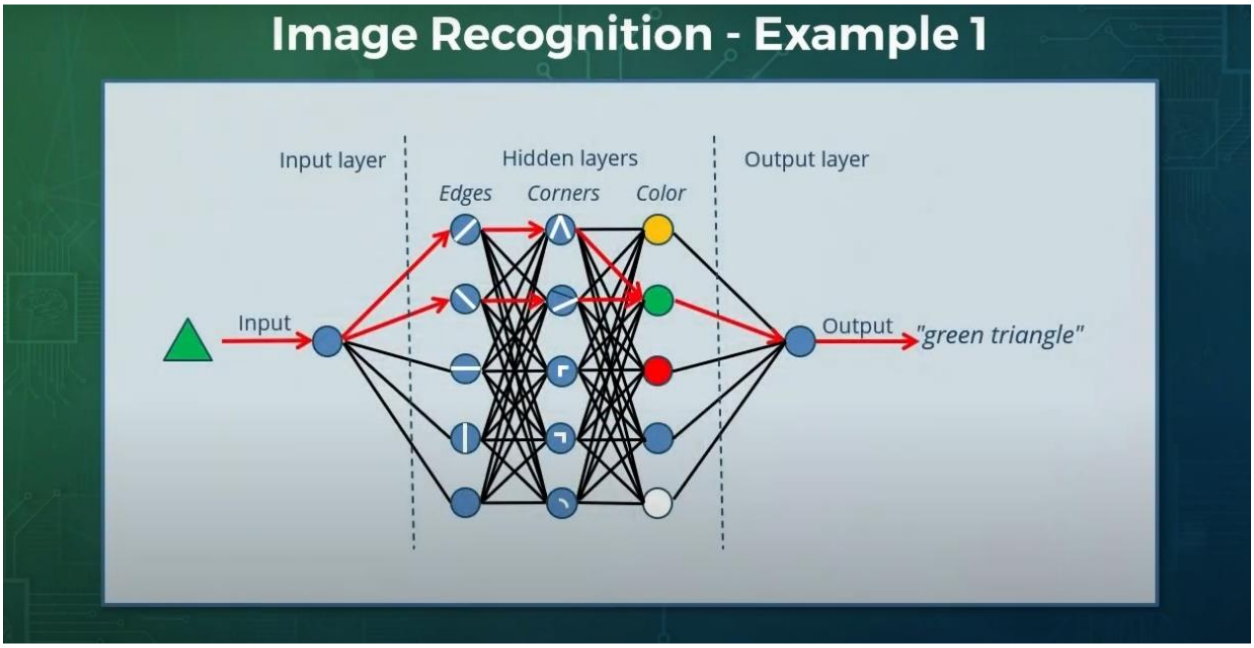
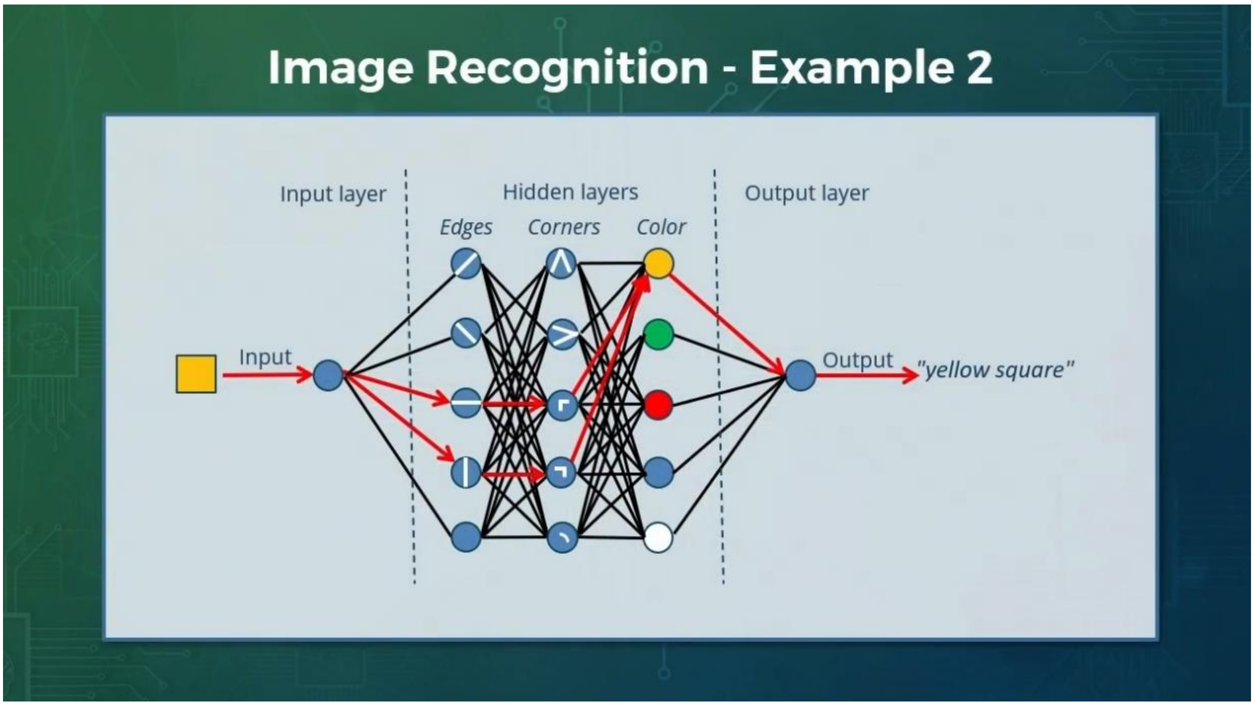
THE INTERPRETABILITY PROBLEM





