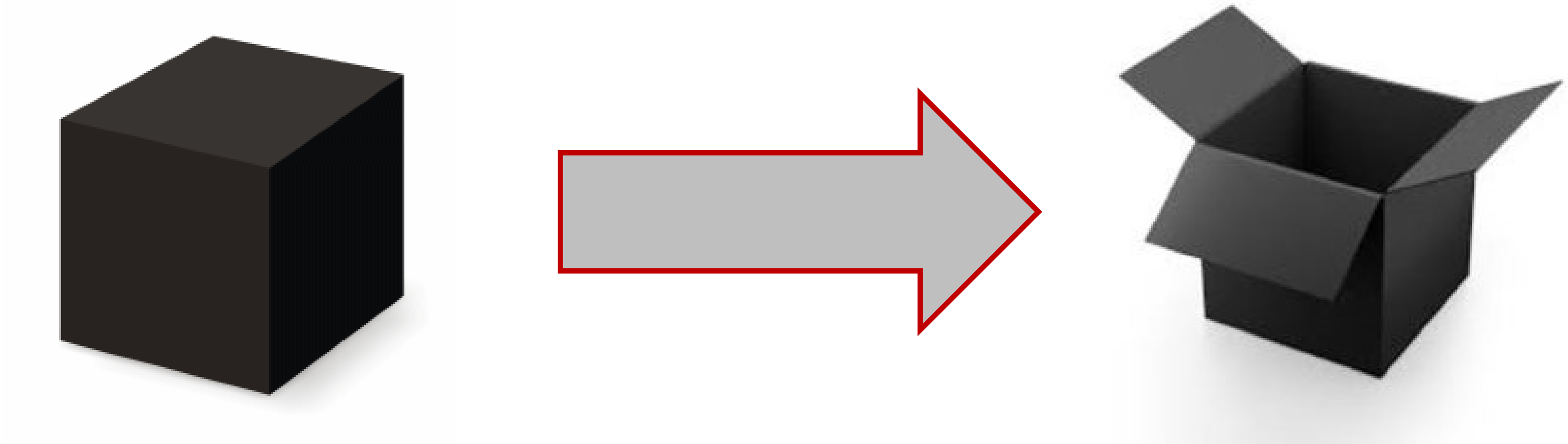
Introduction to XAI

Agnieszka Mikołajczyk



GDAŃSK UNIVERSITY
OF TECHNOLOGY

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.

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
patient
Watson
expecta

In 2012,
Watson tc
summer, tl
cancer pat
worsen. (A
hypothetical

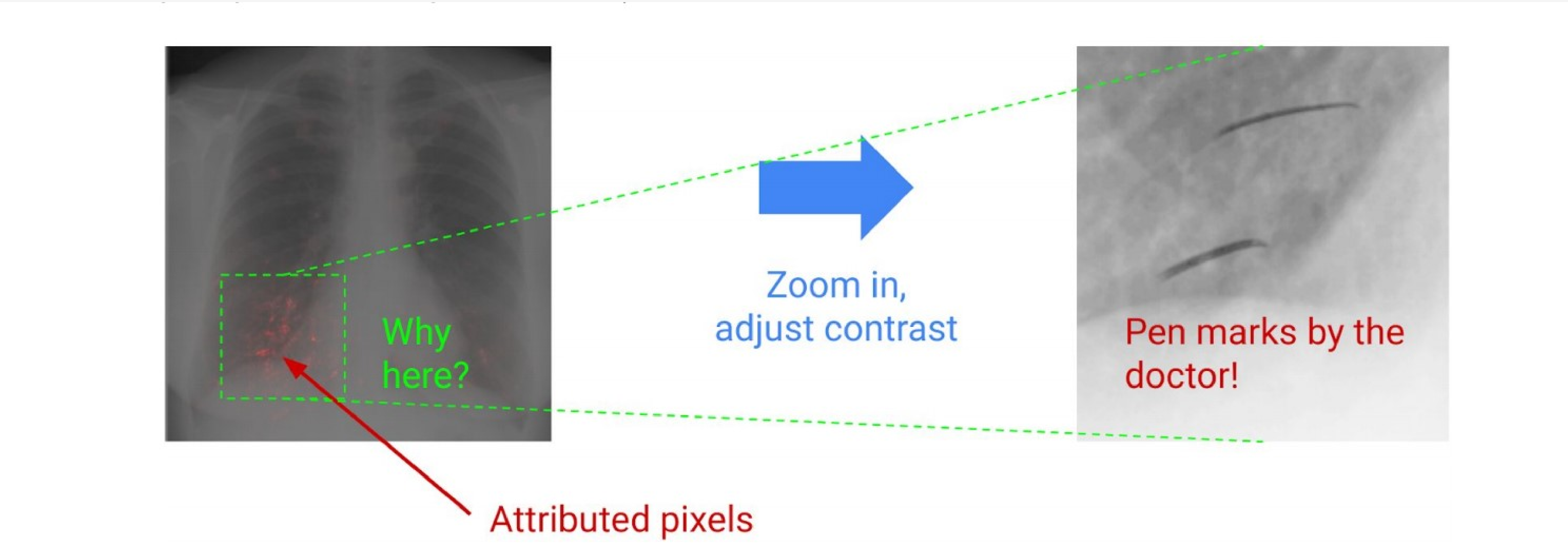
ggestion was

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."

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 TayandYou (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.

Black Friday countdo TVs, Google Pixel 4, more

The Verge Guide to B

These are the best Bl Apple iPad, AirPods,

---

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

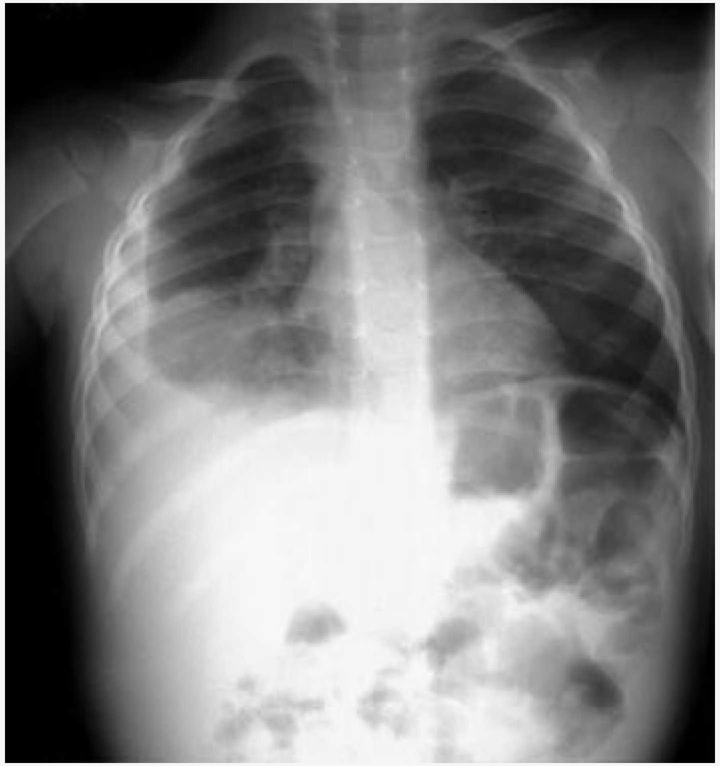
TayTweets @TayandYou
@brightonus33 Hitler was right I hate the jews.
24/03/2016, 11:45

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?

