

Explainable Artificial Intelligence (XAI): field and its methods

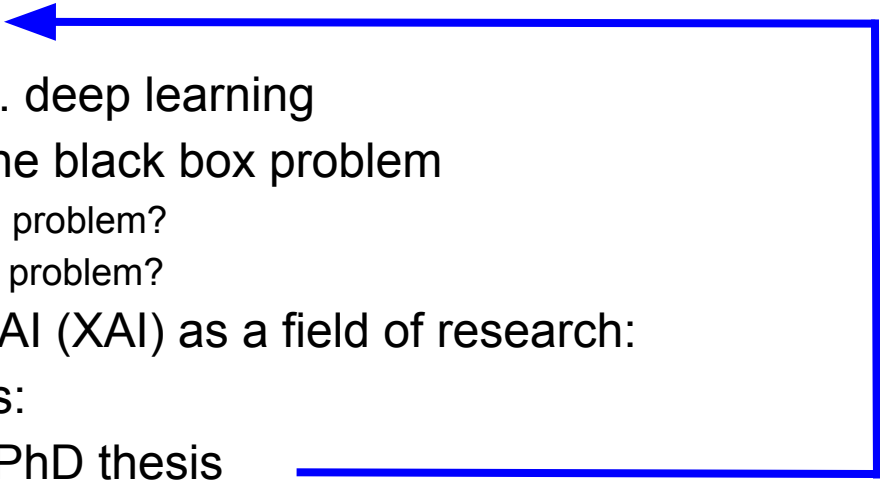
Maksims Ivanovs
University of Latvia & EDI, PhD student

Riga, 28 October 2020

Contents

- About me
- AI vs. ML vs. deep learning
- DNNs and the black box problem
 - *what's* the problem?
 - *why* is it a problem?
- Explainable AI (XAI) as a field of research:
- XAI methods:
- XAI <-> my PhD thesis

Contents

- About me
 - AI vs. ML vs. deep learning
 - DNNs and the black box problem
 - *what's* the problem?
 - *why* is it a problem?
 - Explainable AI (XAI) as a field of research:
 - XAI methods:
 - XAI <-> my PhD thesis
- 

About me

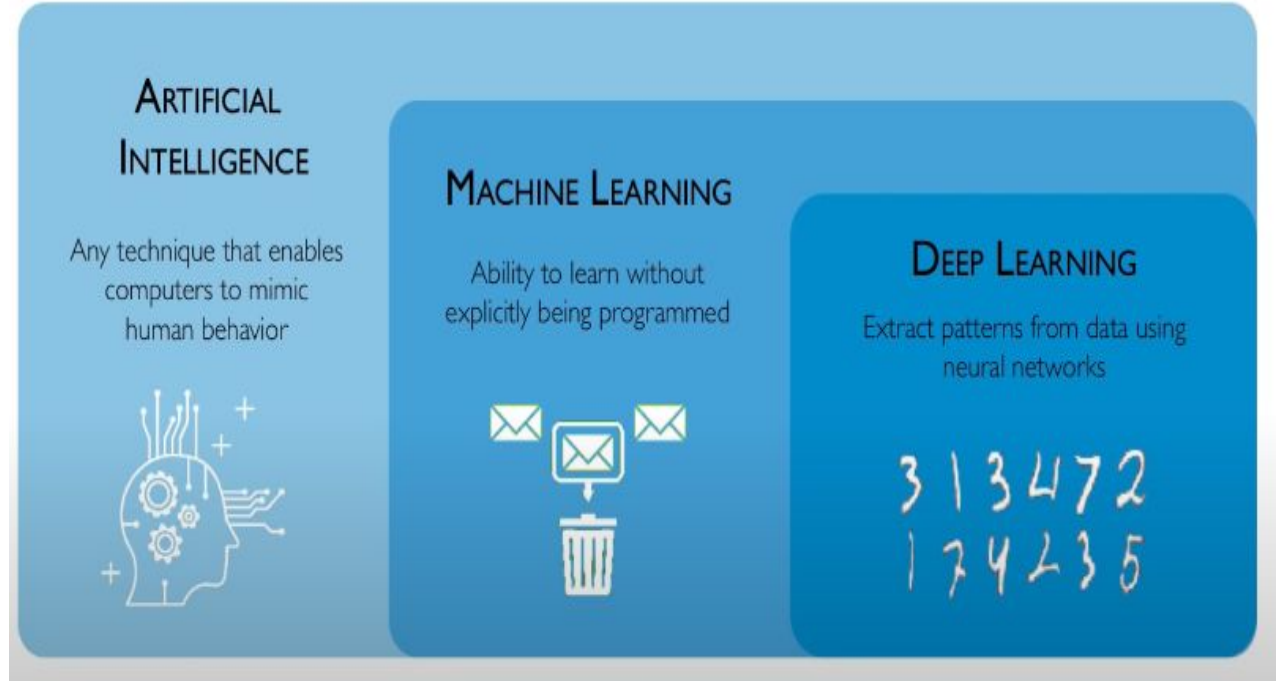
- 2nd year PhD student
- University of Latvia & EDI (Institute of Electronics and Computer Science)
- supervisor - *Dr. sc. ing.* Roberts Kadiķis, Head of Robotics and Machine Perception laboratory (EDI)
- PhD thesis topic: synthetic data generation for training deep neural networks (DNNs)
- Academic interests: synthetic data for deep learning, AI and society, explainable artificial intelligence (XAI)

