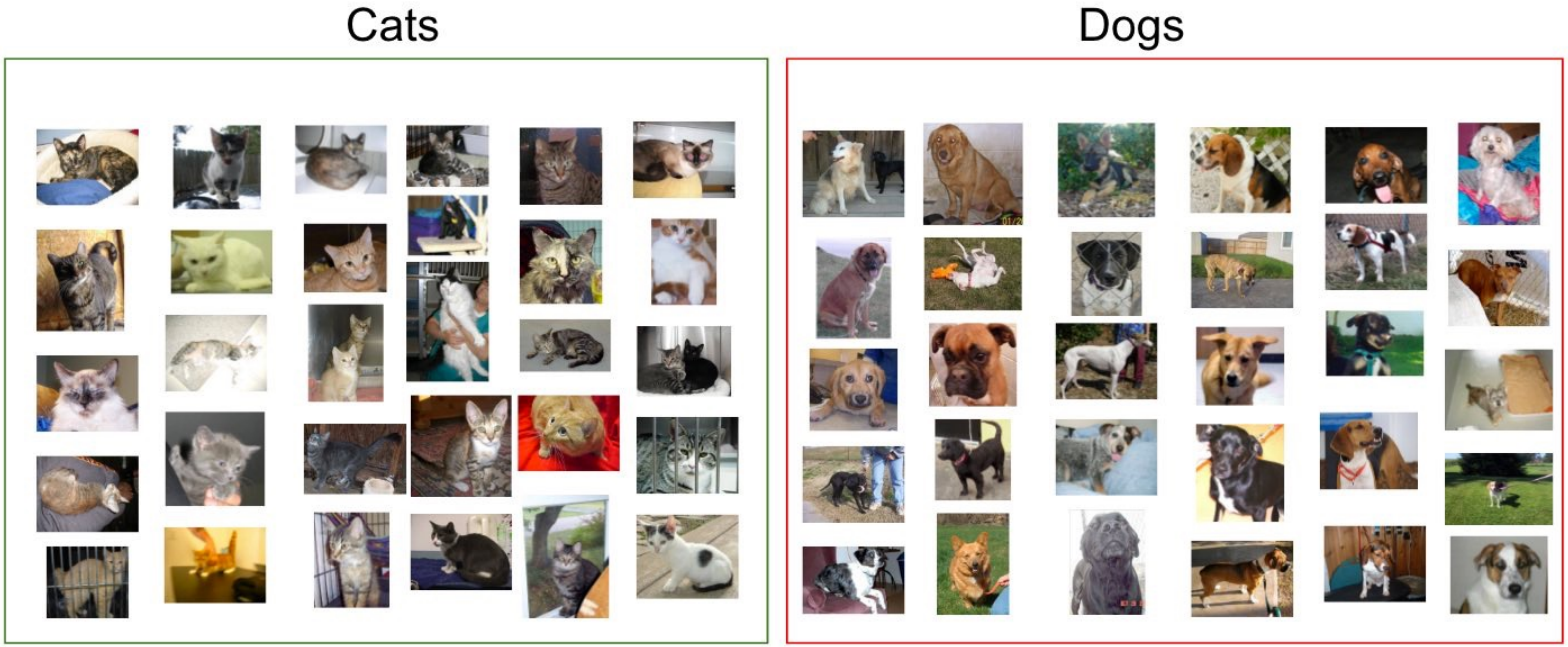


ReLU



Activation

SoftPlus, Sigmoid, …
Remapping/reshaping

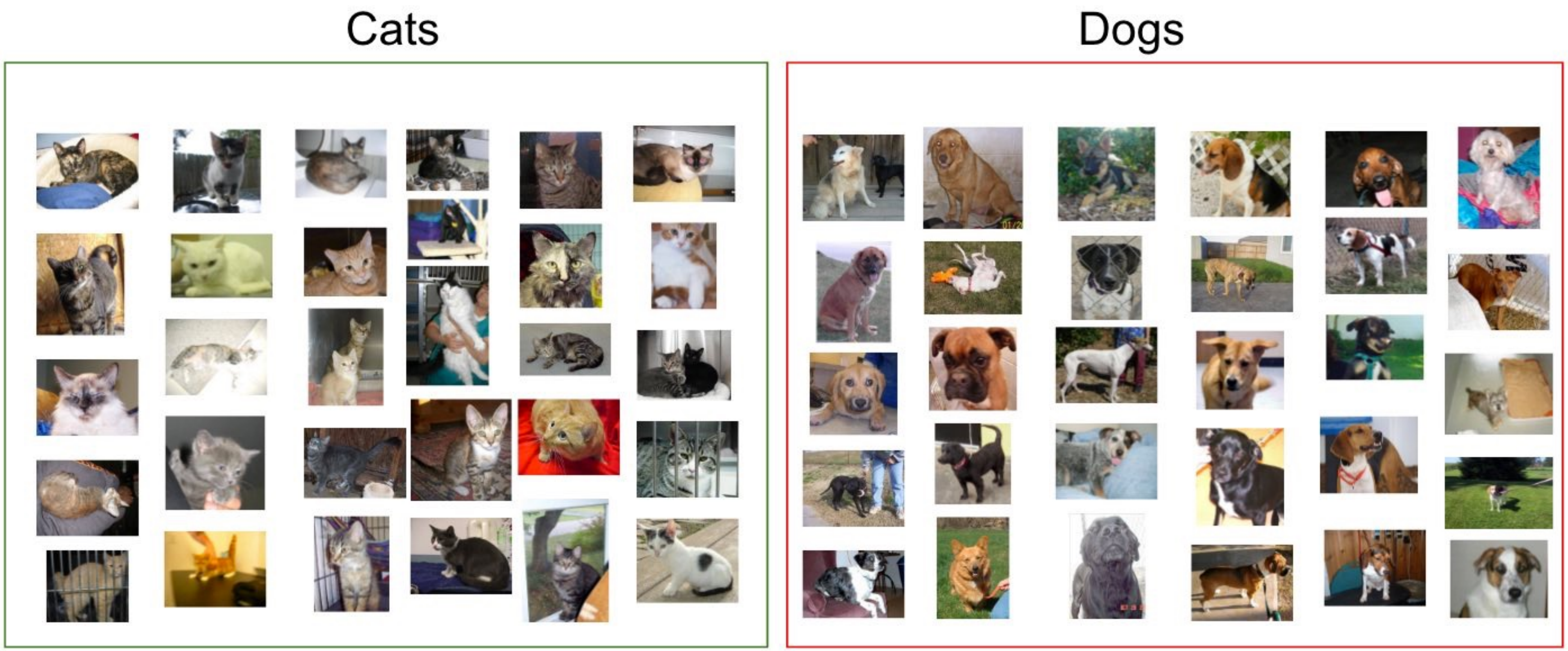


Sample of cats & dogs images from Kaggle Dataset

Traditional Machine Learning Flow

Deep Learning Flow

Adil Moujahid

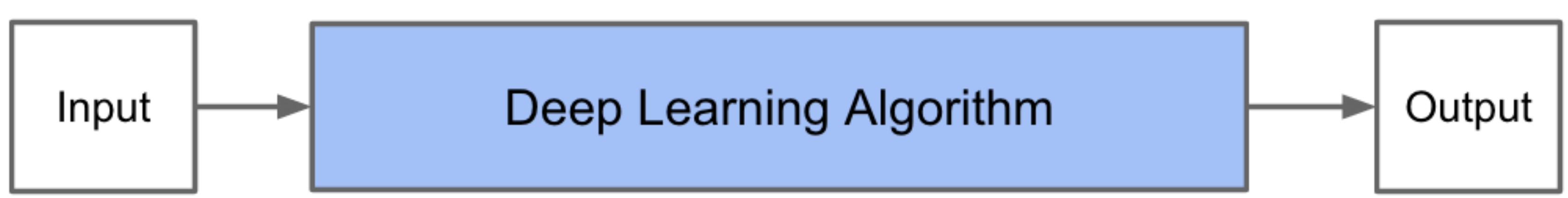


Sample of cats & dogs images from Kaggle Dataset

Promise:
Works better

Traditional Machine Learning Flow

Pitfall:
Blacker box

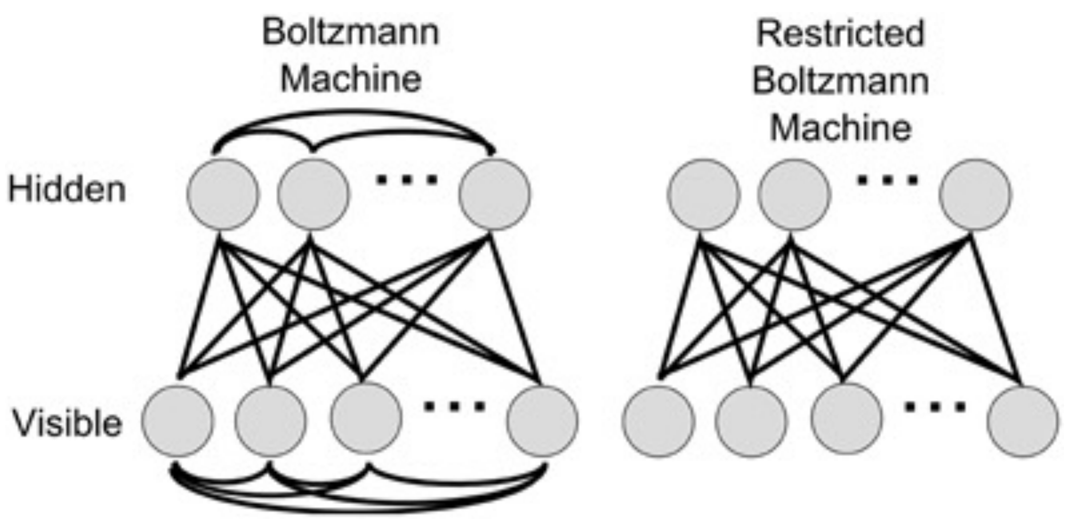


Deep Learning Flow

Adil Moujahid

Deep Learning

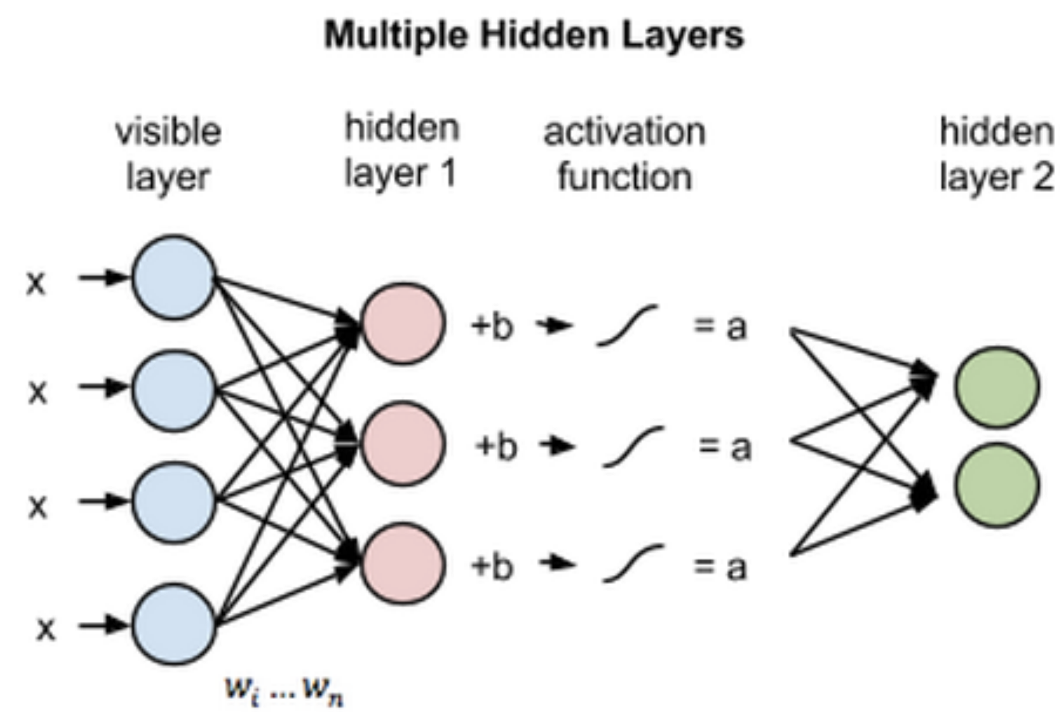
- Neural Networks in a new garb (but it works)



Restricted Boltzman Machines (RBMs)

Hinton

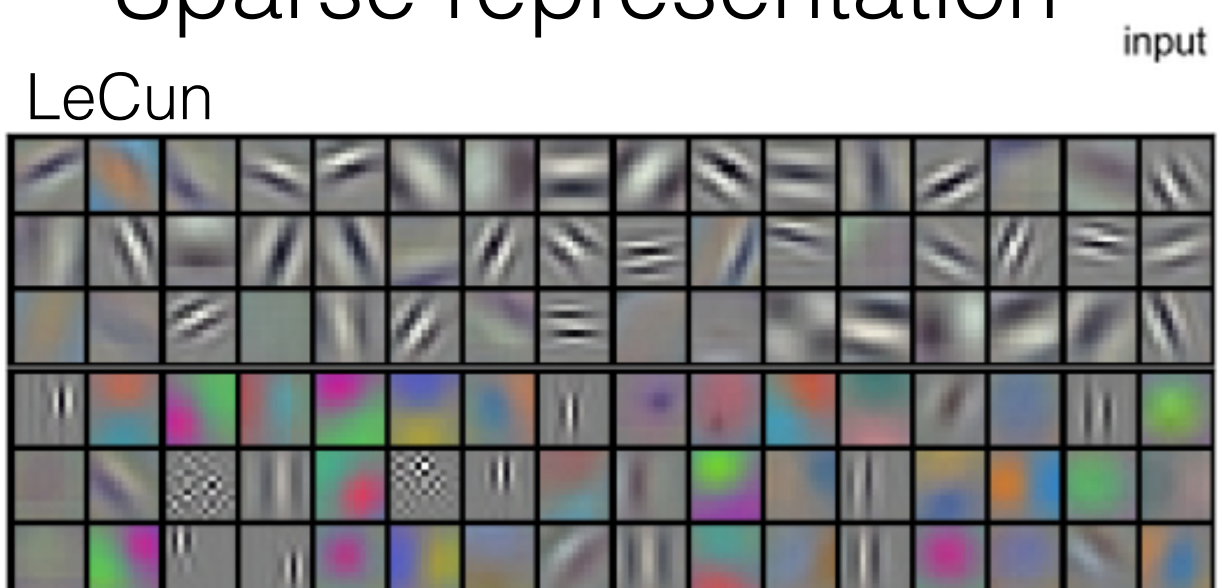
Hidden layers can not interconnect



Simple FeedForward

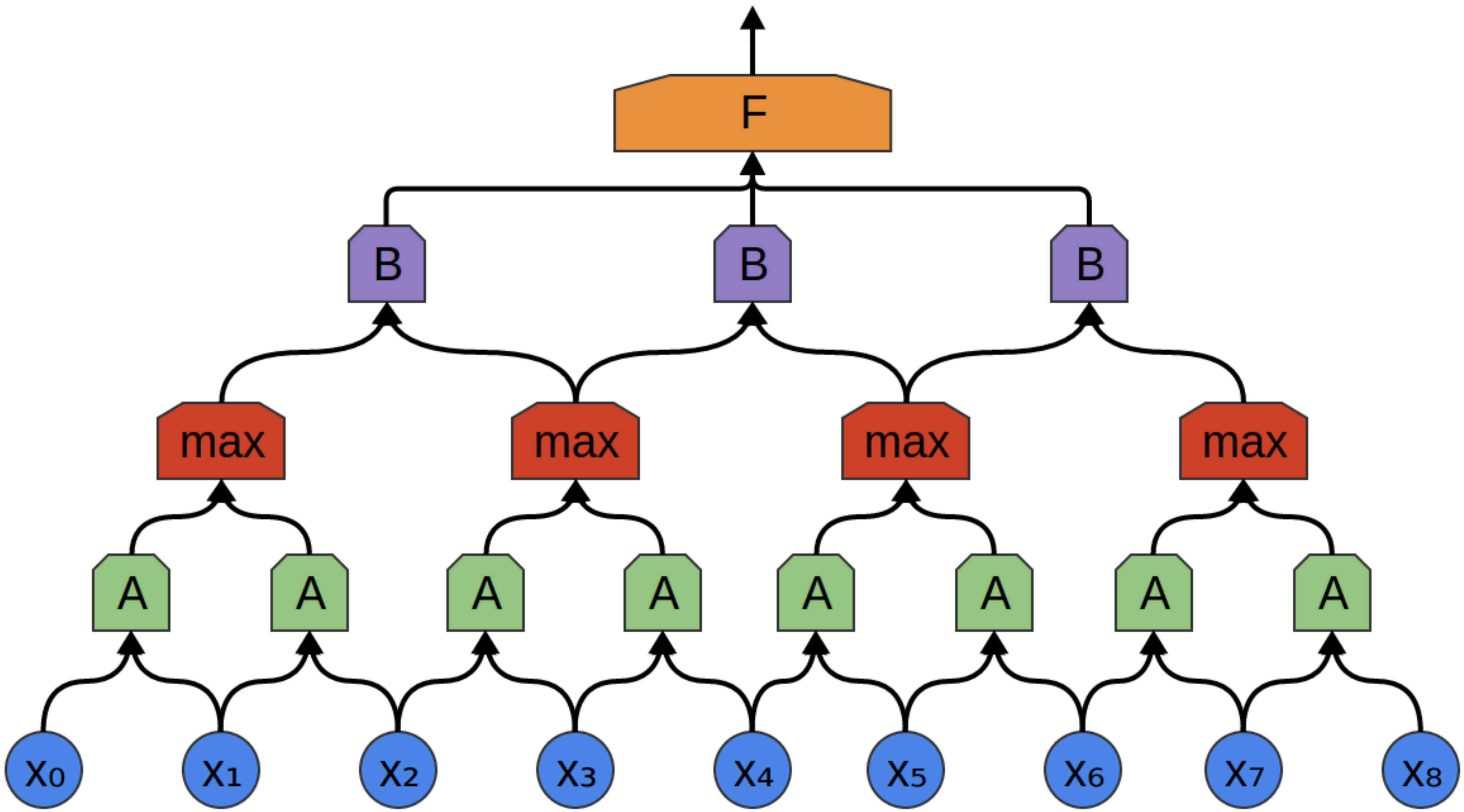
Sparse representation

LeCun

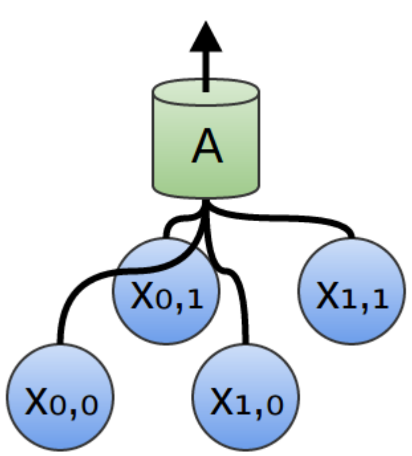


Joint probability: p(x|a) and p(a|x) expressed as the shared weights

Krizehvsky et al. 2012



conv + pool + conv + connected

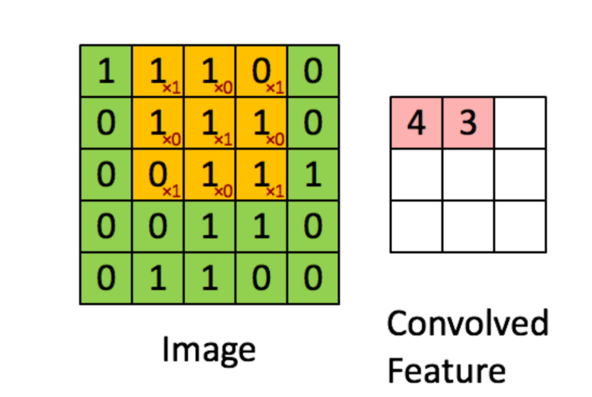


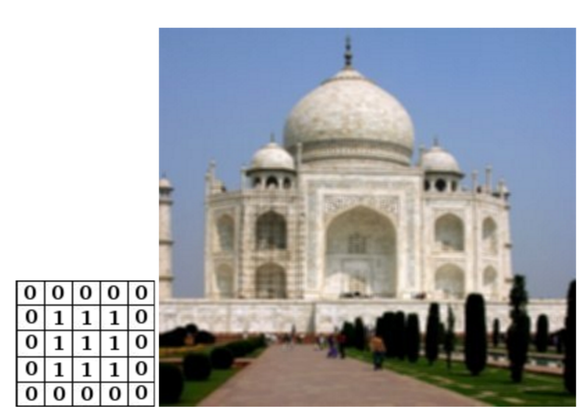
http://colah.github.io/posts/2014-07-Conv-Nets-Modular/

2D version of convolution

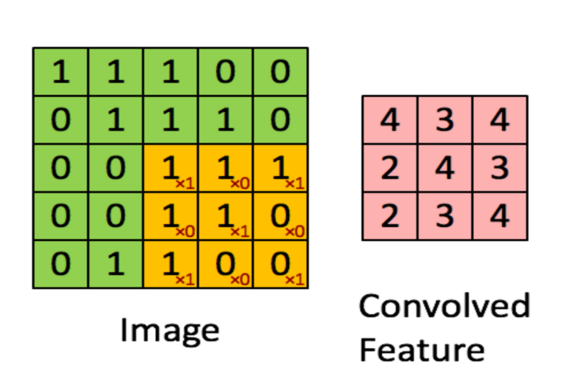
Enhancing angled edges

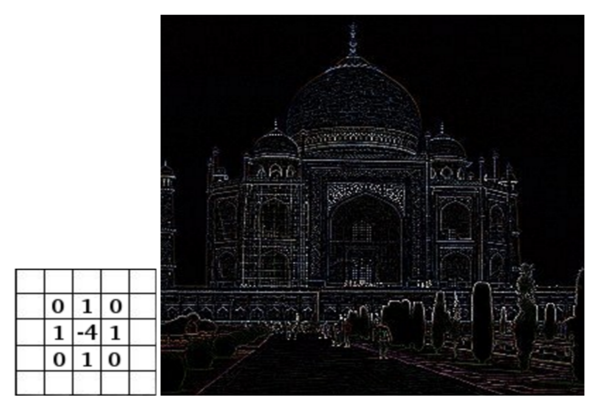
Here's an example of how to make a convolution layer with 3x3 filter.
You run across the matrix to get convolved feature of 3x3.





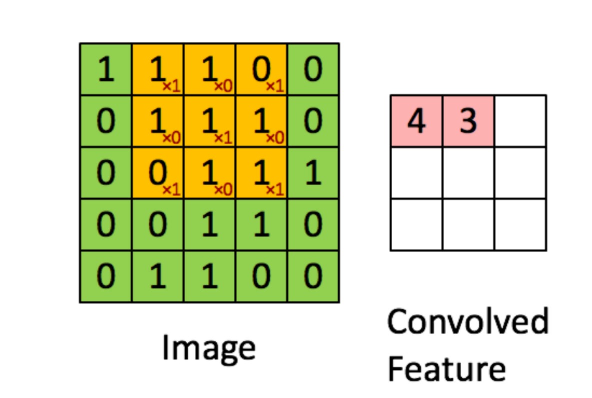
- This will be the resulting convolution layer once matrix is complete. In this example, the layer will detect edges.

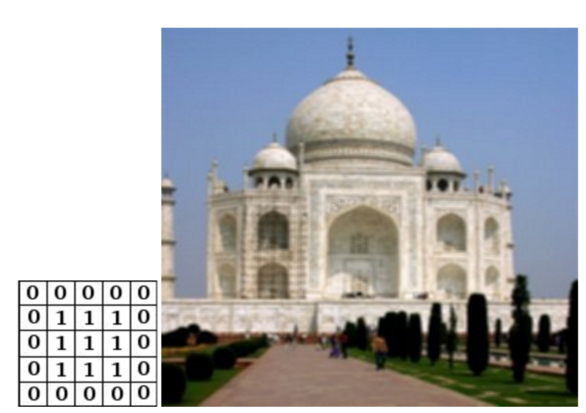




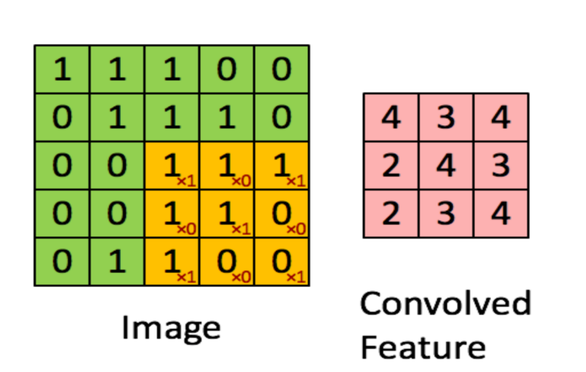
Enhancing angled edges

Here's an example of how to make a convolution layer with 3x3 filter.
You run across the matrix to get convolved feature of 3x3.

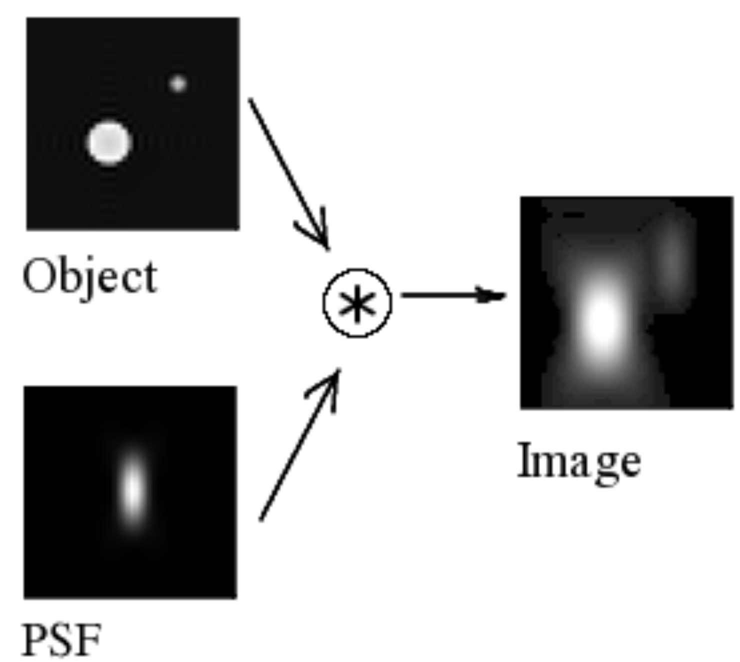




- This will be the resulting convolution layer once matrix is complete. In this example, the layer will detect edges.



