Connect w/ Data Science After Dark

1 **Meetup**

meetup.com/Data-Science-After-Dark

Upcoming Events + RSVP

2 **YouTube**

youtube.com/channel/UC7fA7eMv2745dIez165pYMg

Live streams + past events

3 **Slack**

datascienceafterdark.slack.com

Join us in #data-science

4 **Springfield Tech Calendar**

fwdsgf.com

Local tech-focused events

Welcome to Data Science After Dark!

Attacking a Machine Learning Model
Data Science After Dark, Springfield Missouri

Presented by: Jason Klein, Logic Forte @JasnK
Tuesday, April 21, 2020

meetup.com/Data-Science-After-Dark/events/268654605/

Photo by Shane Rounce

Who am I?

Jason Klein

I have been managing data since 2002. My background is a mix of IT, infosec, software. This led to my interest in security in Machine Learning.

@JasnK
https://jrklein.com/

Attacking a Machine Learning Model

Why we must **protect** Machine Learning models **critical** to our business?

Attackers who can access your model can manipulate inputs to achieve desired outputs. *e.g. Fraud Detection, Profanity and Image Filtering, etc.*

History of Machine Learning

1950 - Turing Test

1958 - Perceptron neural network

1967 - Nearest neighbor algorithm

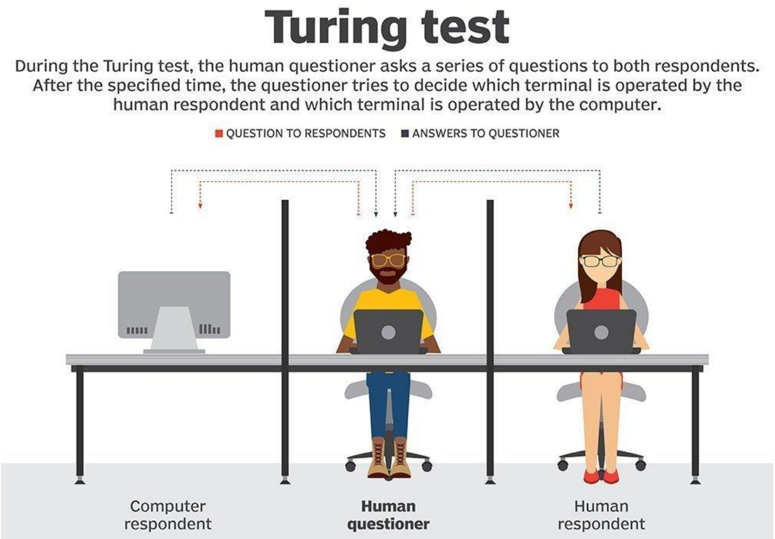
1979 - The Stanford Cart navigates obsticles

1990s - Machine Learning shifts from knowledge-based to data-driven models

2006 - Deep Learning term coined

2010 - Microsoft Kinect tracks 20 features at 30 frames per second

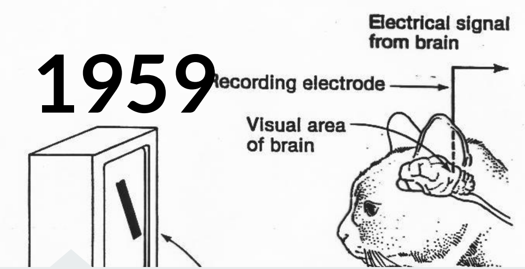
2010 ImageNet, 2011 Google Brain, 2012 YouTube Cat Finder, 2014 DeepFace



Turing test

During the Turing test, the human questioner asks a series of questions to both respondents. After the specified time, the questioner tries to decide which terminal is operated by the human respondent and which terminal is operated by the computer.

History of Image Classification

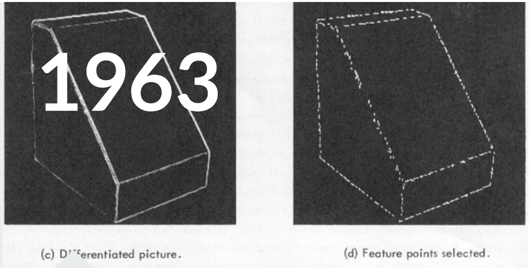


1959

Hubel and Wiesel
discover core principal of edge detection

Russell Kirsch
developed first digital image scanner [1][2]

