Explainable AI (XAI)

11A

CS 3244
Machine Learning

NUS National University of Singapore | Computing

Week 11B: Learning Outcomes

1. Describe multiple methods to interpret **feature importance**
2. Appropriately **interpret** feature attributions from each type of explanation
3. Describe how LIME explanations are generated
4. Describe how Grad-CAM explanations are generated

Week 11B: Lecture Outline

1. Introduction
   1. Motivation for Explainable AI (XAI)
   2. Explaining Why: Feature Importance
2. Explanation techniques
   1. Glassbox Models (Linear Regression, Logistic Regression)
   2. Model-Agnostic Explanations (LIME)
   3. Model-Specific Explanations (Grad-CAM)

Case 1:
Does patient have cancer?

Feature	Value
worst area	1315.00
mean radius	16.13
worst radius	20.96
area error	54.18
worst perimeter	136.80
worst texture	31.48
mean perimeter	108.10
smoothness error	0.01
mean area	798.80
mean concave points	0.10

Prediction: Cancer

Further reading: https://coderzcolumn.com/tutorials/machine-learning/how-to-use-lime-to-understand-sklearn-models-predictions

Image credit

Why??

Explanation: Feature Attributions

Case 2:

Is this skin cancer?

Prediction: Skin Cancer

Why??

Explanation: Highlighted Salient Region

Further reading: https://towardsdatascience.com/medical-image-analysis-using-probabilistic-layers-and-grad-cam-42cc0118711f
Image credit: https://news.yale.edu/2019/11/13/yale-study-reveals-hyperhotspots-identifying-skin-cancer-risk

Feature Importance

- Explains
  - Which features are **important** for the prediction
  - In what way the features **influenced** the prediction
- Implementation
  - **Weights** in Linear / Logistic Regression
  - **Surrogate Weights** from LIME
  - **Saliency Maps** of CNN

Interpreting

Linear Regression

NUS National University of Singapore | Department of Computer Science School of Computing

How would you interpret?

Linear Regression

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \sum_{r=0}^{n} w_r x_r$$
$$= \boldsymbol{w} \cdot \boldsymbol{x} = \boldsymbol{w}^\top \boldsymbol{x}$$

How would you interpret?

Linear Regression

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \sum_{r=0}^{n} w_r x_r$$
$$= \boldsymbol{w} \cdot \boldsymbol{x} = \boldsymbol{w}^\top \boldsymbol{x}$$

<u>Weighted Sum Interpretation</u>

Bigger w_r means

- **Larger** weight

- More **importance** for to x_r
- Direction? Supportive (positive) or opposing (negative) **influence**

<u>Gradient Interpretation</u>

Bigger w_r means

- **Steeper** slope for x_r axis
  - Changes in x_r lead to bigger in $\hat{y}$ changes
- More **importance** for x_r
- Direction indicates increasing or decreasing **influence**

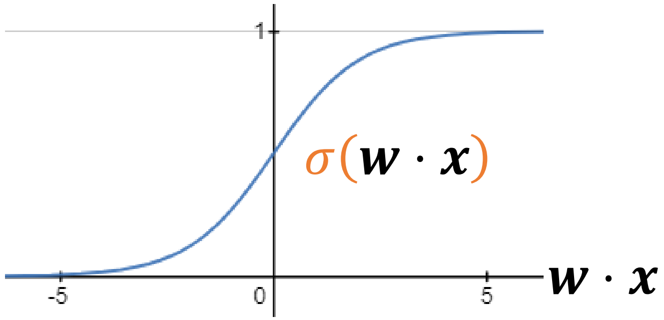
Interpreting
Logistic Regression

NUS National University of Singapore | Department of Computer Science
School of Computing

How would you interpret?
Logistic Regression

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

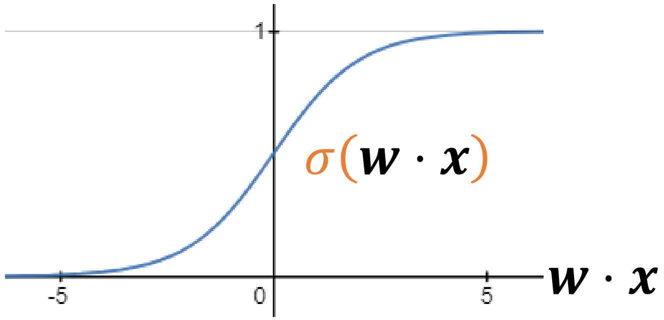
$$z = \boldsymbol{w} \cdot \boldsymbol{x} = \sum_{r=0}^{n} w_r x_r$$



How would you interpret?
Logistic Regression

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$z = \boldsymbol{w} \cdot \boldsymbol{x} = \sum_{r=0}^{n} w_r x_r$$



<u>Weighted Sum Interpretation</u>

Bigger w_r means
- **Larger** importance
- Direction indicates **influence**

<u>Gradient Interpretation</u>

Bigger w_r means?
- Steepness? Sigmoid bounded between 0 and 1
- Direction in 2D (or higher)?

Insert Web Page

This app allows you to insert secure web pages starting with https:// into the slide deck. Non-secure web pages are not supported for security reasons.

Please enter the URL below.

https:// www.desmos.com/calculator/h918gs69t5

Note: Many popular websites allow secure access. Please click on the preview button to ensure the web page is accessible.

Insert Web Page

This app allows you to insert secure web pages starting with https:// into the slide deck. Non-secure web pages are not supported for security reasons.

Please enter the URL below.

https:// www.desmos.com/calculator/crbuwes4ca

Note: Many popular websites allow secure access. Please click on the preview button to ensure the web page is accessible.

Insert Web Page

This app allows you to insert secure web pages starting with https:// into the slide deck. Non-secure web pages are not supported for security reasons.

Please enter the URL below.

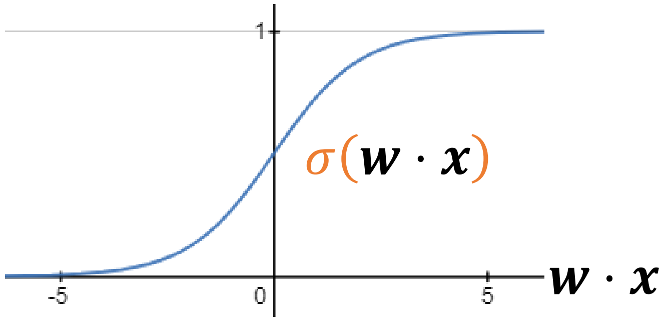
https:// www.desmos.com/calculator/dhckwf0kys

Note: Many popular websites allow secure access. Please click on the preview button to ensure the web page is accessible.

How would you interpret?
Logistic Regression

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$f = \boldsymbol{w} \cdot \boldsymbol{x} = \sum_{r=0}^{n} w_r x_r$$



Weighted Sum Interpretation	Gradient Interpretation
Bigger w_r means • **Larger** importance • Direction indicates **influence**	**Bigger** w_r means? • **Steeper** slope for x_r near decision boundary • **Decision boundary** more *perpendicular* to x_r • Weight sign indicates direction of **pos**/**neg** prediction


?
Questions!