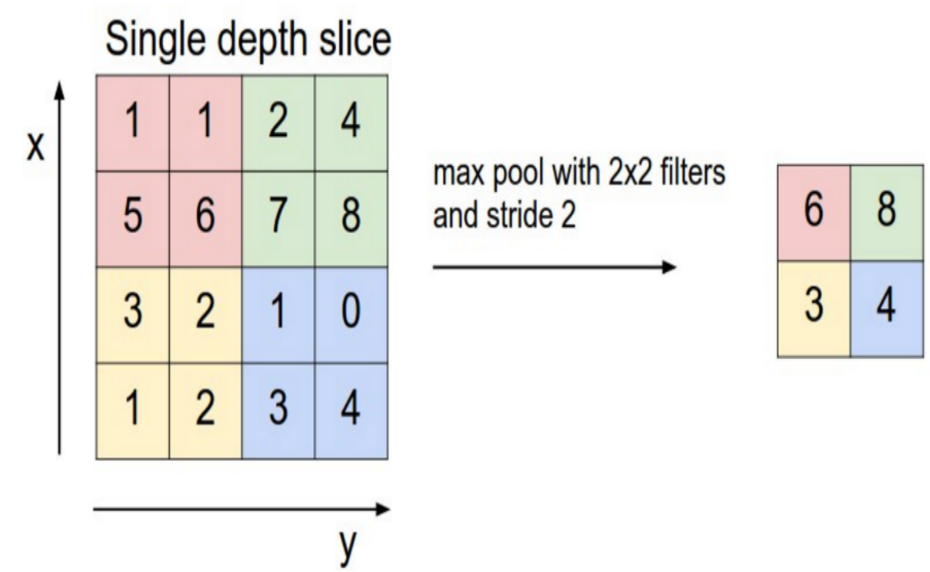
Convolution in Astronomy



Pooling Layers (max pool)

Pooling layers are applied after convolution layers. This is a subsample of your convolution layer.

- Fixed size output matrix
- Reduces output dimensionality with the most salient information

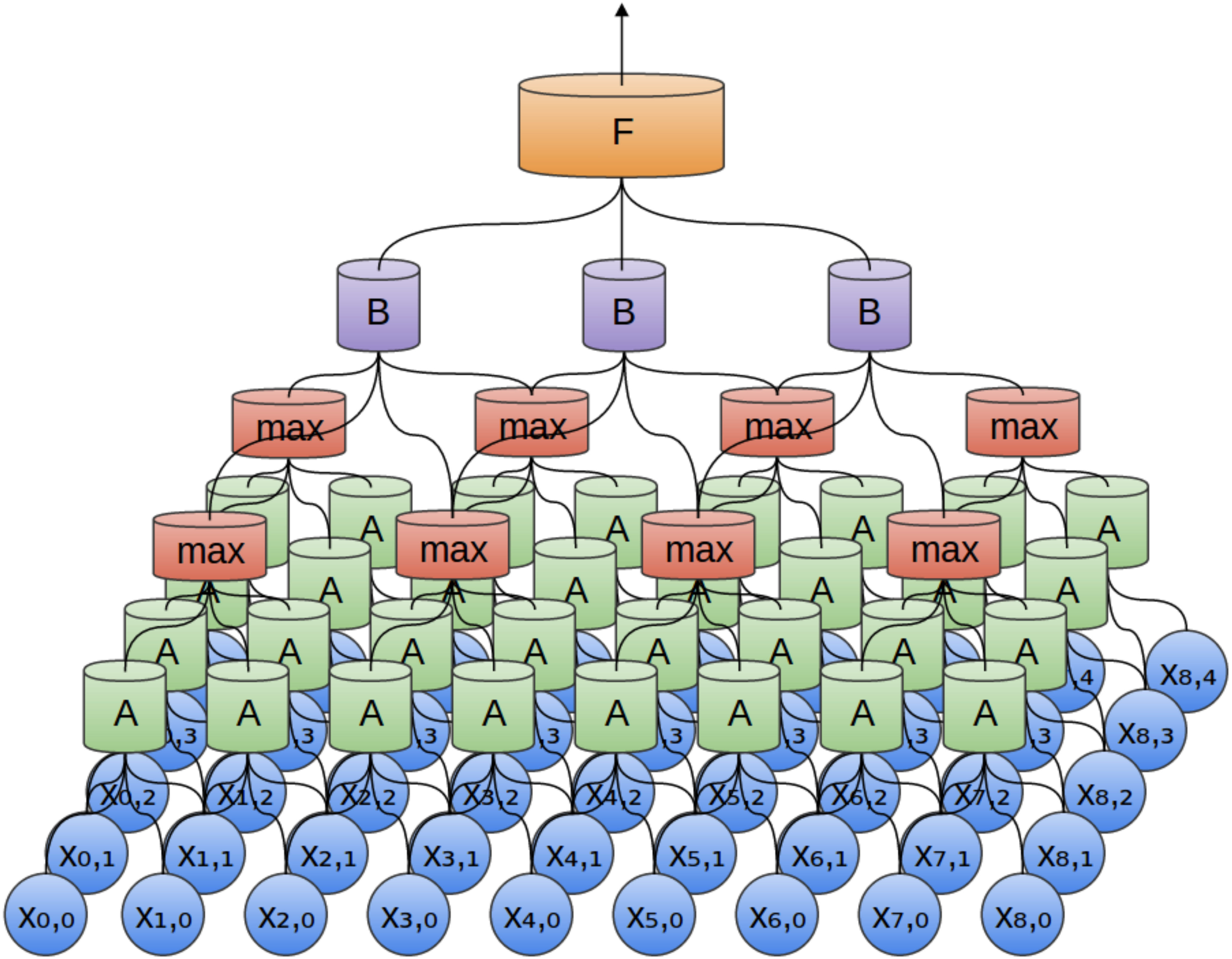


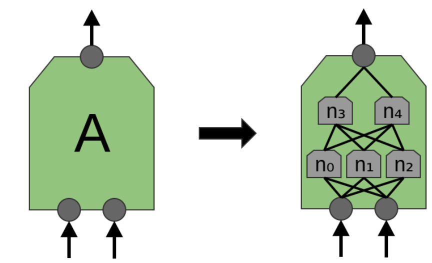
Hyperparameters

Our model has a depth of 3. Pooling blocks are 2x2, with stride of 1 and 0 padded for our model.

Three hyperparameters control the size of the output volume of the convolutional layer: the **depth, stride** and **zero-padding**.

- **Depth** of convolution layers. For example, if the first Convolutional Layer takes the raw image as input, then another may activate edges, or color.
- **Stride** When the stride is 1, a new depth column of neurons is allocated to spatial positions only 1 spatial unit apart.
- **Zero padding** provides control of the output volume spatial size.





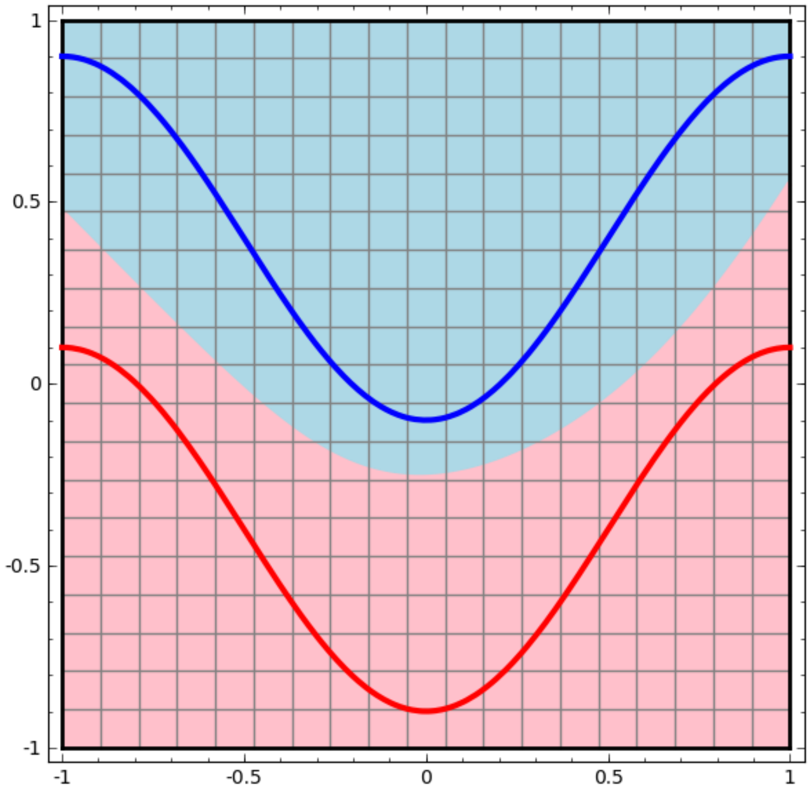
http://colah.github.io/posts/2014-07-Conv-Nets-Modular/

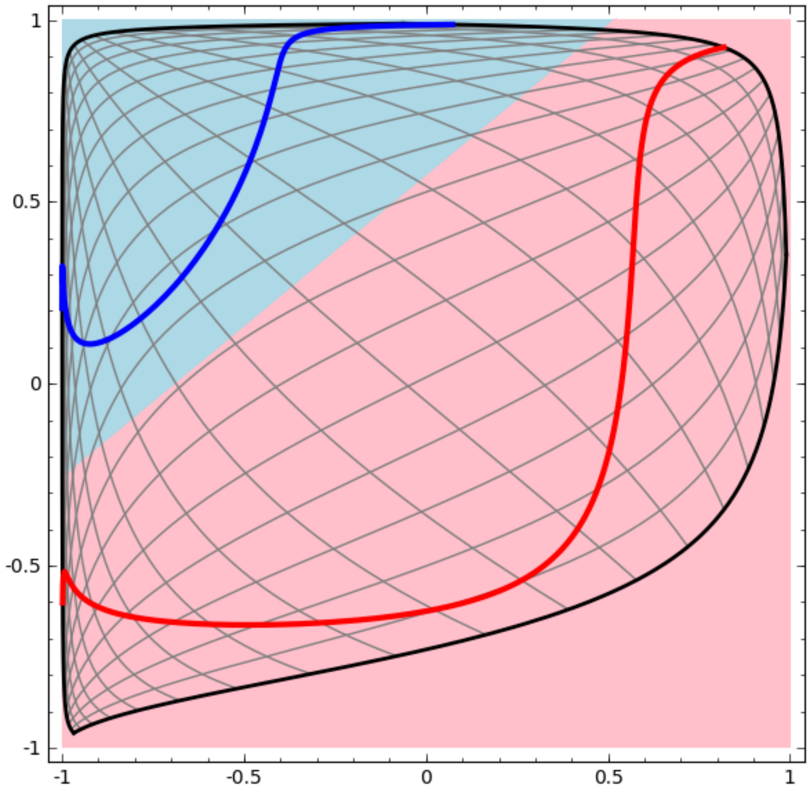
Final layers are fully connected

Several libraries available

- Caffe: http://caffe.berkeleyvision.org/
- Tensorflow: https://www.tensorflow.org/
- Theano: http://deeplearning.net/software/theano/
- PyTorch: https://pytorch.org/
- MxNet: https://mxnet.apache.org/

abstractions, instant gratification

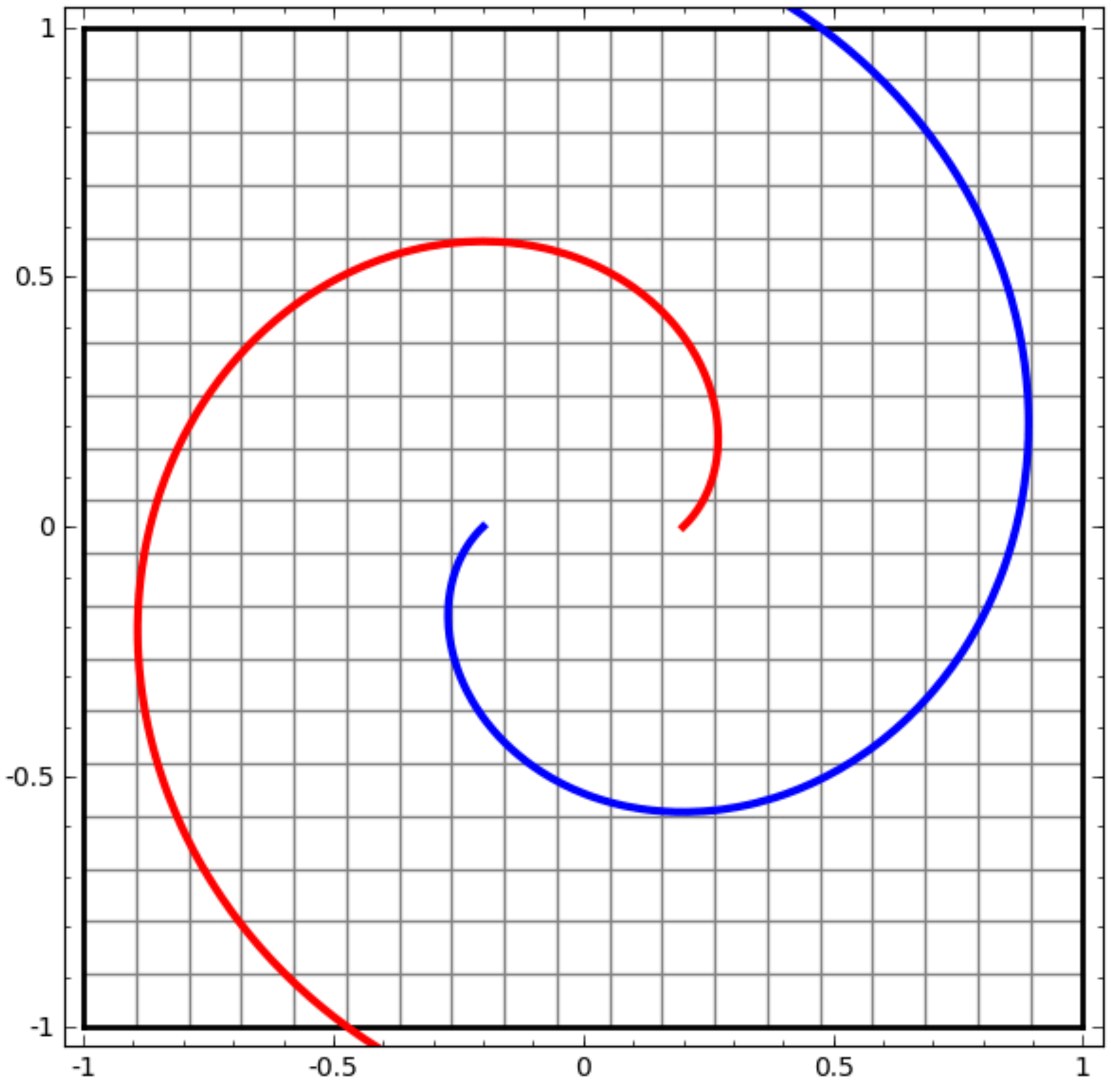


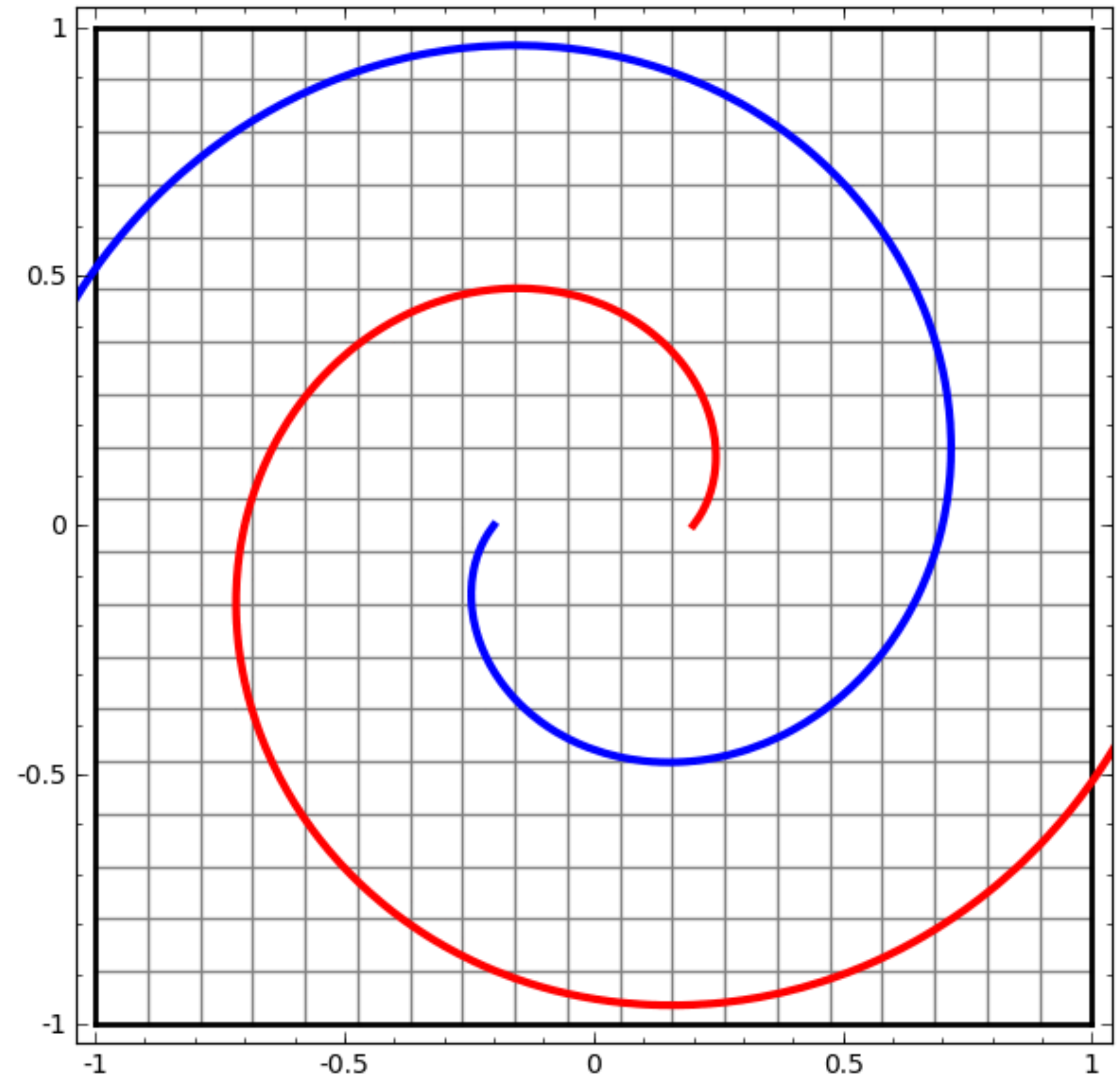


Mapping in order to linearly separate clusters

http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Disentangling with multiple layers



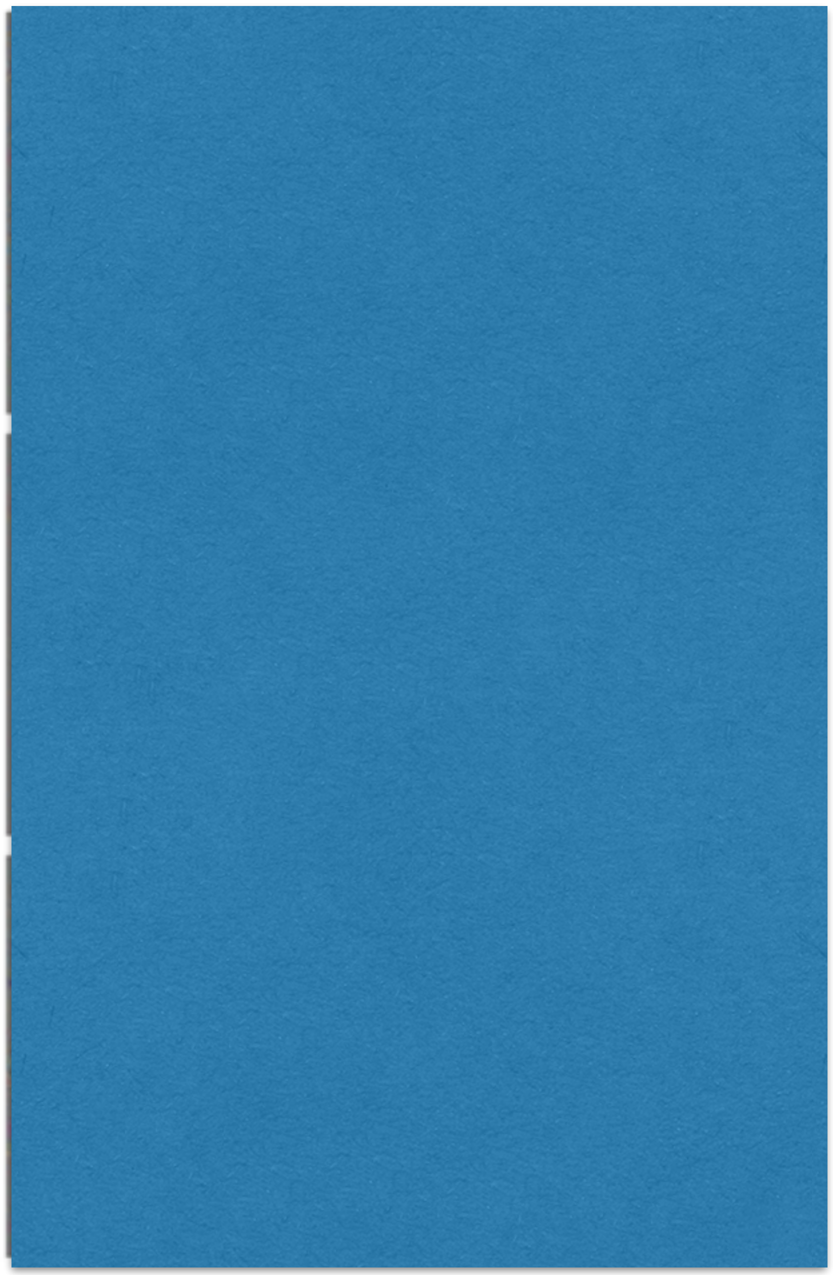


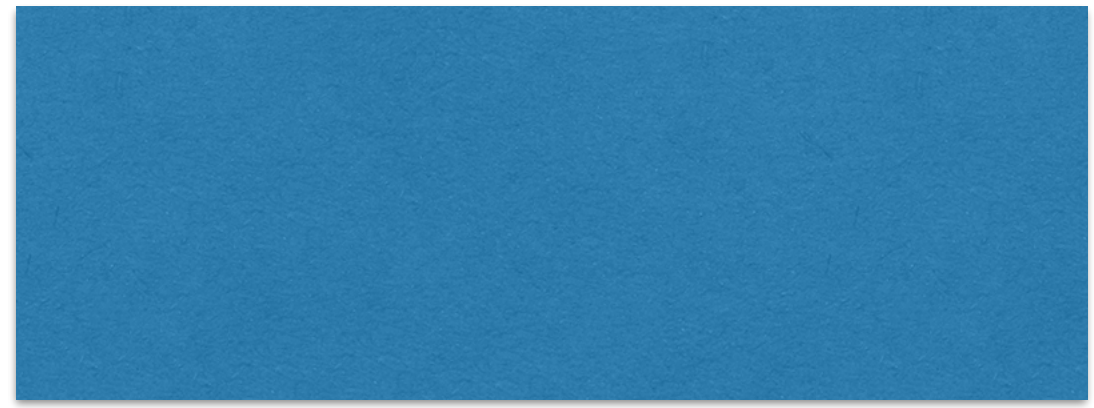
http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

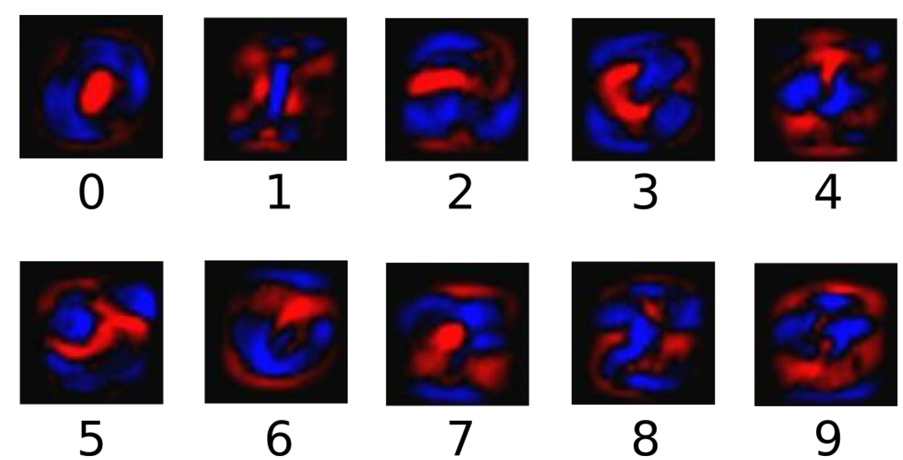
What bird is that?

or: what features is my deep network using?

