Adversarial and Secure Machine Learning

Preliminary meeting

Ching-yu.kao@aisec.fraunhofer.de

Goal

- Combination of machine learning/deep learning and IT Security

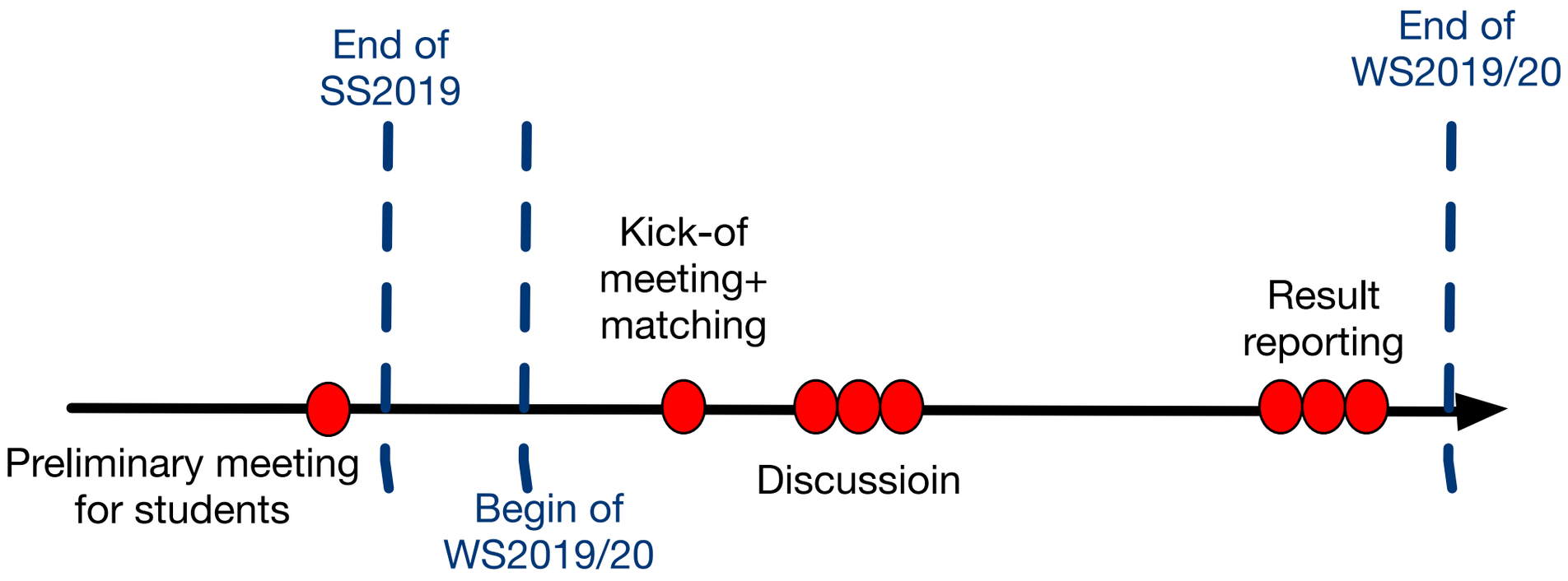
- Read one paper carefully and implement it by yourself

- No written report is needed

- Max. 8 person, 2 person forms a team, each team pick one topic

Arrangement

- Preliminary meeting - one day in this semester
- Kick-off Meeting - one day in October 2019
- Discussion Session - one topic per day, totally 3 days in November 2019
- Reporting - one topic per day, totally 3 days in January 2020



