Safety verification for deep neural networks with provable guarantees

Prof. Marta Kwiatkowska

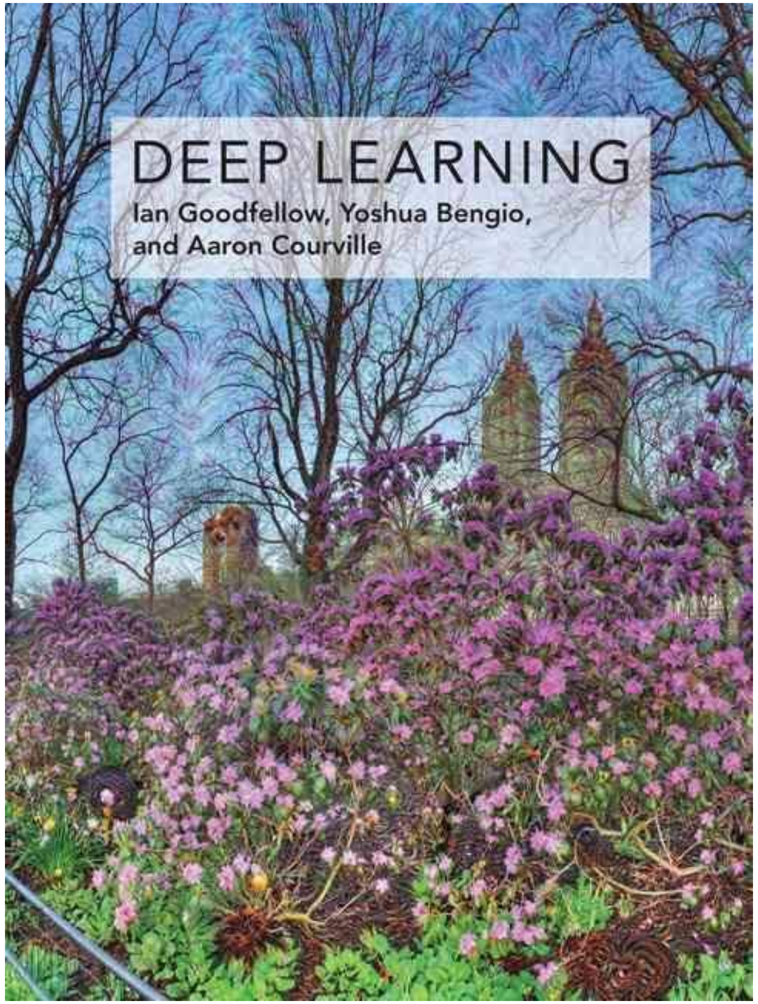
Department of Computer Science
University of Oxford

The unstoppable rise of deep learning

- Neural networks timeline

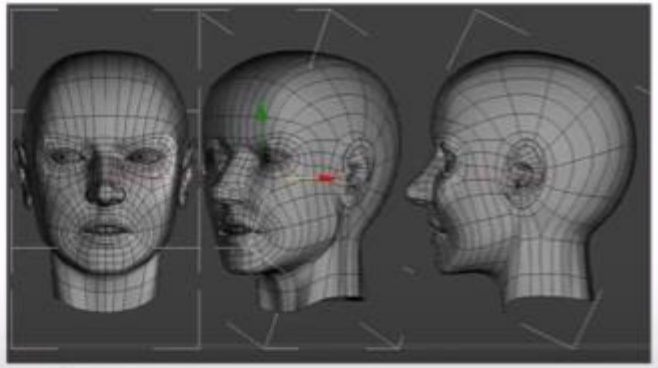
  1940s First proposed
  1998 Convolutional nets
  2006 Deep nets trained
  2011 Rectifier units
  2015 Vision breakthrough
  2016 Win at Go
  2019 Turing Award

- Enabled by
  - Big data
  - Flexible, easy to build models
  - Availability of GPUs
  - Efficient inference



Deep learning with everything

DeepFace
Closing the Gap to Human-Level Performance in Face Verification



Yaniv Taigman
Ming Yang
Marc'Aurelio Ranzato
Lior Wolf
- 2014

97.35% accuracy
Trained on the largest facial dataset – 4M facial images belonging to more than 4,000 identities.

Google Translate—here shown on a mobile phone—will use deep learning to improve its translations between texts.

Deep learning in healthcare

nature International weekly journal of science

Home | News | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 542 > Issue 7639 > Letters > Article > Article metrics > News

Article metrics for:

Dermatologist-level classification of skin cancer with deep l

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Bla

Nature **542**, 115–118 (02 February 2017) | doi:10.1038/nature21056

Last updated: 24 July 2017 10:10:28 EDT

The Stanford University team said the findings were "incredibly exciting" and w
now be tested in clinics.

Eventually, they believe using AI could revolutionise healthcare by turning any
smartphone into a cancer scanner.

Cancer Research UK said it could become a useful tool for doctors.

The AI was repurposed from software developed by Google that had learned
spot the difference **between images of cats and dogs**.

LETTER

https://doi.org/10.1038/s41586-019-1

A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomašev[1]*, Xavier Glorot[1], Jack W. Rae[1,2], Michal Zielinski[1], Harry Askham[1], Andre Saraiva[1], Anne Mottram[1], Clemens Meyer[1], Suman Ravuri[1], Ivan Protsyuk[1], Alistair Connell[1], Cían O. Hughes[1], Alan Karthikesalingam[1], Julien Cornebise[1,12], Hugh Montgomery[3], Geraint Rees[4], Chris Laing[5], Clifton R. Baker[6], Kelly Peterson[7,8], Ruth Reeves[9], Demis Hassabis[1], Dominic King[1], Mustafa Suleyman[1], Trevor Back[1,13], Christopher Nielson[10,11,13], Joseph R. Ledsam[1,13]* & Shakir Mohamed[1,13]

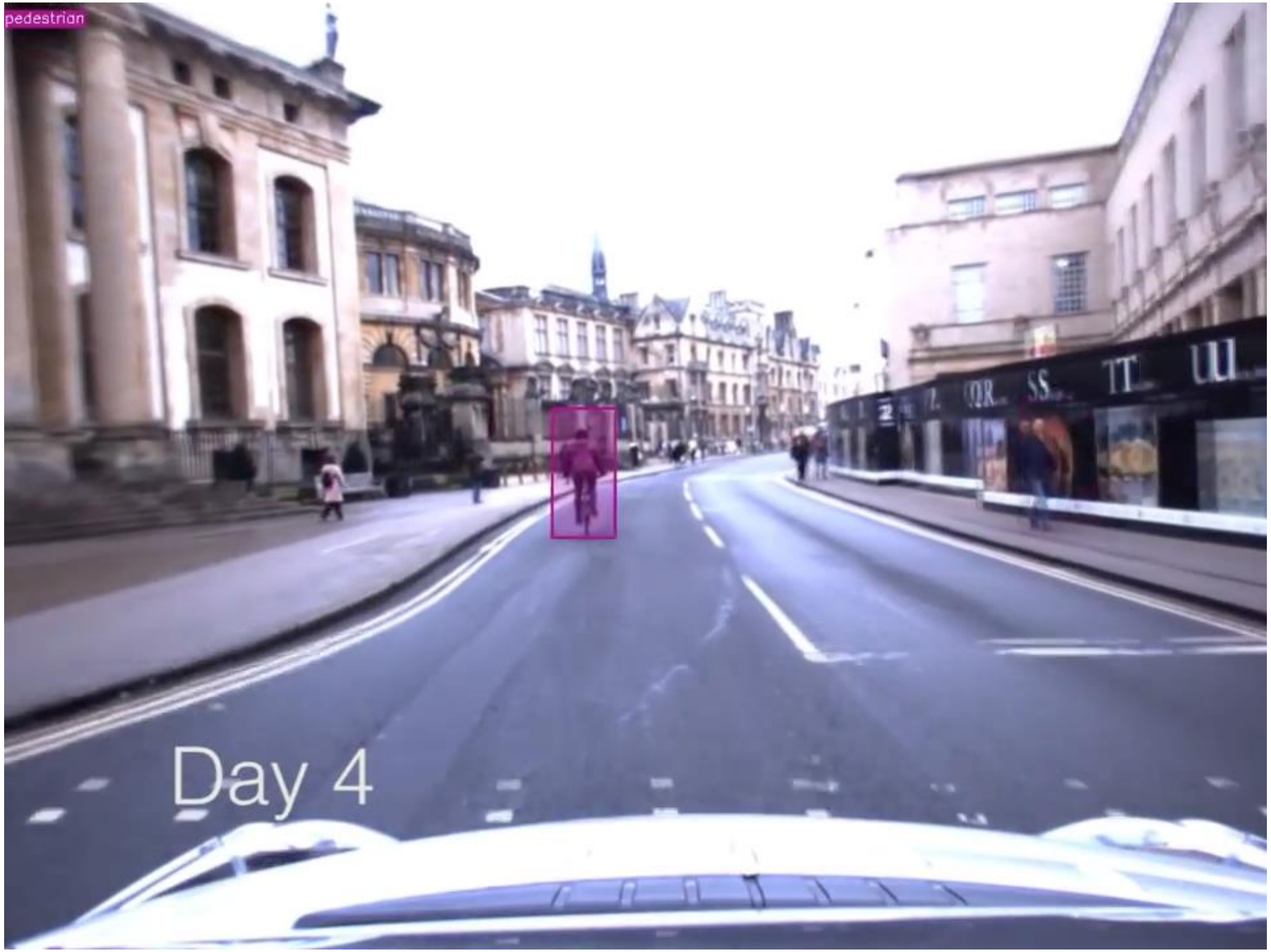
The early prediction of deterioration could have an important role in supporting healthcare professionals, as an estimated 11% of deaths in hospital follow a failure to promptly recognize and treat deteriorating patients[1]. To achieve this goal requires predictions

Promising recent work on modelling adverse events from elec
health records[2–17] suggests that the incorporation of machine lea
may enable the early prediction of AKI. Existing examples of sequ
AKI risk models have either not demonstrated a clinically appl

Much excitement about self-driving

www.bsfilms.me – Black Sheep Films

Self-driving in Oxford….



Would you trust a self-driving car?

We're looking to learn from people with diverse transportation needs. Here are some of the first riders who are already using our self-driving cars every day.

Ted and Candace

A typical day in Ted and Candace's household is full of busy activities across both the parents and their four children: Abbi, Brielle, Izzy and Trey. This lively family is now using our self-driving cars to get to work, shuttle four kids to school and juggle everything from the parents' weekly date night to their children's soccer practice. They are excited about giving everyone in their home a greater sense of freedom and independence.

