INTERPRETABLE DEEP LEARNING
A SUBJECTIVE OVERVIEW

Anna Dawid

Some definitions

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

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

Reliability
Allows to trust the model predictions

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

The **form of data** determines how you want the machine to explain its predictions

Images
They contain interpretable objects, features have spatial structure, and make sense only in relation to their neighbors

Tabular data
They have interpretable features! Like age, velocity, mass…

What is your data??
???

Outline

INTERPRETABILITY
BY ML COMMUNITY

INTERPRETABILITY
IN PHYSICS

1/3

OVERVIEW OF INTERPRETABILITY METHODS

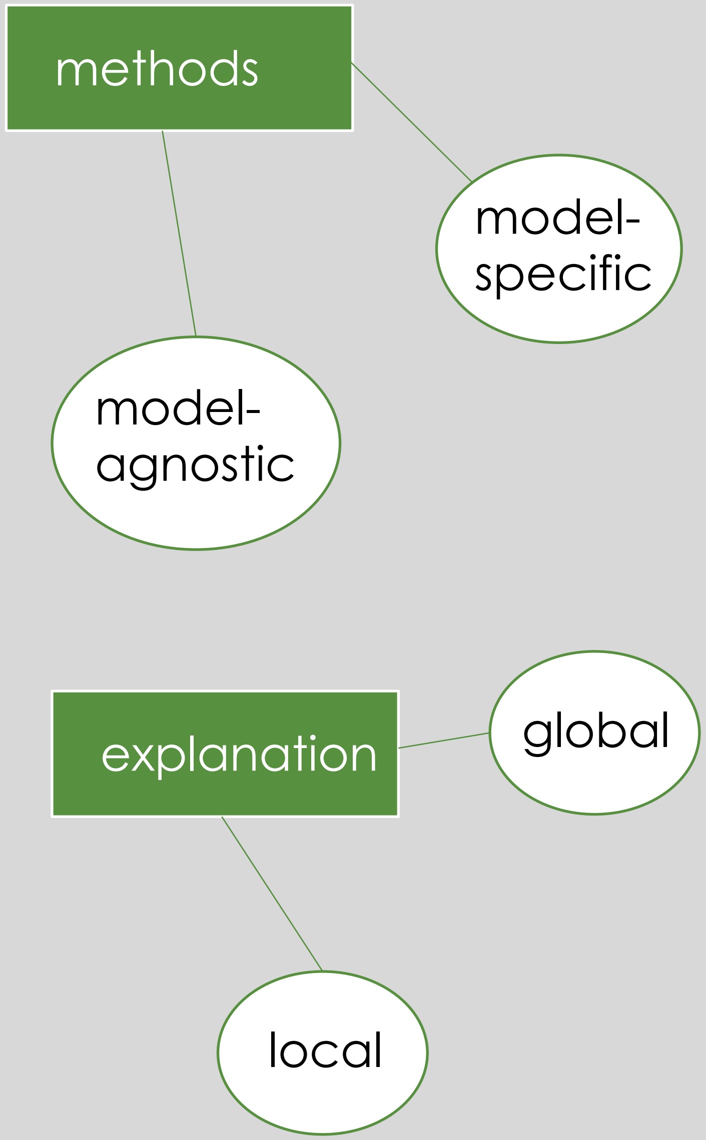
in the ML community

https://christophm.github.io/interpretable-ml-book/

- Feature visualization
- Influence functions
- Class Activation Mapping (CAM)
- Locally Interpretable Model-Agnostic Explanations (LIME)
- Shapley values

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

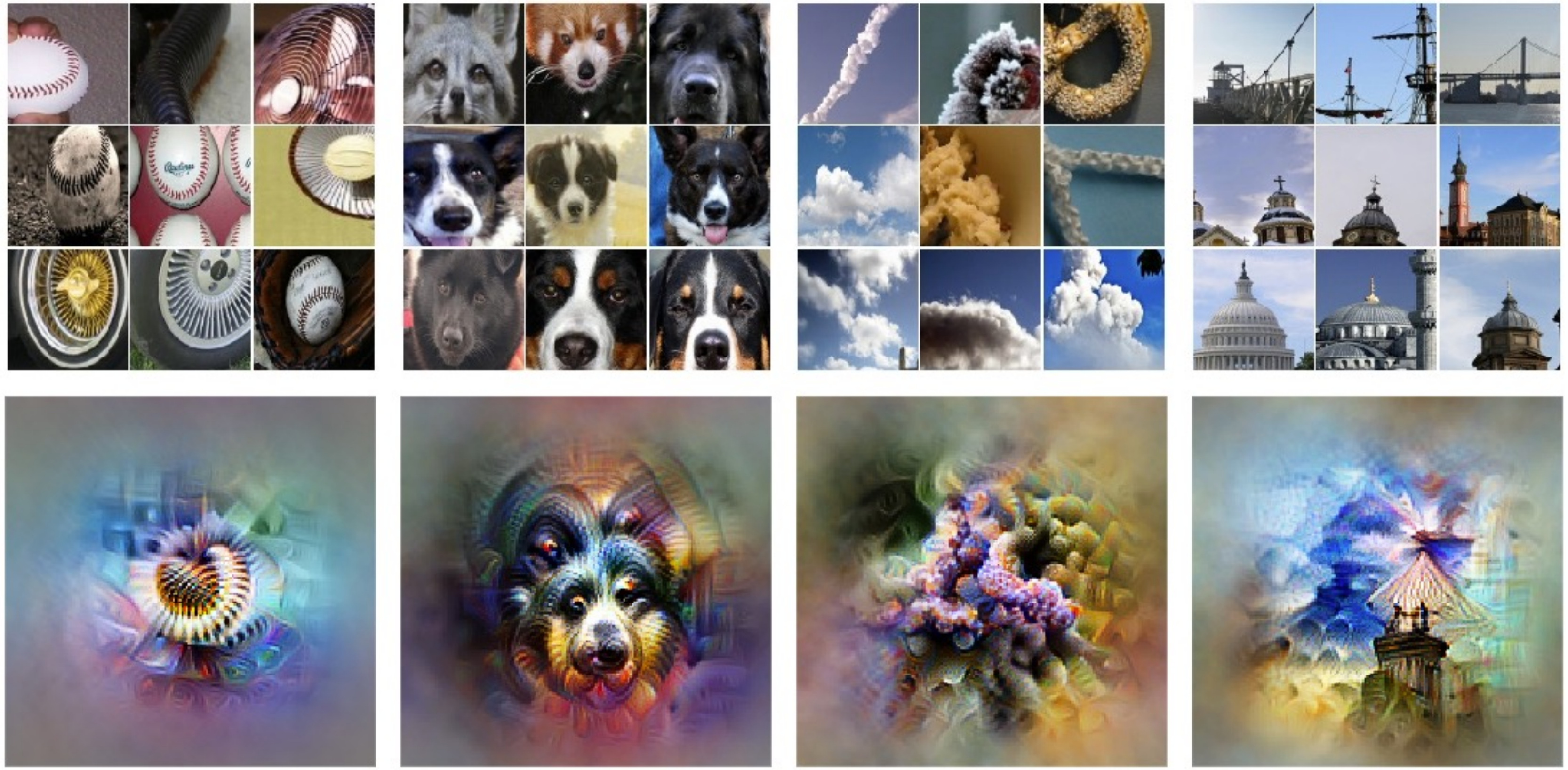
- Feature vizualization
- Influence functions
- Class Activation Mapping (CAM)
- Locally Interpretable Model-Agnostic Explanations (LIME)
- Shapley values

Feature visualization

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

Dataset Examples show us what neurons respond to in practice

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.