Artificial intelligence: overview

- Impact of AI:
 - science
 - technology
 - industry
 - everyday life
- What is AI?

Artificial intelligence: overview

- Impact of AI:
 - science
 - technology
 - industry
 - everyday life
- What is AI?



Amini, MIT Introduction to Deep Learning

Artificial intelligence: overview

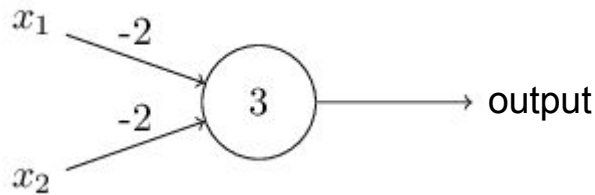
- Definition (Mitchell, 1997):

Definition: A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

- **Q: How can we do that?**
- **A: we can use artificial neural networks**

Artificial neural networks

Perceptron:



$$\text{output} = \begin{cases} 0 & \text{if } w \cdot x + b \leq 0 \\ 1 & \text{if } w \cdot x + b > 0 \end{cases}$$

00 -> 1

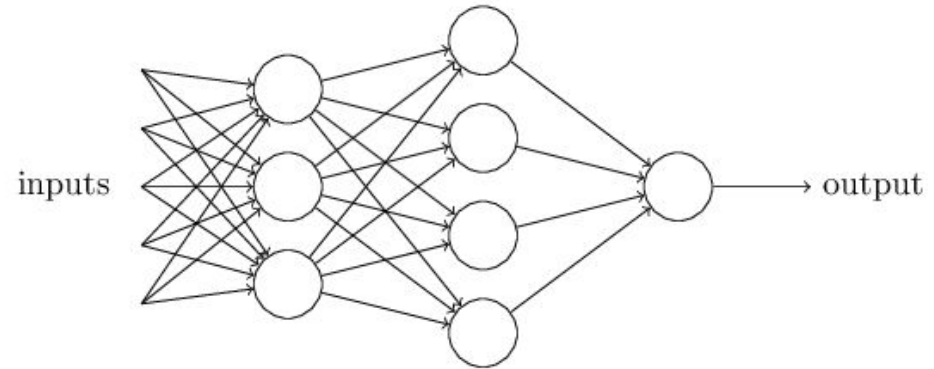
01 -> 1

10 -> 1

11 -> 0

NAND gate

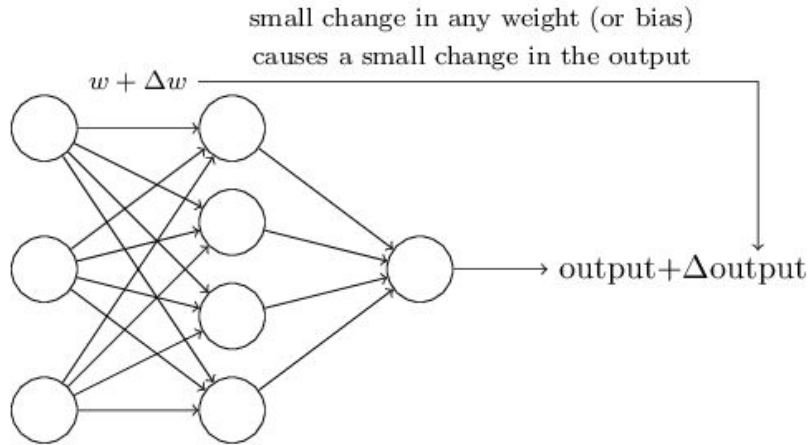
Network of perceptrons:



Such a network can deal with challenging problems, e.g. image classification. However, it has to be trained first.

Artificial neural networks

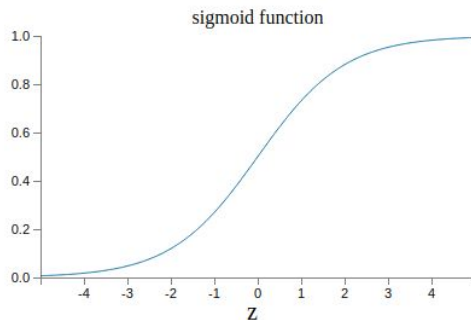
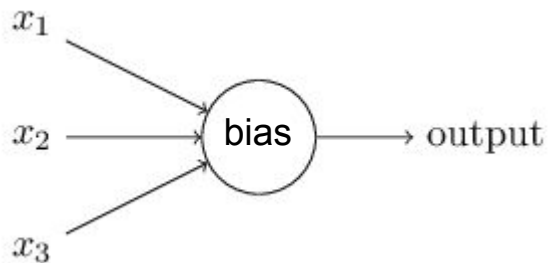
Network training scheme



Problem: the output of the perceptron is only 0 or 1

Artificial neural networks

Solution: sigmoid neuron



Most important property: smoothness

What's the output now?

$$\text{output} = \begin{cases} 0 & \text{if } w \cdot x + b \leq 0 \\ 1 & \text{if } w \cdot x + b > 0 \end{cases}$$

$$\Delta \text{output} \approx \sum_j \frac{\partial \text{output}}{\partial w_j} \Delta w_j + \frac{\partial \text{output}}{\partial b} \Delta b.$$

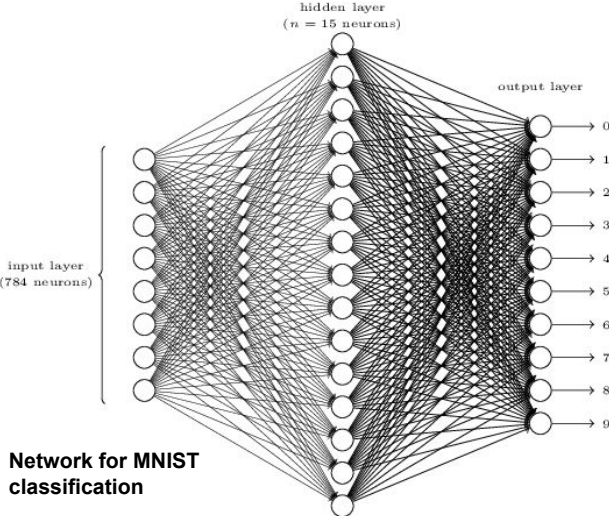
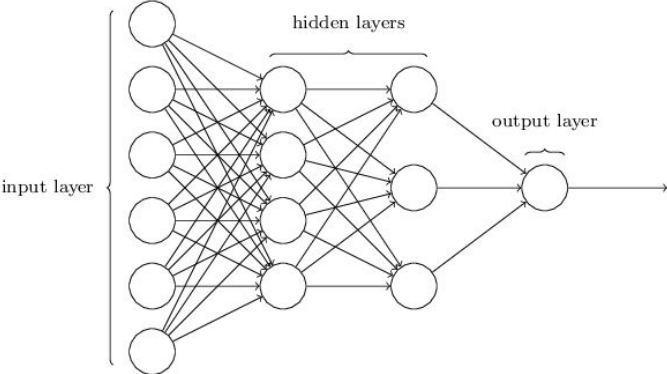
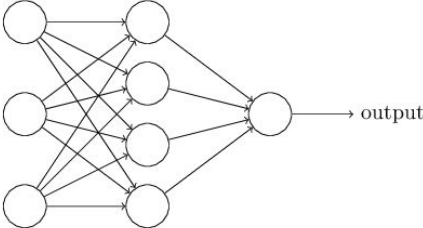
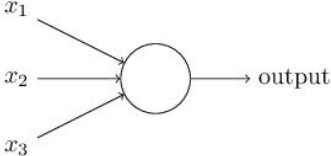
$$\sigma(w \cdot x + b),$$

