Snowmass2021 – Letter of Interest

Self-driving data trigger, filtering, and acquisition systems for high-throughput physics facilities

Instrumentation Frontier – Thematic Areas:

☐ (IF1) Quantum Sensors
☐ (IF2) Photon Detectors
☐ (IF3) Solid State Detectors and Tracking
■ (IF4) Trigger and DAQ
☐ (IF5) Micro Pattern Gas Detectors (MPGDs)
☐ (IF6) Calorimetry
■ (IF7) Electronics/ASICs
☐ (IF8) Noble Elements
☐ (IF9) Cross Cutting and Systems Integration

Computational Frontier – Thematic Areas:

☐ (CompF1) Experimental Algorithm Parallelization
☐ (CompF2) Theoretical Calculations and Simulation
■ (CompF3) Machine Learning
■ (CompF4) Storage and processing resource access
☐ (CompF5) End user analysis
☐ (CompF6) Quantum computing
☐ (CompF7) Reinterpretation and long-term preservation of data and code

Contact Information:
David W. Miller, University of Chicago, Chicago, IL
Contact Email: David.W.Miller@uchicago.edu

Authors:
Yuxin Chen, University of Chicago, IL
Kristin Dona, University of Chicago, IL
Chinmaya Mahesh, University of Illinois at Urbana-Champaign, IL
David W. Miller, University of Chicago, Chicago, IL
Nhan Tran, Fermi National Accelerator Laboratory, Batavia, IL

Abstract:

Data-intensive physics facilities are increasingly reliant on real-time processing capabilities and machine learning workflows, in order to filter and analyze the extreme volumes of data being collected. This is especially true at the energy and intensity frontiers of particle physics where bandwidths of raw data can exceed 100 Tb/s of heterogeneous, high-dimensional data sourced from >300M individual sensors. Data triggering and filtering algorithms targeted at the discovery science performed at future facilities must operate at the level of 1 part in 10^5. Once executed, these algorithms drive the data curation process, funneling event records with certain features into categories that are predefined based on the labels extracted by the trigger algorithms. The design, implementation, monitoring, and usage of these trigger algorithms is very high-dimensional, resource-intensive, and can include significant blindspots.

This *Letter of Interest* aims to express the need to investigate the concept of a *self-driving trigger system* that is able to learn the hyper-dimensional space of data that are processed – and potentially discarded – and thereby autonomously and continuously learn to more efficiently and effectively select, filter, and process data from a particular facility. This concept has the potential to not only increase the performance of such systems, but also to increase discovery potential by moving beyond previous paradigms of fixed menus of carefully hand-curated data.

Letter of Interest:

Data filtering algorithms – so-called trigger algorithms – targeted at discovery science (e.g. identification of a data event containing evidence of dark matter produced at the Large Hadron Collider) – must operate at the level of 1 part in 10^5 due to numerous bandwidth, compute, and storage-related constraints. Once executed, these algorithms often drive the data curation process, funneling event records with certain features into categories that are predefined based on the labels extracted by those algorithms. The design, implementation, monitoring, and usage of these trigger algorithms is resource-intensive and can include significant blindspots, in part because the menu of trigger algorithms is manually designed based on domain knowledge [13]. Although discovery science has been based on this approach for decades, those observations relied heavily on excellent prior knowledge of the feature space being probed. Future particle physics facilities and discoveries may not follow this trend, and thus a new paradigm is needed in which specifically engineered features defined with respect to manually identified categories are not required.

In this *Letter of Interest* we aim to communicate the need to consider more *dynamic* approaches to sampling the hyper-dimensional space probed by the detection systems at particle colliders. Specifically, we would like to posit the idea to automatically design and refine the trigger and data filtering algorithms at future physics facilities by weaving together recent advances in explainable AI [1, 10, 14, 18], active learning [2, 3, 4, 7, 11, 12, 20], reinforcement learning [8, 17], and other approaches that take into account the vast availability of simulated and real data, along with the traditional approach to producing a hand-designed trigger menu.

