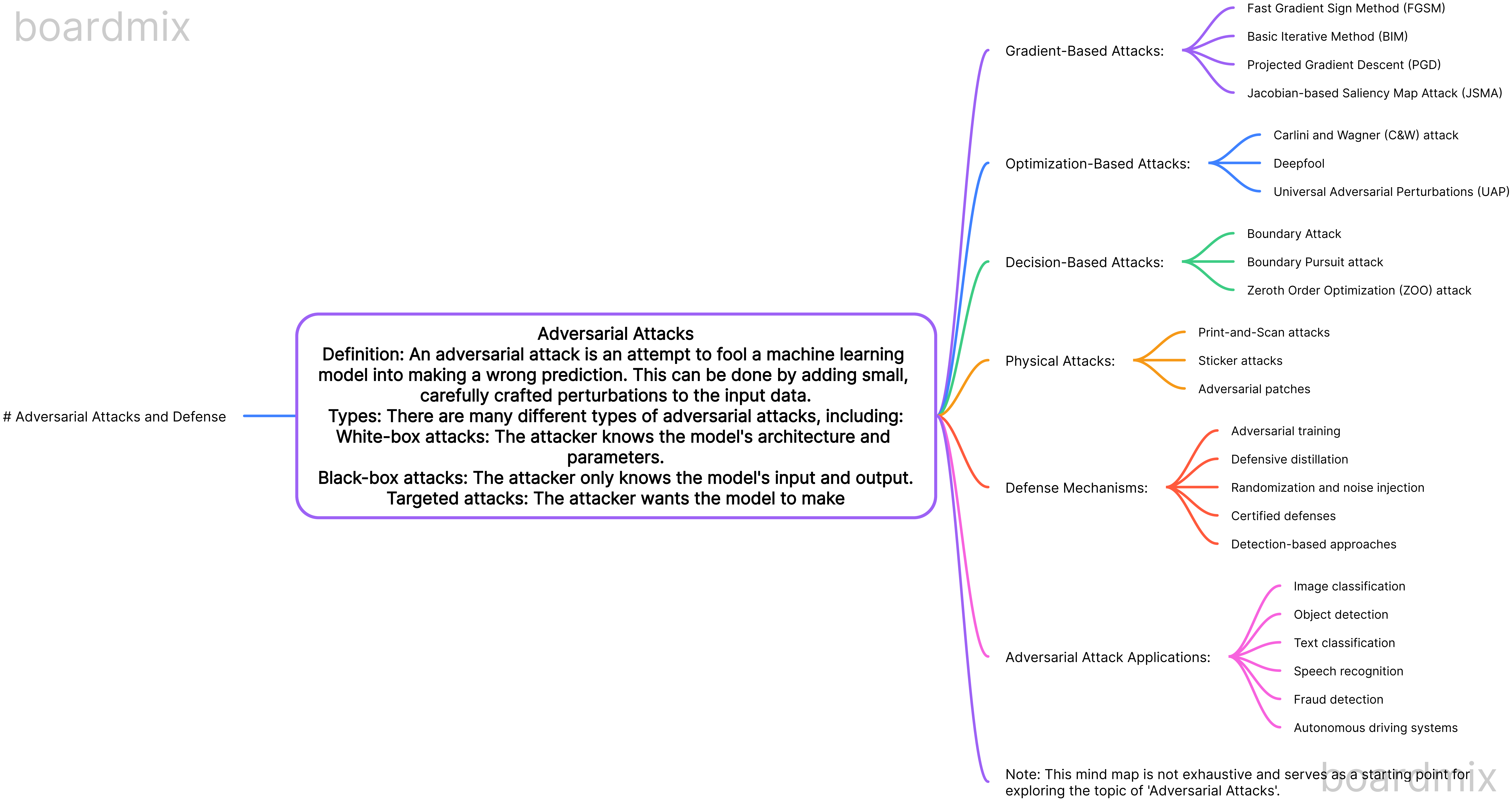

# Adversarial Attacks and Defense

Adversarial Attacks
Definition: An adversarial attack is an attempt to fool a machine learning model into making a wrong prediction. This can be done by adding small, carefully crafted perturbations to the input data.
Types: There are many different types of adversarial attacks, including:
White-box attacks: The attacker knows the model's architecture and parameters.
Black-box attacks: The attacker only knows the model's input and output.
Targeted attacks: The attacker wants the model to make

- Gradient-Based Attacks:
  - Fast Gradient Sign Method (FGSM)
  - Basic Iterative Method (BIM)
  - Projected Gradient Descent (PGD)
  - Jacobian-based Saliency Map Attack (JSMA)
- Optimization-Based Attacks:
  - Carlini and Wagner (C&W) attack
  - Deepfool
  - Universal Adversarial Perturbations (UAP)
- Decision-Based Attacks:
  - Boundary Attack
  - Boundary Pursuit attack
  - Zeroth Order Optimization (ZOO) attack
- Physical Attacks:
  - Print-and-Scan attacks
  - Sticker attacks
  - Adversarial patches
- Defense Mechanisms:
  - Adversarial training
  - Defensive distillation
  - Randomization and noise injection
  - Certified defenses
  - Detection-based approaches
- Adversarial Attack Applications:
  - Image classification
  - Object detection
  - Text classification
  - Speech recognition
  - Fraud detection
  - Autonomous driving systems
- Note: This mind map is not exhaustive and serves as a starting point for exploring the topic of 'Adversarial Attacks'.