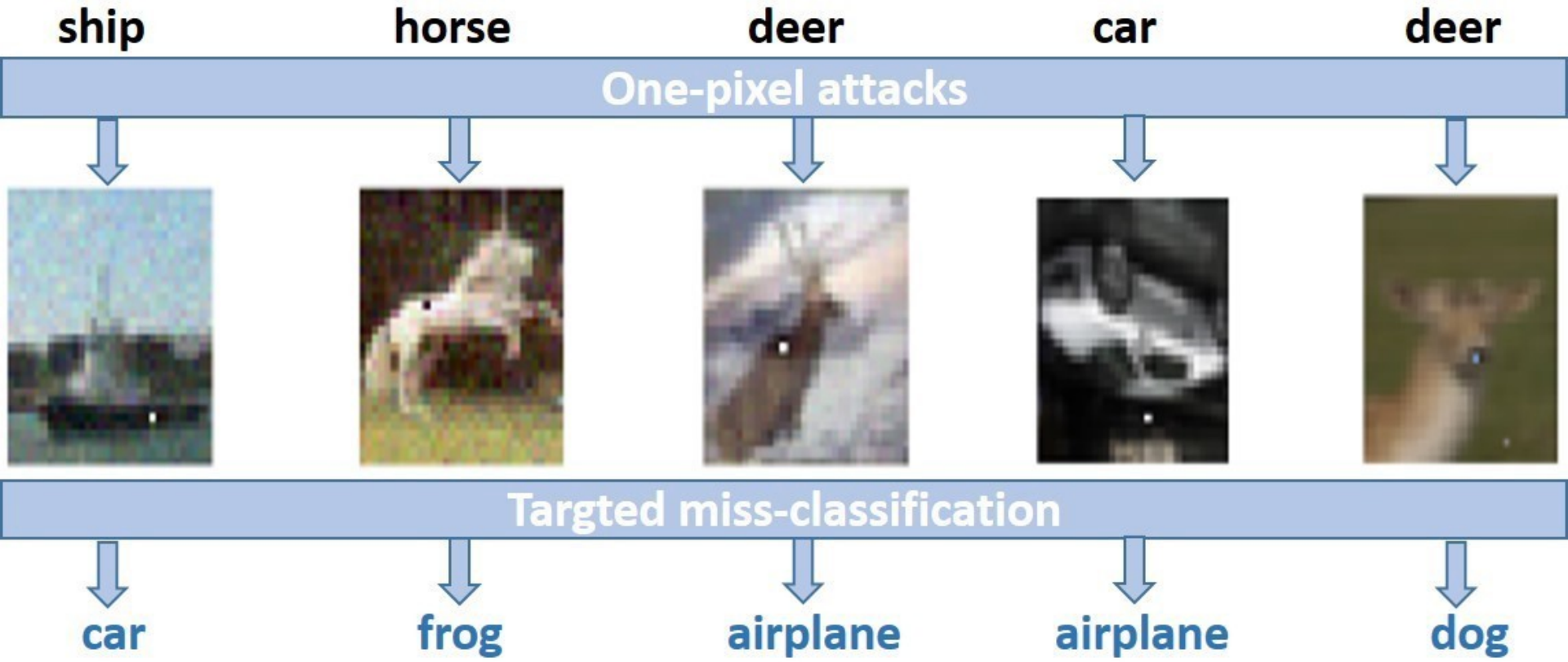
