COMP813 Artificial Intelligence

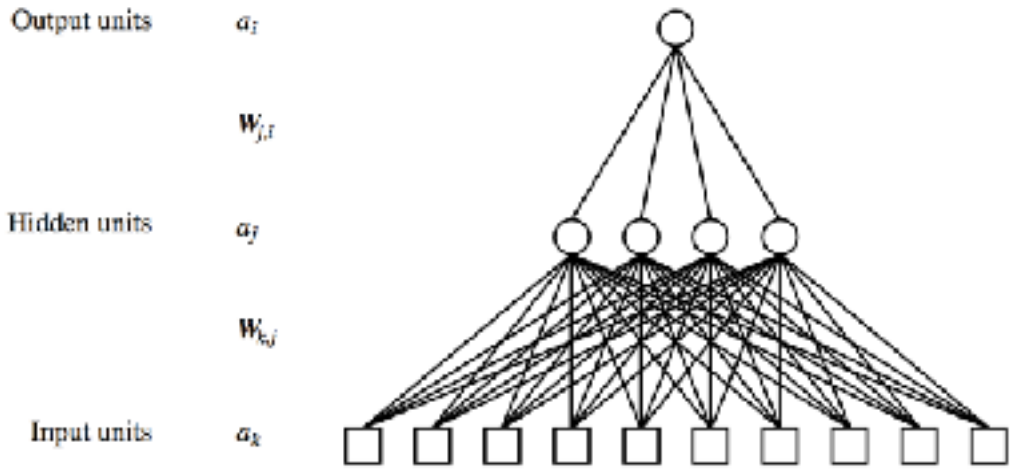
Module 6: AI Safety and Explainability

Auckland University of Technology

Multilayer perceptrons (Deep Learning)

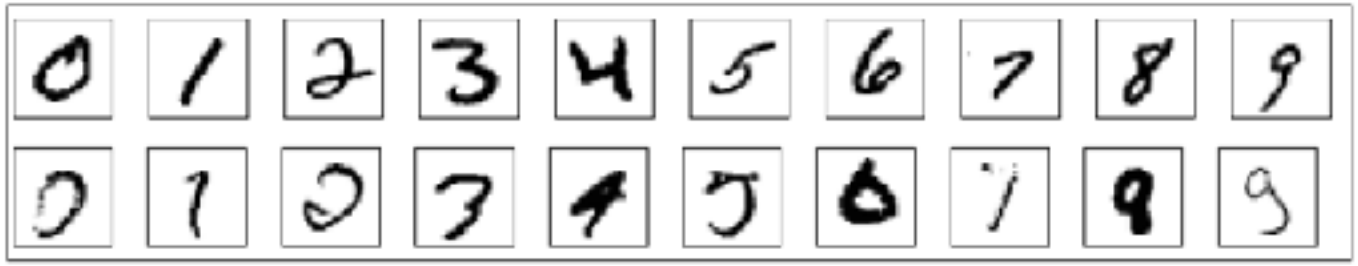
- Layers are usually fully connected;

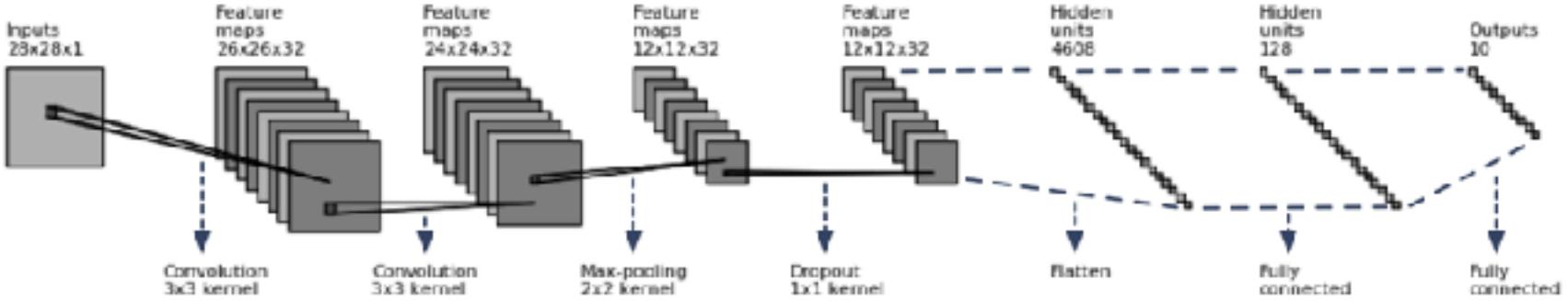
numbers of **hidden units** typically chosen by hand



Deep learning applies in computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation and bioinformatics

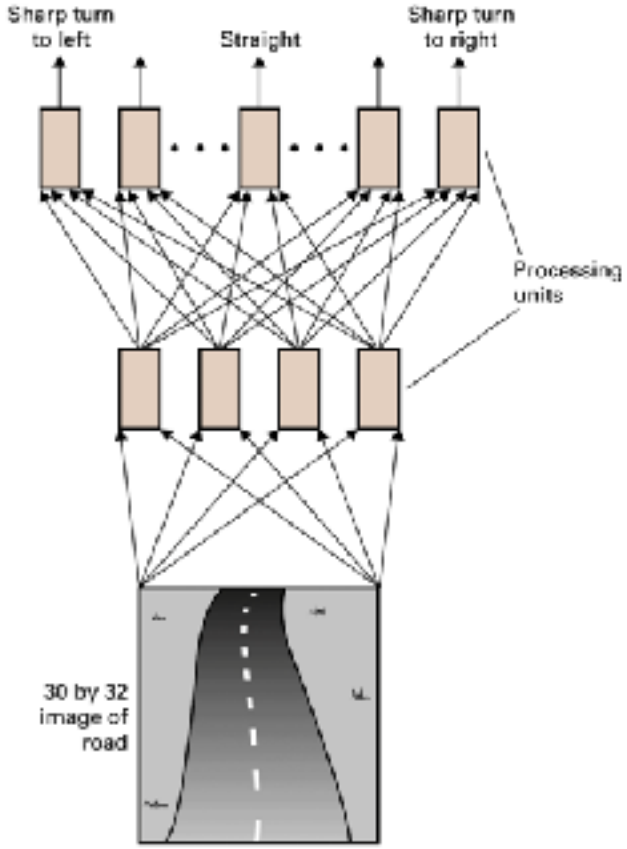
Handwritten digit recognition MNIST



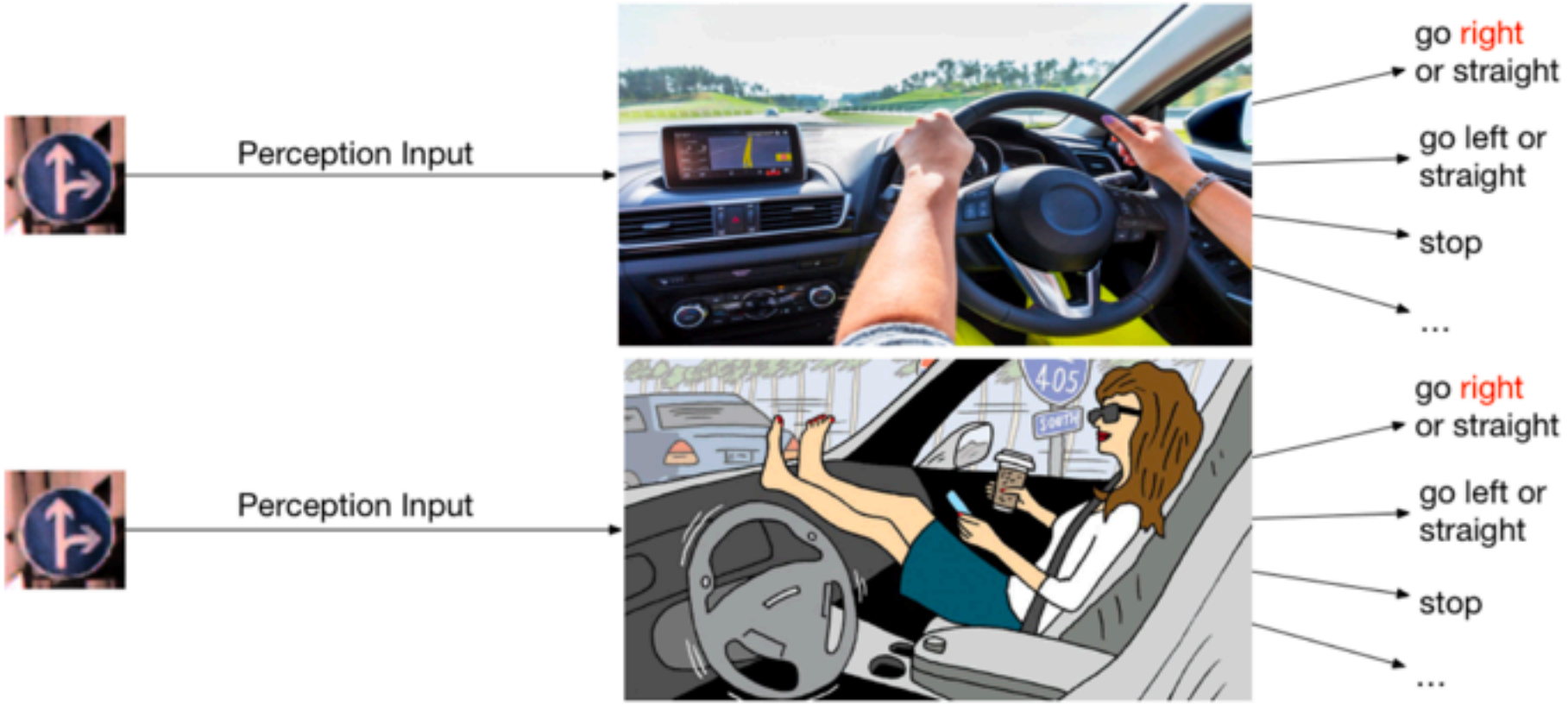


Autonomous Driving

by Dean Pomerleau (CMU)