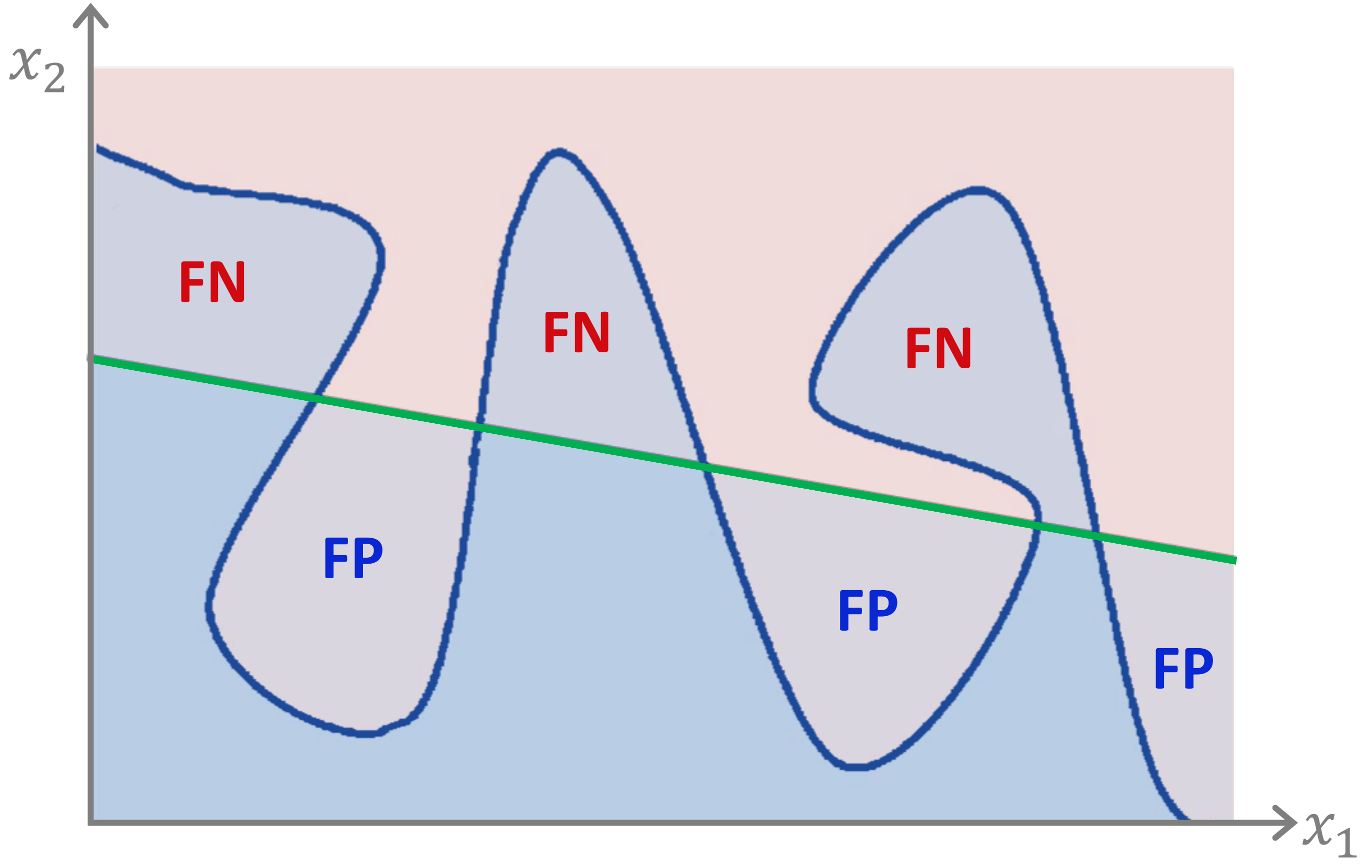
Local Interpretable Model-agnostic Explanations

LIME

NUS National University of Singapore | Department of Computer Science School of Computing

How to describe with just x_1 and x_2?

Non-Linear Decision Boundary $f(\boldsymbol{x})$



Prediction Model

$f(\boldsymbol{x})$

- Non-linear model of $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix}$
- Shown as curvy decision boundary

Explanation: **Linear Model**

$$g(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} = \sum_{r=0}^{n} w_r x_r$$

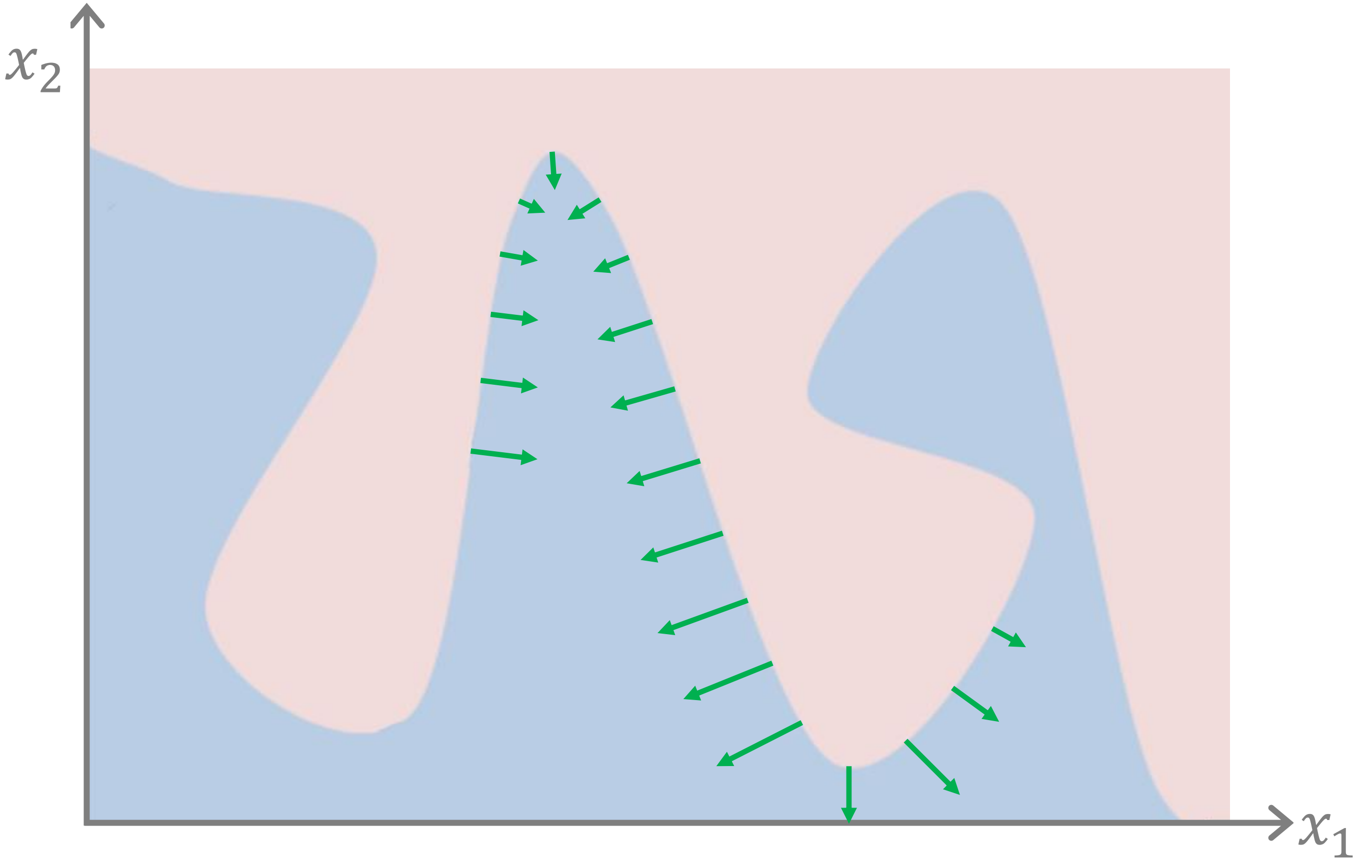
- Simple to interpret
- But too many errors between g and f

