Challenges for Safe and Trustworthy Cyber-Physical Systems

Ezio Bartocci

http://www.eziobartocci.com/

TU WIEN

Faculty of Informatics
Cyber-Physical Group

Cyber-Physical Systems (CPS)

A Major Technology Driver

Amazon drone

Kiva robots

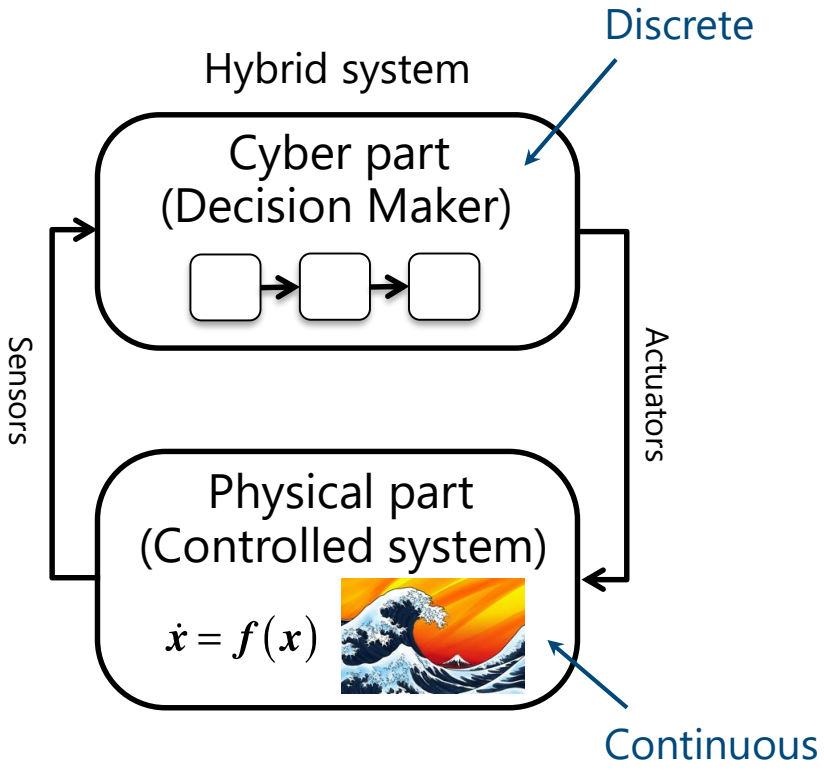
Google self-driving car

Insulin pump

Defibrillator

Cyber-Physical Systems (CPS)