1959

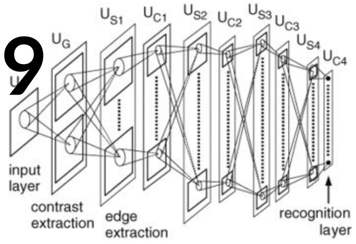


1963

Lawrence Roberts
converts 2D photos into line drawings

History of Image Classification

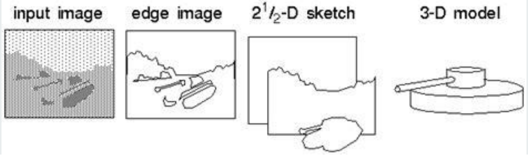
1979



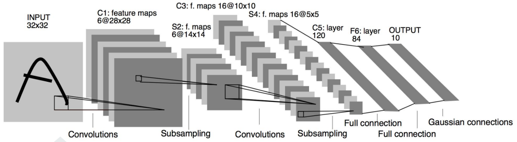
Kunihiko Fukushima created neocognitron neural network [1] [2]

David Marr established that vision is hierarchical [3]

1982

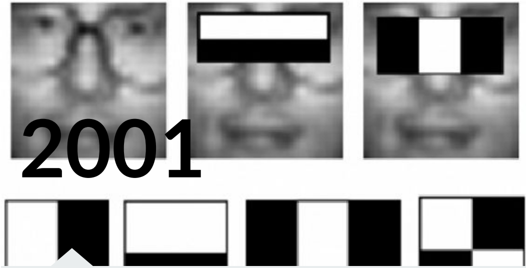


1989



Yann LeCun LeNet-5, first modern convnet, leading to MNIST imageset [4][5]

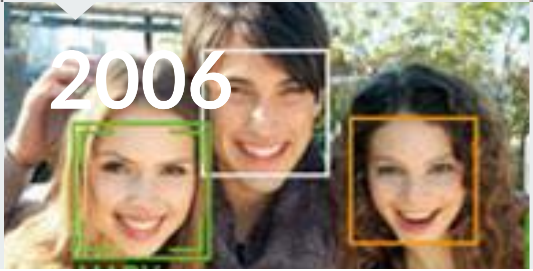
History of Image Classification



2001

Viola and Jones

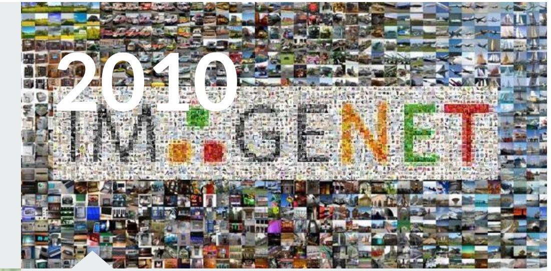
create real-time face detection framework[1]



2006

Fujitsu

releases camera with real-time Viola/Jones face detection

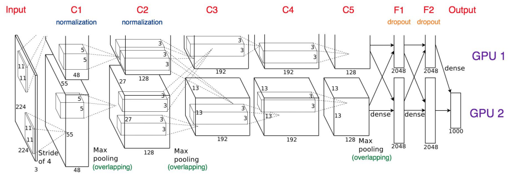


2010

ImageNet project

begins annual contest, recognition of 1000 image categories [2][3]

History of Image Classification



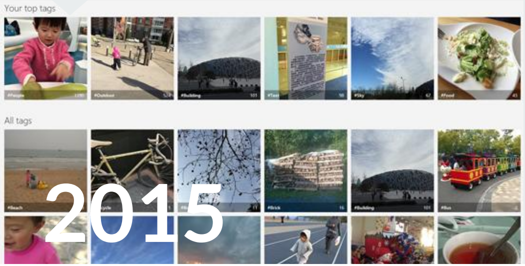
2012

Alex Krizhevsky

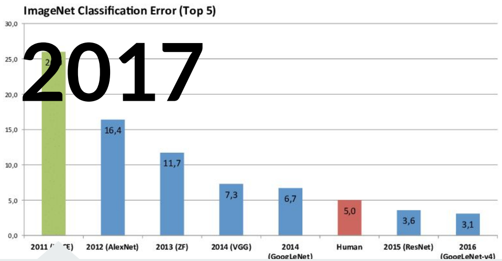
AlexNet wins ImageNet, GPU model accuracy up from 75% to 84.7% [1][2]

Microsoft

team wins ImageNet with 95.1% accuracy [3][4]



