# Don’t Red Team AI like a Chump

Ariel Herbert-Voss
@adversariel

# Intro

- Who am I?
- What is this talk?

# A story about hacking an AI system IRL



+



+

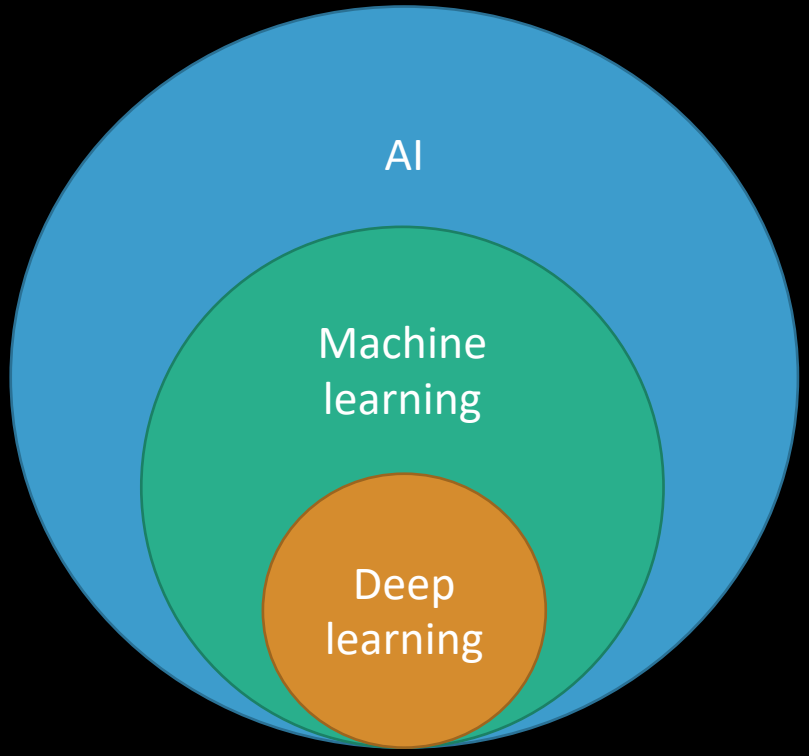


# Key points in this talk (in case you zone out)

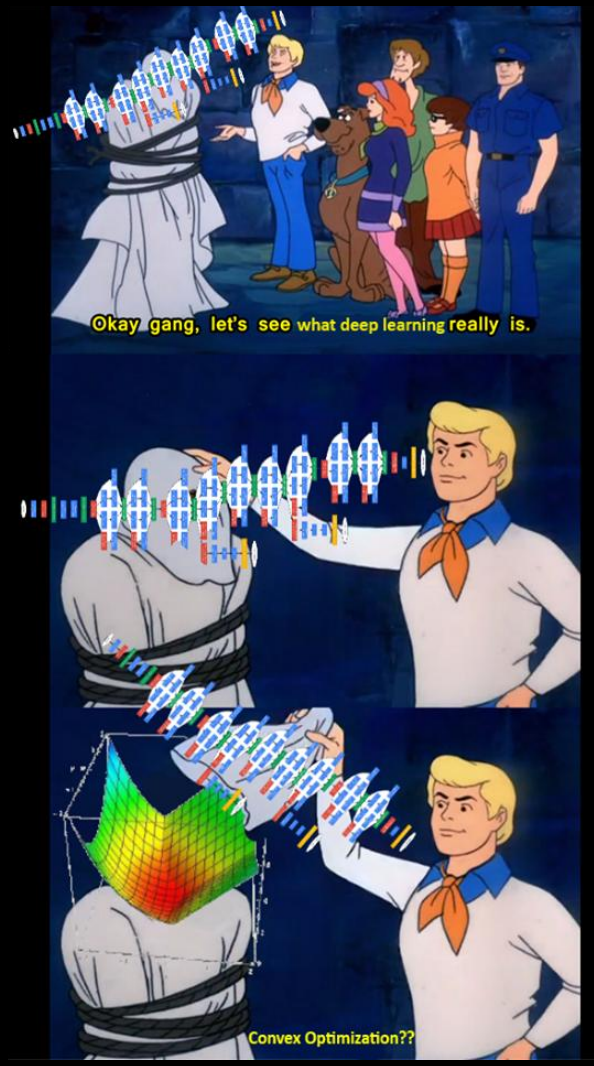
- **Focus on the threat model**
- Go for the ~~jugular~~ feature extraction process
- Think about the data supply chain

# What is an AI system?

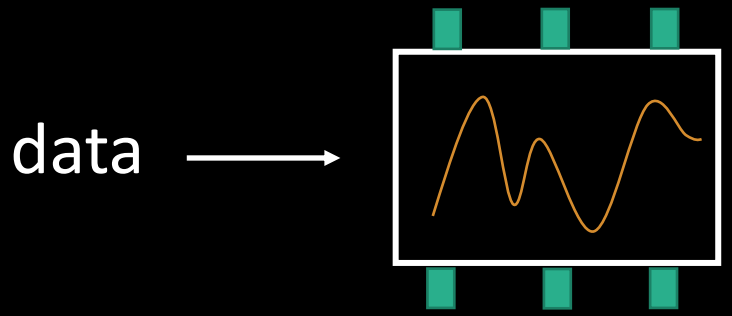
- AI is a class of algorithms that we use to extract actionable information from data
- AI is not new and the hype is real
- In this talk, AI == ML



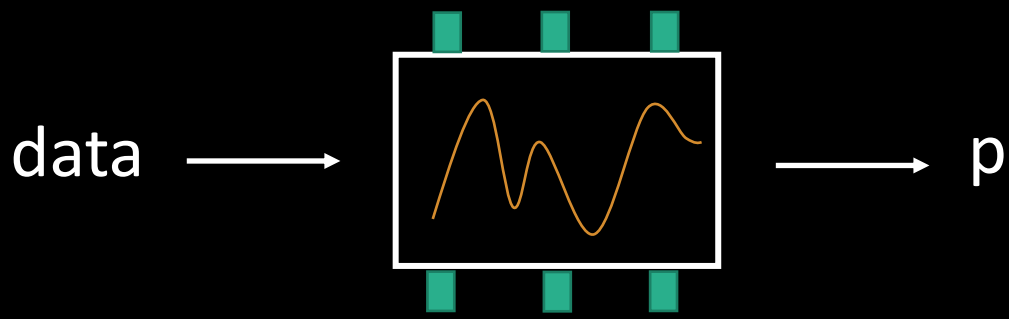
# What is an AI system?



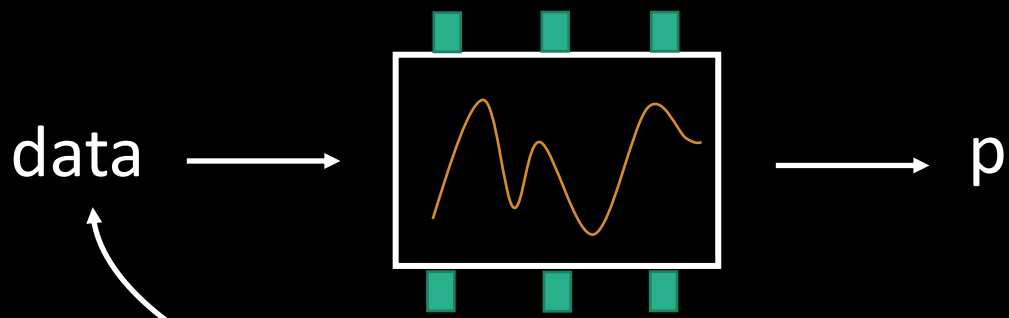
# What is an AI system?