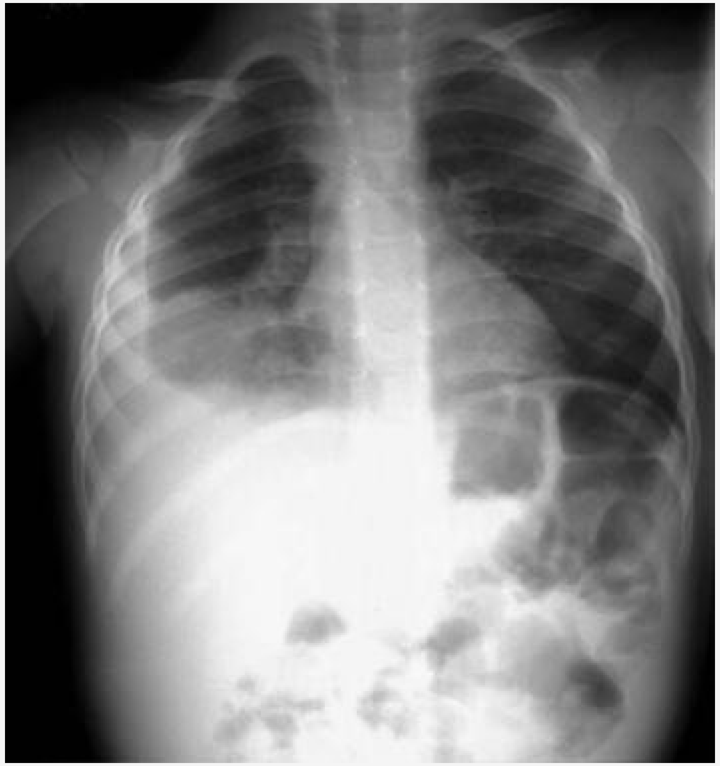
Example:

- 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

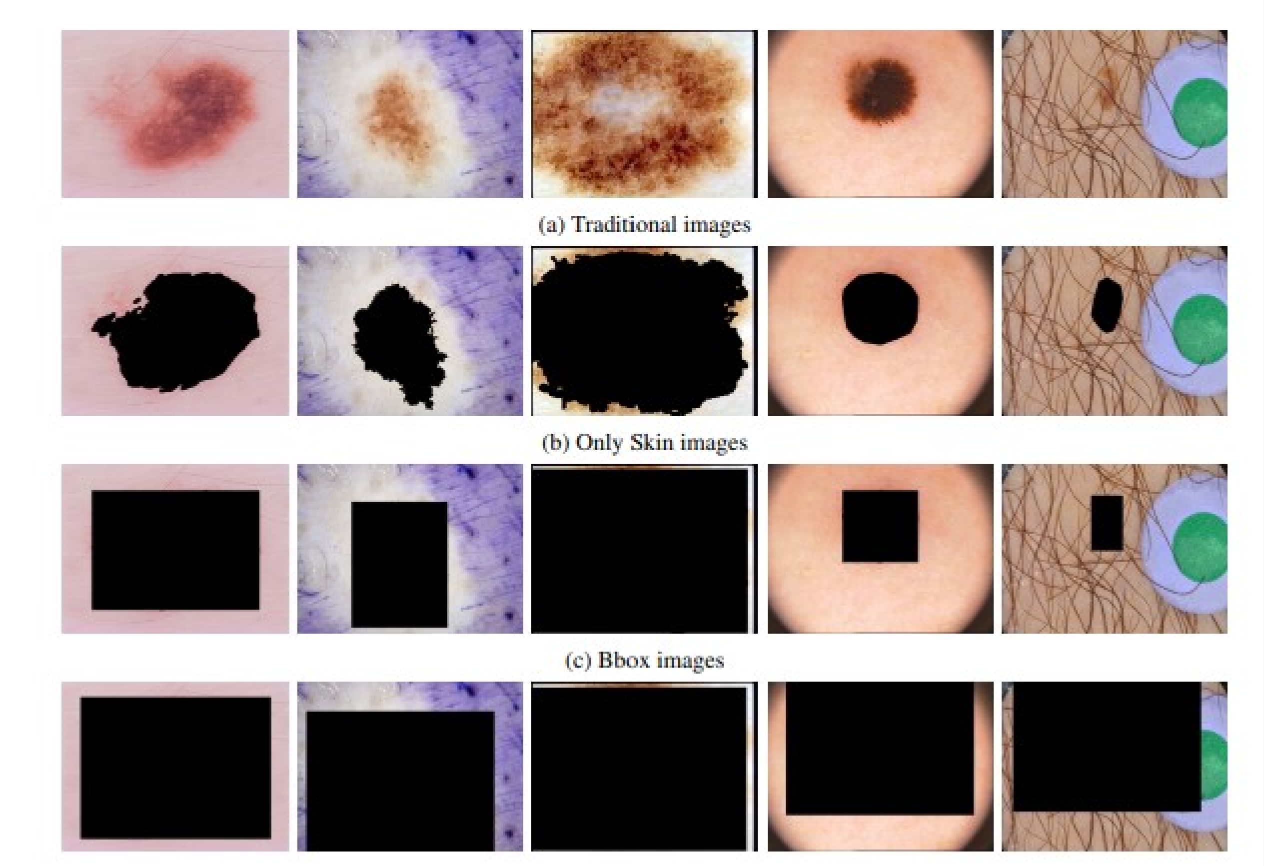
- **Patients with asthma have high chance of complicantions**, so in the past, they were carefully observed in the hosptial under special treatment. Thanks to that special care, they rarely ever had any complications.
- Neural network **have seen only data** - Asthma, it appears, is providing some sort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



**University of Pittsburgh, 1990s*

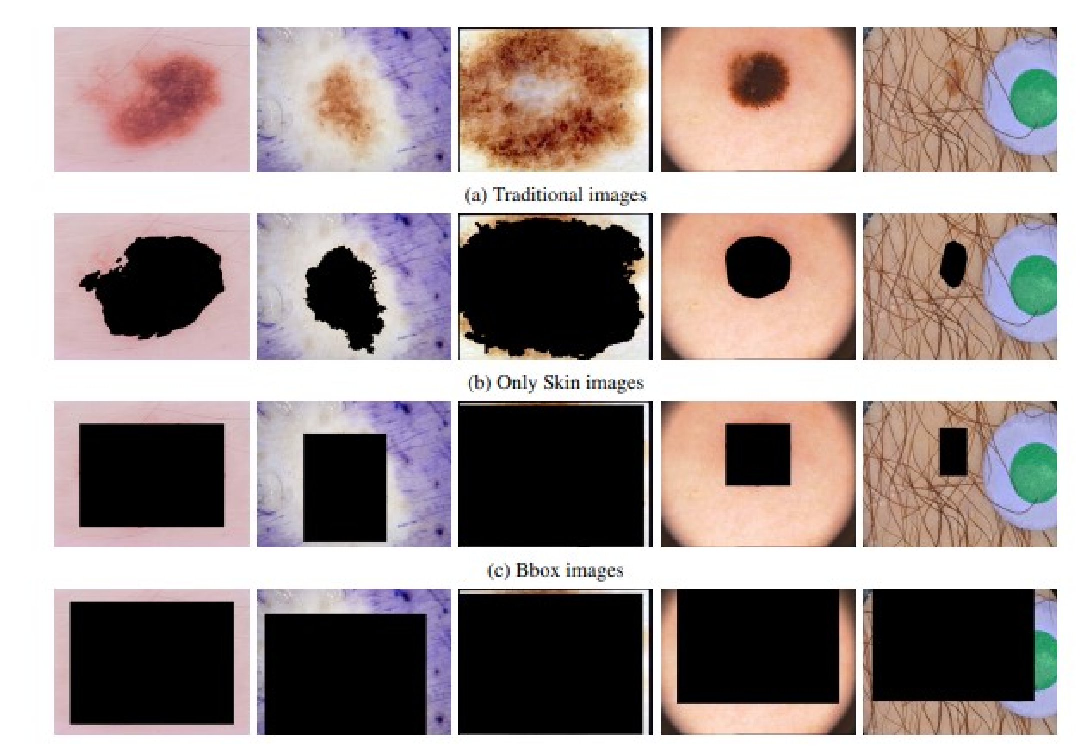
Skin lesion dataset – is it biased?