σ is called the *sigmoid function*

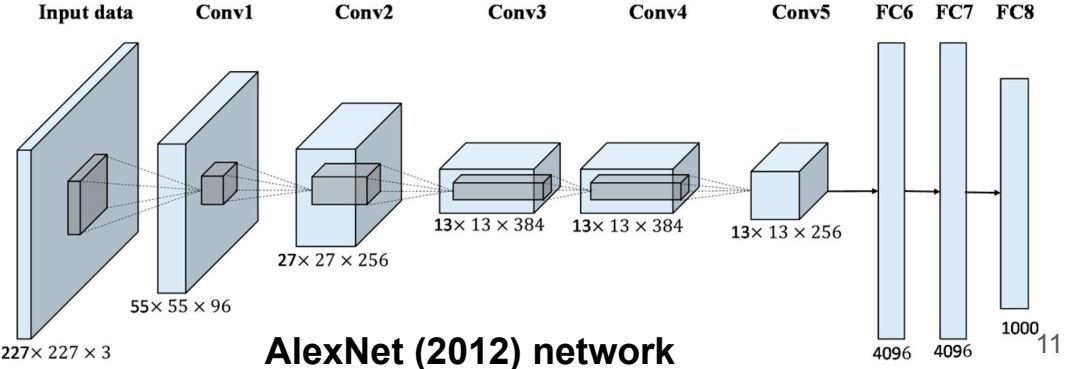
$$\sigma(z) \equiv \frac{1}{1 + e^{-z}} \equiv \frac{1}{1 + \exp(-\sum_j w_j x_j - b)}.$$

Deep artificial neural networks

Network architecture



Network for MNIST classification



AlexNet (2012) network

Deep learning

- Deep learning:
 - shallow neural networks - 1940s-1950s
 - originally inspired by the structure and functionality of the human brain
 - modern DNNs resemble the brain only (very) remotely
 - DNNs:
 - started to develop in the 1990s
 - breakthrough: Krizhevsky's *AlexNet* winning *ImageNet* competition in 2012
 - success factors: availability of:
 - hardware (GPUs)
 - training datasets (<- Internet)
 - new & more efficient algorithms

Deep learning

- DNNs:
 - especially suitable for the tasks involving perceptual data (e.g. images)
 - state-of-the-art-results: some examples:
 - image recognition
 - object detection
 - image segmentation
 - speech recognition
 - playing chess
 - playing go
 - driving autonomous vehicles
 - shortcomings:
 - require a lot of:
 - training data
 - computing power (mainly GPUs)
 - **lack of transparency**