Waymo early riders, Tesla, Uber, …

In the UK FiveAI, Oxbotica, …

Unwelcome news recently…

Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam

Leer en español

By DAISUKE WAKABAYASHI MARCH 19, 2018

Tesla Says Crashed Vehicle Had Been on Autopilot Before Fatal Accident

By GREGORY SCHMIDT MARCH 31, 2018

RELATED COVERAGE

Tesla Looked Like the Fu
Ask if It Has One. MARC

Fatal Tesla Crash Raises New Questions About Autopilot System

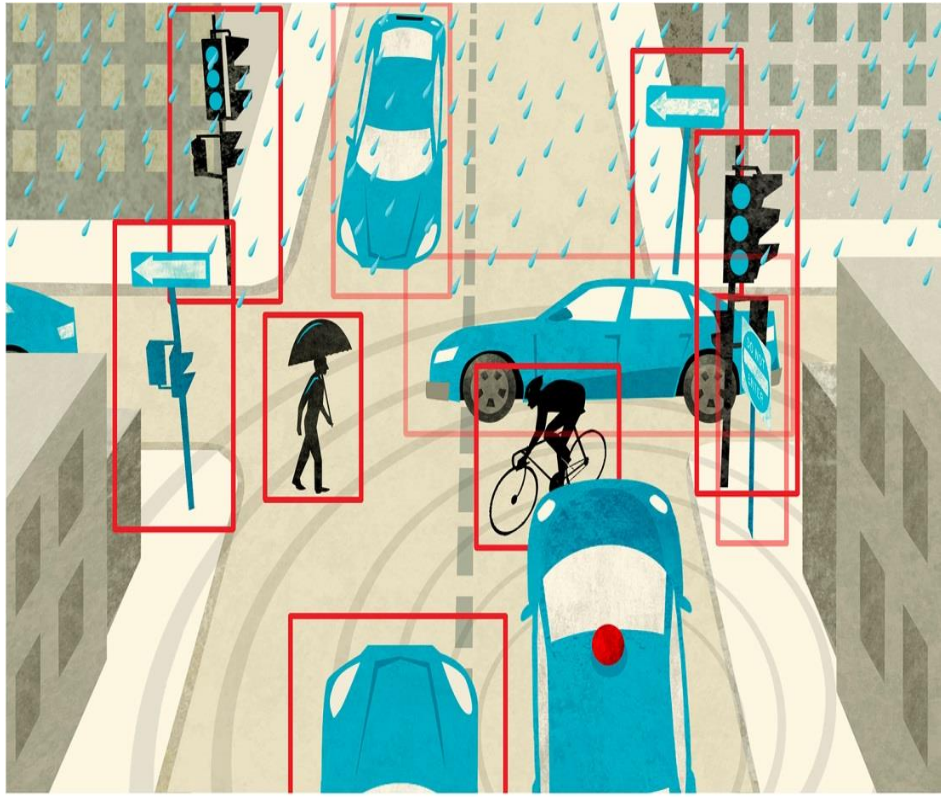
U.S. Safety Agency Criticizes Tesla Crash Data Release

How can this happen if we have 99.9% accuracy?

An AI safety problem…

- Complex scenarios
  - goals
  - perception
  - autonomy
  - situation awareness
  - context (social, regulatory)
  - trust
  - ethics

- Safety-critical, so guarantees needed

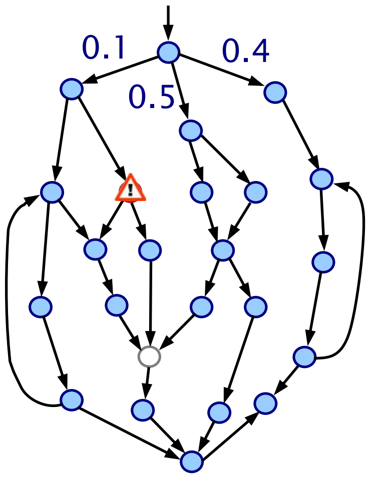


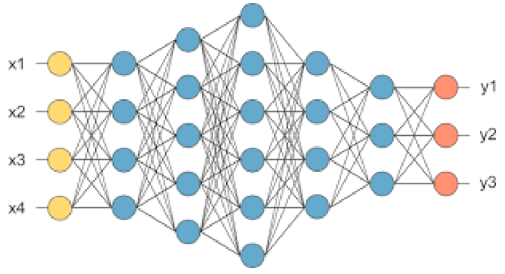
Credit: Anita Dufala/Public source

- Should failure occur, accountability needs to be established

Modelling challenges

- Cyber-physical systems
  - hybrid combination of continuous and discrete dynamics, with stochasticity
  - autonomous control



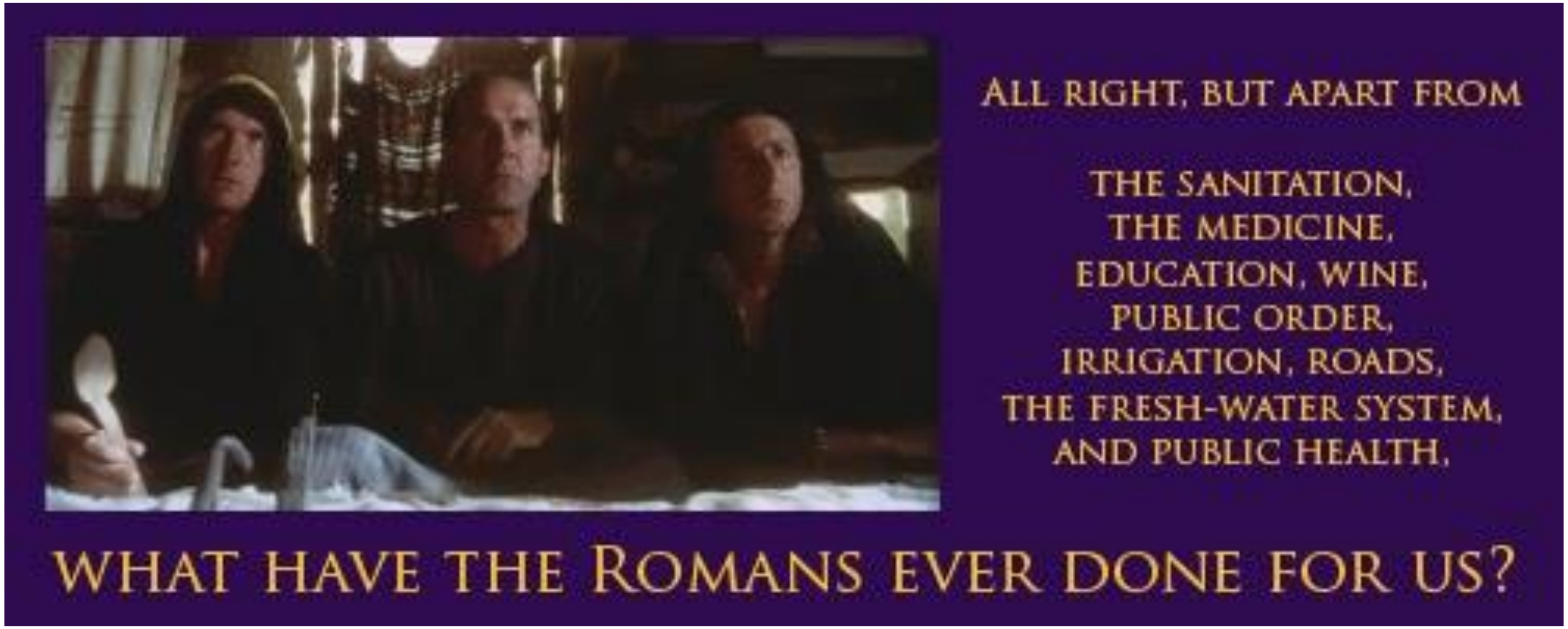


- Data rich, data-enabled models
  - achieved through learning
  - parameter estimation
  - continuous adaptation
- Heterogeneous components, including learning based
  - model-based design
  - automated verification via model checking
  - correct-by-construction model synthesis from specifications

Probabilistic verification and synthesis

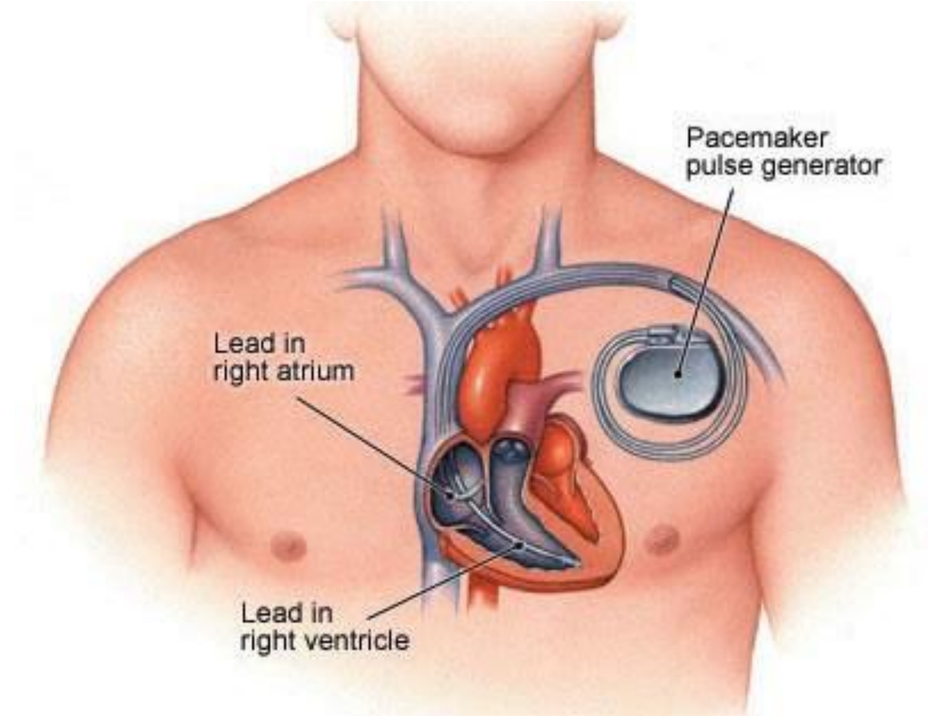
- Stochasticity ever present
  - randomisation, uncertainty, risk
- Need quantitative, probabilistic guarantees for:
  - safety, security, reliability, performance, resource usage, trust, authentication, …
- Examples
  - (reliability) "the probability of the car crashing in the next hour is less than 0.001"
  - (energy) "energy usage is below 2000 mA per minute"
- My focus is on automated, tool-supported methodologies
  - probabilistic model checker PRISM, www.prismmodelchecker.org
  - HVC 2016 Award (joint with Dave Parker and Gethin Norman)
- Applied to a wide range of systems…

OK, but what is probabilistic verification good for?



Case study: Cardiac pacemaker

- How it works
  - reads electrical signals through sensors in the right atrium and right ventricle
  - monitors the timing of heart beats and local electrical activity
  - generates artificial pacing signal as necessary

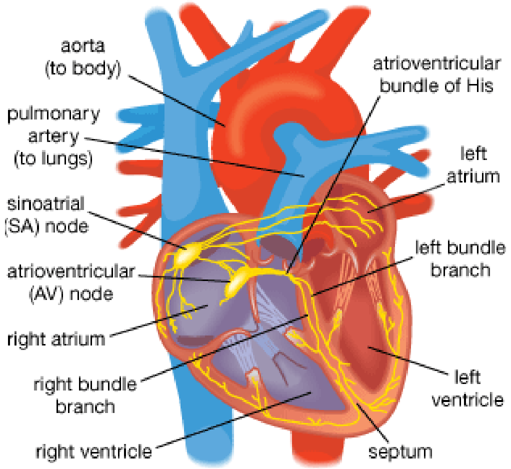


- Safety-critical real-time system!
- The guarantee
  - (basic safety) maintain 60-100 beats per minute

– Killed by code: FDA recalls 23 defective pacemaker devices because of adverse health consequences or death, six likely caused by software defects (2010)

