

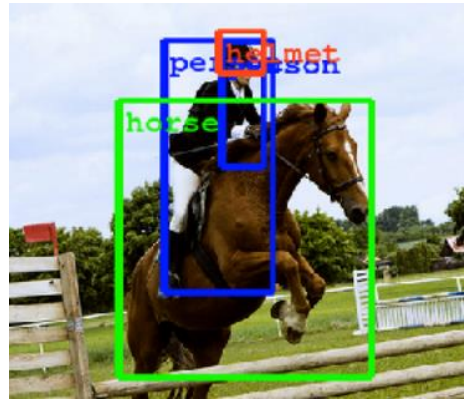
# Exploring New Attack Space on Adversarial Deep Learning

**Chang Liu**, University of California, Berkeley  
FORCE PI Meeting, Jan-25, 2017

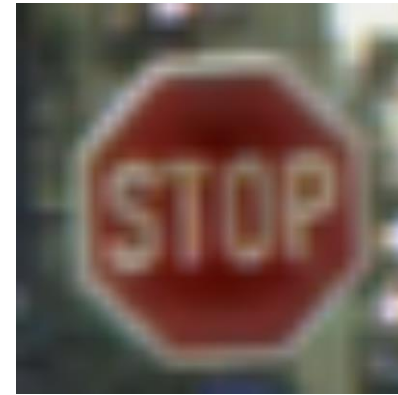


# Deep neural networks have matched human performance at...

- Recognizing objects
- Recognizing faces
- Reading addresses
- Classifying images
- and other tasks



