# Feature importance explanations

CSEP 590B: Explainable AI

Ian Covert & Su-In Lee
University of Washington

# Course announcements

- HW0 due last night
- HW1 posted today, due in three weeks (Monday April 25th, 11:59pm)

# About HW0

- **Q:** Why was there so much math?
- **A:** We want you to truly understand and become educated users/developers of XAI tools

## What to expect

- HW1 will have more math, including the application of ideas from HW0
- Later assignments (HW2, HW3) will focus more on programming

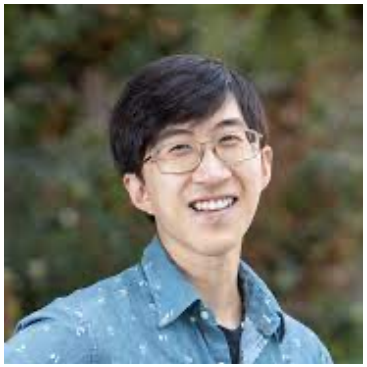
# Office hours

- Our office hours times and locations are now posted on Ed
- If these times don’t work for you, please let us know

# Office hours

- **Su-In Lee**
  - Paul G. Allen Prof.
  - UW Allen School
  - **OH: Thu 5:00-6:00 pm @ Zoom**

- **Ian Covert**
  - Ph.D. candidate
  - UW Allen School
  - **OH: Mon 5:00-6:00 pm @ Zoom**



- **Hugh Chen**
  - Ph.D. student
  - UW Allen School
  - **OH: Tue 5:30-6:30 pm @ Gates 131**

- **Chris Lin**
  - Ph.D. student
  - UW Allen School
  - **OH: Fri 4:30-5:30 pm @ Zoom**

# Machine learning resources

- We added a list of ML resources to the course website
  - Andrew Ng's lecture notes from Stanford CS 229
  - Kevin Jamieson's course website for UW CSE 546
  - The Elements of Statistical Learning
  - Computer Age Statistical Inference

# Thank you, discussion leaders!

| Lecture | Volunteer 1 | Volunteer 2 | Topic |
|---|---|---|---|
| 1 | Hugh Chen | Ian Covert | Introduction |
| 2 | | | Removal-based explanations |
| 3 | | | Shapley values |
| 4 | | | Propagation explanations |
| 5 | | | Evaluating interpretability |
| 6 | | | Inherently interpretable models |
| 7 | | | Concept explanations |
| | | | Granular interpretation |
| 8 | | | Instance explanations |
| 9 | | | Human-AI collaboration |
| 10 | | | Model improvement, applications in industry |

# Feature importance explanations

CSEP 590B: Explainable AI

Ian Covert & Su-In Lee
University of Washington

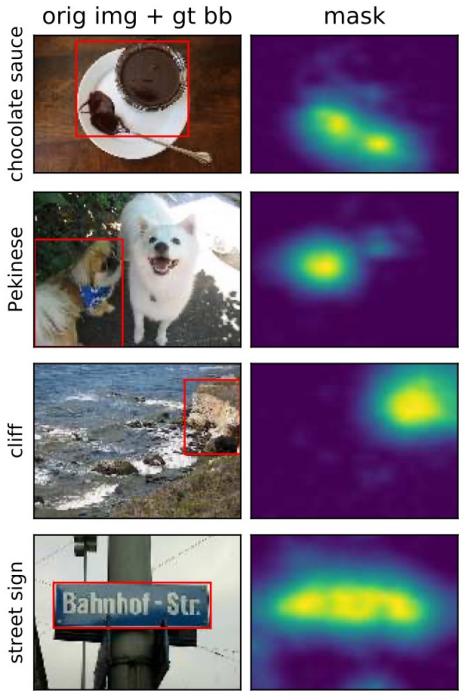
# From last time

- You trained a ML model
- People will ask questions about how it works
  - **Engineer:** *why did we get this prediction wrong?*
  - **Domain expert:** *how did the model come to this conclusion?*
  - **User:** *why was my loan request denied?*

