Julian Martin Rodemann

Uncertainty Quantification in Data-Centric Machine Learning: Some Statistical Perspectives

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München
zur Erlangung des akademischen Doktorgrads der Naturwissenschaften (Dr. rer. nat.)

Eingereicht am 31.10.2025

Julian Martin Rodemann

Uncertainty Quantification in Data-Centric Machine Learning: Some Statistical Perspectives

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Vorgelegt von Julian Martin Rodemann aus Tübingen am 31.10.2025

Betreuer: Prof. Dr. Thomas Augustin
Erster Berichterstatter: Prof. Dr. Thomas Nagler
Zweiter Berichterstatter: Prof. Sébastien Destercke, PhD.
Dritter Berichterstatter: Prof. Lev V. Utkin, PhD., DSc.

Tag der Einreichung: 31.10.2025
Tag der Disputation: 06.03.2026

To Lotti

Acknowledgments

I would like to express my deepest gratitude to everyone who supported me on the journey of my doctoral studies. Among the many people who helped me to complete my dissertation, an especially heartfelt thank you goes to ...

- ... *my supervisor Thomas Augustin for supporting me throughout the years, always having my back, sparking my interest in decision theory and imprecise probabilities and—most importantly—for trusting and believing in me even when I did not. I would like to thank you for always taking the time for my (many...) questions, for sharing fun (statistical) anecdotes and for inspiring me to always try to get to the root of a problem. You fundamentally shaped me as a statistician, but also as a person. Thank you, Thomas, for your patient, thoughtful, and generous supervision that made this work possible!*
- ... *Thomas Nagler, Sébastien Destercke and Lev Utkin for their willingness to serve as reviewers of my PhD thesis.*
- ... *Volker Schmid for chairing the examination committee as well as Fabian Scheipl for serving as alternate.*
- ... *my co-authors Thomas Augustin, Christoph Jansen, Esteban Garcés Arias, Hannah Blocher, Georg Schollmeyer, Matthias Aßenmacher, Christian Heumann, Meimingwei Li, Eyke Hüllermeier, Malte Nalenz, James Bailie, Dominik Kreiß, Federico Croppi, Philipp Arens, Yusuf Sale, Julia Herbinger, Bernd Bischl, Conor J. Walsh, Giuseppe Casalicchio, Jann Goschenhofer, Emilio Dorigatti, Thomas Nagler, Chongsheng Zhang, Stefan Dietrich, Christoph Luther, Yuanhao Ding, Danlu Chen, Gaojuan Fan, Fabian Bongratz, Vladimir Golkov, Lukas Mautner, Luca Della Libera, Frederik Heetmeyer, Felix Czaja and Daniel Cremers. Thank you for your valuable hints, your continuous support and all the constructive feedback. Above all, however, I am deeply grateful for all the stimulating and long discussions we had. Research is a collaborative effort that lives from the* ***people*** *who pursue it. Without you, I would neither have published any papers nor completed this dissertation—and I would have had only half as much fun.*
- ... *all former and current members of the working group: Thomas Augustin, Hannah Blocher, Cornelia Fütterer, Polina Gordienko, Lea Höhler, Christoph Jansen, Gilbert Kiprotich, Dominik Kreiß, Malte Nalenz, Daniel Schlichting, Georg Schollmeyer and Patrick Schwaferts. Thank you for welcoming me so warmly in the beginning, for your valuable advice and countless enjoyable lunch breaks! I would like to especially thank the (former) postdoctoral researchers Christoph Jansen and Georg Schollmeyer for supporting and guiding me at the beginning of my PhD. I have benefited a lot from your mentorship and enjoy our fruitful ongoing collaborations!*
- ... *the graduate center of the Bavarian Institute for Digital Transformation (bidt) within the Bavarian Academy of Sciences (BAS) for supporting my doctoral studies through a scholarship. The graduate center was a great place to learn about the many facets of the digital world that are beyond my own research. Moreover, I am very grateful for the support from the mentoring program of the Faculty of Mathematics, Computer Science and Statistics at LMU Munich.*

... *James Bailie and Xiao-Li Meng for hosting me during my research stay at the Department of Statistics at Harvard University in the fall of 2024. Thank you, both, for also visiting Germany in return!*

... *Hannah Blocher, Christoph Jansen and James Bailie for proofreading (parts of) this thesis. Thank you so much for all your valuable remarks! It goes without saying that all remaining errors are solely my own.*

... *Brigitte Maxa, Martina Brunner and Elke Höfner for their administrative support throughout my employment at the Department of Statistics at LMU Munich.*

... *all students who attended my tutorials, classes and lectures at LMU Munich. Thank you for your curiosity, all your great questions and your fresh perspectives! You reminded me why I love what I do; and I am grateful for everything you taught me in return—it really was a* ***reciprocal learning*** *(see Contribution 1, pun intended) process.*

... *the Department's cleaning staff for keeping my office clean and for sharing sweets and cakes with me during coffee breaks in the old library. Without all the sugar accompanying my coffee, I would not have made it through my PhD...*

... *Dominik Kreiß and Lea Höhler for being great office mates!*

... *all former and current colleagues at the Department of Statistics for the pleasant, open-minded and relaxed work atmosphere.*

... *all the open-source developers whose software (libraries) enabled me to implement machine learning and statistical methods, thus putting theory into practice (Kurt Lewin: "There is nothing so practical as a good theory").*

... *my former study mates Esteban Garcés Arias, Federico Croppi, Sebastian Fischer, Alex Piehler and Leonard Rosen for enthralling discussions in study groups that sparked my interest in statistics and made me pursue a PhD in this wonderful, mind-bending field.*

... *all my friends and my whole family (in particular, my siblings) for your unwavering support, for taking my mind off things when I was deeply immersed in research and for reminding me that there are other things in life than statistics—and way more important things than a PhD.*

... *Lotti for being the love of my life. Thank you for sharing your calm and serenity in the midst of chaos, and for your unconditional support, patience, reassurance and comfort, as well as your unshakable optimism.*

... *my parents. For everything.*

Summary

Machine learning and statistics allow for conclusions about something unknown (population) based on limited observations thereof (data) and assumptions thereon (model). These conclusions come with uncertainty, which can originate in any of the three elements: the unknown population itself (hence irreducible uncertainty) or the data and the model (both reducible). Data contributes to this uncertainty in a quantitative and a qualitative way. While the former contribution is self-explanatory and monocausal (too few observations), the latter contribution can be due to complex data collection, pre-processing, merging and the like.

Data-centric machine learning refers to methods that account for or directly entail such data *selection* steps. Here, data is used for two purposes: to draw conclusions about the population in the first place (training) as well as to evaluate these conclusions' quality later (testing). This cumulative dissertation studies the *selection* of these two kinds of data: training data (Part III, Contributions 1–9) and testing data (Part IV, Contributions 10–14). By quantifying the involved uncertainty, the dissertation aims to advance the reliability and trustworthiness of data-centric machine learning.

As it will turn out, this endeavor requires a closer look at interactions and feedback loops among all three elements from above: population, data and model. The stylized separation among the three will prove illusory. In particular, Part III of the dissertation will demonstrate that various machine learning algorithms let the model *self*-select the training data, with far-reaching consequences for statistical inference from such data. The dissertation answers the questions of whether and to what degree reliable conclusions about the population are still possible in this scenario. In a similar spirit, Part IV investigates how the selection of testing data for (multicriteria) benchmarking algorithms affects the validity of the benchmarking results. Here, a special emphasis is also put on how this validity depends on multiple criteria and on the way they are aggregated. Part III and Part IV fundamentally rely on decision-theoretic embeddings of training and testing data selection, respectively.

All in all, the dissertation offers novel insights into quantifying uncertainty originating from data selection in statistics and machine learning. They lead to more robust and reliable methods as well as critical assessments of existing ones. Both aspects contribute to a safer, more sustainable and less harmful usage of machine learning and statistics.

Zusammenfassung

Maschinelles Lernen und Statistik ermöglichen Einsichten in etwas Unbekanntes (Population) durch begrenzte Beobachtungen des Unbekannten (Daten) und Annahmen über das Unbekannte (Modell). Diese Einsichten sind mit Unsicherheiten behaftet. Die Quellen der Unsicherheiten können in allen drei Elementen liegen: in der unbekannten Population selbst (nicht reduzierbare Unsicherheit) oder in den Daten und dem Modell (beide reduzierbar). Daten tragen quantitativ und qualitativ zu letzterer Unsicherheit bei. Während die quantitative Komponente selbsterklärend und monokausal ist (zu wenige Beobachtungen), kann die qualitative Komponente auf komplexe Datenerfassung, Vorverarbeitung, Zusammenführung und Ähnliches zurückzuführen sein.

Datenzentriertes maschinelles Lernen bezieht sich auf Methoden, die solche Datenauswahlschritte berücksichtigen oder direkt mit sich bringen. Hier werden Daten für zwei Zwecke verwendet: zunächst, um Schlussfolgerungen über die Population zu ziehen (Training) und später dann, um die Qualität dieser Schlussfolgerungen zu bewerten (Test). Diese aus 14 Beiträgen bestehende kumulative Dissertation untersucht die *Auswahl* dieser beiden Arten von Daten: Trainingsdaten (Teil III, Beiträge 1–9) und Testdaten (Teil IV, Beiträge 10–14). Durch die Quantifizierung der damit verbundenen Unsicherheit setzt sich die vorliegende Dissertation zum Ziel, die Zuverlässigkeit und Vertrauenswürdigkeit des datenzentrierten maschinellen Lernens zu verbessern.

Wie sich herausstellen wird, erfordert dieses Unterfangen eine genauere Betrachtung der Wechselwirkungen und Rückkopplungsschleifen zwischen allen drei oben genannten Elementen: Population, Daten und Modell. Die stilisierte Trennung zwischen den drei Elementen wird sich als illusorisch erweisen. Insbesondere wird Teil III der Dissertation zeigen, dass verschiedene Algorithmen des maschinellen Lernens das Modell die Trainingsdaten selbst auswählen lassen, was weitreichende Konsequenzen für die statistische Inferenz aus solchen Daten hat. Die Dissertation beantwortet die Frage, ob und inwieweit in diesem Szenario noch zuverlässige Schlussfolgerungen über die Population möglich sind. In ähnlicher Weise untersucht Teil IV, wie sich die Auswahl von Testdaten für das multikriterielle Benchmarking von Algorithmen auf die Validität der Benchmarking-Ergebnisse auswirkt. Dabei wird auch darauf eingegangen, wie diese Validität von mehreren Kriterien abhängt und insbesondere davon, wie diese Kriterien aggregiert werden. Teil III und Teil IV basieren grundlegend auf entscheidungstheoretischen Einbettungen der Auswahl von Trainings- beziehungsweise von Testdaten.

Insgesamt bietet die Dissertation neue Erkenntnisse zur Quantifizierung von Unsicherheiten, die aus der Datenauswahl in der Statistik und im maschinellen Lernen resultieren. Diese führen zu robusteren und zuverlässigeren Methoden sowie zu einer kritischen Bewertung bestehender Methoden. Beides trägt zu einer sichereren, nachhaltigeren und weniger schädlichen Nutzung von maschinellem Lernen und Statistik bei.

Contents

List of Figures

List of Tables

Contributing Publications

This dissertation is composed of the following contributions. They will be referred to as Contribution 1 to Contribution 14 in what follows.

1. **Julian Rodemann**, Christoph Jansen, and Georg Schollmeyer (2024). "Reciprocal Learning". In *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37, 1686–1724. (Rodemann et al., 2024)

2. **Julian Rodemann** and Thomas Augustin (2024). "Imprecise Bayesian Optimization". *Knowledge-Based Systems* 300:112186. (Rodemann and Augustin, 2024)

3. **Julian Rodemann**, Federico Croppi, Philipp Arens, Yusuf Sale, Julia Herbinger, Bernd Bischl, Eyke Hüllermeier, Thomas Augustin, Conor J. Walsh, and Giuseppe Casalicchio (2025). "Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration for Exosuit Personalization". In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, Research Track and Applied Data Science Track. Ed. by R. P. Ribeiro, B. Pfahringer, N. Japkowicz, P. Larrañaga, A. M. Jorge, C. Soares, P. H. Abreu, J. Gama. Springer, 525–542. (Rodemann et al., 2025b)

4. **Julian Rodemann**, Jann Goschenhofer, Emilio Dorigatti, Thomas Nagler, and Thomas Augustin (2023). "Approximately Bayes-Optimal Pseudo Label Selection". In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by R. J. Evans, and I. Shpitser. PMLR, 1762–1773. (Rodemann et al., 2023b)

5. **Julian Rodemann**, Christoph Jansen, Georg Schollmeyer, and Thomas Augustin (2023). "In All Likelihoods: Robust Selection of Pseudo-Labeled Data". In *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by E. Miranda, I. Montes, E. Quaeghebeur, and B. Vantaggi. PMLR, Vol. 215. 412–425. (Rodemann et al., 2023c)

6. Stefan Dietrich, **Julian Rodemann**, and Christoph Jansen (2024). "Semi-Supervised Learning Guided by the Generalized Bayes Rule Under Soft Revision". In *International Conference on Soft Methods in Probability and Statistics (SMPS)*. Ed. by J. Ansari, S. Fuchs, W. Trutschnig, M. A. Lubiano, M. A. Gil, P. Grzegorzewski, O. Hryniewicz. Springer, 110–117. (Dietrich et al., 2024)

7. **Julian Rodemann**, Dominik Kreiß, Eyke Hüllermeier, and Thomas Augustin (2022). "Levelwise Data Disambiguation by Cautious Superset Learning". In *International Conference on Scalable Uncertainty Management (SUM)*. Ed. by F. Dupin de Saint-Cyr, M. Öztürk-Escoffier, N. Potyka. Springer, 263–276. (Rodemann et al., 2022b)

8. **Julian Rodemann** and James Bailie (2025). "Generalization Bounds and Stopping Rules for Learning with Self-Selected Data". *arXiv preprint* arXiv:2505.07367 (last accessed October 4 2025). *Under review at the Journal of Machine Learning Research (JMLR)*. (Rodemann and Bailie, 2025)

9. Malte Nalenz, **Julian Rodemann**, and Thomas Augustin (2024). "Learning De-Biased Regression Trees and Forests from Complex Samples". *Machine Learning* 113:3379—3398. (Nalenz et al., 2024)

10. Christoph Jansen, Georg Schollmeyer, Hannah Blocher, **Julian Rodemann**, and Thomas Augustin (2023). "Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement". In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by R.J. Evans, and I. Shpitser. PMLR, 941—952. (Jansen et al., 2023b)

11. Christoph Jansen*, Georg Schollmeyer*, **Julian Rodemann***, Hannah Blocher*, and Thomas Augustin (2024). "Statistical Multicriteria Benchmarking via the GSD-Front". **Spotlight Award.** In *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37, 98143—98179. (Jansen et al., 2024)

12. **Julian Rodemann*** and Hannah Blocher* (2024). "Partial Rankings of Optimizers". In *International Conference on Learning Representations (ICLR), Tiny Papers Track*. Ed. by T. F. Burns and K. Maughan. OpenReview.net. (Rodemann and Blocher, 2024)

13. Esteban Garcés Arias, Hannah Blocher, **Julian Rodemann**, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher (2025). "Towards Better Open-Ended Text Generation: A Multicriteria Evaluation Framework". In *ACL Workshop on Generation, Evaluation and Metrics (GEM)*. Ed. by O. Arviv, M. Clinciu, K. Dhole, R. Dror, S. Gehrmann, E. Habba, I. Itzhak, S. Mille, Y. Perlitz, E. Santus, J. Sedoc, M. Shmueli Scheuer, G. Stanovsky, O. Tafjord. Association for Computational Linguistics, 631-–654. (Garcés Arias et al., 2025b)

14. Esteban Garcés Arias, Hannah Blocher, **Julian Rodemann**, Matthias Aßenmacher, Christoph Jansen (2025). Statistical Multicriteria Evaluation of LLM-Generated Text. *arXiv preprint arXiv:2506.18082* (last accessed October 15 2025). In *18th International Natural Language Generation Conference (INLG)* (forthcoming). (Garcés Arias et al., 2025a)

*These authors contributed equally to this work.

Brief Summaries of Contributions

Contribution 1: Julian Rodemann, Christoph Jansen, and Georg Schollmeyer (2024). "Reciprocal Learning". In *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37, 1686–1724.

This Contribution generalizes Bayesian optimization, boosting, active learning, bandits, self-training in semi-supervised learning and superset learning to *reciprocal learning*. In this way, it sets the stage for Contributions 2 and 3 on Bayesian optimization, Contributions 4 through 6 on self-training, and Contribution 7 on superset learning. In Contribution 1, we show that all these algorithms not only learn parameters from data but also vice versa: Learning goes both ways, as the algorithms iteratively change the sample in response to the model fit. Generalizing all these algorithms to reciprocal learning allows us to give novel convergence results for all of them, relying on a decision-theoretic embedding. The key is to guarantee that reciprocal learning constitutes a contraction mapping such that the Banach fixed-point theorem applies. We find that reciprocal learning converges at linear rates to an approximately optimal model under some specific conditions, which we relate to concrete active learning, self-training and bandit algorithms through corollaries.

Contribution 2: Julian Rodemann and Thomas Augustin (2024). "Imprecise Bayesian Optimization". *Knowledge-Based Systems* 300:112186.

While Contribution 1 studies *in-sample* convergence of reciprocal learning generally, this work turns to how fast reciprocal learning's special case of Bayesian optimization converges to *the population (global) optimum* under ambiguity about the model choice. Our approach is driven by both empirical and theoretical analyses of how the specification of the Gaussian Process (GP) prior affects the convergence of Bayesian Optimization (BO). A comprehensive simulation study reveals that among all prior components, the mean parameters of the prior have the greatest impact on BO's convergence. Accordingly, we focus specifically on this aspect of the GP prior. We establish regret bounds for BO when the mean parameters of the GP prior are misspecified. We demonstrate that while sublinear regret bounds become linear under GP misspecification, they remain sublinear if the misspecification-induced error is upper bounded by the GP's variance. Motivated by these findings, we introduce Prior-mean Robust Bayesian Optimization (PROBO), an extension of BO utilizing imprecise probabilities to avoid prior mean parameter misspecification. This is accomplished by explicitly modeling imprecision in the GP prior mean using a prior near-ignorance model. We apply PROBO to the problem of graphene production — a real-world optimization challenge in materials science — and observe that PROBO converges more rapidly than standard BO.

Contribution 3: Julian Rodemann, Federico Croppi, Philipp Arens, Yusuf Sale, Julia Herbinger, Bernd Bischl, Eyke Hüllermeier, Thomas Augustin, Conor J. Walsh, and Giuseppe Casalicchio (2025). "Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration for Exosuit Personalization". In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, Research Track and Applied Data Science Track. Ed. by R. P. Ribeiro, B. Pfahringer, N. Japkowicz, P. Larrañaga, A. M. Jorge, C. Soares, P. H. Abreu, J. Gama. Springer, 525-542.

Building on Contribution 2, this work focuses on Bayesian Optimization, too, but in a more specific application, namely human-in-the-loop Bayesian Optimization for personalizing wearable robotic devices. Here, a human can intervene and potentially reject proposals by the BO in case they do not align with their reasoning. We reduce BO's opacity by proposing ShapleyBO, a framework for interpreting BO's proposals by game-theoretic Shapley values. They quantify each parameter's contribution to BO's acquisition function. Exploiting the linearity of Shapley values, we are further able to identify how strongly each parameter drives BO's exploration and exploitation for additive acquisition functions like the confidence bound. We also show that ShapleyBO can disentangle the contributions to exploration into those that explore aleatoric and epistemic uncertainty. We demonstrate this information is useful for humans deciding on whether to accept or reject BO proposals in the use case of personalizing wearable robotic devices (assistive back exosuits) by human-in-the-loop BO. Results suggest human-BO teams with access to ShapleyBO can achieve lower regret than teams without.

Contribution 4: Julian Rodemann, Jann Goschenhofer, Emilio Dorigatti, Thomas Nagler, and Thomas Augustin (2023). "Approximately Bayes-Optimal Pseudo Label Selection". In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by R. J. Evans, and I. Shpitser. PMLR, 1762–1773.

This Contribution turns to another special case of reciprocal learning as introduced in Contribution 1: Self-training in semi-supervised learning, which strongly relies on pseudo-label selection (PLS). We introduce BPLS, a Bayesian framework for PLS, aiming to address the confirmation bias, a ubiquitous challenge in semi-supervised learning: Since labeled data are scarce, models tend to overfit, which can then be propagated via self-training to the final model, even after adding unlabeled data, which is typically available to a much greater extent. We embed PLS into decision theory and derive a Bayes criterion for selecting instances to label: an analytical approximation of the posterior predictive of pseudo-samples. We prove the Bayes-optimality of this criterion and overcome computational hurdles by approximating the criterion analytically. Its relation to the marginal likelihood allows us to come up with an approximation based on Laplace's method and the Gaussian integral. We empirically assess BPLS on simulated and real-world data. When faced with high-dimensional data prone to overfitting, BPLS outperforms traditional PLS methods.

Contribution 5: Julian Rodemann, Christoph Jansen, Georg Schollmeyer, and Thomas Augustin (2023). "In All Likelihoods: Robust Selection of Pseudo-Labeled Data". In *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by E. Miranda, I. Montes, E. Quaeghebeur, and B. Vantaggi. Vol. 215. PMLR, 412—425.

After having embedded PLS into decision theory in Contribution 4, this work proposes several robust extensions of Bayesian PLS under complex uncertainty, harnessing the rich literature on decision theory and imprecise probabilities. The idea is to select pseudo-labeled data that maximize a multi-objective utility function. The latter can be constructed to account for different sources of uncertainty. We provide three examples: model selection, error accumulation, and covariate shift. We highlight the use of three of our robust extensions on both simulated data and three real-world datasets. In a benchmarking study, we compare these extensions to conventional PLS methods. The results indicate that increased robustness in model selection can yield significant improvements in test accuracy.

Contribution 6: Stefan Dietrich, **Julian Rodemann**, and Christoph Jansen (2024). "Semi-Supervised Learning Guided by the Generalized Bayes Rule Under Soft Revision". In *International Conference on Soft Methods in Probability and Statistics (SMPS)*. Ed. by J. Ansari, S. Fuchs, W. Trutschnig, M.A. Lubiano, M.A. Gil, P. Grzegorzewski, O. Hryniewicz. Springer, 110—117.

Motivated by empirical findings in Contribution 5, we expand one of the therein proposed PLS methods, namely the Gamma-Maximin method with soft revision. It uses credal sets of priors ("generalized Bayes") to represent the epistemic modeling uncertainty. These credal sets are then updated by the Gamma-Maximin method with soft revision. That is, only those priors from the credal set are updated whose marginal likelihood is above a relative threshold. We then select pseudo-labeled data that are most likely in light of the least favorable posterior distribution from the credal set that we obtained in this manner. Inspired by the Laplace approximations based on the Gaussian integral in Contribution 4, we further simplify the corresponding optimization problem, allowing us to implement our method for logistic regression models. Empirically, we observe our Gamma-Maximin method with soft revision to outperform several competing approaches in terms of test accuracy, especially when the proportion of labeled data is low.

Contribution 7: Julian Rodemann, Dominik Kreiß, Eyke Hüllermeier, and Thomas Augustin (2022). "Levelwise Data Disambiguation by Cautious Superset Learning". In *International Conference on Scalable Uncertainty Management (SUM)*. Ed. by F. Dupin de Saint-Cyr, M. Öztürk-Escoffier, N. Potyka. Springer, 263–276.

This Contribution deals with superset learning, yet another special case of reciprocal learning (see Contribution 1). Superset learning can be seen as a generalization of semi-supervised learning (see Contributions 4 through 6). It aims to incorporate set-valued data in the learning process, (re)interpreting and utilizing its information in different manners. In Contribution 7, we propose a way to construct a hierarchical family of subsets within set-valued categorical observations. Each subset corresponds to a level of cautiousness, the smallest one as a singleton representing the most optimistic choice in the sense that the element of the observed set is chosen which is most beneficial for empirical risk minimization. We achieve this by finding instantiations whose corresponding empirical risks are below context-depending thresholds. Varying this threshold induces a hierarchy among those instantiations. In order to rule out ties, we select those instantiations whose optimal separations have the greatest generality. We apply our method to the prototypical example of yet undecided political voters pondering between different options.

To this end, we use both simulated data and pre-election polls by Civey including undecided voters for the 2021 German federal election.

Contribution 8: Julian Rodemann and James Bailie (2025). "Generalization Bounds and Stopping Rules for Learning with Self-Selected Data". *arXiv preprint* arXiv:2505.07367 (last accessed October 29 2025). *Under review at the Journal of Machine Learning Research (JMLR).*

While Contributions 2 through 7 advance specific instances of reciprocal learning (Contribution 1), this article takes a more principled perspective on reciprocal learning. Specifically, it answers the question of how effectively reciprocal learning algorithms *generalize* from their self-selected samples. We establish universal generalization bounds for reciprocal learning by leveraging covering numbers and Wasserstein ambiguity sets. Importantly, our results do not rely on any assumptions about the distribution of self-selected data; instead, they require only verifiable conditions on the reciprocal learning algorithms themselves. We provide guarantees for both convergent solutions and solutions after a finite number of iterations. As the latter are "anytime valid" in the sense of Ramdas et al. (2023), they enable practitioners to follow stopping rules that ensure out-of-sample performance for their reciprocal learning algorithms. Finally, we demonstrate our generalization bounds and stopping rules in the special context of semi-supervised learning as an instance of reciprocal learning.

Contribution 9: Malte Nalenz, **Julian Rodemann**, and Thomas Augustin (2024). "Learning De-Biased Regression Trees and Forests from Complex Samples". *Machine Learning* 113:3379—3398.

Reciprocal learning algorithms (see Contributions 1 through 7) self-select training data, leading to complex, i.e., not independently and identically distributed (*i.i.d.*) samples. While previously mentioned Contribution 8 studied the limits of learning from such data ("What can go wrong?"), this Contribution studies the possibilities of learning from such data ("How can we get things right?"). Instead of focusing on self-selected data from reciprocal learning specifically, we consider complex samples in general. Regarding the model class, we turn our attention to regression trees and forests. Here, we start with the simple observation that a 'naive estimation' that assumes *i.i.d.* may be substantially biased under complex sampling designs. Motivated by this observation, we then propose methods for de-biasing. The key idea is to leverage the relationship between the mean-squared-error criterion used in tree induction and population variance estimation to transfer results from survey statistics to tree induction. We extend our methodology to random forests in two ways. First, we simply plug in our de-biased tree induction into standard random forests giving rise to "Hájek-Forests". Second, we keep standard tree induction, but modify the bootstrap step by inverse probability weights. Moreover, we empirically demonstrate that accounting for complex sampling designs substantially improves predictive accuracy and interpretability. Interestingly, corrected forests can even outperform those trained on *i.i.d.* samples.

Contribution 10: Christoph Jansen, Georg Schollmeyer, Hannah Blocher, **Julian Rodemann**, and Thomas Augustin (2023). "Robust Statistical Comparison of Random Variables with Locally

Varying Scale of Measurement". In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by R.J. Evans, and I. Shpitser. PMLR, 941-–952.

This Contribution is the first of a series that shifts the attention from training data (as in Contributions 1 through 9) to test data—specifically, to benchmarking. It focuses on a very principled problem in statistics and machine learning: How to compare random variables with locally different scales of measurement? A prominent example (see Contribution 11) is multi-criteria/task[1] benchmarking with differently scaled criteria (such as partly ordinal, partly cardinal) or multidimensional poverty measurement, where poverty is measured by both income (cardinal) and education or health insurance status (ordinal). We tackle this issue by introducing an order based on (sets of) expectations of random variables that map into these spaces with locally varying scales of measurement. This order encompasses both stochastic dominance and expectation order as special cases, depending on whether there is no cardinal structure or a perfect one, respectively. We develop a (regularized) statistical test for our generalized stochastic dominance (GSD) order, implement it via linear optimization, and enhance its robustness using imprecise probability models. In Contribution 10, we do not yet apply our methodology to multidimensional benchmarking, but illustrate our statistical test by applying it to real world data sets from poverty measurement, finance, and medicine.

Contribution 11: Christoph Jansen*, Georg Schollmeyer*, **Julian Rodemann***, Hannah Blocher*, and Thomas Augustin (2024). "Statistical Multicriteria Benchmarking via the GSD-Front". **Spotlight Award.** In *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37, 98143-–98179.

The general question from Contribution 10 ("How to compare random variables with locally different scales of measurement?") can be easily translated to one of the most pervasive problems in the benchmarking literature: "How to compare multiple algorithms on multiple instances (typically datasets or prompts) with respect to multiple criteria of mixed scales, e.g. ordinal and cardinal?" This is a common issue, since modern algorithms are deployed in complex use cases, where one criterion is hardly enough to capture all aspects of algorithm performance. When faced with this problem in multi-criteria benchmarking, many practitioners retreat to the Pareto-front of algorithms. That is, only algorithms whose performance is undominated with respect to all criteria are considered. Building on the GSD-order as introduced in Contribution 10, we propose the GSD-front as an information-efficient improvement to the popular Pareto-front. We further propose a set-valued estimator for the GSD-front and provide sufficient conditions for its consistency. Moreover, we develop (static and dynamic) statistical hypothesis tests (permutation tests) to test whether an ML algorithm is in the GSD-front. We further quantify how robust the test decisions are under deviations from the underlying assumption of identically and independently distributed (*i.i.d.*) samples using techniques from robust statistics and imprecise probabilities. We apply our tests to the problem of classifier benchmarking using the PMLB benchmark suite and the OpenML platform. Moreover, we offer an efficient implementation in `R` that is freely available and easily adaptable to comparable problems.

[1] Throughout this dissertation, the terms multi-criteria and multi-task benchmarking are used synonymously, presupposing the different criteria reflect an inherently multidimensional concept, thus representing different *tasks*, rather than different metrics measuring the same latent construct, see Jansen et al. (2024) for background on this distinction.

Contribution 12: Julian Rodemann* and Hannah Blocher* (2024). "Partial Rankings of Optimizers". In *International Conference on Learning Representations (ICLR), Tiny Papers Track*. Ed. by T. F. Burns and K. Maughan. OpenReview.net.

We address the problem of Contribution 11 from a different angle: Instead of aiming for inferential statistics, we propose a *descriptive* method for analyzing multi-criteria benchmarking results, specifically for optimizers. Instead of the GSD-ordering, we simply aim to describe the distribution of partial orders/rankings arising from multiple criteria. To this end, we utilize the recently introduced union-free generic depth function for partial orders/rankings (Blocher et al., 2022). Our method describes the distribution of all partial orders/rankings, avoiding the notorious shortcomings of aggregation. This permits the identification of test functions that produce central or outlying rankings of optimizers and to assess the quality of benchmarking suites. We illustrate our framework on various use cases such as benchmark suites for deep learning optimizers, black box optimizers and evolutionary algorithms.

Contribution 13: Esteban Garcés Arias, Hannah Blocher, **Julian Rodemann**, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher (2025). "Towards Better Open-Ended Text Generation: A Multicriteria Evaluation Framework". In *ACL Workshop on Generation, Evaluation and Metrics (GEM)*. Ed. by O. Arviv, M. Clinciu, K. Dhole, R. Dror, S. Gehrmann, E. Habba, I. Itzhak, S. Mille, Y. Perlitz, E. Santus, J. Sedoc, M. Shmueli Scheuer, G. Stanovsky, O. Tafjord. Association for Computational Linguistics, 631—654.

Instead of the general multi-criteria benchmarking problem (Contributions 10 and 11), we turn to a specific problem in Contribution 13: text generation by language models. Specifically, we turn to multi-criteria comparisons of open-ended text generation. The latter is often challenging due to trade-offs among widely used metrics such as coherence, diversity, and perplexity. This Contribution employs benchmarking approaches based on partial orders (inspired by Contribution 12) and presents a new summary metric (Q*Text) to balance existing automatic indicators, providing a more holistic evaluation of text generation quality. Our experiments demonstrate that the proposed approaches offer a robust way to compare decoding strategies and serve as valuable tools to guide model selection for open-ended text generation tasks. We suggest future directions for improving evaluation methodologies in text generation and make our code, datasets and models publicly available. Moreover, we offer a concrete guideline for practitioners differentiating between two (stylized) use cases of multi-criteria benchmarking of open-ended text generation: a pragmatic and a methodological one. For the former, consider a practitioner with a specific use case (e.g., a customer support bot). Typically, the practitioner runs benchmarking with the simple goal of selecting *one* text generator (decoding model), which renders the metric information about the model performance a means to an end: one is primarily interested in a ranking. In this scenario, we recommend the ufg-depth for partial orders (Contribution 12). In the methodological scenario, one is interested in gaining structural insights into how hyperparameters affect the text generation quality. Here, it is of utmost importance to know *how much* better one method is. In this scenario, we recommend Q*Text to summarize multiple criteria to obtain an absolute (as opposed to relative) ranking on a cardinal scale.

Contribution 14: Esteban Garcés Arias, Hannah Blocher, **Julian Rodemann**, Matthias Aßenmacher, Christoph Jansen (2025). Statistical Multicriteria Evaluation of LLM-Generated Text. Available at *arXiv preprint arXiv:2506.18082* (last accessed October 4 2025). In *18th International Natural Language Generation Conference (INLG)* (forthcoming).

We apply the GSD-Methodology from Contributions 10 and 11 to the problem encountered in Contribution 13 of how to evaluate the quality of text generated by large language models (LLMs) simultaneously with respect to coherence, diversity, fluency and other automatic metrics as well as human evaluation. In this way, we address three major shortcomings of current benchmarking practices: the limitations of single-metric evaluation, the mismatch between cardinal automatic metrics and ordinal human judgments, as well as the absence of inferential statistical guarantees, not to mention ways of quantifying the robustness of inference under potential deviations from *i.i.d.* assumptions. We validate our approach using the Qwen 2.5 - 7B model (Yang et al., 2024) with a curated selection of prompts from Wikitext (Merity et al., 2016) and Wikinews (Wikinews contributors, 2025) data. Specifically, we conduct a comparative analysis of five common decoding strategies against human-written text. Text generated from these five decoding strategies was compared to human-written text by two human annotators and the automatic Q*Text score proposed in Contribution 13. This setup exemplifies a common problem in language model benchmarking: The simultaneity of human judgment (often ordinal) and automatic evaluation (often cardinal scores). The GSD-methodology introduced in Contributions 10 and 11 fits this setup like a glove.

Declaration of the Author's Specific Contributions

Fruitful and close collaboration with many co-authors from various disciplines resulted in the articles presented in this dissertation. The contributions of each author are clarified below. As the collaborations were very close, describing the contributions requires some level of detail. This should not obscure the fact that all aspects of the works were intensively discussed among and analyzed by all authors.

Contribution 1: Julian Rodemann, Christoph Jansen, and Georg Schollmeyer (2024). "Reciprocal Learning". In *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37, 1686–1724.

Julian Rodemann conceived the ideas of generalizing active learning, self-training, multi-armed bandits, superset learning and Bayesian optimization to *reciprocal learning* and to show convergence in this joint framework. Before writing the paper, the idea and its formal embedding into sequential decision theory was discussed at length with Christoph Jansen and Georg Schollmeyer. Julian Rodemann wrote the whole paper (first draft and all revisions) except for Appendix D, which was written by Christoph Jansen. All claims and proofs are due to Julian Rodemann, except for Lemma 1, which was jointly proven by Georg Schollmeyer and Julian Rodemann, and Theorem 3, which was jointly proven by Christoph Jansen and Julian Rodemann. Julian Rodemann implemented and ran the experiments presented in Appendix E. Christoph Jansen designed the illustrative figures 1 through 5. Julian Rodemann created figures 6 and 7. All authors contributed by discussing and proofreading the paper at various stages.

Contribution 2: Julian Rodemann and Thomas Augustin (2024). "Imprecise Bayesian Optimization". *Knowledge-Based Systems* 300:112186.

Using the CRediT roles for authorship as reference: Julian Rodemann: Writing – original draft, Writing – review and editing, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Thomas Augustin: Validation, Supervision, Project administration, Funding acquisition, Conceptualization. Their individual contributions are described in more detail below. As stated in the paper, a prior version of parts of this work had been presented at the Ninth International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM) in March 2022 and published under the title "Accounting for Gaussian Process Imprecision in Bayesian Optimization" in the conference proceedings as part of the Lecture Notes in Computer Science book series (LNAI, volume 13199),

see Rodemann and Augustin (2022a), which in turn is based on the Master's thesis "Robust Generalizations of Stochastic Derivative-Free Optimization" by Julian Rodemann. As such, this article is based on ideas by Julian Rodemann from back in 2021, continuously sharpened and improved through the close supervision by Thomas Augustin. Julian Rodemann drafted the paper and made all revisions after receiving feedback from Thomas Augustin. All claims and proofs are due to him. Julian Rodemann also implemented PROBO and ran all experiments. Thomas Augustin gave extensive feedback on all parts of the paper. Both authors contributed by discussing and proofreading the paper at various stages.

Contribution 3: Julian Rodemann, Federico Croppi, Philipp Arens, Yusuf Sale, Julia Herbinger, Bernd Bischl, Eyke Hüllermeier, Thomas Augustin, Conor J. Walsh, and Giuseppe Casalicchio (2025). "Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration for Exosuit Personalization". In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, Research Track and Applied Data Science Track. Springer, 525–542.

As stated in the paper, this work builds and heavily relies on the master's thesis "Explaining Sequential Model-Based Optimization" by Federico Croppi, supervised by Giuseppe Casalicchio (Croppi, 2021). Federico Croppi ideated the method of `ShapleyBO`, implemented it and tested it on synthetic functions. Julian Rodemann ideated and implemented the extensions to risk-averse confidence bounds and uncertainty-aware confidence bounds. Julian Rodemann also implemented the simulation study on exosuit personalization, based on user data provided by Philipp Arens and Conor J. Walsh. Julian Rodemann wrote the paper, except for the subsection 6.1 on "Personalizing Soft Exosuits", which was written by Philipp Arens, and the paragraph on Shapley values in Section 2, which was written by Julia Herbinger. Conceptually, Sections 4 and 5 are condensed versions of Federico Croppi's master's thesis. Giuseppe Casalicchio supported the project through continuous and close supervision, from the master's thesis to the final camera-ready version. Julia Herbinger assisted in supervision and Yusuf Sale helped with the conceptual embedding into the literature on uncertainty quantification. Eyke Hüllermeier, Bernd Bischl and Thomas Augustin provided very valuable feedback at various stages of the project. All authors contributed by discussing and proofreading the paper.

Contribution 4: Julian Rodemann, Jann Goschenhofer, Emilio Dorigatti, Thomas Nagler, and Thomas Augustin (2023). "Approximately Bayes-Optimal Pseudo Label Selection". In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by R. J. Evans, and I. Shpitser. PMLR, 1762-1773.

Julian Rodemann ideated the method and wrote the whole paper (draft and revisions in light of co-authors' comments) except for the Section on the approximations. Jann Goschenhofer and Emilio Dorigatti contributed by brainstorming and discussions at an early stage of the project, while Thomas Augustin improved the paper by commenting on drafts of the paper. Thomas Nagler developed the approximations and wrote the respective Section of the paper. He also revised the paper and the code substantially and suggested the final structure of the article. Julian Rodemann designed the experiments and implemented the method. He also ran all experiments, except for those using Bayesian neural networks, which were run by Emilio

Dorigatti assisted by Jann Goschenhofer and Julian Rodemann. All authors interpreted the results and contributed by proofreading, discussing and commenting on the paper.

Contribution 5: Julian Rodemann, Christoph Jansen, Georg Schollmeyer, and Thomas Augustin (2023). "In All Likelihoods: Robust Selection of Pseudo-Labeled Data". In *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by E. Miranda, I. Montes, E. Quaeghebeur, and B. Vantaggi. Vol. 215. PMLR, 412—425.

Julian Rodemann developed the main idea of robust PLS extensions that account for model selection, accumulation of error and covariate shift. He drafted and wrote the majority of the paper. Julian Rodemann further implemented robust PLS. He also conceived and conducted the experimental analyses. Christoph Jansen contributed the idea of deploying α-cut ("soft revision") updating rules for robust PLS. Its regret-based adaption was developed by Julian Rodemann. Christoph Jansen contributed several passages on solving robust PLS problems with respect to generalized stochastic dominance. Georg Schollmeyer and Christoph Jansen also aided with making technical notations more concise. Thomas Augustin, Georg Schollmeyer, Christoph Jansen and Julian Rodemann further contributed by stimulating discussions and detailed proof-reading.

Contribution 6: Stefan Dietrich, **Julian Rodemann**, and Christoph Jansen (2024). "Semi-Supervised Learning Guided by the Generalized Bayes Rule Under Soft Revision". In *International Conference on Soft Methods in Probability and Statistics (SMPS)*. Ed. by J. Ansari, S. Fuchs, W. Trutschnig, M.A. Lubiano, M.A. Gil, P. Grzegorzewski, O. Hryniewicz. Springer, 110—117.

Ideated by Julian Rodemann and Christoph Jansen, Stefan Dietrich implemented the generalized Bayes rule under soft revision (also referred to as "α-cuts") as a decision criterion for pseudo label selection in semi-supervised learning. In particular, Stefan Dietrich ideated the usage of the COBYLA (Constrained Optimization by Linear Approximations) algorithm for making the method computationally feasible and wrote the code for implementing and testing the method, based on Julian Rodemann's software as part of Contribution 5 (see above). Furthermore, Stefan Dietrich wrote the first draft of the paper. Julian Rodemann and Christoph Jansen revised and finalized the paper. All authors contributed to discussing and proofreading the article.

Contribution 7: Julian Rodemann, Dominik Kreiß, Eyke Hüllermeier, and Thomas Augustin (2022). "Levelwise Data Disambiguation by Cautious Superset Learning". In *International Conference on Scalable Uncertainty Management (SUM)*. Ed. by F. Dupin de Saint-Cyr, M. Öztürk-Escoffier, N. Potyka. Springer, 263-276.

In most parts, the paper was drafted and written by Julian Rodemann, who also contributed the idea of twisting-the-tuning and implemented the main approach. Dominik Kreiß contributed the idea of a step-wise narrowing down procedure as well as the application to the undecided voters and wrote the application Section. The idea of how to formally narrow down supersets in Section 3 was developed by Julian Rodemann and Dominik Kreiß together. The paper was made

possible and improved by the comments of Eyke Hüllermeier and Thomas Augustin. All authors contributed by proofreading the paper.

Contribution 8: Julian Rodemann and James Bailie (2025). "Generalization Bounds and Stopping Rules for Learning with Self-Selected Data". *arXiv preprint* arXiv:2505.07367 (last accessed October 12 2025.) *Under review at the Journal of Machine Learning Research (JMLR).*

The project was initialized and ideated by Julian Rodemann in discussions with James Bailie. The exact setup was first proposed by Julian Rodemann and then thoroughly revised and edited by James Bailie. Julian Rodemann wrote the first draft of the paper. James Bailie made several essential changes. For instance, he worked out that all the results are pathwise-valid. James Bailie further simplified the conditions and assumptions for the theorems. He also reformulated some of the theorems. All proof strategies were ideated together. Spelling out the proofs was executed by Julian Rodemann, except for Lemma 1, which is mainly due to James Bailie. All figures were created by Julian Rodemann. The experiments in Section 6 were also conducted by Julian Rodemann. James Bailie checked and simplified all proofs. Both authors contributed to discussing and proofreading the article.

Contribution 9: Malte Nalenz, **Julian Rodemann**, and Thomas Augustin (2024). "Learning De-Biased Regression Trees and Forests from Complex Samples". *Machine Learning* 113:3379—3398.

The idea was developed jointly by Malte Nalenz and Julian Rodemann. Malte Nalenz drafted and revised the whole paper except for Section 4.3 and Section 7, which were written by Julian Rodemann. Malte Nalenz and Julian Rodemann implemented the methods in R. Malte Nalenz ran all experiments. He also created all figures in the article. Julian Rodemann and Thomas Augustin revised and commented on the article. The final revisions were made by Malte Nalenz. The experimental results were interpreted jointly by all authors, who also contributed to discussing and proofreading the article.

Contribution 10: Christoph Jansen, Georg Schollmeyer, Hannah Blocher, **Julian Rodemann**, and Thomas Augustin (2023). "Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement". In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by R. J. Evans, and I. Shpitser. PMLR, 941—952.