$$L = f(\boldsymbol{x}) - g(\boldsymbol{x})$$

Image Credit: https://santiagof.medium.com/model-interpretability-making-your-model-confess-lime-89db7f70a72b

How to describe with just x_1 and x_2?

Non-Linear Decision Boundary



Prediction Model

$$f(\boldsymbol{x})$$

- Non-linear model of $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix}$
- Shown as curvy decision boundary

Explanation: **Gradients**

$$g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) = \frac{df}{d\boldsymbol{x}} = \begin{pmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ \vdots \end{pmatrix}$$

- **Steepness** for each feature x_r
- Difficult to remember, since gradients are different for each instance (point)

Image Credit: https://santiagof.medium.com/model-interpretability-making-your-model-confess-lime-89db7f70a72b

LIME

Local Interpretable Model-agnostic Explanations

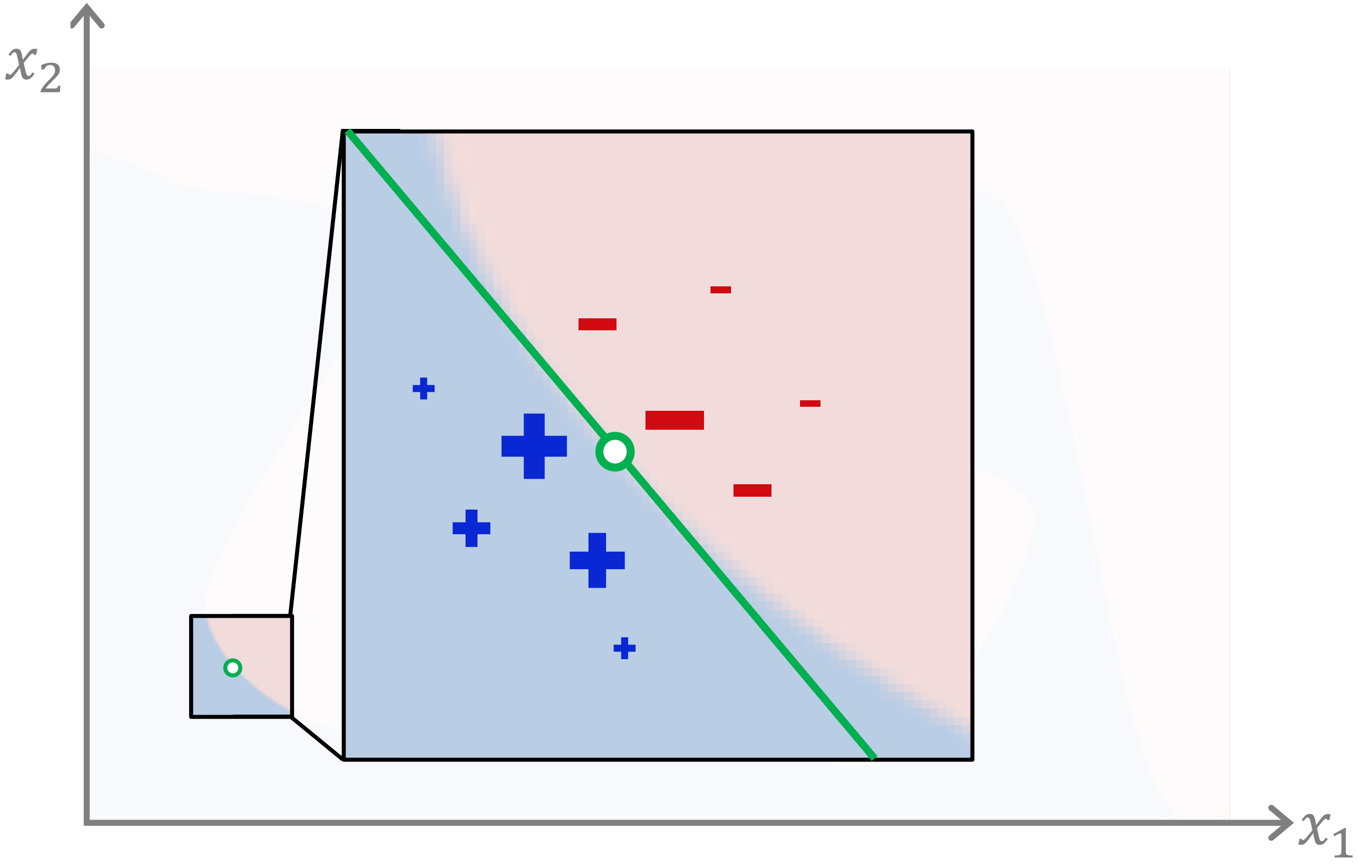


Image Credit: https://santiagof.medium.com/model-interpretability-making-your-model-confess-lime-89db7f70a72b

Prediction Model

$$f(\boldsymbol{x})$$

- Non-linear model of $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \end{pmatrix}$
- Shown as curvy decision boundary

Explanation: LIME

1. Starting with **instance** $\boldsymbol{x}$ to explain
2. Focus on **Local** region
3. Training set as **neighbors** $\boldsymbol{x}^{\langle \eta \rangle} \in X^{\langle \eta \rangle}$
4. Train **surrogate model**, e.g., linear:

$$g(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} = \sum_{r=0}^{n} w_r x_r$$

LIME

Find "best" explainer g that minimizes $\xi(\boldsymbol{x})$

Python API:
https://github.com/marcotcr/lime

$$\underset{g \in G}{\operatorname{argmin}}\Big(\xi(\boldsymbol{x}) = \overbrace{L(f, g, \pi_x)}^{\text{"Faithful"}} + \overbrace{\Omega(g)}^{\text{"Simple"}}\Big)$$

***Locally-weighted* error loss function**

$$L(f, g, \pi_x) = \sum_{\boldsymbol{x}^{\langle\eta\rangle} \in X^{\langle\eta\rangle}} \pi_x\left(\boldsymbol{x}^{\langle\eta\rangle}\right)\left(f\left(\boldsymbol{x}^{\langle\eta\rangle}\right) - g\left(\boldsymbol{x}^{\langle\eta\rangle}\right)\right)^2$$

Neighbor **proximity** function e.g., $e^{-\left(d\left(\boldsymbol{x}, \boldsymbol{x}^{\langle\eta\rangle}\right)\right)^2}$

Predictor model

Explainer model

Sparsity regularization

$$\Omega(g) = \|\boldsymbol{w}\|_1 = \sum_{r=1}^{n} |w_r|$$

- Want simpler explanation
  - ⇒ *fewer weights*
  - ⇒ L1 norm (LASSO)
- Penalizes large total weights

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. *KDD'16*.

