# Explainable AI (XAI)

## AI College

Taegyun Jeon (SI Analytics)

# Taegyun Jeon

- **Founder and Chief Executive Officer**
  - SI Analytics, Jul 2018 – Present
- **AI Team Leader**
  - Satrec Initiative, Sep 2016 – Jun 2018
- **Google Developer Experts – Machine Learning**
  - Google, Sep 2017 – Present
- **PhD – Machine Learning and Computer Vision**
  - Gwangju Institute of Science and Technology (GIST)
- **Co-administrator – TensorFlow User Group Korea**
  - Korea, May 2017 – Oct 2018

**Github**: https://github.com/tgjeon
**Facebook**: https://www.facebook.com/taegyun.jeon

# Agenda

1. Prerequisites
2. Problem Definition
   1. Methods
   2. Evaluations
3. Grading Rule
4. References

# 01 Prerequisites

# Prerequisite

- Key concept of
  **"Explaining Deep Model via Attribution: Method and Evaluation"**
  - Intro: "Explaining Explanations: An Overview of Interpretability of Machine Learning" (arXiv:1806.00069)

- **Python**

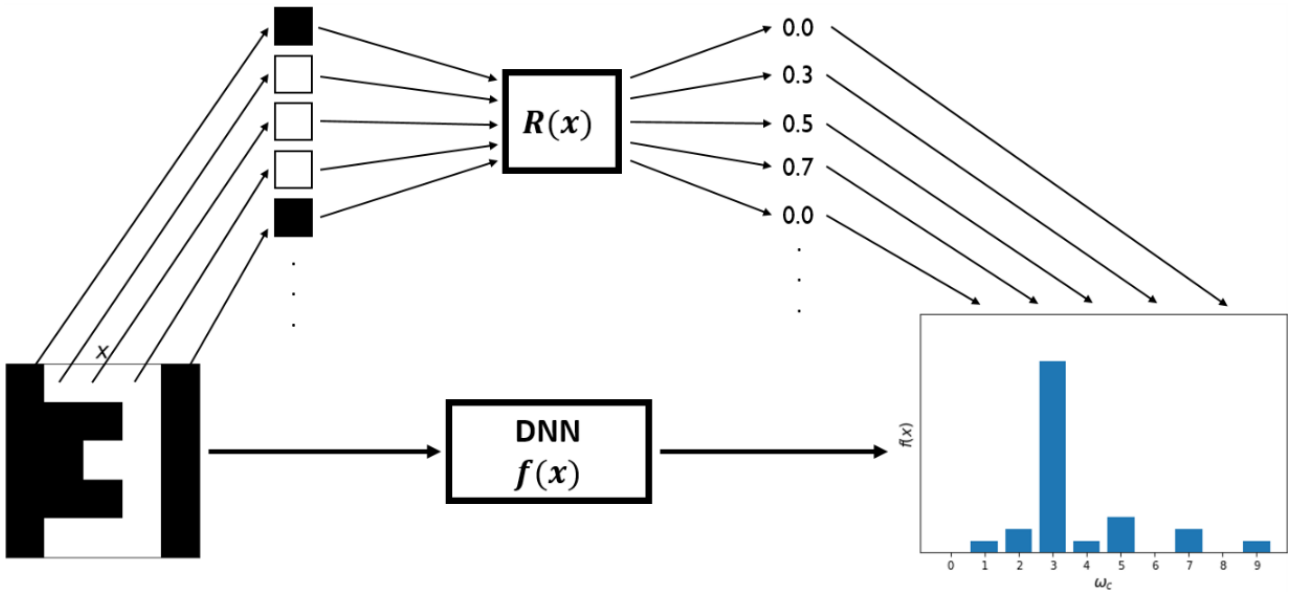
- **Deep Learning Framework** (Ex. TensorFlow, PyTorch)

# 02 Problem Definition (Methods)

# Explaining Deep Model via Attribution Method

- Problem Definition

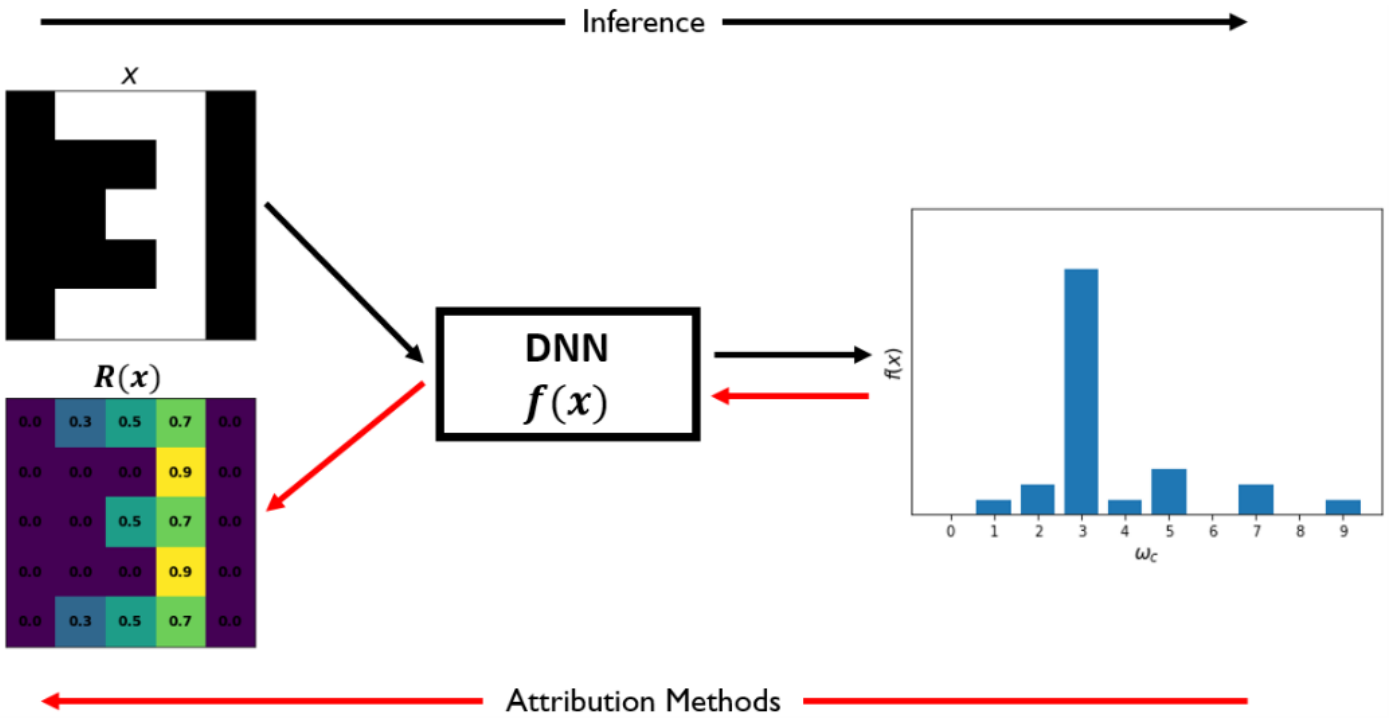
Given an image $\boldsymbol{x} \in \mathbb{R}^n$ and a decision $f(\boldsymbol{x})$,
assign to each pixel $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ attribution values $R_1(\boldsymbol{x}), R_2(\boldsymbol{x}), \ldots, R_n(\boldsymbol{x})$.