Grading

- Reimplementing the discussed paper -> score 2.0

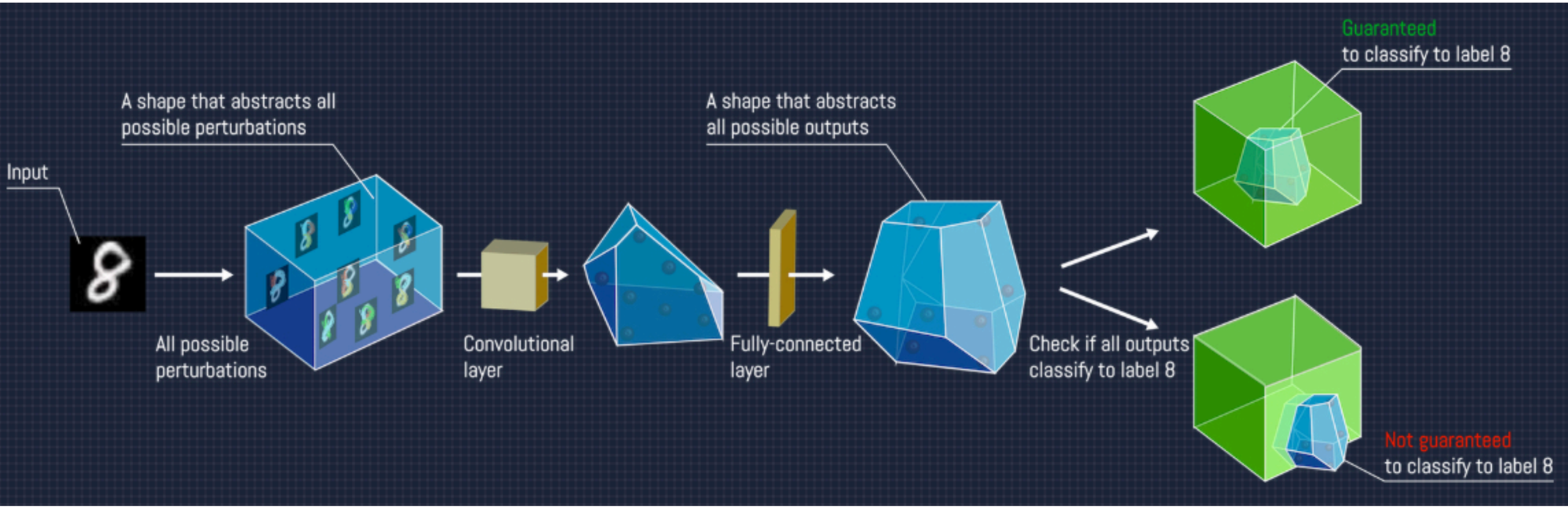
- Each new idea -> -0.3 (small contributions, for example: use security datasets, time efficiency, improvements)

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
- Domain learning
- Anomaly detection
- Deep reinforcement learning

Possible topics

- Certification of deep learning



Source:AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation, IEEE S&P 2018

Possible topics

- Certification of deep learning
- Explainable AI : Attention + Seq-to-Seq

S2VT: A cat is trying to get a small board.

Source: http://www.cs.utexas.edu/users/ml/papers/venugopalan.iccv15.pdf
https://arxiv.org/pdf/1810.02851.pdf

Possible topics

- Certification of deep learning
- Explainable AI : Attention + Seq-to-Seq

Source Text: south korea issued a stern warning monday against illegal labor disputes and campus protests and announced the arrest of ### radicals for violent weekend disturbances .	
Ground Truth: south korea issues stern warning against labor and campus activists	**(A-1)Supervised Result:** south korea issues stern warning against illegal labor disputes
(C-2)WGAN: south korea issued stern warning against illegal labor disputes	**(C-3)Adversarial REINFORCE:** south korea issued stern warning against illegal labor disputes campus arrest
(E-2)WGAN : south korea issued stern warning against illegal labor disputes and arrest	**(E-3)Adversarial REINFORCE:** south korea issued stern warning against illegal labor disputes campus protests

Source: http://www.cs.utexas.edu/users/ml/papers/venugopalan.iccv15.pdf
https://arxiv.org/pdf/1810.02851.pdf