A Major Technology Driver



Amazon drone

Kiva robots

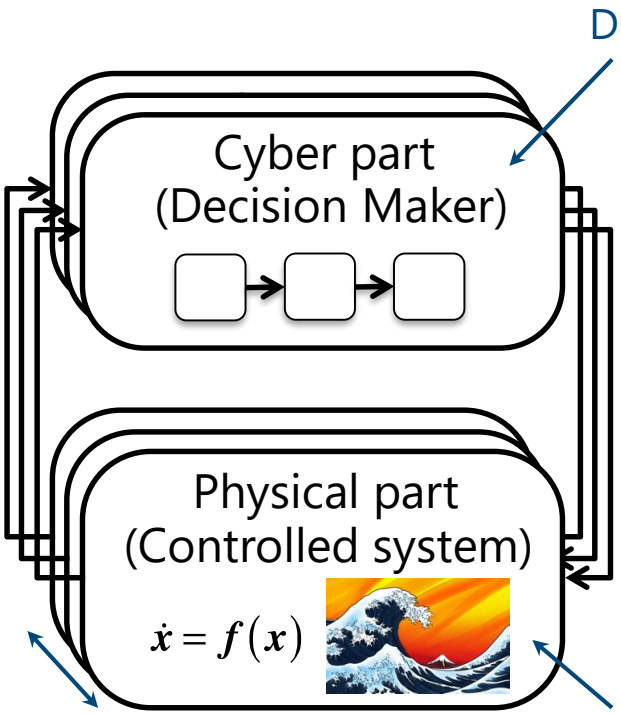
Google self-driving car

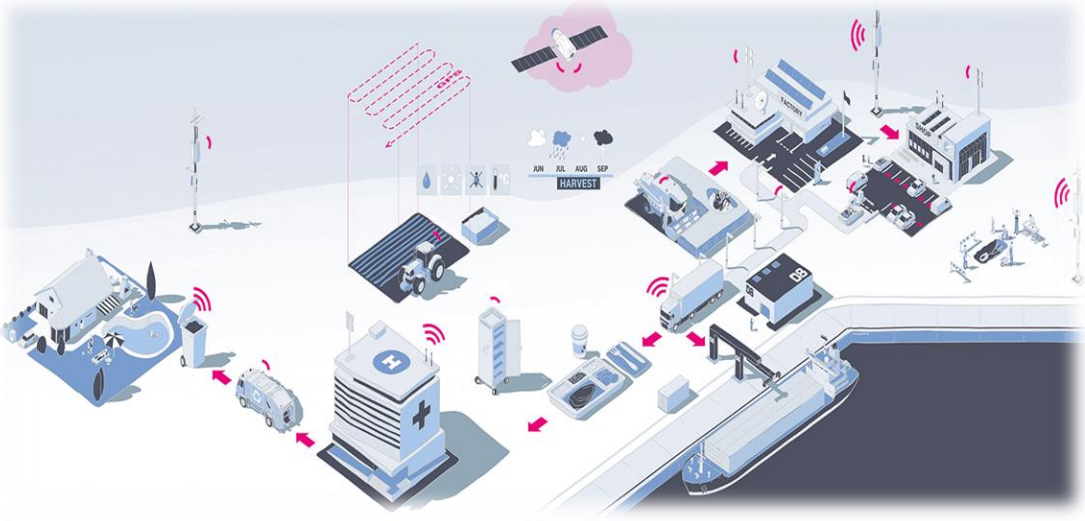
Insulin pump

Defibrillator

Cyber-Physical Systems (CPS)

CPS change the way we interact with the physical world

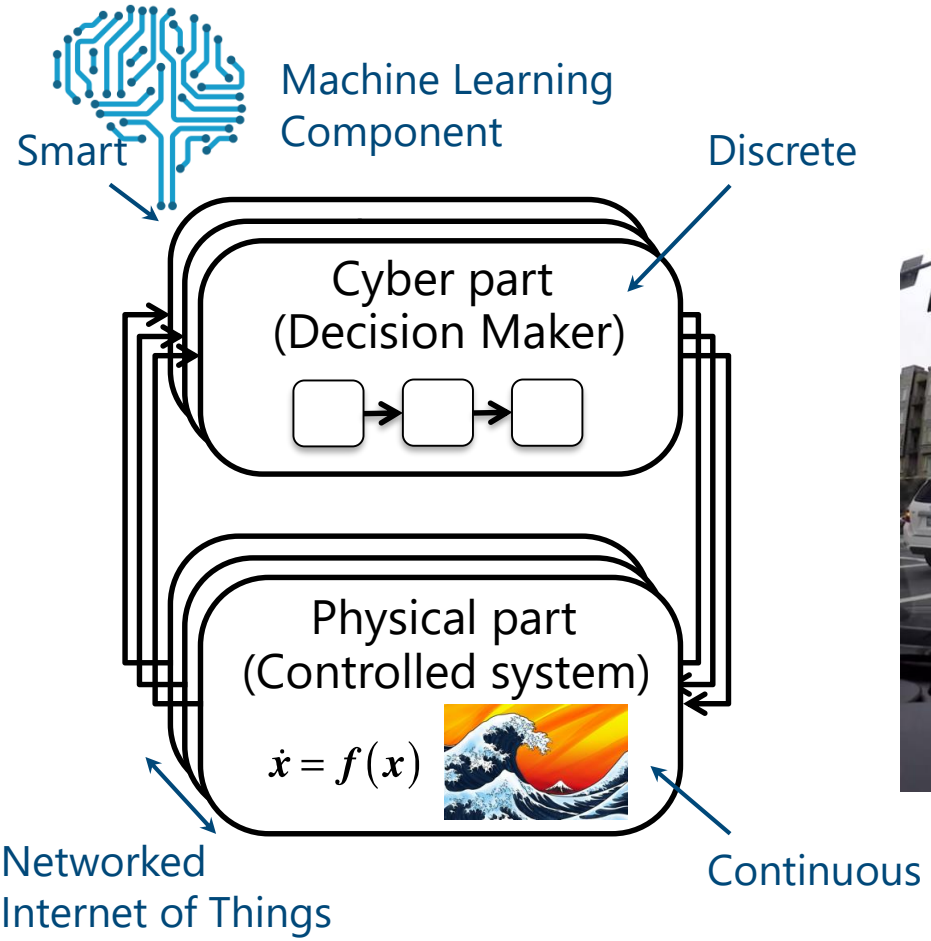




Credits for this picture: http://kayarvizhy.com/

Cyber-Physical Systems (CPS)

