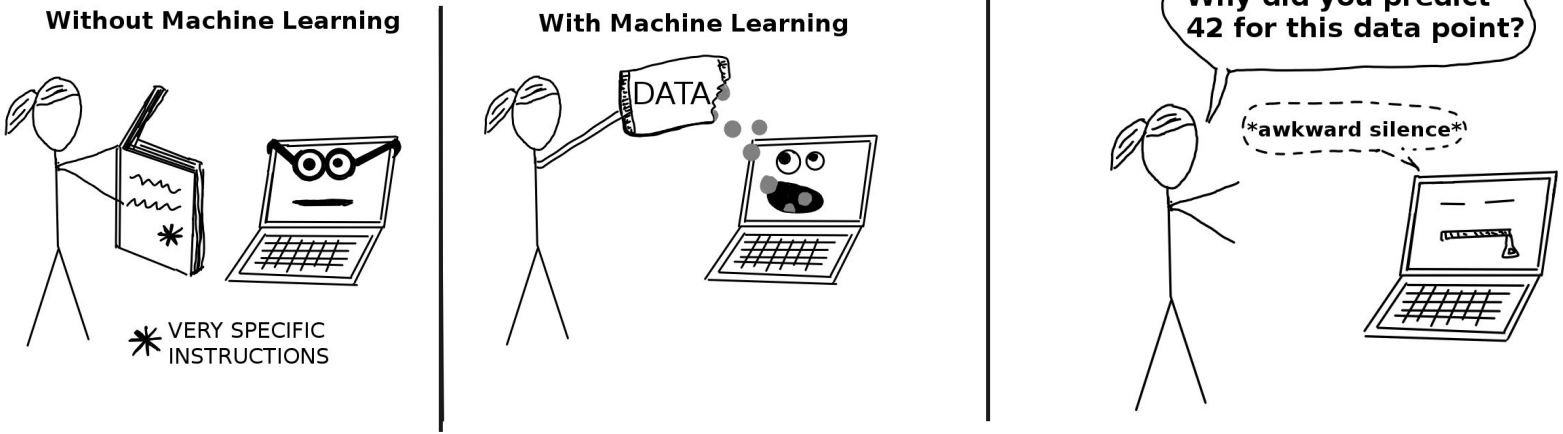
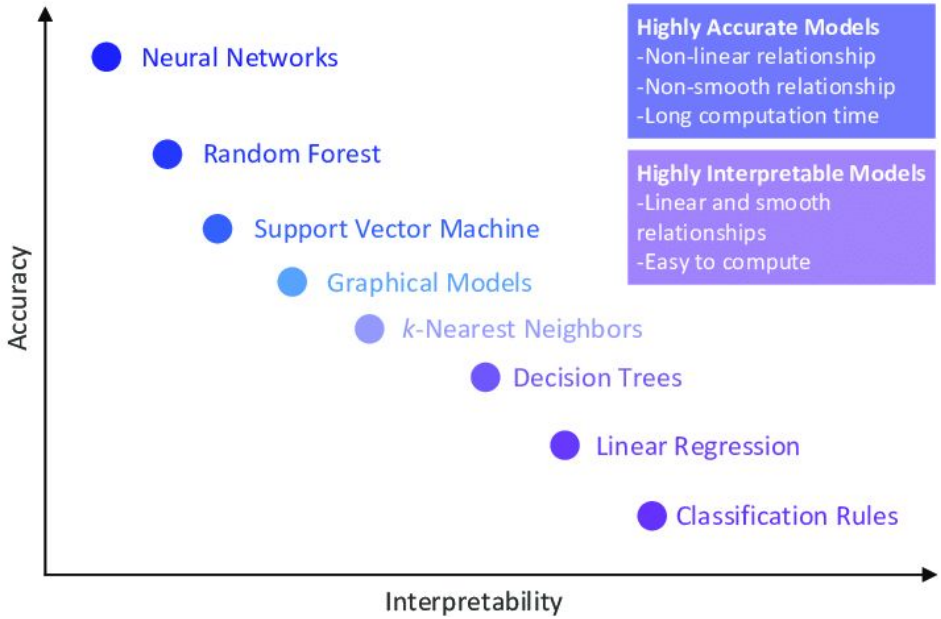
Interpretability in Machine Learning

Interpretability



Interpretability



Interpretability vs Explainability

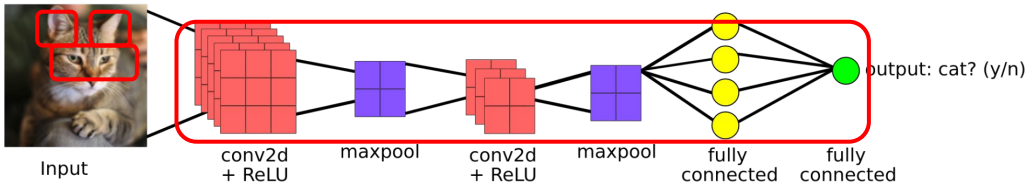
Interpretability

Understand the cause of a model decision.

≠

Explainability:

Understand the internal mechanism leading to a model decision.



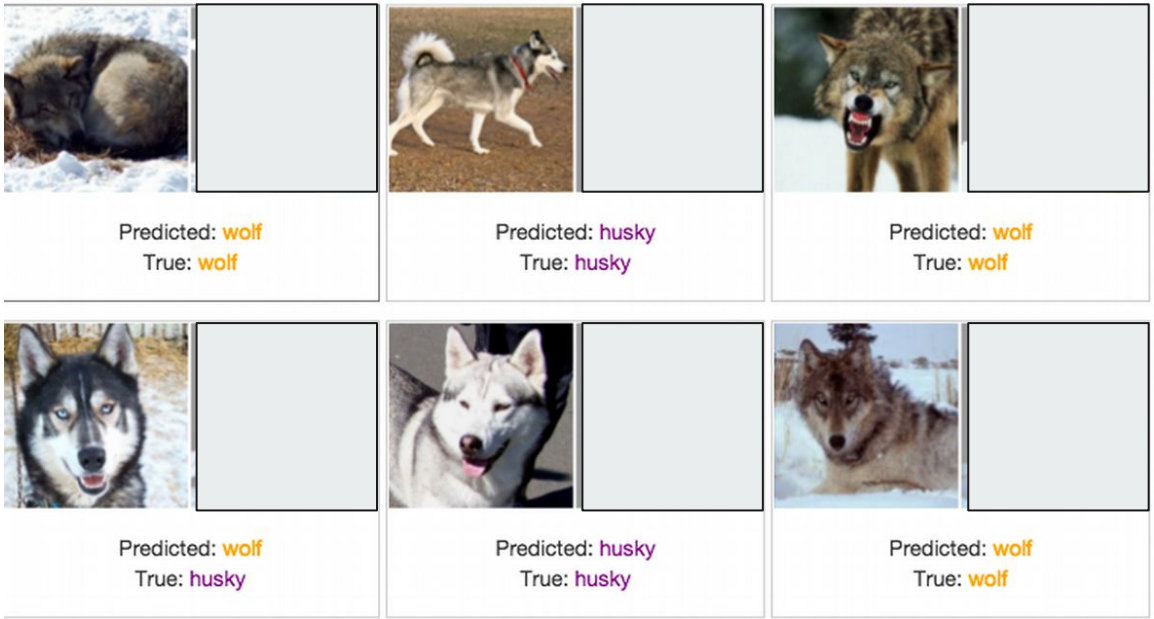
Interpretable Machine Learning:

Methods and models that make the decisions taken by machine learning systems humanly understandable

Why do we want interpretability?

- Human nature
- Acceptance and trust
- Legal issues
- Fairness issues
- Improve human knowledge
- Debugging

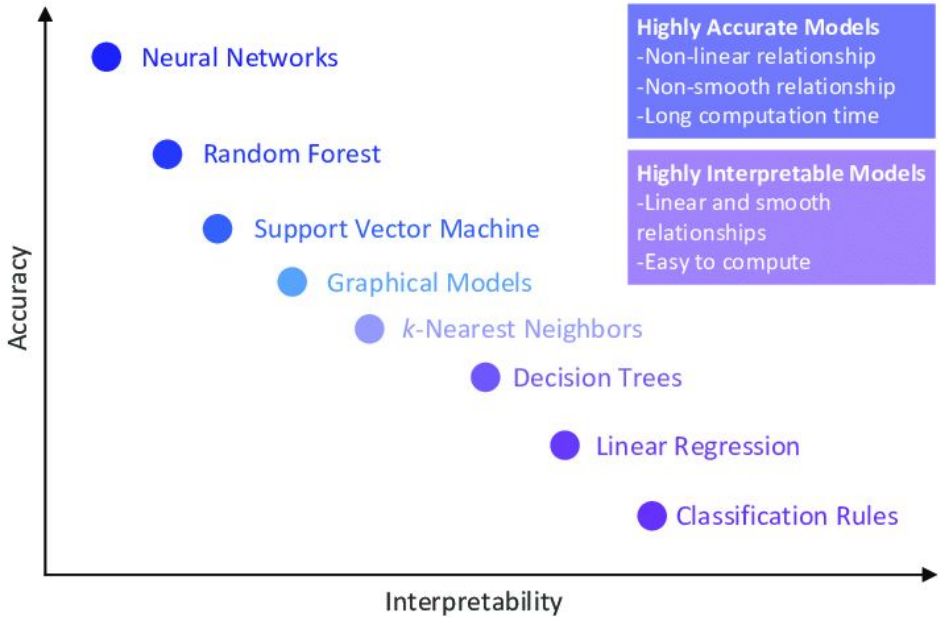
Why do we want interpretability?