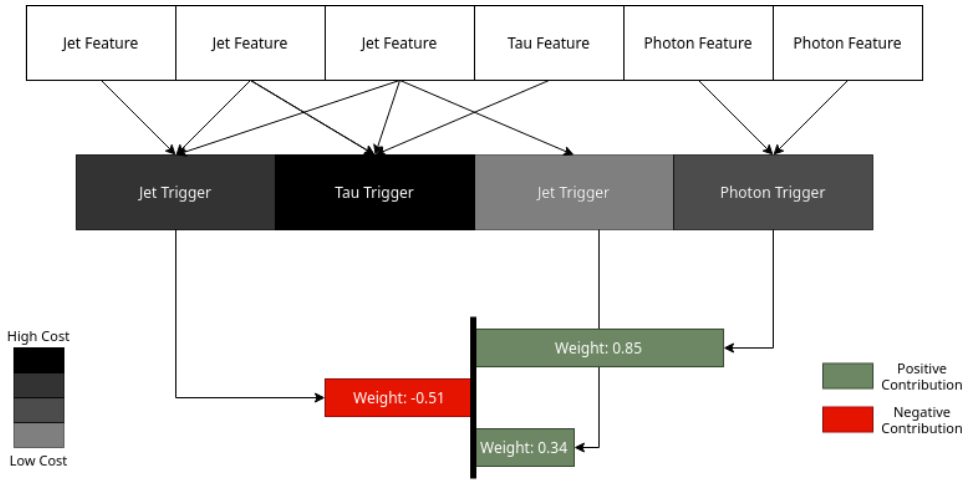

Figure 1: An "open-box" predictive model that deciphers the trigger menu with automated explanations and an associated cost model. Our data-driven trigger system interprets the trigger decision for a given event record by (1) learning a mapping from the *physics features* (top row) to the *labels* extracted by the trigger algorithms from the existing trigger menu (middle row), and (2) generating *explanations* of the trigger decisions (i.e. to keep or discard) by automatically identifying an efficient set of trigger algorithms that contribute the most to the decision (bottom row). In the explanation diagram, larger weight implies that the corresponding label contributes more to the decision.

Data-driven Modeling and Optimal Design of Trigger Menu

To accomplish such a daunting task, it is essential that the potential for *explainability* first be established. As illustated in Figure 1, we advocate to first conceive of a space of cost models by which the data filtering and curation

process can be quantifiably assessed. Progress towards a fully functional automated data filtering and processing system must start by constructing a learning-based filter system that is able to reproduce the current, hand-designed, "trigger menus." More importantly, though, in order to be useful, this new system must be able to *explain* its selections in the context of both low and high-level features (e.g. physics object multiplicities and kinematics). This is the first step towards constructing an active continuous learning model that is able to update itself and provide *explanations* for those updates. A distinguishing aspect of this research effort is its focus on model *interpretability*: Instead of a closed-box model that is capable of recovering the original data distribution, we aim to design an "open-box" predictive model, which, for any given input, not only outputs a decision (e.g., "keep this data point"), but also explains why we should do so, by associating the decision with the existing rules in the hand-designed trigger menu.

In addition to the overall development of this "open-box" predictive model is the desire to minimize the latency of the trigger system. Given an incoming data event, each trigger algorithm incurs a latency at runtime – assuming that algorithms are run in parallel, the latency of the trigger system depends on the worst-case running time of all trigger algorithms. Thus, for each data event, finding the most efficient set of trigger algorithms at run time is crucial for a real-time trigger system. Concretely, to address the real-time data processing challenge, we investigate the following combinatorial optimization problem: given a ground set of candidate trigger algorithms from the existing trigger menu and the latency cost for each trigger algorithm, we seek an optimal subset of trigger algorithms for each incoming data event, such that the selected algorithms can jointly make the correct filtering decision with the *minimal latency cost*. We will then employ the solution of the above optimization problem as the explanation of our "open-box" predictive model, which in turn will be used to optimize the latency of the existing trigger system. Here, we highlight a significant challenge during this process: Since the effectiveness of any subset of trigger algorithm is often unknown *a priori* and needs to be learned from data, this task essentially amounts to building a data-driven, cost-sensitive explainable model—an emerging topic at the frontier of machine learning and operations research.

Automated Trigger Menu Refinement via Active Learning

In order to investigate *new types of event data*, which may not be captured by the existing trigger menu, a long term goal is to adaptively update the trigger menu as new data comes in. Based upon the open-box predictive model constructed from existing event data, our self-driving trigger system will draw upon another two promising threads of AI research: *online learning* [5, 16] and *active learning* [2, 7, 9, 11, 12, 19, 20, 21], to automatically refine the selections of trigger algorithms by adapting to the new data stream. Both the online and active learning literature deal with streaming data—where unlabeled data points arrive one at a time—a setting that fits the high-throughput particle physics application particularly well. In comparison to the classical supervised data-driven models, online learning predictive models constantly evolve, updating their decision rules as each new data point comes in. On the other hand, active learning models are built to autonomously decide which data points are the most promising to keep to help make future decisions. By leveraging tools from these fields, we aim to properly capture the uncertainty of our "open-box" predictive model on the incoming, unseen data distribution. We will then use such uncertainty measure to design a principled active sampling framework to explore the high-dimensional data space to avoid significant blind spots, and hence refine the trigger system on the fly.