CPS change the way we interact with the physical world



Credits for the picture: http://fortune.com/

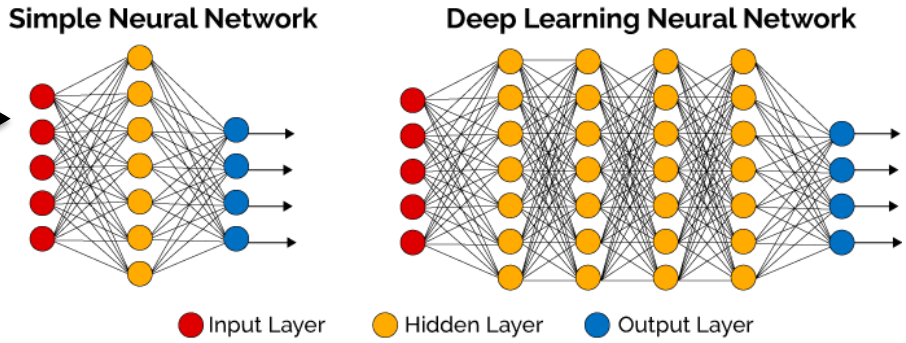
The Rise of Deep Learning and AI in CPS

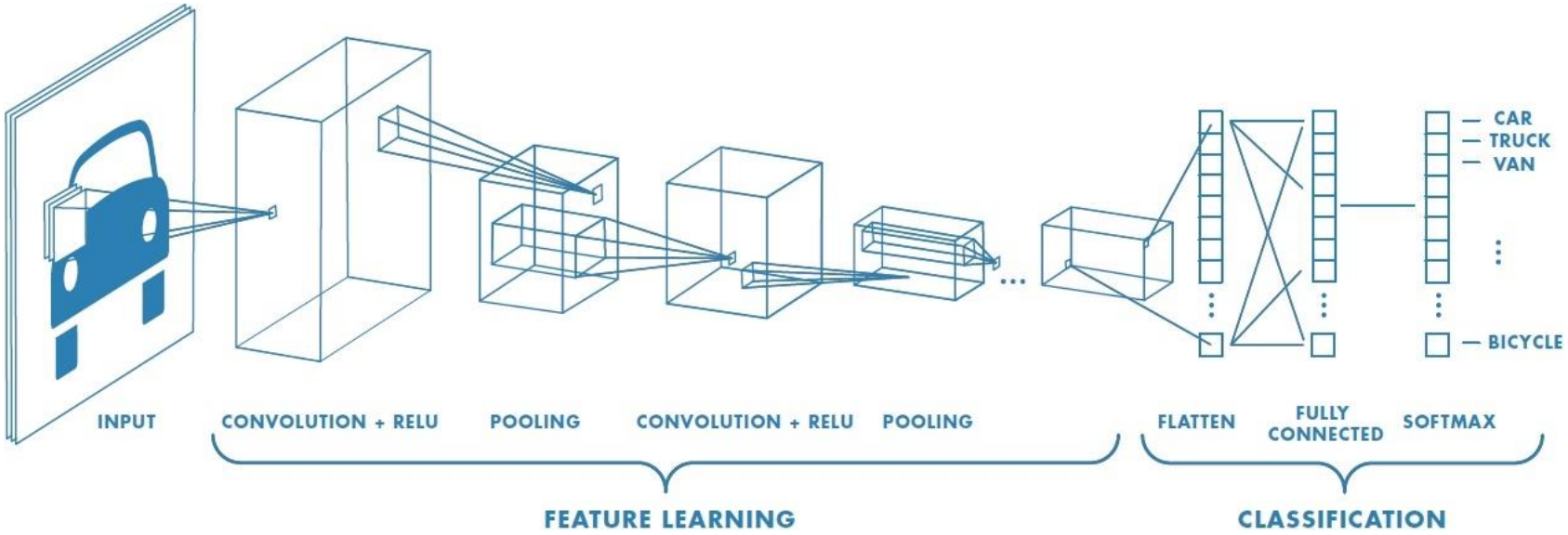
1940s — Perceptron

1980 — **Neocognitron** Fukushima

1998 — **ConvNets** LeCun