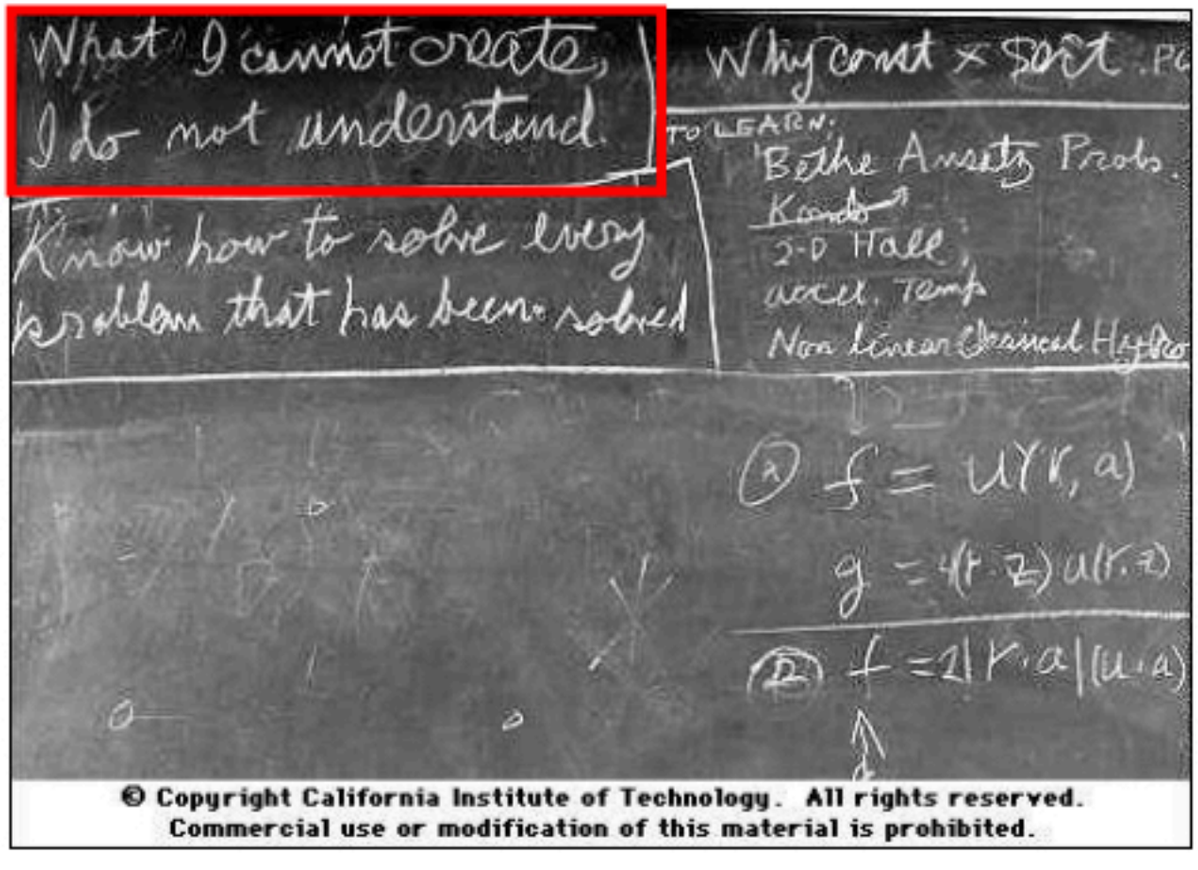
Possible topics

- Certification of deep learning
- Explainable AI
- GAN
  - Text to Image
  - CycleGAN for Steganography
- Domain learning
- Anomaly detection
- Deep reinforcement learning

Possible topics

Creation

- Generative Models: https://openai.com/blog/generative-models/



What I cannot create,
I do not understand.

Richard Feynman

https://www.quora.com/What-did-Richard-Feynman-mean-when-he-said-What-I-cannot-create-I-do-not-understand

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
  - Text to Image
  - CycleGAN for Steganography
- Domain learning
- Anomaly detection
- Deep reinforcement learning

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
  - Text to Image

this small bird has a pink breast and crown, and black primaries and secondaries.

  - CycleGAN for Steganography
- Domain learning
- Anomaly detection
- Deep reinforcement learning

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
  - Text to Image
  - CycleGAN for Steganograp
- Domain learning
- Anomaly detection
- Deep reinforcement learning

this small bird has a pink breast and crown, and black primaries and secondaries.

Source: https://arxiv.org/pdf/1605.05396.pdf

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
  - Text to Image
  - CycleGAN for Steganography
- Domain learning
- Anomaly detection
- Deep reinforcement learning

Possible topics

- Certification of deep
- Explainable AI
- GAN
  - Text to Image
  - CycleGAN for Steganography
- Domain learning
- Anomaly detection
- Deep reinforcement learning

(a) Aerial photograph: x. (b) Generated map: Fx. (c) Aerial reconstruction: GFx.

Figure 1: Details in x are reconstructed in GFx, despite not appearing in the intermediate map Fx.

