

# Automotive Security in Adversarial Learning Environments

---

CS463/ECE424

University of Illinois



# In The News: Crashing of Self-Driving Car (Uber 2018)

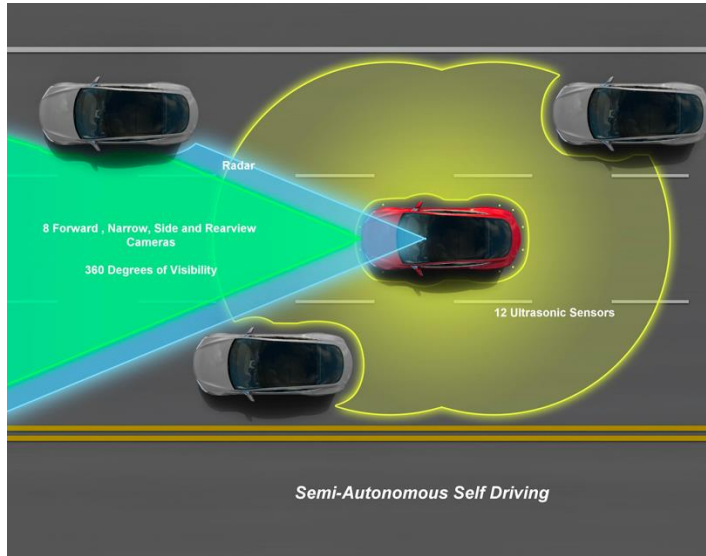
- “Inadequate safety risk assessment procedures”
- The system is not trained to react to pedestrians crossing the street outside of designated crosswalks
- Vehicle operator distracted by personal cellphone

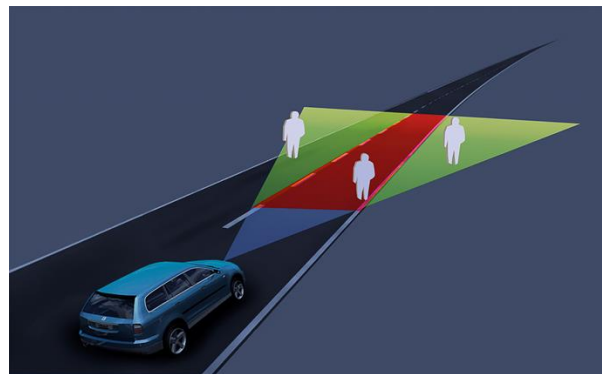


<https://www.theverge.com/2019/11/19/20972584/uber-fault-self-driving-crash-ntsb-probable-cause>

# In The News: Crashing of Self-Driving Car (Tesla 2016)

- Car's cameras failed to pick out a **white trailer** against a **bright sky** in Florida





# Case Study 1: Adversarial Examples to Attack Vision Sensors

---

Robust Physical-World Attacks on Deep Learning Visual Classification. Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song. Computer Vision and Pattern Recognition (CVPR 2018)

---

# Perils of Stationary Assumption

---

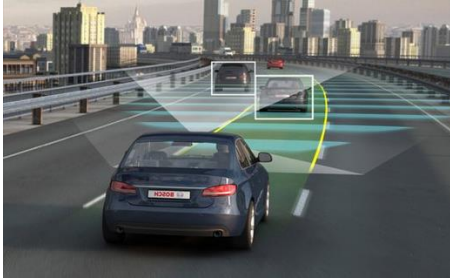
Traditional machine learning approaches assume

Training Data 

≈

Testing Data 

# Autonomous Driving in Practice



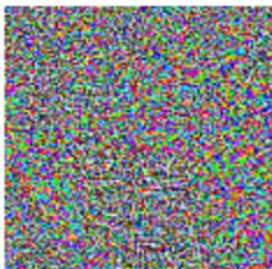
# Adversarial Examples

---



“panda”  
57.7% confidence

+ .007 ×



“nematode”  
8.2% confidence

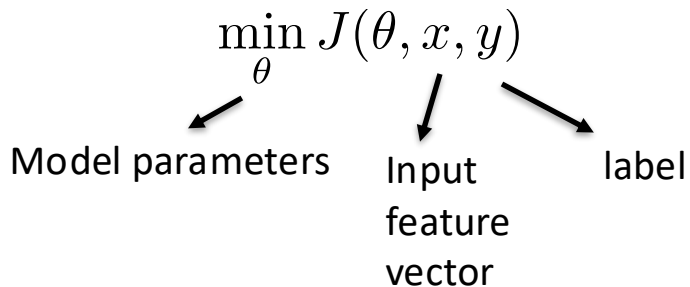
=



“gibbon”  
99.3 % confidence

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples.” *ICLR 2014*.