2006 — **Deep Learning** Hinton & Salakhutdinov



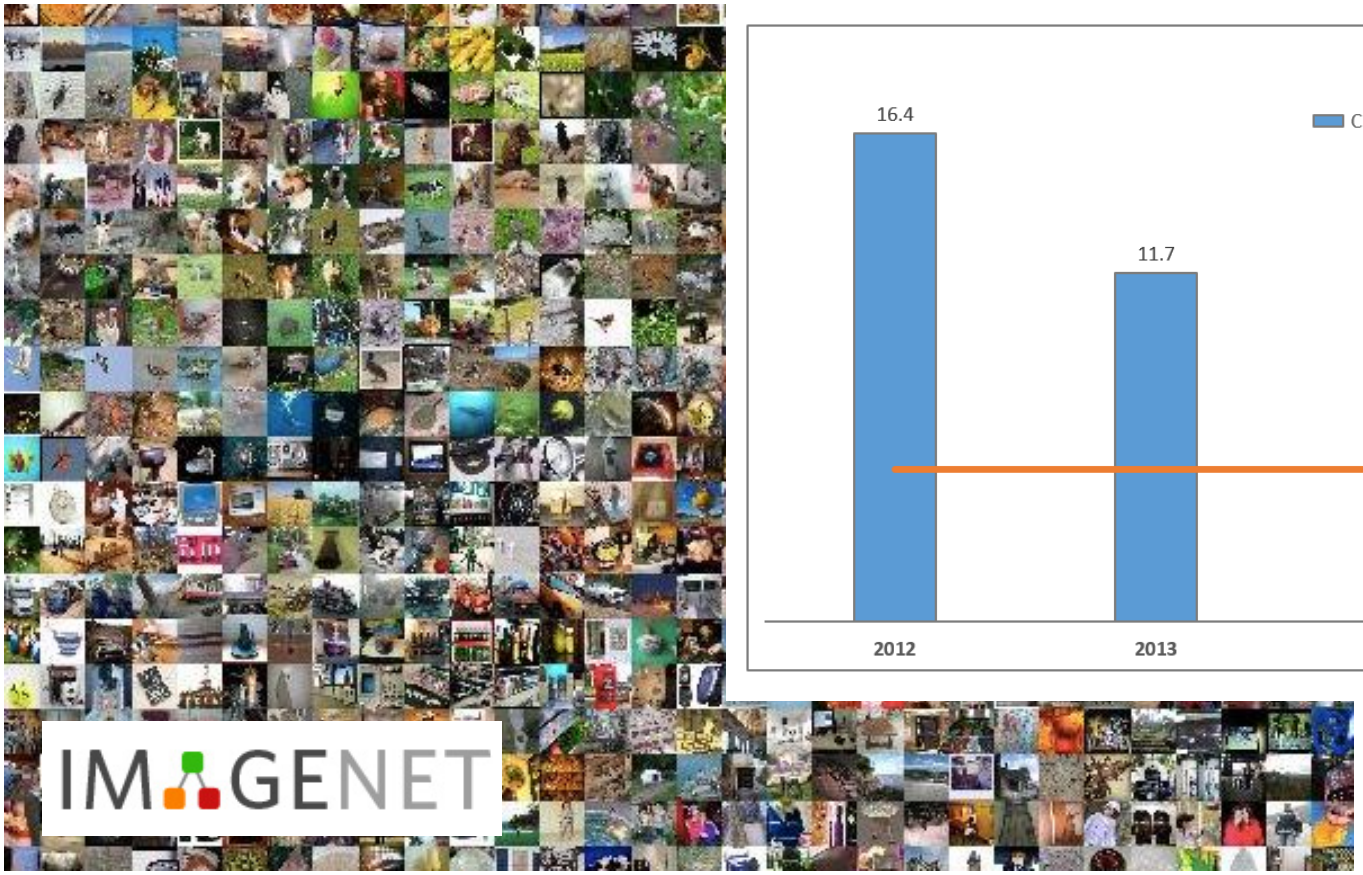


https://de.mathworks.com/discovery/convolutional-neural-network.html

Convolutional Neural Network - Hierarchical multilayered neural network capable of robust visual pattern recognition through learning (inspired by the visual cortex)