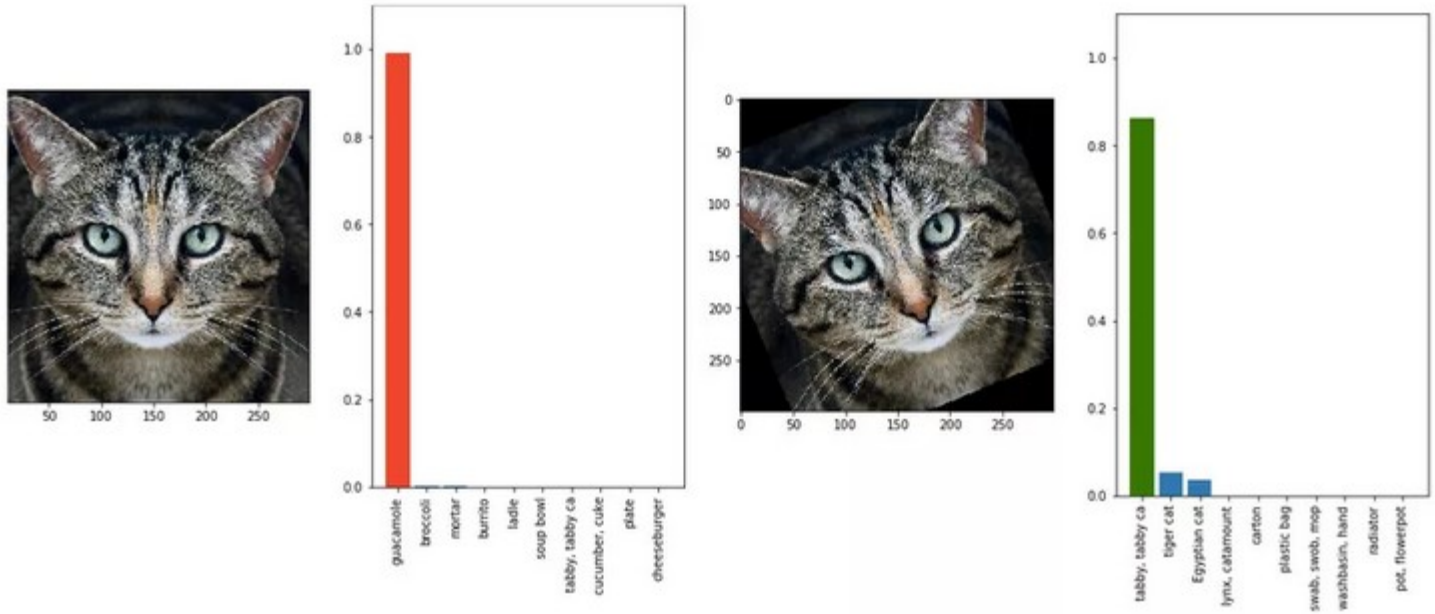
ML Challenge #3 Results