# What is an AI system?



# What is an AI system?



# Parts of an AI system we can exploit

- Data
  - Training/testing sets
  - Deployed environment data
- Model
  - Algorithm
  - Parameters

# Parts of an AI system we can exploit

- Data
  - Training/testing sets
  - Deployed environment data
- Model
  - Algorithm
  - Parameters

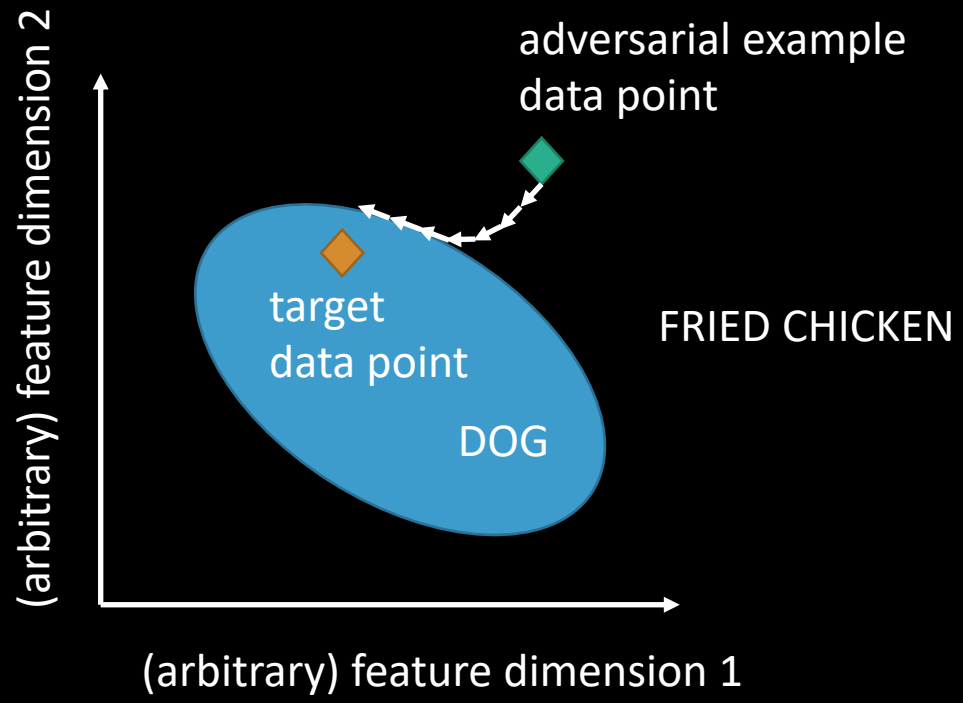


# Parts of an AI system we can exploit

- Data
  - Training/testing sets
  - Deployed environment data
- Model
  - Algorithm
  - Parameters



# Adversarial examples

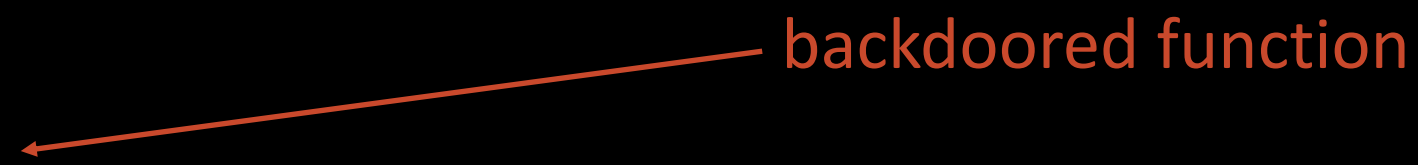





[Biggio et al, 2018]

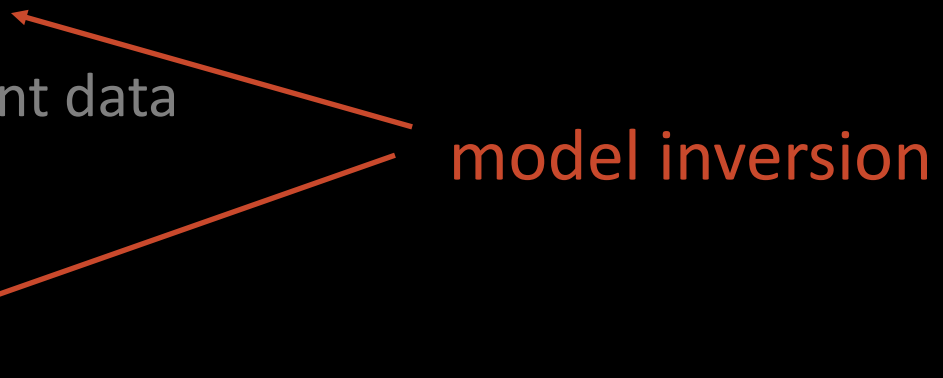
# Parts of an AI system we can exploit

- Data
  - Training/testing sets
  - Deployed environment data
- Model
  - Algorithm
  - Parameters



# Parts of an AI system we can exploit

- Data
  - Training/testing sets
  - Deployed environment data
- Model
  - Algorithm
  - Parameters