[Image credit](#)

Case 1: Does patient have cancer?

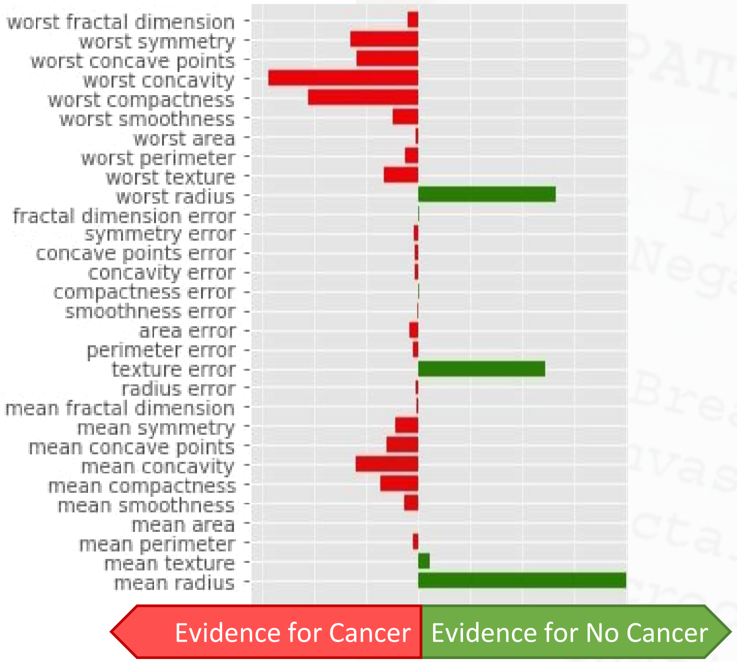
Why do the two set of weights **differ**?

Instance $\boldsymbol{x}$

Feature	Value
worst area	1315.00
mean radius	16.13
worst radius	20.96
area error	54.18
worst perimeter	136.80
worst texture	31.48
mean perimeter	108.10
smoothness error	0.01
mean area	798.80
mean concave points	0.10

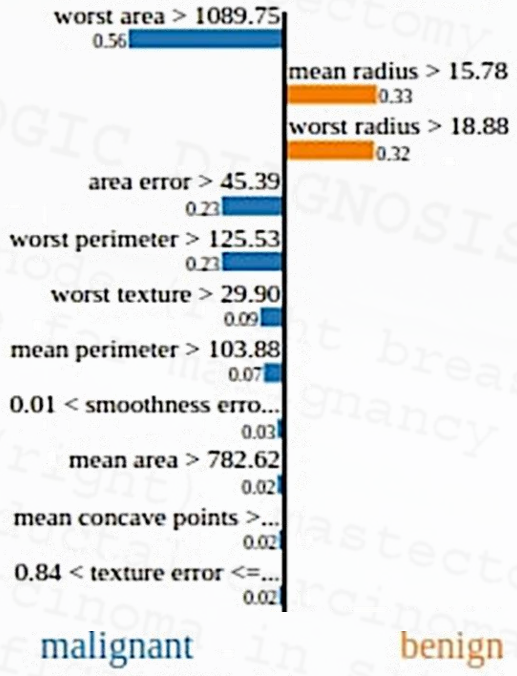
Prediction $\hat{y}$ = Cancer

Logistic Regression



Weights $\boldsymbol{w}$ of surrogate explanation f

LIME



Weights $\boldsymbol{w}$ of surrogate explanation g

Further reading: https://coderzcolumn.com/tutorials/machine-learning/how-to-use-lime-to-understand-sklearn-models-predictions


?
Questions!

Gradient-weighted Class Activation Mapping

Grad-CAM

NUS National University of Singapore
Department of Computer Science
School of Computing

Explaining Image Predictions

- LIME to explain image prediction?
- What are the input features?
    - Feature = Pixels?
        - Too many features
    - Need “super pixels”

- Another way: Attribution → Saliency Map
    - Feature = Activation Map
    - Grad-CAM

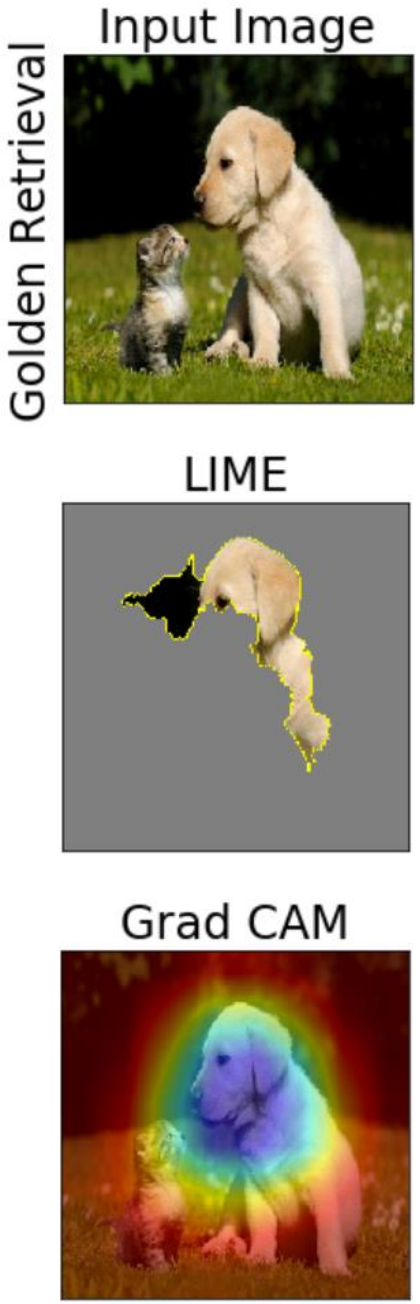
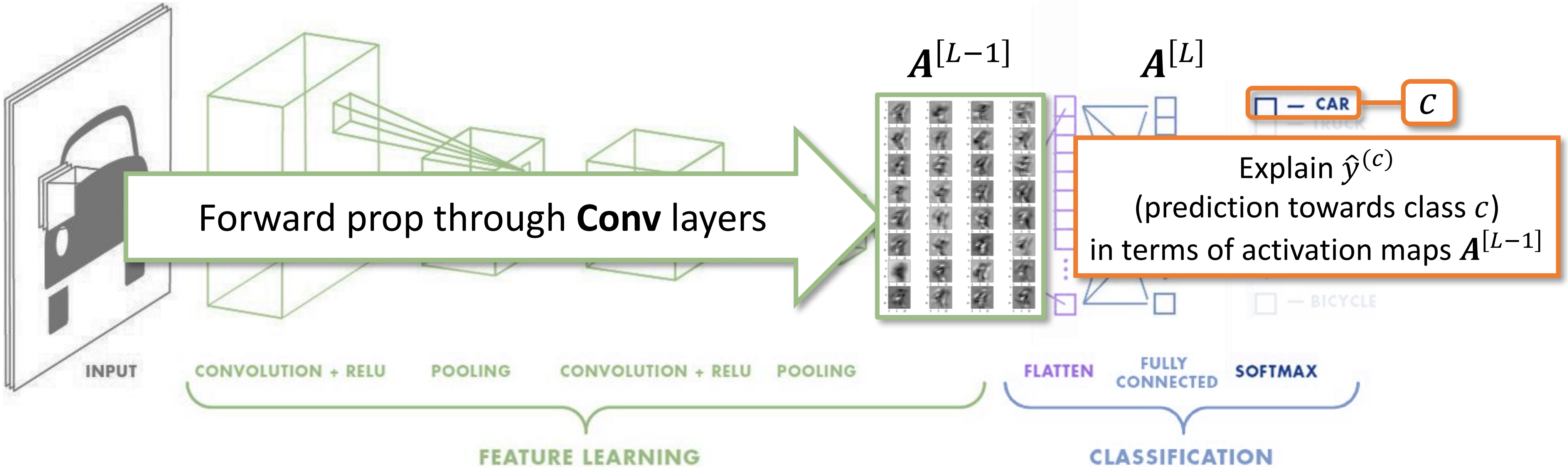