While an active sampling system greatly encourages exploration of novel data events, the cost of computing the optimal active sampling strategy often involves solving non-trivial planing and optimization problems, and can be prohibitive to run (e.g. when the trigger system demands real-time processing capabilities). In such cases, it is desirable to design a decision making system that can learn from the previous decision histories and make efficient predictions at run time [6, 15] (as opposed to solving expensive optimization problems for each incoming data point). Therefore, our long-term goal is to fully exploit the potential of the high throughput data steam, to developed a learning-based framework for stream-based active learning, which takes historical trigger decision records as training data, and learns an efficient active sampling policy on unseen dataset. We view this research as laying the foundations of a scalable, data-driven real-time trigger system. Successful extraction of a succinct set of physical intuitions behind a decision will greatly facilitate human understanding of the phenomenon. As such, it serves as a crucial bridge between the *functionality* and the *accessibility* of the model, and hence is a deciding factor for the real-world deployment of the data-driven trigger system.

References

[1] O. M. Aodha, S. Su, Y. Chen, P. Perona, and Y. Yue. "Teaching Categories to Human Learners with Visual Explanations". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. Apr. 2018.

[2] M.-R. Bouguelia, Y. Belaïd, and A. Belaïd. "A stream-based semi-supervised active learning approach for document classification". In: *2013 12th International Conference on Document Analysis and Recognition*. IEEE. 2013. 611–615

[3] Y. Chen, S. H. Hassani, A. Karbasi, and A. Krause. "Sequential Information Maximization: When is Greedy Near-optimal?" In: *Proc. International Conference on Learning Theory (COLT)*. July 2015.

[4] Y. Chen and A. Krause. "Near-optimal Batch Mode Active Learning and Adaptive Submodular Optimization". In: *International Conference on Machine Learning (ICML)*. 2013.

[5] Y. Chen, J.-M. Renders, M. H. Chehreghani, and A. Krause. "Efficient Online Learning for Optimizing Value of Information: Theory and Application to Interactive Troubleshooting". In: *Proc. Conference on Uncertainty in AI (UAI)*. Aug. 2017.

[6] S. Chernova and A. L. Thomaz. "Robot learning from human teachers". *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **8**:3, 1–121, 2014.

[7] W. Chu, M. Zinkevich, L. Li, A. Thomas, and B. Tseng. "Unbiased online active learning in data streams". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011. 195–203

[8] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, et al. "Bayesian Reinforcement Learning: A Survey". *Foundations and Trends® in Machine Learning*, **8**:5-6, 359–483, 2015.

[9] A. G. Hauptmann, W.-H. Lin, R. Yan, J. Yang, and M.-Y. Chen. "Extreme video retrieval: joint maximization of human and computer performance". In: *Proceedings of the 14th ACM international conference on Multimedia*. ACM. 2006. 385–394

[10] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, et al. "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model". *The Annals of Applied Statistics*, **9**:3, 1350–1371, 2015.

[11] Y. Liu. "Active learning with support vector machine applied to gene expression data for cancer classification". *Journal of chemical information and computer sciences*, **44**:6, 1936–1941, 2004.

[12] C. C. Loy, T. M. Hospedales, T. Xiang, and S. Gong. "Stream-based joint exploration-exploitation active learning". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. 2012. 1560–1567

[13] A. P. Note. "Trigger Menu in 2017". 2018.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?: Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. 2016. 1135–1144

[15] S. Ross, G. Gordon, and D. Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011. 627–635

[16] S. Shalev-Shwartz et al. "Online learning and online convex optimization". *Foundations and trends in Machine Learning*, **4**:2, 107–194, 2011.

[17] C. Szepesvári. "Algorithms for reinforcement learning". *Morgan and Claypool*, 2009.

[18] R. Tibshirani. "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288, 1996.

[19] S. Tong and E. Chang. "Support vector machine active learning for image retrieval". In: *Proceedings of the ninth ACM international conference on Multimedia*. ACM. 2001. 107–118

[20] S. Tong and D. Koller. "Support vector machine active learning with applications to text classification". *Journal of machine learning research*, **2**:Nov, 45–66, 2001.

[21] G. Tur, D. Hakkani-Tür, and R. E. Schapire. "Combining active and semi-supervised learning for spoken language understanding". *Speech Communication*, **45**:2, 171–186, 2005.