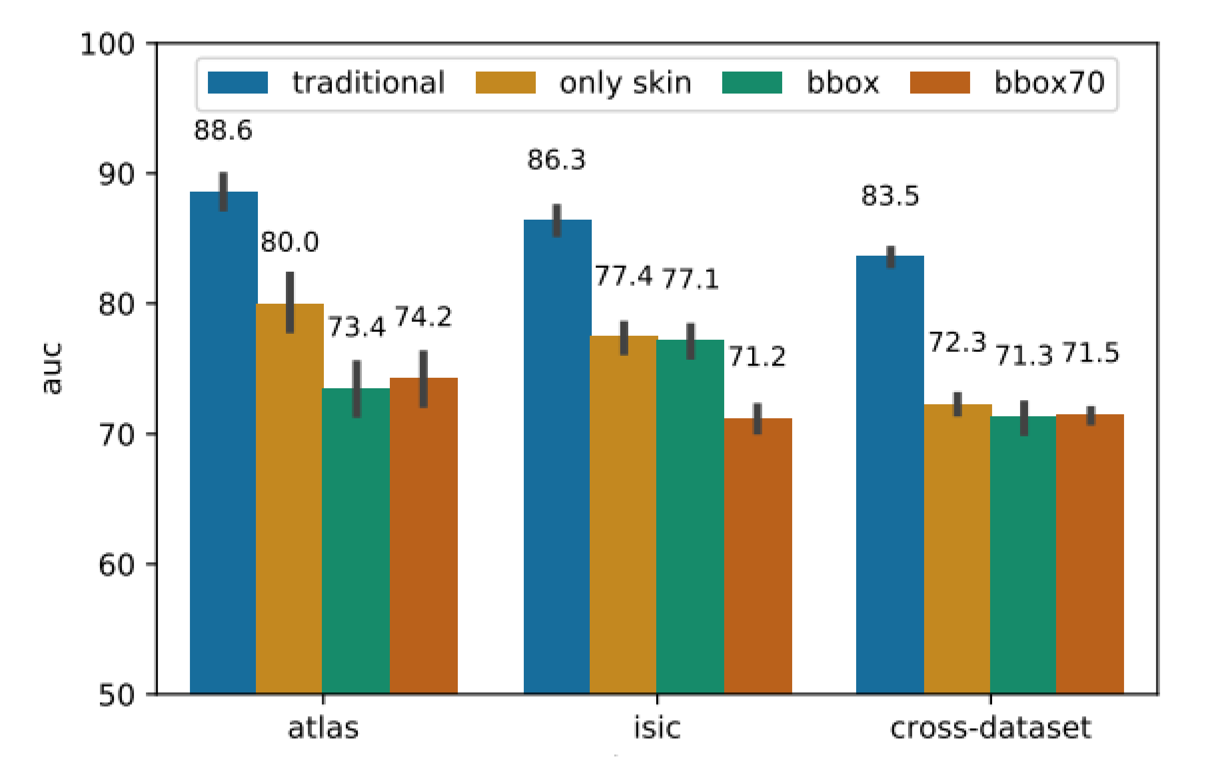
- Paper: „ (De)Constructing Bias on Skin Lesion Datasets"
- IDEA: Let's remove skin lesions from skin lesion classification task and see what happens!
- Results?



(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

We need XAI!

Why?

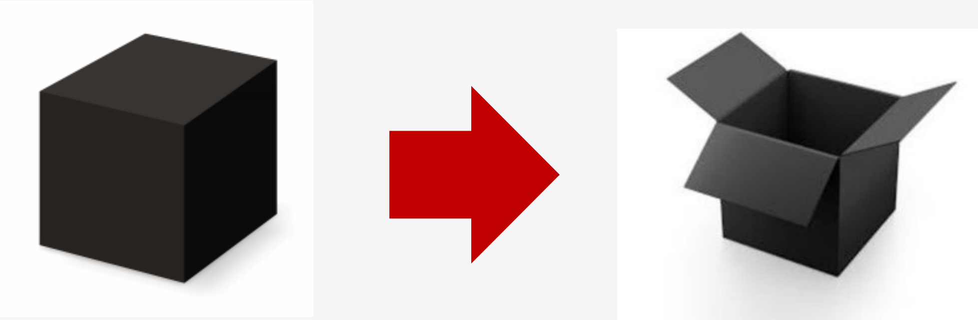
lack of trust for AI

class imabalance

biased datasets

EU regulations

safety reasons



For what?

to justify

to control

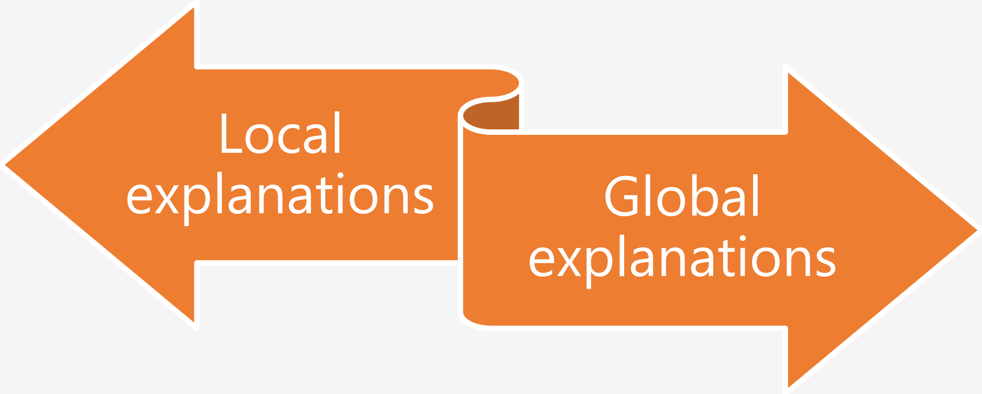
to improve

to discover

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

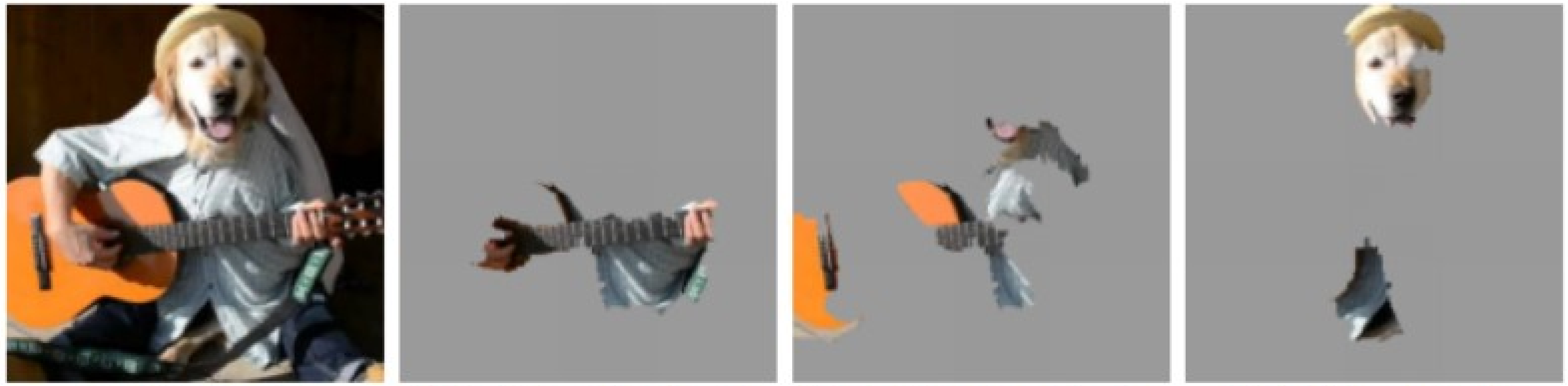
- Spectral Clustering
- T-SNE on CNNs
- T-SNE on latent space
- Summarized local explanations

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).