Baseball—or stripes? *mixed4a, Unit 6*

Animal faces—or snouts? *mixed4a, Unit 240*

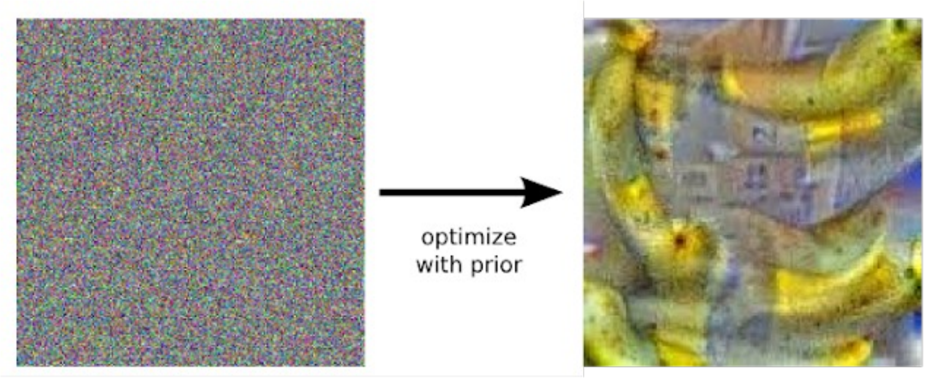
Clouds—or fluffiness? *mixed4a, Unit 453*

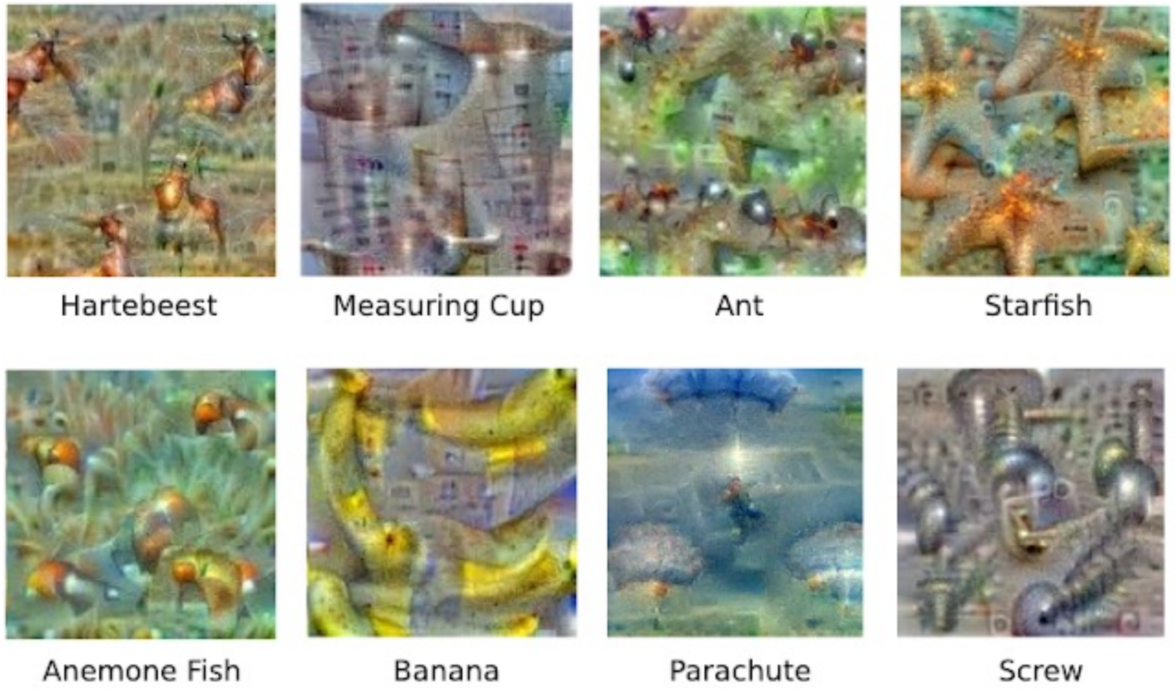
Buildings—or sky? *mixed4a, Unit 492*

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

Feature visualization

finding the image that maximizes the output corresponding to a selected class



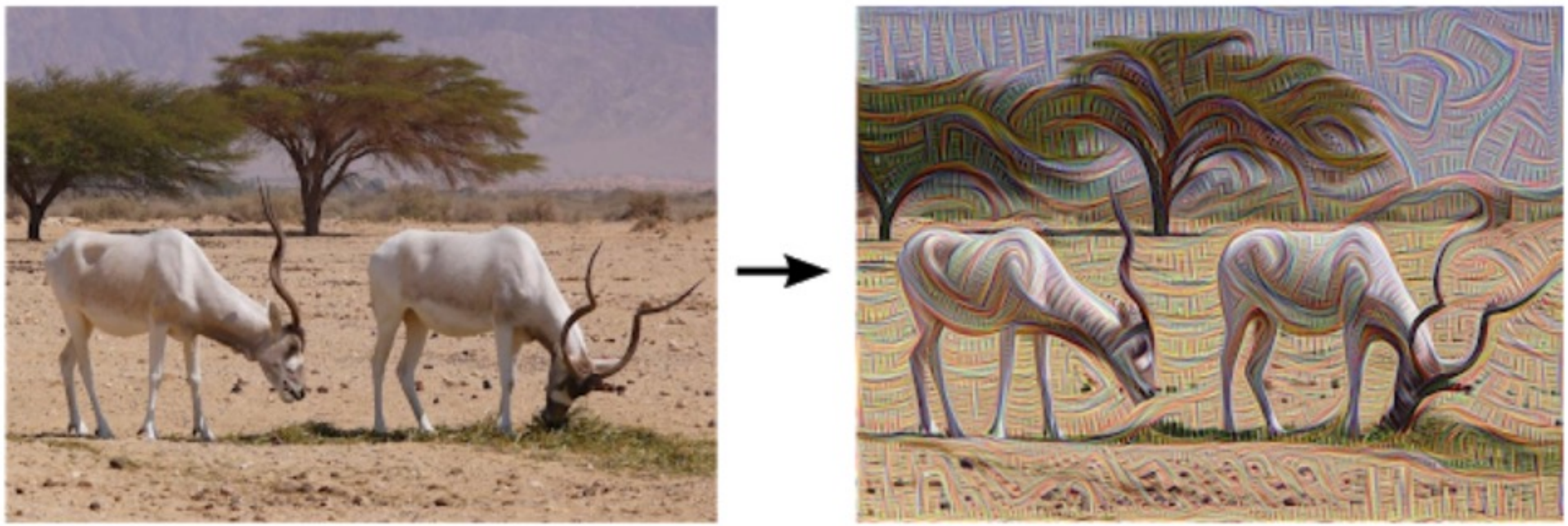


https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

We can look at whole layers!

Deep Dream

enhance what was detected by a chosen layer and visualise



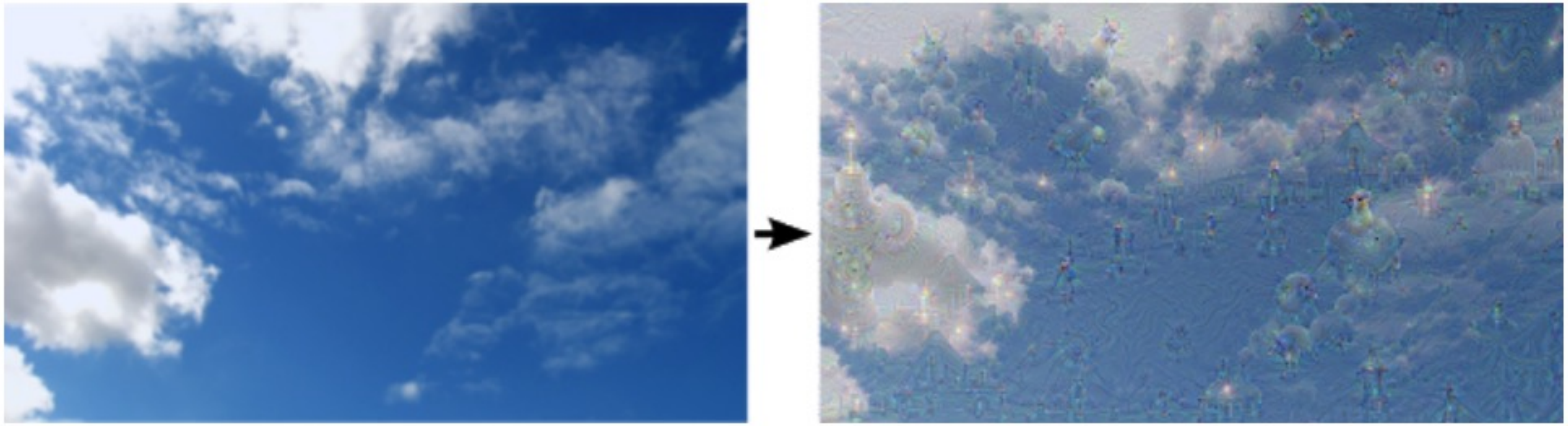
Left: Original photo by Zachi Evenor. Right: processed by Günther Noack, Software Engineer

