INTERPRETABLE MACHINE LEARNING

Anna Dawid

Outline

WHY DO WE NEED INTERPRETABILITY AND WHAT IT IS

OVERVIEW OF INTERPRETABILITY METHODS

TUTORIAL EXAMPLE: CLASS ACTIVATION MAP (CAM)

1/4

WHY DO WE NEED INTERPRETABILITY?

And what it is?

'AI IS THE NEW ELECTRICITY'

"Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform in the next several years."

Andrew Ng

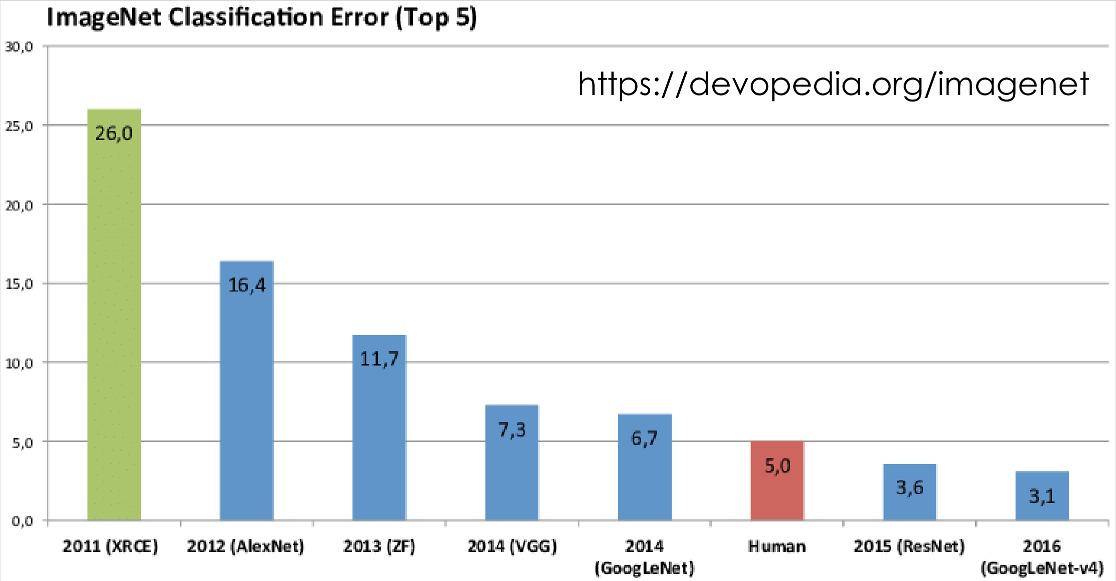
Former chief scientist at Baidu, Co-founder at Coursera

https://www.sharper.ai/ai-is-the-new-electricity/

https://github.com/karolmajek/Mask_RCNN

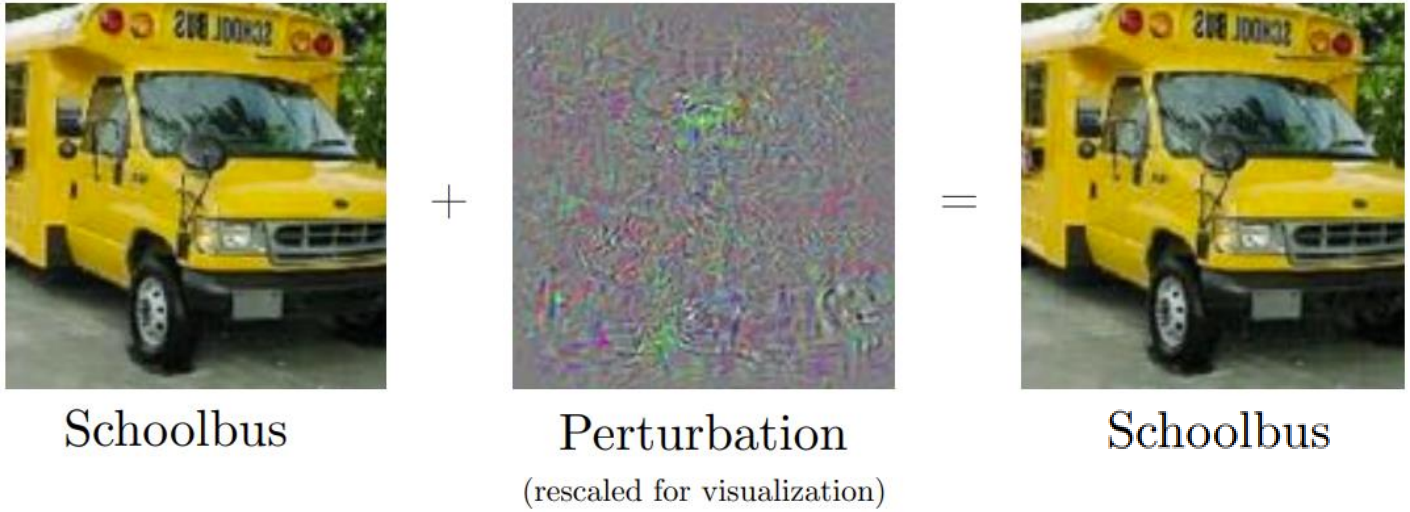
Computer vision for self-driving cars

Super-human performance
on image classification

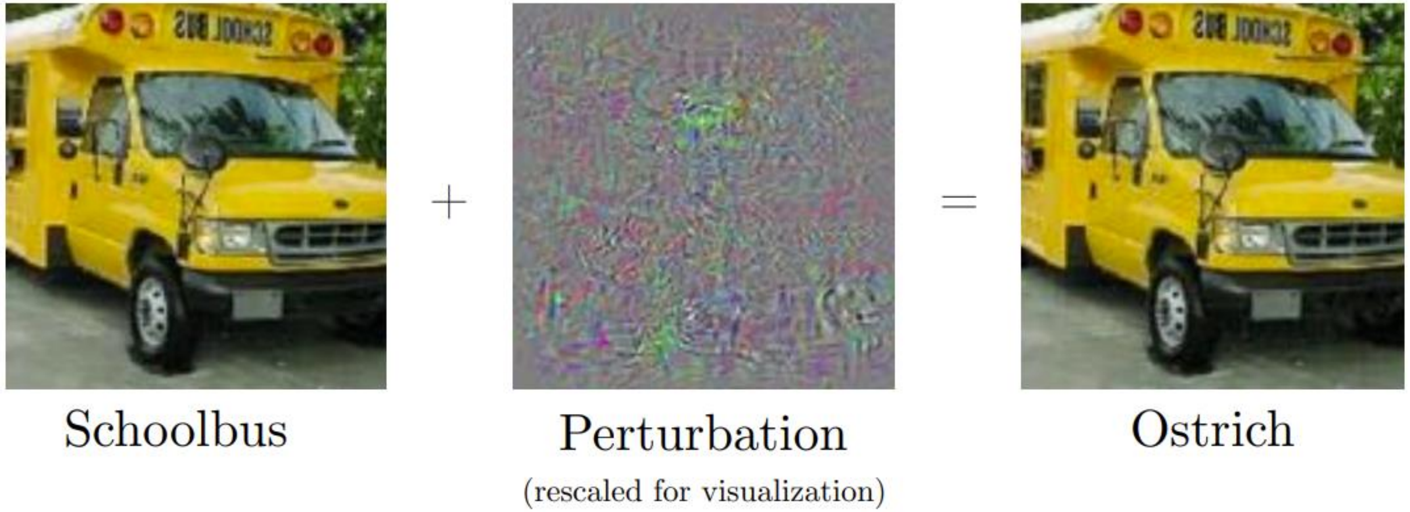


All is well?

What humans see:



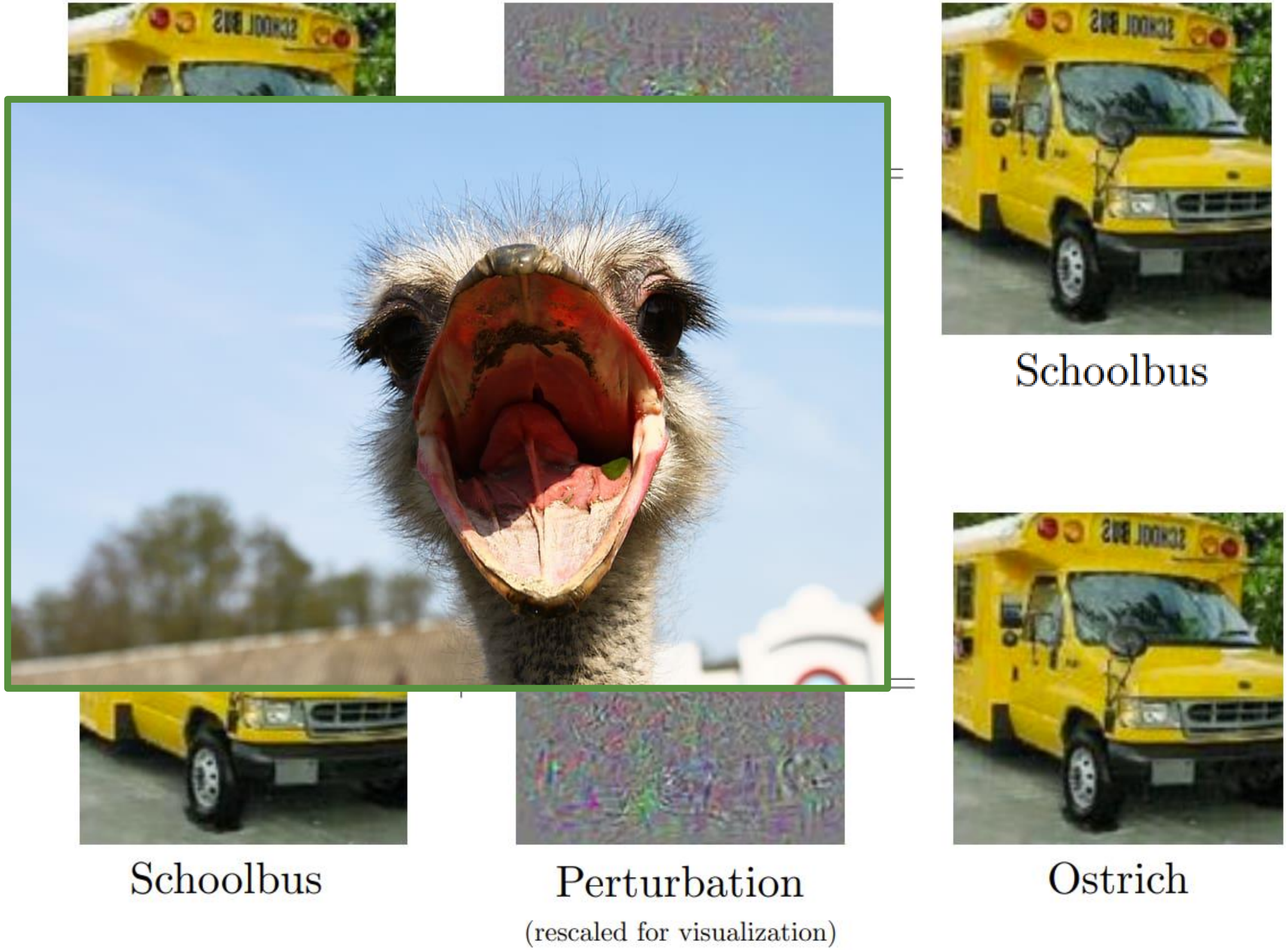
What (some) machines see:



modified from Goodfellow (2018), arXiv:1806.04169v1

What humans see:

What (some) machines see:

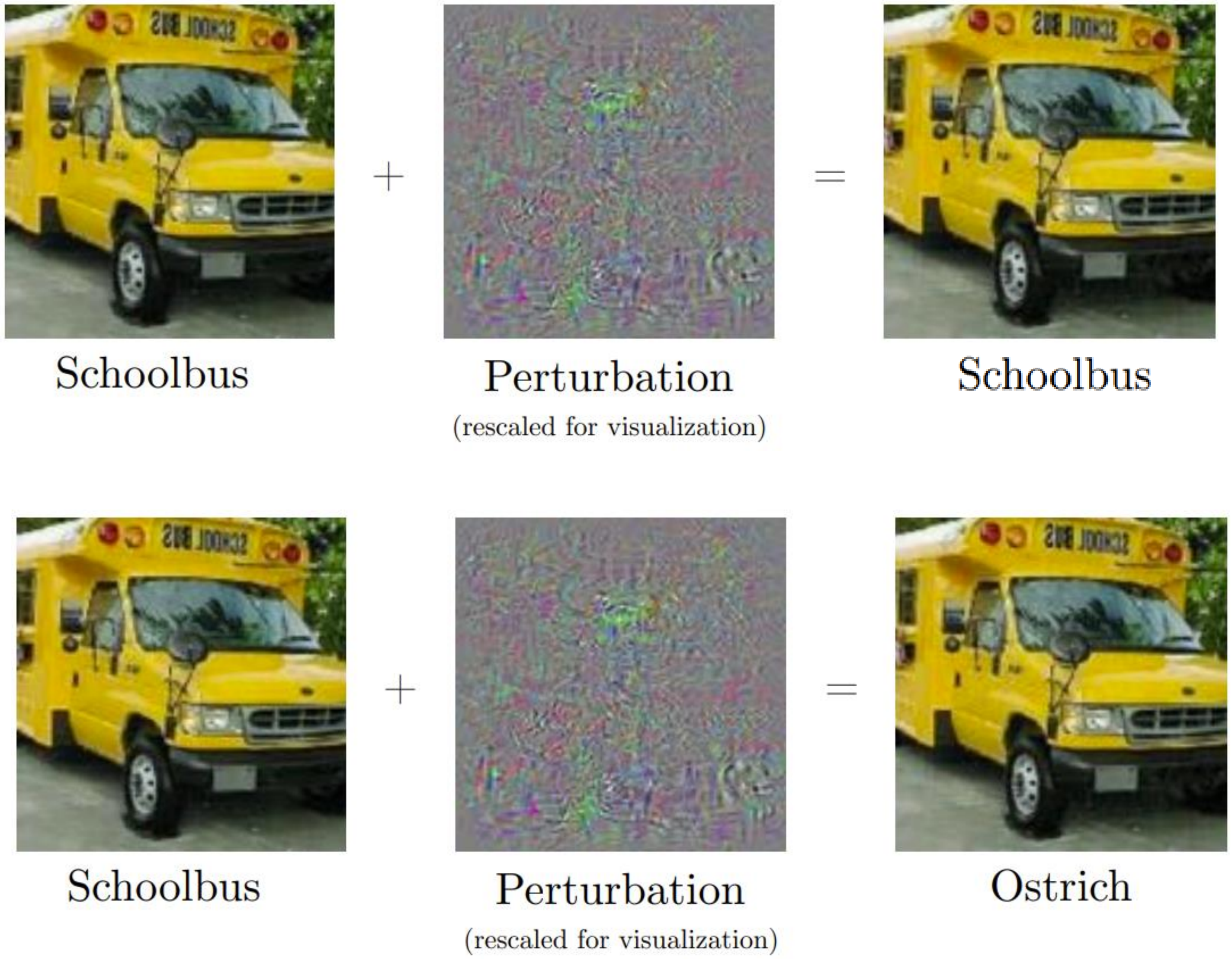


modified from Goodfellow (2018), arXiv:1806.04169v1

open to
adversarial attacks!

What humans see:

What (some)
machines see:

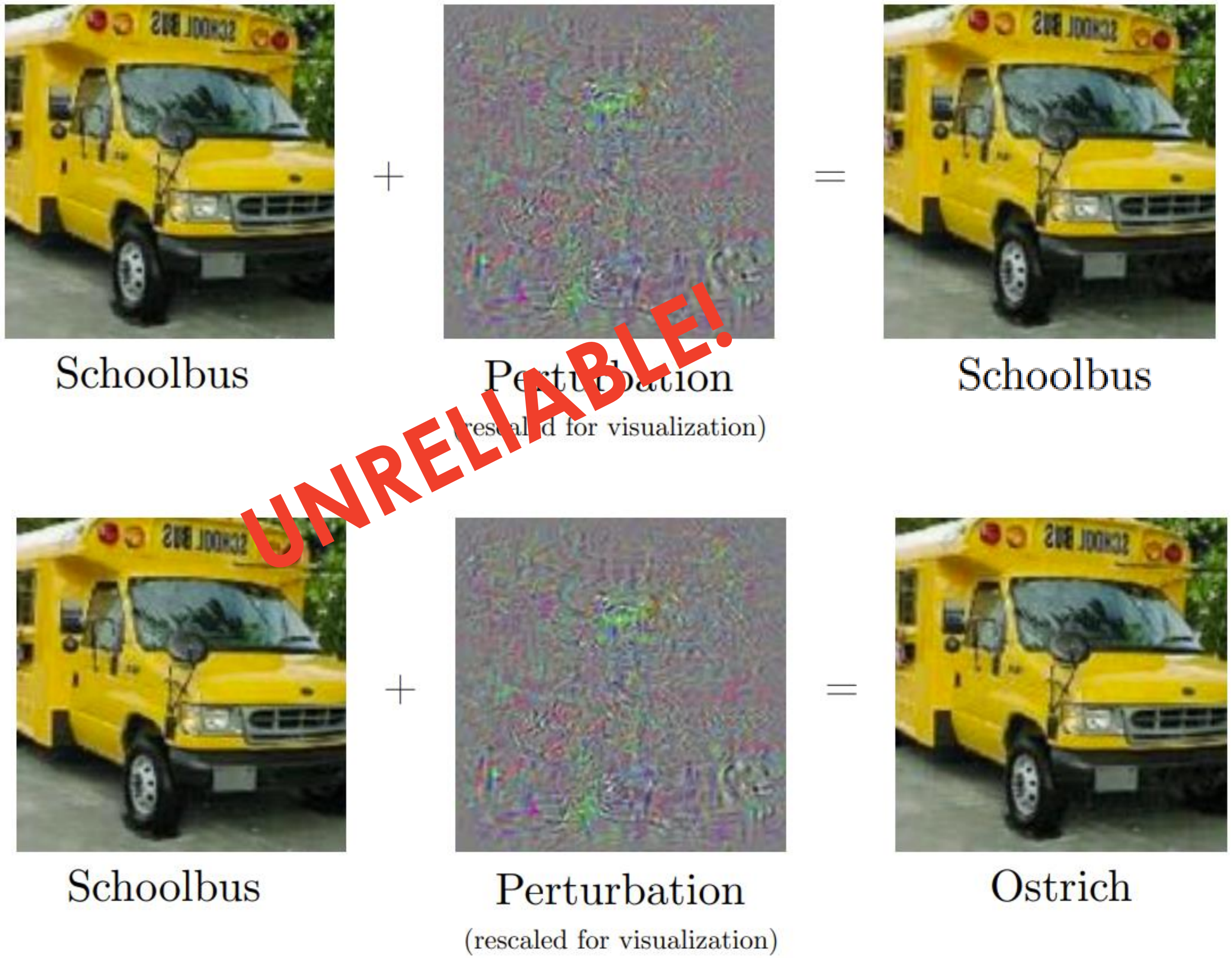


modified from Goodfellow (2018), arXiv:1806.04169v1

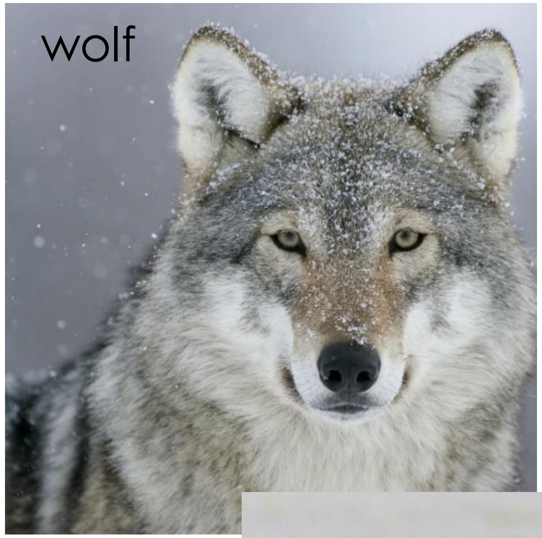
open to
adversarial attacks!

What humans see:

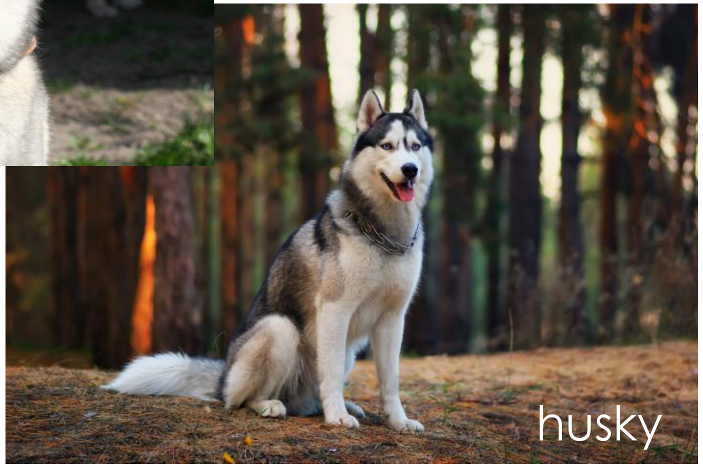
What (some) machines see:



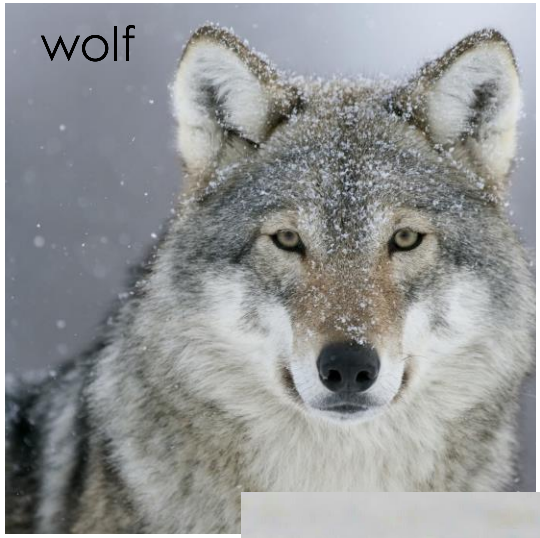
modified from Goodfellow (2018), arXiv:1806.04169v1



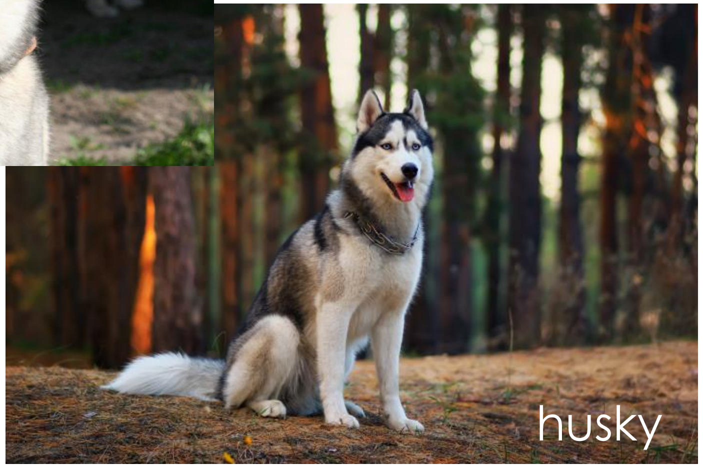
Husky vs. wolf classifier



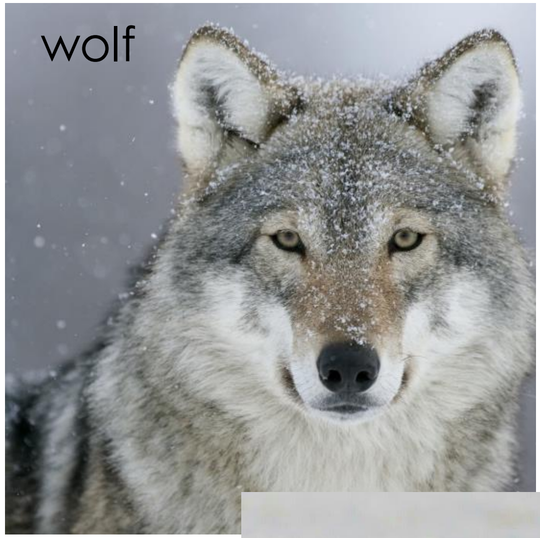
Ribeiro (2016), arXiv:1602.04938v3



Classifier learnt the background…



Ribeiro (2016), arXiv:1602.04938v3



Classifier learnt the background...

Bad generalization!

Ribeiro (2016), arXiv:1602.04938v3

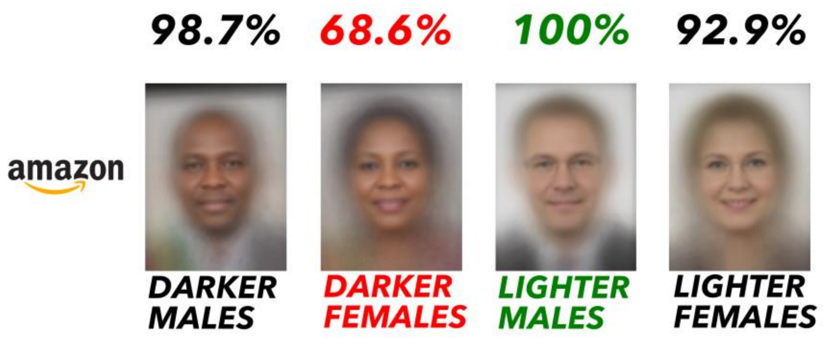
AI trained by Amazon inherited biases from historical hiring decisions

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

TayTweets, a chat bot trained by Microsoft on Twitter data started spouting racist tweets…

https://towardsdatascience.com/biases-in-machine-learning-61186da78591

Amazon Rekognition failures in facial recognition correlate with skin color



https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

AI trained by Amazon inherited biases from historical hiring decisions

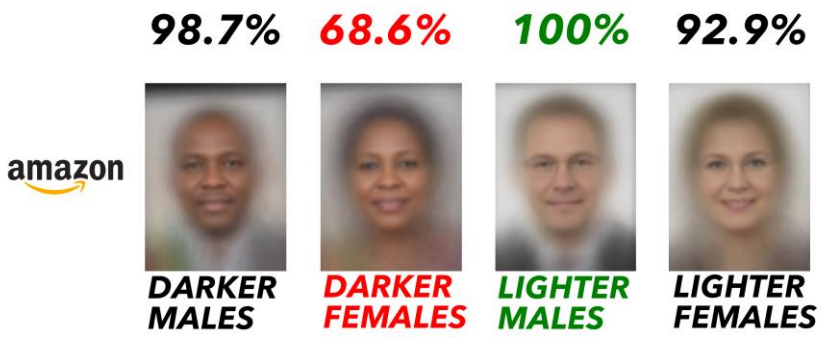
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

UNFAIR!

TayTweets, a chat bot trained by Microsoft on Twitter data started spouting racist tweets…

https://towardsdatascience.com/biases-in-machine-learning-61186da78591

Amazon Rekognition failures in facial recognistion correlate with skin color



https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

People worry that the computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world.

Pedro Domingos "The Master Algorithm"

All is well?

No.

We need **interpretability**.