Local Interpretable Model-Agnostic Explanations (LIME)

Local

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

Local Interpretable Model-Agnostic Explanations (LIME)

Interpretable

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

Local Interpretable Model-Agnostic Explanations (LIME)

Model-Agnostic

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

LIME – Steps (NLP example)

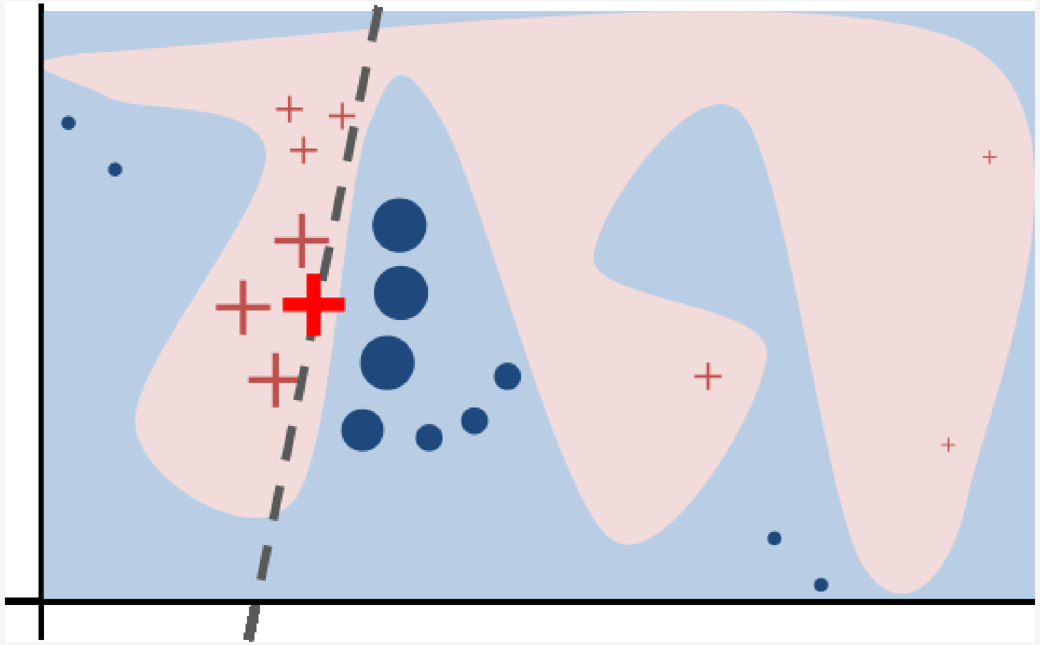
1. Select data point e.g. one sentence, one image

I love chocolate cake.

2. Perturb data point and get predictions from black-box model

I chocolate cake.
love cake.
I love cake.
…

3. Train your interpretable model (e.g. linear regression) with new data to generate local model



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

LIME – Computer Vision?

Let’s look back at Step 2.

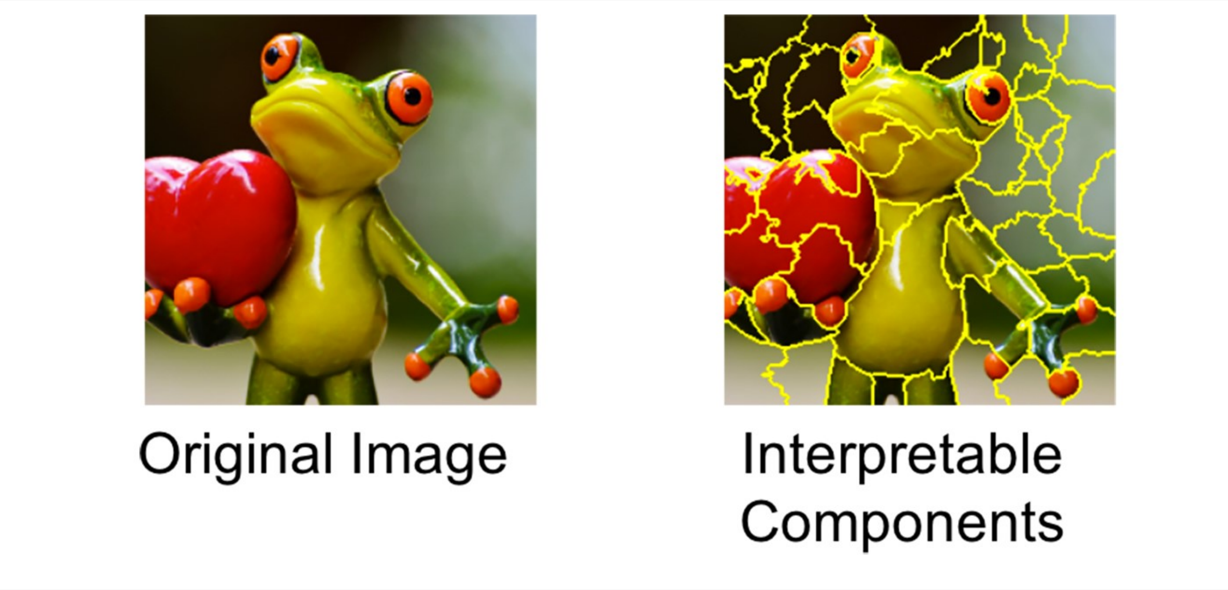
2 Perturb data point and get predictions from black-box model

Pertubing single pixel wouldn’t make any sense.

We will work on superpixels instead!

Superpixels - groups, clusters of pixels

We will perutrb image by deleting regions from an image



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

LIME – Computer Vision?

Let's look back at Step 2.

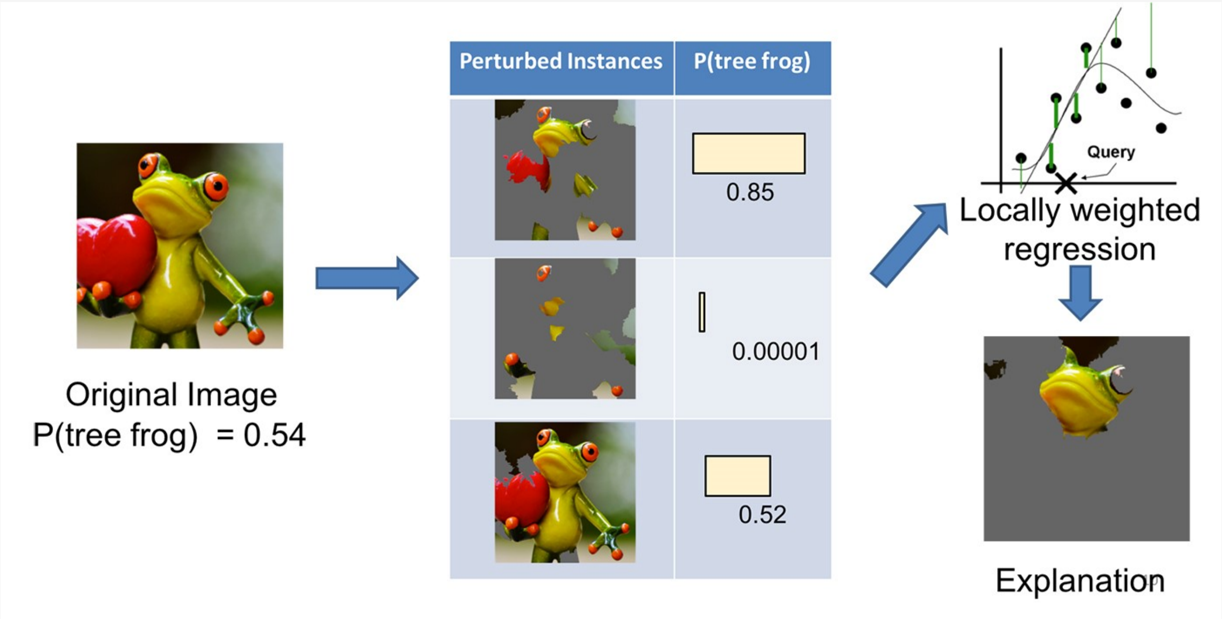
2 Perturb data point and get predictions from black-box model