The idea and most parts of the original draft of the paper come from Christoph Jansen. Georg Schollmeyer supplied parts of Proposition 7 and Proposition 8 including parts of the idea for their proof. He also discussed and drafted parts of the aspects related to regularization. Hannah Blocher wrote Section 8.1 and Appendix B. The R code for performing the permutation-based tests in all three applications as well as the GitHub repository are also due to Hannah Blocher. Julian Rodemann wrote parts of the introduction to Section 6, most parts of Section 8.2 as well as Appendix D. Furthermore, he analyzed and visualized the test results in the paper. The idea for Figure 3 in the main paper as well as Figures 2 and 4 in the appendix was jointly developed by Christoph Jansen and Julian Rodemann. Julian Rodemann proposed the applications on

finance and medicine. Thomas Augustin wrote some parts on related literature and parts of the introduction to Section 7. All authors contributed to revising the paper in several discussion rounds.

Contribution 11: Christoph Jansen*, Georg Schollmeyer*, **Julian Rodemann***, Hannah Blocher*, and Thomas Augustin (2024). "Statistical Multicriteria Benchmarking via the GSD-Front". **Spotlight Award.** In *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37, 98143-–98179.

The idea and most parts of the original draft of the paper come from Christoph Jansen. Section 5 was jointly written by Julian Rodemann, Christoph Jansen and Hannah Blocher (in this order). The sufficient condition of a finite VC-dimension and the proof of Theorem 1 are mostly due to Georg Schollmeyer. Corollary 1 was jointly delivered by Georg Schollmeyer and Christoph Jansen. The idea of the dynamic versions of the tests in Section 4 was jointly developed by Christoph Jansen and Georg Schollmeyer. The R code for performing the permutation-based tests in both applications as well as the GitHub repository are due to Hannah Blocher. The idea of using computing time as an ordinal performance measure is due to Hannah Blocher. The R code for enabling the benchmark analysis of the PMLB experiments, in particular the hyperparameter tuning and integration of compressed rule ensemble learning (CRE), is due to Julian Rodemann and Georg Schollmeyer. The idea of using feature- and class robustness as ordinal performance measures is due to Georg Schollmeyer. Figures 1, 3, 5, 6, 7 were created by Julian Rodemann, Figures 2 and 4 were created by Christoph Jansen. Thomas Augustin wrote Section 1.1 of the paper. All authors contributed to revising the paper.

Contribution 12: Julian Rodemann* and Hannah Blocher* (2024). "Partial Rankings of Optimizers". In *International Conference on Learning Representations (ICLR), Tiny Papers Track*. Ed. by T. F. Burns and K. Maughan. OpenReview.net.

Julian Rodemann and Hannah Blocher contributed equally to this work. Using the CRediT roles for authorship as reference: Supervision, writing, data curation, investigation, formal analysis, validation and review were contributed equally by both authors. Project administration was mainly done by Julian Rodemann. Both authors contributed to revising the article. Their individual contributions are described in more detail below. The idea of applying the union-free generic depth for partial order-valued data to partial orders describing the performance of optimizers evaluated on a benchmark suite comes from Julian Rodemann (Concept). All three benchmark suites considered were proposed by Julian Rodemann (Resources). Hannah Blocher introduced the ufg-depth for partial order-valued data in previous work (Blocher et al., 2022), where she wrote and ran the R code for this Contribution (Methodology, Software). Julian Rodemann wrote the first draft of the abstract and the introduction. The part on depth functions was added by Hannah Blocher. Together, the authors stated the contributions of the article at the end of the introduction. Hannah Blocher wrote the method section. In the results section, the embedding of the result into the theory of optimization and the characteristics of the benchmark suite DeepOBS was done by Julian Rodemann. Hannah Blocher analyzed the results and the general meaning of the depth values. Both authors wrote the outlook/result sections together.

Appendices A, B and C were written by Hannah Blocher. Appendix D was written by Julian Rodemann. The first draft and idea for Appendices E and F was written by Julian Rodemann. The final version of Appendices E and F underwent several changes made by both authors. Each of these parts is based on intensive discussion between the two authors, with both authors revising each part several times.

Contribution 13: Esteban Garcés Arias, Hannah Blocher, **Julian Rodemann**, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher (2025). "Towards Better Open-Ended Text Generation: A Multicriteria Evaluation Framework". In *ACL Workshop on Generation, Evaluation and Metrics (GEM)*. Ed. by O. Arviv, M. Clinciu, K. Dhole, R. Dror, S. Gehrmann, E. Habba, I. Itzhak, S. Mille, Y. Perlitz, E. Santus, J. Sedoc, M. Shmueli Scheuer, G. Stanovsky, O. Tafjord. Association for Computational Linguistics, 631—654.

The project was initialized and ideated by Esteban Garcés Arias in discussions with Julian Rodemann. Esteban Garcés Arias wrote the whole paper, except for the second part of Section 1 and the first part of Section 2 as well as Section 8, which were written by Julian Rodemann, and Section 4 and Appendix B through E, which were written by Hannah Blocher. The implementation and experiments are due to Hannah Blocher, Meimingwei Li and Esteban Garcés Arias. Matthias Aßenmacher and Christian Heumann gave extensive feedback at various stages of the project. Matthias Aßenmacher further substantially revised the introduction. The final structure is also due to him. All authors contributed to revising and discussing the paper.

Contribution 14: Esteban Garcés Arias, Hannah Blocher, **Julian Rodemann**, Matthias Aßenmacher, Christoph Jansen (2025). Statistical Multicriteria Evaluation of LLM-Generated Text. *arXiv preprint arXiv:2506.18082* (last accessed October 15 2025). In *18th International Natural Language Generation Conference (INLG)* (forthcoming).

The project was initialized and ideated by Esteban Garcés Arias and Julian Rodemann. Esteban Garcés Arias, Hannah Blocher and Matthias Aßenmacher wrote the introductory parts, while Christoph Jansen and Hannah Blocher contributed the methodological background and embedding. The experimental sections are due to Hannah Blocher, Esteban Garcés Arias and Christoph Jansen. Julian Rodemann, Matthias Aßenmacher and Esteban Garcés Arias surveyed the related work and wrote the respective parts of the paper. The summarizing and conceptual parts of the paper were written by Esteban Garcés Arias. Implementation and experimental analysis were conducted by Hannah Blocher, Esteban Garcés Arias and Christoph Jansen. Matthias Aßenmacher and Christoph Jansen gave extensive feedback at various stages of the project and substantially revised parts of the paper. All authors contributed to revising and discussing the paper.

Further Publications by the Author (Not Included in the Dissertation)

1. **Julian Rodemann**, Esteban Garcés Arias, Christoph Luther, Christoph Jansen, Thomas Augustin (2025). A Statistical Case Against Empirical Human-AI Alignment. *arXiv preprint arXiv:2502.14581* (last accessed October 4 2025). (Rodemann et al., 2025a)

2. Esteban Garcés Arias, **Julian Rodemann**, Christian Heumann (2025). The Geometry of Creative Variability: How Credal Sets Expose Calibration Gaps in Language Models. *Second Workshop on Uncertainty-Aware NLP at EMNLP 2025* (forthcoming). Association for Computational Linguistics (ACL). (Garcés Arias et al., 2025c)

3. Yuanhao Ding*, Esteban Garcés Arias*, Meimingwei Li*, **Julian Rodemann**, Matthias Aßenmacher, Danlu Chen, Gaojuan Fan, Christian Heumann, Chongsheng Zhang (2025). GUARD: Glocal Uncertainty-Aware Robust Decoding for Efficient Self-Adaptive Text Generation. *Findings of the Association for Computational Linguistics: EMNLP 2025* (forthcoming). Association for Computational Linguistics (ACL). (Ding et al., 2025)

4. Fabian Bongratz, Vladimir Golkov, Lukas Mautner, Luca Della Libera, Frederik Heetmeyer, Felix Czaja, **Julian Rodemann**, Daniel Cremers (2024). How to Choose a Reinforcement-Learning Algorithm. *arXiv preprint arXiv:2407.20917* (last accessed October 4 2025). (Bongratz et al., 2024)

5. **Julian Rodemann** (2024). Towards Bayesian Data Selection. *ICML Workshop on Data-Centric Machine Learning Research (DMLR)*. Ed. by A. Mahdi, L. Schmidt, A. Dimakis, R. Dror, G. Gkioxari, S. Truong, L. Bat-Leah, F. Alzamzami, G. Smyrnis, T. Nguyen, N. Gürel, P. Climaco, L. Oala, H. Schoelkopf, A. Bean, B. Isik, V. Shankar, M. Chen, A. Dave. Available as arXiv:2406.12560 (last accessed October 16 2025). (Rodemann, 2024)

6. Esteban Garcés Arias, **Julian Rodemann**, Meimingwei Li, Christian Heumann, Matthias Aßenmacher (2024). Adaptive Contrastive Search: Uncertainty-Guided Decoding for Open-Ended Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Ed. by Y. Al-Onaizan, M. Bansal, Y.-N. Chen. Association for Computational Linguistics, 15060—15080. (Garcés Arias et al., 2024)

7. **Julian Rodemann** (2023). Pseudo-Label Selection Is a Decision Problem. In *Advances in Artificial Intelligence. 46th German Conference on Artificial Intelligence (KI)*. Ed. by D. Seipel, and A. Steen, Springer, 261–264. (Rodemann, 2023b)

8. **Julian Rodemann**, Thomas Augustin (2022). Accounting for Gaussian Process Imprecision in Bayesian Optimization. *9th International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM)*. Ed. by K. Honda, T. Entani, S. Ubukata, V.-N. Huynh, M. Inuiguchi, Springer, 92–104. (Rodemann and Augustin, 2022a)

*These authors contributed equally to this work.

Eidesstattliche Versicherung (Affidavit)

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt worden ist.

Während des Überarbeitungsprozesses habe ich ChatGPT (GPT-5) und Claude (Sonnet 4.5) als Hilfsmittel verwendet, um bestimmte Sätze Korrektur lesen zu lassen und mir Vorschläge zur Verbesserung der Lesbarkeit einzelner Sätze generieren zu lassen. Diese Vorschläge wurden nicht wortwörtlich übernommen, sondern von mir überprüft, angepasst und umformuliert. Außerdem habe ich GitHub Copilot und ChatGPT (GPT-5) verwendet, um bestimmte Teile der SPECTER-basierten Literaturauswertung in 3.3 zu programmieren. Ferner wurde der TikZ-Code für die Grafiken 1.1 und 1.2 mithilfe von ChatGPT (GPT-5) vereinfacht und überarbeitet.[1]

München, den 31.10.2025 Julian Martin Rodemann

[1] **(Informal) English Translation:**

I hereby declare, in lieu of an oath, that the dissertation was prepared by me independently and without unauthorized assistance.

During the revision process, I used ChatGPT (GPT-5) and Claude (Sonnet 4.5) as aids to have certain sentences proofread and to obtain suggestions for improving the readability of individual sentences. These suggestions were not adopted verbatim, but were reviewed, adjusted, and rephrased by me. I also used GitHub Copilot and ChatGPT (GPT-5) to program certain parts of the SPECTER-based literature analysis in Section 3.3. Furthermore, the TikZ code for Figures 1.1 and 1.2 was simplified and revised with the assistance of ChatGPT (GPT-5).

Part I.

Prologue

1. Data: The Dark Side of the Moon?

> "The activity of the sensible object and that of the percipient sense is one and the same activity [...] the actuality of the sensible object and that of the sensitive subject are both realized in the latter."
>
> Aristotle (approx. 350 BC): *De Anima (On the soul)*, Book III, Chapter 2.
>
> — (Aristotle, 1907, 350 BC)

The Latin word *data* means *(things) given* (Harper, 2001). And we indeed often take data as a given; that is, as something we passively perceive (Heuer, 1999; Jones et al., 2019; Williamson, 2024). Practitioners know this is an illusion: Data is rarely just passively perceived, it is actively selected, subsetted and pre-processed before being analyzed.

In the epistemological words of Leonelli (2019), "there is no such thing as 'raw data,' since data are forged and processed through instruments, formats, algorithms and settings that embody specific theoretical perspectives on the world" (Leonelli, 2019, Page 2). As Popper (1959, 1962) famously argues, every observation is theory-laden, i.e., guided by hypotheses that should be open to falsification. Going further, Horkheimer and Adorno (2002) describe the illusion of "objective observations" as the positivist "myth of that which is the case" (Horkheimer and Adorno, 2002, Page vii), potentially being (mis)used for the legitimization of power.[1] For a deeper dive into the epistemology of data, the interested reader is referred to the programmatic collection of essays titled "Raw Data is an Oxymoron" (Gitelman, 2013).

To put it bluntly, this practical perspective on data as a medium of active engagement rather than passive perception appears to never have percolated up (or trickled down?) in its entirety from practice to mainstream theory.[2] Despite practitioners knowing better, one is inclined to say, many researchers in statistics and machine learning (including myself) still often cast data as an untouched sample of an appropriately defined population. However, construing data as a medium of active engagement does connect to two old branches of statistical science: experimental design and sampling theory, which will be reviewed extensively in Section 3.1 and 3.2 of this dissertation, respectively.

The modern "modeling cultures" (Breiman, 2001b, Page 205) in *both* statistics and machine learning (see below), however, rather focus on the model than the data. Instead of turning the full attention to the complexities of where data comes from and goes to, see Figure 1.1, much effort has gone into building ever better and bigger models. Data, it appears, constitutes the dark side of the moon.

[1] Or at least for the legitimization of *rule* in the sense of Weber (1904, 1922).

[2] A very notable exception is the recent line of work by Ramdas et al. (2023); Grünwald et al. (2024); Vovk and Wang (2021); Ramdas and Wang (2025); De Heide (2024), as detailed below.

1.1. A Process, Not a Thing

In his thought-provoking analysis of the "rhetorics of machine learning", Williamson (2024) identifies the following all too familiar phrase as the rhetorical premise of many research papers in statistics and machine learning:

"Data is drawn independently from some [identical, J.R.] probability distribution."

Williamson's remark is somewhat reminiscent of Leo Breiman's amusement when reading the *Annals of Statistics* and the *Journal of the American Statistical Association (JASA)* upon his return to the University of California, Berkeley, after several years outside of academia. In his seminal paper on "the two cultures" (Breiman, 2001b) contrasting machine learning with statistics, he describes his bewilderment when noticing that in both the *Annals* and in *JASA* "virtually every article contains a statement of the form: 'Assume that the data are generated by the following model: ...'" (Breiman, 2001b, Page 202). Breiman goes on to compare this "data model culture" (Breiman, 2001b, Page 199) with the "algorithmic modeling culture" (ibid.) in machine learning. The latter assumes "that nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable" (Breiman, 2001b, Page 205). However, as Breiman (2001b) acknowledges, the one assumption that is *not* dropped "is that the data is drawn *i.i.d.* from an unknown multivariate distribution" (Breiman, 2001b, Page 205). This assumption, it appears, is required in *both* statistics and machine learning—and this is exactly where the dissertation at hand enters the picture.

The (in)famous *i.i.d.* (independent and identically distributed) assumption is problematic in at least two ways. First, albeit its fundamental role, there seems to be no formal consensus on its exact meaning. According to Williamson (2024), there is no complete mathematical description of what it actually means to draw a sample from a probability distribution "in any statistics text" (Williamson, 2024, Page 4). Second, the strength of this assumption often appears to be severely underestimated. The bulk of survey statistics literature indicates that there is hardly any *i.i.d.* sample in the real world; see Kreuter and Valliant (2007); Groves et al. (2011) for an overview. As with any stylized assumption, the scholar does not need to trust the *i.i.d.* assumption to hold *entirely*. Yet, nature should not deviate too strongly for the scientific conclusions the scholar draws still to be reliable—see, e.g., Contributions 10 and 11 of this dissertation for an exact quantification of this deviation in the context of statistical hypotheses tests. However, this dissertation in general and Contribution 8 in particular will show that deviations[3] of the actually used (self-selected) sample from a hypothetical *i.i.d.* one are quite substantial throughout some popular methodology.

Due to these severe limitations, Williamson (2020, 2024) argues that we should understand data as a *process* rather than a *(given) thing*. Yu (2020); Yu and Barter (2024) explicitly consider a *data lifecycle*, consisting of data collection, cleaning, exploration and visualization, modeling, post hoc analysis, interpretation, which eventually leads to updating domain knowledge, triggering new research questions that precede data collection, completing the circle pictured in Figure 1.1. Their key argument makes sense to any practitioner: Data analysis, machine learning and statistical inference consist of more than just finding and fitting an appropriate model to some data that miraculously fell into the analyst's lap. Notably, construing data as a process goes beyond scientifically modeling data *production* or *acquisition* as is common in social and official

[3] Measured by, e.g., the Wasserstein distance between probability distributions, see Contributions 1 and 8 (Rodemann et al., 2024; Rodemann and Bailie, 2025).

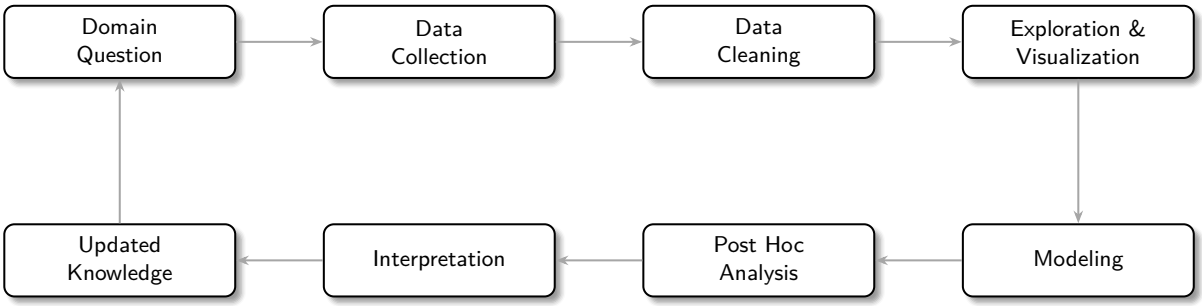

Figure 1.1.: The data lifecycle according to Yu (2020); Yu and Barter (2024). Figure is replicated from (Yu, 2020, Figure 1).

statistics (see e.g., Krug et al. (2001)). Instead, Williamson (2024) argues that data *itself* should be understood as a process.

Treating data as a given might also be responsible for statistics' long neglect of causality. Pearl and Mackenzie (2018) state that modern statistics "hatched out of the causal questions that Galton and Pearson asked [...]. Unfortunately they failed in this endeavor, and rather than pause to ask why, they declared those questions off limits, and turned to develop a thriving, causality-free enterprise called statistics," (Pearl and Mackenzie, 2018, Page 5); see also Pearl (2009, 2010). In his inimitably provocative style, Pearl contends that modern statistics began with causal questions but, after failing to answer them from observational data, he proclaims, the field mostly turned itself into a causality-free calculus of association (Pearl, 2009; Pearl and Mackenzie, 2018). This view is controversial, as the statistical experimental design (as detailed in Section 3.1) can be viewed as part of causal inference. Besides, the causal potential outcomes framework was basically already proposed by the eminent Statistician Jerzey Neyman in his Master's thesis (Neyman, 1923).

Nevertheless, construing data as a dynamic process rather than a static and given thing clearly helps to focus on the causal mechanisms generating the data in the first place, e.g., by modeling it via directed acyclic graphs (DAGs) (Pearl, 1995) or the potential outcomes framework (Rubin, 1974, 1978). In fact, "experimental data"—often considered the gold standard in causal inference—requires explicit design of how data is collected based on a well-defined domain question; see Figure 1.1 and Section 3.1. Starting the statistical analysis with given "observational data" hides these designs and thus the causal mechanisms involved.

1.2. Explanans and Explanandum: Two Sides of the Same Coin?

Summing up so far, one might reasonably conclude that objective "raw data" is a myth crafted by convenience: It avoids dealing with all the intricacies of data collection, subselection, cleaning, pre-processing, curation, sampling and the like.

However, there are some deeper philosophical and historical underpinnings. In particular, I argue that treating data as a static and accurate depiction of nature is an implicit manifestation of the idealized separation between observant and observer in the classical sense of differentiating *explanans* (the explaining thing) from *explandandum* (the explained thing) (Hempel and Oppenheim, 1948). It is self-evident that violating this dichotomy is a potential reason for violations of the *i.i.d.* assumption as pondered upon by Williamson (2024). No longer perceiving *explanans*

(statistical or ML model) and *explanandum* (population) as completely separated entities puts the *i.i.d.* assumption under scrutiny. Repeatedly drawn samples from a probability distribution can hardly be independent, if these samples (or conclusions being drawn from them) affect this distribution or, more severely, are intertwined or even coincide with it.

This static dichotomy between scientific subject and object has been called into question already around 350 BC, when none other than Aristotle recognized that "the actuality of the sensible object and that of the sensitive subject are both realized in the latter" (Aristotle, 1907, Book III, Chapter 2) (see quote at the start of this Chapter). In plain terms, Aristotle implictely claimed there is no object without a subject, because the former requires the latter to exist. Or, to put it in a slightly more nuanced way, sensory perception of an object presupposes the existence of a subject endowed with the corresponding faculties of sense. For our purposes, it suffices to conclude that subject and object are mutually *inter*dependent.

Taking a small tangent here, rejecting a clear dichotomy between the passively observing scholar and the active nature being observed also tends to align with Paul Feyerabend's anarchistic theory of knowledge (Feyerabend, 1975). The famous philosopher of science especially questions the scholar being passive and neutral, and instead describes science as a social and partly manipulative enterprise. In a detailed case study on the shift from the Ptolemaic (geocentric) to the heliocentric paradigm, Feyerabend shows how Galileo Galilei lacked empirical facts to support his theory due to indeterminate and double images produced by his telescope in Padua. Feyerabend goes on to show how Galileo instead used rhetoric, politics and even manipulation in advancing his new ideas (Feyerabend, 1975, Chapters 6,9, and 10). For instance, Feyerabend shows how Galileo presented his ideas as a debate between three characters in *Dialogo sopra i due massimi sistemi del mondo* (Dialogue Concerning the Two Chief World Systems, Galilei (1967)). This piece by Galileo, as Feyerabend argues, is much closer to a strategic rhetorical tool than a neutral presentation of facts. Feyerabend's famous case study reveals that Galileo's theory (scientific *explanans*) had an undeniable effect on how he perceived the object of his scholarly investigations (scientific *explanandum*). More generally, Feyerabend was convinced that science is not a detached, hallowed enterprise that takes an objective, "view from nowhere" (Nagel, 1986) perspective on the universe but instead is a social enterprise, deeply interwoven in society with all its imperfections.[4] This is in contrast to Popper (1962), who asserted that science can still be objective—despite his recognition of empirical observations as theory-laden (see above). Instead of from a passive recording of facts, he argued, objectivity can arise from the openness of scientific claims to intersubjective criticism (cf. ibid.).

1.3. The Model-Data Dichotomy in Statistics and Machine Learning

In the dissertation at hand, I will leave these debates to the philosophers and focus on the implications of giving up the static separation between subject (i.e., a model generated by the scholar) and object (i.e., data generated by nature) for statistics and machine learning. Upon closer examination, there are two (related) aspects at play: The *static* nature of the setup and the strict *separation*. The static setup was early questioned by, most prominently, Wald (1945a) when proposing his sequential "probability ratio" test; see also Wald (1947b). Today,

[4] As an aside, reading Feyerabend's work can be rather discouraging at the outset of one's doctoral studies, given how universal his criticism of the scientific endeavour is. Yet, I would recommend "Against method: Outline of an anarchistic theory of knowledge" (Feyerabend, 1975) to anyone interested in understanding the deep limitations of science.

sophisticated approaches like alpha spending (DeMets and Lan, 1994) allows for interim analyses, i.e., gathering more data after inference while controlling error rates. More generally, online learning (Littlestone, 1988; Cesa-Bianchi and Lugosi, 2006), time series analysis (Box et al., 2015; Hamilton, 1994), and, even more generally, stochastic processes (Ross, 1995; Karatzas and Shreve, 1991) allow for drafting a dynamic rather than a static modelling setup. However, as De Heide (2024) notes, methods like alpha spending still require the scientific model to be set beforehand, including the research question, statistical model and the sample size. In her "Plea for a New Statistical Paradigm", De Heide (2024) goes on to point out a ubiquitous Achilles's heel of statistics:

> "However, at a fundamental level, the current state-of-the-art in statistical methodology still reflects the use-context of [Sir Ronald A, J.R.] Fisher's day, where researchers would engage in well-controlled, well-planned, single experiments with a prespecified research question and a set end date."
>
> Rianne de Heide (2024): *A plea for a new statistical paradigm*,
>
> — (De Heide, 2024, Page 183)

The strand of literature around game-theoretic statistics (Ramdas et al., 2023) and safe testing (Grünwald et al., 2024) with e-variables (Vovk and Wang, 2021; Ramdas and Wang, 2025) offers principled remedies that allow for *anytime-valid* inferential statements, even under optional continuation of data acquisition. Notably, in Contribution 8 of this dissertation, we prove generalization bounds that are anytime-valid in this very sense of Ramdas et al. (2023). Loosely speaking, conclusions from safe testing remain valid even if the *explanans* interacts with the (sampled) *explanandum*.[5]

This points to the arguably most problematic nuance of the strict model-data dichotomy: By not allowing for any mutuality between subject and object, neither *explanandum* nor *explanans* can simultaneously influence and be influenced. This renders the relationship between scientific conclusion on the one hand and statistical or ML models on the other hand a one-way-street: Nature generates data, which then leads to scientific statements.

1.4. If Data is a (Cyclic) Process, Beware the Feedback Loops

While the above mentioned techniques of sequential testing and alpha spending offer some remedies by considering dynamic rather than static setups, they do not focus on formal instruments to explicitly account for direct (feedback) loops between model and data or model and population. The dissertation at hand contributes to developing such instruments—particularly for automated, algorithmic feedback loops, as detailed below.

There are strong arguments to do so. In fact, this deceiving idealization of data as a static representation of reality has been called into question by both the natural and social sciences. Quite prominently, *performative prediction* (Perdomo et al., 2020) construes machine learning as a reflexive problem when applied in a social context. Here, the population changes in response to a model's predictions. These types of feedback can be self-fulfilling or self-defeating. Examples

[5] While e-variables and game-theoretic statistics offer principled solutions to statistical inference, this dissertation will focus on (algorithmic) feedback loops within machine learning, where we often have direct control and exact knowledge about these loops as detailed below.

comprise traffic route predictions making drivers avoid congestions (self-defeating) or predictive policing triggering more police presence in neighborhoods with high crime predictions, which leads to more crimes being registered there (self-fulfilling). Morgenstern (1928) was the first to recognize performativity as a fundamental problem in the social sciences. Later work by Grunberg and Modigliani (1954) and Simon (1954) popularized performativity and delivered sufficient conditions on whether predictions can be accurate under performativity (Grunberg-Modigliani-Simon-Theorem); see Hardt and Mendler-Dünner (2025) for details. Rediscovered as a pervasive problem in the deployment of machine learning models in the social world, e.g., on social media, extensive research has emerged on whether machine learning can be stable under these kinds of feedback loops. (Mendler-Dünner et al., 2020; Miller et al., 2021; Brown et al., 2022)

In recent years, such feedback loops have become pervasive even outside of social contexts, where *human* entities of a population react to predictions being made about them. Large-scale generative AI models quite explicitly alter the population, from which their training data is drawn. The procedure is simple: These *non-human* entities generate tons of data like images or text that becomes freely available on the internet, which is where those same models typically scrape their training data from. This feedback loop can be understood as performative, as the models in some sense change the ground truth (population) by adding statistical units to it.[6] To the best of my knowledge, Hataya et al. (2023) were first to explicitly ask: "Will large-scale generative models corrupt future datasets?" (Hataya et al., 2023, Page 1). Martínez et al. (2023) partially answer this question by studying the interplay of generative computer vision models and the internet. Providing a more theoretical (yet not genuinely statistical) perspective, Bertrand et al. (2024) and Fu et al. (2025) study whether these self-consuming loops can be stable and how to prevent model collapse (see also Van Breugel et al. (2023); Alemohammad et al. (2024) for similar works).

Recent interest in agentic AI (see e.g., Belcak et al. (2025)) further highlights the importance and prevalence of such feedback loops. Agentic AI refers to systems (involving at least one (language) model) that pursue goals with a higher degree of autonomy (in particular, lesser degree of human oversight) than classical machine learning applications. Agentic AI systems plan, decompose tasks into subgoals, act and observe in an environment, and adapt using feedback, rather than merely reacting to instructions; see Acharya et al. (2025); Sapkota et al. (2025) for definitions and taxonomies as well as Wooldridge and Jennings (1995); Sutton and Barto (2018); Russell and Norvig (2021) for pathbreaking early works. Recent prototypes illustrate these loops with iterative self-improvement in open-ended settings (Park et al., 2023). Increased autonomy and lack of oversight have the potential to expedite feedback loops. Seeking reliability, quantifying these system's uncertainty over acts and predictions thus requires giving up the static data-model dichotomy to an even greater extent.

As a whole, the cumulative dissertation at hand is motivated from an angle akin to the performative one sketched above. The central argument, however, originates in the *natural* instead of the *social* sciences—in physics, specifically. In its core, the argument dates back to Dicke (1961, 1957) and goes as follows: All feasible observations of the universe are constrained by the simple fact that they can only be made in a universe capable of developing intelligent life. This is often referred to as the "anthropic principle" or "observation selection effect" (Carter, 1983; Bostrom, 2000, 2002); see also Rodemann et al. (2025a, Section 3.2). In other words, the very act of

[6] It can, however, also be cast as an instance of *reciprocal learning*, see below and Part III, where models change samples only with static ground truth (population).

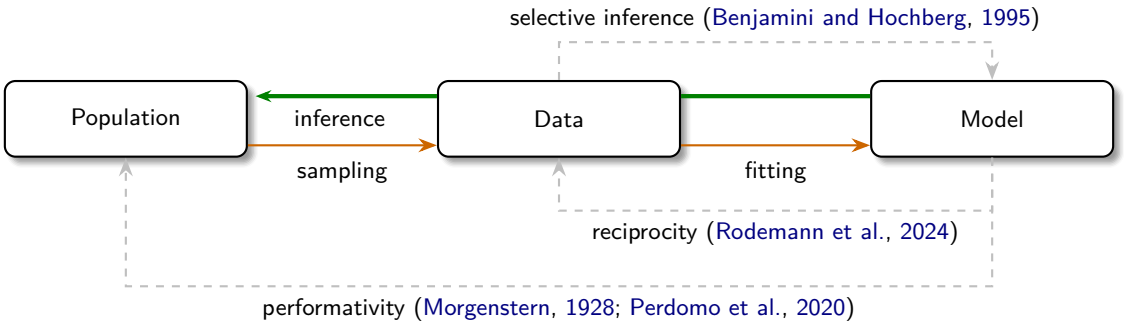

Figure 1.2.: A simplified view of statistical inference, highlighting three kinds of (feedback) loops. The model allows the scholar to learn something about the population *through* the data. Dotted arrows represent feedback loops in this process, via selective inference (Benjamini and Hochberg, 1995; Benjamini, 2020), reciprocal learning (Rodemann et al., 2024; Rodemann and Bailie, 2025) and performative prediction (Perdomo et al., 2020; Perdomo, 2023), respectively. As the *observation selection effect* (Dicke, 1957, 1961) and the anarchistic theory of knowledge (Feyerabend, 1975) show, these information flows do not need to follow, but can also pre-date the main flows.

observing is constrained by the conditions that allow the observer to exist, reemphasizing the depth of Aristotle's insight in 350 BC, as detailed above.

Statistically speaking, performative prediction describes feedback loops between a model and the *population* it aims at explaining. Through the same lens, the observation selection effect is concerned with feedback loops between a model and the *sample* it is based on.[7] As illustrated by Figure 1.2, both effects *feed back* from data to model and vice versa after inference (selective inference) or after model fitting (reciprocal learning), creating a potential *loop*, hence the term feedback loop, originating in cybernetics and control theory (Wiener, 1948).[8]

The present dissertation—in particular, Contributions 1–9—will focus on this second type of feedback loops: the learning algorithm or model (scientific subject) altering the data by (sub)selection. Notably, Dwork et al. (2015) have addressed such feedback loops in the non-algorithmic context of adaptive data analysis, where new analyses rely on both data exploration and the results of previous analyses of the same data. Different to Dwork et al. (2015), the emphasis in this dissertation lies on algorithmic (instead of scholarly) data selection in machine learning. This is in line with growing interest in "data-centric machine learning", which emphasizes automated data selection and curation (Oala et al., 2024). Section 3.3 will review this literature.

Inspired by this statistical legacy of experimental design (*intentional* data selection) and sampling theory (*inadvertent* data selection), this dissertation particularly turns to a third type of data selection that is *intentional* at first sight but, as it turns out, can have *inadvertent* consequences: automated data selection by learning algorithms themselves. Contribution 1 will demonstrate that automated data selection is ubiquitous across several popular learning algorithms like active learning, self-training in semi-supervised learning, superset learning, bandits and Bayesian optimization.

[7] Obviously, feedback loops on the population also trickle down to a sample being drawn from that population.

[8] Norbert Wiener (Wiener, 1948) founded cybernetics eight years prior to the landmark Dartmouth conference on AI, in order to study "circular [...] feedback mechanisms" (Von Foerster, 1952). As detailed in Rodemann and Bailie (2025, Appendix B), cybernetics explored the *dynamic* foundations of system stability, adaptability, and feedback-based control theory, while AI research concentrated on *static* pattern recognition.

Not only have the social and natural sciences, but also the statistics community itself, noticed that construing data and model as separate, independent entities can have serious pitfalls. In statistical practice, the same data is often used multiple times for testing similar or the same hypotheses. Practitioners also derive hypotheses from exploratory data analysis, again using the data more than once. Already in the 1950s, no other than John W. Tukey identified problems when making multiple inferences from a single set of data (Tukey, 1953, 1991). However, it was not before the seminal[9] work by Benjamini and Hochberg (1995) on the false discovery rate that the problem received full attention (Benjamini and Braun, 2002). The problem is now known as "selective inference" and, in Benjamini's words, refers to deteriorating inferential guarantees that occur when "focusing statistical inference on some findings that turned out to be of interest only after viewing the data" (Benjamini, 2020, Page 7). Figure 1.2 depicts selective inference as another type of feedback loop in the classic data-model dichotomy. While reciprocal learning algorithms allow the model to change the data, selective inference is concerned with data changing the model, i.e., affecting the model choice.

Summary There are strong arguments to question the model-data dichotomy in statistics and machine learning. First and foremost, we have learned that this dichotomy is based on the idealized epistemological separation between subjective *explanans* and objective *explanandum*. The latter was already questioned by Aristotle and, roughly leaning on Karl Popper's account of the empirical sciences, "does not rest upon solid bedrock" (Popper, 1959, Pages 93-94). Second, if deployed in social contexts, public predictions made by statistical or machine learning models (*explanans*) can shape the population (*explanandum*), turning the passively describing model into one that actively shapes the *explanandum*, which is no longer just an object of scholarly investigation, but a subject reacting to it. Third, such performative effects also exist on the sample level with observations (*explanandum*) presupposing the observer's existence (*explanans*). Crucially, this very feedback loop on the sample level is not a mere abstraction, but can be observed in a wide range of machine learning algorithms that we unify under the umbrella of *reciprocal learning* in Contribution 1 and study more closely in Contribution 8.

Outlook This dissertation concerns inference when observer and observation are interconnected. Despite connections to old, well-established fields like experimental design (Section 3.1) and sampling theory (Section 3.2), this kind of interconnectedness has by and large been ignored by the prevailing view in statistics and machine learning. I argued that this is mainly due to the bulk of statistics and machine learning research shedding more light on the model rather than on the data, rendering the model (as the last part of the *modeling workflow* in Figure 1.2) the bright side of the moon. Recently, however, growing interest in performativity, reciprocity and, more generally, data-centric machine learning has shifted the focus on what had appeared to be the dark side of the moon: the data and, specifically, how it can be *affected* by the model—both on the population- (performativity) and the sample-level (reciprocity) as visualized by Figure 1.2. Aiming to illuminate the dark side of the moon further, this cumulative dissertation will offer some statistical insights into the selection of training data (Part III) and testing data (Part IV). As it turns out, selecting data can be understood as a decision problem, allowing to harvest the rich literature on (statistical) decision theory under different sources of uncertainty. The following Chapter 2 sets the stage for this endeavor.

[9] One of the most cited articles in statistics of all time according to Gelman (2014).

Part II.

Introduction

2. Motivation and Background

If only one sentence was to endure from this dissertation, let it be this one:

We do not passively perceive data, we actively shape and select it.

As argued in Chapter 1, a substantial share of statistics and machine learning (ML) traditionally revolves around finding an appropriate model provided some—occasionally quite specific—data scenario. In this idealized framework, data just happens to fall in the analyst's lap. In reality, however, data is collected, chosen, merged, pre-processed and the like (Yu and Barter, 2024). In the words of Rainforth et al. (2024), "methods for optimizing data acquisition have been far less explored in the statistics and machine learning literatures than how to utilize data once we have it" (Rainforth et al., 2024, Page 100). This apparent gap between theory and practice has sparked a lot of research on "data-centric machine learning" in recent years, spanning works on data subset selection (Liu et al., 2023; Yin et al., 2024; Chhabra et al., 2024), data pruning (Marion et al., 2023; Ben-Baruch et al., 2024), benchmarks (Zhang and Hardt, 2024; Madaan et al., 2024) and sample efficiency (Kwon et al., 2023; Udandarao et al., 2024).

The cumulative dissertation at hand builds and expands on this line of work by studying the selection of data in its two most prominent facets: training (Part III) and testing (or inference[1]) data (Part IV).

For the latter, this dissertation especially focuses on selecting testing data for what Hardt (2025) calls the "iron rule"[2] of machine learning research: benchmarks. In particular, Part IV will study *multi-criteria* or *-task*[3] benchmarking problems, where algorithms are compared on a selection of testing data with respect to multiple criteria or tasks. After developing appropriate statistical theory in Contribution 10 and 11, see Sections 7.1–7.2, this dissertation also presents three concrete applications to benchmarking optimizers (Contribution 12) and language models (Contribution 13 and 14), see Sections 8.1–8.3.

Part III will demonstrate that various machine learning algorithms iteratively *self*-select training data, with far-reaching consequences for generalization (and statistical inference) from such self-selected data. Examples comprise semi-supervised learning (Chapelle et al., 2006), active learning (Settles, 2010), Bayesian optimization (Močkus, 1975), and multi-armed bandits (Lattimore and Szepesvári, 2020). In Contribution 1, we show that learning goes both ways in these methods: Parameters are not only learned from data, the latter is also chosen in light of previously learned parameters. We unify these classes of learning algorithms under the umbrella of *reciprocal*

[1] The term "inference data" has become popular in the machine learning community only recently, see, e.g., Zha et al. (2025). This dissertation, however, sticks with the classic train/test-terminology, in order to distinguish between testing setups in machine learning and *statistical* inference.

[2] In the sense of Feyerabend (1975).

[3] As explained above, throughout this dissertation, the terms multi-criteria and multi-task benchmarking are used synonymously, presupposing the different criteria reflect an inherently multidimensional concept, thus representing different *tasks*, rather than different metrics measuring the same latent construct. Refer to Jansen et al. (2024) for background on this distinction.

learning, allowing for a convergence analysis in this superstructure. Part III further contains methodological advances to some of reciprocal learning's examples (see Contributions 2-7 in Chapter 4). Moreover, it delivers generalization bounds for learning with such algorithmically self-selected data (see Contribution 8 in Chapter 5). Part III concludes with proposals of how to proceed with such self-selected data (Contribution 9 in Chapter 6) in the concrete case of regression trees and forests.

The remainder of Part II will set the stage by reviewing related literature and building the formal groundwork. After a principled motivation of imprecise probabilities in data-centric machine learning in Section 2.1, a terse review of decision theory is offered in Section 2.2. The literature review in subsequent Chapter 3 is structured as follows. After reviewing the two statistical pillars of "data centricity", namely experimental design (Section 3.1) and sampling theory (Section 3.2), the young and emerging field of data-centric machine learning is summarized (Section 3.3).

2.1. Imprecise Probabilities

Chapter 1 motivated why many research questions require the statistical scientist to construe data as a process rather than a given thing in the sense of Williamson (2024). At first sight, we have the mathematical tools at hand to construe data as a process. In statistics and machine learning, data is commonly modeled as random variables, i.e., as *functions* mapping from a sample space to (a subset of) the real numbers. However, we typically forego this sample space, abstaining from explicitly modeling the source of variation. Instead, we commonly base our analysis on assuming *one single* probability measure according to which data is generated.

In order to loosen this restriction, this dissertation will rely on remedies from the literature on imprecise probabilities (Walley, 1991; Augustin et al., 2014a). Specifically, parts of the present dissertation will use sets of probability measures—so-called credal sets (Levi, 1975)—and corresponding random variables, allowing us to, e.g., jointly model multiple possible data generation processes or multiple empirical distributions (as in Chapter 5 of this dissertation) of a set of samples.

Besides credal sets, the broader field of imprecise probability models comprises capacities (Choquet, 1954), interval probabilities (Weichselberger, 2000; Weichselberger and Pöhlmann, 1990), lower previsions (Walley, 1991; Troffaes and De Cooman, 2014), linear partial information (Kofler and Menges, 1976), possibility distributions (Dubois and Prade, 1988; Destercke and Dubois, 2014), belief functions (Dempster, 1967; Shafer, 1976) and more.[4]

Among these approaches, particularly credal sets have recently attracted increasing interest in statistics and, even more so, in machine learning. Successful applications range from Bayesian and causal networks (Cozman, 2000; Maua and de Campos, 2021; Cabanas et al., 2020; Maua and Cozman, 2020) over trees and random forests (Utkin, 2020; Utkin and Konstantinov, 2022; Sutton-Charani et al., 2012, 2013; Abellán et al., 2018) to deep learning (Caprio et al., 2024; Marquardt et al., 2023; Wang et al., 2024; Löhr et al., 2025), Gaussian processes (Benavoli and Zaffalon, 2015; Mangili, 2015; Mangili and Benavoli, 2015; Benavoli et al., 2021) and hypothesis

[4]This line of work includes connections among different theories (Augustin, 2005; Destercke et al., 2008a,b).

testing (Chau et al., 2025; Augustin and Hable, 2010; Augustin, 1999, 2002b; Huber and Strassen, 1973) as well as conformal prediction (Caprio et al., 2025; Caprio, 2025).[5]

To get a grasp of why imprecise probabilities are more expressive for representing and quantifying (subjective) uncertainty than classical probability theory, consider the famous distinction between epistemic ignorance and chance, or, more scholarly, between epistemic and aleatoric uncertainty (see also Section 3.3.3). A probability assessment of a coin flip can be 0.5 both due to absence of knowledge about the coin's design (ignorance) or perfect knowledge about a fair coin (chance). Being "Janus-faced" (Hacking, 2006, Page 12), classical probability represents both scenarios by putting probability mass 0.5 on heads, thus cannot adequately distinguish the two sources of uncertainty. Relying on imprecise probabilities, however, the scholar can represent a situation of epistemic ignorance about the coin with two point masses (two distinct probabilities of 1 on heads and tails, respectively), while the setup of an entirely fair coin can be uniquely characterized by one probability measure assigning 0.5 on heads.

When construing data as a process, it is self-evident that we are faced with both sources of uncertainty. In Yu's data lifecycle (Figure 1.1), for instance, we might encounter irreducible measurement error (aleatoric uncertainty) at the data collection step as well as ambiguity about the model choice (epistemic uncertainty) at the modeling step.

2.2. Decision Theory

In what follows, the decision-theoretic foundations for understanding sample (*aka* training data) self-selection in reciprocal learning (Part III, in particular Chapter 4 and 5) and algorithm selection in the benchmark problem (Part IV) will be laid out, drawing tangents to decision theory's adjacent disciplines of game theory (Part III) and social choice theory (Part IV).

