# Autonomous Cars

Daria Zahaleanu

# Self-Driving Cars

Utopian view

- Save lives (1.3 million die every year in manual driving)
- 4D's of human folly: drunk, drugged, distracted, drowsy driving
- Eliminate car ownership
- Increase mobility and access
- Save money
- Make transportation personalized, efficient, and reliable

Dystopian view

- Eliminate jobs in the transportation sector
- Failure (even if much rarer) may not depend on factors that are human interpretable or under human control
- Artificial intelligence systems may be biased in ways that do not coincide with social norms or be ethically grounded
- Security

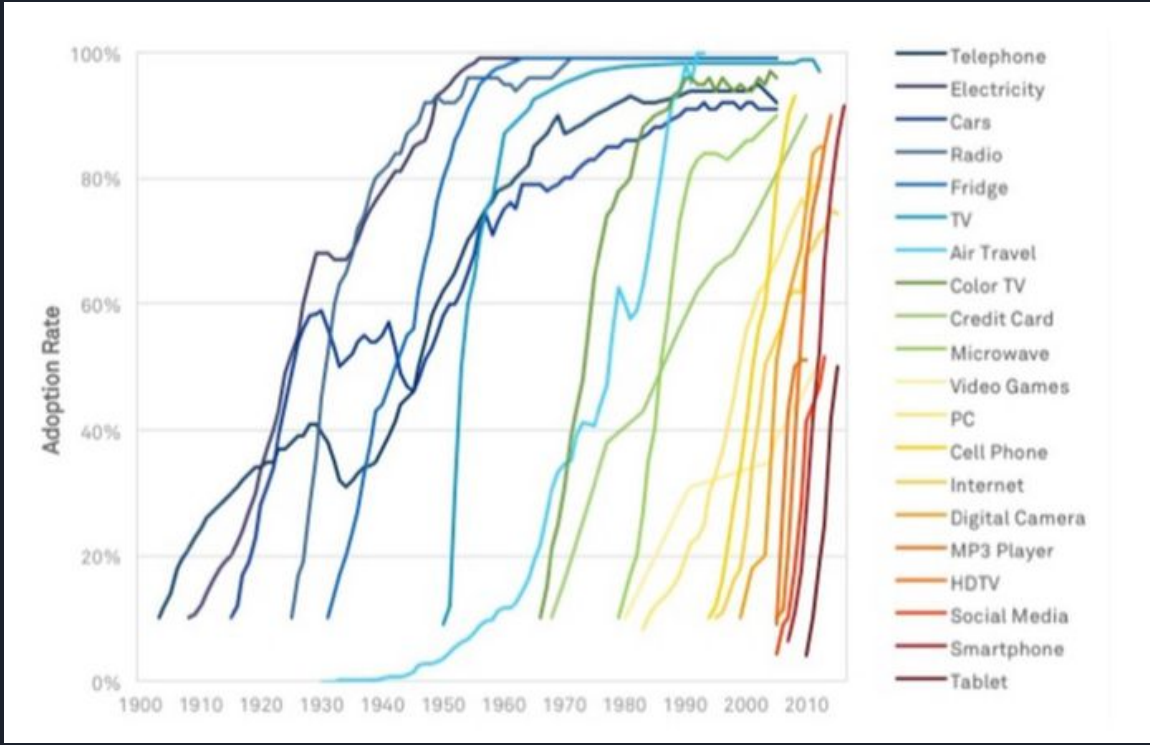
# Autonomous Cars: Grain of Salt

- Our intuition about what is hard or easy for AI is flawed
- Carefully differentiate between:

A. Doubtful: Promises for future vehicles (in 2+ years)
B. Skeptical: Promises for future vehicles (in 1 year)
C. Possible: Actively testing vehicles on public roads at scale·
D. Real: Available for consumer purchase today