# Model inversion



Image of Bill:
*exists in training data set*



Recovered image

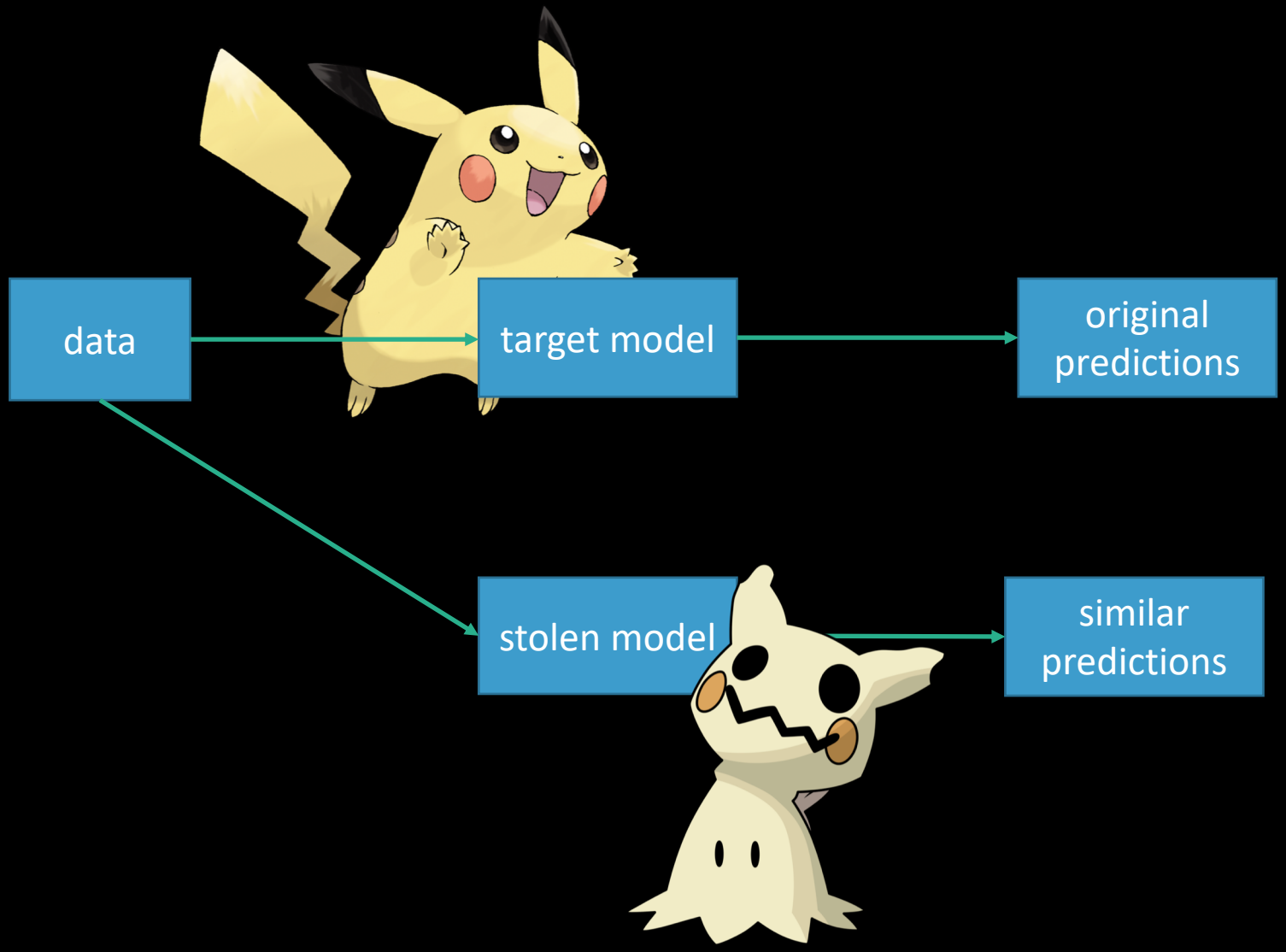
[Fredrikson et al, 2015]

# Parts of an AI system we can exploit

- Data
  - Training/testing sets
  - Deployed environment data
- Model
  - Algorithm
  - Parameters



# Model theft




[Tramèr et al, 2016]

# How to design an attack for an AI system

1. What model are you attacking?
2. Where do the data come from?
3. Where do the predictions go?

# Fooling AI-powered video surveillance

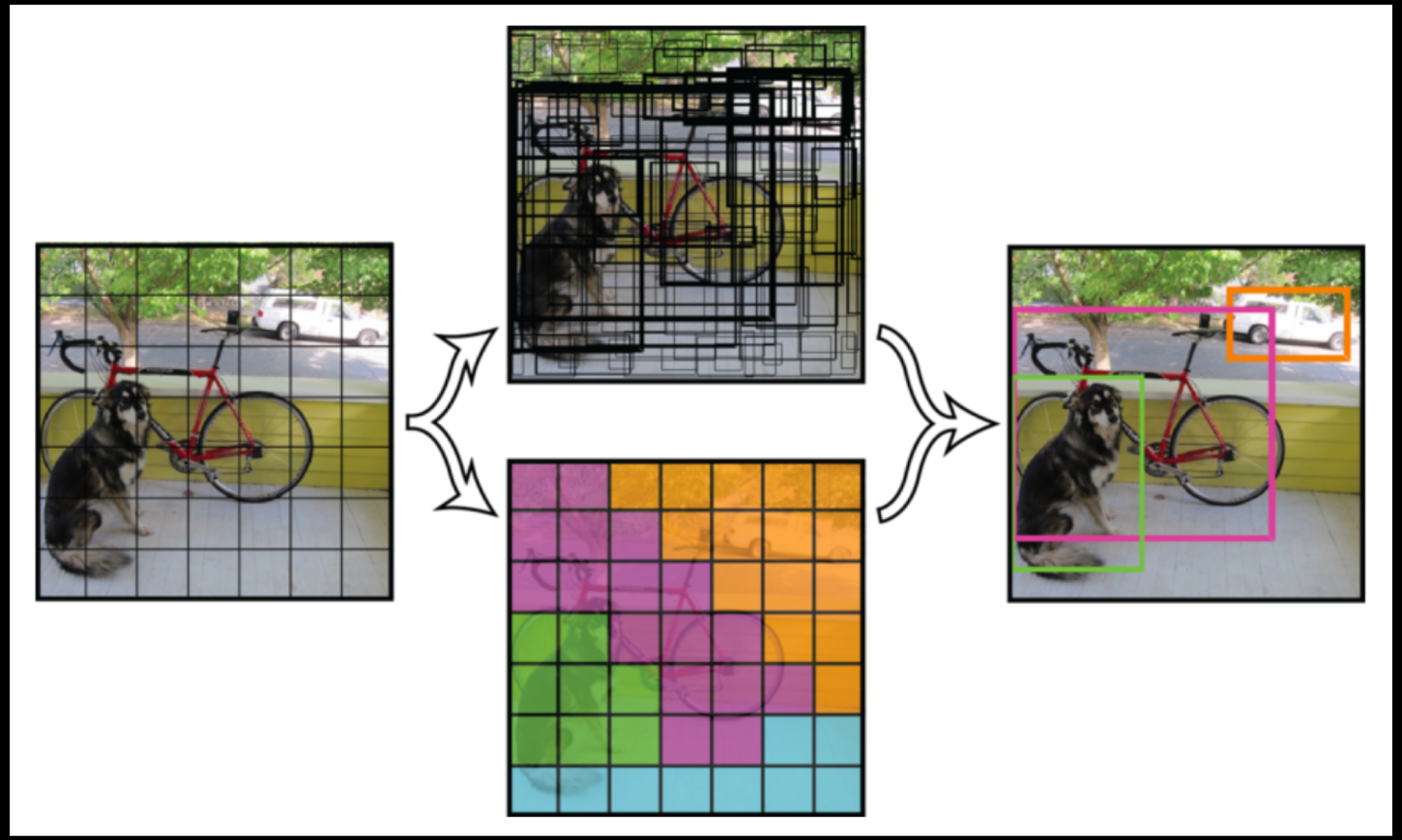
- Common system components
  - Detection: on-premises algorithm that checks each still image to see if there is a face for it to forward to the recognition system
  - Recognition: off-site look-up of flagged images using a data base of known faces



