Imperial College London

BioMedIA

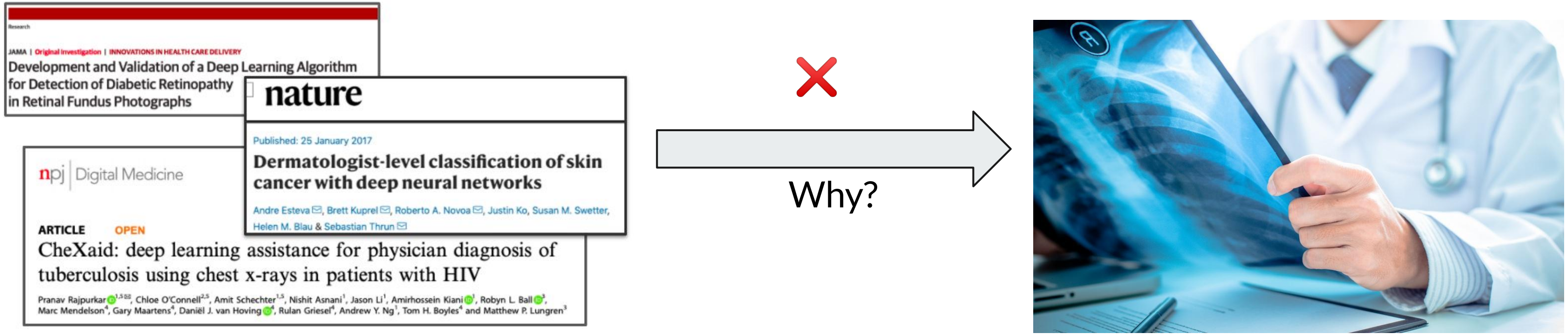
Brief Introduction to Explainable/Interpretable AI

Avinash Kori

What is Interpretability?

"**Interpretability is the degree to which a human can understand the cause of a decision**" or in other terms "**Interpretability is the degree to which a human can consistently predict the model's result.**" [1]

- Higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made
- A model is better interpretable than another model if its decisions are easier for a human to comprehend

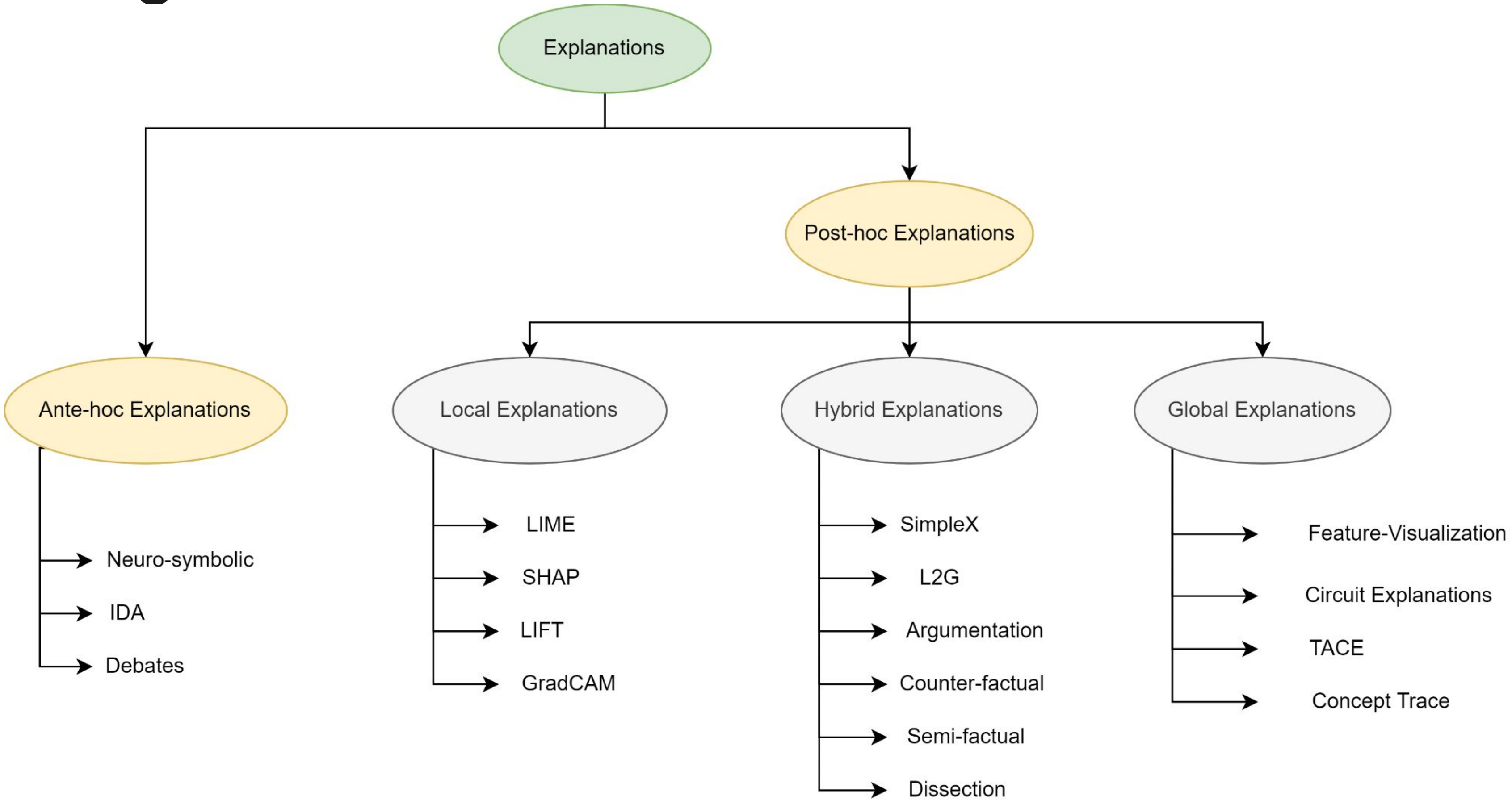


[1] Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, pp.1-38.