2.2.1. What Wald Wrought: A (Nested) Zero-Sum Game Against Nature

Studying how to choose data requires a formal framework akin to those used to choose models. Contributions 1, 4, 5 and 6 of this dissertation (see Chapter 4 and Section 4.4) rest on the simple yet far-reaching insight that selecting data is a *decision problem*—very much like (selecting model parameters in) statistical inference is. Expressing data selection as a similar decision problem turns out to have similar benefits. Most prominently, it allows for a common treatment of various ways of altering data (see Contribution 1). Moreover, it allows harnessing the rich literature on decision theory—in particular by transferring decision-theoretic ideas from one concrete decision problem (like parameter selection) to another (like data selection), but also by transferring abstract (interpretation-free) decision theoretic ideas to the concrete decision problem of data selection. In this sense, the sheer recognition of data selection as an instance of decision theory itself (i.e. irrespective of parallels to statistical decision theory) already has

[5] More generally, credal sets provide a versatile framework for uncertainty quantification and representation (De Bock et al., 2014; De Campos and Antonucci, 2015; Antonucci et al., 2011; Bronevich and Rozenberg, 2019; Sale et al., 2023; Lienen and Hüllermeier, 2021; Hüllermeier and Waegeman, 2021; Hüllermeier et al., 2022; Carranza and Destercke, 2021; Fröhlich et al., 2023; Bailie and Gong, 2023; Rodemann et al., 2023a; Bailie and Derr, 2025; Fröhlich, 2025). This recent line of research builds on foundational work such as Walley and Fine (1982); Pal et al. (1992, 1993); Moral (1992); Walley and Moral (1999); Weichselberger and Augustin (1998); Couso et al. (1999); Weichselberger and Augustin (2003); Abellán and Klir (2005); Abellán et al. (2006); Abellán and Gómez (2006); Bronevich and Klir (2008).

some value. Historically, the framing of statistical inference as a decision problem has arguably paved the way for substantial parts of modern mathematical statistics, as evidenced by standard textbooks like Berger (1985); Witting (2013); French and Insua (2000); Liese and Miescke (2008). Decision theory as a common framework for statistical inference dates back to early seminal work by Wald (1947a, 1949), Savage (1951) and Hodges and Lehmann (1952).

Famously, Abraham Wald was the first to explicitly cast statistics as a two-player zero-sum game of the statistician playing against nature:

> "The true distribution [...] is chosen, we may say, by Nature, and Nature's choice is usually entirely unknown to the statistician. Thus, the situation that arises here is very similar to that of a zero-sum two-person game." (Wald, 1949, Page 173)

On a high level, the decision-theoretic setup (of statistics) is as follows: An agent (statistician) chooses an action (e.g., a realized estimator or test decision) $a \in \mathbb{A}$ from a known set $\mathbb{A}$ of actions (e.g., all potential realizations of unbiased estimators $\tilde{\theta} = a \in \mathbb{A}$). After having made this decision, the agent (statistician) observes a real-valued loss (estimation error). This loss depends on the chosen action (estimated parameter of a distribution) and the unknown state of nature (true parameter of the distribution) $\theta \in \Theta$. Formally, this gives rise to a loss function $\ell : \mathbb{A} \times \Theta \to \mathbb{R}$, fundamental to both statistics and machine learning. We will call any triplet of the type $(\mathbb{A}, \Theta, \ell)$ a (statistical) decision problem.

	θ_1	$\dots$	θ_m
a_1	c_{11}	$\dots$	c_{1m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	c_{k1}	$\dots$	c_{km}

Table 2.1.: A straightforward and expressive framework: Decision theory drafts a decision problem as a collection of states $\Theta = \{\theta_1, \dots, \theta_k\}$ like "rain" and "no rain", actions $\mathbb{A} = \{a_1, \dots, a_m\}$ like "take umbrella" and "leave umbrella at home", consequences $\{c_{11}, \dots, c_{1m}, c_{21}, \dots, c_{2m}, \cdots c_{km}\}$ like "protected by umbrella", "umbrella unnecessary","getting wet", "no rain, no umbrella", and a loss (or utility) function $\ell : \mathbb{A} \times \Theta \to \mathbb{R}$ (see Augustin and Jansen (2023)). We call the triplet $(\mathbb{A}, \Theta, \ell)$ a decision problem.

The attentive reader might have observed that this embedding is without any consideration of data $z \in \mathcal{Z}$ from some data space $\mathcal{Z}$. The statistician simply chooses a specific estimation value $\tilde{\theta} = a \in \mathbb{A}$. The typical setup in statistical inference, however, is to choose among statistical estimators $\hat{\theta} : \mathcal{Z} \to \Theta$, i.e., functions mapping from $\mathcal{Z}$ to the parameter space. Wald (1947a, 1949) directly starts with such estimators and calls them *statistical decision functions*.

Invoking the venerable master himself, Wald (1947a, Page 279) describes them as follows: "In other words, given a class Q of [probability mass, J.R.] functions and the space D of all possible decisions d, the problem is to construct a function $d(x)$, called statistical decision function, which associates with each sample point x an element $d(x)$ of D so that the decision $d(x)$ is made when the sample point x is observed."

Loosely inspired by Wald, we can denote a set of candidate estimators $\hat{\theta} : \mathcal{Z} \to \Theta$ (e.g., all unbiased ones from a reference class) by $\mathcal{D} \subseteq \Theta^{\mathcal{Z}}$. I will refer to these functions as acts in what follows (as opposed to act*ions* that were defined as specific elements from an action space above). This gives rise to the statistical decision problem $(\mathcal{D}, \Theta, \mathscr{R})$, where $\mathscr{R}$ is a risk function

$$\mathscr{R} : \mathcal{D} \times \Theta \to \mathbb{R}; (\hat{\theta}, \theta) \mapsto \int \ell(\hat{\theta}(z), \theta) \, d\mathbb{P}_\theta(z) = \mathbb{E}_{\mathbb{P}_\theta}[\ell(\hat{\theta}(Z), \theta)] \tag{2.1}$$

with $\ell(\cdot, \cdot)$ a loss function as above and $\mathbb{E}_{\mathbb{P}_\theta}$ the expectation with respect to a probability measure $\mathbb{P}_\theta$ parametrized by $\theta \in \Theta$ from a probability space $(\mathcal{Z}, \sigma(\mathcal{Z}), \mathbb{P}_\theta)$ with appropriate σ-field $\sigma(\mathcal{Z})$.

In our example of $\mathcal{D}$ consisting of all unbiased estimators from a reference class and $\ell(\cdot, \cdot)$ the quadratic loss, we have

$$\mathscr{R}(\hat{\theta}, \theta) = \mathbb{E}_{\mathbb{P}_\theta}\left[(\hat{\theta}(Z) - \theta)^2\right] = \mathrm{Var}(\hat{\theta}(Z)) + \underbrace{\mathrm{Bias}(\hat{\theta}(Z))^2}_{=0, \text{ per assumption on } \mathcal{D}} \tag{2.2}$$

due to classical bias-variance decomposition of the mean squared error. Minimizing this risk then gives the well-known uniformly minimum-variance unbiased estimator (UMVU estimator), as evidenced by the right-hand side of Equation 2.2, see, e.g., Augustin and Jansen (2023).

Leaving this simple illustration behind, the "zero-sum game against nature" by Wald (1949) can be solved by various decision criteria. For detailed accounts of a decision-theoretic embedding of likelihood-based inference and Bayesian inference, the reader is referred to Cattaneo (2005, 2013) and Berger (1985); Good (1983), respectively. Beyond such generic embeddings of statistical inference, there exist more specific decision-theoretic embeddings of statistical and machine learning methodology like interval-valued regression (Utkin and Coolen, 2011) and classification (Utkin et al., 2015). Moreover, Guillaume and Dubois (2019) apply min-max-regret criteria to classification problems with set-valued data and Utkin and Augustin (2007) generalize the imprecise Dirichlet model to allow for decision making under imprecise data. Generally, these and several more works (Jaffray, 1999; Augustin, 2001, 2002a, 2003, 2004; Utkin and Augustin, 2003, 2005; Troffaes, 2007; Huntley et al., 2014; Jansen et al., 2018, 2022a,b) bridge decision making and uncertainty representations by imprecise probabilities as discussed in Section 2.1.

Wald (1947a, 1949) solved the "zero-sum game against nature" by applying classical decision criteria such as the pessimistic minimax-criterion (Wald, 1945b), which selects the best act in the worst state of the world. Notably, he also considered the scenario of playing this game, i.e., making (statistical) decisions, *sequentially* as early as 1945 in his seminal work "Sequential Tests of Statistical Hypotheses" (Wald, 1945a), see also Wald (1947a,b) for a generalization of sequential tests to sequential decision functions. Wald's underlying idea was to derive optimal (statistical) decision functions for streams of incoming data. In addition to the usual inferential statement in the form of $\hat{\theta}$, the statistician has to decide when to stop data collection. This gave rise to a stream of research on sequential statistical analyses (Siegmund, 1985), "group sequential" methods (Pocock, 1977; Kim and Demets, 1987), adaptive designs (Bauer and Köhne, 1994), and alpha spending (DeMets and Lan, 1994), as outlined above, mostly in the medical and biometric or epidemiological context of longitudinal studies.

As touched upon earlier, this clearly extends the classical *static* setup (that was criticized in Chapter 1) to a dynamic one. However, it does not give up the data-model dichotomy nor does it account for feedback loops. Information still flows in a one-way street. Apart from the one-time stopping decision, there is no effect of the statistician and their model (*explanans*) on nature (actual *explanandum*) or on data initially obtained from nature (sampled *explanandum*) in this zero-sum game among the two. This changes in Contributions 1 through 8 presented in Part III of this dissertation.

The feedback loops induced by performative predictions and reciprocal learning (see Section 2) render the information flows (between model and nature on the one hand as well as between model and sample on the other hand) a two-way street, respectively. In addition to streams of incoming data as in Wald (1947a,b), we are faced with streams of sequentially learned models that affect the streams of incoming data. In the case of performativity, this has been studied in algorithmic game theory as *two-player games in strategic form*, see, e.g., Nisan et al. (2007). Here, nature—often composed of intelligent entities—(strategically) reacts to predictions being made about itself. In case of reciprocal learning, the game essentially becomes a one-player game *nested in* the classical two-player game. The statistician's (or machine learner's) decision affects the sample instead of the whole population, i.e., the sampled *explanandum* instead of the actual one. After an initial draw from nature's population, the nested one-player game starts: The statistician can change the sample without nature intervening. This game is nested inside the classical two-player game, since in the end, the model should perform well on the unseen populations, not on the self-selected sample. Nature still has the final say.

Let us turn to this sample-level feedback loop in more detail. In Contribution 1 (Rodemann et al., 2024), we generalize several popular machine learning algorithms like active learning, bandits and self-training under the umbrella of reciprocal learning. We also embed reciprocal learning into sequential decision making (Rodemann et al., 2024, Section 2.2) as follows.[6] First, the learner decides for a $\hat{\theta}_t$ from class $\mathcal{D}$ by minimizing an empirical version of the risk (equation 2.1). This corresponds to solving a decision problem characterized by the triple $(\mathcal{D}, \Theta, \mathscr{R})$ as above. Second, features $x_t \in \mathcal{X}$ are chosen and data points $\mathcal{Z} \ni z_t = (x_t, y_t)$ are added to or removed from the sample inducing a new empirical distribution. Depending on the concrete instance of reciprocal learning, y_t is predicted (self-training), queried (active learning) or observed (bandits) based on x_t.[7] Note that for some of these instances, e.g., in bandits, nature intervenes inside the nested one-player game by determining the label y_t of x_t, while in others, e.g., in self-training, nature does not intervene (the label y_t is self-predicted by the model here).

The features x_t are found by solving a decision problem $(\mathcal{X}, \Theta, \ell_{\hat{\theta}_t})$. Two things stand out. First and foremost, the loss function $\ell_{\hat{\theta}_t}$ depends on the previous decision problem's solution $\hat{\theta}_t$. Second, this time, the act space corresponds to the feature space $\mathcal{X}$, not the parameter space Θ. Loosely speaking, the data is judged in light of the parameters here. The perspective is hence symmetric to classical likelihood approaches in statistics or machine learning, where parameters are judged in light of the data. This twist in perspective will clear the way for another type of regularization—that of data, not of parameters—in Contribution 1, see Chapter 4.

This sequential formulation explicitly models the reciprocal feedback loops discussed in Section 2: Data selection in iteration $t \in \{1, \dots, T\}$ depends on previous model (identified by parameters) selection $\hat{\theta}_t$ via the family of loss functions

$$\ell_{\hat{\theta}_t} : \mathcal{X} \times \Theta \to \mathbb{R}; (x, \theta) \mapsto \ell_{\hat{\theta}_t}(\theta, x). \tag{2.3}$$

To make this family more tangible, consider reciprocal learning's special case of self-training in a semi-supervised learning setup, where unlabeled data, denoted by $\mathscr{U}_t \subseteq \mathcal{X}$, is available in

[6] Boosting can be considered another special case of reciprocal learning, see Rodemann and Bailie (2025, Page 2). Notably, Breiman (1999) already recognized boosting's character as a sequential "prediction game", akin to our embedding of reciprocal learning into sequential decision theory presented here.

[7] As we note in Rodemann et al. (2024, Section 2.2), formally, y_t is drawn from some model $\mathbb{P}_{Y|X=x_t}$ which might be degenerate or unknown, depending on the concrete algorithm.

addition to standard labeled data $\mathscr{D}_t \subseteq \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.[8] Self-training algorithms first fit a model $\hat{\theta}_t$ on the labeled dataset. This model is then used to predict labels $\hat{y} = \hat{\theta}_t(x)$ for $x \in \mathscr{U}_t$. In the next stage, a subset of the unlabeled instances is selected and incorporated into the training set together with their predicted labels $\hat{y}$, referred to as pseudo-labels. In Contribution 4, see Section 4.4, we identify this selection of pseudo-labeled data as a decision problem $(\mathscr{U}_t, \Theta, \ell)$ with an act space of unlabeled data $\mathscr{U}_t$ to be selected from, i.e., instances x_i as acts, a space of unknown states of nature (parameters) Θ and a loss function[9]

$$\ell_{\hat{\theta}_t} : \mathscr{U}_t \times \Theta \to \mathbb{R}; \tag{2.4}$$

$$(x_i, \theta) \mapsto \ell_{\hat{\theta}_t}(x_i, \theta) = -p(\mathscr{D}_t \cup (x_i, \hat{y}_i(\hat{\theta}_t)) \mid \theta), \tag{2.5}$$

where $p(\mathscr{D}_t \cup (x_i, \hat{y}_i(\hat{\theta}_t)) \mid \theta)$ denotes the likelihood function, i.e., the probability of observing $\mathscr{D}_t \cup (x_i, \hat{y}_i(\hat{\theta}_t))$ given a model θ. This likelihood function depends on the previous decision problem's solution $\hat{\theta}$ via the pseudo-labels $\hat{y}_i(\hat{\theta}_t)$. In Contribution 4 we go on to derive Bayes-optimal, minimax, and max-max acts for this data selection problem (Rodemann et al., 2023b, Theorem 1–3), which are extended to robust Bayesian criteria in Contribution 5, see Section 4.4.

Generally, the family of loss functions in Equation 2.3 describes all potential loss functions in the data selection problem arising from respective solutions of the previous parameter selection problem. Redefining this family of functions as a single one $\tilde{\ell} : \mathcal{X} \times \Theta \times \Theta \to \mathbb{R}$ makes it clear that we can retrieve decision criteria $c : \mathcal{X} \times \Theta \to \mathbb{R}$ from it, which are generalizations of classical decision criteria $c : \mathcal{X} \to \mathbb{R}$ retrieved from classical losses $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$ as introduced above. We work with these kinds of decision criteria in Contribution 1 (see Chapter 4).

Summing up, reciprocal learning can be thought of as a sequential decision-making problem (Rodemann et al., 2024, Section 2.2):

$t = 1$: $\hat{\theta}_1$ solves decision problem $(\mathcal{D}, \Theta, \ell)$ $\quad$ $t = 2$: $\hat{\theta}_2$ solves decision problem $(\mathcal{D}, \Theta, \ell_{x_1(\theta_1)})$ $\quad \ldots$

$\quad$ x_1 solves decision problem $(\mathcal{X}, \Theta, \ell_{\hat{\theta}_1})$ $\quad$ x_2 solves decision problem $(\mathcal{X}, \Theta, \ell_{\hat{\theta}_2})$ $\quad \ldots$

As touched upon, performative prediction can be cast as a sequential decision problem in a similar fashion, this time with two players. It becomes a zero-sum Stackelberg game (Von Stackelberg, 2010), where the *explanans* (the statistician) moves first, and the *explanandum* (nature) moves second. A broad body of literature in algorithmic game theory (Brückner and Scheffer, 2011; Kuleshov and Precup, 2014), strategic classification (Hardt et al., 2016; Miller et al., 2020), and—more broadly—strategic learning (Shavit et al., 2020; Bechavod et al., 2021; Vo et al., 2024, 2025) hinges upon such a game-theoric embedding of strategic reactions to predictions.

Rounding off this Section, it shall be noted than none other than Abraham Wald himself motivated his pessimistic minimax-criterion by a reasoning akin to zero-sum Stackelberg games (yet not using the name "Stackelberg"), where nature moves second: "if the statistician is in complete ignorance as to Nature's choice, it is perhaps not unreasonable to base the theory [...] on the assumption that Nature wants to maximize [the risk, J.R.]" (Wald, 1945b, Page 285).

As recognized by Goktas and Greenwald (2021, Page 3), a *local* variant of Wald's minimax-model is (via Scarf (1957)) the main paradigm underlying (distributionally) robust optimization (Wiesemann et al., 2014; Gao and Kleywegt, 2023) generally and (distributionally) robust

[8] For exposition, we assume absence of duplicated observations such that we can understand $\mathscr{U}_t$ and $\mathscr{D}_t$ as sets, respectively.

[9] In Contribution 4, we use the equivalent formulation of Equation 2.4 as utility, considering the likelihood directly instead of the negative likelihood, see Rodemann et al. (2023b, Section 2.2).

empirical risk minimization (Shafieezadeh Abadeh et al., 2015) specifically. Here, nature can still play second, but is restricted to a specific set of acts known to the *explanans* (statistician). This is typically formalized by divergence- (e.g., Wasserstein-)balls around some base distributions, with respect to which a supremum risk functional is minimized. Distribution maps (Perdomo et al., 2020) that model how nature reacts to predictions typically come with Lipschitz-conditions that enforce such a locality in the distributions space.[10] We rely on exactly this embedding for deriving generalization bounds for reciprocal learning in Contribution 8 (Rodemann and Bailie, 2025) as summarized in Chapter 5 of this dissertation.

Abraham Wald wrought and anticipated, but could not witness the full materialization of his visionary ideas. He tragically passed away ahead of his time in an airplane crash in the Nilgiri mountains in southern India, see the obituary by Oskar Morgenstern in *Econometrica*: "Thus, suddenly, in a remote and wild corner of a distant land, ended one of the most brilliant careers in the social sciences. A tremendous loss was felt among economists and statisticians alike, for in both fields Wald had made great contributions and, indeed, produced a decisive turn in method and purpose" (Morgenstern, 1951, Page 361).

2.2.2. Yet Another Decision Problem: Benchmarking

A more general approach to formalizing decision problems is as follows. As above, acts shall be defined as functions, but this time, with the set of states of nature Θ as domain and a general consequence space C as codomain, i.e., $X : \Theta \to C$. A decision problem $\mathcal{G}$ shall be defined as a subset of all possible acts $C^{\Theta} = \{X : \Theta \to C\}$; see Jansen (2018, Section 1.1). Intuitively, in this setup, an agent takes a decision X, whose outcome C depends on an unknown state of the world Θ. Decision theory then revolves around finding appropriate choice functions $ch : 2^{\mathcal{G}} \to 2^{\mathcal{G}}$ satisfying $ch(\mathcal{D}) \subseteq \mathcal{D}$ for all non-empty $\mathcal{D} \in 2^{\mathcal{G}}$. We refer to Jansen (2018, Section 2.2) on how statistics can be embedded into this formalization akin to the embedding discussed in Section 2.2.1 above.

The sets $ch(\mathcal{D})$ are referred to as *choice sets*. Their interpretation depends on the quality of the information underlying the choice function. Jansen et al. (2025b) differentiate the strong and weak view. Under the *strong view*, $ch(\mathcal{D})$ represents the set of *optimal* acts within $\mathcal{D}$. In contrast, the *weak view* regards $ch(\mathcal{D})$ as the set of acts in $\mathcal{D}$ that *cannot be ruled out* given the available information.

The appeal of this formalization lies in both its simplicity and its natural account for two different information sources that are required to specify non-trivial choice functions, see Jansen (2018, Section 1.1): The agent's preferences on C and the uncertainty about the world Θ. As pondered upon in Section 3.3.3, these are two different kinds of model uncertainty within epistemic uncertainty. While the machine learning literature mostly tackles the uncertainty about Θ (e.g., by credal sets (Singh et al., 2024; Caprio et al., 2024; Sale et al., 2023), our point of attack in Part IV will be the agent's preferences.[11] Intuitively, to come up with decision criteria for any agent, we have to elicit this agent's preferences regarding the consequences of their decisions and their (absence of) knowledge about nature that affects these very consequences.

[10] This enforcement is a direct consequence of the Kantorovich-Rubinstein Lemma (Kantorovich and Rubinstein, 1958).

[11] This stylized separation, however, does not apply to the axiomatic approaches to finding decision criteria by Savage (1951) and Anscombe and Aumann (1963), as these two approaches require axioms on Θ and C in a joint way, see Jansen (2018).

The acclaimed Von Neumann–Morgenstern utility theorem (von Neumann and Morgenstern, 1944), for instance, specifies four axioms (completeness, transitivity, continuity, and independence) on the agent's preferences on C[12] that allow representing them by a cardinal (measurable) *utility function*, or symmetrically, by a loss function $\ell : C \to \mathbb{R}$. If, additionally, the uncertainty about Θ can be adequately described by a probability measure π on Θ, we can specify the classical choice function based on the expected loss (i.e., risk) as follows

$$ch_{\ell,\pi}(\mathcal{D}) = \Big\{X' \in \mathcal{D} : \mathbb{E}_\pi(\ell \circ X') \leq \mathbb{E}_\pi(\ell \circ X) \text{ for all } X \in \mathcal{D}\Big\}, \tag{2.6}$$

for all $\mathcal{D} \subseteq \mathcal{G}$. In words, $ch_{\ell,\pi}(\mathcal{D})$ chooses acts from $\mathcal{D}$ that minimize the expected loss, corresponding to the strong view of choice functions selecting optimal acts. I shall note in passing that both Θ and C were implicitly equipped with appropriate σ-fields, extending them to measurable spaces $(\Theta, \mathcal{S}_\Theta)$ and $(C, \mathcal{S}_C)$ and rendering our acts $X, X' \in \mathcal{D}$ random variables with well-defined π-expectations.

The alert and conscientious reader might have already noticed that $ch_{\ell,\pi}(\mathcal{D})$ defined nothing less than risk minimization (Vapnik, 1991, 1998), the backbone of nearly all of machine learning (Shalev-Shwartz and Ben-David, 2014). For $ch_{\ell,\pi}(\mathcal{D})$ to correspond to *empirical* risk minimization, all we have to do is to replace π by its empirical analogue, the empirical measure $\hat{\pi}$.

Now imagine our acts to have multivariate consequences. For the classical machine learning example, C might not only consist of, say, accuracy, but also of interpretability or fairness. Generally, multivariate consequence spaces are everywhere. Consider, for instance, multivariate poverty measurement (Garcia-Gomez et al., 2019).[13]

The main motivation the contributions in part IV of this dissertation, however, is an essential part of data-centric machine learning (see Eyuboglu et al. (2022); Mazumder et al. (2022); Zha et al. (2025); Chen et al. (2023); Huang et al. (2024)), namely the benchmark problem:[14] How to compare multiple algorithms (like classifiers, optimizers or language models) on multiple instances (typically testing datasets, or problems, tasks, prompts etc.) with respect to multiple criteria (like prediction error, compute time, sparsity, coherence, diversity, and many more)? Notably, these latter make up another example of a multivariate C. It is straightforward to see that in such multivariate consequence spaces, the Von Neumann–Morgenstern axioms are often violated, since competing criteria may conflict and prevent preferences from being complete, transitive, or independent. Here, the perspective of (computational) social choice theory (Brandt et al., 2016; Conitzer, 2010) naturally comes into play. Broadly speaking, social choice theory offers methods for aggregating individual preferences, judgments, or interests into collective decisions. Originating in social welfare and voting theory (Borda, 1781; Condorcet, 1785; Arrow, 1950), the transfer from voting to multicriteria benchmarking is self-explanatory: algorithms are candidates and multiple criteria make up multiple voters, see e.g., Zhang and Hardt (2024), giving rise to multivariate C.

[12] More specifically, the axioms relate to the set of lotteries over C, where a "lottery" assigns a probability to each of the mutually exclusive outcomes in C, see von Neumann and Morgenstern (1944) for details.

[13] Mostly owing to Sen (1985), it has become consensus that poverty has more facets than income or wealth. It is perceived as a multidimensional concept, involving variables like education or health status, that are often ordinally scaled. In Contribution 10, we illustrate our methodology, among other applications, by this kind of multidimensional poverty measurement, see Section 7.1.

[14] Which has received strongly growing interest in recent years (Demšar, 2006; Benavoli et al., 2017; Jansen et al., 2024; Zhang and Hardt, 2024; Zhang et al., 2025; Hardt, 2025).

Think of two competing language models $X : \Theta \to C$ and $X' : \Theta \to C$. While X produces more semantically coherent (criteria c_1) text, X' outputs text with higher token diversity (criteria c_2). How should they be ranked? This depends on what is commonly referred to as *aggregation* (John, 2006). In practice, two extreme cases can be distinguished.

Firstly, many benchmarks apply a real-valued weighting function $w_1 c_1 + \cdots + w_d c_d$ with (often times arbitrary) weights $w_i, i \in \{1, \ldots, d\}$ for components c_i of c, allowing to harness the total order just like $ch_{\ell,\pi}$ does, see Equation 2.6. Second, some benchmarks retreat to the overly conservative Pareto front by only choosing language models that are undominated with respect to all d criteria $c \in C \subseteq \mathbb{R}^d$. In other words, the (weak) Pareto dominance relation $\preceq$ is given by $c' \preceq c \iff c'_i \leq c_i$ for all $i \in \{1, \ldots, d\}$.

But how to account for the randomness in the output of our two language models $X, X' \in C^\Theta$? The classical Pareto front is very strict in this regard. It considers only those language models worthy to be selected who are not only Pareto-undominated on multivariate C, but also with respect to all states of the world. Thus, we define $\preceq_\Theta$ as follows: $X' \preceq_\Theta X \iff X'(\theta) \preceq X(\theta)$ for all $\theta \in \Theta$. This relation is also called point-wise Pareto dominance relation. Formally, the *Pareto* choice function selects admissible acts

$$ch_{\text{Pareto}}(\mathcal{D}) = \left\{ X' \in \mathcal{D} : \nexists\, X \in \mathcal{D} \text{ with } X \prec_\Theta X' \right\} \tag{2.7}$$

for all $\mathcal{D} \subseteq \mathcal{G}$. In words, $ch_{\text{Pareto}}(\mathcal{D})$ returns precisely those acts (here: language models) for which no alternative performs at least as well in all criteria and strictly better in at least one criterion for at least one state θ.

This strict choice function can be weakened to one based on (first-order) stochastic dominance. Instead of requiring one language model to dominate another in every state of the world, it requires one language model to yield at least as low expected loss for all monotone loss functions, with strict inequality for at least one (see also Lehmann (1955)). Concretely, given any preorder[15] $\precsim$ (like the Pareto dominance relation $\preceq$ from above[16])

$$ch_{\precsim,\pi}(\mathcal{D}) = \left\{ X' : \nexists X \;\; \begin{array}{l} \forall \ell \in \mathcal{L}_{\precsim} : \mathbb{E}_\pi(\ell \circ X - \ell \circ X') \leq 0 \\ \exists \ell \in \mathcal{L}_{\precsim} : \mathbb{E}_\pi(\ell \circ X - \ell \circ X') < 0 \end{array} \right\} \tag{2.8}$$

for all $\mathcal{D} \subseteq \mathcal{G}$, where $\mathcal{L}_{\precsim}$ is the set of all $\precsim$-isotone (and measurable) loss functions $\ell : C \to [0, 1]$. Thus, we choose all acts that are not excluded by every compatible risk-minimizer. This corresponds to the weak interpretation from above: the agent is *indifferent* between acts in $ch_{\precsim,\pi}(\mathcal{D})$ or deems them *incomparable* (see Jansen et al. (2025b,a)).

This stochastic dominance relation will be the starting point for our contributions to benchmarking theory in Chapter 7 of Part IV. Specifically, these contributions rely on generalizing the stochastic dominance relation to the generalized stochastic dominance relation (GSD). While the GSD can applied to a myriad of applications, see Contribution 10, we are mainly motivated by situations where different dimensions of C have different scales of measurement, e.g., some cardinal (semantic coherence measure) and some ordinal (human rankings), as illustrated, e.g., by Contribution 14 in Section 8.3. While the general setup in benchmarking is, as has become evident, the selection of algorithms (like language models), we will also study how to select instances (i.e., testing data)

[15] A preorder is a binary relation $R \subseteq M \times M$, $M \neq \emptyset$, if $(a, a) \in R$, (*reflexive*) and $(a, b), (b, c) \in R \Rightarrow (a, c) \in R$ (*transitive*).

[16] It can be easily verified that $\preceq$ is reflexive and transitive, thus a preorder.

for the design of benchmark suites, based on analyzing (via robust statistics in the classic sense (Huber, 1981)) how sensitive the statistical conclusions from our benchmark analysis are under perturbed testing data.

All in all, this Section demonstrated how decision theory provides a solid theoretical framework for studying the problems outlined in Section 2 above: Training data selection in reciprocal learning (Part III), algorithm selection via benchmarking (Part IV) and testing data selection for benchmark suite design (ibid.). Along the way, it has become clear how universal the decision-theoretic perspective really is: From the simple umbrella example (Table 2.1) via basic statistical testing and estimation (Equation 2.2) to modern nested empirical risk minimization (Equation 2.3) and self-training (Equation 2.4) as well as benchmarking of large language models (Equation 2.7). It almost seems like there are hardly any problems in statistics and machine learning that cannot, at least in a degenerate way, be embedded into decision theory. Of course, this does not answer the question of how fruitful such embeddings really are. In our cases, as we have seen, they indeed prove useful as several ideas and theorems can be transferred from one decision problem to another. For instance, applying the Bayes criterion to pseudo-labeled data selection as sketched above, see Equations 2.3–2.5 and Rodemann et al. (2023b, Theorems 1–3), allowed us to deploy the whole *approximate Bayesian inference (ABI)* machinery (Kass et al., 1989; Alquier, 2020)—developed for Bayesian estimation, i.e., *parameter* selection—to (pseudo-labeled) *data* selection, allowing our theoretical results and methods to be implemented by making them computationally feasible; see Chapter 4.4 for details.

In order to grasp how these decision-theoretic embeddings align with modern data-centric machine learning, we now have to leave the formalism behind and explore the existing literature in the following chapter. However, we shall return to this decision-theoretic framework in both Part III and Part IV. Ultimately, decision theory can indeed be regarded as the unifying formal thread or, if you allow, the glue that holds together this dissertation.

3. Literature Review

> "Now it is the twenty-first century when, as the paper reminds us, we are being asked to face problems that never heard of good experimental design."
>
> — From Bradley Efron's comment (Efron, 2001) on Leo Breiman's famous article "Statistical Modeling: The Two Cultures" (Breiman, 2001b)

In what follows, I will briefly review related work on data-centric machine learning and uncertainties therein, including a tentative definition and a data-driven survey of data-centric machine learning, respectively. Before that, however, I shall start by discussing the two fundamental pillars of any statistical perspective on data: (sequential) experimental design and sampling theory. A formal introduction into decision theory was already provided by Chapter 2.2. Literature that relates to the concrete contributions of this dissertation in a more specific way will be reviewed in Chapters 4 through 6 (Part III) and Chapters 7 and 8 (Part IV) as well as in the attached publications themselves.

As it will turn out, already early pioneers of experimental design (Smith, 1918; Fisher, 1926, 1935) and sampling theory (Cochran, 1942) employed a "data-centric" view on what are now called "learning" problems, rendering this literature review highly relevant for subsequent contributions presented in this dissertation. This comes as little surprise, since across a wide range of fields in the empirical sciences researchers deliberately (Smith, 1918; Fisher, 1935) or inadvertently (Cochran, 1942) determine the sample upon which their inferential conclusions rest, see also Rodemann and Bailie (2025, Section 3) and previous Chapter 2, aligning closely with definitions of *data-centric machine learning* (Zha et al., 2025; Oala et al., 2024; Singh, 2023), including the one presented in Section 3.3.1 later.

3.1. Sequential Experimental Design

> "The whole art and practice of scientific experimentation is comprised in the skillful interrogation of Nature. Observation has provided the scientist with a picture of Nature in some aspect, which has all the imperfections of a voluntary statement. He wishes to check his interpretation of this statement by asking specific questions..."
>
> Joan Fisher Box: *R. A. Fisher, The Life of a Scientist*, cited after Pearl and Mackenzie (2018, Page 144)
>
> — Fisher Box (1978)

Sir Ronald A. Fisher is widely regarded the founding father of statistical experimental design, the study of how to optimally design experiments for data collection, or, in the vivid, picturesque words of his daughter Joan Fisher Box, the "skillful interrogation of Nature" (Fisher Box, 1978). Following Dempster (1979), Fisher "introduced randomization and factorial experimentation,

thus founding modern statistical design and analysis of experimental data" (Dempster, 1979, Page 537).

Fisher's works titled "The Arrangement of Field Experiments" (Fisher, 1926) and "The Design of Experiments" (Fisher, 1935) are indeed often considered the founding document of experimental design. However, it was actually Kirstine Smith, a Danish student of Karl Pearson, who pioneered modern experimental design by publishing her seminal paper "On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance they give Towards a Proper Choice of the Distribution of Observations" (Smith, 1918) in *Biometrika* as early as 1918. Even earlier, Peirce and Jastrow (1885) already designed a randomized (and blinded) experiment, building on Charles S. Peirce's seminal "Theory of Probable Inference" (Peirce, 1883), see also Peirce (2014). For a detailed account of the history of experimental design, and particularly of the role of randomization therein, the interested reader is referred to Hacking (1988).

One reason why many scholar nowadays primarily associate Fisher with experimental design might be the *Fisher* information matrix (FIM), which still is the pivotal quantity in (frequentist) experimental design. Given some design φ affecting the experiment's result y, from which we want to learn something about a parameter θ, the FIM can be written as

$$\mathrm{FIM}(\theta, \varphi) = \mathbb{E}_{p(y|\theta,\varphi)} \left[\left(\frac{\partial}{\partial \theta} \log p(y \mid \theta, \varphi) \right) \left(\frac{\partial}{\partial \theta} \log p(y \mid \theta, \varphi) \right)^{\top} \right], \tag{3.1}$$

where $p(y \mid \theta, \varphi)$ is an assumed model that can be seen as the likelihood function and $\frac{\partial}{\partial \theta}(\cdot)$ the derivative with respect to θ, see the standard textbooks by Pukelsheim (2006); Ryan (2007) or surveys by Chaloner and Verdinelli (1995); Ryan (2007); Atkinson and Doney (1992); Rainforth et al. (2024). The FIM can be expressed as the second derivative of the likelihood, as already sketched on page 328 in Fisher's seminal "Mathematical Foundations of Theoretical Statistics" (Fisher, 1922).

The Fisher Information Matrix tells the scholar about the amount of information an experimental design φ delivers about θ in expectation in a very distinguished (that is, component-wise) way: For multivariate θ, the FIM has a column/row for each element of the parameter vector. From a computational point of view, this multivariate aspect of the Fisher information can also be seen as a bug rather than a feature, since it produces a multi-objective optimization problem when trying to optimize with respect to φ for designing an experiment. This is typically circumvented by optimizing a summary statistic like the trace or the determinant of the FIM.

In fact, many modern statistical approaches to experimental design can be roughly summarized by three criteria, namely A-, D-, and E-optimality, all of which refer to the FIM, see Pukelsheim (2006) for an overview. A-optimality means minimal average variance of parameter estimates by reducing the trace of the inverse information matrix. D-optimality aims to maximize the determinant of the information matrix, thereby minimizing the (generalized) variance and shrinking the volume of the confidence region. E-optimality instead maximizes the smallest eigenvalue of the information matrix, improving precision in the (geometrically interpreted) weakest direction of the FIM.

A more severe drawback than this multivariate aspect is the simple fact that the Fisher info depends on the unknown θ, see Equation 3.1. Fisher introduced it for linear Gaussian models, where the θ neatly drops, rendering the FIM independent of the true parameter. In general,

however, this is not the case. To the best of my knowledge, Box and Lucas (1959) were first to explicitly address this limitation by introducing approximate information matrices and *locally* optimal experimental designs for non-linear models.

The statistical development of sequential experimental design since Fisher, Smith and Peirce can be characterized by two strands, both of which have quite some relevance for our non-static perspective on *data lifecycles*, see Chapter 2. The first strand directly extended Fisher's and Smith's work to the sequential case, while sticking to the frequentist paradigm. The other line of works developed a Bayesian theory of how to design experiments, where the sequential perspective is natively embedded via Bayesian updating, see, e.g., Rainforth et al. (2024, Figure 1), which is why I will review the Bayesian approach in a little more detail later.

Apart from these sequential extensions, the general field of experimental design has folded into the thriving field of causal inference as part of what Judea Pearl calls the "causal revolution" (Pearl, 2018, Page 3), see, e.g., Pearl (2009), Rubin (1974), Imbens and Rubin (2015), Rosenbaum and Rubin (1983), Holland (1986) and Angrist et al. (1996). Designing interventions to estimate causal effects (like the average treatment effects of a drug) has become a central pillar of this literature on causal inference.

The biggest leap ahead[1] along the frequentist strand of work certainly was the "Sequential Design of Experiments" by Chernoff (1959), which derives an asymptotically optimal strategy for experimental design φ given "a finite number of states of nature" (Sen, 1992, Page 341). This work builds on the introduction of the two-armed bandit problem by Robbins (1952), but with a statistical focus on hypothesis testing loosely leaning on the sequential "probability ratio" test by Wald (1945a, 1947b), see Sen (1992, Pages 340-343) for a brief comparison to both Robbins (1952) and Wald (1947b). The relevance of the problem is self-evident. In Chernoff's own words in the paper's introduction,

> "[...] considerable scientific research is characterized as follows. The scientist is interested in studying a phenomenon. At first he is quite ignorant and his initial experiments are preliminary and tentative. As he gathers relevant data, he becomes more definite in his impression of the underlying theory. This more definite impression is used to construct more informative experiments. Finally after a certain point he is satisfied that his evidence is sufficient to allow him to announce certain conclusions and he does so." (Chernoff, 1959, Page 755)

Chernoff's article builds on earlier work of his own that derives Fisher-information–based locally optimal designs for nominal parameters (Chernoff, 1953), see also his later textbook (Chernoff, 1987) summarizing his contributions. Modern textbooks of experimental design (Pukelsheim, 2006; Ryan, 2007) indicate how central the sequential extension of Chernoff still is.

The second strand of work takes a Bayesian perspective on experimental design. What several articles and books by Peirce, Smith and Fisher (Peirce, 1883; Peirce and Jastrow, 1885; Smith, 1918; Fisher, 1926, 1935) provided for frequentist experimental design, one single scholar delivered for the Bayesian counterpart: Dennis Victor Lindley. In what appears to have been a single-handed effort, he introduced the pivotal quantity of Bayesian experimental design, the (expected) information gain (see Equations 3.2 through 3.4 below), in his paper programmatically titled

[1] As a matter of fact, the "Sequential Design of Experiments" by Chernoff (1959) was included the second volume of the collection of "Breakthroughs in Statistics" (Kotz and Johnson, 1992).

"On a measure of the information provided by an experiment" (Lindley, 1956), see also Raiffa and Schlaifer (1961); Lindley (1972); Lindley and Smith (1972).[2]

As always in the Bayesian universe, this presupposes some prior $p(\theta)$—representing pre-experimental beliefs—on the parameter of interest θ. Just like above, we consider an experiment with a controllable design φ independent of $p(\theta)$. Following the recent survey by Rainforth et al. (2024), Bayesian approaches to experimental design then define the Information Gain (IG) about θ from observing y under design φ as the reduction in Shannon entropy $H(\cdot)$ (Shannon, 1948) from prior to posterior:

$$\begin{aligned}\mathrm{IG}_\theta(\varphi, y) &:= H[p(\theta)] - H[p(\theta \mid y, \varphi)] \\ &= \mathbb{E}_{\mathbb{P}_{\theta|y,\varphi}}[\log p(\theta \mid y, \varphi)] \;-\; \mathbb{E}_{\mathbb{P}_\theta}[\log p(\theta)],\end{aligned} \tag{3.2}$$

where $p(\theta \mid y, \varphi) \propto p(\theta)\, p(y \mid \theta, \varphi)$. Because y is unknown at design time, one typically considers the expected information gain (EIG) rather than the IG. The EIG results from marginalizing out y via the marginal predictive $p(y \mid \varphi) := \mathbb{E}_{p(\theta)}[p(y \mid \theta, \varphi)]$:

$$\begin{aligned}\mathrm{EIG}_\theta(\varphi) &:= \mathbb{E}_{\mathbb{P}_{y|\varphi}}[\mathrm{IG}_\theta(\varphi, y)] \\ &= \mathbb{E}_{\mathbb{P}_\theta}\mathbb{E}_{\mathbb{P}_{y|\theta,\varphi}}[\log p(\theta \mid y, \varphi) - \log p(\theta)] \qquad (3.3) \\ &= \mathbb{E}_{\mathbb{P}_\theta}\mathbb{E}_{\mathbb{P}_{y|\theta,\varphi}}[\log p(y \mid \theta, \varphi) - \log p(y \mid \varphi)], \qquad (3.4)\end{aligned}$$

see Rainforth et al. (2024, Section 2.1). Equations (3.3) through (3.4) show that $\mathrm{EIG}_\theta(\varphi)$ equals the mutual information $I(\theta, y \mid \varphi)$ of θ and y given φ, i.e., the expected Kullback–Leibler divergence from posterior to prior. Based on Equation 3.4, some authors equivalently call this the expected reduction in predictive uncertainty. The target θ need not be explicit model parameters; it can represent, for example, the output of an algorithm or future predictions, see, e.g., Hennig and Schuler (2012); Wang and Jegelka (2017); Kleinegesse and Gutmann (2021). This is reminiscent of our embedding of data acquisition into (Bayesian) decision theory, see Contributions 4 through 6 in Section 4.4 and preliminaries sketched in Section 2.2.

As foreshadowed above, the sequential setting is now captured by Bayesian experimental design in a very natural way. Further following Rainforth et al. (2024), let the designs φ and outcomes y decompose as $\varphi = \{\varphi_1, \ldots, \varphi_T\}$ and $y = \{y_1, \ldots, y_T\}$. Denote the history up to iteration $t-1$ as $h_{t-1} = \{(\varphi_k, y_k)\}_{k=1}^{t-1}$ with $h_0 = \emptyset$ and absence of ties per assumption. The *incremental* expected information gain at step t is

$$\mathrm{EIG}_\theta(\varphi_t \mid h_{t-1}) := \mathbb{E}_{\mathbb{P}_{\theta|h_{t-1}}}\mathbb{E}_{\mathbb{P}_{y_t|\theta,\varphi_t,h_{t-1}}}\left[\log \frac{p(y_t \mid \theta, \varphi_t, h_{t-1})}{p(y_t \mid \varphi_t, h_{t-1})}\right]. \tag{3.5}$$

This coincides with a standard EIG evaluated under the updated prior, see Rainforth et al. (2024, Sec. 2.2). Trivially, this strategy weakly dominates non-adaptive (i.e., static) designs that fix all $\{\varphi_t\}_{t=1}^T$ upfront because each decision exploits newly acquired data. Acquisition functions in Bayesian optimization (see Contributions 2 and 3 in Section 4.3) are natural examples of $\mathrm{EIG}_\theta(\varphi_t \mid h_{t-1})$ in case θ is the target function's optimum. Moreover, our Bayesian selection criteria in pseudo-label selection in self training (see Contribution 4 in Section 4.4) can be motivated by minimizing the $\mathrm{EIG}_\theta(\varphi_t \mid h_{t-1})$, because they aim at selecting pseudo-labels with low predictive uncertainty, i.e., those with lowest reduction in predictive uncertainty, i.e., with highest predictive confidence.

[2] The latter was also included in a volume of the "Breakthroughs in Statistics" (Kotz and Johnson, 1998), just like its frequentist counterpart by Chernoff (1959) discussed above.

