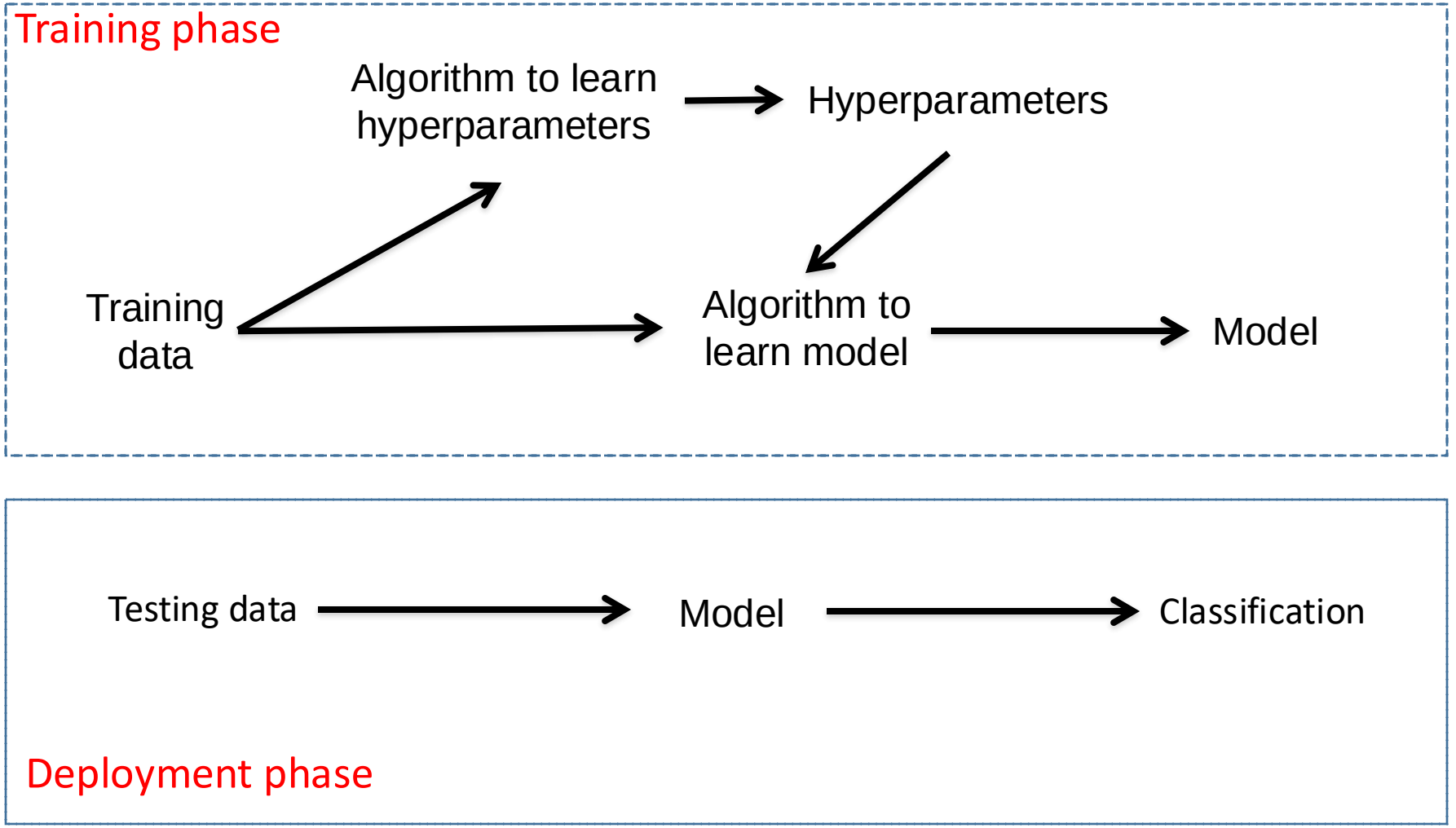
Data Poisoning Attacks to Classifiers

Neil Gong

Machine Learning Pipeline

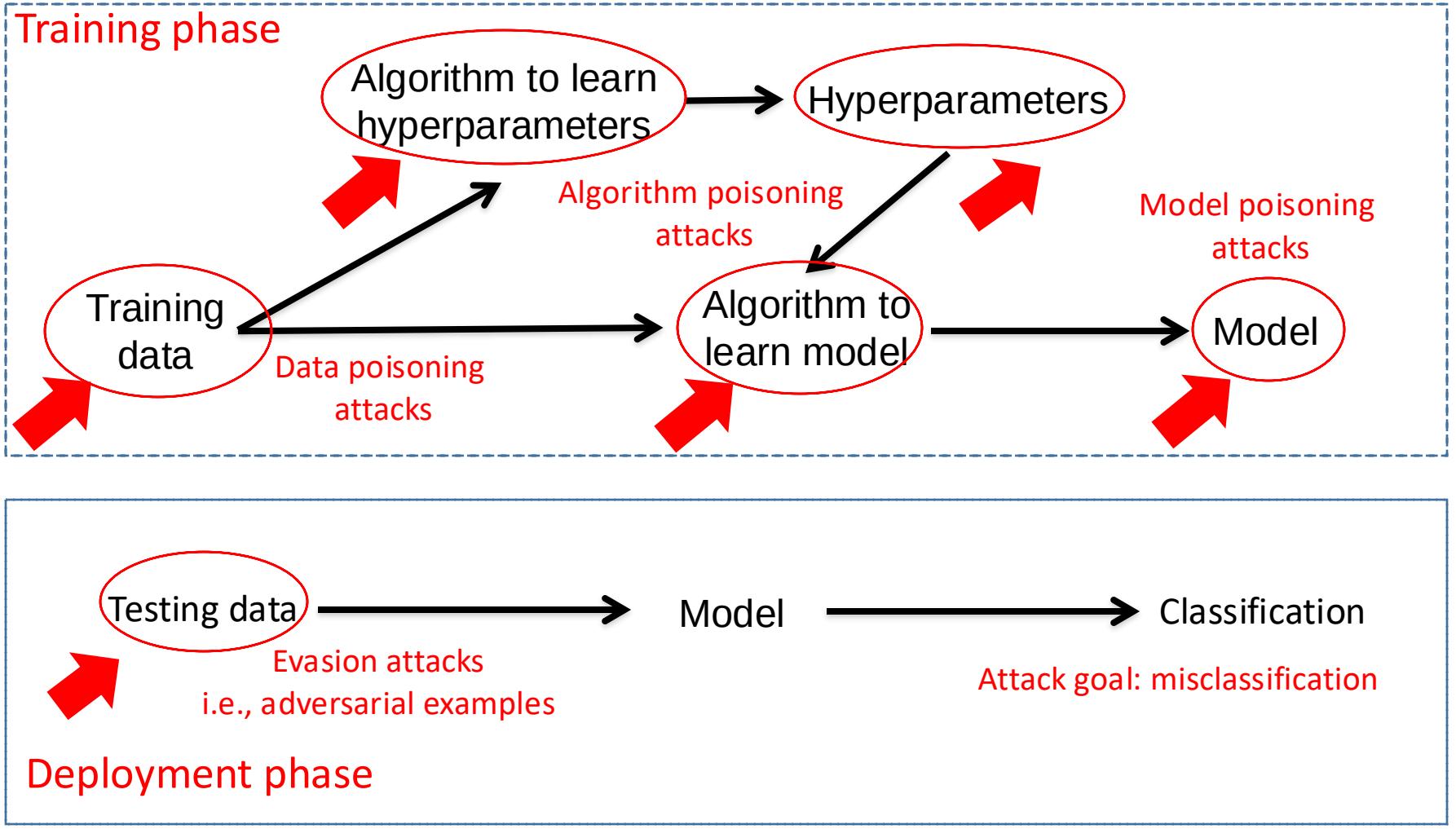


Security of Machine Learning

- Integrity
  - **Training phase**
  - Deployment phase

- Confidentiality
  - Training/testing data
  - Model parameters
  - Hyperparameters
  - Algorithms

Integrity of Machine Learning



Threat Model

- Attacker’s goal
  - Untargeted: large testing error rate
  - Targeted: target label for target inputs
- Attacker’s background knowledge
  - Training data
  - Algorithm
  - Neural network architecture
- Attacker’s capability
  - Add training data
  - Delete training data
  - Modify training data

How to Perform Data Poisoning Attacks

- Untargeted attacks
  - Label flipping
  - Feature perturbation
  - Bi-level optimization problem

- Targeted attacks
  - Clean-label attacks in the feature space
  - Bi-level optimization problem

Untargeted Data Poisoning Attacks