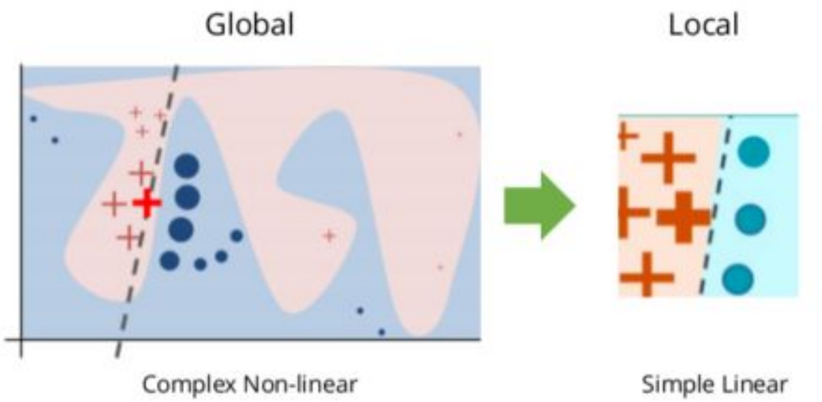
Interpretability vs Explainability

- Interpretability focuses on understanding the model
- Explainability focuses on explaining models reasoning
- Interpretability -> Explainability

Categorization



Post-hoc: Local Explanations: LIME[2]

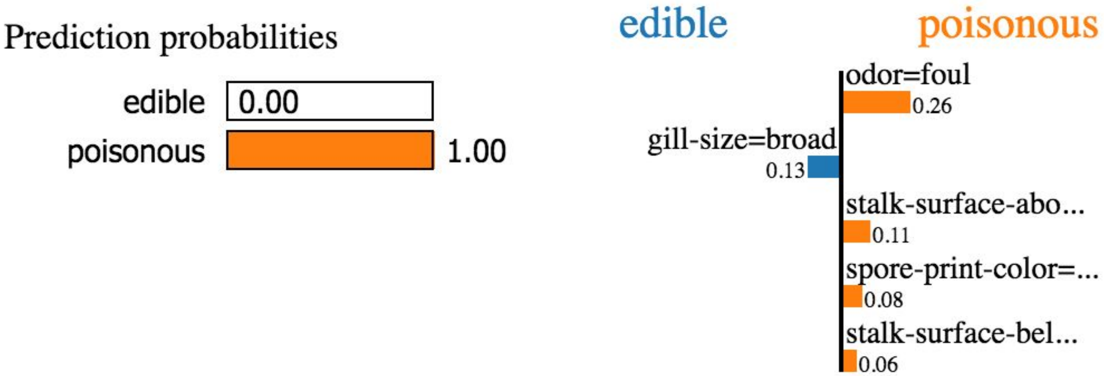


- LIME- Local Interpretable Model-agnostic Explanations
- Constructs data based on local small scale perturbations around a selected point
- Constructs simple linear model g(.), trained on perturbed data
- Explanations are feature importance/contribution in making certain decision

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

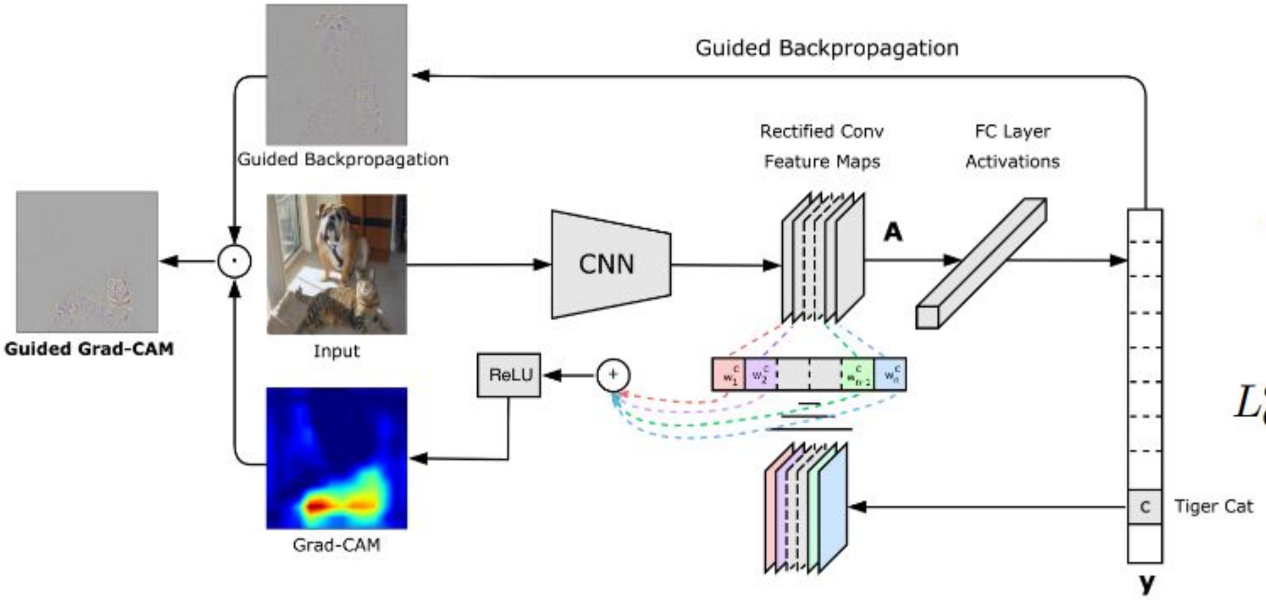
[2] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Post-hoc: Local Explanations: LIME



Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

Post-hoc: Local Explanations: GradCAM[4]

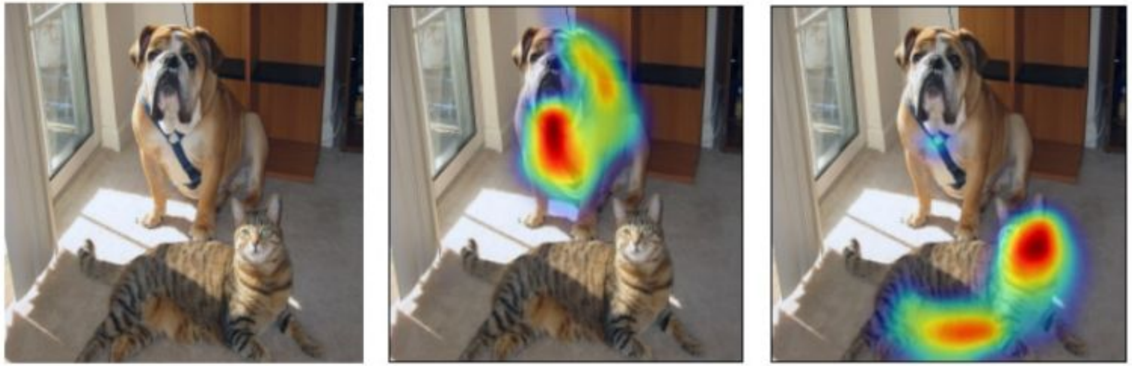


$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

$$L^c_{\text{Grad-CAM}} = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

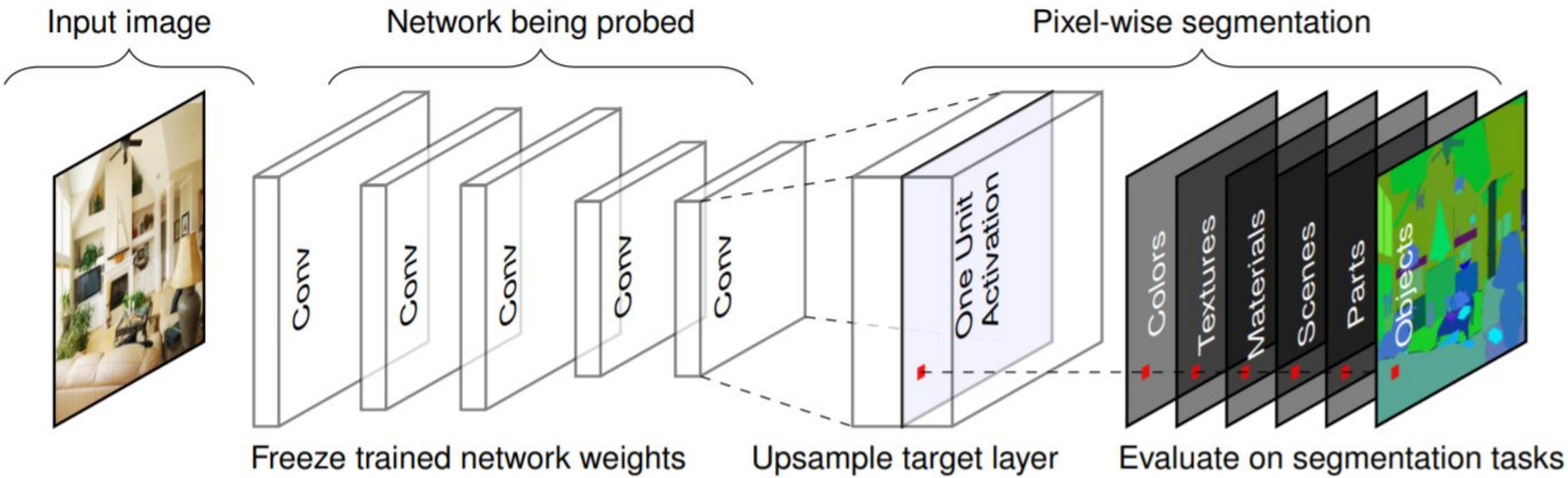
[4] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