Summing up, both frequentist and Bayesian accounts of how to acquire data sequentially provide the methodological foundation for data-centric methods in machine learning. As mentioned above, we refer the reader to a review of more specific literature related to the contributions presented in this dissertation to Parts III and IV, particularly to the literature reviews in the contributions themselves.

3.2. Sampling Theory

Roughly speaking, sampling theory derives methodology for how to proceed with given samples. This is at odds with previously discussed experimental design, which aims at methodological answers on how to arrive at samples in the first place. Thus, the perspective of sampling theory is—at first sight—a manifestation of viewing data as a given, and consequently not relatable to or even incompatible with construing data as a process or lifecycle, as described by Figure 1.1 specifically and Chapter 2 more generally.

At second glance, however, a more nuanced picture emerges. While sampling theory indeed asks how to proceed with data that just happens to be there, after all, the answers given by the theory *expand* the scope to the data collection steps. In fact, explicitly modeling the data collection (mostly survey-based) is at the core of sampling-theoretical methodology. In this way, sampling theory provides information regarding the possibilities and limitations associated with the available samples. Notably, precisely the aspects that are not feasible (and, in particular, the reasons for their infeasibility) allow one to infer how the samples ought to have been produced in order to avoid these limitations. In a nutshell, only understanding where the data comes from allows for drawing meaninful conclusions from it.

This principle can be neatly illustrated by one of the most popular estimators from sampling theory, namely the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). Assume we know that statistical units y_i from a (finite or infinite) population are included in a sample $y_1, \dots, y_n$ with probabilities π_i each. These probabilites are commonly referred to as selection probabilities, see, e.g., Kauermann and Küchenhoff (2010). Horvitz and Thompson (1952) proved the unbiasedness of a mean estimator $\frac{1}{n}\sum_{i=1}^{n} y_i \pi_i^{-1}$, later named Horvitz-Thompson estimator and generalized to two-step M-estimation correction by Heckman (1979).

As has become evident, it is only an explicit modeling assumption about the sampling process via inclusion probabilities that allows for deriving the Horvitz-Thompson estimator. This principle extends beyond simple first-order inclusion probabilities π_i to second-order probabilities $\pi_{i,j}$ of both y_i and y_j being present in the sample and corresponding design weights (Cochran, 1977; Lumley, 2010). While simple random (*i.i.d.*) sampling yields unbiasedness (e.g., of the sample mean), sampling theory typically considers (arguably more realistic) *complex samples*. This term captures *inter alia* stratification, clustering or multistage selection, all potentially giving rise to unequal probabilities.

For exposition, I use probability–proportional–to–size (*PPS*) sampling as an illustrating example following Deville and Tille (1998); Valliant et al. (2018) and Nalenz et al. (2024, Section 2), while noting in passing that the methodology developed here carries over to a broad class of complex sampling designs. In PPS, an auxiliary variable A with realizations $a_i > 0$ that are available for all statistical units and are proportional to y will be used to estimate first–order inclusion as

$$\hat{\pi}_i \;=\; n \, \frac{a_i}{\sum_{j=1}^{n} a_j} \,, \tag{3.6}$$

so that units with larger a_i are more likely to be included. This allows for Horvitz–Thompson estimation of the population mean (Horvitz and Thompson, 1952), as introduced above. Beyond this simple Horvitz–Thompson estimators, stratified designs often (additionally) use Neyman allocation to optimize precision at fixed size (Horvitz and Thompson, 1952; Neyman, 1934). Further approaches include Taylor linearization or resampling methods like the bootstrap or the jackknife (Rust and Rao, 1996) to improve efficiency (Deville and Särndal, 1992; Särndal et al., 2003).

A large literature studies *selection bias*, long recognized as pervasive in survey and observational settings. In this line of work, selection bias arises whenever the selection rule departs from simple random sampling, whether due to survey-design choices (Kreuter and Valliant, 2007; Kreuter et al., 2010; Kreuter, 2013) or self-selection survey participants (Hidiroglou, 1986), and has been documented and synthesized in classic collections and reviews (Wainer, 2013; Heckman, 1979, 2018; Yerushalmy, 1972), see also Rodemann and Bailie (2025, Section 2). Inference procedures based on inverse-probability weighting like the Horvitz–Thompson estimator (see above) provide a unifying remedy across different sources of selection biases, see, e.g., Lumley et al. (2011). For example, Chauvet (2015) establishes a central limit theorem for the Horvitz–Thompson estimator under multistage sampling with simple random sampling at the first stage and a broad class of second-stage designs via coupling methods. Han and Wellner (2021) provide Glivenko–Cantelli and Donsker theorems for Horvitz–Thompson empirical processes under complex designs (including calibrated weights). Already hinting at Contribution 9 (Nalenz et al., 2024) in Chapter 6, we show how to de-bias regression trees and forests learned from complex samples by building on this bridge between algorithmic self-selection and sample designs from the survey literature. Specifically, we incorporate design weights into splitting/fit criteria in tree induction and use weighted bootstrap resampling in random forests. What is more, inverse probability weighting methods yield consistent estimators even under time-varying confounding in causal applications (Robins et al., 2000; Hernán et al., 2000).

As outlined in Rodemann and Bailie (2025, Section 2), Part III (Chapters 4 and 5 specifically) emphasizes a third driver of distorted selection rules: *algorithmic self-selection*, in which machine-learning procedures (e.g., semi-supervised learning, bandits, boosting, active learning) adaptively decide which units to query, label, or upweight. Many design-based results from sampling theory are agnostic to the source of selection and thus remain informative in such settings. Loosely speaking, the probabilities of sample inclusion matter, not which entity (be it a survey designer of a machine learning algorithm) enforces them.

3.3. Data-Centric Machine Learning

3.3.1. An Attempt at a Definition

Stating the obvious, machine learning and statistics are always data-centric in the sense that they draw conclusions from given observations, i.e., data. However, this literal interpretation of the Latin *datum* as a given (observation) and corresponding focus on models rather than data was exposed as somewhat insufficient for various scientific enterprises by Chapter 2. The main reason is that treating observations as a given implicitly implies a static model-data (*explanans-explanandum*) dichotomy, ignorant of feedback loops, see Chapter 2. This is in line

with early definitions of data-centricity by Ng (2021) that contrasts data-centric with model-centric approaches. In this spirit, I will define data-centric machine learning for the remainder of this dissertation roughly as follows.

The term *data-centric machine learning* shall encapsulate *machine learning methods* that (implicitly) entail a *model* of how data is (repeatedly) generated. The dissertation at hand will specifically focus on methods that additionally entail an explicit and non-trivial account of how this model affects the method's outcome. Crucially, this model can, but does not need to be the primary model of the machine learning method. Previously reviewed methods in statistical experimental design (Section 3.1) mostly directly derive the model of data acquisition from the primary model, while sampling theory (Section 3.2) includes examples like reweighting schemes that aim at approximating *i.i.d.* samples suitable for a wide range of statistical models to be employed on them later. Explicit and non-trivial accounts of how this *model* affects the method's outcome typically come in the form of (additional) methods that quantify the uncertainty of the method's outcome. This puts uncertainty quantification at the center of our attention within data-centric machine learning. A short review of sources of uncertainty in data-centric ML in Section 3.3.3 will set the stage for our contributions to uncertainty quantification in data-centric ML in Part III and IV.

The above Definition is more specific than the very broad Definition of data-centric machine learning *research* as "infrastructure, methods and communities revolving around [...] data (and its immediate representations)" by Oala et al. (2024, Page 1). At the same time, it is more abstract than the one found in Singh (2023), which defines data-centric articifical intelligence and machine learning methods by a common goal, which is to "continuously improve data quality" (Singh, 2023, Page 145). Starting from the less specific term of Artificial Intelligence (AI) rather than machine learning, Zha et al. (2025, Page 4) define data-centric AI as "a framework to develop, iterate, and maintain data for AI systems [...]. Data-Centric AI involves the tasks and methods for building effective training data, designing proper inference data, and maintaining the data." Ng (2025, Main Page) simply defines data-centric AI as the "discipline of systematically engineering the data used to build an AI system", see also Ng (2021).

Roughly leaning on Zha et al. (2025), this dissertation distinguishes training data (Part III) development and testing data (Part IV) development. While training data development aims to "collect and produce rich and high-quality training data to support the training of machine learning models" (Zha et al., 2025, Page 5), testing data development is mainly concerned with designing "novel evaluation sets that can provide more granular insights into the model or trigger a specific capability of the model with engineered data inputs" (Zha et al., 2025, Page 6). In particular, selecting testing datasets for designing benchmark suites is wideley considered a core area within data-centric machine learning (Eyuboglu et al., 2022; Mazumder et al., 2022; Zha et al., 2025; Chen et al., 2023; Huang et al., 2024).[3]

As became apparent, this perspective casts training and testing data "development" as an engineering task, being part of a pipeline designed to boost model performance on specific tasks. At first sight, this lies at odds with statistics, where data is typically construed as a sample from some true, unknown population. In light of the two previous Sections, however, the perspectives

[3] In particular, see Mazumder et al. (2022, Section 2.3) for a benchmark-suite framing of data-centric pipelines; Eyuboglu et al. (2022) for defining a specific benchmark for data-centric AI systems; Zha et al. (2025, Section 6) for a survey chapter cataloguing data-centric benchmarks; Chen et al. (2023, Fig. 1 caption; Section 1) for the DAM benchmark within DataPerf as part of data-centric acquisition; and Huang et al. (2024, Section 3.1, "Benchmark Overview") for a dataset-curation benchmark, which operationalizes data-centric quality control.

from engineering and computer science on the one hand and statistics on the other align more closely, as demonstrated, e.g., by the rich statistical legacy on how to design (i.e., develop) experiments, see Section 3.1.

Pondering upon this relation a little further from the statistical perspective, the question then arises quite naturally whether data-centric machine learning is nothing but old wine (experimental design, see Section 3.1) in new bottles (machine learning). Certainly, the general problem of finding informative samples for a hypothesis or problem at hand in data-centric machine learning is akin to the question answered by statisticians like Kirstine Smith, Sir Ronald A. Fisher or Dennis V. Lindley, see above.

However, I contend that the situation is more intricate than that for two reasons. First, modern works within data-centric machine learning take feedback loops like performative effects (Perdomo et al., 2020; Hardt and Mendler-Dünner, 2025; Perdomo, 2025) in social applications, reciprocity (Rodemann et al., 2024; Rodemann and Bailie, 2025) in adaptive algorithms or self-generative loops (Bertrand et al., 2024; Fu et al., 2025) in computer vision into account, as detailed in Section 2. Second, as has become evident from our little historical detour above, the early bandit literature, in particular Robbins (1952), influenced frequentist sequential experimental design (Chernoff, 1953) and, in particular, was introduced *prior to* pathbreaking statistical work by Chernoff (1959); refer to Sen (1992) for a detailed historical account. The multi-armed bandits literature is widely considered a machine learning branch. So, in some sense, machine learning sometimes also makes up the old wine, not only the shiny new bottles.

3.3.2. A Data-Centric Survey

Existing surveys on data-centric machine learning typically employ a top-down perspective, see, e.g., Singh (2023); Zheng et al. (2023); Guo et al. (2025); Miranda (2021); Kumar et al. (2024); Adeoye et al. (2023); Pan et al. (2022); Zha et al. (2025) and Roscher et al. (2023, Section 3). That is, they first define subfields of data-centric ML into which they then categorize existing works. In what follows, I will employ a bottom-up approach instead, which—not without a dash of irony—might be dubbed "data-centric" itself.

After having scraped 1000 recent papers on data-centric machine learning (exact search string below) from Semantic Scholar (Kinney et al., 2023) through a privately requested research-purpose Application Programming Interface (API), I employed k-means clustering (MacQueen, 1967; Lloyd, 1982) on vector embeddings of these papers' titles and abstracts that takes the citation graph into account. This approach to surveying the emerging field of data-centric machine learning aims at (without fully achieving[4]) an unbiased perspective on the field.[5]

More specifically, I retrieved the 1000 most recent academic articles fitting the search query "data-centric machine learning" from the Semantic Scholar database. Among these 1000 papers, only articles with non-trivial abstracts (defined as having more or equal than 80 characters), unique Digital Object Identifiers (DOIs) and non-missing publication date are kept. This leaves

[4] Think of the inductive bias in k-means clustering or in the model generating the vector embeddings. As has become evident in Section 3.3.1, any definition entails a certain inductive bias on the literature—and so does a data-centric one, as it presupposes implicit assumptions that manifest in the model used to generate the vector embeddings.

[5] Jupyter notebooks to reproduce the data-centric survey are available at `https://github.com/rodemann/data-centric-ml-specter-embeddings-clustering`.

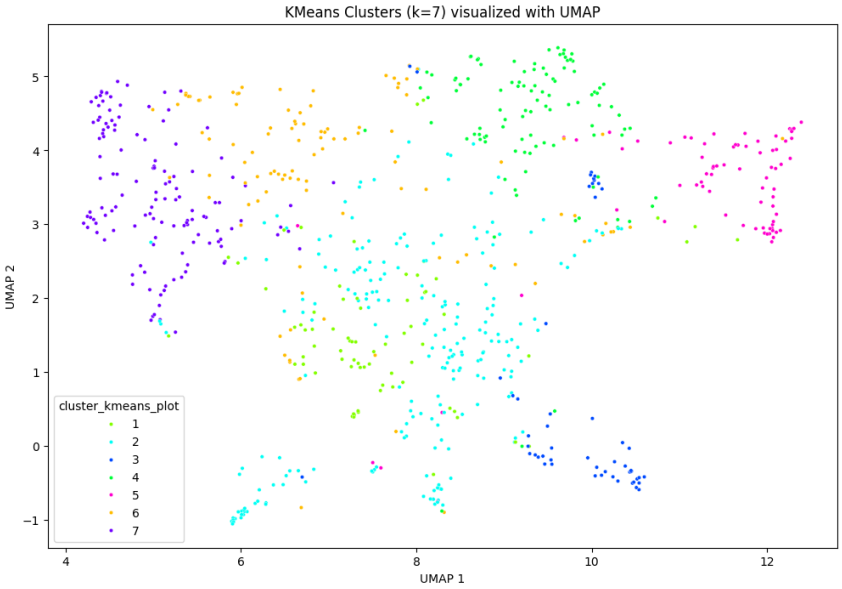

Figure 3.1.: Visualization of clusters of papers on data-centric machine learning found by the unsupervised k-means clustering algorithm ($k = 7$) in two-dimensional UMAP embedding.

Cluster	Top keywords
1	`language`, `text`, `translation`, `sentiment`, `graph`, `medium`, `emotion`, `ood`, `tabular`, `email`
2	`material`, `explanation`, `healthcare`, `analytics`, `objectcentric`, `graph`, `privacy`, `business`, `object`, `datadriven`
3	`memory`, `energy`, `quantum`, `hardware`, `graph`, `device`, `cpu`, `movement`, `computation`, `virtual`
4	`iot`, `fl`, `wireless`, `device`, `federated`, `traffic`, `privacy`, `attack`, `metaverse`, `networking`
5	`weather`, `agriculture`, `air`, `climate`, `urban`, `mediumrange`, `wind`, `ensemble`, `hyperspectral`, `rmse`
6	`health`, `healthcare`, `disease`, `sensor`, `patient`, `wearable`, `heart`, `care`, `anomaly`, `diabetes`
7	`patient`, `clinical`, `disease`, `auc`, `mortality`, `medical`, `cancer`, `risk`, `mri`, `healthcare`

Table 3.1.: Top keywords per cluster allow for a contentual interpretation of the clusters found by k-means clustering.

us with 758 unique papers with non-trivial abstracts, 94 % of which appeared after January 1, 2018.

I then used SPECTER (Cohan et al., 2020) to generate numeric vector representations of each of these title-abstract pairs, so-called embeddings. SPECTER is a transformer-based deep learning model which is pre-trained on citation graphs. The input of the SPECTER model are the token strings of the 758 title-abstract pairs. Its output is a 768-dimensional vector for each title-abstract pair. These embeddings capture topic similarity. That is, closer vectors mean more similar papers. Since the deep learning model is pretrained on citation graphs, these topic similarities are not only based on textual similarity of the titles and abstracts of the papers, but also on their relations through citations.

After standardizing the embeddings, I employ classic k-means clustering (MacQueen, 1967; Lloyd, 1982) on these paper embeddings. The hyperparameter k (number of clusters) was selected from

$k \in \{6, \ldots, 21\}$ to maximize the average silhouette value by simple grid-based hyperparameter-tuning. The silhouette value is based on comparing distances within clusters to the smallest distances to other clusters. As such, the silhouette value is a measure of how similar a statistical unit (here: paper) is to its own cluster compared to other clusters (Rousseeuw, 1987). In other words, k was tuned to achieve clusters as distinct as possible.

Figure 3.1 illustrates the so-found clusters in their two-dimensional UMAP (Uniform manifold approximation and projection) (McInnes et al., 2018) embedding. Note that these two dimensions have no inherent interpretation. The visualization thus only aims to highlight the *relative* positions of clusters to other clusters and positions of papers within clusters. Clusters 5, 3, 7 and 4 appear to be the most salient (in this order) clusters, while clusters 1, 2 and 6 are less clearly separated.

Seeking a contentual understanding of the so-found clusters, I compare the relative frequencies of keywords of papers within clusters across clusters. Due to missing keywords, I employ TF-IDF (term frequency–inverse document frequency) (Jones, 1972), a common method in information retrieval, to get keyword proxies. Table 3.1 highlights the most frequent keywords per cluster. Evidently, the most distinct clusters 5, 3, 7 and 4 and cluster 1 can be clearly related to application domains of data-centric machine learning: cluster 5 comprises meteorology, cluster 3 summarizes hardware systems and computer architecture, cluster 7 is the medial domain, cluster 4 contains distributed systems and connected mobility and cluster 1 clearly contains the natural language generation and processing domain. The less distinct clusters 2 and 6 mix (public) health with (bio)medial technology.

Summing up, this little data-centric survey provides two main insights into the growing body of data-centric machine learning literature. First and foremost, the identified clusters primarily and almost exclusively relate to *applications* of data-centric machine learning. This finding emphasizes the engineering perspective deeply interwoven into data-centric machine learning, as mentioned above in Section 3.3.1. It seems like, firstly, various applications call for efficient sample designs or data subset selection, data pruning, data acquisition and the like and, secondly, several of these calls are answered by ad-hoc engineering of data pre-processing and acquisition pipelines. Might this pragmatic approach, one wonders, be due to inaccessibility of existing methodology in experimental design (Section 3.1) and sampling theory (Section 3.2), sheer ignorance about the latter or due to a void of appropriate methodology? I leave this to further research. Irrespective of the latter, this overall finding of the field (or, at least, the terminology "data-centric machine learning") being mainly application-driven comes with some surprise, as the newly founded journal on data-centric machine learning research (DMLR) seem to be rather methodology-focused, see, e.g., the inaugural volume of DMLR (Ardalani et al., 2024). This also applies to the workshops on data-centric machine learning at NeurIPS 2021 and ICML 2023 as well as other publications at machine learning conferences like Seedat et al. (2022a,b); Oala et al. (2024).

A second insight from this unsupervised literature review is the following. The field of data-centric machine learning, as a whole, still seems to be at an early, developing stage. Even the most salient and distinct clusters, see Figure 3.1 and Table 3.1 again, have several outliers and partly overlap in the latent UMAP-space. This insight, arguably, is less surprising than the first one. "Forming and storming" phases (Tuckman, 1965; Tuckman and Jensen, 1977) can be seen as natural for any young, emerging discipline or field.

3.3.3. Sources of Uncertainty in Data-Centric Machine Learning

Strongly relying on Hüllermeier and Waegeman (2021), three (potentially non-exhaustive) sources of predictive[6] uncertainty in machine learning are very briefly reviewed: Aleatoric uncertainty as well as epistemic model uncertainty and epistemic approximation uncertainty. I will focus on uncertainty *quantification*, deliberately setting aside other, non-numeric representations of uncertainty, see Kirchhof et al. (2025a) for a modern example. Moreover, I explain the consequences of giving up the strong data-model dichotomy (see Part II) on this popular categorization and discuss another categorization of uncertainties that might prove more useful in non-static data lifecycles as discussed in Chapter 2.

Predictive uncertainty can be organized hierarchically. *Aleatoric* uncertainty refers to irreducible randomness of the data-generating process (sensor noise or label ambiguity in the *ontic* (Couso and Dubois, 2014) sense) and therefore persists asymptotically. Typically, it is modeled in the conditional predictive distribution, e.g., via heteroscedastic likelihoods and mixture models (Hüllermeier and Waegeman, 2021; Kiureghian and Ditlevsen, 2009; Guo et al., 2017). As an inherent property of a data generating process, aleatoric uncertainty is not *directly* concerned with choices made in a data lifecycle (Figure 1.1) before or after collecting data. Invoking Fisher Box (1978) again, see Section 3.1, aleatoric uncertainty persists in nature's answer to our questions, no matter how skillful we interrogate. However, the way we model the aleatoric uncertainty that is present in the observations involves modeling assumptions (there is no free statistical lunch) that are subject to epistemic uncertainty, see below.

Epistemic uncertainty, in contrast, captures lack of knowledge and can, in principle, be reduced by more informative data or better inductive biases, i.e., model choices. Within epistemic uncertainty, Hüllermeier and Waegeman (2021) distinguish *model uncertainty* from *approximation uncertainty*. Model uncertainty (often also referred to model imprecision and related to ambiguity (Ellsberg, 1961)) refers to uncertainty over hypotheses within or across model classes, while *approximation uncertainty* (or, statistically speaking, sampling uncertainty) addresses the classical uncertainty in statistics that stems from the fact that the available sample is smaller in quantity than the population we make statements about. Hüllermeier and Waegeman (2021) emphasize that these two sources of epistemic uncertainty are often conflated in uncertainty quantification in deep learning practice like via Monte Carlo dropout or deep ensembles: variability across stochastic forward passes in neural networks, for instance, can reflect both model uncertainty or limited samples (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Ovadia et al., 2019; Blei et al., 2017; Welling and Teh, 2011). A very natural way to represent epistemic model uncertainty is via imprecise probability models, as briefly surveyed in Chapter 1.

A second, structural source of model uncertainty (which is sometimes swept under the carpet in both statistics and machine learning) is the underlying order structures on the outcome space and whether they allow a specification and identification of a single, cardinal loss function. (Jansen, 2025; Jansen et al., 2023b, 2024) Here, *model uncertainty* stems from weakly structured order information, leading to a non-singleton *set* of candidate scales that are compatible with the structure on the codomain of the random variables describing estimators or learners, respectively. This type of uncertainty motivates contributions 10 and 11 in this dissertation, see Chapter 7 in Part IV.

[6] In Part III, especially in contributions 2, 3, 4, and 5, we also discuss how predictive uncertainty trickles down to uncertainty over acts (like proposals in Bayesian optimization or pseudo-label selection in semi-supervised learning) being taken.

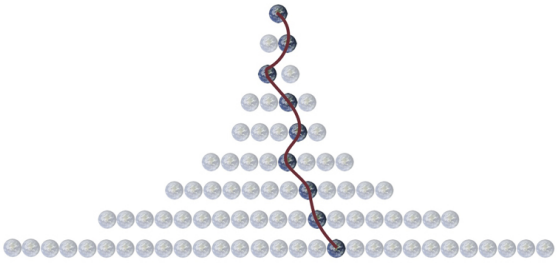

Figure 3.2.: "Parallel universes" illustrating the difference of ensemble averages (horizontal) and time averages (vertical along realized red path). Image credits: https://neurosparkle.com/ergodicity/, originally: London Mathematical Laboratory, www.lml.org.uk/research/economics (last accessed September 29 2019 (sic!)).

While the aleatoric-epistemic distinction has attracted much attention in machine learning recently, little attention is commonly paid to a related, yet different dimension along which uncertainties could be handily categorized: ergodicity. It refers to stochastic processes where ensemble averages (across statistical units at fixed time) equal time averages (across time point for a fixed statistical unit), see Figure 3.2 for an illustration.

A little more formally, consider a standard probability space $(\Omega, \mathcal{B}, \mathbb{P})$ equipped with a shift operator $T : \Omega \to \Omega$ such that $\mathbb{P}(T^{-1}(A)) = \mathbb{P}(A)$ for all $A \in \mathcal{B}$ (measure-preserving T). One then calls $(\Omega, \mathcal{B}, \mathbb{P}, T)$ ergodic if every T-invariant set has probability 0 or 1. In other words, $T^{-1}(A) = A \implies \mathbb{P}(A) \in \{0, 1\}$. In this case, it holds (per Birkhoff's acclaimed ergodic theorem, Birkhoff (1931)) for any measurable random variable X on $(\Omega, \mathcal{B}, \mathbb{P})$ that

$$\frac{1}{T}\sum_{t=1}^{T} X(T^t(\omega)) \xrightarrow[T\to\infty]{\text{a.s.}} \mathbb{E}_{\mathbb{P}}[g(X)], \tag{3.7}$$

see, e.g., Walters (1982); Peters (2019). In other words, time averages almost surely converge to ensemble population averages and the estimation error decreases at the usual statistical rate of $1/\sqrt{T}$ (Walters, 1982). *Ergodic uncertainties* are those arising from such an ergodic data-generating process. In other words, ergodic uncertainties can be averaged away along a single long trajectory (Kiureghian and Ditlevsen, 2009; Peters, 2019).

For instance, measurement noise or *i.i.d.* draws with a fixed distribution could be regarded ergodic. On the other hand, *non-ergodic* uncertainties (e.g., regime shifts (Hamilton, 1989; Bai and Perron, 1998), multiplicative dynamics (Peters and Adamou, 2018), structural breakpoints (Dette and Wied, 2016)) do not average out along one path; time averages need not match ensemble averages, so they cannot be "learned away" iteratively in sequential setups as introduced in Section 2.2 that underlie many of the contributions in this dissertation. The question naturally arises whether ergodicity offers a more fruitful dimension for differentiation types of uncertainties in non-static and data-centric machine learning methods like reciprocal learning or performative prediction. I shall leave this question to future work, not without loosely pointing to Wheeler (2025), and return to the established distinction between aleatoric and epistemic uncertainty.

As recognized by the literature, it remains unclear whether this conceptual epistemic-aleatoric distinction actually offers structural insights into modern agentic models (Kirchhof et al., 2025b),

is aligned with empirical loss minimization in machine learning (Bengs et al., 2022, 2023; Jiménez et al., 2025), captures real-world sources of uncertainty (Gruber et al., 2025), is relevant for decisions and acts (Smith et al., 2025), nor whether it adds up to total uncertainty (Wimmer et al., 2023; Mucsányi et al., 2024).

In a nutshell, one of the main drawbacks of this aleatoric-epistemic dichotomy is the fact that we use models (that come with epistemic uncertainty) to measure aleatoric uncertainty. The dissertation at hand identifies a similar drawback of the model-approximation distinction within epistemic uncertainty: As Section 2 revealed, accounting for cyclic model-data dependencies calls this distinction within epistemic uncertainty into question, since not only the approximation (sampling) uncertainty in previously collected data can inform model selection and thus model uncertainty, but also vice versa, as reciprocity (Part III of this dissertation) and performativity (Morgenstern, 1928; Perdomo et al., 2020; Hardt and Mendler-Dünner, 2025) lead to models changing the sample and the populations from where the sample is drawn, respectively.

Due to these drawbacks, the main emphasis in the remainder of this dissertation will thus not lie on trying to disentangle the inherently interwoven concept of uncertainty. Instead, this dissertation will focus on "measuring, controlling and communicating uncertainty" (American Statistical Association, 2012; Davidian and Louis, 2012), or, in other words: Statistics.[7]

[7] Indeed, the American Statistical Association (ASA) defines Statistics as "the science of learning from data, and of measuring, controlling and communicating uncertainty" (American Statistical Association, 2012; Davidian and Louis, 2012), where emphasis is put on the second part of the definition, abstaining from a critical discussion of the first part's implicit treatment of data as a given ("learning *from* data").

Part III.

Statistical Perspectives on Training Data Selection

4. Reciprocal Learning

This is the first of five chapters (Chapters 4 through 8) summarizing the contributing material of this dissertation. The contributions are clustered into training data selection (Contributions 1 through 9 in Part III) and testing data selection (Contributions 10 through 14 in Part IV).

The classic conceptualization of (parametric) statistics and machine learning tries to find optimal parameters of a model in light of observed data, allowing for generalization or inference to the population, see Figure 1.2 in Chapter 1. In Contribution 1 (Rodemann et al., 2024), however, we show that the relationship between data and parameters is in fact *reciprocal* in several well-established learning paradigms. Examples comprise Bayesian optimization (Section 4.3), self-training in semi-supervised learning (Section 4.4), superset learning (Section 4.5), bandits, boosting and active learning.

This gives rise to a new learning paradigm: Reciprocal Learning, a unifying framework, allowing for a principled analysis of all these methods. After a model fit to the training data, reciprocal learning algorithms alter the latter in a way *that depends* on the fit. This dependence can have various facets, ranging from predicting labels (self-training) over taking actions (bandits) to querying an oracle (active learning), all based on the current model fit. A simple example of reciprocal learning is self-training in semi-supervised learning, in which part of the training data is unlabeled—for example, images without captions. The algorithm labels this data itself and ultimately learns from these "pseudo-labels" by iteratively adding pseudo-labeled variants of unlabeled data to the labeled training data. The pseudo-labels are predicted by the current model, and thus depend on the parameters learned by the model from the labeled data in the first place.

In all these methods summarized by *reciprocal* learning, parameters are not only learned from data but also data is added or removed based on parameters. In particular, we demonstrate that this reciprocity corresponds to two interdependent decision problems and explicitly study how learned parameters affect the subsequent training data. In this way, we present a unifying perspective of many data-centric learning paradigms, which provides the conceptual hinge for the entire Chapter 4, including all its Sections 4.1 through 4.5.

4.1. Reciprocal Learning (Contribution 1)

Contribution 1

Julian Rodemann, Christoph Jansen, and Georg Schollmeyer (2024). "Reciprocal Learning." In: *Neural Information Processing Systems (NeurIPS)*, Vol. 37, pages 1686–1724.

Slightly increasing the level of formality, we call a machine learning algorithm *reciprocal* if it performs iterative ERM on training data that depends on the previous ERM. This dependence can be induced by any kind of data collection, removal, or generation that is affected by the model fit. In particular, it can be stochastic (think of Thompson-sampling in multi-armed bandits) as well as deterministic in nature (think of maximizing a confidence score in self-training). For ease of exposition, we restrict ourselves to *parametric* reciprocal learning. Thus, consider a parameter space Θ of parameter vectors $\theta \in \Theta$.

Preparing a formal definition of *reciprocal learning*, consider further a set $\mathcal{P}$ of probability measures P defined on Borel σ-algebras with finite second moments on a bounded subset of the Euclidean space $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$ with $\mathcal{Y} \subset \mathbb{R}^{d_y}$, $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $d = d_y + d_x$. As is customary, we denote labels by $y \in \mathcal{Y}$, features by $x \in \mathcal{X}$ and random variables on these spaces by Y and X, respectively. For a probability measure $P \in \mathcal{P}$ and a parameter vector $\theta \in \Theta$, define the **risk** of θ under P as

$$\mathscr{R}(P, \theta) := \mathbb{E}_P \left[\ell(Y, X, \theta)\right] = \int_{\mathcal{Z}} \ell(y, x, \theta) \, \mathrm{d}P(z), \tag{4.1}$$

where $\ell : \mathcal{Y} \times \mathcal{X} \times \Theta \to \mathbb{R}$ is a loss function and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. In words, the risk $\mathscr{R}(P, \theta)$ quantifies the expected loss incurred by using θ when the data is drawn from some P.

Given a finite sample $(x_1, y_1), \ldots, (x_n, y_n)$ with empirical law $\hat{\mathbb{P}}_t$, the corresponding **empirical risk minimizer** at iteration $t \in \{1, \ldots, T\}$ is denoted by $\hat{\theta}_t$ and defined via

$$\hat{\theta}_t \in \arg\min_{\theta \in \Theta} \mathscr{R}\Big(\hat{\mathbb{P}}_t, \theta\Big) = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i, \theta). \tag{4.2}$$

We assume there is a unique empirical risk minimizer in what follows. *Reciprocal learning* can be understood as a special variant of multi-shot ERM. Specifically, it refers to any iterative process in which empirical risk minimization is performed repeatedly, where the sample used at each iteration depends both on the previous empirical risk minimizer and on the previous sample.

To formalize this sample adaptation mechanism, define the *sample adaptation function* f_s as

$$f_s : \begin{cases} \Theta \times \mathcal{P} \to \mathcal{P}, \\ (\hat{\theta}_t, \hat{\mathbb{P}}_t) \mapsto \hat{\mathbb{P}}_{t+1}. \end{cases} \tag{4.3}$$

In plain terms, f_s takes the current model parameters together with the current empirical distribution and produces a new empirical distribution that will serve as the basis for the next iteration. For a concrete example of f_s, the reader is referred to Section 2.4 in Contribution 1.

Using the above components, we can finally define reciprocal learning as

$$R : \begin{cases} \Theta \times \mathcal{P} \to \Theta \times \mathcal{P}, \\ (\hat{\theta}_t, \hat{\mathbb{P}}_t) \mapsto (\hat{\theta}_{t+1}, \hat{\mathbb{P}}_{t+1}). \end{cases} \tag{4.4}$$

The update proceeds in two stages. First, the empirical distribution is adapted according to $\hat{\mathbb{P}}_{t+1} = f_s(\hat{\theta}_t, \hat{\mathbb{P}}_t)$, and then the next model is obtained by empirical risk minimization on the adapted sample: $\hat{\theta}_{t+1} \in \arg\min_{\theta \in \Theta} \mathscr{R}(\hat{\mathbb{P}}_{t+1}, \theta)$. For details of this Definition and how the examples of active learning, self-training in semi-supervised learning, bandits, Bayesian optimization and superset learning fit into this framework, we refer to Section 2 of Contribution 1.

As outlined in Section 2.2, reciprocal learning is a form of sequential decision-making. In the first step, a parameter $\hat{\theta}_t$ is obtained via empirical risk minimization, which can be cast as solving

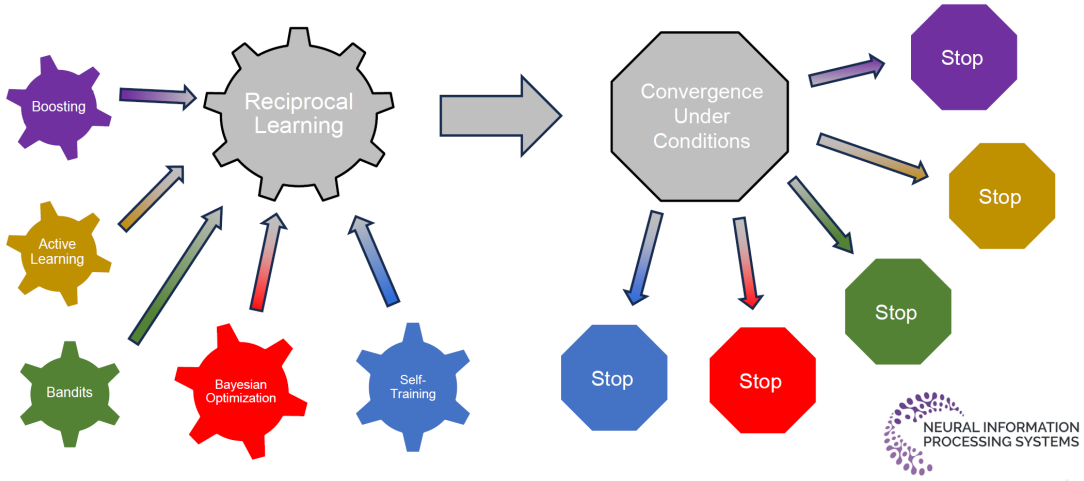

Figure 4.1.: Outline of Contribution 1: Several learning paradigms are unified in the eponymous framework of *reciprocal learning*, which allows for a general convergence analysis subsuming examples of boosting, active learning, bandits, Bayesian optimization, and self-training, to name only few. After convergence theorems are established in this abstract setting, Corollaries allow to interpret these convergence results in the concrete setting of all the examples, giving rise to *stopping* criteria. Figure replicated from poster presented at NeurIPS 2024, available at `https://neurips.cc/virtual/2024/poster/93740` (not included in Rodemann et al. (2024))

a decision problem $(\mathbb{A}, \Theta, \ell)$ characterized by the triple $(\mathcal{D}, \Theta, \ell)$, see Section 2.2. Here Θ is the set of states of nature, the act space is $\mathbb{A} = \mathcal{D}$ (the candidate parameter estimators), and $\ell : \mathbb{A} \times \Theta \to \mathbb{R}$ is a loss function, in direct analogy to classical statistical decision theory (Berger, 1985), see Section 2.2. In the second step, features $x_t \in \mathcal{X}$ are selected and data points (x_t, y_t) are added to or removed from the training set, inducing an updated empirical distribution $\mathbb{P}_{t+1}$; the label y_t may be predicted (self-training), queried (active learning), or observed through interaction (bandits). The choice of x_t itself arises from a second decision problem $(\mathcal{X}, \Theta, \ell_{\hat{\theta}_t})$, where the loss $\ell_{\hat{\theta}_t}$ depends on the previously obtained solution $\hat{\theta}_t$, and the act space is now the feature space, $\mathbb{A} = \mathcal{X}$. For details, refer to Illustration 1 in Section 2.2 in Contribution 1. Notably, this decision-theoretic embedding will be instrumental in Contributions 4, 5 and 6, see Section 4.4.

The embedding of Bayesian optimization (Section 4.3), self-training in semi-supervised learning (Section 4.4), superset learning (Section 4.5), bandits, boosting and active learning into the framework of reciprocal learning allows for a general convergence analysis of all these methods, see Figure 4.1. In particular, we establish via the Banach fixed-point theorem that reciprocal learning algorithms converge at a linear rate if the sample adaptation is sufficiently Lipschitz-continuous with respect to the $L2$-norm on Θ and the Wasserstein-1-distance on $\mathcal{P}$ (Theorem 3 in Contribution 1). The Lipschitz-continuity of the sample adaptation, in turn, follows from probabilistic predictions and regularized or randomized data selection, if the loss function is strongly convex and continuously differentiable in parameters and features (Theorems 1 and 2 in Contribution 1). Moreover, we show that the so-obtained convergent solution is close to the optimal one with respect to the $L2$-norm on Θ (Theorem 4 in Contribution 1).

As visualized in Figure 4.1, we can then relate these general results back to individual special cases of reciprocal learning, see Corollaries 1 through 3 in Contribution 1. For instance, we show that multi-armed bandits with stochastic data selection strategies like the popular Thompson

sampling are guaranteed to converge if the loss function is strongly convex and continuously differentiable in parameters and features, while deterministic data selection strategies might diverge. This directly follows from the requirement of randomized data selection in Theorem 3. In a similar manner, our results allow for distinguishing between soft oracles (which provide probabilistic predictions) and hard oracles (which provide hard labels) in active learning. This is due to the requirement of probabilistic predictions in Theorem 3. Other Corollaries in the same spirit can be found in Contribution 1.

4.2. Outlook and Perspectives

The convergence results in Contribution 1 require some conditions on the loss function. Weakening those to lift the convergence guarantees to an even broader range of machine learning algorithms is a promising line of future work. In particular, the requirement of strongly convex loss prevents the application of the results to multi-layer perceptrons ("neural networks"). Finding convergence guarantees without this requirement is thus an important avenue for future studies.

Besides, a detailed investigation of the generalization performance of reciprocal learning algorithms appears worthy of closer consideration. Contribution 1 only has results on whether the adapted sample is efficient with respect to the *training* task. Will it also turn out to be useful when tested on unseen data? Empirical evidence in active learning (Settles, 2010) and self-training in semi-supervised learning (Van Engelen and Hoos, 2019) suggests so, but can we establish general theoretical results? Contribution 8 will answer in the affirmative.

Another promising line of future work is to identify (approximately) optimal samples. Our convergence results currently only come with optimality guarantees with respect to the parameters. One way to reason about the convergent *sample*'s approximate optimality could be to relate it to the approximately optimal parameters by techniques from data attribution. In fact, it is easy to see from a finite sample analysis that a single optimal empirical distribution might not be identifiable, but a set of probability distributions in a Wasserstein ball around the convergent sample. Theorem 4 in Contribution 1 gives a similar result on the optimal parameters. In particular, it states that reciprocal learning's convergent solution delivers an ε-ball around the optimal solution. In other words, reciprocal learning algorithms like active learning or self-training inherently produce imprecise (i.e., set-valued) solutions, offering exciting avenues for future work, exploiting the rich literature on imprecise probabilities as outlined in Chapter 1. For instance, one could study reciprocal learning's solutions through the lens of cluster points of relative frequencies (Walley and Fine, 1982; Fröhlich et al., 2023) or algorithmic randomness (see e.g., Persiau et al. (2022)).

Moving beyond existing algorithms, the unified framework of reciprocal learning offers exciting avenues for designing new methods. One promising line of research is to derive methodological innovations from our theoretical insights on convergence of reciprocal learning algorithms. Specifically, the sufficient conditions for convergence in Contribution 1 pave the way for a theory-informed design of novel algorithms. They may serve as design principles for self-training, active learning, or bandit algorithms that shall converge. In particular, our results emphasize the importance of regularization of *both* parameters and data for convergence.

Parameter regularization is well-studied in statistics and machine learning and has been heavily applied, see e.g., Vapnik (1998); Hastie et al. (2009). The concept of data regularization might bear similar practical potential. Parameter regularization entails adding a norm on Θ to the empirical

risk to smooth out the effect of the sample on the empirical risk minimizer. Symmetrically, data regularization adds a norm on $\mathcal{X}$ to smooth out the effect of the model on the selected data. We refer the interested reader to Definition 3 and Figure 2 in Contribution 1. For a tangible example, reconsider reciprocal learning's special case of self-training in semi-supervised learning from above. Here, data regularization translates to attenuating the impact of the predicted pseudo-labels on the inclusion of unlabeled data, in order to guarantee convergence. This form of pseudo-label regularization bears considerable practical potential: With the majority of training data scraped from the web being unlabeled, such methods can be pivotal in reliable modern machine learning with applications ranging from language models to computer vision.

In a similar spirit, the decision-theoretic embedding of reciprocal learning offers possibilities for increasing these methods' sample efficiency through cost-effective data selection. The key is to cast reciprocal learning as a two-player game instead of a sequential one-player decision problem. It is then relatively straightforward to see that the convergent solutions of reciprocal learning in Contribution 1 correspond to subgame perfect Nash equilibria in Stackelberg games, see, e.g., Nisan et al. (2007). Instead of one agent solving a bi-variate minimization problem, a Stackelberg game casts reciprocal learning as an iteration of empirical risk minimization (leader) and sample adaptation (follower), allowing to incorporate a second—potentially conflicting—loss/utility for the follower. For instance, this loss/utility can reflect the cost of data acquisition. Leaning on the literature on market design (e.g., Werner et al. (2024)), this embedding could allow for modeling more nuanced notion of sample efficiency, i.e., the trade-off between predictive accuracy and data acquisition.

On a much more speculative level, one could further ponder on the general interpretation and the implications of reciprocity in learning, roughly inspired by recent interest in "learning to learn" (Zhao et al., 2025; Shafayat et al., 2025; Zhou et al., 2025). Discovering reciprocal learning mechanisms in artificial intelligence quite naturally gives rise to the question whether such mechanisms are also prevalent in natural intelligent systems like animals. In psychology, tangential work has questioned educational supervision as a one-way learning process and suggested reciprocity between supervisors and supervisees (Carrington, 2004). Beyond education, studying reciprocity between tasks (data) and learning outcomes (parameters) appears a promising avenue for further investigation, given the fact that reciprocal learning has proven successful in artificial intelligence. Has evolution favored task-efficient (data-efficient) learning? If so, has it given rise to reciprocal learning or other concurring learning paradigms like meta-learning (Vanschoren, 2019; Wang, 2021; Vettoruzzo et al., 2024), where tasks (data) are not learned from learning outcomes (parameters), but from other tasks (meta-data)? These questions call for interdisciplinary research between the fields of artificial intelligence on the one hand and brain science and neurology on the other hand.

