Interpretable Models

University of Calgary’s Statistical Learning Group (2019)

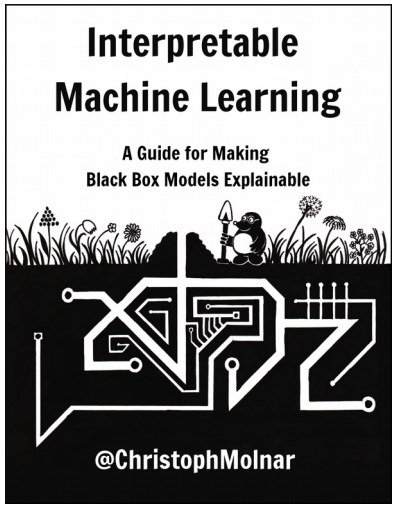
Luis Souto Maior, MSc Student

Seaman Family MR Research Centre / Vascular Imaging Laboratory

Supervisor: Dr. Richard Frayne

Interpretable Machine Learning

“**Interpretable Machine Learning** refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans.” – *Christopher Molnar, Interpretable Machine Learning Book, 2019*

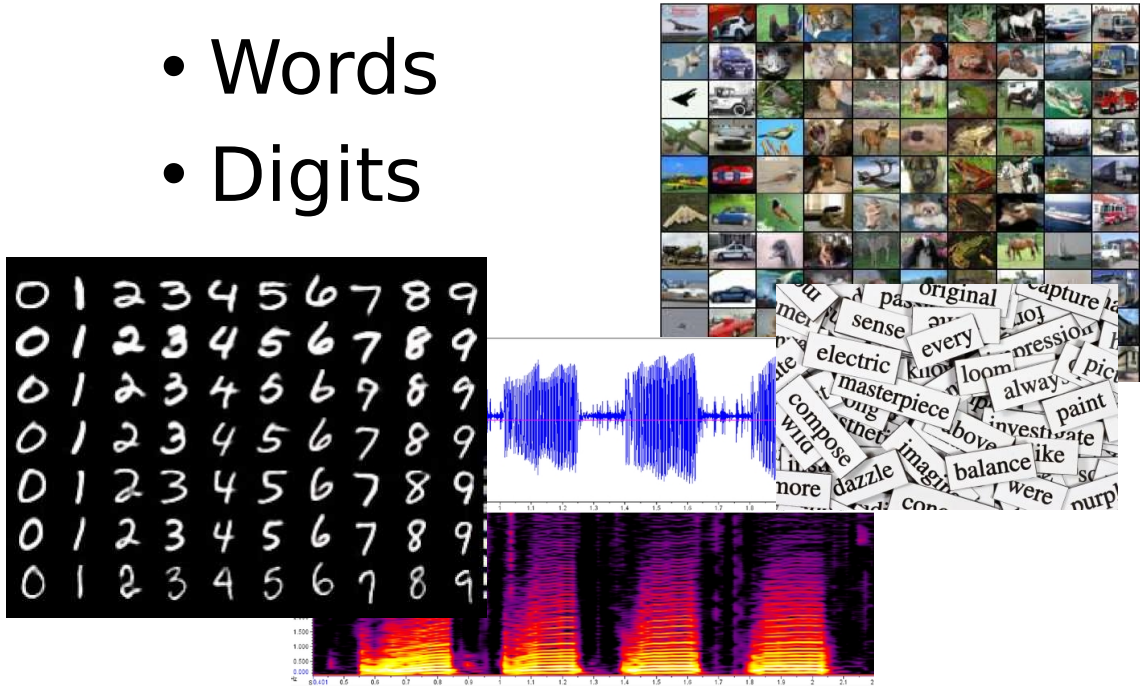


https://christophm.github.io/interpretable-ml-book/intro.html

Interpretable *versus* non-interpretable data

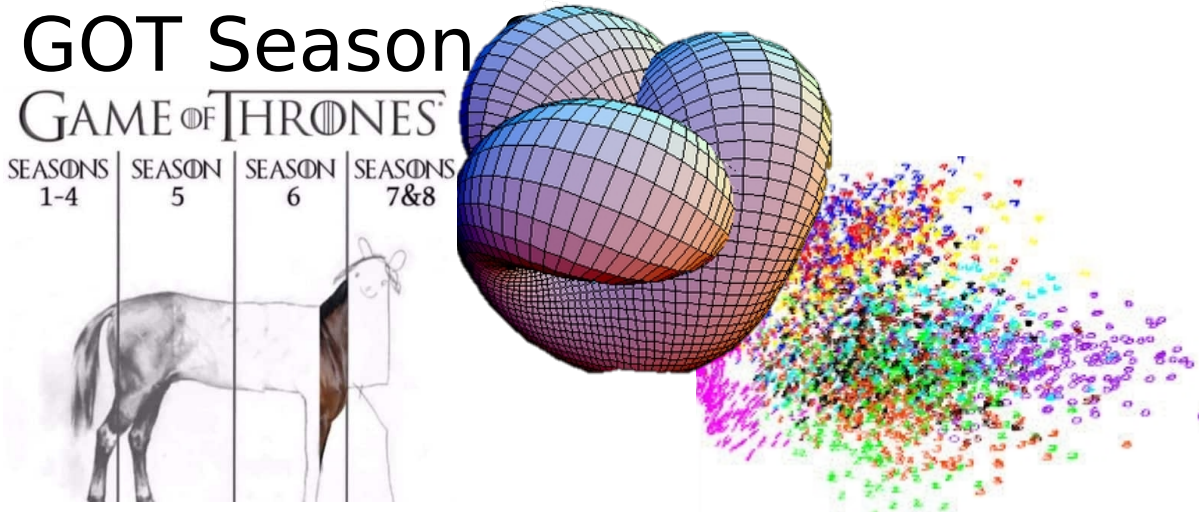
Interpretable

- Natural Images
- Sounds
- Words
- Digits



Non-interpretable

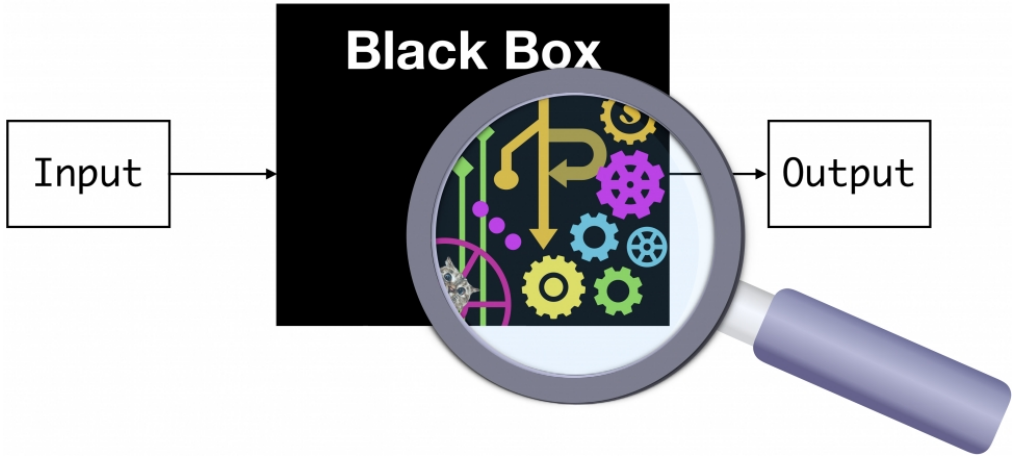
- Nonparametric multimodal distributions
- High dimensional manifolds
- GOT Season