# Explaining Deep Model via Attribution Method

- Problem Definition

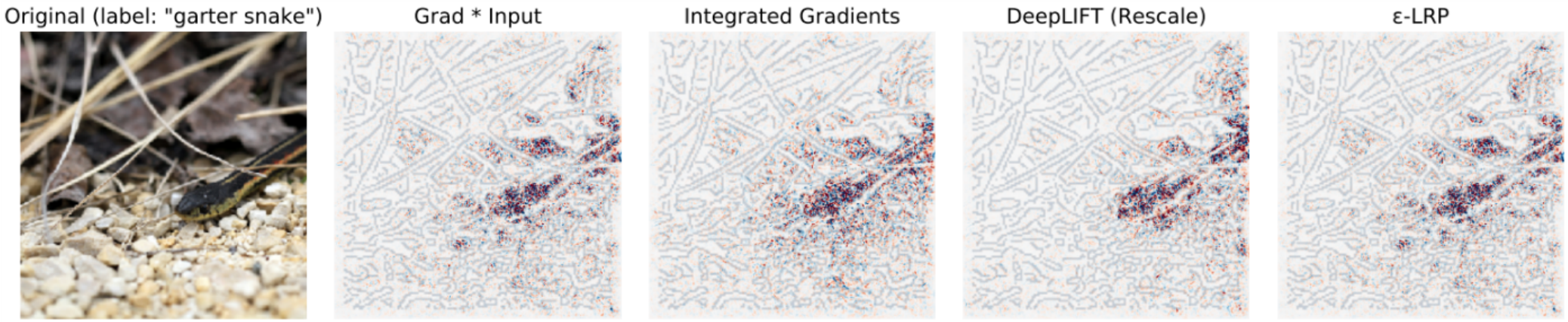
  Usually visualized as heatmaps



# Explaining Deep Model via Attribution Method

- Problem Definition

Usually visualized as heatmaps



# Explaining Deep Model via Attribution Method

- Problem Definition

**The Baseline Attribution Method Saliency Map**

- Gradient of the decision $f(\boldsymbol{x})$ with respect to the input image $\boldsymbol{x}$:

$$Saliency(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$$

- Can be calculated through backpropagation.

"Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps", https://arxiv.org/pdf/1312.6034.pdf

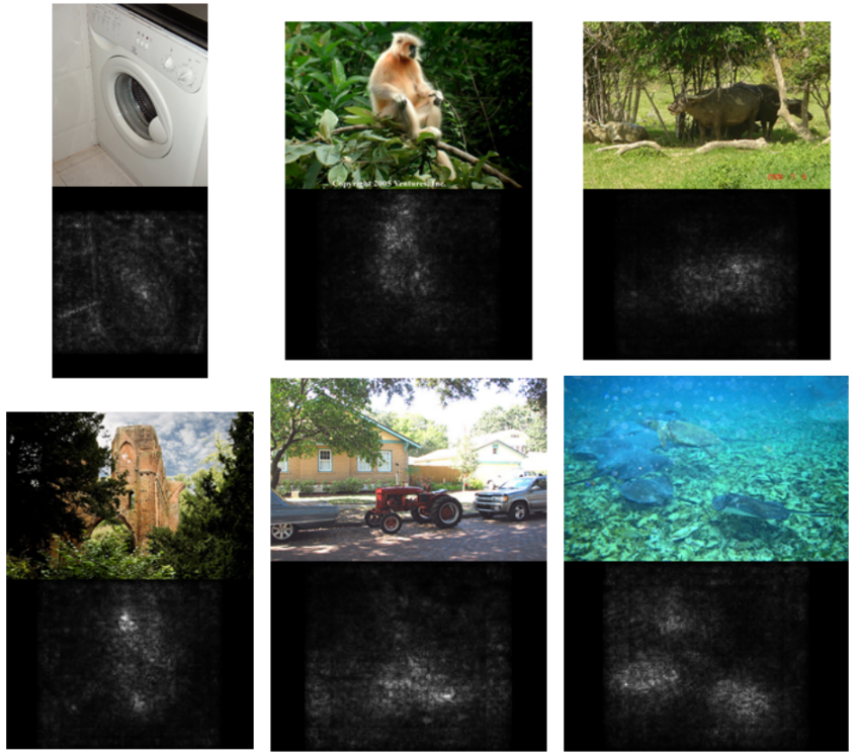
# Explaining Deep Model via Attribution Method