4.3. Bayesian Optimization

From all special cases of reciprocal learning (like, e.g., active learning, bandits or self-training, see above), Bayesian optimization certainly stands out. This is due to the simple fact that its overall goal is optimization rather than learning. Without loss of generality, we shall consider a minimization problem $\min_{x\in\mathcal{X}} \Psi(x)$, where the target function $\Psi : \mathcal{X} \to \mathbb{R}, \mathcal{X} \subseteq \mathbb{R}^d$ is analytically unknown. Nevertheless, as illustrated by Algorithm 1, Bayesian optimization follows the very principles of reciprocal learning laid out above: After an initial fit to the

sample $\mathscr{D} = \{(x_0^{(i)}, \Psi_0^{(i)})\}_{i=1,\dots,n_{init}}$, the surrogate model is refitted on an enhanced sample $\mathscr{D} \leftarrow \mathscr{D} \cup (x_t, \Psi(x_t))$ with $t \in \{1, \dots, T\}$, where x_t maximizes an acquisition function, which depends on the surrogate model's predictions. This latter dependence exemplifies a sample adaptation function f_s, see above: the empirical distribution $\hat{\mathbb{P}}_{t+1}$ depends on its predecessor $\hat{\mathbb{P}}_t$ and on the previous surrogate model $\hat{\theta}_t$ through the acquisition function.

Algorithm 1 Bayesian Optimization (according to Rodemann and Augustin (2024))

Require: initial sample $\mathscr{D} = \{(x^{(i)}, \Psi^{(i)})\}_{i=1,\dots,n_{init}}$

1: **while** termination condition is not fulfilled **do**
2: fit a Surrogate Model $\hat{\theta}_t$ on the observed data $\mathscr{D}$
3: propose x_t that minimizes the Acquisition Function $AF(\hat{\theta}_t(x))$
4: evaluate Ψ on x_t and update $\mathscr{D} \leftarrow \mathscr{D} \cup (x_t, \Psi^(x_t))$
5: **end while**

Bayesian optimization is used in a myriad of different applications and has also gained popularity as hyperparameter tuning strategy in machine learning in recent years. The principled idea to use Bayes criteria for optimizing unknown functions dates back to pioneering work by Jonas Močkus in the 1970s in Lithuania (at that time, Soviet Union), see Močkus (1975); Močkus et al. (1978); Močkus (1989). Its modern day applications, however, are mainly due to Jones et al. (1998), who transferred Močkus' decision-theoretic principles to real-world optimization of black box functions. The latter are systems that produce one or more outputs from several input parameters. They are typically expensive to evaluate and lack an analytical description of their internal workings, as the name suggests. For instance, such problems occur when producing materials through an experimental design that consists of various input parameters, see the illustrative application in Contribution 2 (Section 4.3.1). The output is represented by one or more measurements regarding the quality of the fabricated material. Personalizing wearable electronic devices like exosuits constitutes another example, see Contribution 3 (Section 4.3.2). Here, the aim is to find a suitable configuration of the device, personalized to the individual user. The target function's output is either a human user feedback or a metabolic measurement as a proxy for physical fitness.

4.3.1. Imprecise Bayesian Optimization (Contribution 2)

Contribution 2

Julian Rodemann and Thomas Augustin (2024). "Imprecise Bayesian Optimization." In: *Knowledge-Based Systems* 300:112186.

As pondered upon in Chapter 1 and illustrated by Figure 1.1, data life cycles (Yu and Barter, 2024) are subject to various different sources of uncertainty, ranging from aleatoric (i.e. irreducible) uncertainty in the measurements to epistemic (i.e. reducible) uncertainty in the initial sample or the surrogate model, see also Section 3.3.3. Bayesian optimization is no exception and prior work has already addressed some of these uncertainties. For instance, Makarova et al. (2021) model aleatoric uncertainty by taking into account heteroscedastic noise of the target function, which affects the optimization rationale of risk-averse decision-makers. In Contribution 2 (Rodemann and Augustin, 2024), we address model uncertainty, which constitutes a type of

epistemic uncertainty. In line with Hüllermeier and Waegeman (2021, Section 3.4), see also Hüllermeier et al. (2022) and Section 3.3.3 above, we distinguish two types of epistemic uncertainty: approximation (also called statistical) uncertainty, which stems from the simple fact that the sample is a proper (strict) subset of the population, and model uncertainty (or imprecision), which results from the choice of the model and can be understood as the degree of variation in the conclusion one would obtain if one had specified a different model in the first place.

In particular, we turn to Gaussian Process (GP) models, which are arguably the most common surrogate models in Bayesian optimization for continuous input space $\mathcal{X}$, also referred to as feature or parameter space.[1] The rough idea of GPs is to specify a Gaussian process *a priori* (a Gaussian prior distribution), then observe data (operationalized by a likelihood function) and eventually receive a posterior distribution over functions, from which inference is drawn, usually by mean and variance prediction. More specifically, a function $f(x)$ is said to be generated by a *Gaussian process* $\mathcal{GP}\left(m(x), k(x, x')\right)$ if for any finite vector of data $(x_1, ..., x_n)$, the associated vector of function values $f = (f(x_1), ..., f(x_n))$ has a multivariate Gaussian distribution: $f \sim \mathcal{N}\left(\mu, \Sigma\right)$, where $\mu = m(x_1, ..., x_n)$ is the mean vector and $\Sigma = k((x_1, ..., x_n), (x_1, ..., x_n)')$ the covariance matrix, see, e.g., Rasmussen (2003); Williams and Rasmussen (2006).

The epistemic model uncertainty—in the sense of Hüllermeier and Waegeman (2021, Section 3.4)—in Gaussian processes can be pinned down to choosing a prior, in line with classical Bayesian sensitivity analysis (Weiss, 1996; Augustin et al., 2014b). As recognized by prior work (e.g., Malkomes and Garnett (2018); Schmidt et al. (2008); Lu et al. (2023)) this is particularly relevant in Bayesian optimization, since the latter is typically applied to small sample regimes.[2] Classical Bayesian sensitivity analyses of Gaussian processes (see, e.g., Schmidt et al. (2008)) focus on the effect of the prior choice on the posterior inference. In Bayesian optimization, however, posterior inference is just a means to an end: It is used to propose x_t via the acquisition function. The pivotal quantity is not the posterior, but the cumulative regret $\sum_{t=1}^{T} r_t$, where $r_t = \Psi(x_t) - \min_{x \in \mathcal{X}} \Psi(x)$ is the instantaneous regret in iteration $t \in \{1, \ldots, T\}$ with $\Psi(x_t)$ the target value of proposal x_t in iteration t and $\min_{x \in \mathcal{X}} \Psi(x)$ the universal optimum.

This is why, in Contribution 2, we first conduct an extensive simulation study to investigative which part of the GP prior has the biggest impact on the cumulative regret of Bayesian optimization. We compare the functional form of the mean vector $m(x)$, the parameters of the mean vector $m(x)$, the functional form of the kernel $k(x, x')$ and the parameters of the kernel $k(x, x')$. Our simulation study reveals the mean parameters to have the highest impact, see Section 3 in Contribution 2.

In response to this key finding, we investigate this part of the GP prior specification in greater detail. In particular, we prove that misspecifcation of the GP prior mean parameters make BO's regret bounds grow linearly instead of (under correct specification) sublinearly. Technically, we rely on the concept of information gain from the bandit literature (Dani et al., 2008; Srinivas et al., 2010).

As revealed by the unified framework of reciprocal learning proposed in Contribution 1, bandits and Bayesian optimization are closely connected. The act space and the reward function in the bandit setup corresponds to the parameter space $\mathcal{X}$ and the unknown target function in Bayesian

[1] In case of discrete $\mathcal{X}$, practitioners typically prefer random forests, as evidenced by default settings in popular software libraries like `mlr3MBO` (Bischl et al., 2017).

[2] For target functions that are computationally cheap to evaluate, allowing for large sample sizes, Bayesian optimization is typically outperformed by heuristic optimizers like random search or evolutionary algorithms.

optimization, respectively, see also (Garnett, 2023, Section 10) for details. This allows us to transfer techniques like the information gain from the bandit setup to Bayesian optimization in a relatively straightforward away.

Informed by both the empirical findings from our sensitivity analysis and our theoretical results on the regret bounds, we propose a robust variant of BO that *avoids* prior mean parameter misspecification: Prior-mean-RObust Bayesian Optimization (PROBO). We achieve this by specifying a set of GP prior mean parameters instead of a single one, relying on imprecise Gaussian processes (Mangili, 2015, 2016). The rough idea is to incorporate the epistemic uncertainty (imprecision) regarding the choice of the prior's mean function parameter. Given a (single) base kernel $k(x, x')$ and a fixed degree of imprecision $c > 0$, Mangili (2015, Definition 2) introduces a constant mean imprecise Gaussian process as a set of GP priors:

$$\mathcal{G}_c = \left\{ \mathcal{GP}\left(Mh, k(x,x') + \frac{1+M}{c} \right) : h = \pm 1, M \geq 0 \right\}. \tag{4.5}$$

It can be shown that $\mathcal{G}_c$ verifies prior near-ignorance (Mangili, 2015, Page 194) in the sense of Benavoli and Zaffalon (2015) and that $c \to 0$ yields the precise model (Mangili, 2015, Page 189). As can be observed in Equation 4.5, the mean functional form (constant) as well as both kernel functional form and its parameters do not vary in set $\mathcal{G}_c$, but only the mean parameter $Mh \in]-\infty, \infty[$. Roughly speaking, this variation induces the set of priors.

Updating each of the priors in $\mathcal{G}_c$ via a single likelihood of observed data then gives a set of posteriors. Each of these posterior GPs has a GP mean and variance estimate, respectively. Mangili (2015, 2016) shows that upper and lower bounds of all the posterior means can be derived in analytical form, see also Rodemann (2021a,b). They shall be denoted by $\overline{\hat{\mu}}(x)_c$ and $\underline{\hat{\mu}}(x)_c$ in what follows, where c indicates that they depend on the degree of imprecision in the set $\mathcal{G}_c$ of GP priors. In a relatively straightforward way, we then generalize the lower confidence bound (LCB)[3] $lcb_\lambda(x) = \hat{\mu}(x) - \lambda \cdot \hat{\sigma}(x)$, a popular acquisition function in Bayesian optimization, where $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ are the GP posterior mean and variance estimates, respectively. Intuitively, the LCB balances exploitation of promising (i.e., low) mean predictions $\hat{\mu}(x)$ and exploration of regions in $\mathcal{X}$ with high variance $\widehat{\sigma}(x)$ (less knowledge). Imprecise BO uses the following *generalized* lower confidence bound (GLCB):

$$glcb_{\lambda,\rho}(x) = \widehat{\mu}(x) \; - \underbrace{\lambda \cdot \widehat{\sigma}(x)}_{\text{approximation uncertainty}} \underbrace{- \rho \cdot (\overline{\widehat{\mu}}(x)_c - \underline{\widehat{\mu}}(x)_c)}_{\text{model uncertainty}}. \tag{4.6}$$

As detailed already in Rodemann and Augustin (2021, 2022a,b), the GLCB generalizes the explore-exploit trade-off by explicitly accounting for the prior-induced imprecision. The parameter $\lambda > 0$ controls the classical "mean vs. approximation uncertainty" trade-off (degree of risk aversion) and $\rho > 0$ controls the "mean vs. model imprecision/uncertainty[4]." trade-off (degree of ambiguity aversion). Notably, in some cases, $\overline{\widehat{\mu}}(x)_c - \underline{\widehat{\mu}}(x)_c$ simplifies to an expression only dependent on the kernel between x and the training data. The two parameters ρ and c then collapse to one, see Equation 22 in Contribution 2.

[3] Also referred to as (upper) confidence bound.

[4] Unlike in Contribution 2 itself, we follow the terminology "model uncertainty" from Hüllermeier and Waegeman (2021) here to foster a better comparison to other sources of uncertainty described in Hüllermeier and Waegeman (2021)

We test our method on a univariate target function generated from a data set that describes the quality of experimentally produced graphene, an allotrope of carbon. The experimental manufacturing process of graphene is a as follows. High-performance plastics like polyimide are irradiated with a laser in a reaction chamber. The quality then depends on the characteristics of this process like gas and pressure in the reaction chamber or laser irradiation time. Details can be found in Section 6 of Contribution 2 and Table 3 therein. We compare GLCB against the classic LCB and six other state-of-the-art acquisition functions. We observe GLCB to significantly outperform its competitors. Further benchmark studies (see Appendices F through J in Contribution 2) show that the GLCB performs better than its competitors, if Ψ is sufficiently wiggly and multimodal, see for instance the results on the noisy and multimodal drop-wave function in Appendix I of Contribution 1. For smooth, uni- or bimodal target functions, however, PROBO is outperformed by competing acquisition functions.

4.3.2. Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration for Exosuit Personalization (Contribution 3)

Contribution 3

Julian Rodemann, Federico Croppi, Philipp Arens, Yusuf Sale, Julia Herbinger, Bernd Bischl, Eyke Hüllermeier, Thomas Augustin, Conor J. Walsh, and Giuseppe Casalicchio (2025). "Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration for Exosuit Personalization". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, Research Track and Applied Data Science Track. Springer, pages 525-542. (In press.)

Construing data as a process rather than a given thing in the spirit of Williamson (2024) and considering the data lifecycle by Yu (2020) (see Figure 1.1 in Chapter 1) implies that humans interact with machine learning models. They are active participants rather than passive percipients of such techniques.

Bayesian optimization is no exception. Established in recent years, Human-AI collaborative Bayesian optimization considers setups, where humans and the automated BO jointly optimize an unknown function, see Anjanapura Venkatesh et al. (2025); Chakraborty et al. (2025); Gupta et al. (2023); Anjanapura Venkatesh et al. (2022) and early visionary work by Borji and Itti (2013). As illustrated by the pseudo code in Algorithm 2, humans interact with the Bayesian optimization loop in human-AI collaborative BO by having the opportunity to reject and replace BO's proposals. One typical reason for humans to reject proposals is for them not to align with their own reasoning.

In Contribution 3 (Rodemann et al., 2025b), we enhance this human-machine interface (Lines 4–6 in Algortihm 2) by techniques from Interpretable Machine Learning (IML). We use Shapley values to explain the BO's proposals. Our framework, called ShapleyBO, quantifies the contribution of each parameter to BO's acquisition function. By leveraging the linearity of Shapley values, ShapleyBO can identify the influence of each parameter on BO's exploration and exploitation behaviors for additive acquisition functions like the (lower/upper) confidence bound. We also show that ShapleyBO can disentangle the contributions to exploration into those that explore aleatoric and those that explore epistemic uncertainty. After being shared with the user, this interpretation allows for a better understanding of *why* certain parameters have been proposed,

Algorithm 2 Human-AI Collaborative BO (according to Rodemann et al. (2025b))

Require: initial sample $\mathscr{D} = \{(x^{(i)}, \Psi^{(i)})\}_{i=1,\dots,n_{init}}$
1: **while** termination condition is not fulfilled **do**
2: fit a Surrogate Model $\hat{\theta}_t$ on the observed data $\mathscr{D}$
3: propose x_t that maximizes the Acquisition Function $AF(\hat{\theta}_t(x))$
4: **if** intervention criterion is fulfilled
5: $x_t \leftarrow x_t^{(human)}$, where $x_t^{(human)}$ is a human proposal
6: **end if**
7: evaluate Ψ on x_t and update $\mathscr{D} \leftarrow \mathscr{D} \cup (x_t, \Psi^(x_t))$
8: **end while**

and thus for an informed, well-grounded decision of whether these proposals align with the user's own reasoning, i.e., whether to reject and replace proposals or not.

Notably, ShapleyBO can be applied to any Bayesian optimization that uses additive acquisition functions, see Section 4 in Contribution 3—be it a collaborative setting or a classic fully automated setup. However, the ability to interpret Bayesian optimization can obviously be particularly useful for human-AI collaborative applications, where users observe each step in the sequential optimization procedure. In this case, ShapleyBO can inform users online (that is, while the optimization is still running) about why certain acts were taken over others, instead of providing such explanations after the experiment has concluded.

As shown in the simulation study in Section 5 of Contribution 2, Shapley values can provide structural insights on the relative importance of parameters for the optimization by filtering out uncertainty contributions. Informed by this simulation study on synthetic functions, we develop the following hypothesis: Basing the decision to intervene on this information will speed up the optimization in a collaborative setup. The underlying idea is that users can reject proposals in case the insights from Shapley values do not align with the user's knowledge about the optimization problem.

To test this hypothesis, we benchmark a ShapleyBO-assisted human-AI team against teams without access to Shapley values, as detailed in Section 6 of Contribution 3. We consider the real-world use case of personalizing control parameters of a wearable, assistive back exosuit by Bayesian optimization. Such robotic devices are used for reducing injury risk and supporting rehabilitation, see Siviy et al. (2023); Toxiri et al. (2019) for some background. Amongst others, these studies show that the devices' benefits can differ considerably between individuals, calling for personalization. To find optimal settings for an individual, many studies use BO, see, e.g., Zhang et al. (2017); Ding et al. (2018). This comes as no surprise, as the expensive evaluation of proposals calls for a sample-efficient, query-based method for unknown functional relationship between device settings and the target function Ψ (which consists of either some metabolic measurement or direct user feedback here).

For our experiments, we simulate a utility function for 15 users based on real-world user data from a previous experimental study Arens et al. (2025). In our simulation study, we compare ShapleyBO-assisted human-AI teams not only against trivial competitors (only human, only BO), but also against the intervention criteria proposed by Anjanapura Venkatesh et al. (2022) and a benchmark criterion which uses the parameter values of the proposed exosuit setting directly instead of our Shapley values. We find that ShapleyBO-assisted human-AI teams outperform all

four competitors on all 15 users on average. This outperformance was found to be statistically significant ($\alpha = 0.05$) in 10 out of the 15 individuals.

4.3.3. Outlook and Perspectives

While both Contributions address Bayesian optimization, Contributions 2 and 3 strongly differ along the theory-application dimension and with respect to the concrete scope. Contribution 2 addresses prior mean imprecision in a principled and theoretical way, whereas Contribution 3 solves a concrete practical problem in human-AI collaborative BO without theoretical generality.

At a second glance, however, the two contributions are more aligned than one might think. Both benefit from a (statistical) focus on the involved uncertainties that BO aims to reduce by exploring $\mathcal{X}$. Contribution 2 addresses the modeling uncertainty within epistemic uncertainty, see Section 4.3.1 and Hüllermeier and Waegeman (2021, Section 3.4), while Contribution 3 benefits from filtering out epistemic uncertainty in what parameters contribute to proposals via the linearity of Shapley values.

This comparison calls for a combined, more comprehensive approach to all sources of uncertainty in Bayesian optimization. Specifically, one could combine the risk-averse confidence bound (RACB), see Equation 4 in Contribution 3 and Makarova et al. (2021), and the generalized linear confidence bound (GLCB), see Equation 4.6 above and Definition 11 in Contribution 2, to a more comprehensive uncertainty aware confidence bound (UACB) as follows:

$$uacb_{\lambda,\rho,\alpha}(x) = \widehat{\mu}(x) \; \underbrace{- \underbrace{\lambda \cdot \widehat{\sigma}(x)}_{\text{approximation uncertainty}} \underbrace{- \rho \cdot (\overline{\widehat{\mu}}(x)_c - \underline{\widehat{\mu}}(x)_c)}_{\text{model uncertainty}}}_{\text{epistemic uncertainty}} + \underbrace{\alpha \cdot \widehat{\epsilon}(x)}_{\text{aleatoric uncertainty}}, \tag{4.7}$$

where $\widehat{\epsilon}$ is an on-the-fly estimate of the measurement noise (aleatoric uncertainty) as in Makarova et al. (2021), $\overline{\widehat{\mu}}(x)_c$ and $\underline{\widehat{\mu}}(x)_c$ are upper and lower posterior GP mean estimates, see Section 4.3.1, and $\widehat{\sigma}(x)$ is the classic GP variance estimate. In line with Hüllermeier and Waegeman (2021, Section 3.4), see Section 3.3.3, this acquisition functions balances both facets of epistemic uncertainty (approximation and model uncertainty) with aleatoric uncertainty. Recall that BO internally minimizes acquisition functions like the UACB in order to propose new configurations x_t. The UACB favors points with low[5] mean estimate, i.e., regions where the GP predicts close-to-minimal Ψ values, and with high epistemic uncertainty (aiming to reduce the latter), as well as with low noise, assuming the decision maker is risk-averse.[6] In other words, the UACB aims at proposing x values to minimze Ψ, while simultaneously exploring $\mathcal{X}$ in order to *reduce* epistemic (approximation and model) uncertainty, while *avoiding* high aleatoric uncertainty.

The rationale behind epistemic uncertainty reduction is to hedge against the risk of missing out on regions of $\mathcal{X}$ with potentially lower Ψ values. The motivation behind aleatoric uncertainty avoidance is to hedge against another type of risk, namely the risk of ending up with a higher than average Ψ due to measurement noise. Consequentially, both parameters λ and α can

[5] Recall we consider minimization problems w.l.o.g., see Section 4.3.1.

[6] Intuition for risk-aversion: Among x values with same mean predictions, a risk-averse BO operator prefers those x values with lower variance (noise) to avoid the worst case in the distribution over $\Psi(x)$.

be interpreted as degrees of risk aversion. As explained in Contribution 2, the parameter ρ constitutes the degree of *ambiguity* aversion (in the sense of Ellsberg (1961)).

Given the model uncertainty $\overline{\hat{\mu}}(x)_c - \underline{\hat{\mu}}(x)_c$ (imprecision[7]) induced by prior mean choice, the remaining epistemic uncertainty is represented by the model's predicted variance $\widehat{\sigma}(x)$. Thus, we can interpret it as (an upper bound of) approximation uncertainty. Note that the model uncertainty measured by $\overline{\hat{\mu}}(x)_c - \underline{\hat{\mu}}(x)_c$ is only a lower bound of the total model uncertainty, which also includes, for instance, uncertainty w.r.t. kernel parameter selection, as touched upon in Section 4.3.1. The UACB would allow us to analyze as to why certain values x were proposed, since it disentangles the four rationales of 1) exploitation, 2) exploration to reduce 2.a) approximation uncertainty and 2.b) model uncertainty as well as to avoid 2.c) aleatoric uncertainty. Its linearity would further allow for the application of Shapley values. These latter could inform practitioners as to what degree a specific dimension of $\mathcal{X}$ (a feature) drives exploitation on the one hand or reducing approximation, model or aleatoric uncertainty on the other hand.

On the theoretical end, a rigorous regret analysis in the spirit of Section 4 in Contribution 2 could facilitate a better understanding of why Shapley-assisted human-BO teams outperform team without access to Shapley values in the collaborative setup in Contribution 3. Anjanapura Venkatesh et al. (2022) prove that the human can increase sample efficiency of Bayesian optimization in a human-AI collaborative BO compared to fully automated BO. Their results rely on an embedding in Sobolev spaces (Tartar, 2007), which could prove useful for deriving regret bounds for Shapley-assisted human-BO teams. However, their results refer to sample efficiency only without a comprehensive account for probabilistic regret bounds along the lines of Srinivas et al. (2010). Hence, a direct transfer to our setup might be out of immediate reach and could require substantial extensions of the classic reproducing kernel Hilbert space (RKHS) embedding of GPs. We leave this to future work.

4.4. Self-Training

Compared to Bayesian optimization, self-training is a more straight-forward, almost prototypical example of reciprocal learning. Self-training is one of many approaches to semi-supervised learning (SSL). Arguably, its simplicity is what makes it one of the most popular approaches, see, for instance, Arazo et al. (2020); Rizve et al. (2020); Rodemann (2023a,b, 2024); Rodemann et al. (2023b,c); Li et al. (2020); Rizve et al. (2020); McClosky et al. (2006); Triguero et al. (2014); Rizve et al. (2020); Bordini et al. (2024); Dietrich et al. (2024, 2025)

Generally, the objective in SSL is to learn a predictive classification function $\hat{y}(x, \theta)$, parameterized by θ, from a mixture of labeled and unlabeled data. Self-training procedures begin by fitting $\hat{\theta}$ via ERM (see Equation 4.2) to the labeled data to obtain an initial model. Figure 4.2 illustrates the process for the simple case of binary classification. After the first fit on labeled data, the initial model is used to infer labels for the unlabeled examples. Next, a subset of unlabeled instances—chosen according to a "selection criterion" that typically quantifies predictive uncertainty—is added to the training set together with their predicted labels, the so-called "pseudo-labels." Because these pseudo-labels are produced by the current model $\hat{y}(x, \theta)$, they depend on the parameters θ learned from the labeled data. In any iteration $t > 1$, the collection of labeled and

[7] Recall we follow the terminology "model uncertainty" from Hüllermeier and Waegeman (2021) here to foster a better comparison to other sources of uncertainty, see above.

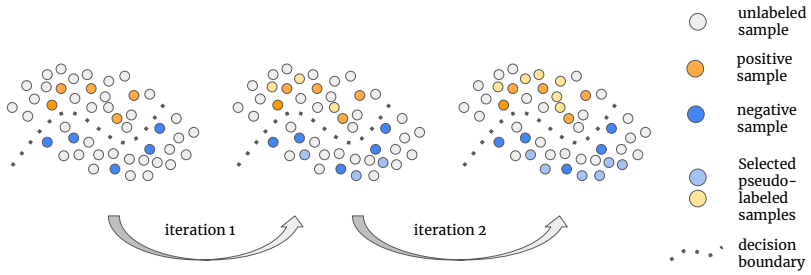

Figure 4.2.: An illustrative example of self-training for binary classification. Features x are chosen based on current model θ, then added to the sample together with self-predicted (pseudo-)labels $\hat{y}(\theta, x)$. Figure replicated from Goschenhofer (2023).

pseudo-labeled samples is determined by the model and its predictions from iteration $t-1$. This dependence instantiates the sample adaptation function introduced in Equation 4.3.

As illustrated by Figure 4.2, self-training crucially depends on the selection criterion for pseudo-label selection (PLS). This is the point of attack for subsequent Contributions 4, 5, and 6. In line with Section 2.2, we embed PLS into decision theory. This allows us to harvest the rich decision theoretic literature to derive novel robust selection criteria with solid theoretical motivation. What is more, this embedding brings us in the fortunate position to use established approximation techniques to render PLS computationally feasible. An example is the Laplace approximation for the Bayes criterion in Contribution 4.

4.4.1. Approximately Bayes-Optimal Pseudo Label Selection (Contribution 4)

Contribution 4

Julian Rodemann, Jann Goschenhofer, Emilio Dorigatti, Thomas Nagler, and Thomas Augustin (2023). "Approximately Bayes-Optimal Pseudo Label Selection". In: *Uncertainty in Artificial Intelligence (UAI)*. PMLR, pages 1762-1773.

Contribution 4 (Rodemann et al., 2023b) casts pseudo-label selection (PLS) as a decision problem and derives a Bayes criterion from it that specifically mitigates the confirmation bias (overconfident, early misjudgments) feared in self-training. More precisely, confirmation bias refers to scenarios where early overfitting propagates to the final model by selecting instances with overconfident but erroneous prediction, see Arazo et al. (2020). We again refer to Figure 4.2 for comparison.

The core idea of our Bayes criterion is not to rely on a point estimate, but to average over plausible parameter values in a Bayesian way and select instances based on the posterior predictive probability of the pseudo-samples. This "pseudo posterior predictive" (PPP) is formally proven to be a Bayes-optimal selection measure. At the same time, we discuss the closely related pseudo-marginal likelihood when the prior distribution is used instead of the posterior distribution. To

make the corresponding integrals tractable, we propose a simple, generalizable approximation via the Laplace method and Gaussian integral, which results in an easy-to-calculate score (intuitively: log-likelihood at the estimated parameter minus half the log-determinant term of the Fisher information, see Equation 3.1 in Section 3.1 for a definition of the Fisher info). Conceptually, the method is model-agnostic as long as likelihood and Fisher information are available (even for non-Bayesian models), and it explicitly aims at selecting data points (not just labels). In this way, it lays the foundations on which the subsequent two Contributions 5 and 6 build their robust extensions. In simulations and on real datasets, BPLS shows robust advantages over common, fit-centered PLS heuristics, especially in high-dimensional settings prone to initial overfitting, which often propagates to the final model in self-training, a problem known as confirmation bias (Arazo et al., 2020).

4.4.2. In All Likelihoods: Robust Selection of Pseudo-Labeled Data (Contribution 5)

Contribution 5

Julian Rodemann, Christoph Jansen, Georg Schollmeyer, and Thomas Augustin (2023). "In All Likelihoods: Robust Selection of Pseudo-Labeled Data". In: *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*. PMLR, pages 412–425.

Building on Contribution 4, we systematically extend the decision-theoretic framework of PLS in Contribution 5 (Rodemann et al., 2023c) to incorporate robustness different multiple sources of uncertainty, see Section 3.3.3 of the dissertation at hand. The main conceptual idea is to cast PLS as a *multi*-objective utility decision problem. Multiple utility functions account for the variation that certain sources of uncertainty induce in PLS.

In addition to error accumulation over iterations and covariate shift, we particularly address epistemic model uncertainty (see taxonomy by Hüllermeier et al. (2022) as touched upon in Section 4.3.2 and discussed in Section 3.3.3) by designing the selection rule across a family of (suitable) models. Where reliable second-order information on the sources of uncertainties is lacking, we further suggest a approach grounded in imprecise probabilities: a generalized Bayesian update rule via Gamma-maximin with α-cuts for credal sets. This proposal is explicitly taken up in Contribution 6, see below.

In benchmarks on simulated and datasets from finance and biology, the robust extension targeting model-selection uncertainty yields especially pronounced accuracy gains. Substantively, the work is closely aligned with BPLS: the idea of parameter marginalization in PLS remains the backbone of the method but is supported more broadly against model and data uncertainties.

4.4.3. Semi-Supervised Learning Guided by the Generalized Bayes Rule Under Soft Revision (Contribution 6)

Contribution 6

Stefan Dietrich, Julian Rodemann, and Christoph Jansen (2024). "Semi-Supervised Learning Guided by the Generalized Bayes Rule Under Soft Revision". In: *Soft Methods in Probability and Statistics (SMPS)*. Springer, pages 110–117.

Contribution 6 (Dietrich et al., 2024) rounds off this little series of Bayesian PLS (Contribution 4), robust Bayesian PLS (Contribution 5) and generalized Bayesian PLS with soft revision (Contribution 6). Being the most recent work in this series, Contribution 6 takes up the credal perspective from Contribution 5 in a targeted way and sharpens it computationally.

Under the heading of generalized Bayes, we formulate PLS with credal sets of priors and update them using Gamma-maximin under soft revision (α-cut) as initially proposed by Cattaneo (2014), see also Augustin and Schollmeyer (2021) for a statistical contextualization. The idea behind Gamma-maximin is to select pseudo-labeled data according to the lowest posterior in the credal set of updated priors (i.e., posteriors). In this way, we hedge against the worst-case prior. In other words, we select the pseudo-labeled instances that would have had the highest expected utility (likelihood) if we had specified the prior in such a way that it contradicted the (potentially overfitted) model's likelihood the most. This is motivated by our mistrust of the model's likelihood due to the confirmation bias, see Section 4.4.1. In order not to overdo it, we weaken the conservativeness of Gamma-maximin by soft revision. The latter's rough idea is to only update those priors whose respective marginal likelihood is larger or equal to $\alpha \in (0, 1)$ times the corresponding maximum marginal likelihood Cattaneo (2014). Operationally, one selects those pseudo-label data that remain most plausible under the least favourable—yet soft-revised, thus data-adapted—prior distribution.

The idea of this method was already touched upon in Contribution 5. Motivated by promising empirical results, Contribution 6 operationalizes pseudo-labeled data selection by Gamma-maximin under soft revision and offers computationally feasible implementations, allowing for a large-scale benchmark study, which compares our method to eight competing selection criteria from the literature. To achieve this, the selection is cast as an optimization problem and instantiated concretely for the class of logistic models (with Laplace approximations for marginal quantities and efficient numerical optimization). Empirically, the soft-revision variant performs particularly strongly when the share of labeled data is low and consistently ranks among the top methods.

Summing up, Contribution 6 closes the loop: It complements the computational feasibility of the pseudo posterior predictive from Contributions 4 with the robustness of our credibly cautious decision rule from Contribution 5. In other words, it carries forward the robustness trajectory from Contribution 5, while reconciling it with the approximations from Contribution 4.

4.4.4. Outlook and Perspectives

Based on the unified framework of reciprocal learning, one should not resist the temptation to transfer the robust PLS criteria across multiple instances of reciprocal learning. e.g., from self-training to active learning, building on existing uncertainty-driven acquisition strategies like Nguyen et al. (2019). Notably, this avenue for future work does not involve extending Contributions 4–6 itself, but rather (based on the unification of several learning paradigms in reciprocal learning) transferring the approaches *from* these Contributions to other instances of reciprocal learning. The multivariate utility for pseudo-label selection from Contribution 5, for instance, might be transferred to multi-agent bandits or to active learning with multiple criteria for data acquisition.

Exploring the exact opposite direction, it would be fascinating to apply the generalization bounds we derived for general reciprocal learning in Contribution 8 to self-training with generalized

Bayesian pseudo-label selection. As a matter of fact, we did something close to this potential line of future research in Rodemann and Bailie (2025, Section 6), where we applied our generalization bounds to classical self-training. Studying how our robust uncertainty treatments in pseudo-label selection affect these bounds makes up an intriguing research agenda. In particular, one could study the exact relation between decision criteria for pseudo-label selection and the Lipschitz-continuity of the so-induced sample change. As we show via the Kantorovich-Rubinstein Lemma (Kantorovich and Rubinstein, 1958), the degree of Lipschitz-continuity directly governs the Wasserstein balls, with respect to which we derive our high probability statements, in a bounded instance space (Rodemann and Bailie, 2025, Lemma 10). This implies there could be a traceable, direct connection between the robustness of pseudo-label selection and the generalization guarantees from so-enhanced samples.

As side notes, one could also study generalization guarantees under model misspecification and PLS criteria that account for the latter. Another goal would be to characterize how prior misspecification and approximations of the Fisher information bias the selection citeria, and to relate the objective to evidence-based model selection. Robustness studies could further examine the behavior under covariate shift, very similar to future work discussed in Chapter 5. Similarly, fairness-aware selection (again embedded via multicriteria utility) deserve potential attention.

Furthermore, there is some potential avenue for future *applied* work. Rather directly related to applying the generalized Bayes rule under soft revision to self-training, one could extend the models from simple logistic regression and generalized additive models (like in Contribution 5) or Bayesian neural networks (like in Contribution 4, see Appendix H) to larger neural networks architecture to keep pace (or even approach pace) with modern transformer-based model architectures that are employed in self-training within semi-supervised learning. First steps towards this direction were taken by Dietrich et al. (2025) To achieve this, the current posterior-predictive criterion could be paired with scalable approximate inference (beyond classical Laplace used in Contributions 5 and 6) such as structured Laplace, low-rank Newton or Kronecker factorizations, or variational families to obtain computationally feasible selection criteria for deep learning. Beyond standard classification, extensions to multi-label and structured outputs, class-imbalance–aware selection, and semi-supervised regression are natural potential future directions. Another path is to encode domain priors directly in the selection stage (e.g., hierarchical priors that reflect feature groups). Moreover, one could try to couple BPLS with modern SSL techniques like consistency regularization (Tarvainen and Valpola, 2017) or other approaches within the student-teacher framework in SSL. The multicriteria nature of our utility in the decision-theoretic embedding could—once more—prove flexible enough fo facilitate such extensions.

4.5. Superset Learning

4.5.1. Levelwise Data Disambiguation by Cautious Superset Learning (Contribution 7)

Contribution 7

Julian Rodemann, Dominik Kreiß, Eyke Hüllermeier, and Thomas Augustin (2022). "Levelwise Data Disambiguation by Cautious Superset Learning". In: *Scalable Uncertainty Management (SUM)*. Springer, pages 263-276.

Superset learning (also known as partial label learning) can be seen as a generalization of Semi-Supervised Learning (SSL) discussed in the previous Section. In addition to labeled data and unlabeled data, superset learning comprises scenarios where data is partially labeled. Here, instead of being either fully labeled (one label per observations) or fully unlabeled (no label per observations), data can have multiple labels. A popular interpretation from an SSL perspective is that some labels can be ruled out for the—in principle—unlabeled data.

Technically, such data is modeled as *set-valued*, that is, as observations with *sets of labels*, comprising classical labeled data (singleton) and unlabeled data (empty or full set) as special cases. Hence, we denote our observations as $\{(x_i, Y_i)\}_{i=1}^n \in \left(\mathcal{X} \times 2^{\mathcal{Y}}\right)^n$, where x_i are singleton observations of covariates and Y_i set-valued observations of target variables with $\mathcal{X}$ and (categorical) $\mathcal{Y}$ as above. Leaning on the idea of Optimistic Superset Learning (OSL) as proposed by Hüllermeier (2014) building on Grandvalet (2002); Hüllermeier and Beringer (2006); Liu and Dietterich (2012)[8], Y_i is regarded as a coarse representation (a superset, hence the name) of a true underlying singleton $y_i \in \mathcal{Y}$. This differentiates superset learning from multi-label learning (Zhang and Zhou, 2007). In what follows, $\mathcal{Y}$ is assumed to be categorical. Let $\mathbf{Y} = Y_1 \times Y_2 \times \cdots \times Y_n$ be the Cartesian product of the observed supersets. We call any singleton vector $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_n)' \in \mathbf{Y}$ an *instantiation* of the set-valued data.

In Contribution 7 (Rodemann et al., 2022b), we introduce a method for constructing a hierarchical family of subsets within such set-valued observations. Each subset reflects a level of cautiousness, starting with the smallest one—a singleton—representing the most optimistic choice. To achieve this, we build on OSL, which resolves ambiguity in set-valued data by identifying the singleton associated with the most predictive model. Specifically, we employ a variant of OSL for classification under 0/1 loss, selecting instantiations whose empirical risks fall below context-dependent thresholds. Adjusting this threshold naturally yields a hierarchy among the instantiations. To break ties that arise from identical classification errors, we leverage a hyperparameter of Support Vector Machines (SVM) that governs model complexity. We adapt the tuning of this hyperparameter to identify instantiations whose optimal separations are maximally general. Finally, we demonstrate our approach on the prototypical case of undecided political voters, treated as set-valued observations. For this purpose, we use both simulated data and pre-election polls from Civey, which included undecided voters in the run-up to the 2021 German federal election.

The attentive reader might have already noticed that this tie-breaking in cautios superset learning via SVM model complexity constitutes another special case of reciprocal learning: A model is fitted on data containing instantiations from a previous model fit. Generally and beyond our levelwise procedure of cautious superset learning in Contribution 7, superset learning can but does not need to be an instance of reciprocal learning. If embedded into the iterative self-training procedure (see Section 4.4.2) via set-valued pseudo labels, it constitutes an obvious special case of reciprocal learning. If, however, superset learning is deployed as one-shot generalized ERM as e.g., in Hüllermeier (2014), it is not subsumed by reciprocal learning.

[8] See also subsequent work by Liu and Dietterich (2014); Hüllermeier and Cheng (2015); Hüllermeier et al. (2019); Destercke (2022).

4.5.2. Outlook and Perspectives

Contribution 7 was published in 2022, long before the decision-theoretic embedding of data selection in self-training (Section 4.5) and—more generally—in reciprocal learning (Chapter 4 and Contribution 1) was ideated. In Contribution 7, we simply select instantiations with respect to empirical risk and, in case of ties, with respect to model complexity. It is thus natural to explore decision criteria beyond this lexicographic one in greater depth to better balance the trade-off between accuracy and generality. For example, one might argue that instantiations associated with higher empirical risk are still justified if they can be separated using sufficiently less complex models, i.e., with more general hyperplanes. The following Chapter 5 might be instructive here, as it specifically studies generalization capabilities of reciprocal learning algorithms.

As already discussed in Contribution 7, we see further potential in a more flexible strategy that avoids enforcing strict disambiguation of inconclusive cases. This could be realized either through two-stage criteria that incorporate additional hyperparameters of the model or a more general measure of model complexity like the Vapnik–Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1968), the Rademacher complexity (Bartlett and Mendelson, 2002; Bartlett et al., 2005) or covering number integrals (Kolmogorov and Tikhomirov, 1959; Talagrand, 2014), see Contribution 8 in Chapter 5 right below.

What is more, one could also use the Bayesian machinery for pseudo-label selection (Sections 4.4.1 through 4.4.3) for *superset* selection. The Laplace approximations derived in Contribution 4 could then be easily transferred to the superset learning setup—with potential computational benefits beyond our specific method of levelwise disambiguation.

5. Reciprocal Learning Theory

5.1. Generalization Bounds and Stopping Rules for Learning with Self-Selected Data (Contribution 8)

Contribution 8

Julian Rodemann and James Bailie (2025). "Generalization Bounds and Stopping Rules for Learning with Self-Selected Data". *arXiv Preprint* arXiv:2505.07367 (last accessed October 24 2025). *Under review at the Journal of Machine Learning Research (JMLR).*

After having introduced reciprocal learning as a unifying framework (Contribution 1) of various machine learning algorithms (exemplified by Contributions 2–7), we are now able to jointly analyze their inferential behavior. This will provide a common statistical perspective on many popular learning algorithms that self-select samples, which can be seen as automated versions of data lifecycles (Yu, 2020), see Figure 1.1 in Chapter 1.

As set out in Chapter 3, there is a broad body of statistical literature on active (experimental design, Section 3.1) and inadvertent (sampling theory, Section 3.2) self-selection of samples. Reciprocal learning constitutes a third type of self-selection mechanism, which we call *algorithmic self-selection* of samples. Its statistical properties like generalization guarantees have been understudied. Presumably, this is due to the self-selection being hidden in the automated procedures of these algorithms. Contribution 8 (Rodemann and Bailie, 2025) addresses this research gap by proving generalization bounds for learning with self-selected data in a principled way, comprising all reciprocal learning instances (see Contribution 1 and Chapter 4) as special cases.

In order to achieve this, we need to first embed reciprocal learning into statistical learning theory (Vapnik, 1998; Vapnik and Chervonenkis, 1968). Concretely, we assume a true but unknown law $\mathbb{P}$. This law is an element of the set $\mathcal{P}$ of probability measures defined on Borel σ-algebras with finite second moments on a bounded subset of the Euclidean space $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$ as introduced above. Furthermore, we fix a metric $d_{\mathcal{Z}}$ on $\mathcal{Z}$ and the Euclidean norm $\|\cdot\|_2$ on $\mathcal{Y}$, and consider the set of measurable functions

$$\mathcal{F} := \{f_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta \in \Theta\},$$

which is uniformly bounded (i.e., $\sup_{\theta\in\Theta, x\in\mathcal{X}} \|f_\theta(x)\|_2 \leq F < \infty$) because $\mathcal{Y}$ is bounded by assumption. As is customary in learning theory, we refer to $\mathcal{F}$ as the *hypothesis space* and to Θ as the *parameter space*, assumed to be a subset of a Euclidean space. Statistical learning theory also requires a complexity notion to quantify the richness of $\mathcal{F}$. We use the covering entropy integral (Kolmogorov and Tikhomirov, 1959; Talagrand, 2014). Our generalization bounds then only depend on this complexity measure, the diameter of $\mathcal{Z}$, and some other known constants.

Figure 5.1.: **Left:** Illustration of bound $W_p(\mathbb{P}, \hat{\mathbb{P}}_0) \leq \beta_0$ on the Wasserstein-p distance W_p between a law $\mathbb{P}$ and an initial *i.i.d.* sample by Fournier and Guillin (2015). **Right:** Lemma 10 in Contribution 8 provides us with a "reciprocal distortion bound" between the initial *i.i.d.* sample $\hat{\mathbb{P}}_0$ and the sample $\hat{\mathbb{P}}_T$ at reciprocal learning iteration T: $W_p(\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_T) \leq \beta_T$. Figure replicated from Rodemann and Bailie (2025).

In particular, we derive bounds on both the generalization gap and the excess risk. The former control the difference between the training error achieved by a reciprocal learning algorithm and its generalization error (Theorems 12 and 15 in Contribution 8). The excess risk is the gap between the learner's risk and the (unknown) minimal risk attained by the best hypothesis in $\mathcal{F}$ (Theorems 13, 14, 16 and 17 in Contribution 8). We further need some generic conditions (see below for a discussion) on the sample adaptation and the loss, so the results encompass the examples of reciprocal learning, see Contributions 1–7 above.

