Automated and Scalable Algorithms and Systems for Unsupervised ML

Yue Zhao Carnegie Mellon University

1. Vision: Enabling High-stakes Applications by Unsupervised Machine Learning

Every day, we generate more than 2.5 quintillion bytes (i.e., 2.5×10^{18}) of data—most of which lacks ground truth labels. For example, we don't know if a sensor on an autonomous vehicle has captured a real hazard until it is verified. However, it is costly and impractical to manually review and label all the data for supervised machine learning (ML). Thus, designing **automated** and **scalable unsupervised algorithms and systems** is the key to the future ML.

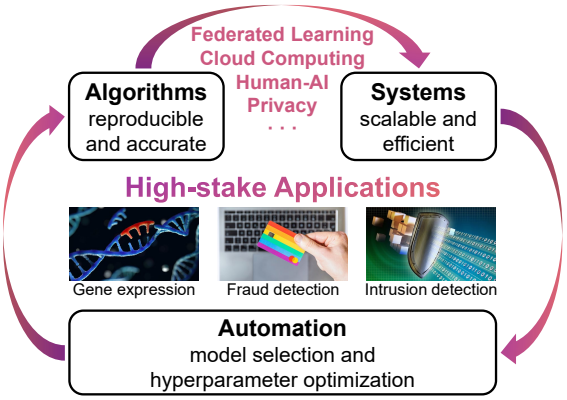

Figure 1: **A Blueprint of My Research**: AD requires the integration of algorithms, systems, and automation. And it interacts with the problems of privacy, human-AI collaboration, federated learning, etc.

My research focuses on using algorithmic and system methods to automate and accelerate unsupervised ML, with notable achievements in *anomaly detection (AD)* and *out-of-distribution (OOD) detection*–the most important *unsupervised* ML task(s). AD and OOD identify deviant and different samples from the general data distribution [19] in a data-driven manner, with many key applications where the automatic discovery of rare and deviant items is valuable. Their applications include robot manipulation detection [13], malware detection [1], and fraud detection [8]. More recently, AD and OOD are used to identify chronic brain infarcts on MRI [30] and real-time risk modeling for autonomous driving [2].

Thus, my long-term goal is to develop *reproducible*, *automated*, and *scalable* detection algorithms and systems to support high-stakes research and applications. Fig. 1 shows the relationship among detection algorithms, systems, and automation—my works cover all of them. In addition to **my interdisciplinary AI studies** in HR management [17], security [16], drug discovery [4, 5], healthcare [22], gene expression [31], and computer vision [12, 33], my research has produced successful results in **unsupervised ML**:

- ***Reproducible* detection benchmarks** [3, 7, 25] to unlock insights and identify new opportunities
- ***Automated* systems** for detection model selection and hyperparameter optimization [11, 14, 23, 27]
- ***Scalable* learning systems** that support detection tasks with big data [6, 19, 21, 26, 28]
- ***Accurate* detection methods** using ensemble learning [9, 18, 20, 24], graph learning [8, 32], and weakly-supervised learning [15, 29]

Summary of outcomes. In addition to **more than 30 papers in leading venues** including *JMLR*, *NeurIPS*, *VLDB*, and *MLSys*, my work has been widely used in academia (**1300+ citations**) and industry (**10M downloads**, **15,000 GitHub Stars**, and **popular ML projects** (PyOD, PyGOD, and TDC)). My machine learning systems works and proposal have won **Norton Fellowship** and **Meta AI4AI Award**.

My future research will make AD and other unsupervised ML algorithms more flexible and accessible, to handle more complex data modalities and stricter privacy requirements. **First**, I plan to design new automated AD algorithms with more flexible objectives that optimize both accuracy and efficiency, which can be applied to time-critical applications. **Additionally**, I will extend (scalable) detection systems to support more complex modalities like streaming data and dynamic graphs. **Finally**, I will work to bridge the gap between AD and OOD detection with other emerging fields, such as robotics and bioinformatics, by proposing domain-specific methods and systems.