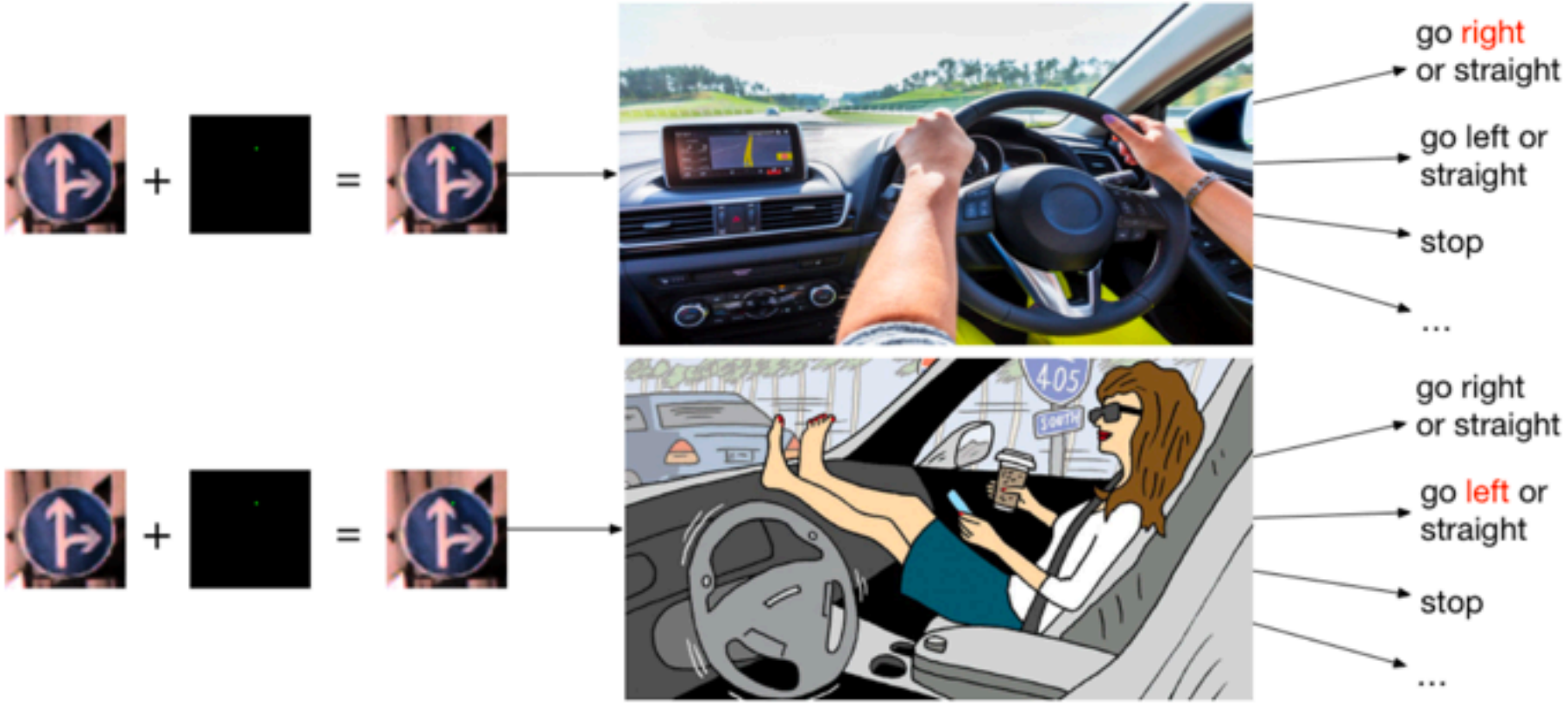


Motivation on AI Safety



Traffic image from “The German Traffic Sign Recognition Benchmark”

Motivation on AI Safety

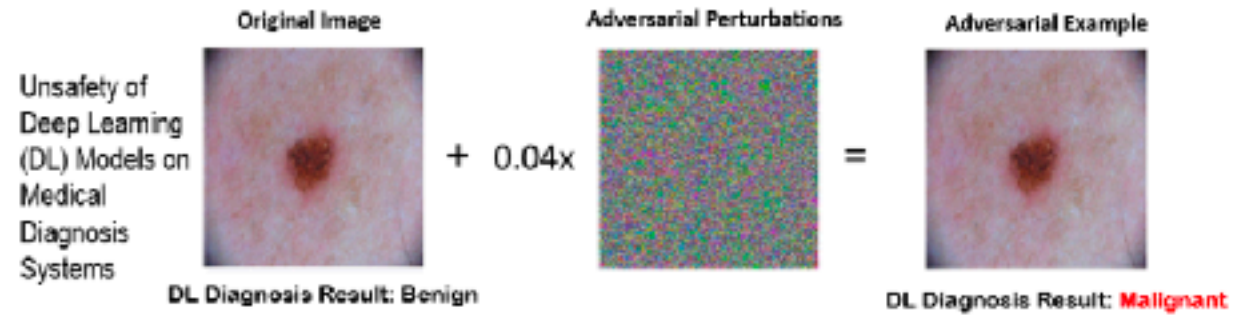


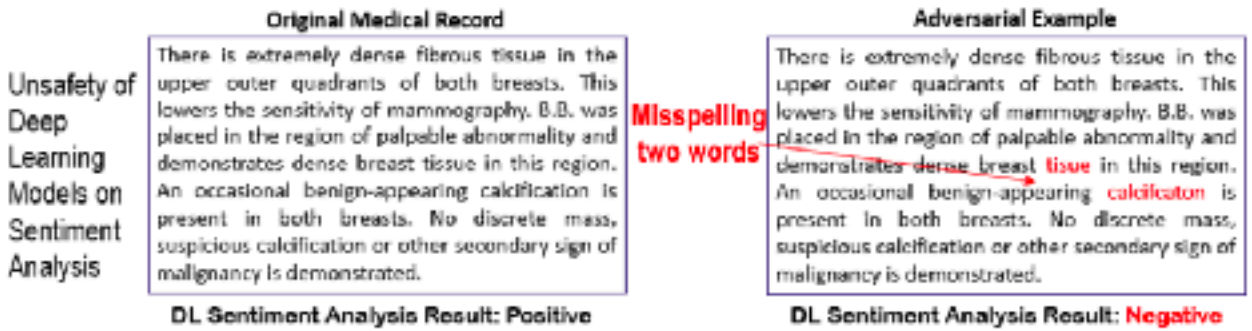
With a very small change to the traffic sign

Motivation on AI Safety

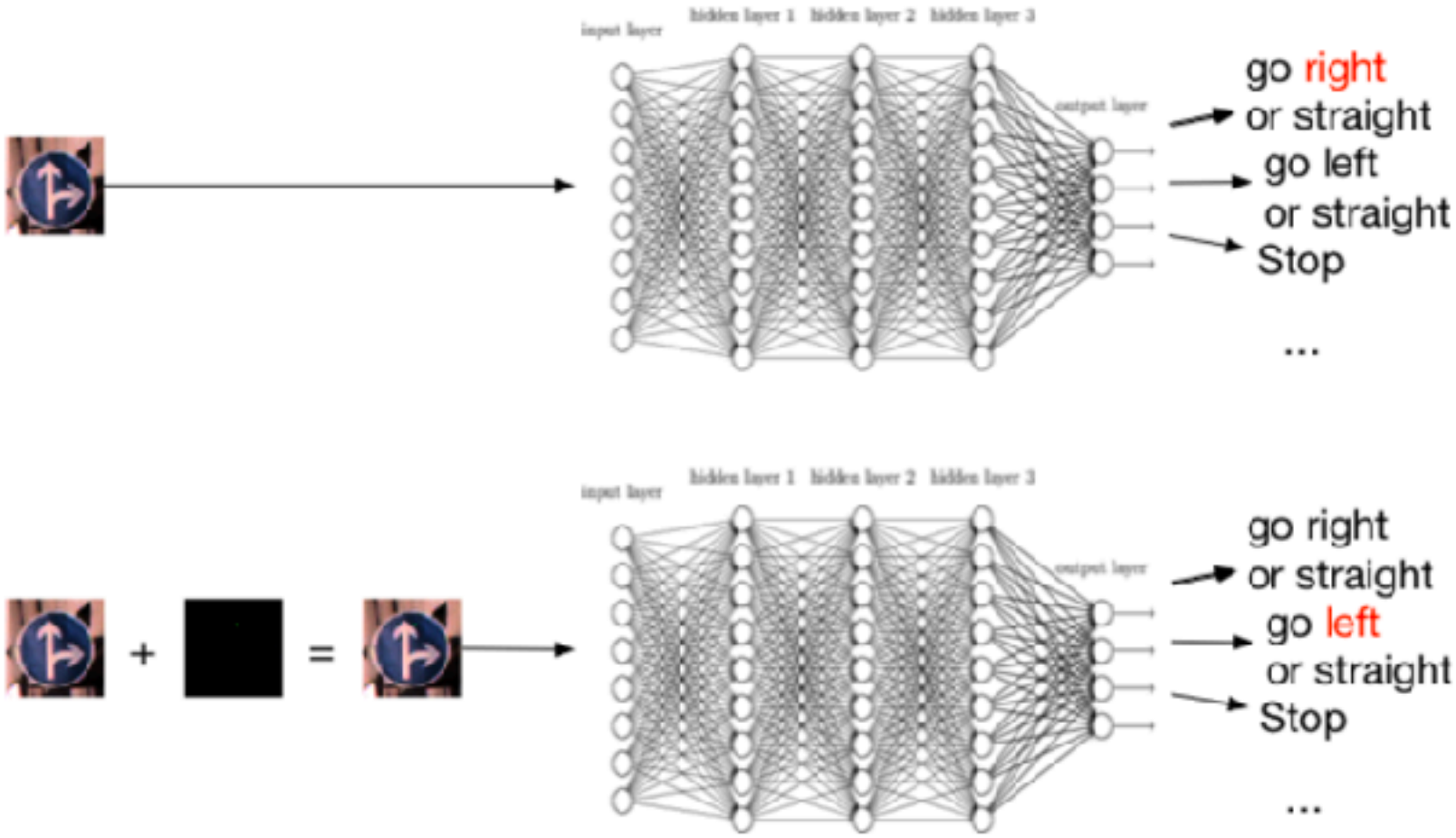
A Tesla Model X driven by Walter Huang is pictured after crashing March 23, 2018, on U.S. 101 in Mountain View. Huang died of injuries suffered in the crash. (LA Times/YouTube)