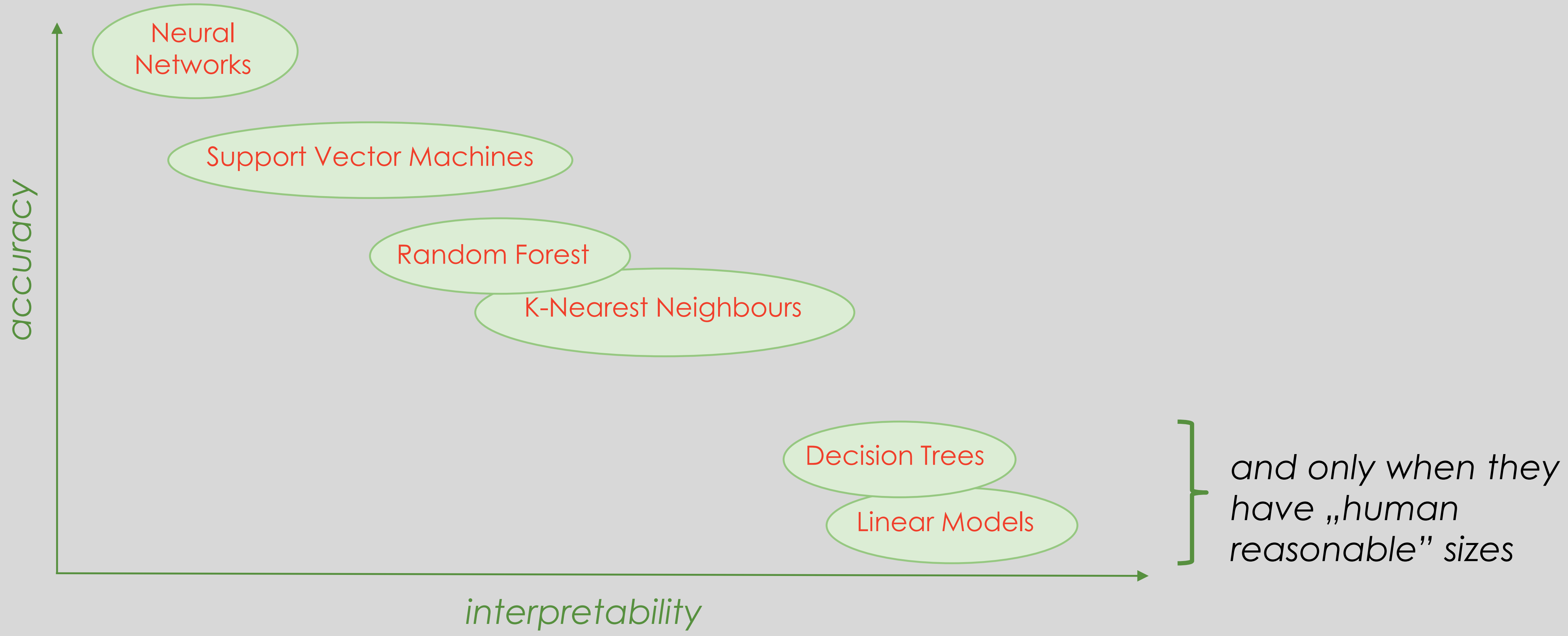
Some definitions

Black boxes
Systems that hide their internal logic to the user (either the internals are unknown or uninterpretable to humans)

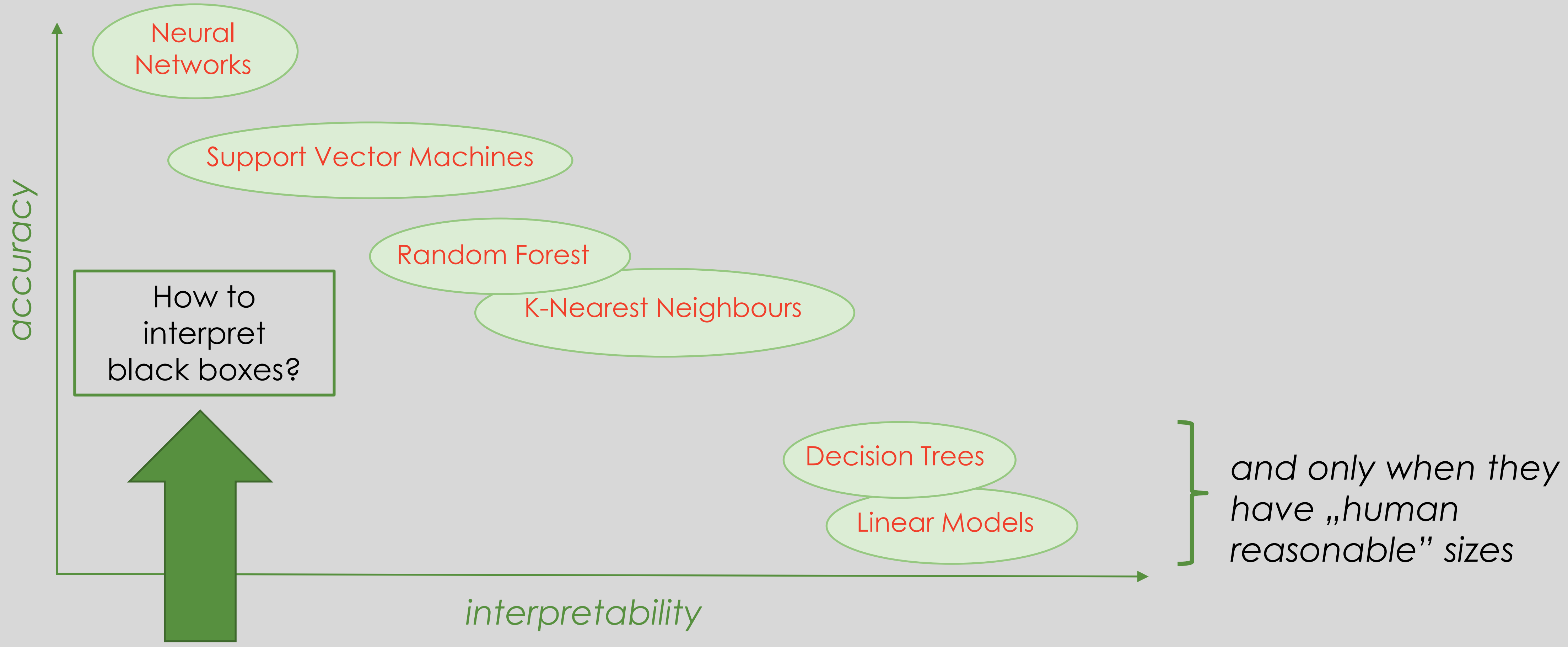
***Interpretability* (also: *comprehensibility*)**
Ability to explain or to present in understandable terms to a human

Guidotti et al. 2018. *A Survey of Methods for Explaining Black Boxes Models*. ACM Comput. Surv. 51, 5, art. 93

Trade-off between complexity and interpretability



Trade-off between complexity and interpretability



OVERVIEW OF INTERPRETABILITY METHODS

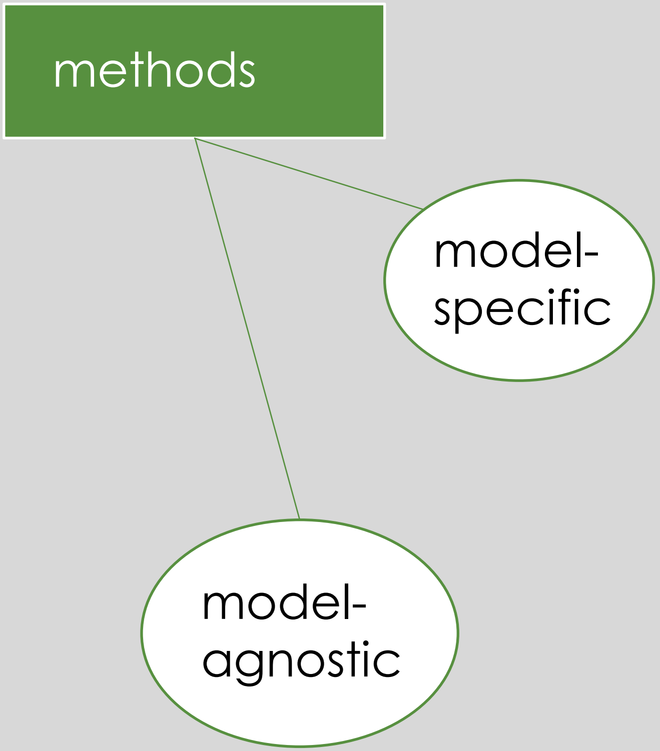
Possible approaches

- methods assigning meaning to individual **model components**
- methods analyzing model predictions when **data is perturbed**
- surrogate approach where the model is approximated by a simpler, more interpretable **surrogate model**

Molnar, Casalicchio, Bischl (2020) arXiv:2010.09337

Possible approaches

- methods assigning meaning to individual **model components** *(always model-specific)*
- methods analyzing model predictions when **data is perturbed** *(mostly model-agnostic)*
- surrogate approach where the model is approximated by a simpler, more interpretable **surrogate model**