2. Past Research: Reproducible, Automated, and Scalable Detection

Reproducible and large-scale analysis for anomaly and OOD detection. Although a long list of detectors has been proposed in the last few decades, it is unclear about their characteristics. A more serious question is that they are often evaluated on an ad-hoc subset of public datasets with simple number comparisons, which makes the advancement and conclusions questionable. In three of my recent *NeurIPS* papers, I designed and ran large-scale AD analysis for tabular data (i.e., ADBench [3]), time-series data (i.e., TSOD [7]), and graph data (i.e., BOND [25]). Taking ADBench as an example, we built the most comprehensive tabular AD benchmark on 57 datasets with 30 detection algorithms, covering different settings including supervision levels, anomaly types, and data corruptions, with 98,436 experiments in total. Under rigorous statistical analysis (for the first time in such benchmarks), we point out that *none of the evaluated unsupervised AD methods are statistically different from the rest*. This surprising finding raises attention to more fair and rigorous evaluation processes. Through these large-scale benchmarks, we unlock many insights into AD, including the role of supervision, the best practice of handling corrupted data, the similarity and difference among data modalities, etc. These valuable observations and insights guide us to design better AD algorithms and tailor them for real-world applications with caution.

Automated detection pipelines. Given none of the AD/OOD methods can consistently outperform as shown in ADBench, *how can we select the "best" one for different applications automatically*? This is extremely critical: the lack of ground truth in unsupervised AD makes model evaluation and validation infeasible, but using a random AD model blindly can cause unsatisfactory accuracy in fraud detection, network intrusion, and rare diseases. In my *NeurIPS' 21* paper [23], for the first time we formalize this challenging unsupervised outlier model selection problem and propose an effective method called MetaOD. It leverages meta-learning (prior knowledge helps future decisions) to intelligently pick the best detection model for a new task. Extensive results show it can consistently pick the top 1%-20% model from hundreds of candidates. Under the umbrella of automated detection, we have designed more novel methods as well as defined more pressing problems, including ELECT [27] for better task-driven model selection, IPMs [11] for unsupervised AD model evaluation, and HPOD [14] for unsupervised AD hyperparameter optimization. Due to their effectiveness, researchers have applied our methods to applications in tunnel engineering [10] and KPI monitoring [34]—huge social-economical gain has been realized.

Scalable detection systems. Applying AD algorithms to applications with high-dimensional, large data is hard. Taking MasterCard (i.e., the global payment system) as an example, they need to do real-time fraud detection for millions of transactions, where the results must be returned within 5–7 seconds. System support has been long ignored in AD and/or other ML algorithms, making it less useful in real-life scenarios. To bridge the gap, I build a series of scalable learning systems for tabular AD (i.e., PyOD [19], SUOD [21], and TOD [28]), time series AD (i.e., TODS [6]), and graph AD (i.e., PyGOD [26]), by leveraging diverse techniques in machine learning system (MLSys). Specifically, I developed Python Outlier Detection (PyOD) [19], which has become the most famous AD platform and gotten published in *JMLR* due to its comprehensiveness (40+ methods) and efficiency (JIT optimization, etc.). Moreover, it enables large-scale comparison of anomaly detection, which has served as the primary baseline/framework in more than 400 academic papers since 2019 (by Google Scholar) in different fields, and in more than 10 data science books (by Google Books). In our recent *VLDB' 23* paper[28], "TOD: GPU-accelerated Outlier Detection via Tensor Operations", we build the first multiple-GPU system to accelerate large-scale AD applications with an on avg. 11× speedup to CPU-based PyOD. Within TOD, we introduce a set of novel techniques to abstract complex AD/OOD algorithms into small tensor operators for GPU acceleration, quantize computations with an accuracy guarantee, and more. I view MLSys as a key venue to bridge AD and system research, and will keep pushing its usage in AD/OOD and other learning applications.

