# Developing Provably Robust Explanation Methods for Image Classifiers

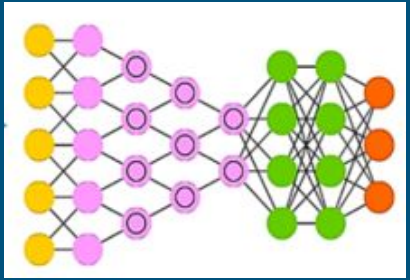
André Shannon

Mentored by

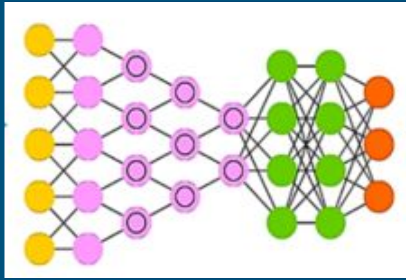
Dr. Douglas Szajda

# Machine Learning Systems

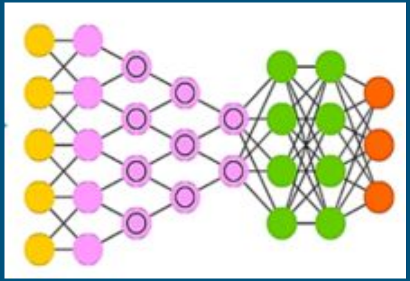
- Train on LOTS of data
- Want to generalize from training data to new instances



# Image Classifier



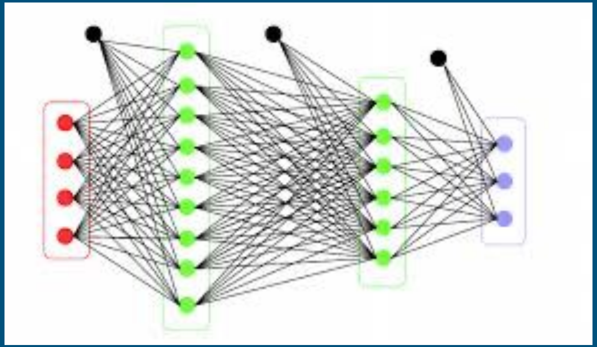
Dog



Human

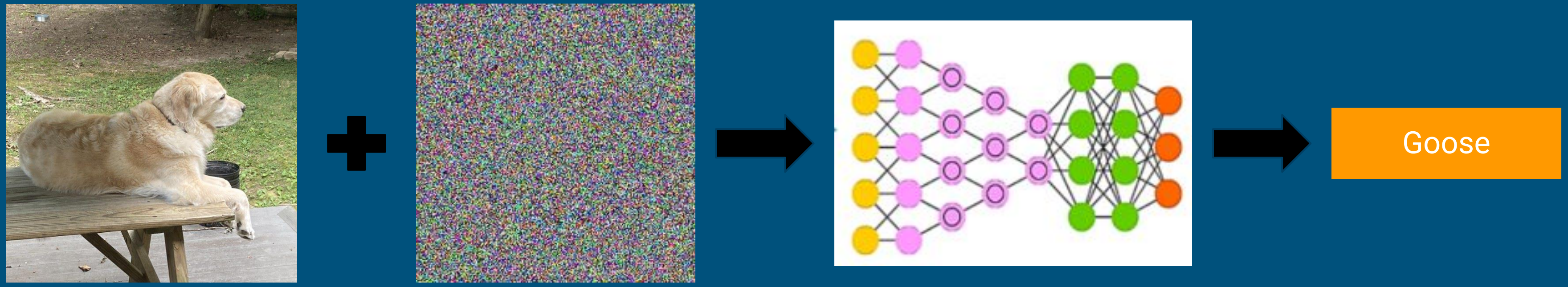
# Training (Supervised)

- Every training sample has a corresponding label
- Start with random parameters
- Optimize for every training sample



