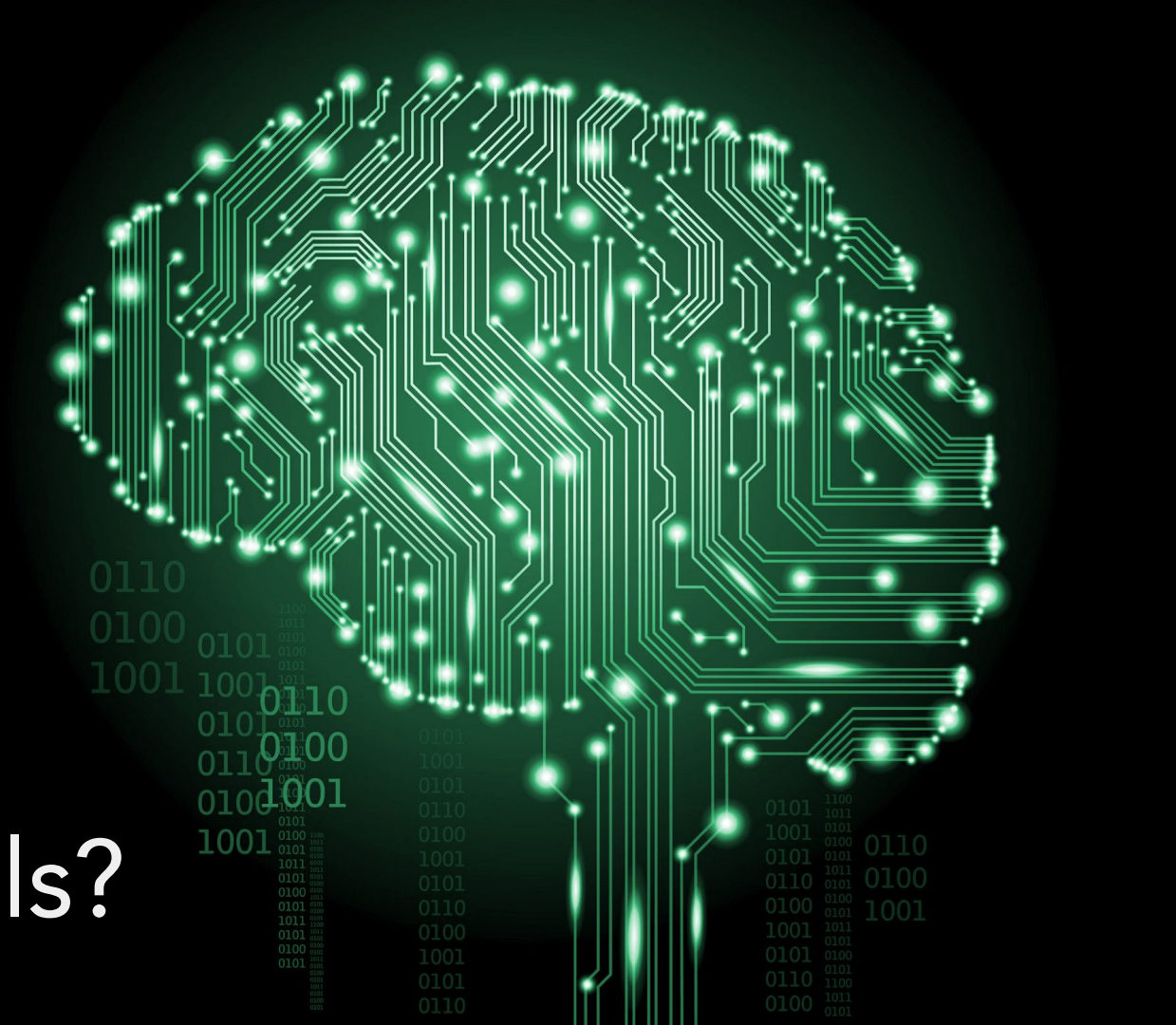
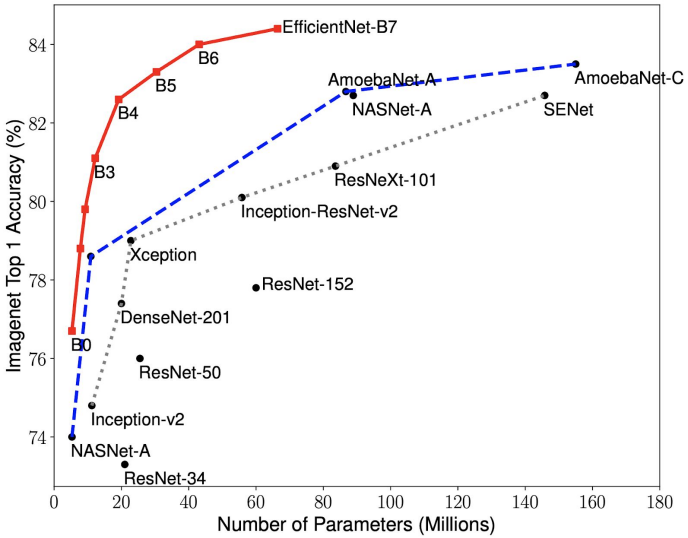
Model Visualization

Understanding models

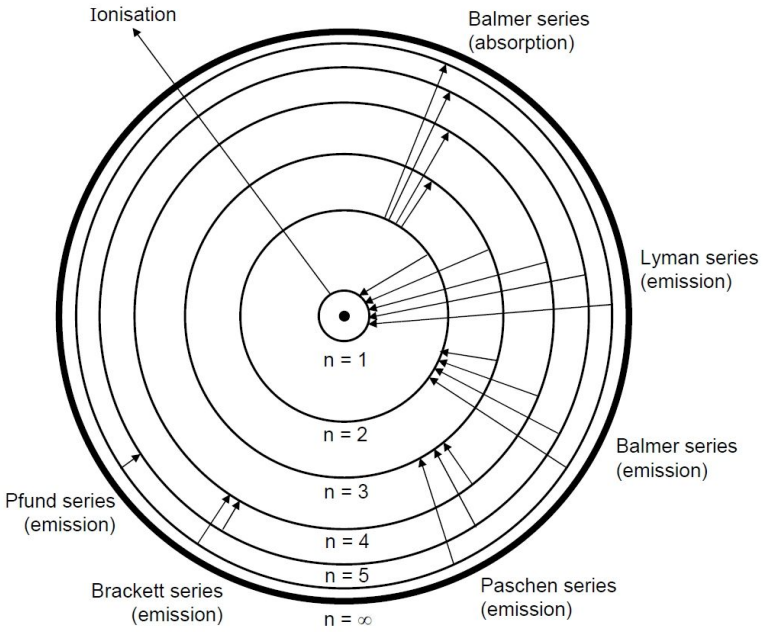
Why do we build models?





Is it to get an extra 0.5% on ImageNet?

No!



We build models to explain how the world works.


STANLEY TUCCI
KIERNAN SHIPKA
THE
SILENCE
A NETFLIX FILM
THEY'RE LISTENING.
APRIL 10 | NETFLIX


FROM THE DIRECTOR OF SUICIDE SQUAD AND END OF WATCH
WARD
WILL SMITH
JOEL EDGERTON
BRIGHT
A NETFLIX FILM
NETFLIX | DEC 22


JASON BATEMAN
LAURA LINNEY
THERE IS NO OFF SEASON
OZARK S2
08.31 | NETFLIX


IN THE NEAR FUTURE
WE MUST SAVE EARTH
OR LEAVE IT
MARGARET QUALLEY
ANTHONY MACKIE
IO
A NETFLIX FILM
JAN 18 | NETFLIX


A NETFLIX ORIGINAL SERIES
QUICKSAND
ALL EPISODES
APRIL 5 | NETFLIX


THE ACCLAIMED SERIES RETURNS
THE
CROWN
A NETFLIX ORIGINAL SERIES
NEW EPISODES
DECEMBER 8 | NETFLIX

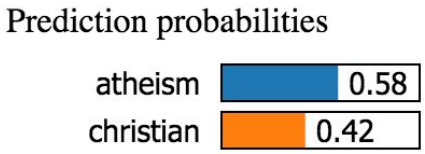

A NETFLIX FILM
ANTHONY ANDERSON
KHALIL EVERAGE
BEATS
RISE ABOVE THE NOISE
JUNE 19 | NETFLIX


SAM WORTHINGTON
TAYLOR SCHILLING
TOM WILKINSON
A NETFLIX FILM
THE
TITAN
MAR 30 | NETFLIX

Model explanations are often more useful than the predictions...

https://www.strong.io/blog/deep-neural-networks-go-to-the-movies

...interpreting also makes us better data scientists.



