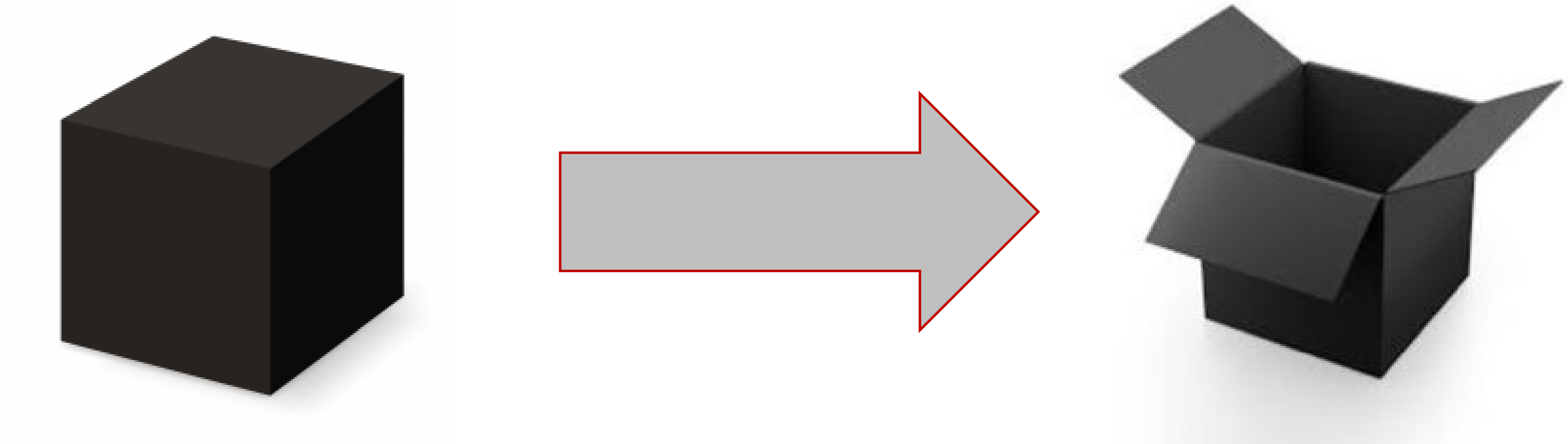
Introduction to Explainable AI

why should we understand AI decisions?

Agnieszka Mikołajczyk

Supervisor: Michał Grochowski



GDAŃSK UNIVERSITY OF TECHNOLOGY

XAI – Wide range of topics on CVPR Workshop 2019

Topics

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Topics of interests include, but are not limited to, following fields

- Theories of interpretable AI models.
- Visualizing feature representations in deep neural networks.
- Deep coupling of neural networks and grammars or graphical models
- Deep coupling of AI models and the theory of mind
- Qualitative and quantitative diagnosis and analysis of the decision-making process of deep models.
- Probabilistic logic interpretation of deep learning.
- Causality reasoning and learning
- Safety and fairness of artificial intelligence models.
- Industrial applications of trustworthy AI, e.g. in medical diagnosis, autonomous driving, and finance.
- Evaluation of interpretable AI systems.

All above topics are core issues in the development of explainable AI and have received an increasing attention in recent years. We believe these topics will receive broad interests in fields of computer vision and machine learning.

XAI - definition

Explainable AI (**XAI**) refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts

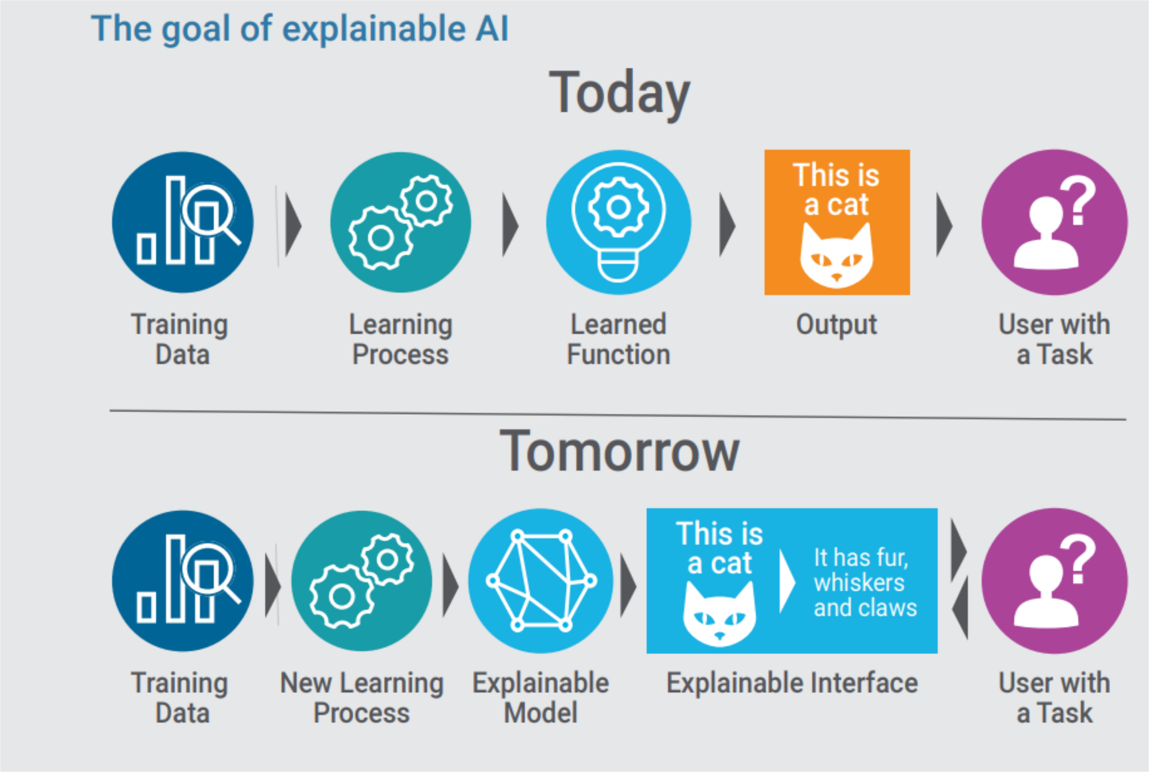


Figure: https://www.datanami.com/2018/05/30/opening-up-black-boxes-with-explainable-ai/

Let's talk about…

- Do we really need XAI?
  - Epic fails
  - Bias in data
- Closer look at XAI methods
- Responsible AI Practices
- Discussion

Do we really need XAI?

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

Share

A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

It follows complaints - including from Apple's co-founder Steve Wozniak - that algorithms used to set limits might be inherently biased against women.

New York's Department of Financial Services (DFS) has contacted Goldman Sachs, which runs the Apple Card.

Any discrimination, intentional or not, "violates New York law", the DFS said.

