**Overview**

**Trustworthy Machine Learning**
- Imperfect Data

- **TMLR Group**
- Trustworthy Machine Learning
  - Noisy Labels
  - Adversarial Examples
  - Out-of-distribution Data
  - New Directions

**TML with Noisy Labels**

**What are Label Noise?**
- Class-Conditional Noise (CCN)
- Instance-Dependent Noise (IDN)

**What are Out-of-distribution Data?**

**New Directions in TML**

**Trustworthy Foundation Models**
- DeepInception: A jailbreak attack method to reveal the safety risks of foundation models by concealing the attack intention with nested instructions for LLM.

**Trustworthy Federated Learning**
- SFAT: Pursue the adversarial robustness of a server model, while reducing the exacerbation of the data heterogeneity.

**Trustworthy Casual Learning**
- COAT (Causal-Observational Approach to Training) using LLMs to generate useful high-level factors and crafting their measurements. COAT also adopts causal discovery methods (CDs) to find causal relations among the identified variables and provide feedback for LLMs to iteratively refine the proposed factors.

**TML under Out-of-distribution Data**

**What are Out-of-distribution Data?**
- Deep Noise: Do Not Necessarily Know What They Don’t Know...
- Open-world (O, L, T): Memory-based (M), prior knowledge (K), test data (T)

**Learn a Watermark**
- Original
- Watermarked

**Detect with the NegLabel**
- Score overlapping is shrink!

**New Diagrams**
- Geometric View on Adversarial Data
- Causal View on Adversarial Data
- Illustration of NegMining

**What are Adversarial Examples?**
- Geometric View on Adversarial Data
- Causal View on Adversarial Data

**Trustworthy Machine Learning under Imperfect Data**
- Dr. Bo Han
- Assistant Professor @ HKBU TMLR Group
- BAIHO Visiting Scientist @ RIKEN AIP Team
- bhanml@comp.hkbu.edu.hk

**What are Adversarial Examples?**
- Geometric View on Adversarial Data
- Causal View on Adversarial Data

More attackable/guarded data are closer to/further away from the decision boundary.

**What are Label Noise?**
- Class-Conditional Noise (CCN)
- Instance-Dependent Noise (IDN)

**What are Out-of-distribution Data?**

**New Directions in TML**

**Trustworthy Foundation Models**
- DeepInception: A jailbreak attack method to reveal the safety risks of foundation models by concealing the attack intention with nested instructions for LLM.

**Trustworthy Federated Learning**
- SFAT: Pursue the adversarial robustness of a server model, while reducing the exacerbation of the data heterogeneity.

**Trustworthy Casual Learning**
- COAT (Causal-Observational Approach to Training) using LLMs to generate useful high-level factors and crafting their measurements. COAT also adopts causal discovery methods (CDs) to find causal relations among the identified variables and provide feedback for LLMs to iteratively refine the proposed factors.