Conceptually, the key of our analysis is a reinterpretation of Wasserstein ambiguity sets $\mathcal{A}_\rho(P)$, which are defined as the p-Wasserstein ball of radius $\rho \geq 0$ centered at P:

$$\mathcal{A}_\rho(P) := \{Q \in \mathcal{P} : W_p(P, Q) \leq \rho\},$$

with $P, Q \in \mathcal{P}$ and $1 \leq p \leq 2$. The Wasserstein-p distance between $P, Q \in \mathcal{P}$ is

$$W_p(P, Q) := \inf_{c \in \mathcal{C}} \left(\int_{\mathcal{Z}} d^p_{\mathcal{Z}}(Z, Z') \, \mathrm{d}c(z, z') \right)^{1/p},$$

where the infimum is taken over all couplings $\mathcal{C}$ of P and Q.[1] Notably, Wasserstein ambiguity sets are a type of credal sets, i.e., sets of probability measures as touched upon in Chapter 1. The fact that we naturally encounter them here and, in fact, require them to capture the sample-related uncertainty in reciprocal learning further strengthens this dissertation's principled argument that credal sets are required for construing data as a process rather than a given thing, see Chapter 1.

Whereas these sets are typically employed to robustify ERM against uncertainty in the distributional *assumption* (e.g., Shafieezadeh Abadeh et al. (2015)), we use them to study shifts in the *empirical* distribution induced by the reciprocal learning algorithms themselves. This change in viewpoint is beneficial from a technical point of view: The algorithm under investigation specifies the shift precisely, allowing us to define the ambiguity sets accordingly rather than choosing them "judiciously" (Shafieezadeh Abadeh et al., 2015, Page 1576) as part of our assumptions.

As illustrated in Figure 5.1, we begin with a classical bound β_0 on the Wasserstein-p distance between the law $\mathbb{P}$ and an initial *i.i.d.* sample $\hat{\mathbb{P}}_0$. Our Lemma 10 in Contribution 8 then

[1] A coupling is a probability measure $c \in \mathcal{C}$ on $\mathcal{Z} \times \mathcal{Z}$ with marginals P and Q.

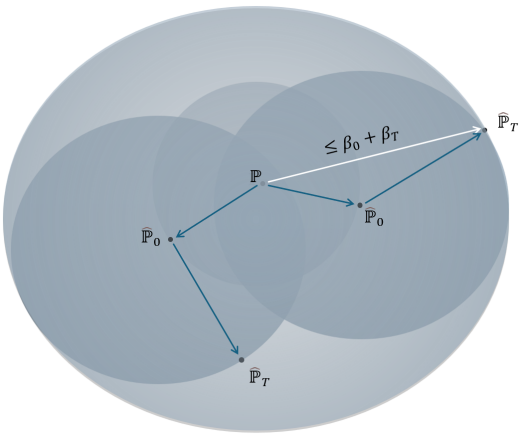

Figure 5.2.: Illustration of our core conceptual approach: Our bounds rely on a Wasserstein ball on the sample distortion in reciprocal learning. It results from bounding the Wasserstein distance between the sample $\hat{\mathbb{P}}_T$ in T and the law $\mathbb{P}$ by $\beta_0 + \beta_T$ (as illustrated in Figure 5.1) via the triangle inequality. Figure replicated from Rodemann and Bailie (2025).

bounds how far reciprocal learning algorithms move this initial *i.i.d.* sample in Wasserstein space, yielding a bound β_T on $W_p(\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_T)$ that holds for any iteration T. By the triangle inequality, we immediately obtain a bound on $W_p(\mathbb{P}, \hat{\mathbb{P}}_T)$ of the form $\beta_0 + \beta_T$ (see Figure 5.2). The remaining step is to connect Wasserstein distances between distributions to differences in the corresponding risks via the Kantorovich–Rubinstein Lemma (Kantorovich and Rubinstein, 1958). The detailed reasoning is explained in Section 5 of Contribution 8 and in its Appendix, containing all proofs.

We go on to illustrate our bounds in the concrete instance of self-training in Section 6 of Contribution 8. Inspired by a real-world application of pipeline failure prediction (Alobaidi et al., 2022) via self-training with pipe age and operational pressure as covariates, we compute our bounds to inform users when to stop the self-training process in order to retain generalization guarantees. This highlights the applicability of our theoretical analysis and allows for a tangible embodiment of how our bounds look like in practice.

5.2. Outlook and Perspectives

An obvious and clear limitation of the triangle-inequality step just sketched is that it ignores the possibility that the two sample displacements in Wasserstein space (those associated with β_0 and β_T in Figures 5.1 and 5.2) may cancel, see Rodemann and Bailie (2025, Section 2). From a generalization perspective, this would be the ideal case: reciprocal learning's sample adaptation could move the sample towards $\mathbb{P}$. Yet ensuring this typically requires assumptions about the distribution of the self-selected data, and these are fundamentally unverifiable. Practitioners may *hope* that the query function effectively accesses $\mathbb{P}$ (active learning) or that estimated pseudo-labels (semi-supervised learning) are correct. Yet, invoking Emil Heinrich Du Bois-Reymond, who famously[2] declared: "*Ignoramus et ignorabimus*" — "We do not know and we will not know" Du Bois-Reymond (1872, Page 464).

[2]Commonly attributed to Emil Heinrich Du Bois-Reymond, but popularized by the counter opinion of none less than David Hilbert (1900, 1902).

As argued for in Rodemann and Bailie (2025, Section 2), rather than relying on conjectural assumptions, we ground our analysis in *verifiable conditions* (see Conditions 1–4 in Contribution 8) on the one object we fully control: the reciprocal learning algorithms themselves. These conditions can be checked for whatever algorithm is used. In this way, we anchor our analysis in the initial *i.i.d.* sample $\hat{\mathbb{P}}_0$ and avoid any additional distributional assumptions on $\hat{\mathbb{P}}_t$ for $t \in 1, \ldots, T$. It suffices to specify conditions describing how the algorithms transform $\hat{\mathbb{P}}_t$ into $\hat{\mathbb{P}}_{t+1}$. Some of these conditions may seem strong; crucially, however, they constrain particular algorithms rather than *entire classes* of algorithms. That is, we do not exclude any learning paradigms (e.g., boosting, bandits, active or semi-supervised learning) encompassed by reciprocal learning—only specific algorithms within these classes. Taken together, our approach yields generalization bounds that are universal in two respects: They neither restrict the classes of algorithms subsumed by reciprocal learning nor impose assumptions on the distribution of newly added data.

This universality, on the other hand, also leaves room for future work that considers a more specific setup. As we make no distributional assumptions about the newly added data, accordingly, stronger bounds may be attainable once such assumptions are introduced. See Rodemann and Bailie (2025, Section 7) for an illustration: Suppose the new data is drawn *i.i.d.* from $\tilde{\mathbb{P}}$. Then $\hat{\mathbb{P}}_t$ becomes a mixture of *i.i.d.* samples from $\mathbb{P}$ and $\tilde{\mathbb{P}}$. Although $\tilde{\mathbb{P}}$ is typically unknown, prior knowledge may justify assuming a particular type of shift from $\mathbb{P}$ to $\tilde{\mathbb{P}}$—e.g., covariate, target, or conditional shift (Zhang et al., 2013).

For instance, consider covariate shifts. Here, $\mathbb{P}(X) = \tilde{\mathbb{P}}(X)$, but $\mathbb{P}(Y \mid X) \neq \tilde{\mathbb{P}}(Y \mid X)$. Obviously, the reciprocal distortion bound from Lemma 10 in Contribution 8 will be tighter, since we can replace the diameter $\mathscr{D}_{\mathcal{Z}}$ by $\mathscr{D}_{\mathcal{Y}} := \sup_{y,y'} d_{\mathcal{Y}}(y, y') < \infty$. (To see this, consider Equation 13 in the proof of Lemma 10.)

Beyond this basic example, richer prior knowledge about the shift from $\mathbb{P}$ to $\tilde{\mathbb{P}}$, e.g., via invariant feature representations (Muandet et al., 2013, 2017), can yield even tighter bounds, albeit—again—under stronger assumptions than those used in our general analysis, see Rodemann and Bailie (2025, Section 7). A possible compromise between our assumption-free setting and specifying a precise $\tilde{\mathbb{P}}$ is offered by imprecise probabilities (Walley, 1991; Augustin et al., 2014a), as outlined in Chapter 1 and in Section 3.3.3. The recently proposed "imprecise domain generalization" framework allows learners to remain noncommittal—to neither $\mathbb{P}$ nor $\tilde{\mathbb{P}}$ nor any fixed aggregation of the two—(Singh et al., 2024). Subsequent selection among one or multiple learners under such second-order uncertainty can then be guided by decision-theoretic criteria (De Campos and Ji, 2008; Antonucci and De Campos, 2011; Jansen et al., 2022b, 2023a,b, 2024). Closely related, Caprio et al. (2024) derive generalization bounds under ambiguity modeled by sets of distributions—more general than classical learning theory yet still more specific than our analysis, which makes no assumptions at all about the distribution of acquired data.

Section 7.3 will outline a further avenue of future work related to Contribution 8. The generalization bounds for learning from self-selected data could be applied to adaptive benchmarking, the very topic of Part IV of the present dissertation. The rough idea is to transfer the generalization analysis from algorithmic self-selection of samples in reciprocal learning to self-selection of instances in a benchmarking suite. We refer the reader to Section 7.3 for the details.

6. De-Biasing Trees Trained on Complex Samples as in Reciprocal Learning

6.1. Learning De-Biased Regression Trees and Forests from Complex Samples (Contribution 9)

Contribution 9

Malte Nalenz, Julian Rodemann, and Thomas Augustin (2024). "Learning De-Biased Regression Trees and Forests from Complex Samples". In: *Machine Learning* 113:3379–3398.

While Contribution 8 analyzes what can go wrong when learning from self-selected (thus generally non-*i.i.d.*) data, Contribution 9 (Nalenz et al., 2024) takes a more constructive approach. It proposes to de-bias regression trees and random forests when trained on such non-*i.i.d.* data, also referred to as complex samples, see Section 3.2 for background. This is achieved by leveraging (Hájek-type) estimators from survey statistics. Notably, our techniques work independently of the source of the bias, as long as we know the inclusion probabilities, rendering our method applicable to a wide range of data scenarios, including survey[1] data.

Even in the age of deep learning, regression trees and random forests are still widely used for tabular prediction problems. Standard training, however, assumes *i.i.d.* samples and can be biased when data comes from complex sampling designs with unequal inclusion probabilities (as in many surveys). We analyze how such "naive" tree induction that ignores sampling weights distorts all three levels of a tree model: split selection via the MSE criterion, node sizes and leaf predictions. We introduce design-aware corrections that de-bias tree induction as follows.

Bridging tree learning with survey statistics, we recast the MSE splitting objective as a variance-estimation problem and derive practical estimators (including a Hájek-type variant) that plug seamlessly into tree builders. Extending these ideas to random forests, we show in simulation studies and in a housing data application that design-corrected forests yield substantially more accurate predictions and more trustworthy tree structures. We further propose to incorporate the weights in the bootstrapping step of random forests (Breiman, 2001a) via inverse probability weighting. In other words, we up-weight those samples (via higher drawing probabilities in bootstrapping) that have low inclusion probability and vice versa. Notably, the corrected forests can even surpass models trained on *i.i.d.* samples in some cases, underscoring that principled use of sampling information can improve both accuracy and interpretability.

Beyond complex surveys, our results suggest a broader lesson for machine learning with adaptively self-collected data (like in reciprocal learning, see Contribution 1 and 8 in Chapter 4) as well as

[1] Motivated by this latter use case, this work has been supported by the Federal Statistical Office of Germany within the joint research project "Machine Learning in Official Statistics".

Rodemann et al. (2022a): Incorporating known or estimable inclusion probabilities and design-based estimators can stabilize learning and reduce bias, much like classical Horvitz–Thompson ideas do in sampling theory, see Section 3.2. The proposed corrections are simple to implement (as demonstrated in the software accompanying Contribution 9) and offer a principled path to de-biased trees and forests whenever sampling departs from *i.i.d.* assumptions.

6.2. Outlook and Perspectives

Looking ahead, there are three particularly fertile avenues for future work. On the theoretical front, one could try to establish risk consistency (Köhler, 2024) or convergence rates for design-aware trees and forests under complex samples. Moreover, one could try to clarify when finite-population versus super-population views coincide, and develop post-selection corrected inference in case of subsequent analyses motivated by earlier results.

Second, there are some direct practical extensions within reach. For instance, our unbiased estimators could be extended from random forests (averaged tree prediction from many models trained independently on bootstrap samples in parallel) to boosting (tree predictions from many models trained sequentially). Practically, this extension appears straightforward, as our implementation of our de-biased trees and forests is based on the `xgboost` package (Chen et al., 2015) in `R` (R Core Team, 2025).

Third, reliability of trees and forests trained on complex samples could be improved by building design-respecting uncertainty quantification. Concretely, explicit confidence intervals for leaf predictions are within reach thanks to the de-biased variance estimators from survey statistics, see Sections 3.2 and 3.3 in Contribution 9. In a similar spirit, reliability could benefit from analyzing how our debiasing schemes affects random forest interpretability beyond partial dependence and variable importance, as discussed in Section 4.3 in Contribution 9.

More generally, there appears to be a lot of potential for methodological innovation in reciprocal learning inspired by population variance estimation approaches or Horvitz-Thompson-type reweighting from survey statistics and sampling theory (see Section 3.2) beyond the special case of trees and random forests. In reciprocal learning's special case of bandits, inferential guarantees for weighted M-estimation have been established by Zhang et al. (2021); Zhang (2023). The weighting scheme is similar to the one used in Contribution 8, which becomes evident by comparing Zhang et al. (2021, Section 3.1) and Section 3.3 in Contribution 9. Our unified framework of reciprocal learning, as introduced in Contribution 1, allows transferring such innovations from one instance of reciprocal learning (like bandits) to another one (like, e.g., self-training, active learning and many more).

Finally, large-scale benchmarks covering both complex surveys and adaptively collected samples in reciprocal learning would accelerate adoption and help surface trade-offs between bias reduction and computational effort. Speaking of which, the following part IV of this dissertation will shed light to benchmarks more generally.

Part IV.

Statistical Perspectives on Testing Data: The Benchmark Problem

7. Theory

After having studied training data selection within reciprocal learning almost ad nauseam, this part of the dissertation shifts the attention to testing data used to analyzing and comparing machine learning methods, specifically to benchmarking. Benchmarks are widely considered an integral part of data-centric machine learning, see e.g., Eyuboglu et al. (2022); Mazumder et al. (2022); Zha et al. (2025); Chen et al. (2023); Huang et al. (2024), as detailed in Section 3.3.1.

Just like before, we adopt a statistical perspective. In the benchmarking setup, this perspective is as follows. The datasets under consideration are construed as statistical units making up a sample drawn from an (in its entirety unknown) population. The statistical lens then differentiates between a *descriptive* and an *inferential* analysis of benchmarking results. While the former *describes* the benchmark result as is, the latter *infers* statements about the population. Since this population is unknown, these statements come with uncertainty. Statistics offers rigorous and principled methods for quantifying this uncertainty, see again Davidian and Louis (2012); American Statistical Association (2012) as discussed in Section 3.3.3.

Recall from Section 2.2 that the benchmark aggregation problem comes down to this single question: **How to compare multiple algorithms on multiple instances (typically datasets or prompts) with respect to multiple criteria?** The alert reader might have already observed that this question once again gives rise to a decision problem (Section 2.2) under complex uncertainty (Section 2.1): We need to decide for one or multiple algorithms based on observed performances (that are subject to uncertainty) along multiple criteria.

In various applications, these criteria have different scales of measurements. As touched upon in Section 2.2, one might, for instance, be interested in comparing language models with respect to established metrics like semantic coherence and diversity scores (cardinal[1] scale) on the one hand and human Likert-scale rankings (ordinal[2] scale) on the other. In the decision-theoretic embedding of benchmarking from Section 2.2 these metrics with different scales were embodied by a multivariate consequence space C *with locally varying scale of measurement*. The language models (or any other algorithms like classifiers, see Contribution 11) were represented by acts $X : \Theta \to C$ mapping into these non-standard spaces. Technically, theses "acts" are random variables, as we have equipped both X and C with appropriate σ-fields in Section 2.2. In benchmarking, we want to compare these random variables to one another.

In a nutshell, Contribution 10 lays the groundwork for doing just that; namely, by proposing an information-efficient order of random variables mapping into such non-standard spaces. The proposed order interpolates between stochastic dominance and expectation order depending on the available cardinal information. What is more, we derive a regularized statistical test for empirically testing this order. We implement the test via linear optimization and robustify it through imprecise probability models, as motivated in Sections 1 and 3.3.3. We conclude by illustrating its usefulness in applications like poverty measurement, finance, and medicine.

[1] Differences between elements are meaningful.
[2] Only the sign of differences between elements carries meaning.

To elaborate a bit further, recall that we defined decision theory in Section 2.2 as the theory of finding appropriate choice functions $ch : 2^{\mathcal{G}} \to 2^{\mathcal{G}}$ satisfying $ch(\mathcal{D}) \subseteq \mathcal{D}$ for all non-empty $\mathcal{D} \in 2^{\mathcal{G}}$ with $\mathcal{G} \subseteq C^{\Theta}$. In benchmarking, this means nothing less than selecting algorithms based on their results on a benchmark suite. Starting with the overly conservative point-wise Pareto choice function (equation 2.7), we study (first-order) stochastic dominance as a weaker and more appropriate alternative. Given any preorder[3] $\precsim$, its choice function selects

$$ch_{\precsim,\pi}(\mathcal{D}) = \left\{ X' : \nexists X \begin{array}{l} \forall \ell \in \mathcal{L}_{\precsim} : \mathbb{E}_{\pi}(\ell \circ X - \ell \circ X') \leq 0 \\ \exists \ell \in \mathcal{L}_{\precsim} : \mathbb{E}_{\pi}(\ell \circ X - \ell \circ X') < 0 \end{array} \right\} \tag{7.1}$$

for all $\mathcal{D} \subseteq \mathcal{G}$, where $\mathcal{L}_{\precsim}$ is the set of all $\precsim$-isotone (and measurable) loss functions $\ell : C \to [0,1]$. We further interpreted $ch_{\precsim,\pi}(\mathcal{D})$ as choosing all acts that are not excluded by every compatible risk-minimizer. We can equivalently, in line with Contribution 10, express this stochastic dominance in terms of utility maximization rather than loss minimization by defining $\mathcal{U}_{\precsim}$ as the set of all $\precsim$-isotone (and measurable) utility functions $u : C \to [0,1]$. With this, the stochastic dominance choice sets equal

$$ch_{\precsim,\pi}(\mathcal{D}) = \left\{ Y : \nexists X \begin{array}{l} \forall u \in \mathcal{U}_{\precsim} : \mathbb{E}_{\pi}(u \circ X - u \circ Y) \geq 0 \\ \exists u \in \mathcal{U}_{\precsim} : \mathbb{E}_{\pi}(u \circ X - u \circ Y) > 0 \end{array} \right\} \tag{7.2}$$

for all $\mathcal{D} \subseteq \mathcal{G}$. As can be seen in both equivalent formulations (equations 7.1 and 7.2), stochastic dominance considers expectations with respect to π instead of point-wise dominance like the Pareto order. While this makes it more appropriate than the strict and thus information-inefficient Pareto choice function, stochastic dominance still relies on a preorder on the consequence space C. In the words of Section 3.3.3 on different sources of uncertainty, it does not weaken the *structural* source of epistemic model uncertainty, which arises from weakly structured order information on C. This wastes information. To understand why, consider the motivating example of mixed ordinal-cardinal scaled C again. A preorder will rank algorithm X above algorithm X' if X dominates X' along all criteria, including the cardinal ones. Thus, it will not be able to represent the cardinal *intensity*, i.e., information beyond the ordering or—in plain terms— the information about *how much better* algorithm X is than X'.

This is exactly our point of attack in Contribution 10 (Jansen et al., 2023b). Building on earlier work (Jansen et al., 2018, 2023a), we propose to benchmark algorithms based on generalizing the stochastic dominance ordering to the generalized stochastic dominance (GSD) ordering that ensures exploiting the entire information in C. We achieve this via so-called *preference systems* $\mathcal{A} = [C, R_1, R_2]$ with a preorder R_1 on C (ordinal content) and a preorder R_2 on *differences* $(a,b) \in R_1$ (cardinal content/intensity).[4] Concretely, assume C is $r \in \mathbb{N}$–dimensional and (w.l.o.g.) the first $0 \leq z \leq r$ dimensions are on a *cardinal scale* (so that differences between elements are meaningful), while the remaining dimensions are purely *ordinal* (only sign of differences has meaning). In Contribution 10, we then work with (bounded subsystems of) the preference system

$$\mathit{pref}(\mathbb{R}^r) = \left[\mathbb{R}^r, R_1^*, R_2^*\right], \tag{7.3}$$

with preorders

$$R_1^* = \left\{(x,y) : x_j \geq y_j \;\; \forall j \leq r\right\}, \tag{7.4}$$

[3] Recall that a preorder is a binary relation $R \subseteq M \times M$, $M \neq \emptyset$, if $(a,a) \in R$, (*reflexive*) and $(a,b),(b,c) \in R \Rightarrow (a,c) \in R$ (*transitive*).

[4] Less sloppy, let $A \neq \emptyset$ be a set, $R_1 \subseteq A \times A$ a preorder on A, and $R_2 \subseteq R_1 \times R_1$ a preorder on R_1. The triplet $\mathcal{A} = [A, R_1, R_2]$ is then called a **preference system** (Jansen et al., 2018) on A.

and

$$R_2^* = \left\{ ((x,y),(x',y')) : \begin{array}{lllll} x_j - y_j & \geq & x'_j - y'_j & & \forall j \leq z, \\ x_j & \geq & x'_j & \geq & y'_j \geq y_j \quad \forall j > z \end{array} \right\}, \tag{7.5}$$

where $x, y \in \mathbb{R}^r$, see Jansen et al. (2023b, Section 7). Evidently, R_1^* is a straightforward component-wise dominance relation. R_2^* prefers one pair of consequences to another if, in the ordinal dimensions, the exchange induced by the first pair does not represent a deterioration compared to the exchange induced by the second, and if, in addition, the differences in the cardinal dimensions exhibit component-wise dominance. For a more precise characterization of the GSD-relation in multidimensional settings, we refer the reader to Jansen et al. (2023b, Proposition 5) in Contribution 10.

Proposition 7 i) in Contribution 10 shows that $[\mathbb{R}^r, R_1^*, R_2^*]$ on $\mathbb{R}^r$ is consistent, i.e., there exists a representation $u : \mathbb{R}^r \to \mathbb{R}$ such that for all $a, b, c, d \in \mathbb{R}^r$ we have:

i) If we have that $(a,b) \in R_1^*$, then it holds that $u(a) \geq u(b)$, where equality holds if and only if $(a,b) \in I_{R_1^*}$ (with I_R denoting the indifference part of a preorder R, i.e., $(a,b) \in I_R$ iff $(a,b) \in R \wedge (b,a) \in R$).

ii) If we have that $((a,b),(c,d)) \in R_2^*$, then it holds that $u(a) - u(b) \geq u(c) - u(d)$, where equality holds if and only if $((a,b),(c,d)) \in I_{R_2^*}$.

The set of all representations of the preference system $\mathit{pref}(\mathbb{R}^r) = [\mathbb{R}^r, R_1^*, R_2^*]$ on $\mathbb{R}^r$ (i.e., on multidimensional C with mixed scales) is denoted by $\mathcal{U}_{\mathit{pref}(\mathbb{R}^r)}$. Finally, we can formally define our GSD ordering relation for multicriteria benchmarking, which will also be the basis for Contribution 11, as follows.

For π a probability measure on $(\Theta, \mathcal{S}_1)$, we define

$$\mathcal{F}_{(\mathit{pref}(\mathbb{R}^r),\pi)} := \Big\{ X \in A^\Theta : u \circ X \in \mathscr{L}^1(\Theta, \mathcal{S}_1, \pi) \ \forall u \in \mathcal{U}_{\mathbb{R}^r} \Big\}, \tag{7.6}$$

where $\mathscr{L}^1(\Theta, \mathcal{S}_1, \pi)$ is the usual space of Lebesgue integrable functions. Set $X, Y \in \mathcal{F}_{(\mathit{pref}(\mathbb{R}^r),\pi)}$. Define Y to be $(\mathit{pref}(\mathbb{R}^r), \pi)$-dominated by X if

$$\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y) \tag{7.7}$$

for all $u \in \mathcal{U}_{\mathbb{R}^r}$. The induced relation is denoted by $R_{(\mathit{pref}(\mathbb{R}^r),\pi)}$ and called generalized stochastic dominance (GSD). It will be the backbone of Contributions 10, 11, and 14.

7.1. Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement (Contribution 10)

Contribution 10

Christoph Jansen, Georg Schollmeyer, Hannah Blocher, Julian Rodemann, and Thomas Augustin (2023). "Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement". In: *Uncertainty in Artificial Intelligence (UAI)*. PMLR, 941–952.

In Contribution 10, we go on to propose a regularized and robustified (two-sample) hypothesis test (a permutation test) that allows practitioners to statistically test whether a random variable GSD-dominates another one. Since directly testing $H_0^{\text{id}} : (X, Y) \notin R_{(\mathcal{A},\pi)}$ vs. $H_1^{\text{id}} : (X, Y) \in R_{(\mathcal{A},\pi)}$ is "too broad" on the null, we propose to test in both directions to account for the inferential asymmetry inherent in partial (and thus, pre-)orders. The population signal of our test is

$$D(X, Y) \;=\; \inf_{u \in N_{\mathit{pref}(\mathbb{R}^r)}} \Big\{ \mathbb{E}_\pi[u(X)] - \mathbb{E}_\pi[u(Y)] \Big\},$$

where $N_{\mathit{pref}(\mathbb{R}^r)}$ is the set of all normalized (Jansen et al., 2023b, Definition 3) representations $\mathcal{U}_{\mathit{pref}(\mathbb{R}^r)}$ (see above) of our preference system $\mathit{pref}(\mathbb{R}^r)$. The empirical test statistic is its sample analog $d_{X,Y}$ computed on the finite support of X, Y. In simple terms, $d_{X,Y}$ asks whether *even under the most adverse admissible representation* u the expected advantage of X over Y is nonnegative. Computationally, we reduce the testing procedure to mixed-integer linear programs embedded in a permutation scheme, see Jansen et al. (2023b, Section 6.2)

Amongst other scientific domains like finance and medicine, we apply our test to multidimensional poverty measurement, see also Section 2.2 for a motivation. We use data from the German General Social Survey (ALLBUS) (Diekmann et al., 2019; GESIS, 2018) that accounts for three dimensions of poverty: income (numeric), health (ordinal, 6 levels) and education (ordinal, 8 levels). To illustrate our testing procedure, we test the null hypothesis that women are poorer than men regarding any compatible utility representation of income, health and education. That is, we test the null of women being GSD-dominated by men with respect to multidimensional poverty.

In line with our deliberations in Chapter 1 about the (often times underestimated) strength of the ubiquitous *i.i.d.*-assumption, we robustify our hypothesis test against deviations from this assumption, building on ε-contamination models (Walley, 1991, Page 147), see also Huber (1981) and related overviews by Destercke et al. (2022); Montes et al. (2020a,b).

Loosely speaking, we perturb the observed empirical distribution in such a way that we can relate each perturbation to a share of data points being arbitrarily distributed.[5] More formally speaking, we hedge against violations of *i.i.d.* by replacing empirical laws by (empirical) credal sets $\widehat{\mathcal{M}}_X, \widehat{\mathcal{M}}_Y$ and, instead of $d_{X,Y}$, use the lower envelope

$$\underline{d}_{X,Y}(\omega) = \inf_{\pi_1 \in \widehat{\mathcal{M}}_X,\, \pi_2 \in \widehat{\mathcal{M}}_Y} \inf_{u \in N_{\mathit{pref}(\mathbb{R}^r)}} \sum_z u(z)(\pi_1(\{z\}) - \pi_2(\{z\}))\,.$$

For ε–contamination (*linear–vacuous*) models $\widehat{\mathcal{M}}_Z = \{\pi : \pi \geq (1-\varepsilon)\hat{\pi}_Z\}$, least favorable extreme points exist in closed form, see Jansen et al. (2023b, Section 6.2). Hence $\underline{d}_{X,Y}$ is obtained by linear programs. In our permutation test, we then compare this to the permutation distribution of the corresponding *upper* envelope.

The ε–contamination model allows us to analyze how robust (or sensitive) our conclusions are to deviations in the sample. This aligns with the perspective of sampling theory (Section 3.2) and experimental design (Section 3.1). More generally, it helps to shed some light on what is usually just taken as a *given* (thing): *i.i.d.* data.[6]

[5] Note that we are not *actually* perturbing the sample itself but what we do with the observed empirical distribution is in fact equivalent to perturbing the sample and then computing this perturbed sample's empirical distribution, given a distributional assumption. Note that working with an "observed empirical distribution" presupposes a distributional assumption.

[6] Cf. its Latin meaning, see Chapter 1.

Summing up in the most crude way, GSD exploits all ordinal and cardinal information in mixed-scale outcomes and our two-sample permutation test allows practitioners to detect GSD-dominance empirically. Robustification via credal sets (with closed-form least favorable points under γ–contamination) is feasible and impactful in practice, as demonstrated by the case study on multidimensional poverty measurement.

7.2. Statistical Multicriteria Benchmarking via the GSD-Front (Contribution 11)

Contribution 11

Christoph Jansen*, Georg Schollmeyer*, Julian Rodemann*, Hannah Blocher*, and Thomas Augustin (2024). "Statistical Multicriteria Benchmarking via the GSD-Front". In: *Neural Information Processing Systems (NeurIPS)*. Spotlight Award, 98143–98179.

Motivated by multicriteria benchmarking, the above Section 7.1 laid the foundations for robust and computable hypothesis tests for random variables with differently scaled dimensions. Alone, the actual applications to multicriteria benchmarks were missing. Contribution 11 (Jansen et al., 2024) fills exactly this gap.

However, in order to fully apply the GSD-machinery to large scale benchmarks of algorithms (illustrated by classifiers in the following), some further methodological innovations are required. Namely, we develop and substantially extend the GSD-based ordering by introducing the concept of the GSD-*front* as an information-efficient improvement of the Pareto-front based on Pareto-ordering as introduced in Section 2.2. We further propose a set-valued estimator for the GSD-front and provide sufficient conditions for its consistency. Building on the permutation test introduced in Contribution 10, we develop (static and dynamic) statistical (permutation-)tests if an algorithm is in the GSD-front. Moreover, we directly quantify how robust the test decisions are under deviations from the underlying assumption of identically and independently distributed (*i.i.d.*) samples. To put this theory into practice, we also offer an efficient implementation that is freely available and easily adaptable to comparable problems.

As touched upon in Section 2.2, traditional approaches to benchmark aggregation either reduce multidimensional performance to arbitrary scalar values by weighting schemes or retreat to the overly conservative (point-wise) Pareto-front, see Equation 2.7. Moreover, existing approaches typically fail to account for statistical uncertainty arising from finite benchmark suites and do not verify robustness under violations of underlying distributional assumptions like the *i.i.d.* sampling assumption.

In Contribution 11, we address these shortcoming by introducing the GSD-front (building on the GSD-based ordering, see above) as a principled middle ground between these extremes. Building on decision-theoretic foundations in Sections 2.2 and 7.1, we embed multidimensional quality metrics into preference systems that formalize both ordinal information (rankings) and cardinal information (intensity of differences). An algorithm $A \in \mathcal{A}$ then belongs to the GSD-front $gsd(\mathcal{A})$ if it is not strictly dominated by any competitor $A' \in \mathcal{A}$ with respect to all compatible utility representations of the quality metrics. In Jansen et al. (2024, Theorem 1), we establish that the (regularized by parameter r and based on n observations) empirical GSD-front $egsd_n^{r(n)}(\mathcal{A})$

*These authors contributed equally to this work.

constitutes a consistent statistical estimator of the true GSD-front under appropriate conditions, i.e.,

$$\pi\left(\left\{\theta \in \Theta : \lim_{n\to\infty} egsd_n^{r(n)}(\mathcal{A}) = gsd(\mathcal{A})\right\}\right) = 1. \tag{7.8}$$

We further prove in Jansen et al. (2024, Theorem 2) that the GSD-front is always a subset of the Pareto-front, making it more discriminative.

Beyond estimation, we develop a comprehensive inferential framework building on the decision-theoretic perspective outlined in Section 2.2. We propose both static and dynamic permutation tests to determine whether a given classifier lies in the GSD-front. These tests are level-α-valid and consistent, see Jansen et al. (2024, Theorem 3). The static test rejects the global null hypothesis if all pairwise tests reject at level α, while the dynamic test identifies the maximal subset of c algorithms for which the candidate significantly lies in the GSD-front by conducting pairwise tests at level α/c. Crucially, based on Contribution 10, we address a major practical concern in benchmarking that resonates with the data-centric perspective of this dissertation: the questionable *i.i.d.* sampling assumption for benchmark suites, as discussed in Sections 1, 3.2 and 7.1. We quantify how our static and dynamic tests' decisions change when a known number of datasets may be arbitrarily distributed, allowing practitioners to assess the stability of their conclusions under realistic (and tangible, see Section 7.1) violations of assumptions.

We demonstrate our methodology on two well-established benchmark suites, namely OpenML (Bischl et al., 2021, 2025) and PMLB (Olson et al., 2017). On OpenML[7], we find that SVM significantly lies in the GSD-front of a subset of algorithms even under contamination of up to 7 datasets. On PMLB[8], we find that the CRE (compressed rule ensemble classifier, Nalenz and Augustin (2022)) cannot be confirmed as lying in the GSD-front of state-of-the-art methods.

These applications evidently connect to the broader theme of reliable testing data selection for benchmarking. Our GSD-based methodology informs benchmark suite designers how to a) handle multiple criteria with mixed scales of measurement, both ordinal and cardinal, b) represent statistical uncertainty through formal hypothesis testing and c) quantify robustness under assumption violations. As demonstrated in Contribution 14 (see Section 8.3), this framework extends naturally to benchmarking (large) language models, completing the bridge between foundational statistical methodology grounded in decision theory and modern machine learning applications.

7.3. Outlook and Perspectives

In order to keep up with the immense speed of algorithm (especially ML model[9]) development, benchmark suites have to be continuously updated and maintained. In line with our dynamic, cyclic perspective in Chapter 1 and its embedding into decision theory as sequential *nested* zero-sum games in Section 2.2, we could aim to adapt our inferential guarantees like consistency

[7] 80 binary classification datasets, comparing 7 algorithms on accuracy, train time and test time, see Jansen et al. (2024, Section 5.1 and Appendix C.1).

[8] 62 datasets, evaluating a compressed rule ensemble classifier against 5 competitors on robust accuracy (composed of classical accuracy and robustness of accuracy w.r.t. perturbed features and classes), see Jansen et al. (2024, Section 5.2 and Appendix C.2).

[9] In recent years, one is inclined to add, especially transformer-based deep learning model development.

of estimation (Proposition 7 in Contribution 10, Theorem 1 in Contribution 11) and validity of tests (Theorem 3) to a sequential setup.

Classical post-selection corrections (Benjamini and Hochberg, 1995) might help, but potentially not suffice here. Recall the comparison of feedback loops in Chapter 1, particularly Figure 1.2: Post-selection corrections allow for valid inference when the latter is focused "on some findings that turned out to be of interest only after viewing the data" (Benjamini, 2020, Page 7). In other words, modeling assumptions (like defining the parameter space) are made after preliminary analysis. If benchmark suites (made up of the domain of algorithms to be compared and the criteria considered) are adapted in light of intermediate findings, one can retain valid inference via standard post-selection correction methods such as Bonferroni corrections or family wise error rates (Bonferroni, 1936; Holm, 1979; Romano and Wolf, 2007). If, however, these preliminary findings inform dataset curation, i.e., the subsequent sample of datasets (which are the statistical units here) in the benchmark suite, we are faced with what appears to be an instance of *reciprocal learning*: inference in t affects the sample in $t+1$. In case the latter is understood as ground-truth (as in a complete survey, i.e., equating the population), we are faced with an example of performativity. Arguably, however, the former scenario of only having samples is more common in benchmark analyses.

To gain an instructive illustration for a better understanding as to why post-selection corrections are not always sufficient here, consider the following analogy of benchmarking to a very classical statistical testing example, where selective inference is well understood and frequently applied: Assume we want to compare the efficacy of two drugs (e.g., newly developed and established antihypertensives) with respect to some parameter (e.g., blood pressure in patients after taking the drug). To get a license by regulators like the federal drug agency (FDA), pharmaceutical companies typically have to conduct statistical hypotheses tests to provide evidence for their drugs' efficacy (Bates et al., 2022). In our simplified example of antihypertensives, the pharmaceutical firm might conduct a two-sample permutation test—just like in Contribution 10 and 11—to test whether there is a significant difference between the old and the new drug's efficacy.

If the parameter considered is chosen or adapted after first tests or multiple hypotheses are tested on the same data, post-selective inference is the appropriate framework. If, however, the intermediate results inform the inclusion of more patients (statistical units), the pharmaceutical should turn to the adaptive design or bandit literature (Wald, 1947b; Robbins, 1952; Siegmund, 1985; Pocock, 1977; Kim and Demets, 1987; Bauer and Köhne, 1994) as well as anytime valid inference and e-values (Grünwald et al., 2024; Ramdas et al., 2023; Howard et al., 2021; Vovk and Wang, 2021; Ramdas and Wang, 2025). In particular, if the subsequent sample-extension decisions (of what units to include or exclude) are made in a controllable, algorithmic way, reciprocal learning (Rodemann et al., 2024), see Chapter 4, is the right framework.

Statistically rigorous benchmarking mirrors this setup. We compare algorithms, not drugs, and we typically do so with respect to more than one parameter (i.e., criterion). But the inference challenge is the same: We want to rule out with high confidence (i.e., low probability of Type I error) that the observed dominance of drug/algorithm A over drug/algorithm A' is due to chance. When preliminary findings inform the design and development—as is often the case—of new drugs/algorithms or new criteria in the comparison, the parametrization of our statistical inference problem is altered. Hence, post-selection correction is required for following inferential statements to remain valid. If, however, datasets (i.e., the sample making up the benchmark suite) are changed in response to previous analyses—see, for instance, Section 8.1 or Section 8.4—generalization bounds for reciprocal inference, i.e., learning with self-selected

data (Contribution 8) could help to obtain valid inferential guarantees. They bound the excess risk (and thus, the generalization error) with high probability given the sample adaptation is Lipschitz in Wasserstein space and the initial (untouched) sample is *i.i.d.*

In order to actually be of practical help for hypothesis testing, however, the bounds on the risk of a learner would have to be equipped with an interpretation suitable for statistical hypothesis testing. This, however, appears within reach when using—once more— a decision-theoretic stance: One can easily formulate testing as a decision rule $\delta : \mathcal{X} \to \{0, 1\}$ with $\delta(x) = 1$ meaning "reject H_0" and $\delta(x) = 0$ meaning "do not reject H_0", see, e.g., Augustin and Jansen (2023). Consequently, Type I and Type II error rates are $\alpha(\delta) := \mathbb{P}_{H_0}(\delta(X) = 1)$, and $\beta(\delta) := \mathbb{P}_{H_1}(\delta(X) = 0)$, respectively. We can retrieve a risk function $R(\delta)$ very easily now. For example, under standard 0/1-loss, the Bayes risk with class priors $\pi_0 = \mathbb{P}(H_0)$ and $\pi_1 = \mathbb{P}(H_1)$ would be $R_{\text{Bayes}}(\delta) = \pi_0\,\alpha(\delta) + \pi_1\,\beta(\delta)$, which gives a risk expression of Type I and Type II errors, allowing for a testing interpretation of risk bounds. A more detailed study, obviously, is required to allow for explicit disentanglement of the two error types in the risk formulation.

Not without some excitement, the following general remark should be made. This very avenue for future work just sketched could bridge the two main parts of this dissertation (Part III and IV). It connects the reciprocal perspective on sample selection to the samples (namely, datasets) in benchmarking studies. Specifically, it connects our generalization guarantees for drawing conclusions from self-selected, model-informed observations to statistically valid benchmarking analyses with quantified uncertainty.

A complementary direction for future research related to Contributions 10 and 11 is to extend the GSD-framework toward regression-style analyses, as suggested by Jansen (2025). In its current form, GSD-front analyses treat datasets as exchangeable and largely ignore dataset-level metadata. The latter, however, is often present in benchmarking suites, see, e.g., (Olson et al., 2017; Bischl et al., 2021, 2025; Mucsányi et al., 2024). A natural next step is to incorporate dataset metadata as covariates and either stratify the GSD comparison by them or adjust for them directly. Doing so yields situation- or task-specific results and, in turn, more informative conclusions than a single comparison. From a statistical perspective, this translates to modeling conditional instead of marginal distributions. Instead of our two-sample permutation test, this would call for t-tests or F-tests within (generalized) linear models to test the significance of the covariates' effects on the benchmarking results. As a stylized example, imagine each dataset is annotated with its class-imbalance ratio. One can compute GSD-fronts within strata defined by imbalance (e.g., low, moderate, severe) and then examine how dominance relations shift across data-regime boundaries. A potential research hypothesis could be the following. In severely imbalanced settings, calibrated probabilistic models may be contained in the GSD-front, while in balanced, large-sample regimes, deep learning models or kernel methods might rise to the (GSD-)front.

8. Application

The previous Chapter 7 contributed to the theoretical foundations of benchmarking by studying it from a principled, decision-theoretic and statistical perspective. In what follows, we put the so-derived methodologies to work.

In particular, Contribution 14 (Section 8.3) benchmarks language models via the GSD-methodology derived in Contributions 10 and 11, see above. Before turning to this Contribution in more detail, we shall, however, turn to an application of depth functions to multicriteria benchmarking results, building on Blocher et al. (2023); Blocher and Schollmeyer (2024). Different to the GSD-approach, Contribution 12 will only capture the ordinal structure of benchmarking results. Here, the motivation is that often times *rankings* of algorithms are the practitioner's ultimate goal, rendering the metric information a means to an end. In the end, bluntly put, practitioners sometimes simply have to pick *one* algorithm for the task they are faced with. Contribution 13 will ponder upon the motivation for benchmarking analyses further, working out two idealized perspectives on benchmarking: choosing an algorithm for a given task and gaining structural insights into algorithm performance to, e.g., design a new one, both of which will be addressed by our practical recommendations in Contribution 13.

8.1. Partial Rankings of Optimizers (Contribution 12)

Contribution 12

Julian Rodemann* and Hannah Blocher* (2024). "Partial Rankings of Optimizers". In: *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Reiterating the common thread within Part IV of this dissertation, the central benchmark question is: How to compare multiple algorithms (here: optimizers) on multiple instances (here: test functions) with respect to multiple criteria (here: e.g., fixed-budget (performance) and fixed-target (speed) evaluation)?

Contribution 12 (Rodemann and Blocher, 2024) provides a fresh perspective on this ubiquitous question with a special focus on applications in multicriteria benchmarking of *optimizers*. Different to previous Contributions 10 and 11, this perspective is *descriptive* only. That is, instead of aiming at inferential statement about an unknown population, we are primarily interested in describing fully available benchmarking results at hand.

Our perspective is grounded in the theory of depth functions (Liu, 1990; Zuo and Serfling, 2000; Mosler and Mozharovskyi, 2022), which formalize notions of centrality. Given a set of observations, a depth function assigns a measure of centrality (or, inversely, outlyingness) to

*These authors contributed equally to this work. Since partial ordering was not permitted, a total order was enforced by a fair coin flip.

each element in this set. Building on Blocher et al. (2022, 2023); Blocher and Schollmeyer (2024), we use a specific depth function that does just that for partially ordered sets (posets) (see also Blocher and Schollmeyer (2025)). Partial orders are preorders[1] that are also antisymmetric. A binary order relation p is antisymmetric if $(y_1, y_2) \in p$ with $y_1 \neq y_2$ implies $(y_2, y_1) \notin p$.[2] As it will turn out, such posets are an accurate and almost assumption-free representation of multicriteria benchmarking results.

Our fresh perspective is motivated by two observations in practice. First, in many setups, benchmarking is often conducted with the primary goal of *ranking* optimizers. This downgrades raw metric values to a means rather than an end. Second, the use of multiple criteria naturally produces incomparabilities—an optimizer can be better on one metric and worse on another—which classical aggregation schemes that enforce total orders fail to capture (Dewancker et al., 2016). Note that this also happens in scenarios where, unlike above in Contribution 10 and 11, all criteria are cardinal. In fact, under natural axioms it is impossible to always consolidate a collection of total orders into a single total order without contradiction (Arrow, 1950), see also the little excursus to social choice theory in Section 2.2.