2015



2017

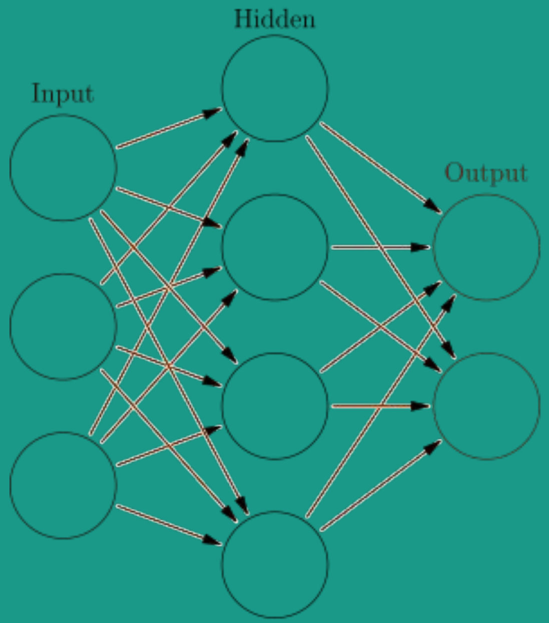
ImageNet

29 of 38 teams >95% accuracy. Developing 3D competition. [5]

Basics of Image Classification

1- Train a Neural Network

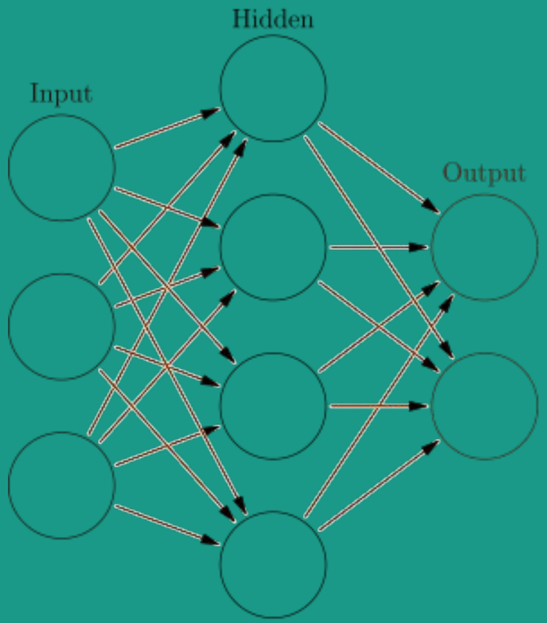
- Input Tagged Images
- Train Hidden Layers
- Output Neurons



Basics of Image Classification

2- Use the Neural Network

- Input Model File
- Input Image
- Output Prediction %

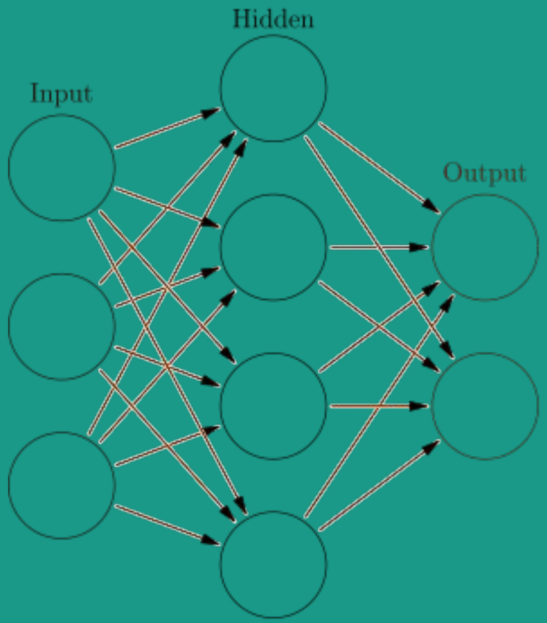


Basics of Image Classification

3- Profit

If you trained your own model, keep your model secret to help avoid attacks.

Ready to learn more? Deep Dive [1][2]



Attacking a Machine Learning Model

I will demonstrate how easily we can attack an image classification model.

I will feed an image of a specific animal into the model and demonstrate how we can modify a single pixel in the original image to convince the model that the image is a different specific/desired animal.

Demo

Photo by Foo Bar

Dog or Frog [1][2][3]

I first learned the potential of this attack when I encountered the “Dog or Frog” problem during the picoCTF 2018 competition:

“I'm going to a Halloween Party soon, and need a costume. Could you use one of those fancy new Convolutional Neural Networks and help me out? I'd like a photo of me as a tree frog, and have the CNN recognize me as such, but still have the photo look like me.”