Molnar, Casalicchio, Bischl (2020) arXiv:2010.09337

Possible approaches

- methods assigning meaning to individual **model components** *(always model-specific)*
- methods analyzing model predictions when **data is perturbed** *(mostly model-agnostic)*
- surrogate approach where the model is approximated by a simpler, more interpretable **surrogate model**

explanation

- global
- local

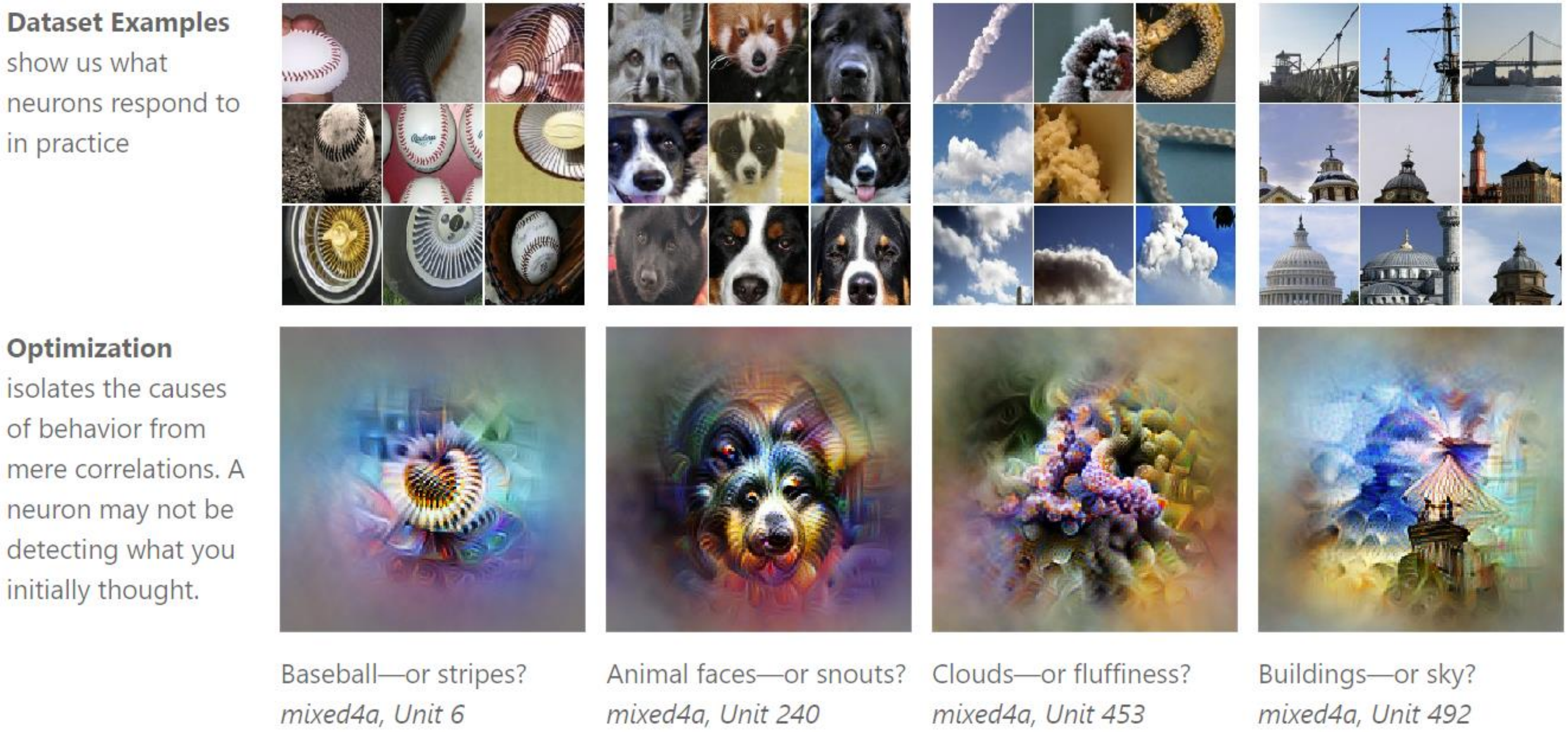
Molnar, Casalicchio, Bischl (2020) arXiv:2010.09337

EXAMPLES

subjective choice of mine!

Feature visualisation

finding the image that maximizes the activation of certain part of the model

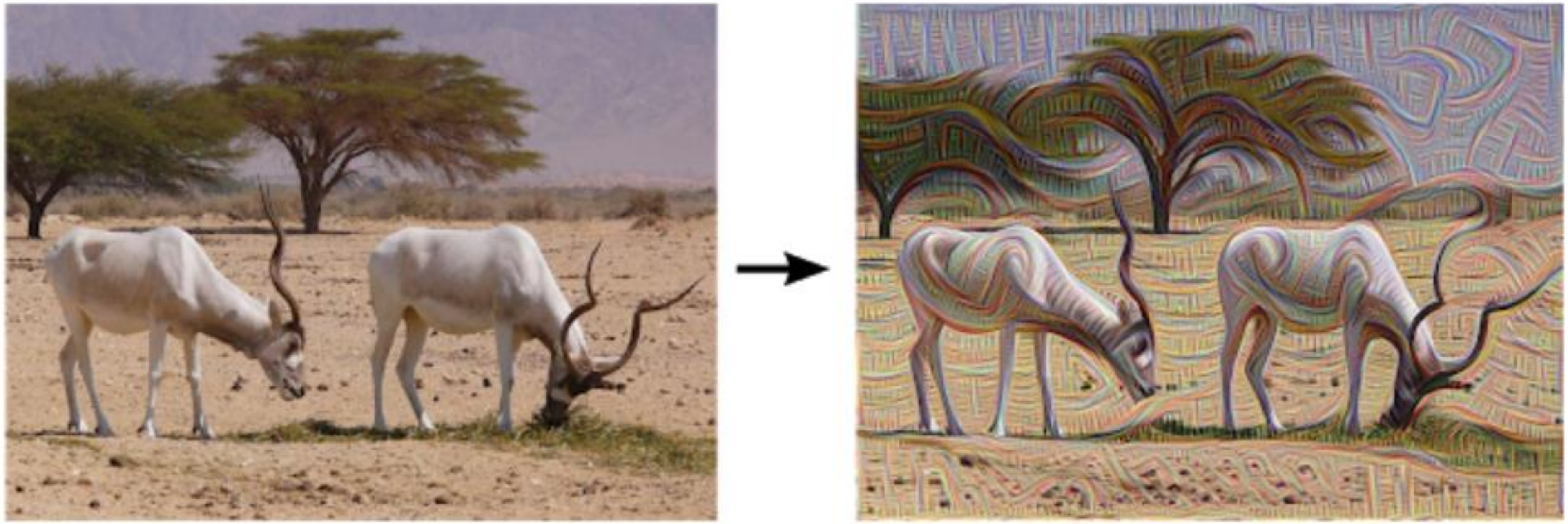


https://distill.pub/2017/feature-visualization/
https://distill.pub/2017/feature-visualization/appendix/

We can look at whole layers!

Deep Dream

enhance what was
detected by a chosen layer
and visualise

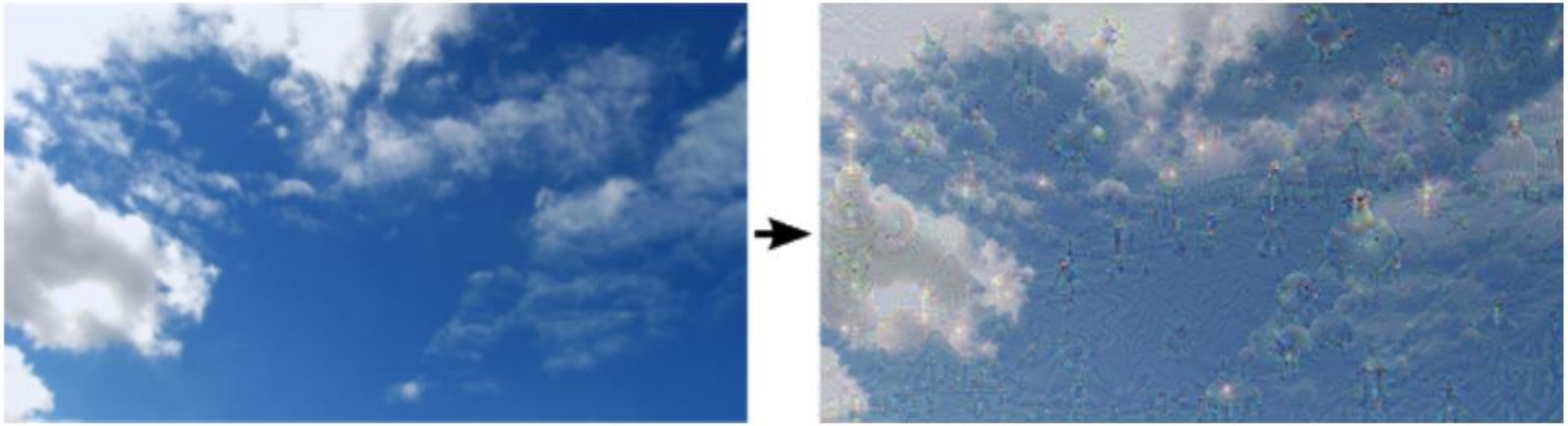


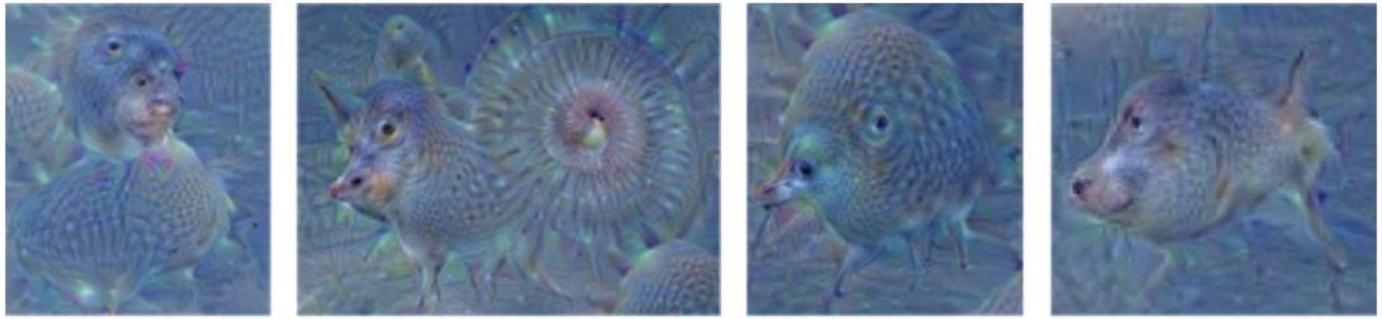
Left: Original photo by Zachi Evenor. Right: processed by Günther Noack, Software Engineer