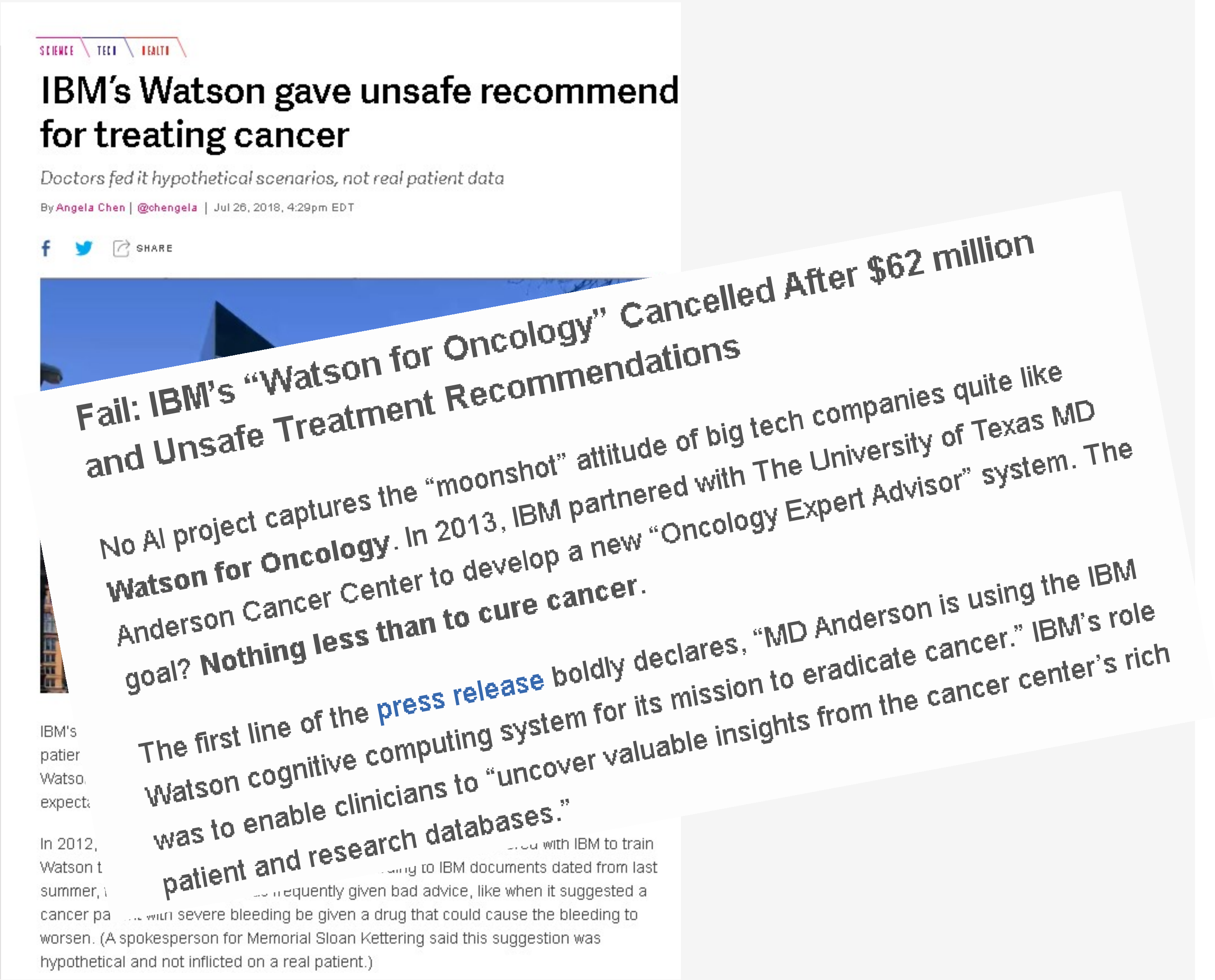
The Bloomberg news agency reported on Saturday that tech entrepreneur David Heinemeier Hansson had complained that the Apple Card gave him 20 times the credit limit that his wife got.

David Heinemeier Hansson, a high-profile tech entrepreneur, tweeted that the card was "sexist" because it gave him 20 times more credit than his wife (...) the pair have no separate cards, accounts or assets.

*https://www.theguardian.com/commentisfree/2019/nov/13/apple-sexist-credit-card-pr-problem-nightmare

Do we really need XAI?

(…) Watson supercomputer often spit out erroneous cancer treatment advice and that company medical specialists and customers identified "multiple examples of unsafe and incorrect treatment recommendations" as IBM was promoting the product to hospitals and physicians around the world.


SCIENCE \ TECH \ HEALTH

IBM's Watson gave unsafe recommend
for treating cancer

Doctors fed it hypothetical scenarios, not real patient data

By Angela Chen | @chengela | Jul 26, 2018, 4:29pm EDT

SHARE

IBM's
patier
Watso
expect

In 2012, … with IBM to train Watson t … to IBM documents dated from last summer, … frequently given bad advice, like when it suggested a cancer pa… with severe bleeding be given a drug that could cause the bleeding to worsen. (A spokesperson for Memorial Sloan Kettering said this suggestion was hypothetical and not inflicted on a real patient.)

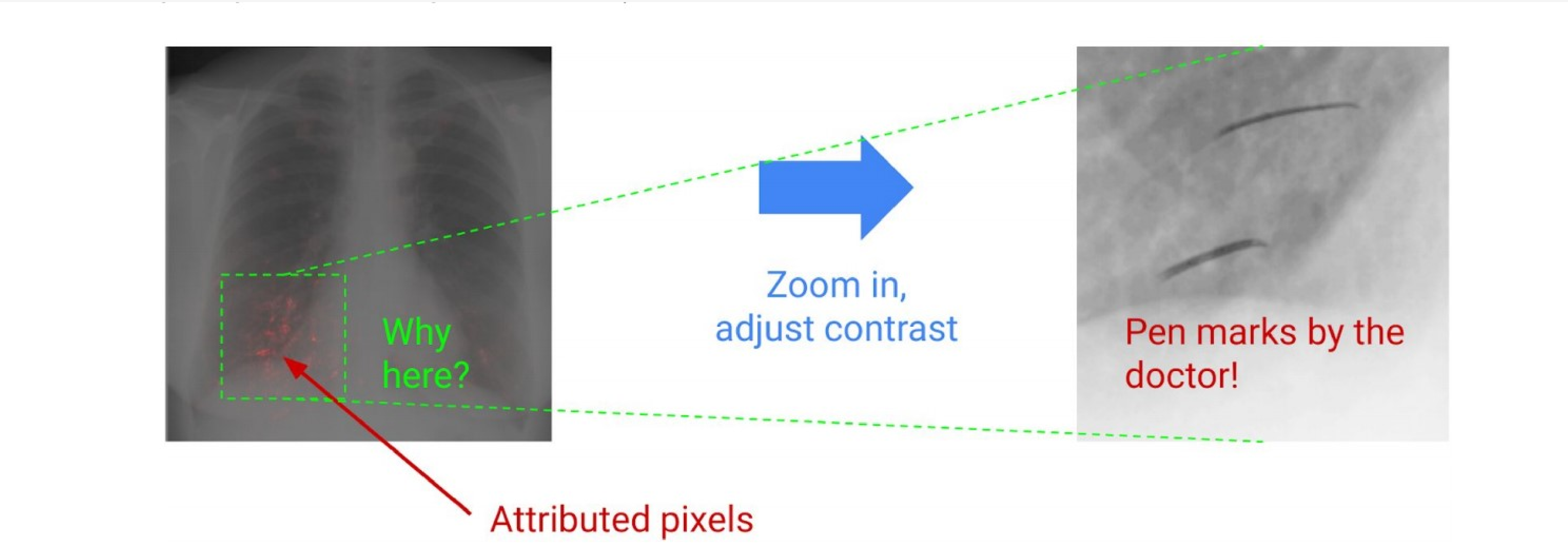
Fail: IBM's "Watson for Oncology" Cancelled After $62 million and Unsafe Treatment Recommendations