- Problem Definition

**The Baseline Attribution Method Saliency Map**

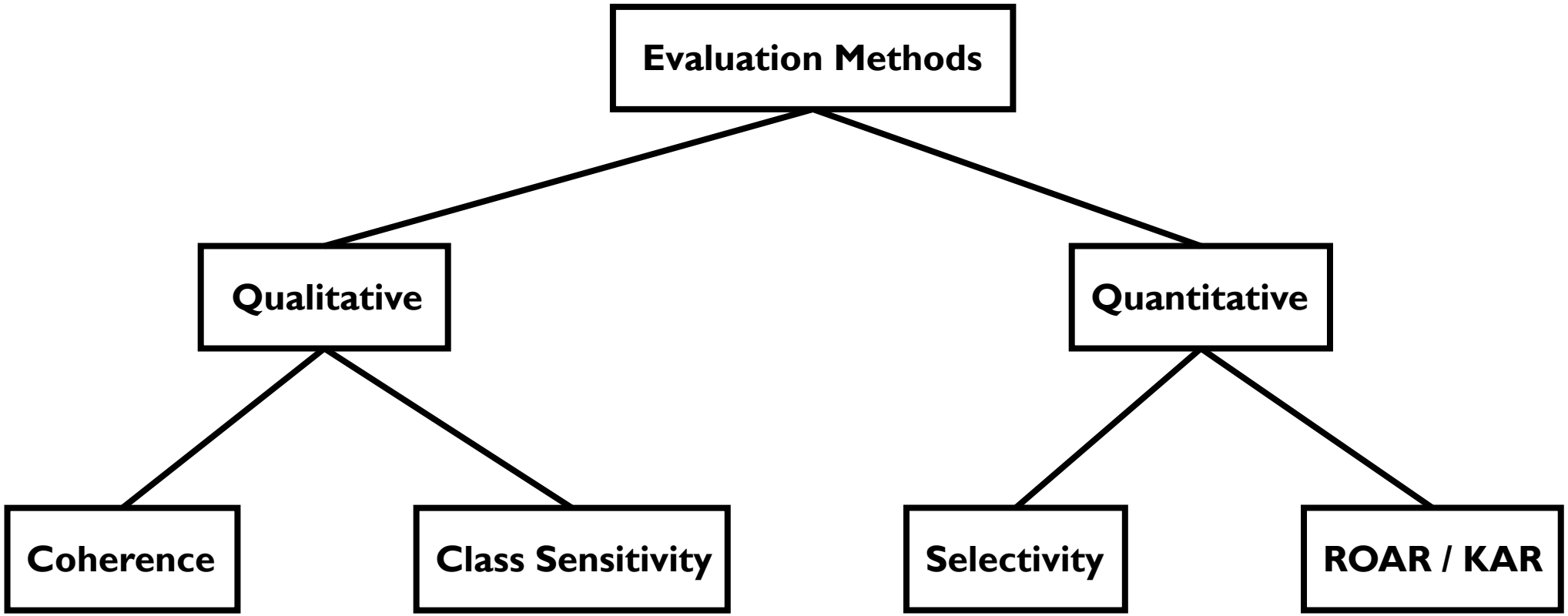
- Saliency maps are very noisy!
- Only roughly correlated with the object(s) of interest.