# Adversarial examples in the cyber-physical world

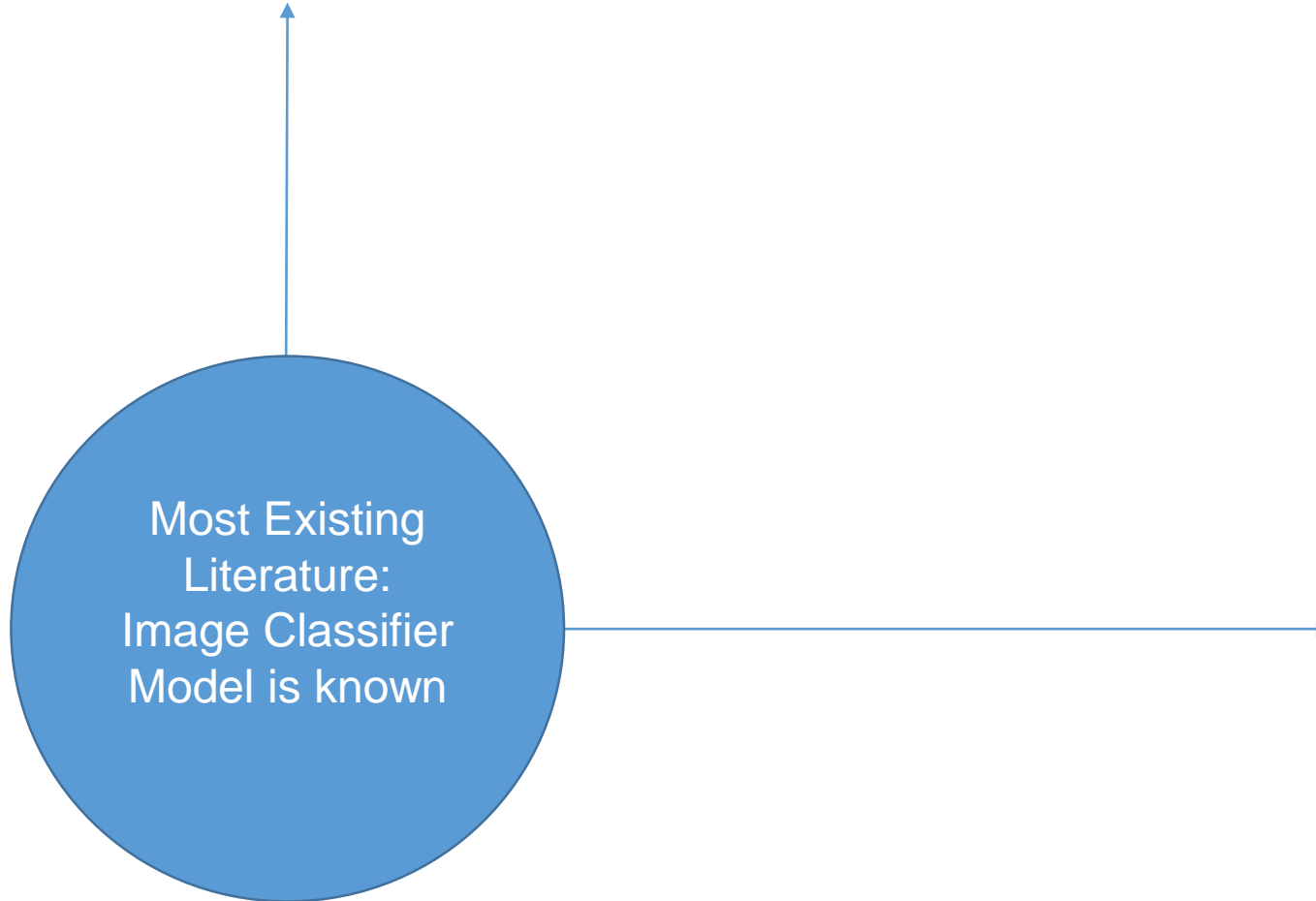


Stop Sign



Yield Sign

Weaker threat model:  
The machine learning  