Motivation on AI Safety



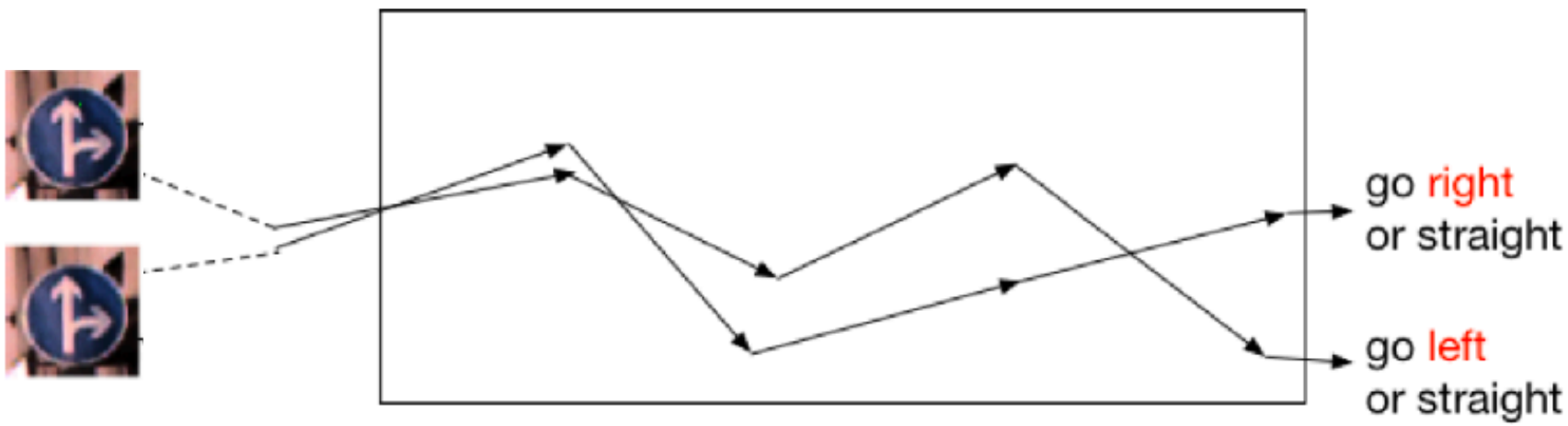


Motivation on AI Safety



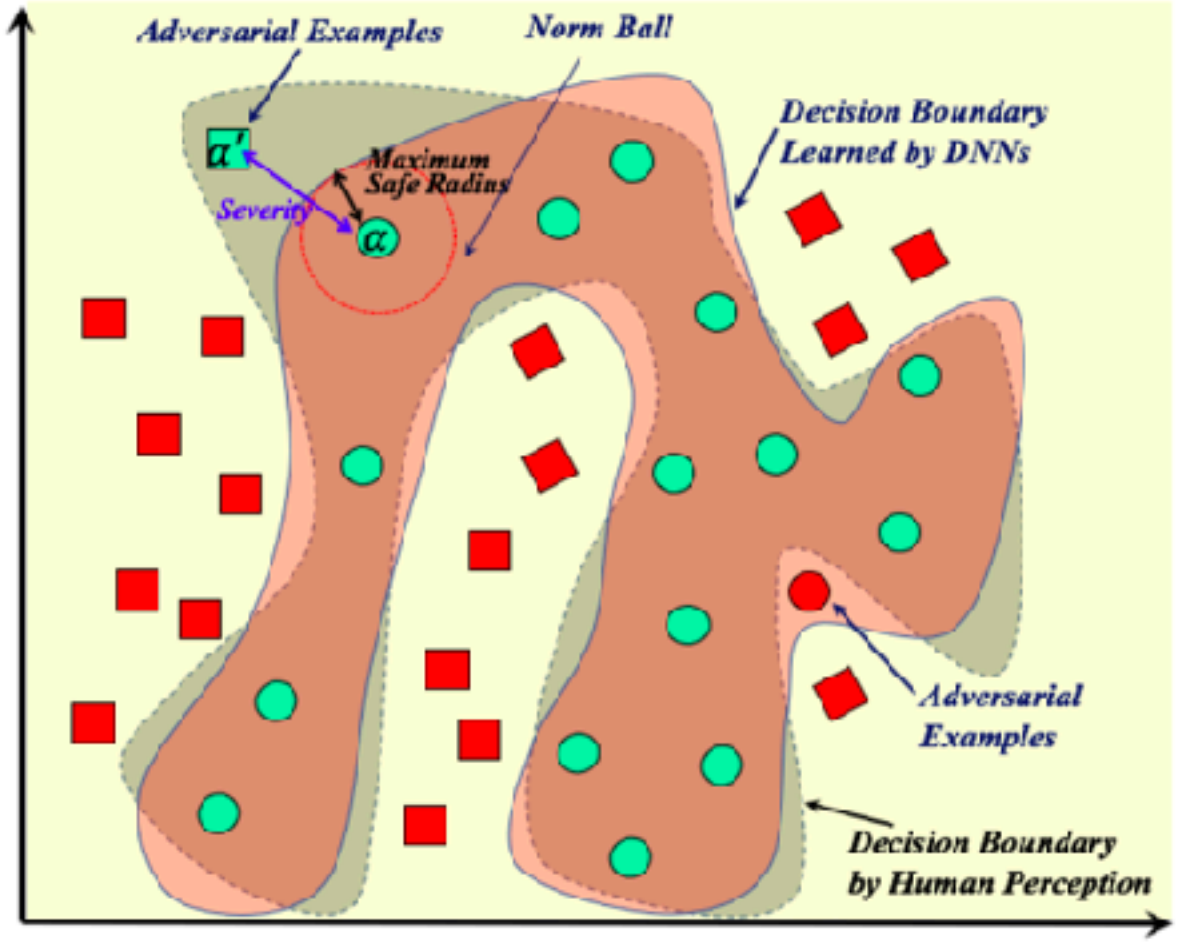
Safety Verification of Deep Neural Networks (SVDNN)

Safety verification of deep neural networks*. Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. In CAV 2017, pages 3–29, 2017.



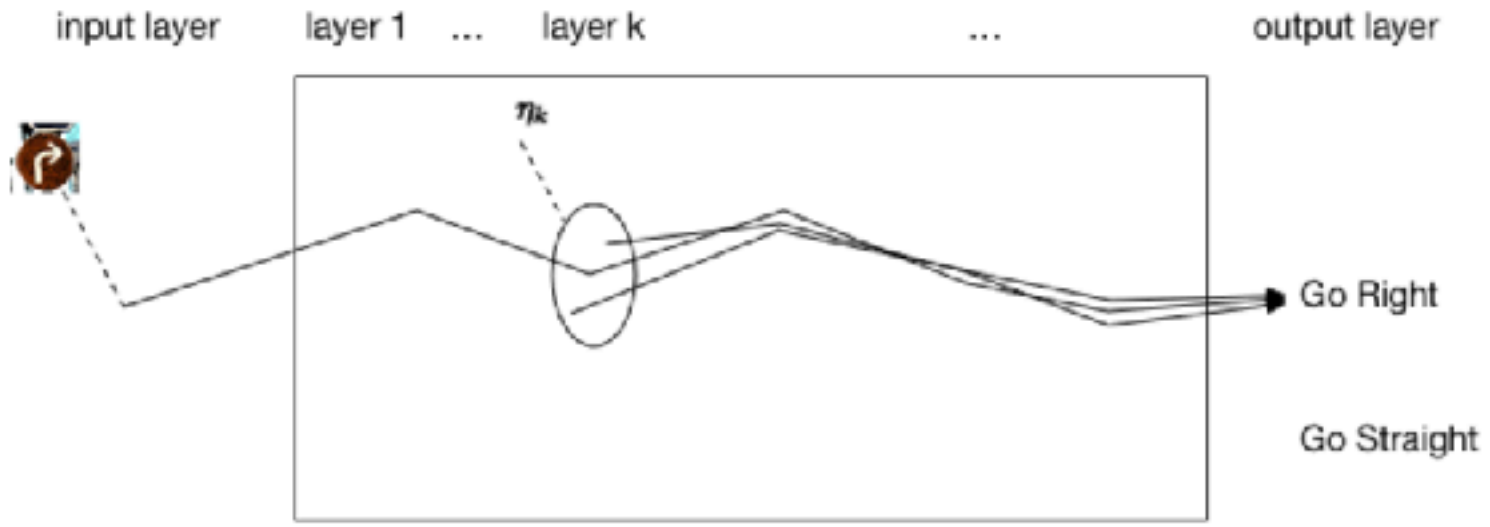
* https://link.springer.com/chapter/10.1007/978-3-319-63387-9_1