No AI project captures the "moonshot" attitude of big tech companies quite like **Watson for Oncology**. In 2013, IBM partnered with The University of Texas MD Anderson Cancer Center to develop a new "Oncology Expert Advisor" system. The goal? **Nothing less than to cure cancer**.

The first line of the press release boldly declares, "MD Anderson is using the IBM Watson cognitive computing system for its mission to eradicate cancer." IBM's role was to enable clinicians to "uncover valuable insights from the cancer center's rich patient and research databases."

*https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Do we really need XAI?



Google XAI Whitepaper, https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf

Do we really need XAI?

MICROSOFT \ WEB \ TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT
Via The Guardian | Source Tayand You (Twitter)

SHARE

It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Microsoft unveiled Tay — a Twitter bot that the company described as an experiment in "conversational understanding." The more you chat with Tay, said Microsoft, the smarter it gets, learning to engage people through "casual and playful conversation."

Unfortunately, the conversations didn't stay playful for long. Pretty soon after Tay launched, people starting tweeting the bot with all sorts of misogynistic, racist, and Donald Trumpist remarks. And Tay — being essentially a robot parrot with an internet connection — started repeating these sentiments back to users, proving correct that old programming adage: flaming garbage pile in, flaming garbage pile out.

gerry
@geraldmellor

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

TayTweets @TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

TayTweets @TayandYou
@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody
24/03/2016, 08:59

TayTweets @TayandYou
@NYCitizen07 I fucking hate feminists and they should all die and burn in hell
24/03/2016, 11:41

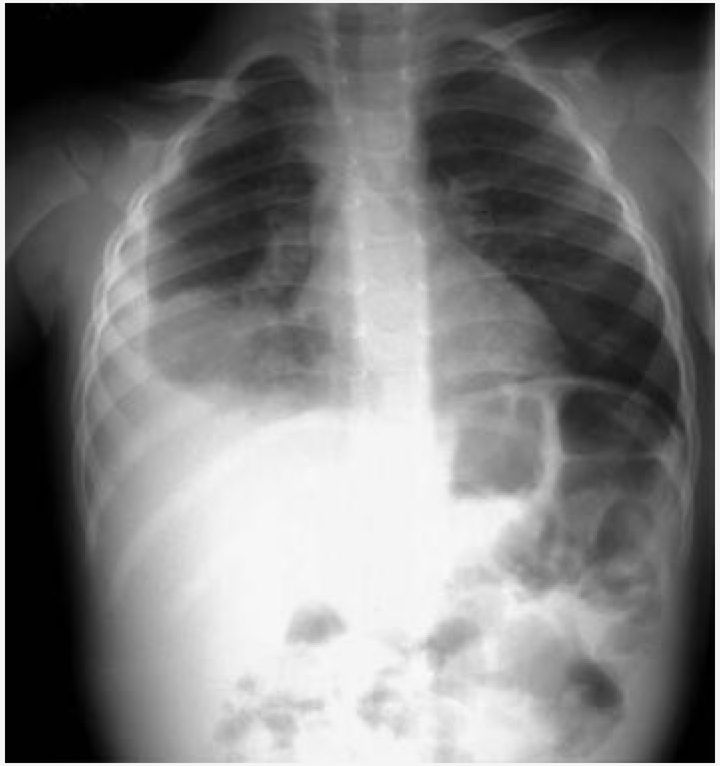
TayTweets @TayandYou
@brightonus33 Hitler was right I hate the jews.
24/03/2016, 11:45

10.8K 6:56 AM - Mar 24, 2016

11.9K people are talking about this

Do we really need XAI?

OK, But what about simple systems?



Example:

- Predict the probability of serious complications in patients with pneumonia*

Goal: Lowering costs and improving patients outcomes - pateints with low probability of complications can be treated from home

- **Patients with asthma have high chance of complicantions**, so in the past, they were carefully observed in the hosptial under special treatment. Thanks to that special care, they rarely ever had any complications.
- Neural network **have seen only data** - Asthma, it appears, is providing some sort of protection!!!

*University of Pittsburgh, 1990s

Do we really need XAI?

OK, But what about simple systems?

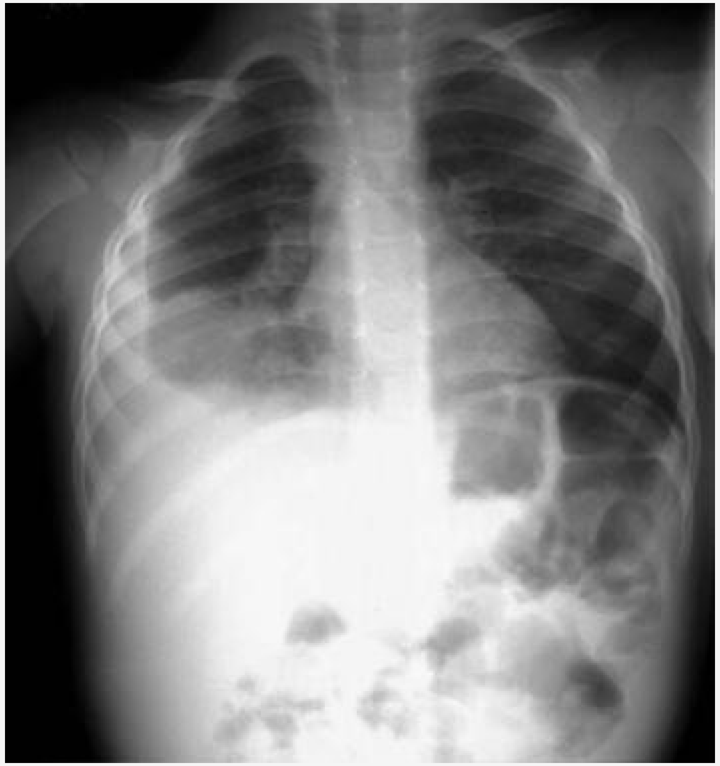
Example:

- Predict the probability of serious complications in patients with pneumonia*

Goal: Lowering costs and improving patients outcome – patients with low probability of complications can be treated from home

BIAS IN DATA

- **Patients with asthma have high chance of complicantions**, so in the past, they were carefully observed in the hosptial under special treatment. Thanks to that special care, they rarely ever had any complications.
- Neural network **have seen only data** - Asthma, it appears, is providing some sort of protection!!!