# Fooling AI-powered video surveillance

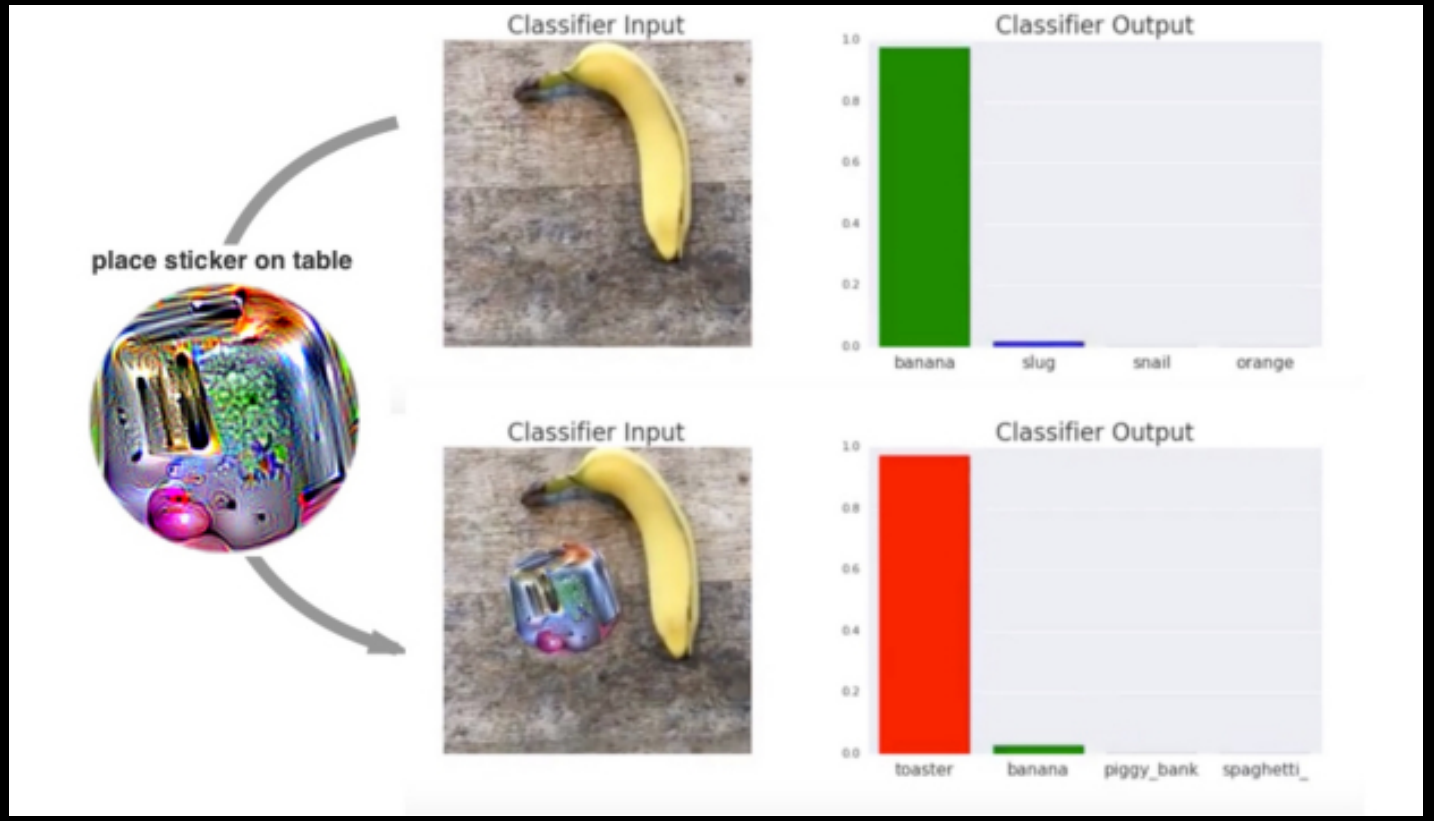
1. What model are you attacking? YOLO v2
2. Where do the data come from? video frames
3. Where do the predictions go? stored as set of flagged frames

# Fooling AI-powered video surveillance



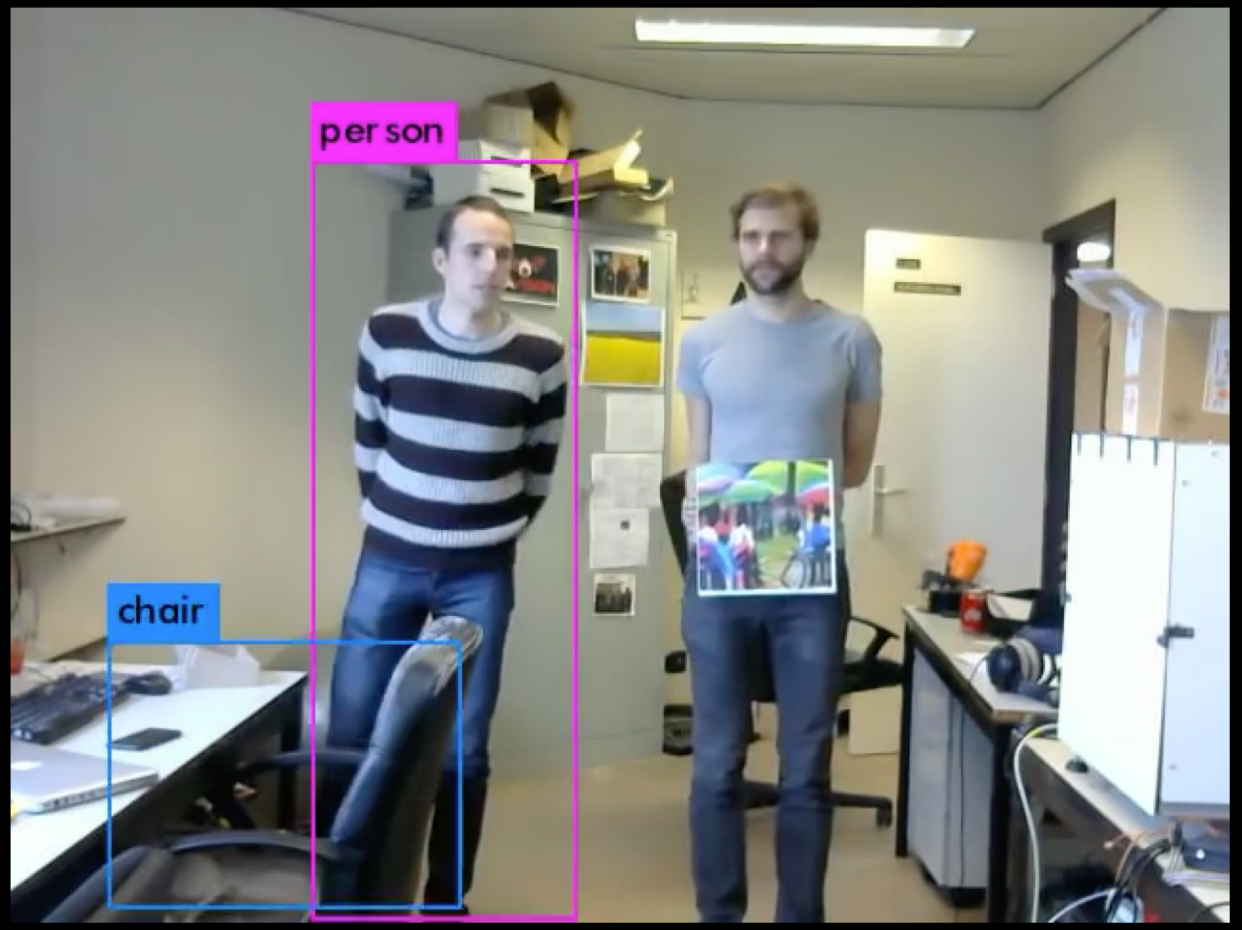
[Thys et al, 2019]