SVDNN: basic ideas



SVDNN: basic ideas

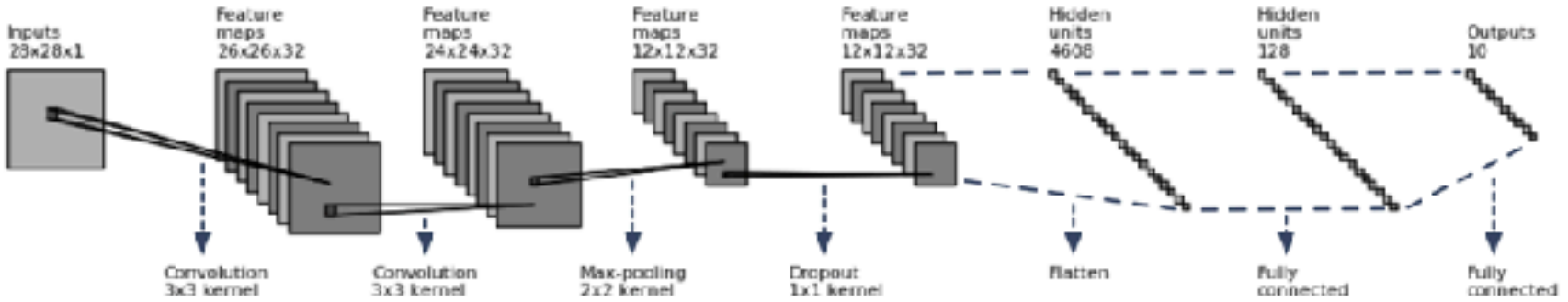
[General Safety] Let $\eta_k(\alpha_{x,k})$ be a region in layer L_k of a neural network N such that $\alpha_{x,k} \in \eta_k(\alpha_{x,k})$. We say that N is *safe for input x and region* $\eta_k(\alpha_{x,k})$, written as $N, \eta_k \models x$, if for all activations $\alpha_{y,k}$ in $\eta_k(\alpha_{x,k})$ we have $\alpha_{y,n} = \alpha_{x,n}$.



* https://link.springer.com/chapter/10.1007/978-3-319-63387-9_1

SVDNN: Experiments MNIST

Image Classification Network for the MNIST Handwritten Numbers 0 – 9

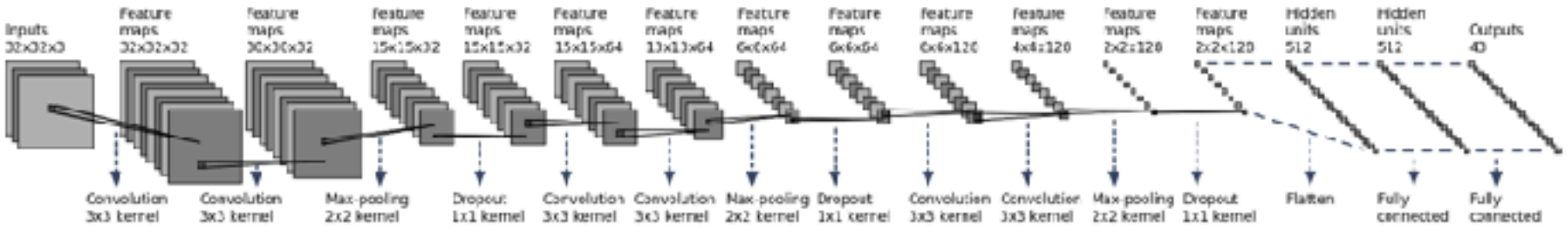


Total params: 600,810

SVDNN: Experiments MNIST

SVDNN: Experiments GTSRB

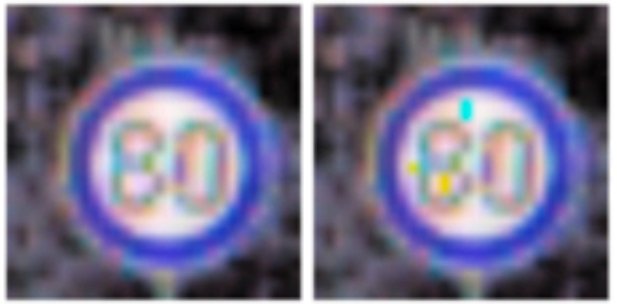
Image Classification Network for The German Traffic Sign Recognition Benchmark



Total params: 571,723

SVDNN: Experiments GTSRB

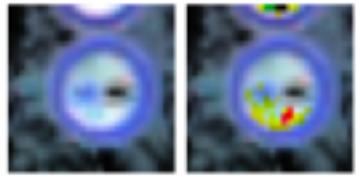
"stop"
to "30m speed limit"



"80m speed limit"
to "30m speed limit"

"go right"
to "go straight"

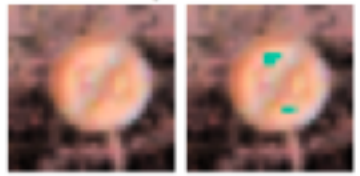
SVDNN: Experiments GTSRB



no overtaking (prohibitory) to go straight (mandatory)

speed limit 50 (prohibitory) to stop (other)

road narrows (danger) to construction (danger)



restriction ends 80 (other) to speed limit 80 (prohibitory)

no overtaking (trucks) (prohibitory) to speed limit 80 (prohibitory)

no overtaking (prohibitory) to restriction ends (overtaking (trucks)) (other)

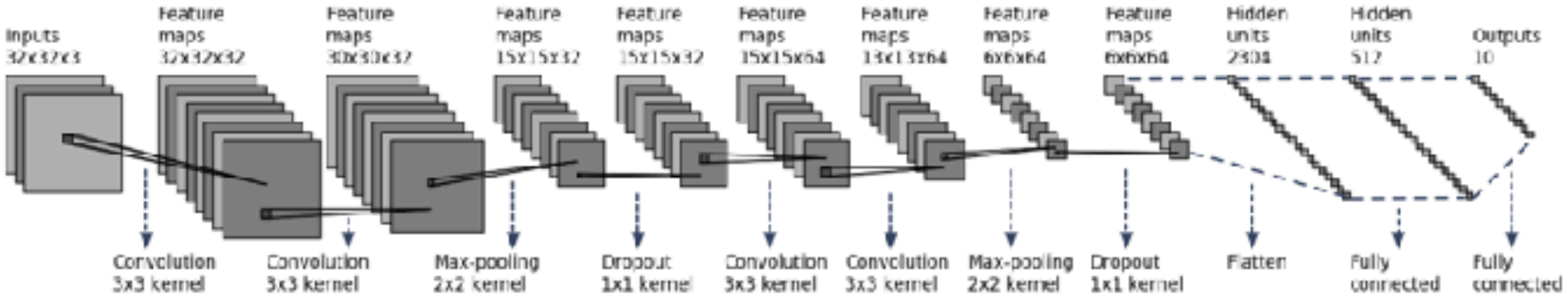
priority at next intersection (danger) to speed limit 30 (prohibitory)

uneven road (danger) to traffic signal (danger)

danger (danger) to school crossing (danger)