DNNs and the black box problem

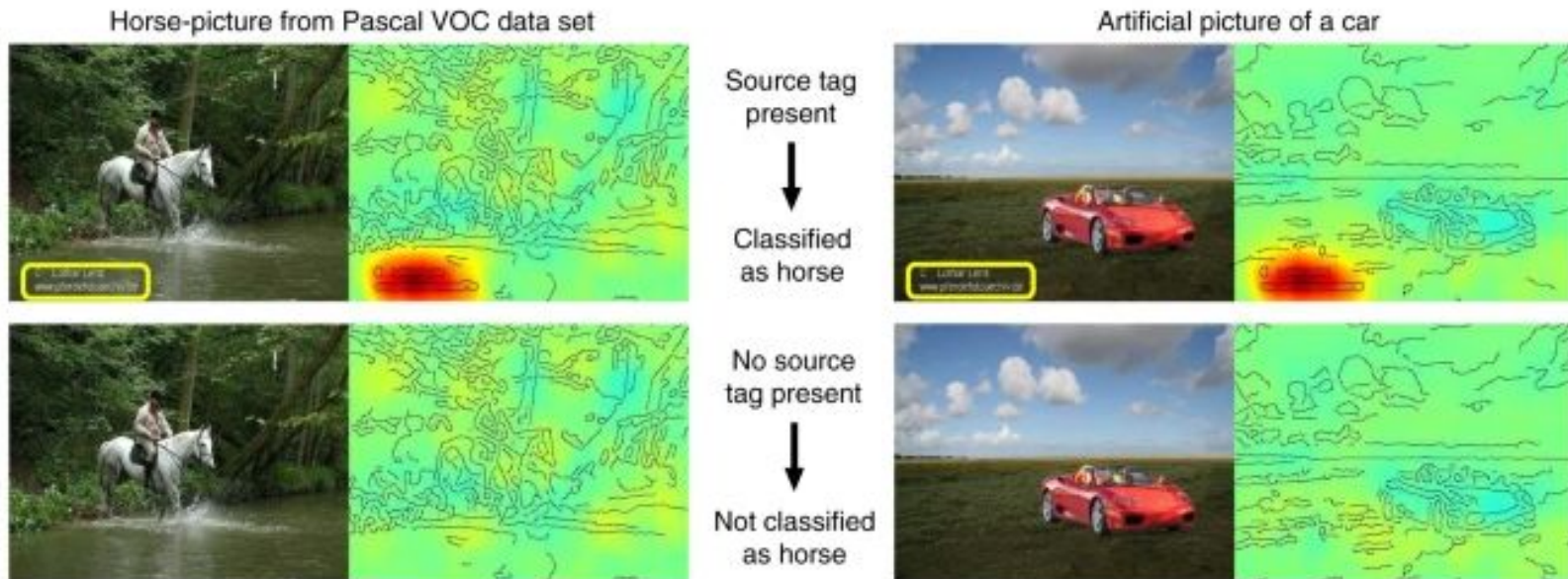
- DNN model is a function $f: X \rightarrow Y$ (X is an input space, Y is an output space)
- f is obtained by means of an opaque learning process (Fong & Vedaldi, 2017)
→ f is opaque itself
- **In other words: f is a black box**
- Terms:
 - black box
 - white box (also 'glass box', Holzinger et al., 2017)
 - gray box
- **Is it a problem!?**

DNNs and the black box problem

- Some safety-critical / high-stakes possible applications of DNNs:
 - medicine
 - driverless vehicles
 - power grid control
 - finance (e.g. mortgage applications assessment)
 - criminal justice (e.g. relapse risk assessment)
- **The black box problem makes it difficult to deploy DNNs for these purposes**
- Legal requirements: GDPR (in EU) -> right to explainability