Open-source contributions and industrial impact. I am a firm believer that ML research should be open and accessible. Being widely used by academia and industry, my open-source systems have changed

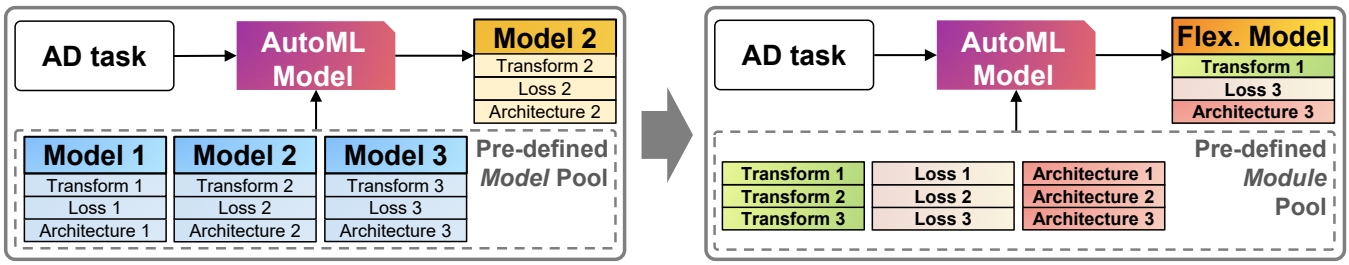

Figure 2: Future research (right) can make current AD model selection (left) more flexible

AD research and related industries. For instance, `PyOD` has got 10,000,000 downloads, 6,000 GitHub stars, and enabled numerous applications — Walmart Labs using it for dynamic pricing, IBM Watson for IoT, Morgan Stanley for risk modeling, etc. An estimate of $35,000,000 savings has been achieved by using PyOD and SUOD at a major payment platform. Due to my ML system works, I am awarded the 2022 Norton Fellowship and Meta AI4AI Award. It is an honor to give invited talks for my research and open-source works at leading institutes including Morgan Stanley, Wells Fargo, E&Y, etc.

3. Future Plans for Flexible and Customized Detection Algorithms and Systems

My past research has laid out the foundations for reproducible, automated, and scalable AD, while I detail three important future steps to further empower AD/OOD with more flexibility and applicability.

Flexible objectives and design choices in automated anomaly detection. **First**, existing research only focuses on maximizing the detection rate, while failing to consider the runtime and memory consumption of the selected AD model. However, these efficiency metrics are as important as accuracy in time-critical and low-resource applications. In future research, I plan to include more practical metrics in the objective function of automated AD model selection and hyperparameter optimization, so that the selection can smartly penalize accurate but costly models if specified. **Second**, current automated AD merely chooses a model from a pre-defined model pool, other than finding an AD model that actually tailors for the underlying task. If we consider all design choices/modules of an AD model, including data transformation, loss function, network architectures, etc., as tunable modules, then we gain the flexibility of intelligently selecting each module, resulting in more customized solutions with better performance. Fig. 2 compares the difference between existing approaches (on the left) to the proposed future plan (on the right). Note the proposed approach is more flexible to dynamically build a new AD model by selecting underlying modules, while the existing methods can only select from the static pool of models.

Broader system support for emerging data modalities. **First**, existing detection systems are primarily for static data settings where anomaly behaviors are assumed not to evolve, while the importance of streaming and dynamic data should not be overlooked. For instance, network intrusion behaviors are not static—intruders dynamically adapt their strategies based on the failure rate. With new AD systems to handle such dynamic settings, more interesting scenarios can be enabled, and better efficacy can be achieved. **Second**, data privacy and security have gained more attention nowadays, which stresses additional challenges for AD systems. For instance, AD applications often require *all* (client/patient/transaction) samples to measure their outlyingness, which may cause challenges in aggregating sensitive information. I propose to design federated-based AD systems, e.g., detecting rare diseases collaboratively by using patients' diagnosis results independently without aggregation—such a system is absent and carries significant value.

