Adversarial Examples and Adversarial Training

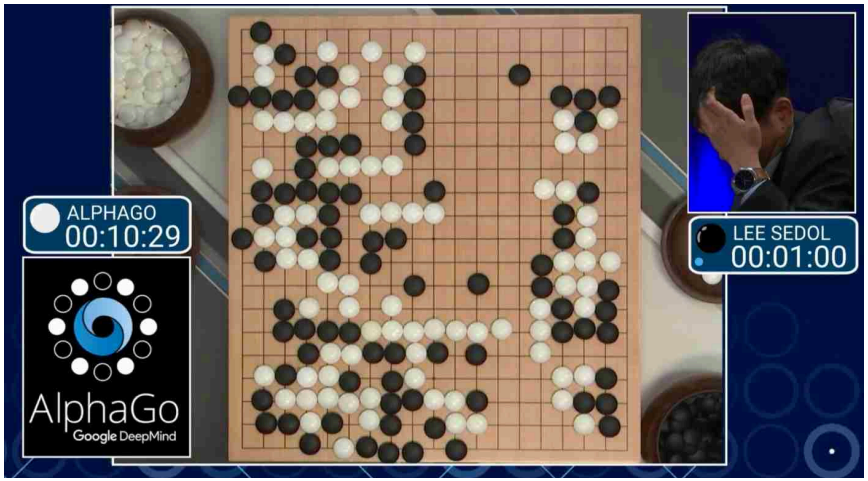
Innovative Technology Leader program

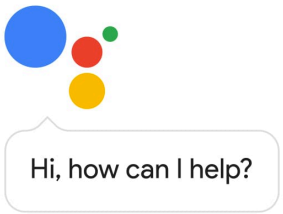
January 22nd 2018

Florian Tramèr

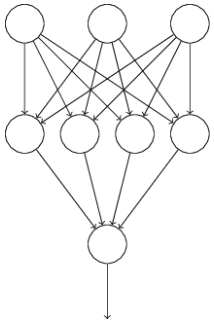
Stanford

Deep Learning is Super Smart!





Is it really?



I'm sure this is a panda

(Goodfellow et al. 2015)

Adversarial Examples in ML

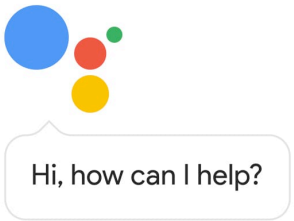
- **Images**
  Szegedy et al. 2013, Nguyen et al. 2015, Goodfellow et al. 2015, Papernot et al. 2016, Liu et al. 2016, Kurakin et al. 2016, …

- **Physical Objects**
  Sharif et al. 2016, Kurakin et al. 2017, Evtimov et al. 2017, Lu et al. 2017, Athalye et al. 2017

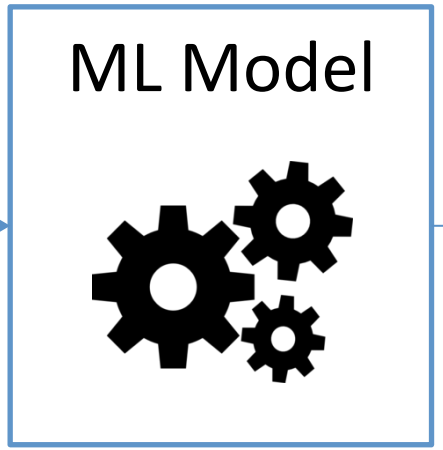
- **Malware**
  Šrndić & Laskov 2014, Xu et al. 2016, Grosse et al. 2016, Hu et al. 2017

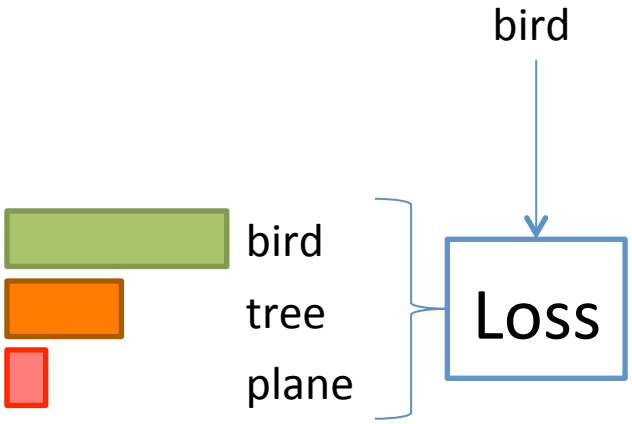
- **Text Understanding**
  Papernot et al. 2016, Jia & Liang 2017