# Adversarial Perturbation In ML



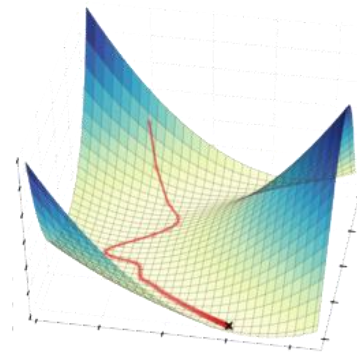
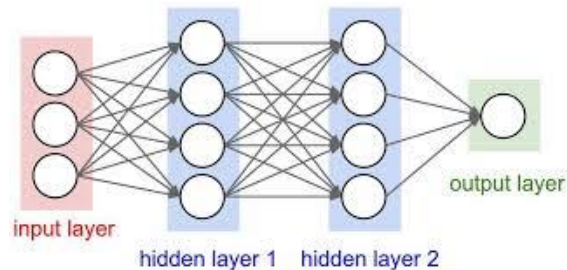
$$\max_{\epsilon} J(\theta, x + \epsilon y)$$

Adversarial perturbation

How to solve the adversary strategy

- Local search
- Combinatorial optimization
- Convex relaxation

Deep Neural Networks



Gradient Descent



# Optimization Based Attack

Large probability of  $x+\delta$  belonging to a target class  $t$

Small perturbation

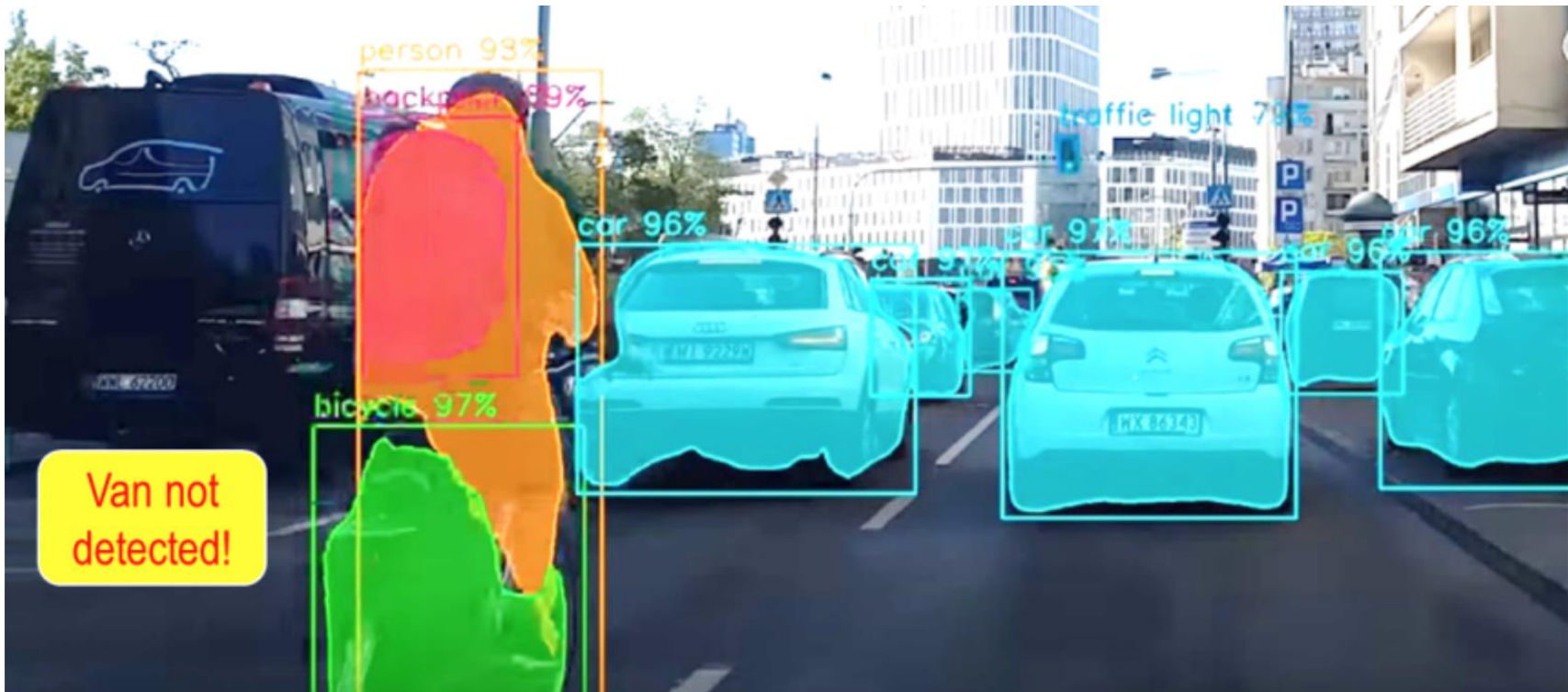
minimize  $\mathcal{D}(x, x + \delta)$   
 such that  $C(x + \delta) = t$   
 $x + \delta \in [0, 1]^n$



minimize  $\mathcal{D}(x, x + \delta) + c \cdot f(x + \delta)$   
 such that  $x + \delta \in [0, 1]^n$