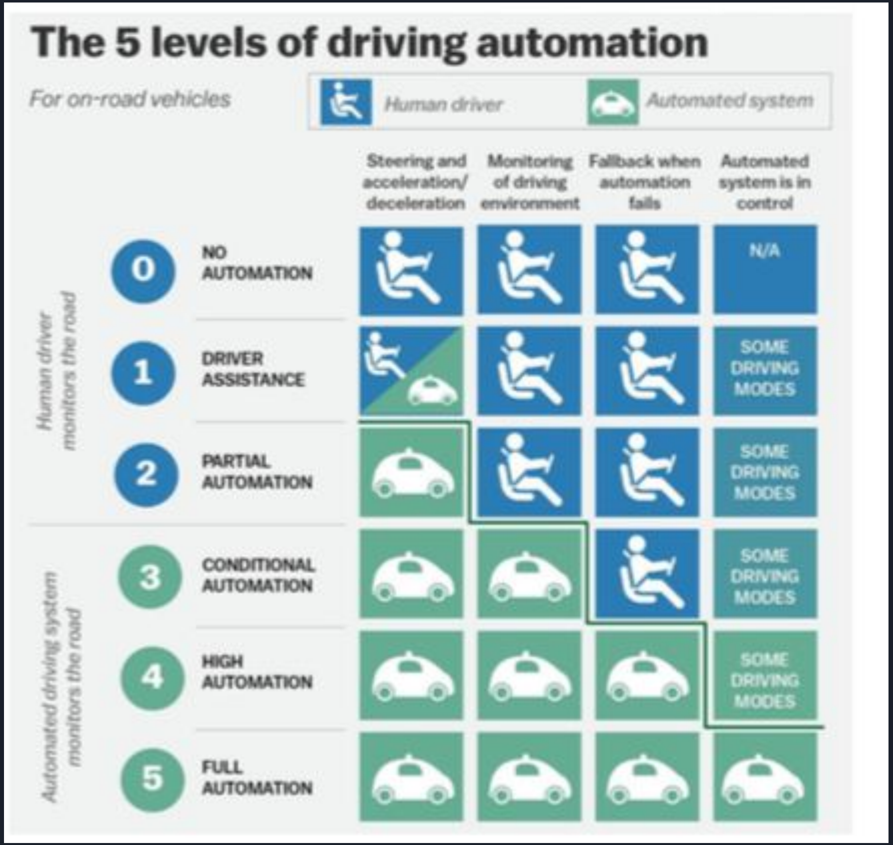
- Rodney Brooks prediction in “My Dated Predictions”:
- >2032: A driverless "taxi" service in a major US city with arbitrary pick and drop off locations, even in a restricted geographical area.
- >2045: The majority of US cities have the majority of their downtown under such rules

# Evolution




*Source: MIT*

# Levels of Automation




*Source: MIT*

# Autonomous Cars and Safety

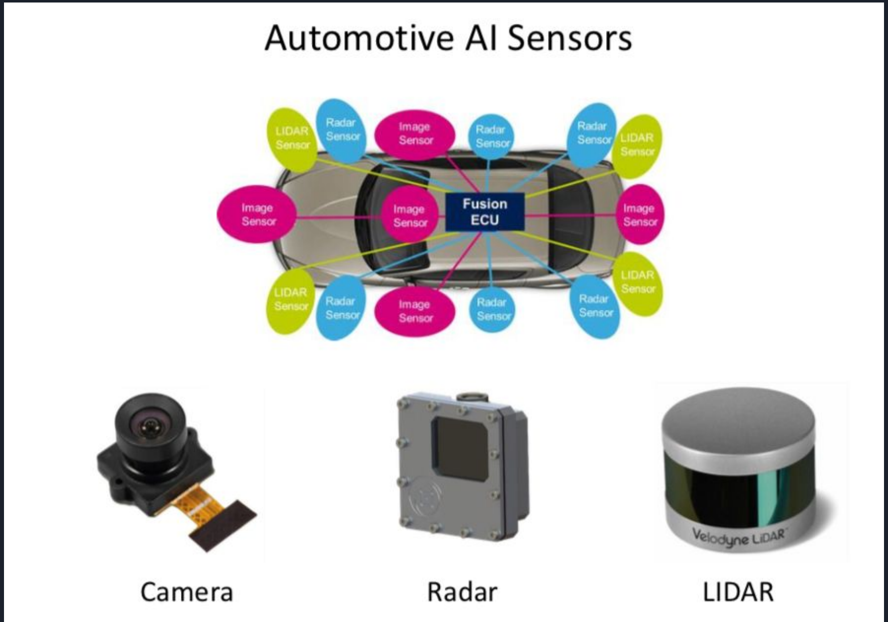
- 30% of the Americans cited safety concerns when asked if they would like to ride in a self-driving car
- Autonomous cars must fight cyber attacks under California's new rules
- Should the government be responsible for the cyber attacks against AVs? Where do automakers stand in this regard?