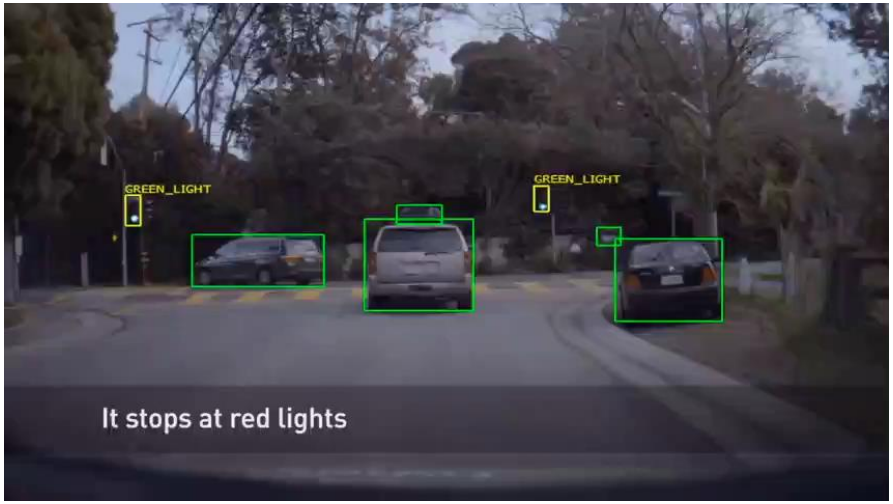
Big Data and Hardware Acceleration

1940s	1980	1998	2006	2009	2015
Perceptron	Neocognitron	ConvNets	Deep Learning	**First GPU-based Learning Algorithms**	**Deep Learning outperform Humans in vision**



The Rise of Deep Learning and AI in CPS

1940s	1980	1998	2006	2009	2015	2016
Perceptron	Neocognitron	ConvNets	Deep Learning	First GPU-based Learning Algorithms	Deep Learning outperform Humans in vision	**AlphaGo** **NVDIA PilotNet** **Tesla Autopilot 8.0**



https://blogs.nvidia.com/blog/2017/04/27/how-nvidias-neural-net-makes-decisions/

Are we safer ?

Man says Tesla Autopilot saved his life by driving him to the hospital

Robert Ferris | @RobertoFerris
Published 3:15 PM ET Fri, 5 Aug 2016
CNBC

Susana Bates | AFP | Getty Images

Tesla Model X is presented during a launch event in Fremont, California last September.

A Missouri man says his Tesla helped saved his life by driving him to the hospital during a life-threatening emergency.

https://www.cnbc.com/2016/08/05/man-says-tesla-autopilot-saved-his-life-by-driving-him-to-the-hospital.html

Europe: 1 fatal crash every 60 Millions of miles

US: 1 fatal crash every 100 Millions of miles

Tesla Autopilot:
1 fatal crash after 130 Millions of miles

Can we fully trust ?

Misclassification can still happens

Tesla driver dies in first fatal crash while using autopilot mode

The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky

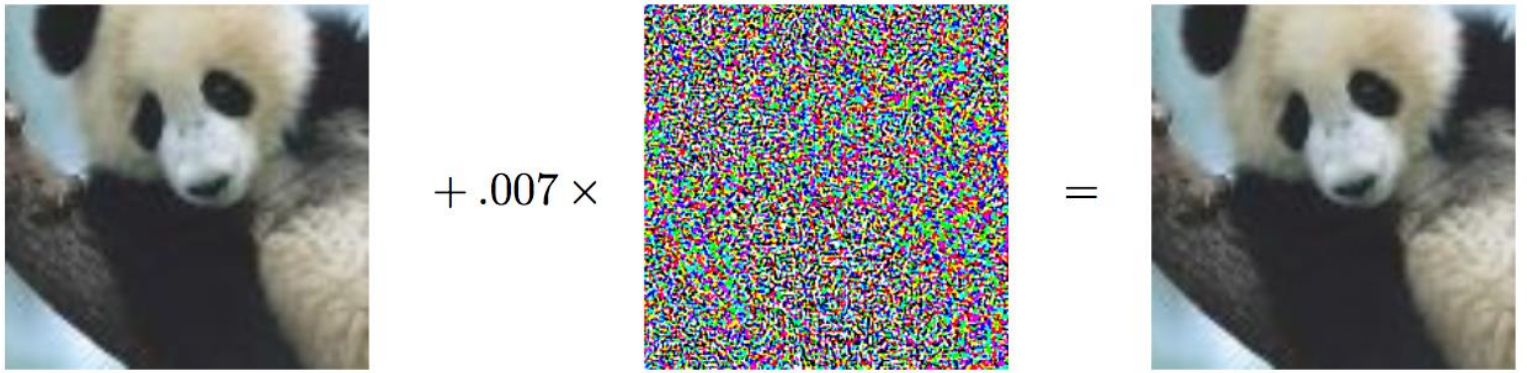
The first known death caused by a self-driving car was disclosed by Tesla Motors on Thursday, a development that is sure to cause consumers to second-guess the trust they put in the booming autonomous vehicle industry.