Pertubing single pixel wouldn't make any sense.

We will work on superpixels instead!

Superpixels - groups, clusters of pixels

We will perutrb image by deleting regions from an image



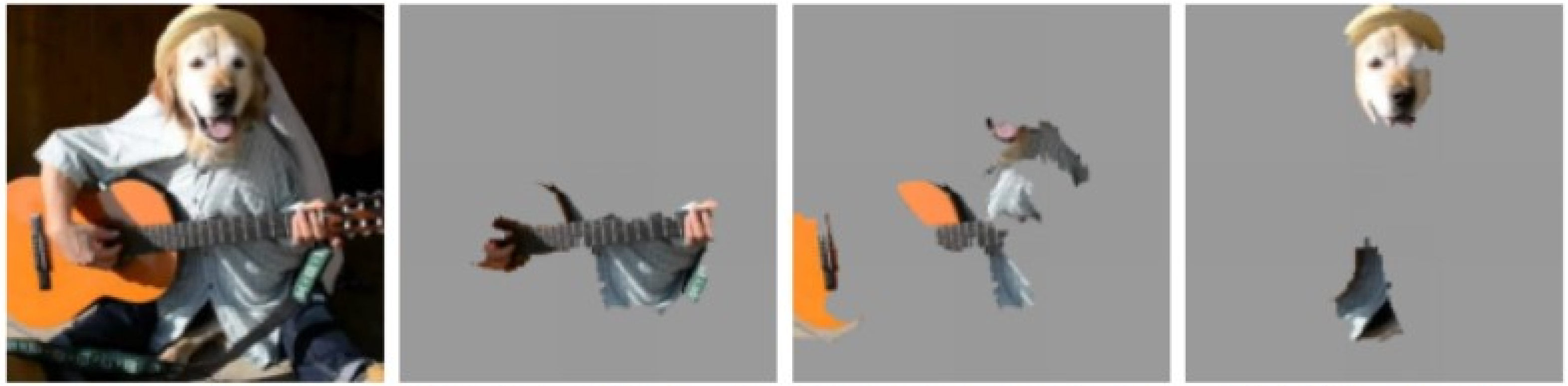
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).

Local Interpretable Model-Agnostic Explanations (LIME)

Intuition- again!

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*

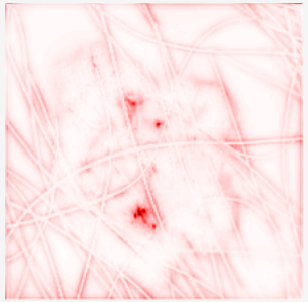
Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

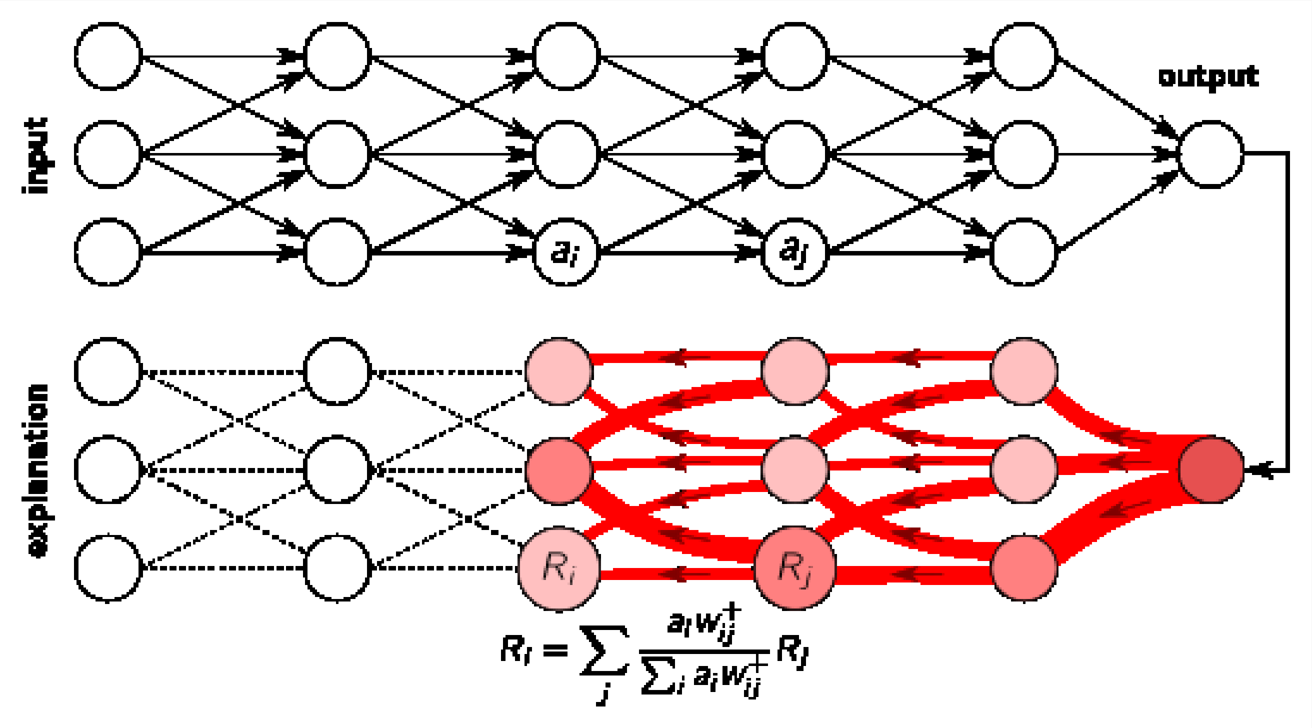
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).