Domain-specific anomaly detection algorithms. Although AD/OOD methods have been used in diverse applications, the general practice is still "one ring to rule them all". However, using generic AD/OOD methods is often insufficient as shown in ADBench [3]. In future research, I will work with domain experts to design specialized detection algorithms and systems—I look forward to working with economists for complex social behavioral signals, healthcare researchers for high-dimensional biological data, as well as AI-robotics teams for complex physical world environments. Deep fusion with other domains can bring anomaly detection and general ML to the next level, which requires interdisciplinary collaboration.

References

[1] Ni An, Alexander Duff, Mahshid Noorani, Steven Weber, and Spiros Mancoridis. Malware anomaly detection on virtual assistants. In *2018 13th International Conference on Malicious and Unwanted Software (MALWARE)*, pages 124–131. IEEE, 2018.

[2] Daniel Bogdoll, Maximilian Nitsche, and J Marius Zöllner. Anomaly detection in autonomous driving: A survey. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4488–4499, 2022.

[3] Songqiao Han*, Xiyang Hu*, Hailiang Huang*, Mingqi Jiang*, and **Yue Zhao***. ADBench: Anomaly detection benchmark. *Neural Information Processing Systems (NeurIPS)*, 2022. (**Equal contribution & the corresponding author**).

[4] Kexin Huang, Tianfan Fu, Wenhao Gao, **Yue Zhao**, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Neural Information Processing Systems (NeurIPS)*, 2021.

[5] Kexin Huang, Tianfan Fu, Wenhao Gao, **Yue Zhao**, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*, 2022. URL `https://doi.org/10.1038/s41589-022-01131-2`.

[6] Kwei-Herng Lai, Daochen Zha, Guanchu Wang, Junjie Xu, **Yue Zhao**, Devesh Kumar, Yile Chen, Purav Zumkhawaka, Minyang Wan, Diego Martinez, et al. TODS: An automated time series outlier detection system. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 16060–16062, 2021.

[7] Kwei-Herng Lai, Daochen Zha, Junjie Xu, **Yue Zhao**, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. *Neural Information Processing Systems (NeurIPS)*, 2021.

[8] Meng-Chieh Lee, **Yue Zhao**, Aluna Wang, Pierre Jinghong Liang, Leman Akoglu, Vincent S Tseng, and Christos Faloutsos. AutoAudit: Mining accounting and time-evolving graphs. In *IEEE International Conference on Big Data (Big Data)*, pages 950–956. IEEE, 2020.

[9] Zheng Li, **Yue Zhao**, N Botta, C Ionescu, and Xiyang Hu. Copula-based outlier detection. In *IEEE International Conference on Data Mining (ICDM)*, pages 17–20, 2020.

[10] Jinquan Liu and Tongtong Zou. Identifying the outlier in tunnel monitoring data: An integration model. *Computer Communications*, 188:145–155, 2022.

[11] Martin Q Ma*, **Yue Zhao***, Xiaorong Zhang, and Leman Akoglu. The need for unsupervised outlier model selection: A review and evaluation of internal evaluation strategies. *ACM SIGKDD Explorations Newsletter*, 25(1), 2023. (**Equal contribution**).

[12] Yiqun Mei, **Yue Zhao**, and Wei Liang. DSR: An accurate single image super resolution approach for various degradations. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[13] Daehyung Park, Zackory Erickson, Tapomayukh Bhattacharjee, and Charles C Kemp. Multimodal execution monitoring for anomaly detection during robot manipulation. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 407–414. IEEE, 2016.

[14] **Yue Zhao** and Leman Akoglu. Towards unsupervised HPO for outlier detection. *arXiv preprint arXiv:2208.11727*, 2022. **Under review**.

[15] **Yue Zhao** and Maciej K Hryniewicki. XGBOD: improving supervised outlier detection with unsupervised representation learning. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Rio de Janeiro, Brazil, July 2018.

[16] **Yue Zhao**, Zhongtian Qiu, Yiqing Yang, Weiwei Li, and Mingming Fan. An empirical study of touch-based authentication methods on smartwatches. In *ACM International Symposium on Wearable Computers, (ISWC)*, pages 122–125. ACM, 2017.