**Question: How to improve saliency maps?**



# 02 Problem Definition (Evaluation)

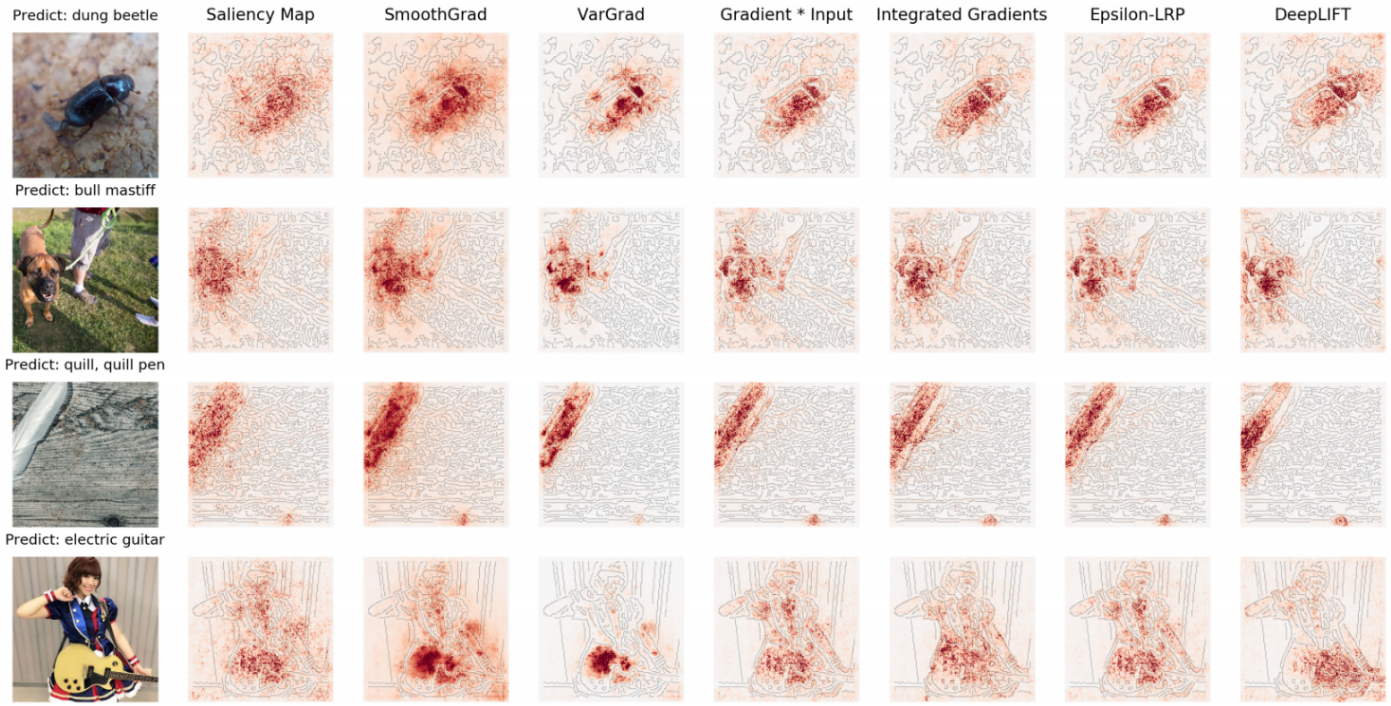
# Evaluation for Attribution Methods



# Evaluation for Attribution Methods

**Coherence** | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**

- Attributions should fall on discriminative features (e.g. the object of interest)

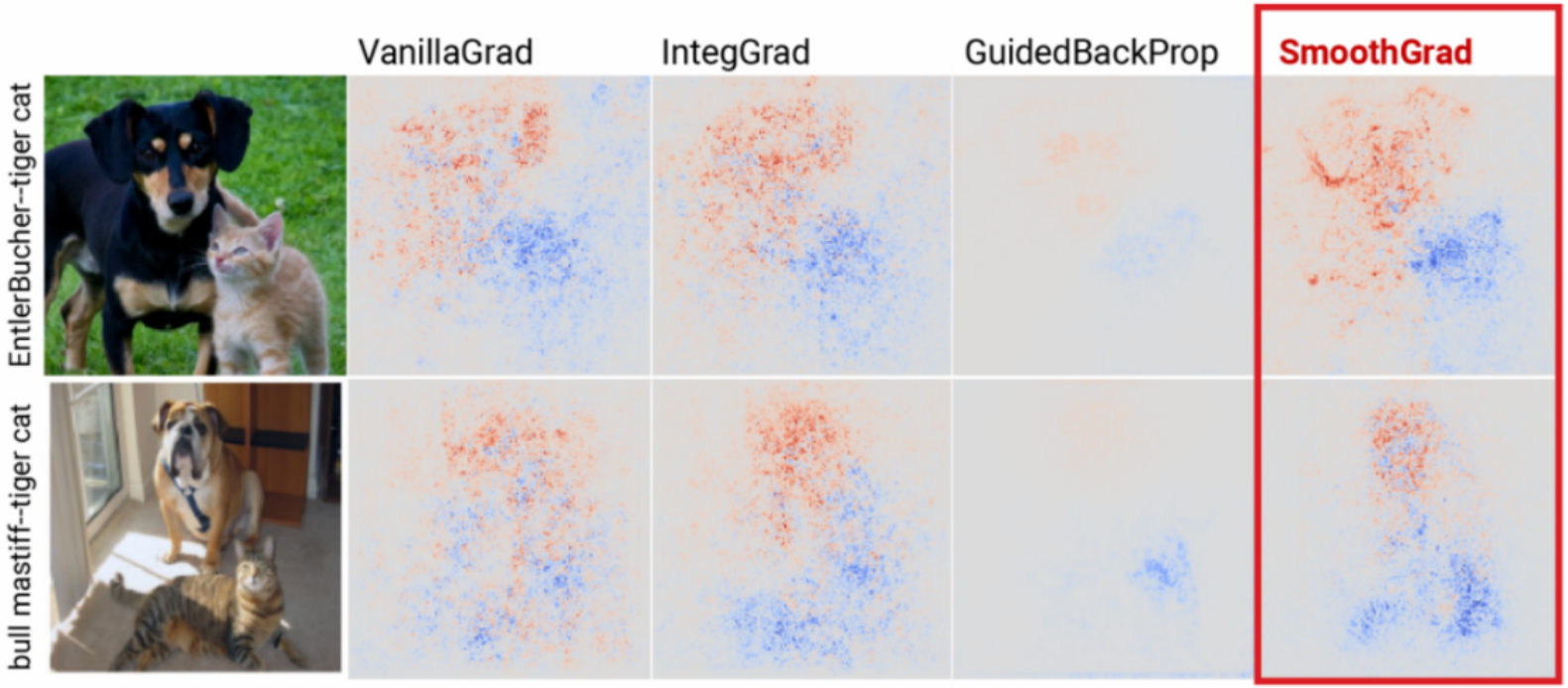



"Noise-adding Methods of Saliency Maps as Series of Higher Order Partial Derivative", https://arxiv.org/pdf/1806.03000.pdf

# Evaluation for Attribution Methods

**Coherence** | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**

- Attributions should be sensitive to class labels




"SmoothGrad: removing noisy by adding noise", https://arxiv.org/pdf/1706.03825.pdf

# Evaluation for Attribution Methods

**Coherence** | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**

- Removing feature with high attribution should cause large decrease in class probability

**Algorithm**

Sort pixel attribution values $R_i(\boldsymbol{x})$

Iterate:

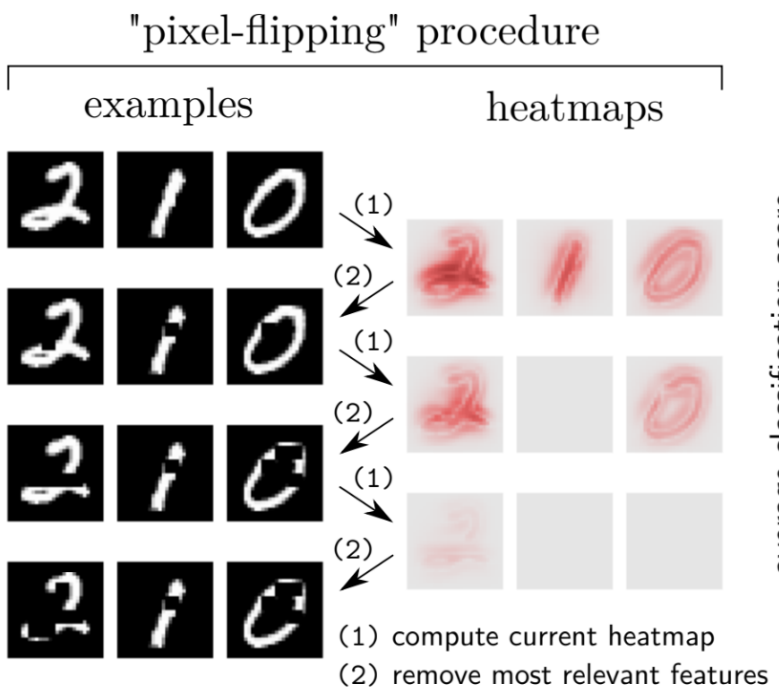
Remove pixels

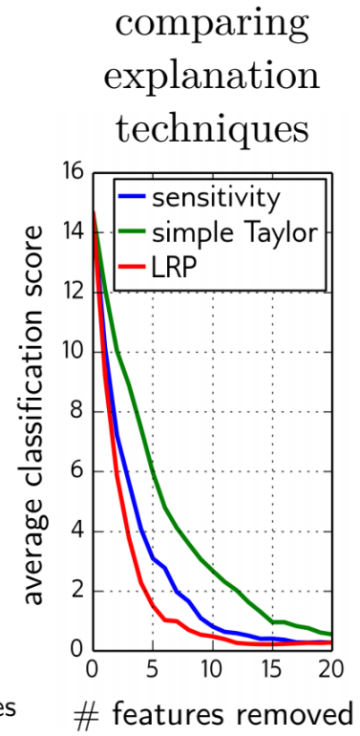
Evaluate $f(\boldsymbol{x})$

Measure decrease of $f(\boldsymbol{x})$

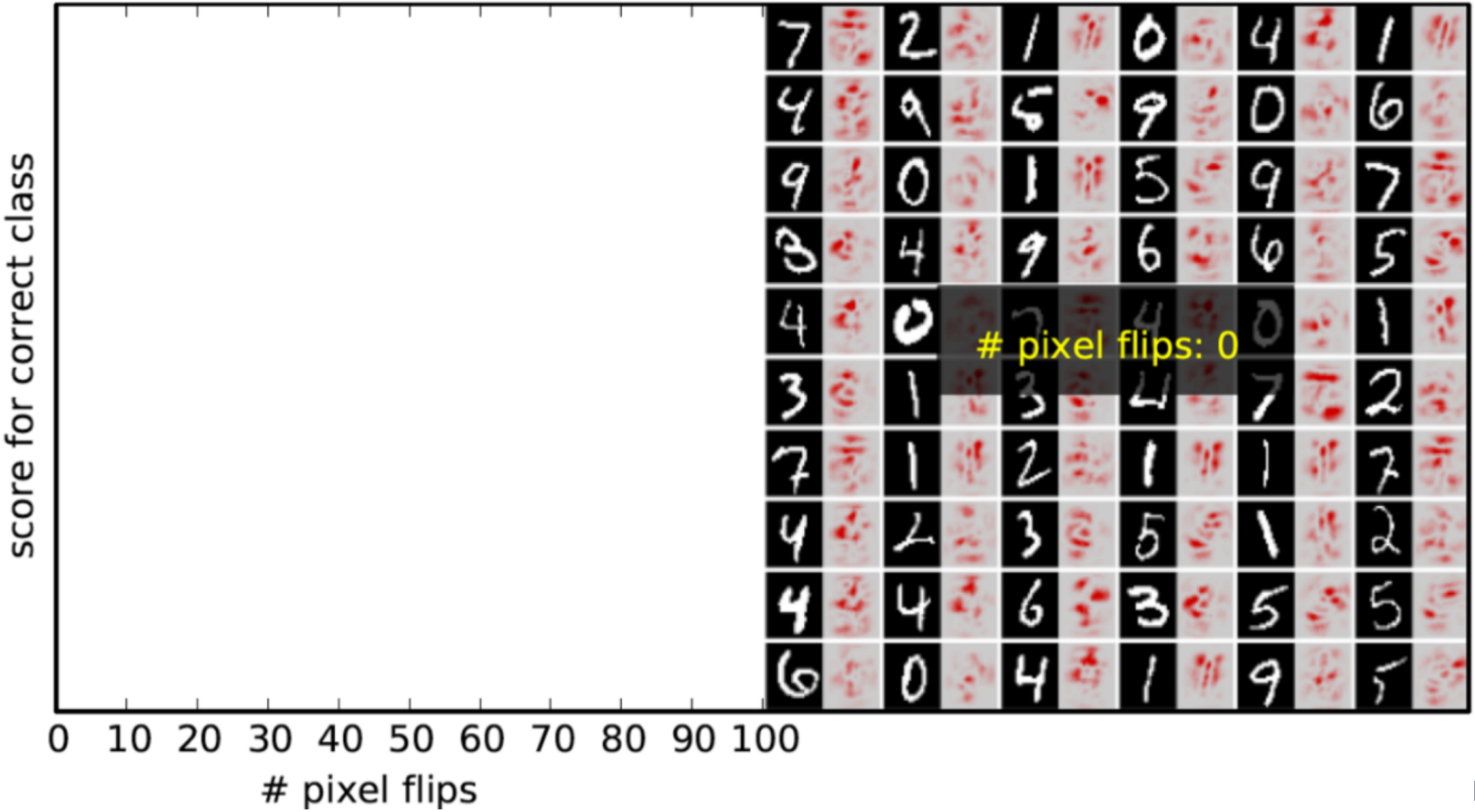
# Evaluating Attribution Methods

**Coherence** | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**

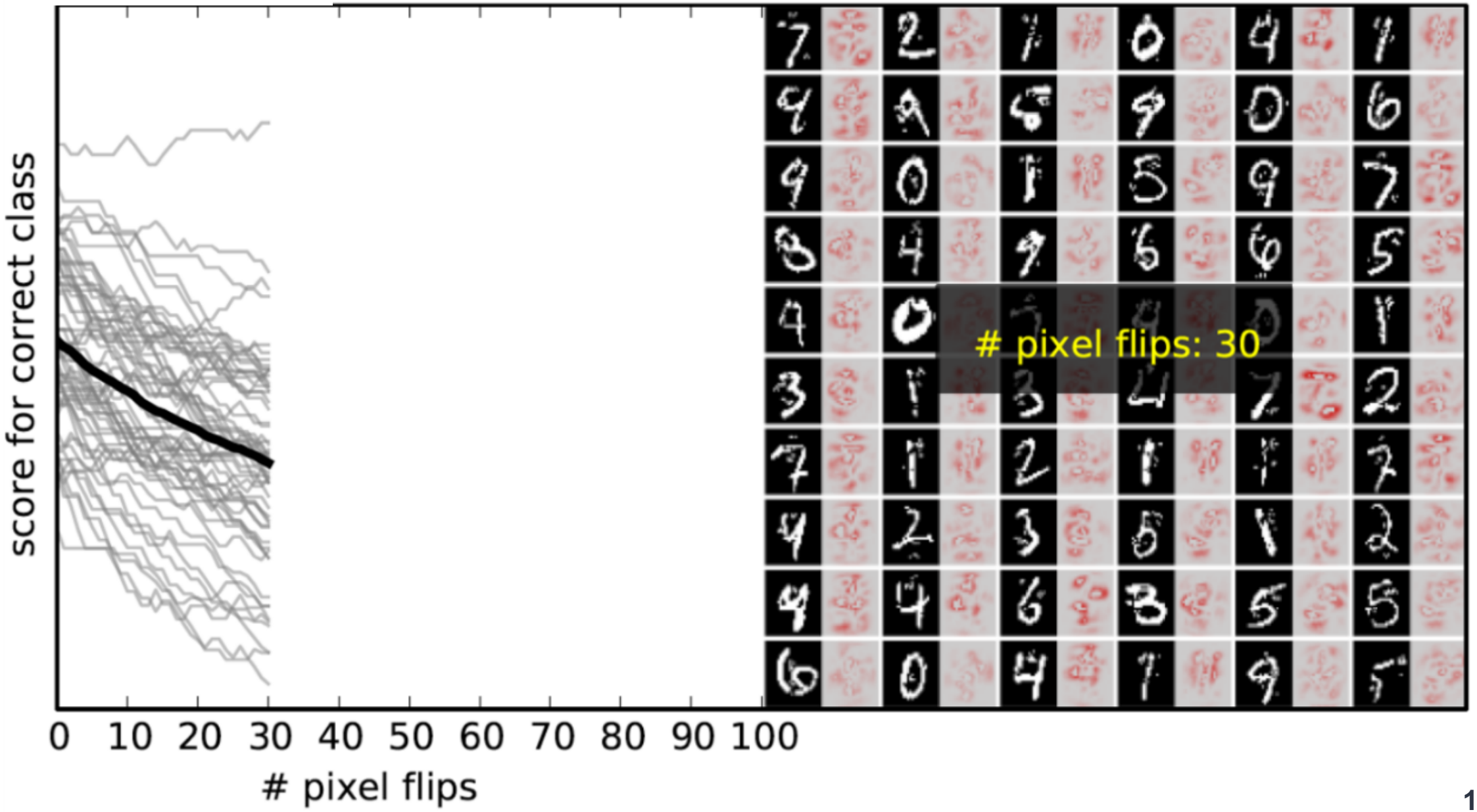