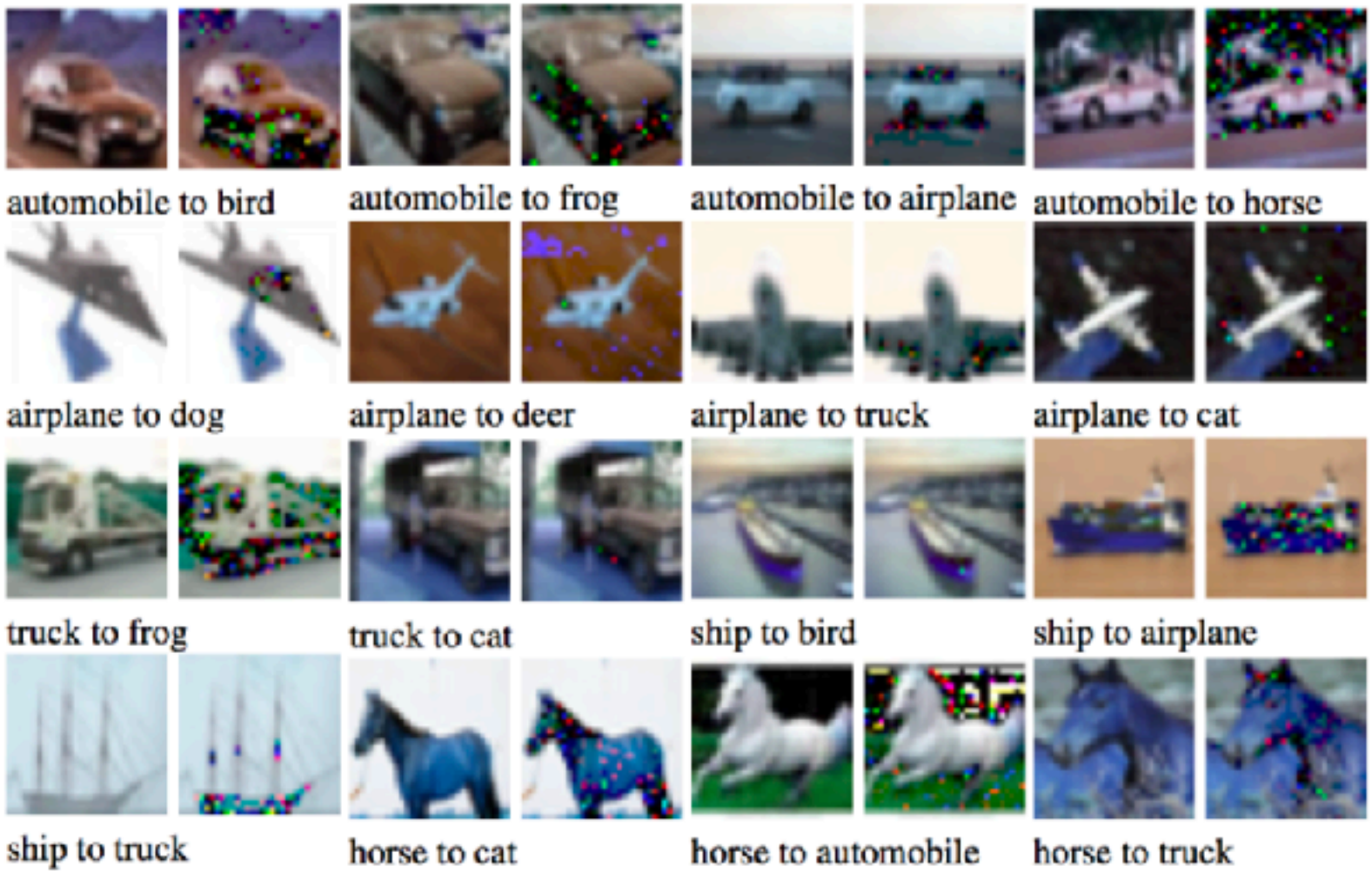
SVDNN: Experiments CIFAR-10

Image Classification Network for the CIFAR-10 small images



Total params: 1,250,858

SVDNN: Experiments CIFAR-10



Motivation on Explainability XAI

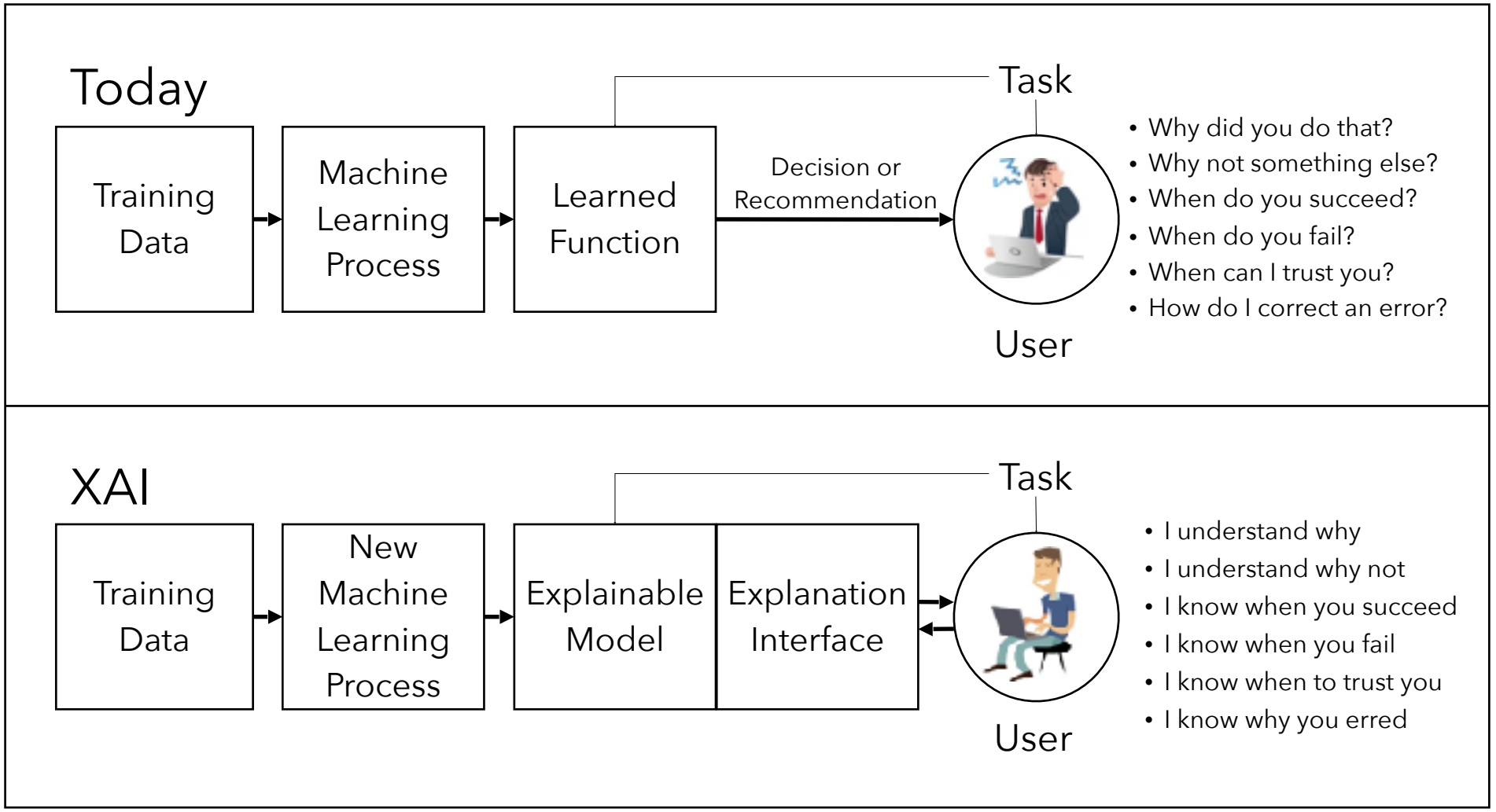
US DARPA (Defense Department Research Arm)
Explainable AI (XAI) program:

1) Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)

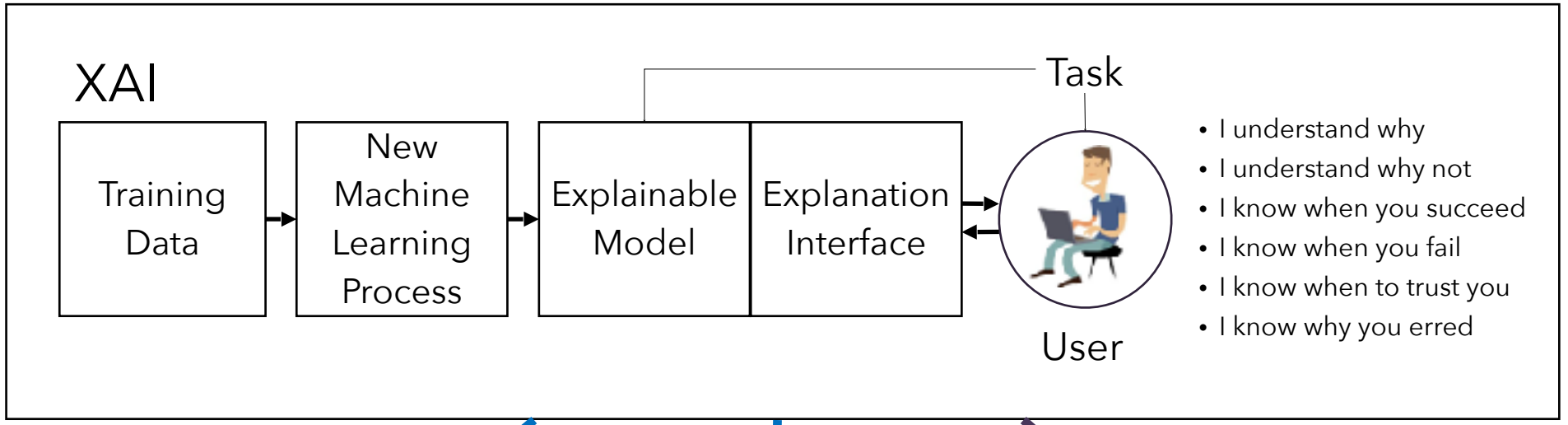
2) Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

https://www.darpa.mil/program/explainable-artificial-intelligence

B. Program Scope – XAI Concept



B. Program Scope – XAI Development Challenges



Explainable Models

- develop a range of new or modified machine learning techniques to produce more explainable models