model is a black-box

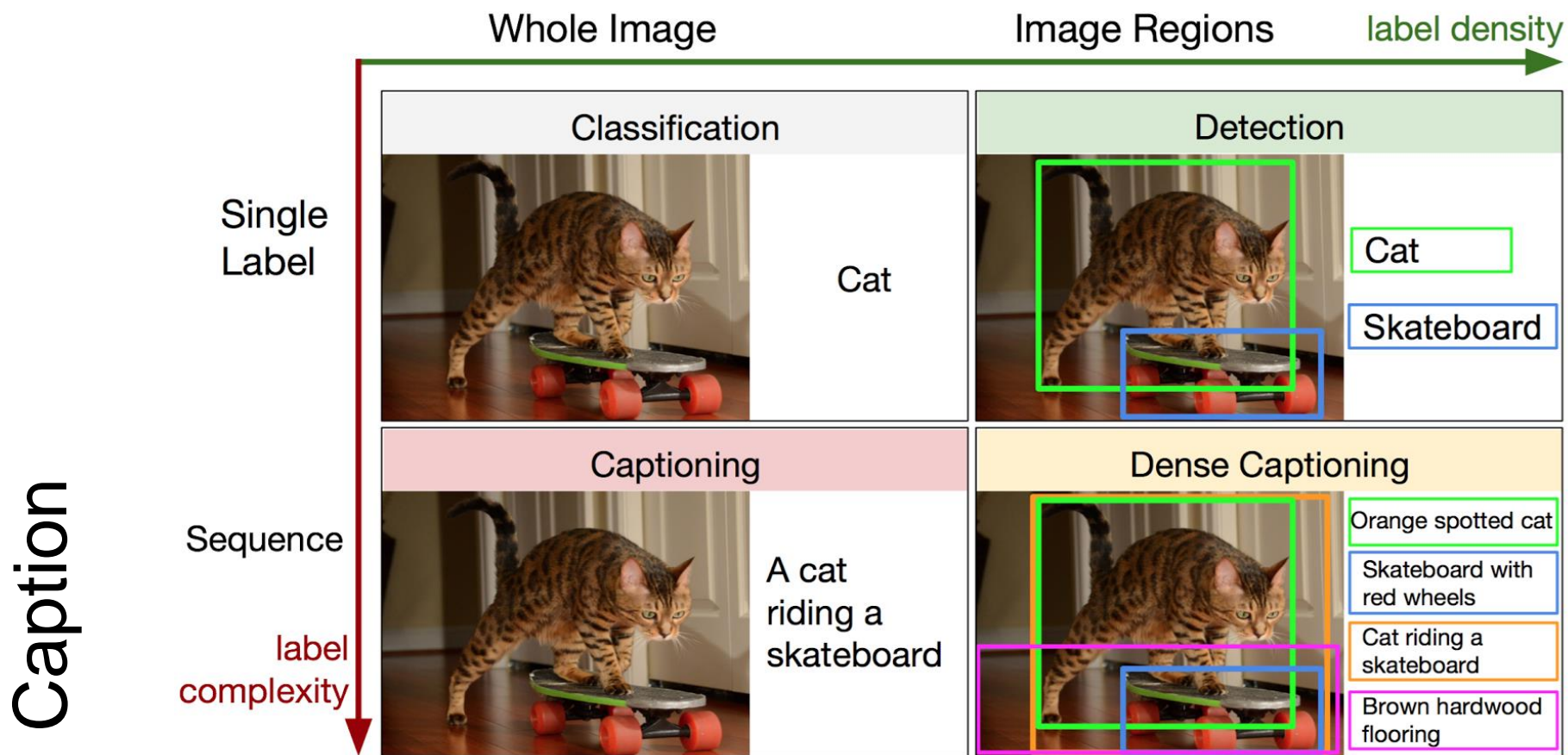


Most Existing  
Literature:  
Image Classifier  
Model is known

More models:  
Object detection  
Captioning  
Etc.