Image credit: https://arxiv.org/pdf/1908.04389.pdf

Convolutional Neural Network



Key concepts

❶ Learn Spatial Feature
- Series of multiple convolution + pooling layers
- Progressively learn more diverse and higher-level features

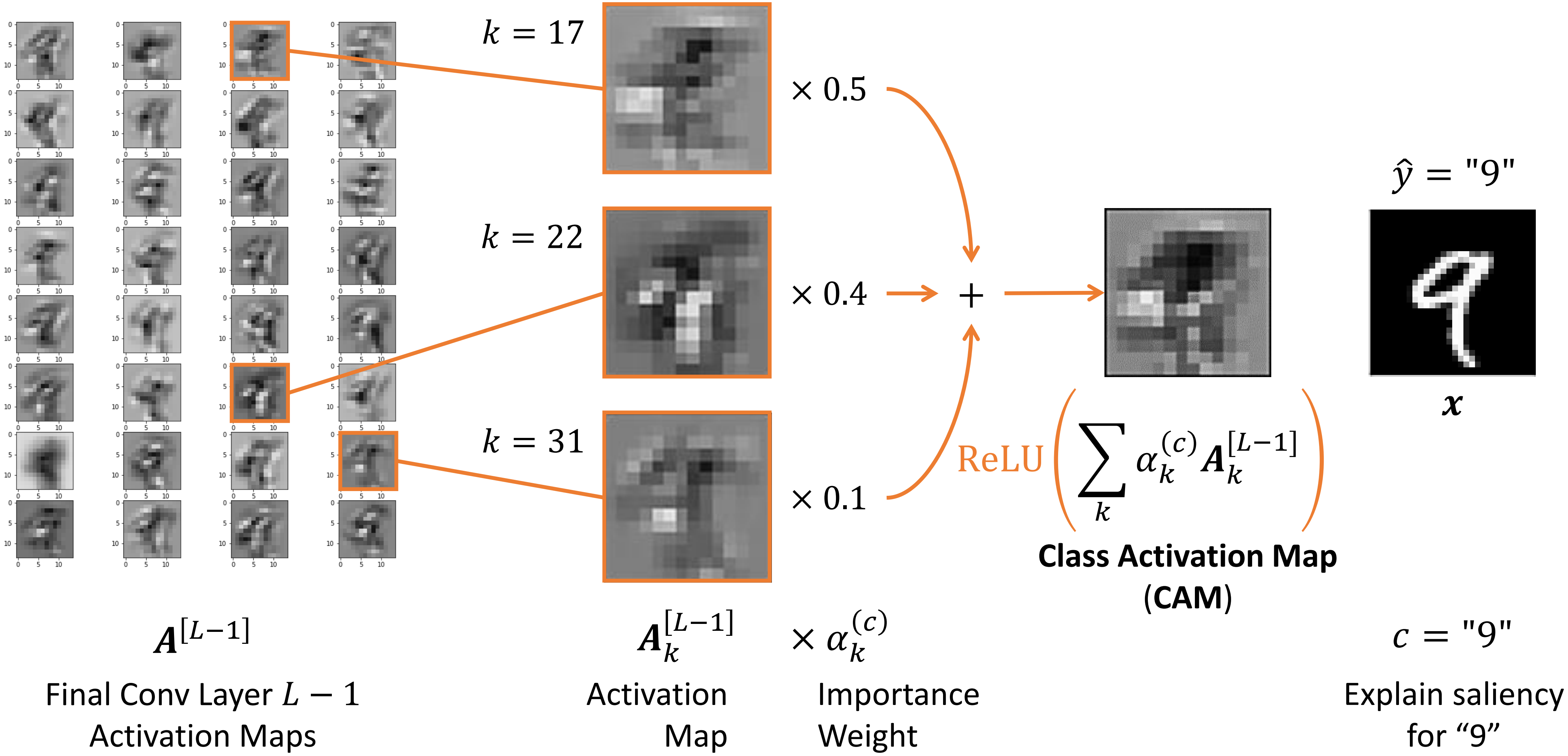
❷ Flattening
- Convert to fixed-length 1D vector

❸ Learn Nonlinear Features
- With fully connected layers (regular neurons)
- Learns nonlinear relations with multiple layers