Explanation Interface

- integrate state-of-the-art HCI with new principles, strategies, and techniques to generate effective explanations

Psychology of Explanation

- summarize, extend, and apply current psychological theories of explanation to develop a computational theory

TA 1: Explainable Learners

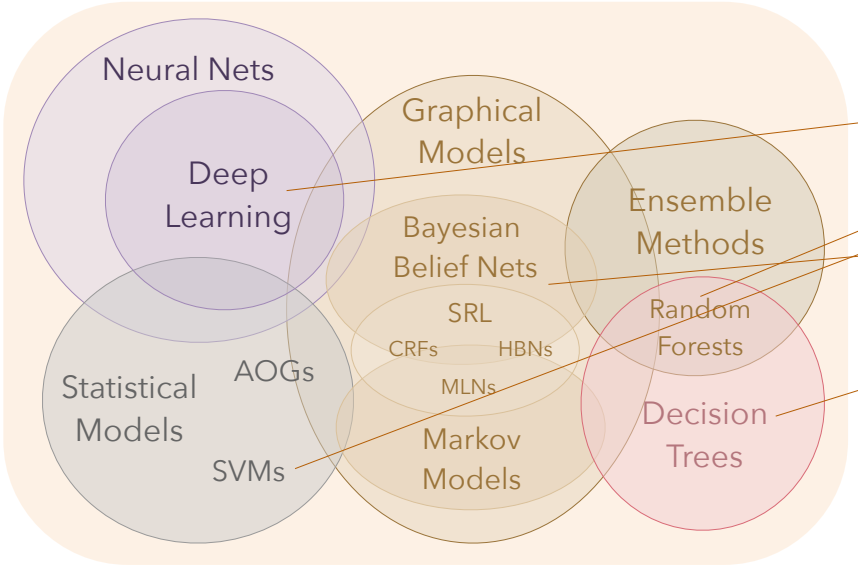
TA 2: Psychological Models

B.1 Explainable Models

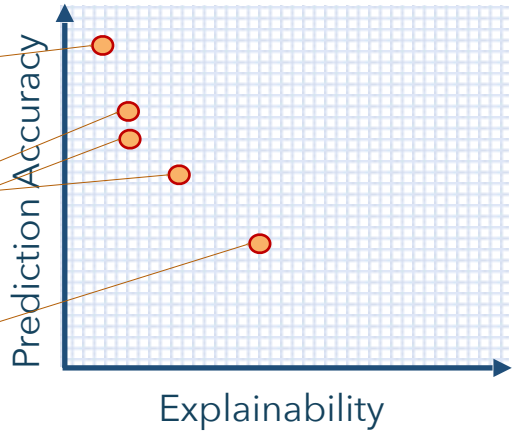
New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)

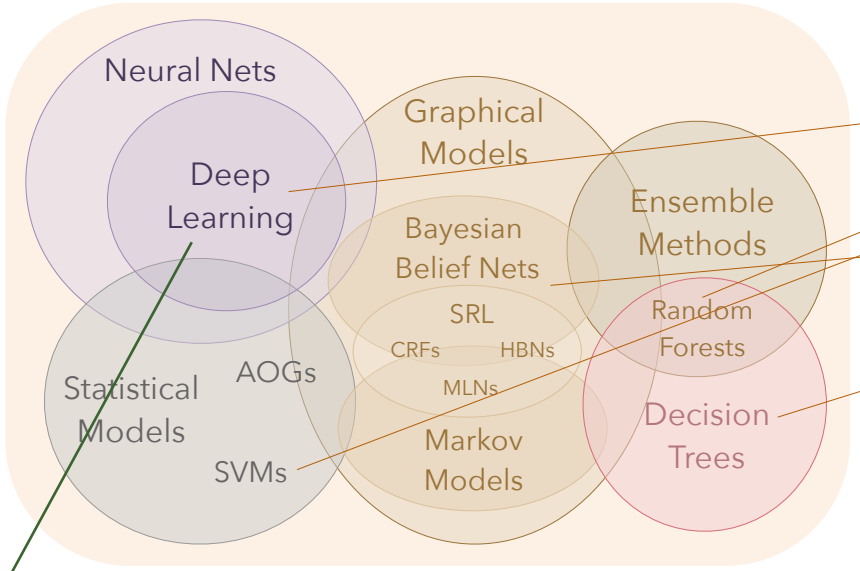


B.1 Explainable Models

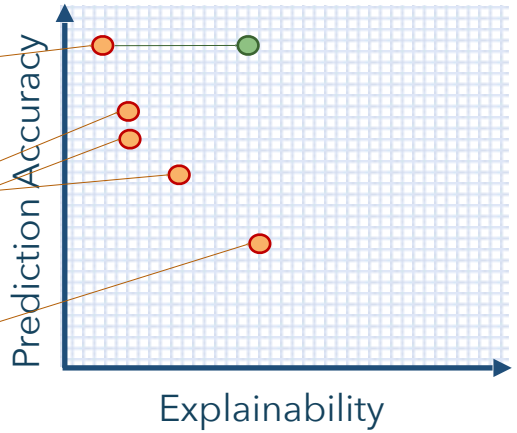
New Approach

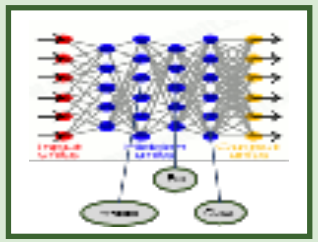
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)





Deep Explanation

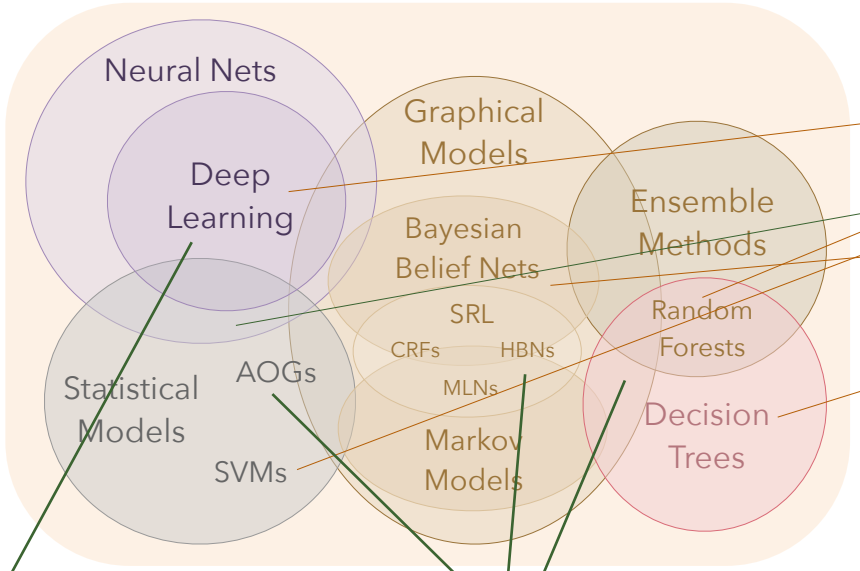
Modified deep learning techniques to learn explainable features

B.1 Explainable Models

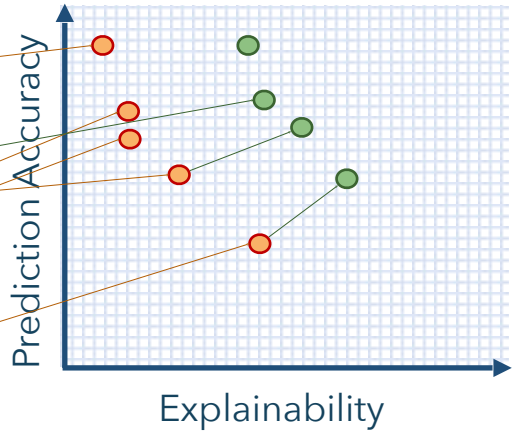
New Approach

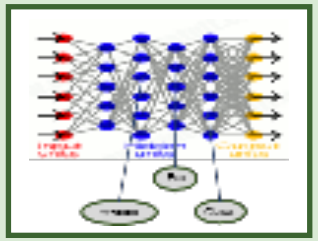
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



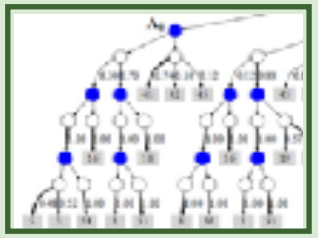
Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models

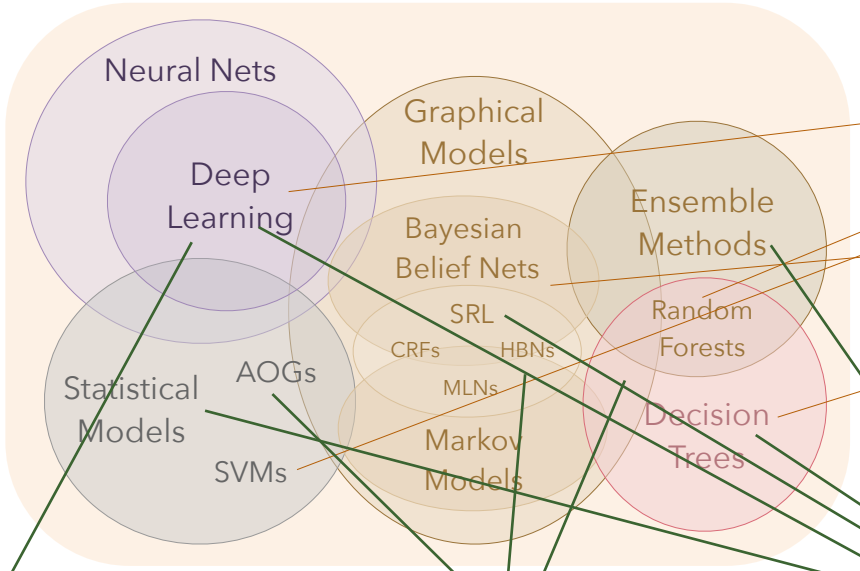
https://www.darpa.mil/program/explainable-artificial-intelligence