Post-hoc: Local Explanations: GradCAM



(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

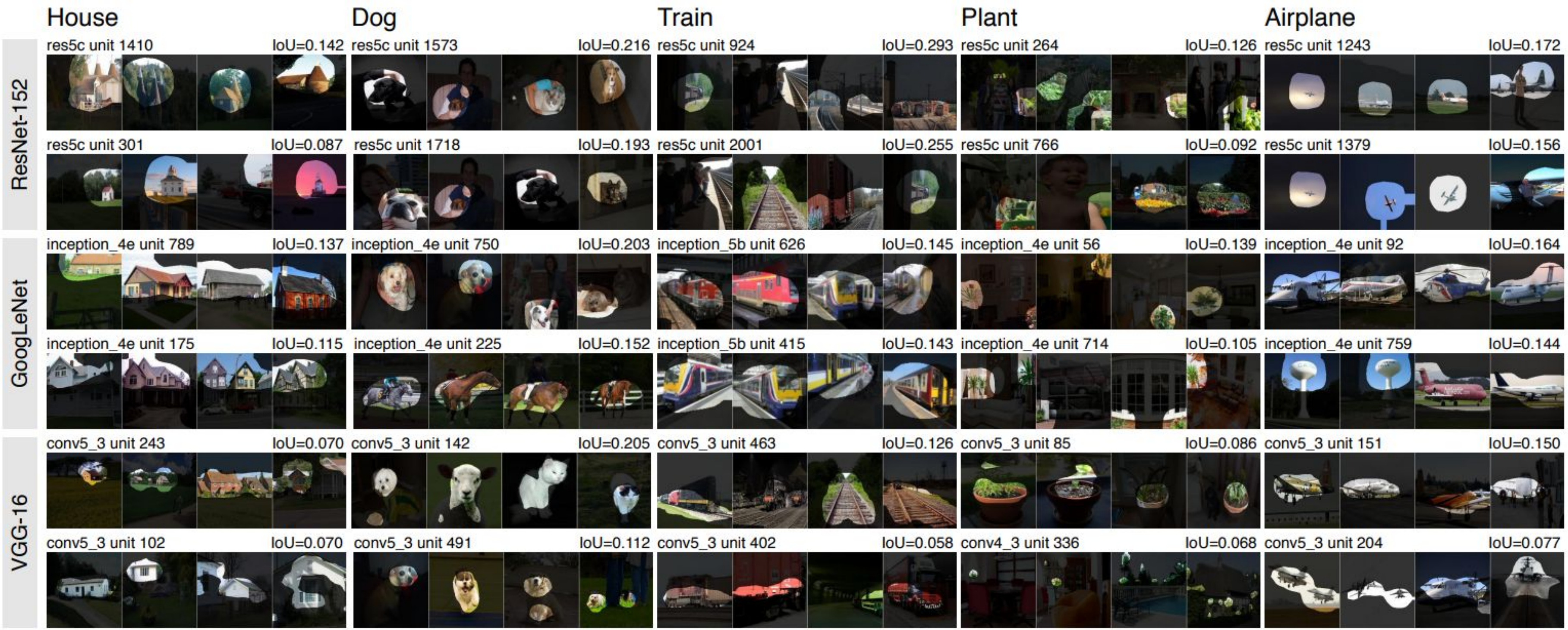
Post-hoc: Hybrid Explanations: Dissection[3]



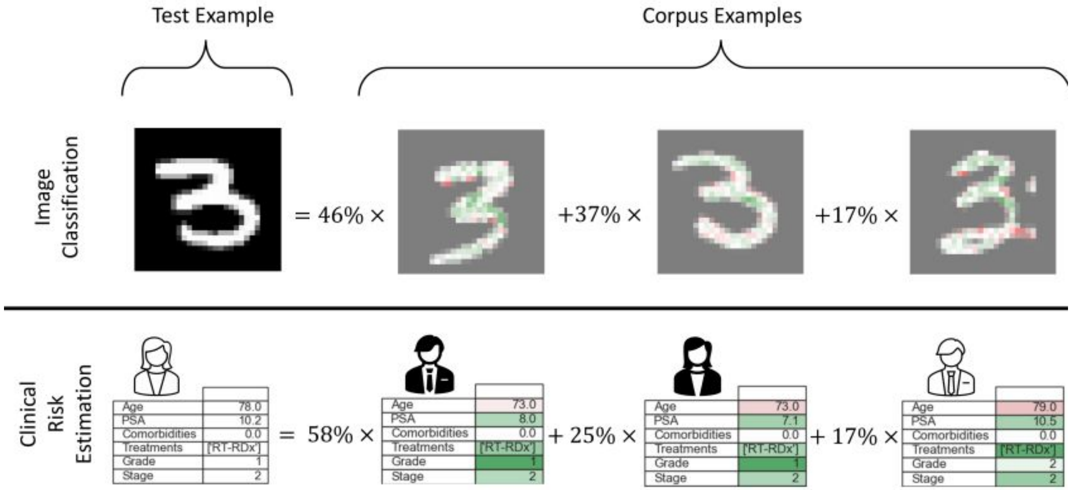
$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|},$$

[3] Bau, D., Zhou, B., Khosla, A., Oliva, A. and Torralba, A., 2017. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6541-6549).

Post-hoc: Hybrid Explanations: Dissection



Post-hoc: Hybrid Explanations: SimplEx[5]



[5] Crabbé, J., Qian, Z., Imrie, F. and van der Schaar, M., 2021. Explaining Latent Representations with a Corpus of Examples. Advances in Neural Information Processing Systems, 34.