Black box *versus* interpretable models

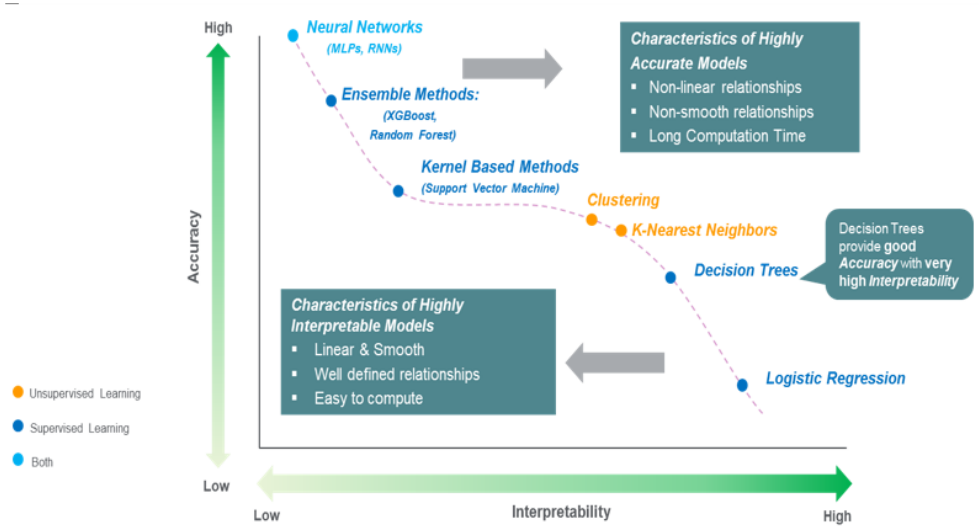
Black box models

- Artificial Neural Networks
- Restricted Boltzmann Machines
- Complex Model



Interpretable models

- Linear Regression
- Logistic Regression
- Support Vector Machines

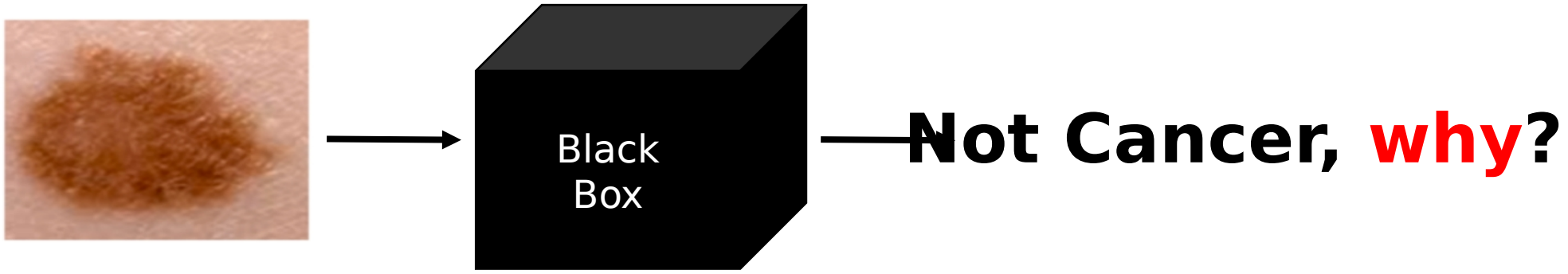


https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

Interpretable Machine Learning

Why pursue interpretable models?

- In high risk areas, **such as medicine**, it is of outmost importance to understand the rationale of predictive models, otherwise it could lead to life threatening situations.
- Knowing the **“why” of given prediction** can enable us to understand more of the problem in hands.

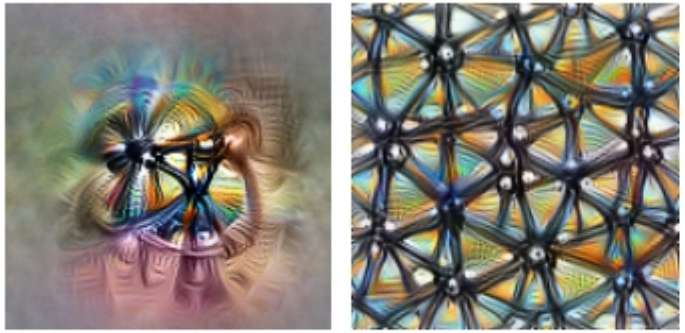


https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

Interpretable *versus* Explainable models

Interpretability

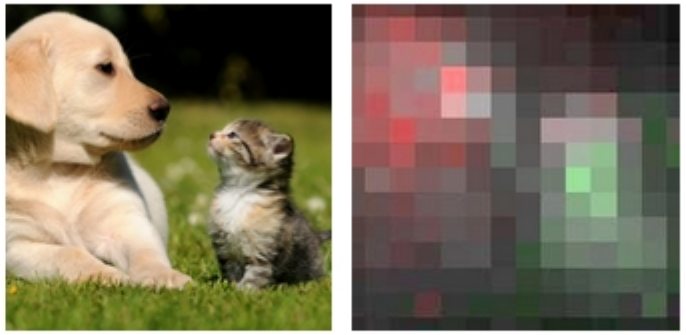
- Ability to describe learned features



Feature visualization answers questions about what a network — or parts of a network — are looking for by generating examples.

Explainability

- Ability to explain a given prediction



Attribution [1] studies what part of an example is responsible for the network activating a particular way.

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

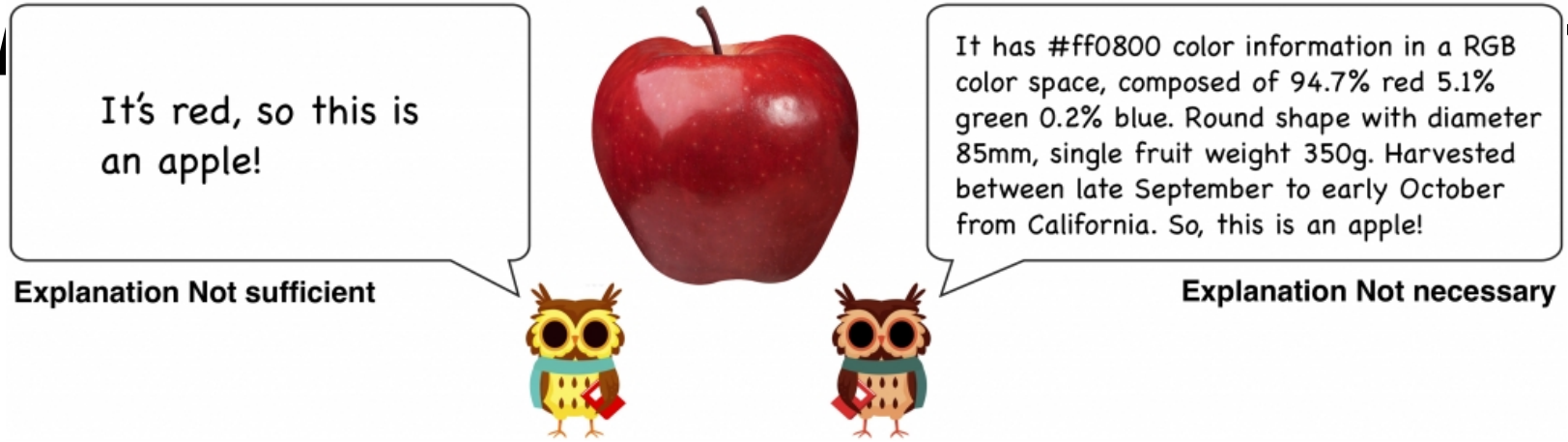
Interpretable Machine Learning

Questions to think about when developing interpretable ML:

1- How much explanation do we need?

2- Do I have to change my model to enable interpretability?

3- W ?



https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

Model-agnostic *versus* model-specific methods

Model Agnostic

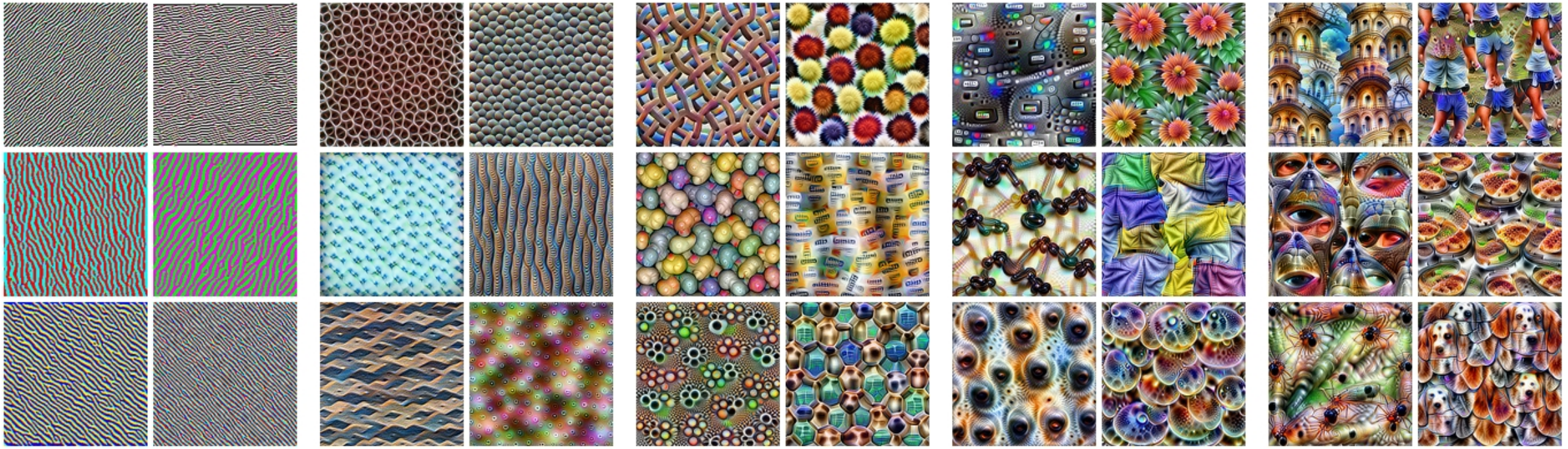
- Can be used independent of the model's structure
- Thus, **no performance on impact**
- Examples:
  - Grad-CAM,
  - Guided Backpropagation,
  - Feature Visualization
  - Activation Maximization
  - LIME
  - Layer-wise Relevance Propagation

Model Dependant

- The model itself explains its features and embeddings
- Baseline model needs to be modified, **possibly impacting performance**
- Examples:
  - Class Activation Mapping (CAM)
  - Autoencoder-based techniques
  - Explainer model-based techniques
  - Uncertainty prediction

Activation-based methods

- **Feature visualization** techniques allow us to see the specialization of each neuron or nodes in a model



