Explainable AI (XAI)

11A

CS 3244 Machine Learning

NUS National University of Singapore | Computing

Week 11A: 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 11A: 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)
3. Human-centered XAI (not in exam)

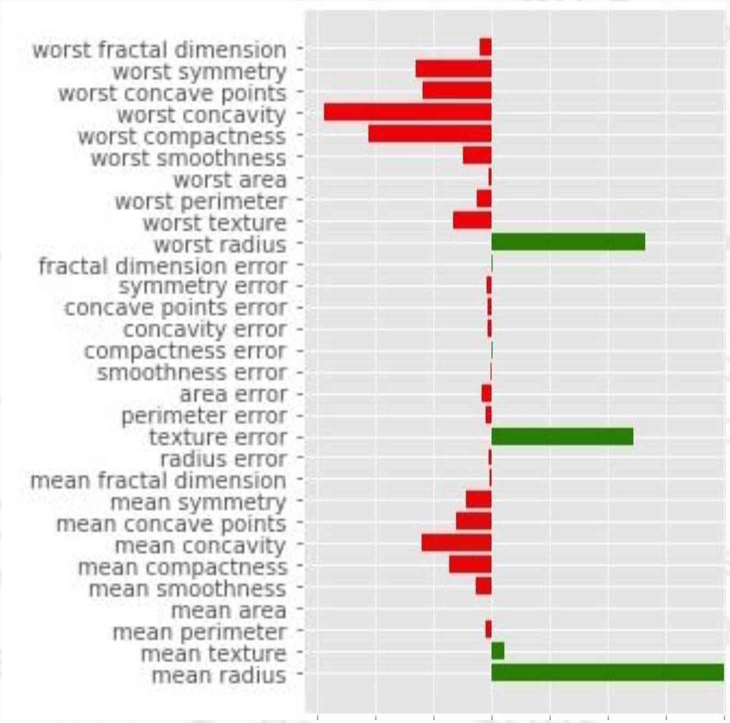
EXPLAINABLE AI EXPLAINED!

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

Why??



Evidence for Cancer | Evidence for No Cancer

Explanation: Feature Attributions

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

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

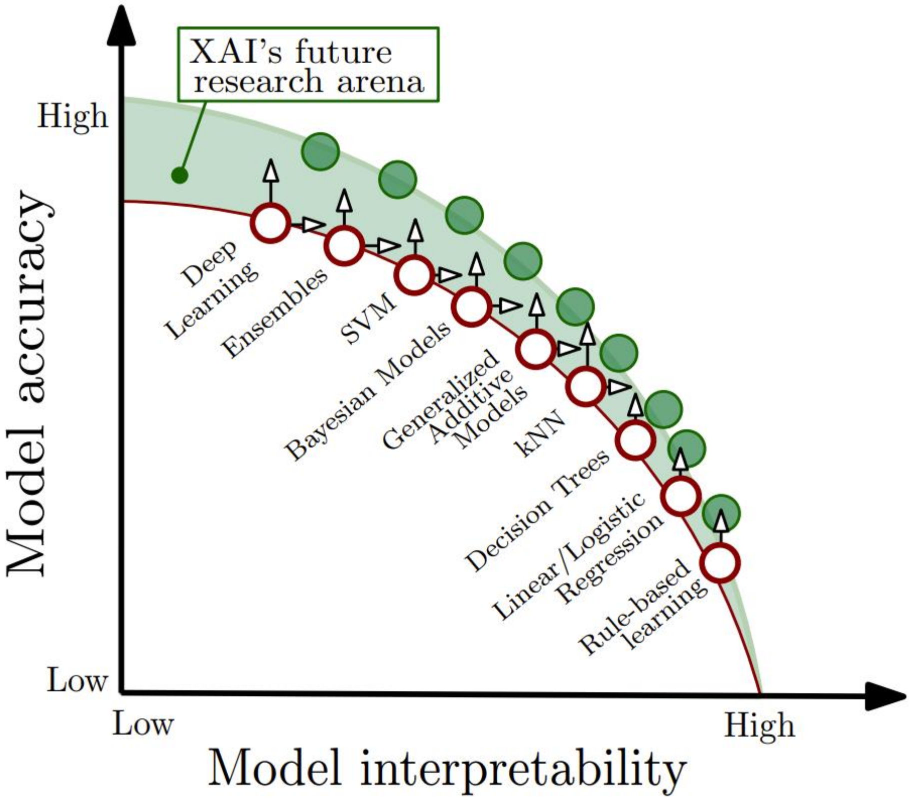
Why? Feature Importance

- Question
  - **Why?**
  - What caused this prediction?
- 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 Map of CNN

Performance-Interpretability Trade-off

High performance
(Low interpretability)

- Non-Linear relationships
  - $y = w_1 x_1^2 + \log(x_2)$
- Interacting features
  - $y = x_1 + x_2 + x_1 x_2$
- Many features or parameters
- Unrelatable features



High interpretability
(Lower performance)

- Linear relationships
  - $y = w_1 x_1 + w_2 x_2$
- Independent features
  - $y = x_1 + x_2$
  - $y = f_1(x_1) + f_2(x_2)$
- Few features (x_r), and parameters (w_r, θ_r)
- Semantically meaningful features

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Explainable AI (XAI) Categorization

<table>
<tr><th colspan="2"></th><th colspan="4">Explanability</th></tr>
<tr><th colspan="2"></th><th>Glassbox</th><th>Model-Agnostic</th><th>Model-Specific</th><th>Blackbox</th></tr>
<tr><td rowspan="2">Scope</td><td>Global</td><td>• Linear Regression
• Logistic Regression
• Decision Tree</td><td colspan="2">• Collection of local explanations</td><td rowspan="2">• Deep Neural Networks
• Highly non-linear models</td></tr>
<tr><td>Local</td><td>• Examples (e.g., kNN)</td><td>• LIME</td><td>• Grad-CAM</td></tr>
</table>

Definitions

Glassbox model
- Prediction model is implicitly interpretable

Model-Agnostic explanation
- Uses surrogate model to provide simple explanations

Model-Specific explanation
- Derived from calculations in specific prediction model

Blackbox model
- Very uninterpretable, not transparent

Global explanation
- Explains general behavior for all instances
- Does not change for different instances

Local explanation
- Explains behavior for specific instance

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

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

$\hat{y} \propto w_r x_r, \forall r$

If $w_1 = k w_2$,
then w_1 is k times more important than w_2

<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

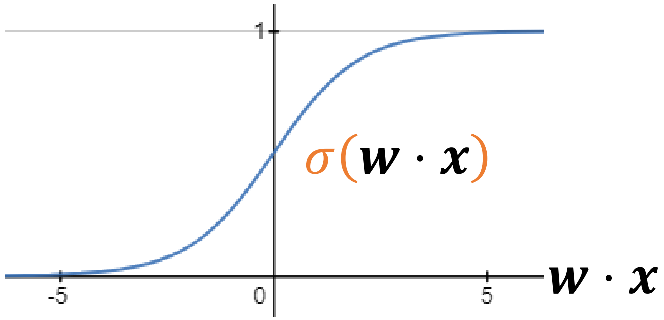
Logistic Regression

Department of Computer Science
School of Computing

How would you interpret?
Logistic Regression

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

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



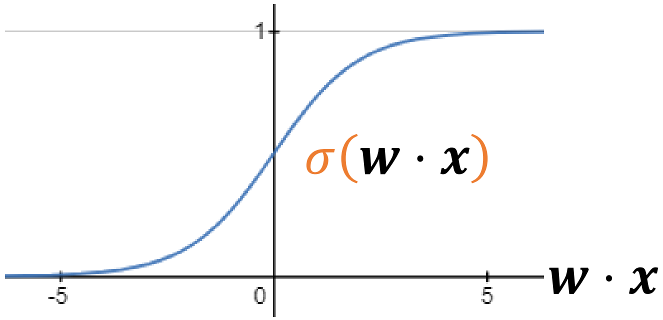
How would you interpret?
Logistic Regression

Not proportional,
since non-linear
If $w_1 = k w_2$,
then what is the relationship between w_1 and w_2?

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

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

$$\boxed{\hat{y} \not\propto w_r x_r, \forall r}$$



<u>Weighted Sum Interpretation</u>

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

$\hat{y} = \sigma(f)$ is
positively monotonic, i.e.,
$f_1 > f_2 \Rightarrow \sigma(f_1) > \sigma(f_2)$
Bigger $f \Rightarrow$ bigger $\hat{y}$

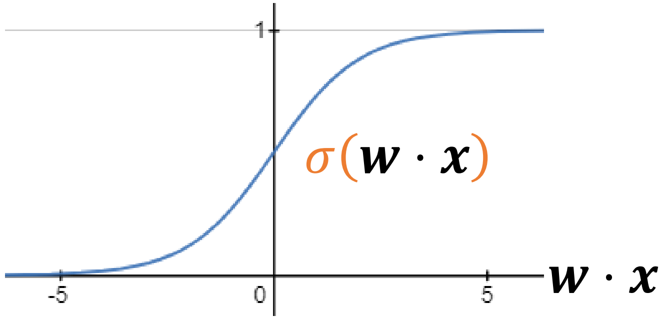
Interpreting weights in Logistic Regression

Probability

$$P(\hat{y}) = p = \hat{y} = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}} = \sigma(p)$$

$$\frac{1}{p} = 1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}$$

$$\frac{1-p}{p} = \frac{1}{p} - 1 = \cancel{1 +} e^{-\boldsymbol{w}\cdot\boldsymbol{x}} = \frac{1}{e^{\boldsymbol{w}\cdot\boldsymbol{x}}}$$

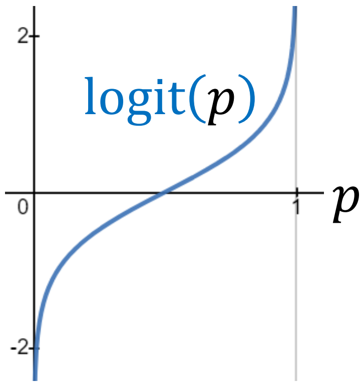


Odds Ratio

$$\frac{p}{1-p} = e^{\boldsymbol{w}\cdot\boldsymbol{x}}$$

Log Odds Ratio

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right) = \boldsymbol{w}\cdot\boldsymbol{x}$$



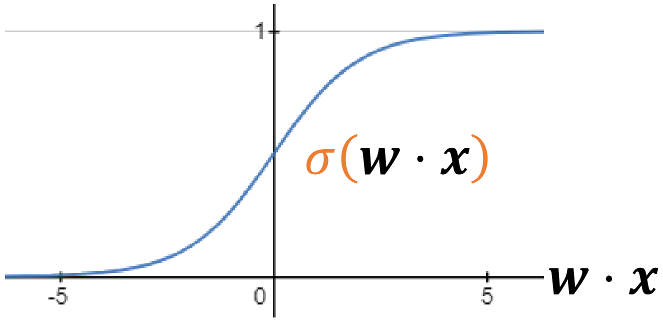
$$\boxed{\mathrm{logit}\big(P(\hat{y})\big) \propto w_r x_r, \forall r}$$

If $w_1 = k w_2$,
then log odds ratio of w_1 is
k times bigger than of w_2

How would you interpret?
Logistic Regression

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

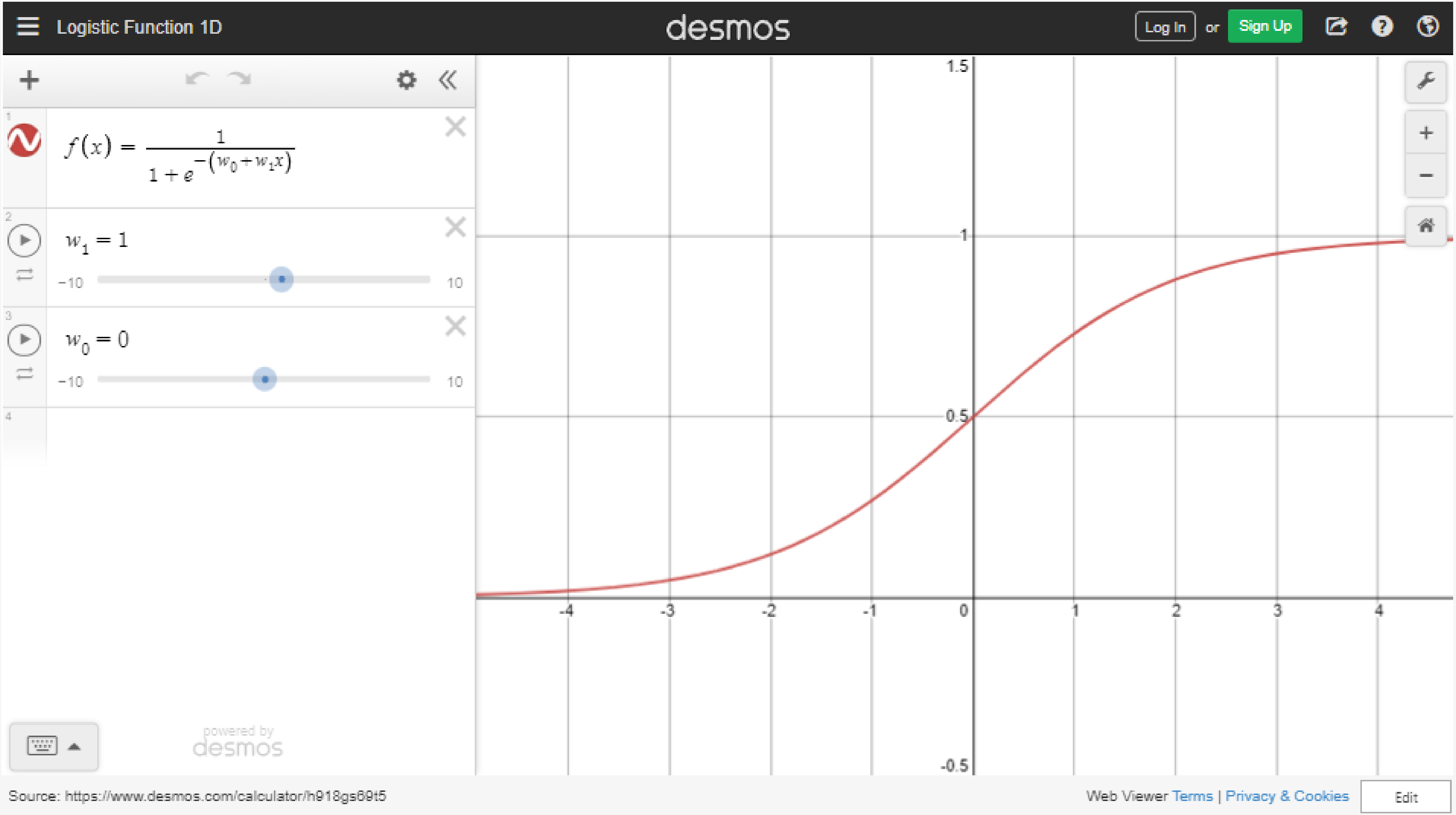
$$f = \boldsymbol{w} \cdot \boldsymbol{x} = \sum_{r=0}^{n} w_r x_r$$



$$\boxed{\text{logit}(P(\hat{y})) \propto w_r x_r, \forall r}$$

If $w_1 = k w_2$,
then log odds ratio of w_1 is
k times bigger than of w_2

Weighted Sum Interpretation	Gradient Interpretation
Bigger w_r means • **Larger** importance • Direction indicates **influence**	**Bigger** w_r means? • Steepness? Sigmoid bounded between 0 and 1 • Direction in 2D (or higher)?