# Why are ACs vulnerable to cyber attacks?

- Electronic sensors commanded remotely using software
- Increase in communication channels
- Security Testing approaches

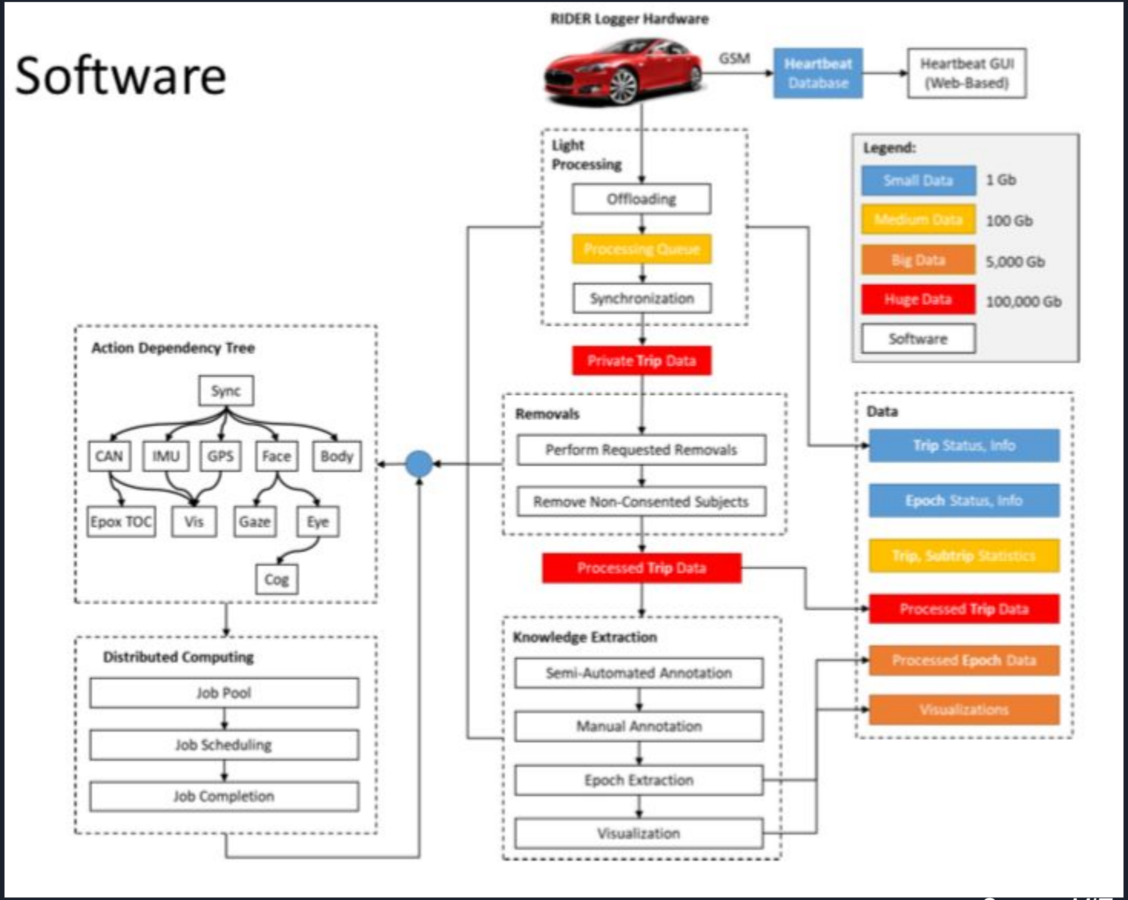
Possible attack consequences:

- Manipulation of biometric authentication
- Car crashes
- Theft of illicit or illegal content



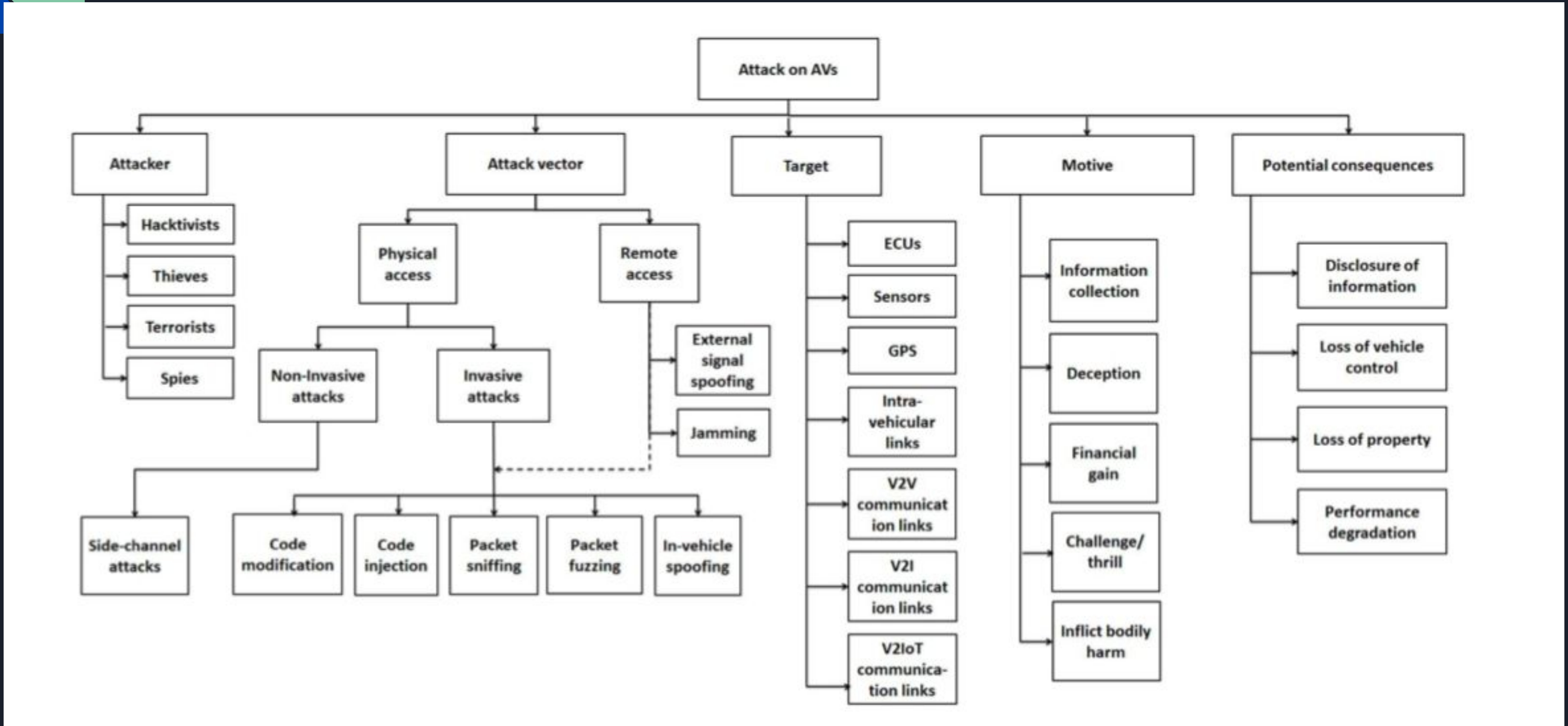

*Source: MIT*

# Software used by AVs



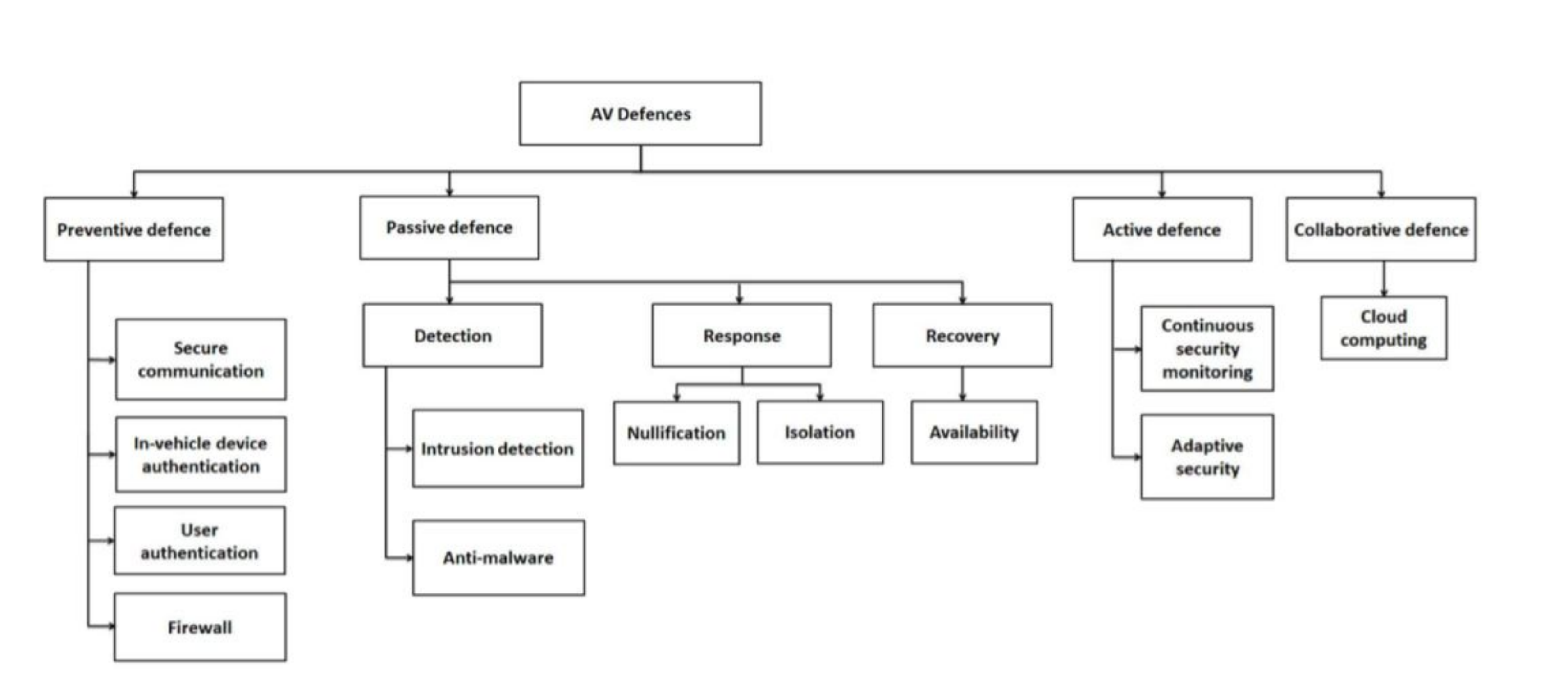

Source: MIT