[17] **Yue Zhao**, Maciej K Hryniewicki, Francesca Cheng, Boyang Fu, and Xiaoyu Zhu. Employee turnover prediction with machine learning: A reliable approach. In *Intelligent Systems Conference*, pages 737–758. Springer, 2018.

[18] **Yue Zhao**, Zain Nasrullah, Maciej K Hryniewicki, and Zheng Li. LSCP: locally selective combination in parallel outlier ensembles. In *SIAM International Conference on Data Mining (SDM)*, pages 585–593. SIAM, May 2019.

[19] **Yue Zhao**, Zain Nasrullah, and Zheng Li. PyOD: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research (JMLR)*, 20(96):1–7, 2019.

[20] **Yue Zhao**, Xuejian Wang, Cheng Cheng, and Xueying Ding. Combining machine learning models and scores using combo library. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[21] **Yue Zhao**, Xiyang Hu, Cheng Cheng, Cong Wang, Changlin Wan, Wen Wang, Jianing Yang, Haoping Bai, Zheng Li, Cao Xiao, Yunlong Wang, Zhi Qiao, Jimeng Sun, and Leman Akoglu. SUOD: Accelerating large-scale unsupervised heterogeneous outlier detection. *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

[22] **Yue Zhao**, Zhi Qiao, Cao Xiao, Lucas Glass, and Jimeng Sun. Pyhealth: A python library for health predictive models. *arXiv preprint arXiv:2101.04209*, 2021.

[23] **Yue Zhao**, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[24] **Yue Zhao***, Zheng Li*, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2022. (**Equal contribution**).

[25] **Yue Zhao***, Kay Liu*, Yingtong Dou*, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, et al. BOND: Benchmarking unsupervised outlier node detection on static attributed graphs. *Neural Information Processing Systems (NeurIPS)*, 2022. (**Equal contribution**).

[26] **Yue Zhao***, Kay Liu*, Yingtong Dou*, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, et al. PyGOD: A python library for graph outlier detection. *arXiv preprint arXiv:2204.12095*, 2022. (**Equal contribution**).

[27] **Yue Zhao**, Sean Zhang, and Leman Akoglu. ELECT: Toward unsupervised outlier model selection. In *IEEE International Conference on Data Mining (ICDM)*. IEEE, 2022. **Accepted, to appear**.

[28] **Yue Zhao**, George H Chen, and Zhihao Jia. TOD: Gpu-accelerated outlier detection via tensor operations. In *International Conference on Very Large Databases (VLDB)*, 2023.

[29] **Yue Zhao**, Guoqing Zheng, Subhabrata Mukherjee, Robert McCann, and Ahmed Awadallah. ADMoE: Anomaly detection with mixture-of-experts from noisy labels. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[30] Kees M van Hespen, Jaco JM Zwanenburg, Jan W Dankbaar, Mirjam I Geerlings, Jeroen Hendrikse, and Hugo J Kuijf. An anomaly detection approach to identify chronic brain infarcts on mri. *Scientific Reports*, 11(1):1–10, 2021.

[31] Changlin Wan, Dongya Jia, **Yue Zhao**, Wennan Chang, Sha Cao, Xiao Wang, and Chi Zhang. A data denoising approach to optimize functional clustering of single cell rna-sequencing data. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 217–222. IEEE, 2020.

[32] Zhiming Xu, Xiao Huang, **Yue Zhao**, Yushun Dong, and Jundong Li. Contrastive attributed network anomaly detection with data augmentation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2022.

[33] Ling Yang, Zhilong Zhang, Shenda Hong, Runsheng Xu, **Yue Zhao**, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022. **Under review**.

[34] Shenglin Zhang, Zhenyu Zhong, Dongwen Li, Qiliang Fan, Yongqian Sun, Man Zhu, Yuzhi Zhang, Dan Pei, Jiyan Sun, Yinlong Liu, et al. Efficient kpi anomaly detection through transfer learning for large-scale web services. *IEEE Journal on Selected Areas in Communications*, 2022.