Logistic Function 1D
desmos
Log In
or
Sign Up
$f(x) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$
$w_1 = 1$
−10
10
$w_0 = 0$
−10
10
powered by desmos
1.5
1
0.5
-0.5
-4
-3
-2
-1
0
1
2
3
4
Source: https://www.desmos.com/calculator/h918gs89t5
Web Viewer Terms | Privacy & Cookies
Edit

3D Plotter v1.0 - Logistic R...

desmos

Log In or Sign Up

1 Display controls

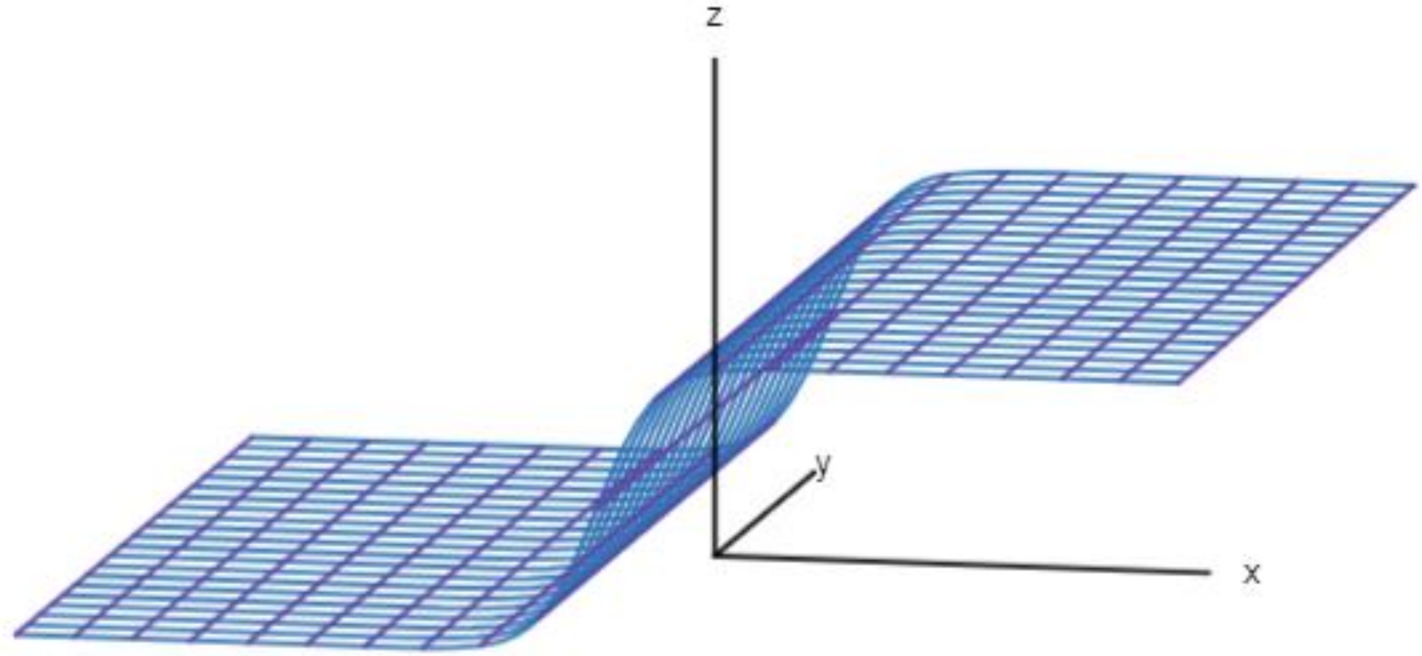
8

9 This graph has lots of different ways of plotting 3D objects.

10 To turn on a plotter, click on its folder icon on the left hand side. Settings for the plotters can be configured inside the respective folders.

11 To plot more than one function at once, simply replace the single function with a list of functions. E.g. below, change g(x, y) = cos(xy+T) to g(x, y) = [cos(xy+T), x^2] .

12 Some things work very well and run very smoothly, others not so much. If something is running particularly slowly, then turn off all unnecessary features i.e. arrowheads on the axes or on vector fields. If this doesn't help, then try deleting unnecessary parts of the program e.g. if you are only using the function plotter then delete the vector field plotter.



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

Web Viewer Terms | Privacy & Cookies

Edit

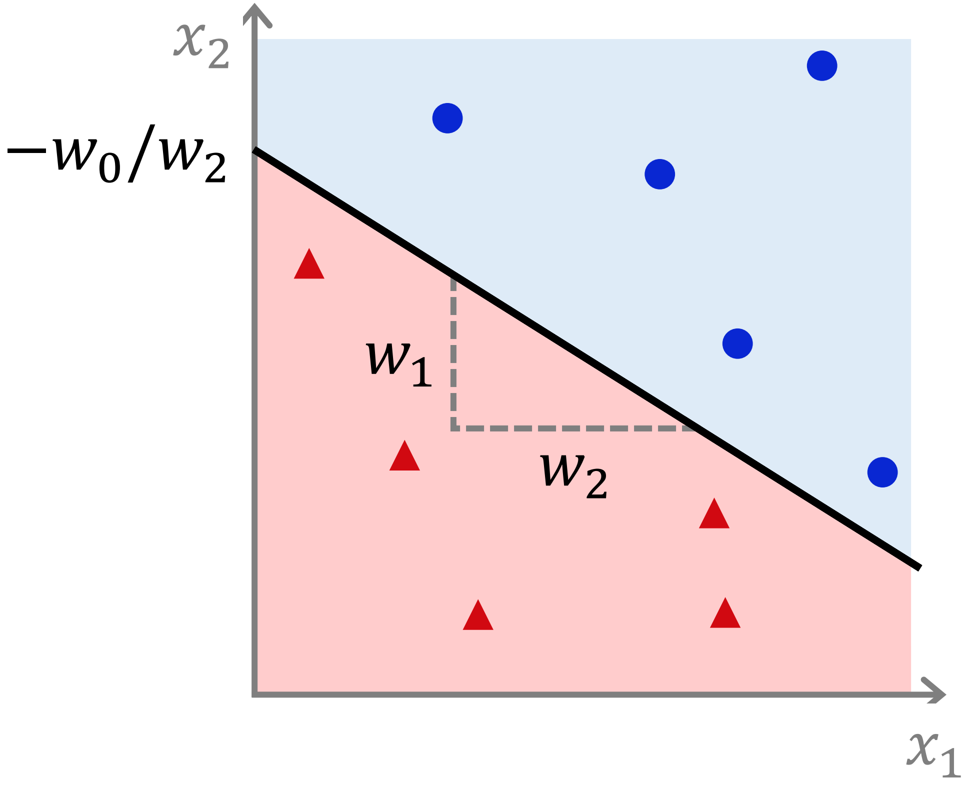
Linear Classification

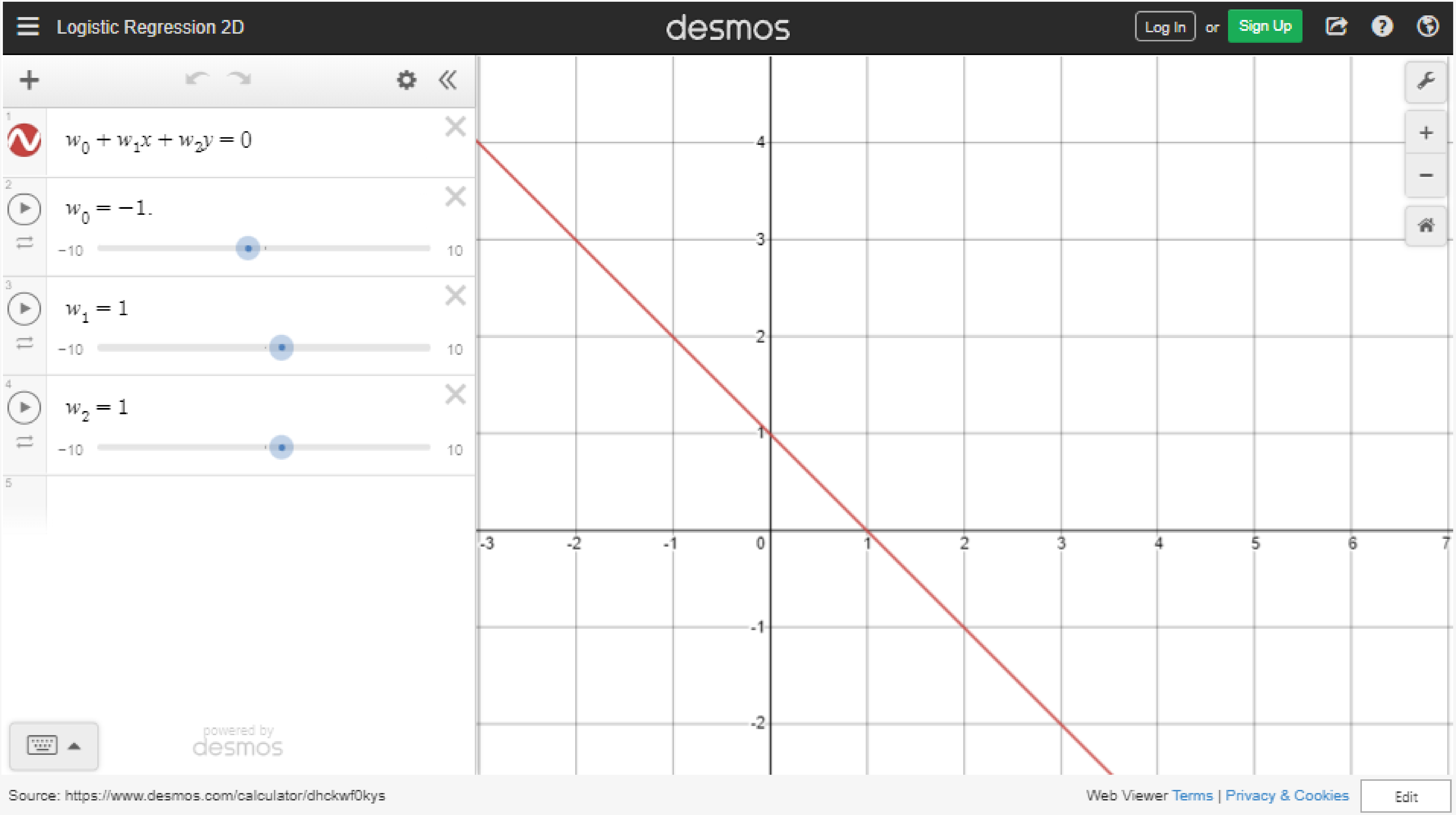
$$w_2x_2 + w_1x_1 + w_0 = 0$$

$$\sum_{r=0}^{n} w_r x_r = 0 \quad x_0 = 1$$

$$\sum_{r=0}^{n} w_r x_r > 0 \qquad \sum_{r=0}^{n} w_r x_r \leq 0$$

$$\hat{y} = \sigma\left(\sum_{r=0}^{n} w_r x_r\right)$$




Logistic Regression 2D
desmos
Log In
or
Sign Up
$w_0 + w_1x + w_2y = 0$
$w_0 = -1.$
−10
10
$w_1 = 1$
−10
10
$w_2 = 1$
−10
10
powered by
desmos
Source: https://www.desmos.com/calculator/dhckwf0kys
Web Viewer Terms | Privacy & Cookies
Edit