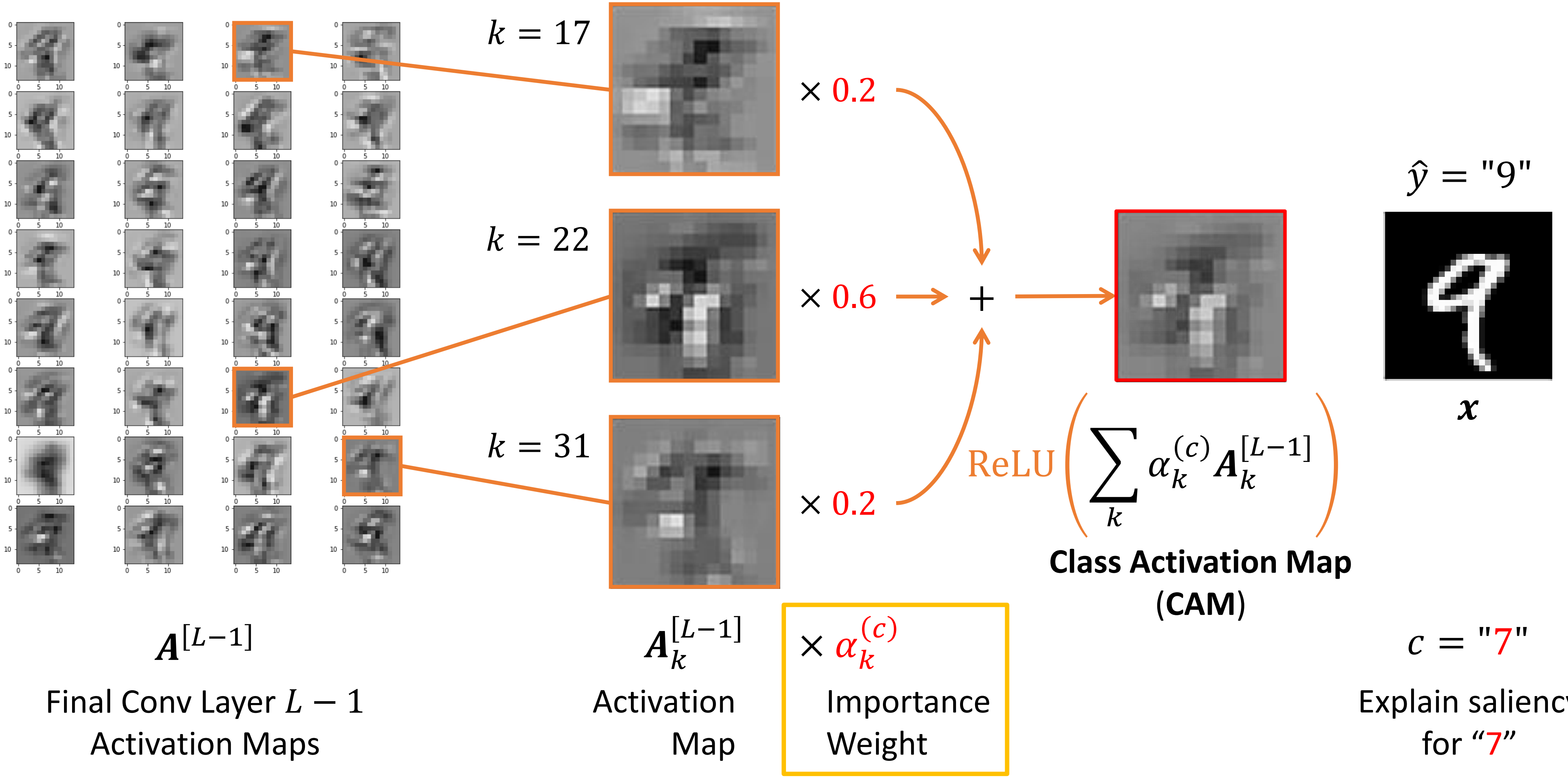
❹ Classification
- Softmax ≔ Multiclass Logistic Regression
- Feature input = image embedding vector (typically large vector)

Image credit: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

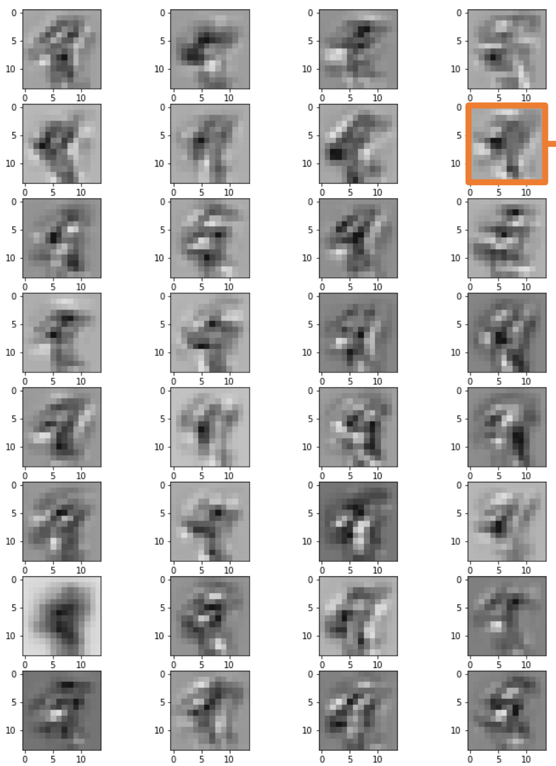
Grad-CAM Example: Why did the CNN predict “9”?



Grad-CAM Example: Why did the CNN predict "7"?



Importance Weight $\alpha_k^{(c)}$ of Activation Map



$\boldsymbol{A}^{[L-1]}$

Final Conv Layer $L-1$
Activation Maps

$$\boldsymbol{A}_k^{[L-1]} = \begin{pmatrix} a_{11k}^{[L-1]} & \cdots & a_{1wk}^{[L-1]} \\ \vdots & \ddots & \vdots \\ a_{h1k}^{[L-1]} & \cdots & a_{hwk}^{[L-1]} \end{pmatrix}$$

Activation of
kth channel
at pixel (h, w)
in layer $L-1$

$$\frac{\partial \hat{y}^{(c)}}{\partial \boldsymbol{A}_k^{[L-1]}} = \begin{pmatrix} \frac{\partial \hat{y}^{(c)}}{\partial a_{11k}^{[L-1]}} & \cdots & \frac{\partial \hat{y}^{(c)}}{\partial a_{1wk}^{[L-1]}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}^{(c)}}{\partial a_{h1k}^{[L-1]}} & \cdots & \frac{\partial \hat{y}^{(c)}}{\partial a_{hwk}^{[L-1]}} \end{pmatrix}$$

$\neq \boldsymbol{W}_k^{[L-1]}$

$$\frac{\partial f^{[L-1]}}{\partial a_{hwk}^{[L-1]}} \frac{d g^{[L-1]}}{d f^{[L-1]}} \frac{\partial f^{[L]}}{\partial g^{[L-1]}} \frac{d \boldsymbol{\sigma}}{d f^{[L]}} \frac{d \hat{y}^{(c)}}{d \boldsymbol{\sigma}} =$$

Sum of *all* gradients in activation map

$$\left\| \frac{d \hat{y}^{(c)}}{d \boldsymbol{A}_k^{[L-1]}} \right\| = \sum_{ij} \frac{\partial \hat{y}^{(c)}}{\partial a_{ijk}^{[L-1]}} = \frac{\partial \hat{y}^{(c)}}{\partial a_{11k}^{[L-1]}} + \cdots + \frac{\partial \hat{y}^{(c)}}{\partial a_{hwk}^{[L-1]}} = \alpha_k^{(c)}$$