Including adversarial examples during training



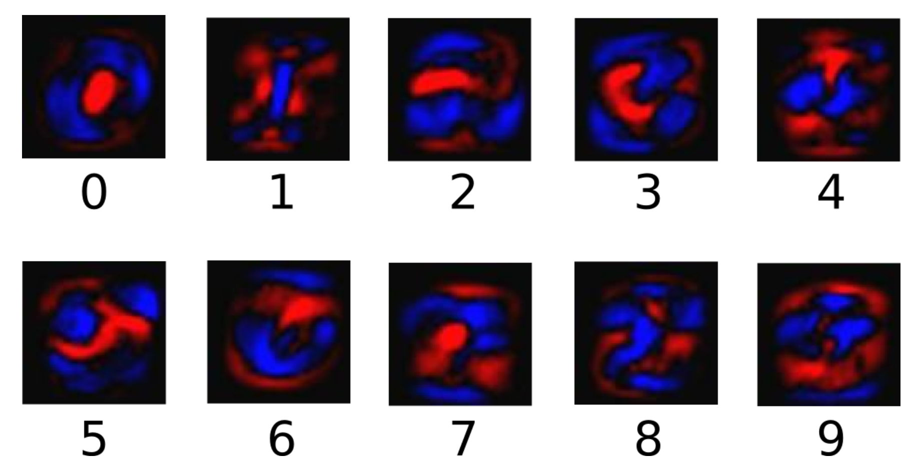




MNIST digits significance map

https://arxiv.org/pdf/1312.6199v4.pdf

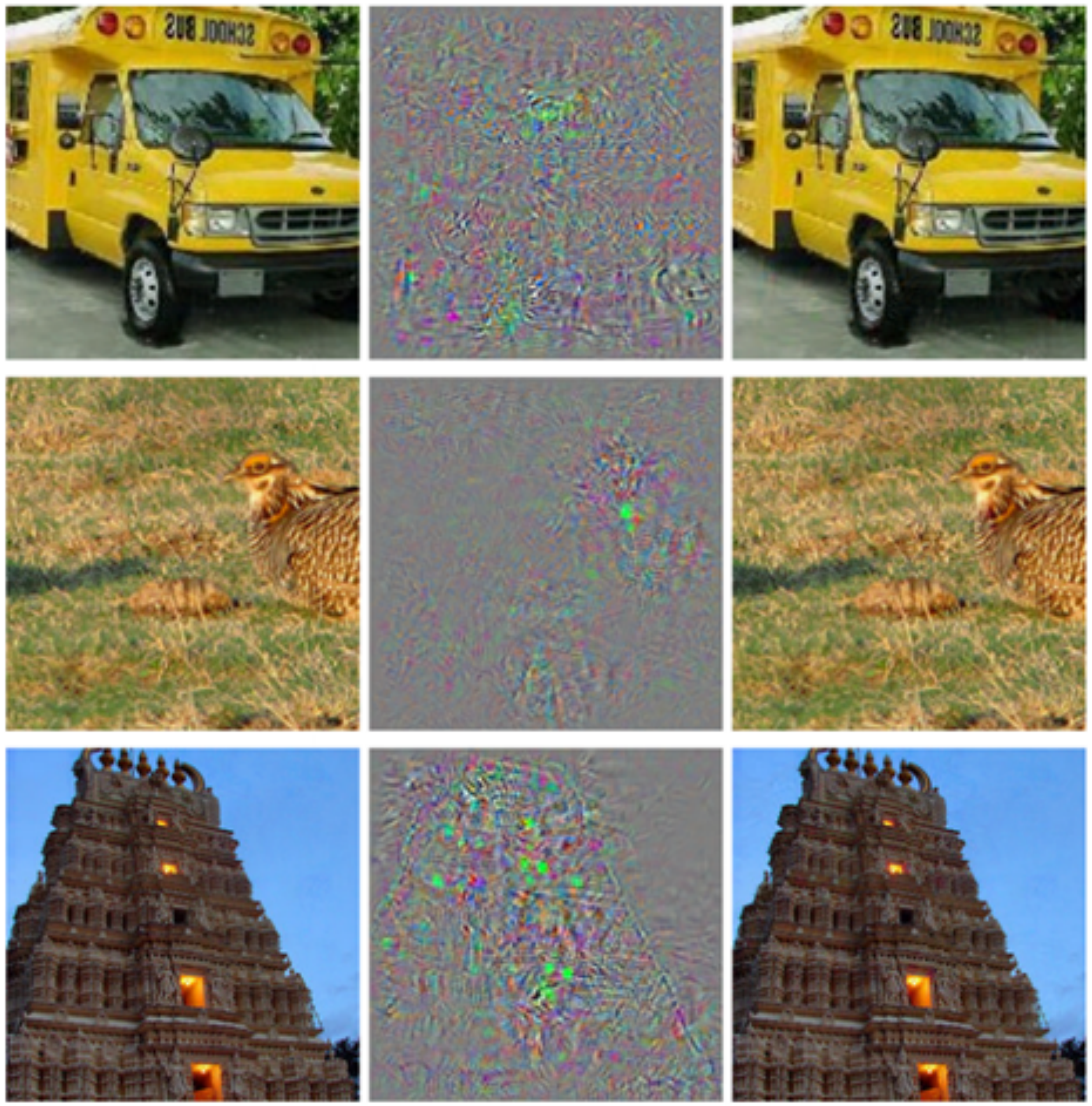
Including adversarial examples during training



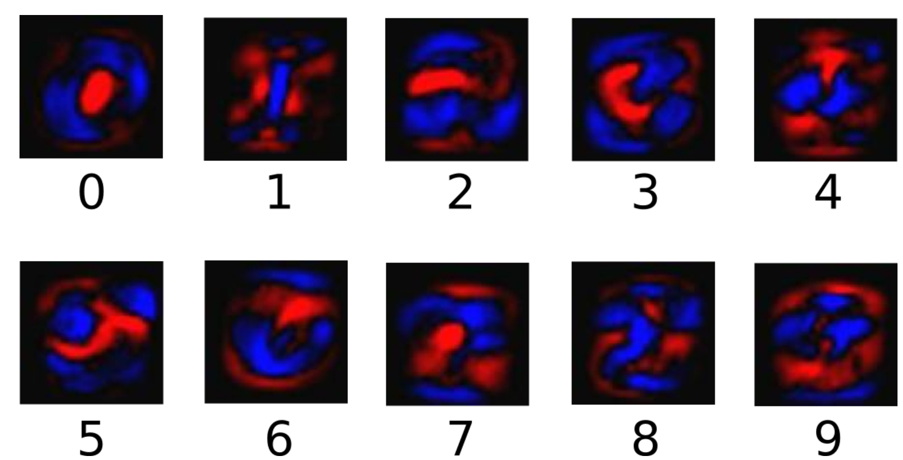
MNIST digits significance map

https://arxiv.org/pdf/1312.6199v4.pdf

Including adversarial examples during training



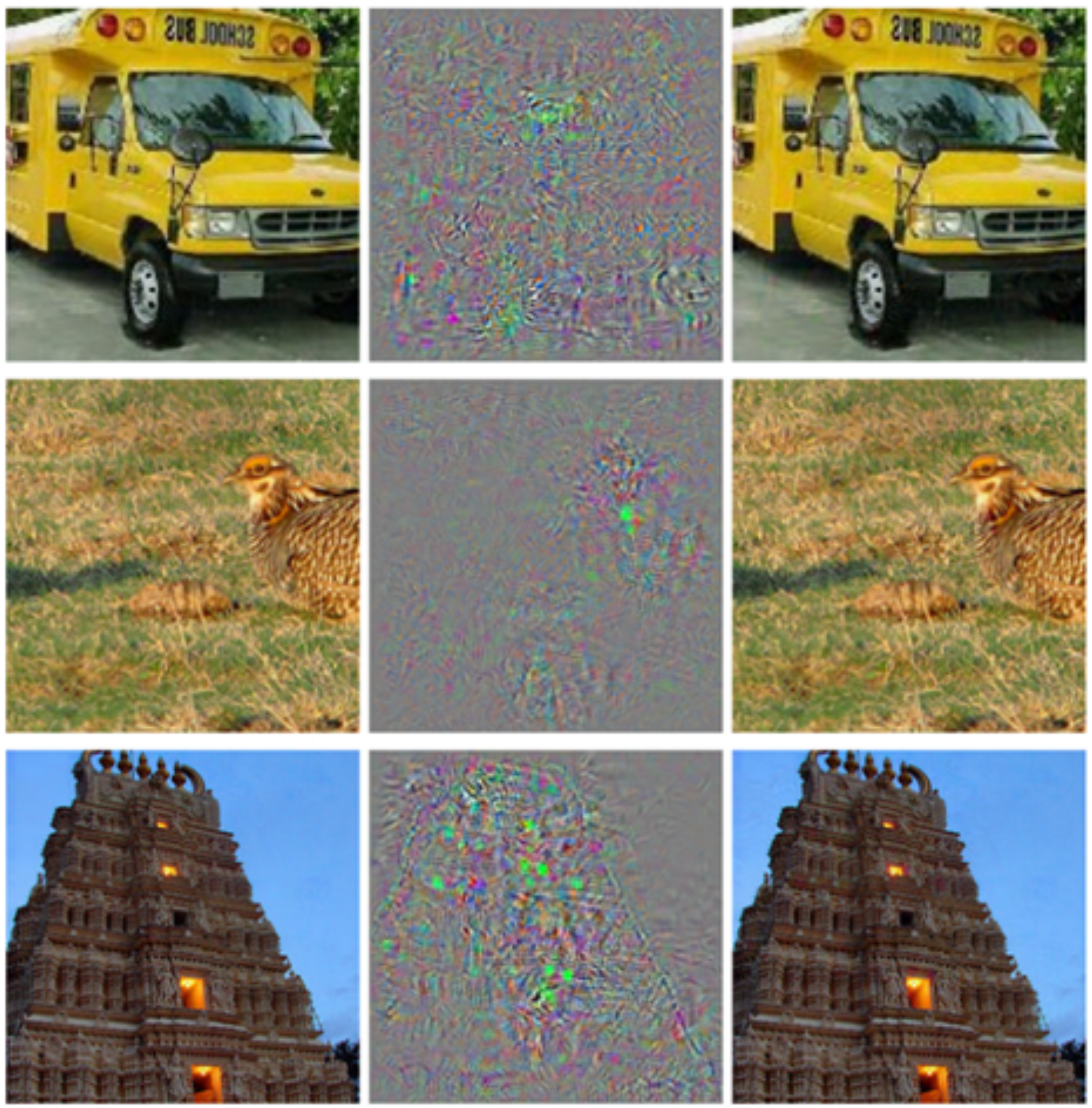
The images in the left most column are correctly classified examples. The middle column represents the distortion between the left and right images. The images in the right most column are predicted to be of the class ostrich! Even though the difference between the images on the left and right is imperceptible to humans, the ConvNet makes drastic errors in classification.



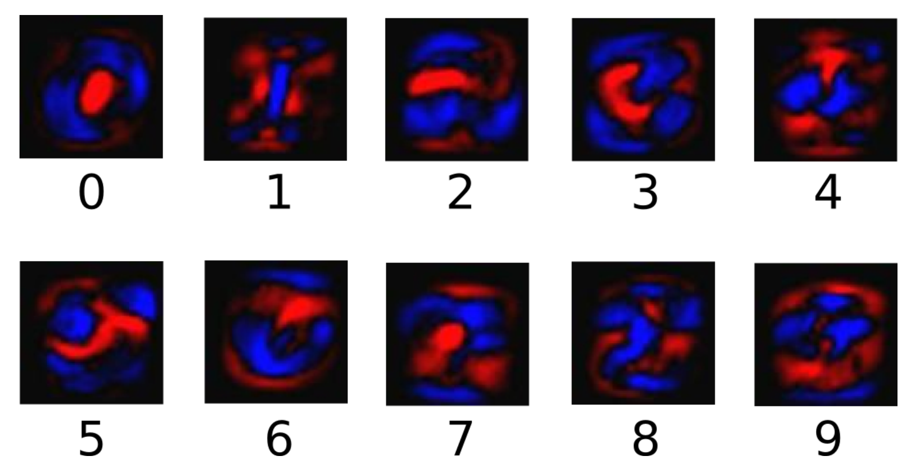
MNIST digits significance map

https://arxiv.org/pdf/1312.6199v4.pdf

Including adversarial examples during training



The images in the left most column are correctly classified examples. The middle column represents the distortion between the left and right images. The images in the right most column are predicted to be of the class ostrich! Even though the difference between the images on the left and right is imperceptible to humans, the ConvNet makes drastic errors in classification.



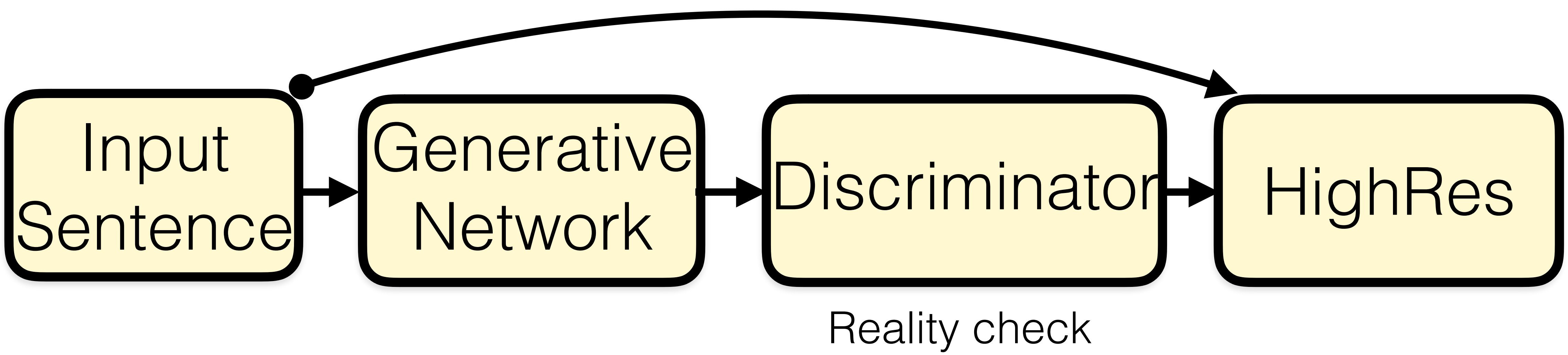
MNIST digits significance map

https://arxiv.org/pdf/1312.6199v4.pdf

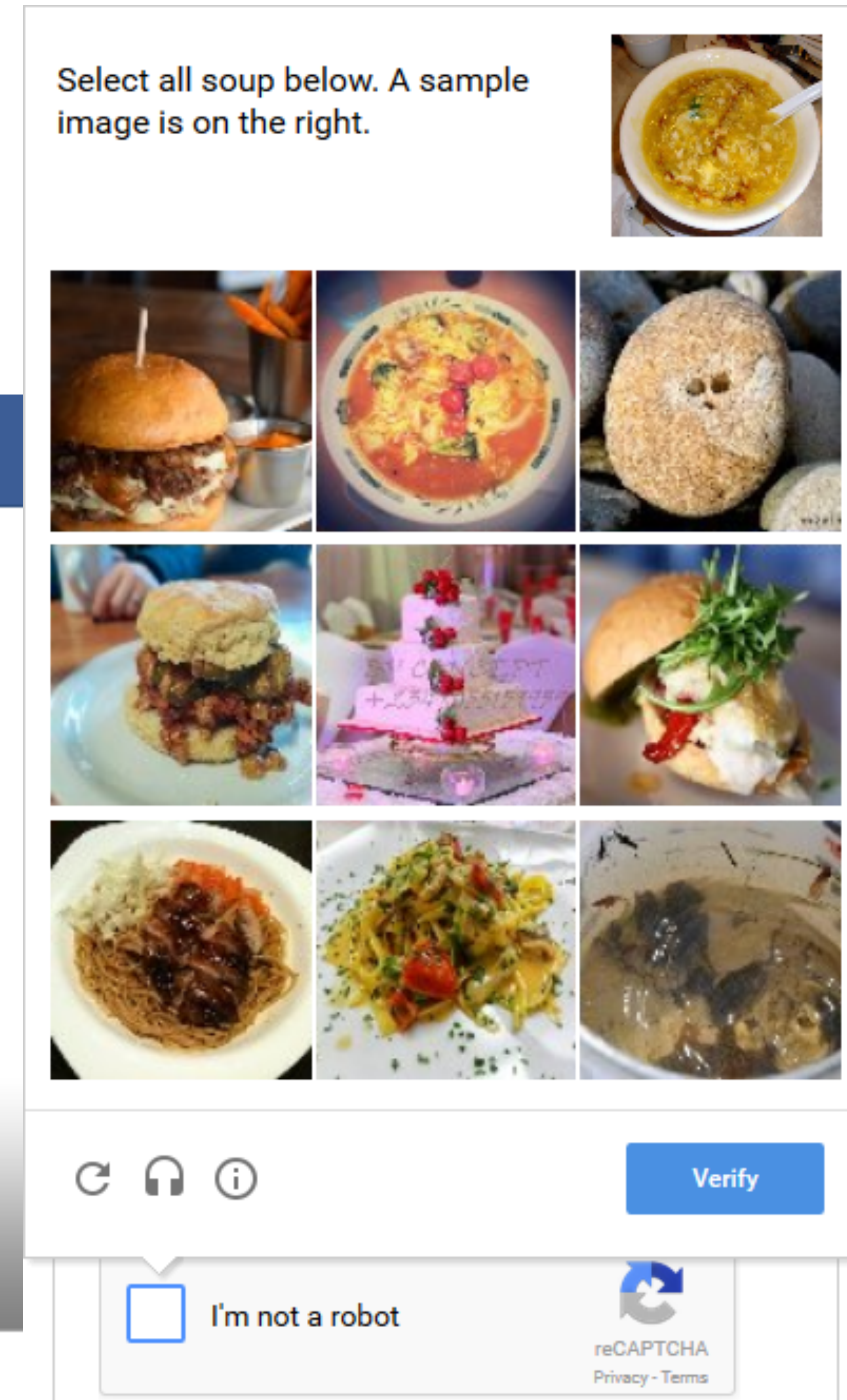
Pitfall: Overlearning

Generative Adversarial Networks

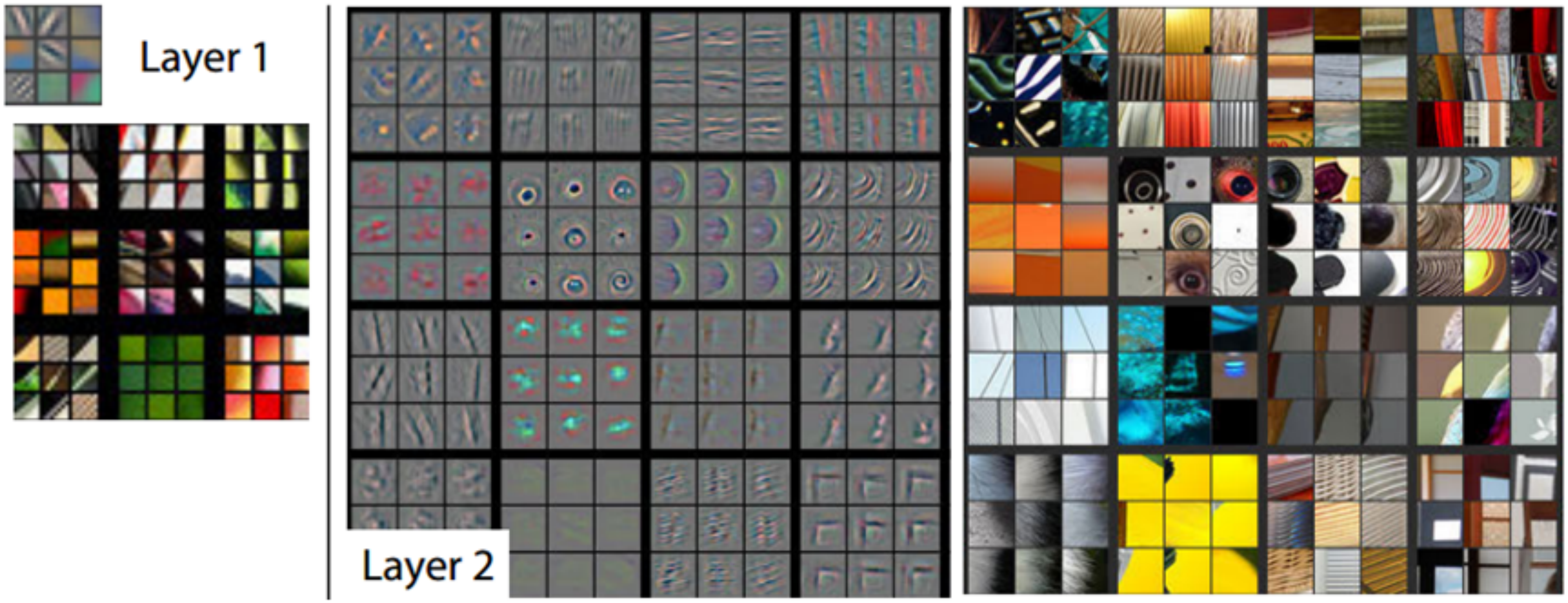
Zhang et al. 2016



Labels are everywhere

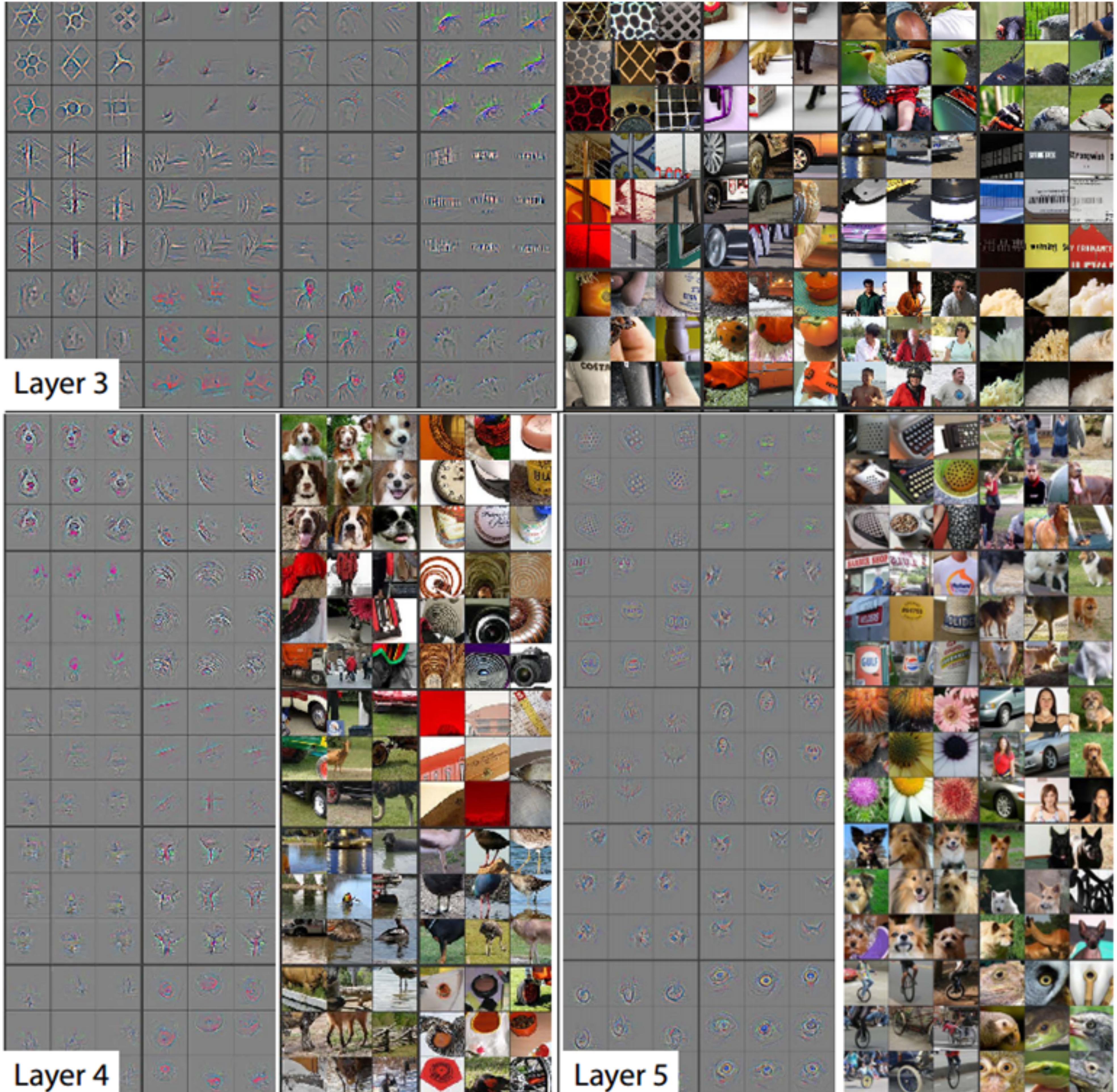


deconvnets

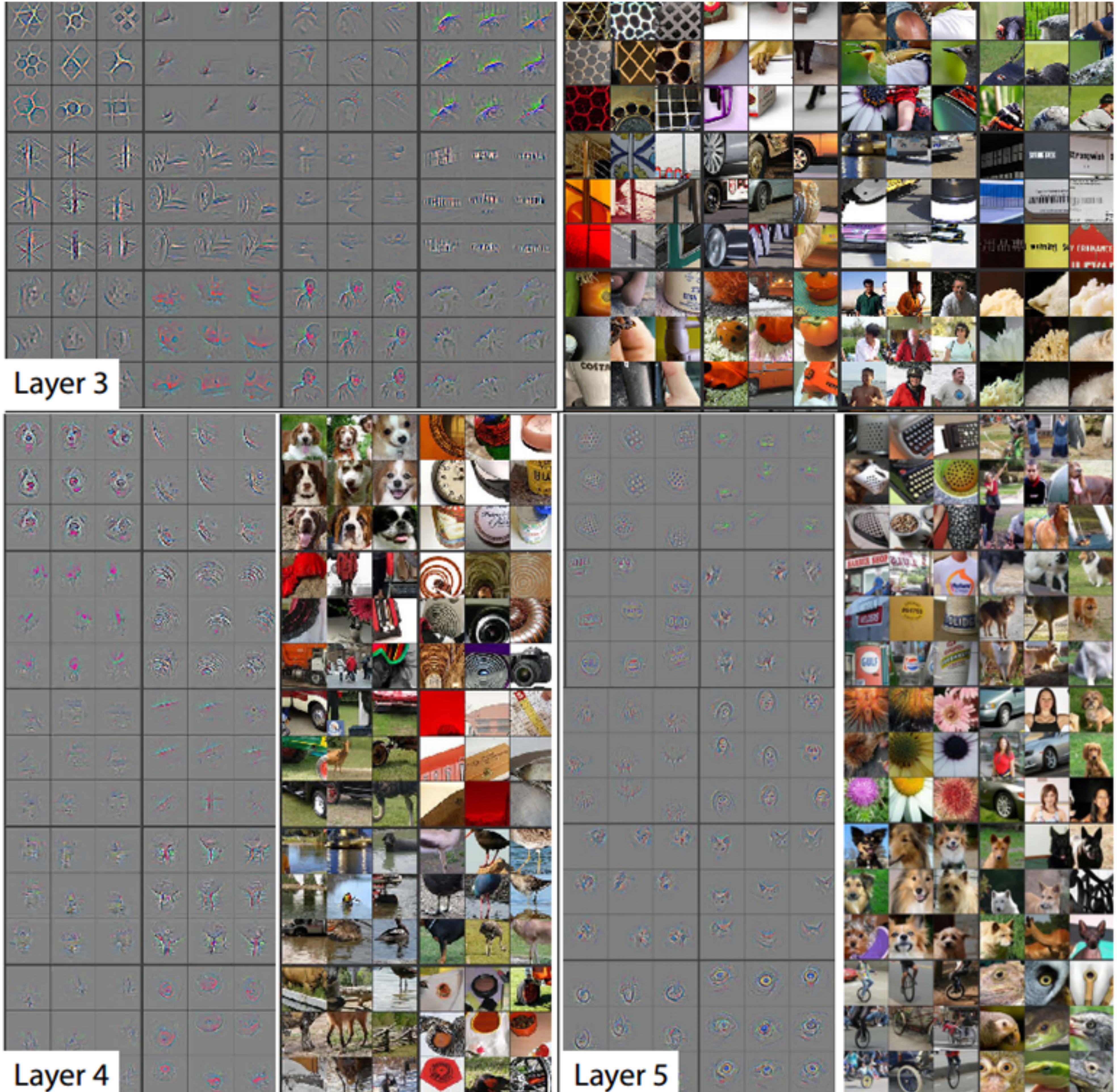


Visualizations of Layer 1 and 2. Each layer illustrates 2 pictures, one which shows the filters themselves and one that shows what part of the image are most strongly activated by the given filter. For example, in the space labled Layer 2, we have representations of the 16 different filters (on the left)