# Feature importance explanations

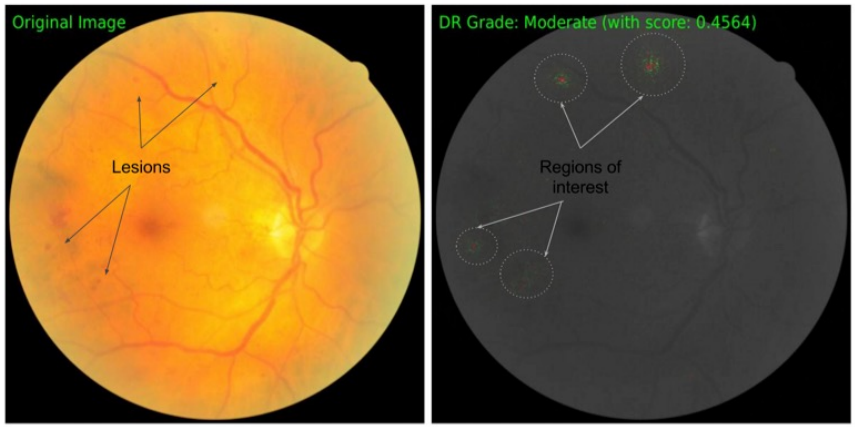
- Many ways to answer such questions
- Feature importance perspective: **how does each feature affect the model?**
- This is what many people associate with XAI

# Example 1: natural images



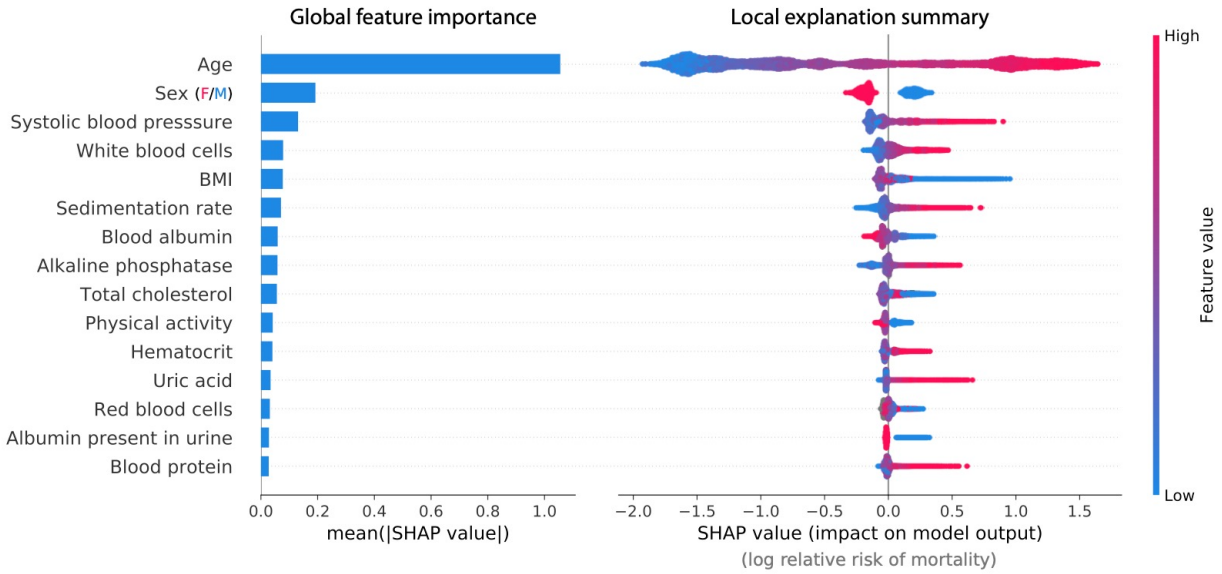

Fong & Vedaldi, "Interpretable explanations of black boxes by meaningful perturbations" (2017)