Thomas Britton
David Lawrence

Problem 3 Background



Problem 3 Preview



An example from labsix of how fragile adversarial attacks often are. The image on the left has been altered so that it's identified as guacamole. Tilting it slightly means it's identified once more as a cat.

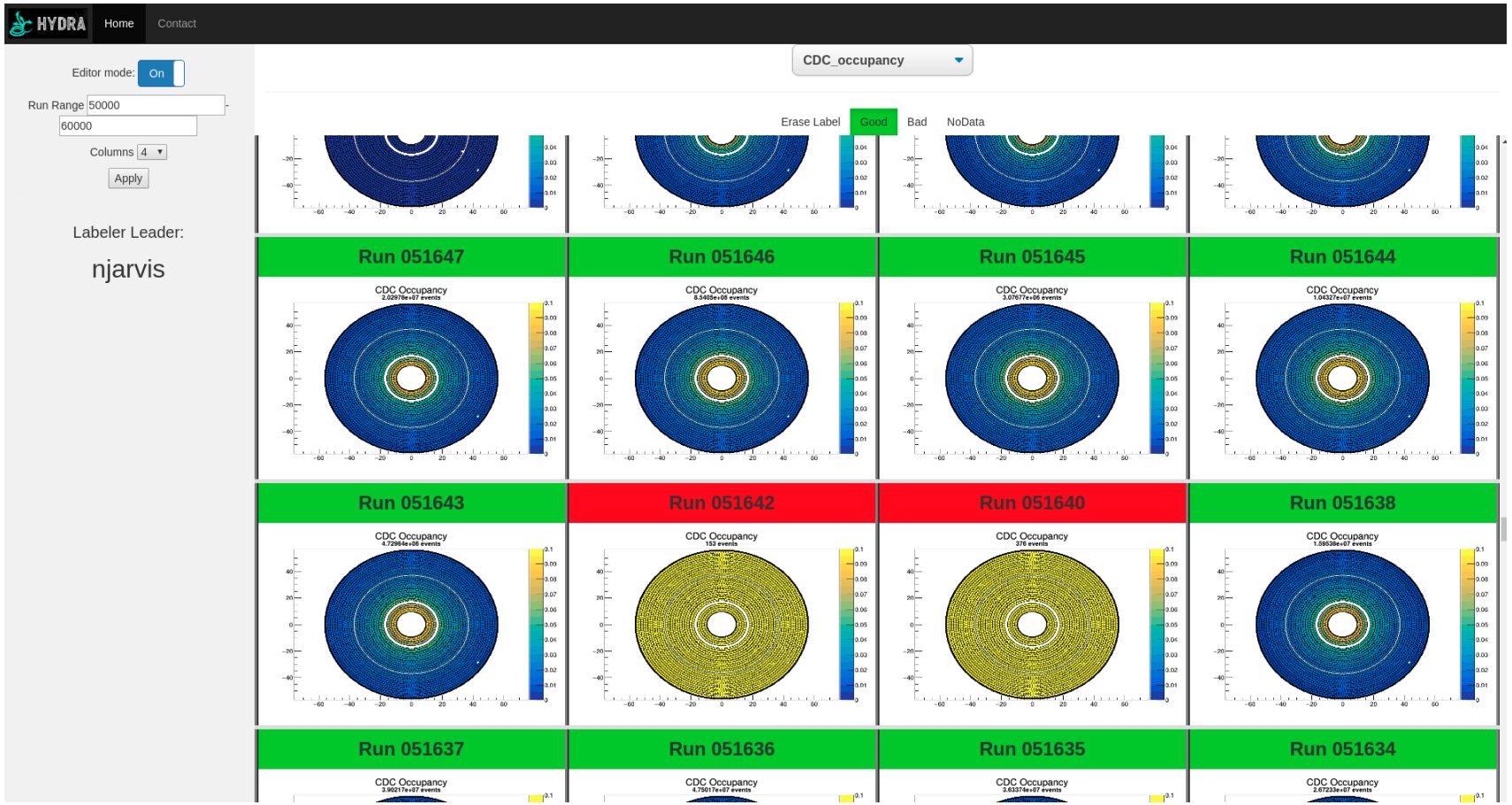

I WANT YOU
TO FOOL MY MODEL
imgflip.com

Introducing Hydra

- Hydra aims to be an extensible framework for training and managing A.I. for near real time monitoring
  - If you need it to tell a dog from cat I can have hydra do that without system modification now
- Most importantly, Hydra allows me to embrace my inner sloth:

Koboldpress.com

My classification problem

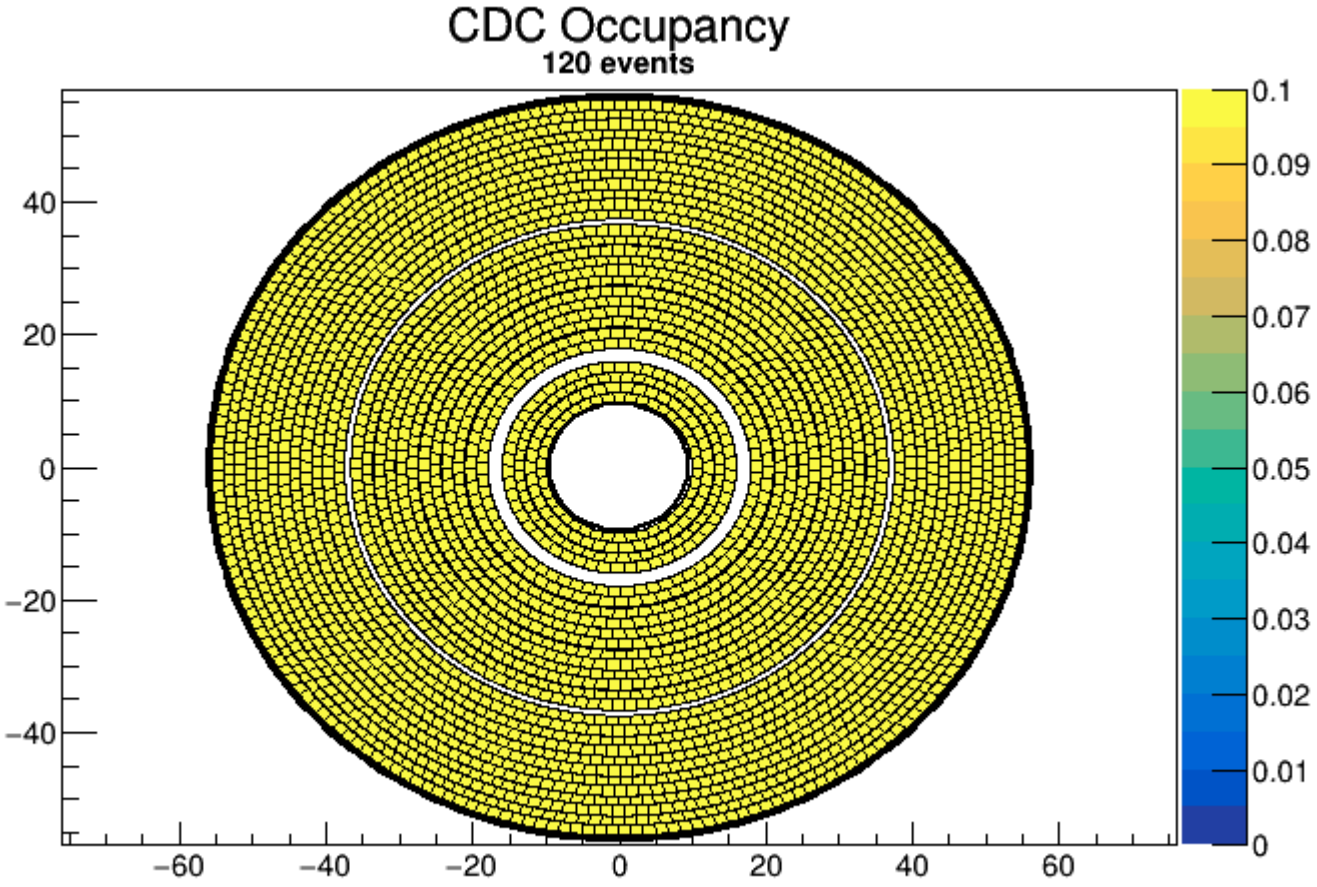


I have created a system to look at plots and classify them

The problem

- Given a model and the target image I want to see modified images that fool my model
  - The winner will have the smallest difference to the original picture while changing the model's confidence and classification
- The target image is "Bad" make the model think it is "Good"

The Target


CDC Occupancy
120 events
40
20
0
−20
−40
−60
−40
−20
0
20
40
60
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

The Materials

- A model file:
  /group/halld/www/halldweb/html/talks/ML_lunch/Nov2019/CDC_occupancy-1569856232_566866.h5
- A target image: **TARGET_IMG.png**
- A script to take in an image and classify it:
  hydra_predict.py -D TARGET_IMG.png
- A set of images for study/experimentation:
  example_images/[Good/Bad/NoData]
- An example of a target with submission:
  EXAMPLE_OF_THE_TARGET_IMG.png and
  EXAMPLE_SUBMISSION_AGAINST_EXAMPLE_TARGET.png

Due

- A valid png image to be analyzed….that’s it

- Due: **February 5th 2020** at **noon**

- Find the needed files:

  https://halldweb.jlab.org/talks/ML_lunch/Nov2019/

My Hope

- There are some adversarial examples found which would NOT fool a person but does fool the model.
  - I want a guacamole-CDC….
- Some helpful hints:
  - Start by making sure you can run the script on the target image
  - DO NOT run the script with the model sitting on /work/ (file locking issue)
  - DO NOT just swap the labels….I'll be running on an unedited copy of the script ;)

Due

- A valid png image to be analyzed….that’s it

- Due: **February 5th 2020** at **noon**

- Find the needed files:

  https://halldweb.jlab.org/talks/ML_lunch/Nov2019/