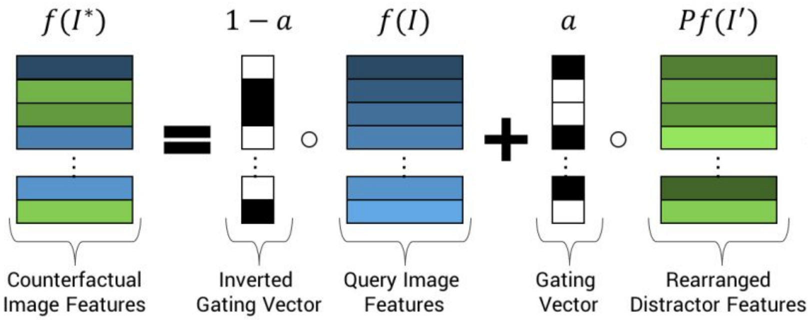
Post-hoc: Hybrid Explanations: Counterfactuals[6]

- “What-if” explanations
- what region in the image made the model predict class *c* instead of class *c′*?

$$\underset{P,\mathbf{a}}{\text{minimize}} \quad ||a||_1$$

$$\text{s.t.} \quad c' = \operatorname{argmax} \; g((\mathbb{1} - \mathbf{a}) \circ f(I) + \mathbf{a} \circ P f(I'))$$

$$a_i \in \{0, 1\} \;\; \forall i \;\; \text{and} \;\; P \in \mathcal{P}$$



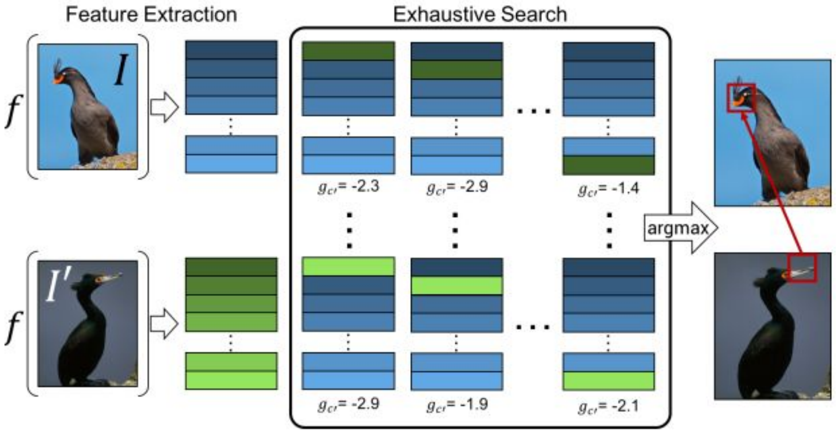
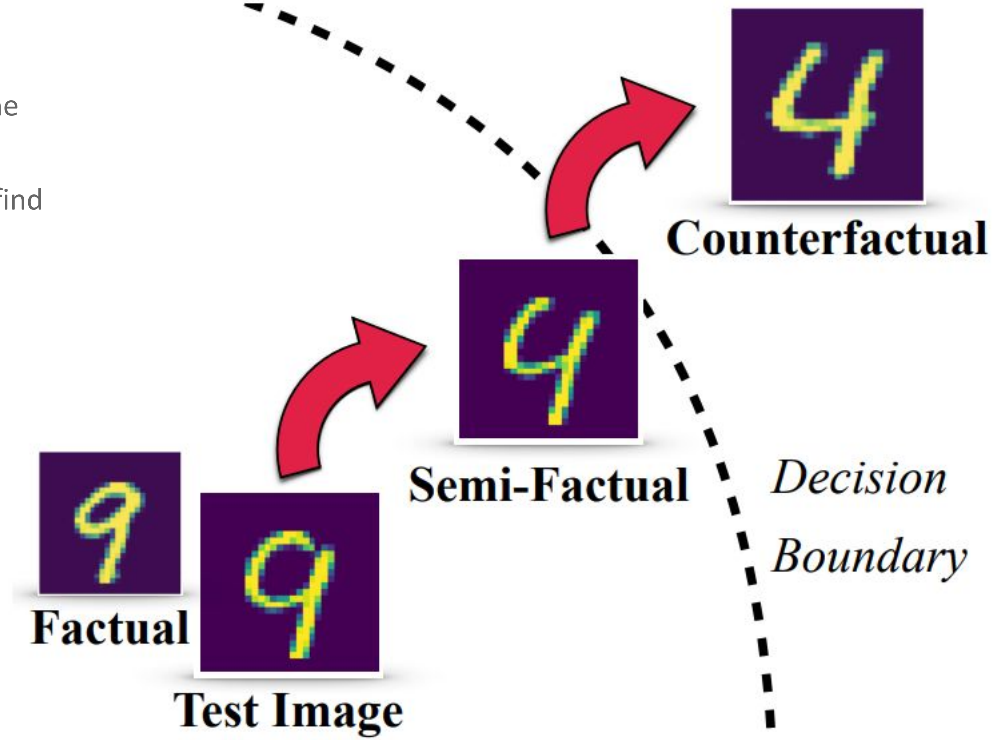


Figure 4. In our exhaustive best-edit search, we check all pairs of query-distractor spatial locations and select whichever pair maximizes the log probability of the distractor class c'.

[6] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D. and Lee, S., 2019, May. Counterfactual visual explanations. In International Conference on Machine Learning (pp. 2376-2384). PMLR.

Post-hoc: Hybrid Explanations: Semi-factuals[7]

- "Even-if" explanations
- Even if the feature value is changed from *a* to *b* the image would still be classified as *c*
- This paper proposes a gradient based method to find the decision boundary



[7] Kenny, E.M. and Keane, M.T., 2021. On generating plausible counterfactual and semi-factual explanations for deep learning. AAAI-21, pp.11575-11585.