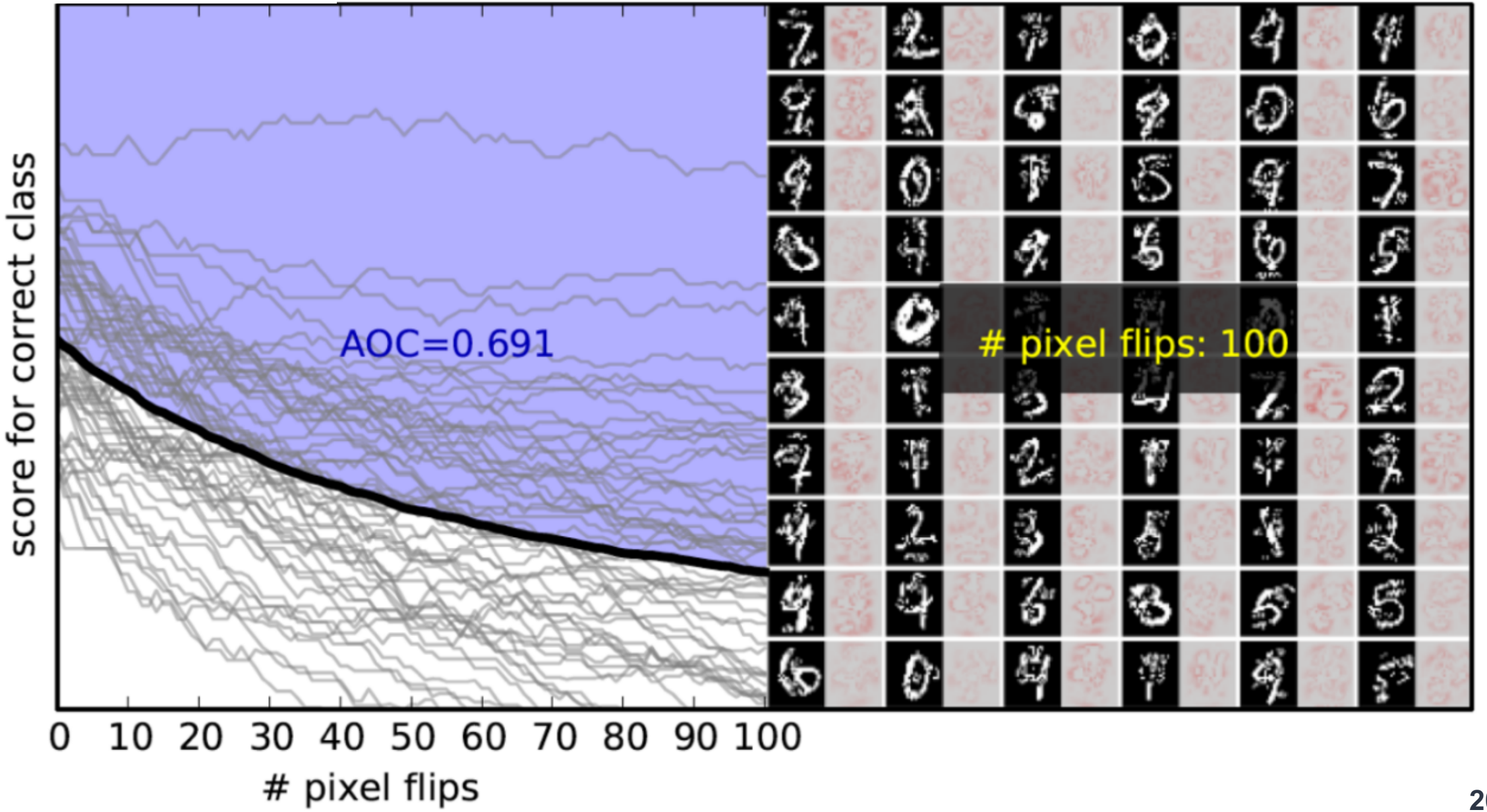
# Selectivity on Saliency Map



# Selectivity on Saliency Map



# Selectivity on Saliency Map



# Evaluating Attribution Methods

**Coherence** | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**

- Sensitivity may not be accurate
- Class probability may decrease because the DNN has never seen such image