The 7 May accident occurred in Williston, Florida, after the driver, Joshua Brown, 40, of Ohio put his Model S into Tesla's autopilot mode, which is able to control the car during highway driving.

Against a bright spring sky, the car's sensors system failed to distinguish a large white 18-wheel truck and trailer crossing the highway, Tesla said. The car attempted to drive full speed under the trailer, "with the bottom of the trailer impacting the windshield of the Model S", Tesla said in a blogpost.

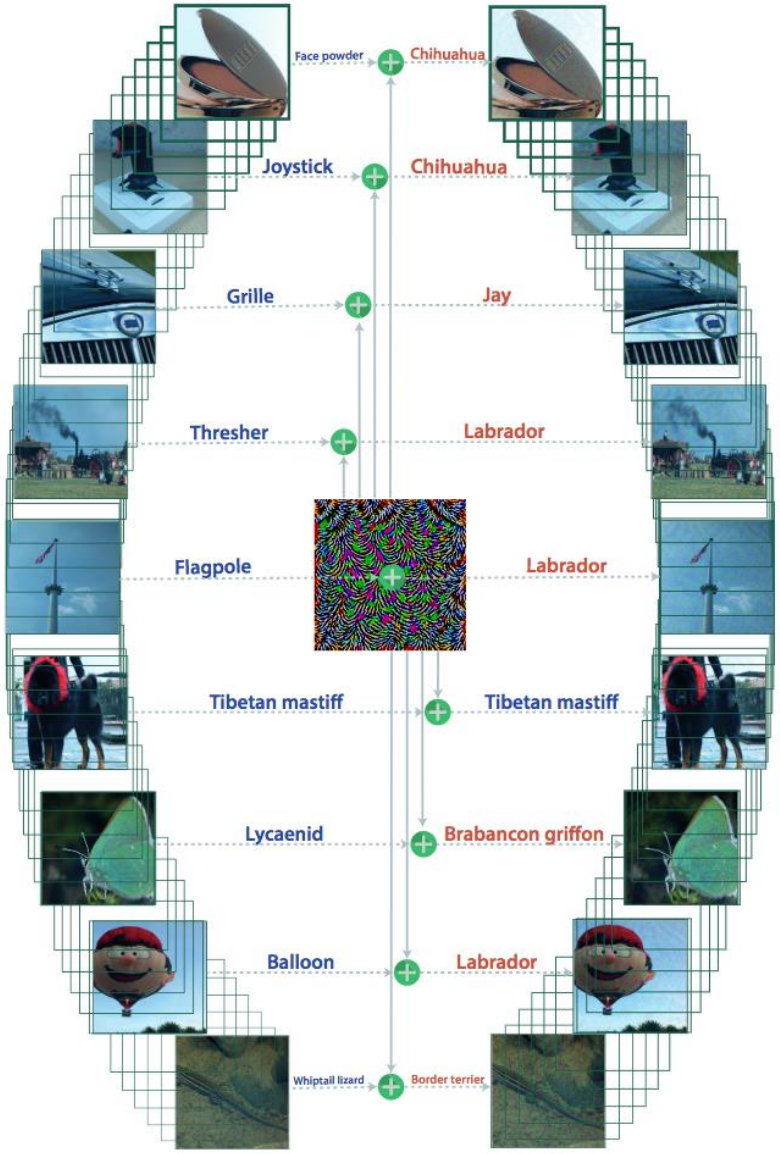
Robustness of classifiers to perturbations

Real-world images undergo perturbations



http://www.srl.ethz.ch/riai2017/Explaining%20and%20Harnessing%20Adversarial%20Examples.pdf

Deep Neural Networks can be fooled



https://arxiv.org/pdf/1610.08401.pdf

- Adversarial perturbations [Szegedy et al. ICLR 2014], [Biggio et al. 2013]
- Random noise [Szegedy at al. 2014]
- **Existence of a universal (image-agnostic) adversarial perturbation** [Moosavi-Dezfooli et al. 2017]