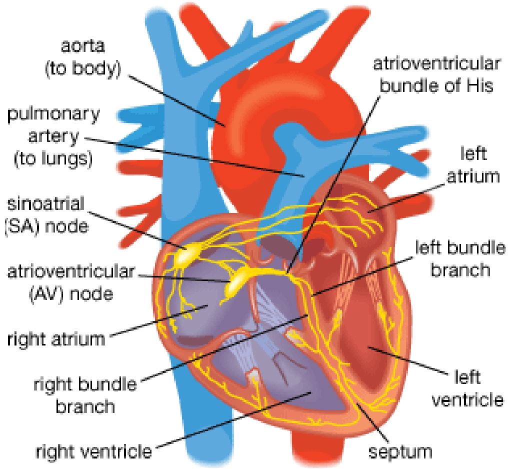
Modelling framework

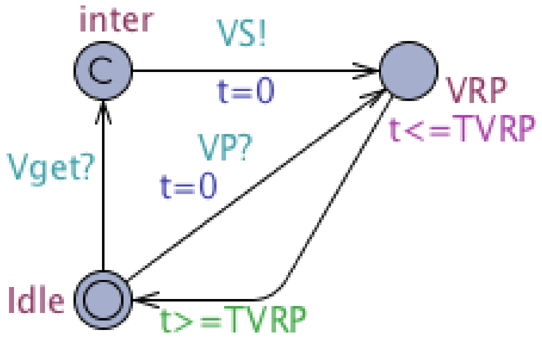
Model the pacemaker and the heart, compose and verify



Quantitative verification of implantable cardiac pacemakers over hybrid heart models. Chen *et al*, Information and Computation 2014

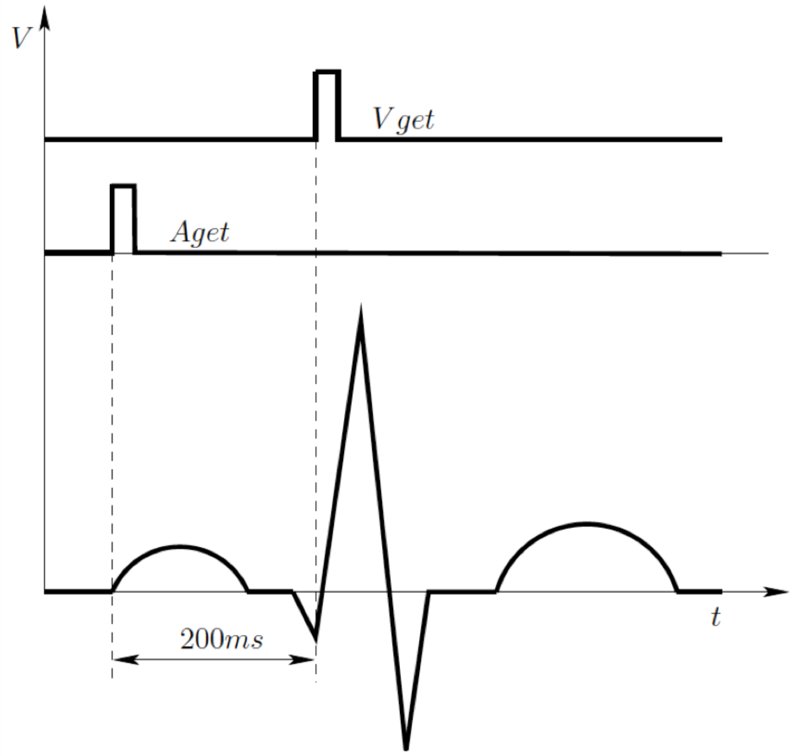
Modelling framework

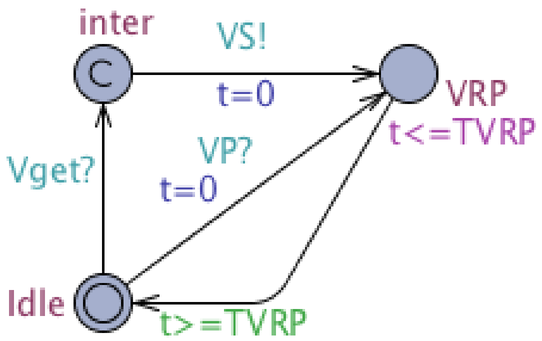





```
module VRP

s_vrp:[0..2] init 0;
t_vrp : clock;

// Invariants for clock t_vrp
  invariant
       (s_vrp = 2 => (t_vrp <= TVRP)) &
       (s_vrp = 1 => (t_vrp <= 0 ))
  endinvariant

[Vget] (s_vrp = 0) -> (s_vrp' = 1) & (t_vrp'=0);
[VP]  (s_vrp = 0) -> (s_vrp' = 2) & (t_vrp' = 0);
```

Modelling framework

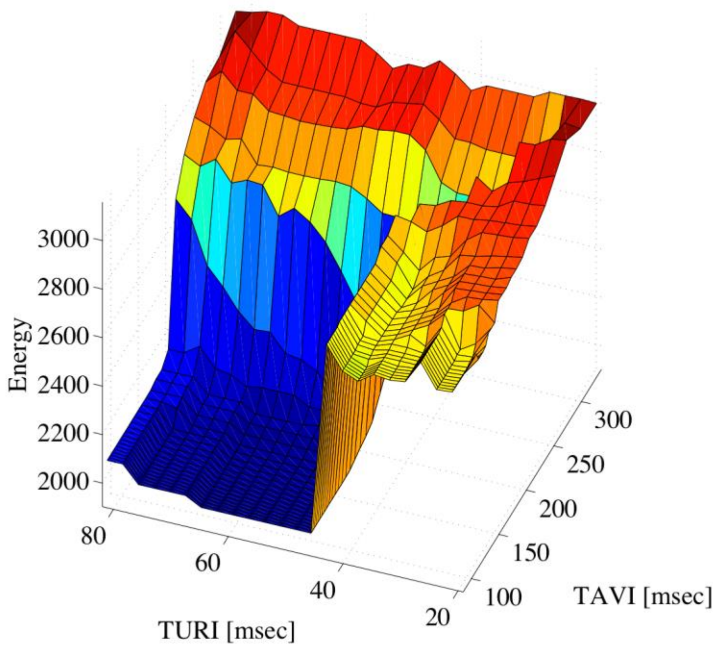





```
module VRP

s_vrp:[0..2] init 0;
t_vrp : clock;

// Invariants for clock t_vrp
  invariant
       (s_vrp = 2 => (t_vrp <= TVRP)) &
       (s_vrp = 1 => (t_vrp <= 0 ))
  endinvariant

[Vget] (s_vrp = 0) -> (s_vrp' = 1) & (t_vrp'=0);
[VP]  (s_vrp = 0) -> (s_vrp' = 2) & (t_vrp' = 0);
```

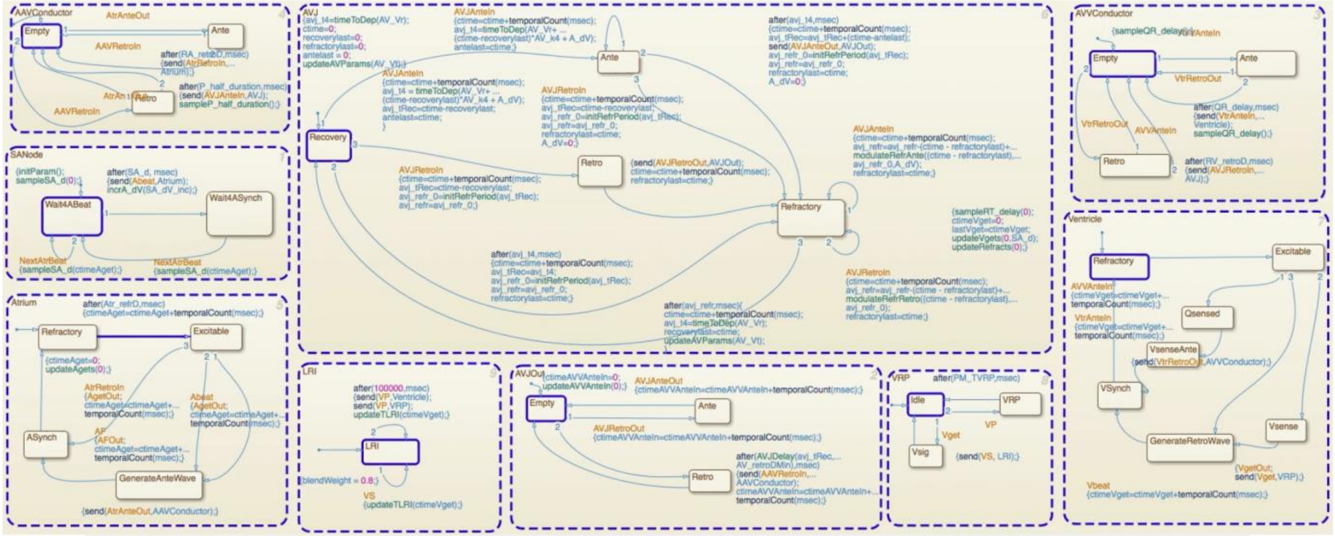
Pacemaker verification

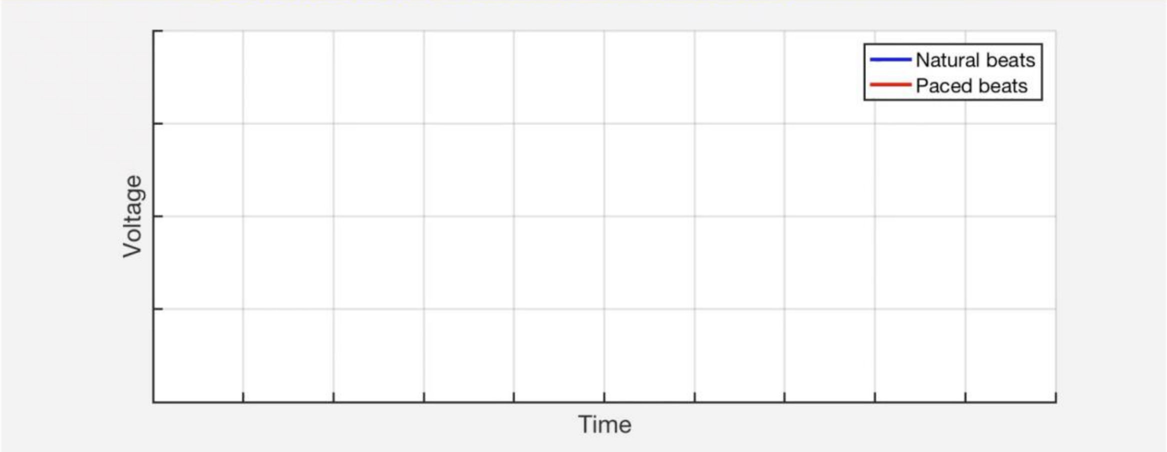
- Basic guarantees
  - (basic safety) maintain 60-100 beats per minute
  - (energy usage) detailed analysis, plotted against timing parameters of the pacemaker
- Advanced guarantees
  - rate-adaptive pacemaker, for patients with chronotropic deficiency
  - (advanced safety) adapt the rate to exercise and stress levels
  - in silico testing