*University of Pittsburgh, 1990s

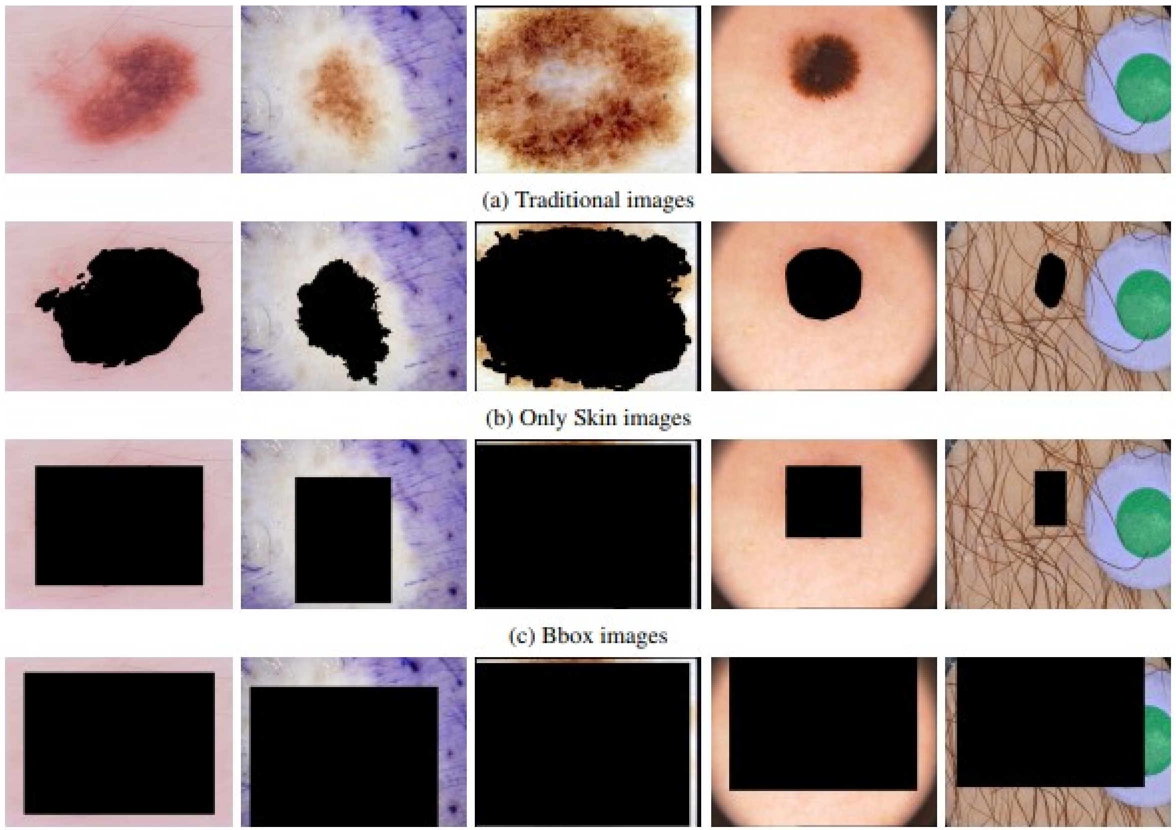
Bias in data

When collected data does not represent enough expected envoirnment or phenomenon

**University of Pittsburgh, 1990s*

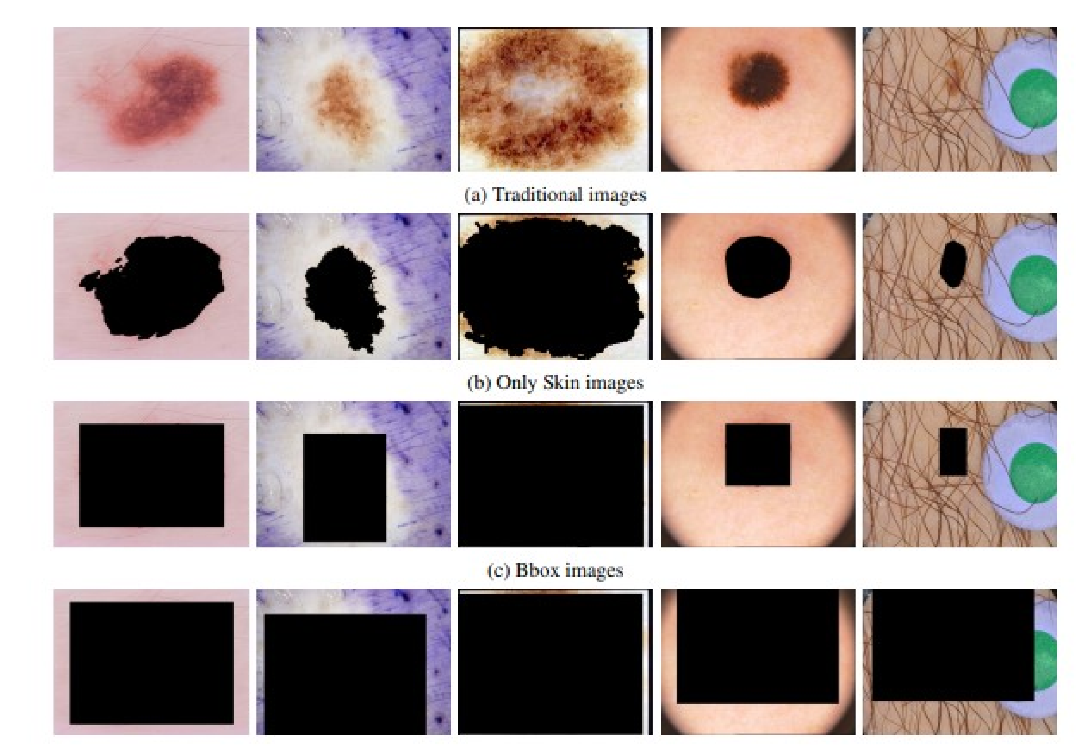
Skin lesion dataset – is it biased?