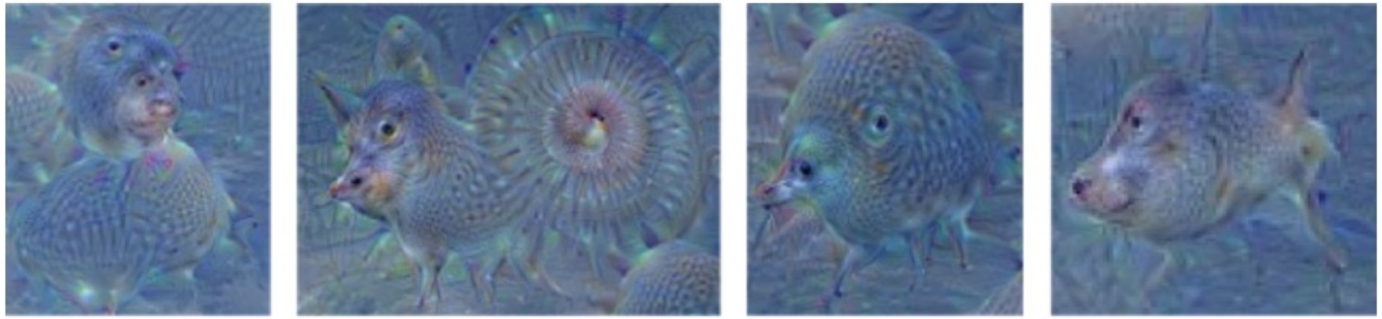
https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

We can look at whole layers!

Deep Dream

enhance what was detected by a chosen layer and visualise



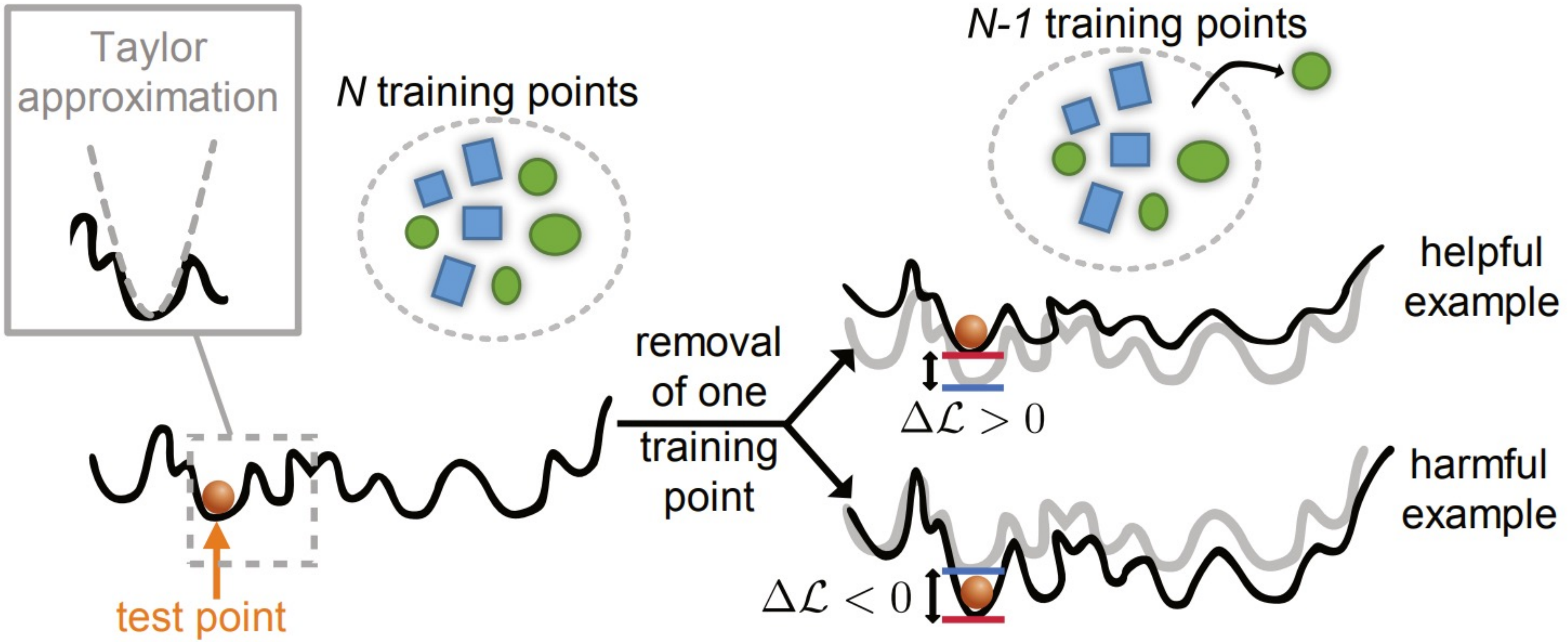


"Admiral Dog!" "The Pig-Snail" "The Camel-Bird" "The Dog-Fish"

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

- Feature vizualization
- Influence functions
- Class Activation Mapping (CAM)
- Locally Interpretable Model-Agnostic Explanations (LIME)
- Shapley values

Leave-one-out training



super expensive!

Dawid et al. *New J. Phys.* 22, 115001 (2020)

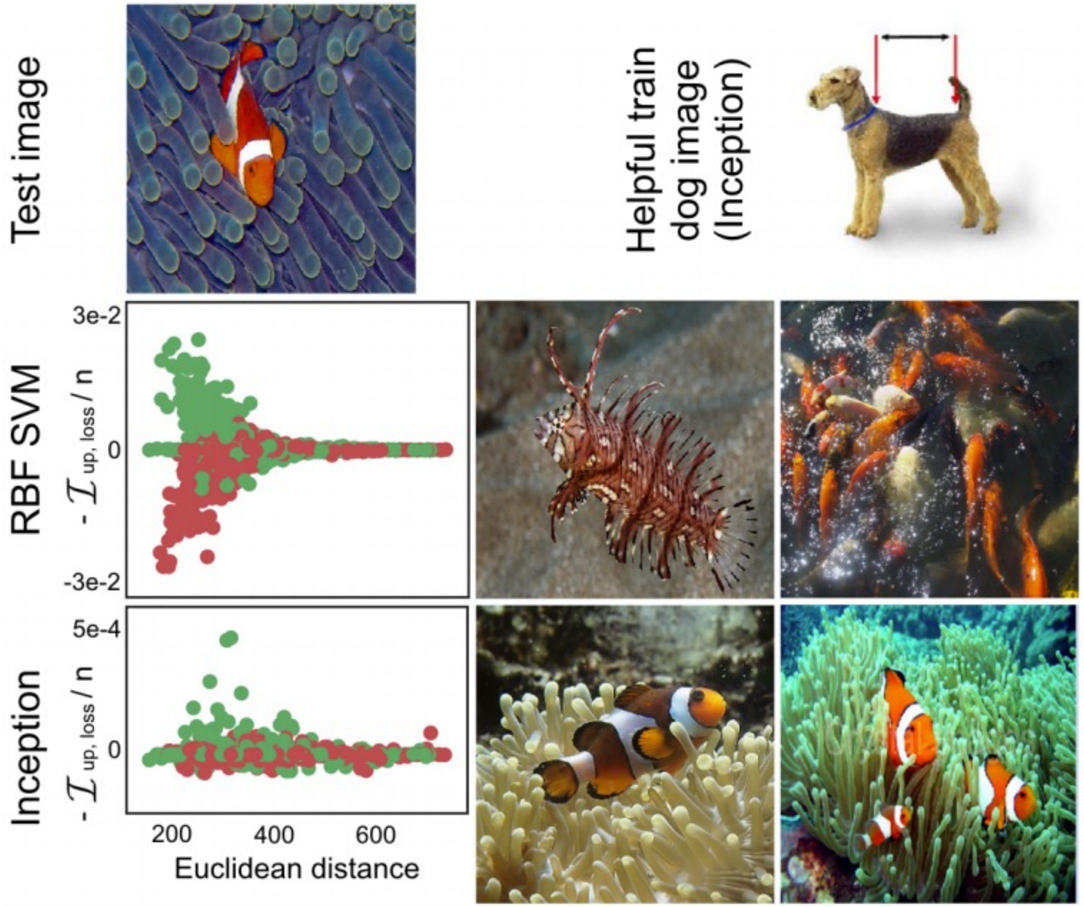


Figure 4. **Inception vs. RBF SVM. Bottom left:** $-\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}})$ vs. $\|z - z_{\text{test}}\|_2^2$. Green dots are fish and red dots are dogs. **Bottom right:** The two most helpful training images, for each model, on the test. **Top right:** An image of a dog in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