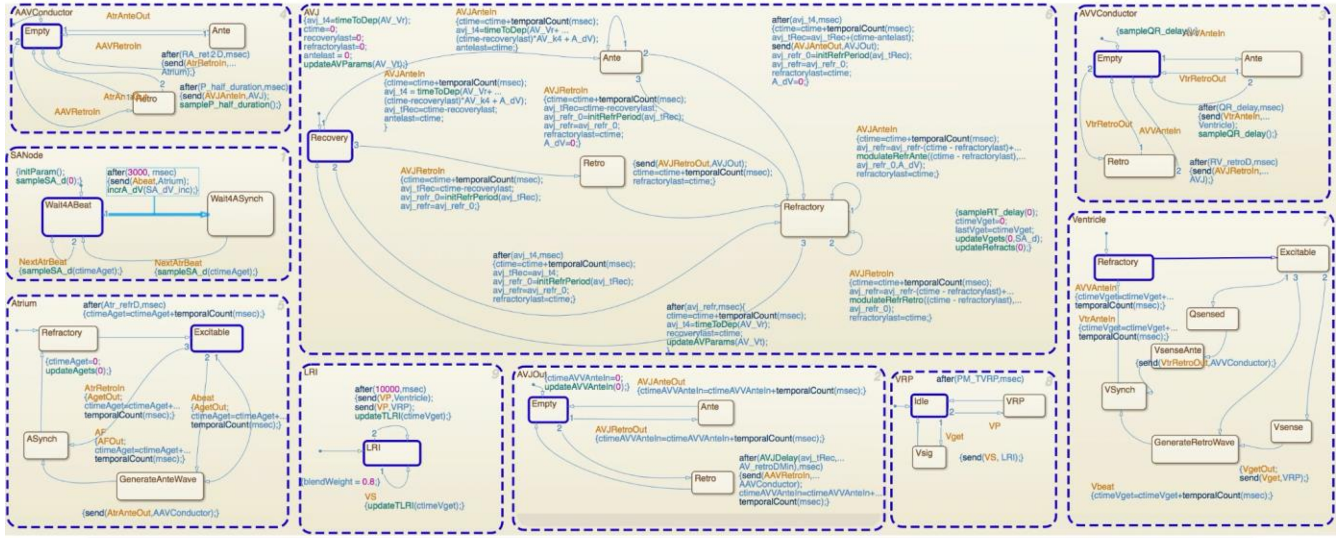
Closed-Loop Quantitative Verification of Rate-Adaptive Pacemakers. Paoletti *et al*, ACM Transactions on Cyber-Physical Systems 2018

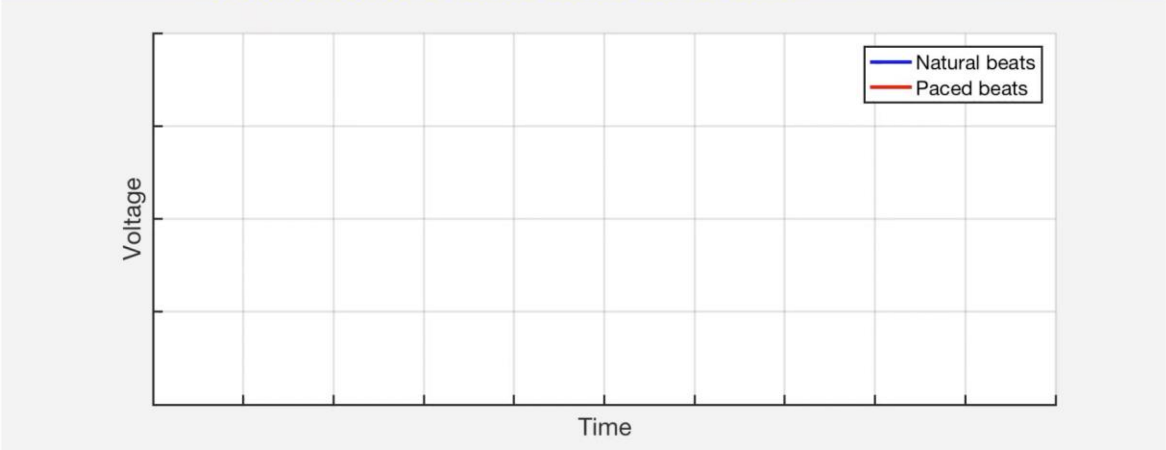
Synthetic ECG: healthy heart



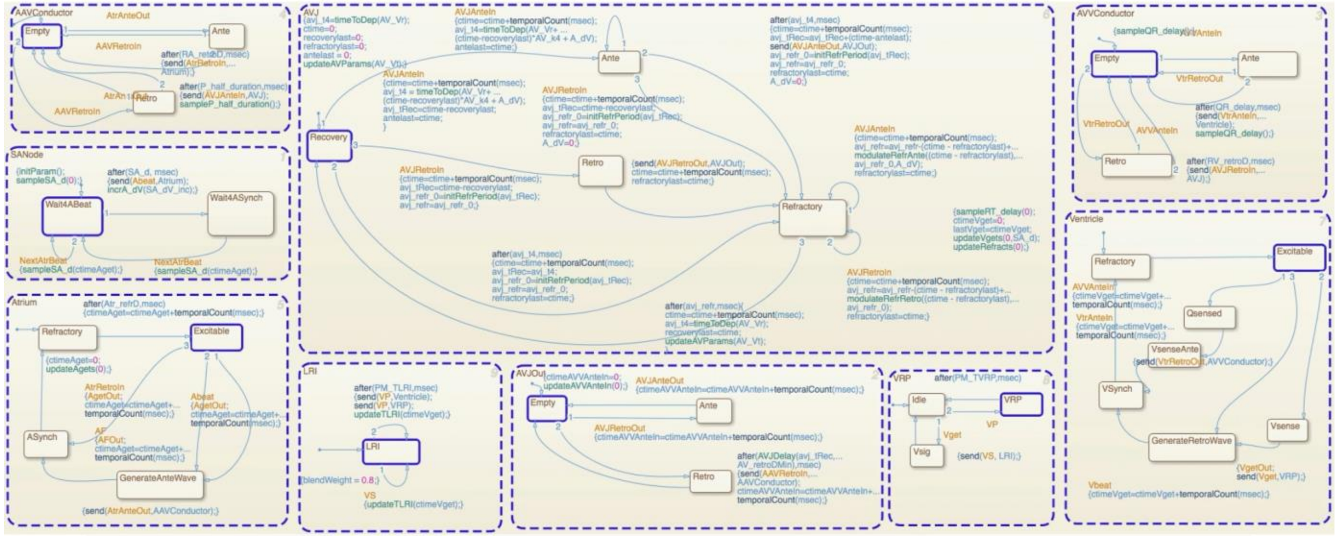


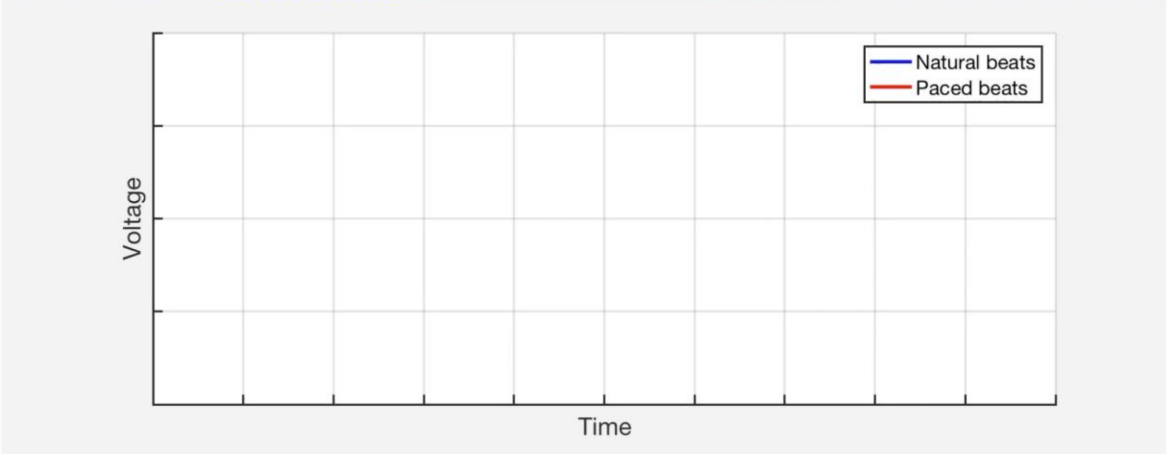
Bradycardia (slow heart rate)





Bradycardia heart, paced



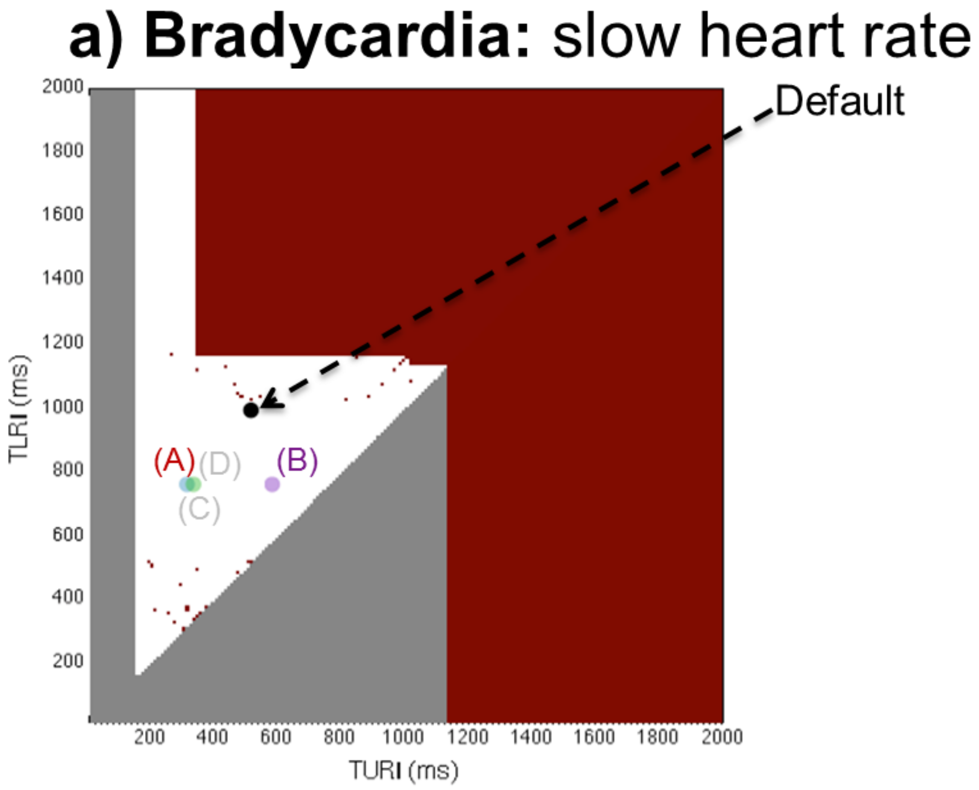


Parameter synthesis for pacemakers

- Can we adapt the pacing rate to patient's ECG to
  - minimise energy usage?
  - maximise cardiac output?
  - explore trade offs?