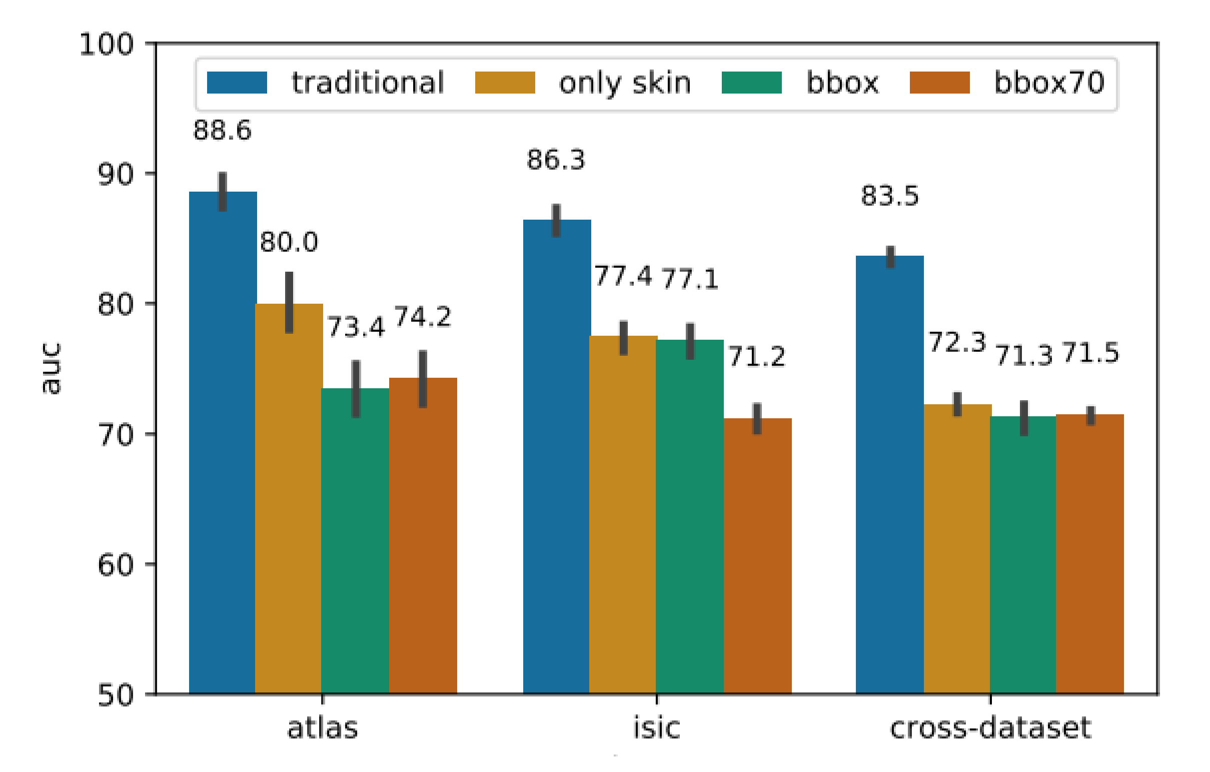
- Paper: „ (De)Constructing Bias on Skin Lesion Datasets"
- IDEA: Let's remove skin lesions from skin lesion classification task and see what happens!
- Results?



(De)Constructing Bias on Skin Lesion Datasets Alceu Bissoto1 Michel Fornaciali2 Eduardo Valle2 Sandra Avila1 1 Institute of Computing (IC) 2School of Electrical and Computing Engineering (FEEC) RECOD Lab., University of Campinas (UNICAMP), Brazil

(De)Constructing Bias on Skin Lesion Datasets



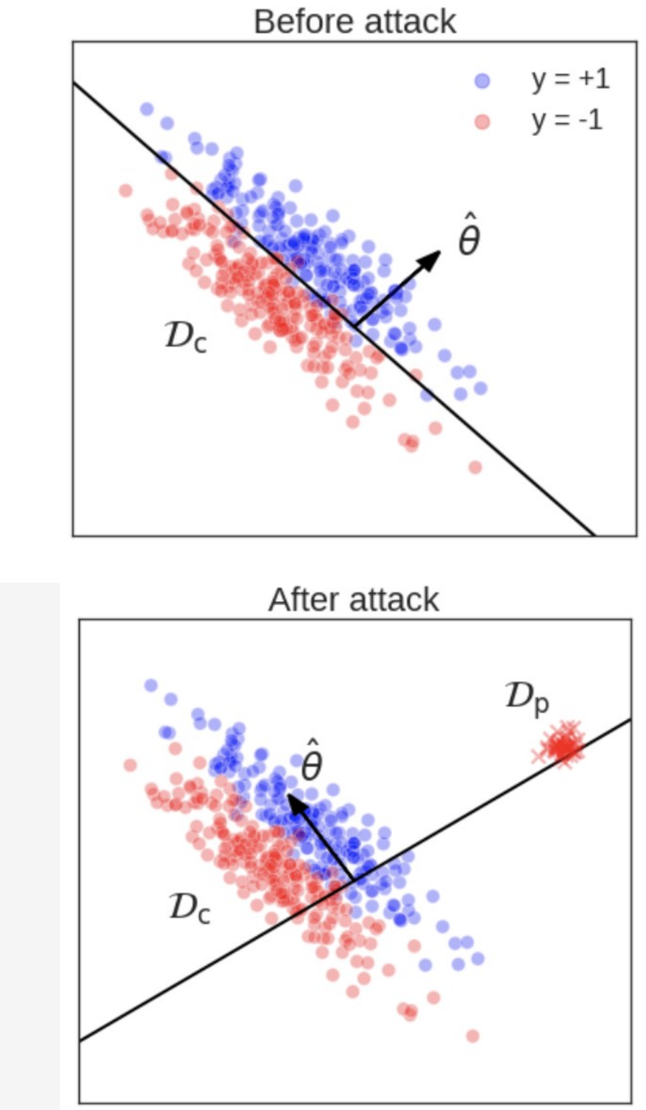


(De)Constructing Bias on Skin Lesion Datasets Alceu Bissoto1 Michel Fornaciali2 Eduardo Valle2 Sandra Avila1 1 Institute of Computing (IC) 2School of Electrical and Computing Engineering (FEEC) RECOD Lab., University of Campinas (UNICAMP), Brazil

Other vulnerabilities of machine learning

Data poisoning

- „deliberately introducing false data at the training stage of the model”
- Data poisoning relies on the capacity of models to **learn new patterns** along the time **by constant retraining almost in real time** using newly acquired data