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

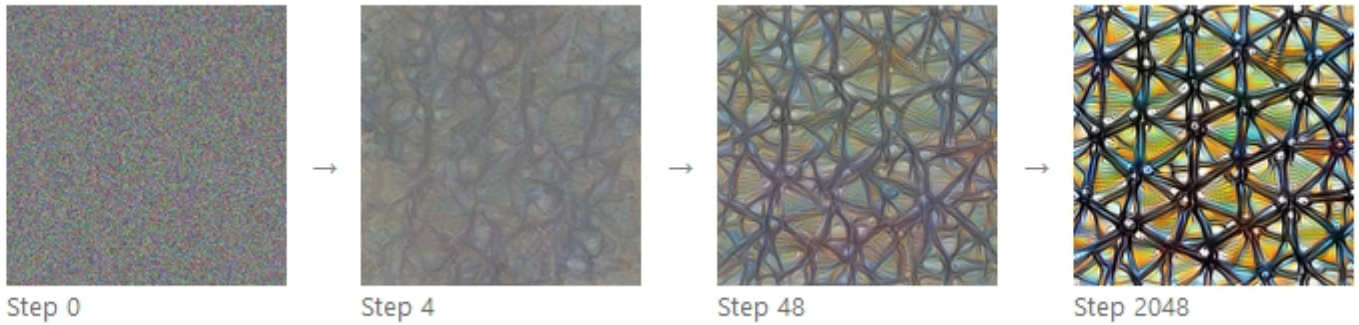
Activation-based methods

- Consist of **maximizing an activation** of a trained model
- Use **optimization algorithms** such as gradient descent
- [illegible] ent [illegible]

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

Activation-based methods

- Example: maximizing a certain neuron activation



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

Activation-based methods

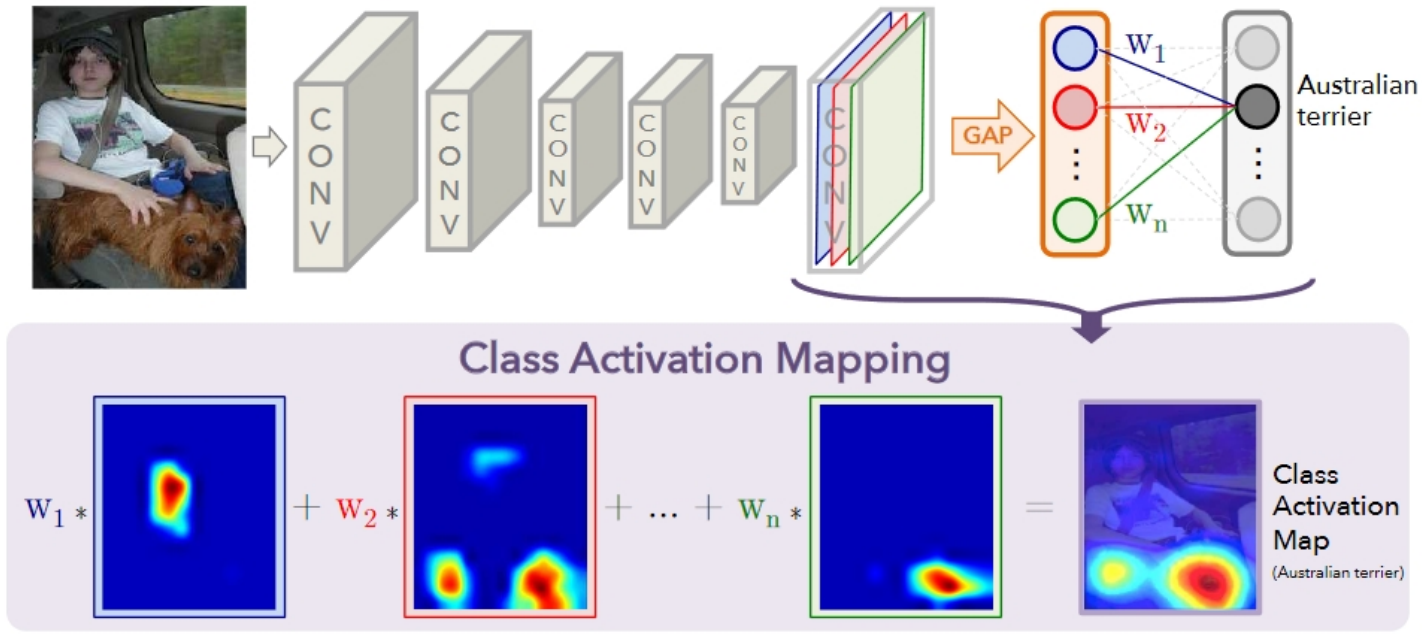
- By maximizing more than one neuron, it is possible to visualize how they interact

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

Activation-based methods

Class Activation Mapping (CAM)

- Algorithm:
  1. Feed input and store feature maps from last layer
  2. Apply global average pooling to find weights $Fk = \sum_{x,y} f_k(x, y)$
  3. Compute softmax for a class c with Sc $\sum_k w_k^c F_k$ w w_k^ce are learnable



Images: https://arxiv.org/abs/1512.04150

Gradient-based methods

Sensitivity Analysis

"Measures the relative importance of input features by calculating the gradient of the output decision with respect to those input features." *

- Algorithm:
  1. Forward-pass input
  2. Compute gradients per
  3. Backpropagate

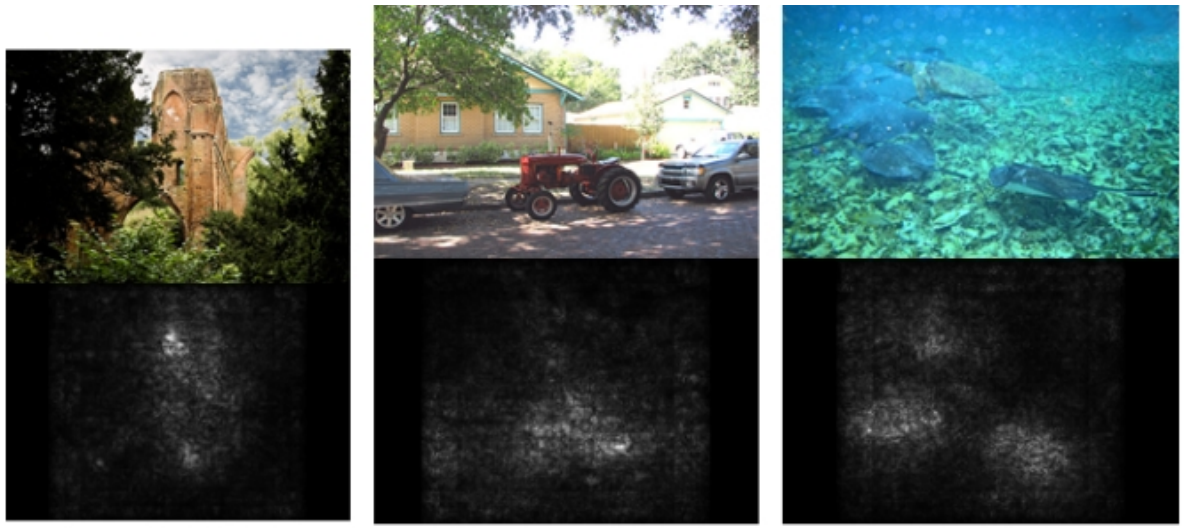


Figure 2: **Image-specific class saliency maps for the top-1 predicted class in ILSVRC-2013 test images.** The maps were extracted using a single back-propagation pass through a classification ConvNet. No additional annotation (except for the image labels) was used in training.