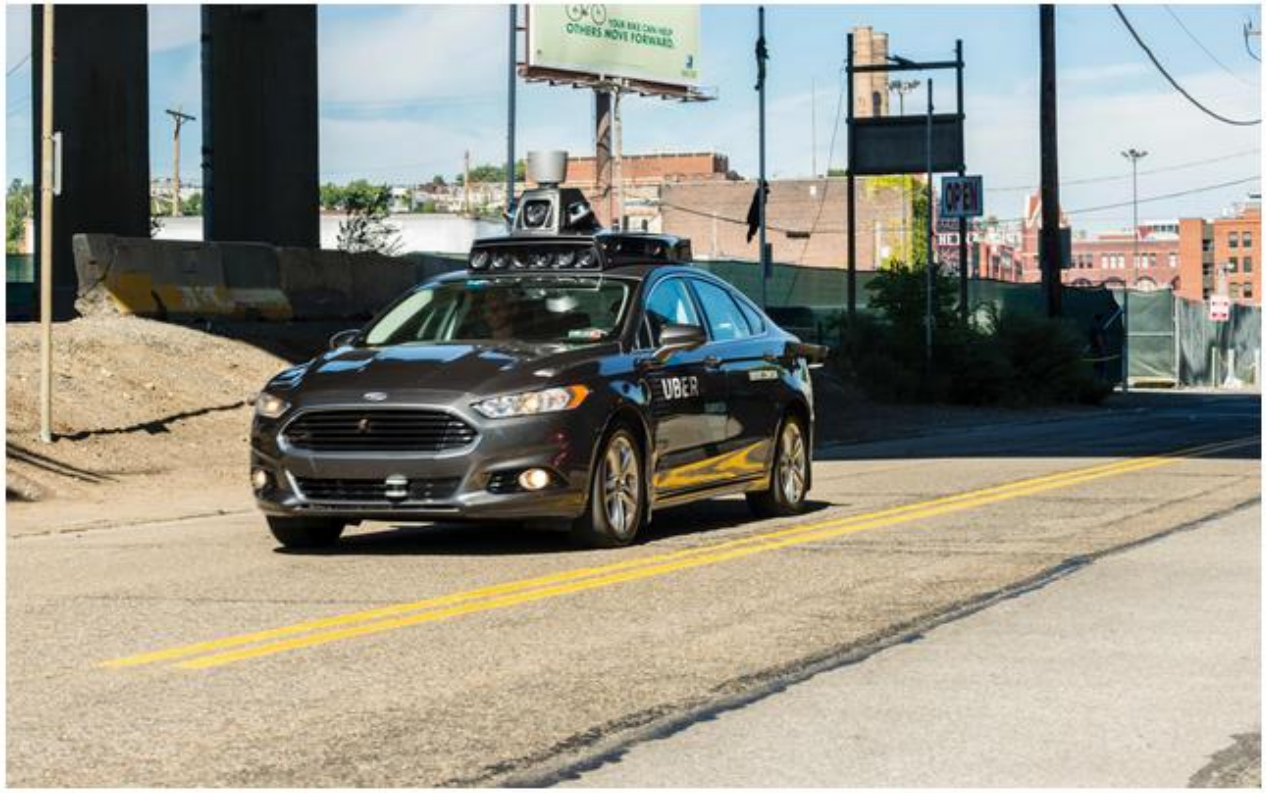
Car Hacking

Technology

Graffiti on stop signs could trick driverless cars into driving dangerously

share

0



http://www.telegraph.co.uk/technology/2017/08/07/graffiti-road-signs-could-trick-driverless-cars-driving-dangerously/