# Example 2: medical images



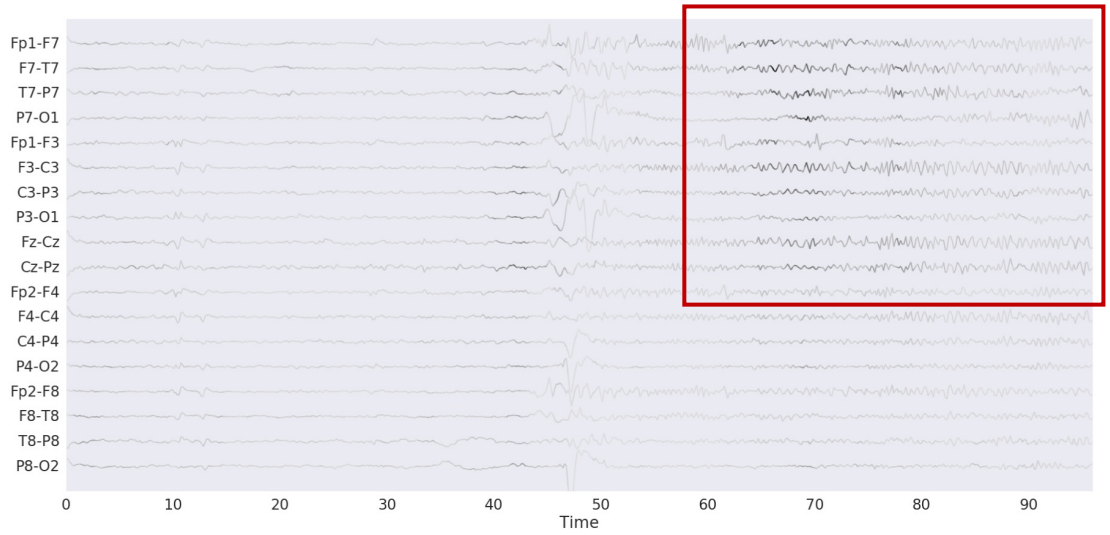

Sundararajan et al., "Axiomatic attribution for deep networks" (2017)

# Example 3: tabular data




Lundberg et al., "Explainable AI for trees: from local explanations to global understanding" (2019)

# Example 4: time series




Covert et al., "Temporal graph convolutional networks for automatic seizure detection" (2019)

# Our teaching approach

- There are too many methods to cover (100+ papers)
- We'll prioritize well-known methods
  - E.g., SHAP, LIME, IntGrad, GradCAM
- Focus on shared principles
  - Tools/perspectives to understand many methods

# Split into three lectures

1. **Removal-based explanations**
   - Analyze impact of removing information
2. **Shapley values**
   - Requires extra game theory background
3. **Propagation-based explanations**
   - Analyze model's sensitivity to small changes

# Split into three lectures

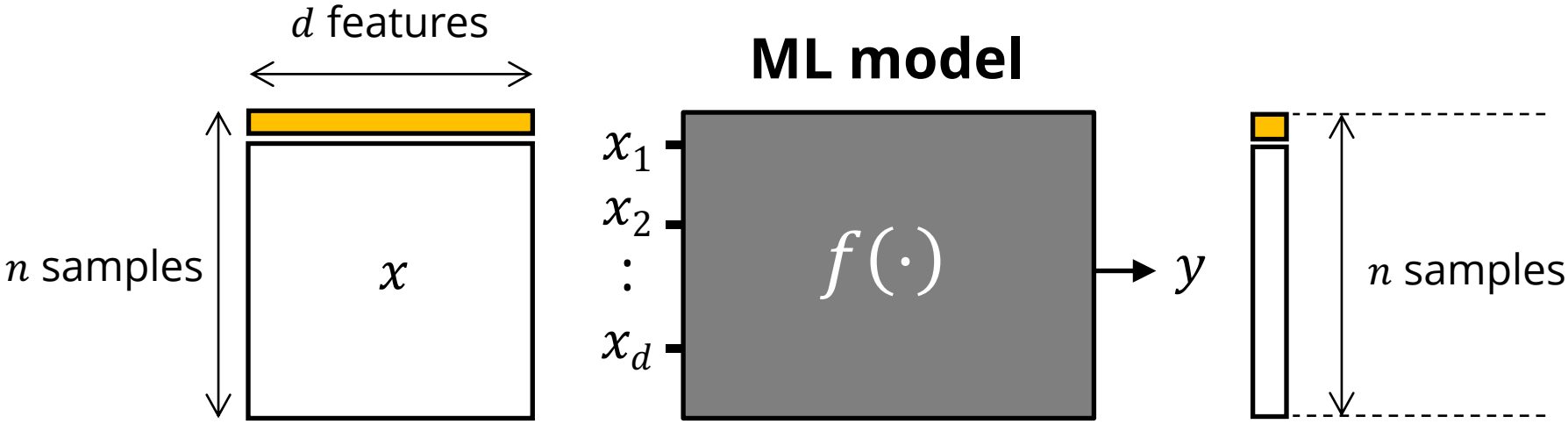
1. **Removal-based explanations** **(today)**
   - Analyze impact of removing information
2. **Shapley values** **(next lecture)**
   - Requires extra game theory background
3. **Propagation-based explanations** **(after that)**
   - Analyze model's sensitivity to small changes

