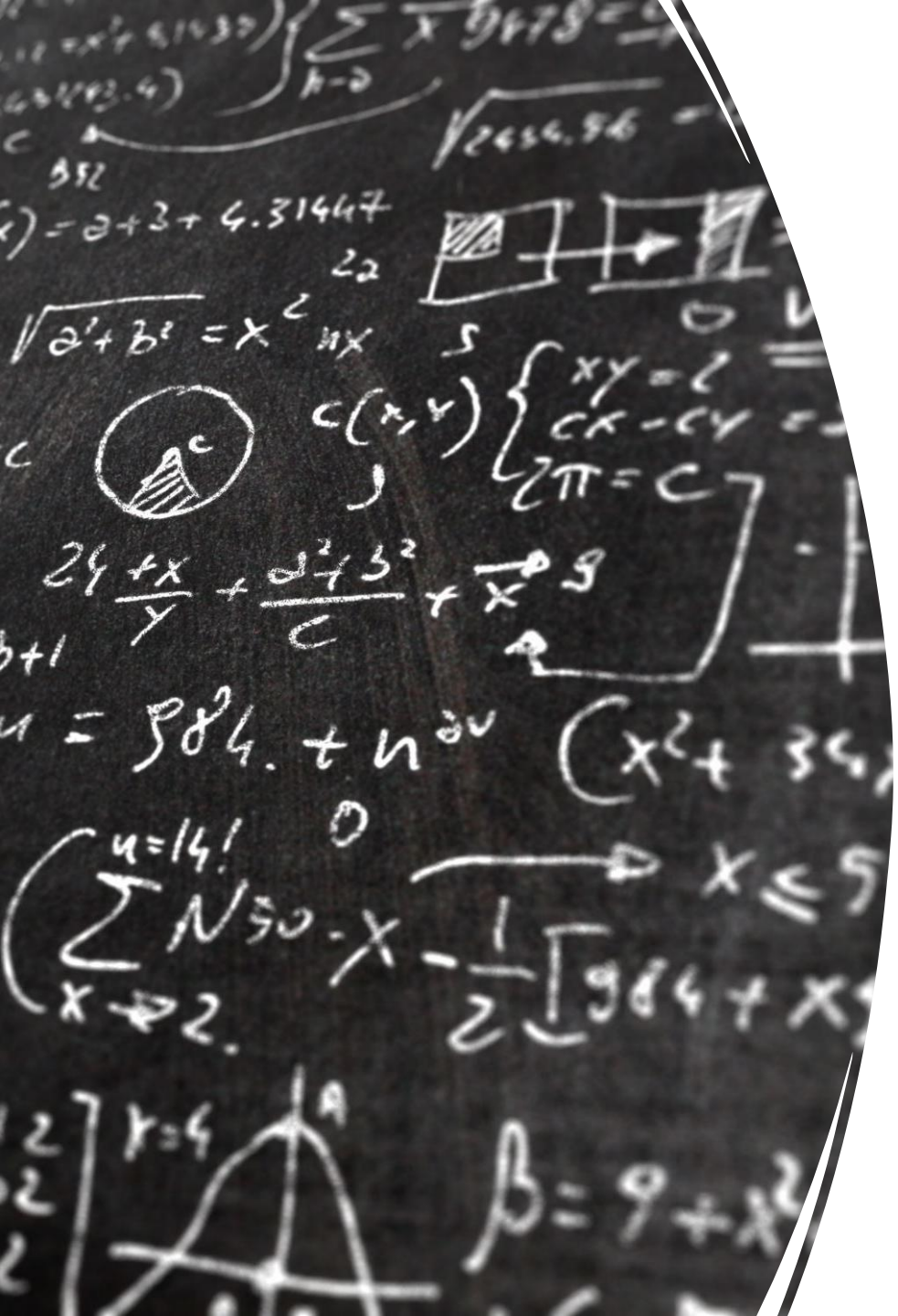


Introduction To Explainable AI (XAI)

Rohan Mitra

Machine Learning Engineer @ Bayut & Dubizzle

Part-time Research Assistant @ CUD

Organizer & Speaker @ Google Developer Groups Sharjah

Case Study: Mount Sinai Hospital

Mount Sinai Hospital

- Wanted to use AI to detect pneumonia from chest X-Rays
- Built a CNN model
  - Trained on lots of X-Rays from their own hospital
  - 158,323 X-Rays used!
  - Made sure to have a proper train-test split
- 99.95% accuracy!!

[1] J. R. Zech *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, Nov. 2018. doi:10.1371/journal.pmed.1002683

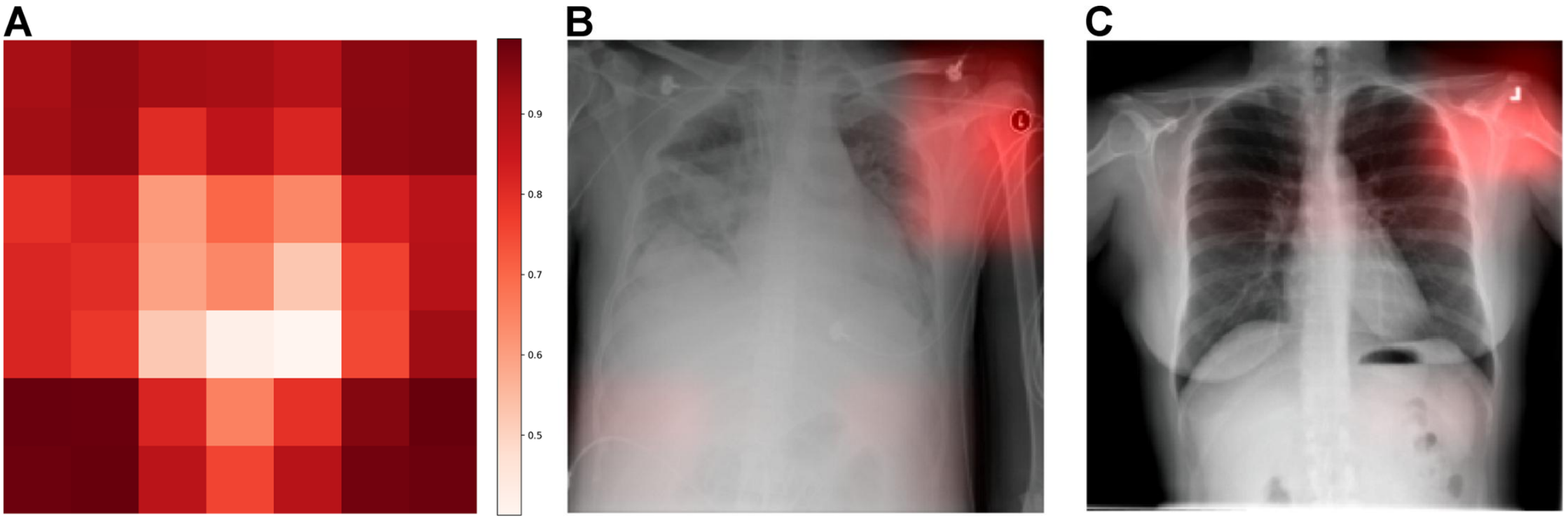
Case Study: Mount Sinai Hospital

- “We want to share this amazing model!”
- Contact National Institutes of Health Clinical Center (NIH)
- They loved it and got it implemented!
- 72% accuracy…

[1] J. R. Zech *et al.*, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,” *PLOS Medicine*, vol. 15, no. 11, Nov. 2018. doi:10.1371/journal.pmed.1002683

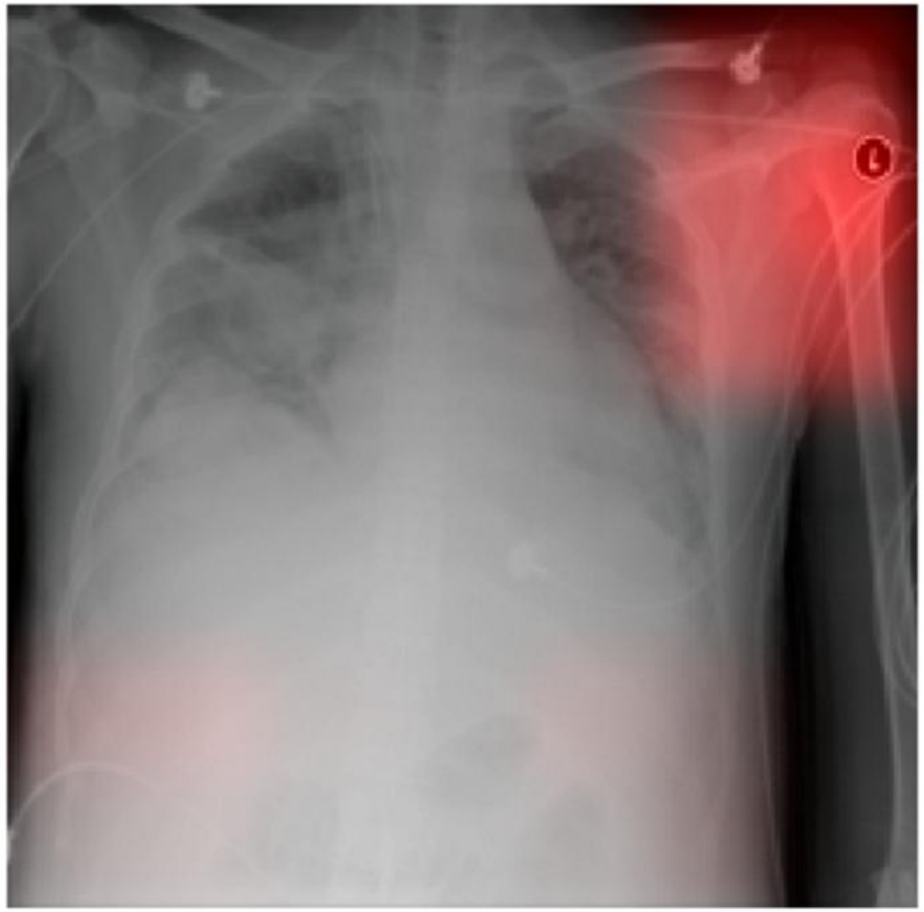
Case Study: Mount Sinai Hospital

- XAI to the rescue!
- Examine activation maps and overlay on the image



[1] J. R. Zech *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, Nov. 2018. doi:10.1371/journal.pmed.1002683

Case Study: Mount Sinai Hospital



- Mount Sinai physicians put a metal physical marker on the patients' shoulders!
- The CNN took a shortcut!

[1] J. R. Zech *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, Nov. 2018. doi:10.1371/journal.pmed.1002683

Impact of XAI

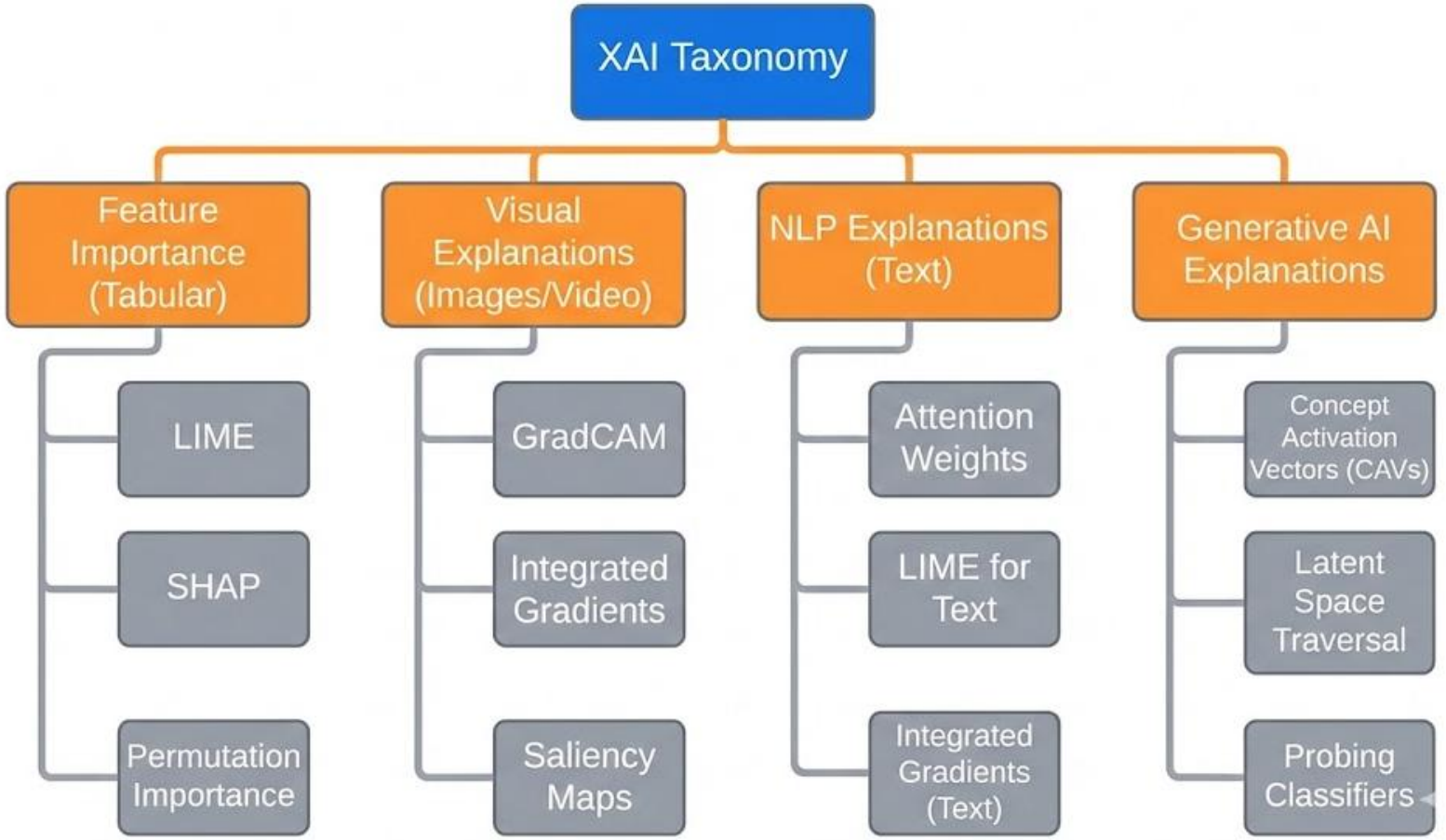
Trust & Transparency (External)

- **Focused on Users:** Doctors, Regulators, Buyers
- Users demand explanations
- High stakes predictions: Medicine, Finance, Law
- Ethical introspection

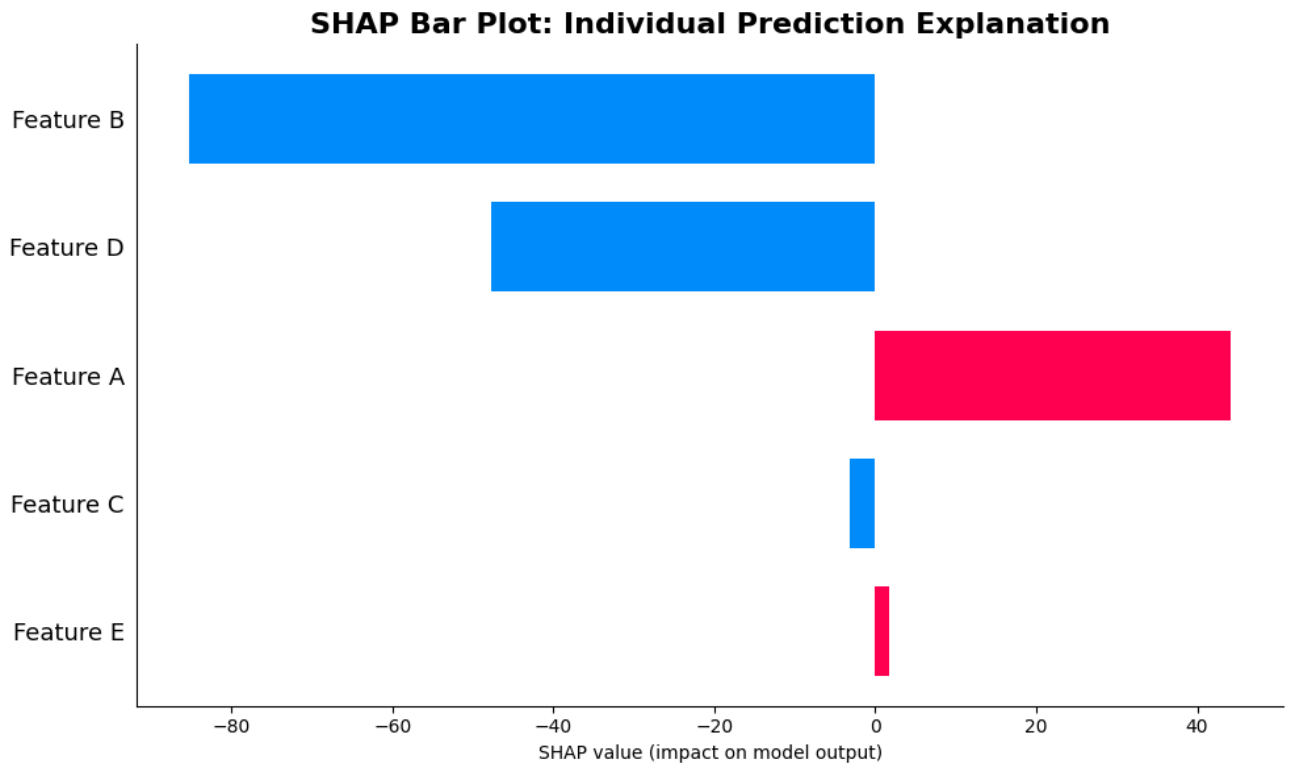
Debugging & Model Hygiene (Internal)

- **Focused on Builders:** Data Scientists, Machine Learning Engineers, YOU!
- Treated as a unit-test to ensure model learns useful features
- Models tend to find shortcuts
- Avoid spurious correlations

Taxonomy


XAI Taxonomy
Feature Importance (Tabular)
Visual Explanations (Images/Video)
NLP Explanations (Text)
Generative AI Explanations
LIME
SHAP
Permutation Importance
GradCAM
Integrated Gradients
Saliency Maps
Attention Weights
LIME for Text
Integrated Gradients (Text)
Concept Activation Vectors (CAVs)
Latent Space Traversal
Probing Classifiers

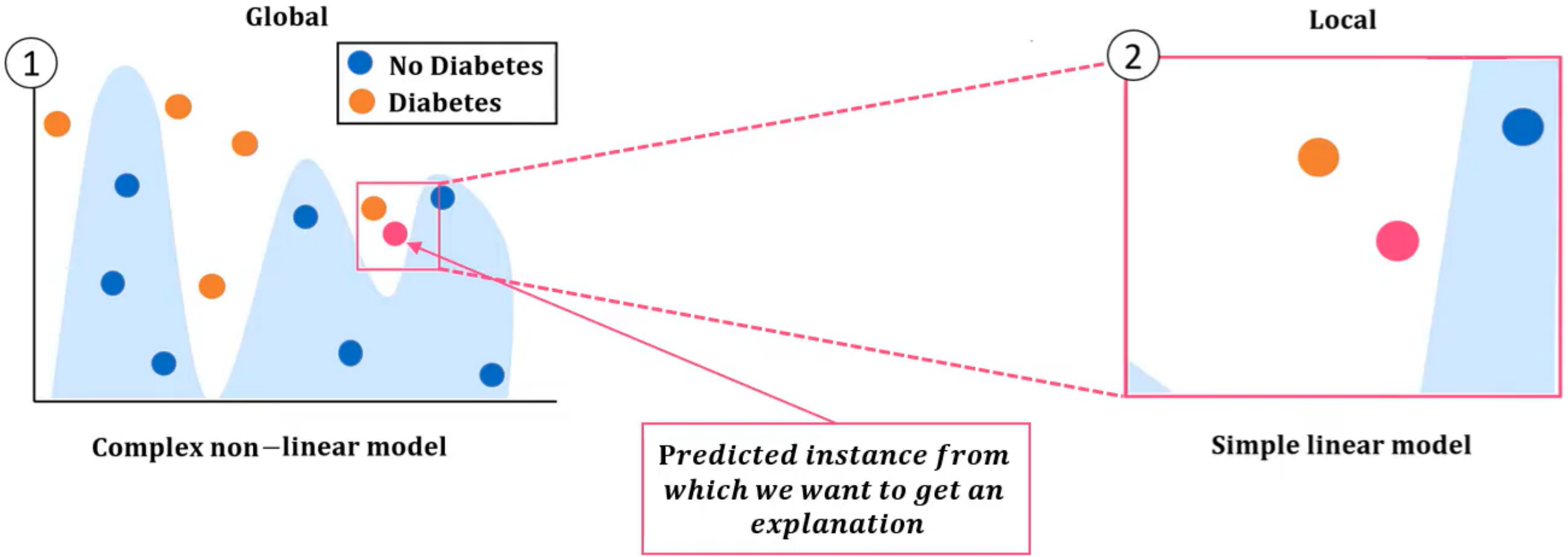
XAI Method: SHAP



- **Origin:** Based on Cooperative Game Theory (Lloyd Shapley, Nobel Prize)
- For a given prediction, much did each feature contribute to the final prediction?
- Calculates the **Marginal Contribution** of a feature across all possible combinations of features
- **Global AND Local:**
  - *Local:* Why did **this** specific person get denied a loan?
  - *Global:* What drives the model generally?

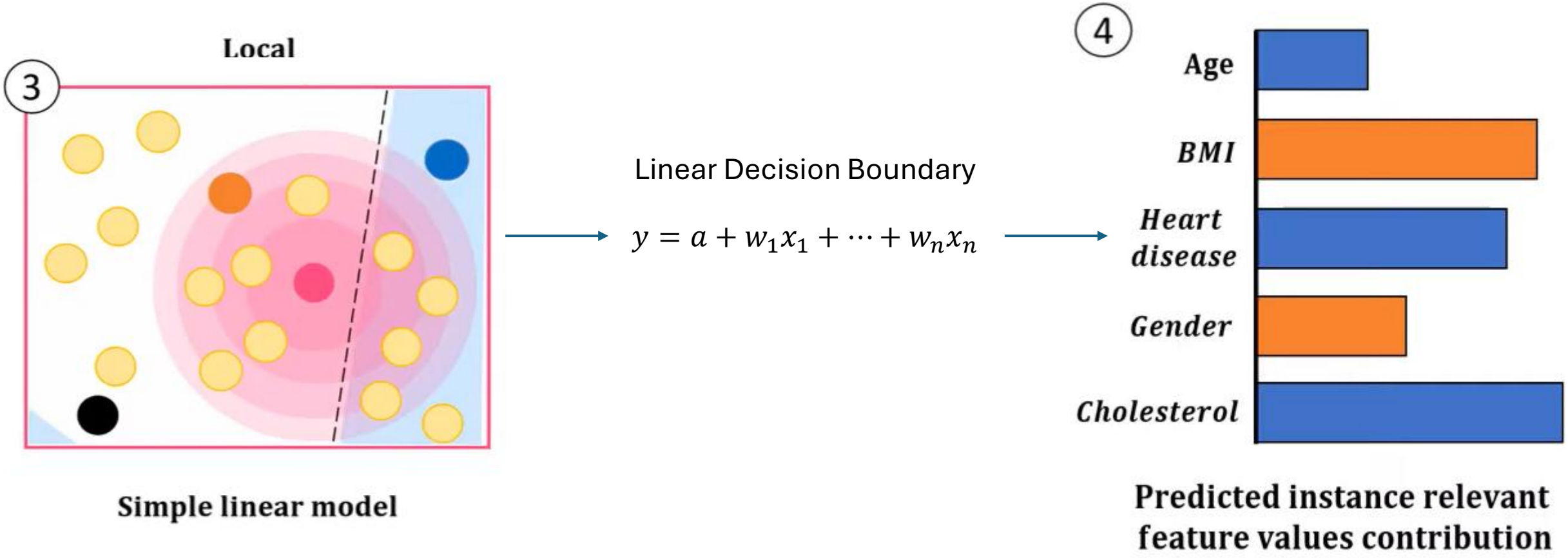
XAI Method: LIME

- **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
- Approximate the global decision boundary with linear approximations near a point
- We don't need to understand how the entire model works, just how it behaves near this particular prediction



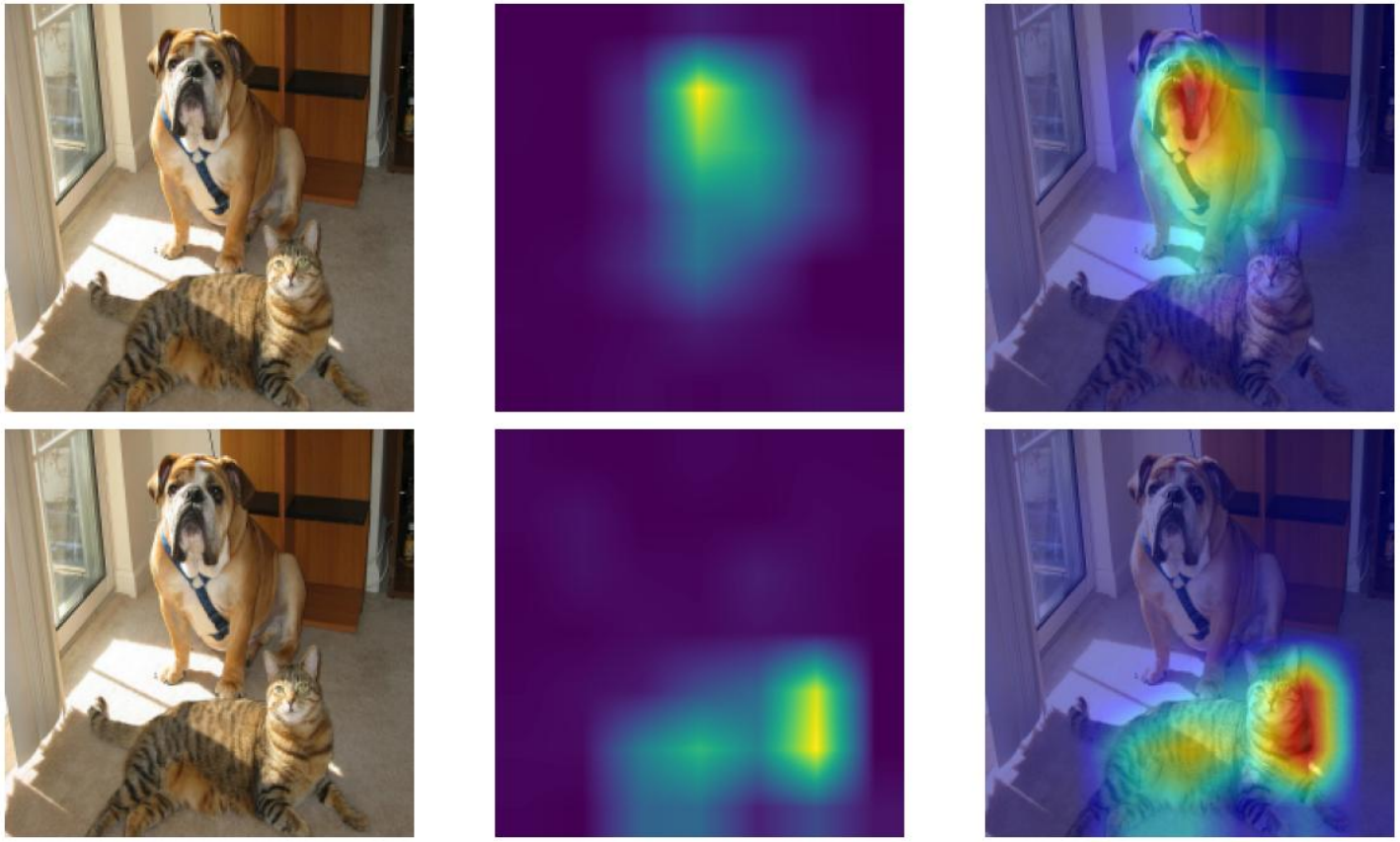
XAI Method: LIME

- **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
- Approximate the global decision boundary with linear approximations near a point
- We don't need to understand how the entire model works, just how it behaves near this particular prediction

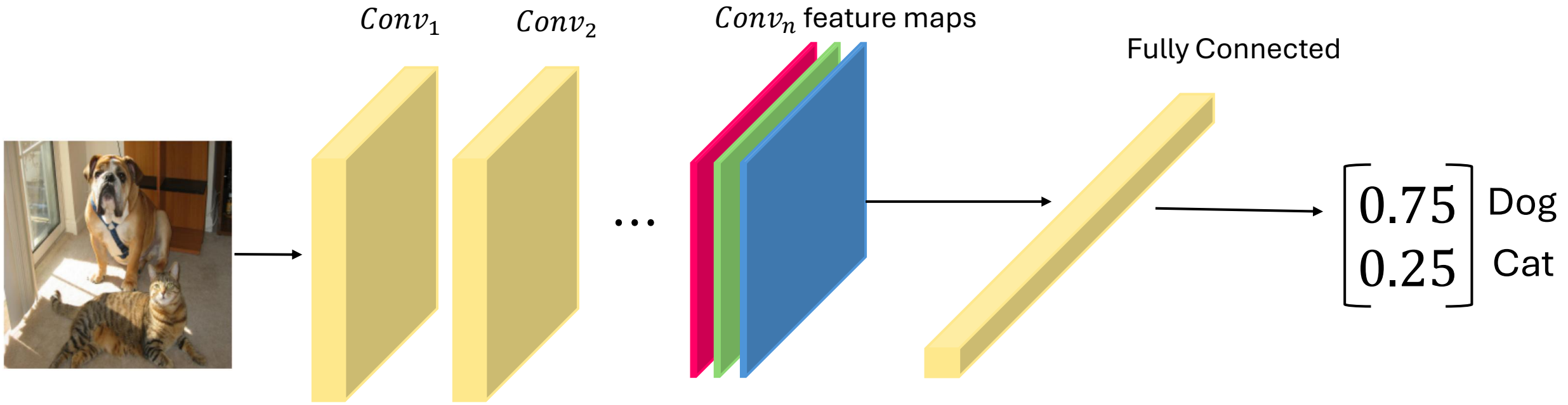


XAI Method: Grad-CAM

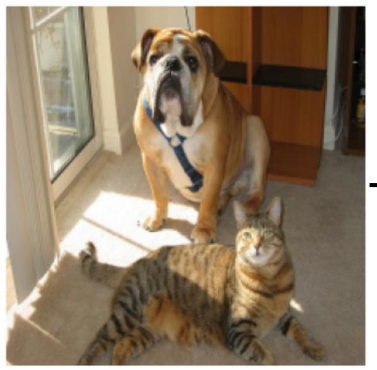
- **Grad**ient-weighted **C**lass **A**ctivation **M**apping
- For image/video data
- Focuses on inner workings of the models
  - Looks at how the final class probability changes with respect to the features changing
- Generates a heatmap of the areas in the image that were important to determine the class



XAI Method: Grad-CAM



XAI Method: Grad-CAM



$Conv_1$ $Conv_2$ ... $Conv_n$ feature maps

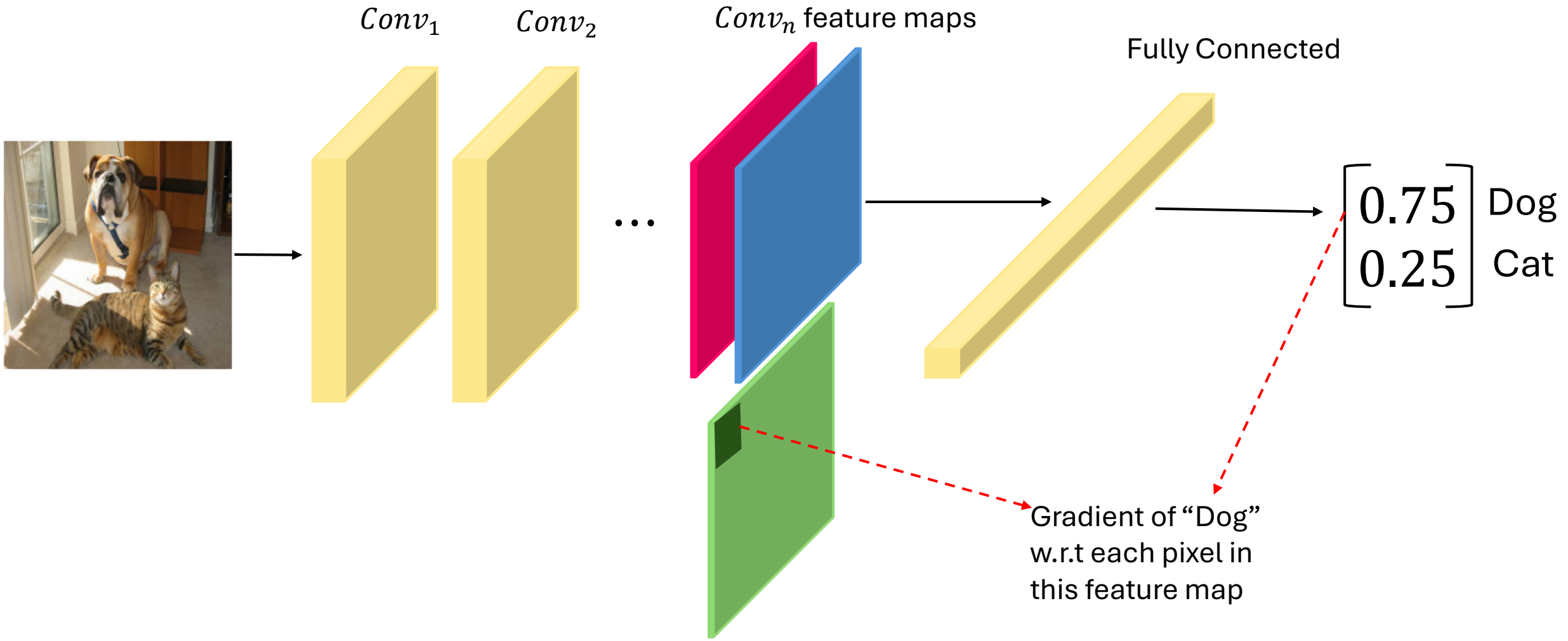
Fully Connected

$$\begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix} \begin{matrix} \text{Dog} \\ \text{Cat} \end{matrix}$$

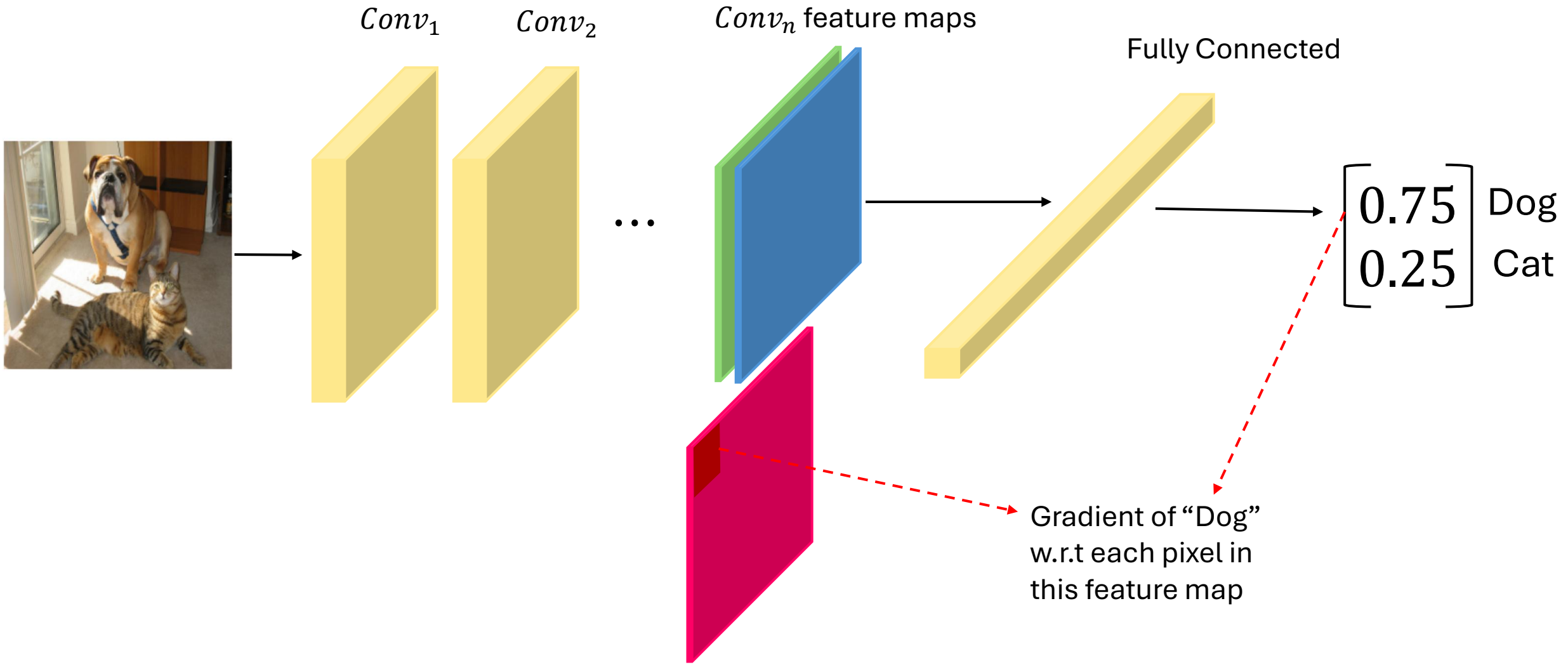
Gradient of “Dog” w.r.t each pixel in this feature map

How much does the probability of “Dog” change when this pixel changes?

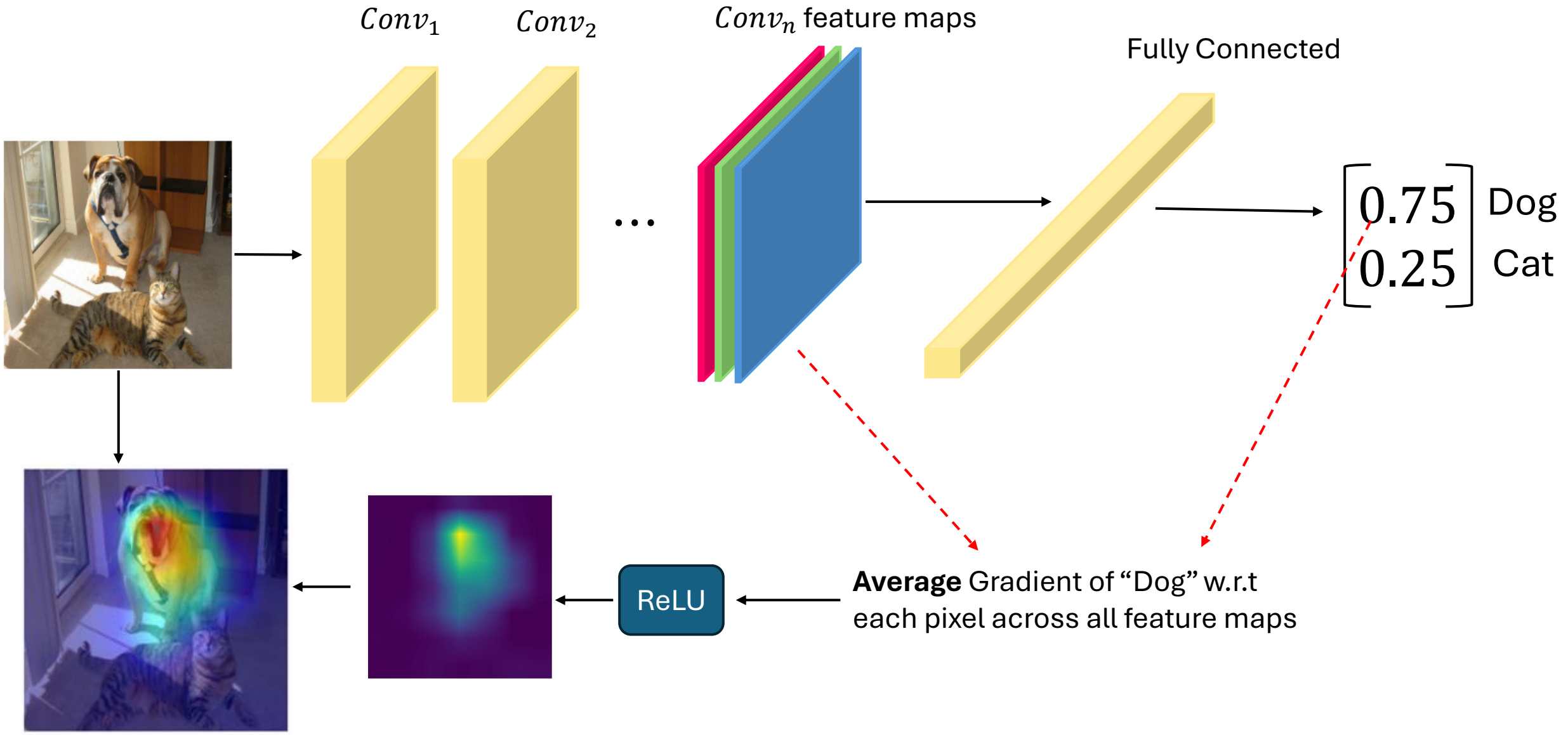
XAI Method: Grad-CAM



XAI Method: Grad-CAM



XAI Method: Grad-CAM



Let’s see some code!

github.com/ro1406/xai-tutorials

Thank You!

My Website

rohanmitra.dev

linkedin.com/in/rohan-mitra14/