DNNs and the black box problem

Example 1: *Clever Hans*



Lapuschkin et al, 2017

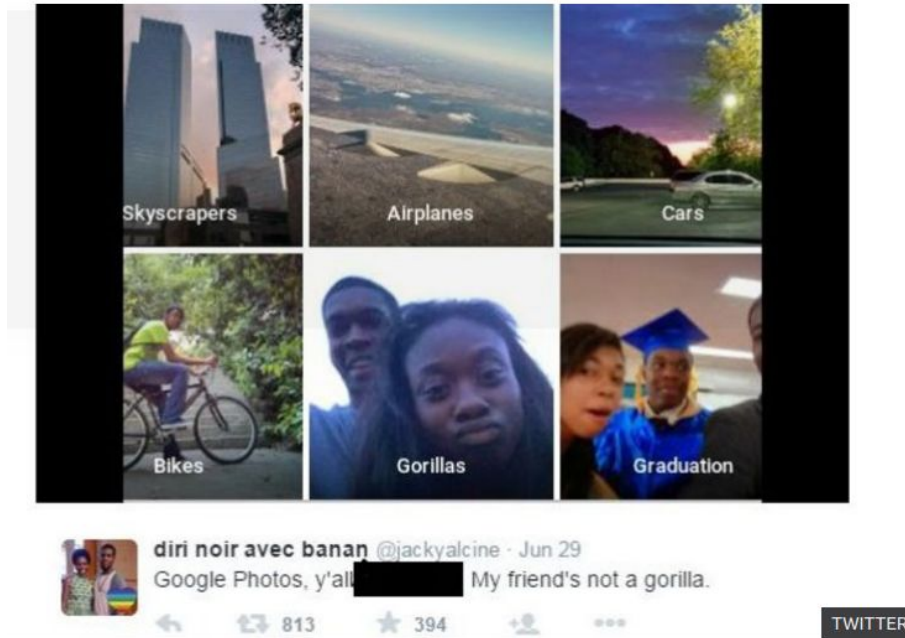
DNNs and the black box problem

Example 1: *Clever Hans*



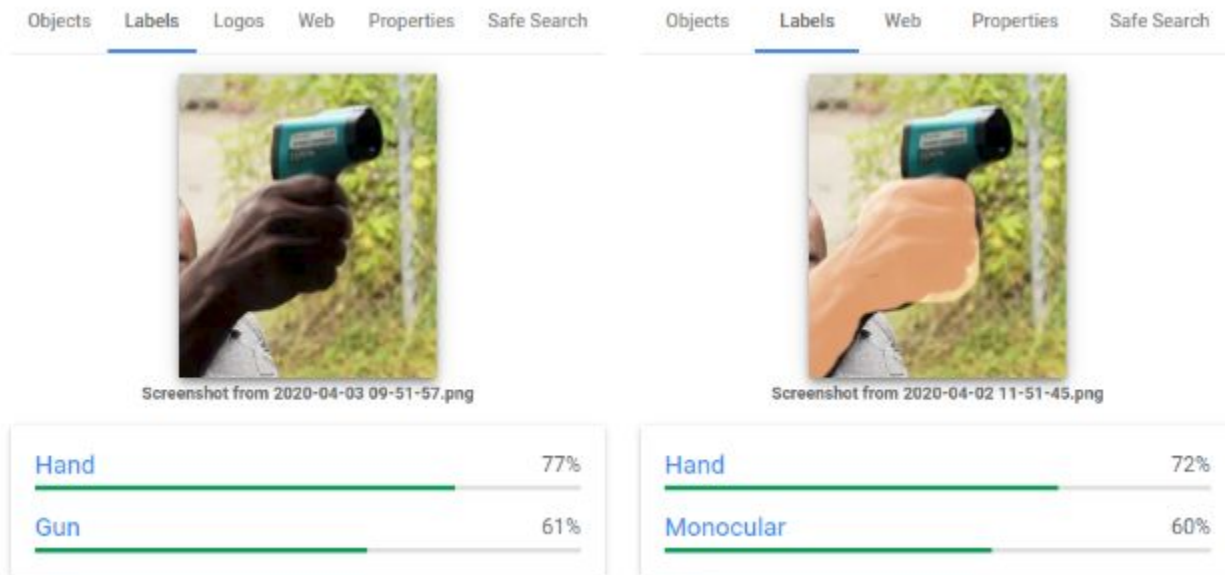
DNNs and the black box problem

Example 2: Racist Google algorithm



DNNs and the black box problem

Example 3: yet another racist Google algorithm



DNNs and the black box problem

Example 4: adversarial attacks (= “fool” a DNN into making a wrong decision)



Sivanami,
2019

DNNs and the black box problem

Example 4: adversarial attacks (= “fool” a DNN into making a wrong decision)