	Best Case				Average Case				Worst Case			
	MNIST		CIFAR		MNIST		CIFAR		MNIST		CIFAR	
	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob
Our $L_0$	8.5	100%	5.9	100%	16	100%	13	100%	33	100%	24	100%
JSMA-Z	20	100%	20	100%	56	100%	58	100%	180	98%	150	100%
JSMA-F	17	100%	25	100%	45	100%	110	100%	100	100%	240	100%
Our $L_2$	1.36	100%	0.17	100%	1.76	100%	0.33	100%	2.60	100%	0.51	100%
Deepfool	2.11	100%	0.85	100%	-	-	-	-	-	-	-	-
Our $L_\infty$	0.13	100%	0.0092	100%	0.16	100%	0.013	100%	0.23	100%	0.019	100%
Fast Gradient Sign	0.22	100%	0.015	99%	0.26	42%	0.029	51%	-	0%	0.34	1%
Iterative Gradient Sign	0.14	100%	0.0078	100%	0.19	100%	0.014	100%	0.26	100%	0.023	100%

# Vulnerabilities of Perceptron Systems of Automobiles



# However, What We Can See Everyday...

---



# The Physical World Is... Messy

---

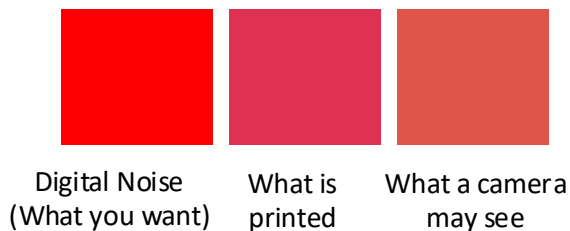
Varying Physical Conditions (Angle, Distance, Lighting, ...)



Physical Limits on Imperceptibility



Fabrication/Perception Error (Color Reproduction, etc.)



Background Modifications




# Creating Robust Physical Adversarial Examples

$$\operatorname{argmin}_{\delta} \lambda \|\delta\|_p + J(f_{\theta}(x + \delta), y^*)$$

Perturbation/Noise Matrix  $\rightarrow$   $\delta$   $\rightarrow$   $\lambda \|\delta\|_p$   $\rightarrow$   $J(f_{\theta}(x + \delta), y^*)$   $\rightarrow$  Adversarial Target Label

Lp norm (L-0, L-2, ...)

Loss Function

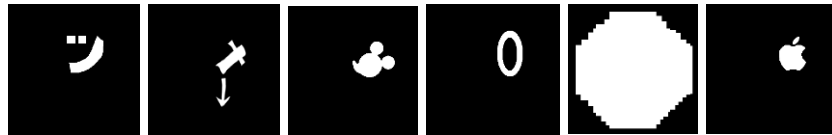
$$\operatorname{argmin}_{\delta} \lambda \|\delta\|_p + \frac{1}{k} \sum_{i=1}^k J(f_{\theta}(x_i + \delta), y^*)$$


The image shows a sequence of six stop signs, each with a different perturbation applied to it. The signs are arranged horizontally and enclosed in large curly braces. A red arrow points from the summation symbol in the equation above to the fourth stop sign from the left.

# Optimizing Spatial Constraints

(Handling Limits on Imperceptibility)

$$\operatorname{argmin}_{\delta} \lambda \|M_x \cdot \delta\|_p + \frac{1}{k} \sum_{i=1}^k J(f_{\theta}(x_i + M_x \cdot \delta), y^*)$$



Subtle Poster  
Camouflage Sticker

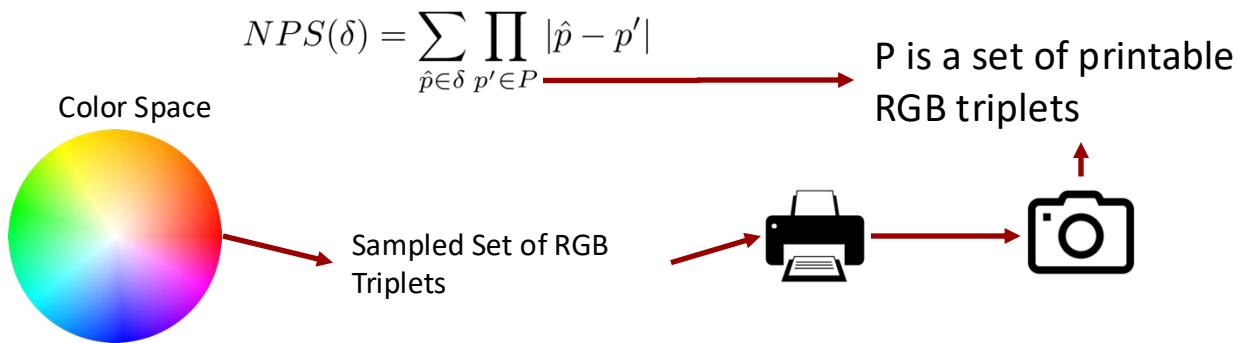
Mimic vandalism

“Hide in the  
human psyche”