https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

We can look at whole layers!
Deep Dream

enhance what was detected by a chosen layer and visualise

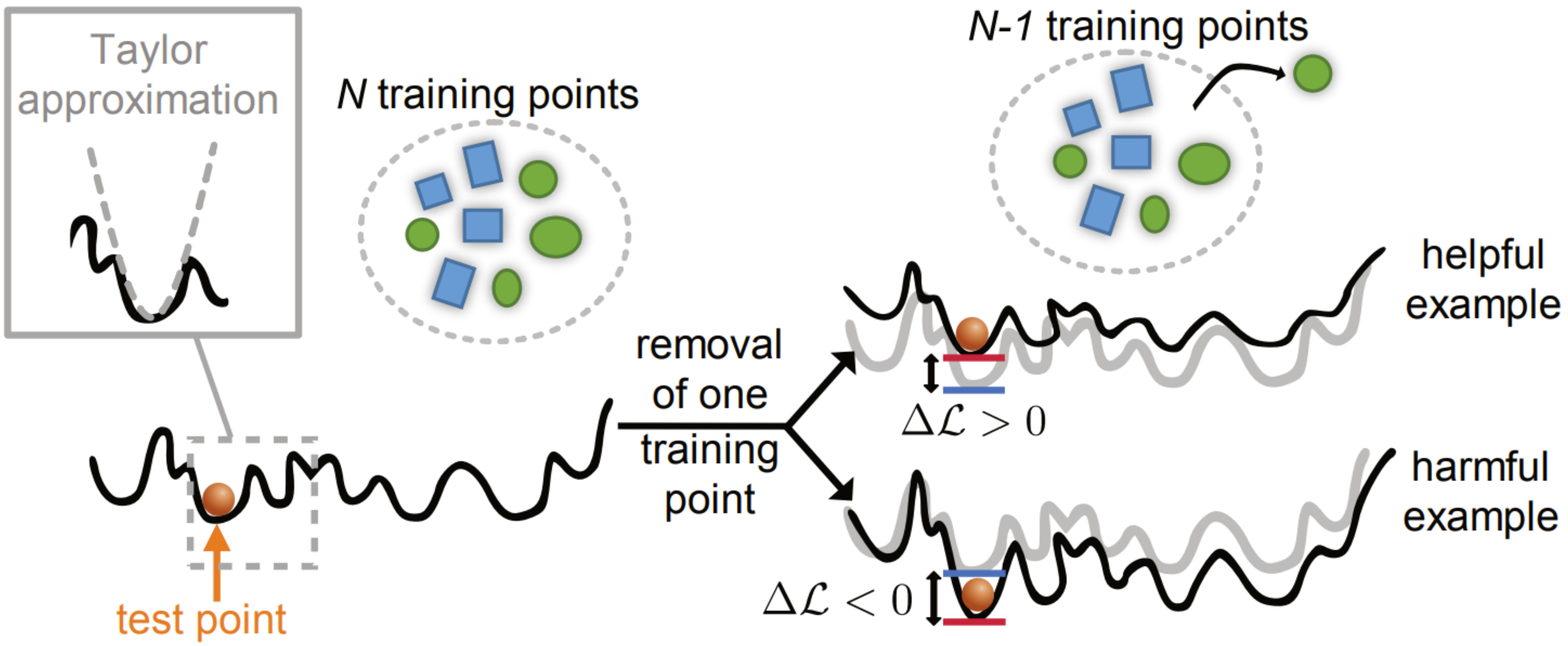




"Admiral Dog!" "The Pig-Snail" "The Camel-Bird" "The Dog-Fish"

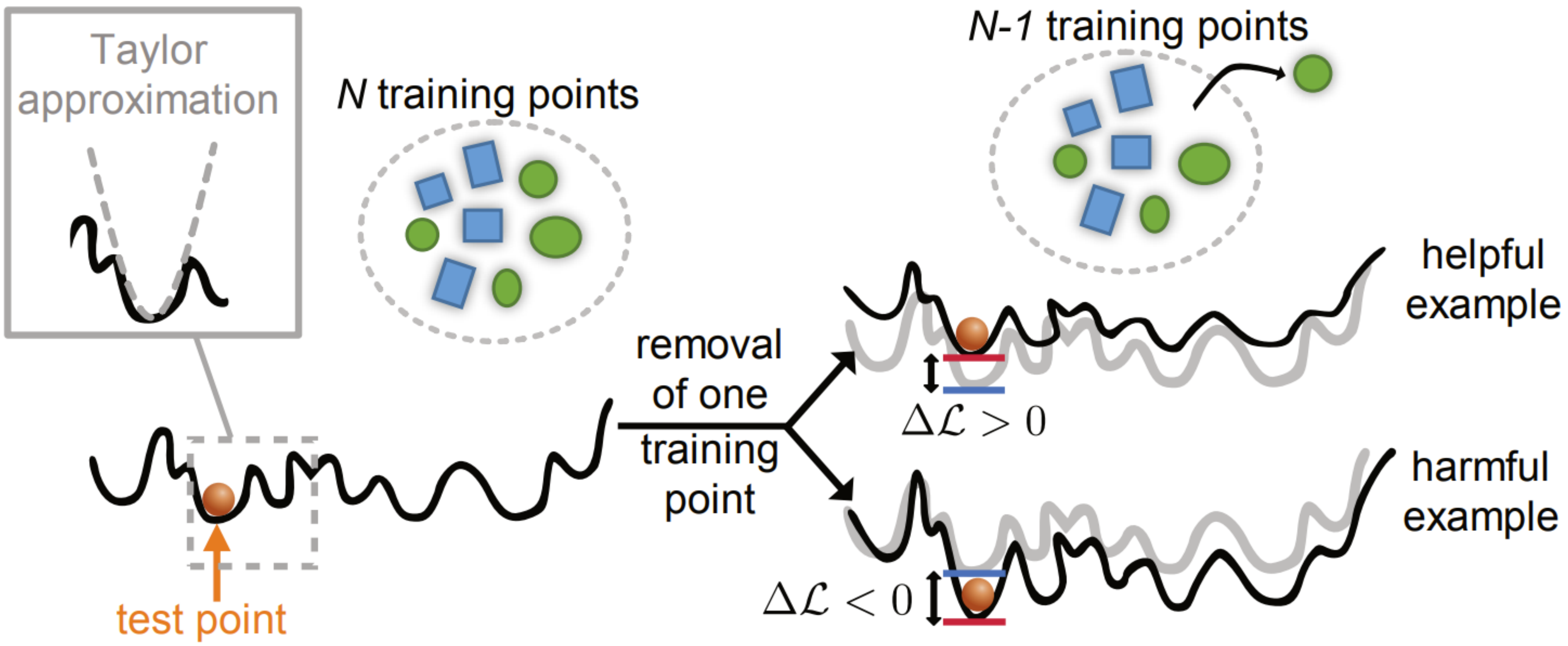
https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Leave-one-out training



Dawid et al. (2020), arXiv:2004.04711v2

Leave-one-out training



super expensive!