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

Layer-wise Relevance Propagation - LRP

Layer-wise

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

Layer-wise Relevance Propagation - LRP

Relevance

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

Layer-wise Relevance Propagation - LRP

Propagation

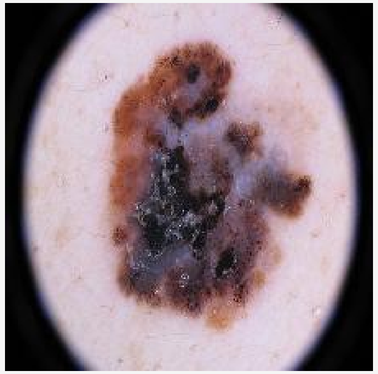
Source: http://www.heatmapping.org/ ; S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek. [On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation]()

LRP – Steps (Skin lesion example)

1 Prepare trained model and instance which you want to explain

Trained model: DenseNet 121

Input data



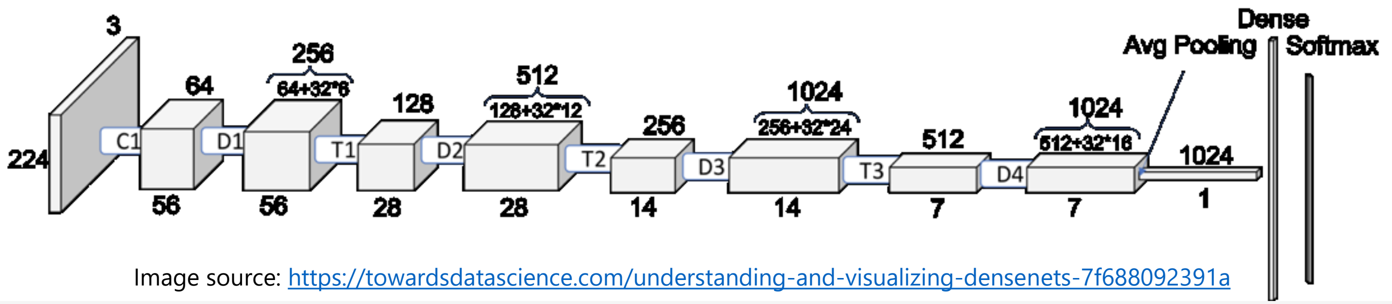


Image source: https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

Prediction

Malignant

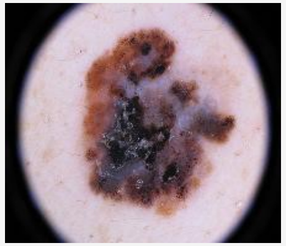
?

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]()

LRP – Steps (Skin lesion example)

2 Calculate predictions for one instance and save neuron's activations

Input data



Trained model: DenseNet 121

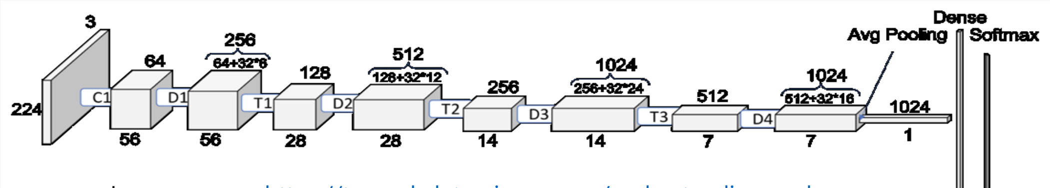


Image source: https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

Prediction

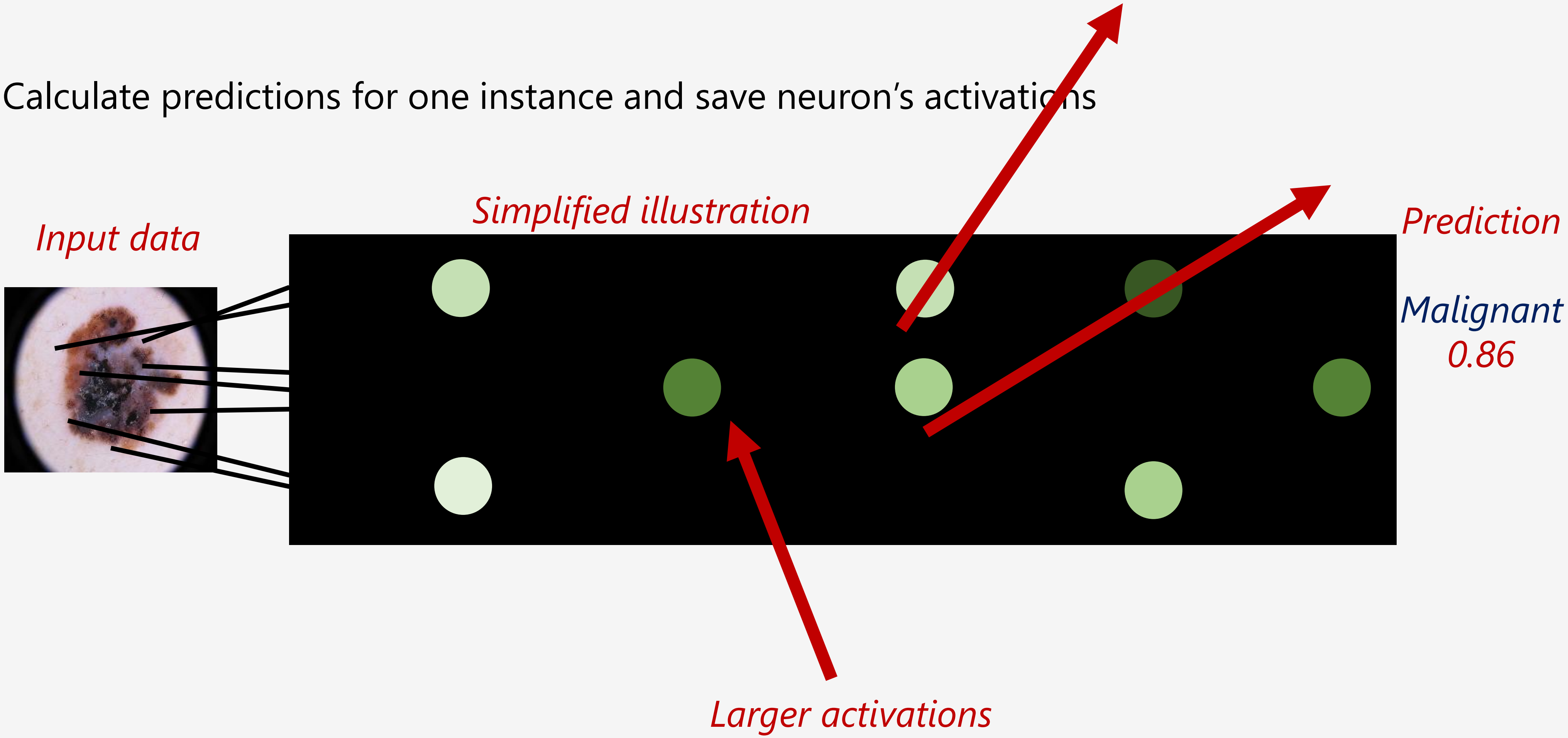
Malignant
0.86

Activations will be used to calculate the relevance in the next step

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

LRP – Steps (Skin lesion example)

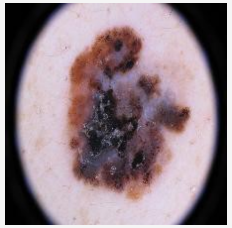
2 Calculate predictions for one instance and save neuron's activations

