RATIONAL INTELLIGENCE

CISPA HELMHOLTZ-ZENTRUM FÜR INFORMATIONSSICHERHEIT

Bachelor and Master Theses Opportunities

Saarbrücken, August 10th, 2025

The Rational Intelligence Lab (https://ri-lab.org/) at the CISPA Helmholtz Center for Information Security in Saarbrücken, Germany (https://cispa.de/en), is seeking motivated bachelor's and master's students to work on compelling thesis topics in the field of **rational machine learning**, with the aim of enhancing the rationality, efficiency, and reliability of intelligent systems. Research topics of interest include, but are not limited to, out-of-distribution generalization, causality, incentive-aware machine learning, distributed and collaborative learning, imprecise probabilistic machine learning, human-AI collaboration, learning with preference and choice data, and learning in games.

For more information on our past and current research, please visit our publications page at https://ri-lab.org/pubs/. You can also check out the core values we are adhering to at https://ri-lab.org/values.

What We Look For

We are looking for highly motivated bachelor's or master's students with strong mathematical skills, coding proficiency, or both. A strong background in basic machine learning is required.

We are committed to fostering a diverse workforce and strongly encourage applications from individuals in underrepresented communities, including but not limited to women, people of color, LGBTQ+ individuals, persons with disabilities, and those from economically disadvantaged backgrounds.

What We Offer

Guided by experienced researchers, the students will conduct a cutting-edge research in the areas situated at *the intersection of machine learning, statistics, and computer science* with the goal of enhancing the rationality, efficiency, and reliability of intelligent systems. Specifically, we are seeking bachelor's and master's students to work on the following topics:

Project 1: Hilbert Space Embedding of Probability Distributions.

Embedding probability distributions into reproducing kernel Hilbert spaces (RKHS) has enabled powerful nonparametric methods such as the maximum mean discrepancy (MMD), a statistical distance with strong theoretical and computational properties [Borgwardt et al., 2006, Gretton et al., 2012, Muandet et al., 2016]. At its core, the MMD relies on kernel mean embeddings to represent distributions as mean functions in RKHS. However, it remains unclear if the mean function is the only meaningful RKHS representation. In Naslidnyk et al. [2025], we introduced the notion of **kernel quantile embeddings (KQEs)** and constructed a family of distances that: (i) are probability metrics under weaker kernel conditions than MMD; (ii) recover a kernelised form of the sliced Wasserstein distance; and (iii) can be efficiently estimated with near-linear cost. Through hypothesis testing, we show that these distances offer a competitive alternative to MMD and its fast approximations.

In this project, the students will build on our previous work by focusing on one or more of the following objectives:

- **Improving Estimates**: Explore more sophisticated methods to improve the empirical estimates of Kernel Quantile Embeddings (KQEs).
- **Generalizing KQEs**: Extend KQEs to represent conditional distributions and important probabilistic rules such as sum, product, and Bayes rules.
- **Expanding Applications**: Apply KQEs to existing applications in a wide range of domains, including not only nonparametric two-sample testing, but also (conditional) independence testing, causal inference, reinforcement learning, learning on distributions, generative modeling, and robust parameter estimation, among others.

Project 2: Out-of-Distribution and Compositional Generalisation.

The capability to generalise knowledge, a hallmark of both biological and artificial intelligence (AI), has seen remarkable progress in recent years, but despite notable achievements, contemporary AI systems may catastrophically fail when operated on out-of-domain (OOD) data because theoretical guarantees for their generalisation hinge on the assumption of independent and identically distributed (IID) training and deployment data, with empirical risk minimisation (ERM) being the dominant learning algorithm. In Singh et al. [2024], we pointed out an **institutional separation** between learners (aka forecasters) and operators (aka decision makers) of ML models as the key challenge in OOD generalisation and proposed an **Imprecise Risk Optimisation (IRO)** algorithm as the first step to addressing this problem. In [Singh et al., 2025], we frame this problem as one of eliciting imprecise forecasts. Our research shows that by allowing for strategic communication between forecasters and decision-makers, we can design **strictly proper scoring rules** that lead to truthful elicitation. This approach successfully overcomes previous impossibility results.