Gradient Interpretation: Steeper ⇒ More important

Grad-CAM Steps

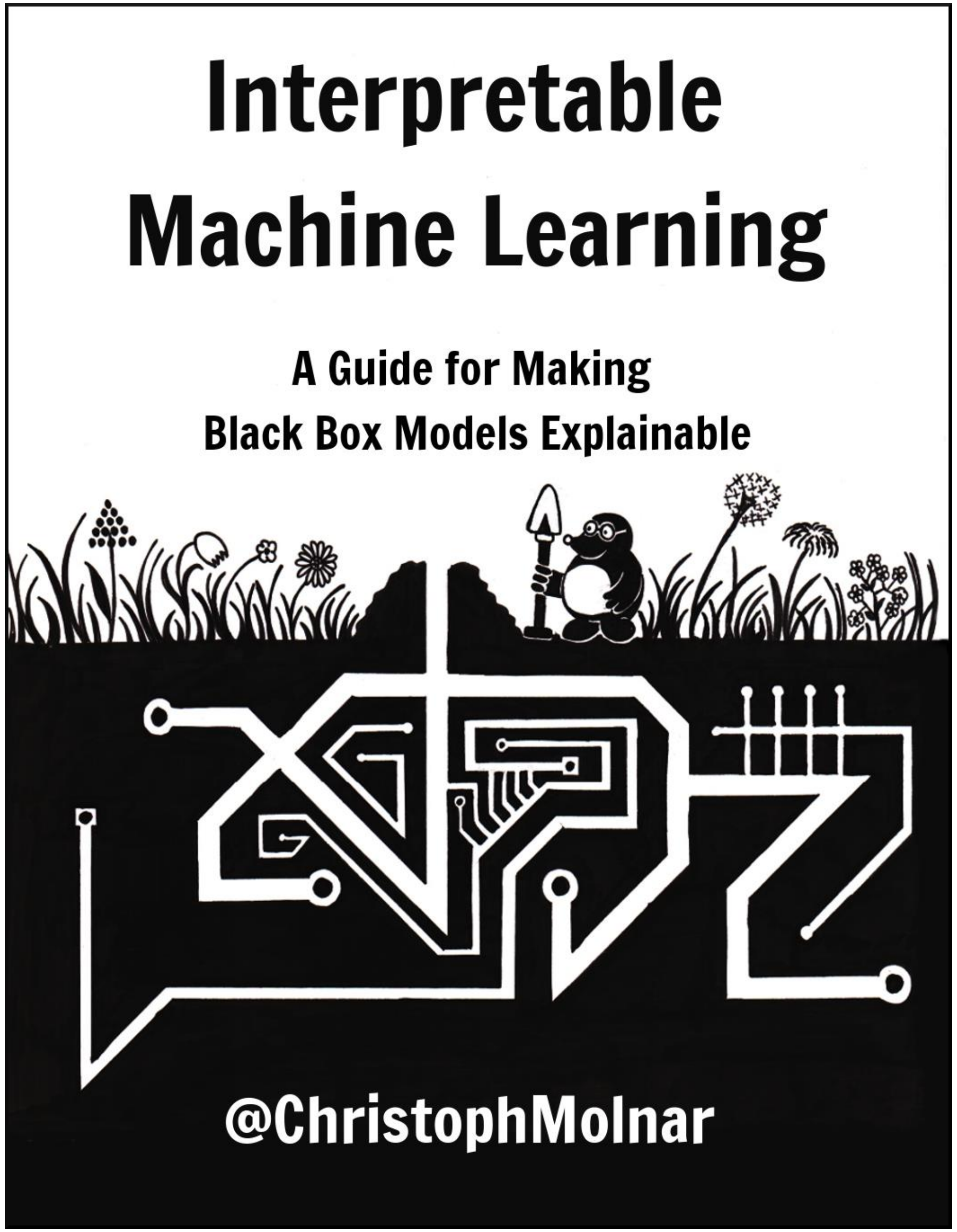
Python API:
https://github.com/jacobgil/pytorch-grad-cam

1. Compute Activation Maps $\boldsymbol{A}^{[L]}$ of last conv layer L
   1. via Forward Propagation
2. Choose class label c to explain about (e.g., predict "9", "car")
3. Filter prediction $\hat{y}$ to be about class c
   1. Given: $\hat{\boldsymbol{y}} = \begin{pmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(c)} \\ \hat{y}^{(n)} \end{pmatrix}$, $\boldsymbol{e}^{(c)} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$, then $\hat{\boldsymbol{y}}^{(c)} = \hat{\boldsymbol{y}} \circ \boldsymbol{e}^{(c)} = \begin{pmatrix} 0 \\ 0 \\ y^c \\ 0 \end{pmatrix}$
   2. To generate explanation only for that class c
4. Compute importance weight $\alpha_k^{(c)}$ for each Activation Map $\boldsymbol{A}_k^{[L]}$
   1. Backprop from $\hat{\boldsymbol{y}}^{(c)}$ to get gradients (relative to activations) at last conv layer
5. Compute weighted sum with ReLU to get **Class Activation Map**


?
Questions!

Further Reading

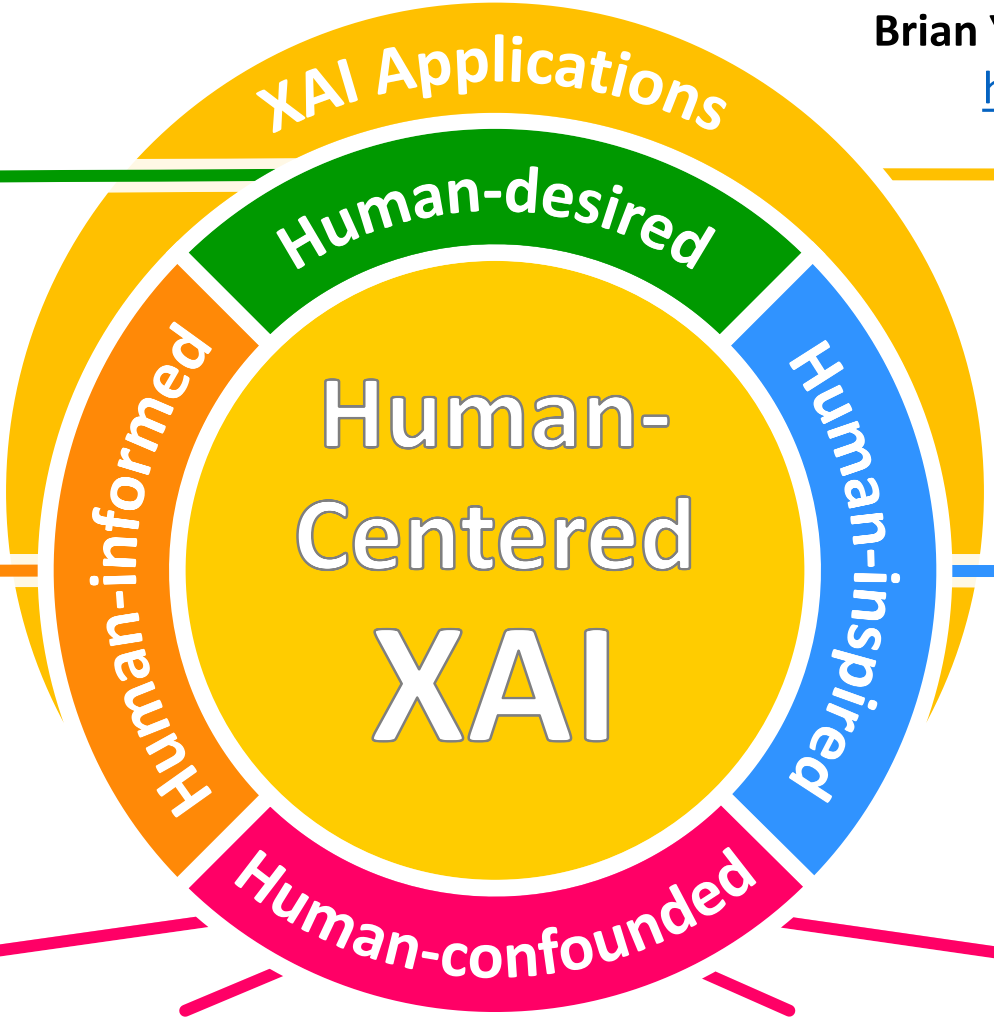
- https://christophm.github.io/interpretable-ml-book