https://arxiv.org/pdf/1311.2901v3.pdf



 Visualizations of Layers 3, 4, and 5 https://arxiv.org/pdf/1311.2901v3.pdf

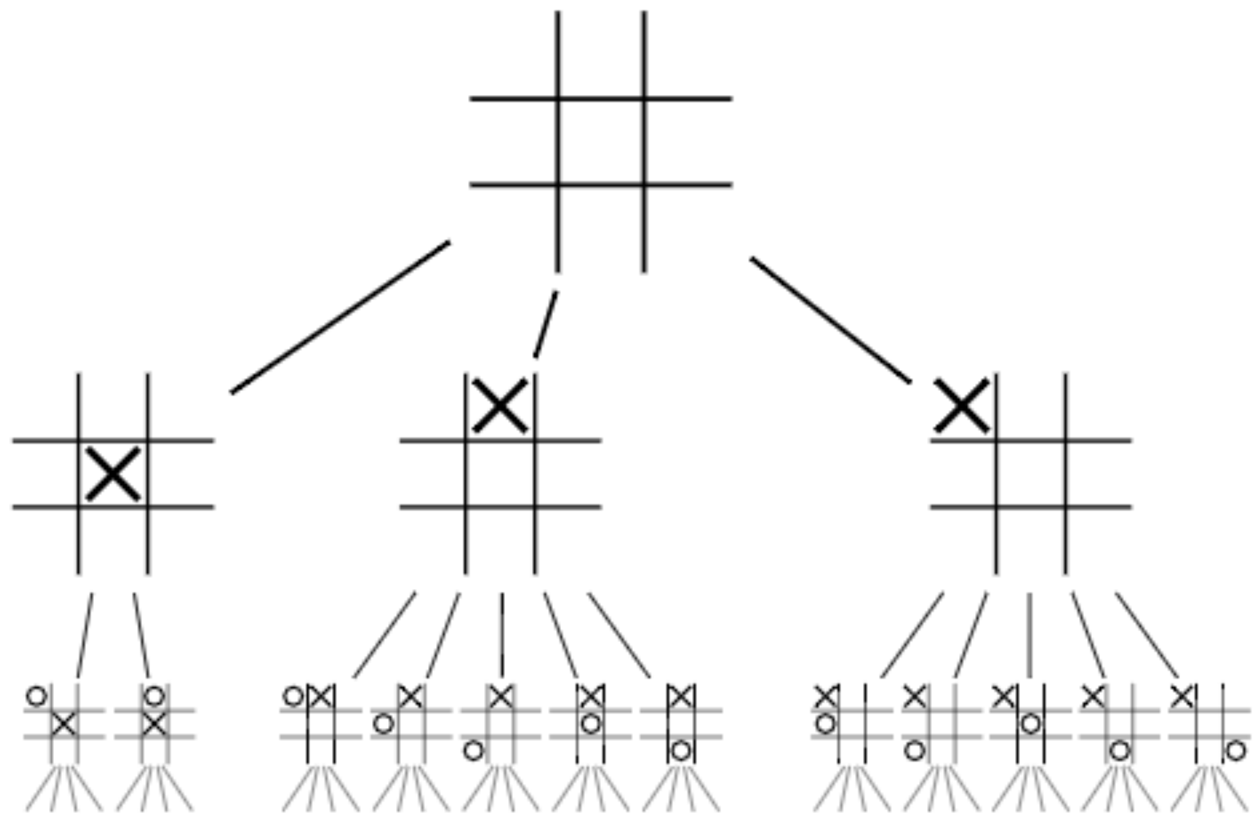


Promise: New Features

Ashish Ma Visualizations of Layers 3, 4, and 5

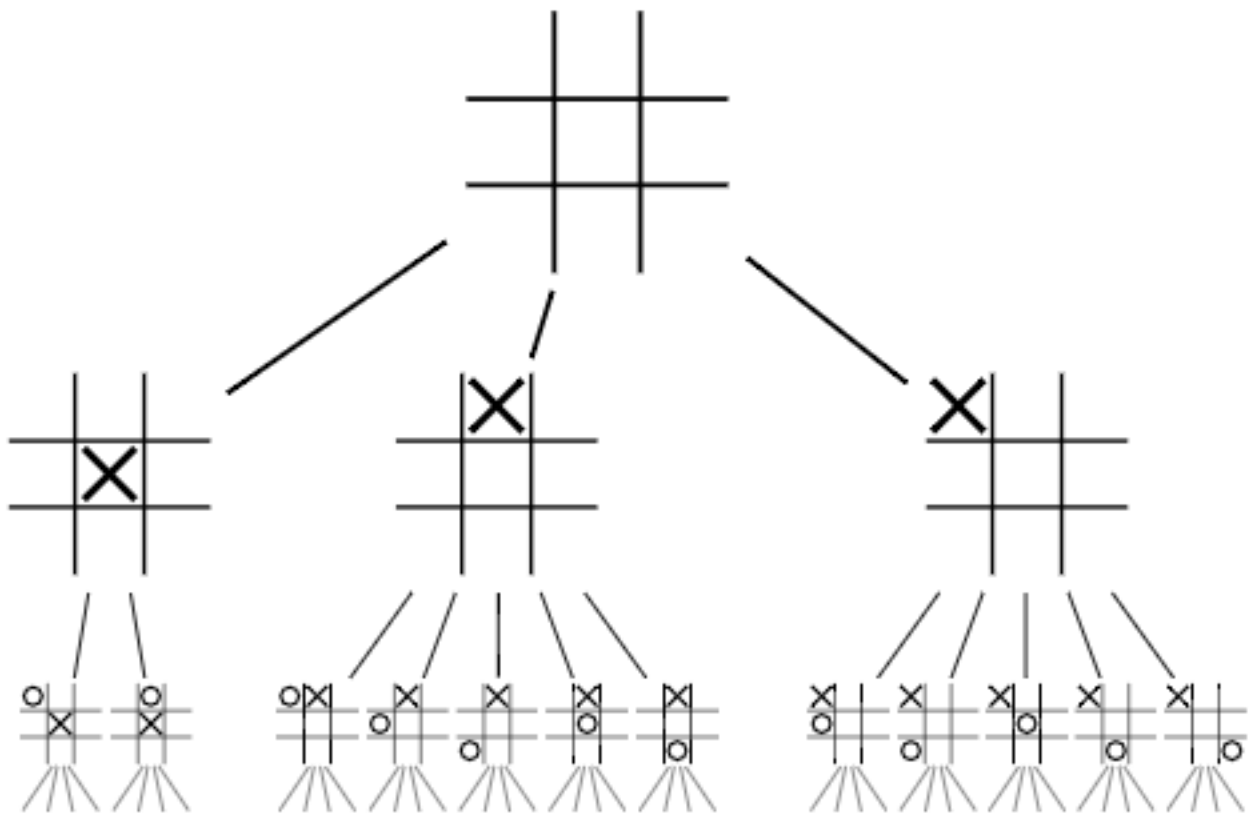
https://arxiv.org/pdf/1311.2901v3.pdf

Tic-Tac-Toe with Deep Learning?



Wikipedia

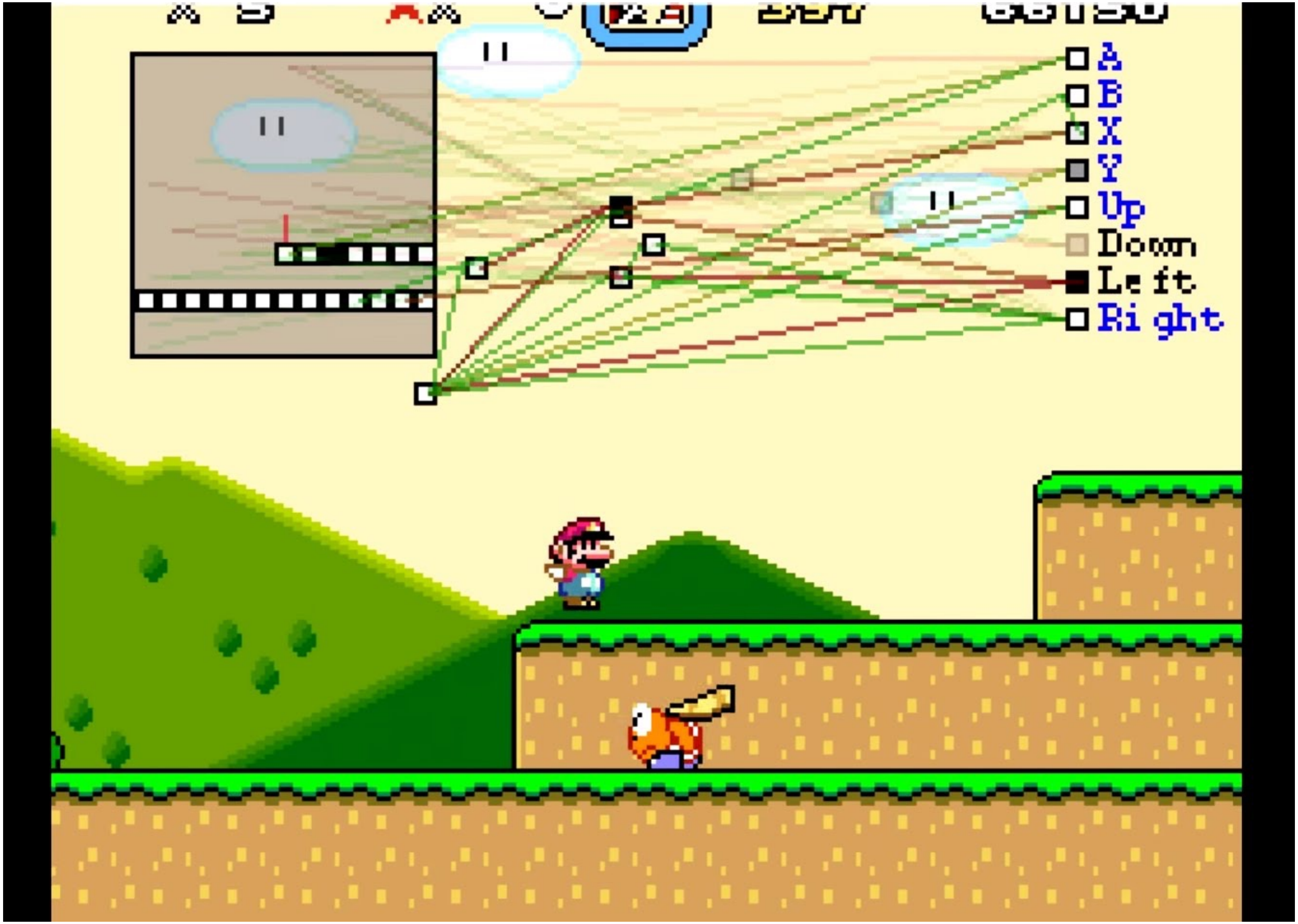
Tic-Tac-Toe with Deep Learning?



Wikipedia

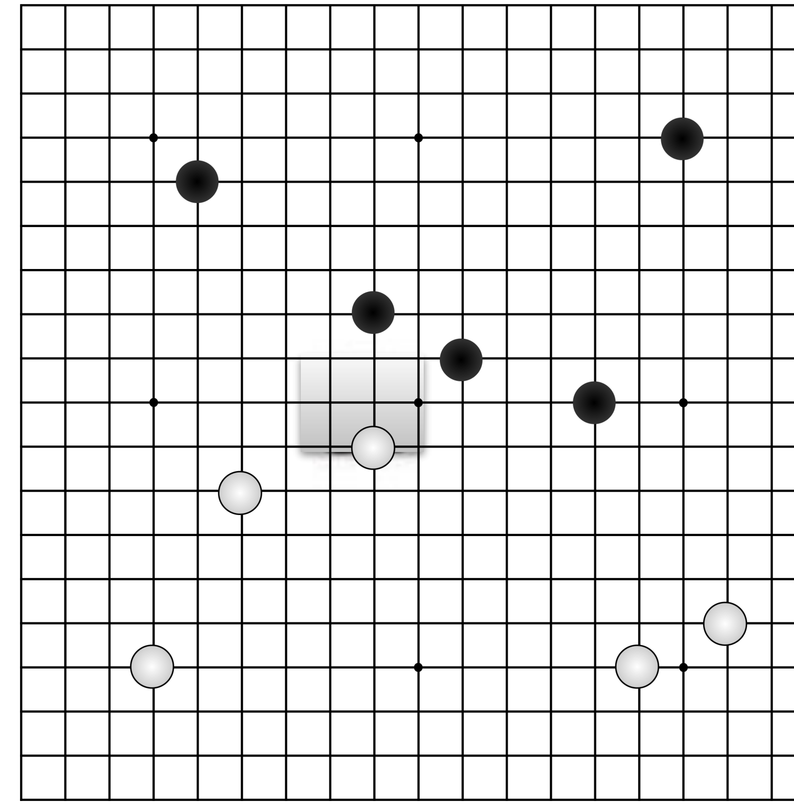
Too shallow!

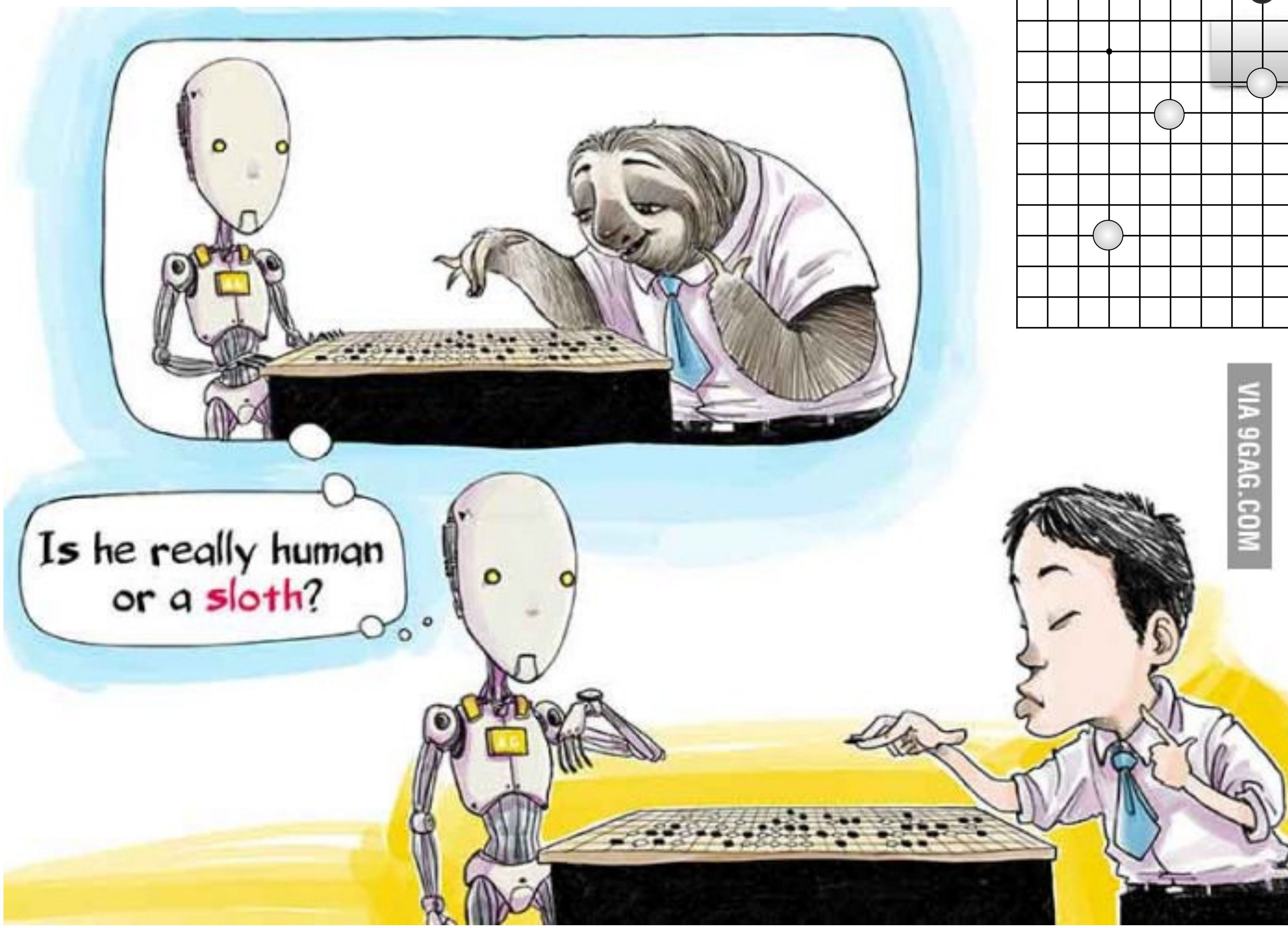
What about Mario? (Deep Mind)