B.1 Explainable Models

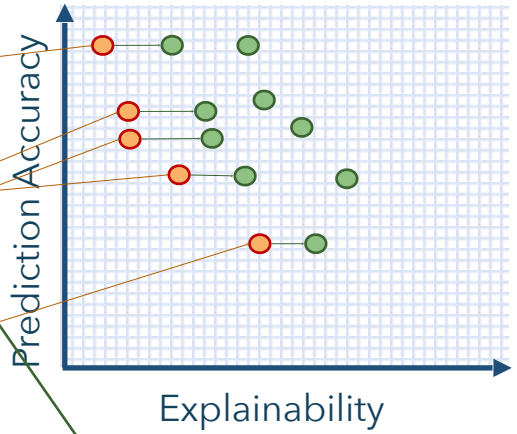
New Approach

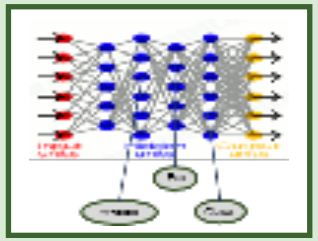
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



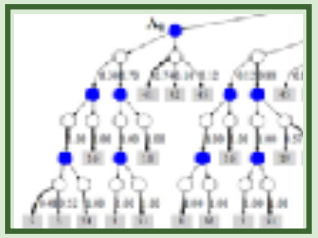
Explainability (notional)





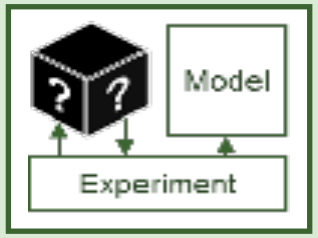
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models



Model Induction

Techniques to infer an explainable model from any model as a black box

https://www.darpa.mil/program/explainable-artificial-intelligence

Model Induction by LIME

LIME (Local Interpretable Model-agnostic Explanations) :

Ribeiro, Singh, and Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of (KDD '16).

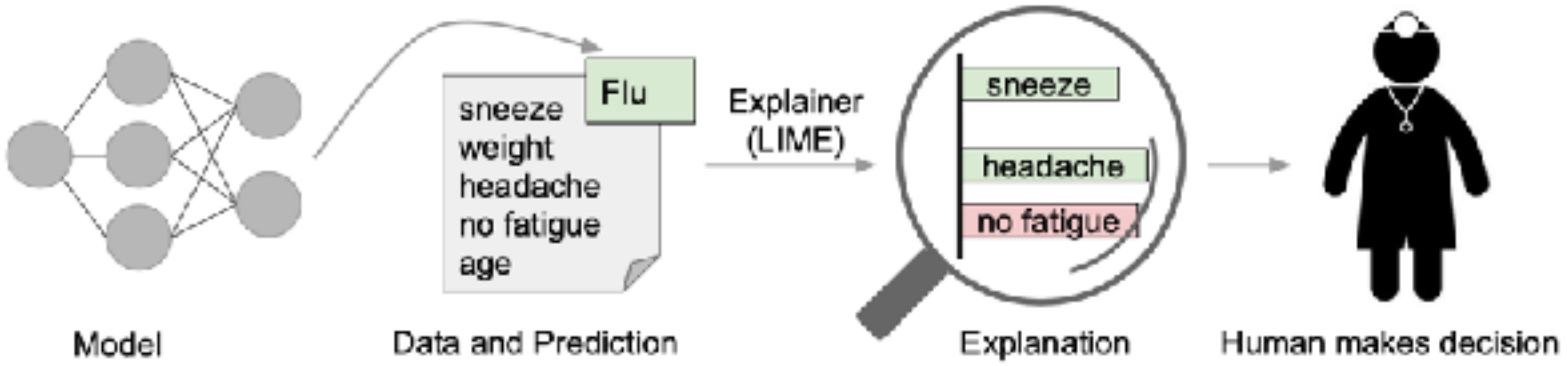


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

LIME: Desired Characteristics for Explainers

- **Interpretable**, i.e., provide qualitative understanding between the input variables and the response.
- **Locally faithful**, i.e. it must correspond to how the model behaves in the vicinity of the instance being predicted.
- **Model-agnostic**, i.e. treat the original model as a black box (for generality)

LIME: key components

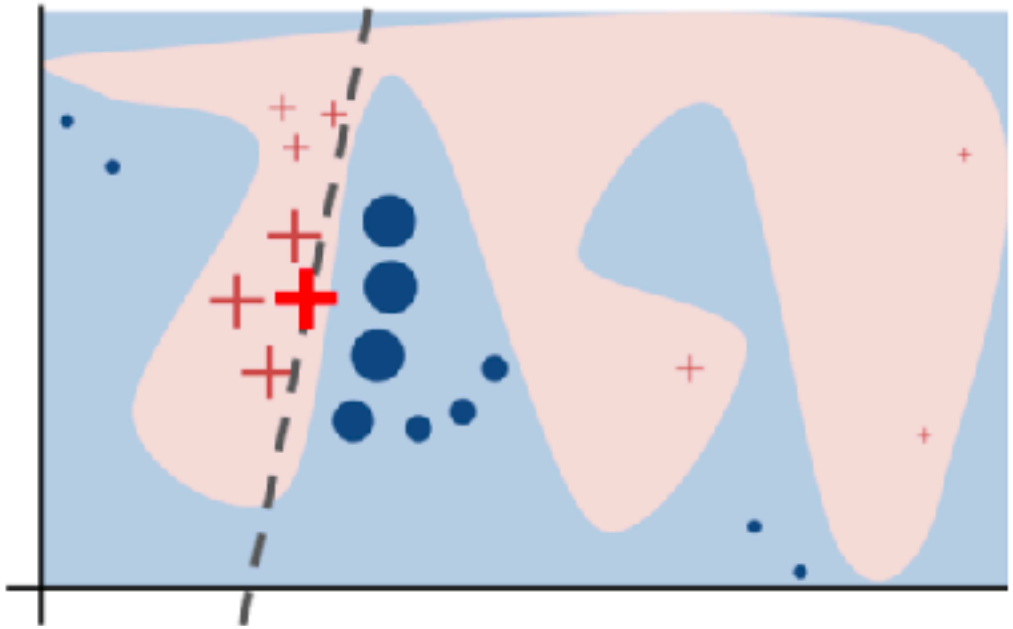
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Function to be explained