Text with highlighted words

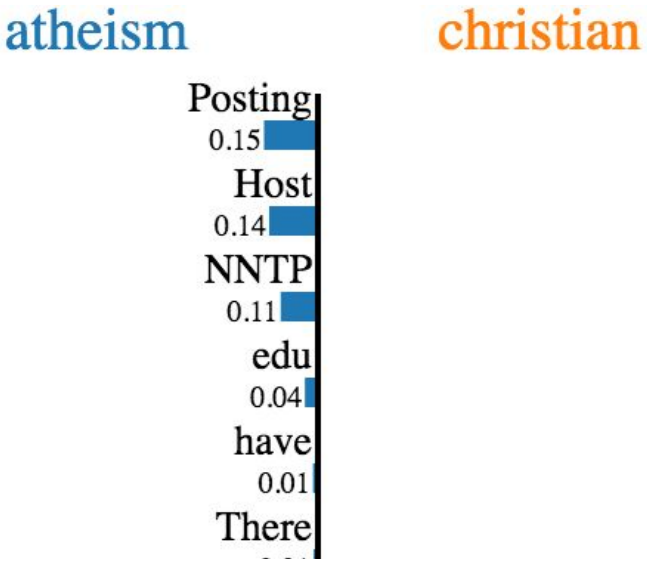
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

This model has 92.4% accuracy.

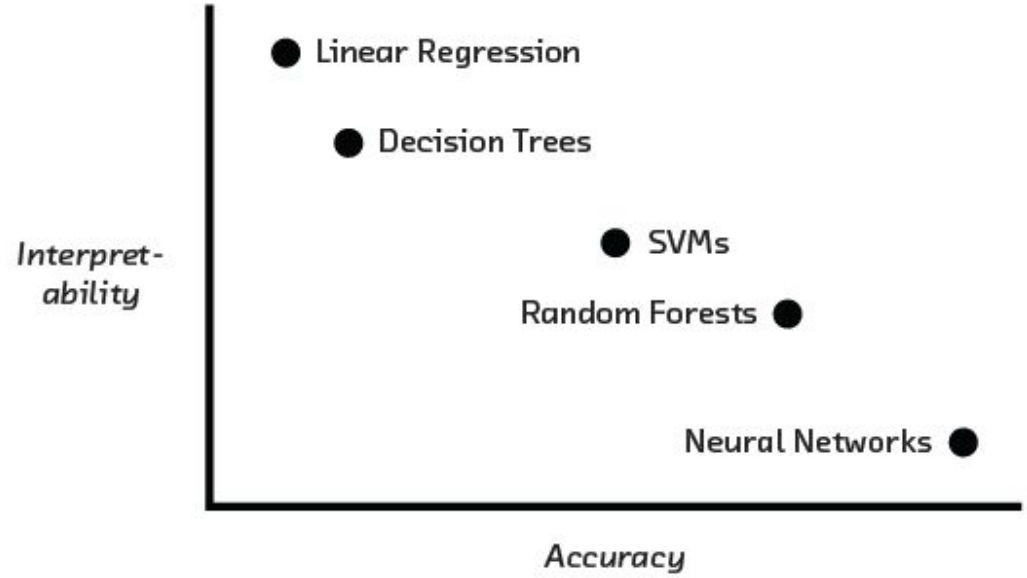
...interpreting also makes us better data scientists.



This model has 92.4% accuracy.

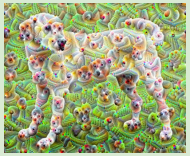
Even though it's looking at junk.

Our problem for this lecture:

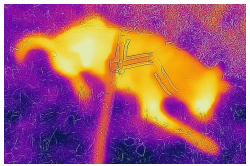


Neural Networks are powerful models, but harder to interpret.

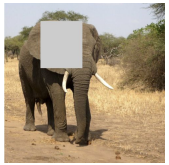
Today:



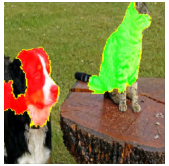
Layer Visualization



Saliency Maps



Occlusion Maps



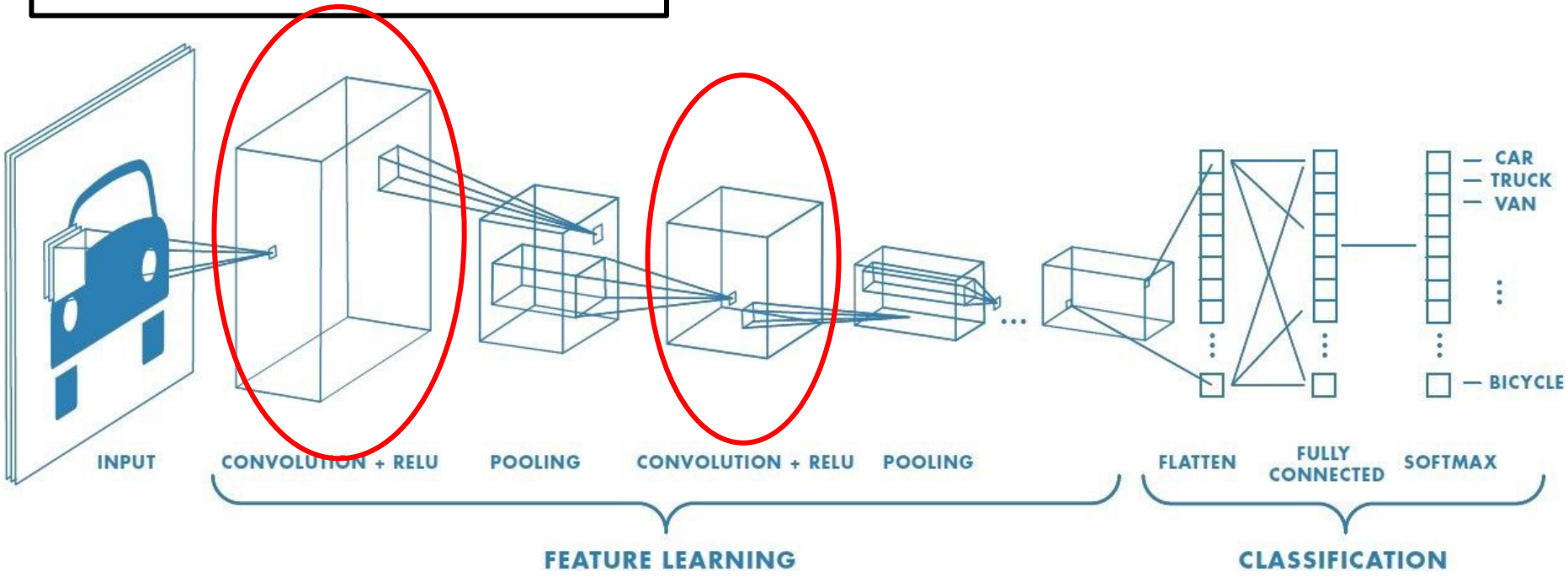
LIME

Layer Visualization

- Techniques shown in this section generalize to any NN architecture
- We will focus on CNNs because their visualizations are more interesting
- Visualization is just the end product of broader interpretation analysis