Taxonomy of Interpretability Methods

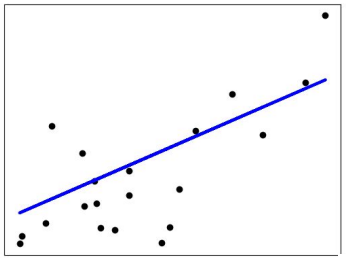
- Intrinsic or Post Hoc

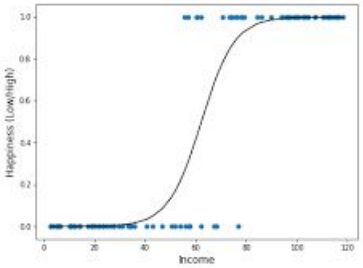
Interpretability

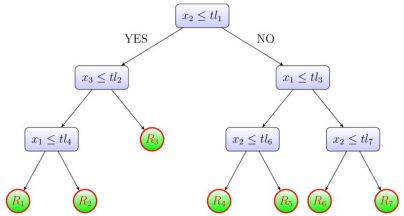


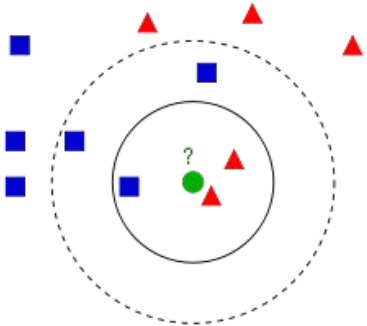
Intrinsic: Interpretable models

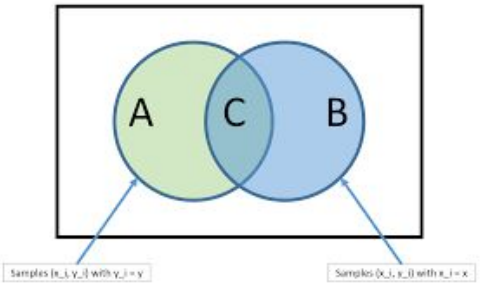
- Linear regression
- Logistic regression
- Decision trees
- Naive Bayes
- K-nearest neighbors



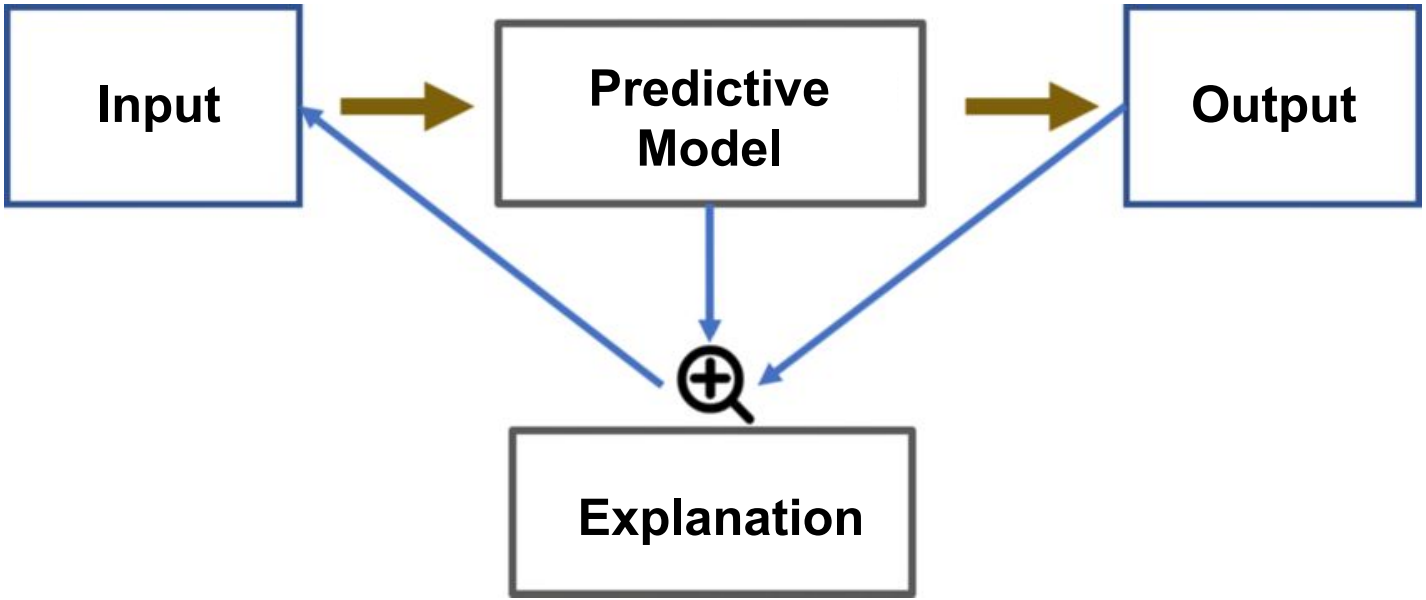








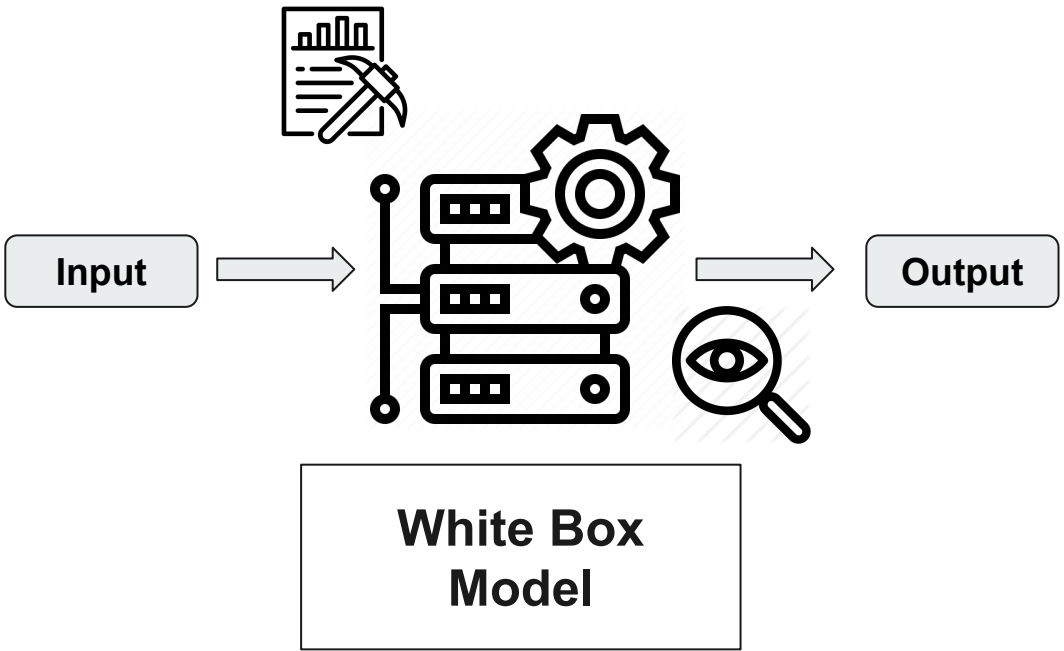
Post Hoc



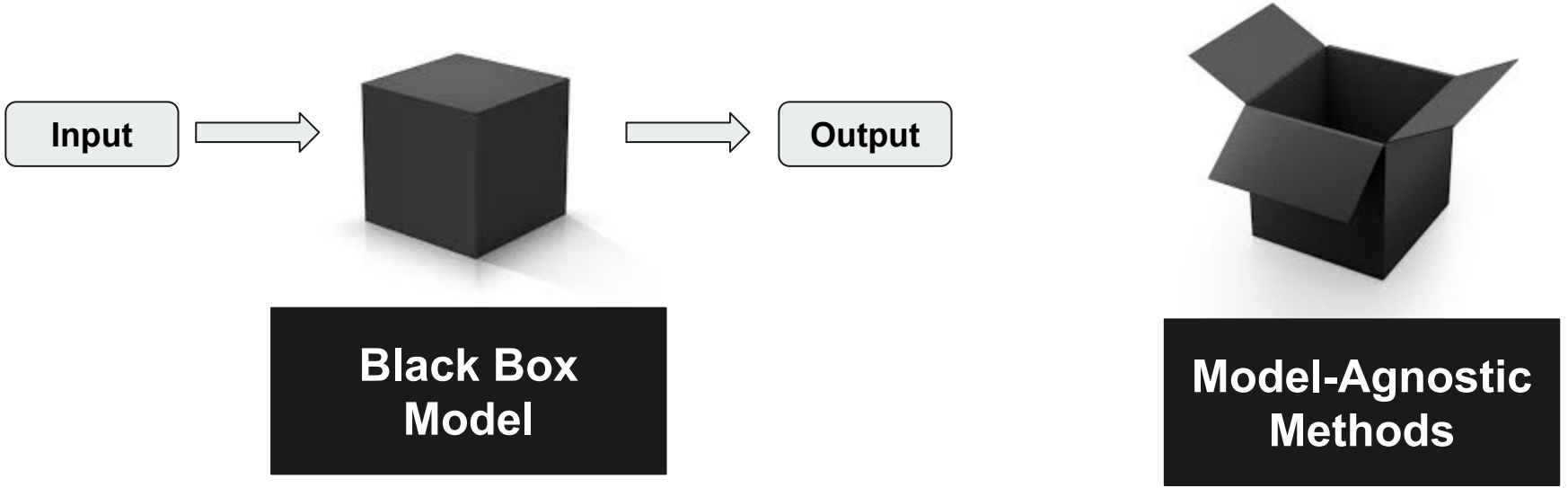
Taxonomy of Interpretability Methods

- Intrinsic or Post Hoc
- Model-specific or Model-agnostic

Model-Specific Methods



Model-Agnostic Methods



Taxonomy of Interpretability Methods

- Intrinsic or Post Hoc
- Model-specific or Model-agnostic
- Local or global

Local vs Global Methods

Global Methods:

- Describe the average behavior
- Discover biases
- Useful for debugging at high level

Local Methods:

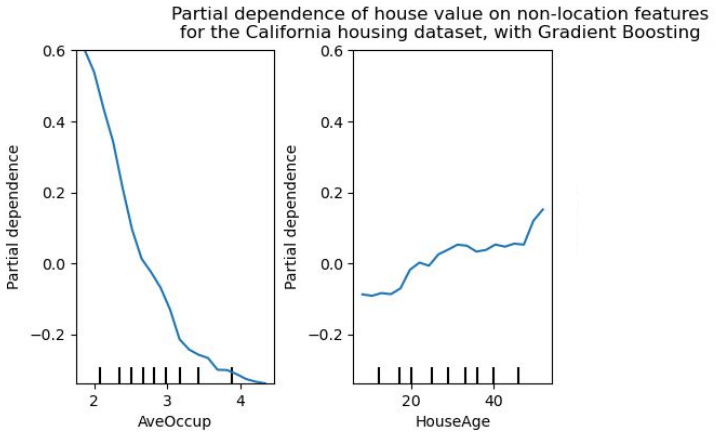
- Individual predictions insights
- Provide explanations to the final user
- Useful for debugging at the instance level

Model-agnostic methods

Partial Dependence Plot (PDP)

- Global Method
- Show the relationship between the target and a specific feature
- Monte Carlo method:

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_S, x_C^{(i)}\right)$$



Greedy function approximation: A gradient boosting machine.

Jerome H. Friedman

PDP-based Feature Importance

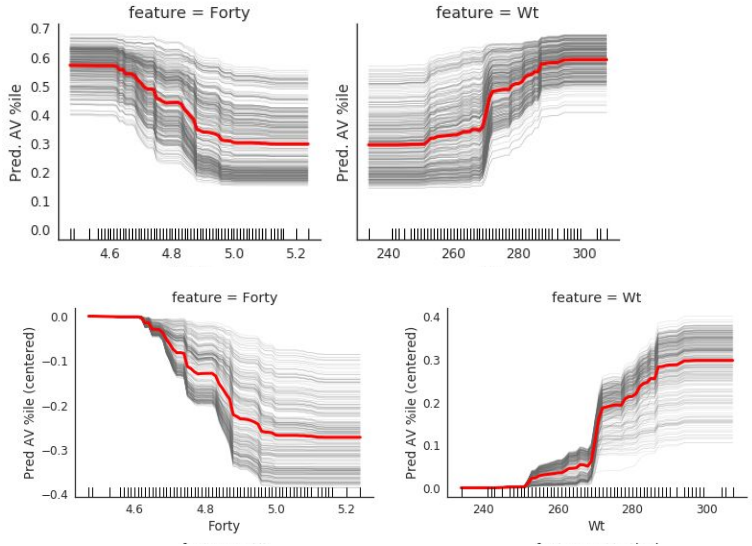
$$I\left(x_S\right) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{f}_S\left(x_S^{(k)}\right) - \frac{1}{K}\sum_{k=1}^{K}\hat{f}_S\left(x_S^{(k)}\right)\right)^2}$$

A Simple and Effective Model-Based Variable Importance Measure

Brandon M. Greenwell, Bradley C. Boehmke, Andrew J. McCarthy

Individual Conditional Expectation (ICE)

- Local Method
- Show the relationship between the target and a specific feature AND instance
- Centered ICE plots



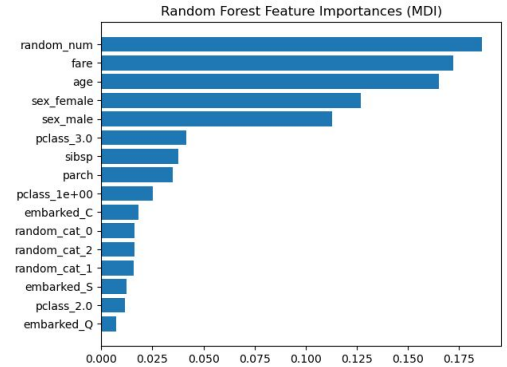
Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation

Alex Goldstein, Adam Kapelner, Justin Bleich, Emil Pitkin

Permutation Feature Importance

- How much permuting a feature increases the prediction error?
- If error increases -> important feature
- If error decreases -> unimportant feature

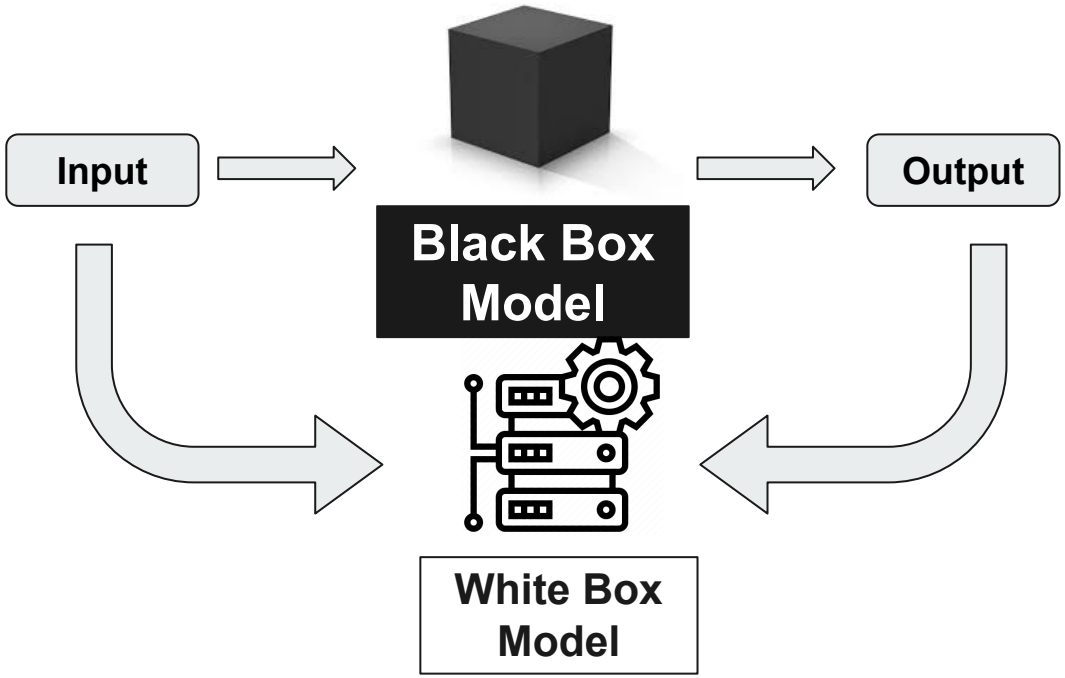
X_A	X_B	X_C	Y
xa1	xb1	xc1	y1
xa2	xb2	xc2	y2
xa3	xb3	xc3	y3
xa4	xb4	xc4	y4
xa5	xb5	xc5	y5
xa6	xb6	xc6	y6



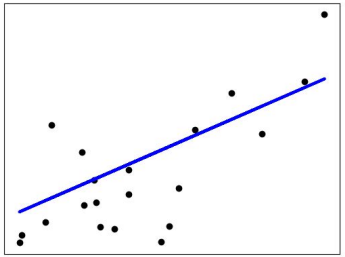
Random Forests (2001)

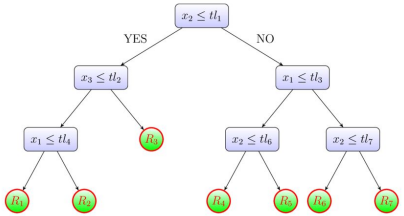
Leo Breiman

Model Surrogates



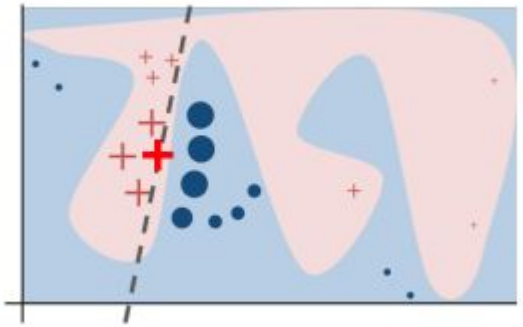
- Global surrogate
- Local surrogate





Local interpretable model-agnostic explanations (LIME)

Training local surrogate to explain individual predictions



Principle:

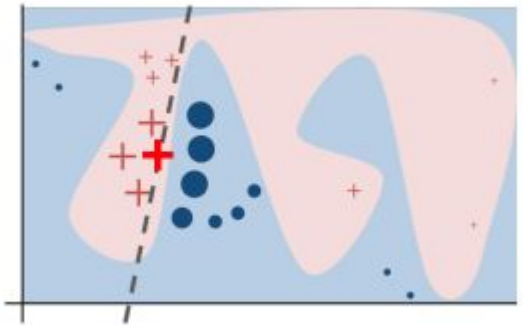
- Choose an instance
- Create local perturbations of the instance -> X
- Get the predictions of the black-box model on X -> Y
- Train a surrogate model on the (X,Y) dataset
  (Using a weighted objective function according to the proximity with the original instance)