Dawid et al. (2020), arXiv:2004.04711v2

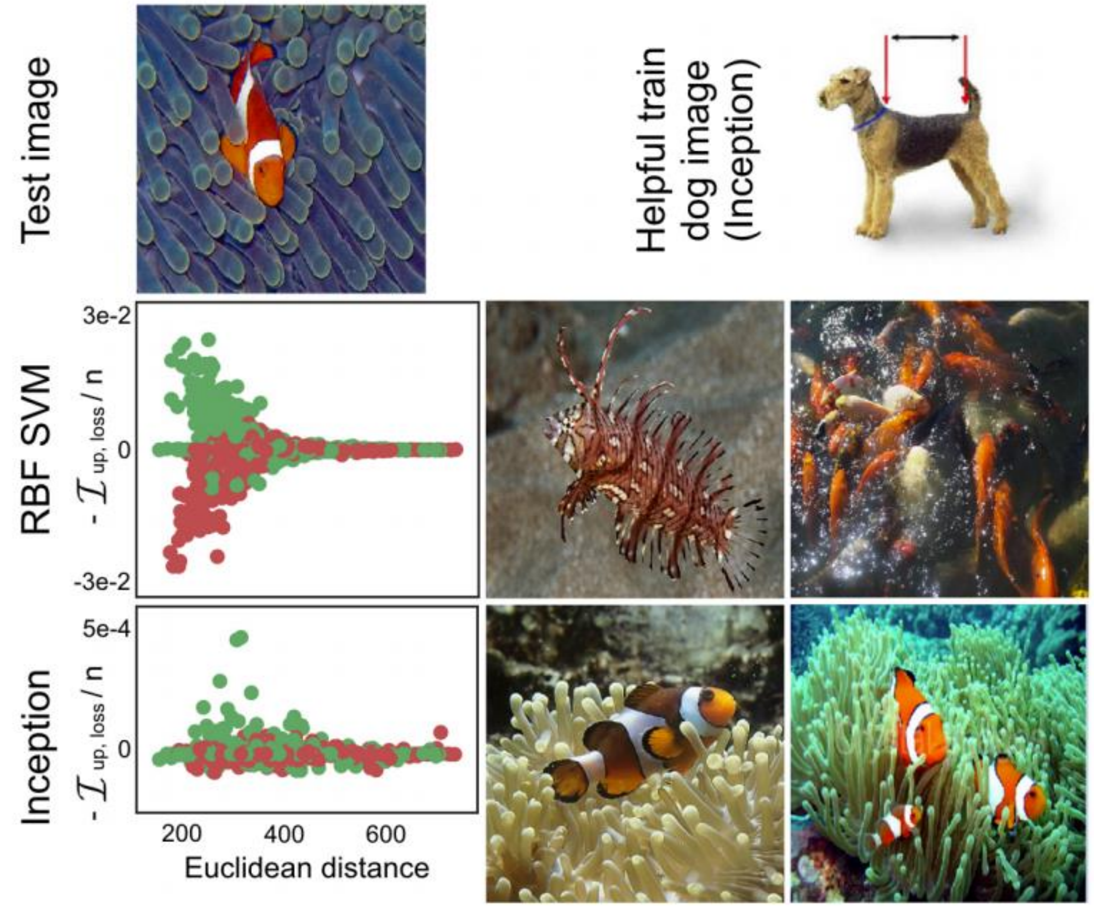


Figure 4. **Inception vs. RBF SVM. Bottom left:** $-\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}})$ vs. $\|z - z_{\text{test}}\|_2^2$. Green dots are fish and red dots are dogs. **Bottom right:** The two most helpful training images, for each model, on the test. **Top right:** An image of a dog in the training set that helped the Inception model correctly classify the test image as a fish.

Its approximation are:
influence functions

(They can be used to NNs after generalizing to non-convex problems)

Koh & Liang (2017), arXiv:1703.04730v1

Go to www.menti.com and use the code: 40 91 76

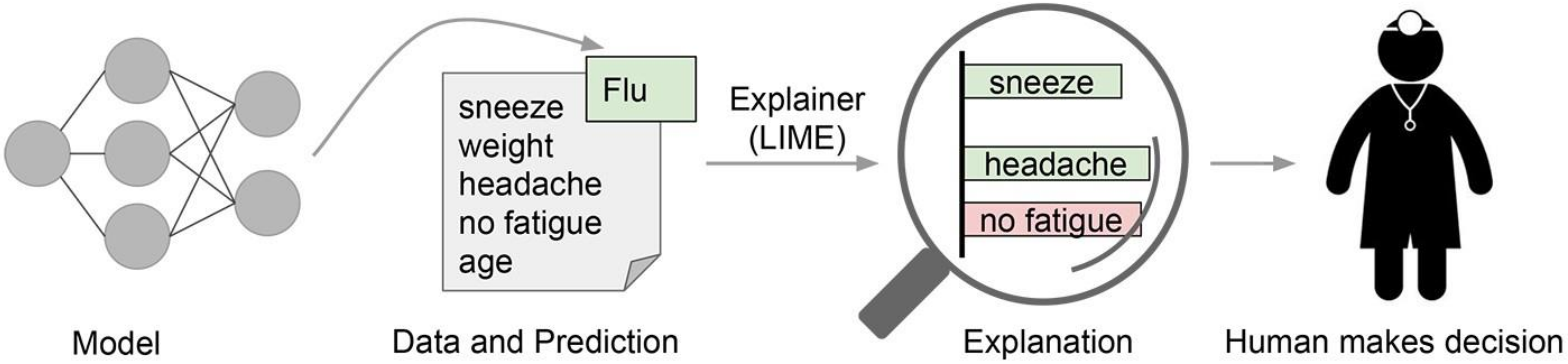
Possible approaches

- methods assigning meaning to individual **model components**
- methods analyzing model predictions when **data is perturbed**
- surrogate approach where the model is approximated by a simpler, more interpretable **surrogate model**

methods
- model-specific
- model-agnostic

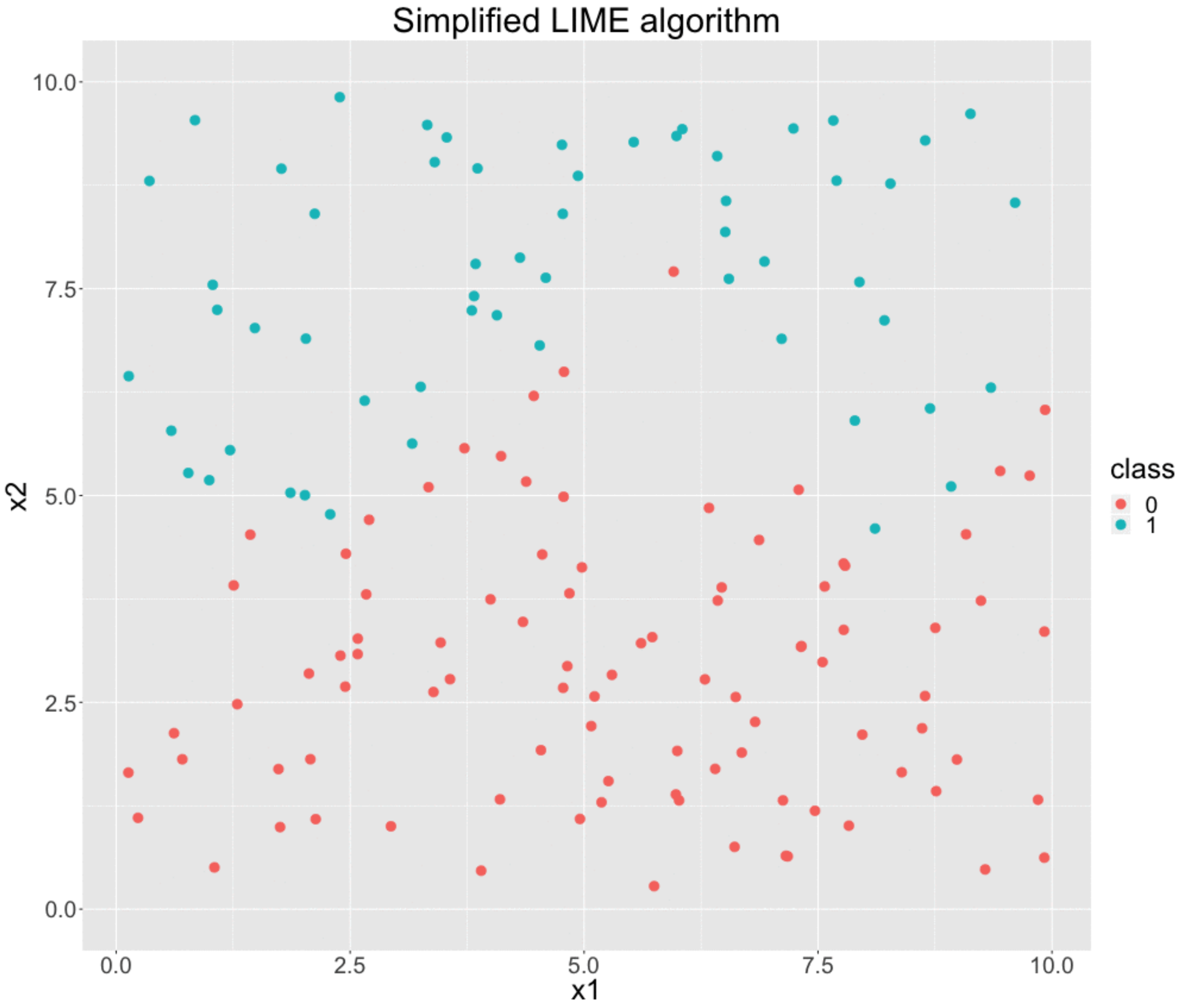
explanation
- global
- local

LIME (Local Interpretable Model-Agnostic Explanations)



