NeuralDivergence

Georgia Tech

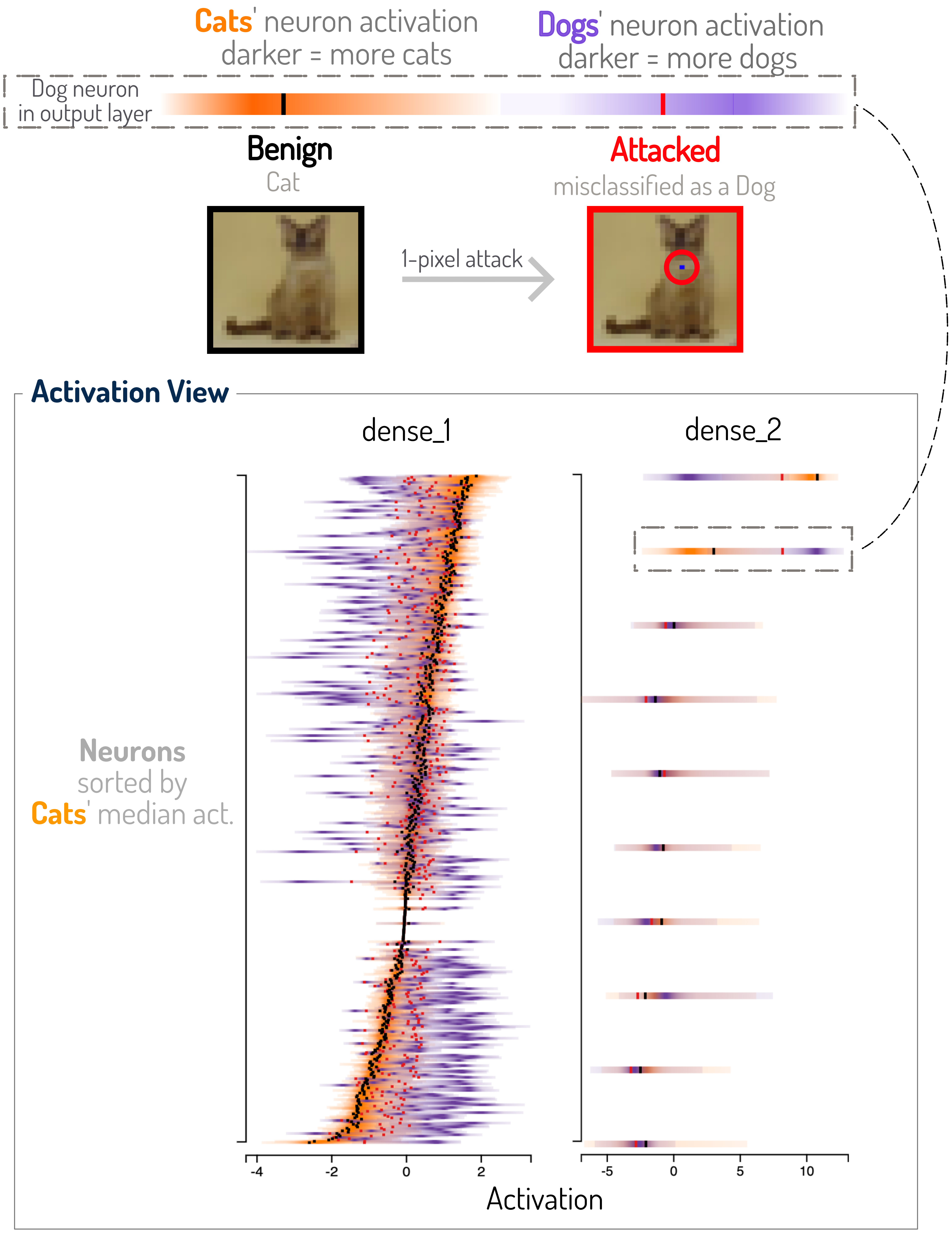
Exploring and Understanding Neural Networks by Comparing Activation Distributions