Post-hoc: Global Explanations: FeatureVis[8]

$$x^* = \arg\max_x \left(\Phi_{k,l}(x) - R_\theta(x) - \lambda \|x\|_2^2 \right)$$

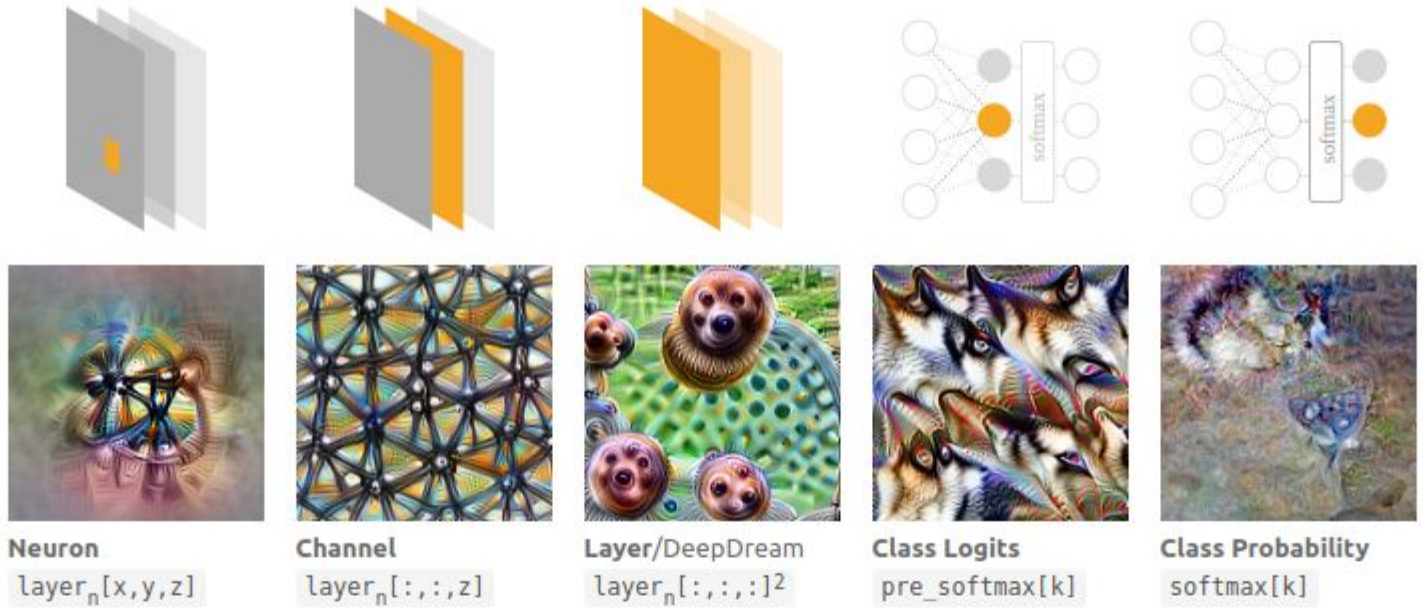
$$R_{TV}(I) = \sum_{k=0}^{c} \sum_{u=0}^{h} \sum_{v=0}^{w} ([I(u, v+1, k) - I(u, v, k)] + [I(u+1, v, k) - I(u, v, k)])$$

$$L(x, s) = \sum_i \sum_j (k(x_i, x_j) + k(s_i, s_j) - 2k(x_i, s_j))$$

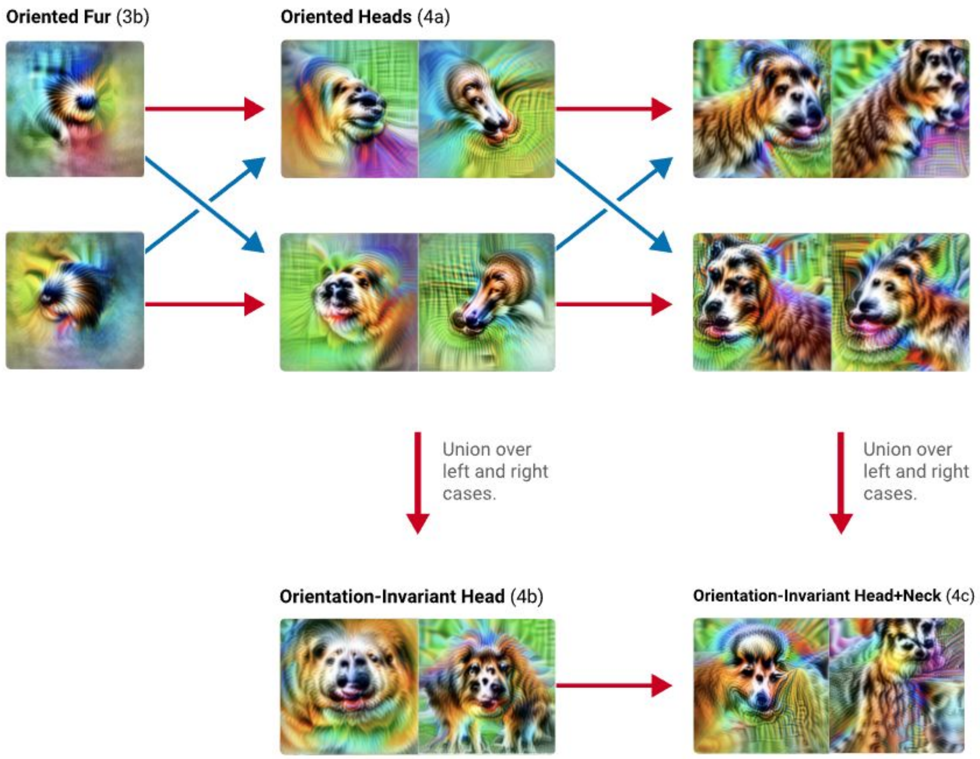
- Mechanistic form of interpretability
- Hand engineer an explainable model by interpreting trained complex model

[8] Olah, C., Mordvintsev, A. and Schubert, L., 2017. Feature visualization. Distill, 2(11), p.e7.