Images: https://arxiv.org/abs/1312.6034

* http://blog.qure.ai/notes/deep-learning-visualization-gradient-based-methods

Gradient-based methods

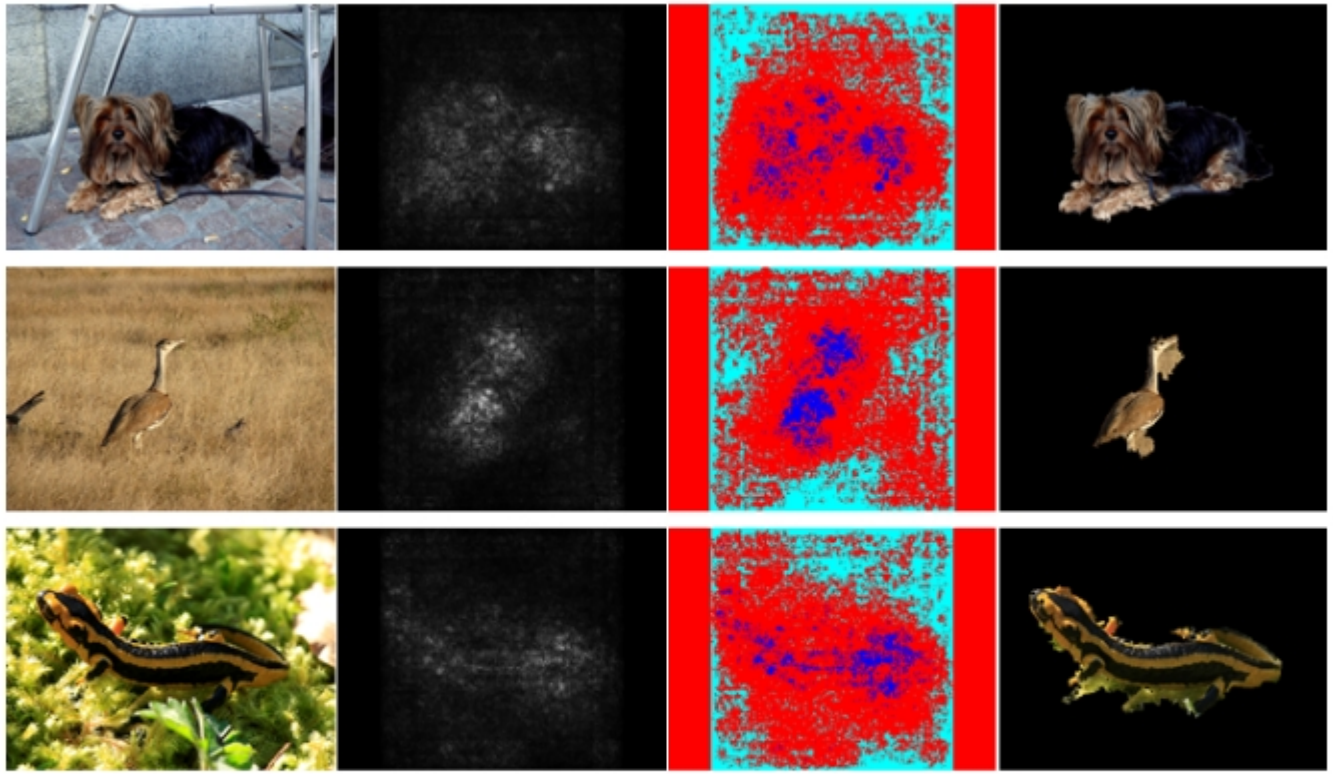


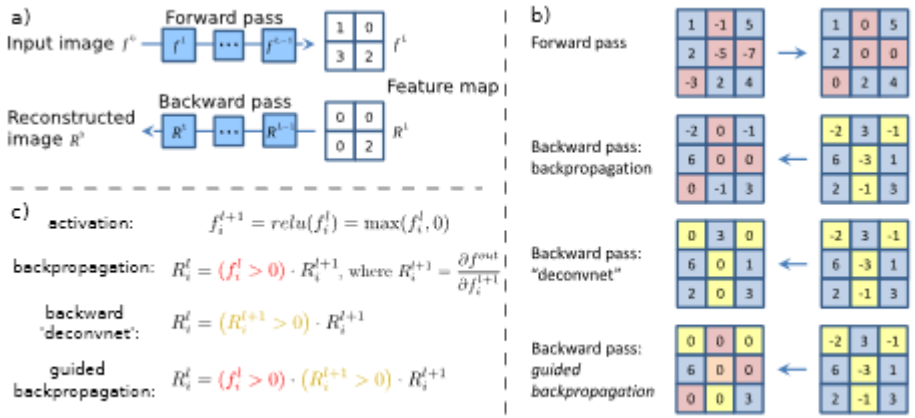
Figure 3: **Weakly supervised object segmentation using ConvNets (Sect. 3.2).** *Left:* images from the test set of ILSVRC-2013. *Left-middle:* the corresponding saliency maps for the top-1 predicted class. *Right-middle:* thresholded saliency maps: blue shows the areas used to compute the foreground colour model, cyan – background colour model, pixels shown in red are not used for colour model estimation. *Right:* the resulting foreground segmentation masks.

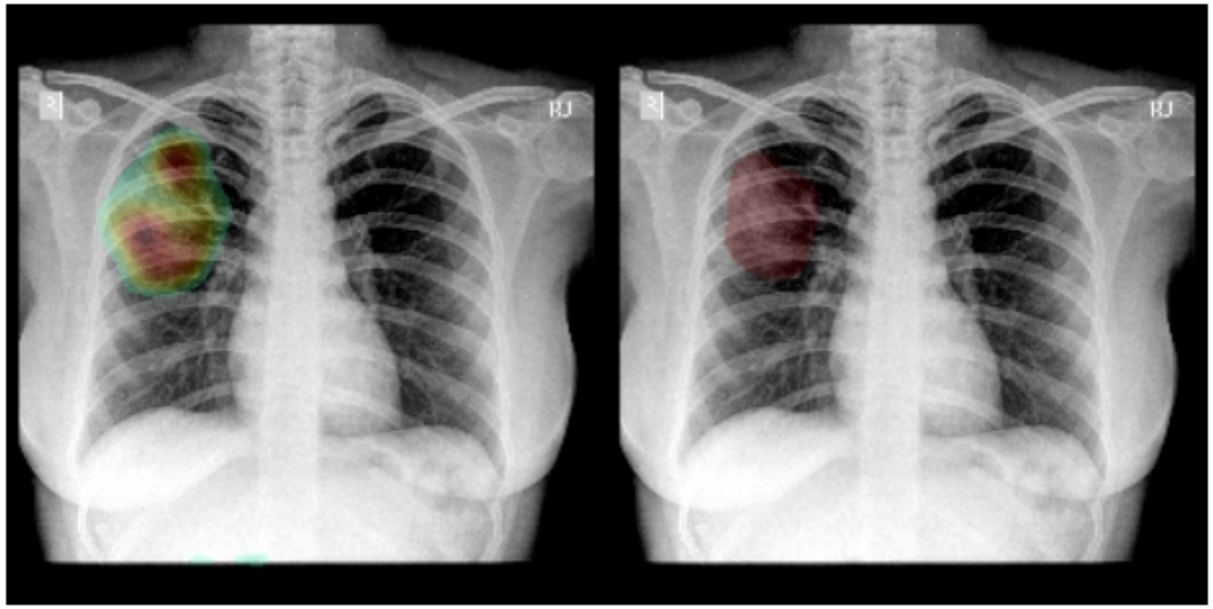
Images: https://arxiv.org/abs/1312.6034

Gradient-based meth

Guided Backprop

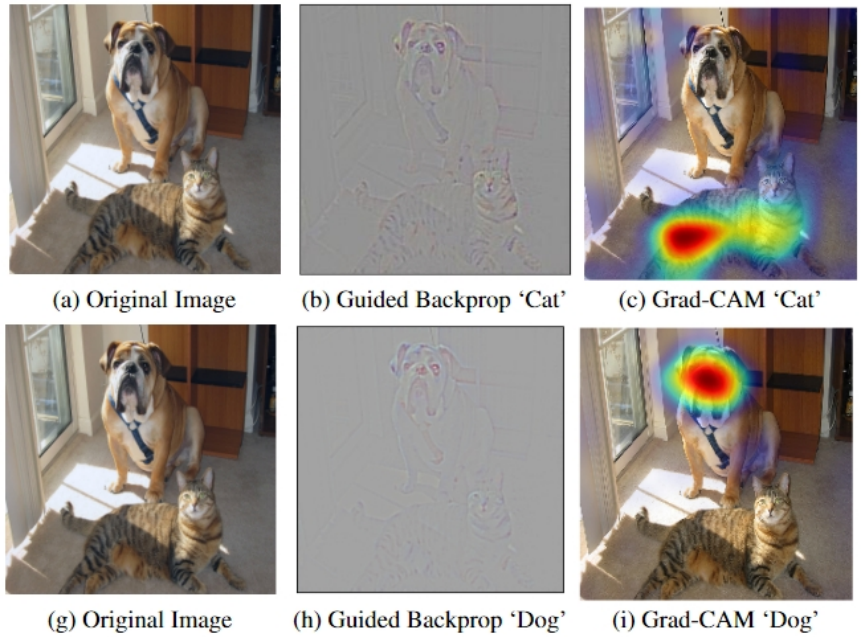
- Algorithm:
  1. Forward-pass input
  2. Zero-out all activations bu one
  3. Compute gradients per lay
  4. Zero-out negative gradien
  5. Backpropagate