# Attack Taxonomy



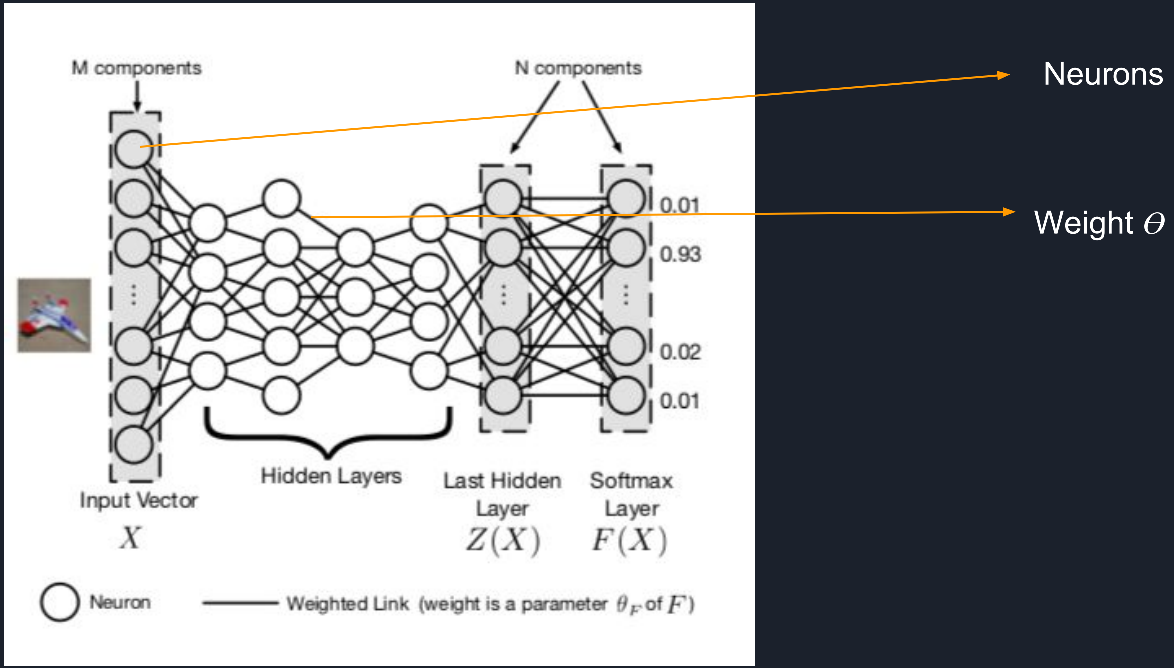

Source: L. L Ting et al. [1]

# Defense Taxonomy




*Source: L. L Ting et al. [1]*

# Deep Neural Networks, Classifiers, Adversarial Samples

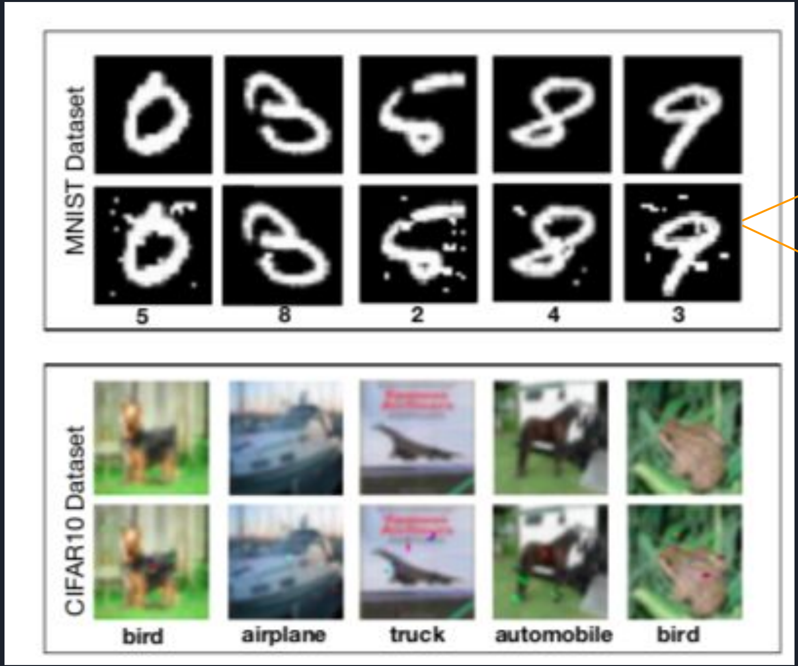



DNN Architecture

*Source: Papernot et al.*

# Adversarial Samples

In most of the cases, the adversary's goal is to produce a minimally altered version of the input x (image, video, text etc) such that it changes the output of the DL model, without being perceptible to the human eye.