A Brief Review of CNNs

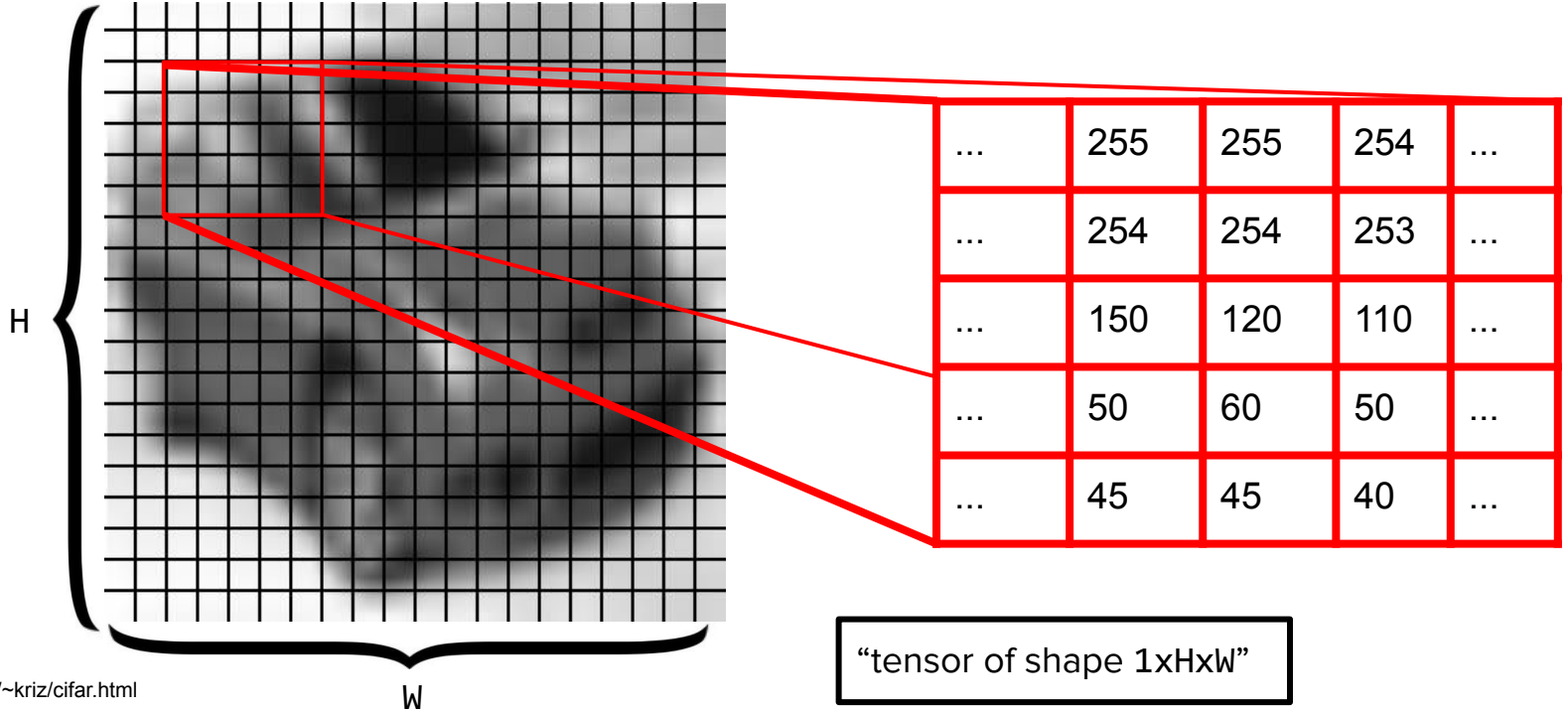
What does it mean to “visualize” a layer?



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

First: What's in an Image?

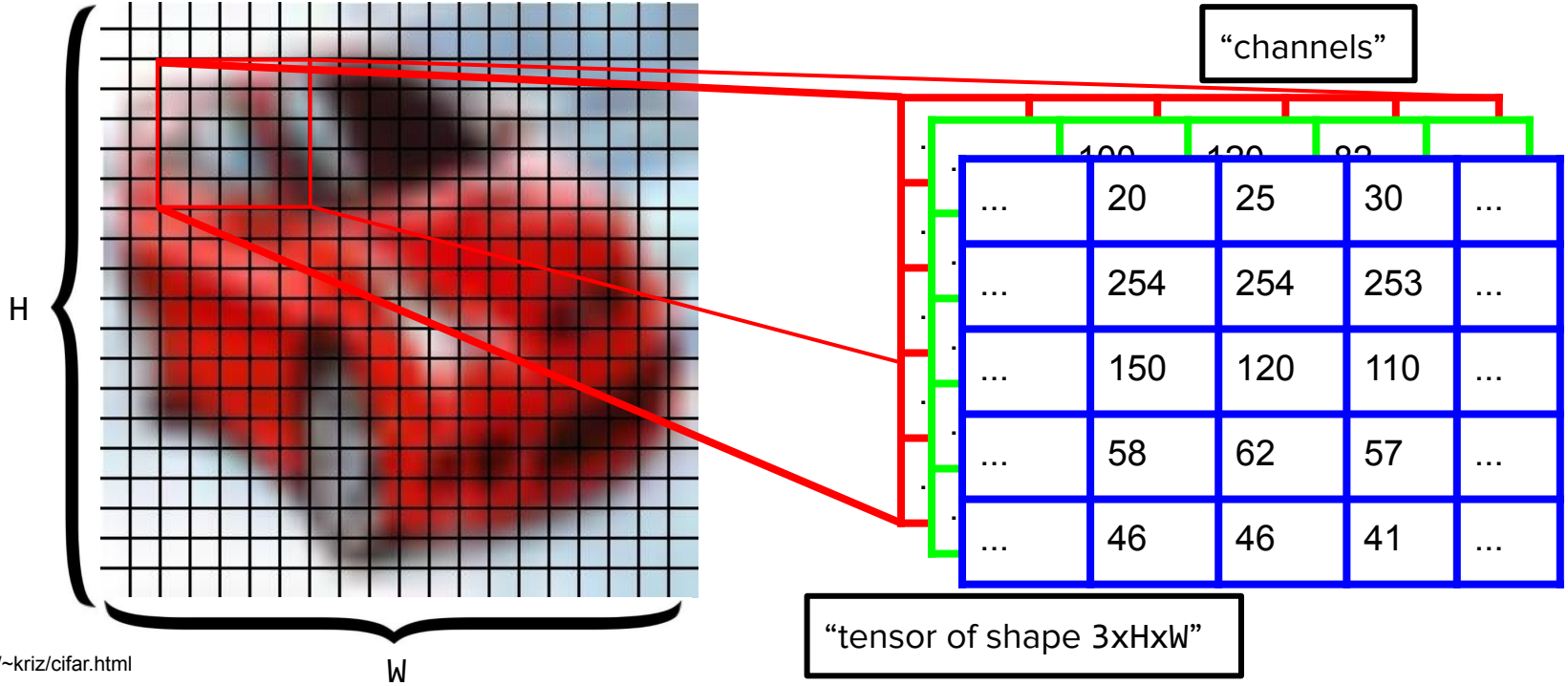
An image is a multidimensional array (a.k.a. "tensor"):



https://www.cs.toronto.edu/~kriz/cifar.html

First: What's in an image?

What about the other colors?