- **Speech**
  Carlini et al. 2015, Cisse et al. 2017



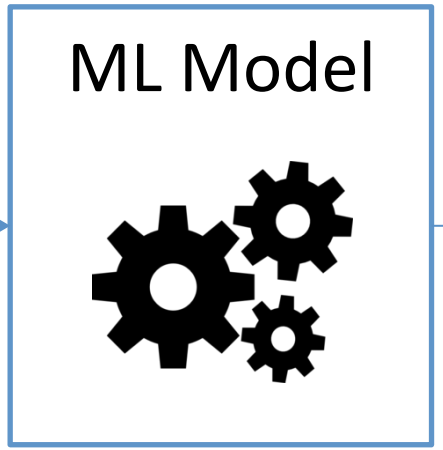
Creating an adversarial example

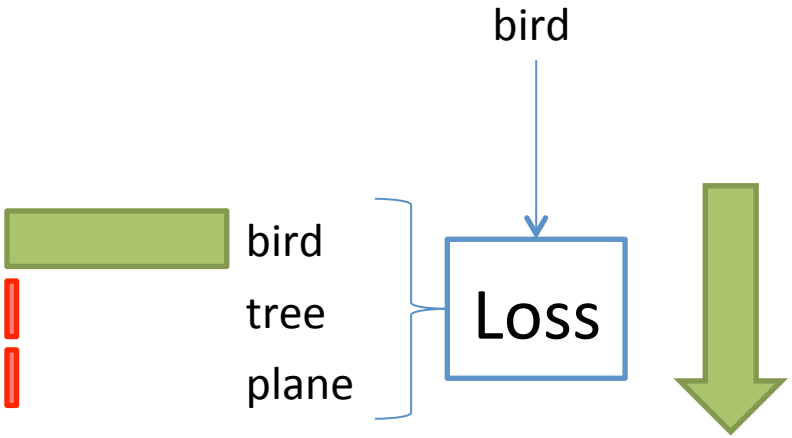




What happens if I nudge this pixel?

Creating an adversarial example

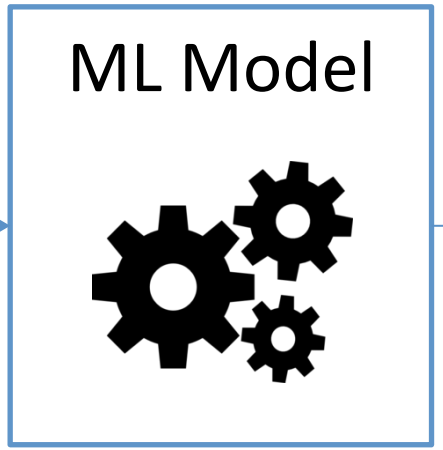


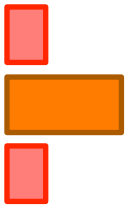


What happens if I nudge this pixel?

Creating an adversarial example

What about this one?