GO: Even Harder

Chess: 20 options
GO: 200 options





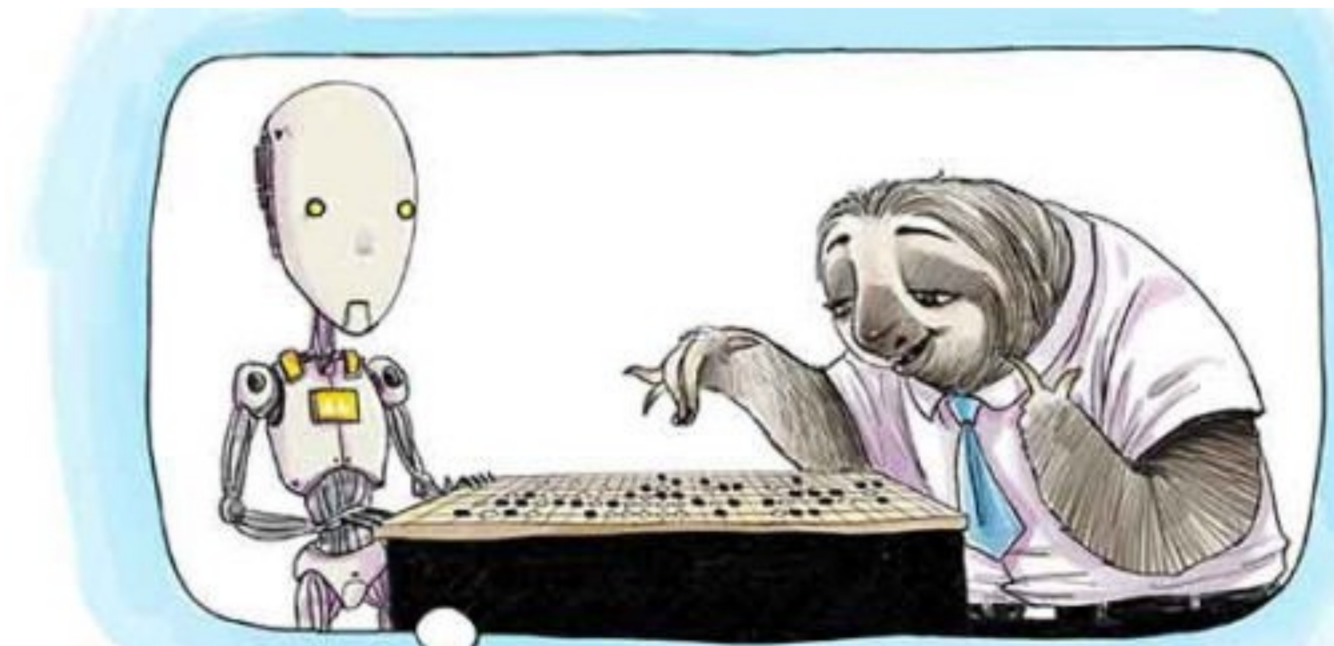
10^170 positions

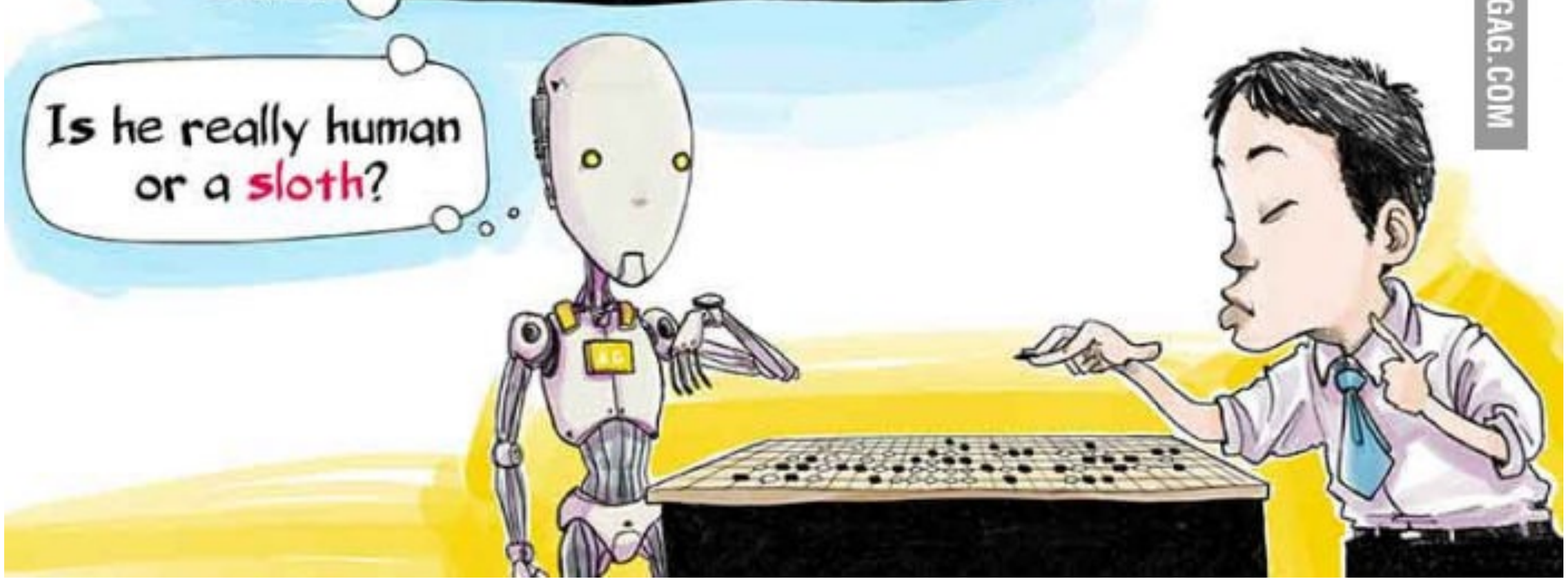
9gag.com

Demis Hassabis,
CEO Deep Mind

Lee Sedol,
18 World Titles

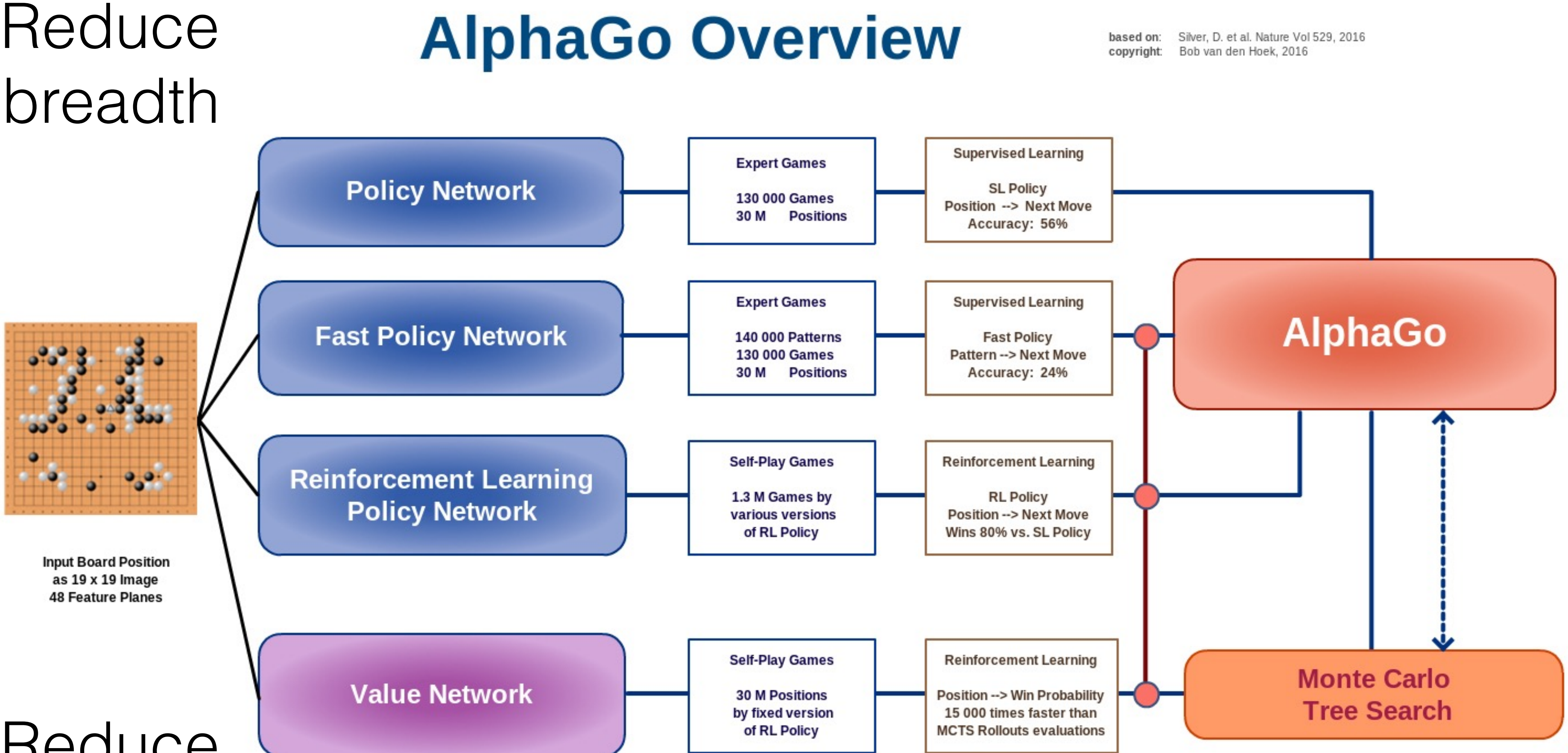
GO: Even Harder





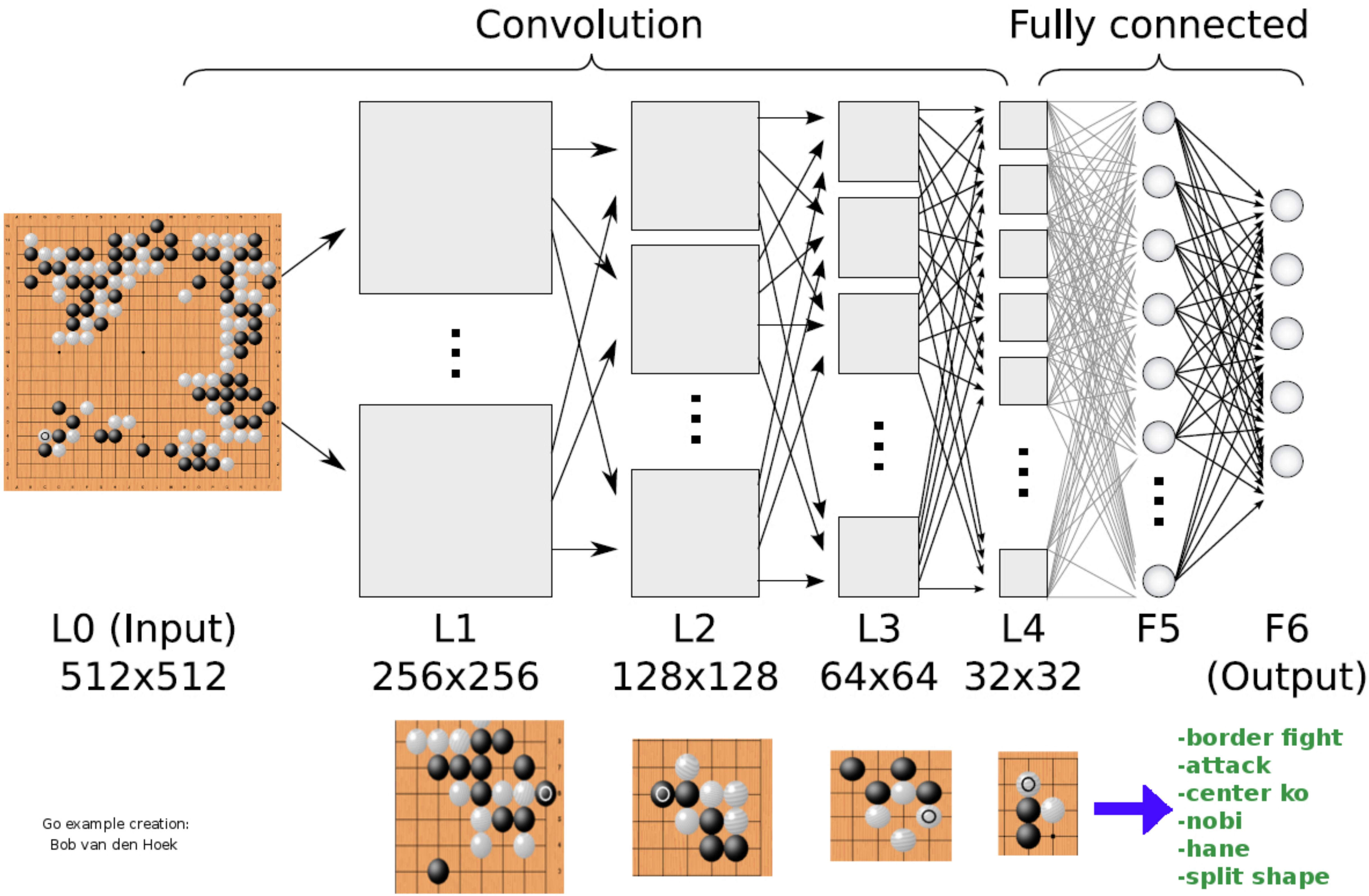
9gag.com

Reduce breadth



Reduce depth

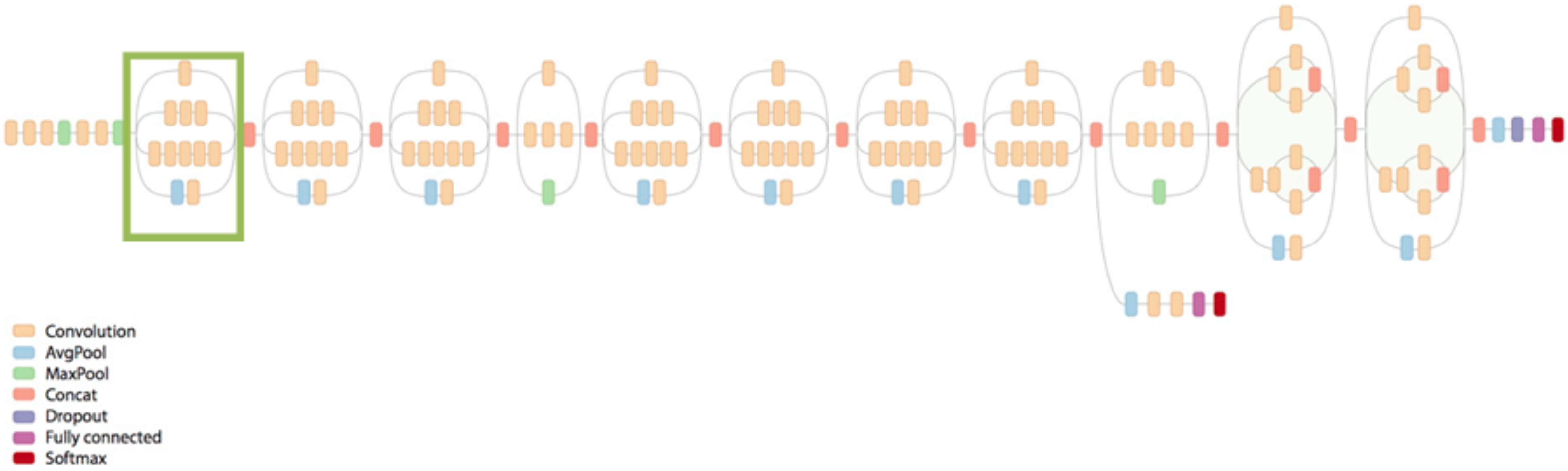
http://deeplearningskysthelimit.blogspot.com/2016/04/part-2-alphago-under-magnifying-glass.html



Go example creation:
Bob van den Hoek

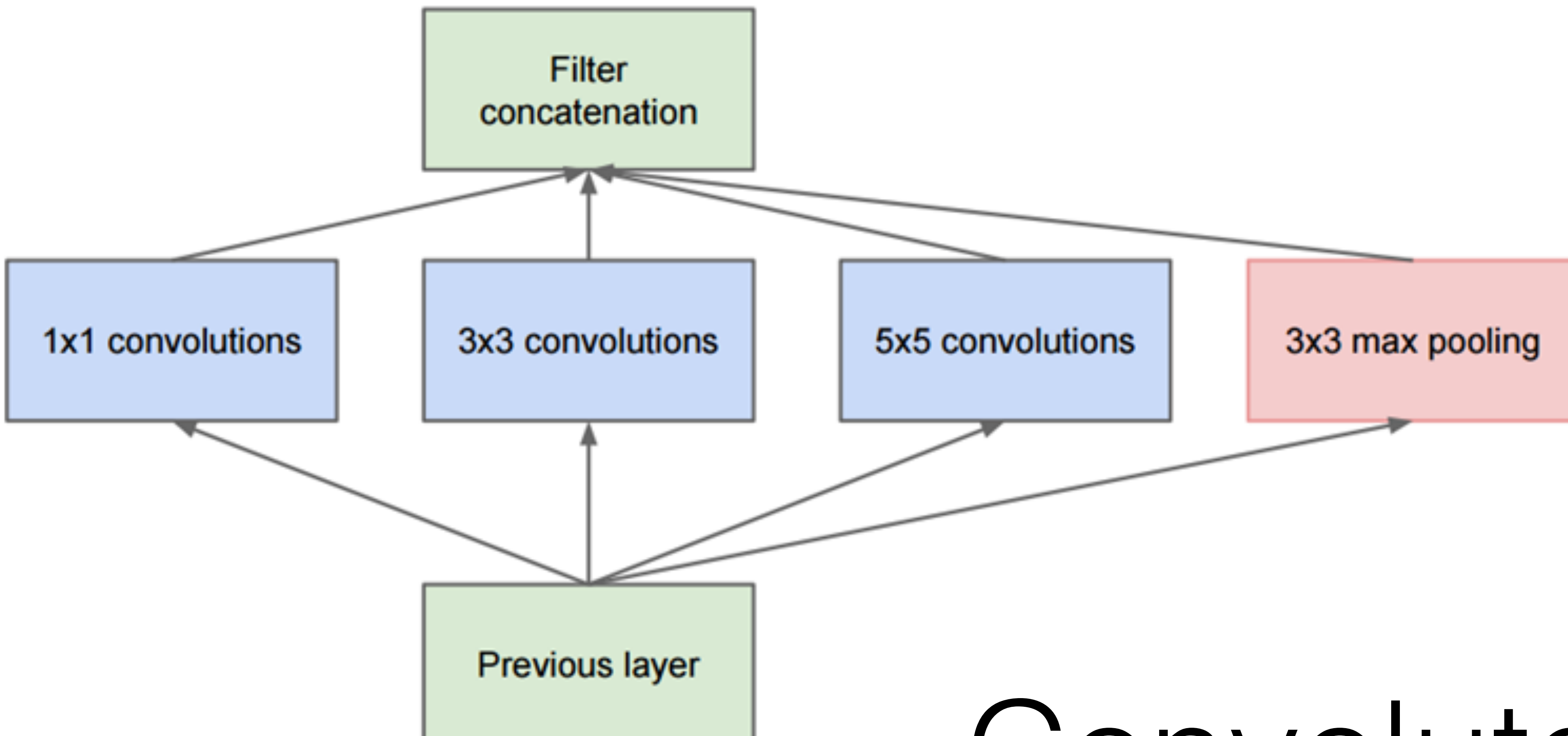
http://gobase.org/online/intergo/?query=%22hane%20nobi%22

http://deeplearningskysthelimit.blogspot.com/2016/04/part-2-alphago-under-magnifying-glass.html



Green box shows parallel region of GoogLeNet

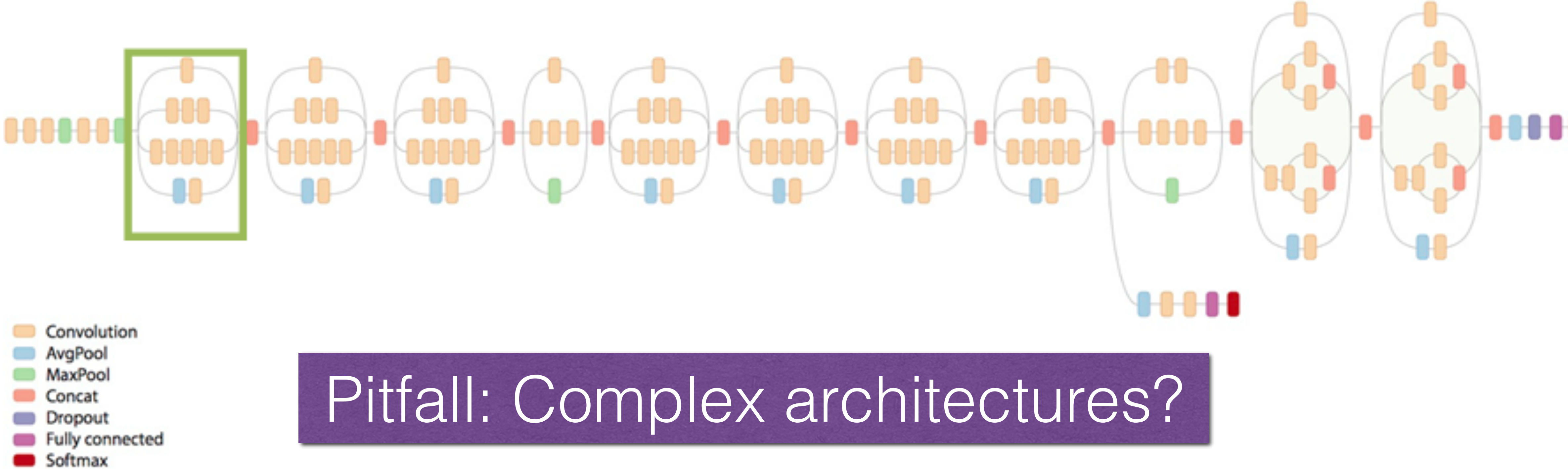
Inception module



Naïve idea of an Inception module

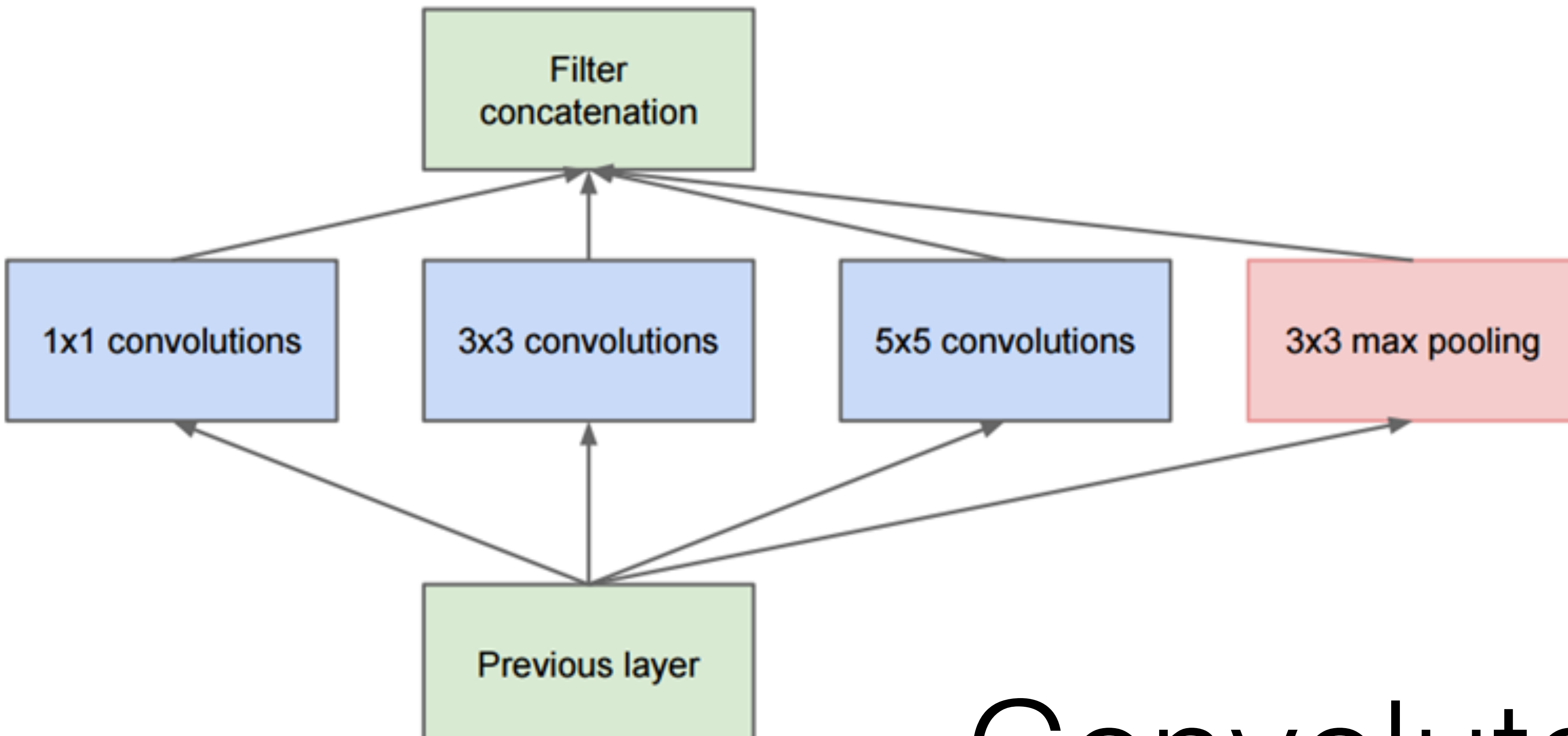
Convoluted?

Pitfall: Complex architectures?



Green box shows parallel region of GoogLeNet

Inception module

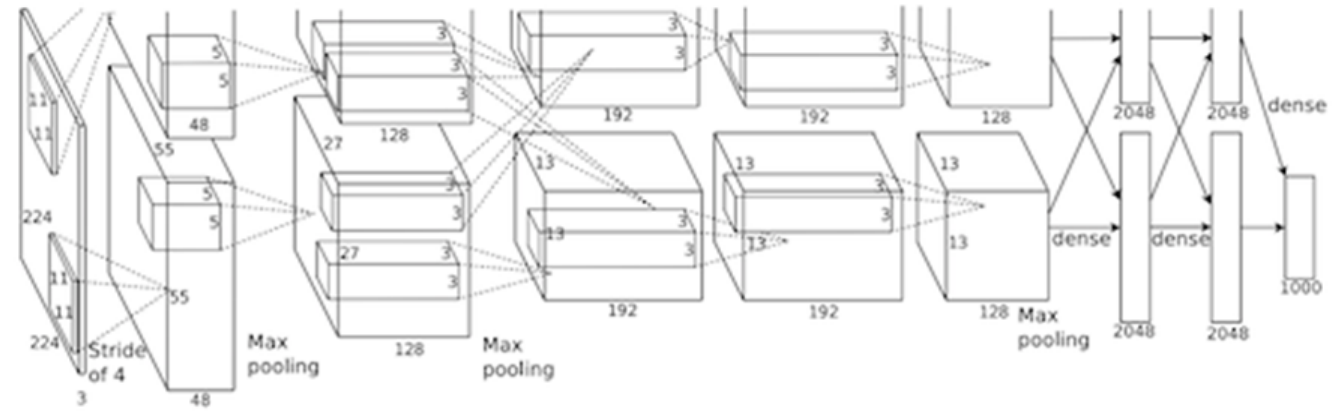


Naïve idea of an Inception module

Convoluted?

IMAGENET

Large Scale Visual Recognition Challenge

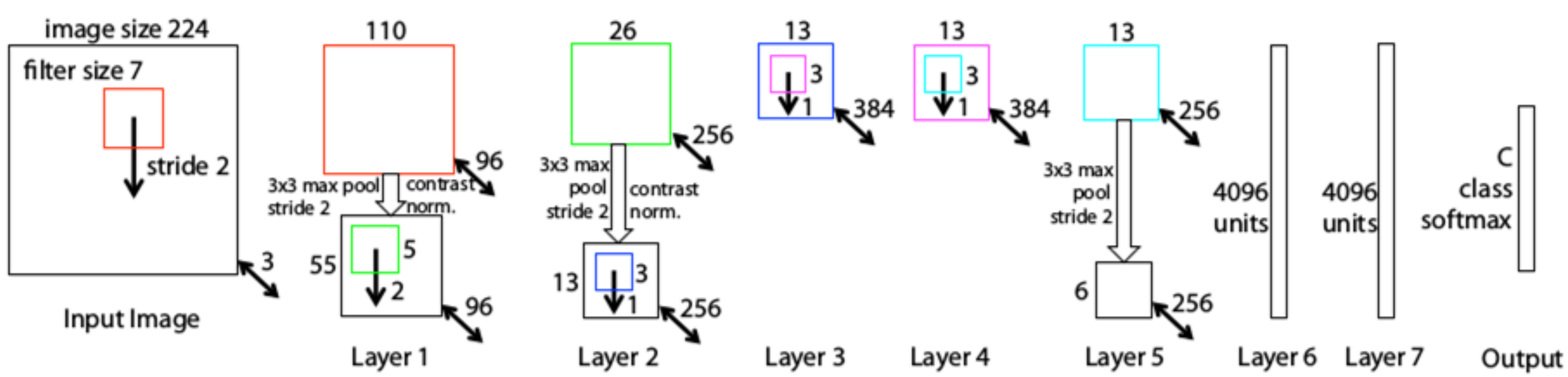


AlexNet architecture (May look weird because there are two different “streams”. This is because the training process was so computationally expensive that they had to split the training onto 2 GPUs)

- 2012: Alexnet (error rate 15.4%)
- 2013: ZFnet (error rate 11.12%)

ILSVRC

- DeConvNets (Caffe)

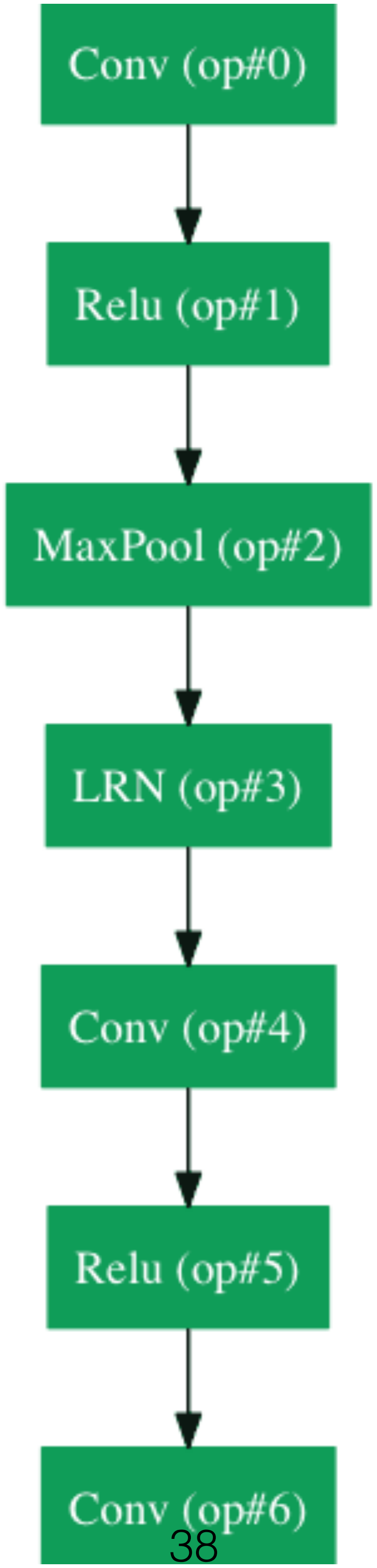


ZF Net Architecture

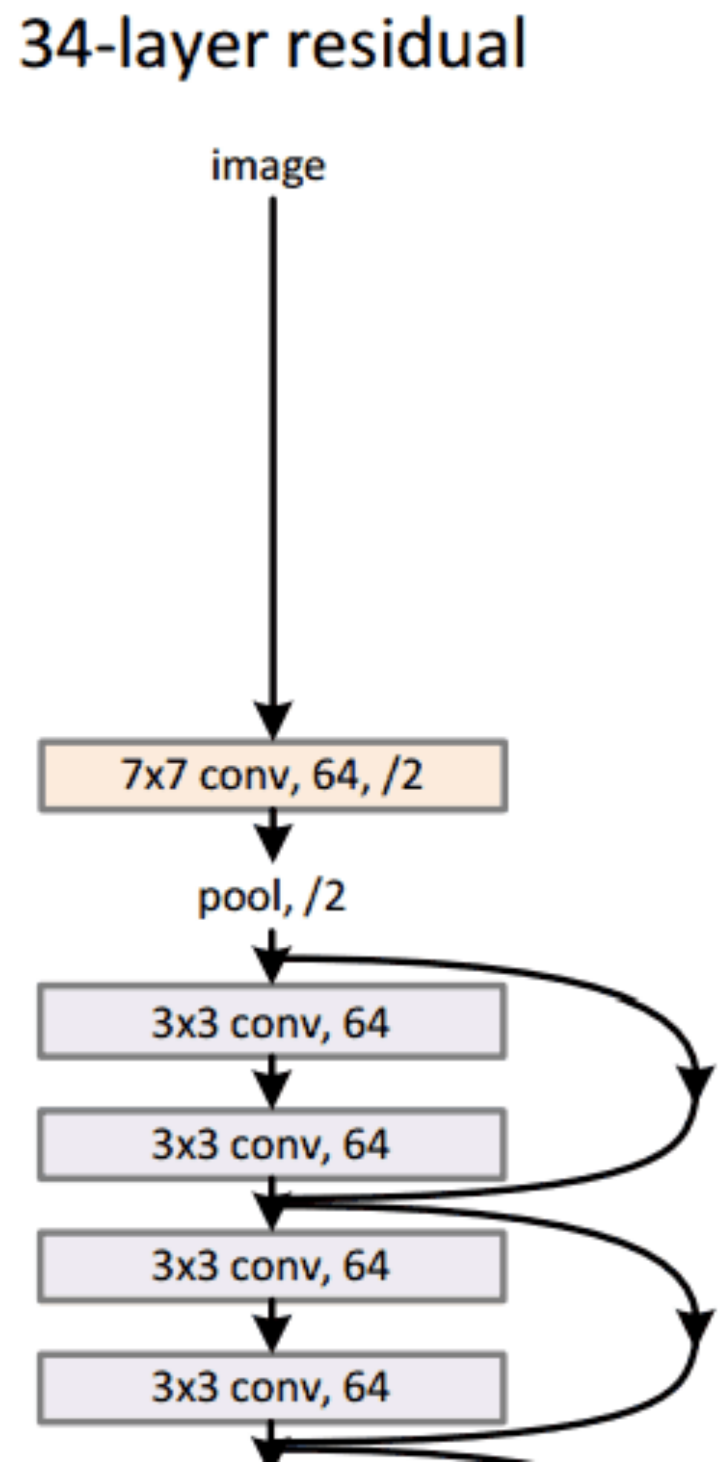
Adit Deshpande

https://adeshpande3.github.io/adeshpande3.github.io/The-9-Deep-Learning-Papers-You-Need-To-Know-About.html

GoogLeNet (2014) 6.7%



ResNET (2015) error rate: 3.6%



2015 ILSVRC leaderboard

Team name	Entry description	Number of object categories won	mean AP
MSRA	An ensemble for detection.	194	0.620741
Qualcomm Research	NeoNet ensemble with bounding box regression. Validation mAP is 54.6	4	0.535745
CUImage	Combined multiple models with the region proposals of cascaded RPN, 57.3% mAP on Val2.	2	0.527113
The University of Adelaide	9 models	0	0.514434
MCG-ICT-	2 models on 2 proposals without category information: {[SS+EB]+	[illegible]	[illegible]

Classification error: 0.03567

Yellow: Winner in category
Yellow/White: Reveal code
Gray: Won't reveal code

2016 ILSVRC leaderboard

Team name	Entry description	Number of object categories won	mean AP
CUImage	Ensemble of 6 models using provided data	109	0.662751
Hikvision	Ensemble A of 3 RPN and 6 FRCN models, mAP is 67 on val2	30	0.652704
Hikvision	Ensemble B of 3 RPN and 5 FRCN models, mean AP is 66.9, median AP is 69.3 on val2	18	0.652003
NUIST	submission_1	15	0.608752
NUIST	submission_2	9	0.607124
Trimps-Soushen	Ensemble 2	8	0.61816
360+MCG-ICT-CAS_DET	9 models ensemble with validation and 2 iterations	4	0.615561
360+MCG-ICT-CAS_DET	Baseline: Faster R-CNN with Res200	4	0.590596
Hikvision	Best single model, mAP is 65.1 on val2	2	0.634003
CIL	Ensemble of 2 Models	1	0.553542
360+MCG-ICT-CAS_DET	9 models ensemble	0	0.613045