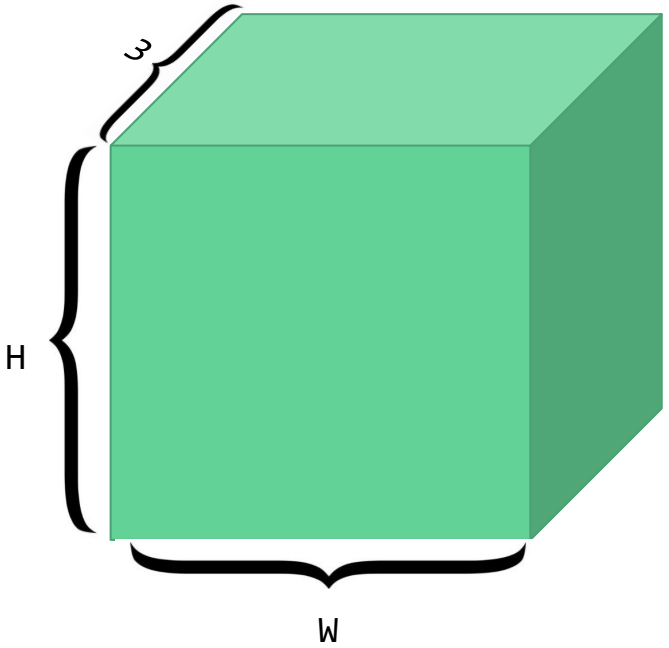


https://www.cs.toronto.edu/~kriz/cifar.html

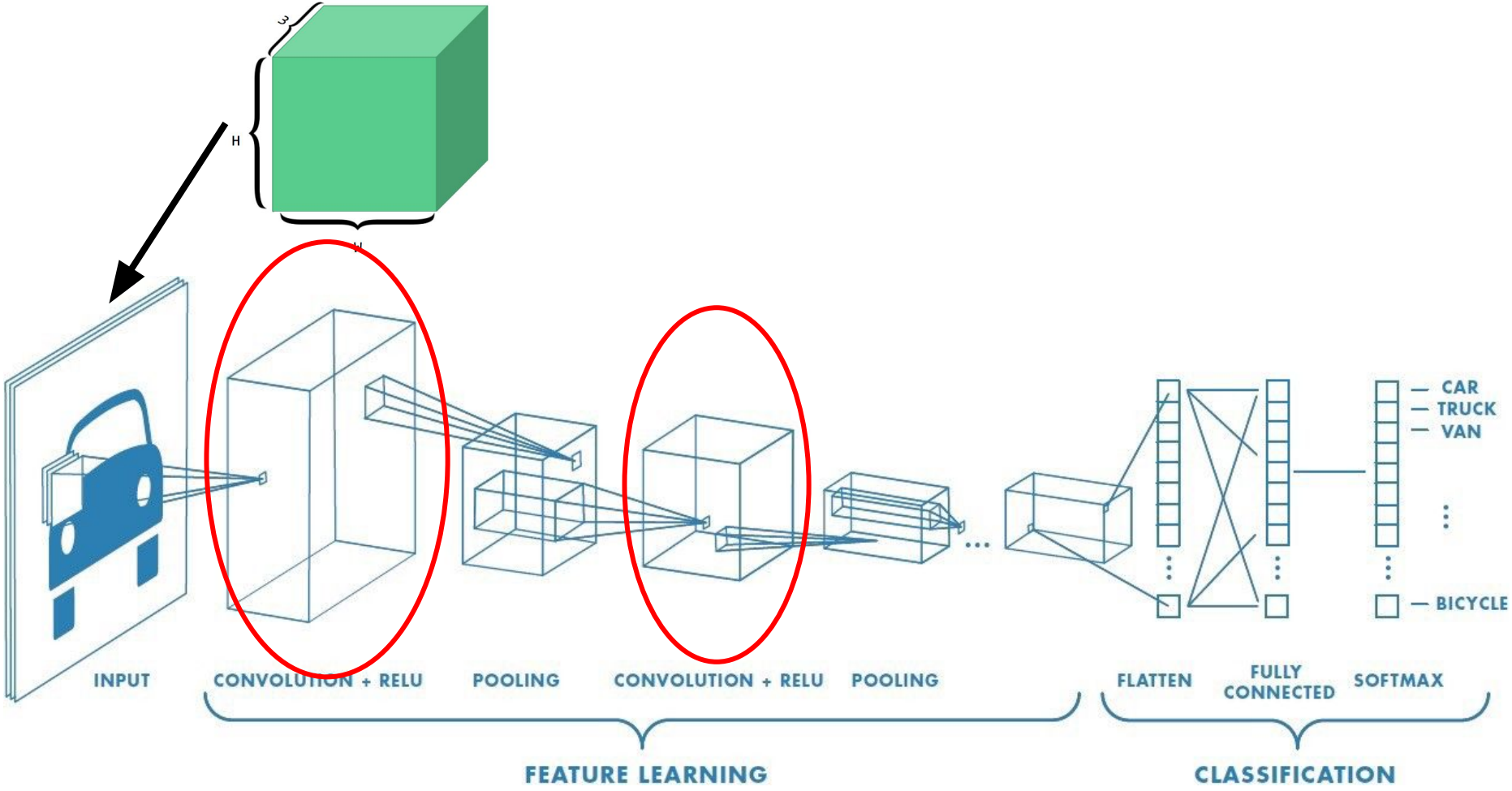
First: What's in an Image?

Takeaway:

=

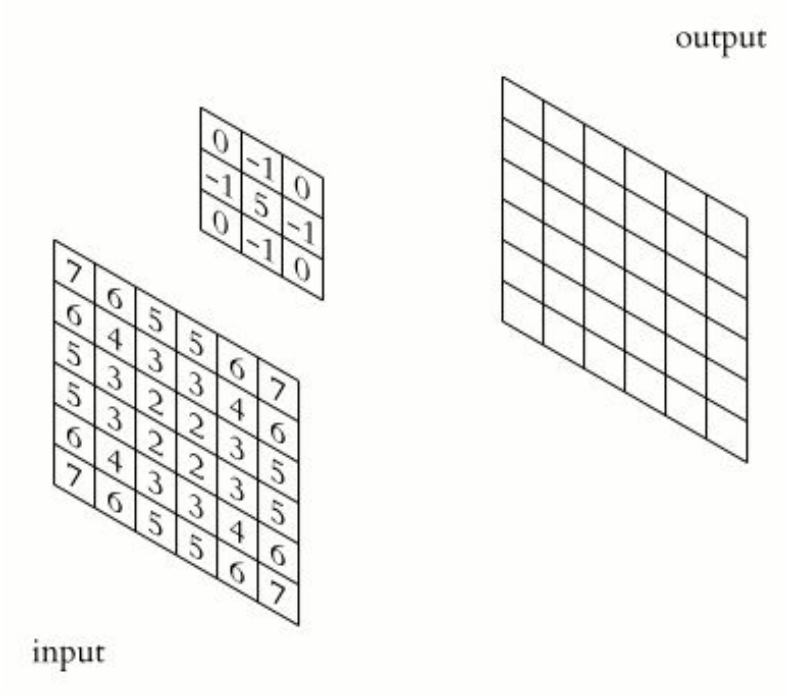


https://www.cs.toronto.edu/~kriz/cifar.html

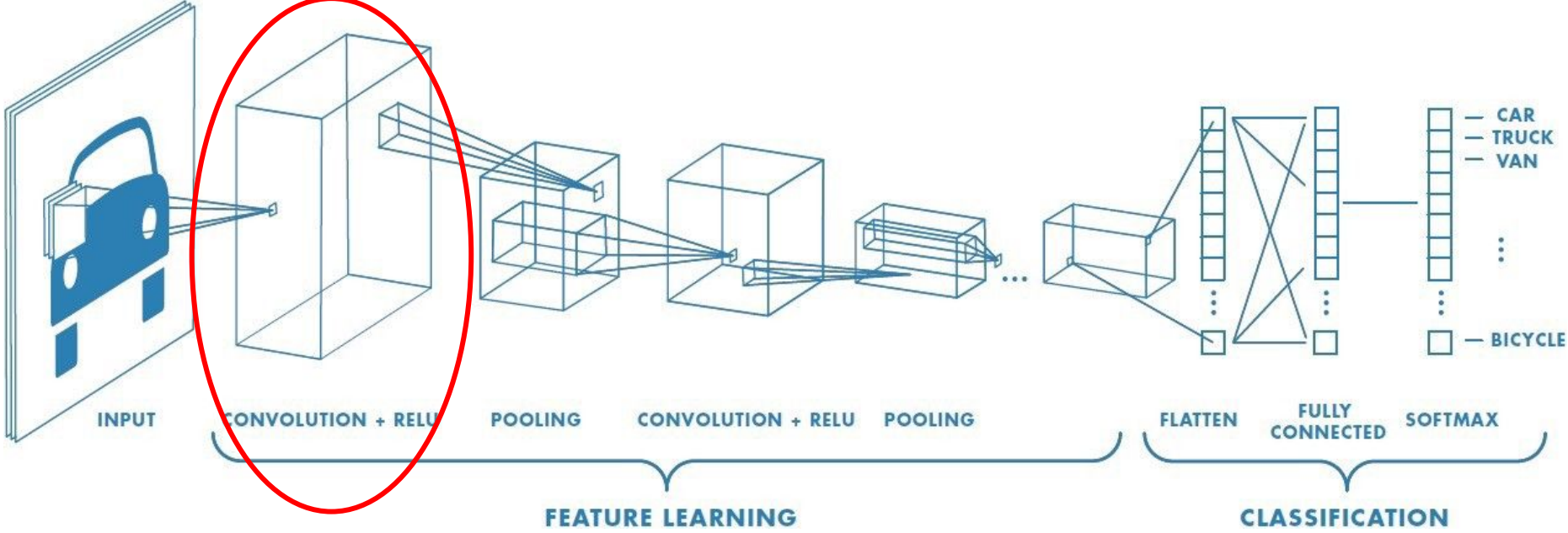

H
INPUT
CONVOLUTION + RELU
POOLING
CONVOLUTION + RELU
POOLING
FLATTEN
FULLY CONNECTED
SOFTMAX
CAR
TRUCK
VAN
BICYCLE
FEATURE LEARNING
CLASSIFICATION