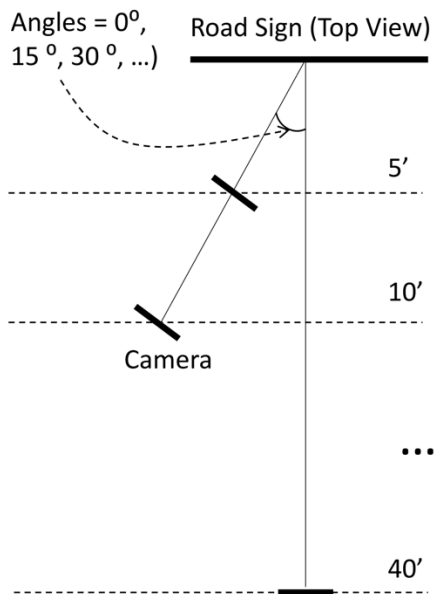
# Handling Fabrication/Perception Errors

$$\operatorname{argmin}_{\delta} \lambda \|M_x \cdot \delta\|_p + \frac{1}{k} \sum_{i=1}^k J(f_{\theta}(x_i + M_x \cdot \delta), y^*) + NPS(M_x \cdot \delta)$$

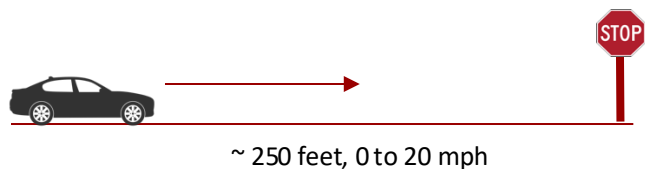


# How Can We Realistically Evaluate Attacks?

## Lab Test (Stationary)



## Field Test (Drive-By)



- Record video
- Sample frames every  $k$  frames
- Run sampled frames through DNN





Subtle  
Poster

Subtle  
Poster

Camo  
Graffiti

Camo Art

Camo Art

Lab Test (Stationary)

Target Class:

**Speed Limit 45**

# Art Perturbation

---



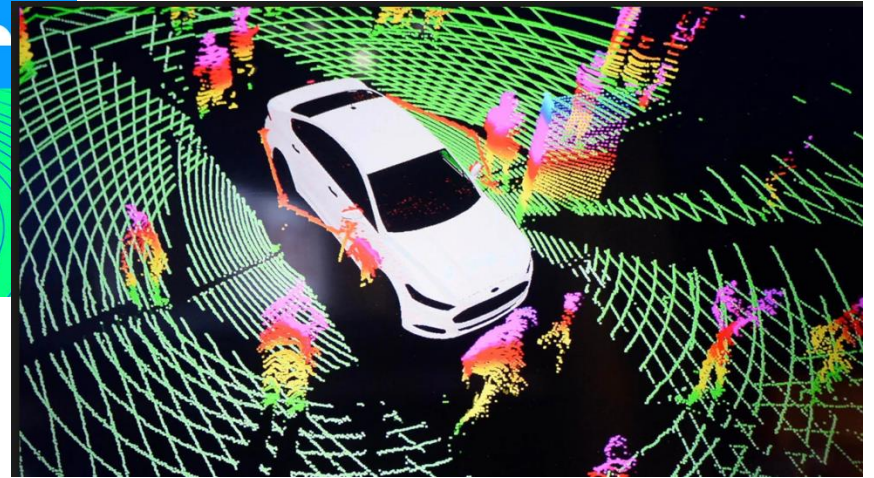
# Subtle Perturbation

---

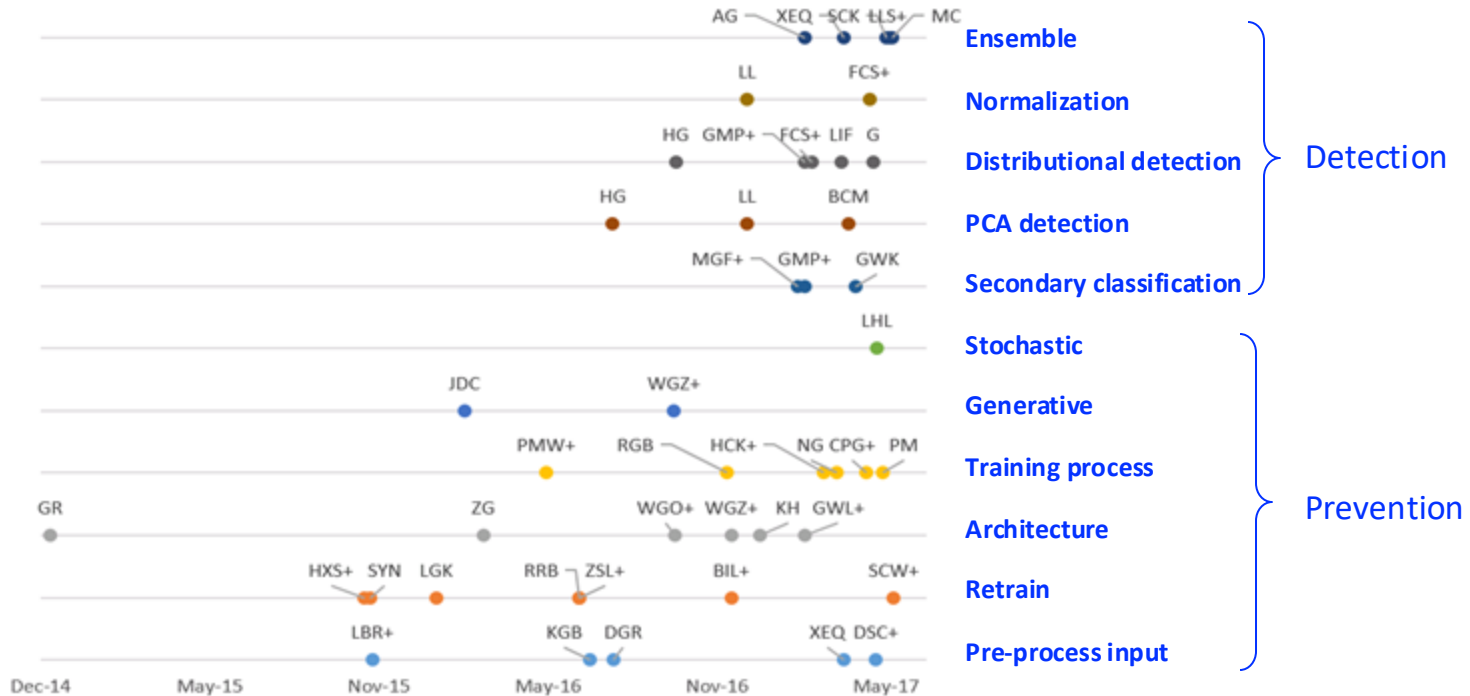


# Thinking more about Physical objects

Similar attack against LiDAR sensors



# Numerous Defenses Proposed





# Case Study 2: Attacking GPS Sensors

---

All Your GPS Are Belong To Us: Towards Stealthy Manipulation of Road Navigation Systems. Kexiong (Curtis) Zeng, Shinan Liu, Yuanchao Shu, Dong Wang, Haoyu Li, Yanzhi Dou, Gang Wang, Yaling Yang Proceedings of *The 27th USENIX Security Symposium (USENIX Security)*, Baltimore, MD, August 2018.

---

# GPS Navigation Systems used by 1+ Billion Users

---

- GPS navigation is widely used by drivers around the world



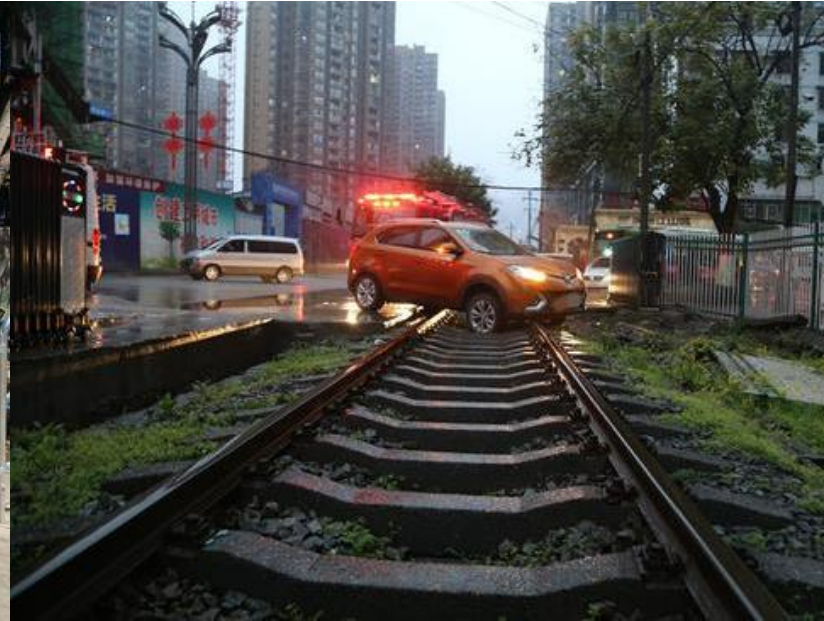
- Self-driving cars rely on GPS for navigation and on-road decisions