- Feature vizualization
- Influence functions
- **Class Activation Mapping (CAM)**
- Locally Interpretable Model-Agnostic Explanations (LIME)
- Shapley values

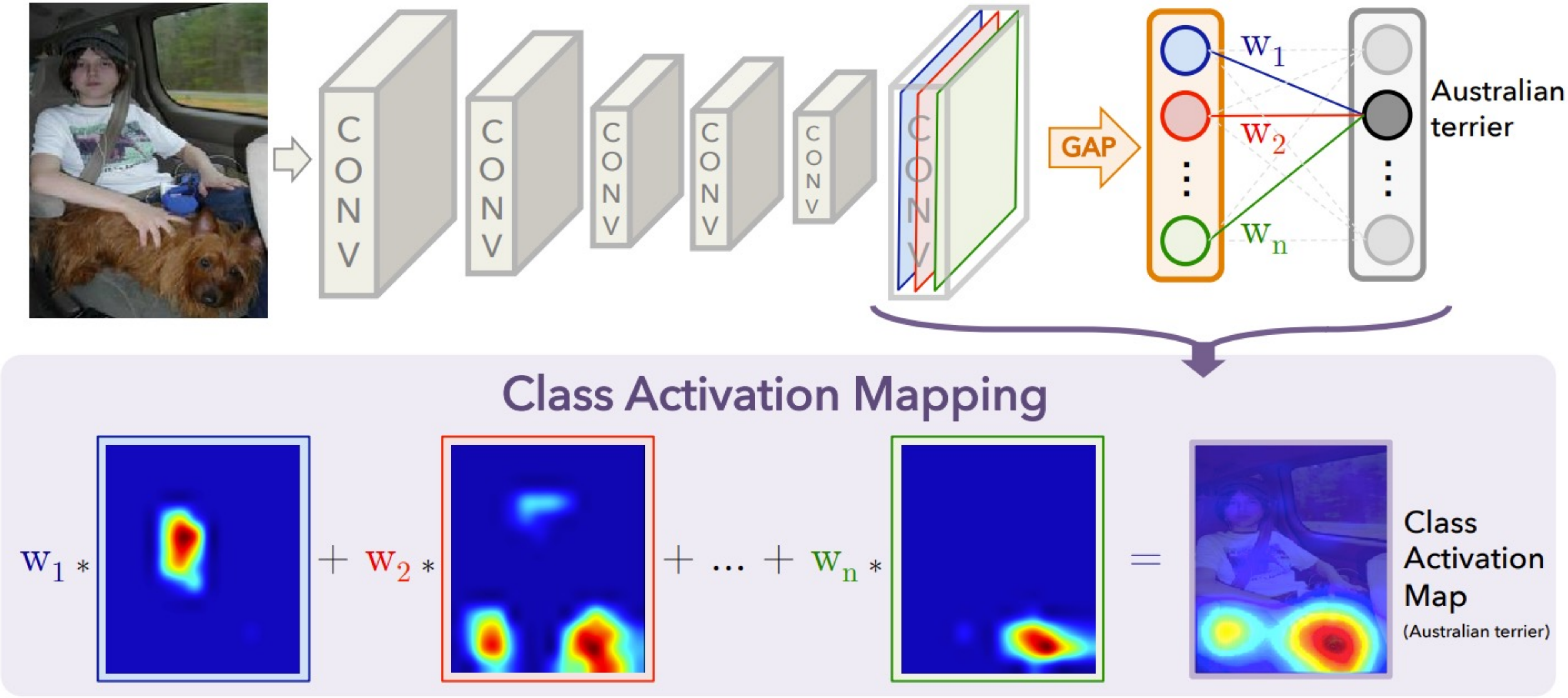


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

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

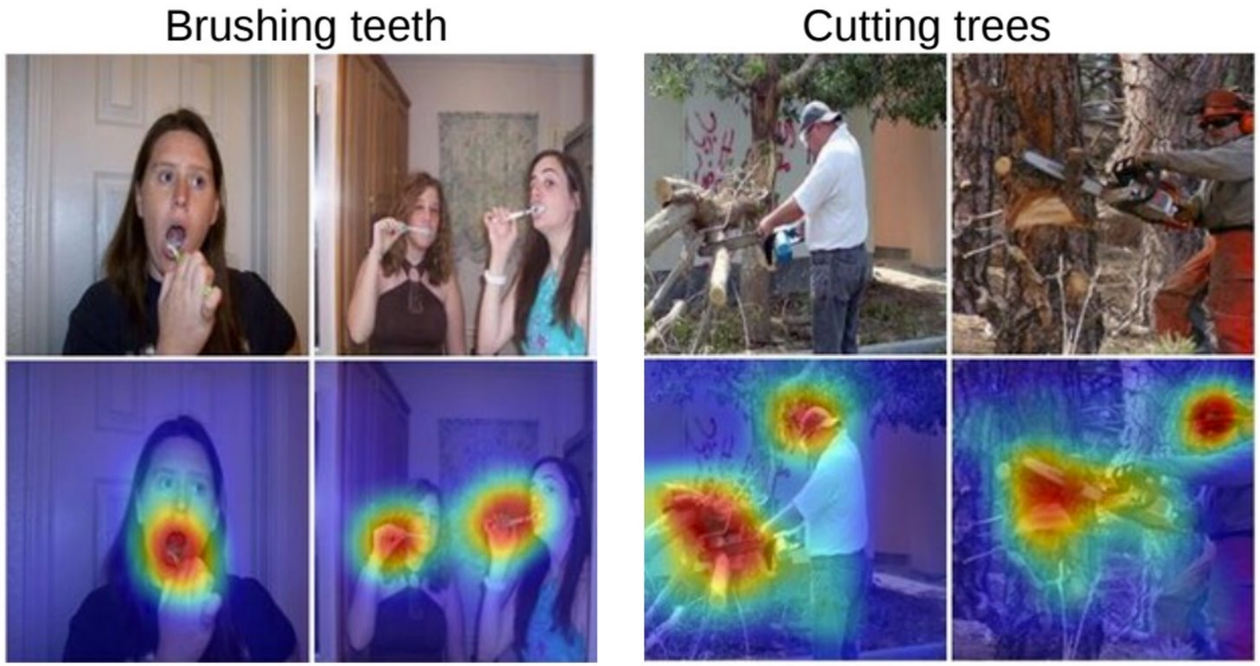
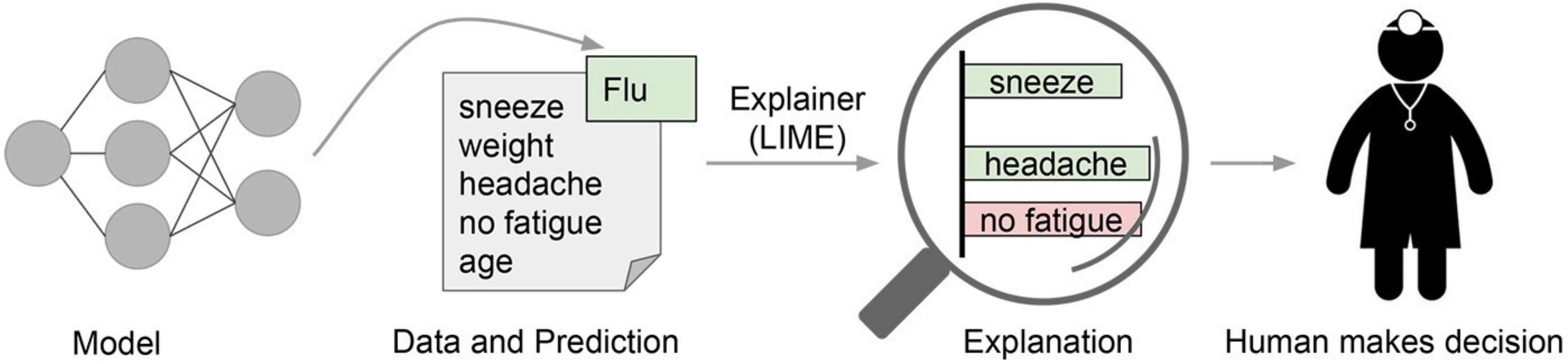