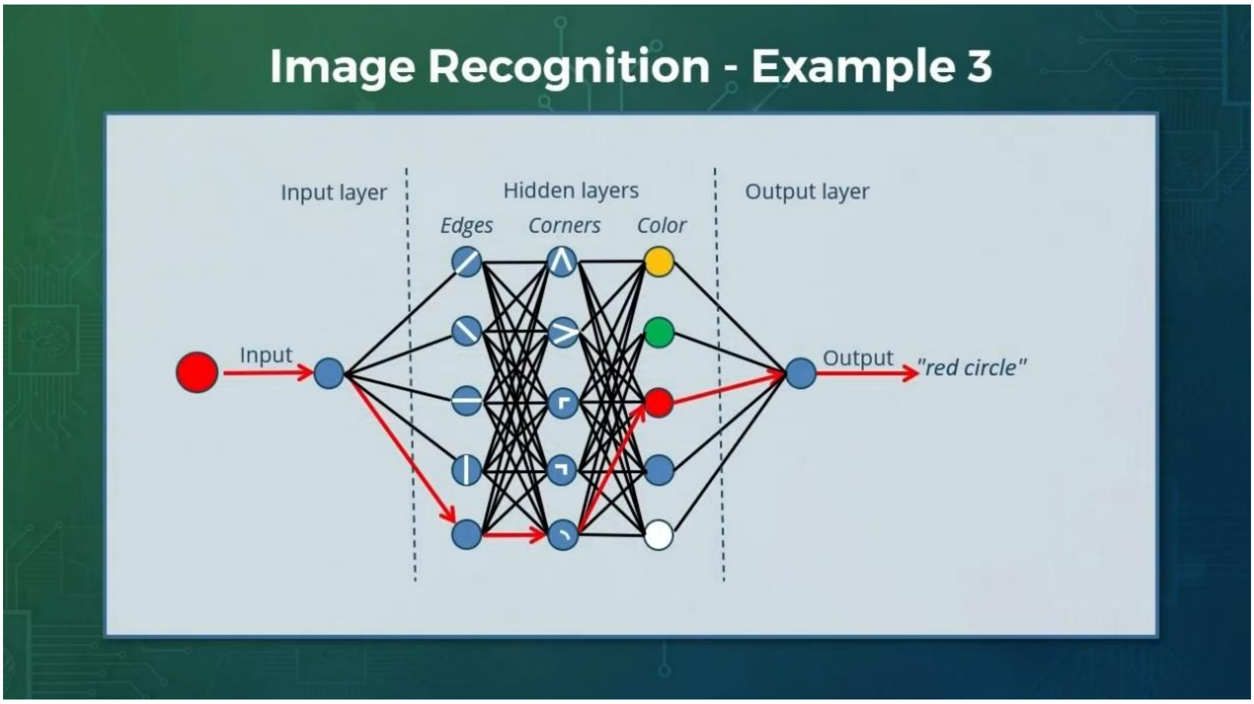
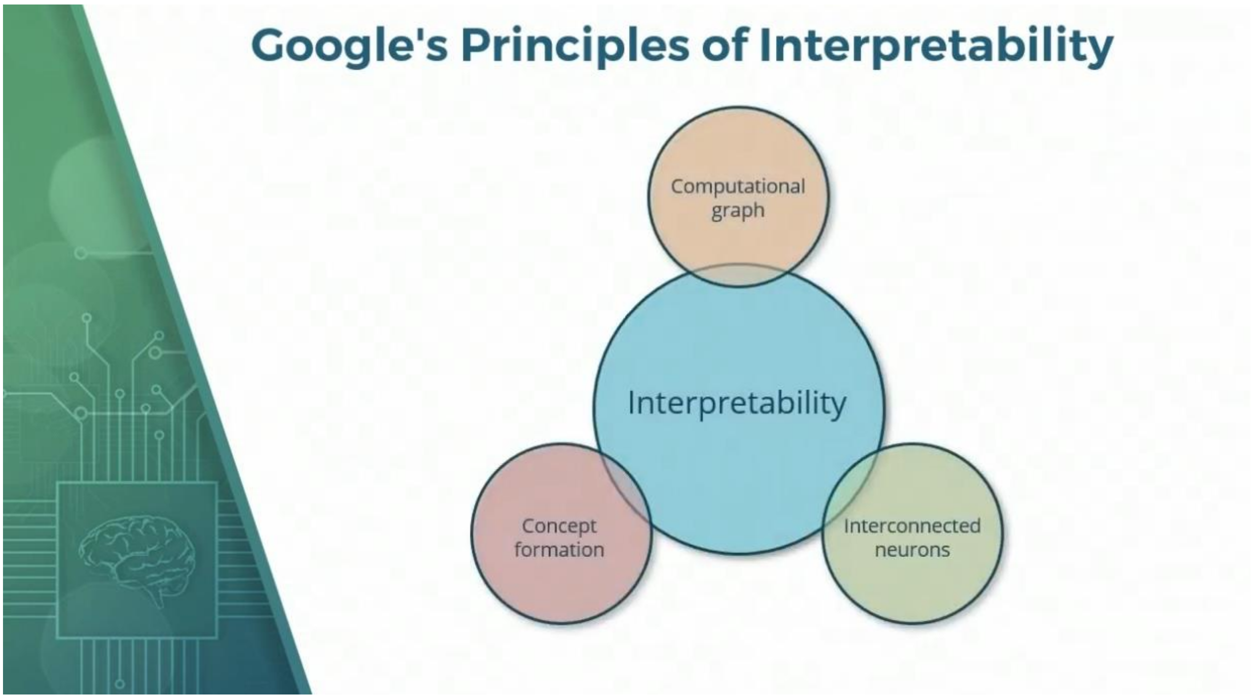