# GPS Navigation Systems used by 1+ Billion Users

---

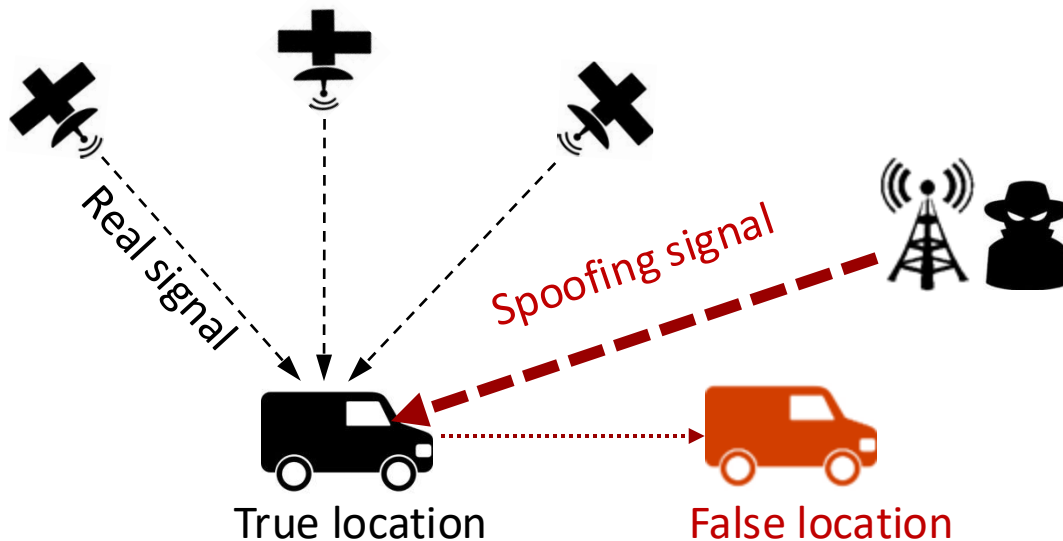


GPS malfunction can lead to real-world consequences

# Known Threat: GPS Spoofing

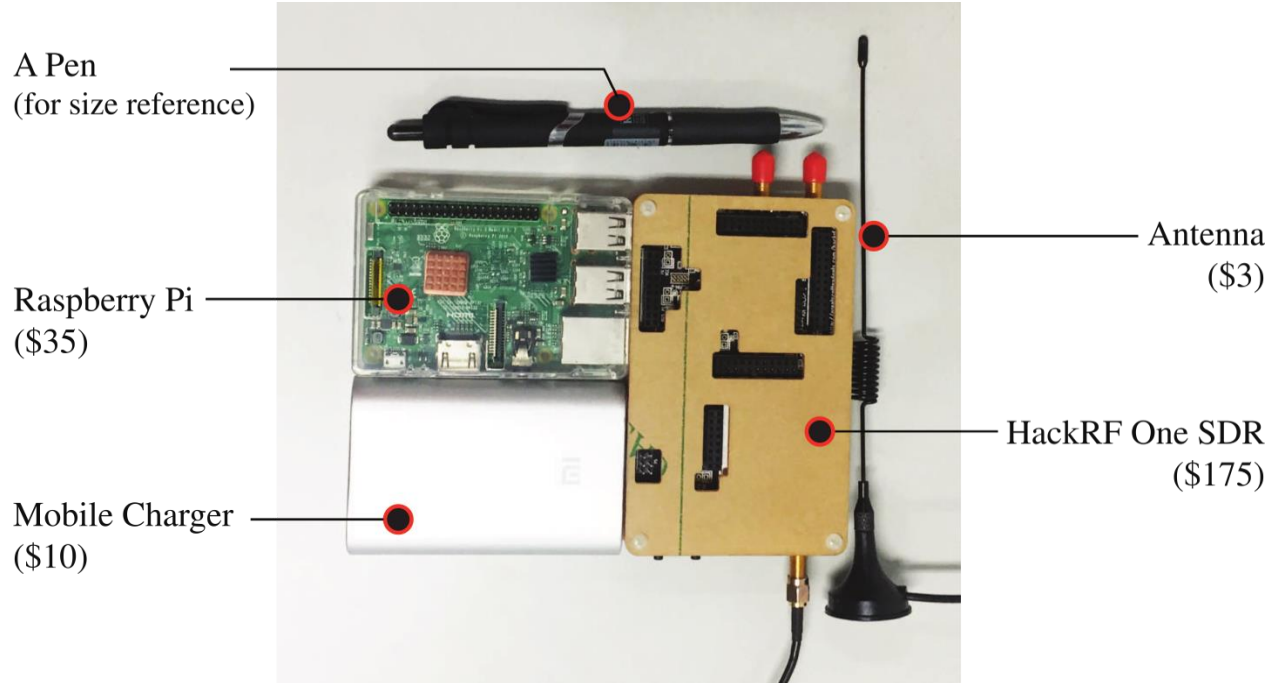
---

- **Civilian GPS** is vulnerable to spoofing attacks
  - A lack of authentication of signal source
- Take over victim GPS via brute-force jamming or smooth methods