Ribeiro, Singh, Guestrin (2016), arXiv:1602.04938v3

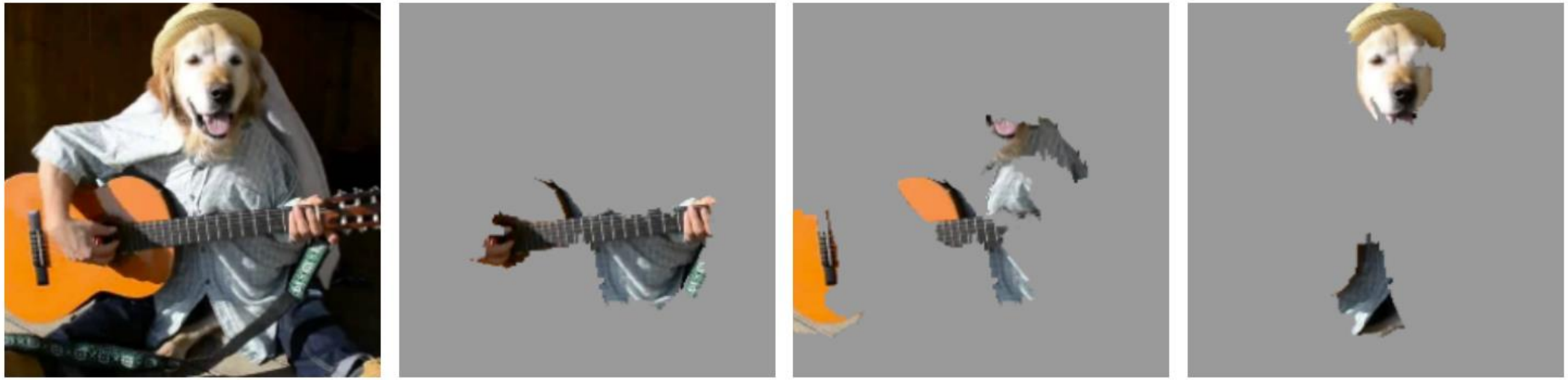
LIME (Local Interpretable Model-Agnostic Explanations)



https://compstat-lmu.github.io/iml_methods_limitations/lime.html

(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.



(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" ($p = 0.21$)

Ribeiro (2016), arXiv:1602.04938v3

TUTORIAL EXAMPLE

Class Activation Map (CAM)

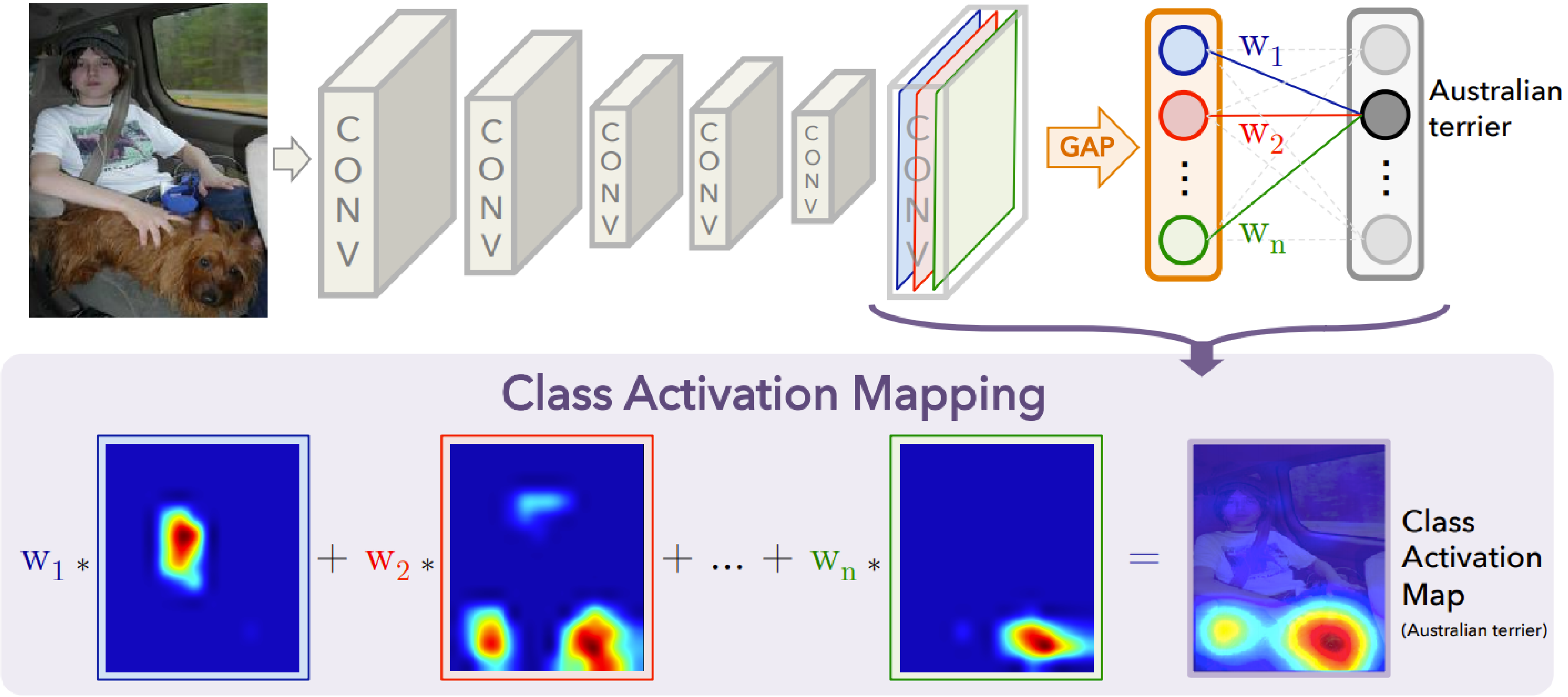


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Zhou et al. (2015), arXiv:1512.04150v1

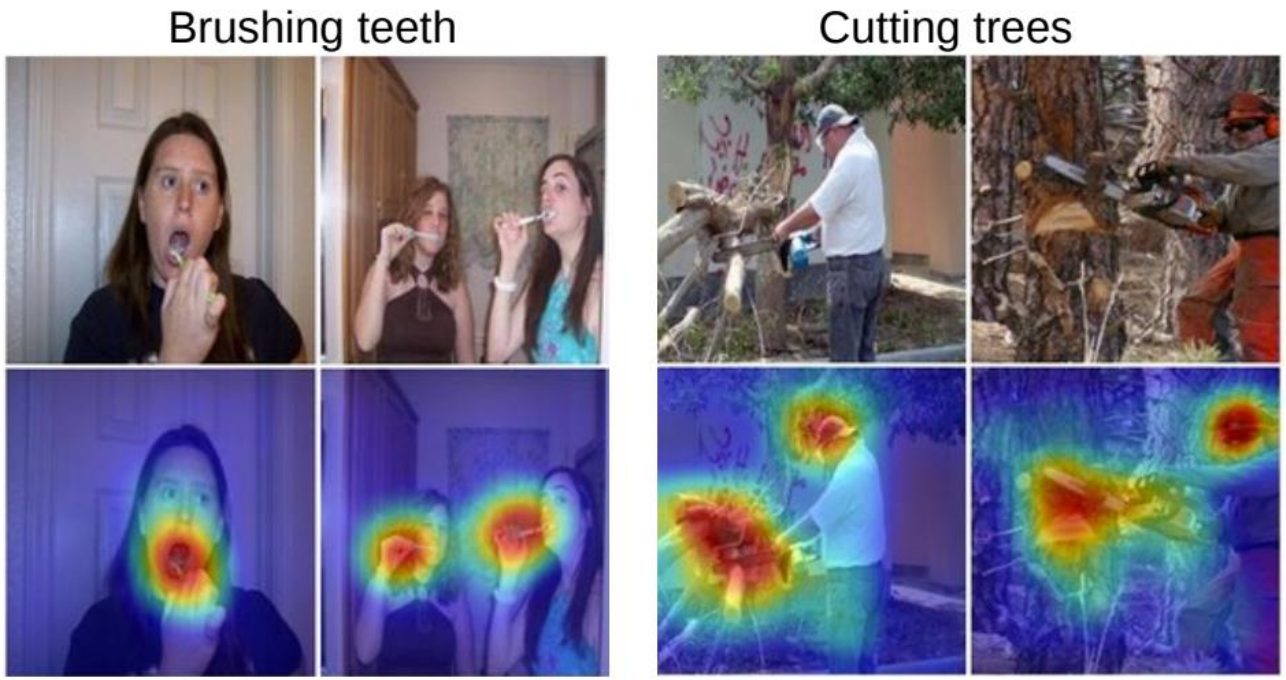


Figure 1. A simple modification of the global average pooling layer combined with our class activation mapping (CAM) technique allows the classification-trained CNN to both classify the image and localize class-specific image regions in a single forward-pass e.g., the toothbrush for *brushing teeth* and the chainsaw for *cutting trees*.

Zhou et al. (2015), arXiv:1512.04150v1

Go to www.menti.com and use the code: 40 91 76

Possible approaches

- methods assigning meaning to individual **model components**
- methods analyzing model predictions when **data is perturbed**
- surrogate approach where the model is approximated by a simpler, more interpretable **surrogate model**

methods

- model-specific
- model-agnostic

explanation

- global
- local

Take-home messages

ML needs to be interpretable if it is to be trusted

Is it reliable? Is it fair? Does it generalizes well? Can we extract the learnt knowledge?

Interpretability methods

Some look at the components of the model (CAM, feature visualisation), some are based on perturbing the data and looking at how it changed the model (LOO training), there are also surrogate approaches (LIME).

Class Activation Map (CAM)

Shows which parts of the image were decisive for the prediction of the given class.

Many other methods exist!