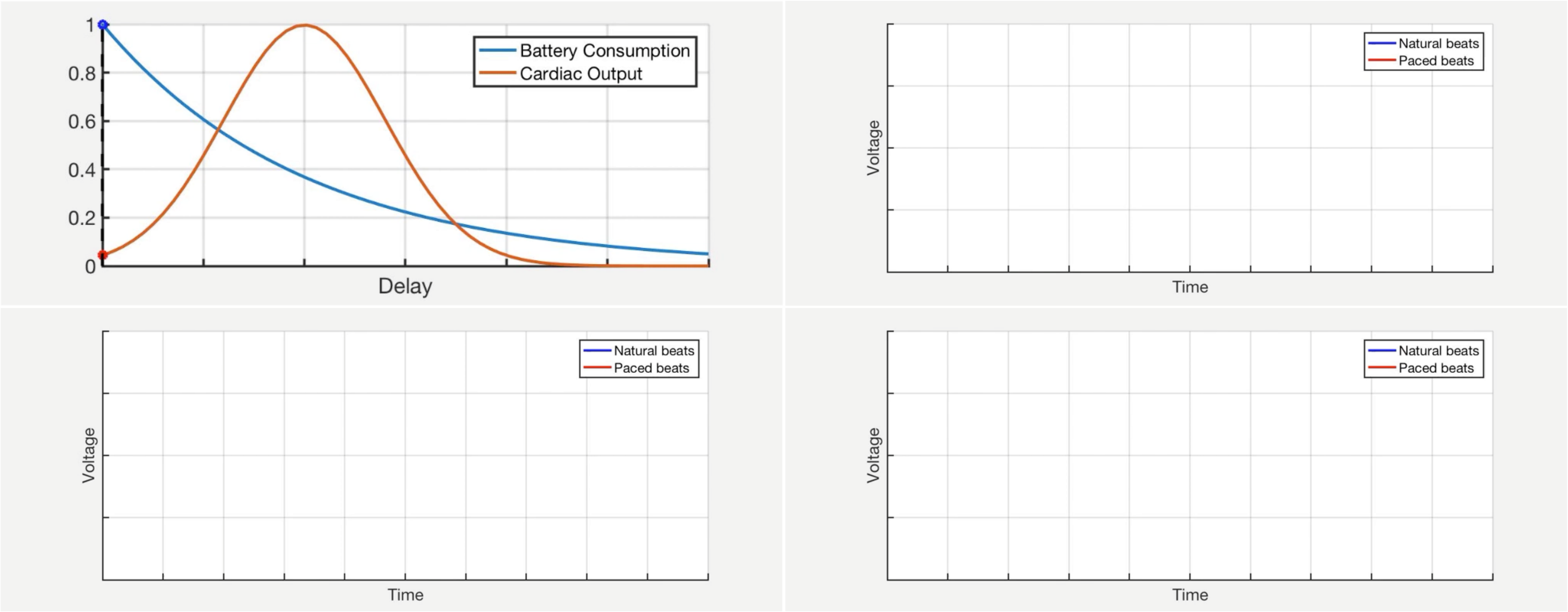
- The guarantee
  - (optimal timing delay synthesis): find values for timing delays that optimise a given objective, adapted to patient's ECG

- Significant improvement over default values



Synthesising robust and optimal parameters for cardiac pacemakers using symbolic and evolutionary computation techniques. Kwiatkowska *et al*, HSB'16

Trade offs in optimal delay synthesis



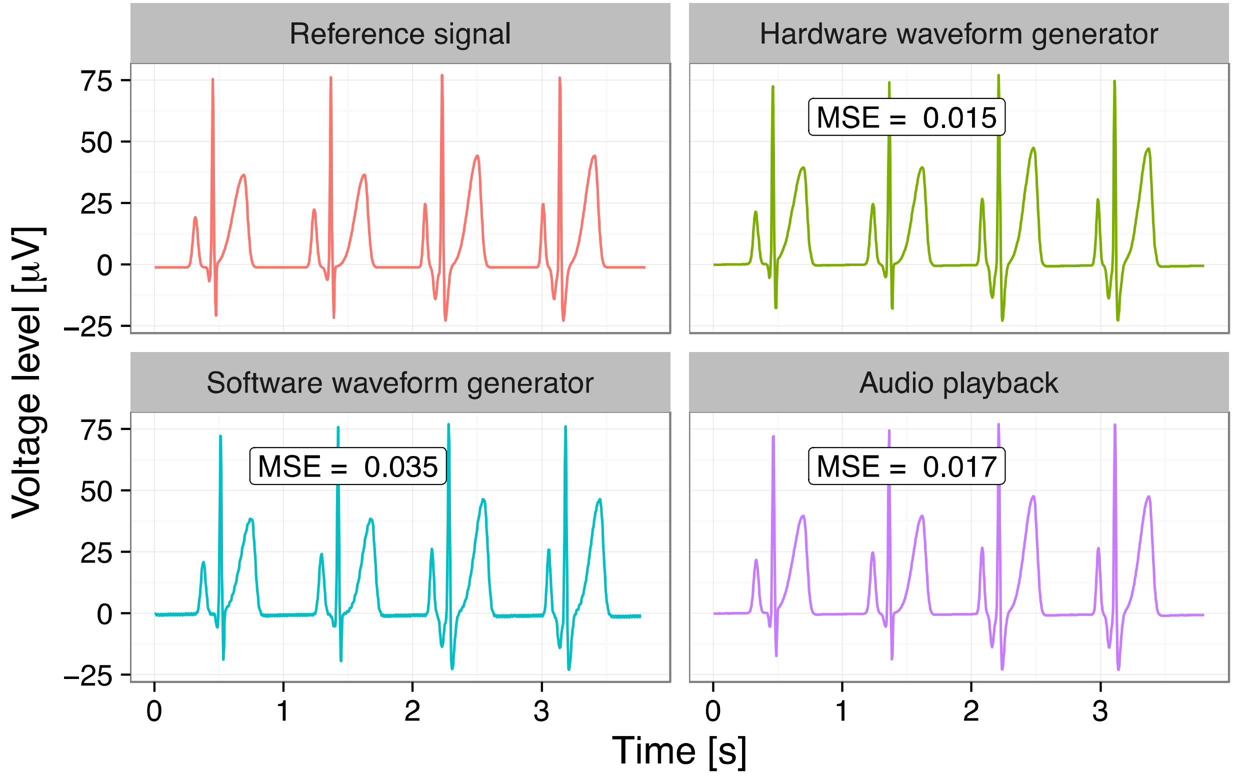
Case study: ECG biometrics

- Biometrics increasing in popularity
  - are they secure?
- Nymi band
  - ECG used as a biometric identifier
  - biometric template created first
  - compared with real ECG signal
- Proposed uses
  - for access into buildings and restricted spaces
  - for payment
  - etc

Broken Hearted: How to Attack ECG Biometrics, Ebertz et al., In *Proc* NDSS 2017

Attack on ECG biometrics

- We use synthetic ECGs to impersonate a user
  - build model from data, 41 volunteers
  - inject synthetic signals to break authentication
  - 80% success rate

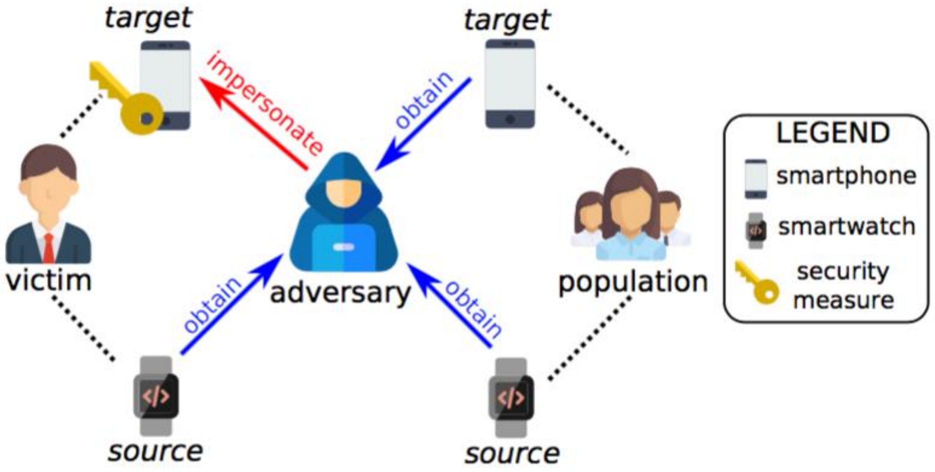


- Results
  - serious weakness
  - countermeasures needed

- Modelling essential, good for attacks…

Case study: Transferability of attack

- Beware your fitness tracker!
- How easy it is to predict attacks when collecting data from different sources
  - ECG
  - eye movements
  - mouse movements
  - touchscreen dynamics
  - gait
  - etc



- Human study
  - easy for eye movements
  - ECG more chaotic

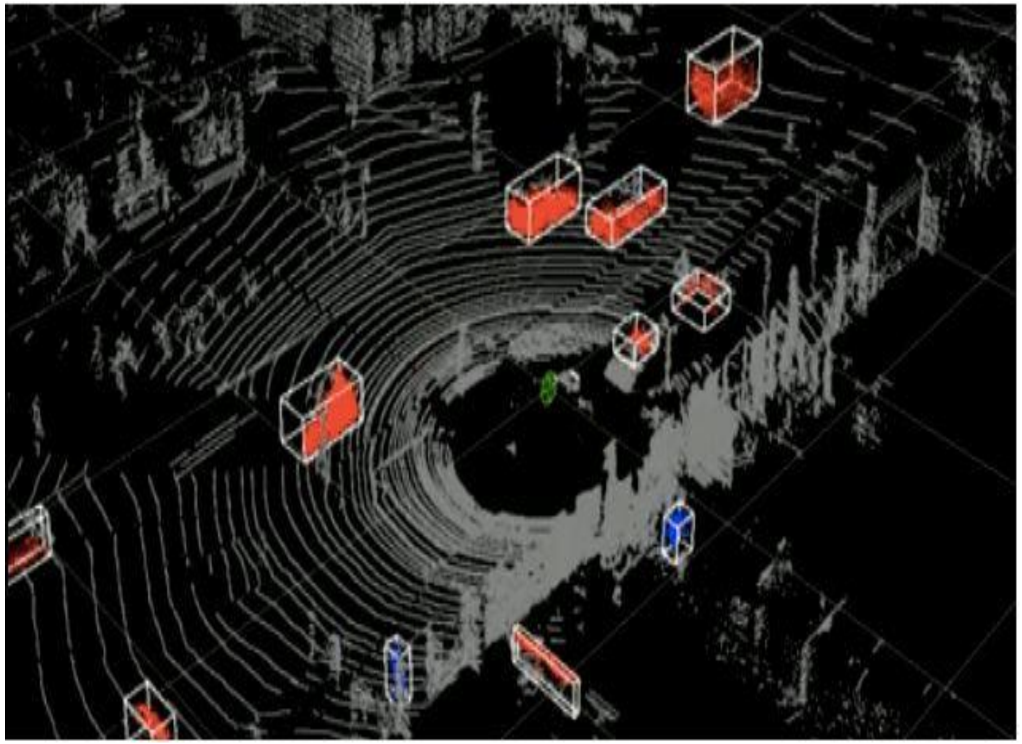
When your fitness tracker betrays you, Ebertz et al., In *Proc* S&P 2018

Back to the challenge of autonomous driving…

- Things that can go wrong in perception software
  - sensor failure
  - object detection failure

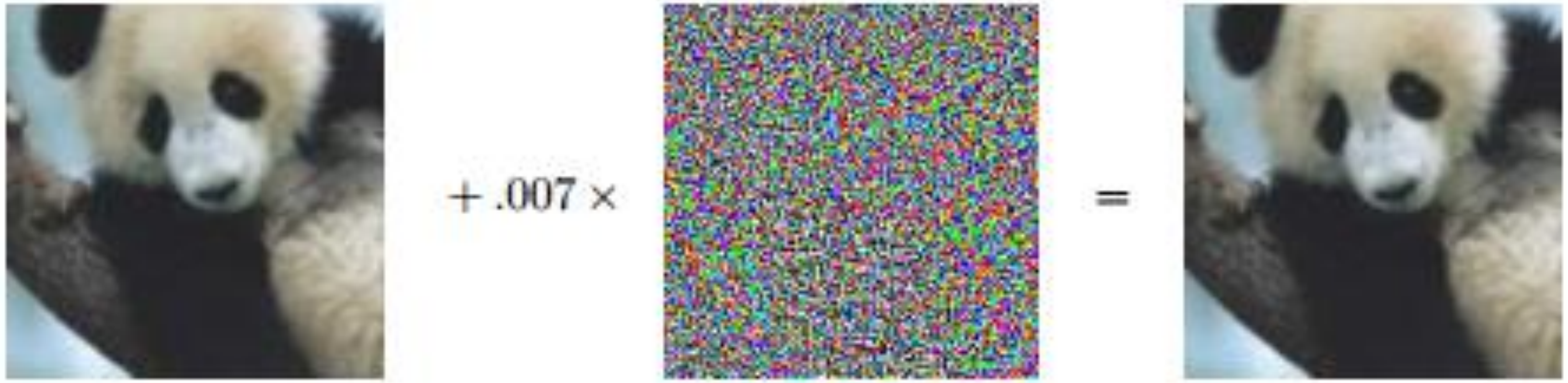
- Machine learning software
  - not clear how it works
  - does not offer guarantees