Linear Classification

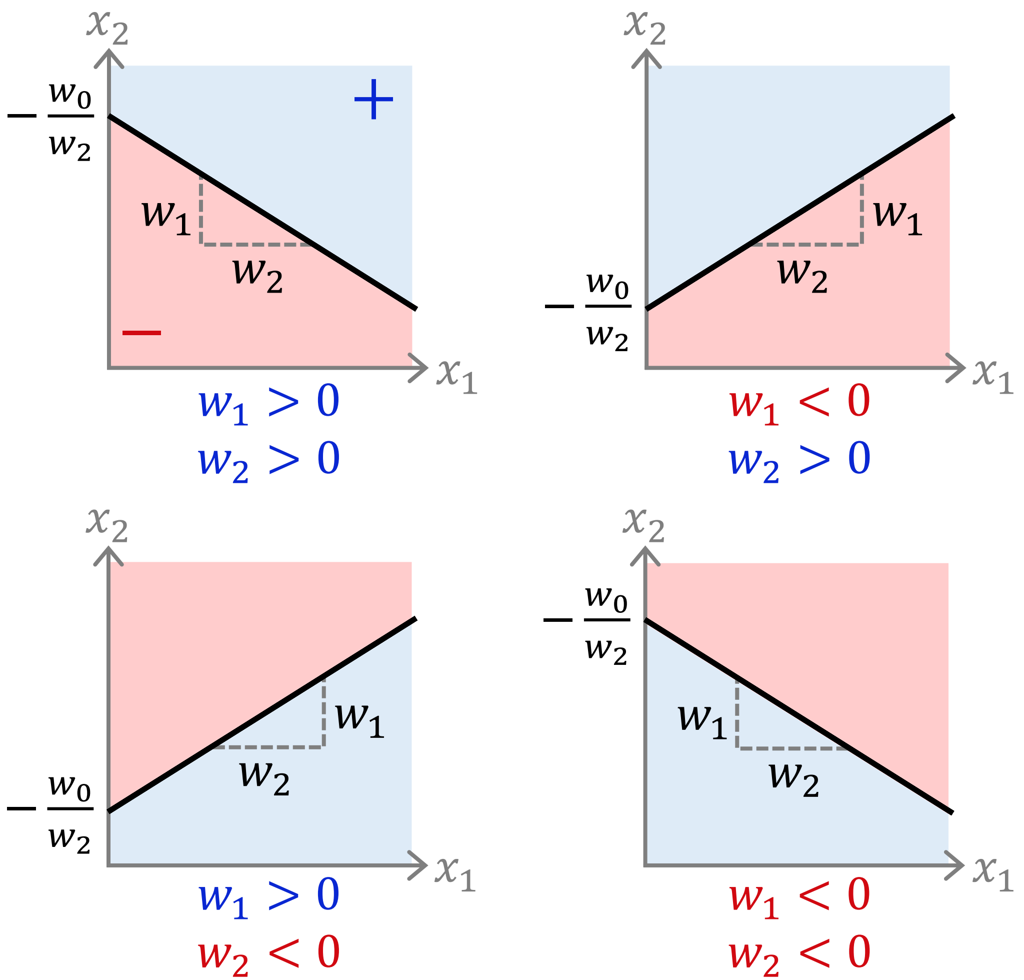
$$w_2 x_2 + w_1 x_1 + w_0 = 0$$

$$\sum_{r=0}^{n} w_r x_r = 0 \quad x_0 = 1$$

$$\sum_{r=0}^{n} w_r x_r > 0 \qquad \sum_{r=0}^{n} w_r x_r \leq 0$$

$$\hat{y} = \sigma\left(\sum_{r=0}^{n} w_r x_r\right)$$

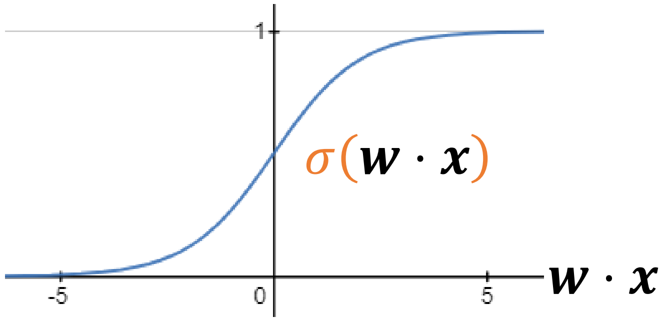
Weight sign indicates direction towards **pos**/**neg** prediction



How would you interpret?
Logistic Regression

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

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



$$\mathrm{logit}(P(\hat{y})) \propto w_r x_r, \forall r$$

If $w_1 = k w_2$, then log odds ratio of w_1 is k times bigger than of w_2

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


Questions!

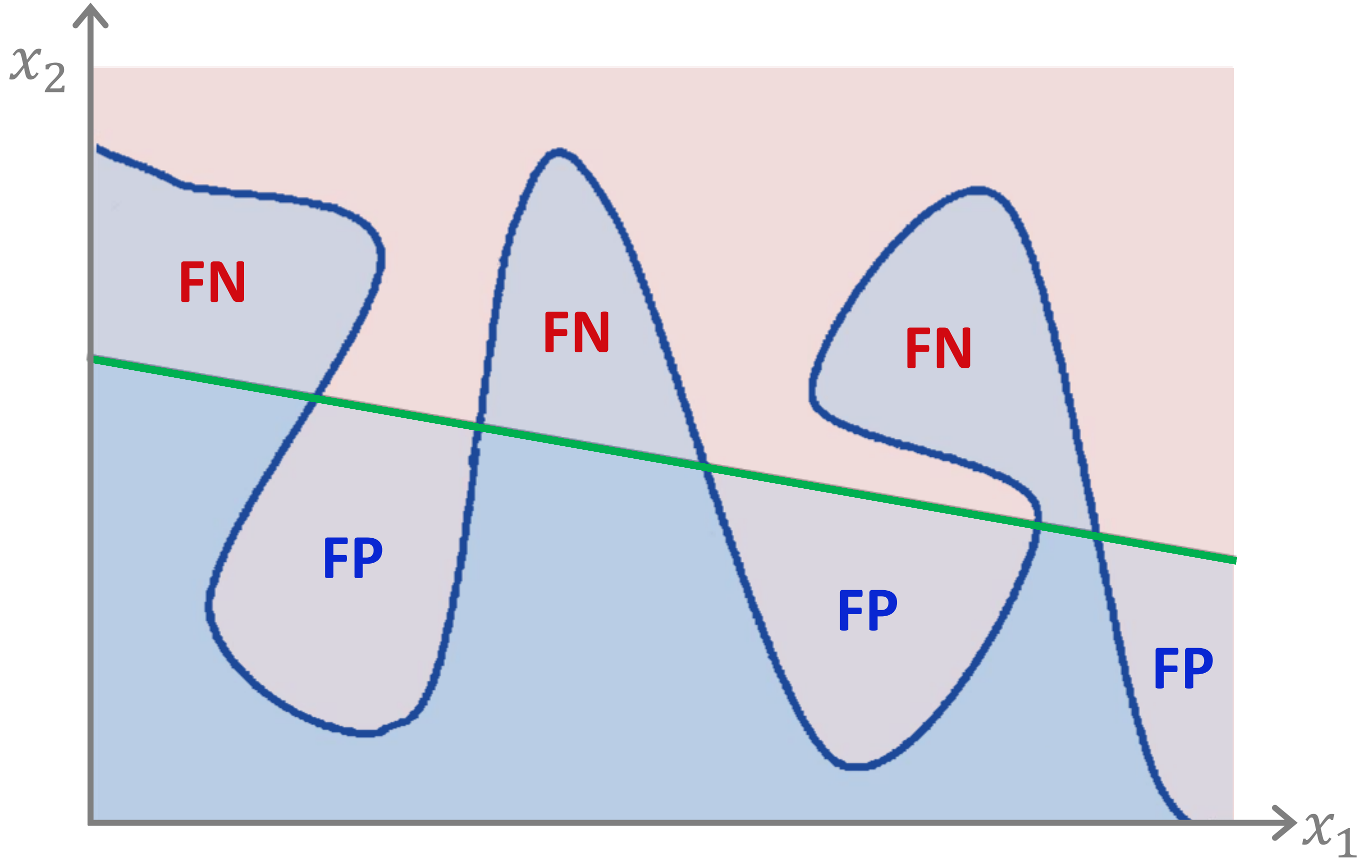
Local Interpretable Model-agnostic Explanations

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

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

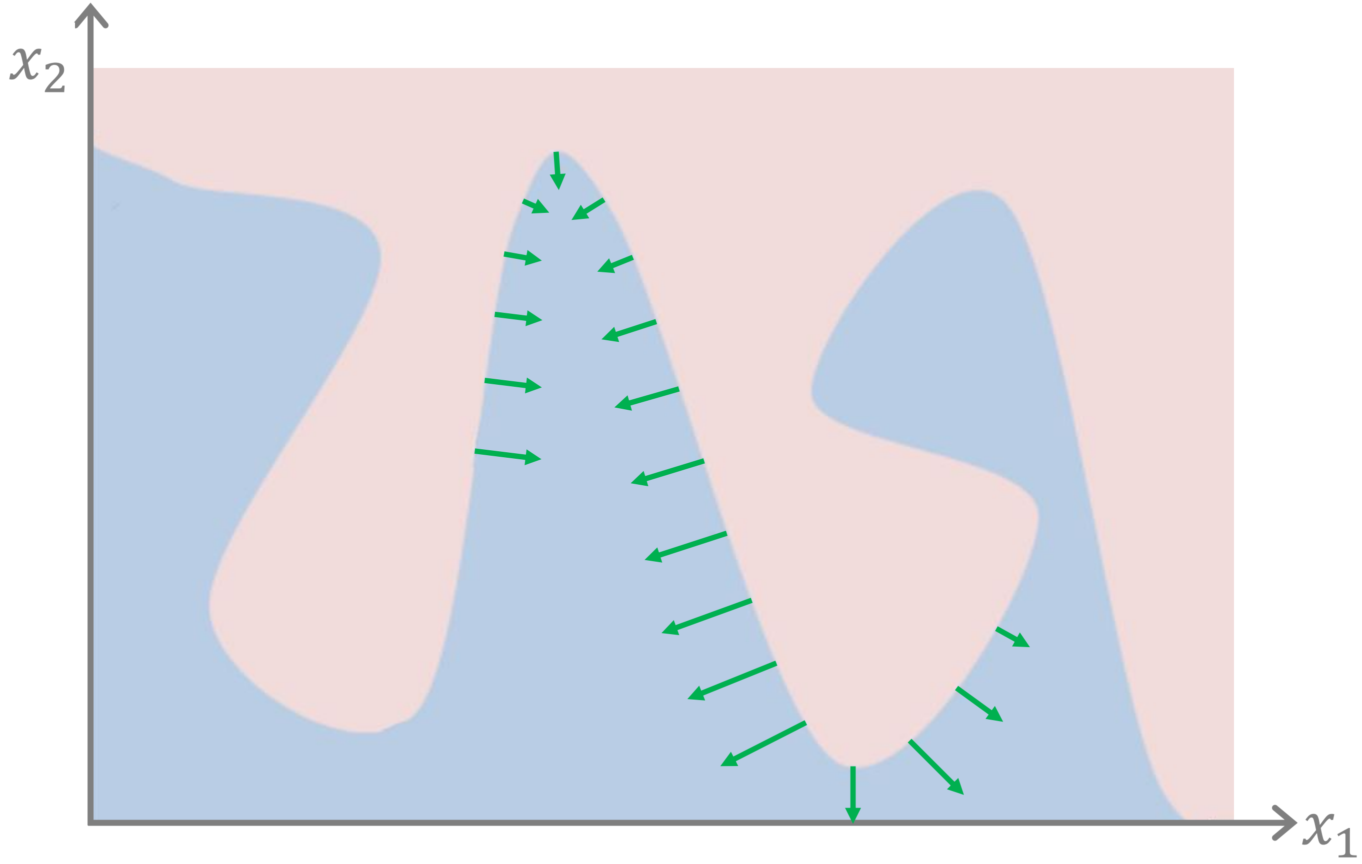
Image Credit: https://santiagof.medium.com/model-interpretability-making-your-model-confess-lime-89db7f70a72b

How to describe with just x_1 and x_2?

Non-Linear Decision Boundary

$$\nabla f(\boldsymbol{x}) \neq \begin{pmatrix} w_1 \\ w_2 \\ \vdots \end{pmatrix} = \boldsymbol{w}$$

- Since f is not linear



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**

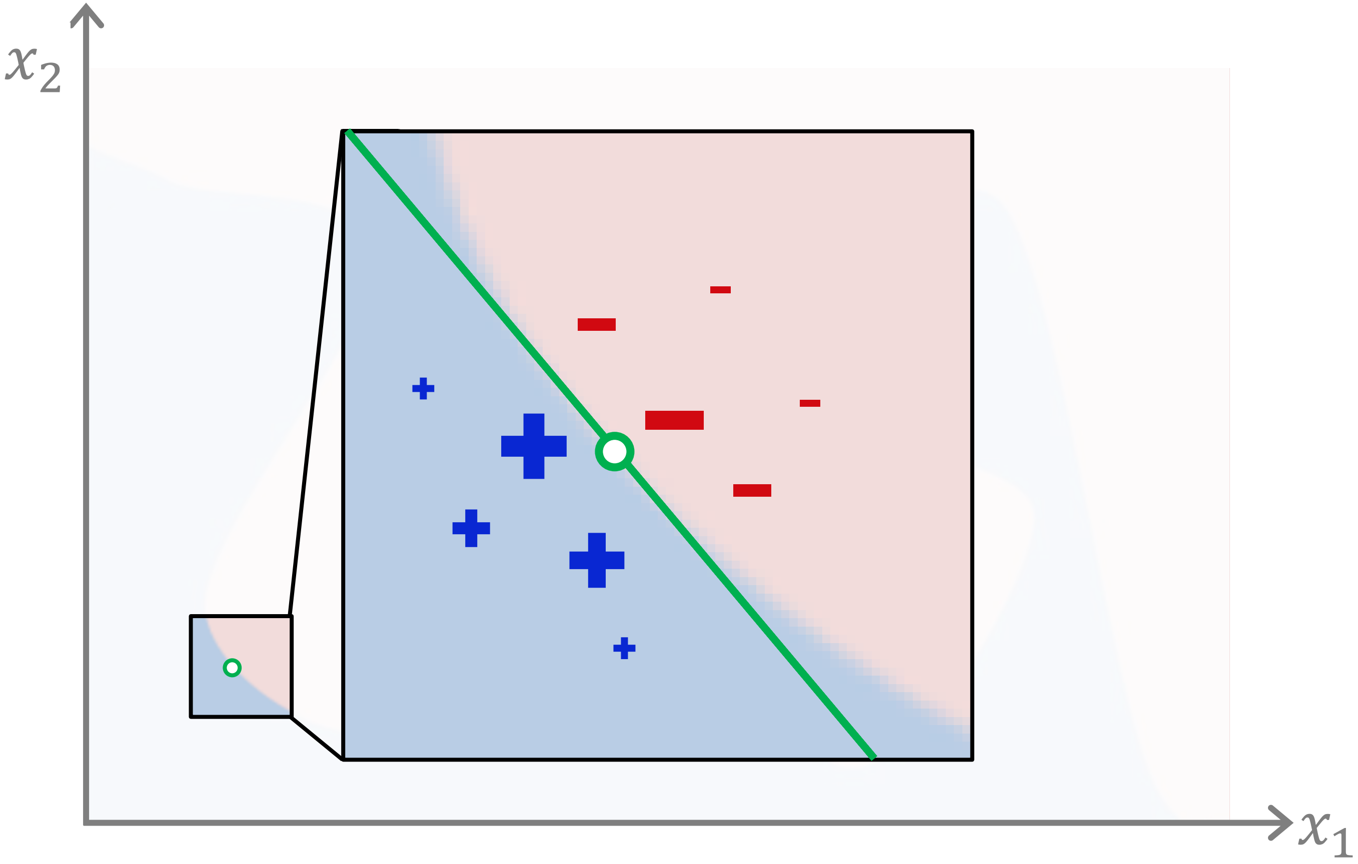
$$g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) = \frac{\partial f}{\partial \boldsymbol{x}} = \begin{pmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \\ \vdots \end{pmatrix}$$

- **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



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:

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

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

LIME

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

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

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

f is the **predictor** function (model)

g is the **explainer** function (model)

$\pi_x\left(\boldsymbol{x}^{\langle\eta\rangle}\right)$ is the neighbor **proximity** function

- E.g., exponential decay $\exp\left(-\left(d\left(\boldsymbol{x}, \boldsymbol{x}^{\langle\eta\rangle}\right)\right)^2\right)$

$L(f, g, \pi_x)$ is the ***locally-weighted error loss*** function between predictor f and explainer g

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

$\Omega(g)$ is the **sparsity regularization**

- Want simpler explanation
  - ⇒ *fewer weights*
  - ⇒ Lasso (L1 norm)
- Penalizes if total weights is too large
- λ is hyperparameter on how much to penalize

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

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?