Classification error:
0.02991

Identifying streaking asteroids

DeepStreaks: identifying FMOs in ZTF data 5

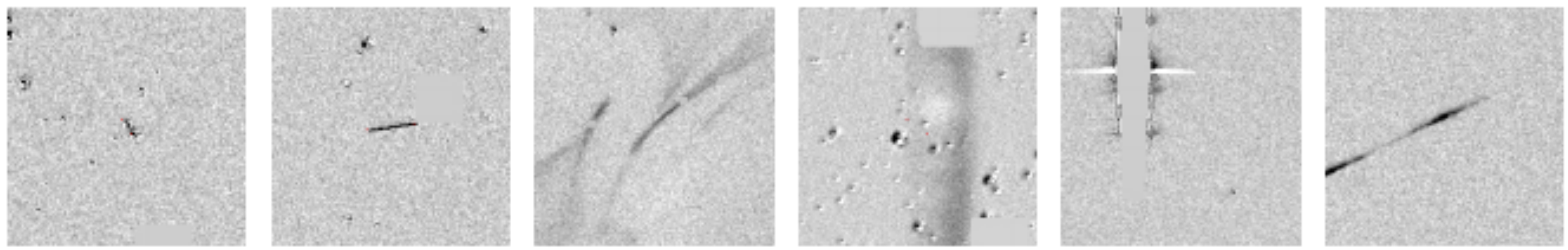
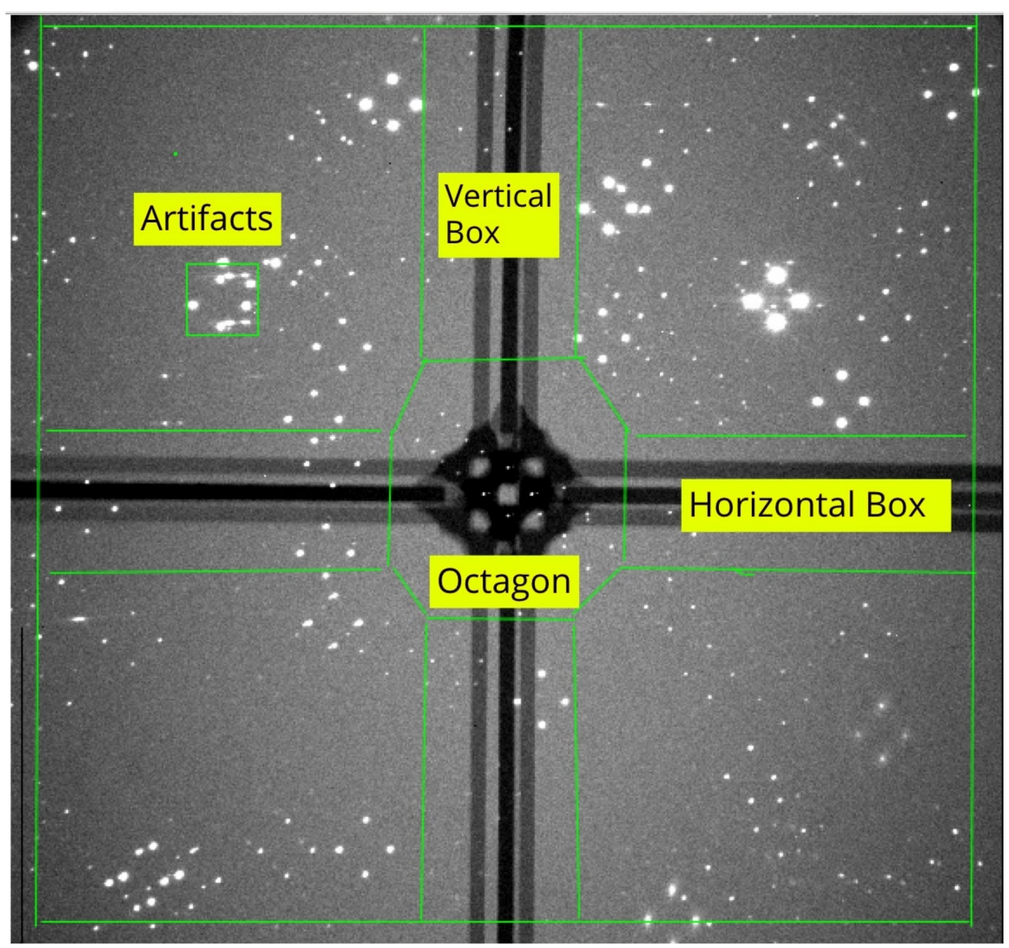


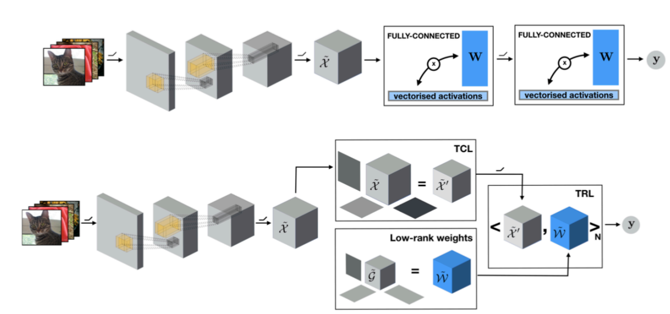
Figure 4. Examples of different classes of bogus streak detections.

Duev, Mahabal, … 2019
arxiv:1904.05920

Robopol



Tensor Regression Networks - a new concept!



Source: Tensor Regression Networks (2017) - J. Kossaifi et al

Dhruv Paranjpye
Gina Panopoulou
Robopol team

Extendable to Gattini data (large pixels, bright upper limits)

braai

`braai` architecture

We will use a simple custom VGG-like sequential model ($VGG6$; this architecture was first proposed by the Visual Geometry Group of the Department of Engineering Science, University of Oxford, UK). The model has six layers with trainable parameters: four convolutional and two fully-connected. The first two convolutional layers use 16 3x3 pixel filters each while in the second pair, 32 3x3 pixel filters are used. To prevent over-fitting, a dropout rate of 0.25 is applied after each max-pooling layer and a dropout rate of 0.5 is applied after the second fully-connected layer. ReLU activation functions (Rectified Linear Unit – a function defined as the positive part of its argument) are used for all five hidden trainable layers; a sigmoid activation function is used for the output layer.

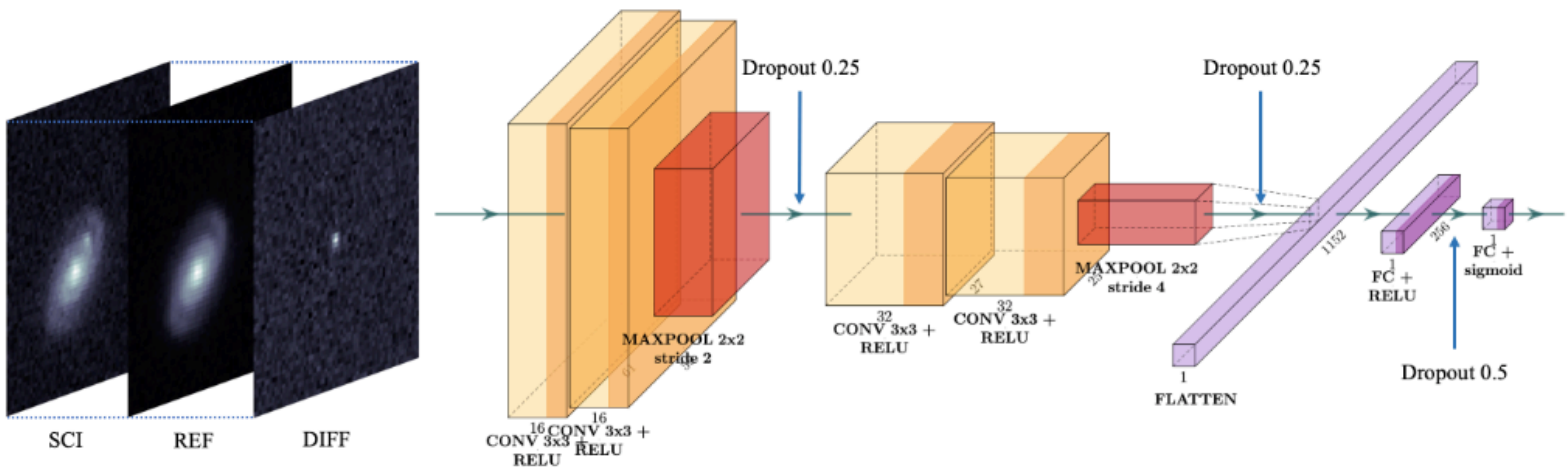
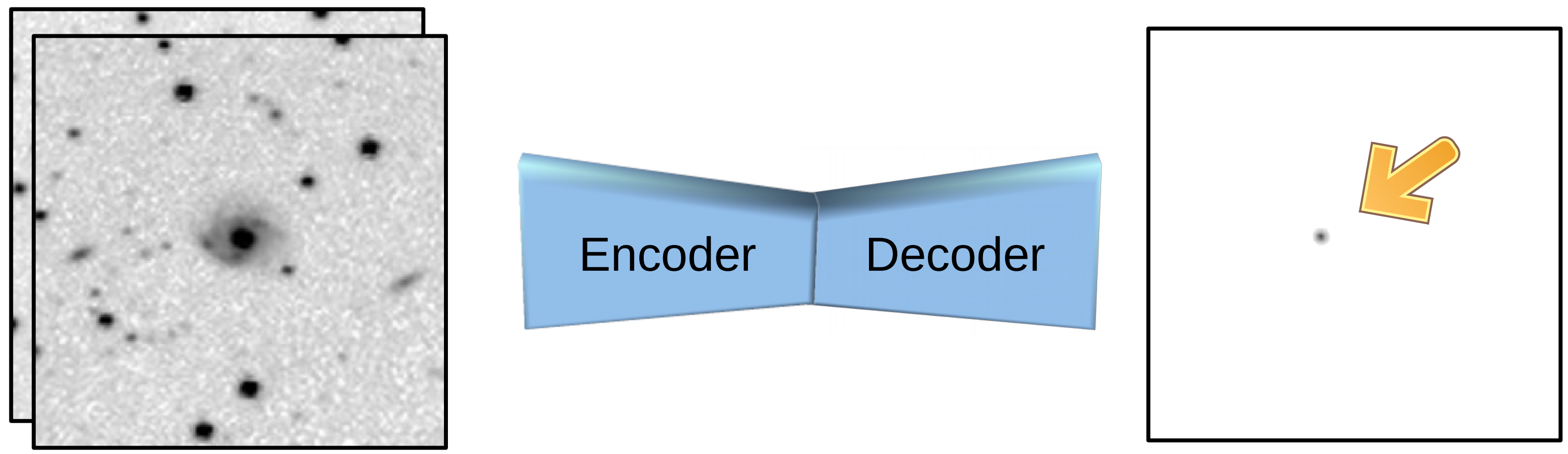
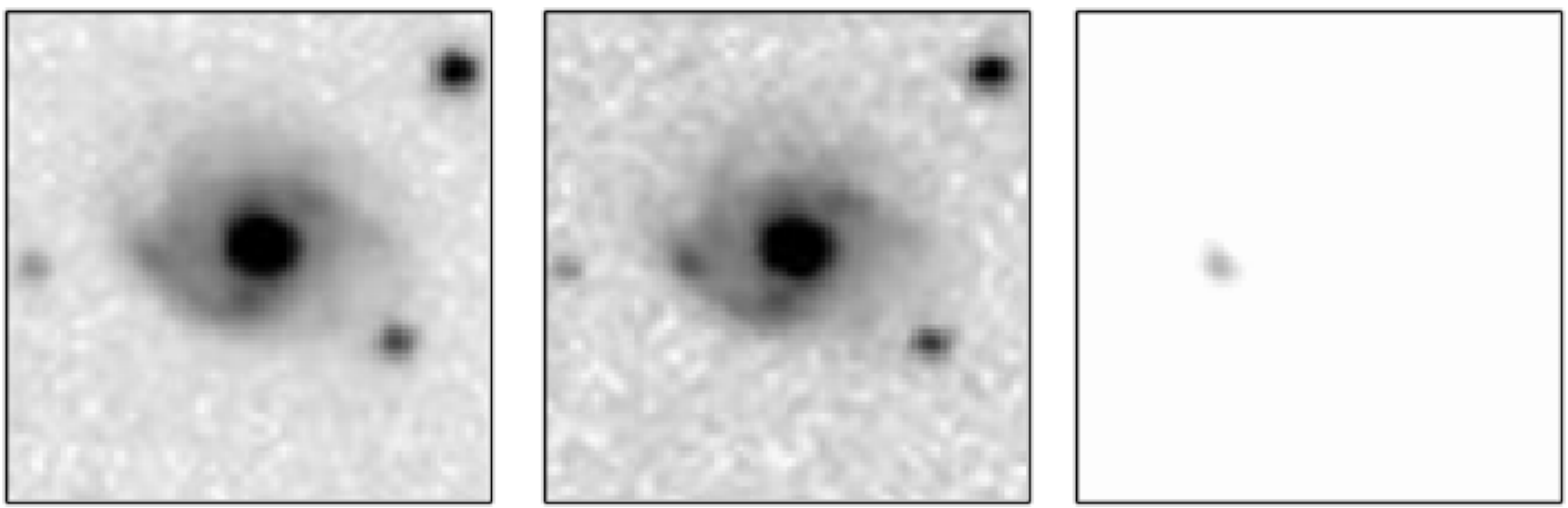


Image subtraction for hunting transients without subtraction





Sedaghat and Mahabal, 2017
arXiv:1710.01422

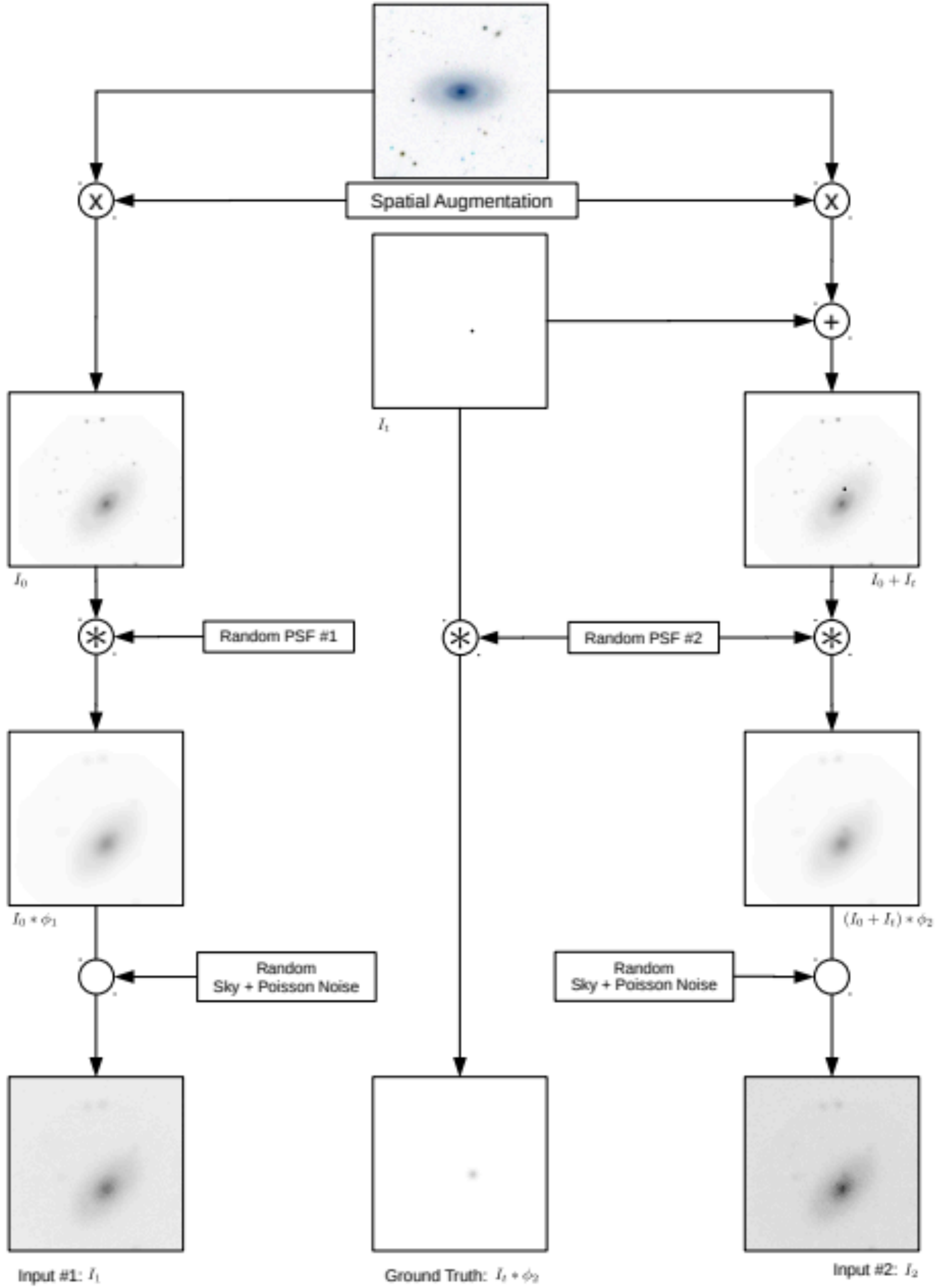
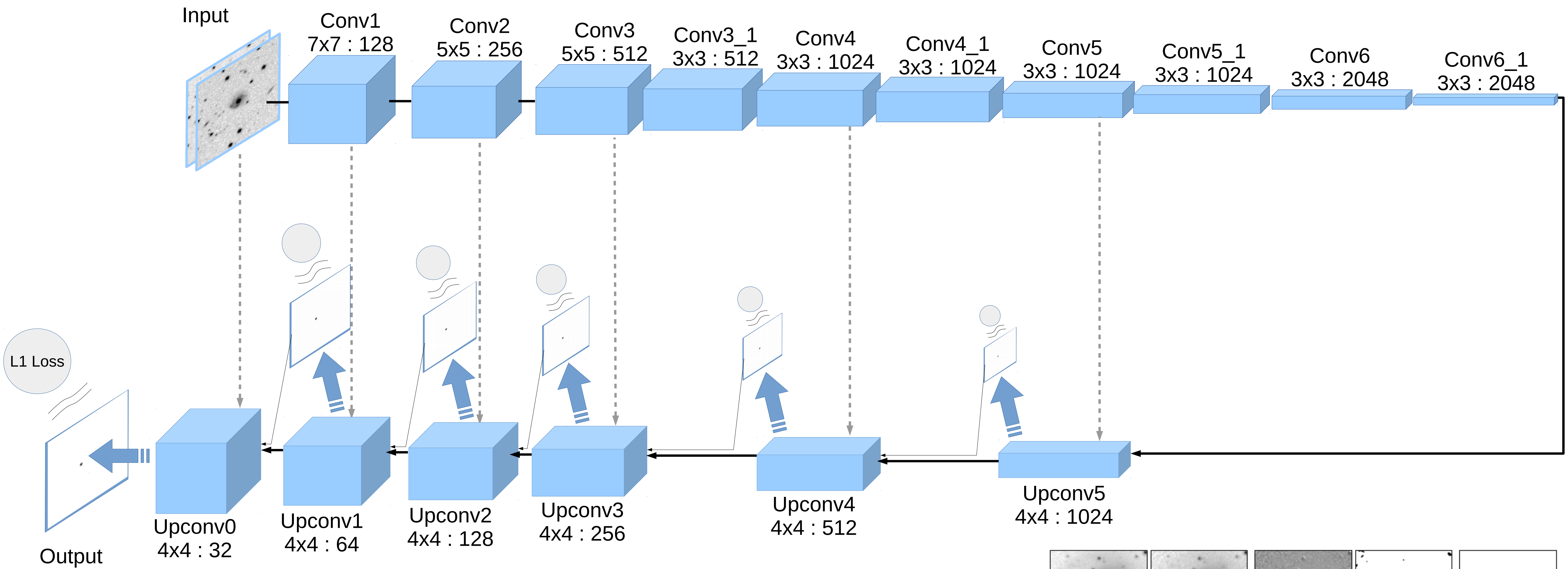
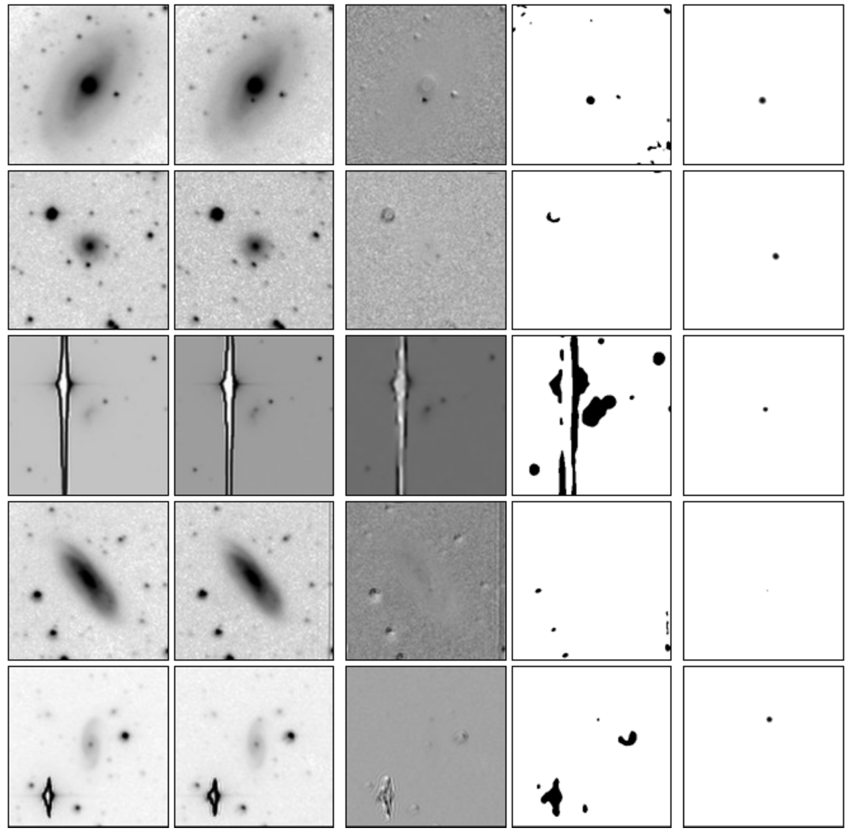


Figure 5. The synthetic sample generation procedure. The notations used here are described in Equations (1) and (2).

Training cycles involving different PSFs

Encoder-decoder network (fully convolutional)