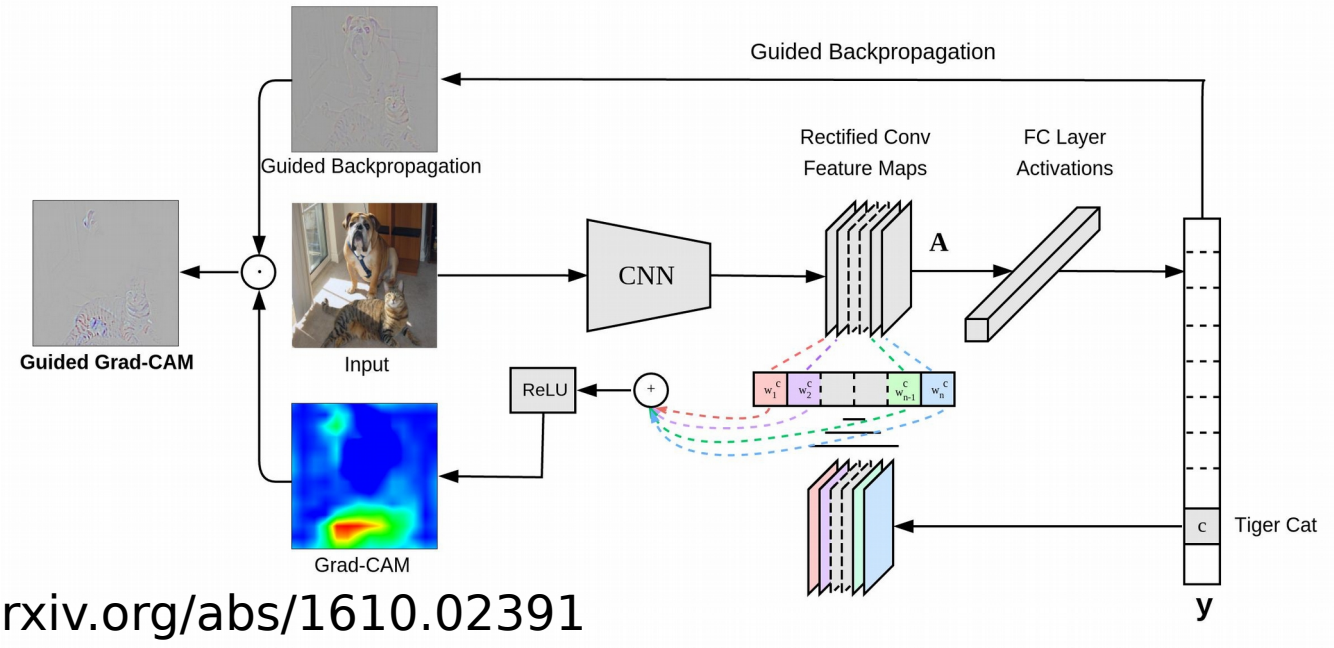
Images: https://arxiv.org/abs/1412.6806

Gradient-based methc



Gradient CAM (Grad-CAM)

- Algorithm:
  1. Compute gradient for class c with respect to each feature map
  2. Calculate global average of gradients for each feature
  3. $$L^c_{\text{Grad-CAM}} = ReLU\left(\sum_k \alpha^c_k A^k\right)$$



Images: https://arxiv.org/abs/1610.02391

Gradient-based methods

Main disadvantages:

1. **Different methods may yield different saliency maps**
2. Exact meaning of highlighted features is unexplained
3. Maps are corrupted with noise and artifacts due to architecture and other unknown phenomena

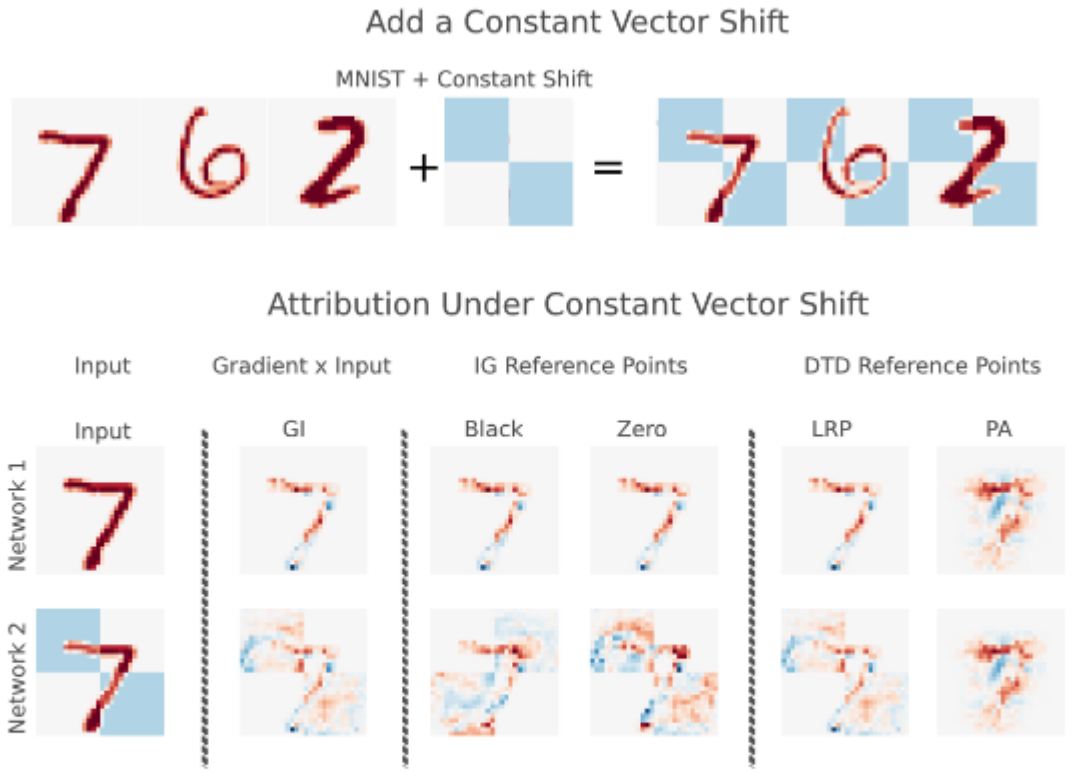


Figure 4: Evaluation of attribution method sensitivity using MNIST. Gradient x Input, all IG reference points and DTD with a LRP reference point do not satisfy input invariance and produce different attributions for each network. DTD with a PA reference point is not sensitive to the transformation of the input.

Images: https://arxiv.org/abs/1711.00867

Gradient-based methods

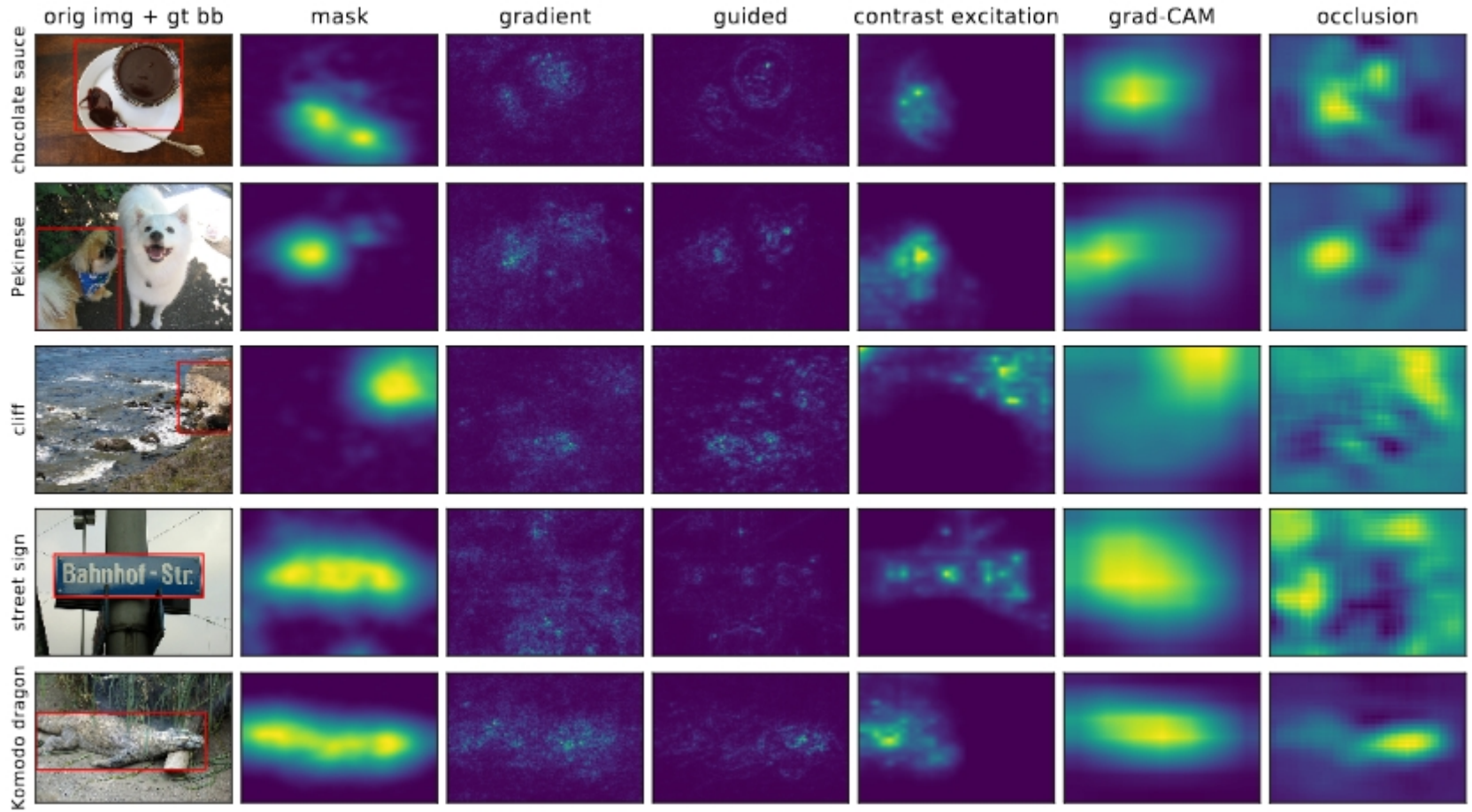
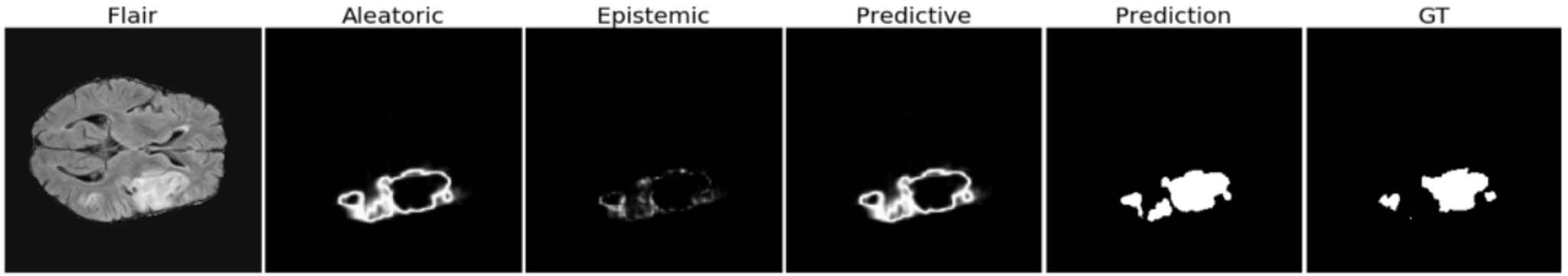