**Remove and Retrain** (ROAR) / **Keep and Retrain** (KAR)

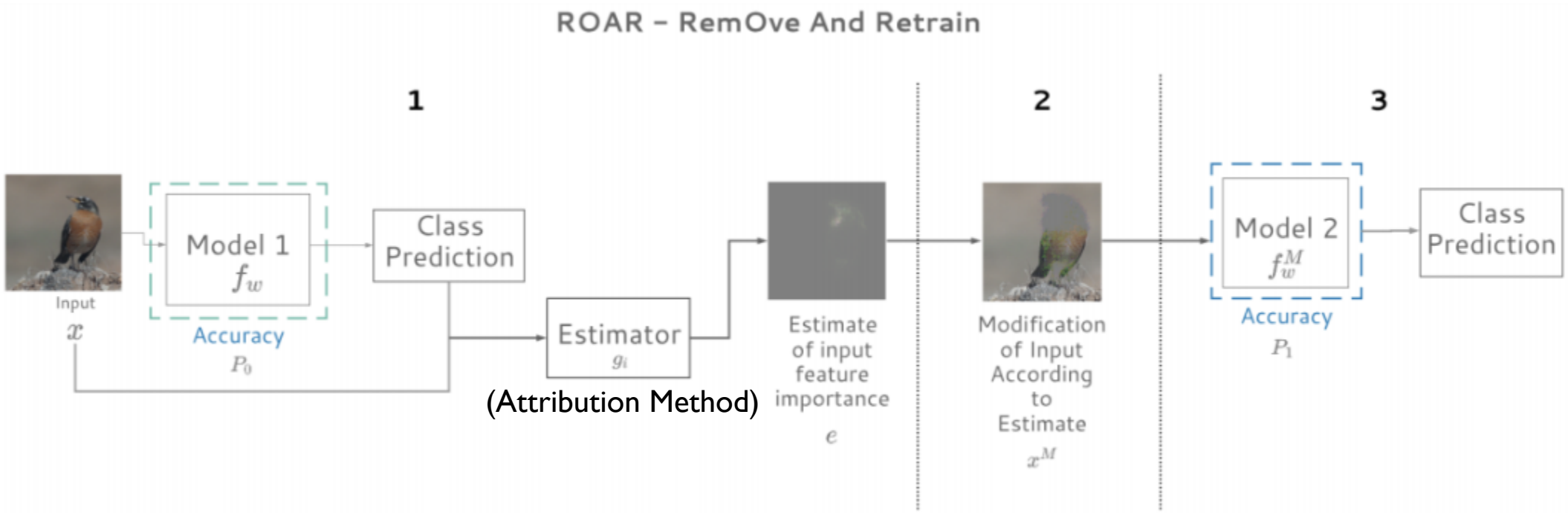
Measure how the performance of the classifier changes as features are removed based on the attribution method

- ROAR: replace $N\%$ of pixels estimated to be *most* important
- KAR: replace $N\%$ of pixels estimated to be *least* important
- Retrain DNN and measure change in test accuracy