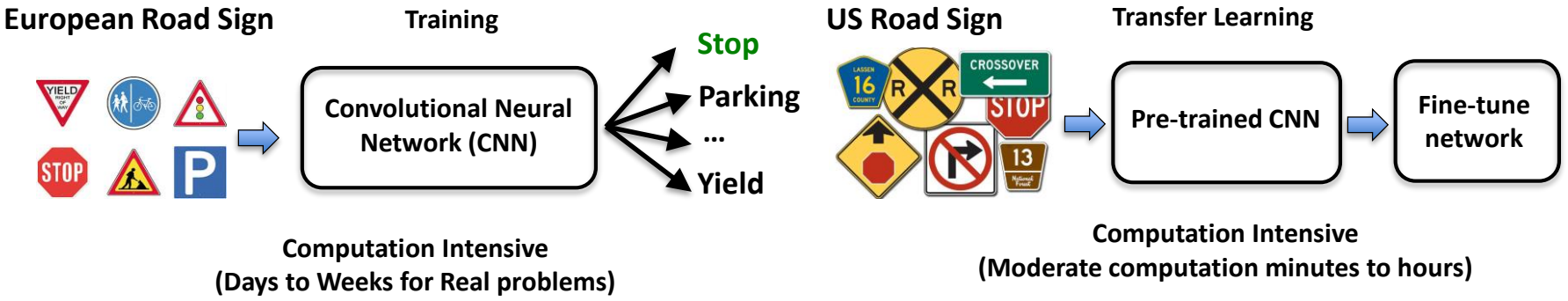
Car Hacking

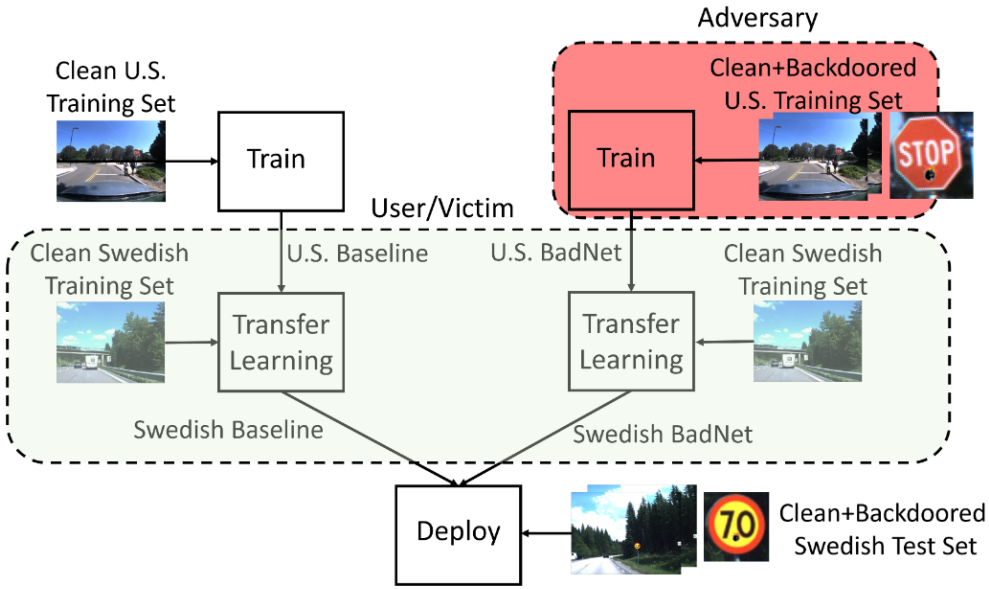
The sign stop is misread as speed limit of 45, by adding a sticker graffiti “Love/Hate”

“…Researchers at the University of Washington demonstrated how car hackers who had gained access to the visual recognition software within the vehicle could create simple alterations to road signs that would cause the car to misread them…”

http://www.telegraph.co.uk/technology/2017/08/07/graffiti-road-signs-could-trick-driverless-cars-driving-dangerously/

Transfer Learning and Backdoors





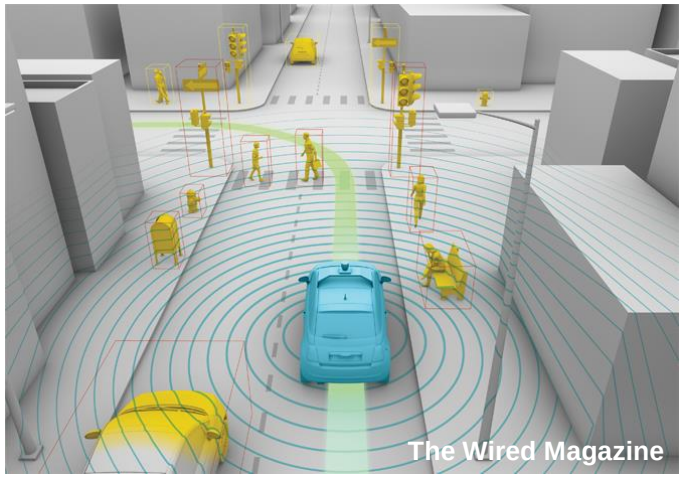
BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

https://arxiv.org/abs/1708.06733

Engineering Safe and Resilient CPS

Exhaustive verification of safety/security properties CPS is intractable:

- **Openness, environmental change**
- **Uncertainty, spatial distribution**
- **ML components can be fooled**
- **Autonomy and machine ethics**
- **Classic state-space explosion problem**



Google Cars

Some of the open challenges:

- **Falsification/formal analysis of CPS with machine learning components**
- **Runtime verification techniques to online detect attacks**