- Yet safety-critical applications



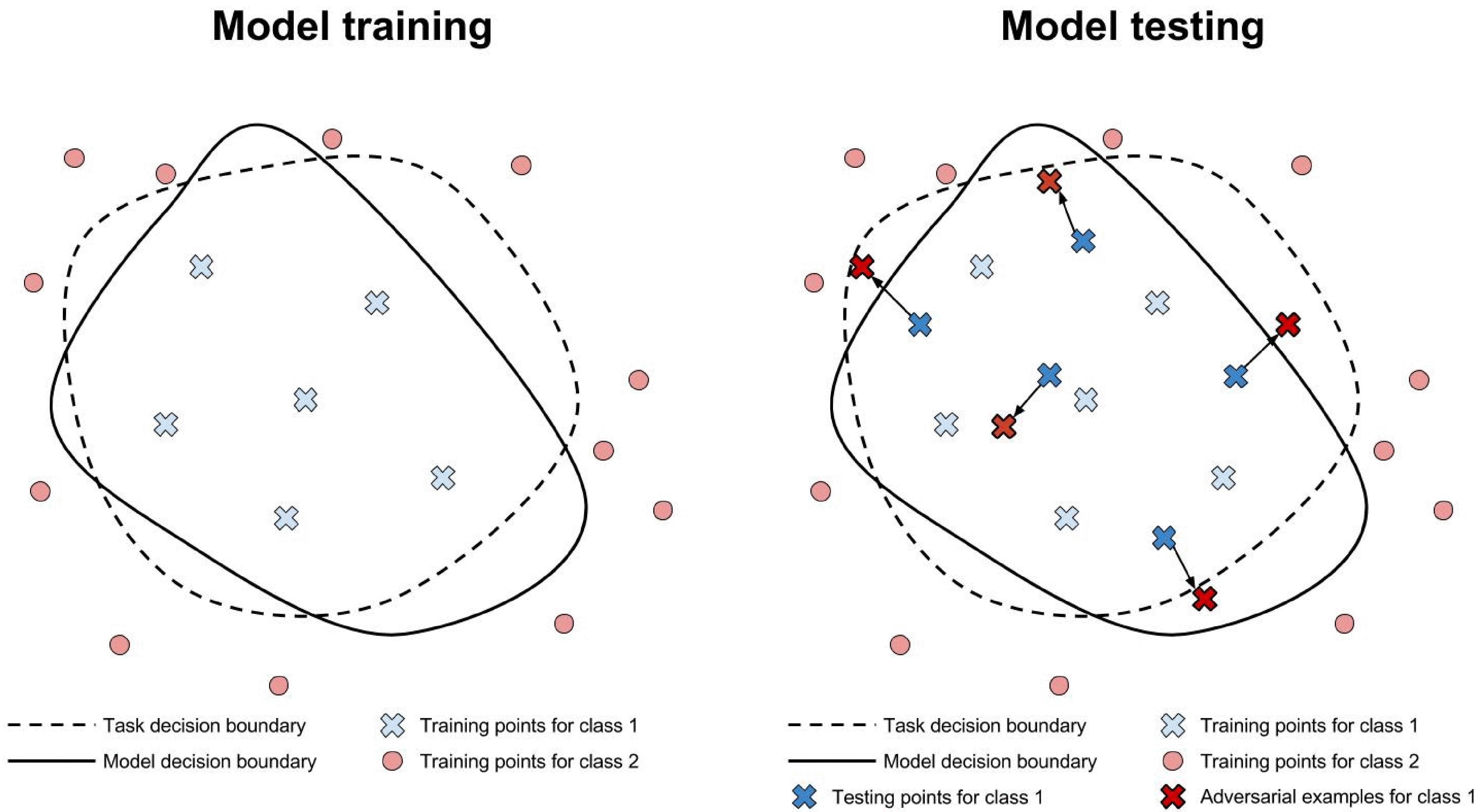
Lidar image, Credit: Oxford Robotics Institute

Deep neural networks can be fooled!



- They are unstable wrt adversarial perturbations
  - often imperceptible changes to the image [Szegedy et al 2014, Biggio et al 2013 …]
  - sometimes artificial white noise
  - practical attacks, potential security risk
  - transferable between different architectures
  - not just image classification: also images segmentation, pose recognition, sentiment analysis…

Training vs testing



Should we worry about safety of self-driving?

(a) (b) (c)

- Deep neural networks are unstable wrt adversarial perturbations
  - Nexar Traffic Light Challenge: red light classified as green with 68%/95%/78% confidence after one pixel change

Feature-Guided Black-Box Safety Testing of Deep Neural Networks. Wicker *et al*, In Proc. TACAS, 2018.

German traffic sign benchmark…

stop	30m speed limit	80m speed limit	30m speed limit	go right	go straight

Confidence 0.999964 0.99

[Safety Verification of Deep Neural Networks.](#) Huang *et al*, In Proc. CAV, 2017.

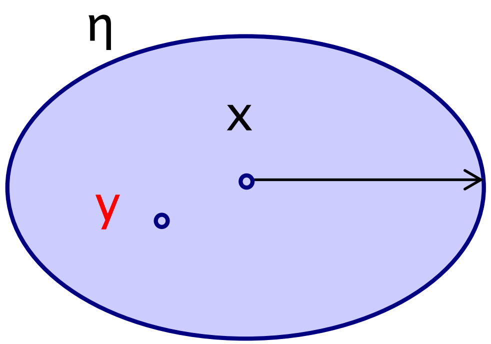
Aren’t these artificial?

Real traffic signs in Alaska!

Need to consider physical attacks, not only digital…

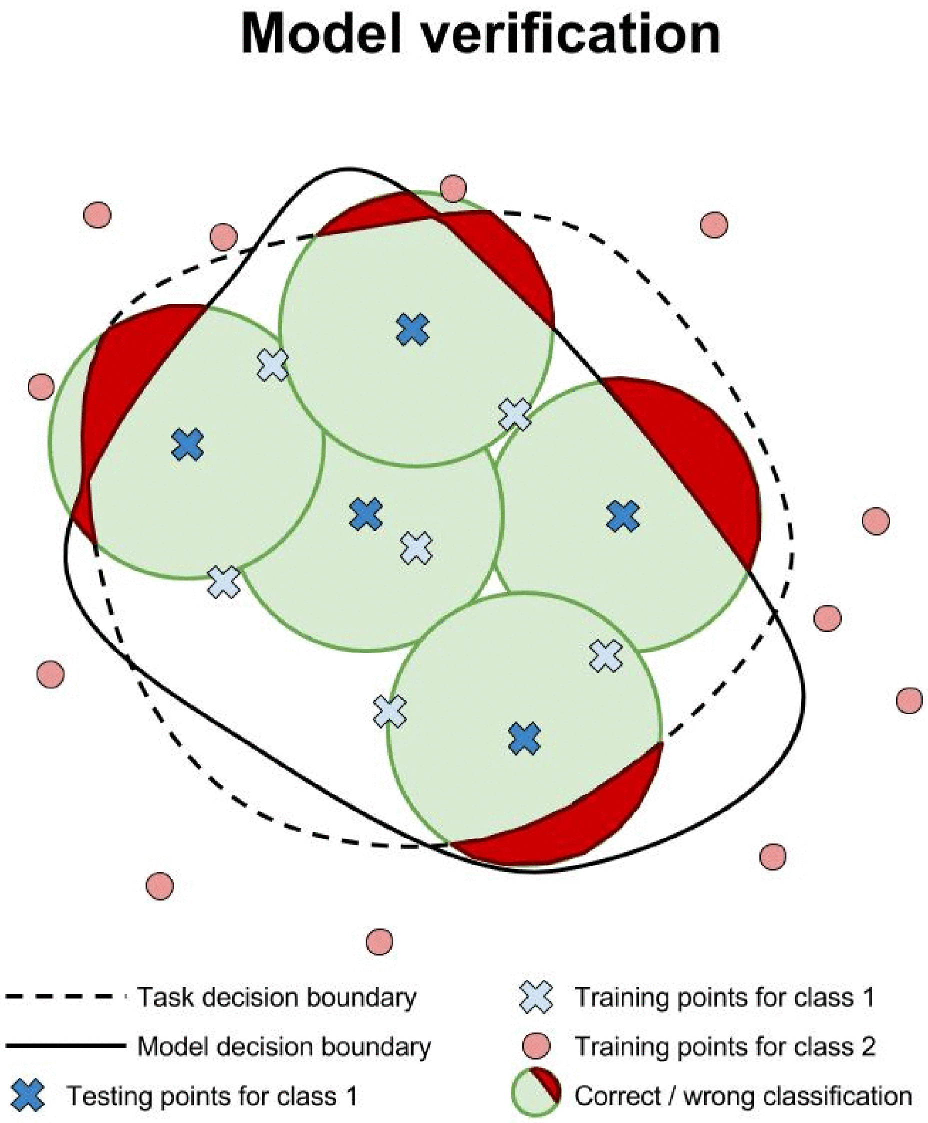
Safety of classification decisions

- Safety assurance process is complex
- Here focus on safety at a point as part of such a process
  - same as pointwise robustness…



- Assume given
  - trained network $f : D \rightarrow \{c_1, \ldots c_k\}$
  - diameter for support region η
  - norm, e.g. L^2, L^∞
- Define safety as invariance of classification decision over η
  - i.e. $\nexists y \in \eta$ such that $f(x) \neq f(y)$
- Also wrt family of safe manipulations
  - e.g. scratches, weather conditions, camera angle, etc

Training vs testing vs verification



Searching for adversarial examples…

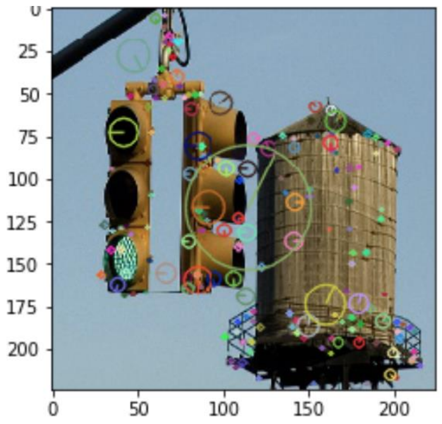
- Input space for most neural networks is high dimensional and non-linear
- Where do we start?
- How can we apply structure to the problem?