Source:https://arxiv.org/pdf/1712.02950.pdf

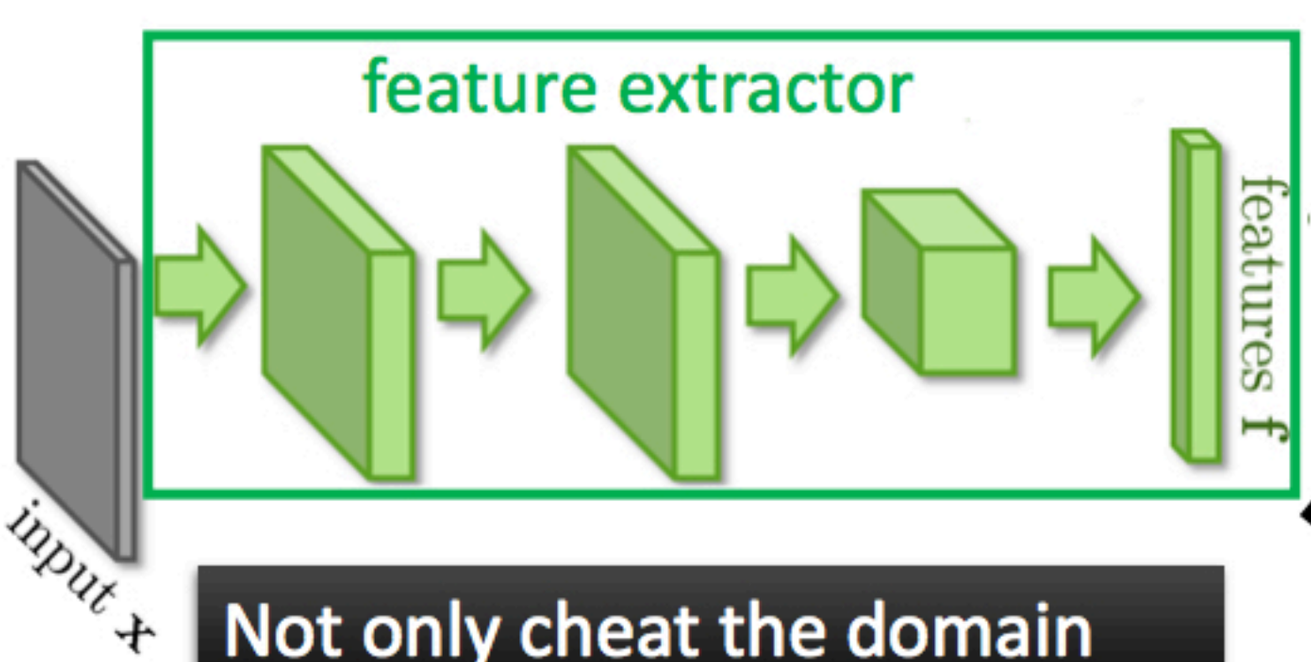
Possible topics

- Certification of deep learning
- Explainable AI
- GAN
- Domain learning
- Anomaly detection
- Deep reinforcement learning