Our approach circumvents this dilemma by assigning to each test function a *partial* order on the set of optimizers, explicitly allowing incomparability, and preserving the ordinal structure induced by multi-criterion performance. A benchmarking suite then yields a collection of such partially ordered sets (posets). To evaluate these posets, we employ the union-free generic (ufg) depth function (Blocher et al., 2023), a general depth function that can be applied to posets.

The ufg depth function provides us with a description of the distribution of order structures. Specifically, applying ufg depth to the posets generated by a benchmark suite reveals which test functions produce well-supported, central performance orderings and which produce outliers (see Appendices C and D in Contribution 12), thereby illuminating the diversity of problem behaviors represented in the suite. In this way, our descriptive perspective supports principled curation and analysis of benchmarks. We can identify representative tasks and diagnose atypical ones. Ultimately, the goal is to understand a landscape of relative relations rather than declaring a universal winner. While computational scalability and statistical inference for posets remain open challenges,[3] this approach establishes a principled foundation for multi-criterion evaluation and motivates further integration of relational perspectives into the study of optimizers, see outlook and perspectives discussed further below.

Empirical studies on deep learning optimizers (DeepOBS) (Schneider et al., 2019), black-box optimization benchmarks (BBOB) (Hansen et al., 2010), and multi-objective evolutionary algorithms (Wu et al., 2023) illustrate the method's insights. In DeepOBS, one test function (training an LSTM on Leo Tolstoi's *War and Peace*) produces a highly atypical poset in which vanilla stochastic gradient decent outperforms Adam, suggesting its limited representativeness for the suite. In BBOB, depth values are uniformly low, indicating substantial diversity in optimizer orderings, while central posets highlight consistent dominance of quasi-Newton methods such as BFGS. For the multi-objective evolutionary algorithms, our depth analysis reveals that the

[1] Recall their Definition from Section 2.2: A preorder is a binary relation $R \subseteq M \times M$, $M \neq \emptyset$, if $(a, a) \in R$, (*reflexive*) and $(a, b), (b, c) \in R \Rightarrow (a, c) \in R$ (*transitive*).

[2] For the sake of completeness, a partial order p is a subset of $M \times M$ which is reflexive (for all $y \in M$ we have $(y, y) \in p$), antisymmetric (if $(y_1, y_2) \in p$ with $y_1 \neq y_2$ then $(y_2, y_1) \notin p$) and transitive (if $(y_1, y_2), (y_2, y_3) \in p$ then $(y_1, y_3) \in p$). A partial order that is strongly connected (for all $y_1, y_2 \in M$ either $(y_1, y_2) \in p$ or $(y_2, y_1) \in p$ holds) is called a *total/linear order*.

[3] With first steps taken by Blocher and Schollmeyer (2024) recently.

newly proposed algorithms may dominate in representative tasks while underperforming in some rather outlying ones, challenging claims of superiority made by Wu et al. (2023), see Rodemann and Blocher (2024, Appendix D) for details.

Overall, our goal is not to declare a single "best" optimizer (that potentially hides some aspects of the multidimensional outcomes) but to understand the range and structure of comparative behaviors. Moreover, depth-based analyses can guide benchmark design by identifying redundant or unrepresentative test functions and quantifying the diversity of ranking patterns. This can be seen as a substantial contribution to efficient training data selection in overall data-centric machine learning, see Section 3.3.1.

8.2. Towards Better Open-Ended Text Generation: A Multicriteria Evaluation Framework (Contribution 13)

Contribution 13

Esteban Garcés Arias, Hannah Blocher, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher (2025). "Towards Better Open-Ended Text Generation: A Multicriteria Evaluation Framework". In *ACL Workshop on Generation, Evaluation and Metrics (GEM)*. Association for Computational Linguistics, 631–654.

Echoing the descriptive approach in Contribution 12 (see previous Section 8.1) but shifting the application, Contribution 13 (Garcés Arias et al., 2025b) tackles the problem of comparing *decoding methods* (:= model × decoding strategy × hyperparameters) for open-ended text generation when quality is inherently multidimensional. To measure this multidimensional quality, we focus on three automatically computable criteria for generated text: (semantic) coherence, diversity, and generation perplexity. They capture the core trade-offs in open-ended generation and remain compatible with instance-wise aggregation.[4] The result is a unified framework that matches benchmarking *use cases* (what decision the idealized practitioner wants to make) from benchmarking *machinery* (how evidence across metrics is combined).

Elaborating in more detail, this paper takes a benchmarking *practitioner*'s perspective, thus partly relying on similar methods as in Contribution 11. In particular, we differentiate two benchmark scenarios and associated goals in natural language processing: *Scenario 1* seeks an *ordinal* ranking of methods for a given content domain, whereas *Scenario 2* demands a *cardinal* assessment that quantifies how much one method outperforms another.

To motivate *Scenario 1*, consider a practitioner deploying open-ended generation for a specific application (e.g., a customer-support bot). Their operational need is to *choose* among available methods for that scenario; hence the metric values are a *means* to the end of producing a *relative* ranking. In this setting, collapsing multiple metrics into an ordinal preference is appropriate, provided we do not force comparability that is not supported by the evidence.

By contrast, researchers *designing* new decoding methods make up *Scenario 2*. They need to know not only which method wins, but by how much. This cardinal signal can guide development of new methods or give structural insights on hyperparameter tuning.

[4] Distributional metrics like MAUVE (Pillutla et al., 2021) are deliberately set aside here.

For Scenario 1 (relative ranking), we offer two ways forward. First, an extended Bradley–Terry model (Bradley and Terry, 1952; Davidson, 1970) turns pairwise wins, ties or losses between methods into classic "worth" parameters and, by construction, a total order. This is attractive computationally and yields interpretable dominance tendencies, but it collapses incomparabilities and rests on independence assumptions across pairwise comparisons. Also, it does not provide a description of the distributions of relations. Second, to preserve ordinal structure without enforcing comparability, each prompt-level set of pairwise results is treated as a partial order and analyzed via the union-free generic depth (ufg-depth) introduced for poset-valued data, similar to Contribution 12.

As was outlined in Section 8.1, ufg-depth identifies the most supported ordering patterns (central posets) and least supported ones (outlier posets) across prompts. Again, this aims at describing the distribution of ranking structures rather than aggregating them into a single consensus order. On a large-scale[5] study, the most central structure on WikiText-103 (Merity et al., 2016) strikingly turns out to be the *empty* dominance relation among the top four contenders. This results shows that many method comparisons are genuinely incomparable once multiple criteria need to be taken into account. Generally, the price for being informative and not assumption heavy is computational: exact ufg-depth scales poorly, which motivates a restriction to a smaller method subset for the poset analysis in practice.

For Scenario 2 (cardinal assessment), the paper introduces *Q*Text*, a single-score summary that harmonizes the three metrics through normalization, weighting, and Gaussian penalty terms that softly pull each metric toward target values and discourage extremes (e.g., repetitive yet coherent text or highly diverse but incoherent text). The nine parameters (three weights, three targets, three penalties) are selected to maximize (Spearman) correlation alignment with public human ratings, yielding a score that is both data-driven and directly optimizable. Q*Text provides a straightforward total ordering.

Empirically, Q*Text appears to favor balanced sampling configurations (e.g., moderate k/p or top-k with moderate temperature) and systematically penalizes degenerate settings (e.g., beam search with long beams or extreme nucleus/top-k values). Human-written references sit near the top of the induced scale, and smaller models with well-chosen decoding can approach larger models' performance, which constitutes an insightful piece of evidence for the value of tuning decoding over model size alone.

8.3. Statistical Multicriteria Evaluation of LLM-Generated Text (Contribution 14)

Contribution 14

Esteban Garcés Arias, Hannah Blocher, Julian Rodemann, Matthias Aßenmacher, Christoph Jansen (2025). Statistical Multicriteria Evaluation of LLM-Generated Text. *arXiv preprint arXiv:2506.18082* (last accessed October 16 2025). In *International Natural Language Generation Conference (INLG)* (forthcoming).

[5] Made up of six language models, five decoding strategies with 59 hyperparameter settings, three datasets and > 1.8 million continuations.

8.3 Statistical Multicriteria Evaluation of LLM-Generated Text (Contribution 14)

Large language models (LLMs) have become increasingly popular. Assessing the quality of their outputs, however, remains a core challenge; namely, a *benchmarking* challenge for both end-users (Scenario 1) and model developers (Scenario 2), just like outlined above in Section 8.2. Many evaluations still use single measures or simple averages. Such methods miss the fine trade-offs between coherence, diversity, fluency, and other facets of quality.

Contribution 14 (Garcés Arias et al., 2025a) adapts the principled framework for benchmarking inference based on Generalized Stochastic Dominance (GSD) from Contribution 10 and 11. The raison d'être is threefold: First, single-metric evaluation is inadequate. Second, cardinal automatic scores clash with ordinal human judgments. Third, most setups lack proper inferential guarantees.

Applied to LLMs with common decoding strategies and human text, Contribution 14 finds statistically significant differences. It also remains valid when sampling is not perfectly *i.i.d.*, which is—as was argued in Chapter 1—crucial in real-world applications of LLM comparisons. In fact, Madaan et al. (2024) demonstrate that minor variations in initialization or sampling can alter rankings of LLMs substantially.

Elaborating in a bit more detail, we build on both the decision-theoretic foundations of generalized stochastic dominance (GSD) from Contribution 10 in Section 7.1 and the GSD-front methodology for benchmarking from Contribution 11 presented in Section 7.2. In this way, we carry the GSD-methodology into open-ended text generation, which allows us to move from description (Contributions 12 and 13, see Sections 8.1–8.2) to *inference*. The key idea is to compare decoding strategies across multiple quality dimensions—combining a cardinal automatic score (Q*Text) with ordinal human ratings—*without* imposing ad-hoc weights or collapsing scales. Technically, Contribution 14 adopts the GSD-front as the minimal set of non-dominated strategies under all utility representations consistent with a mixed-scale preference system, see Equations 7.6 and 7.7 above. This is based on two relations on quality scores obtained for each prompt: an ordinal dominance relation R_1 and an intensity relation R_2 for cardinal dimensions, see Equations 7.3 through 7.4.

Further building on Contributions 12 and 13 from Sections 8.1 and 8.2, we can equip the GSD-front with hypothesis tests and robustness analyses against deviations from *i.i.d.* assumptions. This extends the poset-centric, descriptive lens of Section 8.1 and the two-scenario framework of Section 8.2 with statistical guarantees suitable for principled benchmarking analyses.

Experimentally, Contribution 14 evaluates five common decoding strategies (beam search, contrastive search, temperature, top-k, nucleus) against human continuations on WikiText/WikiNews prompts. Quality is measured by Q*Text—summarizing perplexity, coherence, and diversity via normalization—and two independent 5-point ratings on Likert scales by two humans.[6] Average scores place human continuations at the top, while sampling-based decoders (top-p, top-k, temperature) generally outperform deterministic ones (contrastive, beam).

Conceptually, Contribution 14 completes the bridge between Contribution 12 (Section 8.1) that advocated descriptive, incomparability-aware summaries (depth function for partial orders) and Contribution 13 (Section 8.2) which separated ordinal-vs.-cardinal *use cases* (Scenario 1 vs. Scenario 2).

[6] Among the authors of Contribution 14.

8.4. Outlook and Perspectives

Practically, the key takeaways from Chapter 8 can be summarized as follows: use ufg-depth when characterizing distributions of order structures (Section 8.1); use Bradley–Terry or Q*Text for quick total orders (Section 8.2); and use the GSD-front (Section 8.3) with our permutation tests when you need robust, multicriteria decisions that generalize beyond a finite benchmarking suite at hand. Overall, the result is an end-to-end toolkit both for 1) curating and improving benchmarks from a designer perspective and 2) selecting algorithms like optimizers (Section 8.1) or decoding strategies (Section 8.3), and making robust, statistically valid claims about superiority.

The language model application of the GSD-based benchmarking methodology in Contribution 14 directly motivates another example of the regression-style GSD-extensions discussed in Section 7.3. As touched upon in Chapter 1, (small (Belcak et al., 2025)) language models are becoming increasingly popular in the form of AI agents within the emerging field of agentic AI (Acharya et al., 2025; Sapkota et al., 2025). Recall from Chapter 1 that agentic AI systems are defined as systems that pursue goals with a higher degree of autonomy and lesser degree of human oversight than classical machine learning applications: Agentic AI systems plan, decompose tasks into subgoals, act and observe in an environment, and adapt using feedback, rather than merely reacting to instructions (cf. ibid.). In order to benchmark AI agents, it is naturally to differentiate between tasks.

As a stylized example consider the use of agentic AI within a university travel department. The travel office—hypothetically, of course—deploys several planning agents that operate in the same environment. Each instance corresponds to a travel request by the university's employees and is labeled by task subtype (e.g., conference attendance, research visit, fieldwork, etc.) and covariates such as traveler role, budget constraints, visa complexity, and carbon footprint targets.

A stratified GSD analysis within each subtype can reveal stable dominance relations with respect to multiple criteria that potentially have varying scales of measurement—like, for instance, human-rated traceability and transparency as well as saved costs or execution time. For example, a policy-aware agent may dominate for short-notice international conferences, whereas a sustainability-weighted agent may dominate for domestic trips under binding CO_2 consumption constraints.

A regression-type GSD extension, along the lines of future work outlined in Section 7.3 would look like the following. In this context, it would allow for estimating pairwise dominance probabilities between AI agents as a function of the covariates, yielding task-conditioned rankings along with a provable uncertainty quantification through hypothesis tests. Both approaches would render the GSD framework task-aware. It would support statements that are specific to pre-defined travel scenarios.

On a different note, there is quite some potential in using our descriptive analysis tools derived in Contribution 13 and especially the specific depth function in Contribution 12 to design new benchmark suites. Loosely inspired by survey design or sampling theory (Section 3.2), we might want to subset a benchmark suite to increase runtime efficiency (critical in practical benchmarking scenarios) for an existing, fixed set of algorithms. The ufg-depth, as we laid out in Contribution 12, can help here. Specifically, it can inform us about the centrality of the orderings produced by specific datasets (statistical units) in the benchmark suite. A benchmark suite designer can exploit this information to adapt the suite by, e.g., removing datasets that produce same or similar orderings or outlying ones.

However, in this case, the inferential guarantees can deteriorate. The changed sample of datasets might be more informative with respect to the algorithms and criteria considered, but that does not say much about how informative it is with respect to the actual *population* of datasets.[7] As touched upon in Section 7.3, Part III of this dissertation can come to the rescue here, bridging the two main parts of this dissertation.

If the initial sample of datasets could be assumed to be *i.i.d.*, the reciprocal learning framework (Chapter 4) can be helpful to derive inferential guarantees. In particular, the generalization bounds we proved in Contribution 8 can be applicable. Recall that they require—besides initial *i.i.d.* distribution—assumptions on the *mechanism* of sample change (namely, Lipschitz-continuity) only, not on the actual distribution of altered data. This is a very natural perspective here, since the benchmark suite designer can control the way the suite changes directly, while assumptions about the (unobserved) distribution of altered or modified samples appear much stronger and harder to specify. Please refer to Section 7.3 for a more detailed sketch of this avenue for future work.

[7] Recall our statistical perspective on benchmarking sketched at the very beginning (second paragraph) of Chapter 7.

Part V.

Conclusion

9. Limitations and Future Work

Invoking Fisher Box's metaphor from Section 3.1 again, this dissertation aimed to enhance our "skillful interrogation of Nature" (Fisher Box, 1978) by acknowledging that we actively shape what we observe rather than passively perceive it. Specifically, the thesis answered the question how much an observer can still learn from observations that depend (in different ways) on the very observer itself. In this spirit, the dissertation carried implications for how we conceptualize machine learning research itself. As discussed in Chapter 1, emerging fields like agentic AI (Belcak et al., 2025) and performative prediction (Perdomo et al., 2020; Hardt and Mendler-Dünner, 2025) underscore that feedback loops between models and data—whether at the sample or population level, see Figure 1.2—are no longer exceptional edge cases but central features of deployed systems. By developing statistical tools that remain valid under such reciprocity, this dissertation contributes to foundations for reliable AI systems operating with reduced human oversight. Generally, the data-centric perspective advocated here aligns with growing empirical recognition (Zha et al., 2025; Oala et al., 2024) that improving data quality, curation, and acquisition may offer greater returns than solely focusing on model architectures.

There are some more avenues for further work, most of which already have been touched upon in Parts III and IV. On the algorithmic side, theory-informed data regularization could suggest new families of stable reciprocal learning methods. This is a direct consequence of the unification of active learning, self-training, boosting, bandits, Bayesian optimization and superset learning under the joint decision-theoretic superstructure of reciprocal learning. By discovering symmetries between ERM and data selection within reciprocal learning, we found that regularizing ERM (as is customary) is not sufficient for convergence. Additionally regularizing data selection, however, was found sufficient (besides general technical assumptions) for convergence at linear rates, see Theorems 1 and 3 in Contribution 1. While convergence, as discussed in Chapter 4, can say little about how well reciprocal learning methods generalize *per se*, we did find a connection between the sample adaptation's Lipschitz constant (which is i.a. governed by data regularization) and reciprocal learning's generalization bounds. On the more principled side, the ergodicity dimension of uncertainty categorization (Section 3.2) deserves deeper investigation: distinguishing uncertainties that average away along trajectories from those that persist may prove more fruitful than the aleatoric-epistemic dichotomy in non-static, sequential learning settings. To the best of my knowledge, such a distinction has not yet been axiomatically explored within the uncertainty quantification literature.

Several limitations warrant acknowledgment. They have been mentioned within the respective contributions already. Moreover, the "Outlook and Perspectives" Sections in Parts III and IV demonstrated how these limitations delineate clear avenues for future research. Nevertheless, a few central drawbacks and restrictions of this thesis' contributions shall be emphasized once more. Due to its central role in Part III, some special attention is paid to the framework of reciprocal learning.

First, the convergence results for reciprocal learning in Contribution 1 require strong convexity of the loss function, precluding direct application to modern models like multi-layer neural

networks. However, Contribution 8 partly addresses this gap, see Section 7 in Contribution 8 for a discussion. Moreover, the majority of our results in both Contribution 1 and 8 require the sample adaptation function to be Lipschitz, which again precludes several algorithms. Lipschitz continuity of sample adaptation might be perceived as a strong requirement. However, I would like to emphasize the object of this requirement rather than its subjective strength or weakness: It refers to an object we have full control over, namely the learning algorithms themselves. So instead of hypothesizing about an unknown population's distribution, our conditions address the concrete algorithm at hand. Practically, this means they are verifiable. I shall note in passing that this is also why the Lipschitz conditions can be seen as a feature rather than a bug: They can serve as design principles for (novel or modified) reciprocal learning algorithms that shall converge or generalize well. Moreover, while we established generalization bounds on learning from self-selected data in under quite general assumptions (Contribution 8), tighter bounds may be attainable under specific distributional assumptions about newly acquired data—for instance, under known covariate, target, or conditional shifts, as noted in Section 5.

Second, our GSD-based benchmarking methodology (Contributions 10–11) largely ignores dataset-level metadata that could inform a stratified analyses, as discussed in Section 7.3. Third, computational scalability remains challenging for some methods like, e.g., the ufg-depth for partial orders (Contribution 12) and the credal set updates in robust pseudo-label selection (Contributions 5–6). The latter especially scale badly with the dimension of the covariates (features) and thus become expensive for applications on image data.

On a more principled level, there is a structural limitation of construing data as a process as opposed to a fixed snippet of reality. Explicitly modeling a data lifecycle potentially requires more assumptions about parts of it that are unknown to the scholar, as compared to a static one-time sampling scenario, which only requires assumption about a fixed population, a fixed sample and how they relate. Trivially, extending the focus of an analysis can require more assumptions. If made explicit, however, the scientific conclusions drawn from the analysis can still be regarded more meaningful than the ones obtained from a narrow analysis.

This is somewhat reminiscent of early critique of causal modeling (as pioneered by Rubin (1974), Pearl (2010) and others) from the statistical community. Causal modeling also requires the scholar to explicitly state assumptions about the causal mechanisms underlying the data generating process. A simple correlation analysis does not require these assumptions. The latter's results, however, are of much narrower scope, which can tempt practitioners to over-interpret them.

This dissertation by and large argued that extending the scope of an analysis (from given data to data lifecycles) can be fruitful. The dissertation demonstrated that many modern data-centric machine learning setups involve explicit information about sample adaptation, thus *not* requiring the scholar to make additional assumptions. However, in the case when increasing the set of assumptions is required (due to e.g., closed source language models), this should be made explicit and taken into account when deriving conclusions from the analysis.

10. Concluding Remarks

At this dissertation's outset I argued that data is not merely observed but shaped, i.e., selected, curated, and transformed by human choices and methods that involve specific assumptions. This dissertation's contributions 1 through 14 operationalized this stance: They embedded data selection into decision theory, adopted imprecise-probability formalisms to account for complex uncertainty, and quantified how algorithmic procedures feed back into the very samples on which they are trained and tested.

By construing data as a dynamic process of active involvement rather than a passive given, Part III revealed feedback loops that affect generalization and thus statistical validity. Furthermore, Part III demonstrated that numerous popular methods, ranging from Bayesian optimization (Section 4.3) to self-training (Section 4.4), constitute instances of reciprocal learning (Contribution 1), where models iteratively self-select training data. Through a decision-theoretic embedding (Section 2.2) and statistical analysis, we established convergence guarantees (Contribution 1), generalization bounds (Contribution 8), and debiasing procedures (Contribution 9) for learning with (algorithmically) self-selected samples.

Part IV shifted the attention from training to testing data, developing methodology based on Generalized Stochastic Dominance (GSD) (Contributions 10 and 11) for statistically valid multicriteria benchmarking that respects mixed scales of measurement while quantifying robustness under distributional deviations from the idealized and static *i.i.d.* sampling assumption. Together, these contributions advance both reliability and trustworthiness of data-centric machine learning by explicitly accounting for uncertainties arising from how data is collected, selected, and evaluated—thus illuminating what Chapter 1 called "the dark side of the moon."

The decision-theoretic perspective emerged as a formal thread connecting all contributions across Part III and IV. Section 2.2 set the stage by recasting Abraham Wald's "zero-sum game against nature" (Wald, 1949) as a sequential, *nested* decision problem. The dissertation did cast both data selection (e.g., pseudo-label selection in semi-supervised learning (Contributions 4–6)) and algorithm selection (Contributions 10–14) as decision problems. The synergy between Parts II and III, as outlined in Section 7.3, suggests promising future directions: applying generalization bounds for reciprocal learning to adaptive benchmark suite curation, or extending GSD-based inference to task-conditioned comparisons via regression-style analyses.

The contributions of this cumulative dissertation were presented in Parts III and IV and can be summarized as follows.

Part III of this dissertation offered a fresh perspective on *training* data selection, which makes up a considerable share of data-centric machine learning methods (Section 3.3). It heavily relied on concepts from experimental design (Section 3.1), sampling theory (Section 3.2) and, most crucially, on a decision-theoretic embedding (Section 2.2) of data selection, in order to model model-data feedback loops (see Chapter 1). This embedding hinged on the strikingly simple insight that not only the selection of models, but also the selection of data constitutes a decision problem. Unlike standard model-data dichotomizations, it allowed us to explicitly account for

feedback loops, arising when the selection of samples depends on the model, see Figure 1.2 in Part II.

Chapter 4 turned to a peculiar, yet apparently quite common special case of training data selection, where such feedback loops occur: algorithmic *self*-selection of training data. Chapter 4 demonstrated that a wide range of learning paradigms, in fact, after an initial fit to provided data, *self*-select data in subsequent rounds informed by model fits. This is tangential to the statistical literature on the optimal design of experiments, as sketched in Section 3.1. Examples of such learning paradigms comprise bandits, Bayesian optimization, superset learning, active learning or self-training in semi-supervised learning, to name only a few. We connected the dots between these methods by subsuming all of them in a unified framework using the language of decision theory (Section 2.2). As learning goes both ways here, we called this framework *reciprocal learning*, which is also the programmatic title of Contribution 1 and the whole Chapter 4. This unification set the stage for several contributions to three specific instances of reciprocal learning: Bayesian optimization (Section 4.3), self-training in semi-supervised learning (Section 4.4) and superset learning (Section 4.5).

Section 4.3 entailed both a principled theoretical contribution to reciprocal learning's special case of Bayesian optimization (Section 4.3.1) and an applied one (Section 4.3.2). Section 4.4 turned to self-training in semi-supervised learning setups. Specifically, Section 4.4.1 proposed Bayes-optimal pseudo-label selection in self-training, Section 4.4.2 robustified this method along several dimensions and Section 4.4.3 made one of these robustifications computationally feasible and implemented it. Section 4.5 studied yet another special case of reciprocal learning, namely superset learning, by proposing a levelwise disambiguation procedure within this paradigm.

Chapter 5 took a step back and studied the bigger picture of reciprocal learning by answering the fundamental question of how well all these reciprocal learning methods can generalize from their self-selected samples. Chapter 6 added a constructive perspective by proposing a method to improve generalization and inference from non-*i.i.d.* samples (like the self-selected ones in reciprocal learning) for the concrete case of decision trees and random forests. This is achieved by bridging sampling theory, as introduced in Section 3.2, and tree learning.

Part IV of this dissertation addressed data's second pivotal role in data-centric machine learning: testing data, in particular for benchmarking setups. Construing testing data as a *given thing* (see Chapter 1) has proven misleading in a similar way as for training data in Part III. However, Part IV did not require the explicit study of feedback loops as in Part III. Instead, it investigated how the selection of testing data affects the scholar's conclusions in a benchmarking scenario. In particular, it quantified how deviations from the idealized and static *i.i.d.* assumption of testing datasets affected the results of multitask (inducing multicriteria) benchmarking.

Chapter 7 laid the theoretical groundwork for benchmarking algorithms with respect to multiple criteria. Similar to Chapter 4 in Part III on reciprocal learning, it benefited from a decision-theoretic perspective. Contrary to Chapter 4, however, it did not embed data selection, but algorithm selection into decision theory. Specifically, Section 7.1 modeled any algorithm's performance on multiple tasks (and respective metrics) by multivariate random variables. To account for metrics with different scales (for instance, ordinal ones like human scores for generated text on a Likert scale and cardinal ones like real-valued text coherence metrics), these random variables map to so-called *preference systems*, allowing for a representation of both the available ordinal and cardinal information. Section 7.2 extended this framework and propose an information-efficient improvement of the popular Pareto-front for multi-criteria benchmarking. Moreover, it

applied this framework to the concrete problem of comparing classifiers with respect to differently scaled metrics. By using imprecise probabilities to account for epistemic uncertainty, as touched upon in Section 2.1, we were further able to quantify how robust such benchmarking results are with respect to the selection of testing datasets.

Chapter 8 did expand the focus from benchmarking classifiers to benchmarking optimizers (Section 8.1) and large language models (Sections 8.2 and 8.3). While Section 8.1 and 8.2 used depth functions for partial orders to analyze benchmarking results in a descriptive way, Section 8.3 also allowed for inferential statements, relying on the framework from Section 7.2. In this way, it completed the bridge forged between foundational statistical methodologies with philosophical underpinnings described in Chapter 1 on the one hand and modern problems of benchmarking state-of-the-art large language models on the other hand.

References

Abellán, J. and Klir, G. J. (2005). Additivity of uncertainty measures on credal sets. *International Journal of General Systems*, 34(6):691–713.

Abellán, J. and Gómez, M. (2006). Measures of divergence on credal sets. *Fuzzy Sets and Systems*, 157(11):1514–1531.

Abellán, J., Klir, G. J., and Moral, S. (2006). Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, 35(1):29–44.

Abellán, J., Mantas, C., Castellano, J., and Moral-Garcia, S. (2018). Increasing diversity in random forest learning algorithm via imprecise probabilities. *Expert Systems with Applications*, 97:228–243.

Acharya, D. B., Kuppan, K., and Divya, B. (2025). Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access*, 13:18912–18936.

Adeoye, J., Hui, L., and Su, Y.-X. (2023). Data-centric artificial intelligence in oncology: a systematic review assessing data quality in machine learning models for head and neck cancer. *Journal of Big Data*, 10(1):28.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. (2024). Self-consuming generative models go MAD. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Alobaidi, M. H., Meguid, M. A., and Zayed, T. (2022). Semi-supervised learning framework for oil and gas pipeline failure detection. *Scientific Reports*, 12(1):13758.

Alquier, P. (2020). Approximate Bayesian inference. *Entropy*, 22(11):1272.

American Statistical Association (2012). Statistics. ASA Website statement. https://www.amstat.org/asa-newsroom (last accessed October 27 2025).

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Anjanapura Venkatesh, A. K., Rana, S., Shilton, A., and Venkatesh, S. (2022). Human-AI collaborative Bayesian optimisation. In *Neural Information Processing Systems*, volume 35, pages 16233–16245.

Anjanapura Venkatesh, A. K., Shilton, A., Gupta, S., Ryan, S., Abdolshah, M., Le, H., Rana, S., Berk, J., Rashid, M., and Venkatesh, S. (2025). Accelerated experimental design using a human–AI teaming framework. *Knowledge-Based Systems*, 315:113138.

Anscombe, F. J. and Aumann, R. J. (1963). A definition of subjective probability. *The Annals of Mathematical Statistics*, 34:199–205.

Antonucci, A., Cattaneo, M., and Corani, G. (2011). Likelihood-based naive credal classifier. In *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 21–30.

Antonucci, A. and De Campos, C. P. (2011). Decision making by credal nets. In *Third International Conference on Intelligent Human-Machine Systems and Cybernetics*, volume 1, pages 201–204. IEEE.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks*, pages 1–8. IEEE.

Ardalani, N., Guyon, I., Gürel, N. M., Lawrence, N., Vanschoren, J., and Zhang, C. (2024). Data-centric machine learning research (DMLR). *Journal of Data-centric Machine Learning Research (DMLR)*, 1. Inaugural volume.

Arens, P., Quirk, D. A., Pan, W., Yacoby, Y., Doshi-Velez, F., and Walsh, C. J. (2025). Preference-based assistance optimization for lifting and lowering with a soft back exosuit. *Science Advances*, 11(15).

Aristotle (350 BC/1907). De anima (on the soul), book iii. Translated by J. A. Smith. https://psychclassics.yorku.ca/Aristotle/De-anima/de-anima3.htm. (Last Accessed October 27 2025).

Arrow, K. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, 58:328–346.

Atkinson, A. C. and Doney, A. N. (1992). *Optimum Experimental Designs*. Oxford University Press.

Augustin, T. (1999). Globally least favorable pairs and Neyman–Pearson testing under interval probability. In *International Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, pages 15–23.

Augustin, T. (2001). On decision making under ambiguous prior and sampling information. In *International Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, pages 9–16.

Augustin, T. (2002a). Expected utility within a generalized concept of probability—a comprehensive framework for decision making under ambiguity. *Statistical Papers*, 43(1):5–22.

Augustin, T. (2002b). Neyman–Pearson testing under interval probability by globally least favorable pairs: Reviewing Huber–Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, 105(1):149–173.

Augustin, T. (2003). On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view—a cautionary note on updating imprecise priors. In *International Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, pages 31–45.

Augustin, T. (2004). Optimal decisions under complex uncertainty–basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics*, 84(10-11):678–687.

Augustin, T. (2005). Generalized basic probability assignments. *International Journal of General Systems*, 34(4):451–463.

Augustin, T., Coolen, F. P., de Cooman, G., and Troffaes, M. C. M., editors (2014a). *Introduction to Imprecise Probabilities*. Wiley, Chichester.

Augustin, T. and Hable, R. (2010). On the impact of robust statistics on imprecise probability models: a review. *Structural Safety*, 32(6):358–365.

Augustin, T. and Jansen, C. (2023). Decision theory / advanced decision theory. Lecture slides, Ludwig-Maximilians-Universität München. Summer Term 2023, accessed via Moodle.

Augustin, T. and Schollmeyer, G. (2021). Comment: on focusing, soft and strong revision of Choquet capacities and their role in statistics. *Statistical Science*, 36(2):205–209.

Augustin, T., Walter, G., and Coolen, F. P. (2014b). Statistical inference. In Augustin, T., Coolen, F. P. A., de Cooman, G., and Troffaes, M. C. M., editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley Online Library.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Bailie, J. and Derr, R. (2025). Property elicitation on imprecise probabilities. *arXiv preprint arXiv:2507.05857 (last accessed October 27 2025)*.

Bailie, J. and Gong, R. (2023). Differential privacy: general inferential limits via intervals of measures. In *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 11–24.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *Annals of Statistics*, 33:1497–1537.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Bates, S., Jordan, M. I., Sklar, M., and Soloff, J. A. (2022). Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812 (last accessed October 27 2025)*.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50:1029–1041.

Bechavod, Y., Ligett, K., Wu, S., and Ziani, J. (2021). Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1234–1242.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., and Molchanov, P. (2025). Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153 (last accessed October 29 2025)*.

Ben-Baruch, E., Botach, A., Kviatkovsky, I., Aggarwal, M., and Medioni, G. (2024). Distilling the knowledge in data pruning. *arXiv preprint arXiv:2403.07854* (last accessed October 12 2025).

Benavoli, A., Azzimonti, D., and Piga, D. (2021). A unified framework for closed-form nonparametric regression, classification, preference and mixed problems with skew Gaussian processes. *Machine Learning*, 110(11):3095–3133.

Benavoli, A., Corani, G., Demšar, J., and Zaffalon, M. (2017). Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36.

Benavoli, A. and Zaffalon, M. (2015). Prior near ignorance for inferences in the k-parameter exponential family. *Statistics*, 49(5):1104–1140.

Bengs, V., Hüllermeier, E., and Waegeman, W. (2022). Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Neural Information Processing Systems*, volume 35, pages 29205–29216.

Bengs, V., Hüllermeier, E., and Waegeman, W. (2023). On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning*, pages 2078–2091.

Benjamini, Y. (2020). Selective Inference: The Silent Killer of Replicability. *Harvard Data Science Review*, 2(4). https://hdsr.mitpress.mit.edu/pub/l39rpgyc.

Benjamini, Y. and Braun, H. (2002). John W. Tukey's contributions to multiple comparisons. *The Annals of Statistics*, 30(6):1576–1594.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* Springer., 2nd edition.

Bertrand, Q., Bose, J., Duplessis, A., Jiralerspong, M., and Gidel, G. (2024). On the stability of iterative retraining of generative models on their own data. In *International Conference on Learning Representations (ICLR).* OpenReview.net.

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660.

Bischl, B., Casalicchio, G., Das, T., Feurer, M., Fischer, S., Gijsbers, P., Mukherjee, S., Müller, A. C., Németh, L., Oala, L., Purucker, L., Ravi, S., van Rijn, J. N., Singh, P., Vanschoren, J., van der Velde, J., and Wever, M. (2025). OpenML: Insights from 10 years and more than a thousand papers. *Patterns*, 6(7):101317.

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Mantovani, R., van Rijn, J., and Vanschoren, J. (2021). OpenML: A benchmarking layer on top of OpenML to quickly create, download, and share systematic benchmarks. *Neural Information Processing Systems – Track on Datasets and Benchmarks.*

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., and Lang, M. (2017). mlrMBO: A modular framework for model-based optimization of expensive black-box functions. *arXiv preprint arXiv:1703.03373 (last accessed October 12 2025).*

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blocher, H. and Schollmeyer, G. (2024). Union-free generic depth for non-standard data. *arXiv preprint arXiv:2412.14745 (last accessed October 12 2025)*.

Blocher, H. and Schollmeyer, G. (2025). Data depth functions for non-standard data by use of formal concept analysis. *Journal of Multivariate Analysis*, 205:105372.

Blocher, H., Schollmeyer, G., and Jansen, C. (2022). Statistical models for partial orders based on data depth and formal concept analysis. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 17–30. Springer.

Blocher, H., Schollmeyer, G., Jansen, C., and Nalenz, M. (2023). Depth functions for partial orders with a descriptive analysis of machine learning algorithms. In *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 59–71.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*, volume 8 of *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*. Seeber, Firenze.

Bongratz, F., Golkov, V., Mautner, L., Della Libera, L., Heetmeyer, F., Czaja, F., Rodemann, J., and Cremers, D. (2024). How to choose a reinforcement-learning algorithm. *arXiv preprint arXiv:2407.20917 (last accessed October 12 2025)*.

Borda, J. C. d. (1781). Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 12.

Bordini, V. M., Destercke, S., and Quost, B. (2024). Self-learning from pairwise credal labels. In *27th European Conference on Artificial Intelligence, Including 13th Conference on Prestigious Applications of Intelligent Systems (ECAI-PAIS 2024)*.

Borji, A. and Itti, L. (2013). Bayesian optimization explains human active search. In *Neural Information Processing Systems*, pages 55–63.

Bostrom, N. (2000). *Observational Selection Effects in Science and Philosophy*. PhD thesis, London School of Economics.

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, New York, USA.

Box, G. E. and Lucas, H. L. (1959). Design of experiments in non-linear situations. *Biometrika*, 46(1/2):77–90.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. Wiley & Sons, Hoboken, NJ, 5th edition.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors (2016). *Handbook of Computational Social Choice*. Cambridge University Press.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Bronevich, A. and Klir, G. J. (2008). Axioms for uncertainty measures on belief functions and credal sets. In *NAFIPS 2008-2008 Annual Meeting of the North American Fuzzy Information Processing Society*, pages 1–6. IEEE.

Bronevich, A. G. and Rozenberg, I. N. (2019). The contradiction between belief functions: its description, measurement, and correction based on generalized credal sets. *International Journal of Approximate Reasoning*, 112:119–139.

Brown, G., Hod, S., and Kalemaj, I. (2022). Performative prediction in a stateful world. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6045–6061.

Brückner, M. and Scheffer, T. (2011). Stackelberg games for adversarial prediction problems. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 547–555.

Cabanas, R., Antonucci, A., Huber, D., and Zaffalon, M. (2020). CREDICI: A Java library for causal inference by credal networks. In *International Conference on Probabilistic Graphical Models*, volume 138, pages 597–600.

Caprio, M. (2025). The joys of categorical conformal prediction. *arXiv preprint arXiv:2507.04441 (last accessed October 12 2025)*.

Caprio, M., Dutta, S., Jang, K. J., Lin, V., Ivanov, R., Sokolsky, O., and Lee, I. (2024). Credal Bayesian deep learning. *Transactions on Machine Learning Research (TMLR)*.

Caprio, M., Sale, Y., and Hüllermeier, E. (2025). Conformal prediction regions are imprecise highest density regions. *arXiv preprint arXiv:2502.06331 (last accessed October 12 2025)*.

Carranza, Y. and Destercke, S. (2021). Imprecise Gaussian discriminant classification. *Pattern Recognition*, 112:107739.

Carrington, G. (2004). Supervision as a reciprocal learning process. *Educational Psychology in Practice*, 20(1):31–42.

Carter, B. (1983). The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 310(1512):347–363.

Cattaneo, M. (2005). Likelihood-based statistical decisions. In *International Symposium on Imprecise Probabilities and their Applications (ISIPTA)*, pages 107–116.

Cattaneo, M. E. (2013). Likelihood decision functions. *Electronic Journal of Statistics*, 7:2924–2946.

Cattaneo, M. E. (2014). A continuous updating rule for imprecise probabilities. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 426–435. Springer.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

Chakraborty, T., Wirth, C., and Seifert, C. (2025). Explainable bayesian optimization. In *World Conference on Explainable Artificial Intelligence*, pages 53–77. Springer.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304.

Chapelle, O., Scholkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press.

Chau, S. L., Schrab, A., Gretton, A., Sejdinovic, D., and Muandet, K. (2025). Credal two-sample tests of epistemic uncertainty. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 127–135.

Chauvet, G. (2015). Coupling methods for multistage sampling. *Annals of Statistics*, 43(6):2484–2506.

Chen, L., Acun, B., Ardalani, N., Sun, Y., Kang, F., Lyu, H., Kwon, Y., Jia, R., Wu, C.-J., Zaharia, M., and Zou, J. (2023). Data acquisition: A new frontier in data-centric AI. *arXiv preprint arXiv:2311.13712 (last accessed October 12 2025)*.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 24(4):586–602.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

Chernoff, H. (1987). *Sequential Analysis and Optimal Design*, volume 8 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Chhabra, A., Li, P., Mohapatra, P., and Liu, H. (2024). What data benefits my classifier? Enhancing model performance and interpretability through influence-based data selection. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Choquet, G. (1954). Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295.

Cochran, W. G. (1942). Sampling theory when the sampling-units are of unequal sizes. *Journal of the American Statistical Association*, 37(218):199–212.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley & Sons, New York, 3rd edition.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. (2020). Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Condorcet, M. J. A. N. d. C. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris.

Conitzer, V. (2010). Making decisions based on the preferences of multiple agents. *Communications of the ACM*, 53(3):84–94.

Couso, I. and Dubois, D. (2014). Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7):1502–1518.

Couso, I., Moral, S., and Walley, P. (1999). Examples of independence for imprecise probabilities. In *International Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, pages 121–130.

Cozman, F. G. (2000). Credal networks. *Artificial intelligence*, 120(2):199–233.

Croppi, F. (2021). Explaining sequential model-based optimization. Master's thesis, LMU Munich.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*, volume 2.

Davidian, M. and Louis, T. A. (2012). Why statistics? *Science*, 336(6077):12.

Davidson, R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65:317–328.

De Bock, J., De Campos, C. P., and Antonucci, A. (2014). Global sensitivity analysis for MAP inference in graphical models. In *Neural Information Processing Systems*, pages 2690–2698.

De Campos, C. P. and Antonucci, A. (2015). Imprecision in machine learning and AI. *IEEE Intelligent Informatics Bulletin*, 16(1):20–23.

De Campos, C. P. and Ji, Q. (2008). Strategy selection in influence diagrams using imprecise probabilities. In *Uncertainty in Artificial Intelligence (UAI)*, pages 121–128.

De Heide, R. (2024). A plea for a new statistical paradigm. *Nieuw Archief voor Wiskunde*, 25(3):183.

DeMets, D. L. and Lan, K. K. G. (1994). Interim analysis: the alpha spending function approach. *Statistics in Medicine*, 13(13-14):1341–1352. Discussion 1353–1356.

Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339.

Dempster, A. (1979). Review of "life and work of Ronald Fisher: R.A. Fisher. the life of a scientist." Joan Fisher Box. Wiley, New York, 1978. xiv, 512 pp.+ plates. *Science*, 203(4380):537–537.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30.

Destercke, S. (2022). Uncertain data in learning: challenges and opportunities. In *Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, pages 322–332.

Destercke, S. and Dubois, D. (2014). Other uncertainty theories based on capacities. In Augustin, T., Coolen, F. P. A., de Cooman, G., and Troffaes, M. C. M., editors, *Introduction to Imprecise Probabilities*, pages 93–113. Wiley Online Library.

Destercke, S., Dubois, D., and Chojnacki, E. (2008a). Unifying practical uncertainty representations–I: Generalized p-boxes. *International Journal of Approximate Reasoning*, 49(3):649–663.

Destercke, S., Dubois, D., and Chojnacki, E. (2008b). Unifying practical uncertainty representations. II: Clouds. *International Journal of Approximate Reasoning*, 49(3):664–677.

Destercke, S., Montes, I., and Miranda, E. (2022). Processing distortion models: A comparative study. *International Journal of Approximate Reasoning*, 145:91–120.

Dette, H. and Wied, D. (2016). Detecting relevant changes in time series models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(2):371–394.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Deville, J.-C. and Tille, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101.

Dewancker, I., McCourt, M., Clark, S., Hayes, P., Johnson, A., and Ke, G. (2016). A strategy for ranking optimization methods using multiple criteria. In *ICML Workshop on Automatic Machine Learning*, pages 11–20.

Dicke, R. H. (1957). Gravitation without a principle of equivalence. *Reviews of Modern Physics*, 29(3):363.

Dicke, R. H. (1961). Dirac's cosmology and mach's principle. *Nature*, 192(4801):440–441.

Diekmann, A., Hadjar, A., Kurz, K., Rosar, U., Wagner, U., and Westle, B. (2019). Allgemeine bevölkerungsumfrage der sozialwissenschaften allbus 2018. *GESIS Datenarchiv, Köln. ZA5270 Datenfile Version*, 2(0).