Haekyu Park **Fred Hohman** **Polo Chau**
haekyu.com / haekyu@gatech.edu

Try at:
haekyu.com

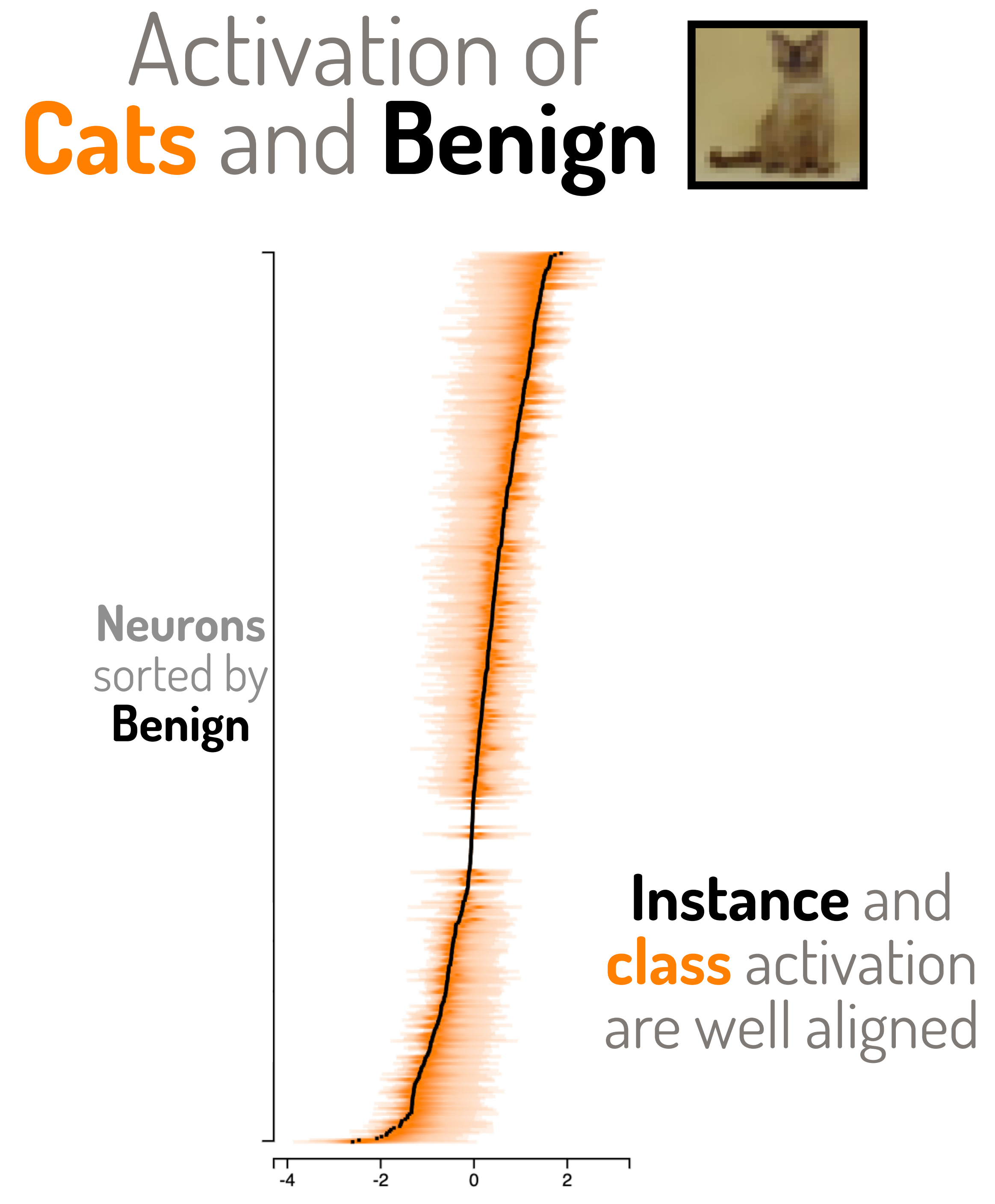
Summary

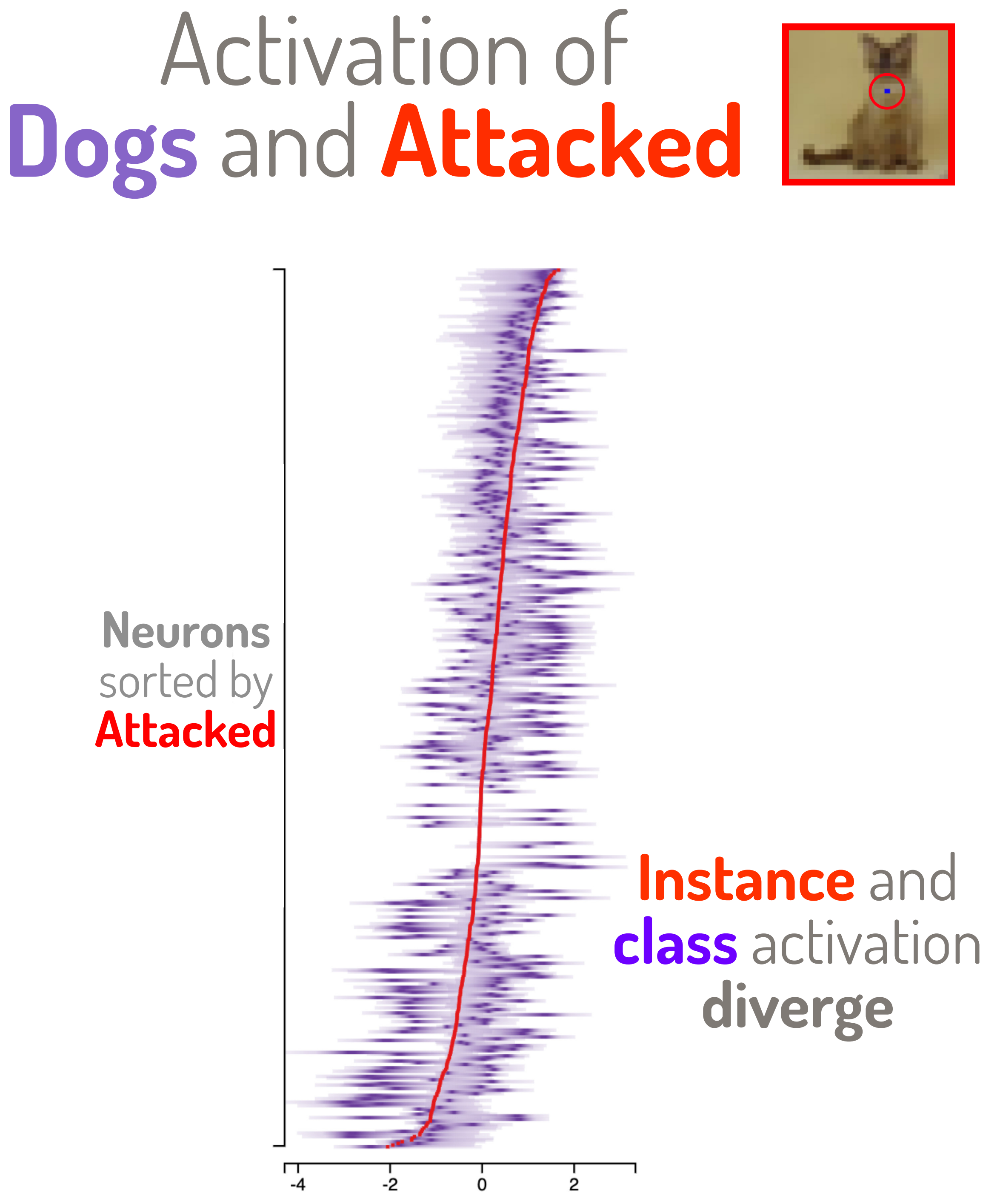
To **understand** neural networks, NeuralDivergence enables users to explore the models through interactive **summarization** of all neuron activation distribution and **comparison** across layers, classes, and instances (e.g. pairs of adversarial attacked and benign images).



Scenario: Deciphering Attack on Deep Learning Model

We can use the amount of "**neural divergence**" between an image and its predicted class **to detect one-pixel attack**. The example below shows that an attacked cat image (misclassified as dog) significantly diverges from the "norm" of the real dog class.





This work was supported by NSF grants CNS-1704701, TWC-1526254, IIS-1563816, the ISTC-ARSA Intel gift, and a NASA Space Technology Research Fellowship.