# Today

- Section 1
  - Notation and definitions 
  - Removal-based explanations
  - Example methods
- Section 2
  - Unified framework for removal-based methods
  - (Bonus) Meaningful perturbations

# Notation

- Input features $x \in \mathcal{X}$, response variable $y \in \mathcal{Y}$
  - Total of $d$ features, or $x = (x_1, \ldots, x_d)$
- Predictive model $f: \mathcal{X} \mapsto \mathcal{Y}$
  - For classification, $f_y(x)$ is probability for class $y$



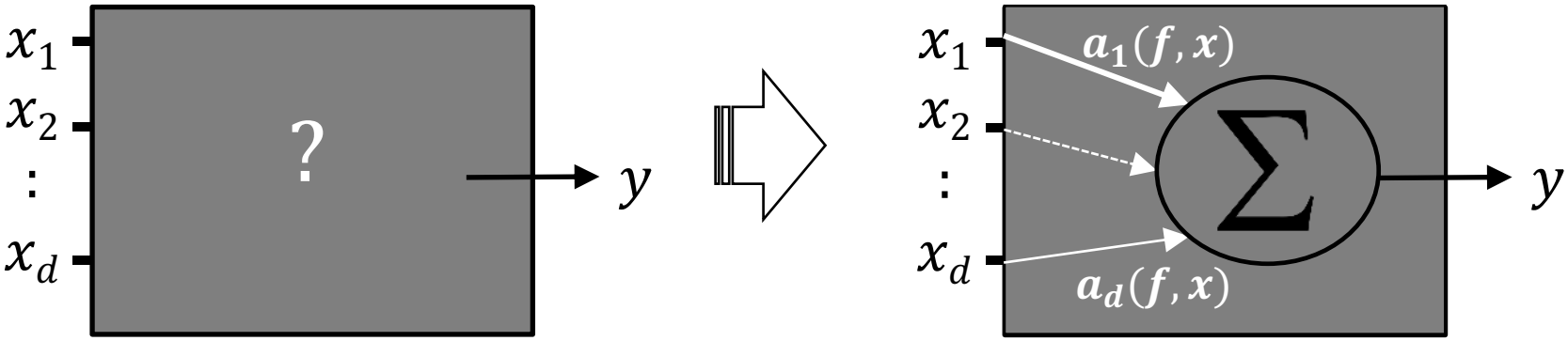
# Notation (cont.)

- For training, use loss function $\ell(\hat{y}, y)$
  - Log-loss for classification, MSE for regression

- Subsets $S \subseteq D = \{1, \ldots, d\}$, complement $\bar{S} = D \setminus S$
- Feature subsets $x_S = \{x_i : i \in S\}$

# Definitions (1)

- A **model explanation** attempts to highlight why a model made a prediction
- A **feature importance explanation** focuses on each *feature's* role
- Explanations may relate to an individual prediction (**local**) or a broader model behavior (**global**)



# Definitions (2)

- Feature importance explanations are typically either **feature attribution** or **feature selection**
  - **Feature attribution:** each feature $x_i$ receives a score $a_i \in \mathbb{R}$
  - **Feature selection:** a subset of important features $x_S \subseteq \{x_1, \ldots, x_d\}$

# Definitions (3)

- An **explanation algorithm** is a method that generates explanations given input data and an ML model

# Today

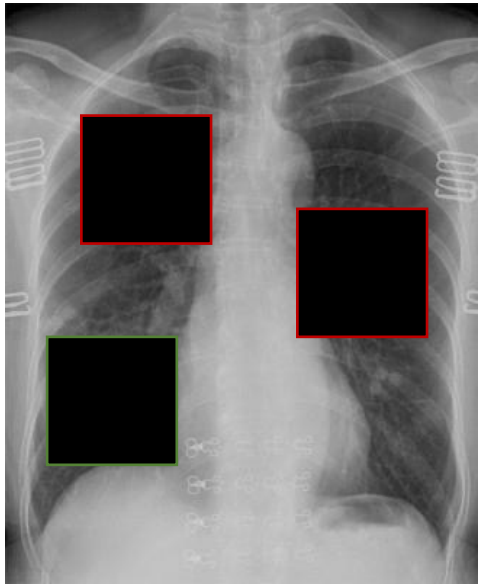
- Section 1
  - Notation and definitions
  - Removal-based explanations 
  - Example methods
- Section 2
  - Unified framework for removal-based methods
  - (Bonus) Meaningful perturbations