"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Local interpretable model-agnostic explanations (LIME)

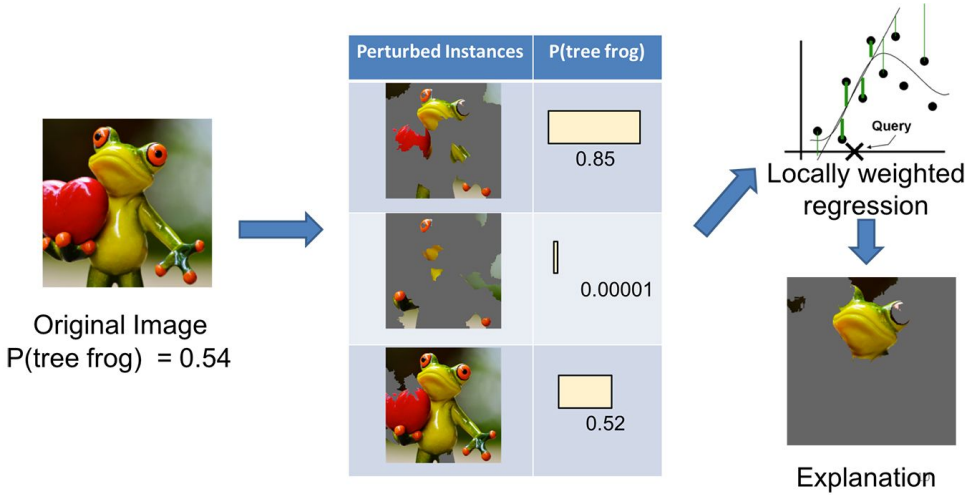
- How to generate perturbations?
- How to compute the distance to the original instance?
- Use several surrogates
- Fidelity measure
- Instability of the explanations



"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

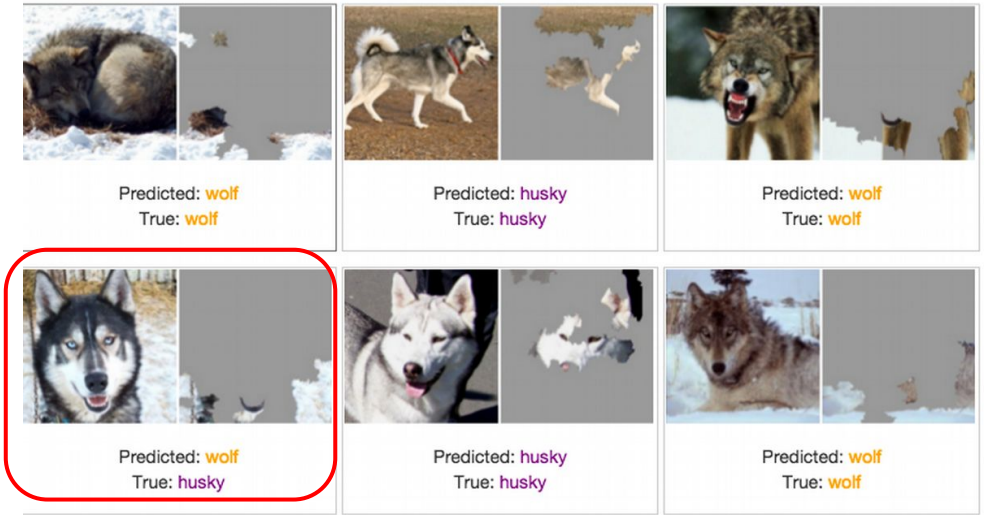
Local Surrogate (LIME)



"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

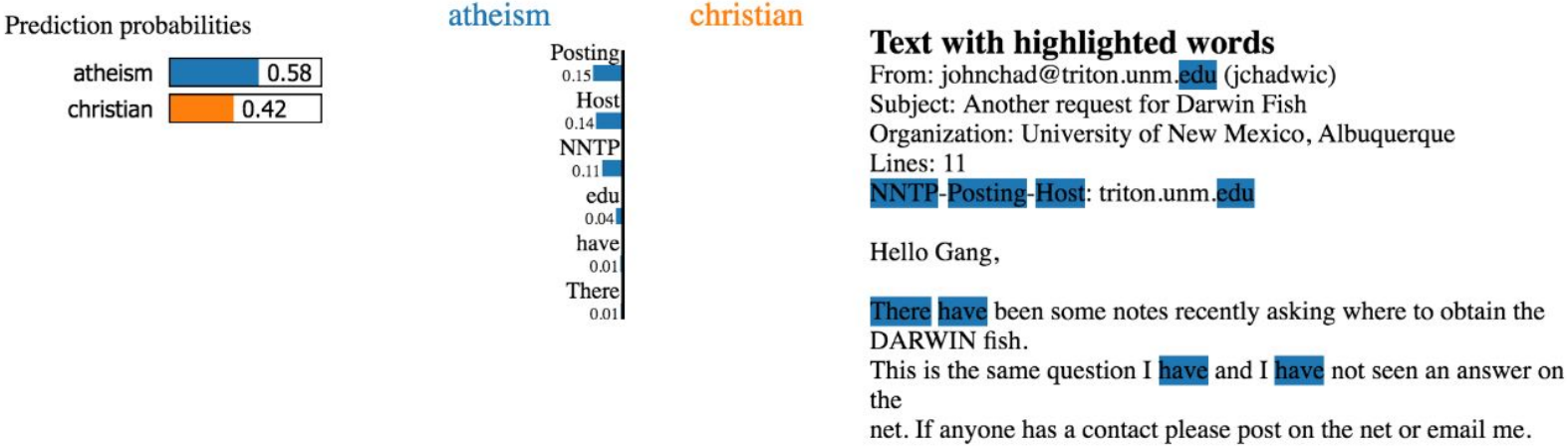
Local Surrogate (LIME)



"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

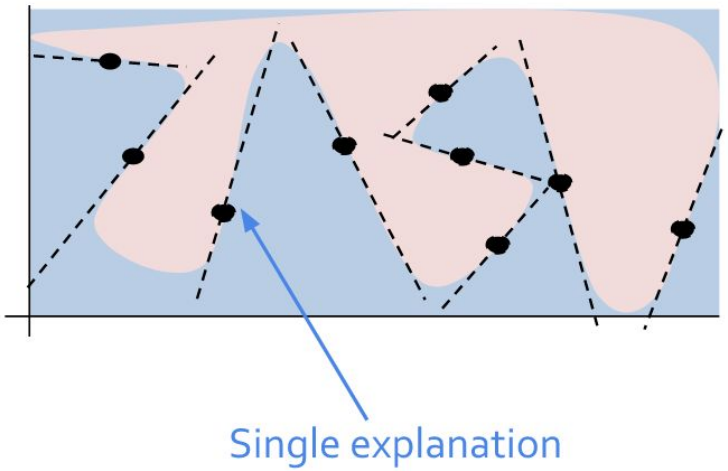
Local Surrogate (LIME)



"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

SP-LIME



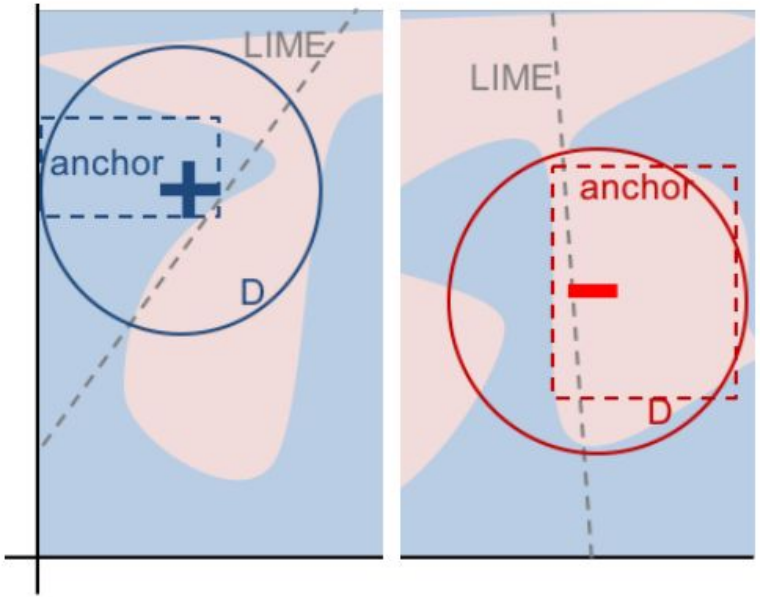
"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Scoped Rules (Anchors)

- Local explanation
- Set of IF-THEN rules (Anchors)
- Precision
- Coverage
- Multi-armed bandit KL-LUCB algorithm

IF HEIGHT < 60cm
AND WIDTH < 70cm
THEN PREDICT Class = Cat
WITH PRECISION 97% AND COVERAGE 15%



Anchors: High-Precision Model-Agnostic Explanations (2018)