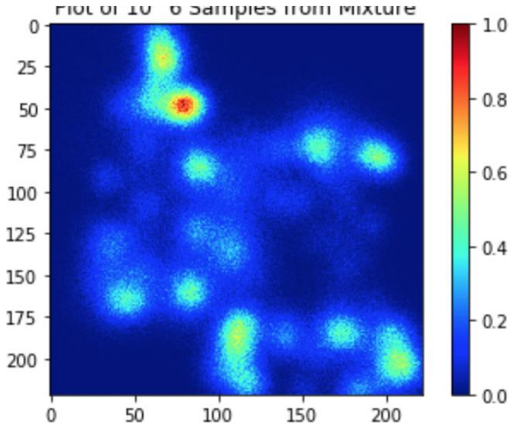
- Image of a tree has 4,000 x 2,000 x 3 dimensions = 24,000,000 dimensions
- We would like to find a very 'small' change to these dimensions

Feature-based representation

- Employ the SIFT algorithm to extract features
- Reduce dimensionality by focusing on salient features
- Use a Gaussian mixture model in order to assign each pixel a probability based on its perceived saliency

$$\mathcal{G}_{i,x} = \frac{1}{\sqrt{2\pi\lambda_{i,s}^2}} exp\left(\frac{-(p_x - \lambda_{i,x})^2}{2\lambda_{i,s}^2}\right) \qquad \mathcal{G}_{i,y} = \frac{1}{\sqrt{2\pi\lambda_{i,s}^2}} exp\left(\frac{-(p_y - \lambda_{i,y})^2}{2\lambda_{i,s}^2}\right)$$

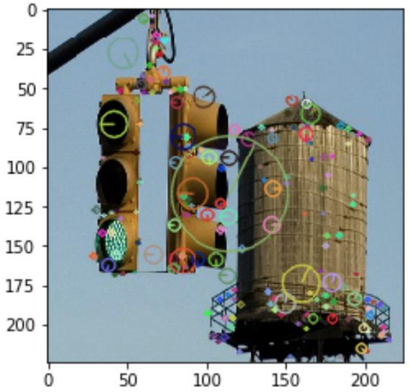




TACAS 2018, https://arxiv.org/abs/1710.07859

Game-based search

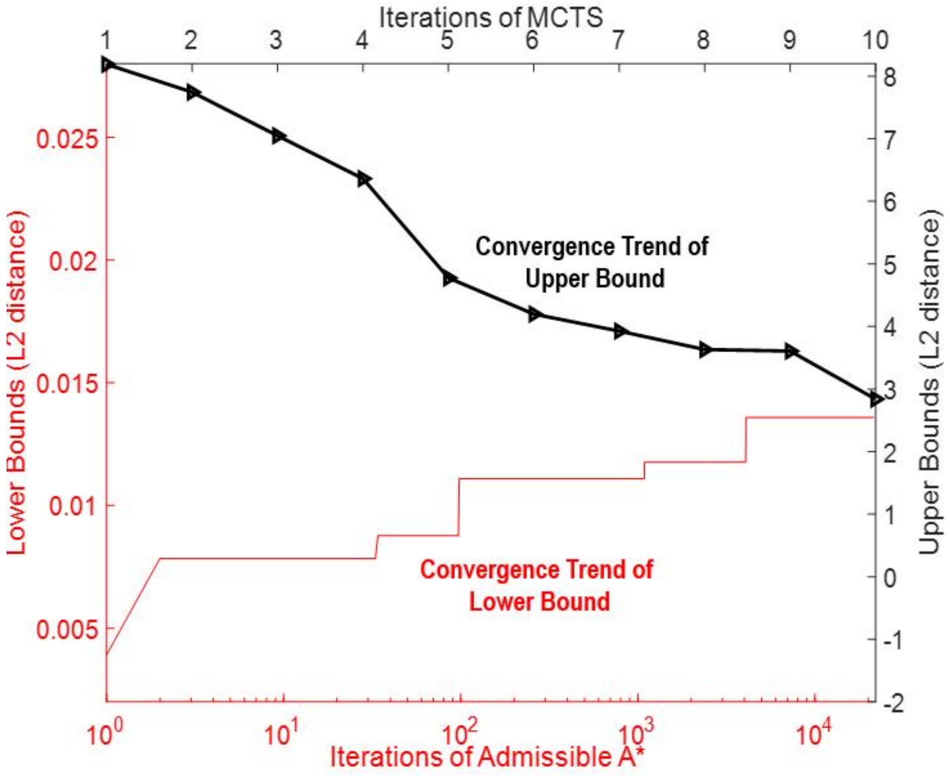
- Goal is finding adv. example, reward inverse of distance
- Player 1 selects the feature that we will manipulate

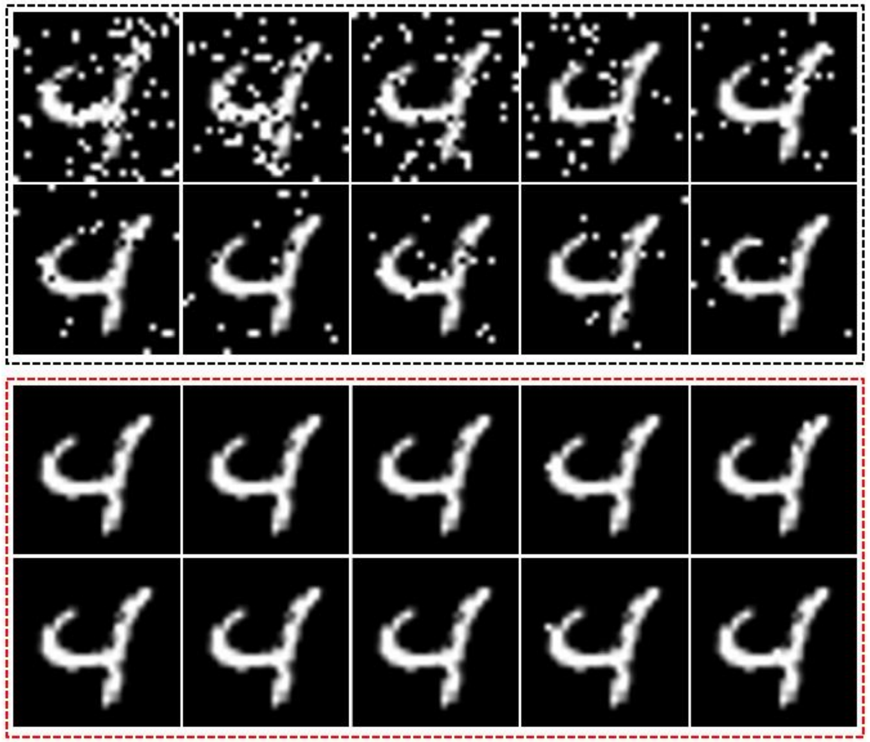


- Each feature represents a possible move for player 1
- Player 2 then selects the pixels in the feature to manipulate
- Use Monte Carlo tree search to explore the game tree, while querying the network to align features
- Method black/grey box, can approximate the maximum safe radius for a given input

Guarantees for deep learning!

- Prove that no adversarial examples exist in a neighbourhood around an input
- Compute lower and upper bounds on maximal safety radius





A Game-Based Approximate Verification of Deep Neural Networks with Provable Guarantees. Wu *et al*, CoRR abs/1807.03571, 2018.

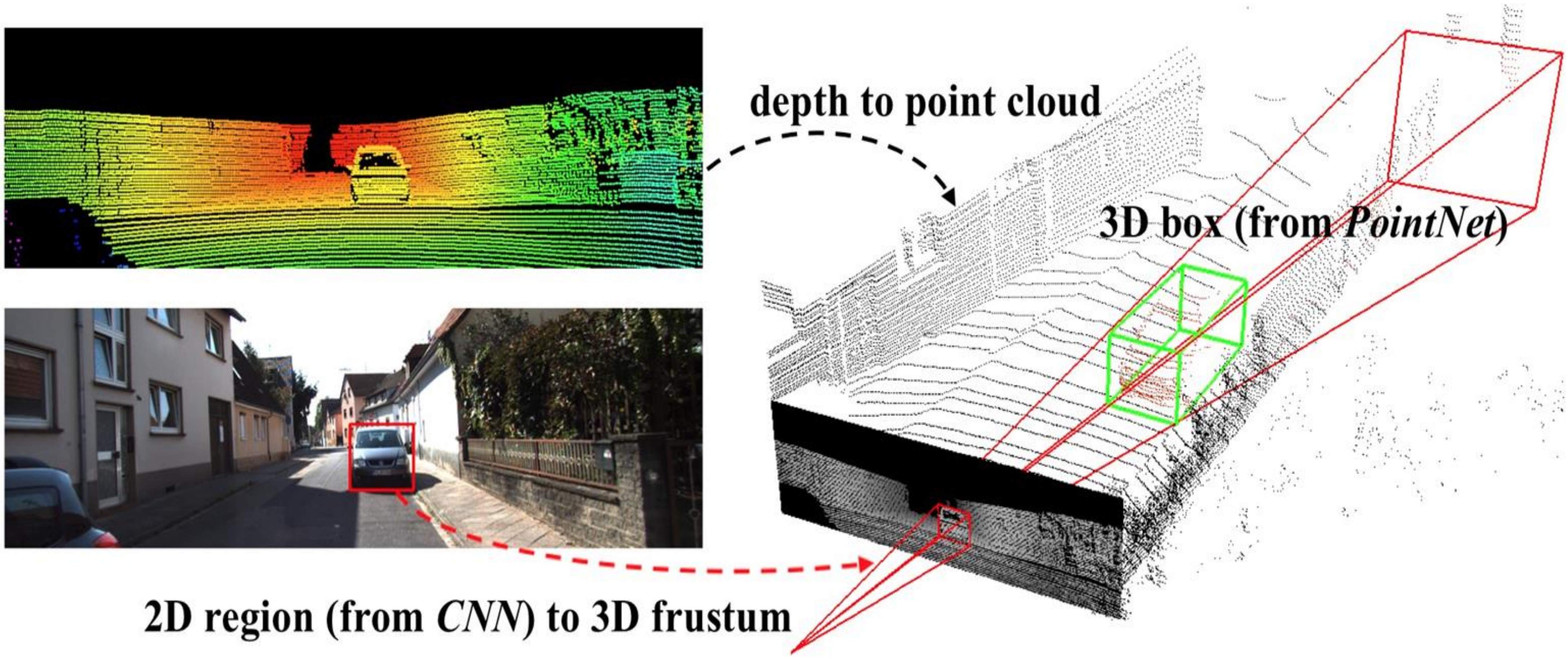
Evaluating safety-critical scenarios: Nexar

- Using our Game-based Monte Carlo Tree Search method we were able to reduce the accuracy of the network form 95% to 0%
- On average, each input took less than a second to manipulate (.304 seconds)
- On average each image was vulnerable to 3 pixel changes

(a) (b) (c)

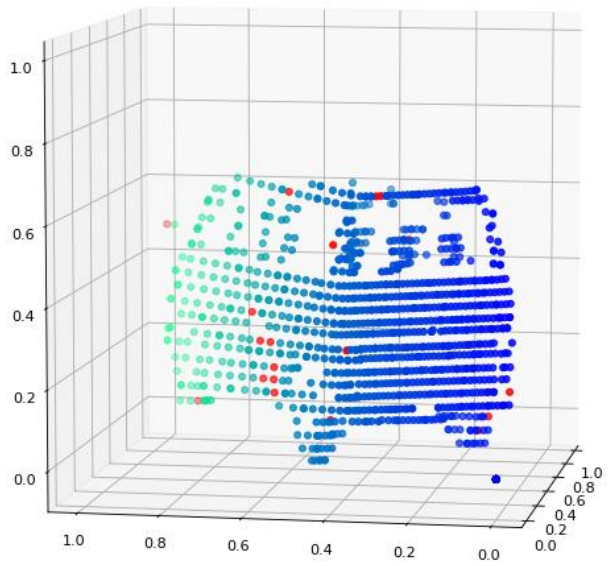
(a) (b) (c)