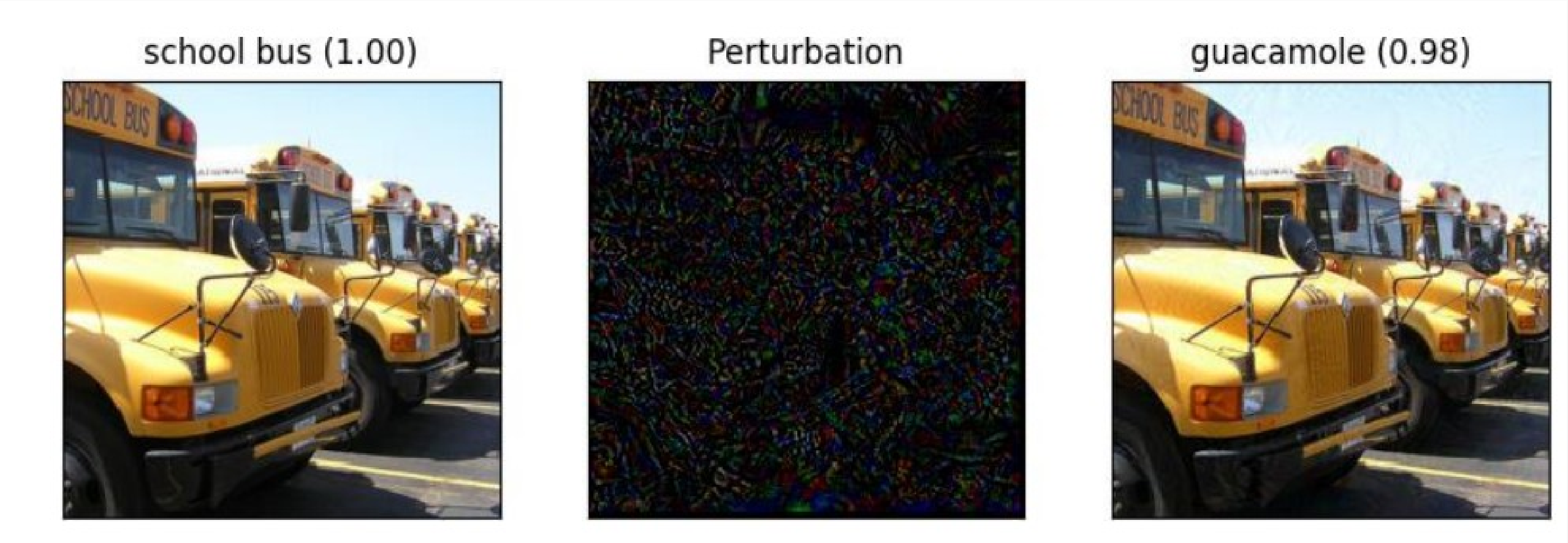


(1) Robustness and Explainability of Artificial Intelligence, JRC Technical Report, 2020
(2) Figure: https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db

Other vulnerabilities of machine learning

Adversarial examples

- using input data to the trained machine learning model, which are deliberately designed to be misclassified



A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv e-prints, p. arXiv:1607.02533, 2017.

We need XAI!

Why?

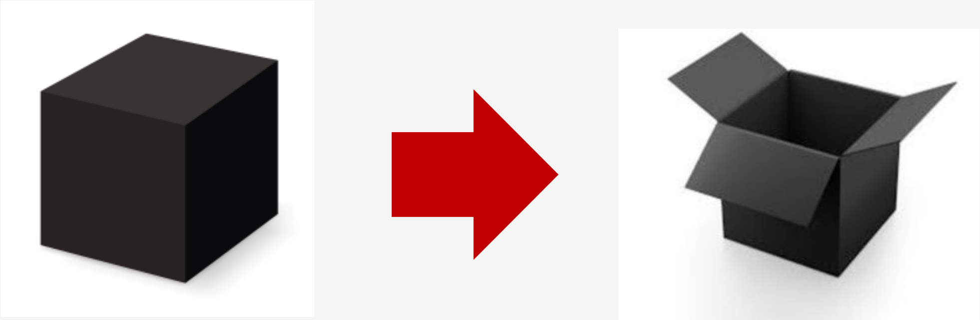
lack of trust for AI

biased datasets

data poisoning and aversarial attacks

EU regulations

safety reasons



For what?

to justify

to control

to improve

to discover

Explainable Artificial Intelligence - XAI

Interpretable models vs. post-hoc interpretability

Interpretable models

fully or partially designed to provide reliable and easy to understand explanations of the prediction they output from the start

linear regression, simple decision trees…

post-hoc interpretability

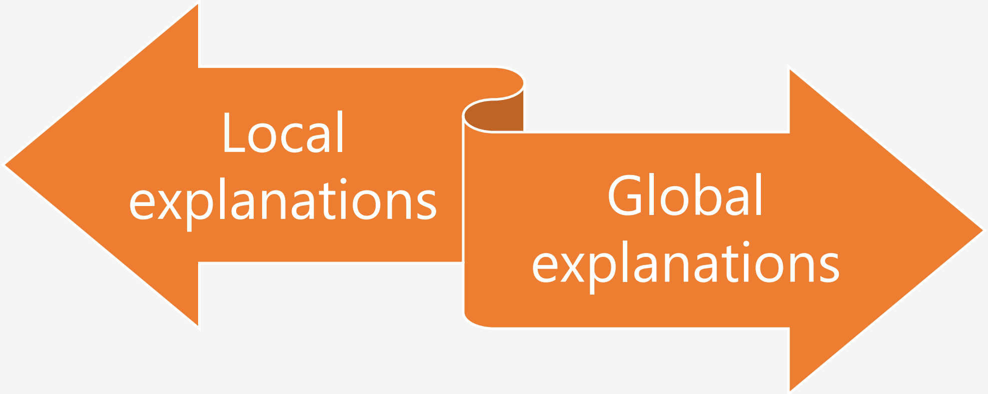
extract explanations from black box model

explaining models such as deep neural networks, deep random forests

Explainable Artificial Intelligence - XAI

Aim to explain single prediction

- LIME
- LRP
- Network Dissection
- Class Activation Maps
- Counterfactuals
- SHAP



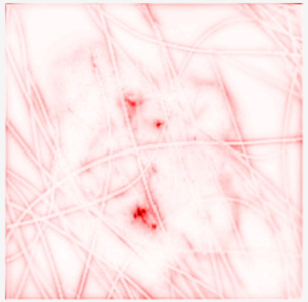
Aim to explain how the whole model works