In this project, the students will build on our previous work by applying the algorithms developed by Singh et al. [2024] and Singh et al. [2025] to real-world applications. The key objectives include:

- **Evaluating Performance**: Applying the proposed algorithms to real-world benchmark datasets to demonstrate their practical benefits and drawbacks.
- **Developing Improvements**: Using the insights gained to develop new learning algorithms that improve upon the existing ones.
- **Extending the Framework**: Alternatively, extending the current framework to handle slightly different settings such as *compositional* generalization.

Project 3: Causal and Counterfactual Machine Learning.

Causal and counterfactual reasoning has been recognised as a hallmark of human and machine intelligence, and in the recent years, the machine learning community has taken up a rapidly growing interest in the subject, in particular in representation learning and natural language processing. In the past years, we have developed sophisticated methods for causal effect estimation [Muandet et al., 2021, Park et al., 2021], instrumental variable (IV) regression [Muandet et al., 2020, Kremer et al., 2022, Zhang et al., 2023], proximal causal learning [Mastouri et al., 2021], causal strategic learning [Vo et al., 2024], and counterfactual prediction [Quinzan et al., 2024], to name a few. Causal inference and analysis provides a dual benefit to the field of machine learning. It not only leverages the efficient computational tools developed by the machine learning community but also offers its own rigorous frameworks to tackle critical challenges like out-of-distribution generalization, algorithmic fairness, and privacy.

In this project, the students will work on the projects that will advance the field of **causal and counterfactual machine learning** by building on one of our previous papers. The key objectives include:

- **Evaluating Performance**: Applying one of the proposed algorithms to large-scale benchmark datasets to demonstrate its practical benefits and drawbacks.
- **Improving Algorithms**: Using the insights gained to develop new algorithms in similar or slightly different settings.
- **Exploring New Applications**: Investigating the potential benefits of causal and counterfactual reasoning for emerging research areas, such as generative modeling, large language models (LLMs), in-context learning (ICL), and security-related topics like memorization and membership inference attacks.

Other topics. Apart from the proposed projects, we are open to supervising students who have their own project ideas that align with our current research interests. Please contact muandet@cispa.de to discuss whether your project is a good fit.

> “ *If your thesis is accepted for publication at a top-tier international conference, you'll receive financial support to attend and present your paper.* ”

How To Apply

The students can apply by sending (1) a **resume** and (2) an up-to-date **transcript** to muandet@cispa.de. In your email, tell us which project you're most excited about and why you're the ideal candidate for it. The position will remain open until filled. Inquiries about the projects are always welcome.

References

K. Borgwardt, A. Gretton, M. Rasch, H.-P. Kriegel, B. Schölkopf, and A. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

H. Kremer, J.-J. Zhu, K. Muandet, and B. Schölkopf. Functional generalized empirical likelihood estimation for conditional moment restrictions. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11665–11682. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/kremer22a.html`.

A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. Kusner, A. Gretton, and K. Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7512–7523. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/mastouri21a.html`.

K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyonds. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2016. URL `https://www.nowpublishers.com/article/Details/MAL-060`.

K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/1c383cd30b7c298ab50293adfecb7b18-Abstract.html`.

K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021. URL `https://www.jmlr.org/papers/volume22/20-185/20-185.pdf`.

M. Naslidnyk, S. L. Chau, F.-X. Briol, and K. Muandet. Kernel quantile embeddings and associated probability metrics. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=9LqXn0Izwk`.

J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8401–8412. PMLR, July 2021. URL `https://proceedings.mlr.press/v139/park21c.html`. ISSN: 2640-3498.

F. Quinzan, C. Casolo, K. Muandet, Y. Luo, and N. Kilbertus. Learning counterfactually invariant predictors. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=pRt1Vw1DPs`.

A. Singh, S. L. Chau, S. Bouabid, and K. Muandet. Domain generalisation via imprecise learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. URL `https://proceedings.mlr.press/v235/singh24a.html`.

A. Singh, S. L. Chau, and K. Muandet. Truthful elicitation of imprecise forecasts, 2025. URL `https://arxiv.org/abs/2503.16395`.

K. Q. Vo, M. Aadil, S. L. Chau, and K. Muandet. Causal strategic learning with competitive selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15411–15419, 2024. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29466`.

R. Zhang, M. Imaizumi, B. Schölkopf, and K. Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1):20220073, 2023. URL `https://www.degruyterbrill.com/document/doi/10.1515/jci-2022-0073/html?lang=en&srsltid=AfmBOoq-QX6alxUAg5XtS9ei4_OGu5coSOlAGuM-BWI1opxCLMYIJ6aw`.