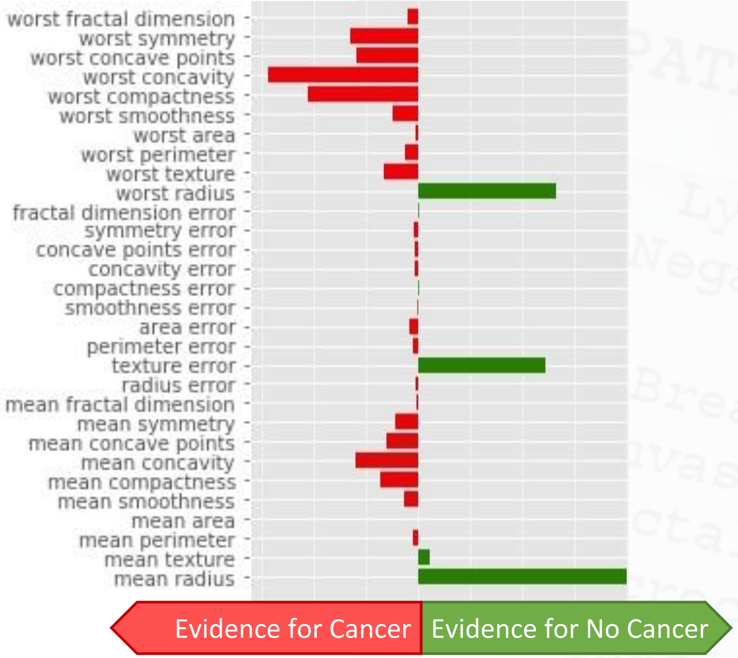
Why do the two set of weights differ?

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

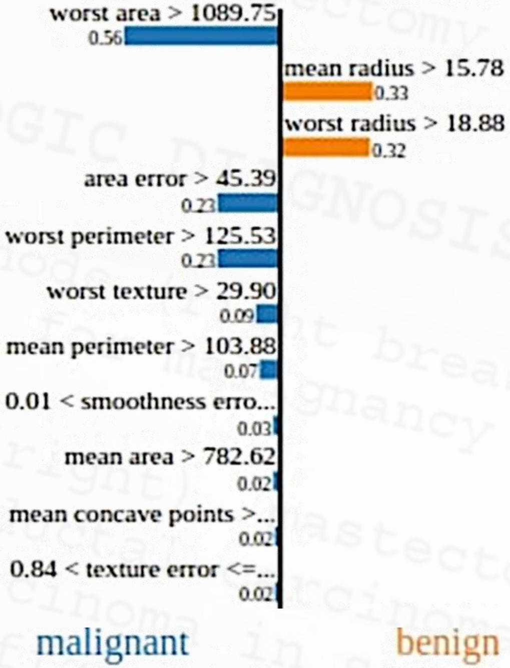
Prediction $\hat{y} =$ Cancer

Logistic Regression



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

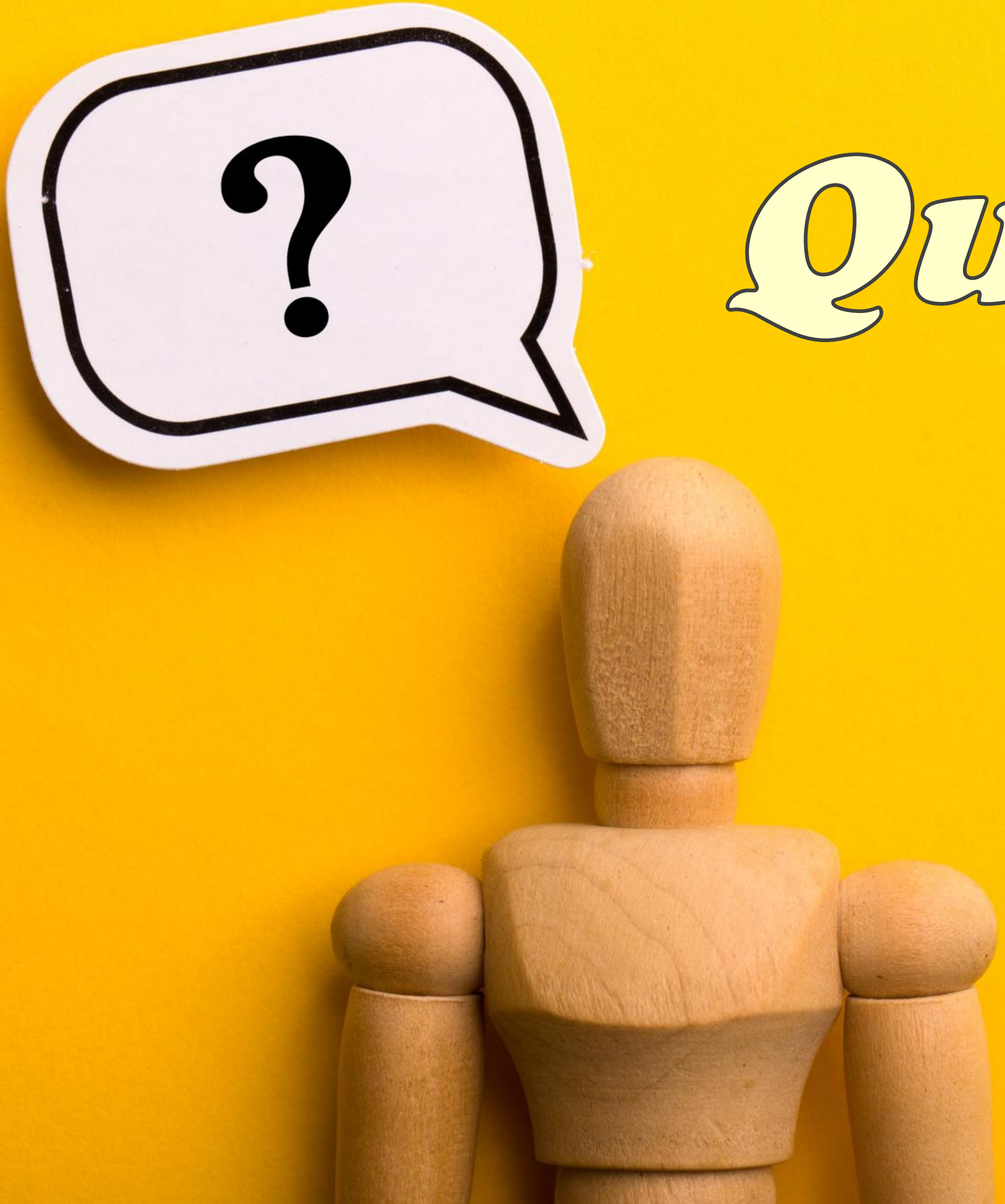
LIME



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

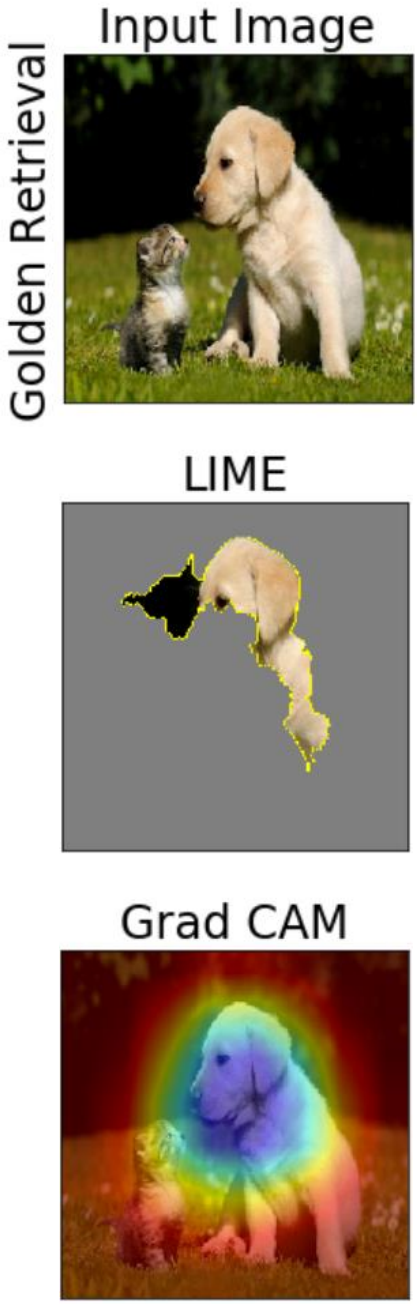
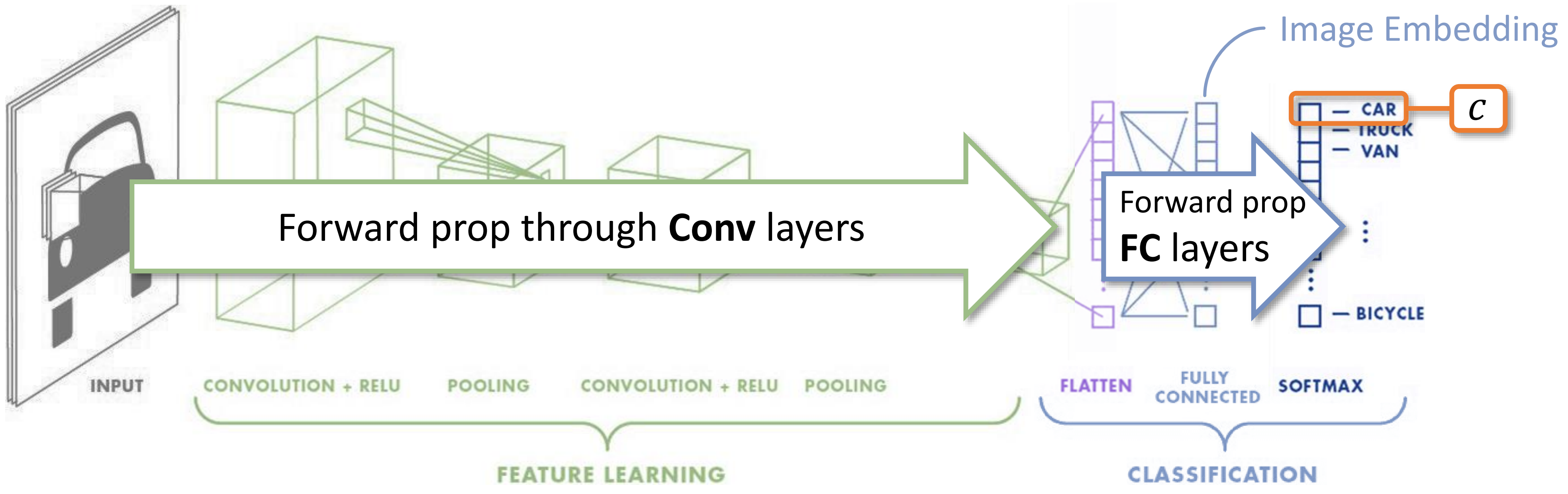


Image credit: https://arxiv.org/pdf/1908.04389.pdf

Convolutional Neural Network



Key concepts

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

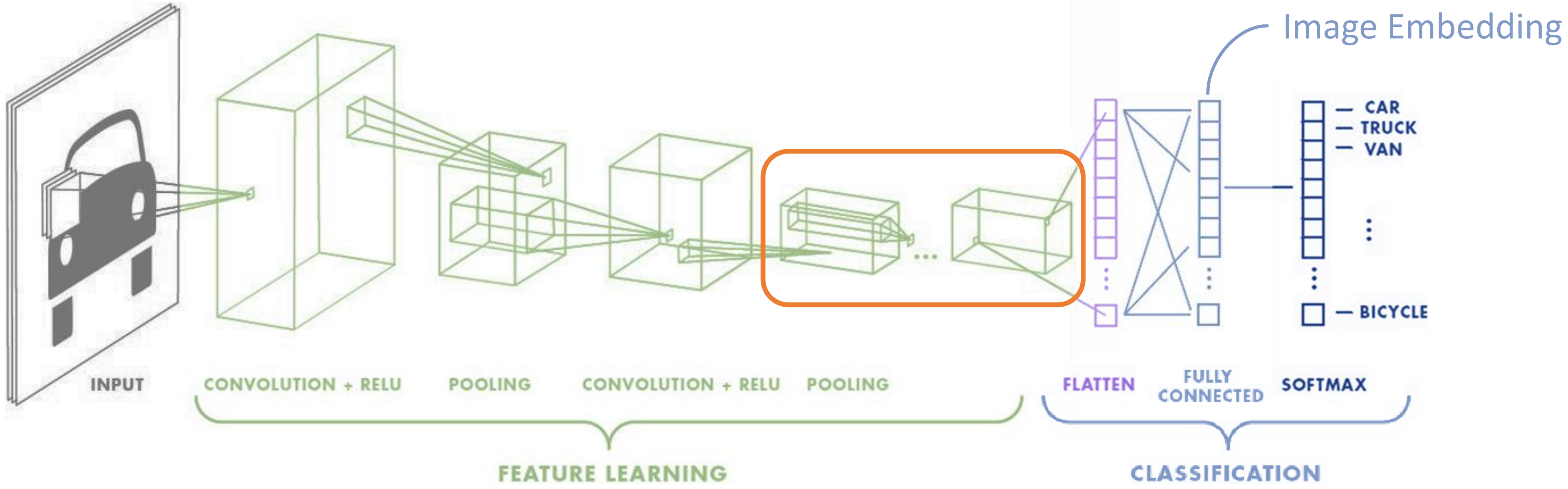
❷ Flattening
- Convert to fixed-length 1D vector

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

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

Convolutional Layer:
Feature Kernels & Feature Maps

$$\boldsymbol{X}^{[0]} \to g^{[1]}\left(\boldsymbol{W}^{[1]} * \boldsymbol{X}^{[0]}\right) = \boldsymbol{A}^{[1]} \xrightarrow{\text{Pooling}} g^{[2]}\left(\boldsymbol{W}^{[2]} * \boldsymbol{X}^{[1]}\right) = \boldsymbol{A}^{[2]}$$

1 channel

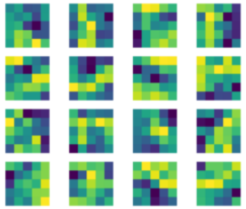
16 filters

16 channels

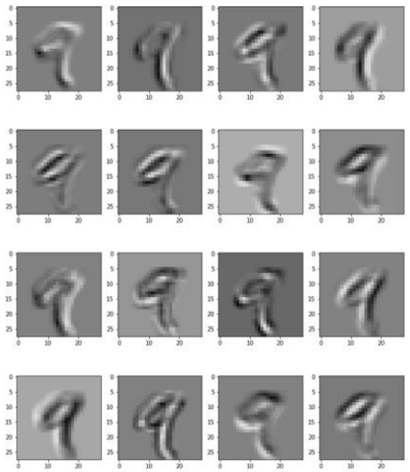
32 filters

32 channels

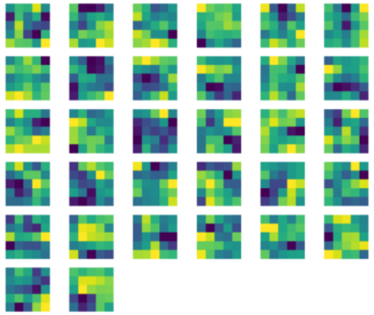
Input



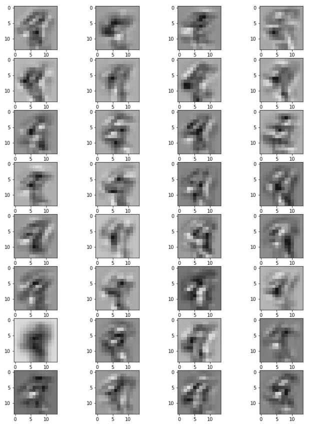
Layer 1 Feature Kernels



Layer 1
Activation Maps



Layer 2
Feature Kernels



Layer 2
Activation Maps

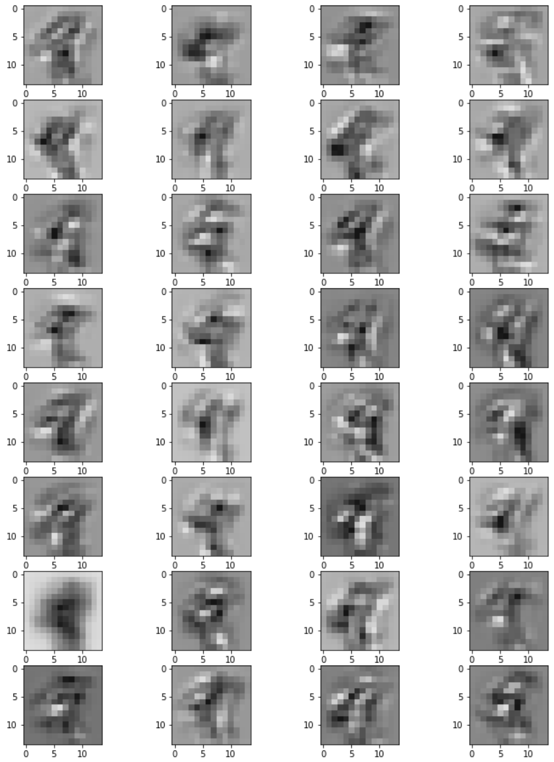
Want to explain with
Activation Maps

Image credit: https://medium.com/dataseries/visualizing-the-feature-maps-and-filters-by-convolutional-neural-networks-e1462340518e

Multi-Channel Activation Maps (layers diagram)

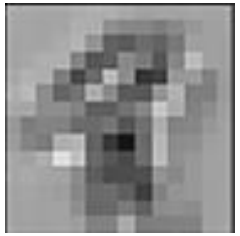
We know the Activation Maps,
But which is more **important**?

Calculate with **Gradients**:
Steeper ⇒ More **important**



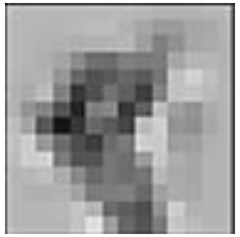
$$\boldsymbol{A}^{[l]} = \begin{pmatrix} \boldsymbol{A}_1^{[l]} & \cdots & \boldsymbol{A}_k^{[l]} \end{pmatrix}$$

$$\boldsymbol{A}^{[l]} = \begin{matrix} \begin{pmatrix} a_{111}^{[l]} & \cdots & a_{1w1}^{[l]} \\ \vdots & \ddots & \vdots \\ a_{h11}^{[l]} & \cdots & a_{hw1}^{[l]} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} a_{11k}^{[l]} & \cdots & a_{1wk}^{[l]} \\ \vdots & \ddots & \vdots \\ a_{h1k}^{[l]} & \cdots & a_{hwk}^{[l]} \end{pmatrix} \end{matrix}$$



$\boldsymbol{A}_1^{[l]}$

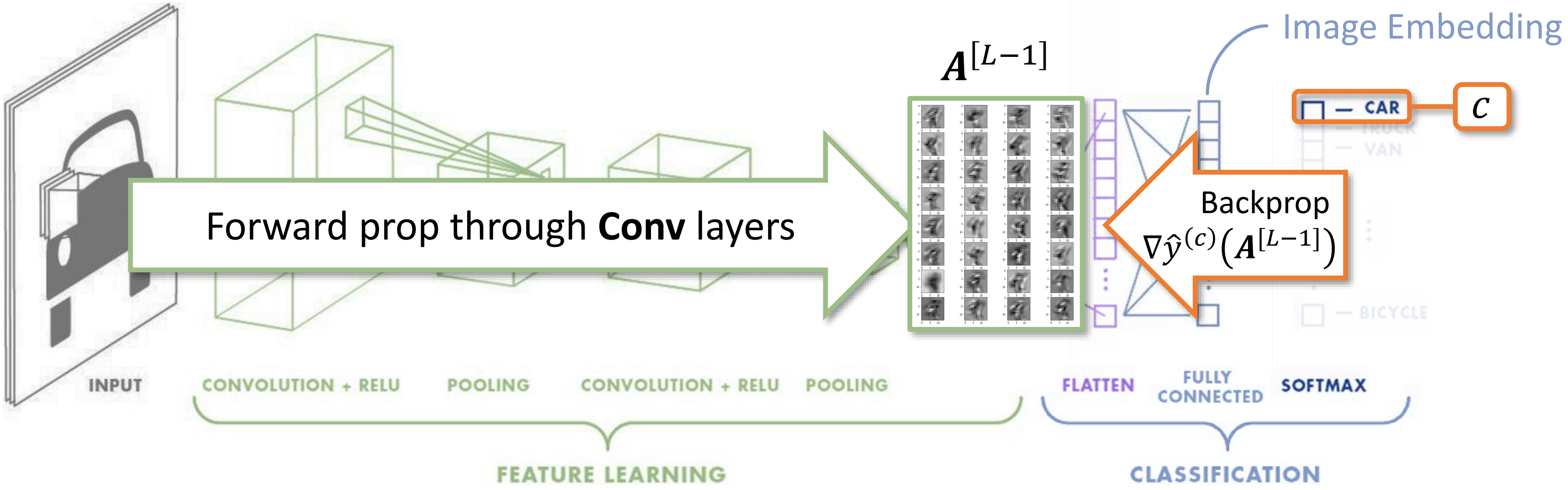
$\vdots$



$\boldsymbol{A}_k^{[l]}$

Activation of kth channel at pixel position (h, w) in layer l

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

Gradient-Weighted Class Activation Maps

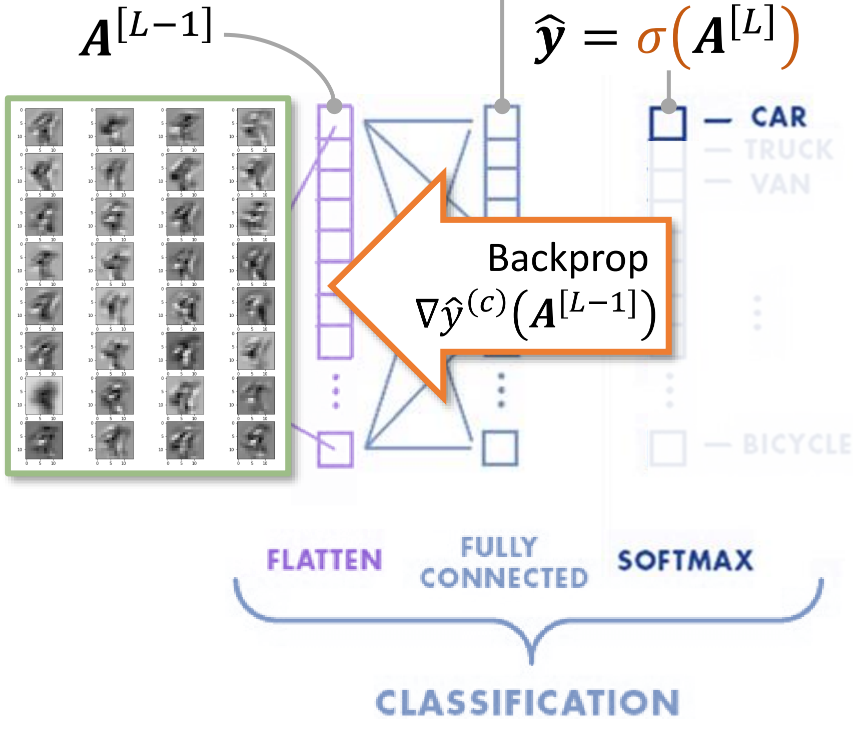
$$\frac{\partial \hat{y}^{(c)}}{\partial a^{[L]}_{hwk}} = \frac{\partial f^{[L]}}{\partial a^{[L]}_{hwk}} \frac{\partial g^{[L]}}{\partial f^{[L]}} \frac{\partial \hat{y}^{(c)}}{\partial g^{[L]}}$$

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

$$\left\| \frac{\partial \hat{y}^{(c)}}{\partial \boldsymbol{A}^{[L]}_{k}} \right\|_{1,1} = \sum_{ij} \frac{\partial \hat{y}^{(c)}}{\partial a^{[L]}_{ijk}} = \frac{\partial \hat{y}^{(c)}}{\partial a^{[L]}_{11k}} + \cdots + \frac{\partial \hat{y}^{(c)}}{\partial a^{[L]}_{hwk}}$$

$$\boldsymbol{A}^{[L]} = g^{[L]}\left(f^{[L]}\left(\boldsymbol{A}^{[L-1]}\right)\right)$$

$$\hat{\boldsymbol{y}} = \sigma\left(\boldsymbol{A}^{[L]}\right)$$



CAM

2D Matrix $\{h^{[L]} \times w^{[L]}\}$

$$\boxed{g\left(\hat{y}^{(c)}\right) = \mathrm{ReLU}\left(\sum_k \alpha^{(c)}_k \boldsymbol{A}^{[L-1]}_k\right), \quad \alpha^{(c)}_k = \left\| \frac{\partial \hat{y}^{(c)}}{\partial \boldsymbol{A}^{[L-1]}_{k}} \right\|_{1,1}, \boldsymbol{A}^{[L]}_k = \begin{pmatrix} a^{[L-1]}_{11k} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & a^{[L-1]}_{hwk} \end{pmatrix}}$$

Weighted Sum
Keep positive activations only

Importance Weight
of kth filter for cth class

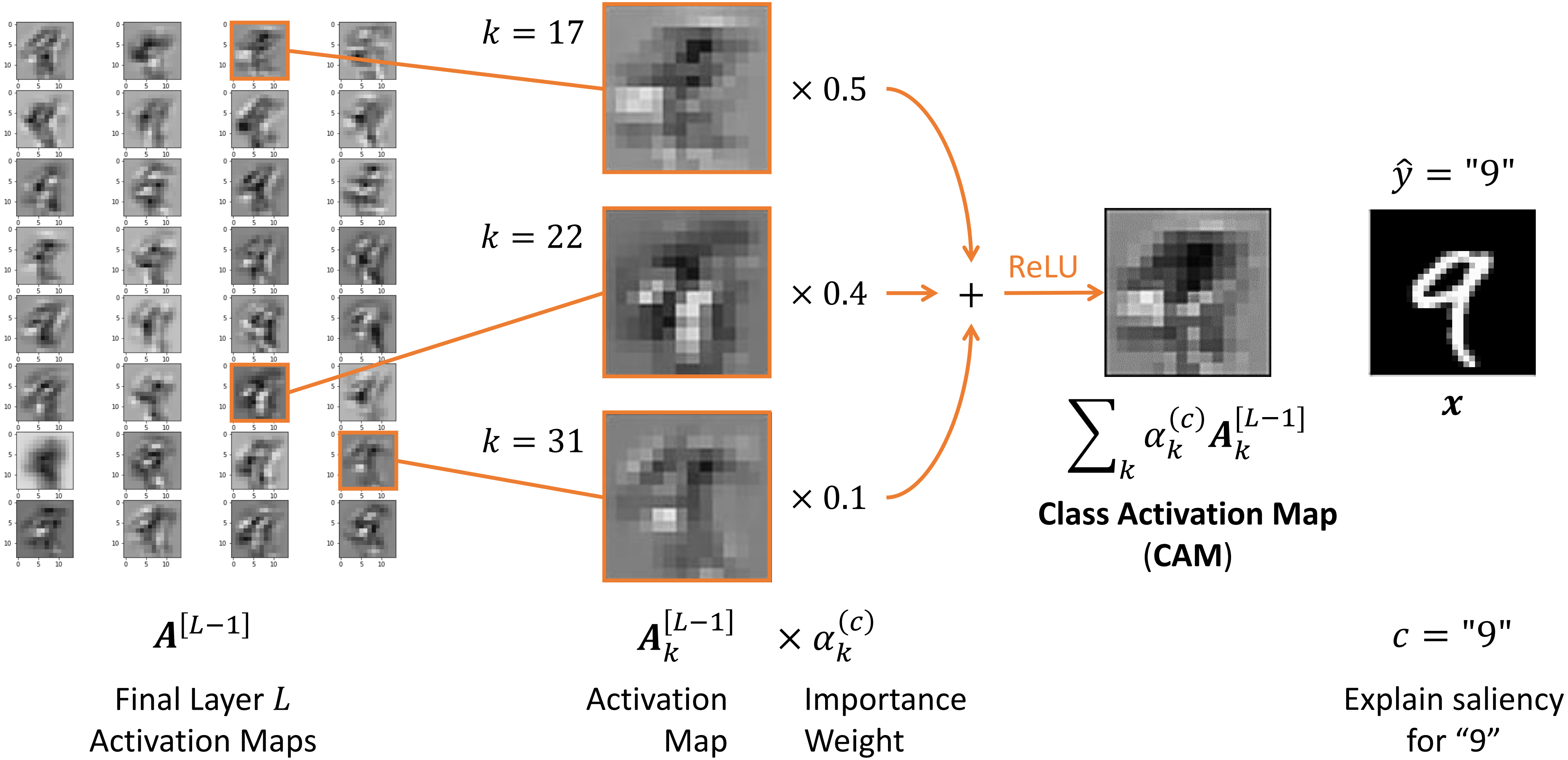
Activation Map
of kth filter in the Lth layer

Selvaraju, R. R., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *ICCV'17*.

Grad-CAM Steps

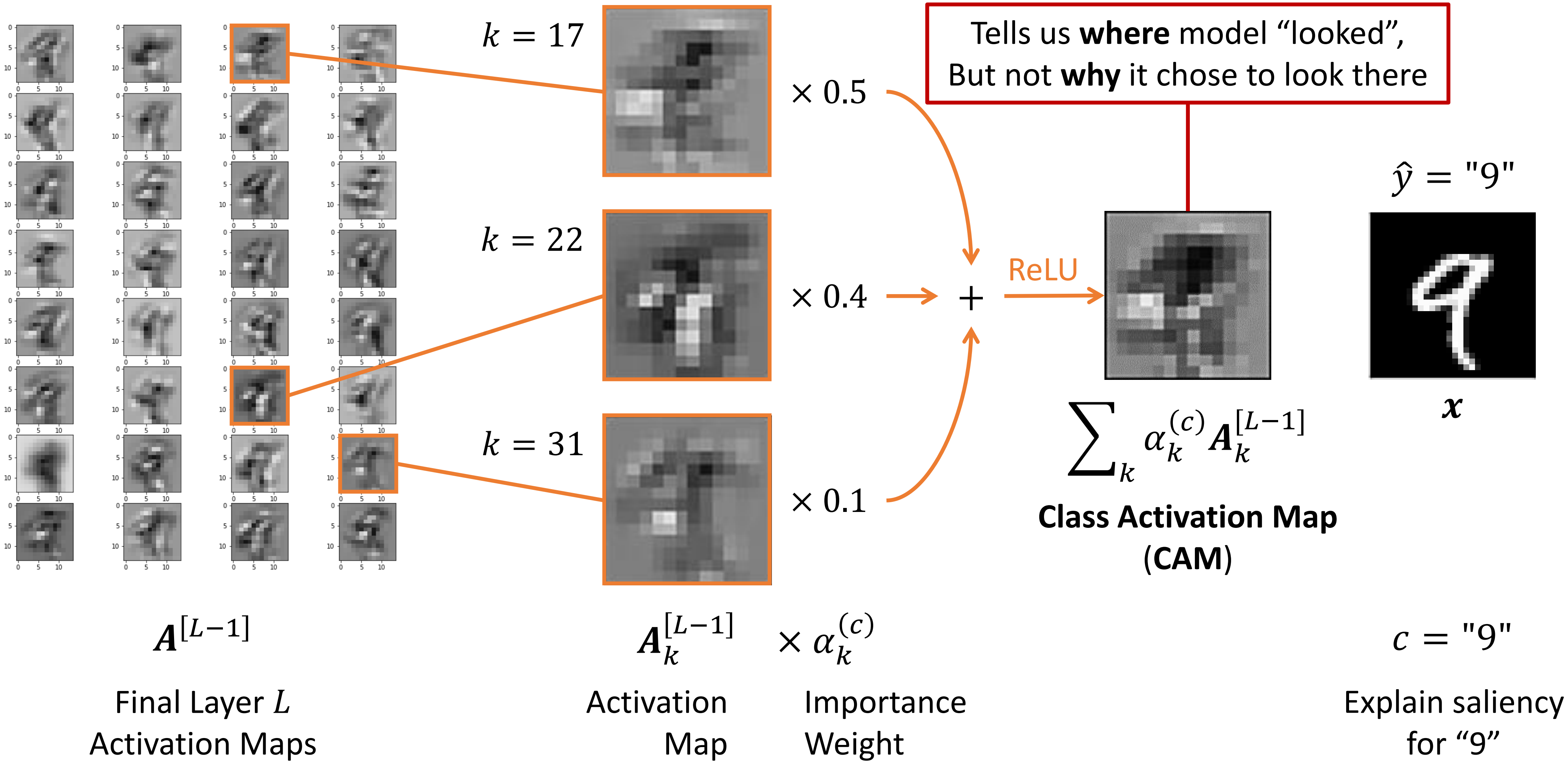
1. Compute Activation Maps $\boldsymbol{A}^{[L]}$ of last conv layer L
   1. via Forward Propagation
2. Choose class label c to explain about (e.g., predict “9”, “car”)
3. Filter prediction $\hat{y}$ to be about class c
   1. Given: $\hat{\boldsymbol{y}} = \begin{pmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(c)} \\ \hat{y}^{(n)} \end{pmatrix}$, $\boldsymbol{e}^{(c)} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$, 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 at last conv layer
   - Note that gradient is relative to activations, not weights
5. Compute weighted sum with ReLU to get **Class Activation Map**

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




?
Questions!

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



W10 Pre-Lecture Task (due before next Mon)

Watch

- Who Invented A.I.? - The Pioneers of Our Future by ColdFusion

Play

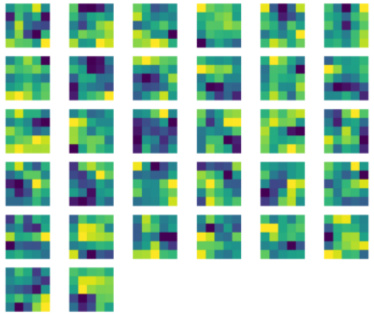
- https://distill.pub/2018/building-blocks/
  - Don’t worry about reading the whole article

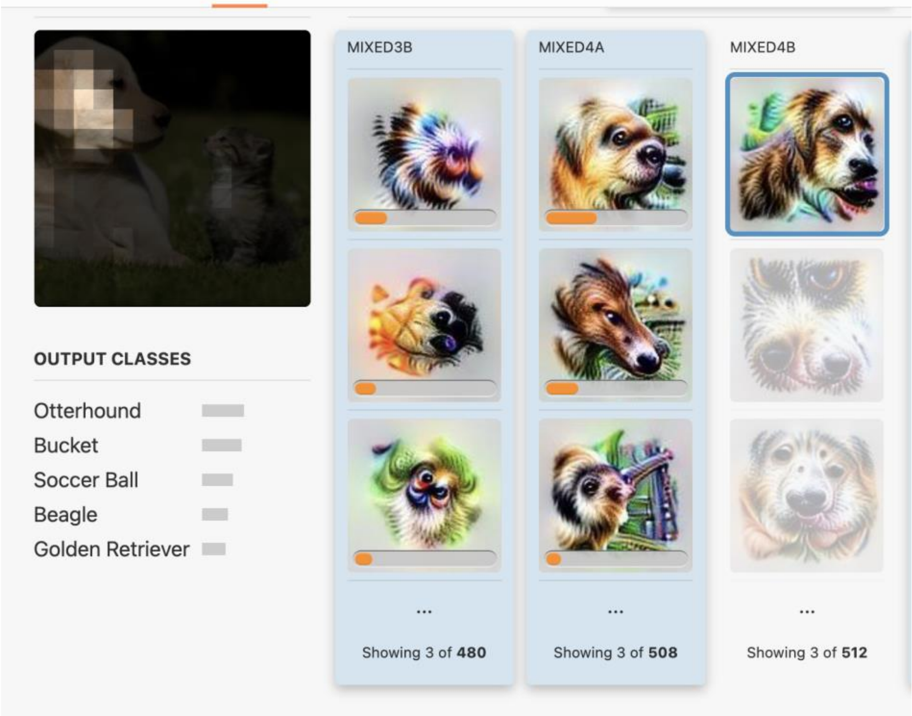
Discuss

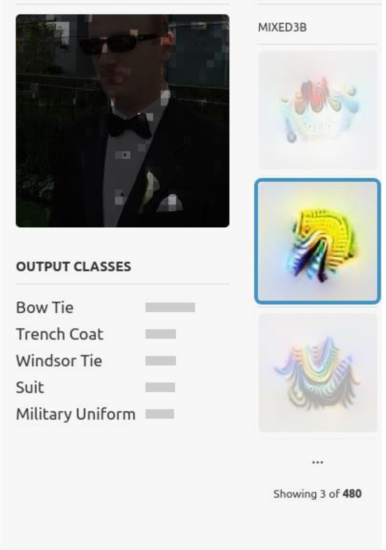
1. Identify what is strange, funny, or erroneous in the deep learning model in Building-Blocks
2. Take a screenshot of the issue and share with your tutorial mates
3. Try to explain why the model was behaving as identified
3. Post a 2–3 sentence description to the topic in your tutorial group: #tg-xx

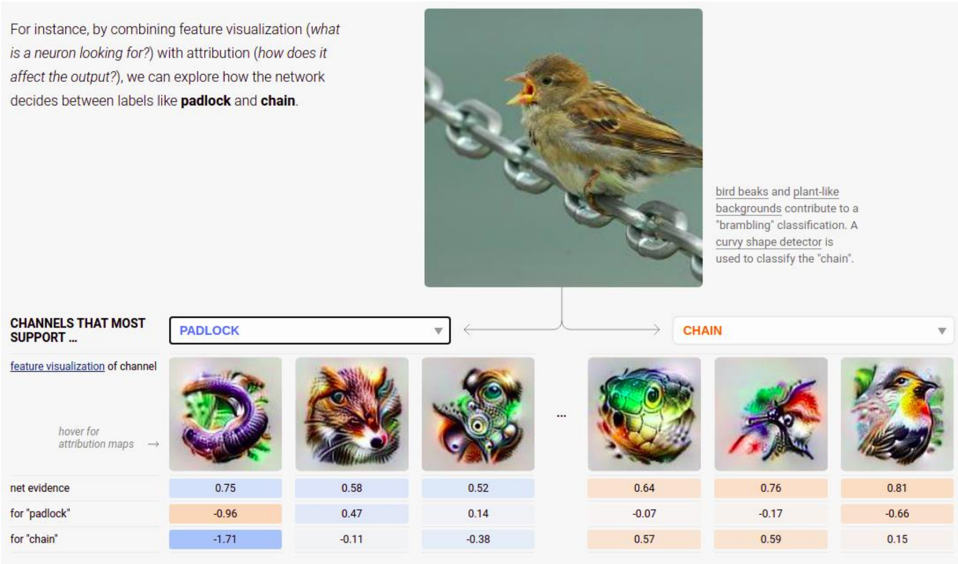
Understanding Filters

Hard to interpret kernels.
They are just **matrices** used
for **convolution**.









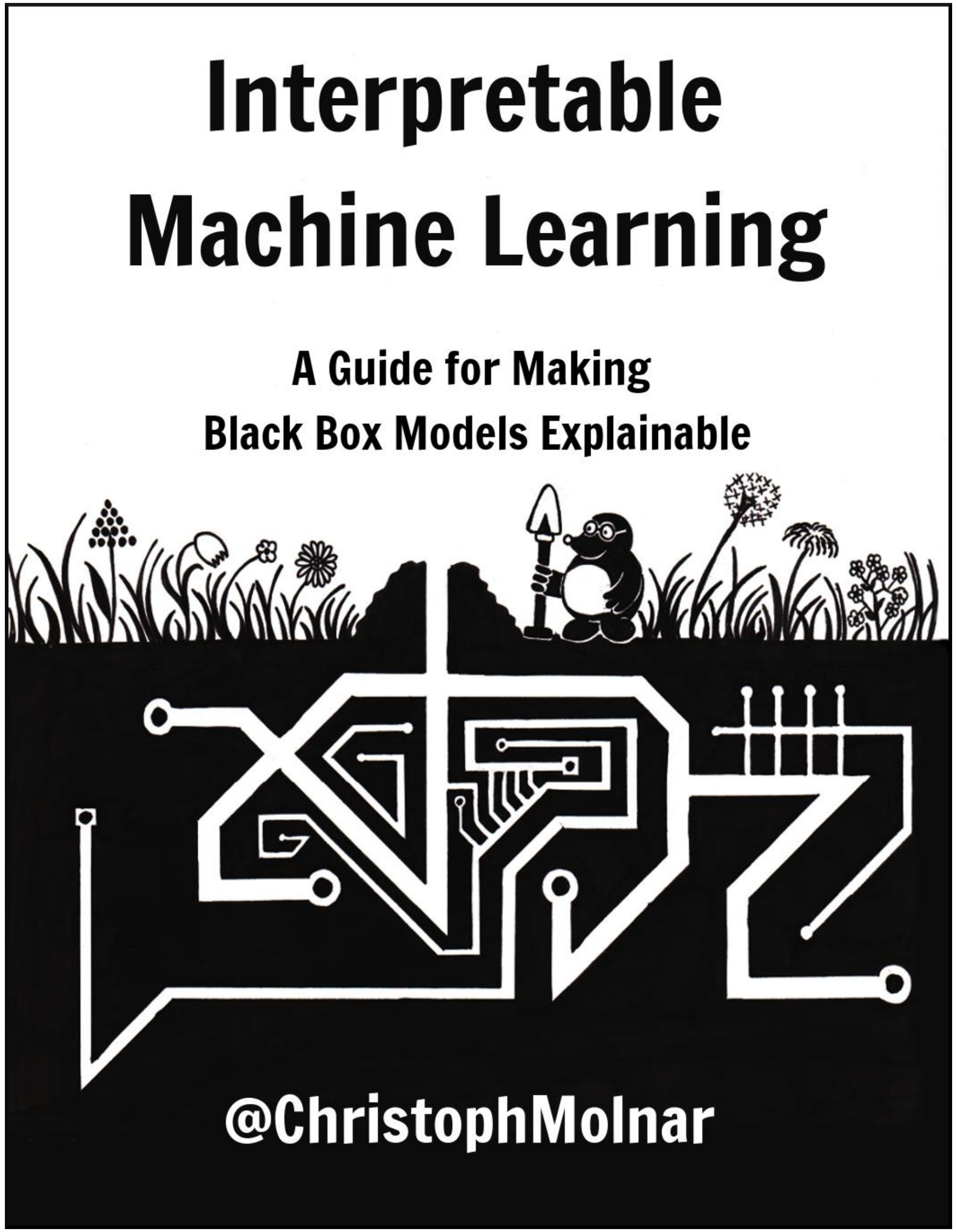
Types of Dogs

Bowtie?

Chain?

Further Reading

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



Research in

Human-Centered XAI

(not in exam!)

NUS National University of Singapore

iHealthtech Institute for Health Innovation & Technology

N-CRiPT NUS Centre for Research in Privacy Technologies

NUS Ubicomp Lab

Apps and Analytics for Smart Cities and Health

https://ubiquitous.comp.nus.edu.sg

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

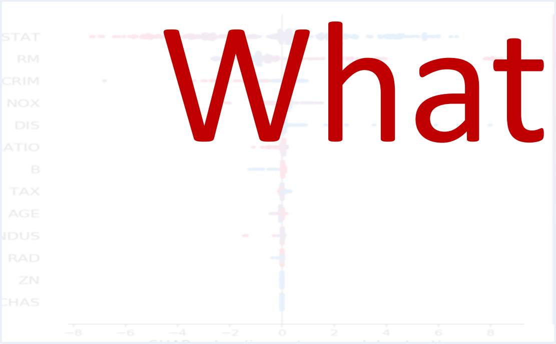
What types of explanation do users need?

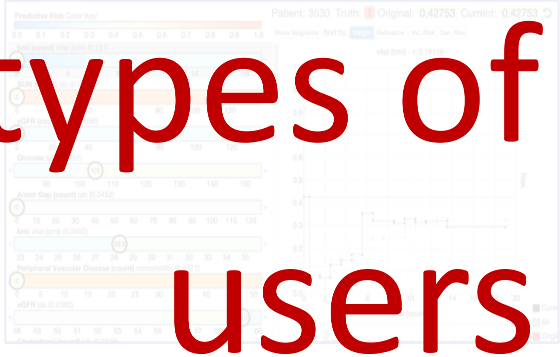
[Ribeiro 2016]

[Zhou 2016]

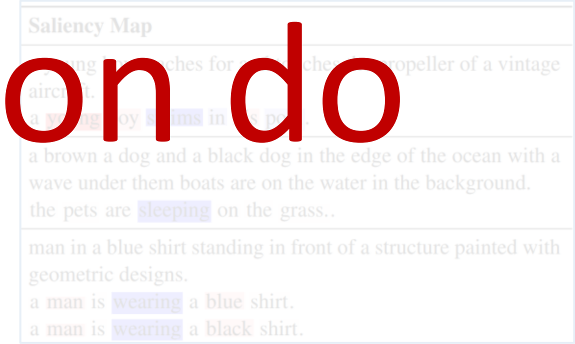
[Lakkaraju 2016]

[Caruana 2016]


[Lundberg 2017]


[Krause 2016]

[Kim 2016]


[Wallace 2018]

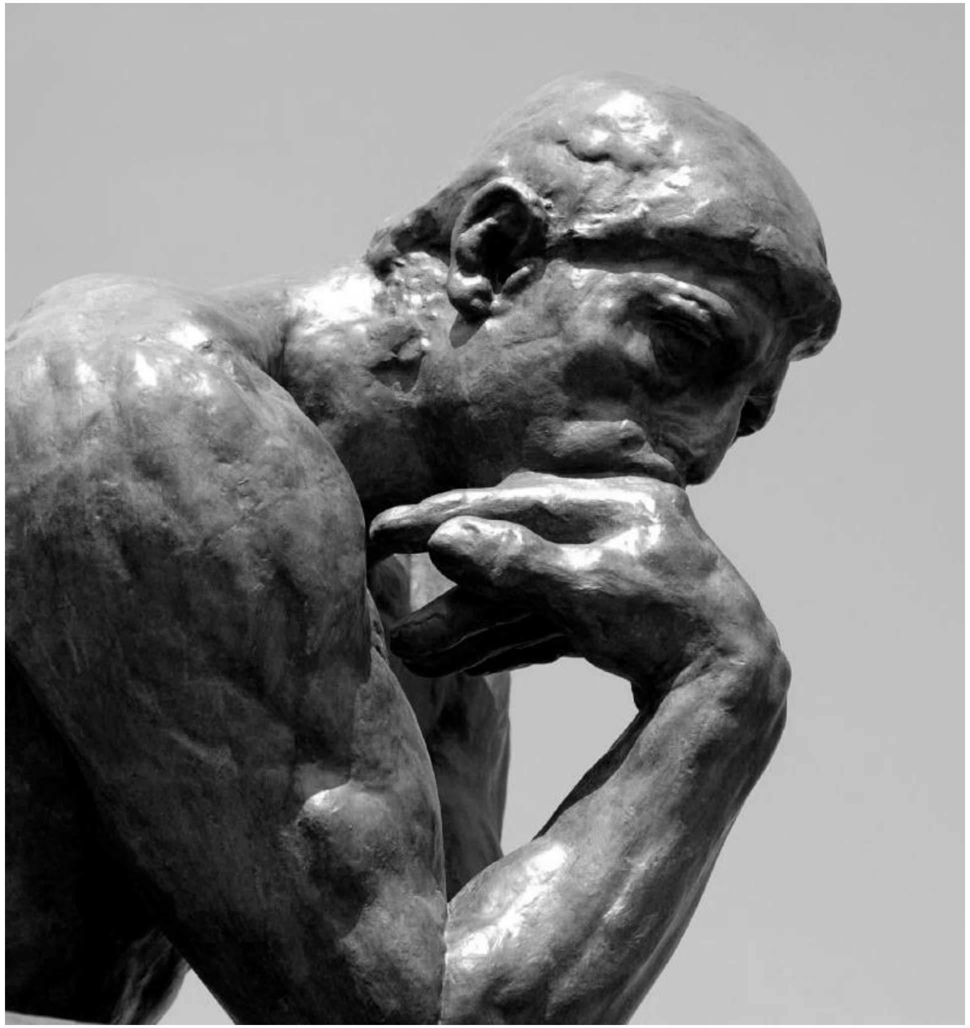
[Cai 2019]

[Koh 2017]

[Kim 2019]

...

How do people think and explain?



Philosophy

Psychology

Human reasoning theories to inform XAI applications

Understanding People	**Informs**	**Explaining AI**
How People should Reason and Explain		How XAI Generates Explanations
How People actually Reason (with Errors)		How XAI Supports Reasoning (and Mitigates Errors)

Wang, D., Yang, Q., Abdul, A., Lim, B. Y. 2019. **Designing Theory-Driven User-Centric Explainable AI.** *CHI '19*.

Human reasoning theories to inform XAI applications

Understanding People

informs

Explaining AI

How People should Reason and Explain

- Inquiry and reasoning
  induction | analogy | deduction
  abduction | hypothetico-deductive model
- Causal explanation and causal attribution
  contrastive | counterfactual | attribution

How People do Reason (with Errors)

- System 1 Heuristic Biases
  representativeness | anchoring | confirmation
- …

How XAI Generates Explanations

- Bayesian probability
  prior | conditional | posterior
- XAI elements
  attribution | name | value | clause
- Visualizations
  tornado plot | saliency heatmap | …

How XAI Mitigate Errors

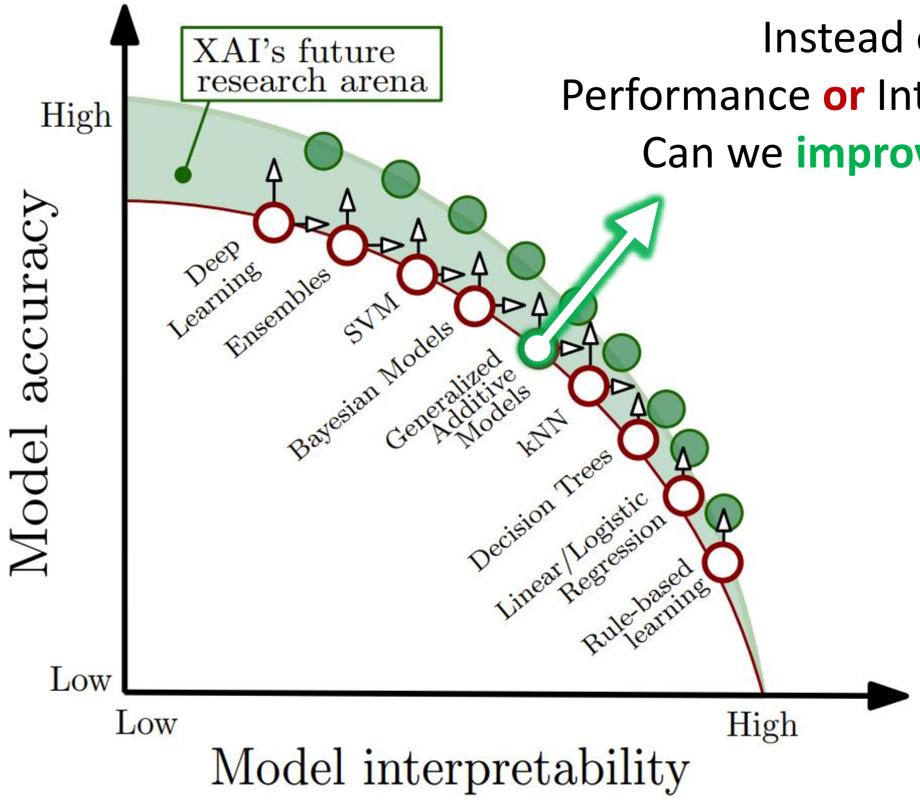
- Mitigate representative bias
  similar prototype | input attributions | contrastive
- Mitigate confirmation bias
  prior probability | input attributions
  1 2

Wang, D., Yang, Q., Abdul, A., Lim, B. Y. 2019. **Designing Theory-Driven User-Centric Explainable AI.** *CHI '19*.

User-Centered Explanations

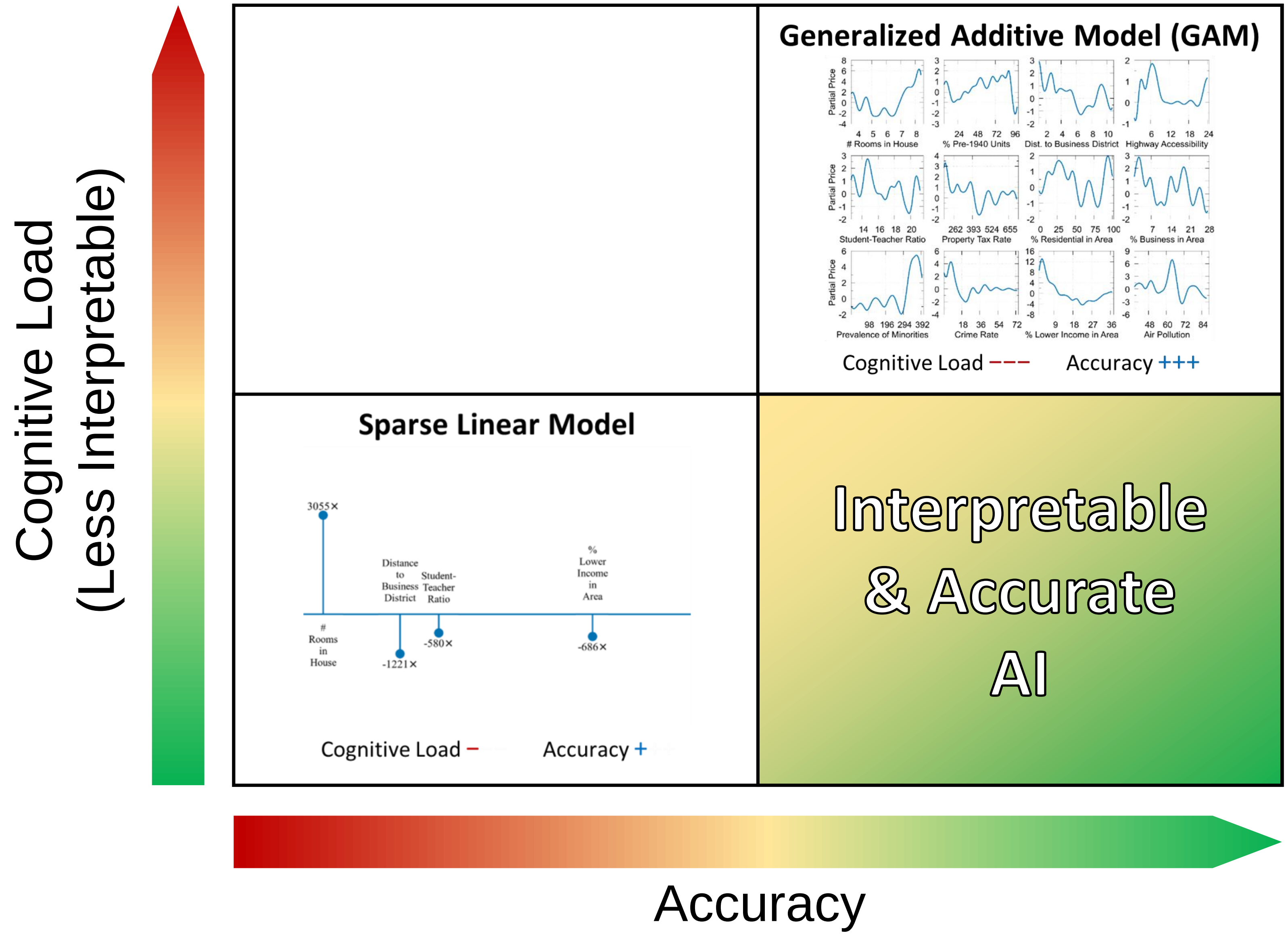
- **Why?** Attribution Explanations
- **Why Not?** Contrastive Explanations
- **How To?** Counterfactual Explanations

Performance-Interpretability Trade-off

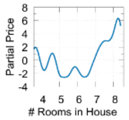


Instead of
Performance **or** Interpretability
Can we **improve both**?

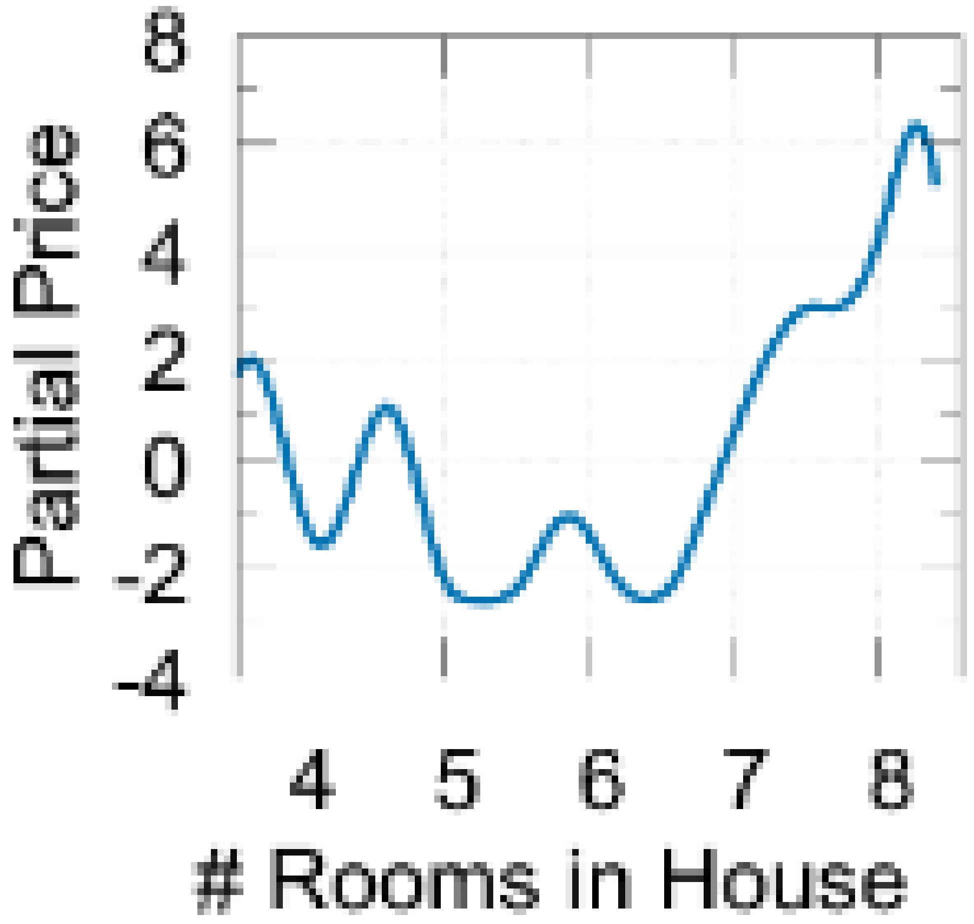
Moderating **cognitive load** in Model Explanations



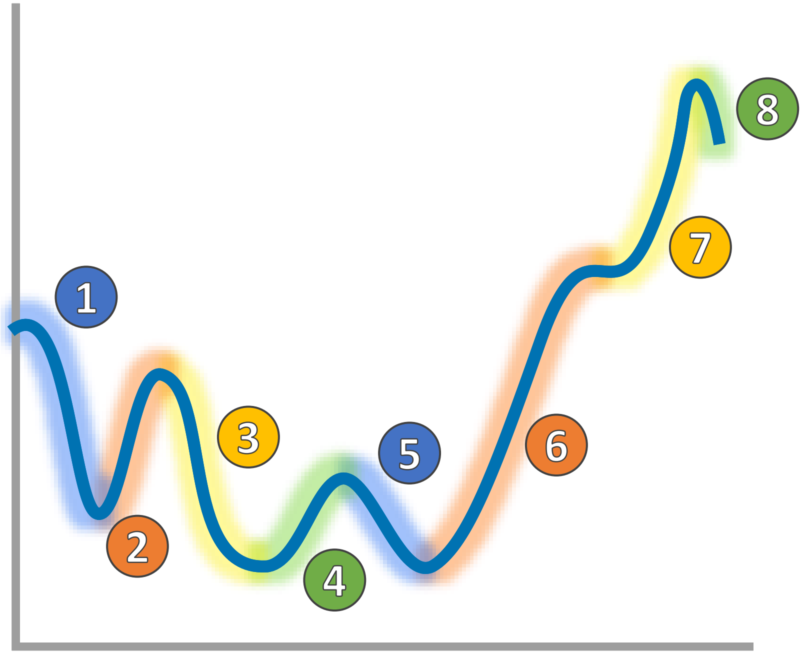
Moderating **cognitive load** in Model Explanations



Moderating **cognitive load** in Model Explanations



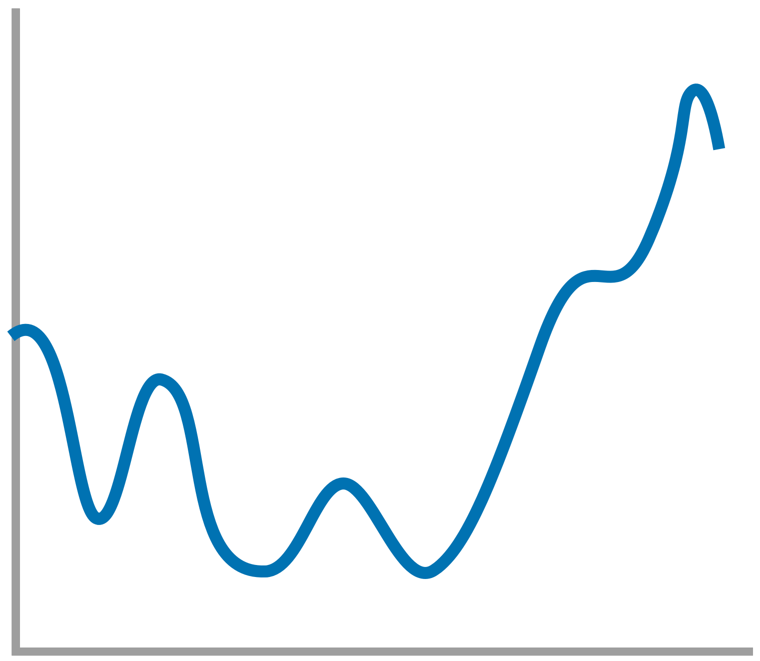
Moderating **cognitive load** in Model Explanations



Cognitive Load = Number of **Visual "Chunks"**

Abdul, A., von der Weth, C., Kankanhalli, M., and Lim, B. Y. 2020.
COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. *CHI '20.*

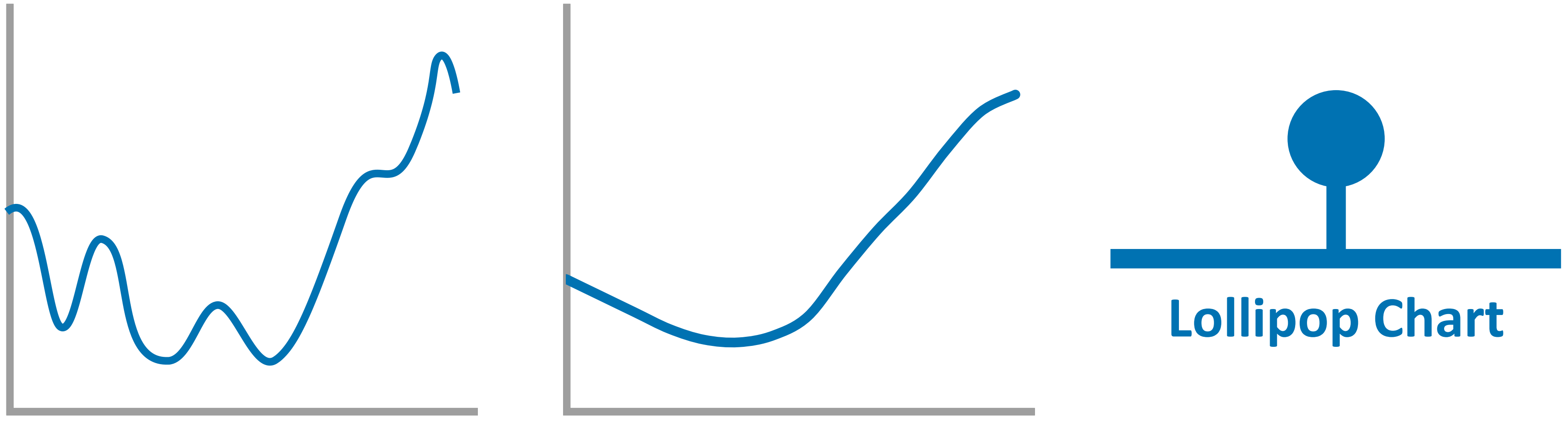
Moderating **cognitive load** in Model Explanations



Cognitive Load = Number of **Visual "Chunks"**

Abdul, A., von der Weth, C., Kankanhalli, M., and Lim, B. Y. 2020.
COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. *CHI '20.*

Moderating **cognitive load** in Model Explanations



8 Chunks | 2 Chunks | 1 Chunk | 0 Chunks

Minimize

Accurate

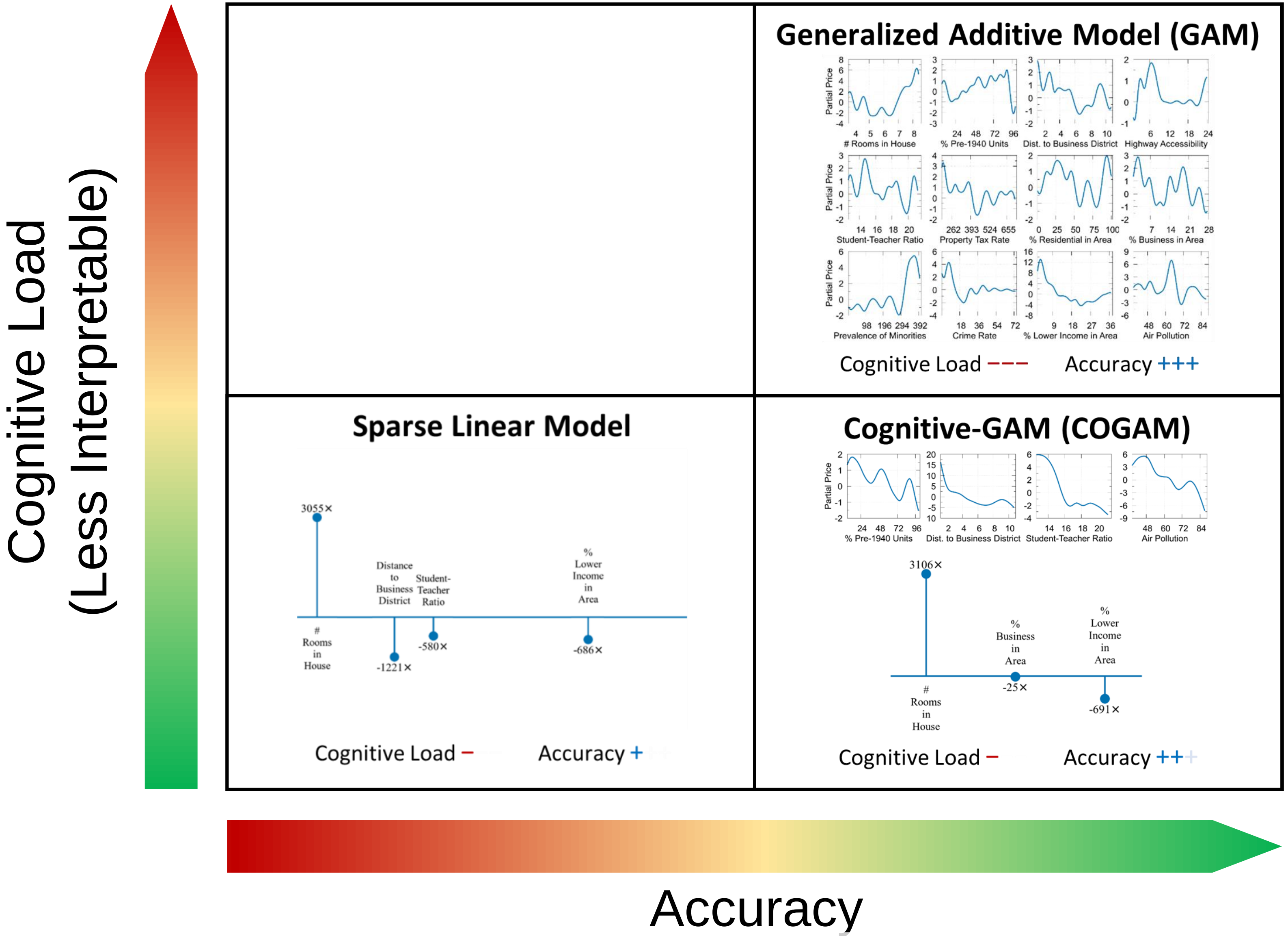
$$\sum_j^n \left(y_i - \sum_i^d f_i\left(x_j^{(i)}\right)\right)^2 + \lambda \sum_i^d \int f_i''(t)^2\, dt$$

Simple

Less Curviness

Abdul, A., von der Weth, C., Kankanhalli, M., and Lim, B. Y. 2020.
COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. *CHI '20.*

Moderating **cognitive load** in Model Explanations



Abdul, A., von der Weth, C., Kankanhalli, M., and Lim, B. Y. 2020.
COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. *CHI '20.*

Wrapping Up

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