Image: https://arxiv.org/abs/1704.03296

Miscellaneous: Uncertainty Prediction

- Existing DL models do not provide confidence estimates, which leads to an inability to identify incorrect outputs or out-of-distribution inputs.
- These uncertainties can be classified into two types
  1. ***data dependent (aleatoric)*** caused by incompleteness in input data (e.g. noise, absence of visual features) -> **Loss = $\mathcal{L}$ + log(σ)**
  2. ***model dependent (epistemic)*** due to incompleteness in the trained model (e.g. type of implant not seen in training) -> **Monte Carlo Dropout**



https://openreview.net/forum?id=Sk_P2Q9sG

Miscellaneous: Explainer Models

DeConvNet

- Algorithm:
  1. Attach a deconv layers to conv layers
  2. Force deconv layers to predict feature maps of regular conv layers
  3. DeConvNet will learn to predict inputs corresponding to a class

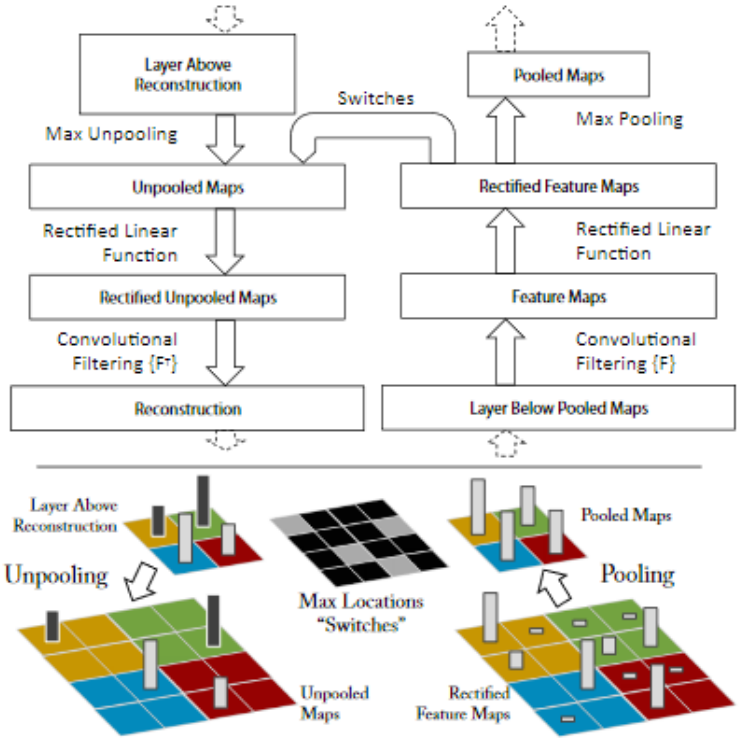
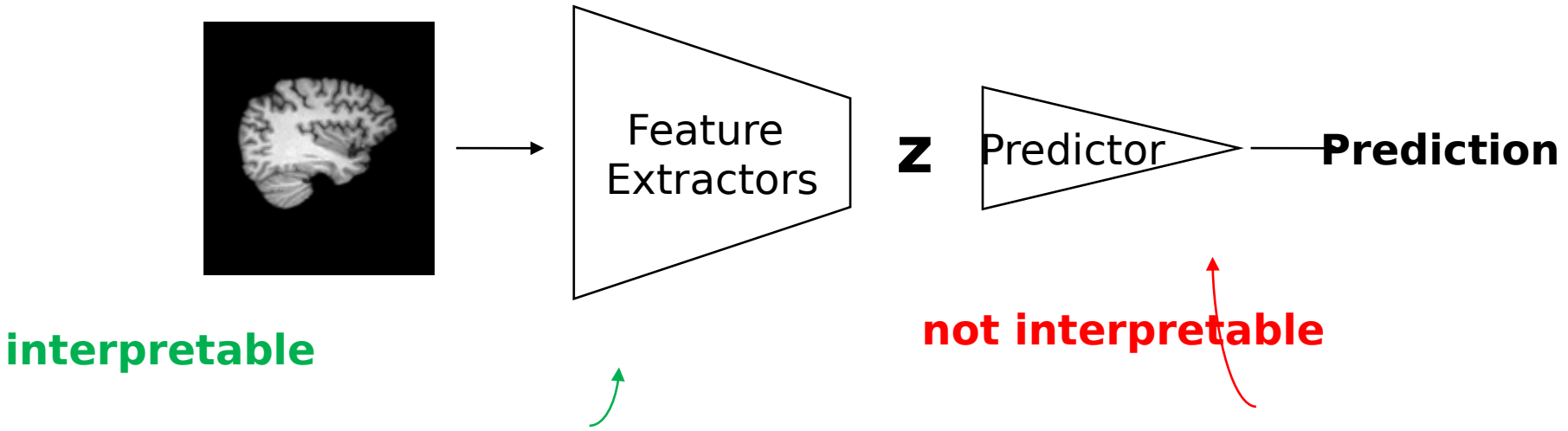


Figure 1. Top: A deconvnet layer (left) attached to a convnet layer (right). The deconvnet will reconstruct an approximate version of the convnet features from the layer beneath. Bottom: An illustration of the unpooling operation in the deconvnet, using *switches* which record the location of the local max in each pooling region (colored zones) during pooling in the convnet.

Images: https://arxiv.org/abs/1311.2901

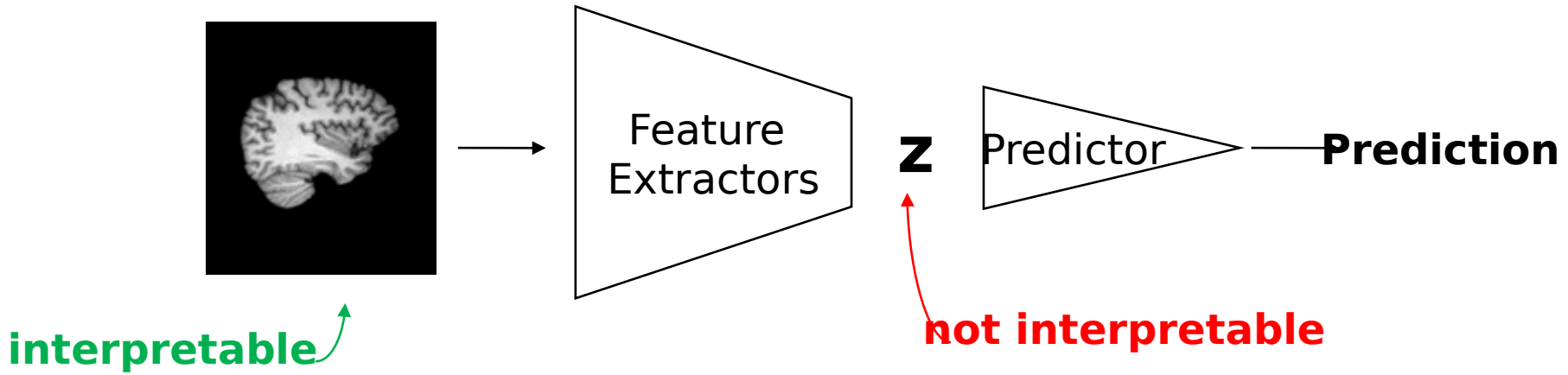
Miscellaneous: Explainer Models

- **Embedding a decoding model to translate from non-interpretable latent to interpretable input space**