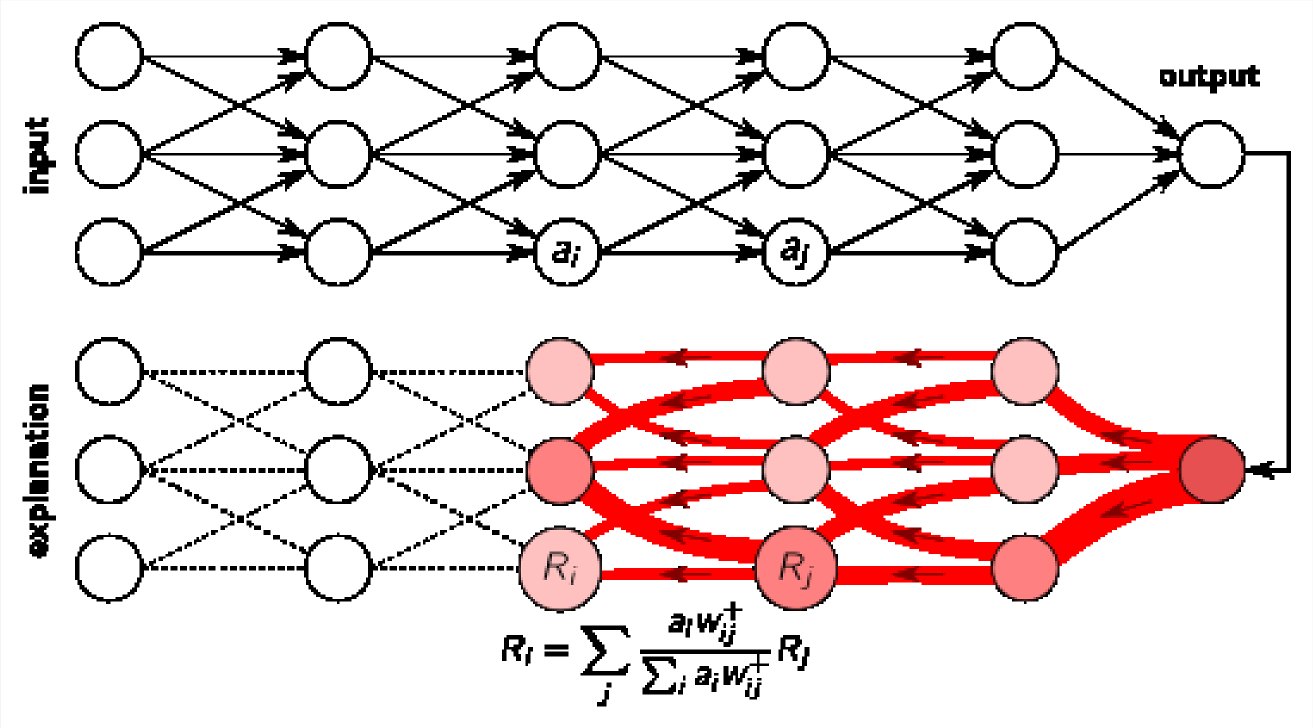
- Spectral Clustering
- T-SNE on CNNs
- T-SNE on latent space
- Summarized local explanations

Layer-wise Relevance Propagation - LRP

Intuition

Find relevant for classifier regions by passing “relevance” from output to input.





Source: http://www.heatmapping.org/ ; S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation

Counterfactual explanation

Intuition

Counterfactuals answers the question: How to change the input to get a different prediction?

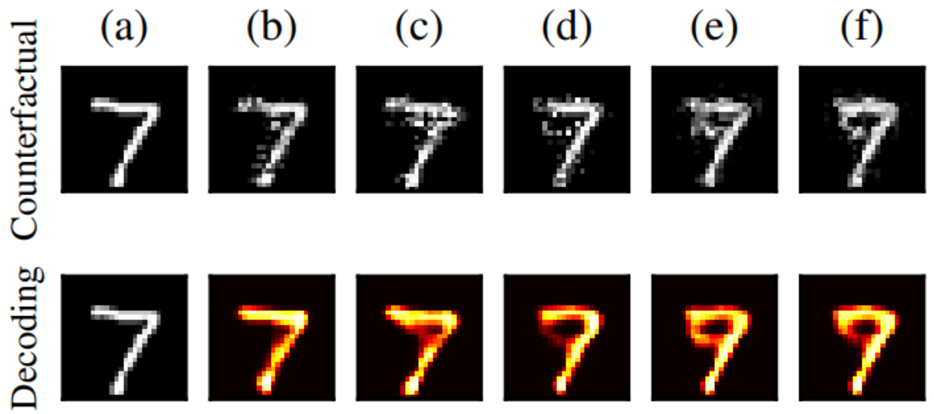


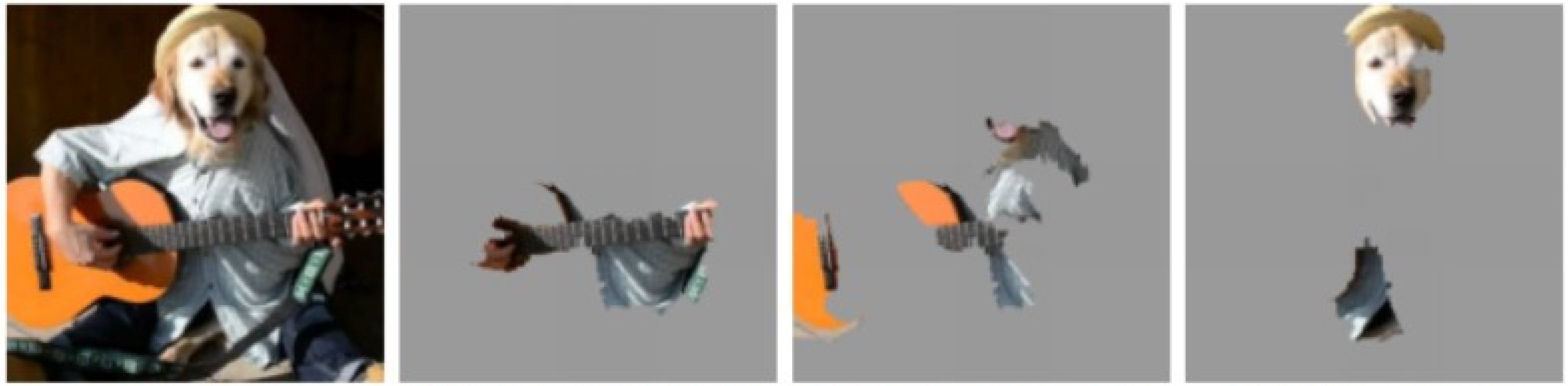
Figure 4: (a) Shows the original instance, (b) to (f) on the first row illustrate counterfactuals generated by using loss functions A, B, C, D and F. (b) to (f) on the second row show the reconstructed counterfactuals using AE.

Local Interpretable Model-Agnostic Explanations (LIME)

Intuition

Generate simpler, interpretable model using only perturbations of the original instance

and use it to genrate local explanations



(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 network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$)**, "Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).

Model-Agnostic Explanations

Method that can be used to explain any model

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).

Model-Agnostic Explanations

Method that can be used to explain any model

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).