$$\vec{x}^* = \vec{x} + \arg\min\{\vec{z} : \tilde{O}(\vec{x}+\vec{z}) \neq \tilde{O}(\vec{x})\} = \vec{x} + \delta_{\vec{x}}$$

$$\tilde{O}(\vec{x}^*) \neq \tilde{O}(\vec{x})$$

Adversarial Samples Perturbation

*Source: Papernot et al.*

# Adversarial Goals and Capabilities

Goals:

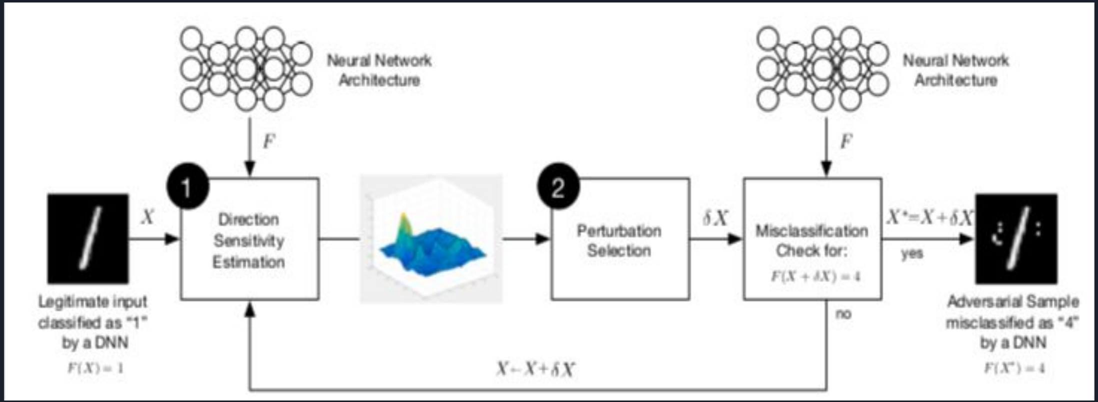
1. Confidence reduction
2. Misclassification
3. Targeted misclassification
4. Source/target misclassification

# Types of Attacks

1. Defensive Distillation
2. Robust Physical - World Attacks
3. Black - Box Attack

# Defensive Distillation

- distillation is a training method designed to train a DNN using knowledge transferred from a different DNN
- defensive distillation is a type of distillation that uses the knowledge from a DNN to improve its resilience to adversarial samples
- defensive distillation reduces the effectiveness of adversarial samples from 95% to 0.5% (*Papernot et al.*) and it smoothes out classifier models, by reducing the sensitivity of a DNN to input perturbations by a factor of 10^30