Photo by Foo Bar

Dog or Frog .. 2

The goal here is to make an adversarial example using the photo of Trixi as a starting point. The website will classify it using the Keras pre-trained MobileNet network, for which you can download a copy. The image needs to do the following:

- Classify as a Tree Frog, at 95% confidence
- Be similar to the original image (max 2 bit difference using p hash)

This vulnerability is not specifically for MobileNet, and works against others as well. MobileNet was chosen as it uses less memory.

Photo by Foo Bar

Dog or Frog .. 3

I want to make it clear that this isn't a stego, web, or "find a photo of the author's dog wearing a frog hat" problem. The intended solution is a photo that is clearly Trixi, but trick MobileNet into thinking there'sa tree frog in it, rather than a dog.

A sample attack image is providedin the source code, that's recognized as a sealion. It looks squished due to the preprocessing of the network. It looks nearly identical to the preprocessed image without any attack present.

Photo by Foo Bar

Dog or Frog .. 4

The linked Google article [1] is about their challenge on trying to build a defenses against this class of attack, and shows where academia is at in terms of both attacks and defenses.

The solution runs in < 10s on my CPU.

I can get 99+% confidence, and 0 bit difference.

Photo by Foo Bar

Dog or Frog .. 5

```
Photo category                          tree_frog
Photo is of a frog                      True
Photo confidence                        0.52180797
P hash distance from original photo     13
Top Preds
    [('n01644373', 'tree_frog', 0.52180797),
    ('n01644900', 'tailed_frog', 0.4586374),
    ('n01675722', 'banded_gecko', 0.018022738),
    ('n01641577', 'bullfrog', 0.0010392488),
    ('n01694178', 'African_chameleon',
    0.0001692069)]
```

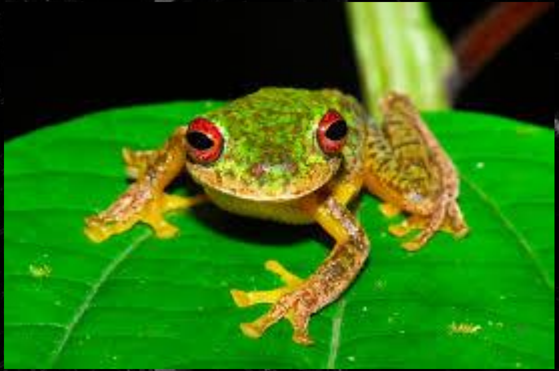


Photo by Foo Bar

Dog or Frog .. 6

```
VERSIONS: macOS 10.15.4, python 3.7.0, pip 20.0.2
SOURCE: solution.py

$ cd ~/Code/dsad-picoctf-2018-dog-or-frog/
$ pip install tensorflow keras Pillow numpy ImageHash
$ python3 solution.py

Using TensorFlow backend.
Model's predicted likelihood that the image is a tree frog: 1.0298706e-09%
Model's predicted likelihood that the image is a tree frog: 1.7842248%
Model's predicted likelihood that the image is a tree frog: 87.817872%
Model's predicted likelihood that the image is a tree frog: 98.974693%
Model's predicted likelihood that the image is a tree frog: 90.498799%
Model's predicted likelihood that the image is a tree frog: 99.204844%

$ ls -l
-rw-r--r--@   1 jrk  staff  17271048 Apr 21 15:42 model.h5
-rw-r--r--    1 jrk  staff      3856 Apr 21 15:54 solution.py
-rw-r--r--@   1 jrk  staff      1641 Apr 21 15:49 solution_template.py
-rw-r--r--@   1 jrk  staff   2534464 Sep 24  2018 trixi.png
-rw-r--r--    1 jrk  staff    110468 Apr 21 16:17 trixi_frog.png
-rw-r--r--@   1 jrk  staff    116308 Sep 24  2018 trixi_sealion.png
```