"Evaluating Feature Importance Estimates", https://arxiv.org/pdf/1806.10758.pdf

# Evaluating Attribution Methods

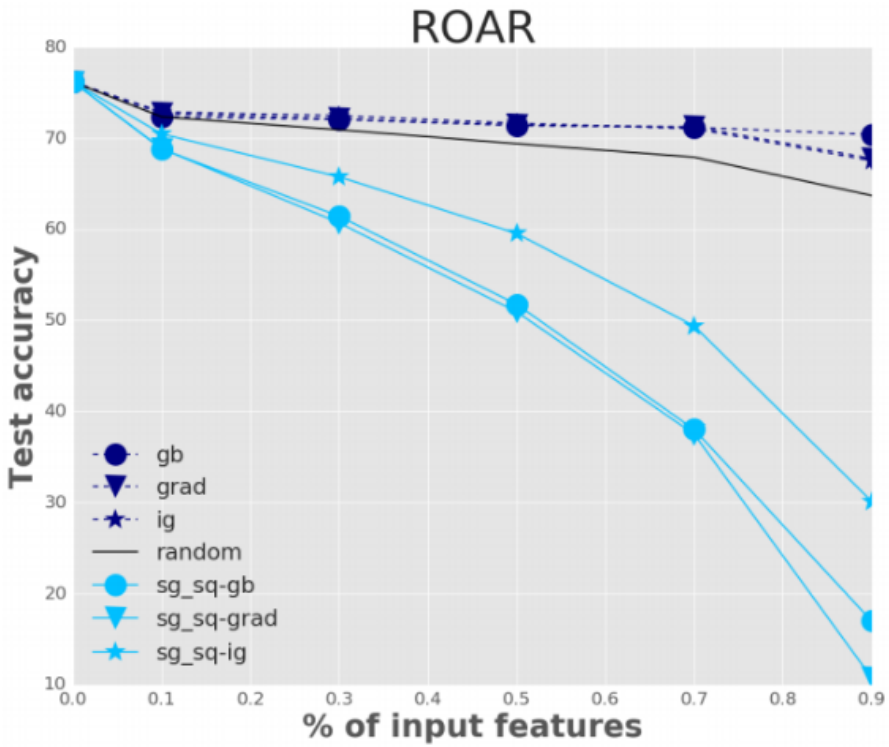
**Coherence** | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**

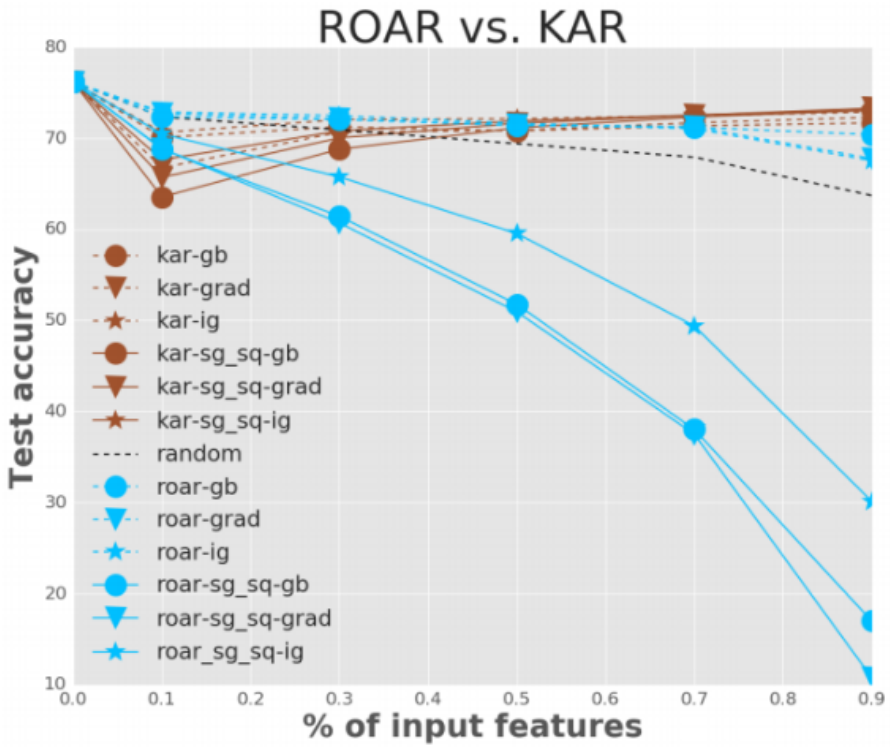



"Evaluating Feature Importance Estimates", https://arxiv.org/pdf/1806.10758.pdf

# Evaluating Attribution Methods

**Coherence** | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**




"Evaluating Feature Importance Estimates", https://arxiv.org/pdf/1806.10758.pdf

# Summary

**1. Qualitative: Coherence**

- Attributions should highlight discriminative features / objects of interest

**2. Qualitative: Class Sensitivity**

- Attributions should be sensitive to class labels

**3. Quantitative: Sensitivity**

- Removing feature with high attribution should cause large decrease in class probability

**4. Quantitative: ROAR & KAR**

- Problem: class probability may decrease because the DNN has never seen such image
- Solution: remove pixels, retrain and measure drop in accuracy

# 03 Grading Rules

# Grading Rules

- Reproduce explaining deep model via attribution methods
  - Implement at least two evaluation methods for MNIST, CIFAR-10 dataset
  - Compare each score and explain the results

# 04 References

# References

- Attribution Methods
  
  - Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." ECCV. 2014.
  - Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." ICML Workshop. 2017.
  - Fong, Ruth C., and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation." CVPR. 2017.
  - Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." CVPR. 2017.
  - Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu. "Interpretable convolutional neural networks." CVPR. 2018.
  - Wagner, Jorg, et al. "**Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks**." CVPR. 2019.
  

# References

- Evaluations
  
  - Ancona, Marco, et al. "Towards better understanding of gradient-based attribution methods for deep neural networks." ICLR. 2018.
  - Hooker, Sara, et al. "Evaluating feature importance estimates." ICML Workshop. 2018.
  - Nie, Weili, Yang Zhang, and Ankit Patel. "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations." ICML. 2018.
  - Adebayo, Julius, et al. "Sanity checks for saliency maps." NIPS. 2018.
  - Yang, Mengjiao, and Been Kim. "BIM: Towards Quantitative Evaluation of Interpretability Methods with Ground Truth." arXiv preprint arXiv:1907.09701 (2019).
  

주식회사 에스아이에이
SI ANALYTICS

Thank you for attention!

**SI Analytics Co., Ltd.** (Satrec Initiative Group)
441Expo-ro, Yuseong-gu, Daejeon, 34051, Korea

tgjeon@si-analytics.ai
+82-42-330-6800 / +82-10-6584-7321

www.si-analytics.ai