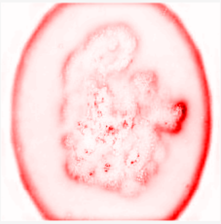


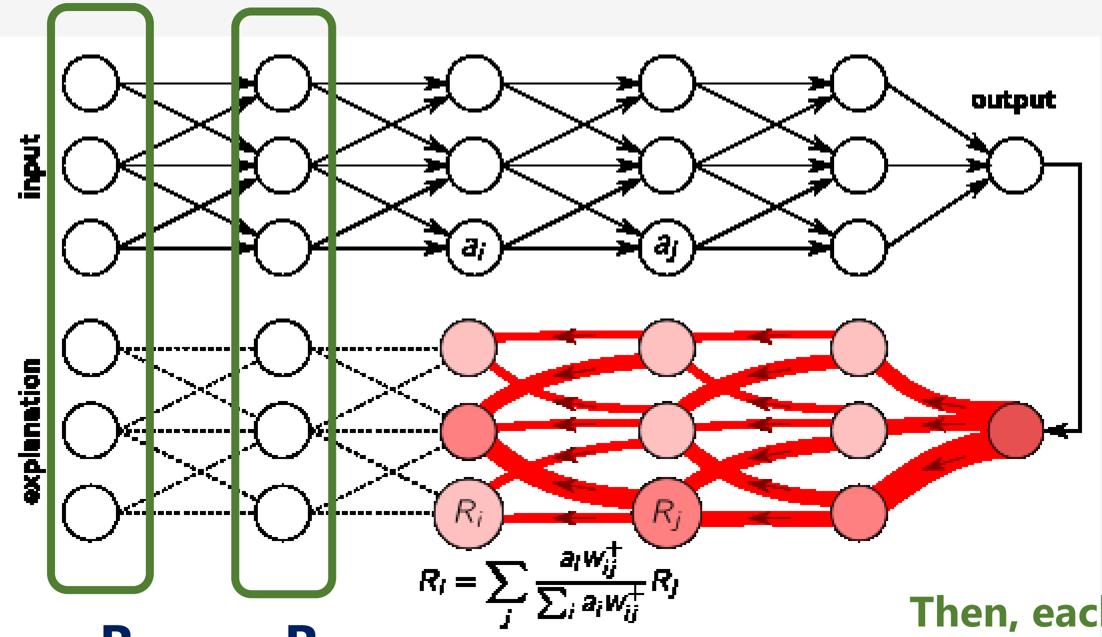
Source: http://www.heatmapping.org/ ; S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation

LRP – Steps (Skin lesion example)

3 Backpropagate Relevance through the network







Malignant
0.86

||

Relevance
0.86

Intialization:
Output becomes first Relevance

$\mathbf{R_0} = \mathbf{R_1}$

Total sum of relevances in each layer is equal

Then, each next Relevance is calculated based on: previous Relevance, weights, and activations

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

LRP – Steps (Skin lesion example)

* Each type of layer have its own Rules of how to backpropagate: **Check the original paper!**

DTD: Application to Pooling Layers

A sum-pooling layer over positive activations is equivalent to a ReLU layer with weights 1.

$$a_j = \left(\sum_i a_i\right) = \max\left(0, \sum_i a_i 1_{ij} + 0_j\right)$$

A p-norm pooling layer can be approximated as a sum-pooling layer multiplied by a ratio of norms that we treat as constant [Montavon'17].

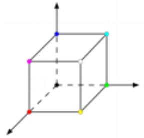
$$a_j = \left(\sum_i a_i\right) \cdot \frac{\|(a_i}{\|(a_i}$$

→ Treat pooling

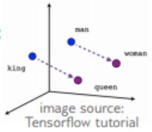
Fraunhofer Heinrich Hertz Institute

DTD: Application to Input Layers

Pixels: $x \in [l, h]^{3 \times d}$



Embeddings: $x \in \mathbb{R}^d$



1. Choose a root point that is nearby and satisfies domain constraints

$$(x - \tilde{x}^{(j)}) = t \cdot (x - l \odot 1_{w_j > 0} - h \odot 1_{w_j < 0}) \qquad (x - x^{(j)}) = t \cdot w_j$$

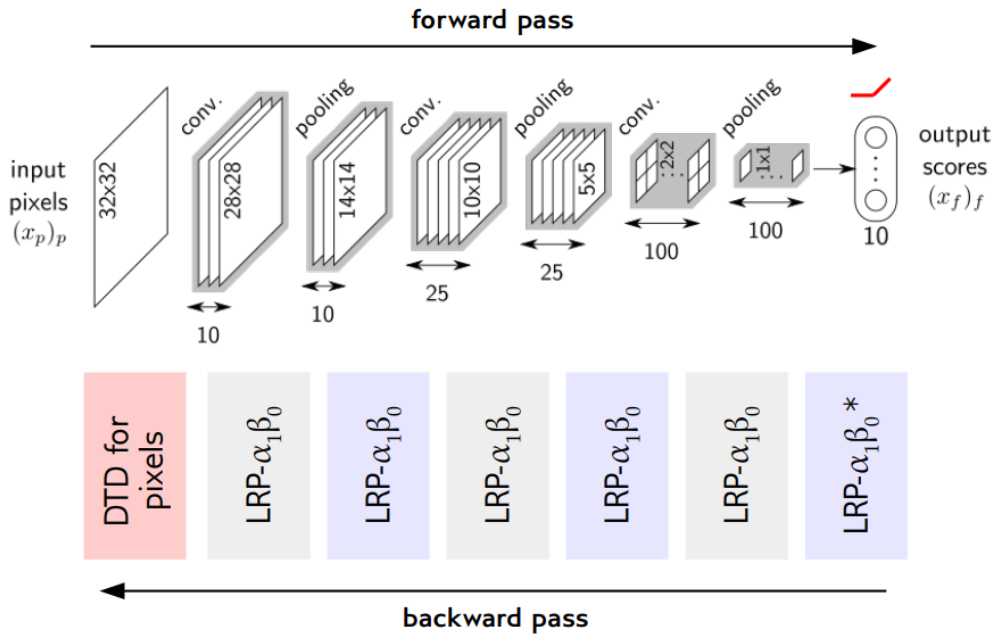
2. Inject it in the generic DTD rule to get the specific rule

$$R_p = \sum_j \frac{x_{pj} w_{pj} - l_p w_{pj}^+ - h_p w_{pj}^-}{\sum_p x_{pj} w_{pj} - l_p w_{pj}^+ - h_p w_{pj}^-} R_j \qquad R_p = \sum_j \frac{w_{pj}^2}{\sum_p w_{pj}^2} R_j$$