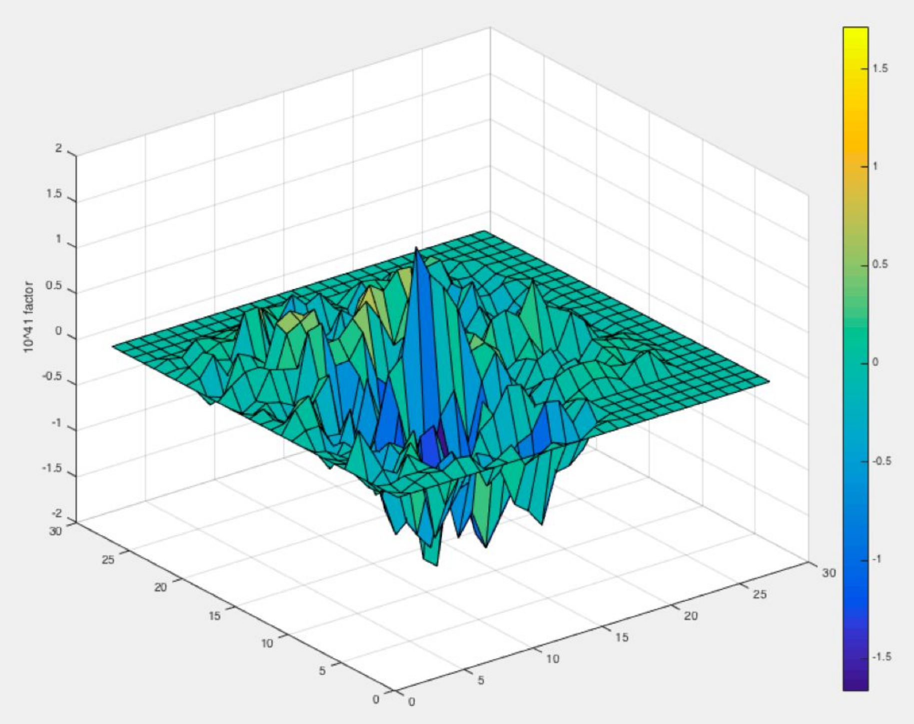
# Adversarial Samples

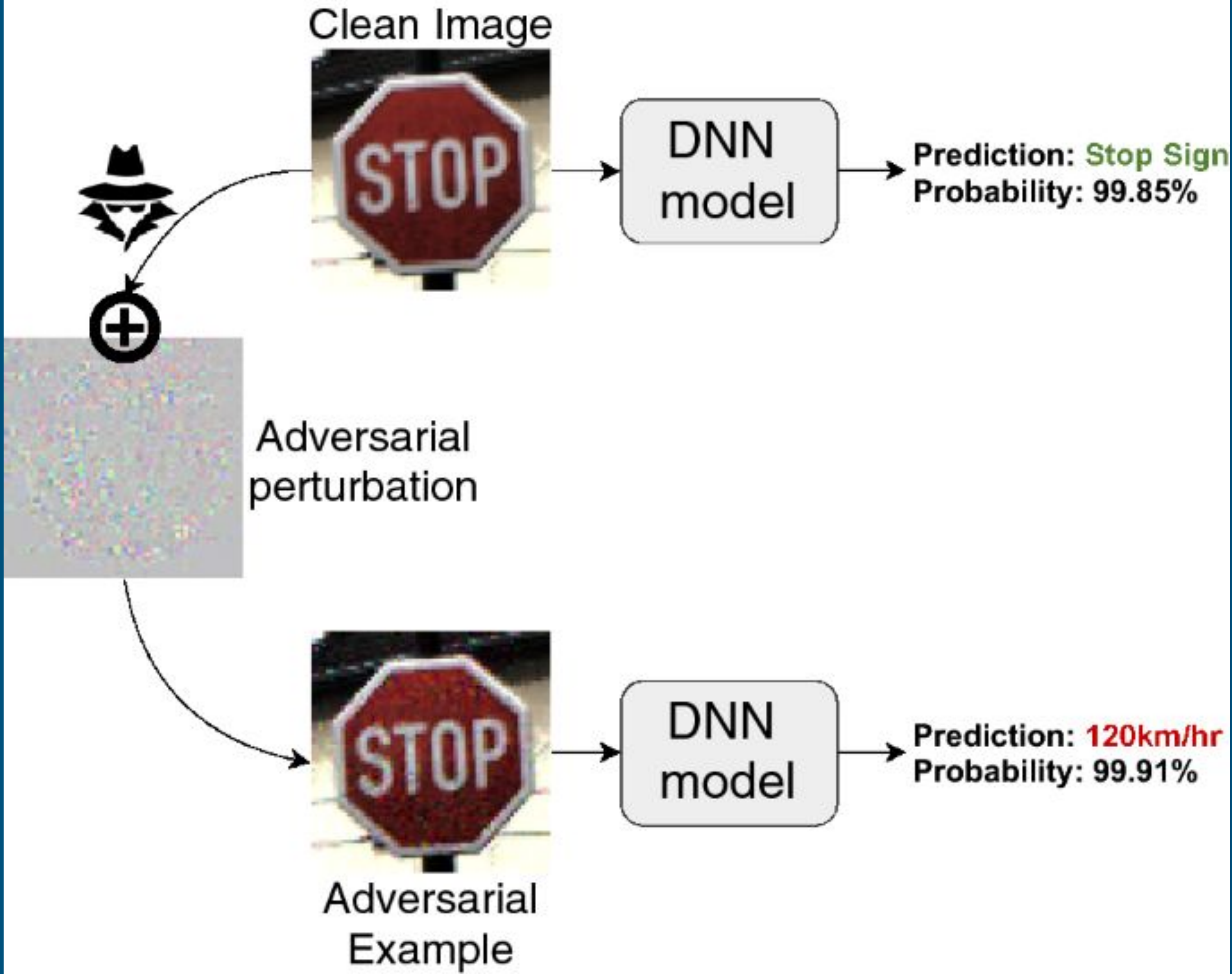
- Want to fool the classifier and not the human
- Untargeted or Targeted