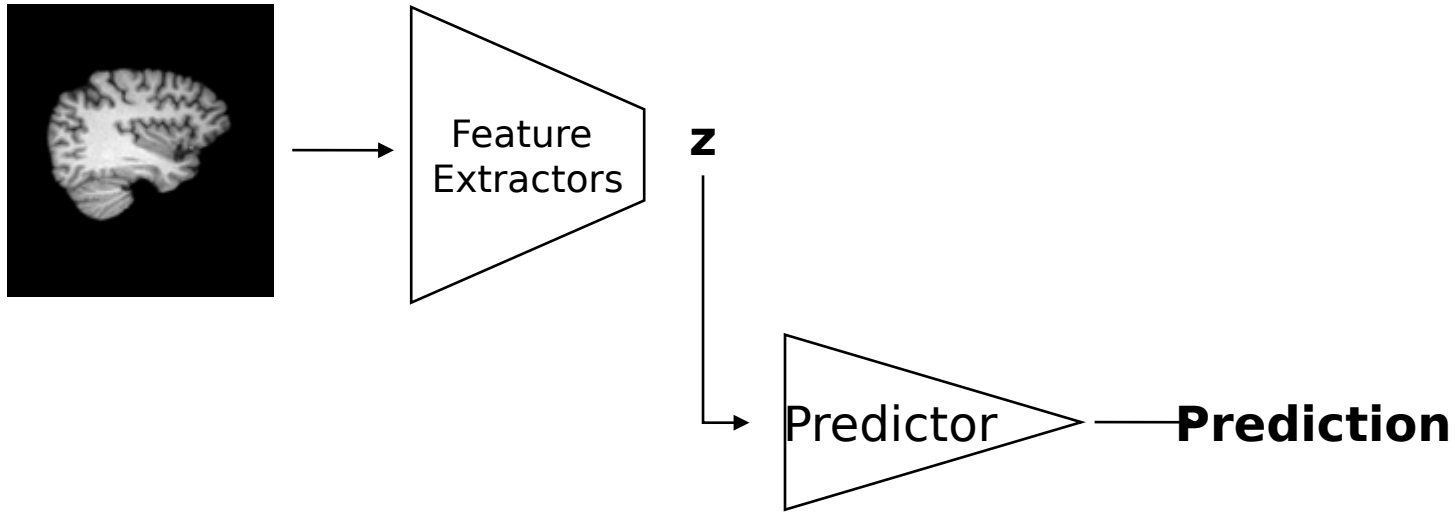
Miscellaneous: Explainer Models

- **Embedding a decoding model to translate from non-interpretable latent to interpretable input space**



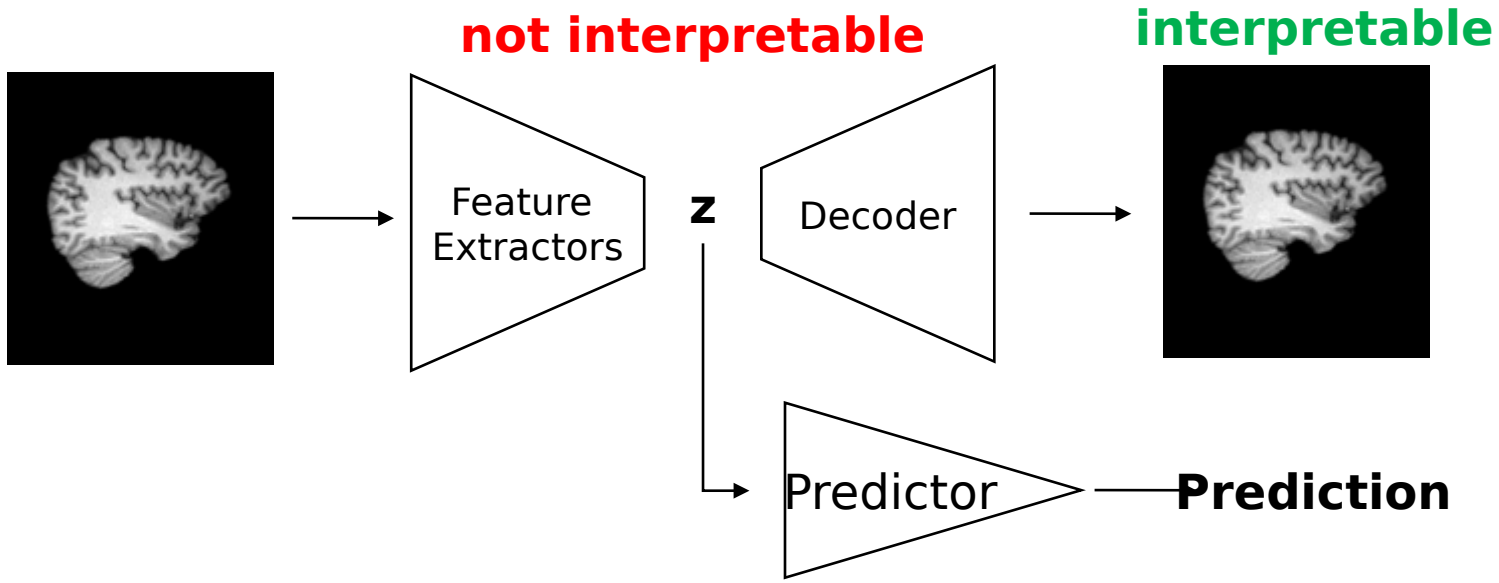
Miscellaneous: Explainer Models

- **Embedding a decoding model to translate from non-interpretable latent to interpretable input space**



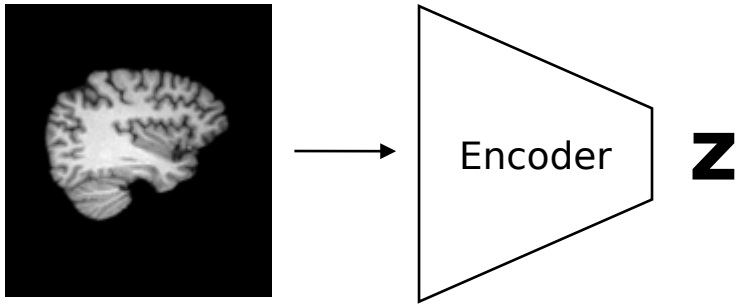
Miscellaneous: Explainer Models

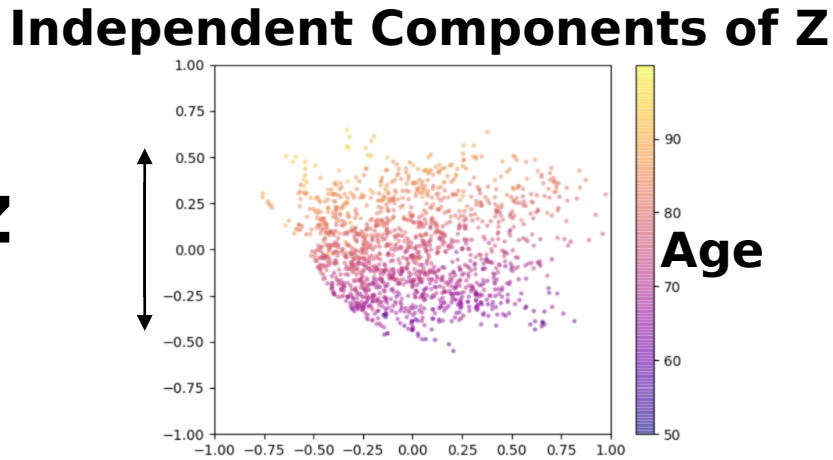
- **Embedding a decoding model to translate from non-interpretable latent to interpretable input space**



Miscellaneous: Explainer Models

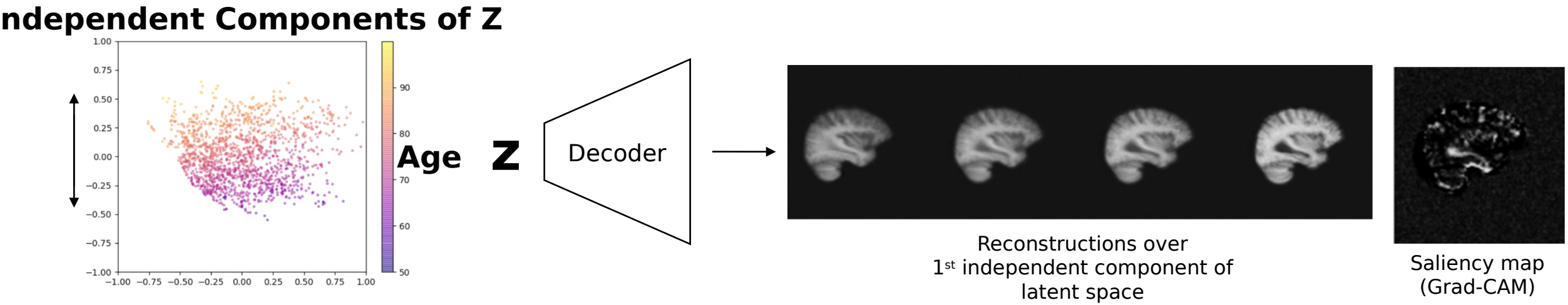
- **Latent space exploration**





Miscellaneous: Explainer Models

- **Latent space exploration**



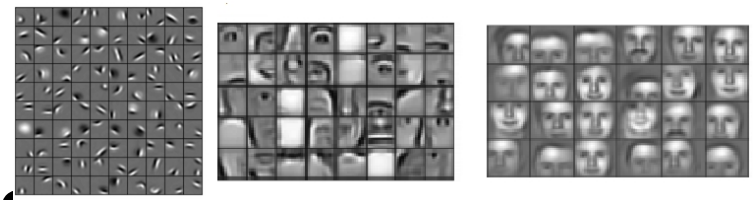
Reconstructions over
1st independent component of
latent space