# Go beyond image classifier: DenseCap

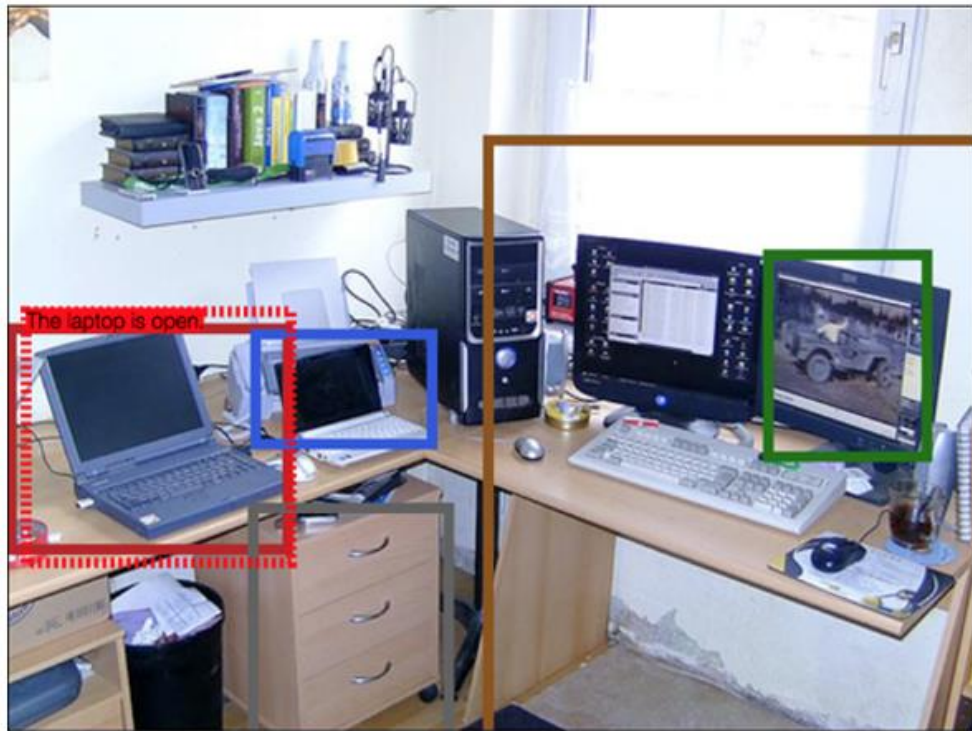
## Object Detection



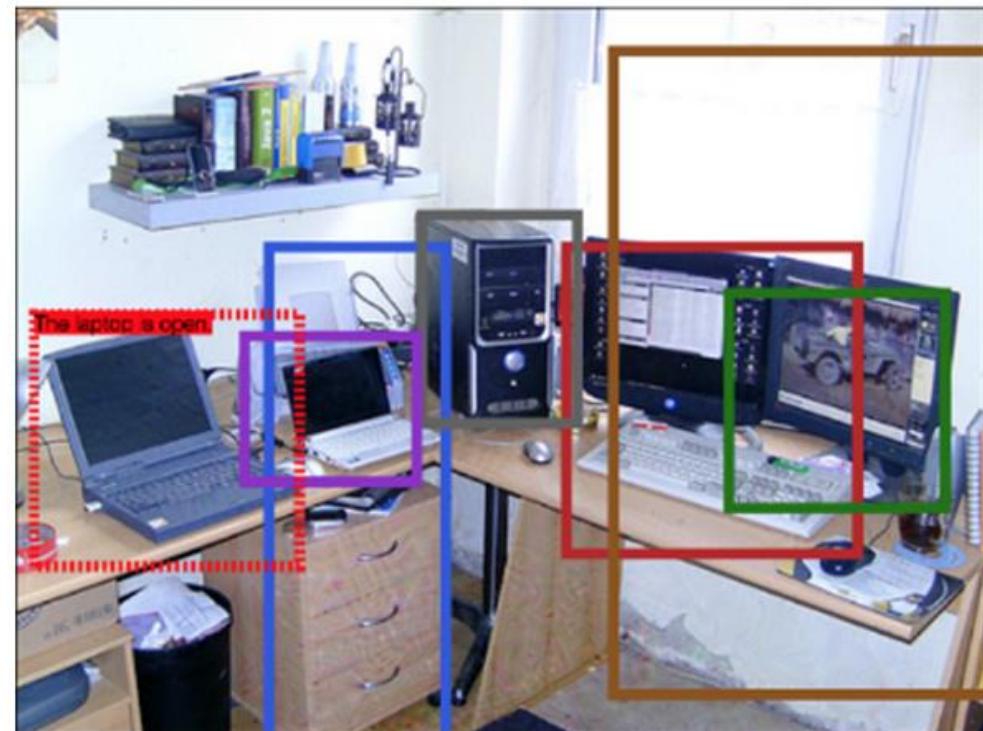


# Adversarial Example for Object Detection

Original Detected Objects



Adversarial Image



# Adversarial Examples for Captioning

Original