# Adversarial Samples (cont.)

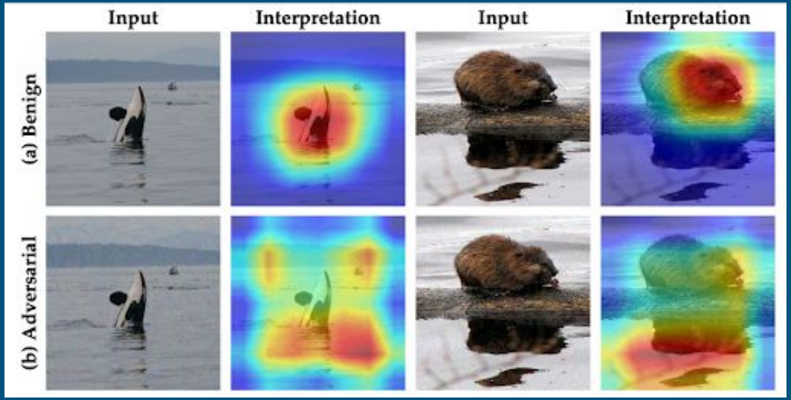
- Projected Gradient Descent
  - Look of gradient of output with respect to the inputs
  - Tweak the pixels corresponding to big gradients




Clean Image
DNN model
Prediction: Stop Sign
Probability: 99.85%
Adversarial perturbation
DNN model
Prediction: 120km/hr
Probability: 99.91%
Adversarial Example

# Explanation Methods

- ❖ Models can have billions of parameters
- ❖ Want to know model's "reasoning"
  - ➢ Also want to detect adversarial samples



# LIME (Local Interpretable Model-agnostic Explanations)

- ❖ Seeks to explain the classification of specific inputs
- ❖ Creates a linear approximation of the model around the input
- ❖ Create dataset
  - ➢ Sample around original input
  - ➢ Classify each sample
- ❖ Create linear model (explanation) based on dataset

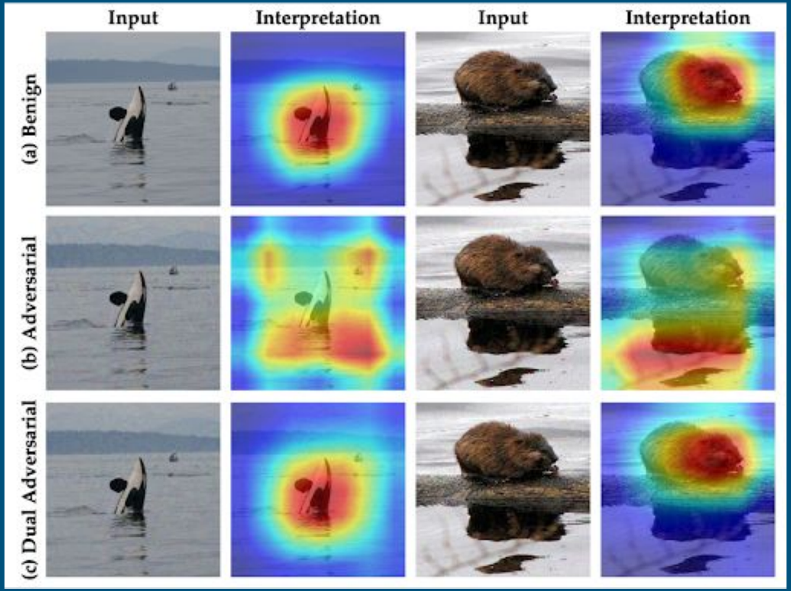
# Trigger Warning! Equation!

- Classifier Model: $f$
- Input image: $\boldsymbol{x}$
  - Consists of feature values $(x_1, x_2, \ldots, x_n)$

$$f(\mathbf{x}) \approx \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

# More Adversarial Samples

- Fool both the classifier and the explanation method



# Certified Adversarial Robustness via Randomized Smoothing

- Certified radii around points for which all points in “ball” around that point are classified the same as the certified point
- Created “smoothed” classifier
  - To classify input, gather samples close to input, classify them, and return the label that shows up the most
  - Calculate radius using probability of being top label
    - Bigger probability -> bigger radii

# Adding Robustness to Explanations

- Create “smoothed” explanation method

# Current Challenge

- How can we say two explanations are the same?

$$\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

  - Rank Coefficients?

$$\beta_3, \beta_{14}, \ldots, \beta_8, \beta_{50}$$

  - Look at just first 10?
  - Edit Distance?

# Questions?

# Thanks Dr. Szajda!

# Credits

- Thanks Grant for the picture of Frosty!


- J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. CoRR, abs/1902.02918, 2019.
- M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.
- F. C. M. Rodrigues, M. Espadoto, R. Hirata, and A. C. Telea. Constructing and visualizing high-quality classifier decision boundary maps. Information, 10(9):280, Sep 2019.