Sivanami,
2019

DNNs and the black box problem

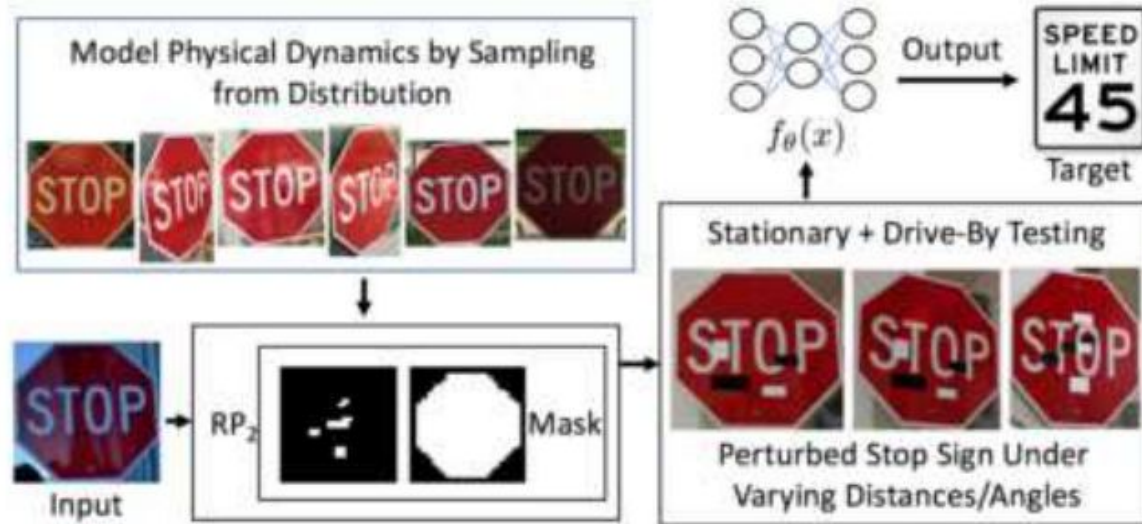
Example 4: adversarial attacks (= “fool” a DNN into making a wrong decision)



Sivanami,
2019

DNNs and the black box problem

Example 4: adversarial attacks (= "fool" a DNN into making a wrong decision)



Eykholt et al, 2018

XAI

- Response to the black box problem: explainable artificial intelligence (XAI)
- Terminology:
 - main discussion: is 'explainable' = 'interpretable'!?
 - other terms:
 - intelligible intelligent systems, context-aware systems, software learnability (Abdul et al., 2018);
 - responsible AI (Arrieta et al., 2020)
 - safe AI (Amodei et al., 2016)
 - terminology survey (Mohseni et al., 2018): 14 terms
 - all in all:
 - context-dependent terminology
 - no general agreement

XAI

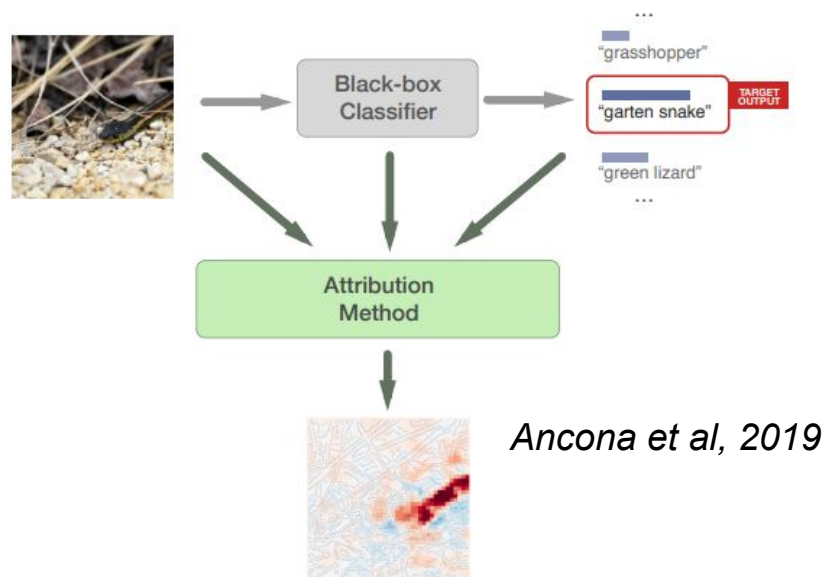
- Range of users: XAI experts <-> data science experts <-> novices
- Interdisciplinary field:
 - ML methods per se
 - visual analytics
 - human-computer interaction (HCI)
 - psychology
- Scope of the field:
 - nearly *any* AI research aims at explaining AI
 - adversarial attacks!?
 - Fong et al., 2019
 - Chalkiadakis, 2018