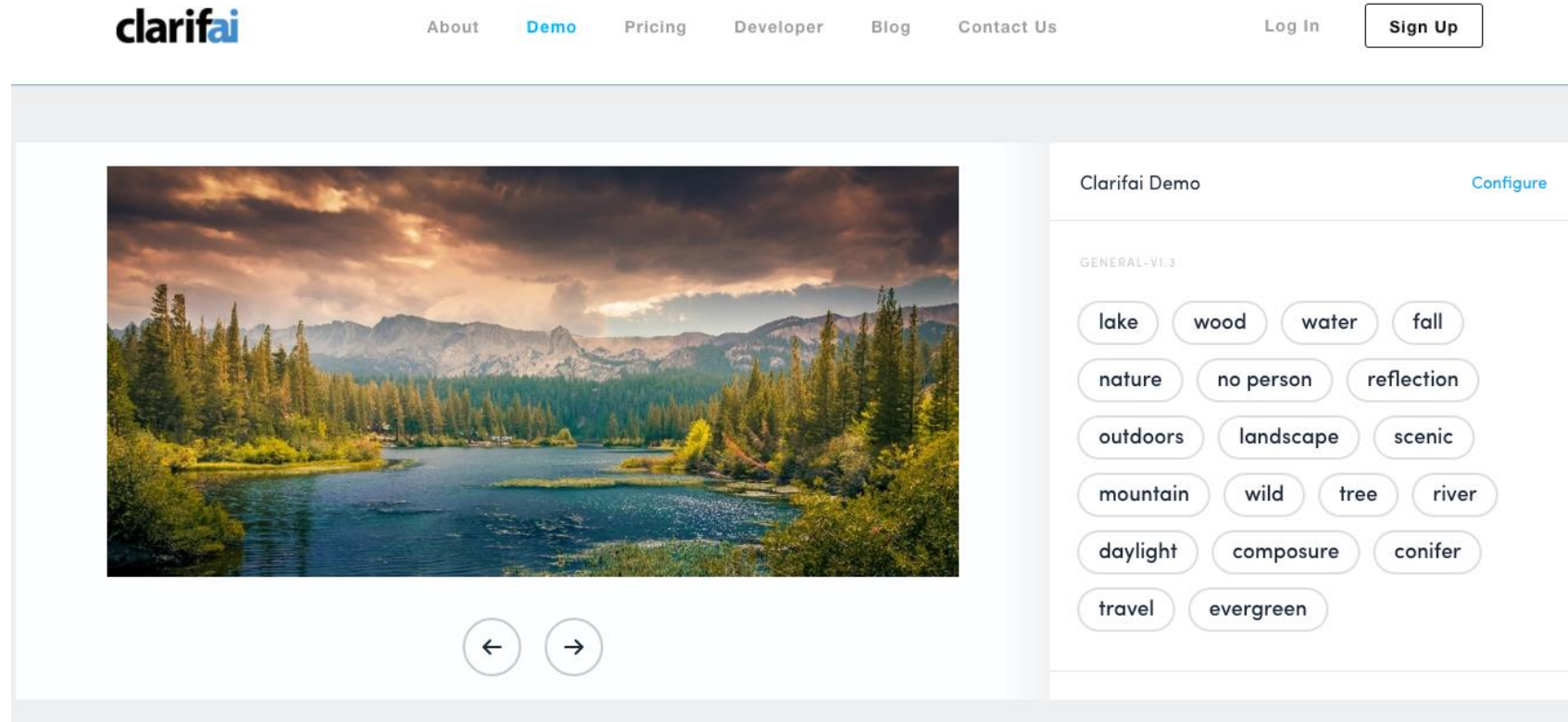
a towel hanging on a rack  
a trash can on the floor  
a mirror on the wall  
a white bathtub  
white cabinets under sink

Adversarial Image



a white and red cup  
front window of a bus  
a dog in a window  
a large mirror on the wall  
a sign on the side of the bus