Dietrich, S., Rodemann, J., and Jansen, C. (2024). Semi-supervised learning guided by the generalized Bayes rule under soft revision. In *International Conference on Soft Methods in Probability and Statistics*, pages 110–117. Springer.

Dietrich, S., Rodemann, J., and Jansen, C. (2025). Robust pseudo-label-selection in semi-supervised learning with soft revision applied to Bayesian networks. *Poster presented at International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA). Available at https://www.researchgate.net/publication/394227190_Robust_Pseudo-Label-Selection_in_Semi-Supervised_Learning_with_Soft_Revision_applied_to_Bayesian_Networks (last accessed October 29 2025).*

Ding, Y., Arias, E. G., Li, M., Rodemann, J., Aßenmacher, M., Chen, D., Fan, G., Heumann, C., and Zhang, C. (2025). Guard: Glocal uncertainty-aware robust decoding for efficient hyperparameter-free text generation. Findings of the Association for Computational Linguistics: EMNLP 2025 (forthcoming). Association for Computational Linguistics (ACL). Available at *arXiv preprint arXiv:2508.20757* (last accessed October 29 2025).

Ding, Y., Kim, M., Kuindersma, S., and Walsh, C. J. (2018). Human-in-the-loop optimization of hip assistance with a soft exosuit during walking. *Science Robotics*, 3(15):eaar5438.

Du Bois-Reymond, E. (1872). *Über die Grenzen des Naturerkennens*. Von Veit.

Dubois, D. and Prade, H. (1988). *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638.

Efron, B. (2001). Comment on "statistical modeling: The two cultures". *Statistical Science*, 16(3):218–219. Comment on L. Breiman (2001).

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669.

Eyuboglu, S., Karlaš, B., Ré, C., Zhang, C., and Zou, J. (2022). dcbench: a benchmark for data-centric AI systems. In *Proceedings of the Sixth Workshop on Data Management for End-to-End Machine Learning (DEEM '22) at SIGMOD/PODS*. ACM.

Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. NLB, London.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33:503–513.

Fisher, R. A. (1935). *The Design of Experiments*, volume 21. Springer.

Fisher Box, J. (1978). *R. A. Fisher: The Life of a Scientist*. Wiley Series in Probability and Mathematical Statistics. Wiley & Sons, New York.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738.

French, S. and Insua, D. R. (2000). Statistical decision theory. In *Kendall's Library of Statistics*. Oxford University Press.

Fröhlich, C., Derr, R., and Williamson, R. C. (2023). Towards a strictly frequentist theory of imprecise probability. In *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 230–240.

Fröhlich, C. (2025). *Imprecise Probabilities in Machine Learning: Structure and Semantics*. PhD thesis, University of Tübingen.

Fu, S., Wang, Y., Chen, Y., Tian, X., and Tao, D. (2025). A theoretical perspective: How to prevent model collapse in self-consuming training loops. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. and Weinberger, K. Q., editors, *International Conference on Machine Learning*, pages 1050–1059.

Galilei, G. (1967). *Dialogue concerning the two world systems*. Berkeley CA: University of California Press. (Original work published in 1632). Translated by Drake, Stillman.

Gao, R. and Kleywegt, A. (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655.

Garcia-Gomez, C., Perez, A., and Prieto-Alaiz, M. (2019). A review of stochastic dominance methods for poverty analysis. *Journal of Economic Surveys*, 33(5):1437–1462.

Garcés Arias, E., Blocher, H., Rodemann, J., Aßenmacher, M., and Jansen, C. (2025a). Statistical multicriteria evaluation of llm-generated text. In *18th International Natural Language Generation Conference (INLG), forthcoming*.

Garcés Arias, E., Blocher, H., Rodemann, J., Li, M., Heumann, C., and Aßenmacher, M. (2025b). Towards better open-ended text generation: A multicriteria evaluation framework. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 631–654, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Garcés Arias, E., Rodemann, J., and Heumann, C. (2025c). The geometry of creative variability: How credal sets expose calibration gaps in language models. *arXiv preprint arXiv:2509.23088 (last accessed October 12 2025)*.

Garcés Arias, E., Rodemann, J., Li, M., Heumann, C., and Aßenmacher, M. (2024). Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15060–15080, Miami, Florida, USA. Association for Computational Linguistics.

Garnett, R. (2023). *Bayesian optimization*. Cambridge University Press.

Gelman, A. (2014). The most-cited statistics papers ever. Statistical Modeling, Causal Inference, and Social Science blog. https://statmodeling.stat.columbia.edu/2014/03/31/cited-statistics-papers-ever/ (Last Accessed October 21 2025).

GESIS (2018). Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2014. GESIS Datenarchiv, Köln. ZA5240 Datenfile Version 2.2.0, https://search.gesis.org/research_data/ZA5240?doi=10.4232/1.13141 (last accessed October 27 2025).

Gitelman, L., editor (2013). *Raw data is an oxymoron*. MIT press.

Goktas, D. and Greenwald, A. (2021). Convex-concave min-max stackelberg games. In *Neural Information Processing Systems*, pages 2991–3003.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*. University of Minnesota Press.

Goschenhofer, J. (2023). *Reducing the Effort for Data Annotation*. PhD thesis, LMU.

Grandvalet, Y. (2002). Logistic regression for partial labels. In *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'02)*, pages 1935–1941.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey methodology*. Wiley & Sons.

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., and Kauermann, G. (2025). Sources of uncertainty in machine learning–a statisticians' view. *Statistical Science (forthcoming). arXiv preprint arXiv:2305.16703 (last accessed October 12 2025)*.

Grunberg, E. and Modigliani, F. (1954). The predictability of social events. *Journal of Political Economy*, 62(6):465–478.

Grünwald, P., De Heide, R., and Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128.

Guillaume, R. and Dubois, D. (2019). A maximum likelihood approach to inference under coarse data based on minimax regret. In Destercke, S., Denoeux, T., Gil, M. Á., Grzegorzewski, P., and Hryniewicz, O., editors, *Uncertainty Modelling in Data Science*, pages 99–106. Springer.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330.

Guo, Y., Bo, D., Yang, C., Lu, Z., Zhang, Z., Liu, J., Peng, Y., and Shi, C. (2025). Data-Centric Graph Learning: A Survey . *IEEE Transactions on Big Data*, 11(01):1–20.

Gupta, S., Shilton, A., Anjanapura Venkatesh, A. K., Ryan, S., Abdolshah, M., Le, H., Rana, S., Berk, J., Rashid, M., and Venkatesh, S. (2023). BO-muse: A human expert and AI teaming framework for accelerated experimental design. *arXiv preprint arXiv:2303.01684 (last accessed October 27 2025)*.

Hacking, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis*, 79(3):427–451.

Hacking, I. (2006). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Han, Q. and Wellner, J. A. (2021). Complex sampling designs: Uniform limit theorems and applications. *Annals of Statistics*, 49(1):459–485.

Hansen, N., Auger, A., Ros, R., Finck, S., and Pošík, P. (2010). Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009. In *Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation*, pages 1689–1696.

Hardt, M. (2025). The emerging science of machine learning benchmarks. Online at https://mlbenchmarks.org (last accessed October 27 2025). Manuscript.

Hardt, M., Megiddo, N., Papadimitriou, C. H., and Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 111–122, New York, NY, USA. ACM.

Hardt, M. and Mendler-Dünner, C. (2025). Performative Prediction: Past and Future. *Statistical Science*, 40(3):417 – 436.

Harper, D. (2001). Data. In *Online Etymology Dictionary*. The Sciolist. https://www.etymonline.com/word/data (last accessed October 12 2025).

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hataya, R., Bao, H., and Arai, H. (2023). Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20555–20565.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Heckman, J. J. (2018). Selection bias and self-selection. In *The new Palgrave dictionary of economics*, pages 12130–12147. Springer.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175.

Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, 13(1):1809–1837.

Hernán, M. A., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570.

Heuer, R. J. (1999). *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency, Washington, DC. Accessed October 21, 2025.

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40(1):27–31.

Hilbert, D. (1900). Mathematische Probleme: Vortrag, gehalten auf dem internationalen Mathematiker-Kongreß zu Paris 1900. Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse (pp. 253–297).

Hilbert, D. (1902). Mathematical problems. *Bulletin of the American Mathematical Society*, 9(3):437–479.

Hodges, J. and Lehmann, E. (1952). The use of previous experience in reaching statistical decisions. *Annals of Mathematical Statistics*, 23:396–407.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Horkheimer, M. and Adorno, T. W. (2002). *Dialectic of Enlightenment: Philosophical Fragments*. Stanford University Press, Stanford.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080.

Huang, B., Yu, Y., Huang, J., Zhang, X., and Ma, J. (2024). Dca-bench: A benchmark for dataset curation agents. *arXiv preprint arXiv:2406.07275 (last accessed October 12 2025)*.

Huber, P. J. (1981). Robust statistics. *Wiley*.

Huber, P. J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, pages 251–263.

Hüllermeier, E. (2014). Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55:1519–1534.

Hüllermeier, E. and Beringer, J. (2006). Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439.

Hüllermeier, E. and Cheng, W. (2015). Superset learning based on generalized loss minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 260–275. Springer.

Hüllermeier, E., Destercke, S., and Couso, I. (2019). Learning from imprecise data: adjustments of optimistic and pessimistic variants. In *International Conference on Scalable Uncertainty Management*, pages 266–279. Springer.

Hüllermeier, E., Destercke, S., and Shaker, M. H. (2022). Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence (UAI)*, pages 548–557.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Huntley, N., Hable, R., and Troffaes, M. C. M. (2014). Decision making. In Augustin, T., Coolen, F. P. A., de Cooman, G., and Troffaes, M. C. M., editors, *Introduction to Imprecise Probabilities*, pages 190–206. Wiley.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jaffray, J.-Y. (1999). Rational decision making with imprecise probabilities. In *International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA)*, pages 324–332.

Jansen, C. (2018). *Some contributions to decision making in complex information settings with imprecise probabilities and incomplete preferences*. PhD thesis, LMU.

Jansen, C. (2025). Contributions to the decision theoretic foundations of machine learning and robust statistics under weakly structured information. Habilitation thesis, Ludwig-Maximilians-Universität München (last accessed October 28 2025), avaliable at https://arxiv.org/abs/2501.10195v1 (last accessed October 27 2025).

Jansen, C., Blocher, H., Augustin, T., and Schollmeyer, G. (2022a). Information efficient learning of complexly structured preferences: Elicitation procedures and their application to decision making under uncertainty. *International Journal of Approximate Reasoning*, 144:69–91.

Jansen, C., Nalenz, M., Schollmeyer, G., and Augustin, T. (2023a). Statistical comparisons of classifiers by generalized stochastic dominance. *Journal of Machine Learning Research*, 24(231):1–37.

Jansen, C., Schollmeyer, G., and Augustin, T. (2018). Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. *International Journal of Approximate Reasoning*, 98:112–131.

Jansen, C., Schollmeyer, G., and Augustin, T. (2022b). Quantifying degrees of E-admissibility in decision making with imprecise probabilities. In Augustin, T., Cozman, F. G., and Wheeler, G., editors, *Reflections on the Foundations of Probability and Statistics: Essays in Honor of Teddy Seidenfeld*, pages 319–346. Springer.

Jansen, C., Schollmeyer, G., Blocher, H., Rodemann, J., and Augustin, T. (2023b). Robust statistical comparison of random variables with locally varying scale of measurement. In *Uncertainty in Artificial Intelligence (UAI)*, pages 941–952.

Jansen, C., Schollmeyer, G., Rodemann, J., and Augustin, T. (2025a). Empirical decision problems. Poster presented at the International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA). Available at https://www.researchgate.net/profile/Christoph-Jansen/publication/394053214_Empirical_decision_problems/links/688727d300a2407910a4bd62/Empirical-decision-problems.pdf (last accessed October 29 2025).

Jansen, C., Schollmeyer, G., Rodemann, J., and Augustin, T. (2025b). Empirical decision theory. *Unpublished manuscript*.

Jansen, C., Schollmeyer, G., Rodemann, J., Blocher, H., and Augustin, T. (2024). Statistical multicriteria benchmarking via the GSD-front. In *Neural Information Processing Systems*, volume 37, pages 98143–98179.

Jiménez, S., Jürgens, M., and Waegeman, W. (2025). Why machine learning models fail to fully capture epistemic uncertainty. *arXiv preprint arXiv:2505.23506 (last accessed October 12 2025)*.

John, L. K. (2006). Aggregating performance metrics over a benchmark suite. In John, L. K. and Eeckhout, L., editors, *Performance Evaluation and Benchmarking*, pages 47–58. CRC Press Boca Raton.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Jones, M., Blackwell, A., Prince, K., Meakins, S., Simpson, A., and Vuylsteke, A. (2019). Data as process: From objective resource to contingent performance. In Trish Reay, Tammar B. Zilber, A. L. and Tsoukas, H., editors, *Institutions and organizations: A process view*, volume 13. Oxford University Press.

Kantorovich, L. and Rubinstein, G. (1958). On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer, New York, 2nd edition.

Kass, R. E., Tierney, L., and Kadane, J. B. (1989). Approximate Bayesian inference in conditionally independent hierarchical models. *Journal of the American Statistical Association*, 84(406):717–726.

Kauermann, G. and Küchenhoff, H. (2010). *Stichproben: Methoden und praktische Umsetzung mit R*. Springer.

Kim, K. and Demets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, 74(1):149–154.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., et al. (2023). The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140 (last accessed October 27 2025)*.

Kirchhof, M., Füger, L., Goliński, A., Dhekane, E. G., Blaas, A., and Williamson, S. (2025a). Self-reflective uncertainties: Do LLMs know their internal answer distribution? *arXiv preprint arXiv:2505.20295 (last accessed October 27 2025)*.

Kirchhof, M., Kasneci, G., and Kasneci, E. (2025b). Position: Uncertainty quantification needs reassessment for large-language model agents. *arXiv preprint arXiv:2505.22655 (last accessed October 12 2025)*.

Kiureghian, A. D. and Ditlevsen, O. (2009). Aleatoric or epistemic? Does it matter? *Structural Safety*, 31(2):105–112.

Kleinegesse, S. and Gutmann, M. U. (2021). Gradient-based Bayesian experimental design for implicit models using mutual information lower bounds. *arXiv preprint arXiv:2105.04379 (last accessed October 29 2025)*.

Kofler, E. and Menges, G. (1976). *Entscheiden bei unvollständiger Information*. Springer, Berlin.

Köhler, H. (2024). On the connection between Lp-and risk consistency and its implications on regularized kernel methods. *Journal of Machine Learning Research*, 25(213):1–33.

Kolmogorov, A. N. and Tikhomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86.

Kotz, S. and Johnson, N. L. (1992). *Breakthroughs in Statistics. Volume II: Methodology and distribution*. Springer.

Kotz, S. and Johnson, N. L. (1998). *Breakthroughs in Statistics. Volume III*. Springer.

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information*. Wiley & Sons.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 173(2):389–407.

Kreuter, F. and Valliant, R. (2007). A survey on survey statistics: What is done and can be done in stata. *The Stata Journal*, 7(1):1–21.

Krug, W., Nourney, M., and Schmidt, J. (2001). *Wirtschafts-und Sozialstatistik: Gewinnung von Daten*. Oldenbourg Verlag.

Kuleshov, V. and Precup, D. (2014). Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028 (last accessed October 12 2025)*.

Kumar, S., Datta, S., Singh, V., Singh, S. K., and Sharma, R. (2024). Opportunities and challenges in data-centric AI. *IEEE Access*, 12:33173–33189.

Kwon, Y., Wu, E., Wu, K., and Zou, J. (2023). Datainf: Efficiently estimating data influence in LoRa-tuned LLMs and diffusion models. *arXiv preprint arXiv:2310.00902 (last accessed October 12 2025)*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Neural Information Processing Systems (NeurIPS)*, pages 6402–6413.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lehmann, E. (1955). Ordered families of distributions. *Annals of Mathematical Statistics*, 26:399–419.

Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review*, 1(1).

Levi, I. (1975). On indeterminate probabilities. *The Journal of Philosophy*, 71(13):391–418.

Li, S., Wei, Z., Zhang, J., and Xiao, L. (2020). Pseudo-label selection for deep semi-supervised learning. In *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 1–5. IEEE.

Lienen, J. and Hüllermeier, E. (2021). Credal self-supervised learning. In *Neural Information Processing Systems*, pages 14370–14382.

Liese, F. and Miescke, K. (2008). *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

Lindley, D. V. (1972). *Bayesian statistics: A review*. SIAM.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(1):1–18.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318.

Liu, L. and Dietterich, T. (2012). A conditional multinomial mixture model for superset label learning. In *Neural Information Processing Systems*, volume 25, pages 548–556.

Liu, L. and Dietterich, T. (2014). Learnability of the superset label learning problem. In *International Conference on Machine Learning*, pages 1629–1637.

Liu, R. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18:405–414.

Liu, W., Zeng, W., He, K., Jiang, Y., and He, J. (2023). What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685 (last accessed October 27 2025)*.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137.

Löhr, T., Hofman, P., Mohr, F., and Hüllermeier, E. (2025). Credal prediction based on relative likelihood. *arXiv preprint arXiv:2505.22332 (last accessed October 12 2025)*.

Lu, Q., Polyzos, K. D., Li, B., and Giannakis, G. B. (2023). Surrogate modeling for Bayesian optimization beyond a single Gaussian process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11283–11296.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley, Hoboken, NJ.

Lumley, T., Shaw, P. A., and Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79(2):200–220.

MacQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297.

Madaan, L., Singh, A. K., Schaeffer, R., Poulton, A., Koyejo, S., Stenetorp, P., Narang, S., and Hupkes, D. (2024). Quantifying variance in evaluation benchmarks. In *Proceedings of the 2nd Workshop on Regulatable ML at NeurIPS 2024*. available at *arXiv:2406.10229* (last accessed October 29 2025).

Makarova, A., Usmanova, I., Bogunovic, I., and Krause, A. (2021). Risk-averse heteroscedastic Bayesian optimization. In *Neural Information Processing Systems*, volume 34, pages 17235–17245.

Malkomes, G. and Garnett, R. (2018). Automating Bayesian optimization with Bayesian optimization. In *Neural Information Processing Systems*, volume 31, pages 5984–5994.

Mangili, F. (2015). A prior near-ignorance Gaussian process model for nonparametric regression. In *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 187–196.

Mangili, F. (2016). A prior near-ignorance Gaussian process model for nonparametric regression. *International Journal of Approximate Reasoning*, 78:153–171.

Mangili, F. and Benavoli, A. (2015). New prior near-ignorance models on the simplex. *International Journal of Approximate Reasoning*, 56:278–306.

Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., and Hooker, S. (2023). When less is more: Investigating data pruning for pretraining LLMs at scale. *arXiv preprint arXiv:2309.04564 (last accessed October 27 2025)*.

Marquardt, A., Rodemann, J., and Augustin, T. (2023). An empirical study of prior-data conflicts in Bayesian neural networks. Poster presented at the International Symposium on Imprecise Probability: Theories and Applications (ISIPTA). Available at https://isipta23.sipta.org/accepted-papers/short-marquard/ (last accessed October 29 2025).

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., and Sarkar, R. (2023). Towards understanding the interplay of generative artificial intelligence and the internet. In *International Workshop on Epistemic Uncertainty in Artificial Intelligence*, pages 59–73. Springer.

Maua, D. and Cozman, F. (2020). Thirty years of credal networks: Specification, algorithms and complexity. *International Journal of Approximate Reasoning*, 126:133–157.

Maua, D. and de Campos, C. (2021). Editorial to: Special issue on robustness in probabilistic graphical models. *International Journal of Approximate Reasoning*, 137:113.

Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Gaviria Rojas, W., Diamos, S., Diamos, G., He, L., Parrish, A., Kirk, H. R., et al. (2022). Dataperf: Benchmarks for data-centric AI development. *arXiv preprint arXiv:2207.10062 (last accessed October 12 2025).*

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426 (last accessed October 27 2025).*

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. (2020). Stochastic optimization for performative prediction. In *Neural Information Processing Systems*, volume 33, pages 4929–4939.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv.1609.07843* (last accessed October 12 2025).

Miller, J., Milli, S., and Hardt, M. (2020). Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926.

Miller, J. P., Perdomo, J. C., and Zrnic, T. (2021). Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*.

Miranda, L. J. (2021). Study notes on data-centric machine learning. https://ljvmiranda921.github.io/notebook/2021/07/30/data-centric-ml/ (Last accessed October 27 2025).

Močkus, J. (1975). On Bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer.

Močkus, J. (1989). The Bayesian approach to local optimization. In *Bayesian Approach to Global Optimization*, pages 125–156. Springer.

Močkus, J., Tiesis, V., and Zilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2.

Montes, I., Miranda, E., and Destercke, S. (2020a). Unifying neighbourhood and distortion models: Part I–new results on old models. *International Journal of General Systems*, 49(6):602–635.

Montes, I., Miranda, E., and Destercke, S. (2020b). Unifying neighbourhood and distortion models: Part II–new models and synthesis. *International Journal of General Systems*, 49(6):636–674.

Moral, S. (1992). Calculating uncertainty intervals from conditional convex sets of probabilities. In *Uncertainty in Artificial Intelligence (UAI)*, pages 199–206.

Morgenstern, O. (1928). Wirtschaftsprognose: Eine untersuchung ihrer voraussetzungen und möglichkeiten [Economic forecast: An examination of its assumptions and possibilities]. *Wissenschaftstheorie in Ökonomie und Wirtschaftsinformatik, Deutscher Universitäts-Verlag, Wiesbaden*, pages 171–190.

Morgenstern, O. (1951). Abraham Wald, 1902–1950. *Econometrica*, 19(4):361–367.

Mosler, K. and Mozharovskyi, P. (2022). Choosing among notions of multivariate depth statistics. *Statistical Science*, 37:348–368.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*, volume 28, pages 10–18.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Mucsányi, B., Kirchhof, M., and Oh, S. J. (2024). Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In *Neural Information Processing Systems*, pages 50972–51038.

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

Nalenz, M. and Augustin, T. (2022). Compressed rule ensemble learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9998–10014.

Nalenz, M., Rodemann, J., and Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine Learning*, 113(6):3379–3398.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Ng, A. (2021). A chat with Andrew on MLOps: From model-centric to data-centric AI. Presentation held at www.deeplearning.ai. https://www.youtube.com/watch?v=06-AZXmwHjo (last accessed October 24 2025).

Ng, A. (2025). Data-centric AI resource hub. https://www.datacentricai.org/. (last accessed October 28 2025).

Nguyen, V.-L., Destercke, S., and Hüllermeier, E. (2019). Epistemic uncertainty sampling. In *International Conference on Discovery Science*, pages 72–86. Springer.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V., editors (2007). *Algorithmic Game Theory*. Cambridge University Press, Cambridge, UK; New York, NY.

Oala, L., Maskey, M., Bat-Leah, L., Parrish, A., Gürel, N. M., Kuo, T.-S., Liu, Y., Dror, R., Brajovic, D., Yao, X., et al. (2024). Dmlr: Data-centric machine learning research-past, present and future. *Journal of Data-centric Machine Learning Research*, 1(5).

Olson, R., La Cava, W., Orzechowski, P., Urbanowicz, R., and Moore, J. (2017). PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, 10:36.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Neural Information Processing Systems*, pages 13991–14002.

Pal, N. R., Bezdek, J. C., and Hemasinha, R. (1992). Uncertainty measures for evidential reasoning I: A review. *International Journal of Approximate Reasoning*, 7(3-4):165–183.

Pal, N. R., Bezdek, J. C., and Hemasinha, R. (1993). Uncertainty measures for evidential reasoning II: A new measure of total uncertainty. *International Journal of Approximate Reasoning*, 8(1):1–16.

Pan, I., Mason, L. R., and Matar, O. K. (2022). Data-centric engineering: integrating simulation, machine learning and statistics. challenges and opportunities. *Chemical Engineering Science*, 249:117271.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *ACM symposium on user interface software and technology*, pages 1–22.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, 40(1):75–149.

Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016 (last accessed October 27 2025)*.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Peirce, C. S. (1883). *A theory of probable inference*. Little, Brown and Co.

Peirce, C. S. (1887/2014). *Illustrations of the Logic of Science*. Open Court.

Peirce, C. S. and Jastrow, J. (1885). On small differences of selection. *Memoirs of the National Academy of Science*, 3:73–83.

Perdomo, J. (2023). *Performative Prediction: Theory and Practice*. PhD thesis, UC Berkeley.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609.

Perdomo, J. C. (2025). Revisiting the predictability of performative, social events. In *International Conference on Machine Learning (ICML)*.

Persiau, F., De Bock, J., and de Cooman, G. (2022). On the (dis) similarities between stationary imprecise and non-stationary precise uncertainty models in algorithmic randomness. *International Journal of Approximate Reasoning*, 151:272–291.

Peters, O. (2019). The ergodicity problem in economics. *Nature Physics*, 15(12):1216–1221.

Peters, O. and Adamou, A. (2018). The time interpretation of expected utility theory. *arXiv preprint arXiv:1801.03680 (last accessed October 29 2025)*.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. (2021). Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Neural Information Processing Systems*, volume 34, pages 4816–4828.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Popper, K. R. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, London, England.

Popper, K. R. (2002/1959). *The Logic of Scientific Discovery*. Routledge Classics.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raiffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory*. Boston: Harvard Business School.

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. (2024). Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601.

Ramdas, A. and Wang, R. (2025). Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1-2):1–390.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. (2020). In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Rodemann, J. (2021a). Robust generalizations of stochastic derivative-free optimization. Master's thesis, LMU Munich.

Rodemann, J. (2021b). Robust surrogate models in Bayesian optimization. Department of Statistics, Ludwig-Maximilians-Universität München (LMU); Slides available at https://www.foundstat.statistik.uni-muenchen.de/personen/mitglieder/rodemann/summer_retreat_talk_21-1.pdf (last accessed October 29 2025).

Rodemann, J. (2023a). Learning under weak supervision: Some insights from decision theory. Young Statistician Lecture Series by International Biometric Society (IBS), German region.

Rodemann, J. (2023b). Pseudo label selection is a decision problem. In *Proceedings of the 46th German Conference on Artificial Intelligence*. Springer.

Rodemann, J. (2024). Towards Bayesian data selection. *5th Workshop on Data-Centric Machine Learning Research (DMLR) at ICML 2024*. available at *arXiv:2406.12560* (last accessed October 19 2025).

Rodemann, J., Arias, E. G., Luther, C., Jansen, C., and Augustin, T. (2025a). A statistical case against empirical human-AI alignment. *arXiv preprint arXiv:2502.14581 (last accessed October 12 2025)*.

Rodemann, J. and Augustin, T. (2021). Accounting for imprecision of model specification in Bayesian optimization. Poster presented at the International Symposium on Imprecise Probabilities (ISIPTA). Available at https://isipta21.sipta.org/abstracts/1_9-67.pdf (last accessed October 29 2025).

Rodemann, J. and Augustin, T. (2022a). Accounting for Gaussian process imprecision in Bayesian optimization. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM)*, pages 92–104. Springer.

Rodemann, J. and Augustin, T. (2022b). Prior-mean-robust Bayesian optimization (PROBO). Young Statisticians Lecture Series (IBS-DR Early Career Working Group, German Region of the International Biometric Society); May 4, 2022. https://www.biometrische-gesellschaft.de/fileadmin/AG_Daten/Nachwuchs/PDFs/Abstract_YSL_Rodemann.pdf (last accessed October 29 2025).

Rodemann, J. and Augustin, T. (2024). Imprecise Bayesian optimization. *Knowledge-Based Systems*, 300:112186.

Rodemann, J., Augustin, T., and De Heide, R. (2023a). Interpreting generalized Bayesian inference by generalized Bayesian inference. Poster presented at the Thirteenth International Symposium on Imprecise Probabilities (ISIPTA). Available at https://isipta23.sipta.org/accepted-papers/short-rodemann/ (last accessed October 29 2025).

Rodemann, J. and Bailie, J. (2025). Generalization bounds and stopping rules for learning with self-selected data. *arXiv preprint arXiv:2505.07367 (last accessed October 29 2025)*.

Rodemann, J. and Blocher, H. (2024). Partial rankings of optimizers. In *International Conference on Learning Representations (ICLR), Tiny Papers Track*. OpenReview.net.

Rodemann, J., Croppi, F., Arens, P., Sale, Y., Herbinger, J., Bischl, B., Hüllermeier, E., Augustin, T., J. Walsh, C., and Casalicchio, G. (2025b). Explaining Bayesian optimization by Shapley values facilitates human-AI collaboration for exosuit personalization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 525–542. Springer. (In press).

Rodemann, J., Fischer, S., Schneider, L., Nalenz, M., and Augustin, T. (2022a). Not all data are created equal: Lessons from sampling theory for adaptive machine learning. In *International Conference on Statistics and Data Science (ICSDS)*.

Rodemann, J., Goschenhofer, J., Dorigatti, E., Nagler, T., and Augustin, T. (2023b). Approximately Bayes-optimal pseudo label selection. In *Uncertainty in Artificial Intelligence (UAI)*.

Rodemann, J., Jansen, C., and Schollmeyer, G. (2024). Reciprocal learning. *Neural Information Processing Systems*, 37:1686–1724.

Rodemann, J., Jansen, C., Schollmeyer, G., and Augustin, T. (2023c). In all likelihoods: Robust selection of pseudo-labeled data. In *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 412–425.

Rodemann, J., Kreiss, D., Hüllermeier, E., and Augustin, T. (2022b). Levelwise data disambiguation by cautious superset classification. In *International Conference on Scalable Uncertainty Management (SUM)*, pages 263–276. Springer.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4):1378–1408.

Roscher, R., Rußwurm, M., Gevaert, C. M., Kampffmeyer, M. C., dos Santos, J. A., Vakalopoulou, M., Hänsch, R., Hansen, S., Nogueira, K., Prexl, J., and Tuia, D. (2023). Better, not just more: Data-centric machine learning for earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 12:335–355.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Ross, S. M. (1995). *Stochastic processes*. Wiley & Sons.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58.

Russell, S. and Norvig, P., editors (2021). *Artificial Intelligence: A Modern Approach*. Pearson, 4 edition.

Rust, K. F. and Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5(3):283–310.

Ryan, T. P. (2007). *Modern Experimental Design*. Wiley & Sons.

Sale, Y., Caprio, M., and Hüllermeier, E. (2023). Is the volume of a credal set a good measure for epistemic uncertainty? In *Uncertainty in Artificial Intelligence (UAI)*, pages 1795–1804.

Sapkota, R., Roumeliotis, K. I., and Karkee, M. (2025). AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468 (last accessed October 29 2025)*.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer, New York.

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical association*, 46(253):55–67.

Scarf, H. E. (1957). A min–max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209.

Schmidt, A. M., Conceição, M. d. F. d. G., and Moreira, G. A. (2008). Investigating the sensitivity of Gaussian processes to the choice of their correlation function and prior specifications. *Journal of Statistical Computation and Simulation*, 78(8):681–699.

Schneider, F., Balles, L., and Hennig, P. (2019). Deepobs: A deep learning optimizer benchmark suite. In *International Conference on Learning Representations (ICLR)*. Amerst, MA. OpenReview.net.

Seedat, N., Crabbé, J., Bica, I., and van der Schaar, M. (2022a). Data-IQ: Characterizing subgroups with heterogeneous outcomes in tabular data. In *Neural Information Processing Systems*, pages 23660–23674.

Seedat, N., Crabbé, J., and van der Schaar, M. (2022b). Data-suite: Data-centric identification of in-distribution incongruous examples. In *International Conference on Machine Learning (ICML)*.

Sen, A. (1985). *Commodities and Capabilities*. Oxford University Press India.

Sen, P. (1992). Introduction to Chernoff (1959): sequential design of experiments. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 339–344. Springer.

Settles, B. (2010). Active learning literature survey. Technical report, University of Wisconsin–Madison. Computer Sciences Technical Report 1648. https://minds.wisconsin.edu/handle/1793/60660 (last accessed October 29 2025).

Shafayat, S., Tajwar, F., Salakhutdinov, R., Schneider, J., and Zanette, A. (2025). Can large reasoning models self-train? *arXiv preprint arXiv.2505.21444* (last accessed October 12 2025).

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. *Neural Information Processing Systems*, 28:1576–1584.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shavit, Y., Edelman, B., and Axelrod, B. (2020). Causal strategic linear regression. In *International Conference on Machine Learning*, pages 8676–8686.

Siegmund, D. (1985). Sequential analysis. *Springer Series in Statistics*.

Simon, H. A. (1954). Bandwagon and underdog effects and the possibility of election predictions. *Public Opinion Quarterly*, 18(3):245–253.

Singh, A., Chau, S. L., Bouabid, S., and Muandet, K. (2024). Domain generalisation via imprecise learning. In *International Conference on Machine Learning*, pages 45544–45570.

Singh, P. (2023). Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*, 6(3):144–157.

Siviy, C., Baker, L. M., Quinlivan, B. T., Porciuncula, F., Swaminathan, K., Awad, L. N., and Walsh, C. J. (2023). Opportunities and challenges in the development of exoskeletons for locomotor assistance. *Nature Biomedical Engineering*, 7(4):456–472.

Smith, F. B., Kossen, J., Trollope, E., van der Wilk, M., Foster, A., and Rainforth, T. (2025). Rethinking aleatoric and epistemic uncertainty. In *International Conference on Machine Learning*.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *27th International Conference on Machine Learning*, pages 1015–1022.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, 2 edition.

Sutton-Charani, N., Destercke, S., and Denœux, T. (2012). Classification trees based on belief functions. In *Belief Functions: Theory and Applications: Proceedings of the 2nd International Conference on Belief Functions, Compiègne, France 9-11 May 2012*, pages 77–84. Springer.

Sutton-Charani, N., Destercke, S., and Denoeux, T. (2013). Learning decision trees from uncertain data with an evidential EM approach. In *International Conference on Machine Learning and Applications (ICMLA)*, volume 1, pages 111–116. IEEE.

Talagrand, M. (2014). *Upper and lower bounds for stochastic processes*. Springer.

Tartar, L. (2007). *An introduction to Sobolev spaces and interpolation spaces*. Springer.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Neural Information Processing Systems*, pages 1195–1204.

Toxiri, S., Näf, M. B., Lazzaroni, M., Fernández, J., Sposito, M., Poliero, T., Monica, L., Anastasi, S., Caldwell, D. G., and Ortiz, J. (2019). Back-support exoskeletons for occupational use: an overview of technological advances and trends. *IISE Transactions on Occupational Ergonomics and Human Factors*, 7(3-4):237–249.

Triguero, I., Sáez, J. A., Luengo, J., García, S., and Herrera, F. (2014). On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. *Neurocomputing*, 132:30–41.

Troffaes, M. C. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29.

Troffaes, M. C. and De Cooman, G. (2014). *Lower previsions*. Wiley & Sons.

Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin*, 63(6):384–399.

Tuckman, B. W. and Jensen, M. A. C. (1977). Stages of small-group development revisited. *Group & Organization Studies*, 2(4):419–427.

Tukey, J. W. (1953). The problem of multiple comparisons. In Braun, H. I., editor, *The Collected Works of John W. Tukey, Volume VIII: Multiple Comparisons, 1948–1983*, pages 1–300. Chapman and Hall, New York. Unpublished manuscript.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6(1):100–116.

Udandarao, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P. H., Bibi, A., Albanie, S., and Bethge, M. (2024). No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint arXiv:2404.04125 (last accessed October 27 2025)*.

Utkin, L. and Konstantinov, A. (2022). Attention-based random forest and contamination model. *Neural Networks*, 154:346–359.

Utkin, L. V. (2020). An imprecise deep forest for classification. *Expert Systems with Applications*, 141:112978.

Utkin, L. V. and Augustin, T. (2003). Decision making with imprecise second-order probabilities. In *International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA)*, pages 545–559.

Utkin, L. V. and Augustin, T. (2005). Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In *International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA)*, pages 349–358.

Utkin, L. V. and Augustin, T. (2007). Decision making under incomplete data using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 44(3):322–338.

Utkin, L. V., Chekh, A. I., and Zhuk, Y. A. (2015). Classification svm algorithms with interval-valued training data using triangular and epanechnikov kernels. In *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 295–303.

Utkin, L. V. and Coolen, F. P. A. (2011). Interval-valued regression and classification models in the framework of machine learning. In *Symposium on Imprecise Probability: Theory and Applications (ISIPTA)*, pages 371–380.

Valliant, R., Dever, J. A., and Kreuter, F. (2018). *Practical tools for designing and weighting survey samples*, volume 2. Springer.

Van Breugel, B., Qian, Z., and Van Der Schaar, M. (2023). Synthetic data, real errors: how (not) to publish and use synthetic data. In *International Conference on Machine Learning*, pages 34793–34808.

Van Engelen, J. E. and Hoos, H. H. (2019). A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Vanschoren, J. (2019). Meta-learning. In Hutter, F., Kotthoff, L., and Vanschoren, J., editors, *Automated machine learning: methods, systems, challenges*, pages 35–61. Springer.

Vapnik, V. (1991). Principles of risk minimization for learning theory. In *Neural Information Processing Systems*, volume 4, pages 831–838.

Vapnik, V. (1998). *Statistical learning theory*. Wiley & Sons.

Vapnik, V. N. and Chervonenkis, A. Y. (1968). The uniform convergence of frequencies of the appearance of events to their probabilities. In *Doklady Akademii Nauk*, volume 181, pages 781–783. Russian Academy of Sciences.

Vettoruzzo, A., Bouguelia, M.-R., Vanschoren, J., Rögnvaldsson, T., and Santosh, K. (2024). Advances and challenges in meta-learning: A technical review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4763–4779.

Vo, K. Q., Aadil, M., Chau, S. L., and Muandet, K. (2024). Causal strategic learning with competitive selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15411–15419.

Vo, K. Q., Chau, S. L., Kato, M., Wang, Y., and Muandet, K. (2025). Explanation design in strategic learning: Sufficient explanations that induce non-harmful responses. *arXiv preprint arXiv:2502.04058 (last accessed October 12 2025)*.

Von Foerster, H. (1952). *Cybernetics; circular causal and feedback mechanisms in biological and social systems*. Josiah Macy, Jr. Foundation.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

Von Stackelberg, H. (2010). *Market structure and equilibrium*. Springer. Translated from the 1934 original by Damien Bazin, Lynn Urch and Rowland Hill.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

Wainer, H., editor (2013). *Drawing Inferences From Self-Selected Samples*. Routledge, New York.

Wald, A. (1945a). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wald, A. (1945b). Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46(2):265–280.

Wald, A. (1947a). Foundations of a general theory of sequential decision functions. *Econometrica*, pages 279–313.

Wald, A. (1947b). *Sequential Analysis*. Wiley & Sons, New York.

Wald, A. (1949). Statistical decision functions. *Annals of Mathematical Statistics*, 20:165–205.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman & Hall.

Walley, P. and Fine, T. L. (1982). Towards a frequentist theory of upper and lower probability. *The Annals of Statistics*, 10(3):741–761.

Walley, P. and Moral, S. (1999). Upper probabilities based only on the likelihood function. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):831–847.

Walters, P. (1982). *An Introduction to Ergodic Theory*, volume 79 of *Graduate Texts in Mathematics*. Springer.

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38:90–95.

Wang, K., Cuzzolin, F., Shariatmadar, K., Moens, D., Hallez, H., et al. (2024). Credal deep ensembles for uncertainty quantification. *Neural Information Processing Systems*, 37:79540–79572.

Wang, Z. and Jegelka, S. (2017). Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning*, pages 3627–3635.

Weber, M. (1904). Die" Objektivität" sozialwissenschaftlicher und sozialpolitischer Erkenntnis. *Archiv für Sozialwissenschaft und Sozialpolitik*, 19(1):22–87.

Weber, M. (1978/1922). *Economy and Society: An Outline of Interpretive Sociology*. University of California Press, Berkeley. Originally published posthumously in 1922.

Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2-3):149–170.

Weichselberger, K. and Augustin, T. (1998). Analysing Ellsberg's paradox by means of interval-probability. In *Econometrics in Theory and Practice: Festschrift for Hans Schneeweiß*, pages 291–304. Springer.

Weichselberger, K. and Augustin, T. (2003). On the symbiosis of two concepts of conditional interval probability. In *International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA)*, pages 606–628.

Weichselberger, K. and Pöhlmann, S. (1990). *A Methodology for Uncertainty in Knowledge-based Systems*. Springer, Heidelberg.

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(4):739–750.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, pages 681–688.

Werner, M., Karimireddy, S. P., and Jordan, M. I. (2024). Defection-free collaboration between competitors in a learning system. *arXiv preprint arXiv:2406.15898 (last accessed October 12 2025)*.

Wheeler, G. (2025). Function-coherent gambles. In *International Symposium on Imprecise Probabilities: Theories and Applications*, pages 285–295.

Wiener, N. (1948). Cybernetics. *Scientific American*, 179(5):14–19.

Wiesemann, W., Kuhn, D., and Sim, M. (2014). Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376.

Wikinews contributors (2025). Wikinews: A free news source. https://en.wikinews.org. last accessed October 27 2025.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT press Cambridge, MA.

Williamson, R. C. (2020). Process and purpose, not thing and technique: How to pose data science research challenges. *Harvard Data Science Review*, 2(3).

Williamson, R. C. (2024). The rhetoric of machine learning. Talk, Persuasive Algorithms? A Symposium on the Rhetoric of Generative AI. Available at https://fm.ls/files/documents/Rhetoric%20of%20ML%20talk.pdf, last accessed October 12 2025.

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., and Hüllermeier, E. (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence (UAI)*, pages 2282–2292.

Witting, H. (2013). *Mathematische Statistik I: Parametrische Verfahren bei Festem Stichprobenumfang*. Springer.

Wooldridge, M. and Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152.

Wu, F., Wang, W., Chen, J., and Wang, Z. (2023). A dynamic multi-objective optimization method based on classification strategies. *Scientific Reports*, 13(1):15221.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z. (2024). Qwen2 technical report. *arXiv preprint arXiv.2407.10671* (last accessed October 12 2025).

Yerushalmy, J. (1972). Self-selection—a major problem in observational studies. In Le Cam, L. M., Neyman, J., and Scott, E. L., editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume IV: Biology and Health, pages 329–342. University of California Press, Berkeley.

Yin, Z., Pu, J., Wan, R., and Xue, X. (2024). Embrace sustainable ai: Dynamic data subset selection for image classification. *Pattern Recognition*, 151:110392.

Yu, B. (2020). Veridical data science. In *ACM International Conference on Web Search and Data Mining*, pages 4–5.

Yu, B. and Barter, R. L. (2024). *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. (2025). Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42.

Zhang, G., Dorner, F. E., and Hardt, M. (2025). How benchmark prediction from fewer data misses the mark. *arXiv preprint arXiv:2506.07673 (last accessed October 12 2025)*.

Zhang, G. and Hardt, M. (2024). Inherent trade-offs between diversity and stability in multi-task benchmarks. In *Proceedings of the 41st International Conference on Machine Learning*, pages 58984–59002.

Zhang, J., Fiers, P., Witte, K. A., Jackson, R. W., Poggensee, K. L., Atkeson, C. G., and Collins, S. H. (2017). Human-in-the-loop optimization of exoskeleton assistance during walking. *Science*, 356(6344):1280–1284.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning (ICML)*, pages 819–827.

Zhang, K. W. (2023). *Statistical Inference for Adaptive Experimentation*. PhD thesis, Harvard University.

Zhang, K. W., Janson, L., and Murphy, S. (2021). Statistical inference with M-estimators on adaptively collected data. *Neural Information Processing Systems*, 34:7460–7471.

Zhang, M.-L. and Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048.

Zhao, X., Kang, Z., Feng, A., Levine, S., and Song, D. (2025). Learning to reason without external rewards. *arXiv preprint arXiv.2505.19590* (last accessed October 12 2025).

Zheng, X., Liu, Y., Bao, Z., Fang, M., Hu, X., Liew, A. W.-C., and Pan, S. (2023). Towards data-centric graph machine learning: Review and outlook. *arXiv preprint arXiv:2309.10979 (last accessed October 27 2025)*.

Zhou, X., Liu, Z., Sims, A., Wang, H., Pang, T., Li, C., Wang, L., Lin, M., and Du, C. (2025). Reinforcing general reasoning without verifiers. *arXiv preprint arXiv.2505.21493* (last accessed October 12 2025).

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics*, 28(2):461–482.