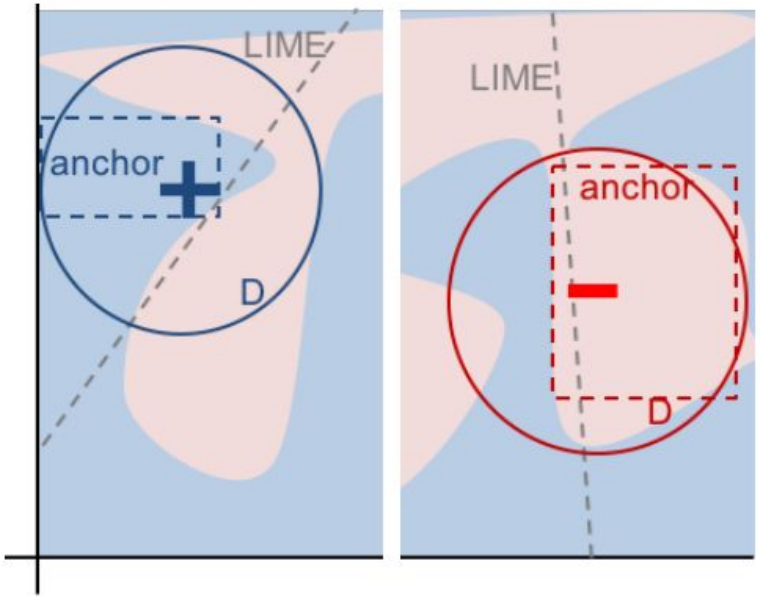
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Scoped Rules (Anchors)

- Local explanation
- Set of IF-THEN rules (Anchors)
- Precision
- Coverage

IF HEIGHT < 60cm
AND WIDTH < 70cm
THEN PREDICT Class = Cat

With Precision=95%
And Coverage=35%



Anchors: High-Precision Model-Agnostic Explanations (2018)

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

SHAP (SHapley Additive exPlanations)

Shapley values:

- Game theory
- Collaborative multi-player game
- Fair repartition of a collaborative game winnings
- Measure individual contributions

A Unified Approach to Interpreting Model Predictions (2017)

SM Lundberg, SI Lee

SHAP (SHapley Additive exPlanations)

- Features = Players
- Winnings = Model's predictions
- Coalitions
- Local instance to explain
- Local perturbations of coalitions
- Estimate 'winnings' with a linear model
- Proximity based on SHAP kernel:

$$\pi_x\left(x'\right) = \frac{(M-1)}{\binom{M}{|x'|}\,|x'|\,(M-|x'|)}$$

X	X_1=7.9	X_2=10.8	...	X_M=1.5
Coalition	1	1	...	1

X'	X_1=7.9	X_2=9.9	...	X_M=2.4
Coalition	1	0	...	0

A Unified Approach to Interpreting Model Predictions (2017)

SM Lundberg, SI Lee

Counterfactual Explanations

- How to change the model's prediction on a specific instance
- Find the smallest change that changes the prediction to a predefined output.

Example: Loan application
Instance: (male, 30 years old, salary 1500€, 5 years professional experience, 2 kids, divorced, ….) -> rejected
-> Increase salary to 2000 €
-> Increase professional experience to 6 years

Counterfactual explanation desired property:

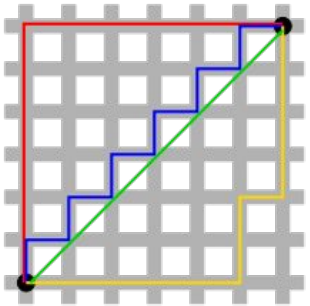
- Be as similar as possible to the original instance
- Change as few features as possible
- A plausible example

How to generate counterfactuals?

Counterfactual Explanations

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

Using scaled Manhattan distance: $$d\left(x, x'\right) = \sum_{j=1}^{p} \frac{\left|x_j - x'_j\right|}{MAD_j}$$



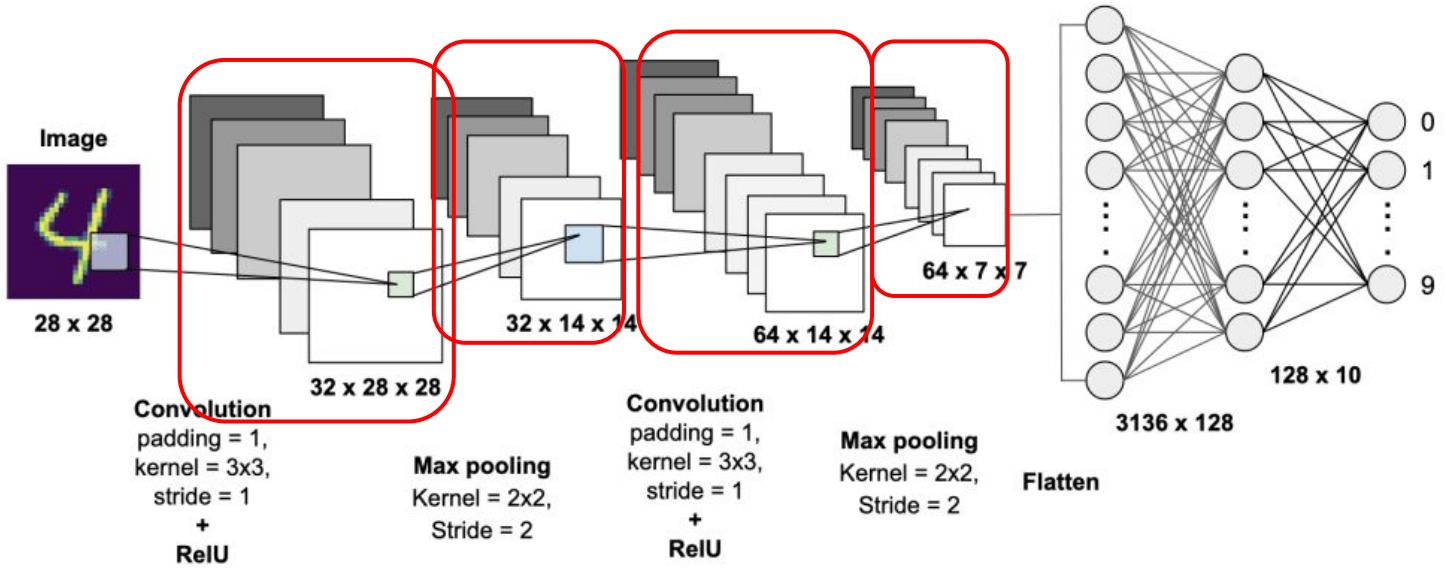
With median absolute deviation: $$MAD_j = \text{median}_{i \in \{1,\ldots,n\}}\left(\left|x_{i,j} - \text{median}_{l \in \{1,\ldots,n\}}\left(x_{l,j}\right)\right|\right)$$

Counterfactual explanations without opening the black box: Automated decisions and the GDPR(2017)

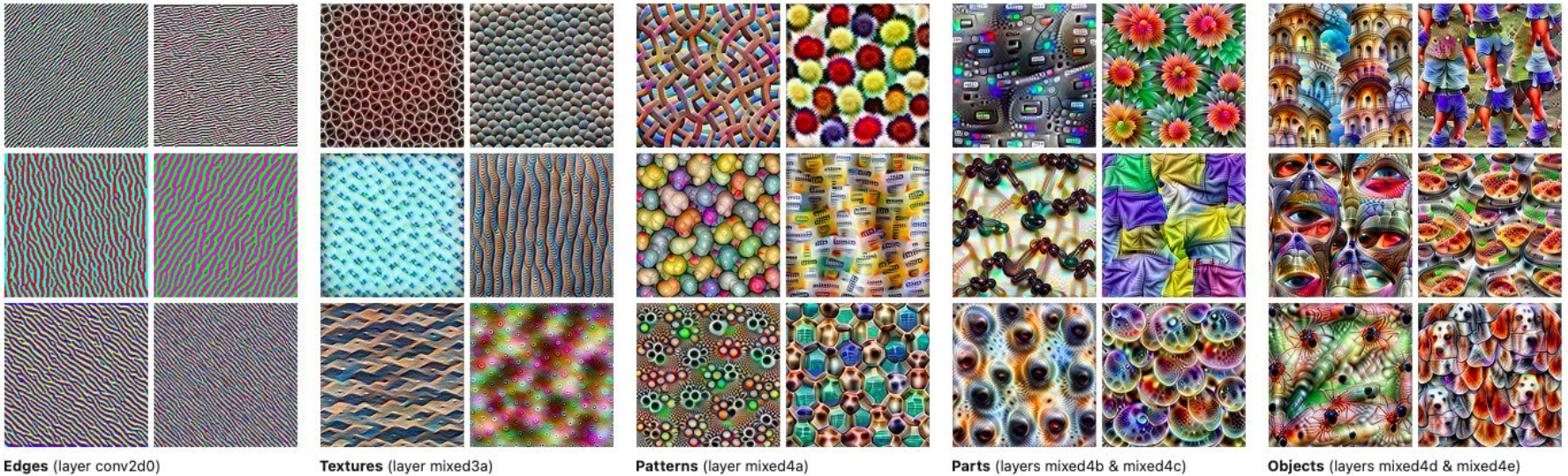
Wachter, Sandra, Brent Mittelstadt, and Chris Russell

Model-specific methods (Neural networks)

Feature visualization



Feature visualization



Feature VisualizationHow neural networks build up their understanding of images

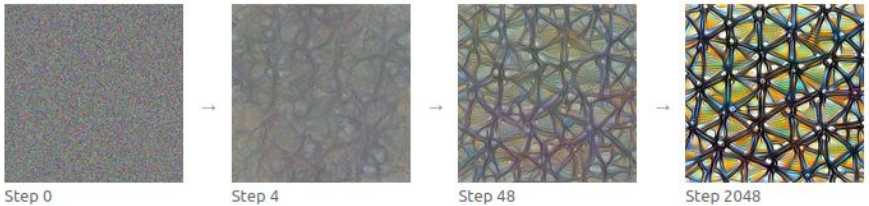
Chris Olah, Alexander Mordvintsev and Ludwig Schubert