Breaking Down a Convolutional Layer

- Convolutional layer applies a set of **independent** transformations to its input
- Each transformation is characterized by a “kernel”
- Mathematically: convolution is an easy arithmetic operation - **differentiable with respect to the kernels**



https://commons.wikimedia.org/wiki/File:3D_Convolution_Animation.gif

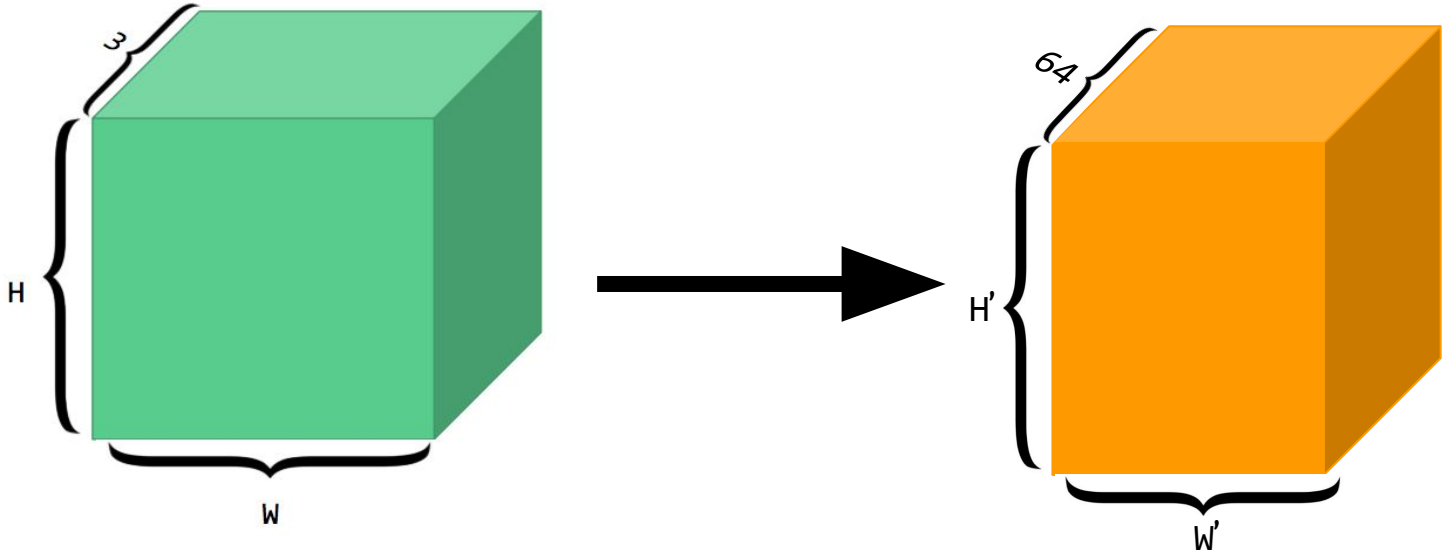



```
Conv2d(3, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
```

3 input channels

64 output channels = 64 kernels

https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53




```
Conv2d(3, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
```

3 input channels

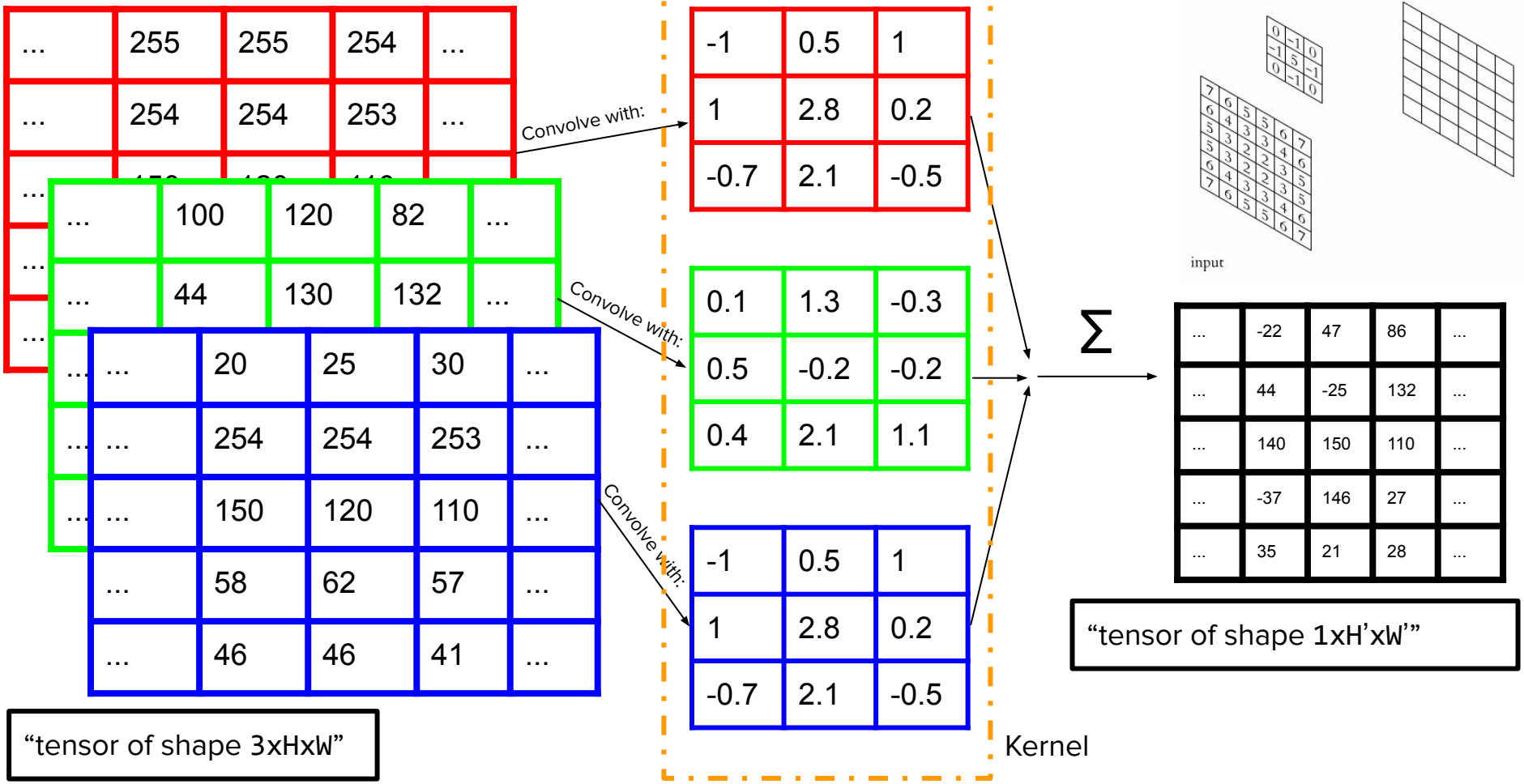
64 output channels
=
64 kernels

Each kernel amplifies a different characteristic of its input.

This is an example of “automatic feature engineering.”

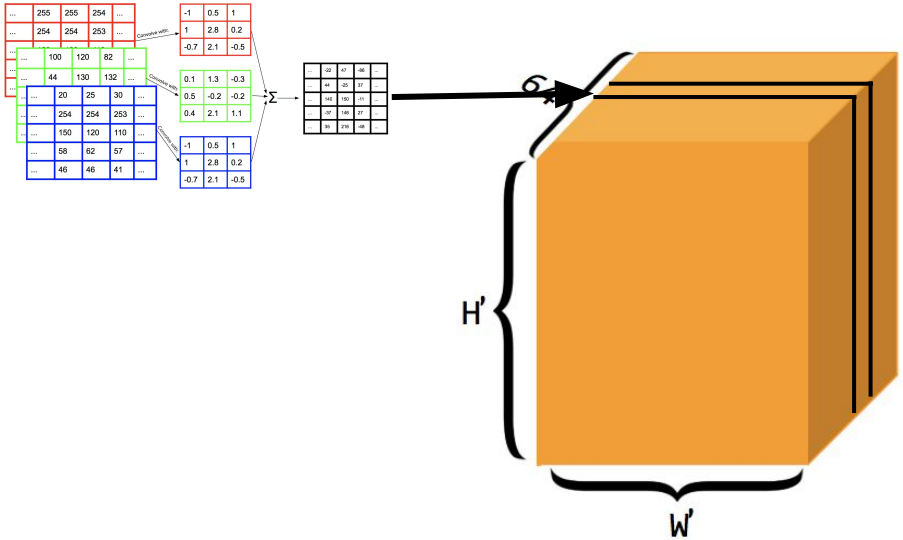
The precise values of these kernels are determined during training via backpropagation -

They are the essence of a trained CNN!



https://commons.wikimedia.org/wiki/File:3D_Convolution_Animation.gif

```
Conv2d(3, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
```



Result of layer given input

What does it mean to “visualize” a layer?