# Portable GPS Spoofer is Affordable (\$223)

---



# GPS Spoofing in Free Space (Air, Water)

---

In 2012, a drone was diverted in White Sands, New Mexico



In 2013, a yacht was diverted on the way from Monaco to Greece

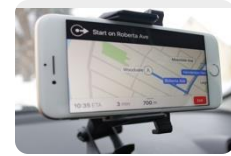


# Spoofing Road Navigation: More Challenging

Physical World



Digital Map



“Turn left” - physically impossible instruction!  
Easily alert human drivers

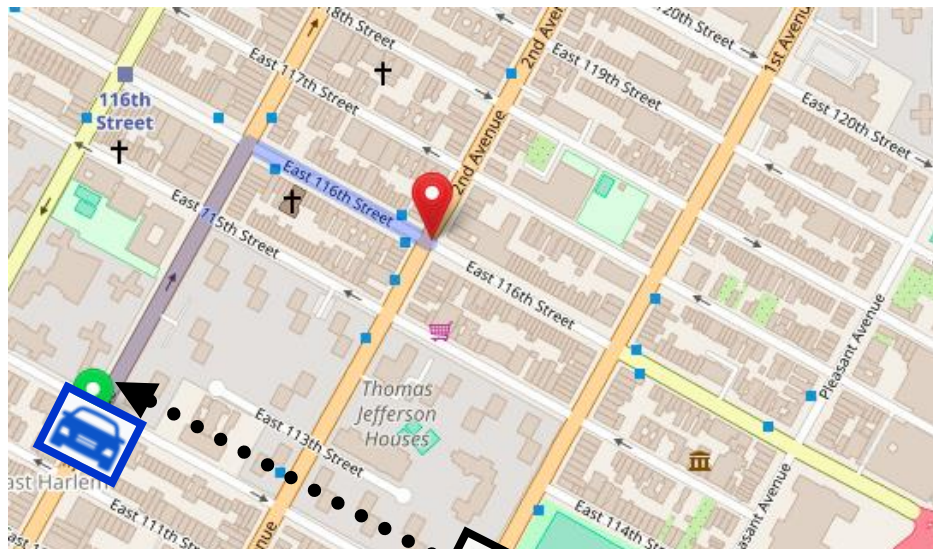
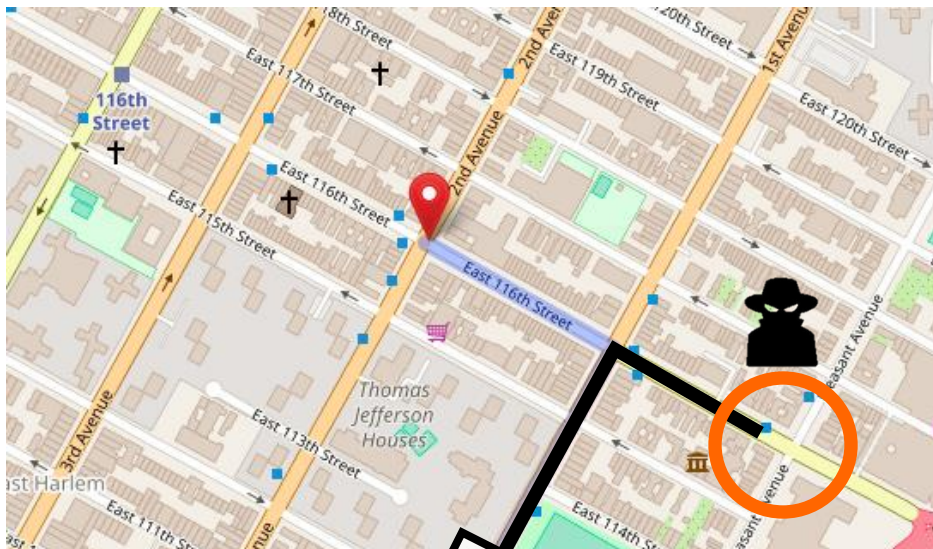
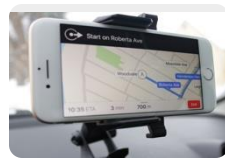


# Making the Attack More Stealthy

Physical World



Digital Map



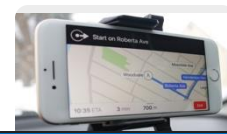
Navigation instructions lead to attacker's pre-defined location