*Source: Papernot et al.*

# Robust Physical - World Attacks

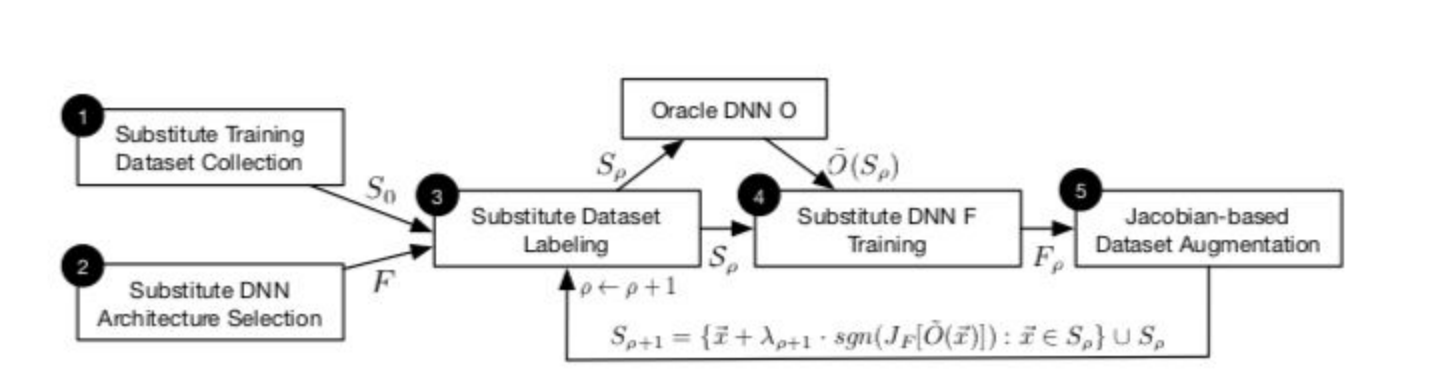
- robust physical perturbation attacks generated random perturbations by adding stickers and graffiti that would lead to the misclassification of object by DNN, without arousing suspicion in human operators
- the goal of this type of attack is to effectively create adversarial samples where the object itself is physically perturbed by placing stickers on it
- Physical challenges for this attack include: environmental conditions, spatial constraints, physical limits on imperceptibility, fabrication errors and nonetheless the angle and distance from the camera of the generated perturbations.

| Distance/Angle | Subtle Poster | Subtle Poster Right Turn | Camouflage Graffiti | Camouflage Art (LISA-CNN) | Camouflage Art (GTSRB-CNN) |
| --- | --- | --- | --- | --- | --- |
| 5′ 0° |  |  |  |  |  |
| 5′ 15° |  |  |  |  |  |
| 10′ 0° |  |  |  |  |  |
| 10′ 30° |  |  |  |  |  |
| 40′ 0° |  |  |  |  |  |
| Targeted-Attack Success | 100% | 73.33% | 66.67% | 100% | 80% |

*Source: Papernot et al.*

# Black-Box Attacks

1. in black-box attacks, the adversary doesn't have any knowledge about the model, except for the the labels
2. the goal is to produce a minimal perturbation to input X, sufficient to determine the DNN to misclassify it, but imperceptible enough for the human eye
3. approach: use DNN as an oracle to synthesized data and generate a synthetic data set S0 to build a model F that approximates the oracle's decision




*Source: Papernot et al.*

# References


1. L.L. Thing, Vrizlynn & Wu, Jiaxi. (2016). Autonomous Vehicle Security: A Taxonomy of Attacks and Defences. 164-170. 10.1109/iThings-GreenCom-CPSCom-SmartData.2016.52.
2. Papernot, Nicolas & McDaniel, Patrick & Jha, Somesh & Fredrikson, Matt & Berkay Celik, Z & Swami, Ananthram. (2015). The Limitations of Deep Learning in Adversarial Settings.
3. Papernot, Nicolas & McDaniel, Patrick & Wu, Xi & Jha, Somesh & Swami, Ananthram. (2016). Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. 582-597. 10.1109/SP.2016.41.
4. Papernot, Nicolas & McDaniel, Patrick & Goodfellow, Ian & Jha, Somesh & Berkay Celik, Z. & Swami, Ananthram. (2016). Practical Black-Box Attacks against Machine Learning. eprint arXiv:1602.02697.
5. Szegedy, Christian & Zaremba, Wojciech & Sutskever, Ilya & Bruna, Joan & Erhan, Dumitru & Goodfellow, Ian & Fergus, Rob. (2013). Intriguing properties of neural networks.
6. Evtimov, Ivan & Eykholt, Kevin & Fernandes, Earlence & Kohno, Tadayoshi & Li, Bo & Prakash, Atul & Rahmati, Amir & Song, Dawn. (2017). Robust Physical-World Attacks on Machine Learning Models.
7. MIT 6.S094: Deep Learning for Self-Driving Cars Lex Fridman January, https://selfdrivingcars.mit.edu, January 2018