# Removal-based explanations

- **Idea:** to understand a feature's importance, remove it and see how the prediction changes
- This is the underlying idea behind many popular approaches

# Doctor analogy

- Suppose we want to understand a doctor's diagnosis
- We can probe the doctor's reasoning by covering parts of the scan
- The diagnosis should change when we cover important regions



# Translation to ML

- We can analyze models by withholding features
  - Understand each feature's influence via the impact of removing it
  - Removing important features creates large changes
  - The directionality of the change matters

# Today

- Section 1
  - Notation and definitions
  - Removal-based explanations
  - Example methods 
- Section 2
  - Unified framework for removal-based methods
  - (Bonus) Meaningful perturbations

# Case study I: permutation tests

- An “old” method introduced for random forests
- Determines overall (global) importance of each input feature

Leo Breiman, “Random forests” (2001)

# Permutation test procedure

- First, evaluate the model's accuracy using the original data
- Then, one at a time, corrupt features and record the drop in accuracy
  - To *corrupt*, randomize/permute a column of the dataset (corresponding to the feature)
  - Hence, **permutation** test

# Mathematical definition

- Intuitive view

$$a_i = Acc(\text{original}) - Acc(x_i \text{ corrupted})$$

# Mathematical definition

- Detailed view

Arbitrary loss function

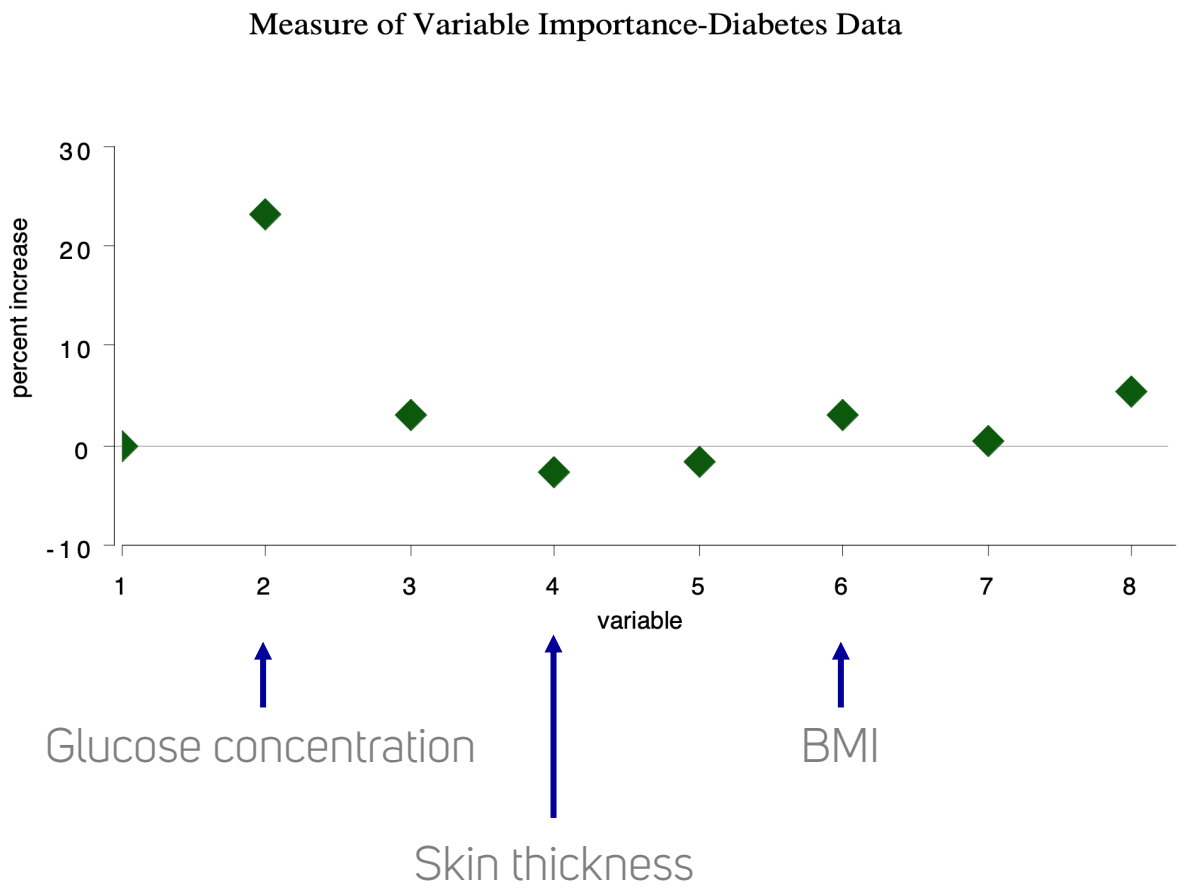
$$a_i = \frac{1}{n}\sum_{j=1}^{n} \ell\left(f\left(x_1^j, \ldots, \tilde{x}_i^j, \ldots, x_d^j\right), y^j\right) - \frac{1}{n}\sum_{j=1}^{n} \ell\left(f\left(x_1^j, \ldots, x_d^j\right), y^j\right)$$