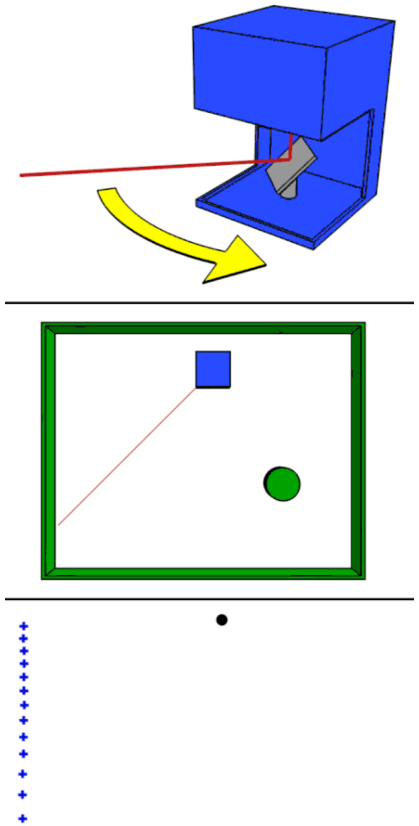
3D deep learning



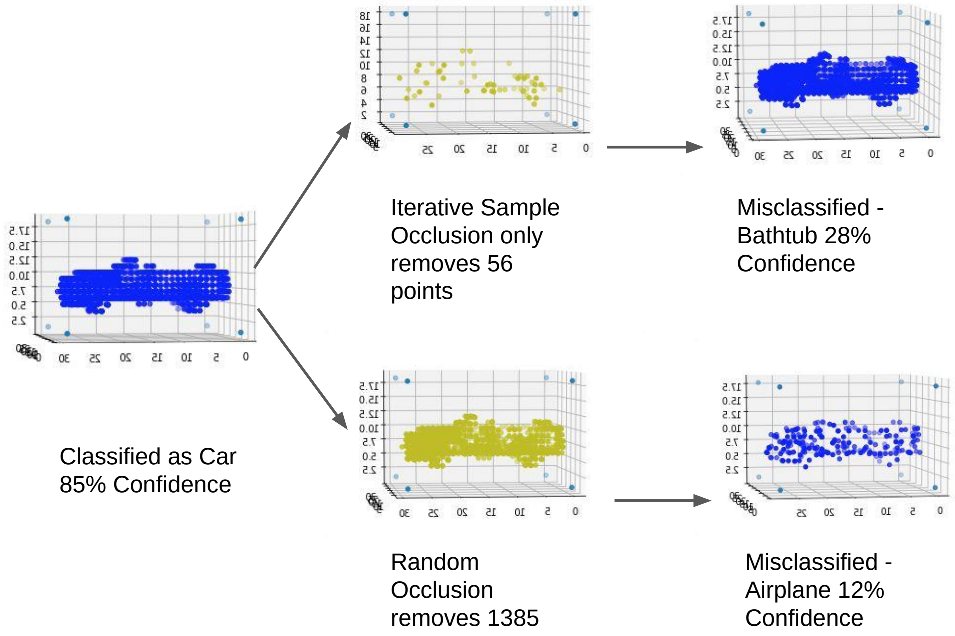
LiDAR and inherent error in point clouds

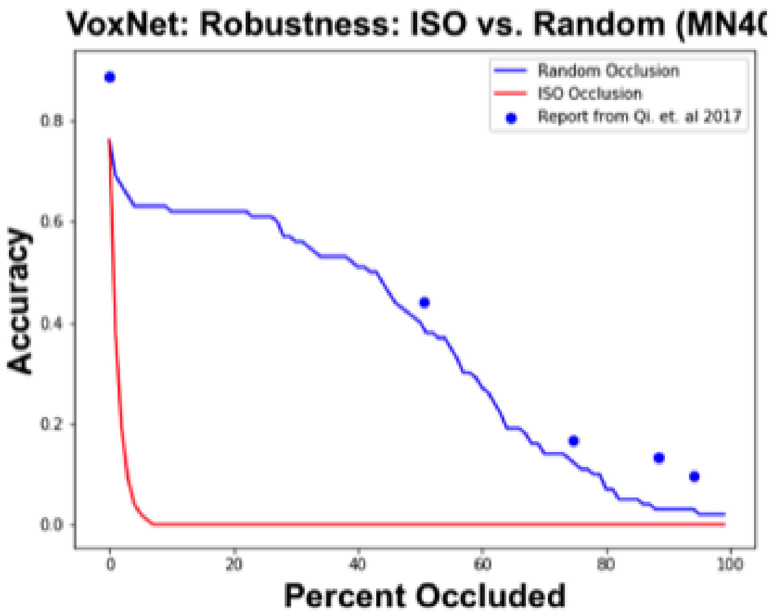
- Point ordering matters
- Partial occlusion of contiguous points
- Dark black could affect the reliability of sensor
- Misoriented sensors
- Need sub-second decision making





Can also attack 3D deep learning (Lidar)





…reduce accuracy to 0% after occlusion of 6.5% of the occupied input space, targeting the critical set

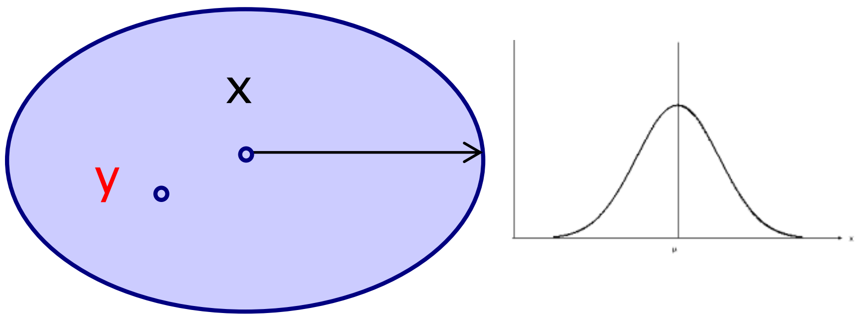
Robustness of 3D Deep Learning in an Adversarial Setting. Wicker & K, In Proc. CVPR 2019.

Probabilistic guarantees

- Requiring that no adversarial examples exist too strict!
- Need to probabilistic guarantees: probability that local perturbations result in predictions that are close to original
- Taking account of the learning process
- Bayesian neural networks have prior on weights
  - account for noise, uncertainty, etc
  - return an uncertainty measure along with the output
- Need to compute posterior probability
  - often intractable
  - can we do better?

Statistical robustness guarantees

- Work with Bayesian neural networks
- Define safety with prob $1-\varepsilon$



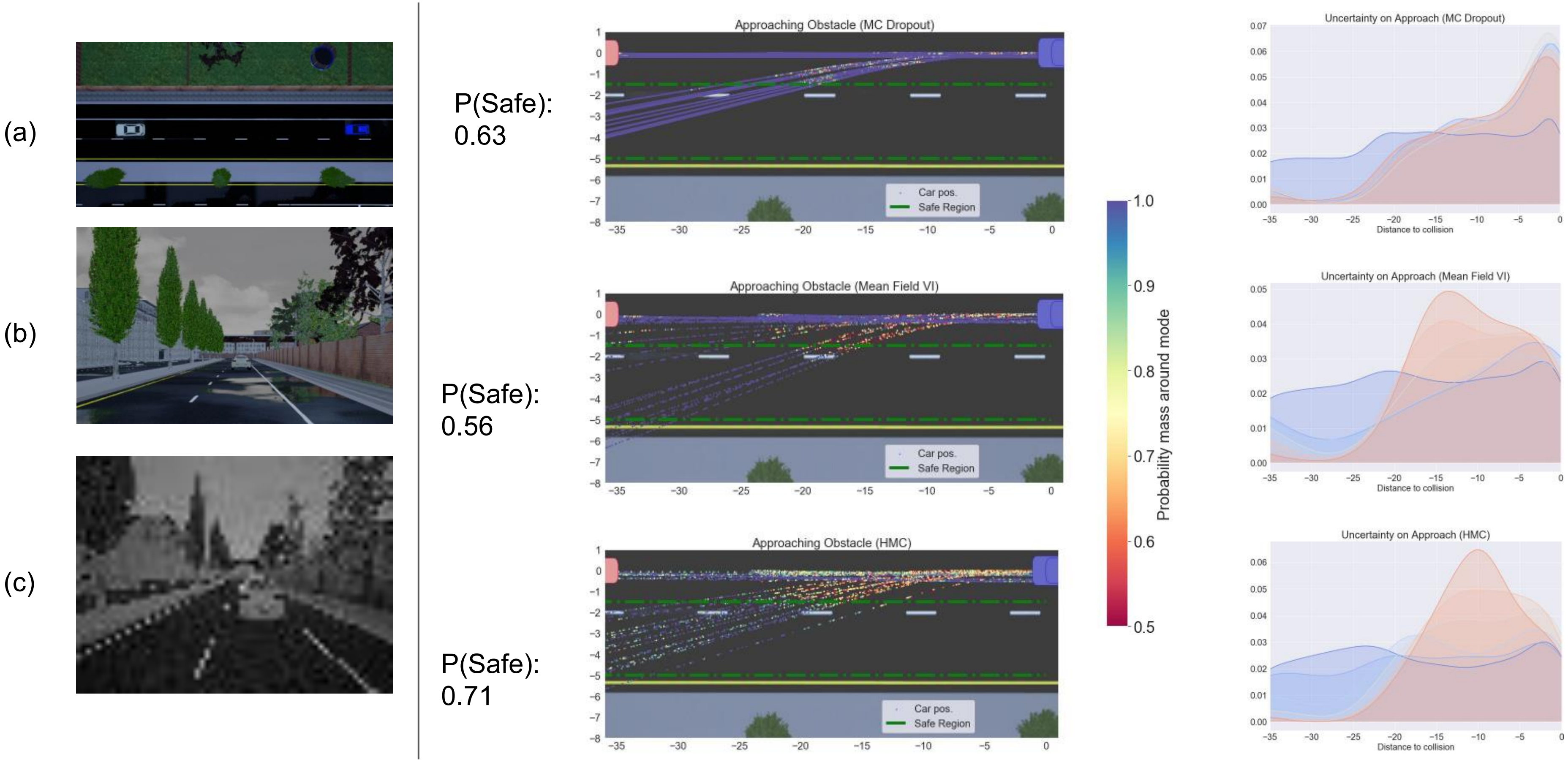
$$Prob(\exists y \in \eta \text{ s.t. } f(x) \neq f(y) \mid D) \leq \varepsilon$$

- i.e. conditioned on training data D
- Method: sample the weights, then employ statistical model checking (Massart bounds, sequential test)
  - compare robustness and accuracy trade offs for different inference methods

IJCAI 2019, https://arxiv.org/abs/1903.01980

Uncertainty quantification with guarantees

- Safety verification for Bayesian neural network autonomous driving controllers



ICRA 2020, https://arxiv.org/abs/1909.09884

But more progress needed…

'I hate them': Locals reportedly are frustrated with Alphabet's self-driving cars

- Alphabet's self-driving cars are said to be annoying their neighbors in Arizona, where Waymo has been testing its vehicles for the last year.
- More than a dozen locals told The Information they they hated the cars, which often struggle to cross a T-intersection near the company's office.
- The anecdotes highlight how challenging it is for self-driving cars, which are programmed to drive conservatively, to handle certain situations.

Published 3:04 PM ET Tue, 28 Aug 2018 | Updated 12:53 PM ET Wed, 29 Aug 2018

CNBC

Source: Waymo

Self-driving cars should be allowed to mount pavements and break speed limit in emergencies

share

28

A Tesla Model S

Concluding remarks

- Much excitement about potential of the developments in AI
- and exciting opportunities!

- But deep learning should be more critically evaluated when put into practice in safety-critical situations

- We must have guarantees for safety, security, privacy, etc
  - formal verification, safety assurance
- and need to know know the limits, also for deep learning
  - rigorous foundations, methodology
- and social implications
  - ethics, fairness and morality

- Many challenges remain

Acknowledgements

- My group and collaborators in this work
- Project funding
  - ERC Advanced Grant VERIWARE
  - EPSRC Mobile Autonomy Programme Grant

- See also
  - PRISM www.prismmodelchecker.org

- New ERC Advanced Grant FUN2MODEL

  "From FUNction-based TO MOdel-based automated probabilistic reasoning for DEep Learning"
- Postdoctoral and PhD positions