Photo by Foo Bar

Dog or Frog .. 7

trixi.png

trixi_sealion.png

trixi_frog.png

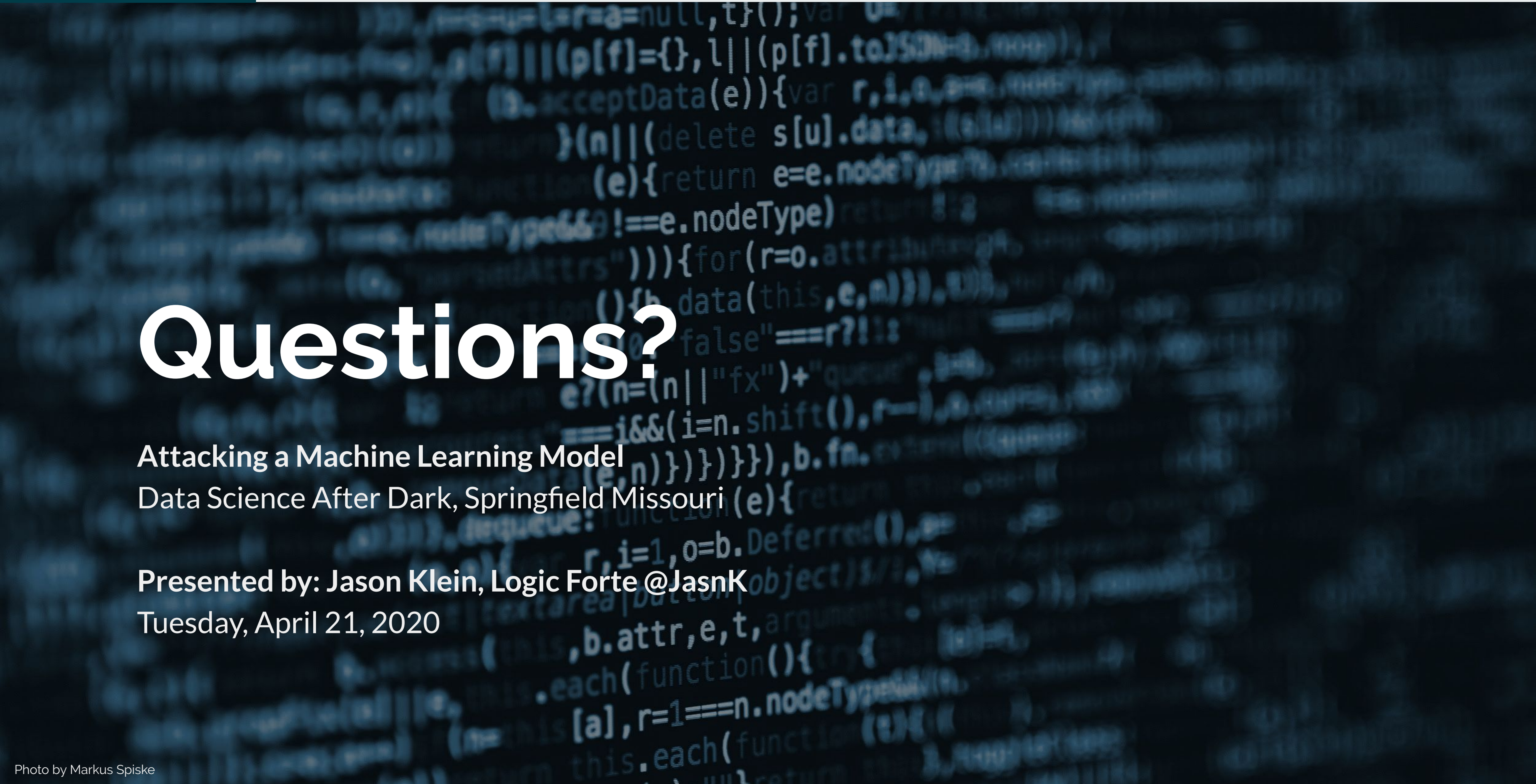
Photo by Foo Bar

Summary

If you train ANY type of model for your organization, be aware that an attacker can use similar techniques to bypass your model if they can directly access your model.

For example, an attacker could feed a fraudulent transaction into a fraud detection model and determine what transaction detail can be changed to fool the model into believing the transaction is NOT fraudulent.

Photo by Markus Spiske



Questions?

Attacking a Machine Learning Model
Data Science After Dark, Springfield Missouri

Presented by: Jason Klein, Logic Forte @JasnK
Tuesday, April 21, 2020

Photo by Markus Spiske

References

A Brief History of Computer Vision and Convolutional Neural Networks (2019)
hackernoon.com/a-brief-history-of-computer-vision-and-convolutional-neural-networks-8fe8aacc79f3

A Short History of Machine Learning -- Every Manager Should Read (2016)
forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/

PicoCTF 2018: Dog or Frog Question
2018shell2.picoctf.com:11889

PicoCTF 2018 Writeup: General Skills (Dog or Frog Solution)
tcode2k16.github.io/blog/posts/picoctf-2018-writeup/general-skills/#dog-or-frog

Photo by Markus Spiske

Thank you for attending Data Science After Dark!

Attacking a Machine Learning Model
Data Science After Dark, Springfield Missouri

Presented by: Jason Klein, Logic Forte @JasnK
Tuesday, April 21, 2020

meetup.com/Data-Science-After-Dark/events/268654605/

Photo by Pero Kalimero