Feature visualization

- Neuron
- Layer Channel
- Layer

For a fixed network:

- Find the image that would maximize unit output using



Feature VisualizationHow neural networks build up their understanding of images

Chris Olah, Alexander Mordvintsev and Ludwig Schubert

Network Dissection

Link feature map channels with human concepts

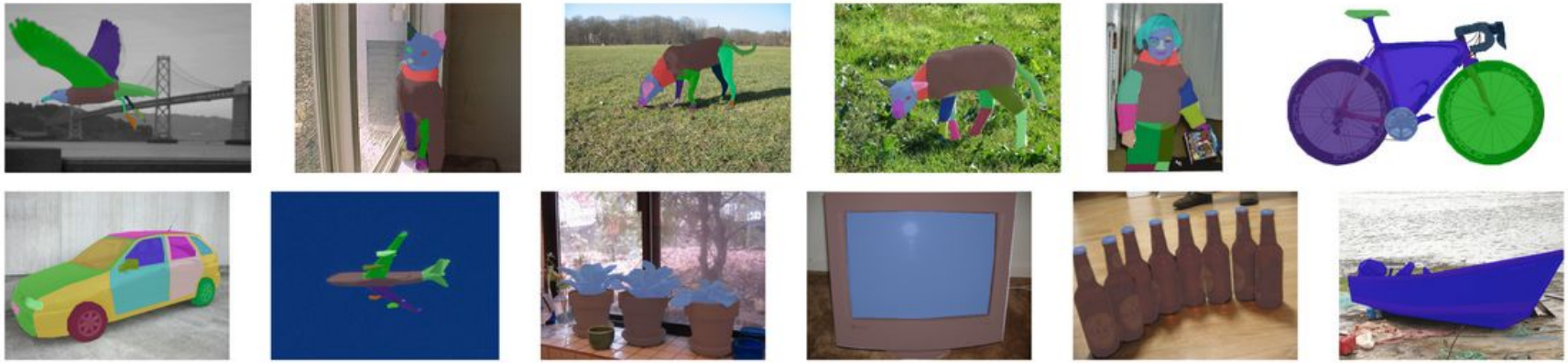
Principle:

- Define a large set of concepts
- Find units responding to these pre-defined concepts
- Measure the Activation-concept alignment

Network Dissection: Quantifying Interpretability of Deep Visual Representations (2017)

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba

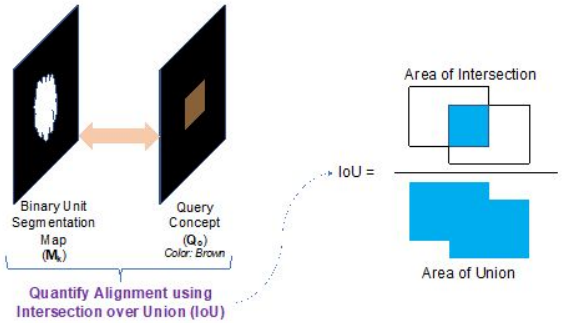
Network Dissection

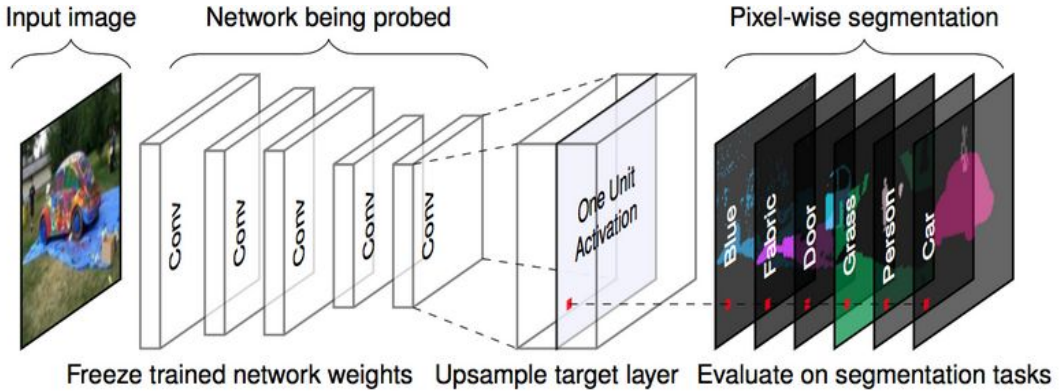


Network Dissection

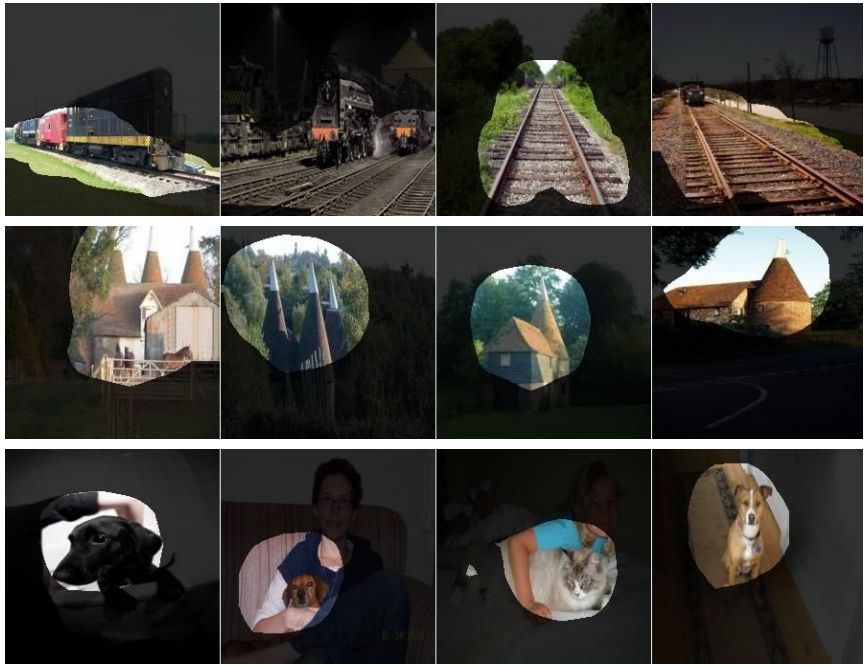
For every channel:

- Compute the distribution of the feature map pixels output on the entire concept dataset
- Compute the 0.995-quantile
- For every image create a mask of feature map pixels > 0.995-quantile
- Scale the feature map output
- Compute IoU