Mean loss

Corrupted feature
(should increase loss)

Original data

# Example (Breiman, 2001)



# Remarks

- Permutation tests work for any model
- Can use with continuous or categorical features
- Fast, easy to implement

# Case study II: occlusion

- An early approach for deep learning models
- Explain individual predictions for image classifiers
  - Calculates pixel (or superpixel) importance

Zeiler & Fergus, "Visualizing and understanding convolutional networks" (2014)

# Occlusion procedure

- Make prediction given full image
- Occlude various image regions, record how the prediction changes
  - *Occlude* by replacing with uninformative (zero) pixels
  - Potentially occlude 2x2, 4x4, etc. superpixels

# Mathematical definition

- Intuitive view

$$a_i = f_y(x) - f_y\left(x_{(-i)}\right)$$

# Mathematical definition

- Detailed view

$$a_i = f_y(x_1, \ldots, x_d) - f_y(x_1, \ldots, 0, \ldots, x_d)$$

Replace with zero

# Example (Ancona et al. 2018)

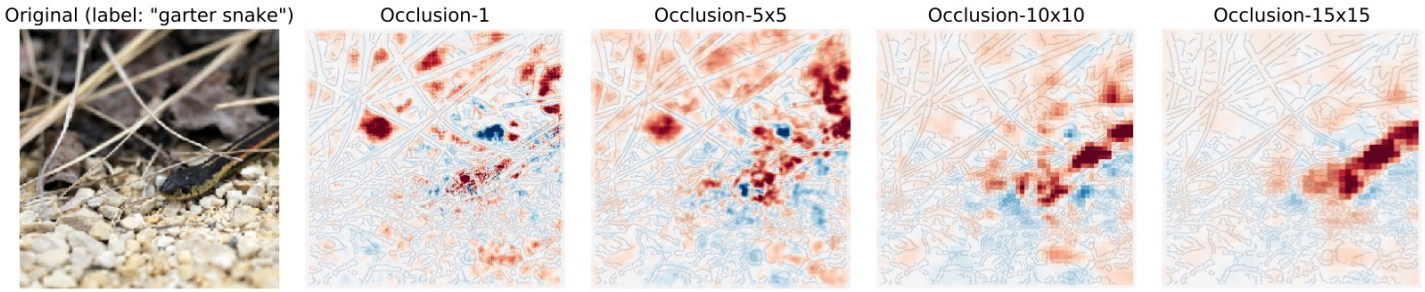



Figure 1: Attributions generated by occluding portions of the input image with squared grey patches of different sizes. Notice how the size of the patches influence the result, with focus on the main subject only when using bigger patches.

# Remarks

- The occlusion idea works with any model, even non-image data
- Moderately fast: $d + 1$ model evaluations to explain each prediction
- Simple, easy to implement

# Perm. tests vs. occlusion

- Seemingly different methods
  - Designed for random forests vs. CNNs
  - Global vs. local explanations
- However, some notable parallels

# Today

- Section 1
  - Notation and definitions
  - Removal-based explanations
  - Example methods

  **10 min break**
- Section 2
  - Unified framework for removal-based methods
  - (Bonus) Meaningful perturbations