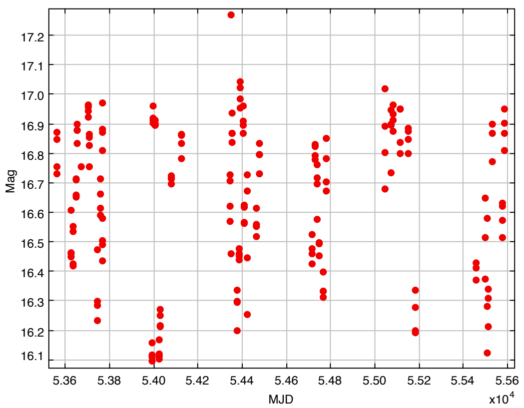


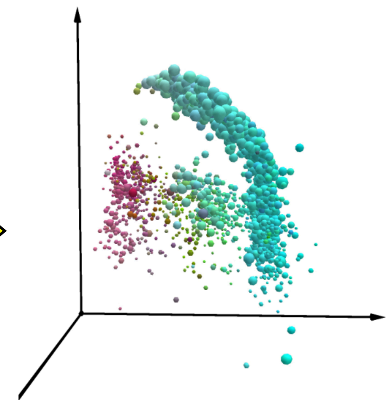


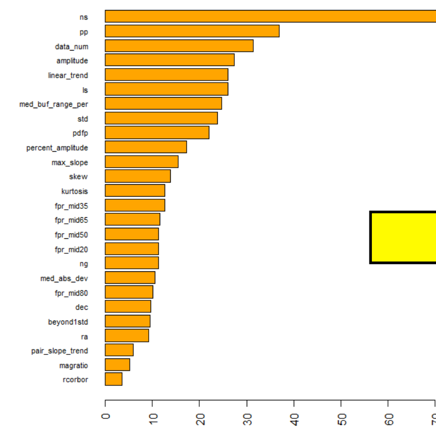
Sedaghat and Mahabal, 2017

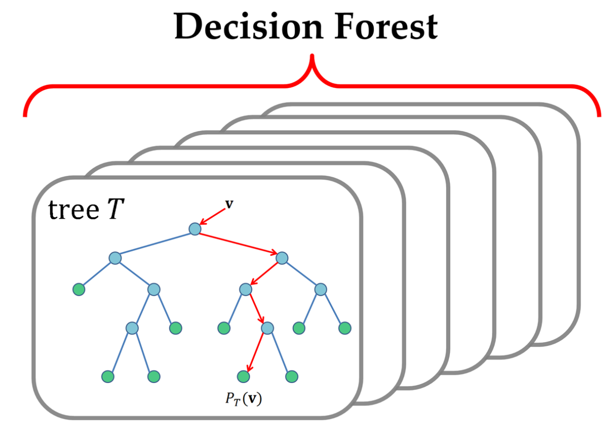
arXiv:1710.01422

Classification Workflow









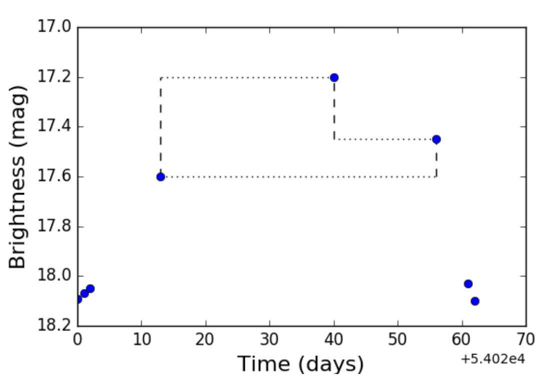
Light curves

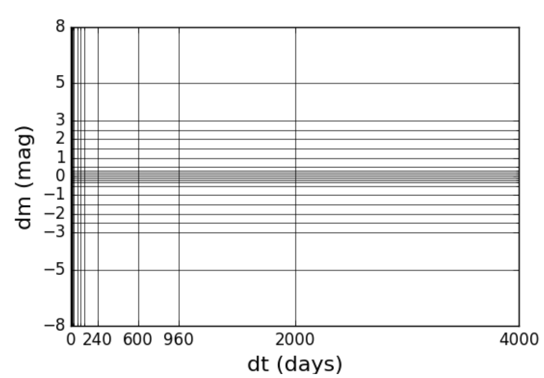
Feature vectors

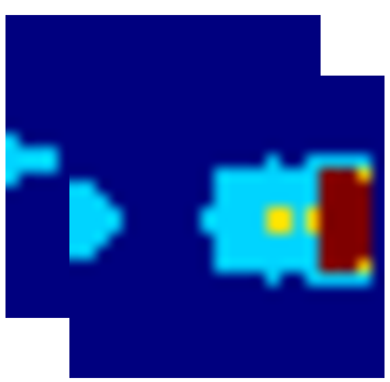
Dimensionality Reduction

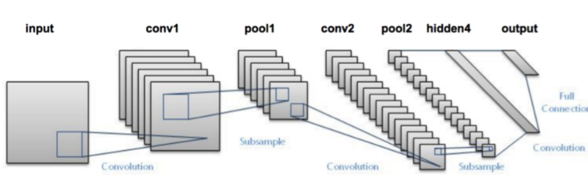
Classification

Domain knowledge/subjectivity







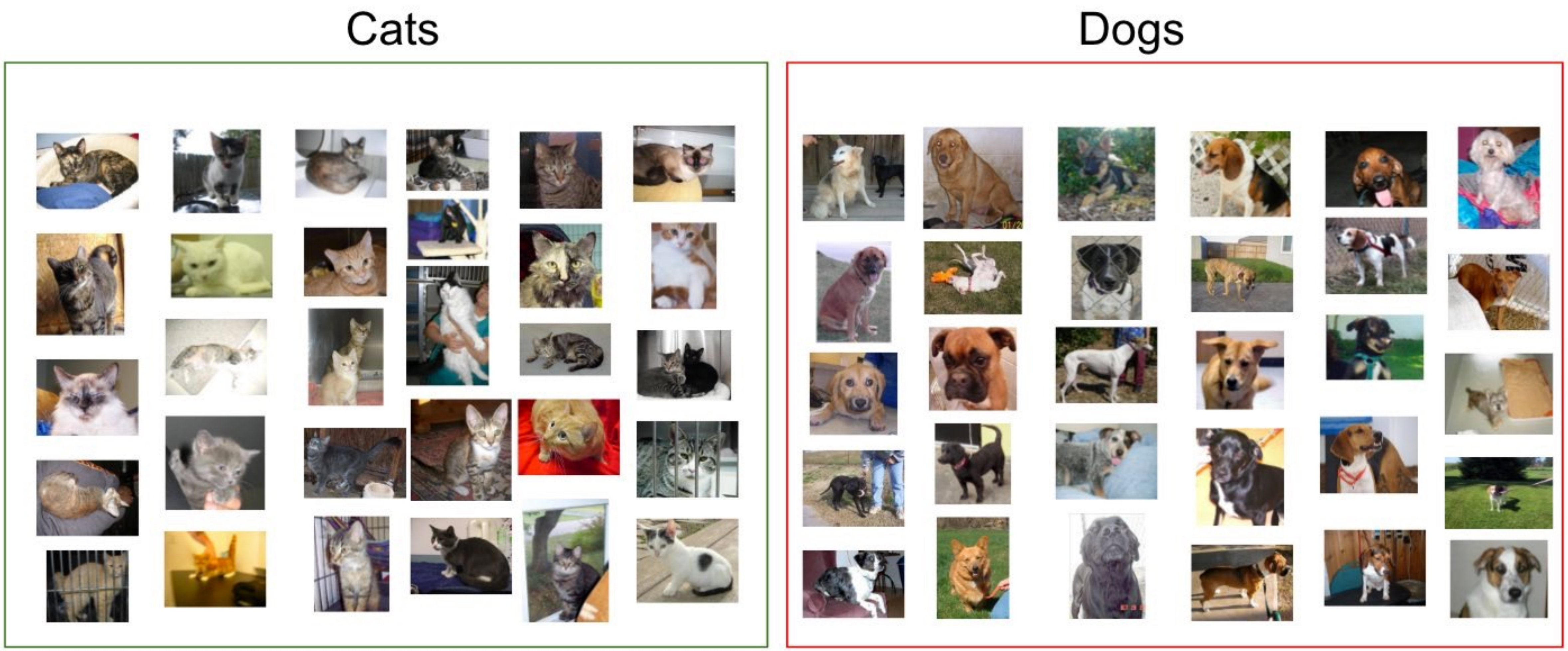


Light curves

Density representation

Equi-area images

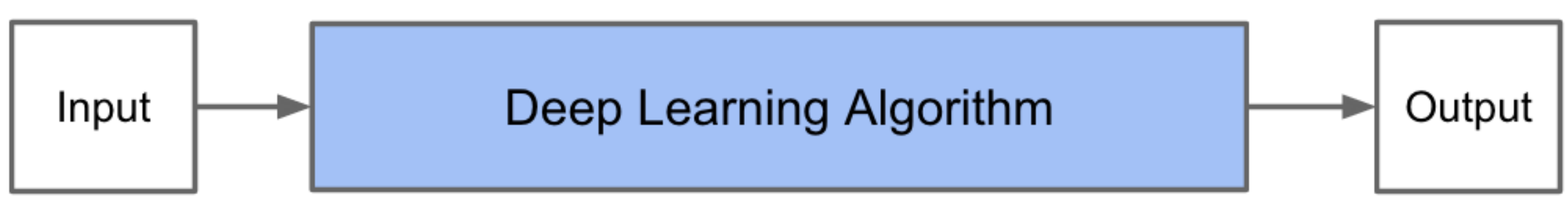
Convolutional Neural Network



Sample of cats & dogs images from Kaggle Dataset

Traditional Machine Learning Flow

Promise: Works better



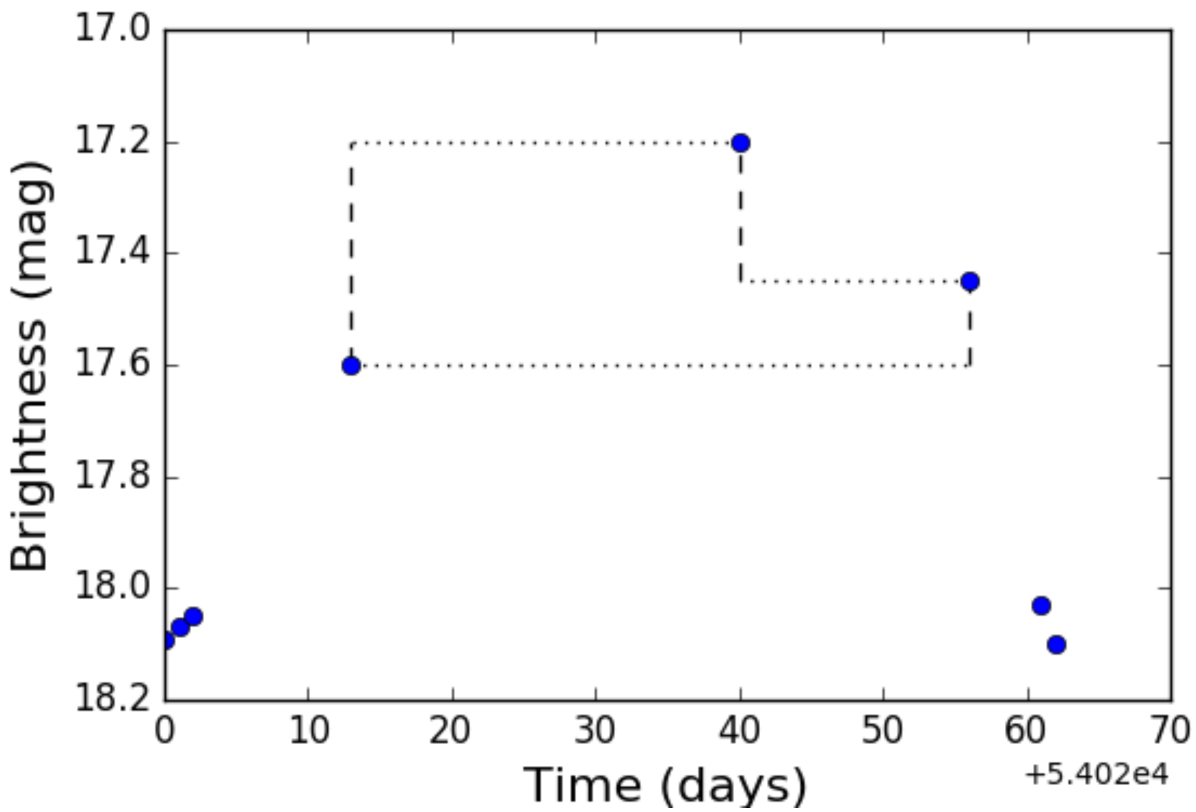
Deep Learning Flow

Pitfall: Blacker box

Adil Moujahid

(dmdt) Image representation

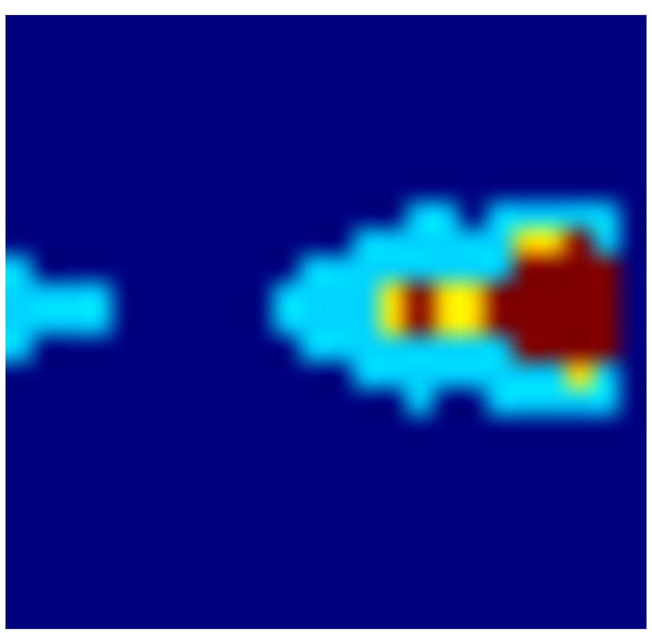
Mahabal, Sheth et al., 2017
1709.06257

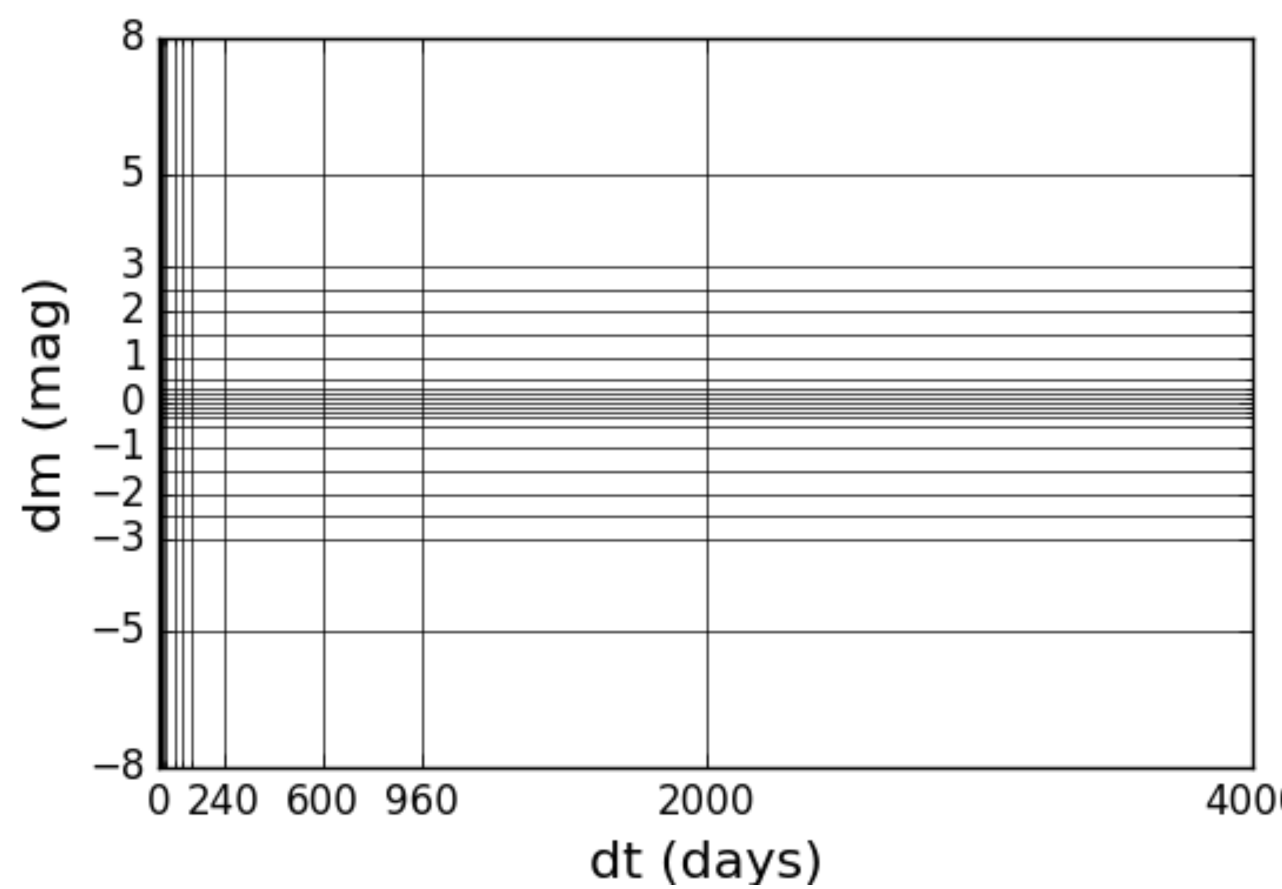


light curve with n points

23 x 24 output grid

n * (n-1)/2 points

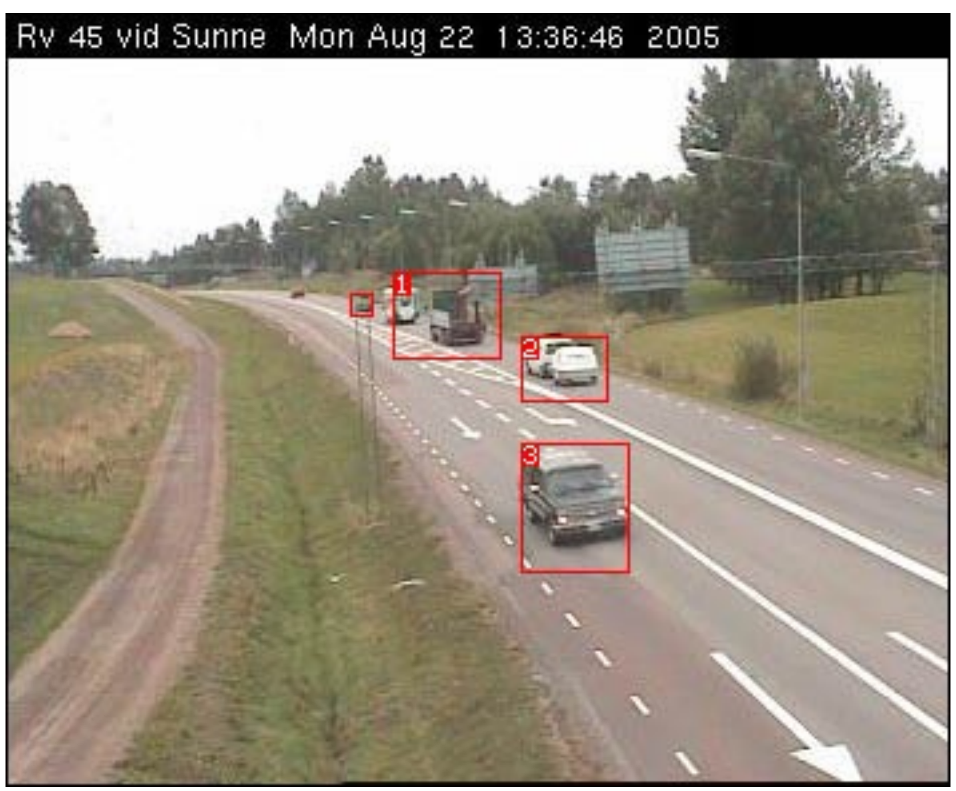




Area equalized pixels

Video Surveillance Analogy

non-convex robust PCA
Netrapalli et al., 2014



Andrew Kirillov

Each class is like a different road
Each individual object has/is perturbations over it

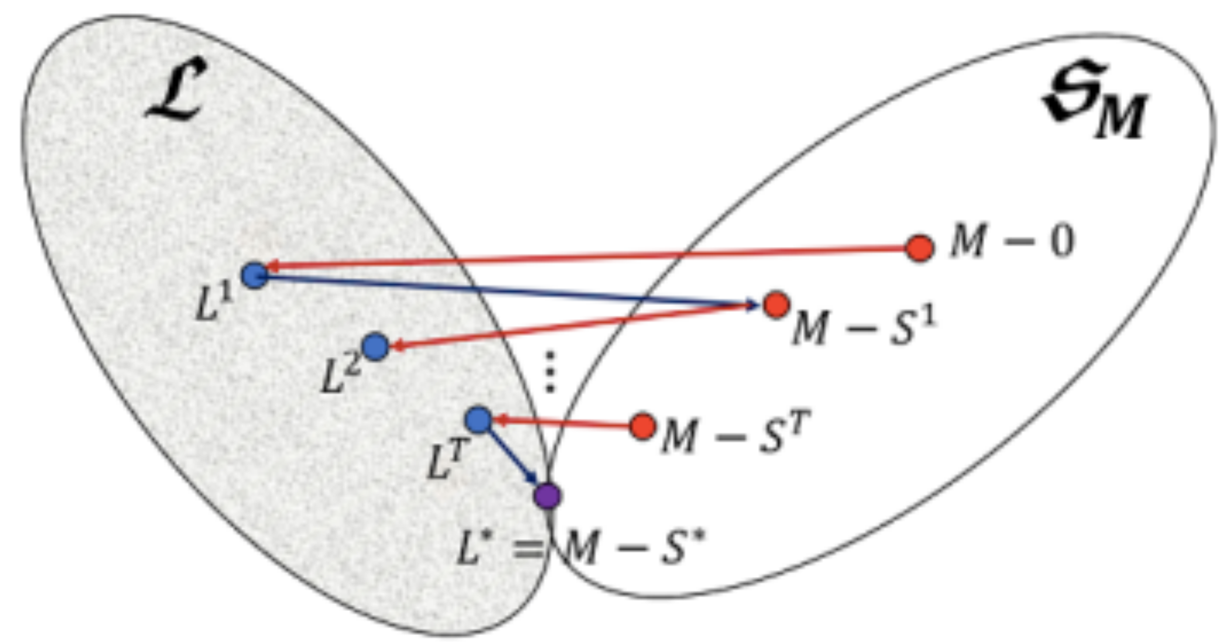
7 classes with at least 500 examples

$dmdt$–image = $b + ci + s$

- background (survey, cadence)
- class background
- individual object (specific)

$$\underset{L,S}{\mathbf{Min}} \|M - L - S\|_2$$

1. L lies in the set of low-rank matrices,
2. S lies in the set of sparse matrices.



non-convex robust PCA
Netrapalli et al., 2014

Diagnosing LIGO lockloss using auxiliary channels

Motivation

- Lockloss events due to environmental events lead to loss of observation time
- Monitor and diagnose lockloss events as they occur

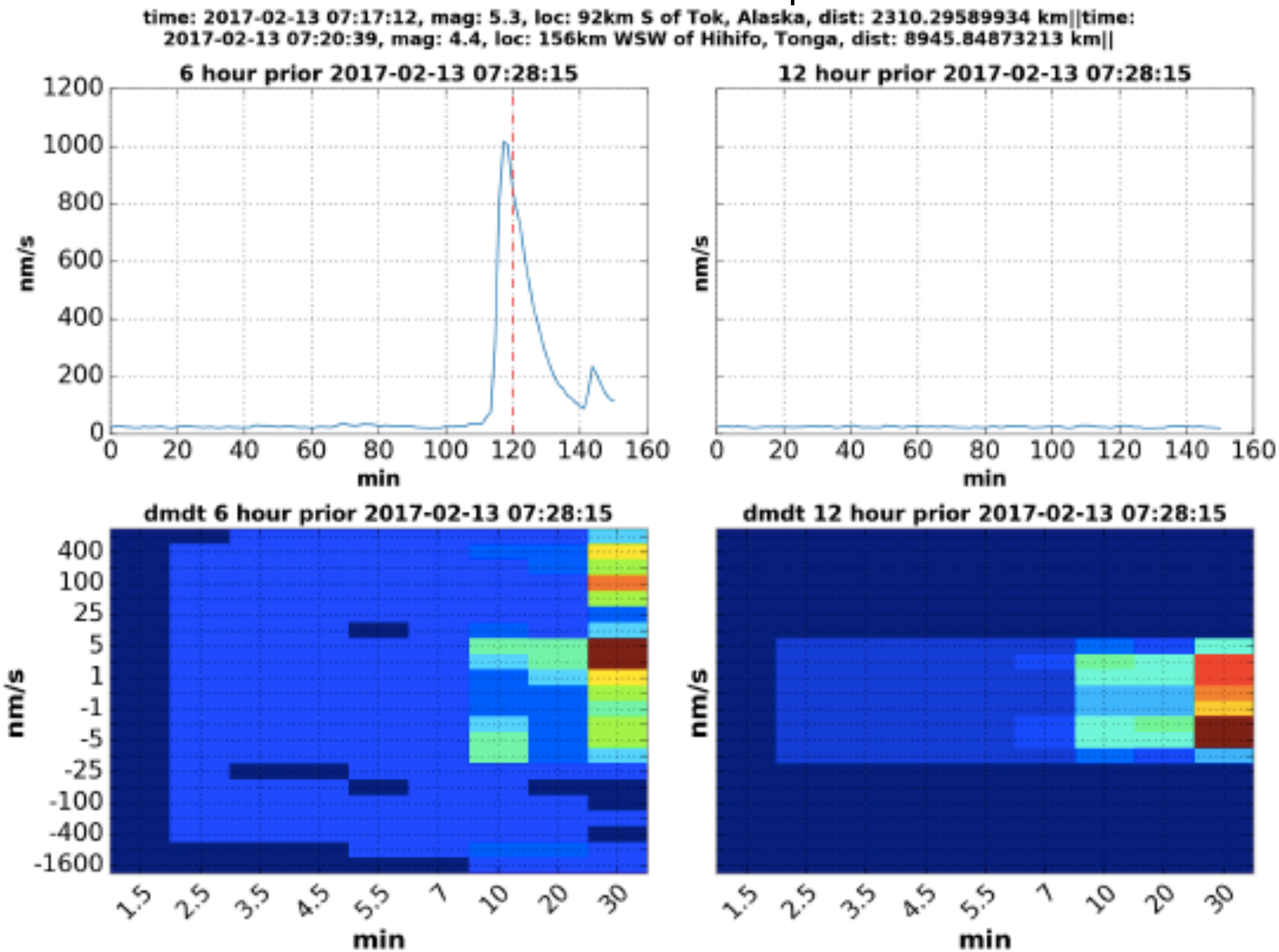
Goals

- To find a minimal set of auxiliary channels that serve as good predictors for lockloss events
- Diagnosis of interferometer behavior leading to lockloss events

With Ayon Biswas and Jess McIver

arXiv: 1910.12143

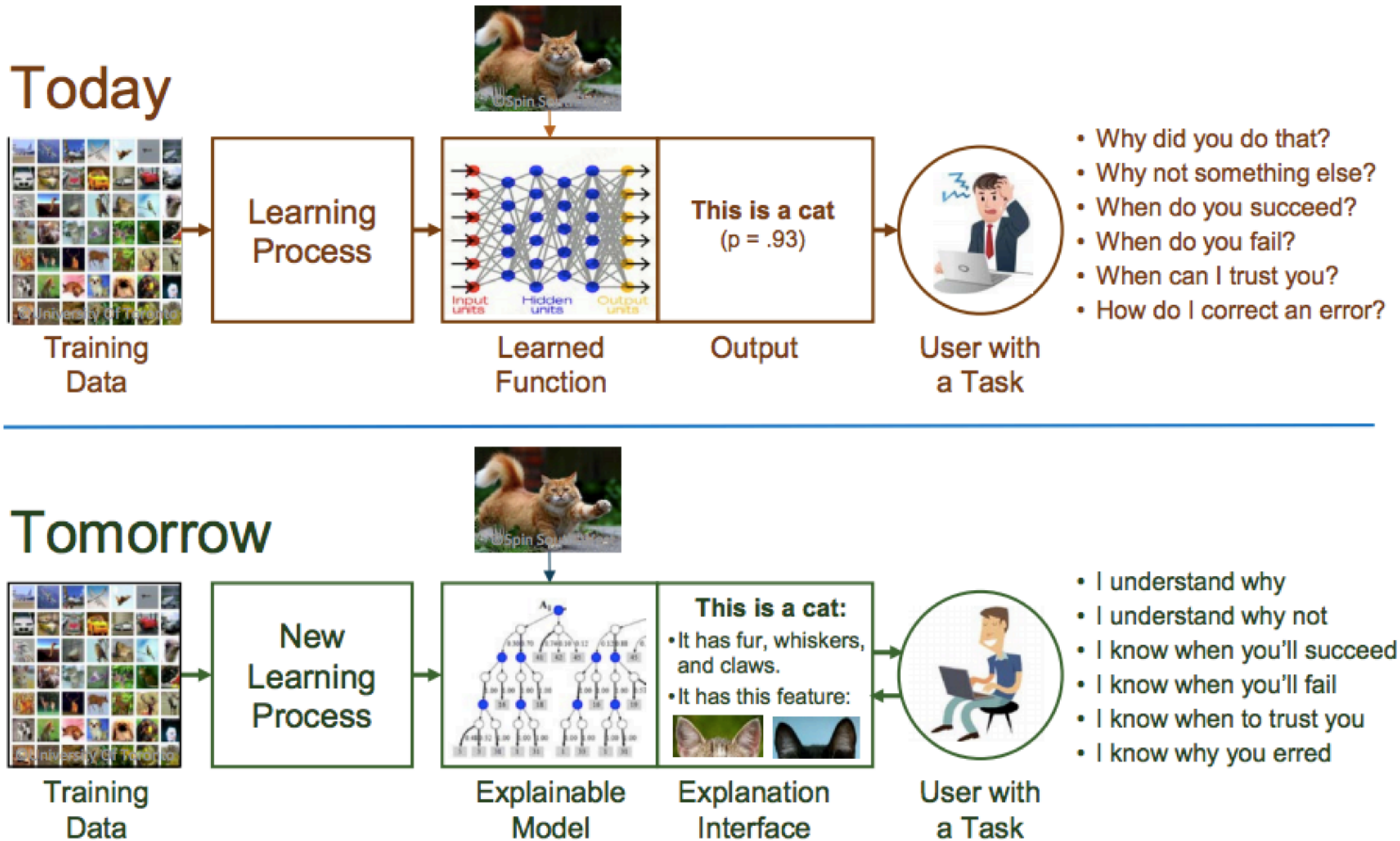
Effect of earthquakes



How to choose bins?
Histogram equalization in both axes?