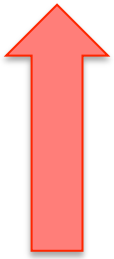




bird

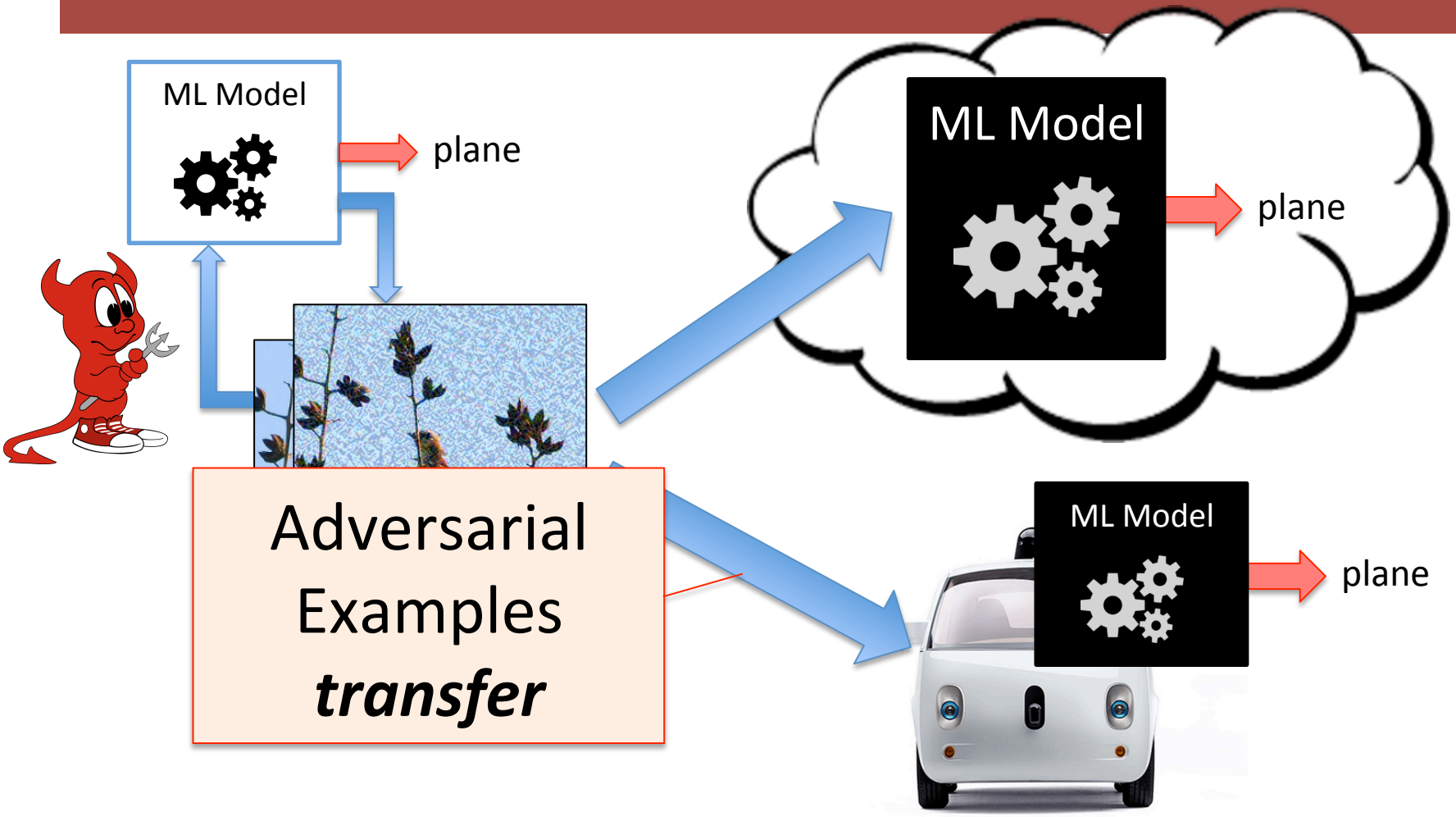
bird
tree
plane

Loss



Maximize loss with gradient *ascent*

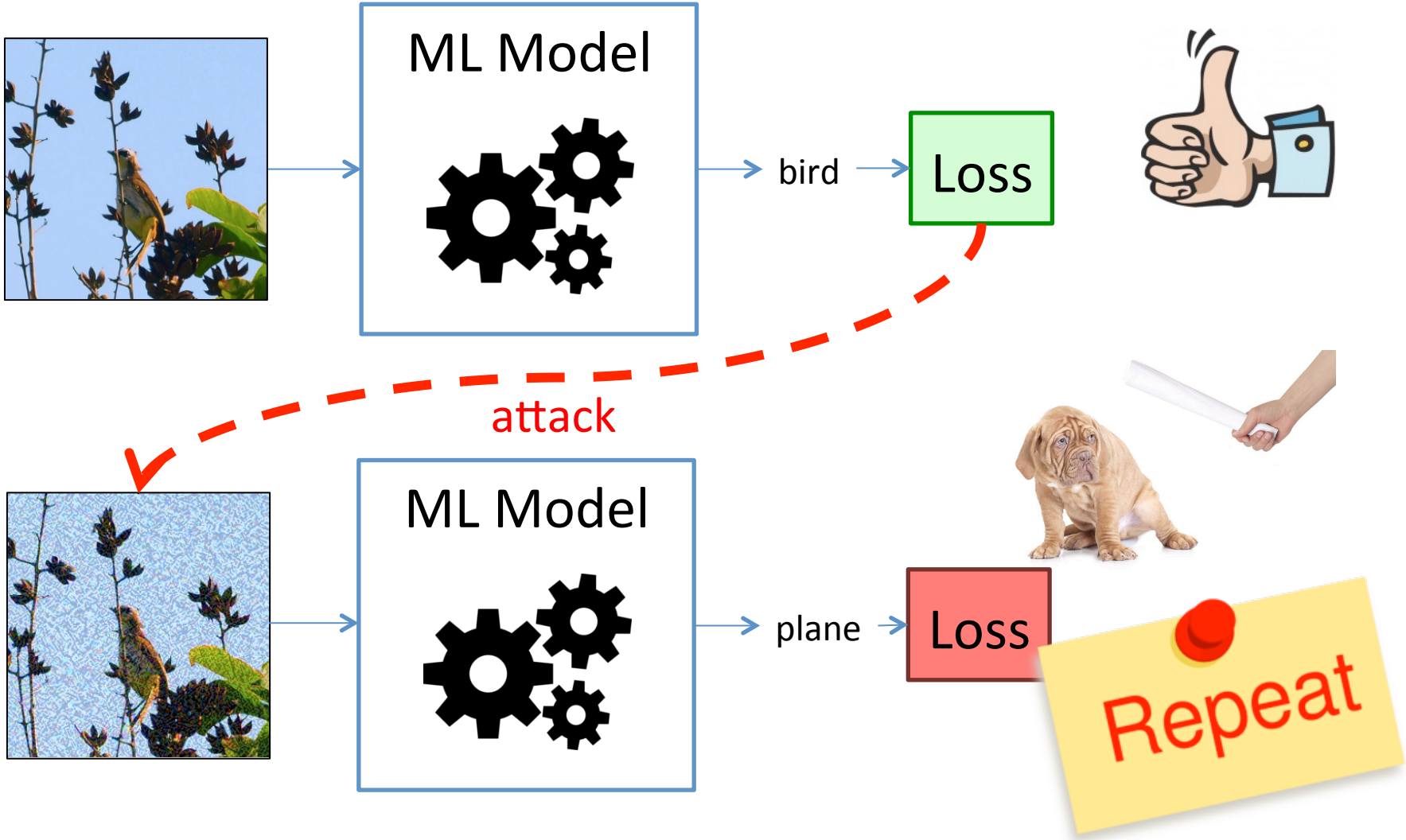
Threat Model: Black-Box Attacks



Defenses?

- Ensembles ✗
- Preprocessing (blurring, cropping, etc.) ✗
- Distillation ✗
- Generative modeling ✗
- <u>Adversarial training</u> ✗

Adversarial Training



Adversarial Training +/-

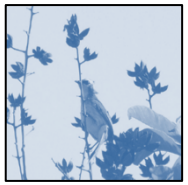
- Pros
  - Intuitive approach
  - Gives strong **formal** and **empirical** guarantees