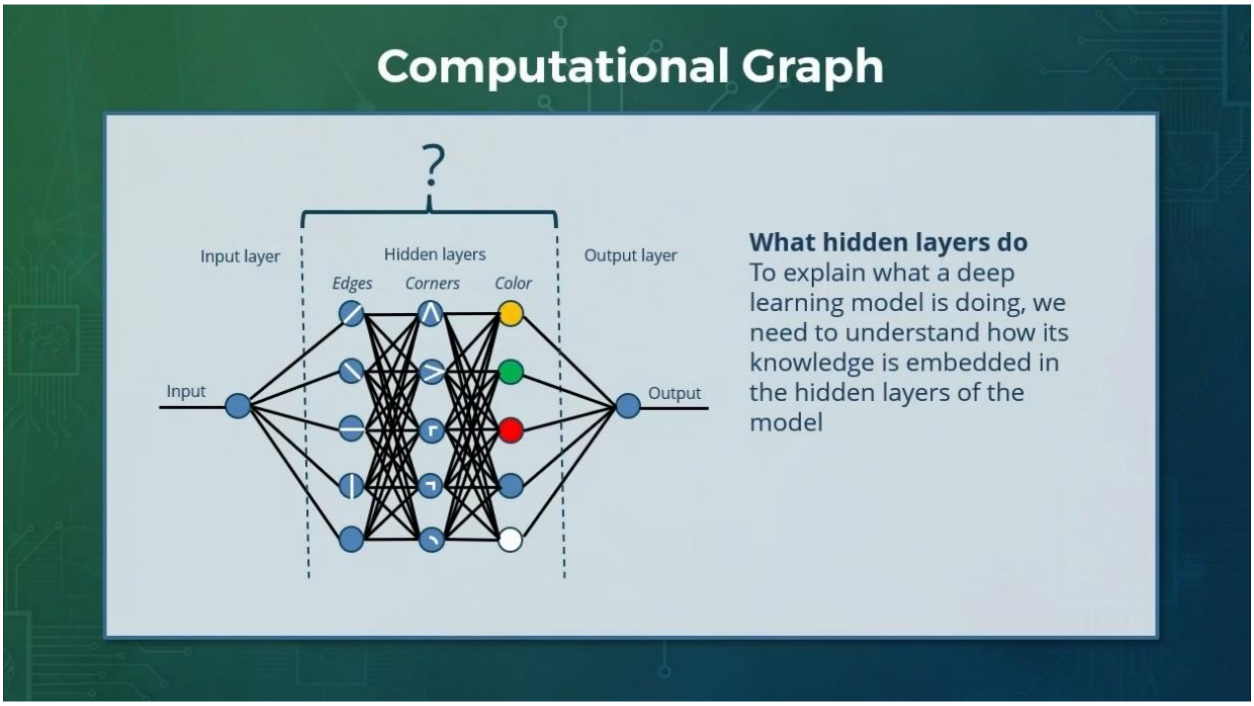


Image Recognition - Example 3



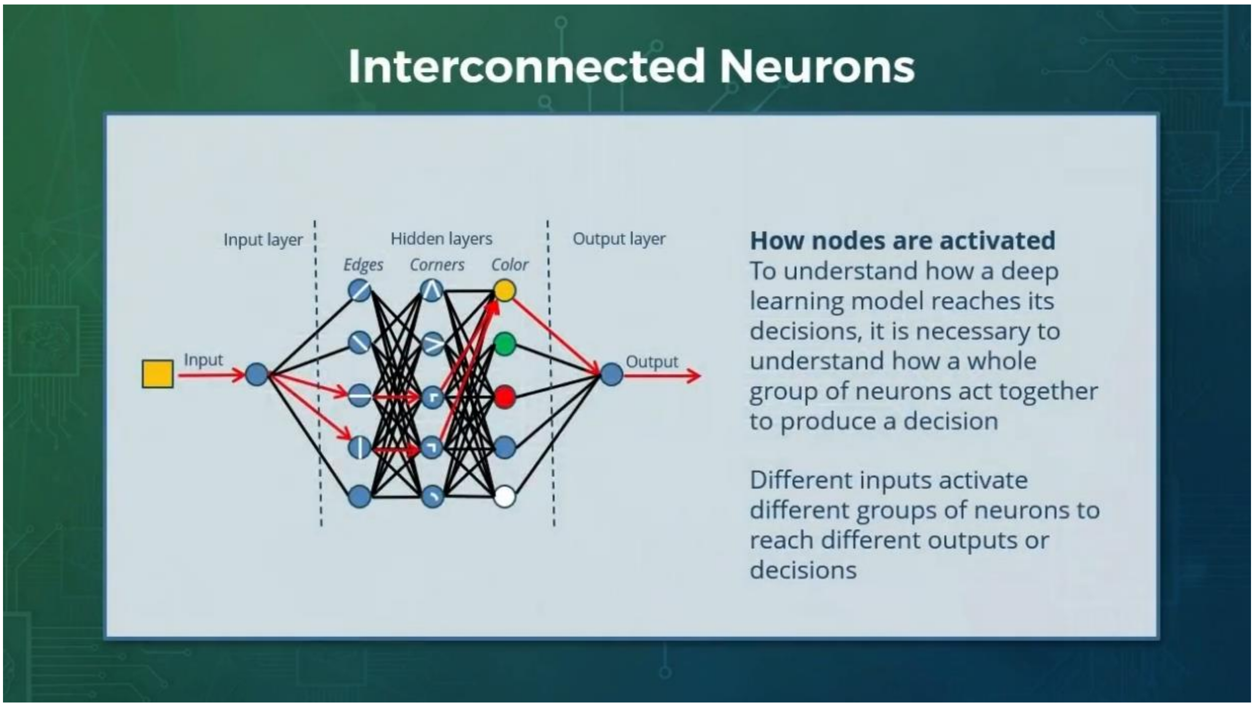
Google's Principles of Interpretability



Computational Graph

What hidden layers do

To explain what a deep learning model is doing, we need to understand how its knowledge is embedded in the hidden layers of the model

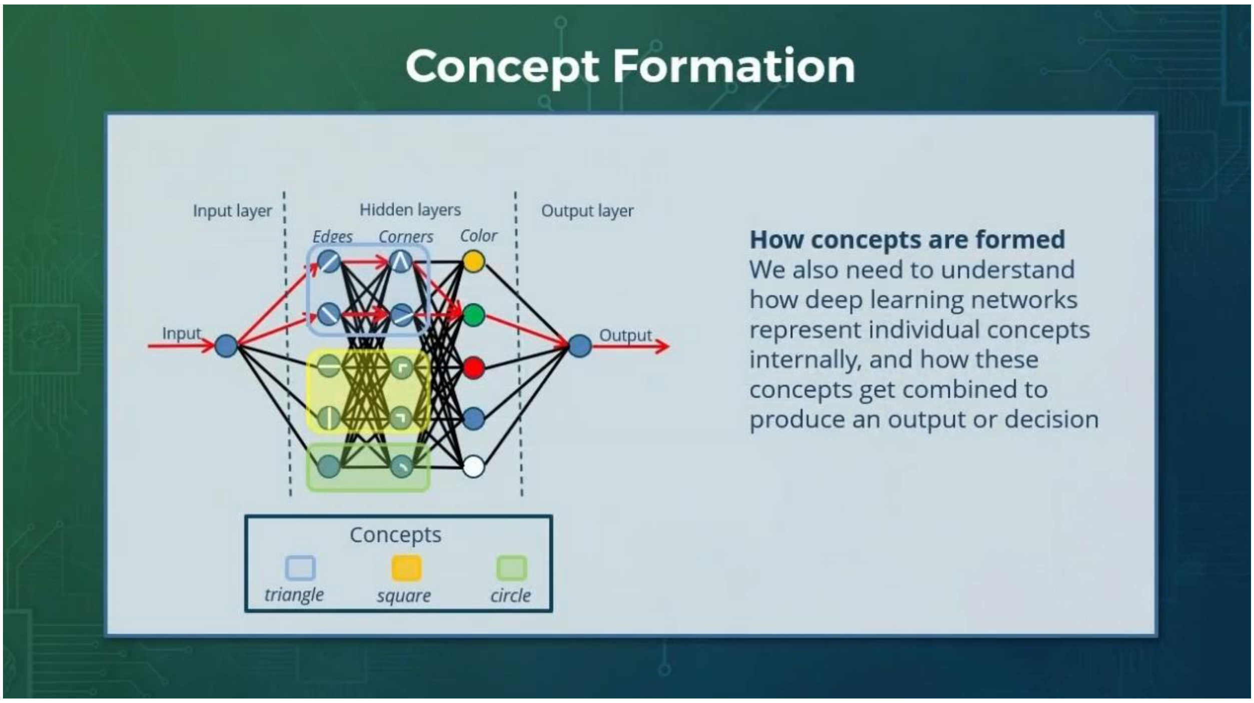


Interconnected Neurons

How nodes are activated

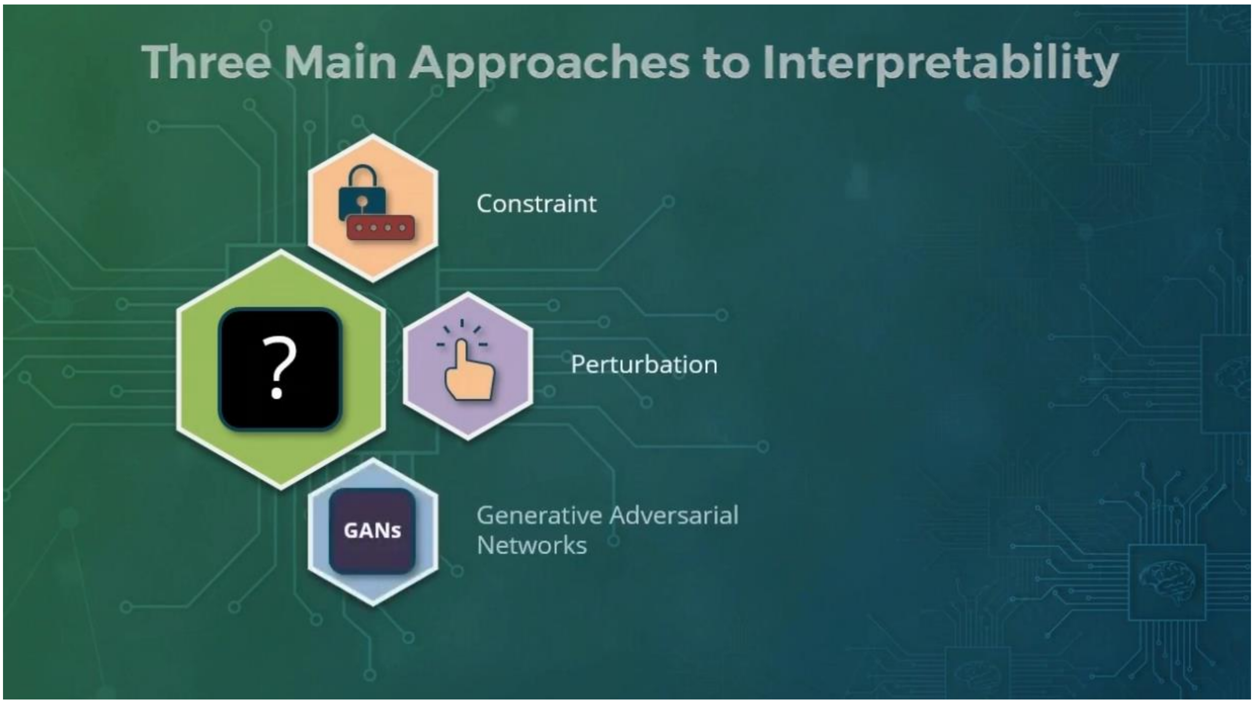
To understand how a deep learning model reaches its decisions, it is necessary to understand how a whole group of neurons act together to produce a decision

Different inputs activate different groups of neurons to reach different outputs or decisions



Concept Formation

How concepts are formed
We also need to understand how deep learning networks represent individual concepts internally, and how these concepts get combined to produce an output or decision

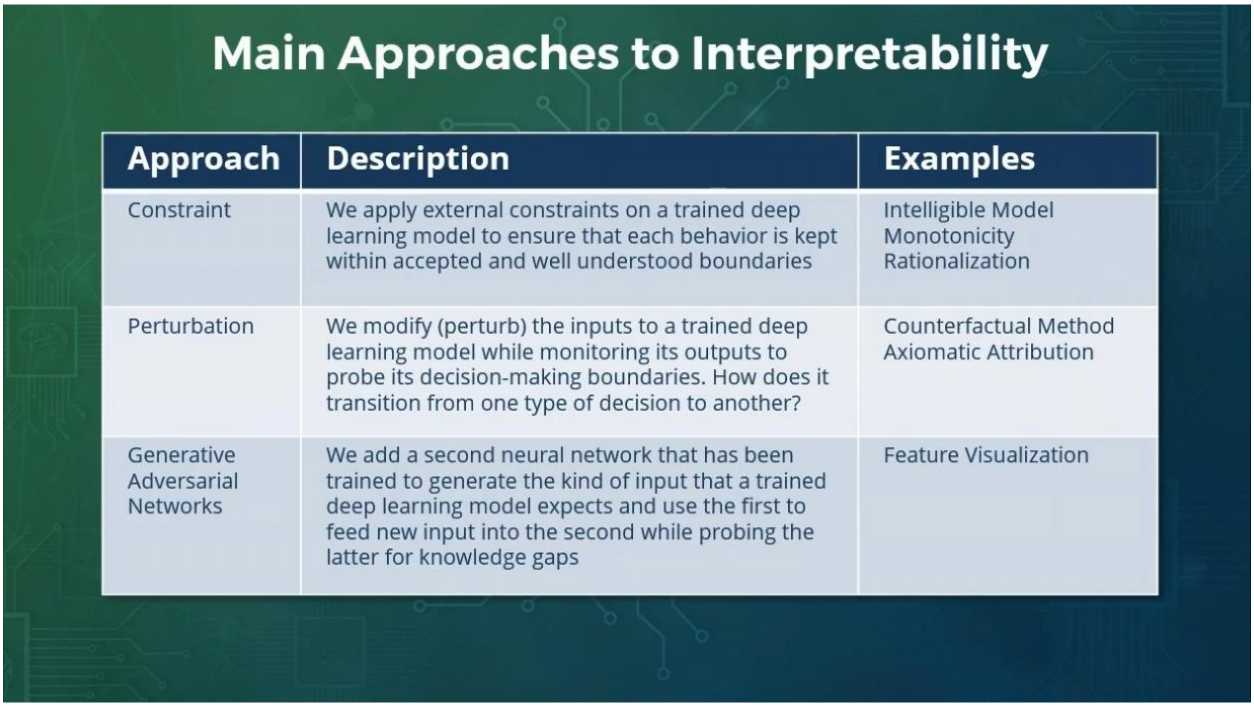


Three Main Approaches to Interpretability

Constraint

Perturbation

Generative Adversarial Networks



Main Approaches to Interpretability

Approach	Description	Examples
Constraint	We apply external constraints on a trained deep learning model to ensure that each behavior is kept within accepted and well understood boundaries	Intelligible Model Monotonicity Rationalization
Perturbation	We modify (perturb) the inputs to a trained deep learning model while monitoring its outputs to probe its decision-making boundaries. How does it transition from one type of decision to another?	Counterfactual Method Axiomatic Attribution
Generative Adversarial Networks	We add a second neural network that has been trained to generate the kind of input that a trained deep learning model expects and use the first to feed new input into the second while probing the latter for knowledge gaps	Feature Visualization