One answer: see what each kernel in that layer “reacts” strongly to; look for broader themes

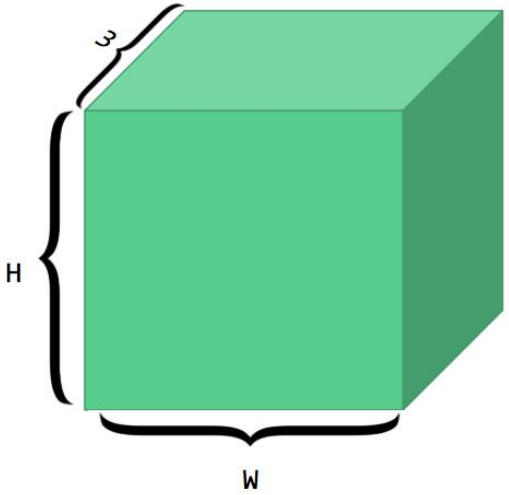
Tying Things Together: Channel Visualizations

- Premise: create an artificial image that maximizes the activation of a channel in a convolutional layer
- Process is exactly the same as training a neural network!
    - i.e. gradient descent

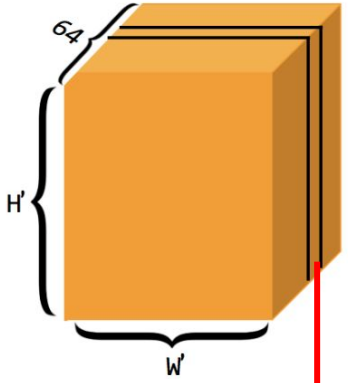
Optimizing over Input Images

Given a pretrained CNN:

Optimize this input tensor:



... to maximize an objective - e.g. :



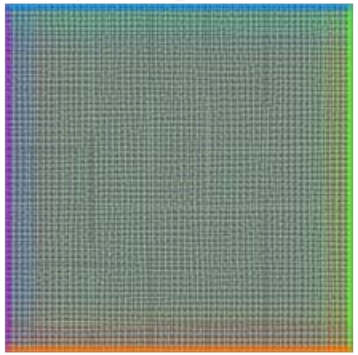
$$\frac{1}{H_{conv} W_{conv}} \sum_{h=0}^{H_{conv}} \sum_{w=0}^{W_{conv}} \left(i^{th} \text{ layer convolution result}_{[h,w,c]} \right)$$

for some layer **`i` and** channel `c` (in that layer) of our choosing.

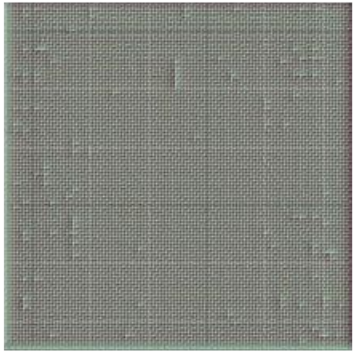
Starting Shallow

VGG Layer block2_conv1, Channel 1

VGG Layer block2_conv1, Channel 3

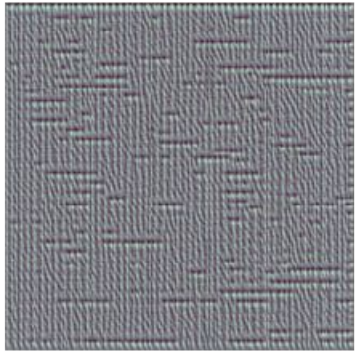


VGG Layer block2_conv1, Channel 5



VGG Layer block2_conv1, Channel 7

VGG Layer block2_conv1, Channel 8



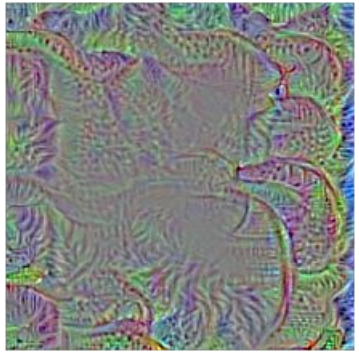
VGG Layer block2_conv1, Channel 11

VGG Layer block2_conv1, Channel 16

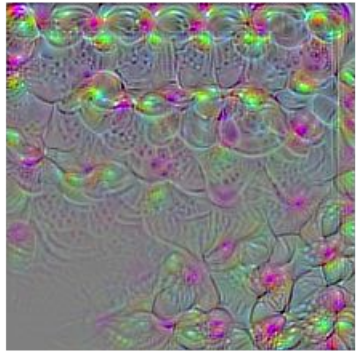
VGG Layer block2_conv1, Channel 18

Jumping in the Deep End

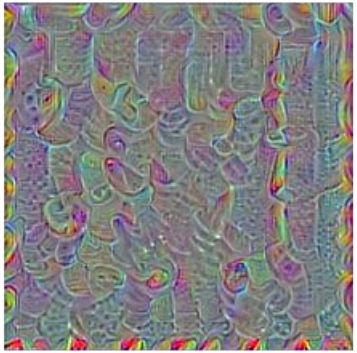
VGG Layer block5_conv1, Channel 0



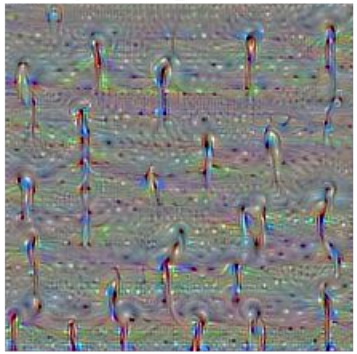
VGG Layer block5_conv1, Channel 1



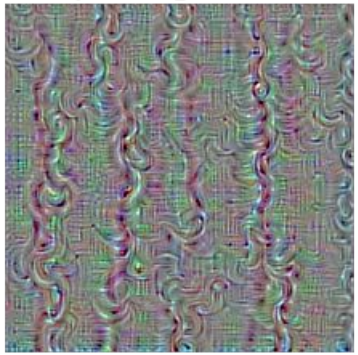
VGG Layer block5_conv1, Channel 3



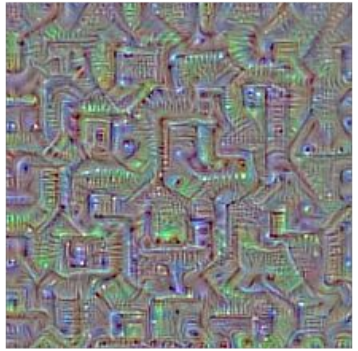
VGG Layer block5_conv1, Channel 7



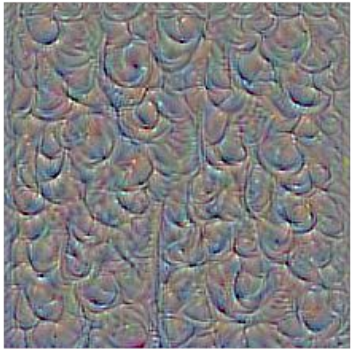
VGG Layer block5_conv1, Channel 8



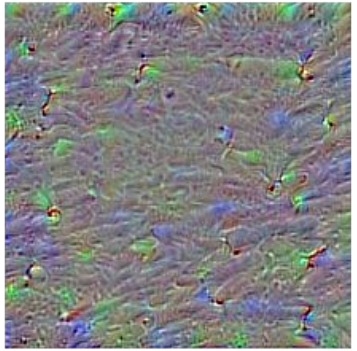
VGG Layer block5_conv1, Channel 11



VGG Layer block5_conv1, Channel 18



VGG Layer block5_conv1, Channel 19

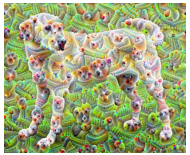


Further Exploration

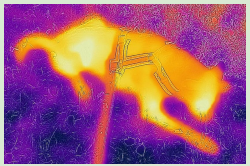
- Image augmentations
  - Gaussian + other types of blurring
  - Stochastic jitter
  - Affine transformations

- Different objective functions
  - Entire layer, Deep Dream
  - Class-wise visualizations - optimize input image to maximize probability of a given class

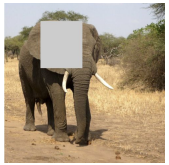
Today:



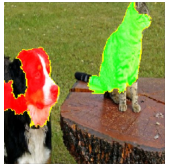
Layer Visualization



Saliency Maps



Occlusion Maps



LIME

Saliency Maps

Can you identify
this object?