Figure 1. A simple modification of the global average pooling layer combined with our class activation mapping (CAM) technique allows the classification-trained CNN to both classify the image and localize class-specific image regions in a single forward-pass e.g., the toothbrush for *brushing teeth*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

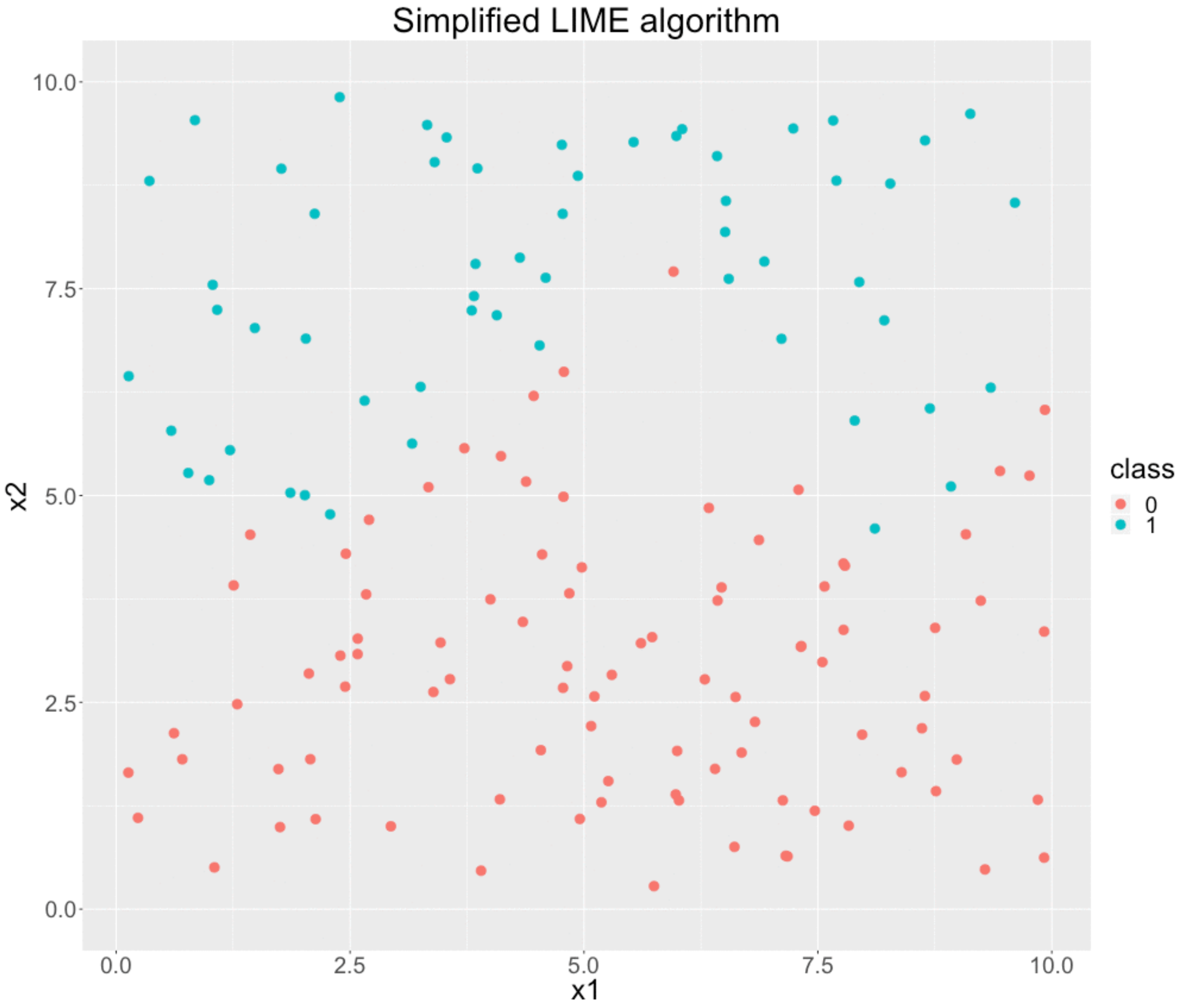
- Feature vizualization
- Influence functions
- Class Activation Mapping (CAM)
- **Locally Interpretable Model-Agnostic Explanations (LIME)**
- Shapley values

LIME (Local Interpretable Model-Agnostic Explanations)



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

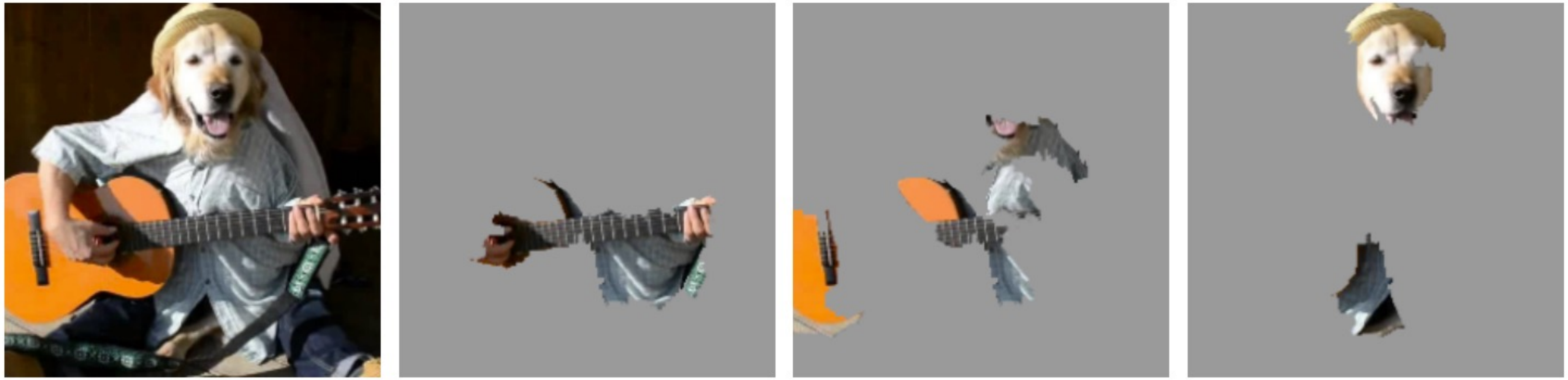
LIME (Local Interpretable Model-Agnostic Explanations)



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.



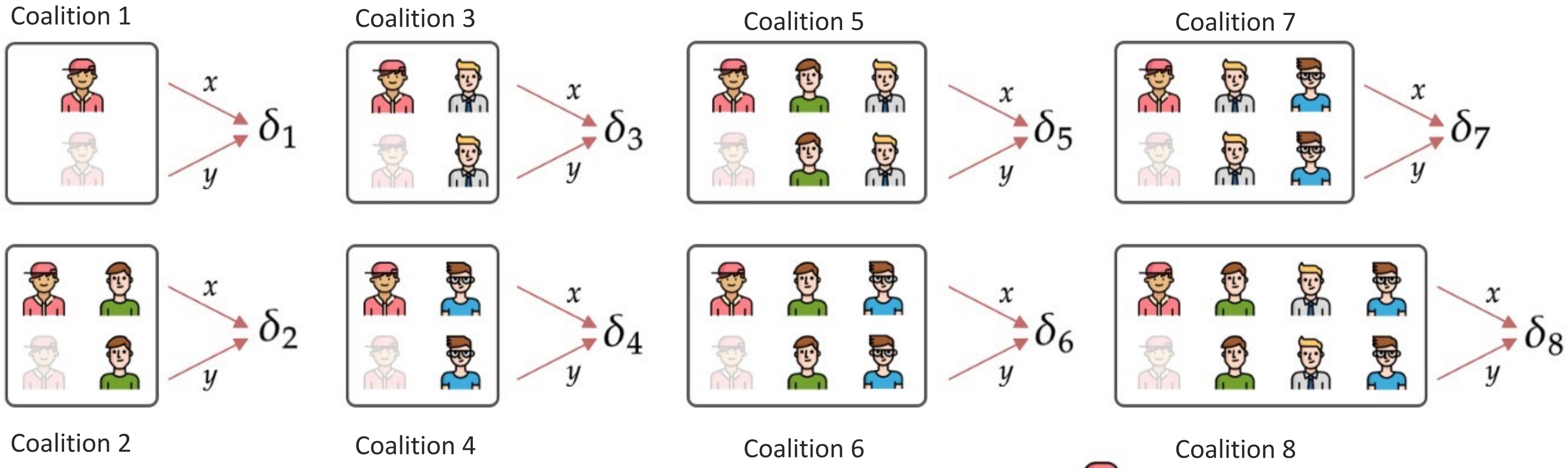
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