# Fooling AI-powered video surveillance




[Brown et al, 2017]

# Fooling AI-powered video surveillance



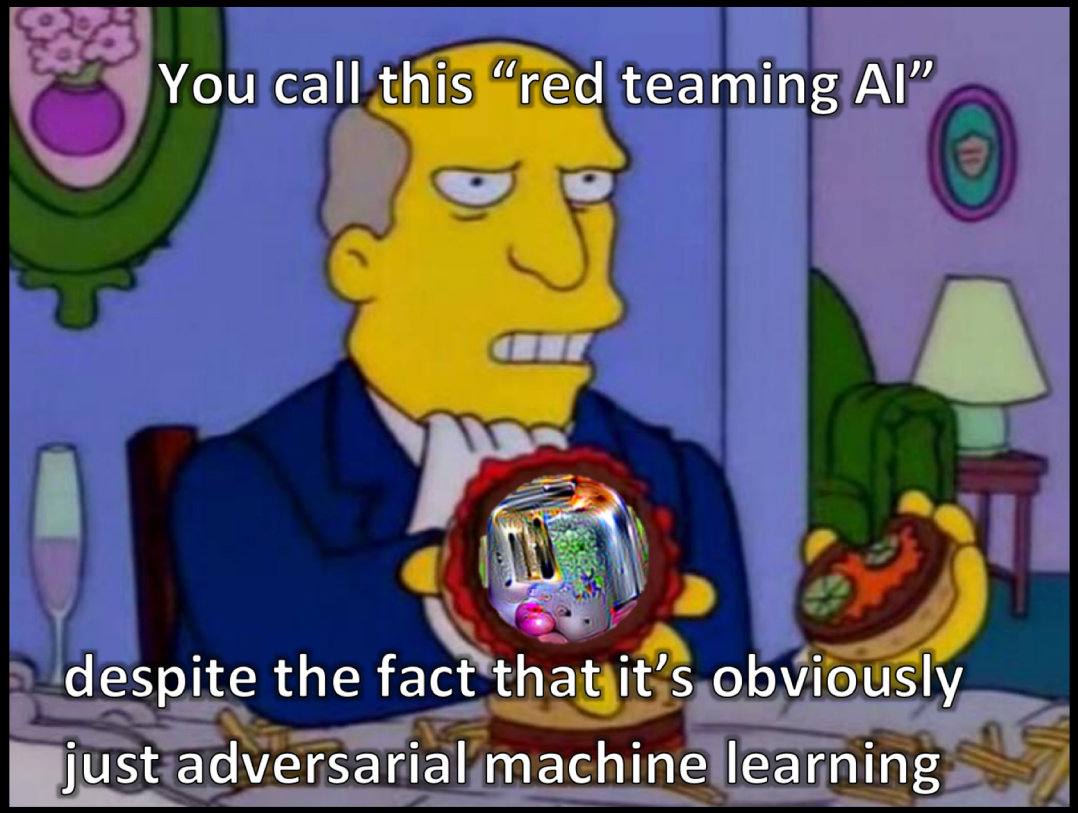

[Thys et al, 2019]

# AI red teaming is out of whack

- Red teaming AI is often conflated with the academic discipline of adversarial machine learning

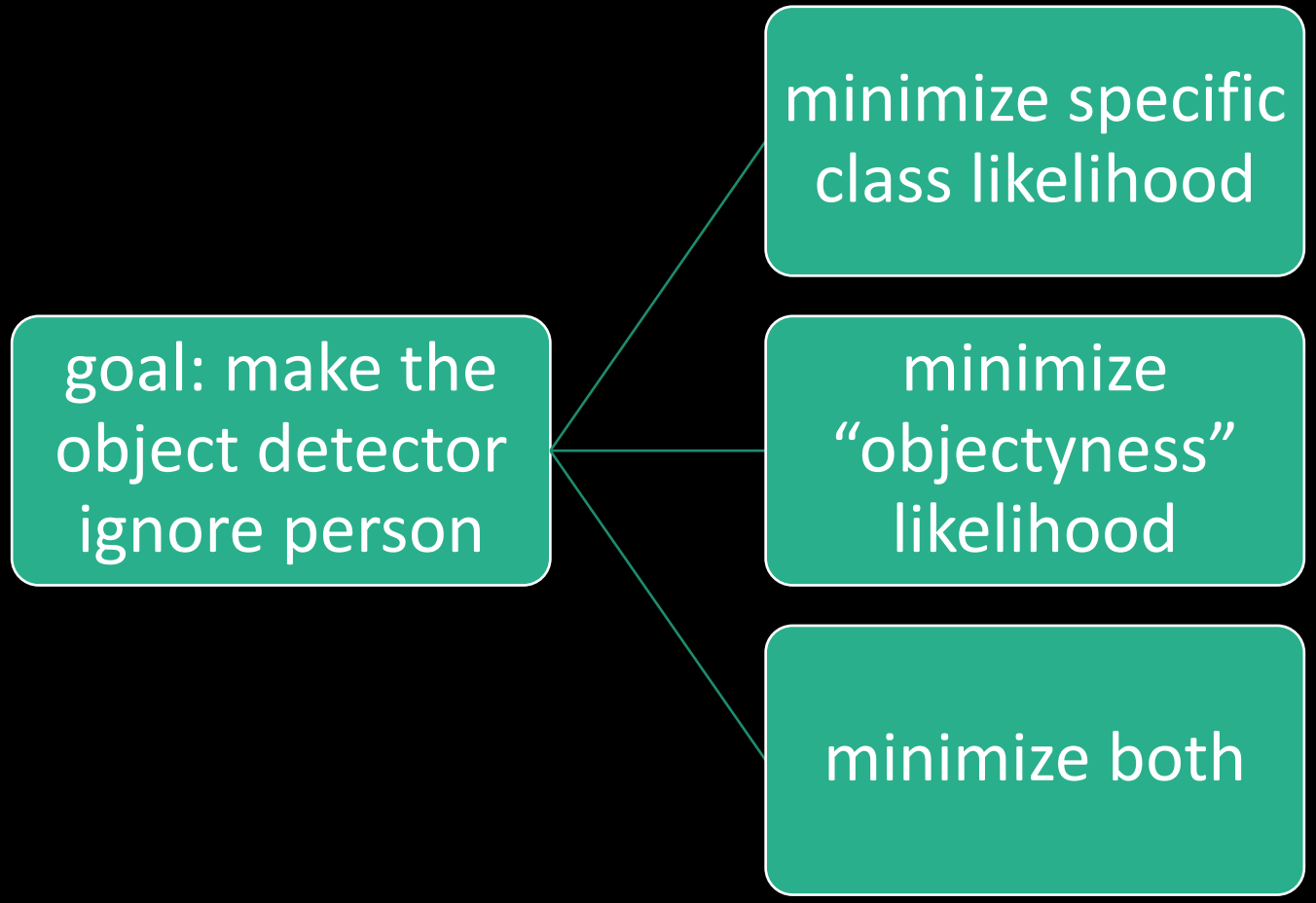
Cool ways to attack an AI model with math
== adversarial ML