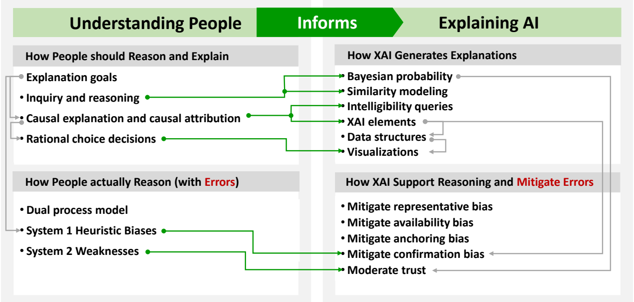
NUS Ubicomp Lab

Apps and Analytics for Smart Cities and Health

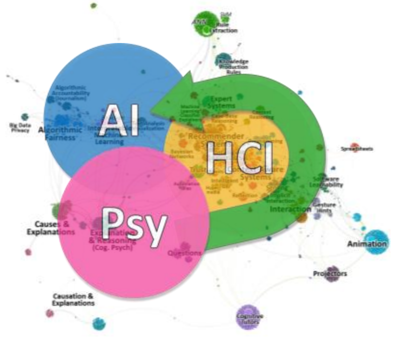
Brian Y. Lim | brianlim@comp.nus.edu.sg
https://ubiquitous.comp.nus.edu.sg



XAI Reasoning Framework [CHI'19]



XAI Gap and Trends [CHI'18]



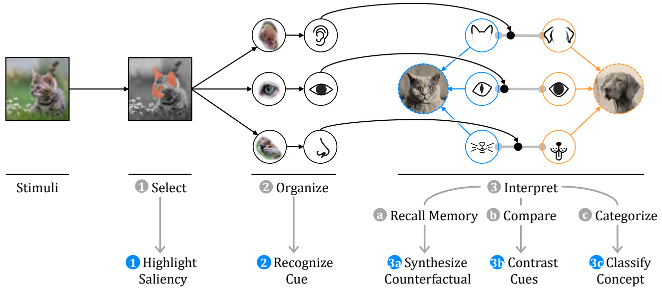
Interpretable Directed Diversity [CHI'22]

Attempt	Message	Score
1	Physical activity is good for health. Let's go for some exercise.	37%
2	Physical activity is good for health	54%
3	Physical activity is good for health. Let's walk more and reduce sitting time.	55%

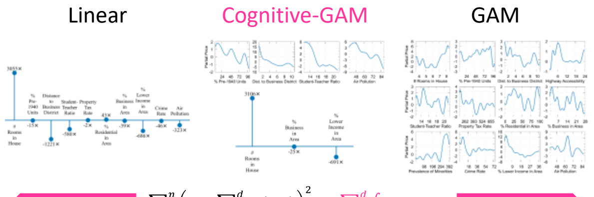
Related to	dreamlining	+1%
Is a	musical time	+2%
Has prerequisite	playing game	+1%

XAI Perceptual Process [CHI'22]

Best Paper Award

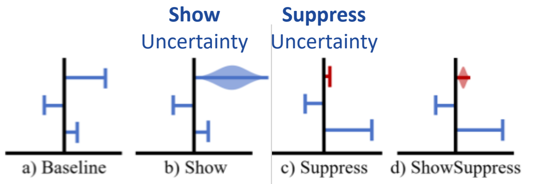


Parsimony *vs.* Performance [CHI'20]



$$\sum_j^n \left(y_i - \sum_i^d f_i\left(x_j^{(i)}\right)\right)^2 + \lambda \sum_i^d \int f_i''(t)^2\, dt$$

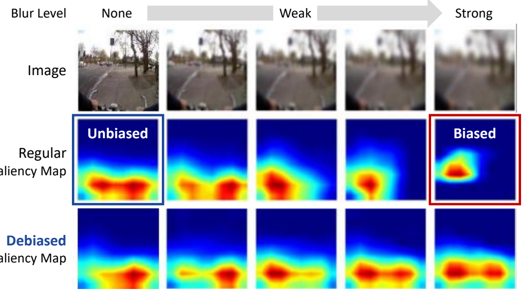
Show *or* Suppress Uncertainty [AIJ'21]



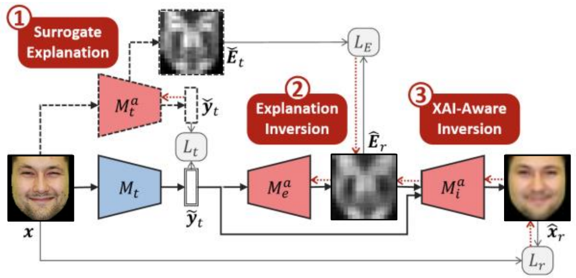
$$\bigcup_{\epsilon} \sum_{x \in s_{x_0}} c_{x_0}(x+\epsilon) \cdot \left(f(x+\epsilon) - \tilde{g}_f(x+\epsilon)\right)^2 + \lambda \|\sigma \circ w\|_2^2$$

Attribution Uncertainty penalty

Privacy harms Explanations [CHI'22]



Explanations harm Privacy [ICCV'21]



Wrapping Up

Department of Computer Science
School of Computing

What did we learn?
Feature Importance Explanations

Feature Attribution

Weights $\boldsymbol{w}$ of f ← **Glassbox** model f
e.g. Linear Regression, Logistic Regression

Weights $\boldsymbol{w}$ of g ← **Model-agnostic** explainer model g
e.g. LIME
← **Blackbox** nonlinear model f

Saliency Map

Grad-CAM explanation g ← **Blackbox** CNN model f

Next week:
Unsupervised Learning

Image credit: https://hip2save.com/2019/11/27/lego-classic-creative-fun-900-piece-set-only-20-at-walmart-regularly-40/

W12 Pre-Lecture Task (due before next Mon)

Read

1. Clustering With More Than Two Features? Try This To Explain Your Findings by Mauricio Letelier

Task

1. Describe other use cases where you need to **apply domain knowledge** with data-driven **unsupervised learning** to better understand your business or engineering problem

   Tip: you can your own projects too; you don't have to be correct

2. Post a 1–2 sentence answer to the topic in your tutorial group: #tg-xx