Saliency Maps

How about now?

Same source, same zoom.

Saliency Maps

- Saliency maps were originally a model for human attention
- Refers to the likelihood that the observer will focus on certain parts of an image

Source: https://unsplash.com/photos/UTw183oZPf0

Saliency Maps

Input image

Gradients across RGB channels

Max gradients

Overlay

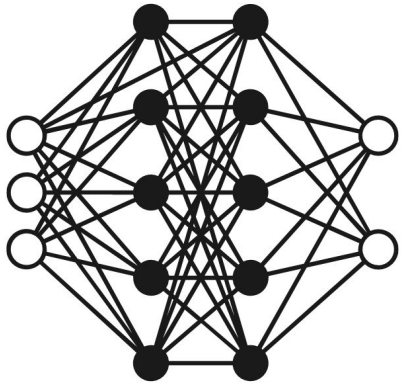
For neural nets - the gradient in the predicted class wrt pixel value, using backpropagation:

$$\frac{\Delta\ Predicted\ Class}{\Delta\ Pixel\ Value}$$

Source: https://medium.com/@ODSC/visualizing-your-convolutional-neural-network-predictions-with-saliency-maps-9604eb03d766

Saliency Maps

Get predicted class

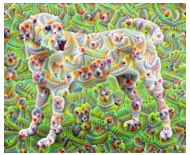


Predicted Class:
Owl

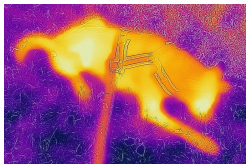
Backpropagate to get the gradient
w.r.t. each pixel of the image

Source: https://arxiv.org/pdf/1312.6034.pdf

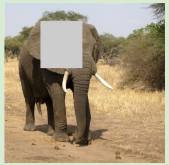
Today:



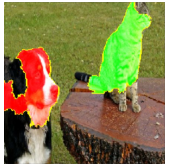
Layer Visualization



Saliency Maps



Occlusion Maps



LIME

Occlusion Maps

What is brand
is this car?

Source: https://unsplash.com/photos/UTw183oZPf0

Occlusion Maps

What is brand
is this car?

Source: https://unsplash.com/photos/UTw183oZPf0

Occlusion Maps

Source: https://unsplash.com/photos/UTw183oZPf0

Occlusion Maps

Source: https://unsplash.com/photos/UTw183oZPf0

Occlusion Maps

Source: https://unsplash.com/photos/UTw183oZPf0

Occlusion Maps

Source: https://unsplash.com/photos/UTw183oZPf0

Occlusion Maps

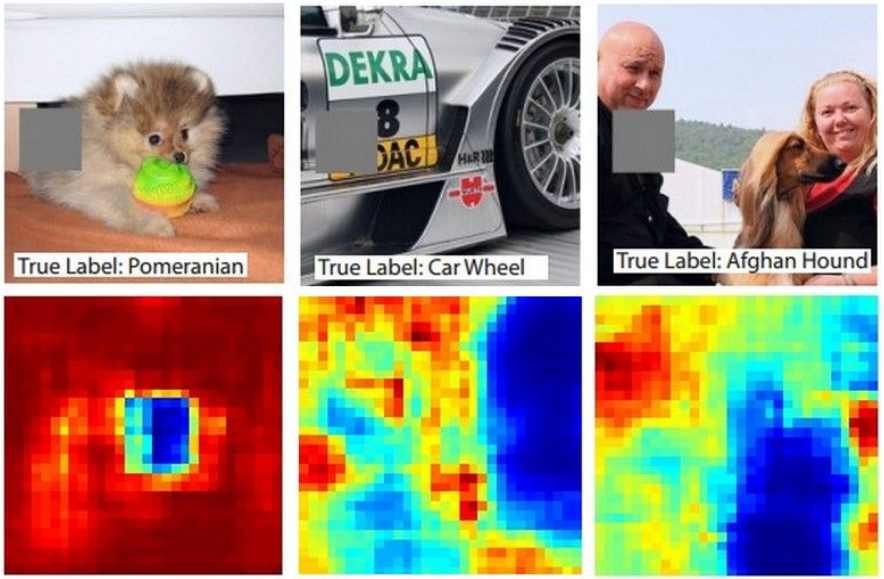
Source: https://unsplash.com/photos/UTw183oZPf0

Occlusion Maps

Source: https://unsplash.com/photos/UTw183oZPf0

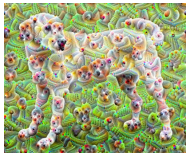
Occlusion Maps

- Zero out a square patch of image, predict and measure accuracy
- Need correct class *a priori*
- nb Saliency is a mathematical technique, Occlusion is an engineering technique

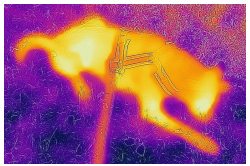


Source: http://cs231n.github.io/understanding-cnn/

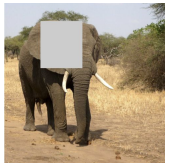
Today:



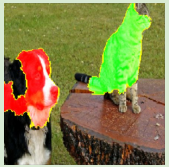
Layer Visualization



Saliency Maps

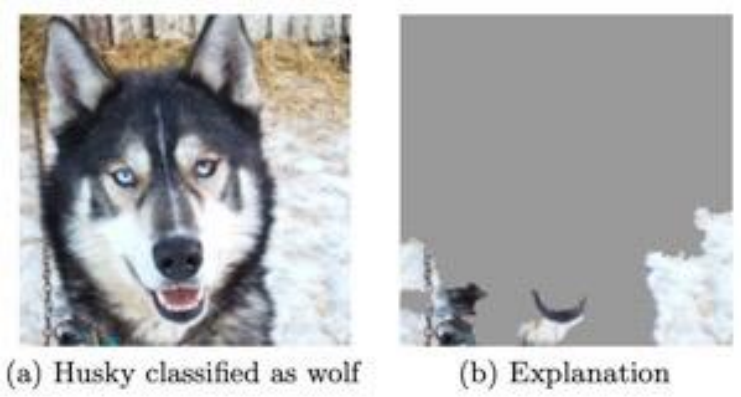


Occlusion Maps



LIME

LIME



(a) Husky classified as wolf (b) Explanation

LIME reveals this (bad) CNN thinks snow=wolf

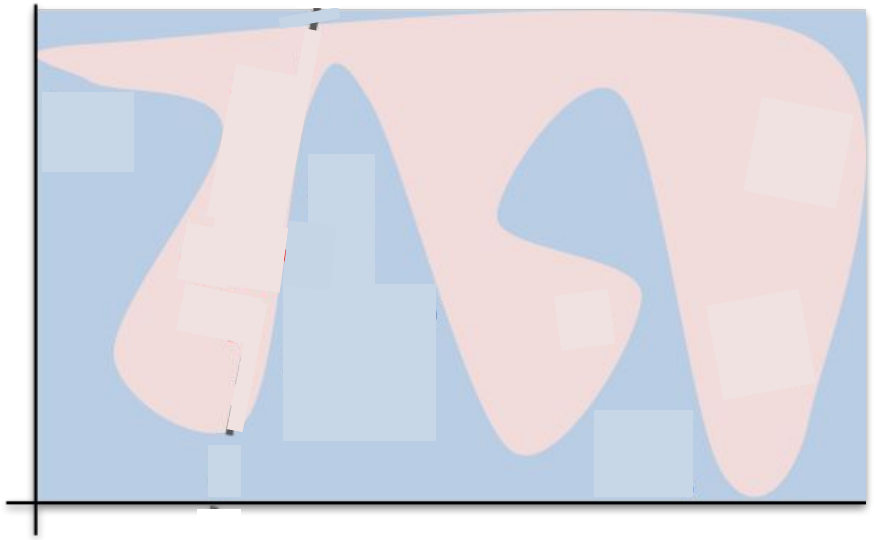
- Local Interpretation for Model-Agnostic Explanations
- Like Occlusion Mapping, perturb input and watch model accuracy
- Unlike Occlusion Mapping - it's not just for CNNs! Works with general model types