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

- Feature vizualization
- Influence functions
- Class Activation Mapping (CAM)
- Locally Interpretable Model-Agnostic Explanations (LIME)
- **Shapley values**

Shapley values and SHAP



The Shapley value for member 

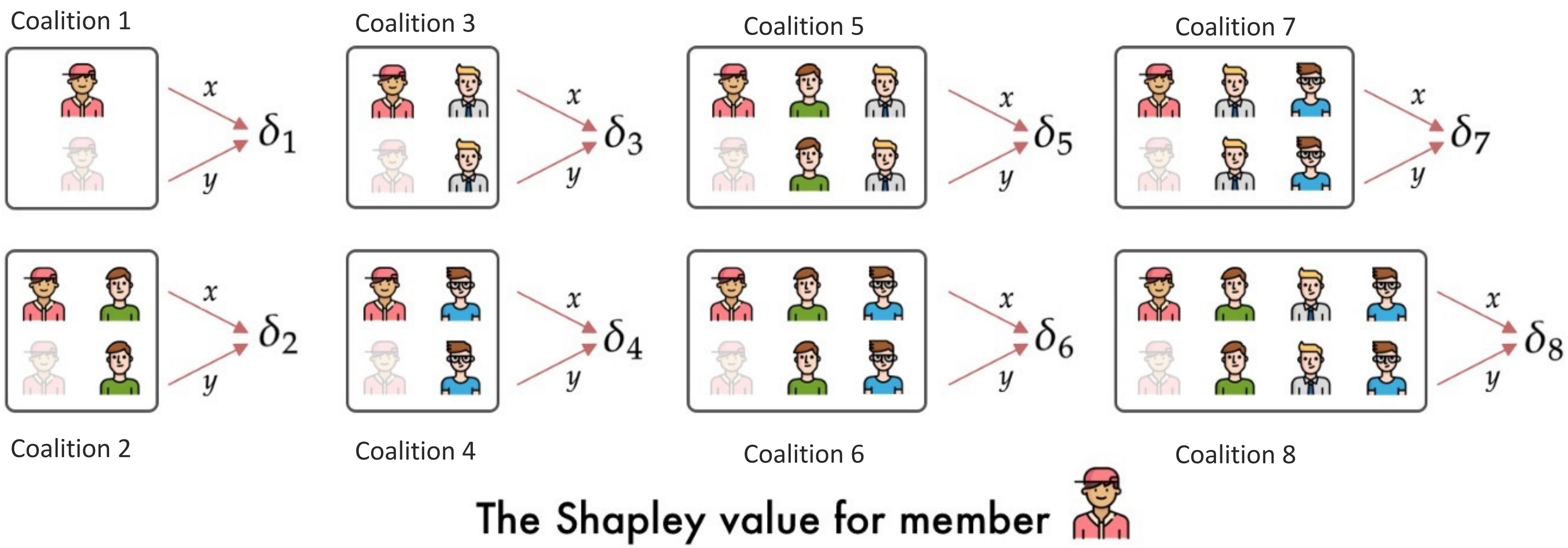
is given by:

Shapley value is the average marginal contribution of an instance of a feature among all possible coalitions.

$$\phi_i = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$

https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

Shapley values and SHAP



Shapley value is the average marginal contribution of an instance of a feature among all possible coalitions.

is given by:

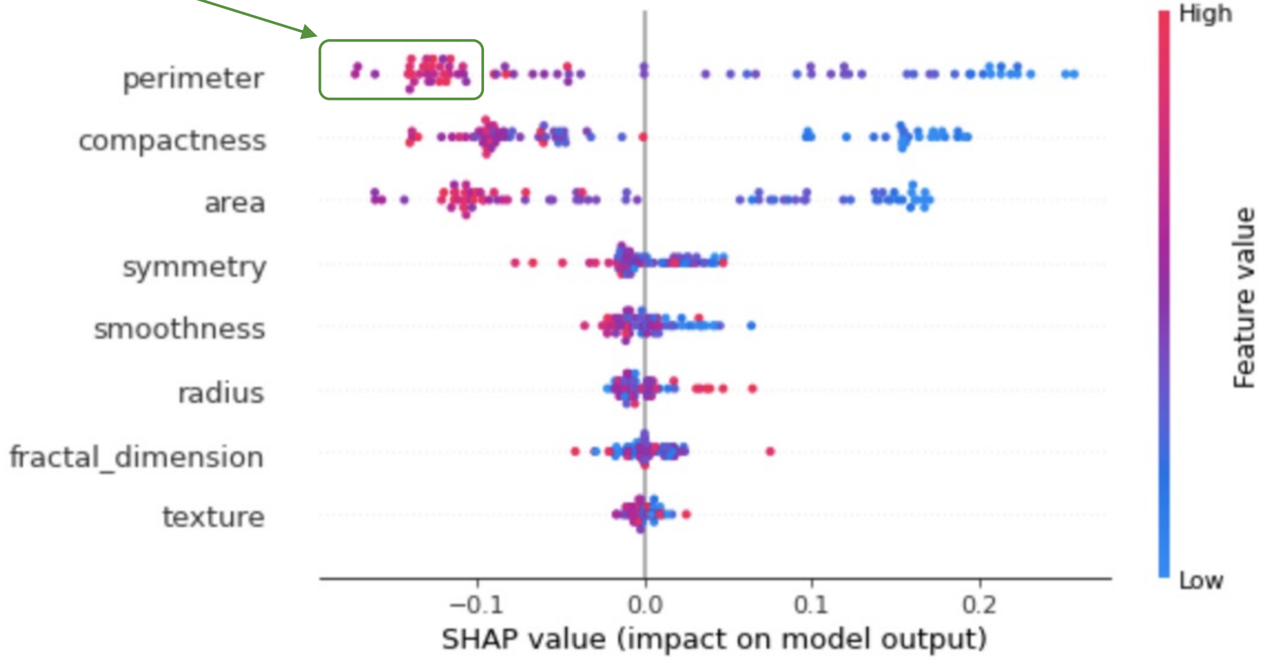
$$\phi_i = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$

NP-hard for ML problems!
SHAP trains a linear classifier on various coalitions and model predictions