- Cons
  - Makes assumptions on attacks
  - **Can overfit (gradient masking)**

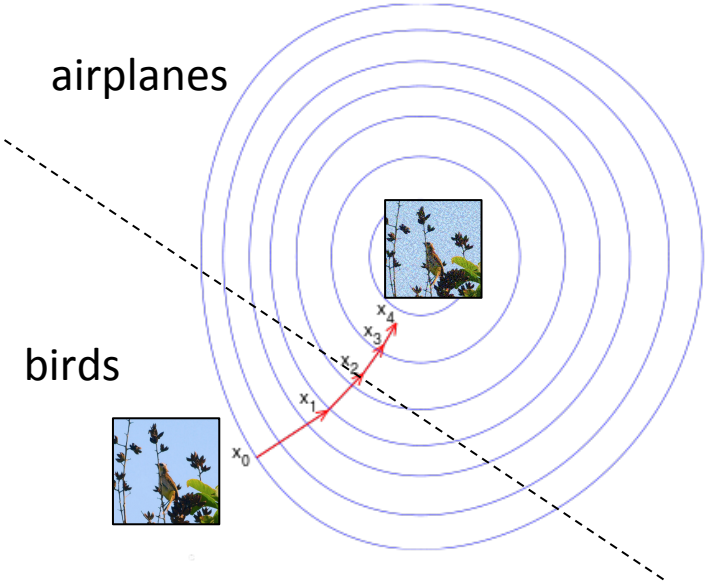
l_p noise

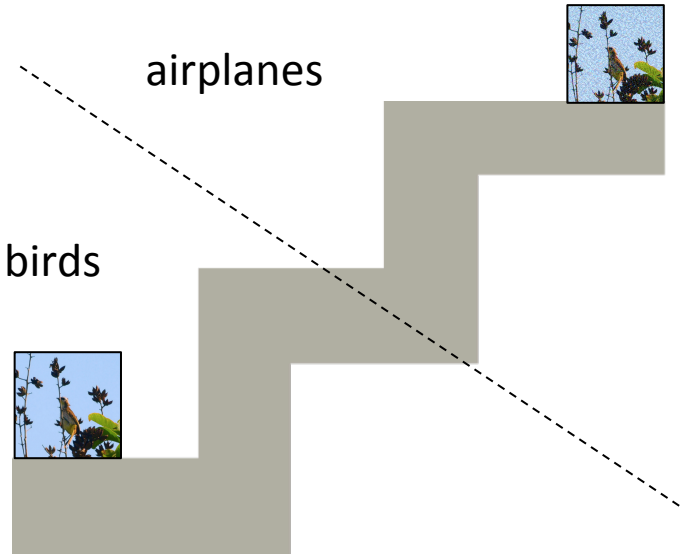
rotations

lighting



Gradient-Masking: A non-defense





"smooth" model

- Gradient-based attacks work
- Black-box attacks work
- Model is not robust!

"non-smooth" model

- Model has no useful gradients
- Black-box attacks still work!
- Model is not robust either!

TKPBM, *"Ensemble Adversarial Training: Attacks and Defenses"*, 2017