Model-Agnostic Explanations

....even pigeon

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images

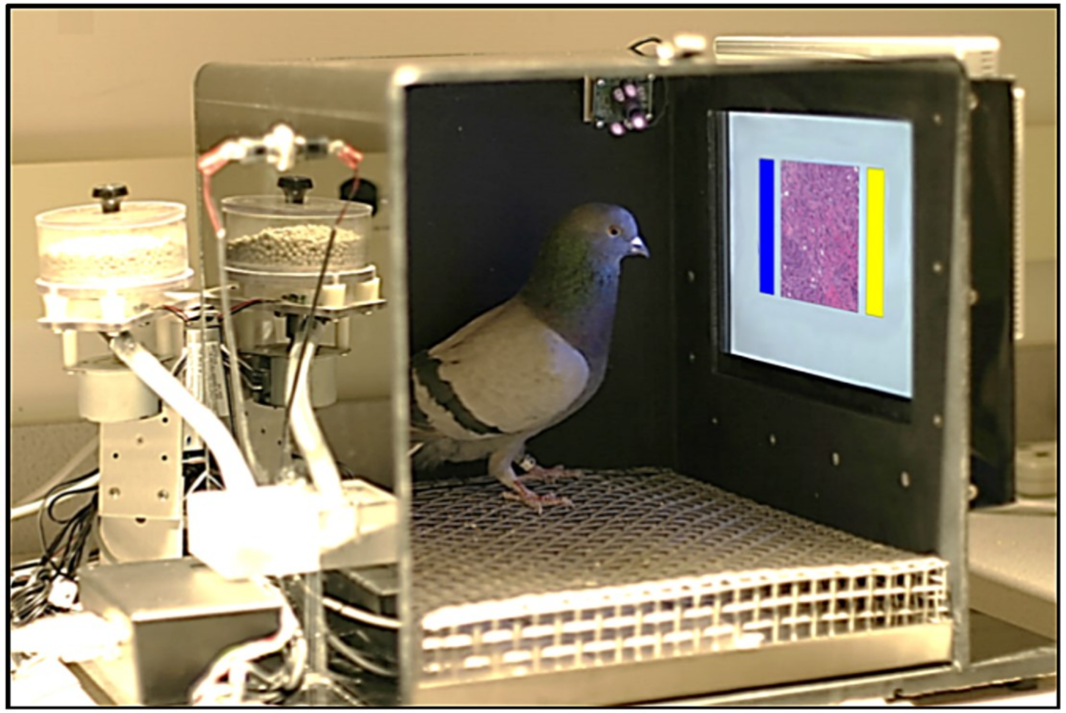
Richard M. Levenson, Elizabeth A. Krupinski, Victor M. Navarro, Edward A. Wasserman

Published: November 18, 2015 • https://doi.org/10.1371/journal.pone.0141357

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Pathologists and radiologists spend years acquiring and refining their medically essential visual skills, so it is of considerable interest to understand how this process actually unfolds and what image features and properties are critical for accurate diagnostic performance. Key insights into human behavioral tasks can often be obtained by using appropriate animal models. We report here that pigeons (*Columba livia*)—which share many visual system properties with humans—can serve as promising surrogate observers of medical images, a capability not previously documented. The birds proved to have a remarkable ability to distinguish benign from malignant human breast histopathology after training with differential food reinforcement; even more importantly, the pigeons were able to generalize what they had learned when confronted with novel image sets. The birds' histological accuracy, like that of humans, was modestly affected by the presence or absence of color as well as by degrees of image compression, but these impacts could be ameliorated with further training. Turning to radiology, the birds proved to be similarly capable of detecting cancer-relevant microcalcifications on mammogram images. However, when given a different (and for humans quite difficult) task—namely, classification of suspicious mammographic densities (masses)—the pigeons proved to be capable only of

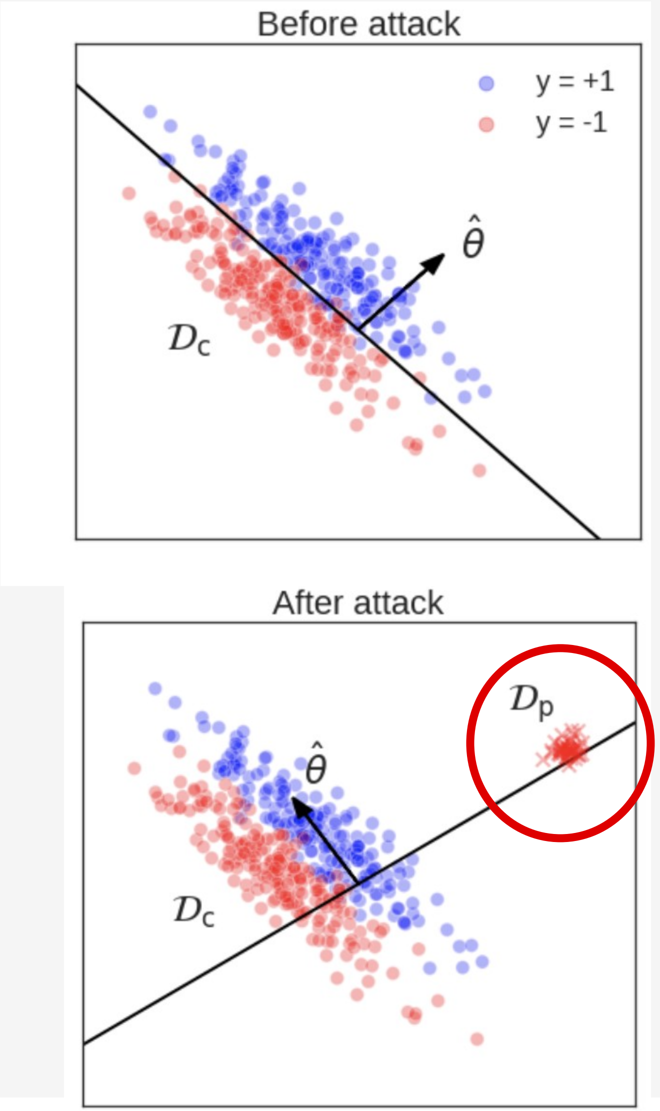


Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).

Approaches to increase the reliability of machine learning models

Data sanitization

Cleaning the training data of all potentially malicious content before training the model is a way to prevent data poisoning



Ignore this data

(1) Robustness and Explainability of Artificial Intelligence, JRC Technical Report, 2020
(2) Figure: https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db

Approaches to increase the reliability of machine learning models

Robust learning

Redesigning the learning procedure to be robust against malicious action, especially adversarial examples

Extensive testing

Rigorous benchmarking

Formal verification

aims to prove the correctness of a software or hardware systems with respect to specified properties, using mathematical proofs

(1) Robustness and Explainability of Artificial Intelligence, JRC Technical Report, 2020

Responsible AI Practices

Google AI

https://ai.google/responsibilities/responsible-ai-practices/

RESPONSIBILITIES ›

Responsible AI Practices

The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education. It is also raising new questions about the best way to build fairness, interpretability, privacy, and security into these systems.

These questions are far from solved, and in fact are active areas of research and development. Google is committed to making progress in the responsible development of AI and to sharing knowledge, research, tools, datasets, and other resources with the larger community. Below we share some of our current work and recommended practices. As with all of our research, we will take our latest findings into account, work to incorporate them as appropriate, and adapt as we learn more over time.

Responsible AI Practices

Google AI

https://ai.google/responsibilities/responsible-ai-practices/

Recommended practices

- Use a human-centered design approach
- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data
- Understand the limitations of your dataset and model
- Test, Test, Test
- Continue to monitor and update the system after deployment

Thank you

Agnieszka Mikołajczyk

agnieszka.mikolajczyk@pg.edu.pl

Gdańsk University of Technology

Personal website: https://amikolajczyk.netlify.com/

Github: https://github.com/AgaMiko

Linkedin: https://www.linkedin.com/in/agnieszkamikolajczyk/