https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

Shapley values and SHAP

large "perimeter" pushes the prediction towards class 0



https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

INTERPRETABILITY METHODS IN PHYSICS

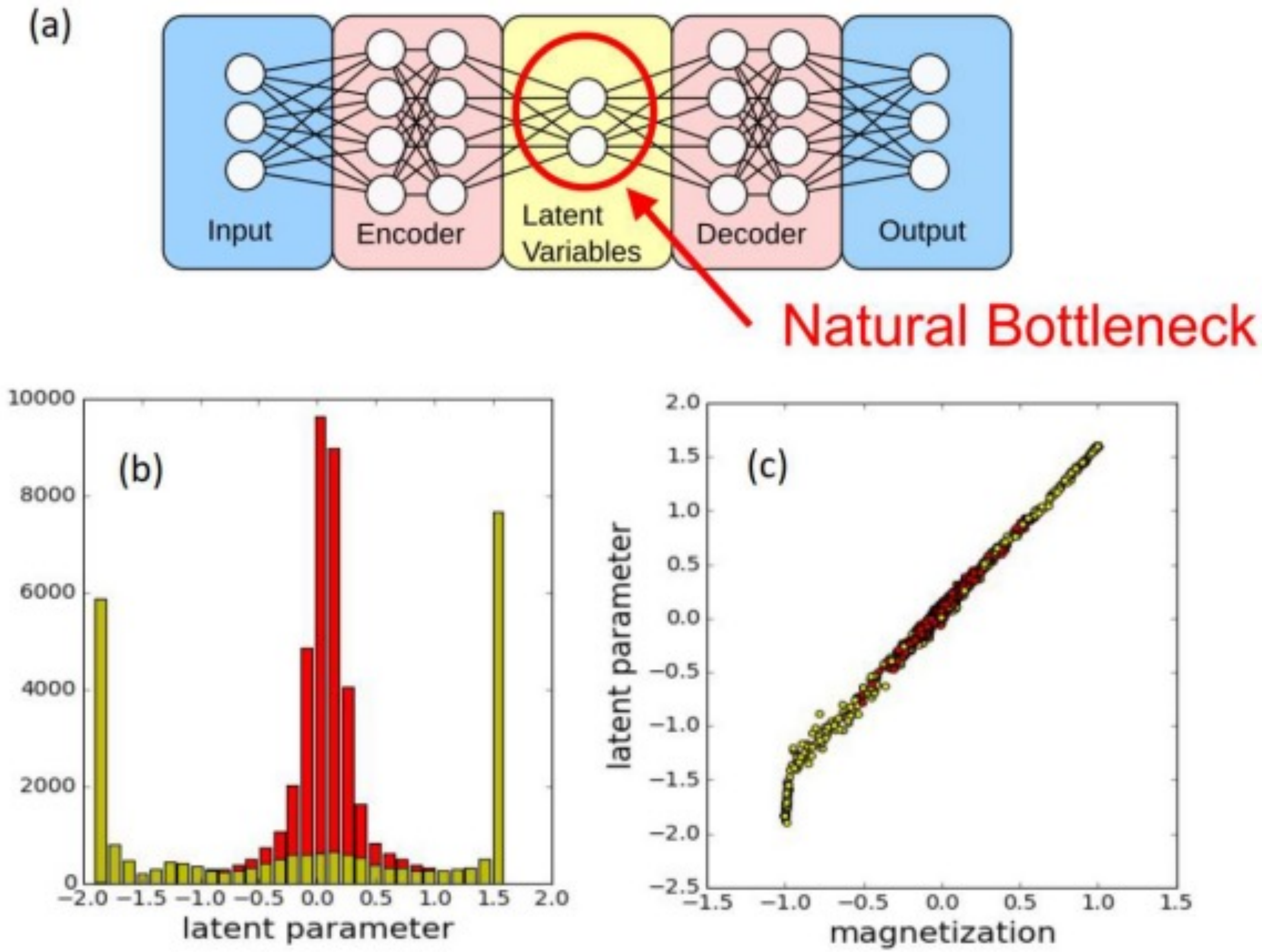
So far

The **form of data** determines how you want the machine to explain its predictions

Images
They contain interpretable objects, features have spatial structure, and make sense only in relation to their neighbors

Tabular data
They have interpretable features! Like age, velocity, mass…

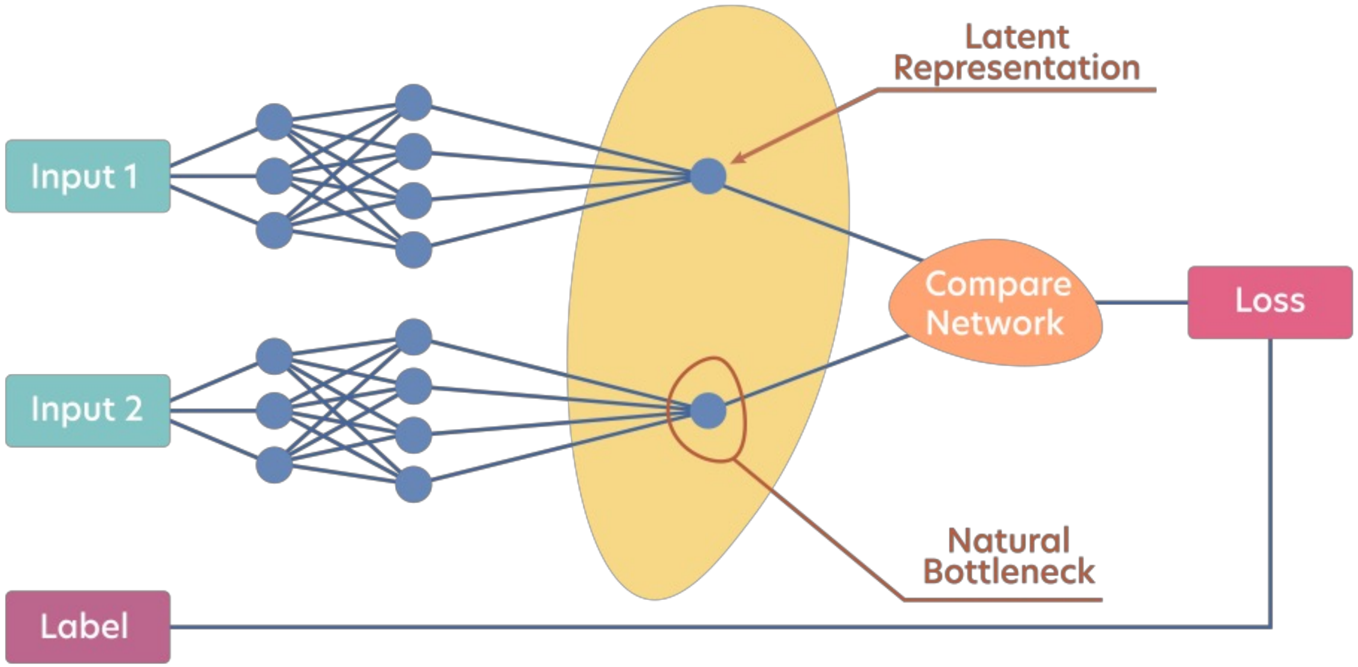
What is your data??
Spin configurations



Autoencoders

S. Wetzel (2017)
Phys. Rev. E 96, 022140

Siamese neural networks

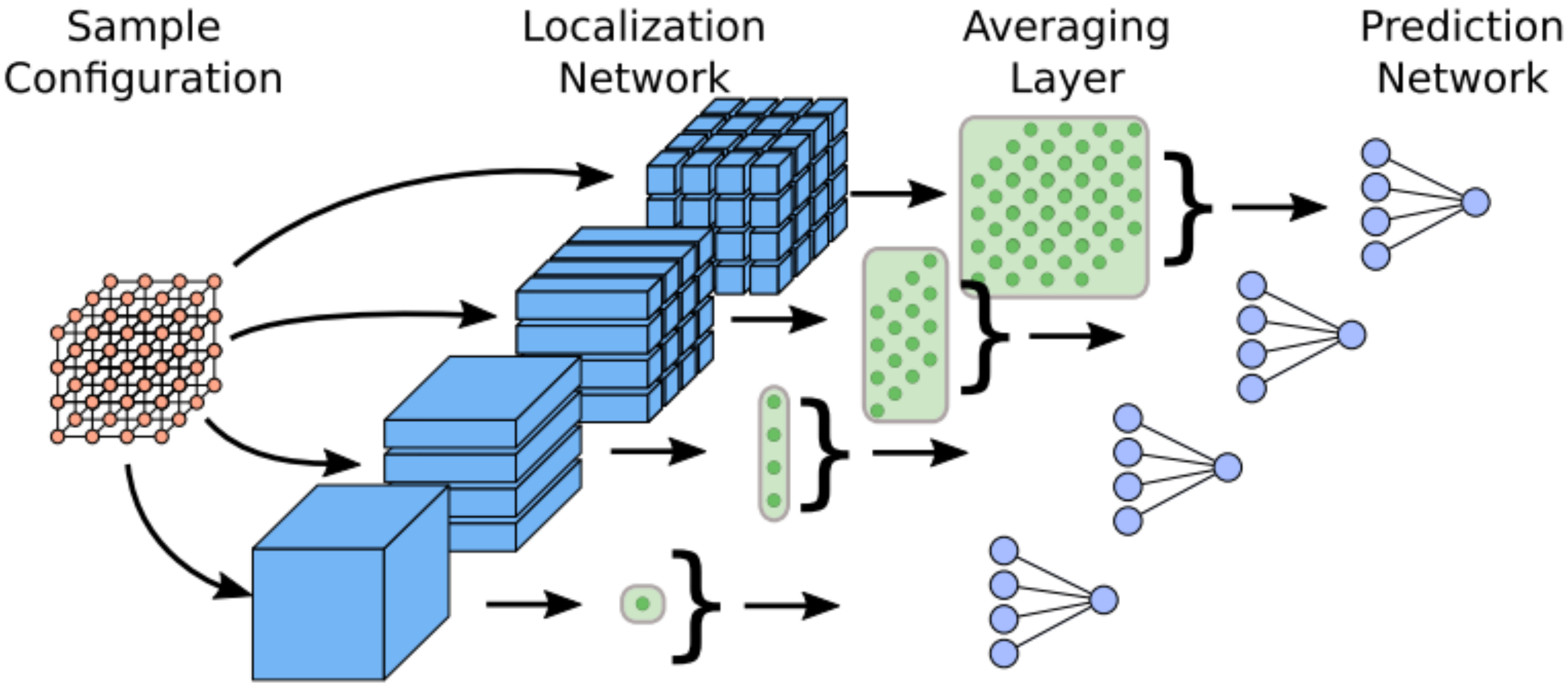


motion of a particle in a central potential

dominant regression term has a connection to the angular momentum of the particle