Post-hoc: Global Explanations: FeatureVis

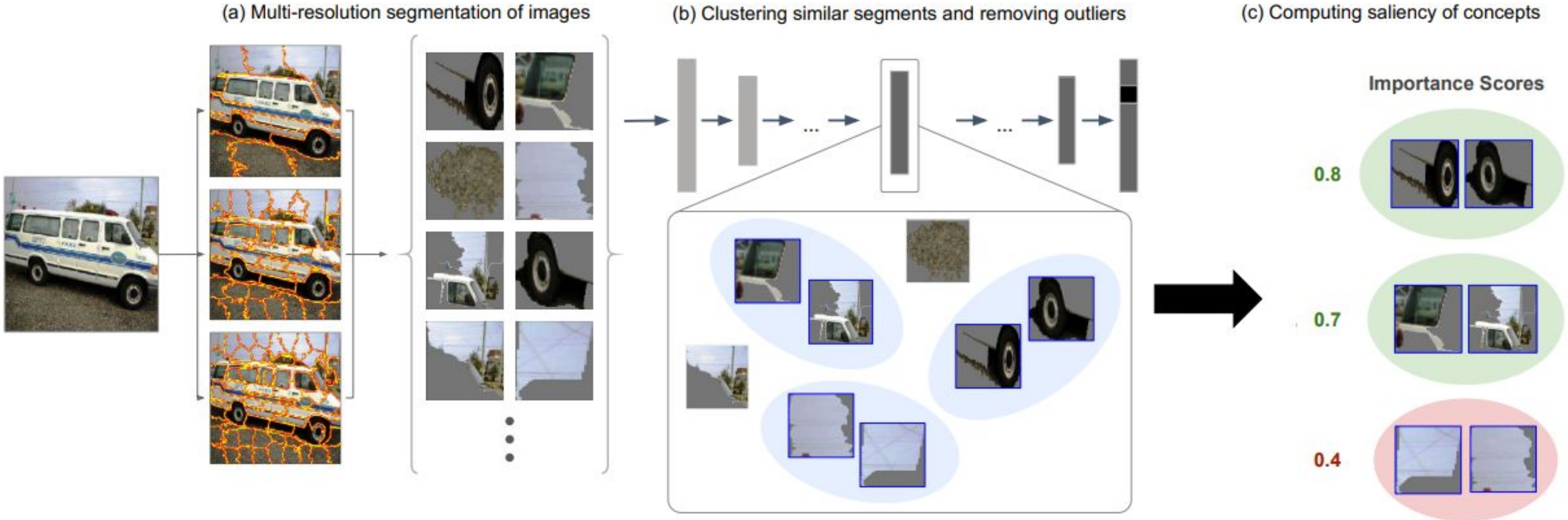


Post-hoc: Global Explanations: Circuits[9]



[9] Olah, C., Mordvintsev, A. and Schubert, L., 2017. Feature visualization. Distill, 2(11), p.e7.

Post-hoc: Global Explanations: TACE[10]