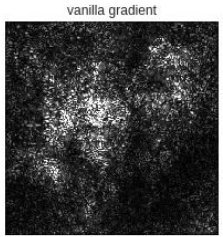


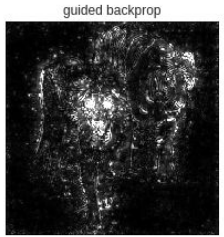


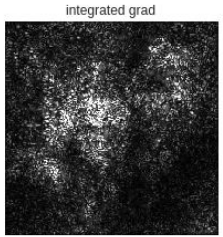
Network Dissection

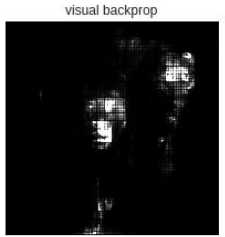


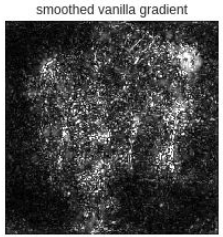
Saliency Maps

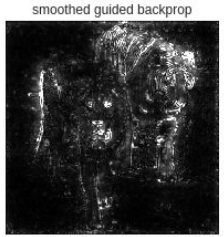


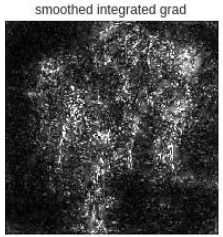


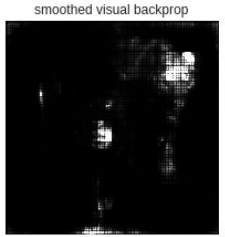




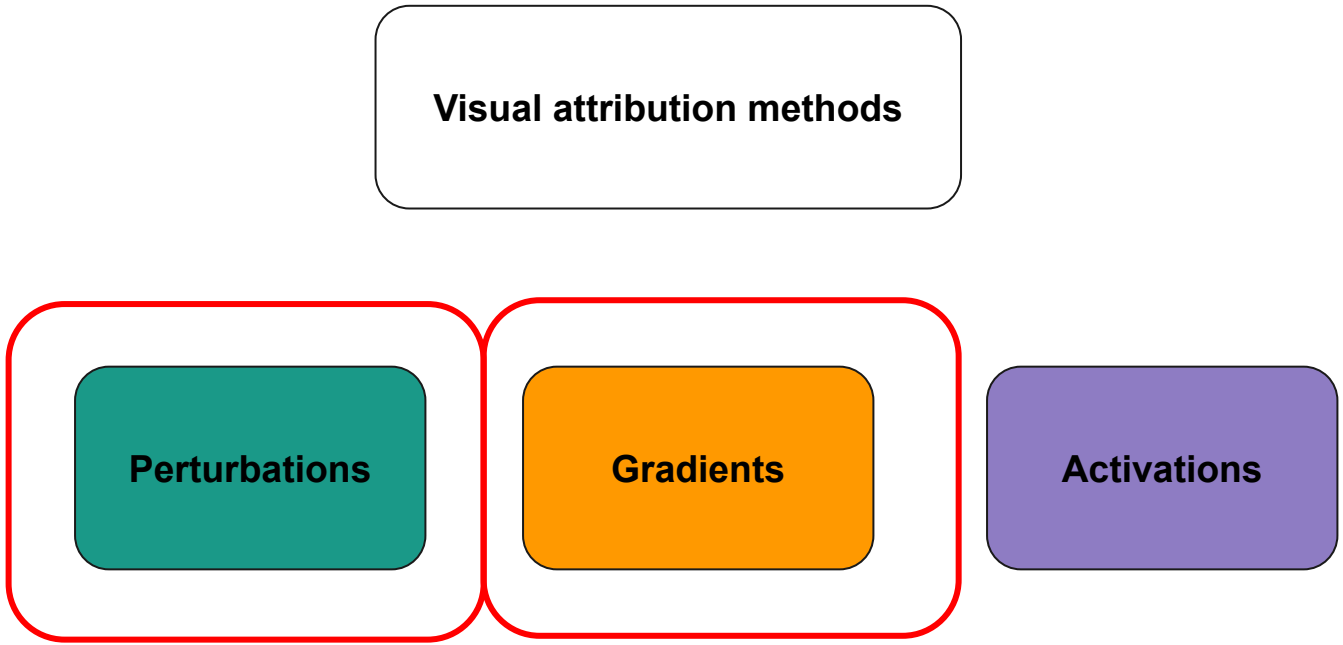




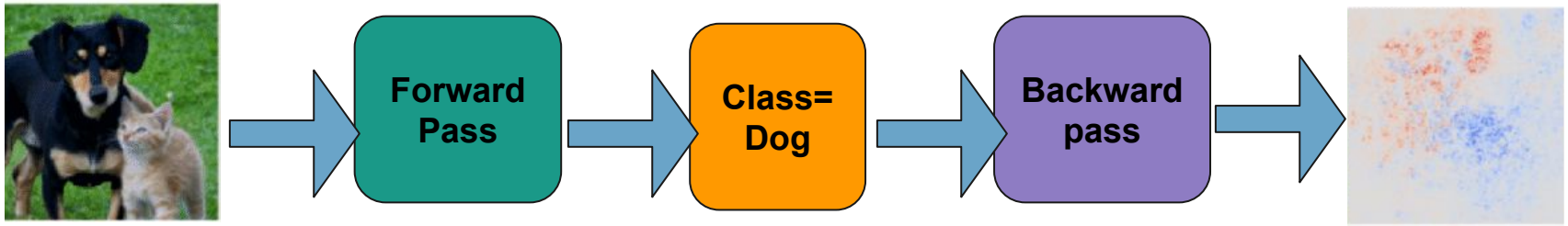




Saliency Maps



Vanilla gradient Saliency Maps

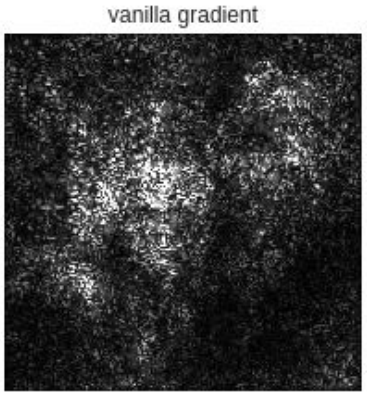


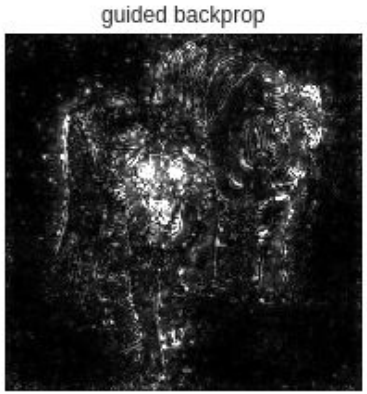
Deep inside convolutional networks: Visualising image classification models and saliency maps.”(2013)

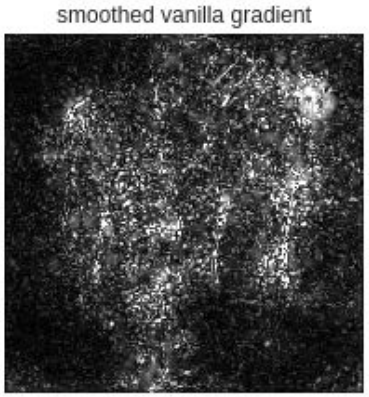
Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman

Gradient Saliency Maps

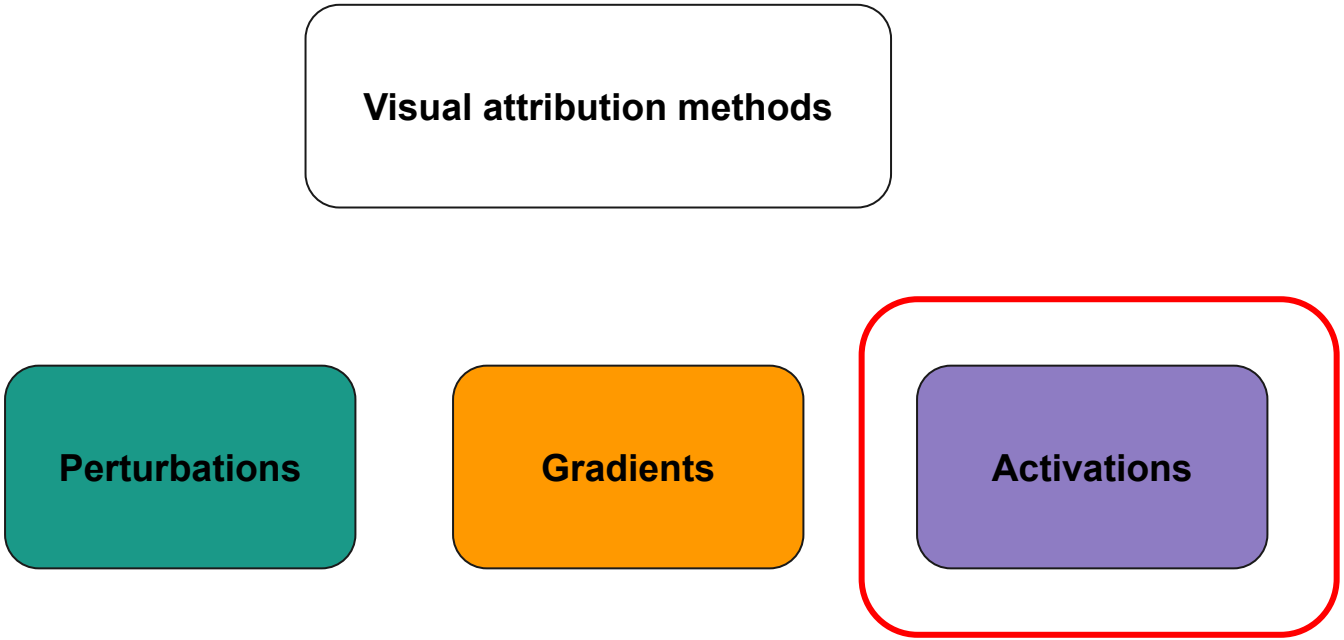
- Guided Backpropagation:
  Striving for Simplicity: The All Convolutional Net (2014)
  Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, Martin Riedmiller
- Smooth Grad:
  SmoothGrad: removing noise by adding noise (2017)
  Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, Martin Wattenberg







Saliency Maps



Grad-cam

- Gradient-weighted Class Activation Map
- Gradient back-propagated to the last convolutional layer
- Produce coarse localization map

Original Image | Grad-CAM ‘Cat’ | Grad-CAM ‘Dog’

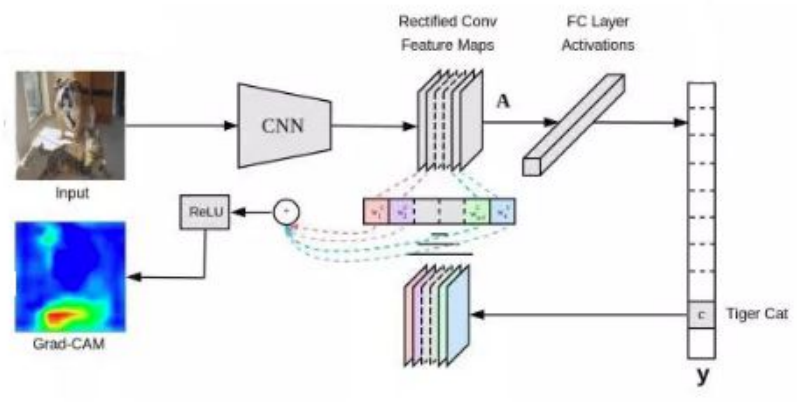
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

Grad-cam

Principle:

- Compute feature maps A^k with forward pass



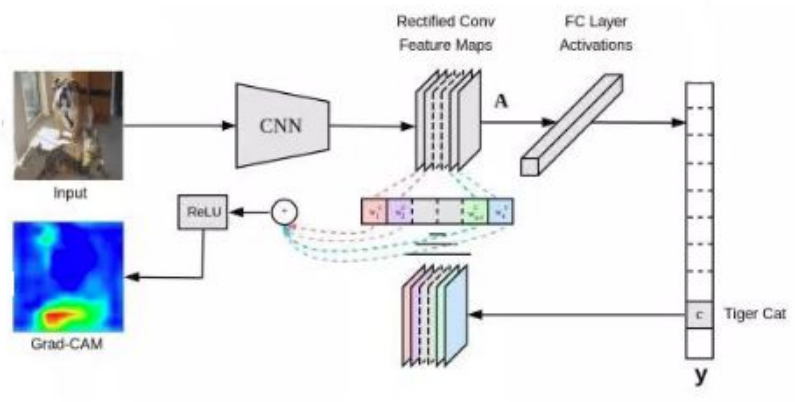
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