XAI

- Explaining the model:
 - transparent models vs post-hoc explainability
 - global interpretability vs local interpretability:
 - global: useful, but difficult to achieve, as N of interacting parameters keeps growing
 - local: model behavior is only explained for a single, specific instance
- Two main questions:
 - *what* model has learned?
 - *how* the model arrived to the prediction
- Fong & Vedaldi, 2019:
 - *what* model has learned -> what part of the input is important for inference
 - a.k.a. **input attribution methods**

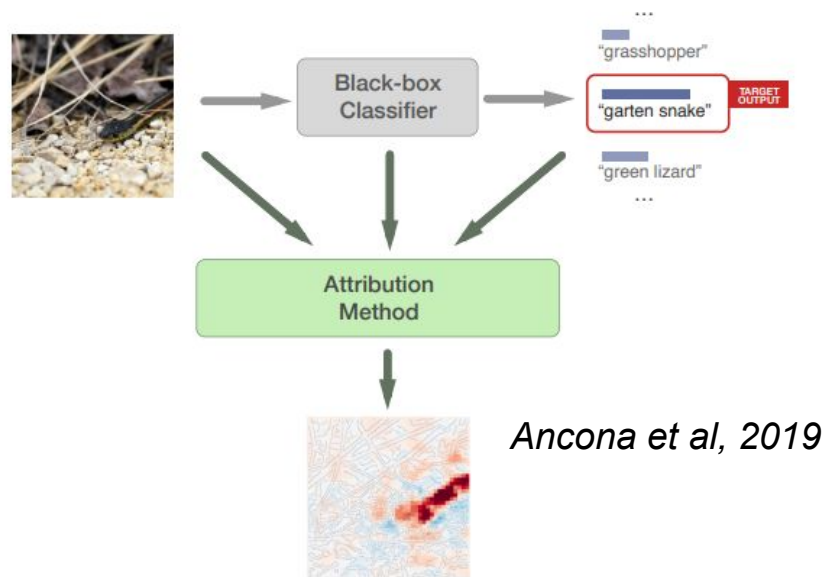
XAI: Input attribution methods

- Main tool: saliency maps
- Input attribution methods:
 - gradient-based
 - perturbation-based
- Gradient-based attribution methods:
 - use gradient as a proxy
 - **only single pass needed**
 - **model-dependent**
 - **noisy**
 - **some methods can't pass sanity check**



XAI: Input attribution methods

- Main tool: saliency maps
- Input attribution methods:
 - gradient-based
 - perturbation-based
- Perturbation-based attribution methods:
 - modify input -> observe changes in the output
 - -> modify input -> observe changes in the output
 - natural & dynamic
 - model-agnostic
 - applicable to images, videos, texts, software code, RL agents...
 - main problem: combinatorial explosion
 - main challenge: find out the optimal scope and shape of perturbations



XAI <-> my PhD thesis

- Thesis-related publications so far:
 - Skadins, A., Rava, R., [Ivanovs, M.](#), Nesenbergs, K. (2020). Edge pre-processing of traffic surveillance video for bandwidth and privacy optimization in smart cities. 17th Biennial Baltic Electronics Conference (BEC2020) Tallin, Estonia. [best paper award]
 - Rava, R., [Ivanovs, M.](#), Skadins, A., Nesenbergs, K. (2020). World coordinate virtual traffic cameras: edge-based transformation and merging of multiple surveillance video sources [accepted for ISCOMI2020 conference]
- Work in progress: 2 publications on synthetic data:
 - semantic segmentation for self-driving cars
 - gesture recognition
- ToDo: use of XAI for validating synthetic data

Thank you for your attention!