Source:https://arxiv.org/pdf/1505.07818.pdf

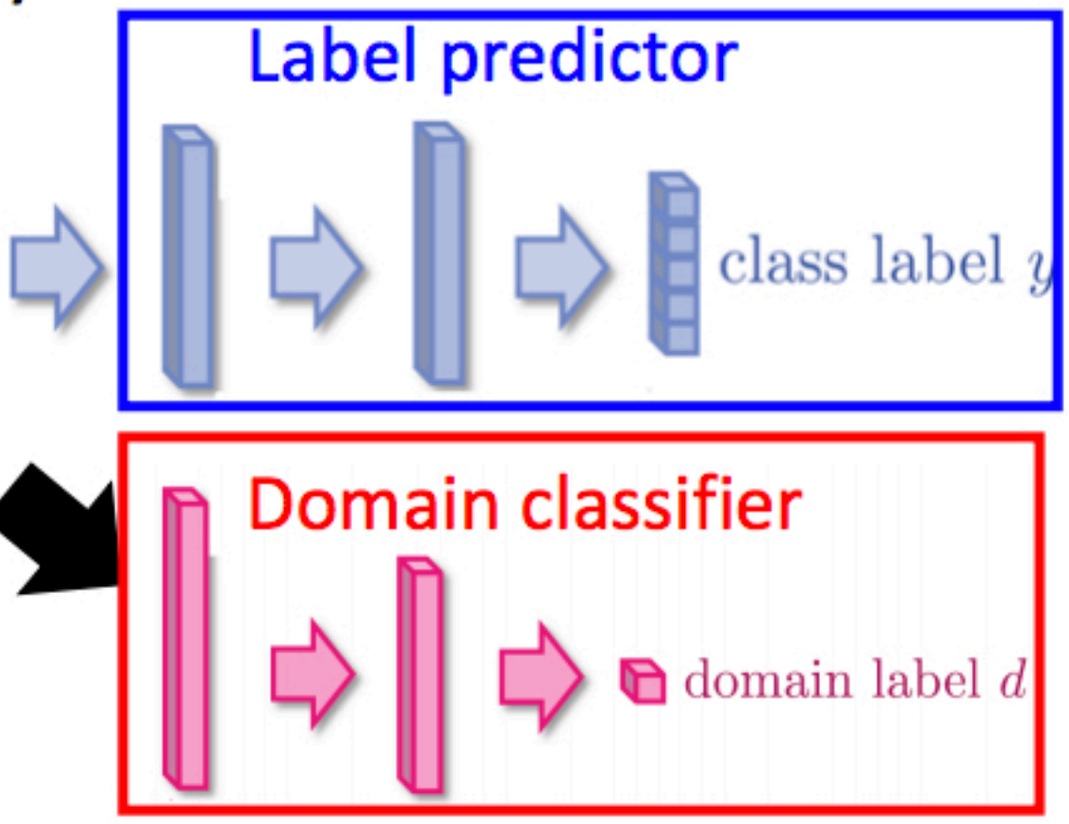
Domain-adversarial training

Maximize label classification accuracy + minimize domain classification accuracy



Not only cheat the domain classifier, but satisfying label classifier at the same time

Maximize label classification accuracy



Maximize domain classification accuracy

This is a big network, but different parts have different goals.

Source:https://arxiv.org/pdf/1505.07818.pdf

Possible topics

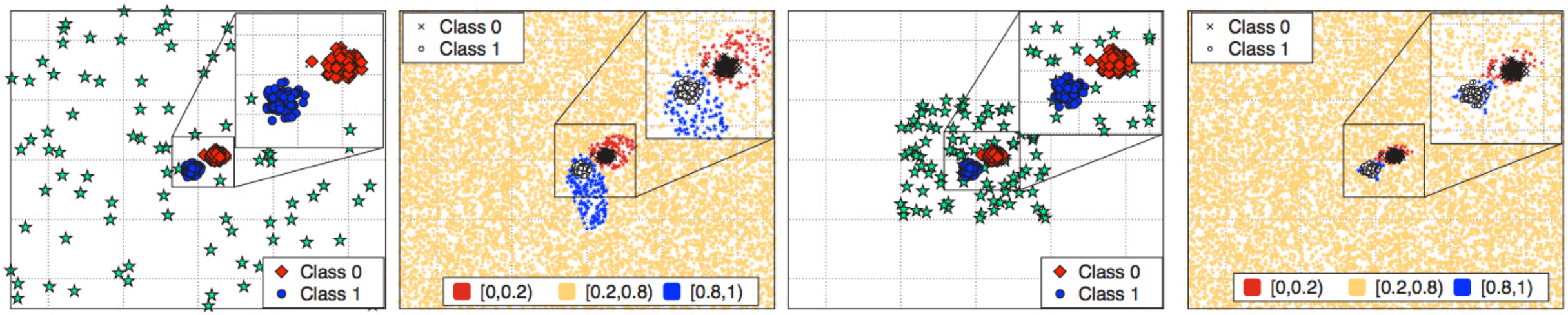
- Certification of deep learning
- Explainable AI
- GAN
- Domain learning
- Anomaly detection
- Deep reinforcement learning

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
- Domain learning
- Anomaly detection
  - Supervised Learning
  - Unsupervised Learning
- Deep reinforcement learning

Source:https://arxiv.org/pdf/1711.09325.pdf

Possible topics



(a) (b) (c) (d)

- Anomaly detection
  - Supervised Learning
  - Unsupervised Learning
- Deep reinforcement learning

Source:https://arxiv.org/pdf/1711.09325.pdf

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
- Domain learning
- Anomaly detection
  - Supervised Learning
  - Unsupervised Learning
- Deep reinforcement learning

Source:https://arxiv.org/pdf/1711.09325.pdf

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
- Domain learning
- Anomaly detection
  - Supervised Learning
  - Unsupervised Learning: Gaussian process, Autoencoder
- Deep reinforcement learning

Possible topics

- Certification of deep learning
- Explainable AI
- GAN
- Domain learning
- Anomaly detection
- Deep reinforcement learning

Questions?