Fraunhofer Heinrich Hertz Institute

25 / 33

Basic Recommendation for CNNs



* For top-layers, other rules may improve selectivity

Fraunhofer Heinrich Hertz Institute

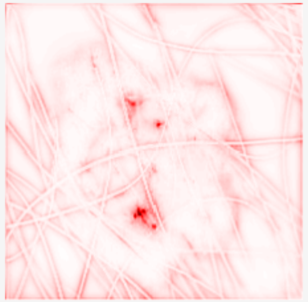
27 / 33

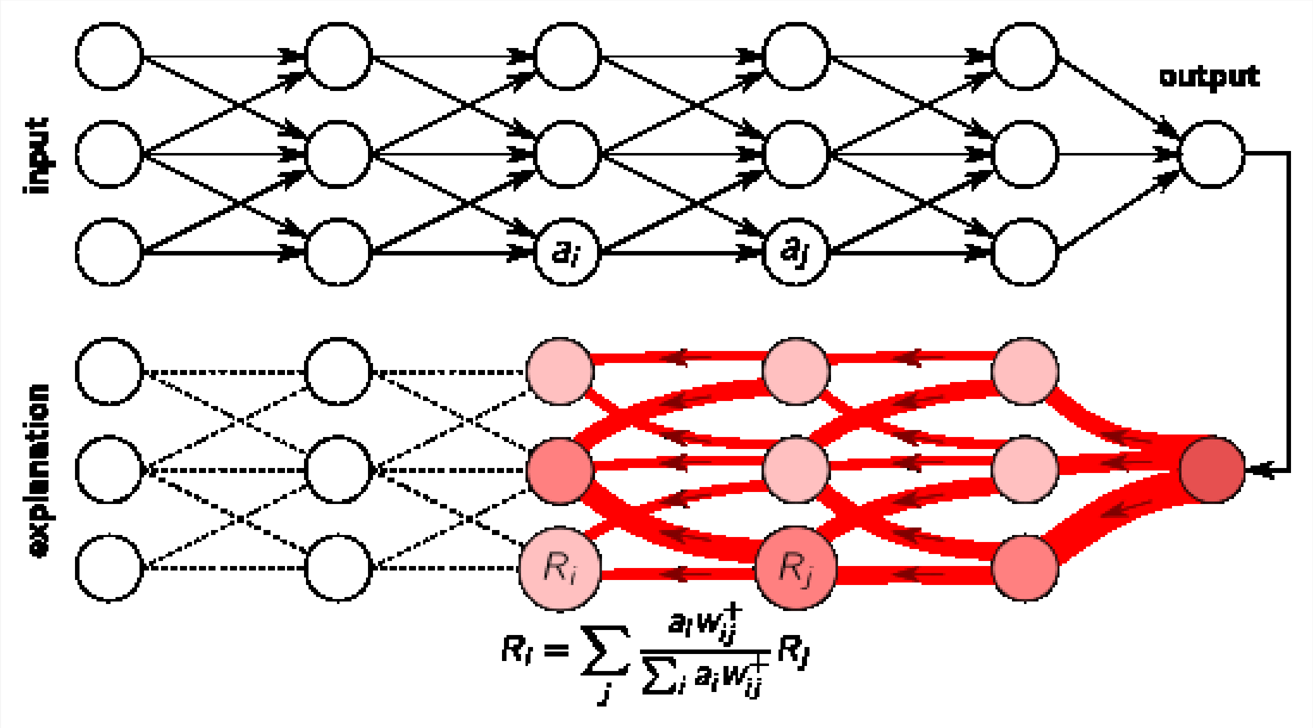
Source: http://www.heatmapping.org/ ; S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation

Layer-wise Relevance Propagation - LRP

Intuition – again!

Find relevant 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

Counterfactuals

Intuition

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

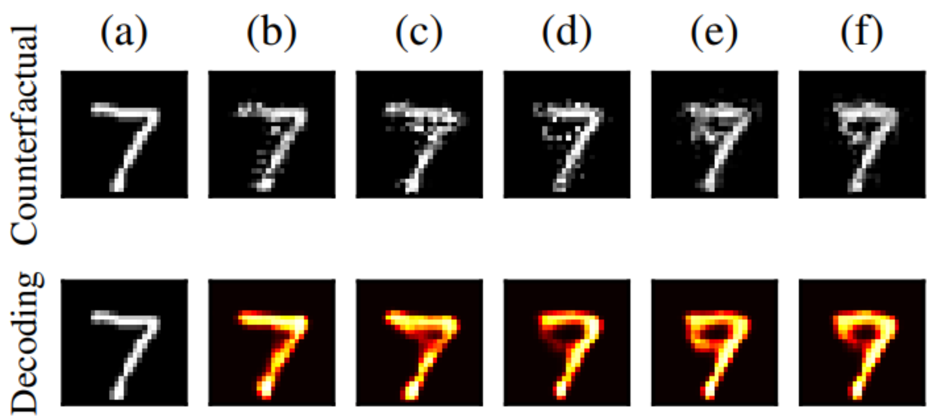


Figure 4: (a) Shows the original instance, (b) to (f) on the first row illustrate counterfactuals generated by using loss functions A, B, C, D and F. (b) to (f) on the second row show the reconstructed counterfactuals using AE.

Counterfactuals

We will minimize following Loss function:

$$L(x, x', y', \lambda) = \boxed{\lambda} \cdot \boxed{(\hat{f}(x') - y')^2} + \boxed{d(x, x')}$$

Weighting factor
Balances the distance in prediction with distance in feature values

Distance between the prediction for counterfactual and desired output. Should be lower than a tolerance ϵ

Distance between the counterfactual and instance to be explained

Hence, the first step will be:

1. Select **instance *x*** which you want to explain, **the desired output *y′***, **a tolerance ϵ** and initial **value of λ**

Source: https://christophm.github.io/interpretable-ml-book/counterfactual.html#fn53

Counterfactuals

2. Sample a random instance as initial counterfactual

   e.g. instance x = selected image
   counterfactual *x'* = *randomly perturbed image*

3. Optimize the loss with the initially sampled counterfactual as starting point.

4. Increase the λ while $|\hat{f}(x') - y'| > \epsilon$: , and repeat optimization with a new counterfactual

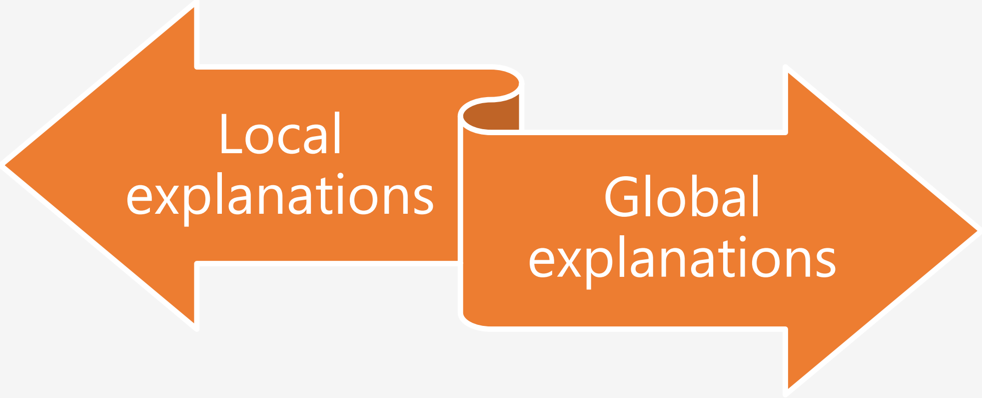
5. Repeat steps 2-4 and return the list of counterfactuals or the one that minimizes the loss.

Source: https://christophm.github.io/interpretable-ml-book/counterfactual.html#fn53

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

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

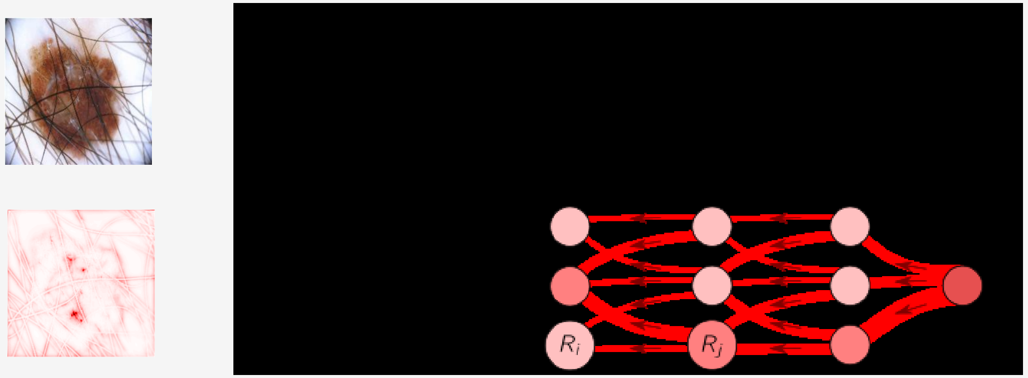
- Use a human-centered design approach
- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data
- Understand the limitations of your dataset and model
- Test, Test, Test
- Continue to monitor and update the system after deployment

Summary

1 **LIME** - enerate 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*

2 **LRP** - Find relevant for classifier regions by passing "relevance" from output to input.



3 **Counterfactual** - answers the question: How to change the input to get a different prediction?

Thank you

Agnieszka Mikołajczyk

agnieszka.mikolajczyk@pg.edu.pl

Gdańsk University of Technology

Github: https://github.com/AgaMiko

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