Evaluate the security of an AI system
== red teaming AI



# Adversarial ML attack tree

Fooling AI-powered video surveillance



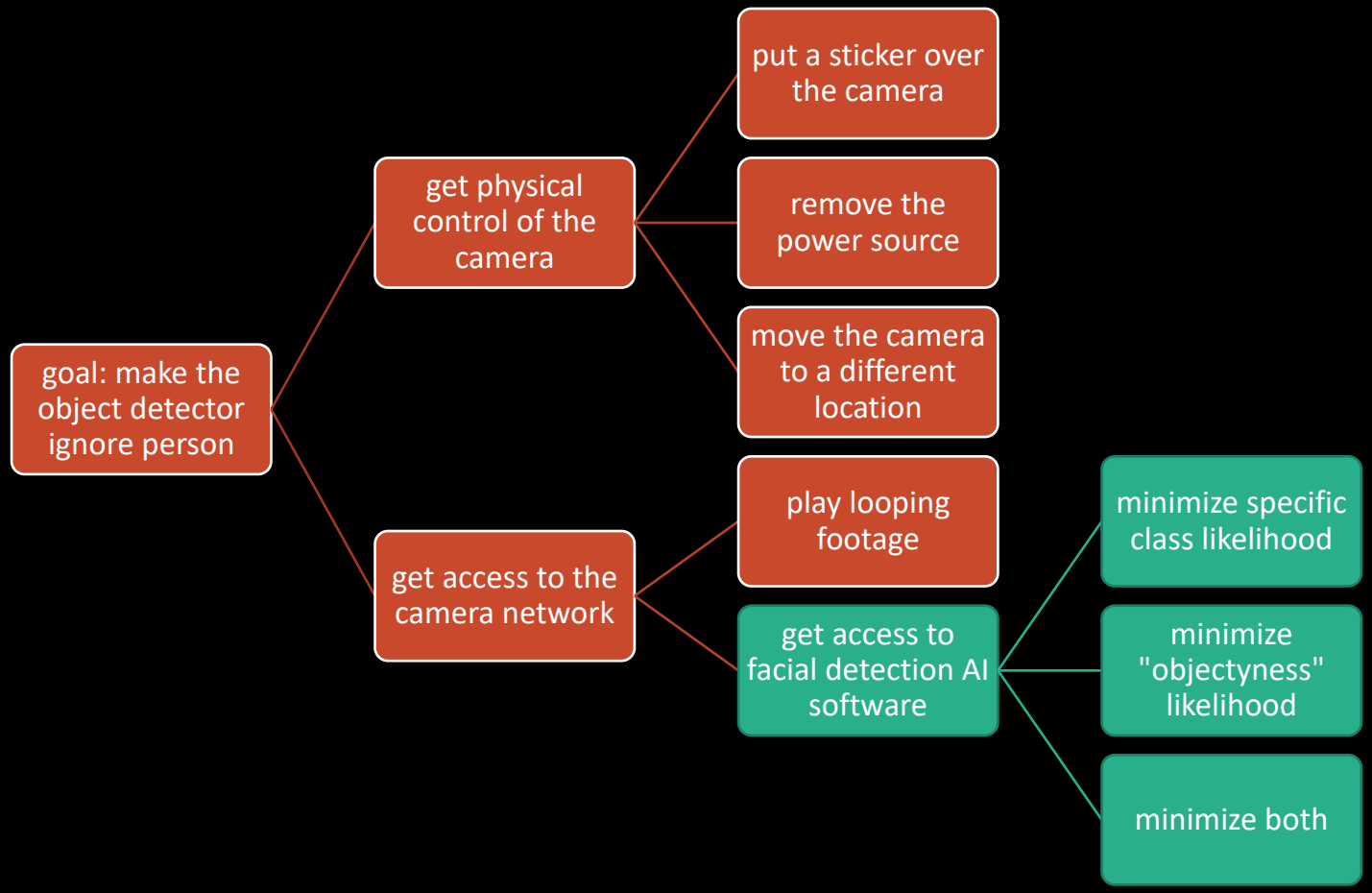
# AI red team attack tree

## Fooling AI-powered video surveillance



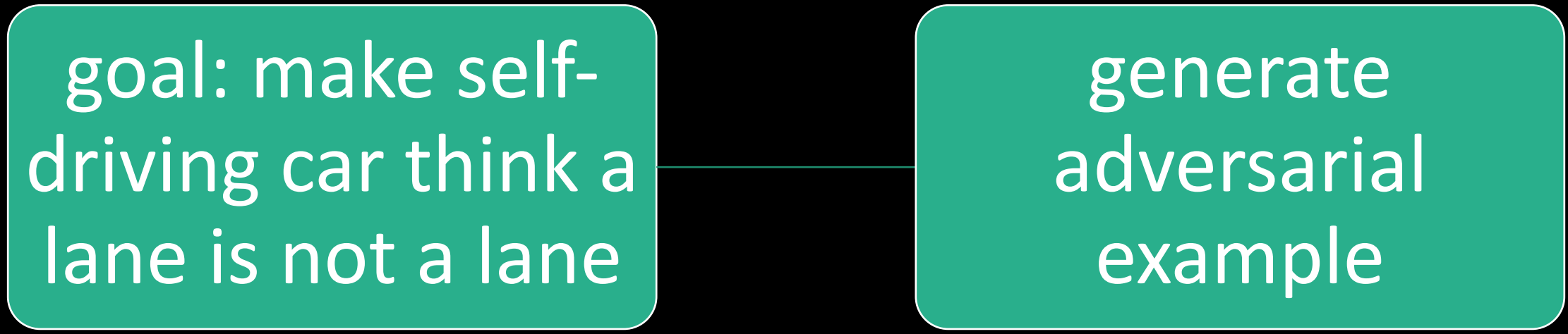
# AI red team attack tree

## Fooling AI-powered video surveillance



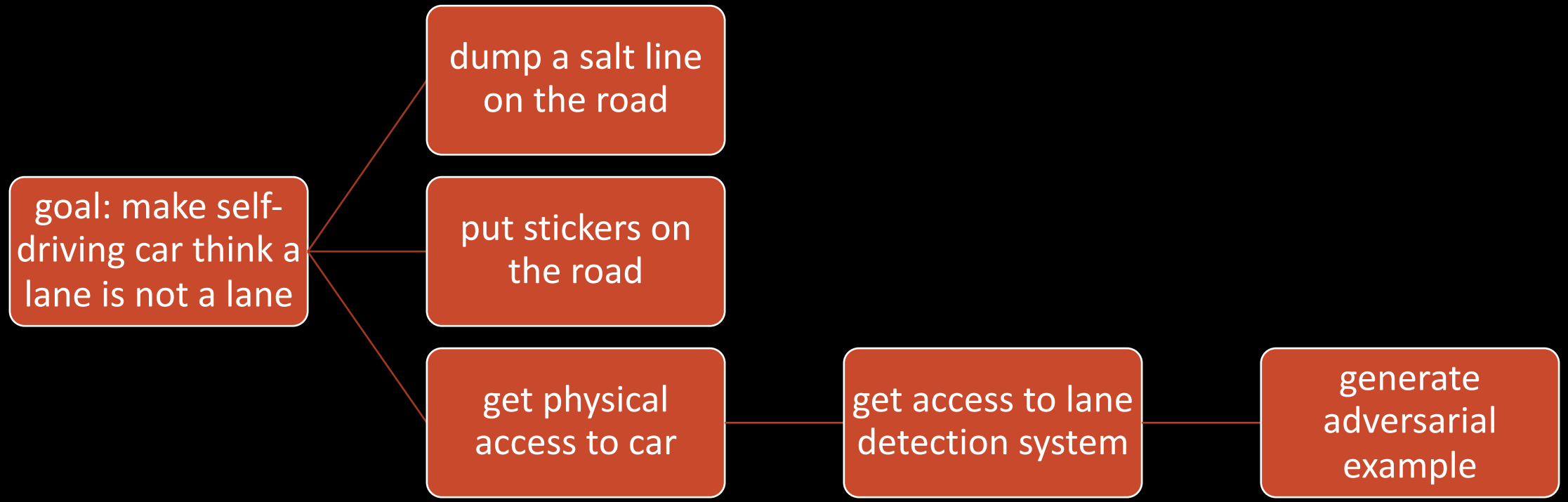
# Adversarial ML attack tree

Breaking self-driving car lane detection



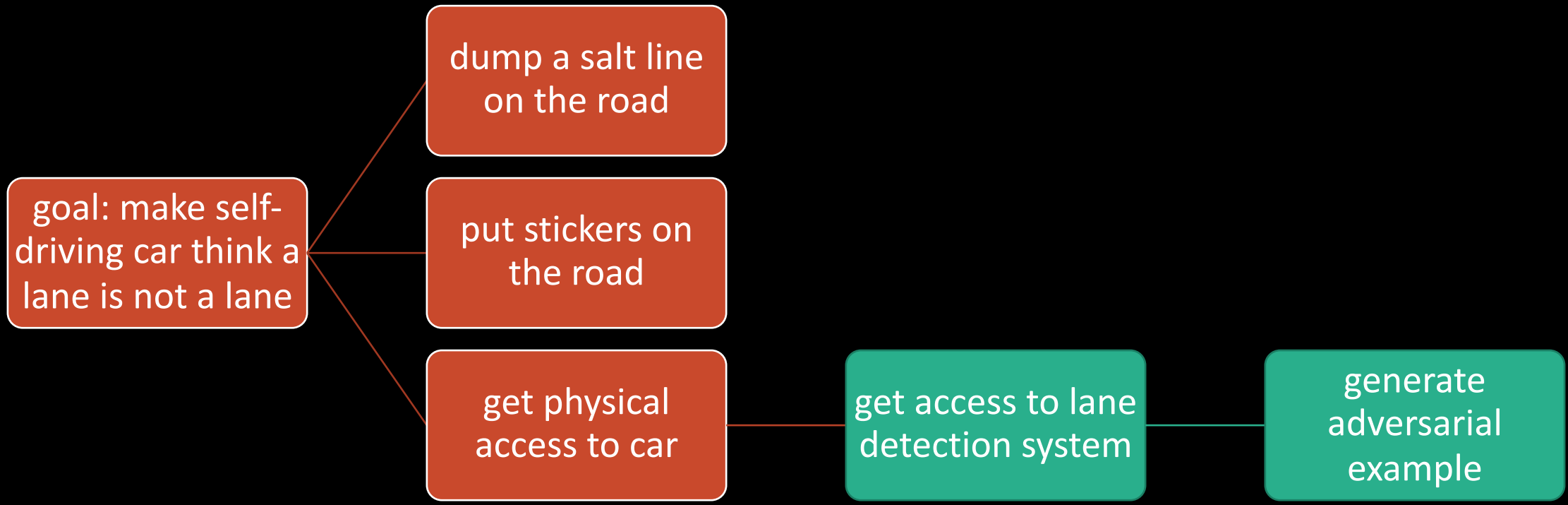
# AI red team attack tree

## Breaking self-driving car lane detection



# Adversarial ML attack tree

## Breaking self-driving car lane detection



# How to design an attack for an AI system

1. What model are you attacking?
2. Where do the data come from?