Source: http://cs231n.github.io/understanding-cnn/

LIME - How it works

Local interpretability:

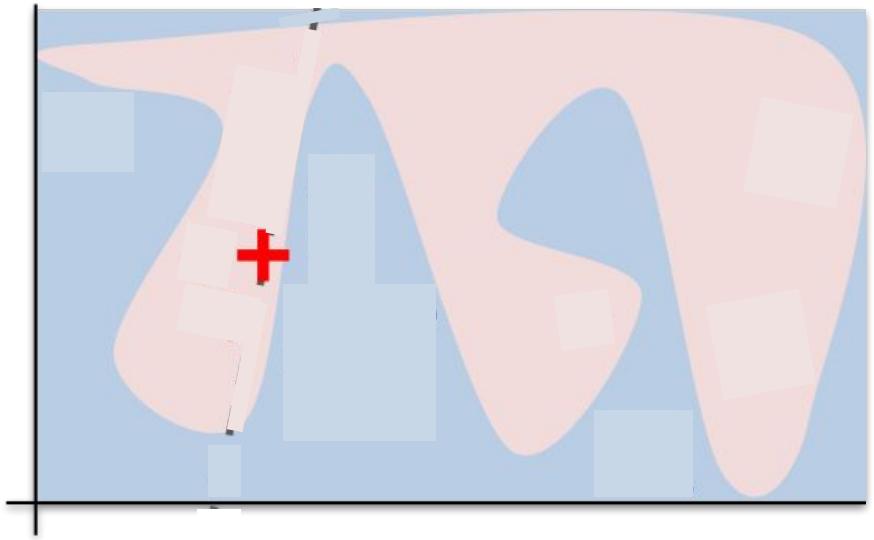
1: Take a black-box model with arbitrary decision boundary



Source: Ribiero, Singh, Guestrin, 2016. '"Why Should I Trust You?" Explaining the Predictions of Any Classifier', https://arxiv.org/pdf/1602.04938v1.pdf

LIME - How it works

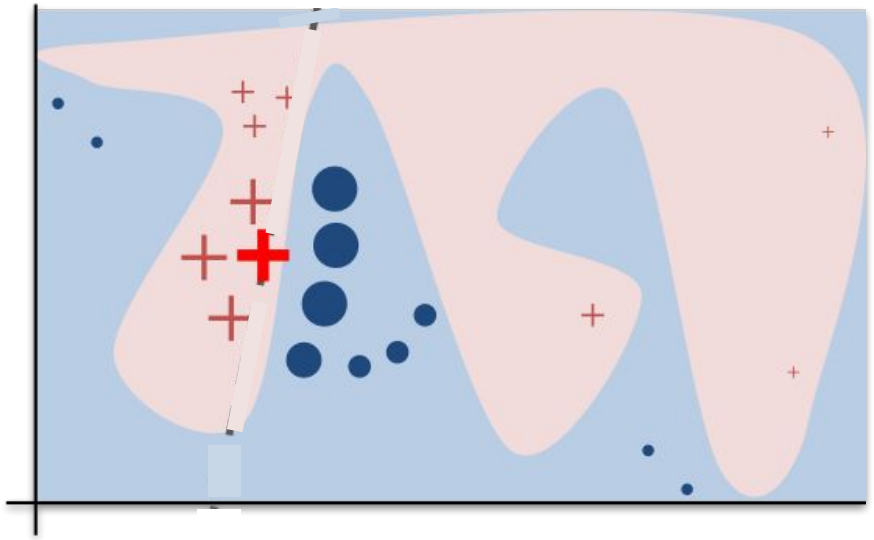
2: Select observation to explain



Source: Ribiero, Singh, Guestrin, 2016. '"Why Should I Trust You?" Explaining the Predictions of Any Classifier', https://arxiv.org/pdf/1602.04938v1.pdf

LIME - How it works

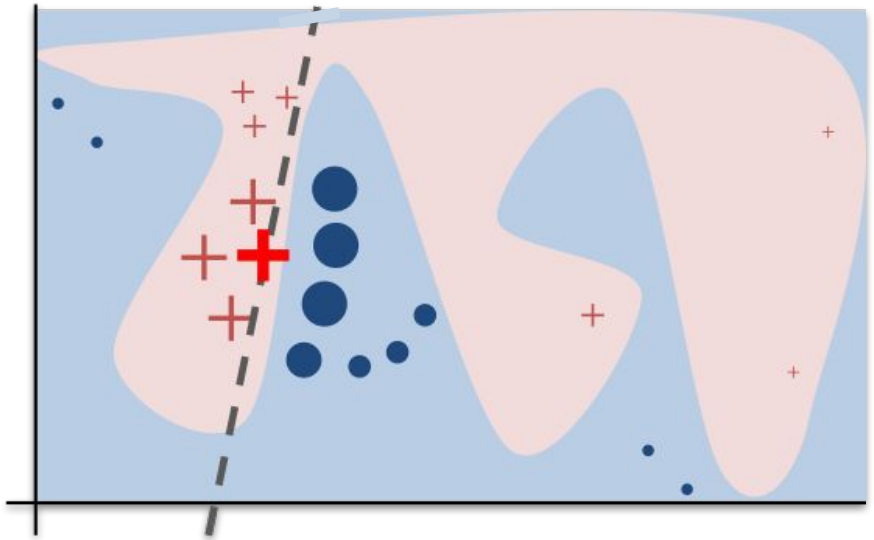
3: Sample around the chosen selected point



Source: Ribiero, Singh, Guestrin, 2016. '"Why Should I Trust You?" Explaining the Predictions of Any Classifier', https://arxiv.org/pdf/1602.04938v1.pdf

LIME - How it works

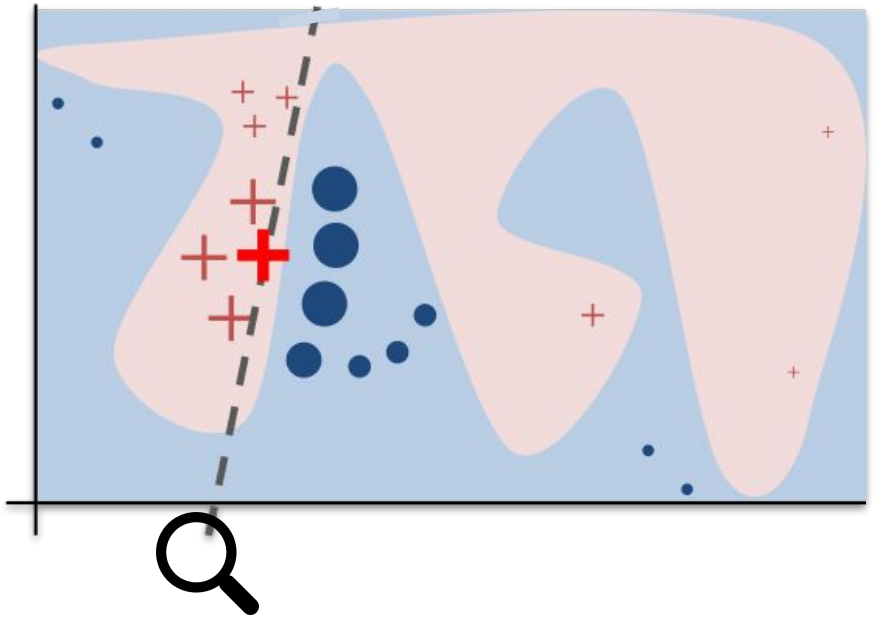
4: Build a local linear regression with the samples



Source: Ribiero, Singh, Guestrin, 2016. "'Why Should I Trust You?" Explaining the Predictions of Any Classifier', https://arxiv.org/pdf/1602.04938v1.pdf

LIME - How it works

5: Feature weights correspond to local explanation!



Source: Ribiero, Singh, Guestrin, 2016. '"Why Should I Trust You?" Explaining the Predictions of Any Classifier', https://arxiv.org/pdf/1602.04938v1.pdf

In closing...

All of these are inherently estimations of a high-dimensional space

Don’t take these as ground truth. Interpretations require interpretations

Interpret a thorough number of samples, and think critically about each

When used properly, these techniques are essential for us to validate and understand our models

In closing...

A herd of sheep grazing on a lush green hillside
Tags: grazing, sheep, mountain, cattle, horse

...interpret every single model you build!

Happy modeling, everyone!!

Source: https://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep