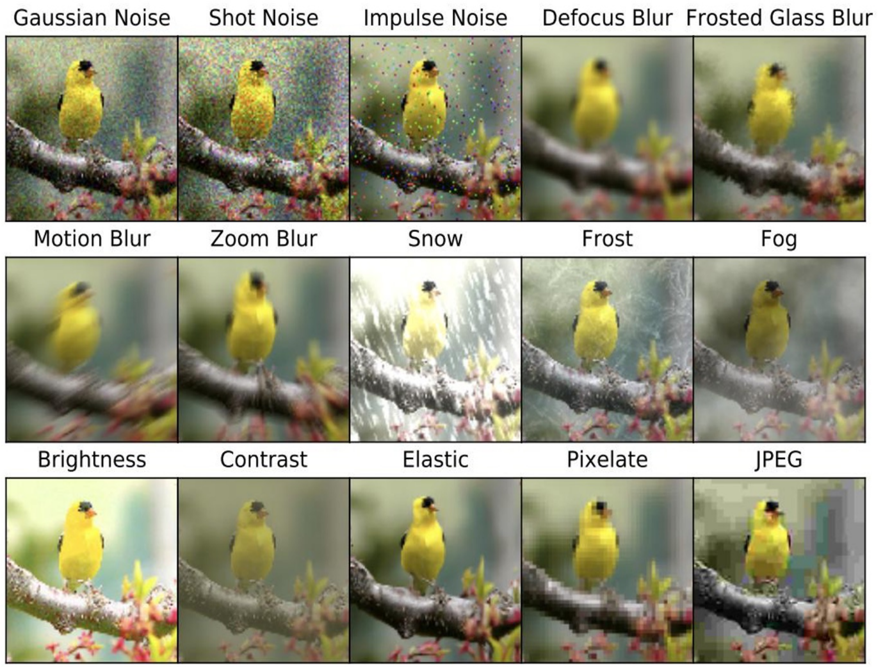
Uncovering Adversarial Risks of Test-time Adaptation

Tong Wu, Feiran Jia, Xiangyu Qi, Jiachen T. Wang,
Vikash Sehwag, Saeed Mahloujifar, Prateek Mittal

PRINCETON UNIVERSITY

PennState

Distribution/Domain Shifts

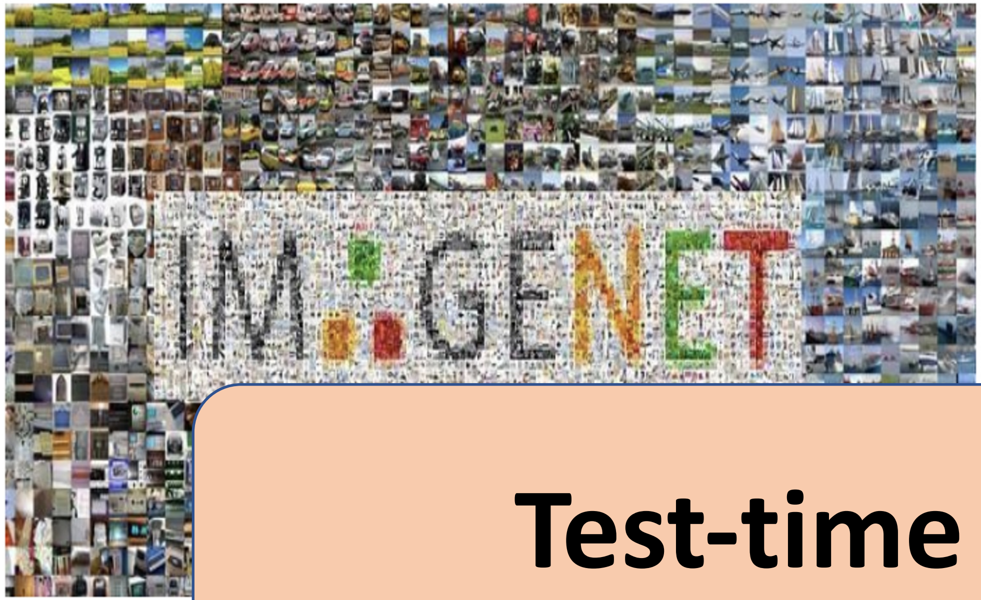


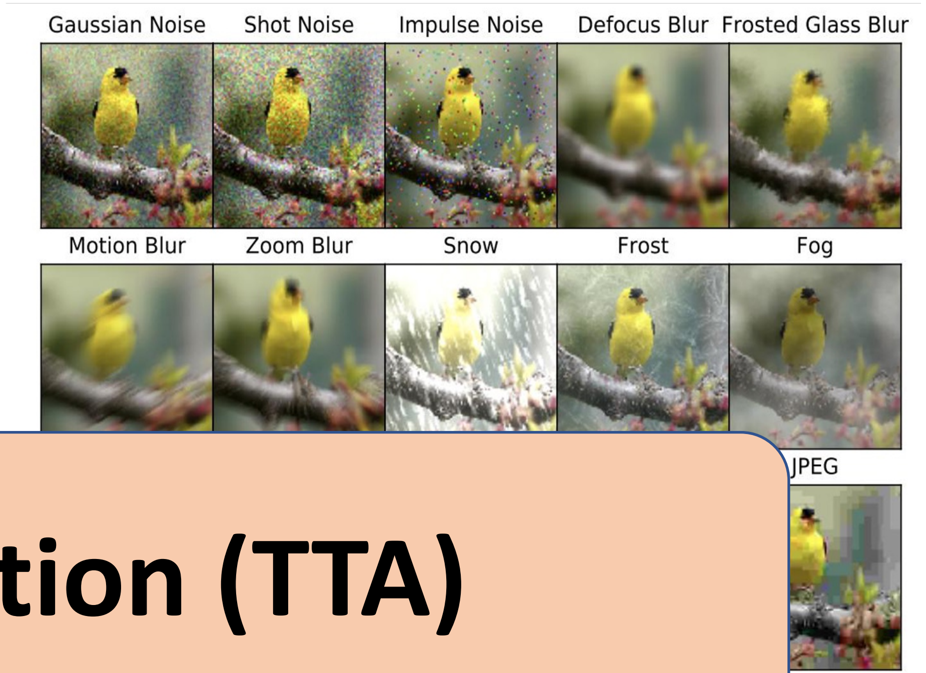
Training Stage:
ImageNet

Deployment Stage:
ImageNet-Corruption

Hendrycks et al, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ICLR 2019

Distribution/Domain Shifts





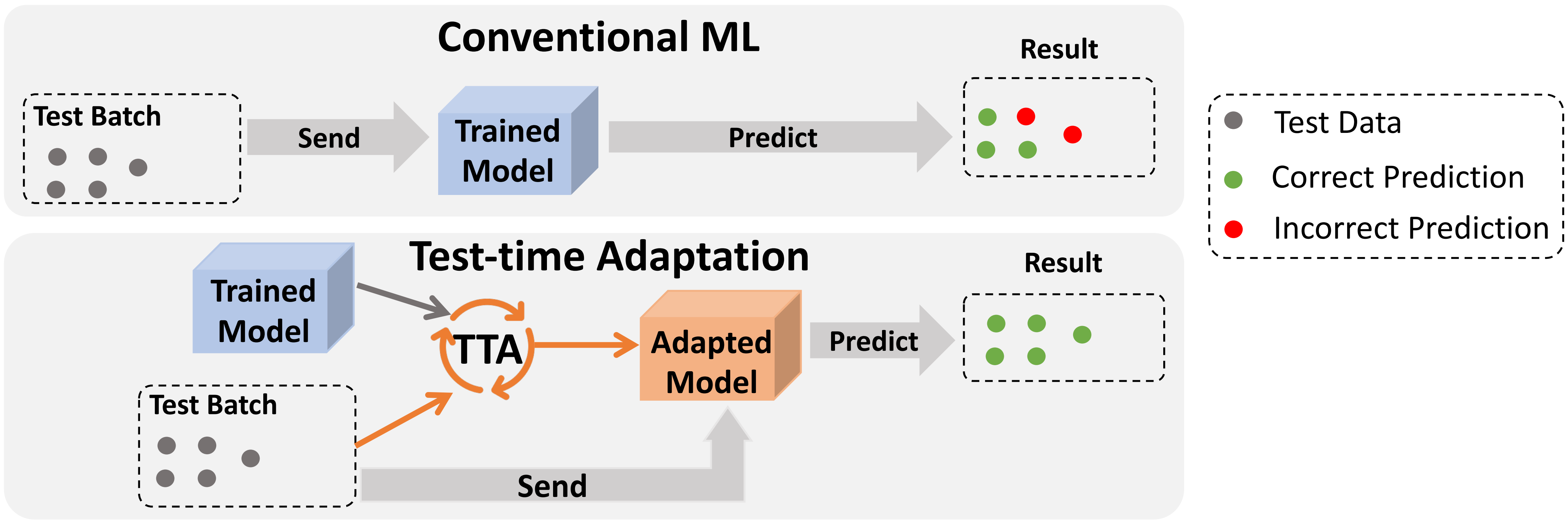
Test-time Adaptation (TTA)

Training Stage:
ImageNet

Deployment Stage:
ImageNet-Corruption

Hendrycks et al, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ICLR 2019

Learning Paradigm Shifts: Test-time Adaptation (TTA)



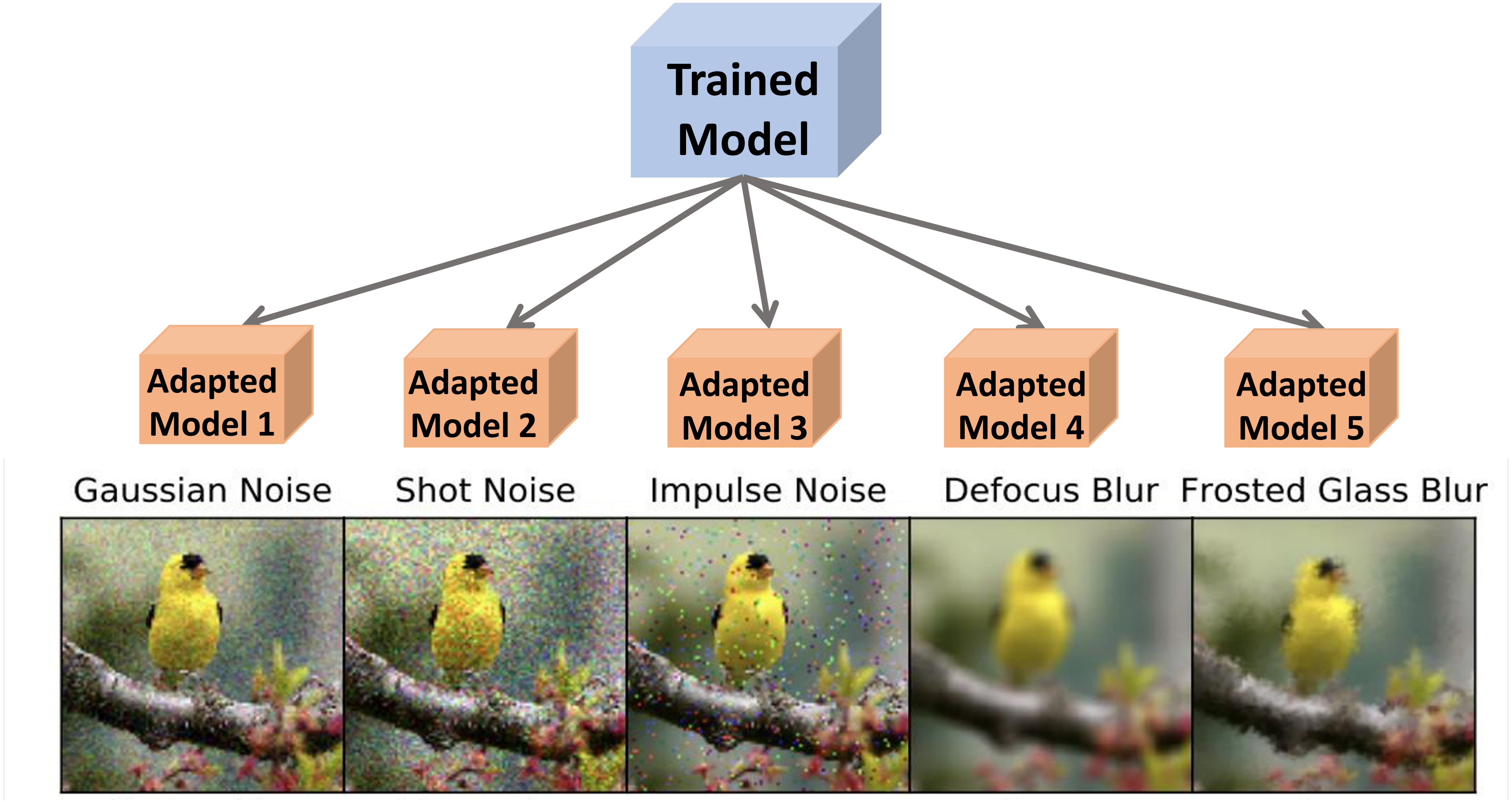
Test-time Adaptation learns the distribution knowledge from test batch

Out of Distribution Test Data

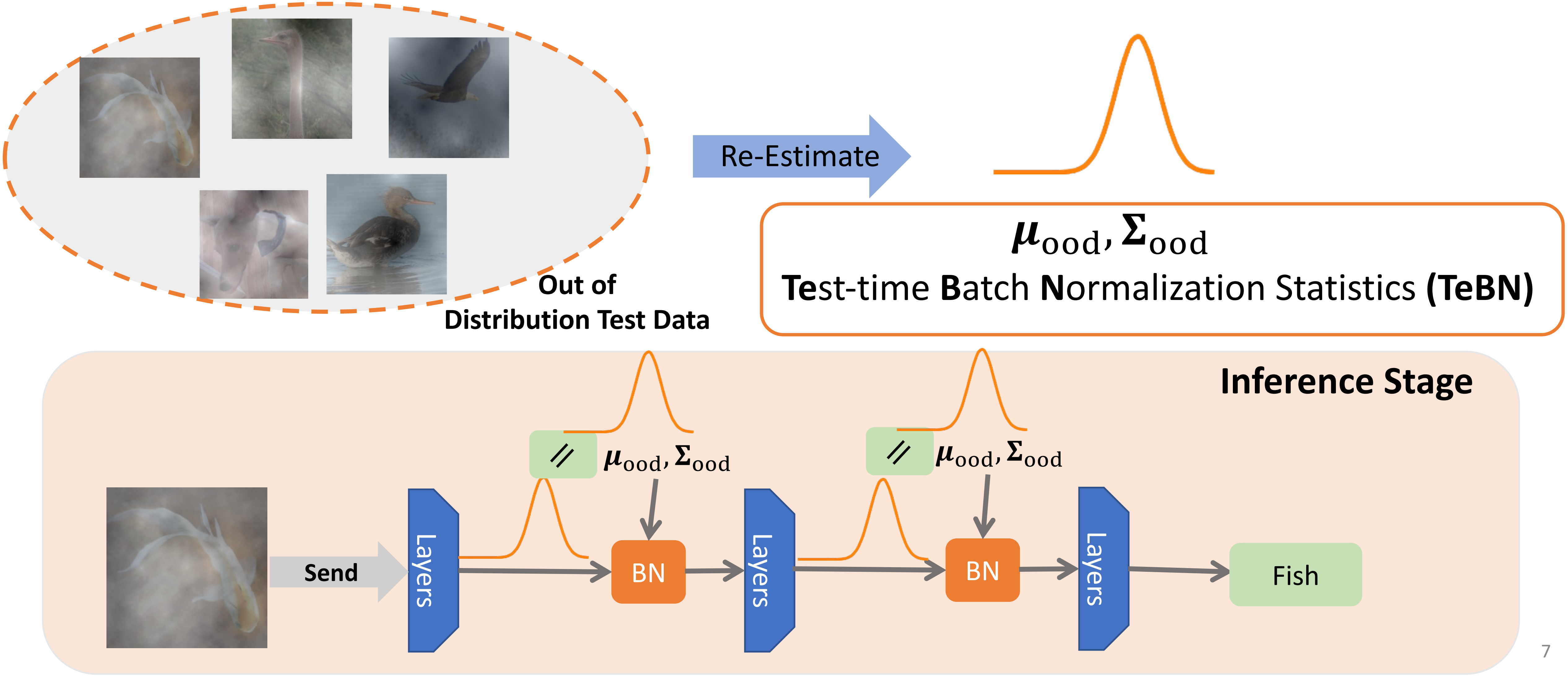
Distribution knowledge:

Fog Experiment

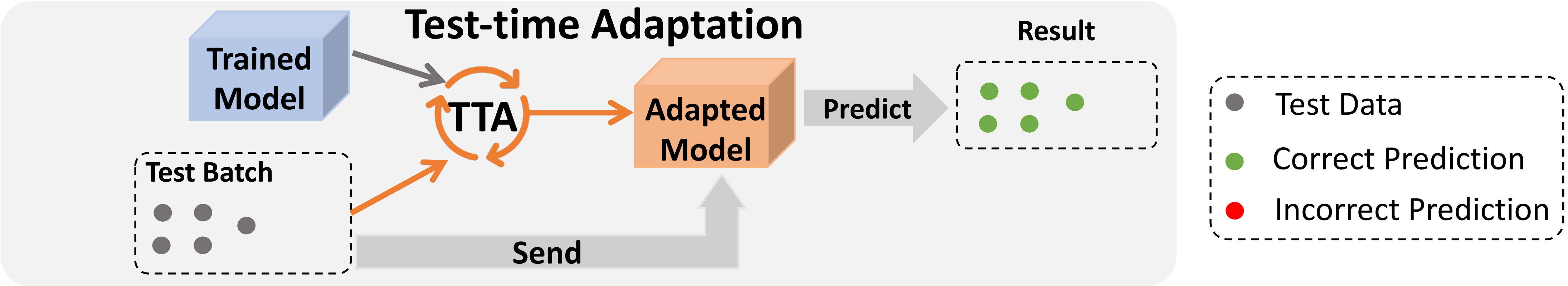
TTA can adjust the model adaptively



Test-time Adaptation: Test Batch Normalization



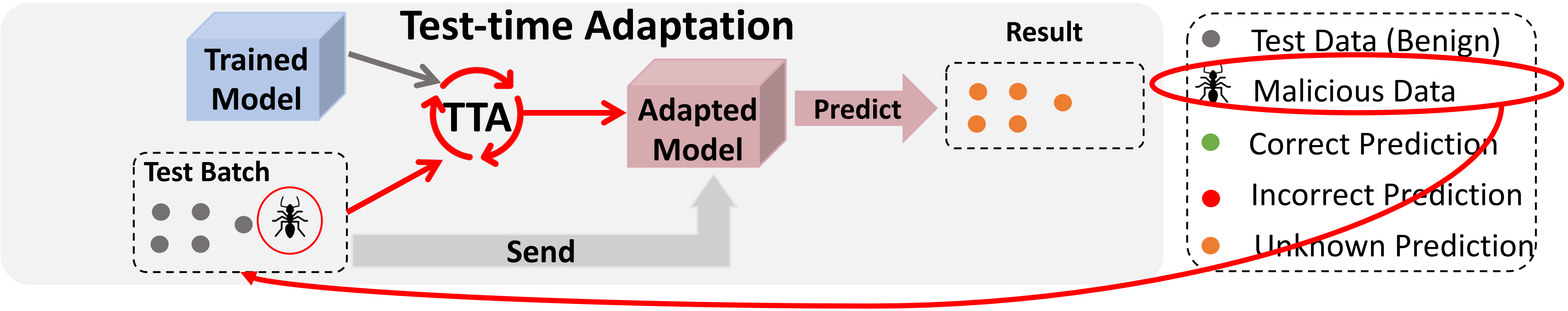
Observations on Test-time Adaptation (TTA)



Adapted model is generated based on the **entire test batch**

The prediction for **one entry** in a batch can be influenced by **other entries**

Test-time Adaptation (TTA) from Adversarial Lens

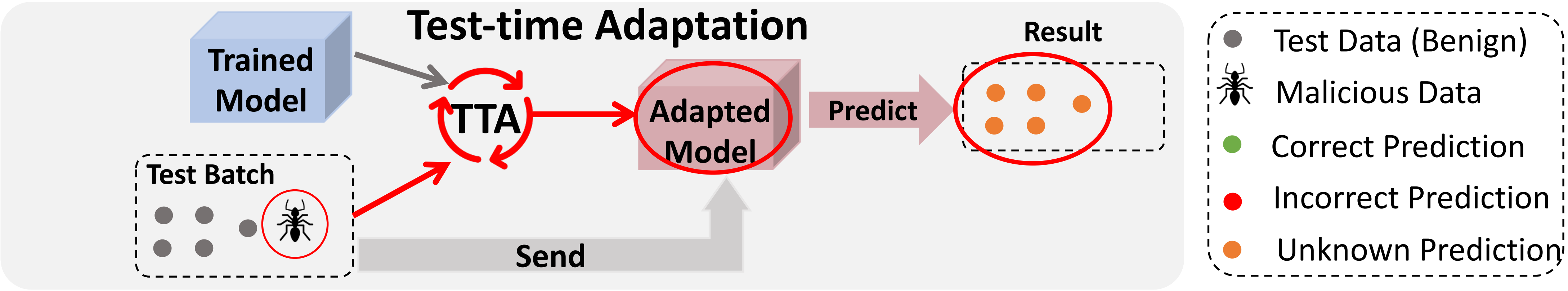


Malicious data at test time can interfere with the generation of adapted model, consequently disrupting predictions on other **unperturbed** data

Introducing Distribution Invading Attacks (DIA)

General Attack Framework: **Distribution Invading Attacks**

An adversary can introduce malicious behaviors into the adapted model by crafting samples in the test batch



Introducing Distribution Invading Attacks (DIA)

General Attack Framework: **Distribution Invading Attacks**

Adversary's Objective

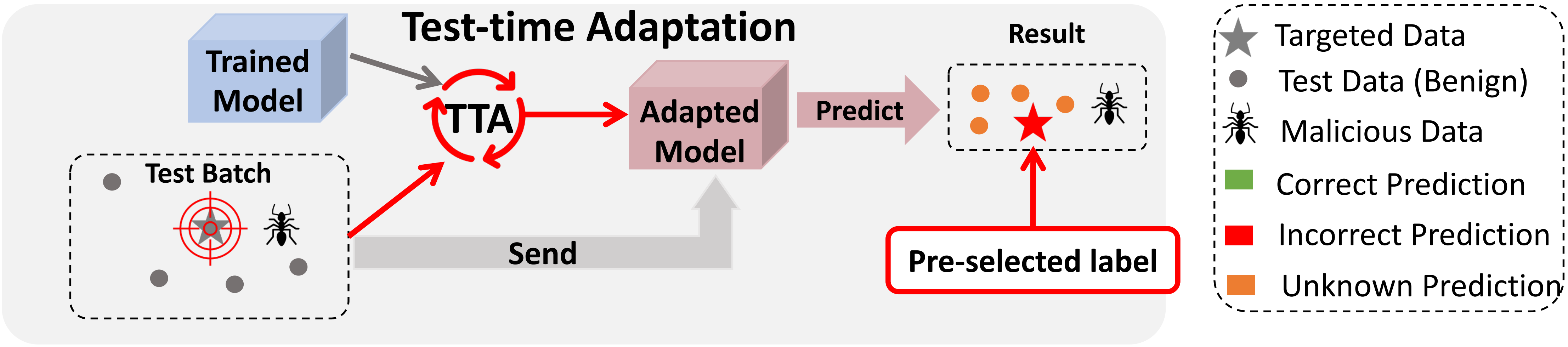
- **Targeted Attack**
- Indiscriminate Attack
- Stealthy Targeted Attack

Input Constraints

- **Unconstrained**
- ℓ_∞ Constraints
- Simulated Corruptions

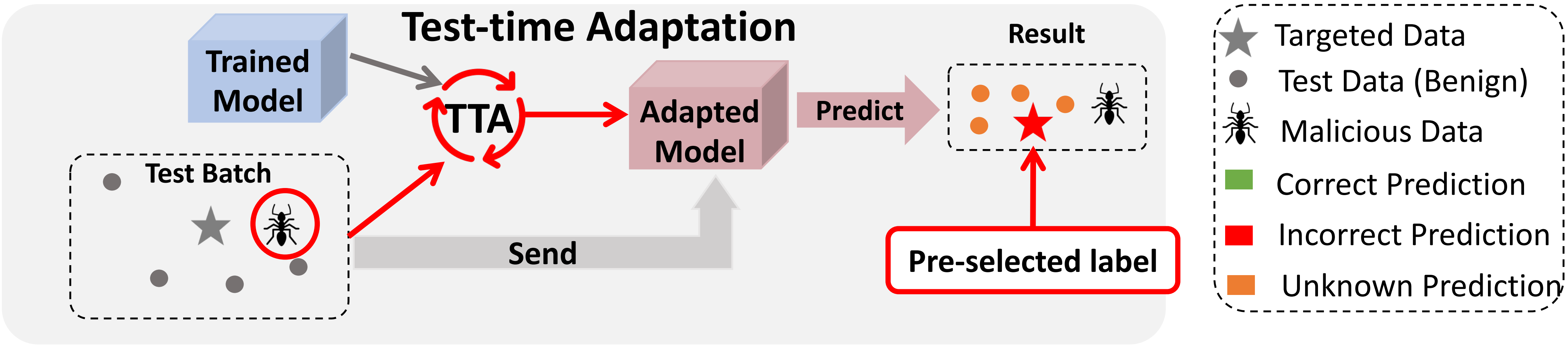
Introducing Targeted Distribution Invading Attacks

- **Adversary's Objective:**
  - Misclassifying a crucial **targeted** sample as a **pre-selected** label



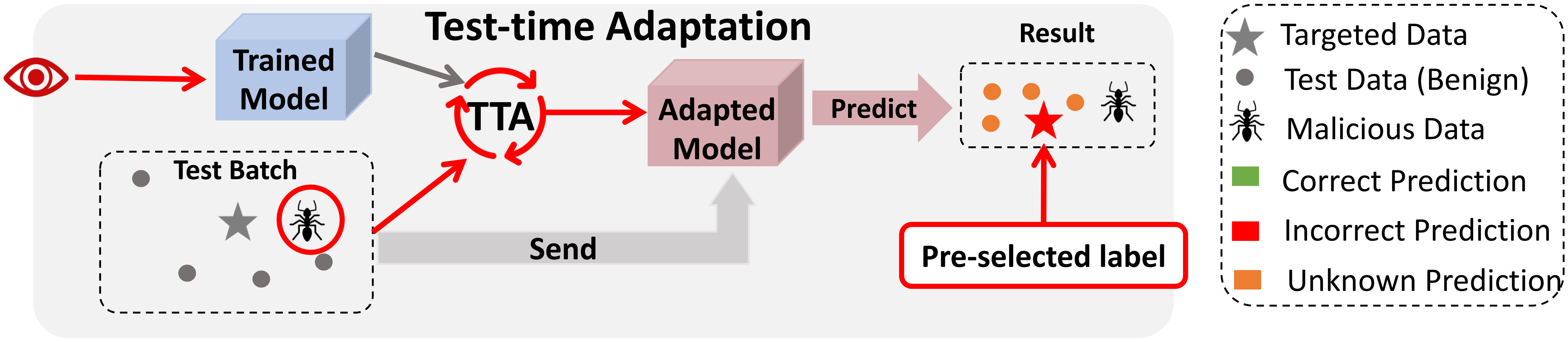
Introducing Targeted Distribution Invading Attacks

- **Adversary's Objective:**
  - Misclassifying a crucial **targeted** sample as a **pre-selected** label
- **Adversary's Capability:**
  - Inject/craft a small portion of **unconstrained** samples to the test batch

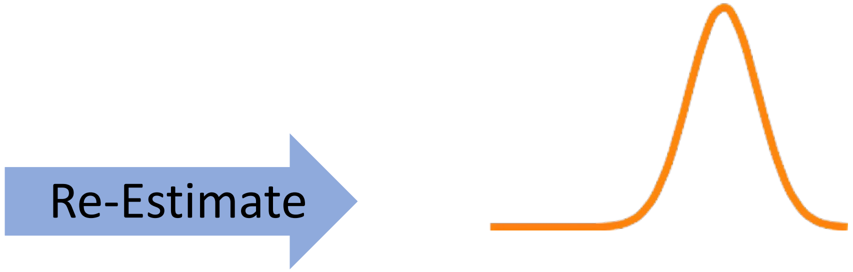


Introducing Targeted Distribution Invading Attacks

- **Adversary's Objective:**
  - Misclassifying a crucial **targeted** sample as a **pre-selected** label
- **Adversary's Capability:**
  - Inject/craft a small portion of **unconstrained** samples to the test batch
- **Attacker's Knowledge:**
  - Model Architecture and parameters



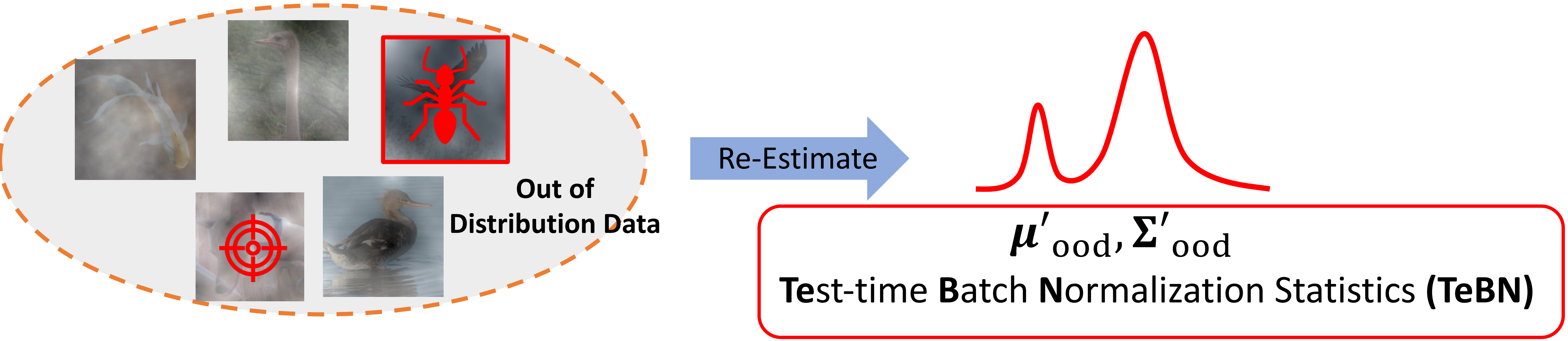
Case Study: Targeted DIA on Test Batch Norm



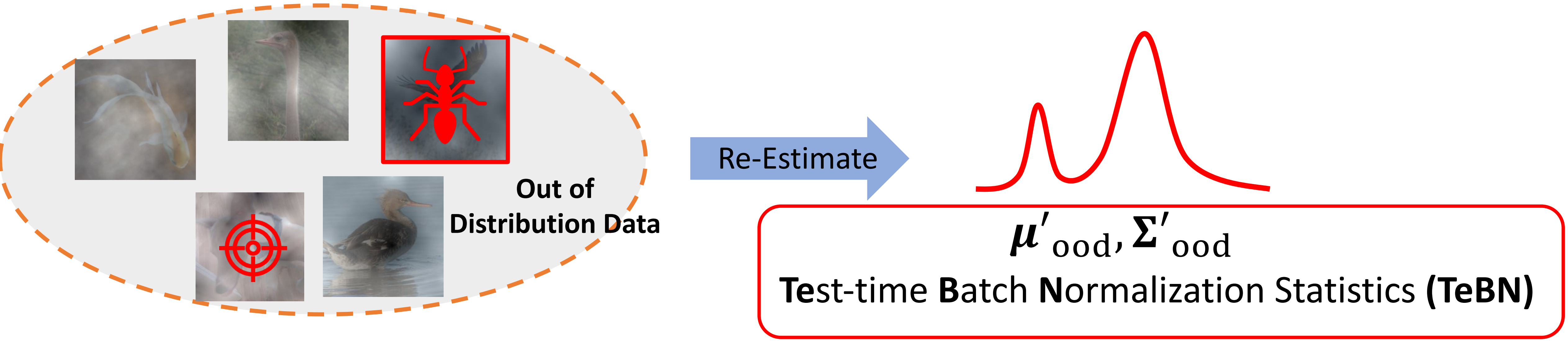
$$\boldsymbol{\mu}_{\text{ood}}, \boldsymbol{\Sigma}_{\text{ood}}$$

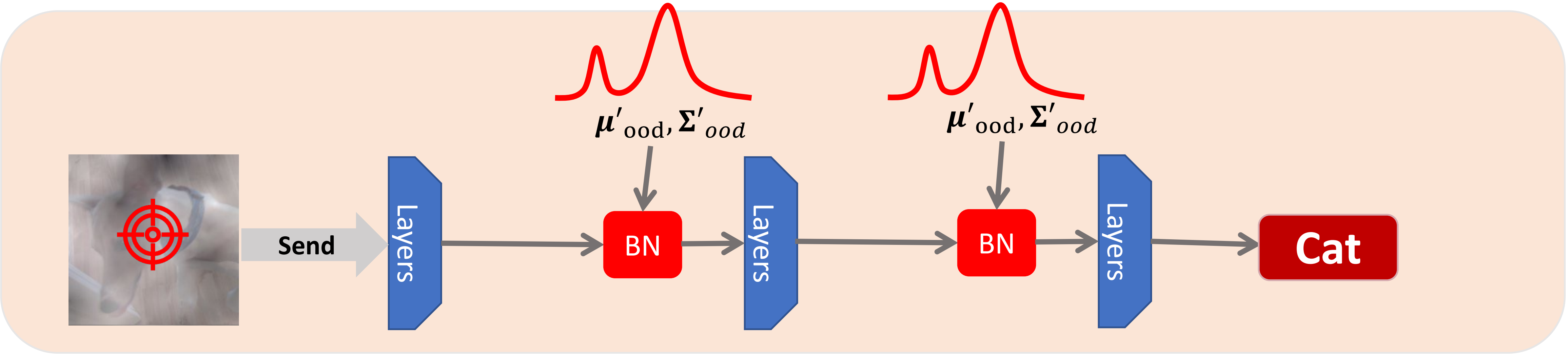
Test-time **B**atch **N**ormalization Statistics **(TeBN)**

Case Study: Targeted DIA on Test Batch Norm



Case Study: Targeted DIA on Test Batch Norm





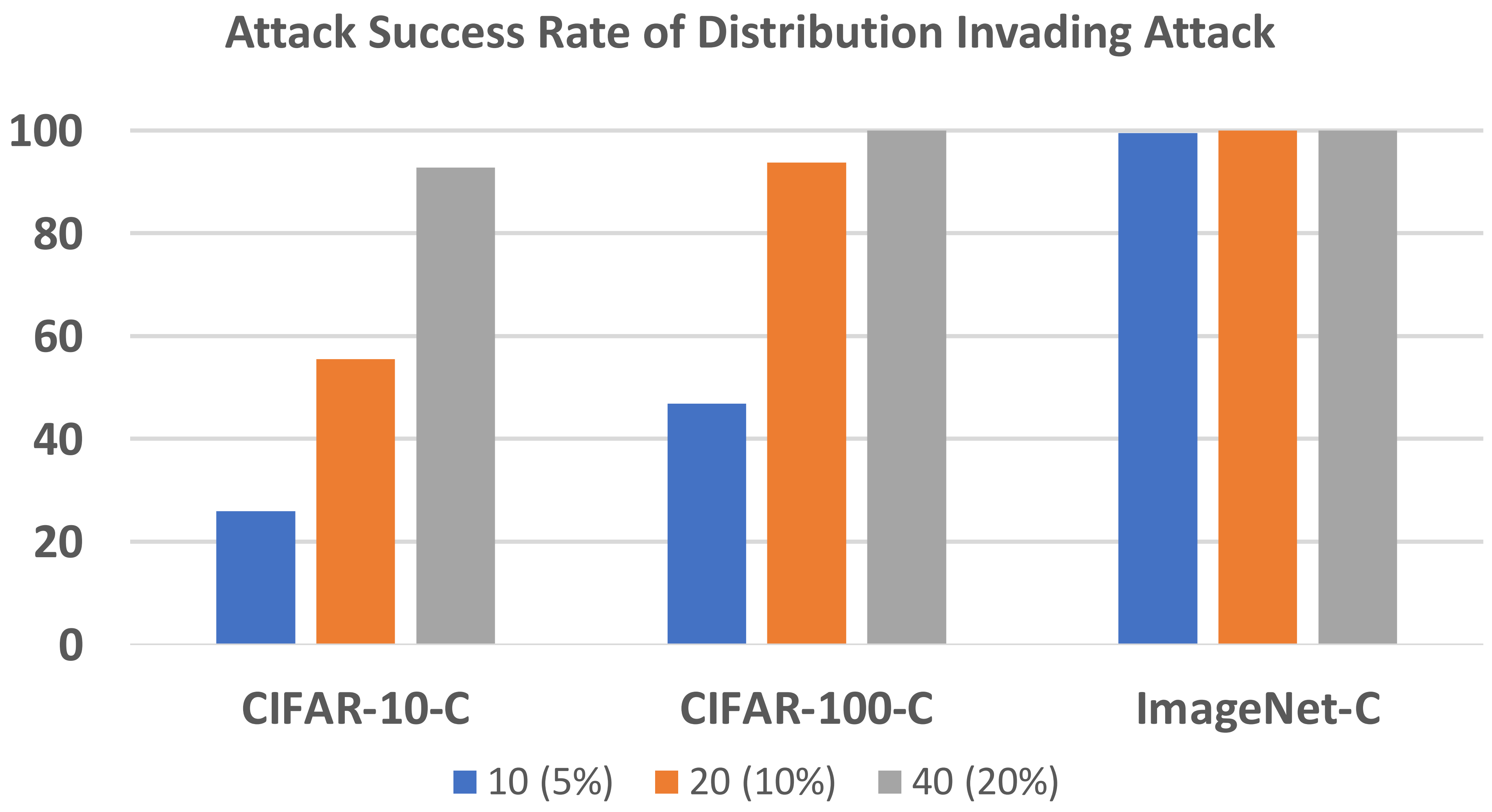
Case Study: Targeted DIA on Test Batch Norm

$$\min_{\boxed{\mathbf{x}_m}} \mathcal{L}\left(\boldsymbol{f}\left(\mathbf{x}_{tgt}; \boxed{\boldsymbol{\theta}_{BN}}\right), \boldsymbol{y}_{tgt}\right)$$

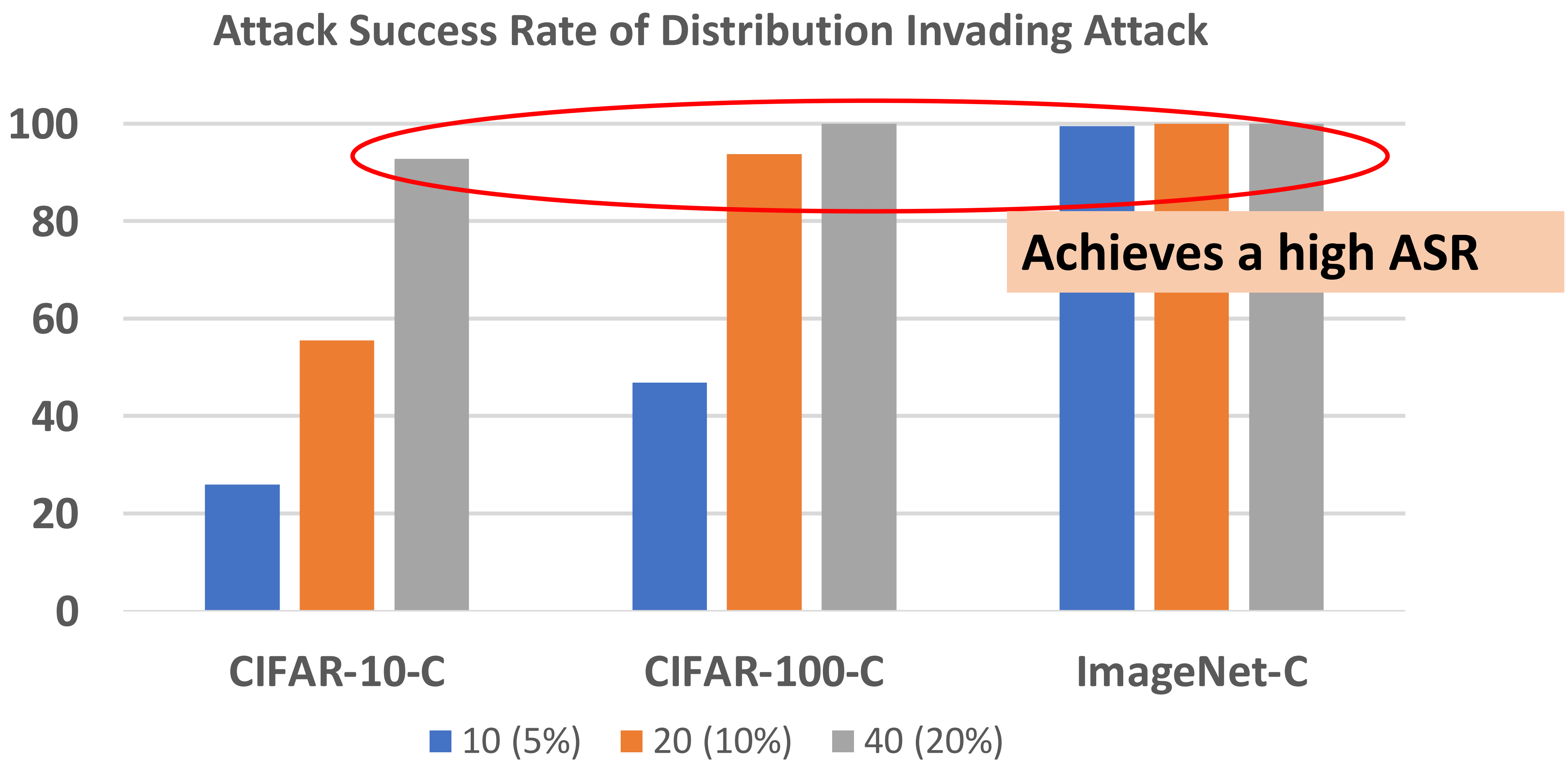
Malicious data → Batch Norm

- Using **Gradient Descent** to minimize the loss of targeted sample.

Experiment Results: (DIA) on Test Batch Norm



Experiment Results: (DIA) on Test Batch Norm



Conclusion

- While TTA achieves better performance on OOD data, it has a novel security risk
- **Distribution Invading Attacks** exploit the risks of TTA.
  - Adversary's objectives
  - Input constraints
  - Eight other TTA methods (check our paper)
- We investigate mitigation strategies (check our paper)
  - adversarially trained model
  - robustly estimating BN statistics
- Our findings inspire building robust and effective TTA techniques.