S. J. Wetzel et al. (2020) Phys. Rev. Res. 2, 033499

Correlation probing neural network

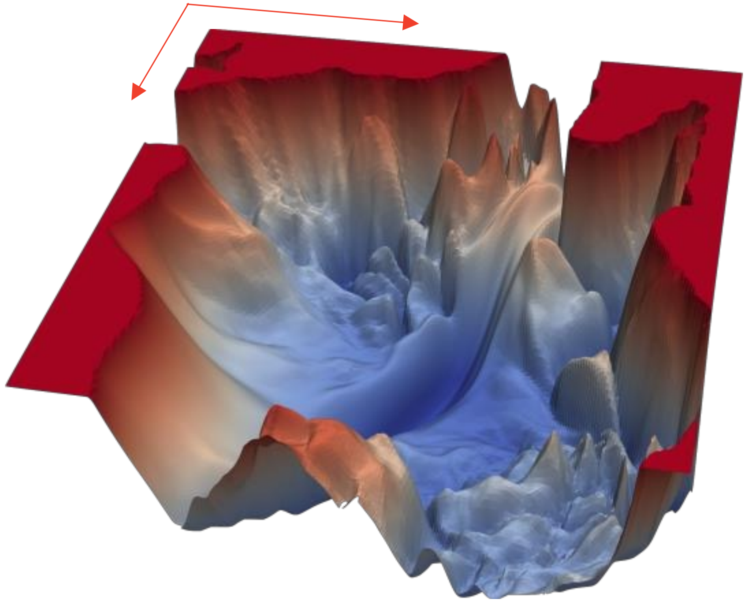


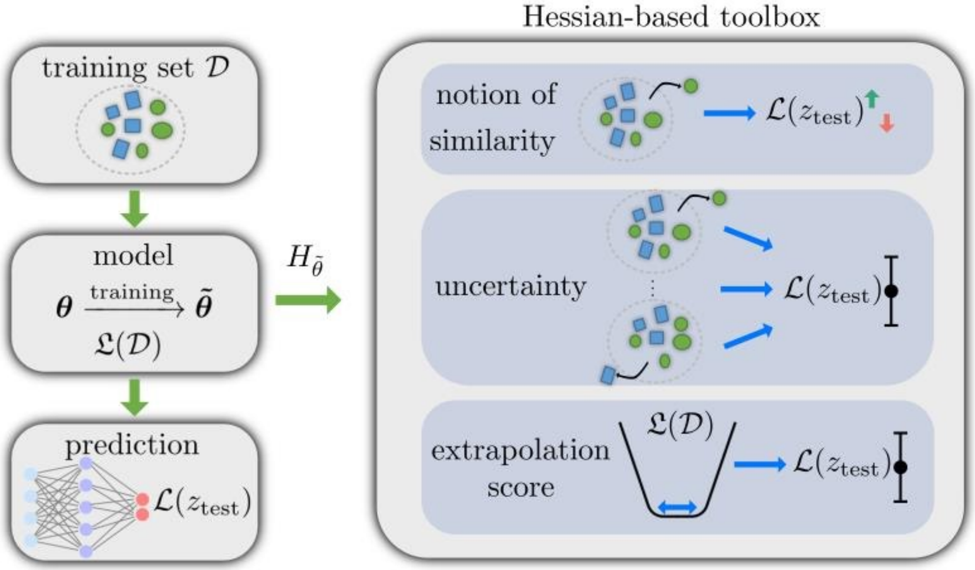
9.02.2023

S. Wetzel (2017) Phys. Rev. B 96, 184410

Hessian-based analysis

parameter space

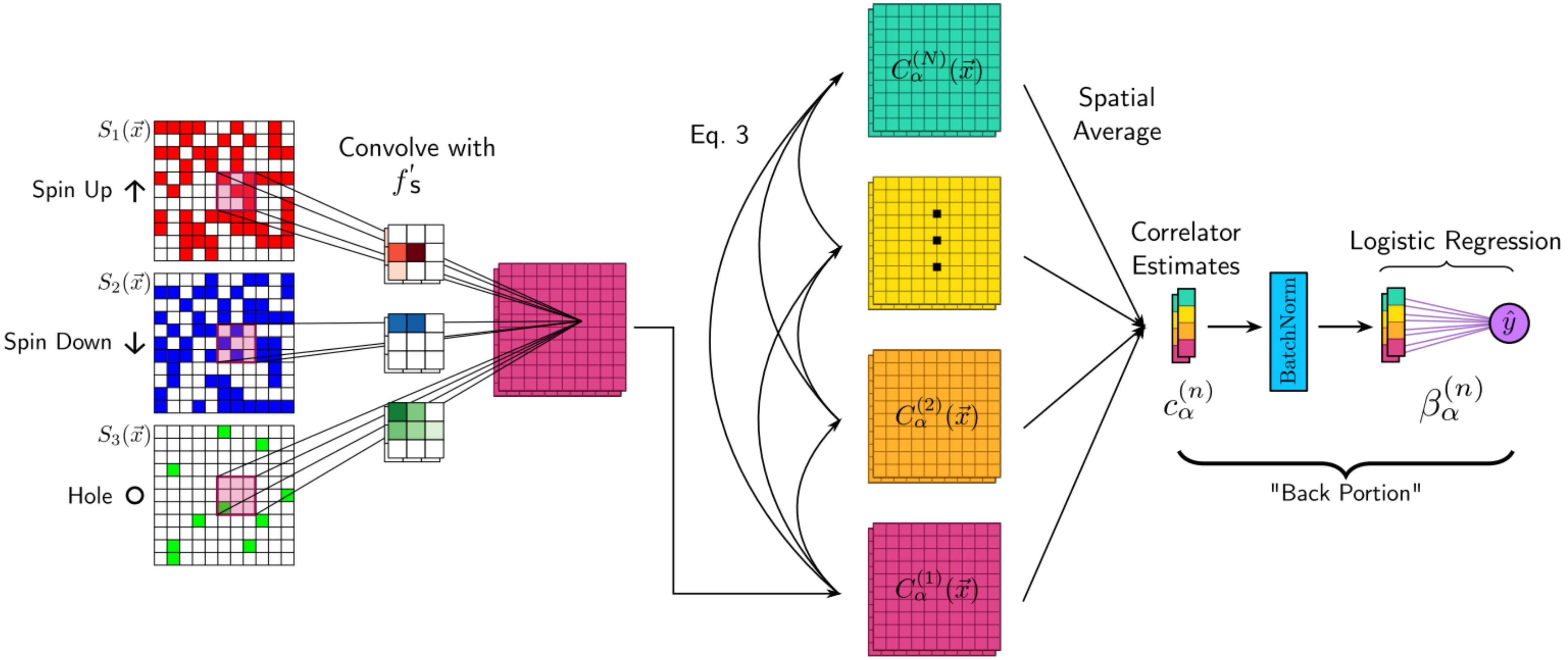




9.02.2023

A. Dawid et al. (2022) Mach. Learn.: Sci. Technol. 3,015002

Correlator convolutional neural networks



C. Miles et al. (2021) Nat. Comm. 12, 3905