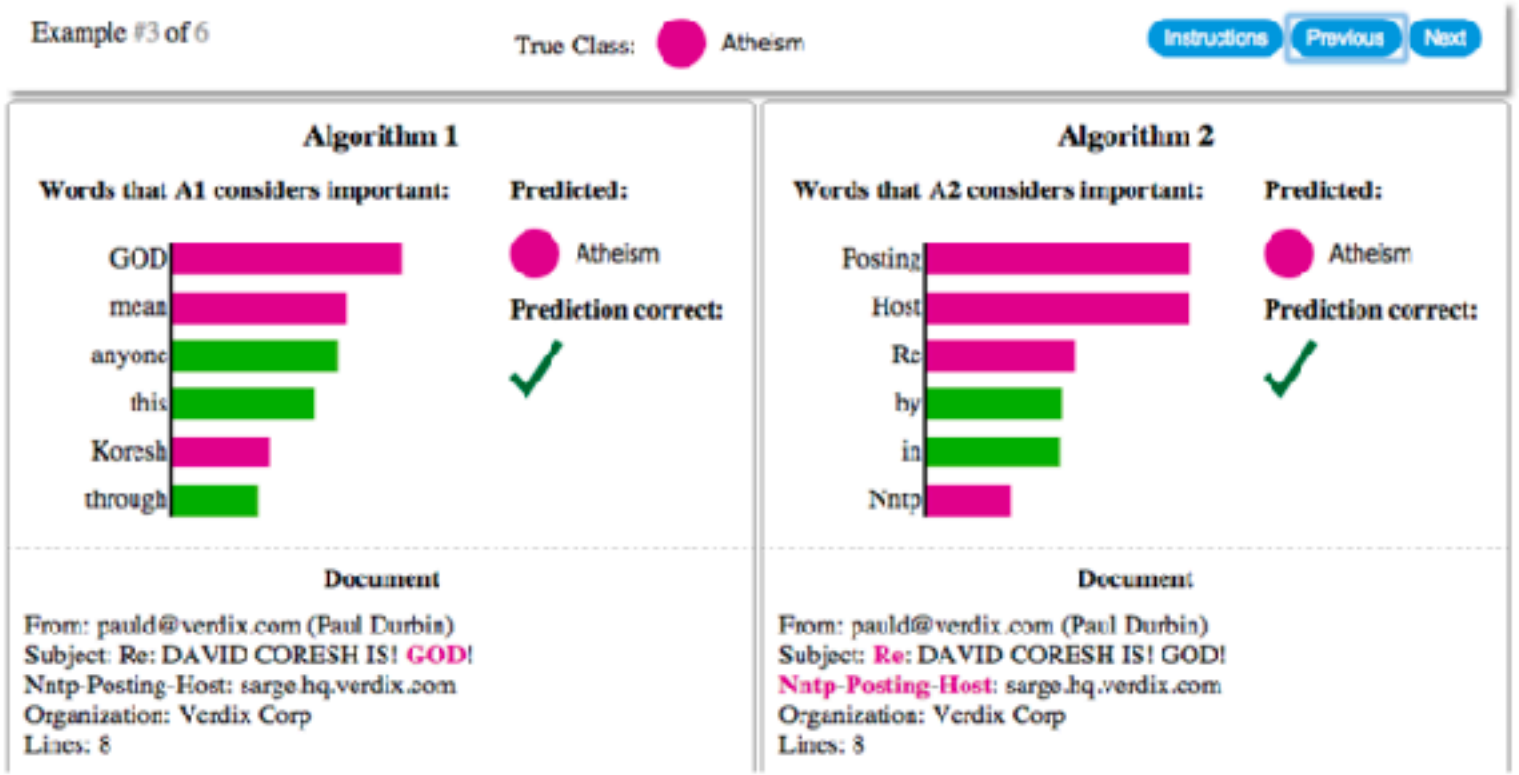
Function to explain

Locality around x

Complexity measure

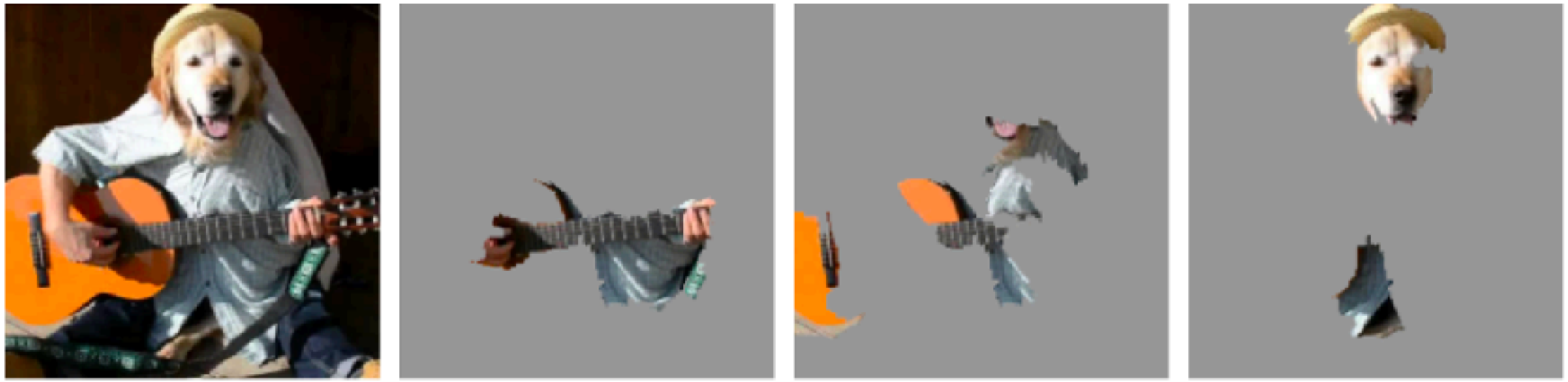


LIME: Text classification with SVMs



Algorithm 2: this classifier achieves 94% held-out accuracy.
The word “Posting” appears in 22% of examples in the training set, 99% of them in the class “Atheism”.

LIME: Deep networks for images

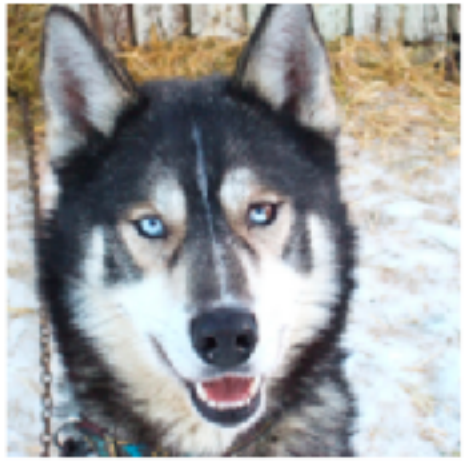


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

The super-pixels with positive weight towards a specific class

LIME:Do explanations lead to insights?



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Geoff Hinton 2018 Wired Interview

"**I'm an expert on trying to get the technology to work**, not an expert on social policy. One place where I do have technical expertise that's relevant is [whether] regulators should insist that you can explain how your AI system works. I think that would be **a complete disaster**.

People can't explain how they work, for most of the things they do. When you hire somebody, the decision is based on all sorts of things you can quantify, and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decision, you are forcing them to make up a story.

Neural nets have a similar problem. When you train a neural net, it will learn a billion numbers that represent the knowledge it has extracted from the training data. If you put in an image, out comes the right decision, say, whether this was a pedestrian or not. **But if you ask "Why did it think that?" well if there were any simple rules for deciding whether an image contains a pedestrian or not, it would have been a solved problem ages ago.**"

Critical Responses

"**I'm an expert on trying to get the technology to work**, not an expert on social policy. One place where I do have technical expertise that's relevant is [whether] regulators should insist that you can explain how your AI system works. I think that would be **a complete disaster**."

For **Dr. Heather Roff, Associate Fellow from the Leverhulme Centre for the Future of Intelligence, University of Cambridge**, the responsibility needs to be broadened. She asserts,

"This is a dangerous position to take. **An expert on technology who feels themselves divorced from social or policy implications does not understand that technology is not value neutral**, and that their decisions—even seemingly basic ones on how many gradient descents to take in a system — have socio-political implications. If one thinks they are only Scientists doing Science, but then simultaneously think that regulators should take an interest has fundamentally misunderstood their role as scientists engaging in socially and morally important questions. If your work requires legislation then you should think about that at the design stage… period."

Critical Responses

"**People can't explain how they work**, for most of the things they do. When you hire somebody, the decision is based on all sorts of things you can quantify, and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decision, you are forcing them to make up a story."

Timothy Miller, Associate Professor in computer science at the University of Melbourne, Australia, whose specializes in explainable AI and human-agent collaboration, disputes Hinton's claim on the limitations of human explanation:

"**His quoted paragraph is itself an explanation**: an explanation of why he has reached the decision that explainability for AI would be a disaster. Is he making up a story about this? I imagine he would claim that he is not and that it is based on careful reasoning. But in reality, it is based on neurons in his brain firing in a particular way that nobody understands. The ability to communicate his reasons to others is a strength of the human brain. Philosopher Daniel Dennett claims that consciousness itself is simply our brain creating an `edited digest' of our brains inner workers for precisely the purpose of communicating our thoughts and intentions (including explanations) to others."

Critical Responses

"Neural nets have a similar problem. When you train a neural net, it will learn a billion numbers that represent the knowledge it has extracted from the training data. If you put in an image, out comes the right decision, say, whether this was a pedestrian or not. **But if you ask "Why did it think that?" well if there were any simple rules for deciding whether an image contains a pedestrian or not, it would have been a solved problem ages ago.**"

For the DOD, where precision in aspects of war require investigation and justification, David Gunning introduces the work being done on explainable ML that will allow future warfighters to "understand, appropriately trust and manage an emerging generation of AI Machine partners" :

"**There are techniques to explain deep nets: DARPA's Explainable AI (XAI) program**, and a growing community of researchers, are developing techniques that can be used to explain, at least partially, deep nets: (1) there are techniques that can select the training examples that were most influential in a decision; (2) there are techniques to identify the most salient input features used in a decision; (3) there are network dissection techniques that can identify meaning features inside the layers of a deep net that can be used for explanation; and: (4) there are deep learning researchers who are developing deep learning techniques to generate explanations. "

Trustworthy AI

Trustworthy AI = Certification + Explanation

- **Certification:** held before the deployment of an AI system to make sure that it functions correctly (and safely).
- **Explanation**: held whenever needed during the lifetime of the AI system. An investigation can be conducted, with a formal report produced, to understand any unexpected behaviour of the system.

Reference

- AIMA book: Chapter 18.7, 26, 27
- http://en.wikipedia.org/wiki/Machine_learning
- https://www.coursera.org/course/ml

- SVDNN: Safety Verification of Deep Neural Networks. Huang, X., Kwiatkowska, M., Wang, S., Wu, M. (2017). In: Majumdar, R., Kunčak, V. (eds) Computer Aided Verification. CAV 2017. Lecture Notes in Computer Science, vol 10426. Springer, Cham. https://doi.org/10.1007/978-3-319-63387-9_1 https://arxiv.org/abs/1610.06940

- A Survey of Safety and Trustworthiness of Deep Neural Networks https://arxiv.org/abs/1812.08342

- US DARPA (Defense Department Research Arm) Explainable AI (XAI) program: https://www.darpa.mil/program/explainable-artificial-intelligence

- LIME: “Why Should I Trust You?": Explaining the Predictions of Any Classifier. Ribeiro, Singh, and Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of (KDD ’16). https://dl.acm.org/doi/10.1145/2939672.2939778

- Interpretable Machine Learning https://christophm.github.io/interpretable-ml-book/