Grad-cam

Principle:

- Compute feature maps A^k with forward pass
- Backpropagate gradients of class C to the last convolutional layer



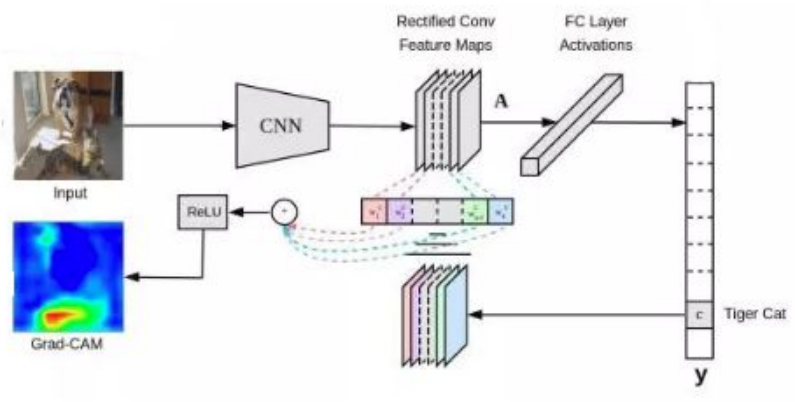
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

Grad-cam

Principle:

- Compute feature maps A^k with forward pass
- Backpropagate gradients of class C to the last convolutional layer
- Average gradients : $\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \frac{\partial y^c}{\partial A_{ij}^k}$



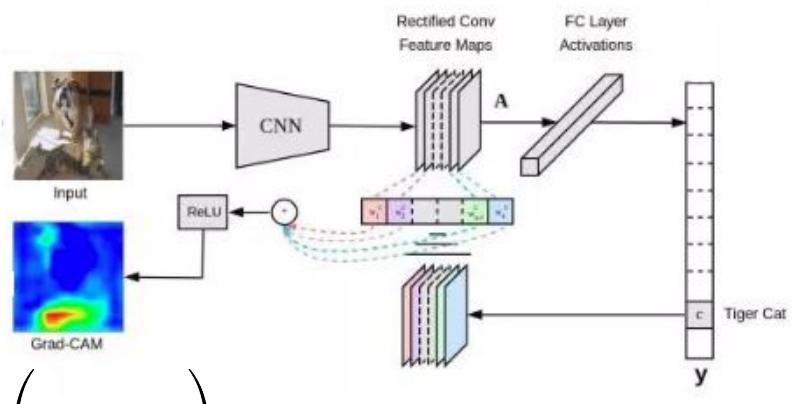
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

Grad-cam

Principle:

- Compute feature maps A^k with forward pass
- Backpropagate gradients of class c to the last convolutional layer
- Average gradients : $\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \frac{\partial y^c}{\partial A_{ij}^k}$



- Compute Grad-CAM heatmap: $L^c_{\text{Grad-CAM}} = \text{Re}\, LU\left(\sum_k \alpha_k^c A^k\right)$

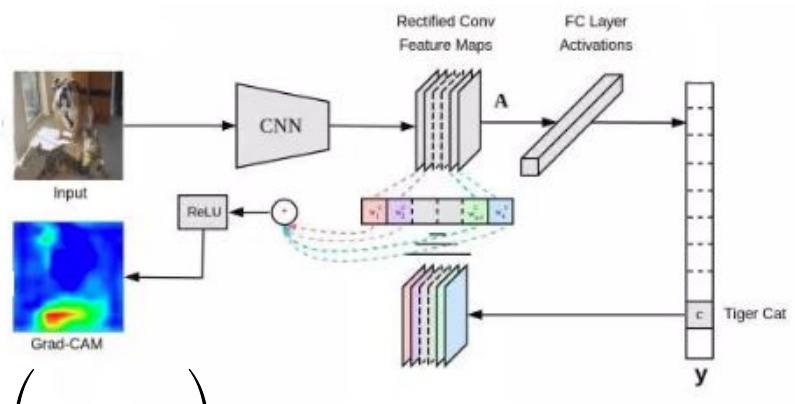
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

Grad-cam

Principle:

- Compute feature maps A^k with forward pass
- Backpropagate gradients of class c to the last convolutional layer
- Average gradients : $\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \frac{\partial y^c}{\partial A_{ij}^k}$



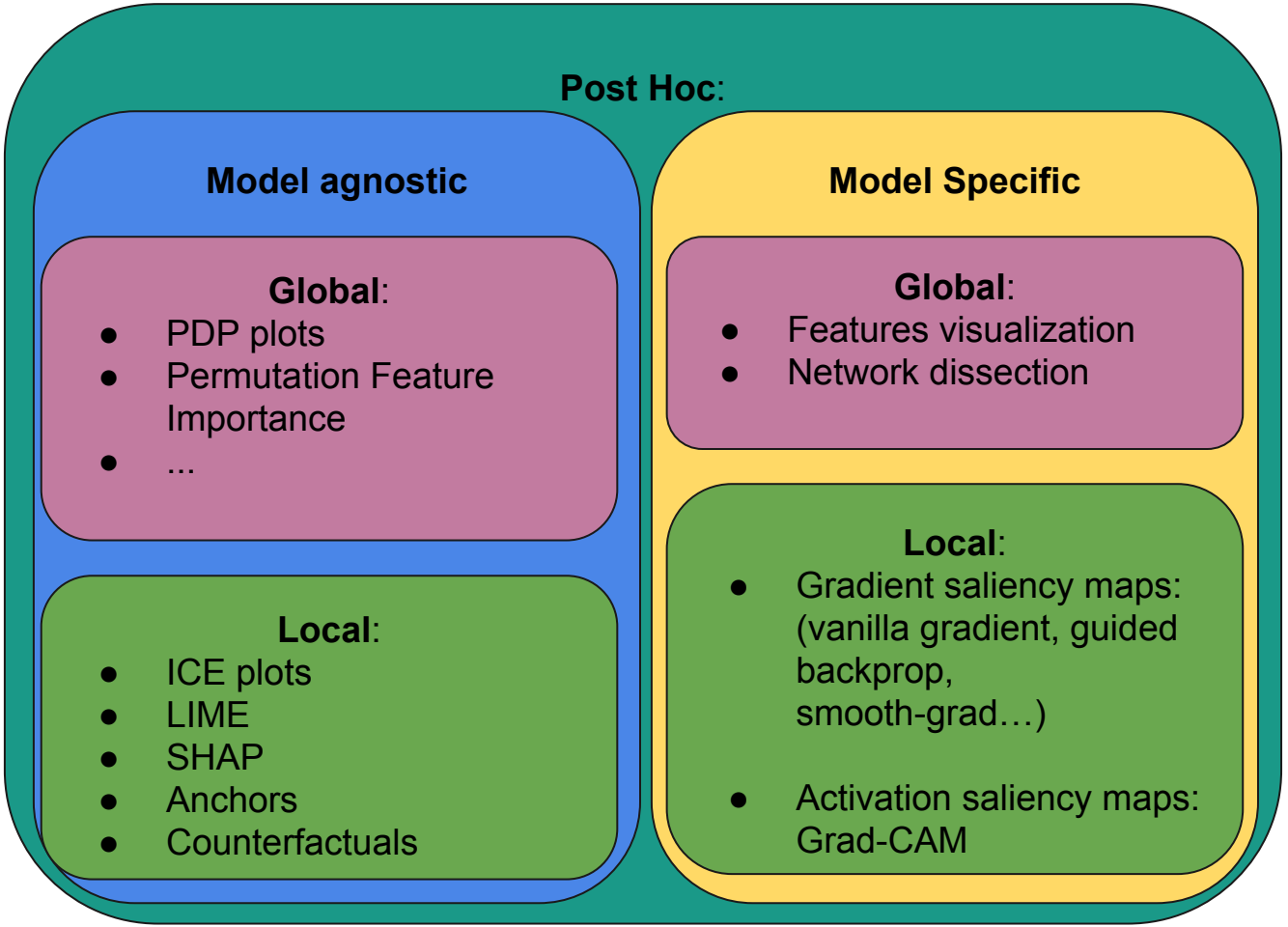
- Compute Grad-CAM heatmap: $L^c_{\text{Grad-CAM}} = \text{Re}\, LU\left(\sum_k \alpha_k^c A^k\right)$
- Upsample the heatmap

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra

Recap:

Intrinsic:
- Linear models
- Decision trees
- Naive Bayes
- K-nn
- ...



References

Molnar, Christoph. “Interpretable machine learning. A Guide for Making Black Box Models Explainable”, 2019. https://christophm.github.io/interpretable-ml-book/.

Ajay Thampi: Interpretable AI

- Greedy function approximation: A gradient boosting machine.(2001)
- A Simple and Effective Model-Based Variable Importance Measure.(2018)
- Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation
- Random Forests (2001)
- "Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)
- Anchors: High-Precision Model-Agnostic Explanations (2018)
- A Unified Approach to Interpreting Model Predictions (2017)
- Counterfactual explanations without opening the black box: Automated decisions and the GDPR(2017)
- Feature VisualizationHow neural networks build up their understanding of images(2018)
- Network Dissection: Quantifying Interpretability of Deep Visual Representations (2017)
- Deep inside convolutional networks: Visualising image classification models and saliency maps.”(2013)
- Striving for Simplicity: The All Convolutional Net (2014)
- SmoothGrad: removing noise by adding noise (2017)
- Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2017)