3. Where do the predictions go?

# How to design an attack for an AI system

1. What ~~model~~ system are you attacking?
2. Where do the data come from?
3. Where do the predictions go?

# How to design an attack for an AI system

1. What system are you attacking?
2. What is the data processing pipeline?
    i. Where do the data come from?
    ii. What is the data representation?
    iii. What is the output?

# How to design an attack for an AI system

1. What system are you attacking?
2. What is the data processing pipeline?
    i. Where do the data come from?
    ii. What is the data representation?
    iii. What is the output?
3. What is the threat model?

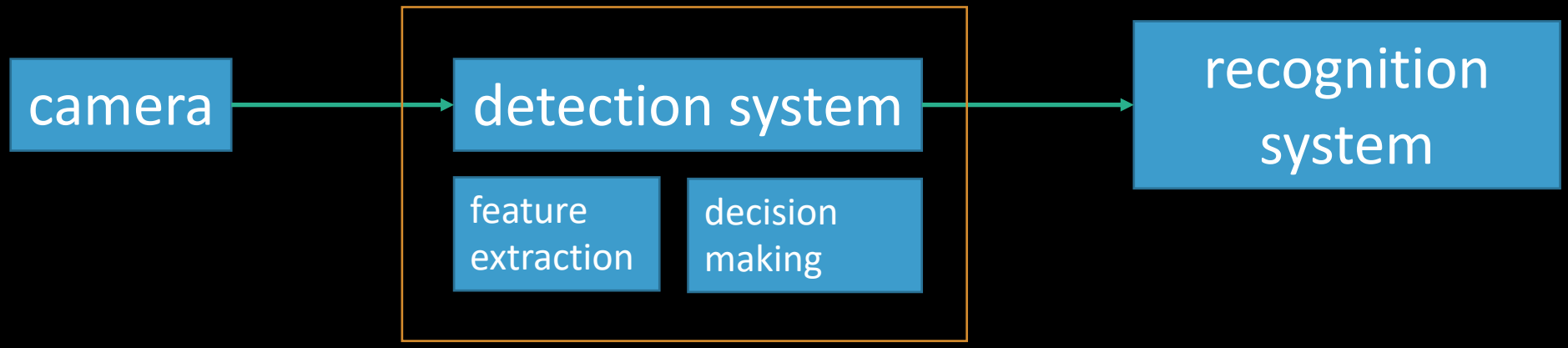
# How to design an attack for an AI system