# Adversarial images for a black-box system



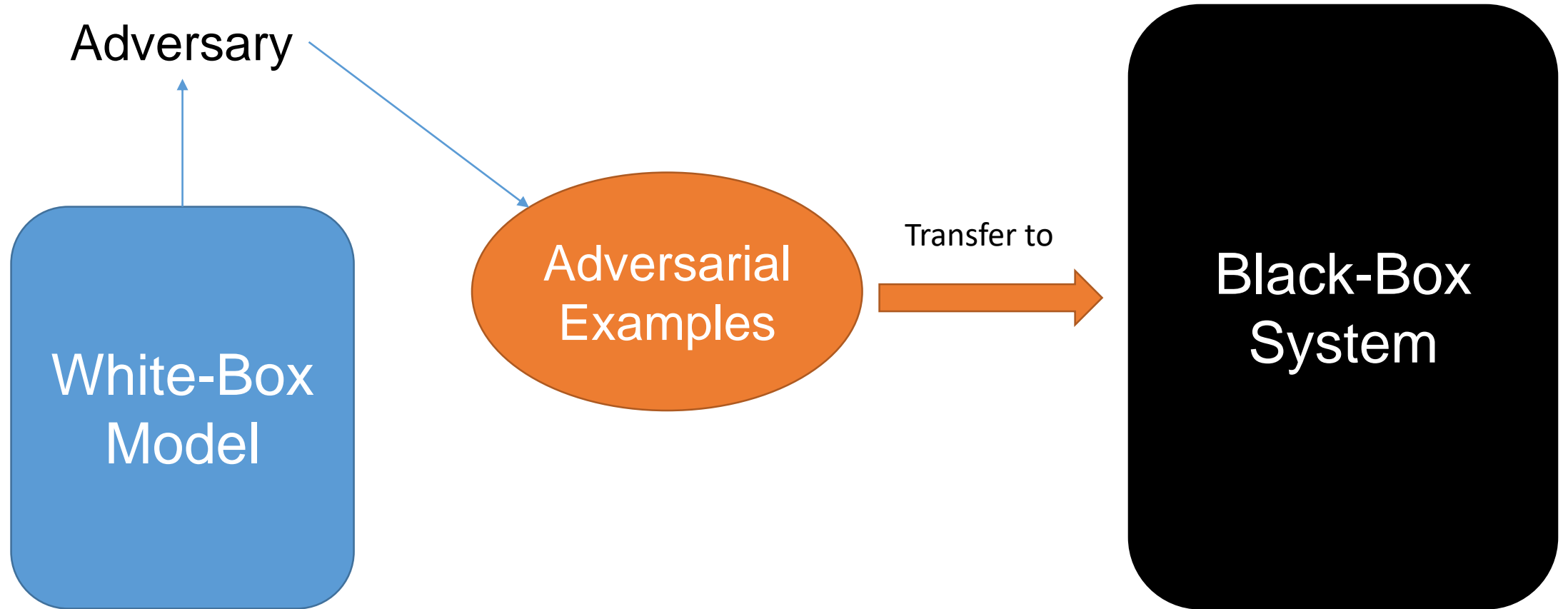
Unknown:  
*Model*  
*Training data*  
*Label set*

Yanpei Liu, Xinyun Chen, **Chang Liu**, Dawn Song. Delving into Transferable Adversarial Examples and Black-box Attacks. Submitted to ICLR 2017.

[arXiv:1611.02770](https://arxiv.org/abs/1611.02770)




# Black-box attack based on transferability



# Clarifai.com

hip, rose hip,

☰ clarifai



Clarifai Demo [Configure](#)

GENERAL-V1.3

no person nature wildlife little

fall fruit food outdoors

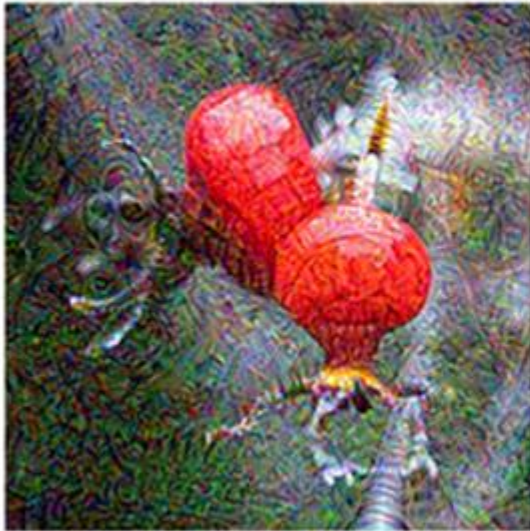


stupa



# Clarifai.com

- Ground truth: **hip, rose hip, rosehip** Target label: **stupa**



GENERAL-V1.3

decoration

art

gold

temple

design

desktop

pattern

religion

traditional

ancient

color

bright

culture

celebration

illustration

old

symbol

Buddha

artistic

NSFW-V1.0

sfw

# Non-targeted adversarial images misclassified by Clarifai.com

Original



case container box  
briefcase luggage

Single-network  
Optimization Approach



case box design old  
bag retro vintage

Ensemble Approach



family furniture design  
sofa inside chair

# Conclusion

- Adversarial deep learning is important for cyber-physical systems
- It is easy to find adversarial examples for deep neural networks for many application domains
- Adversarial examples can be found with even only black-box access to the model.