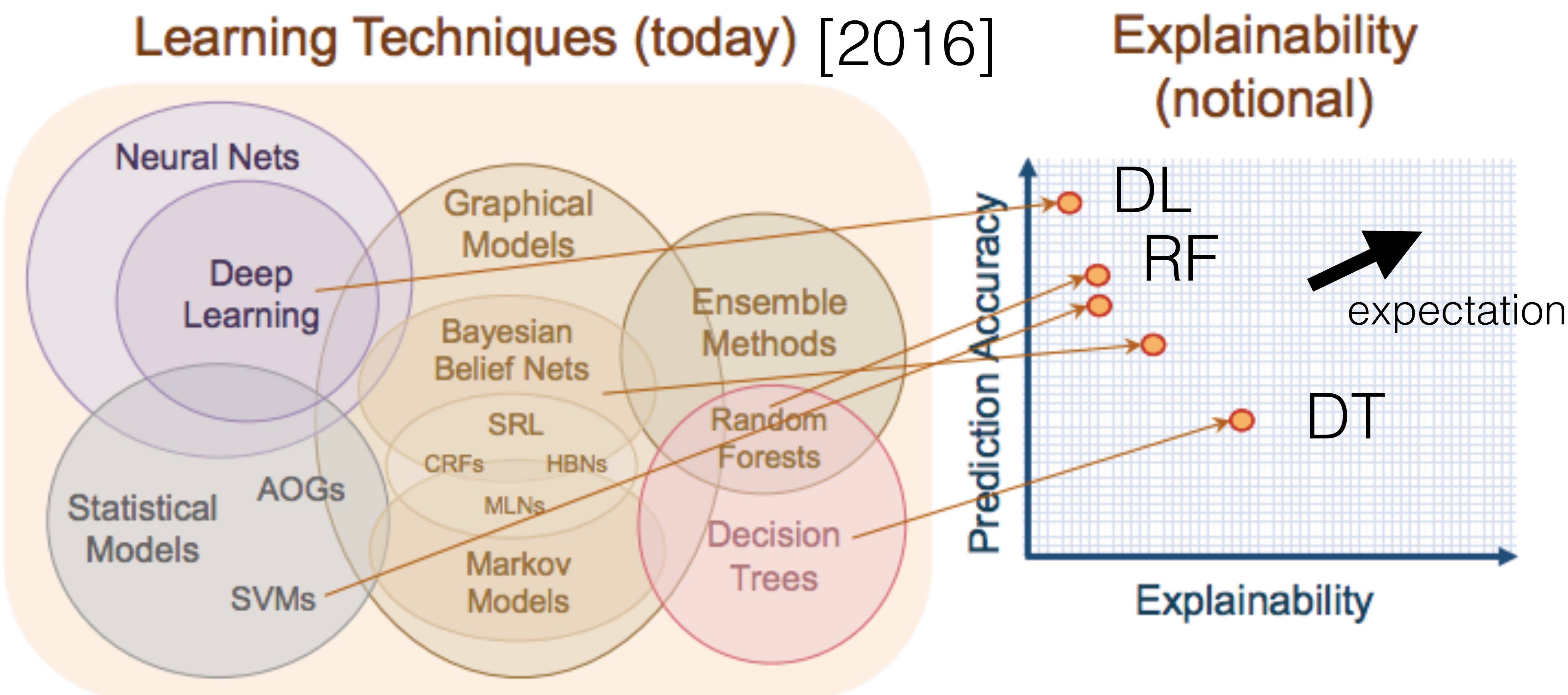
(See the talk by Jess for details)

Interpretability



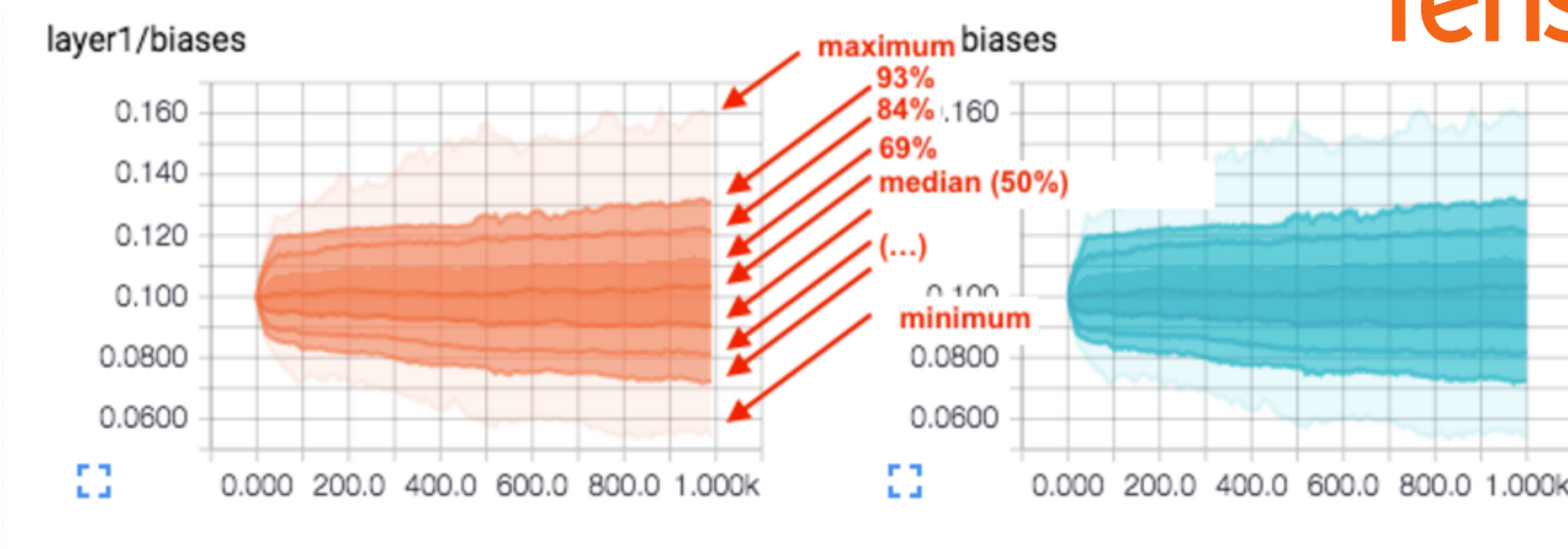
David Gunning (DARPA/I2O)

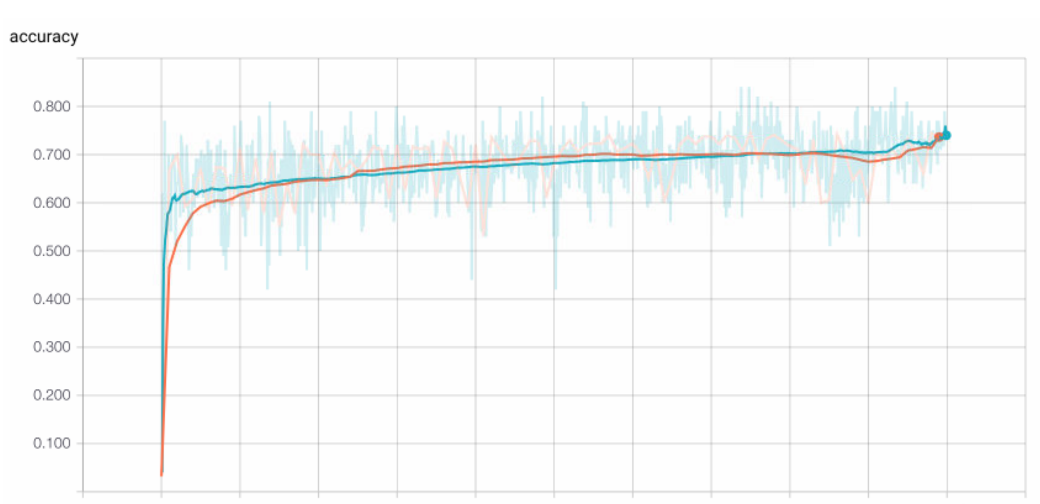
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf



David Gunning (DARPA/I2O)

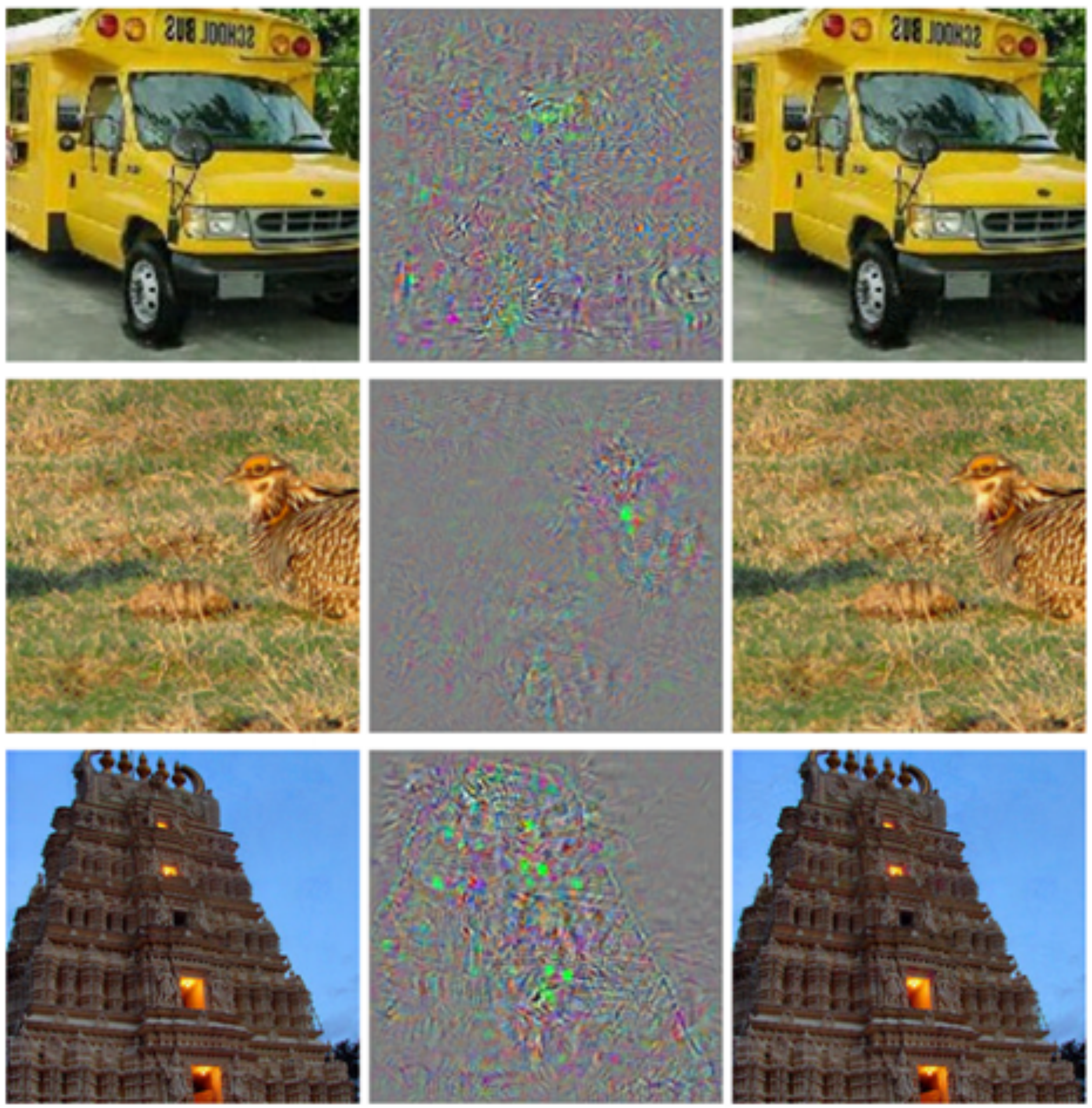
Distribution Summaries



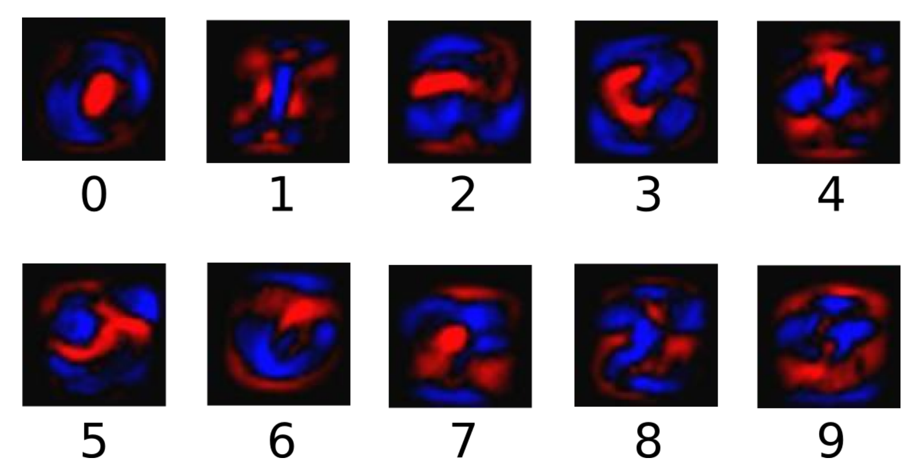


Percentile distributions over the data:
max, 93, 84, 69, 50, 31, 16, 7, min

Including adversarial examples during training

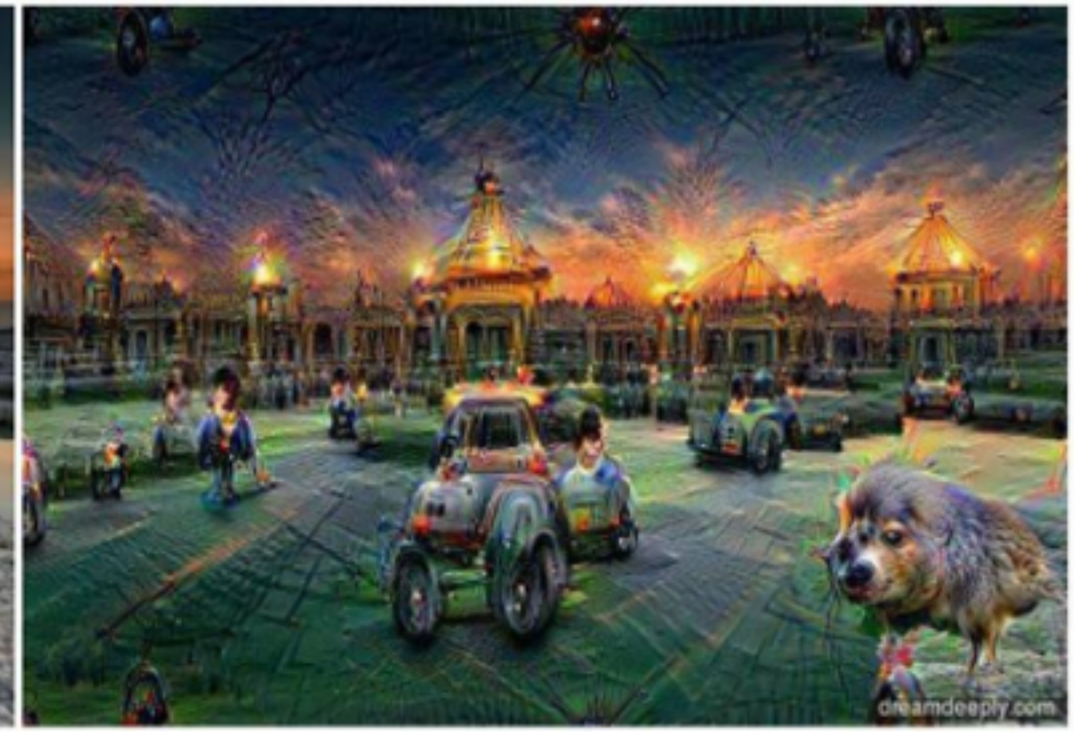


The images in the left most column are correctly classified examples. The middle column represents the distortion between the left and right images. The images in the right most column are predicted to be of the class ostrich! Even though the difference between the images on the left and right is imperceptible to humans, the ConvNet makes drastic errors in classification.



MNIST digits significance map

https://arxiv.org/pdf/1312.6199v4.pdf



Before

After

fromthegrapevine.com

telegraph.co.uk

Style transfer

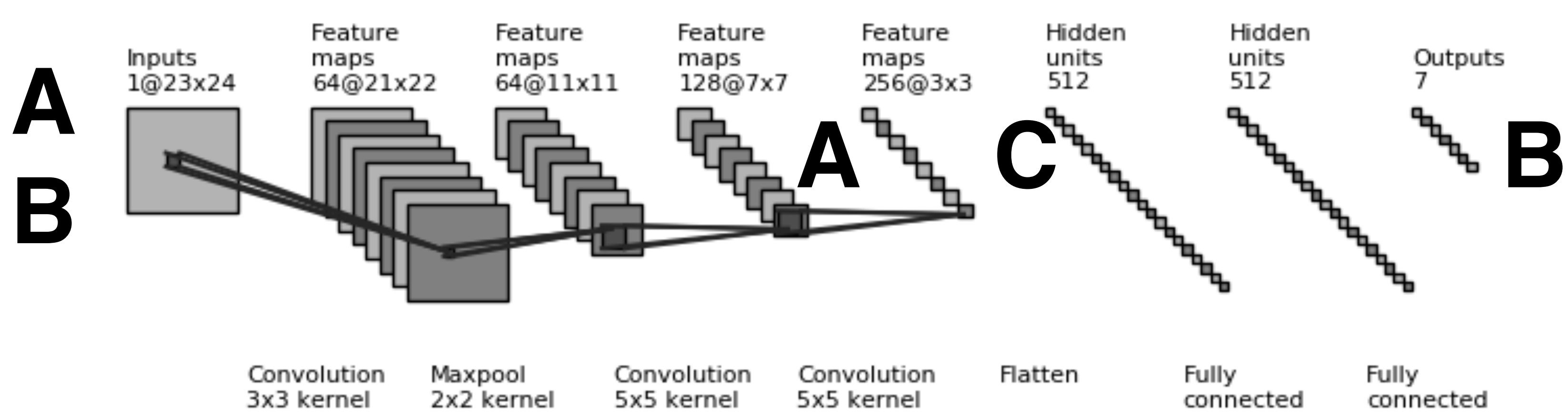
Deep Dreams Gatys et al. 2015

Visualization for interpretability

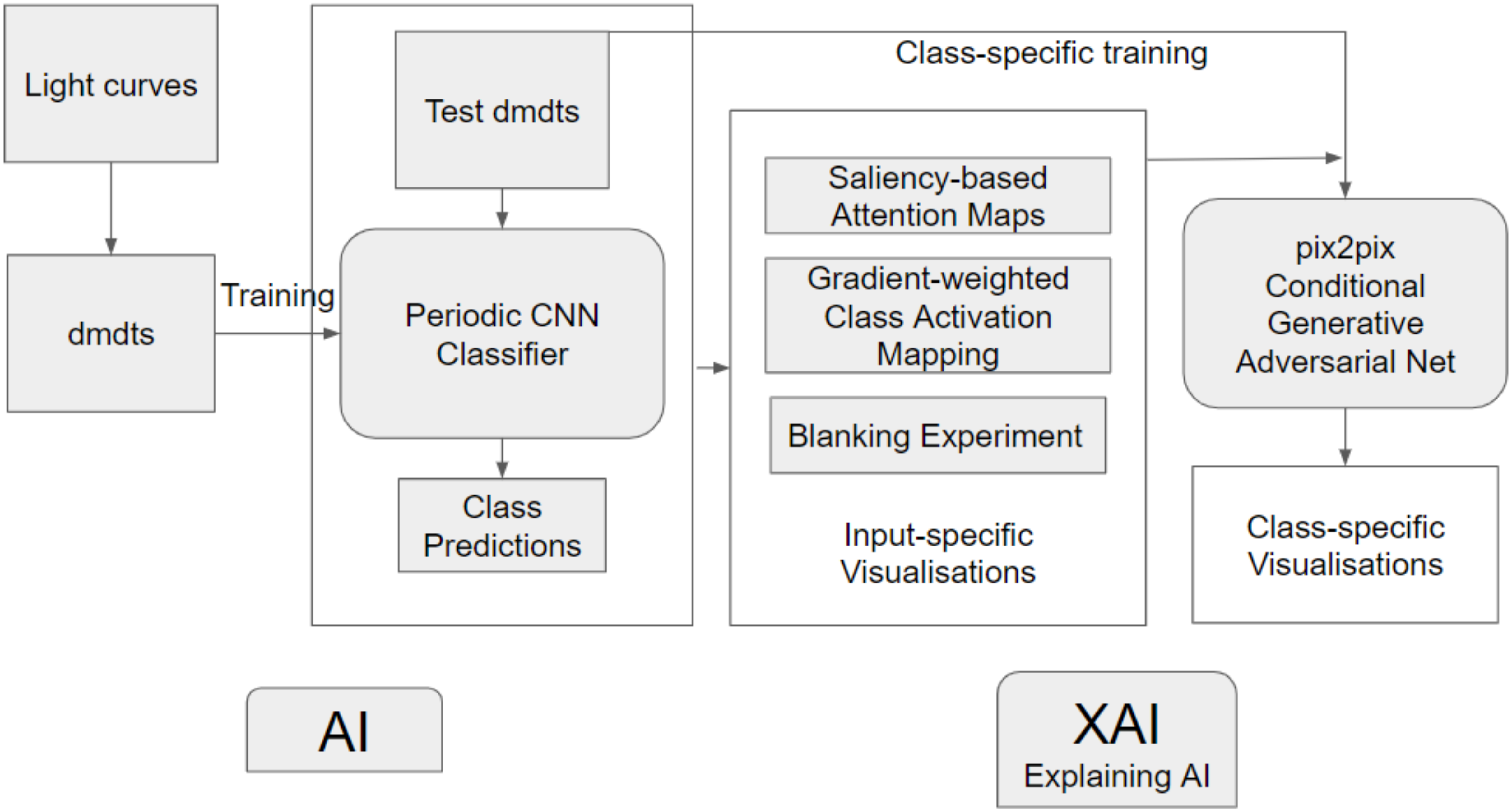
A. Activation Maximization
   - Initial layer filters easy to visualize
   - Generate input image that activates later filters

B. Saliency Maps
   - Gradient of o/p category wrt input image
   - Understanding attention of the classifier

C. Class Activation Maps
   - Gradients based on first dense layer
   - Spatial information still intact

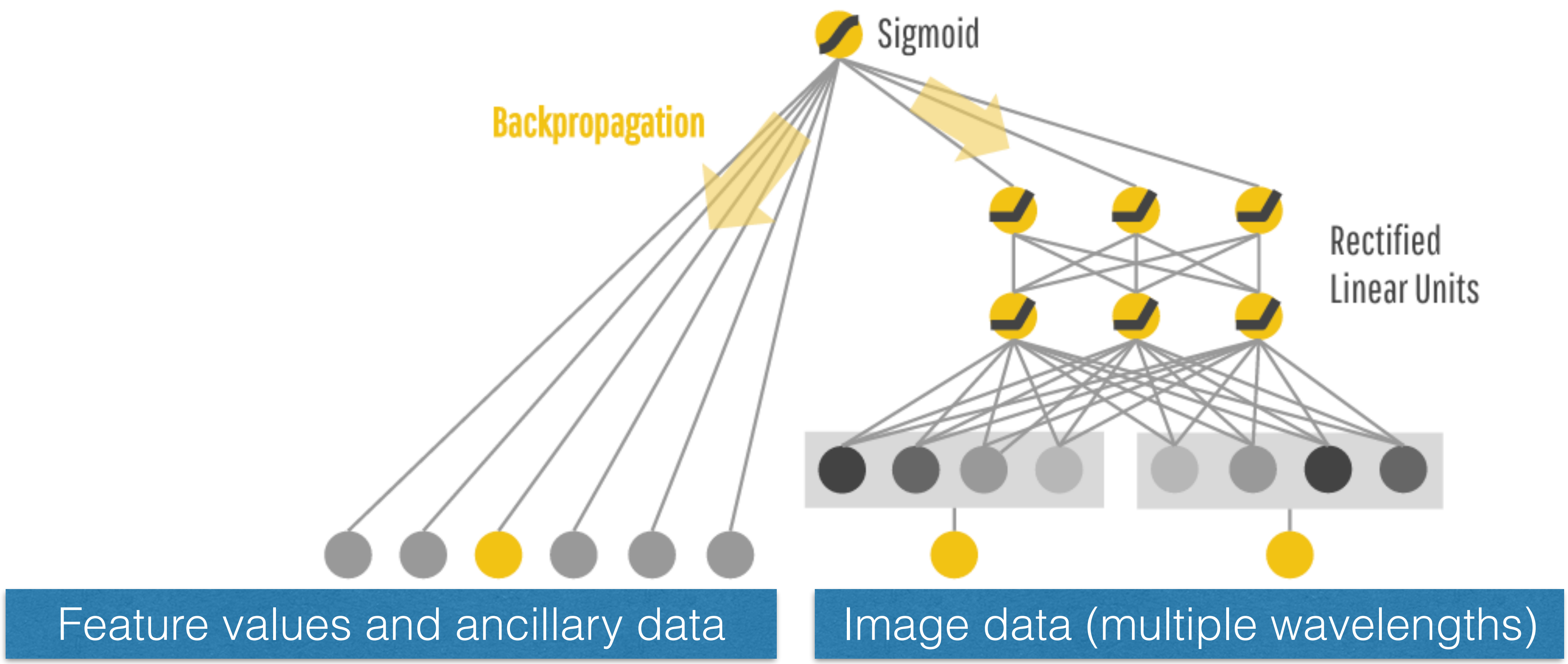


https://raghakot.github.io/keras-vis/


Light curves
dmdts
Training
Test dmdts
Periodic CNN Classifier
Class Predictions
Class-specific training
Saliency-based Attention Maps
Gradient-weighted Class Activation Mapping
Blanking Experiment
Input-specific Visualisations
pix2pix Conditional Generative Adversarial Net
Class-specific Visualisations
AI
XAI
Explaining AI

Meet Gandhi

Combining with unstructured data



The “comments” or metadata become additional features (GoogLeNet)

https://research.googleblog.com/2016/06/wide-deep-learning-better-together-with.html

In <your area of interest> what can you apply deep learning to?

- One speculative example
- One more directly related to your work

Summary

It has become easy to apply DI to astronomy data
There are many low-hanging fruit
Data massaging is required
So is formulating the problem correctly with domain knowledge

Applications to biology are also waiting to be exploited
Larger number of hurdles due to deidentification issues
Also of data fusion

Interpretability and reproducibility are critical