1. What system are you attacking?



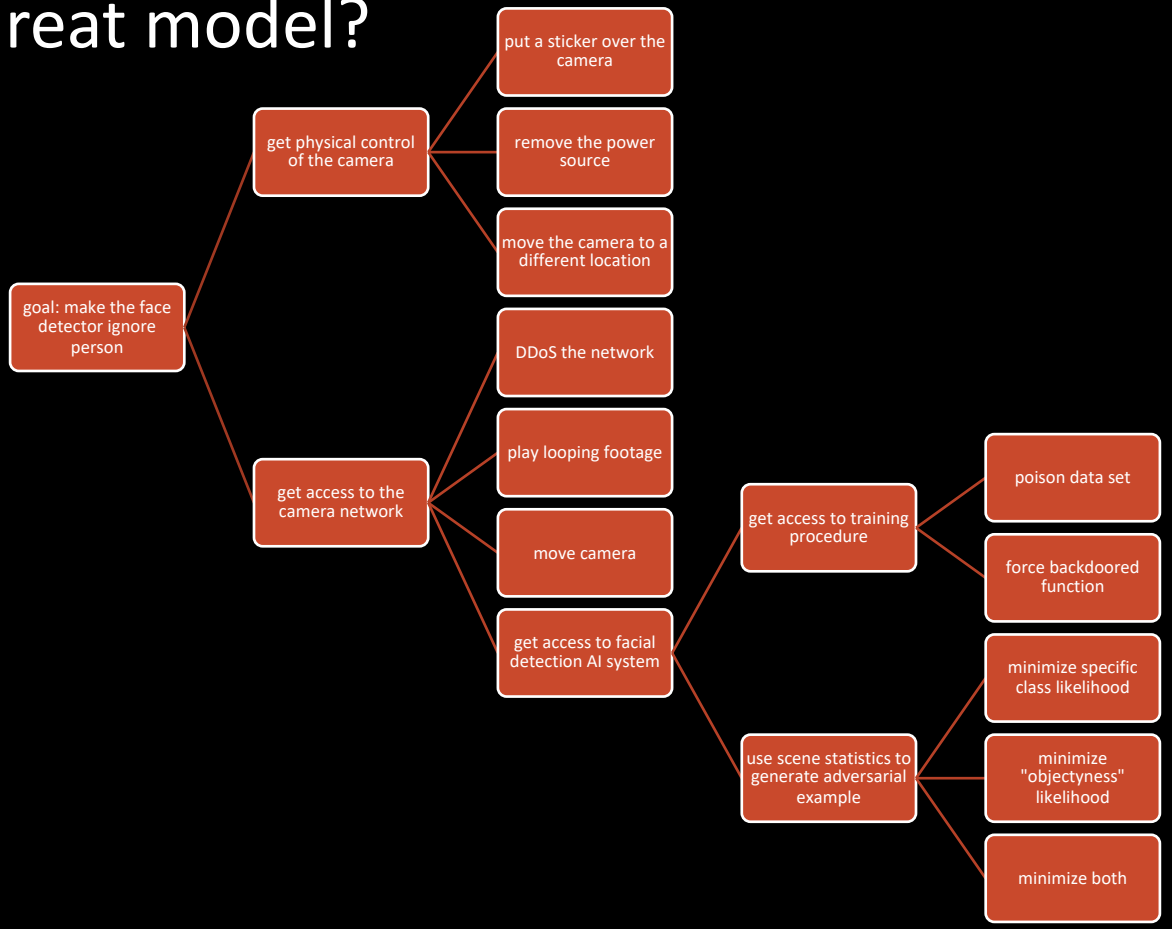
# How to design an attack for an AI system

2. What is the data processing pipeline?
   i. Where do the data come from?
   ii. What is the data representation?
   iii. What is the output?



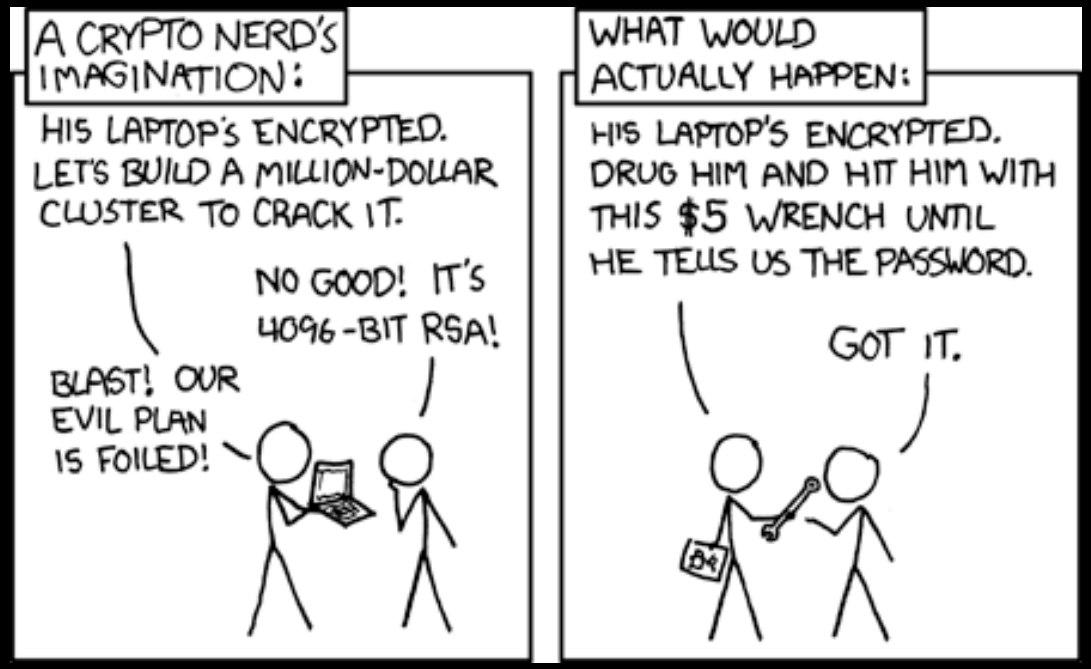
# How to design an attack for an AI system

3. What is the threat model?



# Tl;dw

- Go for the ~~jugular~~ feature extraction process
- Think about the data supply chain
- Focus on the threat model



# For more information (and math!)

Feel free to reach out via Twitter (@adversariel) and/or grab a beer with me :)

- Attacks:
  - Tencent Keen Security Lab, 2019 - https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf
  - [Biggio et al, 2018] - https://pralab.diee.unica.it/sites/default/files/biggio18-pr.pdf
  - [Fredrikson et al, 2015] - https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf
  - [Tramèr et al, 2016] - https://arxiv.org/pdf/1609.02943.pdf
  - [Thys et al, 2019] - https://arxiv.org/pdf/1904.08653.pdf
  - [Brown et al, 2017] - https://arxiv.org/pdf/1712.09665.pdf
- Threat modeling:
  - Schneier, 1999 - https://www.schneier.com/academic/archives/1999/12/attack_trees.html