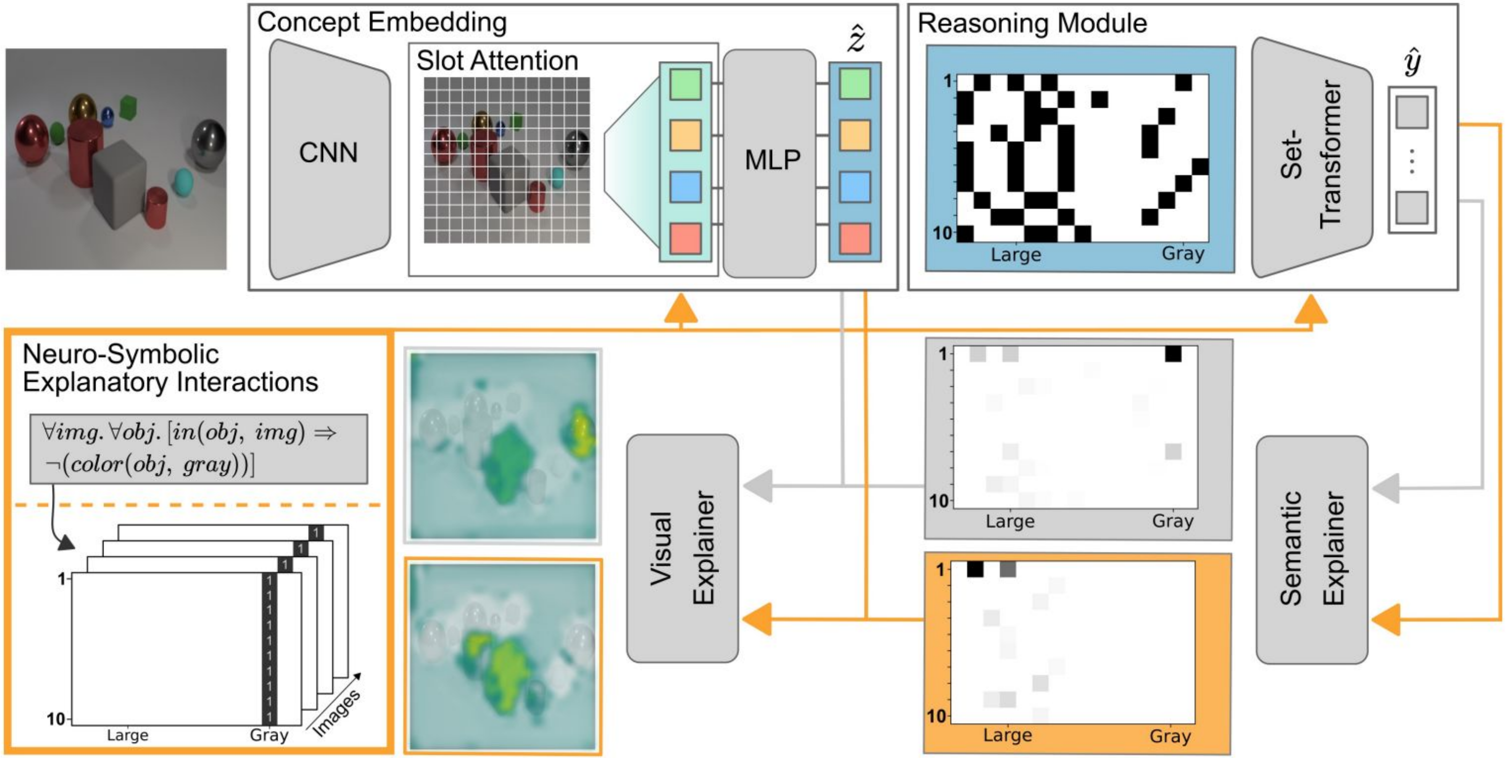


$$\mathrm{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}$$

$$S_{C,k,l}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(\boldsymbol{x}) + \epsilon \boldsymbol{v}_C^l) - h_{l,k}(f_l(\boldsymbol{x}))}{\epsilon}$$

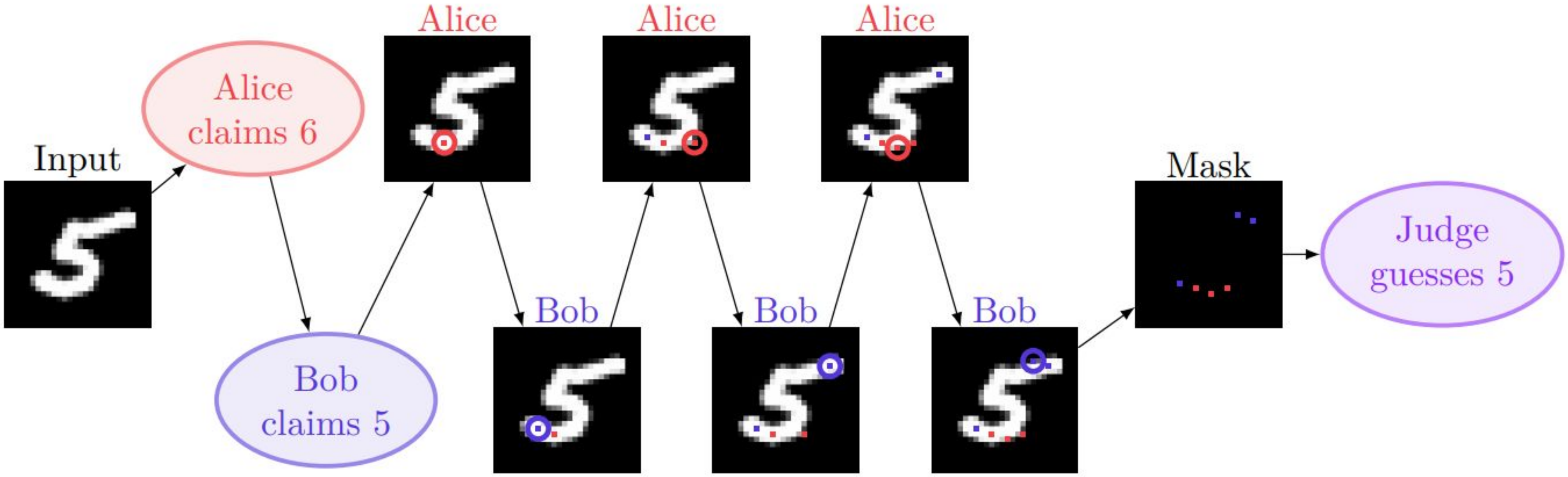
[10] Ghorbani, A., Wexler, J., Zou, J.Y. and Kim, B., 2019. Towards automatic concept-based explanations. Advances in Neural Information Processing Systems, 32.

Ante-hoc: Neuro-Symbolic[11]



[11] Stammer, W., Schramowski, P. and Kersting, K., 2021. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3619-3629).

Ante-hoc: Debate



Desired Properties for Explanations

- **Faithfulness:**
  - Measures the contribution of a model in making model specific explanations
  - An explanation is faithful to the model if it represents the true reasoning process of the model

- **Robustness:**
  - Measures the effect of small scale perturbations on explanations

- **Stability:**
  - Measures variability across runs
  - Model is supposed to follow same reasoning for similar examples

- **Coherence :**
  - Measures the degree of contradicting reasoning made by a model

Desired Properties for Explanations

Methods	**faithfulness**	**stability**	**Coherence**	**Robustness**
Dissection	✔	✔	✔	❌
GradCAM	✔	✔	❌	❌
SHAP	✔	❌	❌	❌
LIME	✔	❌	❌	❌
TACE	❌	❌	✔	✔
Counter/Semi factuals	❌	✔	✔	❌

Questions?