# Core Idea: Calculate Spoofing Location and Timing

Physical World

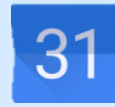


Digital Map



Goal: find ghost route to mimic the shape of victim route

Assumption: know rough destination area or checkpoint



Victim route

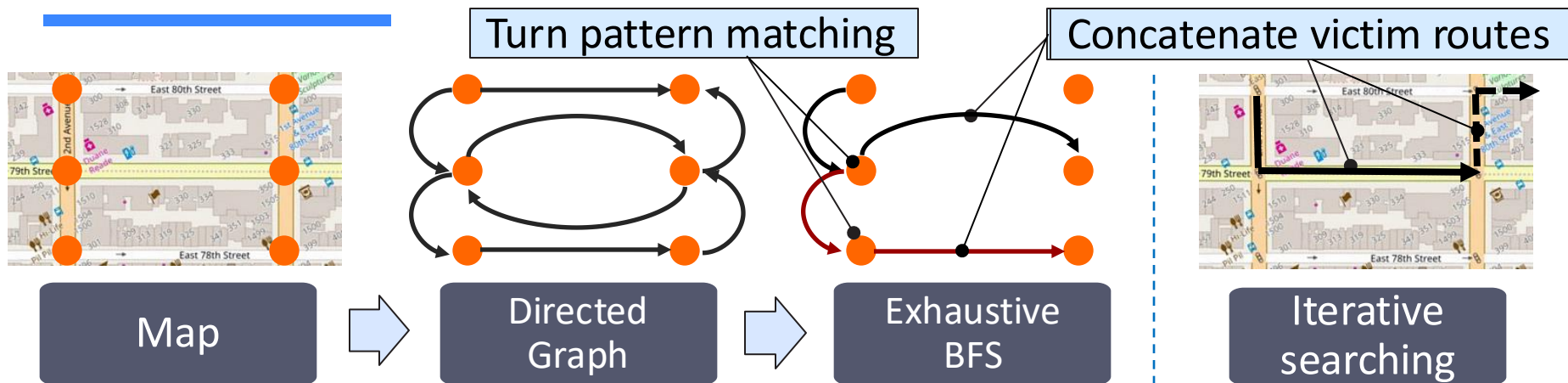


Ghost route



Ghost location

# Route Searching Algorithm

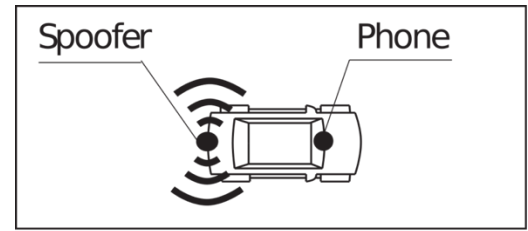


## Trace Driven Evaluation

- 600 real-world trips from the taxi datasets of New York City and Boston
- Deviating attack: 3,507 qualified victim routes per trip
- Endangering attack: 599 out of 600 (99.8%) contains wrong-way

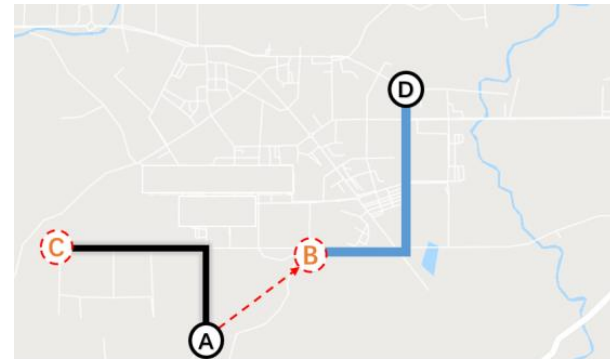
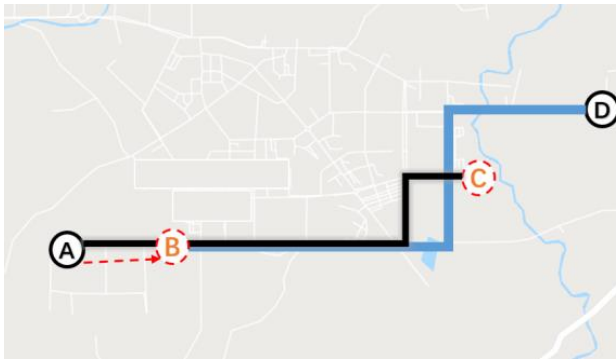


# Real-world Experiments



- Experiments with **legal permission** from local authority and **IRB approval**
  - After midnight, spoofing signals do not affect outside (-127.41 dBm)

Trigger instructions in time and divert to 2.1 & 2.5 km away



# Can Human Users Detect the Attack?

---

- Let users drive in a simulator
  - Play truck drivers to “deliver packages” from location A to B
  - Advertise the study as a usability study, spoof locations in real time
  - **Post-study interview** to know why users can/cannot detect the attack



Experiment setup



Simulator view



Google Street View

# User Study Results

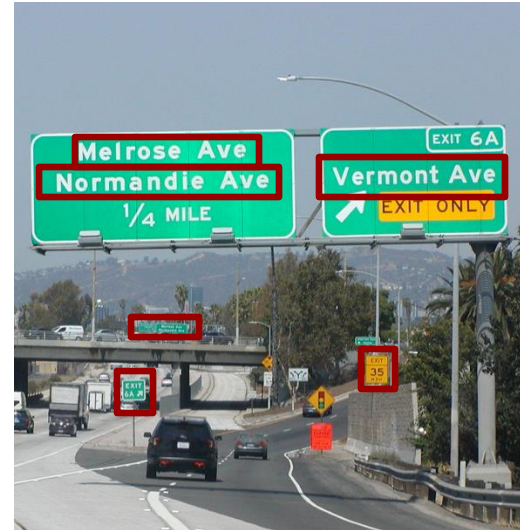
---

- Attack success rate: 95% (38 out of 40)
  - Two users detect it by cross-checking surrounding environment and the map to find inconsistency (Highway vs. local way)
- Users are more likely to use GPS in unfamiliar areas
  - Not enough pre-knowledge/time to check the inconsistency
  - Heavily rely on voice and turn-by-turn instructions
- Most users experienced GPS malfunction in real life
  - Unstable GPS signal does not alert users

# Take-aways: Learning from Users

---

- It is feasible to manipulate road navigation systems
  - Advanced GPS spoofing strategies
  - Works even when humans are in the loop
- Defense ideas inspired by the user study
  - Cross-check data from digital and physical worlds
    - Computer vision-based localization
  - GPS-free localization & navigation





# Remarks

---

- Different sensors in automobile could be vulnerable against adversarial attacks
- Different attacks are optimized differently but they have common adversarial goals
- General/universal defense is hard, but we can leverage certain properties of learning tasks and develop more robust models

# Discussion Questions

---

- What does it take to make you feel safe to ride in a self-driving vehicle?
- Do you prefer a world of autonomous vehicles or the coexistence of human drivers and autonomous driving (or human drivers only)?
- Can you point out other security/privacy challenges faced by autonomous driving systems?