Saliency map
(Grad-CAM)

Miscellaneous: Others

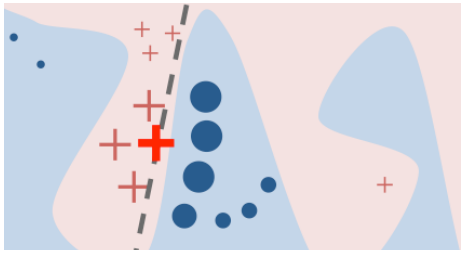
Early feature exploration

- Visualizing raw activations
- Factors of variation analysis



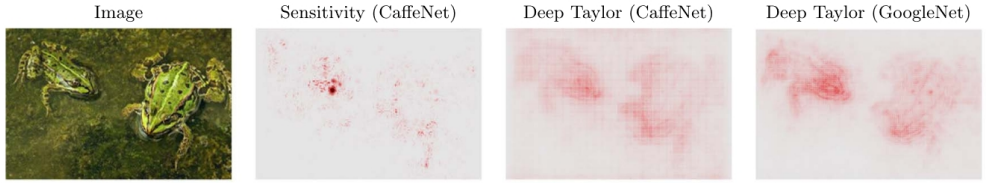
Local interpretable model-agnostic Explanations (LIME)

- Linearizing models for explainability
- Training an interpretable model to copy a non-linear one



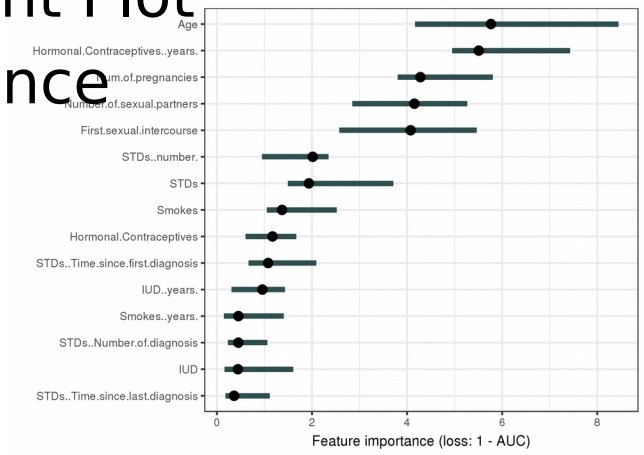
Deep Taylor Decomposition/LRP

- Direct attribution method
- Similar to gradient methods



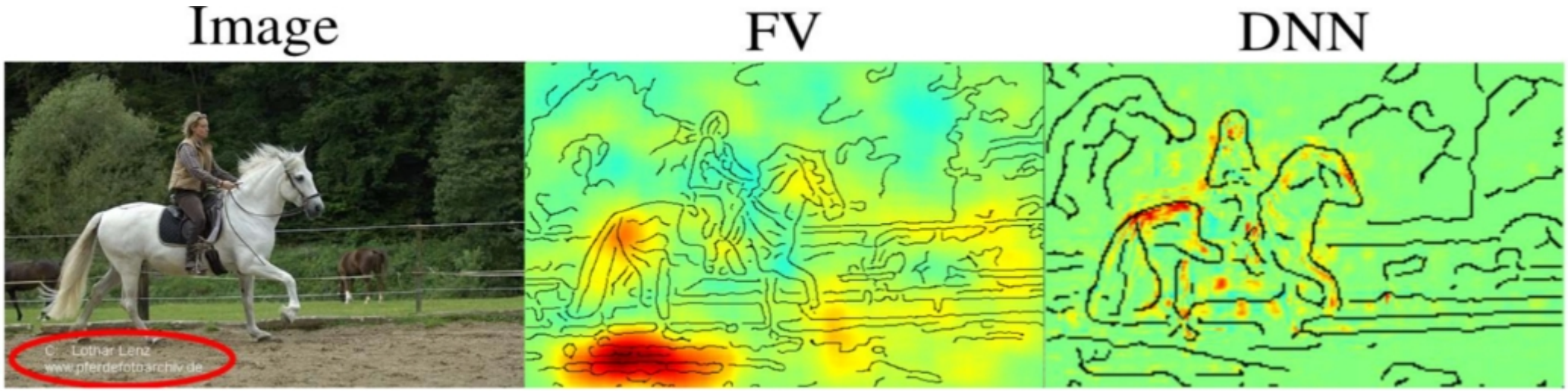
Other non-Deep Learning stuff

- Partial Dependent Plot
- Feature importance
- Shapley Values



Conclusion

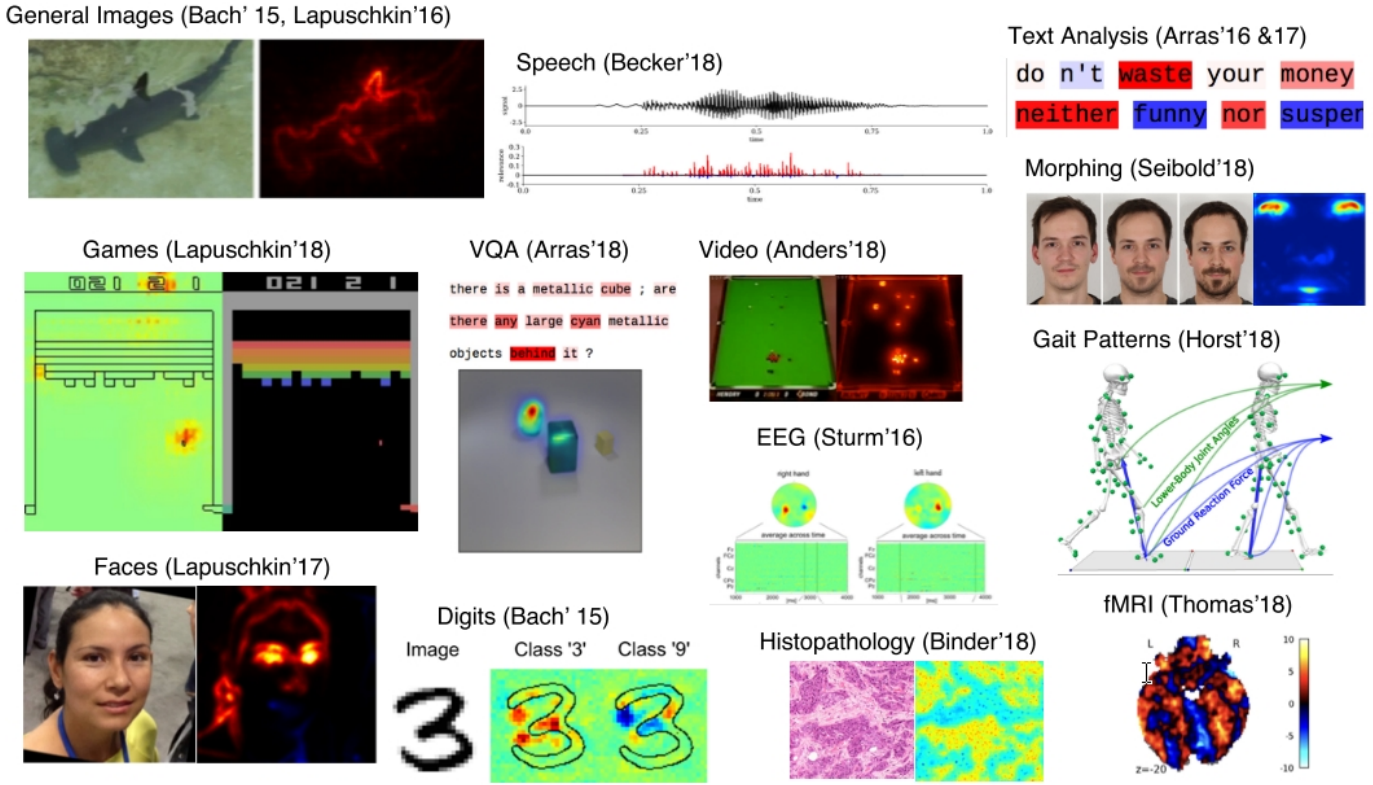
- **Explanations** helps us find flaws in models



Taken from: ICIP’18 Tutorial on Interpretable Deep Learning

Conclusion

- **Explanations** helps us find flaws in models
- **Interpretations** allow us to obtain new scientific insights



Taken from: ICIP'18 Tutorial on Interpretable Deep Learning

Conclusion

- **Explanations** helps us find flaws in models
- **Interpretations** allow us to obtain new scientific insights on data
- **Performance impact** and **practicality** of implementation are important concerns when looking for making a model more interpretable
- **Very early stage** of interpretable deep learning: new algorithms and methods in a fast pace
- Current state-of-the-art is still **not ideal** for providing both **explanations and interpretations** without hurting performance

Thank you! Questions?